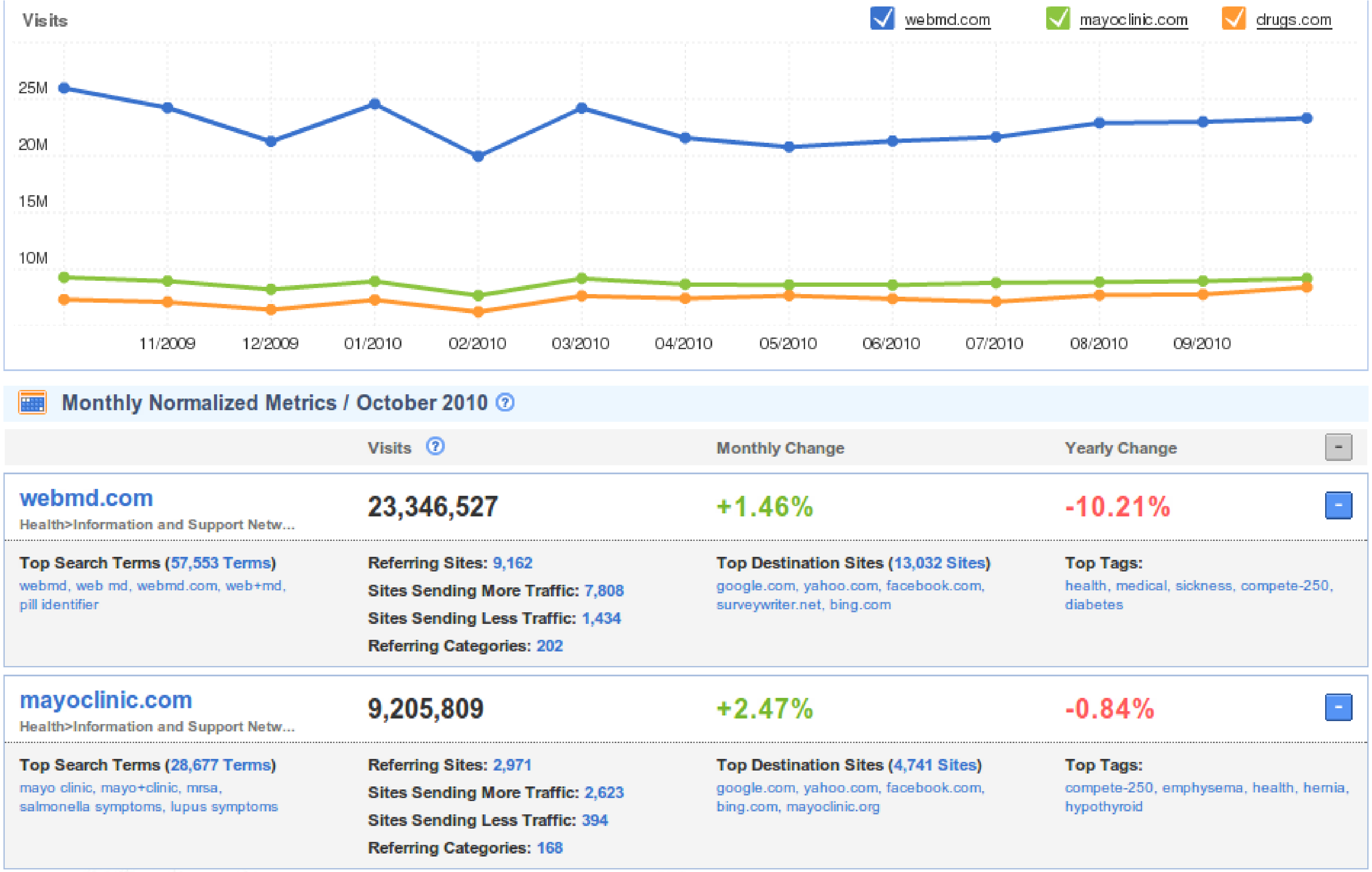
Using Twitter to Estimate H1N1 Activity

Alessio Signorini
<alessio-signorini@uiowa.edu>

Alberto Maria Segre
<alberto-segre@uiowa.edu>

Philip Polgreen
<philip-polgreen@uiowa.edu>

Thousands of Health Websites



Monthly Normalized Metrics / October 2010 ?

	Visits ?	Monthly Change	Yearly Change
webmd.com Health>Information and Support Netw...	23,346,527	+1.46%	-10.21%
Top Search Terms (57,553 Terms) webmd, web md, webmd.com, web+md, pill identifier	**Referring Sites:** 9,162 **Sites Sending More Traffic:** 7,808 **Sites Sending Less Traffic:** 1,434 **Referring Categories:** 202	**Top Destination Sites (13,032 Sites)** google.com, yahoo.com, facebook.com, surveywriter.net, bing.com	**Top Tags:** health, medical, sickness, compete-250, diabetes
mayoclinic.com Health>Information and Support Netw...	9,205,809	+2.47%	-0.84%
Top Search Terms (28,677 Terms) mayo clinic, mayo+clinic, mrsa, salmonella symptoms, lupus symptoms	**Referring Sites:** 2,971 **Sites Sending More Traffic:** 2,623 **Sites Sending Less Traffic:** 394 **Referring Categories:** 168	**Top Destination Sites (4,741 Sites)** google.com, yahoo.com, facebook.com, bing.com, mayoclinic.org	**Top Tags:** compete-250, emphysema, health, hernia, hypothyroid

Playing Doctor on Google

Flu Searches

flu symptoms, stomach flu, flu duration, flu treatment, how long does the flu last, cold, pneumonia, fever, bronchitis, influenza, tamiflu, strep throat

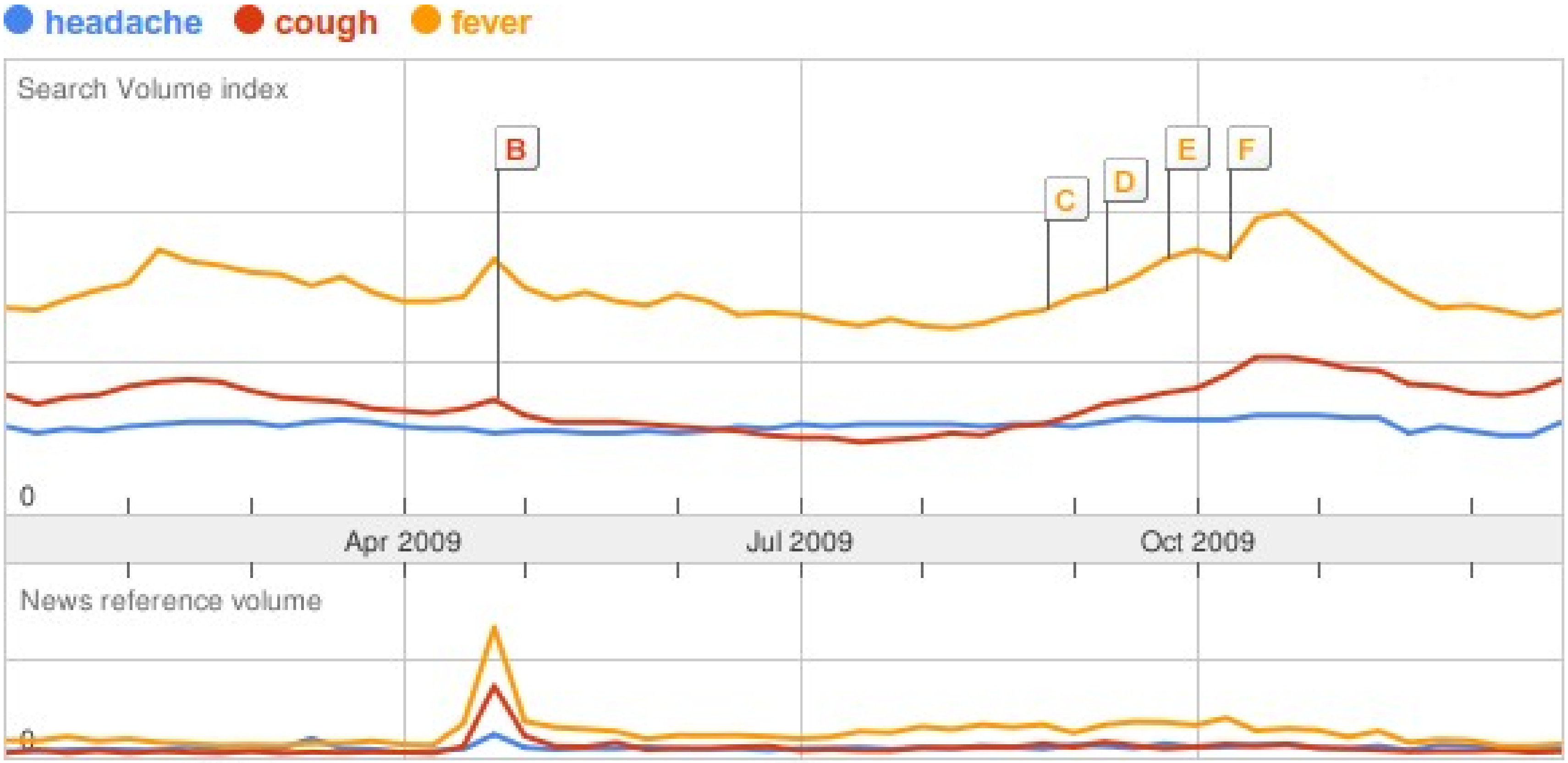
Cough Searches

bronchitis, pneumonia, cold, tuberculosis, flu, sneeze, dry cough, cough medicine, whooping cough, chronic cough, cough remedies, cough treatment, acute cough

Headache Searches

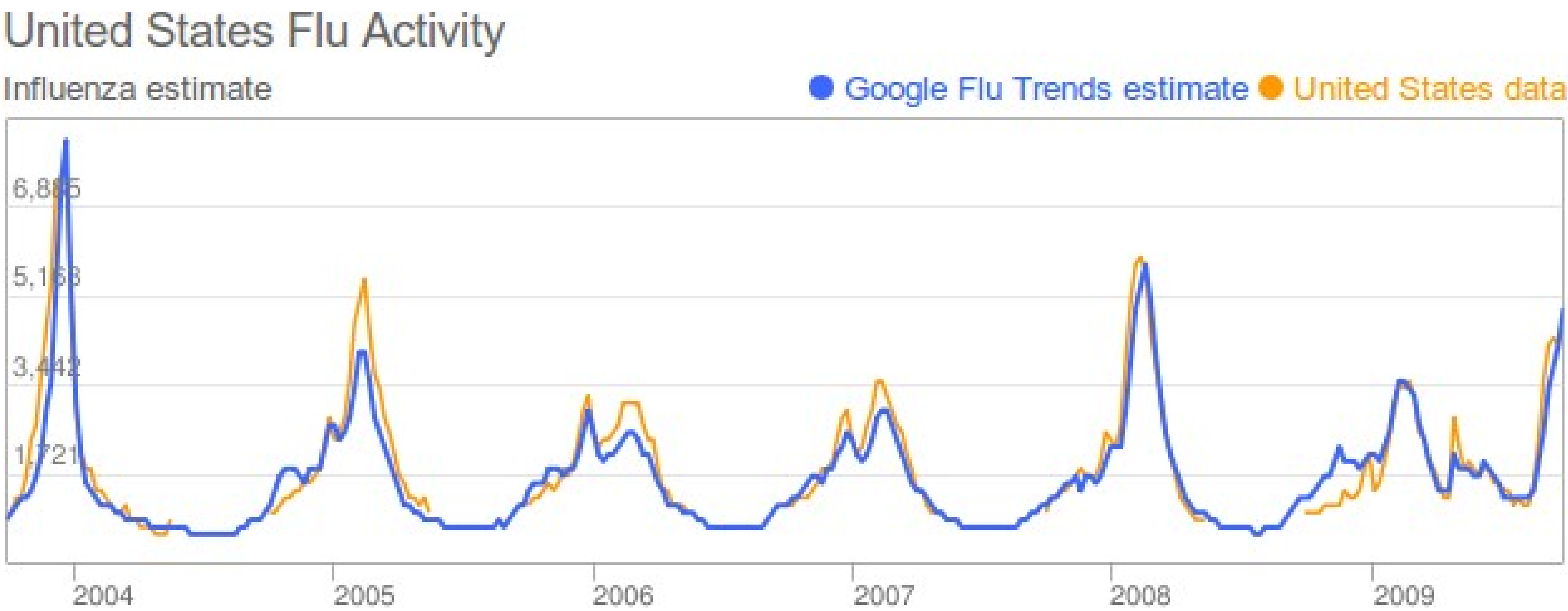
sinus headache, headache causes, headache types, headache remedies, headache cures, headache treatment, headache back of head, migraine, brain tumor, meningitis

More Health Queries = Sick?



http://www.google.com/trends

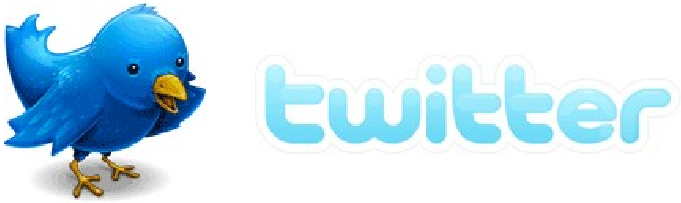
Google Flu Trends (2009)



http://www.google.org/flutrends/

Philip Polgreen and Yahoo! Research published similar results in 2008.

Luckily, Twitter was Invented

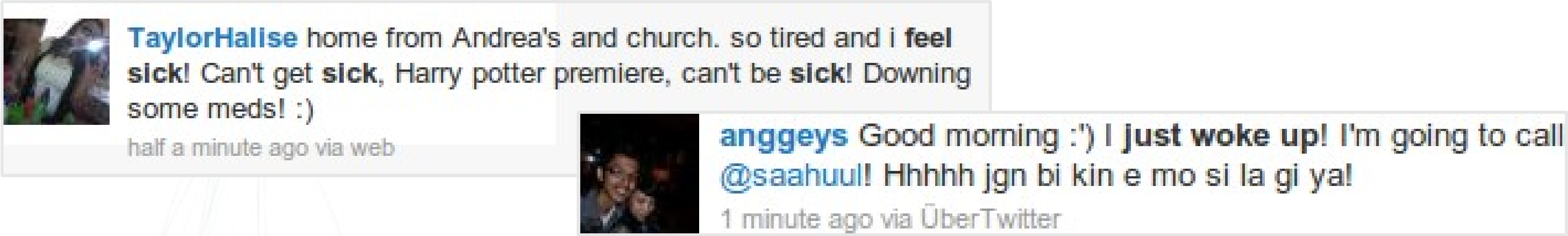


Personal Micro-Blog for Short Status Updates

(~ 60 Million per day!)

People share lots of information:

where they are, what they are doing, with whom,
what they are eating, how they feel, ...



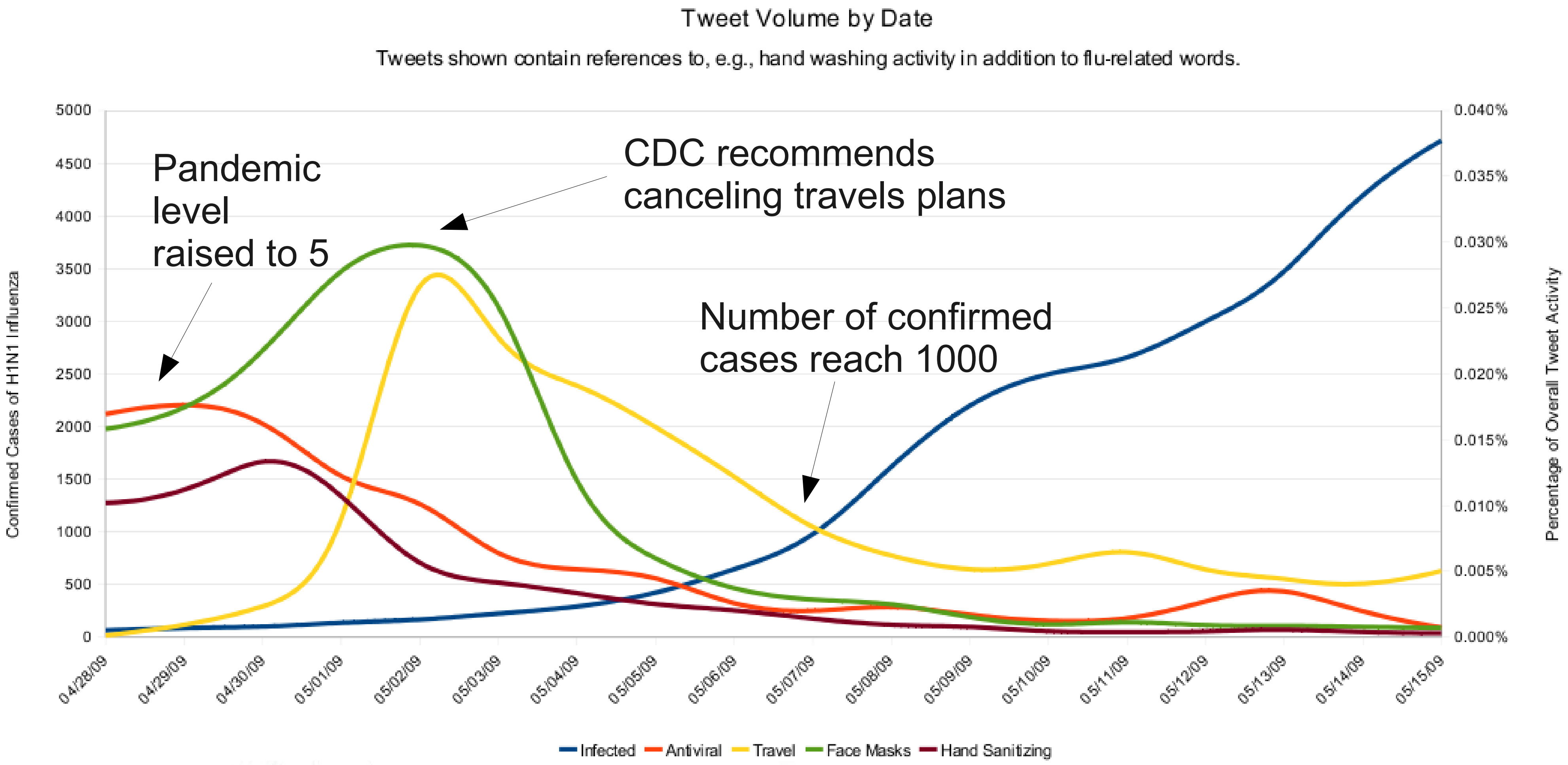
TaylorHalise home from Andrea's and church. so tired and i **feel sick**! Can't get **sick**, Harry potter premiere, can't be **sick**! Downing some meds! :)

half a minute ago via web

anggeys Good morning :') I **just woke up**! I'm going to call @saahuul! Hhhhh jgn bi kin e mo si la gi ya!

1 minute ago via ÜberTwitter

H1N1 2009: Tweets Volume



Pandemic level raised to 5

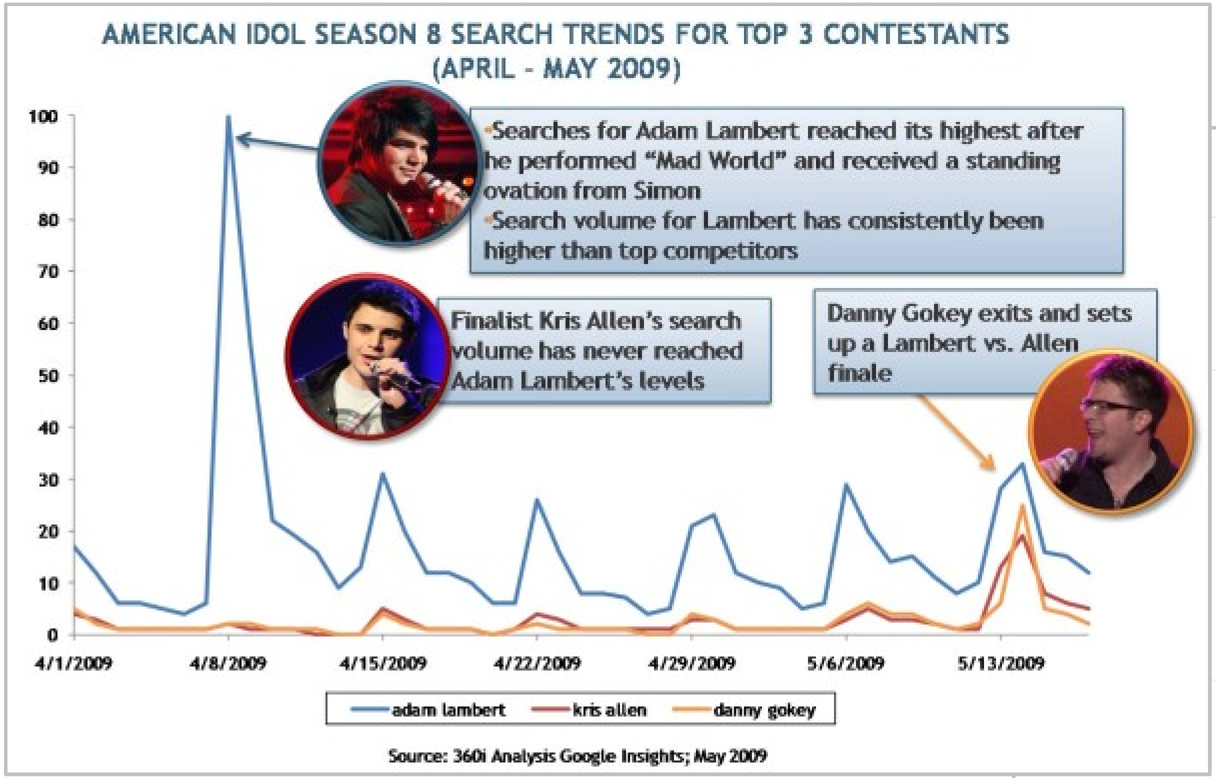
CDC recommends canceling travels plans

Number of confirmed cases reach 1000

comp|epi
computational epidemiology research

The University of Iowa

American Idol: Queries vs. Twitter



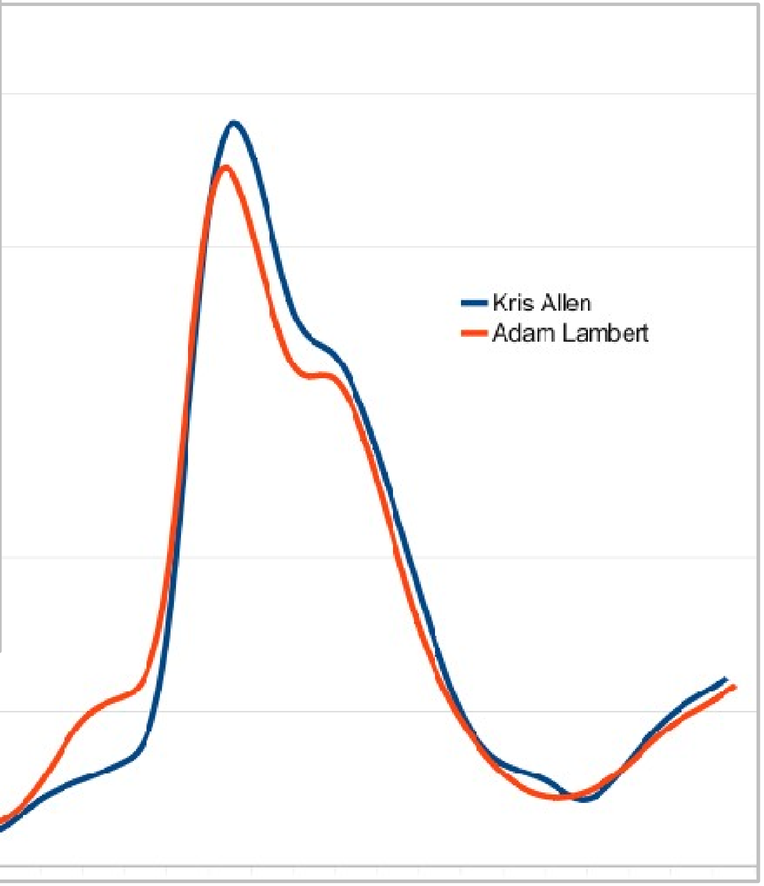
AMERICAN IDOL SEASON 8 SEARCH TRENDS FOR TOP 3 CONTESTANTS (APRIL - MAY 2009)

- Searches for Adam Lambert reached its highest after he performed "Mad World" and received a standing ovation from Simon
- Search volume for Lambert has consistently been higher than top competitors

Finalist Kris Allen's search volume has never reached Adam Lambert's levels

Danny Gokey exits and sets up a Lambert vs. Allen finale

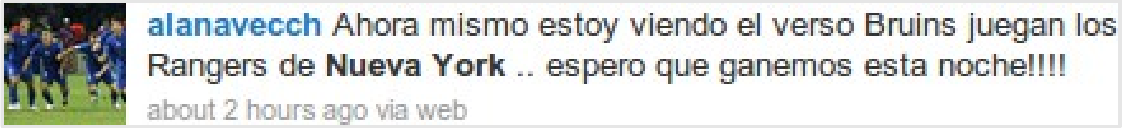
Source: 360i Analysis Google Insights; May 2009



Google query volume declared Adam Lambert as winner but tweet sentiment analysis suggested Kris Allen would win.

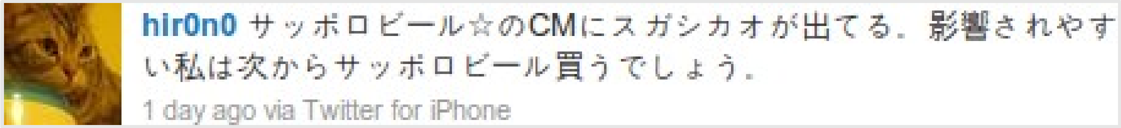
Tweets are Often Messy

Non-English


alanavecch Ahora mismo estoy viendo el verso Bruins juegan los Rangers de **Nueva York** .. espero que ganemos esta noche!!!!

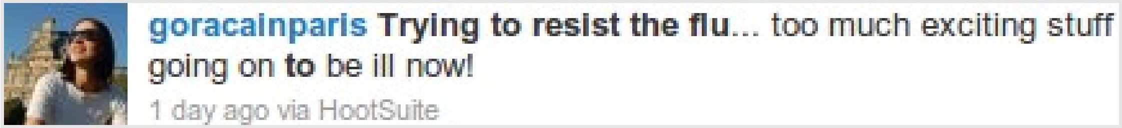
about 2 hours ago via web

Non-ASCII


hir0n0 サッポロビール☆のCMにスガシカオが出てる．影響されやすい私は次からサッポロビール買うでしょう．

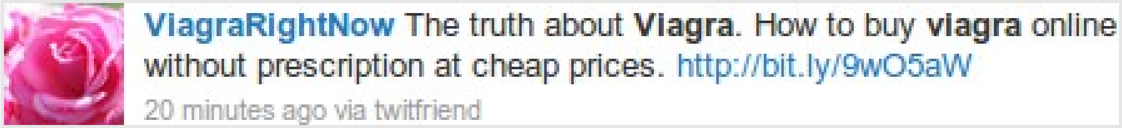
1 day ago via Twitter for iPhone

Out of US


goracainparis **Trying to resist the flu**... too much exciting stuff going on **to** be ill now!

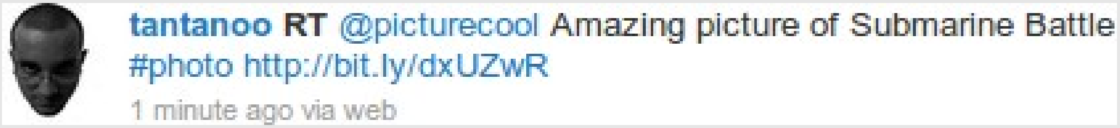
1 day ago via HootSuite

Spam


ViagraRightNow The truth about **Viagra**. How to buy **viagra** online without prescription at cheap prices. http://bit.ly/9wO5aW

20 minutes ago via twitfriend

Jargon


tantanoo **RT** @picturecool Amazing picture of Submarine Battle #photo http://bit.ly/dxUZwR

1 minute ago via web

Typos and Stemming

Tweets contains plenty of typos and misspellings
(e.g., migrane, flue, cought, …)

We decided to eliminate any term with only a few occurrences in each week.

Words can be Inflected or Derived
(e.g., ill, illness, sick, sickest, ...)

The process of reducing words to their root is called Stemming. Many algorithms exists, we used the well-known Porter Algorithm.

Few Words are Really Important

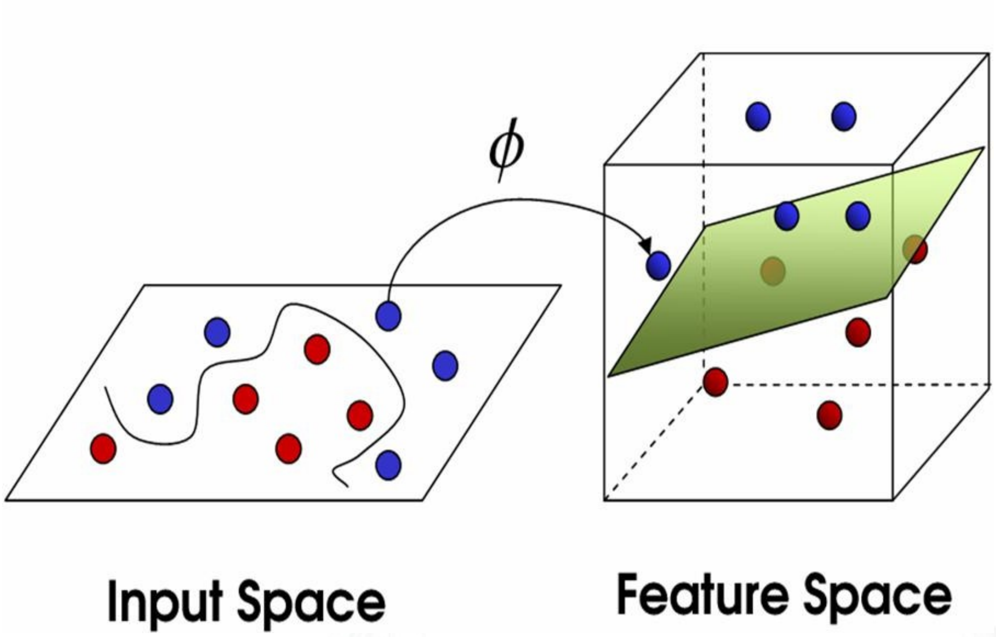
Stopwords are usually not relevant
we excluded the most common during our analysis
(e.g., the, and, with, of, …)

Our first experiment tracked only tweets containing
words correlated with influenza
(e.g., flu, h1n1, influenza, cough, tamiflu, …)

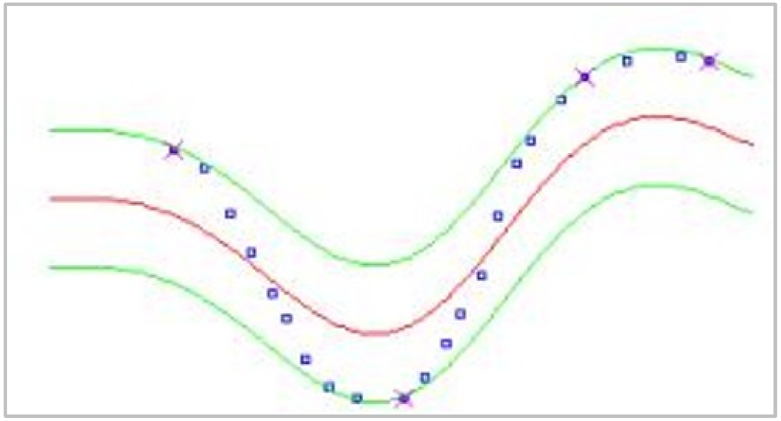
A later experiment tracked a
random 5% sample of all tweets
but noise was overwhelming.

Support Vector Regression

Support Vector Machines (SVM) are a set of supervised learning methods used for classification and regression.



Classification
http://www.imtech.res.in/raghava/rbpred



Regression
http://kernelsvm.tripod.com/

Training and Testing

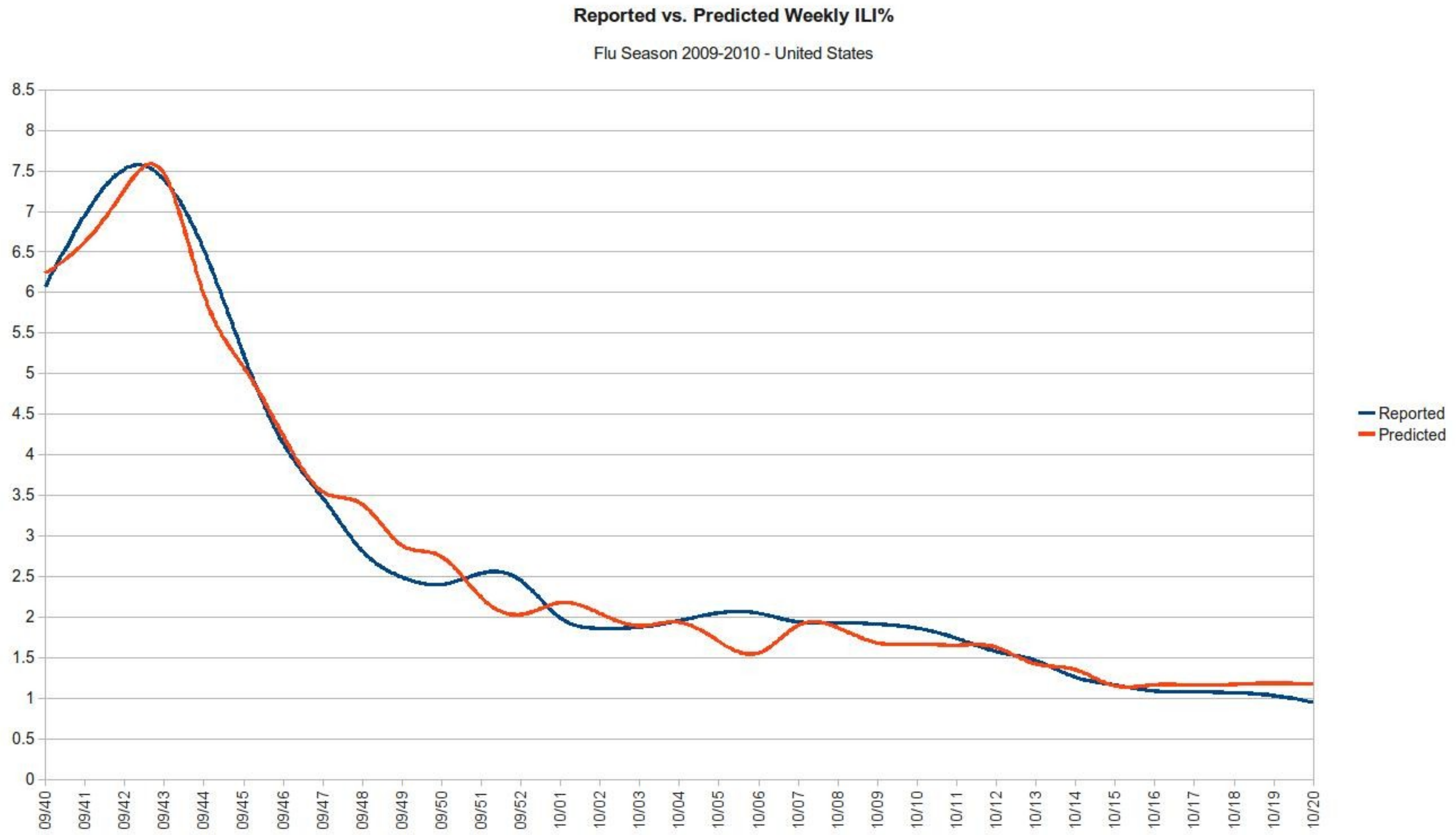
We used the popular libSVM library
and a polynomial kernel.

The dataset included 32 weeks of data,
about 4.2M tweets. We used n-fold validation.

Our target was the weighted ILI% for each week.
at first of the entire US, then of each HHS region.

Examples of highly-correlated terms:
flu, cough, shot, immun, sick, vaccin, school,
sneez, virus, germ, wash, pregnant, ...

Reported vs. Predicted (US)



1-fold validation ~ error avg=0.28%, min=0.04%, max=0.93%. Std=0.23%

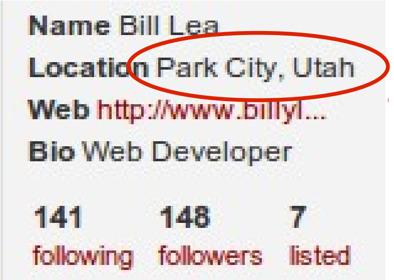
User/Tweet Geolocation

Tweets are often tagged with the geographical coordinates of the user who sent them.

Last year this technology was not widely adopted.

When geolocation was not available, we used the location declared in the user's profile.

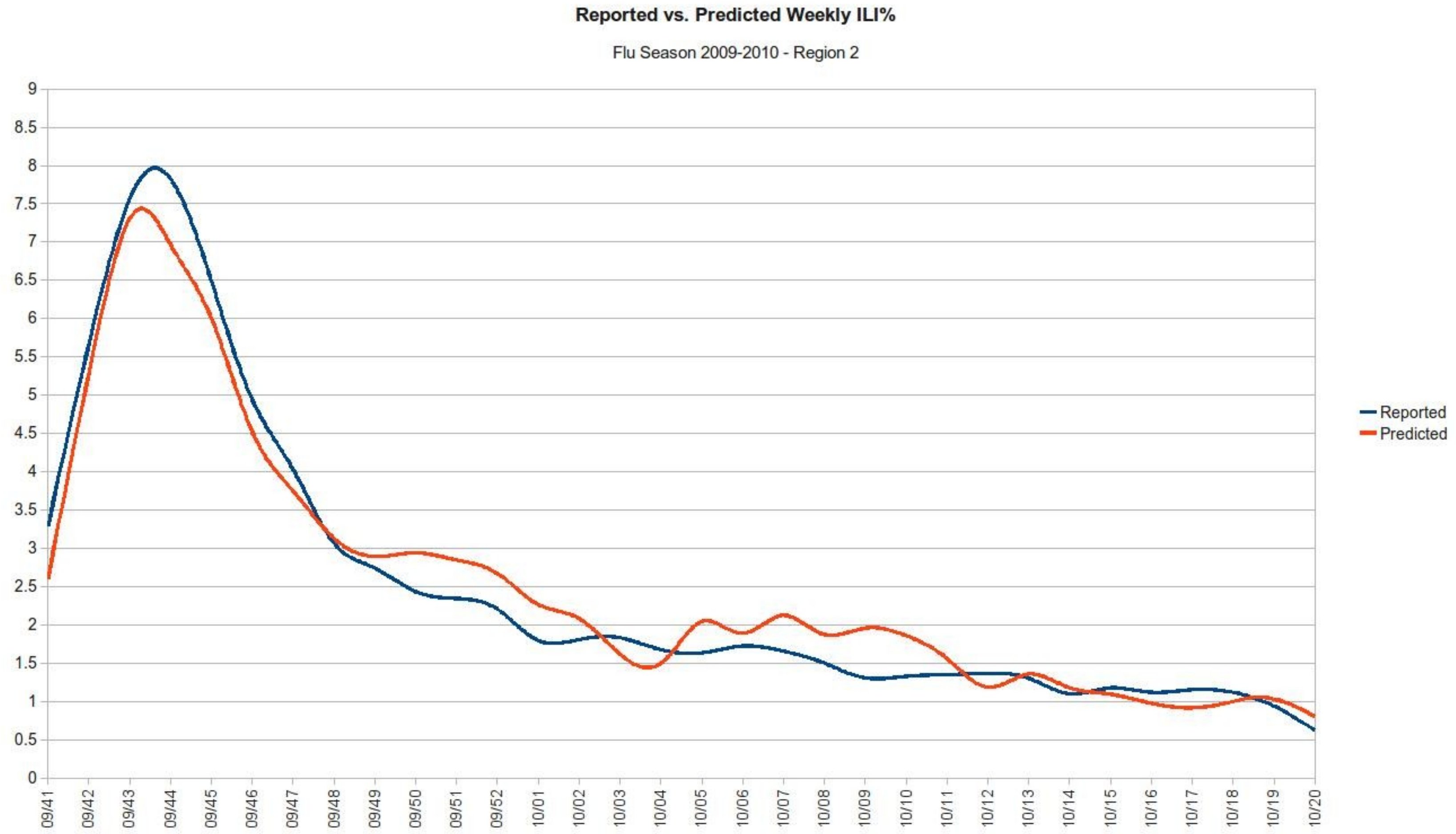
billylea



Name Bill Lea
Location Park City, Utah
Web http://www.billyl...
Bio Web Developer

141	148	7
following	followers	listed

Reported vs. Predicted (NY+NJ)



Out-of-sample Prediction ~ error avg=0.37%, min=0.01%, max=1.25%. Std=0.26%

Where to Get More Information

Alessio Signorini
alessio-signorini@uiowa.edu
http://www.cs.uiowa.edu/~asignori/

UIOWA Computational Epidemiology Group
http://compepi.cs.uiowa.edu

paper and datasets will be soon available
on the CompEpi website