Cool Things You Can Do with Internet for Diseases Forecasting

April 21th, 2011

Alessio Signorini
alessio-signorini@uiowa.edu

Alessio Signorini – Who am I?

Born in Pisa, Italy and played professional soccer until seven years ago. No coffee, wine or cigarettes for me.

Director of Technology for Ask.com, then started OneRiot with a similar role. PhD Candidate at the University of Iowa, often work with Alberto Segre and Phil Polgreen.

Recently founded company which uses facial recognition and AI to target advertising on mall/airport billboards. Freaky but interesting, I will tell you later about it.

Research Interests – Everything?

I have a very broad range of interests and always find a way to sneak one or two more projects in my schedule:

- Web Search
- Natural Language Processing
- Clustering/Classification of News
- Artificial Intelligence
- Computer Vision
- Optimization
- Personalization of Search/Things
- World Peace

Random Personal Projects

Decided to optimize a keyboard layout for my personal use because DVORAK was not enough. Fun project and statistics were great. Too lazy to re-learn how to type.

Zappos has 52 colors for men shoes (e.g., “Tan Mad Cat Goat”?). I just wanted some brown shoes! Downloaded all shoe images, clustered by color, got a job offer.

Boulder County Schools get only 65c for each kid meal. Using weather, flu and attendance data, plus past sales, can reduce waste and food costs to improve meal quality.

“ Talk About Something Cool ”

Web is Growing: Users and Content

By the end of 2008 more than 82% of the household had Internet access. Users spend online 48h/week, 75% have Facebook/MySpace profiles and ~15% use blogs/forums.

Historical data, maps, graphs, and many other resources are available online for free. Many Encyclopedias and other publications exist today only in electronic form.

More than 20% of Americans look for medical advices online. Health domains (e.g., WebMD, MayoClinic, …) are among the most popular sites of the Internet.

The Web in Numbers (March 2011)

23.3 Billion

Minutes/day spent on Facebook

16.9 Billion

Searches/month

27.2 Million

Blog Posts/month

140 Million

Twitter Messages/day

Google Tracks you Around the Web

As soon as you visit a site with some Google's stuff on it, a cookie is saved on your machine and you are being tracked. Examples:

Browser makes JS/IFrame request to Google's server and they use “Referral URL” to identify originating page. When you log into something data is associated to your profile.

Browser Signature: Tracking w/o Cookies

I wrote “Tab Cookies”, a Google Chrome Extension that deletes unused cookies when you close a browser tab.

The combination of resolution, plugins, OS, browser, etc, provides a pretty unique ID of your computer. Check out the work of the guys at

http://panopticlick.eff.org

Surrender, you can and are tracked! Even easier/better if somebody has access to the proxy logs of your company or university.

From Query/Posts/URLs to User Infos

Plenty of research (e.g., Microsoft/Yahoo) show how much can be inferred from the query logs of somebody: gender, age, location, income, education, health, …

Other researches show how something similar can be done examining the posts of a user on a blog, Twitter, MySpace or Facebook.

Examining the URLs visited by a person allows to infer similar data and to create a profile of the user.

“Apache”: Indians or Web Server?

The query “apache” is frequent in search engine's logs. If you are a geek, it is synonymous of “web server”.

But 70% of times what users are looking for are information on the Indian tribe. About 8% of the times, they want the helicopter.

One could dedicate 7 results to the Indians, 2 to the web server and 1 to the helicopter. Using your profile results could be personalized.

Mining Profiles and Query/URLs for Health

Intersecting user profiles, IP geolocation and URLs visited could reveal interesting data. If you are visiting

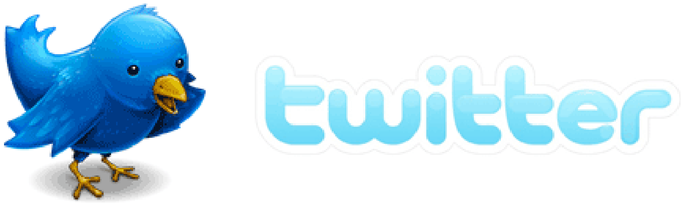
www.mayoclinic.com/health/cold-sore/DS00358

you probably have or suspect to have a cold sore. Where do you go next? Your clicks may reveal if you are looking for symptoms or remedies.

Big universities and companies can do this kind of analysis on their proxy logs. Wikipedia's proxy logs are public and often show interesting peaks in traffic.

What if you do not have logs?

Alternative to Google Logs: Twitter

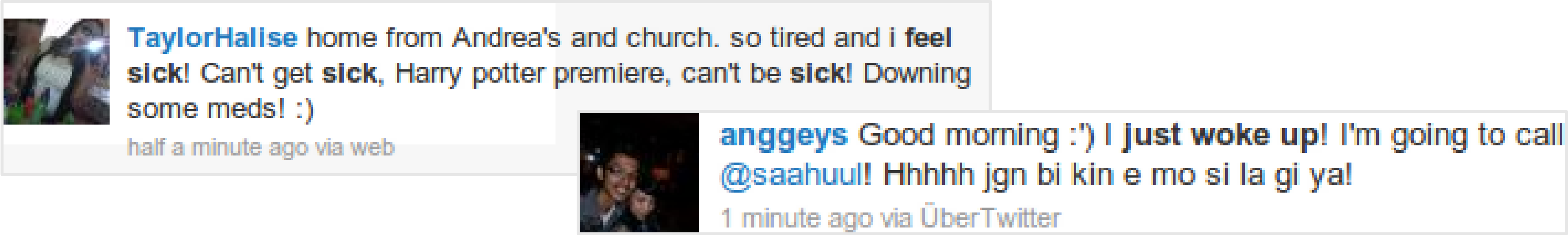


Personal Micro-Blog for Short Status Updates

(~ 140 Million per day!)

People share lots of information:

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



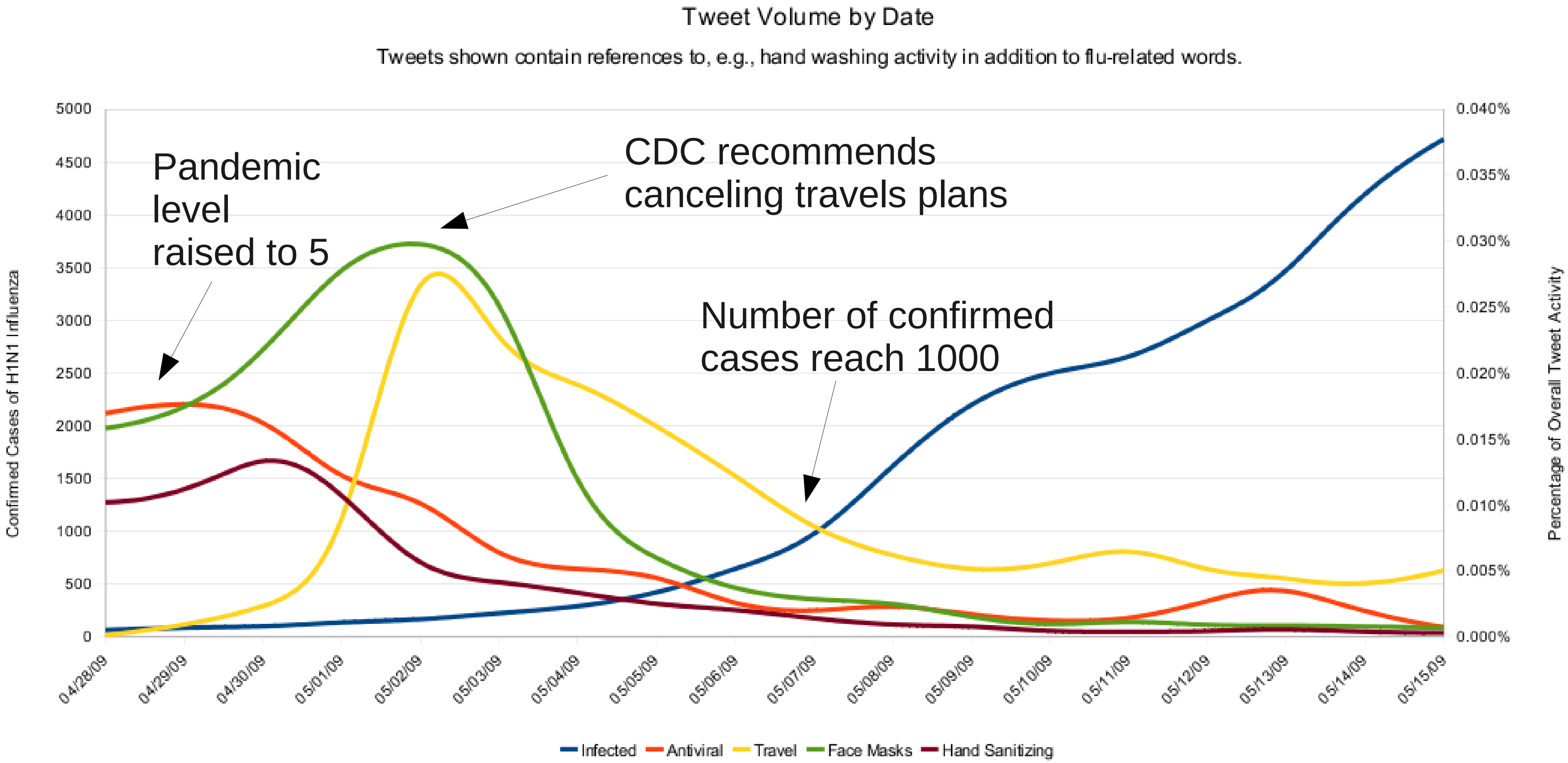
TaylorHalise home from Andrea's and church. so tired and i **feel sick**! Can't get **sick**, Harry potter premiere, can't be **sick**! Downing some meds! :)

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

Number of Tweets during H1N1

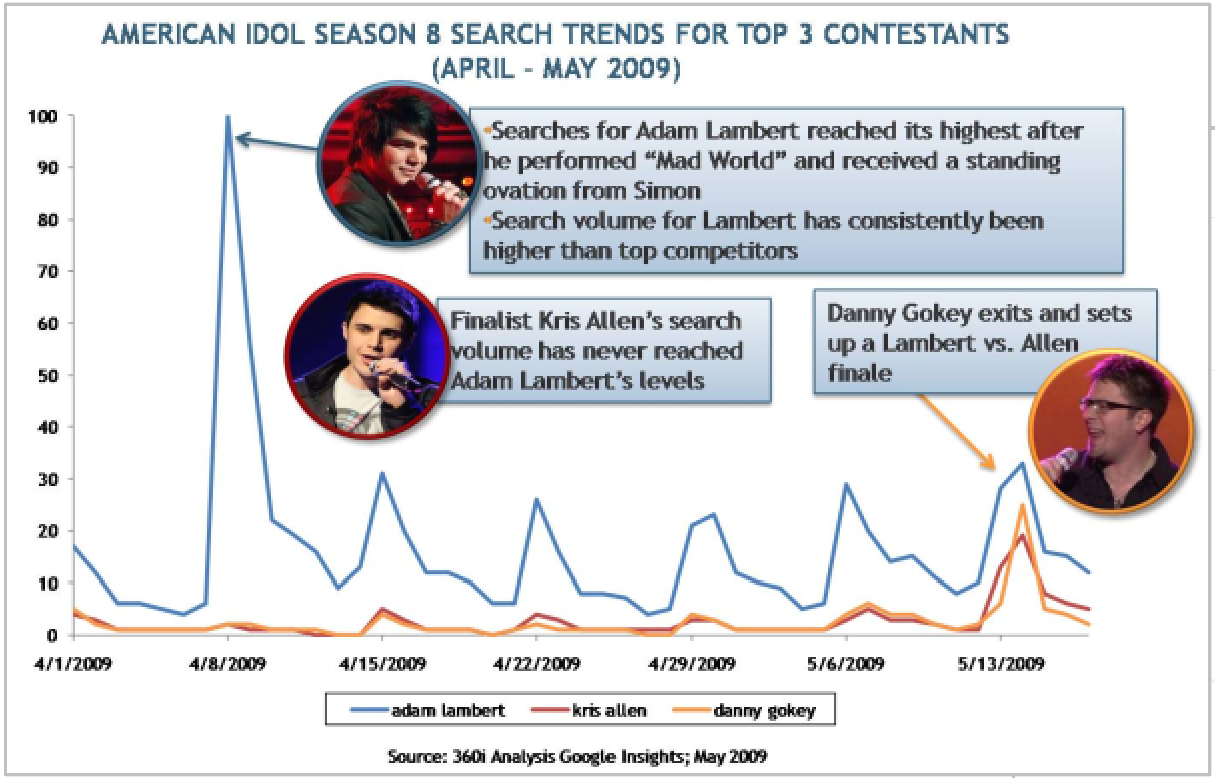


Pandemic level raised to 5

CDC recommends canceling travels plans

Number of confirmed cases reach 1000

American Idol: Queries vs. Twitter



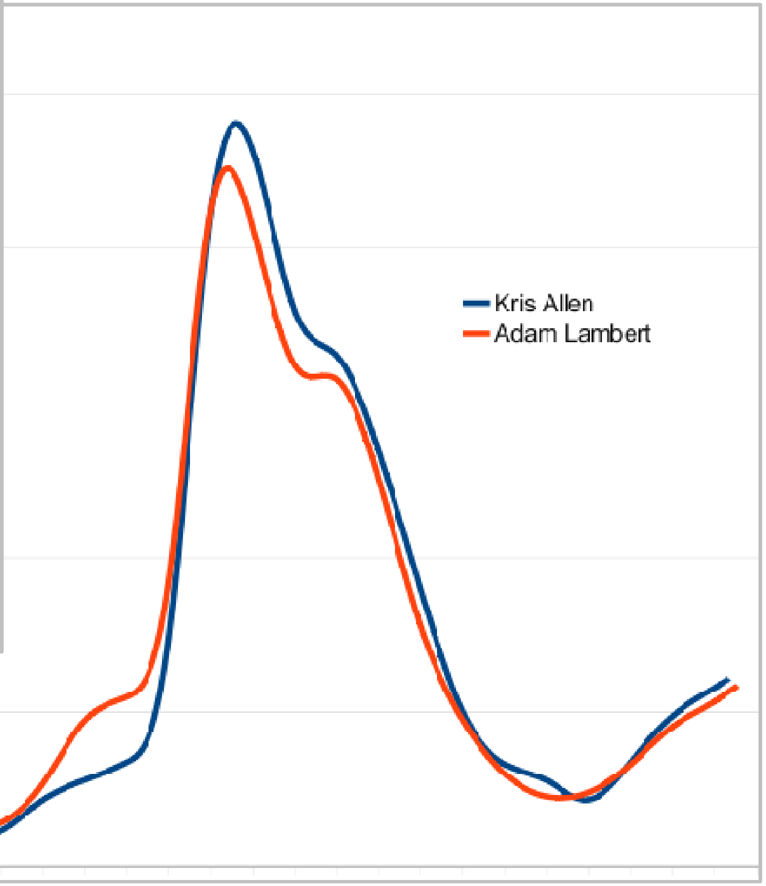
AMERICAN IDOL SEASON 8 SEARCH TRENDS FOR TOP 3 CONTESTANTS (APRIL - MAY 2009)

- 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

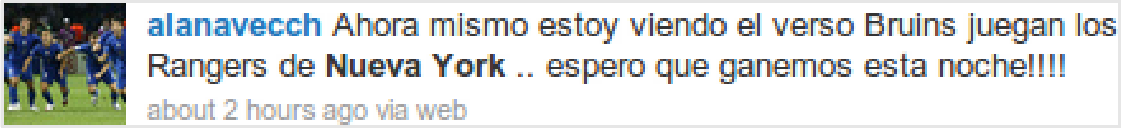
Source: 360i Analysis Google Insights; May 2009



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

Tweets are Often Messy

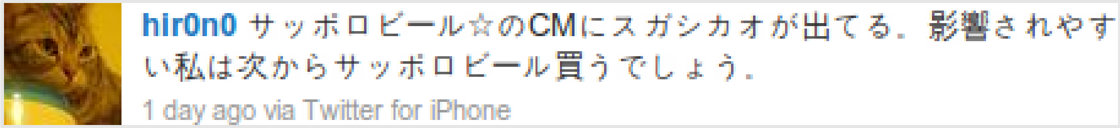
Non-English



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

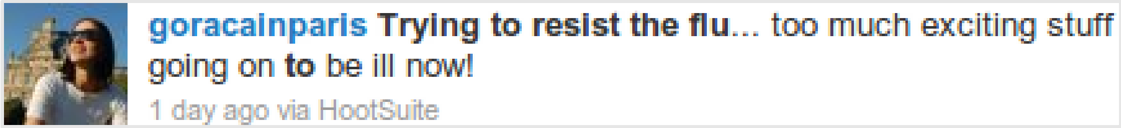
Non-ASCII



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

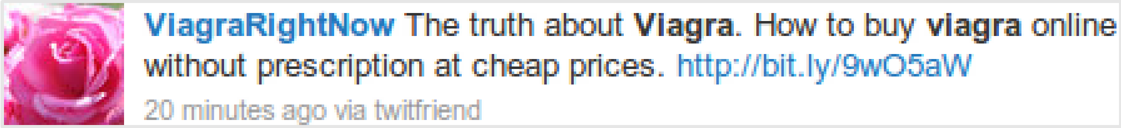
Out of US



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

1 day ago via HootSuite

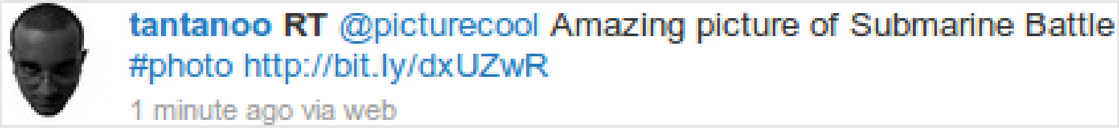
Spam



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

More Cleanup: Stopwords and Stemming

Original:

I feel sicker and sicker, this flu is never going to go away!

Removal of Stopwords (very common words):

feel sicker sicker flu never going go away

Stemming (reducing words to root):

feel sick sick flu never go go away

Duplicate Removal:

feel sick flu never go away

From Tweets to ILI%: Training

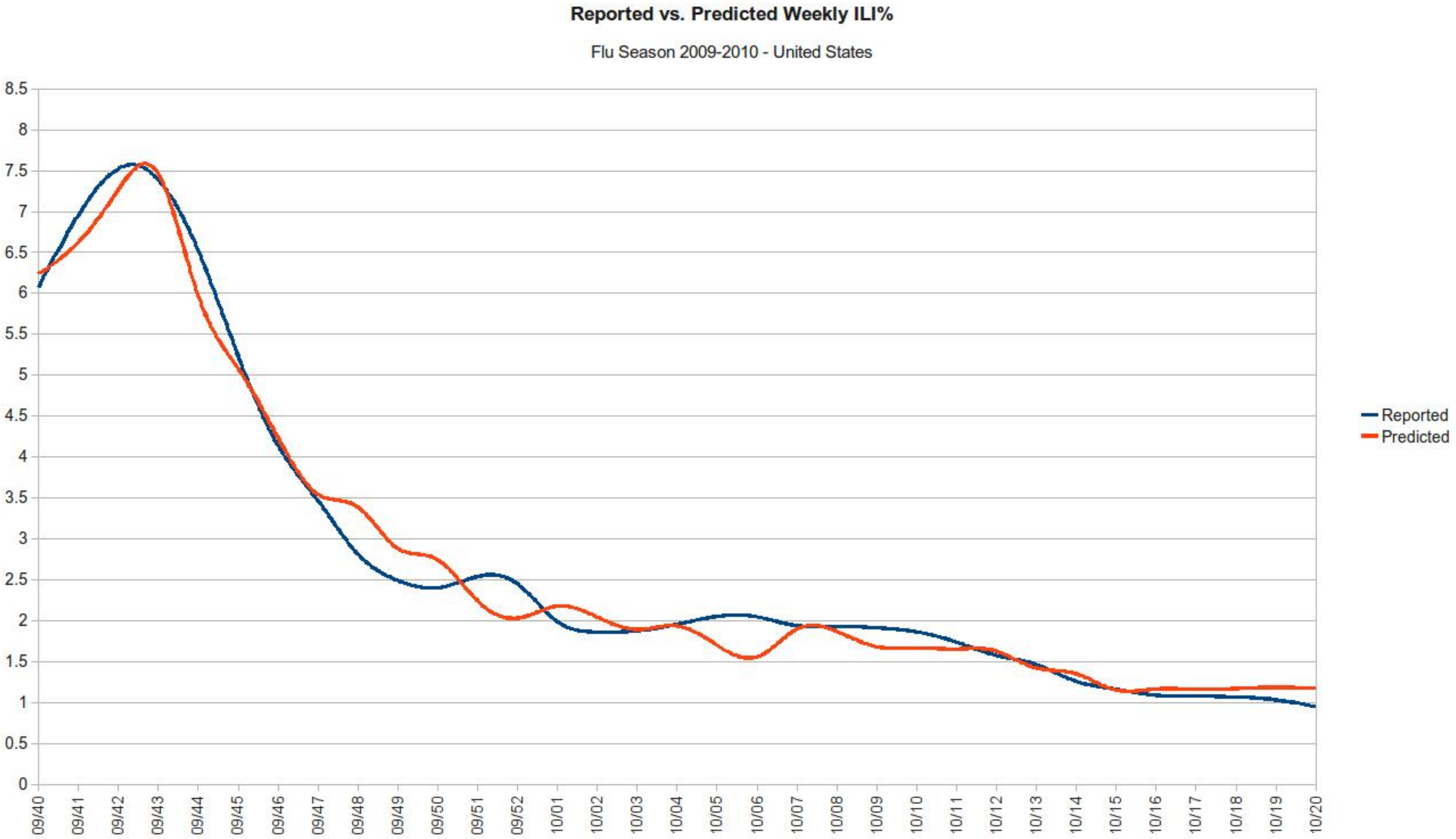
We used the popular library libSVM with a polynomial kernel. The dataset included 32 weeks of data, about 4.2M tweets. We used n-fold validation.

Each term was a feature and its value was the normalized #occurrences. 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, ...

ILI% Reported vs. Estimated (US)



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

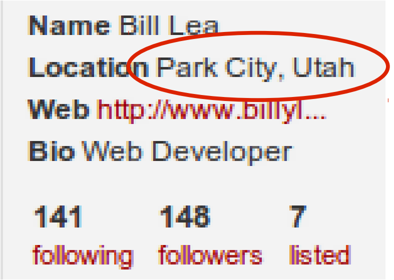
Users Tweet Geolocation

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.

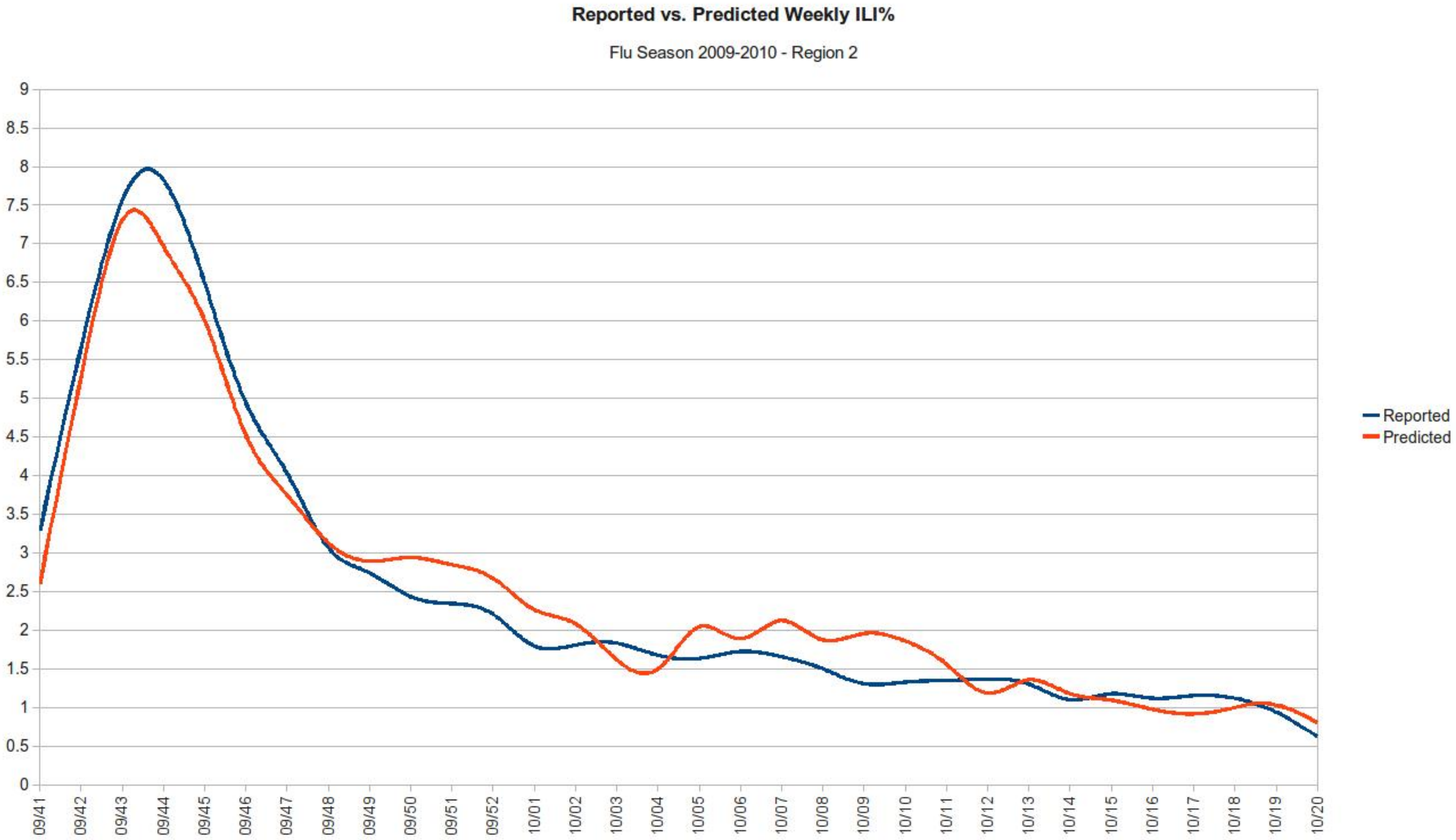
billylea



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

141	148	7
following	followers	listed

ILI% Reported vs. Estimated (NY+NJ)



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

Where will it go next?

Travel Models without Airlines/GSM

A few years ago it was possible to work with airline companies and get tickets data to create travel models. After 9/11 this is very-very difficult, if not impossible.

GSM towers data could be a good alternative, but phone companies are super-secretive about those and almost never release them to the public.

Recent studies used “Where is George” data to create in-town probabilistic travel models. Others, used speedway traffic data.

Travel Models using Check-in's

Luckily, the recent popularity of GPS receiver on phones allowed the creation of dozens of "check-in" applications.

foursquare

Gowalla

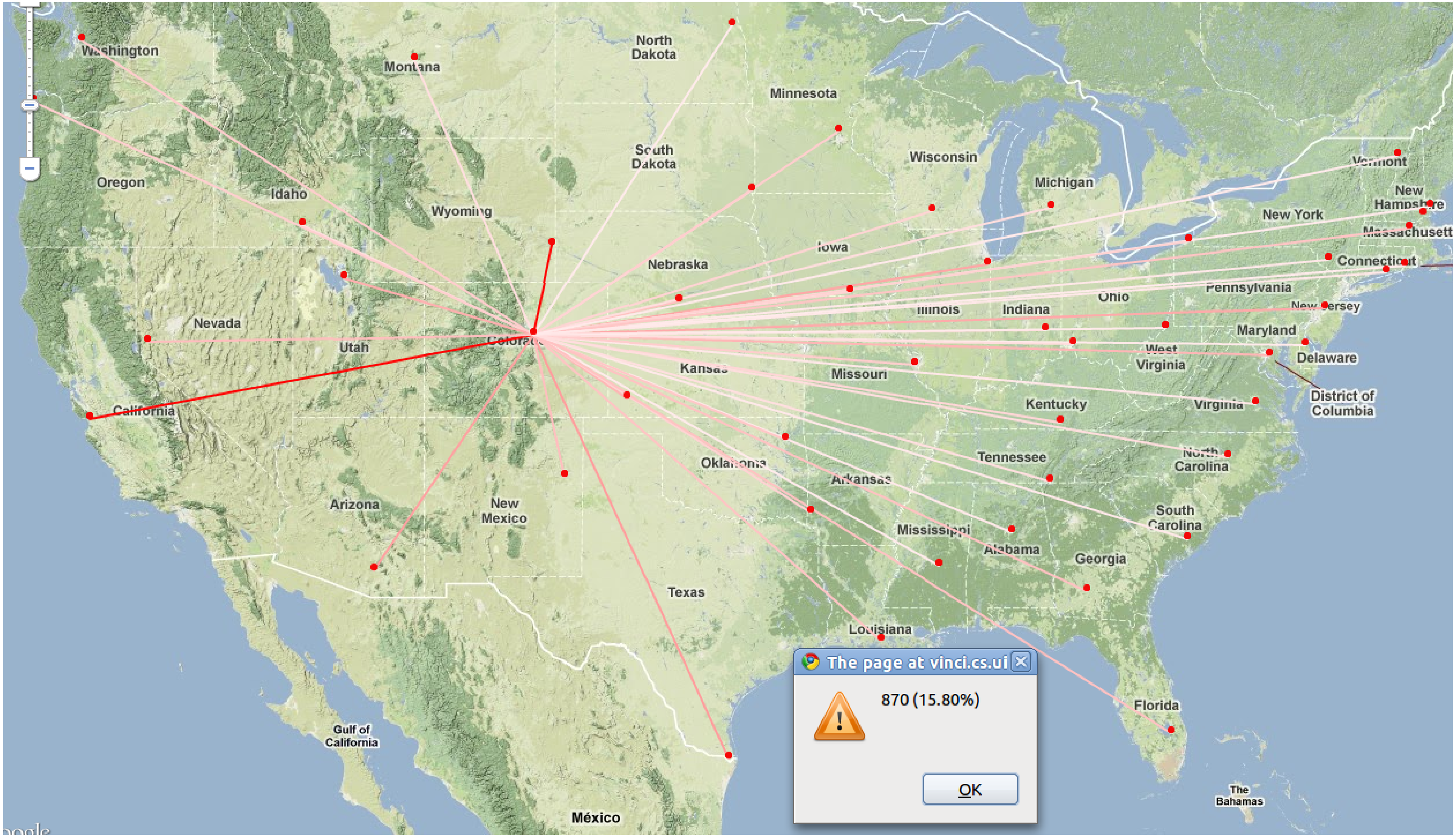
Places

Who. What. When. And now **Where.**

Every check-in is associated with some specific GPS coordinates, or an area (e.g., if you are in a park). Foursquare alone receives more than 3 Million check-in's per day.

These data can be obtained using the Foursquare API or through Twitter's Streaming API.

Example of Travels Data from Colorado



870 (15.80%)

http://vinci.cs.uiowa.edu/~alessio/twitter/travel-paths/

Have you seen Minority Report?

Current Status of Digital Billboards

There are more than 3 Million pedestrian digital signs in the US. Unfortunately, they are no more than slideshows, changing the Ad (randomly) every 15 seconds.

Buying is hard, since they are fragmented in 400 different networks. There is also no accountability, mostly relies on the traffic details the owner provides.

Finally, although 70% are Internet connected, distribution of the creatives is still mostly manual, with guys walking around with USB keys and CDs loading things up.

Google Ads for the Real World?

Lots of progresses have been made in computer vision (e.g., gender, age, race, height, ...) in the last years. In addition, good webcams and computers are now cheap.

FourSquare, PlaceIQ, SimpleGeo, …, aggregate user information and provide great demographic information given an area.

We combine all those, plus weather, ambiance noise, and much more, and use AI to optimize the Ads displayed. We also monitor user attentions and learn from it.

Analytics: the “click” of Billboards

Given some variables (e.g., time, place, weather) with enough samples and some multivariate analysis we can estimate the expected attention time given a user/Ad.

Ads are selected trying to maximize the attention time of the crowd. We check if people looked “long enough” and learn from it.

Many screens support other interactions methods like a touch, the scan of a QR code, sending a text message, etc...

Not Bored Yet?

Alessio Signorini

alessio-signorini@uiowa.edu

www.alessiosignorini.com
blog.alessiosignorini.com

@a_signorini