a pattern for digital disease detection

- ground truth data from official source
- proxy data from digital source
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
- validate model performance

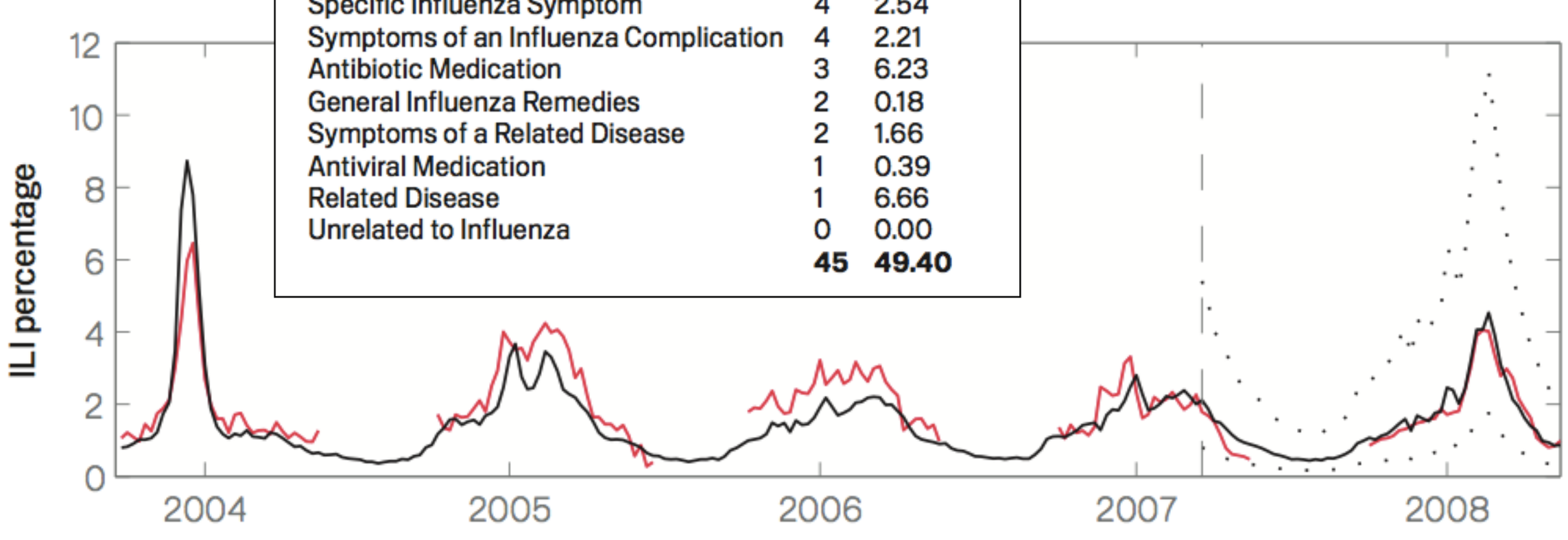
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

Search Query Topic	Top 45 Queries N	Weighted
Influenza Complication	11	18.15
Cold/Flu Remedy	8	5.05
General Influenza Symptoms	5	2.60
Term for Influenza	4	3.74
Specific Influenza Symptom	4	2.54
Symptoms of an Influenza Complication	4	2.21
Antibiotic Medication	3	6.23
General Influenza Remedies	2	0.18
Symptoms of a Related Disease	2	1.66
Antiviral Medication	1	0.39
Related Disease	1	6.66
Unrelated to Influenza	0	0.00
	45	**49.40**

google.org/flutrends



J. Ginsberg *et al.*, Nature 457, 1012 (2009)

http://www.google.org/flutrends

google.org Flu Trends

Language: English (United States)

Google.org home

Dengue Trends

Flu Trends

Home

Select country/region

How does this work?

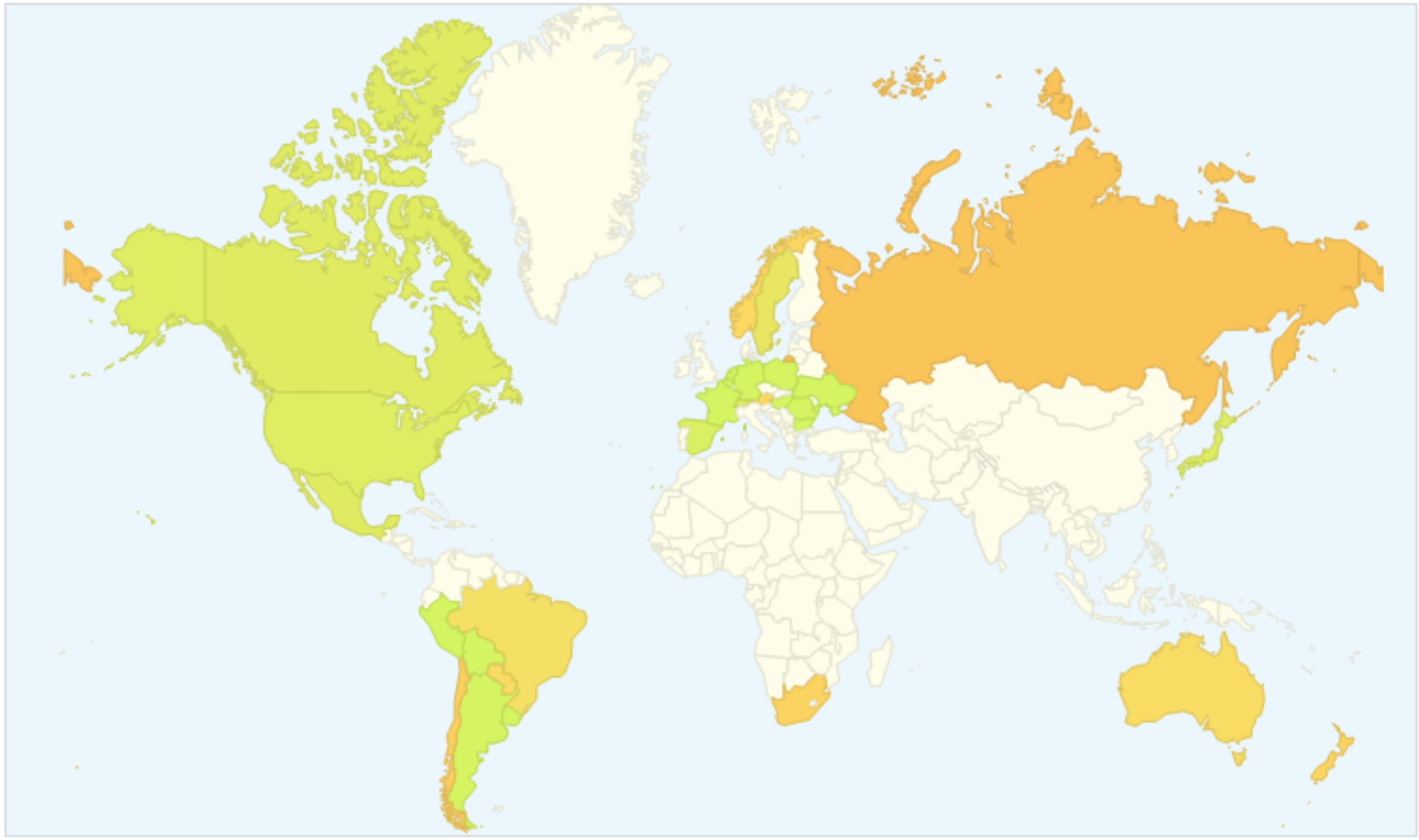
FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

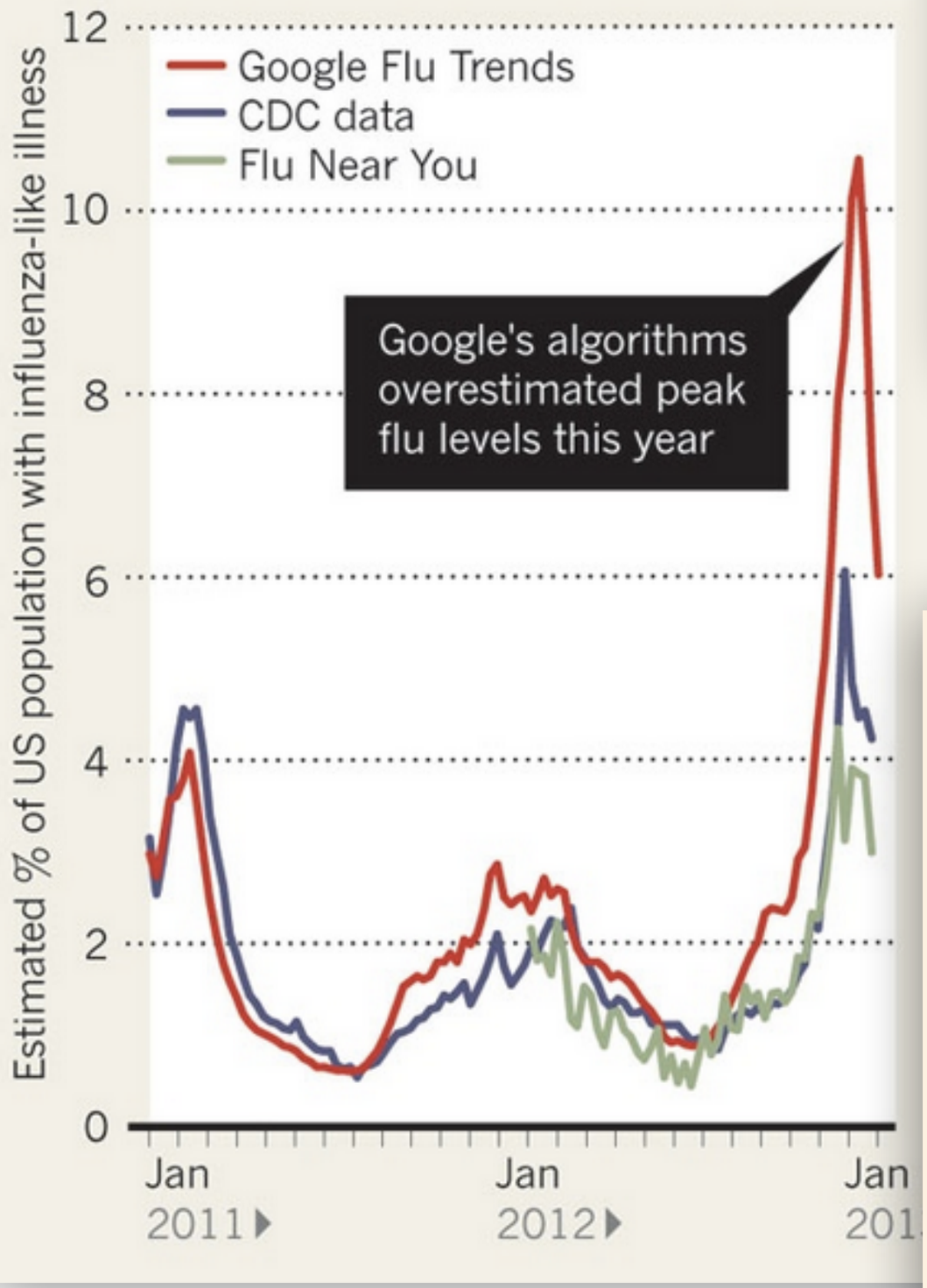
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »



Download world flu activity data - Animated flu trends for Google Earth - Compare flu trends across regions in Public Data Explorer

FEVER PEAKS

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



nature
International weekly journal of science

NATURE | NEWS

عربي

When Google got flu wrong

Science 14 March 2014:
Vol. 343 no. 6176 pp. 1203–1205
DOI: 10.1126/science.1248506

POLICY FORUM

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]

FT Magazine

March 28, 2014 11:38 am

Big data: are we making a big mistake?

By Tim Harford

Big data is a vague term for a massive phenomenon that ha rapidly become an obsession with entrepreneurs, scientists governments and the media

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Monitoring Influenza Epidemics in China with Search Query from Baidu

Qingyu Yuan, Elaine O. Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, John S. Brownstein

Published: May 30, 2013 • http://dx.doi.org/10.1371/journal.pone.0064323

Abstract

Several approaches have been proposed for near real-time detection and prediction of the spread of influenza. These include search query data for influenza-related terms, which has been explored as a tool for augmenting traditional surveillance methods. In this paper, we present a method that uses Internet search query data from Baidu to model and monitor influenza activity in China. The objectives of the study are to present a comprehensive technique for: (i) keyword selection, (ii) keyword filtering, (iii) index composition and (iv) modeling and detection of influenza activity in China. Sequential time-series for the selected composite keyword index is significantly correlated with Chinese influenza case data. In addition, one-month ahead prediction of influenza cases for the first eight months of 2012 has a mean absolute percent error less than 11%. To our knowledge, this is the first study on the use of search query data from Baidu in conjunction with this approach for estimation of influenza activity in China.

Subject Areas

Influenza
Swine influenza
Influenza viruses
H1N1
China
Influenza A virus
Infectious disease s…

influenza & search engine queries

- ground truth data from official source
  - routine flu surveillance from China's Ministry of Health
- proxy data from digital source
  - search query logs from Baidu search engine
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
  - search index composition & multiple regression
- validate model performance
  - validation set

why influenza ?

Seasonal influenza epidemics result in an estimated 3 to 5 million cases of severe illness and 250,000 to 500,000 deaths worldwide each year

http://www.who.int/mediacentre/factsheets/fs211/en/

methodology

- search for keywords or terms which might be related to influenza
- process keywords by eliminating those unrelated to influenza epidemics
- define weights and composite search index
- fit regression model using selected keyword index to influenza case data

official data from China's Ministry of Health

Month	ICD*	Month	ICD	Month	ICD	Month	ICD	Month	ICD
2009–03	8015	2009–12	29977	2010–09	5114	2011–06	3065	2012–03	21625
2009–04	6794	2010–01	10415	2010–10	4121	2011–07	2654	2012–04	10707
2009–05	7769	2010–02	6595	2010–11	5323	2011–08	3243	2012–05	8520
2009–06	7999	2010–03	8488	2010–12	6529	2011–09	4360	2012–06	6195
2009–07	7791	2010–04	6357	2011–01	6072	2011–10	5525	2012–07	6738
2009–08	14548	2010–05	3865	2011–02	5930	2011–11	7055	2012–08	6793
2009–09	43596	2010–06	2642	2011–03	7299	2011–12	11631		
2009–10	25132	2010–07	2627	2011–04	5727	2012–01	10046		
2009–11	43018	2010–08	3588	2011–05	4130	2012–02	17421		

*ICD is the abbreviation for influenza case data.
doi:10.1371/journal.pone.0064323.t001

- network of physicians report laboratory confirmed cases to the MOH on a daily basis
- released 1–2 week after the end of each month
- data is solely laboratory confirmed influenza cases and does not include ILI cases.

query log data

http://index.baidu.com/

- daily counts of search queries are publicly available
- monthly aggregation to match time scale of official data

find related keywords

流感 (“flu”)

流感(flu)	***新流感(new flu)***	h1n1流感(h1n1 flu)	**本山快乐营猪流感 (Benshan Happy camp swine flu)**	***流感吃什么药(influenza drugs)***	甲型流感疫苗(type a influenza vaccine)	甲流感预防知识*(the knowledge of type a influenza)*
流感疫苗(influenza vaccine)	预防流感知识(knowledge of influenza prevention)	**上流感{up influenza(song)}**	季节性流感疫苗 (seasonal influenza vaccine)	**qq流感大盗下载(qq flu game download)**	***甲型流感症状(type a flu symptom)***	甲型h1n1流感的症状(the symptoms of type a h1n1 flu)
甲型*h1n1*流感*(type a h1n1 flu)*	**关颖 上流感{Guan Ying up influenza(song)}**	甲型流感(type a flu)	**流感概念股(flu concept stock)**	流感疫苗价格(the price of influenza vaccine)	如何预防猪流感(how to prevent swine flu)	*甲型h1n1流感防治(type h1n1 influenza prevention and control)*
山东流感*(Shandong flu)*	***a型流感病毒(type a influenza virus)***	*西班牙流感(Spanish flu)*	***香港流感 (Hongkong flu)***	如何预防流感(how to prevent flu)	*h1n1流感的症状(h1n1 flu symptom)*	***甲型h1n1流感治疗(type a h1n1 flu therapy)***
流感的预防措施 (prevention measures of influenza)	甲型h1n1流感防控(the prevention of h1n1 flu)	甲流感(h1n1 influenza)	**羊流感(goat flu)**	季节性流感(seasonal flu)	***h3n2流感病毒(h3n2 flu virus)***	*甲型h1n1流感资料(type a h1n1 flu information)*
流感的预防 (prevention of flu)	甲型h1n1流感预防 (prevention of h1n1 flu)	流感预防措施(the prevention measures of influenza)	***预防猪流感(prevent swine flu)***	***北京流感(Beijing influenza)***	***甲型h1n1流感知识(the knowledge of type a h1n1 influenza)***	甲型流感的预防(prevention of type a flu)
流感症状(influenza symptom)	甲型h1n1流感症状(influenza symptom of h1n1)	**情流感(love flu)**	预防甲型h1n1流感 (prevention of type a h1n1 flu)	*甲流感预防(prevention of type a influenza)*	**甲型h1n1流感作文 (composition of type a h1n1 flu)**	*甲型流感预防知识(prevention knowledge of type a influenza)*
流感病毒(influenza virus)	流感疫情(Flu epidemic)	预防流感(prevent the flu)	*h1n1流感症状(h1n1 influenza symptom)*	流感传播途径 (transmission way of flu)	***甲型h3流感(type a h3 flu)***	*流感疫苗不良反应(Influenza vaccine adverse reaction)*
qq流感大盗 (qq influenza game)	甲型h1n1流感疫苗(influenza vaccine of h1n1)	流感的传播途径(the transmission way of flu)	***甲流感的症状(the symptom of h1n1 flu)***	流感大流行(influenza pandemic)	***甲型流感 症状(type a flu symptoms)***	*流感疫苗接种(influenza vaccinations)*
流感的症状(the influenza symptom)	*甲型h1n1流感(type a h1n1 flu)*	流感疫苗副作用(Influenza vaccine side effects)	甲流感症状(symptom of h1n1 flu)	流感预防(prevent influenza)	预防甲型流感(prevention of type a flu)	*流感最新疫情(Latest outbreak of influenza)*
流感防治知识(the knowledge of influenza prevention)	***流感疫苗接种时间(Influenza vaccination time)***	**h1n1流感手抄报(h1n1 flu Shouchao Bao)**	甲型h1n1流感疫情 (epidemic situation of type a h1n1 flu)	流感治疗(flu treatment)	*防控甲型h1n1流感 (prevention and control type a h1n1 flu)*	人感染猪流感症状(Human infection with swine flu symptoms)
情流感菌(love flu virus)	*北京 流感(Beijing flu)*	**上流感 关颖(up influenza(song) Guan Ying)**	*甲型h3n2流感病毒(type a h3n2 flu virus)*	***新型流感(new influenza)***	**狗流感(dog influenza)**	*如何预防甲型流感(How to prevent influenza a)*
山东猪流感(Shandong swine flu)	**预防甲型流感手抄报 (Shouchao Bao of prevention type a flu)**	***副流感病毒(Para-influenza)***	***甲型流感的症状(the symptom of h1n1 flu)***	a型流感(type a influenza)	甲型h1n1流感病毒(type a h1n1 flu virus)	***流感疫苗有必要打吗(The flu vaccine is necessary to play)***
h1n1流感预防(h1n1 influenza prevention)	*h1n1流感预防知识(h1n1 knowledge of influenza prevention)*	H1n1流感(h1n1 flu)				

filter related keywords

- should represent factors that might influence the influenza epidemic
- search query data for keyword should be a sequential time series with a daily, weekly or monthly resolution
- low bound on maximum cross-correlation coefficient with influenza case data:

$$\gamma_{XY}(\tau) = \mathrm{E}[(X_t - \mu_X)(Y_{t+\tau} - \mu_Y)]$$

define a query-based index

$$y^{(j)} = \alpha_0 + \alpha_1 \sum_{i=1}^{j} \omega_i x_i^*$$

- need to select number of *j* of keywords we use
- order keywords by maximal correlation with flu time series
- this is a “nested” model
- for a given choice of *j*, do regression & compute partial F-test statistic
- add keywords one by one until F-test shows that there is no improvement

partial F-test:

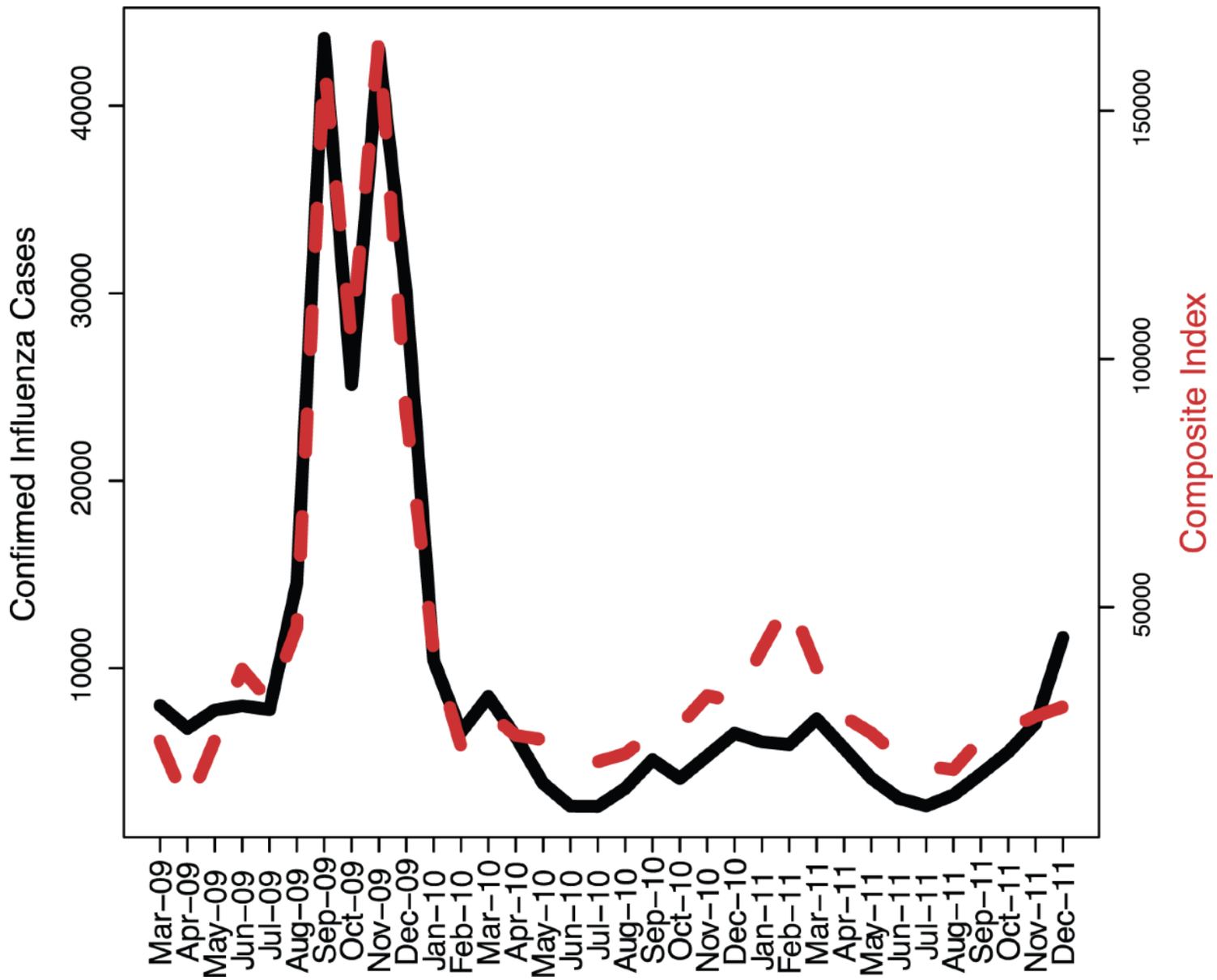
Full model SSE:	SSE_{Full}	$\leftarrow df_{\text{Full}} = n-k-1$
Reduced model SSE:	SSE_{Reduced}	$\leftarrow df_{\text{Reduced}} = n-k-1+m$
Extra SSE:	$SSE_{\text{Reduced}} - SSE_{\text{Full}}$	$\leftarrow df = (n-k-1+m) - (n-k-1) = m$

The **partial F test statistic** is the ratio of two variances. The numerator is the difference in error sums of squares (the "extra sum of squares") between the two models, divided by the number of predictors eliminated. The denominator is the mean squared error for the full model (SSE_{Full}) divided by its degrees of freedom.

$$F_{\text{calc}} = \frac{\left(\frac{SSE_{\text{Reduced}} - SSE_{\text{Full}}}{m}\right)}{\left(\frac{SSE_{\text{Full}}}{n-k-1}\right)} \leftarrow \text{if } m \text{ predictors are eliminated}$$

query-based index: result

- trained only on part of the data (2012- left aside for validation)
- stepwise regression keeps only 8 keywords out of 40 selected from the original 94 keywords)
- cross-correlation coefficient = 0.96 at lag 0



time series model

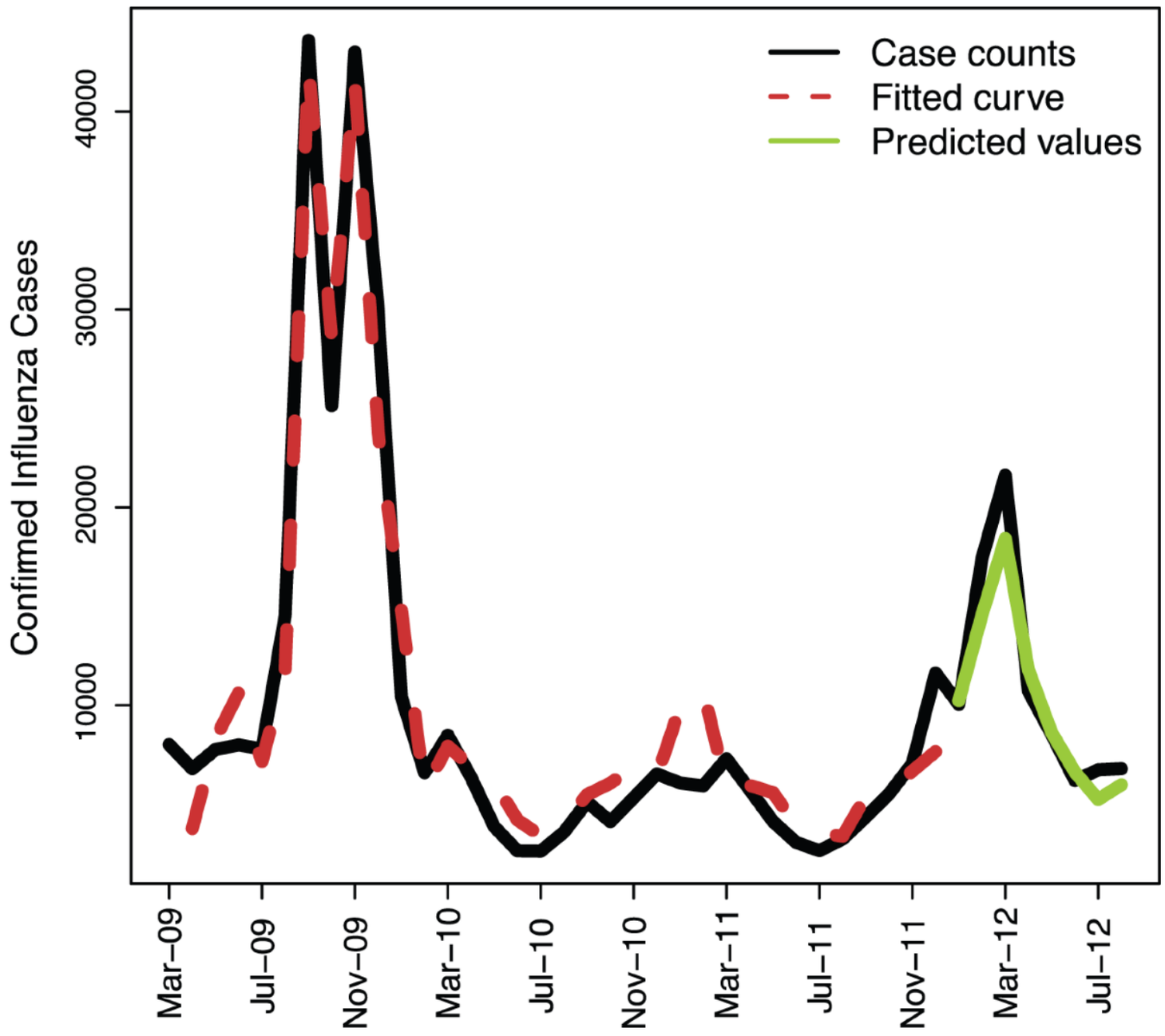
- most significant correlations between the query-based index *y* and the case data are observed at lag 0 and lag 1
- we fit the following autoregressive model:

$$I(t) = \beta_0\, I(t-1) + \beta_1\, y(t) + \beta_2\, y(t-1) + \epsilon$$

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
index[*t*]	0.253	0.015	17.455	<0.001
index[*t*−1]	−0.138	0.044	−3.159	0.0036
ICD[*t*−1]	0.555	0.157	3.534	0.0013
residual	ADF	MacKinnon threshold		
	t-Stat	1%	5%	10%
	−5.685	−3.654	−2.957	−2.617

model validation

- tested on validation set (2012-)
- mean % error of prediction for the 8 months of validation is 10.6%
- more complex models do not perform better on validation set



influenza & search engine queries

- ground truth data from official source
  - routine flu surveillance from China's Ministry of Health
- proxy data from digital source
  - search query logs from Baidu search engine
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
  - search index composition & multiple regression
- validate model performance
  - validation set

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time

David J. McIver*, John S. Brownstein

Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Circulating levels of both seasonal and pandemic influenza require constant surveillance to ensure the health and safety of the population. While up-to-date information is critical, traditional surveillance systems can have data availability lags of up to two weeks. We introduce a novel method of estimating, in near-real time, the level of influenza-like illness (ILI) in the United States (US) by monitoring the rate of particular Wikipedia article views on a daily basis. We calculated the number of times certain influenza- or health-related Wikipedia articles were accessed each day between December 2007 and August 2013 and compared these data to official ILI activity levels provided by the Centers for Disease Control and Prevention (CDC). We developed a Poisson model that accurately estimates the level of ILI activity in the American population, up to two weeks ahead of the CDC, with an absolute average difference between the two estimates of just 0.27% over 294 weeks of data. Wikipedia-derived ILI models performed well through both abnormally high media coverage events (such as during the 2009 H1N1 pandemic) as well as unusually severe influenza seasons (such as the 2012–2013 influenza season). Wikipedia usage accurately estimated the week of peak ILI activity 17% more often than Google Flu Trends data and was often more accurate in its measure of ILI intensity. With further study, this method could potentially be implemented for continuous monitoring of ILI activity in the US and to provide support for traditional influenza surveillance tools.

Citation: McIver DJ, Brownstein JS (2014) Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol 10(4): e1003581. doi:10.1371/journal.pcbi.1003581

Editor: Marcel Salathé, Pennsylvania State University, United States of America

Received December 20, 2013; **Accepted** March 11, 2014; **Published** April 17, 2014

Funding: This work was funded by the National Institutes of Health and National Library of Medicine 1R01LM010812-03. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: david.mciver@childrens.harvard.edu

influenza & Wikipedia page views

- ground truth data from official source
  - CDC ILI (Influenza-Like Illness) data
- proxy data from digital source
  - hourly page view data on Wikipedia articles
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
  - Generalized Linear Model (+ LASSO)
- validate model performance
  - validation set

CDC Influenza Surveillance Report

National and Regional Summary of Select Surveillance Components

HHS Surveillance Regions*	Data for current week			Data cumulative since October 2, 2016 (week 40)						
	Out-patient ILI†	Number of jurisdictions reporting regional or widespread activity§	% respiratory specimens positive for flu in clinical laboratories‡	A(H1N1)pdm09	A (H3)	A (Subtyping not Performed)	B Victoria lineage	B Yamagata lineage	B lineage not performed	Pediatric Deaths
				Influenza test results from public health laboratories only						
Nation	Normal	2 of 54	1.6%	25	184	8	3	7	12	0
Region 1	Normal	0 of 6	0.4%	0	14	0	0	0	0	0
Region 2	Normal	1 of 4	0.8%	0	2	0	0	0	2	0
Region 3	Normal	0 of 6	0.4%	2	9	1	0	0	0	0
Region 4	Normal	0 of 8	3.6%	1	19	0	1	2	6	0
Region 5	Normal	0 of 6	0.7%	0	23	4	0	1	1	0
Region 6	Normal	0 of 5	1.2%	3	0	0	1	3	0	0
Region 7	Normal	0 of 4	0.2%	0	5	1	1	0	0	0
Region 8	Normal	0 of 6	0.9%	16	25	0	0	0	0	0
Region 9	Normal	1 of 4	1.2%	3	39	0	0	1	1	0
Region 10	Normal	0 of 4	3.7%	0	48	2	0	0	2	0

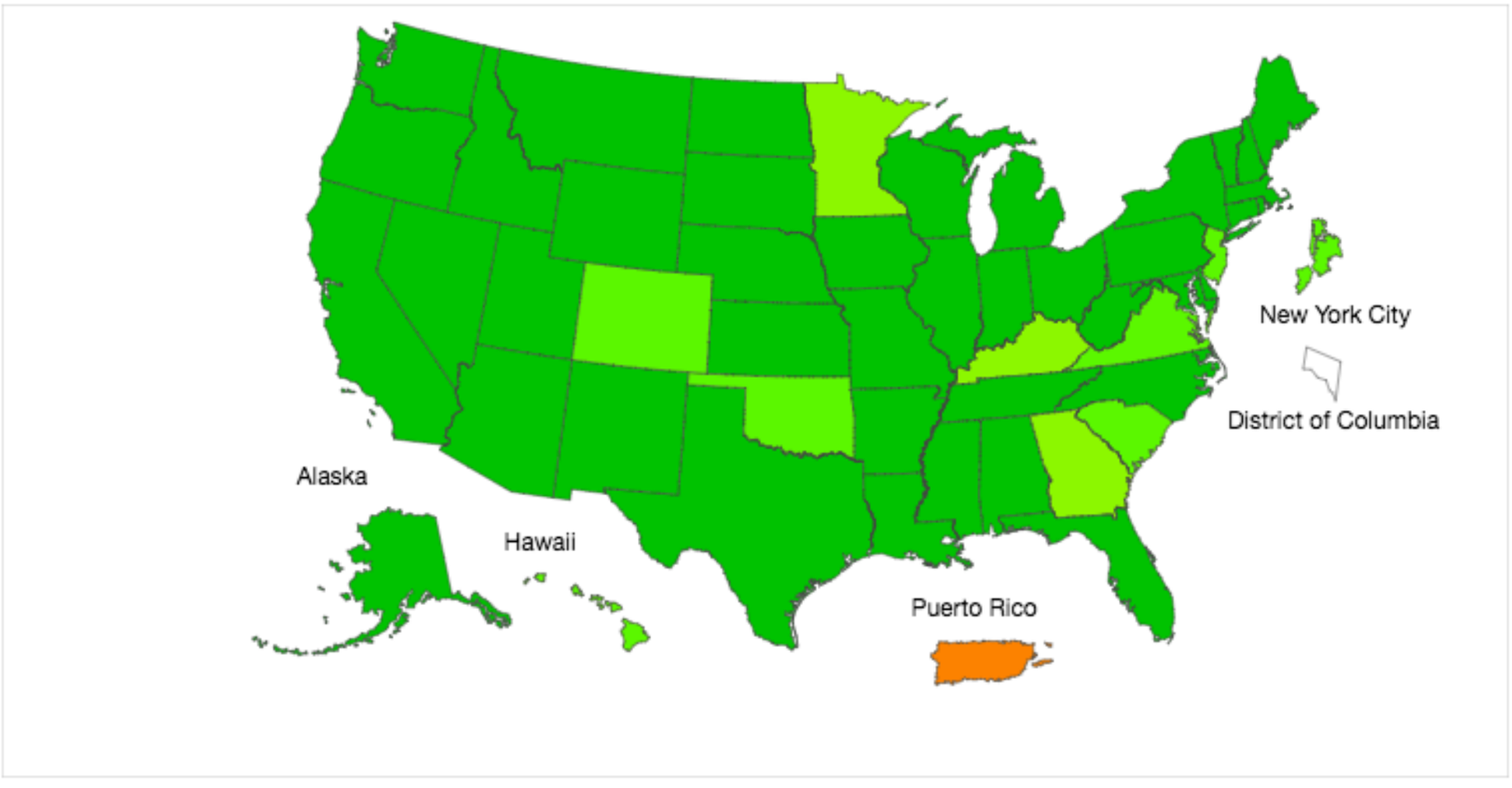
http://www.cdc.gov/flu/weekly/

CDC Influenza Surveillance Report

FLUVIEW A Weekly Influenza Surveillance Report Prepared by the Influenza Division
Influenza-Like Illness (ILI) Activity Level Indicator Determined by Data Reported to ILINet

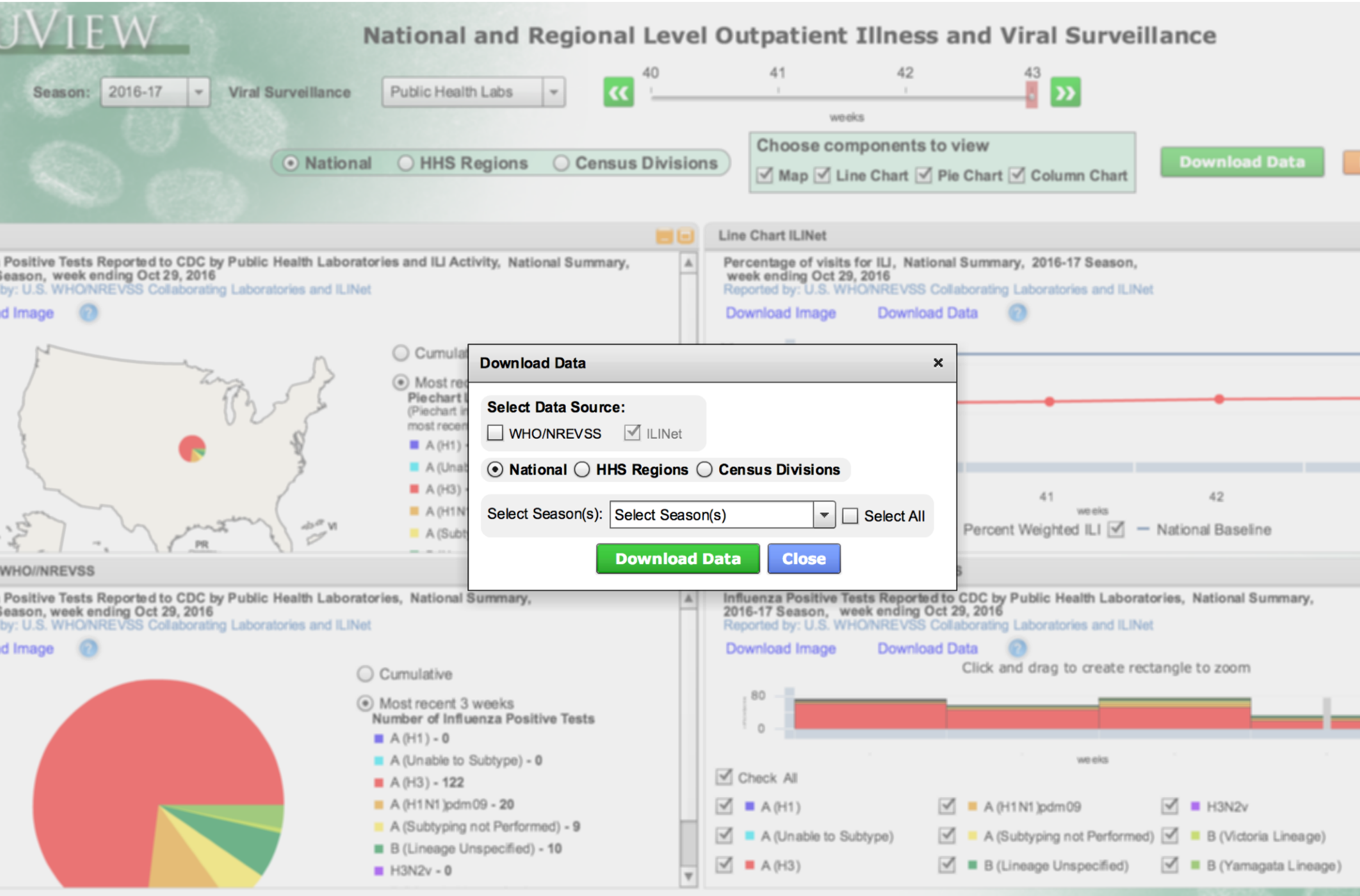
CDC

2016-17 Influenza Season Week 43 ending Oct 30, 2016



http://www.cdc.gov/flu/weekly/

http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html



Wikipedia page view data

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item

Print/export
Create a book
Download as PDF
Printable version

Wikipedia:Pageview statistics

From Wikipedia, the free encyclopedia

See also: Wikipedia:Web statistics tool, Wikipedia:Statistics § Page views, and Wikipedia article traffic

This information page describes the editing community's consensus on some aspect or aspects of Wikipedia's norms and practices. It is not one of Wikipedia's policies or guidelines.

Shortcuts:
WP:PAGEVIEW
WP:PVS

Pageview stats refers to how often a page is viewed by others.
This is **not** a measure of notability.

See: HERE for the pageview statistics tool

Page view statistics (or **Pageview stats**) is a tool available for Wikipedia pages, which allows one to see how many people have visited an article during a given time period. However, like the Search engine tests, there are limitations to it. Before using such statistics to make conclusions about an ongoing discussion, there are things that must be considered. There are software limitations (or, more exactly, conclusions that may not be taken from the provided data) and circumstances that may influence them, both from inside and outside Wikipedia. Typically, the item which ranks first in the Wikipedia Page View Statistics is Special:Export/SynchronizationStartTime:. However, the article which would actually qualify as an article in the English Wikipedia, which typically has the highest page view statistics, is the Main Page of Wikipedia.

It refers to the number of times a particular page has been requested. Using toollabs:pageviews, it is possible to see statistics on how often Wikipedia pages have been viewed during various times. These figures do not reflect the number of unique visitors a page has received.[1]

The pageview stats tool is available from any page, in two ways: 1) see the toolbox in the sidebar, which shows *page information*; the external link is in the last section of page information; and 2) look behind the page's history tab and select *page view statistics*.

Page stats can help determine how popular a page is, but are not an indication of a topic's notability. Wikipedia's inclusion guidelines are based on coverage found in reliable sources. If a page's stats are low, it is not a reason to consider it for deletion, and if high it is not a reason to save it from deletion.

https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics

https://tools.wmflabs.org/pageviews

Pageviews Analysis BETA!

Comparison of pageviews across multiple pages

Options

Dates

17/10/2016 - 05/11/2

Project

en.wikipedia.org

Platform

All

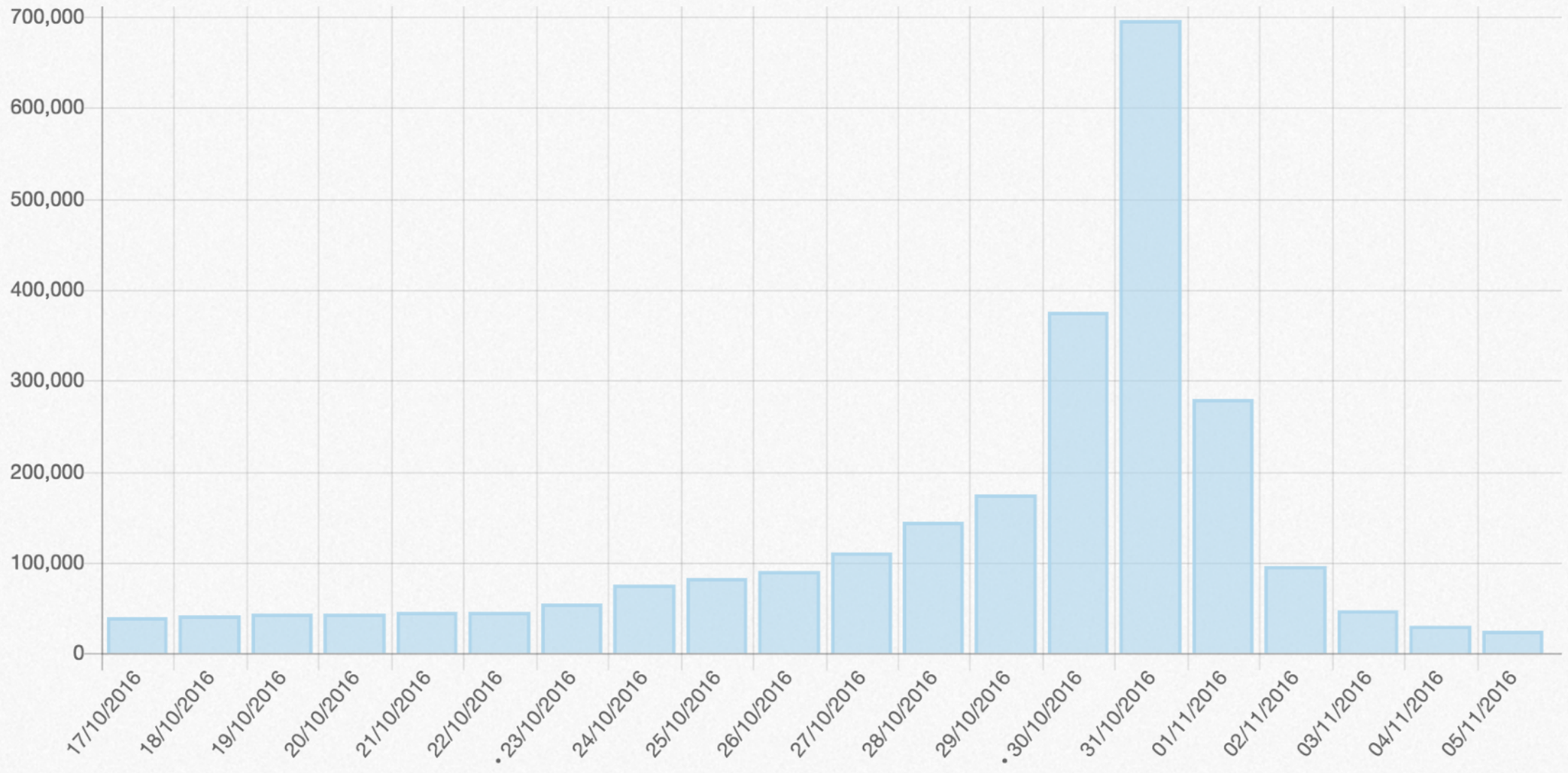
Agent

User

Pages Enter up to 10 pages

× Halloween

Chart type | Permalink | Download | Begin at zero | Logarithmic scale



Halloween · 17/10/2016 - 05/11/2016 · **2,544,840 pageviews (127,242/day)**

Wikipedia page view data

- manually selected by domain experts
- a few pages chosen to provide control for global activity (Wikimedia main oage) and European vs US activity (CDC and ECDC pages)
- weekly temporal aggregation

Avian influenza*	Influenza Virus B*
Centers for Disease Control and Prevention*	Influenza Virus C*
Common Cold*	Influenza Virus Subtype H1N1
Epidemic*	Influenza Virus Subtype H2N2*
European Centers for Disease Control and Prevention	Influenza Virus Subtype H2N9*
Fever*	Influenza Virus Subtype H3N1*
Flu Season*	Influenza Virus Subtype H3N2*
Human Influenza*	Influenza Virus Subtype H5N1*
Influenza	Influenza Virus Subtype H5N2*
Influenza-like Illness*	Oseltamivir*
Influenza Pandemic	Pandemic
Influenza Research*	Swine Influenza
Influenza Treatment*	Tamiflu*
Influenza Vaccine*	Vaccine
Influenza Virus*	Wikipedia Main Page
Influenza Virus A*	1918 Flu Pandemic*

Generalized Linear Model (GLM)

- a probability distribution from the exponential family
- a linear predictor $\eta = \boldsymbol{X\beta}$
- *a* link function *g* such that $E(Y) = \mu = g^{-1}(\eta)$

Generalized Linear Model (GLM)

- we are modeling *counts* in an interval of fixed duration
- link functions for the generalized linear model:

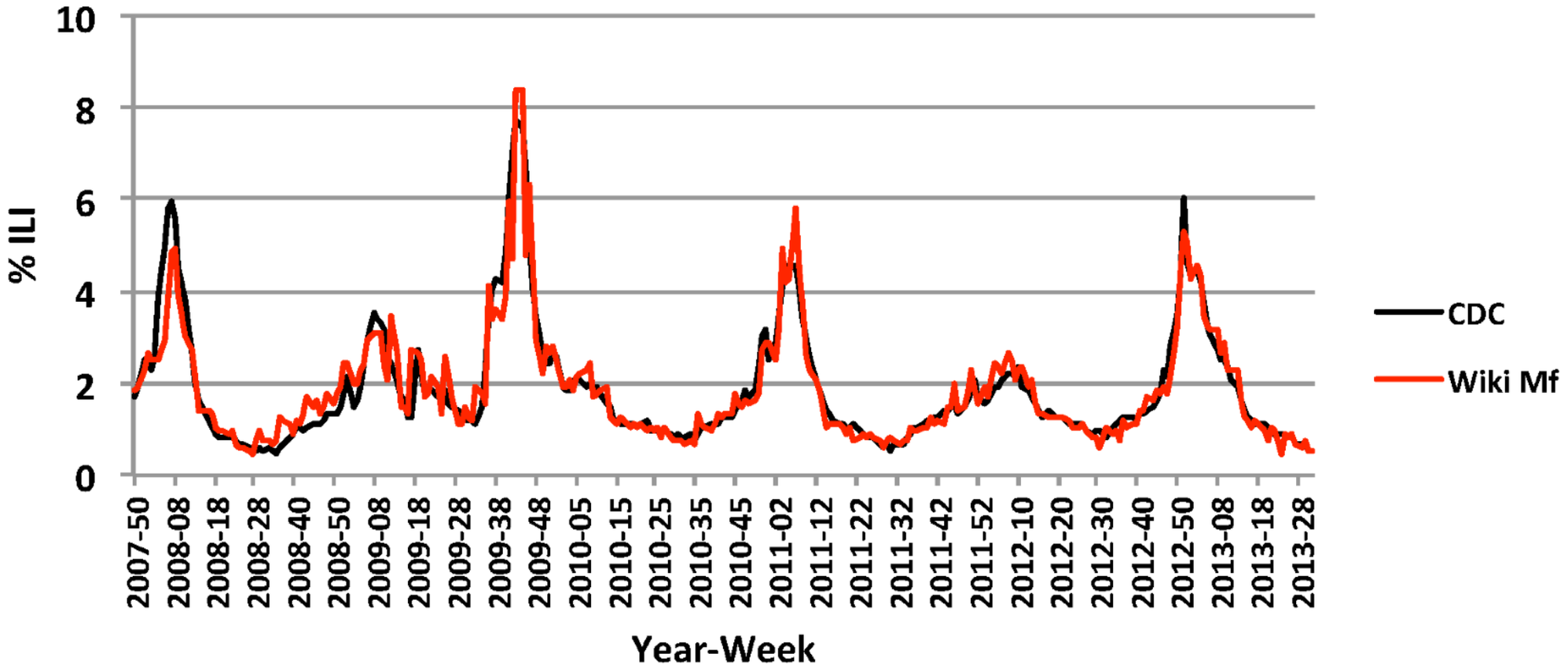
Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

Generalized Linear Model (GLM)

- a probability distribution from the exponential family
- a linear predictor $\eta = \boldsymbol{X\beta}$
- *a* link function *g* such that $E(Y) = \mu = g^{-1}(\eta)$
- $\boldsymbol{\beta}$ includes weekly view counts on selected Wikipedia pages (32 pages) + calendar year + month
- fitted by regression (or regression + LASSO)
- model selection through cross-validation

result & validation



Pearson correlation coefficient between CDC ILI values and model 0.946 ($p<0.001$)

mean absolute difference 0.27%

influenza & Wikipedia page views

- ground truth data from official source
  - CDC ILI (Influenza-Like Illness) data
- proxy data from digital source
  - hourly page view data on Wikipedia articles
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
  - Poisson Generalized Linear Model (+ LASSO)
- validate model performance
  - validation set

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic

Alessio Signorini, Alberto Maria Segre, Philip M. Polgreen

Published: May 4, 2011 • http://dx.doi.org/10.1371/journal.pone.0019467

Abstract

Twitter is a free social networking and micro-blogging service that enables its millions of users to send and read each other's “tweets,” or short, 140-character messages. The service has more than 190 million registered users and processes about 55 million tweets per day. Useful information about news and geopolitical events lies embedded in the Twitter stream, which embodies, in the aggregate, Twitter users' perspectives and reactions to current events. By virtue of sheer volume, content embedded in the Twitter stream may be useful for tracking or even forecasting behavior if it can be extracted in an efficient manner. In this study, we examine the use of information embedded in the Twitter stream to (1) track rapidly-evolving public sentiment with respect to H1N1 or swine flu, and (2) track and measure actual disease activity. We also show that Twitter can be used as a measure of public interest or concern about health-related events. Our results show that estimates of influenza-like illness derived from Twitter chatter accurately track reported disease levels.

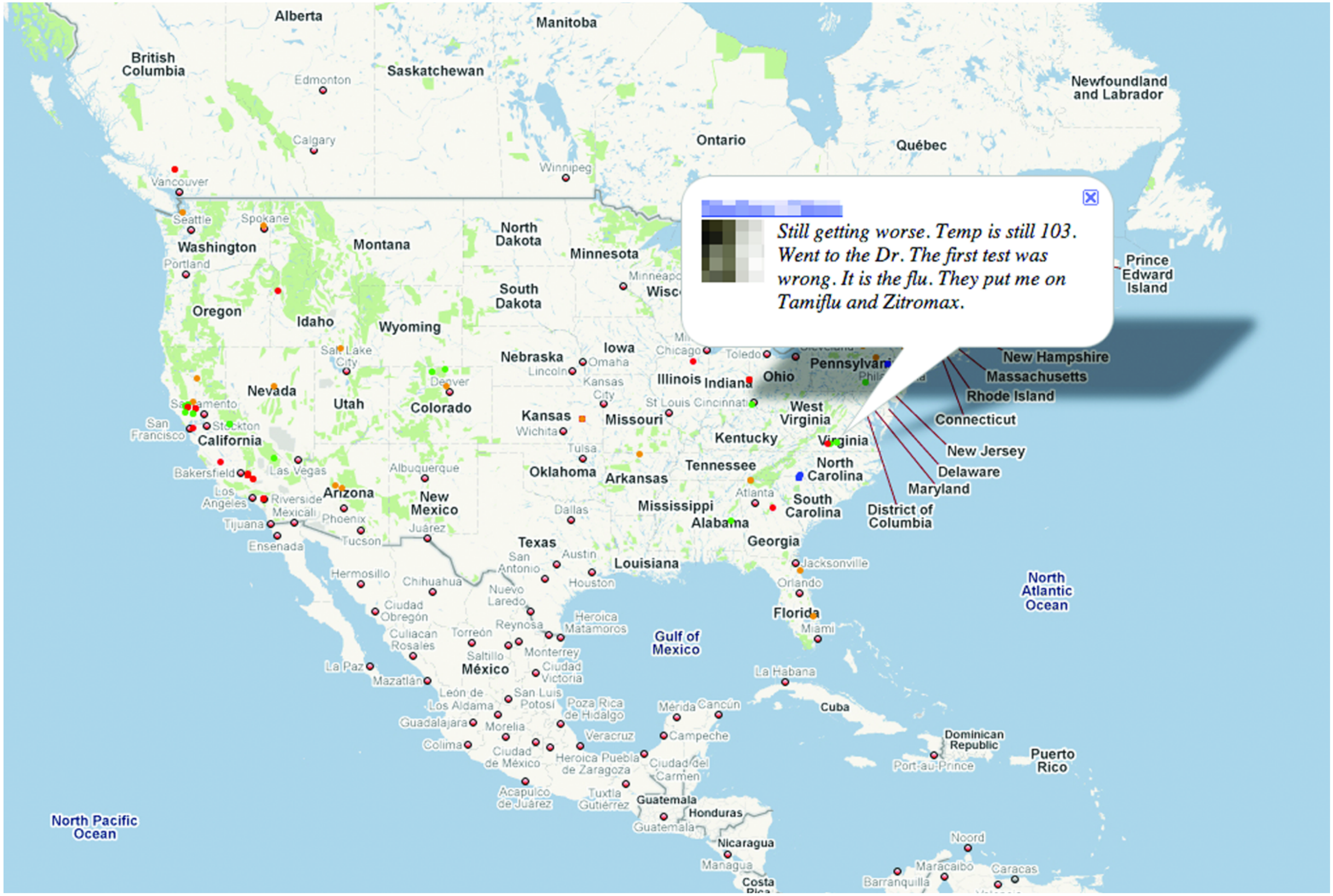
Subject Areas

Twitter | Influenza | H1N1 | Swine influenza | Infectious disease s… | Public and occupati… | Vaccines

influenza & Twitter

- ground truth data from official source
  - CDC ILI (Influenza-Like Illness) data
- proxy data from digital source
  - messages in Twitter mentioning ILI symptoms
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
  - Support Vector Regression
- validate model performance
  - leave-1-out cross-validation

influenza & Twitter



https://dev.twitter.com/streaming/overview

Twitter Developer Documentation

Streaming APIs

Overview

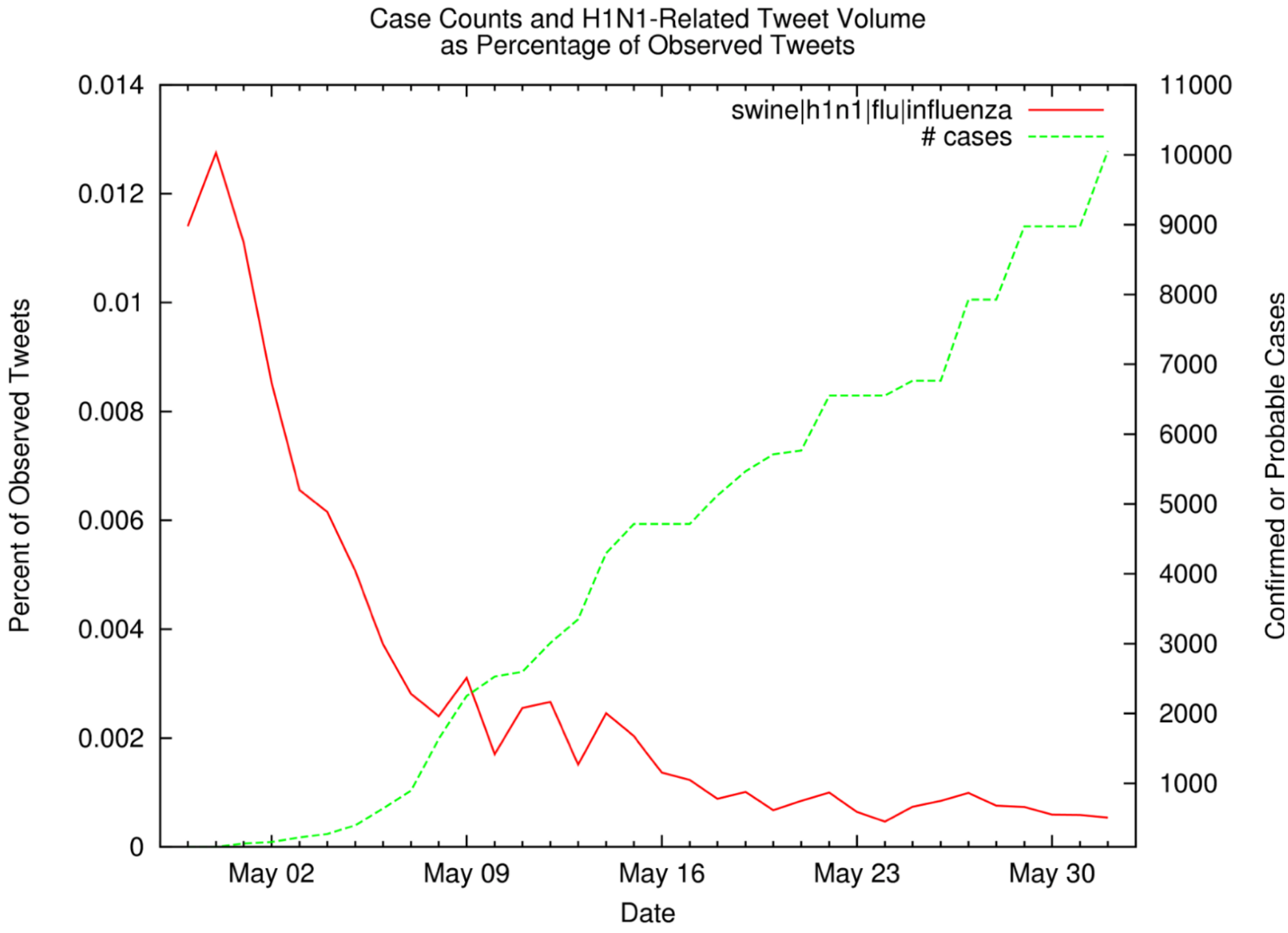
The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data. A proper implementation of a streaming client will be pushed messages indicating Tweets and other events have occurred, without any of the overhead associated with polling a REST endpoint.

If your intention is to conduct singular searches, read user profile information, or post Tweets, consider using the REST APIs instead.

Twitter offers several streaming endpoints, each customized to certain use cases.

Public streams	Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining.
User streams	Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.
Site streams	The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users.

influenza & Twitter



methodology

- use Twitter Streaming APIs to collect tweets mentioning flu-related keywords
- keywords chosen by domain experts
- filter tweets (geographic information)
- process content of tweets (stemming)
- compute normalized feature vector of term frequencies for each time interval
- fit Support Vector Regression model using keyword frequencies as feature vector

Support Vector Regression

sklearn.svm.SVR

class sklearn.svm. **SVR** (*kernel='rbf', degree=3, gamma='auto', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=-1*) [source]

Epsilon-Support Vector Regression.

The free parameters in the model are C and epsilon.

The implementation is based on libsvm.

Read more in the User Guide.

Parameters:

C : float, optional (default=1.0)

Penalty parameter C of the error term.

epsilon : float, optional (default=0.1)

Epsilon in the epsilon-SVR model. It specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value.

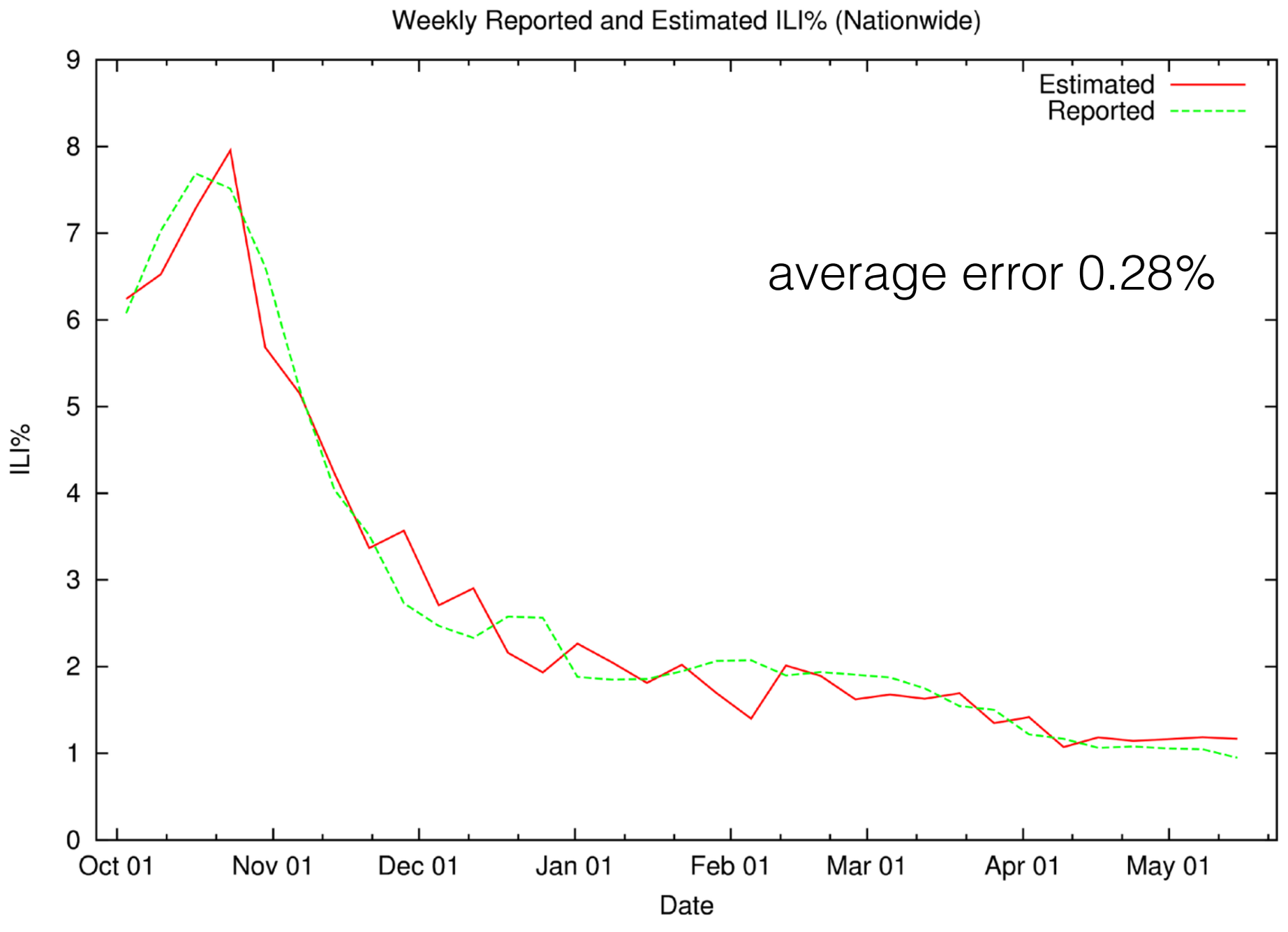
kernel : string, optional (default='rbf')

Specifies the kernel type to be used in the algorithm. It must be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable. If none is given, 'rbf' will be used. If a callable is given it is used to precompute the kernel matrix.

degree : int, optional (default=3)

Degree of the polynomial kernel function ('poly'). Ignored by all other kernels.

result


Weekly Reported and Estimated ILI% (Nationwide)
Estimated
Reported
average error 0.28%
ILI%
0
1
2
3
4
5
6
7
8
9
Oct 01
Nov 01
Dec 01
Jan 01
Feb 01
Mar 01
Apr 01
May 01
Date

influenza & Twitter

- ground truth data from official source
  - CDC ILI (Influenza-Like Illness) data
- proxy data from digital source
  - messages in Twitter mentioning ILI symptoms
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
  - Support Vector Regression
- validate model performance
  - leave-1-out cross-validation

a pattern for digital disease detection

- ground truth data from official source
- proxy data from digital source
- problem: predict ground truth data (target) from proxy data (features)
- train a statistical/learning model to solve the problem
- validate model performance