

PUBLIC HEALTH MEETS SOCIAL MEDIA: MINING HEALTH INFO FROM TWITTER

Michael Paul (@mjp39)
Johns Hopkins University



Crowdsourcing and Human Computation
University of Pennsylvania
October 30, 2013

Learning about the real world through Twitter

- Millions of people share on the web what they are doing every day
- Can analyze social media to infer what is happening in a population
 - Can make inferences about the population's **health**
- **Passive** data monitoring
 - Work with data that's already out there
 - vs active methods: soliciting data from people (e.g. surveys)
- Faster, cheaper than traditional data collection – but noisier

This lecture: Key ideas

- Applications
 - What can we learn about health?
(and why would we want to do that?)
- Methods
 - How do you mine Twitter?
- Evaluation
 - How accurate is the mined data?
- Ethics
 - How does social media mining fit in with current medical research practices?

Twitter: Data



- Free streams of data provide 1% random sample of public status messages (tweets)
- Search streams provide tweets that match certain **keywords**
 - Still capped at 1%, but more targeted
 - We collect tweets matching any of 269 health keywords

```
,sick,cold,body,pain,hurts,sore,nose,hospital,doctor,cancer,u  
,burning,flu,exhausted,medicine,surgery,knee,cough,fever,doct  
ise,cure,eaten,dentist,vision,bedtime,physical,treatment,pill  
t,ache,ankle,pill,numb,recovery,physically,wrist,depression,h  
y,clinic,pains,jaw,sneeze,lungs,swollen,puke,anxiety,appt,rec  
petite,resting,coughing,infection,diabetes,migraine,sickness,
```

- <https://dev.twitter.com/docs/streaming-apis/keyword-matching>
- https://github.com/mdredze/twitter_stream_downloader

Twitter: Location data



- **Geolocation:** often we need to identify where the authors of tweets are located
- Some tweets tagged with GPS coordinates
 - Only 2-3% of tweets/users
 - Can improve coverage by tenfold by also considering self-reported location in user profiles



Twitter: Location data



- **Geolocation:** often we need to identify where the authors of tweets are located

- *Carmen*

- Identifies where a tweet is from using GPS + user profile info, e.g.

```
{"city": "Baltimore",  
"state": "Maryland",  
"country": "United States"}
```



- Java (python coming soon) software available:
 - <https://github.com/mdredze/carmen>

Twitter: Health data?

- Twitter is a **noisy** data source

Mashable

SOCIAL MEDIA ▾

TECH ▾

BUSINESS ▾

ENTERTAINMENT ▾

MORE ▾

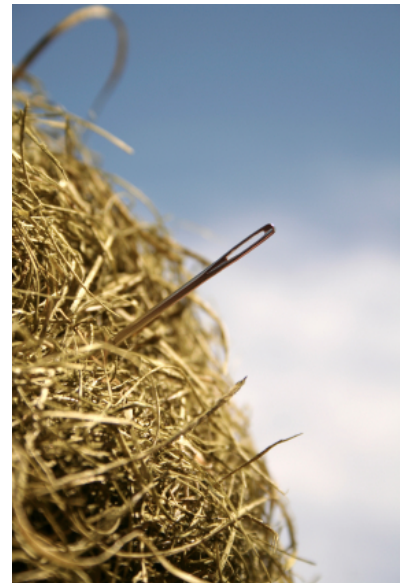
TWITTER ANALYSIS: 40% of Tweets Are Pointless Babble

- 2012 study (André, Bernstein, Luther):

ratings on Twitter updates. Using our dataset of over 43,000 voluntary ratings, we find that nearly **36% of the rated tweets are worth reading**, 25% are not, and 39% are middling. These results suggest that users tolerate a large amount of less-desired content in their feeds. We find that

Twitter: Health data?

- My estimate: about **0.1%** of tweets are about tweeters' health
 - (1.6 million out of 2 billion tweets in an earlier study)
- 0.1% of Twitter is still a lot of data!
 - ~ **half a million** tweets per day
- Lots of data, but hard to find in noise
 - Absolutely huge
 - Relatively tiny



Finding health tweets

- Step 1: keyword filtering
 - Filter out tweets unlikely to be about health
 - Large set of 20,000 keywords
- Not all tweets containing keywords are actually about someone's health
 - This tweet contains lots of health keywords:



- Step 2: supervised machine learning

Finding health tweets

- Step 2: supervised machine learning
- Labeled data
 - **5,128** tweets
 - *About health | Unrelated to health | Not English*
- Labels collected through Mechanical Turk
 - Each tweet labeled by 3 annotators
 - Final label determined by majority vote
 - 10 labels per HIT
 - Each HIT contained 1 gold-labeled tweet to identify poor-quality annotators



Finding health tweets

- About **1%** of tweets contained the 20,000 health keywords
- About **15%** of those were tagged as relevant by the health machine learning classifier



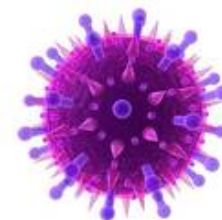
about **0.1%** of all tweets are health-related

- **1.6 million** health tweets from 2009-2010
- Over **150 million** collected since Aug 2011

Health tweets

- So we can we do with health tweets?

Flu surveillance



- Idea: people tweet about being sick
- More sick tweets will appear when the flu is going around
 - <https://twitter.com/search?q=flu&src=typd&f=realtime>
- Why do we care?
 - Cheap data source to complement primary disease surveillance systems (e.g. hospital data, lab work)
 - Real-time, can be automated
 - Lofty goal: early detection of novel, serious epidemics

Flu surveillance

- Goal: identify and count tweets that indicate the user is sick with the flu
 - Proxy for how many people in the population have the flu
- Challenge: not all tweets that mention “flu” actually indicate a person is sick



Ellen DeGeneres @TheEllenShow

21h

Long story short; I didn't have the swine **flu**, but what I do have is a serious case of dance fever. [#UnusedEllenDanceLines](#)

 Favorited 1,012 times

Expand

Finding flu tweets

- As before: supervised machine learning
- Labeled data
 - **11,990** tweets
 - *Flu infection | General flu awareness | Unrelated to flu*
- Same quality control measures as before
 - Also hand-verified all labels in the end
 - Changed 14% of labels



Finding flu tweets

- Machine learning classifiers identify tweets that indicate **flu infection**
- Many features beyond n-grams:
 - Retweets, user mentions, URLs
 - Part-of-speech information
 - Word classes:

Infection	getting, got, recovered, have, having, had, has, catching, catch, ...
Disease	bird, the flu, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, terrified
Treatment	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
...	...

Flu surveillance

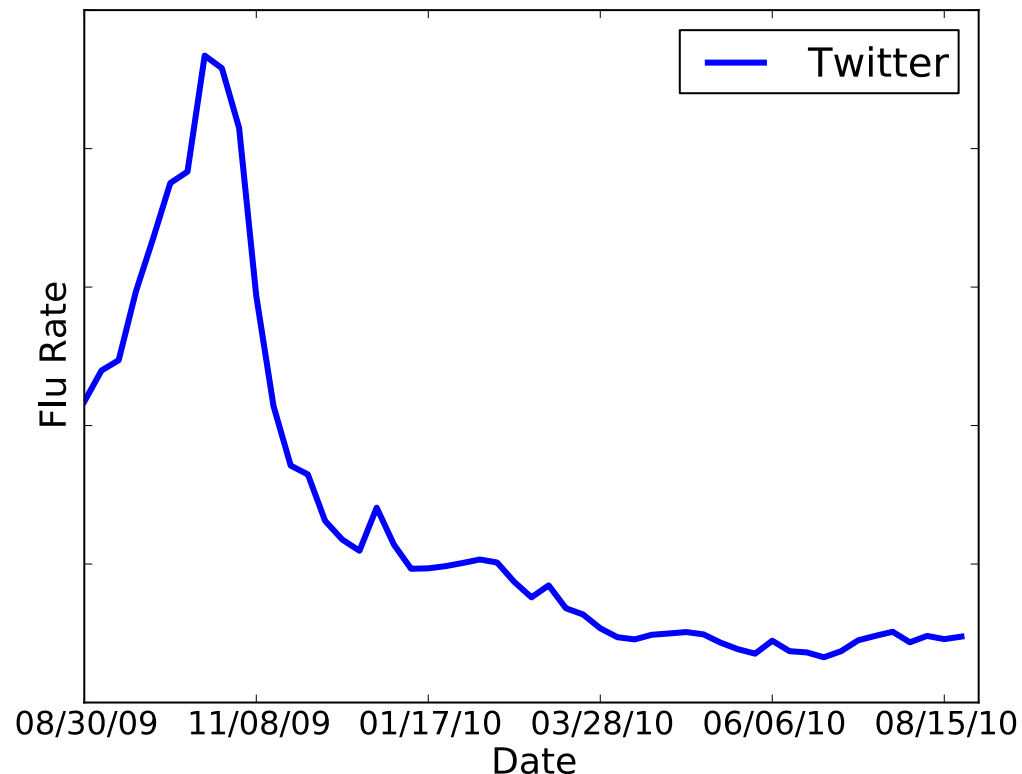
- Estimated weekly rate of flu on Twitter:

$$\frac{\# \text{ tweets about flu infection that week}}{\# \text{ of all tweets that week}}$$

- Normalize by number of all tweets to adjust for change in Twitter volume over time

Flu surveillance (2009-10)

- Large spike of flu activity around October
 - This was during the swine flu pandemic



- Is this accurate?

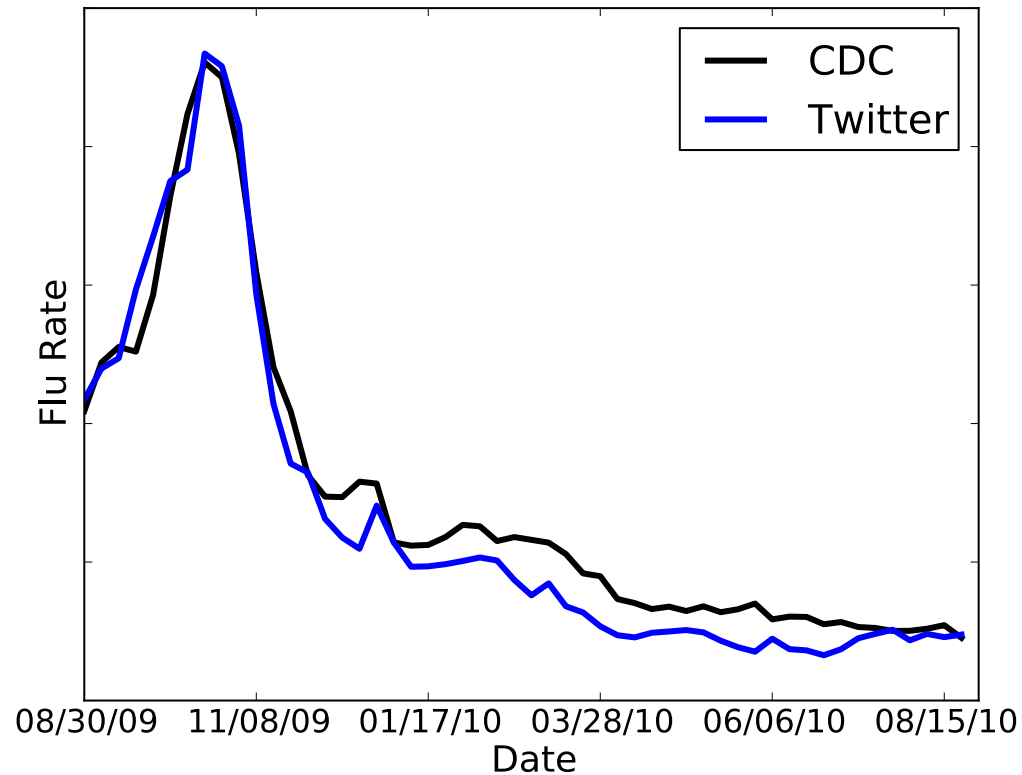
Flu surveillance: Evaluation

- Compare our estimates to “ground truth” data
- We take government surveillance data to be ground truth
 - from the CDC (Centers for Disease Control and Prevention)
 - weekly counts of hospital outpatient visits for influenza-like symptoms
- Common metric: Pearson correlation
 - compare temporal trend of Twitter estimates against CDC data

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

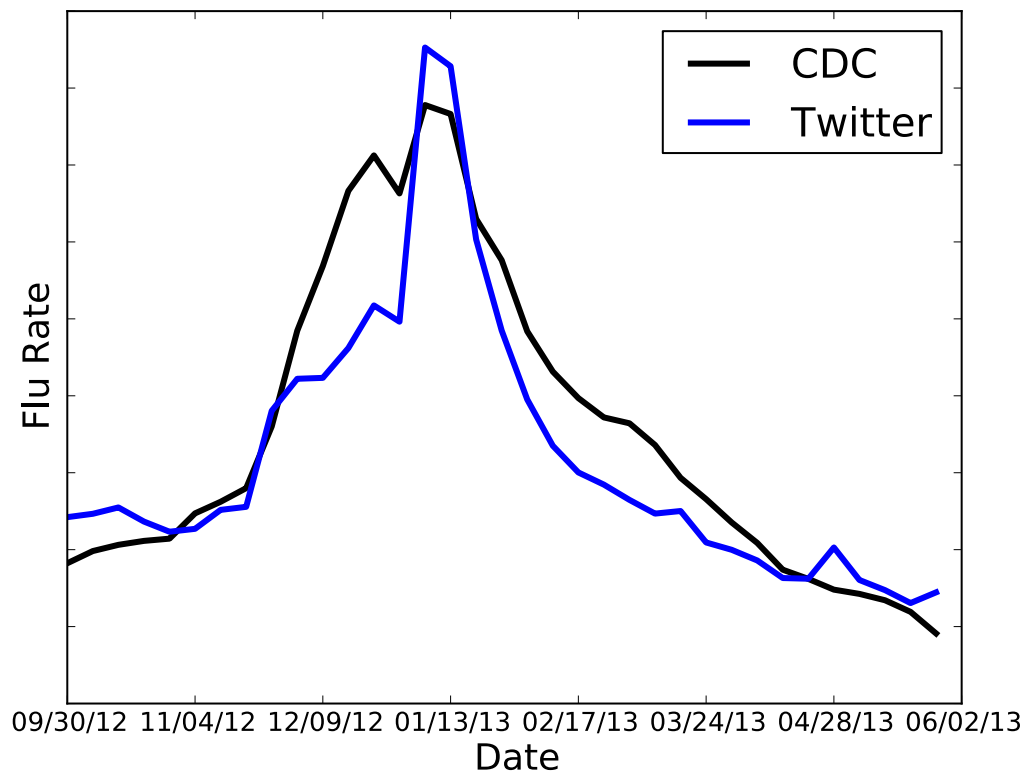
Flu surveillance (2009-10)

- Correlation with CDC: **0.99**



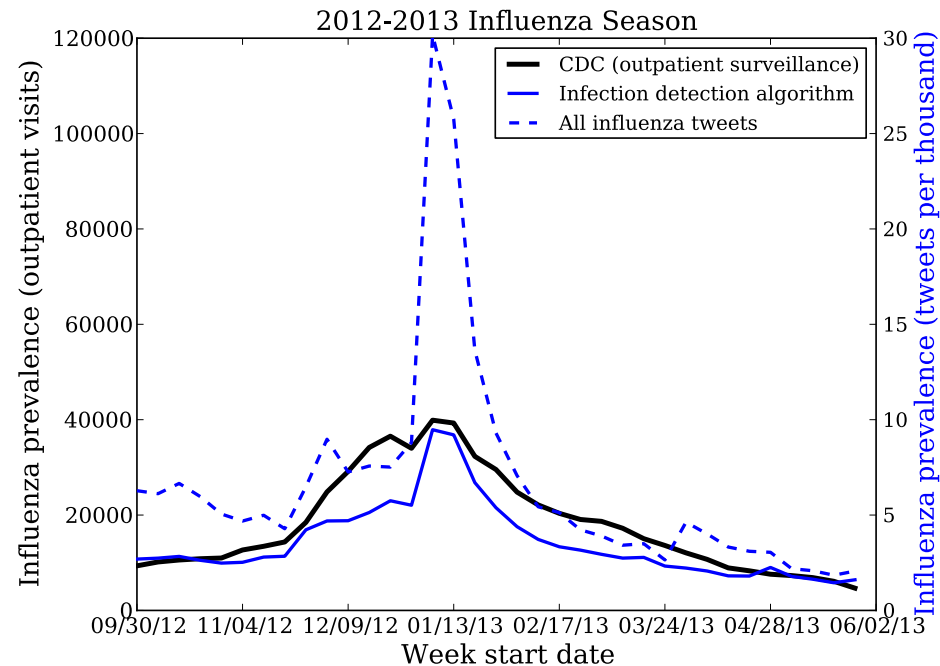
Flu surveillance (2012-13)

- Correlation with CDC: **0.93**



Flu surveillance: More evaluation

- What if we just estimate the flu rate by counting tweets containing the words “flu” or “influenza”?
- Not as highly correlated:
 - 2009-10: 0.97 (2% reduction)
 - 2012-13: 0.75 (20% reduction)



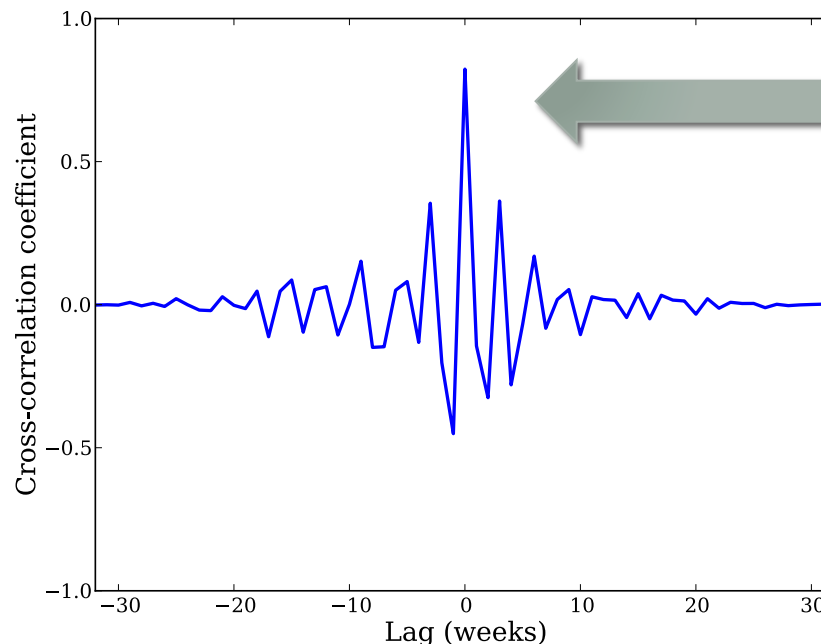
- More spurious spikes from keyword matching

Flu surveillance: More evaluation

- Cross-correlation

- Measures similarity between curves when one of the trends is offset by some number of weeks (lead/lag)

$$(f \star g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f^*(\tau) g(t + \tau) d\tau$$



Twitter neither leads/lags CDC (but maybe certain keywords do?)

Flu surveillance: More evaluation

- Basic correlation may overstate how good you are doing
 - As long as the peak weeks have above-average rates and the off-season weeks are below-average, you'll get a pretty high number
 - Especially true if trend has high **autocorrelation** (cross-correlation with itself) at nonzero lag
- Trend **differencing**
 - Subtract previous week's rate from current week
 - Measures correlation of week-to-week increase/decrease
 - More directly measures what you probably care about
- Box-Jenkins methods
 - Guidelines for applying differencing

Flu surveillance: More evaluation

- Simple accuracy
 - How often does the weekly direction of the trend (up or down) match CDC?
- Maybe more interpretable than correlation
- Our Twitter infection classifier:
 - **85%** direction accuracy (2012-13)
 - Simple keyword matching: 46%

Beyond flu

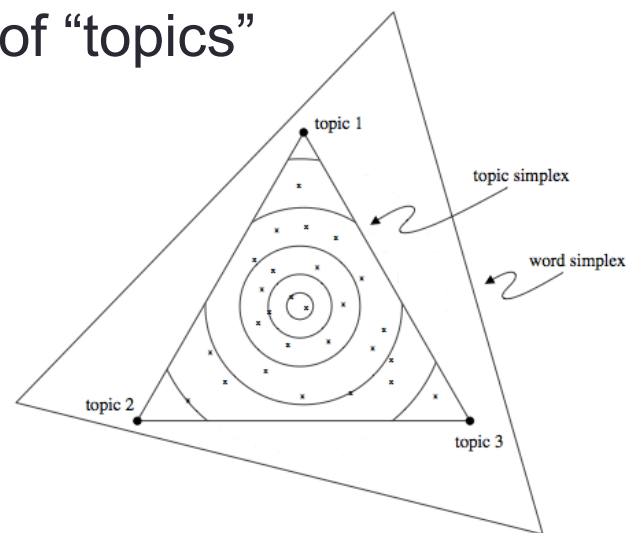
- The flu project was an in-depth study of one disease
 - Machine learning with human annotations
 - Time/labor intensive
 - Rich set of features

Beyond flu

- The flu project was an in-depth study of one disease
 - Machine learning with human annotations
 - Time/labor intensive
 - Rich set of features
- Alternative approach: broad, exploratory analysis
 - Find lots of diseases on Twitter
 - **Unsupervised** machine learning
 - No human input
 - Simple keyword-based models

Topic modeling

- Statistical model of text generation
 - decomposes data set into small number of “topics”
 - the topics are not given as labels
 - **unsupervised** model
- Two types of parameters:
 - **$p(\text{topic}|\text{document})$** for each document
 - **$p(\text{word}|\text{topic})$** for each topic
- Optimize parameters to fit model to data (a collection of documents)



Topic modeling

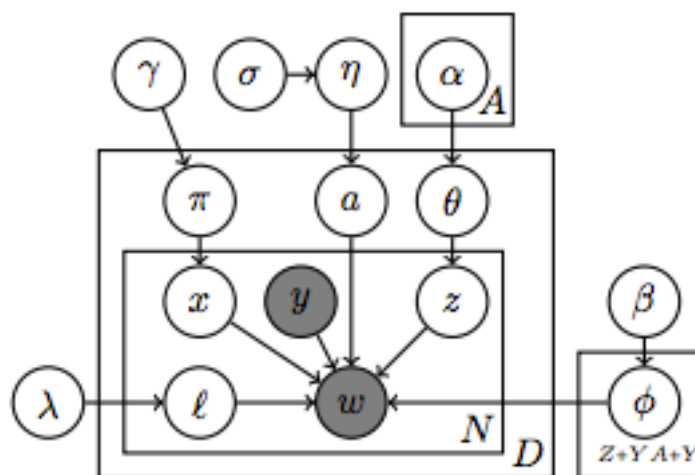
- Automatically groups words into topics
- Automatically labels documents with topics
- Example when applied to New York Times articles:

music band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach
theater play production show stage street broadway director musical directed	clinton bush campaign gore political republican dole presidential senator house	stock market percent fund investors funds companies stocks investment trading	restaurant sauce menu food dishes street dining dinner chicken served

- from Hoffman, Blei, Wang, Paisley

Topic modeling health tweets

- We created a topic model specifically for finding health topics in Twitter
- Ailment Topic Aspect Model (ATAM)
 - Distinguishes health topics from other topics in the data
 - Breaks down health topics by general words, symptom words, treatment words



Topic modeling health tweets

“Insomnia”



Topic modeling: Evaluation

- How accurately do these word clusters correspond to real-world concepts?
- As before: find existing data sources to compare to

Topic modeling: Diet and exercise

- Compare the “diet and exercise” health topic to government survey data about lifestyle factors
- Track rates across U.S. states
 - Geographic trends (vs temporal trends)
- Positively correlated with rates of physical activity and aerobic exercise
 - **0.61** and **0.53**
- Negatively correlated with rates of obesity
 - **-0.63**

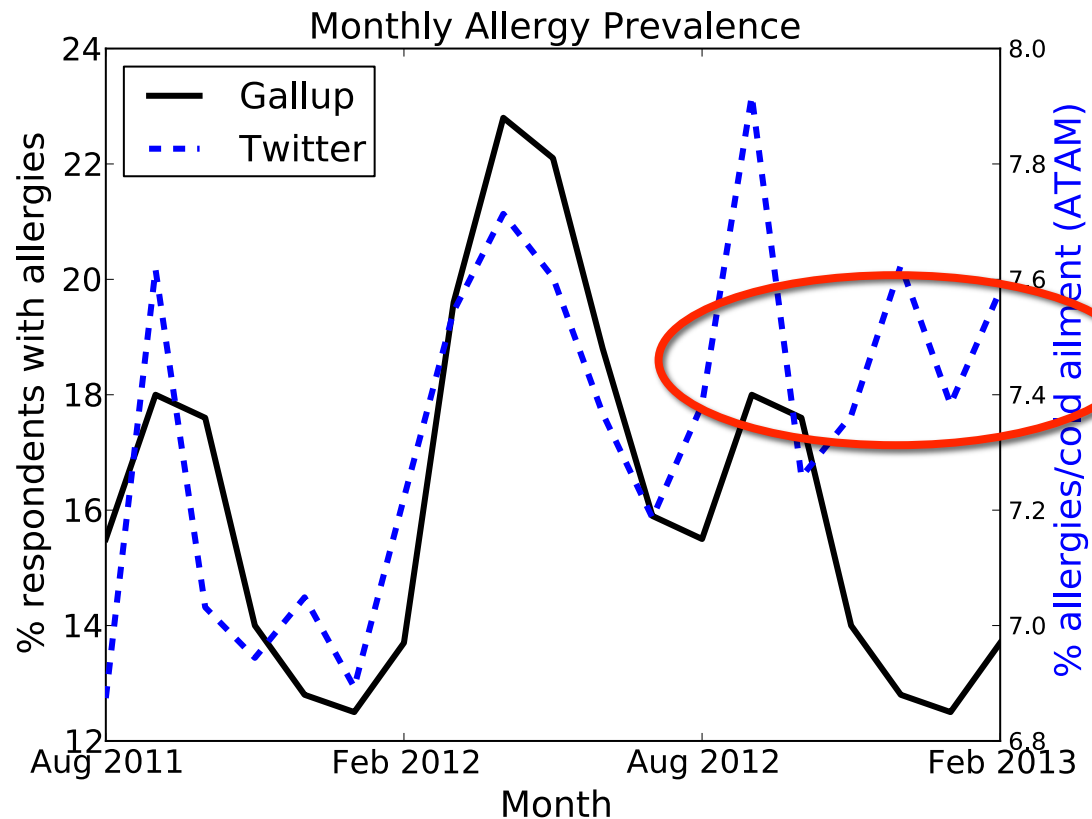
Topic modeling: Allergies

- Allergies aren't part of CDC surveillance systems
 - But private data sources exist
- We compared to phone survey results from Gallup
 - “Were you sick with allergies yesterday?”



Topic modeling: Allergies

- Correlation: **0.48**
 - 2011-12 season only: **0.84**



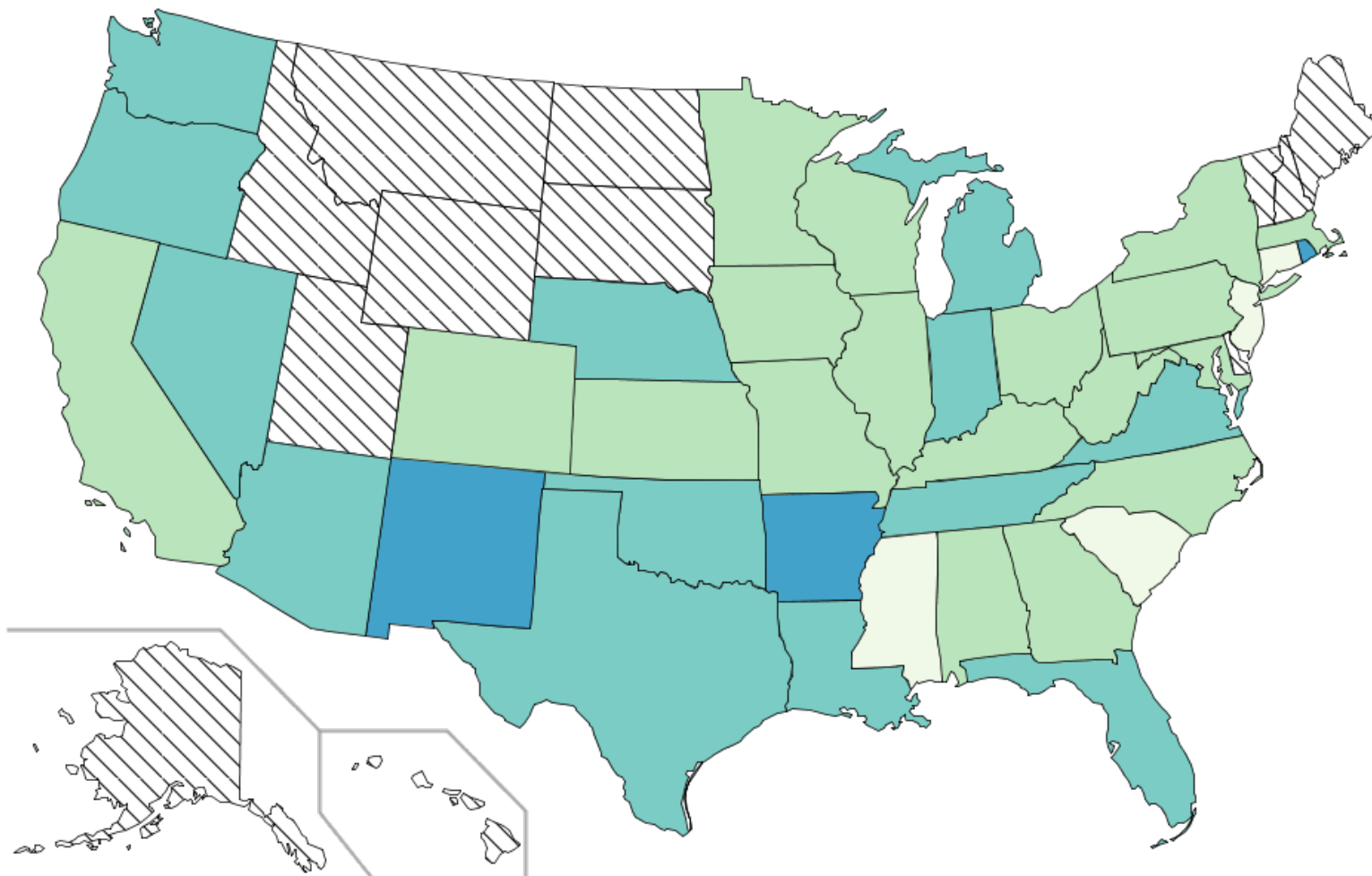
higher in 2012-13;
conflated with
strong flu season
(similar symptoms:
coughing, sneezing)

Topic modeling: Evaluation

- Informal evaluation: visualize, check against intuition
- Do word clusters make sense?
 - Is there obvious noise?
 - e.g. cold/flu symptoms mixed with allergies
- Do geographic patterns make sense?
 - Are rates similar in nearby states?

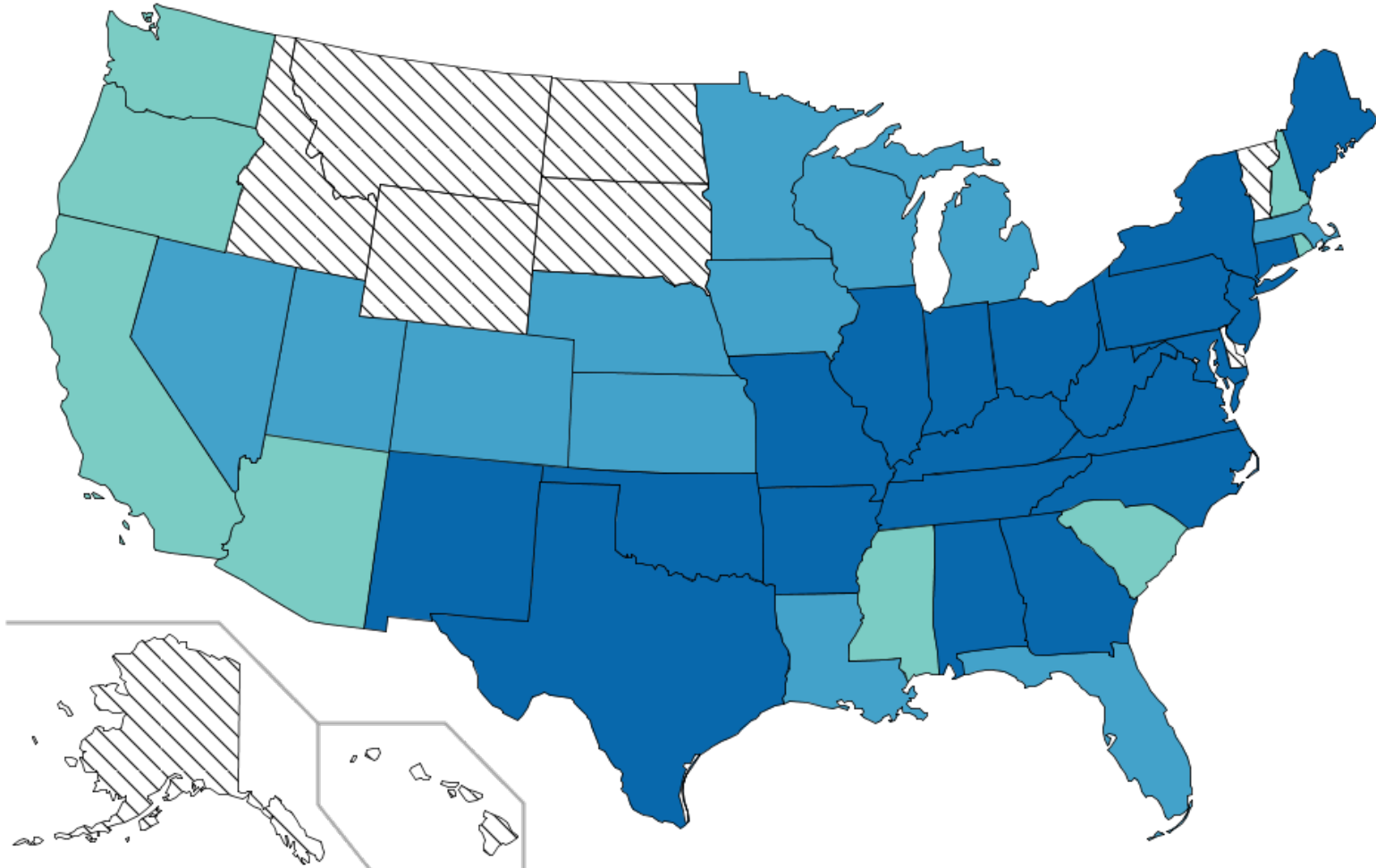
Topic modeling: Allergies

February 2010



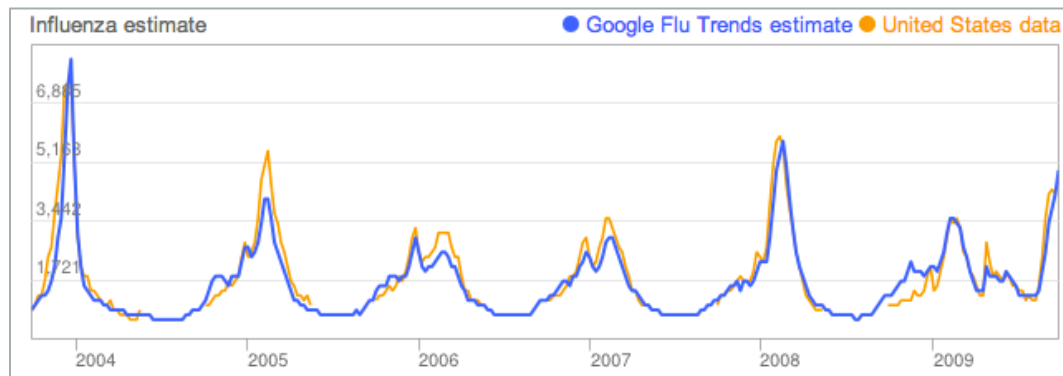
Topic modeling: Allergies

April 2010



Other methods

- (Linear) Regression
 - Explicitly train your system to predict ground-truth data
 - Approach used by Google Flu Trends (and others)



- Manual analysis by humans
 - “small data” approach
 - Good option for ideas too difficult to automate with machine learning

Other applications

- Broad health applications seem to work well enough
 - but the data may not support deeper medical questions
 - people share a lot but not everything
 - demographics differ between real world and social media
 - and can vary a lot between different social media sites
- More hype than substance in many applications
 - Can Twitter predict the **stock market**?
 - I don't know of anyone who has gotten rich from this yet...
 - Can Twitter predict **elections**?
 - If it could, Ron Paul would be president...

Ethics/Privacy



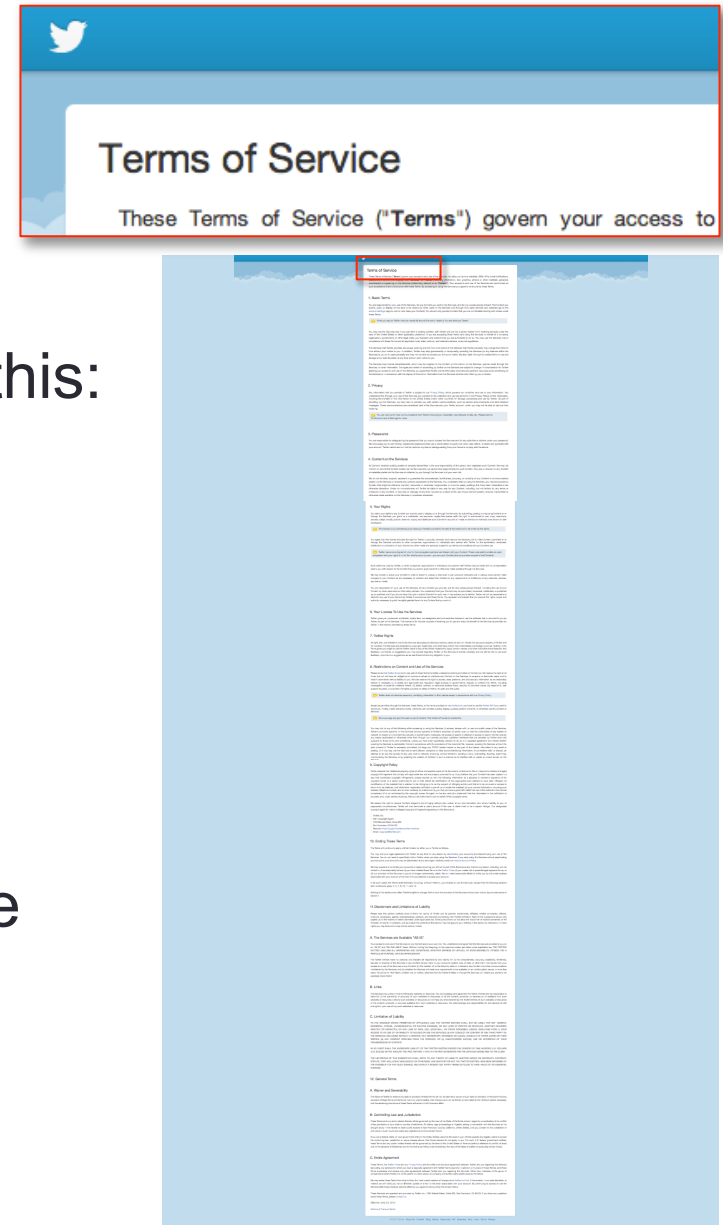
- Ethics is an important part of human subject research
- But there aren't comparable guidelines for "human computation" research
- So here are some things to think about...

Ethics: Guidelines

- Modern medical research guided by IRBs (institutional review board)
- IRB policy on social media data mining:
 - Data is public, therefore no approval required
 - This applied to what I showed you today
 - This policy may eventually change
- There are still privacy norms surrounding public data
 - and it's possible to use public data in creepy ways

Ethics: Guidelines

- Informed consent?
- Technically, if people actually read this:
- Tweets are public by default
 - not all users realize this
- Even if users are aware the data is public, they might not expect it to be used in certain ways



Ethics: Privacy

- Aggregate statistics generally preserve privacy
 - as long as they aren't aggregated in a town of 20 people
- Analysis of individual users may deviate from privacy expectations



- Sadilek, Kautz, Silenzio 2012

Ethics: Privacy

- Health information is particularly sensitive and should be treated differently



- Courtesy of MappyHealth by Social Health Insights

Ethics: Privacy

- Hard to fully anonymize data
 - Same issue with medical records
- Can be de-anonymized with more data, more context

Forbes ▾	New Posts +2 posts this hour	Popular A Walk Down Silk Road	Lists TV's Highest-Paid Actors
-----------------	--	---	--

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Ethics: Rules of thumb

- Stick to large-scale aggregate analysis
- Don't use more data than a user likely intended to share
- Don't use data that you don't need for what you're doing
- Think about how users will react to your app or research

- Good reading:
 - Slides by Caitlin Rivers: "Ethical user of Twitter for digital disease detection"
 - http://figshare.com/articles/Ethical_use_of_Twitter_for_DigDisDet/805198
 - danah boyd and Kate Crawford. 2012. Critical questions for Big Data. *Information, Communication & Society*: 15(5).

See more

- socialmediahealthresearch.org
- CIKM 2013 tutorial: Twitter and the real world
 - <https://sites.google.com/site/twitterandtherealworld/>
- Conference on Digital Disease Detection
 - <http://healthmap.org/ddd/>
- Apps/startups:
 - GermTracker – germtracker.org
 - MappyHealth – mappyhealth.com
 - Sickweather – sickweather.com