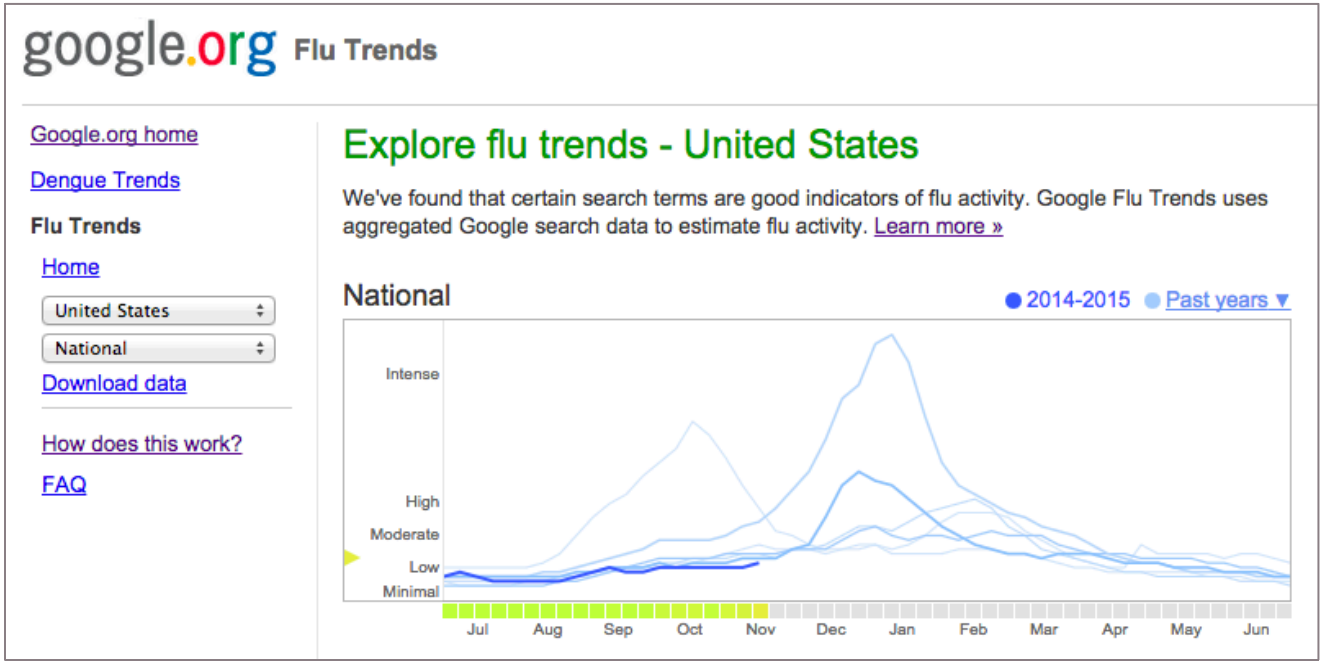
MAKING SENSE OF THE WEB FOR PUBLIC HEALTH USING NLP

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY

December 3, 2014. University of Maryland College Park.

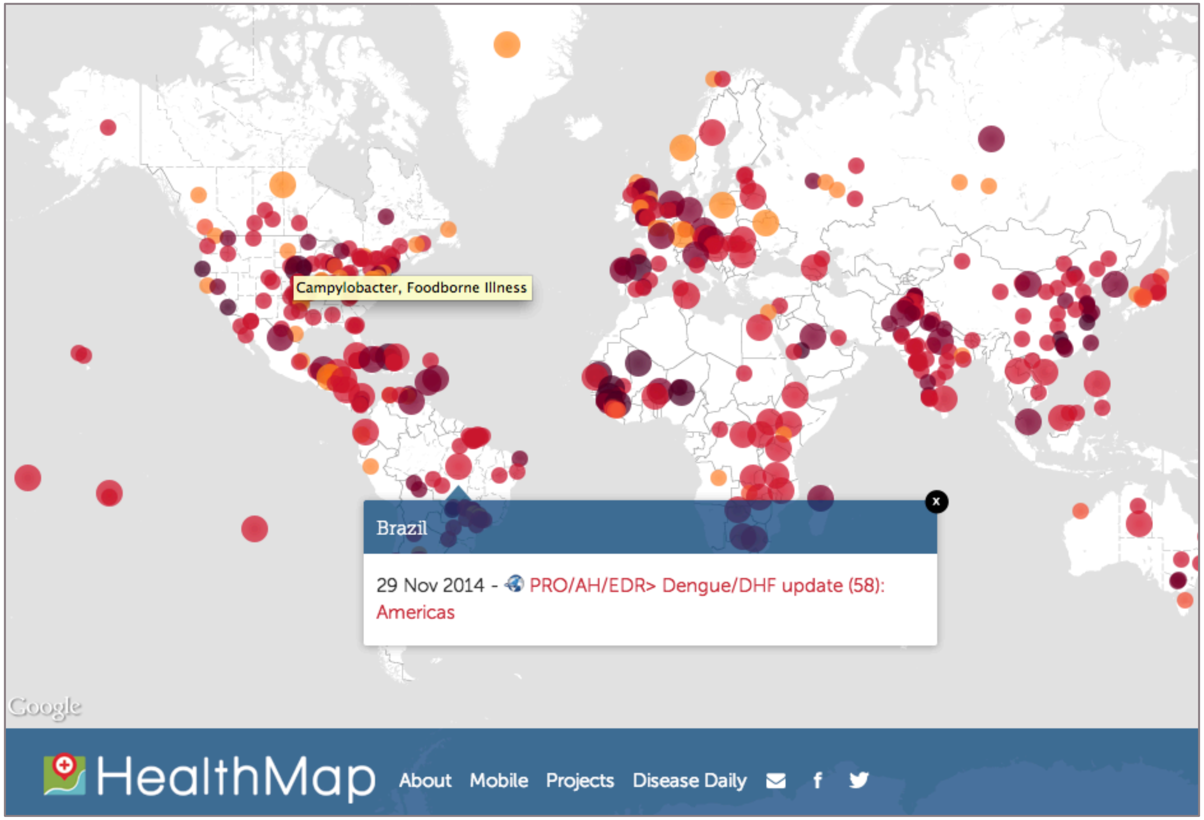
PUBLIC HEALTH + WEB

Google Flu Trends:

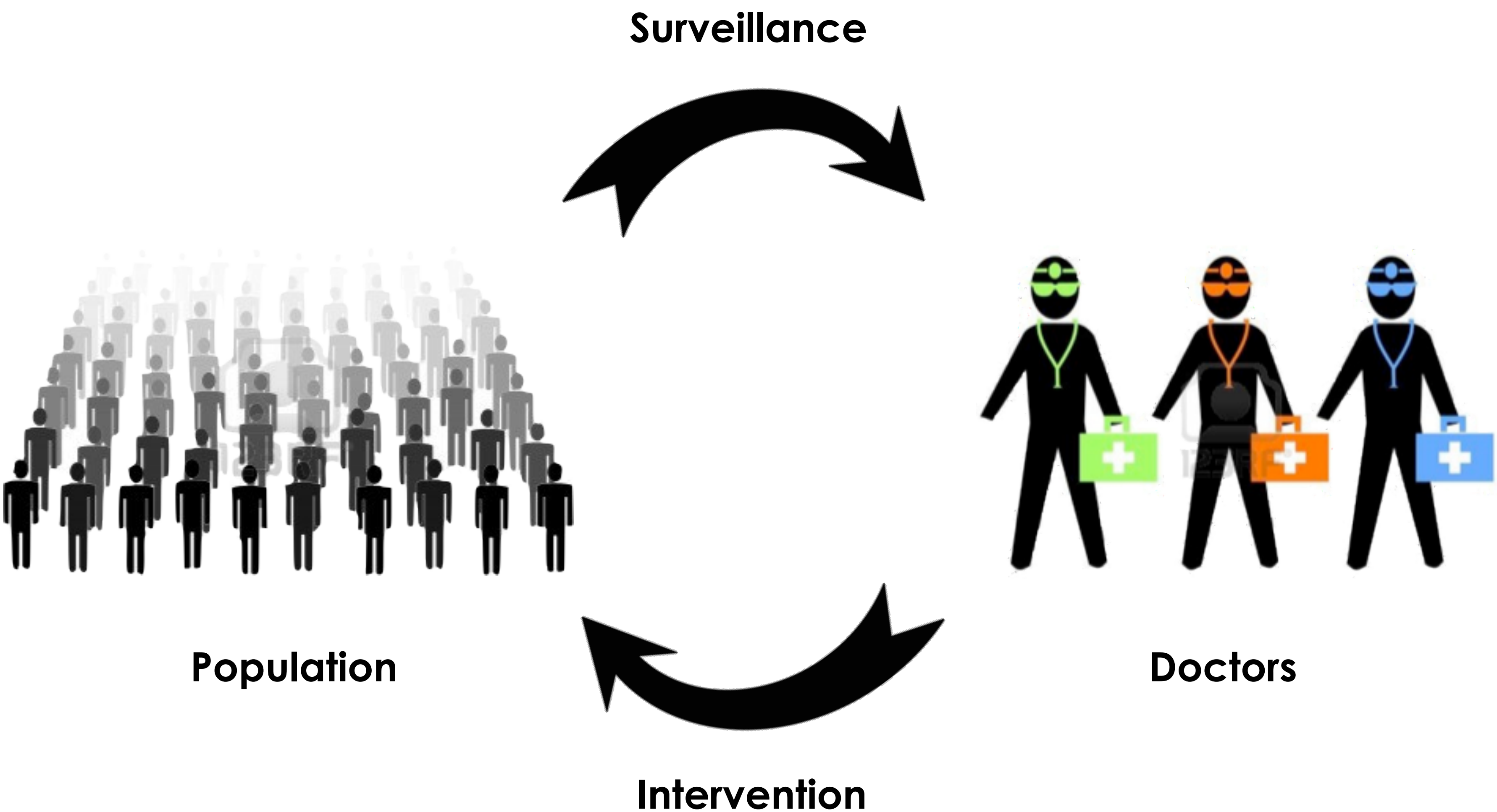


PUBLIC HEALTH + WEB

HealthMap:

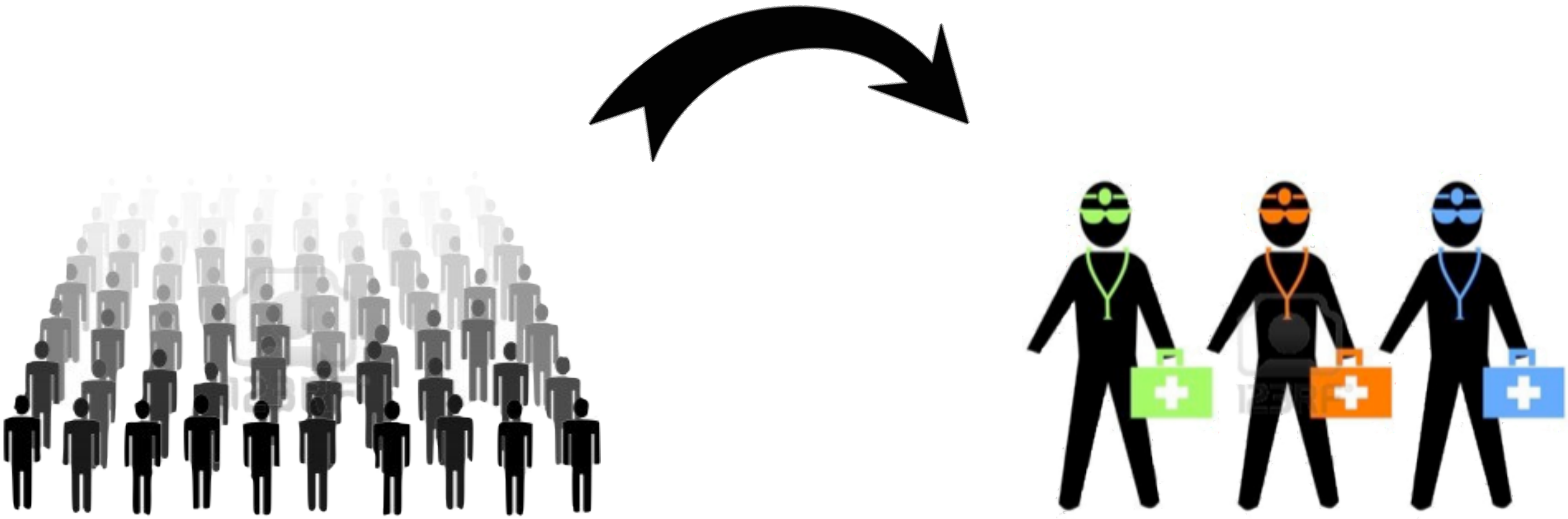


PUBLIC HEALTH



EPIDEMIOLOGY

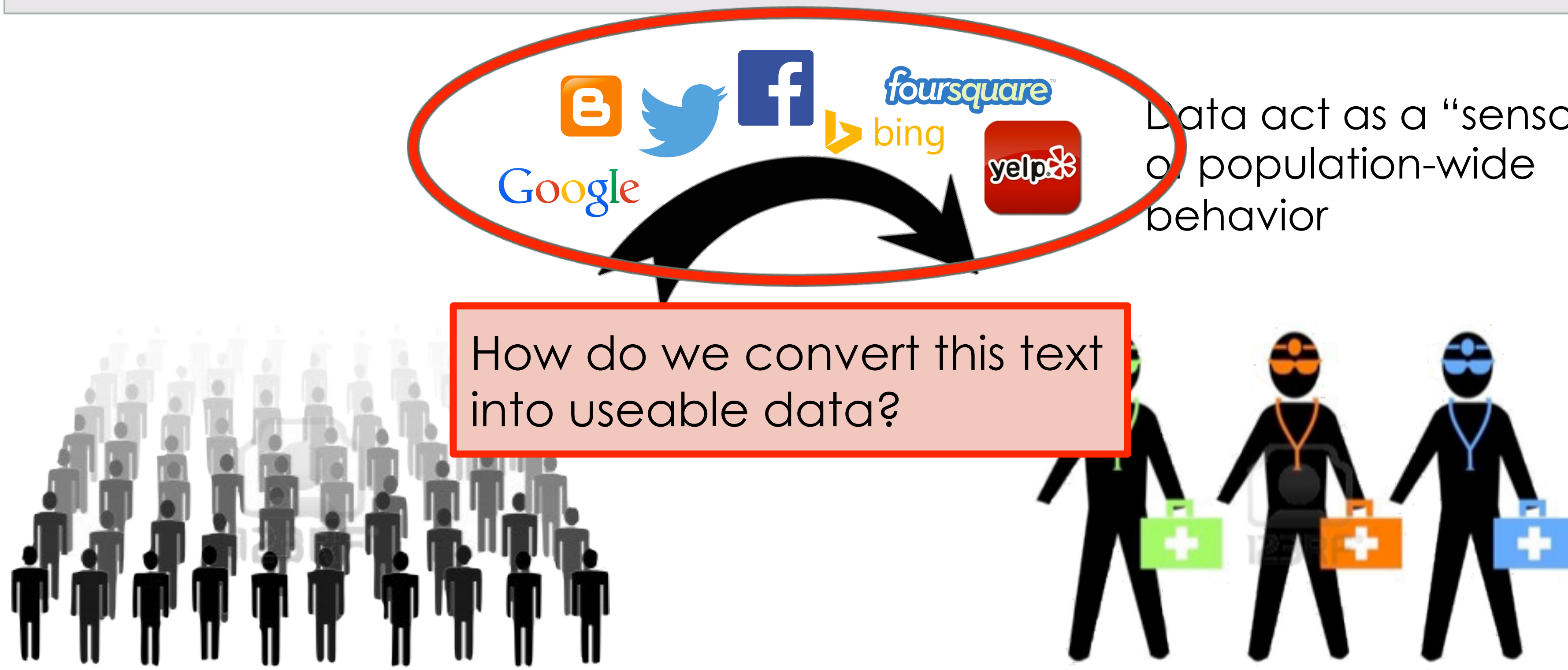
Surveillance



COMPUTATIONAL EPIDEMIOLOGY

Data act as a "sensor" of population-wide behavior

COMPUTATIONAL EPIDEMIOLOGY



Data act as a “sensor” of population-wide behavior

How do we convert this text into useable data?

FROM TEXT TO DATA

- Simplest approach: keyword counting

A. Seifter et al. - Geospatial Health 4(2), 2010, pp. 135-137 137

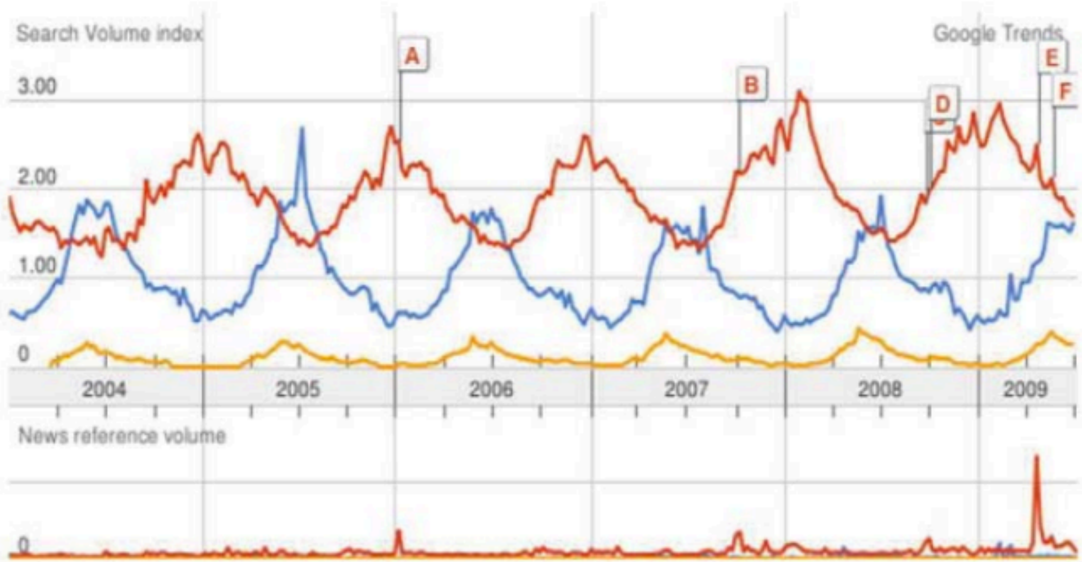
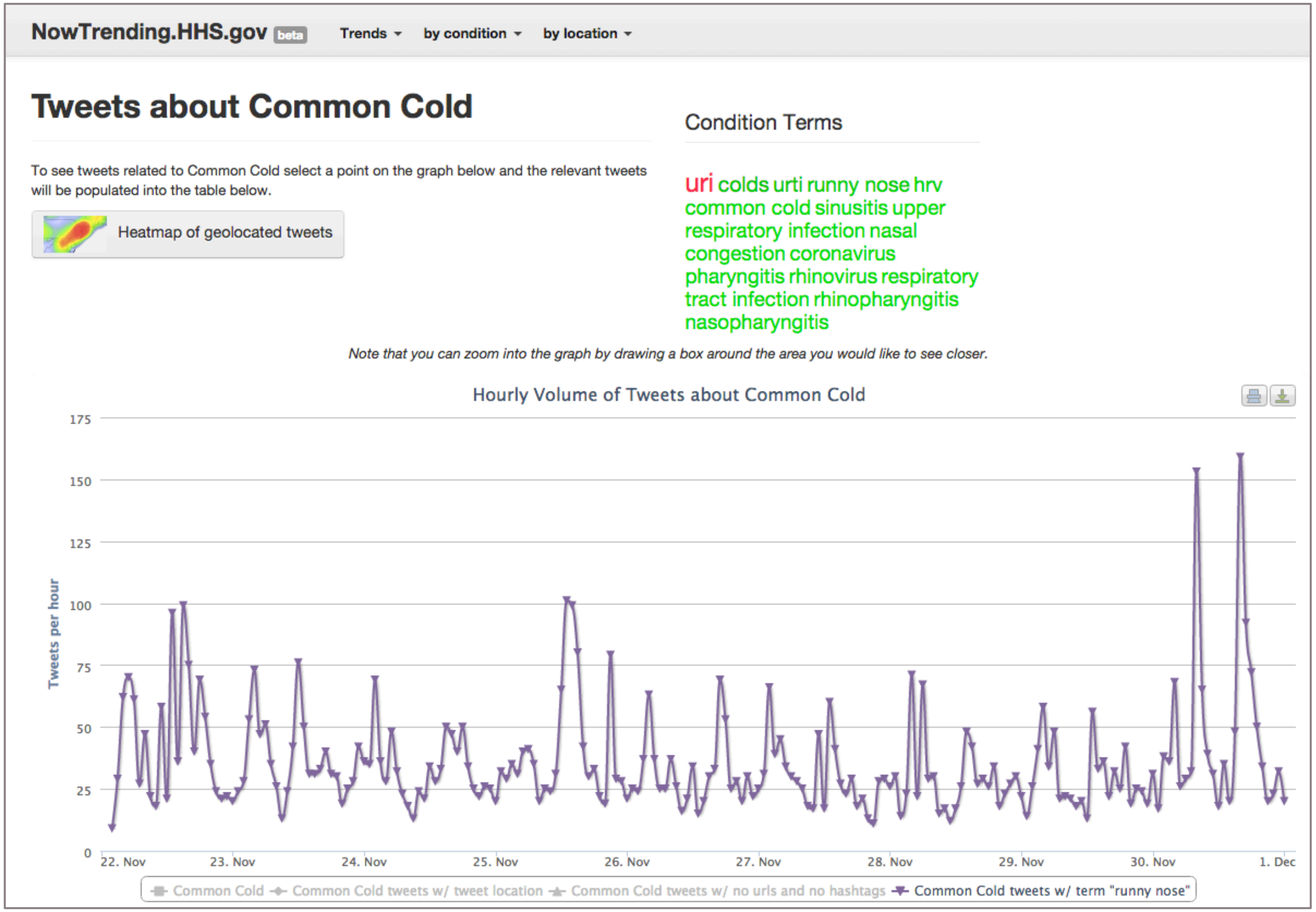


Fig. 2. Google Trends graph depicting tendency over time to search for "Lyme disease", "tick bite", and "cough" (http://www.google.com/trends)[a,b,c]

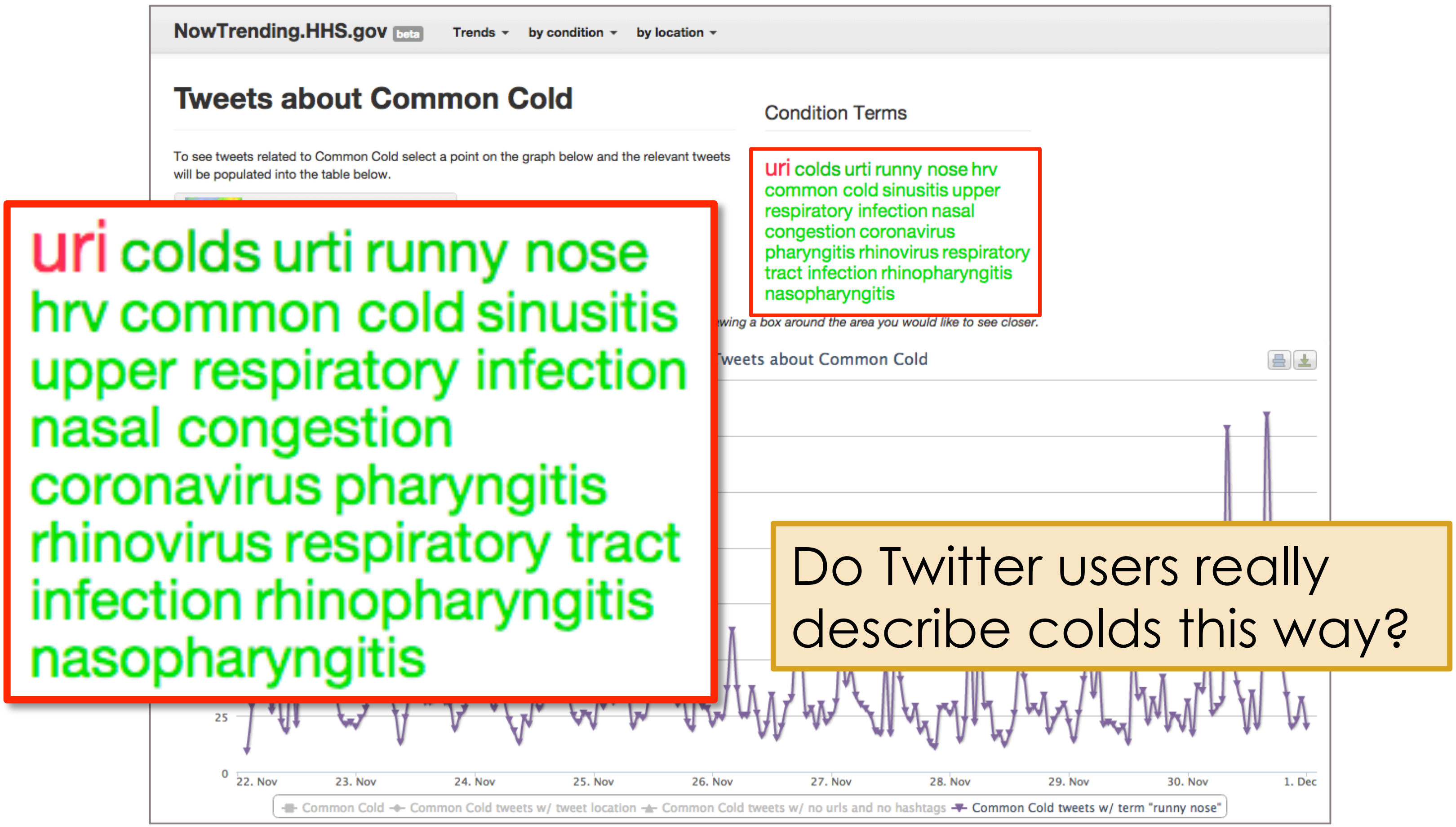
FROM TEXT TO DATA

- Simplest approach: keyword counting



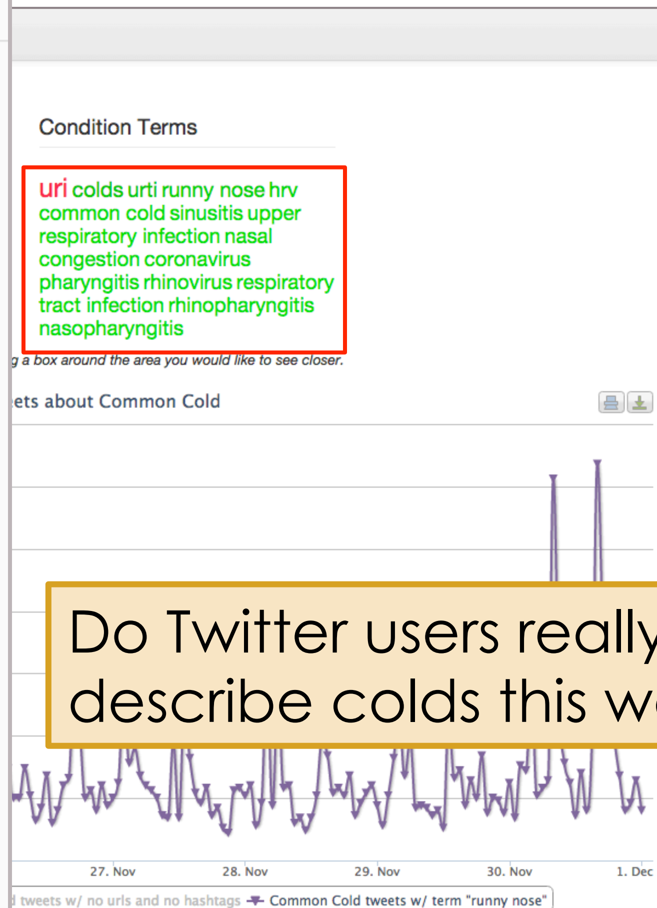
FROM TEXT TO DATA

- Simplest approach: keyword counting



FROM TEXT TO DATA

ord counting



Do Twitter users really describe colds this way?

FROM TEXT TO DATA

- Most common approach: regression

Google

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$

where P is the percentage of ILI physician visits, Q is the ILI-related query fraction, β_0 is the intercept, β_1 is the multiplicative coefficient, and ε is the error term. $logit(P)$ is the natural log of $P/(1-P)$.

FROM TEXT TO DATA

- Most common approach: regression

Google

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$

This is a scalar.
Seems crazy to an NLPer!

where ***P*** is the percentage of ILI ph sician visits, ***Q*** is the ILI-related query
β_1 is the multiplicativ term.
logit(P) is the natural log of *P/(1-P)*.

FROM TEXT TO DATA

- Most common approach: regression

Multivariate models have problems too:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_V x_{iV}$$

Flu rate in week i (given by CDC)

Count of word 2 in week i

2009-2010	2012-2013
flu	christmas
sick	sick
swine	flu
shot	strong
cancer	processing
fever	snow
h1n1	new
#beatcancer	want
better	hard
getting	better
home	body
halloween	best
breast	coughing
cough	festivities
throat	eve

words with highest β values

WE NEED LANGUAGE UNDERSTANDING!

(This is the point of my talk)

TALK OVERVIEW

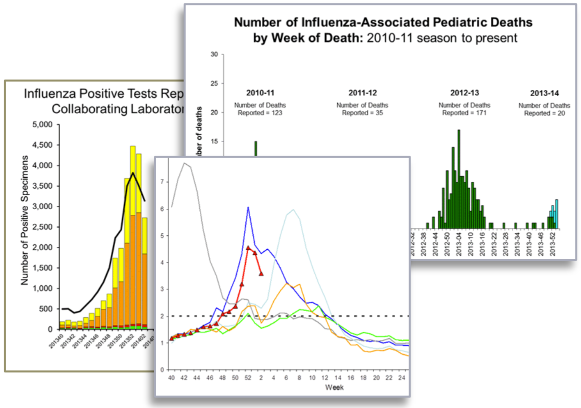
- Three applications for NLP:
  - Influenza surveillance
  - Air pollution monitoring
  - Medical search behavior
- What’s next?

TALK OVERVIEW

- Three applications for NLP:
  - **Influenza surveillance**
  - Air pollution monitoring
  - Medical search behavior
- What's next?

INFLUENZA SURVEILLANCE

- Government flu monitoring is the gold standard
  - But reports have a delay of ~2 weeks
    (or longer, if the government shuts down ☺)



- Text-driven systems can produce estimates **immediately**
  - This talk: let’s use **tweets**
    - advantage: huge, public, free

TWITTER FLU PREDICTION

We only want to count tweets about the flu

- Not about Christmas or breast cancer

We want to include only tweets that are **experiential**

“think I’m coming down with the flu”

vs “tired of hearing about the flu”

TWITTER FLU PREDICTION

We only want to count tweets about the flu

- Not about Christmas or breast cancer

We want to include only tweets that are **experiential**

"think I'm coming down with the flu"

vs "tired of hearing about the flu"

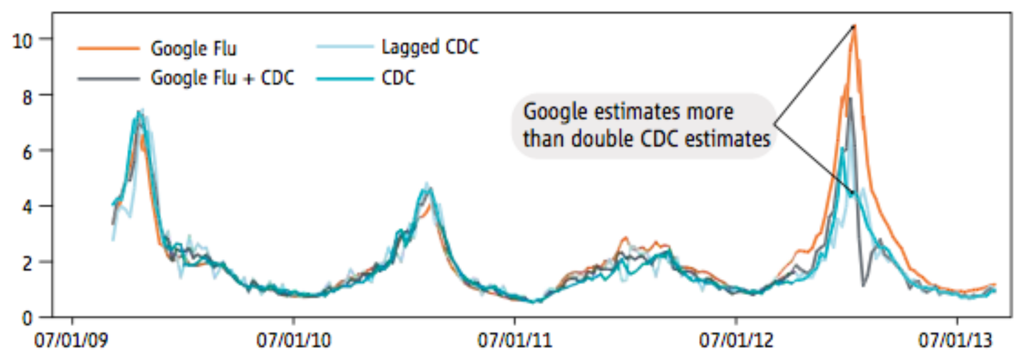
Our labeled data: **Infection** vs **Awareness**

What we're trying to measure

Affected by panic, undue media attention

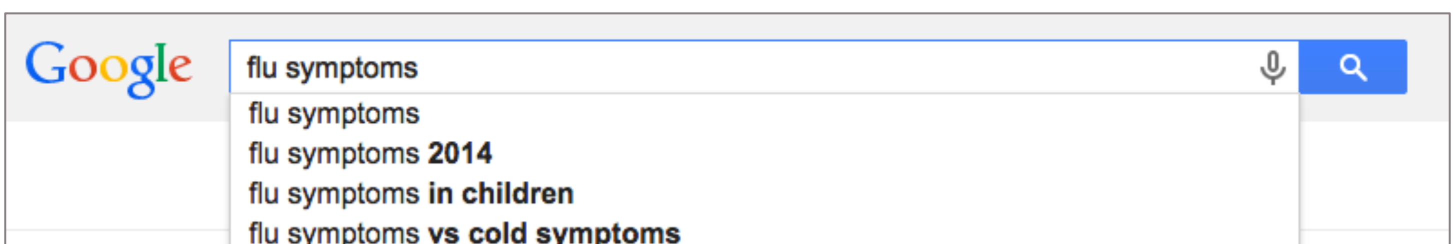
TWITTER FLU PREDICTION

The infection vs awareness distinction matters!



From Lazer et al., *Science*, 2014

Google concluded that media attention was a primary cause of their huge overestimate in 2012-2013



"flu symptoms" – not an experiential query

TWITTER FLU PREDICTION

Our current system uses a cascade of 3 MaxEnt classifiers:

- **about health** vs **not about health**
- **about flu** vs **not about flu**
- **flu infection** vs **flu awareness**

Training data:
11,900 labeled tweets collected through MTurk

Estimated weekly flu rate:

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

TWITTER FLU PREDICTION

Features:

- Stylometry
  - Retweets, user mentions, URLs, emoticons
- 8 manually created word classes

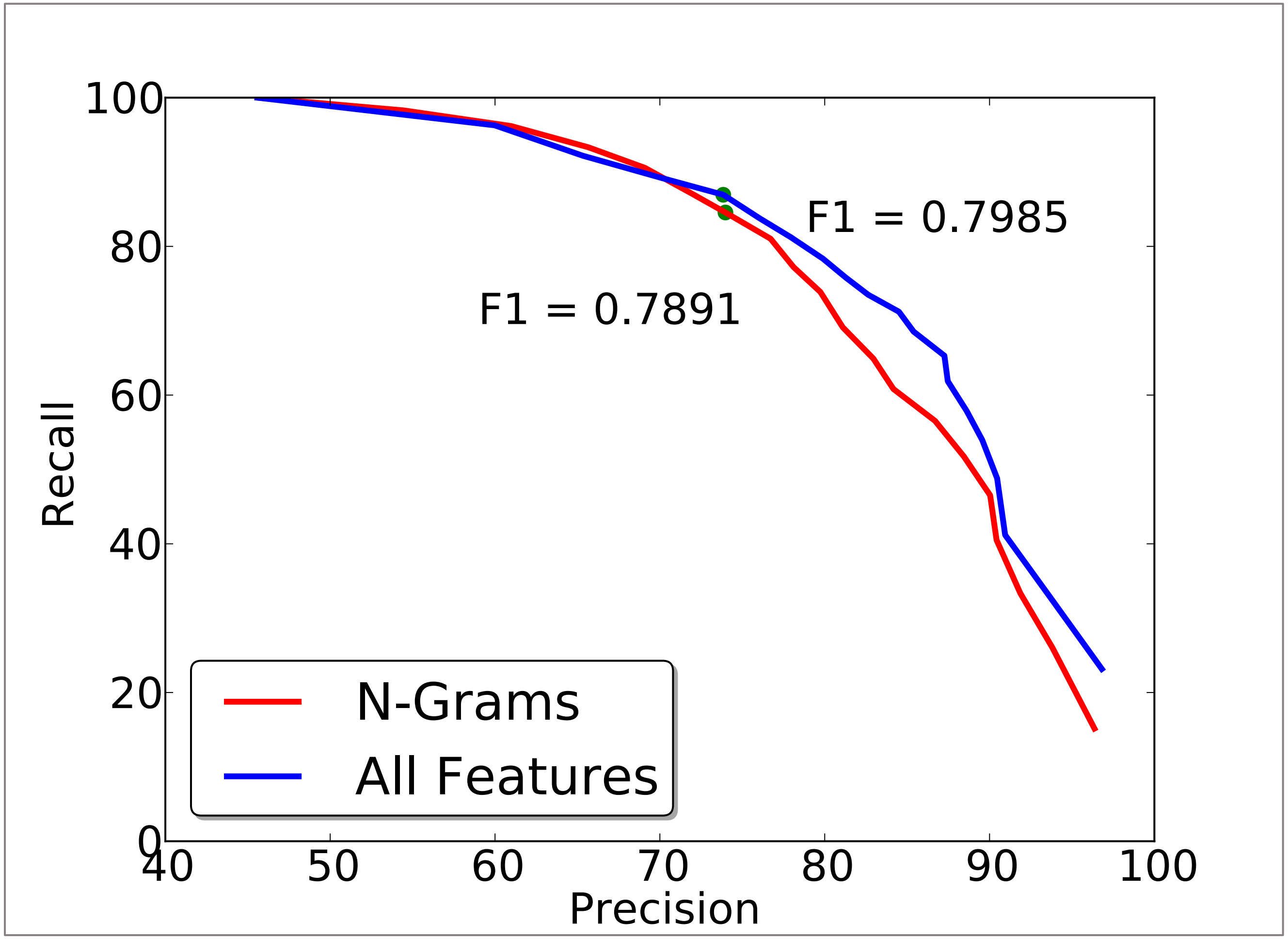
Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
Disease	bird, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
Treatment/ Prevention	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
…	…

TWITTER FLU PREDICTION

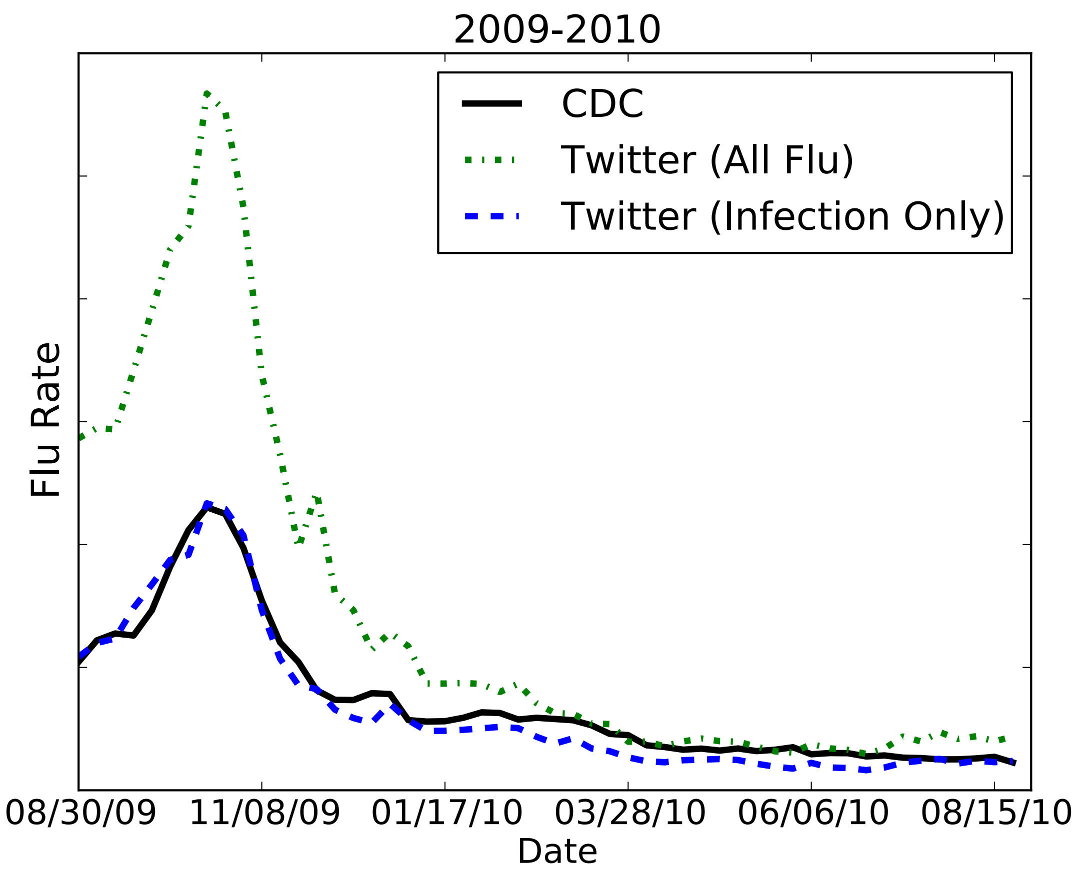
Features:

- Part of speech templates
  - (subject,verb,object) tuples
    - always a good feature, IMO
  - numeric references
    - "100 more cases of swine flu"
  - whether "flu" is a noun or adjective
    - "tired of the flu" vs "tired of the flu hype"
  - whether "flu" is the subject or object
    - "I have the flu" vs "the flu is going around"
  - … and others

TWITTER FLU PREDICTION

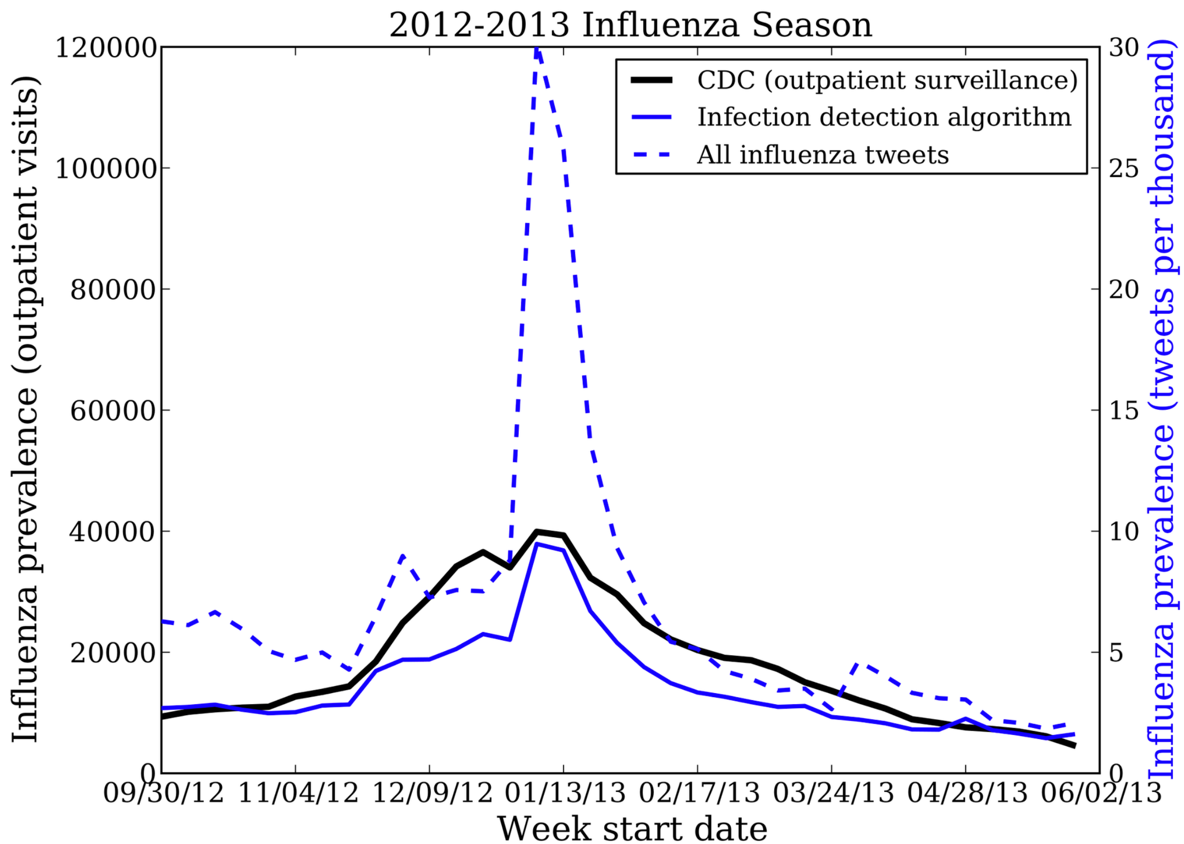


TWITTER FLU PREDICTION



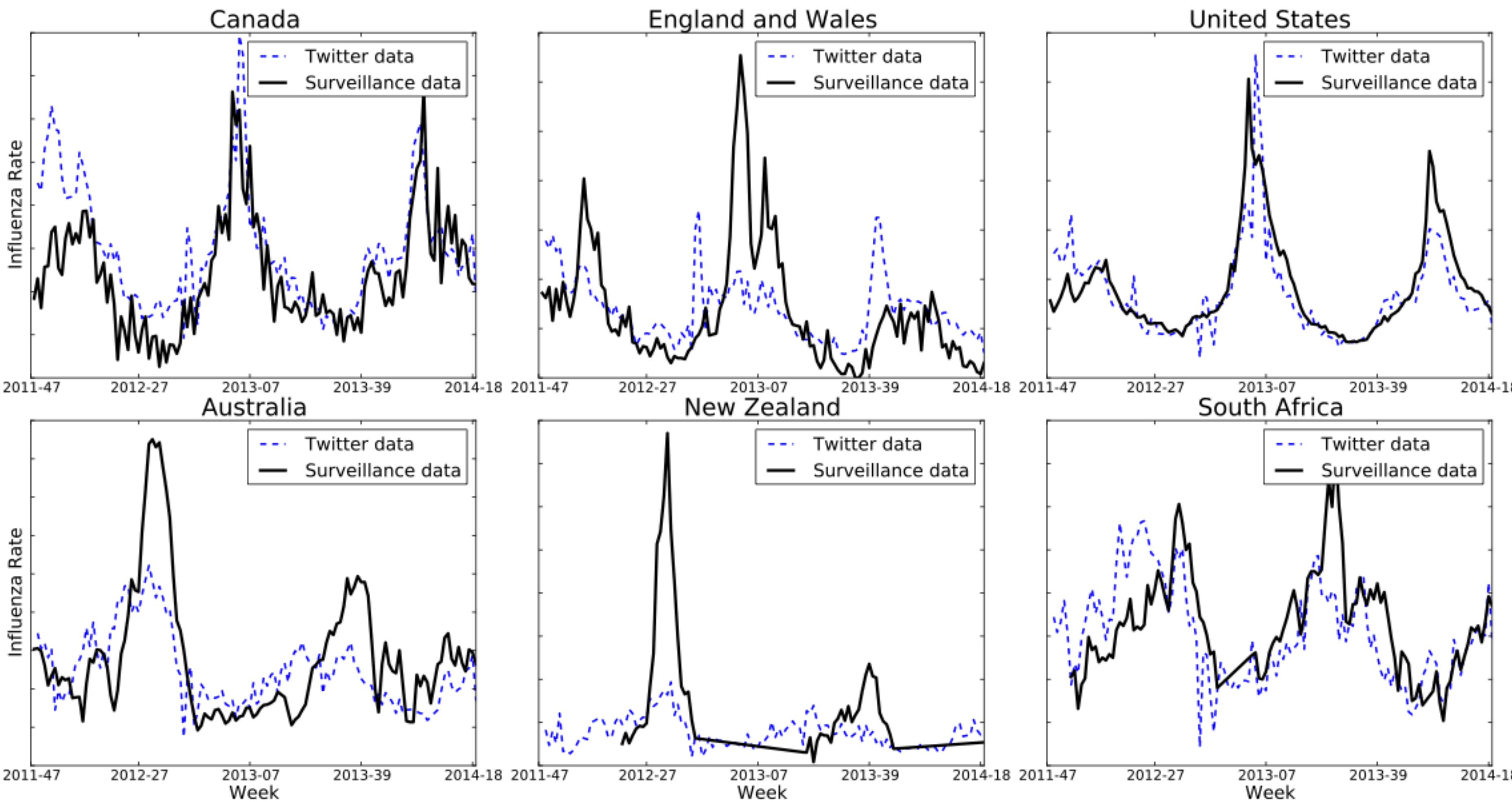
Correlation with classifier: **0.990**
Correlation with keywords: **0.977**

TWITTER FLU PREDICTION



Correlation with classifier: **0.93**
Correlation with keywords: **0.75**

TWITTER FLU PREDICTION



TALK OVERVIEW

- Three applications for NLP:
  - Influenza surveillance
  - **Air pollution monitoring**
  - Medical search behavior
- What’s next?

AIR POLLUTION IN CHINA

What do people have to say about air quality on **Sina Weibo?**

- Can social media detect pollution levels?
- Can we learn about health effects and behavioral response?

AIR POLLUTION IN CHINA

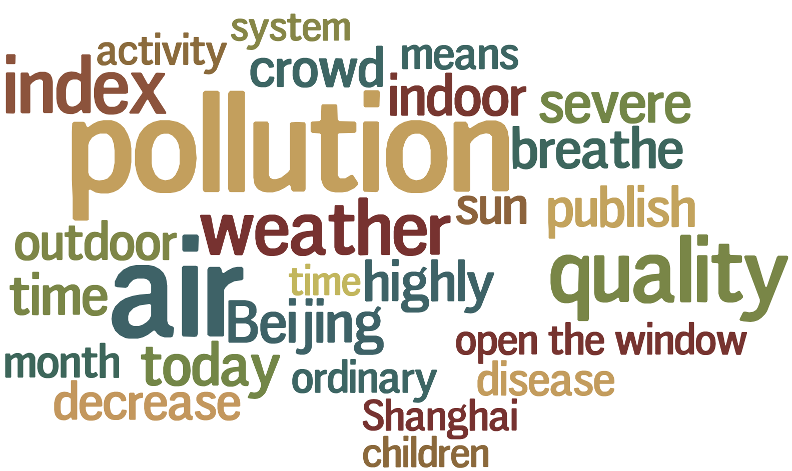
What do people have
air quality on **Sina Wei**

- Can social media de
  pollution levels?
- Can we learn about
  health effects and
  behavioral response

AIR POLLUTION IN CHINA

Data pipeline:

1. Started with 93 million crawled Weibo messages
2. Filtered to 1 million messages with health keywords
3. Ran LDA with 100 topics
4. Analyzed messages with the air pollution topic



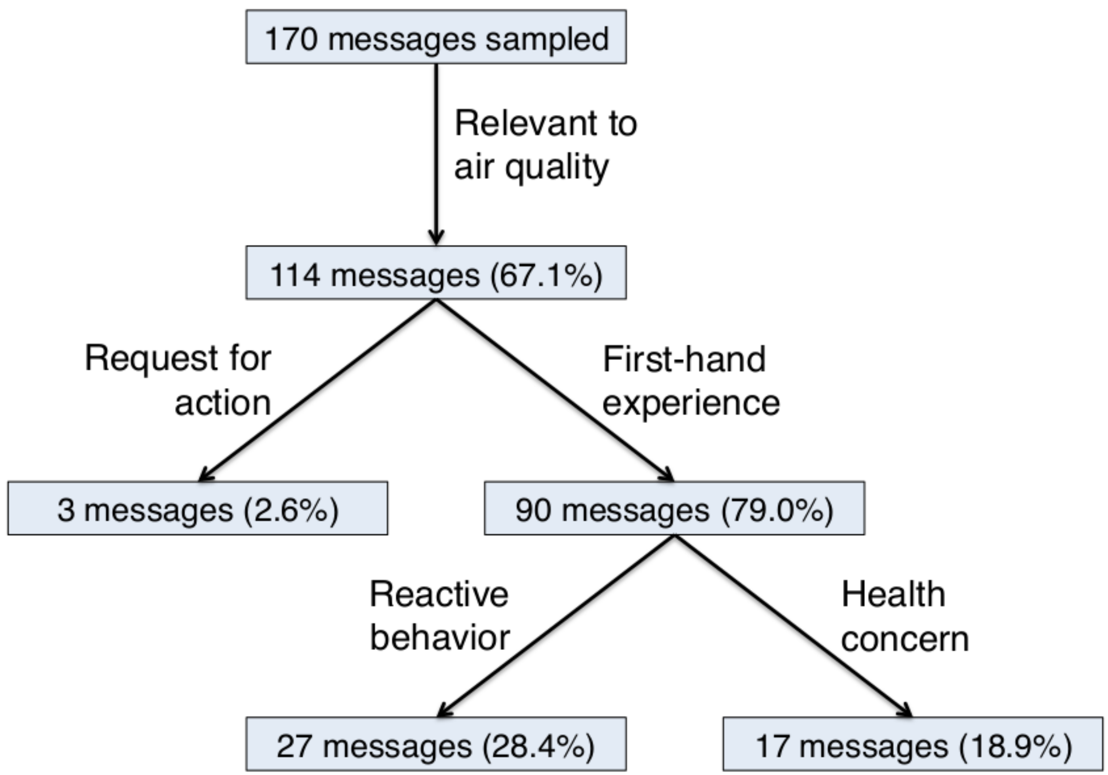
AIR POLLUTION IN CHINA

Validation: We compared the volume of messages with this topic to government-provided pollution rates

Correlation across 74 cities: **.583**

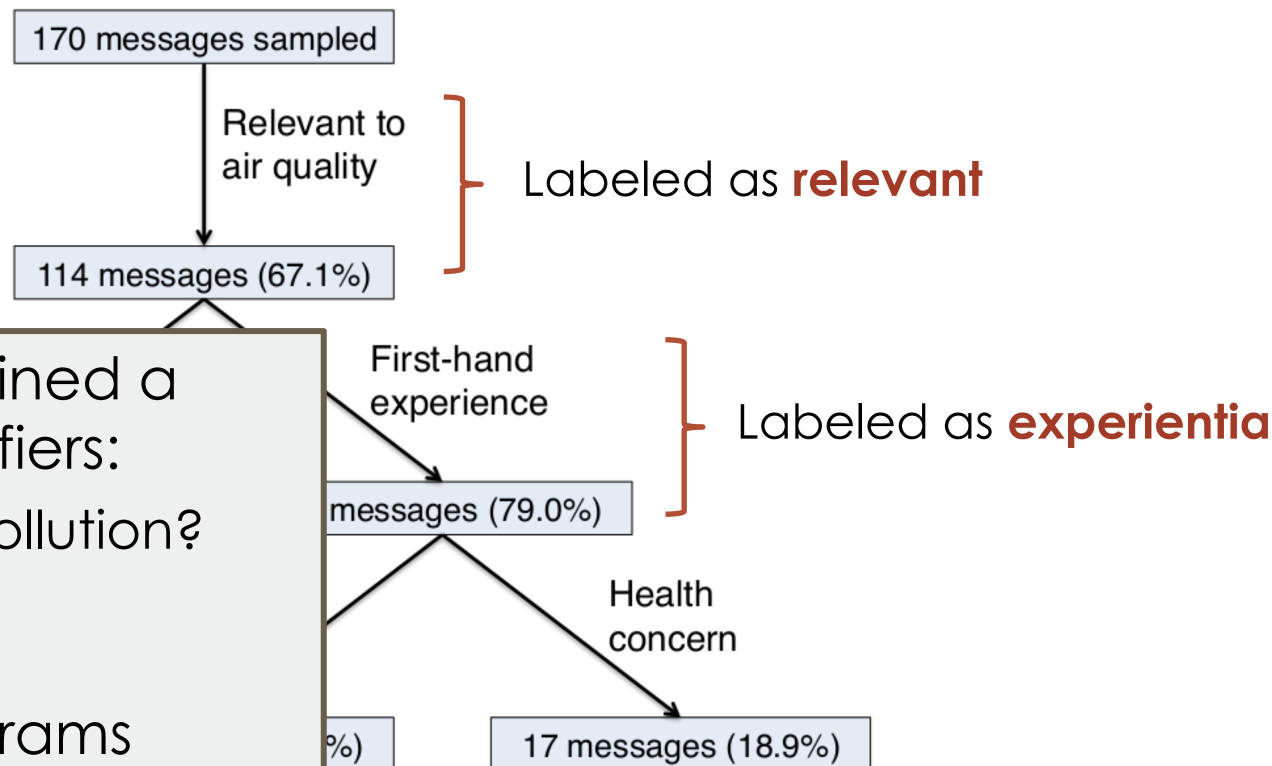
AIR POLLUTION IN CHINA

We then annotated a small sample of topical messages with detailed codes



AIR POLLUTION IN CHINA

We then annotated a small sample of topical messages with detailed codes



As with flu, we trained a cascade of classifiers:

1. **Relevant** to air pollution?
2. **Experiential**?

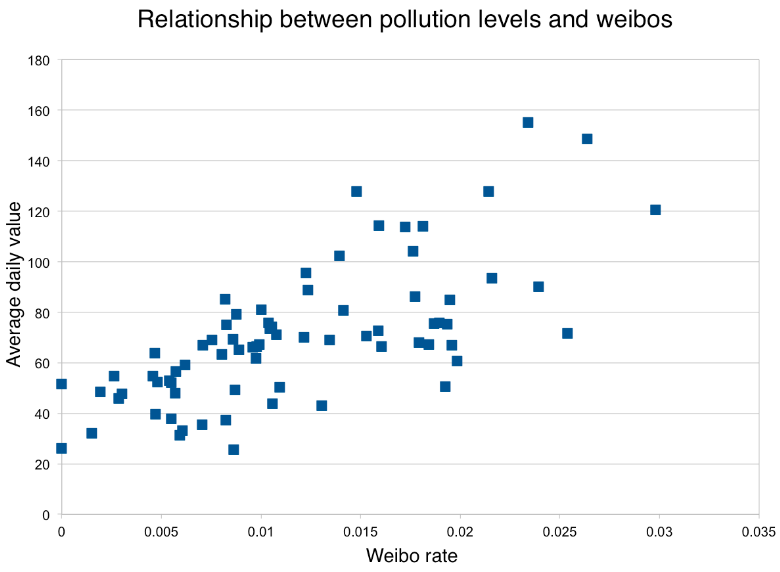
Features: 1, 2, 3-grams

AIR POLLUTION IN CHINA

Validation: We compared the volume of messages with this topic to government-provided pollution rates

Correlation across 74 cities: **.583**

with experiential classifiers: **.718**



TALK OVERVIEW

- Three applications for NLP:
  - Influenza surveillance
  - Air pollution monitoring
  - **Medical search behavior**
- What's next?

MEDICAL SEARCH

Scientific questions:

- **What information** do patients need?
  - and **when?**
- How do people use the web to make **decisions?**
  - e.g. choice of treatment, choice of doctor

Engineering goals:

- How can we make **search engines better** to support these goals?

MEDICAL SEARCH

What do people search when confronted with a **major illness**?

Our project focused on breast and prostate cancer

- I'll just talk about the first today

Approach: large scale analysis of anonymized logs

- Step 1: retrieve search histories about breast cancer

SEARCH AND BREAST CANCER

Starting point: filter for search histories containing "breast cancer" ≥ 3 times

SEARCH AND BREAST CANCER

Starting point: filter for search histories containing "breast cancer" ≥ 3 times

- But people search for lots of reasons…

81° Don't worry about it

the ONION® America's Finest News Source

VIDEO · POLITICS · SPORTS · SCIENCE/TECH · LOCAL · ENTERTAINMENT

Internet Opens Up Whole New World Of Illness For Local Hypochondriac

NEWS · Local · Old Internet · Health · Healthcare · Internet · ISSUE 36·16 · May 3, 2000

Share on Facebook 285 Share on Twitter 165 g+ 0

MERIDEN, CT–All her life, Janet Hartley has suffered from a host of ill-defined viruses and inexplicable aches and pains, diagnosing herself with everything from diabetes to cancer. But ever since discovering such online medical resources as WebMD, drkoop.com, and Yahoo! Health, the 41-year-old hypochondriac has had a whole new world of imaginary illnesses opened up to her.

Janet Hartley learns more about her suspected case of arteriovascular malformation on Yahoo! Health.

"The Internet has really revolutionized my ability to keep on top of my medical problems," said Hartley, speaking from her bed. "For instance, I used to think my headaches were just really bad migraines. But then last week, while searching Mt. Sinai Hospital's online medical database, I learned about something much more serious called cranial AVM, or arteriovascular malformation, which, along with headache pain, may also result in dizziness, loss of concentration, and

EXPERIENTIAL SEARCH PREDICTION

As before, we need to identify experiential search

Classifier:

- Annotated 480 partial histories
  - filtered for relevant queries
- Trained with boosted decision trees

EXPERIENTIAL SEARCH PREDICTION

Features:

- Ontology of terms
  - each category is a feature

Category			
Level 1	**Level 2**	**Level 3**	**Terms**
Cosmetic	Post-Surgery	Post-Surgery	{cosmetic,plastic} {surgery,surgeon}, prosthesis, prosthetic(s), implant(s), reconstruction
Cosmetic	Hair Loss	Hair Loss	wig(s), head {scarf,scarves,covering(s)}, hair (re)grow(th)
Description	Type	Cancer Type	DCIS, LCIS, IDC, ILC, lobular, ductal, in situ, metaplastic, mucinous, inflammatory
Description	Staging/Grading	Staging/Grading	what stage, stages, staging, what grade, grades, grading, differentiated
Description	Staging/Grading	Stage	pre( )cancer, early stage, stage {[0–4],zero–four,[I–IV]}({a,b,c})
Description	Staging/Grading	Grade	grade {[1–3],[I–III]}, {low,moderate,intermediate,high} grade
Diagnosis	Diagnosis	Diagnosis	diagnosis, diagnosed
Diagnosis	Diagnostics	Biopsy	biopsy, biopsies
Diagnosis	Screening	Mammagraphy	mammogram(s), mammography
Diagnosis	Screening	Ultrasound	ultrasound(s)
Lifestyle	Lifestyle	Diet	diet(s), eat(ing), food(s), vitamin(s), supplements, nutrition, protein, recipe(s), cookbook
Lifestyle	Lifestyle	Fitness	fitness, exercise(s), yoga
Professional	Healthcare	Provider	clinic(s), hospital(s), cancer center(s)
Professional	Healthcare	Doctor	doctor(s), physician(s)
Professional	Healthcare	Oncologist	oncologist(s)
Treatment	Treatment	Treatment	treatment(s), medication(s), meds
Treatment	Treatment	Side Effects	side effect(s)
Treatment	Chemotherapy	Chemotherapy	chemotherapy, chemo, cemo, kemo
Treatment	Chemotherapy	Side Effects	hair loss, hair fall(ing), {lose,losing} {my,your} hair

EXPERIENTIAL SEARCH PREDICTION

Features:

- Language features:
  - First/second person pronouns (including possessives)
  - Experiential phrases (e.g. "i have", "i was diagnosed")
  - Starts with a question word
- Volume and temporal patterns:
  - % of queries/sessions containing ontology terms
  - Length of cancer-related sessions
  - Time between cancer-related sessions
  - Ordering of categories searched
  - … and a lot more

EXPERIENTIAL SEARCH PREDICTION

External validation

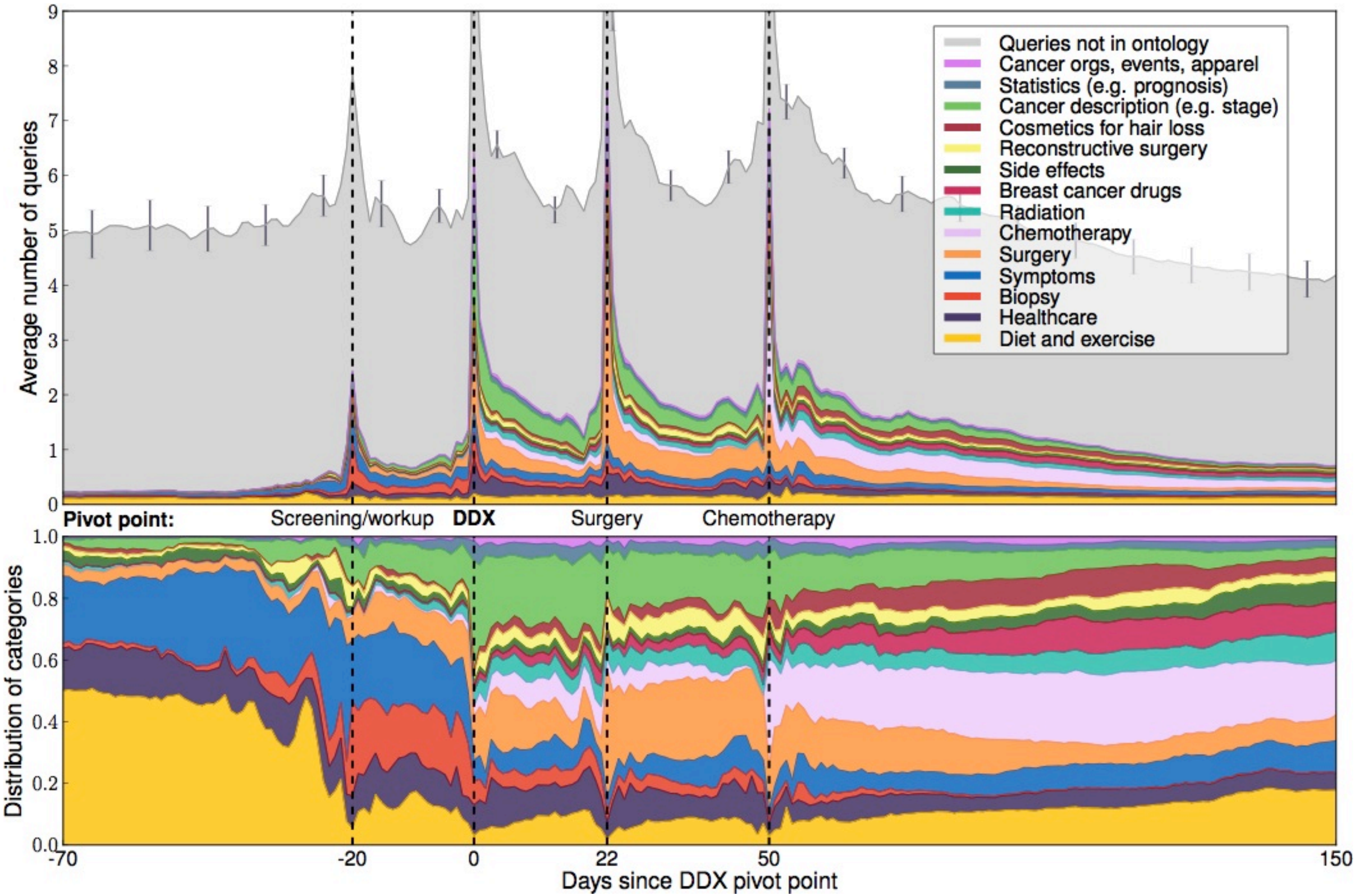
- Geography: correlation with state rates
  - Keyword filter: **.036** (i.e. "breast cancer" 3 times)
  - With classifier: **.348** (a tenfold increase!)
- Gender (100 times more common in women):
  - Keyword filter: **70.0%** women
  - With classifier: **88.9%** women
- Age (6 times more common in elderly):
  - Keyword filter: **5.4%** aged 65+
  - With classifier: **22.2%** aged 65+

SEARCH TIMELINE ALIGNMENT

We also built classifiers to identify the inferred day of diagnosis **(DDX)**

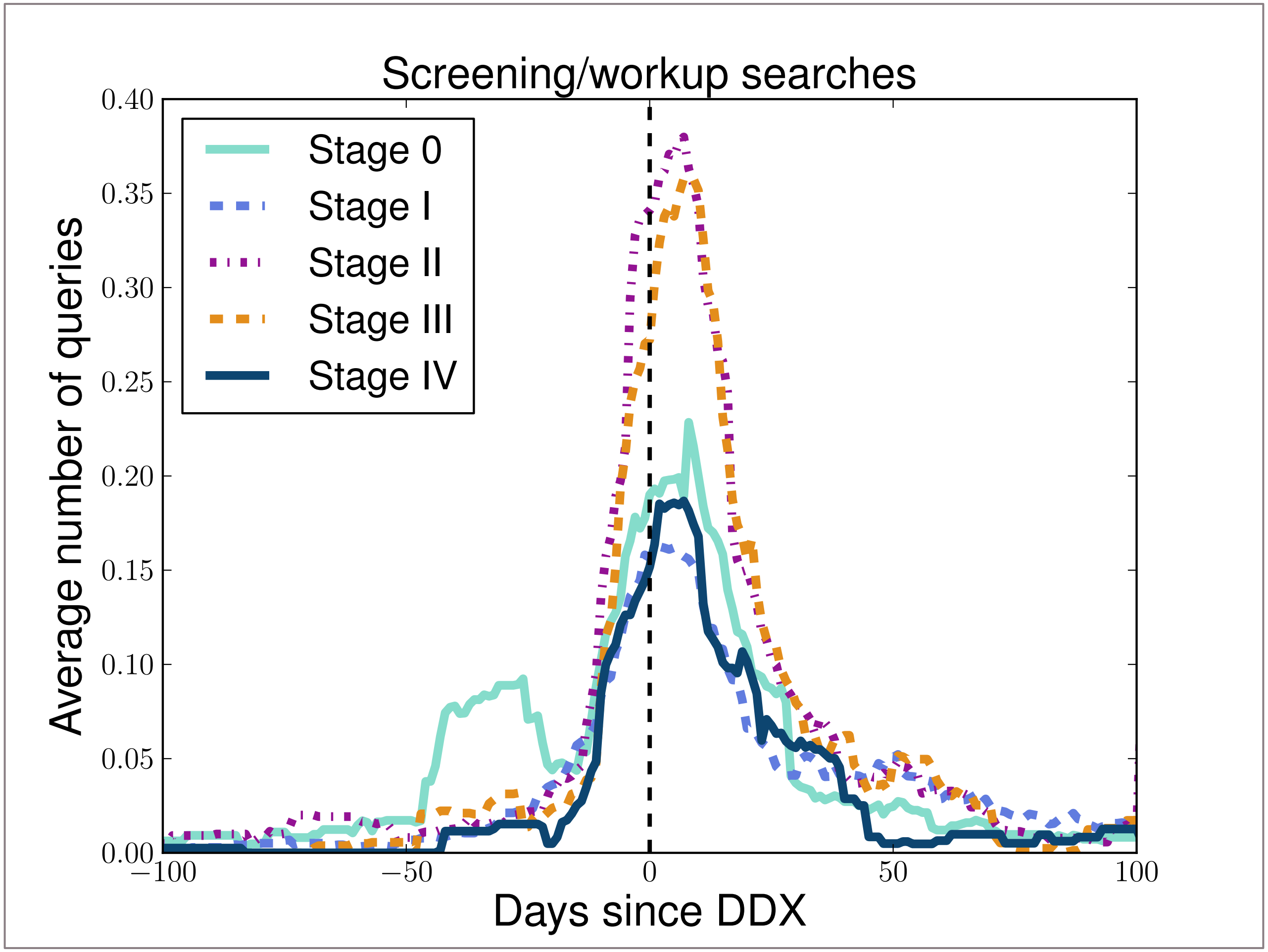
- I'll skip the details today

This gives a common point for **aligning** the 1700 histories tagged by the classifier

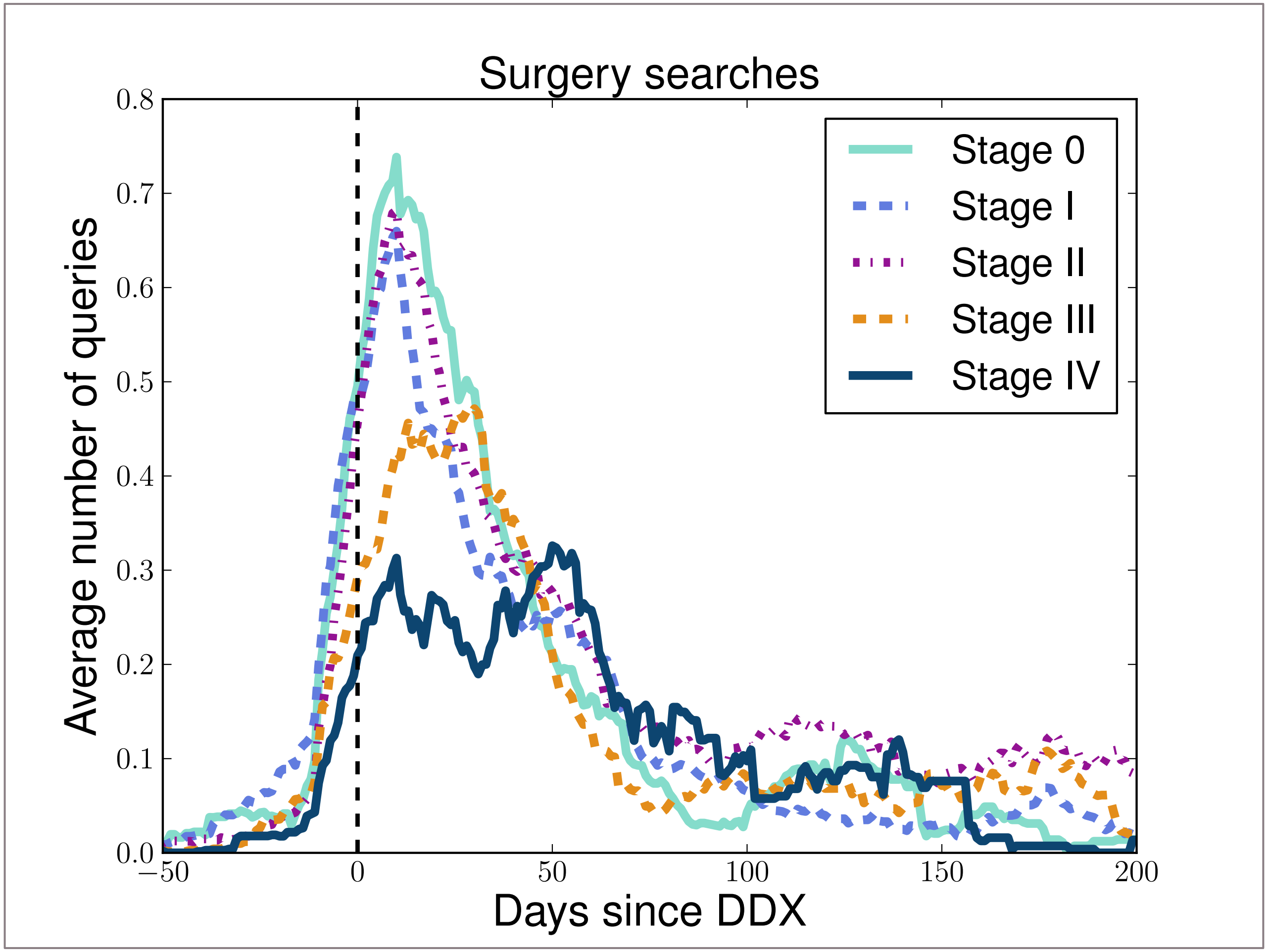


Queries not in ontology
Cancer orgs, events, apparel
Statistics (e.g. prognosis)
Cancer description (e.g. stage)
Cosmetics for hair loss
Reconstructive surgery
Side effects
Breast cancer drugs
Radiation
Chemotherapy
Surgery
Symptoms
Biopsy
Healthcare
Diet and exercise
Average number of queries
Pivot point:
Screening/workup
DDX
Surgery
Chemotherapy
Distribution of categories
Days since DDX pivot point
-70
-20
0
22
50
150

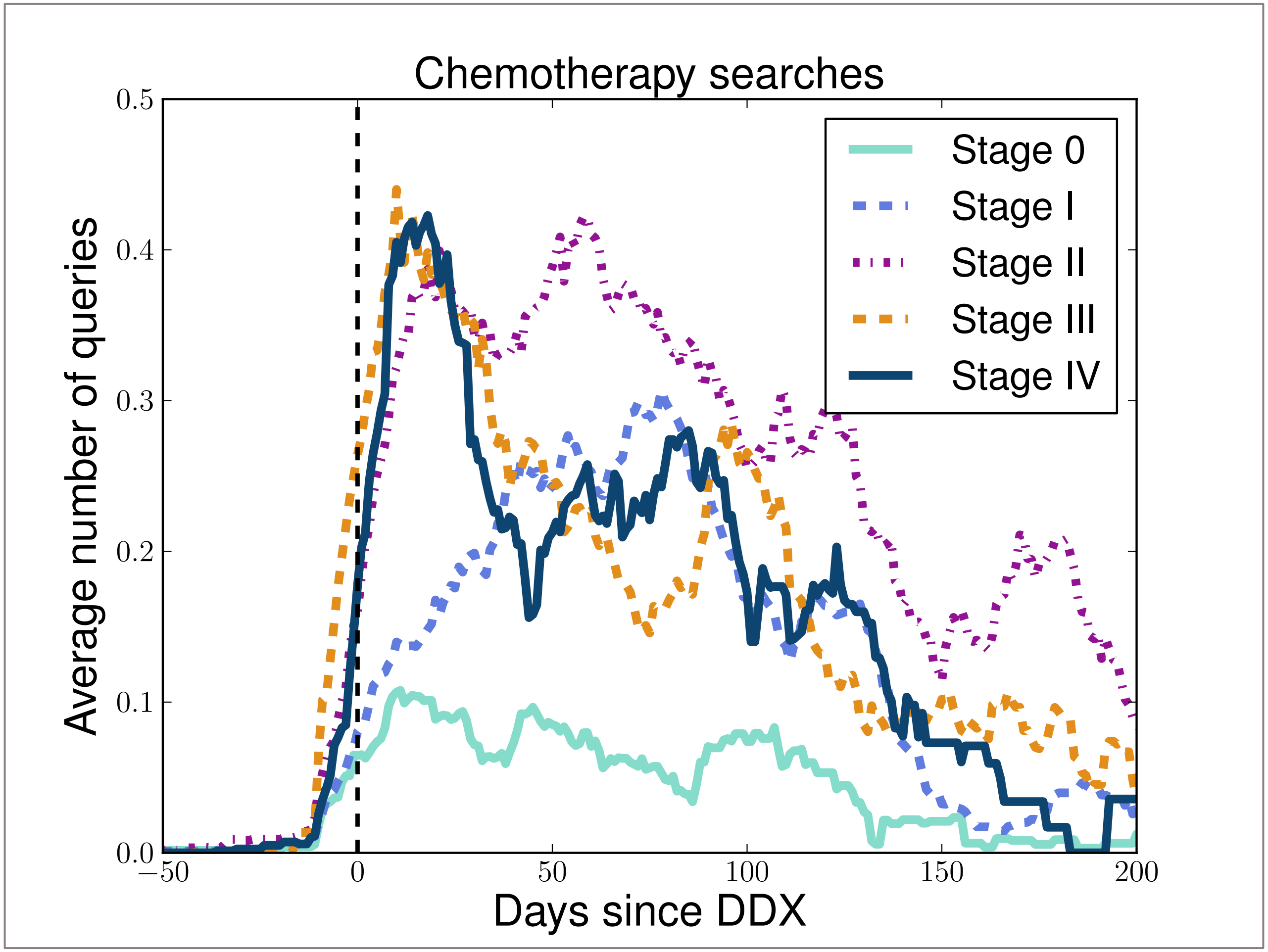
BREAKDOWN BY STAGE



BREAKDOWN BY STAGE



BREAKDOWN BY STAGE



WHAT'S NEXT?

(And what role will NLP play?)

WE STILL NEED BETTER NLP

Another application: medical mistakes in Twitter

- important public health issue
- not well understood

Our *qualitative* study:

- just need to find some relevant tweets to examine
- so we came up with reasonable search terms…

WE STILL NEED BETTER NLP

Surprisingly many false positives…

- I hope the doctor was wrong and a miracle happens
- The antibiotics were just to prevent surgery infection.
- I think the hospital gave me the wrong kid lol
- I hate going to the stupid doctor
- on my way to the hospital fucked up my knee
- I'm just drowsy... I bought the wrong meds.
- You must be on some wrong pills bro

WE STILL NEED BETTER NLP

Why Big Data Missed the Early Warning Signs of Ebola

Hint: Ils ne parlent pas le français.

BY KALEV LEETARU | SEPTEMBER 26, 2014

WE STILL NEED BETTER NLP

It's clear that experiential classification is important. This requires NLP. But there's much more to do!

Interesting problems for language understanding:

- mining **attitudes**, **perceptions**, **and behaviors**

THANKS TO MANY PEOPLE

- Mark Dredze (advisor)
- Microsoft Research (funding)

Flu:

- David Broniatowski
- Alex Lamb
- Nicholas Generous

Pollution:

- Shiliang Wang
- Angie Chen
- Brian Schwartz

Medical search:

- Eric Horvitz
- Ryen White
- Sara Javid
- Janice Tsai

Patient safety:

- Sarah Bell
- Atul Nakhasi
- Ralph Passarella
- Peter Pronovost