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. Millions of people worldwide remain on big data's periphery. Their data are not regularly collected or analyzed, because they do not routinely engage in the sorts of behaviors big data is designed to capture. Consequently, their preferences and behaviors risk being routinely ignored when governments and private industry use big data to shape public policy and the marketplace. Because big data poses a unique threat to equality, not just privacy, the essay argues that a new "data antisubordination" doctrine may be needed.*

I.

The big data revolution has arrived. Every day, a new book or blog post, op-ed or white paper arrives casting big data, for better or worse, as groundbreaking, disruptive, a game-changer. We are constantly told that big data[1] is fundamentally altering countless aspects of modern life, from medicine to commerce to national security.[2] It may even have the power to change our understanding of existence: in the future, "we will no longer regard our world as a string of happenings that we explain as natural and social phenomena, but as a universe comprised essentially of information."[3]

This revolution has its dissidents. Critics worry the world is increasingly being datafied in ways that ignore, or even smother, the unquantifiable and immeasurable parts of human experience.[4] They warn of big data's other dark sides, too: potential government abuses of civil liberties, erosion of long-held public norms around personal privacy, and even damage to the

---

* Attorney-Adviser, Office of the Legal Adviser, U.S. Department of State. The views expressed in this essay are my own and not necessarily those of the Department of State or the United States government. I thank Paul Schwartz for spurring my interest in this subject.

1. By *big data* I mean the 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 INSTITUTE, BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 1 (June 2011). *See also* CHRIS EATON ET AL., UNDERSTANDING BIG DATA 3 (2012) (explaining that "the term Big Data applies to information that can't be processed or analyzed using traditional processes or tools"). 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* (Jan. 2012), at 3, http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

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

3. *See id.* at 96.

4. David Brooks, for example, frets about the rise of "data-ism." David Brooks, *The Philosophy of Data*, N.Y. TIMES, Feb. 5, 2013, at A23. Similarly, Leonard Weiseltier decries 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.

natural environment (the so-called "server farms" on which big data depends consume huge amounts of energy).[5]

Legal debates over big data tend to focus on the privacy and civil liberties concerns of those people swept up in its net, and on whether existing safeguards such as data minimization, consent, and anonymization provide sufficient protection. It is a perspective of *inclusion*. In many ways, that perspective makes sense. Most people, at least in the industrialized world, experience the impact of, and routinely contribute to, the big data revolution. Under this conception, big data is the whale, and we are all of us Jonah.

This short essay takes a different approach, exploring big data from a perspective of *exclusion*. Big data poses risks also to those persons who are *not* swallowed up by it—that is, whose information is not regularly harvested, farmed, or mined. (Pick your anachronistic metaphor.) Although big data's boosters and critics alike tend to view this revolution as totalizing and universal, the reality is that millions of people remain on its margins, or are excluded from these technological developments altogether, because they do not routinely engage in activities that big data and advanced analytics are currently designed to capture and evaluate.[6]

Rather than considering big data solely from the perspective of the persons whom these technologies most often target, I argue we should consider it also from a perspective of exclusion: Whom does big data omit? What are the consequences of omission, for those individuals, for big data as a technology, and for societies? These underexplored questions deserve greater attention. Furthermore, because big data poses unique dangers to equality, and not just privacy, a new legal doctrine may be needed to protect the groups whom the big data revolution risks leaving behind. I call it data antisubordination.

---

5. *See* Maureen Mackey, *The Dark Side of Today's 'Big Data' Revolution*, FISCAL TIMES, Jan. 17, 2013, *available at* http://www.thefiscaltimes.com/Articles/2013/01/17/The-Dark-Side-of-Todays-Big-Data-Revolution.aspx; Joshua Keating, *The Verizon Records and the Dark Side of 'Big Data'*, FOREIGN POL'Y, June 6, 2013, *available at* http://ideas.foreignpolicy.com/posts/2013/06/06/the_verizon_records_and_the_dark_side_of_big_data; Will Oremus, *Big Data's Dark Side: A Massive, Polluting Drain on the Nation's Power Supply*, SLATE, Sept. 24, 2012, http://www.slate.com/blogs/future_tense/2012/09/24/big_data_pollution_cloud_servers_waste_electricity_on_massive_scale_new_york_times_finds_.html. Viktor Mayer-Schönberger and Kenneth Neil Cukier, who can fairly be characterized as proponents of the big data revolution, recognize big data's dangers, particularly to privacy and liberty. However, they argue that those dangers can be mitigated effectively through regulation. *See* BIG DATA, *supra* note 2, at 150–170; Viktor Mayer-Schönberger & Kenneth Neil Cukier, *Big Data—and Its Dark Side* (Lecture), Harvard Law School, Mar. 6, 2013, video and audio available at http://cyber.law.harvard.edu/events/2013/03/bigdata.

6. These activities include, for example, using the Internet, especially for email, 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."

II.

Big data, for all its technological complexity, springs from a simple idea: by gathering enough information about the past, then applying the right analytical tools to it, users can find connections and correlations, which can then be analyzed to make unusually accurate predictions about the future—how shoppers decide which foods to buy, how terrorists communicate, how a disease spreads. Predictions based on big data already shape decisions made by businesses and governments every day, all over the globe. And if experts are correct, big data's impact on the world is only growing.[7]

If big data, as both an epistemological phenomenon and a new booming industry, increasingly shapes government and business decisionmaking, then one might assume that much attention is paid to who and what shapes big data—the input. In general, however, big data's practitioners express a surprising nonchalance about the precision or provenance of data. In fact, they tend to embrace "messiness" as a virtue.[8] Big data is so big—terabytes, petabytes, exabytes—that the sources or reliability of particular data points cease to be too important. Patterns and trends, not granularity or exactness, are the goal. Datasets need not be pristine. Build a big enough haystack, the thinking goes, and you have a better chance of finding needles; where the hay comes from is less important than how much hay you gather.[9]

Such sentiments presume, however, that the errors creeping into datasets are random and absorbable, and can be factored in to the ultimate analysis. But there is another type of error that can creep into these massive datasets, too: the nonrandom, systemic omission of people who live on datafication's margins, whether due to poverty, geography, or lifestyle. In important sectors, their omission and marginalization risks distorting the datasets and, consequently, skewing the analysis on which private and public actors depend. They are big data's exclusions.

Consider two hypothetical people.

The first is a 30-year-old upper-middle-class woman in New York City. She participates in modern life in all the ways typical of her demographic: smartphone, Google, Netflix, Gmail, Spotify. She uses Facebook, with its

---

7. *See, e.g.*, MCKINSEY GLOBAL INSTITUTE, BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY (May 2011), http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

8. *See* BIG DATA, *supra* note 2, at 32–49.

9. "In a world of small data," write Viktor Mayer-Schönberger and Kenneth Cukier Mayer-Schönberger, "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 that "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.

default privacy settings, to stay in touch with friends. Occasionally she tweets and posts pictures on Instagram and Flickr using geotags. She travels frequently. Her wallet holds a debit card, two credit cards, and a SmarTrip card for New York's subway and bus system. On her keychain, in addition to her house and car keys, are two small plastic barcoded cards for the "customer rewards" programs of her primary grocery and drugstore. In her car, a GPS device sits on the dash, and an E-ZPass transponder (for bridge, tunnel, and highway tolls) hangs from the windshield.

The data she generates every day—and that governments and private industry can harvest from her activities to learn about her and people like her—are nearly incalculable. In addition to data collected by private industry about her spending habits, online activities, and communications, government agencies also know a great deal about her. New York, like London and many other large cities, has transformed itself in the last decade into a supercharged generator of big data. Manhattan alone has a network of some 3,000 closed-circuit television cameras that record terabytes of data every day and are accessible to local and federal law enforcement agencies.[10] The city can also track her movements through her SmarTrip card and FasTrak transponder—data that could be used not only for law enforcement purposes, but to plan new subway schedules and routes, roads, and rates for bridge and tunnel tolls. New York has a dedicated team of quantitative analysts who sift through the terabytes of city data on subjects as diverse as parking, electricity use, commuting habits, children's test scores, fire alarms, and stop-and-frisk statistics.[11]

For our Manhattan subject, avoiding capture by big data is impossible. To begin even to limit her exposure—to curb her contributions to these various dataflows—she would need to fundamentally reconstruct the way she goes about her everyday life. And she would need to move to a new place—a fate anathema to many New Yorkers. Thus, unless she takes relatively drastic steps to limit her contributions to the big data vacuum, our Manhattanite will continue to generate a steady dataflow for government and corporate consumption.

Now consider a second hypothetical person. He lives east of Manhattan in Camden, New Jersey, America's poorest city. He is underemployed, working twenty hours a week at a restaurant, where he is paid under the table, in cash. He has no cell phone, no computer, no cable television. He does not often travel and has no passport, no car, and no GPS. He occasionally

---

10. The New York City Police Department's new Domain Awareness System, a $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.

11. Alan Feuer, *The Mayor's Geek Squad*, N.Y. TIMES, Mar. 23, 2013, at MB1 (describing the work of New York City's Office of Policy and Strategic Planning).

uses the Internet, but only at the local library on a public terminal. When he rides the bus, he pays the fare in cash.

Currently, many of big data's tools are calibrated for our Manhattan subject and people like her—those who generate a large amount of harvestable data in their day-to-day lives. A world shaped by big data will therefore take into account her habits and preferences. But big data currently overlooks our Camden subject almost entirely. (And even he, simply by virtue of living in the United States, would have a larger data footprint than a person living in, say, rural Kazakhstan.) In a future where public policy and the marketplace will be shaped significantly by big data and the predictions it makes possible, the exclusion of poor and marginalized people has troubling implications: for economic opportunity, for social mobility, and even for equal citizenship. These technologies have the potential to create a new form of voicelessness, one in which the preferences and behaviors of poor and otherwise marginalized people receive little or no consideration when companies and governments make decisions, both large and small, about how public institutions and the marketplace should evolve.

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

First, those excluded from the big data revolution may suffer tangible political and economic harms. Businesses will not fully take into account the preferences and behaviors of consumers who do not shop in a way that big data can easily capture, aggregate, and analyze. Stores may not open in their neighborhoods; certain promotions may not be offered to them. Of course, the poor, as well as minority groups, are in many ways already marginalized in the marketplace.[12] But big data could exacerbate such problems, because these groups are even more cut off from electronic commerce than they are from the traditional brick-and-mortar economy.

Politicians and governments, too, may come to depend on big data tools to such a degree that exclusion from the universe of big data equates to exclusion from civic and political life—and a barrier to equal citizenship. Political campaigns, most famously President Obama's in 2008 and 2012, have already come to rely heavily on big data analytics to raise money, plan voter-turnout efforts, and shape their messaging. And big data tools are quickly making the leap from politics to policy: in 2012, for example, the White House announced the National Big Data Research and Development Initiative, which committed more than $200 million to helping federal agencies

12. Grocery stores, for example, are notoriously rare in "minority and low-income areas," both urban and rural—a market failure that has led to phenomena known as "food deserts" and "food swamps." *See* Paul A. Diller, *Combating Obesity with a Right to Nutrition*, 101 GEO. L.J. 969, 971–72 (2013). These phenomena predate the big data revolution.

improve their ability "to access, organize, and glean discoveries from huge volumes of digital data."[13]

Just as U.S. election districts, and thus U.S. democracy, depend on accurate Census data, so too will policy decisions in the future likely depend on the accuracy of big data and advanced analytics. Being left out of government datasets, then, could mean missing out on important government services and public goods. In that way, the big data revolution could trigger broader democracy concerns, and create the possibility of a datafied form of social inequality and subordination.

III.

In the United States, as Justice John Marshall Harlan wrote in his *Plessy v. Ferguson* dissent, "[t]here is no caste" and "no superior, dominant, ruling class of citizens."[14] But big data has the potential to solidify existing inequalities and stratifications and to create new ones. It could reshape societies so that the only people who matter—quite literally the only people who count—are those who regularly contribute to the right dataflows.

As Mayer-Schönberger and others have argued, it seems likely that existing data privacy laws—whether the U.S. patchwork-quilt approach or the more comprehensive European approach—will prove inadequate for managing the privacy risks posed by the big data revolution. But big data is not just a threat to privacy; it also a potential threat to equal citizenship. Existing 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 under existing doctrine, and thus it would do little to ensure that they share in big data's benefits. And equal protection 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."[15]

Moreover, it is not clear what formal equality would even mean in the context of big data. It would be a strange law indeed that compelled public and private consumers of big data to collect *everyone's* information, all the time, in the name of equal protection, or somehow to collect information from different racial and socioeconomic groups proportionally. Two of big

---

13. The White House, Office of Science and Technology Policy, *Big Data Is a Big Deal* (blog post), Mar. 29, 2012, http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal; see also Executive Office of the President, Big Data Across the Federal Government (fact sheet), 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); The White House, Office of Science and Technology Policy, *Unleashing the Power of Big Data* (blog post), Apr. 18, 2013, http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data (providing an update on the progress of the Big Data Initiative).

14. 163 U.S. 537, 559 (1896) (Harlan, J., dissenting).

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

data's supposed built-in safeguards, after all, are anonymization and randomization. To make big data resemble other processes governed by the U.S. equal protection principles (redistricting, for example) would mean requiring collectors of data to determine the race or class of a person before deciding whether to collect the underlying data from him—a sort of double privacy intrusion.

Because existing legal regimes are insufficient to mitigate big data's potential for social stratification and systemic bias, a new antidiscrimination doctrine may be necessary: a principle of data antisubordination. Traditional theorists of antisubordination under U.S. constitutional law 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."[16] This antisubordination approach—what Owen Fiss called the "group disadvantaging principle"[17]—may need to be revised, given big data's potential to impose new forms of stratification and subordination and to reinforce the subordinate status of already-disadvantaged groups.[18]

Under a principle of data antisubordination, groups who exist outside or on the margins of existing dataflows would receive some guarantee that their status—i.e., as persons with light data footprints—would not subject them to unequal treatment by government in the allocation of public services. To be most effective, data antisubordination would need to extend beyond state action. Private players like Google have an outsize influence on societies and a power over communications and the flow of information communication that in previous generations was reserved for governments. Thus, a data antisubordination principle would be incomplete unless it extended, at least to some degree, to the private sector as well. Once fully developed by theorists, the judiciary, and lawmakers, a data antisubordination principle could be enshrined in U.S. law by statute. Like the Genetic Information Nondiscrimination Act,[19] it would be a civil rights law designed to address potential and real threats from powerful new technologies—threats that neither the Framers nor civil rights activists of the past envisioned.[20]

In our age of big data, the "right to be left alone" is an obsolete and insufficient protection. Even modern information privacy principles, such as

---

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

17. *Id.* at 10 (quoting Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 157 (1976)).

18. 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 crisscross the globe.

19. Pub. L. 110-233, 122 Stat. 881 (May 21, 2008).

20. I explore this and related subjects in a forthcoming article. Jonas Lerman, *Democracy and Big Data* (draft on file with author).

consent and the so-called "right to be forgotten," may turn out to have limited utility. Surely new information privacy laws and norms are needed. But they are not sufficient. Ensuring that the big data revolution is a just one—a revolution whose benefits are broadly and equitably shared—may also require, paradoxically, a right *not* to be forgotten.