Microsoft Research

UCL

UNIVERSITY OF COPENHAGEN

Learning about health and medicine from Internet data

Elad Yom-Tov, Microsoft Research Israel

Ingemar Johansson Cox, University College London and University of Copenhagen

Vasileios Lampos, University College London

About the authors

Elad Yom-Tov, Senior Researcher, Microsoft Research
Research interests: Large-scale IR & ML for medicine
Website: www.yom-tov.info

Ingemar J. Cox, Professor of CS, U. Copenhagen and University College London
Research interests: IR & application of data mining methods to online resources for medical purposes
Website: http://mediafutures.cs.ucl.ac.uk/people/IngemarCox/

Vasileios Lampos, Research Associate, University College London
Research interests: Statistical Natural Language Processing, Social Media Research, Computational Social Science
Website: http://lampos.net/

Outline

- When is Internet data useful for medical research?
- Data sources
- Linking to ground truth
- Identifying a cohort
- Learning from Internet data
- Privacy and ethics
- Some open questions

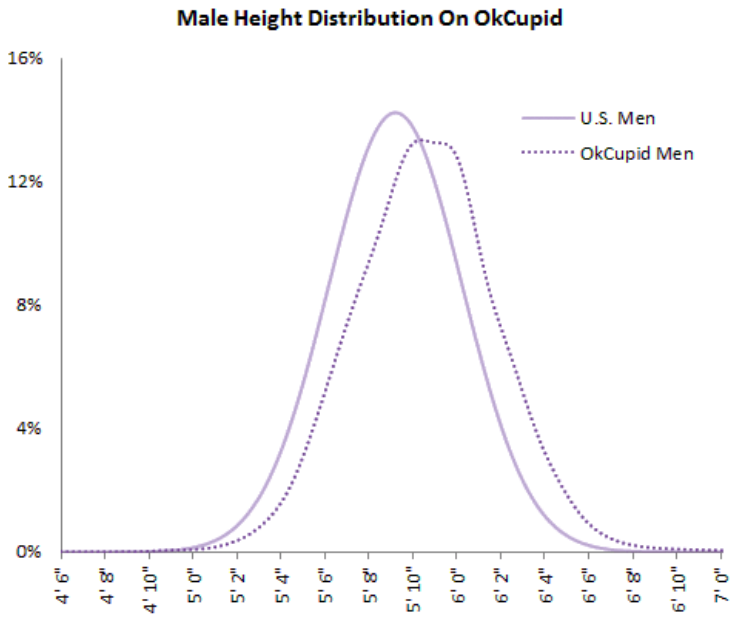
When is Internet data useful for medical research?

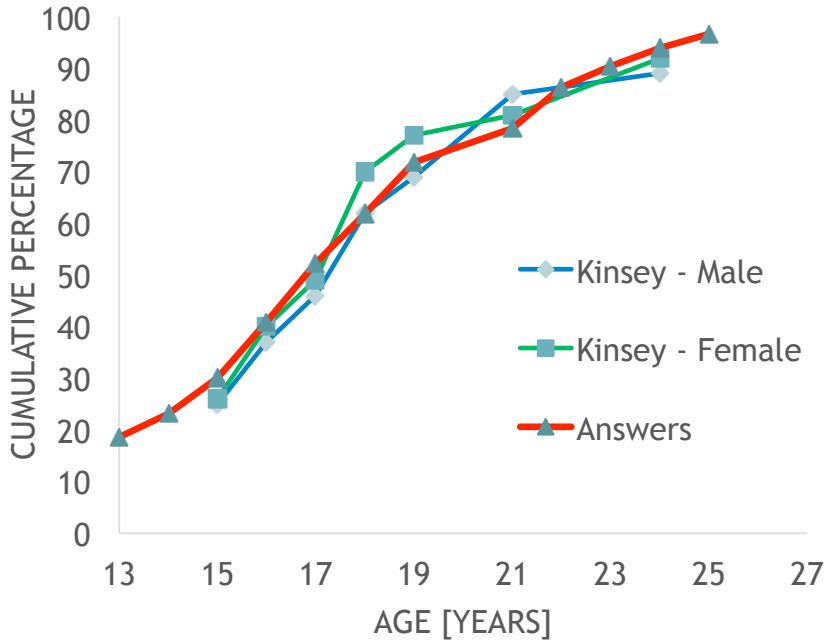
When is Internet data useful for medical research?

- **If** it is harder to collect (unbiased) data in the physical world
- **If** a more delicate sensor is needed
- **If** the activity is largely web-driven
- **If** people have a difficulty reporting associations

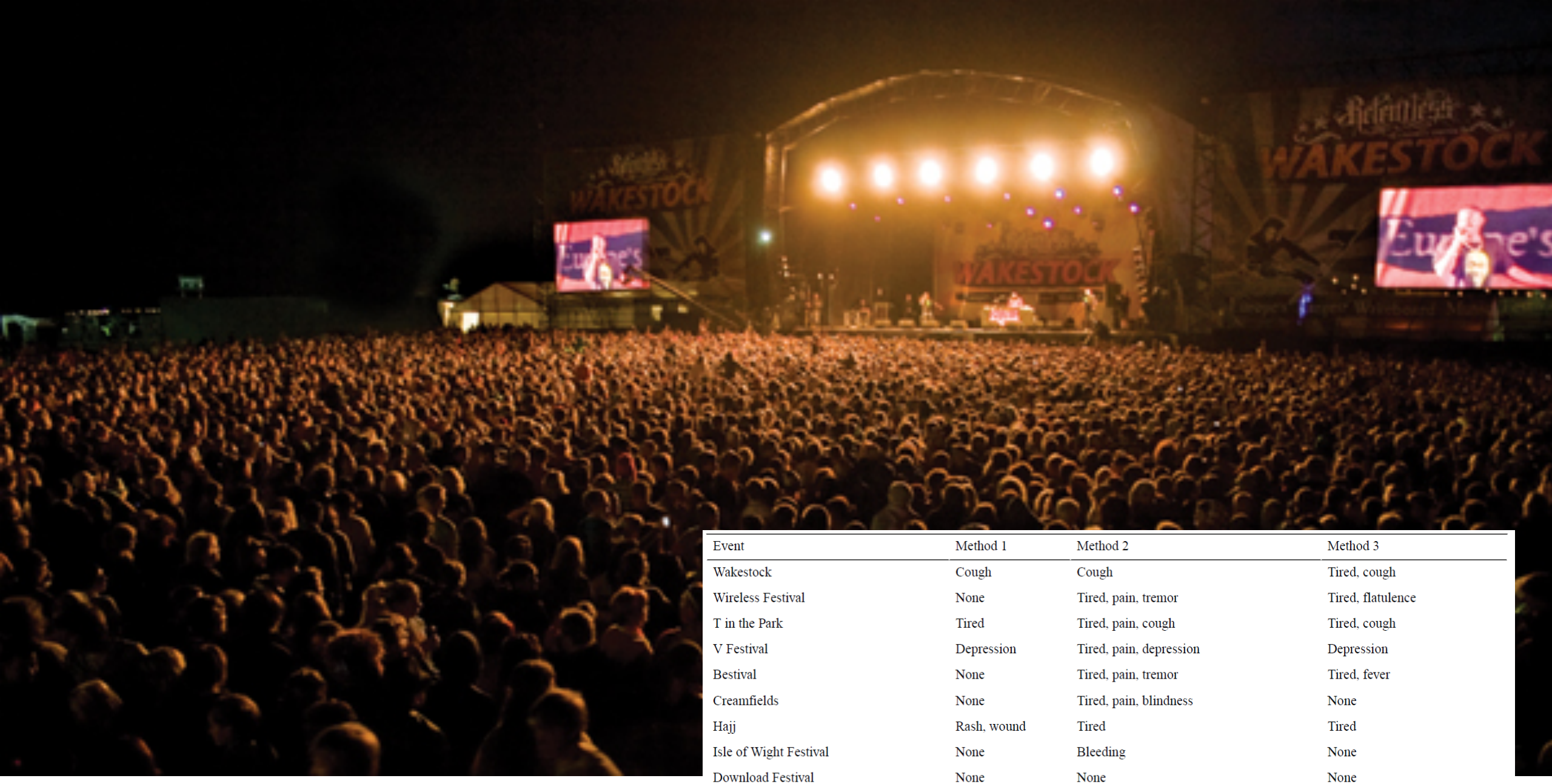
When is it worthwhile doing?

- **If** it is harder to collect (unbiased) data in the physical world





Pelleg et al., 2012

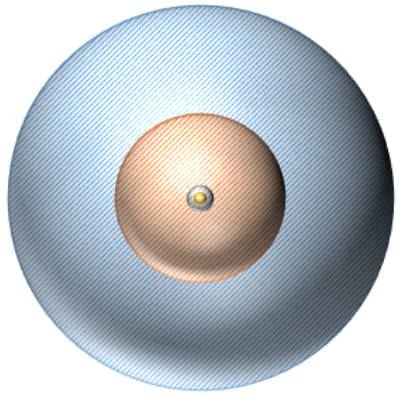


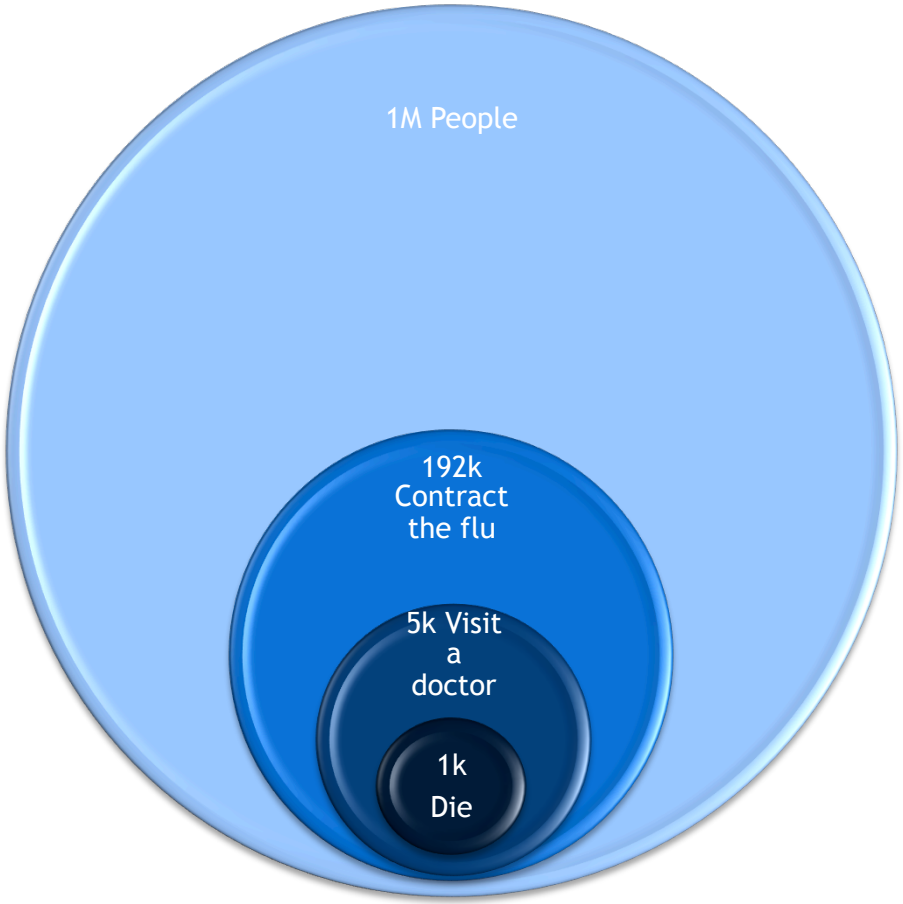
Event	Method 1	Method 2	Method 3
Wakestock	Cough	Cough	Tired, cough
Wireless Festival	None	Tired, pain, tremor	Tired, flatulence
T in the Park	Tired	Tired, pain, cough	Tired, cough
V Festival	Depression	Tired, pain, depression	Depression
Bestival	None	Tired, pain, tremor	Tired, fever
Creamfields	None	Tired, pain, blindness	None
Hajj	Rash, wound	Tired	Tired
Isle of Wight Festival	None	Bleeding	None
Download Festival	None	None	None
RockNess	None	Phobia, swelling	None

Yom-Tov et al., 2014

When is it worthwhile doing?

- **If** a more delicate sensor is needed





google.org Flu Trends

Google.org home

Dengue Trends

Flu Trends

Home

Select country/region

How does this work?

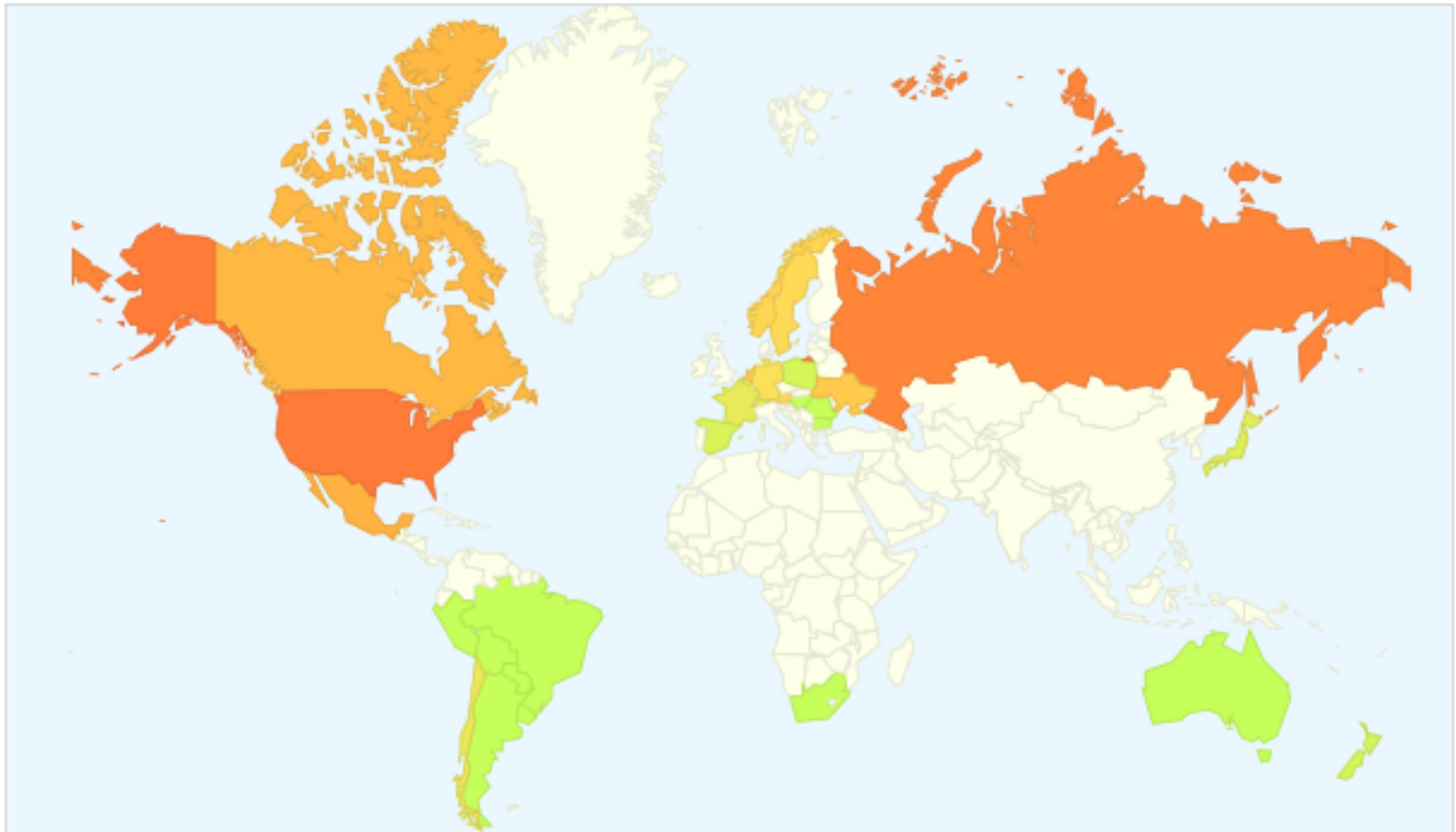
FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

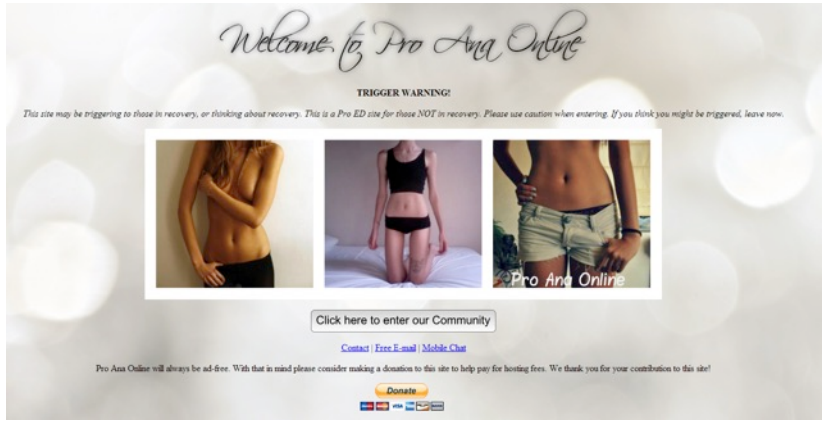
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

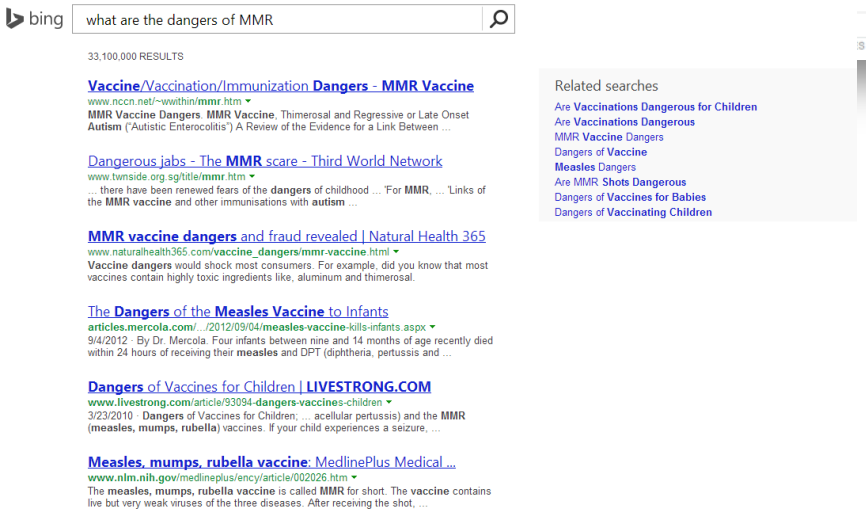


Download world flu activity data - Animated flu trends for Google Earth - Compare flu trends across regions in Public Data Explorer

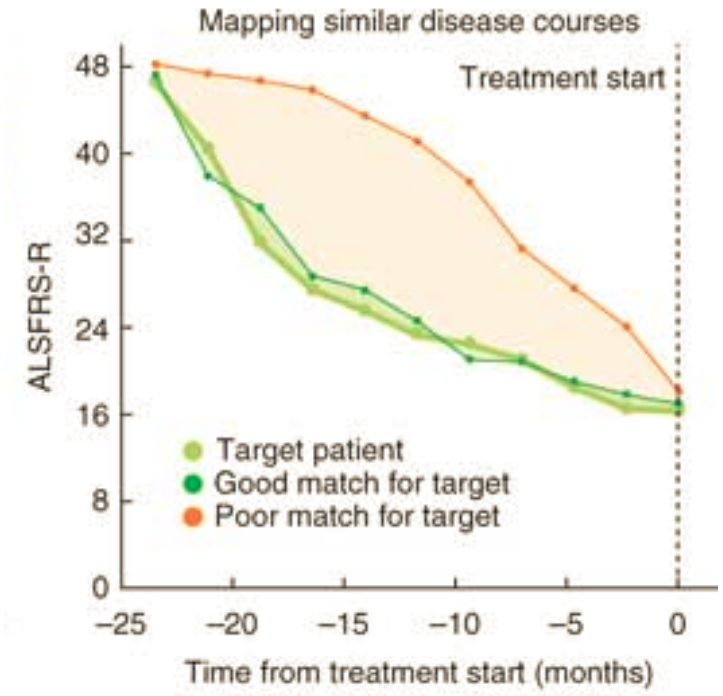
When is it worthwhile doing?

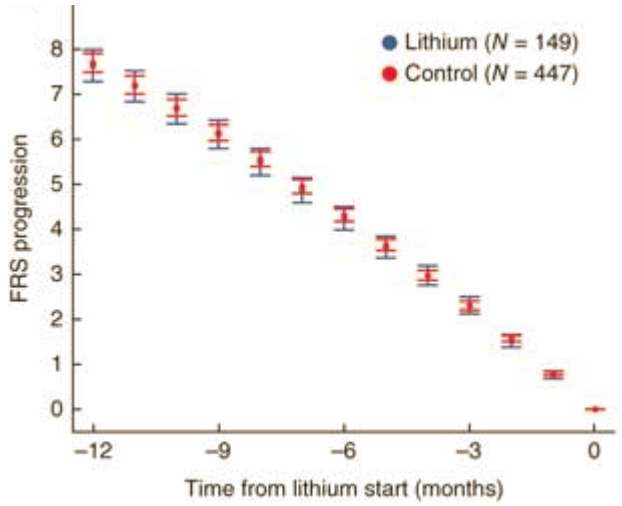
- **If** the activity is largely web-driven





Is Lithium a good treatment for ALS?

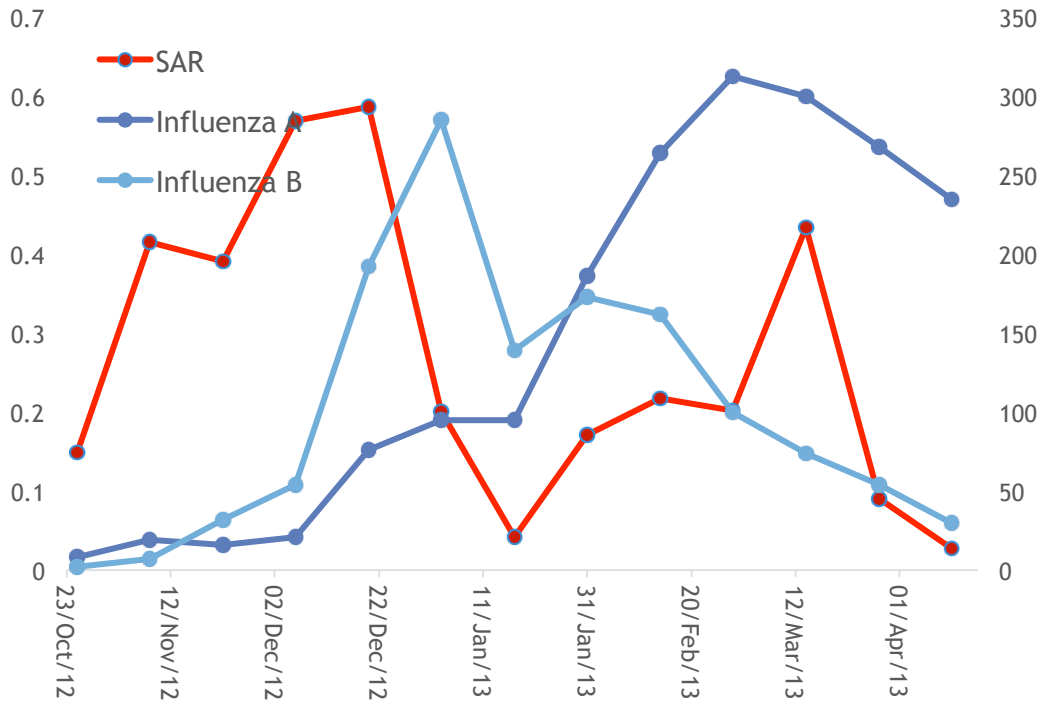


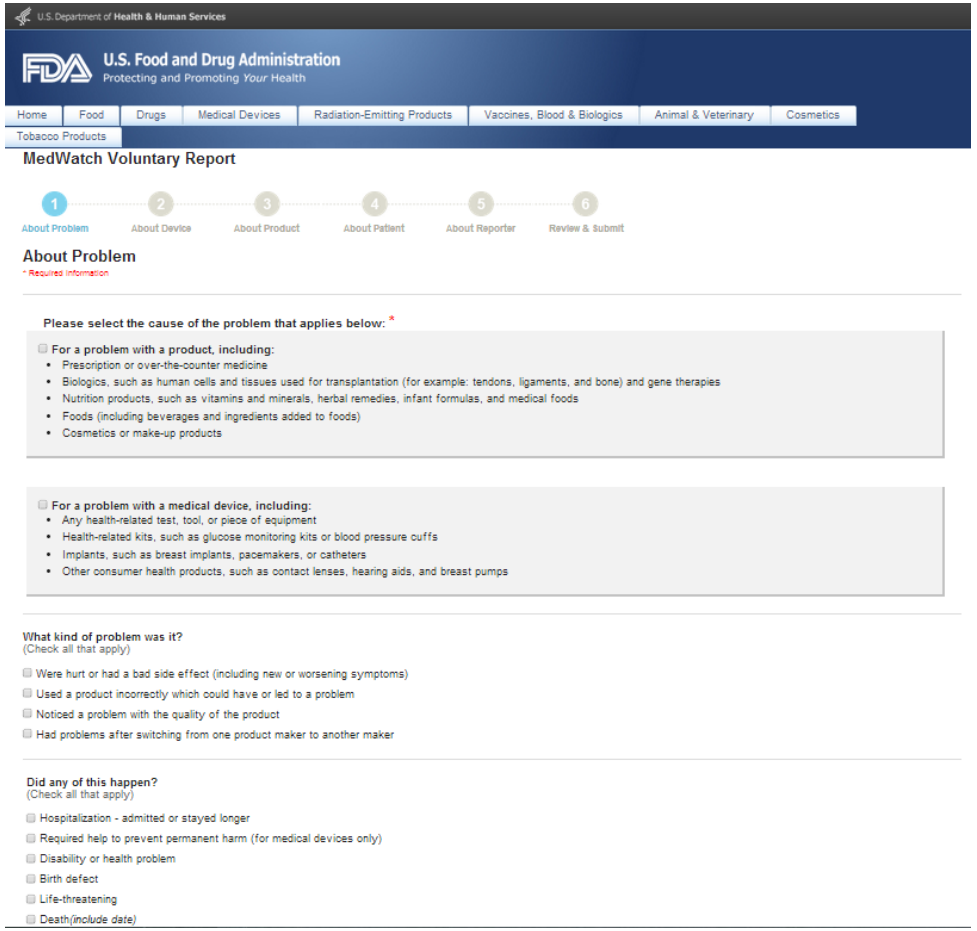


Wicks et al., Nature Biotechnology 2011

When is it worthwhile doing?

- **If** people have a difficulty reporting associations





Vocabulary

- ***Incidence***: The rate of occurrence of new cases of a particular disease in a population
- ***Prevalence***: The percentage of a population that is affected with a particular disease at a given time
- ***Cohort***: A group of people with a shared characteristic (i.e., a disease)

Data sources

Data sources

- Web search
- General social media: Twitter, Facebook, Flickr
- Medical social media: eHealthMe, PatientsLikeMe, TUdiabetes
- Medical Internet aggregators: HealthMap
- Online advertisements
- Public health data
- Other data: Smartphone interaction, Fitness monitors

What we’re not going to talk about

- Small-scale observational studies
  - Qualitative studies and ones based on a very small, subjective, sample
- Studies with a limited CS aspect
  - Limited modelling, small data, only summary statistics, etc.
- (Most likely) Your favorite example

Characteristics of data sources

- Truthfulness
  - Are people providing real information?
- Anonymity and usefulness:
  - What do people say on each? What do they feel comfortable discussing?
  - Personal interest (news, gossip) versus personal medical need
  - Real or imagined?
- Metadata
  - Demographics, medical diagnosis, etc.
- Explicit vs. implicit creation
  - Patient groups versus location data
- Accessibility for research

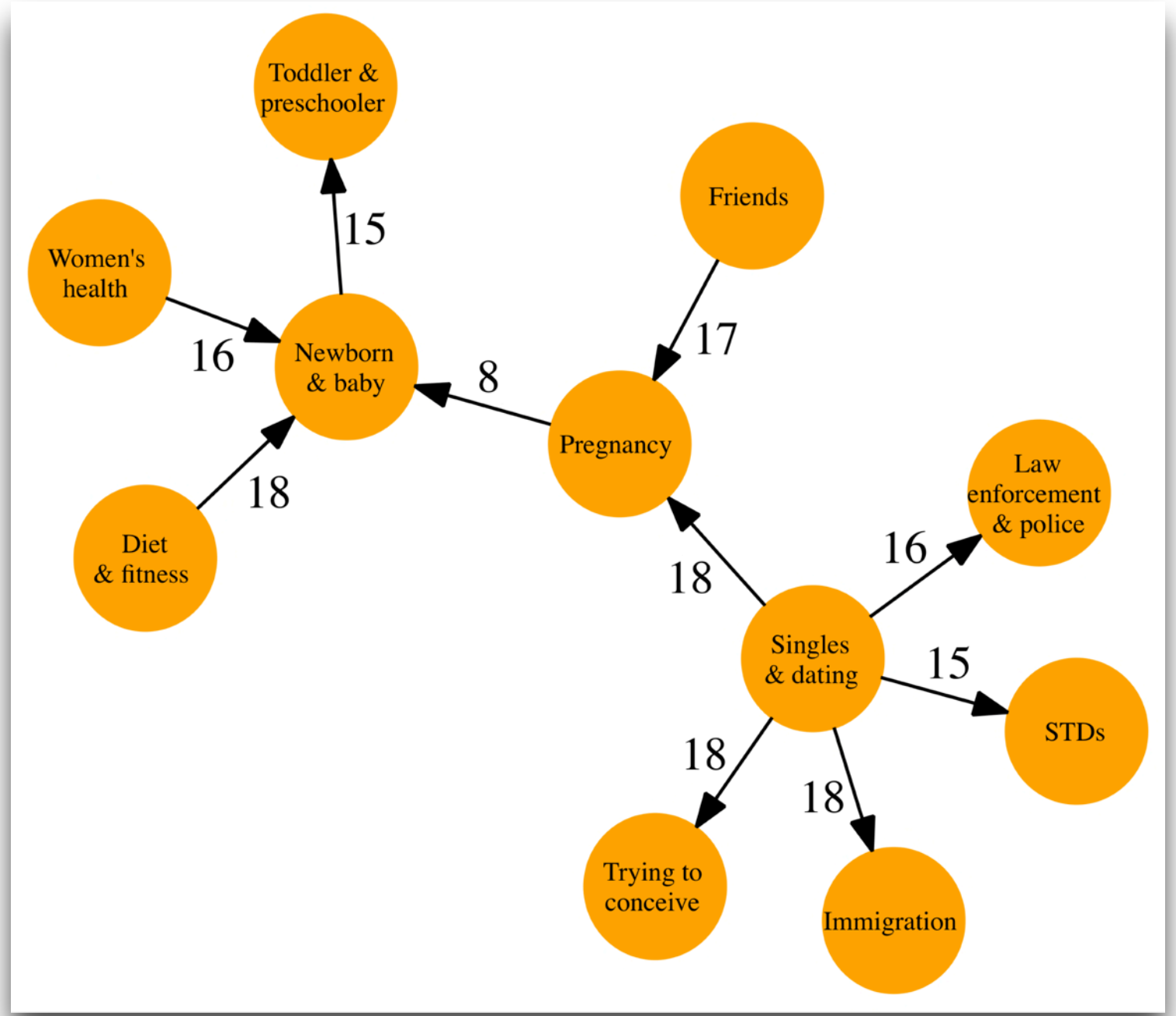
Truthfulness on social media (Pelleg et al., 2012)

- An asker is **truthful** if she reveals her true needs in the question she asks, while an answerer is truthful if she answers to the best of her knowledge in the goal of satisfying the asker

- When truthfulness is **attained**, **social welfare**, the **amount of trade** (volume of user engagement) and **users' utility functions** are maximized.

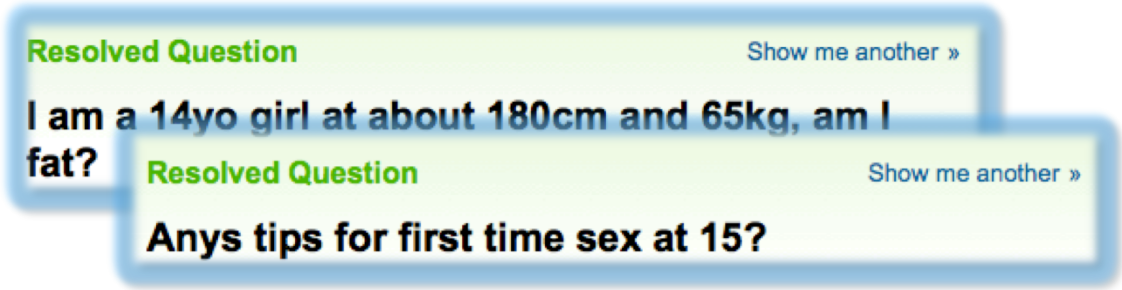
- People are generally more truthful in anonymous media, or when they can take steps to anonymize their identity. They are more careful about truthfulness in topics that (in the WEIRD countries) are:
  - Personal
  - Financial
  - Socially undesirable

- (How do we deal with context: sarcasm, humor, etc. ("Bieber fever")?)


Toddler & preschooler
15
Women's health
16
Newborn & baby
8
Friends
17
Pregnancy
18
Diet & fitness
18
Law enforcement & police
16
Singles & dating
15
STDs
18
Trying to conceive
18
Immigration

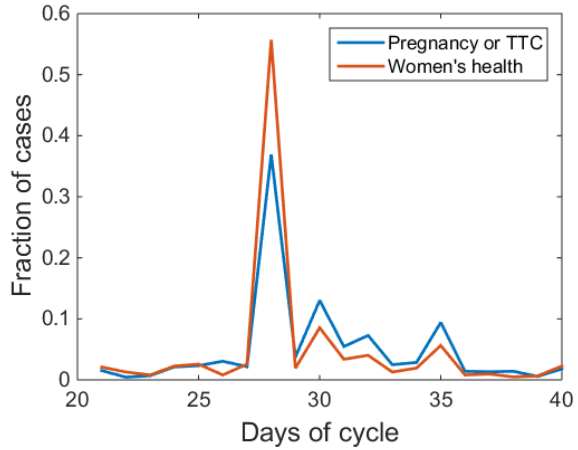
Some anecdotal evidence on truthfulness

	Source	Match
Anthropomorphic data as a function of age	YAnswers	R^2>0.85
BMI per county	YAnswers	R^2=0.31
Age of first intercourse	YAnswers	R^2=0.98
Financial information per county	YAnswers	No statistically significant difference
Gender on registration data	YAnswers	96%
Popularity of medical drugs	Query log	R^2=0.69
Incidence of cancer	Query log	R^2=0.66



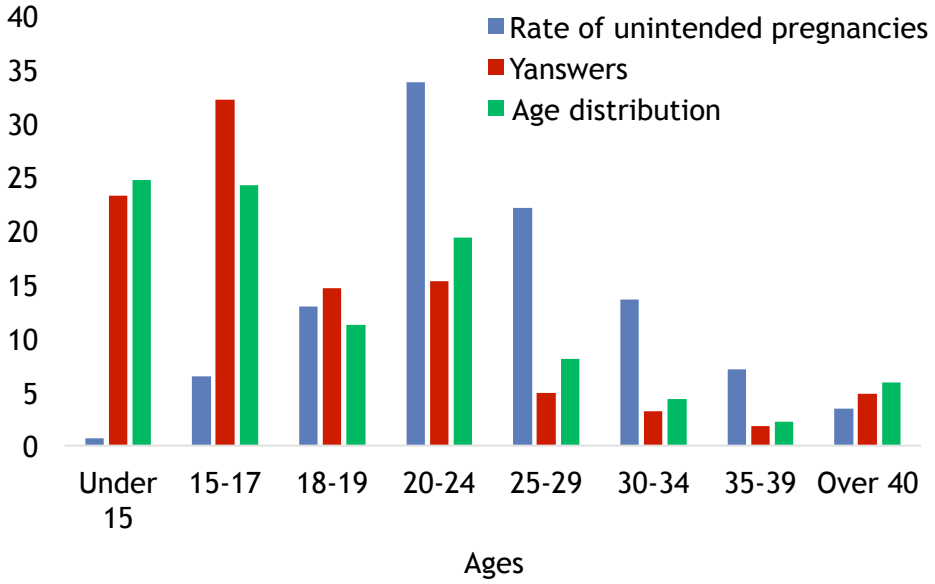
Not all is rosy

- It's important to ask:
  - Why are people posting their data?
  - What is their incentive?
  - What is their demographic distribution?



Yahoo Answers, 4300 questions, unpublished

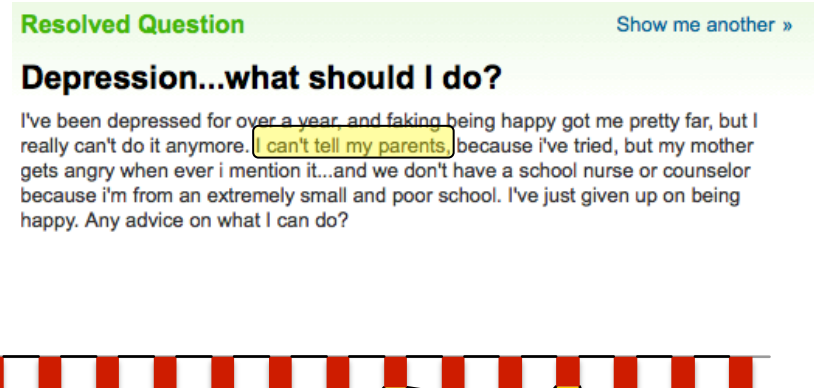
- Outside of patient groups, it is usually easier to find data on:
  - Incidence, not prevalence
  - Abnormal events
  - Acute, not chronic

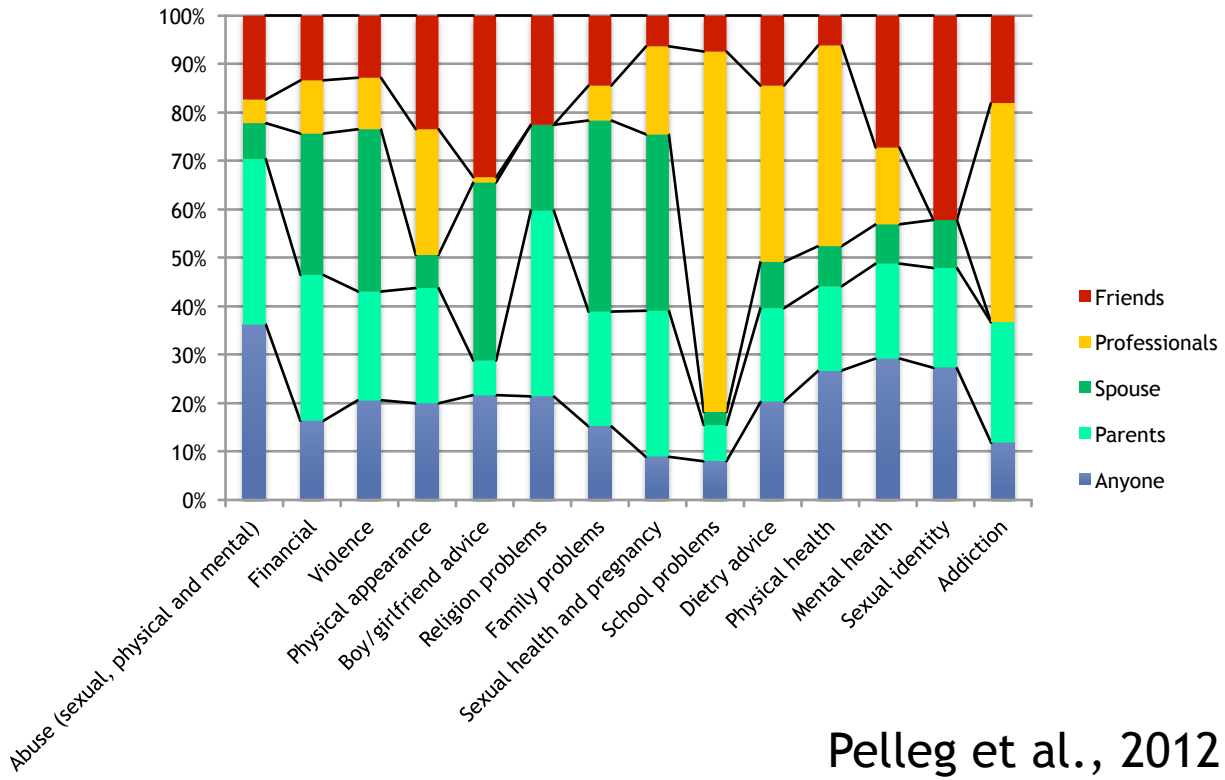


Yahoo Answers, 6200 questions, unpublished`

Anonymity and usefulness

- What do people say on each? What do they feel comfortable discussing?
- Personal interest (news, gossip) versus personal medical need
- Real or imagined?

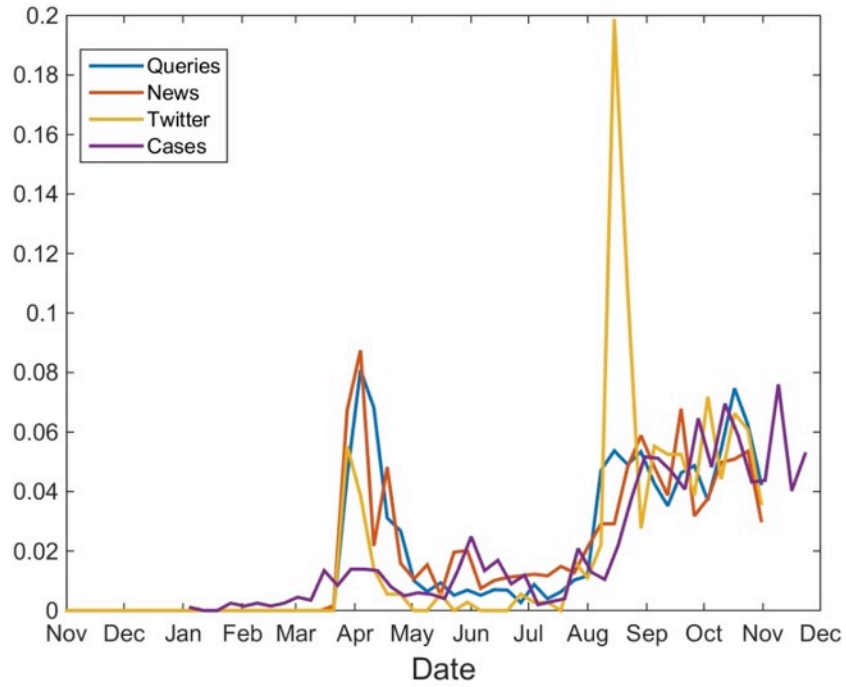




Pelleg et al., 2012

Anonymity and usefulness

- What do people say on each? What do they feel comfortable discussing?
- Personal interest (news, gossip) versus personal medical need
- Real or imagined?



Guinea, unpublished data

Anonymity and usefulness

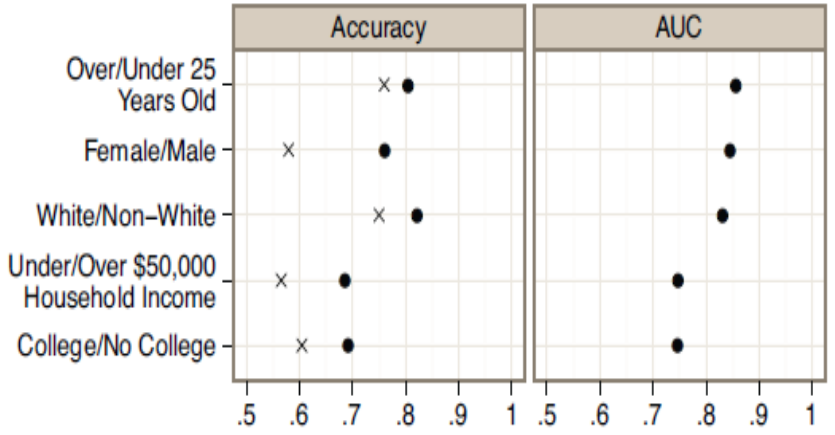
- What do people say on each? What do they feel comfortable discussing?
- Personal interest (news, gossip) versus personal medical need
- Real or imagined?

Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search

RYEN W. WHITE and ERIC HORVITZ

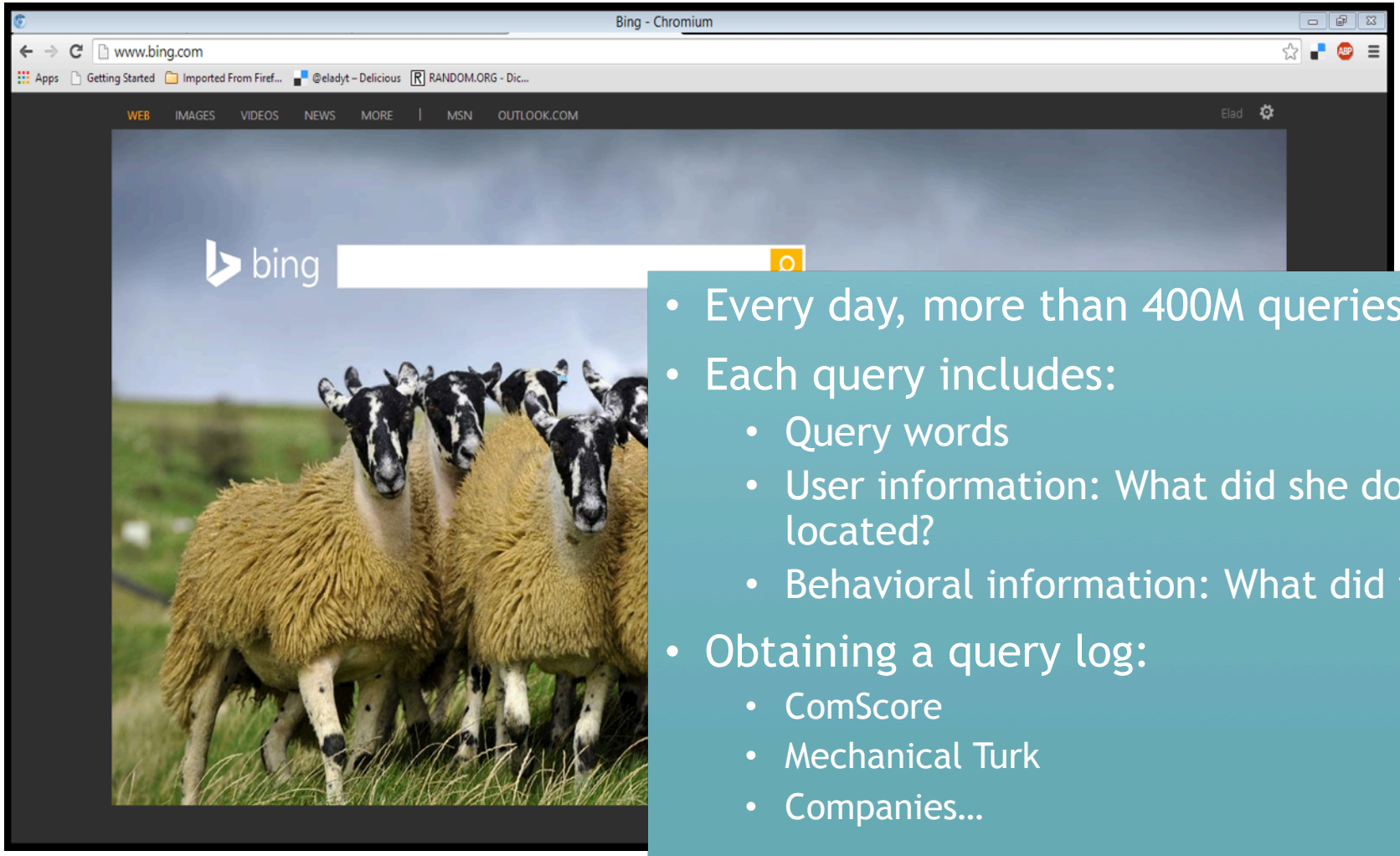
Metadata

- Demographics: Age, gender, location (race, income, education)
- Medical status: Are they the patients?



Goel et al.: *Who does what on the web*

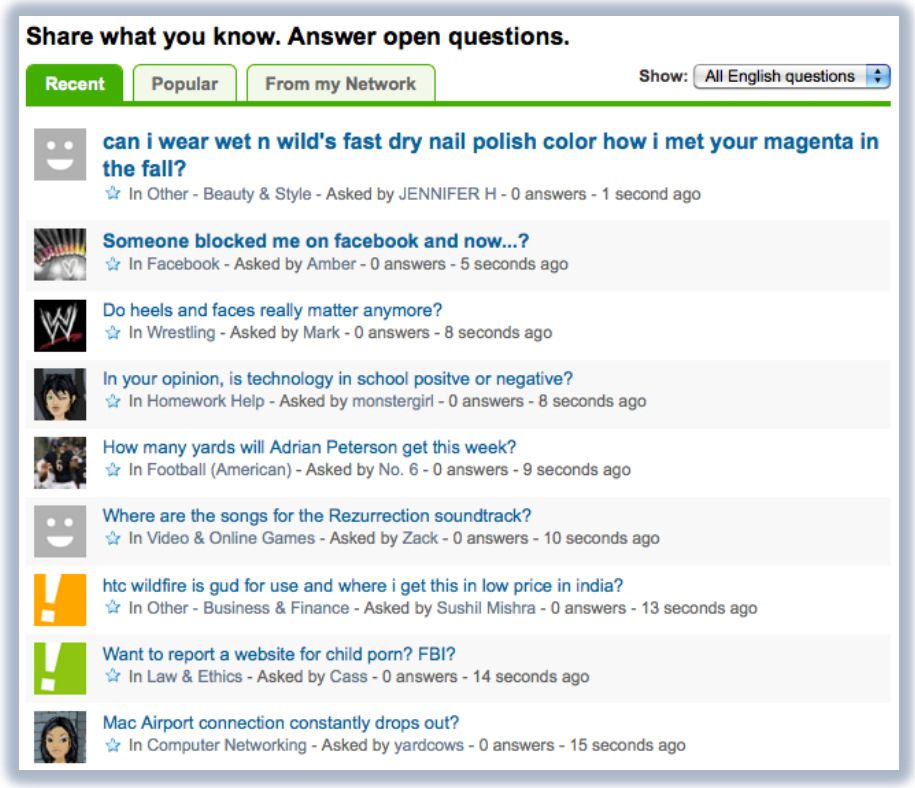
Data sources: Web search



- Every day, more than 400M queries are submitted in the USA
- Each query includes:
  - Query words
  - User information: What did she do in the past? Where is she located?
  - Behavioral information: What did the user do?
- Obtaining a query log:
  - ComScore
  - Mechanical Turk
  - Companies…

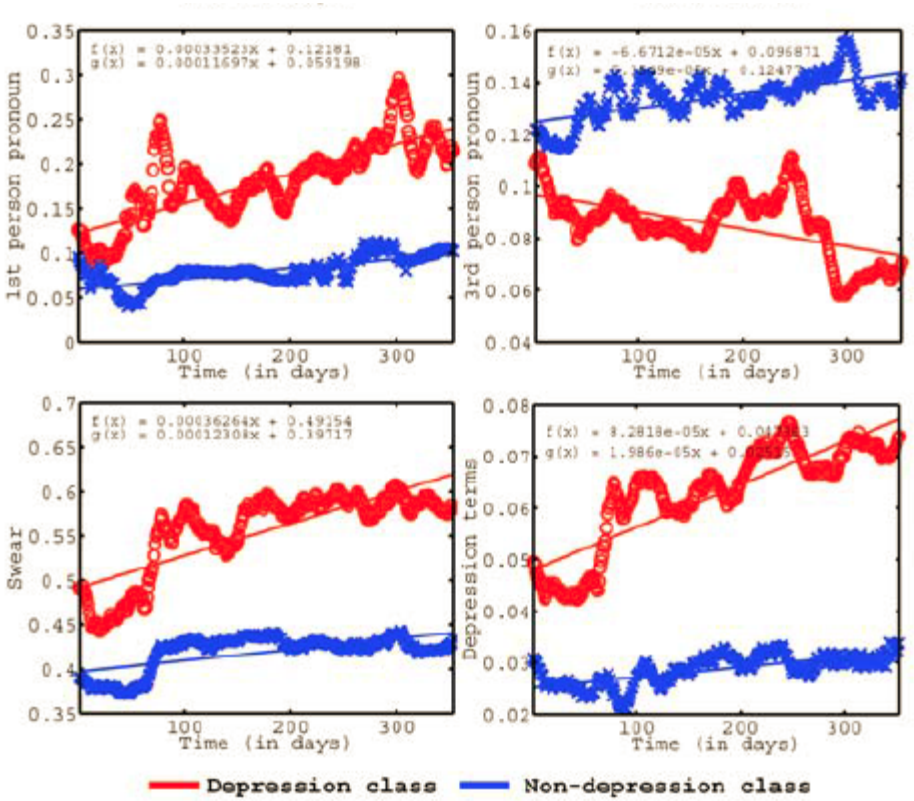
(Manual) web search

- Over 200M questions
- About 10 years of data
- Categorized into ~1700 categories



General social media: Twitter, Facebook, Flickr

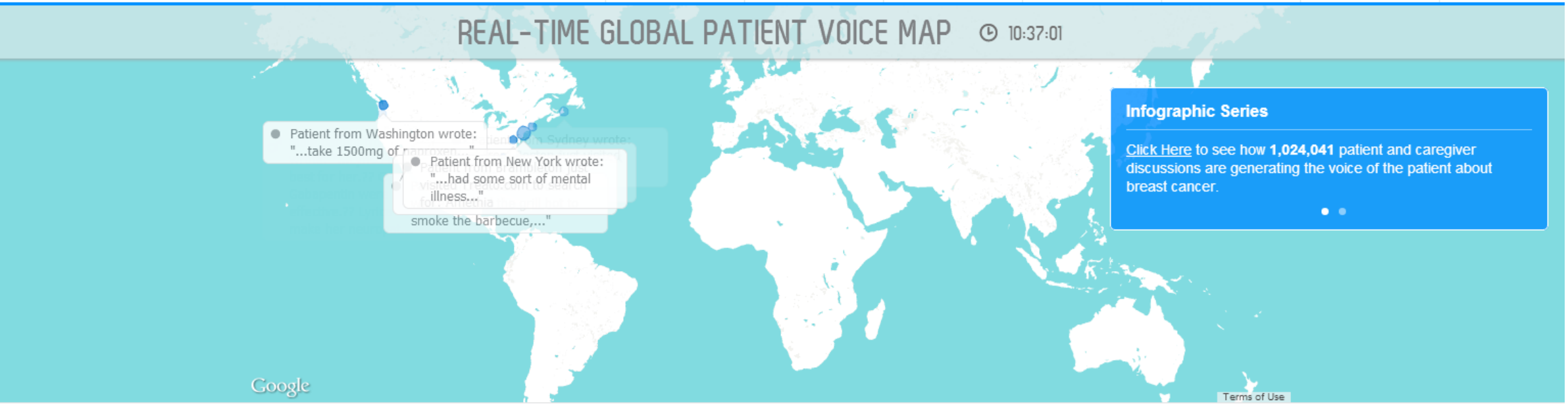
- **Truthfulness**: Dependent on anonymity and sensitivity
- Both **explicit** (patient groups, disease support) and **implicit** (flu reports) data
- Small scale data is generally available (in collated datasets or through crawl)



De Choudhury et al., 2012

TREATO
The voice of the patient

PLATFORM | SOLUTIONS | TECHNOLOGY | LEARN MORE | MEET TREATO | CONTACT | LOG IN

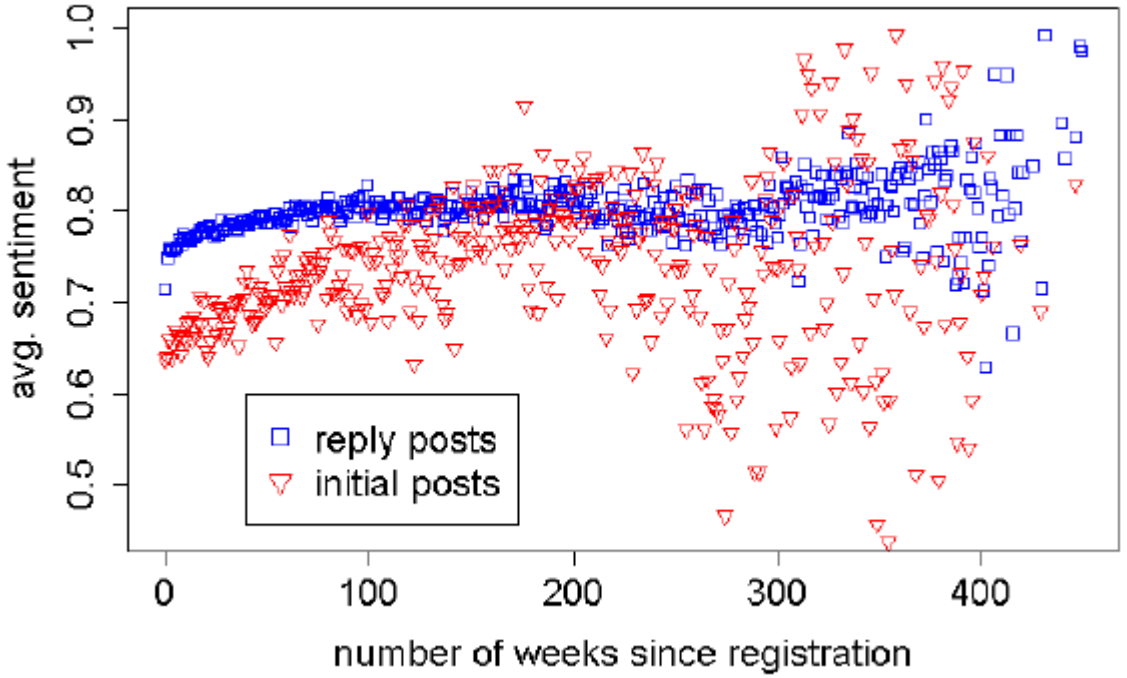


1 , 751 , 220 , 050 Patient conversations world-wide

BRINGING YOU THE VOICE OF THE PATIENT

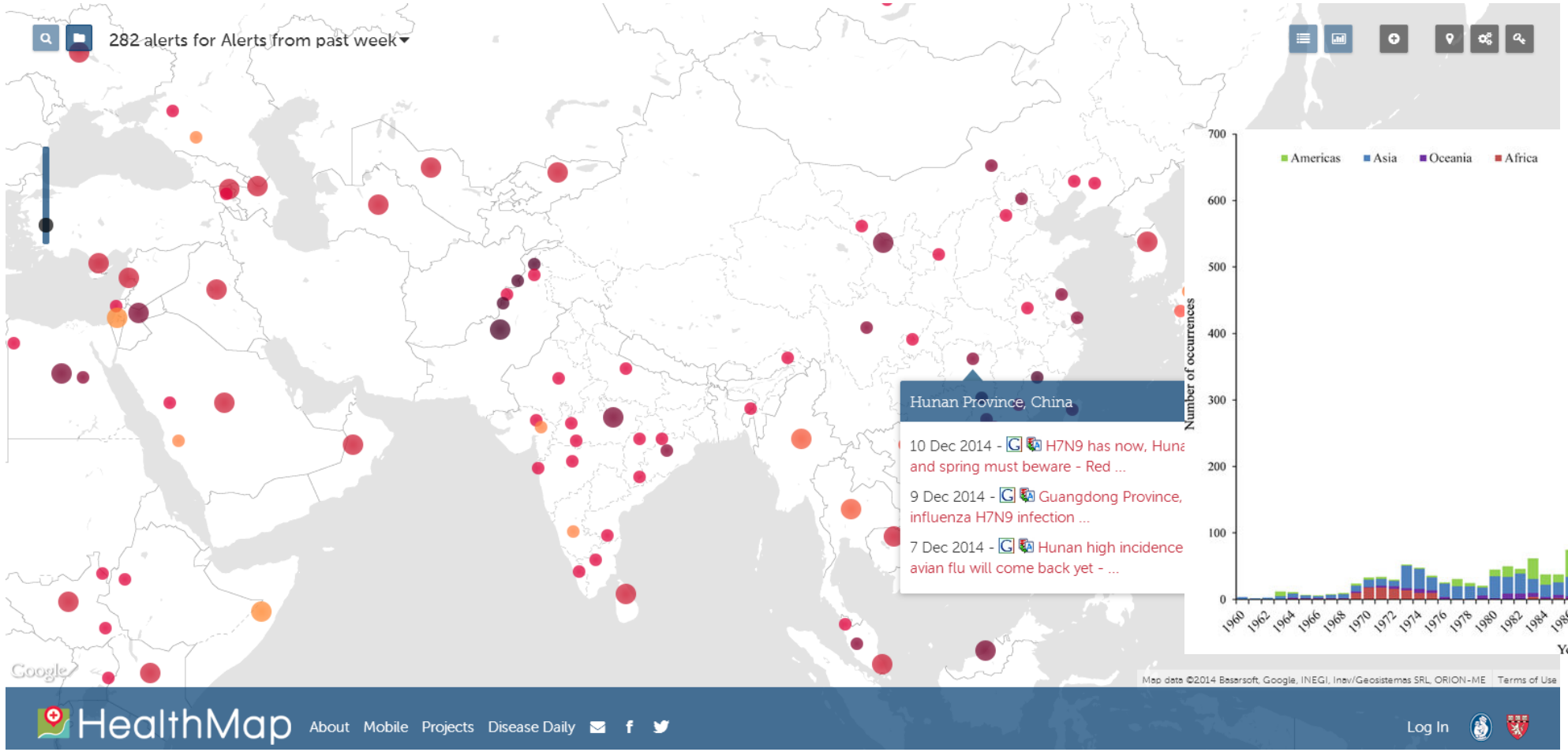
Medical social media: People gathering to discuss their specific predicament

- Examples: eHealthMe, PatientsLikeMe, TUdiabetes
- Truthfulness is usually high.
- Data availability can be a (legal) problem

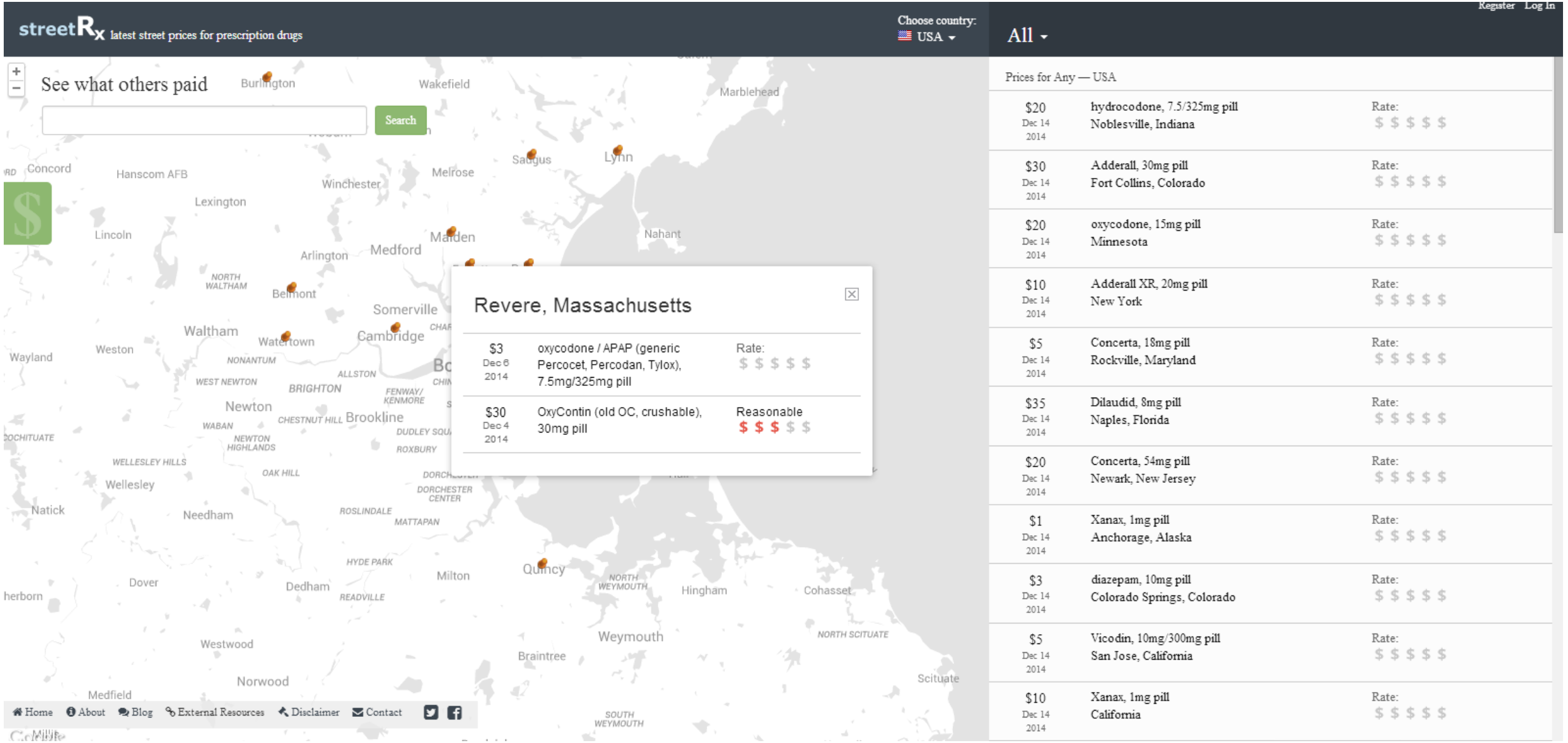


Zhang et al., 2014

Medical Internet aggregators: HealthMap



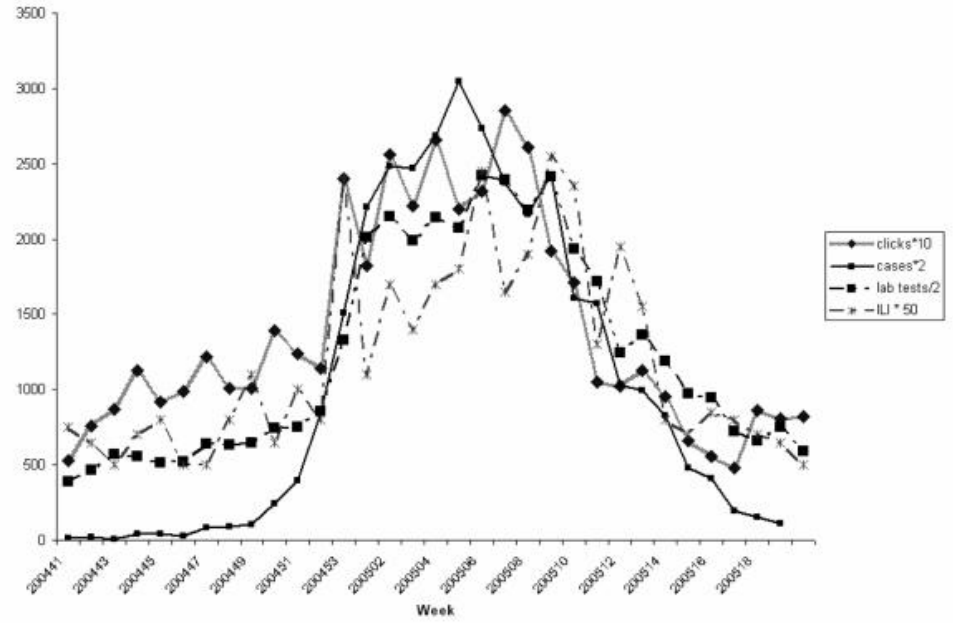
Messina et al., 2014


streetRx latest street prices for prescription drugs
Choose country:
USA
Register
Log In
All
See what others paid
Search
Burlington
Wakefield
Marblehead
Saugus
Lynn
Concord
Hanscom AFB
Melrose
Winchester
Lexington
Lincoln
Malden
Nahant
Arlington
Medford
NORTH WALTHAM
Belmont
Somerville
Waltham
Cambridge
Watertown
Wayland
Weston
NONANTUM
ALLSTON
WEST NEWTON
BRIGHTON
FENWAY/ KENMORE
Newton
CHESTNUT HILL
Brookline
WABAN
NEWTON HIGHLANDS
DUDLEY SQUA
ROXBURY
WELLESLEY HILLS
OAK HILL
Wellesley
DORCHESTER CENTER
Natick
Needham
ROSLINDALE
MATTAPAN
HYDE PARK
Milton
Quincy
Dover
Dedham
READVILLE
NORTH WEYMOUTH
Hingham
Cohasset
Weymouth
NORTH SCITUATE
Braintree
Westwood
Norwood
Scituate
Medfield
SOUTH WEYMOUTH
Revere, Massachusetts
$3 Dec 6 2014
oxycodone / APAP (generic Percocet, Percodan, Tylox), 7.5mg/325mg pill
Rate: $ $ $ $ $
$30 Dec 4 2014
OxyContin (old OC, crushable), 30mg pill
Reasonable $ $ $ $ $
Prices for Any — USA
$20 Dec 14 2014
hydrocodone, 7.5/325mg pill
Noblesville, Indiana
Rate: $ $ $ $ $
$30 Dec 14 2014
Adderall, 30mg pill
Fort Collins, Colorado
Rate: $ $ $ $ $
$20 Dec 14 2014
oxycodone, 15mg pill
Minnesota
Rate: $ $ $ $ $
$10 Dec 14 2014
Adderall XR, 20mg pill
New York
Rate: $ $ $ $ $
$5 Dec 14 2014
Concerta, 18mg pill
Rockville, Maryland
Rate: $ $ $ $ $
$35 Dec 14 2014
Dilaudid, 8mg pill
Naples, Florida
Rate: $ $ $ $ $
$20 Dec 14 2014
Concerta, 54mg pill
Newark, New Jersey
Rate: $ $ $ $ $
$1 Dec 14 2014
Xanax, 1mg pill
Anchorage, Alaska
Rate: $ $ $ $ $
$3 Dec 14 2014
diazepam, 10mg pill
Colorado Springs, Colorado
Rate: $ $ $ $ $
$5 Dec 14 2014
Vicodin, 10mg/300mg pill
San Jose, California
Rate: $ $ $ $ $
$10 Dec 14 2014
Xanax, 1mg pill
California
Rate: $ $ $ $ $
Home
About
Blog
External Resources
Disclaimer
Contact

Actively collecting data

- Mechanical Turk \ CrowdFlower
- eLance \ oDesk
- Online advertising
- Online surveys

Online advertisements



From: Eysenbach, 2006

MMR vaccine

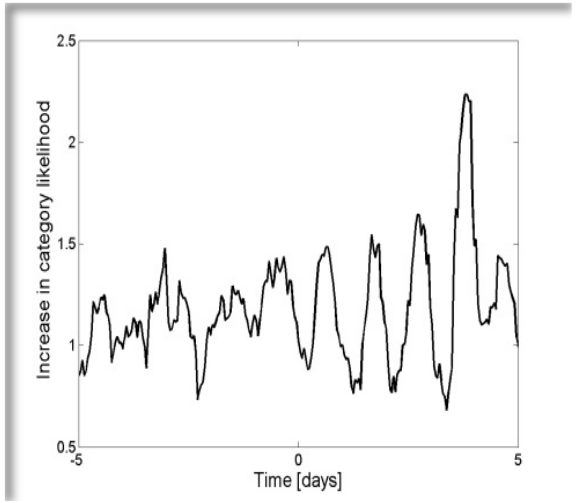
http://tiny.cc/wf497w

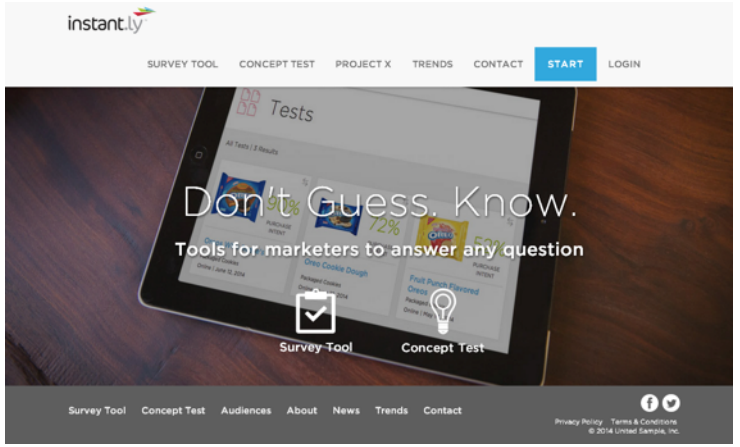
Do you want to learn about the ~~importance~~ of this vaccine?

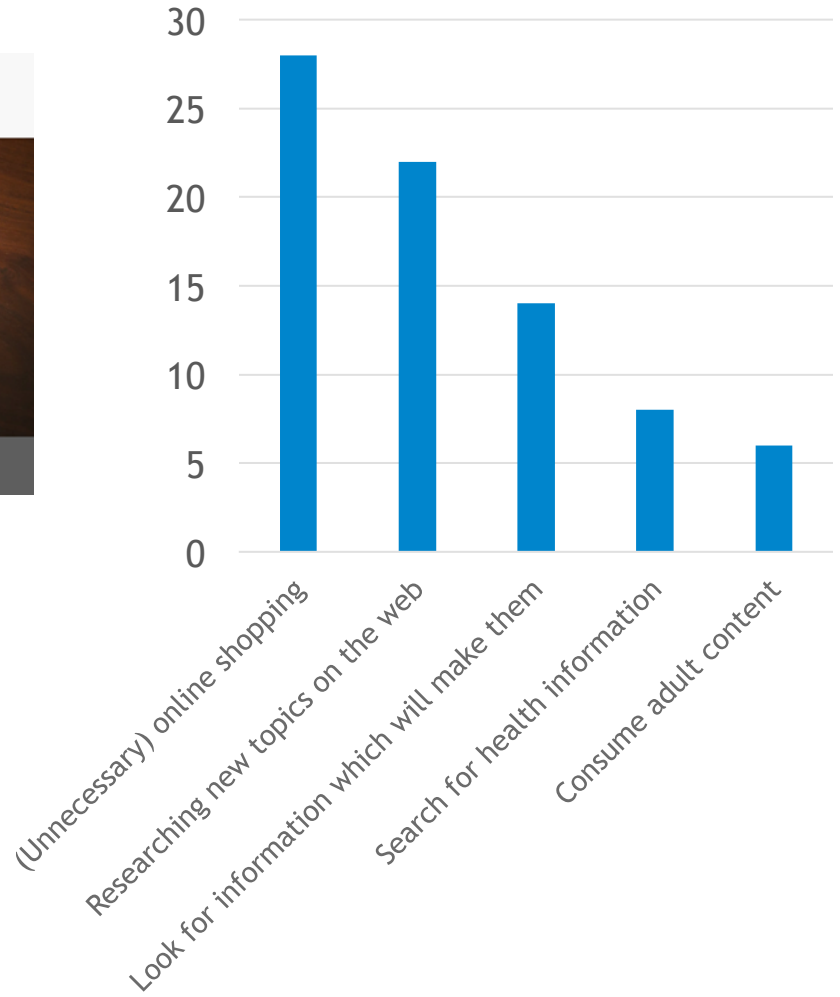
dangers

	Advertisement	
	Anti	Pro
Low VAS	0.556	0.468
High VAS	0.472	1.197

Validating findings using online surveys

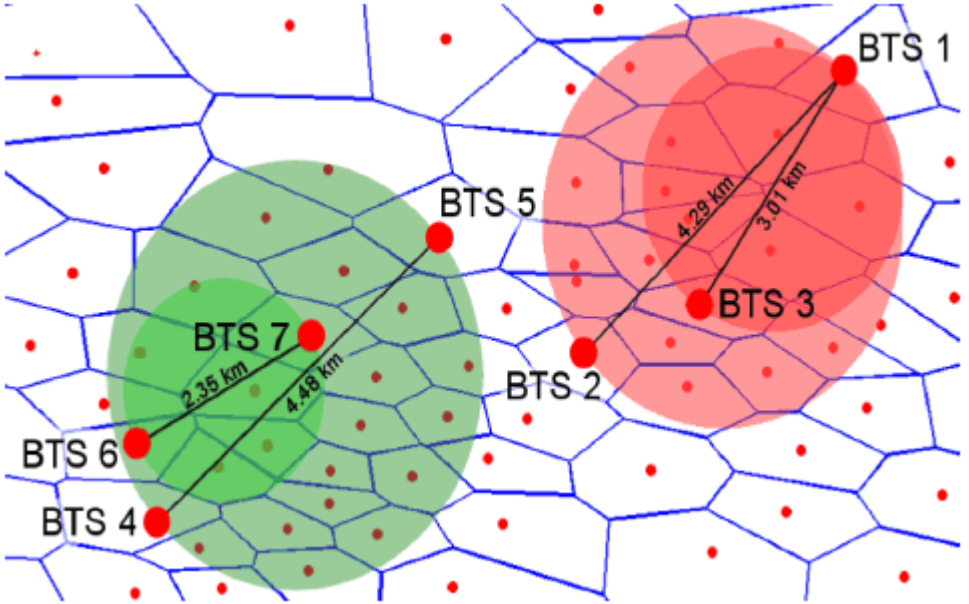




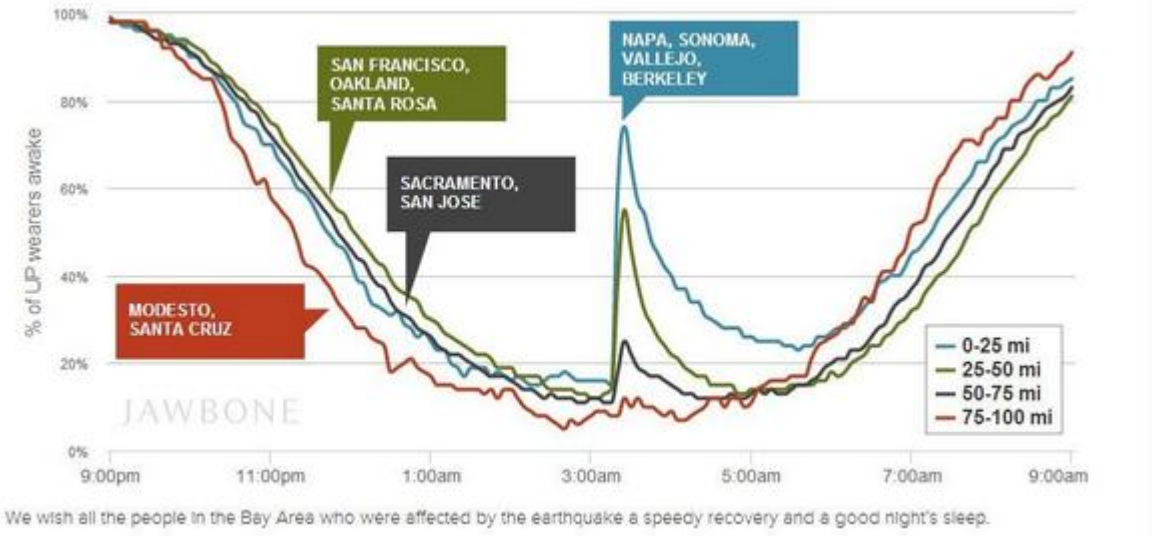


Other data

- Smartphone interaction:
  - Human mobility patterns during the 2009 Mexico influenza pandemic
  - Surveys (Hygiene and Tropical Medicine)
- Fitness monitors
- Internet of Things (IoT)



Frias-Martinez et al., 2012

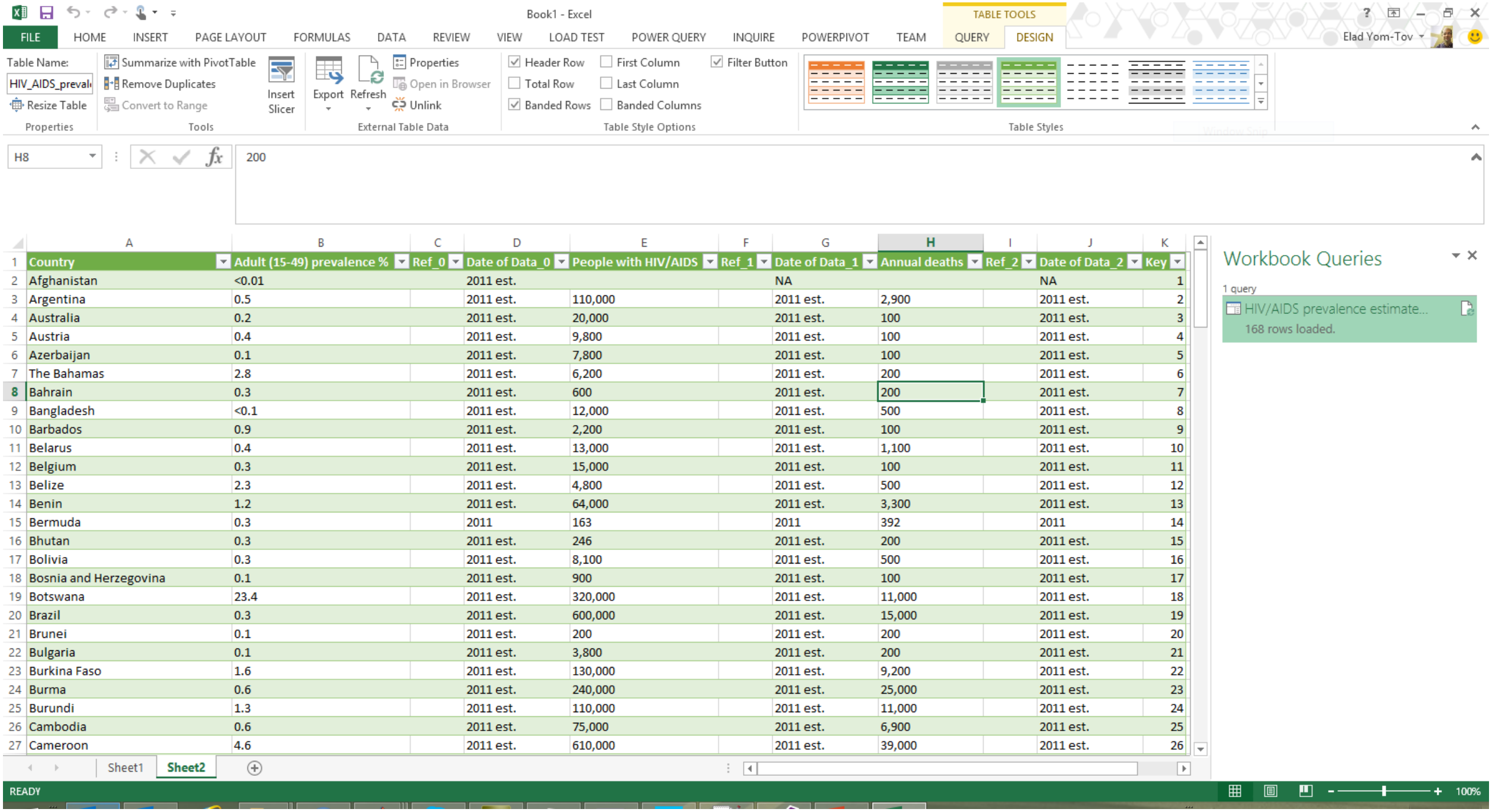


Summary

Source	Truthfulness	Anonymity and usefulness	Metadata	Creation	Accessibility for research
Web search	High	High	Rare	Implicit	Within companies or via toolbars
General social media	Low	Low-medium	Available	Explicit	Through hoses or scraping
Medical social media	Medium-High	High	Common	Explicit	Usually via scraping
Medical internet aggregators	High	Medium	--	Explicit	?
Smartphone interaction	High	Medium	None	Implicit	Very difficult
Actively collecting data	Variable	Medium	Available	Explicit	Easy – Make your own!

Public health data: Linking to ground-truth data

Authority	Links
Centers for Disease Control (CDC)	http://wonder.cdc.gov/ http://www.cdc.gov/datastatistics/index.html http://www.cdc.gov/flu/weekly/ http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm https://www.healthdata.gov/dataset/search
World Health Organization (WHO)	http://www.who.int/healthinfo/global_burden_disease/en/ http://apps.who.int/gho/data/?theme=main
Dartmouth College	http://www.dartmouthatlas.org/
Public Health England	https://www.gov.uk/government/collections/seasonal-influenza-guidance-data-and-analysis
Dbpedia	http://wiki.dbpedia.org/Datasets
Other	http://www.ehdp.com/vitalnet/datasets.htm http://phpartners.org/health_stats.html



	A	B	C	D	E	F	G	H	I	J	K
1	Country	Adult (15-49) prevalence %	Ref_0	Date of Data_0	People with HIV/AIDS	Ref_1	Date of Data_1	Annual deaths	Ref_2	Date of Data_2	Key
2	Afghanistan	<0.01		2011 est.			NA			NA	1
3	Argentina	0.5		2011 est.	110,000		2011 est.	2,900		2011 est.	2
4	Australia	0.2		2011 est.	20,000		2011 est.	100		2011 est.	3
5	Austria	0.4		2011 est.	9,800		2011 est.	100		2011 est.	4
6	Azerbaijan	0.1		2011 est.	7,800		2011 est.	100		2011 est.	5
7	The Bahamas	2.8		2011 est.	6,200		2011 est.	200		2011 est.	6
8	Bahrain	0.3		2011 est.	600		2011 est.	200		2011 est.	7
9	Bangladesh	<0.1		2011 est.	12,000		2011 est.	500		2011 est.	8
10	Barbados	0.9		2011 est.	2,200		2011 est.	100		2011 est.	9
11	Belarus	0.4		2011 est.	13,000		2011 est.	1,100		2011 est.	10
12	Belgium	0.3		2011 est.	15,000		2011 est.	100		2011 est.	11
13	Belize	2.3		2011 est.	4,800		2011 est.	500		2011 est.	12
14	Benin	1.2		2011 est.	64,000		2011 est.	3,300		2011 est.	13
15	Bermuda	0.3		2011	163		2011	392		2011	14
16	Bhutan	0.3		2011 est.	246		2011 est.	200		2011 est.	15
17	Bolivia	0.3		2011 est.	8,100		2011 est.	500		2011 est.	16
18	Bosnia and Herzegovina	0.1		2011 est.	900		2011 est.	100		2011 est.	17
19	Botswana	23.4		2011 est.	320,000		2011 est.	11,000		2011 est.	18
20	Brazil	0.3		2011 est.	600,000		2011 est.	15,000		2011 est.	19
21	Brunei	0.1		2011 est.	200		2011 est.	200		2011 est.	20
22	Bulgaria	0.1		2011 est.	3,800		2011 est.	200		2011 est.	21
23	Burkina Faso	1.6		2011 est.	130,000		2011 est.	9,200		2011 est.	22
24	Burma	0.6		2011 est.	240,000		2011 est.	25,000		2011 est.	23
25	Burundi	1.3		2011 est.	110,000		2011 est.	11,000		2011 est.	24
26	Cambodia	0.6		2011 est.	75,000		2011 est.	6,900		2011 est.	25
27	Cameroon	4.6		2011 est.	610,000		2011 est.	39,000		2011 est.	26

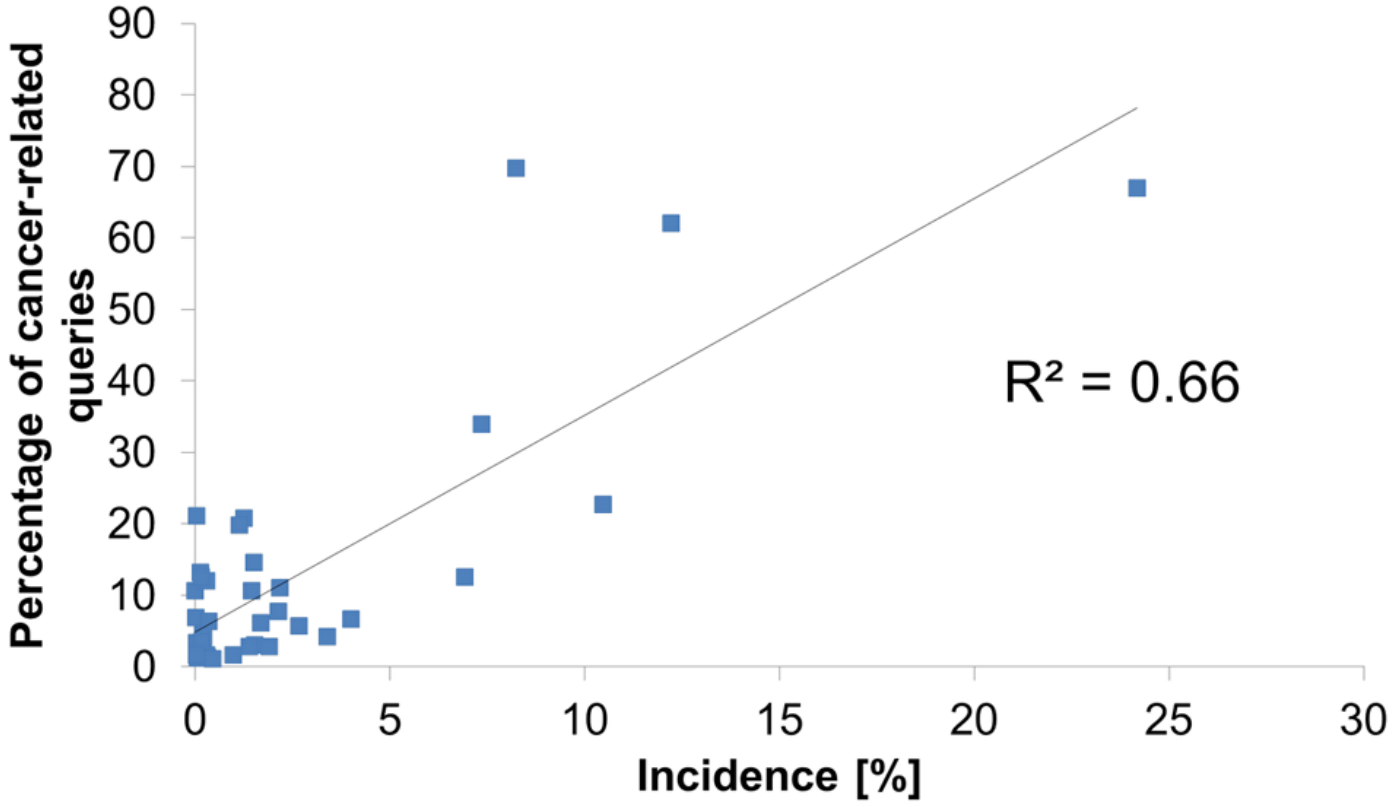
Linking to ground truth

Linking to ground truth

- Validate a cohort
- Train a predictive model
- Validate the prediction model
- Find interesting disagreements with the prediction model

Using ground truth data

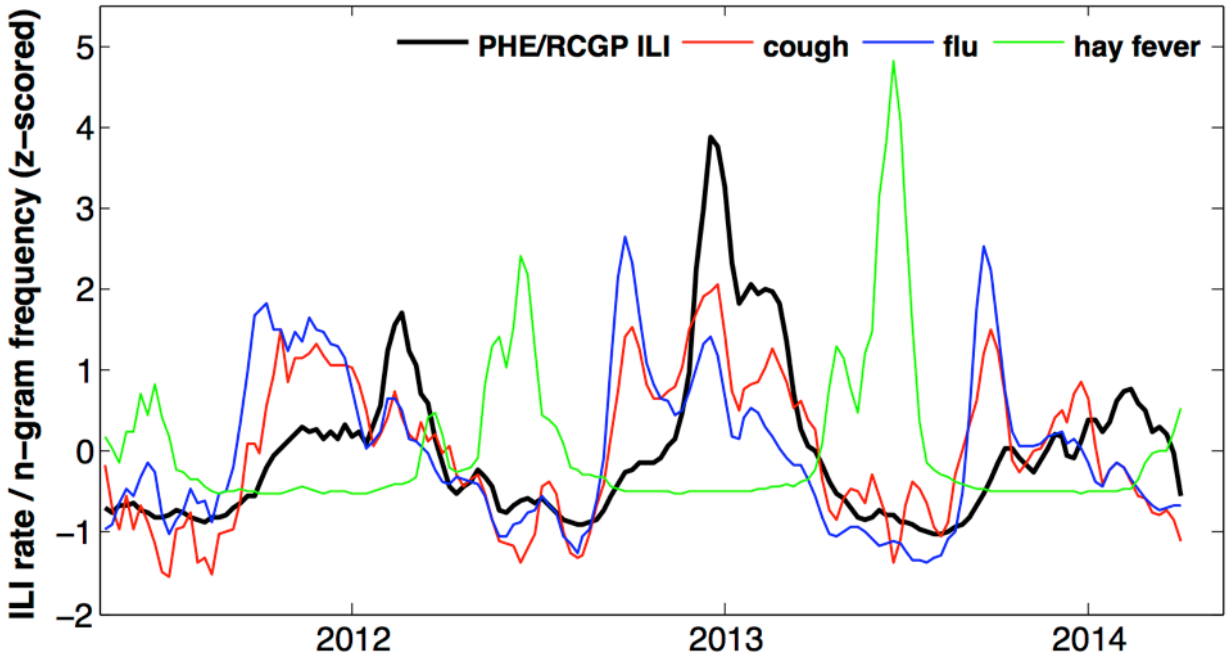
To validate a cohort, that is, that the population under study is (mostly) of patients:

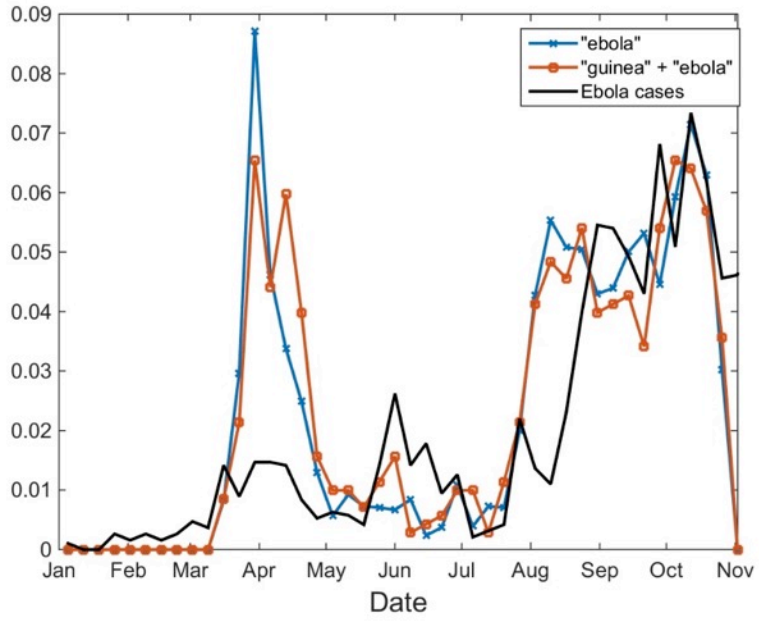


Ofran et al., 2012

Using ground truth data (2)

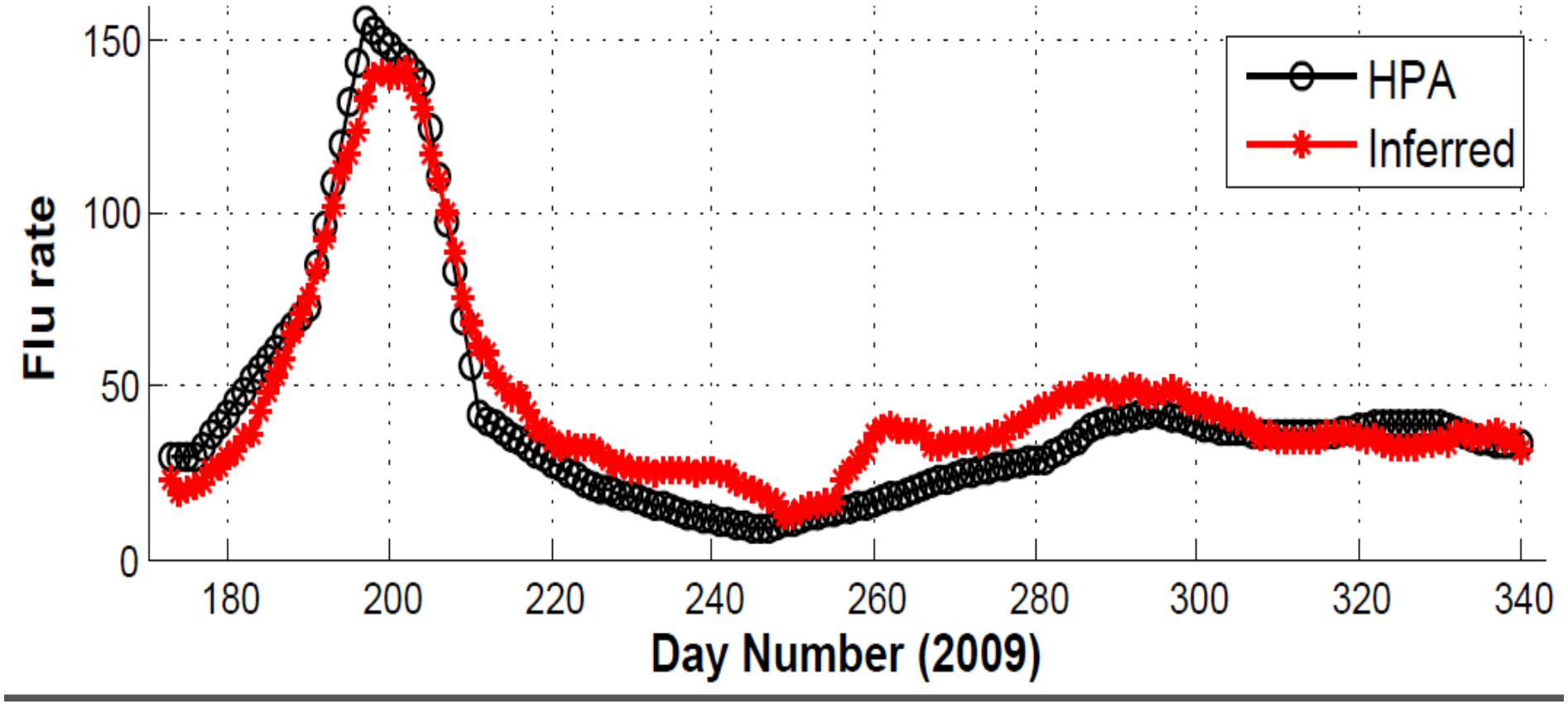
To train a predictive model:





Using ground truth data (3)

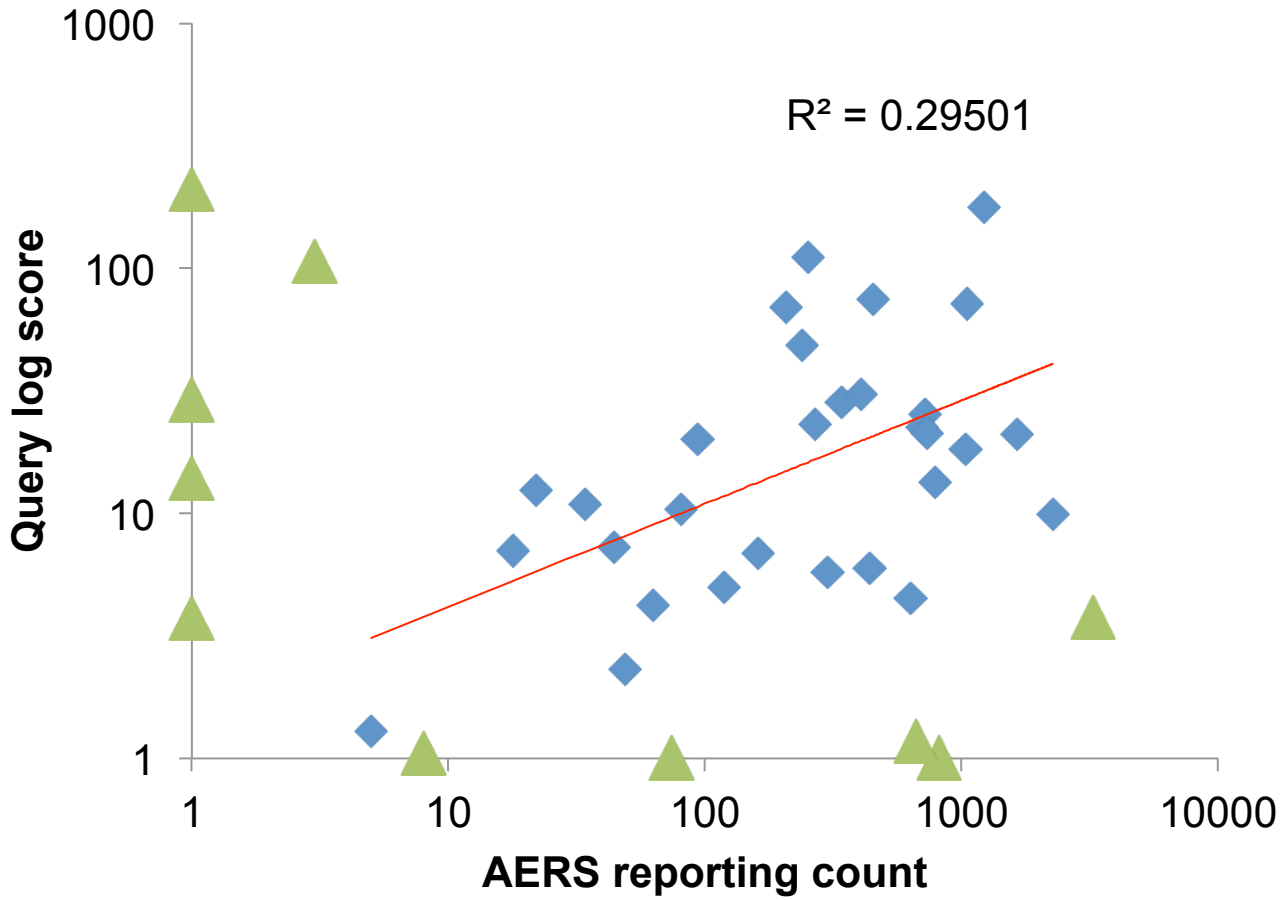
To validate the prediction model:



Lampos and Cristianini, 2010

Using ground truth data (4)

To find interesting disagreements with the prediction model:



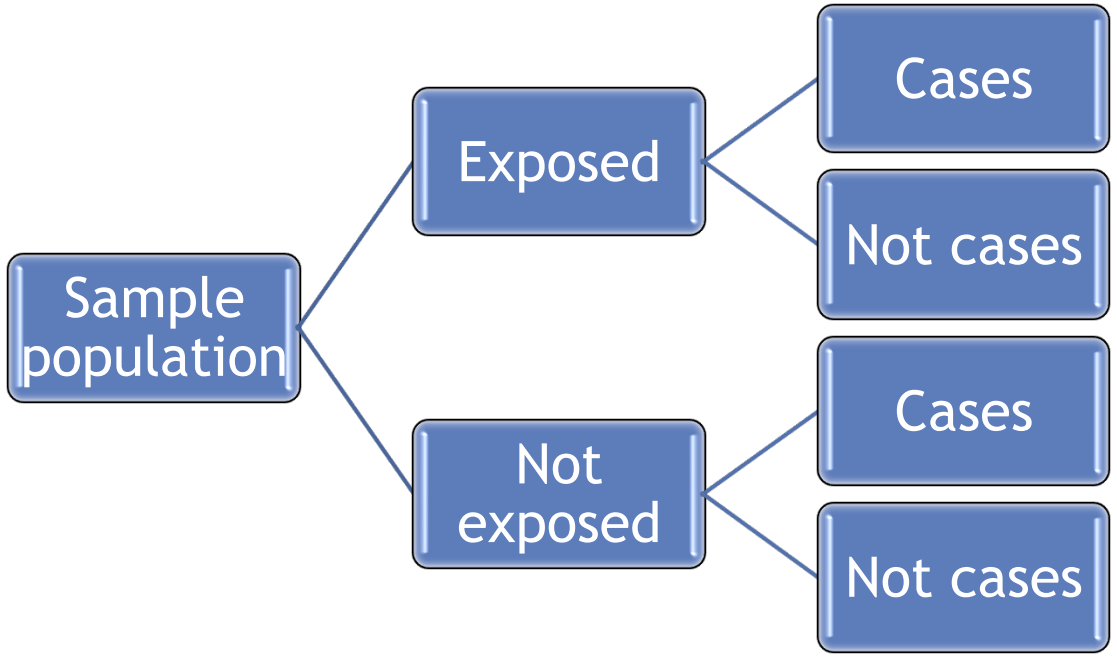
Identifying a cohort

Study Types

- Cross-Sectional Studies
- Cohort Studies
- Case-Control Studies
- Intervention Studies

Cross-Sectional Study - Definition

- Observational study
- Data is collected at a defined time, not long term
- Typically carried out to measure the prevalence of a disease in a population



Cross-Sectional Studies - Self-Selection

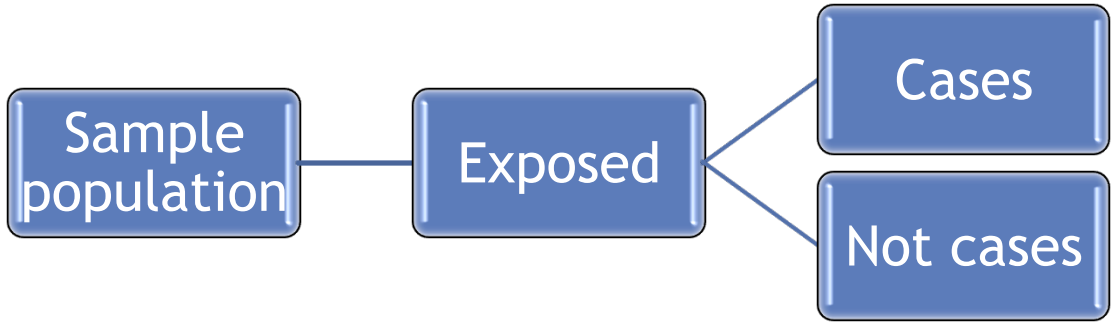
- Selection bias
  - Self-selected participants might not be representative of the population of interest
- Use cases
  - Hypothesis building
  - Reaching hidden populations
- Example: Simmons et al. used a cross-sectional study for hypothesis building. They posted an anonymous questionnaire on websites targeted multiple sclerosis patients. The patients were asked which factors in their opinion were improving or worsening their multiple sclerosis symptoms.

Cross-Sectional Study – Digital Trail

- Mislove (2011) looks at the demographic distribution of Twitter users in the U.S. based on information about Twitter users representing 1% of the U.S. Population
  - Their is an over-representation of people living in highly populated areas, while sparsely populated regions are under-represented
  - Male bias, but it is declining
  - The distribution of races differs from each county, but does not follow the actual distribution
- Knowing the demographics makes is possible to adjust the bias of the collected data
- Example:
  - Messina (2014) used aggregated information from medical journals together with news articles to build a map of the prevalence of dengue fever across the world

Cohort study - Definition

- Observational study
- Studies a group of people with some common characteristic or experience for a period of time



Cohort studies - Self-Selection

- Well suited for an internet based approach
- Inexpensive and efficient follow-up
- Can easily be ported to other geographical locations
- Example: NINFEA a multipurpose cohort study investigating certain exposures during prenatal and early postnatal life on infant, child and adult health. 85–90% response rate when using both email and phone calls.

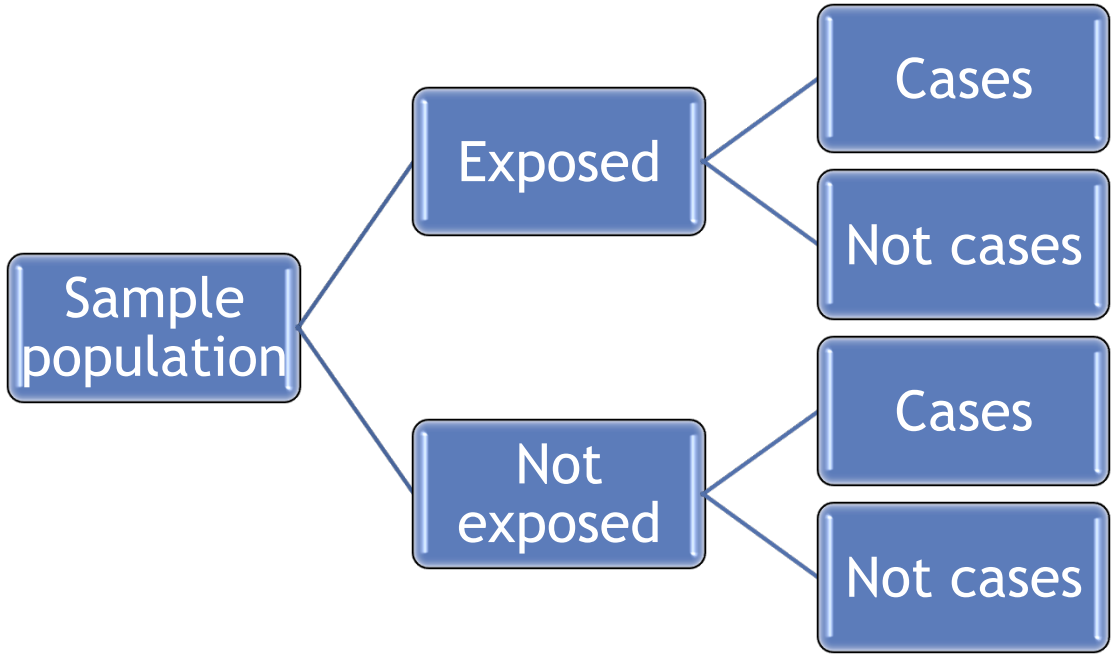
Cohort studies – Digital Trail

- Selecting the cohort
  - Geo-location
  - Self diagnosis, e.g. querying “I have a bad knee”
  - Showing interest in a topic, e.g. querying about specific cancer types

- Examples
  - Ofran et al. (2012) used query logs to identify the information needs of cancer patients
  - Yom-Tov et al. (2015) used query logs to identify people with specific health events and afterwards evaluated whether specific online behavior was predictive of the event
  - Lampos (2010) used tweets to predict the prevalence of ILI in several regions in UK. http://geopatterns.enm.bris.ac.uk/epidemics/

Case-Control Study - Definition

- Observational study
- Studies two groups; cases and controls
  - Cases – people with the condition of interest
  - Controls – people at risk of becoming a case
- Both groups should be from the same population



Case-Control Study - Self-Selection

- Not well suited for an internet-based approach
- Difficult to assess whether the determinants for self-selection are related to the exposure of interest
- Difficult to obtain cases and controls from the same source population

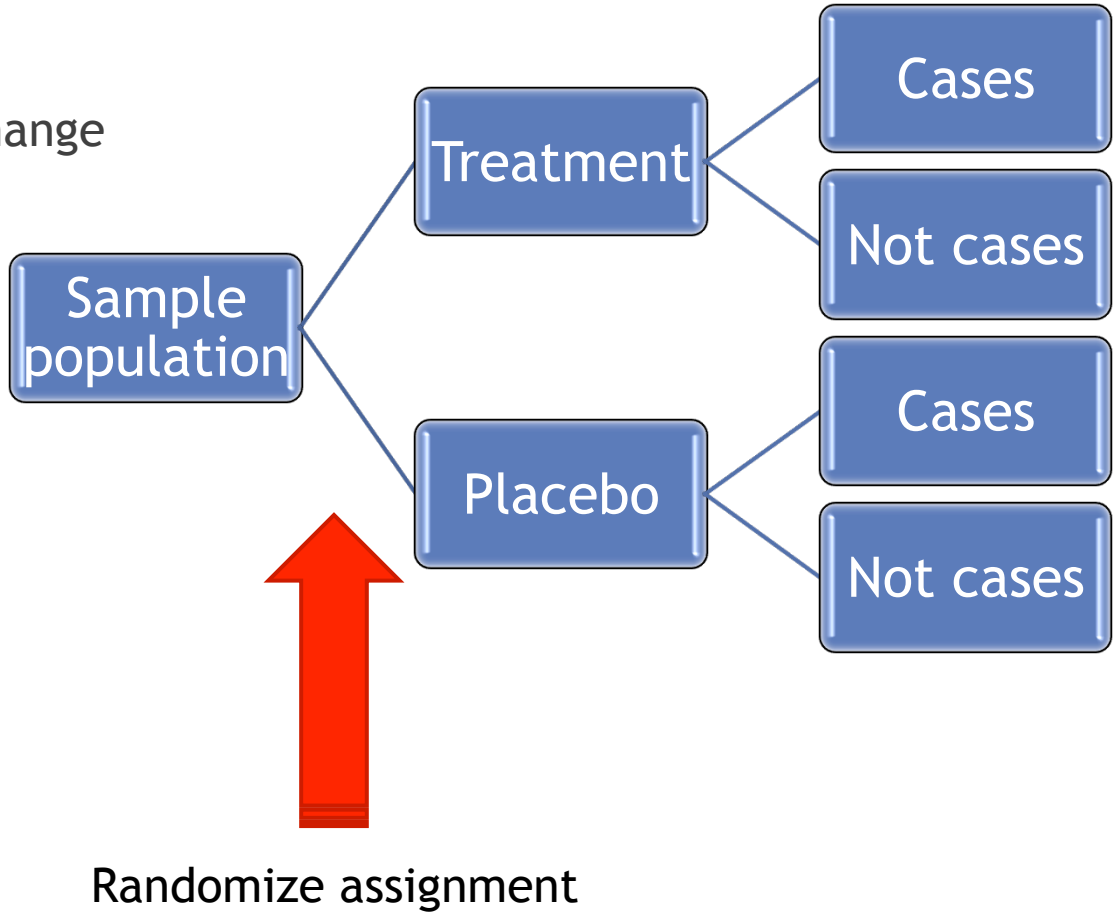
Case-Control Study – Digital Trail

- Use the available data to identify the group of interest and afterwards identify a control group

- Example:
  - Lampos (2014) used Twitter and Bing data to evaluate effectiveness of a vaccination campaign made by Public Health England

Intervention Study - Definition

- Experimental study
- Participants are divided into two groups
  - Treatment – exposed to medicine or behavioral change
  - Placebo – no exposure or inactive placebo



Intervention Studies - Self-Selection

- Internet recruitment fits well with intervention studies
- A review of 20 internet-based smoking cessation interventions shows low long-term benefits (Civljak et al. 2010)
- High dropout

Intervention Study – Digital trail

- Intervention types are limited
- Ethical concerns
- Example:
  - Kramer (2013) used modified Facebook “News Feed” to provide evidence for emotional contagion through social media

Learning from Internet data

Two lines of research

Category A

- many manual operations
- fine grained data set creation, feature formation / selection
- harder for methods to generalize, hard to replicate
- provide a good insight on a specific problem

Category B

- fewer (or zero) manual operations
- more noisy features
- applied statistical methods may generalize to related concepts
- solve a class of problems but provide fewer opportunities for qualitative analysis
- still hard to replicate (data availability is ambiguous)

Flow of the presentation

Aims and motivation

- What is the aim of this work?
- Why is it useful?

Data

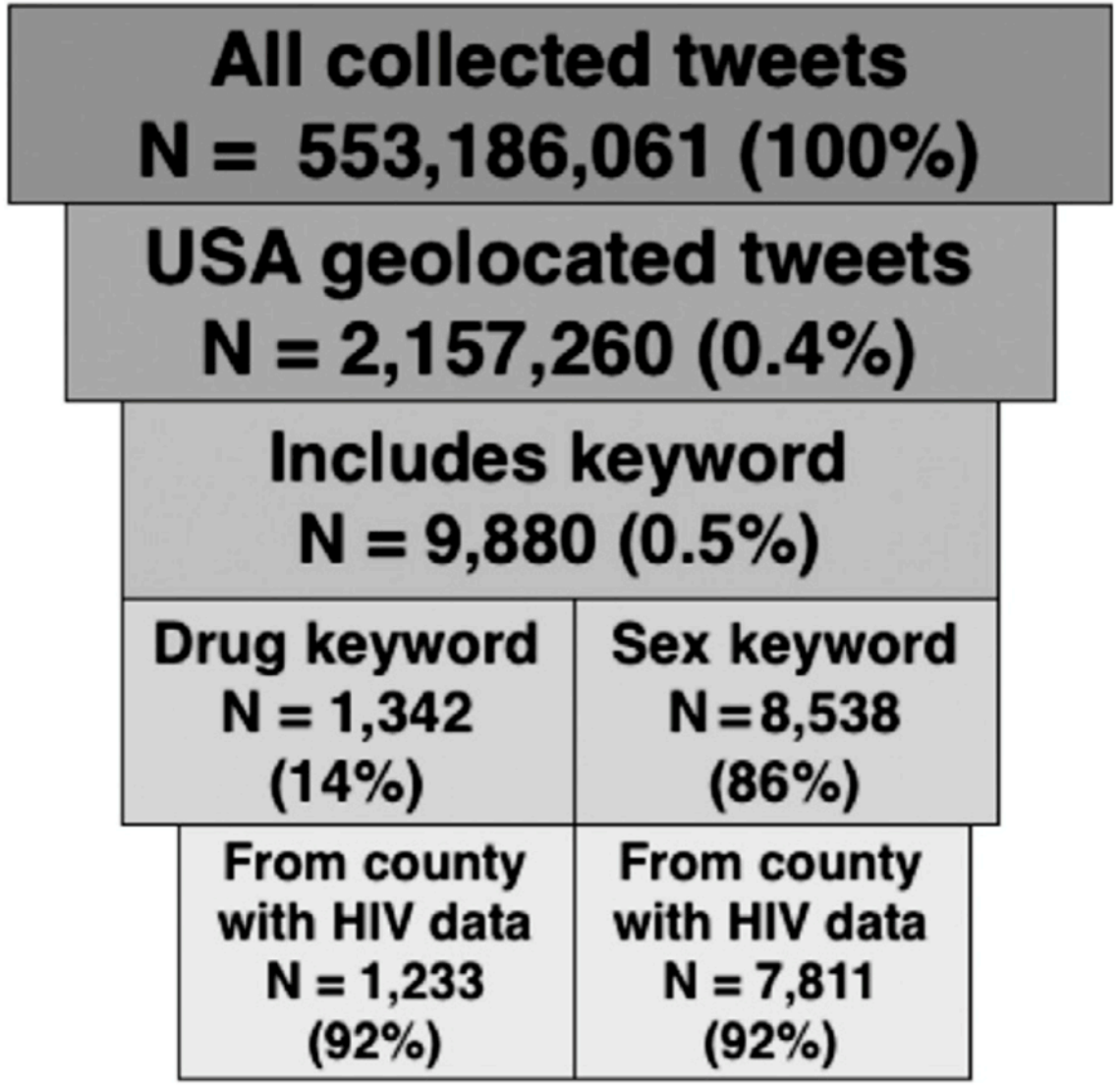
- What data have been used in this task?
- Were there any interesting data extraction techniques?

Methods and Results

- What are the main methodological points
- Present a subset of the results

HIV detection from Twitter

- as **simple** approach as possible
- Data: 550 million tweets (1% sample) from May to December 2012
- Filtered out non geolocated content, kept US content only (2.1 million tweets), geolocation at the county level
- manual list of risk related words suggestive of **sex** and **substance use**
- stemming applied
- county level US 'ground truth' from http://aidsvu.org (HIV/AIDS cases)
- incl. socio-economic status + GINI index (wealth inequality measure)



Young et al., 2014

HIV detection from Twitter

- **univariate regression analysis** using proportion of sex and drug risk-related tweets: significant positive relationship with HIV prevalence
- **multivariate regression analysis** of factors associated with county HIV prevalence (see Table below)

	Coefficient	Standard error	p-value
Proportion of HIV-related tweets (sex and drugs)	265	12.4	<.0001
% living in poverty	2.1	0.4	<.0001
GINI index	4.6	0.6	<.0001
% without health insurance	1.3	0.4	<.01
% with a high school education	-1.1	-3.1	<.01

Young et al., 2014

Predicting Depression from Twitter

- **Mental illness** leading cause of **disability** worldwide
- 300 million people suffer from depression (WHO, 2001)
- Services for identifying and treating mental illnesses: **NOT adequate**
- Can content from social media (Twitter) assist?

- Focus on **Major Depressive Disorder** (MDD)
  - low mood
  - low self-esteem
  - loss of interest or pleasure in normally enjoyable activities

De Choudhury et al., 2013

Predicting Depression from Twitter

Data set formation

- **crowdsourcing** a depression survey, share Twitter username
- determine a **depression score** via a formalized questionnaire (Center for Epidemiologic Studies Depression Scale; **CES-D**):
  - from 0 (no symptoms) to 60
- **476 people**
  - diagnosed with depression with onset between September 2011 and June 2012
  - agreed to monitor their public Twitter profile
  - 36% with CES-D > 22 (definite depression)
- **Twitter** feed collection ~ 2.1 million tweets
  - depression-positive users (from onset and one year back)
  - depression-negative users (from survey date and one year back)

De Choudhury et al., 2013

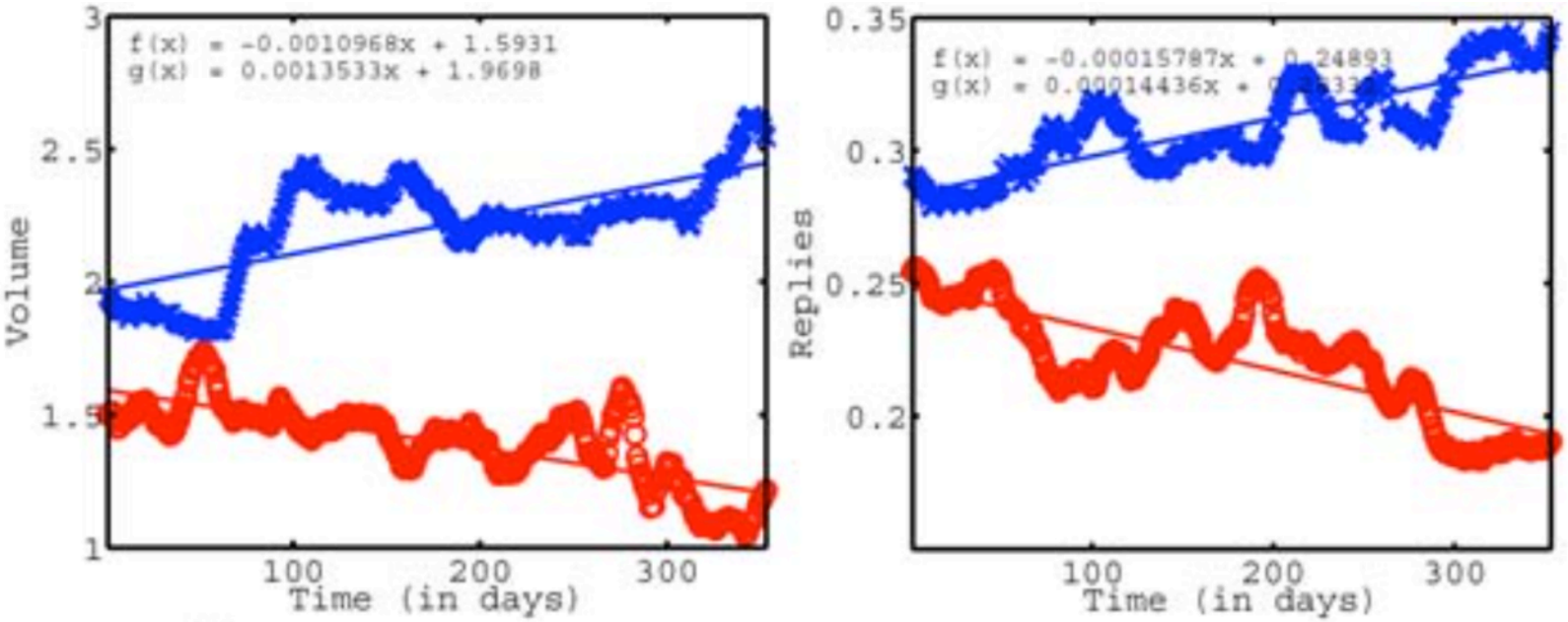
Predicting Depression from Twitter

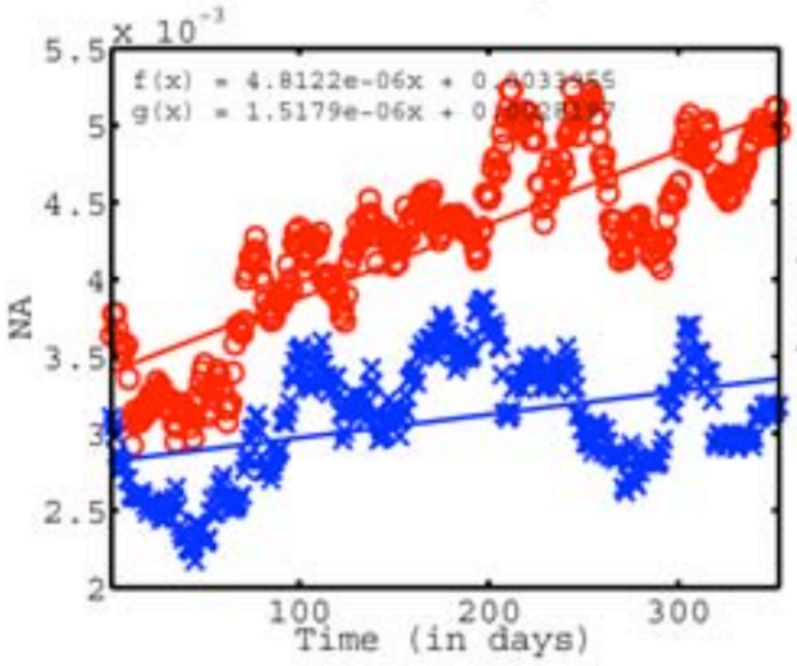
Examples of feature categories (overall 47)

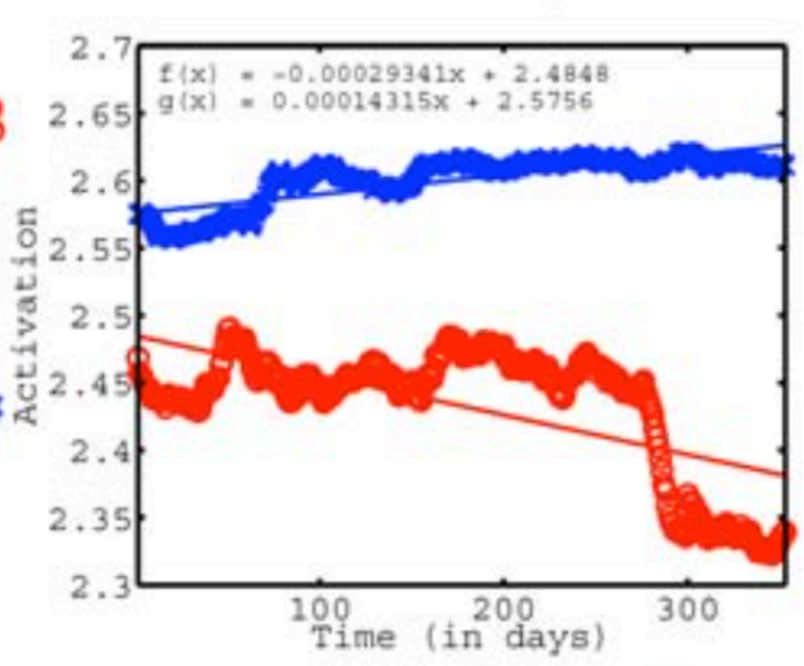
- **Engagement** ~ daily volume of tweets, proportion of @reply posts, retweets, links, question-centric posts, normalized difference between night and day posts (insomnia index)
- **Social network properties (ego-centric)** ~ followers, followees, reciprocity (average number of replies of U to V divided by number of replies from V to U), graph density (edges / nodes in a user's ego-centric graph)
- Linguistic Inquiry and Word Count (**LIWC** – http://www.liwc.net)
  - features for emotion: positive/negative affect, activation, dominance
  - features for linguistic style: functional words, negation, adverbs, certainty
- **Depression lexicon**
  - Mental health in Yahoo! Answers
  - Pointwise-Mutual-Information + Likelihood-ratio between 'depress*' and all other tokens (top 1%)
  - TF-IDF of these terms in Wikipedia to remove very frequent terms:1,000 depression words
- **Anti-depression language**: lexicon of antidepressant drug names

De Choudhury et al., 2013

Predicting Depression from Twitter







Depressive user patterns:

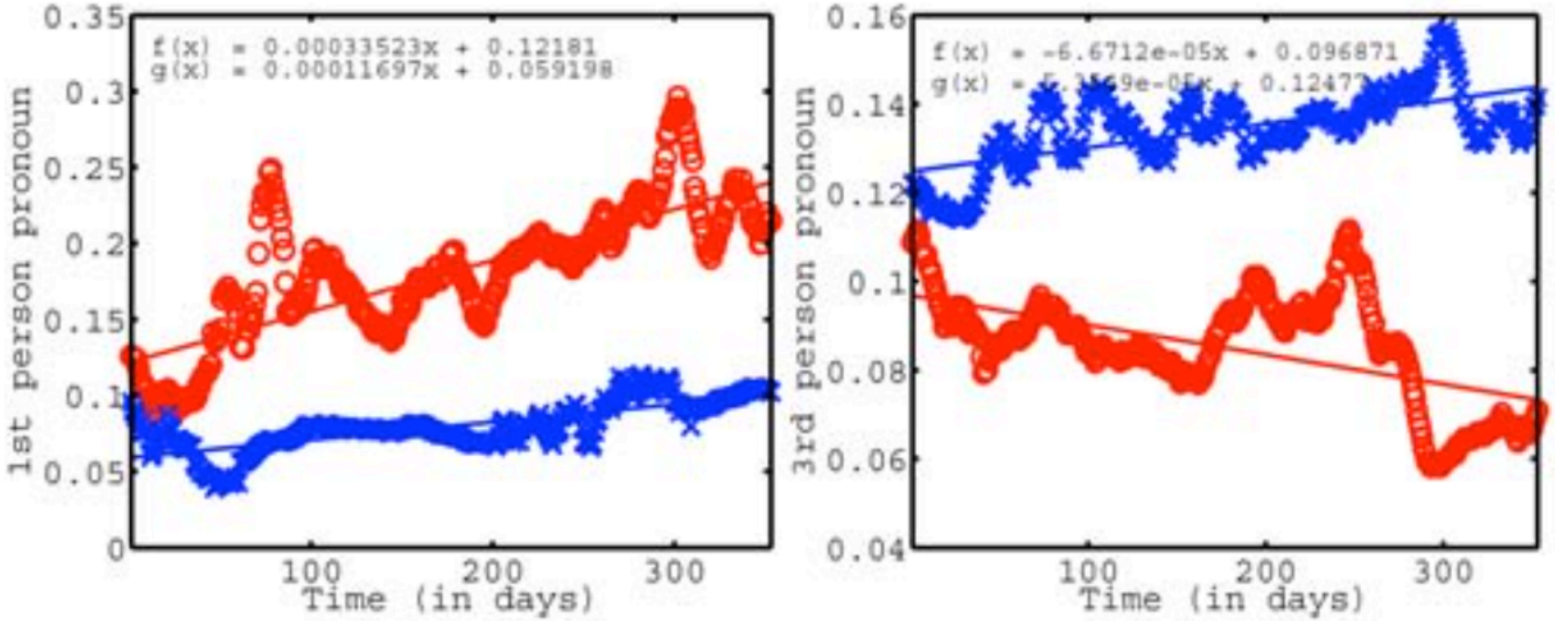
- decrease in user engagement (volume and replies)
- higher Negative Affect (NA)
- low activation (loneliness, exhaustion, lack of energy, sleep deprivation)

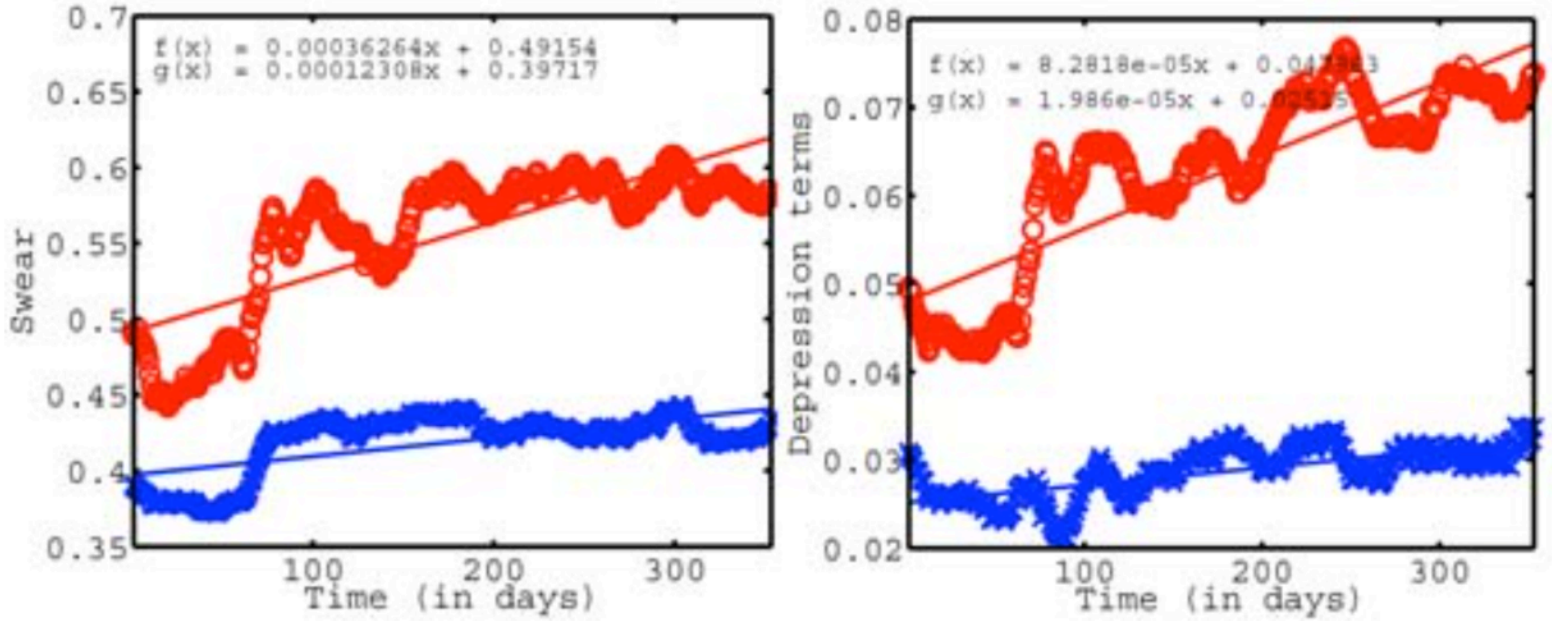
RED: depression class

BLUE: non-depression class

De Choudhury et al., 2013

Predicting Depression from Twitter





Depressive user patterns:

- increased presence of 1st person pronouns
- decreased for 3rd person pronouns
- use of depression terms higher (examples: anxiety, withdrawal, fun, play, helped, medication, side-effects, home, woman)

RED: depression class

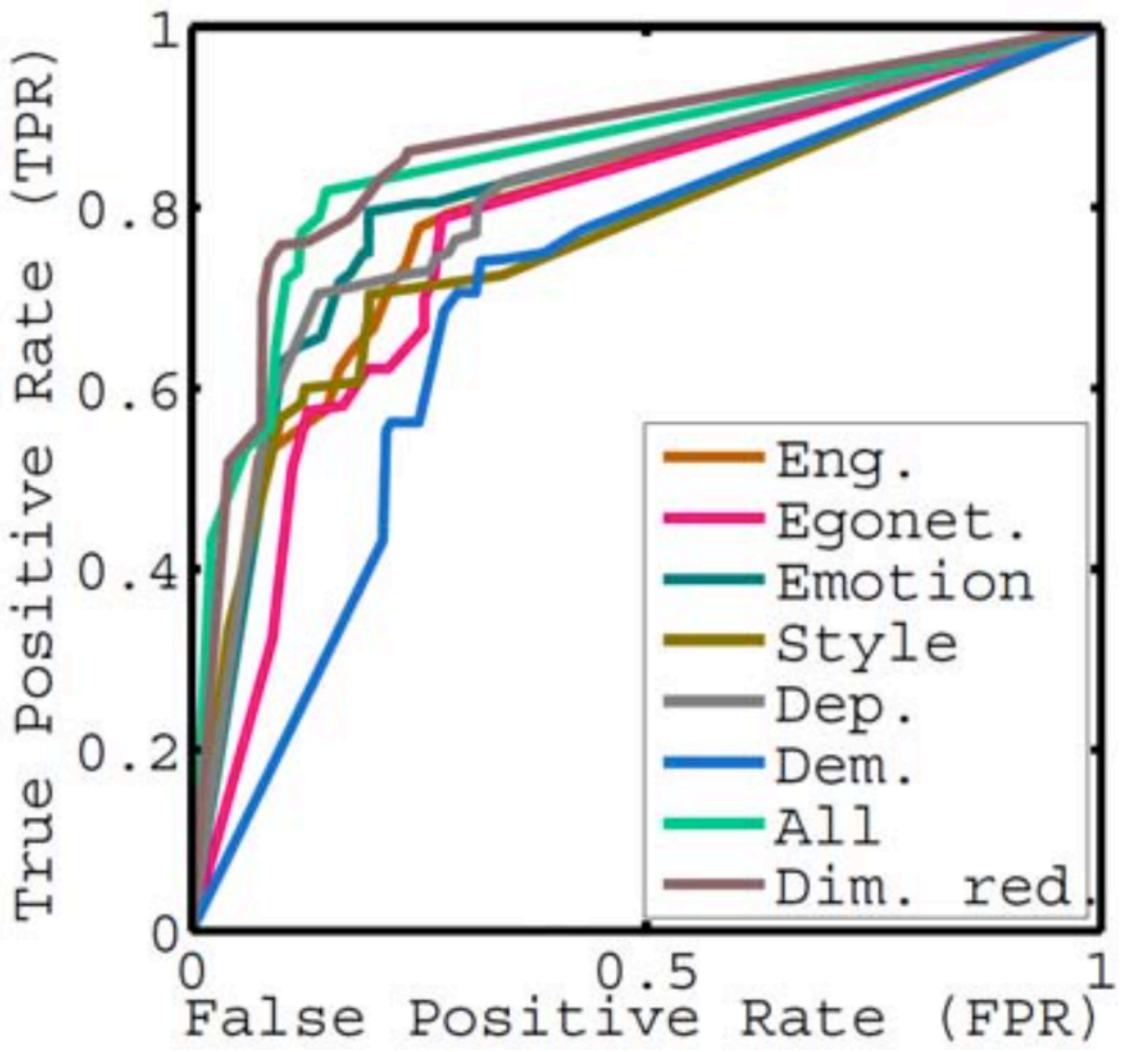
BLUE: non-depression class

De Choudhury et al., 2013

Predicting Depression from Twitter

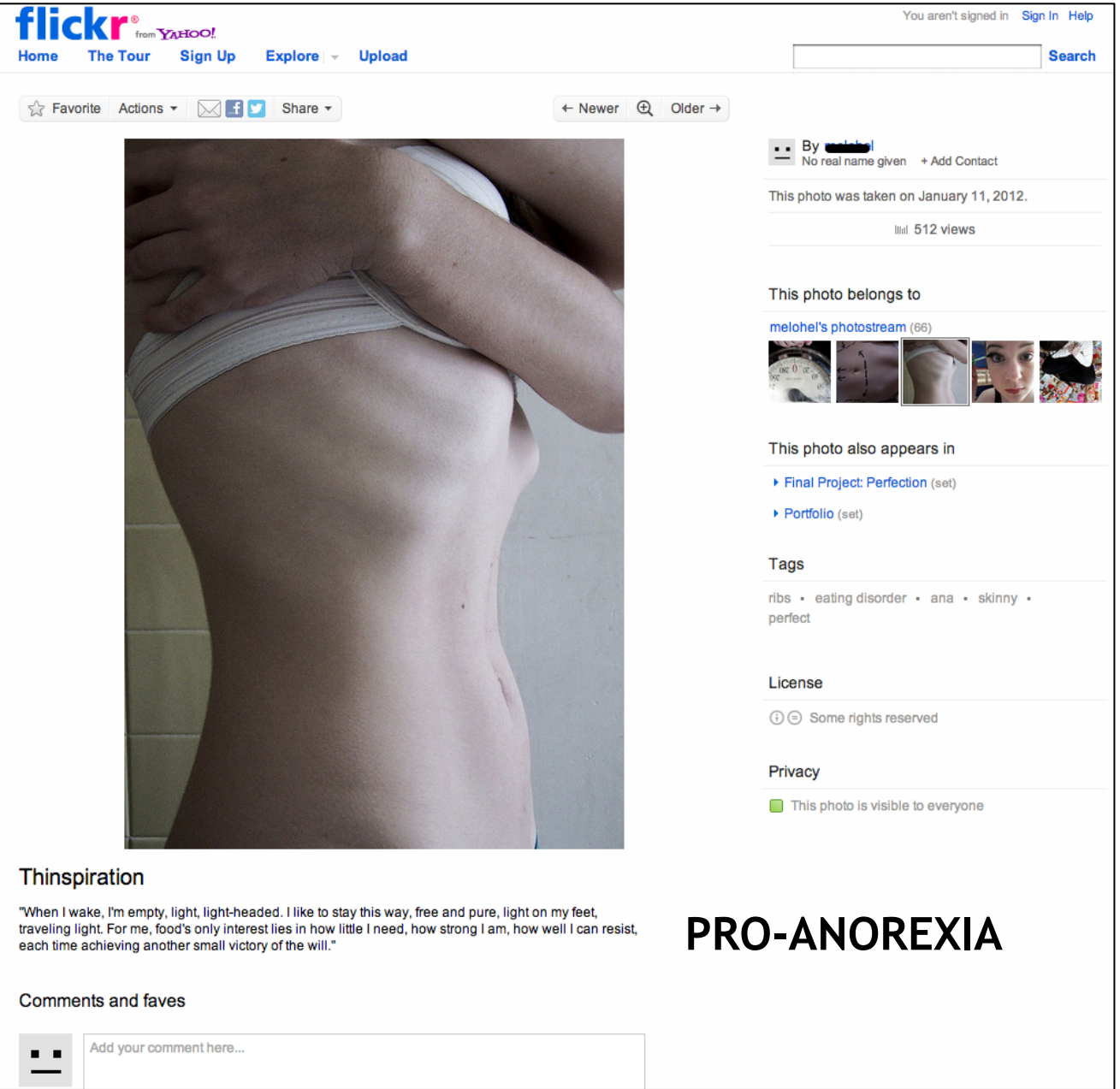
- 188 features (47 features X mean frequency, variance, mean momentum, entropy)
- Support Vector Machine with an RBF kernel
- Principal Component Analysis (PCA)

	accuracy (positive)	accuracy (mean)
BASELINE	NA	64%
engagement	53.2%	55.3%
ego-network	58.4%	61.2%
emotion	61.2%	64.3%
linguistic style	65.1%	68.4%
depressive language	66.3%	69.2%
all features	68.2%	71.2%
all features (PCA)	70.4%	72.4%



De Choudhury et al., 2013

Pro-anorexia and pro-recovery content on Flickr



PRO-ANOREXIA

PRO-RECOVERY

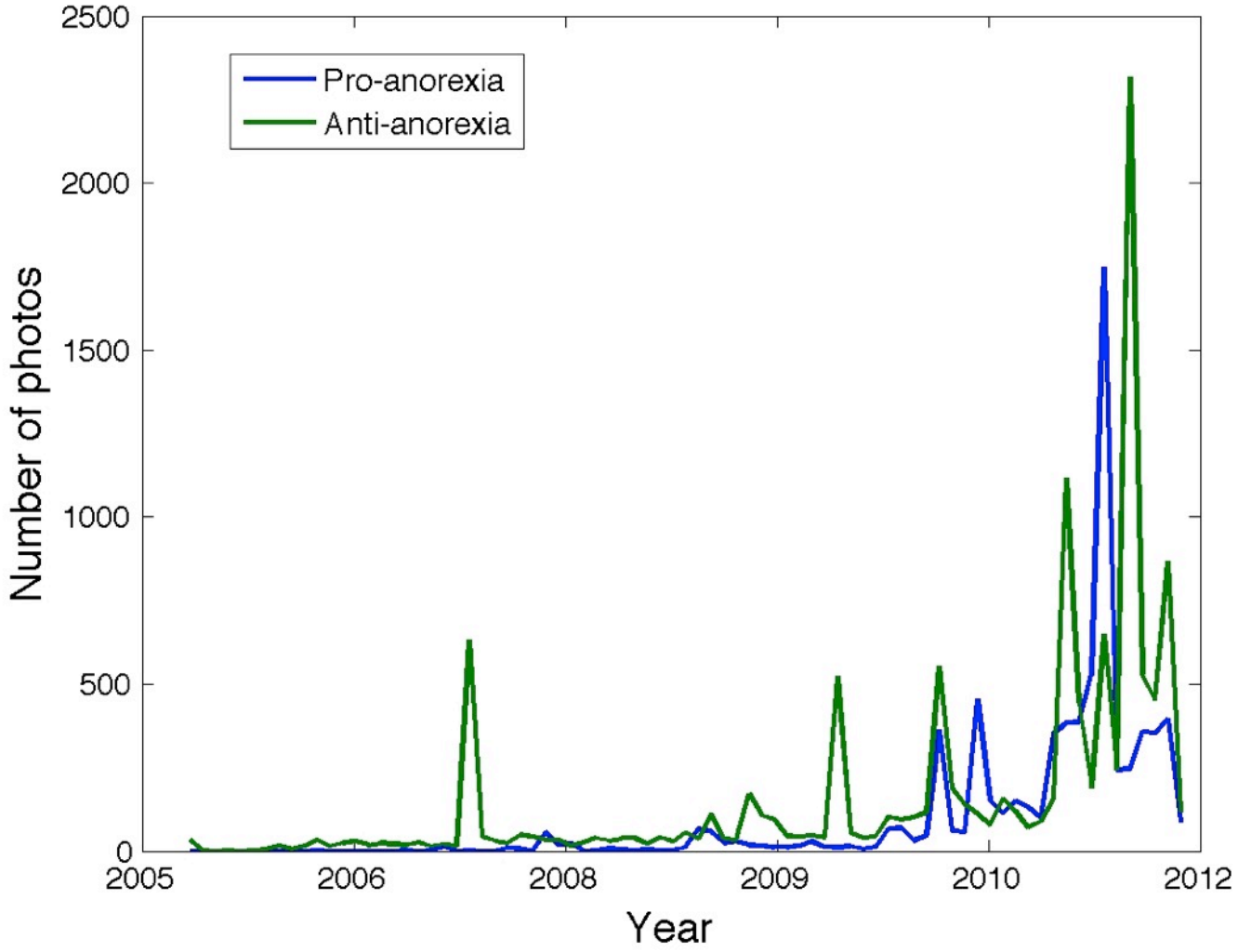
Yom-Tov et al., 2012

Pro-anorexia and pro-recovery content on Flickr

- Study the relationship between **pro-anorexia (PA)** and **pro-recovery (PR)** communities on Flickr – can the PR community affect PA?
- **Data**: Pro-anorexia and pro-recovery **photos**
  - contacts, favorites, comments, tags
  - multi-layered data set creation with many manual steps
- **Filtered by**
  - anorexia keywords (‘thinspo’, ‘pro-ana’, ‘thinspiration’) in photo tags
  - who commented
  - who favorited or groups (such as ‘Anorexia Help’)
- 543K photos, 2.2 million comments for 107K photos by 739 users
- **172 PR**, **319 PA users** (labeled by 5 human judges)

Yom-Tov et al., 2012

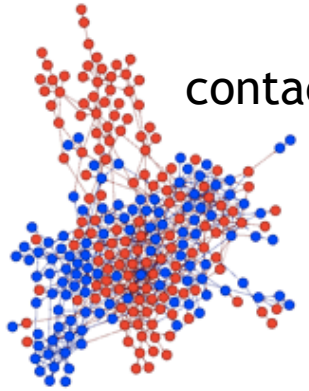
Pro-anorexia and pro-recovery content on Flickr

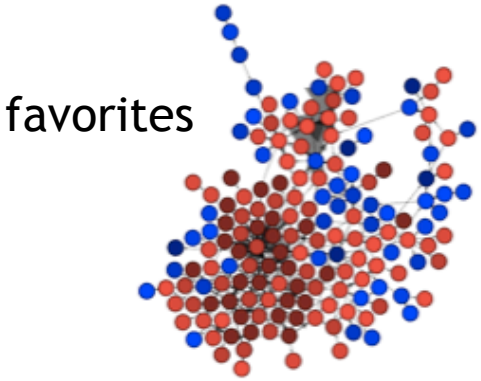


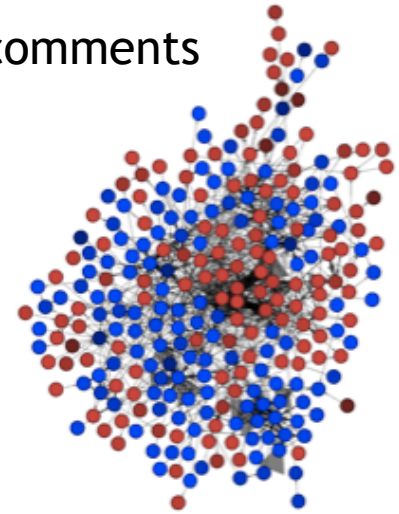
- number of photos time series from these classes correlate (Spearman correlation ρ = .82)
- pro-anorexia most frequent tags: ‘thinspiration’, ‘doll’, ‘thinspo’, ‘skinny’, ‘thin’
- pro-recovery: ‘home’, ‘sign’, ‘selfportrait’, ‘glass’, ‘cars’ (*no underlying theme*)

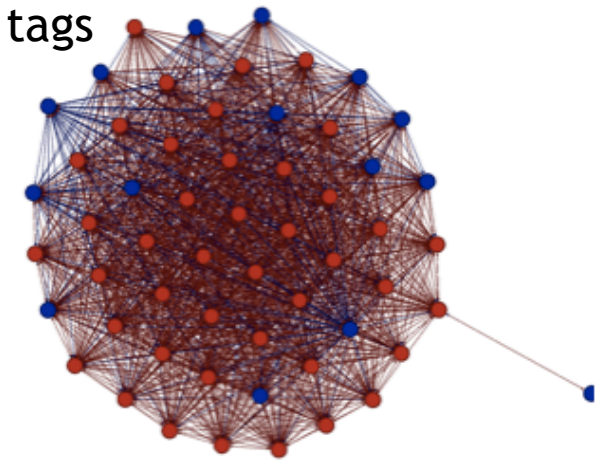
Yom-Tov et al., 2012

Pro-anorexia and pro-recovery content on Flickr









red: pro-anorexia

blue: pro-recovery

- how users are connected based on contacts, favorites, comments, tags
- main connected component shown
- classes intermingled especially when observing tags
- best separated through contacts

Yom-Tov et al., 2012

Pro-anorexia and pro-recovery content on Flickr

Did pro-recovery interventions help? ***Not really.***

(PA = Pro-Anorexia, PR = Pro-Recovery)

	Commented by			
	Cessation rate		**Avg days to cessation**	
	PA	**PR**	**PA**	**PR**
PA	61%	46%	225	329
PR	61%	71%	366	533

Yom-Tov et al., 2012

Postmarket drug safety surveillance via search queries

Why?

- Current postmarket drug surveillance mechanisms depend on patient reports
- Hard to identify if an adverse reaction happens after the drug is taken for a long period
- Hard to identify if several medications are taken at the same time

Therefore,

- Could we complement this process by looking at search queries?

Yom-Tov and Gabrilovich, 2013

Postmarket drug safety surveillance via search queries

Data

- queries submitted to Yahoo search engine during 6 months in 2010
- 176 unique million users (search logs anonymized)

Drugs under investigation

- 20 top-selling drugs (in the US)

Symptoms lexicon

- 195 symptoms from the international statistical classification of diseases and related health problems (WHO)
- filtered by Wikipedia (http://en.wikipedia.org/wiki/List_of_medical_symptoms)
- expanded with synonyms acquired through an analysis of the most frequently returned web pages when a symptom was forming the query

Aim

- quantify the prevalence of adverse drug reports (ADR) for a given drug

Yom-Tov and Gabrilovich, 2013

Postmarket drug safety surveillance via search queries

- 'ground truth': reports to repositories for safety surveillance for approved drugs mapped to same list of symptoms
- **score of drug-symptom** pair

When user queried for drug	User queried for the drug?	
	NO	YES
Before Day 0	n_{11}	n_{12}
After Day 0	n_{21}	n_{22}

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_{i1} - n_{i2})^2}{n_{i2}}$$

n_{ij}: how many times a symptom was searched

Day 0: first day user searched for a drug D

- if the user has not searched for a drug, then day 0 is the midpoint of his history

Yom-Tov and Gabrilovich, 2013

Postmarket drug safety surveillance via search queries

- Comparison of drug-symptom scores based on query logs and 'ground truth'
- Which symptoms reduce this correlation the most? (*most discordant ADRs*)
  - discover previously unknown ADRs that patients do not tend to report

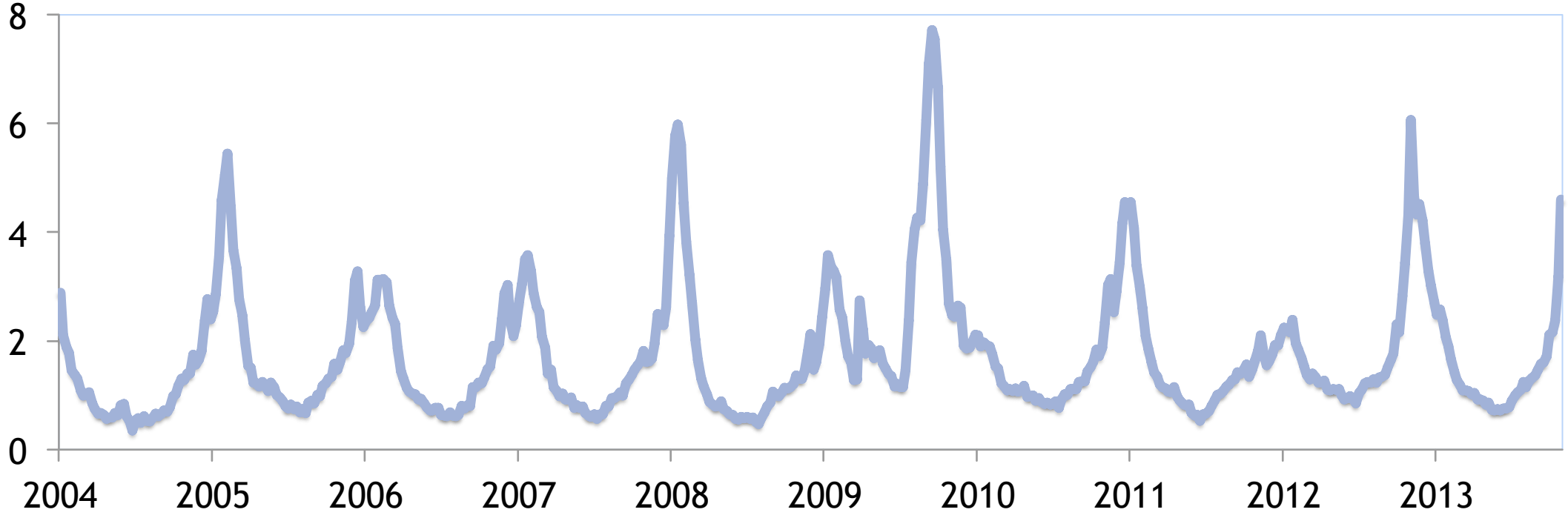
Drug	ρ	p-value	most discordant ADRs
Zyprexa	.61	.002	constipation, diarrhea, nausea, paresthesia, somnolence
Effexor	.54	<.001	nausea, phobia, sleepy, weight gain
Lipitor	.54	<.001	asthenia, constipation, diarrhea, dizziness, nausea
Pantozol	.51	.006	chest pain, fever, headache, malaise, nausea
Pantoloc	.49	.001	chest pain, fever, headache, malaise, nausea

- **Class 1**
  ADRs recognized by patients and medical professionals (acuteness, fast onset)
- **Class 2**
  later onset, less acute

Yom-Tov and Gabrilovich, 2013

Modeling ILI from search queries (Google Flu Trends)

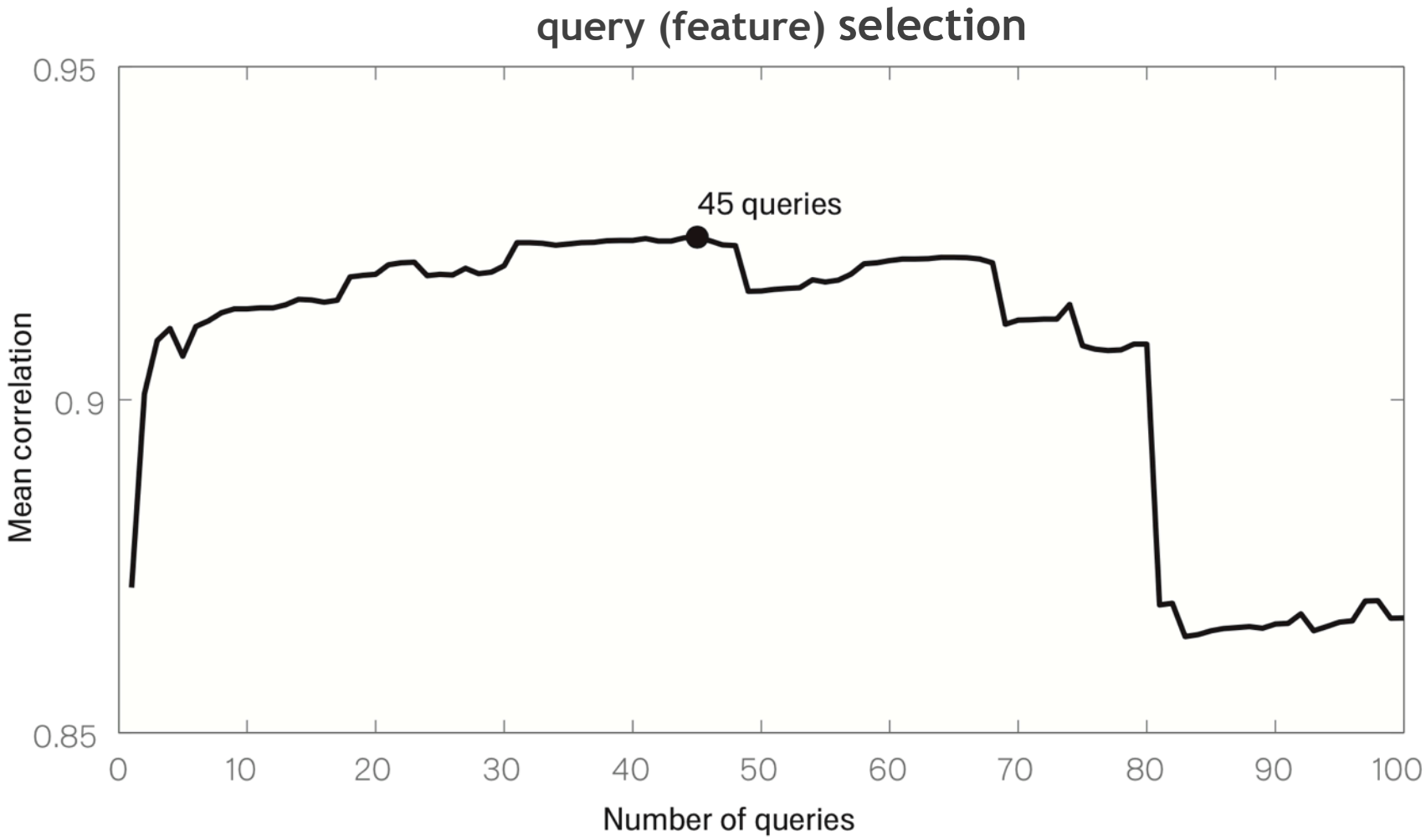
- **Motivation:** Early-warnings for the rate of an infectious disease
- **Output:** Predict influenza-like illness rates in the population (as published by health authorities such as CDC)



Ginsberg et al., 2009

Modeling ILI from search queries (Google Flu Trends)

- test the goodness of fit between the frequency of **50 million** candidate search queries and CDC data across 9 US regions
- get the **N** top-scoring queries
- decide optimal **N** using held-out data
- **N = 45 (!!)**

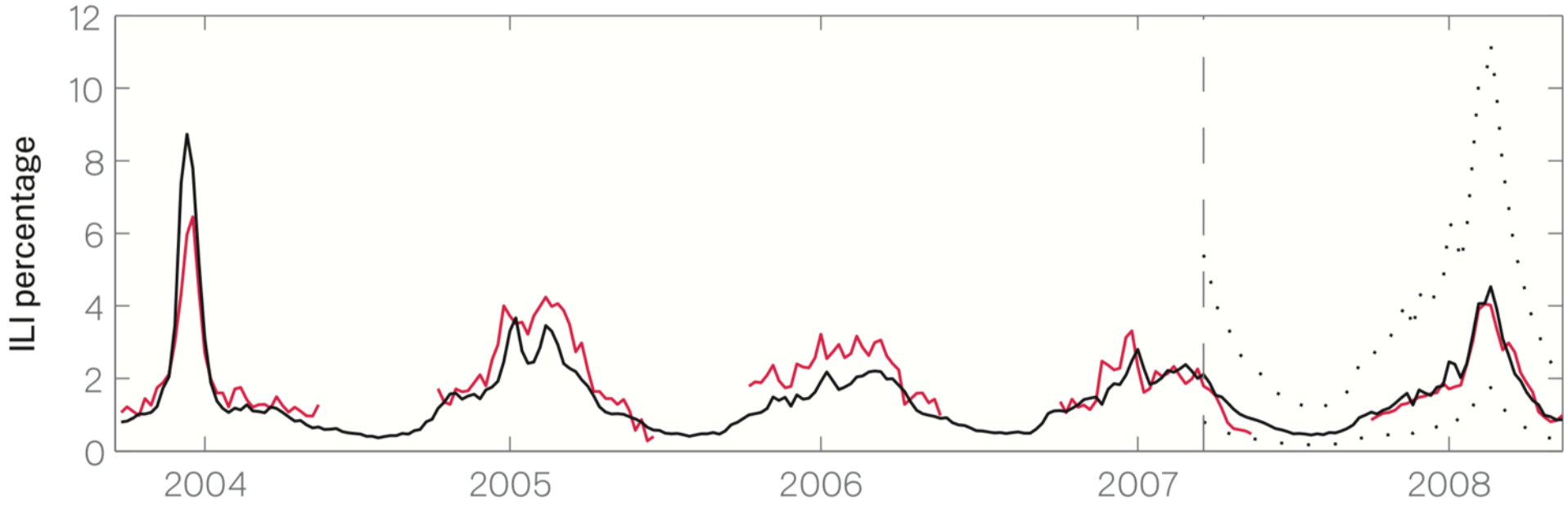


Ginsberg et al., 2009

Modeling ILI from search queries (Google Flu Trends)

- Google flu trends model
  - **q** is the aggregate query frequency among the selected queries and ILI rates (CDC) across US regions [just one variable!]
  - linear correlation was enhanced in the logit space

$$\text{logit}(\text{ILI}) = \alpha \times \text{logit}(\mathbf{q}) + \beta$$

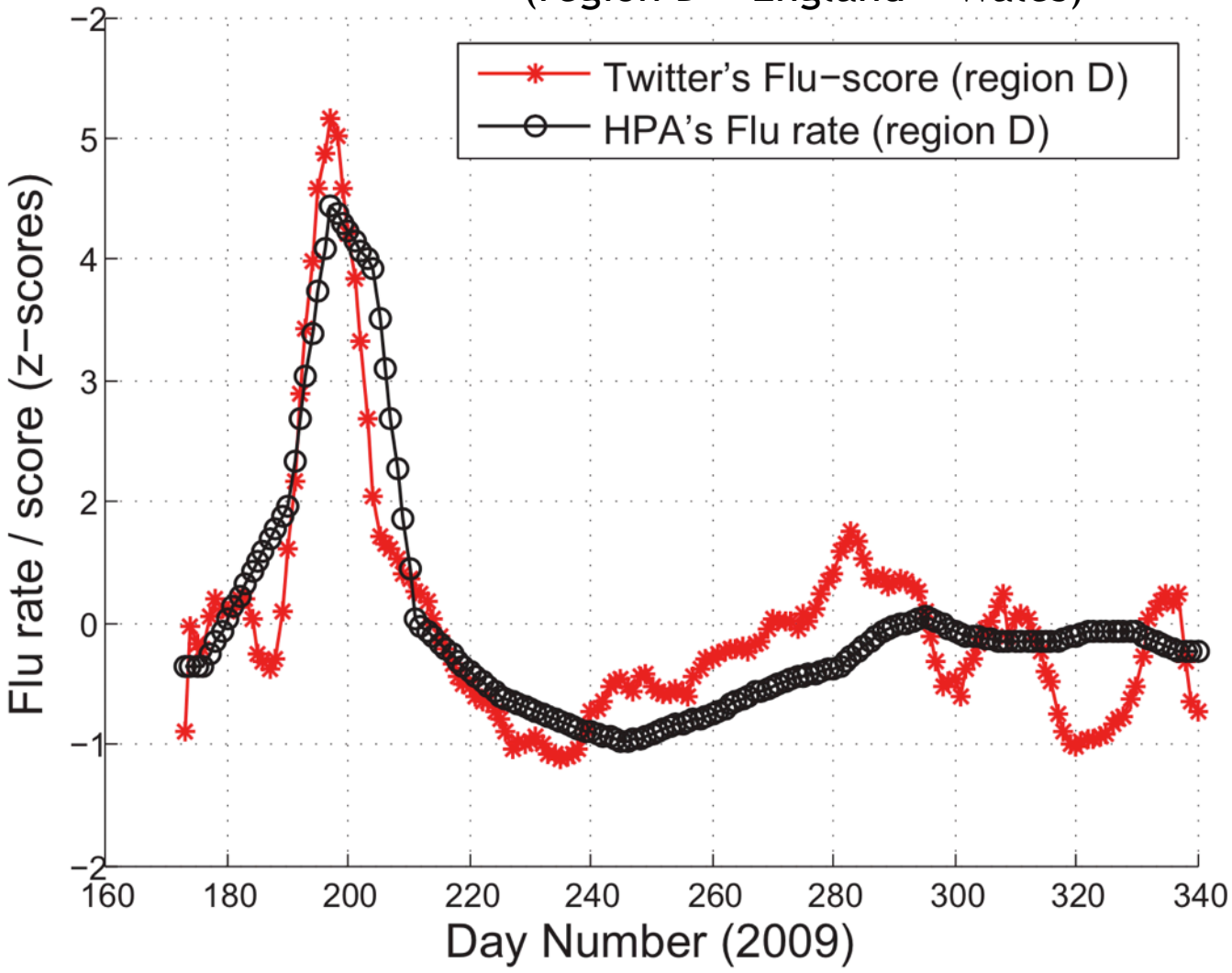


Ginsberg et al., 2009

Modeling ILI from Twitter (take 1)

- Is it possible to **replicate** the previous finding using a different user-generated source? (**Twitter**)
- 25 million tweets from June to December 2009
- Manually create a list of **41 flu related terms** ('fever', 'sore throat', 'headache', 'flu')
- Plot their frequencies against 'ground truth' from Health Protection Agency (HPA; official health authority in the UK)

(region D = England + Wales)



Lampos and Cristianini, 2010

Modeling ILI from Twitter (take 1)

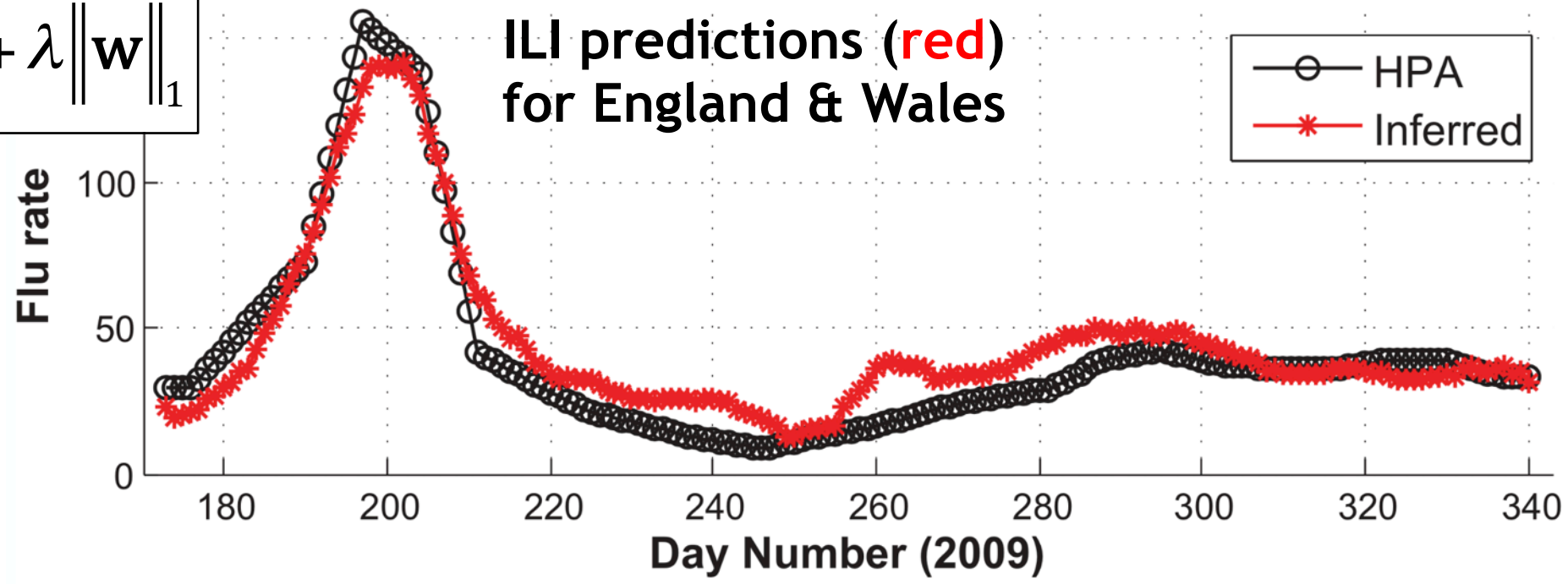
- Can we automate feature selection?
- Generate a pool of **1560 candidate stemmed flu markers** (1-grams) from related web pages (Wikipedia, NHS forums etc.)
- **Feature selection** and ILI **prediction**
  - **X** expresses normalized time series of the candidate flu markers
  - L1 norm regularization via the '**lasso**' (λ is the reg. parameter)
  - feature selection, tackles overfitting issues

$$\underset{\mathbf{w}}{\operatorname{argmin}} \left\| \mathbf{X}\mathbf{w} - \mathbf{y} \right\|_2^2 + \lambda \left\| \mathbf{w} \right\|_1$$

Lampos and Cristianini, 2010

Modeling ILI from Twitter (take 1)

$$\underset{\mathbf{w}}{\operatorname{argmin}} \left\| \mathbf{X}\mathbf{w} - \mathbf{y} \right\|_2^2 + \lambda \left\| \mathbf{w} \right\|_1$$

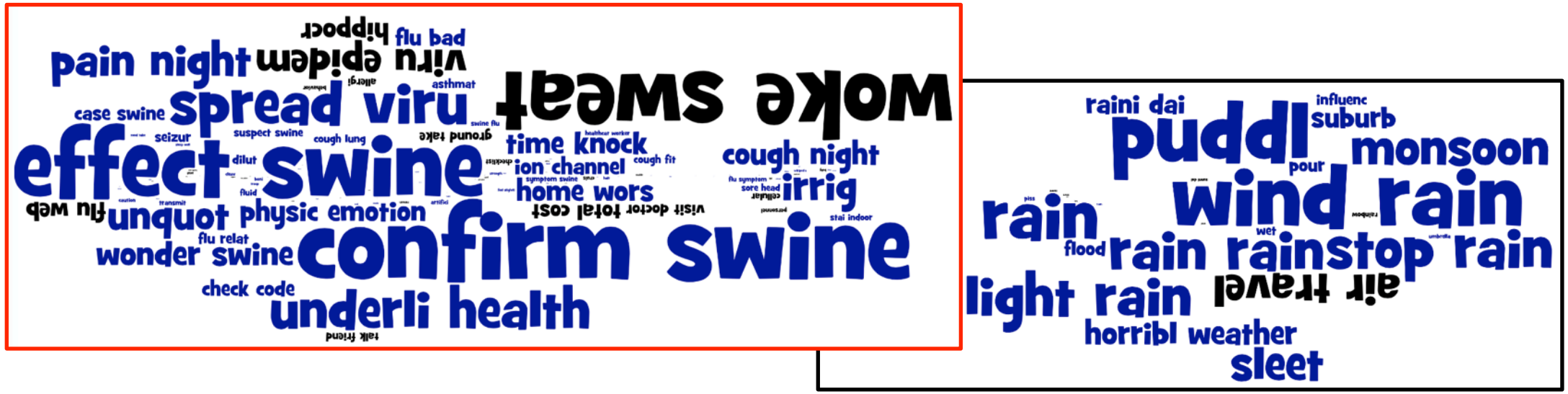


Examples of selected 1-grams: muscl, appetit, unwel, throat, nose, immun, phone, swine, sick, dai, symptom, cough, loss, home, runni, wors, diseas, diarrhoea, pregnant, headach, cancer, fever, tired, temperatur, feel, ach, flu, sore, vomit, ill, thermomet, pandem

Lampos and Cristianini, 2010

Modeling ILI from Twitter (take 2)

- 2048 1-grams and 1678 2-grams (by indexing web pages relevant to flu)
- more consistent feature selection (**bootstrap lasso**)
  - N (~= 40) bootstraps, create N sets of selected features
  - learn optimal consensus threshold (>= 50%)
  - hybrid combination of 1-gram and 2-gram based models

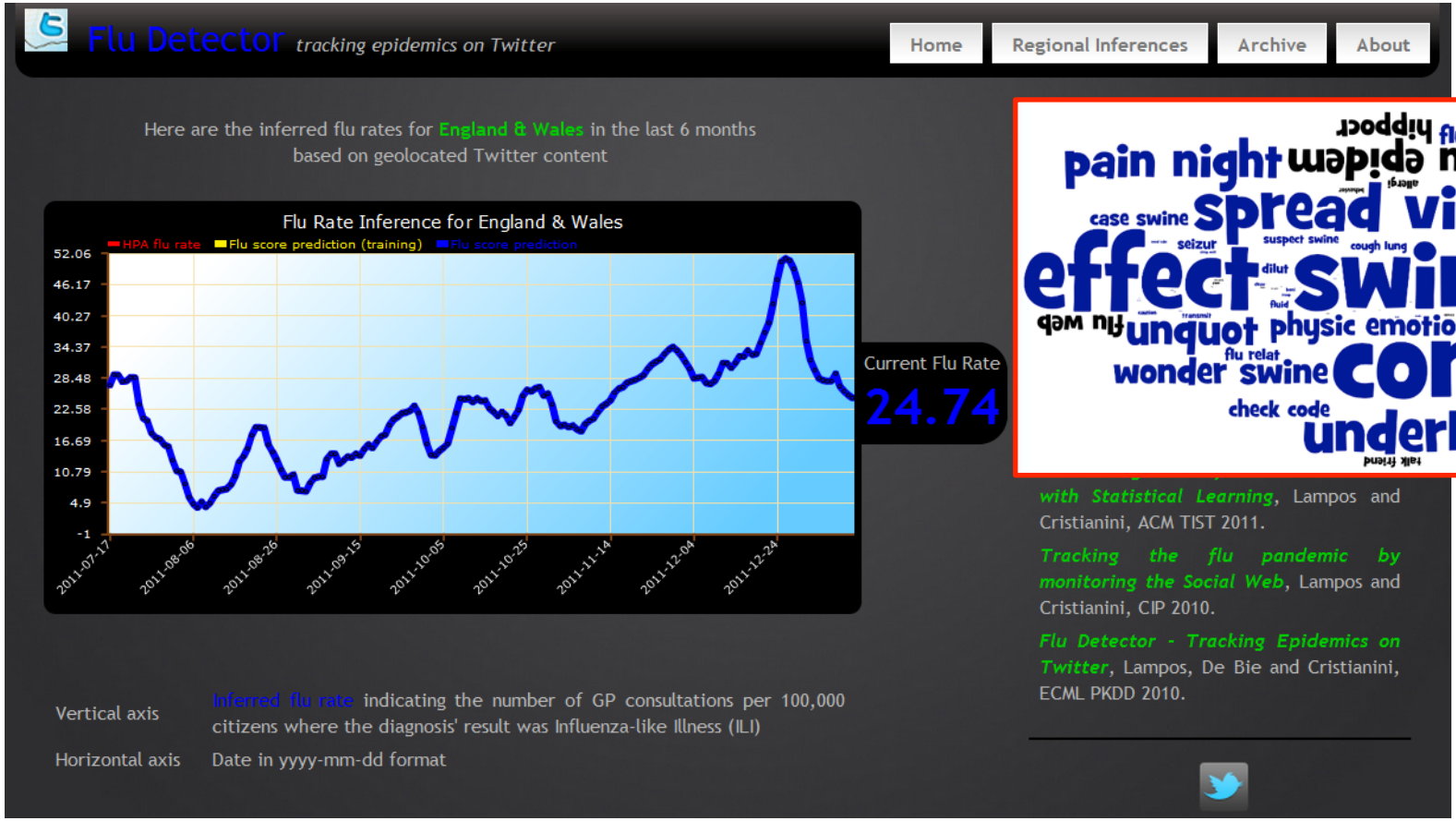


Data: June 2009 – April 2010 (50 million tweets)

Lampos and Cristianini, 2012

Modeling ILI from Twitter (take 2)

Flu Detector (the 1st web application for tracking ILI from Twitter)

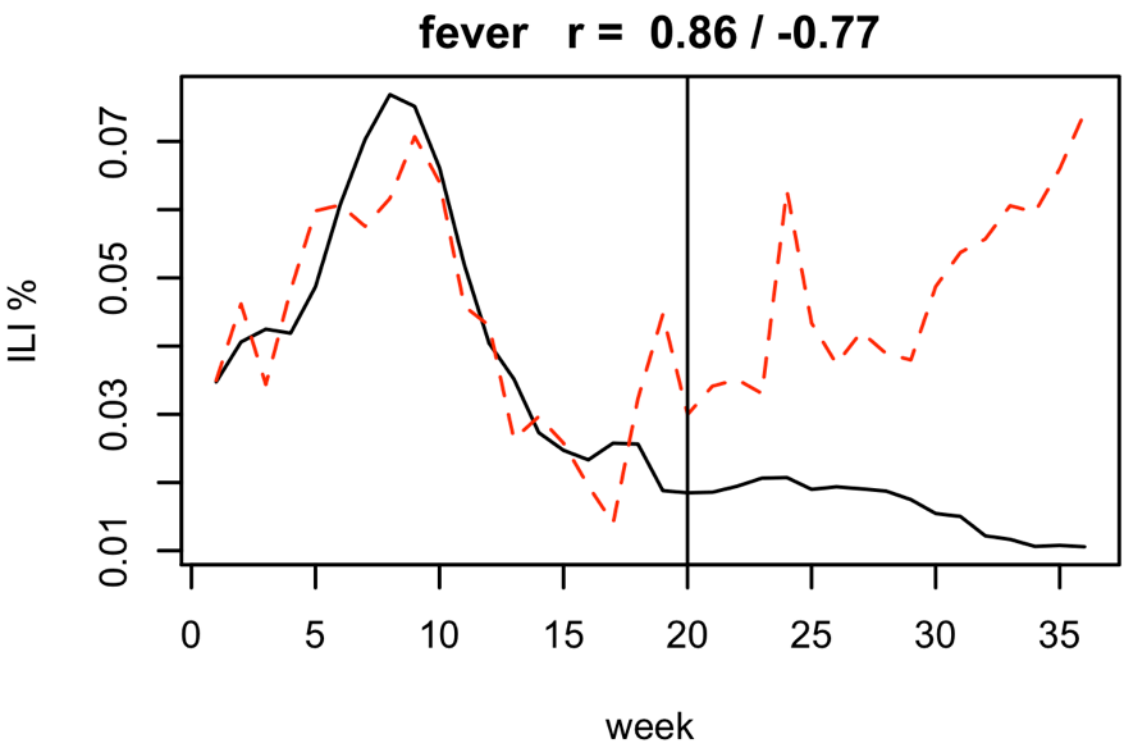


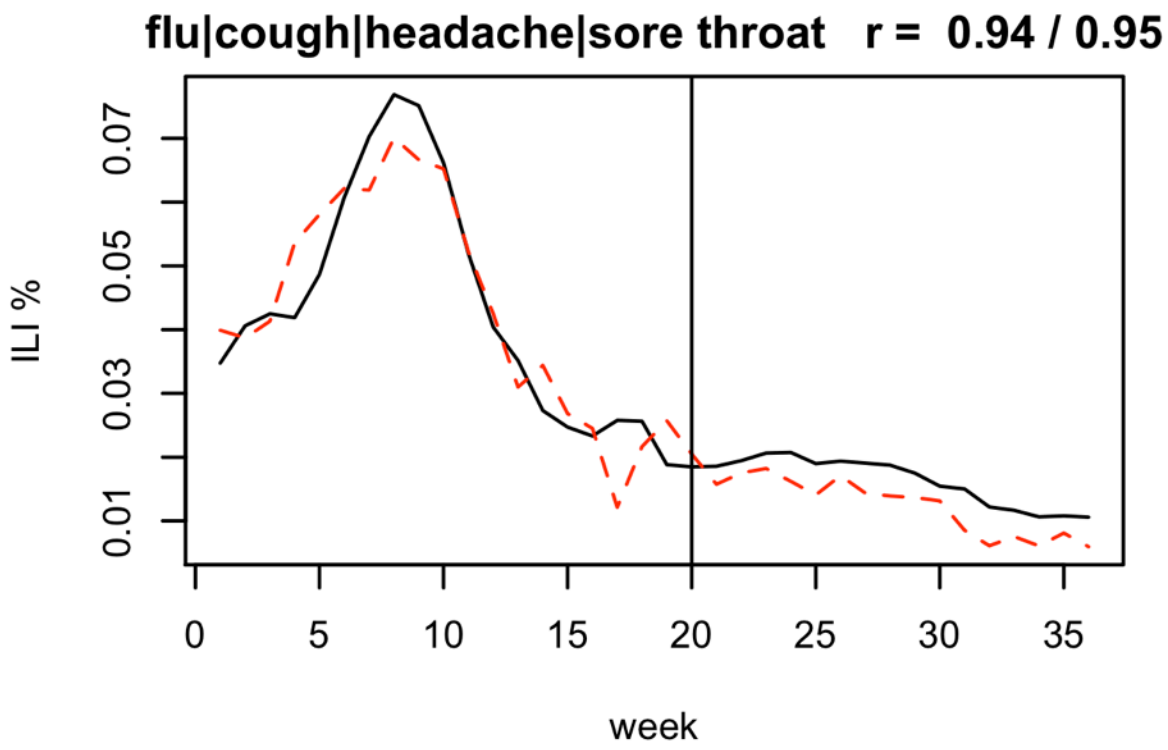
Lampos et al., 2010

Modeling ILI rates from Twitter (take 3)

- data: 570 million tweets, 8-month period
- **light-weight** approach: ‘flu’, ‘cough’, ‘headache’, ‘sore throat’ (term matching)
- aggregate frequency (**T**) of selected tweets into a GFT model

$$\text{logit}(\text{ILI}) = \alpha \times \text{logit}(\text{T}) + \beta$$

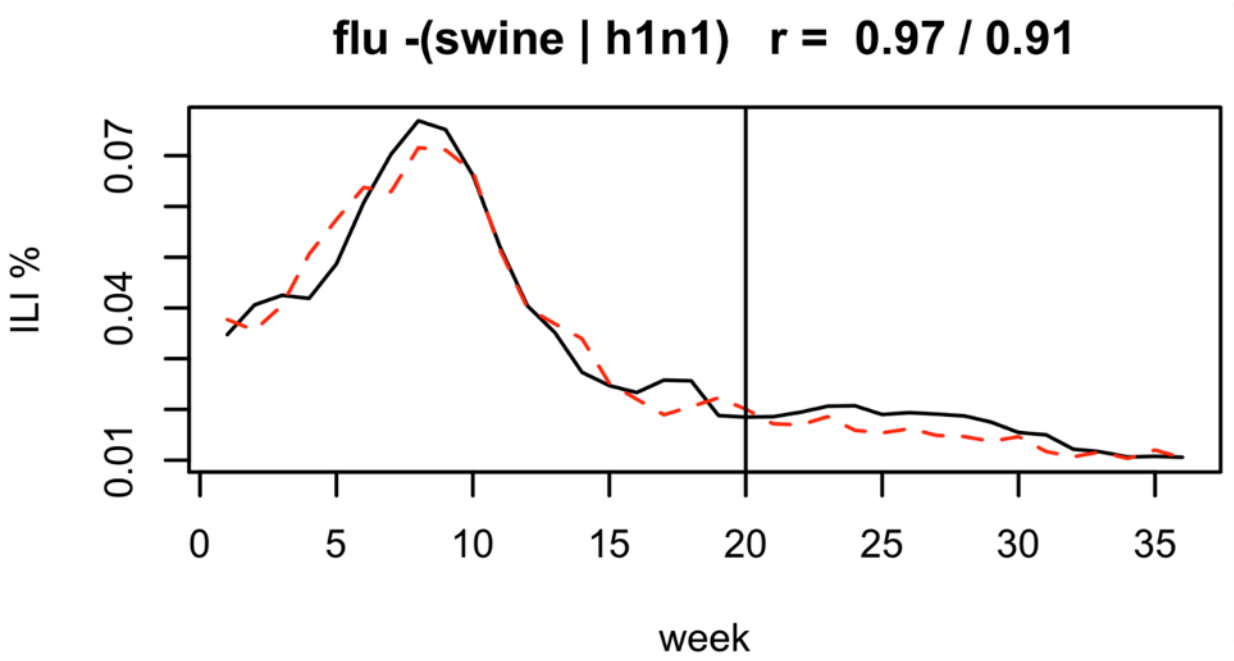


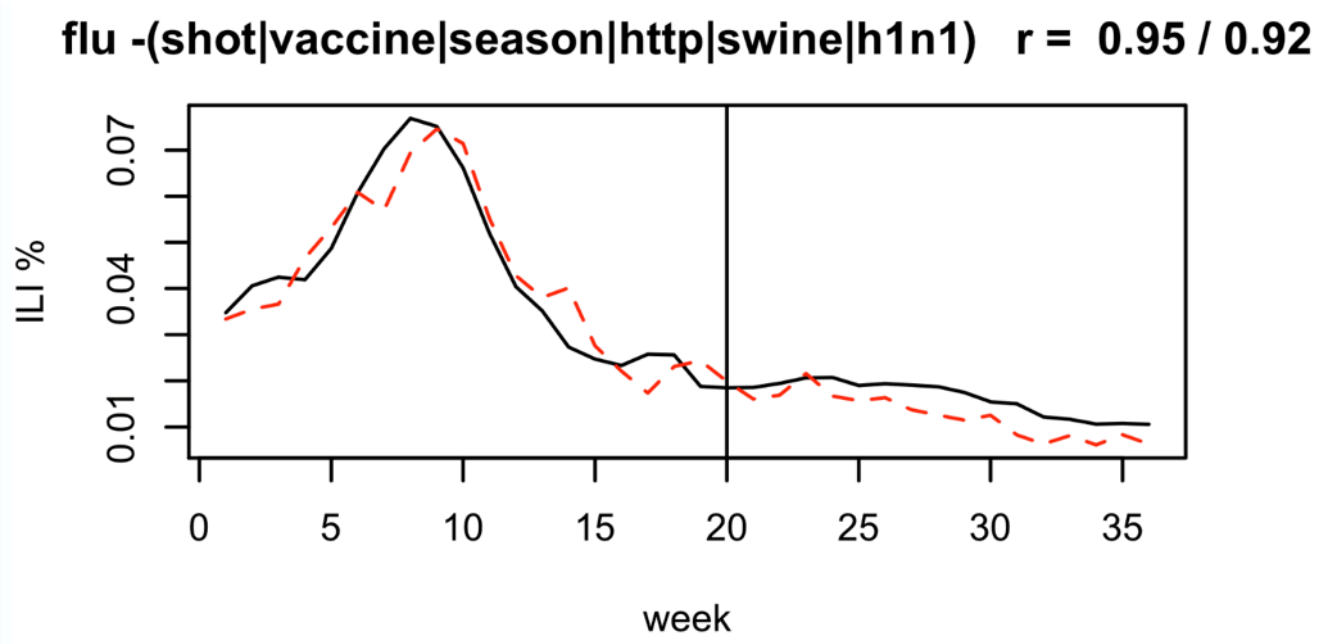


Culotta, 2013

Modeling ILI rates from Twitter (take 3)

- if ambiguous terms are removed (shot, vaccine, swine, h1n1 etc.)
- fit of training data may improve, prediction performance on held-out data may not

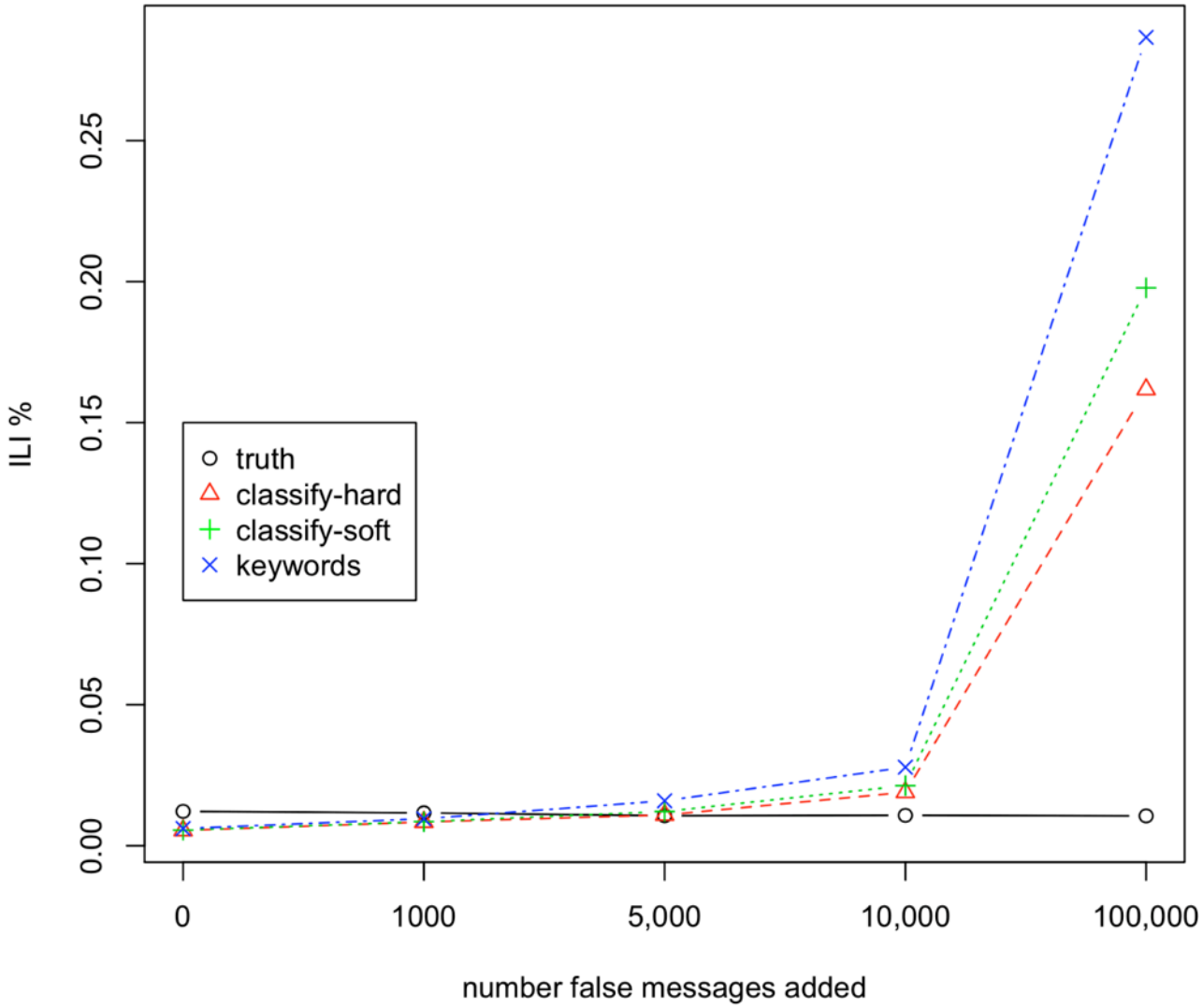




Culotta, 2013

Modeling ILI rates from Twitter (take 3)

- bag-of-words logistic regression **classifier** (related/unrelated to ILI tweets, 206 labeled samples)
- **84% accuracy**, easy-to-build
- did not improve, but also did not hurt performance
- simulation of ‘false’ indicators (injection of likely to be spurious tweets in the data) – **classification helps**
- SVM (RBF kernel) instead of did not improve performance (however, model too simplistic to give SVM a chance)



Culotta, 2013

Modeling ILI rates from Twitter (take 4)

- A **different approach**
  - NO supervised learning of ILI, but **intrinsic learning**
  - modeling based on natural language processing operations

- Why this may be **useful?**
  - syndromic surveillance is not the perfect ‘ground truth’
  - however, syndromic surveillance rates are used for evaluation!

- **Data**
  - 2 billion tweets from May 2009 to October 2010
  - 1.8 billion tweets from August 2011 to November 2011

Lamb et al., 2013

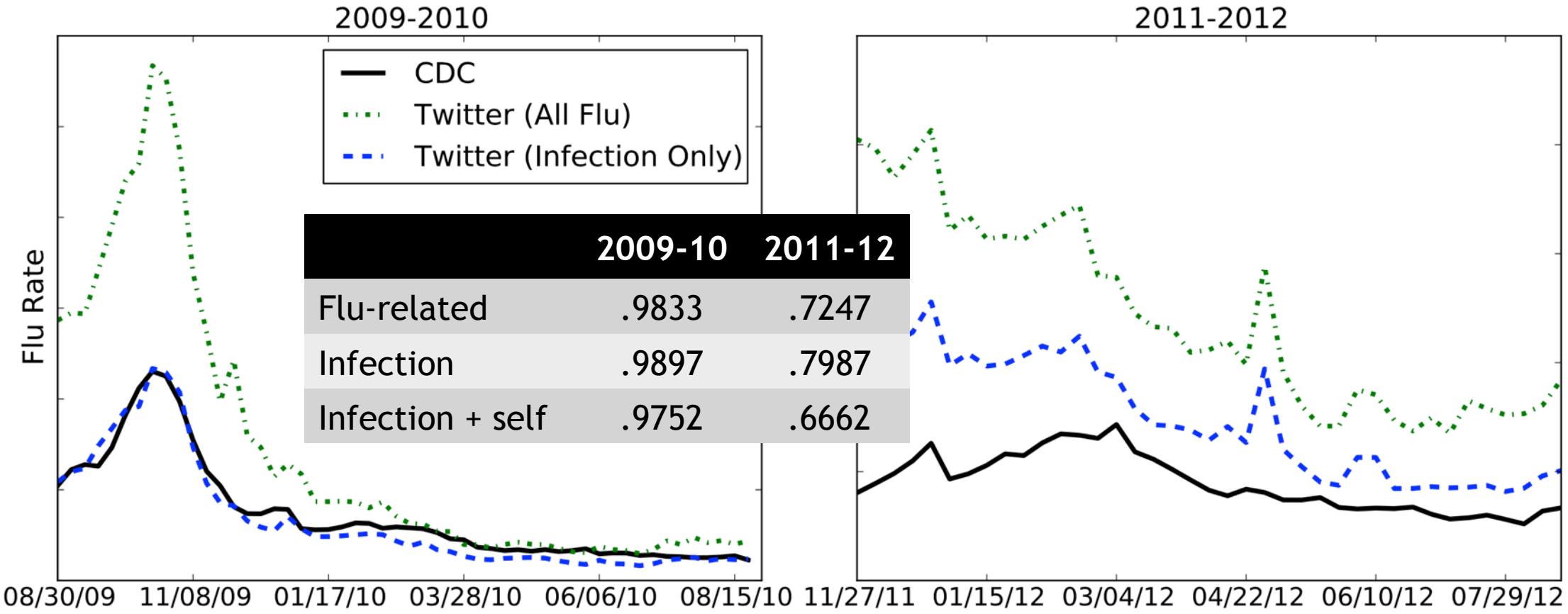
Modeling ILI rates from Twitter (take 4)

- **word classes** defined by manually configured identifiers, e.g.,
  - *infection* ('infected', 'recovered')
  - *concern* ('afraid', 'terrified')
  - *self* ('I', 'my')
- **Twitter specific** features, e.g.,
  - #hashtag, @mentions, emoticons, URLs
- **Part-of-Speech** templates, e.g.,
  - verb-phrase, flu word as noun OR adjective, flu word as noun before first phrase
- All above used as features in a **2-step classification task** using log-linear model with L_2 norm regularization
  - identify illness-related tweets
  - classify **awareness *vs*. infection**
  - then, classify **self-tweets *vs*. tweets for others**

Lamb et al., 2013

Modeling ILI rates from Twitter (take 4)

- **separating infection from awareness** improved correlation with CDC rates, but identification of self tweets did not help



	2009-10	2011-12
Flu-related	.9833	.7247
Infection	.9897	.7987
Infection + self	.9752	.6662

Lamb et al., 2013

Forecasting ILI rates using Twitter

$$y_{t+k} = \gamma \mathrm{ILI}_t^{\mathrm{Twitter}} + \alpha_1 \mathrm{ILI}_{t-1}^{\mathrm{CDC}} + \alpha_2 \mathrm{ILI}_{t-2}^{\mathrm{CDC}} + \alpha_3 \mathrm{ILI}_{t-3}^{\mathrm{CDC}}$$

Twitter-based inference for time instance *t*

Autoregressive components based on ILINet data from CDC for time instances *t*-1, *t*-2 and *t*-3

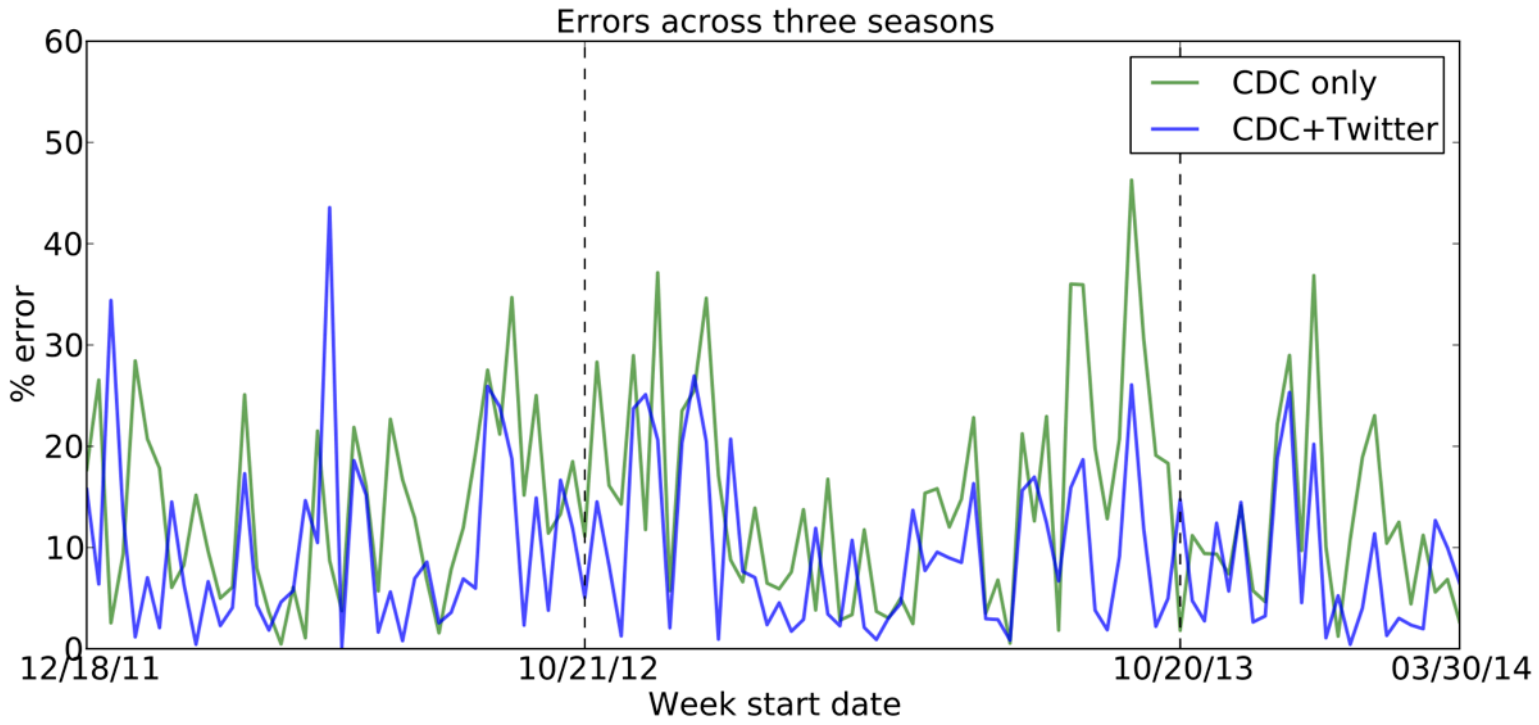
Data / Flu Season	2011-12	2012-13	2013-14
Forecasting using CDC ILI rates with 1-week lag	.20	.30	.32
Nowcasting using Twitter	.33	.36	.48
Nowcasting using Twitter and CDC ILI rates with 1-week lag	.14	.21	.21

Twitter content improves Mean Absolute Error

Paul and Dredze, 2014a

Forecasting ILI rates using Twitter

Lag in weeks	CDC	CDC +Twitter
0	.27 (.06)	.19 (.03)
1	.40 (.12)	.29 (.07)
2	.49 (.17)	.37 (.08)
3	.59 (.22)	.46 (.11)

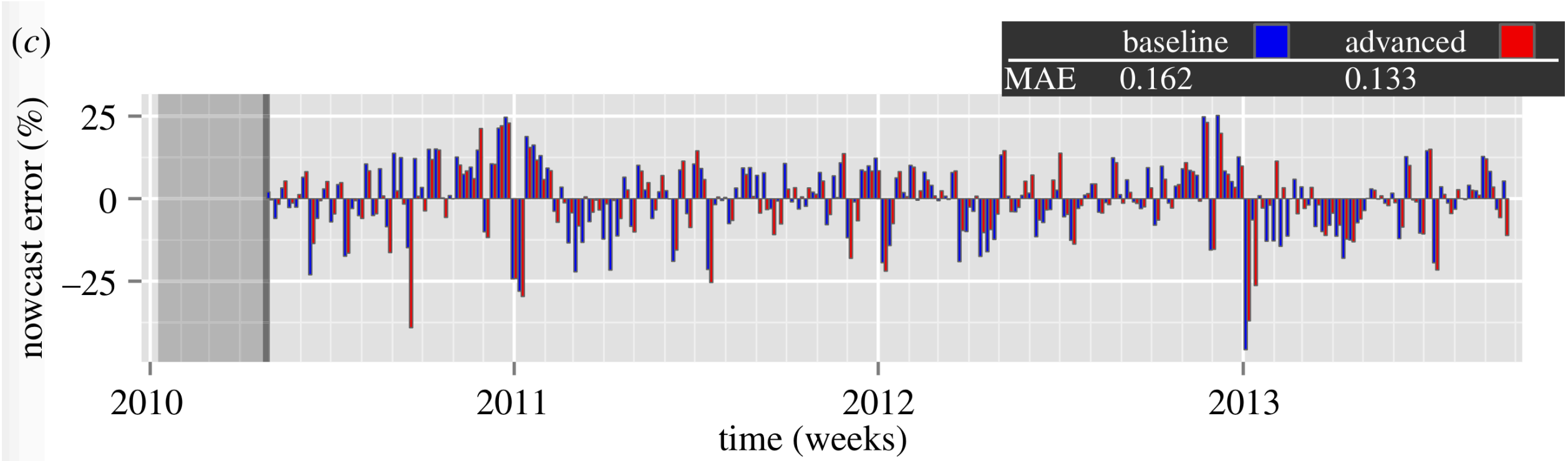


performance measured
by Mean Absolute Error

Paul and Dredze, 2014a

Forecasting ILI using Google Flu Trends

- same story, different source (GFT) and a more advanced better autoregressive model (ARIMA)



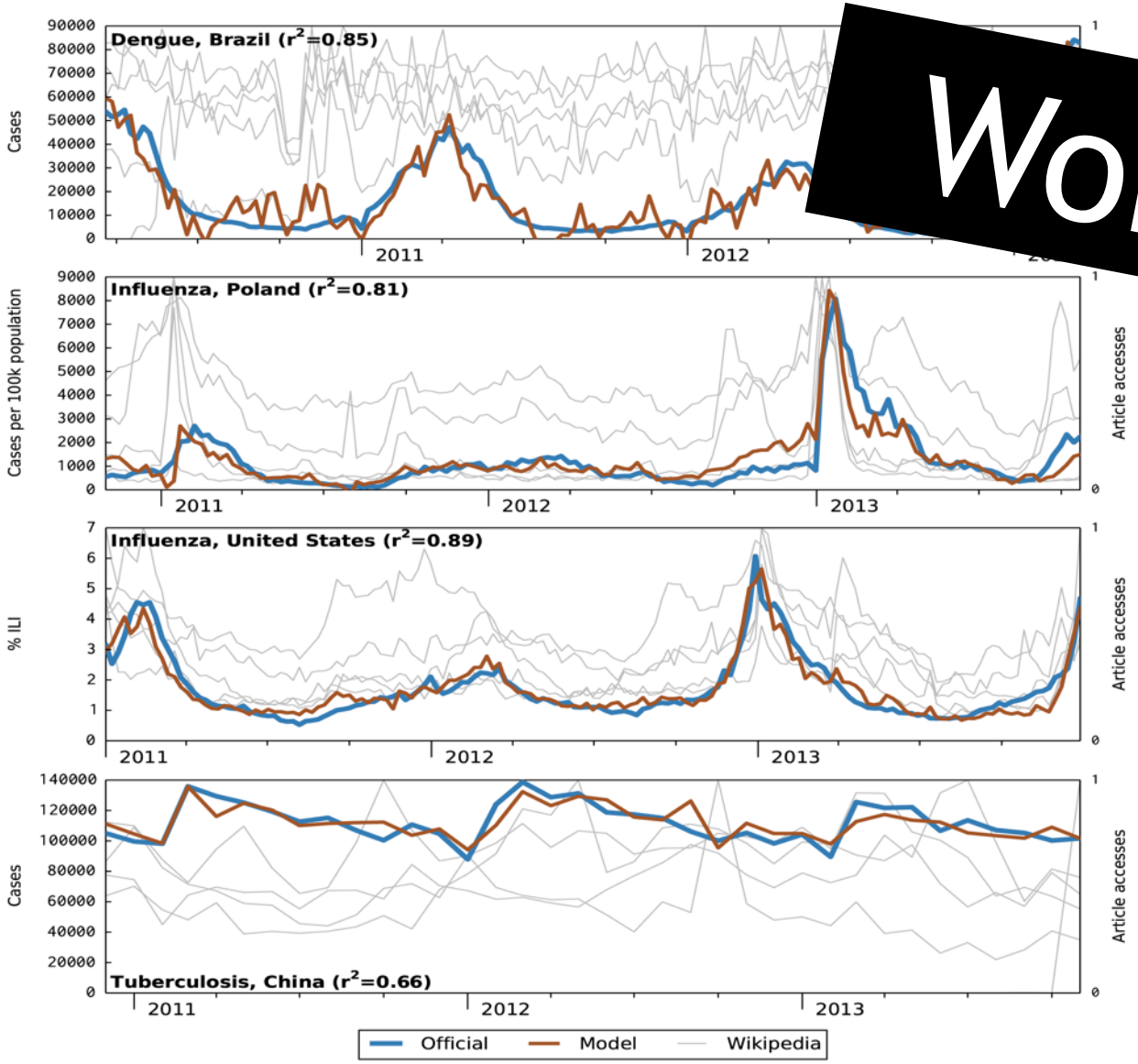
Preis and Moat, 2014

Nowcasting and forecasting diseases via Wikipedia

- explore a different source: **Wikipedia**
- major limitation: use **language** as a proxy for **location**
- number of requests per article (proxy for human views)
- which Wikipedia articles to include?
  - unresolved, manual selection of a pool of articles
  - use the 10 best historically correlated with the target signal (Pearson's *r*)
- ordinary least squares using these 10 "features"
- **not clear what kind of training-testing was performed**
  - performance measured by correlation only
- however, able to test a lot of interesting scenarios

Generous et al., 2014

Nowcasting and forecasting diseases via Wikipedia



Works! ???

Dengue, Brazil (r^2 = .85)

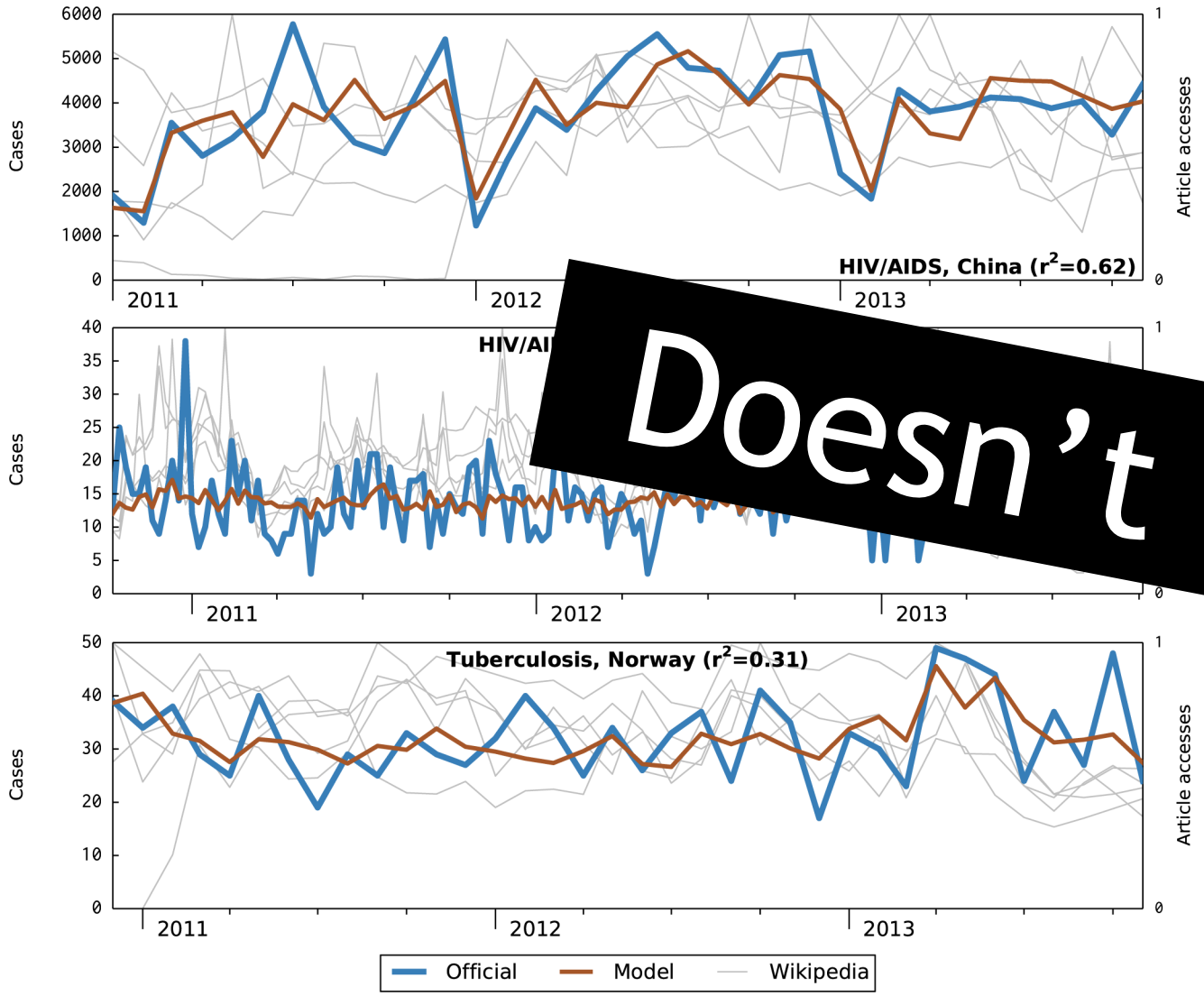
Influenza-like illness, Poland (r^2 = .81)

Influenza-like illness, US (r^2 = .89)

Tuberculosis, China (r^2 = .66)

Generous et al., 2014

Nowcasting and forecasting diseases via Wikipedia



Doesn't work! ???

HIV/AIDS, China (r^2 = .62)

HIV/AIDS, Japan (r^2 = .15)

Tuberculosis, Norway (r^2 = .31)

Generous et al., 2014

Modeling health topics from Twitter

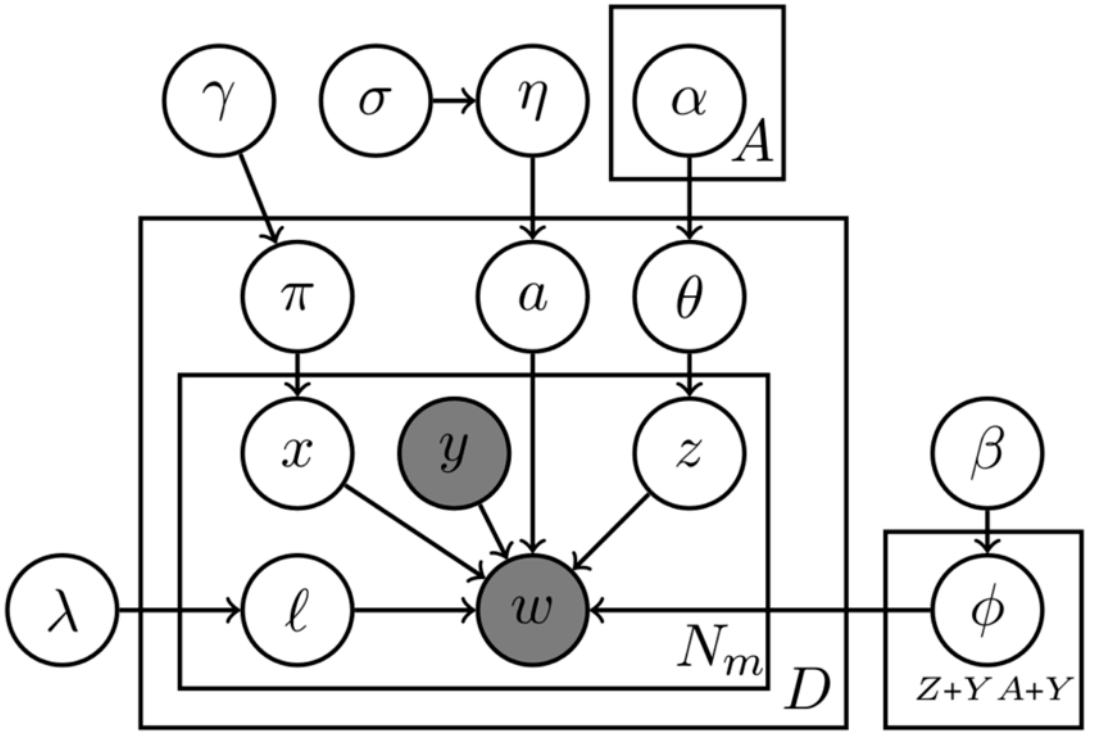
- Instead of focusing on one disease (flu), try to **model multiple health signals**
- (again this is based on intrinsic modeling, not supervised learning)
- **Data**
  - 2 billion tweets from May 2009 to October 2010
  - 4 million tweets/day from August 2011 to February 2013
- **Filtering** by keywords
  - 20,000 keyphrases (from 2 websites) related to illness used to identify symptoms & treatments
  - articles for 20 health issues from WebMD (allergies, cancer, flu, obesity, etc.)
- Mechanical Turk to construct **classifier** to identify health related tweets
  - binary logistic regression with 1-2-3-grams (68% precision, 72% recall)
- Final data set: 144 million health tweets for this work
  - geolocated approximately (Carmen)

Paul and Dredze, 2014b

Modeling health topics from Twitter

Ailment Topic Aspect Model (ATAM)

- variant of Latent Dirichlet Allocation (**LDA**) model, document ~ topics, topic ~ words
- draw focus on health topics
- incorporate background noise
  - word generated under ATAM ~ λ
    background noise ~ 1-λ
- *x* switch: ailment OR common topic
- ℓ switch: background noise or NOT
- each ailment has 3 separate word distributions (***y***): general words, symptoms, treatments



Paul and Dredze, 2014b

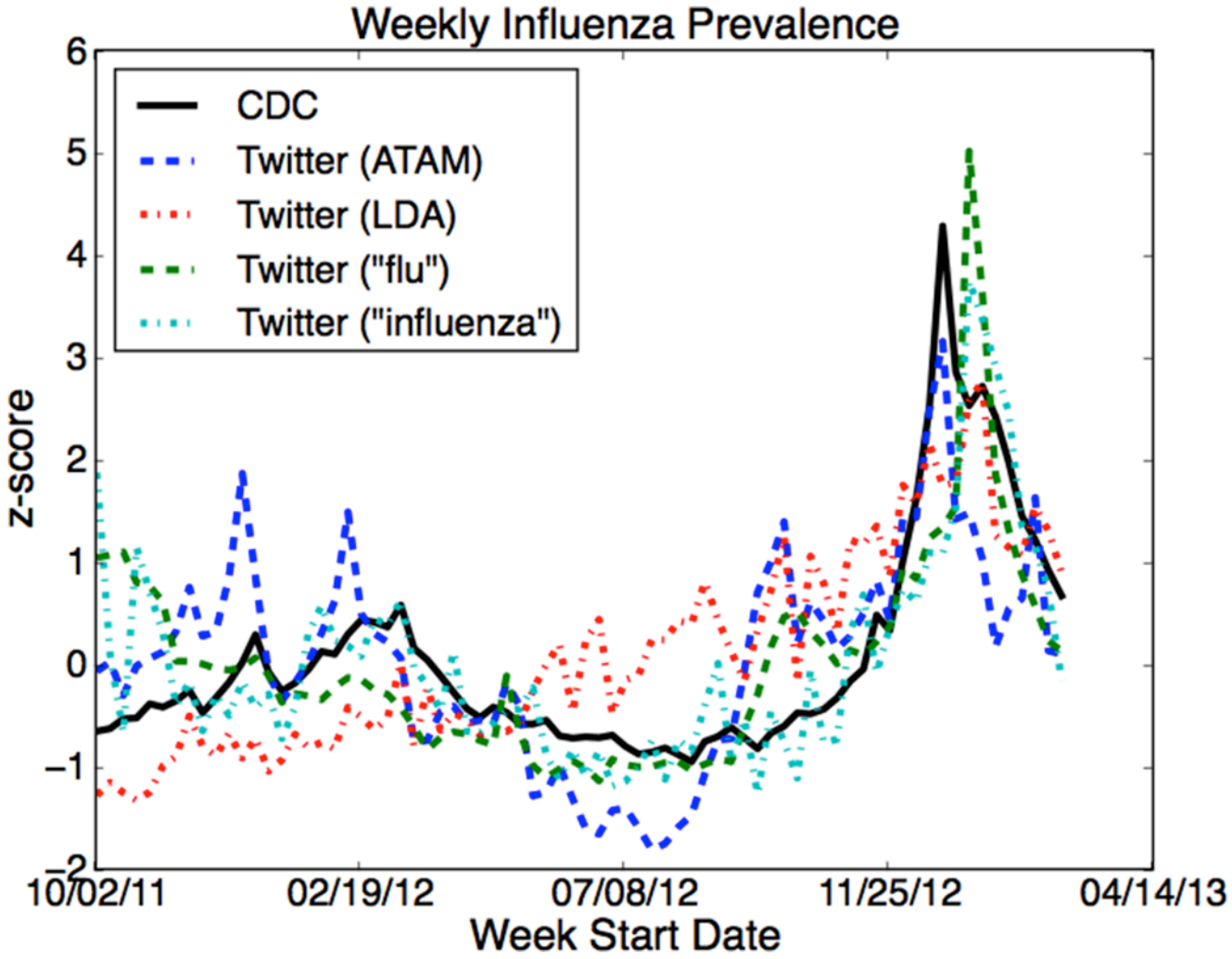
Modeling health topics from Twitter

Non-Ailment Topics				
Conversation	**TV & Movies**	**Games & Sports**	**Family**	**Music**
ok, haha, ha, fine, yeah, thanks	watch, watching, tv, killing, movie, seen	play, game, win, boys, fight, lost, team	mom, shes, dad, says, hes, sister	voice, hear, feelin, night, bit, listening, sound

Ailments				
	Influenza-like illness	**Insomnia & Sleep Issues**	**Diet & Exercise**	**Cancer & Serious Illness**
General words	better, hope, soon, feel, feeling	night, bed, body, tired, work, hours	body, pounds, gym, weight, lost, workout	cancer, help, pray, died, family, friend
Symptoms	sick, sore, throat, fever, cough	sleep, headache, insomnia, sleeping	sore, pain, aching, stomach	cancer, breast, lung, prostate, sad
Treatments	hospital, surgery, paracetamol, antibiotics	sleeping, pills, caffeine, tylenol	exercise, diet, dieting, protein	surgery, hospital, treatment, heart

Paul and Dredze, 2014b

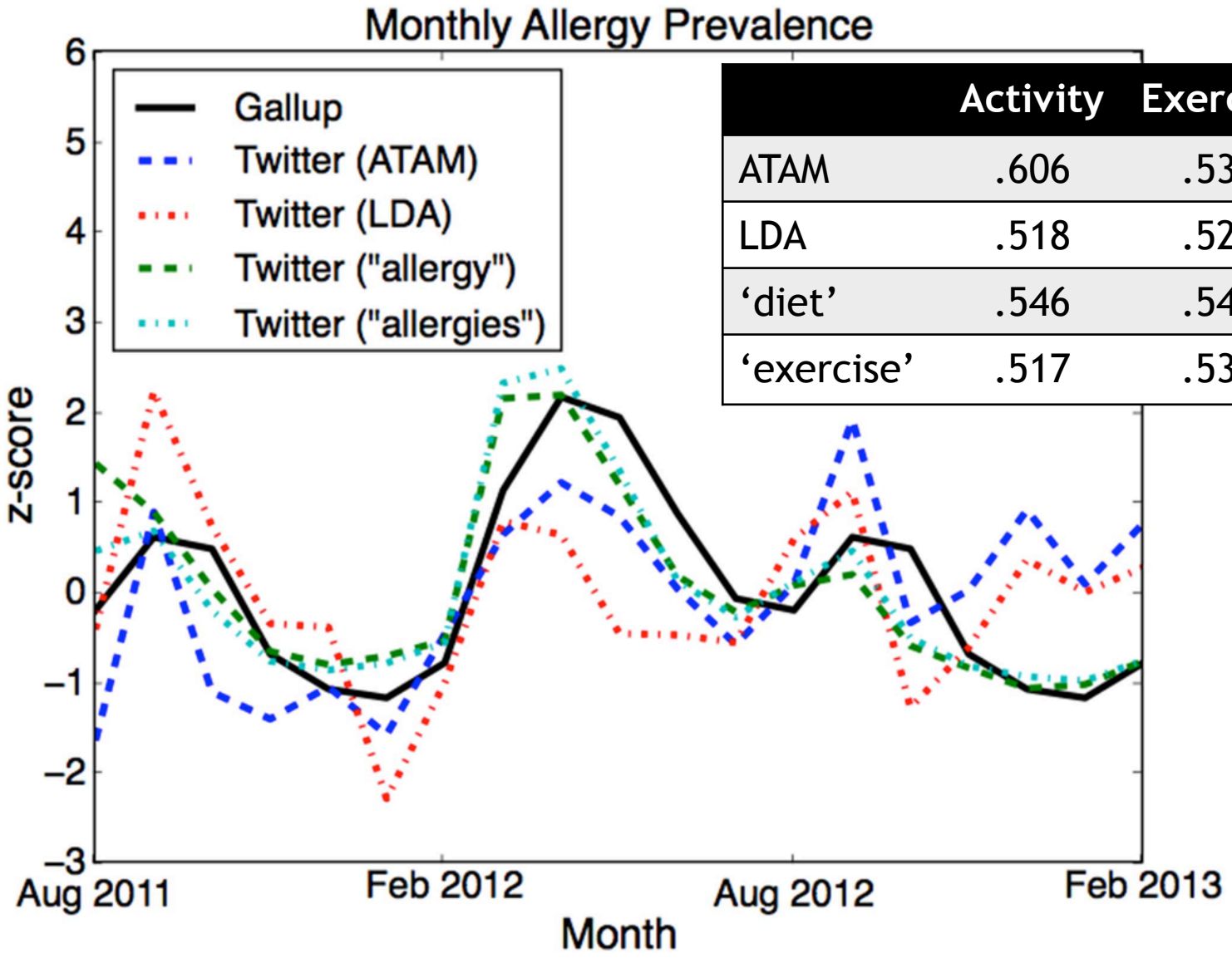
Modeling health topics from Twitter



	2011-12	2012-13	2011-13
ATAM	.613	.643	.689
LDA (1)	**.670**	.198	.455
LDA (2)	−.421	.698	.637
'flu'	.259	.652	.717
'influenza'	.509	**.767**	**.782**

Paul and Dredze, 2014b

Modeling health topics from Twitter

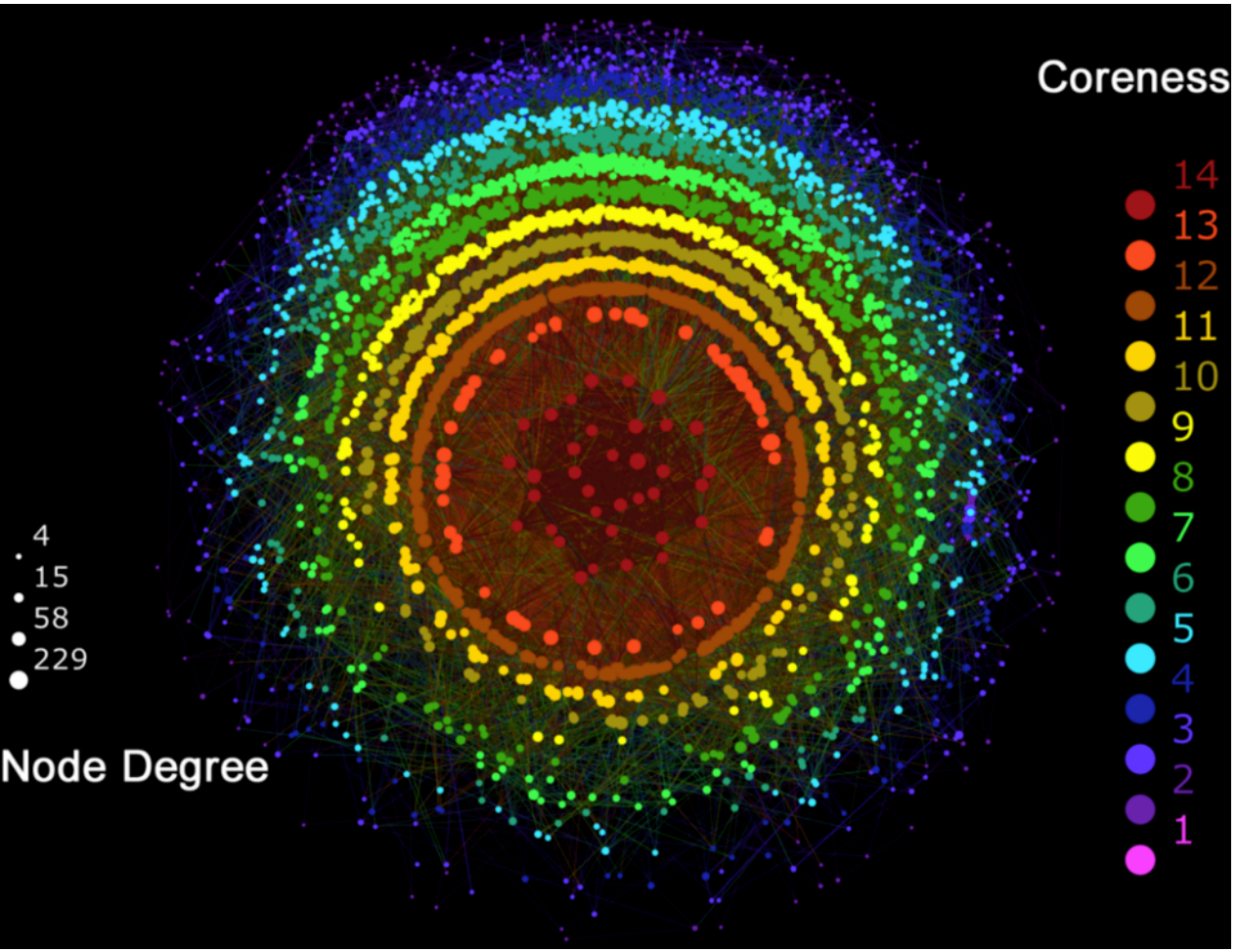


	Activity	Exercise	Obesity	Diabetes	Cholesterol
ATAM	.606	.534	−.631	−.583	-.194
LDA	.518	.521	−.532	−.560	-.146
'diet'	.546	.547	−.567	−.579	-.214
'exercise'	.517	.539	−.505	−.611	-.170

	08/2011 to 04/2012	08/2011 to 02/2013
ATAM	.810	.479
LDA	.705	.366
'allergy'	.873	.823
'allergies'	.922	.877

Paul and Dredze, 2014b

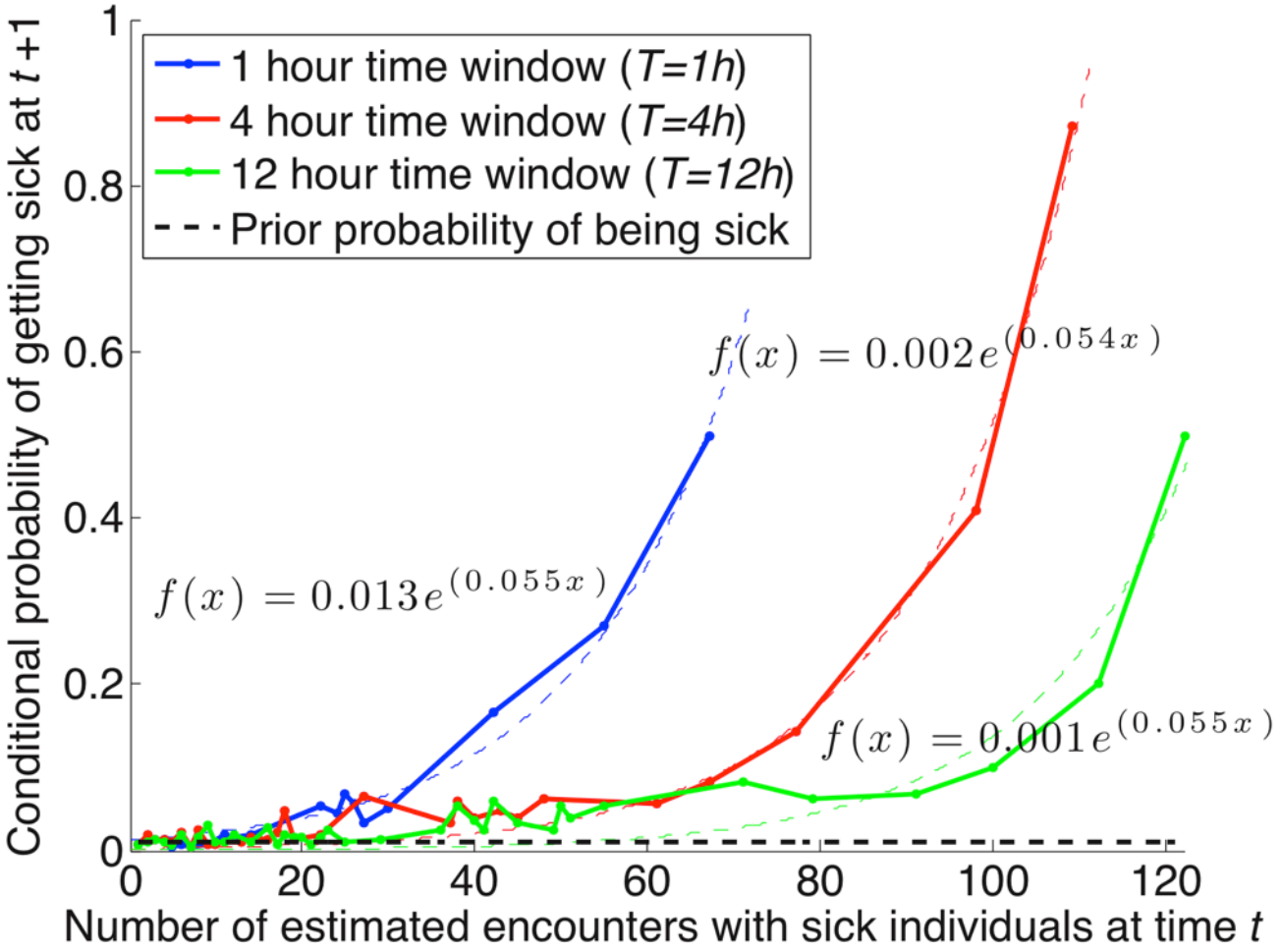
Modeling disease spread from Twitter



- exploring the **social network structure**
- 6,237 geo-active users (NYC)
- 2,535,706 tweets (~ 85K tweets/day)
- 2,047 classified ‘sick’ tweets
  - start from labeled tweets (Mechanical Turk)
  - learn two SVM classifiers: penalized for false positives and negatives
  - feature space: 1-2-3-grams
  - use ROCArea SVM (class imbalance)

Sadilek et al., 2012

Modeling disease spread from Twitter



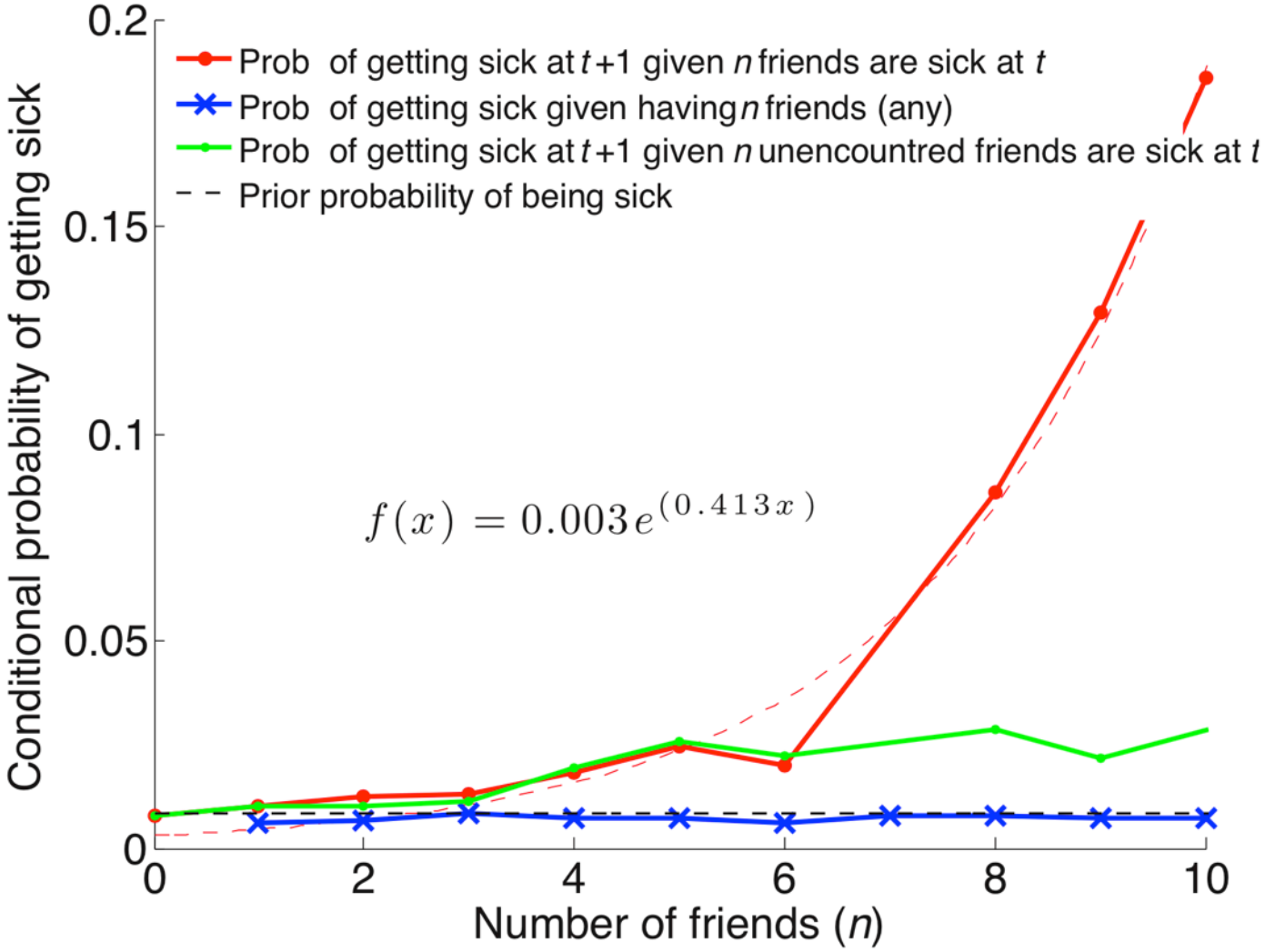
- r = .73 with Google Flu Trends for NYC
- **co-located users:** visit same 100x100 meter cell within T time window
- user considered ill for 2 days after posting a 'sick' tweet
- **probability of getting sick as a function of encounters with sick individuals**

$$f(x) = (0.011 / T) \times e^{0.055x}$$

- proportional to $1/T$
- 100 encounters within T = 4 hours, 40% prob. of getting sick

Sadilek et al., 2012

Modeling disease spread from Twitter



- probability of getting sick as a function of the number of sick friends

Sadilek et al., 2012

Cox hazard models

- Incidence (hazard) rate: number of new cases of disease per population at-risk per unit time (or mortality rate, if outcome is death)
- Hazard:

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}$$

(The probability that ***if you survive to t***, you will succumb to the event in the next instant.)

- Censored vs. non-censored data: Censored data have survived throughout the observation period.

- D.R. Cox (1972) “Regression Models and Life-Tables”

Anorexia and the media

Toolbar data over a period of 5 months, in which we identified two types of behavior:

Celebrity queries

- One of 3640 known celebrities
- Each scored for the probability of them appearing in conjunction with the word “anorexia”
- We refer to this probability as the **Perceived Anorexia Score (PAS).**

Anorexia queries

We define anorexic activity searching (AAS) as one of the following:

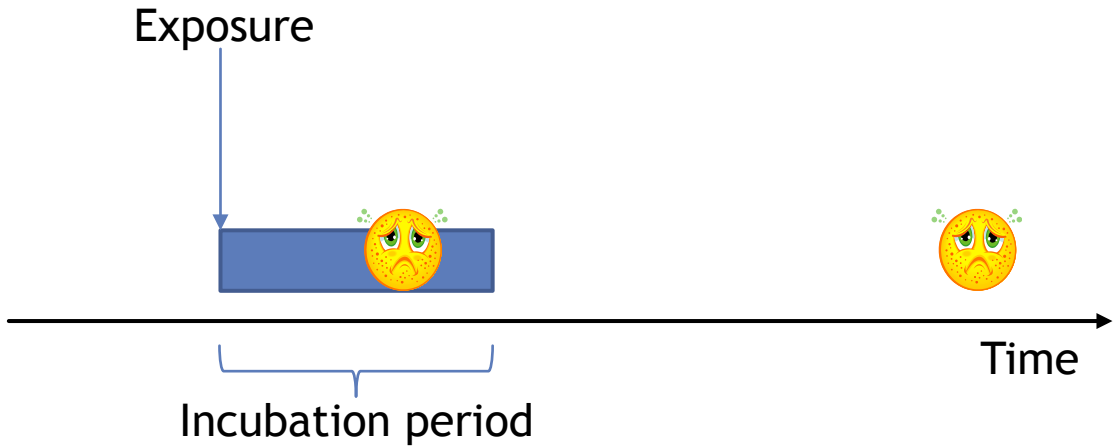
1. Tips for proana or anorexia
2. “how to … ” and proana or anorexia.
3. Proana buddy

A total of 5,800,270 users searched for least one celebrity in the top 2.5% of PAS, of which 3,615 also made AASs.

Hazard models

Attributes	N = 1	
	Weight (s.e.)	Exp(weight)
Number of all searches	1.35*10-3 (5.31*10-5)	1.00
Number of celebrity searches	$-2.06*10^{-3}$ ($1.10*10^{-2}$) N.S.	1.00
Number of searches for top PAS celebrities	$3.24*10^{-3}$ ($1.10*10^{-2}$)	1.03
Number of (unique) top PAS celebrities searched	0.61 ($5.70*10^{-2}$)	1.84
Peak in all Twitter activity	0.29 (0.11)	1.33
Peak in Twitter activity related to anorexia	-0.25 (0.13) N.S.	0.78

Finding precursors:
The Self-Controlled Case Series (SCCS)



$$P(Condition \mid Exposure) = e^{-\lambda_{i,d}} \lambda_{i,d} / y_{i,d}!$$

$$L \propto \prod_{i=1}^{N} \prod_{d=1}^{D} (e^{\beta x_{i,d}} / Z)^{y_{i,d}}$$

$$\lambda_{i,d} = e^{\phi_i + \beta x_{i,d}}$$

Baseline rate

Exposure

Precursors identified

Condition	Precursors	Category or query	Relative hazard
Abortion	Methods of abortion	Category	6.37
Allergy	Petco Pet stores Crops originating from the Americas	Query Category Category	3.88 3.34 2.88
Eating disorder	Image search Bipolar spectrum Depression	Category Category Category	8.14 8.01 6.66
Herpes simplex	Military brats Plenty of fish Redtube	Category Query Query	2.52 2.34 1.49
HIV	Xtube Same sex online dating Adam4adam	Query Category Query	5.50 3.54 3.42
Myocardial infarction	Fast food hamburger restaurants Theme restaurants	Category Category	5.28 4.22

Yom-Tov et al., 2015

Limitations I

- User-generated **data can be biased**
  - very young or very old people are under-represented on social media
  - not all social classes are covered
  - people that post content about topic X may also be a biased subset with characteristics that are difficult to specify

- **Data collection** / formation / extraction can also be biased
  - filtering by approximated location information
  - filtering by specific keywords
  - restrictions due to data sampling (no full data access)

Limitations II

- **Ground truth** from health authorities is not always the “ground truth”
  - syndromic surveillance data are based on people that use medical facilities
  - trained models may not provide new (the correct) information when needed

- Data sets are ‘**big**’ but not always ‘**long**’
  - time-span of the data is also important, not only in the volume
  - in many works, models are not assessed properly
  - strange (unrealistic) training / testing setups

Limitations III

- Using the **loss measure** that benefits my algorithm
  - e.g., predictions measured by Pearson correlation only
  - multiple measures must be applied to cover all angles

- Computer scientists **isolate** themselves from other communities
  - apart from GFT, I have not seen a solid work that health authorities have tried to adapt
  - motivation, aim, results must be defined in collaboration with the health community
  - (it can be a mutual isolation!)

Reducing sampling bias for Twitter studies

- Social media content **NOT representative** of entire population
- Can we address this issue?

Data

- 27 health statistics (e.g., obesity, smoking, uninsured, unemployment) for 100 most populous counties in the US
- 4.31 million tweets from 1.46 million unique users (in approx. 9 months)

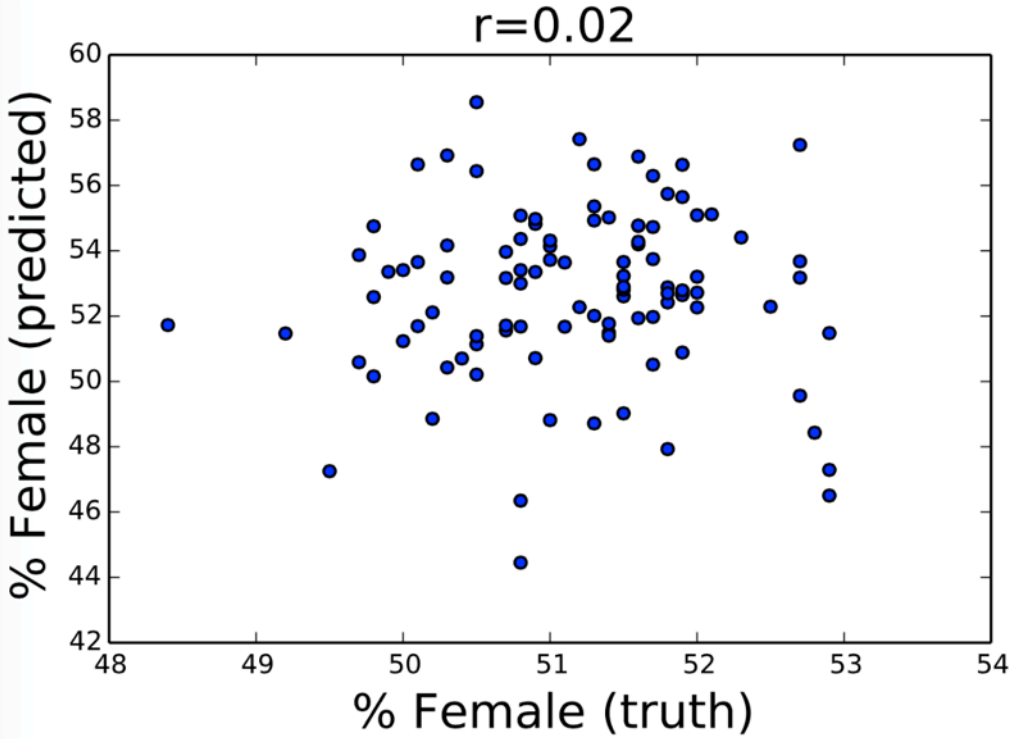
Features - Method

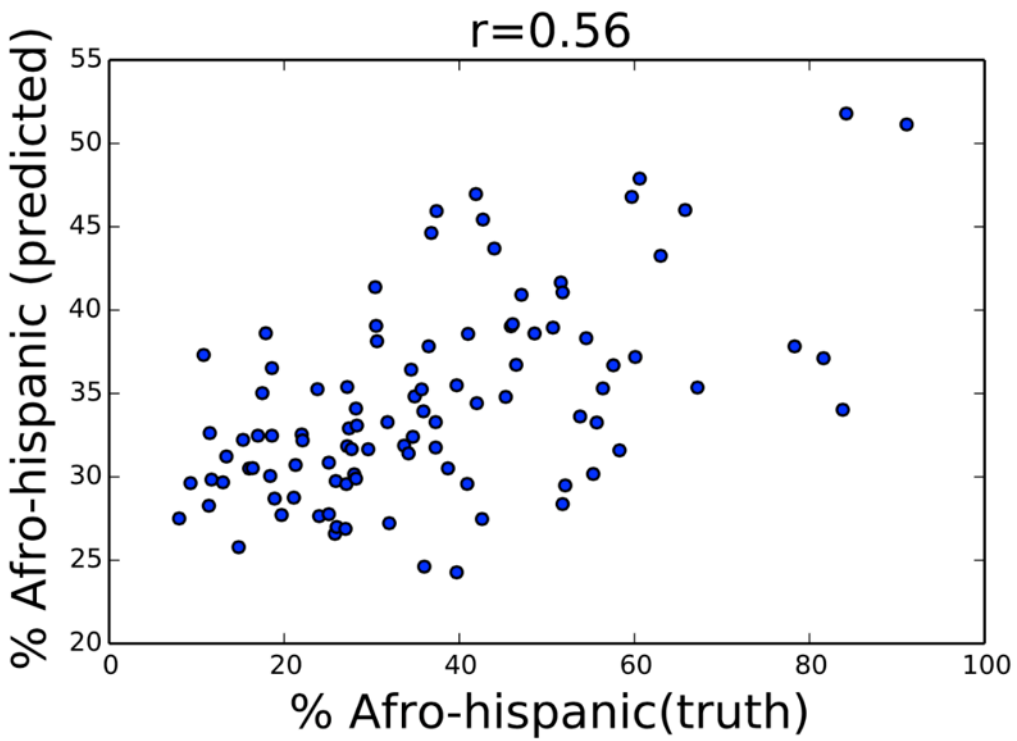
- 70 LIWC (Positive Affect, Family, I) and 10 PERMA (Engagement, Achievement) categories
- 160 features (80+80 for text in tweets and bio description)
- Ridge regression (L2-norm regularization); 5-fold validation; train on 80 counties, predict 20
- Then: **Reweighting** of Twitter features based on gender and race

Culotta, 2014

Reducing sampling bias for Twitter studies

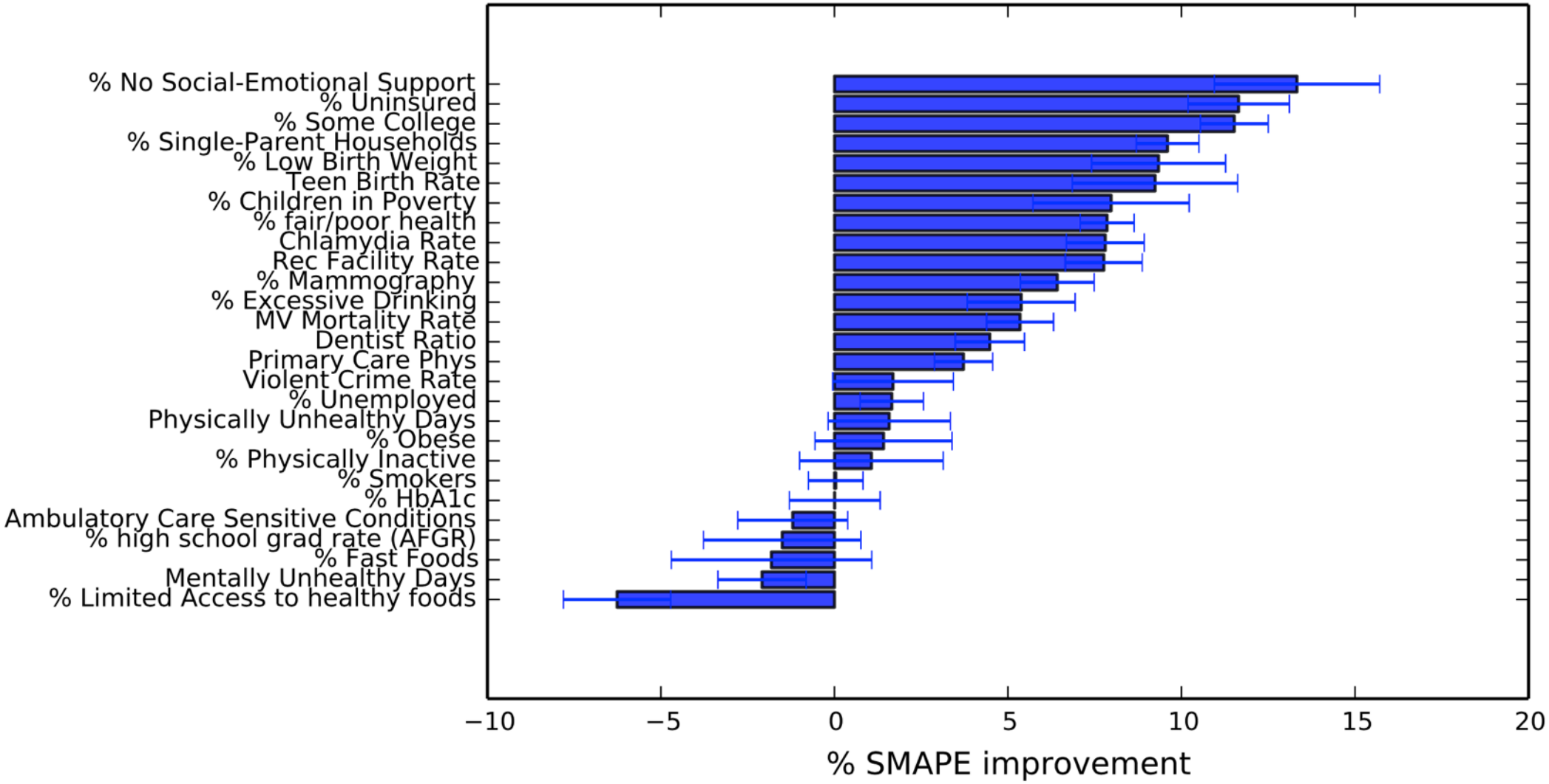
- gender inferred using first names
- race (African American, Hispanic, Caucasian) inferred via a classifier (manually-labeled) using bio information
- Reweighting example: *county's record indicates 60% female, but Twitter estimates 30% female, then tweets from females for this county are counted twice*





Culotta, 2014

Reducing sampling bias for Twitter studies



Predictions are improved on average

Culotta, 2014

Privacy and ethics

Outline

- Some examples
- What is private information?
- What law governs privacy?
- Ethics
- ACM Ethics
- Medical Ethics

Some problems

- Phone records
  - Economist’s ebola article
- Samaritan's suicide prevention app
- http://www.wired.co.uk/news/archive/2014-11/10/samaritans-radar-twitter-app-pulled
- Facebook – emotion engineering PNAS

Institutional Review Boards (IRBs)

- An IRB is a committee that has been formally designated to approve, monitor, and review biomedical and behavioral research involving humans.
- Most countries have some form of IRBs. See http://archive.hhs.gov/ohrp/international/HSPCompilation.pdf
- Human subject research is subject to IRB review in the USA only when it is conducted or funded by any of the Common Rule agencies, or when it will form the basis of an FDA marketing application.

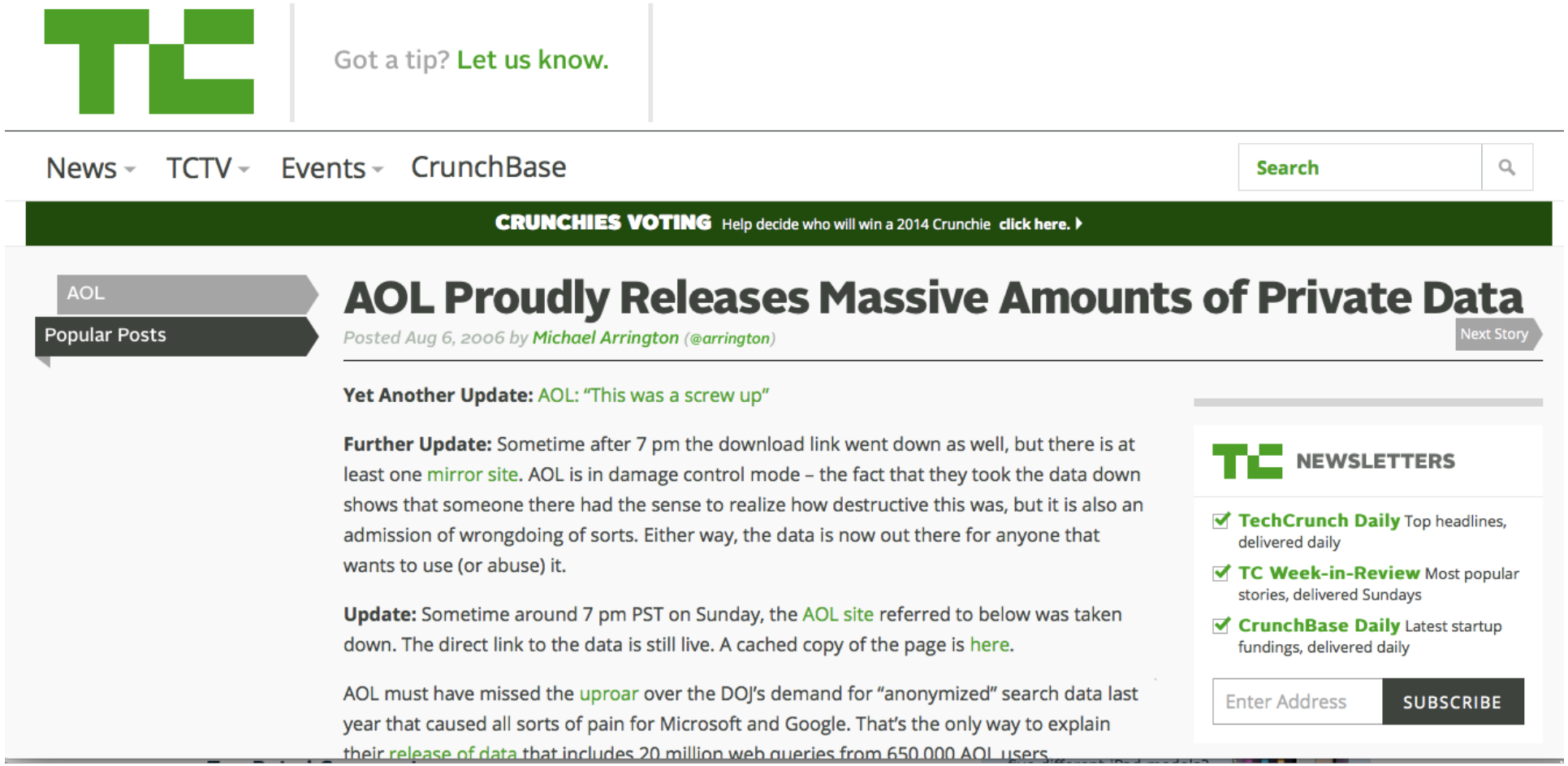
IRB exemptions in the USA

- Research in conventional educational settings, such as those involving the study of instructional strategies or effectiveness of various techniques, curricula, or classroom management methods. In the case of studies involving the use of educational tests, there are specific provisions in the exemption to ensure that subjects cannot be identified or exposed to risks or liabilities.
- Research involving the analysis of **existing data and other materials if they are already publicly available, or where the data can be collected such that individual subjects cannot be identified in any way**.
- Studies intended to assess the performance or effectiveness of public benefit or service programs, or to evaluate food taste, quality, or consumer acceptance.

The chief executive officer of Sun Microsystems said Monday that consumer privacy issues are a "red herring." "**You have zero privacy anyway**," Scott McNealy told a group of reporters and analysts Monday night at an event to launch his company's new Jini technology. "**Get over it**."

http://archive.wired.com/politics/law/news/1999/01/17538

AOL Query Log



Got a tip? Let us know.

News TCTV Events CrunchBase Search

CRUNCHIES VOTING Help decide who will win a 2014 Crunchie click here.

AOL

Popular Posts

AOL Proudly Releases Massive Amounts of Private Data

Posted Aug 6, 2006 by Michael Arrington (@arrington)

Next Story

Yet Another Update: AOL: "This was a screw up"

Further Update: Sometime after 7 pm the download link went down as well, but there is at least one mirror site. AOL is in damage control mode – the fact that they took the data down shows that someone there had the sense to realize how destructive this was, but it is also an admission of wrongdoing of sorts. Either way, the data is now out there for anyone that wants to use (or abuse) it.

Update: Sometime around 7 pm PST on Sunday, the AOL site referred to below was taken down. The direct link to the data is still live. A cached copy of the page is here.

AOL must have missed the uproar over the DOJ's demand for "anonymized" search data last year that caused all sorts of pain for Microsoft and Google. That's the only way to explain their release of data that includes 20 million web queries from 650,000 AOL users

NEWSLETTERS

TechCrunch Daily Top headlines, delivered daily

TC Week-in-Review Most popular stories, delivered Sundays

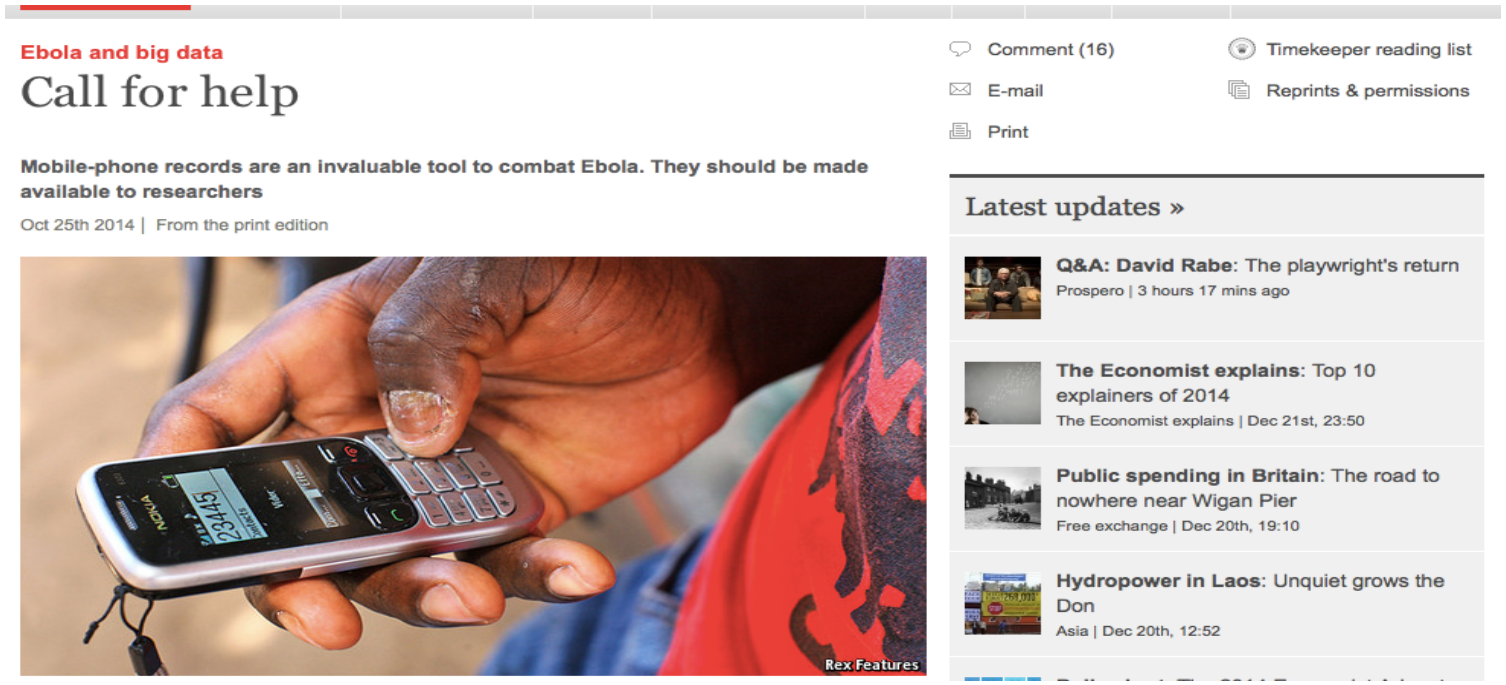
CrunchBase Daily Latest startup fundings, delivered daily

Enter Address SUBSCRIBE

http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

Economist: Call for help

"Governments should require mobile operators to give approved researchers access to their CDRs."



Ebola and big data

Call for help

Mobile-phone records are an invaluable tool to combat Ebola. They should be made available to researchers

Oct 25th 2014 | From the print edition

Rex Features

Comment (16)
Timekeeper reading list
E-mail
Reprints & permissions
Print

Latest updates »

Q&A: David Rabe: The playwright's return
Prospero | 3 hours 17 mins ago

The Economist explains: Top 10 explainers of 2014
The Economist explains | Dec 21st, 23:50

Public spending in Britain: The road to nowhere near Wigan Pier
Free exchange | Dec 20th, 19:10

Hydropower in Laos: Unquiet grows the Don
Asia | Dec 20th, 12:52

http://www.economist.com/news/leaders/21627623-mobile-phone-records-are-invaluable-tool-combat-ebola-they-should-be-made-available

Samaritans pull Twitter app

WIRED.CO.UK TWITTER TECHNOLOGY MENTAL HEALTH

Samaritans pulls Twitter suicide-prevention app

TWITTER / 10 NOVEMBER 14 / by KATIE COLLINS

40 Tweet 2 +1 18 Share submit

Only a week after launching a dedicated Twitter app dedicated to helping those in distress, the suicide prevention charity Samaritans has made the decision to pull Samaritans Radar.

The decision, the organisation writes in a blog post, has been taken due to "serious concerns raised by some people with mental health conditions using Twitter". Samaritans has apologised, saying that its primary concern was to support vulnerable people.

Shutterstock

http://www.wired.co.uk/news/archive/2014-11/10/samaritans-radar-twitter-app-pulled

Facebook

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks. This work also suggests that, in contrast to prevailing assumptions, in-person interaction and nonverbal cues are not strictly necessary for emotional contagion, and that the observation of others' positive experiences constitutes a positive experience for people.

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook's "News Feed" product (8). Because people's friends frequently produce much more content than one person can view, the News Feed filters posts, stories, and activities undertaken by friends. News Feed is the primary manner by which people see content that friends share. Which content is shown or omitted in the News Feed is determined via a ranking algorithm that Facebook continually develops and tests in the interest of showing viewers the content they will find most relevant and engaging. One such test is reported in this study: A test of whether posts with emotional content are more engaging.

The experiment manipulated the extent to which people ($N =$

Facebook

“What corporations can do at will to serve their bottom line, and non-profits can do to serve their cause, we shouldn’t make (even) harder—or impossible—for those seeking to produce generalizable knowledge to do.”

http://www.wired.com/2014/06/everything-you-need-to-know-about-facebooks-manipulative-experiment/

What is privacy?

- EU defines personal data as

Personal data is any information relating to an individual, whether it relates to his or her private, professional or public life.

European Convention

Article 8 of the European Convention on Human Rights, which was drafted and adopted by the Council of Europe in 1950 and meanwhile covers the whole European continent except for Belarus and Kosovo, protects the right to respect for private life: "**Everyone has the right to respect for his private and family life, his home and his correspondence**." Through the huge case-law of the European Court of Human Rights in Strasbourg, privacy has been defined and its protection has been established as a positive right of everyone.

http://en.wikipedia.org/wiki/Privacy_law

United Nations

Article 17 of the International Covenant on Civil and Political Rights of the United Nations of 1966 also protects privacy: "**No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation**. Everyone has the right to the protection of the law against such interference or attacks."

http://en.wikipedia.org/wiki/Privacy_law

Laws

- Privacy laws vary by jurisdiction (EU – Constitution, USA – laws)

- Specific privacy laws that are designed to regulate specific types of information. Some examples include:
  - Communication privacy laws
  - Financial privacy laws
  - Health privacy laws
  - Information privacy laws
  - Online privacy laws
  - Privacy in one's home

OECD

GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA

Adopted by the Council of Ministers of the Organisation for Economic Co-operation and Development (OECD) on 23 September 1980

http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

OECD Guidelines

GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA

Adopted by the Council of Ministers of the Organisation for Economic Co-operation and Development (OECD) on 23 September 1980

http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

1. **Collection Limitation Principle**: There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.
2. **Data Quality Principle**: Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.
3. **Purpose Specification Principle**: The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

OECD Guidelines

4. **Use Limitation Principle**: Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Paragraph 9 [3] except:
    a. with the consent of the data subject; or
    b. by the authority of law.
5. **Security Safeguards Principle**: Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorised access, destruction, use, modification or disclosure of data.
6. **Openness Principle**: There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

OECD Guidelines

7. **Individual Participation Principle**—An individual should have the right:

a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him;

b) to have communicated to him, data relating to him within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him;

c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and

d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.

8. **Accountability Principle**—A data controller should be accountable for complying with measures which give effect to the principles stated above.

Jurisdiction

- Data in the cloud
- Export of data

Bits

LEGAL CASES

Judge Rules That Microsoft Must Turn Over Data Stored in Ireland

By NICK WINGFIELD JULY 31, 2014 3:50 PM 40 Comments

Microsoft is challenging the authority of federal prosecutors to force it to hand over a customer's email stored in a data center in Ireland. Microsoft

General guidelines

- Use anonymous data
- Do not try to de-anonymize
- Wherever possible, use aggregate data
- Only collect what you need

Ethics

ACM Code of Ethics

Consists of:

1. General Moral Imperatives.
2. More Specific Professional Responsibilities.
3. Organizational Leadership Imperatives.
4. Compliance with the Code.
5. Acknowledgments.

http://www.acm.org/about/code-of-ethics?searchterm=ethics

ACM Code of Ethics

1.7 Respect the privacy of others.

Computing and communication technology enables the collection and exchange of personal information on a scale unprecedented in the history of civilization. Thus there is increased potential for violating the privacy of individuals and groups. **It is the responsibility of professionals to maintain the privacy and integrity of data describing individuals.** This includes taking precautions to ensure the accuracy of data, as well as protecting it from unauthorized access or accidental disclosure to inappropriate individuals. Furthermore, procedures must be established to allow individuals to review their records and correct inaccuracies.

This imperative implies that only the necessary amount of personal information be collected in a system, that retention and disposal periods for that information be clearly defined and enforced, and that personal information gathered for a specific purpose not be used for other purposes without consent of the individual(s). These principles apply to electronic communications, including electronic mail, and prohibit procedures that capture or monitor electronic user data, including messages, without the permission of users or bona fide authorization related to system operation and maintenance. User data observed during the normal duties of system operation and maintenance must be treated with strictest confidentiality, except in cases where it is evidence for the violation of law, organizational regulations, or this Code. In these cases, the nature or contents of that information must be disclosed only to proper authorities.

WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects

- 1. The World Medical Association (WMA) has developed the Declaration of Helsinki as a statement of **ethical principles for medical research involving human subjects**, including research on identifiable human material and data.
- 23. The research protocol must be submitted for consideration, comment, guidance and approval to the concerned research ethics committee before the study begins.

- http://www.wma.net/en/30publications/10policies/b3/

Medical ethics: four principles plus attention to scope

Raanan Gillon

The "four principles plus scope" approach provides a simple, accessible, and culturally neutral approach to thinking about ethical issues in health care. The approach, developed in the United States, is based on four common, basic prima facie moral commitments—respect for autonomy, beneficence, non-maleficence, and justice—plus concern for their scope of application. It offers a common, basic moral analytical framework and a common, basic moral language. Although they do not provide ordered rules, these principles can help doctors and other health care workers to make decisions when reflecting on moral issues that arise at work.

in committing ourselves to four prima facie moral principles plus a reflective concern about their scope of application. Moreover, these four principles, plus attention to their scope of application, encompass most of the moral issues that arise in health care.

The four prima facie principles are respect for autonomy, beneficence, non-maleficence, and justice. "Prima facie," a term introduced by the English philosopher W D Ross, means that the principle is binding unless it conflicts with another moral principle —if it does we have to choose between them. The four principles approach does not provide a method for choosing, which is a source of dissatisfaction to people

http://www.bmj.com/content/309/6948/184

Four principles

- **Respect for autonomy:** The patient has the right to refuse or choose their treatment.
- **Beneficence: A** practitioner should act in the best interest of the patient.
- **Non-maleficence:** "first, do no harm"
- **Justice: C**oncerns the distribution of scarce health resources, and the decision of who gets what treatment (fairness and equality).

Open questions

Some open questions

- Generalization
- Moving to interventions
- Is online surveillance worth it? Is early detection worth it?
- Integration of multiple data sources for more accurate prediction
- Social networks and health
- Models:
  - We know when anonymous users are ill. How do we know when they get better?
  - Dynamic modelling: How do systems change with time?
- Policy:
  - Dealing with privacy in a more principled manner
  - Access to data for research

That’s all folks!

References I

Jiang Bian, Umit Topaloglu, Fan Yu (2012) Towards Large-scale Twitter Mining for Drug-related Adverse Events

Brennan, Sadilek, Kautz (2013) Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions

John S. Brownstein, Clark C. Freifeld, Lawrence C. Madoff, (2010) Digital disease detection--harnessing the Web for public health surveillance

Chew, Cynthia and Eysenbach, Gunther (2010) Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak

Nicholas A Christakis, James H Fowler (2007) The spread of obesity in a large social network over 32 years

Cook, Samantha and Conrad, Corrie and Fowlkes, Ashley L and Mohebbi, Matthew H (2011) Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic

Civljak M, Sheikh A, Stead LF, Car J (2010) Internet-based interventions for smoking cessation. Cochrane Databse Syst Rev:CD007078

Glen A. Coppersmith, Craig T. Harman, Mark H. Dredze (2014) Measuring Post Traumatic Stress Disorder in Twitter

Aron Culotta (2010) Towards detecting influenza epidemics by analyzing Twitter messages

Aron Culotta (2013) Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages

Aron Culotta (2014) Reducing Sampling Bias in Social Media Data for County Health Inference

Sean D. Young, Caitlin Rivers, Bryan Lewis (2014) Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes

Munmun De Choudhury, Meredith Ringel Morris, Ryen W. White (2014) Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media

Munmun De Choudhury, Scott Counts, Eric Horvitz (2013) Predicting postpartum changes in emotion and behavior via social media

Munmun De Choudhury Michael Gamon Scott Counts Eric Horvitz (2013) Predicting depression via social media

Gunther Eysenbach (2006) Tracking flu-related searches on the Web for syndromic surveillance

Vanessa Frias-Martinez, Alberto Rubio, Enrique Frias-Martinez (2012) Measuring the impact of epidemic alerts on human mobility using cell-phone network data

Generous, Fairchild, Deshpande, Del Valle and Priedhorsky (2014) Global Disease Monitoring and Forecasting with Wikipedia

References II

Ginsberg, Jeremy and Mohebbi, Matthew H. and Patel, Rajan S. and Brammer, Lynnette and Smolinski, Mark S. and Brilliant, Larry (2009) Detecting influenza epidemics using search engine query data

Cassandra Harrison, Mohip Jorder, Henri Stern, Faina Stavinsky, Vasudha Reddy, Heather Hanson, HaeNa Waechter, Luther Lowe, Luis Gravano, Sharon Balter (2014) Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness — New York City, 2012–2013

Meghan Kuebler, Elad Yom-Tov, Dan Pelleg, Rebecca M. Puhl, Peter Muennig (2013) When Overweight Is the Normal Weight: An Examination of Obesity Using a Social Media Internet Database

Lamb, Alex and Paul, Michael J. and Dredze, Mark (2013) Separating fact from fear: Tracking flu infections on Twitter

A. D. I. Kramer, J. E. Guillory, J. T. Hancock (2013) Experimental evidence of massive-scale emotional contagion through social networks

Vasileios Lampos, Nello Cristianini (2010) Tracking the flu pandemic by monitoring the Social Web

Vasileios Lampos, Tijl De Bie, Nello Cristianini (2010) Flu Detector - Tracking Epidemics on Twitter

Vasileios Lampos, Nello Cristianini (2012) Nowcasting Events from the Social Web with Statistical Learning

Lazer, Kennedy, King and Vespignani (2014) The Parable of Google Flu: Traps in Big Data Analysis

Russell Lyons (2011) The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis

Milinovich, Gabriel J and Williams, Gail M and Clements, Archie C A and Hu, Wenbiao (2013) Internet-based surveillance systems for monitoring emerging infectious diseases

Jane P Messina, Oliver J Brady, David M Pigott, John S Brownstein, Anne G Hoen & Simon I Hay (2014) A global compendium of human dengue virus occurrence

Jane P. Messina, Oliver J. Brady, David M. Pigott, John S. Brownstein, Anne G. Hoen and Simon I. Hay (2014) A global compendium of human dengue virus occurrence

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, J. Niels Rosenquist (2011) Understanding the Demographics of Twitter Users

Yishai Ofran, Ora Paltiel, Dan Pelleg, Jacob M. Rowe, Elad Yom-Tov (2012) Patterns of Information-Seeking for Cancer on the Internet: An Analysis of Real World Data

References III

Olson, Donald R and Konty, Kevin J and Paladini, Marc and Viboud, Cecile and Simonsen, Lone (2013) Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales

Paul and Dredze (2011) You Are What You Tweet: Analyzing Twitter for Public Health

Paul and Dredze (2014) Twitter Improves Influenza Forecasting (a)

Paul and Dredze (2014) Discovering Health Topics in Social Media Using Topic Model (b)

Dan Pelleg, Elad Yom-Tov, Yoelle Maarek (2012) Can you believe an anonymous contributor? On truthfulness in Yahoo! Answers

Dan Pelleg, Denis Savenkov, Eugene Agichtein (2013) Touch Screens for Touchy Issues: Analysis of Accessing Sensitive Information from Mobile Devices

Polgreen, Philip M and Chen, Yiling and Pennock, David M and Nelson, Forrest D (2008) Using internet searches for influenza surveillance

Preis and Moat (2014) Adaptive nowcasting of influenza outbreaks using Google searches

L. Richiardi, C Pizzi, D. Paolotti (2014) Internet-Based Epidemiology

Sadilek, Kautz and Silenzio (2012) Modeling Spread of Disease from Social Interactions

Adam Sadilek, Henry Kautz (2013) Modeling the Impact of Lifestyle on Health at Scale

Simmons RD, Ponsonby AL, van der Mei IA, Sheridan P (2004) What affects your MS? Responses to an anonmous, internet-based epidemiological survey. Mult Scler 10:202-211

Ryen R. White, Eric Horvitz (2012) Studies on the onset and persistence of medical concerns in search logs.

Ryen R. White, Eric Horvitz (2009) Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search

Paul Wicks, Timothy E Vaughan, Michael P Massagli, James Heywood (2011) Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm

Elad Yom-Tov, Luis Fernandez-Luque, Ingmar Weber, Steven P Crain (2012) Pro-Anorexia and Pro-Recovery Photo Sharing: A Tale of Two Warring Tribes

Elad Yom-Tov, Evgeniy Gabrilovich (2013) Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries

References IV

Elad Yom-Tov, danah boyd (2014) On the link between media coverage of anorexia and pro-anorexic practices on the web

Elad Yom-Tov, Ryen W White, Eric Horvitz (2014) Seeking Insights About Cycling Mood Disorders via Anonymized Search Logs

Elad Yom-Tov, Diana Borsa, Ingemar J Cox, Rachel A McKendry (2014) Detecting Disease Outbreaks in Mass Gatherings Using Internet Data

Elad Yom-Tov, Luis Fernandez-Luque (2014) Information is in the eye of the beholder: Seeking information on the MMR vaccine through an Internet search engine

Young, Rivers and Lewis (2014) Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes

Shaodian Zhang, Erin Bantum, Jason Owen, Noémie Elhadad (2014) Does Sustained Participation in an Online Health Community Affect Sentiment?

Further reading

- http://www.hhs.gov/ohrp/policy/engage08.html
  - Guidance on Engagement of Institutions in Human Subjects Research
- “Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines”, F. H. Cate, P. Cullen, V. Mayer-Schönberger, (2014)
- NINFEA Project (2011) www.progettonifea.it
- Influenzanet https://www.influenzanet.eu/