

Online Social Media Sensing 1: Sensing the Physical and Social World

CSE 40437/60437-Spring 2015

Prof. Dong Wang

Online Social Media: A New Information Age



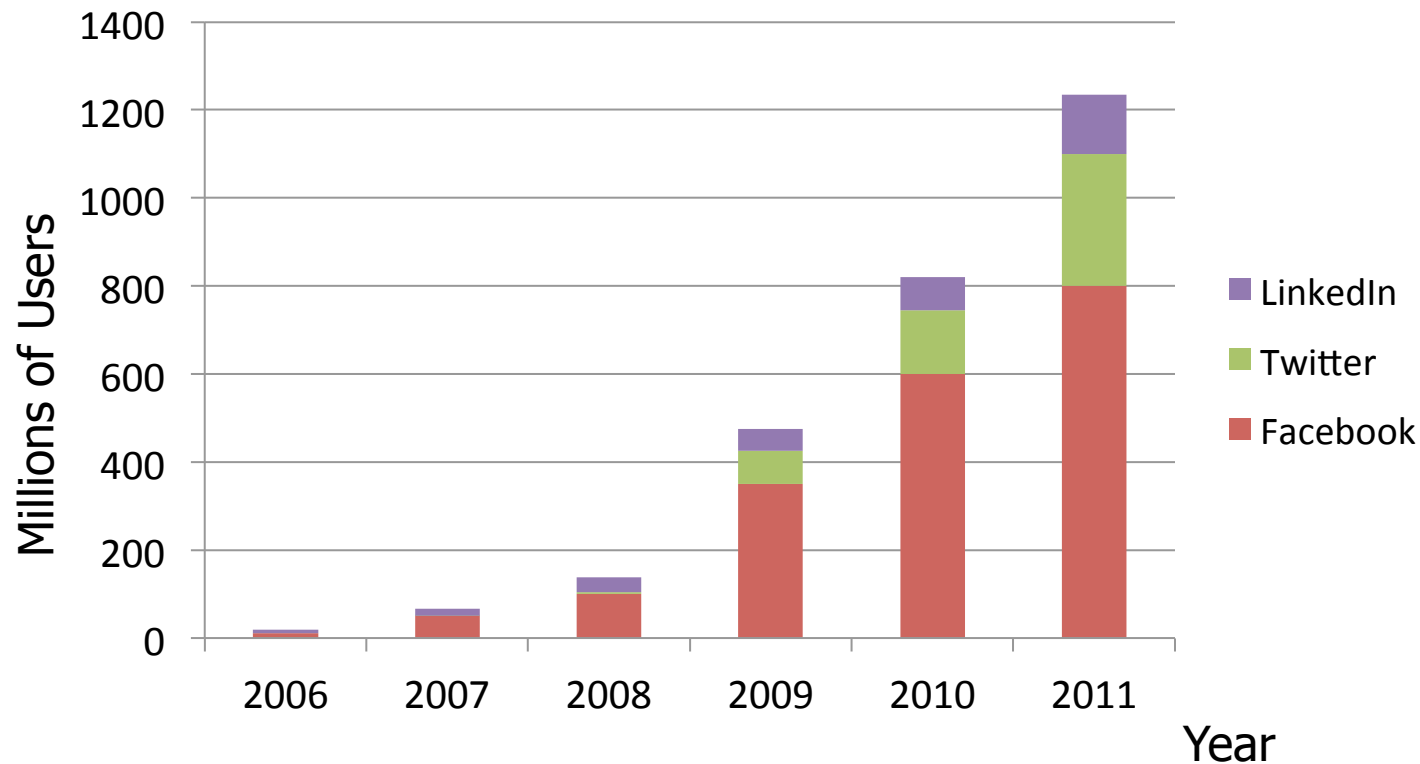
Empower the analysts and decision makers to quickly understand ongoing events!



The Decade of Social Media

Platforms for Information Dissemination

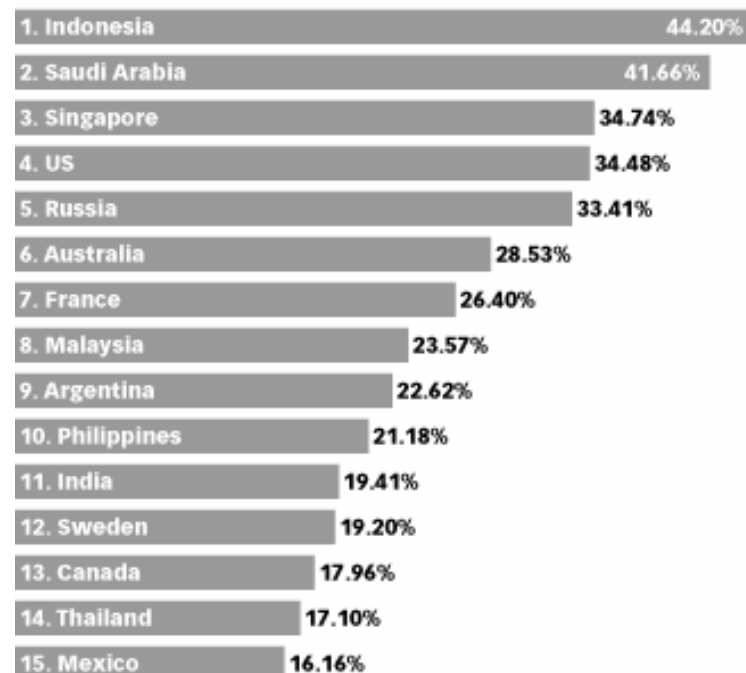
- Game changing trend of the decade: The rise of social media



Social Media Statistics

- Social media growth charts

Top 15 Countries on Twitter, Ranked by Growth in Account Owners, Q1 2013
% change vs. Q2 2012



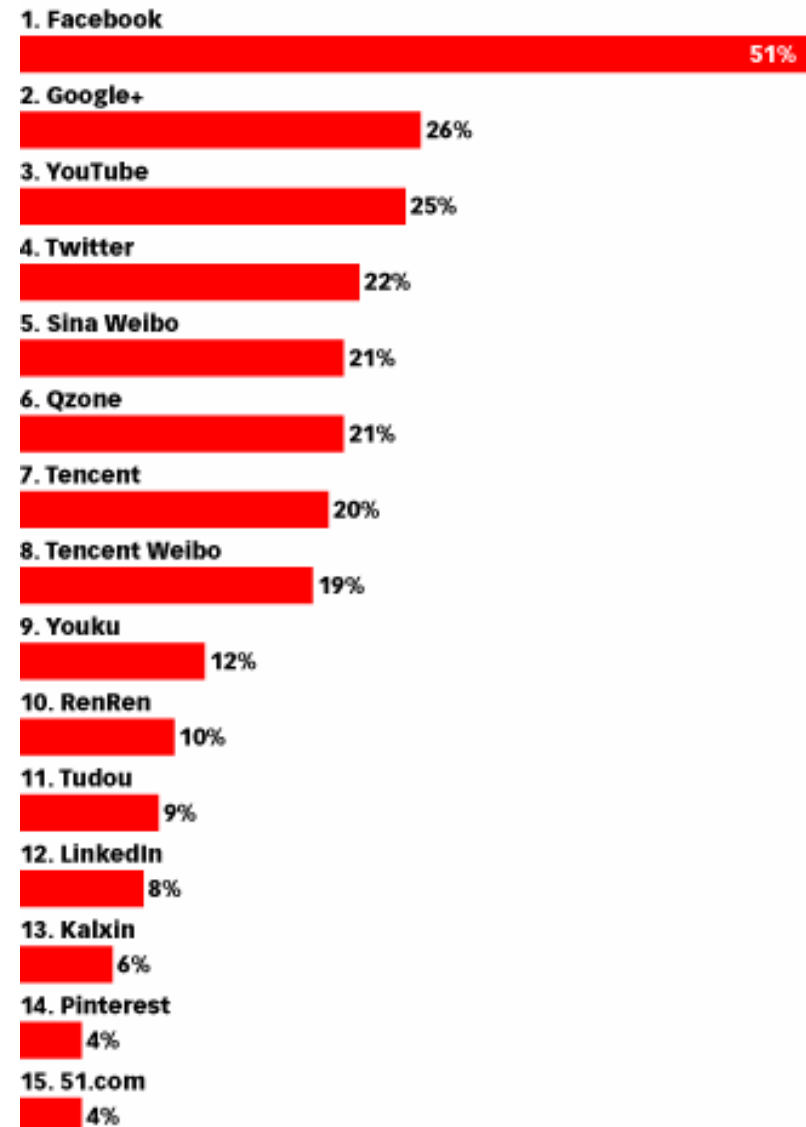
Note: ages 16-64; used or contributed in the past month
Source: GlobalWebIndex, "Stream Social: Quarterly Social Platforms Update Q1 2013," April 26, 2013

156844

www.eMarketer.com

Top 15 Social Media Sites Worldwide, Ranked by Penetration of Active Users, Q1 2013

% of internet users



Note: ages 16-64; used or contributed in the past month
Source: GlobalWebIndex, "Stream Social: Quarterly Social Platforms Update Q1 2013," April 26, 2013

156798

www.eMarketer.com

A Wealth of Social Sensing Content

Every Minute:

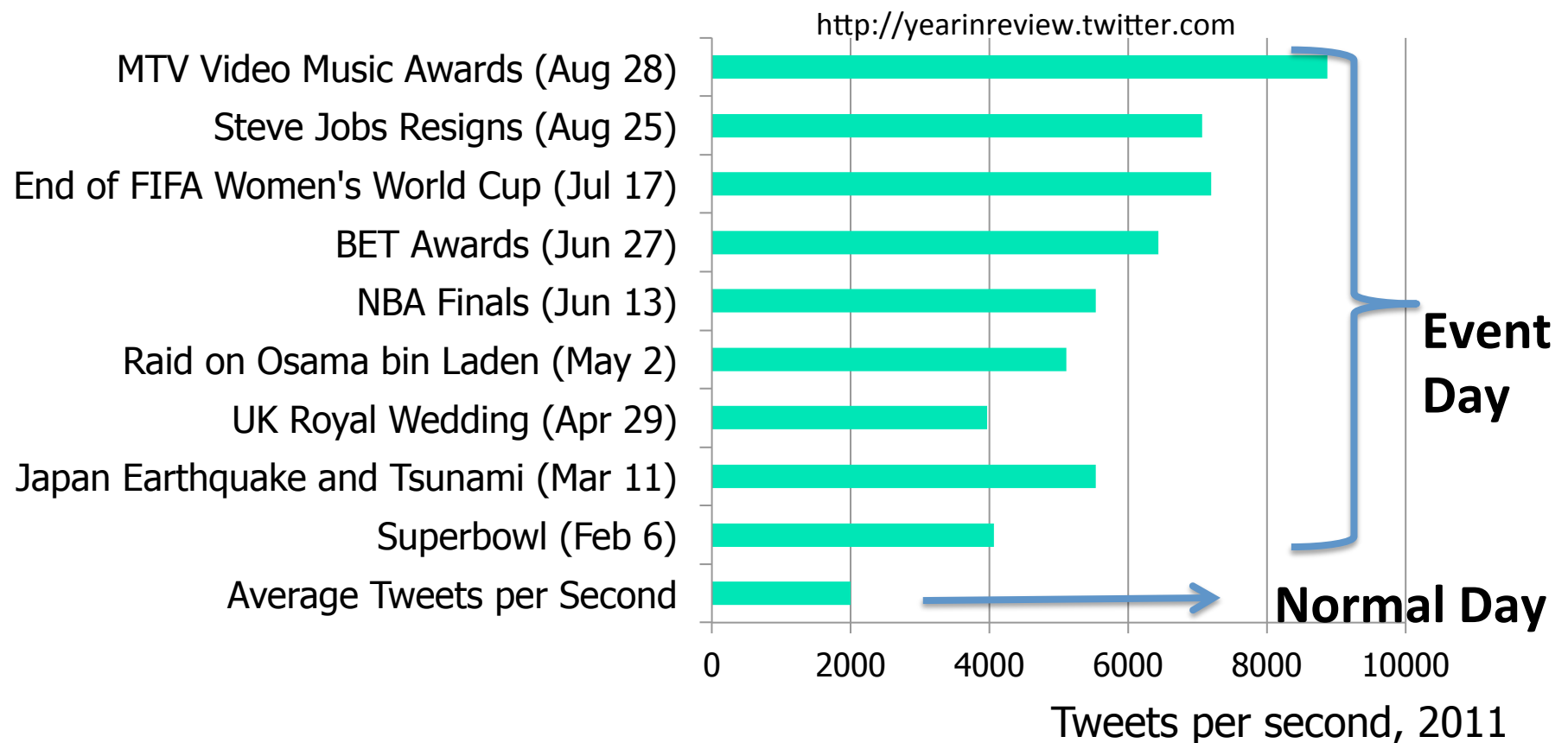


- More than **350,000 tweets** made on Twitter
<http://www.internetlivestats.com/twitter-statistics/>
- Almost **700,000 status updates** made on Facebook
<http://www.themarketingbit.com/infographics/online-for-one-minute/>
- More than **3,500 images** uploaded to Flickr
<http://www.pcmag.com>
- More than **2000 check-ins** on FourSquare
<http://articles.businessinsider.com>
- More than **100 hours of video** uploaded to YouTube
http://www.youtube.com/t/press_statistics



Significant Information on Ongoing Events

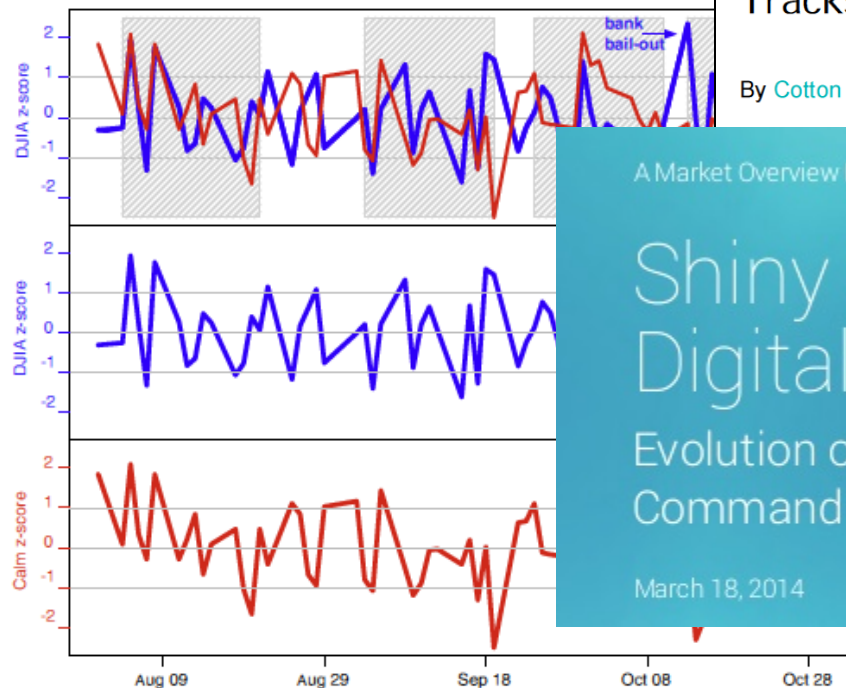
- Thousands of tweets per second during key events



Examples of Real-time Content Analytics

Twitter mood predicts stock market trends

J. Bollen et al. / Journal of Computational Science 2 (2011) 1–



Wells Fargo Opens Command Center to Handle Surge of Social Content

Tracks Anywhere From 2,000 to 4,000 Mentions a Day

By [Cotton Delo](#). Published on April 08, 2014. 1

A Market Overview Report

Shiny Object or
Digital Intelligence Hub?
Evolution of the Enterprise Social Media
Command Center

March 18, 2014

Reconstructing Event Timelines

A Twitter Example

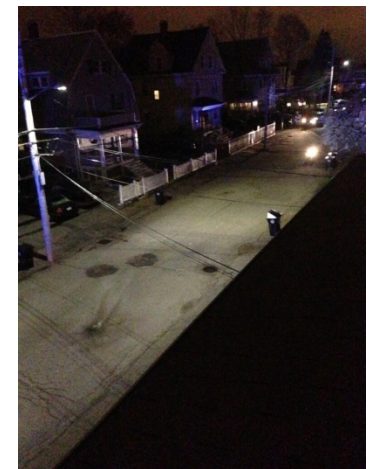


RT @Boston_Police: Do you know these individuals? Contact boston@ic.fbi.gov or 1-800-CALL-FBI (1-800-225-5324), prompt #3 <http://t.co/QJias1Kywe>



RT @TomlinM: Unreal. Officer down in Cambridge. Scene centered around a MIT building. Does NOT appear related to #bostonmarathon. <http://t.co/DQuXQuP7wK>

@AKitz: Site of the bomb explosion on laurel st. bomb detectors are out #mitshooting #boston #mit <http://t.co/lhOjCiBgY5>




2:50 PM- 15 April 13

RT @MassStatePolice: Photos taken from State Police Air Wing on Watertown manhunt.Media, please credit MSP for pics. <http://t.co/Qzafbp4MBE>



FILED UNDER [Internet](#)

Japan considers using social networks in disaster situations

By Jamie Rigg  posted Aug 30th 2012 1:41AM



Emergency services are embracing technology as ways to [investigate](#), [send alerts](#) and [receive reports](#) on crises. And now, the Japanese are looking at social networks to support communication in disaster scenarios, especially when traditional services [fail](#). The local [Fire and Disaster Management Agency](#) [put](#) together a panel discussion on just that topic, with representatives attending from the likes of Twitter, Yahoo, Mixi and NHN Japan, as well as various government and emergency bodies. The talk was motivated, in part, by the March tsunami, when the internet was the sole means of information for some and with initiatives like [Google's Person Finder](#) playing a role in the aftermath. Any formal implementation of the ideas discussed is probably

Japan's Tsunami and Nuclear Event

Dow Jones Hickup

- Dow Jones lost 150 points on a rumor of two explosions in the White House on April 23rd, 2013



Challenge:

Data Cleaning and Summarization

- Much like Google organizes (relatively static) world content, we need an engine for organizing real-time/streaming data feeds and:

Reconstructing the
“State of the World”,
Physical and Social!

Information distillation

Clean structured
representation,
high quality of
information

A firehose of text, images, video,
sound, and time-series data



Papers discuss today

Paper 1: "Earthquake shakes Twitter users: real-time event detection by social sensors." Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. Proceedings of the 19th international conference on World wide web. ACM, 2010.

Paper2: "From tweets to polls: Linking text sentiment to public opinion time series." O'Connor, Brendan, et al. ICWSM 11 (2010): 122-129.

Physical Events

Paper 1: "Earthquake shakes Twitter users: real-time event detection by social sensors." Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. Proceedings of the 19th international conference on World wide web. ACM, 2010.



Outline

Introduction

Event Detection

Model

Experiments And Evaluation

Application

Conclusions

What's happening?

- Twitter
 - is one of the most popular microblogging services
 - has received much attention recently
- Microblogging
 - is a form of blogging
 - that allows users to send brief text updates
 - is a form of micromedia
 - that allows users to send photographs or audio clips
- In this research, they focus on an important characteristic

real-time nature

Real-time Nature of Microblogging

social events
parties
baseball games
presidential campaign

disastrous events
storms
fires
traffic jams
riots
heavy rain-falls
earthquakes

- Twitter users write tweets several times a day
- There is a large number of tweets, which results in many reports related to events
- We can know how other users are doing in real-time
- We can know what happens around other users in real-time.

The Motivation

- Adam Ostrow, an Editor in Chief at Mashable wrote the possibility to detect earthquakes from tweets in his blog

We can know earthquake occurrences from tweets
-> The motivation of this research

So
the



bachocha
Ricardo Duran

traveling

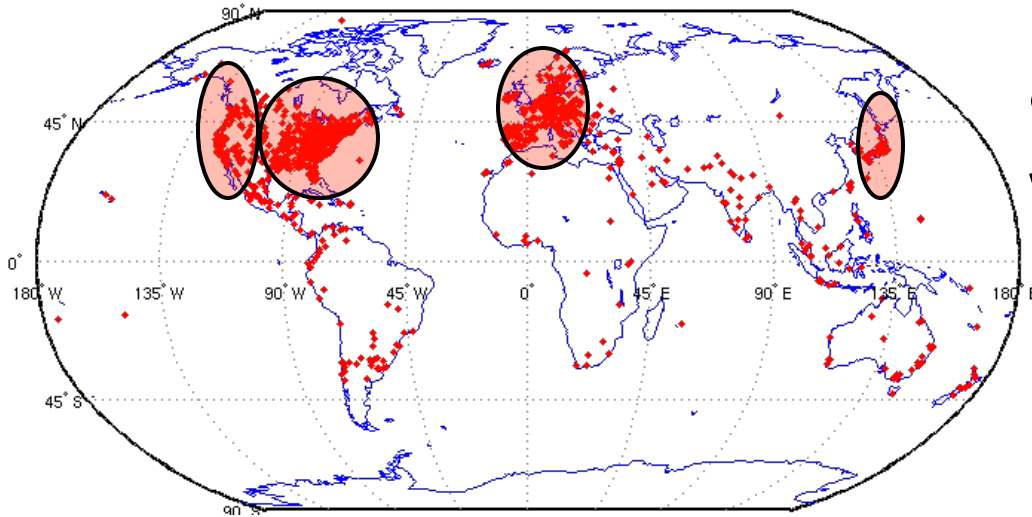
*USGS : United States Geological Survey

The Goals

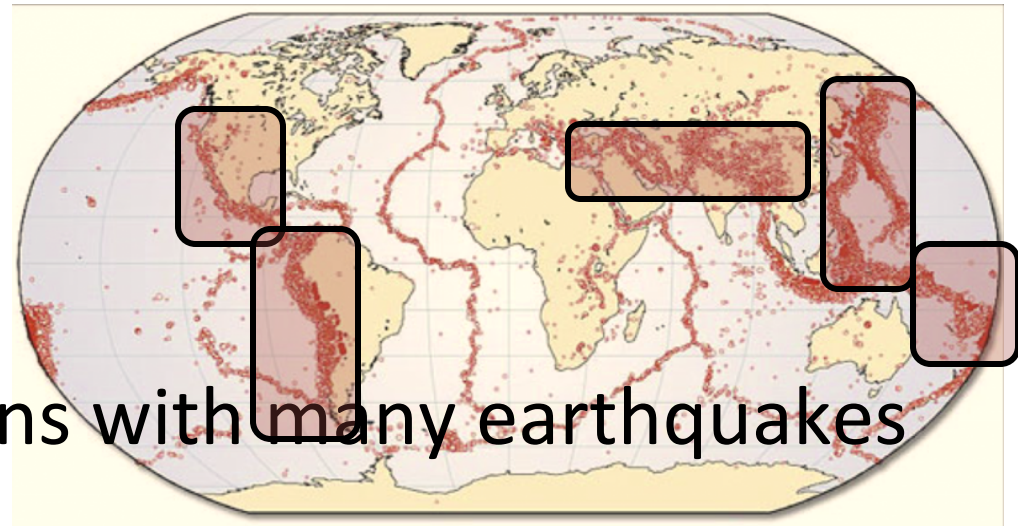
- Propose an algorithm to detect a target event
 - do semantic analysis on Tweet
 - to obtain tweets on the target event precisely
 - regard Twitter user as a sensor
 - to detect the target event
 - to estimate location of the target
- Produce a probabilistic spatio-temporal model for
 - event detection
 - location estimation
- Propose Earthquake Reporting System using Japanese tweets

Twitter and Earthquakes in Japan

a map of Twitter user world wide

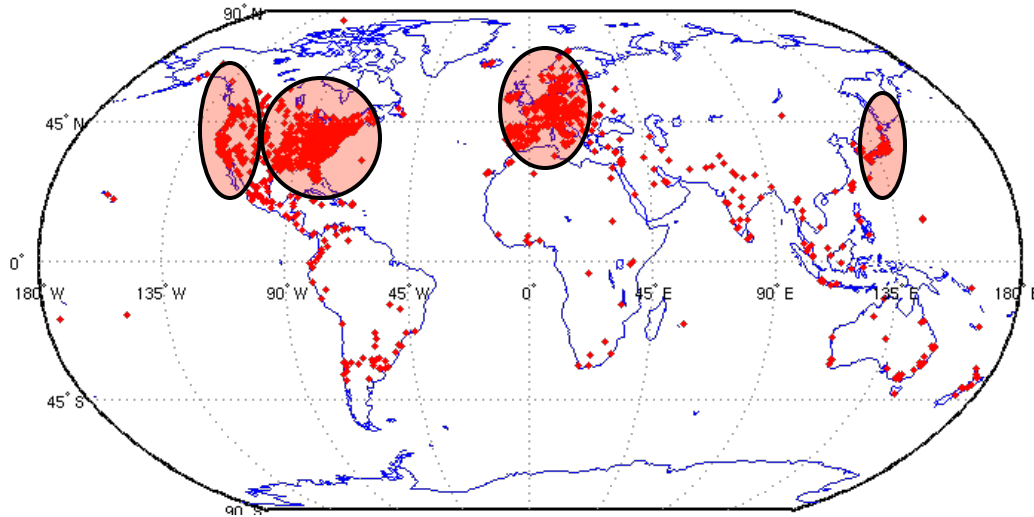


a map of earthquake occurrences world wide



The intersection is regions with many earthquakes and large twitter users.

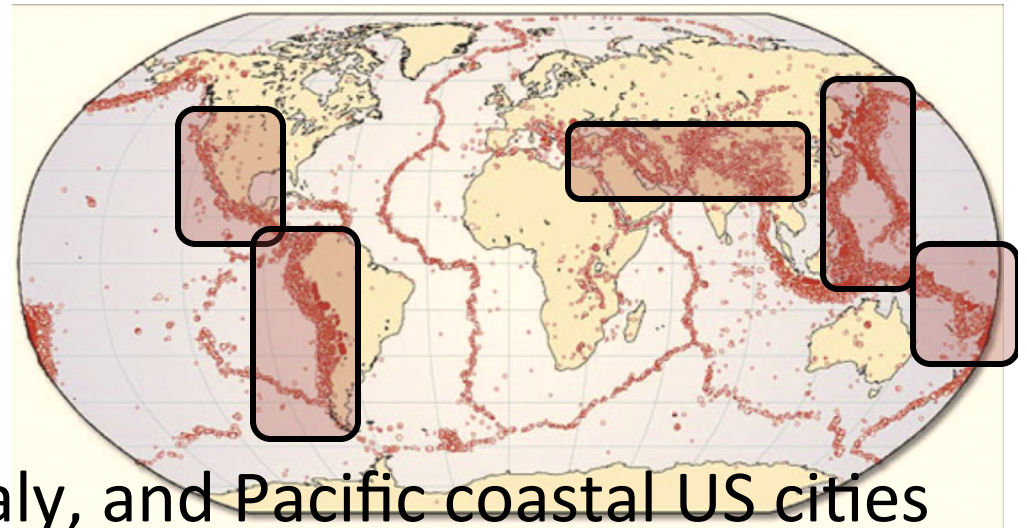
Twitter and Earthquakes in Japan



0

Other regions:

Indonesia, Turkey, Iran, Italy, and Pacific coastal US cities



Event detection algorithms

- Do semantic analysis on Tweet
 - to obtain tweets on the target event precisely
- Consider a Twitter user as a sensor
 - to detect the target event
 - to estimate location of the target

Semantic Analysis on Tweet

- Search tweets including **keywords** related to a target event
 - Example: In the case of earthquakes
 - “shaking”, “earthquake”
- Classify tweets into a **positive** class or a **negative** class
 - Example:
 - “Earthquake right now!!” ---positive
 - “Someone is shaking hands with my boss” --- negative
 - Create a classifier

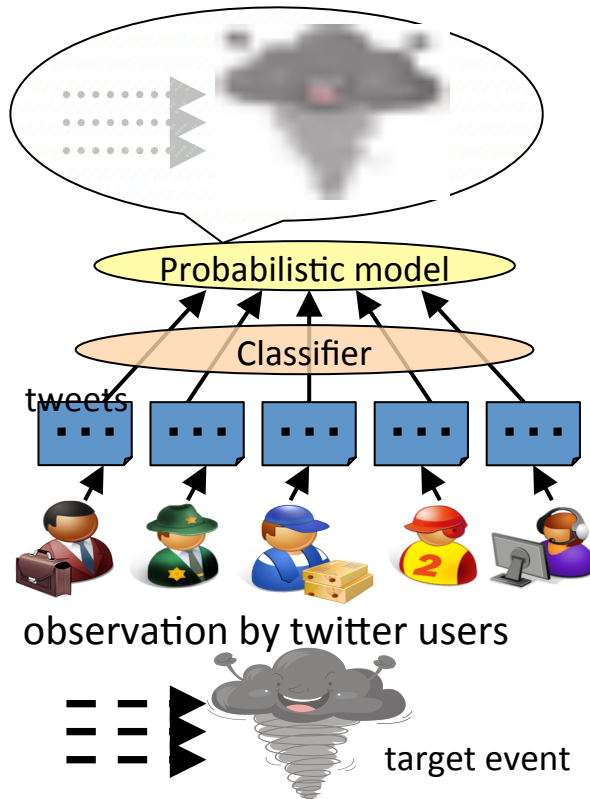
*How would you build such a classifier?
What kinds of features would you use?*

Semantic Analysis on Tweet

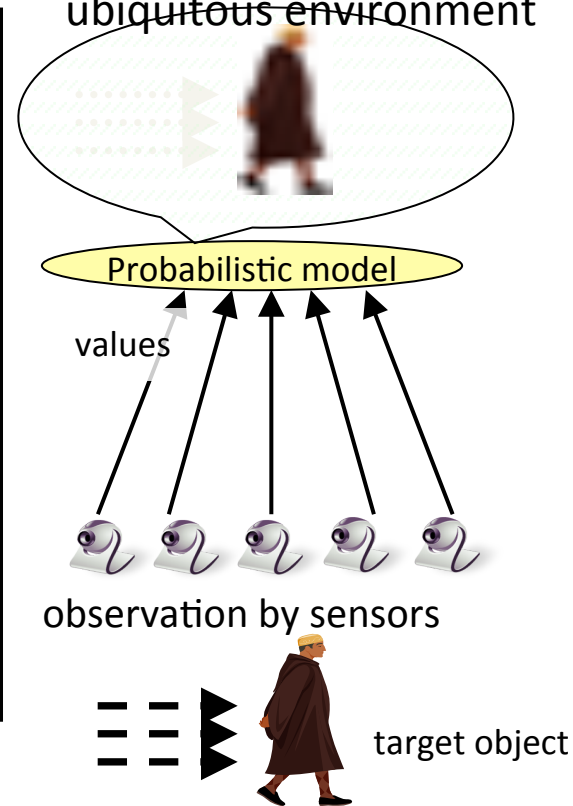
- Create classifier for tweets
 - use Support Vector Machine (SVM)
- Features (Example: I am in Japan, earthquake right now!)
 - **Statistical features** (7 words, the 5th word)
the number of words in a tweet message and the position of the query within a tweet
 - **Keyword features** (I, am, in, Japan, earthquake, right, now)
the words in a tweet
 - **Word context features** (Japan, right)
the words before and after the query word

Tweet as a Sensory Value

Event detection from twitter

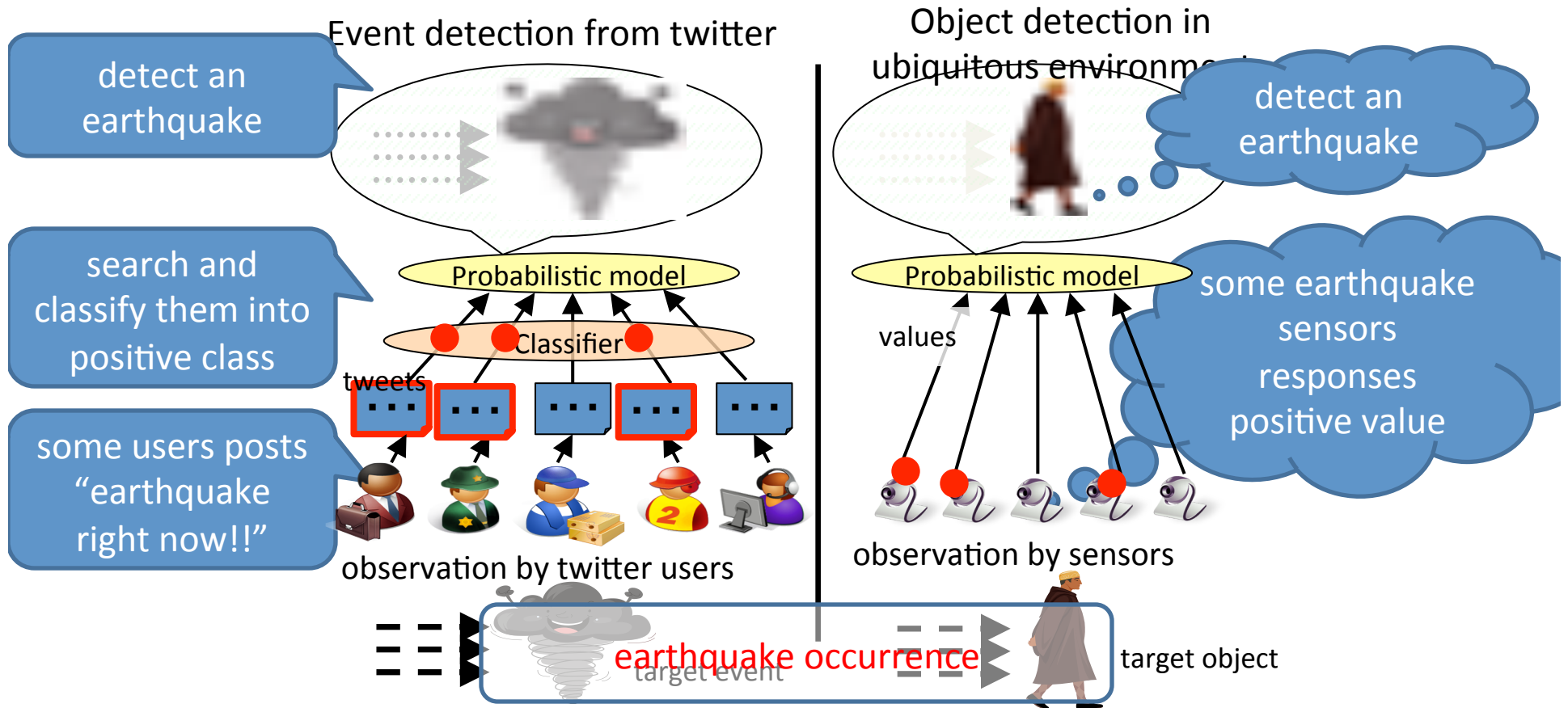


Object detection in ubiquitous environment



the correspondence between **tweets processing** and **sensory data detection**

Tweet as a Sensory Value



We can apply methods for sensory data detection to tweets processing

Tweet as a Sensory Value

- **The location assumption may not always hold:**
 1. Very few tweets may have geo-locations embedded.
 2. The user's profile location may not match the exact location of the user.
- **Assumption 2:** Each tweet is associated with a time and location
 - a time : post time
 - location : GPS data or location information in user's profile

Processing the time and location information of the positive tweets, we can detect target events and estimate location of target events

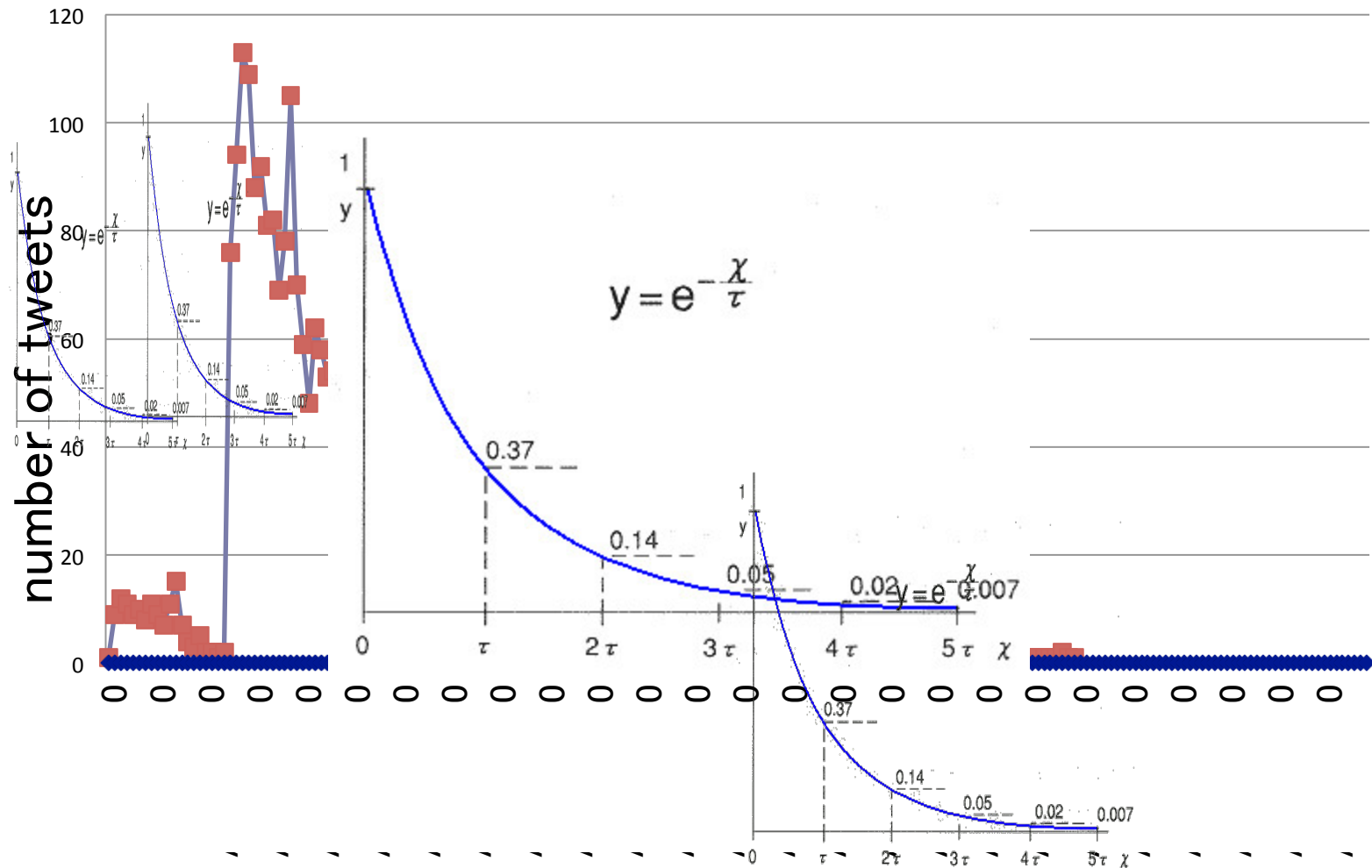
Probabilistic Model

- Why we need probabilistic models?
 - Sensor values are noisy and sometimes sensors work incorrectly (e.g., **unreliable human sources** and non-perfect classifier)
 - We cannot judge whether a target event occurred or not from a single tweet
 - We have to calculate the probability of an event occurrence from a series of data
- Propose a probabilistic model for
 - **event detection** from time-series data
 - **location estimation** from a series of spatial information

Temporal Model

- Calculate the probability of an event occurrence from multiple sensor values
- Examine the actual time-series data to create a temporal model

Temporal Model



Temporal Model

- The data fits very well to an exponential function

$$f(t; \lambda) = \lambda e^{-\lambda t} (t > 0, \lambda > 0) \quad \lambda = 0.34$$

- Design the alarm of the target event probabilistically, which was based on an exponential distribution

Temporal Model

- The probability of an event occurrence at time t

$$p_{occur}(t) = 1 - p_f^{n_0(1-e^{-\lambda(t+1)})/(1-e^{-\lambda})}$$

- the false positive ratio of a sensor p_f
- the probability of all n sensors returning a false alarm p_f^n
- the probability of event occurrence $1 - p_f^n$
- n_0 sensors at time 0 $\rightarrow n_0 e^{-\lambda t}$ sensors at time t
- the number of sensors until t $n_0(1 - e^{-\lambda(t+1)})/(1 - e^{-\lambda})$

- Expected wait time t_{wait} to deliver notification for 1% false positives

$$t_{wait} = (1 - (0.1264/n_0))/0.7117 - 1$$

- parameter

$$\lambda = 0.34, p_f = 0.35, p_{occurr} = 0.99$$

Less than 0.4 s

Spatial Model

- We must calculate the probability distribution of location of a target
- We apply Bayes filters to this problem which are often used in location estimation by sensors
 - Kalman Filters
 - Particle Filters

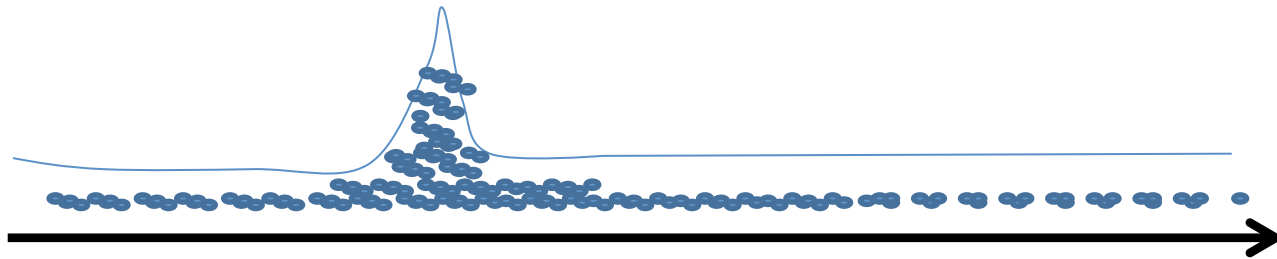
Bayesian Filters for Location Estimation

- Kalman Filters
 - are the most widely used variant of Bayes filters
 - approximate the probability distribution which is virtually identical **to a uni-modal Gaussian** representation
 - advantages: the computational efficiency
 - disadvantages: being limited to accurate sensors or sensors with high update rates

Bayesian Filters for Location Estimation

- Particle Filters

- represent the probability distribution by sets of samples, or particles
- advantages: the ability to **represent arbitrary probability densities**
 - particle filters can converge to the true posterior even in non-Gaussian, nonlinear dynamic systems.
- disadvantages: the difficulty in applying to high-dimensional estimation problems



A Short Intro on Particle Filter



A short tutorial: <http://www.it.uu.se/katalog/andsv164>

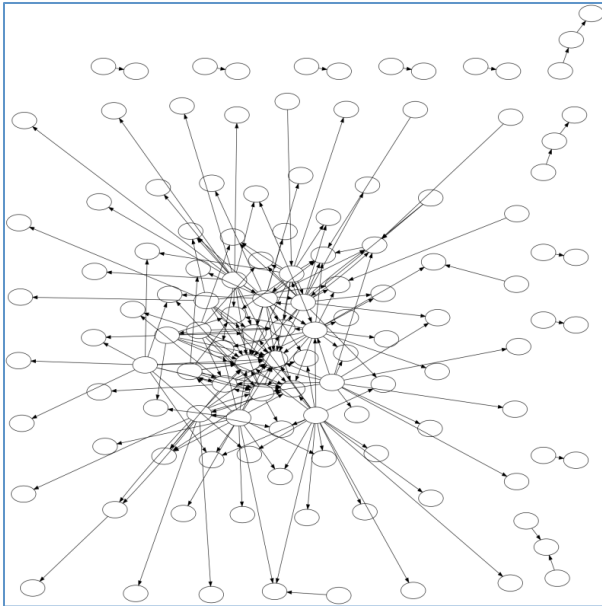
Information Diffusion Related to Real-time Events

- Proposed spatiotemporal models need to meet one condition that
 - Sensors are assumed to be independent
- We check if information diffusions about target events happen because
 - if an information diffusion happened among users, Twitter user sensors are not independent. They affect each other

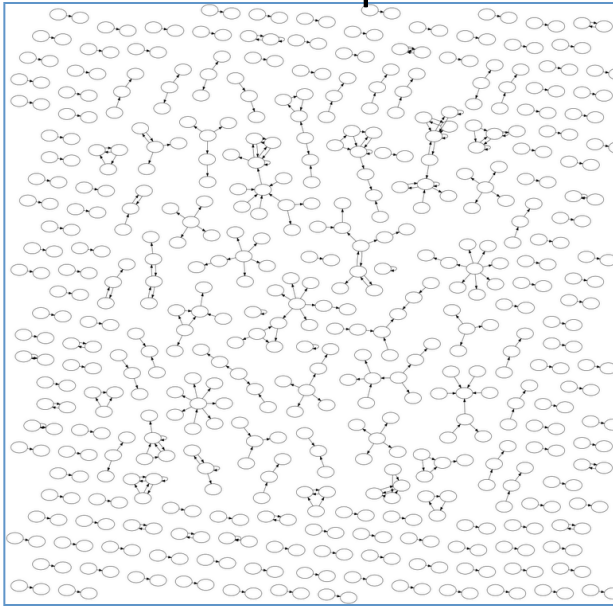
Information Diffusion Related to Real-time Events

Information Flow Networks on Twitter

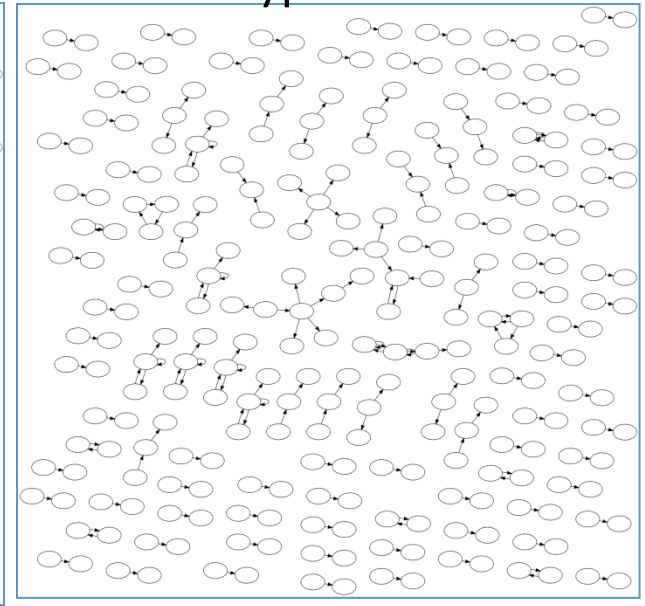
Nintendo DS Game



an earthquake



a typhoon



In the case of an earthquakes and a typhoons, very little information diffusion takes place on Twitter, compared to Nintendo DS Game

— Source independency may not always hold for twitter case studies, even for disaster scenarios!

Experiments And Evaluation

- Demonstrate performances of
 - **tweet classification**
 - **event detection** from time-series data
 - show this results in “application”
 - **location estimation** from a series of spatial information

Evaluation of Semantic Analysis

- Queries
 - Earthquake query: “shaking” and “earthquake”
 - Typhoon query: “typhoon”
 -
- Examples to create classifier
 - 597 positive examples

Evaluation of Semantic Analysis

- “earthquake” query

Features	Recall	Precision	F-Value
Statistical	87.50%	63.64%	73.69%
Keywords	87.50%	38.89%	53.85%
Context	50.00%	66.67%	57.14%

- “sh

“Is this an earthquake or a truck passing?”

Features	Recall	Precision	F-Value
Statistical	66.67%	68.57%	67.61%
Keywords	86.11%	57.41%	68.89%
Context	52.78%	86.36%	68.20%
All	80.56%	65.91%	72.50%

Discussions of Semantic Analysis

Features	Recall	Precision	F-Value
Statistical	87.50%	63.64%	73.69%
Keywords	87.50%	38.89%	53.85%
Context	50.00%	66.67%	57.14%
All	87.50%	63.64%	73.69%

- It obtains highest F-value when we use **Statistical features** and **all features**.
- Keyword features and Word Context features don't contribute much to the classification performance
- A user becomes surprised and might produce a very short tweet
- It's apparent that the precision is not so high as the recall

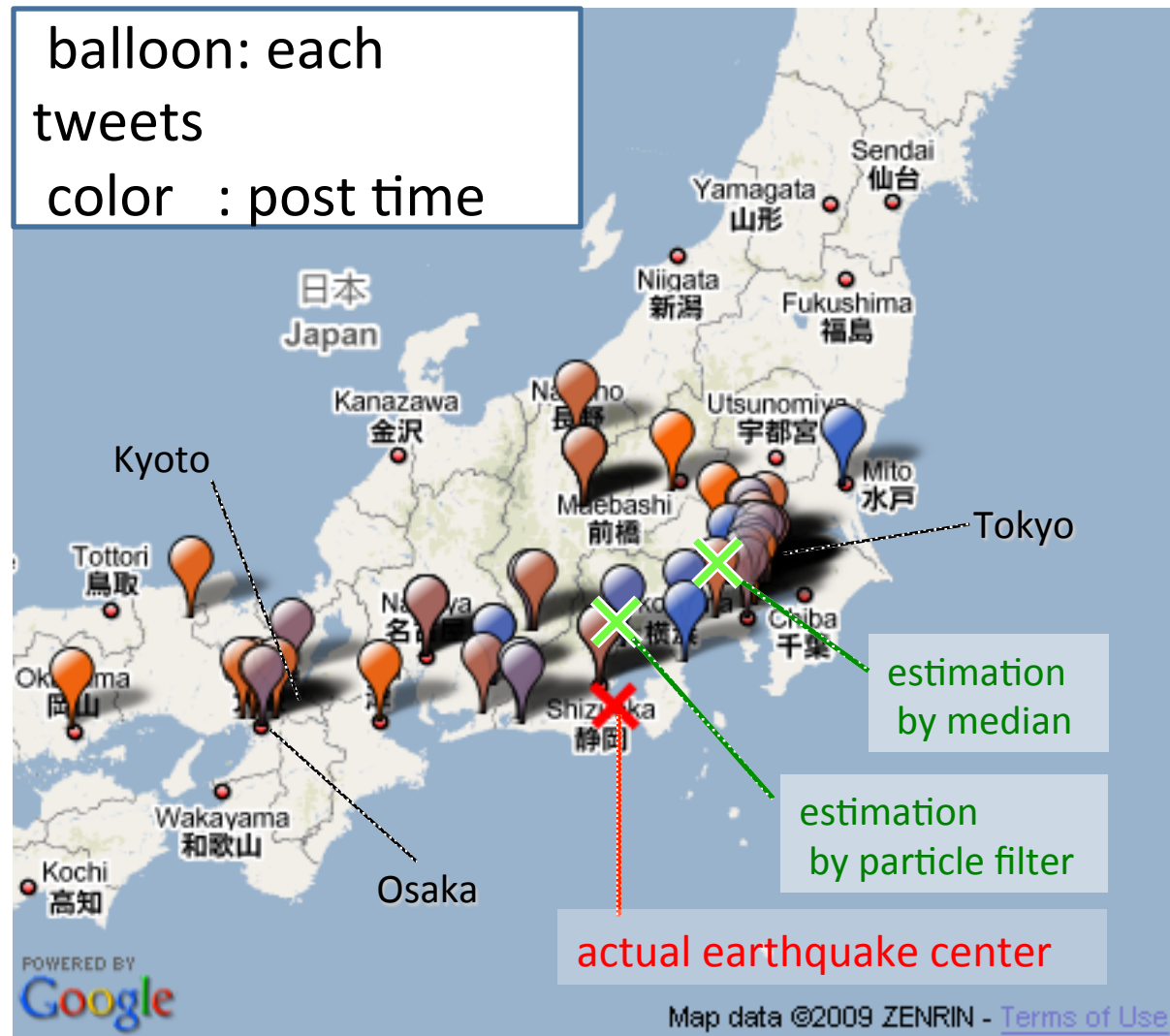
Experiments And Evaluation

- Demonstrate performances of
 - tweet classification
 - event detection from time-series data
 - show this results in “application”
 - location estimation from a series of spatial information

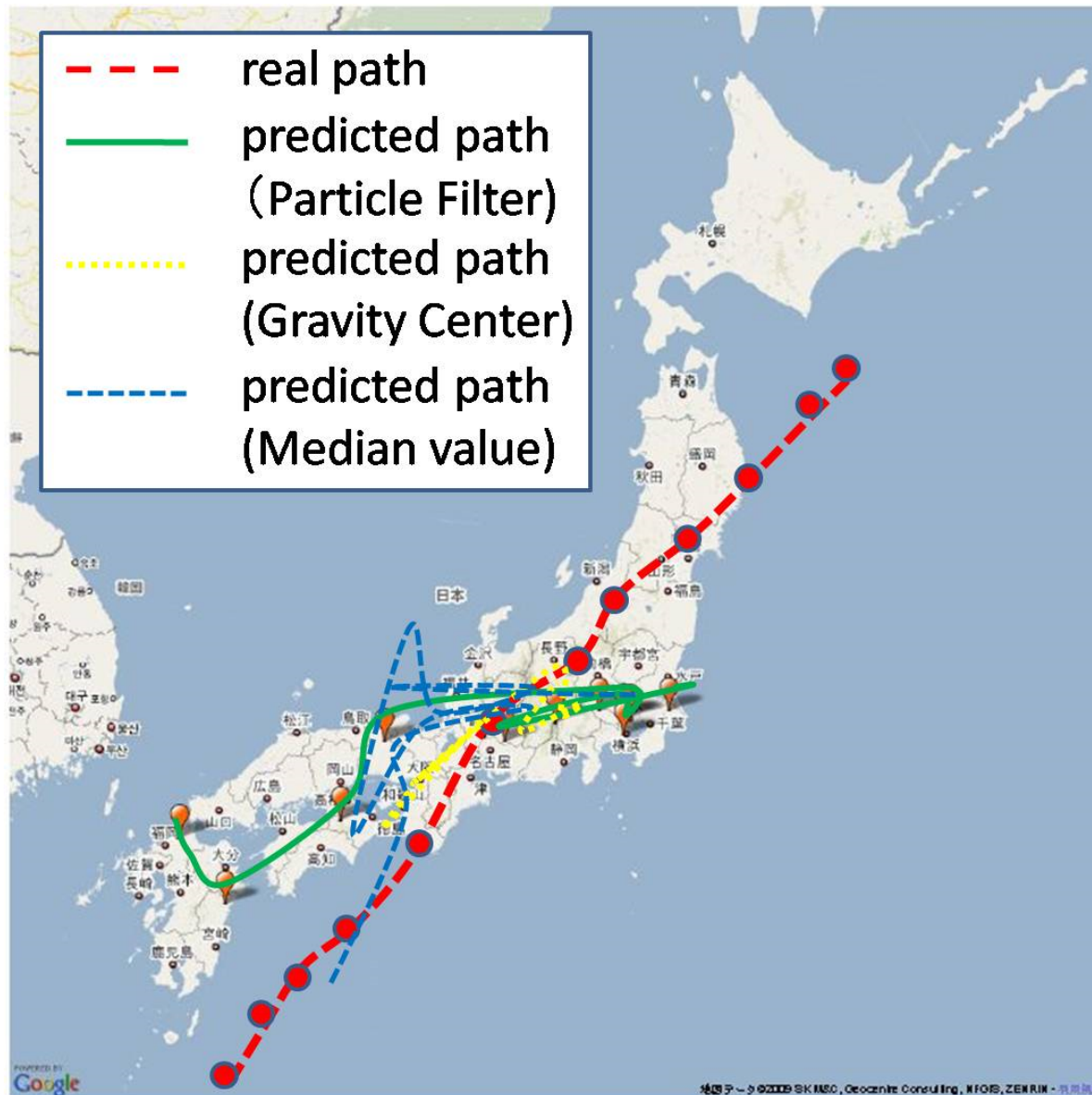
Evaluation of Spatial Estimation

- Target events
 - earthquakes
 - 25 earthquakes from August.2009 to October 2009
 - typhoons
 - name: Melor
- Baseline methods
 - weighed average
 - simply takes the average of latitudes and longitudes
 - the median
 - simply takes the median of latitudes and longitudes
- Evaluate methods by distances from actual centers
 - a distance from an actual center is smaller, a method works better

Evaluation of Spatial Estimation



Evaluation of Spatial Estimation



Evaluation of Spatial Estimation Earthquakes

Date	Actual Center		Median			Weighed Average			Kalman Filter			Particle Filter		
Aug. 10 01:00	33.10	138.50	3.40	-0.80	3.49	2.70	-0.10	2.70	2.67	-0.50	2.72	2.60	0.50	2.65
Aug. 11 05:00	34.80	138.50	0.90	-0.90	1.27	0.70	-0.30	0.76	0.60	-0.20	0.63	0.30	-0.90	0.95
Aug. 13 07:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 08:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 09:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 10:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 11:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 12:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 13:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 14:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 15:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 16:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 17:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 18:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 19:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 20:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 21:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 22:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 13 23:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	3.25	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 25 20:19	35.40	140.40	-1.80	2.41	2.20	-0.70	2.31	0.70	-1.60	1.75	1.40	0.10	1.40	0.42
Aug. 31 00:46	37.20	141.50	-3.60	3.62	-1.10	-2.30	2.55	-1.30	-2.20	2.56	-0.30	-0.30	0.42	1.71
Aug. 31 21:11	33.40	130.90	-3.60	5.76	0.50	2.10	2.16	0.70	1.90	2.02	-0.20	-1.70	1.71	3.19
Sep. 3 22:26	31.10	130.30	-0.10	6.20	4.00	5.00	6.40	4.90	7.20	8.71	2.40	2.10	3.19	1.61
Sep. 4 11:30	35.80	140.10	-1.70	3.54	0.20	-0.90	0.92	0.00	-1.00	1.00	0.80	1.40	1.61	6.17
Sep. 05 10:59	37.00	140.20	-2.10	8.30	8.73	-1.40	-3.10	3.40	-1.30	-3.30	3.55	-2.10	-5.80	3.83
Sep. 08 01:24	42.20	143.00	-3.60	9.90	9.60	-2.50	-3.90	4.63	-4.50	-6.00	7.50	1.30	-3.60	7.06
Sep. 10 18:29	43.20	146.20	-5.90	11.78	-4.90	-7.10	8.63	-4.50	-7.20	8.49	-0.90	-7.00	2.51	8.36
Sep. 16 21:38	33.40	130.90	1.10	1.12	0.90	2.10	2.28	0.50	1.40	1.49	-0.20	-2.50	0.76	2.02
Sep. 22 20:40	47.60	141.70	-11.10	13.40	-10.80	-3.10	11.24	-11.30	-3.80	11.92	-7.80	-3.00	0.58	2.73
Oct. 1 19:43	36.40	140.70	0.70	3.86	-0.60	-1.80	1.90	-0.30	-1.50	1.53	-0.70	0.30	0.58	2.73
Oct. 5 09:35	42.40	141.60	-3.70	4.83	-2.70	-2.00	3.36	-2.60	-1.60	3.05	1.10	-1.70	0.58	2.73
Oct. 6 07:49	35.90	137.60	0.50	1.30	-0.20	0.80	0.82	-0.10	0.90	0.91	0.30	0.50	0.58	2.73
Oct. 10 17:43	41.80	142.20	-3.50	6.44	-1.40	-2.10	2.52	-2.20	-2.60	3.41	2.40	-1.30	0.58	2.73
Oct. 12 16:10	35.90	137.60	2.80	2.84	0.80	1.20	1.44	0.80	1.60	1.79	3.60	1.40	0.58	2.73
Average	—		5.47			3.62			3.85			3.01		

Particle filters works better than other methods

Evaluation of Spatial Estimation

A typhoon

Date	Actual Center		Median			Weighed Average			Kalman Filter			Particle Filter		
	lat.	long.	lat.	long.	dist.	lat.	long.	dist.	lat.	long.	dist.	lat.	long.	dist.
Oct. 7 12:00	29.00	131.80	-1.90	-1.90	2.69	-5.20	-3.60	6.32	-3.90	-1.10	4.05	-4.70	1.10	4.83
Oct. 7 15:00	29.90	132.50	-3.70	-2.60	4.52	-3.80	-2.40	4.49	3.20	3.10	4.46	-2.70	0.90	2.85
Oct. 7 18:00	29.80	133.80	-1.90	-1.90	4.52	-4.40	-3.50	5.62	-6.40	5.40	8.37	-3.20	-0.70	3.28
Oct. 7 21:00	29.80	135.50	-3.50	-3.50	5.24	-3.60	-3.30	4.88	-10.90	-1.60	11.02	-3.70	-0.50	3.73
Oct. 8 00:00	29.80	137.10	-2.10	-2.10	2.30	-2.30	-0.90	2.47	-12.60	-20.40	23.98	-2.90	-3.50	4.55
Oct. 8 03:00	29.80	138.00	-1.00	-1.00	3.40	0.80	1.70	1.88	4.20	16.00	16.54	-0.60	-2.50	2.57
Oct. 8 06:00	29.80	139.60	-1.60	-1.60	3.65	0.00	0.50	0.50	0.50	2.60	2.65	0.70	-0.80	1.06
Oct. 8 09:00	29.80	141.90	-1.90	-1.90	4.25	1.50	1.20	1.92	2.10	1.60	2.64	1.40	0.10	1.40
Oct. 8 12:00	38.00	140.00	-3.20	-3.20	3.94	2.40	2.20	3.26	1.70	7.60	7.79	2.40	2.70	3.61
Oct. 8 15:00	39.00	142.30	-3.70	-3.70	7.97	3.50	5.10	6.19	2.10	-18.80	18.92	3.70	5.10	6.30
Oct. 8 21:00	40.00	143.60	4.30	4.30	5.81	4.00	5.30	6.64	1.60	4.50	4.78	4.20	3.10	5.22
Average	—		4.39			4.02			9.56			3.58		

Particle Filters works better than other methods

Discussions of Experiments

- Particle filters performs better than other methods
- If the center of a target event is in an oceanic area, it's more difficult to locate it precisely from tweets
- It becomes more difficult to make good estimation in less populated areas

Earthquake Reporting System

- Toretter (<http://toretter.com>)
 - Earthquake reporting system using the event detection algorithm
 - All users can see the detection of past earthquakes

– Re Dear Alice,

ea We have just detected an earthquake
around Chiba. Please take care.

Toretter Alert System

Screenshot of Toretter.com



The screenshot shows the Toretter.com website with a header featuring the logo, the text '日本の地震' (Earthquake in Japan), and a cartoon fish. Below the header is a table of earthquake reports. Each row contains a timestamp, location, title in Japanese, a screen name, and a URL. English translations of the Japanese titles are provided in speech bubbles.

Published	Location	Title	Screen_name	URL
2009-08-11 05:08:57	Saitama, Japan	地震おおいわー	tondol	http://twitter.com/tondol
2009-08-11 05:08:56	unknown	地震。	tirolly	http://twitter.com/tirolly
2009-08-11 05:08:53	iPhone: 35.509506,139.615601	揺れたね	Hakkan	http://twitter.com/Hakkan
2009-08-11 05:08:53	Mie Prefecture	すごい地震だ【mb】	narude531 masu	http://twitter.com/narude531 masu
2009-08-11 05:08:52	Kawasaki city	地震だ！！	yaketasamma	http://twitter.com/yaketasamma
2009-08-11 05:08:52	unknown	地震こわいですかんぺん	wzzc	http://twitter.com/wzzc
2009-08-11 05:08:52	Kansai	あら、地震？	Haru_Iro	http://twitter.com/Haru_Iro
2009-08-11 05:08:52	Sakado, Saitama, Japan	地震だ	d_wackys	http://twitter.com/d_wackys
2009-08-11 05:08:51	unknown	愛知も揺れたw	edomain	http://twitter.com/edomain
2009-08-11 05:08:51	unknown	また地震 長いな...	laukaz	http://twitter.com/laukaz
2009-08-11 05:08:51	JP	地震なう	echomun	http://twitter.com/echomun

Q: What would be a key feature of this system?

Earthquake Reporting System

- Effectiveness of alerts of this system
 - Alert E-mails urges users to prepare for the earthquake if they are received by a user **shortly before** the earthquake actually arrives.
- Is it possible to receive the e-mail before the earthquake actually arrives?
 - An earthquake is transmitted through the earth's crust at about **3~7 km/s**.
 - a person has about **20~30 sec** before its arrival at a point that is **100 km distant** from an actual center

Results of Earthquake Detection

Date	Magnitude	Location	Time	E-mail sent time	time gap [sec]	# tweets within 10 minutes	Announce of JMA
Aug. 18	4.5	Tochigi	6:58:55	7:00:30	95	35	7:08
Aug. 18	3.1	Suruga-wan	19:22:48	19:23:14	26	17	19:28
Aug. 21	4.1	Chiba	8:51:16	8:51:35	19	52	8:56
Aug. 25	4.3	Uraga-oki	2:22:49	2:23:21	31	23	2:27
Aug.25	3.5	Fukushima	2:21:15	22:22:29	73	13	22:26
Aug. 27	3.9	Wakayama	17:47:30	17:48:11	41	16	1:7:53
Aug. 27	2.8	Suruga-wan	20:26:23	20:26:45	22	14	20:31
Ag. 31	4.5	Fukushima	00:45:54	00:46:24	30	32	00:51
Sep. 2	3.3	Suruga-wan	13:04:45	13:05:04	19	18	13:10
Sep. 2	3.6	Bungo-suido	17:37:53	17:38:27	34	3	17:43

In all cases, we sent E-mails before announces of JMA (Japan Meteorological Agency)

In the earliest cases, it can sent E-mails in **19 sec**.

On average, JMA announces **6 min** after the earthquake

Experiments And Evaluation

- We demonstrate performances of
 - tweet classification
 - event detection from time-series data
 - show this results in “application”
 - location estimation from a series of spatial information

Results of Earthquake Detection

JMA intensity scale	2 or more	3 or more	4 or more
Num of earthquakes	78	25	3
Detected	70(89.7%)	24(96.0%)	3(100.0%)
Promptly detected*	53(67.9%)	20(80.0%)	3(100.0%)

JMA intensity scale: the original scale of earthquakes by Japan Meteorology Agency
Promptly detected: detected in a minute

Period: Aug.2009 – Sep. 2009
Tweets analyzed : 49,314 tweets
Positive tweets : 6291 tweets by 4218 users

It detected **96%** of earthquakes that were stronger than scale 3 or more during the period.

Conclusions

- It investigated **the real-time nature of Twitter** for event detection
- **Semantic analysis** was applied to tweets classification
- It considers each Twitter user as a sensor and set a problem to detect an event based on **sensory observations**
- Location estimation methods such as **Kaman filters and particle filters** are used to estimate locations of events
- It developed **an earthquake reporting system**, which is a novel approach to notify people promptly of an earthquake event
- Authors plan to expand their system to **detect events of various kinds** such as rainbows, traffic jam etc.

Social Events

Paper2: "From tweets to polls: Linking text sentiment to public opinion time series." O'Connor, Brendan, et al. ICWSM 11 (2010): 122-129.



Measuring public opinion through social media?

People in U.S.

Query



Think about three possible benefits of using social media to measure public opinion

Can we derive a similar measurement?



Write



Query

Aggregate
Text Sentiment
Measure

Contributions

- Correlations between
 1. Very simple text sentiment analysis
 2. Telephone public opinion polls
 - Consumer confidence and Presidential job approval
- Negative results as well!
- Also
 - Time-series smoothing is a critical issue
 - Topic selection, topic volumes, text leads polls, stemming, election polling

Text Data: Twitter

- A large and public social media service
 - People volunteer their thoughts on various topics at a large scale
- Sources
 1. Archiving the Twitter Streaming API
 - “Gardenhose”/”Sample”: ~10-15% of public tweets
 2. Scrape of earlier messages
- ~2 billion messages before topic selection, Jan 2008 – May 2010

It did not justify tweet users are good representations of the general population.

Message data

```
{
  "text": "Time for the States to fight back !!!      Tenth Amendment Movement: Taking On the
Feds http://bit.ly/14t1RV  #tcot #teaparty",
  "created_at": "Tue Nov 17 21:08:39 +0000 2009",
  "geo": null,
  "id": 5806348114,
  "in_reply_to_screen_name": null,
  "in_reply_to_status_id": null,

  "user": {
    "screen_name": "TPO_News",
    "created_at": "Fri May 15 04:16:38 +0000 2009",
    "description": "Child of God - Married - Gun carrying NRA Conservative - Right Winger
hard Core Anti Obama (Pro America), Parrothead - www.ABoldStepBack.com #tcot #nra #iPhone",
    "followers_count": 10470,
    "friends_count": 11328,
    "name": "Tom O'Halloran",
    "profile_background_color": "f2f5f5",
    "profile_image_url": "http://a3.twimg.com/profile_images/295981637/
TPO_Balcony_normal.jpg",
    "protected": false,
    "statuses_count": 21147,
    "location": "Las Vegas, Baby!!",
    "time_zone": "Pacific Time (US & Canada)",
    "url": "http://www.tpo.net/1dollar",
    "utc_offset": -28800,
  }
}
```

Message data used

```
{  
  "text": "Time for the States to fight back !!!    Tenth Amendment Movement: Taking On the  
Feds http://bit.ly/14t1RV    #tcot #teaparty",  
  "created_at": "Tue Nov 17 21:08:39 +0000 2009"  
}
```

1. Text
2. Timestamp

- Message data not used:
 - Locations from GPS
 - *Locations from IP addresses – not public*
 - User information (name, description, self-described location)
 - Conversation structure: retweets, replies
 - Social structure: follower network

Poll Data

- Consumer confidence, 2008-2009
 - Index of Consumer Sentiment (Reuters/Michigan)
 - Gallup Daily (free version from gallup.com)
- 2008 Presidential Elections
 - Aggregation, Pollster.com
- 2009 Presidential Job Approval
 - Gallup Daily

Poll Data: Sample Questions

- “We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?”
- “Now turning to business conditions in the country as a whole—do you think that during the next twelve months we’ll have good times financially, or bad times?”

*Ask human sensors what they are good at:
Binary Observations!*

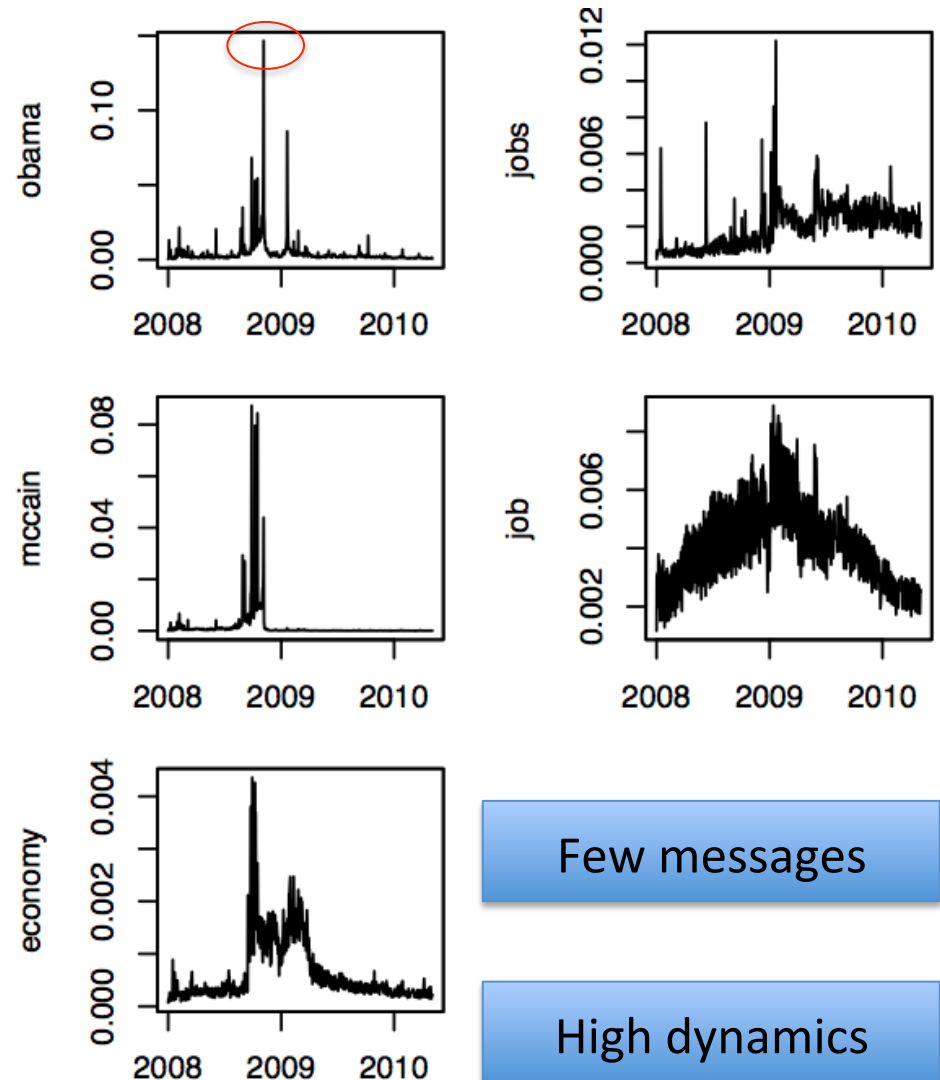
Text sentiment analysis method

- Two-stage process
 1. Topic Selection: select topical tweets, via hand-selected keywords
 2. Sentiment Analysis: For one topic' s messages, count “sentiment polarity” words, from pre-existing lexicon

(More sophisticated methods are possible, for both stages)

Message selection via topic keywords

- Analyzed subsets of messages that contained manually selected topic keyword
 - “economy”, “jobs”, “job” -> **consumer confidence**
 - “obama” -> **presidential approval**
 - “obama”, “mccain” -> **elections**
- High day-to-day dynamics
 - Fraction of messages containing keyword
 - Nov 5 2008: 15% of tweets contain “obama”



Opinion Estimation: Word Counting

- Subjectivity Clues lexicon from OpinionFinder (Univ. of Pittsburgh)

<http://mpqa.cs.pitt.edu/opinionfinder/>

- Wilson et al 2005
 - 1600 positive, 1200 negative words

Sentiment
Analysis Tool

- Procedure
 1. Within topical messages,
 2. Count messages containing these positive and negative words

A note on the sentiment list

- This list is not well suited for social media English.

– “sucks”, “ :) ”, “ :(”, “lol”

False Negatives

- Examples for one day.

(Top examples)

<i>word</i>	<i>valence</i>	<i>count</i>
will	positive	3934
bad	negative	3402
good	positive	2655
help	positive	1971

False Positives

(Random examples)

<i>word</i>	<i>valence</i>	<i>count</i>
funny	positive	114
fantastic	positive	37
cornerstone	positive	2
slump	negative	85
bearish	negative	17
crackdown	negative	5

Sentiment Ratio over Messages

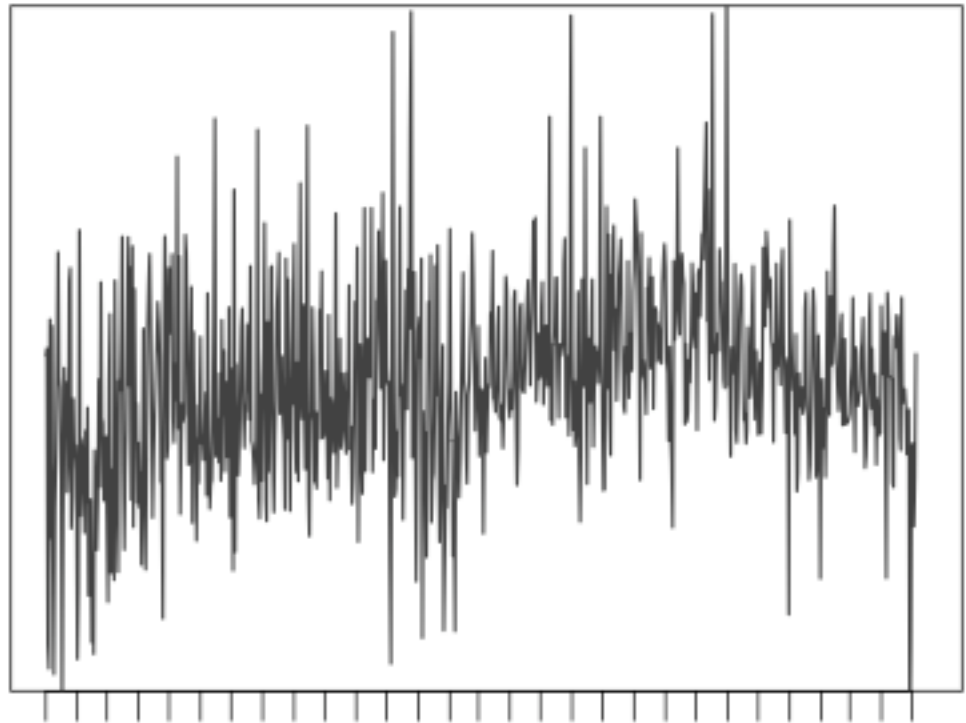
For **one day** t and a **particular *topic word***, compute score

SentimentRatio($topic_word, t$) =

$$\begin{aligned} & \frac{\text{MessageCount}_t(\text{pos. word AND topic word})}{\text{MessageCount}_t(\text{neg. word AND topic word})} \\ = & \frac{p(\text{pos. word} \mid \text{topic word}, t)}{p(\text{neg. word} \mid \text{topic word}, t)} \end{aligned}$$

Sentiment Ratio Moving Average

- High day-to-day dynamics.
- Average last k days.
- SentRatio (“jobs”),
 $k = 1$
- (Gallup tracking polls: 3 or 7-day smoothing)

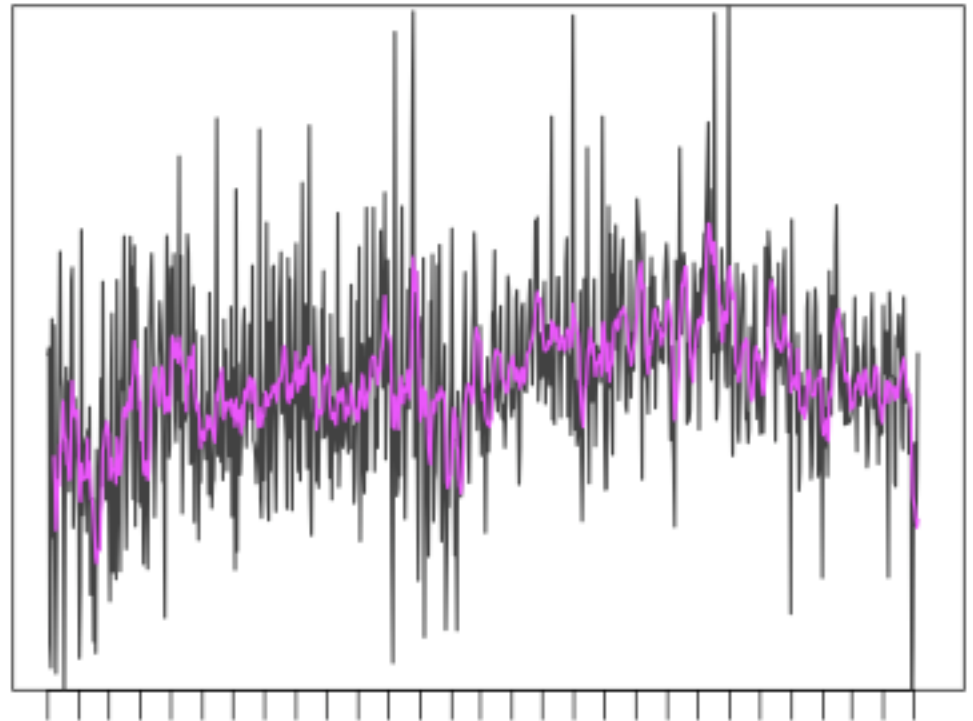


Q: How to observe consistent behaviors over a longer period of time?

$$MA_t = \frac{1}{k} (x_{t-k+1} + x_{t-k+2} + \dots + x_t)$$

Sentiment Ratio Moving Average

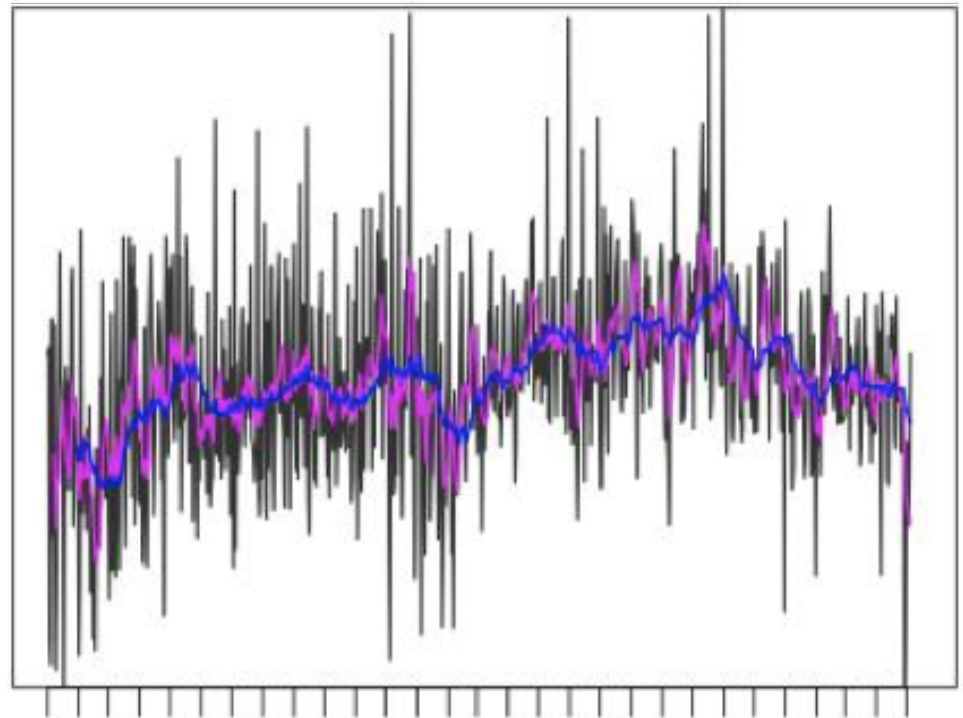
- High day-to-day volatility.
- Average last k days.
- SentRatio("jobs"),
 $k = 1, 7$
- (Gallup tracking polls: 3 or 7-day smoothing)



$$MA_t = \frac{1}{k} (x_{t-k+1} + x_{t-k+2} + \dots + x_t)$$

Sentiment Ratio Moving Average

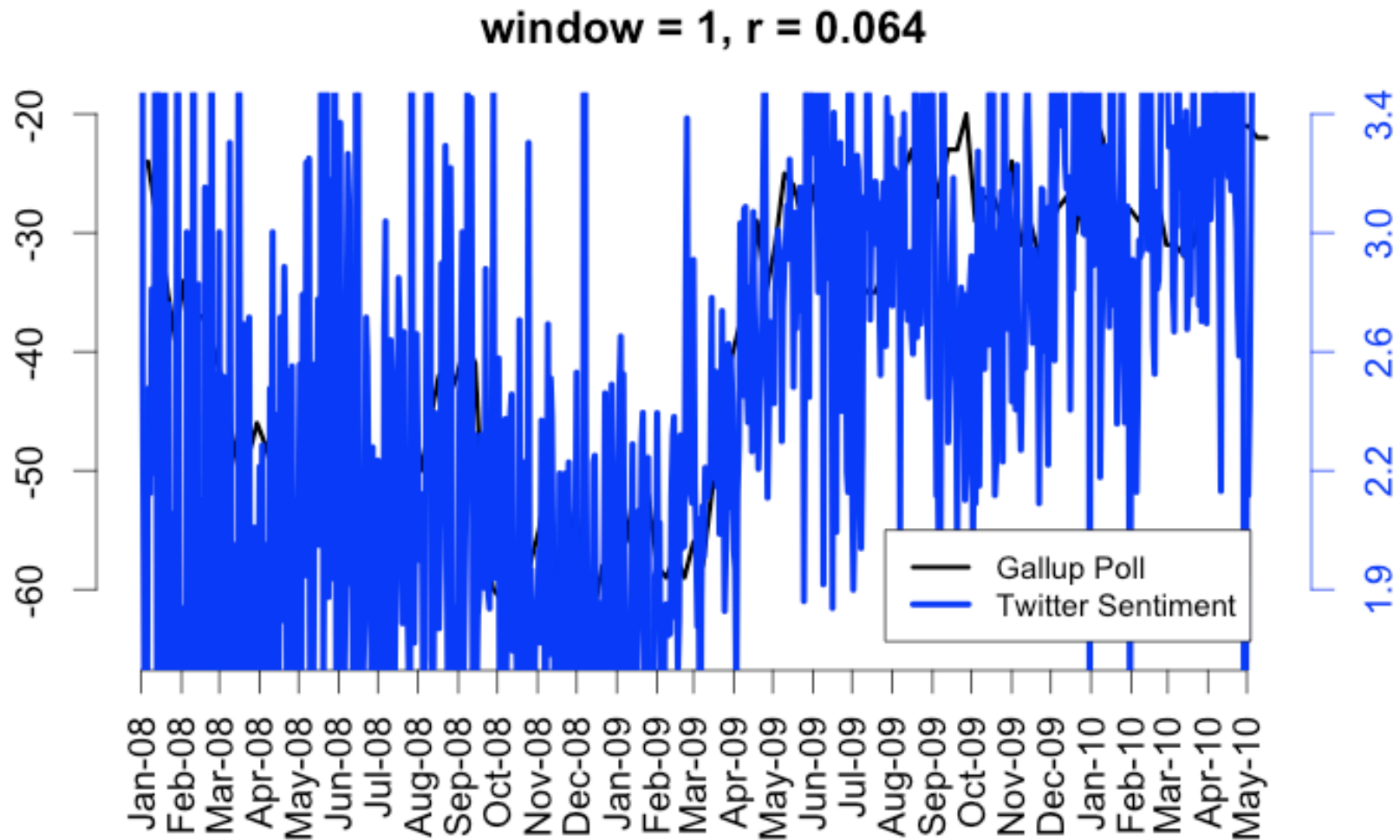
- High day-to-day volatility.
- Average last k days
- SentRatio("jobs"),
k = 1, 7, 30
- (Gallup tracking polls: 3 or 7-day smoothing)



$$MA_t = \frac{1}{k} (x_{t-k+1} + x_{t-k+2} + \dots + x_t)$$

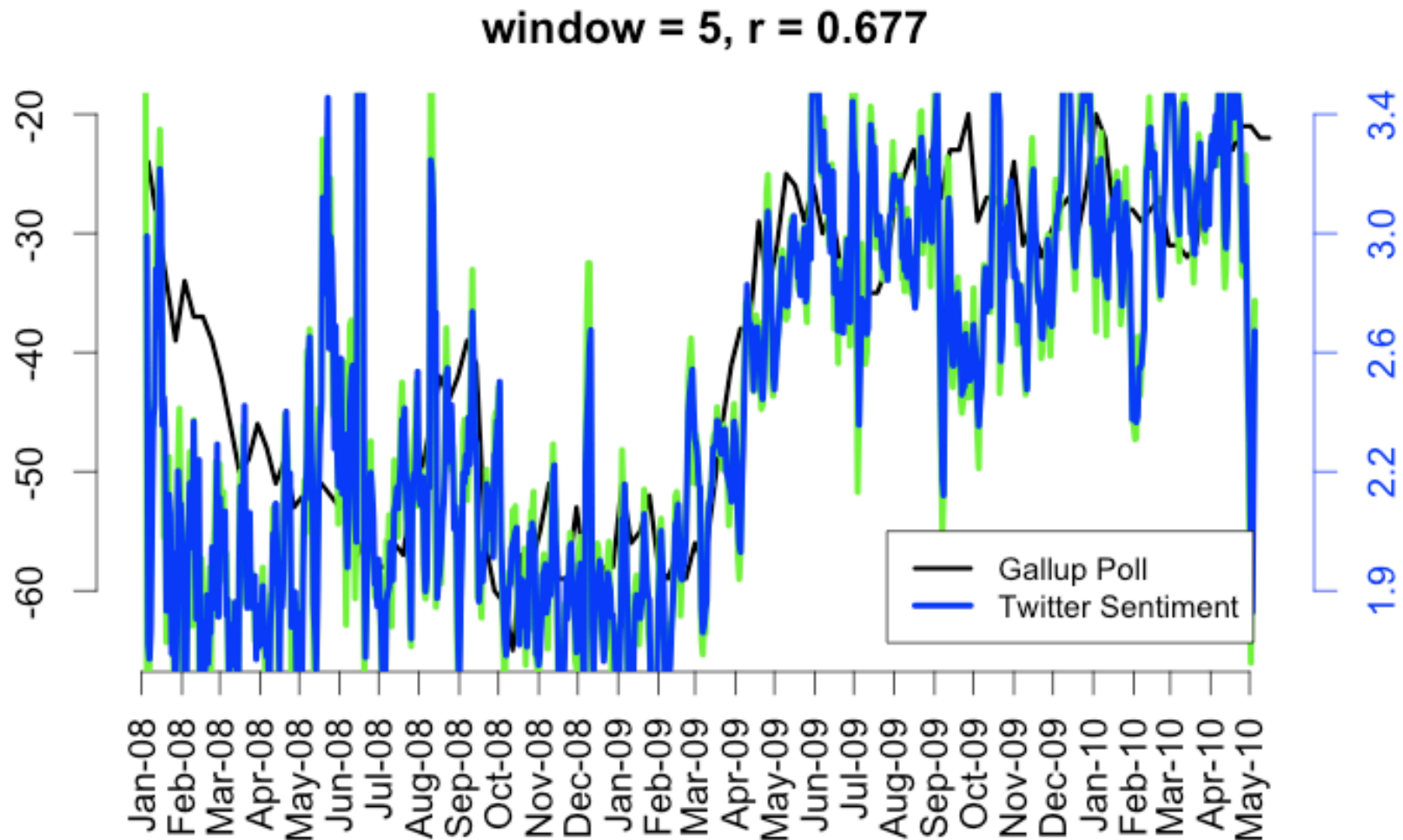
Smoothed comparisons

SentimentRatio("jobs")



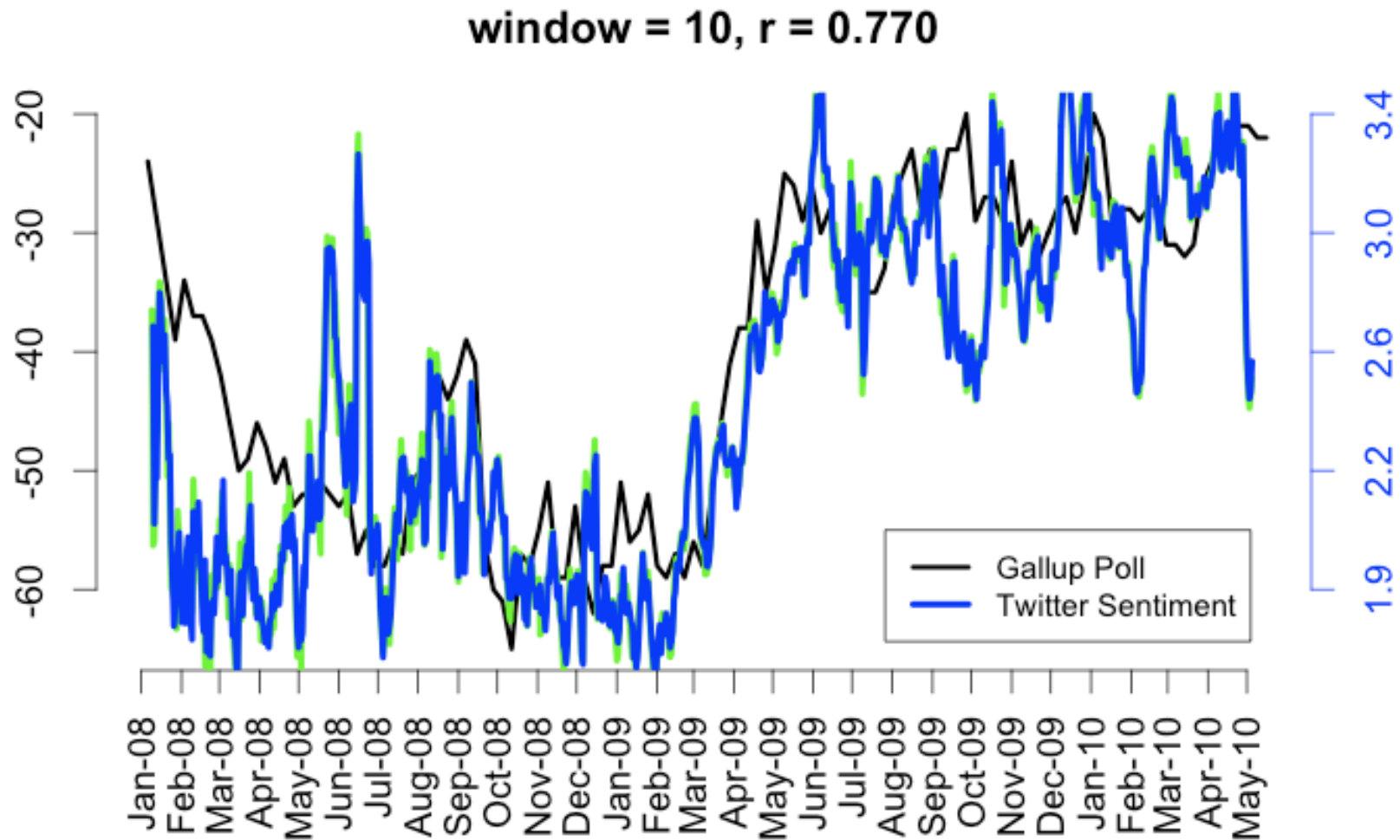
Smoothed comparisons

SentimentRatio("jobs")



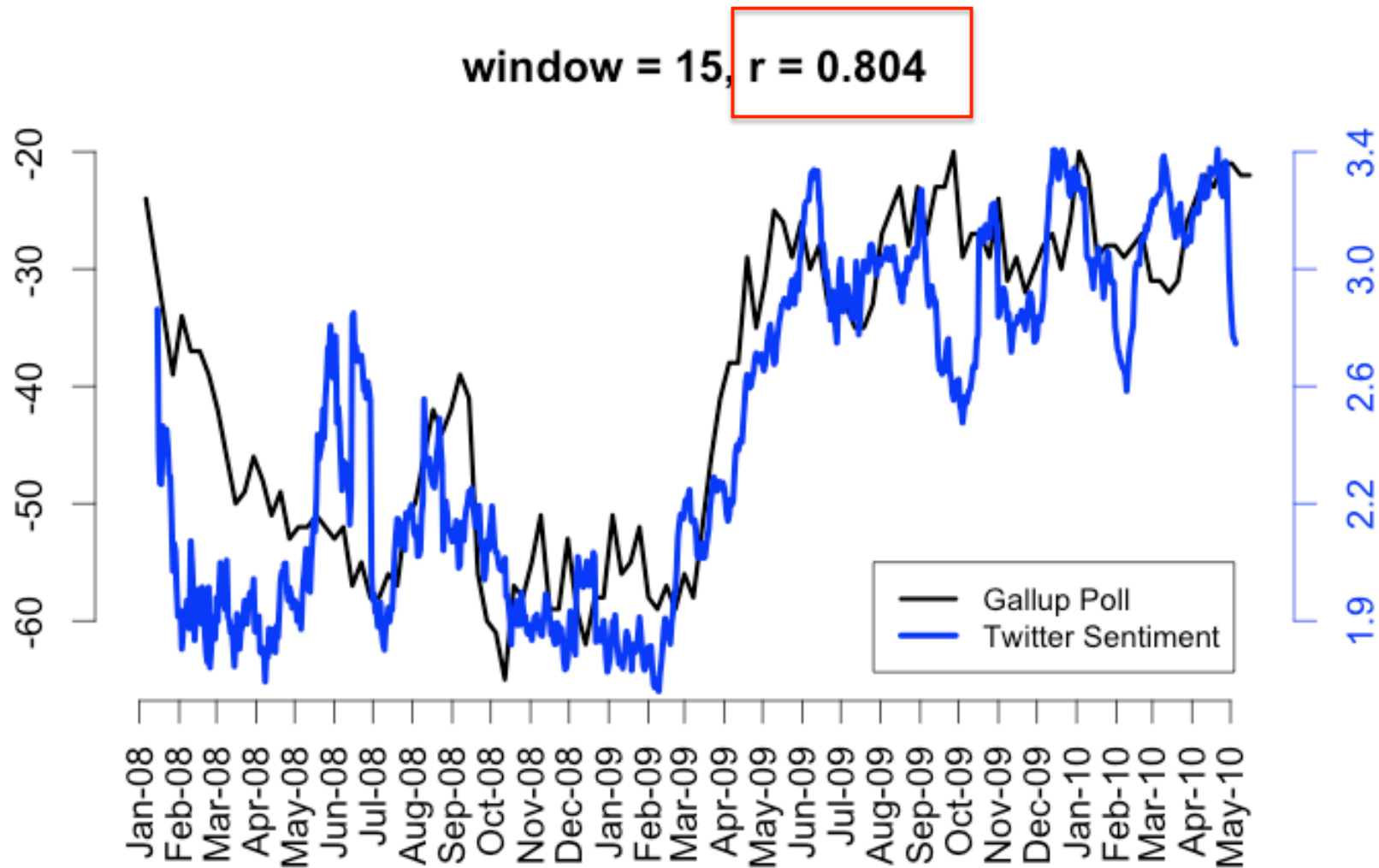
Smoothed comparisons

SentimentRatio("jobs")



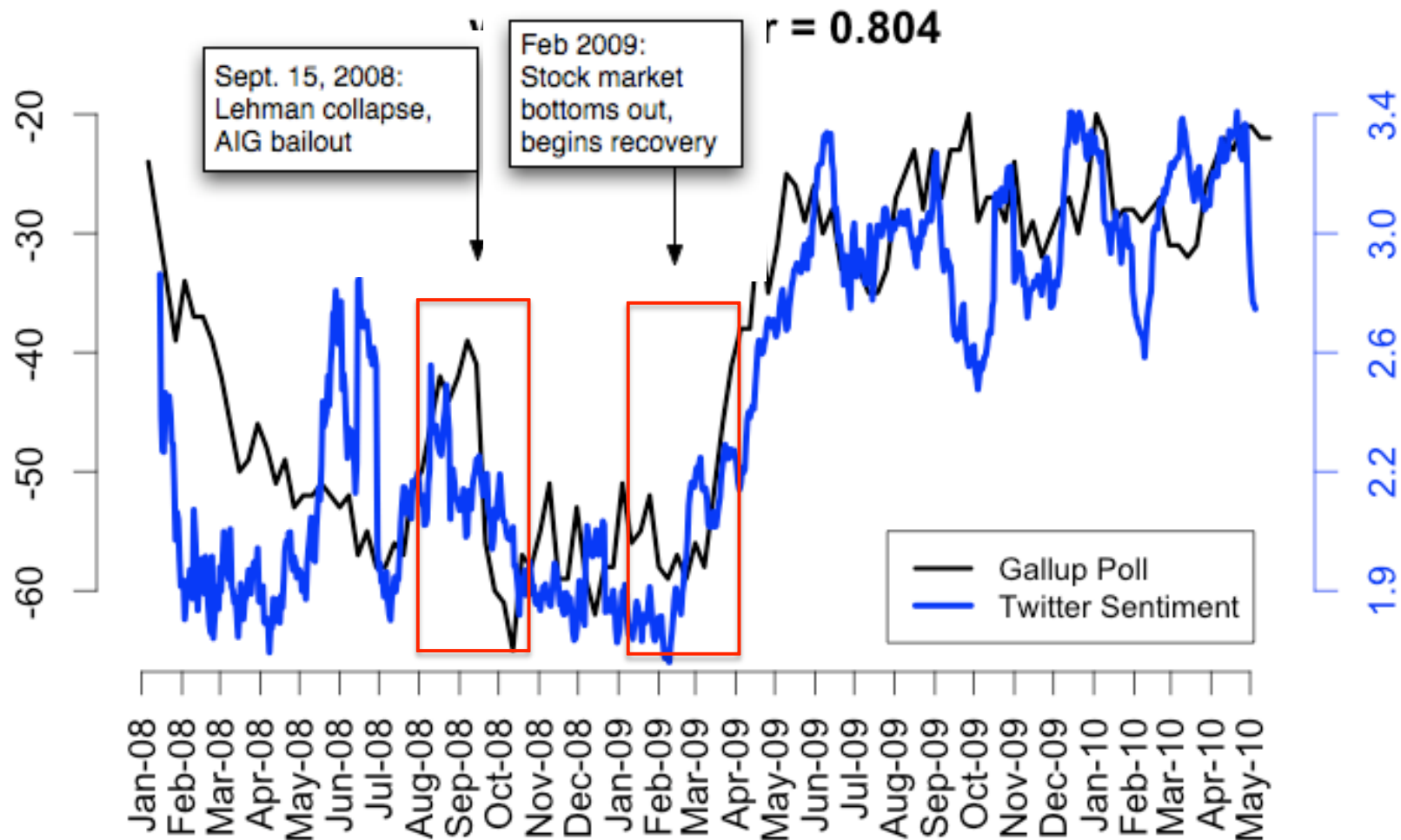
Smoothed comparisons

SentimentRatio("jobs")



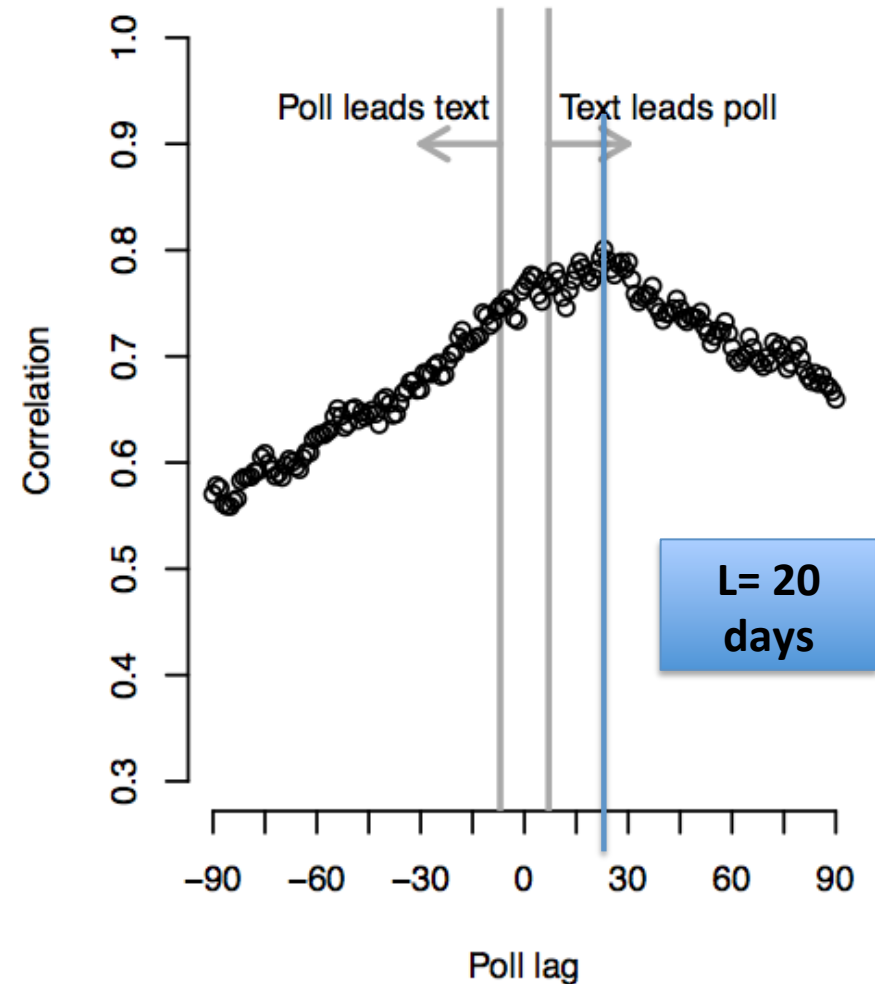
Smoothed comparisons

SentimentRatio("jobs")

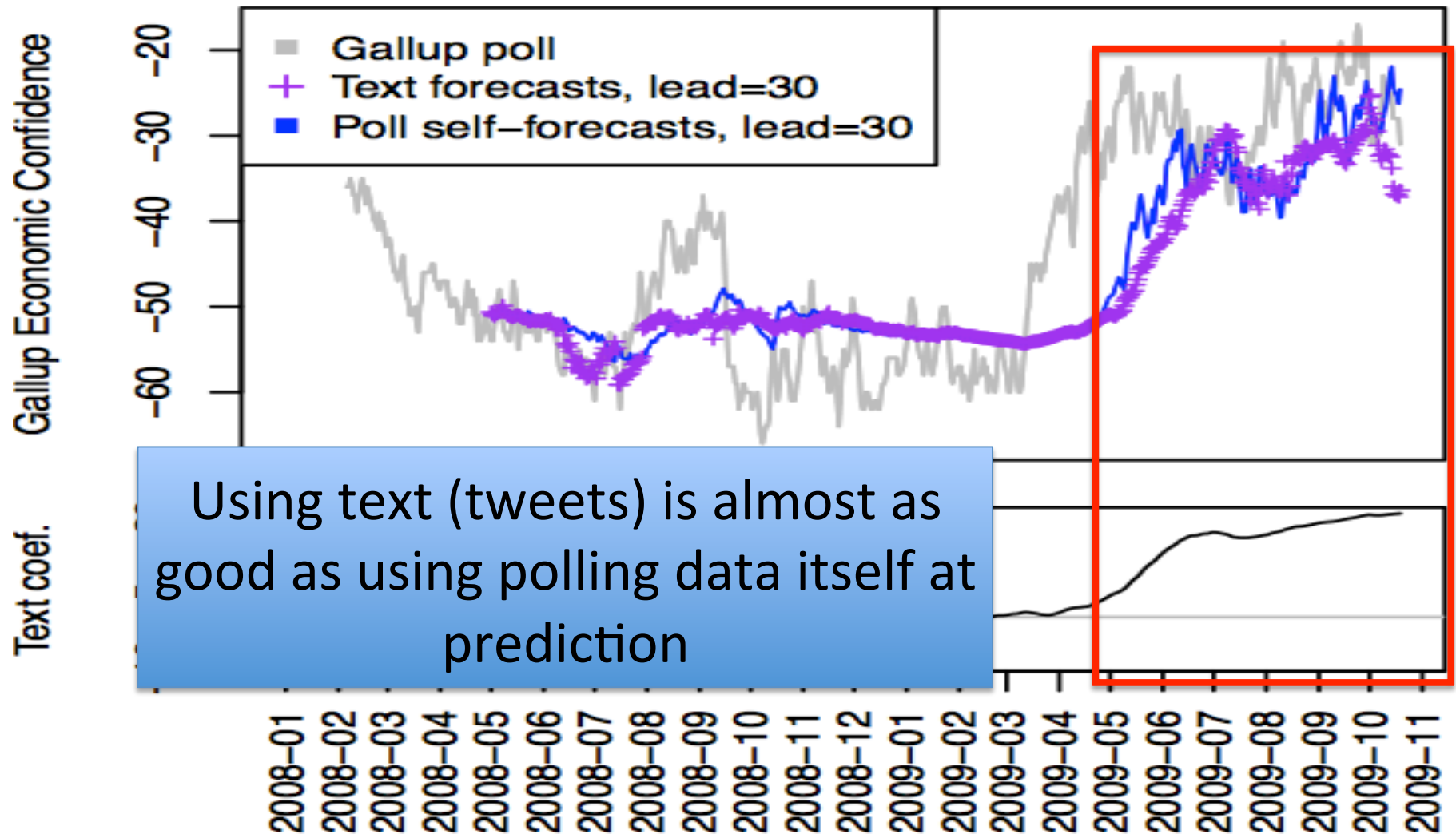


Which leads, poll or Twitter?

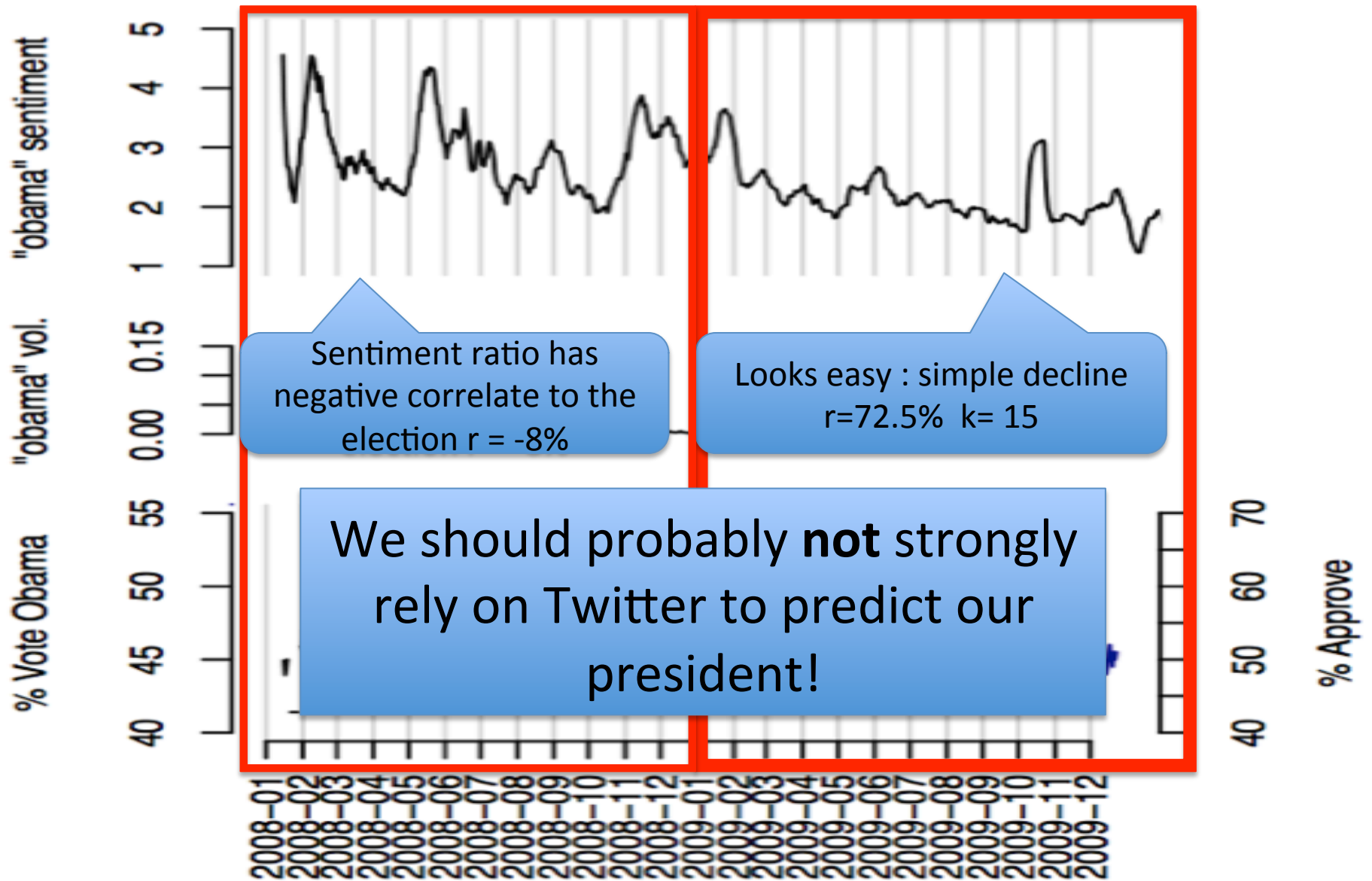
- Cross-correlation analysis: between
 - SentimentRatio(“jobs”) on day t
 - Poll for day $t+L$
- Twitter text is a leading indicator for polls



Predicting polls



Presidential elections and job approval



Remained Challenges

- Who is using Twitter?
 - Massive changes over time (2009 Twitter != 2015 Twitter)
- Is Polling the Gold Standard?
 - Other more reliable sources (e.g., face-to-face interviews)
- Better text analysis
 - How to handle retweets and duplicated information (next lecture)
 - Word sense ambiguity: “steve jobs”
- Less biased data
 - Use demographic information on Twitter to reduce bias