The Parable of Google Flu: Traps in Big Data Analysis

D. Lazer, R. Kennedy, G. King, and A. Vespignani

Science, vol. 343, March 14, 2914, pp. 1203-1205

In February 2013, **Google Flu Trends (GFT)** made headlines: *Nature* reported that GFT was predicting ***more than double*** the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC)

CDC data is considered to be ground truth

-- based on surveillance reports from labs across the US

GFT was built to predict CDC reports

Given that GFT is often held up as an exemplary use of big data, what lessons can we draw from this error?

Big Data Hubris

Algorithm Dynamics

Big Data Hubris

- “Big data hubris” is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis
- Quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and *dependencies* among data
- 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

Big Data Hubris: GFT versions

- The initial version of GFT:
  - The methodology was to find the best matches among 50 million search terms to fit 1152 data points
  - The odds of finding search terms that match the propensity of the flu but are structurally unrelated (and so do not predict the future) were quite high
- GFT developers report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data (e.g., those regarding high school basketball)
  - **Warning!** Big data were overfitting the small number of cases—a standard concern in data analysis
  - This **ad hoc method** of throwing out peculiar search terms failed when GFT completely missed the non-seasonal 2009 influenza A–H1N1 pandemic
- The initial version of GFT was part flu detector, part winter detector
  - a particularly problematic marriage of big and small data

Big Data Hubris: GFT versions 2

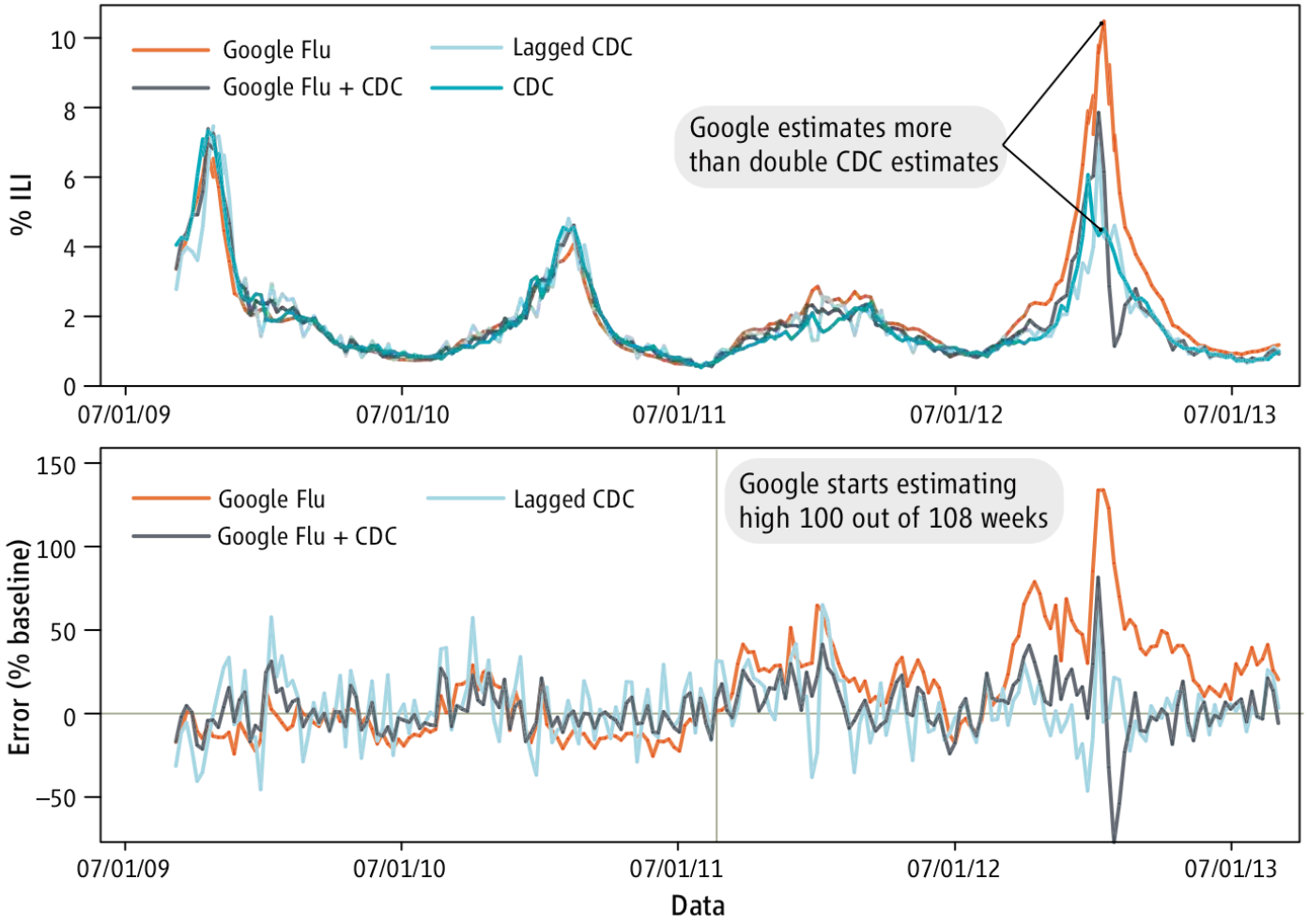
- Algorithm was updated in 2009
- With a few changes announced in October 2013

Big Data Hubris: Results 1

- The new GFT has been persistently overestimating flu prevalence
- GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011
- These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal auto-correlation), and the direction and magnitude of error varies with the time of year (seasonality)
- These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods

Big Data Hubris: Results 2

- The comparison has become even worse since 2010, with lagged models significantly outperforming GFT
  - Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT



GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage {[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Does this mean that the current version of GFT is not useful?

Big Data Hubris: How to Improve

- Greater value can be obtained by combining GFT with other near–real-time health data
  - combining GFT and lagged CDC data
  - dynamically recalibrating GFT

Algorithm Dynamics

- Algorithm dynamics are the ***changes*** made by engineers to improve the commercial service and by consumers in using that service

Why is algorithm dynamics a significant issue?

Algorithm Dynamics: Foundation of Measurement

- All empirical research stands on a foundation of measurement
  - Is the instrumentation actually capturing the theoretical construct of interest?
  - Is measurement stable and comparable across cases and over time?
  - Are measurement errors systematic?
- At a minimum, it is quite likely that **GFT was an unstable reflection of the prevalence of the flu because of algorithm dynamics affecting Google's search algorithm**

Algorithm Dynamics: Changes in Search Algorithm

- **Several changes in Google's search algorithm and user behavior likely affected GFT's tracking**
- The Google search algorithm is *not* a static entity—the company is constantly testing and improving search
  - E.g., the official Google search blog reported 86 changes in June and July 2012 alone

Algorithm Dynamics: Challenges

- There are ***multiple*** challenges to replicating GFT's original algorithm
- GFT has never documented the 45 search terms used, and the examples that have been released appear misleading
- Google does provide a service, Google Correlate, which allows the user to identify search data that correlate with a given time series
  - However, it is limited to national level data, whereas GFT was developed using correlations at the regional level
- The service also fails to return any of the sample search terms reported in GFT-related publications

Algorithm Dynamics: Patterns

- Searches for treatments for the flu and searches for information on differentiating the cold from the flu track closely with GFT's errors
- This points to the possibility that the explanation for changes in relative search behavior is "**blue team**" dynamics
  - where the algorithm producing the data (and thus user utilization) has been modified by the service provider in accordance with their business model

Algorithm Dynamics: Biases in the Data Collected?

- In improving its service to customers, Google is ***also changing the data-generating process***
- Modifications to the search algorithm are presumably implemented so as to support Google's business model
  - E.g., in part, by providing users useful information quickly
  - E.g., in part, to promote more advertising revenue
- **Recommended** searches, usually based on what others have searched, will increase the relative magnitude of certain searches

Algorithm Dynamics: Biases in the Data Collected?

- Because GFT uses the relative prevalence of search terms in its model, improvements in the search algorithm can adversely affect GFT's estimates
- **KEY CONCERN**: GFT assumes that relative search volume for certain terms is statically related to external events, but search behavior is ***not just*** exogenously determined, it is also endogenously cultivated by the service provider

Blue team issues are not limited to Google

- **Platforms such as Twitter and Face-book are always being re-engineered**

Whether studies conducted even a year ago on data collected from these platforms can be replicated in later or earlier periods is an open question

SIDE BAR

Red team dynamics occur when research subjects (in this case Web searchers) attempt to manipulate the data-generating process to meet their own goals, such as economic or political gain

Critical lessons learned?

Transparency and Replicability

- Replication is a growing concern across the academy
  - The supporting materials for the GFT-related papers did ***not*** meet emerging community standards
  - ***Neither were core search terms identified nor larger search corpus provided***
- It is impossible for Google to make its full arsenal of data available to outsiders, nor would it be ethically acceptable, given privacy issues
  - ***However, there is no such constraint regarding the derivative, aggregated data***
- Even if one had access to all of Google's data, it would be impossible to replicate the analyses of the original paper from the information provided regarding the analysis
  - E.g., the few search terms offered in the papers 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)

Transparency and Replicability: WHY?

- First, **science is a cumulative endeavor**
  - requires that scientists be able to continually assess work on which they are building
- Second, **accumulation of knowledge requires fuel in the form of data**
  - There is a network of researchers waiting to improve the value of big data projects and to squeeze more actionable information out of these types of data

SIDE BAR

Google is a business, but it also holds in trust data on the desires, thoughts, and the connections of humanity

Making money "without doing evil" (paraphrasing Google's motto) is not enough when it is feasible to do so much good

It is also incumbent upon academia to **build institutional models to facilitate collaborations with such big data projects**—something that is too often missing now in universities

Use Big Data to Understand the Unknown

- What is more valuable is to understand the prevalence of flu at **very local** levels, which is not practical for the CDC to widely produce, but which, in principle, more finely granular measures of GFT could provide
- Such a finely granular view, in turn, would provide powerful input into generative models of flu propagation and more accurate prediction of the flu months ahead of time

Study the Algorithm

- Twitter, Facebook, Google, and the Internet more generally are constantly changing (e.g., engineers + users)
  - A better understanding of how these changes occur over time; replicate findings; to make sure observed patterns are robust and not evanescent trends
- Studying the ***evolution*** of socio-technical systems embedded in our societies is intrinsically important and worthy of study
  - The algorithms underlying Google, Twitter, and Facebook help determine what we find out about our health, politics, and friends

It’s Not Just About Size of the Data

- There is a tendency for big data research and more traditional applied statistics to live in two different realms
- **Big data** offer enormous possibilities for understanding human interactions at a societal scale, with rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables
  - We contend that these are the most exciting frontiers in studying human behavior.
- Traditional “**small data**” often offer information that is not contained (or containable) in big data, and the very factors that have enabled big data are enabling more traditional data collection
  - The Internet has opened the way for improving standard surveys, experiments, and health reporting

Instead of focusing on a "big data revolution," perhaps it is time we were focused on an "all data revolution," where we recognize that the critical change in the world has been innovative analytics, using data from *all* traditional and new sources, and providing a deeper, clearer understanding of our world