Harvard
Business
Review

Web-Based Technologies

Google Flu Trends' Failure Shows Good Data > Big Data

by Kaiser Fung

March 25, 2014

In their best-selling 2013 book *Big Data: A Revolution That Will Transform How We Live, Work and Think*, authors Viktor Mayer-Schönberger and Kenneth Cukier selected Google Flu Trends (GFT) as the lede of chapter one. They explained how Google's algorithm mined five years of web logs, containing hundreds of billions of searches, and created a predictive model utilizing 45 search terms that "proved to be a more useful and timely indicator [of flu] than government statistics with their natural reporting lags."

Unfortunately, no. The first sign of trouble emerged in 2009, shortly after GFT launched, when it completely missed the swine flu pandemic. Last year, *Nature* reported that Flu Trends overestimated by 50% the peak Christmas season flu of 2012. Last week came the most damning evaluation yet. In *Science*, a team of Harvard-affiliated researchers published their findings that GFT has over-estimated the prevalence of flu for 100 out of the last 108 weeks; it's been wrong since August 2011. The *Science* article further points out that a simplistic forecasting model—a model as basic as one that predicts the temperature by looking at recent-past temperatures—would have forecasted flu better than GFT.

In short, you wouldn't have needed big data at all to do better than Google Flu Trends. Ouch.

In fact, GFT's poor track record is hardly a secret to big data and GFT followers like me, and it points to a little bit of a big problem in the big data business that many of us have been discussing: Data validity is being consistently overstated. As the Harvard researchers warn: "The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis."

The amount of data still tends to dominate discussion of big data's value. But more data in itself does not lead to better analysis, as amply demonstrated with Flu Trends. Large datasets don't guarantee valid datasets. That's a bad assumption, but one that's used all the time to justify the use of and results from big data projects. I constantly hear variations on the "N=All therefore it's good data" argument, from real data analyts: "Since Google has 80% of the search market, we can ignore the other search engines. They don't matter." Or, "Since Facebook has a billion accounts, it has substantively everyone."

Poor assumptions are neither new nor unpredictable. When the mainstream economists collectively failed to predict the housing bubble: their neoclassical model is built upon several assumptions including the Efficient Markets Hypothesis, which suggests that market prices incorporate *all* available information, and, as Paul Krugman says, leads to the "general belief that bubbles just don't happen."

In the wake of epic fails like these, the natural place to look for answers is in how things are being defined in the first place. In the business community, big data's definition is often some variation on McKinsey's widely-circulated big data report (PDF), which defines big data as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.”

Can we do better? I started asking myself and other data analysts what are the key differences between datasets that underlie today’s GFT-like projects and the datasets we were using five to 10 years ago. This has led to what I call **the OCCAM framework**, a more honest assessment of the current state of big data and the assumptions lurking in it.

Big data is:

O*bservational*: much of the new data come from sensors or tracking devices that monitor continuously and indiscriminately without design, as opposed to questionnaires, interviews, or experiments with purposeful design

Lacking ***C****ontrols:* controls are typically unavailable, making valid comparisons and analysis more difficult

Seemingly ***C****omplete:* the availability of data for most measurable units and the sheer volume of data generated is unprecedented, but more data creates more false leads and blind alleys, complicating the search for meaningful, predictable structure

A*dapted:* third parties collect the data, often for a purposes unrelated to the data scientists’, presenting challenges of interpretation

M*erged*: different datasets are combined, exacerbating the problems relating to lack of definition and misaligned objectives

This is far less optimistic a definition, but a far more honest appraisal of the current state of big data.

The worst outcome from the *Science* article and the OCCAM framework, though, would be to use them as evidence that big data’s “not worth it.” Honest appraisals are meant to create honest progress, to advance the discipline rather than fuel the fad.

Progress will come when the companies involved in generating and crunching OCCAM datasets restrain themselves from overstating their capabilities without properly measuring their results. The authors of the *Science* article should be applauded for their bravery in raising this thorny issue. They did a further service to the science community by detailing the difficulty in assessing and replicating the algorithm developed by Google Flu Trends researchers. They discovered that the published information about the algorithm is both incomplete and inaccurate. Using the reserved language of academics, the authors noted: "Oddly, the few search terms offered in the papers [by Google researchers explaining their algorithm] do not seem to be strongly related with either GFT or the CDC data—*we surmise that the authors felt an unarticulated need to cloak the actual search terms identified*." [emphasis added]

In other words, Google owes us an explanation as to whether it published doctored data without disclosure, or if its highly-touted predictive model is so inaccurate that the search terms found to be the most predictive a few years ago are no longer predictive. If companies want to participate in science, they need to behave like scientists.

Like the Harvard researchers, I am excited by the promises of data analytics. But I'd like to see our industry practice what we preach, conducting honest assessment of our own successes and failures. In the meantime, outsiders should be attentive to the challenges of big data analysis, as summarized in the OCCAM framework, and apply considerable caution in interpreting such analyses.

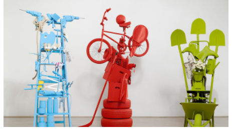
KF

Kaiser Fung is founder and CEO of Principal Analytics Prep, a next-generation data science

bootcamp based out of the HBS Startup Studio. He directed the MS in Applied Analytics at Columbia University, and is the creator of Junk Charts, a blog devoted to the critical examination of data visualization in the mass media. His latest book is *NumberSense: How to Use Big Data to Your Advantage*. He holds an MBA from Harvard Business School, and degrees from Princeton and Cambridge Universities, and was an analytics leader at Vimeo, SiriusXM Radio and American Express.

Recommended For You

How to Launch a Successful Portfolio Career



Why You Should Build a "Career Portfolio" (Not a "Career Path")

Why Facebook Messenger Is a Big Deal for Customer Service

Why Tesco's Strengths Are No Longer Good Enough