



The Human and Ethical Aspects of Big Data

Grady Booch

FOR THE 2011 census, India began a grand experiment wherein every citizen was to be photographed, fingerprinted, and questioned regarding their marital status, education, and occupation. Mind you, such data gathering is not unusual—nations have long conducted censuses—but the scale to which India chose to undertake such a vast digitization of information was unprecedented.

In earlier times, the Roman Empire was particularly detailed in the

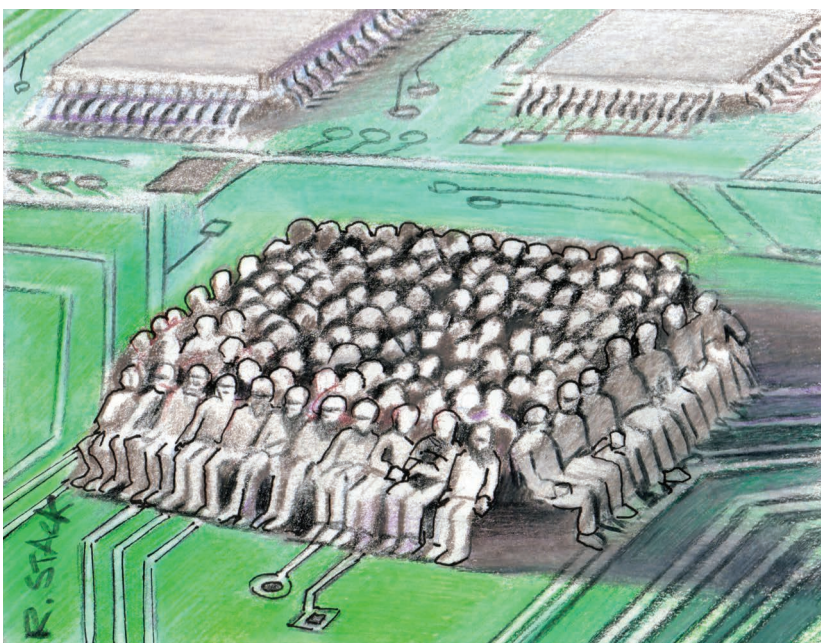
matter, as were the Normans, as reflected in the *Domesday Book*. In the US, Article 1, Section 2 of the Constitution empowers Congress to carry out a census. Data obtained from the census are protected under Title 13 of the US Code, a provision of which calls for census information to be kept private for 72 years. It's particularly interesting to study how the questions asked by that census have changed over the years. In 1790, we counted the number of free

white males as well as the number of slaves in a given household. In the census of 1890, questions included, "Is the person defective of mind, sight, hearing, or speech?" and "Can the person read or write?" In 1990, the questions reflected a very different culture: "What time did this person usually leave home to go to work?" and "What was this person's total income?" For 2000, partly in reaction to the growth of the immigrant population, we asked, "Is this person a citizen of the United States?" And in 2010, we asked even more directly, "Is Person 1 of Hispanic, Latino, or Spanish origin?"

Use and Misuse

Nations use censuses for a variety of honorable reasons, especially as they can relate to revenue planning and social policy. All too often, however, that data is misused. In May 1943, just five months after the bombing at Pearl Harbor, all persons of Japanese ancestry were forced into internment camps in the US, even if they were legitimate citizens. How did we identify such persons? From the census of 1940, although that personalized data was—by law—intended to be kept private.

Norbert Weiner—from whom we get the term "cyber"—observed,



“To live effectively is to live with adequate information.”¹ True, but he also noted, “The penalties for errors of foresight, as great as they are now, will be enormously increased as automatization comes into full use.”² Weiner’s point of view arose in the context of World War II and the Cold War, but as Google’s Eric Schmidt has suggested far more recently, we are in that very place that Weiner predicted.

Schmidt observed, “From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days.”³ Paul Ohm, a lawyer and prolific commentator on big data, has made clear the unintended consequences of this explosion in data: “Databases will grow to connect every individual to at least one closely guarded secret.”⁴ And yet, as Microsoft’s Kate Crawford notes, this leads us to the state of big data fundamentalism, “the idea that with larger data sets, we get closer to objective truth.”⁵

Confidence in Big Data

We can speak of some things concerning big data with confidence. For one, the domains for nonpersonal big data collections are extensive. The Large Hadron Collider, the Square Kilometer Array, and meteorological data are just three examples of domains drawn from the physical world for which prodigious volumes of information exist. The domains for personal big data are equally extensive. The US presidential election in 2012 was an intensely data-driven activity, and the very business models of Facebook, Amazon, Google, eBay, Wal-Mart, and so many others rely on big data from which an individual is both a user as well as an object to be mon-

itized. That these companies are using personal data shouldn’t come as a surprise to any user of such services: we receive value at the cost of (what we hope is) rational collection and analysis of the data our activity generates. The problem, however, is

tice Commission has made it clear that privacy is a basic human right. Equally, the Obama administration (in “Consumer Data Privacy in a Networked World”) states that “consumers have a right to expect that companies will collect, use,

The choices we make in acting upon that data have very real and very human consequences.

that data is often collected innocuously from activity in the physical world—our phone calls, the movement of our smartphones and wired transportation, electrical use as measured by smart meters—and then combined with other information for use beyond our choice. It’s at this intersection of the possible and desirable that the moral and ethical use of big data become unclear.

This leads us to suggest that there are some things for which we can’t yet speak confidently about big data. How do we best develop the models that lie behind our data? Who owns data? Who takes responsibility when the assumptions we make about data are violated at any point along its life cycle?

The Consequences of Big Data

As an insider to computing, I know that data is morally neutral. However, as a person living a very human life in a very human culture, I also know that the means by which we analyze data, as well as the choices we make in acting upon that data have very real and very human consequences. The EU Jus-

and disclose personal data in ways that are consistent with the context in which consumers provide that data.”⁶ Still, this leaves room for considerable interpretation.

As insiders to computing—and especially if we choose to become members of certain professional organizations—we take on the responsibility to make good ethical choices. The International Council on Systems Engineering Code of Ethics states, “The practice of systems engineering can result in significant social and environmental benefits, but only if unintended and undesired effects are considered and mitigated.” The IEEE Code of Ethics states, “We, the members of the IEEE, in recognition of the importance of our technologies in affecting the quality of life throughout the world ... do hereby commit ourselves to technical and professional conduct.” The ACM Software Engineering Code of Ethics is even more direct: “Software engineers have significant opportunities to do good or to cause harm, to enable others to do good or cause harm, or to influence others to good or cause harm.”

Big Data for Good

During the London cholera outbreak of 1854, John Snow mapped the incidents of cholera around the city. His visualization made it evident that there was a root cause—a single, contaminated water pump—and shutting it down signaled the beginning of the end of that terrible scourge. Thanks to the work of Craig Venter and the National Institute on Aging, we now know how to sequence the human genome. The costs of doing so continue to plummet, from millions of dollars per sequence to hundreds of thousands to close to \$1,000 per human today. It won't be long before we can quickly sequence an individual's DNA for \$100. At that price point, the economics of big data genomics changes. Should cross-sectional studies of such data be permitted? Should parents be allowed to genetically engineer their children? What would you do if such analysis deter-

mined your unborn child had a debilitating genetic defect?

Charles Babbage lived in a world in which statistical study was largely uncharted—the Royal Statistical Society prided itself on its 1842 study of “Bastardy in England and Wales as indicated by the Registers of Births.” Today, in our Internet of Things, we gather data on all sorts of stuff. For example, we maintain extensive records on vehicle registration, and we're on the cusp of gathering unprecedented amounts of information for every vehicle, in real time. However, the National Rifle Association in the US—the one organization that probably holds the greatest amount of information about gun ownership—is the same organization that resists any efforts to establish a national gun registry. To be exceedingly clearly, I won't take a political stand either way here, but it does raise the public policy issue of big data: what are the limits that culture or politics can or should place on the life cycle of data?

In 1590, Tycho Brahe and Johannes Kepler got into a very personal battle about access to data. Brahe had collected vast amounts of information about the movement of the planets, but had little mathematical skill to make sense of it. Kepler, on the other hand, was a most clever scientist, but he had no data. Happily (for Kepler, not Brahe), the situation resolved itself after Brahe's death. Today, in the realm of targeted advertising—the very life blood of companies such as Google—we see a similar battle over the ownership of data. What data can companies legitimately collect? How does your answer change if that data point represents you? What if that data point represents your child?

The Amish people are known

for their gentle ways. In the face of encroaching technology, as it has been since their culture took root, they try to be intentional about the use of new technology. As a group of Amish leaders recently observed, “It's not just how you use the technology that concerns us. We're also concerned about what kind of person you become when you use it.”

As software professionals, it isn't enough to simply “do no evil.” Rather, our calling is to do that which brings the most good. However, that's not at all easy, and it is a calling that each of us must face, individually and corporately. ☞

References

1. N. Wiener, *The Human Use of Human Beings: Cybernetics and Society*, Doubleday, 1956.
2. N. Wiener, *God & Golem, Inc.*, MIT, 1966.
3. R. Smolan and J. Erwit, *The Human Face of Big Data*, Against All Odds Productions, 2012.
4. P. Ohm, “Don't Build a Database of Ruin,” *Harvard Business Review*, blog, 23 Aug. 2012; <http://blogs.hbr.org/2012/08/dont-build-a-database-of-ruin>.
5. Q. Hardy, “Why Big Data Is Not Truth,” *New York Times*, blog, 1 June 2013; <http://bits.blogs.nytimes.com/2013/06/01/why-big-data-is-not-truth>.
6. “Consumer Data Privacy in a Networked World,” white paper, The White House; www.whitehouse.gov/sites/default/files/privacy-final.pdf

GRADY BOOCH is an IBM Fellow and one of the UML's original authors. He's currently developing *Computing: The Human Experience*, a major trans-media project for public broadcast. Contact him at grady@computingthehumanexperience.com.

IEEE
Software

FIND US ON
**FACEBOOK
& TWITTER!**

[facebook.com/
ieeesoftware](http://facebook.com/ieeesoftware)

[twitter.com/
ieeesoftware](http://twitter.com/ieeesoftware)



See [www.computer.org/
software-multimedia](http://www.computer.org/software-multimedia)
for multimedia content
related to this article.