RESPONSE

THE UNDERWHELMING BENEFITS OF BIG DATA

PAUL OHM†

In response to Paul M. Schwartz, *Information Privacy in the Cloud*, 161 U. PA. L. REV. 1623 (2013).

The cloud is a hodgepodge, and Paul Schwartz, in his rich Article, *Information Privacy in the Cloud*,[1] tackles many different parts of the confusing combination, giving meaning to mush in his characteristically careful style. Consider his thoughts on the changes being wrought to information privacy law by the move to "networked intelligence in the cloud."[2] This expression refers, at least in part, to what others have been calling "Big Data," the trendy moniker for powerful new forms of data analytics.[3] Professor Schwartz weighs the benefits of Big Data techniques against the risks they pose to privacy. Better than some others, he takes care to point to the benefits that truly matter. Too many commentators have too often overstated the benefits of Big Data, inflating studies and praising the merely trivial.

If I do not acknowledge this near the outset, some will misinterpret or misrepresent the point of my Response, claiming falsely that I do not agree with the following patently true prediction: Big Data will lead to important benefits. Whether applied to crises in medicine, in climate, in food safety, or in some other arena, Big Data techniques will lead to significant, new, life-enhancing (even life-saving) benefits that we would be ill advised to

† Associate Professor, University of Colorado Law School.

1 Paul M. Schwartz, *Information Privacy in the Cloud*, 161 U. PA. L. REV. 1623 (2013).

2 *See id.* at 1644-47 (comparing the European Union's expansionist approach to defining personal information with the United States' reductionist approach).

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).

electively forego. This prediction is both obvious and somewhat uninteresting. It flows from the definitional squishiness surrounding the phrase: "Big Data" has become nearly synonymous with "data analysis," and data analysis is a lynchpin of modern science. To argue against Big Data is to argue against science. That is not my brief.

But some Big Data projects will also lead to bad outcomes, like invasions of privacy and hard-to-detect invidious discrimination. Big Data techniques can help governments spy on their citizens and criminals prey on their victims. As we worry about these negative consequences, and particularly as we consider whether we might forego or shape some forms of Big Data so as to limit their negative effects, we must weigh the associated costs and benefits. In doing so, we should scrutinize carefully claims that the benefits of Big Data outweigh the costs to individuals and society. Too often, when Big Data's cheerleaders talk about its benefits, they blur the significant with the trivial, the important with the frivolous. Big Data's benefits are real and important; we should give less attention to those benefits that are not.

Big Data's touted benefits are often less significant than claimed and less necessary than assumed. Professor Schwartz, however, is not really guilty of overstating Big Data's virtues. The benefits he touts focus in particular on "information-based forms of health research [that] 'have led to significant discoveries, the development of new therapies, and a remarkable improvement in health care and public health.'"[4] He cites sources that herald new discoveries resulting from advanced data analysis in the treatment of breast cancer, Alzheimer's disease, and thrombosis.[5] These benefits seem unimpeachably important, even if the specifics remain somewhat underdeveloped in Schwartz's Article.[6]

Consider in stark contrast a project Schwartz does not mention, but one that has often served as the poster child for the positive benefits of Big Data: Google Flu Trends.[7] Flu Trends is a project that Google's philanthropic arm,

---

4 *See* Schwartz, *supra* note 1, at 1631 (quoting INST. OF MED. OF THE NAT'L ACADS., BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 113 (Sharyl J. Nass et al. eds., 2009)).

5 *See id.* at 1631-32 & n.39 (citing INST. OF MED. OF THE NAT'L ACADS., *supra* note 4; Gina Kolata, *Rare Sharing of Data Led to Results on Alzheimer's*, N.Y. TIMES, Aug. 13, 2010, at A1).

6 Despite his thorough use of examples, Schwartz sometimes falls prey to the tendency to accept the benefits of analytics too conclusorily. For example, he concludes with very little discussion that the European Union's Proposed Data Protection Regulation "creates a potential threat to socially productive uses of analytics—and ones that do not raise significant risks of individual privacy harms." *Id.* at 1647.

7 GOOGLE FLU TRENDS, http://www.google.org/flutrends/ (last visited July 6, 2013).

Google.org, launched in 2008.[8] To test the theory that one might predict the parts of the world suffering from flu outbreaks by watching the symptoms people type into the Google search engine, Google gave its internal researchers access to its users' search queries.[9] It turns out the theory works, and Google reports that it can detect likely flu outbreaks a week or two faster than the physician-reporting surveillance efforts of the Centers for Disease Control and Prevention (CDC).[10] To showcase the project, Google publishes an interactive website displaying maps that reveal flu outbreaks around the world, color coding cities, states, and nations according to the estimated prevalence of the virus, in hues ranging from reassuring greens to ominous reds.[11] Almost nobody has anything bad to say about Flu Trends.[12] It represents the triumph of Big Data over illness and potential death, abetted by pervasive surveillance.

But let us think about Flu Trends a bit more critically and balance its presumed benefits against its privacy-related costs. I am not saying that Flu Trends is an evil experiment that ought to be swept from the earth; I am saying only that we should not assume that it represents a major advance for human health without giving it much more critical scrutiny than it has received. Focus first on its costs to privacy. Here is the bill of particulars against Flu Trends: Google breached a wall of trust by dipping into its users' private search data in ways that went beyond traditional and historically accepted uses for search query data, such as those uses relating to security, fraud detection, and search engine design. While Google's users likely would have acquiesced had Google asked them to add "help avoid pandemics" or "save lives" to the list of accepted uses, they never had the chance for

---

8 *See* Google.org, *Tracking Flu Trends*, THE OFFICIAL GOOGLE.ORG BLOG (Nov. 11, 2008, 1:14 PM), http://blog.google.org/2008/11/tracking-flu-trends.html (explaining the methodology employed by Google to track the spread of influenza).

9 *See id.*

10 *See id.* Then again, Google's method might not work so well after all. Recent reports suggest that Flu Trends overestimated the 2012-13 flu season. *See* Declan Butler, *When Google Got Flu Wrong*, 494 NATURE 155, 155 (2013), *available at* http://www.nature.com/news/when-google-got-flu-wrong-1.12413; David Wagner, *Google Flu Trends Wildly Overestimated This Year's Flu Outbreak*, ATLANTIC WIRE (Feb. 13, 2013), http://www.theatlanticwire.com/technology/2013/02/google-flu-trends-wildly-overestimated-years-flu-outbreak/62113. Some are already using this error as a cautionary tale to illustrate the limits of Big Data. *See, e.g.*, Nick Bilton, *Disruptions: Data Without Context Tells a Misleading Story*, N.Y. TIMES BITS (Feb. 24, 2013, 11:00 AM), http://bits.blogs.nytimes.com/2013/02/24/disruptions-google-flu-trends-shows-problems-of-big-data-without-context (noting that "[d]ata inherently has all of the foible of being human").

11 *See* GOOGLE FLU TRENDS, *supra* note 7.

12 One notable exception is the Electronic Privacy Information Center, which has raised privacy concerns about Google Flu Trends from the outset. See *Google Flu Trends and Privacy*, ELECTRONIC PRIVACY INFO. CENTER, http://epic.org/privacy/flutrends (last visited July 6, 2013) (outlining the privacy concerns implicated by Flu Trends).

a public conversation. Instead, the privacy debate was held—if at all—within the walls of Google alone. By breaching the public's trust, Google has expanded researchers' ability to examine our search queries and given them a motive to focus in particular on some of the most sensitive information about us, our medical symptoms.

But why focus on costs to privacy, and why call for a public debate, if we all agree that the benefits of Flu Trends outweigh the probably slight costs to privacy? On the contrary, as far as I can tell—and I have spoken to Google employees who have not refuted this understanding—the project produced exactly two things: those pretty, color-coded maps on Google's website, and a publication in the journal *Nature* for a few Google employees.[13]

People seem to assume that the color-coded maps provide a tangible benefit to public health—that somehow these charts and the data they represent can save lives. But because Google does not want to impinge too much on privacy, it aggregates its released results. A user can see the prevalence of the flu in the United States or Colorado or Denver, going back a few years, but cannot examine the data at the level of census block or ZIP code.[14] This is very good for privacy, but how does this aggregated, limited release save lives? Who has created an app, therapy, or epidemiological study based on the colors on this map?[15] Has a traveler ever avoided boarding a plane to a city on a distant coast because of the relative difference in the shading of the oranges between home and destination? The answer, I suspect, is that none of these positive results has occurred. Instead, the project's primary mission is to market Google: we are reminded by a colorful map that Google is not evil.

I hope Google is doing more with the data behind the scenes, but I suspect it is not. As far as I know, it has not shared the data with the CDC on a much more granular basis.[16] If it did, the project would impinge more, not less, on privacy, but at least this would provide a more worthy justification for the violation. If this type of data sharing is not occurring, it seems that

---

13 *See* Jeremy Ginsberg et al., *Detecting Influenza Epidemics Using Search Engine Query Data*, 457 NATURE 1012 (2009). One of the six authors, Lynnette Brammer, appears not to work for Google and instead lists the CDC as her institutional affiliation.

14 *See* GOOGLE FLU TRENDS, *supra* note 7.

15 Apparently, Aetna has used Google Flu Trends data to "prepare itself financially" for worse-than-usual flu seasons. Damon Poeter, *Aetna CEO: We Use Google to Track Flu Outbreaks*, PCMAG.COM (Jan. 23, 2013), http://www.pcmag.com/article2/0,2817,2414629,00.asp.

16 This is not to say that the CDC is not paying attention to the publicly available Flu Trends data. *See* Butler, *supra* note 10, at 156 (quoting Lyn Finelli, head of the CDC's Influenza Surveillance and Outbreak Response Team, as saying, "I'm in charge of flu surveillance in the United States and I look at Google Flu Trends and Flu Near You all the time, in addition to looking at US-supported surveillance systems").

Google's data jocks are not using the data to save lives. Rather, they are merely flexing their analytic muscles, exploiting their privileged access to our secrets in order to win fancy publications.

At the very least, the fact that Google had a good idea for how to learn more about health by invading privacy a tiny bit—through breaching the boundaries of what Julie Cohen calls "semantic discontinuity"[17]—should give Google no claim of higher moral standing over companies, like DuckDuckGo, that promise to leave our privacy intact.[18] It is wrong to reward a company for defying the privacy expectations of its users simply because it does so for a good cause.

Big Data's proponents point to more than just Flu Trends to celebrate their cause. Omer Tene and Jules Polonetsky have written a pair of articles that urge a relaxation or reorientation of privacy law to help unleash Big Data, and these articles recite a series of Big Data's triumphs.[19] While some of their examples justly deserve celebration, others do not.

In addition to praising Flu Trends, Tene and Polonetsky commend another group of health researchers at Stanford who peeked at search query logs. These researchers discovered that two drugs, Paxil and Pravachol, interact in a previously unknown way, increasing blood glucose to diabetic levels.[20] The researchers confirmed this result, in part, by noting that people who searched for the names of the two drugs tended also to search for symptoms like "headache" or "fatigue," consistent with the "symptomatic footprint" of diabetes-inducing drugs.[21] But as Tene and Polonetsky point out, the researchers developed their hypothesis through traditional methods, by searching an FDA database of adverse events, and used the novel privacy-invasive step only

---

17 *See* JULIE E. COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE 239 (2012) (defining "semantic discontinuity" as "a function of interstitial complexity within the institutional and technical frameworks that define information rights and obligations and establish protocols for information collection, storage, processing, and exchange").

18 *See* Michael Rosenwald, *Ducking Google in Search Engines*, WASH. POST, Nov. 9, 2012, http://articles.washingtonpost.com/2012-11-09/business/35505935_1_duckduckgo-search-engine-search-results (profiling DuckDuckGo, a search engine that does not track users or generate search results based on their previously expressed interests); *Privacy*, DUCKDUCKGO, https://duckduckgo.com/privacy (last updated Apr. 11, 2012) ("DuckDuckGo does not collect or share personal information.").

19 *See* Tene & Polonetsky, *supra* note 3, at 245-51 (noting the "big benefits" of Big Data in the areas of healthcare, mobile, smart grid information, traffic management, retail, fraudulent payments, and online data); Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63, 64 (2012), http://www.stanfordlawreview.org/online/privacy-paradox/big-data ("The uses of big data can be transformative, and the possible uses of the data can be difficult to anticipate at the time of initial collection.").

20 *See* Tene & Polonetsky, *supra* note 3, at 245-46.

21 *Id.* at 245.

to support the hypothesis.[22] Scientific confirmation is, of course, an important step, but it does change the weight we might give to the benefit of the research.

Rank ordering the supposed benefits of Big Data by decreasing significance, Tene and Polonetsky seem to understand the speciousness of some of the other benefits they herald. After summarizing the benefits related to medical research, they discuss the analysis of mobile phone records to better the lives of people living in slums in developing countries[23] and the use of smart grids to monitor and control electricity usage, leading to better energy conservation.[24] Both are important applications. From there, they begin to slide into much less significant territory, touting benefits that pale in comparison to medical research or benefits that might have been achieved without invading to such an extent the privacy of data subjects.

With Big Data, Tene and Polonetsky note, Wal-Mart can better manage its inventory, Amazon can sell a more tailored product, and payment card companies can detect fraud.[25] These are important to the economy, to be sure, but they seem like the kind of benefits the market can provide through upfront exchange rather than over the objections of the users whose data is mined.

The point is that we need to be much more nuanced in analyzing what we gain in return for invasions of privacy. The benefits of Big Data are real, even if to date a bit unrealized, so we should focus on true benefits and stop talking about minimally interesting results. To help us shift our attention to Big Data's true benefits, consider the following rules of engagement, which distinguish between different classes of Big Data's benefits.

First, we should separate benefits built upon data sets that are full of information about people from those built upon data that has almost nothing to do with personal information, and thus almost nothing to do with personal privacy.[26] Big Data techniques can unlock mysteries of manufacturing, climate change, financial markets, and cybersecurity without delving into data at the individual level. We should be mindful, however, that sometimes data that seems not to involve individuals will often reveal

---

22 *Id.*

23 *See id.* at 247.

24 *See id.* at 248.

25 *See id.* at 249-51 (noting the benefits Big Data provides with respect to retail, online shopping, and electronic payments).

26 *See* Schwartz, *supra* note 1, at 1655 (suggesting exempting from EU data protection law the "mere automation of processing choices," which does not involve "decisions about the individuals whose personal data it is processing").

individual information through inference, a topic I have investigated in prior work.[27]

One potential benefit of this distinction is that it might help us direct researchers toward the kind of Big Data studies that do not threaten privacy. There are enough vitally important problems that we can solve in a manner consistent with individual privacy that it seems a shame that many researchers devote so much energy to privacy invasion.[28]

Second, we should recognize that many of the benefits we care most deeply about, including most medical research, originate in research institutions with an established track record of respecting personal privacy.[29] For example, hospitals and other entities that hold electronic health records *should* collaborate with computer scientists to study their large datasets in order to come to new and better results.[30] But they should do this within the ambit of responsible research; conducted by trained researchers; and subject to controls and protocols of trust and monitoring, most importantly those rules designed to protect data subjects, such as human subjects review and the Common Rule.[31]

But as medical research follows the lead of Google Flu Trends and begins to slip outside these traditional institutions and their concomitant safeguards, we should be concerned about the relative lack of controls. Particularly as more medical research is conducted by profit-driven companies—whether large corporations or small startups—we should worry about forcing the public to accept new risks to privacy with little countervailing benefit and none of the controls. The worst of all worlds would occur if

---

27 *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1716-27 (2010) (noting the ease with which "release-and-forget anonymization" can be undone).

28 In the words of Jeff Hammerbacher, former research scientist for Facebook, "The best minds of my generation are thinking about how to make people click ads . . . . That sucks." Ashlee Vance, *This Tech Bubble Is Different*, BLOOMBERG BUSINESSWEEK, Apr. 14, 2011, http://www.businessweek.com/magazine/content/11_17/b4225060960537.htm.

29 Professor Schwartz's examples of beneficial analytics focus largely on health researchers who, presumably, undergo scrutiny by institutional review boards. *See* Schwartz, *supra* note 1, at 1631-32 & nn.38-43.

30 A prime example is the successful collaboration between researchers at the University of California, Riverside and a physician at Children's Hospital Los Angeles to analyze data collected from the hospital's pediatric intensive care units. *See Using Big Data to Save Lives*, PHYS.ORG (Oct. 22, 2012), http://phys.org/news/2012-10-big.html.

31 The Common Rule, which regulates research involving human subjects, outlines, among other topics, basic provisions for institutional review boards, informed consent, and assurances of compliance. *See* Protection of Human Subjects, 45 C.F.R. pt. 46 (2011); U.S. Dep't of Health & Human Servs., *Federal Policy for the Protection of Human Subjects ('Common Rule')*, HHS.GOV, http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html (last visited July 6, 2013) (describing the current U.S. system of protection for human research subjects).

medical researchers at non-profit institutions began to clamor for relaxed human subjects review in a race to the bottom to compete with their for-profit counterparts.

Third, we should distinguish between research that benefits the public and that which serves only narrow and private gain. Google Flu is less a public health success than a well-executed marketing campaign. This marketing creates not only a general public relations benefit for Google, but also gives Google and others an argument for why companies should be able to monitor behavior more often, store the corresponding data for longer periods of time, and analyze that data for purposes inconsistent with expectations.

This is not to say that only non-profit or public institutions can benefit the public good through Big Data, but it does mean that we should expect research produced by private institutions and built upon the private secrets of users to give something back to the public in exchange, perhaps in the form of new therapies or drugs. And in demanding meaningful returns for the public good, we should not confuse for science the kinds of daily trivia—blurbs and tweets and infographics—that ricochet around the web and die shortly thereafter. Too often, our internet culture does little more than titillate, producing "results" that allow us to feel like armchair scientists and social scientists when, in reality, we are doing little more than playing voyeur. Infographics are the best example of this phenomenon: infographics are the effluent of the information society, transmitting small, amusing facts about the human experience, but doing little else. If the only tangible benefit the public receives from research built upon the secrets of users is a series of infographics, then the harms of that research may well outweigh the gains.

Big Data is coming, like it or not. We have an opportunity to shape it, to ensure it operates *for* us, not *on* us. The coming debate over whether and how we might do this promises to be a vigorous one. Let us have that debate, in a frank and honest way, agreeing at the outset to focus only on what really matters.

Preferred Citation: Paul Ohm, Response, *The Underwhelming Benefits of Big Data*, 161 U. PA. L. REV. ONLINE 339 (2013), http://www.pennlawreview.com/responses/8-2013/Ohm.pdf.