Big Data's Other Privacy Problem

James Grimmelmann[1]

I. The Modern Memex

We still don't have personal jetpacks or lunar clone colonies, but at least we got the memex. In 1945, Vannevar Bush, writing with the kind of foresight usually reserved for mystics and madmen, sketched a design for the dream desk of the future. Built around a microfilm archive, Bush's design lets the user flip through data at will, following associations and creating new ones. The "intricate web of trails" in the researcher's brain is mapped out in the annotations he makes, creating a permanent record of his discoveries:

> Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified. The lawyer has at his touch the associated opinions and decisions of his whole experience, and of the experience of friends and authorities. The patent attorney has on call the millions of issued patents, with familiar trails to every point of his client's interest. The physician, puzzled by a patient's reactions, strikes the trail established in studying an earlier similar case, and runs rapidly through analogous case histories, with side references to the classics for the pertinent anatomy and histology. The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behavior.[2]

Bush called this device the "memex," but only because he had never seen a Bloomberg Terminal.

In a way, any computer with an Internet connection is a memex that also plays cat videos, but in another, more accurate, way the Bloomberg Terminal is its true spiritual heir. Despite costing $1,500 a month to lease, or perhaps in part because of it, the Bloomberg Terminal is the magic sword that turns mere wheeler-dealers into Masters of the Universe[3] who execute trades with extreme prejudice. Bloomberg users mainline real-time market data for anything anyone anywhere has ever done a deal in, engage in warlock-level feats of technical analysis, and enjoy unequalled access to breaking news, regulatory filings, industry reports, gossip, rumor, innuendo,

---

1 Professor of Law, University of Maryland.

2 Vannevar Bush, As We May Think, Atlantic Monthly, July 1945.

3 *See* Tom Wolfe, The Bonfire of the Vanities 11 (1987).

propaganda, and everything else that might possibly conceivably in some imaginable universe affect the price of something worth buying or selling. The Terminal is absurdly customizable, helping traders track every last penny, piastre, and paisa of their portfolios. It even has restaurant reviews and shopping, both optimized for the 1% of the 1% looking to blow some of the wealth they have just extracted from their fellow man.

The Bloomberg Terminal, in short, is Big Data's elite commando strike force. There have always been quants, with their artisanal small-batch hand-crafted voodoo finance. But the Bloomberg Terminal makes financial necromancy accessible to mere mortals. Imagine yourself granted entry to this Olympus of data, striding here and there, seeing connections, making associations, tasting the ambrosia of insight. And now imagine, if you will, that you are being watched.

II. Hellhound on My Research Trail

It's not normally news when a crack addict stops coming around to his dealer. But when the addict is a partner at Goldman Sachs and the crack is the information flowing from his Bloomberg Terminal, that's news.[4] A reporter thought to ring up Goldman and inquire: is so-and-still with the firm? He hasn't logged in to Bloomberg lately. It was a nice bit of journalistic tradecraft, except for one detail: the reporter worked for Bloomberg News, and knew about the partner's terminal use because Bloomberg News employees had access to it. For years, they had been checking when persons of interest logged in, and what features they accessed.[5] "We were told again and again and again, find ways to use what's on the terminal to write stories." [6]

From there, things went pear-shaped in a hurry for Bloomberg. Goldman's management called around to understand the extent of the snooping.[7] The rest of the press found out, and wrote about the snooping with the gleeful ferocity of an athlete who has just discovered the syringes in his archrival's locker.[8] Regulators from Treasury and the Federal Reserve asked pointed questions about whether their employees had been spied on, too.[9] Bloomberg cut off its reporters' access to terminal usage information, and then, when that failed to stanch the reputational bleeding,

---

4 *See* Mark DeCambre, *Terminally Nosy*, N.Y. POST, May 10, 2013, at 41.

5 *See* Amy Chozick and Ben Protess, *Privacy Breach on Bloomberg's Data Terminals*, N.Y. TIMES, May 10, 2013, at A1.

6 Amy Chozick, *Bloomberg Reporters' Practices Become Crucial Issue for Company*, N.Y. TIMES, June 13, 2013, at A1.

7 *See* Susanne Craig and Jessica Silver-Greenberg, *Hunch About Bloomberg Brought Rivals Together*, N.Y. TIMES, June 1, 2013, at B1.

8 *See., e.g.*, DeCambre, *Terminally Nosey*, *supra* note 4.

9 *See* Amy Chozick and Ben Protess, *More Clients Ask Questions Of Bloomberg*, N.Y. TIMES, May 14, 2013, at B1.

commissioned a nominally "independent" review of its privacy and security standards.[10]

Some have downplayed the privacy implications, pointing out that Bloomberg reporters could see only general information about users' activities, not specific searches and stocks.[11] They have a point, given the steady drumbeat of genuinely serious privacy breaches in the news.[12] But even if in this particular case, Bloomberg's reporters stopped short of the most dastardly deeds they were technologically capable of, we should not let their restraint blind us to the full extent of the dastardry Big Data makes possible.

We are accustomed to speaking about Big Data's privacy concerns in terms of the surveillance it enables of data subjects.[13] Anyone high enough to take a ten-thousand-foot view can see over fences. Take a wide-angle shot, zoom and enhance, and you have a telephoto close-up. But consider now the user of the Bloomberg terminal, zipping from function to function, running down a hunch and preparing to make a killing. Perhaps he correlates historical chart data for energy-sector indices with news reports on international naval incidents in the Pacific Rim. He pulls patterns out of after-hours trading data, checking them against SEC filings and earnings calls. He has a theory, about what happens when certain shipbuilders report their quarterlies—two usually-coupled bond funds briefly diverge—and he stands ready to pocket some cash the next time it happens by exploiting this informational advantage with overwhelming financial force. Tell him that someone has been watching every keystroke, and you will see the blood drain from his face.

---

10 Samuel J. Palmisano, the former C.E.O. of IBM, who was named to lead the review, is on the board of the charitable organization run by Bloomberg's founder, Michael Bloomberg. *See* Amy Chozick, *Former IBM Chief to Lead Bloomberg Privacy Review*, N.Y. Times, May 17, 2013, at B6. The review concluded, unsurprisingly to all, that "Bloomberg has an appropriate Client Data compliance framework in place." Hogan Lovells and Promontory Financial Group, Client Data Policies and Practices of Bloomberg L.P. (2013).

11 *See., e.g.*, William McGeveran, *Privacy and the Bloomberg Terminal*, Concurring Opinions (May 11, 2013), http://www.concurringopinions.com/archives/2013/05/bloomberg-term-pvy.html.

12 This is not to say that Bloomberg or its reporters conducted themselves well. For journalists, data-gathering of this sort is an ethical breach of the first order. They were not engaged in the clever acquisition of information about people's activities in other walks of life. No, they were breaching a trust, using information their own organization solicited for radically different purposes. If you like, you can regard Bloomberg News as having burned hundreds of sources. Or, if you prefer, you can treat it as an offense against readers, like putting GPS trackers in every copy a paperboy delivers. And the corruption—for that is what it was—was truly systemic: hundreds of reporters took advantage of what they called a feature but most of us would call a bug.

13 *See, e.g.*, Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701 (2010);

Or for more a more sympathetic figure, go back to Bush's description of a memex user:

> The owner of the memex, let us say, is interested in the origin and properties of the bow and arrow. Specifically he is studying why the short Turkish bow was apparently superior to the English long bow in the skirmishes of the Crusades. He has dozens of possibly pertinent books and articles in his memex. First he runs through an encyclopedia, finds an interesting but sketchy article, leaves it projected. Next, in a history, he finds another pertinent item, and ties the two together. Thus he goes, building a trail of many items. Occasionally he inserts a comment of his own, either linking it into the main trail or joining it by a side trail to a particular item. When it becomes evident that the elastic properties of available materials had a great deal to do with the bow, he branches off on a side trail which takes him through textbooks on elasticity and tables of physical constants. He inserts a page of longhand analysis of his own. Thus he builds a trail of his interest through the maze of materials available to him.[14]

For "bow and arrow," substitute "genital herpes" or "radical politics." For the historian, substitute a lawyer on a major case or a journalist on the City Hall beat. Think about all the people who might find some use in seeing your research trails: opposing counsel, corrupt police chiefs, lovers, rivals, frenemies, talk radio demagogues, creepy bosses, trolls, self-righteous prudes, kooks and zealots of every stripe, and that one petty-minded neighbor with a lot of grudges and a little time to kill. The memex is—*it is designed to be*—an externalized record of its user's every thought. Using it is like plugging yourself into the most perfect brain scanner ever devised. If you care, even just the slightest bit, about your intellectual freedom, then you, I submit, are very interested in who has access to your memex and the memories it holds.[15] Big Data puts the world at your fingertips—so with Big Data, your fingerprints are everywhere.

Big Data's other privacy problem is like its first privacy problem, and also unlike it. Subject privacy is about actions: Big Data knows what you did last summer. User privacy is about thoughts: Big Data knows you watched *I Know What You Did Last Summer*. With enough such data points, it can make a pretty good guess what you're likely to do next summer. Google searches have been used to convict murderers; how long before they're used as evidence of pre-crime?[16] Oh, wait, they already are: the

---

14 Bush, *supra* note 1.

15 *See, e.g.*, Neil Richards, *Intellectual Privacy*, 87 Tex. L. Rev. 387 (2008).

16 *See, e.g.*, Declan McCullagh, *Police Blotter: Web Searches Lead to Murder Conviction*, CNET (Feb. 12, 2010), http://news.cnet.com/8301-13578_3-10452471-38.html.

NSA tells analysts to find targets by identifying people "searching the web for suspicious stuff."[17] Search-query signature strikes cannot be far off.

Indeed, Big Data is recursive: it tends inevitably to convert its users into its subjects. How does Google map flu trends? Not by testing people for infection, but seeing who searches for information about flu.[18] Every visitor to the Land of Data leaves a little of herself behind. Every query is further grist for the mill. Even Vannevar Bush, writing in 1945, bless his prescient and naive heart, understood this much:

> The historian, with a vast chronological account of a people, parallels it with a skip trail which stops only on the salient items, and can follow at any time contemporary trails which lead him all over civilization at a particular epoch. There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record. The inheritance from the master becomes, not only his additions to the world's record, but for his disciples the entire scaffolding by which they were erected.

He might have added that we will all be trailblazers, that the greatest value of the memex—its permanent and globally shared record of every user's research trails—is precisely its greatest curse. We are the spiders spinning Big Data's web of knowledge, and we are also the flies trapped in it.[19] He who works with data should look to it that he himself does not become data. And when you gaze long into Big Data, Big Data also gazes into you.[20]

---

17 *See, e.g.*, XKEYSCORE 15 (NSA PowerPoint Presentation Feb. 25, 2008), *available at* http://www.theguardian.com/world/interactive/2013/jul/31/nsa-xkeyscore-program-full-presentation.

18 *How Does This Work?*, GOOGLE.ORG FLU TRENDS (2011), http://www.google.org/flutrends/about/how.html.

19 Another computer visionary, Douglas Engelbart, noted as an aside during his legendary 1968 demo that "an advantage of being online is that it keeps track of who you are and what you're doing all the time." Presentation of Douglas Engelbart at the 1968 Fall Joint Computer Conference, Dec. 9, 1968, *available at* http://www.youtube.com/watch?v=VScVgXM7lQQ&list=PL76DBC8D6718B8FD3.

20 For another example, consider scientific publisher Elsevier's purchase of the citation-management system Mendeley. To its academic users, Mendeley is a mini-memex: enabling them to craft and share associational trails. To Elsevier, those trails are a commodity—or perhaps an input into a copyright-enforcement system. *See, e.g.*, David Dobbs, *When the Rebel Alliance Sells Out*, THE NEW YORKER ELEMENTS BLOG (Apr. 12, 2013), http://www.newyorker.com/online/blogs/elements/2013/04/elsevier-mendeley-journals-science-software.html; David Banks, *The Mendeley Dilemma*, CYBORGOLOGY (Jan. 22, 2013), http://thesocietypages.org/cyborgology/2013/01/22/the-mendeley-dilemma/.

III. Mutually Assured Privacy Destruction

There is another way of understanding the relationship between Big Data subjects and Big Data users. The fact that users also have privacy interests at stake complicates the project of protecting subject privacy. To understand the problem, it helps to understand something of the debate over how what to do about safeguarding those whose personal information has been hoovered up at terabyte scale.

For a time, it appeared that no restrictions on use might be necessary because there were no data subject privacy interests at stake. Deidentification was the watchword of the day: it was thought that some simple scrubbing—stripping a dataset of names, ranks, and serial numbers—would render these data driftnets dolphin-safe. And the database wranglers would have gotten away with it, too, if it hadn't been for those meddling computer scientists.[21] Personal information always contains something unique. It expresses its singularity even in an IP address, and a very modest grade of data has in it something irreducible, which is one man's alone. That something he may be reidentified from, unless there is a restriction in access to the database. Although there is a lively dispute about where to draw the balance between the needs of the many (as data subjects) and the needs of the many (as research beneficiaries), it is by now painfully clear that some such balance must be struck.[22]

The next line of defense, implicit in the burgeoning discourse of Big Data boosterism, is that only incorruptible researchers who are pure of heart will be plowing through the piles of data in search of ponies. Epidemiologists are the poster children, perhaps because public health officials would never, ever jump to conclusions about poorly understood diseases sweeping through their communities. This ideal of a trusted elite priesthood of data analysts bears an uncanny similarity to National Rifle Association head Wayne LaPierre's invocation of "good guys with guns."[23] When Big Data is outlawed, only outlaws will have Big Data. Actuaries and supply chain optimizers, perhaps, come close to this technocratic ideal.

But Big Data today is probably better embodied by marketers and hedge-fund traders, two professions not known for their generous concern for human flourishing. It is hard to feel sanguine about the Big Swinging Dicks[24] who brought us the

---

21 *See* Ohm, *supra* note 10 (summarizing reidentification literature).

22 *Compare id*. (reidentification = big deal) *with* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV.J.L. & TECH. 1 (2011) (reidentification = big whoop). *See also* Felix T. Wu, *Privacy and Utility in Data Sets*, U. COLO. L. REV. forthcoming 2013); Daniel Barth-Jones, *The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now* (June 4, 2012), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397.

23 *See* Speech to NRA Convention by Wayne LaPierre (May 4, 2013), *available at* http://home.nra.org/pdf/waynelapierre_130504.pdf.

24 *See* MICHAEL LEWIS, LIAR'S POKER 56–57 (1989).

subprime financial Chernobyl or about ad men in the business of running A/B tests to optimize their manipulation of consumers' cognitive biases.[25] Any sufficiently advanced marketing technology is indistinguishable from blackmail. The global phishing industry shows what happens when confidence men scale up their scams.

And all of this is to say nothing about Carnivore, Total Information Awareness, PRISM, EvilOlive, and the other ominously-named trappings of the National Surveillance State.[26] Give the CIA six megabytes of metadata inadvertently emitted by the most honest of men, and it will find something in them to put him on the drone kill list. One might—as the Obama Administration asks—simply trust in the good faith and minimal competence of the Three Letter Agencies that brought us extraordinary rendition, COINTELPRO, and the Clipper Chip. Or, more realistically, one might question the wisdom of creating comprehensive fusion centers accessible to every vindictive cop with a score to settle.[27]

Thus, since Big Data cannot be entirely defanged and its users cannot be entirely trusted, it becomes necessary to watch them at work. It seems like a natural enough response to the problem of the Panopticon. Subject privacy is at risk because Big Data users can hide in the shadows as they train their telescopes not on the stars but on their neighbors. And so we might say, turn the floodlights around: ensure that there are no dark corners from which to spy. We would demand audit trails—permanent, tamper-proof records of every query and computation.

But if we are serious about *user* privacy as well as about *subject* privacy, transparency is deeply problematic. The audit trails that are supposed to protect Big Data subjects from abuse are themselves a perfect vector for abusing Big Data users.[2] Indeed, they are doubly sensitive, because they are likely to contain sensitive information about *both* subjects and users. The one-way vision metaphor of the Panopticon, then, is double-edged. Think about glasses. A common intuition is that mirrorshades are creepy, because the wearer can see what he chooses without revealing where his interest lies. Everyone is up in arms about the Google Glass-holes who wear them into restrooms. But the all-seeing Eye is a window to the soul. The

---

25 *See* Joseph Turow, The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth (2013).

26 *See* Jack M. Balkin, *The Constitution in the National Surveillance State*, 93 Minn. L. Rev. 1 (2008).

27 *See* Danielle Keats Citron and Frank Pasquale, *Network Accountability for the Domestic Intelligence Apparatus*, 62 Hastings L.J. 1441 (2011).

28 *See, e.g.*, Letter from Dione J. Stearns, Assistant General Counsel, Federal Trade Commission, to Ginger McCall (Sept. 25, 2012) (withholding information about Google's privacy assessments from release under the Freedom of Information Act). To be sure, Google requested that the "confidential and proprietary information" be withheld it was "competitively sensitive," but at a finer level of granularity, such a report would include precise and sensitive details of how Google employees use their access to its massive databases. *See generally* Yakowitz, *Data Commons*, *supra* note __, at 17–20 (describing agencies' use of personal privacy arguments to deny FOIA requests).

Segway for your face is also a camera *pointed directly at your brain* that syncs all its data to the cloud. The assumption Glass users are making, presumably, is that no one else will have access to their data, and so no one else will be pondering what they're pondering. But that's what Bloomberg Terminal users thought, too.

This leaves meta-oversight: watching the watchmen. Audit trails don't need to be public; access to them could be restricted to a small and specialized group of auditors. But this privacy epicycle introduces complications of its own. You have a security problem, so you audit your users. Now you have *two* security problems: you are committed to safeguarding and watching over not just your data, but your data about how your data is being used. Whoever looks through the logfiles will be able to gain remarkable insight into users' methods and madnesses. Yes, the auditors will be looking for suspicious access patterns, but they'll need to have access to the full, sensitive range of information. You wouldn't want an insider trading scandal in which an auditor piggybacked on an analyst's research, or a auditor who picks a favorite user and turns into a stalker. Your auditors, in other words, are *also* Big Data users, which means that they too will have to be audited. It's watchmen all the way down.

IV. Crowdsourcing and Power

This convergence between Big Data's two privacy problems brings home the degree to which the Big Data story is a story of centralized control. It is the accumulation of large repositories of data that makes comprehensive surveillance possible, on both sides. When data is lying about raw, in the wild, any forager can pluck some, but the risks are necessarily limited. When that same data has been harvested, processed, and warehoused, two things change. First, it becomes far more threatening to the subjects, precisely because the accumulation of details harmless in themselves can make patterns evident. And second, it becomes necessary, to restrict access to the data: not just anyone can be allowed to run queries against it with anonymous impunity.

That is, centralizing data disempowers both subjects and users. They are both now subject to the policies—or, perhaps, "whims"—of whatever entity controls the dataset. Data ownership is power, of a peculiarly feudal ilk. The data barons of Silicon Valley struggle to ensure that data about people comes to rest in their own servers. Whomsoever would work these vast tracts of data to harvest insight must do the data barons homage. By what feudal incidents does one become a data vassal? Socage, of course—payment in coin—but also data service—giving up yet more information about oneself. The synoptic view that Big Data affords is not, it turns out, something that will be widely shared. That, the barons will hoard for themselves.

Big Data doesn't just grow on trees; it isn't natural or inevitable. It describes a particular configuration of institutional relationships among data's subjects, users, and owners, one in which concatenation and concentration give a small set of actors

disproportionate power to determine who knows what about whom.[29] This is a form of ideology: that there are invaluable insights available in datasets so large they can only be effectively managed by centralized repositories. This ideology of Big Data is explicitly used to justify overriding the privacy concerns of its subjects,[30] and it has the side effect of putting its users in a position where they are subject to the observation and control of the data barons. Not just privacy law, but intellectual property law, contract law, unauthorized access law, and many other bodies of decidedly non-neutral doctrine are used to create a world in which data is always collected and rarely distributed.[31] This *novus ordo datorum* serves the interests of autocrats, bureaucrats, and secret police; it appeals to data-addicted technologists and venture capitalists on the prowl for their next Internet-scale score. But—for all the insights that Big Data offers—we can question whether its arrangements are as good for the rest of us.

Perhaps there is another way. Consider a different possible ideal for managing our relationship to information. Call it Small Data, or Local Data, or Slow Data, or Sustainable Data, or perhaps Democratic Data—enough to go around, enough for everyone to have some. Not perhaps, a lot, but enough. Enough not to be beholden to anyone else, enough to participate meaningfully in society, enough that no one can take away your dignity. Democratic Data is not the opposite of Big Data: these yeoman dataholders will join together their datasets at times to serve the public good, but they will be ever mindful of the risk of data tyranny.[32] Alexander Hamilton would have loved Big Data, with its Enlightenment ambitions, brutally rational economics, and awe-inspiring centralized power. Democratic Data is more of a Thomas Jefferson kind of idea—civic, romantic, uplifting, a little contradictory, and faintly impractical. Perhaps this might mean giving up the Bloomberg Terminal, or dividing out its functions and sharing access to some of them more widely. But would we want to live in a world that sets these ideals aside, rather than seeking, however fitfully and imperfectly, to realize a Republic of Data, where all men and

---

29 *See* Julie Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904, 1918–27 (2013).

30 *See, e.g* Yakowitz, *Data Commons*, *supra* note __, at 66 (calling use of personal data for public-benefiting research projects "he tax we pay to our public information reserves" and arguing that "if taken to the extreme, data privacy can make discourse anemic and shallow").

31 Subject privacy is, in large measure, a creature of tort law. Tort duties supposedly keep us safe from GPS trackers and upskirt drones; tort law hangs like the sword of Damocles over Big Data controllers who guard their super-sized datasets with subpar security (even if it never actually seems to drop). But user privacy is almost entirely a creature of contract law. The boilerplate contracts that let users inside armor-plated data silos spell out what can, or less often can't, be done with the security-camera footage. If they sell your record of searches for panda gangbangs to the FSB, or your detailed map of flu trends and your interest in viral genetics to the FBI, they'll say you have only yourself to blame for clicking where it said "Click here to agree."

32 *Cf.* Omer Tene and Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013).

women are created equal, and are endowed by their databases with certain unalienable rights, among them life, liberty, and the pursuit of privacy?