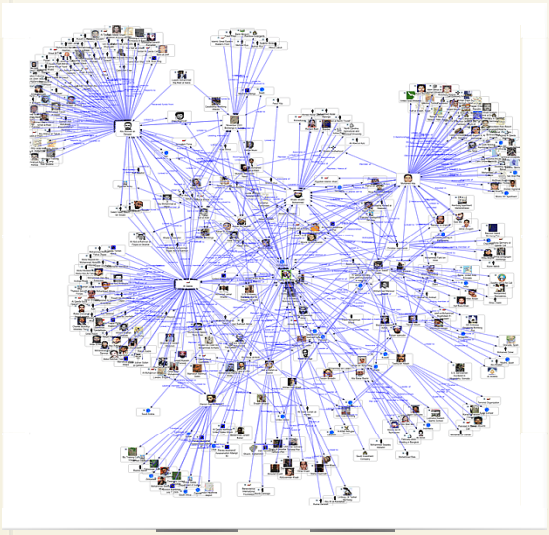
Predicting Flu Trends using Twitter Data

SNEFT – Social Network Enabled Flu Trends



- Harshavardhan Achrekar [1]
- Avinash Gandhe [2]
- Ross Lazarus [3]
- Ssu-Hsin Yu [2]
- Benyuan Liu [1]

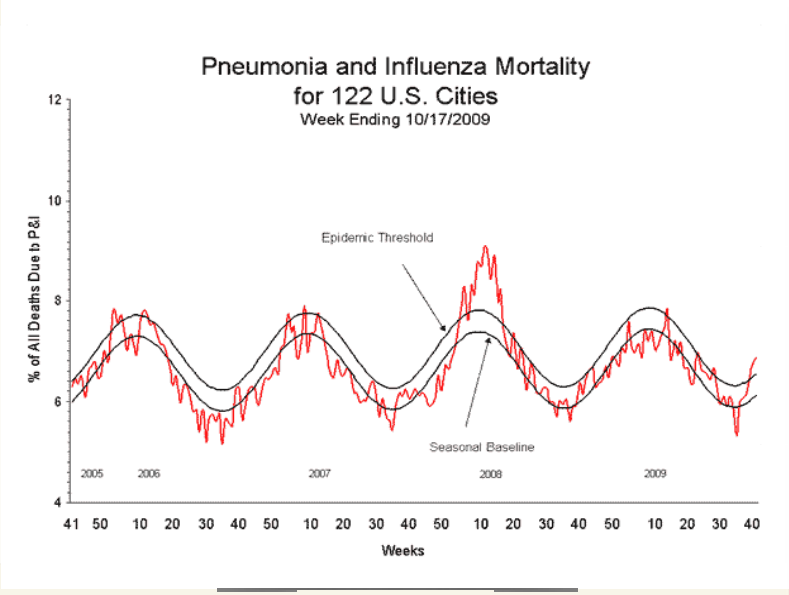
[1] Department of Computer Science, University of Massachusetts Lowell
[2] Scientific Systems Company Inc , Woburn, MA
[3] Department of Population Medicine - Harvard Medical School

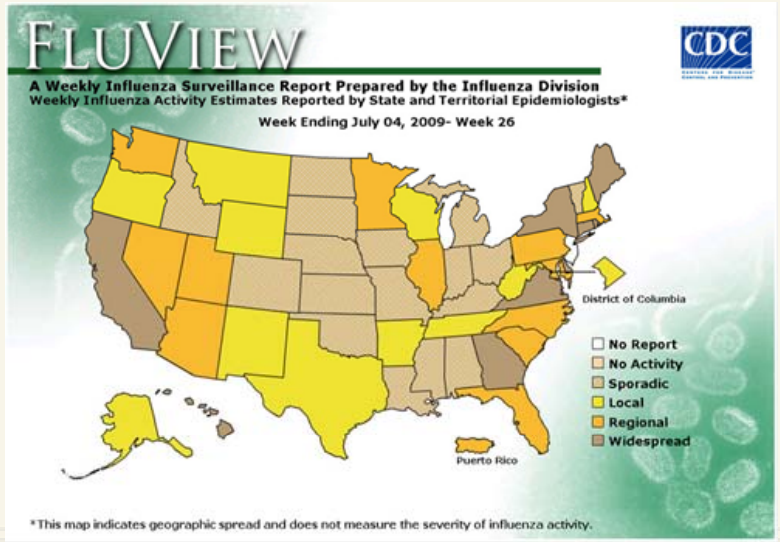
Outline

- Background
- Related Work
- Our Approach
- Twitter Dataset and Analysis
- Conclusion

Seasonal flu

- Influenza (flu) is contagious respiratory illness caused by influenza viruses.
- Seasonal - wave occurrence pattern.
- 5 to 20 % of population gets flu
- ~ 200,000 people hospitalized from flu related complications.
- ~ 30,000 people die from flu every year in USA, worldwide death toll 250,000 to 500,000.
- Epidemiologists use early detection of disease outbreak to reduce number of people affected
- CDC collects Influenza-like Illness (ILI) from its surveillance network and publishes weekly (usually 1-2 weeks delay)





Emerging Flu Epidemics

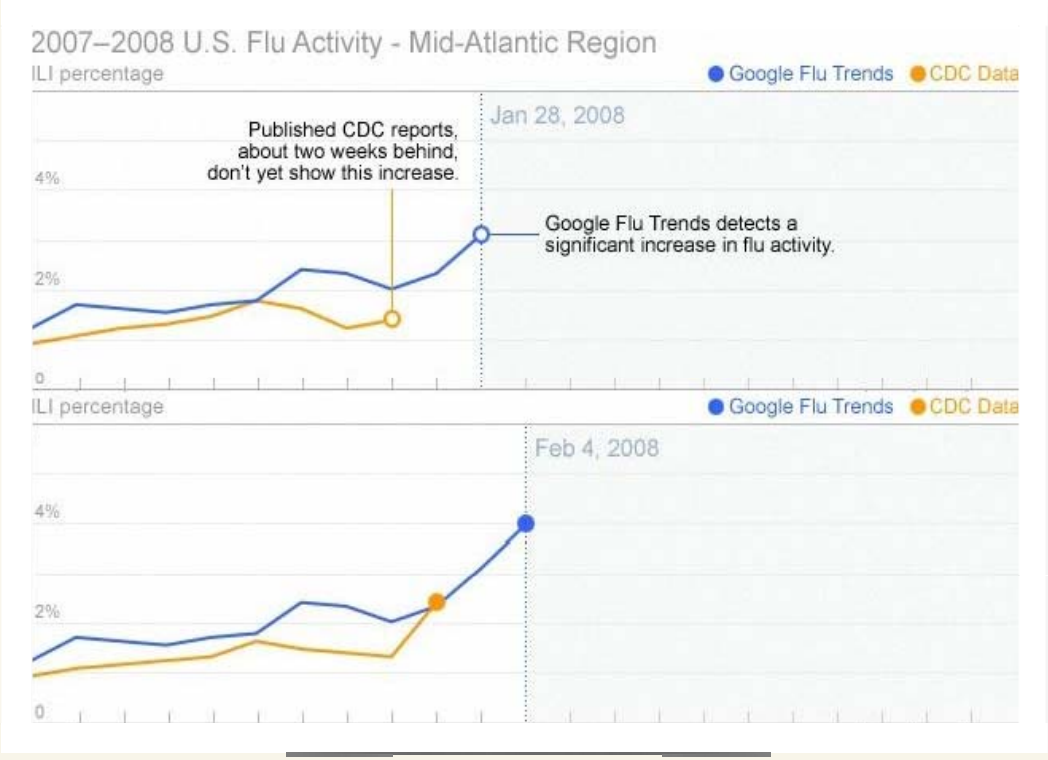
SARS, 2002-2003

H1N1 (swine), 2009-2010

?

Related Work :- Google Flu Trends

- Certain Web Search terms are good Indicators of flu activity.
- Google Trends uses aggregated search data on flu indicators.
- Estimate current flu activity around the world in real time.
- From example :- Google Flu Trend detects increased flu activity two weeks before CDC (Center for Disease Control and Prevention).



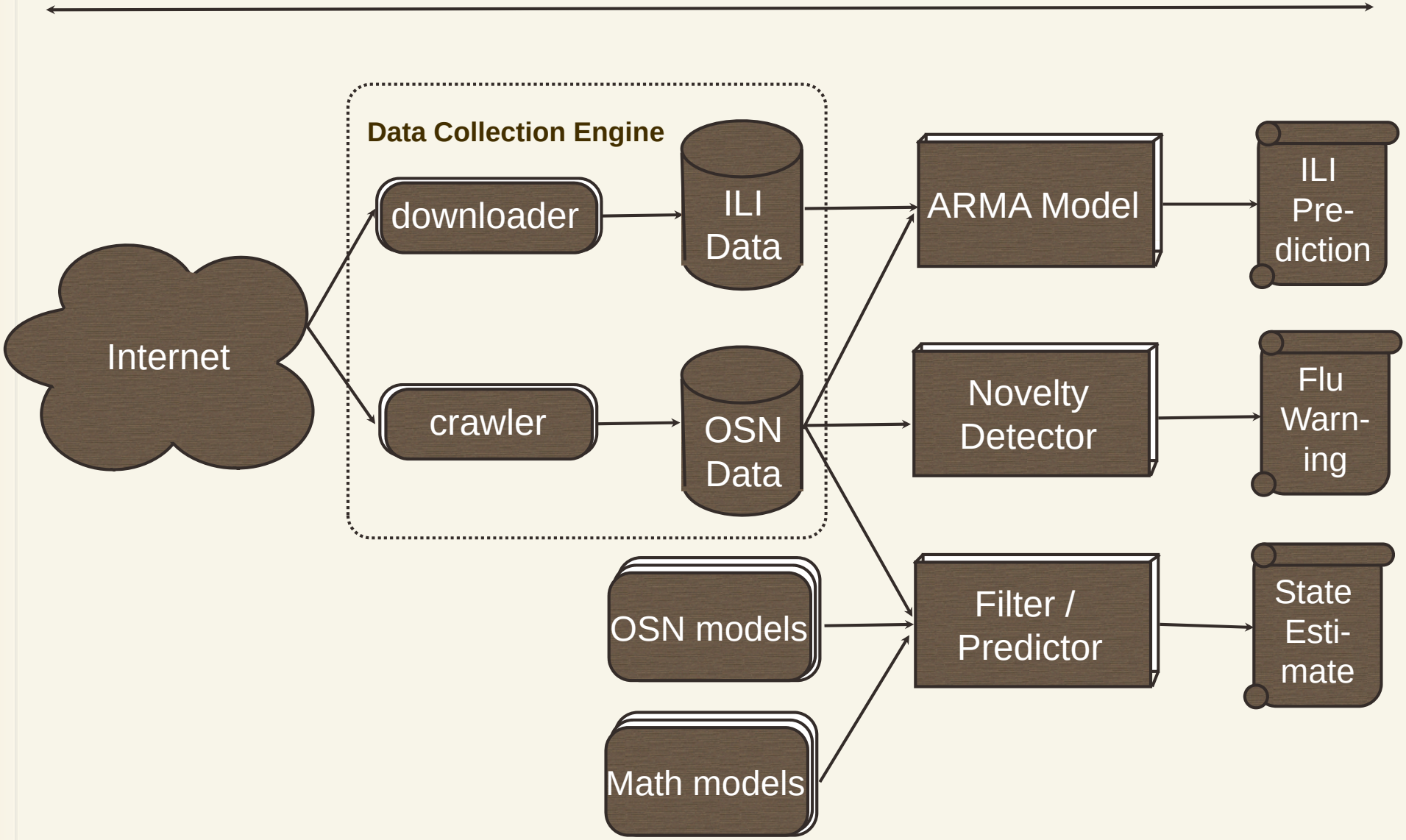
OSN as a Data Source for Detection and Prediction

- OSN emerged as popular platform for people to make connections, share information and interact
- Facebook: ~ 750 million users, Twitter: ~200 million users
- billions of pieces of information being posted and shared on the web every week
- Applications:
  - Real-world outcome of box-office revenues for a movie
  - Earth-quake detection and reporting
  - Online service downtime and disruptions
  - people's mood

Our Approach

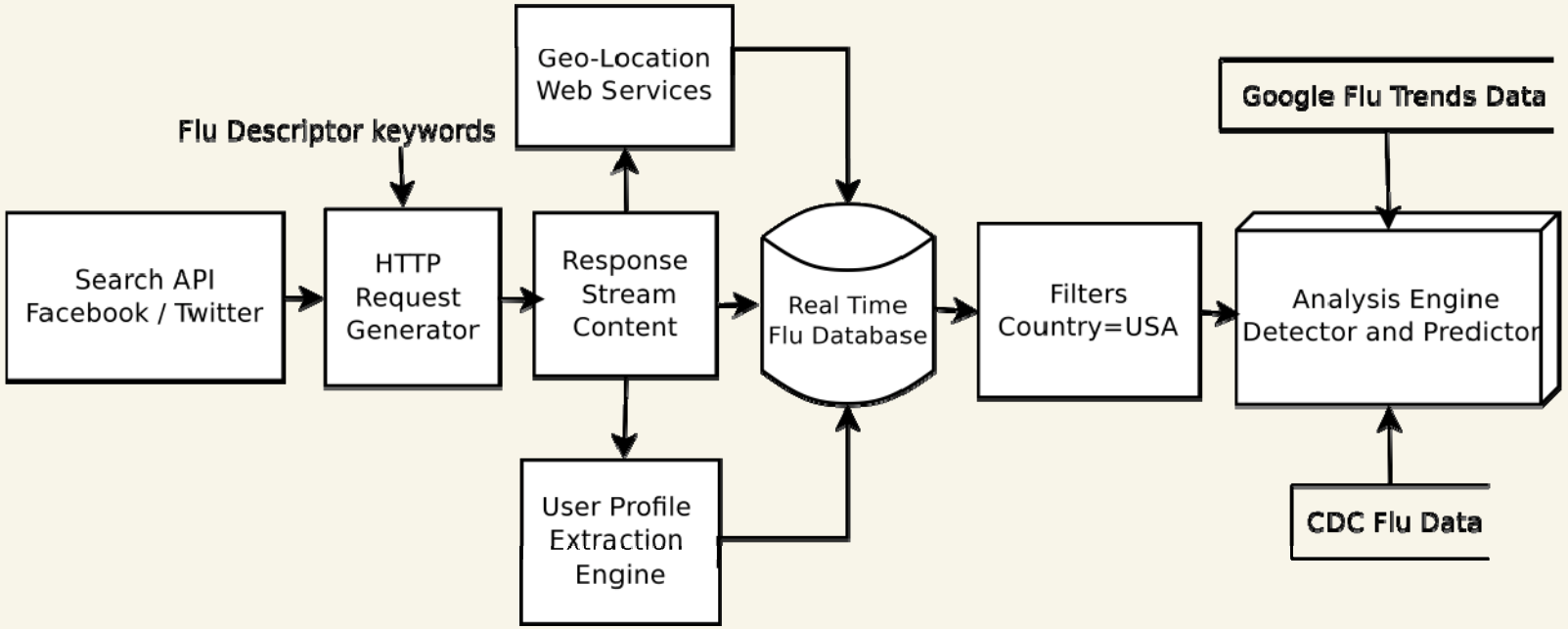
- OSN represent a previously untapped data source for detecting onset of an epidemic and predicting its spread.
- {“i am down with flu”, “got flu.”} msg exchanged between users provide early predictions.
- Twitter/Facebook mobile users tweet/posts updates with their geo-location updates. helps in carrying out refined analysis.
- User demographics like age, gender, location, etc can be obtained or inferred from data.
- Provide snapshot of current epidemic condition and preview on what to expect next on daily or hourly bases.

System Architecture of SNEFT



ILI stands for Influenza-Like Illness

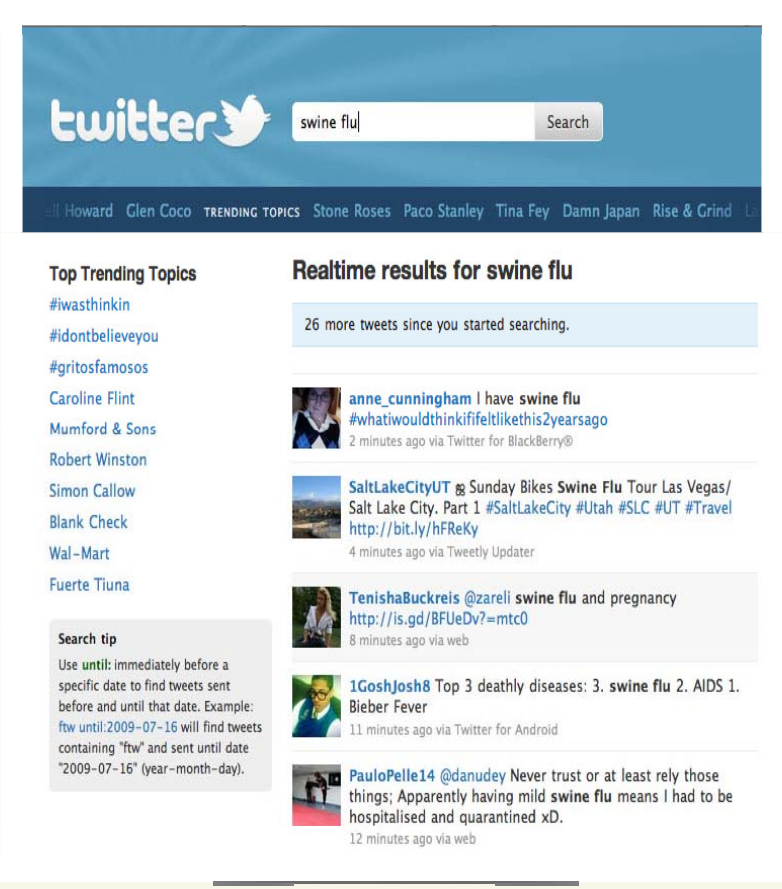
OSN Data Collection



OSN data collection engine / Crawler

Twitter Data Set

- Real Time Response Stream fetches entries relevant to searched keyword having the tweets in reverse-time order.
- Data collection active from October 18, 2009 until present.
  - 2009 – 2010: 4.7 million tweets from 1.5 million unique users
  - 2010 – 2011: 4.5 million tweets from 1.9 unique users

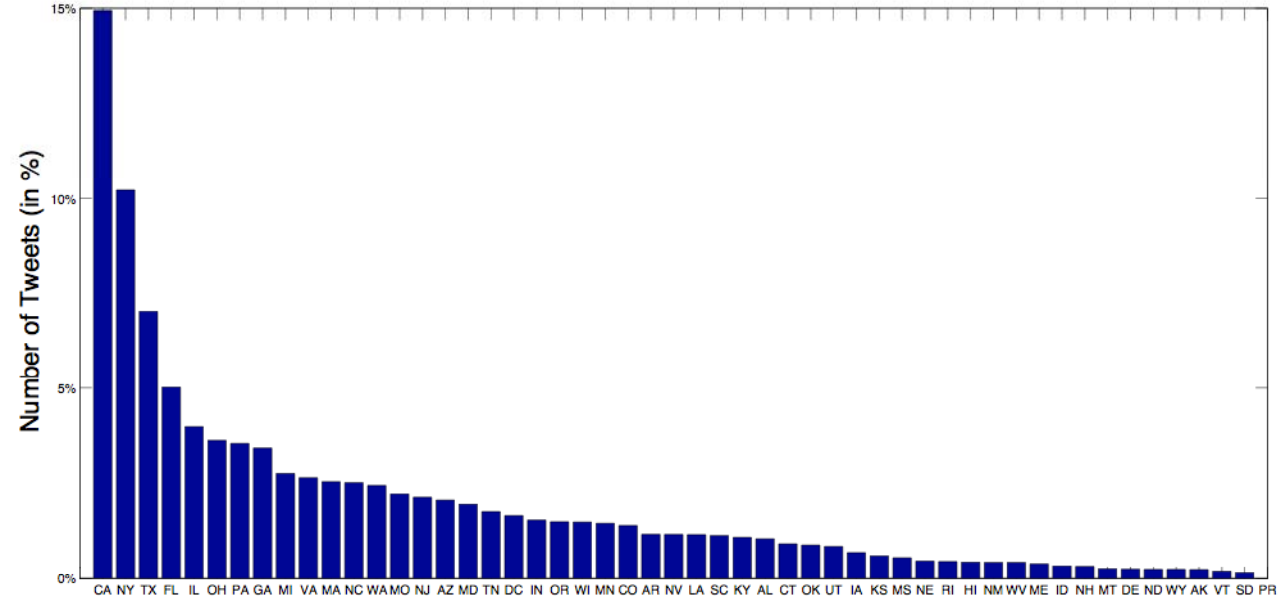


Spatio Temporal Database for Twitter Data Set

- Crawler uses Streaming Real time Search Application Programming Interface (API) to fetch data at regular time intervals. A tweet has the
  - Twitter User Name,
  - the Post with status id
  - Time stamp attached with each post.
- From Twitter's username we can get profile details attached to every user which include
  - number of followers,
  - number of friends,
  - his/her profile creation date,
  - location {public or private from the profile page or mobile client}
  - status updates count
- User's current location is passed as an input to Google's location based web services to get geo-location codes (i.e., latitude and longitude) along with the country, state, city with a certain accuracy scale.

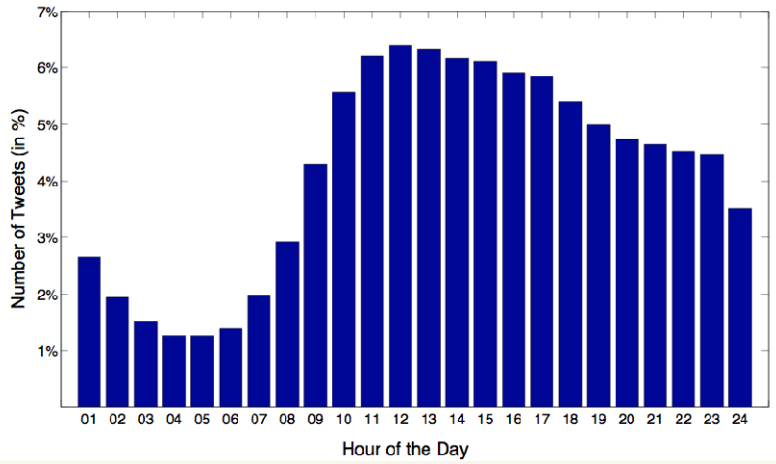
Twitter Data Set

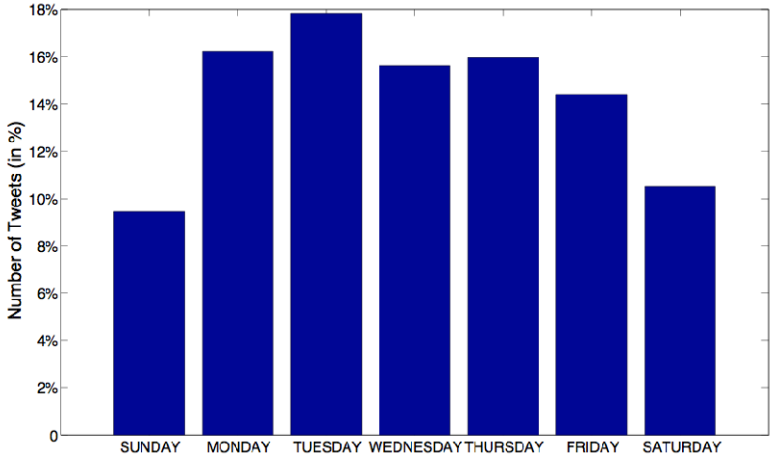
- 22 % users are from USA,
- 46 % users are outside USA
- 32% users have not published their location details.



State-wise Distribution of USA users on Twitter for flu postings

Twitter Data Set





Hourly Twitter usage pattern in USA

Average daily Twitter usage within a week

- The hourly activity patterns observed at different hours of the day are much to our expectations, with high traffic volumes being witnessed from late morning to early afternoon and less tweet posted from midnight to early morning, reflecting people's work and rest hours within a day.
- Average daily usage pattern within a week suggests a trend on OSN sites with more people discussing about flu on weekdays than on weekends.

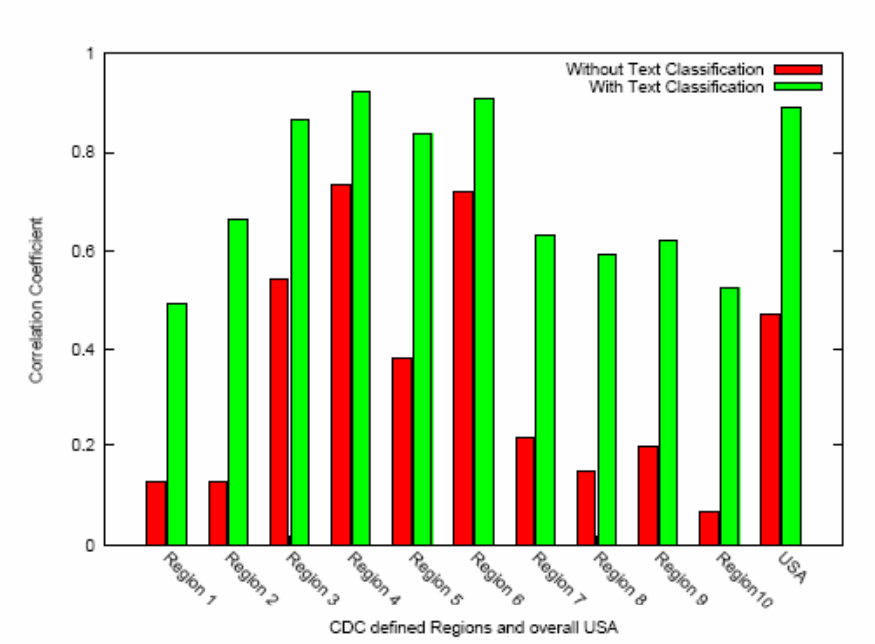
Twitter Data Set Cleaning

- Retweets: A retweet is a post originally made by one user that is forwarded by another user.
- Syndrome elapsed time: An individual patient may have multiple encounters associated with a single episode of illness . To avoid duplication the first encounter for each patient within any single syndrome group is reported to CDC, but subsequent encounters with the same syndrome are not reported as new episodes until more than six weeks has elapsed since the most recent encounter in the same syndrome. We call it syndrome elapsed time.
- Remove retweets and tweets from the same user within a certain syndrome elapsed time, since they do not indicate new ILI cases.

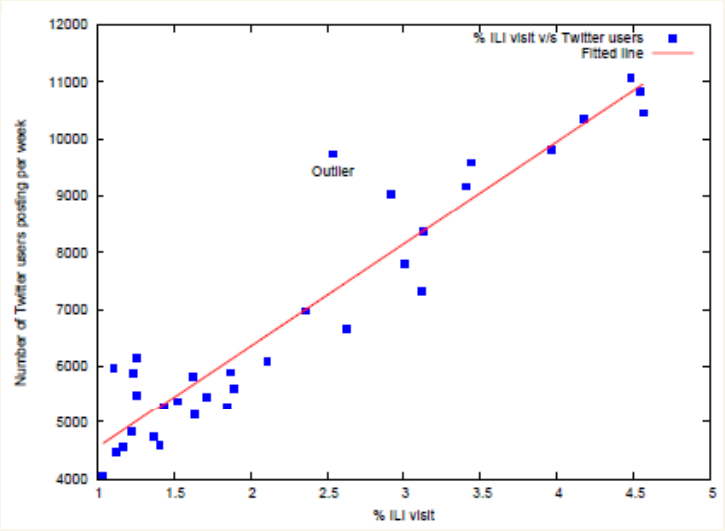
Twitter Data Set Cleaning

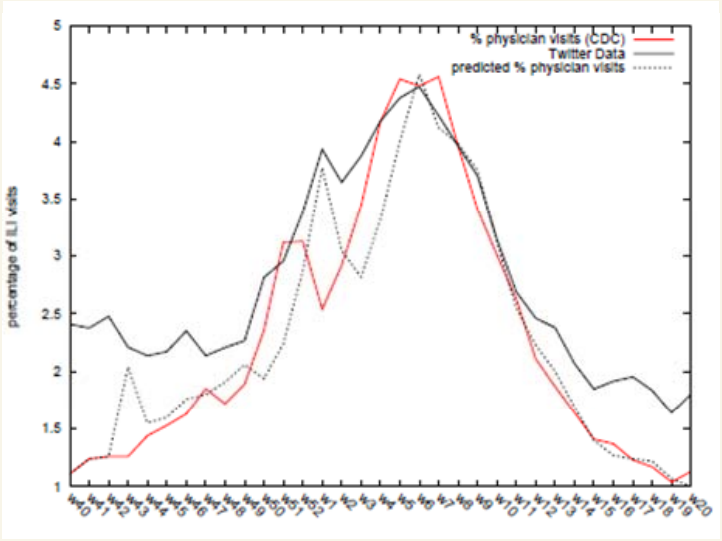
- “I got flu shot”, “got stomach flu”, “flu season” … do not indicate real flu events
- Need to classify the tweets into positive or negative categories
- 10,000 tweets classified by Amazon Mechanical Turk as training set
- Use trained SVM to classify all other tweets
- Significant improvement of correlation between the Twitter data and CDC data.

Classifier	Class	Precision	Recall	F-value
J48 decision tree	Yes	0.801	0.791	0.796
	No	0.813	0.704	0.755
Naïve Bayesian	Yes	0.725	0.829	0.773
	No	0.813	0.704	0.755
Support Vector Machine	Yes	**0.807**	**0.822**	**0.814**
	No	**0.829**	**0.814**	**0.822**



Twitter Data Set Analysis





- Data show strong correlation (Pearson correlation coefficient 0.8907) between Twitter data set and ILI rates from CDC, providing a strong base for accurate prediction of ILI.

Prediction Model

Auto Regressive model with external input (Twitter data)

Logistic ARX Model

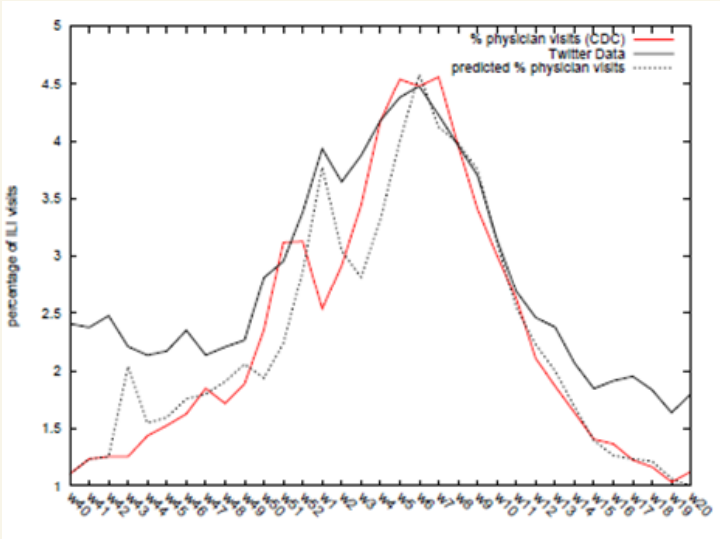
$$\log\left(\frac{y(t)}{1-y(t)}\right) = \sum_{i=1}^{m} a_i \log\left(\frac{y(t-i)}{1-y(t-i)}\right) + \sum_{j=0}^{n-1} b_j \log(u(t-j)) + c + e(t)$$

- y(t) : percentage of physician visits due to ILI in week t
- u(t): number of unique Twitter users with flu related tweets in week t
- e(t): sequence of independent random variables
- c : a constant term to account for offset
- m: previous CDC data in weeks
- n: previous Twitter data in weeks

Cross Validation Results

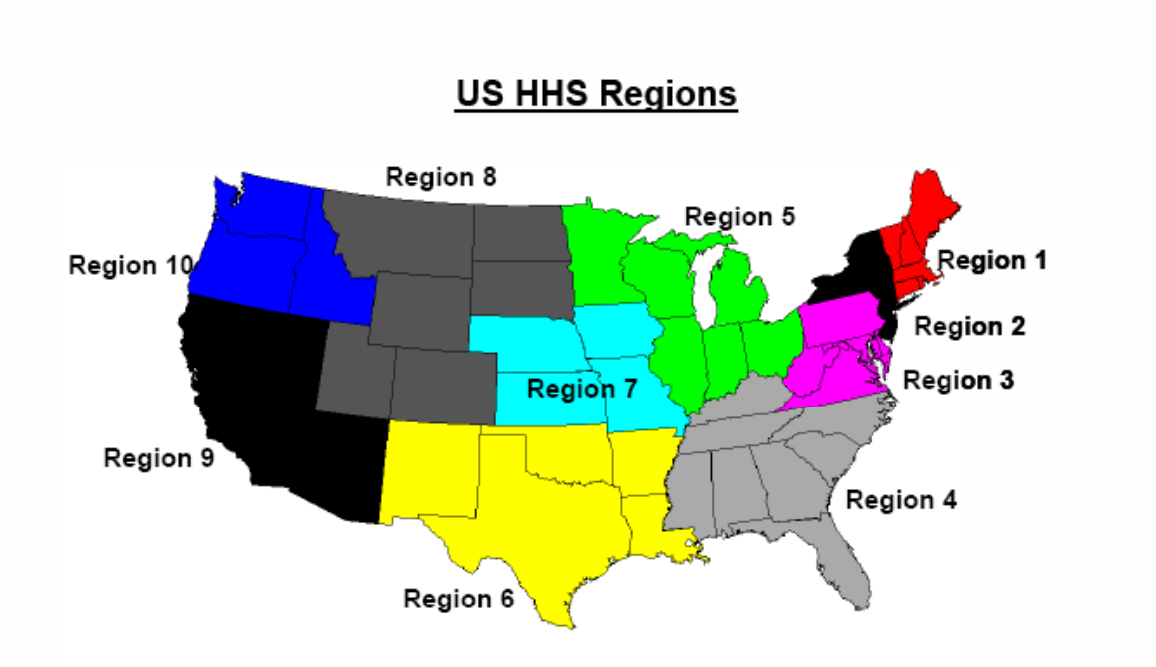
	n=0	n=1	n=2	n=3
m=0		0.5204	0.4636	0.4709
m=1	0.6169	0.4110	0.4225	0.4365
m=2	0.5535	**0.4107**	0.4133	0.4303

Root mean squared errors from 5-fold cross validation

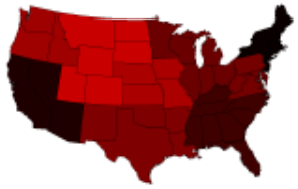
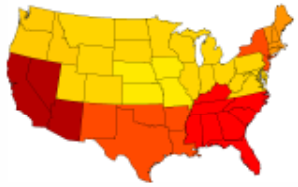
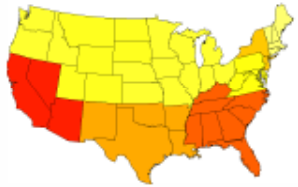


- addition of *Twitter data improves the* prediction with past CDC data alone
- use of Twitter data alone to predict the ILI rates (m = 0) results in poor predictions
- best result when m = 2, n =1: previous 2 week's CDC data, current Twitter data

Regional & Age-based Flu Prediction Analysis



Regional & Age-based Flu Prediction Analysis

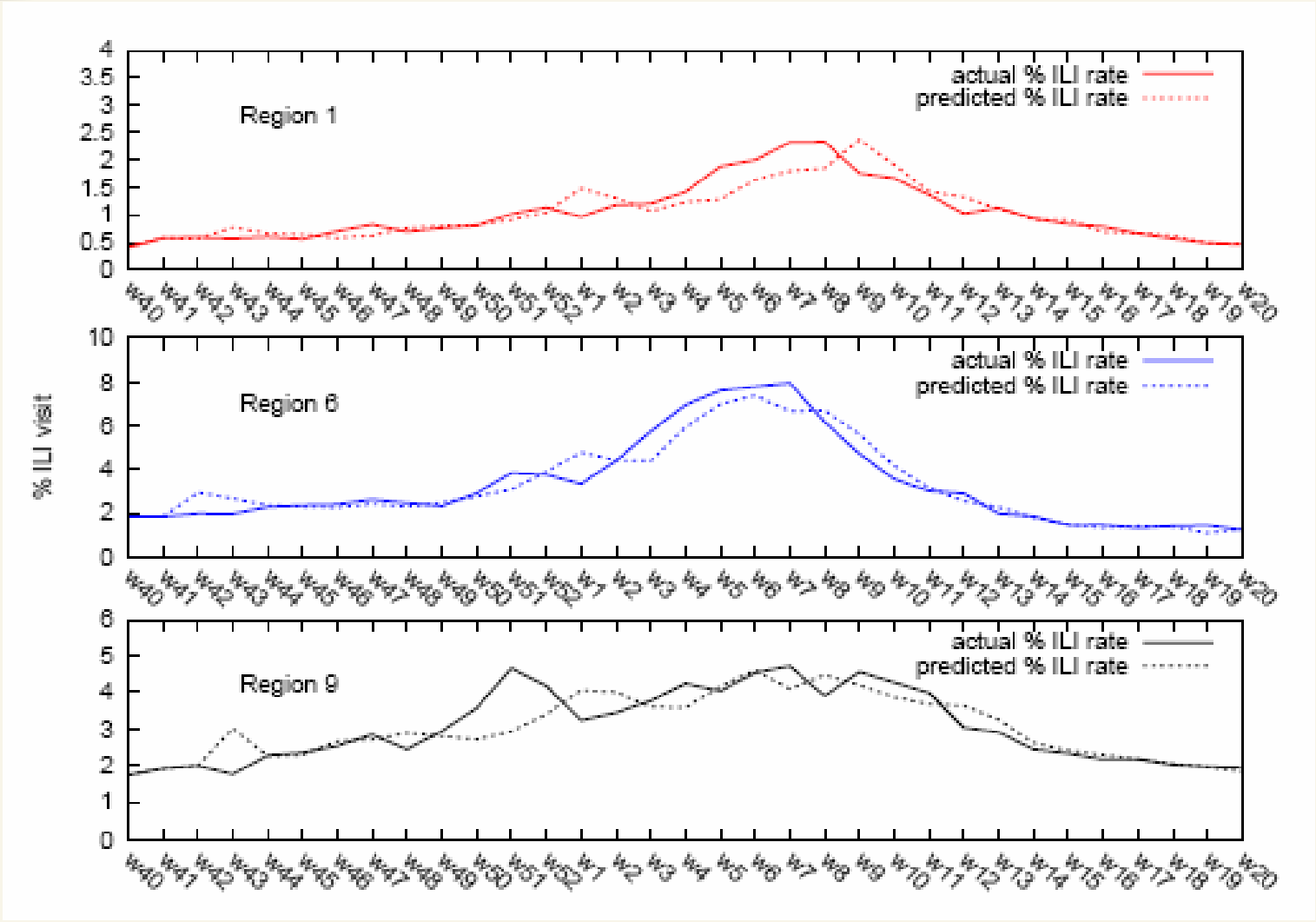
CDC week	CDC ILI data	*Twitter* Reports	Comments
2009 Week 43			Both the ILI and *Twitter* numbers are at their maximum except in the northeast.
2009 Week 44			The ILI and *Twitter* numbers both peak in northeast during this week. The drop in ILI in Regions 2 and 9 is small, which is also reflected in the *Twitter* numbers.
2009 Week 45			Areas that show small drop-off in ILI rates (Regions 1, 2, 4, 10) also show small drops in *Twitter* numbers. The mid-western states show large drops in both ILI and *Twitter* numbers.
2009 Week 47			The ILI incidence in the Southern US remains significant and this reflected to some extent in the *Twitter* numbers.
2009 Week 50			

- ILI seems to peak later in the northeast (Regions 1 and 2) than in the rest of the country by at least week. *The Twitter reports also follow this trend.*

- In Region 9 (CA, NV, AZ …), Region 4 (FL, etc.) and the northeast, the ILI rates seem to drop off fairly slowly in the weeks immediately following the peaks. *This is also reflected in the Twitter reports.*

- Approximately 20-25 weeks after the peak ILI, the northern regions have lower levels relative to the peaks than the southern regions. *This is also true of the Twitter reports.*

Regional & Age-based Flu Prediction Analysis



Twitter data good indicator of ILI rates and can be used to effectively improve the prediction of the current ILI rates.

Regional & Age-based Flu Prediction Analysis

	$0-4yrs$	$5-24yrs$	$25-49yrs$	$50+yrs$
US	0.5285(0-2)	0.4261(2-2)	**0.3577(1-2)**	0.4320(1-1)
Reg1	0.5728(2-1)	0.6000(2-2)	**0.5499(1-1)**	0.7763(1-1)
Reg2	0.6954(0-3)	0.6005(2-1)	**0.4965(0-3)**	0.5171(1-3)
Reg3	0.4423(0-2)	0.3268(2-2)	**0.3066(2-3)**	0.3515(1-2)
Reg4	0.5281(0-3)	**0.3719(0-1)**	0.4792(0-1)	0.5192(0-1)
Reg5	0.6387(1-1)	0.4337(2-3)	**0.4300(0-3)**	0.5198(1-1)
Reg6	**0.3032(0-2)**	0.3407(1-2)	0.3564(0-3)	0.4469(0-3)
Reg7	**0.5426(2-3)**	0.5571(1-3)	0.5492(1-3)	0.6454(2-2)
Reg8	0.6511(1-1)	**0.6133(1-2)**	0.6649(2-2)	0.6445(2-3)
Reg9	0.7453(2-1)	**0.4229(2-1)**	0.4690(1-1)	0.6176(2-1)
Reg10	0.8548(2-1)	**0.5746(2-1)**	0.6462(2-2)	0.7347(2-1)

Prediction performance (mean-squared error)

- For most regions, Twitter data fits the age-groups of 5-24, 25-49 years best, correlates well with the fact that this is likely the most active age groups using Twitter

Conclusions

- Investigated the use of a previously untapped data source, namely, messages posted on Twitter to track and predict influenza epidemic situation in the real world.
- Results show that the number of flu related tweets are highly correlated with ILI activity in CDC data
- Build auto-regression models to predict number of ILI cases in a population as percentage of visits to physicians in successive weeks.
- Verified that Twitter data effectively improves model's accuracy in predicting ILI cases.
- Opportunity to significantly enhance public health preparedness among the masses for influenza epidemic and other large scale pandemic.