Use of Social Media to Monitor and Predict Outbreaks and Public Opinion on Health Topics

Alessio Signorini

Department of Computer Science
University of Iowa

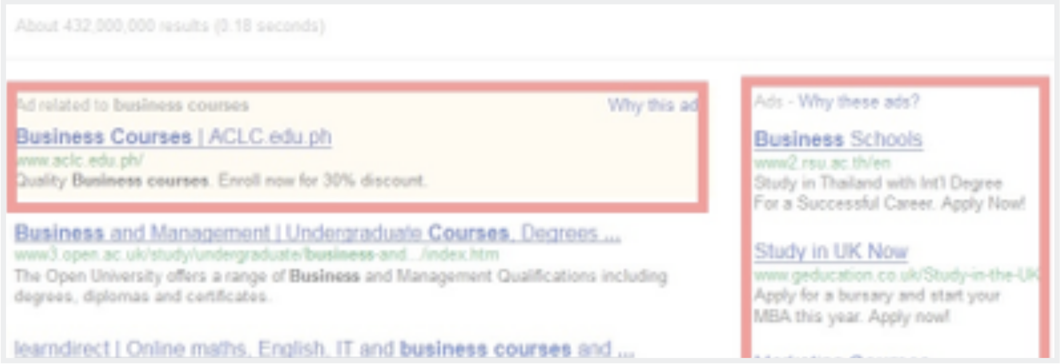
December 3rd, 2014

“ Measurement is the first step
that leads to control
and eventually to improvement.“

- James Harrington

Data Analytics

- Nascar / Formula One
- Sports
- Insurances
- Sales / Marketing
- Online Advertising
- Logistics



in Public Health we have

Disease Surveillance

Surveillance Systems

- Vital Statistics & Registries (e.g., births, deaths, defects)
- Population Surveys (e.g., substance abuse)
- Disease Reporting (e.g., salmonellosis, measles)
- Sentinel Surveillance (e.g., Influenza-Like Illnesses)
- Adverse Events Surveillance (e.g., issues with drugs)
- Laboratory Data

surveillance data

should be a byproduct

of any healthcare operation

Syndromic Surveillance

- Focuses on Early Detection
- Based on disease signs or symptoms, not diagnosis
- Novel sources: Emergency Room data, Drugs sales
- Uses well known Data Mining techniques
- Reduced delay in results

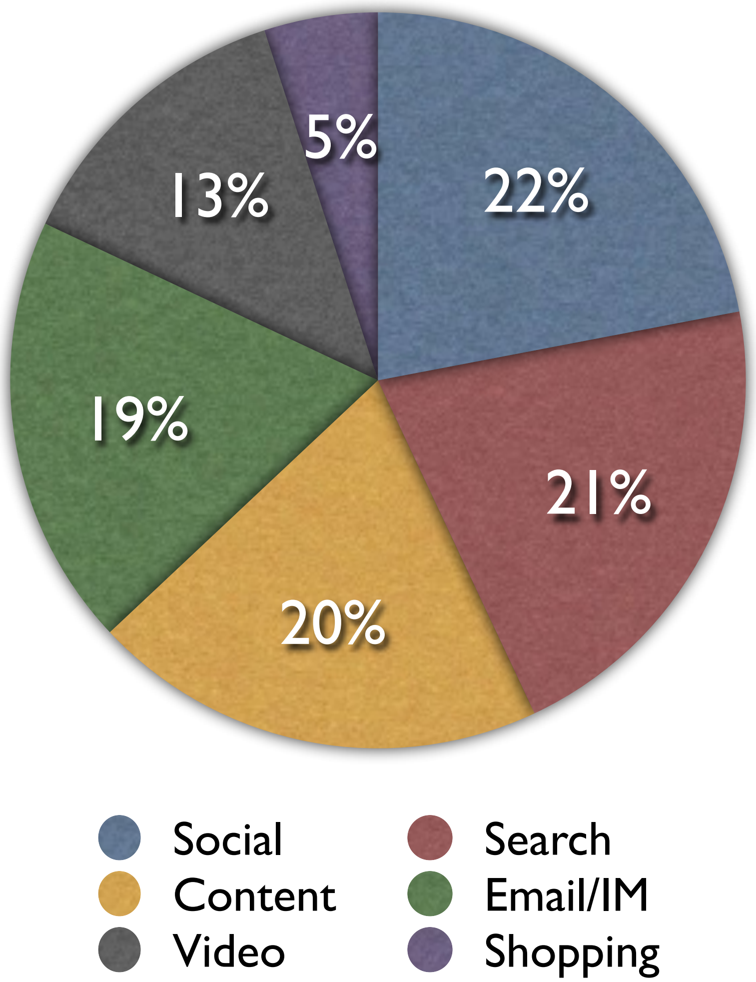
aggregate and analyze

Social Media Data

to monitor and predict health trends

online
~27h/mon

mobile
~34h/mon

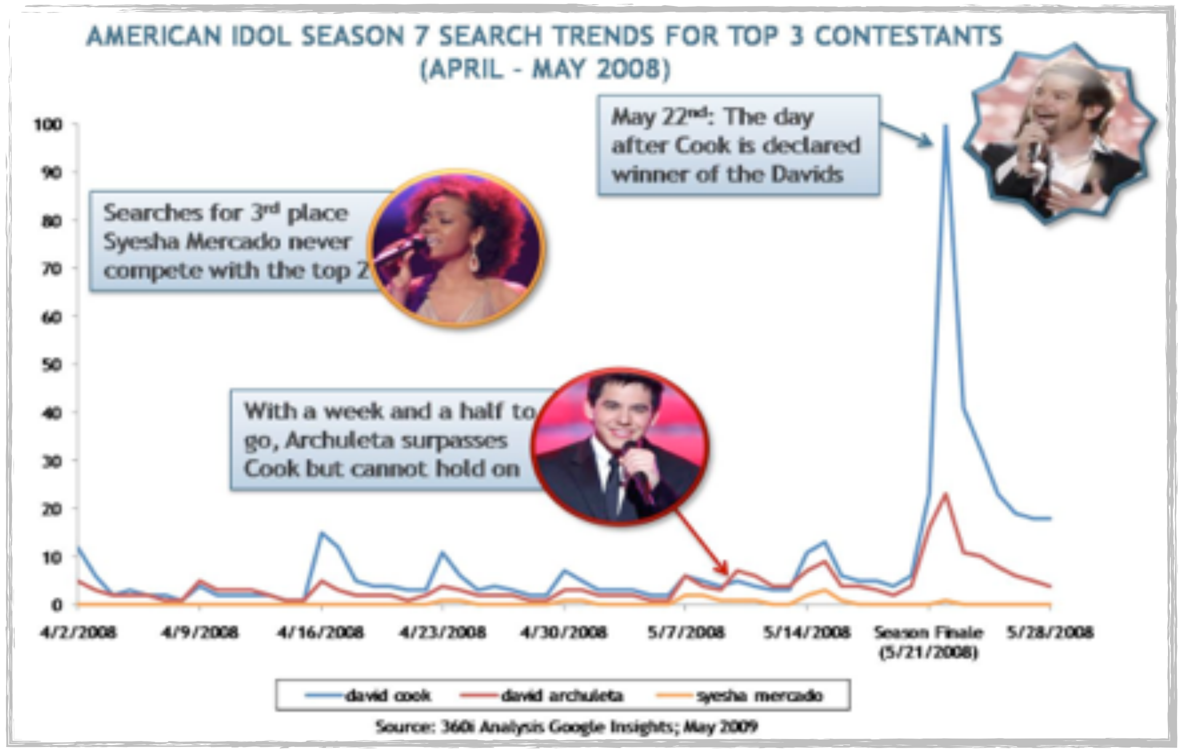


facebook
~5B/day

twitter
~500M/day

foursquare
~7M/day

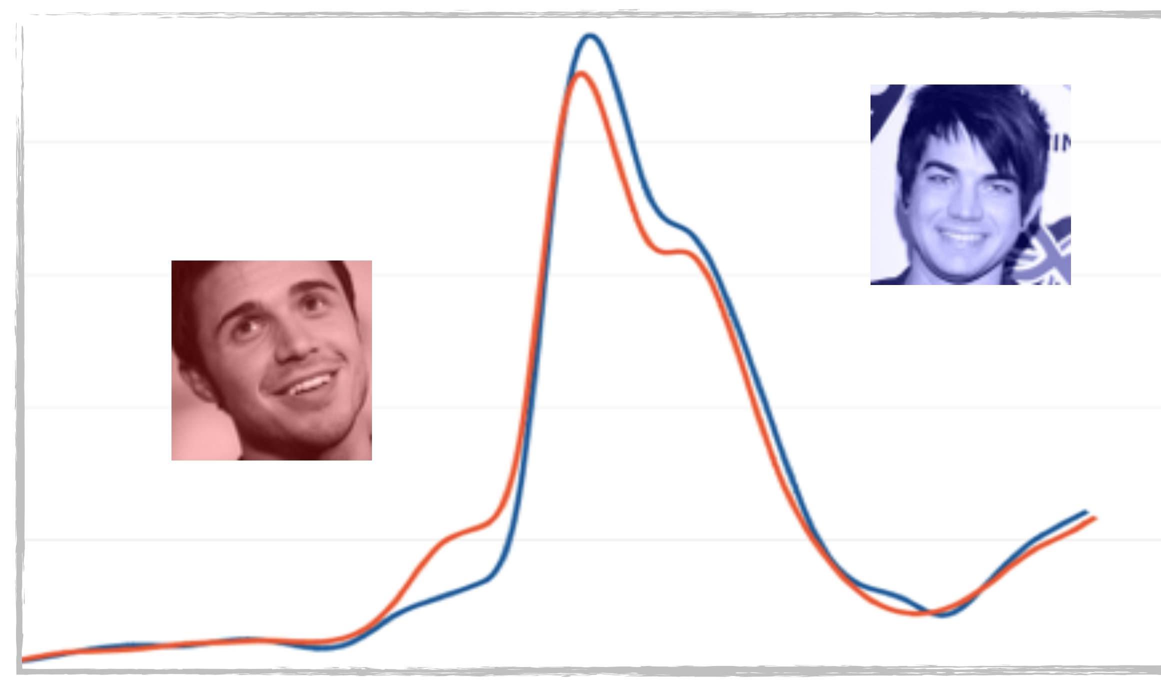
WordPress
~10M/day



Google Searches

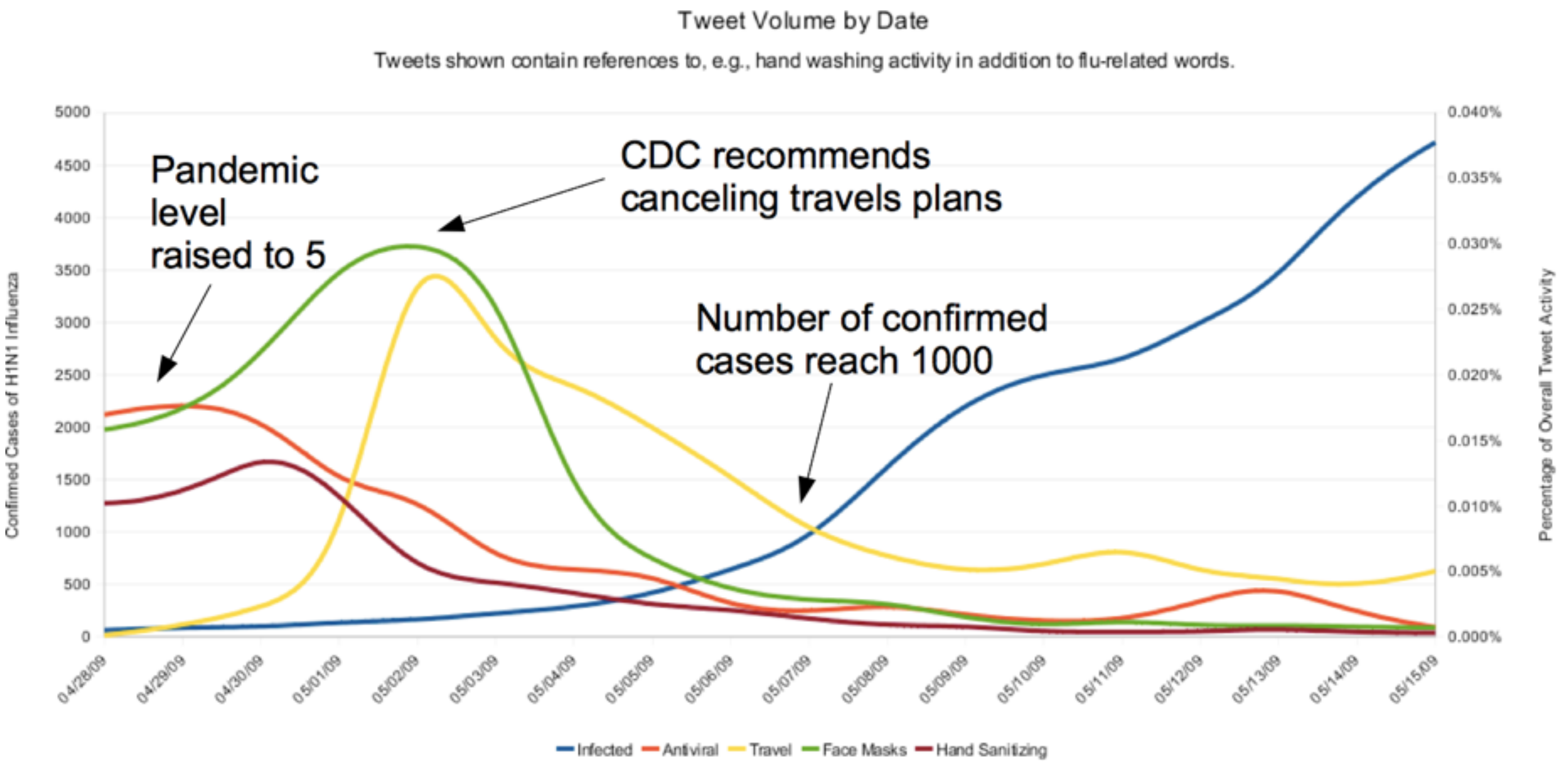
Google vs. twitter

Monitor
Public Opinion

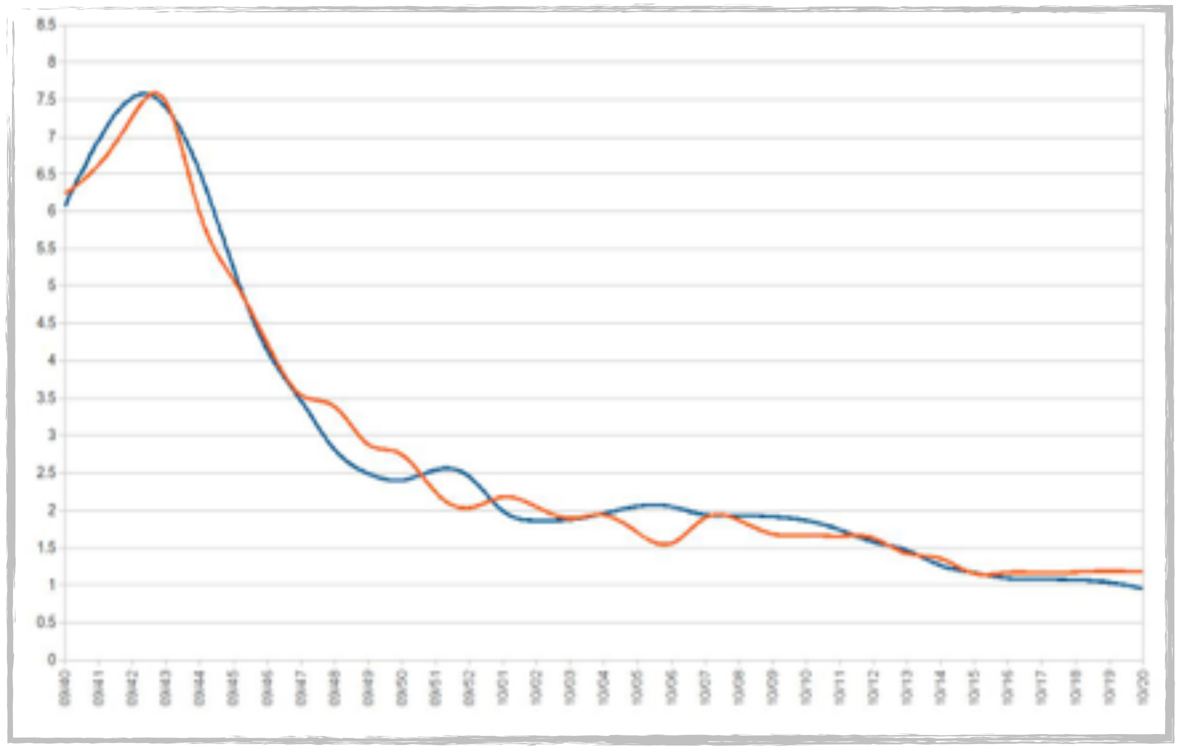


Positive Tweets

Comprehensive Exam
Alessio Signorini
University of Iowa, May 2010

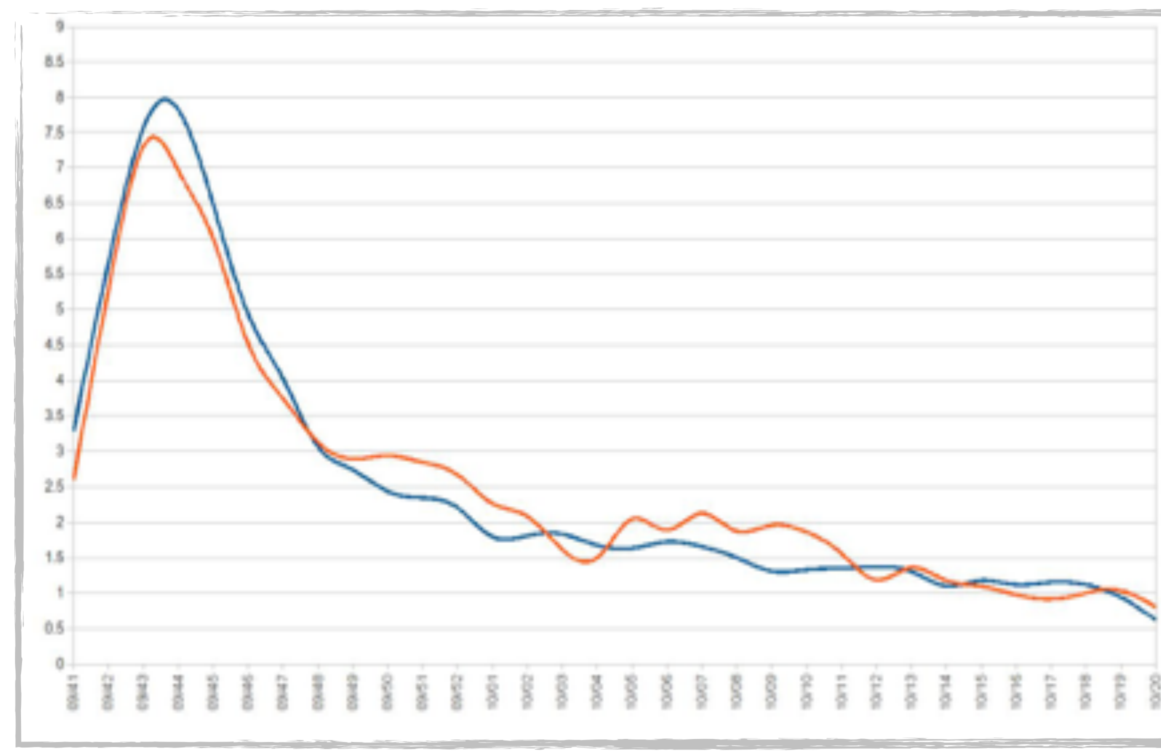


The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic
Alessio Signorini, Alberto Segre, Philip Polgreen
PLoS ONE – Journal, May 2011



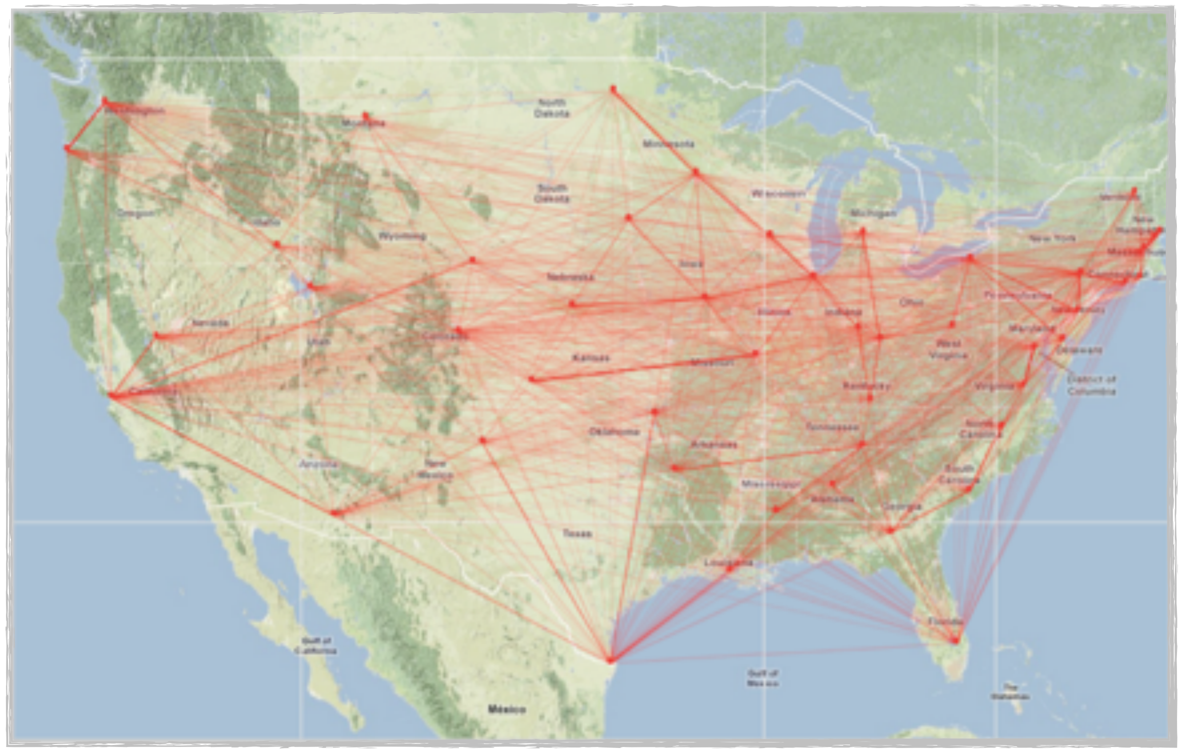
error ~0.28%

Estimate ILI%



error ~0.37%

Using Twitter to Estimate H1N1 Activity
Alessio Signorini, Alberto Segre, Philip Polgreen
ISDS 2010 – 9th Annual Conference of International Society for Disease Surveillance



National

foursquare

Monitor Travels

Local

Inferring Travel from Social Media

Alessio Signorini, Alberto Segre, Philip Polgreen

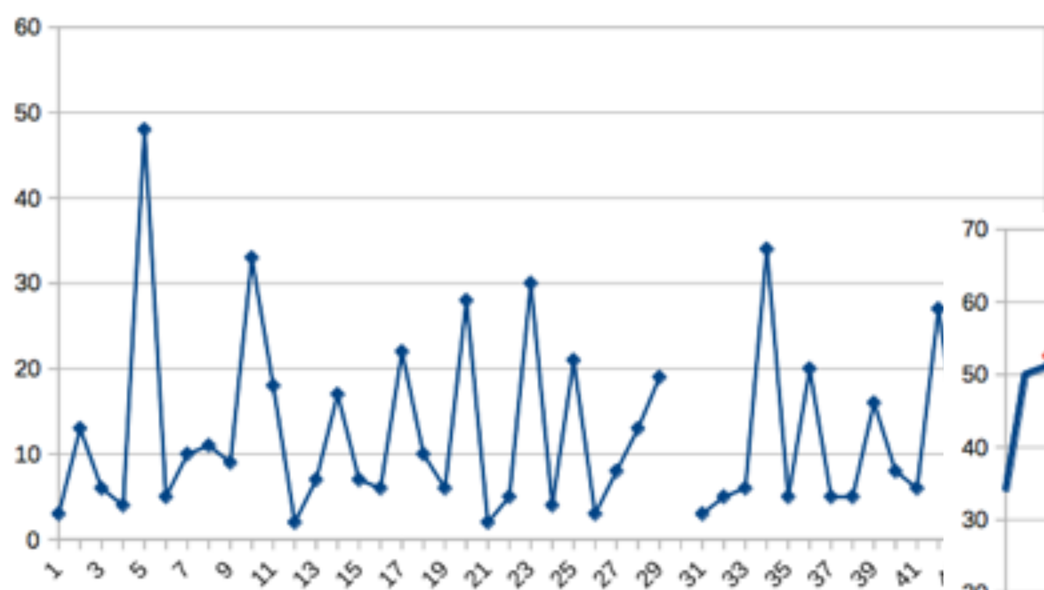
ISDS 2011 – 10th Annual Conference of International Society for Disease Surveillance can we use

“Social Travel Models”

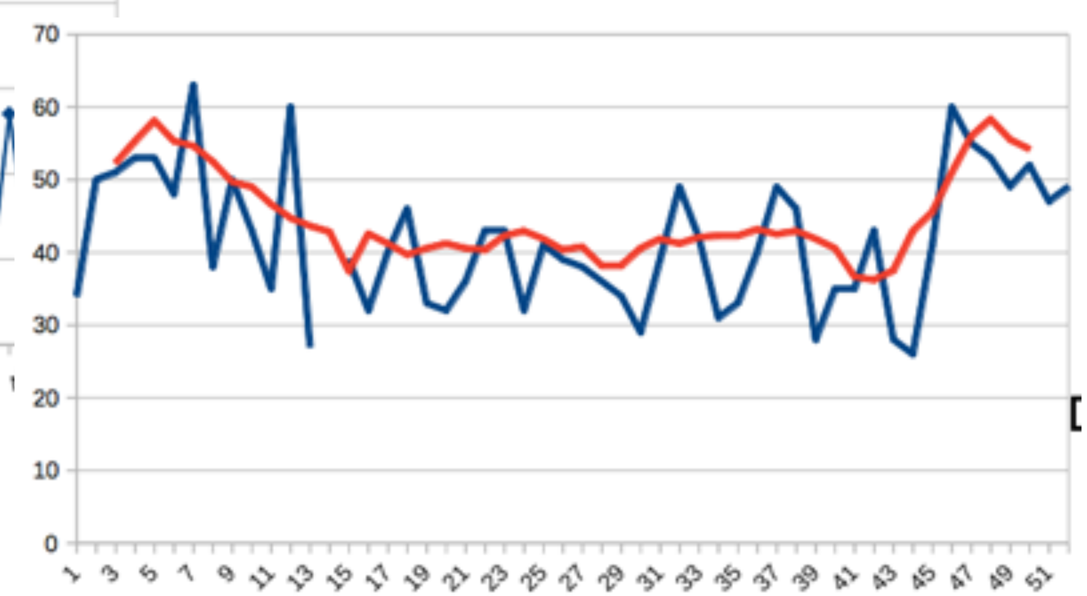
to improve local flu trends prediction?

City-Level Flu Trends

- CDC's MMWR - Flu & Pneumonia Deaths for 122 cities
- Smoothed each week with values of prev/next 2 weeks



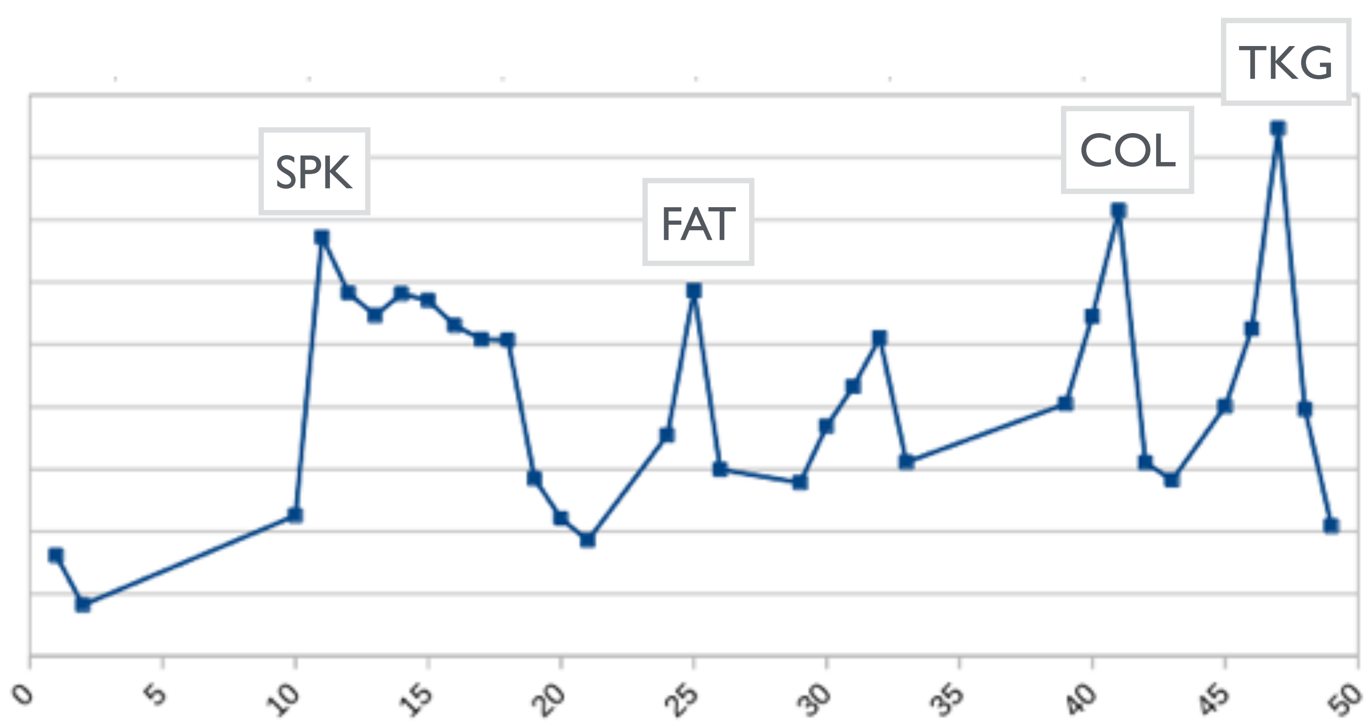
Philadelphia, PA - Deaths for 2012



New York City, NY - Deaths for 2012

Social Travel Data

- 240 Million geolocated tweets posted by 4 Million users
- Mapped over MMWR cities, discarded overlapping ones
- Used Spark cluster of 8 machines to do geo-mapping



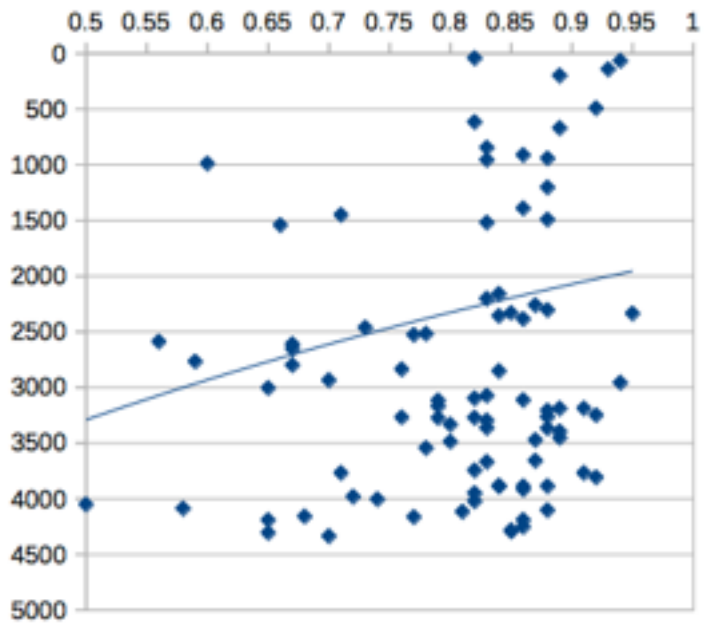
Volume of Trips among MMWR cities 2012

Social Travel Model

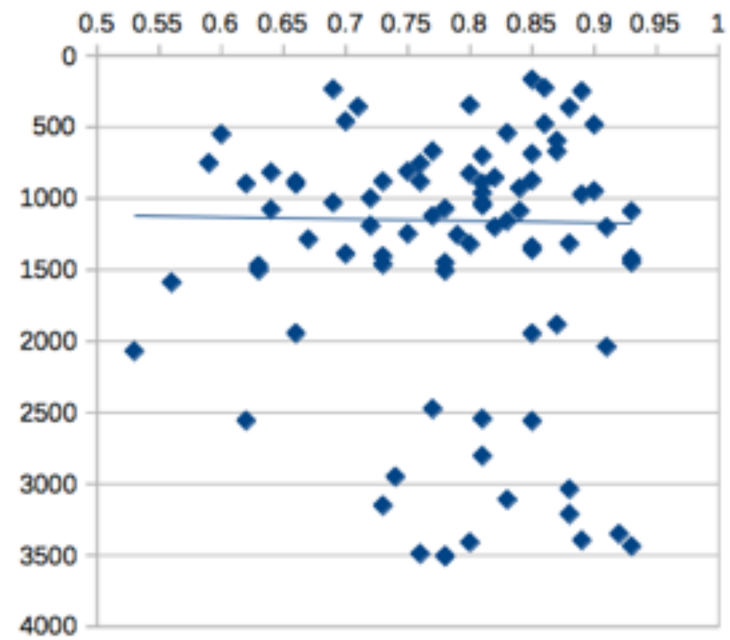
- Final dataset: 78 cities, 124M tweets, 2.2M users
- Assumed "home" the most common location
- A "trip" was a post at home followed by one elsewhere
- Used population to scale volume of trips between cities

City	Population	Home for	%
New York City, NY	8406000	297486	3.54%
Los Angeles, CA	3857799	206378	5.35%
Chicago, IL	2714856	94952	3.50%
Houston, TX	2160821	71780	3.32%
Philadelphia, PA	1547607	77626	5.02%

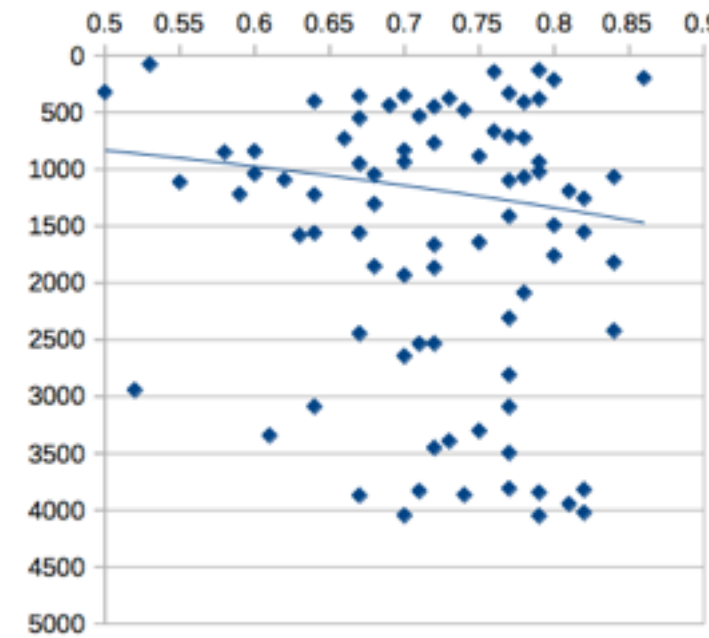
Correlation b/w Cities



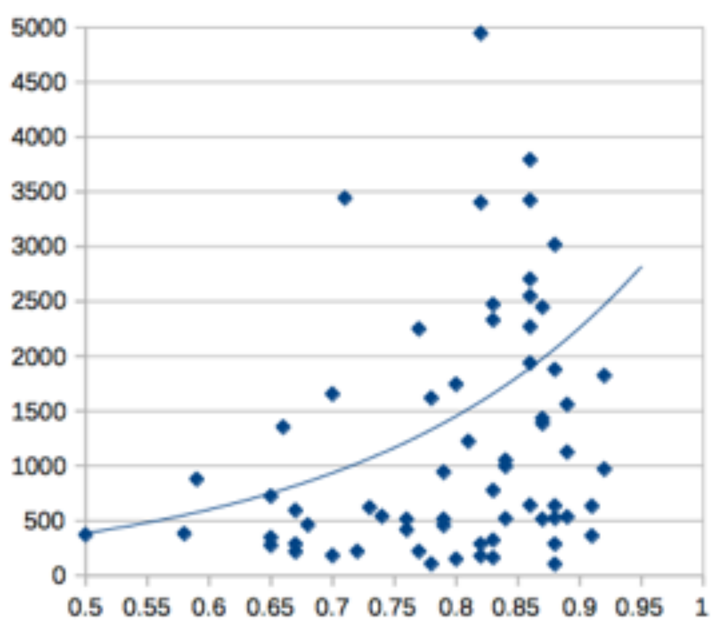
San Jose, CA

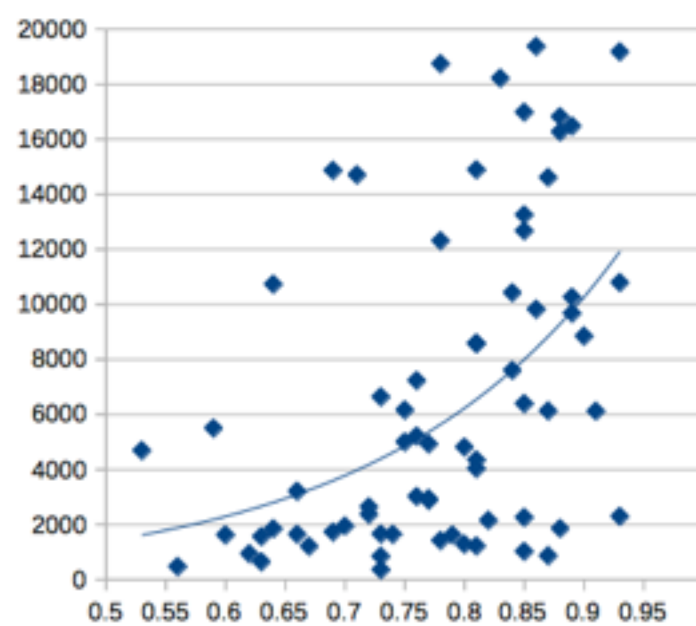


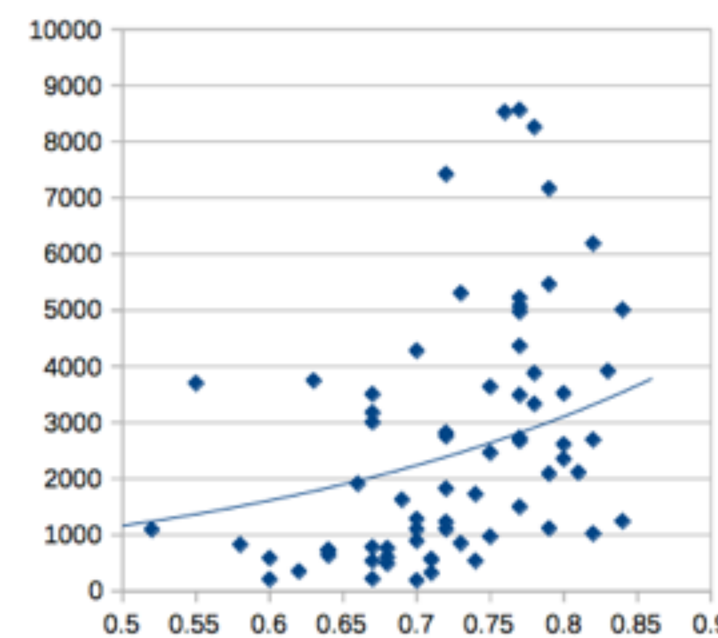
Atlanta, GA



Philadelphia, PA







Predicting Flu Trends

- Flu Trends of 78 cities generated from MMWR data
- Used 2011 for training and 2012 for testing
- Support Vector Regression with polynomial kernel
- Target: value of local flu trend for that week
- Features: value of top 20 correlated cities 2 weeks before

Measures Compared

- Distance closest 20 cities
- Similarity most similar 20 cities on 2011 flu trends
- Flow top 20 cities by number of visitors

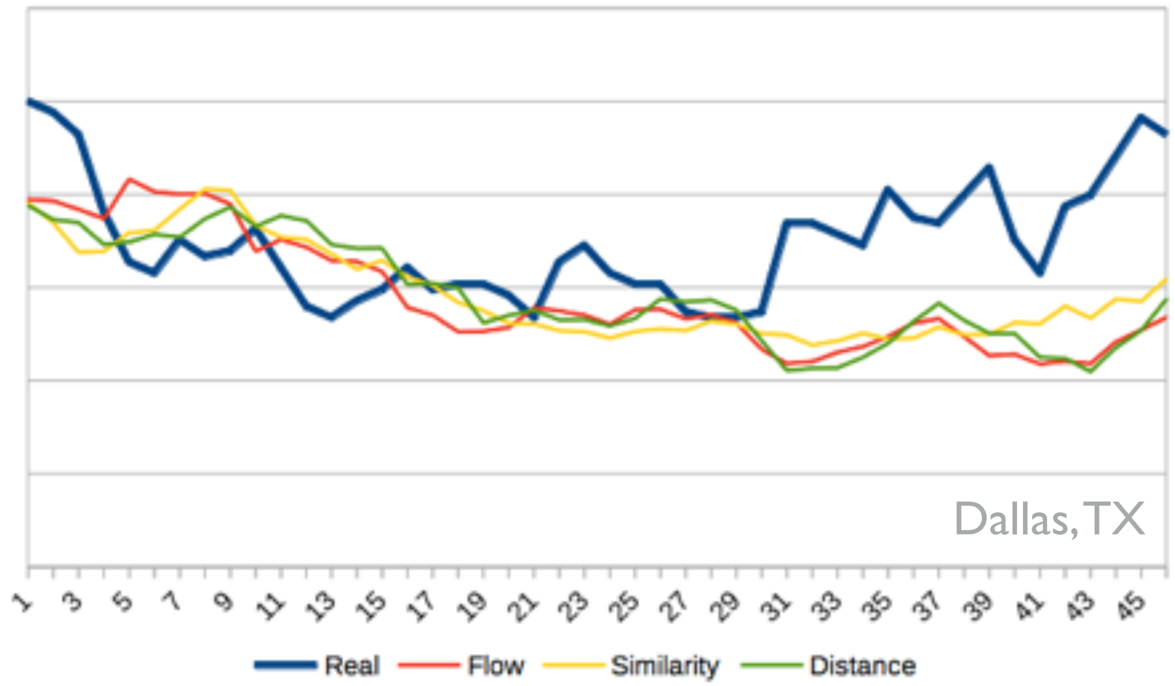
	Distance	**Similarity**	**Flow**	**Combined**
Average	0.23086	0.26971	0.28080	0.28543
Minimum	0.00009	0.00004	0.00027	0.00001
Maximum	0.7780	0.73145	0.83247	0.79387

Table 3.4: Square Correlation Coefficients for each approach

Prediction Results

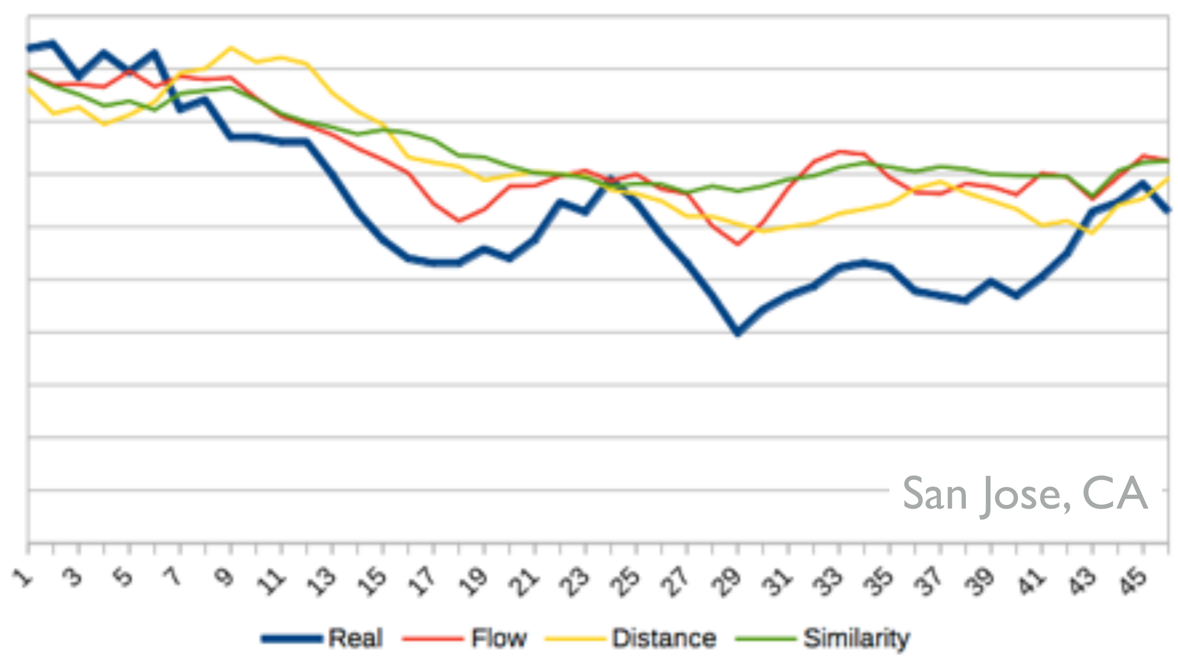
	Distance	Similarity	Flow	Combined
Washington, DC	0.043178	0.066362	0.013697	0.001167
Philadelphia, PA	0.000334	0.057870	0.007827	0.002519
Dallas, TX	0.000088	0.024712	0.005816	0.001576
Waterbury, CT	0.061919	0.004790	0.002763	0.000006
Fort Wayne, IN	0.005689	0.000038	0.000272	0.000518

Table 3.6: Square Correlation Coefficients for most difficult cities



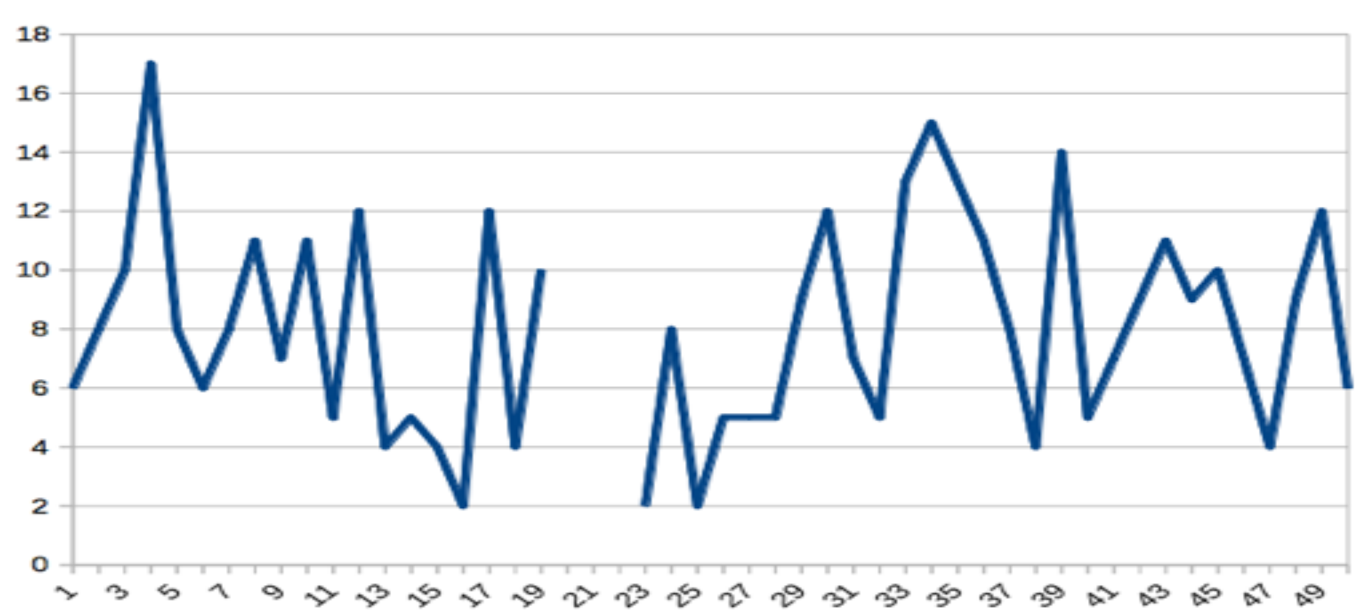
	Distance	Similarity	Flow	Combined
Las Vegas, NV	0.757804	0.562152	0.83247	0.735552
San Jose, CA	0.602464	0.689917	0.801097	0.793866
Albuquerque, NM	0.556896	0.626601	0.711688	0.494305
Tucson, AZ	0.740066	0.5472	0.694016	0.661322
San Antonio, TX	0.467228	0.636067	0.686346	0.612461

Table 3.5: Square Correlation Coefficients for most predictable cities



Failure Hypothesis

- Port-of-entry influenced by international travels
- Noisy data Watebury, CT had only 43 deaths in 2011
- Few data Fort Wayne has 1/50th of Las Vegas' users

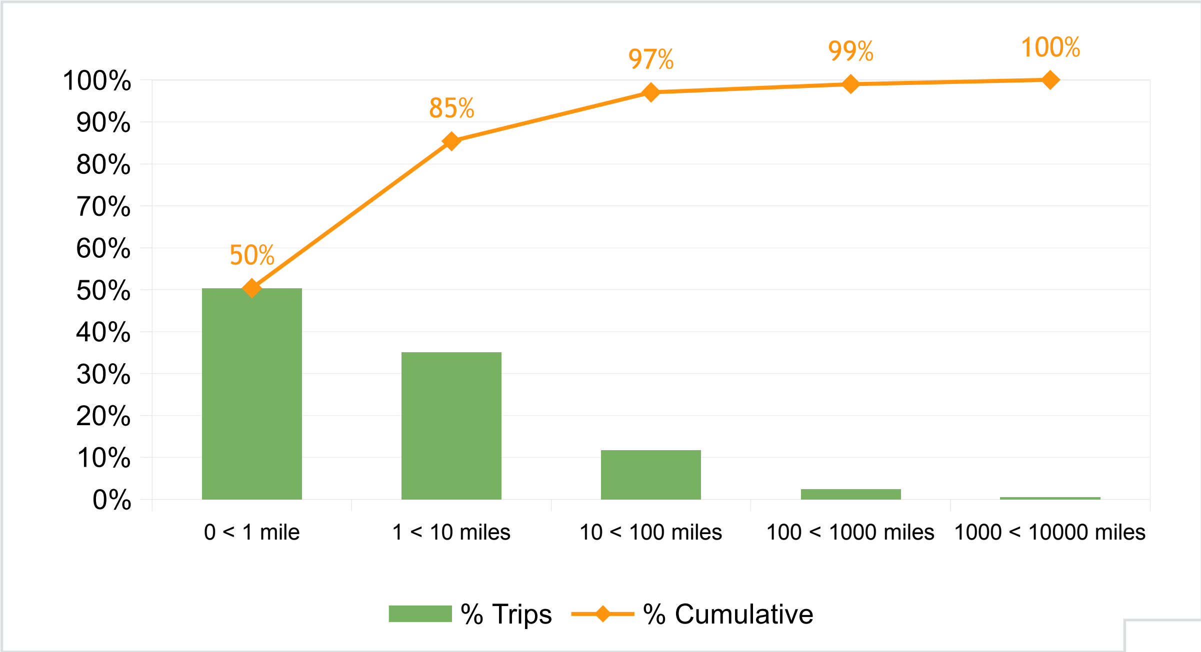


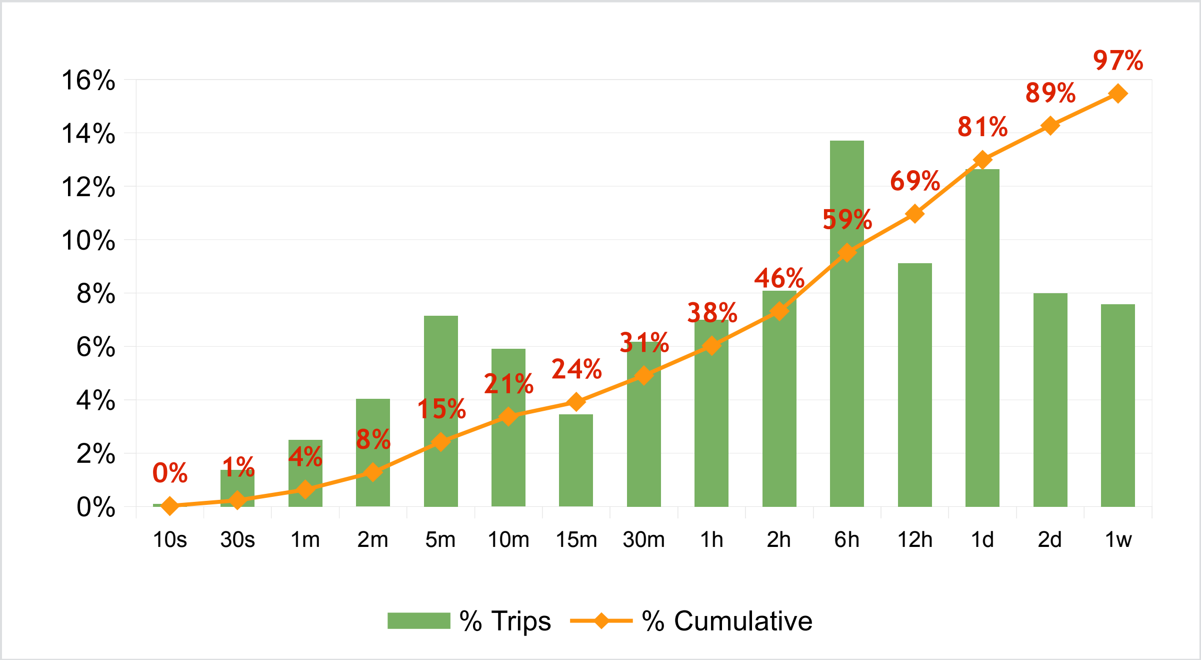
Washington, DC - Flu Deaths 2012

Conclusions

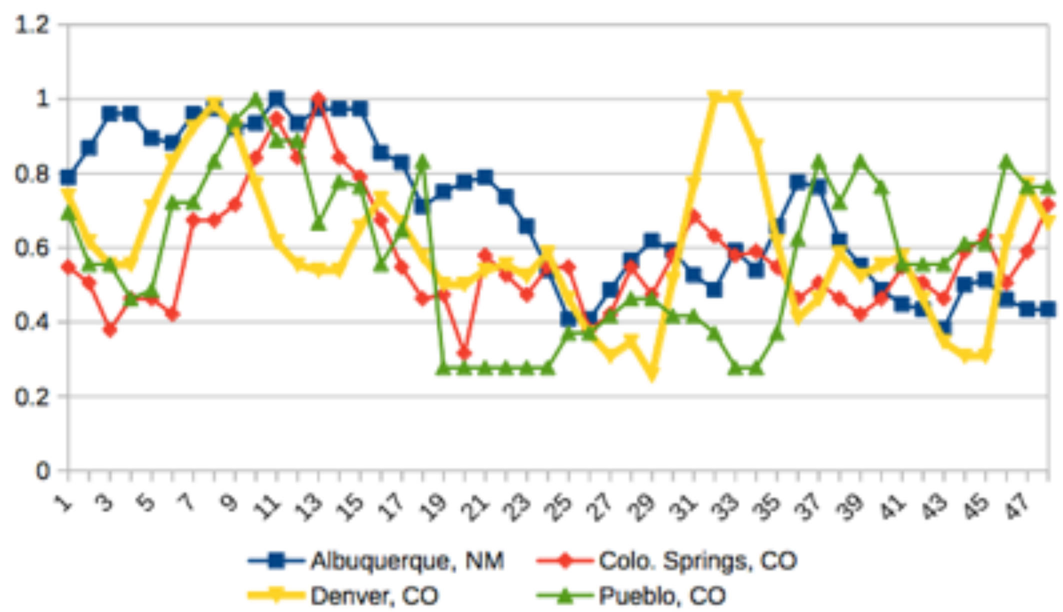
- Social Media can be an important source for surveillance
- Can predict American Idol's winner ;)
- Allows to monitor public sentiment about health topics
- Can effectively be used to monitor ILI% in real time
- Geolocated posts can be used to create travel models
- Social Travel Data provides additional predictive power for flu trends

Checkins Distributions

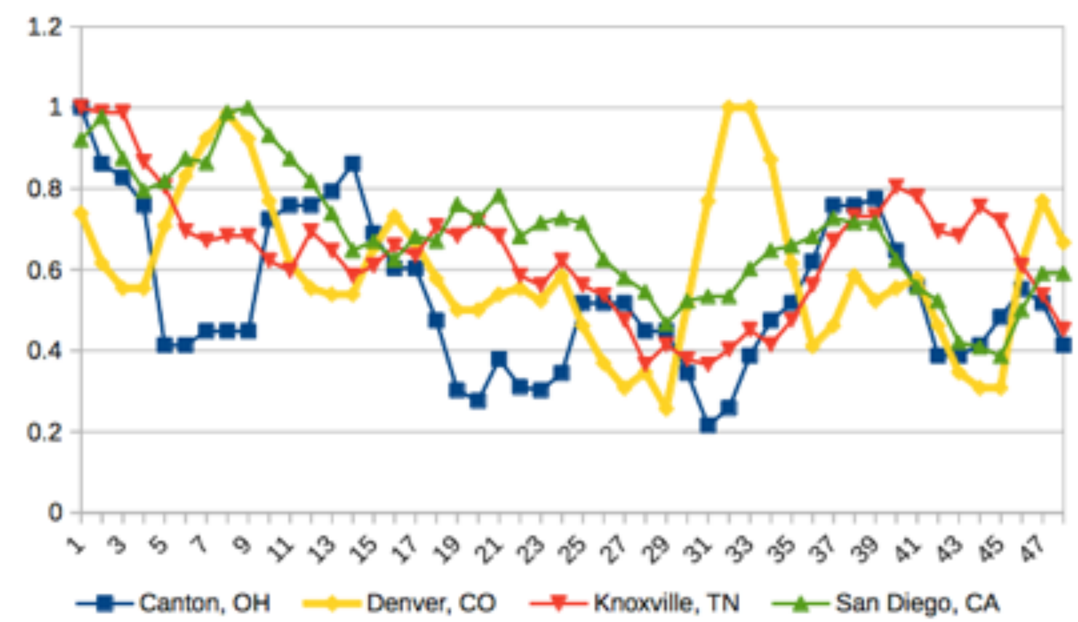




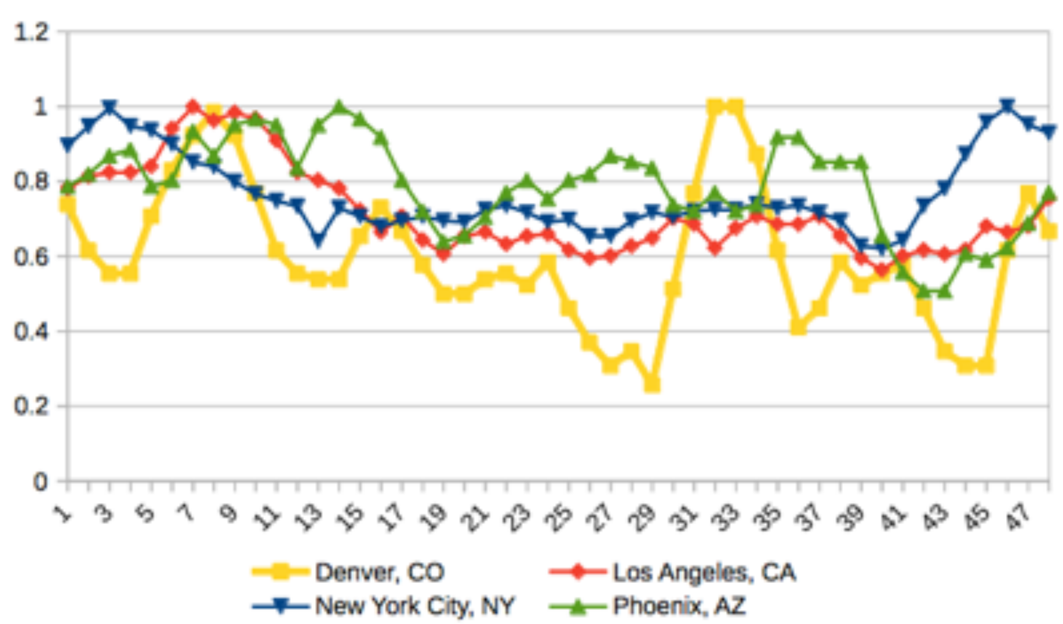
Denver, CO



Distance

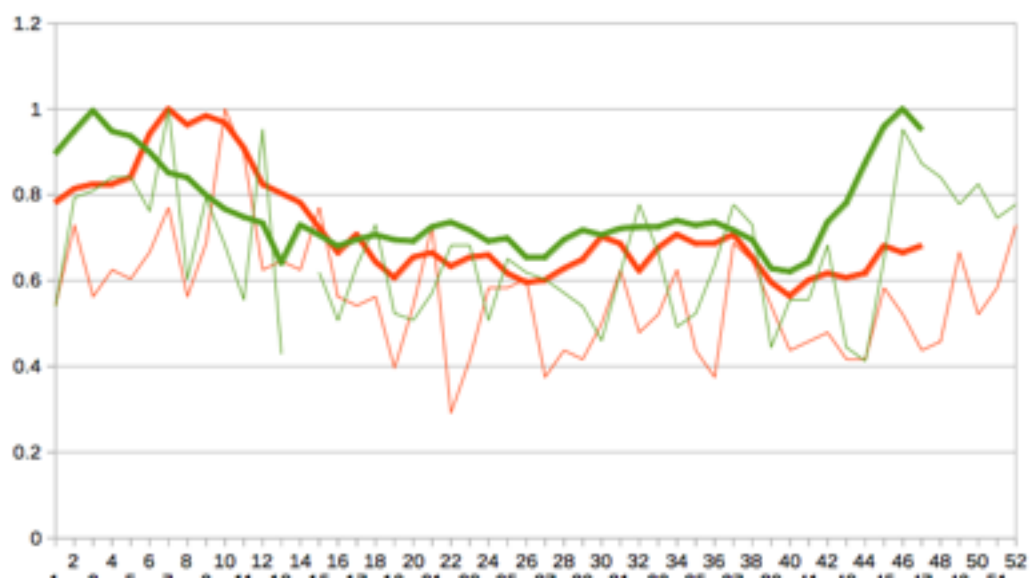


Similarity

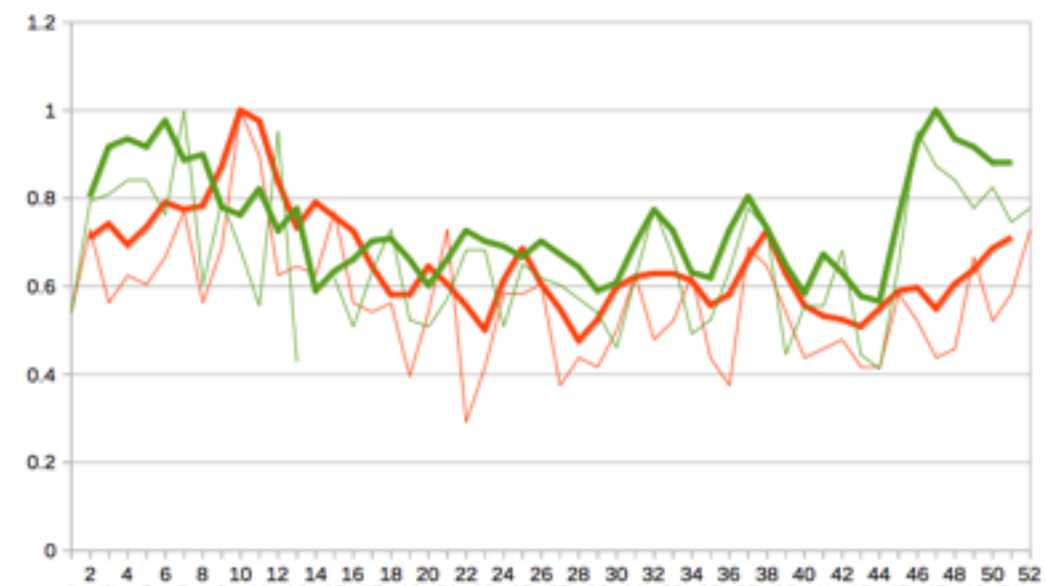


Flow

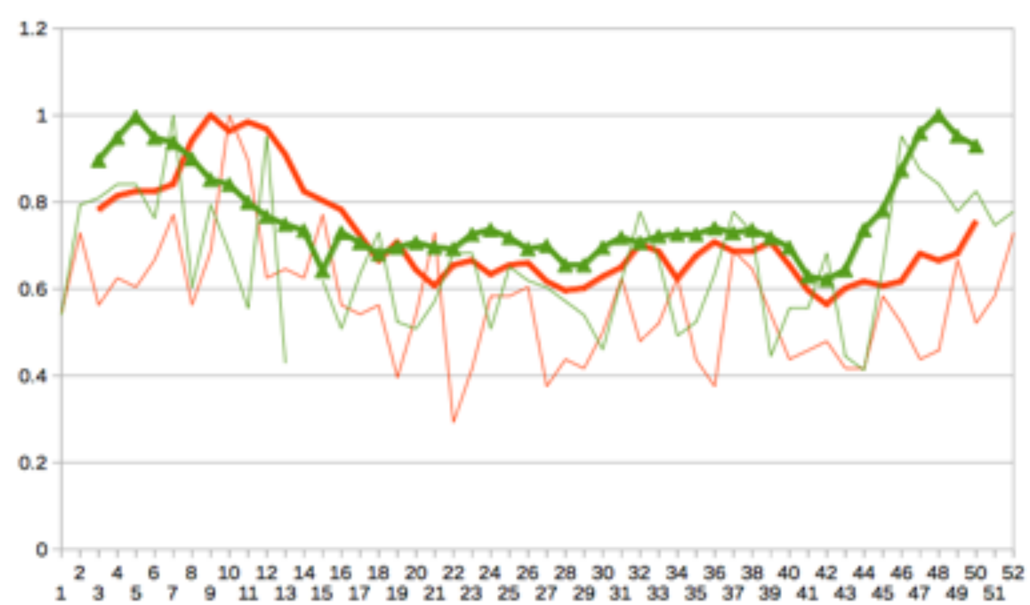
Smoothing Methods



5 weeks ahead



1 week around



2 weeks around