Lecture 19:
CS 5306 / INFO 5306:
Crowdsourcing and Human Computation

Course Projects: Amazon Mechanical Turk

- Responses to come today
- If you're using AMT (80% of you):
  - Still figuring out best was to handle payments
  - In the meantime:
    - Make sure at least one on your team can get an MTurk *requester* account
      - IF NOT, LET ME KNOW ASAP
    - Experiment with what you want to do using the MTurk *requester sandbox*:
      - More info: https://requester.mturk.com/developer/sandbox
      - Sandbox: https://requestersandbox.mturk.com/
    - Can even use the sandbox for your real experiments

Mturk Requester Sandbox

“The Mechanical Turk Developer Sandbox is a simulated environment that lets you test your applications and Human Intelligence Tasks (HITs) prior to publication in the marketplace.

Benefits:

- Free to use for registered Mechanical Turk requesters. Fees will not be withdrawn and payments are not made to Worker accounts.
- Has functional parity with the production website.
- Requires only a URL change to configure your application to work against the developer sandbox or the production website.”

Types of Crowdsourcing

- Overt
  - Collecting (Amazon Reviews)
  - Labor Markets (Amazon Mechanical Turk)
  - Collaborative Decisions (Prediction Markets)
  - Collaborative Creation (Wikipedia)
  - Smartest in the Crowd (Contests)
  - Games with a Purpose
- Covert / Crowd Mining
  - Web page linkage, search logs, social media, collaborative filtering
- Dark side of crowdsourcing and human intelligence
- Collective intelligence in animals

SEO.com
#1 in Digital Marketing

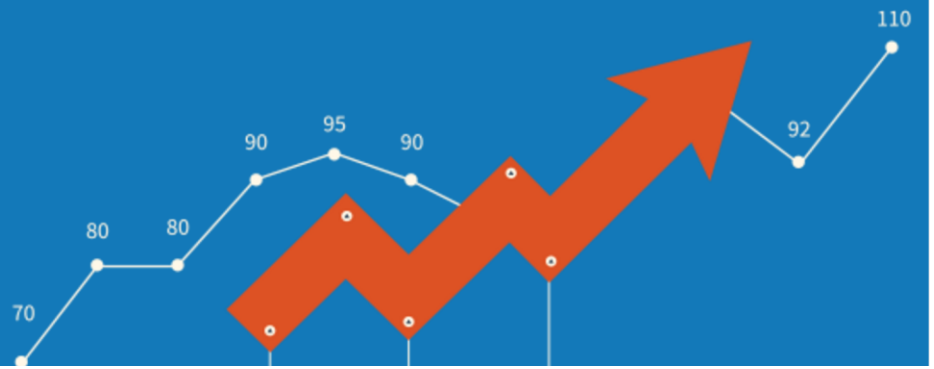
Services Results About Blog Contact Us 800.351.9081

Stress-Free SEO & Online Marketing Solutions for Small Businesses

Done-for-you online marketing services that are everything your business needs to generate awareness, drive traffic, connect with customers, and increase sales.

Get started with a FREE WEBSITE ANALYSIS »

Interested in working with us? REQUEST A PROPOSAL »



Google Bombing

- “more evil than Satan himself“: microsoft.com (1999)
- “French military victories”: page with “Did you mean French military defeats?” (2003)
- “weapons of mass destruction” (2003)
- “miserable failure”: George Bush (2003)
- “waffles”: Al Gore (2004)
- “Jew”: Wikipedia article for “Jew” (2004)
- Amway Quixtar (2006)
- “liar”: Tony Blair (2005)
- “worst band in the world”: Creed (2006)
- “dangerous cult”: Scientology
- “murder”: Wikipedia article for abortion

Google Let JC Penney Spam Search Results For Months

Matt Rosoff | Feb. 13, 2011, 12:10 PM | 5,901 | 15

Share 37 Tweet 142 Recommend 15 Email A A A

The New York Times exposed the dirty side of search engine optimization this morning with a long article about how JC Penney spammed Google so it would appear at the top of search results.

Somebody created thousands of fake pages with the keywords that Penney wanted to game, like "black dresses," and a direct link to Penney's site. This messes with Google's PageRank algorithm, which assumes that a site is useful if it's popular. (A Penney spokesperson denied that the company knew what was going on -- it was probably a guerrilla SEO team or agency working on Penney's behalf.)

The amazing part of the story isn't how Penney tricked Google -- this kind of "black hat" SEO has been around almost since Google began.

Google cofounders Larry Page and Sergey Brin back in more carefree days.

GOOG	Jul 8 2011, 05:20 PM EDT		
	531.99	Change	% Change
		-14.61	-2.67%

A Bully Finds a Pulpit on the Web

By DAVID SEGAL
Published: November 26, 2010

SHOPPING online in late July, Clarabelle Rodriguez typed the name of her favorite eyeglass brand into Google's search bar.

Enlarge This Image

David G. Klein

In moments, she found the perfect frames — made by a French company called Lafont — on a Web site that looked snazzy and stood at the top of the search results. Not the tippy-top, where the paid ads are found, but under those, on Google's version of the gold-medal podium, where the most relevant and popular site is displayed.

RECOMMEND
TWITTER
COMMENTS (317)
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Related

Issue 14.09 - September 2006

Spam + Blogs = Trouble

Splogs are the latest thing in online scams – and they could smother the Internet.

By Charles C. Mann

Page 1 of 4 **next »**

I am aware that spending a lot of time Googling yourself is kind of narcissistic, OK? But there are situations, I would argue, when it is efficiently – even forgivably – narcissistic. When I published a book last year, I wanted to know what, if anything, people were saying about it. Ego-surfing was the obvious way to do that. Which is how I stumbled across Some Title.

Story Tools

PRINT MAIL

Story Images

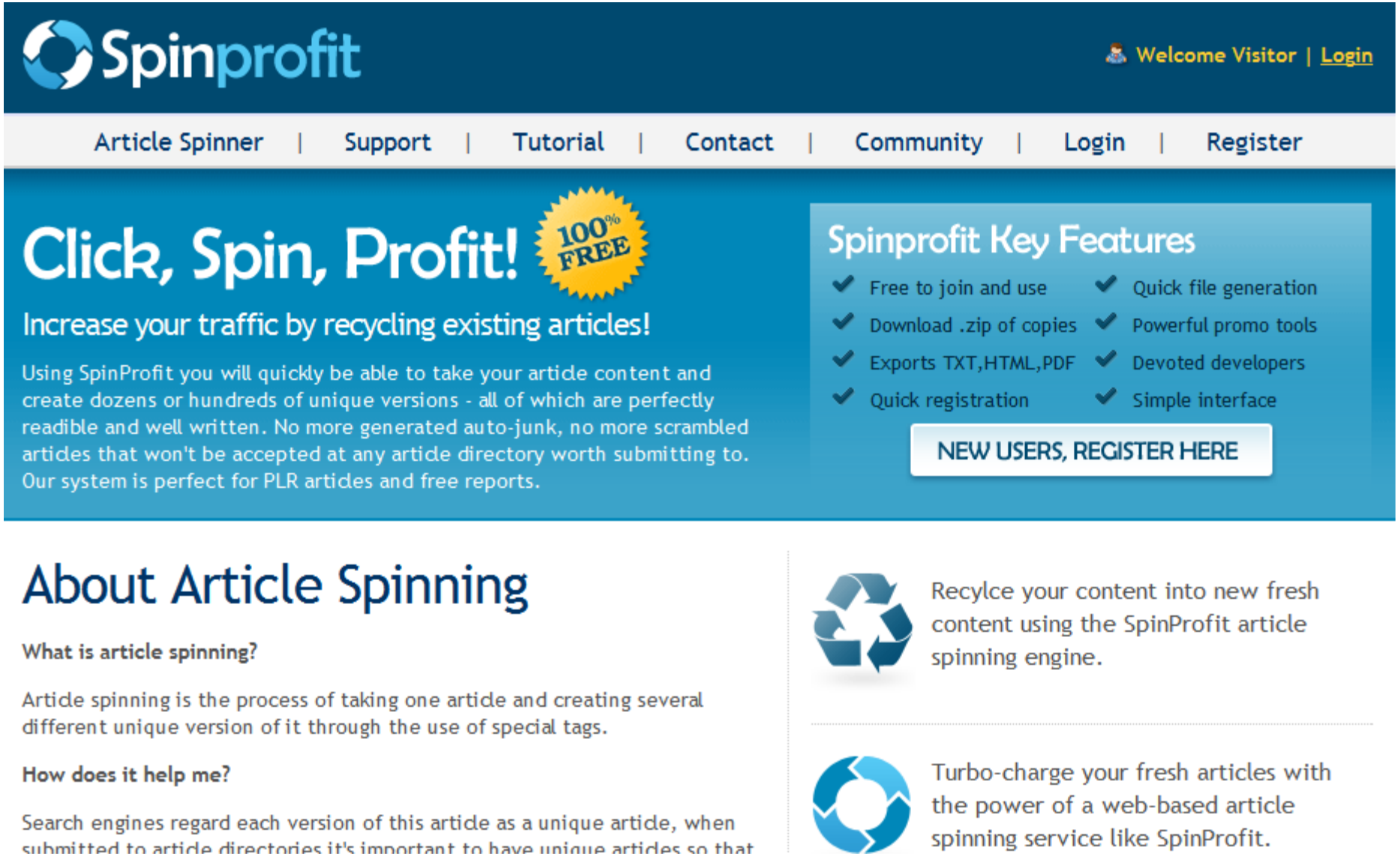
Click thumbnails for full-size image:

Rants + Raves

Some Title identified itself as a blog but obviously wasn't one. Here, reprinted in its entirety, is the paragraph from the site that mentioned me:

Show Disputed Vinland Map Was Made Half Century Before Columbus Trip Audio/Video Columbus: Secrets From The Grave quot;The Last Voyage of Columbus quot;: An Epic Tale Charles Mann's quot;1491 quot; (Audio

In orthodox bloggy style, the paragraph linked to another Web page. When I clicked on the link, I was confronted with more gibberish: "Below," it stated, "you will find some grave



Adversarial Information Retrieval

- Content manipulation
- Link creation and manipulation
- Cloaking: Serving up different content for spiders vs people
- Click fraud
- Query fraud
- Referrer fraud
- Hacked websites

Foundations and Trends® in
Information Retrieval
Vol. 4, No. 5 (2010) 377–486

DOI: 10.1561/1500000021

now
the essence of knowledge

Adversarial Web Search

By Carlos Castillo and Brian D. Davison

Contents

Content Spam

- *repetition* of words to boost weights (including tiny fonts, white on white text, etc.)
- *dumping* unrelated terms or phrases into the page to make the page partially “relevant” for multiple topics
- *weaving* spam phrases into non-spam content copied from other sources
- *stitching* together non-spam content to create new artificial content that might be attractive for search engines

Content Spam

- “Malicious mirroring”
- “302 attacks” – redirects
- Splogs
- Forum spam
- Comment spam

Link Spam

- Link farms
  - Sybil attacks
  - Collusion attacks (“mutual admiration society”)

Usage Spam

- Click fraud
- Query fraud
- Referrer fraud

Combating Web Spam with TrustRank

Zoltán Gyöngyi
Stanford University
Computer Science Department
Stanford, CA 94305
zoltan@cs.stanford.edu

Hector Garcia-Molina
Stanford University
Computer Science Department
Stanford, CA 94305
hector@cs.stanford.edu

Jan Pedersen
Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089
jpederse@yahoo-inc.com

Abstract

Web spam pages use various techniques to achieve higher-than-deserved rankings in a search engine's results. While human experts can identify spam, it is too expensive to manually evaluate a large number of pages. Instead, we propose techniques to semi-automatically separate reputable, good pages from spam. We first select a small set of seed pages to be evaluated by an expert. Once we manually identify the reputable seed pages, we use the link structure of the web to discover other pages that are likely to be good. In this paper we discuss possible ways to implement the seed selection and the discovery of good pages. We present results of experiments run on the World Wide Web indexed by AltaVista and evaluate the performance of our techniques. Our results show that we can effectively filter out spam from a significant fraction of the web, based on a good seed set of less than 200 sites.

Introduction

creation of a large number of bogus web pages, all pointing to a single target page. Since many search engines take into account the number of incoming links in ranking pages, the rank of the target page is likely to increase, and appear earlier in query result sets.

Just as with email spam, determining if a page or group of pages is spam is subjective. For instance, consider a cluster of web sites that link to each other's pages repeatedly. These links may represent useful relationships between the sites, or they may have been created with the express intention of boosting the rank of each other's pages. In general, it is hard to distinguish between these two scenarios.

However, just as with email spam, most people can easily identify the blatant and brazen instances of web spam. For example, most would agree that if much of the text on a page is made invisible to humans (as noted above), and is irrelevant to the main topic of the page, then it was added with the intention to mislead. Similarly, if one finds a page with thousands of URLs referring to hosts like

```
buy-canon-rebel-300d-lens-case.camerasx.com,
buy-nikon-d100-d70-lens-case.camerasx.com,
```

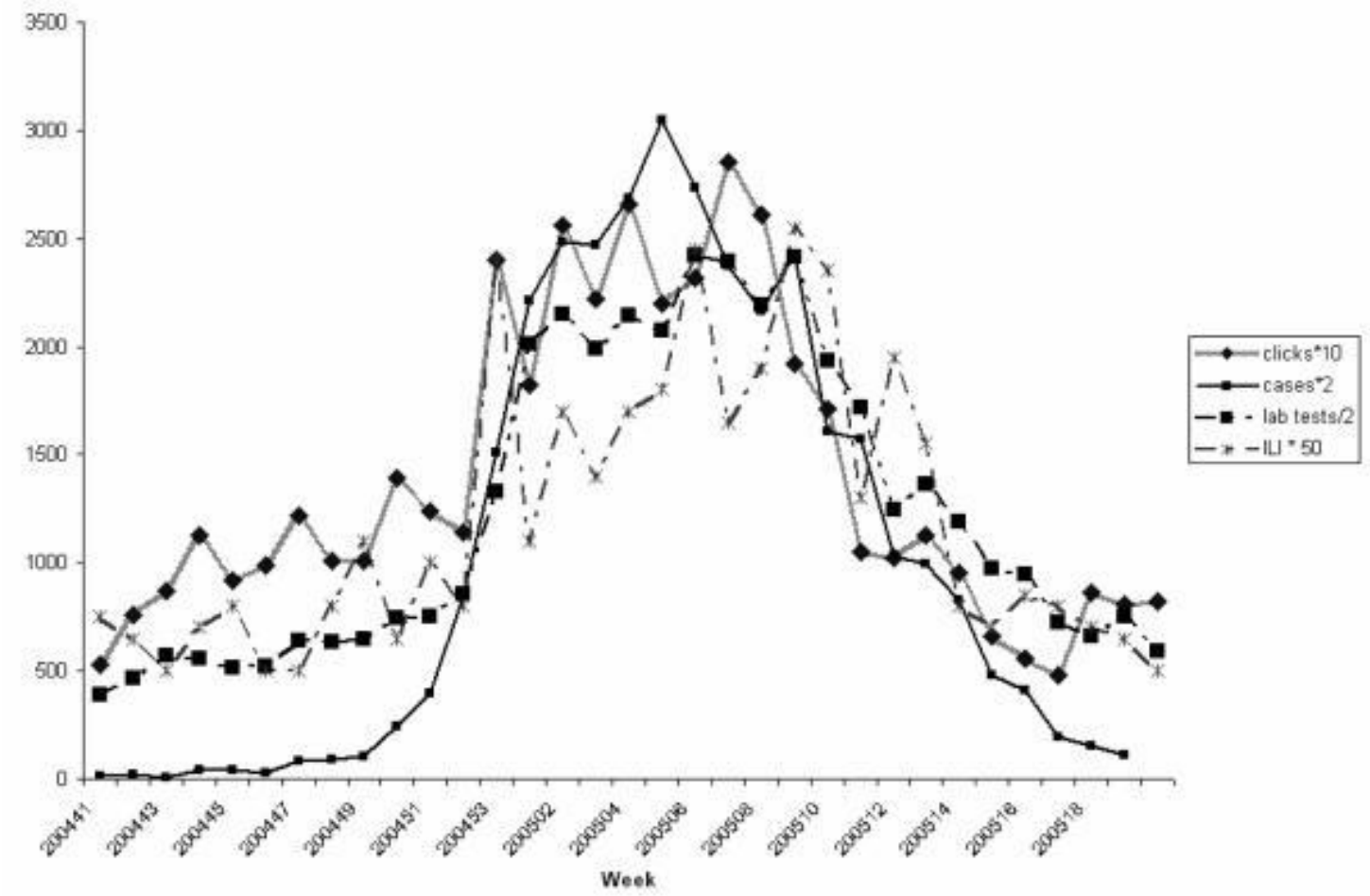
Types of Crowdsourcing

- Overt
  - Collecting (Amazon Reviews)
  - Labor Markets (Amazon Mechanical Turk)
  - Collaborative Decisions (Prediction Markets)
  - Collaborative Creation (Wikipedia)
  - Smartest in the Crowd (Contests)
  - Games with a Purpose
- Covert / Crowd Mining
  - Web page linkage, search logs, social media, collaborative filtering
- Dark side of crowdsourcing and human intelligence
- Collective intelligence in animals

“Tracking flu-related searches on the Web for syndromic surveillance”
Gunther Eysenbach, *AMIA Annu Symp Proc,* November 2006

- Tracked prevalence of search terms using Google Adsense
  - “flu” or “flu symptoms” for users in Canada
  - Ad read: “Do you have the flu? Fever, Chest discomfort, Weakness, Aches, Headache, Cough.”
  - Linked to “generic patient education website”
  - 54,507 impressions and 4,582 clicks
- Compared to Public Health Agency Canada FluWatch reports on number of flu cases seen by doctors, number of lab tests, number of positive lab tests

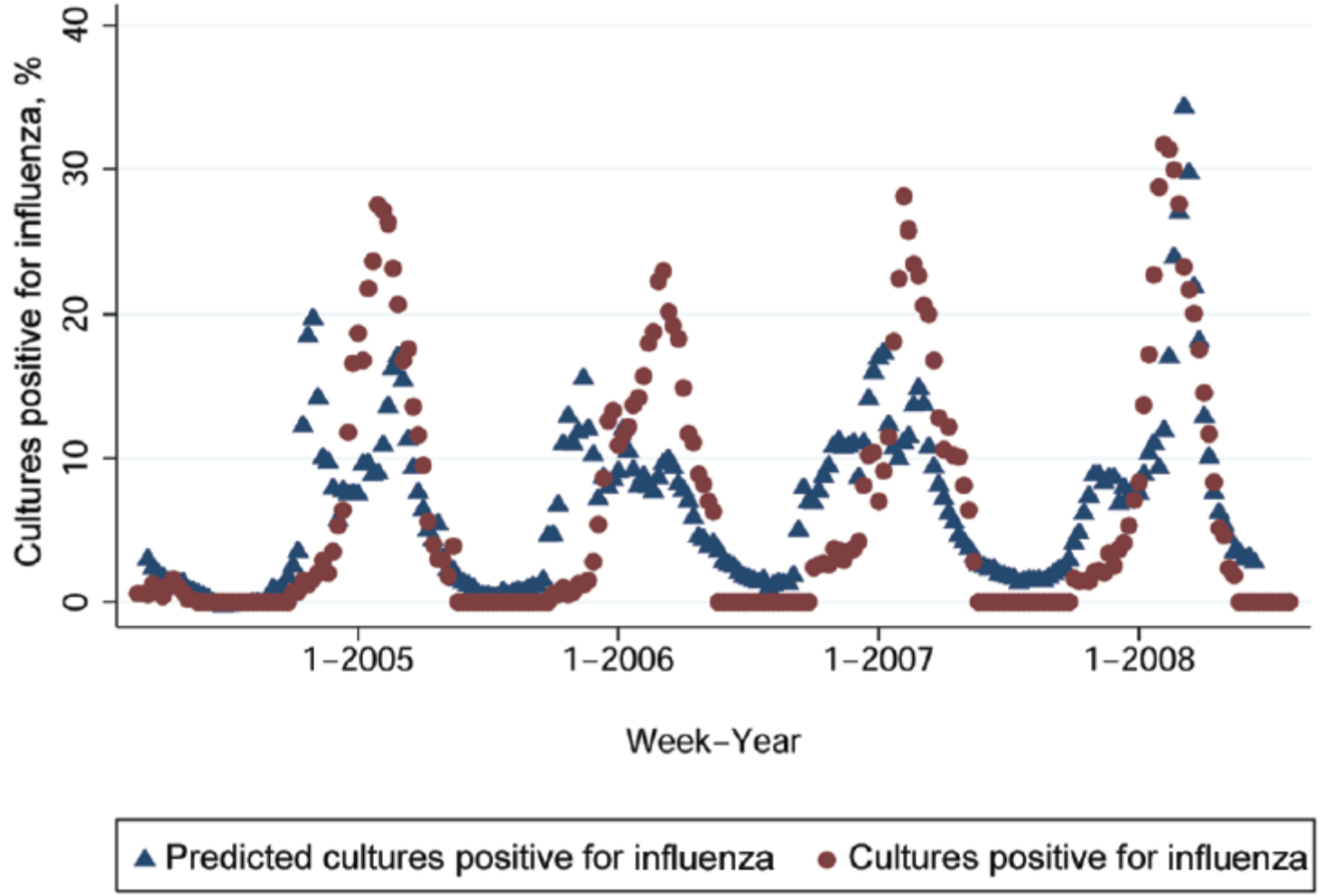
“Tracking flu-related searches on the Web for syndromic surveillance”
Gunther Eysenbach, *AMIA Annu Symp Proc,* November 2006



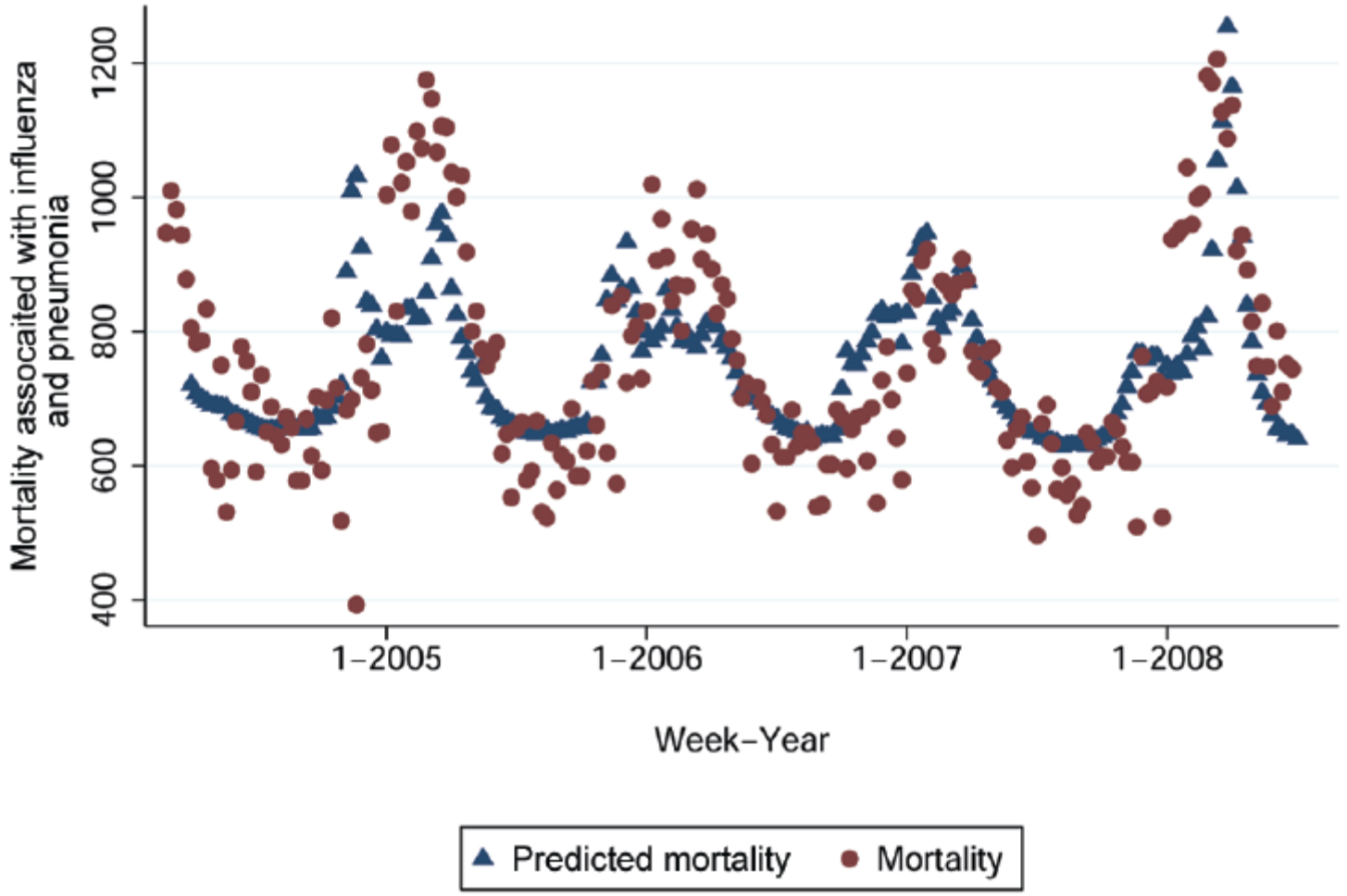
“Using internet searches for influenza surveillance”
Philip M. Polgreen, Yiling Chen, David M. Pennock, Forest D. Nelson,
Clinical Infectious Diseases, 1 December 2008

- Yahoo! U.S. search engine queries that contain the terms “influenza” or “flu” but do not contain the terms “bird,” “avian,” “pandemic,” “vaccine,” “vaccination,” and “shot.”
- Normalized by total number of queries
- Used to predict positive cultures at labs, mortality reports
- Compared to Centers for Disease Control and Prevention data and 122 Cities Mortality Reporting System

Predicing Positive Cultures, 2-Week Lag



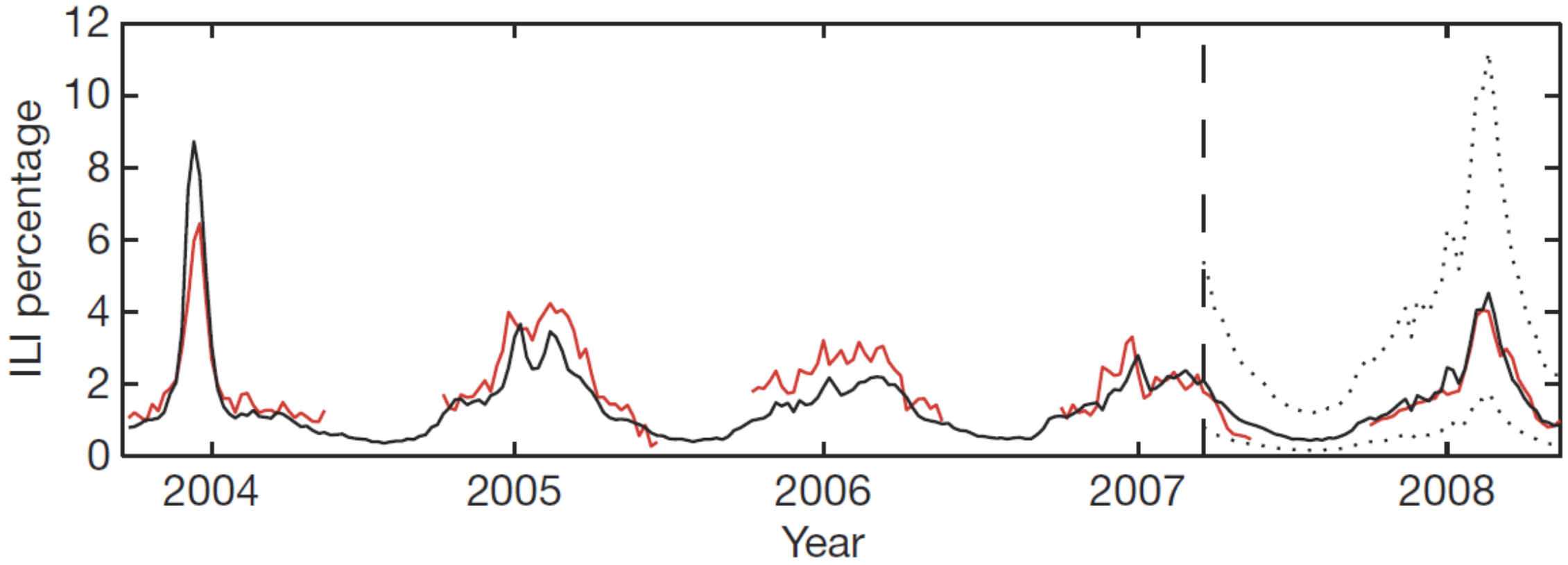
Predicting Mortality, 5-Week Lag



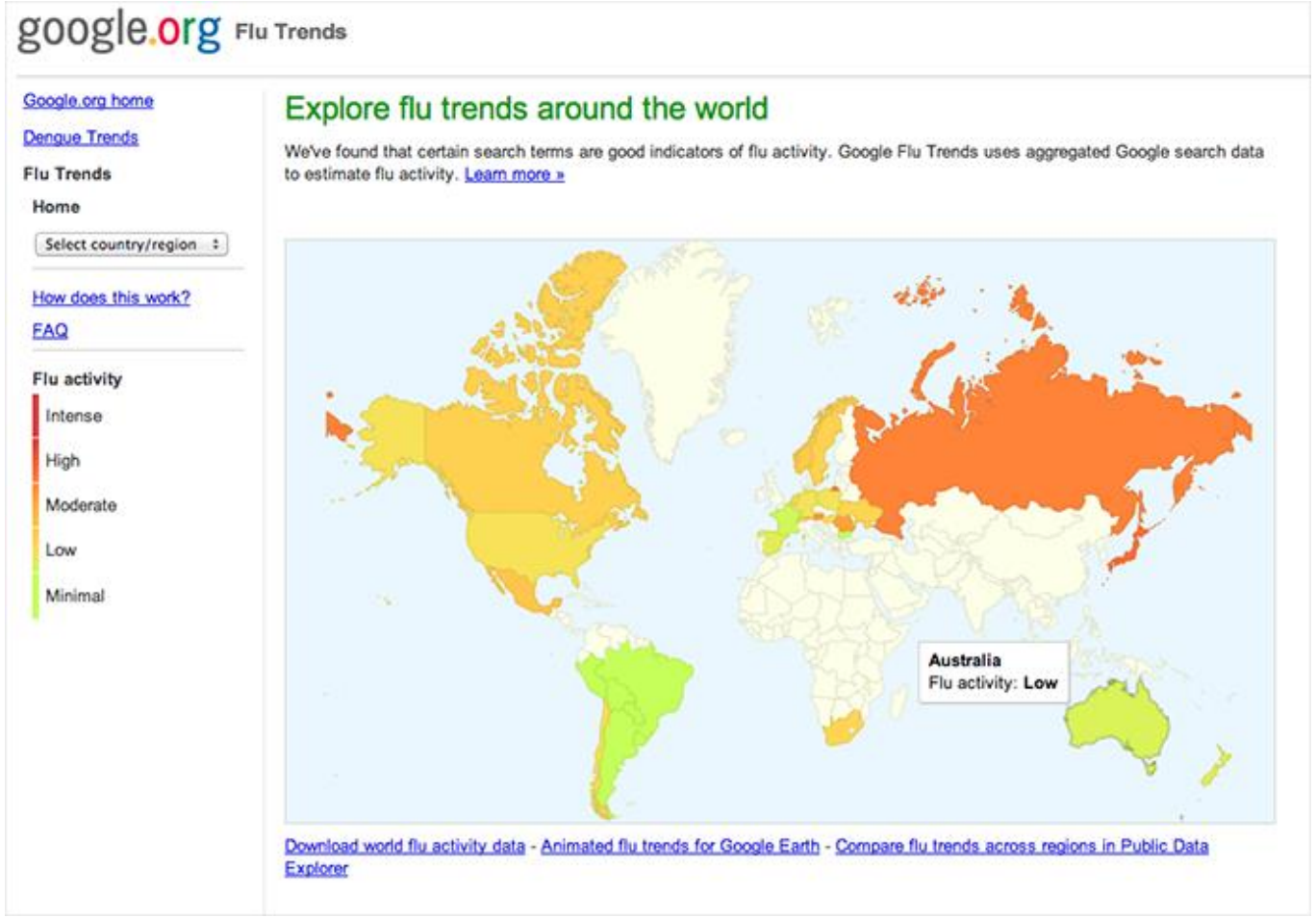
“Detecting influenza epidemics using search engine query data”
J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, *Nature*, 19 February 2009

- Used Google search queries
- Used to predict average percentage influenza-related physician visits
- Compared to Centers for Disease Control and Prevention data
- Query terms discovered automatically

"Detecting influenza epidemics using search engine query data"
J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski,
and L. Brilliant, *Nature*, 19 February 2009



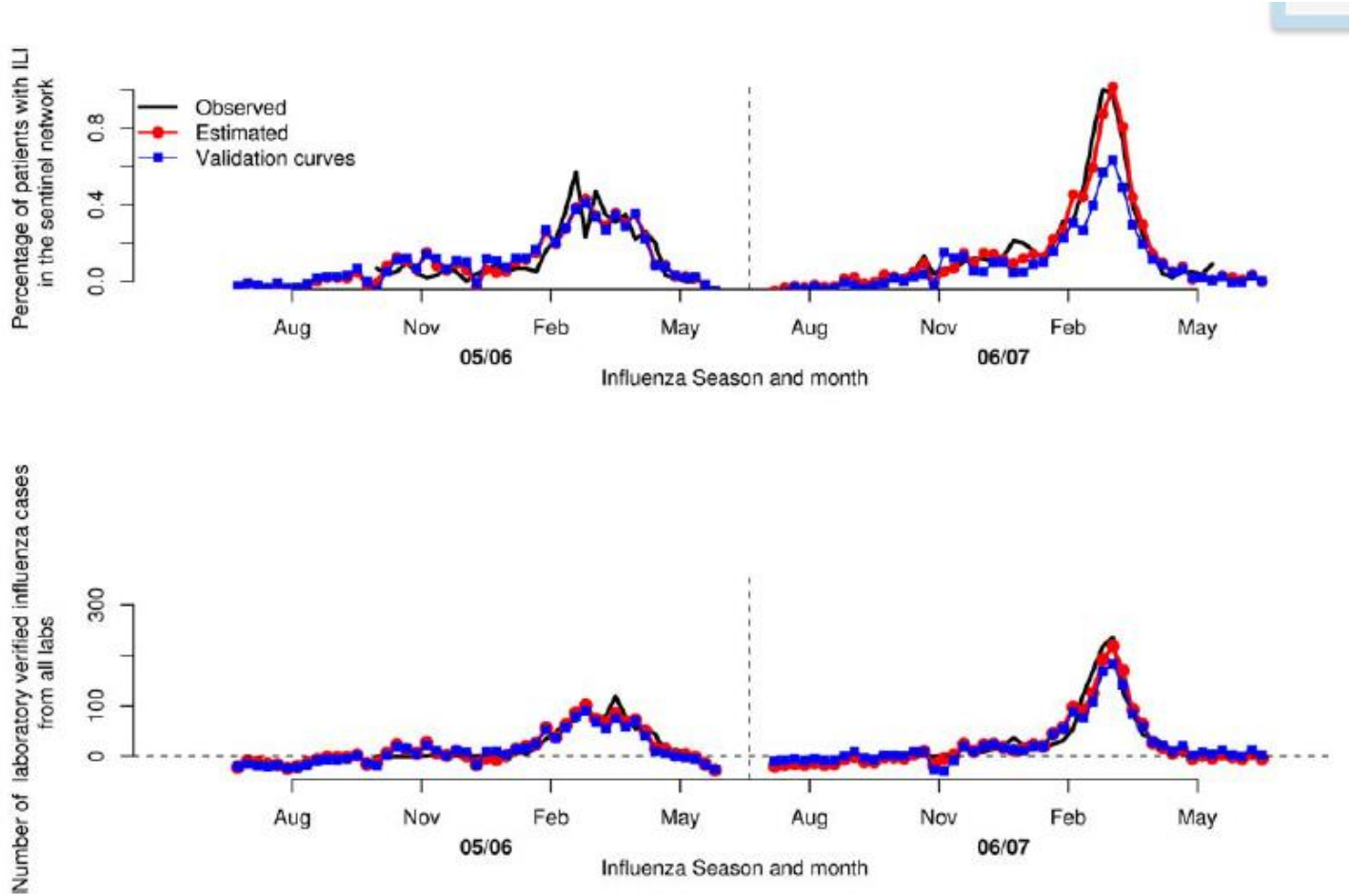
"Detecting influenza epidemics using search engine query data"
J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, *Nature*, 19 February 2009



“Web queries as a source for syndromic surveillance”
A. Hulth, G. Rydevik, and A. Linde, *PLoS One*, 6 February 2009

- Used queries to a medical website (Va°rdguiden.se)
- Used to predict number of lab-verified flu cases, and percentage of doctors visits about flu

“Web queries as a source for syndromic surveillance”
A. Hulth, G. Rydevik, and A. Linde, *PLoS One*, 6 February 2009



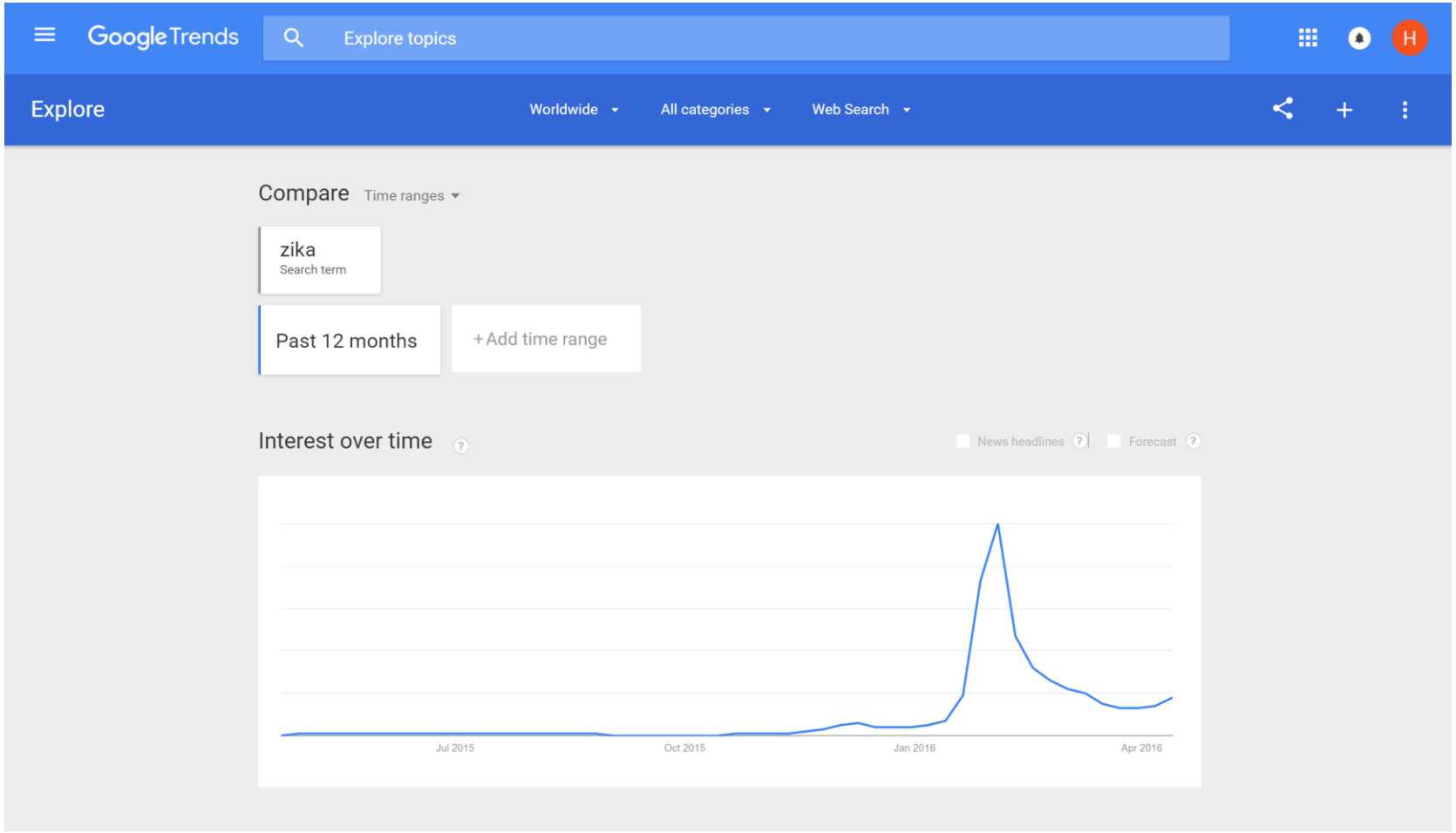
"Google trends: a web-based tool for real-time surveillance of disease outbreaks"
H. A. Carneiro, H.A. and E. Mylonakis,
Clinical infectious diseases, 21 October 2009

- Used Google Trends
- Compared to data from the Centers for Disease Control and Prevention
- Used to predict West Nile virus, respiratory syncytial virus, avian influenza

"Google trends: a web-based tool for real-time surveillance of disease outbreaks"
H. A. Carneiro, H.A. and E. Mylonakis,
Clinical infectious diseases, 21 October 2009

- Used Google Trends
- Compared to data from the Centers for Disease Control and Prevention
- Used to predict West Nile virus, respiratory syncytial virus, avian influenza

Google Trends



"Google trends: a web-based tool for real-time surveillance of disease outbreaks"
H. A. Carneiro, H.A. and E. Mylonakis,
Clinical infectious diseases, 21 October 2009

- Used Google Trends
- Compared to data from the Centers for Disease Control and Prevention
- Used to predict West Nile virus, respiratory syncytial virus, avian influenza

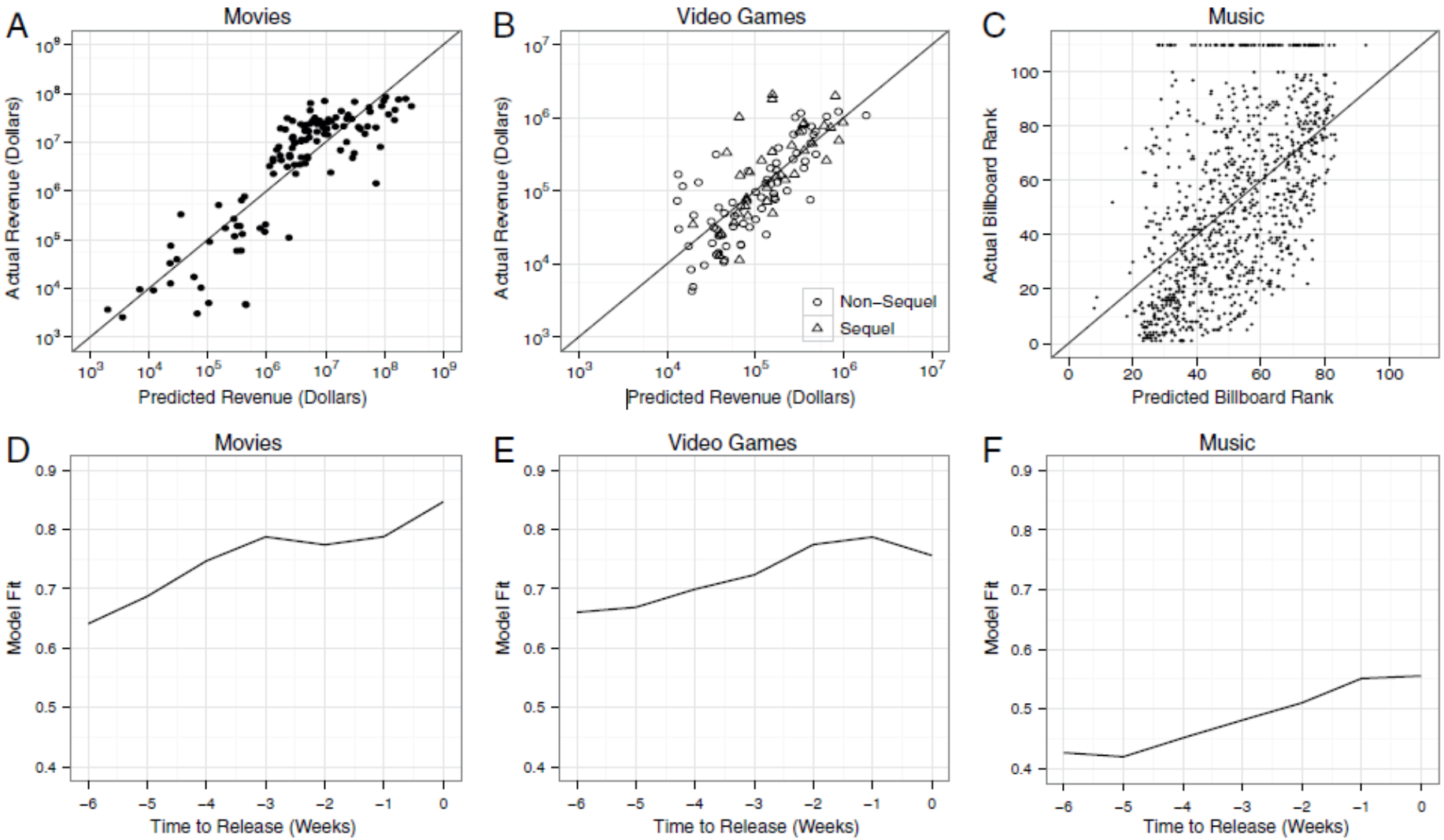
DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

- Used Yahoo! query data for
  - 119 feature films released in the United States between October 2008 and September 2009
  - first-month sales of video games across all gaming platforms (e.g., Xbox, PlayStation, etc.) for 106 games released between September 2008 and September 2009
  - weekly rank of 307 songs that appeared on the Billboard Hot 100 list between March and September 2009


A
Movies
Actual Revenue (Dollars)
Predicted Revenue (Dollars)
B
Video Games
Actual Revenue (Dollars)
Predicted Revenue (Dollars)
Non−Sequel
Sequel
C
Music
Actual Billboard Rank
Predicted Billboard Rank
D
Movies
Model Fit
Time to Release (Weeks)
E
Video Games
Model Fit
Time to Release (Weeks)
F
Music
Model Fit
Time to Release (Weeks)

“Reading tea leaves in the tourism industry: A case study in the gulf oil spill”

H. Choi and P. Liu

2011

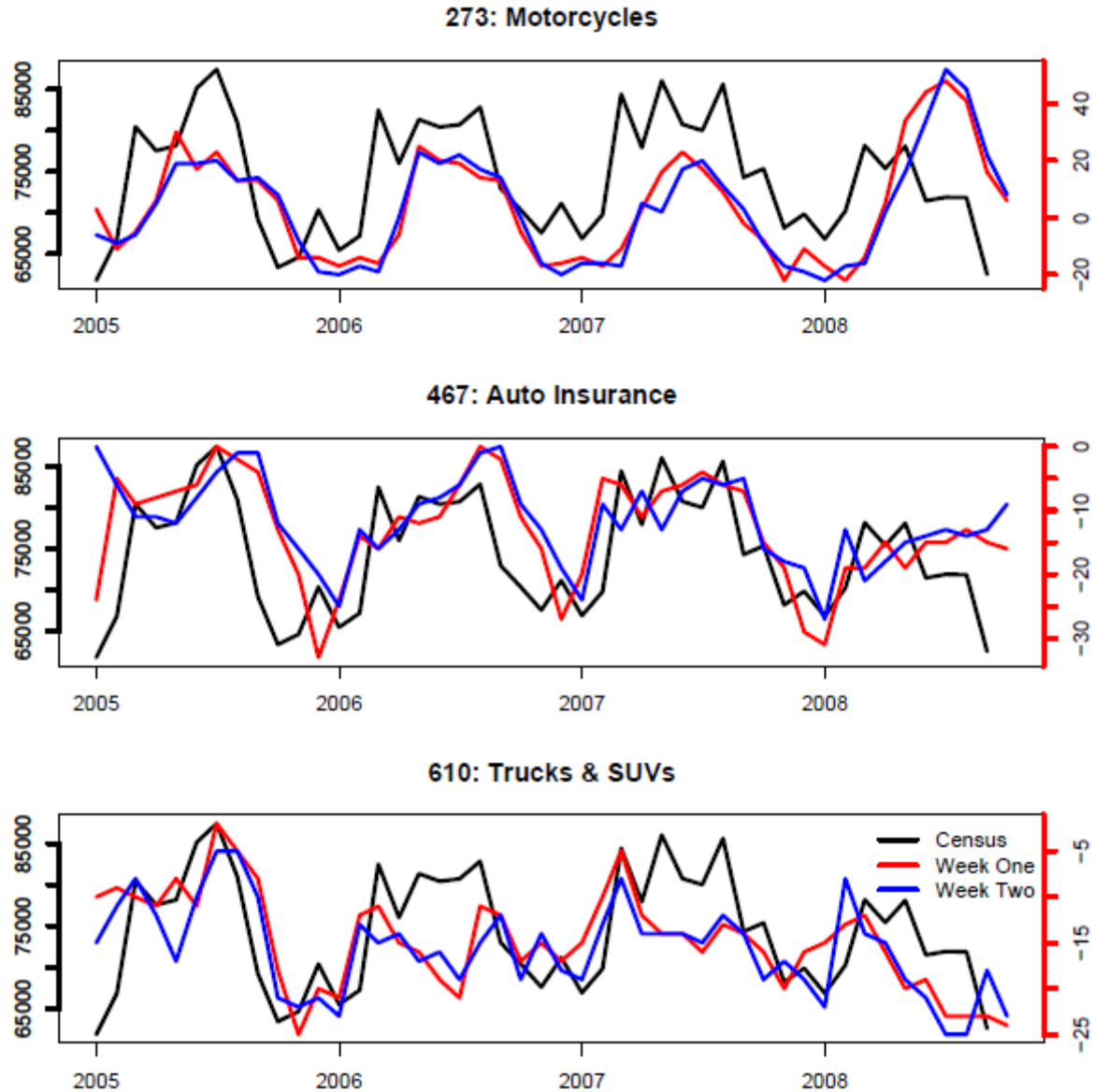
- Used Google Trends search query data for vacation locations
- Compared to travel dataset from Smith Travel Research
- Could predict hotel bookings

“Predicting the present with Google Trends”

H. Choi and H. Varian

2011

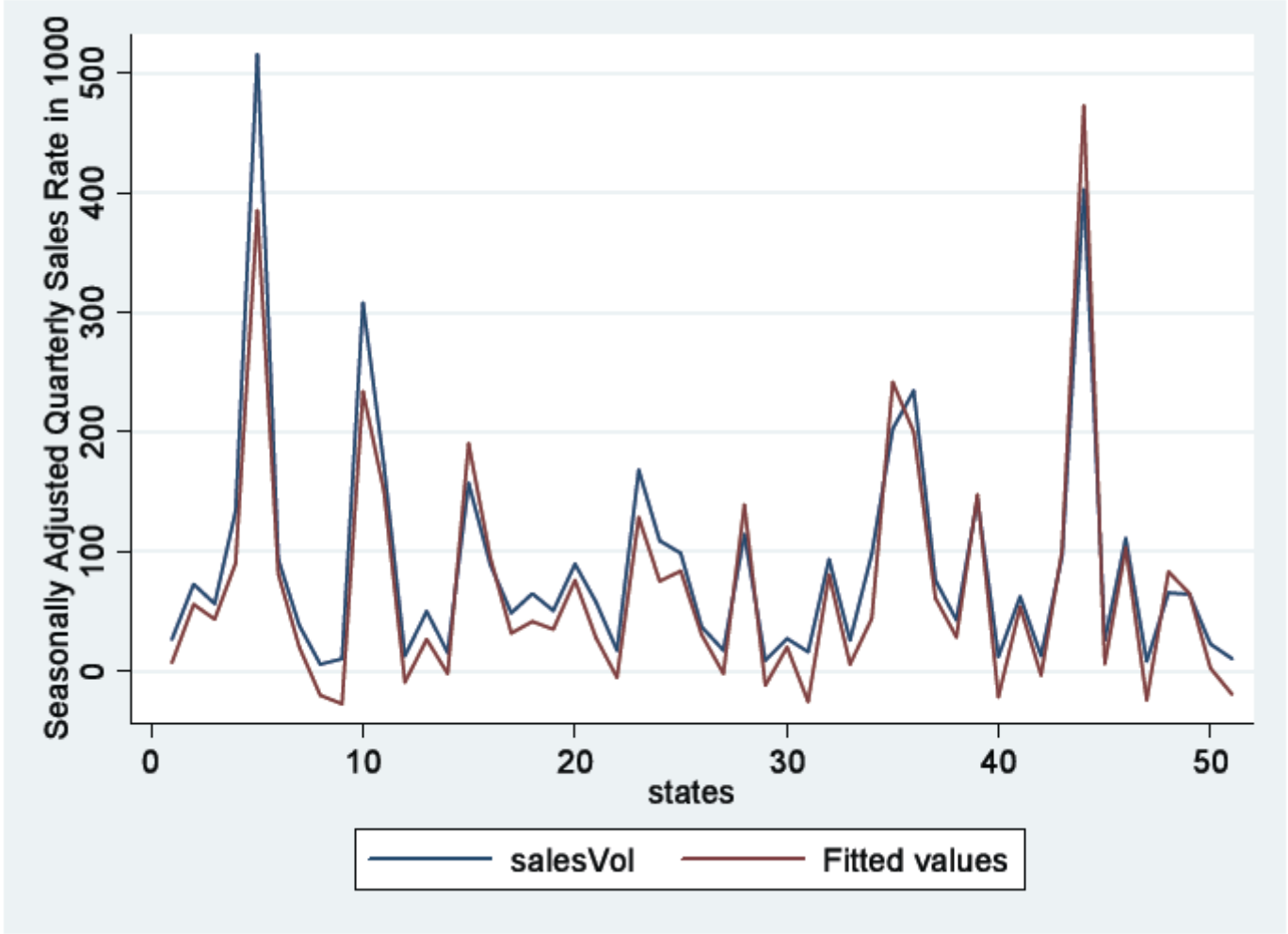
- Uses Google Trends search query data
- Predicts consumer purchasing, car sales, home sales, and travel


273: Motorcycles
65000
75000
85000
-20
0
20
40
2005
2006
2007
2008
467: Auto Insurance
65000
75000
85000
-30
-20
-10
0
2005
2006
2007
2008
610: Trucks & SUVs
65000
75000
85000
-25
-15
-5
2005
2006
2007
2008
Census
Week One
Week Two

“The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities”

L. Wu and E. Brynjolfsson

Workshop on Information Systems and Economics, 2009

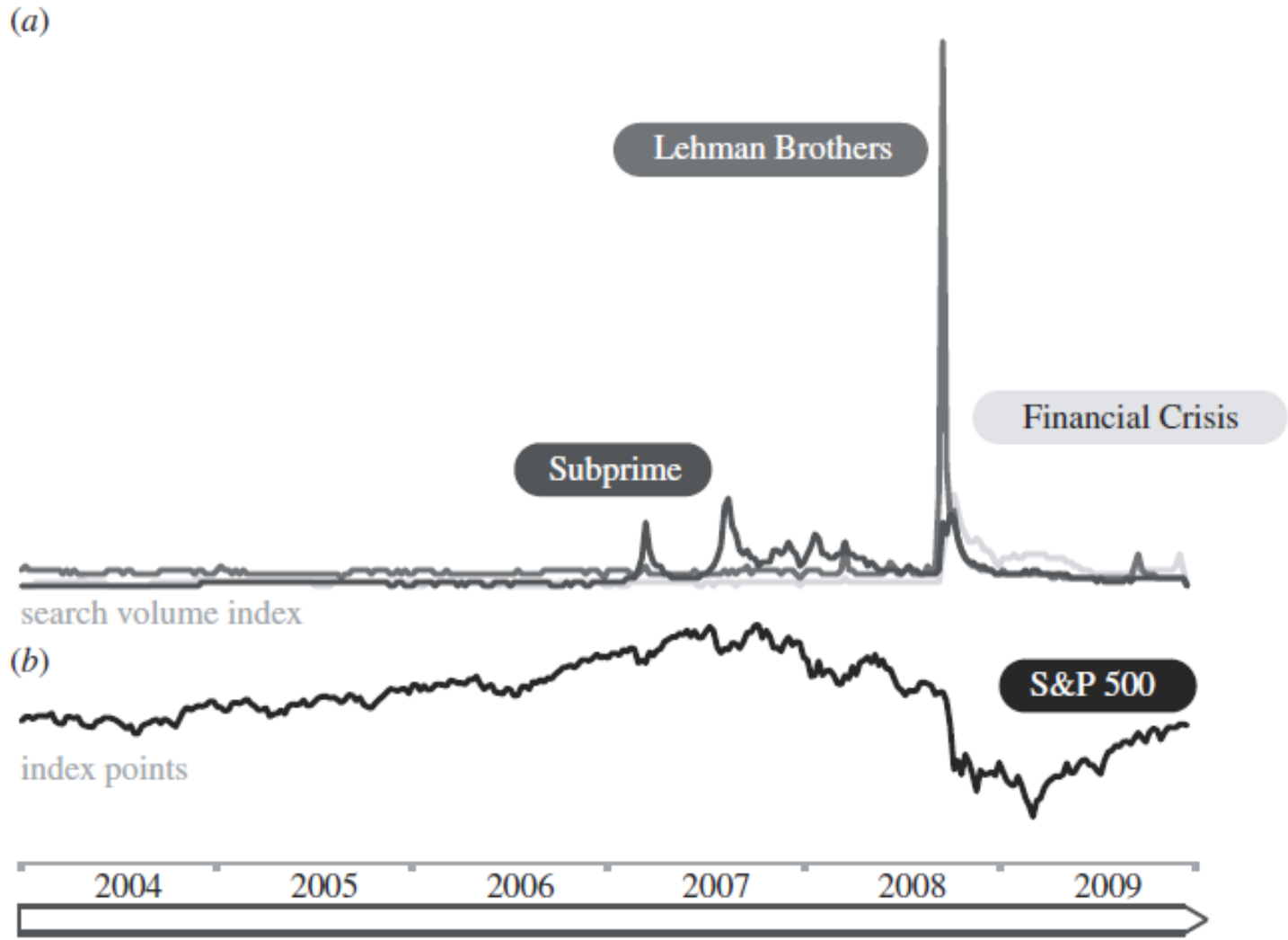

Seasonally Adjusted Quarterly Sales Rate in 1000
0
100
200
300
400
500
0
10
20
30
40
50
states
salesVol
Fitted values

“Complex dynamics of our economic life on different scales: insights from search engine query data”

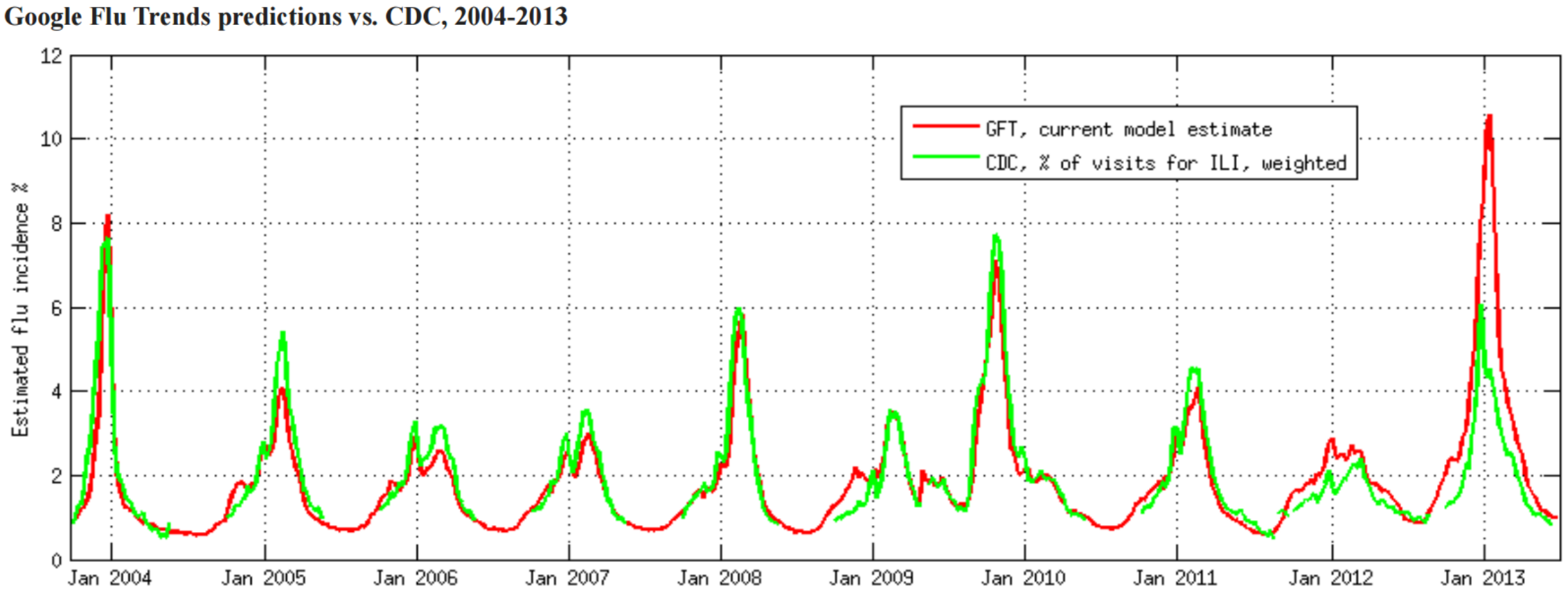
T. Preis, D. Reith, and H. Stanley

Phil Trans R Soc A, 28 December 2010

- Used Google Trends queries of companies
- weekly transaction volumes of S&P 500 companies are correlated with weekly search volume of corresponding company names


(a)
Lehman Brothers
Financial Crisis
Subprime
search volume index
(b)
S&P 500
index points
2004
2005
2006
2007
2008
2009

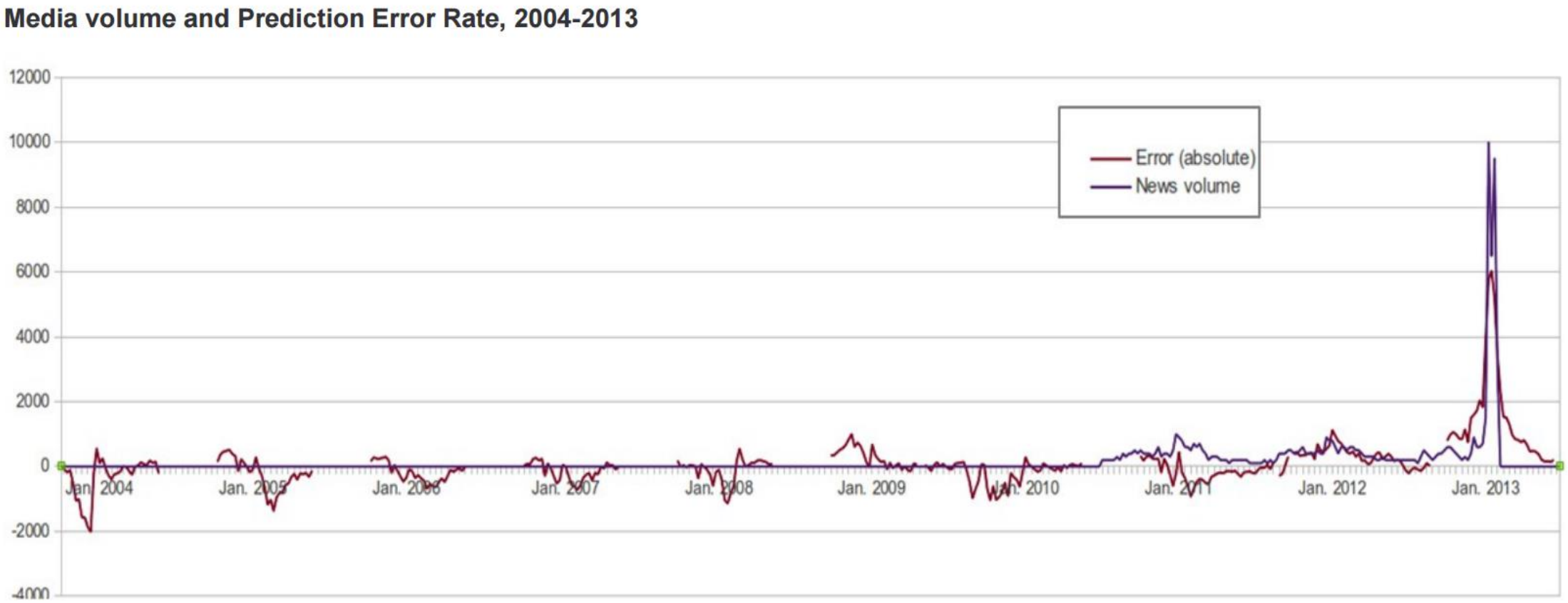
“Google disease trends: an update”
Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). *Nature*, *457*, 1012-1014.



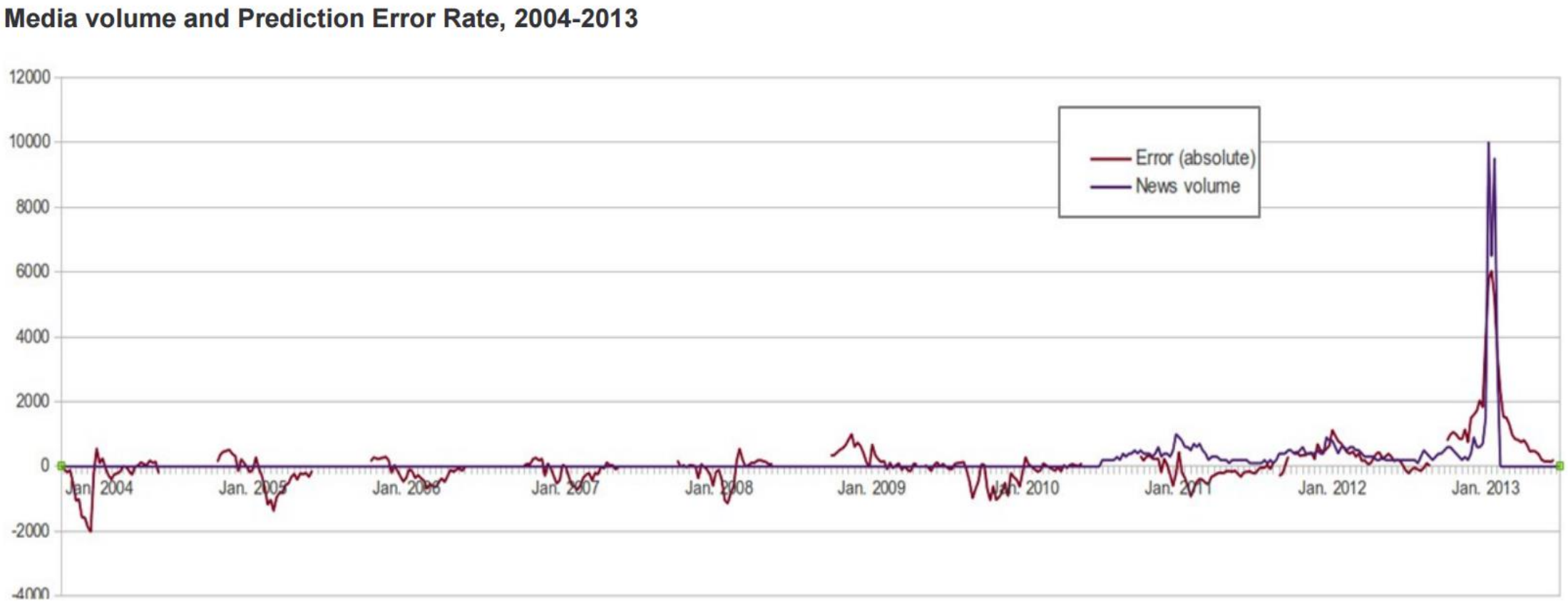
“Google disease trends: an update”
Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). *Nature*, *457*, 1012-1014.

“We have concluded that our algorithm for Flu and Dengue were susceptible to heightened media coverage and have since developed several improvements.”

"Google disease trends: an update"
Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). *Nature*, *457*, 1012-1014.

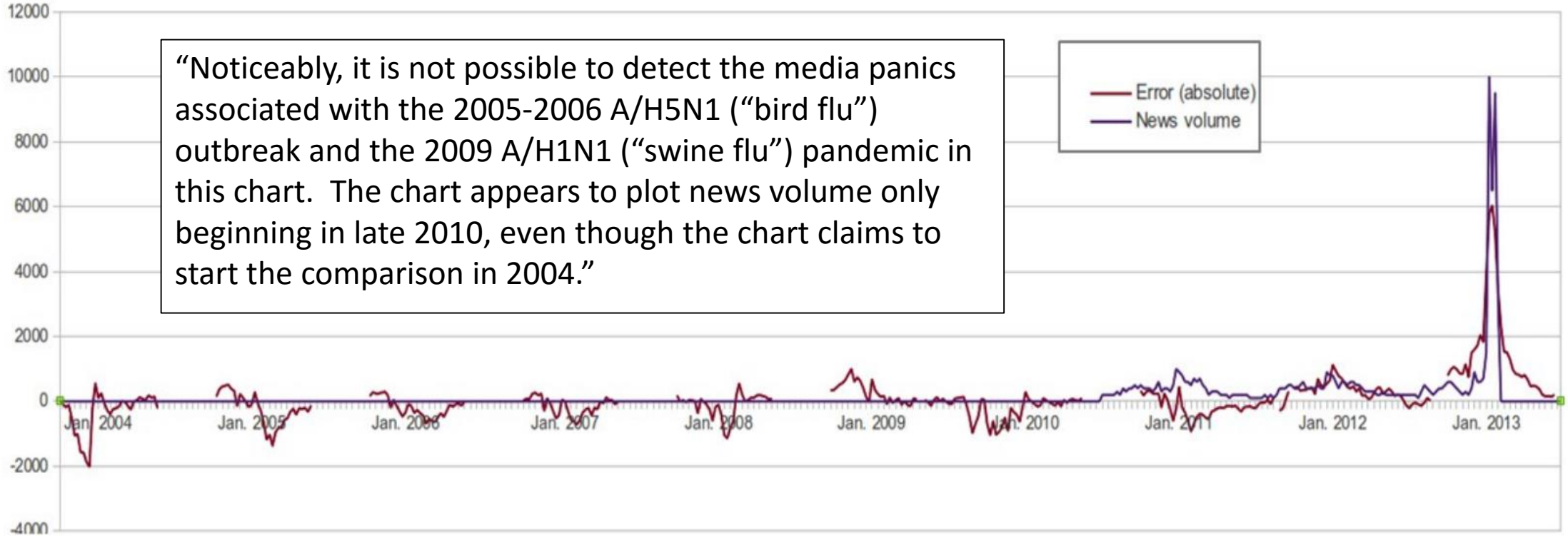


“Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season”, Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. *SSRN Electronic Journal,* January 2014



“Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season”, Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. *SSRN Electronic Journal,* January 2014

Media volume and Prediction Error Rate, 2004-2013



“Noticeably, it is not possible to detect the media panics associated with the 2005-2006 A/H5N1 (“bird flu”) outbreak and the 2009 A/H1N1 (“swine flu”) pandemic in this chart. The chart appears to plot news volume only beginning in late 2010, even though the chart claims to start the comparison in 2004.”

"Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season", Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. *SSRN Electronic Journal,* January 2014

GFT released backdated data based on their new algorithm, but this data does not seem to coincide with the data released on their main page … After much digging, we were able to identify what was wrong with the 2009 algorithm's back data. The file labeled as the "2009 Model Update applied to prior years" is actually the original 2008 algorithm's results. We were only able to confirm this because we located an older version of the GFT website that had been crawled by the Internet Archive (aka. "The Wayback Machine")

"Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season", Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. *SSRN Electronic Journal,* January 2014

When we attempted to contact Google.org about this issue, we found it difficult to get a response, or to identify who to contact. The feedback form provided on the GFT website leads to a page saying that they will "not be able to reply to any feedback or requests for support" (Fig 1). The academic paper outlining the 2013 update provides no contact author, only stating, "For more information, please contact: google.org" (3). We attempted to do so, and also attempted to find e-mail addresses for any of the authors listed in the paper, and were unsuccessful in all these attempts …. There is no readily identifiable contact source for information on how to reconcile these differences in data that is reportedly generated by the same algorithm and data.

“Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season”, Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. *SSRN Electronic Journal,* January 2014

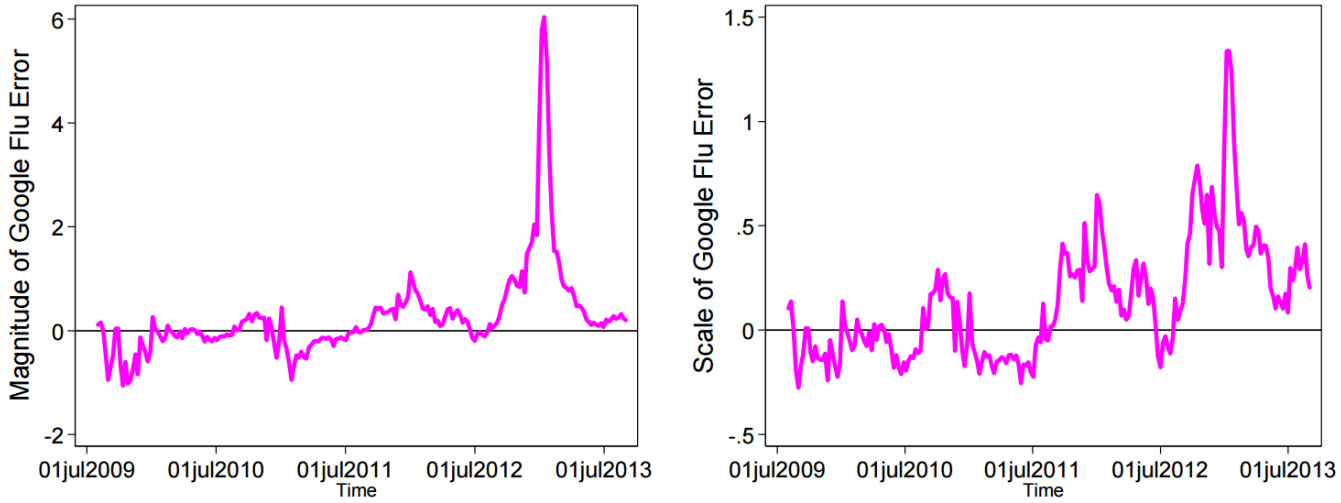


Fig 3. Comparison of absolute and proportional error in the 2010-2013. Absolute error of GFT versus CDC (GFT-CDC) (**Left**) and proportional error of GFT [(GFT-CDC)/CDC] (**Right**). Scaling the errors paints a different picture of GFT errors than looking at absolute error. Absolute error will look larger when the baseline level of flu is lower. This is why most of the focus has been on 2012-2013 as an aberration. Scaled error reveals that GFT also predicted flu prevalence that is about 65% higher than the CDC estimate in 2011-2012, and missed high in 100 out of 108 weeks from August 21, 2011 to September 1, 2013. All of this suggests that GFT’s problems started earlier than is usually thought and might not correlate as highly with spikes in media coverage.

"Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season", Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. *SSRN Electronic Journal,* January 2014

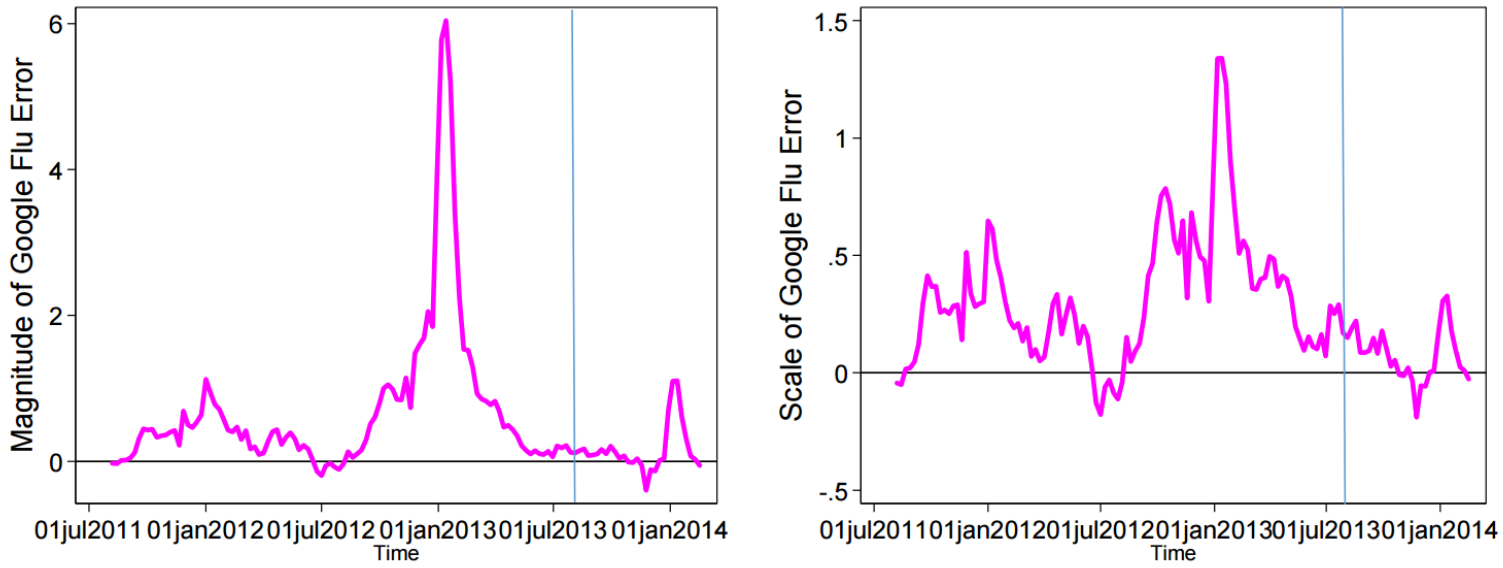


Fig 4. GFT vs. CDC estimates after latest GFT update. Absolute (**Left**) and proportional (**Right**) error of GFT from 2011 to 2014. Observations to the right of the blue line are after GFT started its new algorithm. While the update has dampened the size of GFT estimates (by about 12% for those observations in which we have overlap between the old and new model), GFT is still estimating high almost 75% of the time. It also still estimated about 30% higher than the CDC in the 2013-2014 flu season.