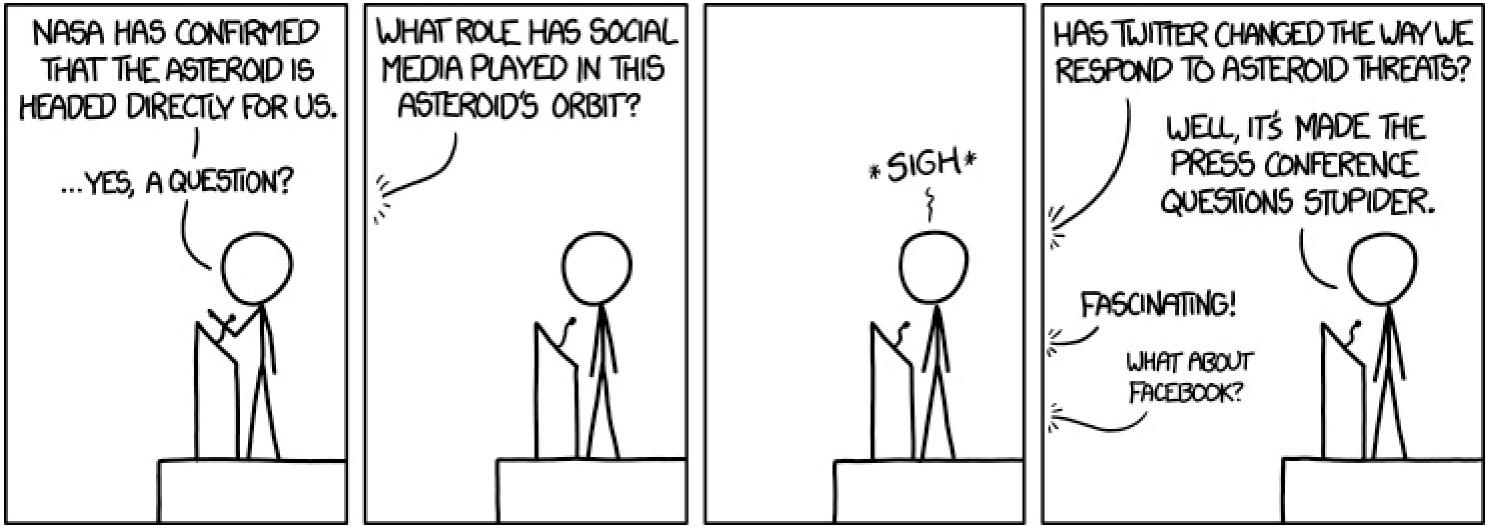
Mining Social Media

Class Algorithmic Methods of Data Mining
Program M. Sc. Data Science
University Sapienza University of Rome
Semester Fall 2015
Lecturer Carlos Castillo http://chato.cl/

Sources:

- Most slides here are from “Twitter and the Real World” CIKM'13 Tutorial and references therein [link]
- See also the 2015 book “Twitter: A Digital Socioscope” by Mejova, Weber, and Macy

Social media changes *everything*



https://xkcd.com/1239/

Digital Humanities

- Research in social science may be supported by social media data
- Many in science, technology and engineering have also interest in the humanities
  - Plus a bit of actual formal education on the subject
  - Plus a ton of intuitions, a few of them correct

Cut-off between 1 and 2 is 30 years old for social media.

An attractive topic

- Social media is a “young” technology (~10 to 15 years old)
- Douglas Adams on new technologies:
  - Anything that is in the world when you're born is normal and ordinary and is just a natural part of the way the world works.
  - Anything that's invented between when you're 15 and 35 is new and exciting and revolutionary and you can probably get a career in it.
  - Anything invented after you're 35 is against the natural order of things.

Definitions

- Social software
  - Software to facilitate or mediate social interactions
- Social networking sites
  - Web applications to maintain social connections
- Social media sites
  - Web applications to create, share, and exchange content
- Social media content
  - The content shared by users in social media platforms

Why mining social media?

- “What do people think / how do they feel about X?”
  - Sentiment analysis and opinion mining
- An alternative to traditional opinion polls?
- Attractive for many reasons including:
  - Lower latency (waiting time)
  - Lower cost
  - Larger population

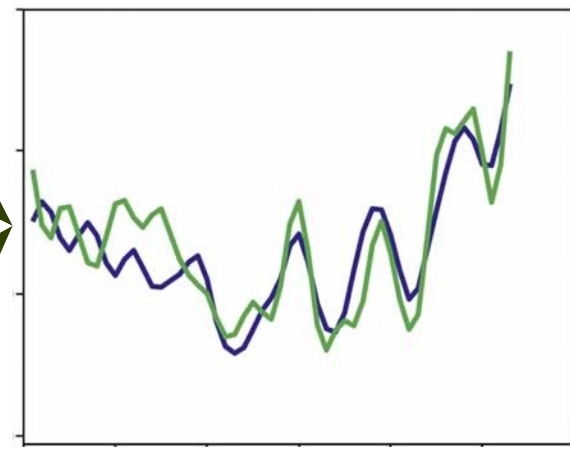
Template: Google Flu Trends

Low latency: people query symptomps before going to the doctor!

Many social media mining papers

Domain-specific data

Correlation/Influence



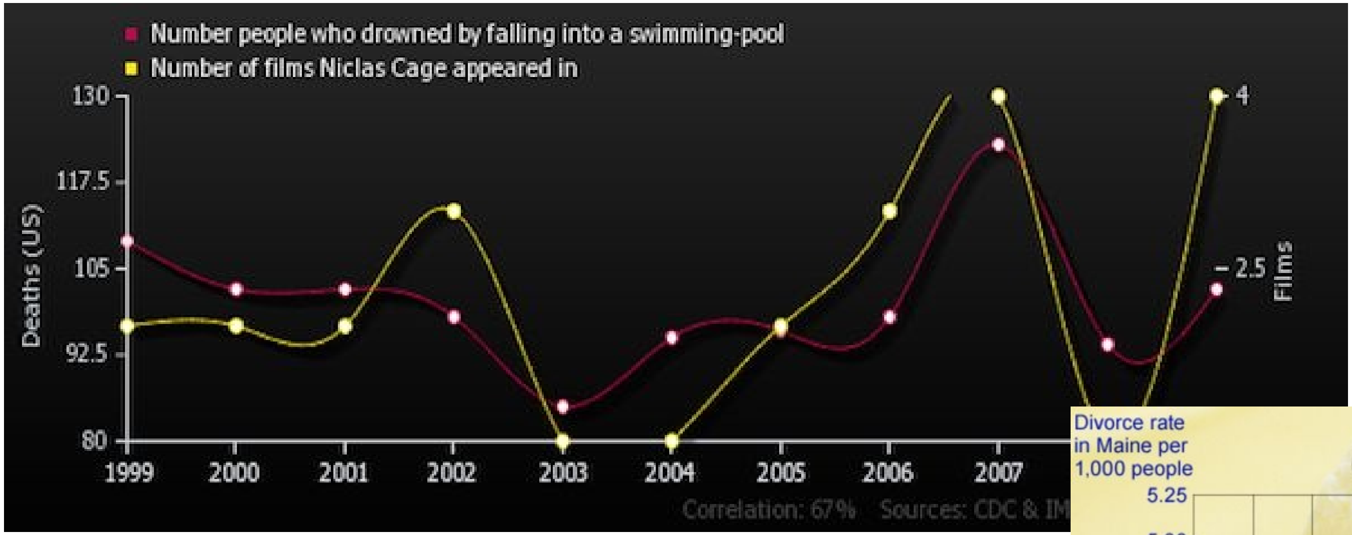
Profit?

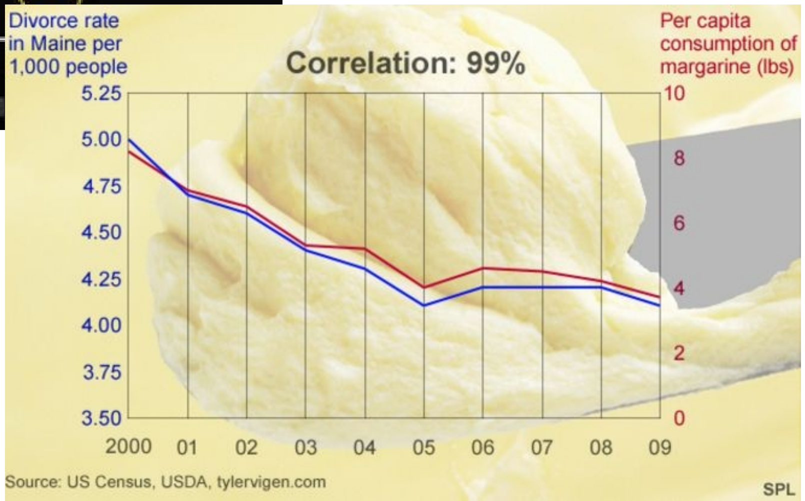
Social media data

The devil is in the details

- Which domain-specific data? This is not always readily available
- Mapping social media data to a time series?
  - Geolocation of messages
  - Mapping to topics/sentiments/intents or other characteristics
  - What is the variable: Volume? Sentiment? Other?
- Measuring correlation/influence
  - Correlation (lagged); Transfer entropy
- Finding a mechanism

Caveat 1: correlation might be spurious



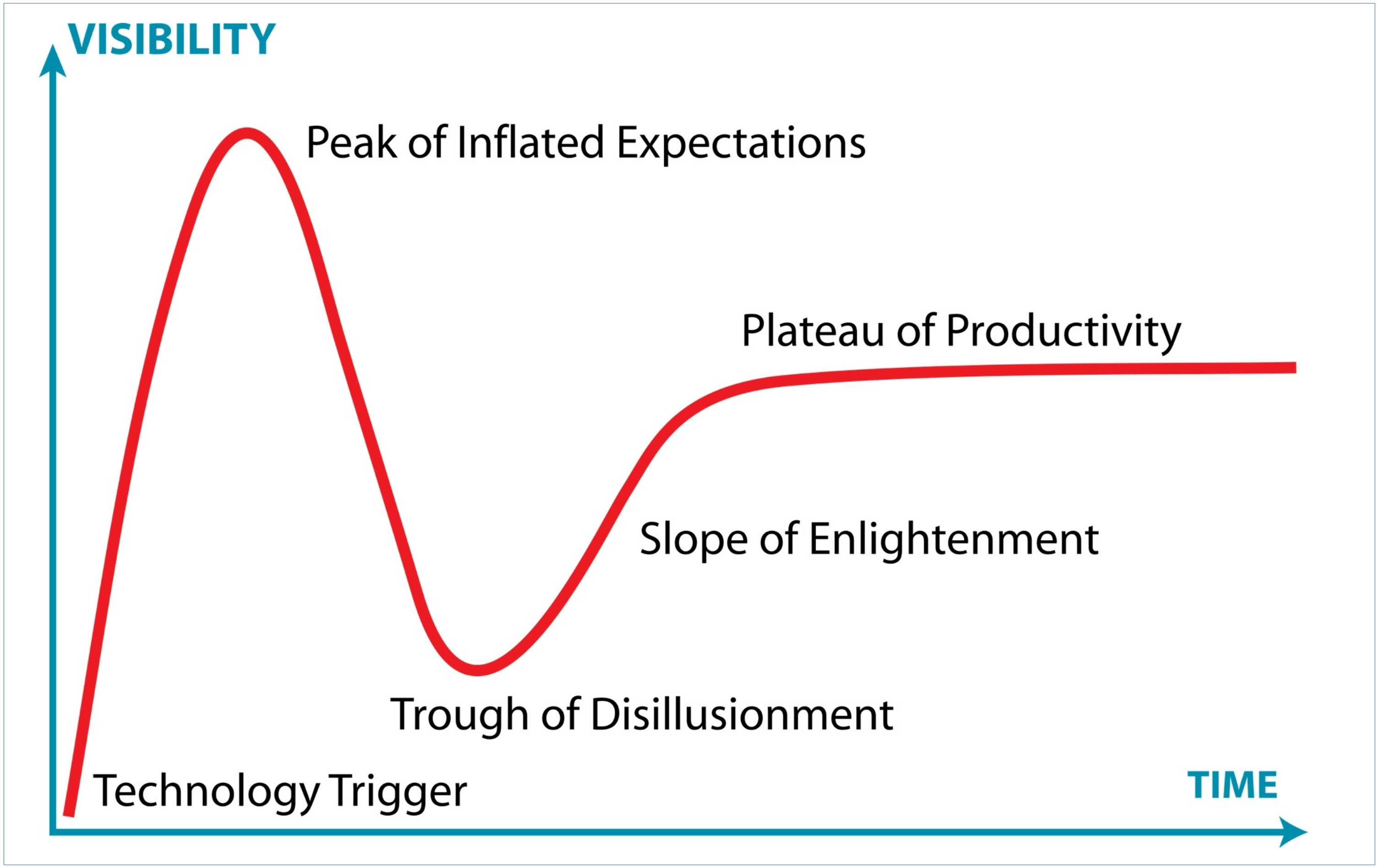


Caveat 2: correlation might be useless

- Sometimes there are much better predictors
- Social media can be used to predict box office revenue
  - But ticket sales on first weekend *almost* always determine total sales, with exceptions: Citizen Kane (1941), Blade Runner (1982), Fight Club (1999)
- Social media can be used to detect earthquakes
  - But seismographic sensors are quite dense in many areas of the world, the exception being underdeveloped areas

Caveat 3: the “war on terror”

The Hype Curve



Examples in the 1980s: (a) AI, (b) online learning.

Example social media mining topics

- Economics
- Politics
- Public health
- Smart cities
- Event detection

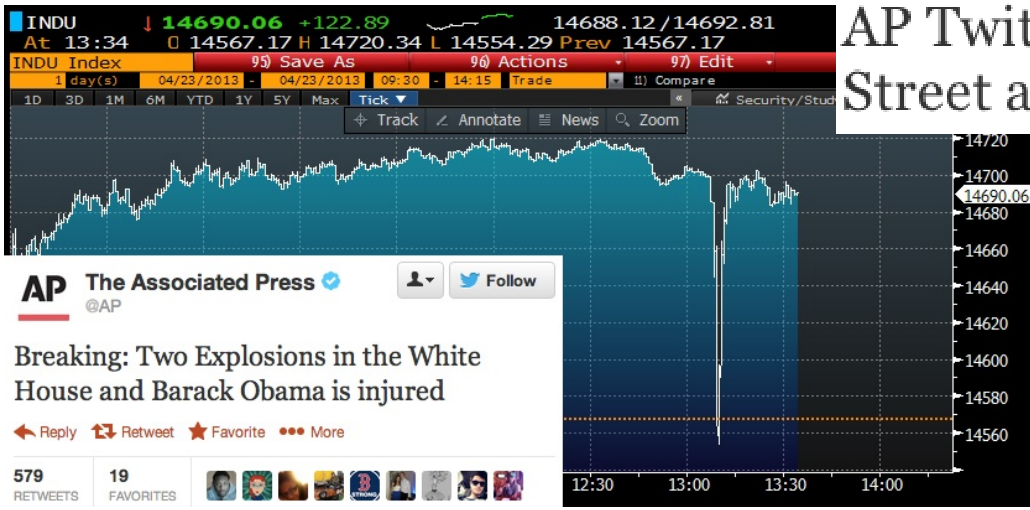
Most examples on this section come from https://sites.google.com/site/twitterandtherealworld/home

Economics

Examples in economics

- Financial success of movies
- Economic indices such as DJIA or NASDAQ
  - Words related to anxiety/worry/calmness/hope
- Stock option prices
  - Centrality in interaction graphs

When Twitter sneezes, the stock market gets the flu ...



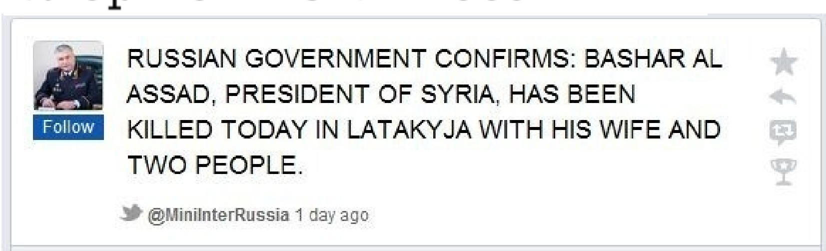
AP Twitter hack causes panic on Wall Street and sends Dow plunging

During those three minutes, the "fake tweet erased $136 billion in equity market value,"

- Bloomberg News

http://www.washingtonpost.com/blogs/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/

Twitter Death Rumor Leads to Spike in Oil Prices



http://mashable.com/2012/08/07/twitter-rumor-oil-price/

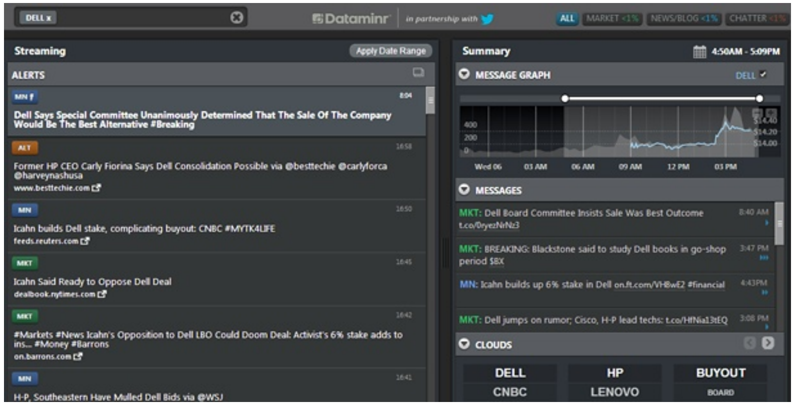
Netflix CEO's Facebook Post Triggered SEC Wells Notice

http://www.cnbc.com/id/100289227

Bloomberg

http://nymag.com/daily/intelligencer/2013/04/bloombergs-vip-terminal-tweeters.html

Dataminr™



http://dataminr.com/

2. specialized providers

3. data analytics

1. content providers

4. traders

Cayman Atlantic
INVESTMENT MANAGEMENT

GNIP

tumblr.

foursquare

WordPress

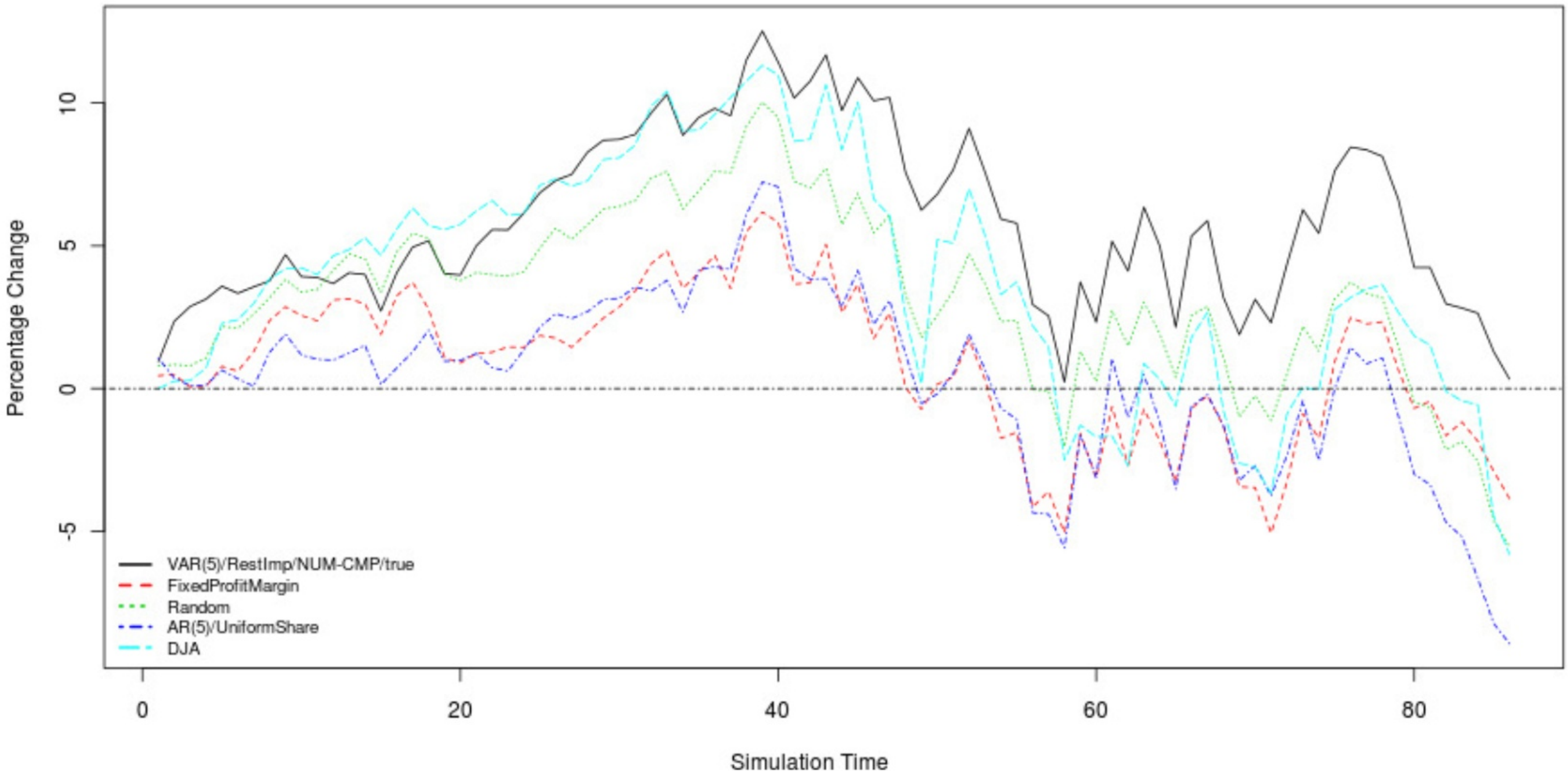
StockTwits

estimize

http://gnip.com/

Self-reported Gains http://www.caymanatlantic.com/

Trading stock using social media



http://dx.doi.org/10.1145/2124295.2124358

Why you can't get rich using this

Efficient Market Hypothesis:

Financial markets are **information efficient**: prices fully reflect all available information

Cannot be predicted

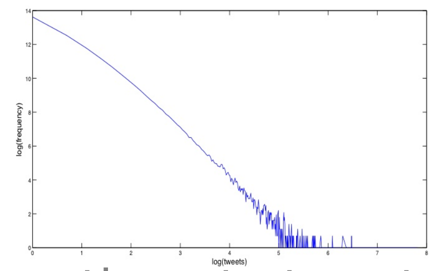
Movies!

Predicting the Future with Social Media

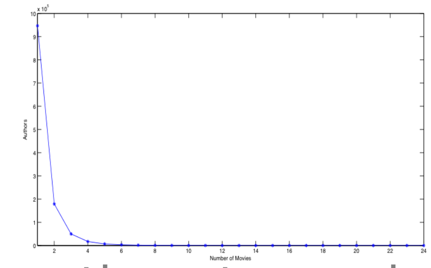
@sitaramasur Asur, Huberman @ WI-IAT 2010

- 2.89 million tweets
- 24 movies (manually compiled keywords)

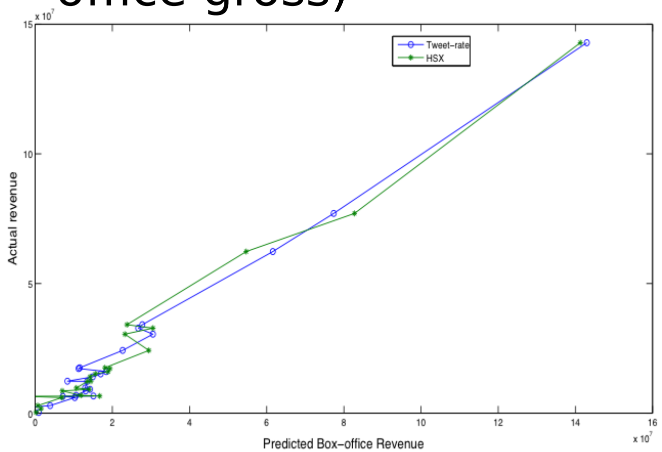
Correel (tweet rate & box office gross) **= 0.90**



authors to tweets



authors to movies

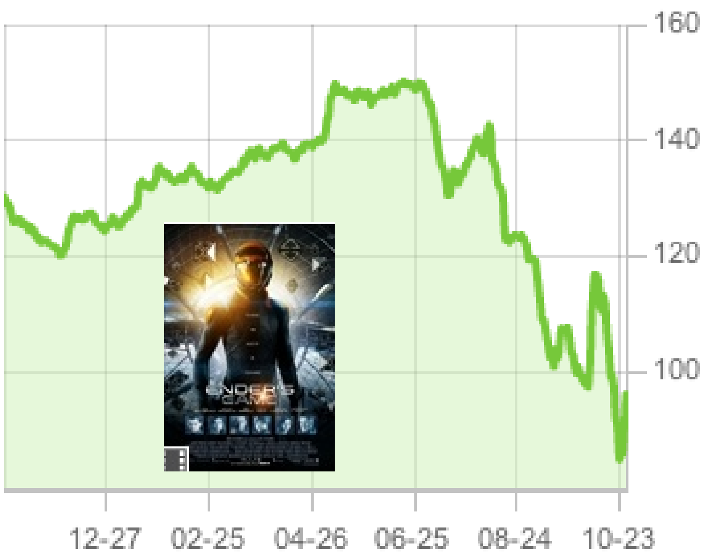


predicted vs actual box office scores

least squares linear regression using previous week's tweets to predict weekend box office gross:

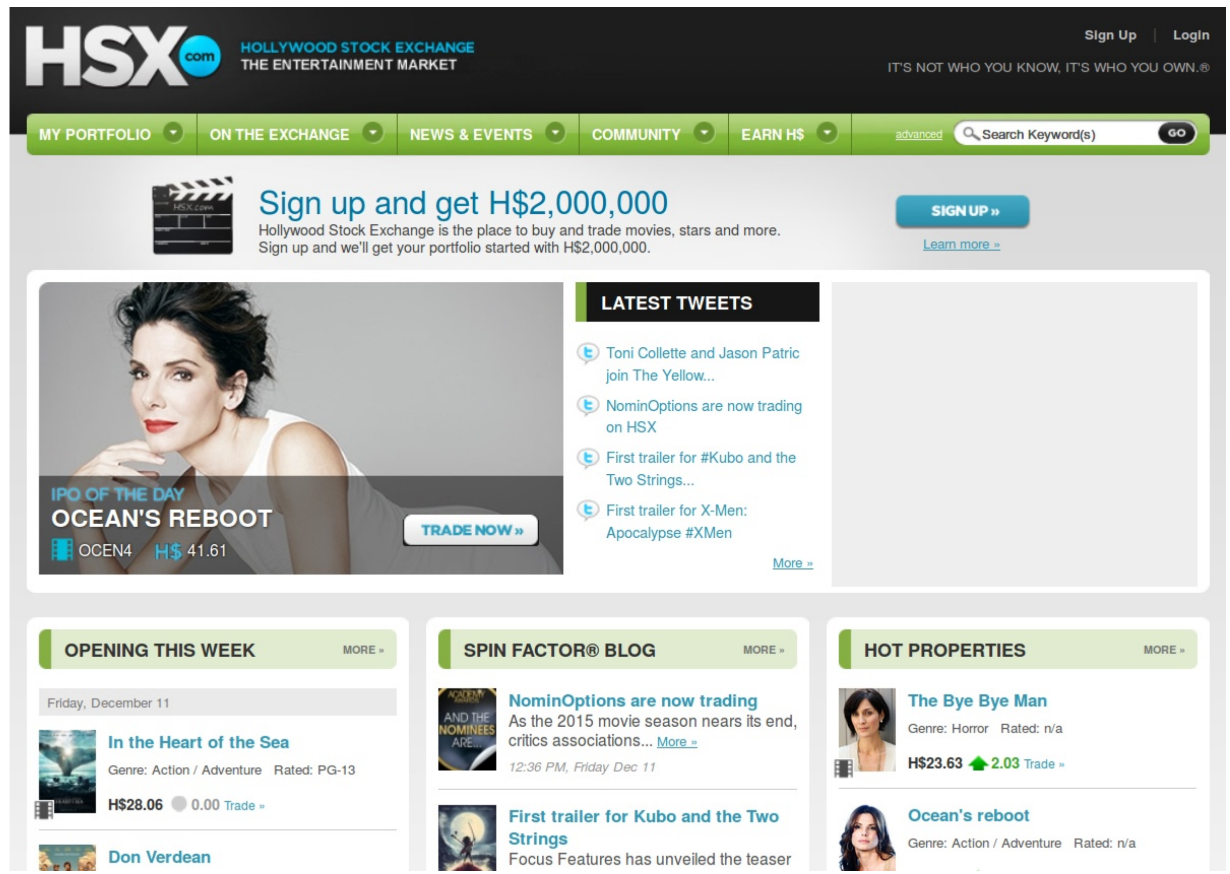
	Adj R^2
Average Tweet-rate	0.80
Tweet-rate time series	0.93
Tweet-rate time series + theater count	**0.973**
HSX time series + theater count	0.965

Movies!



160
140
120
100
12-27
02-25
04-26
06-25
08-24
10-23

THE GREATEST TURKEY MOVIE OF ALL TIME
FREEBIRDS
50
40
30
20
12-27
02-25
04-26
06-25
08-24
10-23



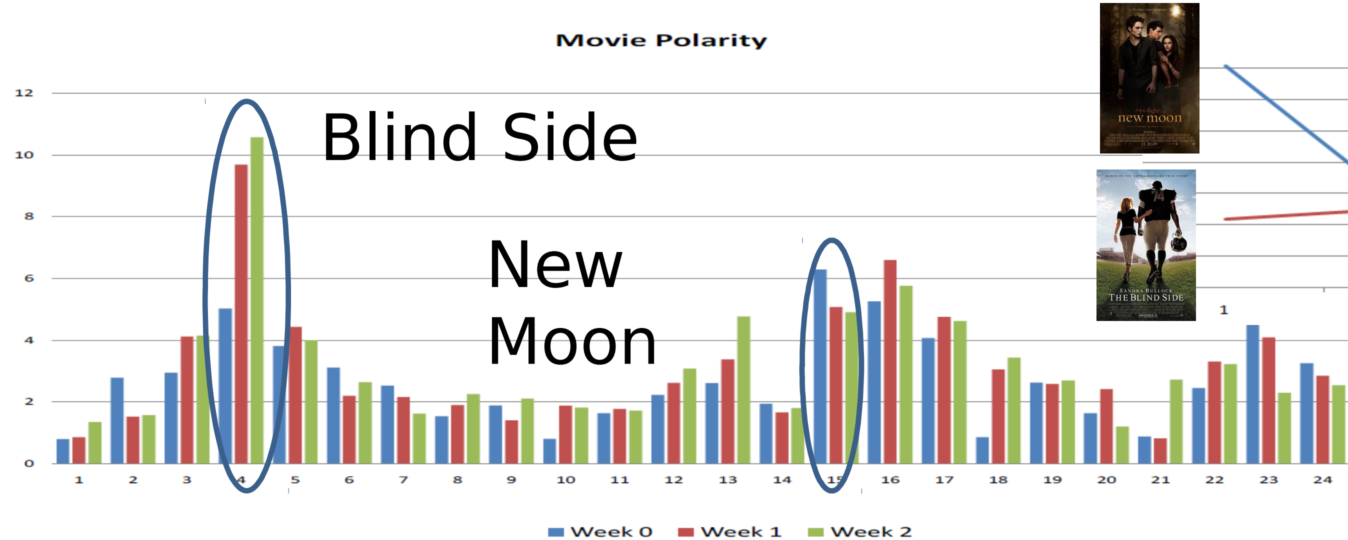
HSX.com
HOLLYWOOD STOCK EXCHANGE
THE ENTERTAINMENT MARKET
Sign Up | Login
IT'S NOT WHO YOU KNOW, IT'S WHO YOU OWN.®
MY PORTFOLIO
ON THE EXCHANGE
NEWS & EVENTS
COMMUNITY
EARN H$
advanced
Search Keyword(s)
GO
Sign up and get H$2,000,000
Hollywood Stock Exchange is the place to buy and trade movies, stars and more.
Sign up and we'll get your portfolio started with H$2,000,000.
SIGN UP »
Learn more »
IPO OF THE DAY
OCEAN'S REBOOT
OCEN4 H$ 41.61
TRADE NOW »
LATEST TWEETS
Toni Collette and Jason Patric join The Yellow...
NominOptions are now trading on HSX
First trailer for #Kubo and the Two Strings...
First trailer for X-Men: Apocalypse #XMen
More »
OPENING THIS WEEK
MORE »
Friday, December 11
In the Heart of the Sea
Genre: Action / Adventure Rated: PG-13
H$28.06 0.00 Trade »
Don Verdean
SPIN FACTOR® BLOG
MORE »
NominOptions are now trading
As the 2015 movie season nears its end, critics associations... More »
12:36 PM, Friday Dec 11
First trailer for Kubo and the Two Strings
Focus Features has unveiled the teaser
HOT PROPERTIES
MORE »
The Bye Bye Man
Genre: Horror Rated: n/a
H$23.63 2.03 Trade »
Ocean's reboot
Genre: Action / Adventure Rated: n/a

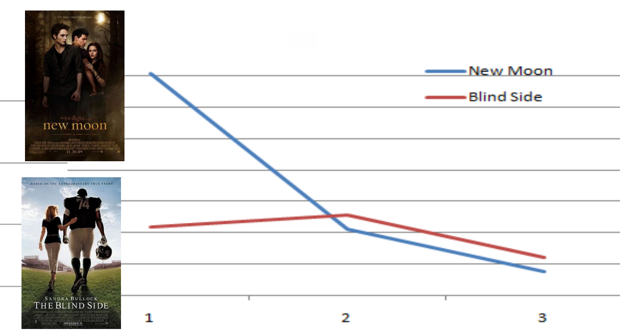
Movies!

Predicting the Future with Social Media

@sitaramasur Asur, Huberman @ WI-IAT 2010

- LingPipe package – language model classifier
- Amazon Mechanical Turk – labeling data
- Positive / Negative / Neutral accuracy 98% (8-grams)
- Predicting second weekend sales
  - using tweet rate time series + P/N ratio: **0.94** Adjusted R^2





[some figures from authors' original slides]

Politics

Examples in politics

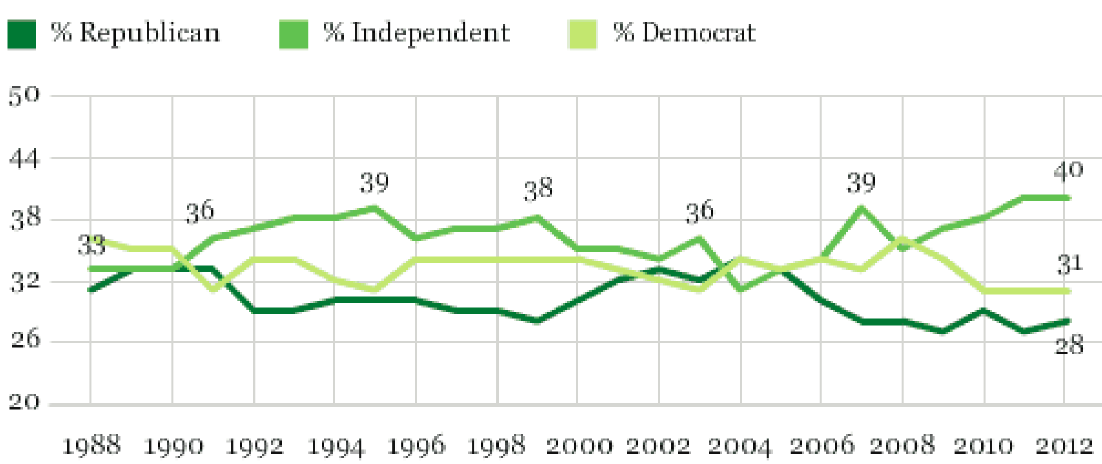
- Hashtags are a good indicator of political topics
- Signs of political leaning
  - Connections, profiles, conversations
- Political manipulation
  - Fake “grassroot” campaigns = “astroturfing”
- “No, you can't predict elections with Twitter”

US Politics

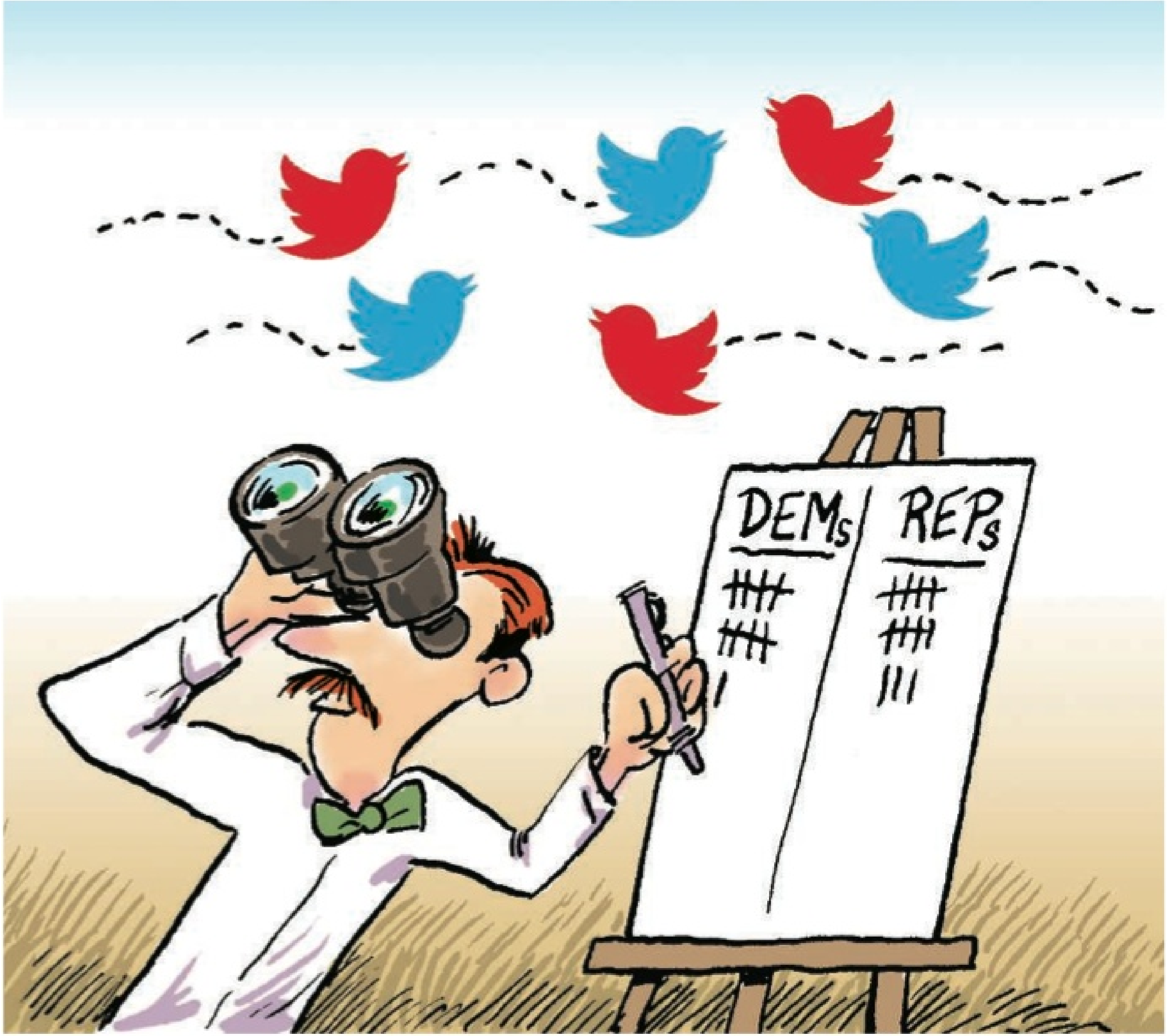
- Most research done so far
- Clear left/right distinction
- Popular political figures
- High(ish) Twitter engagement

DEMOCRAT (left) **REPUBLICAN (right)**

Party Identification, Yearly Averages, Gallup Polls, 1988-2012



Note: Trend is for Gallup polls conducted by telephone.


DEMs
REPs

Political leaning classification

Predicting the political alignment of twitter users

@vagabondjack Conover, Gonçalves, Ratkiewicz, Flammini, Menczer @ SocialCom (2011)

- Bootstrapped hashtag-based sample of political discussion
- Gardenhose Sep 14 - Nov 4, 2010
- Classes: right, left, ambiguous

- Text-based: remove stopwords, hashtags, mentions, urls, all words occurring once in the corpus
- TFIDF weighting:

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{k,j}} \qquad IDF_i = \log \frac{|U|}{|U_i|}$$

- Hashtag-based: remove hashtags used by only one user

Results

Predicting the political alignment of twitter users

@vagabondjack Conover, Gonçalves, Ratkiewicz, Flammini, Menczer @ SocialCom (2011)

- Classifier: Support Vector Machine

Features	Conf. matrix	Accuracy
Full-Text	$\begin{bmatrix} 266 & 107 \\ 75 & 431 \end{bmatrix}$	79.2%
Hashtags	$\begin{bmatrix} 331 & 42 \\ 41 & 465 \end{bmatrix}$	90.8%
Clusters	$\begin{bmatrix} 367 & 6 \\ 38 & 468 \end{bmatrix}$	94.9%
Clusters + Tags	$\begin{bmatrix} 366 & 7 \\ 38 & 468 \end{bmatrix}$	94.9%

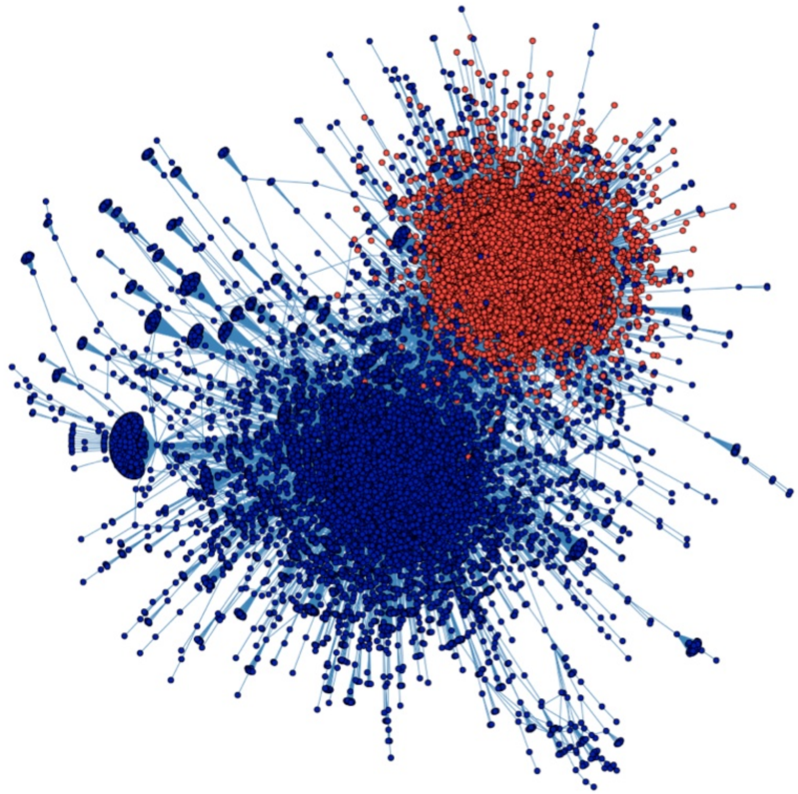
network-based method

Network-based methods

- Label propagation
  - Initialize cluster membership arbitrarily
  - Iteratively update each node's label according to the majority of its neighbors
  - Ties are broken randomly
- Cluster assignment by majority cluster label (using manually labeled data)

Network	Min	Max	Mean
Mention	0.80	1.0	0.89
Retweet	0.94	0.98	0.96

Adjusted Rand Index for 100 label propagation runs on political data



retweet network

$$\text{Clusters} \quad \begin{bmatrix} 367 & 6 \\ 38 & 468 \end{bmatrix} \quad 94.9\%$$

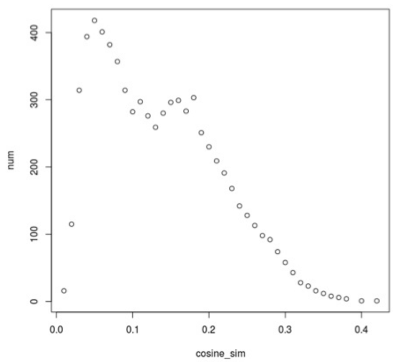
class assignment by cluster majority

Secular vs Islamist

Secular vs. Islamist polarization in Egypt on Twitter

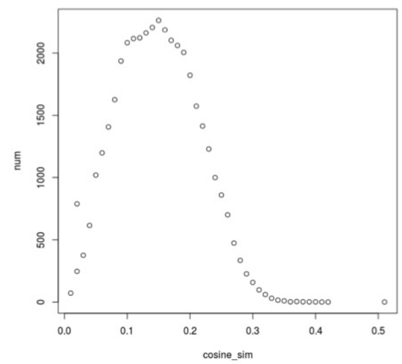
@ingmarweber Weber, Garimella, Batayneh @ ASONAM (2013)

hashtag cosine user-user similarity



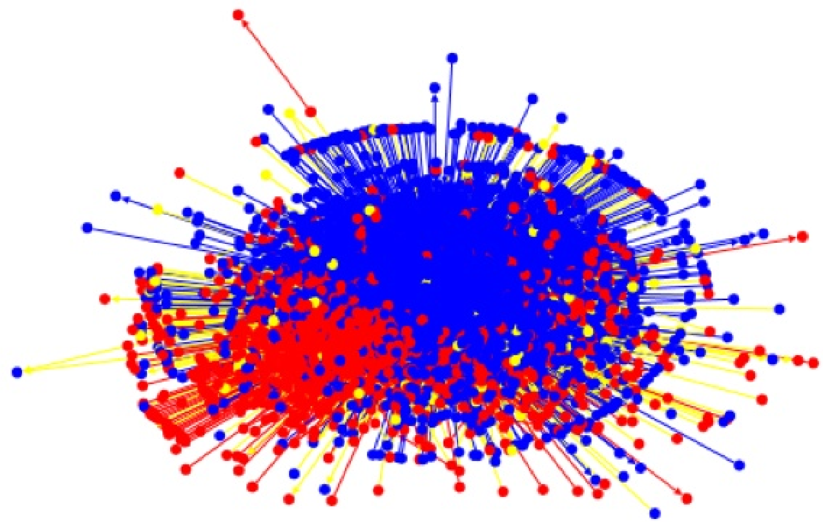
Islamist

average: 0.16
median: 0.11



Secularist

average: 0.16
median: 0.21



Islamist, Secular, intra-ideology

the closer to Islamist, use of

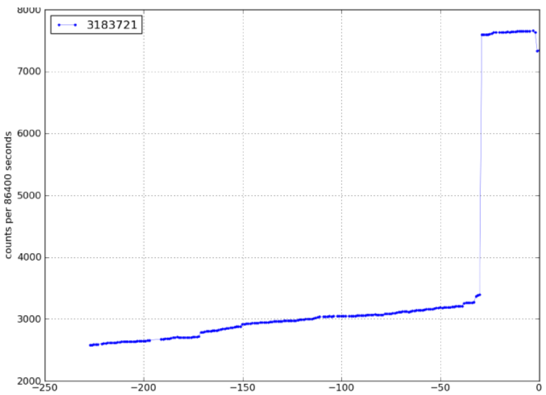
- religious terms *increases*
- charity-related terms *increases*
- derogatory terms *decreases*

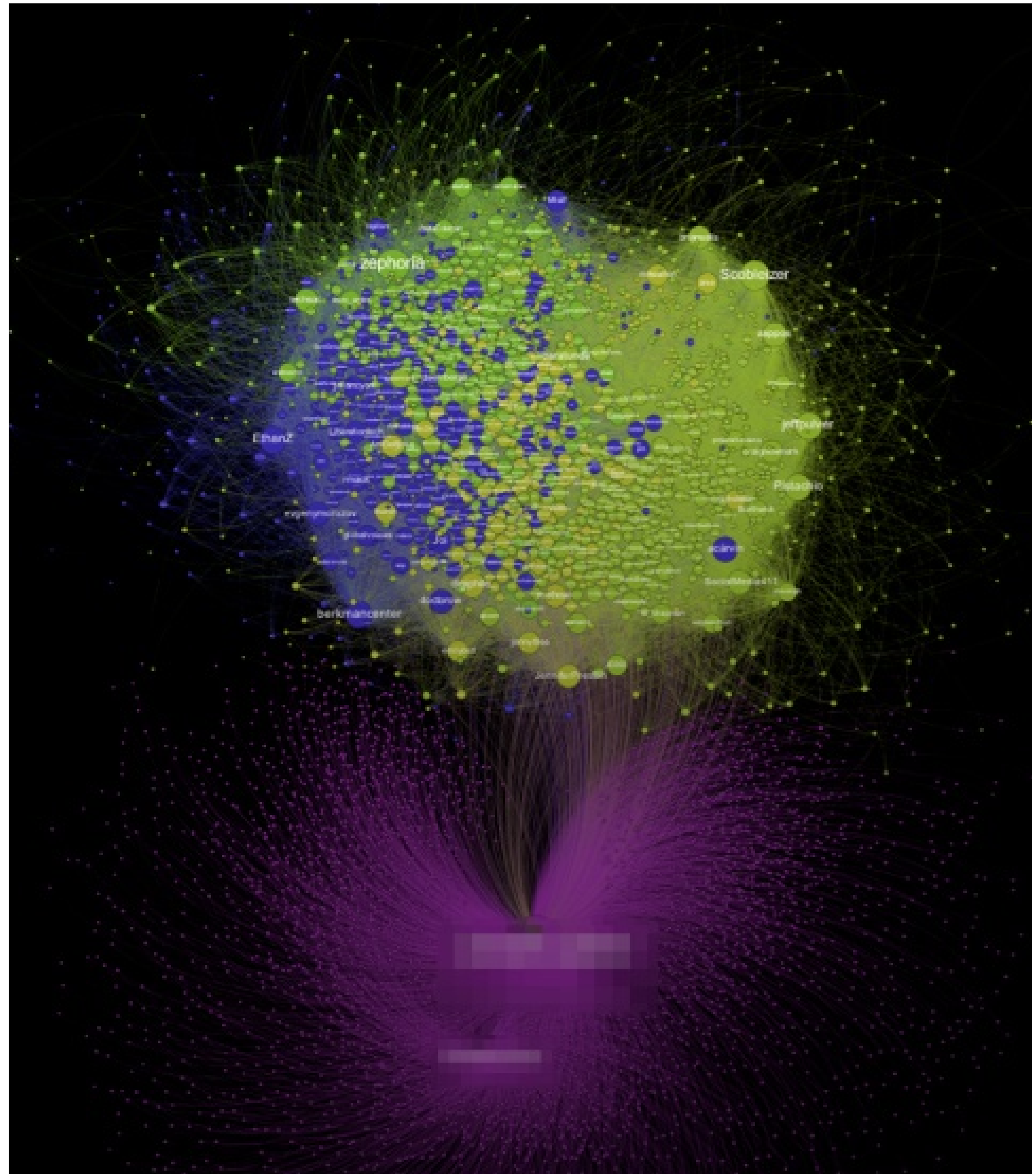
Predicting election results

How (Not) To Predict Elections @takis_metaxas Metaxas et al. @ SocialCom (2011)

- A method of prediction should be an algorithm finalized **before** the election
  - specify data collection, cleaning, analysis, interpretation…
- Data from social media are fundamentally different than data from natural phenomena
  - people change their behavior next time around
  - spammers & activists will try to take advantage
- From a testable theory on *why* and *when* it predicts (avoid self-deception!)
- (maybe) Learn from professional pollsters
  - tweet ≠ user
  - user ≠ eligible voter
  - eligible voter ≠ voter

Astroturfing (4K followers for USD 5)





https://medium.com/i-data/fake-friends-with-real-benefits-eec8c4693bd3

Political Spam (“Truthy”)

Ratkiewicz, Conover, Meiss, Goncalves, Flammini, Menczer @ ICWSM (2011)

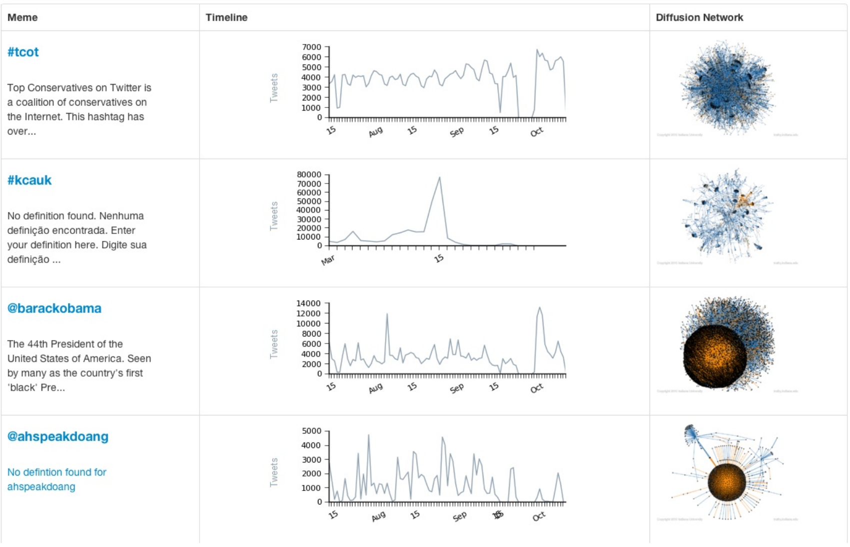
- **Truthiness** is a quality characterizing a "truth" that a person making an argument or assertion claims to know intuitively "from the gut" or because it "feels right" without regard to evidence, logic, intellectual examination, or facts.

`nodes`	Number of nodes
`edges`	Number of edges
`mean_k`	Mean degree
`mean_s`	Mean strength
`mean_w`	Mean edge weight in largest connected component
`max_k(i,o)`	Maximum (in,out)-degree
`max_k(i,o)_user`	User with max. (in,out)-degree
`max_s(i,o)`	Maximum (in,out)-strength
`max_s(i,o)_user`	User with max. (in,out)-strength
`std_k(i,o)`	Std. dev. of (in,out)-degree
`std_s(i,o)`	Std. dev. of (in,out)-strength
`skew_k(i,o)`	Skew of (in,out)-degree distribution
`skew_s(i,o)`	Skew of (in,out)-strength distribution
`mean_cc`	Mean size of connected components
`max_cc`	Size of largest connected component
`entry_nodes`	Number of unique injections
`num_truthy`	Number of times ‘truthy’ button was clicked
sentiment scores	Six GPOMS sentiment dimensions

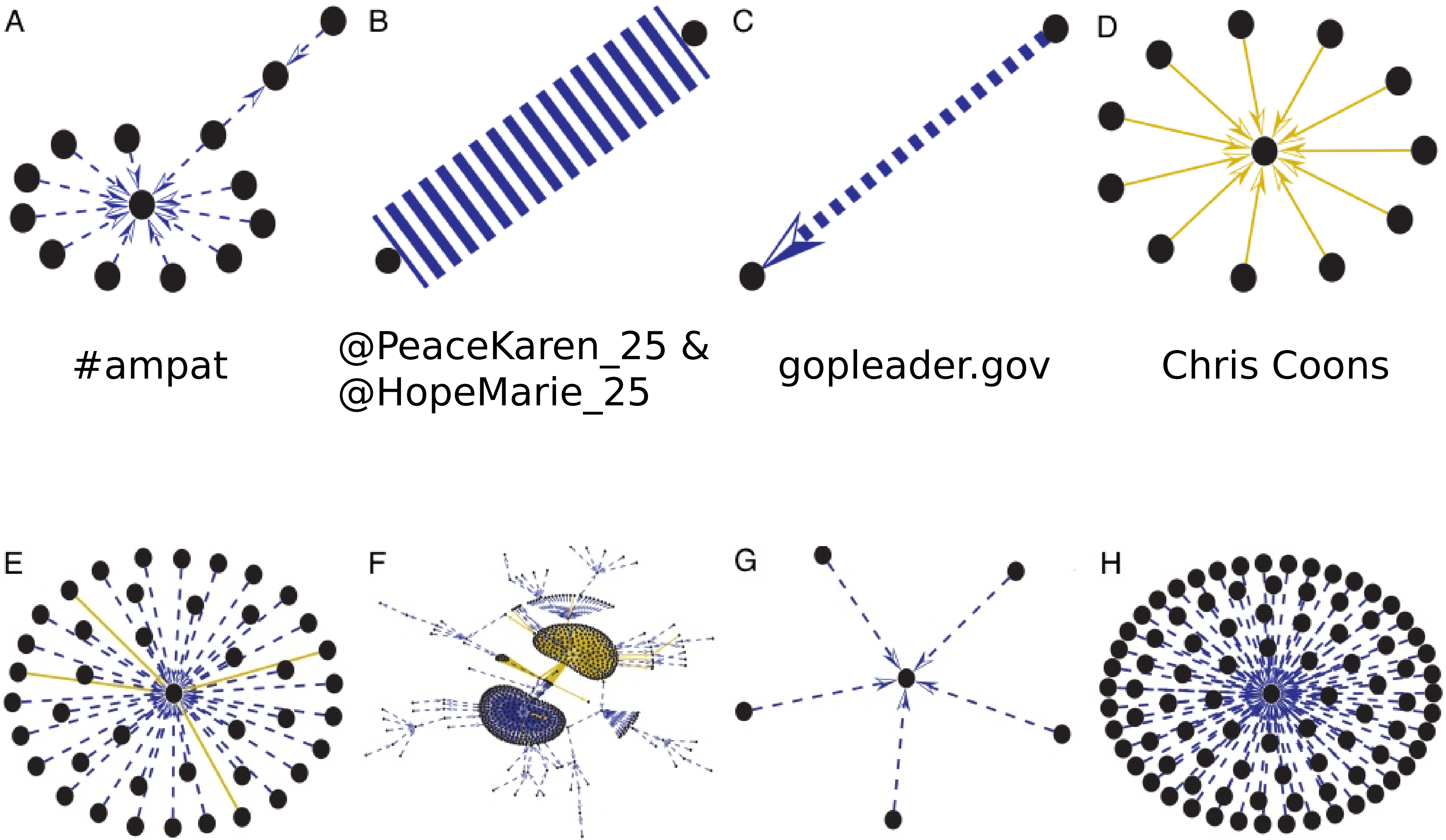
Classifying memes for **astroturf**

Classifier	Resampling?	Accuracy	AUC
AdaBoost	No	92.6%	0.91
AdaBoost	Yes	96.4%	0.99
SVM	No	88.3%	0.77
SVM	Yes	95.6%	0.95

Truthy project by **Indiana University**



Examples



TRUTHY

#ampat

@PeaceKaren_25 & @HopeMarie_25

gopleader.gov

Chris Coons

LEGITIMATE

#Truthy

@senjohnmccain

on.cnn.com/aVMu5y

“Obama said…”

Public Health

Lifestyle and health at scale

Adam Sadilek
& Henry Kautz

@Sadilek &
@HenryKautz

WSDM 2013

Examples in public health

- Many works derived from original Flu Trends
- Increasingly complex models of symptom-messages, treatment-messages
- Allergies, obesity, insomnia
- Mapping well-being in a city

Detecting sick people

Easy: just look for “fever”, “cough”, “sick”, “pain”, and so on! Right?

“I have Bieber *fever*! Justin is amazing”

“I have a horrible *fever*”

“I’m so *sick* of ads on TV”

“No *pain* no gain! Going to the gym for the 5th time this week.”

- Classical binary classification problem
- Used all token unigrams, bigrams and trigrams as tokens
- Used Amazon Mechanical Turk for ground truth labels
- Trained an SVM, got .98/.97 precision/recall

Example features

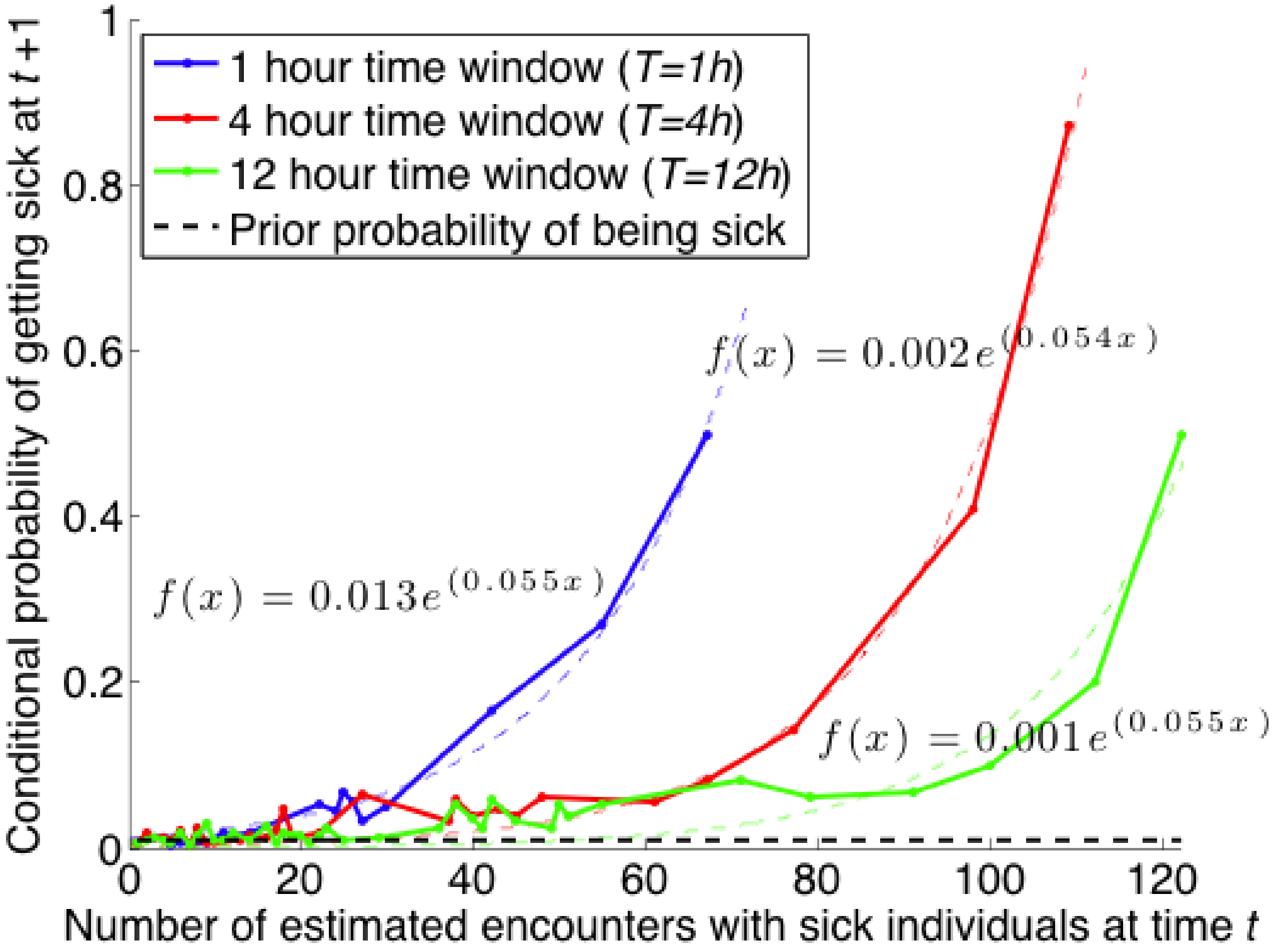
Positive Features		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
sick	0.9579	sick of	−0.4005
headache	0.5249	you	−0.3662
flu	0.5051	lol	−0.3017
fever	0.3879	love	−0.1753
feel	0.3451	i feel your	−0.1416
coughing	0.2917	so sick of	−0.0887
being sick	0.1919	bieber fever	−0.1026
better	0.1988	smoking	−0.0980
being	0.1943	i'm sick of	−0.0894
stomach	0.1703	pressure	−0.0837
and my	0.1687	massage	−0.0726
infection	0.1686	i love	−0.0719
morning	0.1647	pregnant	−0.0639

Table 1: Examples of positively and negatively weighted significant features of our SVM model C.

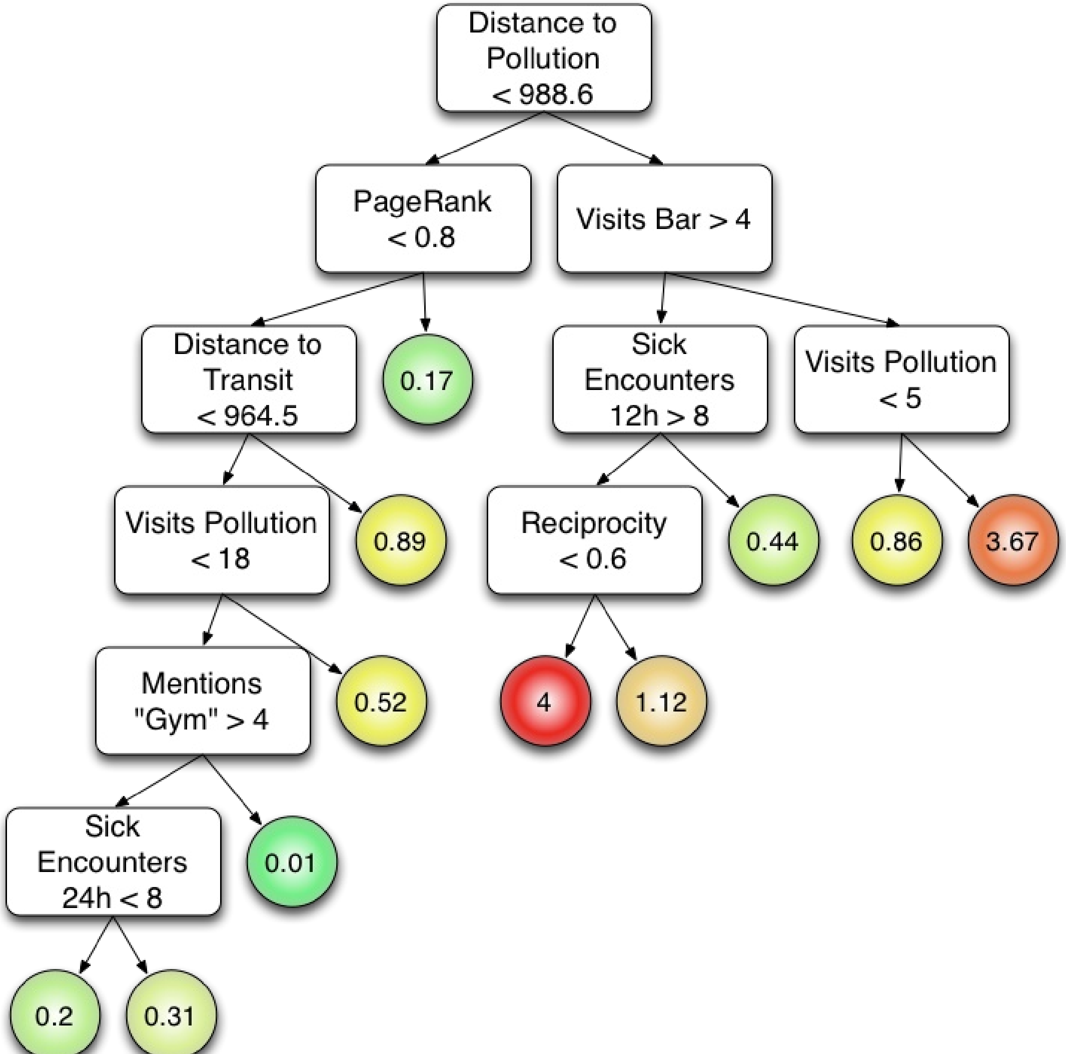
Visits and meetings

- Have mobility traces for the 6k users
- Can infer when they “meet”, i.e. are within a close vicinity of each other
  - Could just be the same bar without actually meeting face to face
- Can also infer visits to places such as gyms, bars, public transportation

Being close to sick people makes you more likely to become sick



Predicting number of sick days



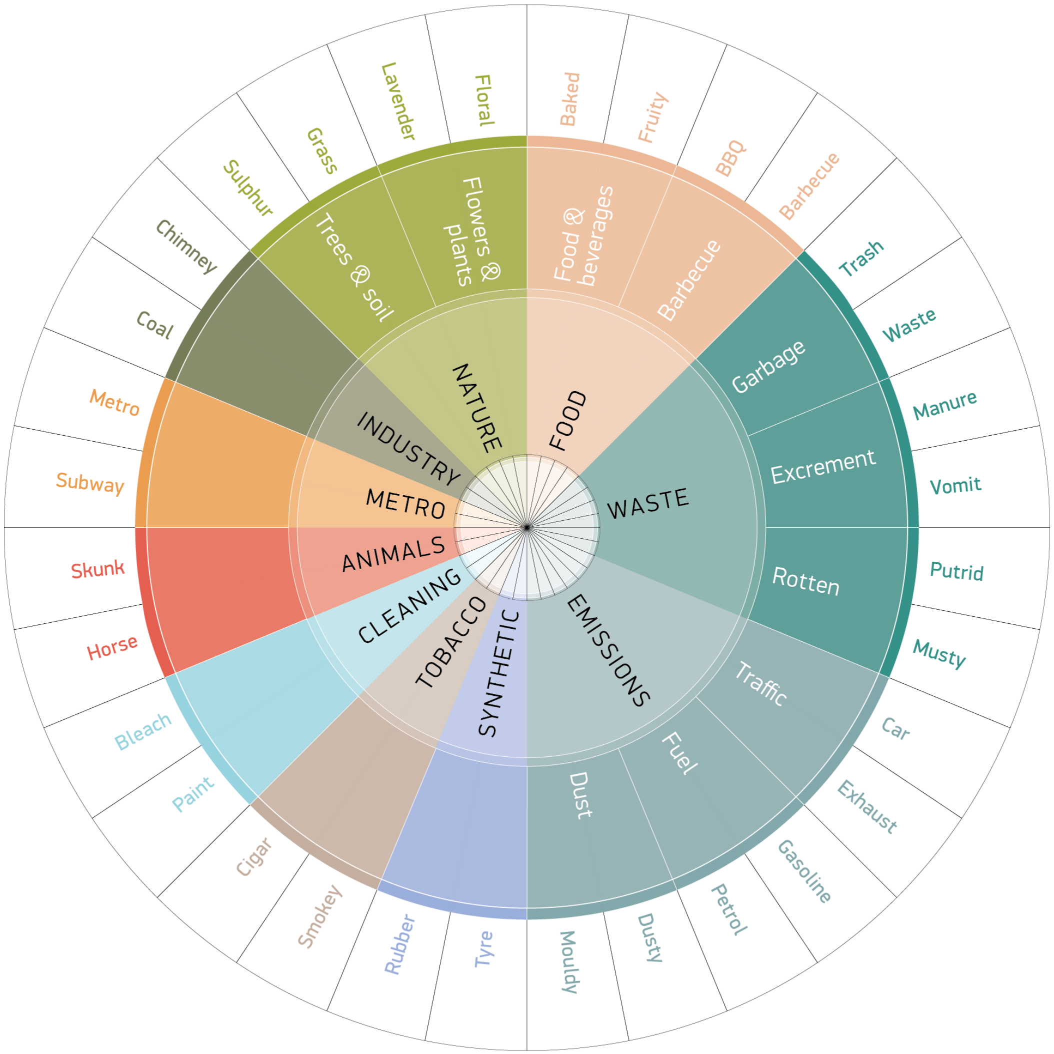
Smart Cities

Examples in “smart cities”

- Data-driven neighborhood boundaries
- Data-driven residencial/commercial zones
- Tourism and beauty

(a) Shortest (b) Beauty

Smell maps

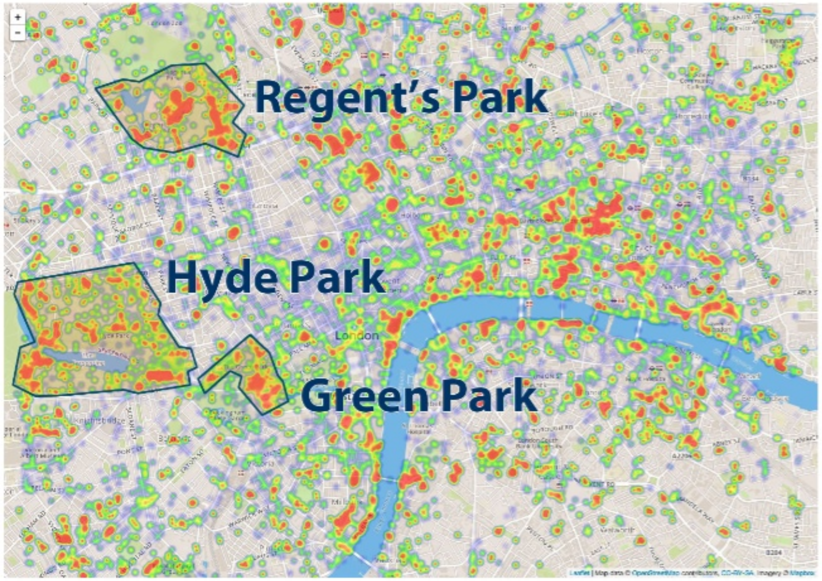


Urban Smellscape Aroma Wheel
(depicting background and episodic aromas only)
Aiello, L., McLean, K., Quercia, D., Schifanella, R., 2015

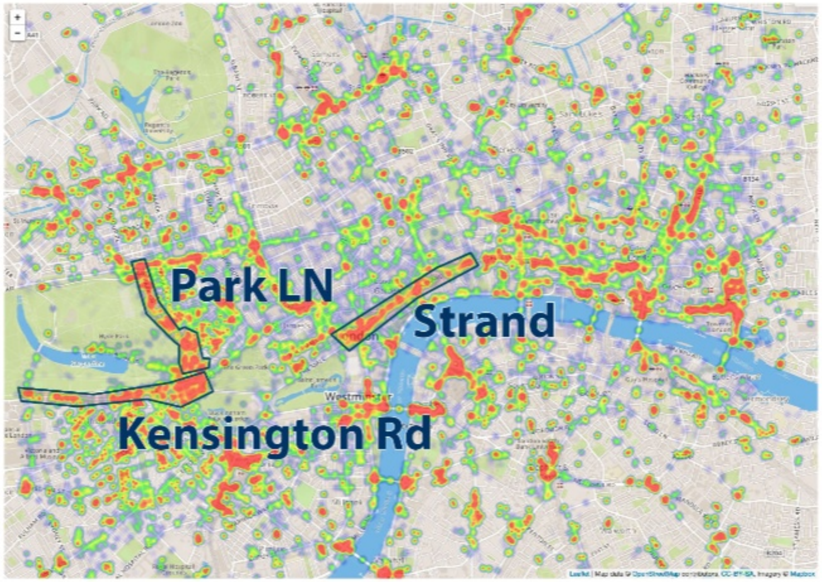
Smells and tags

- Compute P(smell|tags) for places where both a set of tagged photos and a set of smell annotations have been observed
  - Supervised learning approach
- Data from Twitter and Instagram

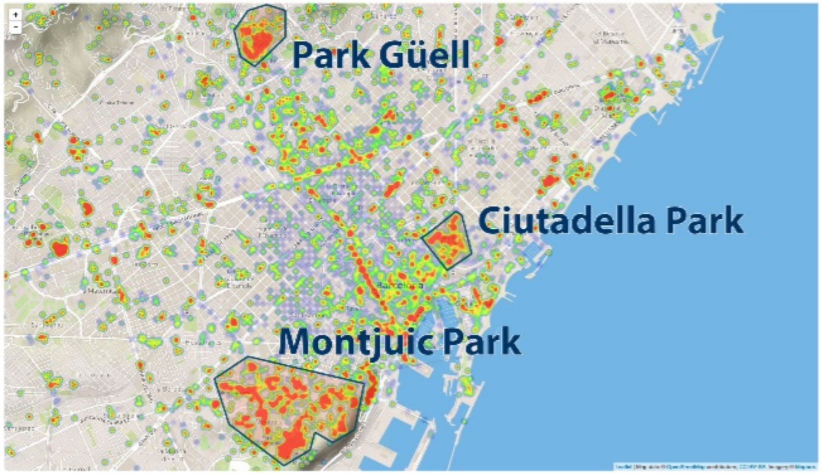
Smells



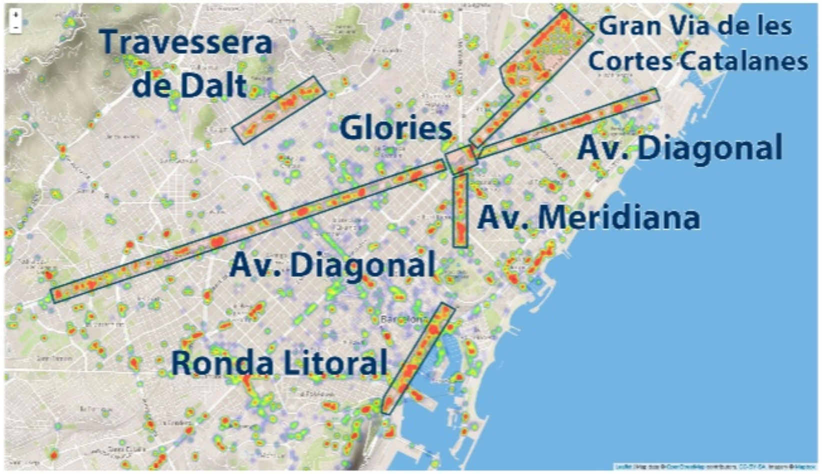
London, nature



London, emissions



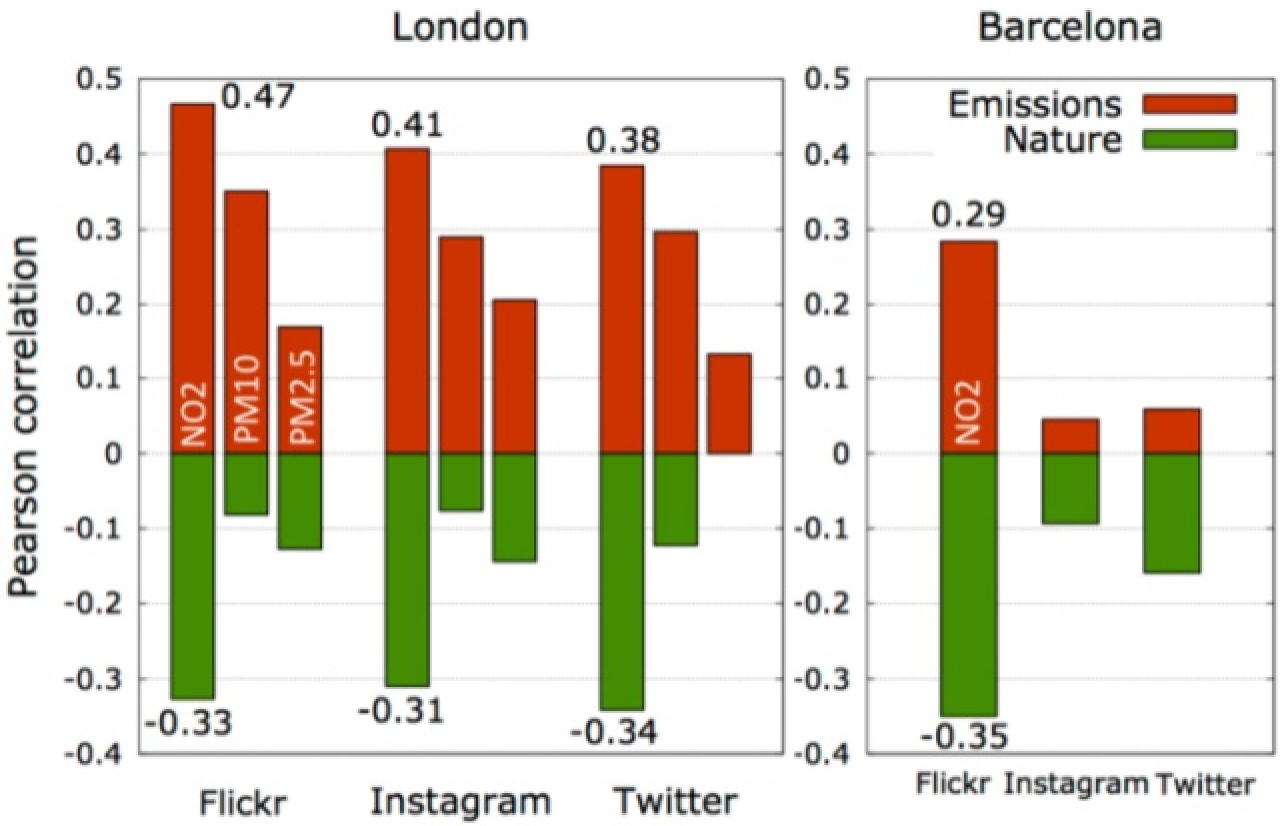
Barcelona, nature



Barcelona, emissions

http://researchswinger.org/smellymaps/

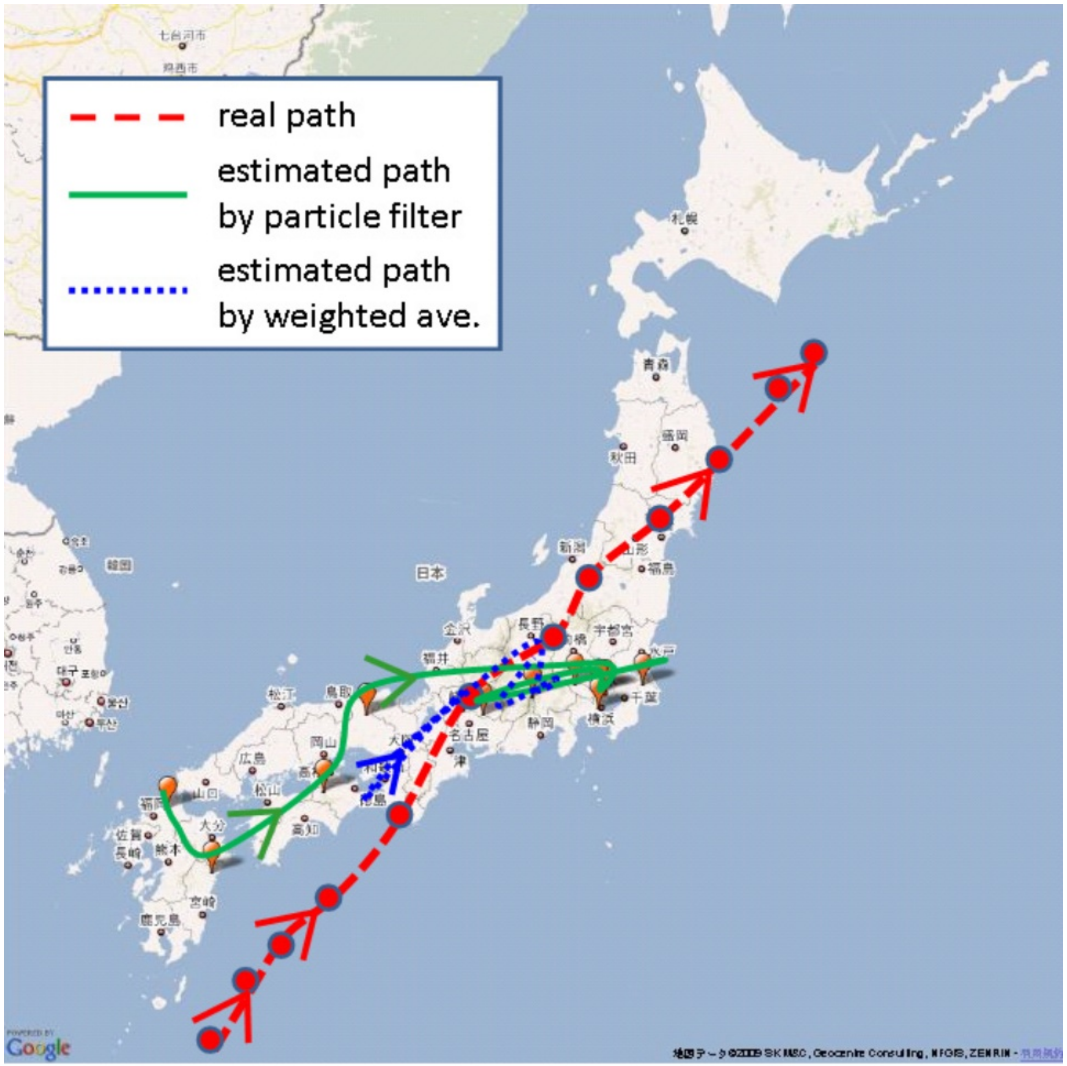
Correlation with pollution



Examples in event detection

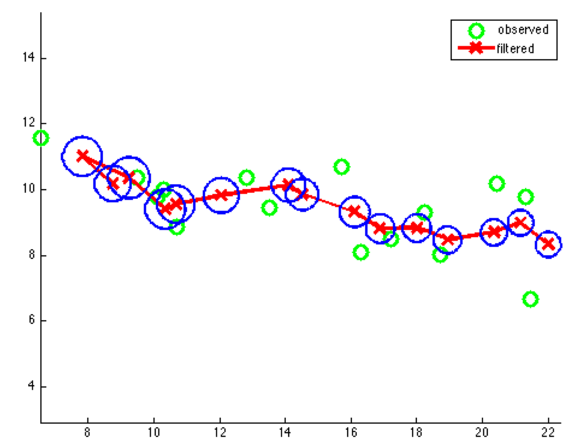
- Mass convergence events, e.g. demonstrations
- Precursors of riots
- Traffic jams, accidents, or road blocks
- Man-made and natural disasters
  - And sub-events

Estimation of typhoon trajectory



Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 851-860. DOI=http://dx.doi.org/10.1145/1772690.1772777

Kalman filter



Best practices in social media mining

- Interdisciplinary work
- Mixed methods: qualitative and quantitative
- Well-grounded in the domains' literature
- Recognize, measure, and possibly counter sample biases
- Robust to different settings, metrics, datasets
- Outcomes provide an advantage to practitioners
  - E.g. to make better decisions than without this data