What We Can Learn From the Epic Failure of Google Flu Trends

David Lazer and Ryan Kennedy Wired 2015 October

EVERY DAY, MILLIONS of people use Google to dig up information that drives their daily lives, from how long their commute will be to how to treat their child's illness. This search data reveals a lot about the searchers: their wants, their needs, their concerns—extraordinarily valuable information. If these searches accurately reflect what is happening in people's lives, analysts could use this information to track diseases, predict sales of new products, or even anticipate the results of elections.

In 2008, researchers from Google explored this potential, claiming that they could "nowcast" the flu based on people's searches. The essential idea, published in a paper in Nature, was that when people are sick with the flu, many search for flu-related information on Google, providing almost instant signals of overall flu prevalence. The paper demonstrated that search data, if properly tuned to the flu tracking information from the Centers for Disease Control and Prevention, could produce accurate estimates of flu prevalence two weeks earlier than the CDC's data—turning the digital refuse of people's searches into potentially life-saving insights.

And then, GFT failed—and failed spectacularly—missing at the peak of the 2013 flu season by 140 percent. When Google quietly euthanized the program, called Google Flu Trends (GFT), it turned the poster child of big data into the poster child of the foibles of big data. But GFT's failure doesn't erase the value of big data. What it does do is highlight a number of problematic practices in its use—what we like to call "big data hubris." The value of the data held by entities like Google is almost limitless, if used correctly. That means the corporate giants holding these data have a responsibility to use it in the public's best interest.

In a paper published in 2014 in Science, our research teams documented and deconstructed the failure of Google to predict flu prevalence. Our team from Northeastern University, the University of Houston, and Harvard University compared the performance of GFT with very simple models based on the CDC's data, finding that GFT had begun to perform worse. Moreover, we highlighted a persistent pattern of GFT performing well for two to three years and then failing significantly and requiring substantial revision.

The point of our paper was not to bury big data—our own research has demonstrated the value of big data in modeling disease spread, real time identification of emergencies, and identifying macro economic changes ahead of traditional methods. But while Google's efforts in projecting the flu were well meaning, they were remarkably opaque in terms of method and data—making it dangerous to rely on Google Flu Trends for any decision-making.

For example, Google's algorithm was quite vulnerable to overfitting to seasonal terms unrelated to the flu, like "high school basketball." With millions of search terms being fit to the CDC's data, there were bound to be searches that were strongly correlated by pure chance, and these terms were unlikely to be driven by actual flu cases or predictive of future trends. Google also did not take into account changes in search behavior over time. After the introduction of GFT, Google introduced its suggested search feature as well as a number of new health-based add-ons to help people more effectively find the information they need. While this is great for those using Google, it also makes some search terms more prevalent, throwing off GFT's tracking.

The issue of using big data for the common good is far more general than Google—which deserves credit, after all, for offering the occasional peek at their data. These records exist because of a compact between individual consumers and the corporation. The legalese of that compact is typically obscure (how many people carefully read terms and conditions?), but the essential bargain is that the individual gets some service, and the corporation gets some data.

What is left out that bargain is the public interest. Corporations and consumers are part of a broader society, and many of these big data archives offer insights that could benefit us all. As Eric Schmidt, CEO of Google, has said, "We must remember that technology remains a tool of humanity." How can we, and corporate giants, then use these big data archives as a tool to serve humanity?

Google's sequel to GFT, done right, could serve as a model for collaboration around big data for the public good. Google is making flu-related search data available to the CDC as well as select research groups. A key question going forward will be whether Google works with these groups to improve the methodology underlying GFT. Future versions should, for example, continually update the fit of the data to flu prevalence—otherwise, the value of the data stream will rapidly decay.

This is just an example, however, of the general challenge of how to build models of collaboration amongst industry, government, academics, and general do-gooders to use big data archives to produce insights for the public good. This came to the fore with the struggle (and delay) for finding a way to appropriately share mobile phone data in west Africa during the Ebola epidemic (mobile phone data are likely the best tool for understanding human—and thus Ebola—movement). Companies need to develop efforts to share data for the public good in a fashion that respects individual privacy.

There is not going to be a single solution to this issue, but for starters, we are pushing for a "big data" repository in Boston to allow holders of sensitive big data to share those collections with researchers while keeping them totally secure. The UN has its Global Pulse initiative, setting up collaborative data repositories around the world. Flowminder, based in Sweden, is a nonprofit dedicated to gathering mobile phone data that could help in response to disasters. But these are still small, incipient, and fragile efforts.

The question going forward now is how build on and strengthen these efforts, while still guarding the privacy of individuals and the proprietary interests of the holders of big data.