

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

"All the News That's Fit to Print"

The New York Times

VOL. CLVIII . . . No. 54,492

© 2008 The New York Times

NEW YORK, WEDNESDAY, NOVEMBER 12, 2008

\$1.50



DAY OF HONOR (AND SNACKS) Members of the military were among some 20,000 people in Manhattan for the Veterans Day parade.

ACHES, A SNEEZE, A GOOGLE SEARCH

Data on Web May Warn of Outbreaks of Flu

By MICHAEL HELETT
SAN FRANCISCO — There is a new common symptom of the flu, in addition to the usual aches, coughs, fevers and sore throats.

Veterans' Families Seek Aid for Caregiver Role

By LESLIE KAUFMAN

They kept me by her husband, Matt, in August 2003 between his first and second tours of duty in Iraq. They survived in January 2005. Six weeks later, Staff Sergeant Koll was shot in the neck while on patrol in Ramadi, Iraq, and remained a quadriplegic.

Because her husband, now 37, could no longer take care of himself, Mrs. Koll, 38, quit her job as an accountant to take care of him.

through four workers in nine months she gave up. She said many of the caregivers from contractors on the government-provided list "were awful." One did not know how to use the lift system that hoists Mr. Koll out of bed; another gossiped about the family's private business.

But the real problem was that even the good caregivers could not help Mr. Koll live as he wanted. Regulations, for example, do not permit them to take him out of the house. "Mom is back to his

doesn't want just a baby sitter." While she has never regretted leaving her job, the financial repercussions have been serious. Although Mr. Koll gets a full disability pension of \$6,800 a month and their home in Parker, Colo., was donated to them, they have lost Mrs. Koll's salary of \$58,000 a year, as well as employer contributions to her retirement account, and her dental plan.

Mrs. Koll has joined a growing group of veterans' families who are asking to be compensated in some way for the care they provide.

Buying Binge Slams to Halt

Crisis of Confidence For U.S. Consumers

Just an eve crisis of confidence may be ending, another may be coming.

The panic on Wall Street has cooled in the last few weeks, and banks have become somewhat more willing to make loans. But in these same few weeks, American house-

DAVID LEONHART
ECONOMIST

holdings appear to have fallen into their own defensive crouch. Suddenly, our consumer society is doing a lot less consuming. The numbers are pretty intractable. Sales of new vehicles have dropped 22 percent in the third quarter. Consumer spending appears likely to fall next year for the first time since 1982 and perhaps by the largest amount since 1942.

With Wall Street edging back from the brink, this crisis of consumer 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 change their behavior too quickly, it could be very painful.

Already, Circuit City has filed for bankruptcy, and General Motors has said that it's in danger of running out of cash. If the consumer slump continues, there is a potential for a dangerous feedback loop, in which spending cuts and layoffs reinforce each other.

It's a scary time. "LEO Allen, 20, a nursing student in Atlanta, told one of The Times reporters who lunched out across the country last weekend to ask people about the economy. "Why can't we make the economy worse. If people want to make that worse?"

Late Edition
Today, scattered sunshine, high 51. Tonight, hickory clouds, low 43. Tomorrow, cloudy, a few scattered showers, some heavy late, high 54. Weather map is on Page A18.

DEMOCRATS SEEK EMERGENCY HELP FOR AUTOMAKERS

CALL FOR AID PACKAGE

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

By DAVID M. HERSHENHORN and CARL HULSE

WASHINGTON — Democratic Congressional leaders said Tuesday that they were ready to push emergency legislation to aid the imperiled auto industry when lawmakers return to Washington next week for the first time after the election, setting the stage for one last showdown with President 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.

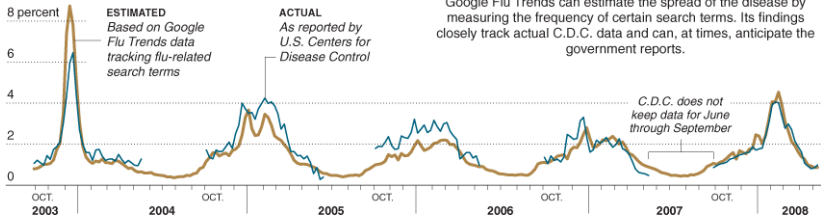
His call for the session came shortly after the House speaker, Nancy Pelosi, said Congress and the administration "must take immediate action" to avert off a possible collapse of the American auto industry.

Mrs. Pelosi stopped short of saying Congress would adopt legislation to provide emergency financial aid to the automakers, giving the Treasury Department the option of using money from the \$300 billion bailout program instead.

But with the White House insisting that the bailout money be reserved for financial institutions, that option seemed unlikely, leading a senior Democratic official to say Democrats would

NYT figure

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



Sources: Google; Centers for Disease Control

THE NEW YORK TIMES

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.

LETTERS

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 year¹. 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 influenza^{3,4}. 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

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

Nature fig. 2

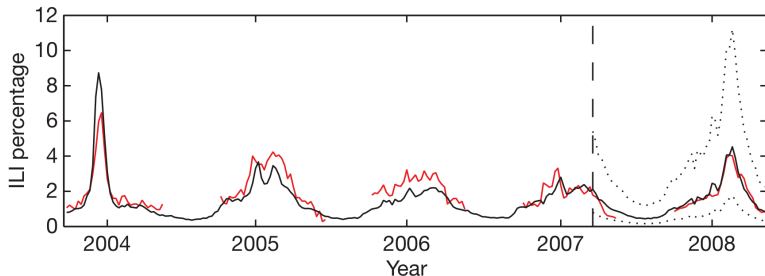


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation 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**)

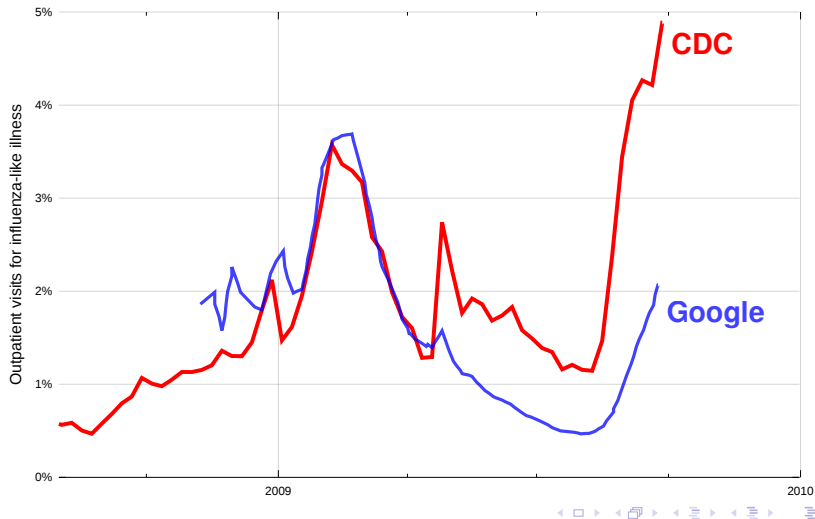
"We intend to update our model each year with the latest sentinel provider ILI data, obtaining a better fit and adjusting as online health-seeking behaviour evolves over time."

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



Aug. 19, 2011: *PLoS ONE* paper

- ▶ 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

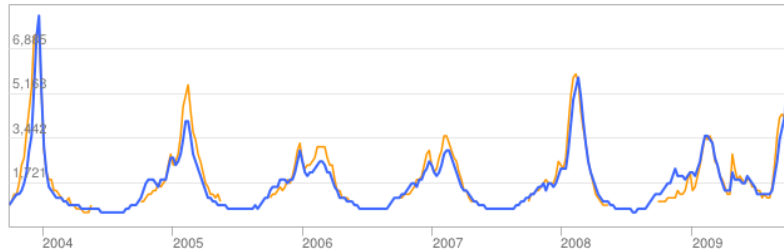
Historical estimates

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

Historical estimates

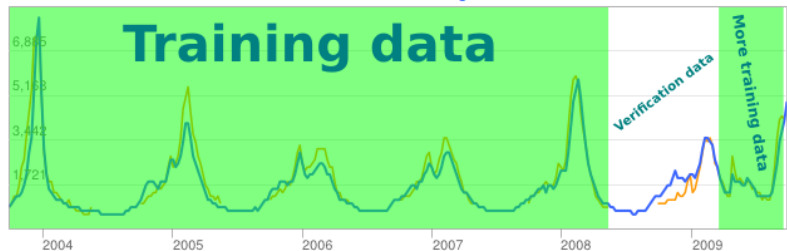
See data for:

United States

United States Flu Activity

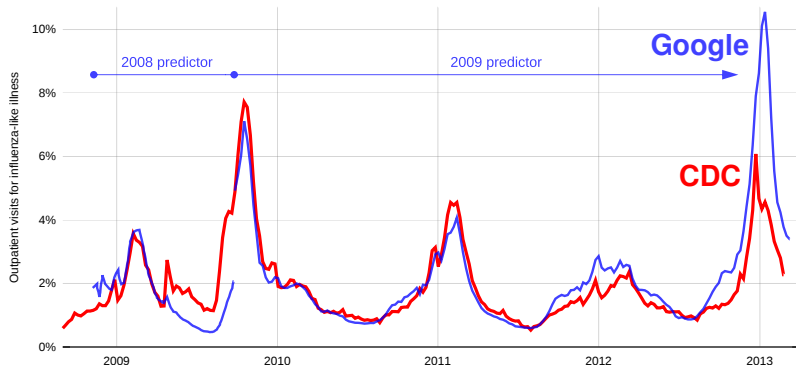
Influenza estimate

● Google Flu Trends estimate ● United States data

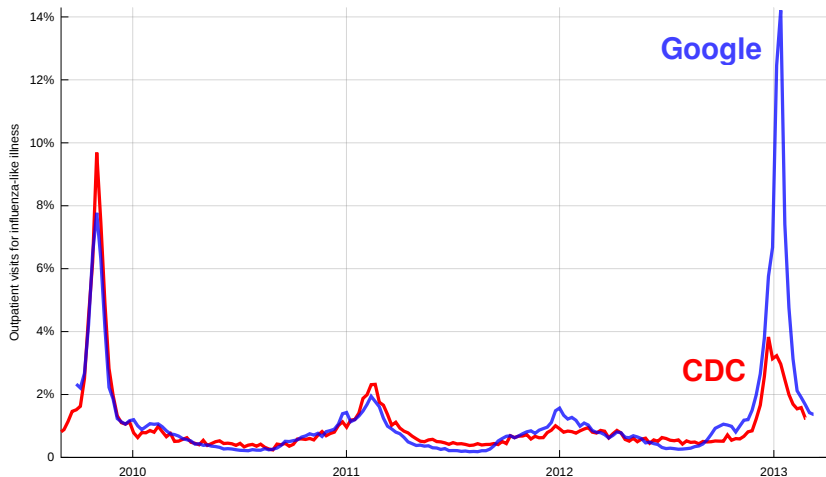


United States: Influenza-like illness (ILI) data provided publicly by the [U.S. Centers for Disease Control](#).

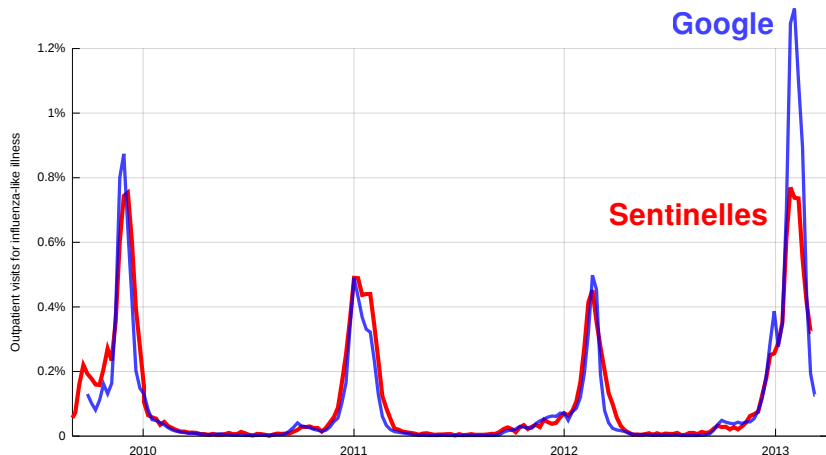
Second divergence in 2012–2013 for U.S.



Large divergence (3.7×) in New England (HHS region 1)



Substantial divergence (+72%) in France



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 \Rightarrow 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, ...