The 2012–2013 Divergence of Google Flu Trends

Keith Winstein

keithw@mit.edu
MIT CSAIL

March 14, 2013
The premise

- CDC ILInet flu surveillance is slow
  - 12 days from start of MMWR week
  - 5 days from end of MMWR week
  - Revised over subsequent weeks

- Idea: Real-time estimate
- Use Google searches
  - roughly 40,000 per second
Disease-agnostic training procedure

- Gets historical ground truth (training data)
- Finds search queries that correlate well
- Evaluated on held-out verification set
- Danger: “oscar nominations”
What is ground truth?

- Virological surveillance
  - Cannot predict well from search data
- Influenza-like illness
  - Fever $\geq 100^\circ F$ and (cough and/or sore throat)
  - Measured as %age of outpatient visits
  - 1,950 sites report weekly to CDC
Nov. 11, 2008 announcement

The New York Times

Vol. Clxviii... No. 54,492  ©2008 The New York Times
New York, Wednesday, November 12, 2008

1.50

Buying Binge Slams to Halt
Crisis of Confidence
For U.S. Consumers

Just as one crisis of confidence may be ending, another may be coming.

The price on Wall Street has eased in the last few weeks, and
banks have become somewhat more willing to make loans. But in those
same few weeks, American households appear to have followed into
their own defensive crunch. Suddenly, our consumer society is doing a lot
less consuming. The numbers are pretty incred-
ible. Sales of new vehicles have dropped 21 percent in the third
quarter. Consumer spending appears likely to fall more than for the
first time since 1991 and perhaps by the largest amount since 1982.

With Wall Street edging back from the brink, this crisis of con-
sumer confidence has become the No. 1 short-term issue for the
economy. Nobody doubts that families need to start saving
more than they saved over the last two decades. But if they
don’t change their behavior too quickly, it could be very painful.

Already, Circuit City has filed for bankruptcy, and General Mo-
tors has said that it’s in danger of running out of cash. If the con-
sumer slump continues, there is a potential for a dangerous feed-
back loop, in which spending cuts and layoffs reduce sales,
which reduces sales, which reduces sales, which reduces
sales, which reduces sales, which reduces sales, which reduces
sales, which reduces sales, which reduces sales, which reduces
sales, which reduces sales, which reduces

Call for Aid Package

Leaders May Try to Use Lame-Duck Session to Press Bush

By David M. Hershey

WASHINGTON — Democratic Congressional leaders said Thurs-
day that they were ready to push emergency legislation to aid the
imperiled auto industry when lawmakers return to Washing-
ton next week for the first time after the election, singling out
for one last showdown with Presi-

dent Bush.

“Next week, during the lame-
duck session of Congress, we are
determined to pass legislation
that will save the jobs of millions of workers whose livelihoods are
On the line,” the majority leader,
Harry Reid of Nevada, said in a
statement.

His call for the session came
shortly after the House speaker,
Nancy Pelosi, and Congress and the administration “must take
immediate action” to stave off a possible collapse of the American
auto industry.

Me, Pelosi said at the start of
saying Congress would adopt
an emergency financial aid to the automakers, giving the Treasury Department
the option of using money from
the $700 billion bailout program
instead.

But with the White House in-
stating that the bail-out money
should be reserved for financial institu-
tions, that option seemed unlikely,
leaving a similar Democratic
effort to pass legislation.
NYT figure

PERCENT OF HEALTH VISITS FOR FLU-LIKE SYMPTOMS  Mid-Atlantic region

8 percent

ESTIMATED
Based on Google Flu Trends data tracking flu-related search terms

ACTUAL
As reported by U.S. Centers for Disease Control

Using Google to Monitor the Flu

Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

C.D.C. does not keep data for June through September

Sources: Google; Centers for Disease Control

THE NEW YORK TIMES
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year². In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities². Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza³. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each state. By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of a query being related to an ILI as well as several other variables.
Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.
Accuracy figures

Index based on 45 queries (e.g. “pnumonia”).

- Training data (2003–2007): $0.80 \leq r \leq 0.96$ (mean 0.90)
- Verification (2007–2008): $0.92 \leq r \leq 0.99$ (mean 0.97)

“We intend to update our model each year with the latest sentinel provider ILI data, obtaining a better fit and adjusting as online health-seeking behaviour evolves over time.”
**High expectations**

**NYT:** “In April 2009, Dr. Brilliant said it epitomized the power of Google’s vaunted engineering prowess to make the world a better place, and he predicted that it would save untold numbers of lives.”

**Brilliant on PBS:** “This one little program, done by three engineers, outperforms CDC or WHO’s very expensive surveillance system by two or three weeks. And CDC is thrilled about that. They’re not unhappy. It’s not a competitive issue. They’re really happy. So you can find less expensive ways to know when the flu season is beginning, what states should get the first shipment of vaccine or antivirals, using these technologies.” (May 2009)
Performance in the first year

The 2012–2013 Divergence of Google Flu Trends

- Training data (2003–2007): Mean correlation **0.90**
- Verification (2007–2008): Mean correlation **0.97**
- Actual (March–August 2009): Mean correlation **0.29**!

Model retrained in September 2009, now 160 queries. “We will continue to perform annual updates of Flu Trends models to account for additional changes in behavior, should they occur.”
Google Flu Trends plot as of today

United States Flu Activity

Historical estimates

See data for: United States

United States Flu Activity

Influenza estimate

Google Flu Trends estimate • United States data

United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.

(http://www.google.org/flutrends/about/how.html)
Most of plot is training data

United States Flu Activity

Influenza estimate

Google Flu Trends estimate

United States data

Training data

Verification data

More training data

United States: Influenza-like illness (ILI) data provided publicly by the [U.S. Centers for Disease Control](http://www.google.org/flutrends/about/how.html).
Second divergence in 2012–2013 for U.S.

Outpatient visits for influenza-like illness

Google

CDC

2008 predictor

2009 predictor

Keith Winstein keithw@mit.edu MIT CSAIL

The 2012–2013 Divergence of Google Flu Trends
Large divergence (3.7×) in New England (HHS region 1)

Outpatient visits for influenza-like illness

Google

CDC

Keith Winstein

The 2012–2013 Divergence of Google Flu Trends
Substantial divergence (+72%) in France

Outpatient visits for influenza-like illness

Google
Sentinelles

Keith Winstein
keithw@mit.edu
MIT CSAIL

The 2012–2013 Divergence of Google Flu Trends
Substantial divergence in Japan
My understanding of Google’s point of view

- **GFT succeeded** at predicting early flu onset
- Correlation and RMS error aren’t the end of the story
- Primary audience is public health authorities
  - Independent index ⇒ value-add
  - Not necessarily trying to get most accurate figure overall
- Method is resilient to confounding by media
- Prefer not to retrain model if still performing well
- Idea is to minimize human influence as much as possible
- Don’t show 2008–09 model, because older versions of software not as relevant for estimating performance of current version.
- Intend to clarify PLoS ONE vs. Nature and training data vs. verification on GFT Web site
- Decline to share 2008–09 data (removed from site)
- Decline to discuss Japanese estimate
My questions re: GFT

- Why did GFT overestimate this year’s flu activity?
- Could several ILInet regions, Réseau Sentinelles, and Japanese NIID have had correlated error?
- In retrospect, were there clues last summer when decision made not to retrain?
- Would more frequent retraining have helped or hindered?
More questions

- Can we develop methods that are robust against whatever befell GFT?
- Is it possible to measure robustness without waiting five years for results?
- Instead of $r$ or RMSE, what about a decision-theoretic measure of accuracy?
  - Method A is earlier but less accurate
  - May still allow us to distribute limited vaccines more appropriately than Method B
  - Model vaccine-distribution policy as function of model estimate
  - Figure of merit: flu cases averted, QALY gained, $ saved, ...