



Launching the next generation of digital disease surveillance tools

Mauricio Santillana

Faculty member, CHIP, Boston Children's Hospital Informatics Program

Associate, Harvard Institute for Computational and Applied Sciences

Instructor, Harvard Medical School



HARVARD
School of Engineering
and Applied Sciences

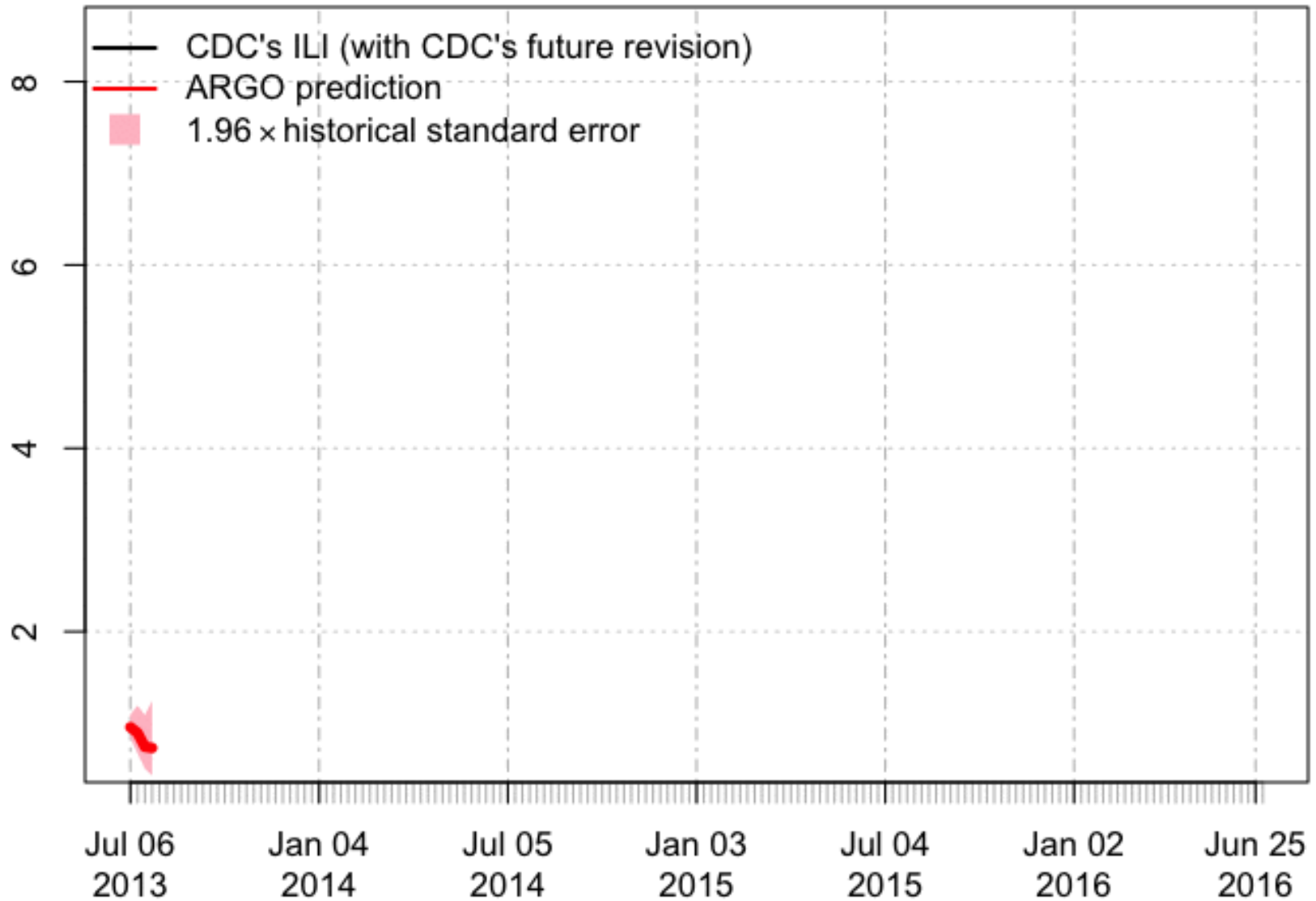


HARVARD
MEDICAL SCHOOL

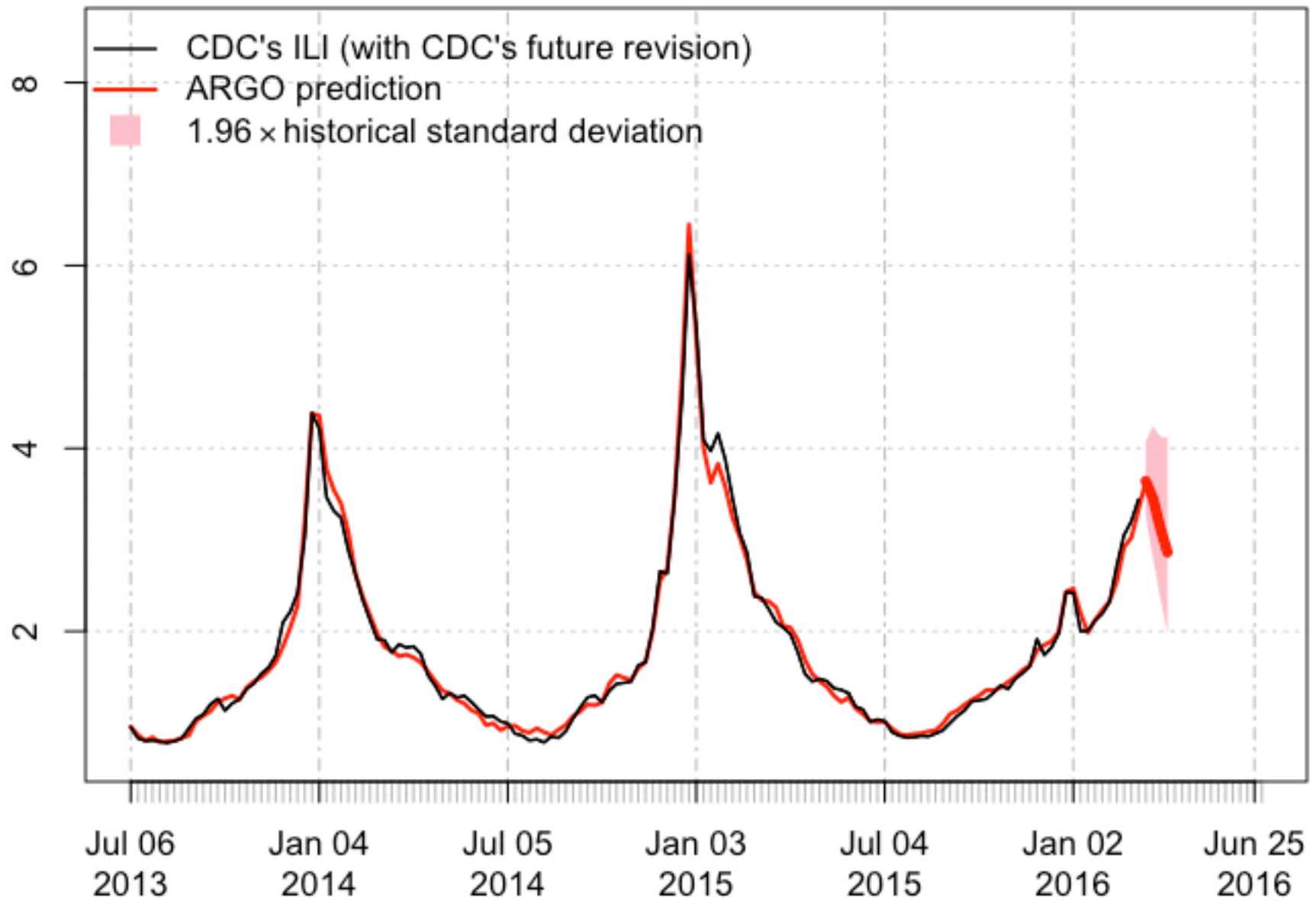


**Boston
Children's
Hospital**

ARGO Prediction vs. CDC's ILI



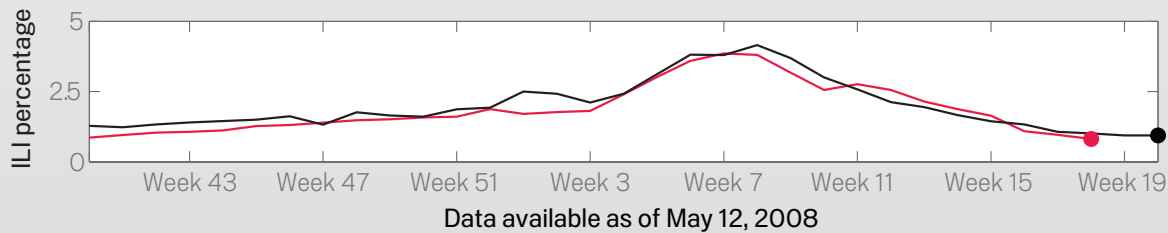
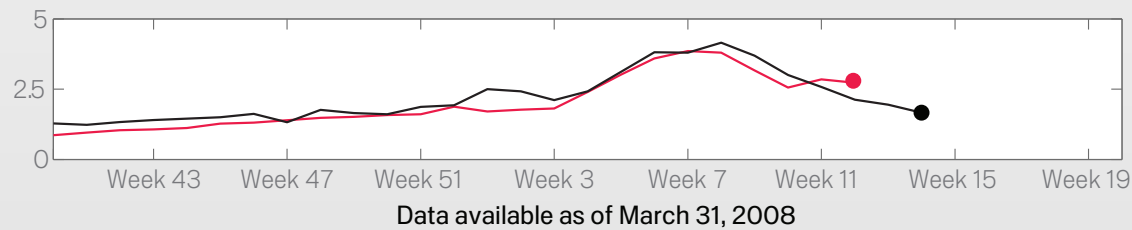
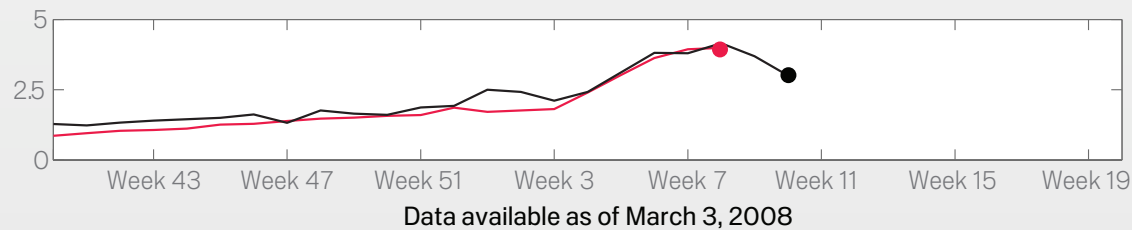
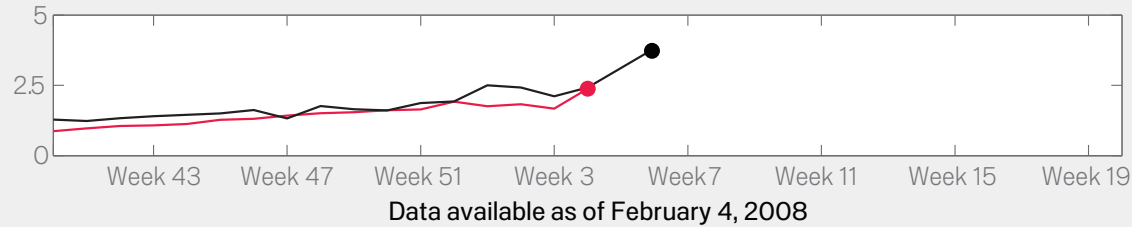
ARGO Prediction vs. CDC's ILI



The promise of big data in public health

GOOGLE FLU TRENDS

Epidemiological information available 2-3 weeks ahead of traditional clinical tracking systems



nature

International weekly journal of science

Letter

Nature **457**, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; [Corrected](#) 19 February 2009

Detecting influenza epidemics using search engine query data

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

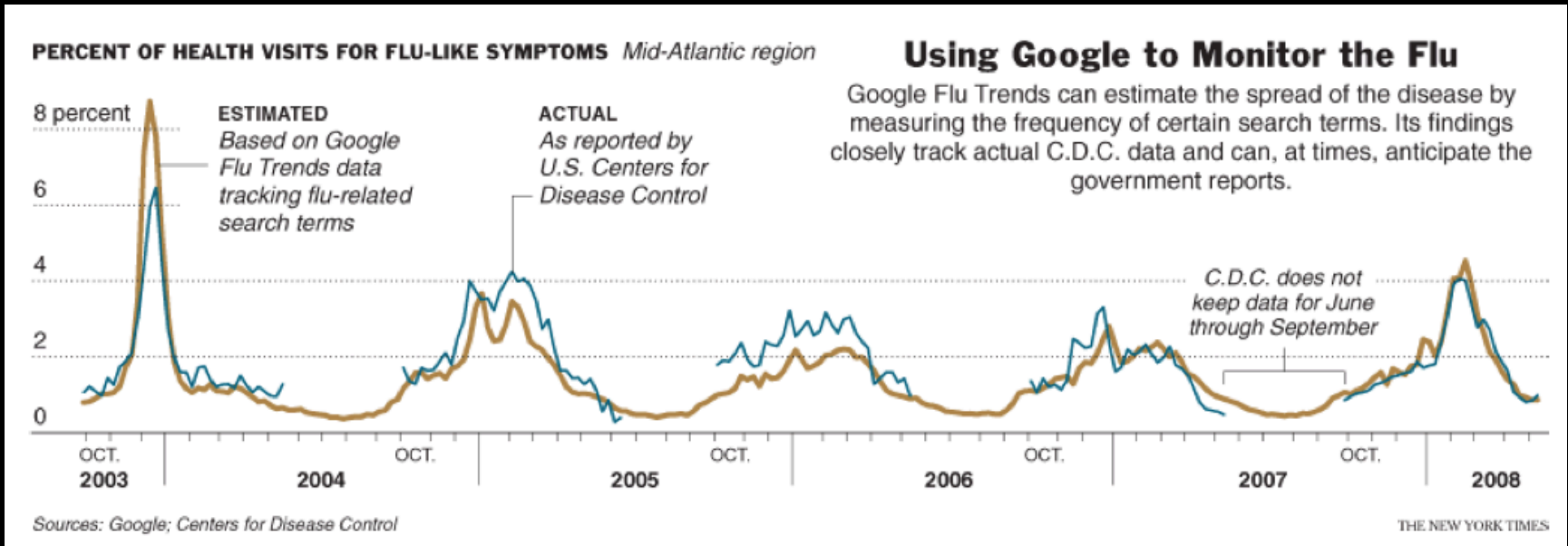
1. Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA

2. Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA

Correspondence to: Matthew H. Mohebbi¹ Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-support@google.com).

The New York Times

Very promising retrospective comparison!



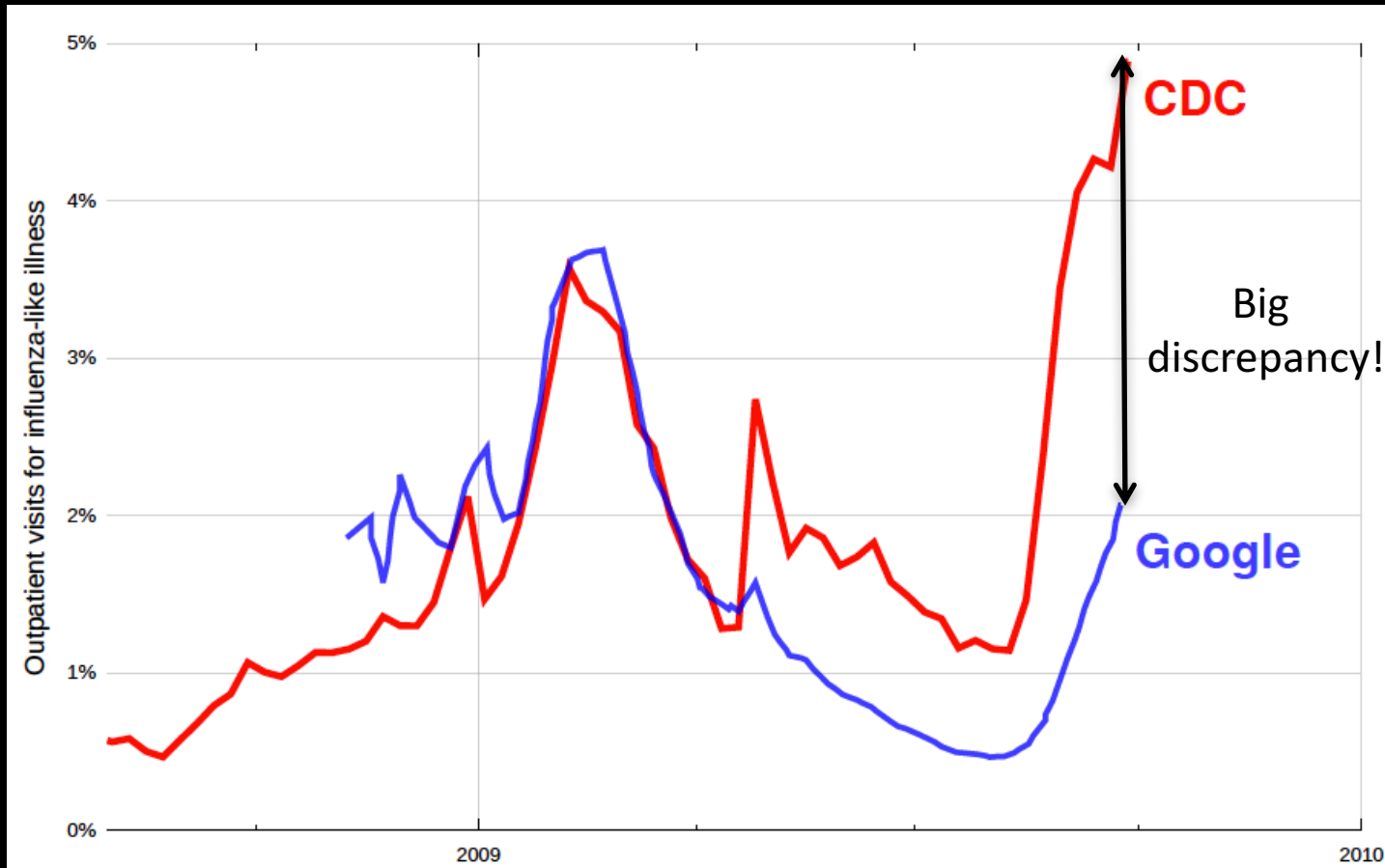
In April 2009, Dr. Brilliant said it epitomized the power of Google's vaunted engineering prowess to make the world a better place, and he predicted that it would save untold numbers of lives.

Google Flu Trends

launched in November 2008

Real-time performance, first year...

Big errors seen during H1N1 pandemic (off-season)



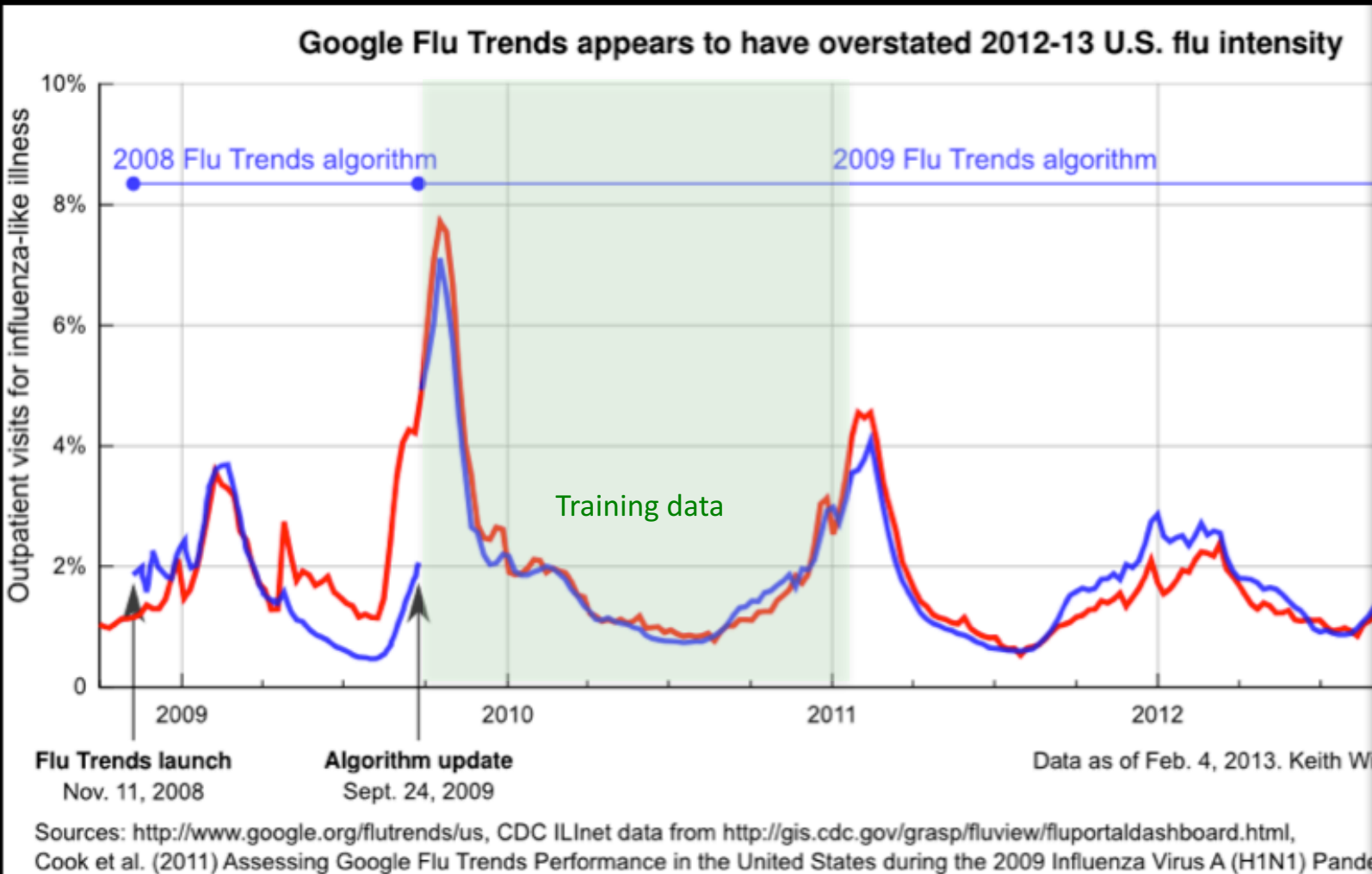
To some extent GFT was good at predicting seasons: fall-winter, not flu!

Plot obtained from:

<http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>

What next?

need to remove (not useful) search terms

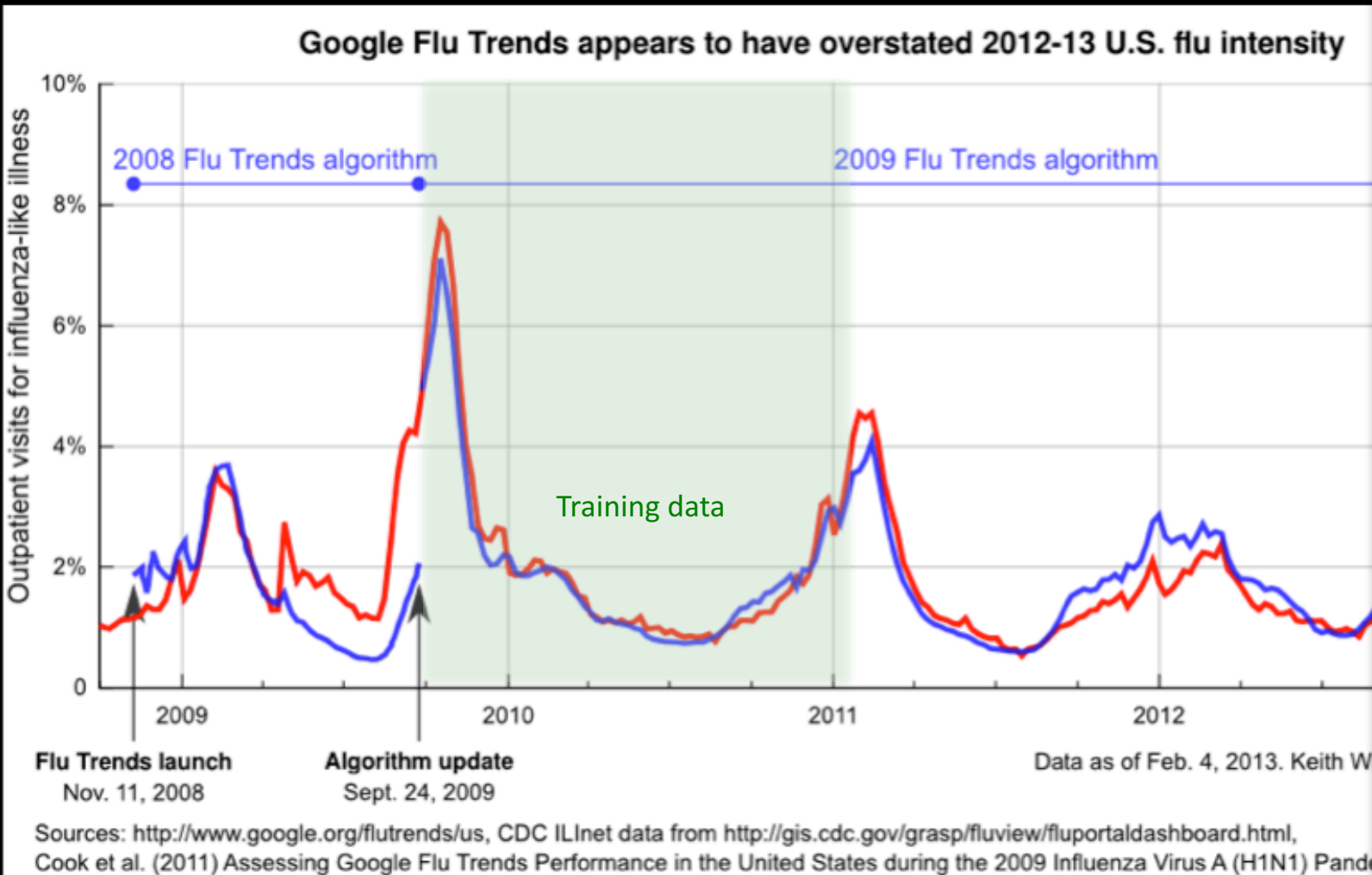


Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: <http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>

What next?

need to remove (not useful) search terms



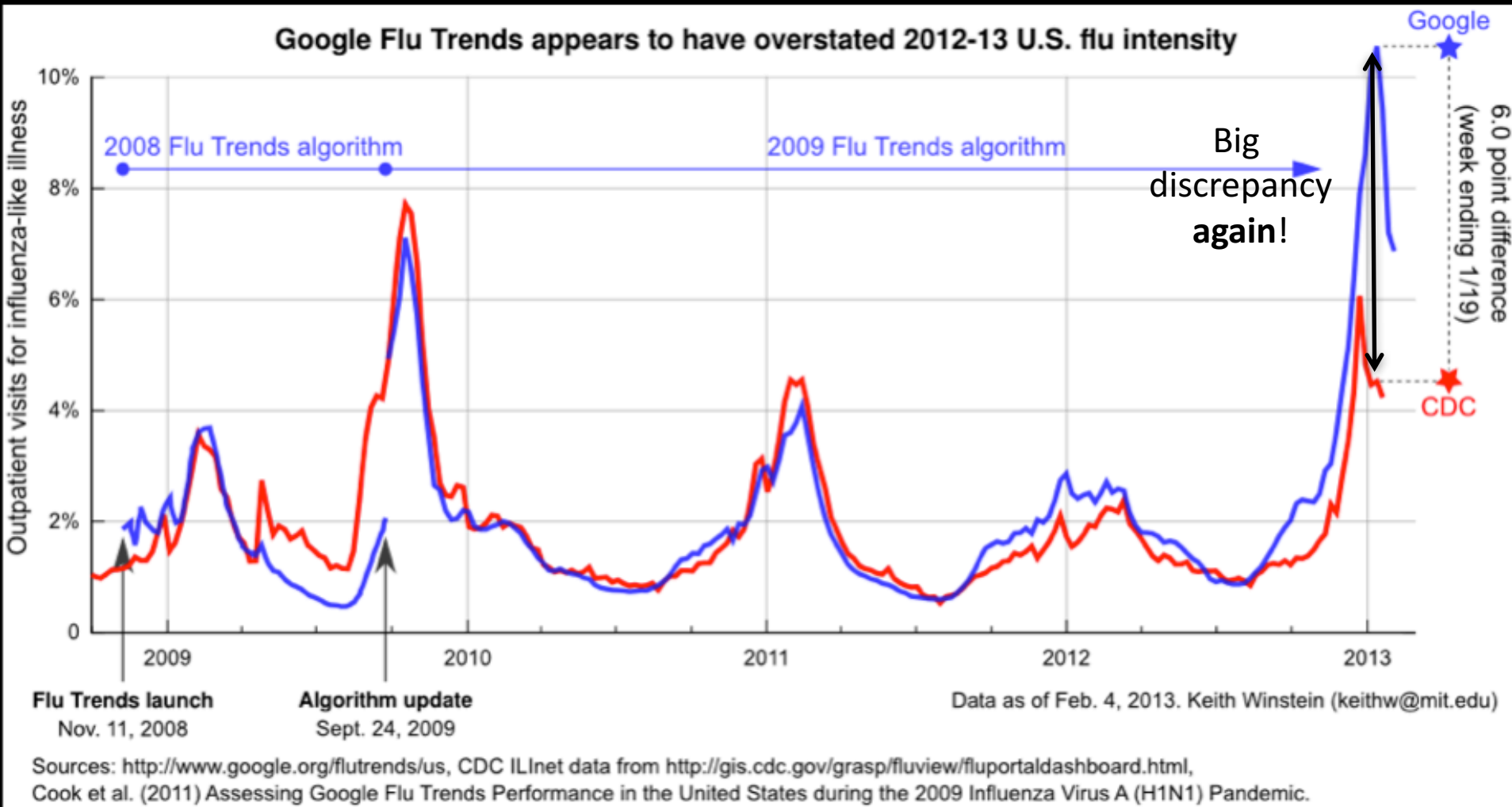
promising
performance

Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: <http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>

What next? need to remove (not useful) terms.

Big discrepancies again!



Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: <http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>

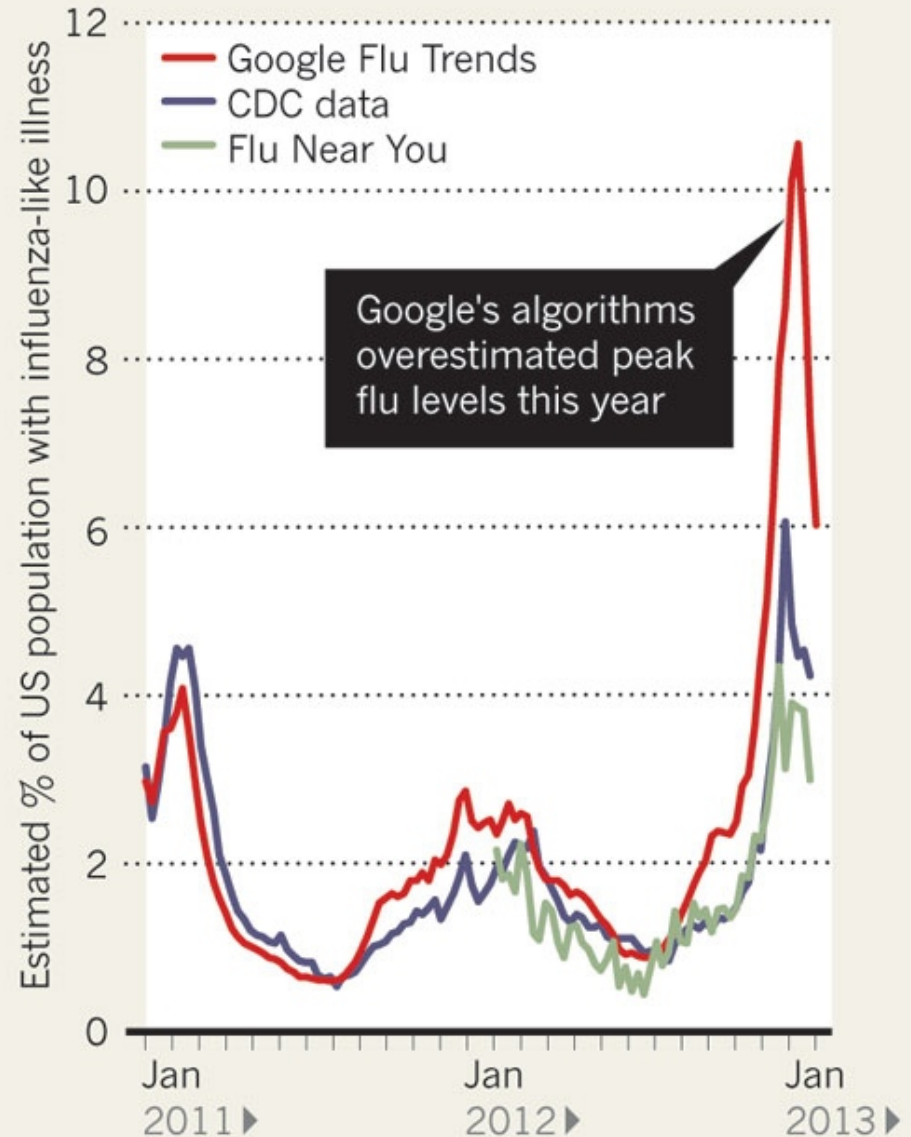


When Google got flu wrong.

nature.com/news/when-google-got-flu-wrong.

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



Snowden And The Challenge Of Intelligence: The Practical Case Against The NSA's Big Data

63

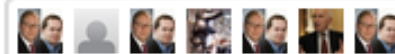
f Share

106

🐦 Tweet

61

in Share



12 comments, 7 called-out

+ Comment Now

+ Follow Comments

“ We should soon be able to keep track of most activities on the surface of the earth, day or night, in good weather or bad.”

s i l i c o n ^ a n g l e

where computer science meets social science

{SILICON ANGLE}

CLOUD

MOBILE

SOCIAL

SERVICES

DEVOPS

RESEARCH

SiliconANGLE » Can Nate Silver's Data Culture Lead Us Out Of The NSA + Public Data Scare?

Can Nate Silver's Data Culture Lead Us Out of the NSA + Public Data Scare?

RYAN COX | SEPTEMBER 18TH

READ MORE

hive | A

al flu.

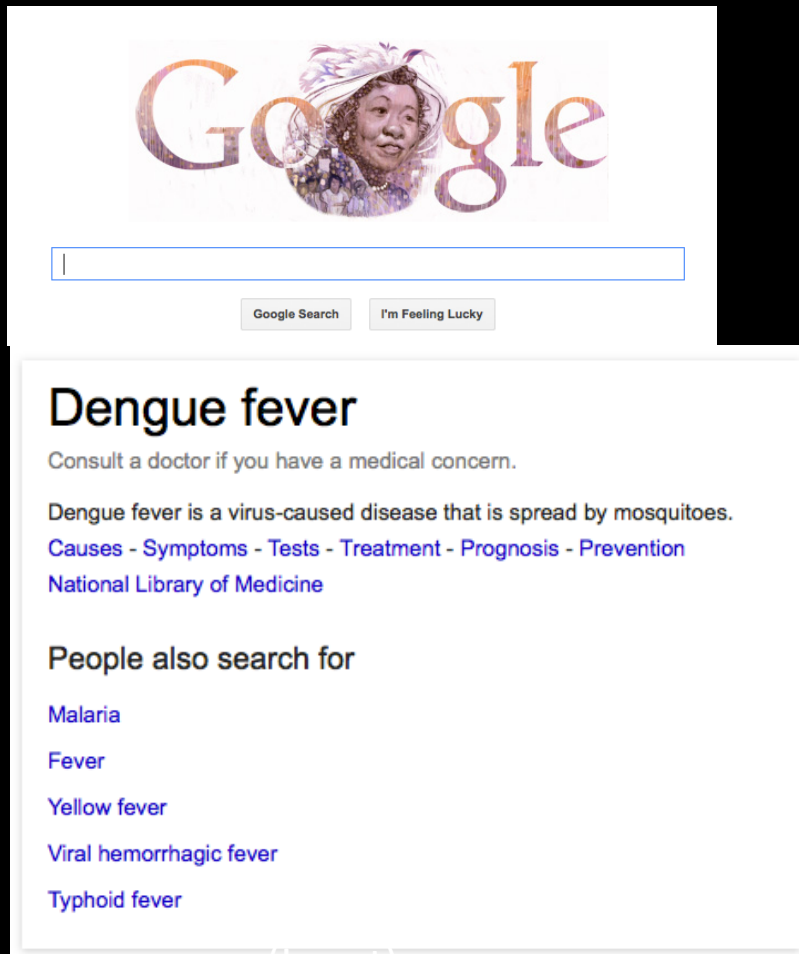
ancy

Lessons learned

Let's work on a short exercise to understand how Google Flu Trends used to work...

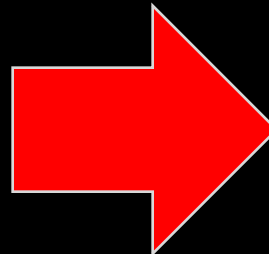
Supervised machine learning examples:

Given the number of Google searches associated to the term “dengue”, and given the number of confirmed cases of dengue in Mexico from 2004 to 2006 (Training period), can we estimate how many people will most likely get dengue based on the number of searches during the subsequent years?



The image shows a Google search interface. At the top, the word "Google" is displayed in its characteristic multi-colored font, with a circular inset showing a woman's face. Below the logo is a search bar with a cursor. Underneath the search bar are two buttons: "Google Search" and "I'm Feeling Lucky". The search results for "Dengue fever" are displayed below. The title "Dengue fever" is in a large, bold, black font. Below the title is a sub-header: "Consult a doctor if you have a medical concern." The main text reads: "Dengue fever is a virus-caused disease that is spread by mosquitoes." Below this text are several blue hyperlinks: "Causes - Symptoms - Tests - Treatment - Prognosis - Prevention" and "National Library of Medicine". Further down, there is a section titled "People also search for" with a list of related terms: "Malaria", "Fever", "Yellow fever", "Viral hemorrhagic fever", and "Typhoid fever".

(Input)



(Output)

4. Least squares in Public Health. (30 points)

Dengue fever is a virus-caused disease that is spread by mosquitoes that affects millions of people in tropical environments around the Globe. In this problem, you are asked to construct a simple version of the digital disease detection tool: “Google Dengue Trends” for Mexico. For this, you will download the spreadsheet `Dengue trends AM 111.xls` from the course website. The first column in the spreadsheet represents the date (in months, from 2004-2011), the second column represents the number of Google searches of the term “dengue” in Mexico, in a given month. The third column represents the number of cases of Dengue in Mexico, as reported by the Mexican Ministry of Health. You may use Matlab or Excel for this problem.

- (a) Plot the number of cases of Dengue as a function of time.
- (b) For the **training period** 2004-2006 (36 months), find the best line that explains the number of cases of Dengue as a function of the number of searches of the term “dengue”. You should do this by solving the least squares problem, and you should obtain the value of the y-intercept and the slope.
- (c) Use the equation of the line you obtained in (b) and plot the number of cases as a function of the number of searches of the term “dengue”, predicted by your method during the training period. Compare your results to the plot in (a) for such time period.
- (d) For the **prediction or validation period** 2007-2011, use the equation of the line you obtained in (b) to predict the number of the dengue cases as a function of the number of searches of the term “dengue” from 2007-2011. Plot your predictions and compare them to the actual number of cases.
- (d) Discuss your results. Could you improve this modeling approach? If so, how?

Did you get it to work?

Using Google searches to track diseases statically

begin

%% Load data %%

CDC=load(CDC ILI Data) (ONE COLUMN OF VALUES)

X=load(Google search Data) (MULTIPLE COLUMNS OF VALUES)

%% initialize output array %%

Y=zeros(1: end.of.predictions) (INITIALIZE ARRAY TO STORE PREDICTIONS)

%% train model and produce predictions %%

CDC ← standardize(CDC) (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)

X ← standardize(X) (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)

model=LASSOroutine.fit(CDC[1 : *training*] ~ X[1 : *training*]) (TRAINING: IN-SAMPLE
MODEL)

Y[1 : *training*]=

LASSOroutine.predict(model, X[1 : *training*])
(IN-SAMPLE PREDICTIONS)

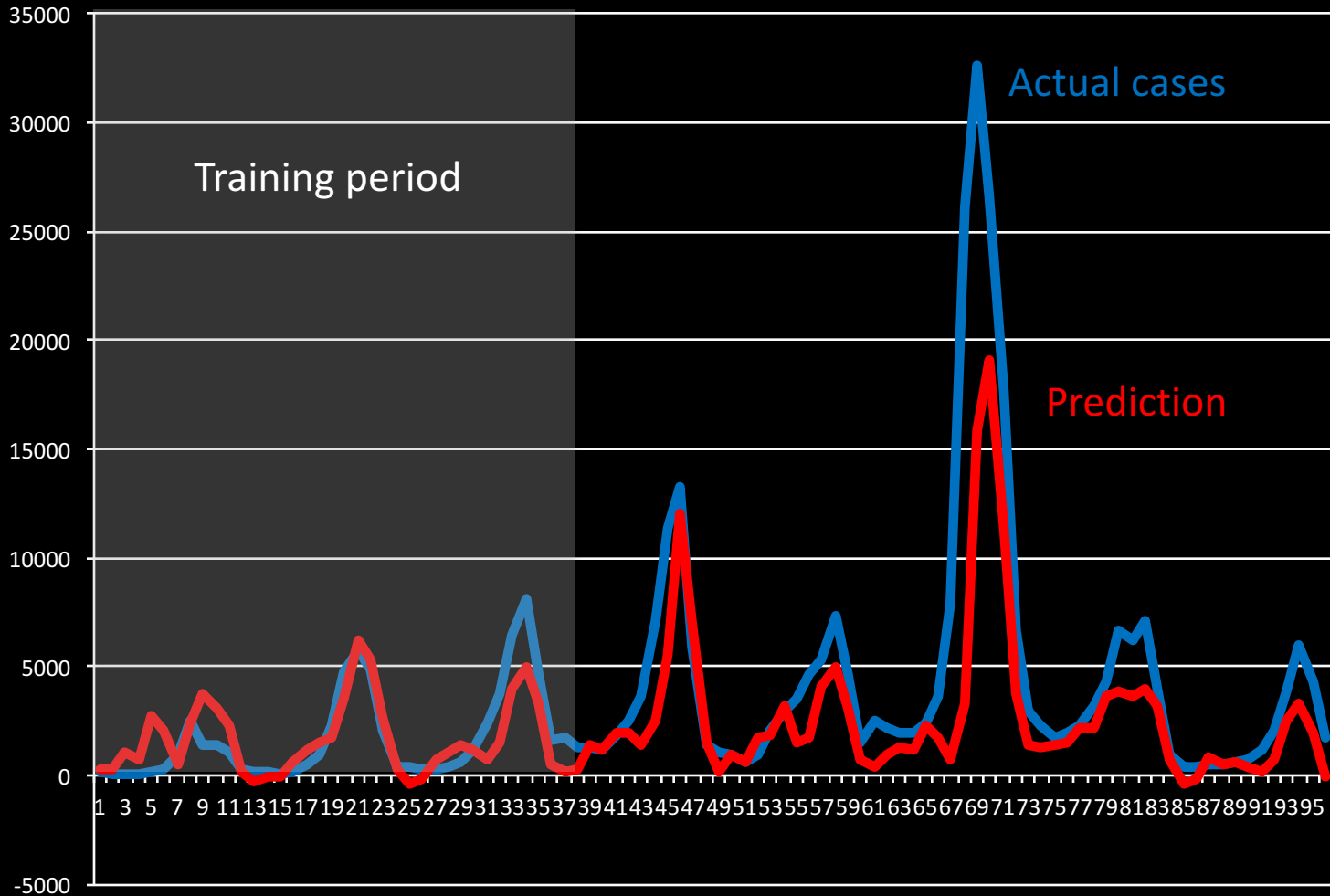
Y[*training* + 1 : *end.of.predictions*]=

LASSOroutine.predict(model, X[*training* + 1 : *end.of.predictions*])
(PRODUCE OUT-OF-SAMPLE PREDICTIONS)

end

Supervised machine learning examples:

Static approach, fixed training set



How could the previous approach be improved with the given information?

Using Google searches to track diseases dynamically

```

begin
%% Load data %%
CDC=load(CDC ILI Data)           (ONE COLUMN OF VALUES)
X=load(Google search Data)       (MULTIPLE COLUMNS OF VALUES)

%% initialize output arrays %%
Y=zeros(1:end.of.predictions)   (INITIALIZE ARRAY TO STORE PREDICTIONS)
coefficients=zeros(1:end.of.predictions) (INITIALIZE ARRAY TO STORE COEFFS)

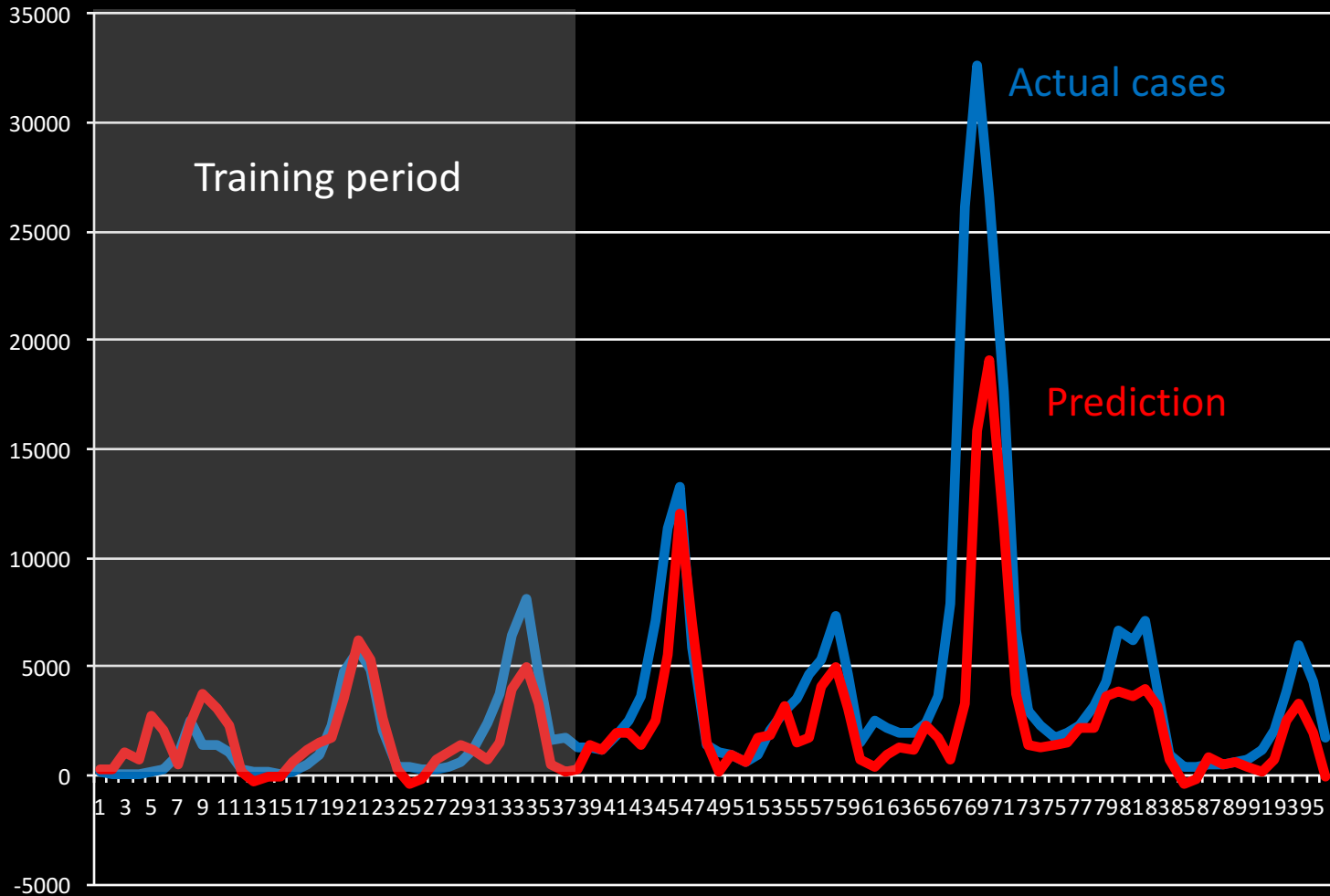
%% train models and produce out-of-sample predictions %%
for i = training : end.of.predictions
    CDC ← standardize(CDC)       (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)
    X ← standardize(X)           (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)
    model=LASSOroutine.fit(CDC[1 : i] ~ X[1 : i]) (TRAINING: IN-SAMPLE MODEL )
    coefficients(i) ← model(coefficients)
    Y(i + 1)=LASSOroutine.predict(model, X(i + 1)) (PRODUCE OUT-OF-SAMPLE
                                                    PREDICTIONS)

    if(i == training)
        Y[1:i]=LASSOroutine.predict(model, X[1:i]) IN -SAMPLE PREDICTIONS
    end
end
end
end

```

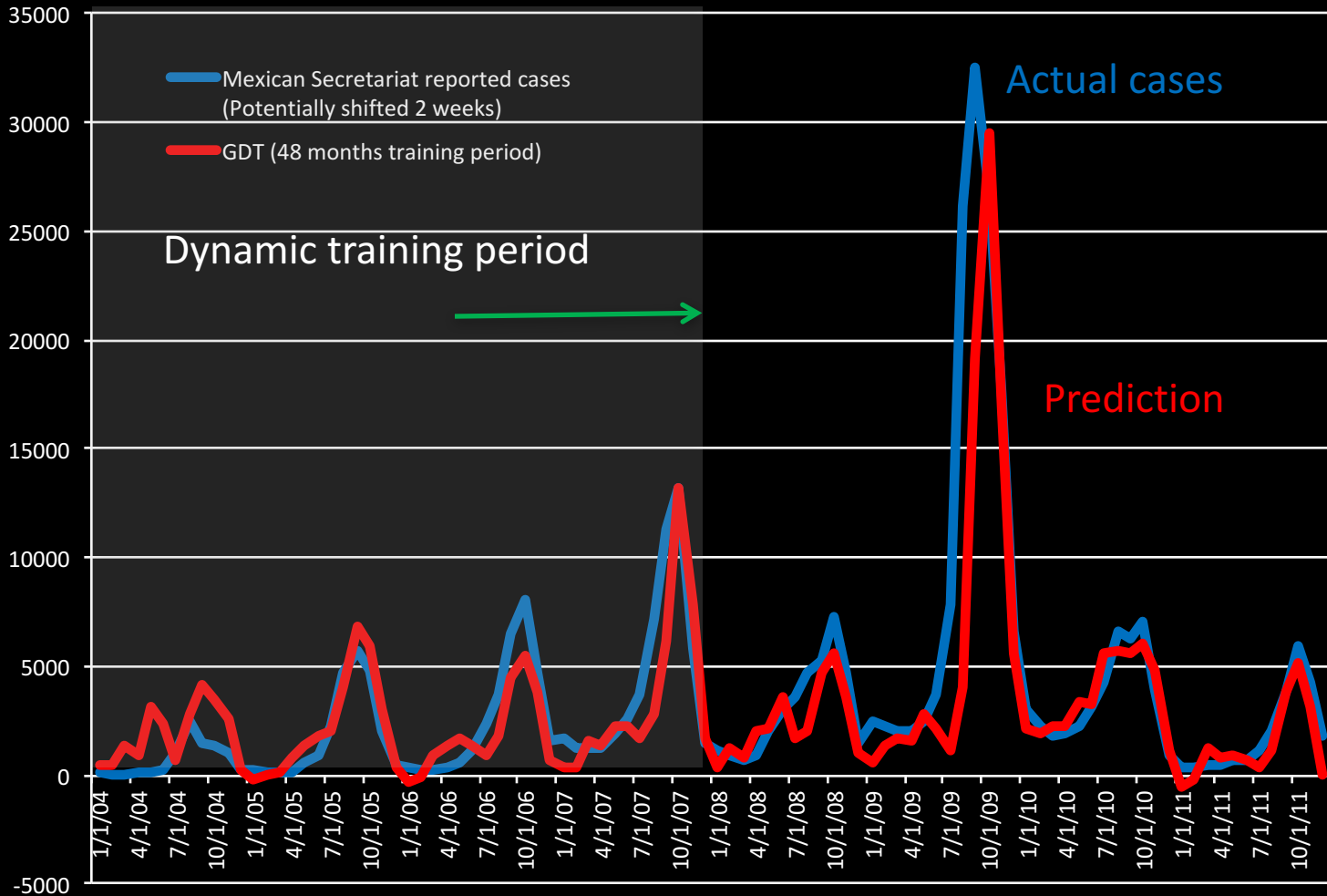
Supervised machine learning examples:

Static approach, fixed training set



Supervised machine learning examples:

Dynamic approach, letting the training set expand as more information becomes available



How could the previous approach be improved?

Assumptions in Google Flu Trends:

1. Number of (influenza-like) ill people proportional to number of **total** searches of (Influenza-like illnesses) related terms

$$\mathit{logit}(P) = \beta_0 + \beta_1 \times \mathit{logit}(Q) + \varepsilon$$

where P is the percentage of ILI physician visits, Q is the ILI-related query fraction, β_0 is the intercept,

Assumptions in Google Flu Trends:

1. Number of (influenza-like) ill people proportional to number of **total** searches of (Influenza-like illnesses) related terms

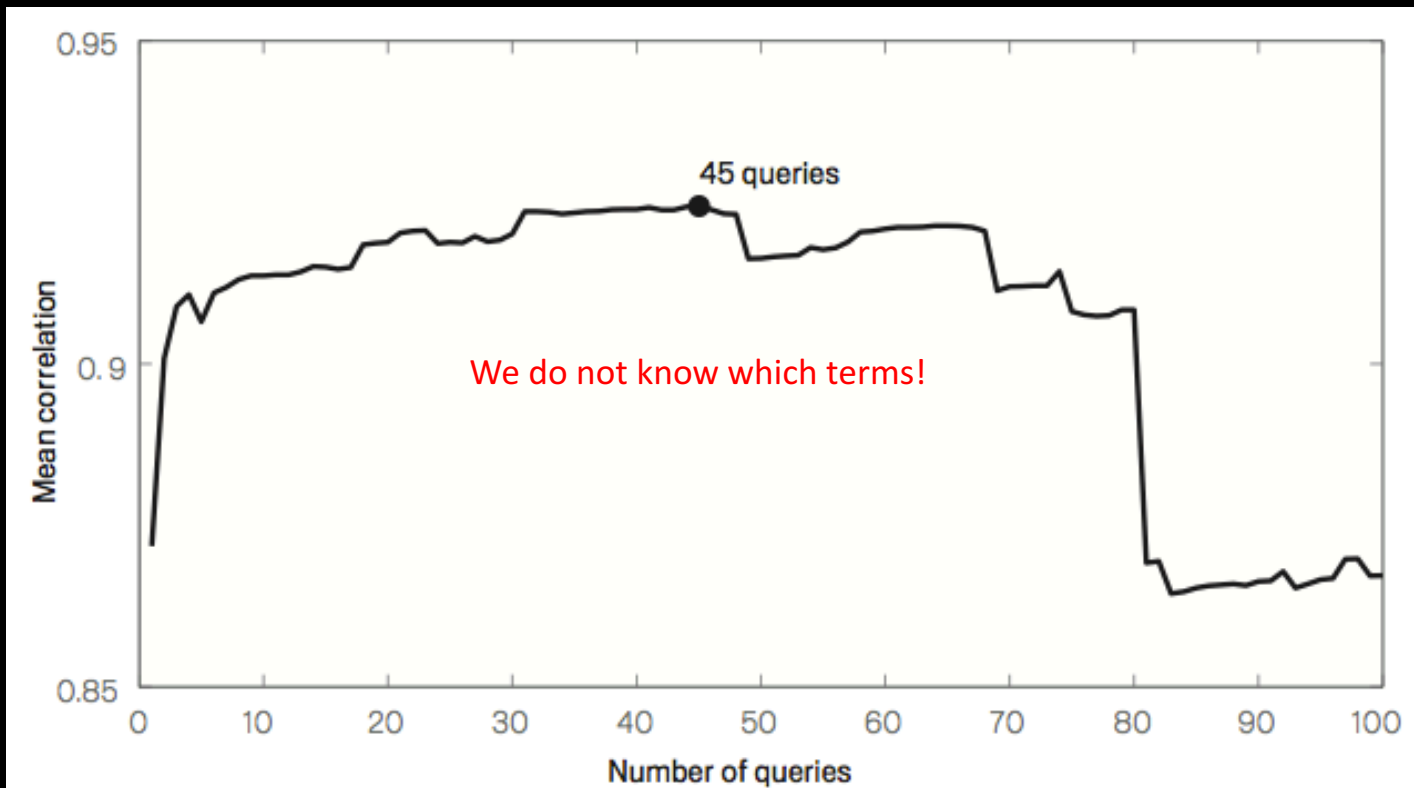


Figure 1: An evaluation of how many top-scoring queries to include in the ILI-related query fraction. Maximal performance at estimating out-of-sample points during cross-validation was obtained by summing the top 45 search queries. A steep drop in model performance occurs after adding query 81, which is "oscar nominations".

Assumptions in Google Flu Trends:

2. Relationship between search volume and proportion of (influenza like) ill people is **static** (during a given year).

Assumptions in Google Flu Trends:

2. Relationship between search volume and proportion of (influenza like) ill people is **static** (during a given year).

Consequences: Model needed constant supervision by human experts

- a. **Human experts** needed to **assess** relevance of individual search terms,
- b. **Human Experts** needed to **recalculate** relationship between total number of searches and ill people, and
- c. It is bound to **deliver poor predictions** at some point in the near future!

We proposed an alternative method and tested it using low quality input from Google Correlate in January 2013.

(with D. Wendong Zhang)

New model:

1. Each search term may contribute to prediction of ILI rate separately (**multi-variate approach**)
2. Relationship between search volume for each individual term and proportion of ill people is **dynamic** and should be found using supervised machine learning optimization techniques.

$$\beta^{lasso} = \arg \min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^M x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^M |\beta_j| \right\}$$

Every week the multiplicative coefficients (β 's) would be automatically updated by expanding the training set (labeled data) as new information from the CDC became available.

Top correlated terms to CDC-reported data from 1/2004- 3/2009 (using Google Correlate)

1	influenza type a	35	is the flu contagious	68	fever in adults
2	bronchitis	36	flu in children	69	decongestant
3	influenza a	37	fever flu	70	normal body
4	symptoms of pneumonia	38	take action tour	71	low body temperature
5	flu incubation	39	flu remedies	72	a fever
6	influenza incubation	40	flu report	73	influenza a symptoms
7	flu contagious	41	nasal congestion	74	dangerous fever
8	influenza contagious	42	fever reducer	75	is flu contagious
9	flu incubation period	43	sinus infections	76	lauderdale florida
10	tussionex	44	rhode island wrestling	77	hotel fort lauderdale
11	benzonatate	45	symptoms of influenza	78	webmail shaw ca
12	influenza symptoms	46	castaway bay	79	high fever
13	a influenza	47	coral by the sea	80	robitussin ac
14	sinus	48	cold or flu	81	bronchitis contagious
15	pneumonia	49	respiratory infection	82	indoor driving
16	flu fever	50	take action	83	tussionex pennkinetic
17	flu duration	51	respiratory flu	84	wrestling report
18	taste of chaos	52	soweto gospel	85	walking pneumonia
19	bronchitis symptoms	53	soweto gospel choir	86	days inn miami
20	symptoms of bronchitis	54	illinois wrestling	87	body temperature
21	how long does the flu last	55	how long is the flu contagious	88	phlegm
22	symptoms of the flu	56	cold symptoms	89	flu relief
23	taste of chaos tour	57	the taste of chaos	90	mt sunapee
24	influenza incubation period	58	is bronchitis	91	harlem globe
25	sinus infection	59	upper respiratory	92	levaquin
26	flu recovery	60	afrin	93	strep throat
27	chaos tour	61	painful cough	94	coughing
28	type a influenza	62	laprepsoccer	95	whistler snow
29	flu symptoms	63	upper respiratory infection	96	fever temperature
30	tessalon	64	amoxicillin	97	sales tax credit
31	type a flu	65	ski harness	98	glitches
32	treat the flu	66	robitussin dm	99	pennkinetic
33	treating the flu	67	treating flu	100	histinex
34	how to treat the flu				

In October 2012 Google decided to discontinue updating
Google Correlate (it was Jan 2013)



In October 2012 Google decided to discontinue updating
Google Correlate (it was Jan 2013)

We used even lower quality data from Google Trends to test
methodology in 2012-2013 recent flu season



What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, PhD, MS, D. Wendong Zhang, MA, Benjamin M. Althouse, PhD, ScM,
John W. Ayers, PhD, MA

© 2014 Published by Elsevier Inc. on behalf of American Journal of Preventive Medicine Am J Prev Med 2014;47(3):341–347 **341**

First week after being published online, it became the second most read paper in journal's history! (After a paper published in 1998)

AMERICAN JOURNAL OF Preventive Medicine

A Journal of the American College of Preventive Medicine and Association for Prevention Teaching and Research

Articles in Press

Most Read

Most Cited

Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study

Vincent J Felitti, Robert F Anda, Dale Nordenberg, David F Williamson, Alison M Spitz, Valerie Edwards, Mary P Koss, James S Marks

Vol. 14, Issue 4

Published in issue: May, 1998

[Abstract](#) | [Full-Text HTML](#) | [PDF](#)

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, D. Wendong Zhang, Benjamin M. Althouse, John W. Ayers

Vol. 47, Issue 3

Published online: July 1, 2014

[Abstract](#) | [Full-Text HTML](#) | [PDF](#)

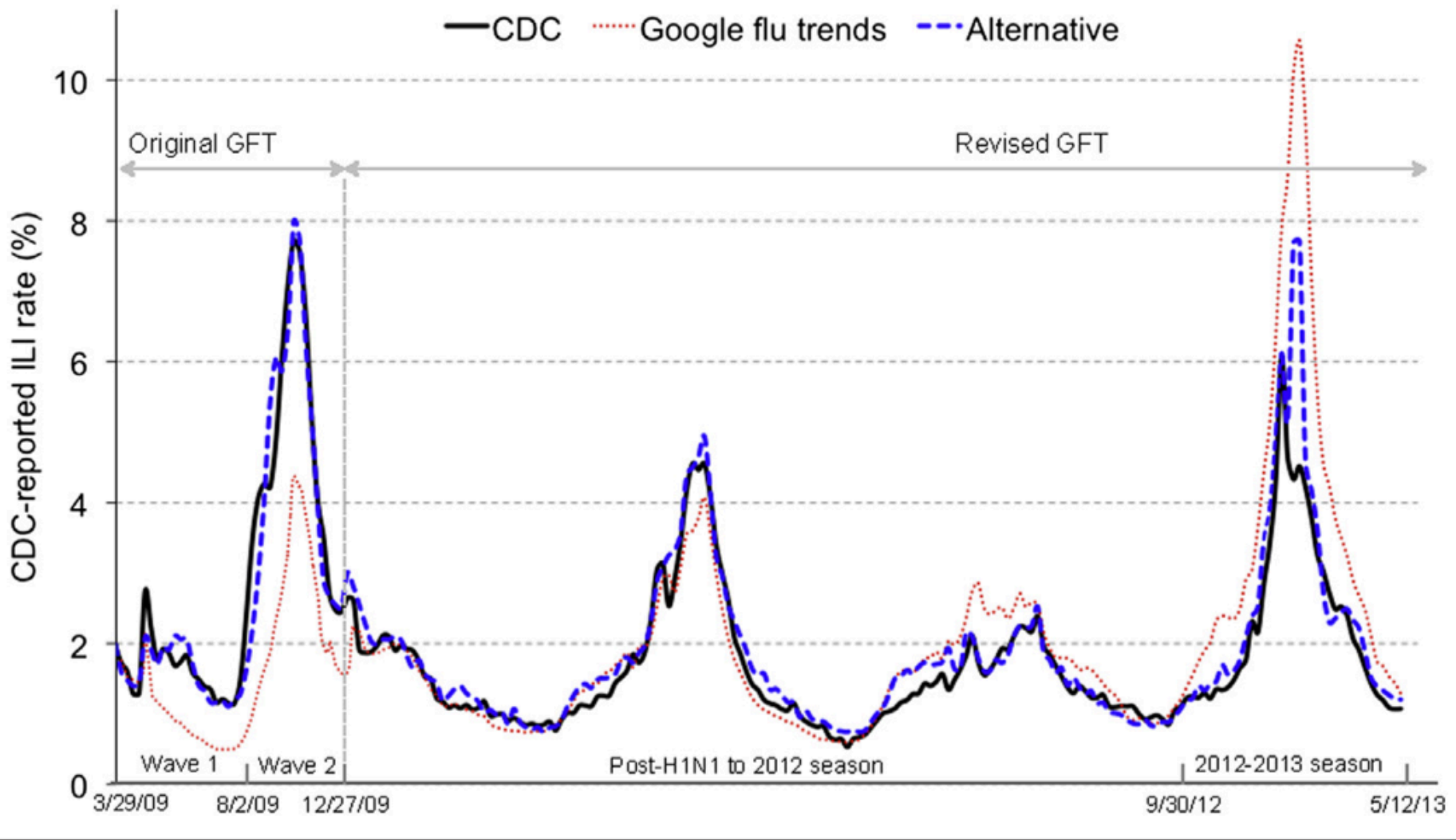
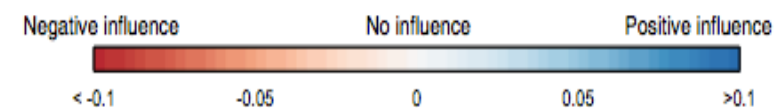
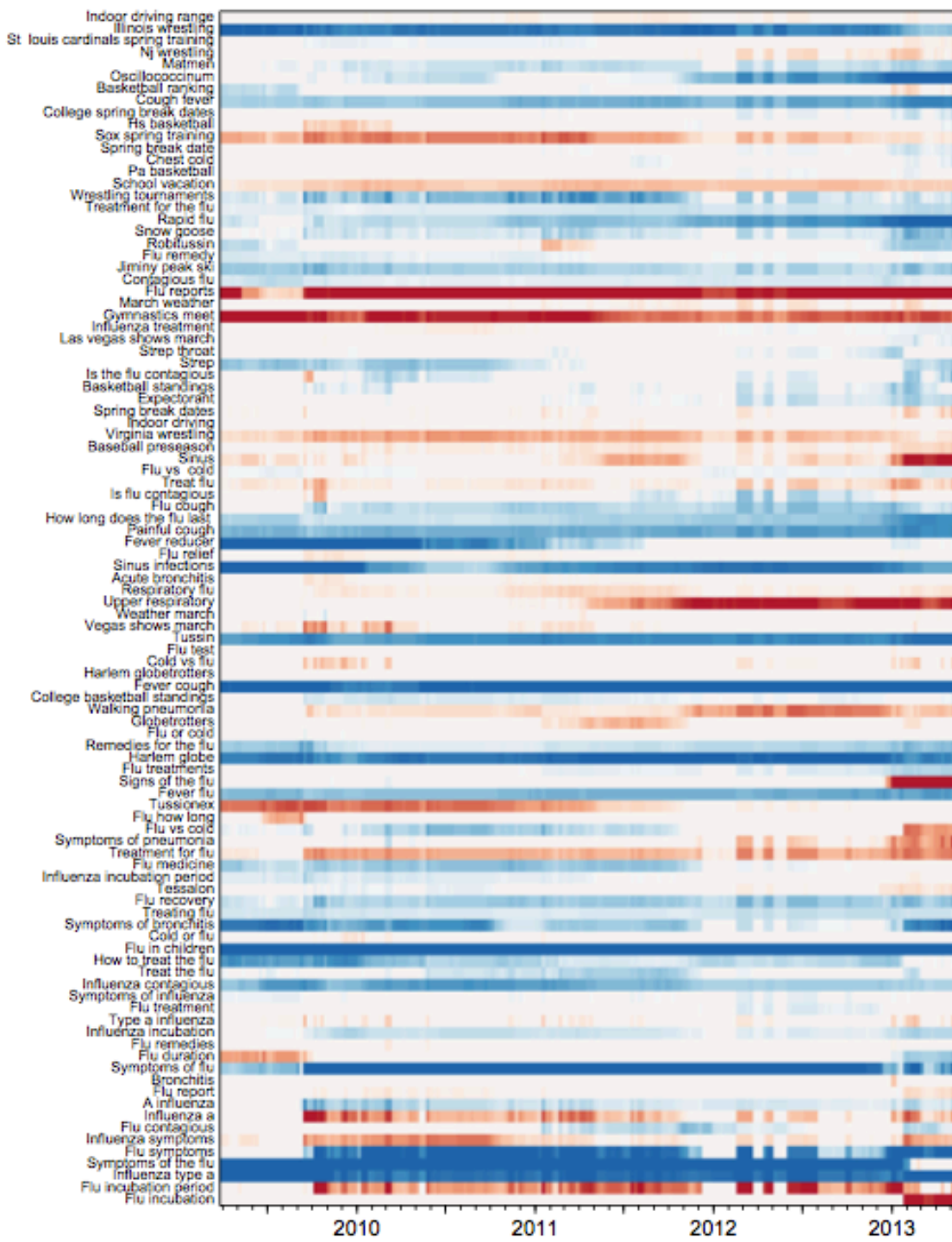
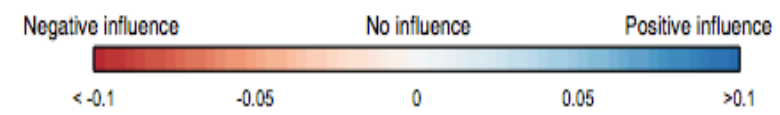
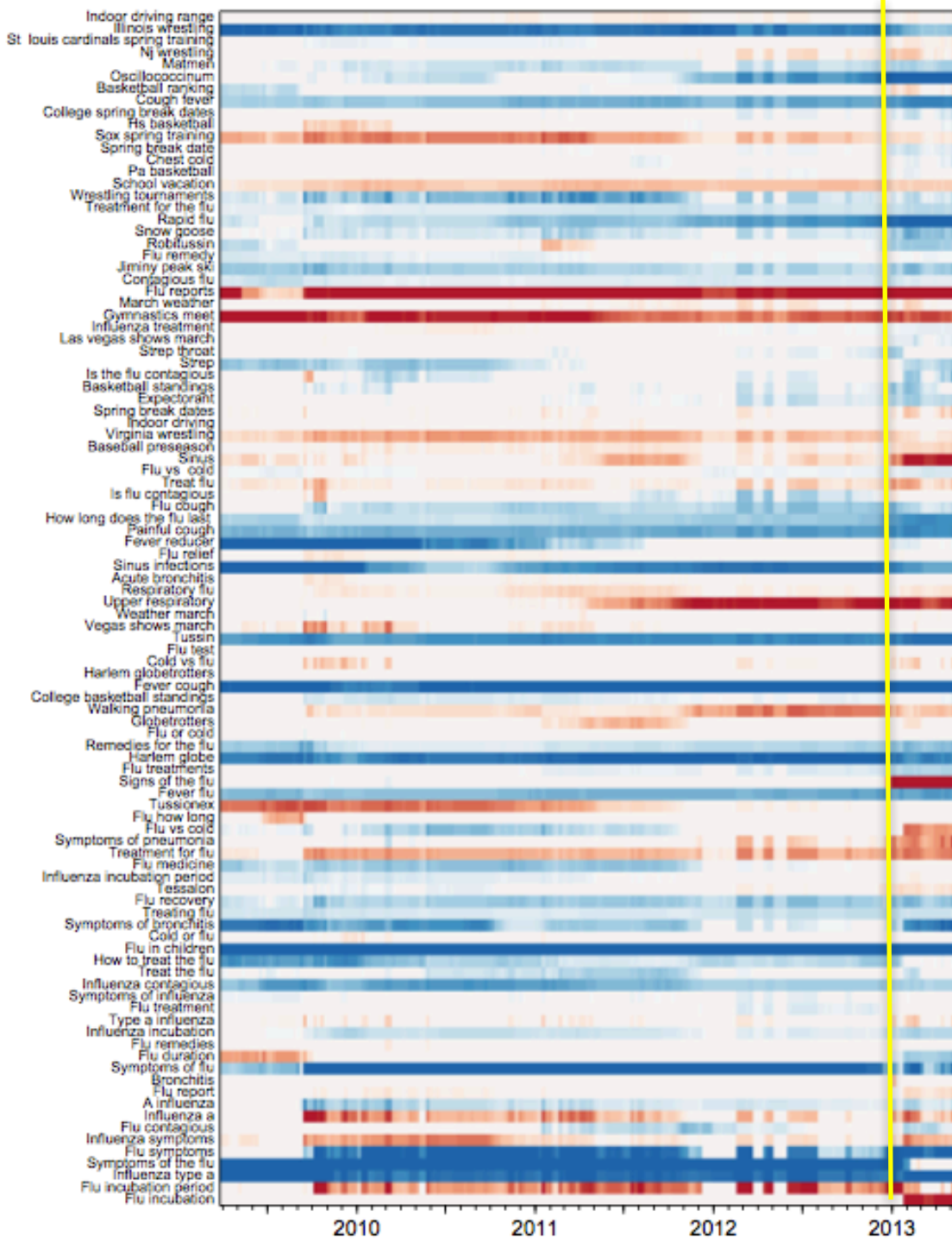


Figure 1. The alternative model outperforms Google Flu Trends

$$\text{logit}[I(t)] = \sum_{i=1}^n a_i(t) \text{logit}[Q_i(t)] + e,$$



Santillana et al. American Journal of Preventive Medicine, 2014; 47 (3) pp 341-347




Santillana et al. American Journal of Preventive Medicine, 2014; 47 (3) pp 341-347

Google Flu Trends promises are overstated, researchers say

New study finds way to improve Google Flu Trends accuracy threefold - but says systems must be more open

Charles Arthur

 Follow @charlesarthur

 Follow @guardiantech

theguardian.com, Friday 4 July 2014 11.44 EDT

Google Flu Trends promises are overstated, researchers say

New study finds way to improve Google Flu Trends accuracy threefold - but says systems must be more open

HealthData Management

NEW

POLICY & REGULATION

EHR

HEALTH INFO EXCHANGE

REVENUE CYCLE & PAYMENTS

CHRONIC CARE

Researchers Suggest Fixes to Google Flu Trends Analytics

A new study concludes that "revising the inner plumbing" of the Google Flu Trends disease surveillance system can improve the accuracy of forecasts for the severity of a flu season.

Featured Research

from universities, journals, and other organizations

Google Flu Trends is overstated, research

New study finds way to improve threefold - but says systems mu

Finding real value in big data for public health

Date: July 2, 2014

Source: San Diego State University

Summary: Media reports of public health breakthroughs from big data have been largely oversold, according to a new study. But don't throw away that data just yet. The authors maintain that the promise of big data can be fulfilled by tweaking existing methodological and reporting standards. In the study, the research team demonstrate this by revising the inner plumbing of the Google Flu Trends (GFT) digital disease surveillance system, which was heavily criticized last year (see here and here) after producing erroneous forecasts.

Share This

- > Email to a friend
- > Facebook
- > Twitter
- > LinkedIn
- > Google+
- > Print this page

HealthData Management

POLICY & REGULATION

EHR

HEALTH EXCHANGE

Related Topics

Health & Medicine

- > Health Policy
- > Public Health Education

Computers & Math

- > Computers and Internet
- > Computer Modeling

Science & Society

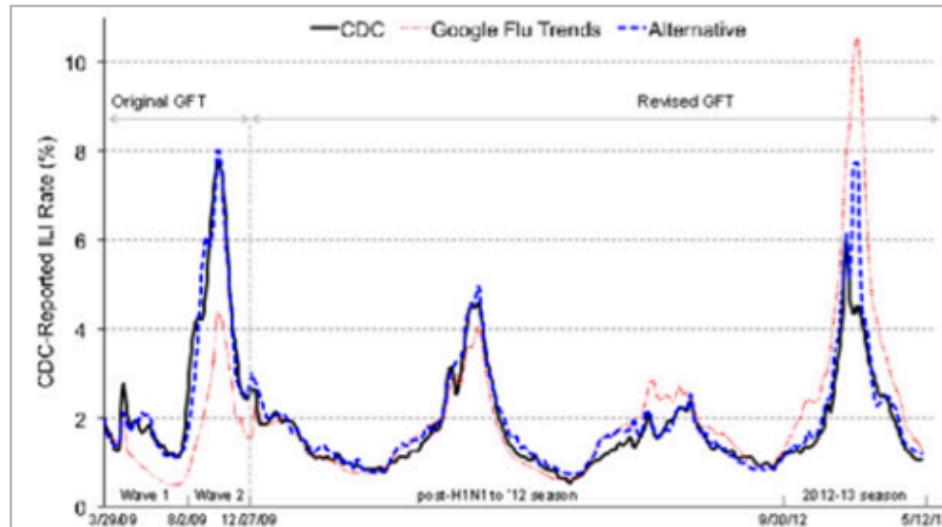
- > Public Health
- > Surveillance

Researchers Suggest Flu Trends Analysis

A new study concludes that "revising the inner surveillance system can improve the accuracy of the system"


Related Articles

- > Public health
- > Data mining



A graph depicting Google Flu Trends.

Google
oversta
New study
threefold - b



iHealthBeat

Reporting Technology's Impact on Health Care

HOME INSIGHT PERSPECTIVES PICTURE OF HEALTH

Hea

Ma

POLICY &
REGULATION

Resea

Flu Tre

A new study concludes that "revising the in... surveillance system can improve the accur

NEWS ARCHIVE

SHARE EMAIL PRINT REPUBLISH

Study: Methodology Changes Improve Google Flu Trend Accuracy

Monday, July 7, 2014

RELATED TOPICS:

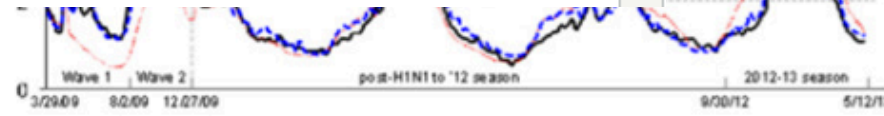
- Public Health

The accuracy of Google Flu Trends' disease surveillance system can be improved through simple changes in three different methodologies used by the system, according to a new study published in the *American Journal of Preventive Medicine, Health Data Management* reports (Goedert, *Health Data Management*, 7/7).

Related Articles

- > Public health
- > Data mining

This
mail to a friend
facebook
twitter
LinkedIn
Google+
Print this page



A graph depicting Google Flu Trends.

and other organization

Big Data's Potential in Public Health: Revisiting Google Flu Trends

July 7, 2014 Written by: Dan Gray 1 Reply



This

- mail to a friend
- facebook
- twitter
- LinkedIn
- google+
- print this page

Research Flu Trends

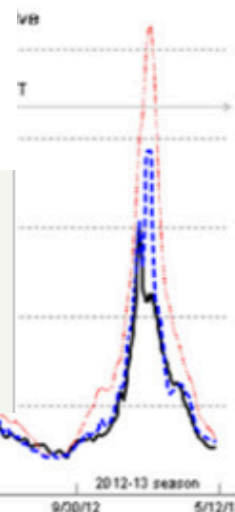
RELATED TOPICS:

- Public Health

The accuracy of Google Flu Trends disease surveillance system can be improved through simple changes in three different methodologies used by the system, according to a new study published in the *American Journal of Preventive Medicine, Health Data Management* reports (Goedert, *Health Data Management*, 7/7).

A new study concludes that "revising the in- surveillance system can improve the accur **Related Articles**

- > Public health
- > Data mining



A graph depicting Google Flu Trends.



Google Research Blog

The latest news from Research at Google

Big
Flu

July 7, 2014

Google Flu Trends gets a brand new engine

Posted: Friday, October 31, 2014

222

Tweet 161

Like 104

Posted by Christian Stefansen, Senior Software Engineer

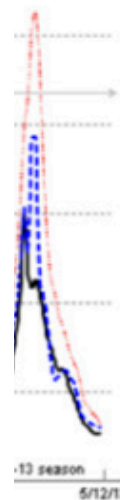
Each year the flu kills thousands of people and affects millions around the world. So it's important that public health officials and health professionals learn about outbreaks as quickly as possible. In 2008 we launched [Google Flu Trends](#) in the U.S., using aggregate web searches to indicate when and where influenza was striking in real time. These models [nicely complement](#) other survey systems—they're more fine-grained geographically, and they're typically more immediate, up to 1-2 weeks ahead of traditional methods such as the CDC's official reports. They can also be incredibly helpful for countries that don't have official flu tracking. Since launching, we've expanded Flu Trends to cover 29 countries, and launched [Dengue Trends](#) in 10 countries.

The original model performed surprisingly well despite its simplicity. It was retrained just once per year, and typically used only the 50 to 300 queries that produced the best estimates for prior seasons. We then left it to perform through the new season and evaluated it at the end. It didn't use the official CDC data for estimation during the season—only in the initial training.

organization

end

ge



Google Flu Trends heavily criticized in a paper
published by Alex's research team

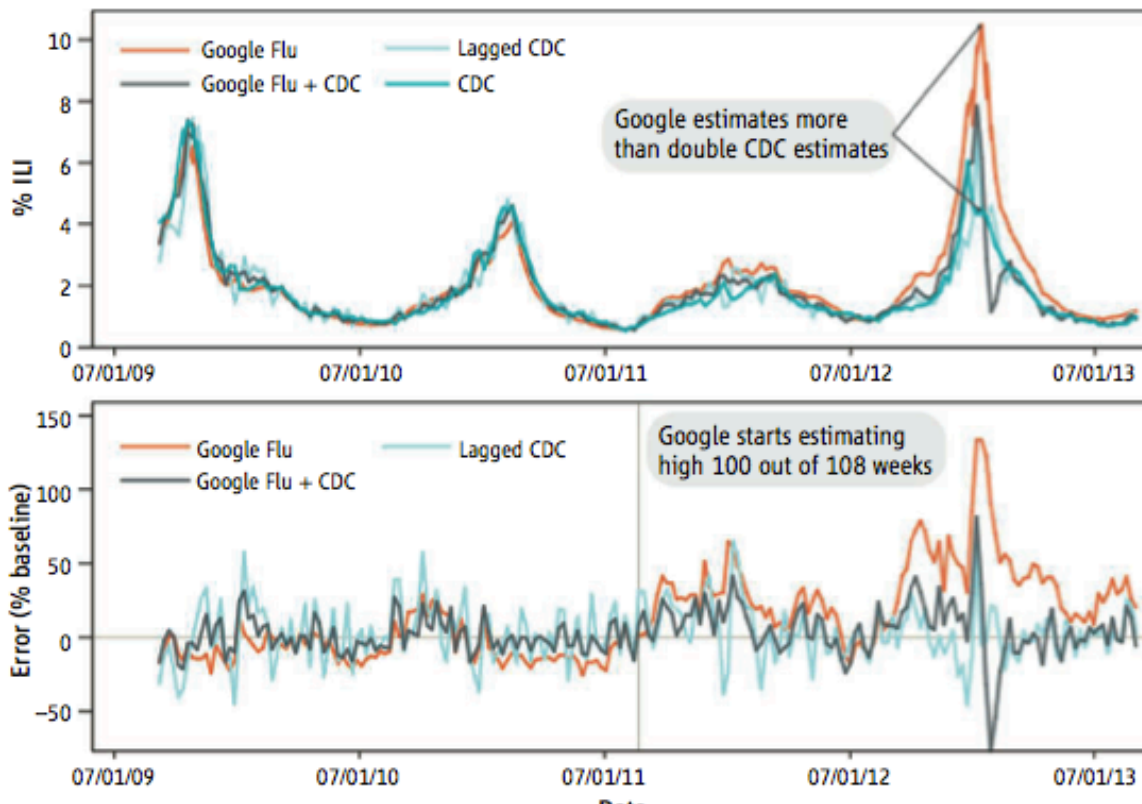
BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,2*} Ryan Kennedy,^{1,3,4} Gary King,³ Alessandro Vespignani^{5,6,3}

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014

Published by AAAS



Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

1. Lagged (CDC-based) models capable of outperforming GFT.
2. GFT + lagged CDC can outperform GFT (recalibrating importance of GFT)
3. Google search engine itself changed 86 times in June and July 2012 potentially leading to changes in Google search results (independent variable)
4. Feedbacks (recommended search terms depend on previous searches)

We recently established a new standard by
Incorporating historical information (via autoregressive terms)



Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang^a, Mauricio Santillana^{b,c,1}, and S. C. Kou^{a,1}

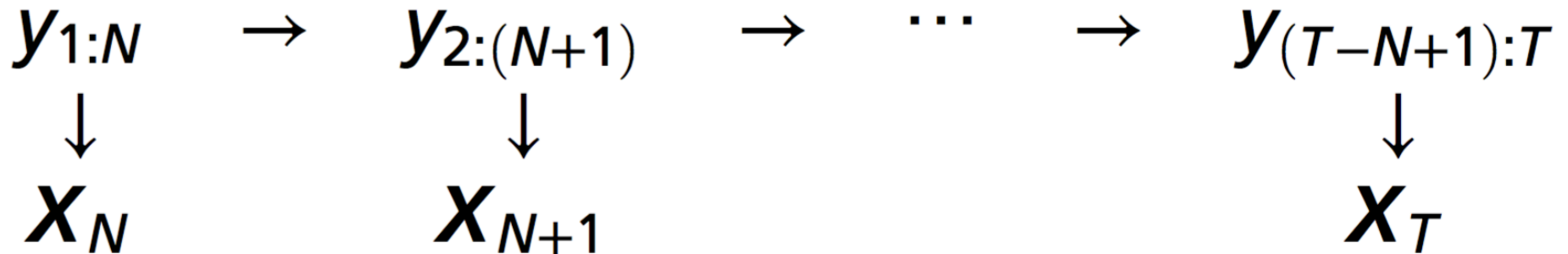
^aDepartment of Statistics, Harvard University, Cambridge, MA 02138; ^bSchool of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and ^cComputational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with Google search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search-based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people's online search behavior over time. ARGO is also flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians' queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26), Flutracking (Australia) (27), and Flu Near You (United States) (28). Among them, GFT has received the most attention and has inspired subsequent digital disease detection systems (3, 8,

We assume there is a Hidden Markov model

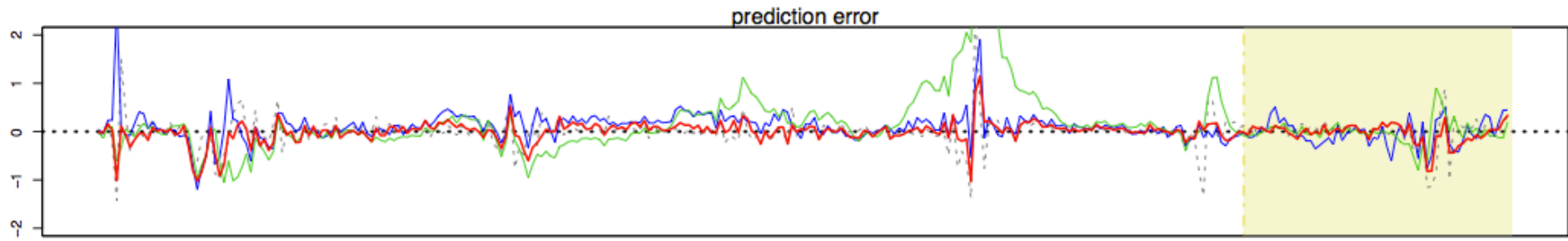
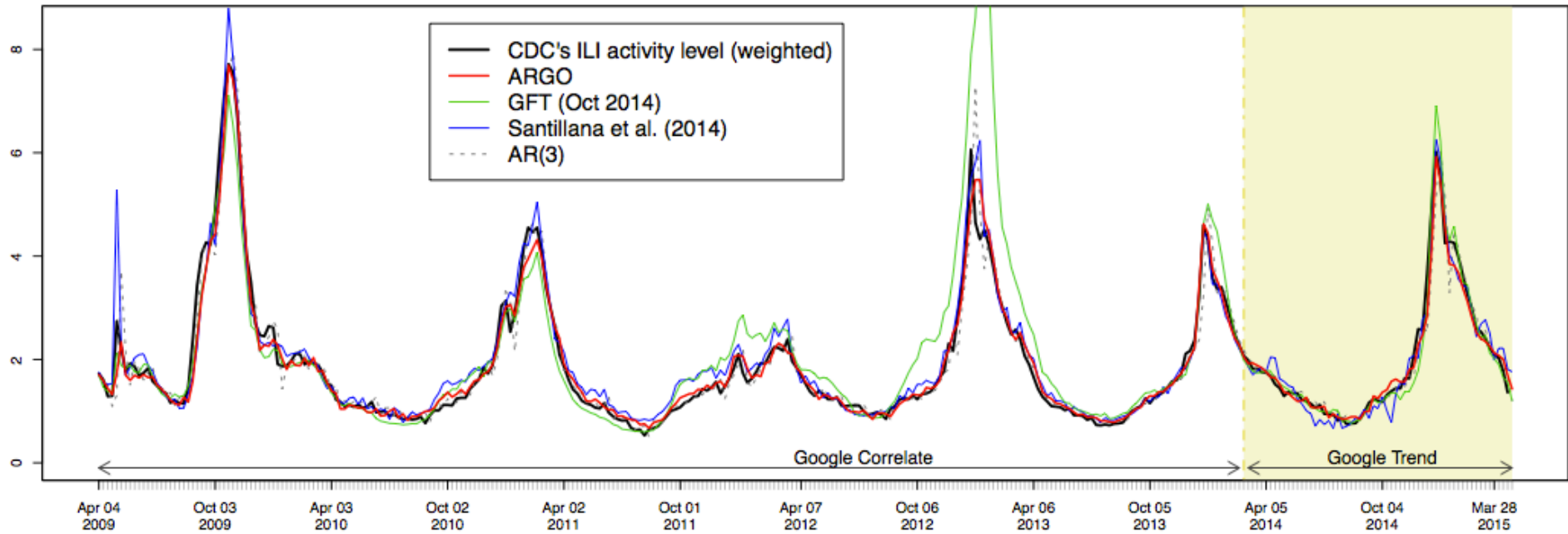


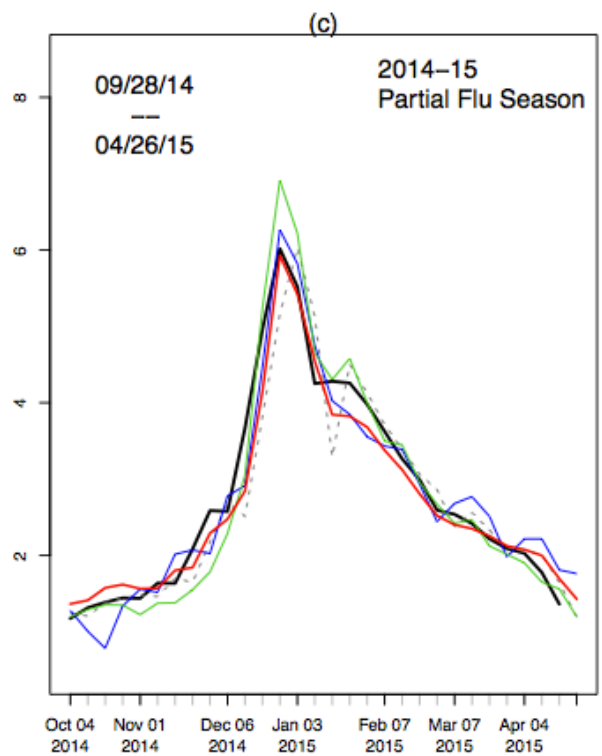
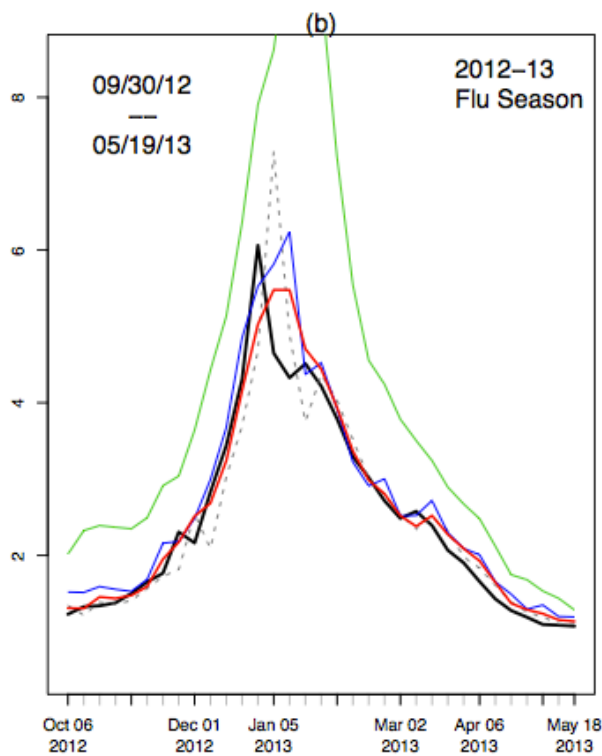
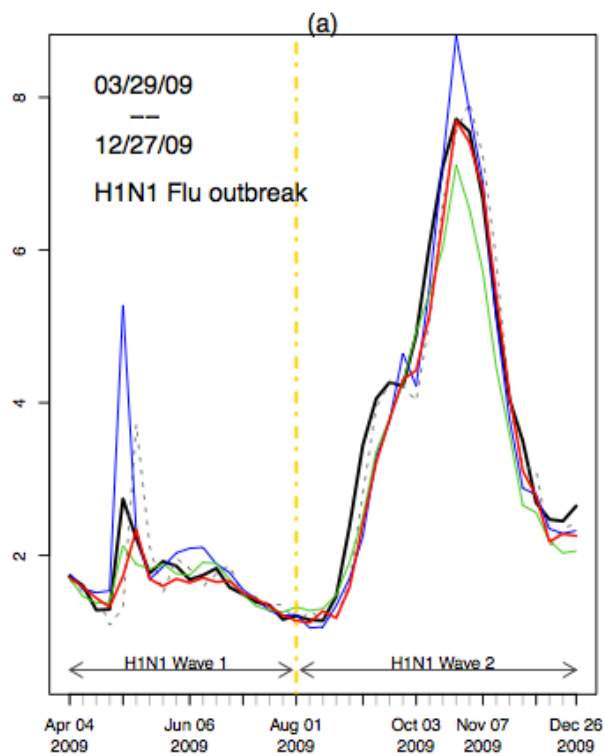
Our formal mathematical assumptions are

(assumption 1) $y_t = \mu_y + \sum_{j=1}^N \alpha_j y_{t-j} + \epsilon_t, \epsilon_t \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2)$

(assumption 2) $\mathbf{X}_t | y_t \sim \mathcal{N}_K(\mu_x + y_t \boldsymbol{\beta}, \mathbf{Q})$

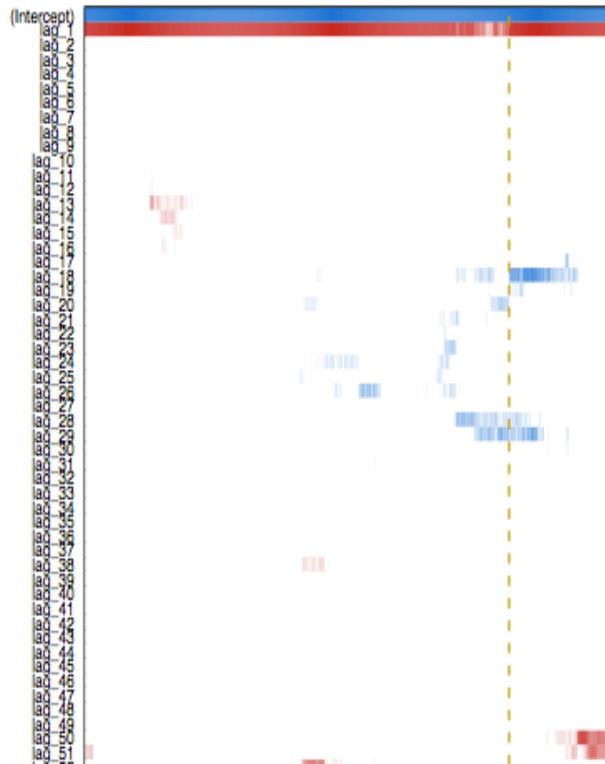
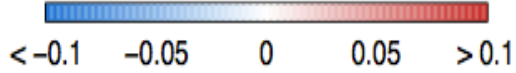
(assumption 3) conditional on y_t , \mathbf{X}_t is independent of $\{y_l, \mathbf{X}_l : l \neq t\}$





	Whole period	Off-season flu	Regular flu seasons (week 40 to week 20 next year)				
		H1N1	2010-11	2011-12	2012-13	2013-14	2014-15 partial
RMSE							
ARGO	0.637	0.655	0.618	0.830	0.679	0.308	0.593
GFT (Oct 2014)	2.213	0.773	1.110	3.023	4.451	0.981	0.683
Santillana et al. (2014)	0.909	0.945	0.864	1.688	0.918	0.495	0.683
AR(3)	0.955	0.813	0.794	1.051	1.191	0.966	0.924
Naive	1.000 (0.354)	1.000 (0.600)	1.000 (0.339)	1.000 (0.163)	1.000 (0.499)	1.000 (0.350)	1.000 (0.500)
MAE							
ARGO	0.680	0.607	0.588	0.760	0.653	0.406	0.673
GFT (Oct 2014)	1.828	0.777	1.260	3.277	5.028	0.884	0.726
Santillana et al. (2014)	1.035	0.793	0.977	1.782	0.897	0.634	0.872
AR(3)	0.920	0.777	0.787	0.951	0.988	0.915	0.924
Naive	1.000 (0.206)	1.000 (0.425)	1.000 (0.259)	1.000 (0.135)	1.000 (0.325)	1.000 (0.213)	1.000 (0.332)
Correlation							
ARGO	0.984	0.984	0.988	0.924	0.968	0.993	0.981
GFT (Oct 2014)	0.874	0.989	0.968	0.833	0.926	0.969	0.984
Santillana et al. (2014)	0.970	0.959	0.982	0.898	0.960	0.982	0.967
AR(3)	0.963	0.968	0.971	0.877	0.903	0.928	0.939
Naive	0.960	0.951	0.954	0.887	0.924	0.923	0.929
Corr. of increment							
ARGO	0.744	0.796	0.793	0.309	0.532	0.944	0.851
GFT (Oct 2014)	0.706	0.863	0.702	0.484	0.502	0.849	0.910
Santillana et al. (2014)	0.671	0.782	0.688	0.599	0.375	0.882	0.738
AR(3)	0.386	0.585	0.569	0.077	0.011	0.414	0.498
Naive	0.438	0.602	0.570	0.095	0.134	0.415	0.518

Negative coefficient Positive coefficient



2010 2012 2014



2010 2012 2014

New flu tracker uses Google search data better than Google

Unlike defunct Flu Trends, the model is self-correcting and close to reality.

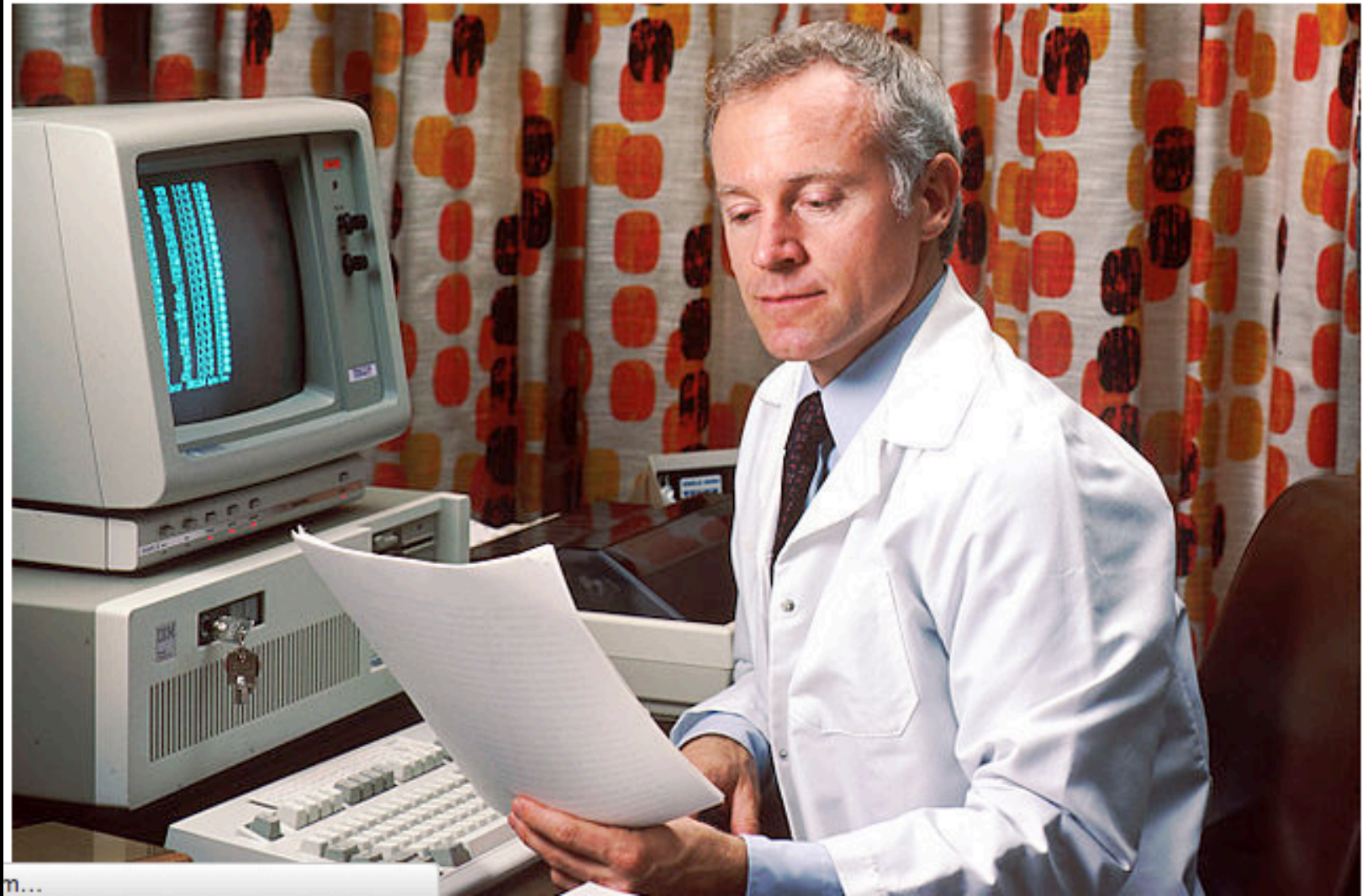
by Beth Mole - Nov 9, 2015 3:35pm EST

[Share](#)

[Tweet](#)

[Email](#)

25



m...

New flu tracker uses Google search data better than Google

MOTS-CLÉS

La revanche du big data : Harvard plus forte que Google pour prédire la grippe

Des chercheurs de la prestigieuse université américaine ont conçu un modèle statistique deux fois plus efficace que la méthode Google. Le géant de l'Internet avait fermé cet été son projet, dont les prédictions avaient tourné au flop.

Par Delphine Cuny Rédactrice en chef adjointe. Publié le 11/11/2015 à 07h09

7 397 VISITES

7 RÉACTIONS

• 1



New Google, estadística y 'big data' para bett cazar brotes de gripe



G+ 20

Me gusta 154

Tweet

Un nuevo modelo que combina información epidemiológica y búsquedas de Google es capaz de predecir los brotes de gripe una o dos semanas antes que los métodos clínicos tradicionales. El modelo podrá servir para mejorar la toma de decisiones, como la distribución de personal y recursos hospitalarios en regiones que más lo necesiten.

Más información sobre: [gripe](#) [brote](#) [epidemia](#) [Google](#) [estadística](#) [big data](#)

SINC | [Seguir a @agencia_sinc](#) | 09 noviembre 2015 21:00



Par Delphine Cuny Rédactric



El modelo es capaz de producir estimaciones más precisas sobre brotes de gripe que cualquier otro método disponible, según los autores. / Sebastian Smit

New Google, estadística y 'big data' para better cazar brotes de gripe



G+1 20

Me gusta 154

Tweet

Un nuevo modelo que combina información epidemiológica y búsquedas de Google es capaz de predecir los brotes de gripe una o dos semanas antes que los métodos clínicos tradicionales. El modelo podrá servir para mejorar la toma de decisiones, como la distribución de personal y recursos hospitalarios en regiones que más lo necesiten.

Más información sobre: [gripe](#) [brote](#) [epidemia](#) [Google](#) [estadística](#) [big data](#)

HARVARD gazette

SCIENCE & HEALTH > HEALTH & MEDICINE

On top of the flu

Chance for advance warning in search-based tracking method



Let's work on writing our own version of **ARGO**

Step-by-step

1. Download Google searches input file from the website "Google correlate"
2. Download the CDC data (gold standard) from course website
3. Repeat dengue exercise, this time make it multi-variables (build a static and dynamic version: Santillana et al, 2014, AJPM)
4. Add historical information in the form of autoregressive terms
5. Let's make sure you succeed on Wed morning.
6. If you need assistance, please download the ARGO package from course website

Using Google searches to track diseases dynamically

```
begin
%% Load data %%
CDC=load(CDC ILI Data)          (ONE COLUMN OF VALUES)
X=load(Google search Data)     (MULTIPLE COLUMNS OF VALUES)

%% initialize output arrays %%
Y=zeros(1:end.of.predictions)  (INITIALIZE ARRAY TO STORE PREDICTIONS)
coefficients=zeros(1:end.of.predictions) (INITIALIZE ARRAY TO STORE COEFFS)

%% train models and produce out-of-sample predictions %%
for i = training : end.of.predictions
    CDC ← standardize(CDC)      (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)
    X ← standardize(X)          (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)
    model=LASSOroutine.fit(CDC[1 : i] ~ X[1 : i]) (TRAINING: IN-SAMPLE MODEL )
    coefficients(i) ← model(coefficients)
    Y(i + 1)=LASSOroutine.predict(model, X(i + 1)) (PRODUCE OUT-OF-SAMPLE
                                                    PREDICTIONS)

    if(i == training)
        Y[1:i]=LASSOroutine.predict(model, X[1:i]) IN -SAMPLE PREDICTIONS
    end
end
end
end
```

And on Aug 20th, 2015

Google discontinues Flu Trends indefinitely!



Google Research Blog

The latest news from Research at Google

The Next Chapter for Flu Trends

Posted: Thursday, August 20, 2015



Instead of maintaining our own website going forward, we're now going to empower institutions who specialize in infectious disease research to use the data to build their own models. Starting this season, we'll provide Flu and Dengue signal data directly to partners including [Columbia University's Mailman School of Public Health](#) (to update their [dashboard](#)), [Boston Children's Hospital/Harvard](#), and [Centers for Disease Control and Prevention \(CDC\) Influenza Division](#). We will also continue to make historical Flu and Dengue estimate data available for anyone to see and analyze.

NEWS

Google Flu Trends calls out sick, indefinitely

Google will pass along search queries related to the flu to health organizations so they can develop their own prediction models

By [Fred O'Connor](#) | [Follow](#)

IDG News Service | Aug 20, 2015 2:07 PM PT

MORE LIKE THIS ::

[Google Begins Tracking Swine Flu in Mexico](#)



[Google's Panicky Flu Estimates Were Dead Wrong](#)

BIG DATA

Google discontinues Flu Trends, starts offering data to researchers

[JORDAN NOVET](#) | AUGUST 20, 2015 12:17 PM

TAGS: [GOOGLE](#), [GOOGLE FLU TRENDS](#)

Our team at Boston Children's Hospital now has access to Google's search volumes, as one of the exclusive Google's partners.

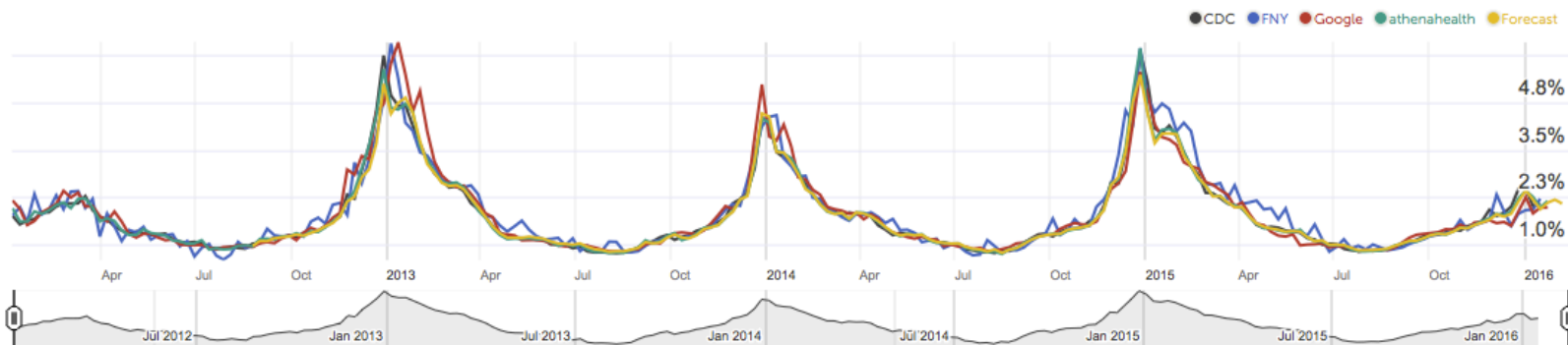
We are creating a new improved disease forecasting platform

United States ▾

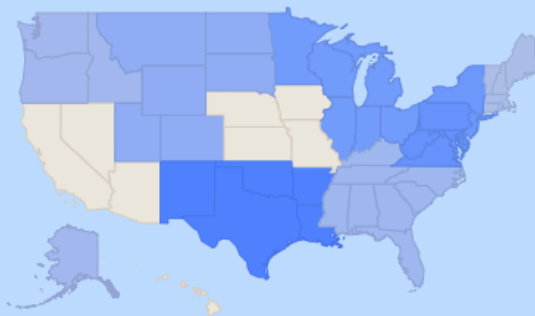
Time range ▾

Flu ▾

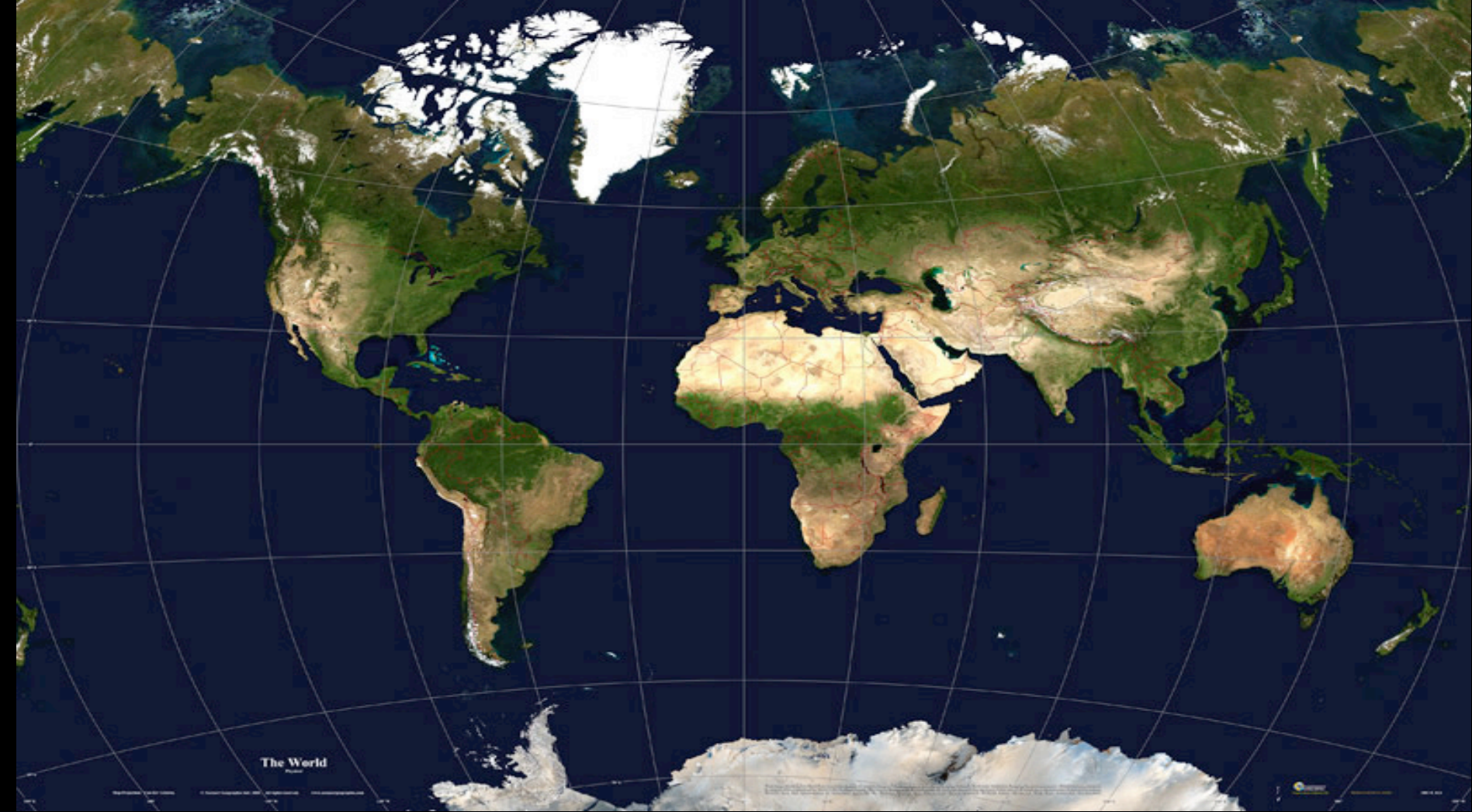
Forecast 2.2% this week 2.1% next week



CDC FNY Google athenahealth Forecast



Thanks to Sue Aman, Rachel Chorney, Jeff Andre, Andre Nguyen, John Brownstein and Healthmap team!



Thank you!

Contact: msantill@fas.harvard.edu