66 STAN. L. REV. ONLINE 55
September 3, 2013

BIG DATA AND ITS EXCLUSIONS

Jonas Lerman*

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

* 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.

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.

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]

---

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.

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.

---

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.

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

---

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.

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]

---

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).

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

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.

doctrine—a principle of *data antisubordination*. Traditionally, U.S. anti-subordination 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, law-making, 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

---

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'").

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

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").

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.

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").