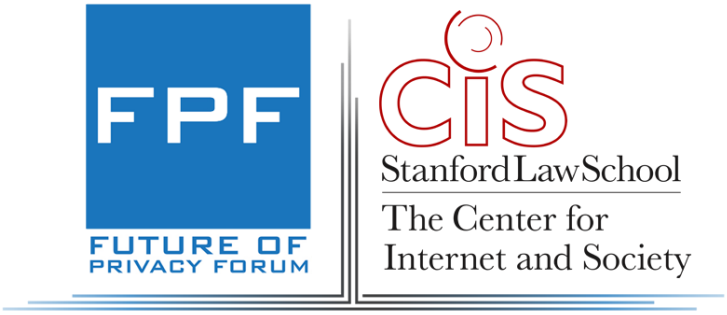
Stanford Law School

The Center for
Internet and Society

BIG DATA AND PRIVACY

MAKING ENDS MEET



BIG DATA AND PRIVACY

MAKING ENDS MEET

Solutions to many pressing economic and societal challenges lie in better understanding data. New tools for analyzing disparate information sets, called Big Data, have revolutionized our ability to find signals amongst the noise. Big Data techniques hold promise for breakthroughs ranging from better health care, a cleaner environment, safer cities, and more effective marketing. Yet, privacy advocates are concerned that the same advances will upend the power relationships between government, business and individuals, and lead to prosecutorial abuse, racial or other profiling, discrimination, redlining, overcriminalization, and other restricted freedoms.

On Tuesday, September 10th, 2013, the Future of Privacy Forum joined with the Center for Internet and Society at Stanford Law School to present a full-day workshop on questions surrounding Big Data and privacy. The event was preceded by a call for papers discussing the legal, technological, social, and policy implications of Big Data. A selection of papers was published in a special issue of the *Stanford Law Review Online* and others were presented at the workshop. This volume collects these papers and others in a single collection.

These essays address the following questions: Does Big Data present new challenges or is it simply the latest incarnation of the data regulation debate? Does Big Data create fundamentally novel opportunities that civil liberties concerns need to accommodate? Can de-identification sufficiently minimize privacy risks? What roles should fundamental data privacy concepts such as consent, context, and data minimization play in a Big Data world? What lessons can be applied from other fields?

We hope the following papers will foster more discussion about the benefits and challenges presented by Big Data—and help bring together the value of data and privacy, as well.

TABLE OF CONTENTS

PRIVACY AND BIG DATA:
MAKING ENDS MEET

*Jules Polonetsky & Omer Tene**

66 STAN. L. REV. ONLINE 25 (2013)

INTRODUCTION

How should privacy risks be weighed against big data rewards? The recent controversy over leaked documents revealing the massive scope of data collection, analysis, and use by the NSA and possibly other national security organizations has hurled to the forefront of public attention the delicate balance between privacy risks and big data opportunities.[1] The NSA revelations crystalized privacy advocates' concerns of "sleepwalking into a surveillance society" even as decisionmakers remain loath to curb government powers for fear of terrorist or cybersecurity attacks.

Big data creates tremendous opportunity for the world economy not only in the field of national security, but also in areas ranging from marketing and credit risk analysis to medical research and urban planning. At the same time, the extraordinary benefits of big data are tempered by concerns over privacy and data protection. Privacy advocates are concerned that the advances of the data ecosystem will upend the power relationships between government, business, and individuals, and lead to racial or other profiling, discrimination, over-criminalization, and other restricted freedoms.

Finding the right balance between privacy risks and big data rewards may very well be the biggest public policy challenge of our time.[2] It calls for momentous choices to be made between weighty policy concerns such as scientific research, public health, national security, law enforcement, and efficient use of resources, on the one hand, and individuals' rights to privacy, fairness, equality, and freedom of speech, on the other hand. It requires deciding whether efforts to cure fatal disease or eviscerate terrorism are worth subjecting human individuality to omniscient surveillance and algorithmic decisionmaking.[3]

Unfortunately, the discussion progresses crisis by crisis, often focusing on legalistic formalities while the bigger policy choices are avoided. Moreover, the debate has become increasingly polarized, with each cohort fully discounting the concerns of the other. For example, in the context of government surveillance, civil libertarians depict the government as pursuing absolute power, while law enforcement officials blame privacy for child pornography and airplanes falling out of the sky. It seems that for privacy hawks, no benefit no matter how compelling is large enough to offset privacy costs, while for data enthusiasts, privacy risks are no more than an afterthought in the pursuit of complete information.

This Essay suggests that while the current privacy debate methodologically explores the *risks* presented by big data, it fails to untangle commensurate *benefits*, treating them as a hodgepodge of individual, business, and government interests. Detailed frameworks have developed to help decisionmakers understand and quantify privacy risk, with privacy impact assessments now increasingly common for government and business undertakings.[4] Yet accounting for *costs* is only part of a balanced value equation. In order to complete a cost-benefit analysis, privacy professionals need to have at their disposal tools to assess, prioritize, and to the extent possible, quantify a project's *rewards*. To be sure, in recent years there have been thorough

* Jules Polonetsky is Co-Chair and Director, Future of Privacy Forum. Omer Tene is Associate Professor, College of Management Haim Striks School of Law, Israel; Senior Fellow, Future of Privacy Forum; Affiliate Scholar, Stanford Center for Internet and Society.

expositions of big data benefits.[5] But the societal value of these benefits may depend on their nature, on whether they are certain or speculative, and on whether they flow to individuals, communities, businesses, or society at large.

The integration of benefit considerations into privacy analysis is not without basis in current law. In fact, it fits neatly within existing privacy doctrine under both the FTC's authority to prohibit "unfair trade practices" in the United States[6] as well as the "legitimate interests of the controller" clause in the European Union data protection directive.[7] Over the past few years, the FTC has carefully recalibrated its section 5 powers to focus on "unfair" as opposed to "deceptive" trade practices. An "unfair" trade practice is one that "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and *is not outweighed by countervailing benefits* to consumers or competition."[8] Clearly, benefit considerations fit squarely within the legal analysis. Moreover, in determining whether an injury is outweighed by countervailing benefits, the FTC typically considers not only the impact on specific consumers but also on society at large.[9]

In the European Union, organizations are authorized to process personal data without an individual's consent based on such organizations' "legitimate interests" as balanced against individuals' privacy rights. In such cases, individuals have a right to object to processing based "on compelling legitimate grounds."[10] Similar to the FTC's "unfairness" doctrine, legitimate interest analysis is inexorably linked to a benefit assessment.

This Essay proposes parameters for a newly conceptualized cost-benefit equation that incorporates both the sizable benefits of big data as well as its attendant costs. Specifically, it suggests focusing on *who* are the beneficiaries of big data analysis, *what* is the nature of the perceived benefits, and with what level of *certainty* can those benefits be realized. In doing so, it offers ways to take account of benefits that accrue not only to businesses but also to individuals and to society at large.

1. BENEFICIARIES

Who benefits from big data? In examining the value of big data, we start by evaluating who is affected by the relevant breakthrough. In some cases, the individual whose data is processed directly receives a benefit. In other cases, the benefit to the individual is indirect. And in many other cases, the relevant individual receives no attributable benefit, with big data value reaped by business, government, or society at large.

A. INDIVIDUALS

In certain cases, big data analysis provides a direct benefit to those individuals whose information is being used. This provides strong impetus for organizations to argue the merits of their use based on their returning value to affected individuals. In a previous article, we argued that in many such cases, relying on individuals' choices to legitimize data use rings hollow given well-documented biases in their decisionmaking processes.[11] In some cases, a particular practice may be difficult to explain within the brief opportunity that an individual pays attention, while in others, individuals may decline despite their best interests. Yet it would be unfortunate if failure to obtain meaningful consent would automatically discredit an information practice that directly benefits individuals.

Consider the high degree of customization pursued by Netflix and Amazon, which recommend films and products to consumers based on analysis of their previous interactions. Such data analysis directly benefits consumers and has been justified even without solicitation of explicit consent. Similarly, Comcast's decision in 2010 to proactively monitor its customers' computers to detect malware,[12] and more recent decisions by Internet service providers including Comcast, AT&T, and Verizon to reach out to consumers to report potential malware infections, were intended to directly benefit consumers.[13] Google's autocomplete and translate functions are based on comprehensive data collection and real time keystroke-by-keystroke analysis. The value proposition to consumers is clear and compelling.

In contrast, just *arguing* that data use benefits consumers will not carry the day. Consider the challenges that proponents of behavioral advertising have faced in persuading regulators that personalized ads deliver direct benefits to

individuals. Behavioral ads are served by grouping audiences with specific web surfing histories or data attributes into categories, which are then sold to advertisers using algorithms designed to maximize revenue. Consumers may or may not perceive the resulting ads as relevant, and even if they do, they may not appreciate the benefit of being targeted with relevant ads.

B. COMMUNITY

In certain cases, the collection and use of an individual's data benefits not only that individual, but also members of a proximate class, such as users of a similar product or residents of a geographical area. Consider 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. Those users who do agree to send crash reports benefit not only themselves, but also other users of the same product. Similarly, individuals who report drug side effects confer a benefit to other existing and prospective users.[14]

C. ORGANIZATIONS

Big data analysis often benefits those organizations that collect and harness the data. Data-driven profits may be viewed as enhancing allocative efficiency by facilitating the "free" economy.[15] The emergence, expansion, and widespread use of innovative products and services at decreasing marginal costs have revolutionized global economies and societal structures, facilitating access to technology and knowledge[16] and fomenting social change.[17] With more data, businesses can optimize distribution methods, efficiently allocate credit, and robustly combat fraud, benefitting consumers as a whole.[18] But in the absence of individual value or broader societal gain, others may consider enhanced business profits to be a mere value transfer from individuals whose data is being exploited. In economic terms, such profits create distributional gains to some actors (and may in fact be socially regressive) as opposed to driving allocative efficiency.

D. SOCIETY

Finally, some data uses benefit society at large. These include, for example, data mining for purposes of national security. We do not claim that such practices are always justified; rather, that when weighing the benefits of national security driven policies, the effects should be assessed at a broad societal level. Similarly, data usage for fraud detection in the payment card industry helps facilitate safe, secure, and frictionless transactions, benefiting society as a whole. And large-scale analysis of geo-location data has been used for urban planning, disaster recovery, and optimization of energy consumption.

E. BENEFITS

Big data creates enormous value for the global economy, driving innovation, productivity, efficiency, and growth. Data has become the driving force behind almost every interaction between individuals, businesses, and governments. The uses of big data can be transformative and are sometimes difficult to anticipate at the time of initial collection. And any benefit analysis would be highly culture-specific. For example, environmental protection may be considered a matter of vital importance in the United States, but less so in China.

In a recent article titled *The Underwhelming Benefits of Big Data*, Paul Ohm critiques our previous articles, arguing that "Big Data's touted benefits are often less significant than claimed and less necessary than assumed."[19] He states that while some benefits, such as medical research, are compelling, others yield only "minimally interesting results."[20] He adds, "Tene and Polonetsky seem to understand the speciousness of some of the other benefits they herald."[21]

While we agree that society must come up with criteria to evaluate the relative weight of different benefits (or social values), we claim that such decisions transcend privacy law. The social value of energy conservation, law enforcement, or economic efficiency is a meta-privacy issue that requires debate by experts in the respective fields. If privacy regulators were the sole decision-makers determining the relative importance of values that sometimes conflict with privacy, such as free speech, environmental protection, public health, or national security, they would become the *de facto* regulators of all things commerce, research, security, and speech.[22] This would be a perverse result, given that even where privacy constitutes a

fundamental human right, it is not an "*über-value*" that trumps every other social consideration.

This Essay does not provide a comprehensive taxonomy of big data benefits. It would be pretentious to do so, ranking the relative importance of weighty social goals. Rather it posits that such benefits must be accounted for by rigorous analysis taking into account the priorities of a nation, society, or culture. Only then can benefits be assessed *within* the privacy framework.

Consider the following examples of countervailing values (*i.e.*, big data benefits) as they are addressed, with little analytical rigor, by privacy regulators. For example, despite intense pushback from privacy advocates, legislative frameworks all over the world give national security precedence over privacy considerations.[23] On the other hand, although mandated by corporate governance legislation in the United States, whistleblower hotlines are not viewed by privacy regulators as worthy of deference.

What is the doctrinal basis for accepting national security as a benefit that legitimizes privacy costs, while denying the same status to corporate governance laws? Such selective, apparently capricious enforcement is detrimental for privacy. Regulators should pursue a more coherent approach, recognizing the benefits of big data as an integral part of the privacy framework through legitimate interest analysis under the European framework or unfairness doctrine applied by the FTC.

F. Certainty

The utility function of big data use depends not only on absolute values, but also on the *probability* of any expected benefits and costs. Not every conceivable benefit, even if highly likely, justifies a privacy loss. Legitimate interest analysis should ensure that lack of certainty of expected benefits is a discounting factor when weighing big data value.

A given level of uncertainty may weigh differently depending on the risk profile of a given culture or society. The United States, for example, established by explorers who pushed the frontier in a lawless atmosphere, continues to highly reward entrepreneurship, innovation, research, and discovery. The quintessential American hero is the lone entrepreneur who against all odds weaves straw into gold. This environment may—and to this day in fact does—endorse practically unfettered data innovation, except in certain regulated areas such as health and financial information, or in cases of demonstrable harm. Failure is considered valuable experience and entrepreneurs may be funded many times over despite unsuccessful outcomes. Conversely, in Europe, the departure point is diametrically opposite, with data processing being prohibited unless a legitimate legal basis is shown.

To critics on either side of the Atlantic, both the U.S. and E.U. approaches have their shortcomings. Taken to their extremes, the E.U. approach, with its risk aversion and regulatory bureaucracy, could stifle innovation and growth of a vibrant technology sector, while the U.S. approach, with its *laissez faire* ideology, risks a rude awakening to a reality of eerie surveillance and technological determinism.

Conclusion

This symposium issue sets the stage for a discussion of big data that recognizes the weighty considerations on both sides of the value scale. The authors deploy different lenses to expose diverse aspects of the big data privacy conundrum. Some authors focus on the macro, debating broad societal effects: Cynthia Dwork and Deirdre Mulligan discuss the impact of big data on classification, discrimination, and social stratification.[24] Neil Richards and Jonathan King uncover three paradoxes underlying the power structure of the big data ecosystem.[25] Joseph Jerome warns that big data may be socially regressive, potentially exacerbating class disparities.[26] Jonas Lerman examines the overlooked costs of being excluded from big data analysis, suffered by "[b]illions of people worldwide [who] remain on big data's periphery."[27] Ian Kerr and Jessica Earle focus on big data's "preemptive predictions," which could reverse the presumption of innocence, upending the power relationships between government and individuals.[28] Other authors concentrate on the micro, focusing on interpersonal relationships in a data-rich environment: Karen Levy argues that big data has transcended the scope of organizational behavior, entering the delicate domain of

individual relationships.[29] Woodrow Hartzog and Evan Selinger predict that absent a robust concept of obscurity, the "data-fication" of personal relationships would strain the social fabric.[30] Other authors seek to harness technology to tame big data effects. Jonathan Mayer and Arvind Narayanan advocate privacy enhancing technologies.[31] Ryan Calo supports organizational measures, such as "consumer subject review boards."[32] Yianni Lagos and Jules Polonetsky stress the importance of a combination of technological and organizational mechanisms to achieve robust de-identification.[33] We hope that the following essays shift the discussion to a more nuanced, balanced analysis of the fateful value choices at hand.

---

[1] Glenn Greenwald, *NSA Collecting Phone Records of Millions of Verizon Customers Daily*, GUARDIAN (June 6, 2013),http://www.guardian.co.uk/world/2013/jun/06/nsa-phone-records-verizon-court-order ; Glenn Greenwald & Ewen MacAskill, *NSA Prism Program Taps in to User Data of Apple, Google and Others*, Guardian (June 7, 2013), http://www.guardian.co.uk/world/2013/jun/06/us-tech-giants-nsa-data.

[2] Ira Rubinstein, *Big Data: The End of Privacy or a New Beginning?*, 3 INT'L DATA PRIVACY L. 74, 77-78 (2013); Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 240-42 (2013).

[3] We are not arguing that these public policy objectives are mutually exclusive. To the contrary, we support the "Privacy by Design" paradigm that aims to integrate privacy safeguards into projects, products, and services. Yet at some point, stark policy choices need to be made—this is where privacy costs need to be balanced against big data benefits. *See* Ann Cavoukian, *Privacy by Design: The Seven Foundational Principles*, INFO. PRIVACY COMM'R, ONT., CAN. (Jan. 2011), http://www.ipc.on.ca/images/resources/7foundationalprinciples.pdf (" *Privacy by Design* seeks to accommodate all legitimate interests and objectives in a positive-sum 'win-win' manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made.").

[4] *See, e.g.*, PRIVACY IMPACT ASSESSMENT 4-5 (David Wright & Paul De Hert eds., 2012); *Privacy Impact Assessments: The Privacy Office Official Guidance*, DEP'T HOMELAND SEC. (June 2010), http://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_guidance_june2010.pdf.

[5] *See, e.g.*, VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013); RICK SMOLAN & JENNIFER ERWITT, THE HUMAN FACE OF BIG DATA (2012); Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63 (2012); *Big Data and Analytics: Seeking Foundations for Effective Privacy Guidance*, CTR. FOR INFO. POL'Y LEADERSHIP (Feb. 2013),http://www.hunton.com/files/Uploads/Documents/News_files/Big_Data_and_Analytics_February_2013.pdf; *Unlocking the Value of Personal Data: From Collection to Usage*, World Econ. F. (Feb. 2013), http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf.

[6] 15 U.S.C. § 45(a)(1) (2011).

[7] Council Directive 95/46, art. 7(f), 1995 O.J. (L 281) 31, 40 (EC), *available at* http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1995:281:0031:0050:EN:PDF.

[8] 15 U.S.C. § 45(n) (emphasis added).

[9] Woodrow Hartzog & Daniel Solove, The FTC and the New Common Law of Privacy (Aug. 19, 2013), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2312913.

[10] Council Directive, *supra* note 7, at art. 14(a).

[11] 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. 281, 285-86 (2012).

[12] Roy Furchgott, *Comcast to Protect Customer's Computers from Malware*, N.Y. TIMES GADGETWISE (Sept. 30, 2010),http://gadgetwise.blogs.nytimes.com/2010/09/30/comcast-to-monitor-customer-computers-for-malware.

[13] Daniel Lippman & Julian Barnes, *Malware Threat to Internet Corralled*, WALL ST. J. (July 9, 2012),http://online.wsj.com/article/SB10001424052702303292204577515262710139518.html.

[14] Nicholas P. Tatonetti et al., *A Novel Signal Detection Algorithm for Identifying Hidden Drug-Drug Interactions in Adverse Event Reports*, 19 J. AM. MED. INFORMATICS ASS'N 79, 79-80 (2012).

[15] CHRIS ANDERSON, FREE: THE FUTURE OF A RADICAL PRICE (2009).

[16] Tim Worstall, *More People Have Mobile Phones than Toilets*, FORBES (Mar. 23, 2013), http://www.forbes.com/sites/timworstall/2013/03/23/more-people-have-mobile-phones-than-toilets.

[17] WAEL GHONIM, REVOLUTION 2.0: THE POWER OF THE PEOPLE IS GREATER THAN THE PEOPLE IN POWER: A MEMOIR (2012).

[18] *A Different Game: Information Is Transforming Traditional Businesses*, ECONOMIST (Feb. 25, 2010), http://www.economist.com/node/15557465.

[19] Paul Ohm, *The Underwhelming Benefits of Big Data*, 161 U. PA. L. REV. ONLINE 339, 340 (2013).

[20] *Id.* at 344.

[21] *Id.*

[22] Currently, privacy regulators appear to be making almost arbitrary decisions when it comes to balancing privacy risks against potential data rewards. In fact, the recent Opinion of the Article 29 Working Party, which required national regulators to assess compatibility "on a case-by-case basis[,]" appears to legitimize an unpredictable decisionmaking process. *Opinion of the Data Protection Working Party on Purpose Limitation*, (Apr. 2, 2013), *available at* http://ec.europa.eu/justice/data-

protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[23] *See, e.g.*, Data Protection Act, 1998, c. 29, § 28 (U.K.).

[24] Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35 (2013).

[25] Neil M. Richards & Jonathan H. King, *Three Paradoxes of Big Data*, 66 STAN. L. REV. ONLINE 41 (2013).

[26] Joseph W. Jerome, *Buying and Selling Privacy: Big Data's Different Burdens and Benefits*, 66 STAN. L. REV. ONLINE 47 (2013).

[27] Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55 (2013).

[28] Ian Kerr & Jessica Earle, *Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy*, 66 STAN. L. REV. ONLINE 65 (2013).

[29] Karen E.C. Levy, *Relational Big Data*, 66 STAN. L. REV. ONLINE 73 (2013).

[30] Woodrow Hartzog & Evan Selinger, *Big Data in Small Hands*, 66 STAN. L. REV. ONLINE 81 (2013).

[31] Jonathan Mayer & Arvind Narayanan, *Privacy Substitutes: A Thought Experiment*, 66 STAN. L. REV. ONLINE 89 (2013).

[32] Ryan Calo, *Consumer Subject Review Boards*, 66 STAN. L. REV. ONLINE 97 (2013).

[33] Yianni Lagos & Jules Polonetsky, *Public vs. Nonpublic Data: The Benefits of Administrative Controls*, 66 STAN. L. REV. ONLINE 103 (2013).

S-M-L-XL DATA:
BIG DATA AS A NEW INFORMATIONAL PRIVACY PARADIGM

*Michael Birnhack**

Can informational privacy law survive Big Data? A few scholars have pointed to the inadequacy of the current legal framework to Big Data, especially the collapse of notice and consent, the principles of data minimization and data specification.[1] These are first steps, but more is needed.[2] One suggestion is to conceptualize Big Data in terms of property:[3] Perhaps data subjects should have a property right in their data, so that when others process it, subjects can share the wealth. However, privacy has a complex relationship with property. Lawrence Lessig's 1999 proposal to propertize personal data, was criticized: instead of more protection, said the critics, there will be more commodification.[4] Does Big Data render property once again a viable option to save our privacy?

To better understand the informational privacy implications of Big Data and evaluate the property option, this comment undertakes two paths. *First*, I locate Big Data as the newest point on a continuum of Small-Medium-Large-Extra Large data situations. This path indicates that Big Data is not just "more of the same", but a new informational paradigm. *Second*, I begin a query about the property/privacy relationship, by juxtaposing informational privacy with property, real and intangible, namely copyright. This path indicates that current property law is unfit to address Big Data.

S-M-L-XL

Context is a key term in current privacy studies. Helen Nissenbaum suggested that in order to evaluate the privacy implications of socio-technological systems, we should ask how these systems affect the informational norms of a given context.[5] This notion fits within the American reasonable expectations test, which indicates whether the interest in a particular context is worthy of legal protection.[6] Accordingly, I draw a continuum of data contexts, and briefly explore several parameters: the archetypical players, their relationship, the volume, source and kind of data, and the kind of privacy harm that is at stake. For each situation, I note the current legal response.

The continuum is not a neat or rigid classification. The points are indicators of a context. The purpose is to show the development of the contexts, culminating with Big Data. Importantly, the appearance of a new point does not negate or exclude previous points. Big Data raises new challenges, but old and familiar contexts have not elapsed.

Small. The typical Small Data situation assumes one party, usually and individual, that harms another person regarding one informational bit, such as disclosure of a private fact. The data subject and the adversary, to borrow computer scientists' parlance, might have a prior relationship (*e.g.*, family members, neighbors, colleagues), or they are in close proximity: physically (Peeping Tom), socially (a Facebook friend's friend), or commercially (a seller).

Privacy torts developed with small data in mind, and form the common denominator of Warren and Brandies' definition of privacy as the right to be let alone,[7] and Dean Prosser's privacy torts

* Professor of Law, Faculty of Law, Tel-Aviv University. Thanks to Omer Tene and Eran Toch for helpful comments and to Dan Largman for research assistance. The research was supported by ISF Grant 1116/12 and Grant 873051-3-9770 of the Israeli Ministry of Science & Technology.

classification.[8] The law attempts to prevent the harm caused to one's need in having a backstage, either physically or mentally. The parties' proximity means that social norms might also be effective.

Medium. Here too there are two parties. The difference is the volume of the data and the balance of power. Unlike the one-time intrusion in the Small Data context, the adversary, now called a data controller, accumulates data and uses it over time. The result is a specified database, created and managed for one purpose, and not further transferred. Typically, the controller is stronger than the subject. Examples are a school that collects data about students, an employer vs. employees, insurance company vs. customers. The technology used can be as simple as a sheet of paper.

In the United States, specific sector-based federal laws apply, *e.g.*, the Family Educational Rights and Privacy Act (FERPA), regulating students' records.[9] The law attempts to assure that the data is not misused. The data controller's legitimate interests are taken into consideration. For example, in the workplace context, the employer has an interest in protecting trade secrets. Thus, the law carves exceptions, limitations, and undertakes various forms of balancing. When the controller is the government, constitutional checks are in operation, under the Fourth Amendment.

Large. As of the 1970s, with technological advancements, it is easier to collect separate bits of data. The volume is much larger, controllers are no longer close to the subjects, and the parties' inequality is enhanced. The paradigmatic situation is a single data collector that processes personal data of many subjects in one database, uses it for multiple purposes, and transfers it to third parties.

Social norms are no longer effective. The concern shifts from the bit to the byte, and then to the megabyte, namely, the database. Once personal data enters a database, the subject can hardly control it. The database contains information without the subject knowing what kinds of data are kept, or how it is used. Moreover, databases may be maliciously abused (nasty hackers, commercial rivals, or enemies), abused to discriminate (by the state, employers, insurance companies, etc.), or reused for new purposes, without the data subject's consent.

The legal response was a new body of law, now called informational privacy or data protection. It assumes that the concentration of the data is dangerous *per se*. Data protection law originated in the 1970s, with the American Ware Report and the Privacy Act of 1974 being an important step,[10] continuing with the influential OECD Guidelines in 1980,[11] and now carried globally by the 1995 EU Data Protection Directive.[12] The common denominator is Fair Information Privacy Principles (FIPPs) that provide data subjects with some (limited) tools to control personal data: notice, consent, limitations on the use of the data, subjects' rights of access and rectification, and the controllers' obligations to confidentiality and data security. In the United States there is a series of sector-based and/or content-based laws that regulate specific contexts. While much of the law is phrased in technologically-neutral language, a close reading reveals that it assumes Large Data.[13]

Extra-Large. Once megabytes turned into terabytes, the risk to personal data shifted yet again. This is Big Data. The volume staggers. There are multiple adversaries. Personal data is gathered from a variety of sources. Data subjects provide a constant stream of accurate, tiny bits of everything they do. It is not always clear who is the data controller. The kind of control also changes. Under Large Data, the way the database was structured mattered. Sensitive kinds of data could be deleted, anonymized, or not collected at all. In contrast, under Big Data, every bit is welcome. The controller does not need to arrange the data at all: all bits are thrown together into one huge bucket. The original context doesn't matter. Bits are constantly collected, taken out of their original context, and mixed. Data is decontextualized only to recontextualize it in a different way. The notion of context-specific laws collapses. Current (mostly European) rules would simply prohibit much of XL databases that contain data about identifiable people.[14] Notice and consent per-use are impossible; Big Data operates under a maximization principle rather

than a minimization principle. The law breaks down.

PROPERTY/PRIVACY

The property option seems quite tempting. In order to share the wealth, we should be able to protect the wealth in the first place. However, current property law that addresses intangible assets, namely copyright law, does not provide the answer. Here is an outline of the privacy/property juxtaposition along the S-M-L-XL continuum.

S. Property and privacy may overlap. If my home is mine, I can effectively exclude unauthorized intruders and reasonably protect my privacy. The Supreme Court recently concluded that the government's use of drug-sniffing dogs is a "search" under the Fourth Amendment. The Court conducted a property analysis; Justice Kagan's concurrence reached the same conclusion under a privacy analysis.[15] However, privacy and property do not always overlap, as the law protects people, not places.[16]

S., M. From a copyright perspective, for both Small and Medium contexts, the single bit of data does not qualify as proper subject matter. It is an unprotected fact.[17] Neither the data subject nor the controller can rely on copyright law. Without protected property, it is difficult to share the wealth.

L. Real property is irrelevant here. Copyright law may protect the database as a whole, if the selection and arrangement of the facts are original.[18] The individual bits of data remain unprotected. The subject has no property in her personal data, but the data controller might have a property right in the aggregated data. Once the database is protected, there is a reference point for sharing the wealth: it is easier to track down how personal data is processed and used.

XL. Copyright law does not provide the controller with legal protection: the data is not arranged in any particular form, let alone in any original way. Unstructured databases fall outside copyright's subject matter. The controller should seek alternative ways for effective control: the use of technological protection measures is one possible avenue, and to the extent that one undertakes reasonable means to keep the data confidential, trade secret law might be another avenue.[19]

*

The continuum of S-M-L-XL data highlights the special characteristics of each data context, the different legal answers, and the ultimate collapse of context under Big Data. Nevertheless, the appearance of Big Data does not mean that previous sizes are eliminated: privacy law is still relevant for the other contexts.

Property law, occasionally suggested as a possible solution for the privacy concerns, is unlikely to offer comfort. Copyright law does not protect the data subject or the controller. Trade secret law might enable the latter effective control, but not assist the data subject.

---

[1] Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013); Ira S. Rubinstein, *Big Data: The End of Privacy or a New Beginning?*, INTERNATIONAL DATA PRIVACY LAW 1 (January 25, 2013).

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

[3] See *e.g.*, Rubinstein, *Big Data*, supra note 1, at 8; OECD, Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value, 220 OECD DIGITAL ECONOMY PAPERS 35 (2013), available at http://dx.doi.org/10.1787/5k486qtxldmq-en.

[4] See LAWRENCE LESSIG CODE AND OTHER LAWS OF CYBERSPACE 159-62 (1999). For criticism, see *e.g.*, Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as an Object*, 52 STAN. L. REV. 1373 (2000). For a complex property/privacy analysis, see Paul M. Schwartz, *Property, Privacy and Personal Data*, 117 HARV. L. REV. 2055 (2004).

[5] HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY AND THE INTEGRITY OF SOCIAL LIFE (2010).

[6] Katz v. United States, 389 U.S. 347, 360 (1967).

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

[8] William L. Prosser, *Privacy (A Legal Analysis)*, 48 CAL. L. REV. 383, 422 (1960).

[9] 20 U.S.C. §1232g.

[10] 5 U.S.C. § 552a.

[11] Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980).

[12] Council Directive 95/46, 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) (EC).

[13] Michael D. Birnhack, *Reverse Engineering Informational Privacy Law*, 15 YALE J.L. & TECH. 24, 87-89 (2013).

[14] Article 29 Data Protection Working Party, Opinion 03/2013 on Purpose Limitation (April 2013) at 35, 45-46.

[15] Florida v. Jardines, 569 U.S. __ (2013).

[16] Katz, 389 U.S. at 351.

[17] 17 U.S.C. §102(b).

[18] 17 U.S.C. §103(b); Feist Publications, Inc. v. Rural Telephone Service Company, Inc., 499 U.S. 340 (1991).

[19] For an early suggestion, in a Large Data context, see Pamela Samuelson, *Privacy as Intellectual Property?*, 52 STAN. L. REV. 1125 (2000).

WHY SURVEILLANCE MATTERS:

SURVEILLANCE AS A DE FACTO PRIVACY HARM

*Justin Brookman & G.S. Hans**

Consumer privacy remains one of the most pressing issues in technology policy. The interactions between individuals and service providers generate a great deal of data, much of it personally identifiable and sensitive. Individual users are transacting more and more data online with each passing year, and companies have begun exploring what insights and lessons they can glean from consumer data, via storage, processing, and analysis of exceedingly large data sets. These practices, loosely described as *big data*, have raised questions regarding the appropriate balance of control between individuals and companies, and how best to protect personal privacy interests.

In terms of privacy protection, some theorists have insisted that advocates must articulate a concrete harm as a prerequisite for legislated rules, or even self-regulation. Others have argued that privacy protections should focus exclusively on curtailing controversial uses rather than on the collection of personal information.

This paper argues that consumers have a legitimate interest in the mere collection of data by third parties. That is, big data collection practices *per se*, rather than bad uses or outcomes, are sufficient to trigger an individual's privacy interests.[1] Today, big data collection practices are for the most part unregulated. As collection, retention, and analysis practices become increasingly sophisticated, these threats will only increase in magnitude, with a concomitant chilling effect on individual behavior and free expression.

I. The Interests Implicated by Data Collection

Commercial collection of personal information necessarily implicates a range of potential threats that should be considered when evaluating the need for collection limitations. This paper focuses on five particular threat models: data breach, internal misuse, unwanted secondary use, government access, and chilling effect on consumer behavior. These scenarios are for the most part outside corporate control — and indeed, contrary to corporate interest — and can never be fully mitigated by internal procedures. As big data becomes more pervasive, the susceptibility of consumer data to these threats will undoubtedly increase.

A. Data Breach

One of the most common threats that arise from the mere collection of personal information is data breach. Companies consistently experience data breaches, either due to inadequate security or aggressive external hacking. As companies collect an increasing amount of data and retain it for future uses, the consequences of a breach become more severe — both for the company and for consumers. Moreover, the more robust a company's database is, the more appealing it may be for malicious actors. The risk of breach will necessarily increase as companies collect more data about their consumers.

* Justin Brookman is Director of Consumer Privacy at the Center for Democracy & Technology. G.S. Hans is the 2012-14 Ron Plesser Fellow at the Center for Democracy & Technology.

The consequences of data breach are obvious. Personal information, including real name, contact information, financial information, health data, and other sensitive data, can fall into the wrong hands. Consumers can therefore become susceptible to financial fraud or inadvertent identification by third parties. However, this interest extends beyond the potential for economic loss; data breach could also reveal private, embarrassing information that a consumer did not want shared with others or published to the world. For this reason, the Federal Trade Commission has increasingly found substantial harm arising from less sensitive disclosures, such as "revealing potentially embarrassing or political images"[2] "impair[ing consumers'] peaceful enjoyment of their homes,"[3] allowing hackers to "capture private details of an individual's life,"[4] and "reduc[ing consumers'] ability to control the dissemination of personal or proprietary information (e.g., voice recordings or intimate photographs)."[5]

B. Internal Misuse

Internal misuse by rogue employees — data voyeurism — is another significant threat implicated by commercial collection of personal data. While the scale of such misuse of data would probably be markedly smaller than a data breach (which would likely be conducted by an external party), employees may possess a more focused desire to access individualized data than external hackers. For example, in one prominent case, an engineer spied on the user accounts of multiple minors, including contact lists, chat transcripts, and call logs, and used that information to manipulate the users whose accounts he had accessed.[6] Consumer reliance on cloud services to store and transmit their personal communications necessarily involves an opportunity for rogue individuals employed by those cloud services to access such data, unless the data is fully encrypted, and the companies do not have access to the encryption keys.

C. Unwanted Secondary Usage and Changes in Company Practices

Companies that collect personal information may decide to use that information in ways that are inimical to consumers' interests. Such usage could range from the merely annoying (say, retargeted advertising) to price discrimination to selling the information to data brokers who could then use the information to deny consumers credit or employment.

Even if companies do not engage in such unwanted uses right away, they may subsequently change their minds. Although the FTC has asserted for years that material retroactive changes to privacy policies constitutes deceptive and unfair business practices,[7] that legal theory has only rarely been tested in court. Moreover, in the United States, companies are not legally required to justify and explain all data usage practices at the time of collection. Companies could in a privacy policy reserve broad rights to utilize data (or potentially just remain silent on the issue), and subsequently repurpose that information without providing notice or an opportunity to opt out of such usage to the user.

D. Government Access

Government access without robust due process protection is arguably the most significant threat posed by the collection of personal information. As the recent NSA revelations aptly demonstrate, much of the data that governments collect about us derives not from direct observation, but from access to commercial stores of data. Even in so-called rule of law jurisdictions such as the United States and Europe, that data is often obtained without transparent process, and without a particularized showing of suspicion — let alone probable cause as determined by an independent judge. Unfortunately, there is almost nothing that consumers can do to guard against such access or in many cases even know when it occurs.[8]

E. Chilling Effects

Finally, all these concerns together —along with others, and even with an irrational or inchoately realized dislike of being observed — has a chilling effect on public participation and free expression. Consumers who don't want to be monitored all the time may be resistant to adopting new technologies; indeed, the Obama administration used this as an explicit

commercial justification in calling for the enactment of comprehensive commercial privacy protections.[9]

More fundamentally, however, citizens who fear that they are being constantly observed may be less likely to speak and act freely if they believe that their actions are being surveilled. People will feel constrained from experimenting with new ideas or adopting controversial positions. In fact, this constant threat of surveillance was the fundamental conceit behind the development of the Panopticon prison: if inmates had to worry all the time that they were being observed, they would be less likely to engage in problematic behaviors.[10] Big Data transposes this coercive threat of constant observation to everyday citizens.

The United States was founded on a tradition of anonymous speech. In order to remain a vibrant and innovative society, citizens need room for the expression of controversial — and occasionally wrong — ideas without worry that the ideas will be attributable to them in perpetuity. In a world where increasingly every action is monitored, stored, and analyzed, people have a substantial interest in finding some way to preserve a zone of personal privacy that cannot be observed by others.

II. Internal Controls Are Necessary — But Not Sufficient

When faced with these threat models, some have argued that they can be sufficiently addressed by internal organizational controls — such as privacy by design, accountability mechanisms, and use limitations. However, of the above threats, only unwanted secondary usage can be fully solved by internal controls, as deliberate secondary usage is the only threat model fully within the control of the organization. Even then, if the data is retained, the organization could eventually change its mind if the internal controls weaken, ownership is transferred, or the organization is dissolved and its assets liquidated.[11]

Data breach, internal misuse, and government access all derive from extra-corporate motivations, and cannot be definitively prevented so long as the data remains within the company's control. Adherence to best practices and strict protections can diminish the threat of data breach and internal misuse, but cannot wholly prevent them. When it comes to government access, internal controls are even less effective. Companies may engage in heroic efforts to prevent disclosure of customer records, but ultimately they can be beholden by law to comply.[12]

Empirically, internal privacy programs have proven to be insufficient to prevent privacy violations. Many of the companies cited to date by the FTC, state Attorneys General, and private suits have been large companies with mature and far-ranging privacy compliance mechanisms in place. Despite these state-of-the-art programs, those companies either lost control of the data or internally justified privacy-invasive practices.

Moreover, internal controls are completely opaque and indistinguishable to the average user, rendering them rather ineffective in diminishing the chilling effect of surveillance. However, as noted above, even if consumers could discern and evaluate the full range of internal controls over their data, their fears would not be assuaged.[13]

III. Conclusion

The ambition of this paper is deliberately modest. We merely endeavor to articulate (beyond allegations of *creepiness*) why consumers have a privacy interest in controlling commercial collection of their personal information, rather than relying entirely on best practices in use limitations. We do not mean to argue that this interest should always outweigh legitimate commercial interests in that same data, or that consumers' interest always necessitates express consent for all data collection. However, it is an important interest, deserving of consideration in evaluating the appropriate framework for commercial data protection.

---

[1] Setting aside entirely the issue of whether consumers have privacy *rights* over their data, which this paper does not address.

[2] Facebook Inc., Docket No. C-4365, File No. 0923184 (Fed. Trade Comm'n July 27, 2012) (complaint), http://www.ftc.gov/os/caselist/0923184/120810facebookcmpt.pdf.

[3] Aspen Way Enterprises, Inc., Docket No. C-4392, File No. 1123151 (Fed. Trade Comm'n Apr. 15, 2013) (complaint), http://www.ftc.gov/os/caselist/1123151/aspenway/130415aspenwaycmpt.pdf.

[4] HTC America Inc., File No. 122 3049 (Fed. Trade Comm'n February 22, 2013) (complaint), http://www.ftc/gov/os/caselist/1223049/130222htccmpt.pdf.
[5] Frostwire LLC, Docket No. 23643 , File No. 112 3041(Fed. Trade Comm'n October 11, 2011) (complaint), http://www.ftc.gov/os/caselist/1123041/111011frostwirecmpt.pdf.

[6] Adrian Chen, *GCreep: Google Engineer Stalked Teens, Spied on Chats,* Gawker (Sept. 14, 2010, 3:26 PM) http://gawker.com/5637234/gcreep-google-engineer-stalked-teens-spied-on-chats.

[7] Gateway Learning Corp., File No. 042-3047, (Fed. Trade Comm'n September 17, 2004) (complaint), http://www.ftc.gov/os/caselist/0423047/040917comp0423047.pdf.

[8] For a more expansive exploration of the privacy threats implicated by government surveillance, *see* Daniel J. Solove, *Nothing to Hide: The False Tradeoff between Privacy and Security* (2011).

[9] 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, http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

[10] Michel Foucault, *Discipline and Punish: The Birth of the Prison* (1977).

[11] Toysmart.com, LLC, Docket No. 00-11341-RGS, File No. X000075 (Fed. Trade Comm'n July 21, 2000) (complaint), http://www.ftc.gov/os/2000/07/toysmartcomplaint.htm.
[12] Claire Cain Miller, *Secret Court Ruling Put Tech Companies in Data Bind*, N.Y. Times (June 14, 2013), at A1, *available at* http://www.nytimes.com/2013/06/14/technology/secret-court-ruling-put-tech-companies-in-data-bind.html.

[13] Which is not to say that internal controls are not privacy-enhacing, or indeed essential, to preserving data that has been collected. Moreover, some internal controls are more effective than others. Data deletion (and to a lesser extent aggregation and anonymization) is almost certainly the most effective internal control in eliminating the privacy threat posed by static stores of consumer data. Even then, consumers likely have imperfect visibility into internal deletion practices, and may not fully trust in the adequacy of companies' deletion or deidentification techniques. That said, strong data deletion policies are probably the most effective way to address the harms of collection after the fact.

CONSUMER SUBJECT REVIEW BOARDS:
A THOUGHT EXPERIMENT

*Ryan Calo**

66 STAN. L. REV. ONLINE 97 (2013)

The adequacy of consumer privacy law in America is a constant topic of debate. The majority position is that United States privacy law is a "patchwork," that the dominant model of notice and choice has broken down,[1] and that decades of self-regulation have left the fox in charge of the henhouse.

A minority position chronicles the sometimes surprising efficacy of our current legal infrastructure. Peter Swire describes how a much-maligned disclosure law improved financial privacy not by informing consumers, but by forcing firms to take stock of their data practices.[2] Deirdre Mulligan and Kenneth Bamberger argue, in part, that the emergence of the privacy professional has translated into better privacy on the ground than what you see on the books.[3]

There is merit to each view. But the challenges posed by big data to consumer protection feel different. They seem to gesture beyond privacy's foundations or buzzwords, beyond "fair information practice principles" or "privacy by design." The challenges of big data may take us outside of privacy altogether into a more basic discussion of the ethics of information.[4] The good news is that the scientific community has been heading down this road for thirty years. I explore a version of their approach here.

Part I discusses why corporations study consumers so closely, and what harm may come of the resulting asymmetry of information and control. Part II explores how established ethical principles governing biomedical and behavioral science might interact with consumer privacy.

I. RATIONALES FOR STUDYING BEHAVIOR

There are only a handful of reasons to study someone very closely. If you spot a tennis rival filming your practice, you can be reasonably sure that she is studying up on your style of play. Miss too many backhands and guess what you will encounter come match time. But not all careful scrutiny is about taking advantage. Doctors study patients to treat them. Good teachers follow students to see if they are learning. Social scientists study behavior in order to understand and improve the quality of human life.

Why do corporations study consumers? An obvious reason is to figure out what consumers want so as to be in a position to deliver it—hopefully better and cheaper than a competitor. I assume the reason that Microsoft employs the second greatest number of anthropologists in the world (after the United States government)[5] has to do with designing intuitive and useful software. But is that the only reason companies study consumers? And if not, how should we think about consumers as subjects of scientific scrutiny?

Were you to play the market equivalent of tennis against a corporation, it seems fair to think you would lose. They have several advantages. The first advantage is superior information. The websites and stores you visit gather whatever data

* Assistant Professor, University of Washington School of Law; Faculty Director, the Tech Policy Lab at the University of Washington; Affiliate Scholar, Stanford Law School Center for Internet and Society.

they can about you and may supplement that information with profiles they purchase from third-party data brokers.[6] They also run data through powerful algorithms in a constant quest for novel insight.[7] The second advantage is that firms tend to control the circumstances of their transactions with consumers, sometimes entirely. Apple does not divulge its preferences and travel to a website *you* created from scratch in order to sell you music.[8] Firms hire people with advanced degrees and give them access to cutting-edge technology and rich datasets. These people write the legal terms and design the virtual and physical spaces in which our interactions with the firms occur.

Such advantages are fine in a win-win situation. The truth, however, is that sometimes consumers lose. The well-documented use of software by banks to maximize consumer overdraft fees by manipulating when ATM and debit transactions get processed is a simple enough example.[9] But pause to consider the full universe of possibility. Recent research suggests that willpower is a finite resource that can be depleted or replenished over time.[10] Imagine that concerns about obesity lead a consumer to try to hold out against her favorite junk food. It turns out there are times and places when she cannot. Big data can help marketers understand exactly how and when to approach this consumer at her most vulnerable—especially in a world of constant screen time in which even our appliances are capable of a sales pitch.[11]

If this sort of thing sounds far-fetched, consider two recent stories published by the *New York Times*. The first article—obligatory in any discussion of big data and privacy—focuses on how the retail giant Target used customer purchase history to determine who among its customers was pregnant, following which Target added ads related to babies in their direct marketing to those customers.[12] A second article describes the "extraordinary" lengths to which food manufactures go to scientifically engineer craving.[13] Either story alone raises eyebrows. But taken together they bring us closer than is comfortable to the scenario described in the previous paragraph.

My current writing project, *Digital Market Manipulation*, discusses the incentives and opportunities of firms to use data to exploit the consumer of the future.[14] But it is easy to take such concerns too far. The ascendance of big data will likely improve as many lives as it impoverishes.[15] The same techniques that can figure out an individual consumer's reservation price or pinpoint a vulnerability to a demerit good can filter spam, catch terrorists, conserve energy, or spot a deadly drug interaction.[16] And big data may never deliver on its extraordinary promise. Both its proponents and detractors have a tendency to ascribe near magical powers to big data. These powers may never materialize.[17] Yet the possibility that firms will abuse their asymmetric access to and understanding of consumer data should not be discounted. I believe changes in this dynamic will prove the central consumer protection issue of our age.[18]

II. Ethical Principles

People have experimented on one another for hundreds of years. America and Europe of the twentieth century saw some particularly horrible abuses. In the 1970s, the U.S. Department of Health, Education, and Welfare commissioned twelve individuals, including two law professors, to study the ethics of biomedical and behavioral science and issue detailed recommendations. The resulting Belmont Report—so named after an intensive workshop at the Smithsonian Institute's Belmont Conference Center—is a statement of principles that aims to assist researchers in resolving ethical problems around human-subject research.[19]

The Report emphasizes informed consent—already a mainstay of consumer privacy law.[20] In recognition of the power dynamic between experimenter and subject, however, the Report highlights additional principles of "beneficence" and "justice." Beneficence refers to minimizing harm to the subject and society while maximizing benefit—a kind of ethical Learned Hand Formula. Justice prohibits unfairness in distribution, defined as the undue imposition of a burden or withholding of a benefit. The Department of Health, Education, and Welfare published the Belmont Report verbatim in the Federal Register and expressly adopted its principles as a statement of Department policy.[21]

Today, any academic researcher who would conduct experiments involving people is obligated

to comply with robust ethical principles and guidelines for the protection of human subjects, even if the purpose of the experiment is to benefit those people or society. The researcher must justify her study in advance to an institutional, human subject review board (IRB) comprised of peers and structured according to specific federal regulations.[22] But a private company that would conduct experiments involving thousands of consumers using the same basic techniques, facilities, and personnel faces no such obligations, even where the purpose is to profit at the expense of the research subject.[23]

Subjecting companies to the strictures of the Belmont Report and academic institutional review would not be appropriate. Firms must operate at speed and scale, protect trade secrets, and satisfy investors. Their motivations, cultures, and responsibilities differ from one another, let alone universities. And that is setting aside the many criticisms of IRBs in their original context as plodding or skewed.[24] Still, companies interested in staying clear of scandal, lawsuit, and regulatory action could stand to take a page from biomedical and behavioral science.

The thought experiment is simple enough: the Federal Trade Commission, Department of Commerce, or industry itself commissions an interdisciplinary report on the ethics of consumer research. The report is thoroughly vetted by key stakeholders at an intensive conference in neutral territory (say, the University of Washington). As with the Belmont Report, the emphasis is on the big picture, not any particular practice, effort, or technology. The articulation of principles is incorporated in its entirety in the Federal Register or an equivalent. In addition, each company that conducts consumer research at scale creates a small internal committee comprised of employees with diverse training (law, engineering) and operated according to predetermined rules.[25] Initiatives clearly intended to benefit consumers could be fast-tracked whereas, say, an investigation of how long moviegoers will sit through commercials before demanding a refund will be flagged for further review.

The result would not be IRBs applying the Belmont Report. I suspect Consumer Subject Review Boards (CSRBs) would be radically different. I am not naïve enough to doubt that any such effort would be rife with opportunities to pervert and game the system. But the very process of systematically thinking through ethical consumer research and practice, coupled with a set of principles and bylaws that help guide evaluation, should enhance the salutary dynamics proposed by Mulligan, Bamberger, Swire, and others.

Industry could see as great a benefit as consumers. First, a CSRB could help unearth and head off media fiascos before they materialize. No company wants to be the subject of an article in a leading newspaper with the title *How Companies Learn Your Secrets*. Formalizing the review of new initiatives involving consumer data could help policy managers address risk. Second, CSRBs could increase regulatory certainty, perhaps forming the basis for an FTC safe harbor if sufficiently robust and transparent. Third, and most importantly, CSRBs could add a measure of legitimacy to the study of consumers for profit. Any consumer that is paying attention should feel like a guinea pig, running blindly through the maze of the market. And guinea pigs benefit from guidelines for ethical conduct.[26]

I offer CSRBs as a thought experiment, not a panacea. The accelerating asymmetries between firms and consumers must be domesticated, and the tools we have today feel ill suited. We need to look at alternatives. No stone, particular one as old and solid as research ethics, should go unturned.

---

[1] *See* DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 71 (2006) ("Thus, the federal privacy statutes form a complicated patchwork of regulation with significant gaps and omissions."); Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 Harv. L. Rev. 1880, 1880-82 (2013).

[2] *See* Peter P. Swire, *The Surprising Virtues of the New Financial Privacy Law*, 86 MINN. L. REV. 1263, 1264, 1316 (2002).

[3] *See* Kenneth Bamberger & Deirdre Mulligan, *Privacy on the Books and on the Ground*, 63 STAN. L. REV. 247 (2011); *cf.* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013) (urging a cautious approach to addressing privacy in big data).

[4] My topic here is the intersection of corporate ethics and consumer privacy. There is a rich literature around the ethics of privacy, but it tends to focus on the importance of privacy as a value. *See, e.g.*, ANITA L. ALLEN, UNPOPULAR PRIVACY: WHAT MUST WE HIDE? (2011); James H. Moor, *The Ethics of Privacy Protection*, 39LIBR. TRENDS 69 (1990).

[5] *See* Graeme Wood, *Anthropology Inc.*, THE ATLANTIC (FEB. 20, 2013), http://www.theatlantic.com/magazine/archive/2013/03/anthropology-inc/309218.

[6] *See* Julia Angwin, *The Web's New Gold Mine: Your Secrets*, WALL ST. J. (Jul. 30, 2010), http://online.wsj.com/article/SB10001424052748703940904575395073512989404.html.

[7] *See* Ira S. Rubinstein et al., *Data Mining and Internet Profiling: Emerging Regulatory and Technical Approaches*, 75 U. CHI. L. REV. 261 (2008) (describing the capabilities of data mining).

[8] The ability to design the interface means, for instance, that Apple can update the look of its progress bar to create the appearance of faster download times. *See* Chris Harrison et al., Faster Progress Bars: Manipulating Perceived Duration with Visual Augmentations (2010), *available at* http://www.chrisharrison.net/projects/progressbars2/ProgressBarsHarrison.pdf (finding Apple's new progress bar reduces perceived duration by 11% in subjects). Apple even brings psychology to bear in its physical store. *See, e.g.*, Marcus Morretti, *Revealed: These 10 Extraordinary Rules Make Apple Stores the Most Profitable Retailers in the World*, BUS. INSIDER (June 18, 2012), http://www.businessinsider.com/genius-bar-apple-store-secrets-2012-1?op=1.

[9] *See* Halah Touryalai, *Are Banks Manipulating Your Transactions to Charge You an Overdraft Fee?*, FORBES (Feb. 22, 2012), http://www.forbes.com/sites/halahtouryalai/2012/02/22/are-banks-manipulating-your-transactions-to-charge-you-an-overdraft-fee (reporting on the launch of a Consumer Finance Protection Bureau investigation into how banks process overdraft fees). Several banks eventually settled multimillion-dollar class actions lawsuits.

[10] For a popular account of this literature, see generally ROY BAUMEISTER & JOHN TIERNEY, WILLPOWER: REDISCOVERING THE GREATEST HUMAN STRENGTH (2012).

[11] Objects, from watches to refrigerators, will increasingly be networked and have interfaces. A report by the Swiss mobile device company Ericsson and the Alexandra Institute estimates about fifty billion devices will be networked by 2020 into an "Internet of Things." *See* INSPIRING THE INTERNET OF THINGS! 2 (Mirko Presser & Jan Holler, eds., 2011), *available at* http://www.alexandra.dk/uk/services/publications/documents/iot_comic_book.pdf.

[12] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. Times Mag. (Feb. 16, 2012).

[13] Michael Moss, *The Extraordinary Science of Addictive Junk Food*, N.Y. TIMES MAG. (Feb. 20, 2013).

[14] Ryan Calo, *Digital Market Manipulation* (Univ. of Wash. Sch. of Law, Research Paper No. 2013-27, 2013), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309703.

[15] For a definition of big data and an optimistic account of its impact on society, see VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013).

[16] *See id*; *see also* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1, 8-10 (2011). "Reservation price" and "demerit good" are economic terms referring, respectively, to the highest price a person is willing to pay and a product that is harmful if over-consumed.

[17] *See* Paul Ohm, Response, *The Underwhelming Benefits of Big Data*, 161 U. PA. L. REV. ONLINE 339, 345 (2013), *available at* Already much consumer protection law focuses on asymmetries of information and bargaining power, which big data stands to dramatically enhance.

[18] Already much consumer protection law focuses on asymmetries of information and bargaining power, which big data stands to dramatically enhance.

[19] NAT'L COMM'N FOR THE PROT. OF HUMAN SUBJECTS OF BIOMEDICAL & BEHAVIORAL RESEARCH, THE BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH (1978).

[20] *See* M. Ryan Calo, *Against Notice Skepticism in Privacy (and Elsewhere)*, 87 NOTRE DAME L. REV. 1027, 1028, 1032 (2012).

[21] Protection of Human Subjects, 44 Fed. Reg. 23,192 (Apr. 18, 1979).

[22] *See* Protection of Human Subjects, 45 C.F.R. §§ 46.103, 46.108 (2012) (describing IRB functions and operations).

[23] *Cf.* EVGENY MOROZOV, TO SAVE EVERYTHING, CLICK HERE: THE FOLLY OF TECHNOLOGICAL SOLUTIONISM 148 (2013) ("What institutional research board would approve Google's quixotic plan to send a fleet of vehicles to record private data floating through WiFi networks or the launch of Google Buzz . . . ?"). Morozov's point seems to be that technology companies should think before innovating. I'm not sure I agree with this frame. His examples are also curious—there is no evidence that Google sniffed WiFi on purpose and the problem with Google Buzz was *not enough* advanced consumer testing. *See also* Ohm, *supra* note 16, at 345 (noting that hospitals examining health records should conform to human subject research rules).

[24] *See, e.g.*, Dale Carpenter, *Institutional Review Boards, Regulatory Incentives, and Some Modest Proposals for Reform*, 101 NW. U. L. REV. 687 (2007).

[25] Without delving into issues of standards or structure, Viktor Mayer-Schönberger and Kenneth Cukier briefly suggest that firms employ "internal algorithmists" akin to ombudsman that vet big data projects for integrity and societal impact. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 14, at 181-82.

[26] NAT'L RESEARCH COUNCIL, GUIDE FOR THE CARE AND USE OF LABORATORY ANIMALS (8th ed. 2011).

MO' DATA, MO' PROBLEMS?

PERSONAL DATA MINING AND THE CHALLENGE TO THE DATA MINIMIZATION PRINCIPLE

*Liane Colonna**

1. INTRODUCTION

Data minimization is a bedrock principle of data protection law. It is enshrined in privacy regulations all around the world including the OECD Guidelines, the EU Data Protection Directive, the APEC Privacy Framework and even the recent US Consumer Privacy Bill of Rights. The principle requires that the only personal data that should be collected and stored is that data, which is necessary to obtain certain specified and legitimate goals. It further requires that the personal data should be destroyed as soon as it is no longer relevant to the achievement of these goals.

Data minimization is a rather intuitive and common sense practice: do not arbitrarily collect and store data because this will only lead down a road of trouble consisting of such things as privacy and security breaches. It's the analogue to "mo' money, mo' problems."[1] The predicament is, however, because of recent advances in software development and computer processing power, "mo' data" often means "mo' knowledge" which, like money, can arguably be used to solve many of life's problems.

This paper is about how the concept of personal data mining, a term used to explain the individual use of dynamic data processing techniques to find hidden patterns and trends in large amounts of personal data, challenges the concept of data minimization. It is an attempt to demonstrate that fair information principles like data minimization, while providing a useful starting point for data protection laws, must give way to more nuanced legal rules and models. It stresses that a shift of paradigms from the current paternalistic approach to handling personal towards an empowered-user approach is needed in order to better protect privacy in light of recent advancements in technology.

The outline is as follows. First, the notion of "data minimization" will be commented upon. Second, the technology of data mining will be explained, paying particular attention to a subset of the field that has been dubbed "personalized data mining." Finally, the paper will reflect upon how an unyielding commitment to the principle of data minimization is problematic in a world where the indiscriminate collection and the ad hoc retention of data can lead to many benefits for individuals and society alike.

2. DATA MINIMIZATION

The principle of data minimization first emerged during the 1970s at a time when there was great concern over the large-scale collection and processing of personal data in centralized, stand-alone, governmental computer databases. The idea was simple: limit the collection and storage of personal data in order to prevent powerful organizations from building giant dossiers of innocent people which could be used for purposes such as manipulation, profiling and discrimination. That is, minimizing data collection and storage times, would help protect the individual against privacy intrusions by the

* Doctoral Candidate in Law and Informatics, Swedish Law and Informatics. Research Institute, University of Stockholm.

State or other puissant organizations. After all, data cannot be lost, stolen or misused if it does not exist.

At that time the concept of data minimization was first formulated individuals did not have the software or the processing power to handle large amounts of data themselves. Nor was there a way for ordinary people to collect and distribute limitless amounts of data via an international super network. In other words, while the concern to protect individuals from Big Brother's exploitation of large-scale personal data repositories was palpable, there certainly was little regard for the fact that individuals could somehow benefit from an amassment of their personal data. This is, however, no longer the case.

3. THE TECHNOLOGY OF DATA MINING

3.1 DATA MINING IN GENERAL

Data mining is often thought to be the most essential step in the process of "knowledge discovery in databases", which denotes the entire process of using data to generate information that is easy to use in a decision-making context.[2] The data-mining step itself consists of the application of particular techniques to a large set of cleansed data in order to identify certain previously unknown characteristics of the data set.[3] Data mining techniques can include, for example, classification analysis (takes data and places it into an existing structure[4]), cluster analysis (clumps together similar things, events or people in order to create meaningful subgroups[5]) or association analysis (captures the co-occurrence of items or events in large volumes of data[6]).

A key feature of data mining is that, unlike earlier forms of data processing, it is usually conducted on huge volumes of complex data and it can extract value from such volume.[7] Data mining is also highly automated, sometimes relying on "black boxes."[8] Another interesting feature of data mining is that it creates "new knowledge" such as an abstract description or a useful prediction that did not exist *a priori*.[9] A final important feature about data mining is that it is not necessarily limited by the creativity of humans to create hypotheses because data mining can be used to explore the dataset and generate hypotheses automatically.[10]

In some respect, data mining can be thought of as voodoo science. According to the conventional scientific method, a hypothesis is built and then the data is carefully collected to test the hypothesis. Unlike with the conventional scientific method, the data-mining method involves an exploration of a dataset without a hypothesis in order to discover hidden patterns from data. Instead of being driven by a hypothesis, the process is driven by the data itself and therefore, the results are unanticipated and serendipitous.[11] Here, the concern is that scientific proposals that are derived without a preconceived hypothesis about the data are not valuable, reliable or significant because correlations that appear in the data could be totally random.[12] As such, it is important that data miners understand the risk in the approach and take steps to evaluate the reliability of their findings.[13]

3.2 PERSONALIZED DATA MINING

Individuals today collect and retain large amounts of personal data through a multiplicity of different channels. Through, for example, participating in the so-called Web 2.0, a massive amount of personal data is stored in emails, blogs, Wikis, web browsing history and so on. Social media, a Web 2.0 innovation that introduced web-based sharing with the click of a button, also provides for rich sources of personal data. The information that a user puts onto Twitter and Facebook, for example, can reveal a tremendous amount about a person such as individual's speech patterns, the topics an individual obsesses over and the identity of an individual's "real" friends.[14]

Likewise, individuals are generating a huge amount of data about themselves through using technologies that are embedded in everyday objects that interact with the physical world. Here, there is no need to press any buttons or to self-report: the information is raw and unfiltered. For example, an individual's mobile phone can be used to automatically track location data or Nike+ can be used to record every mile an individual runs.

One way of understanding all of these data is to use the information for personal data mining. That is, this information can be mined to cluster, classify and discover rules in order to assist individuals to extract important insights about themselves and their worlds that might be hidden within these large datasets. For example, if an individual gets frequent headaches then he/she could use data mining to look for patterns that suggest what food or activity that seems to bring the headaches on.[15] Another example is using personal data mining to identify factors that influence weight.[16]

An interesting feature about personal data mining is that the data can be mined either alone or in conjunction with the data of others, possibly collected on multiple platforms, in order to reveal hidden information among the data and the associated users.[17] The question of how precisely an individual shall gain access to this "third-party data" is not straightforward or obvious. For example, in some circumstances, the individual may be able to purchase the data from third parties and in other circumstances the individual may be given free access to the data in the interest of the collective good. The individual is also likely to encounter denial of access to data due to the nature and the value of the information.

While, at first blush, individuals may not appear to have the processing power or computer software that is available to governments and private companies, there are services being offered, which would allow individuals to search for novel and implicit information in large datasets. For example, Google offers a data mining service called Correlate that allows individuals to search for trends by combining individual data with Google's computing power.[18] Likewise, Microsoft has been granted a patent for personal data mining[19] and is currently offering Lifebrowers as a tool "to assist individuals to explore their own sets of personal data including e-mails, Web browsing and search history, calendar events, and other documents stored on a person's computer."[20]

4. PERSONAL DATA MINING AND THE CHALLENGE TO THE NOTION OF DATA MINIMIZATION

Reconciling the principle of data minimization and the notion of personal data mining is difficult. This is because a perquisite to personal data mining is the amassment of huge amounts of data. It is also because the potential benefits of mining this data are unpredictable and can grow exponentially with time, which means there is an interest in storing the data for an indefinite period.

One way of addressing this reality is to focus away from fair information principles such as data minimization towards a misuse model of data protection.[21] That is, instead of the placing the emphasis on limiting data collection, the emphasis could be placed on limiting the misuse of data. This, however, would require a more substantive approach to data protection where individuals can rely upon explicit remedies for the misuse of their personal data.

The main point here is that it matters who uses data and how they use the data and in what context.[22] The individual's mining of personal records in order to fulfill certain personal goals such as becoming more efficient, healthy or knowledgeable about his/her strengths and weaknesses in the context of self-discovery, requires a different reaction from, for example, Facebook mining an individual's personal data to reveal his/her credit worthiness in the context of a mortgage application.[23] While the personal data clearly has value to both the different controllers, it is the way that the data is used where it becomes obvious whether there has been a privacy infraction.

5. CONCLUSION

It is true that limiting the collection and storage of data could help safeguard privacy in certain contexts by, for example, guarding against security breaches. It is also true that the unlimited collection and storage of data can give rise to many individual and societal benefits. Consequently, the current mechanical approach to data protection that presupposes that the haphazard collection of data is always bad for individuals must give way to a more nuanced, relevant and organic model that reflects the recent and dramatic advancements in dynamic data processing and storage techniques.

It is time to recognize that "mo' data" does not always mean "mo' problems" and to create an environment where individuals – and not just

governments and big business – are able to benefit from the analysis of large repositories of personal data. It is time to pay closer attention to the ways that individuals can be empowered with tools to manage and understand their personal data. The current paradigm of "data protection" should be shifted towards "data empowerment" to exhibit greater connection with the technological reality.

[1] "Mo' Money, Mo' Problems" is a song by the legendary rapper and hip hop artist Notorious B.I.G. (also known as Biggie Smalls). It is the second single from his album Life After Death. The single was released posthumously and topped the Billboard Hot 100 for two weeks in 1997. *For more, see* Wikipedia, *Mo Money Mo Problems* retrieved at http://en.wikipedia.org/wiki/Mo_Money_Mo_Problems.

[2] Han, J. and Micheline Kamber, *Data Mining: Concepts and Techniques (Second Edition)(*San Francisco: Morgan Kaufmann Publishers, 2006).

[3] *Id.*

[4] Alexander Furnasapr, *Everything You Wanted to Know About Data Mining but Were Afraid to Ask*, The Atlantic (April 3, 2012).

[5] Amit Kumar Patnaik, *Data Mining and Its Current Research Directions*, a paper presented at International Conference on Frontiers of Intelligent Computing retrieved at http://ficta.in/attachments/article/55/07%20Data%20Mining%20and%20Its%20Current%20Research%20Directions.pdf

[6] Gary M. Weiss and Brian Davison, *Data Mining*, In Handbook of Technology Management H. Bidgoli (ed.), Volume 2 (John Wiley and Sons, 2010), 542-555.

[7] M. Lloyd-Williams, *Discovering the Hidden Secrets in Your Data - the Data Mining Approach to Information*, Information Research: An International Electronic Journal (1997) retrieved at http://informationr.net/ir/3-2/paper36.html

[8] Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives,* 116 Penn State Law Review 285 (*2011*).

[9] K.A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 Columbia Science & Technology Law Review 2 (2003).

[10] Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, In Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases (Bart Custers, Tal Zarsky, Bart Schermer, Toon Calders)(eds.))(Springer 2013).

[11] J.A. McCarty, *Database Marketing*. Wiley International Encyclopedia of Marketing (Wiley 2010).

[12] D.B. Kell, S.G. Oliver, *Here Is the Evidence, Now What is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-Genomic Era*, 26 Bioessays 99–105 (Wiley 2004).

[13] J.A. McCarty, *Database Marketing*. Wiley International Encyclopedia of Marketing (Wiley 2010).

[14] Christopher Mims, *How to Use Twitter for Personal Data Mining*, MIT Technology Review, (October 13, 2010) retrieved at http://www.technologyreview.com/view/421201/how-to-use-twitter-for-personal-data-mining/

[15] Kevin Maney, *Download Net on Your Laptop? Maybe Someday Way storage is Growing, Who knows?* USA Today (July 12, 2006).

[16] Kuner Patal, *Personal Data Mining*, Creativity Online (April 28, 2009) retrieved at http://creativity-online.com/news/personal-data-mining/136077

[17] See *Google's Patent for Personal Data Mining US 20080082467 A1* retrieved at http://www.google.com/patents/US20080082467.

[18] Douglas Perry, *Google Releases Data Mining Engine* (May 26, 2011) retrieved at http://www.tomsguide.com/us/google-org-data-mining-correlate-serach-engine,news-11343.html.

[19] *Google's patent for Personal data mining US 20080082467 A1* retrieved at http://www.google.com/patents/US20080082467.

[20] Tom Simonite, *Microsoft Builds a Browser for Your Past*, MIT Technology Review (March 15, 2012) retrieved at http://www.technologyreview.com/news/427233/microsoft-builds-a-browser-for-your-past/

[21] *See generally*, Fred H. Cate, *The Failure of Fair Information Practice Principles* In Consumer Protection In The Age Of The Information Economy (Jane K. Winn (ed.))(Surry, UK: Ashgate 2006); Peter Seipel, *Privacy and Freedom of Information in Sweden in Nordic Data Protection* Law (First Edition)(Peter Blume (ed..)(Copenhagen: DJØF Publishing, 2001).

[22] Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford Law Books (Stanford, California 2010).

[23] Martha C. White, *Could That Facebook 'Like' Hurt Your Credit Score?*, Time Magazine (June 14, 2012).

CLOUD COMPUTING AND TRANS-BORDER LAW ENFORCEMENT ACCESS TO PRIVATE SECTOR DATA

DATA CHALLENGES TO SOVEREIGNTY, PRIVACY AND DATA PROTECTION.

*Paul de Hert & Gertjan Boulet**

INTRODUCTION

The controversial PRISM programme has uncovered a global reality of trans-border law enforcement access to private sector data, triggered by cloud computing. Law enforcement agencies (LEAs) are indeed increasingly targeting foreign cloud computing service providers[1] and, as put by Europol, cloud computing *"will continue to have a profound impact on law enforcement investigations."*[2] This reality poses challenges to both state interests and individual rights, as it does not only disturb the relations between sovereign states,[3] but also causes legal uncertainty for the individual as regards the applicable privacy and data protection standards for law enforcement access to personal data and metadata in the fight against cybercrime.[4] The distinction between personal data and metadata becomes irrelevant when cross-referencing several sources of data about one individual. Moreover, metadata might be even more revealing than content,[5] so that it can be said that big data *"exacerbate the existing asymmetry of power between the state and the people."*[6]

CHALLENGES TO SOVEREIGNTY

Technology allows state officials to gather evidence and take actions outside their territorial scope without permission from other states. Law can and does acknowledge this either by extending the scope of existing powers or by creating new powers with an explicit extraterritorial reach. In that regard, two Belgian investigative measures are emblematic for a global reality where sovereignty is affected by the trans-border reach of national investigative powers. First, the Belgian Supreme Court held that Article 46bis of the Belgian Code of Criminal Procedure (CCP) can also be applied to a foreign provider of electronic communications services (Yahoo!) to hand over identification data.[7] Secondly, the Belgian lawmaker created the power of the network search (Article 88ter CCP) allowing an investigative judge,[8] when performing a search on a computer system, to extend this search to another computer system even outside the Belgian borders and without formal request for mutual legal assistance. The extraterritorial reach of this network search has been justified by considerations of time and risk of evidence loss in cases of serious crime, but backed by principles of necessity, proportionality and a posteriori notification.[9]

The Belgian Yahoo case and the network search powers raise questions about the scope of territorial jurisdiction, respectively the legality of international hacking and extraterritorial jurisdiction in cyberspace. In that respect, Hildebrandt rightly posits that "[t]he fact that the Internet facilitates remote control across national borders at low costs basically means that the fundamental assumptions of territorial criminal jurisdiction will increasingly fail to describe accurately what is a stake".[10]

Considering the lack of any effective international initiatives for trans-border investigations on the Internet, it would be unrealistic to prohibit national extraterritorial

* Paul de Hert is Professor at the Vrije Universiteit Brussel (Belgium) and the Tilburg University (The Netherlands). Gertjan Boulet is a Doctoral Researcher at the Research Group on Law, Science, Technology and Society (LSTS) at the Vrije Universiteit Brussel.

initiatives for trans-border access.[11] Moreover, current discussions on the international level even seem to endorse such practices.

First, at the invitation of the NATO Cooperative Cyber Defence Centre of Excellence, an International Group of Experts prepared an unofficial (draft) "*Tallinn Manual on the International Law Applicable to Cyber Warfare*".[12] The Manual provides that without prejudice to applicable international obligations, a State may exercise its jurisdiction extraterritorially, in accordance with international law. The Manual further recognizes the impact of cloud computing on jurisdiction, but provides that *"[a] State shall not knowingly allow the cyber infrastructure located in its territory or under its exclusive governmental control to be used for acts that adversely and unlawfully affect other States."*[13] This raises questions about the legality of international hacking, and the role of a posteriori notification duties.

Secondly, the Cybercrime Convention Committee (T-CY) of the Council of Europe has been discussing the development of an Additional Protocol to the Cybercrime Convention on trans-border acces. In a discussion paper of December 6, 2012, the T-CY put that "*the Belgian solution offers great opportunities to handle data stored in 'the cloud'. [...] [and] makes clear that it is not important to know where the data is stored, but from where it is accessible.*"[14] This could mean that the T-CY considers the Belgian network search as an exceptional circumstance under which it would allow hacking.[15] In a guidance note of February 19th, 2013, the T-CY underlined that Parties *"may need to evaluate themselves the legitimacy of a search or other type of [trans-border] access in the light of domestic law, relevant international law principles or considerations of international relations."*[16] In the recent draft elements for an Additional Protocol of April 9th, 2013, the T-CY recalled to avoid international hacking, but at the same time proposed far-reaching options for trans-border access.[17]

Model provisions on notification duties can be found in the Convention on Mutual Assistance in Criminal Matters between the Member States of the European Union[18] which contains a section on the *"Interception of telecommunications without the technical assistance of another Member State*" (MS). The intercepting MS shall inform the notified MS of the interception prior or after the interception depending on whether it knows when ordering or becomes aware after the interception that the subject of the interception is on the territory of the notified MS. Until the notified MS decides if the interception can be carried out or continued, the intercepting MS may continue the interception and use the material already intercepted for taking urgent measures to prevent an immediate and serious threat to public security.

In these documents we see a first set of good ideas about regulating trans-border law enforcement. 'It can be done' and 'it has advantages', the documents seem to suggest, but sovereignty needs to be protected as much as possible, through creating some sort of transparency before or after interventions by the law enforcing state.

We now turn to the challenges that trans-border law enforcement access to private sector data poses to the rights to privacy and data protection.

CHALLENGES TO THE RIGHTS TO PRIVACY AND DATA PROTECTION

A cybercrime report of the United Nations Office on Drugs and Crime contains a section on "extra-territorial evidence from clouds and service providers", which provides that the Cybercrime Convention does not adequately cover situations of trans-border access due to provisions on consent of the person with lawful authority to disclose the data.[19] In its draft elements of April 2013, the T-CY also moved away from a strict condition of consent, but which evoked criticism from academics, private sector and civil society.[20] Considering the apparent lack of trust in the transparency and accountability of governments, it can be said that complementary private sector instruments addressing their own transparency and accountability in law enforcement, would give the individual a least some parts of the puzzle on his or her fundamental rights status on the table. For instance, Google and Microsoft are increasingly providing transparency about their cooperation with LEAs.[21] Moreover, Microsoft hosted a series of five privacy dialogues to

discuss "the role of individual control and notice and consent in data protection today, as well as alternative models that might better protect both information privacy and valuable data flows in the emerging world of Big Data and cloud computing."[22] The final Global Privacy Summit yielded a report underpinned by a respect for information privacy principles.

CONCLUSION

The global reality of trans-border law enforcement access to private sector data, triggered by cloud computing, undermines both state interests and the rights to privacy and data protection. Challenges to sovereignty relate to the scope of territorial jurisdiction, and to the legality of international hacking and extraterritorial jurisdiction in cyberspace. Challenges to the rights to privacy and data protection relate to the existing legal uncertainty for the individual as regards the application of privacy and data protection standards, including the role of individual consent, for law enforcement access to personal data and metadata in the fight against cybercrime. Current international documents seem to suggest full protection of sovereignty, through a priori or a posteriori notification duties for the law enforcing state. Yet, considering the apparent lack of trust in the transparency and accountability of governments, complementary private sector instruments addressing their own transparency and accountability in law enforcement could arguably give the individual a least some parts of the puzzle on his or her fundamental rights status on the table. Although the challenges at stake still need to be addressed in the greatest detail, current documents can and should already be critically assessed in that respect.

---

[1] Omer Tene & Christopher Wolf, "Overextended: Jurisdiction and Applicable Law under the EU General Data Protection Regulation", White Paper, Future of Privacy Forum, January 2013, http://www.futureofprivacy.org/wp-content/uploads/FINAL-Future-of-Privacy-Forum-White-Paper-on-Jurisdiction-and-Applicable-Law-January-20134.pdf.

[2] Europol, SOCTA 2013, Public version, March 2013, p. 14, https://www.europol.europa.eu/content/eu-serious-and-organised-crime-threat-assessment-socta.

[3] Didier Bigo, Gertjan Boulet, Caspar Bowden, Sergio Carrera, Elspeth Guild, Nicholas Hernanz, Paul de Hert, Julien Jeandesboz and Amandine Scherrer, Open Season for Data Fishing on the Web The Challenges of the US PRISM Programme for the EU, CEPS Policy Brief, No. 293, 18 June 2013, p. 3, http://www.ceps.be/book/open-season-data-fishing-web-challenges-us-prism-programme-eu.

[4] Didier Bigo, Gertjan Boulet, Caspar Bowden, Sergio Carrera, Julien Jeandesboz & Amandine Scherrer, Fighting cyber crime and protecting privacy in the cloud, European Parliament, Committee on Civil Liberties, Justice and Home Affairs, PE 462.509, http://www.europarl.europa.eu/committees/en/studiesdownload.html?languageDocument=EN&file=79050, pp. 9, 37.

[5] *Cf* Ann Chavoukian, "A Primer on Medatata: Separating Fact From Fiction", Information and Privacy Commissioner Ontario, Canada, July 2013, http://www.privacybydesign.ca/index.php/paper/a-primer-on-metadata-separating-fact-from-fiction/.

[6] Kenneth Cukier and Viktor Mayer-Schoenberger, "The Rise of Big Data. How It's Changing the Way We Think About the World", 92 Foreign Affairs, May/June 2013, p. 37.

[7] Supreme Court, September 4th, 2012, A.R. P.11.1906.N/2.

[8] An investigative judge is a magistrate charged with the task of gathering evidence in a case. Investigative judges only exist in the inquisitorial system used throughout continental Europe.

[9] Belgian Computer Crime Act of 2000, preparatory works Chamber of Representatives, 1999-2000, nr. 2-392, pp. 23-25.

[10] Mireille Hildebrandt, "Extraterritorial jurisdiction to enforce in cyberspace?: Bodin, Schmitt, Grotius in cyberspace", *University of Toronto Law Journal*, Vol. 63, No. 2, Spring 2013, p. 204.

[11] Paul De Hert, "Cybercrime and Jurisdiction in Belgium and the Netherlands. Lotus in Cyberspace – whose Sovereignty is at Stake?", B.J. Koops & S.W. Brenner, *Cybercrime and Jurisdiction*, The Hague, T.CM. Asser Press, 2006, pp. 73-74, 76, http://www.vub.ac.be/LSTS/pub/Dehert/024.pdf.

[12] M.N. Schmitt (ed.), "Tallinn Manual on the International Law Applicable to Cyber Warfare", , Cambridge/New York/Melbourne/Madrid/Cape Town/Singapore/São Paulo/Delhi/Mexico City, Cambridge University Press 2013, p. 23, http://www.ccdcoe.org/249.html.

[13] Ibid., pp. 27, 33.

[14] Council of Europe (Cybercrime Convention Committee (T-CY), Ad-hoc sub-group on jurisdiction and transborder access to data): Discussion paper: Transborder access and jurisdiction: What are the options? Report of the Transborder Group. Adopted by the T-CY on 6 December 2012, p. 33, § 172, http://www.coe.int/t/dghl/standardsetting/t-cy/TCY2012/TCY_2012_3_transborder_rep_V30public_7Dec12.pdf.

[15] Ibid., pp. 50, 57.

[16] Cybercrime Convention Committee (T-CY), "T-CY Guidance Note # 3 Transborder access to data (Article 32), Proposal prepared by the Bureau for comments by T-CY members and observers and for consideration by the 9th

Plenary of the T-CY (June 2013)", T-CY (2013)7 E, Strasbourg, 19 February 2013, pp. 3 & 5, http://www.coe.int/t/dghl/cooperation/economiccrime/Source/Cybercrime/TCY/TCY%202013/TCY_2013_7E_GN3_transborder_V2public.pdf.

[17] Cybercrime Convention Committee (T-CY), (Draft) elements of an Additional Protocol to the Budapest Convention on Cybercrime regarding transborder access to data, Proposal prepared by the Ad-hoc Subgroup on Transborder Access, T-CY (2013)14, Strasbourg, version 9 April 2013, http://www.coe.int/t/dghl/cooperation/economiccrime/Source/Cybercrime/TCY/TCY%202013/T-CY(2013)14transb_elements_protocol_V2.pdf.

[18] Council Act of 29 May 2000 establishing in accordance with Article 34 of the Treaty on European Union the Convention on Mutual Assistance in Criminal Matters between the Member States of the European Union (2000/C 197/01), *PB.C* 197/1 of 12 July 2000, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2000:197:0001:0023:EN:PDF.

[19] United Nations Office on Drugs and Crime, Comprehensive Study on Cybercrime, Draft – February 2013, United Nations, New York, 2013, p. 216, http://www.unodc.org/documents/organized-crime/UNODC_CCPCJ_EG.4_2013/CYBERCRIME_STUDY_210213.pdf.

[20] http://www.edri.org/edrigram/number11.11/transborder-data-access-cybercrime-treaty; T-CY Public Hearing: 3 June 2013, Strasbourg, France, http://www.coe.int/t/dghl/cooperation/economiccrime/cybercrime/T-CY/Public%20Hearing/TCY_Public_Hearing_en.asp.

[21] Google, Transparency report: http://www.google.com/transparencyreport/userdatarequests/; Microsof, Law Enforcement Request Report: http://www.microsoft.com/about/corporatecitizenship/en-us/reporting/transparency/.

[22] Fred H. Cate and Viktor Mayer-Schönberger, Notice and Consent in a World of Big Data, Microsoft Global Privacy Summit Summary Report and Outcomes, November 2012, p. 4, http://www.microsoft.com/en-au/download/details.aspx?id=35596; Fred H. Cate and Viktor Mayer-Schönberger, "Notice and consent in a world of Big Data", International Data Privacy Law, 2013, Vol. 3, No. 2, pp. 67-73.

TAMING THE BEAST:
BIG DATA AND THE ROLE OF LAW

*Patrick Eggimann & Aurelia Tamò**

MAPPING THE PROBLEM

The term big data has become omnipresent – journalists, privacy scholars and politicians are becoming aware of its importance. The benefits as well as concerns that big data is linked to, are not fundamentally new to privacy advocates. The root and rationale behind big data had earlier been debated under the term data mining or data warehousing. Yet, big data goes beyond the known: with the increased velocity of data processing, the immense volume of generated data and potential to combine a variety of data sets, the so far undeveloped predictive element in our digital world has been released.[1]

Until now, «datafication», or the quantification of information about all things happening, has shifted the focus away from the search for causality. In order to reduce complexity, correlations within big data sets are analyzed. Based on these correlations, predictions are made.[2] A future dimension of big data could well shift the focus of analytics back to causality once again. Overall, the high level of complexity for analyzing the "what or the why" requires complex, autonomous processes which are often opaque for users. Accordingly, the human capacity for understanding how data is being processed within the system, and on what grounds the outcomes are being justified, is seriously challenged.

The user's loss of control over and ignorance of how the data and information is handled and in what ways the resulting knowledge is used, leads to civil liberties concerns. Knowledge extracted from the information provided in the big data sets is in fact the "key to power in the information age".[3] Even if the search for knowledge is in general to be considered a positive goal of the big data phenomenon, knowledge can turn out to be destructive depending on how it is used (a fact that Albert Einstein already recognized). From a social and democratic perspective, the concentration of knowledge as power in the hands of a few together with its potential misuse would represent such a destructive force. Moreover, an increased fear of being observed and analyzed could result in a threat not only to the freedom of speech or freedom of information, but more broadly to the individuals' willingness to participate in public and democratic debates, or even in social interactions on an individual level.[4]

THE ROLE OF DATA PROTECTION LAW IN A BIG DATA AGE

In Europe, the European Convention for Human Rights (ECHR) as well as the Charter for Fundamental Rights (EUCFR) protects the individual's private and family life (Art. 8 ECHR, Art. 7 EUCFR) as well as his or her personal data (Art. 8 EUCFR). These fundamental rights are incorporated into European data protection law (Directive 95/46/EC), which on the basis of the protection of the individual's right to personality, is the main approach when dealing with (big) data processing. In particular, the fundamental principles of consent, transparency, purpose limitation, data minimization, security and proportionality are key to restricting the processing and evaluation of big (personal[5]) data sets.

When talking today about the limitations of data processing the focus lies primarily on private companies, such as Google, Facebook or Amazon. This fact is of special interest because the *ratio legis* behind the introduction of data

* Patrick Eggimann, Executive Manager, Research Center for Information Law (FIR-HSG); Aurelia Tamò, Researcher, Research Center for Information Law (FIR-HSG).

protection law in Europe was the protection of the individual against the superiority of governmental bodies and the potential misuse of citizens' data and census databases rather than the threats from private entities.[6] This scope is also reflected in the famous *Volkszählungsentscheid* of the German Supreme Court of 1983 which is seen as the fundament for the right of informational self-determination.[7]

Even though, the data protection principles in Europe are applicable to both, governmental bodies and private parties that are processing data, the trend that private companies possess and handle a great deal of valuable information about individuals has shifted the balance of knowledge. The recent PRISM and Tempora affairs illustrate the fact that governments want to have what Silicon Valley has: vast amounts of private data and the most sophisticated technology to harvest it.[8]

Distinguishing the actors that interplay in informational relationships is crucial, since the founding rationales governing the relationship are converse: When the government is processing the personal data of citizens, its actions must be democratically legitimized by legal norms, whereas the informational relationships between private entities and consumers are governed by the freedom of contract.

Against this backdrop and in light of big data processing, the principle of purpose limitation is of particular interest. This principle, also referred to in the US as purpose specification,[9] stands in contrast to the mechanism of big data. A rigorous enforcement of purpose limitation would preclude big data since it lies in its logic to evaluate more data for purposes unknown at the moment of collection. The question remains therefore, whether this democratically legitimized principle stands above consent, i.e. the parties' agreements on data processing. Such an extensive application is suggested by the European Data Protection Authority, so-called Working Party 29.[10]

Restrictions among private parties were not conceived within the original purpose of data protection law in Europe. Even if it can be argued that the principle of consent is currently applied in a problematic way,[11] there is no mandate for a state authority to narrow the scope of private consensus by restrictively applying data protection principles. Such an approach results in a hybrid understanding of data protection regulation, which collides with the underlying *ratio legis* of data protection law. By protecting the specified *raison d'être* of data processing, data protection authorities in Europe use a questionable paternalistic approach to overcome the information asymmetry between the data controller and the data subject. State interventions in general, and legal provisions that are protecting the weaker party in particular, are by no means reprehensible and are usefully adopted in many areas of the law.[12] Nevertheless, when it comes to data protection in a big data world such an approach reaches its limits.

Overcoming the Big Challenges

Different approaches toward overcoming the challenges arising out of big data have been discussed by legal scholars.[13] We argue that taking an approach based on consent when personal data is being processed by private entities is not totally amiss. In theory, contract law has the advantage of offering greater flexibility and respects considered, self-determined consumer choices.[14] In practice however, the downside remains the information asymmetry, which in our highly technologized world of big data is increasingly challenging. In addition, the option of negotiation as a vital element of a contract, is underdeveloped and in peril when agreement is already considered to be reached by the mere usage of a service.[15] The question is how to overcome these practical obstacles by other means than strict regulatory intervention.

Overcoming information asymmetries (rather than the superiority of the state as rooted in data protection law outlined above) and creating options for successful negotiations are not singular problems of big data. However, big data accentuates asymmetries due to its complexity, unpredictability and individuals' lack of awareness that data is being processed. Contractual law has already established counter mechanisms to overcome these challenges, such as the principle of *culpa in contrahendo* regarding negotiations or the principle of good faith. Also the courts in civil law countries play an important role in concretizing such principles.

In Switzerland for instance, a court ruling obliged banks to disclose relevant information to its clients in order for them to be able to contractually waive the profits out of retrocession payments by third parties.[16]

Solutions to enhance negotiation between private parties should be centered on improving the choices of the individuals. Here the option to choose the private entity that is processing the personal data is key. Already today, a variety of private entities lure users to their services by providing them with what they need without the exchange of personal information. The search engine duckduckgo, whose increasing user number was further boosted with the PRISM affair, or the software disconnect, as an example for a privacy by design solution provided by a third party, are two examples of how competition and innovation can lead to a more diverse field of options for consumers. Also mechanisms such as labeling could be implemented in an online world to counterbalance the information gap and facilitate more informed consumer choices.[17] Governments then have the responsibility to ensure market conditions that enhance such innovation through appropriate regulation.

As the argument laid out here shows, we are not claiming that governments should not play a role in the current debates on how to regulate our big data world. On the contrary, governments play a crucial role not only in the education of their citizens, but also in setting the underlying structures in which technology can and will flourish. Transparency and choice play an important role in this context: informed individuals should be put in the position to decide what they are willing to give up in order to gain new possibilities and make use of the latest technological advancements.

The willingness and ease with which people make use of new technologies is essentially determined by trust.[18] Trust is key when it comes to establishing a relationship since transparency is almost always only given to a certain degree. Nevertheless, transparency must be measured on its result, which ought to be clarity and not obfuscation. In this sense, the tools of big data are very likely to be not only the cause of the problem but also part of the solution. This can be seen in applications such as disconnect, which graphically captures the potential big data processors. In relation to the government, trust entails the expectation that the former will not fall short on its promise to enforce its laws.

Taking a step back, we believe it is important not to forget the social changes resulting out of the evolving consolidation of the digital and non-digital spheres. As a recent study of online-behavior on social networking sites by the Pew Research Center has shown, adolescents are adapting to the new privacy conditions online. This adaptation is in our opinion an important factor as it reflects an individual change of attitude[19] that has not yet been integrated enough into debates between politicians, industry representatives and consumer protection organizations. We see here the potential for academia to provide further insights into the understanding of the relationship of society, law and technology.

---

[1] Paul C. Zikopoulos et al., IBM Understanding Big Data 15 (2012), https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CMP=is_bdebook1_bdmicrornav.

[2] Mayer-Schönberger & Cukier, Big Data – A Revolution That Will Transform How We Live, Work, and Think 14, 79 et seqq. (2013).

[3] Daniel J. Solove, The Digital Person - Technology and Privacy in the Information Age 74 (2004).

[4] Fred H. Cate & Viktor Mayer-Schönberger, Notice and Consent in a World of Big Data, Microsoft Global Privacy Summit Summary Report and Outcomes 5 (2012), http://www.microsoft.com/en-us/download/details.aspx?id=35596; Lizette Alvarez, Spring Break Gets Tamer as World Wachtes Online, NYTimes (March 16, 2012), http://www.nytimes.com/2012/03/16/us/spring-break-gets-tamer-as-world-watches-online.html?_r=0.

[5] In Europe the 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 is applicable whenever personal data is being processed. The definition of the term "personal data" and the utopia of "anonymous" data have already been discussed in depth by legal scholars: Cf. Paul Ohm, Broken Promises of Privacy to the Surprising Failure of Anonymization, 57 UCLA L. Rev. 1701 (2010); Paul M. Schwartz & Daniel J. Solove, The PII Problem: Privacy and a New Concept of Personally Identifiable Information, 86 N.Y.U. L. Rev. 1814 (2011).

[6] Colin Bennett, Regulating Privacy: Data Protection and Public Policy in Europe and the United States 29 et seqq. (1992).

[7] Horst-Peter Götting, Christian Schertz & Walter Seitz, Handbuch des Persönlichkeitsrechts § 22 N 59 (2008).

[8] Bits New York Times, Deepening Ties Between N.S.A. and Silicon Valley (June 20, 2013), http://bits.blogs.nytimes.com/2013/06/20/daily-report-the-deepening-ties-between-the-n-s-a-and-silicon-valley/?nl=technology&emc=edit_tu_20130620.

[9] Daniel J. Solove, Understanding Privacy 130 (2008).

[10] Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation 11 et seqq. (April 2, 2013), http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[11] Cate & Mayer-Schönberger, *supra* note 4, 3 seq.; Ira S. Rubinstein, Big Data: The End of Privacy or a New Beginning?, NYU Public Law Research Paper No. 12-56, 3, 5 (2013); cf. also Solon Barocas & Helen Nissenbaum, On Notice: The Trouble with Notice and Consent, Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information (October 2009), http://www.nyu.edu/projects/nissenbaum/papers/ED_SII_On_Notice.pdf.

[12] Cf. Jeremy A. Blumenthal, Emotional Paternalism, 35 Fla. St. U. L. Rev. 1 (2007).

[13] Mayer-Schönberger & Cukier, *supra* note 2, 173 argue for holding data controllers accountable for how they handle data and propose a tort law approach; Omer Tene & Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 11 Nw. J. Tech. & Intell. Prop. 239, 263 seq. propose a "sharing the wealth" strategy with data controllers providing individuals access to their data in a usable format and allowing them to profit from data analysis with the help of tools provided; for more references see Ira S. Rubinstein, *supra* note 11, fn. 48.

[14] Cf. Robert Cooter & Thomas Ulen, Law & Economics 355 (6th ed. 2012).

[15] Cf. e.g. Facebook Statement of Rights and Responsibilities as of December 11, 2012, 14, (3) states that "Your continued use of Facebook following changes to our terms constitutes your acceptance of our amended terms", https://www.facebook.com/legal/terms.

[16] Fabien Aepli et al., Landmark decision of the Swiss Federal Supreme Court (November 2, 2012), http://www.lexology.com/library/detail.aspx?g=61377f8b-bd6b-430f-b826-87fe0fed63f3.

[17] The European Interactive Digital Alliance has announced the approval of two technology platform providers to serve an Online Behavioural Advertising Icon, http://www.iabeurope.eu/news/edaa-names-truste-and-evidon-approved-oba-icon-providers.

[18] Cf. World Economic Forum in collaboration with the Boston Consulting Group, Rethinking Personal Data: Strengthening Trust (May 2012), http://www.weforum.org/reports/rethinking-personal-data-strengthening-trust; McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity 116 (June 2011). http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[19] Pew Research Center, Teens, Social Media and Privacy, Berkman Center 8, 41-50, 63 (2013), http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_TeensSocialMediaandPrivacy.pdf.

MANAGING THE MUDDLED MASS OF BIG DATA

*Susan Freiwald**

At the same time that Big Data promises previously unobtainable insights, its use places significant pressure on three significant methods of legal regulation to protect privacy. First, because Big Data merges data from different sources, it makes ineffective legal regulation targeted to the method of data collection. Second, Big Data renders obsolete regulation that relies on identifying a particular data holder. Third, Big Data makes it more difficult to keep data of one type segregated from data of another type and weakens regulations that depend on information segregation.

Managing the muddled mass of Big Data requires law makers to focus not only on how the data got to where it is but also on how it is being used. It requires an evaluation of the value versus the risk of having large databases, which depend on the quality and security of their data, and the dangers from data disclosure. Whenever Big Data projects involve risks to privacy and civil liberties, trustworthy experts should assess the value of the analytics they use in a transparent manner, and those results should be regularly reassessed.

WHAT IS NEW ABOUT BIG DATA?

Prior to the era of Big Data, databases[1] held discrete sets of data, whose collection we could regulate, which were stored by an identifiable and stable source. In the private context, companies that sold goods and services recorded information electronically about their customers, as did health care providers, banks, and credit card companies. Even online companies kept information about our web browsing, our searching, and our "likes" in their own proprietary databases. Law enforcement agents gathered information about a particular target, using a particular technique, such as an electronic pen register or a request for stored emails, and stored those records in a database.[2]

Big Data projects merge data from multiple places, which is how they get to be "Big". In the government context, the perceived need to find potential terrorists in our midst has led to the merger of data from multiple sources in fusion centers[3] and to the FBI joining forces with the NSA to gather up huge quantities of information from multiple sources. Big Data projects in the private sector involve data brokers pulling data from multiple sources to create behavioral profiles to yield the most effective targeted marketing.[4] While Big Data projects need good analytical tools based on sound logic, they work best, at least in theory, when they have the richest and deepest data to mine.

The depth of the data in Big Data comes from its myriad sources. To visualize, think of a Big Data database that has more information about a particular person (or entry) as adding to its length, in the sense that it spans a longer period (i.e., 5 years of John Doe's email records rather than 6 months). Adding entries for more people (e.g., adding in the emails of John Doe's wife and kids) increases its width. But Big Data has greater depth as well, in the sense that it can also analyze John Doe's web browsing data and his tweets. Because Big Data information comes from multiple sources, the entity who analyzes it is quite likely not the one who gathered it.[5]

REGULATION BASED ON COLLECTION

In each of the commercial, law enforcement, and national security contexts, we have traditionally regulated at the point of data collection. Any data that has become

* Professor of Law, University of San Francisco School of Law.

untethered from its collector and the method by which it was collected moves beyond the reach of those laws.[6]

Sectoral privacy laws place limits on what data may be collected, requiring that some personally identifiable data, in some contexts, be gathered only after data subjects give some kind of consent. The Children's Online Privacy Protection Act (COPPA),[7] which regulates the acquisition of information for marketing purposes about those under 13, provides perhaps the most rigorous regime, but regulations in the health care, financial, and cable context provide other examples.[8] Terms of service in the online context also permit, in varying degrees, those who contract with online companies to limit the extent to which those companies may collect and store information.

Those mechanisms are of limited use for those entities that operate outside of the specific parameters of the statutory definitions or outside of the contracts that terms of service arguably create. No sectoral law yet covers data brokers, for example, so their collection practices face no statutory regulation. And those who are covered by either statutory or contractual limits generally find ways to transfer information to third parties who are free of those limits. Once data ends up in the hands of Big Data processors, it has often become free of legal constraints based on collection.

Data Privacy protections in the law enforcement context reside in controls over how law enforcement may conduct surveillance. The Electronic Communications Privacy Act (ECPA) imposes procedural safeguards before agents may use electronic devices to gather up information (email intercepts or modern pen registers) or compel the disclosure of electronic and related communications information from service providers.[9] But ECPA places no limits on buying data in bulk from commercial vendors, or amassing it in fusion centers, both of which enable the use of Big Data analysis for preventative law enforcement.

The recent revelations about Section 215 of the USA PATRIOT Act illustrate the executive branch's use of a terrorism-prevention rationale to avoid regulations geared towards collection. Even though the statute requires that information be gathered only when it is "relevant" to "protect against international terrorism or clandestine intelligence activities"[10] the executive branch has admitted to collecting all telephony metadata (non-content information) for calls within the United States and storing the data for five years; apparently it does not query the database without some suspicion of wrongdoing.[11] By avoiding the statutory collection limit, the executive has apparently been subjecting been itself to its own discretionary limits on its data access. The danger to civil liberties is obvious; through its almost certainly unconstitutional practices, the executive has amassed a gigantic database filled with all of our personal communication information.

Regulation Based on Identification

Big Data also renders ineffective those privacy protections that depend on the identification of a stable data collector. When someone becomes the target of inappropriate or unlawful data collection, she needs to be able to identify the data holder to have that holder purge the improperly collected data. That may be impossible with Big Data.

In the commercial context, for example, COPPA requires that website operators accede to demands by parents to purge their databases of information about their children.[12] From the recently decided *Maryland v. King* case, we know that, under the state statute whose constitutionality the Supreme Court upheld, authorities destroy the DNA information of any arrestee subsequently found to be not guilty.[13] The minimization provisions of the Foreign Intelligence Surveillance Act (FISA) purport to get rid of (some of the) improperly intercepted communications of U.S. persons as soon as it is determined that they are not relevant to foreign intelligence. For all of these mechanisms to work effectively, however, the data holder has to be stable and identifiable, and the data has to remain with that entity.

After data has been copied and sold to other entities, having it purged by the original collector does no good. If fusion centers merge data from private and public sources into one

master database, they presumably would not indicate that to the original subject so that person could bring claims based on inappropriate use. Maryland may purge its own DNA database, but if the defendant's DNA has already been transferred to a central repository, it is unlikely to be purged after the defendant's acquittal. And of the many revelations that have come to light about the FISA minimization procedures, one indicates that the inadvertently collected communications of U.S. persons may be forwarded to the FBI for any law enforcement purpose.[14]

Regulation Based on Segregation

The merger of information in the foreign intelligence and law enforcement context illustrates another method of privacy protection that Big Data renders ineffective. Historically, the law has distinguished between data held by private entities from data held by government entities. It has also treated surveillance for law enforcement purposes under an entirely different set of rules than surveillance for foreign intelligence gathering. Big Data has merged all data together.

Traditionally, we have been more concerned about private data in the hands of the government than we have been about private data in private hands. That is why the Privacy Act[15] regulates government data collection only and does not address private collection. It is also why ECPA permits electronic communications services providers (those who provide email, cell phone services, etc.) to voluntarily divulge records of those services to any non-government entity but not to governmental entities.[16] Once private intermediaries acquire such records, however, they are free to sell or give them to the government, which undoubtedly contributes to how fusion center databases become populated with information.

In the past, we erected virtual walls between the workings of domestic law enforcement and foreign intelligence agents. The former operated under much stricter standards, because citizens have constitutional rights that foreigners lack, and because protecting the nation's security carries more weight than ordinary crime fighting. Recent disclosures indicate that the FBI and the NSA have been working closely together to gather up the giant metadata database described above. The NSA apparently uses metadata databases (of both telephony and internet data) to hone its foreign intelligence queries. These actions mandate reform because it seems clear that the executive is operating under the weaker foreign intelligence standards to further ordinary law enforcement goals. Big Data should be the focus of some reform.

Handling the Muddy Mass

With recognition of the problem the first step towards solving it, the next step does not require reinventing the wheel. Academics[17] and expert commissions[18] have studied data mining at some length and come to several conclusions about how to minimize harm. Those insights themselves need to be mined as we supplement our ineffective legal approaches with ones that are effective for Big Data.

Those who have studied the issue agree on several key principles. Importantly, we must not be intimidated by the technically sophisticated nature of Big Data analysis. Even if we have to engage independent experts to do it, we should subject our data queries to oversight for effectiveness, and make sure we do not attribute unwarranted legitimacy to the results of Big Data queries.[19] Big Data programs must be much more transparent than they now are, so that the efficacy and fairness of their use can be monitored.

In addition, we must better appreciate that the mere accumulation of data in one place creates a risk both from insiders who abuse their access and outsiders who gain access. Because of those risks, data security, immutable audit trails, and meaningful accountability are also crucial features of effective Big Data regulations.

Conclusion

Big Data's depth represents its value and its challenge. By pulling data from a variety of sources into a single source, Big Data promises new answers to questions we may never have thought to ask. But it also fundamentally

challenges regulations based on collection, identification, and segregation. Instead, we need to focus on transparency, expert review, efficacy, security, audit and accountability to reap the benefits of Big Data while minimizing the costs.

[1] Databases are not new. I worked full-time as a database programmer more than 25 years ago.

[2] *See, e.g.,* United States v. Forrester, 512 F.3d 500, 511 (9th Cir. 2008) (finding real-time collection of IP addresses by law enforcement agents to be unprotected by the Fourth Amendment); United States v. Warshak, 631 F.3d 266, 288 (6th Cir. 2010) (holding that acquisition of thousands of stored email without a warrant is unconstitutional). In both of these cases, law enforcement surely stored the electronic records they acquired in a database they could search for evidence.

[3] *See* Danielle Keats Citron and Frank Pasquale, *Network Accountability for the Domestic Intelligence Apparatus,* 62 HASTINGS L. J. 1441 (2011) (describing and critiquing fusion centers).

[4] *See* Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934 (2013).

[5] WhileTwitter stores tweets for some period of time, other public and private entities are engaging in social network scraping, where they collect and store publicly available information, so a compiler of tweets may not be Twitter.

[6] This is in contrast to the European approach, which regulates data processing generally. *See* LOTHAR DETERMANN, DETERMANN'S FIELD GUIDE TO INTERNATIONAL DATA LAW COMPLIANCE xiii (2012) ("European data protection laws are first and foremost intended to restrict and reduce the automated processing of personal data – even if such data is publicly available.").

[7] 15 U.S.C. §§ 6501-6506 (as amended).

[8] Sectoral privacy laws regulate information gathered in the contexts of health care (the Health Insurance Portability and Accountability Act Regulations or HIPAA); banking (the Gramm-Leach Bliley Act of 1999), cable (the Cable Communications Policy Act), videotape rentals (Video Privacy Protection Act), and others.

[9] *See* 18 U.S.C. §§ 2510–2522 (2002) (regulating the interception of electronic communications); at 18 U.S.C. §§ 3121–27 (2010) (regulating the use of pen registers); 18 U.S.C. §§ 2701–11 (2010) (regulating the acquisition of stored communications and records).

[10] 50 U.S.C. § 1861(b)(2)(A) (2006).

[11] *See* Administration White Paper, *Bulk Collection of Telephony Metadata under Section 215 of the USA Patriot Act* (August 9, 2013) (available at https://www.eff.org/document/administration-white-paper-section-215-patriot-act).

[12] 15 U.S.C. § 6502(b)(1)(B)(ii) (requiring the parent to "refuse to permit the operator's further use or maintenance in retrievable form, or future online collection, of personal information from that child").

[13] Maryland v. King, 133 S .Ct. 1958, 1967 (2013).

[14] *See* Foreign Intelligence Surveillance Court Memorandum Opinion and Order of October 3, 2011 (redacted) at 51-52 (available at http://www.dni.gov/files/documents/October%202011%20Bates%20Opinion%20and%20Order%20Part%206.pdf).

[15] 5 U.S.C. § 552a (2006).

[16] 18 U.S.C. §2702 (a)(3) (2006).

[17] *See e.g.*, Fred. H. Cate, *Government Data Mining, the Need for a Legal Framework,* 43 Harv. C.R.-C.L. L. Rev. 435 (2008); K.A. Taipale, *Data Mining and Data Security: Connecting the Dots to Make Sense of Data,* 5 COLUM. SCI. & TECH. L. REV. 2 (2005).

[18] *See, e.g.,* The Task Force on National Security in the Information Age, Markle Found., Creating a Trusted Network for Homeland Security (2003); Task Force on National Security in the Information Age, Markle Found., Mobilizing Information to Prevent Terrorism (2006); Tech. and Privacy Advisory Comm., U.S. Dep't of Def., Safeguarding Privacy in the Fight Against Terrorism (2004).

[19] Some computer experts have questioned the very premise of searching large databases for terrorist-planning patterns because we lack enough terrorist events to know what a plan looks like.

REGULATING THE MAN BEHIND THE CURTAIN

Christina Gagnier, Esq.

"Pay no attention to the man behind the curtain!"

- Frank L. Baum, *The Wonderful Wizard of Oz*

Frank L. Baum's famed novel, *The Wonderful Wizard of Oz,* has been noted as a political allegory for the gold standard amongst other speculation as to Baum's intentions when penning the beloved children's tale. While the early twentieth century novel was written at a time when the conception of privacy itself was nascent, with Samuel Warren and Louis Brandeis' often-cited *The Right to Privacy* being written for Harvard Law Review a mere ten years before, the title character, the Wizard of Oz, the "man behind the curtain," serves as an appropriate analogy for exploring the practice employed by many of the world's most famous brands today of managing Internet user data through the use of third-party "social network intermediary" systems.[1]

The Wizard of Oz is an unknown entity for much of the 1900 novel: he appears in multiple incarnations, but his true nature does not become clear until near the end of the story. He is simply a "man behind the curtain," using machines, illusions and gadgetry unknown to the public on the other side. Despite the illusion, many of the world's most popular brands are not directly interacting with Internet users through their own means on social networks like Twitter, Facebook and YouTube. Their communication is powered by the use of social network intermediaries, third party systems that allow for brands to manage all communications about or with them on multiple social network services across the social Web. While these brands may be using these third party services, the Internet user has no clue as to their existence: these services are a hidden party unknown to the Internet user.

While these social network intermediaries operate legally under arguably some of the strictest standards, such as the 1995 European Privacy Directive, those who constructed this regulation could not have envisioned their existence.[2] Today, as the "right to be forgotten" online is being debated in the European Union, the existence of these social network intermediaries, these "man behind the curtain" systems, may threaten the ability of Internet users to fully preserve their rights.

Why should we care that third parties are processing data that has already been made publicly available by Internet users? It cannot be overlooked that these social network intermediaries do not merely "process" and "store" data. Their systems take publicly available data, and by aggregating Internet users activity across multiple social networks, they enable brands to create a profile of these Internet users and all of their interactions. While the original data may be publicly available, these systems allow for aggregation, commentary, campaigns and brand interactions that form an entirely new set of data that the brand gets to leverage and the intermediary has to store.

The unsettling legal existence of the social network intermediary should be examined in three ways: 1) the ability of the social network intermediary to give meaningful notice to the Internet user whose data is being processed; 2) the ability of the social network intermediary to gain meaningful consent from the Internet user; and 3) the ability of the social network intermediary to engage in data deletion for those Internet users who wish to "be forgotten."

GIVING MEANINGFUL NOTICE AND GAINING MEANINGFUL CONSENT

Much like the man behind the curtain, the social network intermediary's intent is to remain unknown: their business purpose is to empower brands to look like they are masters of social

media and public relations in this digital age. This invisibility to the Internet user, however, smacks against society's notions, regulatory and normative, of meaningful notice and consent when it comes to the collection, management and storage of data.

The classic method of notice, the privacy policy, is rendered wholly ineffective since the Internet user does not know where to go to even find it. Alternate notice mechanisms, as discussed in the literature regarding notice, may also be ineffective for the same reason since the Internet user is likely unaware the third party even exists.[3] The consent problem is relatively straightforward: I cannot say "yes" or " no" if I do not know that you exist.

These social network intermediaries have the same obligations as any other company to comport with domestic and international privacy laws. Many of the companies that operate as social network intermediaries, in fact, do have privacy policies and comply with international privacy standards. In searching the Department of Commerce's EU-US Safe Harbor database of companies that are certified as complying with the 1995 EU Privacy Directive, you can find many of these U.S.-based companies listed as being Safe Harbor compliant.[4] While these companies may have done what they needed to do to comply with the letter of existing laws, the spirit of these laws is not honored since the Internet user does not know the social network intermediary is operating with their information, even if it is publicly available.

The EU Data Privacy Directive appears to account for this relationship between the brand and the social network intermediary: it has set out requirements and obligations for data controllers, those who may be the original source of data input, and companies who act as data processors, merely providing the vehicle for the data to be manipulated within.[5] There is a meaningful distinction when it comes to social network intermediaries between the entity that controls the data in question and the entity that merely processes it. Practically, when the social network intermediary's relationship is executed with the brand, through a vendor contract, normally a licensing agreement of some sort for platform use, it is usually accompanied by a Data Transfer Agreement (DTA) that is executed with provisions known as the Standard Contractual Clauses.[6] These clauses painfully detail the obligation of the data controller and the data processor as well as what types of information are applicable to cross-border transfer in that particular situation.

While the obligations may be clear to the parties involved in the contractual relationship, the public's inability to know of the existence of all parties strips them of their rights to voice concerns or file grievances with the appropriate authorities under these agreements, such as the Federal Trade Commission (FTC), the European data protection authorities (DPAs) or the Swiss Federal Data Protection and Information Commissioner (FDPIC). The reasonable expectation of the data subjects, the Internet user, has received limited treatment as to the liability assigned between the controller and the processor vis-à-vis one another, but this reasonable expectation must also be considered generally in terms of the public's ability to understand the relationship of all companies involved with their data, public or private.[7]

To Be Forgotten: Advances in the European Union's Approach to Data Privacy

The ultimate form of "opting out" of a platform or online system is currently being debated in Europe: data deletion. The European Commission has been exploring comprehensive reform of the EU data protection rules, incorporating the inclusion of the "right to be forgotten."

If such regulation came to pass, data controllers and processors would likely be required to delete the information of users who no longer desired to have their information stored. Spain is already enforcing the "right to be forgotten" when it comes to data that is publicly available through search engines.[8] Spain's Data Protection Agency has ordered search engine Google to delete links and information on nearly ninety people, action that Google continues to challenge. Artemi Rallo, the Director of the Spanish Data Protection Agency makes a fair point: "Google is just 15 years old, the Internet is barely a generation old and they are beginning to detect problems that affect privacy. More and more people are going

to see things on the Internet that they don't want to be there."[9]

All of the cases involving Google share the same genesis – the individuals petitioned the Spanish agency to have the information removed from Google's index. While it is apparent in the case of Google that it was Google that had the power to remove the information about the individuals (after all, they did "google" to find the information about themselves), the data deletion involved in the "right to be forgotten" is contingent upon a party having knowledge of all parties involved in controlling the destiny of their data.

In the case of the social network intermediary, enforcement of data deletion would be reliant on the data controller communicating to the data processor that a particular individual's data must be deleted. The Internet user would be counting on the brand to communicate to the social network intermediary to delete the information. While this obligation is something that could arguably be embedded into the amended regulatory framework, its' practical application is something else altogether. It assumes that the brand companies have invested in robust privacy practices and training practices for their employees who are on the front lines managing these requests. It also assumes that the social network intermediary has done the same.

The right to be forgotten currently faces a variety of challenges, but its adoption, which may take place in 2014, would pose issue for the uncomfortable existence of the intermediary and their responsibility to the Internet user.[10]

What To Do With That Which is Already Public

"Oh, no my dear. I'm a very good man. I'm just a very bad Wizard."

- Frank L. Baum, *The Wonderful Wizard of Oz*

The world's greatest brands utilize social network intermediaries to remain the world's greatest brands. They seek to have relationships and a level of responsiveness to the would-be consumer or fan that would not be possible without the existence of the social network intermediary's powerful "social" platform. Is it that big of a deal that brands want avenues to connect to their most loyal fans on the Web?

Society's privacy debate, as its core, is about trust in relationships. Internet users want to be able to know that the data they put out about themselves online is only being used by parties that they have given consent to and is not being used in a manner or by a party they are unaware of.

Brands using social network intermediaries are hiding something: they are concealing the fact that a third party is involved in the relationship, the man behind curtain. Their privacy policies, if they even exist, may give notice that they are using "third party services" to effectuate their relationship with their consumers and the general public, but most often they do not disclaim who these third parties are.

It must not be forgotten that the data being discussed as subject to protection has already been made public. It is data that is already out in the wild and voluntarily so. Is it not just waiting to be reined in?

The privacy we hope to enjoy is in the perception. We believe these interactions are happening directly with the brands we Like and Tweet, not the "man behind the curtain." We believe that our favorite brands have a Twitter account, and, perhaps these interactions are being stored by Twitter, but that is where it ends. Twitter has a data deletion policy; these social network intermediaries may not. In the civil law world where privacy is based on norms, perception is everything. If we look behind the curtain, we might not like what we see.

---

[1] Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 Harvard L. Rev. 193 (1890).

[2] Council Directive 95/46, 1995 O.J. (L 281) 0031-0050 (EC).

[3] M. Ryan Calo, *Code, Nudge or Notice?,* University of Washington School of Law Research Paper No. 2013-04 (February 13, 2013).

[4] Department of Commerce, EU-US Safe Harbor Home Page, available at http://export.gov/safeharbor/eu/eg_main_018365.asp.

[5] Council Directive 95/46, 1995 O.J. (L 281) 0031-0050 (EC).

[6] Commission Decision, 2010/87/EU, 2010 O.J. (L 39) 5-6, 11 (EU).

[7] Article 29 Working Party, 'Opinion 10/2006 on the processing of personal data by the Society for Worldwide Interbank Financial Telecommunication (SWIFT)' (WP 128, 22 November 2006).

[8] AOL News, Fox News Latino, *Spain Asserts a 'Right to be Forgotten,' Ordering Google to Remove Some Old Links*, April 21, 2011, available at http://noticias.aollatino.com/2011/04/21/right-to-be-forgotten-google/.

[9] *Id*.

[10] Eric Pfanner, *Archivists in France Fight a Privacy Initiative*, The New York Times, June 16, 2013, available at http://www.nytimes.com/2013/06/17/technology/archivists-in-france-push-against-privacy-movement.html?pagewanted=all.

BIG DATA IN SMALL HANDS

*Woodrow Hartzog & Evan Selinger**

66 STAN. L. REV. ONLINE 81

"Big data" can be defined as a problem-solving philosophy that leverages massive datasets and algorithmic analysis to extract "hidden information and surprising correlations."[1] Not only does big data pose a threat to traditional notions of privacy, but it also compromises socially shared information. This point remains underappreciated because our so-called public disclosures are not nearly as public as courts and policymakers have argued—at least, not yet. That is subject to change once big data becomes user friendly.

Most social disclosures and details of our everyday lives are meant to be known only to a select group of people.[2] Until now, technological constraints have favored that norm, limiting the circle of communication by imposing transaction costs—which can range from effort to money—onto prying eyes. Unfortunately, big data threatens to erode these structural protections, and the common law, which is the traditional legal regime for helping individuals seek redress for privacy harms, has some catching up to do.[3]

To make our case that the legal community is under-theorizing the effect big data will have on an individual's socialization and day-to-day activities, we will proceed in four steps.[4] First, we explain why big data presents a bigger threat to social relationships than privacy advocates acknowledge, and construct a vivid hypothetical case that illustrates how democratized big data can turn seemingly harmless disclosures into potent privacy problems. Second, we argue that the harm democratized big data can inflict is exacerbated by decreasing privacy protections of a special kind—ever-diminishing "obscurity." Third, we show how central common law concepts might be threatened by eroding obscurity and the resulting difficulty individuals have gauging whether social disclosures in a big data context will sow the seeds of forthcoming injury. Finally, we suggest that one way to stop big data from causing big, unredressed privacy problems is to update the common law with obscurity-sensitive considerations.

I. BIG, SOCIAL DATA

For good reason, the threat big data poses to social interaction has not been given its due. Privacy debates have primarily focused on the scale of big data and concentrations of power—what big corporations and big governments can do with large amounts of finely analyzed information. There are legitimate and pressing concerns here, which is why scholars and policymakers focus on Fair Information Practice Principles (FIPPs), deidentification techniques, sectoral legislation protecting particular datasets, and regulatory efforts to improve data security and safe international data transfers.[5]

This trajectory fails to address the full scope of big data as a disruptive force in nearly every sector of the patchwork approach to privacy protection in the United States. Individuals eventually will be able to harness big datasets, tools, and techniques to expand dramatically the number and magnitude of privacy harms to themselves and others, perhaps without even realizing it.[6] This is problematic in an age when so many aspects of our social relationships with others are turned into data.

Consider web-scraping companies that dig up old mugshots and showcase them online, hoping embarrassed or anxious citizens will pay to have

* Woodrow Hartzog is Assistant Professor, Cumberland School of Law, Samford University; Affiliate Scholar, Stanford Center for Internet and Society. Evan Selinger is Associate Professor of Philosophy, Rochester Institute of Technology; Fellow, Institute for Ethics and Emerging Technology.

their images taken down. It isn't hard to imagine that the next generation of this business will cast a wider net, capitalizing on stockpiles of aggregated and filtered data derived from diverse public disclosures. Besides presenting new, unsettling detail about behavior and proclivities, they might even display predictive inferences couched within litigation-buttressing weasel wording—e.g., "correlations between X and Y have been known to indicate Z." Everyone, then, will be at greater risk of unintentionally leaking sensitive personal details. Everyone will be more susceptible to providing information that gets taken out of its original context, becomes integrated into a new profile, and subsequently harms a friend, family member, or colleague.

Inevitably, those extracting personal details from big data will argue that the information was always apparent and the law should not protect information that exists in plain sight.[7] The law has struggled with protecting privacy in public long before big data. However, we envision a tipping point occurring whereby some pro-publicity precedent appears more old than wise.

II. More Data, Less Obscurity

Socialization and related daily public disclosures have always been protected by varying layers of obscurity, a concept that we previously defined as follows:

> Obscurity is the idea that when information is hard to obtain or understand, it is, to some degree, safe. Safety, here, doesn't mean inaccessible. Competent and determined data hunters armed with the right tools can always find a way to get it. Less committed folks, however, experience great effort as a deterrent.
>
> Online, obscurity is created through a combination of factors. Being invisible to search engines increases obscurity. So does using privacy settings and pseudonyms. Disclosing information in coded ways that only a limited audience will grasp enhances obscurity, too. Since few online disclosures are truly confidential or highly publicized, the lion's share of communication on the social web falls along the expansive continuum of obscurity: a range that runs from completely hidden to totally obvious.[8]

In the past, individuals have been able to roughly gauge whether aspects of their daily routines and personal disclosures of information would be safeguarded at any appropriate level of privacy protection by (sometimes implicitly) guessing the likelihood their information would be discovered or understood by third parties who have exploitative or undesirable interests. In the age of big data, however, the confidence level associated with privacy prognostication has decreased considerably, even when conscientious people exhibit due diligence.

Increasingly powerful and often secretive (proprietary and governmental) algorithms combined with numerous and massive datasets are eroding the structural and contextual protections that imposed high transactional costs on finding, understanding, and aggregating that information. Consumers got a taste of both the ease and power in which these processes can occur when Facebook rolled out Graph Search, denied it had privacy implications, then also revealed how readily what we "like" gets translated into who we are.

Maintaining obscurity will be even more difficult once big data tools, techniques, and datasets become further democratized and made available to the non-data-scientist masses for free or at low cost. Given recent technological trends, this outcome seems to be gradually approaching inevitability. At the touch of a button, Google's search engine can already unearth an immense amount of information that not too long ago took considerable effort to locate. Looking ahead, companies like Intel are not shy about letting the public know they believe "data democratization is a good bet."[9]

Decreasing confidence in our ability to judge the privacy value of disclosures puts us on a collision course for deepening the problem of "bounded rationality" and, relatedly, what Daniel Solove recognized as the problems of scale, aggregation, and assessing harm.[10] It appears that the courts will need to grapple with a new

wave of allegations of harms arising from behavior that yielded unintended and unforeseeable consequences.

As a thought experiment that crystalizes our guiding intuitions, consider a big data update to the problems that occurred when college students were revealed to be gay to their disapproving parents after a third party added them as members to Facebook's Queer Chorus group.[11] In the original instance, the salient tension was between how Facebook described its privacy settings and what users expected when utilizing the service. But what if someday a parent, teacher, or other authority figure wanted to take active steps to determine if their child, student, or employee was gay? Using democratized big data, a range of individually trivial, but collectively potent, information could be canvassed. Geolocation data conveyed when the child, or, crucially, his or her friends, used services like Foursquare combined with increasingly sophisticated analytical tools could lead to a quick transition from checking in to being outed. People-search services like Spokeo are well positioned to offer such user-friendly big data services.

III. The Common Law Privacy Implications of Big Data for Everyone

Once big data is democratized and obscurity protections are further minimized, peer-to-peer interactions are poised to challenge many traditional common law concepts. Because the courts already make inconsistent rulings on matters pertaining to what reasonable expectations of privacy are, tort law is especially vulnerable.[12]

Here are a few of the fundamental questions we expect the courts will struggle to answer:

What Constitutes a Privacy Interest? A crucial question for both the tort of public disclosure of private facts and the tort of intrusion upon seclusion is whether the plaintiff had a privacy interest in a certain piece of information or context. This determination has varied wildly among the courts, and it is unclear how ubiquitous big data will alter this. For example, some courts have found that a privacy interest exists in involuntary exposure in public.[13] Other courts have found that overzealous surveillance in public that reveals confidential data can be seen to violate a privacy interest.[14] Will invasive "dataveillance" trigger the same protections?[15] Finally, courts have found, albeit inconsistently, a privacy interest in information known only to, and likely to stay within, a certain social group.[16] Does an increased likelihood that such information might be ascertained by outsiders destroy the privacy interest in information shared discreetly in small groups?[17]

What Actions Are Highly Offensive? Directly revealing or gaining access to certain kinds of information has been found to be highly offensive for purposes of the disclosure, intrusion, and false light torts.[18] In an age of predictions based upon data, would indirect disclosures of private information also be considered highly offensive? If not, does the law need to better articulate these limits? Does it matter if the eventual revelation of certain kinds of information that is highly offensive was predictable? Regarding the intrusion tort, can information gleaned from "public" big datasets ever be considered "secluded" and, if so, would using tools to unearth such data ever be considered highly offensive to a reasonable person?[19]

What Kinds of Disclosures Breach a Confidence? When has a confidant disclosed enough indirect information effectively to breach a confidence? If revealing a friend's location more than once a week allows others to determine that he is visiting a doctor for treatment of a communicable disease—a secret you promised to keep confidential—have you breached your promise? Courts would likely be hesitant to find a breach if the link between the disclosure and revealed confidential information were speculative, though inevitably some indirect disclosures will be so likely to compromise the confidentiality of other pieces of information so as to result in a *de facto* disclosure of the information itself. Should contracts with privacy-protective terms between individuals and small groups contemplate potential uses in big data? What lengths must confidants go to protect facts from being uncovered via big data techniques?

IV. Regulating the Big Impact of Small Decisions

Given the powerful debate over large-scale regulation of big data, safeguarding smaller, peer-to-peer interaction may prove to be the most feasible and significant privacy-related protection against big data.[20] The concept of obscurity might be useful in guiding the common law's evolution. If embraced as part of the disclosure and intrusion privacy torts, obscurity would allow socially shared information to fall within the ambit of "private facts" and "secluded" contexts. Contracts could also be used to protect the obscurity of individuals by targeting big data analysis designed to reveal socially shared but largely hidden information. Those charged with interpreting broad privacy-related terms should keep in mind structural and contextual protections that might have been relied upon by those whose privacy was to be protected.

Those forming the common law can now choose one of two paths. They can cling to increasingly ineffective and strained doctrines that were created when structural and contextual protections were sufficient for most of our socialization and obscure activities in public. Or they can recognize the debilitating effect big data has on an individual's ability to gauge whether social disclosures and public activity will later harm themselves and others, and evolve the common law to keep small acts of socialization and our day-to-day activities from becoming big problems.

---

[1] Ira Rubinstein, *Big Data: The End of Privacy or a New Beginning?*, 3 Int'l Data Privacy L. 65, 74 (2013). The term "big data" has no broadly accepted definition and has been defined many different ways. *See* Viktor Mayer-Schönberger & Kenneth Cukier, Big Data: A Revolution that Will Transform how We Live, Work, and Think 6 (2013) ("There is no rigorous definition of big data . . . . One way to think about the issue today . . . is this: big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value . . . .").

[2] *See, e.g.*, Lior Jacob Strahilevitz, *A Social Networks Theory of Privacy*, 72 U. Chi. L. Rev. 919 (2005).

[3] *See* Patricia Sánchez Abril, *Recasting Privacy Torts in a Spaceless World*, 21 Harv. J.L. & Tech. 1, 19-20 (2007); Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 Calif. L. Rev. 1805, 1827 (2010); Andrew Jay McClurg, *Bringing Privacy Law Out of the Closet: A Tort Theory of Liability for Intrusions in Public Places*, 73 N.C. L. Rev. 989, 1057 (1995); Neil M. Richards, *The Limits of Tort Privacy*, 9 J. Telecomm. & High Tech. L. 357, 383 (2011); Neil M. Richards & Daniel J. Solove, *Prossers Privacy Law: A Mixed Legacy*, 98 Calif. L. Rev. 1887, 1889 (2010); Harry Surden, *Structural Rights in Privacy*, 60 SMU L. Rev. 1605 (2007).

[4] A notable exception is Paul M. Schwartz and Daniel J. Solove's *Reworking Information Privacy Law: A Memorandum Regarding Future ALI Projects About Information Privacy Law* (Aug. 2012), http://law.duke.edu/sites/default/files/images/centers/judicialstudies/Reworking_Info_Privacy_Law.pdf. They write:

> People also expect "privacy by obscurity," that is, the ability to blend into a crowd or find other ways to be anonymous by default. This condition is rapidly disappearing, however, with new technologies that can capture images and audio nearly everywhere. As an example, facial recognition technology is constantly improving. Already, Facebook and Apple use technologies that permit the automatic tagging of photographs. One day devices, such as Google Glasses, may permit the identification of passing pedestrians on the street. In short, if the privacy torts are to be rethought, more guidance must be provided as to the underlying concept of privacy.

Id. at 11 (citations omitted).

[5] *See, e.g.*, Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701, 1776 (2010) ("Easy reidentification represents a sea change not only in technology but in our understanding of privacy."); Rubinstein, *supra* note 1, at 74; Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 Nw. J. Tech. & Intell. Prop. 239, 256-57 (2013); Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data*, 64 Stan. L. Rev. Online 63 (2012); Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. Colo. L. Rev. 1117 (2013); danah boyd, Address at the WWW2010 Conference: "Privacy and Publicity in the Context of Big Data" (Apr. 29, 2010),http://www.danah.org/papers/talks/2010/WWW2010.html. *But see* Jane Yakowitz, *Tragedy of the Data Commons*, 25 Harv. J.L. & Tech. 1 (2011).

[6] Although we're focusing on how the law should respond to the dark side of big data, some see mastering quantitative legal prediction as essential to the future of entrepreneurial law firms and the law schools that train students to work in them. *See, e.g.*, Daniel Martin Katz, *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 Emory L.J. 909 (2013).

[7] *See* Mayer-Schönberger & Cukier, *supra* note 1, at 29 (describing one instance of information discovered from big data analysis as "always apparent [as] [i]t existed in plain sight").

[8] Woodrow Hartzog & Evan Selinger, *Obscurity: A Better Way to Think About Your Data than 'Privacy,'* Atlantic (Jan. 17, 2013),http://www.theatlantic.com/technology/archive/2013/01/obscurity-a-better-way-to-think-about-your-data-than-privacy/267283 (explaining how obscurity is the proper conceptual framework for analyzing the privacy implications that follow from the introduction of Graph to Facebook's interface and analytics); *see also* Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity*, 101 Calif. L. Rev. 1 (2013) (identifying four key factors that define an obscurity

continuum); Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. L. REV. 385 (2013) (explaining how obscurity considerations can enhance privacy by design efforts); Fred Stutzman & Woodrow Hartzog, *Boundary Regulation in Social Media*, *in* PROCEEDINGS OF THE ACM 2012 CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK 769 (2012), *available at* http://dl.acm.org/citation.cfm?id=2145320&bnc=1 (observing that the creation of obscurity is part of the boundary regulation process of social media users).

[9] *See* Jordan Novet, *Why Intel Thinks Data Democratization is a Good Bet*, GIGAOM (May 30, 2013), http://gigaom.com/2013/05/30/why-intel-thinks-data-democratization-is-a-good-bet.

[10] Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1879, 1888-93 (2013) ("The point is that it is virtually impossible for a person to make meaningful judgments about the costs and benefits of revealing certain data."); *see, e.g.*, Alessandro Acquisti & Jens Grossklags, *Privacy and Rationality: A Survey*, *in* Privacy and Technologies of Identity: A Cross-Disciplinary Conversation 15, 16 (Katherine R. Strandburg & Daniela Stan Raicu eds., 2006); Danielle Keats Citron, *Reservoirs of Danger: The Evolution of Public and Private Law at the Dawn of the Information Age*, 80 S. CAL. L. REV. 241 (2007); Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1609, 1661 (1999) ("The difficulty with privacy-control in the Information Age is that individual self-determination is itself shaped by the processing of personal data.").

[11] Geoffrey A. Fowler, *When the Most Personal Secrets Get Outed on Facebook*, WALL ST. J. (Oct. 13, 2012), http://online.wsj.com/article/SB10000872396390444165804578008740578200224.html.

[12] *See, e.g.*, Strahilevitz, *supra* note 2, at 921.

[13] *See, e.g.*, Daily Time Democrat v. Graham, 162 So. 2d 474, 478 (Ala. 1964).

[14] *See, e.g.*, Nader v. Gen. Motors Corp., 255 N.E.2d 765, 771 (N.Y. 1970) ("[I]t is manifest that the mere observation of the plaintiff in a public place does not amount to an invasion of his privacy. But, under certain circumstances, surveillance may be so 'overzealous' as to render it actionable . . . . A person does not automatically make public everything he does merely by being in a public place."); Kramer v. Downey, 680 S.W.2d 524, 525 (Tex. App. 1984).

[15] *See* Roger Clarke, ROGER CLARKE'S DATAVEILLANCE AND INFORMATION PRIVACY HOME-PAGE, http://www.rogerclarke.com/DV (last updated Jan. 6, 2013) (defining dataveillance as "the systematic use of personal data systems in the investigation or monitoring of the actions or communications of one or more persons"); *see also* Jerry Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193, 1261 (1998) (arguing that "information collection in cyberspace is more like surveillance than like casual observation"); Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53STAN. L. REV. 1393, 1417 (2001) ("Dataveillance is thus a new form of surveillance, a method of watching not through the eye or the camera, but by collecting facts and data."); Katherine J. Strandburg, *Freedom of Association in a Networked World: First Amendment Regulation of Relational Surveillance*, 49B.C. L. REV. 741, 761 (2008) (observing that "[j]ust as "dataveillance" can chill an individual's experimentation with particular ideas or pastimes, relational surveillance can chill tentative associations and experimentation with various group identities").

[16] *See, e.g.*, Y.G. v. Jewish Hosp., 795 S.W.2d 488 (Mo. Ct. App. 1990).

[17] *See* Strahilevitz, *supra* note 2 at 922.

[18] *See, e.g.*, Nappier v. Jefferson Standard Life Ins. Co., 322 F.2d 502 (4th Cir. 1963) (identity of a rape victim); Crippen v. Charter Southland Hosp. Inc., 534 So. 2d 286 (Ala. 1988) (confidential medical data); Taylor v. K.T.V.B., Inc., 525 P.2d 984 (Idaho 1974) (nude photos); Brents v. Morgan, 299 S.W. 967 (Ky. 1927) (debts); Reid v. Pierce Cnty., 961 P.2d 333 (Wash. 1998) (autopsy photos).

[19] *See* Citron, *Mainstreaming Privacy Torts*, *supra* note 3, at 1827 ("[P]laintiffs probably cannot sue database operators for intrusion on seclusion under current case law. To prevail in an intrusion suit, a plaintiff must show that a defendant invaded his physical solitude or seclusion, such as by entering his home, in a manner that would be highly offensive to the reasonable person. Database operators and data brokers, however, never intrude upon a plaintiff's private space. They do not gather information directly from individuals and, to the extent that they do, the privacy problem involves the failure to secure personal information, not its collection.") (citations omitted). *But see* Jane Yakowitz Bambauer, *The New Intrusion*, 88 Notre Dame L. Rev. 205, 207 (2012) ("Intrusion has great, untapped potential to address privacy harms created by advances in information technology. Though the tort is associated with conduct in real space, its principles apply just as well to operations in the era of Big Data."); Lyrissa Barnett Lidsky, *Prying, Spying, and Lying: Intrusive Newsgathering and What the Law Should Do About It*, 73 TUL. L. REV. 173, 227 (1998) ("[S]everal recent examples indicate that the average citizen's privacy is protected from media intrusions primarily by media disinterest, a tenuous basis at best for privacy protection."); McClurg, *supra* note 3, at 1057 ("The tort of intrusion can be redefined in a way that would allow recovery in suitable cases of public intrusion while also accommodating the competing interests of free social interaction and free speech."); Richards, *supra* note 3 at 383 ("[I]f we are interested in protecting against what we colloquially call 'invasions of privacy,' the intrusion model is a better fit with our intuitive linguistic understandings of that metaphor.").

[20] For a skeptical view on the likelihood of significant regulation limiting how businesses mine data, see Lior Jacob Strahilevitz, *Toward a Positive Theory of Privacy Law*, 126 HARV. L. REV. 2010, 2033 (2013) ("The deck is stacked against restrictions on data mining."). *Cf.* Citron, *Reservoirs of Danger*, *supra* note 10 at 296 (asserting that, as a private law response to privacy harms, "[t]he contours of a negligence regime are simply too uncertain, and inherent problems with its enforcement undermines optimal deterrence," and proposing a strict-liability response instead); Sarah Ludington, *Reigning in the Data Traders: A Tort for the Misuse of Personal Information*, 66 MD. L. REV. 140, 146 (2006) (proposing a tort to target "insecure data practices" and "the use of personal information data for purposes extraneous to the original transaction").

THE GLASS HOUSE EFFECT:

WHY BIG DATA IS THE NEW OIL, AND WHAT TO DO ABOUT IT

*Dennis Hirsch**

"Data is the new oil," Clive Humbly announced in 2006.[1] More recently, IBM CEO Virginia Rometty updated the phrase, explaining that "big data" is the new oil.[2] The analogy resonates. Data flows like oil. One must "drill down" into data to extract value from it. Data is an essential resource that powers the information economy in much the way that oil has fueled the industrial economy. Data promises a plethora of new uses – diagnosis of diseases, direction of traffic patterns, etc. – just as oil has produced useful plastics, petrochemicals, lubricants, gasoline, and home heating. "Big data is the new oil" has not only become a popular meme; it is a banner behind which we can march, an optimistic declaration of the way forward.

Such comparisons ignore oil's negative side. Tankers run aground and spill their deadly black cargo. The Deepwater Horizon platform collapses in flames and raw oil gushes into the Gulf for weeks. This too must be included in the analogy. Data spills occur with the regularity of oil spills. The victim of identity theft, bogged down in unwanted credit cards and bills, is just as trapped and unable to fly as the bird caught in the oil slick, its wings coated with a glossy substance from which it struggles to free itself.

As the data sets get bigger the threat, too, grows. Big data is like a massive oil tanker navigating the shoals of computer-savvy criminals and human error. Yes, big data make us smarter and wealthier and our lives better. But that dream has a dark, viscous underside that threatens to pollute the information ecosystem.

How to proceed? Environmental law reduces oil pollution without undermining the fossil fuel-based economy. Can we look to it for strategies that will allow us to reap big data's many benefits, while reducing its negative impacts?

The history of oil pollution law is highly instructive. In the 19th Century, judges and legislators shaped the law to encourage the production and transportation of oil. Maritime tort law recognized property damage from oil spills, but not injuries to fishing, tourism and other such affected industries. Traditional tort doctrines required plaintiffs to show negligence—a difficult task in a risky field where even the careful could spill their cargo. Collective action and free rider problems further reduced the incentives to bring such a suit since, when many suffer a small harm, no single person has the incentive to bring the suit or to organize the group to do so. Finally, as if tort liability were not yet sufficiently constrained, Congress passed the Limited Liability Act of 1851 which capped oil spill damages at the value of the vessel and freight remaining after the accident.[3] This statute, whose original purpose was to facilitate the transportation of otherwise uninsurable cargo, came to produce absurd results. The 1967 wreck of the Torrey Canyon oil tanker, which spilled over 100,000 tons of crude oil into the English channel and despoiled 100 miles of French and British coasts, resulted in only $50 in damages—the value of the sole remaining lifeboat.[4] Clearly, something needed to be done.

Congress gave its answer in the 1970 Clean Water Act[5] and, responding to the massive Exxon Valdez oil spill, the 1990 Oil Pollution Act.[6] Together, these statutes re-write oil

* Geraldine W. Howell Professor of Law, Capital University Law School. The author would like to thank Professor Paul Ohm for suggesting the idea for this paper, and for early discussions that helped to shape it. Unless otherwise indicated, the author alone is responsible for the paper's content.

pollution law. They allow the government to clean up an oil spill and then bring an action against the responsible party to recoup the clean-up costs.[7] This overcomes the collective action and free rider problems that undermine private tort actions. The Oil Pollution Act recognizes new causes of action for damage to economic, as opposed to property, interests.[8] The statutes provide for strict liability, thereby relieving plaintiffs of the difficult task of demonstrating negligence. They greatly increase the liability limits.[9] Finally, the Oil Pollution Act requires all new oil transportation vessels operating in U.S. waters to employ double hull technology that greatly reduces the chance of an oil spill.[10] The statutory scheme has reduced spills by oil-transporting vessels.

This environmental law success story offers important lessons for big data. Like the early laws governing the oil industry, today's doctrines appear designed to encourage the production and transfer of the "new oil." Following a data spill, courts generally allow damages only for the concrete economic injuries associated with identity theft. They refuse to recognize the other, non-economic damages that data spills create – the increased risk of identity theft and the anxiety that that risk produces; the sense of violation and exposure that comes from release of one's personal data.[11] As in the oil pollution context, negligence is difficult to prove in the complex area of data security. Collective action and free-rider problems abound. Why should any individual bring the suit that requires a company to provide increased data security for *all* its customers? Data breach notification statutes require firms to bear the cost of providing notice to affected persons, but not the full cost of the injuries that their breach has caused. While these laws provide a notice and deterrent function that makes them far more useful than the 1851 Limited Liability Act, the liability that they create is limited. Why should we wait for the big data equivalent of the Exxon Valdez spill to change this system and require companies to internalize the full costs of their data security breaches? Big data has arrived. We no longer need to design the law to subsidize it. Rather, we need laws that require big data to internalize its externalities and so make the information economy sustainable in the long term.

Environmental law provides a possible model for doing this. As with the Clean Water Act and Oil Pollution Act, Congress could pass legislation that authorizes, and provides funding for, a government agency to clean up after data spills (e.g. to identify root causes, assess the extent of the breach, and provide credit monitoring and identity theft recovery services to consumers). The agency could then seek reimbursement from the responsible party. This would overcome the collective action and free-rider problems that would otherwise inhibit private lawsuits. Like the Oil Pollution Act, such legislation could also expand tort liability and require courts to recognize the non-economic damages that data spills create. It could establish strict liability for data spills and so eliminate the need to prove defendant's negligence. Just as the OPA requires ships to adopt an environmentally-protective design, so the legislation could require firms to adopt privacy by design. If oil tankers must use double hulls, perhaps data security systems should have to employ two-factor identification.[12] Taken together, these measures could reduce data spills significantly just as the Oil Pollution Act has lessened oil spills.

While such measures would be productive, they will address only part of the problem. A further exploration of the environmental analogy suggests why this is so. Oil does not only lead to oil spills. It also produces carbon emissions that accumulate in the atmosphere and contribute to the greenhouse effect. This warms the earth, disturbs ecosystems, and makes the world less hospitable to humans and other species.[13] In much the same way, big data is generating layers upon layers of personal information at an extraordinarily rapid pace.[14] This creates, not a greenhouse effect, but a *glass house effect*. It is as if we were increasingly living in a glass house whose transparent walls allowed the hot glare of public scrutiny to reach in and scorch our most private selves. What else can we call it when companies store and mine our search queries, e-mail messages and web activity and share them with each other, or with the National Security Agency (NSA)? Climate change acts on the physical world. The glass house effect acts on our inner lives. It focuses too much hot light on us and, in so doing, stunts the growth of the "inviolate

personality"[15] which requires shade and shelter in which to flourish. Like the greenhouse effect, the glass house effect produces conditions that are less favorable to life – to a full, human life. If the growth of big data continues on its current track we will pass on to our children a depleted ecosystem for the cultivation of the human personality.

The environmental analogy can point us towards solutions to this problem. The long-term solution to climate change is the development of clean energy technologies—solar, wind, hydro and geothermal power--that can substitute for oil and produce far smaller environmental impacts. The same should be true for big data. The answer is not simply to punish those who spill data. It is to prevent such spills, and reduce the glass house effect, through new "clean data" technologies and privacy-protective business models. Recently, the United States and other countries have engaged in a massive push to develop clean energy technologies. They know that these innovations are needed, not only for preserving health and quality of life at home, but for economic competitiveness in the global marketplace. As data sets grow larger and larger, could the desire for privacy and consumer trust ramp up demand for clean data technologies? Could these innovations, too, be technologies of the future that form the basis, not only of better data security and privacy protection, but also of a "clean data" sector that makes us more competitive? Should we fund a push for innovation with respect to encryption, data anonymization and other clean data technologies?[16] Should venture capitalists look to this new field as an important investment opportunity?

The optimistic claim that "big data is the new oil" is indeed helpful. It both shows us the tremendous upside of this new phenomenon, and points to the threats that big data, like oil, poses. It should motivate us to find sustainable ways to utilize this highly valuable new resource—methods that allow us to enjoy the benefits of big data, while preserving fertile ground for personal development.

[1] Clive Humbly, ANA Senior marketer's summit, Kellogg School (2006); *see* Michael Palmer, *Data is the New Oil,* blog post available at http://ana.blogs.com/maestros/2006/11/data_is_the_new_oil.

[2]http://siliconangle.com/blog/2013/03/11/ibms-ceo-says-big-data-is-like-oil-enterprises-need-help-extracting-the-value/

[3] 46 U.S.C. 183 (1988).

[4] Jeffrey D. Morgan, *The Oil Pollution Act of 1990: A Look at its Impact on the Oil Industry,* 6 FORD. ENV. L. J. 1, 2 (1994)

[5] Pub. L. No. 91-224, 84 Stat. 91 (1970) (codified as amended in scattered sections of 33 U.S.C.)

[6] Pub. L. No. 101-380, 104 Stat. 484 (1990).

[7] Water Quality Improvement Act, § 11(c)(1), 84 Stat. at 93.

[8]Oil Pollution Act § 1002(b)(2), 104 Stat. at 490; *see also* Kenneth M. Murchison, *Liability Under the Oil Pollution Act: Current Law and Needed Revisions*, 71 LA. L. REV. 917 (2011).

[9] Oil Pollution Act, § 1004(a)(1), 104 Stat. at 491-92. The original Oil Pollution Act raised the limits to the greater of $1200 per ton, or $10,000,000 for a vessel greater than 3000 tons, or $2,000,000 for a smaller vessel. *Id.*

[10] 46 U.S.C. 3703(a) (1988 & Supp. IV 1992).

[11] *See, e.g., Pinero v. Jackson Hewitt Tax Service, Inc.*, No. 08-3535, 2009 U.S. Dist. LEXIS 660, (E.D. La. January 7, 2009); Kelley, Drye & Warren, Consumer Finance Law Blog, *Fears of Future Identity Theft Generally Not Sufficient to Establish Actual Damages in a Lawsuit,* available at http://www.consumerfinancelawblog.com/2009/03/articles/privacy/fears-of-future-identity-theft-generally-not-sufficient-to-establish-actual-damages-in-a-lawsuit/ (last visited June 30, 2013).

[12] Professor Paul Ohm made this connection and generously shared it with me. Two-factor identification can be defined as "a security process in which the user provides two means of identification, one of which is typically a physical token, such as a card, and the other of which is typically something memorized, such as a security code. In this context, the two factors involved are sometimes spoken of as *something you have* and *something you know*. A common example of two-factor authentication is a bank card: the card itself is the physical item and the personal identification number (PIN) is the data that goes with it." http://searchsecurity.techtarget.com/definition/two-factor-authentication (last visited June 30, 2013).

[13] *See e.g.,* International Panel on Climate Change, Fourth Assessment Report: Climate Change (2007) (describing scientific findings on climate change).

[14] The amount of data in the world is doubling every two years. Steve Lohr, *The Age of Big Data*, N.Y. TIMES (Feb. 11m, 2012), available at http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all (last visited June 30, 2013).

[15] Louis Brandeis & Samuel Warren, *The Right to Privacy,* 4 HARV. L. REV. 193, 205 (1890).

[16] Professor Ohm suggested the importance of such a project.

HOW THE FAIR CREDIT REPORTING ACT REGULATES BIG DATA

*Chris Jay Hoofnagle**

INTRODUCTION

This short essay makes two observations concerning "big data." First, big data is not new. Consumer reporting, a field where information about individuals is aggregated and used to assess credit, tenancy, and employment risks, achieved the status of big data in the 1960s. Second, the Fair Credit Reporting Act of 1970 (FCRA) provides rich lessons concerning possible regulatory approaches for big data.

Some say that "big data" requires policymakers to rethink the very nature of privacy laws. They urge policymakers to shift to an approach where governance focuses upon "the usage of data rather than the data itself."[1] Consumer reporting shows us that while use-based regulations of big data provided more transparency and due process, they did not create adequate accountability. Indeed, despite the interventions of the FCRA, consumer reporting agencies (CRAs) remain notoriously unresponsive and unaccountable bureaucracies.

Like today's big data firms, CRAs lacked a direct relationship with the consumer, and this led to a set of predictable pathologies and externalities. CRAs have used messy data and fuzzy logic in ways that produce error costly to consumers. CRAs play a central role in both preventing and causing identity fraud, and have turned this problem into a business opportunity in the form of credit monitoring. Despite the legislative bargain created by the FCRA, which insulated CRAs from defamation suits, CRAs have argued that use restrictions are unconstitutional.

Big data is said to represent a powerful set of technologies. Yet, proposals for its regulation are *weaker* than the FCRA. Calls for a pure use-based regulatory regime, especially for companies lacking the discipline imposed by a consumer relationship, should be viewed with skepticism.

ORIGINS

Consumer reporting is over a century old.[2] Starting with local efforts to share information about credit risks, consumer reporting agencies began operating regionally in the 1950s and 1960s. Even then, consumer reporting would certainly qualify under any definition of "big data." The volume of data and the increasingly nationwide operations of CRAs necessitated a move from paper records to computers. Computing also enabled deeper analysis of credit risks, enabled the emergence of credit scoring, and created business models around fine-tuned credit offers, extending even into the subprime market.

Consumer reporting is essential to a modern economy. Consumer reporting can reduce credit discrimination, by focusing lenders' attention away from moral considerations to more objective financial risk factors. It reduces transaction costs for consumers, who can shop around for credit without having to establish a deep relationship with each potential creditor.

At the same time, such reporting must be performed fairly for all to enjoy the benefits of credit. Prior to the passage of the FCRA, Robert Ellis Smith recounts that CRAs collected information about sexual orientation, couples that lived out of wedlock, alcohol-consumption

* Lecturer in Residence, UC Berkeley Law.

habits, and rumors of encounters with the police. Investigators even fabricated derogatory information about individuals.[3] Congress recognized that absent a direct relationship with consumers, CRAs had inadequate incentives to treat individuals fairly. A primary purpose thus of the FCRA was to end the collection of "irrelevant" information.[4]

The FCRA is a complex statute that has been amended multiple times. Its primary provisions concern "permissible uses" of consumer credit information, requirements that data be verifiable, and access and correction rights. By complying with these safeguards, CRAs were shielded from defamation suits.

A. Permissible Uses of Consumer Reports

The FCRA's primary regulation comes in the form of "permissible" uses of consumer reports. 15 USC § 1681b specifies a range of uses, including for issuing credit, evaluating a prospective employee, underwriting an insurance policy, and a catch all "legitimate business purpose" exception for transactions initiated by the consumer. Non-enumerated uses are impermissible, thus the FCRA essentially whitelists the scope of permissible uses of data. The FCRA approach is thus very different from proposals for big data, which lean towards permitting any kind of analysis using data, and instead limiting certain decision making from analyses.

B. Maximum Possible Accuracy: A Form of Collection Limitation

In preparing a consumer report, a CRA must, "follow reasonable procedures to assure maximum possible accuracy of the information concerning the individual about whom the report relates."[5] This standard presumably becomes more stringent with time, as data collection and reporting systems improve. It is also supplemented with the duty of a CRA to verify disputed information, and in cases where data are "inaccurate or incomplete or cannot be verified," the CRA must promptly delete the disputed item.[6]

In effect, the interplay between maximum possible accuracy and the duty to verify and delete embeds a collection limitation rule in the FCRA. As noted above, prior to passage of the FCRA, embarrassing and irrelevant derogatory information was collected or fabricated by investigators. After passage of the FCRA, consumer reporting agencies were more restrained in collecting irrelevant information, because this information inherently cannot be verified. The requirement shifted consumer reporting agencies focus to verifiable credit-related information.[7]

C. Transparency and Correction Provisions

Consumers are probably most familiar with the FCRA's transparency provisions, which entitle individuals to obtain a free copy of their consumer report from each nationwide agency once a year. Additionally, consumers have the right to dispute errors on reports; this requires CRAs to conduct a "reasonable" investigation into the disputed item or delete it within thirty days.

Accountability and the FCRA

Despite the duties imposed by the FCRA, the accountability of CRAs to data subjects may charitably be described as problematic. Gone are the days where CRAs reported on couples living in various states of sin. But freed from the discipline created by the threat of defamation liability, and freed from limits upon collection of data, CRA's incentives are to minimize the costs associated with user rights to access and correction or to turn them into profit centers. For instance, after Congress imposed the responsibility to provide free consumer reports, Experian drew consumers away from the free service (annualcreditreport.com) by operating a misleadingly named site (freecreditreport.com) that sold expensive credit monitoring.[8]

The consumer reporting agencies are frequent targets of consumer suits (Westlaw produces over 1,400 suits with CRAs' names in case captions), but the systematic lack of accountability is summarized well by the following survey of Federal Trade Commission litigation against these companies.

A. Unanswered Phones

On the most basic level, it is notoriously difficult to interact with CRAs. The FTC sued all three major CRAs in 2000 because they did not answer their phones and when they did, some consumers were placed on unreasonably long holds. According to the FTC complaints, over one million calls to Experian and Trans Union went unanswered; Equifax neglected "hundreds of thousands of calls."[9] The companies paid fines and agreed to auditing to ensure adequate call availability. But a year later, Equifax paid additional fines for not answering phone calls.

B. A First Amendment Right to Ignore Use Restrictions

More fundamentally, CRAs have flouted the use restrictions imposed by the FCRA. Equifax recently settled a FTC case alleging that the company sold data in violation of use restrictions to a company that resold the data to "third parties that then used it to market products to consumers in financial distress, including companies that have been the subject of law enforcement investigations."[10]

Even more problematic and relevant to the current debate surrounding big data is the rationale for violating use restrictions—the first amendment. For instance, Trans Union was unwilling to follow use restrictions upon its data, and sold it to create target marketing lists. The company challenged use restrictions as an impingement upon its first amendment rights.[11]

C. Inaccuracy

Big data enthusiasts have argued that companies should embrace "messy" data;[12] that errors in databases actually help enhance knowledge discovery.[13] In the consumer reporting context, fuzzy matching and errors have nearly wrecked individuals' lives. One well-known anecdote concerns Judy Thomas, who sued Trans Union for regularly mixing her report with a Judith Upton. As FCRA expert Evan Hendricks explained, "Upton's Social Security number was only one digit different than Thomas' SSN. That, combined with three common letters in the first name, was sufficient to cause a regular merging of the two women's credit histories."[14]

But this problem is not just anecdotal; it is structural. In a landmark and labor intensive study, academics working in conjunction with the FTC studied almost 3,000 credit reports belonging to 1,000 consumers and found that 26 percent had "material" errors—problems serious enough to affect the consumers' credit scores.[15] Under the most conservative definition of error, this means that 23 million Americans have material errors on a consumer report. These errors matter: five percent of the study participants had errors that once corrected, improved their credit score such that they could obtain credit at a lower price.

D. The Externality of Identity Theft

The sine qua non of identity theft is the release of a consumer's report, through the trickery of an impostor. While most identity theft narratives frame this as the wrongdoing of a particular bad actor, a more nuanced look surfaces business incentives that fuel the problem.[16] Simply put, CRAs forgo revenue when they tighten security and sell fewer reports. The lost time and money paid out of pocket to resolve identity theft are externalities imposed upon consumers by CRAs and creditor grantors incentives. CRAs have capitalized on this problem by selling credit monitoring.

Conclusion

Big data enthusiasts argue that data collection rules are antiquated and that future business models should be bound mainly by use restrictions. These arguments ignore our history with FCRA, with its decades-old application of use restrictions to big data. In the FCRA context, use based approaches produced systemic unaccountability, errors that cause people financial harm, and business externalities passed off as crimes.

Like modern big data firms, consumers have no direct relationship with CRAs and no ability to limit CRAs' collection of data. Such a structure gives the individual no exit from odious practices and inadequate accountability.

---

[1] World Economic Forum, Unlocking the Value of Personal Data: From Collection to Usage 4 (Feb. 2013), available at

http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf.

[2] Evan Hendricks, CREDIT SCORES & CREDIT REPORTS 183 (Privacy Times 2007, 3rd Ed.).

[3] Robert Ellis Smith, BEN FRANKLIN'S WEB SITE, PRIVACY AND CURIOSITY FROM PLYMOUTH ROCK TO THE INTERNET 317 (Privacy Journal 2004).

[4] Anthony Rodriguez, Carolyn L. Carter & William P. Ogburn, FAIR CREDIT REPORTING 10 (NCLC Press 2002, 5th ed.).

[5] 15 USC 1681e (2013)

[6] 15 USC 1681i (a)(5)(A) (2013).

[7] Mark Furletti, *An Overview and History of Credit Reporting*, Federal Reserve Bank of Philadelphia Payment Cards Center Discussion Paper No. 02-07, June 2002, available at http://ssrn.com/abstract=927487.

[8] FEDERAL TRADE COMMISSION, MARKETER OF "FREE CREDIT REPORTS" SETTLES FTC CHARGES, "FREE" REPORTS TIED TO PURCHASE OF OTHER PRODUCTS; COMPANY TO PROVIDE REFUNDS TO CONSUMERS, Aug. 15, 2005, available at http://www.ftc.gov/opa/2005/08/consumerinfo.shtm

[9] *U.S. v. Experian Information Solutions, Inc.*, 3-00CV0056-L (N.D. Tx. 2000)(citing complaint), available at http://www.ftc.gov/os/caselist/ca300cv0056l.shtm; *U.S. v. Equifax Credit Information Services, Inc*. 1:00-CV-0087 (N.D. Ga. 2000)(citing complaint), available at http://www.ftc.gov/os/caselist/9923016.shtm; *U.S. v Trans Union LLC*, 00-C- 0235 (ND Il. 2000)(citing complaint), available at http://www.ftc.gov/os/caselist/00c0235.shtm.

[10] FTC SETTLEMENTS REQUIRE EQUIFAX TO FORFEIT MONEY MADE BY ALLEGEDLY IMPROPERLY SELLING INFORMATION ABOUT MILLIONS OF CONSUMERS WHO WERE LATE ON THEIR MORTGAGES, IN SEPARATE ACTIONS, EQUIFAX AND ITS CUSTOMERS WILL PAY A TOTAL OF $1.6 MILLION, FEDERAL TRADE COMMISSION, Oct. 10, 2012, available at http://www.ftc.gov/opa/2012/10/equifaxdirect.shtm.

[11] *Trans Union LLC v. Federal Trade Commission*, 536 U.S. 915 (2002)(J. Kennedy dissenting from denial of certiori).

[12] Viktor Mayer-Schonberger & Kenneth Cukier, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (Eamon Dolan/Houghton Mifflin Harcourt 2013).

[13] Jeff Jonas, *Big Data. New Physics*, Nov. 18, 2010, available at http://jeffjonas.typepad.com/jeff_jonas/2010/11/big-data-new-physics.html

[14] Evan Hendricks, *Oregon Jury, D.C. Circuit Continue Trans Union's Losing Streak*, 22 Privacy Times 15 (Aug. 5, 2002), available at http://www.privacytimes.com/buttons/b3_FCRA.htm.

[15] FEDERAL TRADE COMMISSION, REPORT TO CONGRESS UNDER SECTION 319 OF THE FAIR AND ACCURATE CREDIT TRANSACTIONS ACT OF 2003 (Dec. 2012), available at http://www.ftc.gov/os/2013/02/130211factareport.pdf.

[16] Chris Jay Hoofnagle, *Internalizing Identity Theft*, 13 UCLA J. L. & Tech. 1 (2009), available at http://ssrn.com/abstract=1585564

BUYING & SELLING PRIVACY:

BIG DATA'S DIFFERENT BURDENS & BENEFITS

*Joseph W. Jerome**

66 STAN. L. REV. ONLINE 47

Big data is transforming individual privacy—and not in equal ways for all. We are increasingly dependent upon technologies, which in turn need our personal information in order to function. This reciprocal relationship has made it incredibly difficult for individuals to make informed decisions about what to keep private. Perhaps more important, the privacy considerations at stake will not be the same for everyone: they will vary depending upon one's socioeconomic status. It is essential for society and particularly policymakers to recognize the different burdens placed on individuals to protect their data.

I. THE VALUE OF PRIVACY

Privacy norms can play an important role defining social and individual life for rich and poor. In his essay on the social foundations of privacy law, the dean of Yale Law School, Robert Post, argued that privacy upholds social "rules of civility" that create "a certain kind of human dignity and autonomy which can exist only within the embrace of community norms."[1] He cautioned that these benefits would be threatened when social and communal relationships were replaced by individual interactions with "large scale surveillance organizations."[2]

Today, privacy has become a commodity that can be bought and sold. While many would view privacy as a constitutional right or even a fundamental human right,[3] our age of big data has reduced privacy to a dollar figure. There have been efforts—both serious and silly—to quantify the value of privacy. Browser add-ons such as Privacyfix try to show users their value to companies,[4] and a recent study suggested that free Internet services offer $2,600 in value to users in exchange for their data.[5] Curiously, this number tracks closely with a claim by Chief Judge Alex Kozinski that he would be willing to pay up to $2,400 per year to protect his family's online privacy.[6] In an interesting Kickstarter campaign, Federico Zannier decided to mine his own data to see how much he was worth. He recorded all of his online activity, including the position of his mouse pointer and a webcam image of where he was looking, along with his GPS location data for $2 a day and raised over $2,700.[7]

"Monetizing privacy" has become something of a holy grail in today's data economy. We have seen efforts to establish social networks where users join for a fee and the rise of reputation vendors that protect users' privacy online, but these services are luxuries. And when it comes to our privacy, price sensitivity often dictates individual privacy choices. Because the "price" an individual assigns to protect a piece of information is very different from the price she assigns to sell that same piece of information, individuals may have a difficult time protecting their privacy.[8] Privacy clearly has financial value, but in the end there are fewer people in a position to pay to secure their privacy than there are individuals willing to sell it for anything it's worth.

A recent study by the European Network and Information Security Agency discovered that most consumers will buy from a more privacy-invasive provider if that provider charges a lower price.[9] The study also noted that when two companies offered a product for the same price, the more privacy-friendly provider won out. This was hailed as evidence that a pro-privacy business model could succeed, but this also anticipates that, all things being equal, one company would choose not to collect as much information as a competitor just to

* Legal and Policy Fellow, Future of Privacy Forum.

be seen as "privacy friendly." This defeats much of the benefit that a big data economy promises.

II. The Big Data Challenge

The foundations of big data rest on collecting as much raw information as possible before we even begin to understand what insight can be deduced from the data. As a result, long-standing Fair Information Practices like collection limits and purpose limitations are increasingly viewed as anachronistic,[10] and a number of organizations and business associations have called for privacy protections to focus more on how data might be used rather than limit which data can be collected.[11] The conversation has moved away from structural limitations toward how organizations and businesses can build "trust" with users by offering transparency.[12] Another suggestion is to develop business models that will share the benefits of data more directly with individuals. Online data vaults are one potential example, while the Harvard Berkman Center's "Project VRM" proposes to rethink how to empower users to harness their data and control access to it.[13] In the meantime, this change in how we understand individual privacy may be inevitable—it may be beneficial—but we need to be clear about how it will impact average individuals.

A recent piece in the *Harvard Business Review* posits that individuals should only "sell [their] privacy when the value is clear," explaining that "[t]his is where the homework needs to be done. You need to understand the motives of the party you're trading with and what [he] ha[s] to gain. These need to align with your expectations and the degree to which you feel comfortable giving up your privacy."[14] It could be possible to better align the interests of data holders and their customers, processing and monetizing data both for business and individual ends. However, the big challenge presented by big data is that the value may not be clear, the motives let alone the identity of the data collector may be hidden, and individual expectations may be confused. Moreover, even basic reputation-management and data-privacy tools require either users' time or money, which may price out average consumers and the poor.

III. Big Data and Class

Ever-increasing data collection and analysis have the potential to exacerbate class disparities. They will improve market efficiency, and market efficiency favors the wealthy, established classes. While the benefits of the data economy will accrue across society, the wealthy, better educated are in a better position to become the type of sophisticated consumer that can take advantage of big data.[15] They possess the excellent credit and ideal consumer profile to ensure that any invasion of their privacy will be to their benefit; thus, they have much less to hide and no reason to fear the intentions of data collectors. And should the well-to-do desire to maintain a sphere of privacy, they will also be in the best position to harness privacy-protection tools and reputation-management services that will cater to their needs. As a practical matter, a monthly privacy-protection fee will be easier for the wealthy to pay as a matter of course. Judge Kozinski may be willing and able to pay $200 a month to protect his privacy, but the average consumer might have little understanding what this surcharge is getting him.

The lower classes are likely to feel the biggest negative impact from big data. Historically, the poor have had little expectation of privacy—castles and high walls were for the elite, after all. Even today, however, the poor are the first to be stripped of fundamental privacy protections. Professor Christopher Slobogin has noted what he calls a "poverty exception" to the Fourth Amendment, suggesting that our expectations of privacy have been defined in ways that make the less well-off more susceptible to experience warrantless government intrusions into their privacy and autonomy.[16] Big data worsens this problem. Most of the biggest concerns we have about big data—discrimination, profiling, tracking, exclusion—threaten the self-determination and personal autonomy of the poor more than any other class. Even assuming they can be informed about the value of their privacy, the poor are not in a position to pay for their privacy or to value it over a pricing discount, even if this places them into an ill-favored category.

And big data is all about categorization. Any given individual's data only becomes useful when it is aggregated together to be exploited for good or ill. Data analytics harness vast pools of data in order to develop elaborate

mechanisms to categorize and organize. In the end, the worry may not be so much about having information gathered about us, but rather being sorted into the wrong or disfavored bucket.[17] Take the example of an Atlanta man who returned from his honeymoon to find his credit limit slashed from $10,800 to $3,800 simply because he had used his credit card at places where other people were likely to have a poor repayment history.[18]

Once everyone is categorized into granular socioeconomic buckets, we are on our way to a transparent society. Social rules of civility are replaced by information efficiencies. While this dynamic may produce a number of very significant societal and communal benefits, these benefits will not fall evenly on all people. As Helen Nissenbaum has explained, "the needs of wealthy government actors and business enterprises are far more salient drivers of their information offerings, resulting in a playing field that is far from even."[19] Big data could effectuate a democratization of information but, generally, information is a more potent tool in the hands of the powerful.

Thus, categorization and classification threaten to place a privacy squeeze on the middle class as well as the poor. Increasingly large swaths of people have little recourse or ability to manage how their data is used. Encouraging people to contemplate how their information can be used—and how best to protect their privacy—is a positive step, but a public education campaign, while laudable, may be unrealistic. Social networks, cellular phones, and credit cards—the lifeblood of the big data economy—are necessities of modern life, and assuming it was either realistic or beneficial to get average people to unplug, an overworked, economically insecure middle class does not have the time or energy to prioritize what is left of their privacy.

At present, the alternative to monetizing privacy is to offer individuals the right to make money off their information. Michael Fertik, who runs the online privacy management site, Reputation.com, sees a bright future in allowing companies to "unlock huge value in collaboration with their end users" by monetizing "the latent value of their data."[20] Startups like Personal have tried to set themselves up as individually tailored information warehouses where people can mete out their information to businesses in exchange for discounts.[21] These are projects worth pursuing, but the degree of trust and alignment between corporate and individual interests they will require are significant. Still, it is unlikely we can ever develop a one-to-one data exchange. Federico Zannier sold his personal data at a rate of $2 per day to anyone who would take it as an experiment, but average individuals will likely never be in a position to truly get their money's worth from their personal data. Bits of personal information sell for a fraction of a penny,[22] and no one's individual profile is worth anything until it is collected and aggregated with the profiles of similar socioeconomic categories.

Conclusion

While data protection and privacy entrepreneurship should be encouraged, individuals should not have to pay up to protect their privacy or receive coupons as compensation. If we intend for our economic and legal frameworks to shift from data collection to use, it is essential to begin the conversation about what sort of uses we want to take off the table. Certain instances of price discrimination or adverse employment decisions are an easy place to start, but we ought to also focus on how data uses will impact different social classes. Our big data economy needs to be developed such that it promotes not only a sphere of privacy, but also the rules of civility that are essential for social cohesion and broad-based equality.

If the practical challenges facing average people are not considered, big data will push against efforts to promote social equality. Instead, we will be categorized and classified every which way, and only the highest high value of those categories will experience the best benefits that data can provide.

---

[1] Robert C. Post, *The Social Foundations of Privacy: Community and Self in the Common Law Tort*, 77 Calif. L. Rev. 957, 959 (1989).

[2] *See id.* at 1009 (suggesting that the relationships between individuals and large organizations are "not sufficiently textured or dense to sustain vital rules of civility" and instead emphasize raw efficiency in data collection).

[3] *See, e.g.*, Griswold v. Connecticut, 381 U.S. 479, 485-86 (1965) (suggesting that constitutional guarantees create zones of privacy); Convention for the Protection of Human Rights and Fundamental Freedoms, art. 8, Nov. 4, 1950, 213 U.N.T.S. 222.

[4] Joe Mullin, *How Much Do Google and Facebook Profit from Your Data?*, ARS TECHNICA (Oct. 9, 2012, 6:38 AM PDT), http://arstechnica.com/tech-policy/2012/10/how-much-do-google-and-facebook-profit-from-your-data.

[5] *Net Benefits: How to Quantify the Gains that the Internet Has Brought to Consumers*, ECONOMIST (Mar. 9, 2013), http://www.economist.com/news/finance-and-economics/21573091-how-quantify-gains-internet-has-brought-consumers-net-benefits.

[6] Matt Sledge, *Alex Kozinksi, Federal Judge, Would Pay a Maximum of $2,400 a Year for Privacy*, HUFFINGTON POST (Mar. 4, 2013, 5:51 PM EST),http://www.huffingtonpost.com/2013/03/04/alex-kozinski-privacy_n_2807608.html.

[7] Federico Zannier, *A Bite of Me*, KICKSTARTER, http://www.kickstarter.com/projects/1461902402/a-bit-e-of-me (last visited Aug. 29, 2013).

[8] *See, e.g.,* Alessandro Acquisti et al., What Is Privacy Worth? 27-28 (2010) (unpublished manuscript), *available at* http://www.heinz.cmu.edu/~acquisti/papers/acquisti-ISR-worth.pdf.

[9] NICOLA JENTZSCH ET AL., EUR. NETWORK & INFO. SEC. AGENCY, STUDY ON MONETISING PRIVACY: AN ECONOMIC MODEL FOR PRICING PERSONAL INFORMATION 1 (2012), *available at* http://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/monetising-privacy/at_download/fullReport.

[10] Since their inception three decades ago, the Fair Information Practices, which include principles such as user notice and consent, data integrity, and use limitations, have become the foundation of data protection law. For a thorough discussion and a critique, see Fred H. Cate, *The Failure of the Fair Information Practice Principles*, *in* CONSUMER PROTECTION IN THE AGE OF THE "INFORMATION ECONOMY" 343 (2006).

[11] *See, e.g.*, WORLD ECON. F., UNLOCKING THE VALUE OF PERSONAL DATA: FROM COLLECTION TO USAGE 4 (2013), *available at*http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf. In the lead-up to the National Telecommunications and Information Administration's privacy multi-stakeholder process, the Telecommunications Industry Association demanded that the group's "focus should be on regulating how personal information is used, rather than how it is collected." Press Release, Telecomms. Indus. Ass'n, Telecommunications Industry Association Says NTIA Privacy Code Should Focus on Data Use, Not Collection Method (July 12, 2012),http://www.tiaonline.org/news-media/press-releases/telecommunications-industry-association-says-ntia-privacy-code-should.

[12] Michael Fertik, *Big Data, Privacy, and the Huge Opportunity in the Monetization of Trust*, WORLD ECON. F. BLOG (Jan. 25, 2012, 2:13 AM),http://forumblog.org/2012/01/davos-daily-big-data-privacy-and-the-huge-opportunity-in-the-monetization-of-trust.

[13] VRM stands for "Vendor Relationship Management." According to the Harvard Berkman Center, the goal of the project is to "provide customers with both *independence* from vendors and *better ways of engaging* with vendors." *ProjectVRM*, HARV. UNIV. BERKMAN CTR. FOR INTERNET & SOC'Y,http://cyber.law.harvard.edu/projectvrm/Main_Page (last updated Mar. 27, 2013, 07:07 PM). It hopes Project VRM can improve individuals' relationships with not just businesses, but schools, churches, and government agencies. *Id*.

[14] Chris Taylor & Ron Webb, *A Penny for Your Privacy?*, HBR BLOG NETWORK (Oct. 11, 2012, 11:00 AM),http://blogs.hbr.org/cs/2012/10/a_penny_for_your_privacy.html.

[15] For a discussion of the "winners and losers" of big data, see Lior Jacob Strahilevitz, *Toward a Positive Theory of Privacy Law*, 126 HARV. L. REV. 2010, 2021-33 (2013); Omer Tene, *Privacy: For the Rich or for the Poor?*, CONCURRING OPINIONS (July 26, 2012, 2:05 AM),http://www.concurringopinions.com/archives/2012/07/privacy-for-the-rich-or-for-the-poor.html (discussing the argument that the pervasive collection of personal information allows companies "to make the poor subsidize luxury goods for the rich").

[16] Christopher Slobogin, *The Poverty Exception to the Fourth Amendment*, 55 FLA. L. REV. 391, 392, 406 (2003).

[17] *See* Tene, *supra* note 15.

[18] *See* Lori Andrews, *Facebook Is Using You*, N.Y. TIMES (Feb. 4, 2012), http://www.nytimes.com/2012/02/05/opinion/sunday/facebook-is-using-you.html. Tech analyst Alistair Croll discusses this example, arguing that big data will become a difficult civil rights issue. Alistair Croll, *Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It: What the Data Is Must Be Linked to How It Can Be Used,* O'REILLY RADAR (Aug. 2, 2012),http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html.

[19] HELEN NISSENBAUM, PRIVACY IN CONTEXT 211 (2010).

[20] Fertik, *supra* note 12.

[21] Alexis C. Madrigal, *How Much Is Your Data Worth? Mmm, Somewhere Between Half a Cent and $1,200*, ATLANTIC (Mar. 19, 2012, 3:18 PM ET),http://www.theatlantic.com/technology/archive/2012/03/how-much-is-your-data-worth-mmm-somewhere-between-half-a-cent-and-1-200/254730.

[22] Emily Steel et al., *How Much Is Your Personal Data Worth?*, FIN. TIMES (June 12, 2013, 8:11 PM), http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html.

PREDICTION, PREEMPTION, PRESUMPTION:
HOW BIG DATA THREATENS BIG PICTURE PRIVACY

*Ian Kerr & Jessica Earle**

66 STAN. L. REV. ONLINE 65

Big data's big utopia was personified towards the end of 2012.

In perhaps the most underplayed tech moment in the first dozen years of the new millennium, Google brought The Singularity nearer,[1] hiring Ray Kurzweil not as its chief futurist but as its director of engineering. The man the *Wall Street Journal* dubbed a restless genius announced his new post rather quietly in mid-December, without so much as an official press release from Google.[2] This is remarkable when one considers exactly what Google hired him to do. Kurzweil and his team will try to create a mind—an artificial intellect capable of predicting on a "semantically deep level what you are interested in."[3] With easy access to the search giant's enormous user base and the potential to scour all Google-mediated content, Kurzweil (and apparently Google) aims to turn the very meaning of "search" on its head: instead of people using search engines to better understand information, search engines will use big data to better understand people. As Kurzweil has characterized it, intelligent search will provide information to users before they even know they desire it. This accords precisely with Larry Page's longstanding vision: intelligent search "understands exactly what you mean and gives you back exactly what you want."[4]

Kurzweil's new project reifies society's increasing optimism in harnessing the utility of big data's predictive algorithms—the formulaic use of zetabytes of data to anticipate everything from consumer preferences and customer creditworthiness to fraud detection, health risks, and crime prevention. Through the predictive power of these algorithms, big data promises opportunities like never before to anticipate future needs and concerns, plan strategically, avoid loss, and manage risk. Big data's predictive tool kit clearly offers many important social benefits.[5] At the same time, its underlying ideology also threatens fundamental legal tenets such as privacy and due process.

Contrary to the received view, our central concern about big data is *not* about the data. It is about big data's power to enable a dangerous new philosophy of preemption. In this Essay, we focus on the social impact of what we call "preemptive predictions." Our concern is that big data's promise of increased efficiency, reliability, utility, profit, and pleasure might be seen as the justification for a fundamental jurisprudential shift from our current ex post facto system of penalties and punishments to ex ante preventative measures that are increasingly being adopted across various sectors of society. It is our contention that big data's predictive benefits belie an important insight historically represented in the presumption of innocence and associated privacy and due process values—namely, that there is wisdom in setting boundaries around the kinds of assumptions that can and cannot be made about people.[6]

I. PREDICTION

Since much of the big data utopia is premised on prediction, it is important to understand the

* Ian Kerr is Canada Research Chair in Ethics, Law and Technology, Faculty of Law, University of Ottawa. Jessica Earle is JD/MA Candidate, Faculty of Law, University of Ottawa and Norman Paterson School of International Affairs, Carleton University.

different purposes that big data predictions serve. This Part offers a quick typology.

The nature of all prediction is anticipatory. To predict is to "state or estimate . . . that an action or event will happen in the future or will be a consequence of something."[7] For example, when a lawyer predicts "what the courts will do in fact,"[8] she anticipates the legal consequences of future courses of conduct in order to advise clients whether it is feasible to avoid the risk of state sanction. We call predictions that attempt to anticipate the likely consequences of a person's action *consequential predictions.* As doctors, lawyers, accountants, and other professional advisors are well aware, the ability to make reliable consequential predictions can be profitable—especially in a society increasingly preoccupied with risk. The recent development of anticipatory algorithms within these fields is generally client centered.[9] The aim of these prediction services is to allow individuals to eschew risk by choosing future courses of action that best align with their own self-interest, forestalling unfavorable outcomes.

Of course, not all of big data's predictions are quite so lofty. When you permit iTunes Genius to anticipate which songs you will like or Amazon's recommendation system to predict what books you will find interesting, these systems are not generating predictions about your conduct or its likely consequences. Rather, they are trying to stroke your preferences in order to sell goods and services. Many of today's big data industries are focused on projections of this material sort, which we refer to as *preferential predictions.* Google's bid to create personalized search engines is a prime example of society's increasing reliance on preferential predictions. The company's current interface already uses anticipatory algorithms to predict what information users want based on a combination of data like website popularity, location, and prior search history.

There is a third form of prediction exemplified by a number of emerging players in big data markets. Unlike consequential and preferential predictions, *preemptive predictions* are intentionally used to diminish a person's range of future options. Preemptive predictions assess the likely consequences of allowing or disallowing a person to act in a certain way. In contrast to consequential or preferential predictions, preemptive predictions do not usually adopt the perspective of the actor. Preemptive predictions are mostly made from the standpoint of the state, a corporation, or anyone who wishes to prevent or forestall certain types of action. Preemptive predictions are not concerned with an individual's actions but with whether an individual or group should be permitted to act in a certain way. Examples of this technique include a no-fly list used to preclude possible terrorist activity on an airplane, or analytics software used to determine how much supervision parolees should have based on predictions of future behavior.[10] The private sector is also embracing this approach. For example, companies are increasingly combing through big data to find their job candidates, rather than looking to the traditional format of resumes and interviews.[11]

These three types of prediction—consequential, preferential, and preemptive—are not meant to provide an exhaustive list of all possible predictive purposes. But, as the following section reveals, understanding the different predictive purposes will help locate the potential threats of big data. To date, much of the academic focus on big data and privacy investigates what we have called consequential and preferential predictions in the context of data protection frameworks.[12] In this Essay, we focus on the less understood category of preemptive prediction and its potential impact on privacy and due process values.

II. PREEMPTION

The power of big data's preemptive predictions and its potential for harm must be carefully understood alongside the concept of risk. When sociologist Ulrich Beck coined the term *risk society* in the 1990s, he was not suggesting that society is riskier or more dangerous nowadays than before; rather, he argued that society is reorganizing itself in response to risk. Beck believes that in modern society, "the social production of wealth is systematically accompanied by the social production of risks," and that, accordingly, "the problems and conflicts relating to distribution in a society of scarcity overlap with the problems and conflicts

that arise from the production, definition, and distribution of techno-scientifically produced risks."[13]

On Beck's account, prediction and risk are interrelated concepts. He subsequently describes risk as "the modern approach to foresee and control the future consequences of human action . . . ."[14] This helps to demonstrate the link between prediction and preemption. Prediction industries flourish in a society where anyone and anything can be perceived as a potential threat, because it is lucrative to exploit risk that can later be avoided. In such cases, prediction often precipitates the attempt to preempt risk.

With this insight, an important concern arises. Big data's escalating interest in and successful use of preemptive predictions as a means of avoiding risk becomes a catalyst for various new forms of social preemption. More and more, governments, corporations, and individuals will use big data to preempt or forestall activities perceived to generate social risk. Often, this will be done with little or no transparency or accountability. Some loan companies, for example, are beginning to use algorithms to determine interest rates for clients with little to no credit history, and to decide who is at high risk for default. Thousands of indicators are analyzed, ranging from the presence of financially secure friends on Facebook to time spent on websites and apps installed on various data devices. Governments, in the meantime, are using this technique in a variety of fields in order to determine the distribution of scarce resources such as social workers for at-risk youth or entitlement to Medicaid, food stamps, and welfare compensation.[15]

Of course, the preemption strategy comes at a significant social cost. As an illustration, consider the practice of using predictive algorithms to generate no-fly lists. Before the development of many such lists in various countries, high-risk individuals were generally at liberty to travel—unless the government had a sufficient reason to believe that such individuals were in the process of committing an offense. In addition to curtailing liberty, a no-fly list that employs predictive algorithms preempts the need for any evidence or constitutional safeguards. Prediction simply replaces the need for proof.

Taken to its logical extreme, the preemption philosophy is not merely proactive—it is aggressive. As President George W. Bush famously argued:

> If we wait for threats to fully materialize, we will have waited too long. . . . We must take the battle to the enemy, disrupt his plans, and confront the worst threats before they emerge. . . . [O]ur security will require all Americans to be forward-looking and resolute, to be ready for preemptive action when necessary . . . .[16]

Proponents of this approach argue there is a "duty to prevent," which means the responsible choice requires use of predictive tools to mitigate future risk.[17] But with this, we see that a universalized preemption strategy could challenge some of our most fundamental jurisprudential commitments, including the presumption of innocence. In the following Part, we seek to demonstrate that even more mundane forms of preemption generated by big data can also threaten privacy and due process values.

III. PRESUMPTION

To date, much of the best work on the implications of big data tends to treat the privacy worry as though it were somehow contained within the minutiae of the data itself. As Tene and Polonetsky have meticulously argued: "Information regarding individuals' health, location, electricity use, and online activity is exposed to scrutiny, raising concerns about profiling, discrimination, exclusion, and loss of control."[18] Through the fine-tuned microscope of data privacy frameworks, the central issues tend to be the definition of personally identifiable information, the prospect of de-identifying the data, the nature of consent to the collection, use, or disclosure of the data, and a range of other data privacy rules such as purpose limitation and data minimization.

Our approach examines the privacy issue with a telescope rather than a microscope.

If the legal universe has a prime directive, it is probably the shared understanding that everyone is presumed innocent until proven

guilty. In legal discourse, the presumption of innocence is usually construed, narrowly, as a procedural safeguard enshrined within a bundle of "due process" rights in criminal and constitutional law. These include the right to a fair and impartial hearing, an ability to question those seeking to make a case against you; access to legal counsel, a public record of the proceedings, published reasons for the decision, and, in some cases, an ability to appeal the decision or seek judicial review.[19] Likewise, a corollary set of duties exists in the private sector. Although such duties are not constitutionally enshrined, companies do owe employees and customers the right to full information, the right to be heard, the right to ask questions and receive answers, and the right of redress.[20] Gazing at the bigger picture, the presumption of innocence and related private sector due process values can be seen as wider moral claims that overlap and interrelate with core privacy values.

Taken together, privacy and due process values seek to limit what the government (and, to some extent, the private sector) is permitted to presume about individuals absent evidence that is tested in the individuals' presence, with their participation. As such, these values aim to provide fair and equal treatment to all by setting boundaries around the kinds of assumptions that can and cannot be made about people. This is wholly consistent with privacy's general commitment to regulating what other people, governments, and organizations are permitted to know about us. Among other things, the aim is to prevent certain forms of unwarranted social exclusion.[21]

With all of this, we are finally able to locate the threat that big data poses. Big data enables a universalizable strategy of preemptive social decisionmaking. Such a strategy renders individuals unable to observe, understand, participate in, or respond to information gathered or assumptions made about them. When one considers that big data can be used to make important decisions that implicate us without our even knowing it, preemptive social decision making is antithetical to privacy and due process values.

CONCLUSION

The nexus between big data and privacy is not a simple story about how to tweak existing data protection regimes in order to "make ends meet"; big data raises a number of foundational issues. Since predictability is itself an essential element of any just decisionmaking process, our contention is that it must be possible for the subjects of preemptive predictions to scrutinize and contest projections and other categorical assumptions at play within the decisionmaking processes themselves. This is part of our broader assertion that privacy and due process values require setting boundaries around the kinds of institutional assumptions that can and cannot be made about people, particularly when important life chances and opportunities hang in the balance.

We believe that such considerations will become increasingly significant in both public and private sector settings, especially in light of the kinds of big data prediction machines that Ray Kurzweil and others want to build "to . . . Google scale."[22] These projects must be kept in mind given our emerging understanding that "some uses of probability and statistics serve to reproduce and reinforce disparities in the quality of life that different sorts of people can hope to enjoy."[23]

While it is exciting to think about the power of big data and the utopic allure of powerful prediction machines that understand exactly what we mean and tell us exactly what we want to know about ourselves and others, we believe that privacy values merit the further study and development of potential limitations on how big data is used. We need to ensure that the convenience of useful prediction does not come at too high a cost.

---

[1] Or, not. *See* John Pavlus, *By Hiring Kurtzweil, Google Just Killed the Singularity*, MIT TECH. REV. (DEC. 17, 2012),http://www.technologyreview.com/view/508901/by-hiring-kurzweil-google-just-killed-the-singularity.

[2] William M. Bulkeley, *Kurzweil Applied Intelligence Inc.*, WALL ST. J., June 23, 1989, at A3. In an email sent on July 7, 2013, Google confirmed that the hire was not announced by the search giant, but was posted on Kurzweil's website, http://www.kurzweilai.net/kurzweil-joins-google-to-work-on-new-projects-involving-machine-learning-and-language-processing. E-mail from Jason Freidenfelds, Google Communications Representative, to Jessica Earle

(July 7, 2013, 17:52 UTC) (on file with Stanford Law Review).

[3] Interview by Keith Kleiner with Ray Kurzweil, Director of Engineering, Google, in Moffett Field, Cal. (Jan. 4, 2013), *available at*http://www.youtube.com/watch?v=YABUffpQY9w.

[4] *Our Products and Services*, GOOGLE, http://www.google.com/corporate/tech.html (last visited Aug. 29, 2013) (internal quotation marks omitted).

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

[6] Our argument in this brief Essay is an adaptation of an earlier book chapter, Ian Kerr, *Prediction, Pre-emption, Presumption: The Path of Law After the Computational Turn*, *in* PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN 91 (Mireille Hildebrandt & Katja de Vries eds., 2013).

[7] *See Predict Definition*, OXFORD ENGLISH DICTIONARY, http://www.oed.com/view/Entry/149856 (last visited Aug. 29, 2013).

[8] Oliver W. Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 461 (1897).

[9] *See IBM Watson*, IBM,http://www-03.ibm.com/innovation/us/watson (last visited Aug. 29, 2013); *see also AI Am the Law*, ECONOMIST (Mar. 10, 2005),http://www.economist.com/search/PrinterFriendly.cfm?story_id=3714082.

[10] Soumya Panda, *The Procedural Due Process Requirements for No-Fly Lists*, 4 PIERCE L. REV. 121 (2005); Steve Watson, *Pre-Crime Technology to Be Used in Washington D.C.*, PRISON PLANET (Aug. 24, 2010), http://www.prisonplanet.com/pre-crime-technology-to-be-used-in-washington-d-c.html.

[11] *E.g.*, Max Nisen, *Moneyball at Work: They've Discovered What Really Makes a Great Employee*, BUS. INSIDER (May 6, 2013, 1:00 PM),http://www.businessinsider.com/big-data-in-the-workplace-2013-5.

[12] *E.g.*, Asim Ansari et al., *Internet Recommendation Systems*, 37 J. MARKETING RES. 363 (2000); Tam Harbert, *Big Data Meets Big Law: Will Algorithms Be Able to Predict Trial Outcomes?*, LAW TECH. NEWS (Dec. 27, 2012), http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202555605051 ; Ernan Roman, *BIG Data Must Create BIG Experiences*, DIRECT MKTG. NEWS (Mar. 18, 2013), http://www.dmnews.com/big-data-must-create-big-experiences/article/284831; Daniel Martin Katz, Remarks at Michigan State University College of Law's lawTechCamp: Quantitative Legal Prediction (Or How I Learned to Stop Worrying and Embrace Disruptive Technology) (June 8, 2013), *available at* http://lawtechcamp.com/qualitative-legal-prediction.

[13] ULRICH BECK, RISK SOCIETY: TOWARDS A NEW MODERNITY 19 (1992).

[14] ULRICH BECK, WORLD RISK SOCIETY 3 (1999).

[15] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1256 (2008); Stephen Goldsmith, *Big Data, Analytics and a New Era of Efficiency in Government*, GOVERNING THE STATES AND LOCALITIES (May 22, 2013), http://www.governing.com/blogs/bfc/col-big-data-analytics-government-efficiency.html ; Evgeny Morozov, *Your Social Networking Credit Score*, SLATE (Jan. 30, 2013, 8:30 AM),http://www.slate.com/articles/technology/future_tense/2013/01/wonga_lenddo_lendup_big_data_and_social_networking_banking.html.

[16] President George W. Bush, Graduation Speech at West Point (June 1, 2002, 9:13 AM), *available at* http://georgewbush-whitehouse.archives.gov/news/releases/2002/06/20020601-3.html.

[17] Lee Feinstein & Anne-Marie Slaughter, *Duty to Prevent*, FOREIGN AFF., Jan.-Feb. 2004, at 136.

[18] Tene & Polonetsky, *supra* note 5, at 65.

[19] Henry J. Friendly, *"Some Kind of Hearing"*, 123 U. PA. L. REV. 1267 (1975).

[20] Kerr, *supra* note 6, at 108. *See generally* Lauren B. Edelman, *Legal Environments and Organizational Governance: The Expansion of Due Process in the American Workplace*, 95 AM. J. SOC., 1401, 1405-08 (1990).

[21] OSCAR H. GANDY JR., THE PANOPTIC SORT: A POLITICAL ECONOMY OF PERSONAL INFORMATION (1993); Richard V. Ericson, *The Decline of Innocence*, 28 U.B.C. L. REV. 367 (1994).

[22] Interview with Ray Kurzweil by Keith Kleiner, *supra* note 3.

[23] OSCAR H. GANDY, JR., COMING TO TERMS WITH CHANCE: ENGAGING RATIONAL DISCRIMINATION AND CUMULATIVE DISADVANTAGE 1 (2009).

PUBLIC VS. NONPUBLIC DATA:

THE BENEFITS OF ADMINISTRATIVE CONTROLS

*Yianni Lagos & Jules Polonetsky**

66 STAN. L. REV. ONLINE 103

This Essay attempts to frame the conversation around de-identification. De-identification is a process used to prevent a person's identity from being connected with information. Organizations de-identify data for a range of reasons. Companies may have promised "anonymity" to individuals before collecting their personal information, data protection laws may restrict the sharing of personal data, and, perhaps most importantly, companies de-identify data to mitigate privacy threats from improper internal access or from an external data breach. Hackers and dishonest employees occasionally uncover and publicly disclose the confidential information of individuals. Such disclosures could prove disastrous, as public dissemination of stigmatizing or embarrassing information, such as a medical condition, could negatively affect an individual's employment, family life, and general reputation. Given these negative consequences, industries and regulators often rely on de-identification to reduce the occurrence and harm of data breaches.

Regulators have justifiably concluded that strong de-identification techniques are needed to protect privacy before publicly releasing sensitive information. With publicly released datasets, experts agree that weak technical de-identification creates an unacceptably high risk to privacy.[1] For example, statisticians have re-identified some individuals in publicly released datasets.

None of these publicized attacks, however, have occurred using nonpublic databases. Experts also agree that organizations reduce privacy risk by restricting access to a de-identified dataset to only trusted parties.[2] This Essay builds on this consensus to conclude that de-identification standards should vary depending on whether the dataset is released publicly or kept confidential.

This Essay first describes only technical de-identification (DeID-T) and how policymakers have recognized the benefits of de-identifying data before publicly disclosing a dataset. Second, this Essay discusses how administrative safeguards provide an additional layer of protection to DeID-T that reduces the risk of a data breach. Third, this Essay analyzes the use of de-identification in conjunction with administrative safeguards (DeID-AT). DeID-AT minimizes privacy risks to individuals when compared to using DeID-T or administrative safeguards in isolation. Fourth, this Essay discusses how the different privacy risk profiles between DeID-AT and DeID-T may justify using a reasonably good de-identification standard—as opposed to extremely strict de-identification measures—for non-publicly disclosed databases.

I. TECHNICAL DE-IDENTIFICATION (DEID-T)

DeID-T is a process through which organizations remove or obscure links between an individual's identity and the individual's personal information. This process involves deleting or masking personal identifiers, such as names and social security numbers, and suppressing or generalizing quasi-identifiers, such as dates of birth and zip codes. By using technical de-identification, organizations can transform sensitive information from being fully individually identifiable to being unconnected to any particular person. With publicly disclosed

* Yianni Lagos is Legal and Policy Fellow, Future of Privacy Forum. Jules Polonetsky is Director and Co-Chair, Future of Privacy Forum.

datasets, DeID-T provides the sole line of defense protecting individual privacy.

Policymakers have recognized the benefits of DeID-T by providing regulatory inducements to companies that de-identify publicly disclosed databases. For example, if a company adequately anonymizes a dataset under the 1995 E.U. Data Protection Directive (E.U. Directive), de-identification allows for public disclosure of data without violating individual privacy.[3] Following the E.U. Directive and the U.K. Data Protection Act, the United Kingdom's Information Commissioner's Office (ICO) expressed support for de-identification: "[T]he effective anonymization of personal data is possible, desirable and can help society to make rich data resources available whilst protecting individuals' privacy."[4] The U.S. Department of Health and Human Services (HHS) similarly recognized the benefits of de-identifying health data: "The process of de-identification, by which identifiers are removed from health information, mitigates privacy risks to individuals and thereby supports the secondary use of data . . . ."[5]

There are, however, limits to the protections provided by DeID-T. Two different threat models create a risk of re-identification—i.e., reconnecting an individual with what is usually called "personal data" in the European Union and "personally identifiable information" (PII) in the United States.[6] First, outsiders can potentially re-identify an individual by comparing quasi-identifiers in a de-identified database with an identified database, such as a voter registration list. Outsider attacks can come from bad actors or academics, attempting to exploit or show weaknesses in DeID-T protections. In fact, the highest profile re-identification attacks have come from academics attempting to re-identify individuals in publicly disclosed databases.[7] Second, insiders can potentially re-identify an individual by using knowledge that is not generally known. For instance, a Facebook friend, acquaintance, or "skillful Googler" might exploit information that only a limited set of people know, such as a Facebook post mentioning a hospital visit.[8] Similarly, an employee might be able to search through other information held by the organization to re-identify a person.

The threats posed by outsiders, and insiders with restricted access to information, vary significantly depending on whether the de-identified data is publicly disclosed or kept confidential within an organization. When organizations publicly disclose a dataset, every academic, bad actor, and friend can attempt to re-identify the data with DeID-T providing the sole protection. When organizations keep datasets confidential, in contrast, the risk of potential attackers having access to the de-identified data is minimized due to the additional defense of administrative safeguards.

II. Administrative Safeguards

This Essay uses the term administrative safeguards to mean all non-technical data protection tools that help prevent confidential data from becoming publicly released or improperly used. In the E.U. Directive, these safeguards are referred to as organizational measures. Non-technical protections include two broad categories: 1) internal administrative and physical controls (internal controls) and 2) external contractual and legal protections (external controls).[9] Internal controls encompass security policies, access limits, employee training, data segregation guidelines, and data deletion practices that aim to stop confidential information from being exploited or leaked to the public. External controls involve contractual terms that restrict how partners use and share information, and the corresponding remedies and auditing rights to ensure compliance.

By implementing administrative safeguards, organizations provide important privacy protections independent of DeID-T. A dentist's office, for instance, does not routinely de-identify patient records to protect a person's privacy, which could negatively impact patient care. Instead, privacy law recognizes that a dental office can hold fully identifiable information if it uses appropriate administrative safeguards, such as performing pre-hire background checks on employees, physically locking drawers with patient records, limiting the information on forms to only needed data, and training employees regarding appropriate access, handling, and disposal of patient files. No privacy breach occurs as long as the

confidential patient records do not become disclosed.

The use of administrative safeguards as an additional data protection tool along with DeID-T is consistent with both E.U. and U.S. privacy law. Article 17 of the E.U. Directive requires organizations to "implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access . . . ."[10] The General Data Protection Regulation extends the Directive's existing support for using both technical and organizational measures by incorporating those safeguards into a variety of data protection processes, and by granting the European Commission the power to specify "the criteria and conditions for the technical and organizational measures."[11]

U.S. law similarly requires the use of administrative and technical safeguards. The U.S. Privacy Act of 1974 requires federal agencies to "establish appropriate administrative, technical and physical safeguards to insure the security and confidentiality of records and to protect against any anticipated threats or hazards to their security or integrity."[12] The Gramm-Leach-Bliley Act mandates that financial agencies establish "administrative, technical, and physical safeguards" for financial institutions.[13] Policymakers have thus given value to administrative (or organizational) safeguards as a privacy tool separate from DeID-T that organizations can use to enhance data protection. Similar appreciation for administrative safeguards may be appropriate when applied in the de-identification sphere.

III. Administrative and Technical De-identification

Organizations who use DeID-AT build a two-tiered barrier that significantly enhances individual privacy protections compared with a single layer. One layer, administrative safeguards, reduces the likelihood of personal data being accessed without authorization. If an insider or outsider does get unauthorized access, another layer, technical de-identification, acts as an additional fortification to minimize potential privacy harms. The two-layered defense provided by DeID-AT means that potential bad actors must not only circumvent administrative measures to gain access to data, but also must re-identify that data before getting any value from their malfeasance. Both are low probability events that together greatly reduce privacy risks. Hence, organizations that implement DeID-AT improve individual privacy.

Policymakers have recognized the distinction between DeID-AT and DeID-T. The ICO drew a line of demarcation between public and nonpublic databases: "We also draw a distinction between publication to the world at large and the disclosure on a more limited basis—for example to a particular research establishment with conditions attached."[14] The Canadian De-Identification Working Group also voiced its belief: "Mitigating controls work in conjunction with de-ID techniques to minimize the re-ID risk."[15] These statements appear to support the proposition that DeID-AT provides a different level of privacy protection than when DeID-T is the sole defensive tool used in publicly disclosed databases.

The heightened privacy protection provided by adding de-identification to administrative safeguards is best demonstrated by using simple statistics. Suppose, for example, the probability of a technical attack on a database gives a one percent chance of re-identification. Suppose as well that the probability of a breach of administrative safeguards is also one percent. (In practice, the likelihood of each is generally much lower.) With both technical and administrative protections, the probability of re-identifying data is thus one percent of one percent, or one in 10,000.[16] This simple statistical example shows that the risk of re-identification with DeID-AT may well be orders of magnitude lower than using only technical safeguards in isolation.

IV. Policy Implications

The additional protections provided by DeID-AT compared with DeID-T suggest a different risk profile that may justify the use of fairly strong technical measures, combined with effective administrative safeguards. The Federal Trade Commission recognized this fact when it called in its 2012 report for technical measures that

made a dataset "not reasonably identifiable."[17] The combination of reasonably good technical measures, as well as good administrative measures, likely leads to a lower risk of re-identification than stronger technical measures acting alone. The HIPAA de-identification standard that requires a "very small" risk of re-identification before publicly releasing health data is an example of a relatively strict standard for re-identification, designed for datasets that can be made fully public.[18] A less strict standard, however, achieves a similar or stronger level of protection for non-publicly available databases.

Giving credit to the use of administrative controls also helps prevent an illogical outcome: greater data restrictions for the original collector of the data than downstream recipients or the public. The original collector commonly has more access to data on an individual than it would disclose to another party. A separate nonpublic database containing an individual's name or email address, for example, would normally not be disclosed. That separate database could potentially be used to re-identify an individual, giving the original collector a re-identification advantage over any other party.[19] Thus, if administrative controls do not receive regulatory recognition, the original data collector would be subject to a steeper regulatory burden than potential downstream recipients.

Relying on the data protection benefits of using DeID-AT to justify allowing reasonably strict de-identification comes with a caveat that it can be difficult to assess the efficacy of administrative safeguards. Privacy advocates and academics can test DeID-T used in public data releases. In fact, improvements in DeID-T can result from privacy advocates and academics testing claims of anonymization. Companies, however, keep administrative safeguards proprietary for security purposes, and privacy advocates cannot audit non-transparent privacy protections. The use of third-party auditors is one approach for ensuring that administrative safeguards effectively prevent privacy attacks, but without a certain level of public transparency of such measures, regulators and privacy advocates may find it difficult to assess the exact benefits of administrative safeguards.

CONCLUSION

Non-publicly disclosed datasets have a lessened risk of re-identification than publicly disclosed datasets due to the added protection of administrative controls. The different risk profiles suggest requiring different measures of de-identification for publicly disclosed datasets compared with confidential datasets. This Essay urges regulators to recognize the heightened individual privacy protections provided by DeID-AT compared with DeID-T when developing privacy regulations.

---

[1] *See* Daniel C. Barth-Jones, The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risk and Privacy Protections, Then and Now 5 (July 24, 2012) (unpublished working paper), *available at*https://www.privacyassociation.org/media/pdf/knowledge_center/Re-Identification_of_Welds_Medical_Information.pdf (finding that 29% of individuals examined had a plausible risk of re-identification with full data of birth, gender, and five-digit ZIP code, though actual risk was much lower given incomplete data).

[2] *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1771 (2010).

[3] Council Directive 95/46, 1995 O.J. (L 281) 26 (EC).

[4] INFO. COMM'R'S OFFICE, ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE 7 (2012).

[5] OFFICE OF CIVIL RIGHTS, U.S. DEP'T OF HEALTH AND HUMAN SERVS., GUIDANCE ON DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION 5 (2012), *available at* http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf. HHS is referring to the de-identification provisions found in the Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified at scattered sections of the U.S. Code): "*Standard: de-identification of protected health information.* Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information." 45 C.F.R. § 164.514 (2012).

[6] Felix T. Wu, *Privacy and Utility in the Data Set*, 85 U. COLO. L. REV. 1117 (2013).

[7] C. Christine Porter, *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information*, SHIDLER J.L. COM. & TECH., Sept. 23, 2008, at 1, *available at* http://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/417/vol5_no1_art3.pdf (referri

ng to AOL and Netflix as examples of re-identification attacks).

[8] Wu, *supra* note 6, at 28 (quoting Nw. Mem'l Hosp. v. Ashcroft, 362 F.3d 923, 929 (7th Cir. 2004)).

[9] This Essay combines the administrative and physical safeguards referred to in the Privacy Act of 1974 into one category: administrative safeguards. Privacy Act of 1974, 5 U.S.C. § 552a(e)(10) (2011).

[10] Council Directive 95/46, *supra* note 3, at art. 17(1).

[11] *Commission 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)*, at 56-60, COM (2012) 11 final (Jan. 1, 2012).

[12] 5 U.S.C. § 552a(e)(10).

[13] 15 U.S.C. § 6801(b).

[14] INFO. COMM'R'S OFFICE, *supra* note 4, at 7.

[15] HEALTH SYS. USE TECHNICAL ADVISORY COMM. DATA DE-IDENTIFICATION WORKING GRP., 'BEST PRACTICE' GUIDELINES FOR MANAGING THE DISCLOSURE OF DE-IDENTIFIED HEALTH INFORMATION 19 (2010).

[16] The one in 10,000 statistic is based on the assumption that the probability of the technical and administrative attacks are independent of each other. In practice, under a particular attack scenario, this assumption may not hold. By arguing for a different de-identification standard for public and non-public data, we do not claim that pseudonymization is sufficient to constitute pretty good de-identification. Other factors, such as whether companies maintain the cryptographic key when transforming identifiers, will determine the effectiveness of pseudonymization. It is clear, however, that if a company can easily re-identify every individual from a pseudonymous database, the statistical benefits of combining administrative measures with technical measures are lost.

[17] FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE 22 (2012).

[18] OFFICE OF CIVIL RIGHTS, *supra* note 5, at 6.

[19] KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION 142 (2013).

BIG DATA ANALYTICS:
EVOLVING BUSINESS MODELS AND GLOBAL PRIVACY REGULATION

*Peter Leonard**

At the heart of the current global debate as to how privacy regulation should address big data lie three questions:

- Can national privacy laws and regulation facilitate socially beneficial uses and applications of big data while also precluding 'Big Brother', 'spooky, 'creepy' or otherwise socially or culturally unacceptable big data practices?

- Can diverse national privacy laws and regulation, including markedly different constructs as to what is personally identifying information and sensitive information, be applied or adapted so as to accommodate socially beneficial uses and applications of big data, or is a more fundamental overhaul of law and regulation required?

- If fundamental design precepts of privacy regulation require adaptation or supplementation to address big data, can those changes be made without threatening broader consistency and integrity of privacy protections for individuals? Can any adaptation or changes be made quickly enough to address growing citizen concerns about unacceptable or hidden big data practices?

From the summer of 2012 media and policy attention in the United States as to privacy and big data focussed on data analytics conducted by offline ('bricks and mortar') businesses in relation to their customers and on the nature and range of analytics services offered by third party providers collectively labelled 'data brokers'. Media reportage reinforced unease and a perception of many people that business data analytics principally involves hidden and deliberatively secretive identification and targeting of individual consumers for tailoring of 'one to one' marketing material directed to them, including targeting by marketers with whom the individual has no prior customer relationship. The fact that this has been a U.S. led debate is of itself is not surprising, for at least two reasons. First, in contrast to the European Union and other advanced privacy regulating jurisdictions such as Canada, Australia and Hong Kong, the U.S.A. has not had economy wide collection and notification requirements in relation to PII or as to notification to the data subject as to collection and processing of PII collected about that data subject other than directly from the data subject. Second, the U.S. Do Not Track debate has focussed consumer attention upon online behavioural advertising and probably reinforced perceptions that the dominant reason for offline retailers implementing big data projects is for 'one to one' targeting and marketing.

* Peter G. Leonard B.Ec. (Hons) LLM (Syd.) is a TMT and Projects Partner of Gilbert + Tobin Lawyers, based in the firm's Sydney, Australia office. He is a director of the International Association of Privacy Professionals Australia & New Zealand (iappANZ) and National Chair of the Media and Communications Committee of the Business Law Section of the Law Council of Australia. He is a former Global Chair of the Technology Committee of the Business Law Section of the International Bar Association and a former director of the Internet Industry Association of Australia.

The author acknowledges valuable assistance from comments on an earlier draft of this paper by many friends and privacy colleagues, including Malcolm Crompton, Jennifer Barrett-Glasgow, Carolyn Lidgerwood and Teresa Troester-Falk, and the ever cheerful and questioning assistance of Natalie Lane. The author also acknowledges the influence of Greg Schneider and Mike Briers, who introduced the author to practicalities of business customer analytics. The paper is the better for their assistance, but may not reflect their views.

The European big data debate since early 2012 has been quite differently focussed. The debate has included discussion of the long standing, particularly European concern as to decisions made by automated data processing without significant human judgement – so called 'automated individual decisions', or 'profiling'. The European profiling debate has a philosophical core: is the personal dignity and integrity of individuals compromised by decisions made by automated processes, as contrasted to individual decision making by humans constrained both by laws against discrimination and also, perhaps, human empathy? The profiling debate in the United Kingdom has also included a pragmatic, economic dimension. In response to consumer advocate concerns as to differential pricing online, the U.K. Office of Fair Trading examined possibilities for geo-location based and 'personalised pricing': that is, "the possibility that businesses may use information that is observed, volunteered, inferred, or collected about individuals' conduct or characteristics, such as information about a particular user's browsing or purchasing history or the device the user uses, to set different prices to different consumers (whether on an individual or group basis) based on what the business thinks they are willing to pay."

The commonality of concerns around overly intrusive or 'bad' big data practices has been partially obscured by regional and national differences in privacy regulation and in the detail of technical legal analysis as to the interpretation of privacy law. There is an engaged and continuing global debate as to how fundamental privacy concepts of notice and consent should be adapted to apply in a fully networked world of individuals and of interworking devices (the so called 'internet of things'). There has also been an active debate as to the continuing differences in national regulatory approaches to PII and particularly sensitive information such as health data and how these differences may affect implementation of now common transnational services such as global or regional data centres and software applications delivered as cloud services. Although the debate as to privacy regulation of big data has usefully focussed upon how the business practices of big data analytics can be appropriately risk managed through adaption of regulation and application of privacy by design principles, the discussion has often failed to give due credence to the depth of community concerns as to analytics about individuals conducted by third parties that do not have a direct business or other relationship with the individual and analytics that feel 'spooky' or 'creepy'.

In advanced privacy law jurisdictions privacy interests of individuals are often given effect through privacy regulation and legal sanctions and remedies (at least where these are available and affordable) attaching to breach of collection notices, privacy statements and customer terms. However, citizen concerns are also given practical effect through the significant reputational damage, and in particular adverse media coverage, suffered by governments and businesses that misjudge consumer sentiments and tolerance of perceived privacy invasive practices, regardless of whether those practices contravene laws. Lack of transparency as to activities that may conform to present law can create significant citizen concern, as most recently illustrated in the debates as to acceptable limits to alleged online metadata mining conducted by US intelligence agencies in the PRISM program and as to uses by journalists employed by Bloomberg News of their privileged access to information relating to Bloomberg customers use of Bloomberg Finance services and terminals. Sentiments expressed as dislike of 'creepiness' or 'spookiness' often reflect citizen concerns about lack of transparency and lack of control or accountability of businesses dealing with personal information about them. These concerns are often not expressed in terms of these basic privacy principles and often do not map to existing laws. There is a growing deficit of trust of many citizens in relation to digital participation, as demonstrated by pressure for expansion in profiling restrictions under European privacy law, for 'just in time' notices as to use of cookies, enactment of Do Not Track laws and laws restricting geo-tracking and employers access to social media. That deficit of trust threatens to spill-over to offline data applications and by so doing endanger socially beneficial applications of big data by businesses and by government. The perception of citizen unease has pushed some businesses to be less transparent about their data analytics

projects, which has reinforced the sense of a growing climate of business and government colluding in secrecy.

The counter-view is that a growing sector of the public comfortably live their digital lives reflecting the oft-quoted aphorism that 'privacy is dead' and may therefore be expected to become more accepting of privacy affecting big data analytics as time goes by. However, there is already compelling evidence that many individuals presented with privacy choice will display a more nuanced and contextual evaluation as to what personal information they particularly value or regard as sensitive, as to particular entities with whom they will entrust their personal information and as to the trades that they are willing to make for use of that information. As individuals come to understand the economic value that increasingly accrues around personal information, it is reasonable to expect that these contextual judgements will become even more nuanced and conditional. It may be that the deficit of trust in digital participation is growing and not merely a relic of inter-generational differences.

Application of today's privacy regulation to map a course through big data implementation may miss the mark of sufficiently addressing this deficit of trust. Not infrequently, business customer analytics projects stall at a point where a chief marketing officer has successfully addressed the concerns of the chief information officer, the chief privacy officer and the general counsel, but the chief executive or a consumer advocate within a corporation is then persuasive with her or his view that customers will not trust the business with the proposed implementation. Moreover, the trust deficit can be highly contextual to a particular transaction type, a particular vendor-client relationship, a distinct geography, or a particular culture. Many consumers understand that enabling geo-location on mobile devices for a particular app enables the provider of that app to target content of offers to them based upon that location. Many consumers understand that they derive a benefit from a loyalty card in a value exchange with a vendor who will use that loyalty card data for customer analytics to target offers to that consumer. A direct and proximate vendor-client relationship promotes accountability: consumers may vote with their trade if the vendor betrays the customer's expectations, whether those expectations are based on legal rights or not. A direct and proximate relationship also leads to accountability: many consumers will draw no distinction between a vendor and the vendor's sub-contractors, such as external data analytics providers, in relation to breaches of security or uses or abuses of personal information given to that vendor. By contrast, the term 'data broker' of itself conjures the sense of lack of accountability and lack of transparency, in addition to there being no value exchange between the broker and the affected individual.

Engendering trust requires more than good privacy compliance. Compliance is, of course, a necessary component of responsible business governance for using data about individuals for marketing purposes, but it is only one component. Responsible governance of data analytics affecting citizens, whether by businesses or government, requires a new dialogue to be facilitated to build community understanding as to appropriate transparency and fair ethical boundaries to uses of data. This requires both businesses and government to acknowledge that there is both good big data and bad big data and that transparency as to data analytics practices is necessary for this dialogue and community understanding.

Fundamental failings of many data analytics projects today include unnecessary use of personally identifying information in many applications where anonymised or de-identified transaction information would suffice and omission of technical, operational and contractual safeguards to ensure that risk of re-identification is appropriately risk managed. Both good privacy compliance and sound customer relations requires planning of operational processes to embed, in particular, safeguards against re-identification of anonymised information, in how an organisation conducts its business, manages its contractors, offers its products and services and engages with customers. Privacy by design and security by design is sometimes implemented through a binary characterisation of data as personal and therefore regulated, or not personally identifying and therefore unregulated. The developing

privacy theory adopts a more nuanced, graduated approach. This graduated approach puts re-identification into a continuum between certainty of complete anonymisation and manifestly identifying information and then seeks to answer four implementation questions:

- Can this graduated or 'differential' approach be made to work within diverse national current regulatory regimes and varying definitions of personal information and PII and requirements as to notice and consent, data minimisation and limits to data retention?

- How should a privacy impact assessor or a privacy regulator assess the risk mitigation value of stringent limited access and other administrative, operational and legal safeguards? Are these safeguards only relevant *in addition* to high assurance of technical de-identification?

- Is there a subset of legal obligations that should apply to users of de-identified datasets about individuals to protect against re-identification risk?

- How should citizens be informed about customer data analytics so as to ensure that notices are understandable and user friendly? How can these notices accommodate the dynamic and unpredictable manner in which business insights may be discovered and then given operation in production data analytics?

Privacy theory meets the reality of business and government big data analytics in the way that these questions will be answered in business practices. The last question must be answered sufficiently quickly to build community understanding and engagement as to 'good big data' before concerns by privacy advocates and concerned citizens as to 'bad big data' prompt regulatory over-reach. Although these questions have not been definitively answered by privacy regulators, over the last year regulators in a number of advanced privacy jurisdictions, including the United Kingdom, Australia and Singapore, have published views that usefully and constructively engage the debate.

What is striking from a comparison of these regulatory views is the conceptual similarity between the approach of these regulators in answering the question as to when personal information, or personally identifying information, as diversely defined and interpreted under national laws, should be considered sufficiently de-identified or anonymised as to make re-identification unlikely. The conceptual similarity is of itself is unusual: most areas of national privacy regulation are characterised by marked divergence in national or regional privacy theory and practical application. Each regulatory view requires assessment of the sensitivity of the data, the context and limits of its disclosure and implementation by the data analytics provider of appropriate risk mitigation measures. Once the first assessment has been completed in terms of the possibilities and limits of effective de-identification, the second step of applying additional safeguards will often need to follow. Although the standard for acceptable risk is variously stated, the regulatory views are not dissimilar - 'low', 'remote' or 'trivial'. The possibility of re-identification is contextually assessed, or as the U.K. Information Commissioner puts it, 'in the round'. Risk mitigation measures – being appropriately 'robust' safeguards – are to be implemented before purportedly anonymised data is made available to others. These risk mitigation measures may be a combination of technical, operational and contractual safeguards. The regulatory views also converge in not being prescriptive as to particular safeguards, instead offering a menu board approach for consideration in a privacy and security impact assessment individual to that deployment as to the safeguards appropriate for a particular data analytics deployment.

The menu board of safeguards is relatively long. It includes use of trusted third party arrangements; use of pseudonymisation keys and arrangements for separation and security of decryption keys; contractual limitation of the use of the data to a particular project or projects; contractual purpose limitations, for example, that the data can only be used by the recipient for an agreed purpose or set of purposes;

contractual restriction on the disclosure of the data; limiting the copying of, or the number of copies of, the data; required training of staff with access to data, especially on security and data minimisation principles; personnel background checks for those granted access to data; controls over the ability to bring other data into the environment (allowing the risk of re-identification by linkage or association to be managed); contractual prohibition on any attempt at re-identification and measures for the destruction of any accidentally re-identified personal data; arrangements for technical and organisational security, e.g. staff confidentiality agreements; and arrangements for the destruction or return of the data on completion of the project.

While these regulatory views are being developed and refined, the questions that the regulators are tentatively answering are already being addressed through business practices that, if and when done well, deploy technical de-identification and also embed privacy impact assessment, privacy by design and security by design principles into other operational (administrative, security and contractual) safeguards within data analytics service providers, governments and corporations. But because this area is new, there is no common industry practice as to such safeguards, and sub-standard implementations continue and threaten to further erode citizen trust as to big data. If bad practices and bad media further promote other businesses and government to be less transparent about their data analytics projects, public perception of business and government colluding in secrecy will grow, prompting more prescriptive regulation. Big data and the privacy regulatory and compliance response to it will be one of the most important areas for development of operational privacy compliance for the next five years.

BIG DATA AND ITS EXCLUSIONS

*Jonas Lerman**

66 STAN. L. REV. ONLINE 55

Legal debates over the "big data" revolution currently focus on the risks of inclusion: the privacy and civil liberties consequences of being swept up in big data's net. This Essay takes a different approach, focusing on the risks of exclusion: the threats big data poses to those whom it overlooks. Billions of people worldwide remain on big data's periphery. Their information is not regularly collected or analyzed, because they do not routinely engage in activities that big data is designed to capture. Consequently, their preferences and needs risk being routinely ignored when governments and private industry use big data and advanced analytics to shape public policy and the marketplace. Because big data poses a unique threat to equality, not just privacy, this Essay argues that a new "data antisubordination" doctrine may be needed.

* * *

The big data revolution has arrived. Every day, a new book or blog post, op-ed or white paper surfaces casting big data,[1] for better or worse, as groundbreaking, transformational, and "disruptive." Big data, we are told, is reshaping countless aspects of modern life, from medicine to commerce to national security. It may even change humanity's conception of existence: in the future, "we will no longer regard our world as a string of happenings that we explain as natural or social phenomena, but as a universe comprised essentially of information."[2]

This revolution has its dissidents. Critics worry the world's increasing "datafication" ignores or even smothers the unquantifiable, immeasurable, ineffable parts of human experience.[3] They warn of big data's other dark sides, too: potential government abuses of civil liberties, erosion of long-held privacy norms, and even environmental damage (the "server farms" used to process big data consume huge amounts of energy).

Legal debates over big data focus on the privacy and civil liberties concerns of those people swept up in its net, and on whether existing safeguards—minimization, notice, consent, anonymization, the Fourth Amendment, and so on—offer sufficient protection. It is a perspective of *inclusion*. And that perspective makes sense: most people, at least in the industrialized world, routinely contribute to and experience the effects of big data. Under that conception, big data is the whale, and we are all of us Jonah.

This Essay takes a different approach, exploring big data instead from a perspective of *exclusion*. Big data poses risks also to those persons who are *not* swallowed up by it—whose information is not regularly harvested, farmed, or mined. (Pick your anachronistic metaphor.) Although proponents and skeptics alike tend to view this revolution as totalizing and universal, the reality is that billions of people remain on its margins because they do not routinely engage in activities that big data and advanced analytics are designed to capture.[4]

* Attorney-Adviser, Office of the Legal Adviser, U.S. Department of State. The views expressed in this Essay are my own and do not necessarily represent those of the U.S. Department of State or the United States government. This Essay's title is inspired by Ediberto Román's book *Citizenship and Its Exclusions* (2010). I am grateful to Benita Brahmbatt for her helpful comments on an earlier draft and to Paul Schwartz for spurring my interest in the relationship between privacy and democracy. Any errors are my own.

Whom does big data exclude? What are the consequences of exclusion for them, for big data as a technology, and for societies? These are underexplored questions that deserve more attention than they receive in current debates over big data. And because these technologies pose unique dangers to equality, and not just privacy, a new legal doctrine may be needed to protect those persons whom the big data revolution risks sidelining. I call it *data antisubordination*.

* * *

Big data, for all its technical complexity, springs from a simple idea: gather enough details about the past, apply the right analytical tools, and you can find unexpected connections and correlations, which can help you make unusually accurate predictions about the future—how shoppers decide between products, how terrorists operate, how diseases spread. Predictions based on big data already inform public- and private-sector decisions every day around the globe. Experts project big data's influence only to grow in coming years.[5]

If big data, as both an epistemological innovation and a new booming industry, increasingly shapes government and corporate decisionmaking, then one might assume much attention is paid to who and what shapes big data—the "input." In general, however, experts express a surprising nonchalance about the precision or provenance of data. In fact, they embrace "messiness" as a virtue.[6] Datasets need not be pristine; patterns and trends, not granularity or exactness, are the goal. Big data is so big—terabytes, petabytes, exabytes—that the sources or reliability of particular data points cease to matter.

Such sentiments presume that the inevitable errors creeping into large datasets are random and absorbable, and can be factored into the ultimate analysis. But there is another type of error that can infect datasets, too: the nonrandom, systemic omission of people who live on big data's margins, whether due to poverty, geography, or lifestyle, and whose lives are less "datafied" than the general population's. In key sectors, their marginalization risks distorting datasets and, consequently, skewing the analysis on which private and public actors increasingly depend. They are big data's exclusions.

Consider two hypothetical people.

The first is a thirty-year-old white-collar resident of Manhattan. She participates in modern life in all the ways typical of her demographic: smartphone, Google, Gmail, Netflix, Spotify, Amazon. She uses Facebook, with its default privacy settings, to keep in touch with friends. She dates through the website OkCupid. She travels frequently, tweeting and posting geotagged photos to Flickr and Instagram. Her wallet holds a debit card, credit cards, and a MetroCard for the subway and bus system. On her keychain are plastic barcoded cards for the "customer rewards" programs of her grocery and drugstore. In her car, a GPS sits on the dash, and an E-ZPass transponder (for bridge, tunnel, and highway tolls) hangs from the windshield.

The data that she generates every day—and that governments and companies mine to learn about her and people like her—are nearly incalculable. In addition to information collected by companies about her spending, communications, online activities, and movement, government agencies (federal, state, local) know her well: New York has transformed itself in recent years into a supercharged generator of big data.[7] Indeed, for our Manhattanite, avoiding capture by big data is impossible. To begin even to limit her exposure—to curb her contributions to the city's rushing data flows—she would need to fundamentally reconstruct her everyday life. And she would have to move, a fate anathema to many New Yorkers. Thus, unless she takes relatively drastic steps, she will continue to generate a steady data flow for government and corporate consumption.

Now consider a second person. He lives two hours southwest of Manhattan, in Camden, New Jersey, America's poorest city. He is underemployed, working part-time at a restaurant, paid under the table in cash. He has no cell phone, no computer, no cable. He rarely travels and has no passport, car, or GPS. He uses the Internet, but only at the local library on

public terminals. When he rides the bus, he pays the fare in cash.

Today, many of big data's tools are calibrated for our Manhattanite and people like her—those who routinely generate large amounts of electronically harvestable information. A world shaped by big data will take into account her habits and preferences; it will look like her world. But big data currently overlooks our Camden subject almost entirely. (And even he, simply by living in a U.S. city, has a much larger data footprint than someone in Eritrea, for example.) In a future where big data, and the predictions it makes possible, will fundamentally reorder government and the marketplace, the exclusion of poor and otherwise marginalized people from datasets has troubling implications for economic opportunity, social mobility, and democratic participation. These technologies may create a new kind of voicelessness, where certain groups' preferences and behaviors receive little or no consideration when powerful actors decide how to distribute goods and services and how to reform public and private institutions.

This might sound overheated. It is easy to assume that exclusion from the big data revolution is a trivial concern—a matter simply of not having one's Facebook "likes" or shopping habits considered by, say, Walmart. But the consequences of exclusion could be much more profound than that.

First, those left out of the big data revolution may suffer tangible economic harms. Businesses may ignore or undervalue the preferences and behaviors of consumers who do not shop in ways that big data tools can easily capture, aggregate, and analyze. Stores may not open in their neighborhoods, denying them not just shopping options, but also employment opportunities; certain promotions may not be offered to them; new products may not be designed to meet their needs, or priced to meet their budgets. Of course, poor people and minority groups are in many ways already marginalized in the marketplace. But big data could reinforce and exacerbate existing problems.

Second, politicians and governments may come to rely on big data to such a degree that exclusion from data flows leads to exclusion from civic and political life—a barrier to full citizenship. Political campaigns already exploit big data to raise money, plan voter-turnout efforts, and shape their messaging.[8] And big data is quickly making the leap from politics to policy: the White House, for example, recently launched a $200 million big data initiative to improve federal agencies' ability "to access, organize, and glean discoveries from huge volumes of digital data."[9]

Just as U.S. election districts—and thus U.S. democracy—depend on the accuracy of census data, so too will policymaking increasingly depend on the accuracy of big data and advanced analytics. Exclusion or underrepresentation in government datasets, then, could mean losing out on important government services and public goods. The big data revolution may create new forms of inequality and subordination, and thus raises broad democracy concerns.

* * *

"There is no caste here," Justice Harlan said of the United States, "no superior, dominant, ruling class of citizens."[10] But big data has the potential to solidify existing inequalities and stratifications and to create new ones. It could restructure societies so that the only people who matter—quite literally the only ones who count—are those who regularly contribute to the right data flows.

Recently, some scholars have argued that existing information privacy laws—whether the U.S. patchwork quilt or Europe's more comprehensive approach—may be inadequate to confront big data's privacy risks. But big data threatens more than just privacy. It could also jeopardize political and social equality by relegating vulnerable people to an inferior status.

U.S. equal protection doctrine, however, is ill suited to the task of policing the big data revolution. For one thing, the poor are not a protected class,[11] and thus the doctrine would do little to ensure, either substantively or

procedurally, that they share in big data's benefits. And the doctrine is severely limited in its ability to "address[] disadvantage that cannot readily be traced to official design or that affects a diffuse and amorphous class."[12] Moreover, it is hard to imagine what formal equality or "anticlassification" would even look like in the context of big data.[13]

Because existing equality law will not adequately curb big data's potential for social stratification, it may become necessary to develop a new equality doctrine—a principle of *data antisubordination*. Traditionally, U.S. antisubordination theorists have argued "that guarantees of equal citizenship cannot be realized under conditions of pervasive social stratification," and "that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups."[14] This antisubordination approach—what Owen Fiss called the "group-disadvantaging principle"[15]—may need to be revised, given big data's potential to impose new forms of stratification and to reinforce the status of already-disadvantaged groups.[16]

A data antisubordination principle would, at minimum, provide those who live outside or on the margins of data flows some guarantee that their status as persons with light data footprints will not subject them to unequal treatment by the state in the allocation of public goods or services. Thus, in designing new public-safety and job-training programs, forecasting future housing and transportation needs, and allocating funds for schools and medical research—to name just a few examples—public institutions could be required to consider, and perhaps work to mitigate, the disparate impact that their use of big data may have on persons who live outside or on the margins of government datasets. Similarly, public actors relying on big data for policymaking, lawmaking, election administration, and other core democratic functions could be required to take steps to ensure that big data's marginalized groups continue to have a voice in democratic processes. That a person might make only limited contributions to government data flows should not relegate him to political irrelevance or inferiority.

Data antisubordination could also (or alternatively) provide a framework for judicial review of congressional and executive exploitation of big data and advanced analytics.[17] That framework could be modeled on John Hart Ely's "representation-reinforcing approach" in U.S. constitutional law,[18] under which "a court's ability to override a legislative judgment ought to be calibrated based on the fairness of the political process that produced the judgment."[19] In the context of big data, rather than mandating any particular substantive outcome, a representation-reinforcing approach to judicial review could provide structural, process-based safeguards and guarantees for those people whom big data currently overlooks, and who have had limited input in the political process surrounding government use of big data.

To be most effective, however, a data antisubordination principle would need to extend beyond state action. Big data's largest private players exert an influence on societies, and a power over the aggregation and flow of information, that in previous generations not even governments enjoyed. Thus, a data antisubordination principle would be incomplete unless it extended, in some degree, to the private sector, whether through laws, norms, or standards.

Once fully developed as theory, a data antisubordination principle—at least as it applies to state action—could be enshrined in law by statute. Like GINA,[20] it would be a civil rights law designed for potential threats to equal citizenship embedded in powerful new technologies—threats that neither the Framers nor past civil rights activists could have envisioned.

As lines between the physical and datafied worlds continue to blur, and as big data and advanced analytics increasingly shape governmental and corporate decisionmaking about the allocation of resources, equality and privacy principles will grow more and more intertwined. Law must keep pace. In "The Right to Privacy," their 1890 *Harvard Law Review* article, a young Louis Brandeis and co-author Samuel Warren recognized that "[r]ecent inventions and business methods call attention

to the next step which must be taken for the protection of the person."[21] The big data revolution, too, demands "next steps," and not just in information privacy law. Brandeis and Warren's "right to be let alone"—which Brandeis, as a Supreme Court justice, would later call the "most comprehensive of rights and the right most valued by civilized men"[22] —has become an obsolete and insufficient protector.[23] Even more modern information privacy principles, such as consent and the nascent "right to be forgotten,"[24] may turn out to have only limited utility in an age of big data.

Surely revised privacy laws, rules, and norms will be needed in this new era. But they are insufficient. Ensuring that the big data revolution is a just revolution, one whose benefits are broadly and equitably shared, may also require, paradoxically, a right *not* to be forgotten—a right against exclusion.

---

[1] In this Essay, I use the term *big data* as shorthand for a variety of new technologies used to create and analyze datasets "whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." MCKINSEY GLOBAL INST., BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 1 (2011), *available at* http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Big data can include "[t]raditional enterprise data," such as web store transaction information; "[m]achine-generated/sensor data"; and "[s]ocial data." ORACLE, BIG DATA FOR THE ENTERPRISE 3 (2013), http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

[2] VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 96 (2013).

[3] David Brooks, for example, frets about the rise of "data-ism." David Brooks, Op-Ed., *The Philosophy of Data*, N.Y. TIMES, Feb. 5, 2013, at A23, *available at*http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html. Similarly, Leon Wieseltier rejects the (false, in his view) "religion of information." Leon Wieseltier, *What Big Data Will Never Explain*, NEW REPUBLIC (Mar. 26, 2013), *available at*http://www.newrepublic.com/article/112734/what-big-data-will-never-explain.

[4] These activities include, for example, using the Internet, especially for e-mail, social media, and searching; shopping with a credit, debit, or "customer loyalty" card; banking or applying for credit; traveling by plane; receiving medical treatment at a technologically advanced hospital; and receiving electricity through a "smart meter." A 2013 report by the International Telecommunications Union found that only thirty-nine percent of the world's population uses the Internet. INT'L TELECOMMS. UNION, THE WORLD IN 2013: ICT FACTS AND FIGURES 2 (2013), *available at* http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf. The report found that in Africa, "16% of people are using the Internet—only half the penetration rate of Asia and the Pacific." *Id.* This disparity will likely shrink in coming years. For example, Facebook and several mobile phone companies recently announced internet.org, "a global partnership" aimed at "making internet access available to the next 5 billion people." Press Release, Facebook, Technology Leaders Launch Partnership to Make Internet Access Available to All (Aug. 20, 2013), http://newsroom.fb.com/News/690/Technology-Leaders-Launch-Partnership-to-Make-Internet-Access-Available-to-All. The Obama administration is also working to expand Internet access in underserved communities within the United States. *See* Edward Wyatt, *Most of U.S. Is Wired, but Millions Aren't Plugged In*, N.Y. TIMES, Aug. 19, 2013, at B1, *available at*http://www.nytimes.com/2013/08/19/technology/a-push-to-connect-millions-who-live-offline-to-the-internet.html.

[5] *See* MCKINSEY GLOBAL INST., *supra* note 1, at 16.

[6] *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 2, at 32-49. "In a world of small data," write Mayer-Schönberger and Cukier, "reducing errors and ensuring high quality of data was a natural and essential impulse," but in the world of big data such precision ceases to be necessary: the new datasets are large enough to compensate for the "erroneous figures and corrupted bits" that may find their way into any dataset. *Id.* at 32. Indeed, in the big data world, "allowing for imprecision—for messiness—may be a positive feature, not a shortcoming," because "[i]n return for relaxing the standards of allowable errors, one can get ahold of much more data." *Id.* at 33.

[7] Manhattan alone has a network of some three thousand closed-circuit television cameras recording terabytes of data every day and accessible to local and federal law enforcement. The New York City Police Department's Domain Awareness System, a new $40 million data-collection program developed by Microsoft, can track our subject's movements via the city's CCTV network and hundreds of automated license plate readers "mounted on police cars and deployed at bridges, tunnels, and streets." Michael Endler, *NYPD, Microsoft Push Big Data Policing into Spotlight*, INFORMATION WEEK (Aug. 20, 2012),http://www.informationweek.com/security/privacy/nypd-microsoft-push-big-data-policing-in/240005838. The city can also track her movements through her MetroCard and E-ZPass—useful data not only for law enforcement, but also for determining new transit schedules, planning roads, and setting tolls. To crunch these data, the city employs a dedicated staff of "quants" in the Office of Policy and Strategic Planning. They analyze subjects ranging from commuting habits to electricity use, from children's test scores to stop-and-frisk statistics. *See* Alan Feuer, *The Mayor's Geek Squad*, N.Y. TIMES, Mar. 23, 2013, at MB1, *available at* http://www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html.

[8] President Obama's 2008 and 2012 campaigns are the most famous examples of this phenomenon. *See, e.g.*, Jim Rutenberg, *Data You Can Believe In*, N.Y. TIMES, June 23, 2013, at MM22, *available at* http://www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html.

[9] Tom Kalil, *Big Data Is a Big Deal*, WHITE HOUSE OFFICE OF SCI. AND TECH. POLICY BLOG (Mar. 29, 2012, 9:23

AM),http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (describing the National Big Data Research and Development Initiative); *see also* Fact Sheet, Exec. Office of the President, *Big Data Across the Federal Government* (Mar. 29, 2012),http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf (highlighting ongoing federal programs that seek to exploit big data's potential; Tom Kalil & Fen Zhao, *Unleashing the Power of Big Data*, White House Office of Sci. and Tech. Policy Blog (Apr. 18, 2013, 4:04 PM),http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data (providing an update on the progress of the Big Data Initiative).

[10] *Plessy v. Ferguson*, 163 U.S. 537, 559 (1896) (Harlan, J., dissenting).

[11] *See, e.g.*, Mario L. Barnes & Erwin Chemerinsky, *The Disparate Treatment of Race and Class in Constitutional Jurisprudence*, 72 L. & CONTEMP. PROBS. 109, 110-12 (2009).

[12] Goodwin Liu, *Education, Equality, and National Citizenship*, 116 YALE L.J. 330, 334 (2006).

[13] It would be a strange law indeed that compelled public and private users of big data to collect *everyone's* information, all the time, in the name of equality, or to collect information from different racial and socioeconomic groups proportionally. After all, two of big data's supposed built-in privacy safeguards are anonymization and randomization. Making big data and advanced analytics resemble other processes already governed by U.S. equal protection law—redistricting, for example—could mean requiring collectors of data to initially determine the race and class of a person before collecting his underlying information: a sort of double privacy intrusion, and in any event surely unworkable in practice. Similarly, a strict anticlassification conception of equality would be incompatible with the very idea of big data—a primary purpose of which, after all, is to streamline and enhance users' ability to classify individuals and groups based on their behaviors—and would fail to address the marginalization concerns I have outlined here.

[14] Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 9 (2003).

[15] *See* Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 147-56 (1976).

[16] In addition, the protections provided by existing international law instruments, such as the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights, may need updating to address on a global scale the potential stratifying effects of big data. After all, big data is an international phenomenon, and just as the Internet has blurred borders, so too will big data and its effects traverse the globe.

[17] That review could build on the famous footnote four of *United States v. Carolene Products*. 304 U.S. 144, 152 n.4 (1938) (recognizing that "prejudice against discrete and insular minorities may be a special condition, which tends seriously to curtail the operation of those political processes ordinarily to be relied upon to protect minorities, and which may call for a correspondingly more searching judicial inquiry").

[18] *See* JOHN HART ELY, DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW 101 (1980) (contrasting the representation-reinforcing approach with "an approach geared to the judicial imposition of 'fundamental values'").

[19] Jane S. Schacter, *Ely at the Altar: Political Process Theory Through the Lens of the Marriage Debate*, 109 MICH. L. REV. 1363, 1364 (2011). As Schacter notes, this political process theory functions as "a simple, but central, principle of institutional architecture" in U.S. constitutional law. *Id.* Although I am not proposing the constitutionalization of new rights related to big data, some version of Ely's political process theory could also provide an "institutional architecture" for government use of these technologies.

[20] Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233, 122 Stat. 881 (2008). After its unanimous passage, Senator Edward M. Kennedy called the Act "the first civil rights bill of the new century of the life sciences." *See* David H. Kaye, Commentary, *GINA's Genotypes*, 108 MICH. L. REV. FIRST IMPRESSIONS 51, 51 (2010), http://www.michiganlawreview.org/assets/fi/108/kaye2.pdf.

[21] Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195 (1890).

[22] Olmstead v. United States, 277 U.S. 438, 478 (1928) (Brandeis, J., dissenting).

[23] Paul Schwartz recognized this deficiency almost twenty years ago. *See* Paul M. Schwartz, *Privacy and Participation: Personal Information and Public Sector Regulation in the United States*, 80 IOWA L. REV. 553, 558-63 (1995) (arguing that "privacy as the right to be let alone serves as an incomplete paradigm in the computer age").

[24] *See* Kate Connolly, *Right to Erasure Protects People's Freedom to Forget the Past, Says Expert*, GUARDIAN (APR. 4, 2013),http://www.theguardian.com/technology/2013/apr/04/right-erasure-protects-freedom-forget-past (interview with Viktor Mayer-Schönberger about the right to be forgotten). *But see* Google Spain SL v. Agencia Española de Protección de Datos, No. C-131/12, ¶ 138.3 (E.C.J. June 25, 2013) (opinion of Advocate General) (EUR-Lex) (concluding that a person has no right under the European Union's Data Protection Directive to "be consigned to oblivion" by demanding deletion from Google's "indexing of the information relating to him personally, published legally on third parties' web pages"). *See generally* VIKTOR MAYER-SCHÖNBERGER, DELETE: THE VIRTUE OF FORGETTING IN THE DIGITAL AGE (2009) (arguing for a right to be forgotten in an era of "comprehensive digital memory").

RELATIONAL BIG DATA

*Karen E.C. Levy**

66 STAN. L. REV. ONLINE 73

"Big Data" has attracted considerable public attention of late, garnering press coverage both optimistic and dystopian in tone. Some of the stories we tell about big data treat it as a computational panacea—a key to unlock the mysteries of the human genome, to crunch away the problems of urban living, or to elucidate hidden patterns underlying our friendships and cultural preferences.[1] Others describe big data as an invasive apparatus through which governments keep close tabs on citizens, while corporations compile detailed dossiers about what we purchase and consume.[2] Like so many technological advances before it, our stories about big data generate it as a two-headed creature, the source of both tremendous promise and disquieting surveillance. In reality, like any complicated social phenomenon, big data is both of these, a set of heterogeneous resources and practices deployed in multiple ways toward diverse ends.[3]

I want to complicate matters further by suggesting another way in which data has become big: data now *mediate our day-to-day social relationships* to an unprecedented degree. This other big data revolution relies on the proliferation of new data collection and analysis tools that allow individuals to track easily, quantify, and communicate information about our own behaviors and those of others. This type of big data arguably touches more of us more directly than the big data practices more commonly discussed, as it comes to reshape our relationships across multiple domains of daily life.

In this sense, data is big not because of the number of points that comprise a particular dataset, nor the statistical methods used to analyze them, nor the computational power on which such analysis relies. Instead, data is big because of the depth to which it has come to pervade our personal connections to one another. A key characteristic of this flavor of big data, which I term "relational"[4] (more on this in a moment) is *who* is doing the collection and analysis. In most big data stories, both dreamy and dystopian, collection and analysis are *top-down*, driven by corporations, governments, or academic institutions. In contrast, relational big data is collected and analyzed by *individuals*, inhabiting social roles (as parents, friends, etc.) *as a means for negotiating social life*. In other words, we can understand big data not simply as a methodological watershed, but as a fundamental social shift in how people manage relationships and make choices, with complex implications for privacy, trust, and dynamics of interpersonal control.

Another notable distinction is the multiplicity of sources of relational big data. While most analyses of social big data focus on a few behemoth forums for online information-seeking and interaction, what Zeynep Tufekci describes as "large-scale aggregate databases of imprints of online and social media activity"[5]—Google, Facebook, Twitter, and the like—I suggest that "data-fication" extends well beyond these digital presences, extending into diverse domains and relying on multiple dispersed tools, some of which are household names and some of which never will be.

* Ph.D. Candidate, Princeton University.

In the rest of this Essay, I flesh out the idea of relational big data by describing its conceptual predecessor in economic sociology. I suggest a few domains in which data mediate social relationships and how interactions might change around it. I then consider what analytical purchase this flavor of big data gets us regarding questions of policy in the age of ubiquitous computing.

I. What's Relational about Data?

To say that big data is *relational* borrows a page from economic sociology, particularly from the work of Viviana Zelizer.[6] As its name implies, economic sociology broadly examines the social aspects of economic life, from how markets are structured to the development of money. One of Zelizer's seminal contributions to the field is the idea that economic exchanges do "relational work" for people: through transactions, people create and manage their interpersonal ties. For example, individuals vary the features of transactions (in search of what Zelizer calls "viable matches" among interpersonal ties, transactions, and media) in order to differentiate social relationships and create boundaries that establish what a relationship *is* and *is not*. (Consider, for instance, why you might feel more comfortable giving a coworker a gift certificate as a birthday present rather than cash.) Thus, to construe transactions merely as trades of fungible goods and services misses a good part of what's interesting and important about them.

I suggest that we should do for data practices what Zelizer does for economic practices: we should consider that people use data to create and define relationships with one another. Saying that data practices are relational does more than simply observe that they occur against a background of social networks; rather, people *constitute and enact* their relations with one another *through* the use and exchange of data.[7] Consider, for example, a person who monitors the real-time location of her friends via a smartphone app designed for this purpose. By monitoring some friends but not others, she differentiates among her relationships, defining some as closer. By agreeing to share their locations, her friends communicate that they have no expectation of privacy (to her) as to where they are, perhaps suggesting that they trust her. The acts of sharing and monitoring say a lot about the nature of the relationship; focusing only on the locational data itself, as much big data analysis does, ignores the social negotiations taking place via data practices.

Big data is, at heart, a social phenomenon—but many of the stories we tell about it reduce people to mere data points to be acted upon. A relational framework is appealing because it puts people, their behaviors, and their relationships at the center of the analysis as active agents. Big data and its attendant practices aren't monoliths; they are diverse and socially contingent, a fact which any policy analysis of big data phenomena must consider.

II. Big Data Domains

Data pervade all kinds of social contexts, and the tools available to gather and use data vary tremendously across them. In what types of relationships do data circulate? I touch on a few here.

Children and families. Technologies for data gathering and surveillance within families are proliferating rapidly. A number of these involve monitoring the whereabouts of family members (often, though not always, children). One such product, LockDown GPS, transmits data about a vehicle's speed and location so parents can easily monitor a teen's driving habits. The system can prevent a car from being restarted after it's been shut off, and parents are immediately notified of rule violations by e-mail. The system purports to "[put] the parent in the driver's seat 24 hours a day, from anywhere in the world."[8]

A number of other products and apps (like FlexiSpy, Mamabear, My Mobile Watchdog, and others) allow individuals to monitor data like the calls a family member receives, the content of texts and photos, real-time location, Facebook activity, and the like, with or without the monitored party being aware of it. And not all intra-family monitoring is child-directed: a number of products market themselves as tools for tracking down untrustworthy spouses,[9] while others detect such behaviors as whether an elder parent has taken his or her medicine.[10]

Communities and friendships. Jeffrey Lane's ethnographic account of three years spent living with Harlem youth describes how they manage diverse relationships with friends, rivals, and authority figures using social media.[11] An abundance of other tools enable us to relate to our communities through data by, for instance, finding friends in physical space (Find My Friends), selecting local businesses to patronize (Yelp), or "checking in" to physical locations (Foursquare).

The workplace. The use of productivity metrics to manage employees is far from new, but the proliferation of tools for doing so introduces data into new kinds of employment relationships. Parents can monitor a caretaker's behavior via nanny cam. Fast-growing workplace wellness monitoring programs frequently use health indicators and behavioral data (derived, for instance, from a digital pedometer) to let employers and insurers keep tabs on the health of their workforce.[12] Highly mobile employees like truck drivers, who traditionally are accorded a good deal of occupational autonomy, are increasingly monitored via fleet management and dispatch systems that transmit data about their driving habits, fuel usage, and location to a central hub in real time—practices that have engendered deep concerns about driver privacy and harassment.[13]

Self-monitoring. Finally, individuals increasingly use electronic data gathering systems to control their *own* behavior. The Quantified Self "movement" is the most acute example of this—Quantified Selfers monitor their own biophysical, behavioral, and environmental markers in efforts to measure progress toward health and other goals.[14] Even among those who would not identify with such a movement, a number of self-tracking systems have recently emerged on the consumer electronics market (for example, the FitBit and Nike FuelBand), while popular services like 23AndMe, Mint, and Daytum facilitate tracking of genetic information, personal finance, and myriad other types of data. Even when monitoring is self-directed, however, these data can impact interpersonal relationships (for example, by facilitating comparison and competition within one's personal networks).[15]

In many areas of life, then, individuals use data gathering and analysis tools to manage their relationships with one another in a variety of ways, only a few of which I mention here. In some cases, data help people to *control* the actions of others by serving as a digital site of accountability for action, potentially diminishing the need for social trust (for instance, monitoring a teen's car may effectively undermine the need for parent-child trust by creating a seemingly objective record of compliance or noncompliance with parental rules). In others, technologies facilitate *competition* in relationships: employment metrics are commonly publicized to encourage intra-workforce competition, and many health-centric data services allow and encourage users to compete with peers and strangers. Such competition is not merely an externality of the use of these devices, but a central reason why these techniques can be effective. Third, data practices may help individuals to *distinguish* between relationships and send desired signals to one another (e.g., as suggested earlier, adding certain friends but not others to a find-my-friends service). The meanings and effects of data practices vary considerably within and across life domains.

III. Policy, Privacy, Implications

Big data poses big problems for privacy,[16] which are only compounded by the relational framework I suggest. Top-down data collection programs create the need for strong civil liberties protections, due process, and assurances of data integrity. But the privacy interests implicated by relational big data are bound up in particular social contexts;[17] no single piece of legislation or court ruling would prove a useful tool to protect them.

Instead, it is likely that some privacy interests implicated by relational big data may figure into existing legal frameworks governing personal relationships (for instance, workplace harassment, or tort claims like invasion of privacy) or in some cases via domain-specific rules, such as laws governing the use of medical or genetic information.[18] Gathered data may also come to legal use as evidence, substantiating an alibi or providing proof of a fact like vehicle speed. But in most cases, interpersonal privacy intrusions facilitated by relational data-gathering tools fall outside the realm of legal redress, precisely because the law

is traditionally hesitant to get involved in the minutiae of personal relationships.

Despite the fact that law doesn't provide a clear approach, policymakers and privacy scholars still have much to gain from thinking about relational data practices. The ubiquity of interpersonal data-gathering activities helps us understand people as *both subjects and objects* of big data regimes, not just data points. When people collect and use data to constitute their relationships with one another, social norms around accountability, privacy, veracity, and trust are likely to evolve in complex ways.

In addition, thinking about individuals this way may be instructive when considering public responses to top-down surveillance. For instance, although recent revelations about the NSA's PRISM surveillance program (in which essentially every major technology provider secretly supplied consumer communications to the NSA) excited much outrage among academics and civil libertarians, news of the program's existence engendered a comparatively tepid response from the general public.[19] Part of the reason may be that we have become docile[20] in light of the ubiquity and pervasiveness of data gathering across domains of daily life. Relational data practices may instill in the public a tolerance for watching and being watched, measuring and being measured, that leads us to abide additional surveillance without much complaint.

---

[1] *See* Michael Specter, *Germs Are Us*, NEW YORKER, Oct. 22, 2012, at 32, *available at*http://www.newyorker.com/reporting/2012/10/22/121022fa_fact_specter (human genome); Alan Feuer, *The Mayor's Geek Squad*, N.Y. TIMES, Mar. 24, 2013, at MB1, *available at* http://www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html (city services); Nick Bilton, *Looking at Facebook's Friend and Relationship Status Through Big Data*, N.Y. Times Bits Blog (Apr. 25, 2013), http://bits.blogs.nytimes.com/2013/04/25/looking-at-facebooks-friend-and-relationship-status-through-big-data (interpersonal relationships).

[2] Two much-discussed recent examples are the National Security Agency's wide-ranging PRISM data collection program, Charlie Savage et al., *U.S. Confirms Gathering of Web Data Overseas*, N.Y. TIMES, June 7, 2013, at A1, *available at* ://www.nytimes.com/2013/06/07/us/nsa-verizon-calls.html, and the revelation that Target collected purchasing data that predicted the pregnancy of a teenage girl before her family knew about it, Charles Duhigg,*Psst, You in Aisle 5*, N.Y. TIMES, Feb. 19, 2012, at MM30, *available at* http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[3] The slipperiness of the definition here isn't helped by the vagueness around whether big data is *data* or *practice*—the millions or billions of pieces of information being examined or the methodological tools for its examination. Much of the "new" information to which big data refers isn't actually new (we have always had a genome); what *is* new is our capacity to collect and analyze it.

[4] My use of the term "relational" here is distinct from the computational meaning of the word (i.e., relating to the structure of a database). I also do not mean "relational" in the sense of merely having to do with social networks and communications, though other big data analysis is based on such associations. *See, e.g.*, Katherine J. Strandburg, *Freedom of Association in a Networked World: First Amendment Regulation of Relational Surveillance*, 49 B.C. L. REV. 741 (2008).

[5] Zeynep Tufekci, *Big Data: Pitfalls, Methods, and Concepts for an Emergent Field*, *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2229952(Mar. 7, 2013). A great deal of scholarly work has investigated how digital communication forums like Facebook and Twitter mediate interactions both on- and off-line. *See, e.g.*, Alice Marwick & danah boyd, *I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience*, 13 NEW MEDIA AND SOC'Y 114 (2010).

[6] Viviana Zelizer, *How I Became a Relational Economic Sociologist and What Does That Mean?*, 40 POL. & SOC'Y 145 (2012) [hereinafter Zelizer, *Relational Economic Sociologist*]; Viviana Zelizer, *Pasts and Futures of Economic Sociology*, 50 AM. BEHAV. SCIENTIST 1056 (2007).

[7] Zelizer similarly contrasts her relational perspective with the previous "embeddedness" approach in economic sociology. Zelizer, *Relational Economic Sociologist*, *supra* note 6, at 162.

[8] *Family Protection,* LockDown System, Inc., http://www.lockdownsystems.com/basic-page/family-protection (last visited Aug. 29, 2013).

[9] *See, e.g.*, Sophie Curtis, *'Boyfriend Tracker' App Banished from Google Play*, Telegraph (Aug. 22, 2013),http://www.telegraph.co.uk/technology/news/10259516/Boyfriend-Tracker-app-banished-from-Google-Play.html.

[10] *See, e.g.*, *How the Philips Medication Dispensing Service Works*, PHILIPS, http://www.managemypills.com/content/How_PMD_Works (last visited Aug. 29, 2013).

[11] Jeffrey Lane, Presentation on Code-Switching on the Digital Street at the American Sociological Association Annual Meeting (August 12, 2013);*see also* danah boyd & Alice Marwick, Social Steganography: Privacy in Networked Publics (May 9, 2011) (unpublished manuscript),*available at* http://www.danah.org/papers/2011/Steganography-ICAVersion.pdf.

[12] Workplace health monitoring practices are not without critics; CVS recently faced criticism from privacy advocates for its announcement that workers would be fined $600 per year if they

failed to disclose health metrics to the company's insurer. *See* Christine McConville, *CVS Presses Workers for Medical Information*, BOS. HERALD (Mar. 19, 2013), http://bostonherald.com/business/healthcare/2013/03/cvs_presses_workers_for_medical_information.

[13] For instance, new proposed regulations that would require truckers' work hours to be electronically monitored have been challenged due to the possibility that motor carriers will use the technology to harass drivers. *See Owner-Operators Indep. Drivers Ass'n v. Fed. Motor Carrier Safety Admin.*, 656 F.3d 580 (7th Cir. 2011).

[14] *See* Gary Wolf, *The Data-Driven Life*, N.Y. TIMES, May 2, 2010, at MM38, *available at* http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html. Of course, this phenomenon is not entirely new; self-monitoring has long been an element of many wellness efforts before the digital age (e.g., analog diet and exercise tracking). But, digital tools markedly increase the scale and depth of monitoring programs, as well as facilitating the use of such data for interpersonal competition.

[15] A recent anecdote in the *New Yorker* describes husband-and-wife FitBit users; the husband checks the wife's activity stats at the end of the day and will suddenly take the dog for a walk, while the wife, knowing what her husband is up to, paces around the house while he's gone to prevent him from "winning." Susan Orlean, *The Walking Alive*, NEW YORKER, May 20, 2013, at 44, 47, *available at*http://www.newyorker.com/reporting/2013/05/20/130520fa_fact_orlean.

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

[17] *See* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010).

[18] For instance, laws like Health Insurance Portability and Accountability Act (HIPAA) of 1996, Pub. L. No. 104-91, 110 Stat. 1936, and Genetic Information Nondiscrimination Act (GINA) of 2008, Pub. L. No. 110-233, 122 Stat. 881, protect privacy interests in health-related and genetic information.

[19] Poll data suggest that sixty-six percent of Americans support the government's collection of Internet data via the PRISM program. Brett LoGiurato, *The NSA's PRISM Program is Shockingly Uncontroversial with the American Public*, BUS. INSIDER (June 17, 2013), http://www.businessinsider.com/prism-surveillance-poll-nsa-obama-approval-2013-6.

[20] Michel Foucault famously described how disciplinary techniques create "docile bodies" accustomed to further discipline. For instance, he observed that schools operated as "pedagogical machine[s]," analogous to institutions like the factory and the prison: by inculcating disciplinary systems in students, schools prepare young subjects to encounter similar techniques in other realms. MICHEL FOUCAULT, DISCIPLINE AND PUNISH: THE BIRTH OF THE PRISON 172 (1977).

PRIVACY SUBSTITUTES

*Jonathan Mayer & Arvind Narayanan**

66 STAN. L. REV. ONLINE 89

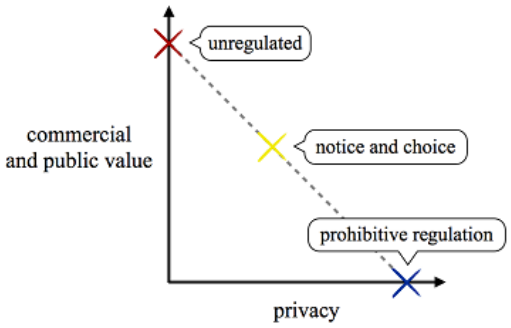
INTRODUCTION

Debates over information privacy are often framed as an inescapable conflict between competing interests: a lucrative or beneficial technology, as against privacy risks to consumers. Policy remedies traditionally take the rigid form of either a complete ban, no regulation, or an intermediate zone of modest notice and choice mechanisms.

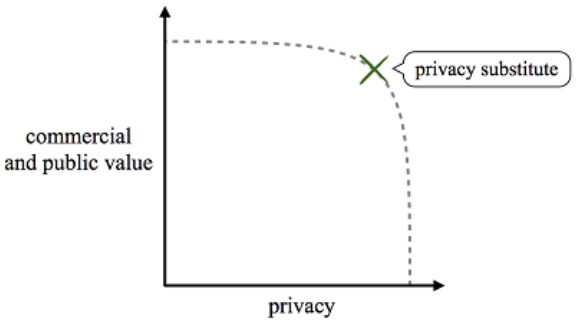
We believe these approaches are unnecessarily constrained. There is often a spectrum of technology alternatives that trade off functionality and profit for consumer privacy. We term these alternatives "privacy substitutes," and in this Essay we argue that public policy on information privacy issues can and should be a careful exercise in both selecting among, and providing incentives for, privacy substitutes.[1]

I. DISCONNECTED POLICY AND COMPUTER SCIENCE PERSPECTIVES

Policy stakeholders frequently approach information privacy through a simple balancing. Consumer privacy interests rest on one side of the scales, and commercial and social benefits sit atop the other.[2] Where privacy substantially tips the balance, a practice warrants prohibition; where privacy is significantly outweighed, no restrictions are appropriate. When the scales near equipoise, practices merit some (questionably effective[3]) measure of mandatory disclosure or consumer control.[4]



Computer science researchers, however, have long recognized that technology can enable tradeoffs between privacy and other interests. For most areas of technology application, there exists a spectrum of possible designs that vary in their privacy and functionality[5] characteristics. Cast in economic terms, technology enables a robust production-possibility frontier between privacy and profit, public benefit, and other values.



The precise contours of the production-possibility frontier vary by technology application

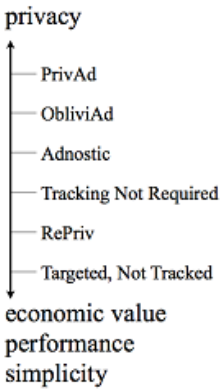
* Jonathan Mayer is Junior Affiliate Scholar, Stanford Center for Internet and Society. Arvind Narayanan is Assistant Professor, Department of Computer Science, Princeton University; Affiliate Scholar, Center for Internet and Society at Stanford Law School.

area. In many areas, privacy substitutes afford a potential Pareto improvement relative to naïve or status quo designs. In some application areas, privacy substitutes even offer a strict Pareto improvement: privacy-preserving designs can provide the *exact same* functionality as intrusive alternatives. The following Subparts review example designs for web advertising, online identity, and transportation payment to illustrate how clever engineering can counterintuitively enable privacy tradeoffs.

A. Web Advertising

In the course of serving an advertisement, dozens of third-party websites may set or receive unique identifier cookies.[6] The technical design is roughly akin to labeling a user's web browser with a virtual barcode, then scanning the code with every page view. All advertising operations—from selecting which ad to display through billing—can then occur on advertising company backend services. Policymakers and privacy advocates have criticized this status quo approach as invasive since it incorporates collection of a user's browsing history.[7] Privacy researchers have responded with a wide range of technical designs for advertising functionality.[8]

Frequent buyer programs provide a helpful analogy. Suppose a coffee shop offers a buy-ten-get-one-free promotion. One common approach would be for the shop to provide a swipe card that keeps track of a consumer's purchases, and dispenses rewards as earned. An alternative approach would be to issue a punch card that records the consumer's progress towards free coffee. The shop still operates its incentive program, but note that it no longer holds a record of precisely what was bought when; the punch card keeps track of the consumer's behavior, and it only tells the shop what it needs to know. This latter implementation roughly parallels privacy substitutes in web advertising: common elements include storing a user's online habits within the web browser itself, as well as selectively parceling out information derived from those habits.



Each design represents a point in the spectrum of possible tradeoffs between privacy—here, the information shared with advertising companies—and other commercial and public values. Moving from top to bottom, proposals become easier to deploy, faster in delivery, and more accurate in advertisement selection and reporting—in exchange for diminished privacy guarantees.

B. Online Identity

Centralized online identity management benefits consumers through both convenience and increased security.[9] Popular implementations of these "single sign-on" or "federated identity" systems include a sharp privacy drawback, however: the identity provider learns about the consumer's activities. By way of rough analogy: Imagine going to a bar, where the bouncer phones the state DMV to check the authenticity of your driver's license. The bouncer gets confirmation of your identity, but the DMV learns where you are. Drawing on computer security research, Mozilla has deployed a privacy-preserving alternative, dubbed Persona. Through the use of cryptographic attestation, Persona provides centralized identity management without Mozilla learning the consumer's online activity. In the bar analogy, instead of calling the DMV, the bouncer carefully checks the driver's license for official and difficult-to-forge markings. The bouncer can still be sure of your identity, but the DMV does not learn of your drinking habits.

C. Transportation Payment

Transportation fare cards and toll tags commonly embed unique identifiers, facilitating intrusive tracking of a consumer's movements. Intuitively, the alternative privacy-preserving design would be to store the consumer's balance on the device, but this approach is vulnerable to cards being hacked for free transportation.[10] An area of cryptography called "secure multiparty computation" provides a solution, allowing two parties to transact while only learning as much about each other as is strictly mathematically necessary to complete the transaction.[11] A secure multiparty computation approach would enable the transportation provider to add reliably and deduct credits from a card or tag—without knowing the precise device or value stored.

II. Nonadoption of Privacy Substitutes

Technology organizations have rarely deployed privacy substitutes, despite their promise. A variety of factors have effectively undercut commercial implementation.

Engineering Conventions. Information technology design traditionally emphasizes principles including simplicity, readability, modifiability, maintainability, robustness, and data hygiene. More recently, overcollection has become a common practice—designers gather information wherever feasible, since it might be handy later. Privacy substitutes often turn these norms on their head. Consider, for example, "differential privacy" techniques for protecting information within a dataset.[12] The notion is to intentionally introduce (tolerable) errors into data, a practice that cuts deeply against design intuition.[13]

Information Asymmetries. Technology organizations may not understand the privacy properties of the systems they deploy. For example, participants in online advertising frequently claim that their practices are anonymous—despite substantial computer science research to the contrary.[14] Firms may also lack the expertise to be aware of privacy substitutes; as the previous Part showed, privacy substitutes often challenge intuitions and assumptions about technical design.

Implementation and Switching Costs. The investments of labor, time, and capital associated with researching and deploying a privacy substitute may be significant. Startups may be particularly resource constrained, while mature firms face path-dependent switching costs owing to past engineering decisions.

Diminished Private Utility. Intrusive systems often outperform privacy substitutes (e.g., in speed, accuracy, and other aspects of functionality), in some cases resulting in higher private utility. Moreover, the potential for presently unknown future uses of data counsels in favor of overcollection wherever possible.

Inability to Internalize. In theory, consumers or business partners might compensate a firm for adopting privacy substitutes. In practice, however, internalizing the value of pro-privacy practices has proven challenging. Consumers are frequently unaware of the systems that they interact with, let alone the privacy properties of those systems; informing users sufficiently to exercise market pressure may be impracticable.[15] Moreover, even if a sizeable share of consumers were aware, it may be prohibitively burdensome to differentiate those consumers who are willing and able to pay for privacy. And even if those users could be identified, it may not be feasible to transfer small amounts of capital from those consumers. As for business partners, they too may have information asymmetries and reflect (indirectly) lack of consumer pressure. Coordination failures compound the difficulty of monetizing privacy: without clear guidance on privacy best practices, users, businesses, and policymakers have no standard of conduct to which to request adherence.

Organizational Divides. To the extent technology firms do perceive pressure to adopt privacy substitutes, it is often from government relations, policymakers, and lawyers. In some industries the motivation will be another step removed, filtering through trade associations and lobbying groups. These nontechnical representatives often lack the expertise to propose privacy alternatives themselves or adequately solicit engineering input.[16]

Competition Barriers. Some technology sectors reflect monopolistic or oligopolistic structures. Even if users and businesses demanded

improved privacy, there may be little competitive pressure to respond.

III. Policy Prescriptions

Our lead recommendation for policymakers is straightforward: understand and encourage the use of privacy substitutes through ordinary regulatory practices. When approaching a consumer privacy problem, policymakers should begin by exploring not only the relevant privacy risks and competing values, but also the space of possible privacy substitutes and their associated tradeoffs. If policymakers are sufficiently certain that socially beneficial privacy substitutes exist,[17] they should turn to conventional regulatory tools to incentivize deployment of those technologies.[18] For example, a regulatory agency might provide an enforcement safe harbor to companies that deploy sufficiently rigorous privacy substitutes.

Policymakers should also target the market failures that lead to nonadoption of privacy substitutes. Engaging directly with industry engineers, for example, may overcome organizational divides and information asymmetries. Efforts at standardization of privacy substitutes may be particularly effective; information technology is often conducive to design sharing and reuse. We are skeptical of the efficacy of consumer education efforts,[19] but informing business partners could alter incentives.

Finally, policymakers should press the envelope of privacy substitutes. Grants and competitions, for example, could drive research innovations in both academia and industry.

Conclusion

This brief Essay is intended to begin reshaping policy debates on information privacy from stark and unavoidable conflicts to creative and nuanced tradeoffs. Much more remains to be said: Can privacy substitutes also reconcile individual privacy with government intrusions (e.g., for law enforcement or intelligence)?[20] How can policymakers recognize privacy substitute pseudoscience?[21] We leave these and many more questions for another day, and part ways on this note: pundits often cavalierly posit that information technology has sounded the death knell for individual privacy. We could not disagree more. Information technology is poised to protect individual privacy—if policymakers get the incentives right.

---

[1] The area of computer science that we discuss is sometimes referenced as "privacy enhancing technologies" or "privacy-preserving technologies." We use the term "privacy substitutes" for clarity and precision.

[2] *See, e.g.*, *Balancing Privacy and Innovation: Does the President's Proposal Tip the Scale?: Hearing Before the Subcomm. on Commerce, Mfg., & Trade of the H. Comm. on Energy & Commerce* , 112th Cong. 4 (2012) (statement of the Hon. Mary Bono Mack, Chairman, Subcomm. on Commerce, Mfg., & Trade) ("When it comes to the Internet, how do we—as Congress, as the administration, and as Americans—balance the need to remain innovative with the need to protect privacy?"), *available at* http://www.gpo.gov/fdsys/pkg/CHRG-112hhrg81441/pdf/CHRG-112hhrg81441.pdf; Fed. Trade Comm'n, Protecting Consumer Privacy in an Era of Rapid Change 36 (2012) ("Establishing consumer choice as a baseline requirement for companies that collect and use consumer data, while also identifying certain practices where choice is unnecessary, is an appropriately balanced model."), *available at* http://ftc.gov/os/2012/03/120326privacyreport.pdf.

[3] Recent scholarship has challenged the efficacy of current notice and choice models for technology privacy. *See, e.g.*, Pedro Giovanni Leon et al., *What Do Online Behavioral Advertising Privacy Disclosures Communicate to Users?*, Proc. 2012 Assoc. for Computing Mach. Workshop on Privacy in the Electronic Soc'y 19, 19 (2012); *see also* Yang Wang et al., *Privacy Nudges for Social Media: An Exploratory Facebook Study*, Proc. 22d Int'l Conf. on World Wide Web 763, 763 (2012).

[4] We depict notice and choice as a straight line since, in many implementations, consumers are given solely binary decisions about whether to accept or reject a set of services or product features. The diagrams in this Essay attempt to illustrate our thinking; they are not intended to precisely reflect any particular privacy issue.

[5] This includes speed, accuracy, usability, cost, technical difficulty, security, and more.

[6] *See* Jonathan R. Mayer & John C. Mitchell, *Third-Party Web Tracking: Policy and Technology*, Proc. 2012 IEEE Symp. on Security & Privacy 413, 415 (2012),*available at* https://cyberlaw.stanford.edu/files/publication/files/trackingsurvey12.pdf.

[7] *E.g.*, *id.* at 416-17.

[8] *E.g.*, Michael Backes et al., *ObliviAd: Provably Secure and Practical Online Behavioral Advertising*, Proc. 2012 IEEE Symp. on Security & Privacy 257, 258 (2012); Matthew Fredrikson & Benjamin Livshits, *RePriv: Re-Imagining Content Personalization and In-Browser Privacy*, Proc.

2011 IEEE SYMP. ON SECURITY & PRIVACY 131, 131 (2011); Saikat Guha et al., *Privad: Practical Privacy in Online Advertising*, Proc. 8th USENIX SYMP. ON NETWORKED SYS. DESIGN & IMPLEMENTATION 169, 170 (2011); Vincent Toubiana et al., *Adnostic: Privacy Preserving Targeted Advertising*, PROC. 17TH NETWORK & Distributed Sys. Symp. 1, 2 (2010); Mikhail Bilenko et al., *Targeted, Not Tracked: Client-Side Solutions for Privacy-Friendly Behavioral Advertising* 13-14 (Sept. 25, 2011) (unpublished manuscript), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1995127; Jonathan Mayer & Arvind Narayanan, *Tracking Not Required: Advertising Measurement*, WEB POLICY (July 24, 2012), http://webpolicy.org/2012/07/24/tracking-not-required-advertising-measurement; Arvind Narayanan et al., *Tracking Not Required: Behavioral Targeting*, 33 BITS OF ENTROPY (June 11, 2012, 2:42 PM), http://33bits.org/2012/06/11/tracking-not-required-behavioral-targeting; Jonathan Mayer & Arvind Narayanan, *Tracking Not Required: Frequency Capping*, WEB POLICY (Apr. 23, 2012), http://webpolicy.org/2012/04/23/tracking-not-required-frequency-capping.

[9] *See Why Persona?*, MOZILLA DEVELOPER NETWORK (May 10, 2013, 3:02 PM), https://developer.mozilla.org/en-US/docs/Mozilla/Persona/Why_Persona.

[10] *See, e.g.*, Loek Essers, *Android NFC Hack Enables Travelers to Ride Subways for Free, Researchers Say*, COMPUTERWORLD (Sept. 20, 2012, 12:29 PM),https://www.computerworld.com/s/article/9231500/Android_NFC_hack_enables_travelers_to_ride_subways_for_free_researchers_say.

[11] Secure multiparty computation has been implemented in various well-known protocols. The area traces its roots to Andrew Yao's "garbled circuit construction," a piece of "crypto magic" dating to the early 1980s. Researchers have used secure multiparty computation to demonstrate privacy-preserving designs in myriad domains—voting, electronic health systems and personal genetics, and location-based services, to name just a few. The payment model we suggest is based on David Chaum's "e-cash." His company DigiCash offered essentially such a system (not just for transportation, but for all sorts of payments) in the 1990s, but it went out of business by 1998. *See generally How DigiCash Blew Everything*, Next Mag., Jan. 1999, *available at*http://cryptome.org/jya/digicrash.htm.

[12] *See generally* Cynthia Dwork, *Differential Privacy: A Survey of Results*, PROC. 5TH INT'L CONF. ON THEORY & APPLICATIONS MODELS COMPUTATION 1, 2 (2008).

[13] Most production systems have data errors, in fact, but they are subtle and underappreciated. Differential privacy is ordinarily a matter of kind and degree of error, not whether error exists at all.

[14] *See, e.g.*, Mayer & Mitchell, *supra* note 6 at 415-16. Some of these misstatements may, of course, reflect intentional downplaying of privacy risks for strategic advantage in public and policy debates.

[15] In theory, uniform privacy-signaling mechanisms or trust intermediaries might assist in informing users. In practice, both approaches have had limited value. *See, e.g.*, Benjamin Edelman, *Adverse Selection in Online "Trust" Certifications and Search Results*, 10 ELECTRONIC COM. RES. & APPLICATIONS 17 (2011) (studying efficacy of website certification providers); Adrienne Porter Felt et al., *Android Permissions: User Attention, Comprehension, and Behavior*, PROC. 8TH SYMP. ON USABLE PRIVACY & SECURITY 2 (2012) (exploring usability of the Android device permissions model).

[16] We have observed firsthand the difficulty imposed by organizational divides in the World Wide Web Consortium's process to standardize Do Not Track. Participants from the online advertising industry have largely been unable to engage on privacy substitutes owing to limited technical expertise, distortions in information relayed to technical staff, and inability to facilitate a direct dialog between inside and outside technical experts.

[17] Sometimes a rigorously vetted privacy substitute will be ready for deployment. Frequently, to be sure, the space of privacy substitutes will include gaps and ambiguities. But policymakers are no strangers to decisions under uncertainty and relying on the best available science.

[18] We caution against requiring particular technical designs. In the future, better designs may become available, or deficiencies in present designs may be uncovered. Cast in more traditional terms of regulatory discourse, this is very much an area for targeting ends, not means.

[19] *See supra* note 3.

[20] The congressional response to Transportation Security Administration full-body scanners might be considered an instance of a privacy substitute. Congress allowed the TSA to retain the scanners, but required a software update that eliminated intrusive imaging. 49 U.S.C. § 44901(*l*) (2011).

[21] For example, some technology companies are lobbying for European Union law to exempt pseudonymous data from privacy protections. *See* CTR. FOR DEMOCRACY & TECH., CDT POSITION PAPER ON THE TREATMENT OF PSEUDONYMOUS DATA UNDER THE PROPOSED DATA PROTECTION REGULATION (2013), *available at*https://www.cdt.org/files/pdfs/CDT-Pseudonymous-Data-DPR.pdf. Information privacy researchers have, however, long recognized that pseudonymous data can often be linked to an individual. *See, e.g.*, Mayer & Mitchell, *supra* note 6, at 415-16.

REVISITING THE 2000 STANFORD SYMPOSIUM IN LIGHT OF BIG DATA

*William McGeveran**

On February 6, 2000, mere weeks into the 21st Century, a collection of the brightest minds considering the regulation of the digital world gathered at Stanford Law School to discuss a cutting-edge question: *Cyberspace and Privacy: A New Legal Paradigm?* Soon after, I purchased a copy of the *Stanford Law Revi*ew containing the writing that emerged from that symposium.[1] (How quaint! A bound volume, made of ink and paper!) Today this remarkable collection remains one of the most consulted books in my collection, printed or digital. Even that early in the internet era, the authors of those articles had already identified the outlines of the crucial issues that continue to occupy us today. (And, indeed, continue to occupy *them*, since almost all remain among the leading scholars specializing in internet-related topics).

Thirteen years later, questions about the emergence of a "new paradigm" often relate to "Big Data" methodologies – the analysis of huge data sets to search for informative patterns that might not have been derived from traditional hypothesis-driven research. Big Data burst into general public consciousness within the last year, and so did its implications for privacy. But the core practices of Big Data go back to 2000 and earlier, albeit at scales not quite as Big. By 2000, Google had already refined its search algorithm by analyzing huge numbers of users' queries. Transportation engineers already planned road improvements by running simulations based on numerous observations of real traffic patterns. Epidemiological research already relied on mass quantities of patient data, including both health and demographic information. And, as demonstrated by Michael Froomkin's inventory of "privacy-destroying technologies" in the 2000 Symposium, we were already experiencing massive data collection and inevitable subsequent processing.[2]

Today's Symposium, cosponsored by Stanford once more, asks whether Big Data represents something entirely new for privacy. Well, leafing through the pages of the 2000 Stanford Symposium, one encounters all the same debates that are arising now in the context of Big Data – perhaps with a few twists, but still quite familiar. This brief essay offers some examples.

I have now heard a number of smart people suggest that treating personal information as a species of property would address many concerns about Big Data. After all, the insights gleaned from Big Data analysis are valuable. They think propertization would require those analyzing data to internalize privacy costs generated by their processing, give individuals leverage, or ensure that resulting windfalls are shared with the people whose information contributed to the profit. We have had this argument before. At the time of the 2000 Symposium, Pamela Samuelson aptly critiqued a portion of the privacy debate as "a quasi-religious war to resolve whether a person's interest in her personal data is a fundamental civil liberty or commodity interest."[3] Up to that point many commentators had similarly suggested that conceiving of personal information as one's property would be an attractive way to secure privacy. There is an initial attraction to the idea. But at the 2000 Symposium and soon thereafter, a growing

* Associate Professor, Vance Opperman Research Scholar, University of Minnesota Law School.

scholarly consensus joined Samuelson in expressing great skepticism about that notion. [4]

Mixing property concepts with privacy concepts brought up doctrinal complications. To begin with, IP regimes such as copyright exist to encourage broad distribution of the underlying content, the very opposite purpose of privacy rules intended to limit the audience for information.[5] Further, complex adjustments to preserve speech interests and the public domain overwhelmed the simplicity of the property model.[6]

At a deeper theoretical level, it wasn't terribly clear what a property rationale really accomplished. The "quasi-religious" dispute often turned on framing without affecting substance. Certainly, as Julie Cohen pointed out in the 2000 Symposium and in much of her later work, the rhetoric of ownership has an effect. If we talk about Big Data organizations "buying" personal information from the willing sellers depicted by that information, we will enshrine assumptions about consent, knowledge, and utility that merit closer inspection.[7] But as a matter of legal design, merely calling an entitlement "property" does not make it any stronger. If the data subject can bargain the right away, all that really matters is the structure of that interaction – default rules, disclosure obligations, imputed duties. Regimes such as the European Union's data protection directive or the HIPAA privacy rules impose significant privacy obligations on data processing without calling the resulting individual rights "property." If I own my data but can sell it to a data miner (Big or Small) by clicking an "I agree" button at site registration, then what difference does that ownership make on the ground? I encourage those who would turn to ownership as the silver-bullet response to Big Data to read those 2000 Symposium articles first.

Another renewed debate that was already in full cry at the 2000 Symposium relates to technological protections. Big Data is made possible by rapid advances in computational power and digital storage capacity. Why not, smart people now ask, use these same features to ensure that downstream Big Data entities respect individuals' preferences about the use of their data? Ideas like persistent tagging of data with expiration dates or use restrictions are in vogue. Internet scholars such as Viktor Mayer-Schönberger and Jonathan Zittrain emphasize the importance of curtailing data permanence through a variety of measures including technological ones.[8] And developments like the European Union's deliberation over a "right to be forgotten" and California's "shine the light" law might create incentives to design Big Data mechanisms that allow individuals to inspect the personal data entities hold about them, and to delete it if they withdraw their consent for processing.

Unlike the propertization strategy, I think this approach has some potential merit, if it is backed by legal rules ensuring adoption and compliance. But nothing about Big Data makes any of these new concepts. Zittrain certainly recognizes this, because he was one of several speakers at the Symposium debating the potential of "trusted systems" to embed privacy protection in the architecture of data systems.[9] And Lawrence Lessig's notion that "code is law" was a centerpiece of the debate by 2000.[10] Proposals for trusted intermediaries or data brokers who handled information with a duty to protect the data subject's privacy interests were already in wide circulation by 2000 as well. These types of techno-architectural responses should be guided by history, such as the failure of P3P and the very slow uptake for other privacy-enhancing technologies, all discussed in the 2000 Symposium. As we already knew in 2000, technology can contribute greatly to addressing privacy problems, but cannot solve them on its own.

A third argument that has flared up with renewed vigor, fueled by Big Data, asks how much speech-related protection might apply to processing of data.[11] This discussion relates to new regulatory proposals, particularly those that advocate increased control at the processing and storage phases of data handling. These rules, it is said, contrast with the collection-focused rules that now dominate privacy law, especially in the US.

Once again, the seminal work was already happening in the 2000 Symposium. In his contribution, Eugene Volokh memorably characterized much of privacy law as "a right to stop people from speaking about you."[12] Others

in the Symposium took up both sides of the argument.[13] The speech aspects of Big Data activities resemble very much the speech aspects of past data mining activities. While downstream regulation may be more attractive, there is still no real sea change in the dissemination of personal information. Neither its larger scale nor its lack of hypothesis should influence application of First Amendment principles to Big Data. There is no more *speaking* in Big Data than there was in Medium-Sized Data, circa 2000.

Finally, some discussion of Big Data emphasizes that, by its nature, the subsequent processing of information is unpredictable. Smart people wonder what this means for the consent that was offered at the time of initial collection. If the purposes for which data would be used later could not be specified then, could there be true consent from the data subject? In the European Union, the answer to this question has long been: no. But for a long time now, the U.S. has embraced an increasingly farcical legal fiction that detailed disclosures to data subjects generated true informed consent. The empirical silliness of this notion was brought home by a recent study calculating that it would take the average person 76 work days to read every privacy policy that applied to her.[14]

Yet again, however, the 2000 Symposium already understood the disconnection between the complexities of data collection and processing and the cognitive abilities of an individual site user to offer meaningful consent.[15] Froomkin explained the economics of "privacy myopia," under which a consumer is unable to perceive the slow aggregation of information in a profile, and therefore its true privacy costs.[16] If Big Data processing might be even more remote, then it might induce even more myopia, but we would have the tools to analyze it from the 2000 Symposium.[17]

Each of these four debates – propertization, technological measures, speech protection, and privacy myopia – takes on new salience because of Big Data. But they are not fundamentally different from the brilliant deliberations at the 2000 Symposium. To see how they apply today one must substitute the names of some companies and update some technological assumptions. But these cosmetic changes don't compromise their theoretical core.

In the end, what is different about Big Data? Basically, that it is Big. The scale of information collected and processed is considerably greater. In addition, the ability to draw inferences from data has become steadily more sophisticated. So there is more data and it is more useful. But by 2000 we already surrendered vast quantities of personal information in our everyday life. It was already mined assiduously in search of insights both aggregate and personalized. We were already worried about all that, and already considering how to respond. I don't mean to suggest that the development of Big Data isn't important. I only emphasize that the ways to think about it, and the policy debates that it generates, have been around for a long time. The 2000 Symposium remains highly relevant today – and that kind of longevity itself proves the enduring value of the best privacy scholarship.

---

[1] Symposium, *Cyberspace and Privacy: A New Legal Paradigm?*, 52 STAN. L. REV. 987 (2000).

[2] A. Michael Froomkin, *The Death of Privacy?*, 52 STAN. L. REV. 1461, 1468-1501 (2000).

[3] Pamela Samuelson, *Privacy as Intellectual Property?*, 52 STAN. L. REV. 1125, 1157-58 (2000).

[4] *See* Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 STAN. L. REV. 1373 (2000); Jerry Kang & Benedikt Buchner, *Privacy in Atlantis*, 18 HARV. J. L. & TECH. 229 (2004); Mark A. Lemley, *Private Property*, 52 STAN. L. REV. 1545 (2000); Jessica Litman, *Information Privacy/Information Property*, 52 STAN. L. REV. 1283 (2000); Samuelson, supra note 3; Paul M. Schwartz, *Internet Privacy and the State*, 32 CONN. L. REV. 815 (2000); Jonathan Zittrain, *What the Publisher Can Teach the Patient: Intellectual Property and Privacy in an Age of Trusted Privication*, 52 STAN. L. REV. 1201 (2000); *see also* William McGeveran, *Programmed Privacy Promises: P3P and Web Privacy Law*, 76 N.Y.U. L. REV. 1812, 1834-45 (2001) (applying this reasoning to the once-promising privacy-enhancing technology known as P3P).

[5] *See, e.g.*, Litman, supra note 4, at 1295-96.

[6] *See, e.g.*, Lemley, supra note 4, at 1548-50.

[7] *See, e.g.*, JULIE E. COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE (2012); Cohen, supra note 4.

[8] *See* VIKTOR MAYER-SCHÖNBERGER, DELETE: THE VIRTUE OF FORGETTING IN THE DIGITAL AGE (2011); JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET--AND HOW TO STOP IT (2009).

[9] *See* Zittrain, supra note 4; *see also* Henry T. Greely, *Trusted Systems and Medical Records: Lowering*

Expectations, 52 STAN. L. REV. 1585 (2000); Jonathan Weinberg, *Hardware-Based ID, Rights Management, and Trusted Systems*, 52 STAN. L. REV. 1251 (2000).

[10] LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999). For another important rendition of this argument, see Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553, 570 (1998). *See also* Jay P. Kesan and Rajiv C. Shah, *Shaping Code*, 18 HARV. J. L. & TECH. 319 (2005) (reviewing history of code-backed restrictions).

[11] *See, e.g.*, Jane Bambauer, *Is Data Speech?*, 66 STAN. L. REV. __ (forthcoming 2013); Neil M. Richards, *Data Privacy and the Right to be Forgotten after Sorrell* (working paper, June 6, 2013, on file with author).

[12] Eugene Volokh, *Freedom of Speech and Information Privacy: the Troubling Implications of a Right to Stop People From Speaking About You*, 52 STAN. L. REV. 1049 (2000).

[13] Compare Richard Epstein, *Privacy, Publication, and the First Amendment: The Dangers of First Amendment Exceptionalism*, 52 STAN. L. REV. 1003 (2000) with Cohen, supra note 4; Paul M. Schwartz, *Free Speech vs. Information Privacy: Eugene Volokh's First Amendment Jurisprudence*, 52 STAN. L. REV. 1559 (2000).

[14] *See* Aleecia M. McDonald & Lorrie Faith Cranor, *The Costs of Reading Privacy Policies*, 4 I/S J. OF L. & POLICY 540 (2008); Alexis C. Madrigal, *Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days*, THEATLANTIC.COM (March 1, 2012) at http://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851/.

[15] *See* Cohen, supra note 4; Froomkin, supra note 2.

[16] Froomkin, supra note 2, at 1501-05; *see also* Neil Weinstock Netanel, *Cyberspace Self-Governance: A Skeptical View From Liberal Democratic Theory*, 88 CAL. L. REV. 395, 476-77 (2000) (discussing ignorance of data aggregation).

[17] Granted, Big Data may result in more decisions and assumptions about an individual to her detriment – such as price discrimination or insurance underwriting. If so, then most likely those decision processes ought to be regulated in themselves, through mechanisms modeled on the Fair Credit Reporting Act, 15 U.S.C. § 1681 et seq., or the Genetic Information Nondiscrimination Act, Pub. L. 110–233 (2008).

POLICY FRAMEWORKS TO ENABLE BIG HEALTH DATA

*Deven McGraw**

The latest hot topic in technology is "big data," with nearly everyone, from corporate executives to academics to policy pundits, espousing its transformative power to help address a broad range of societal challenges.

Big data is a particular focus of health care policy conversations. We know already (from data, of course) that the U.S. health care system is plagued by overall poor outcomes, distressing racial and ethnic health disparities, alarming safety problems, and unsustainable costs. But we know far less about how to effectively address those issues. Big data is seen as crucial in reversing those trends.[1]

The term big data is used in many ways; for some, big data actually must be "big" in terms of the size of the database.[2] In health care, the term is frequently used to refer to analytics of health data across multiple health care databases, such as physician and hospital medical records and insurance claims databases. Whether the database qualifies as "big" by technology standards is less important than having sufficient information to draw scientifically valid conclusions, which depends on the question posed. Health big data is an important component of the "learning health care system," which refers to leveraging health information to improve the knowledge base about effective prevention and treatment strategies, and to disseminate that knowledge more rapidly to improve the quality and efficiency of health care.[3]

There is broad support for leveraging health data for learning purposes. At the same time, concerns have been raised about whether current laws governing "learning" uses of health data are up to the task. Some have questioned whether those laws are sufficiently protective of patient privacy, while others have charged that the rules erect unnecessary and costly barriers.[4] Surveys show that a majority of the public supports research uses of data; at the same time, individuals consistently express concerns about the privacy of their medical information. [5] At the end of the day, the public must trust a vibrant and sustainable ecosystem for health big data.

Are today's debates about health big data just a rehash of old (and still ongoing) debates about policies governing research uses of health information? The issues surrounding privacy protections for research uses of data are not new. But health big data is more than just a new term for an old problem. The health data environment today is vastly different and will likely change more rapidly in the near future. Until recently, researchers frequently needed to extract clinical data from paper files in order to do analytics. In addition, health data research was customarily done by sophisticated academic medical and research centers, health plans (whose claims data has been digital for well over a decade) and well-resourced data mining companies. But the Health Information Technology for Economic and Clinical Health Act of 2009 (HITECH)[6] is changing that dynamic. Among those providers eligible for HITECH's electronic medical record incentive program, more than half of clinicians and 80 percent of hospitals are now capturing clinical data using EMRs.[7] More clinical data is available in digital (and in some cases, standardized) form, due to existing sources of research data going digital and an increase in potential sources of clinical research data.

In addition, digital health data is no longer collected only by traditional health system entities like health care providers and health insurers. Consumers are increasingly collecting

* Deven McGraw, Director, Health Privacy Project, Center for Democracy & Technology.

and sharing data on health and wellness using personal health record tools, mobile health applications, and social networking sites. Individuals also can leave digital health footprints when they conduct online searches for health information. The health data shared by consumers using such tools can range from detailed clinical information, such as downloads from an implantable device and details about medication regimens, to data about weight, caloric intake, and exercise logs.

The worlds of clinical and administrative claims data and consumer health data today are largely separate silos, but efforts to learn from combined health datasets are increasing. Beginning in 2014, patients will begin to have direct access to their clinical health information through portals to their provider's electronic medical records.[8] Such access will include the capability to download this data into tools of the patient's choosing, and to directly transmit this data to other entities.[9] Policymakers are also in discussion about requiring providers participating the HITECH incentive program to incorporate electronic data generated by patients into their clinical workflows.[10]

Building and maintaining public trust in a broader, robust, health big data ecosystem will require the development and implementation of comprehensive, adaptable policy and technology frameworks. Such frameworks should:

- provide protections for health data while still enabling analytics to solve pressing health challenges;
- apply consistently to health data regardless of the type of entity collecting it (be it a hospital or a commercial health app) and yet still be flexible enough to respond to the particular risks to privacy posed by different health data sharing models;
- include mechanisms to hold entities collecting and analyzing health data accountable for complying with rules and best practices;
- provide incentives for the adoption of privacy-enhancing technical architectures/models for collecting and sharing data; and
- be based on thoughtful application of the Fair Information Practice Principles (FIPPs), which have been the foundation for privacy laws and industry best practices both in the U.S. and internationally. Although there are many articulations of the FIPPs, the Markle Foundation led a multi-stakeholder effort to create a version tailored to health data that could be the starting point for health big data frameworks.[11]

Efforts to develop more effective ethical and legal frameworks for learning uses of health data have already begun, although they have largely been focused on a re-consideration of existing policies governing the traditional health care system. For example, in a Hastings Center special report, *Ethical Oversight of Learning Health Care Systems*, renowned bioethicists challenged the traditional treatment of research uses of data as inherently more risky for patients than treatment and called for a new ethics framework for the learning health care system more expressly acknowledges contributing to learning from data as an ethical obligation of both providers and patients.[12] CDT also has participated in discussions regarding the policy and technology needs of a learning health care system sponsored by the Institute of Medicine, the Clinical Trials Transformation Initiative, AcademyHealth, eHealth Initiative, the Bipartisan Policy Center, and the American Medical Informatics Association. Given broad support for leveraging health data to reform health care, there are likely other important conversations taking place on these issues.

However, these efforts have not yet yielded consensus on how to address health big data issues and are not focusing (at least not yet) on developing frameworks that also could apply to health data outside of the traditional healthcare ecosystem. CDT is beginning work to develop a policy and technology framework for health big data. We hope to be a catalyst for merged conversations about consistent policies for health data analytics regardless of the type of entity collecting, sharing and analyzing the data. Although this work is in the early stages, we offer the following to jumpstart a richer dialogue.

- Some consistency for policies governing health big data is desirable, to create predictability and reduce uncertainty. But desire for consistency need not (and probably should not) yield the exact same rules for all circumstances. For example, appropriate policies governing how doctors and hospitals use health data for learning purposes will likely vary from those for a social networking site. As noted above, the FIPPs ideally should be applied thoughtfully, considering the context of a particular health data use (or uses) and the potential "risk" to patients.[13] With respect to health data, the concept of risk needs a broad frame, beyond the typical tangible harms like loss of employment or insurance discrimination and encompassing risks like stereotyping, harms to dignity and harms to trust in the historic confidentiality of the clinician-patient relationship.

- The rules under the Health Insurance Portability and Accountability Act (HIPAA) and the Common Rule (which governs federally funded health data research) are the best place to start for refining policy frameworks for big data uses in the traditional health care system. However, those two legal regimes should be more consistent; efforts to do so have been launched but do not appear close to completion.[14] In addition, both regimes rely disproportionately on patient consent to govern learning uses of data; yet consent shifts the burden of protecting privacy to individuals and may be less effective in protecting privacy in big data analytics.[15] A thoughtful application of the FIPPs should include consideration of whether policies that enhance transparency to individuals about big data uses, or that enable more active engagement and input of individuals in the research enterprise, are more effective at building public trust while facilitating health big data analytics.

- Policies governing health data today provide few, if any, incentives to pursue data analytics using privacy-enhancing technical architectures, such as distributed data networks in lieu of centralized collection of copies of data. For example, in the Mini Sentinel Distributed Database, which facilitates safety surveillance on drugs approved by the FDA, participating data sources format their data into a Common Data Model and perform the analytics; aggregate results (not raw data) are reported out and collectively produce an answer to the research question (sometimes referred to as "bringing the questions to the data").[16] Other models include pushing data to a dedicated edge server, enabling analytics to be performed without releasing the raw data (a model that works particularly well for data sources without the expertise to perform the analytics). The technical model used for research should address the particular analytic needs, so there is no "one size fits all." Nevertheless, incentives to use privacy-enhancing technical architectures (of which there are more than the two examples listed here) should be part of the discussion.

- It's not clear there are sufficient incentives to pursue big data analytics that address the nation's most pressing health care priorities. Within the traditional health care system, rules governing learning uses of health data permit such uses but do not require entities to undertake them. Consequently, entities that engage in research are those whose missions expressly incorporate research and/or who are receiving some financial support for it. With respect to health data collected in the consumer-facing or commercial space, business imperatives likely will drive big data uses. We need additional debate regarding how to provide incentives for big data uses that benefit the public. Of course, the form of those incentives will need to be carefully considered within

the context of creating a trust framework for big data uses.

- Policies around de-identification of health data also need reconsideration. Much of health big data analytics will take place using so-called "de-identified" data. However, there are no standards for de-identification other than those set forth in HIPAA, and non-covered entities are not required to use them. Questions have been raised about whether one of the methodologies for de-identification, the safe harbor, is sufficiently rigorous; and too few entities use the statistical method, which provides more protection and yields greater data utility.[17] In addition, because de-identification does not eliminate risk of re-identification, protections are still needed for the residual re-identification and other privacy risks that remain in the data.[18]

The promise of health big data is clear but will not be realized without the trust of the public. Now is the time to accelerate the hard work of developing the technology and policy frameworks that will achieve that trust.

---

[1] *See, e.g.,* Lorraine Fernandes et al., *Big Data, Bigger Outcomes,* 83 J. OF AM. HEALTH INFO. MGMT. ASS'N 38 (2012).

[2] *Big Data*, WIKIPEDIA, http://en.wikipedia.org/wiki/Big_data (last visited June 30, 2013).

[3] NATIONAL RESEARCH COUNCIL. *THE LEARNING HEALTHCARE SYSTEM: WORKSHOP SUMMARY (IOM ROUNDTABLE ON EVIDENCE-BASED MEDICINE) (2007)*.

[4] *See, e.g.,* Sharyl J. Nass et al., *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research* (Institute of Medicine 2009); Robert E. Litan, *Big Data Can Save Health Care – But at What Cost to Privacy?,* THE ATLANTIC (May 25, 2012); Steve Lohr, *Big Data is Opening Doors, but Maybe Too Many,* N.Y. TIMES (March 23, 2013).

[5] Sally Okun, et al., *Making the Case for Continuous Learning from Routinely Collected Data*, Discussion Paper (Inst. of Med.) (April 15, 2013), http://www.iom.edu/Global/Perspectives/2013/MakingtheCaseforContinuousLearning.aspx (accessed June 29, 2013).

[6] 42 U.S.C. § 300jj et seq. (2012).

[7] Press Release, U.S. Department of Health and Human Services, Doctors' and Hospitals' Use of Health IT More than Doubles in 2012, (May 22, 2013) http://www.hhs.gov/news/press/2013pres/05/20130522a.html (accessed June 29, 2013).

[8] Centers for Medicare and Medicaid Services, Stage 2 Overview Tipsheet 4 (last updated August 2012), https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Stage2Overview_Tipsheet.pdf, (accessed June 29, 2013).

[9] *Id.*

[10] Helen R. Pfister & Susan R. Ingargiola, *Meaningful Use Stage 3 Could Include Patient-Generated Health Data,* iHealthBeat (December 11, 2012), http://www.ihealthbeat.org/features/2012/meaningful-use-stage-3-could-include-patientgenerated-health-data.aspx (accessed June 29, 2013).

[11] *See generally* Markle Foundation, Markle Common Framework, http://www.markle.org/health/markle-common-framework (accessed June 29, 2013).

[12] *See* Nancy E. Kass et al., *The Research-Treatment Distinction: A Problematic Approach for Determining Which Activities Should Have Ethical Oversight*, 43 HASTINGS CENTER REP. 4 (2013); Ruth R. Faden et al., *An Ethics Framework for a Learning Health Care System: A Departure from Traditional Research Ethics and Clinical Ethics,* 43 HASTINGS CENTER REP. 16 (2013).

[13] The work to explore use- and risk-based frameworks for big data has already begun. *See, e.g.,* Fred H. Cate & Viktor Mayer-Schönberger, *Notice and Consent in a World of Big Data,* 3 INT'L DATA PRIVACY L. 67 (2013).

[14] In July 2011, the U.S. Department of Health & Human Services (HHS) released an Advance Noticed of Proposed Rulemaking, seeking comment on potential changes to the Common Rule, some of which were designed to make HIPAA and that Rule more consistent. 76 Fed. Reg. 44512–44531 (July 26, 2011). HHS has published nothing further to advance that effort. In January 2013, HHS amended HIPAA's research provisions to make it easier to obtain patient authorization for research uses of data, but it's not clear whether these changes will be reflected in any future revisions to the Common Rule. 78 Fed. Reg. 5566-5702 (January 25, 2013).

[15] *See* Cate & Mayer-Schonburger, *supra* note 13; Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma,* 126 Harv. L. Rev. 1880 (2013); Helen Nissenbaum, *A Contextual Approach to Privacy Online*, 140 Daedalus 32 (2011).

[16] Mini-Sentinel, About Mini-Sentinel, http://www.mini-sentinel.org/about_us/default.aspx (accessed June 29, 2013).

[17] *See generally* KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION (2013).

[18] Deven McGraw, *Building Public Trust in De-Identified Health Data*, 20 J. AM. MED. INFO. ASS'N 704 (2012), *available at* http://jamia.bmj.com/content/early/2012/06/25/amiajnl-2012-000936.full.html.

IT'S NOT PRIVACY, AND IT'S NOT FAIR

*Deirdre K. Mulligan & Cynthia Dwork**

66 STAN. L. REV. ONLINE 35

Classification is the foundation of targeting and tailoring information and experiences to individuals. Big data promises—or threatens—to bring classification to an increasing range of human activity. While many companies and government agencies foster an illusion that classification is (or should be) an area of absolute algorithmic rule—that decisions are neutral, organic, and even automatically rendered without human intervention—reality is a far messier mix of technical and human curating. Both the datasets and the algorithms reflect choices, among others, about data, connections, inferences, interpretation, and thresholds for inclusion that advance a specific purpose. Like maps that represent the physical environment in varied ways to serve different needs—mountaineering, sightseeing, or shopping—classification systems are neither neutral nor objective, but are biased toward their purposes. They reflect the explicit and implicit values of their designers. Few designers "see them as artifacts embodying moral and aesthetic choices" or recognize the powerful role they play in crafting "people's identities, aspirations, and dignity."[1] But increasingly, the subjects of classification, as well as regulators, do.

Today, the creation and consequences of some classification systems, from determination of tax-exempt status to predictive analytics in health insurance, from targeting for surveillance to systems for online behavioral advertising (OBA), are under scrutiny by consumer and data protection regulators, advocacy organizations and even Congress. Every step in the big data pipeline is raising concerns: the privacy implications of amassing, connecting, and using personal information, the implicit and explicit biases embedded in both datasets and algorithms, and the individual and societal consequences of the resulting classifications and segmentation. Although the concerns are wide ranging and complex, the discussion and proposed solutions often loop back to privacy and transparency—specifically, establishing individual control over personal information, and requiring entities to provide some transparency into personal profiles and algorithms.[2]

The computer science community, while acknowledging concerns about discrimination, tends to position privacy as the dominant concern.[3] Privacy-preserving advertising schemes support the view that tracking, auctioning, and optimizing done by the many parties in the advertising ecosystem are acceptable, as long as these parties don't "know" the identity of the target.[4]

Policy proposals are similarly narrow. They include regulations requiring consent prior to tracking individuals or prior to the collection of "sensitive information," and context-specific codes respecting privacy expectations.[5] Bridging the technical and policy arenas, the World Wide Web Consortium's draft "do-not-track" specification will allow users to signal a desire to avoid OBA.[6] These approaches involve greater transparency.

Regrettably, privacy controls and increased transparency fail to address concerns with the classifications and segmentation produced by big data analysis.

* Deirdre K. Mulligan is Assistant Professor of School of Information, Berkeley Law; Co-Director, Berkeley Center for Law and Technology. Cynthia Dwork is Distinguished Scientist, Microsoft Research.

At best, solutions that vest individuals with control over personal data indirectly impact the fairness of classifications and outcomes—resulting in discrimination in the narrow legal sense, or "cumulative disadvantage" fed by the narrowing of possibilities.[7] Whether the information used for classification is obtained with or without permission is unrelated to the production of disadvantage or discrimination. Control-based solutions are a similarly poor response to concerns about the social fragmentation of "filter bubbles"[8] that create feedback loops reaffirming and narrowing individuals' worldviews, as these concerns exist regardless of whether such bubbles are freely chosen, imposed through classification, or, as is often the case, some mix of the two.

At worst, privacy solutions can hinder efforts to identify classifications that unintentionally produce objectionable outcomes—for example, differential treatment that tracks race or gender—by limiting the availability of data about such attributes. For example, a system that determined whether to offer individuals a discount on a purchase based on a seemingly innocuous array of variables being positive ("shops for free weights and men's shirts") would in fact routinely offer discounts to men but not women. To avoid unintentionally encoding such an outcome, one would need to know that men and women arrayed differently along this set of dimensions. Protecting against this sort of discriminatory impact is advanced by data about legally protected statuses, since the ability to both build systems to avoid it and detect systems that encode it turns on statistics.[9] While automated decisionmaking systems "may reduce the impact of biased individuals, they may also normalize the far more massive impacts of system-level biases and blind spots."[10] Rooting out biases and blind spots in big data depends on our ability to constrain, understand, and test the systems that use such data to shape information, experiences, and opportunities. This requires more data.

Exposing the datasets and algorithms of big data analysis to scrutiny—transparency solutions—may improve individual comprehension, but given the independent (sometimes intended) complexity of algorithms, it is unreasonable to expect transparency alone to root out bias.

The decreased exposure to differing perspectives, reduced individual autonomy, and loss of serendipity that all result from classifications that shackle users to profiles used to frame their "relevant" experience, are not privacy problems. While targeting, narrowcasting, and segmentation of media and advertising, including political advertising, are fueled by personal data, they don't depend on it. Individuals often create their own bubbles. Merely *allowing* individuals to peel back their bubbles—to view the Web from someone else's perspective, devoid of personalization—does not guarantee that they will.[11]

Solutions to these problems are among the hardest to conceptualize, in part because perfecting individual choice may impair other socially desirable outcomes. Fragmentation, regardless of whether its impact can be viewed as disadvantageous from any individual's or group's perspective, and whether it is chosen or imposed, corrodes the public debate considered essential to a functioning democracy.

If privacy and transparency are not the panacea to the risks posed by big data, what is?

First, we must carefully unpack and model the problems attributed to big data.[12] The ease with which policy and technical proposals revert to solutions focused on individual control over personal information reflects a failure to accurately conceptualize other concerns. While proposed solutions are responsive to a subset of privacy concerns—we discuss other concepts of privacy at risk in big data in a separate paper—they offer a mixed bag with respect to discrimination, and are not responsive to concerns about the ills that segmentation portends for the public sphere.

Second, we must approach big data as a sociotechnical system. The law's view of automated decisionmaking systems is schizophrenic, at times viewing automated decisionmaking with suspicion and distrust and at others exalting it as the antidote to the discriminatory urges and intuitions of people.[13] Viewing the problem as one of

machine versus man misses the point. The key lies in thinking about how best to manage the risks to the values at stake in a sociotechnical system.[14] Questions of oversight and accountability should inform the decision of where to locate values. Code presents challenges to oversight, but policies amenable to formal description can be built in and tested for. The same cannot be said of the brain. Our point is simply that big data debates are ultimately about values first, and about math and machines only second.

Third, lawyers and technologists must focus their attention on the risks of segmentation inherent in classification. There is a broad literature on fairness in social choice theory, game theory, economics, and law that can guide such work.[15] Policy solutions found in other areas include the creation of "standard offers"; the use of test files to identify biased outputs based on ostensibly unbiased inputs; required disclosures of systems' categories, classes, inputs, and algorithms; and public participation in the design and review of systems used by governments.

In computer science and statistics, the literature addressing bias in classification comprises: testing for statistical evidence of bias; training unbiased classifiers using biased historical data; a statistical approach to situation testing in historical data; a method for maximizing utility subject to any context-specific notion of fairness; an approach to fair affirmative action; and work on learning fair representations with the goal of enabling fair classification of future, not yet seen, individuals.

Drawing from existing approaches, a system could place the task of constructing a metric—defining who must be treated similarly—outside the system, creating a path for external stakeholders—policymakers, for example—to have greater influence over, and comfort with, the fairness of classifications. Test files could be used to ensure outcomes comport with this predetermined similarity metric. While incomplete, this suggests that there are opportunities to address concerns about discrimination and disadvantage. Combined with greater transparency and individual access rights to data profiles, thoughtful policy, and technical design could tend toward a more complete set of objections.

Finally, the concerns related to fragmentation of the public sphere and "filter bubbles" are a conceptual muddle and an open technical design problem. Issues of selective exposure to media, the absence of serendipity, and yearning for the glue of civic engagement are all relevant. While these objections to classification may seem at odds with "relevance" and personalization, they are not a desire for irrelevance or under-specificity. Rather they reflect a desire for the tumult of traditional public forums—sidewalks, public parks, and street corners—where a measure of randomness and unpredictability yields a mix of discoveries and encounters that contribute to a more informed populace. These objections resonate with calls for "public" or "civic" journalism that seeks to engage "citizens in deliberation and problem-solving, as members of larger, politically involved publics,"[16] rather than catering to consumers narrowly focused on private lives, consumption, and infotainment. Equally important, they reflect the hopes and aspirations we ascribe to algorithms: despite our cynicism and reservations, "we want them to be neutral, we want them to be reliable, we want them to be the effective ways in which we come to know what is most important."[17] We want to harness the power of the hive brain to expand our horizons, not trap us in patterns that perpetuate the basest or narrowest versions of ourselves.

The urge to classify is human. The lever of big data, however, brings ubiquitous classification, demanding greater attention to the values embedded and reflected in classifications, and the roles they play in shaping public and private life.

---

[1] GEOFFREY C. BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES 4 (2000).

[2] *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1308-09 (2008); Lucas D. Introna & Helen Nissenbaum, *Shaping the Web: Why the Politics of Search Engines Matters*, 16 INFO. SOC'Y 169 (2000); Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. ON TELECOMM. & HIGH TECH. L.235 (2011); Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1 (2005).

[3] Vincent Toubiana et al., *Adnostic: Privacy Preserving Targeted Advertising* 1 (17th Annual Network & Distributed Sys. Sec. Symposium Whitepaper, 2010),*available at* http://www.isoc.org/isoc/conferences/ndss/10/pdf/05.pdf ("Some are concerned that OBA is manipulative and discriminatory, but the dominant concern is its implications for privacy.").

[4] Alexey Reznichenko et al., *Auctions in Do-Not-Track Compliant Internet Advertising*, 18 PROC. ACM CONF. ON COMPUTER & COMM. SECURITY 667, 668 (2011) ("The privacy goals . . . are . . . [u]nlinkability: the broker cannot associate . . . information with a single (anonymous) client.").

[5] Multistakeholder Process to Develop Consumer Data Privacy Codes of Conduct, 77 Fed. Reg. 13,098 (Mar. 5, 2012); Council Directive 2009/136, art. 2, 2009 O.J. (L 337) 5 (EC) (amending Council Directive 2002/58, art. 5); FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS 45-46 (2012), available at http://ftc.gov/os/2012/03/12032privacyreport.pdf.

[6] World Wide Web Consortium, *Tracking Preference Expression (DNT), W3C Editor's Draft*, WORLD WIDE WEB CONSORTIUM (June 25, 2013),http://www.w3.org/2011/tracking-protection/drafts/tracking-dnt.html.

[7] Oscar H. Gandy Jr., *Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems*, 12 ETHICS & INFO. TECH.29, 37-39 (2010).

[8] *See generally* ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (2011).

[9] Julie Ringelheim, *Processing Data on Racial or Ethnic Origin for Antidiscrimination Policies: How to Reconcile the Promotion of Equality with the Right to Privacy?*14-15 (Ctr. for Human Rights & Global Justice, Working Paper No. 8/06, 2006) (discussing the use of demographic data to identify disparate impact in "neutral" rules).

[10] Gandy Jr., *supra* note 8, at 33.

[11] *See* Introna & Nissenbaum, *supra* note 2; Pasquale, *supra* note 2.

[12] Recent symposia have begun this process. *E.g.*, Symposium, Transforming the Regulatory Endeavor, 26 BERKELEY TECH. L.J. 1315 (2011); *see also* N.Y. Univ. Steinhardt Sch. of Culture, Educ., & Human Dev., Governing Algorithms: A Conference on Computation, Automation, and Control (May 16-17, 2013),http://governingalgorithms.org.

[13] *See, e.g.*, FED. FIN. INSTS. EXAMINATION COUNCIL, INTERAGENCY FAIR LENDING EXAMINATION PROCEDURES 7-9 (2009).

[14] *See, e.g.*, Roger Brownsword, *Lost in Translation: Legality, Regulatory Margins, and Technological Management*, 26 BERKELEY TECH. L.J. 1321 (2011).

[15] Among the most relevant are theories of fairness and algorithmic approaches to apportionment. *See, e.g.*, the following books: HERVÉ MOULIN, FAIR DIVISION AND COLLECTIVE WELFARE (2003); JOHN RAWLS, A THEORY OF JUSTICE (1971); JOHN E. ROEMER, EQUALITY OF OPPORTUNITY (1998); JOHN E. ROEMER, THEORIES OF DISTRIBUTIVE JUSTICE (1996); H. PEYTON YOUNG, EQUITY: IN THEORY AND PRACTICE (1995). Roemer's approach to equal opportunity embraces (potentially sensitive) information about the individual over which she has no control—genes, family background, culture, social milieu—explicitly taking these into account, in the form of what he calls a "type," when considering how resources should be allocated.

[16] Tanni Haas & Linda Steiner, *Public Journalism: A Reply to Critics*, 7 JOURNALISM 238, 242 (2006).

[17] Tarleton Gillespie, *Can an Algorithm Be Wrong? Twitter Trends, the Specter of Censorship, and Our Faith in the Algorithms Around Us*, Culture Digitally (Oct. 19, 2011), http://culturedigitally.org/2011/10/can-an-algorithm-be-wrong.

SENSOR PRIVACY AS ONE REALISTIC & REASONABLE MEANS TO BEGIN REGULATING BIG DATA

*Scott R. Peppet**

Let us start with a reasonably safe prediction: It is unlikely that the United States will ever enact comprehensive Big Data privacy legislation. Privacy scholars have long lamented the difficulties of enacting *any* comprehensive legislative privacy reform.[1] Beyond that general inertia, Big Data legislation is particularly improbable. Although it is relatively easy to articulate broad principles to control Big Data—such as those in the Obama Administration's Consumer Privacy Bill of Rights—it is hard to imagine how a comprehensive statute would define its scope sufficiently broadly to have impact but not so broadly as to bring every byte of data within its purview. Moreover, the obvious economic value of Big Data means that strong constituents will seek to protect its growth trajectory and limit legislative overreach. Although even ardent proponents of Big Data increasingly acknowledge its privacy implications and seek legal constraints to prevent extreme privacy-violative uses,[2] so far there have been very few concrete proposals in academic work,[3] industry reports,[4] or legislation to regulate Big Data.

This lack of a realistic regulatory agenda is dangerous for both privacy and the Big Data industry. Without some realistic means to constrain Big Data, its proponents' calls for more robust privacy protection will begin to seem unhelpful, at best, or disingenuous, at worst. This risks consumer disengagement and skepticism: as the World Economic Forum recently put it, "the lack of resolution on means of accountability … contributes to a lack of trust throughout the [Big Data] ecosystem."[5] How then, to make real progress on regulating Big Data?

Legislative momentum builds in response to salient, concrete, urgent needs that are easy to understand and act upon. Rather than wait for an unlikely (and potentially unwieldy) comprehensive Big Data law, we should focus on the achievable: implementing data security, data transparency, and data use constraints for sensitive types of information in a localized, sector by sector, input-type by input-type fashion that attends to salient threats caused by particular aspects of the Big Data infrastructure. The key is to regulate uses, not types of data, but in context-specific ways.

I nominate sensor privacy as the first candidate. Concern about Big Data generally focuses on the ways in which inferences can be drawn from the *online* data available about each of us, such as Twitter and Facebook accounts, Google searches, and web surfing patterns. Far more powerful, however, are the new streams of information emanating from the millions of tiny, largely unnoticed, sensors beginning to saturate daily life. Whether in your smart phone, health monitoring bracelet (e.g., FitBit or Nike FuelBand), automobile black box, home or "smart grid" electricity monitor, employee tracking device, or even your baby's Internet-connected and sensor-laden "onesie," sensors are suddenly everywhere.[6] As the cost of such sensors plummeted in the last few years, they have become ubiquitous in consumer products available at scale.[7] Some estimate that by 2025 over one *trillion* consumer and industrial devices will be connected to the Internet or each other.[8]

* Professor of Law, University of Colorado School of Law.

Sensor data are the stuff of Big Data dreams. Unlike information gleaned from online posts, Tweets, or searches, sensor data provide a rich picture of actual behavior, not beliefs or self-projections.[9] Your FitBit shows whether you actually exercise; your Facebook profile shows only that you *say* you exercise. As inputs into Big Data analytic engines, these data are revolutionizing health care, energy efficiency, management productivity analysis, and industrial engineering.

At the same time, sensor data feed Big Data analytics in ways that present serious and particularly pressing privacy risks. Consider three.

First, sensor data are inherently both sensitive and migratory. Sensors may *directly* monitor sensitive information: a health monitor may reveal weight, exercise or eating habits, or stress level, for example.[10] Similarly, electricity sensors—whether as part of a state-wide "smart grid" or a consumer energy-saving device—may show how much time you watch television or how late at night you get home (e.g., just after the local bars typically close). In addition, however, sensors can easily reveal sensitive information by supporting unexpected Big Data inferences. For example, monitoring such electrical signals can also reveal how responsible you are (e.g., by showing whether you leave your children home alone), how forgetful you may be (e.g., by showing whether you leave the oven on while at work), and even your intellectual interests (e.g., research has shown that one can accurately determine exactly what movie someone is watching on television just by monitoring the electrical signals emanating from the person's house).[11]

Most important, sensor data inherently migrate across contexts. Although a consumer may think that an electricity sensor will generate data only to promote energy savings or that a FitBit's biometric information is useful solely for wellness-related purposes, such data could easily help an insurer draw inferences about that consumer to set premiums more accurately (e.g., amount of exercise may influence health or life insurance), aid a lender in assessing the consumer's creditworthiness (e.g., conscientious exercisers may be better credit risks), or help an employer determine whom to hire (e.g., those with healthy personal habits may turn out to be more diligent employees). To the extent that context-violative data use breaks privacy norms—as Helen Nissenbaum and others have argued—such Big Data use of consumer sensors will disrupt consumers' expectations.[12]

Second, sensor data are particularly difficult to de-identify. Without delving into the burgeoning literature on de-identification generally—which has consistently shown that anonymized datasets are easier to re-identify than previously assumed[13]—the point here is that sensor data sets are particularly vulnerable. For example, Ira Hunt, Chief Technology Officer of the Central Intelligence Agency, recently noted that "simply by looking at the data [from a FitBit] … you can be one hundred percent guaranteed to be identified by simply your gait—how you walk."[14] Sensor data sets are prone to what computer scientists call "sparsity"—individuals can be re-identified relatively easily because sensor data measurements are so rich and detailed that each individual in the data set is reasonably unique. For example, researchers at MIT recently analyzed data on 1.5 million cellphone users in Europe over fifteen months and found that it was fairly easy to extract complete location information for a single person from an anonymized data set.[15] To do so only required locating that single user within several hundred yards of a cellphone transmitter sometime over the course of an hour four times in one year. With four such known data points, the researchers could identify 95 percent of the users in the data set. As one commentator put it, "what they are showing here, quite clearly, is that it's very hard to preserve anonymity."[16]

Third, at the moment sensor data seem particularly prone to security flaws. Because sensors often must be small to work in consumer devices, manufacturers currently may forego robust security technology in favor of a compact form factor. For example, a research team recently showed that FitBit health monitoring sensors could be hacked wirelessly from a distance of fifteen feet.[17] Sensors in automobiles—specifically, the tire pressure monitoring systems that are standard in almost all vehicles—can likewise be monitored from a distance as great as one hundred feet.[18] These

sorts of basic security problems threaten consumer privacy, and the current lack of regulatory consequences for breaches of sensor security mean that the growing sensor industry has little incentive to improve the situation.

The good news is that sensor privacy has salience. The power of sensors to capture our movements, behaviors, habits and even personalities in such high resolution—and the ease with which Big Data analysis can draw uncomfortable inferences from such data that could be used across a variety of contexts—are likely to prompt much more public interest and legislative response than "Big Data" in the abstract. No one wants her Nike FuelBand to unknowingly influence her credit score, or her driving habits sold behind the scenes to a prospective employer to assess her risk-taking or sense of responsibility. Sensor privacy may therefore be an easier regulatory target than Big Data generally, and thus a way to begin to ensure that Big Data analysis happens responsibly. Again, the key to realistic but timely progress towards accountability is to find tangible, simple regulatory actions that will constrain out-of-bounds uses without overly limiting Big Data's promise.

Here are some concrete first steps. In the last decade, legislatures in all but a few states have passed data breach notification laws that require companies to disclose publicly serious computer security violations compromising personal information. None of these state laws currently covers sensor data independently of other personally identifiable information.[19] Each should. State legislatures could relatively easily amend such statutes to include biometric and other sensor data so that firms take seriously their obligation to protect such information. Given the uniqueness and difficulty of de-identifying sensor data, if FitBit gets hacked, consumers should know.

Another fairly easy step: limiting the ability of firms to force consumers to disclose sensor data. Arkansas, North Dakota, Oregon, and Virginia, for example, have forbidden auto insurers from requiring consumers to consent to future access to a car's "black box" sensor data as a condition of insurability or payment of a claim.[20] Such sensor data helps the auto industry do data analytics to discover safety problems—it was not meant to influence insurance rates. This is a reasonable step that other jurisdictions should mimic.

More weighty would be restrictions on using sensor data from one domain—such as the information from a personal health monitor—to draw inferences in another domain—such as the financial decision of whether to lend to a given consumer. Just because your coffee pot knows that you are lazy and sleep late (or your car's black box knows that you speed too much), you shouldn't be prevented from getting a mortgage. As an example, several states have limited a utility company's ability to sell smart grid data to third parties.[21] Such use restrictions are reasonable—sensor data firms should not be tempted to exploit migratory uses. (If an informed individual wants to knowingly sell her data for such cross-context use, that is another matter altogether.)

Finally, wherever possible we should enforce the norm that consumers own and have access to sensor data about them. Currently, sensor manufacturers are free to use their privacy policy to claim ownership of users' biometric or other sensor data. Some do—the popular BodyMedia health monitoring armband is an example.[22] If just one state required manufacturers of personal health monitors to concede that consumers own and have access to their sensor information, that would radically clarify expectations in this domain. Similarly, the National Highway Traffic Safety Administration (NHTSA) should rule explicitly that a consumer owns and controls data generated by her automobile's event data recorder, following the lead of several state legislatures.

Each of these first steps is politically feasible precisely because each focuses on a specific, concrete problem in a particular sensor privacy context. This patchwork approach will no doubt seem cumbersome and frustrating to some, but it is the only realistic means to ensure accountability and constraint in Big Data analysis. Sensor privacy can be a model for future context-by-context, precision regulation of other aspects of Big Data infrastructure. If we can find ways to reasonably regulate the manufacturers of consumer sensor devices and

the users of the data those devices generate, both consumers and the Big Data industry will realize that regulation and innovation need not conflict. Good, fair uses will continue unfettered; less reasonable uses will be limited one-by-one. It will not be easy, but it will be worth it. Ultimately, Big Data—if done responsibly—will change the world for the better. Reassuring consumers that basic accountability has been provided is a necessary precondition to that revolution.

---

[1] *See e.g.* Paul M. Schwartz, *Preemption and Privacy*, 118 YALE L.J. 902 (2009).

[2] *See e.g.* WORLD ECONOMIC FORUM, UNLOCKING THE VALUE OF PERSONAL DATA: FROM COLLECTION TO USAGE 3 (2013) (advocating "a shift from focusing away from trying to control the data itself to focusing on the uses of data").

[3] *See e.g.* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013) (providing the most complete analysis of Big Data in the legal literature to date, and stressing legislative reform to provide consumers access to and transparency about information, but not identifying a specific legislative proposal). *See also* Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data*, 64 STAN. L. REV. ONLINE 63 (2012).

[4] *See e.g.* FRED H. CATE & VIKTOR MAYER-SCHONBERGER, NOTICE AND CONSENT IN A WORLD OF BIG DATA: MICROSOFT GLOBAL PRIVACY SUMMIT SUMMARY REPORT AND OUTCOMES (Nov. 2012) (noting the difficulties of applying notice and consent regimes to Big Data, and calling for other means to limit data use rather than collection, but not proposing concrete action).

[5] *See* UNLOCKING THE VALUE OF PERSONAL DATA, *supra* note 2, at 3.

[6] *See* Scott R. Peppet, *Privacy & The Personal Prospectus: Should We Introduce Privacy Agents or Regulate Privacy Intermediaries?*, 97 IOWA L. REV. BULL. 77 (2012) (providing more examples of such sensors).

[7] *See* Alexander Wolfe, *Little MEMS Sensors Make Big Data Sing*, FORBES (Jun. 10, 2013, 10:26 AM), http://www.forbes.com/sites/oracle/2013/06/10/little-mems-sensors-make-big-data-sing/2/ ("With the cost impediment overcome, deployment has caught fire.").

[8] *See* Bill Wasik, *Welcome to the Programmable World*, WIRED (May 14, 2013, 6:30 AM), http://www.wired.com/gadgetlab/2013/05/internet-of-things/.

[9] *See* Alex (Sandy) Pentland, *Reinventing Society in the Wake of Big Data*, EDGE (Aug. 30, 2012), http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data ("[T]he power of Big Data is that it is information about people's behavior instead of information about their beliefs. … This sort of Big Data comes from things like location data off of your cell phone … That's very different than what you put on Facebook.").

[10] *See* Andrew Raij et al., *Privacy Risks Emerging from the Adoption of Innocuous Wearable Sensors in the Mobile Environment*, CHI 2011 (May 7-12, 2011), http://animikh.in/raij-chi2011.pdf (showing that users of a wearable sensor were particularly sensitive to inferences drawn about their stress levels).

[11] *See See* Ann Cavoukian, Jules Polonetsky & Christopher Wolf, *SmartPrivacy for the Smart Grid: Embedding Privacy Into the Design of Electricity Conservation*, 3 IDENTITY IN THE INFORMATION SOCIETY 275, 284 (2010) (providing examples of such inferences); Miro Enev et al., *Inferring TV Content from Electrical Noise*, ACM CONFERENCE '10 1, 1 (2010) ("[W]e show that given a 5 minute recording of the electrical noise unintentionally produced by the TV it is possible to infer exactly what someone is watching (with an average accuracy of 96% …) by matching it to a database of content signatures.").

[12] *See* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010).

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

[14] Mathew Ingram, *Even the CIA is Struggling to Deal with the Volume of Real-Time Social Data*, GIGAOM (Mar. 20, 2013 10:27 AM), http://gigaom.com/2013/03/20/even-the-cia-is-struggling-to-deal-with-the-volume-of-real-time-social-data/.

[15] *See* Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen & Vincent D. Blondel, *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 SCIENTIFIC REPORTS 1 (2013) ("[L]ittle outside information is needed to re-identify the trace of a targeted individual even in a sparse, large-scale, and coarse mobility dataset.").

[16] Larry Hardesty, *How Hard Is It to "De-Anonymize" Cellphone Data?*, MIT NEWS (Mar. 27, 2013).

[17] *See* Mahmudur Rahman, Bogdan Carbunar & Madhusudan Banik, *Fit and Vulnerable: Attacks and Defenses for a Health Monitoring Device*, arXiv:1304.5672 (Apr. 20, 2013).

[18] *See* Ishtiaq Rouf et al., *Security and Privacy Vulnerabilities of In-Car Wireless Networks: A Tire Pressure Monitoring System Case Study*, USENIX SECURITY 21 (2010).

[19] State data breach disclosure statutes are available at National Conference of State Legislatures, http://www.ncsl.org/issues-research/telecom/security-breach-notification-laws.aspx.

[20] *See* Ark. Code § 23-112-107; N.D. Cent. Code § 51-07-28; Ore. Rev. Stat. §§ 105.925-.948; Va. Code § 38.2-2212(C)(s).

[21] *See* S.B. 674, 2011-2012 Reg. Sess. (Cal. 2011); H.B. 1191, 68th Gen. Assemb. (Colo. 2011).

[22] *See* http://www.bodymedia.com/privacy-policy ("All data collected including, but not limited to, food-logs, weight, body-fat-percentage, sensor-data, time recordings, and physiological data … are and shall remain the sole and exclusive property of BodyMedia.").

THREE PARADOXES OF BIG DATA

*Neil M. Richards & Jonathan H. King**

66 STAN. L. REV. ONLINE 41

INTRODUCTION

Big data is all the rage. Its proponents tout the use of sophisticated analytics to mine large data sets for insight as the solution to many of our society's problems. These big data evangelists insist that data-driven decisionmaking can now give us better predictions in areas ranging from college admissions to dating to hiring.[1] And it might one day help us better conserve precious resources, track and cure lethal diseases, and make our lives vastly safer and more efficient. Big data is not just for corporations. Smartphones and wearable sensors enable believers in the "Quantified Self" to measure their lives in order to improve sleep, lose weight, and get fitter.[2] And recent revelations about the National Security Agency's efforts to collect a database of all caller records suggest that big data may hold the answer to keeping us safe from terrorism as well.

Consider *The Human Face of Big Data*, a glossy coffee table book that appeared last holiday season, which is also available as an iPad app. Such products are thinly disguised advertisements for big data's potential to revolutionize society. The book argues that "Big Data is an extraordinary knowledge revolution that's sweeping, almost invisibly, through business, academia, government, healthcare, and everyday life."[3] The app opens with a statement that frames both the promise and the peril of big data: "Every animate and inanimate object on earth will soon be generating data, including our homes, our cars, and yes, even our bodies." Yet the app and the book, like so many proponents of big data, provide no meaningful analysis of its potential perils, only the promise.

We don't deny that big data holds substantial potential for the future, and that large dataset analysis has important uses today. But we would like to sound a cautionary note and pause to consider big data's potential more critically. In particular, we want to highlight three paradoxes in the current rhetoric about big data to help move us toward a more complete understanding of the big data picture. First, while big data pervasively collects all manner of private information, the operations of big data itself are almost entirely shrouded in legal and commercial secrecy. We call this the *Transparency Paradox*. Second, though big data evangelists talk in terms of miraculous outcomes, this rhetoric ignores the fact that big data seeks to identify at the expense of individual and collective identity. We call this the *Identity Paradox*. And third, the rhetoric of big data is characterized by its power to transform society, but big data has power effects of its own, which privilege large government and corporate entities at the expense of ordinary individuals. We call this the *Power Paradox.* Recognizing the paradoxes of big data, which show its perils alongside its potential, will help us to better understand this revolution. It may also allow us to craft solutions to produce a revolution that will be as good as its evangelists predict.

THE TRANSPARENCY PARADOX

* Neil M. Richards is Professor of Law, Washington University. Jonathan H. King is Vice President, Cloud Strategy and Business Development, Savvis, a CenturyLink Company.

Big data analytics depend on small data inputs, including information about people, places, and things collected by sensors, cell phones, click patterns, and the like. These small data inputs are aggregated to produce large datasets which analytic techniques mine for insight. This data collection happens invisibly and it is only accelerating. Moving past the Internet of Things to the "Internet of Everything," Cisco projects that thirty-seven billion intelligent devices will connect to the Internet by 2020.[4] These devices and sensors drive exponentially growing mobile data traffic, which in 2012 was almost twelve times larger than all global Internet traffic was in 2000.[5] Highly secure data centers house these datasets on high-performance, low-cost infrastructure to enable real-time or near real-time big data analytics.

This is the Transparency Paradox. Big data promises to use this data to make the world more transparent, but its collection is invisible, and its tools and techniques are opaque, shrouded by layers of physical, legal, and technical privacy by design. If big data spells the end of privacy, then why is the big data revolution occurring mostly in secret?

Of course, there are legitimate arguments for some level of big data secrecy (just as there remain legitimate arguments for personal privacy in the big data era). To make them work fully, commercial and government big data systems which are constantly pulling private information from the growing Internet of Everything are also often connected to highly sensitive intellectual property and national security assets. Big data profitability can depend on trade secrets, and the existence of sensitive personal data in big databases also counsels for meaningful privacy and security. But when big data analytics are increasingly being used to make decisions about individual people, those people have a right to know on what basis those decisions are made. Danielle Citron's call for "Technological Due Process"[6] is particularly important in the big data context, and it should apply to both government and corporate decisions.

We are not proposing that these systems be stored insecurely or opened to the public *en masse*. But we must acknowledge the Transparency Paradox and bring legal, technical, business, government, and political leaders together to develop the right technical, commercial, ethical, and legal safeguards for big data and for individuals.[7] We cannot have a system, or even the appearance of a system, where surveillance is secret,[8] or where decisions are made about individuals by a Kafkaesque system of opaque and unreviewable decisionmakers.[9]

The Identity Paradox

Big data seeks to *identify*, but it also threatens *identity*. This is the Identity Paradox. We instinctively desire sovereignty over our personal identity. Whereas the important right to privacy harkens from the right to be left alone,[10] the right to identity originates from the right to free choice about who we are. This is the right to define who "I am." I am me; I am anonymous. I am here; I am there. I am watching; I am buying. I am a supporter; I am a critic. I am voting; I am abstaining. I am for; I am against. I like; I do not like. I am a permanent resident alien; I am an American citizen.

How will our right to identity, our right to say "I am," fare in the big data era? With even the most basic access to a combination of big data pools like phone records, surfing history, buying history, social networking posts, and others, "I am" and "I like" risk becoming "you are" and "you will like." Every Google user is already influenced by big-data-fed feedback loops from Google's tailored search results, which risk producing individual and collective echo chambers of thought. In his article, *How Netflix Is Turning Viewers into Puppets*, Andrew Leonard explains how:

> The companies that figure out how to generate intelligence from that data will know more about us than we know ourselves, and will be able to craft techniques that push us toward where they want us to go, rather than where we would go by ourselves if left to our own devices.[11]

Taking it further, by applying advances in personal genomics to academic and career

screening, the dystopian future portrayed in the movie *Gattaca*[12] might not be that outlandish. In *Gattaca*, an aspiring starship pilot is forced to assume the identity of another because a test determines him to be genetically inferior. Without developing big data identity protections now, "you are" and "you will like" risk becoming "you cannot" and "you will not". The power of Big Data is thus the power to use information to nudge, to persuade, to influence, and even to restrict our identities.[13]

Such influence over our individual and collective identities risks eroding the vigor and quality of our democracy. If we lack the power to individually say who "I am," if filters and nudges and personalized recommendations undermine our intellectual choices, we will have become identified but lose our identities as we have defined and cherished them in the past.

The Power Paradox

The power to shape our identities for us suggests a third paradox of big data. Big data is touted as a powerful tool that enables its users to view a sharper and clearer picture of the world.[14] For example, many Arab Spring protesters and commentators credited social media for helping protesters to organize. But big data sensors and big data pools are predominantly in the hands of powerful intermediary institutions, not ordinary people. Seeming to learn from Arab Spring organizers, the Syrian regime feigned the removal of restrictions on its citizens' Facebook, Twitter, and YouTube usage only to secretly profile, track, and round up dissidents.[15]

This is the Power Paradox. Big data will create winners and losers, and it is likely to benefit the institutions who wield its tools over the individuals being mined, analyzed, and sorted. Not knowing the appropriate legal or technical boundaries, each side is left guessing. Individuals succumb to denial while governments and corporations get away with what they can by default, until they are left reeling from scandal after shock of disclosure. The result is an uneasy, uncertain state of affairs that is not healthy for anyone and leaves individual rights eroded and our democracy diminished.

If we do not build privacy, transparency, autonomy, and identity protections into big data from the outset, the Power Paradox will diminish big data's lofty ambitions. We need a healthier balance of power between those who generate the data and those who make inferences and decisions based on it, so that one doesn't come to unduly revolt or control the other.

Conclusion

Almost two decades ago, Internet evangelist John Perry Barlow penned *A Declaration of the Independence of Cyberspace*, declaring the Internet to be a "new home of [the] Mind" in which governments would have no jurisdiction.[16] Barlow was one of many cyber-exceptionalists who argued that the Internet would change everything. He was mostly right—the Internet did change pretty much everything, and it did create a new home for the mind. But the rhetoric of cyber-exceptionalism was too optimistic, too dismissive of the human realities of cyberspace, the problems it would cause, and the inevitability (and potential utility) of government regulation.

We think something similar is happening in the rhetoric of big data, in which utopian claims are being made that overstate its potential and understate the values on the other side of the equation, particularly individual privacy, identity, and checks on power. Our purpose in this Essay is thus twofold.

First, we want to suggest that the utopian rhetoric of big data is frequently overblown, and that a less wild-eyed and more pragmatic discussion of big data would be more helpful. It isn't too much to ask sometimes for data-based decisions about data-based decisionmaking.

Second, we must recognize not just big data's potential, but also some of the dangers that powerful big data analytics will unleash upon society. The utopian ideal of cyberspace needed to yield to human reality, especially when it revealed problems like identity theft, spam, and cyber-bullying. Regulation of the Internet's excesses was (and is) necessary in order to gain the benefits of its substantial breakthroughs. Something similar must happen with big data, so that we can take advantage of the good

things it can do, while avoiding as much of the bad as possible. The solution to this problem is beyond the scope of this short symposium essay, but we think the answer must lie in the development of a concept of "Big Data Ethics"—a social understanding of the times and contexts when big data analytics are appropriate, and of the times and contexts when they are not.

Big data will be revolutionary, but we should ensure that it is a revolution that we want, and one that is consistent with values we have long cherished like privacy, identity, and individual power. Only if we do that will big data's potential start to approach the story we are hearing from its evangelists.

---

[1] *See, e.g.*, Adam Bryant, *In Head-Hunting, Big Data May Not Be Such a Big Deal*, N.Y. Times (June 19, 2013), http://www.nytimes.com/2013/06/20/business/in-head-hunting-big-data-may-not-be-such-a-big-deal.html?pagewanted=all&_r=0.

[2] *See* Emily Singer, *Is "Self-tracking" the Secret to Living Better?*, MIT TECH. REV. (June 9, 2011), http://www.technologyreview.com/view/424252/is-self-tracking-the-secret-to-living-better.

[3] RICK SMOLAN & JENNIFER ERWITT, THE HUMAN FACE OF BIG DATA 3 (2012).

[4] Dave Evans, *How the Internet of Everything Will Change the World . . . for the Better #IoE [Infographic]*, CISCO BLOGS (Nov. 7, 2012, 9:58 AM PST),http://blogs.cisco.com/news/how-the-internet-of-everything-will-change-the-worldfor-the-better-infographic.

[5] *See* CISCO, CISCO VISUAL NETWORKING INDEX: GLOBAL MOBILE DATA TRAFFIC FORECAST UPDATE, 2012-2017, at 1 (2013), *available at* http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf.

[6] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008).

[7] *See* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW J. TECH. & INTELL. PROP. 239, 270-72 (2013).

[8] *See* Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1959-61 (2013).

[9] *Cf.* DANIEL J. SOLOVE, THE DIGITAL PERSON (2005).

[10] *See* Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904, 1906 (2013).

[11] Andrew Leonard, *How Netflix Is Turning Viewers into Puppets*, SALON (Feb. 1, 2013, 7:45 AM EST),http://www.salon.com/2013/02/01/how_netflix_is_turning_viewers_into_puppets.

[12] GATTACA (Columbia Pictures 1997).

[13] *See* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (2009); Richards, *supra* note 8, at 1955-56.

[14] *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 11 (2013).

[15] *See* Stephan Faris, *The Hackers of Damascus*, BLOOMBERG BUSINESSWEEK (Nov. 15, 2012), http://www.businessweek.com/articles/2012-11-15/the-hackers-of-damascus.

[16] John Perry Barlow, *A Declaration of the Independence of Cyberspace*, ELEC. FRONTIER FOUND. (Feb. 8, 1996), https://projects.eff.org/~barlow/Declaration-Final.html.

BIG DATA:
A PRETTY GOOD PRIVACY SOLUTION

*Ira S. Rubinstein**

INTRODUCTION

Big data—by which I mean the use of machine learning, statistical analysis, and other data mining techniques to extract hidden information and surprising correlations from very large and diverse data sets—raises numerous privacy concerns. A growing number of privacy scholars (myself included) have argued that big data casts doubt on the Fair Information Practices ('FIPs'), which form the basis of all modern privacy law.[1] With the advent of big data, the FIPs seem increasingly anachronistic for three reasons. First, big data heralds the shift from data actively collected with user awareness and participation to machine-to-machine transactions (think of electronic toll-collection systems) and passive collection (data collected as a by-product of other activities like searching or browsing the web).[2] Thus, big data nullifies informed choice, undermining the FIPs at their core. Second, big data thrives on comingling and sharing large data sets to create economic value and innovation from new and unexpected uses, making it inimical to collection, purpose, use or retention limitations, without which the FIPs are toothless. Finally, big data seems to make anonymization impossible. Why? The amount of data available for analysis has increased exponentially and while much of it seems non-personal, researchers have shown that almost any attribute, when combined with publicly available background information, can be linked back to an individual.[3] There is a large and growing literature on whether anonymization is no longer an effective strategy for protecting privacy[4] and to what extent this failure makes it impossible to publicly release data that is both private and useful.[5]

This indictment of the FIPs paints big data with a broad brush. And yet a moment's thought suggests that not every big data scenario is necessarily alike or poses the same risk to privacy. Having reviewed dozens of big-data analyses culled from the lay literature, I want to explore whether they have distinguishing characteristics that would allow us to categorize them as having a low, medium, or high risk of privacy violations.[6] In what follows, I offer a tentative and preliminary categorization of big data scenarios and their varying levels of risks. And I emphasize two supplemental FIPs that may help address some (but not all) of the riskier scenarios: first, a default prohibition on the transfer of large data sets to third parties for secondary uses without the explicit, opt-in consent of the data subject; and, second, a broad prohibition on the re-identification of anonymized data, with violators subject to civil and/or criminal sanctions. This approach is partial and imperfect at best but perhaps offers a pretty good privacy solution for the moment.

DISCUSSION

In a recent book explaining big data for the lay reader, Viktor Mayer-Schönberger and Kenneth Cukier describe dozens of scenarios in which big data analytics extract new insights.[7] Several of these scenarios are low-risk and raise no or minimal privacy alarms. As they observe, "Sensor data from refineries does not [contain personal information], nor does machine data from factory floors or data on manhole explosions or airport weather."[8] What about services using billions of flight-price records to predict the direction of prices on specific airline routes or popular web services using billions of text or voice samples and "machine learning" algorithms to develop highly accurate spam

* Senior Fellow and Adjunct Professor of Law, Information Law Institute, New York University School of Law.

filters, grammar and spell checkers, and translation and voice recognition tools? These scenarios are low risk for several reasons: they mainly involve first-party collection and analysis of non-personal or de-identified data, they seek to improve or enhance devices or systems that affect consumers rather than specific individuals, and they involve either very limited or pre-defined data sets that are not shared with others. And the services have little incentive to re-identify individuals; indeed, they may have made binding promises to safeguard data security.

If other risk factors are present, however, first party collection and analysis of limited data sets may be more troubling. Medium-risk scenarios occur when (1) the data is personal and/or the first party contemplates (2) sharing the data with a third party for secondary uses or (3) a broad or public data release. And yet it is possible to reduce the privacy risks in each of these cases.

A good example of (1) is Google Flu Trends, which uses search engine query data and complex models for the early detection of flu epidemics. Although search queries are IP-based and therefore identifiable, Google safeguards privacy by aggregating historical search logs and discarding information about the identity of every user.[9]

A good example of (2) is any smart meter system subject to California's SB 1476, a recently-enacted privacy law that "requires aggregators of energy consumption data to obtain consumer consent before sharing customer information with third parties; mandates that third parties may only have access to such data when they are contracting with the utility to provide energy management-related services; stipulates that data be kept secure from unauthorized parties; and mandates that electricity ratepayers opt in to authorize any sharing of their energy consumption data for any secondary commercial purpose[s]."[10] Absent such protections, utilities might be tempted to sell consumption data for analysis and secondary use by third parties for marketing purposes or to determine insurance risk. SB 1476 permits first party data analysis for operational purposes that benefit both consumers and society while also addressing the risks associated with third party sharing for secondary uses.

A good example of (3) is using anonymized geo-location data derived from GPS-equipped devices to optimize public transit systems. The analysis relied on a research challenge dubbed "Data for Development" in which the French telecom Orange "released 2.5 billion call records from five million cell-phone users in Ivory Coast. . . . The data release is the largest of its kind ever done. The records were cleaned to prevent anyone identifying the users, but they still include useful information about these users' movements."[11] Locational data is highly sensitive and it has proven very difficult to achieve anonymization by removing identifiers from mobility datasets.[12] However, the researchers who gained access to the Orange data set had to be affiliated with a public or private research institution, submit a research proposal for approval, and sign a data-sharing agreement.[13] These agreements typically prohibit re-identification of the data subject and impose additional security and privacy safeguards such as audits, privacy impact assessments, and data destruction upon completion of the research.[14] This contractual approach seems to finesse the "de-identification dilemma"[15] by avoiding both Ohm's Scylla (that anonymized data sets lack either privacy or utility) and Yakowitz's Charybdis (that all useful research requires the public release of anonymized data sets).[16]

High-risk scenarios occur whenever big data analytics result in actions taken regarding groups with sensitive attributes or affecting specific individuals. Mayer-Schönberger and Cukier provide several relevant examples such as startups that would determine a consumer's credit rating based on "behavioral scoring" using rich social media data sets not regulated by fair credit reporting laws; insurance firms that would identify health risks by combining credit scores with various lifestyle data not regulated by any privacy laws; and the notorious Target incident, in which the firm used big data analytics to predict whether female shoppers were newly pregnant and then marketed baby-related products to them, even though they may have delayed sharing this news with family members.[17] Why are these high-risk scenarios?

First, the data sets are large and heterogeneous, increasing the likelihood that analysis will reveal sensitive or intimate attributes, even though we think of the underlying data as non-personal. Second, the data comes from multiple sources, so individuals are unaware of how third parties collect, store or use it and therefore lack any ability to access their data or control detrimental uses of inferred attributes. Third, when firms rely on big data analytics to infer sensitive attributes (creditworthiness, insurability, pregnancy), they often skirt regulations limited to the collection and use of specific types of personal data. Another problem is that these analytic techniques are imperfect and may result in erroneous or unfair decisions.[18] In any case, the underlying privacy issues in high-risk scenarios are far more difficult to address: at a minimum, they require stronger default rules and perhaps a major shift in business models and new and innovative data frameworks.[19]

CONCLUSION

This short essay seeks to characterize big data scenarios according to their level of privacy risks and to identify supplemental FIPs that might help mitigate these risks. Whether this approach is worthwhile requires further study of many more scenarios and development of a more comprehensive set of risk criteria and supplemental privacy principles. A risk-based approach is at best a compromise. Yet is has the virtue of acknowledging that while the anonymous release of useful data is no silver bullet for privacy, neither is big data in all cases a poison pill.

---

[1] *See* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 257-63 (2013); Ira S. *Rubinstein*, *Big Data*: The End of Privacy or a New Beginning? 3 INT'L DATA PRIV. L. 74, 78 (2012).

[2] *See* World Economic Forum, *Unlocking the Value of* Personal Data: From Collection to Usage 7-8 (2013), http://www.weforum.org/reports/unlocking-value-personal-data-collection-usage.

[3] *See* Arvind Narayanan & Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, 2008 PROC. 29TH IEEE SYMP. ON SECURITY & PRIVACY 111.

[4] *Compare* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010) *with* Jane Yakowitz, *Tragedy of the Data Commons* 25 HARV. J. L. & TECH. 1 (2011).

[5] *See* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. (forthcoming 2013).

[6] This paper confines itself to consumer and research scenarios and does not address government data mining.

[7] *See* VIKTOR MAYER-SCHÖNBERGER AND KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013). All of the big data scenarios discussed below are drawn from this book unless otherwise noted.

[8] *Id.* at 152.

[9] *See* Jeremy Ginsberg, et al., *Detecting Influenza Epidemics Using Search Engine Query Data* 457 NATURE 1012 (2009).

[10] *See* John R. Forbush, *Regulating the Use and Sharing of Energy Consumption Data: Assessing California's SB 1476 Smart Meter Privacy Statute*, 75 ALB. L. REV. 341, 343 (2012).

[11] *See* David Talbot, *African Bus Routes Redrawn Using Cell-Phone Data*, MIT TECH. REV. (Apr. 30, 2013).

[12] *See* Y.-A. de Montjoye, et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, SCIENTIFIC REPORTS 3 (March 25, 2013), http://www.readcube.com/articles/10.1038%2Fsrep01376. However, efforts are underway to make large-scale mobility models provably private without unduly sacrificing data accuracy using new techniques based on differential privacy; *see* Darakshan J. Mir, et al., *Differentially Private Modeling of Human Mobility at Metropolitan Scale* (2013) (unpublished paper on file with the author).

[13] *See* D4D Challenge, Learn More, http://www.d4d.orange.com/learn-more (last visited June 25, 2013).

[14] *See* Khaled El Emam, *Risk-Based De-Identification of Health Data,* IEEE SEC & PRIV, May-June 2010, at 66, 64-67. Both Ohm, *supra* note 4 at 1770, and Yakowitz, *supra* note 4 at 48-49, endorse penalizing improper re-identification.

[15] *See* Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 FORDHAM INTELL. PROP. MEDIA & ENT. L. J. 33 (2010).

[16] Dozens of papers describing presumably valuable research results from the D4D Challenge were presented at the 2013 NetMob conference at MIT, *available at* http://perso.uclouvain.be/vincent.blondel/netmob/2013/.

[17] For a fascinating and detailed account, *see* Charles Duhigg, *How Companies Learn Your Secrets*, NY TIMES (Feb. 16, 2012), *available at* http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all.

[18] *See* Kate Crawford & Jason Schultz, *The Due Process Dilemma: Big Data and Predictive Privacy Harms* (2013) (unpublished paper on file with the author).

[19] The World Economic Forum has published several reports championing a new business model based on personal data stores (PDS); *see* http://www.weforum.org/issues/rethinking-personal-data (last visited June 25, 2013). For a privacy-protective implementation of PDS, *see* Y.-A. de Montjoye, et al., *On the Trusted Use of Large-Personal Data*, 35 IEEE DATA ENG. BULL. 5 (2012).

BIG DATA AND THE "NEW" PRIVACY TRADEOFF

*Robert H. Sloan & Richard Warner**

Predictions of transformative change surround Big Data. It is routine to read, for example, that "with the coming of Big Data, we are going to be operating very much out of our old, familiar ballpark."[1] But, as both Niels Bohr and Yogi Berra are reputed to have observed, "Prediction is difficult, especially about the future." And, they might have added, especially regarding the effects of major technological change. In the Railroad Mania of nineteenth century England, for example, some made the typical prediction that a new communication network meant the end of an old one: namely, that that face-to-face communication over the emerging railroad network would entail a drastic drop in postal mail. In fact, mail volume increased.[2] Given the difficulty of forecasting transformative change, we opt for a "prediction" about the present: Big Data *already* presents a "new" and important privacy challenge. As the scare quotes indicate, the challenge is not truly new. What Big Data does is compel confrontation with a difficult trade-off problem that has been glossed over or even ignored up to now. It does so because both the potential benefits and risks from Big Data analysis are so much larger than anything we have seen before.

We confine our inquiry to the private sector. Governmental concerns are critically important, but they require separate treatment.

THE TRADEOFF PROBLEM AND BIG DATA

We claim Big Data greatly exacerbates a now decades old problem about how to balance the benefits of data collection and analysis against the relevant privacy risks. In the 1990s and early 2000s, before the current Big-Data era, commentators typically identified the following benefits of data collection: increased economic efficiency, improved security, better personalization of services, increased availability of relevant information, and innovative platforms for communication.[3] The tradeoff task was to balance that relatively short list of benefits against the loss of informational privacy. (By informational privacy, we mean the ability to control who collects information about you and what they do with it, and data collection and analysis reduces one's control.) Unfortunately, while privacy advocates and policy makers acknowledge tradeoff issues, they typically pay little attention to them.[4] Instead, they concentrate on the—also crucial—task of ensuring free and informed consent to businesses' data collection and use practices. Big Data compels a change: it involves such large and important risks *and* benefits that there is no longer any excuse for setting tradeoff issues aside.

"Big Data" refers to the acquisition and analysis of massive collections of information, collections so large that until recently the technology needed to analyze them did not exist.[5] The analysis can reveal patterns that would otherwise go unnoticed, and this has already yielded an astonishing array of benefits from detecting drug interactions to improving access to social services in India by creating digital IDs for citizens.[6] The risks are equally serious. The risk of a massive loss of informational privacy has become much larger, and there are other risks as well. Consider improving access to social services in India. A significant improvement will increase the demand

* Robert H. Sloan is Professor and Head, Department of Computer Science, University of Illinois at Chicago. Partially supported by National Science Foundation Grant No. DGE-1069311. Richard Warner is Professor of Law, Chicago-Kent College of Law, Visiting Foreign Professor, University of Gdańsk, Poland.

for the services. Meeting that demand may require an increased investment in those services, thereby creating at least two risks: social discontent if the demand is not met; and, the diversion of scarce resources from other critical areas if it is. An acceptable level of information flow into Big Data analysis is one that yields acceptable tradeoffs between risks and benefits. The problem is to find a level of information flow that does that. The current mechanisms for determining the proper level are woefully inadequate.

MID-20TH CENTURY INFORMATION PROCESSING

To see why, it helps to turn back the clock to the mid-twentieth century. Data collection was in its infancy, with only the beginnings of credit reporting practices. Direct marketing was not widely used until the 1970s because prior to that time it was too difficult to differentiate among consumers (the change came when the government began selling census data on magnetic tapes).[7] People did disclose information to businesses, governmental and private licensing agencies, and so on, but the information was typically stored in paper records and geographically scattered. There was no convenient way to search all of it or to retrieve readily storable, reusable information. You could by and large regulate the flow of your information to private businesses in the way you thought best. The sum of the individual decisions about data collection provided the answer to how much information should flow to businesses for analysis.

Did this yield an acceptable level of information flow? The answer did not matter much because mid-twentieth century information processing did not generate significant risks and benefits compared to today, but, in general, summing individual decisions is not a good way to answer "acceptable level" tradeoff questions, as the following example illustrates.[8] Imagine that in a community that does not have a telephone book, everyone would like to have one. However, each person prefers not to have his or her phone number listed and so refuses to consent to listing. No phone book is the result—a result each regards as much worse than having one.

Unfortunately, society has not yet—in the opening decades of the twenty-first century—changed its ways. Summing individual decisions still plays a key role in answering the "acceptable level" question. Summing individual decisions works extremely well for setting prices in highly competitive markets with no externalities, but can work very poorly indeed when results of individual decisions come with significant externalities. For Big Data today, there are tremendous externalities: Decisions by individual consumers to withhold data may have large negative externalities for society's overall ability to reap the benefits of Big Data, and decisions by individual businesses may have large negative externalities for citizens' privacy.

THE CURRENT MECHANISM FOR SUMMING INDIVIDUAL DECISIONS

Outside the health and finance sectors, private businesses are relatively unconstrained in their data collection and analysis practices, and summing individual decisions still plays a key role in determining the level of information that flows to private businesses. We focus on the online context, but similar remarks hold for offline situations. Online, the current summing mechanism is Notice and Choice (sometimes called Notice and Consent). The "notice" is a presentation of terms. The "choice" is an action signifying acceptance of the terms (typically using a website or clicking on an "I agree" button). Implementations of Notice and Choice lie along a spectrum. One extreme is home to implementations that place few restrictions on Notices (how they are presented and what they may or must say) and few restrictions on what counts as choice (using the site, clicking on an "I agree" button); the other extreme is occupied by restrictive implementations requiring conformity to some or all of the Fair Information Practice Principles of transparency, error correction, restriction of use of data to purposes stated at the time of collection, deletion of data when it is no longer used for that purpose, and data security.

Proponents of Notice and Choice make two claims. First: when adequately implemented, (the appropriate version of) Notice and Choice ensures that website visitors can give free and informed consent to businesses' data collection and use practices. For purposes of this essay, we grant the first claim.[9] Our concern is with the second claim: namely, that the sum of the individual consent

decisions determines an acceptable level of information flowing to businesses. We see little reason to think it is true. As the telephone book example illustrates, summing individual decisions can lead to information flows that are inconsistent with what the individuals making those decisions would collectively agree is good overall. We believe Notice and Choice will not yield results good for society as a whole. In all its versions, Notice and Choice leaves tradeoff issues largely to the discretion of private business.[10] The Notices under which they collect consumers' information leave the subsequent uses of that information largely up to the businesses. By way of illustration, consider one well-known example. Microsoft allowed Dr. Russ Altman to analyze Bing searches for search terms correlated with dangerously high blood sugar levels. This was a key step in Altman's confirming that the antidepressant Paxil together with the anti-cholesterol drug Pravachol could result in diabetic blood sugar levels.[11] Our point is that the decision about how to use the Bing searches was *Microsoft's*. The Altman result is a life-saving one, but not all uses of Big Data are so uncontroversially good. Target, for example, infamously uses Big Data analysis to predict which of their customers are pregnant,[12] and it would be remarkable if decisions by businesses about data use reliably yielded acceptable society-wide balances of risks and benefits. Each business will balance in ways that serve its business goals, and there is no reason to think that summing up business decisions will yield an acceptable balance of risks and benefits from the point of view of society as a whole. This is just the "summing" problem over again with businesses making the decisions instead of consumers. Since the businesses do not suffer any of the negative effects on consumers of the loss of informational privacy, they will undervalue consumers' interests and reach an unacceptably biased overall tradeoff.

THE NOT-NEW-BUT-NOW-MORE-DIFFICULT-AND-IMPORTANT PROBLEM

Is there a way to balance risks and benefits that reliably yields acceptable results? We will not answer that question here.[13] Our point is that this problem is not new, but that Big Data does make it both considerably more difficult and considerably more important. We can certainly no longer reasonably rely on an approach that was acceptable in the mid-twentieth century only because back then information processing created relatively small benefits and risks.

---

[1] Alex (Sandy) Pentland, *Reinventing Society in the Wake of Big Data*, EDGE, 2012, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data.

[2] Andrew Odlyzko, *The Volume and Value of Information*, 6 INT. J. COMMUN. 920, 925 (2012).

[3] *See, e.g.*, Jerry Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193–1294 (1998) (emphasizing availability of relevant information, increased economic efficiency, improved security).

[4] *See* Fred Cate, *The Failure of Fair Information Practice Principles*, *in* THE FAILURE OF FAIR INFORMATION PRACTICE PRINCIPLES 342, 361–367 (Jane Winn ed., 2006).

[5] Omer Tene & Jules Polonetsky, *Privacy In The Age Of Big Data: A Time For Big Decision*, 64 STAN. L. REV. ONLINE 63 (2012).

[6] RICK SMOLAN & JENNIFER ERWITT, THE HUMAN FACE OF BIG DATA 34 (2012).

[7] DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 18 (2004).

[8] Amartya Sen, *Social Choice*, *in* THE NEW PALGRAVE DICTIONARY OF ECONOMICS (2nd ed. 2008), http://www.dictionaryofeconomics.com/dictionary.

[9] We criticize and reject the claim in Robert H. Sloan & Richard Warner, *Beyond notice and Choice: Privacy, Norms, and Consent*, ___ SUFFOLK UNIV. J. HIGH TECHNOL. LAW ___ (2014).

[10] The point is widely accepted. We give our reasons for it in Richard Warner & Robert H Sloan, *Behavioral Advertising: From One-Sided Chicken to Informational Norms*, VANDERBILT ENTERTAIN. TECHNOL. LAW J. 15 (2012).

[11] See Peter Jaret, *Mining Electronic Records for Revealing Health Data*, NEW YORK TIMES, January 14, 2013, http://www.nytimes.com/2013/01/15/health/mining-electronic-records-for-revealing-health-data.html?pagewanted=all.

[12] ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE Kindle Locations 1368–1376 (Kindle Edition ed. 2013).

[13] We offer a partial answer in ROBERT H. SLOAN & RICHARD WARNER, UNAUTHORIZED ACCESS: THE CRISIS IN ONLINE PRIVACY AND INFORMATION SECURITY (2013).

PRIVACY IN A POST-REGULATORY WORLD:
LESSONS FROM THE ONLINE SAFETY DEBATES

*Adam Thierer**

No matter how well-intentioned, privacy laws and regulations are being increasingly strained by the realities of our modern Information Age, and that fact should influence our strategies for dealing with the challenges posed by ubiquitous social media, Big Data, and the coming "Internet of Things."[1] Specifically, we need to invert the process of how we go about protecting privacy by focusing more on bottom-up solutions—education, empowerment, media literacy, digital citizenship lessons, *etc.*—instead of top-down legalistic solutions or techno-quick fixes.[2] In this regard, we can draw important lessons from the debates over how best to protect children from objectionable online content.[3]

NEW REALITIES

Lawmakers and policy advocates who worry about how best to protect online privacy today must contend with the fact that, for better or worse, we now live in a world that is ruthlessly governed by two famous Internet aphorisms. First, "information wants to be free." Sometimes that fact is worth celebrating. "Unfortunately," notes computer scientist Ben Adida, "information replication doesn't discriminate: your *personal data*, credit cards and medical problems alike, also want to be free. Keeping it secret is really, really hard," he correctly notes.[4]

A second well-known Internet aphorism explains why this is the case: "The Net interprets censorship as damage and routes around it," as Electronic Frontier Foundation co-founder John Gilmore once noted.[5] But this insight applies to *all* classes of information. Whether we are talking about copyright policy, cybersecurity, state secrets, pornography, hate speech, or personal information, the reality is always the same: *Any* effort to control information flows will be resisted by many other forces or actors in the online ecosystem. Moreover, once the genie is out of the bottle, it is incredibly hard to get it back in.

These two realities are the byproduct of the Internet's decentralized, distributed nature; the unprecedented scale of modern networked communications; the combination of dramatic expansions in computing and processing power (also known as "Moore's Law")[6] alongside a steady drop in digital storage costs; and the rise of widespread Internet access and ubiquitous mobile devices and access.

Compounding matters further still—especially for efforts to protect privacy—is the fact that we are our own worst enemies when it comes to information containment. Ours is a world of unprecedented individual information sharing through user-generation of content and self-revelation of data. On top of that, we now have decentralized peer-on-peer surveillance; new technologies make it easier than ever for us to release information not only about ourselves abut about all those around us.

Traditional information control mechanisms are being strained to the breaking point in this new environment and we need to be discussing how to come to grips with these new realities.

* Senior Research Fellow, Mercatus Center, George Mason University.

A Conversation Few Want to Have

Unfortunately, we're not having that conversation today. Or, to the extent we are, we're focused on the wrong set of issues or solutions. Discussions about protecting online privacy and reputation are still predominately tied up with philosophical ("What privacy rights do we have?) and legalistic ("How should we enforce those rights?") debates. Outside of some very narrow contexts (i.e., sensitive health and financial information), consensus about privacy rights has been elusive here in the United States.

The urge to delineate a tidy set of neatly-defined privacy rights and then protect them by law is completely understandable. But it is becoming more of a pipe dream with each passing year. Call me a defeatist, but esoteric metaphysical debates about the nature of our privacy rights and heated policy debates about how to enforce them are increasingly a waste of time.

Moreover, at some point the costs associated with regulatory controls must be taken into account. If we conduct a careful benefit-cost analysis of various regulatory proposals—something that has been woefully lacking on the privacy front in recent years—we find that many complex economic and social trade-offs are at work.[7] Regulation is not a costless exercise and, as noted, there are reasons to doubt it will even be effective if pursued.

New Approaches

We desperately need a new approach and I believe we can find it by examining the debate we have had about online child protectio over the past 15 years.[8] Since the dawn of the commercial Internet in the early 1990s, online safety and access to objectionable content has been a major public policy concern. As a result, countless regulatory schemes and technical solutions have been proposed. But those efforts were largely abandoned over time as policymakers and online safety advocates came to realize that legal hurdles and practical realities meant a new approach to dealing with access to objectionable online content was needed.

Between 2000 and 2010, six major online safety task forces or blue ribbon commissions were formed to study these concerns and consider what should be done to address them, including legal and technical solutions. Three of these task forces were convened by the United States federal government and issued reports in 2000,[9] 2002[10] and 2010.[11] Another was commissioned by the British government in 2007 and issued in a major report in March 2008.[12] Finally, two additional task forces were formed in the U.S. in 2008 and concluded their work, respectively, in December of 2008[13] and July of 2009.[14]

Altogether, these six task forces heard from hundreds of experts and produced thousands of pages of testimony and reports on a wide variety of issues related to online safety. While each of these task forces had different origins and unique membership, what is striking about them is the general unanimity of their conclusions. In particular, the overwhelming consensus of these expert commissions was that there is no single "silver-bullet" technological solution or regulatory quick-fix to concerns about access to objectionable online content. Many of the task forces cited the rapid pace of change in the digital world when drawing that conclusion.

Instead, each of the task forces concluded that education should be the primary solution to most online child safety concerns. Specifically, these task forces consistently stressed the importance of media literacy, awareness-building efforts, public service announcements, targeted intervention techniques, and better mentoring and parenting strategies.

As part of these efforts to strive for "digital citizenship," experts stressed how vital it is to teach both children and adults smarter online hygiene (sensible personal data use) and "Netiquette" (proper behavior toward others), which can further both online safety and digital privacy goals.[15] More generally, as part of these digital literacy and citizenship efforts, we must do more to explain the potential perils of over-sharing information about ourselves and others while simultaneously encouraging consumers to delete unnecessary online information occasionally and cover their digital footprints in other ways.

These education and literacy efforts are also important because they help us adapt to new technological changes by employing a variety of coping mechanisms or new social norms. These efforts and lessons should start at a young age and continue on well into adulthood through other means, such as awareness campaigns and public service announcements.

THE ROLE OF PRIVACY PROFESSIONALS & THE DIGITAL DESIGNERS OF THE FUTURE

Finally, education and digital citizenship efforts are essential not only because they teach consumers how to navigate new information environments and challenges but also because they can guide the actions of current or future *producers* of new digital technologies.

We've spent a great deal of time in recent years encouraging digital innovators to institute "privacy by design" when contemplating their new products. But *real* privacy by design should be a state of mind and a continuous habit of action that influences how designers think about the impact of their products and services before and after creation.

The role of privacy professionals is equally vital. As Deirdre Mulligan and Kenneth Bamberger have noted, increasingly, it is what happens "on the ground"—the day-to-day management of privacy decisions through the interaction of privacy professionals, engineers, outside experts and regular users—that is really important. They stress how "governing privacy through flexible principles" is the new norm.[16]

We should continue to consider how we might achieve "privacy by design" before new services are rolled out, but the reality is that "privacy on the fly" through those "flexible principle" may become even more essential.

CONCLUSION

So, while law and regulation will likely continue to be pursued and, at the margin, may be able to help with egregious privacy and security harms, the reality is that, outside narrow exceptions such as health and financial information, the case for regulatory control becomes harder to justify as the costs will almost certainly exceed the benefits.

That's why it is so essential to have a good backup plan for when control is impossible or simply too costly. Education is the strategy with the most lasting impact. Education and digital literacy provide skills and wisdom that can last a lifetime, enhancing resiliency. Specifically, education can help teach both kids and adults how to behave in—or respond to—a wide variety of situations. Rethinking privacy from the bottom-up and engaging citizens in this way will ultimately serve us better than the top-down approaches being pursued today.

---

[1] Adam Thierer, Mercatus Center at George Mason University, *Public Interest Comment to the Federal Trade Commission in the Matter of The Privacy and Security Implications of the Internet of Things* (May 31, 2013), http://mercatus.org/publication/privacy-and-security-implications-internet-things; Adam Thierer, *Can We Adapt to the Internet of Things*? IAPP PRIVACY PERSPECTIVES (June 19, 2013), https://www.privacyassociation.org/privacy_perspectives/post/can_we_adapt_to_the_internet_of_things.

[2] Adam Thierer, *Let's Not Place All Our Eggs in the Do Not Track Basket*, IAPP PRIVACY PERSPECTIVES, (May 2, 2013), https://www.privacyassociation.org/privacy_perspectives/post/lets_not_place_all_our_eggs_in_the_do_not_track_basket.

[3] Adam Thierer, *The Pursuit of Privacy in a World Where Information Control Is Failing*, 36 HARV. J.L. & PUB. POL'Y 409 (2013).

[4] Ben Adida, *(your) information wants to be free*, Benlog (Apr. 28, 2011, 12:46 AM), http://benlog.com/articles/2011/04/28/your-information-wants-to-be-free.

[5] *Quoted in* Philip Elmer-Dewitt, *First Nation in Cyberspace*, TIME (Dec. 6, 1993), http://www.chemie.fu-berlin.de/outerspace/internet-article.html.

[6] "Moore's Law" refers to a statement by Intel co-founder Gordon Moore regarding the rapid pace of semiconductor technology. Moore stated, "The number of transistors and resistors on a chip doubles every 18 months." *Definition of Moore's Law*, PC MAGAZINE ENCYCLOPEDIA, http://www.pcmag.com/encyclopedia/term/47229/moore-s-law, (last visited June 29, 2013).

[7] Adam Thierer, *A Framework for Benefit-Cost Analysis in Digital Privacy Debates*, GEO. MASON UNIV. L. REV. (*forthcoming*, Summer 2013).

[8] *See generally* ADAM THIERER, PROGRESS & FREEDOM FOUND., PARENTAL CONTROLS & ONLINE CHILD PROTECTION: A SURVEY OF TOOLS & METHODS (Version 4.0) (2009), http://www.pff.org/parentalcontrols.

[9] COPA Commission, REPORT TO CONGRESS (Oct. 20, 2000), www.copacommission.org.

[10] Computer Science and Telecommunications Board, National Research Council, YOUTH, PORNOGRAPHY AND THE INTERNET (2002), http://www.nap.edu/openbook.php?isbn=0309082749.

[11] Online Safety and Technology Working Group, YOUTH SAFETY ON A LIVING INTERNET (June 4, 2010), http://www.ntia.doc.gov/legacy/reports/2010/OSTWG_Final_Report_060410.pdf.

[12] Byron Review, SAFER CHILDREN IN A DIGITAL WORLD: THE REPORT OF THE BYRON REVIEW (Mar. 27, 2008), http://www.education.gov.uk/ukccis/about/a0076277/the-byron-reviews.

[13] Internet Safety Technical Task Force, ENHANCING CHILD SAFETY & ONLINE TECHNOLOGIES: FINAL REPORT OF THE INTERNET SAFETY TECHNICAL TASK FORCE TO THE MULTI-STATE WORKING GROUP ON SOCIAL NETWORKING OF STATE ATTORNEYS GENERAL OF THE UNITED STATES (Dec. 31, 2008), http://cyber.law.harvard.edu/pubrelease/isttf.

[14] PointSmart, ClickSafe, TASK FORCE RECOMMENDATIONS FOR BEST PRACTICES FOR CHILD ONLINE SAFETY (July 2009), http://www.pointsmartreport.org.

[15] Common Sense Media, Digital Literacy and Citizenship in the 21st Century: Educating, Empowering, and Protecting America's Kids 1 (2009), www.commonsensemedia.org/sites/default/files/CSM_digital_policy.pdf; Anne Collier, *From users to citizens: How to make digital citizenship relevant*, NET FAMILY NEWS, (Nov. 16, 2009, 2:23 PM), www.netfamilynews.org/2009/11/from-users-to-citizen-how-to-make.html; Nancy Willard, *Comprehensive Layered Approach to Address Digital Citizenship and Youth Risk Online*, CTR. FOR SAFE & RESPONSIBLE INTERNET USE (2008), *available at* http://digitalcitizen.wikispaces.com/file/view/yrocomprehensiveapproach.pdf.

[16] Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy on the Books and on the Ground*, 63 STAN. L. REV. 247 (2011).

HAS *KATZ* BECOME QUAINT?

USE OF BIG DATA TO OUTFLANK THE FOURTH AMENDMENT

*Jeffrey L. Vagle**

INTRODUCTION

On December 14, 2010, a federal court, upon a government motion, entered an order pursuant to the Stored Communications Act ("SCA") requiring Twitter to turn over to the government subscriber information concerning the accounts of three Twitter users. The order demanded only "non-content" data: names, addresses, and all records of user activity, including dates, times, and IP address data for all subscriber activity since November 1, 2009.*

The subscribers filed a motion to vacate the order on grounds that it was insufficient under the SCA and violated both the First and Fourth Amendments. The motion was denied by the magistrate judge.[1] The subscribers then filed objections to the magistrate judge's ruling.[2] The district judge denied the subscribers' objections, agreeing with the magistrate judge that the subscribers lacked standing to challenge the SCA-based order on non-Constitutional grounds. The court also rejected the subscribers' Fourth Amendment challenge, stating that "any privacy concerns were the result of private action, not government action," and thus the "mere recording of . . . information by Twitter and subsequent access by the government cannot by itself violate the Fourth Amendment."[3]

The problems illustrated by this case are twofold. First, in the age of big data, the collection and analysis of "non-content" data can yield far more information about someone than was thought when the SCA was first drafted.[4] Properly applied, big data analytics can make record data more illuminating to the analyst than content, heightening concerns over reduced SCA protections for non-content data. Second, since this data is collected by third party providers, the government can obtain this data without dealing with Fourth Amendment protections,[5] possibly bypassing the courts altogether.[6] Furthermore, the government's focus on national security since 2001 has resulted in an increase in requests for such data, some of which remain unexamined due to government claims of state secrecy.[7] This essay argues that the nexus of ubiquitous computing and big data analytics has rendered existing standards of Fourth Amendment protection inadequate, and calls for a reexamination of these doctrines based on today's technologies.

MOSAIC THEORY AND THE AGE OF BIG DATA

In recent years, data storage capacities have increased by orders of magnitude, while associated costs have plummeted. Processing speeds have increased to the point that most people carry smartphones that are far more capable than the computers that sat on their desks a few years ago. These factors have combined to enable real time analysis of massive quantities of data, spurring research advances in fields as diverse as atmospheric science, genomics, logistics, and disease prevention.

These capabilities have not gone unnoticed by governments, which have employed big data analytics to reach previously unheard of dimensions of intelligence analysis.[8] These techniques have spilled over into domestic law enforcement, yielding some positive results[9] while at the same time posing new challenges to Fourth Amendment doctrine. And we are the ones supplying the data.

* Mr. Vagle is an associate with Pepper Hamilton LLP. J.D., Temple University Beasley School of Law; B.A., Boston University.

Most Americans own cell phones. We carry them everywhere, and are generally never more than a few feet from one at any time. We use them to send emails and text messages, post messages on Facebook or Twitter, take photos and share them with friends (or the world), and sometimes even to make calls. They are always on, and always on us. Most cell phone users understand that, in order for a cell phone to work, it must be in constant communication with the provider network. The information that is passed back and forth between the phone and the network includes subscriber and location information, and any content that you send or receive. All of this information is collected and stored by the service provider, often without our knowledge.

In fact, providers of all kinds of services make it their practice to collect every bit of data we generate—explicitly or implicitly—and store it for some amount of time.[10] Various privacy laws exist at the state and federal level to prevent the collection of personally identifiable information ("PII"), but big data analytics obviates the need for personal information by leveraging the vast amounts of non-PII data we constantly provide.[11]

The Shrinking Distinction Between "Record" and "Content" Data Under the SCA

The SCA was enacted in 1986, and was intended to extend privacy protections to new forms of telecommunications and computer technology then just emerging, *e.g.*, cell phones and email.[12] The core of the SCA is 18 U.S.C. § 2703, which articulates procedures by which the government may obtain electronic communications and related information. Section 2703 distinguishes between "content" and (non-content) "records," giving greater protection to the content of a communication.

This distinction is based on Congress's original purpose in enacting the SCA. Because Fourth Amendment privacy protections leave gaps when it comes to information sent to—and stored by—third parties,[13] the SCA was enacted to fill those gaps by providing additional statutory privacy rights against government access to information stored by service providers. It was reasoned that users may have a "reasonable expectation of privacy"[14] in the substance of their stored communications ("content"), but would not enjoy the same expectation in non-content ("record") information shared with their service provider.

Thus, if the government seeks access to non-content subscriber records under the SCA, their agents may get this information without a warrant, using either a subpoena or a "specific or articulable facts" order, and are not required to provide notice of this access to the subscriber.[15] But, armed with the ability to perform real-time analytics over vast amounts of this data, the government can make non-content information more illuminating than content information, thus skirting the original intent of the SCA's content/non-content distinction.

Third-Party Doctrine

Under current doctrine, the Fourth Amendment does not prohibit the government from obtaining information revealed to a third party who then conveys that information to government authorities, even if the information was revealed on the assumption that it will be used only for a limited purpose and the confidence placed in the third party will not be betrayed.[16] This third-party doctrine has been the basis for courts holding that information "voluntarily disclosed" to service providers, including IP addresses, files shared on private peer-to-peer networks, and historical cell phone location records, does not have Fourth Amendment protection.[17]

But courts have begun to question the application of this doctrine as applied to current technologies and use patterns. This nascent recognition of the advent of ubiquitous computing, made possible through Internet-enabled laptops, tablets, and smart phones, and the resulting "voluntary disclosure" by millions of Americans of vast amounts of non-content information to third party service providers, has raised concerns that the aggregation and analysis of these enormous data sets may be more revealing than content information. For example, one court has observed that cell service providers "have records of the geographic location of almost every American at almost every time of the day and night," enabling "mass or wholesale electronic surveillance" by the government, and holding therefore that "an exception to the third-party-disclosure doctrine applies to cell-site-location records."[18]

CONCLUSION

As Judge Garaufis recently observed, "[i]n order to prevent the Fourth Amendment from losing force in the face of changing technology, Fourth Amendment doctrine has evolved . . . and must continue to do so."[19] For most Americans, the use of "always on, always on us" technology has become an indispensable part of everyday life, forcing us to accept the fact that private service providers collect the data we constantly generate. Under existing Fourth Amendment doctrine, this non-content data is afforded few protections, even though it may be more revealing than content data. Courts should therefore recognize that our current Fourth Amendment protections must evolve to adapt to the age of big data analytics.

---

[1] *In re § 2703(d) Order*, 787 F. Supp. 2d 430 (E.D. Va. 2011). In her decision, the magistrate judge reasoned that since the order demanded only "records" and not the "contents" of their electronic communications, the subscribers had no standing to challenge the compelled disclosure under the SCA. Further, she held that the subscribers had no First Amendment claim involving non-content information, and they had no legitimate Fourth Amendment expectation of privacy in this information.

[2] *In re Application of the United States*, 830 F. Supp. 2d 114 (E.D. Va. 2011).

[3] *Id*. at 132-33 (citing *U.S. v. Jacobsen*, 466 U.S. 109, 115-17 (1984)).

[4] The statutory definition of "content" is "any information concerning the substance, purport, or meaning of that communication." 18 U.S.C. § 2711(1). The SCA provides greater protection to the "contents of electronic communications" than to their non-content "records." 18 U.S.C. § 2073(a)-(c).

[5] Charlie Savage and Leslie Kaufman, *Phone Records of Journalists Seized by U.S.*, NY TIMES, May 13, 2013. This instance also illustrates the important First Amendment issues at stake.

[6] The government has solicited cooperation and assistance from multiple private companies under the auspices of national security. *See generally* David Kravets, *Court Upholds Google-NSA Relationship Secrecy*, WIRED, May 11, 2012; Brandan Sasso, *Supreme Court Lets AT&T Immunity Stand in Surveillance Case*, THE HILL, Oct. 9, 2012, *available at* http://thehill.com/blogs/hillicon-valley/technology/260951-supreme-court-lets-atat-immunity-stand-in-surveillance-case.

[7] James Bamford, *The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)*, WIRED, Mar. 15, 2012.

[8] Big data analytics is especially useful under the "mosaic theory" of intelligence gathering, which holds that small, disparate items of information, though individually of little or no use, can become meaningful when combined with other items of information by illuminating relationships between the data. Notably, the U.S. government has recognized that mosaic theory can work against them, as well, resulting in increased assertions of state secrecy in denying FOIA requests. *See* David E. Pozen, *The Mosaic Theory, National Security, and the Freedom of Information Act*, 115 YALE L.J. 628 (2005).

[9] *See* Frank Konkel, *Boston Probe's Big Data Use Hints at the Future*, FCW, Apr. 26, 2013.

[10] Private companies maintain differing data retention policies, which can be based on government regulation, data management best practices, or internal procedures.

[11] Location data alone can make someone's life an open book. "GPS monitoring generates a precise, comprehensive record of a person's public movements that reflects a wealth of detail about her familial, political, professional, religious, and sexual associations. The Government can store such records and efficiently mine them for information years into the future. And because GPS monitoring is cheap in comparison to conventional surveillance techniques and, by design, proceeds surreptitiously, it evades the ordinary checks that constrain abusive law enforcement practices: 'limited police resources and community hostility.'" *United States v. Jones*, 132 S. Ct. 945, 955-956 (2012) (J. Sotomayor concurring) (internal citations omitted).

[12] *See generally* Orin S. Kerr, *A User's Guide to the Stored Communications Act, and a Legislator's Guide to Amending It*, 72 GEO. WASH. L. REV. 1208 (2004).

[13] A key reason behind these gaps, third party doctrine, is discussed in more detail below.

[14] The "reasonable expectation" Fourth Amendment test was first articulated in *Katz v. United States*, 289 U.S. 347, 360 (1967) (J. Harlan, concurring), and has recently been "added to" in *Jones* and *Florida v. Jardines*, 133 S. Ct. 1409 (2013).

[15] 18 U.S.C. § 2703(d); 18 U.S.C. § 2703(c)(3).

[16] *United States v. Miller*, 425 U.S. 435, 443 (1976). *See also Smith v. Maryland*, 442 U.S. 735, 743-44 (1979).

[17] *See In re Application of the United States*, 830 F. Supp. 2d at 135 (IP addresses); *United States v. Brooks*, 2012 U.S. Dist. LEXIS 178453, *6-*7 (E.D.N.Y. 2012) (private peer-to-peer networks); *United States v. Graham*, 846 F. Supp. 2d 384, 389 (D. Md. 2012) (historical cell site location records)

[18] *In re United States for an Order Authorizing the Release of Historical Cell-Site Info.*, 809 F. Supp. 2d 113 (E.D.N.Y. 2011). The full reasoning behind the court's decision is beyond the scope of this essay, but it is worth noting the court's closing observation that "the collection of cell-site-location records, without the protections of the Fourth Amendment, puts our country far closer to [Orwell's] Oceania than our Constitution permits." *Id*. at 127.

[19] *In re United States*, 809 F. Supp. 2d at 126.

BIG DATA THREATS

*Felix Wu**

The pros and cons of big data are the subject of much debate. The "pro" side points to the potential to generate unprecedented new knowledge by gathering, aggregating, and mining data, knowledge that can be used for everything from earlier detection of drug side effects to better management of electricity and traffic.[1] The "con" side says that big data raises privacy issues.[2*]

To talk about a big data privacy problem, however, is far too imprecise. In this context, the concept of "privacy" stands for a diverse set of interests. In order to evaluate those interests, weigh them against competing interests, and design appropriate regulatory responses, we need to disentangle them.

Consider the issue of online behavioral advertising, that is, the targeting of advertising based upon one's prior online activity. Perhaps the problem with behavioral advertising is that the tracking technologies that make such advertising possible cause users to feel surveilled as they go about their business, online or off. The problem could also be that stored tracking information might be revealed, to acquaintances or to the government. Alternatively, it might be the targeting itself that is the problem, and that it is wrong to use tracking information to determine what advertisements a person sees.

Similarly, think about the story of Target, which apparently computed a pregnancy prediction score based upon its customers' purchases and used this score to determine to whom to send coupons for baby products.[3] Maybe it makes Target shoppers "queasy" to think that Target is able to predict whether they are pregnant, or even to think that Target is trying to do so.[4] Target's practices might also lead to inadvertent disclosure, as when a father supposedly learned of his daughter's pregnancy for the first time from seeing the coupons Target sent to her.[5] Perhaps it is a problem for pregnant women to get different offers than non-pregnant women. While there might be nothing wrong with targeting baby products to the people who might actually buy them, perhaps differing offers for other products, or on the basis of other predictions, might be more problematic.

In the context of big data in particular, it is helpful to think not just in terms of privacy in general, but in terms of specific privacy threats.[6] When faced with a big data practice, the key question is: "How could this go wrong?" Even for a single practice, that question has many potential answers.

One can conceive of at least three broad categories of big data threats: surveillance, disclosure, and discrimination. By surveillance, I mean the feeling of being watched, which can result from the collection, aggregation, and/or use of one's information.[7] The feeling of being surveilled might be an intrinsic problem, akin to emotional distress. It might also be a problem because such a feeling can affect how people behave, if they start to think twice about the things they do, read, or search for.[8] On this account, one problem with pervasive web tracking is the possibility that people will avoid certain searches or certain sources of information, for fear that doing so inevitably reveals interests, medical conditions, or other personal characteristics they would rather remain hidden.

* Associate Professor, Benjamin N. Cardozo School of Law.

Surveillance can arise from the mere collection of information, as when visits to sensitive websites are tracked. As in the Target example, however, it can also arise from the particular form of processing that the data undergoes. Presumably any unease that customers feel from receiving baby products coupons comes from the sense that Target "knows" about their pregnancy, rather than from knowing that Target has recorded a list of their purchases. Thus, it can be the data mining itself or the characteristic being mined for that converts a mere collection of information into surveillance.

Other problems arise because of the disclosure of data beyond the entity that initially collected it. One disclosure threat might be the nosy employee who looks up people he knows in a corporate database. Another might be an identity thief who successfully hacks into a database. Problems of insecurity are in this sense problems of disclosure. Less maliciously, information might be revealed to people who happen to be nearby and see the ads on another person's computer. Similarly, as in the Target example, people in the same household might see one another's mail. Disclosure to the government is a different potential threat. Government as threat is also not a monolithic one, and could encompass everything from a rogue government employee to a systematic campaign that harasses people on the basis of lawful activity.

Other big data problems are problems of discrimination, that is, treating people differently on the basis of information collected about them. Again, there are many different kinds of discrimination threats. The most obvious might be those based on predicted membership in some protected class, such as race or religion. Some might further object to any discrimination that is correlated with a protected characteristic, whether or not it forms the explicit basis for the targeting. Beyond the traditionally prohibited forms of discrimination, consumers seem also to have a visceral reaction against certain forms of price discrimination.[9] On this view, a problem with big data is its ability to enable highly personalized pricing.

Personalized persuasion is another form of big-data discrimination that might be problematic.[10] The idea here is that rather than simply altering the price or product being sold, the advertiser alters the sales pitch itself so as to best exploit each individual's own cognitive biases.[11] Big data may make it more possible to identify widely shared biases, and this might already be a source of concern. Even if we are willing to tolerate the exploitation of widely shared biases, however, the exploitation of individual biases raises additional concerns about an imbalance of power between advertisers and consumers.

Lack of transparency can in some ways constitute a fourth category of threats. Without transparency, individuals may be stuck in a world in which consequential decisions about them are made opaquely, and over which they are unable to exercise meaningful control.[12] That sense of helplessness, distinct from the feeling of being surveilled, might itself be a problem with big data.

On the other hand, transparency might also be understood as a tool to mitigate some of the other threats identified above. Appropriate transparency could, at least in theory, make it possible for individuals to choose to deal with companies that minimize disclosure risks. Transparency could also diminish the effectiveness of personalized persuasion, again at least in theory.

Even though the word "threat" implies that there is something problematic about the occurrence of the threat, in speaking about threats, I am not necessarily arguing that everything laid out above should in fact be a cognizable threat. One could, for example, hold the view that certain types of discrimination are perfectly acceptable, even desirable. Similarly, some might argue that some of the negative consequences of big data that I have described are not privacy problems at all, but problems of

a different sort.[13] Here, I am not trying to delimit the boundaries of privacy versus other types of harms.

Nor does distinguishing among threats necessarily mean we need distinct regulatory responses. The threat of discrimination might be dealt with by restricting the practice, but it may be far easier to regulate the collection of the relevant information than to detect its misuse.

My goal here instead is simply to catalogue some of the different meanings that people may have when they say that there is a privacy problem with big data. Doing so helps better to frame the big data privacy analysis. It can help us determine when tools like de-identification can be effective at balancing privacy and utility,[14] and it can help us determine in what contexts the benefits outweigh the burdens of big data analysis.

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

[2] *See, e.g.*, Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1748 (2010).

[3] *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES, Feb. 19, 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[4] *Id.* ("If we send someone a catalog and say, 'Congratulations on your first child!' and they've never told us they're pregnant, that's going to make some people uncomfortable . . . . [E]ven if you're following the law, you can do things where people get queasy.").

[5] *Id.*

[6] *See* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1147–48 (2013).

[7] Ryan Calo calls this a "subjective privacy harm." *See* M. Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L.J. 1131, 1144–47 (2011).

[8] *See* Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387 (2008).

[9] *See* Jennifer Valentino-DeVries et al., *Websites Vary Prices, Deals Based on Users' Information*, WALL ST. J., Dec. 24, 2012, at A1.

[10] *See* M. Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. (forthcoming 2014).

[11] *Id.*

[12] Dan Solove has argued that the appropriate metaphor is to Franz Kafka's The Trial. *See* Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393 (2001).

[13] *See, e.g.*, Calo, *supra* note 7, at 1158.

[14] *See generally* Wu, *supra* note 6.

About the Future of Privacy Forum

The Future of Privacy Forum (FPF) is a Washington, DC based think tank that seeks to advance responsible data practices. The forum is led by Internet privacy experts Jules Polonetsky and Christopher Wolf and includes an advisory board comprised of leading figures from industry, academia, law, and advocacy groups.

About the Center for Internet and Society at Stanford Law School

Founded in 2000 by Lawrence Lessig, the Center for Internet and Society (CIS) is a public interest technology law and policy program at Stanford Law School and a part of Law, Science and Technology Program at Stanford Law School. CIS brings together scholars, academics, legislators, students, programmers, security researchers, and scientists to study the interaction of new technologies and the law and to examine how the synergy between the two can either promote or harm public goods like free speech, innovation, privacy, public commons, diversity, and scientific inquiry. CIS strives to improve both technology and law, encouraging decision makers to design both as a means to further democratic values. CIS provides law students and the general public with educational resources and analyses of policy issues arising at the intersection of law, technology and the public interest.