FORBES > INNOVATION > ENTERPRISE & CLOUD

The Government-Academia Complex and Big Data Religion

Gil Press Senior Contributor
I write about technology, entrepreneurs and innovation.

Follow

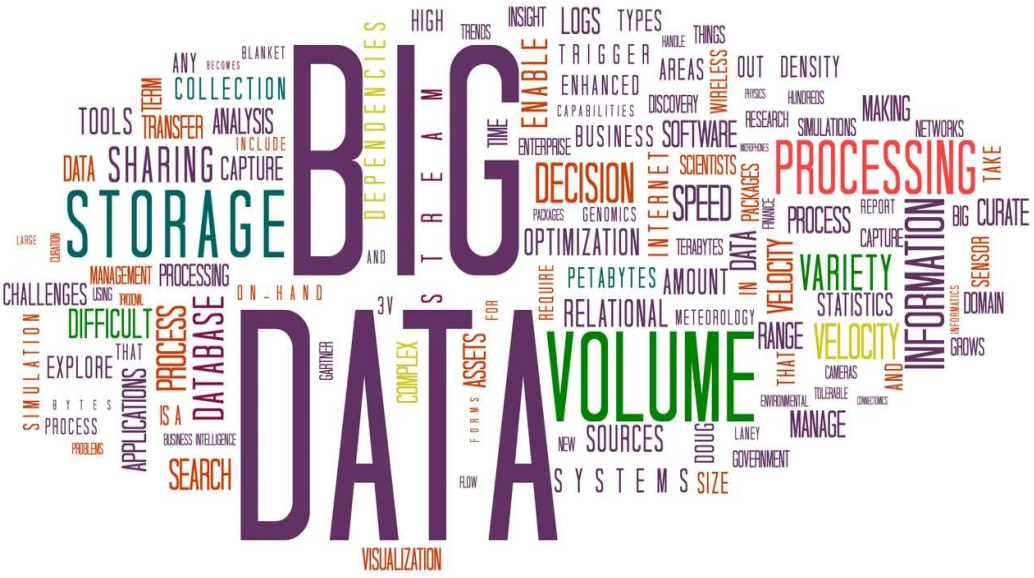
Sep 9, 2014, 11:09am EDT

This article is more than 9 years old.

This was the summer of our discontent with big data. First came the news of the Facebook experiment manipulating the emotions of almost 700,000 of its users in the name of big data "science." Then the *Guardian* told us about similar studies paid for by DARPA, the advanced research arm of the Department of Defense (DoD), in which researchers communicated with "unwitting participants in order to track and study how they responded." [Update: see DARPA's Fact Sheet for its response to the story] And new NSA-related revelations continued to pop up throughout the summer, including a *Washington Post* investigation revealing that ordinary Internet users far outnumber legally targeted foreigners in the communications intercepted by the NSA from U.S. digital networks and Snowden telling *Wired* about MonsterMind, an NSA cyberwarfare program accessing virtually *all* private communications coming in from overseas to people in the U.S.

The dubious ethics of the experiment conducted by Facebook's data scientists and by the DoD-funded researchers is what mostly riled the critics. The questionable constitutionality of the NSA

surveillance has been at the center of that debate. The critics in both cases, however, rarely point to the common factor that makes federal government employees (whether at the NSA or DARPA), academics, and private sector data scientists believe they are not violating any ethical and legal norms: They have been brainwashed by the religion of big data.



Source: Camelia.Boban, Wikimedia

Bigger is better and data possesses unreasonable effectiveness. The more data you have, the more unexpected insights will rise from it, and the more previously unseen patterns will emerge. This is the religion of big data. As a believer, you see ethics and laws in a different light than the non-believers. You also believe that you are part of a new scientific movement which does away with annoying things such as making hypotheses and the assumptions behind traditional statistical techniques. No need to ask questions, just collect lots of data and let it speak.

Kate Crawford writes that "there is a tendency in big data studies to accord merit to massive sample sizes, regardless of the importance of the question or the significance of the findings. But if there's something we've learned from the emotional contagion study, a large number of participants and data points does not necessarily produce good research."

It is not good research for many reasons (see here, for example), but possibly the most important one is that it is based on too much data. Rich Morin, Senior Editor at the Pew Research Center, writes that "studies based on supersized samples can produce results that are statistically significant but at the same time are substantively trivial. It's simple math: The larger the sample size, the smaller any differences need to be to be statistically significant—that is, highly likely to be truly different from each other. ... And when you have an enormous random sample of 689,003, as these researchers did, even tiny differences pass standard tests of significance. That's why generations of statistics teachers caution their students that 'statistically significant' doesn't necessarily mean 'really, really important.'"

No matter. Big data believers ignore the boundaries and limitations of traditional statistical techniques and make sweeping claims such as "Facebook and the data it has, has been able to advance social psychology by quantum leaps over the past decade."

If you would like to find out more about those "quantum leaps," check out Kashmir Hill's survey of the findings of studies of social networks. These include: you're more likely to spread information if you can see friends sharing it; people like spreading rumors; outrageous stuff travels farther and faster than the debunking of that outrageous stuff; we're thinking things that we don't put down to digital paper.

Given what one reads on social networks, the last finding is a bit surprising. Still, it—and the findings of similar studies—certainly don't qualify as "quantum leaps in social psychology." To Hill's "WTF score"—rating the severity of privacy intrusion by these studies—we may add a "stupid science score," rating the banality of the studies' findings or how "substantially trivial" they are.

As I discussed in a previous post, the banality and triviality of the findings does not prevent the big data priests from claiming they have discovered the "mathematical laws that govern society." Still, this kind of preaching works and now you can get your paper into top journals simply because you did your experiment on 700,000 people. Looks like the sheer number of the subjects in the study is the new overriding criterion for judging the merits of a scientific paper.

If you think that this can only happen in the social sciences because they have never had the same rigorous criteria used by the natural sciences, you are not aware of the widespread impact of the religion of big data. Newsweek's Megan Scudellari recently published an excellent survey of arguments for and against the belief that "gathering data first and asking questions second is a new, exciting way to make discoveries about the natural world," a belief attributed by Scudellari to David Van Essen, lead investigator of the $40 million NIH-funded Human Connectome Project (HCP) and Philip Bourne, associate director for data science at the National Institutes of Health, among others.

I'm sure that the big data priests know very well that Charles Darwin said that "without speculation there is no good and original observation" and that Albert Einstein observed that "it is the theory that determines what we can observe." But why bring up the ancients? They didn't have the tools and technologies we have today, the big data priests would say, so they had to limit themselves to speculation and theory. Now we are charting a new scientific "paradigm," where we collect first and ask questions

later, but only if the data has not already given us the answers. Because big data is today's Oracle of Delphi, it speaks to us and reveals the "unknown unknowns."

The key issue, in my mind, with all this talk about "scale," "automating science," and the "unreasonable effectiveness of data," is indeed the issue of "effectiveness." Scudellari quotes J. Anthony Movshon, a neuroscientist at New York University: "The idea that you should collect a lot of information because somewhere in this chaff is a little bit of wheat is a poor case for using a lot of money."

The danger of wasting a lot of money in the fanatical pursuit of more and more data becomes particularly alarming when we realize that big data has become the nexus of the government-academia complex. Unlike the military-industrial complex, which Eisenhower acknowledged the "imperative need" for its development (while warning about its "grave implications"), the main reason for the flourishing of the government-academia complex has been the pursuit of "big," as in "big money," "big government," and "big science," with academia providing the government with more and more reasons to become bigger and, in exchange, benefiting from more and more government funding and employment opportunities. Big data religion has recently become a significant accelerator of these decades-long trends.

A good example (one of many) is the Big Mechanism program of the Department of Defense. The description of the $42 million program (see here and here and here) is fascinating, as the following quotes illustrate:

"The first challenge the Big Mechanism program intends to address is cancer pathways, the molecular interactions that cause cells to become and remain cancerous. The program has three primary technical areas: Computers should read abstracts and papers in cancer biology to extract fragments of cancer pathways.

Next, they should assemble these fragments into complete pathways of unprecedented scale and accuracy, and should figure out how pathways interact. Finally, computers should determine the causes and effects that might be manipulated, perhaps even to prevent or control cancer."

"Although the domain of the Big Mechanism program is cancer biology, the overarching goal of the program is to develop technologies for a new kind of science in which research is integrated more or less immediately—automatically or semi-automatically—into causal, explanatory models of unprecedented completeness and consistency. Cancer pathways are just one example of causal, explanatory models."

"The collection of big data is increasingly automated, but the creation of big mechanisms remains a human endeavor made increasingly difficult by the fragmentation and distribution of knowledge. To the extent that the construction of big mechanisms can be automated, it could change how science is done."

How effective is it for the resource-constrained Department of Defense to provide academics with millions of dollars to change how science is done? Does this program overlap with other government-funded or privately-funded efforts to understand better cancer biology? What happens if after 42 months (the promised lifespan of the program) Big Mechanism ends with no (automatic?) determination of causes and effects in the domain of cancer pathways?

Google developed big data tools because it needed to index the *entire* Web and sampling was no option. That happened to be true for many other Web-based businesses and their specific Data mining needs. As Peter Skomoroch generalizes from his experience at LinkedIn (where he developed its "People You May Know" feature): "Many features and signals can only be observed by collecting massive amounts of data (for example, the

relationships across an entire social network), and would not be detected using smaller samples. Processing large datasets in this manner was often difficult, time consuming, and error prone before the advent of technologies like MapReduce and Hadoop, which ushered in a wave of related tools and applications now collectively called big data technologies."

In some situations, more data is better than a smaller amount of data. In other situations, smaller data (i.e., a carefully constructed sample) is better than lots of data. And in yet other situations, intuition (based on experience or a specific talent) is better than data, big or small.

But this is heresy to big data believers in government and academia, often aided and abetted by private-sector data scientists. Big data preachers, no matter where they work, more often than not are simply drunk with the power and promise of the "revolution" they help bring to society, science, business, anything that will be supposedly "improved" with more data. This is why they don't see any problem with manipulating people's emotions or breaking the law, all in the name of the effectiveness of big data.

Big data religion is powered by the authority in our society of Silicon Valley. It is Silicon Valley that also has the power (and the knowledge) to temper the exaggerated claims, to explain where more data may be of help and where it is a waste of money. It can help stop the government-academia complex from being further inflated by big data.

Eisenhower's warning and admonition are still valid today, although they apply to a different complex in a different context: "The potential for the disastrous rise of misplaced power exists, and will persist. We must never let the weight of this combination endanger our liberties or democratic processes. We should take nothing for granted."

Follow me on Twitter
@GilPress or Facebook or Google+

Gil Press

Follow

I'm Managing Partner at gPress, a marketing, publishing, research and education consultancy. Previously, I held senior marketing and research...

Read More