Understanding Public Health Using Twitter

Aron Culotta
Assistant Professor
Department of Computer Science
Illinois Institute of Technology
Chicago, IL
aculotta@iit.edu

Outline

1. Tracking influenza-like illness rate
2. Tracking alcohol sales
3. Tracking community health
4. Inferring the origin of social media messages

Important Public Health Questions

- How many people have influenza/diabetes/H1N1/...?
  - Right now
  - Where are they?
- What influences incidence of influenza/diabetes/H1N1/...?
  - Social connections
  - Environmental factors
  - Personal behaviors

How these questions are usually answered

- How many people have influenza/diabetes/H1N1/...?
  - Right now
  - Where are they?

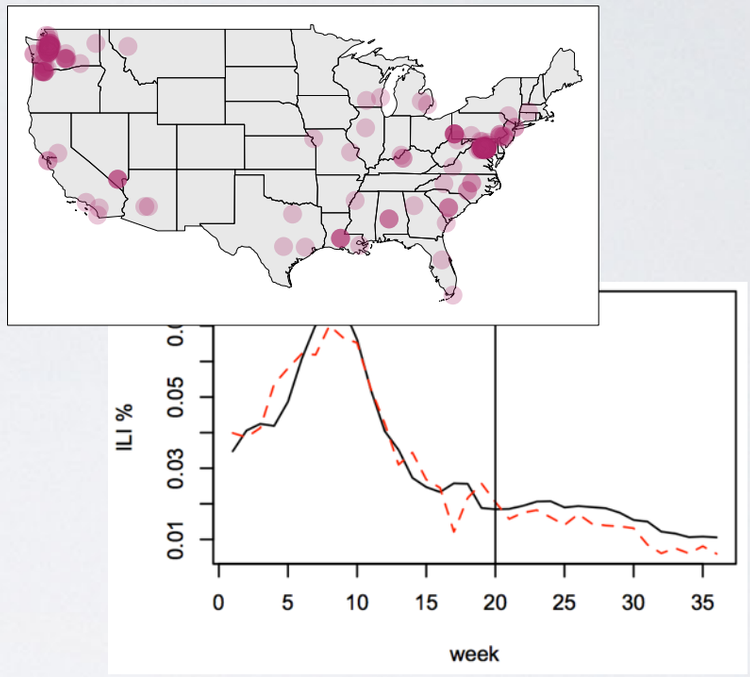
Surveys of hospitals, pharmacies

- What influences incidence of influenza/diabetes/H1N1/...?
  - Social connections
  - Environmental factors
  - Personal behaviors

Controlled experiments
Observational Studies

Can we do this better/faster/cheaper?

Prediction from Social Media



- People have analyzed online chatter to try to predict:
  - How many people will see the new Star Wars movie?
  - Who will win the next presidential election?
  - What will the S&P's value be tomorrow?

Social Media Analysis

...Central to this effort was PhRMA president, CEO and top lobbyist Billy Tauzin, a longtime Democratic member of Congress who switched party affiliations after Republicans gained control of Congress in 1994. By switching parties Tauzin was able to maintain his influence and even rose to be Chairman of the House Committee on Energy & Commerce. Tauzin became the poster child of Washington's mercenary culture. He crafted a bill to provide prescription drug access to Medicare recipients, one that provided major concessions to the pharmaceutical industry. Medicare would not be able to negotiate for lower prescription drug costs and reimportation of drugs from first world countries would not be allowed. A few months after the bill passed, Tauzin announced that he was retiring from Congress and would be taking a job helming PhRMA for a salary of $2 million.

Tauzin's job change became fodder for a campaign ad that then presidential candidate Barack Obama ran in the spring of 2008 simply titled "Billy." It featured the candidate, sleeves rolled up, talking to a salon of gasping Americans about the ways of Washington. "The pharmaceutical industry wrote into the prescription drug plan that Medicare could not negotiate with drug companies. And you know what, the chairman of the committee, who pushed the law through, went to work for the pharmaceutical industry making $2 million a year." The screen fades to black to inform the viewer that, "Barack Obama is the only candidate who refuses Washington lobbyist money," while the candidate continues his lecture, "Imagine that. That's an example of the same old game playing in Washington. You know, I don't want to learn how to play the game better, I want to put an end to the game playing."

Aiding PhRMA in their outreach to Congress would be a squadron of lobbyists to push their health care reform priorities. Over the course of 2009, the drug industry trade group spent over $28 million on in house and hired lobbyists. Aside from PhRMA's massive in-house lobbying operation, the trade group hired 48 outside lobbying firms. The total number of lobbyists working for PhRMA in 2009 reached 165. Some 137 of those 165 lobbyists representing PhRMA were former employees of either the legislative or executive branches. Of these dozens were former congressional staffers including two former chiefs of staff to Max Baucus....

Social Media Analysis

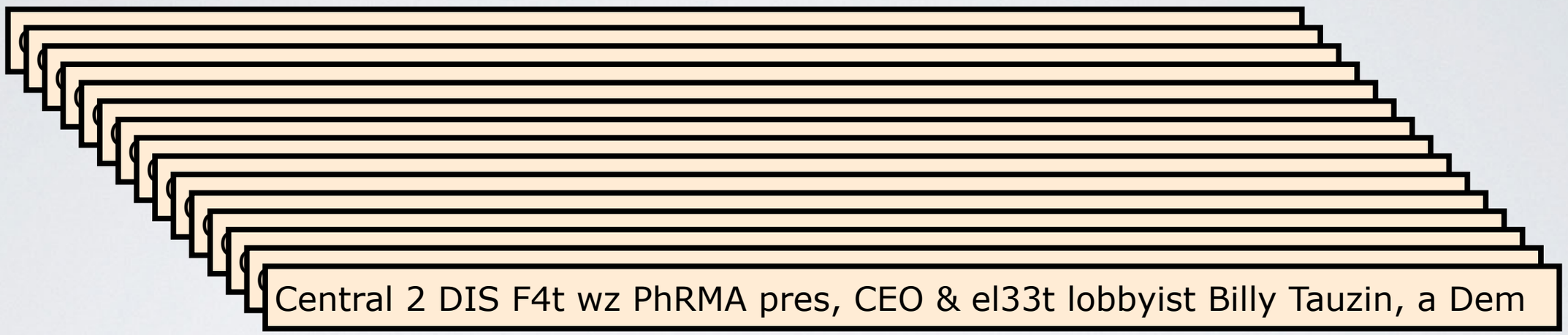
Central to this effort was PhRMA president, CEO and top lobbyist Billy Tauzin, a longtime Democratic member of Congress

Social Media Analysis

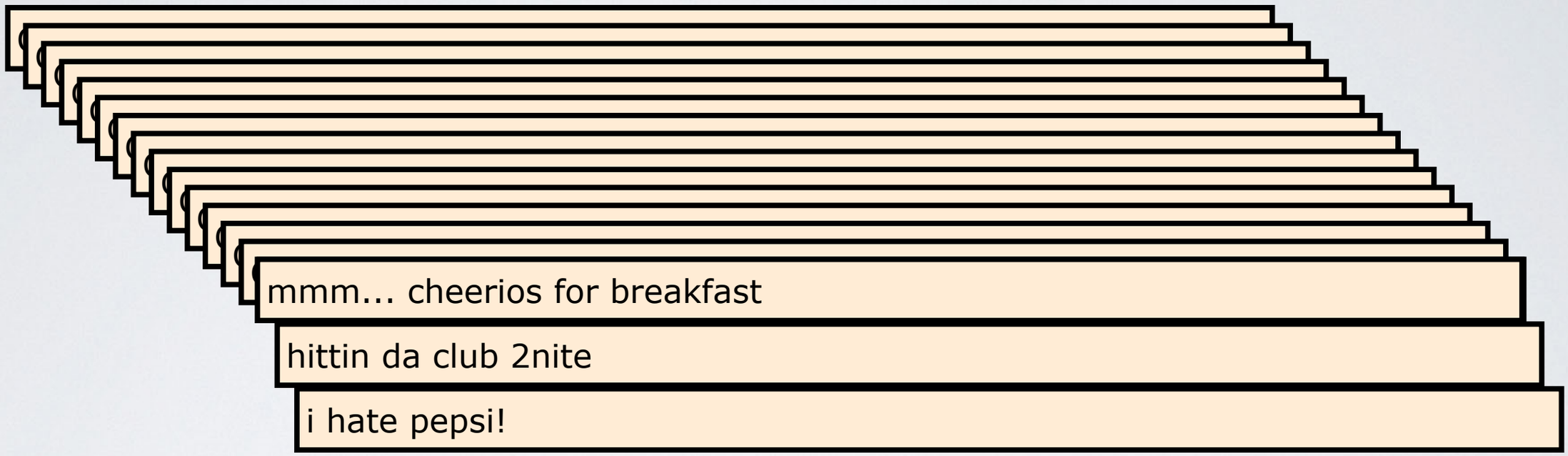
Central 2 DIS F4t wz PhRMA pres, CEO & el33t lobbyist Billy Tauzin, a Dem

English -> SMS translator: http://transl8it.com

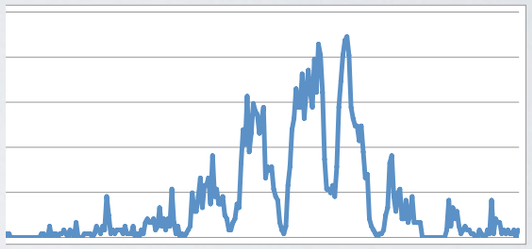
Social Media Analysis



Social Media Analysis



Nowcasting from Social Media



Validation Time-Series

Regression

e.g., Gallup poll

Document Sampling

Document Classification

Aggregation

e.g. All Twitter messages mentioning President Obama in June 2011

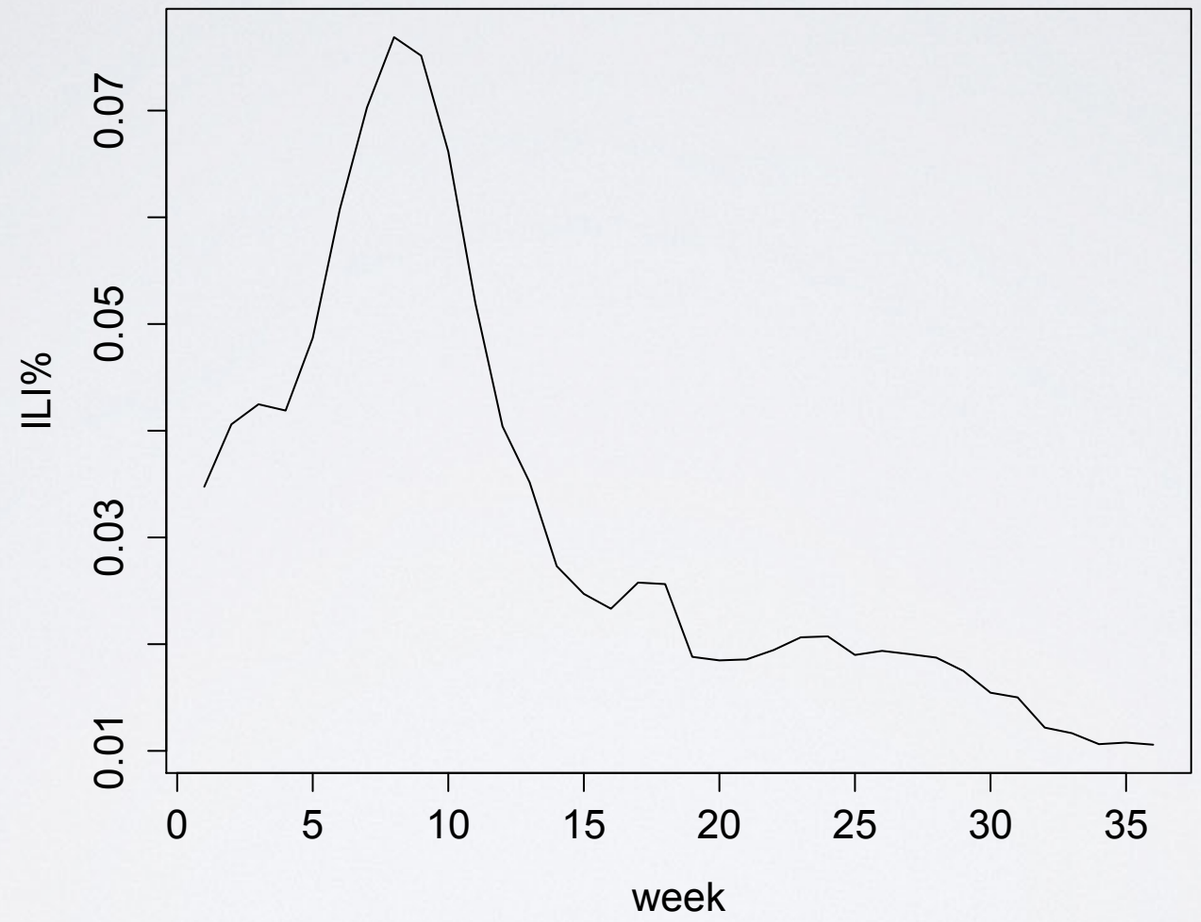
e.g., positive/negative sentiment

e.g., 52% positive

Can we estimate the national flu rate from Twitter?

Nowcasting Rate of Influenza-like Illness (ILI)

[Culotta '10; Culotta '13]



CDC collects weekly data by sampling hospitals

1-2 week reporting delay

Nowcasting Rate of Influenza-like Illness (ILI)

Nowcasting Rate of Influenza-like Illness (ILI)

Sampling: 570M tweets from 9/09-5/10
10% pseudo-uniform sample from Twitter API

Nowcasting Rate of Influenza-like Illness (ILI)

Sampling: 570M tweets from 9/09-5/10
10% pseudo-uniform sample from Twitter API

Classification: simple keyword matching
Does tweet contain flu, cough, headache, sore throat?

Nowcasting Rate of Influenza-like Illness (ILI)

Sampling: 570M tweets from 9/09-5/10
10% pseudo-uniform sample from Twitter API

Classification: simple keyword matching
Does tweet contain flu, cough, headache, sore throat?

Aggregation: fraction of tweets matching keywords

$$f(w, D) = \frac{|D_w|}{|D|}$$

D_w : tweets matching word w
D : all tweets this week

Nowcasting Rate of Influenza-like Illness (ILI)

Sampling: 570M tweets from 9/09-5/10
10% pseudo-uniform sample from Twitter API

Classification: simple keyword matching
Does tweet contain flu, cough, headache, sore throat?

Aggregation: fraction of tweets matching keywords

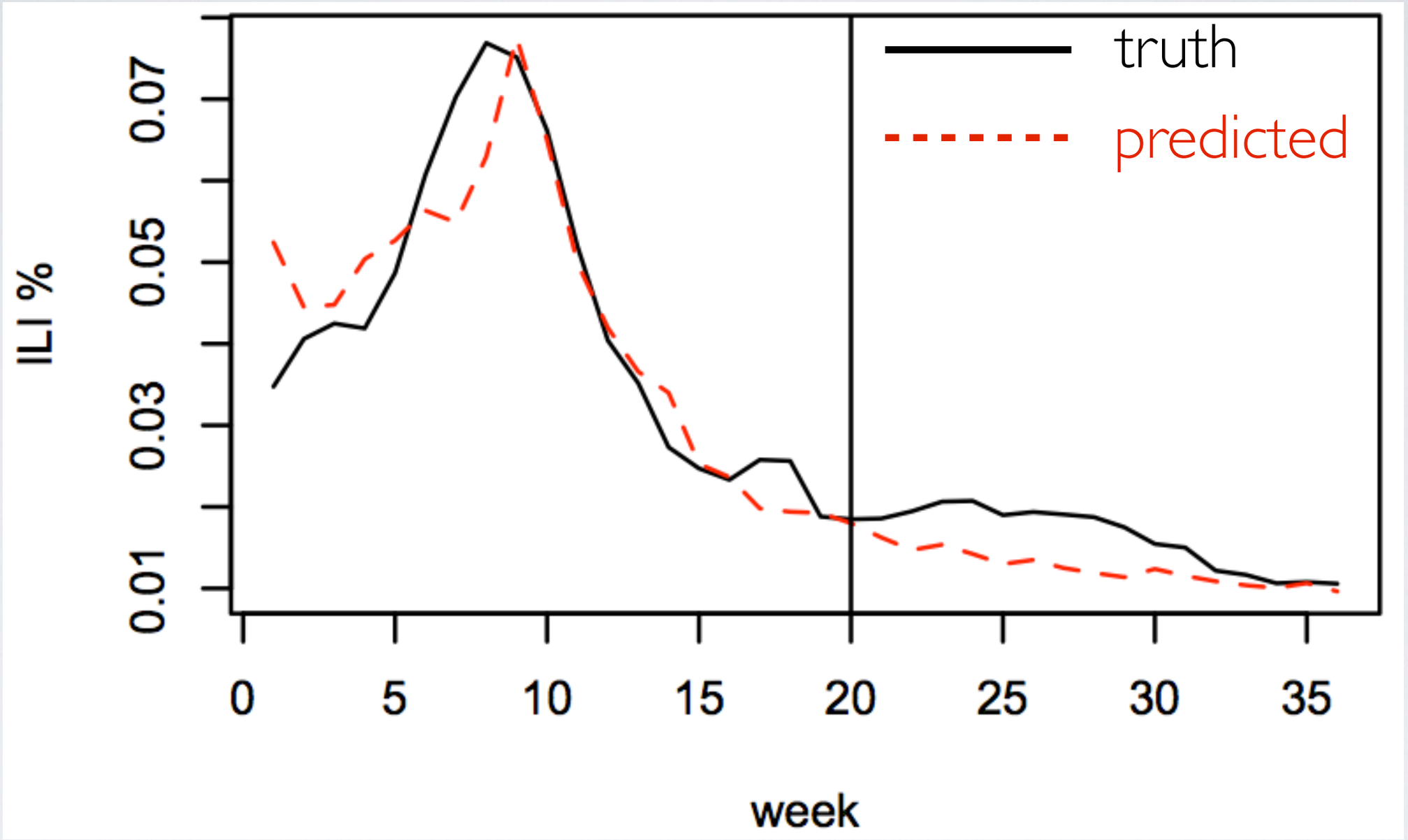
$$f(w, D) = \frac{|D_w|}{|D|}$$

D_w : tweets matching word w
D : all tweets this week

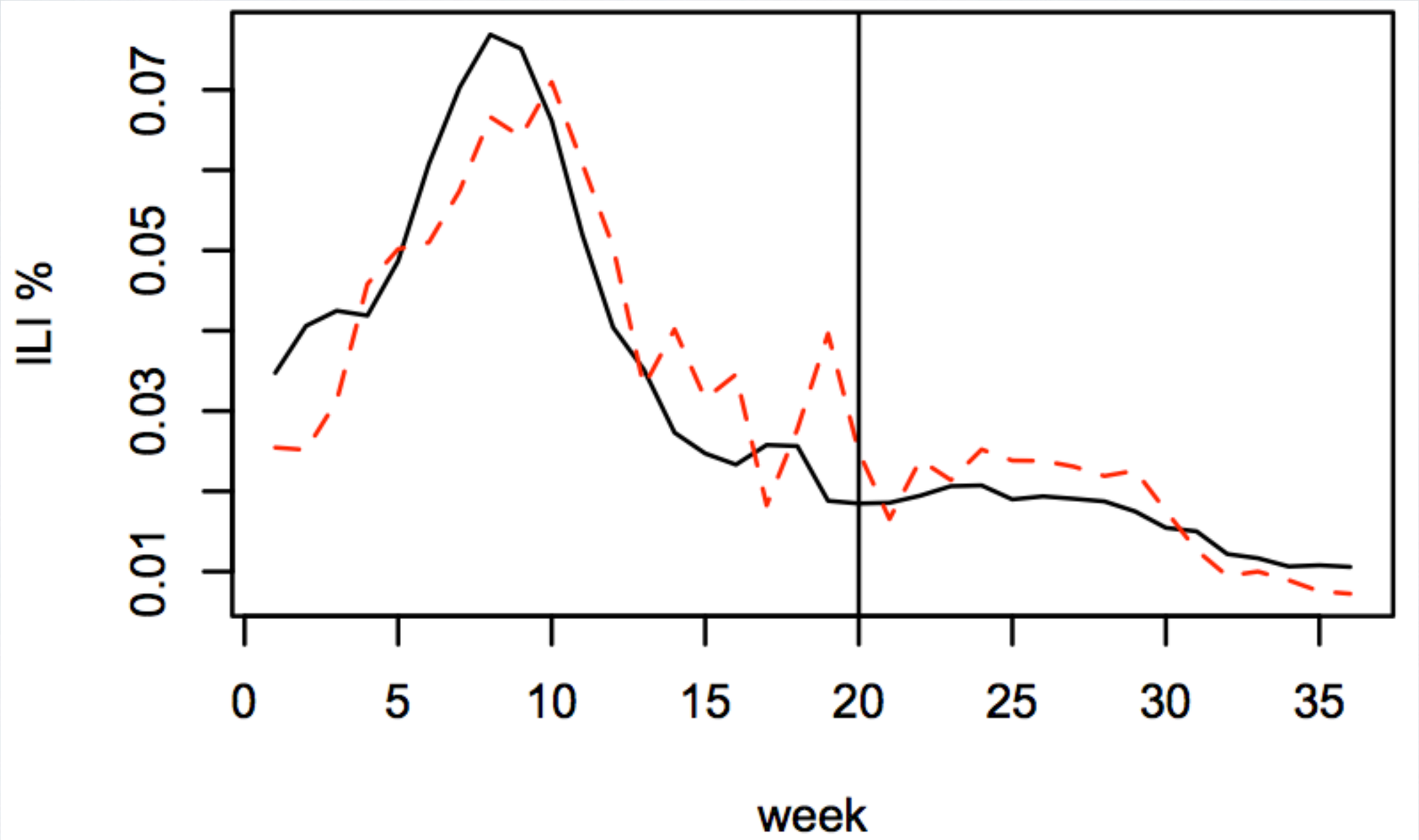
Regression: linear regression (constrained between 0-1)

$$\text{logit}(p) = \beta_1 \, \text{logit}(f(w, D)) + \beta_2 + \epsilon$$

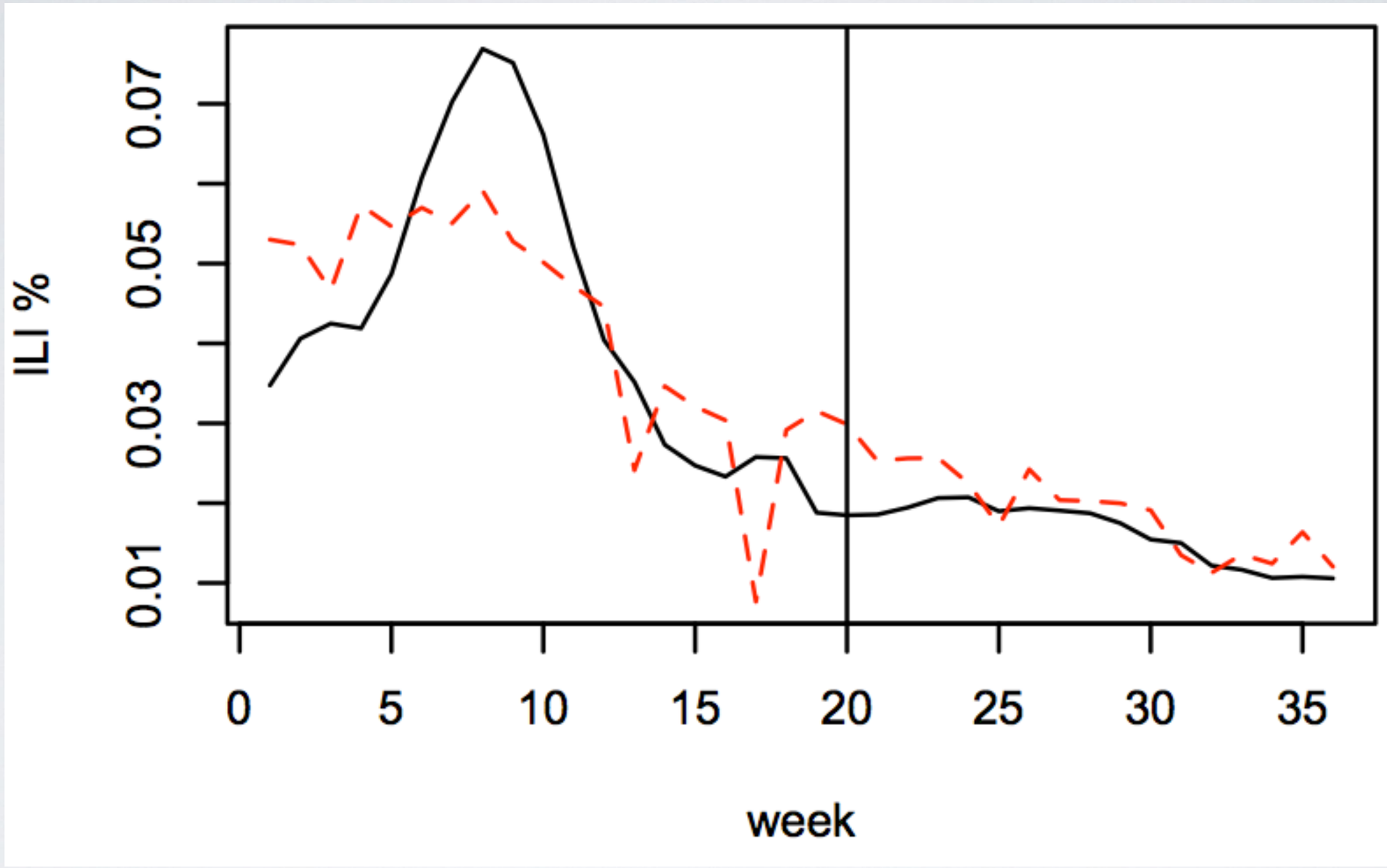
flu r_{train}=0.92 r_{test}=0.84



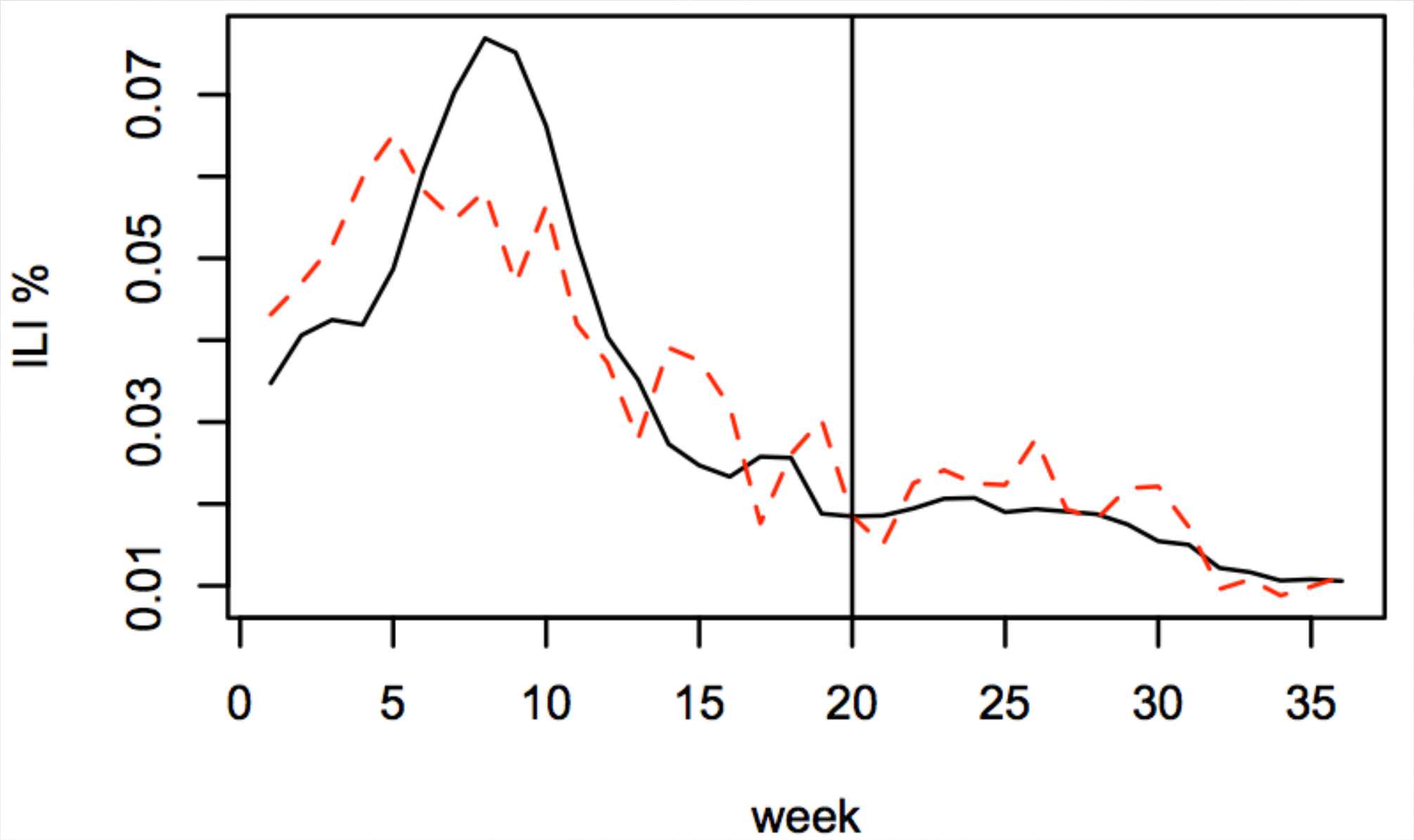
cough r_{train}=0.84 r_{test}=0.95



headache r_{train}=0.74 r_{test}=0.86

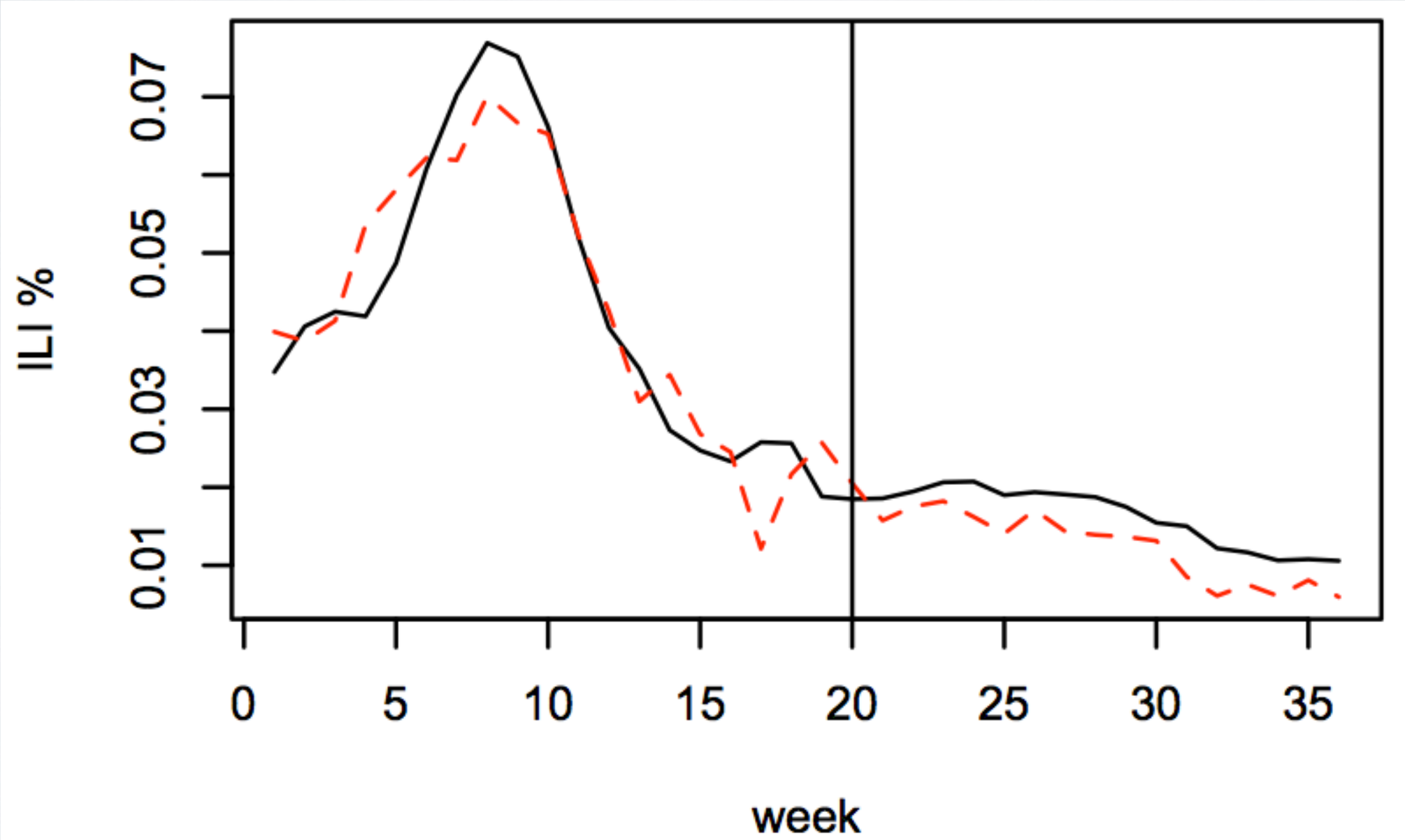


sore throat r_{train}=0.75 r_{test}=0.87

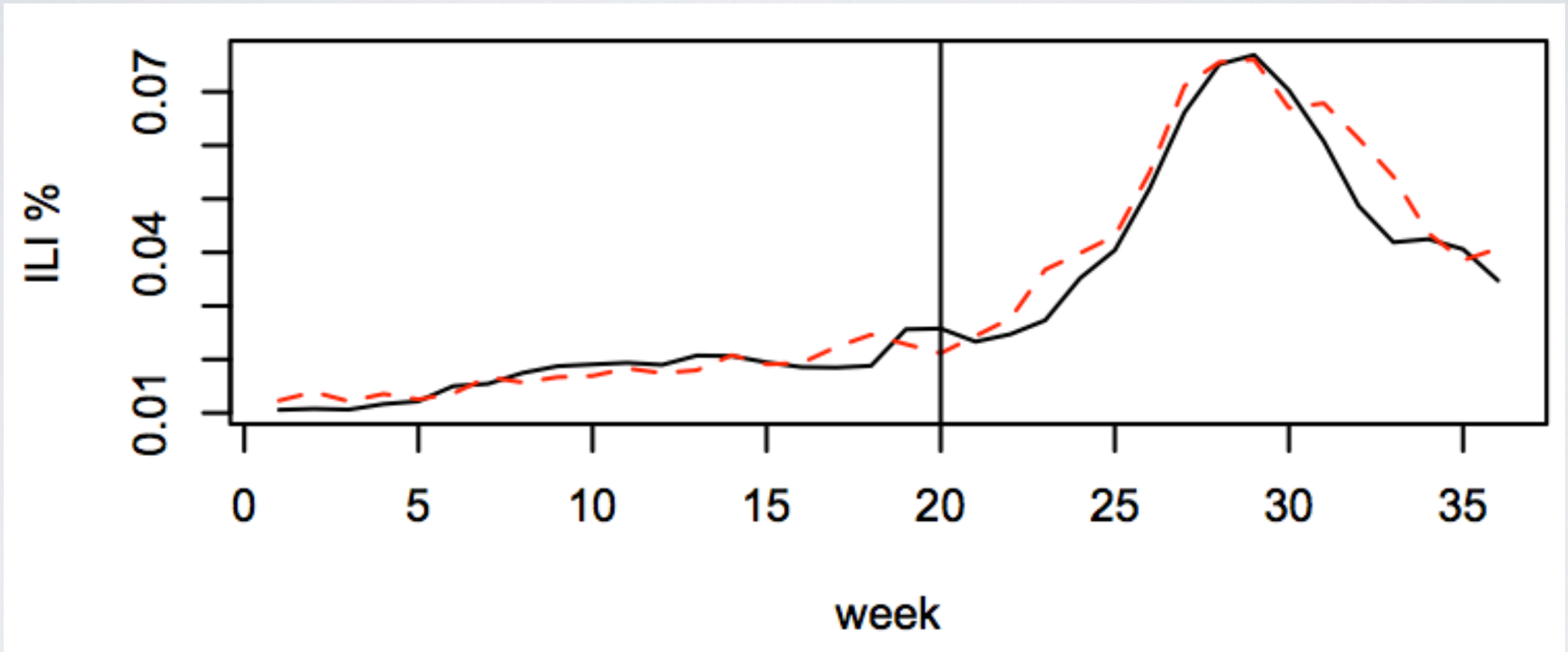


flu or cough or headache or sore throat

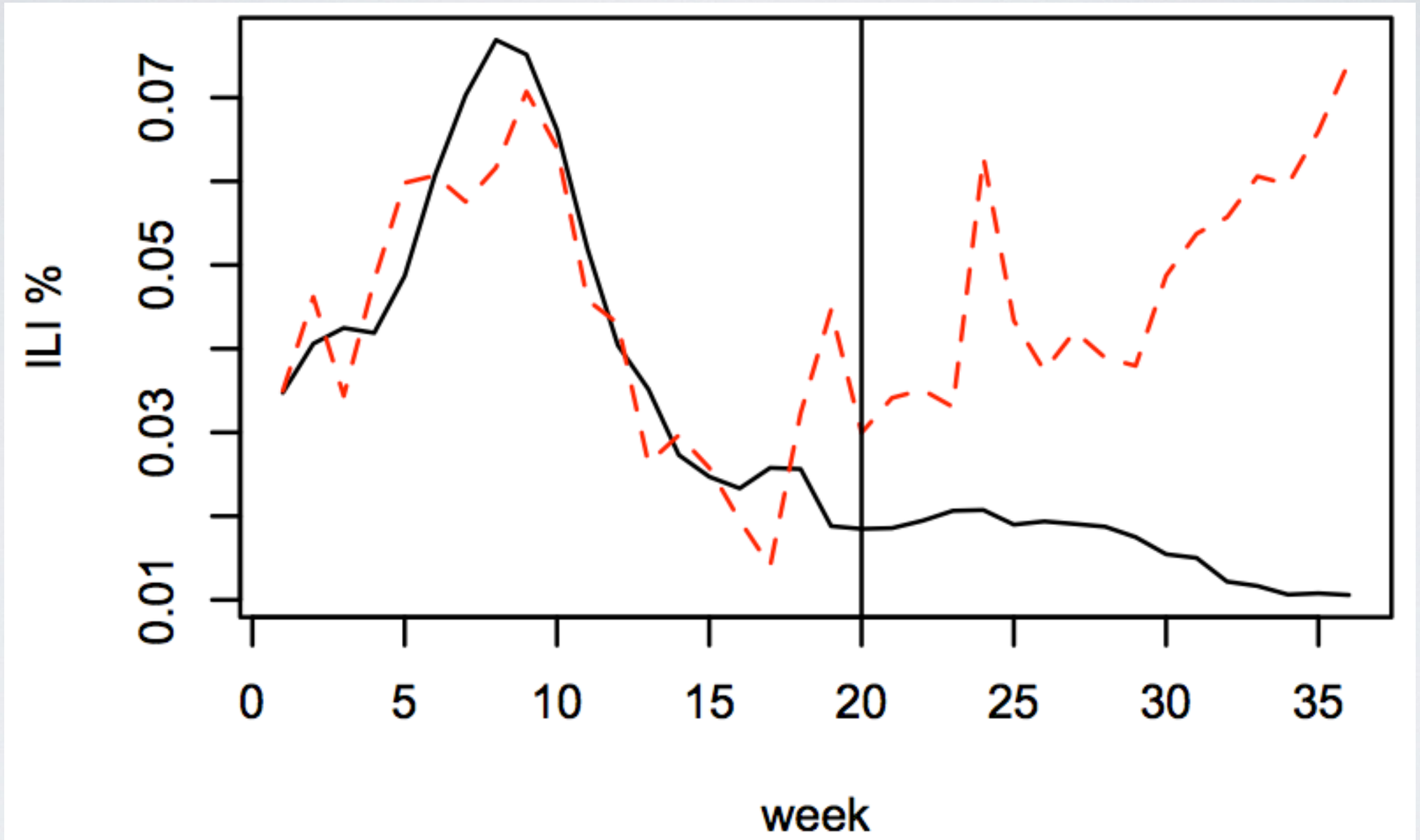
r_{train}=0.94 r_{test}=0.95



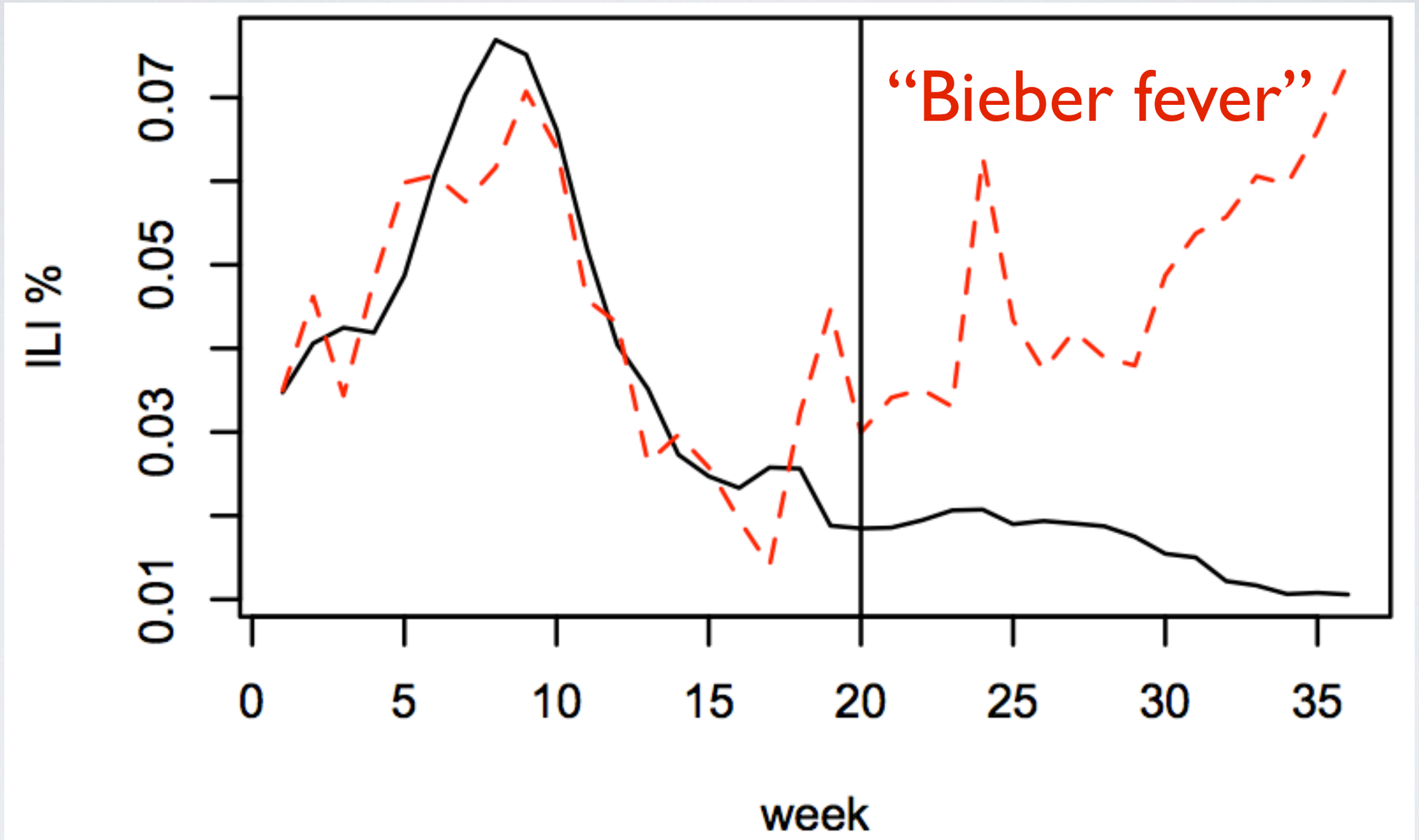
flu r_{train}=0.82 r_{test}=0.96



fever r_{train}=0.86 r_{test}=-0.77



fever r_{train}=0.86 r_{test}=-0.77



Spurious Matches

Are we tracking what we think we're tracking?

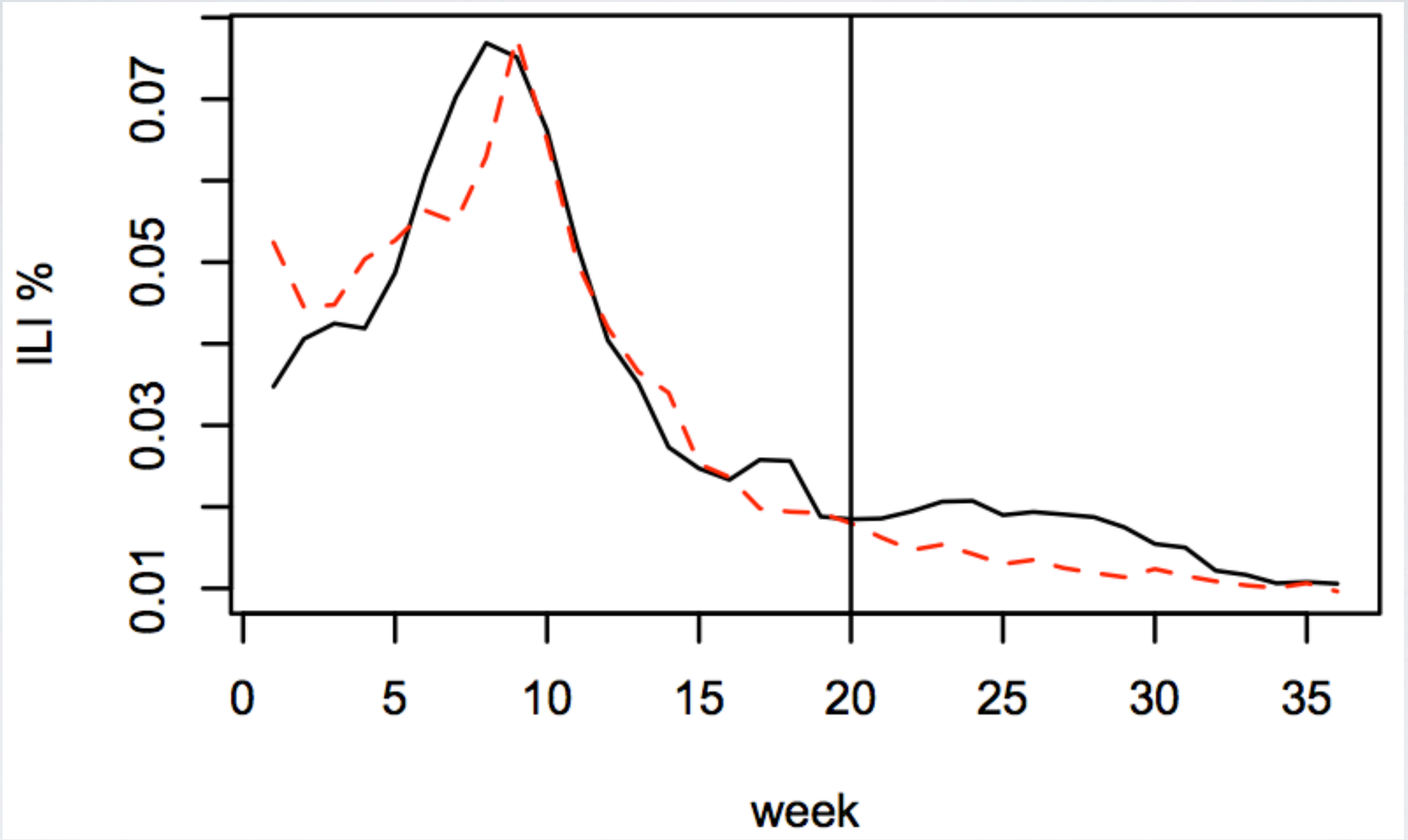
Compare:

"Had a great vacation!! Got home and got the flu. It was terrible!"

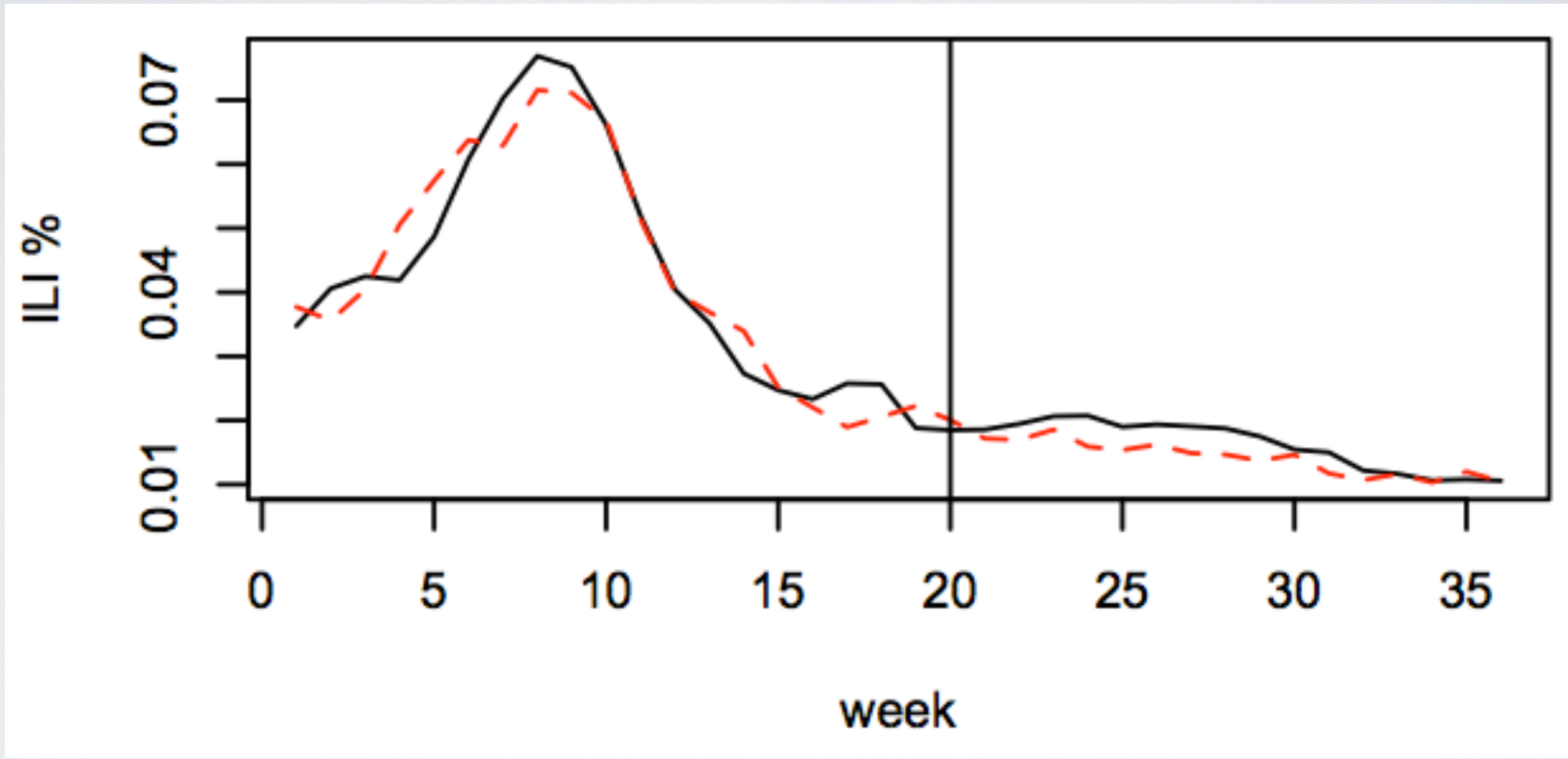
"Getting my flu shot. Hooray!"

"Businesses need to be just as ready for Swine Flu H1N1 this fall"

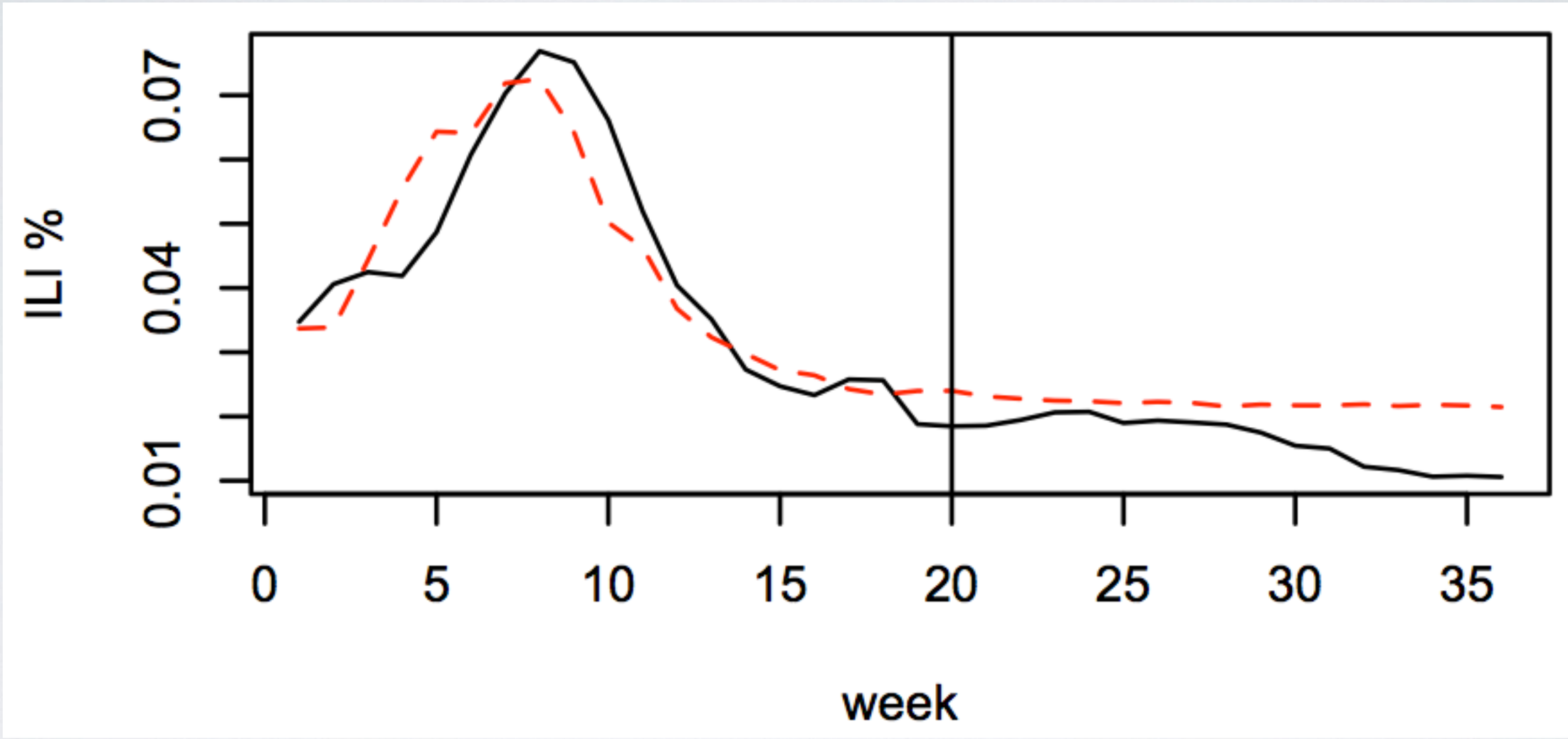
flu r_{train}=0.92 r_{test}=0.84



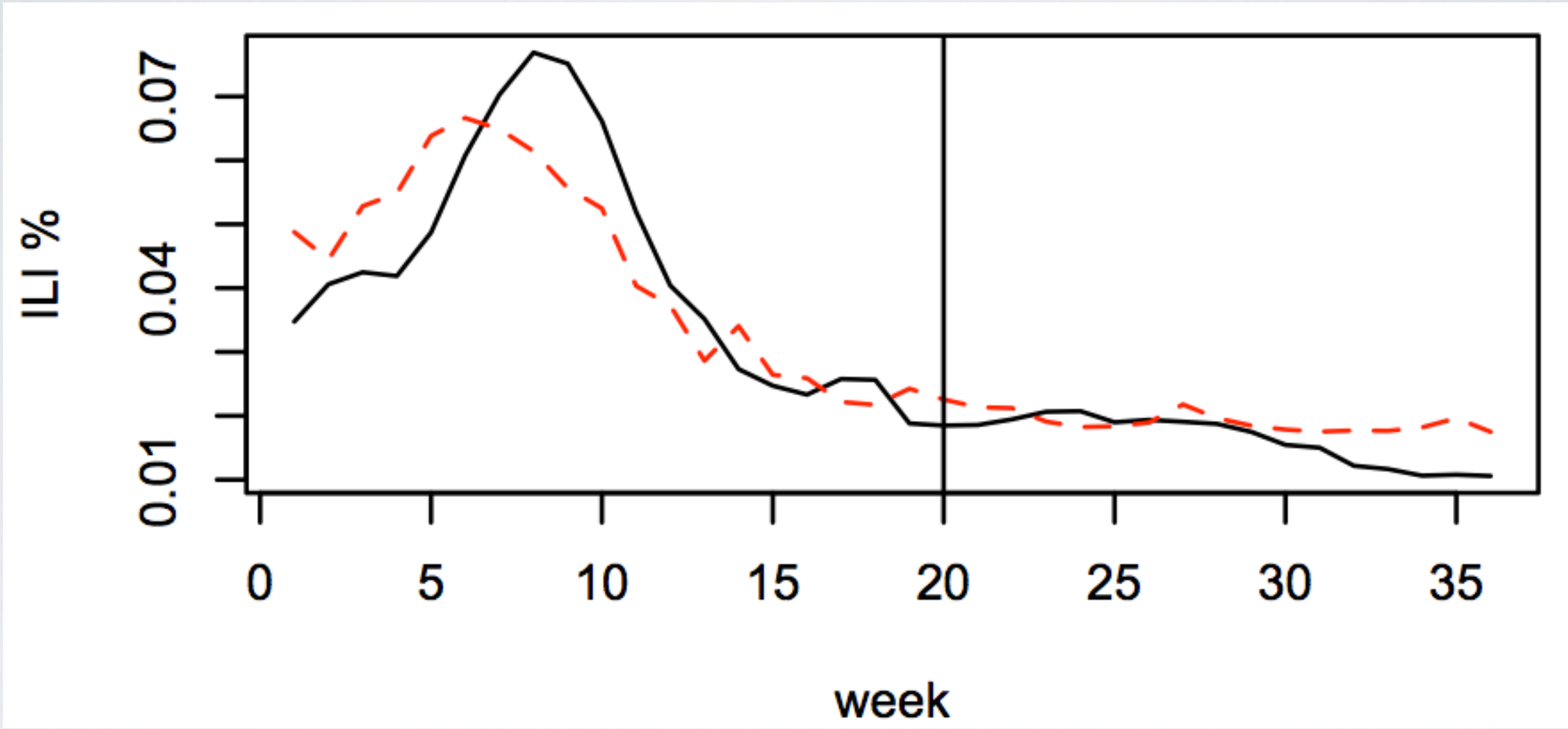
flu -(swine OR h1n1) r_{train}=0.97 r_{test}=0.91



flu AND shot -(swine OR h1n1) r_{train}=0.92 r_{test}=0.64



flu AND season (-swine OR h1n1) r_{train}=0.85 r_{test}=0.49



Nowcasting Rate of Influenza-like Illness (ILI)

Sampling: 570M tweets from 9/09-5/10

Classification: simple keyword matching

Does tweet contain flu, cough, headache, sore throat?

Aggregation: fraction of tweets matching keywords

$$f(w, D) = \frac{|D_w|}{|D|}$$

D_w : tweets matching word w

D : all tweets this week

Regression: linear regression (constrained between 0-1)

$$\text{logit}(p) = \beta_1 \, \text{logit}(f(w, D)) + \beta_2 + \epsilon$$

Nowcasting Rate of Influenza-like Illness (ILI)

Classification: simple keyword matching

Does tweet contain flu, cough, headache, sore throat?

Nowcasting Rate of Influenza-like Illness (ILI)

Classification: document classification
logistic regression (~84% accuracy)

Nowcasting Rate of Influenza-like Illness (ILI)

Classification: document classification
logistic regression (~84% accuracy)

Headache, cold sniffles, sore throat, sick in the tummy. Oh joy!! :' (

are you eating fruit breezers. those other the yummy ones haha. the other ones taste like well, cough drops haha.

Nowcasting Rate of Influenza-like Illness (ILI)

Classification: document classification
logistic regression (~84% accuracy)

Headache, cold sniffles, sore throat, sick in the tummy. Oh joy!! :' (

are you eating fruit breezers. those other the yummy ones haha. the other ones taste like well, cough drops haha.

Aggregation: fraction of tweets matching keywords

$$f(w, D) = \frac{|D_w|}{|D|}$$

D_w : tweets matching word w

D : all tweets this week

Nowcasting Rate of Influenza-like Illness (ILI)

Classification: document classification
logistic regression (~84% accuracy)

Headache, cold sniffles, sore throat, sick in the tummy. Oh joy!! :' (

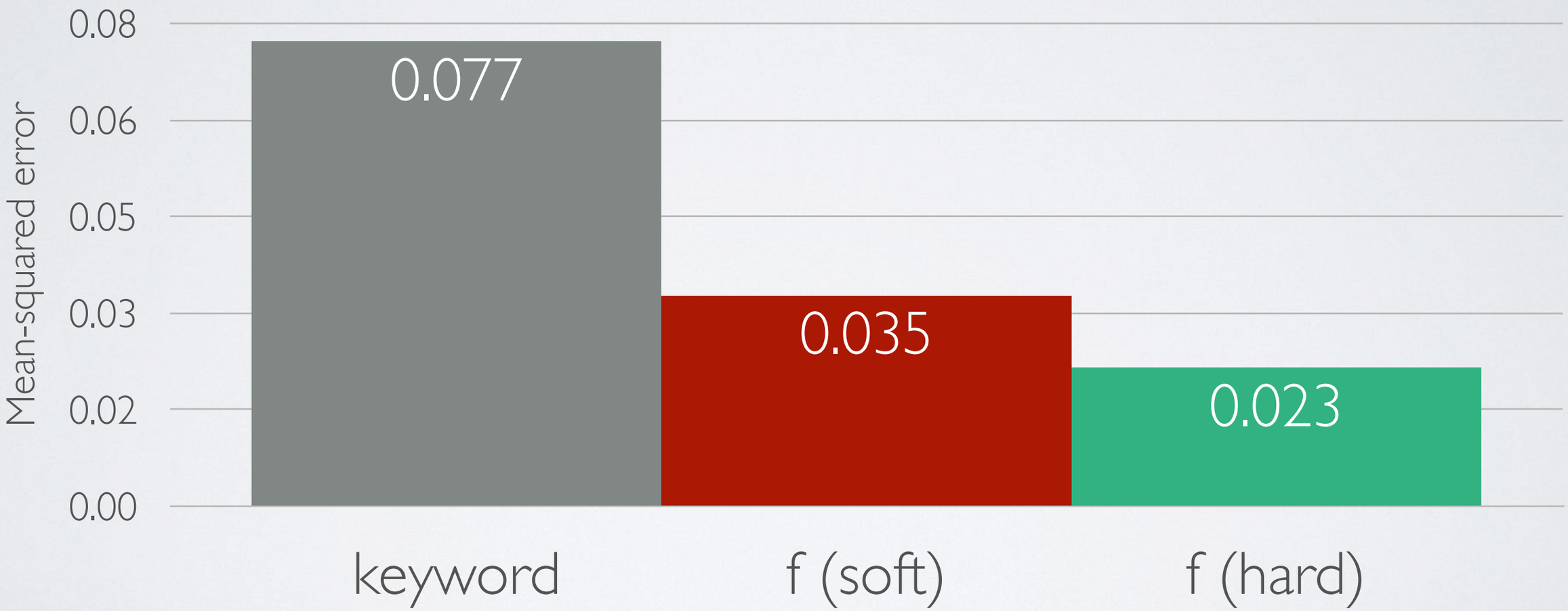
are you eating fruit breezers. those other the yummy ones haha. the other ones taste like well, cough drops haha.

Aggregation: % classified as positive examples

$$f_s(w, D) = \frac{\sum_{d_i} \Pr(y = 1|d_i)}{|D|}$$

$$f_h(w, D) = \frac{\sum_{d_i} \mathbf{1}[\Pr(y = 1|d_i) > 0.5]}{|D|}$$

Filtering Results: Simulated "false outbreak"



Related Work : Online Flu Tracking

News [Grishman et al '02, Mawudeku & Blench '06, Brownstein et al. '08]

Blogs [Corley et al '10]

Web browsing statistics [Johnson et al '04]

Search Engine Queries [Eysenbach '06, Polgreen et al '08, Ginsberg et al '09]

Twitter [Lampos & Cristianini 2010; Signorini et al 2011; Dredze 2011]

Outline

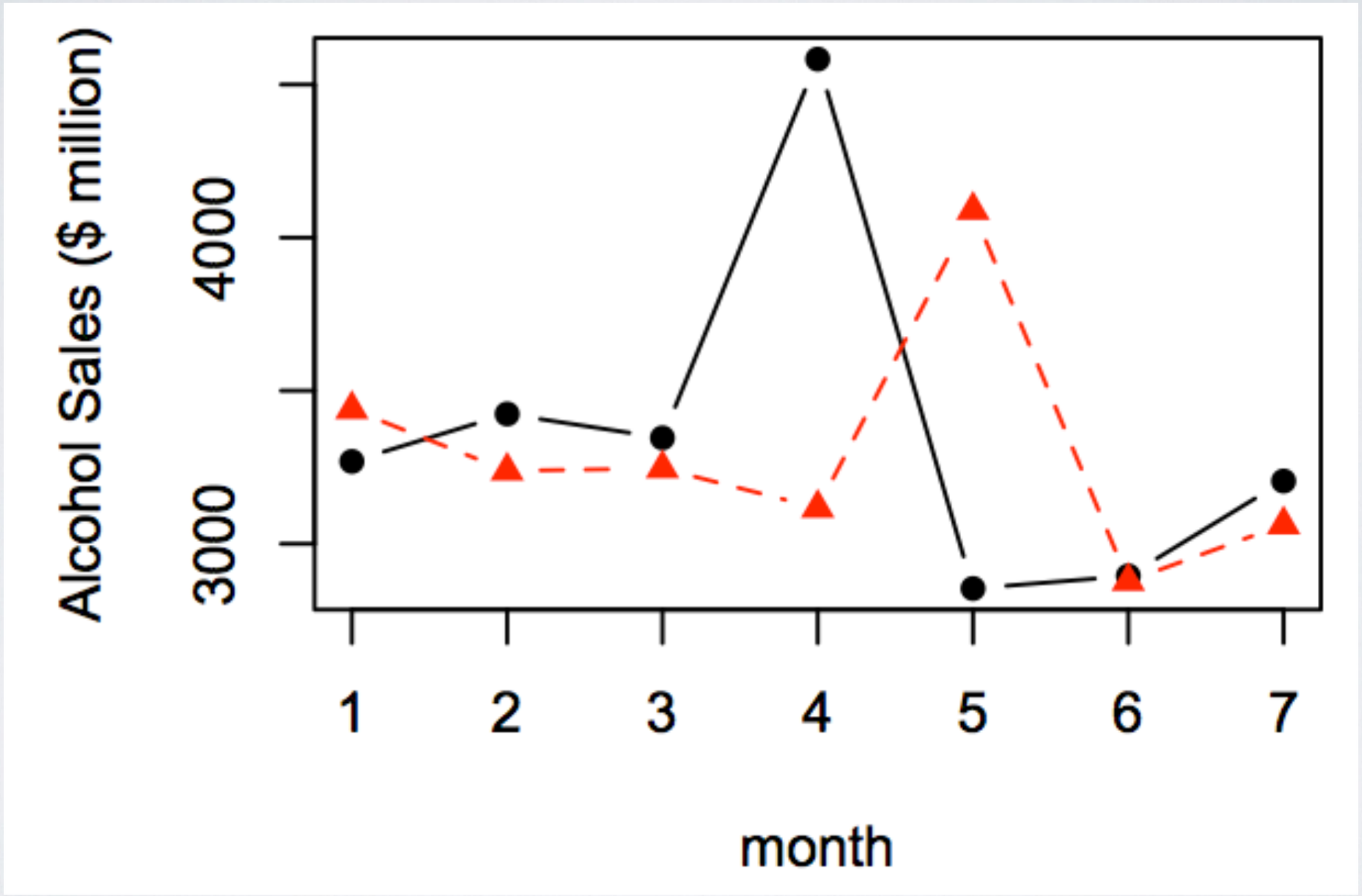
1. Tracking influenza-like illness rate
2. Tracking alcohol sales
3. Tracking community health
4. Inferring the origin of social media messages

Nowcasting Alcohol Sales

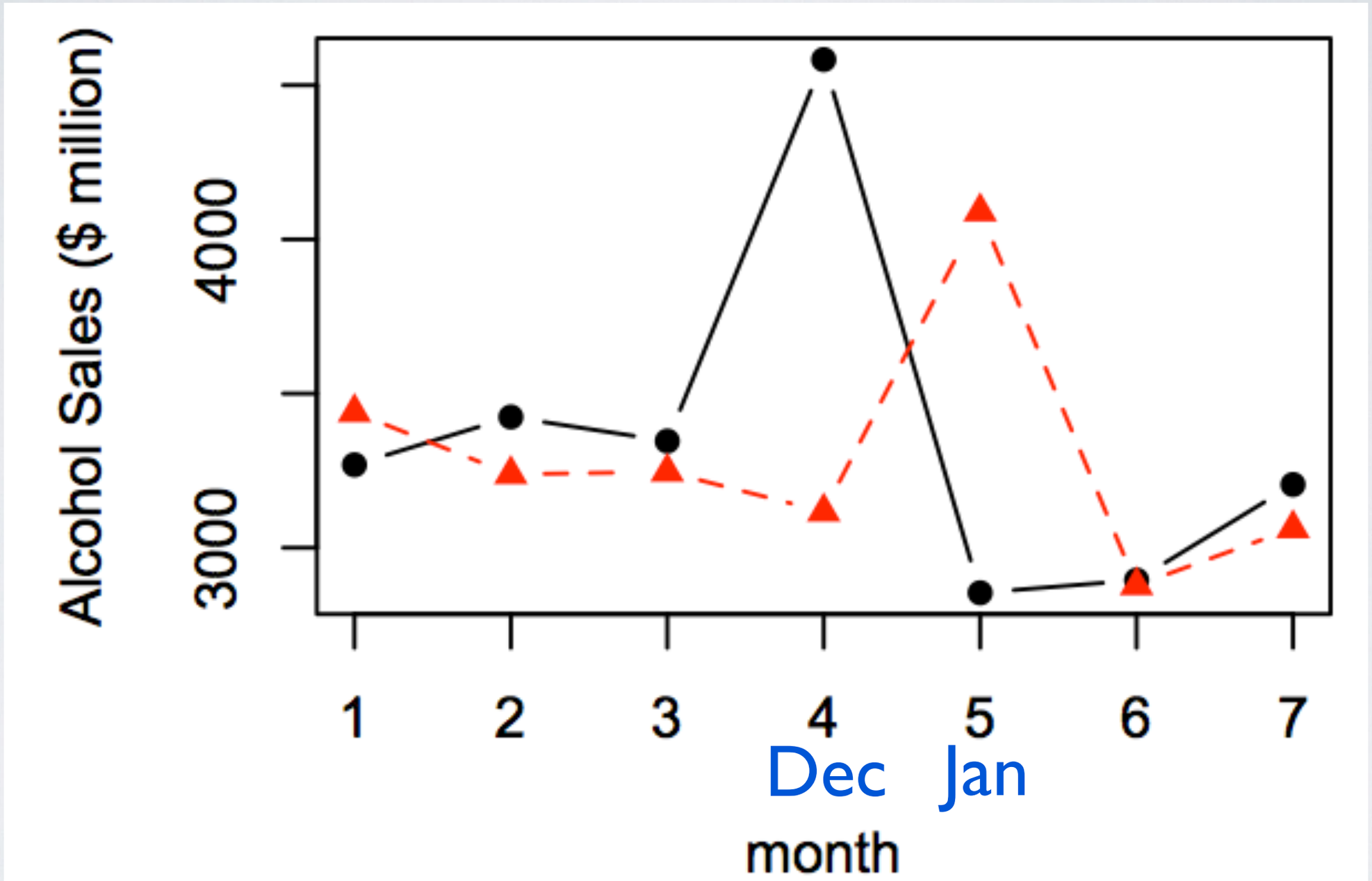
- U.S. Census reports monthly sales volume by industry
    - "Beer, wine, and liquor stores"
    - Collect for 7 months from Sept 2009-March 2010
    - "Leave-one-out" estimation
    - Fit model on 6 months, predict on 7th

drunk r = -0.3

drunk r = -0.3

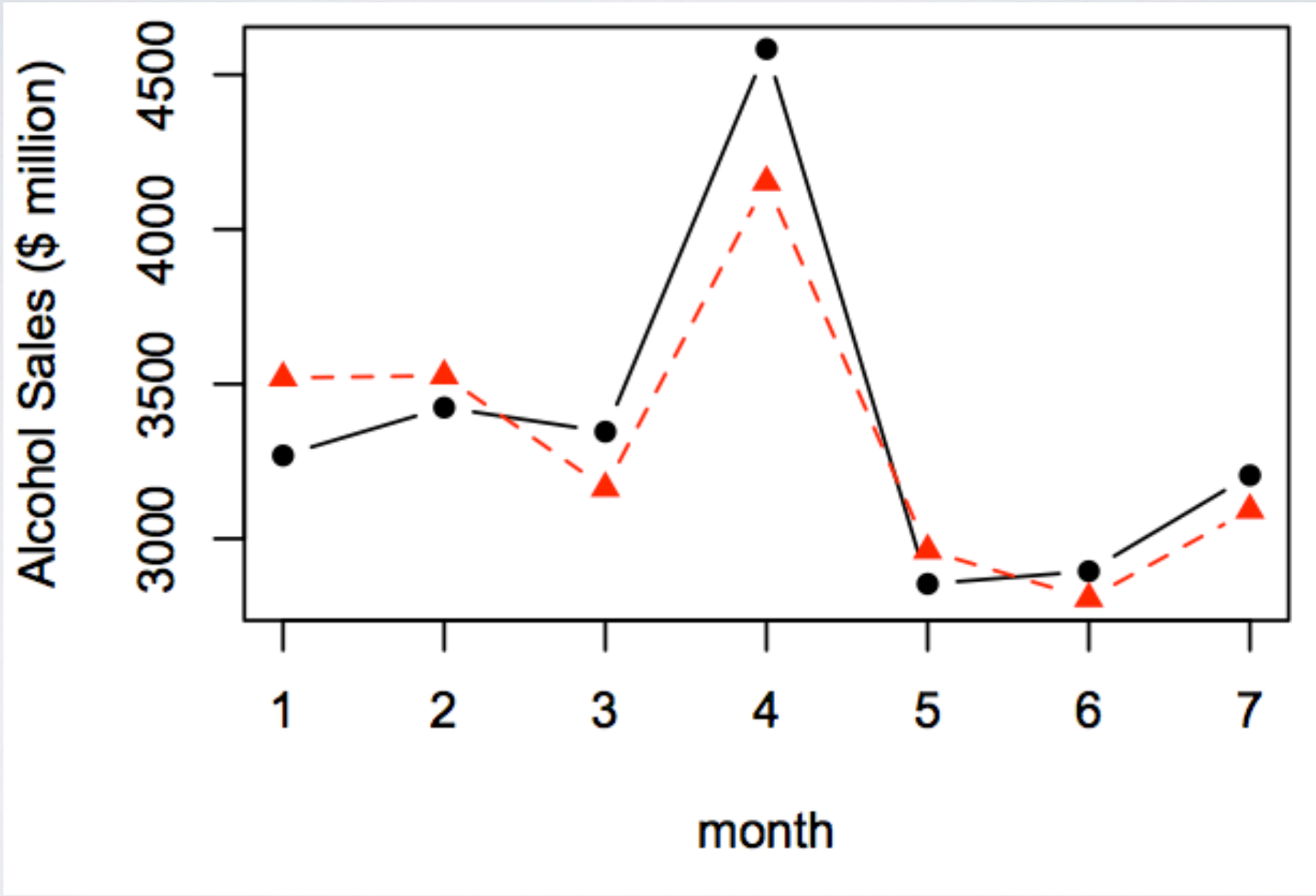


drunk r = -0.3



after 7-day lag

drunk r = 0.93



“Past-casting”?

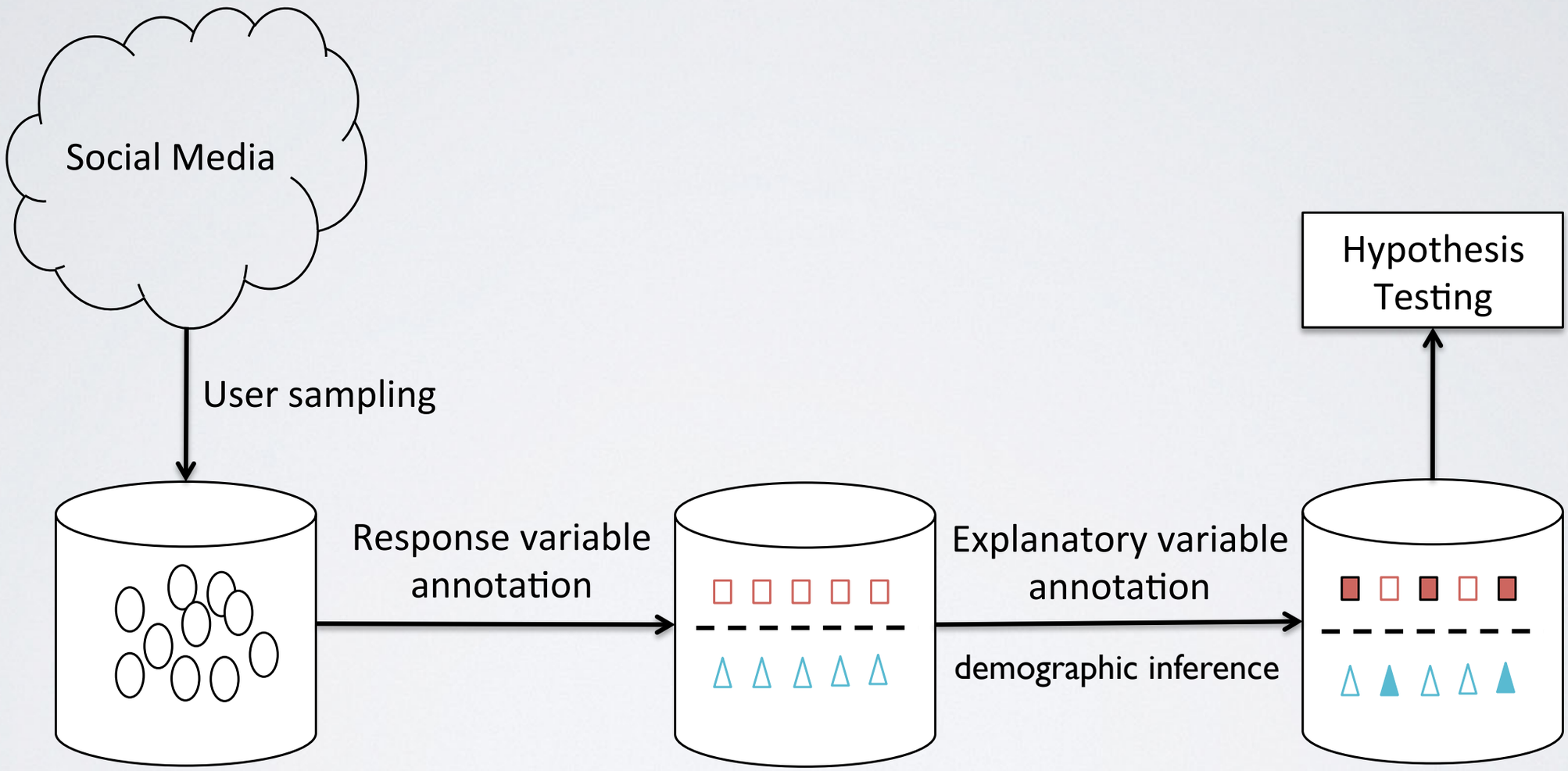
"Past-casting"?

- Validates Twitter data as useful sensor data

"Past-casting"?

- Validates Twitter data as useful sensor data
- Beyond nowcasting:
- **Web-scale observational studies**

Web-scale Observational Studies



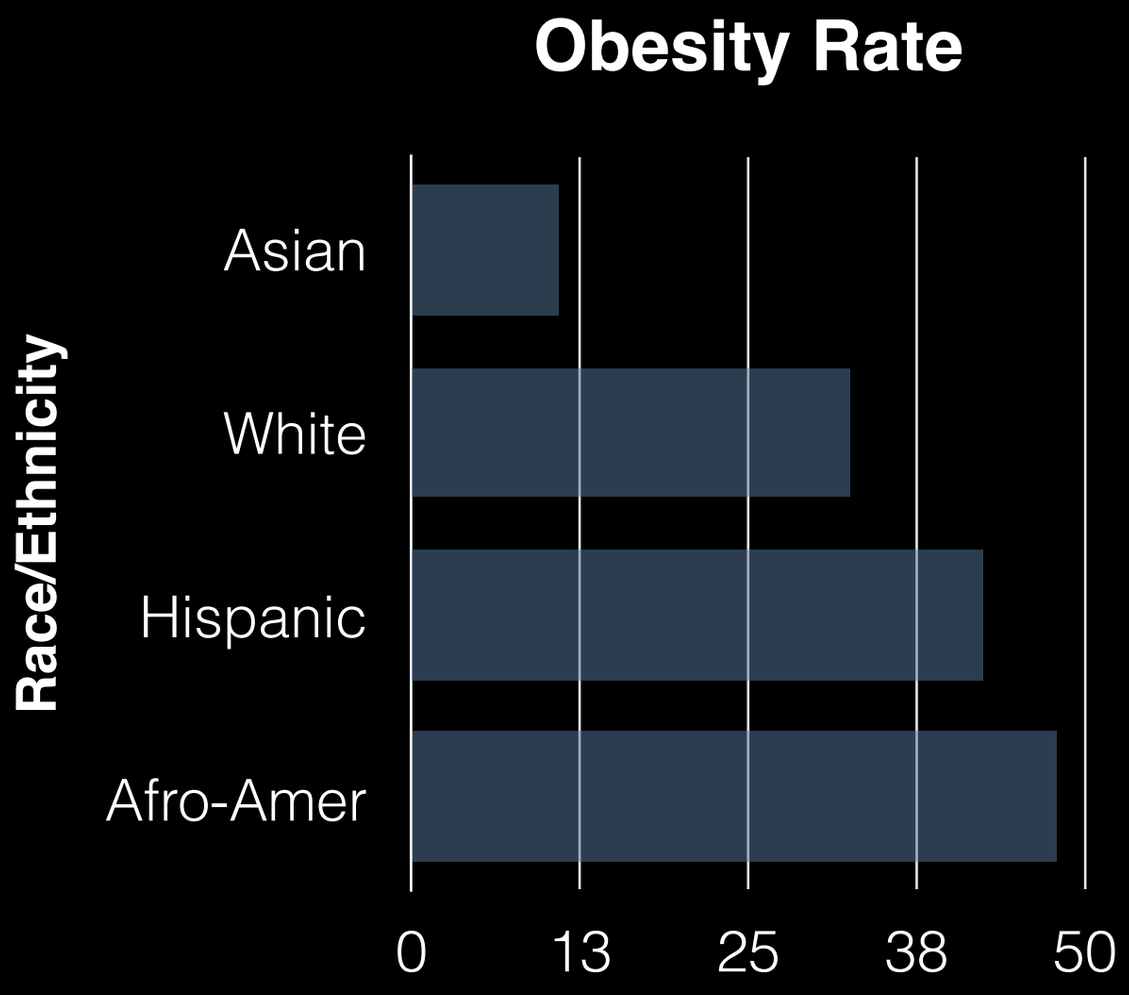
Computational social science [Lazar et al '09; Hopkins & King '10]

Outline

1. Tracking influenza-like illness rate
2. Tracking alcohol sales
3. Tracking community health
4. Inferring the origin of social media messages

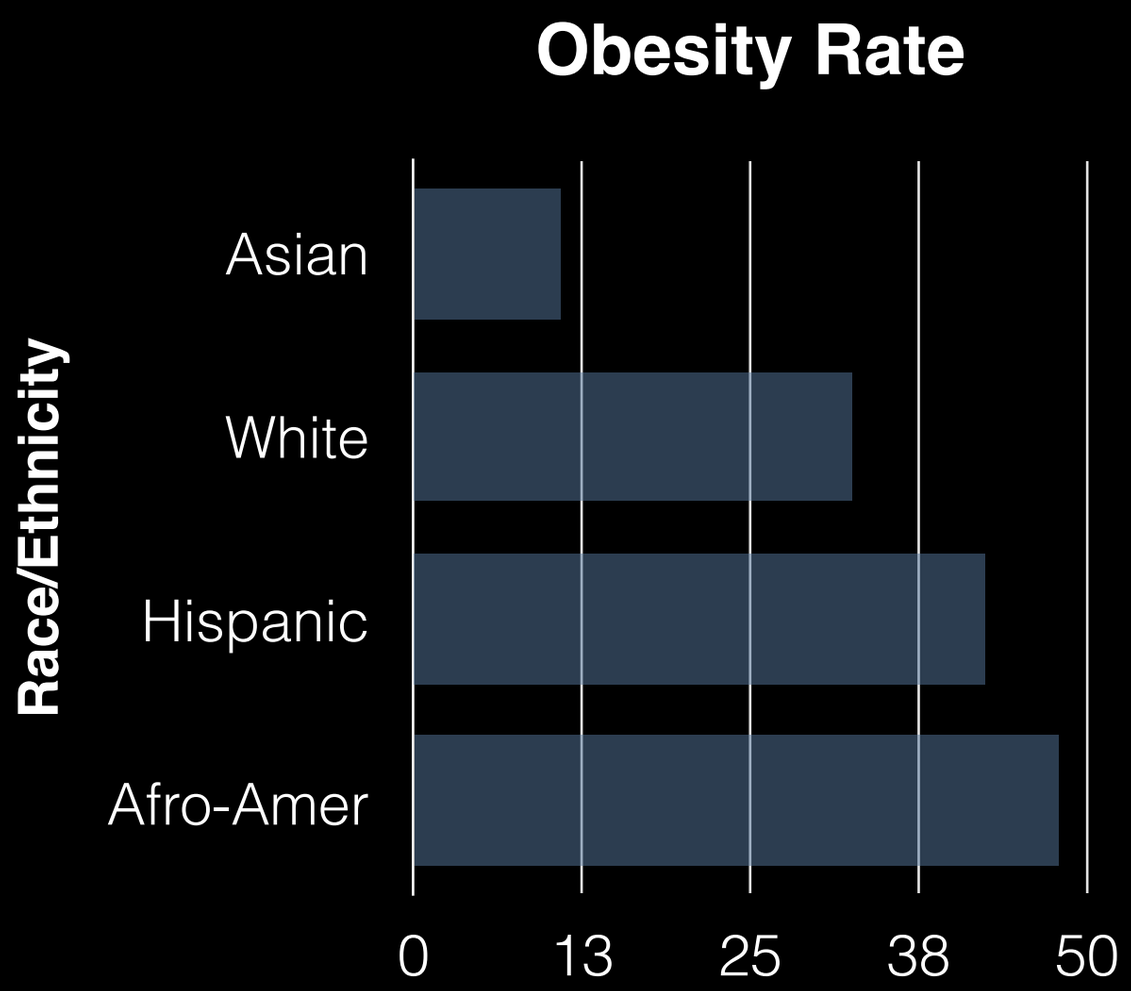
Health varies with socio-cultural factors.

Health varies with socio-cultural factors.

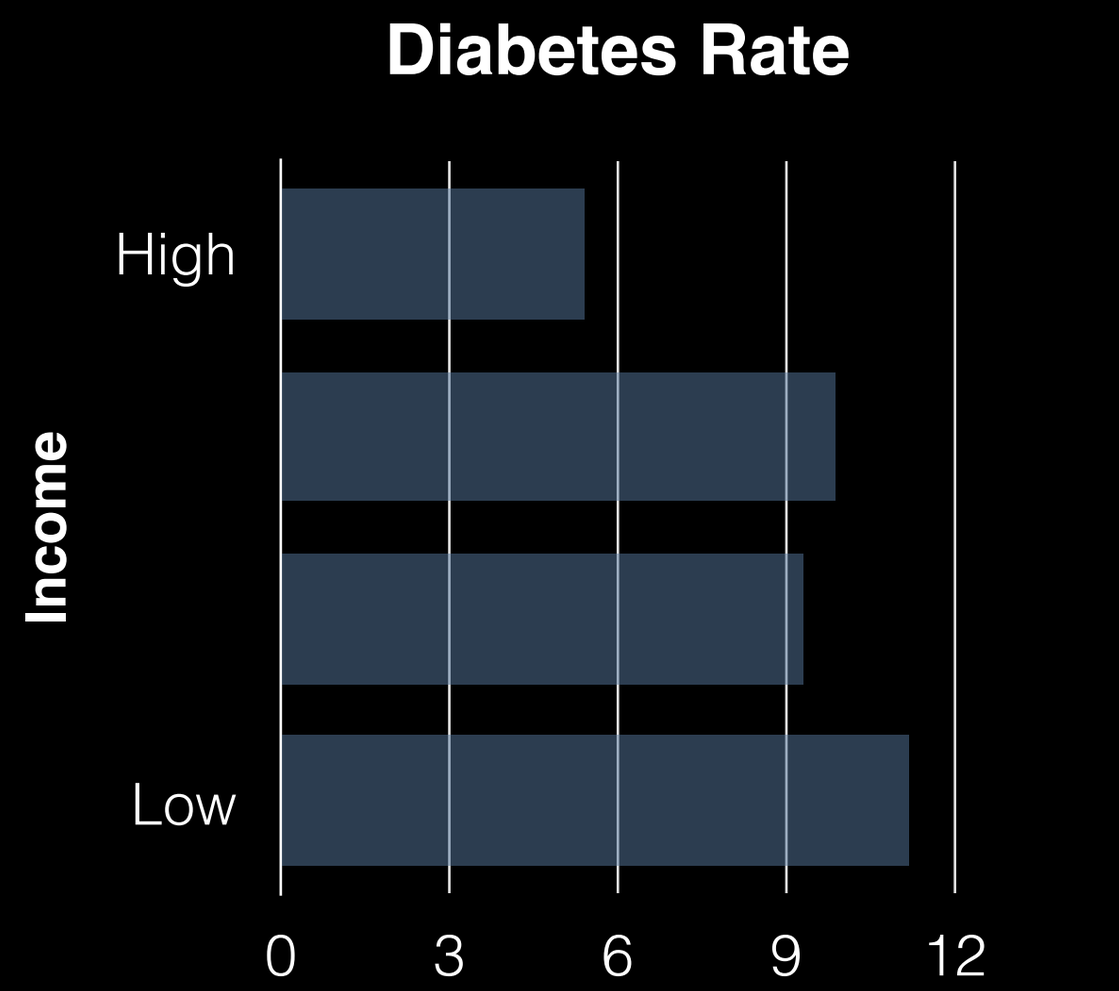


http://www.cdc.gov/obesity/data/adult.html

Health varies with socio-cultural factors.



http://www.cdc.gov/obesity/data/adult.html



http://www.cdc.gov/mmwr/preview/mmwrhtml/su6001a20.htm

Does health vary with language?

Does health vary with language?

Population A

Population B

Does health vary with language?

	Population A	Population B
Obesity Rate	**?**	**?**

Does health vary with language?

	Population A	Population B
Obesity Rate	**?**	**?**
% Afro-Hispanic	**45.3%**	**51.8%**

Does health vary with language?

	Population A	Population B
Obesity Rate	**?**	**?**
% Afro-Hispanic	**45.3%**	**51.8%**
Median Income	**$39K**	**$42K**

Does health vary with language?

	Population A	Population B
Obesity Rate	**?**	**?**
% Afro-Hispanic	**45.3%**	**51.8%**
Median Income	**$39K**	**$42K**
“tired”, “bored”	**7%**	**3%**

Does health vary with language?

	Population A	Population B
Obesity Rate	**?**	**?**
% Afro-Hispanic	**45.3%**	**51.8%**
Median Income	**$39K**	**$42K**
“tired”, “bored”	**7%**	**3%**
profanity	**12%**	**6%**

County Health Rankings & Roadmaps

Building a Culture of Health, County by County

A Robert Wood Johnson Foundation program

County Health Rankings & Roadmaps

Building a Culture of Health, County by County

A Robert Wood Johnson Foundation program

University of Wisconsin Population Health Institute

Robert Wood Johnson Foundation

County Health Rankings & Roadmaps

Building a Culture of Health, County by County

A Robert Wood Johnson Foundation program

University of Wisconsin Population Health Institute

Robert Wood Johnson Foundation

CDC

BRFSS

United States Census Bureau

HRSA

THE Dartmouth INSTITUTE FOR HEALTH POLICY & CLINICAL PRACTICE

27 health-related statistics

27 health-related statistics

Outcomes

Poor Health
Unhealthy Days
Mentally Health
Low Birthweight
Diabetes
Obesity

27 health-related statistics

Outcomes

Poor Health
Unhealthy Days
Mentally Health
Low Birthweight
Diabetes
Obesity

Behaviors

Smoking
Inactivity
Drinking
Driving Deaths
STIs
Teen Birth Rate

27 health-related statistics

Outcomes
Poor Health
Unhealthy Days
Mentally Health
Low Birthweight
Diabetes
Obesity

Care
Ambulatory Care
Uninsured
Primary Care
Dentists
Mammography

Behaviors
Smoking
Inactivity
Drinking
Driving Deaths
STIs
Teen Birth Rate

27 health-related statistics

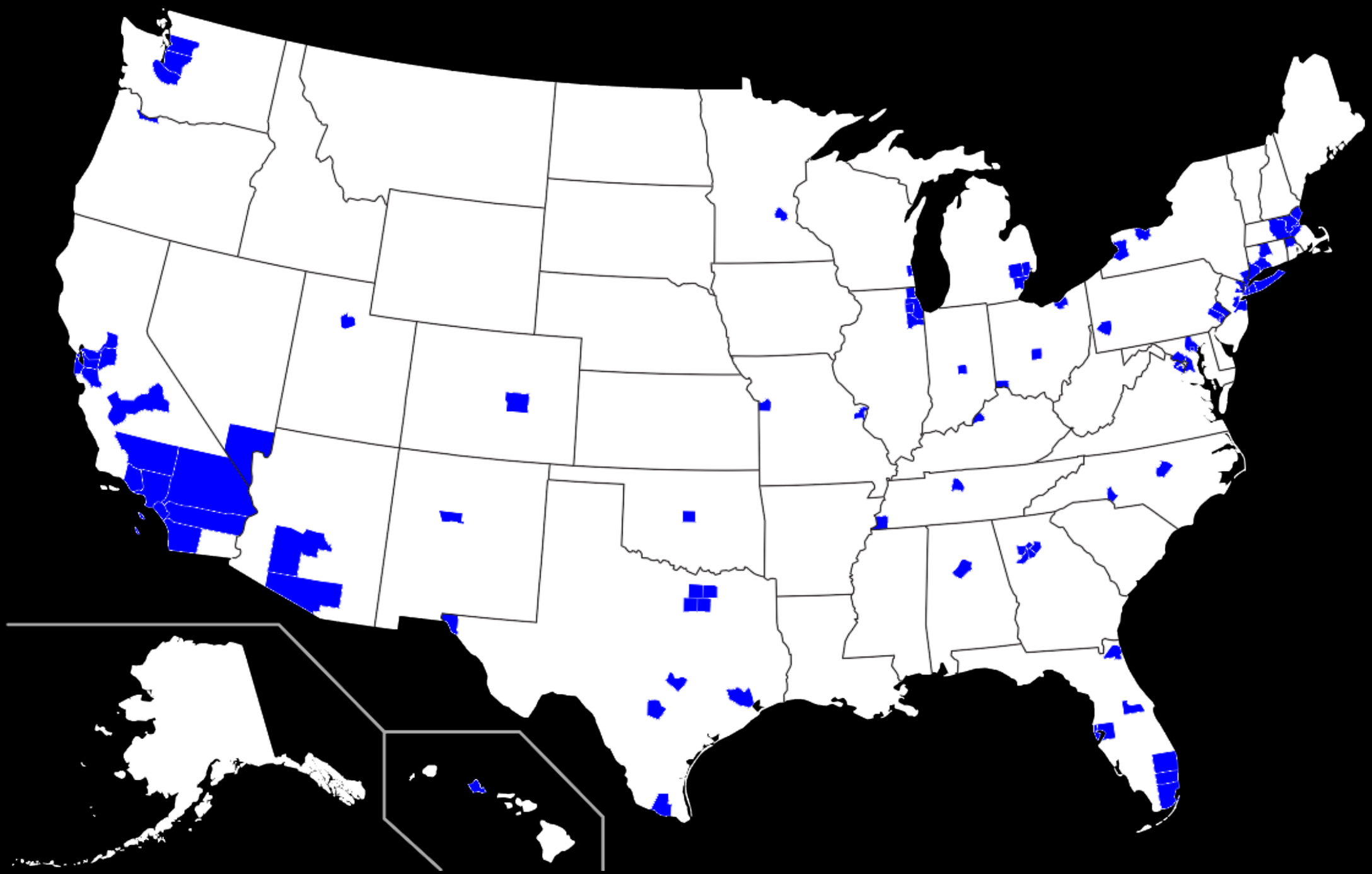
Outcomes
Poor Health
Unhealthy Days
Mentally Health
Low Birthweight
Diabetes
Obesity

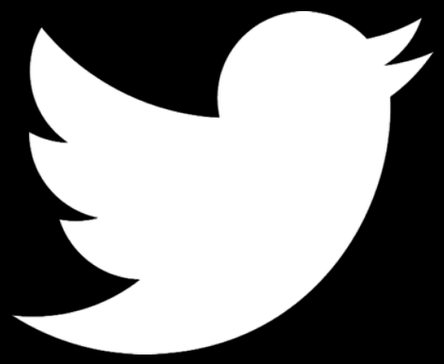
Care
Ambulatory Care
Uninsured
Primary Care
Dentists
Mammography

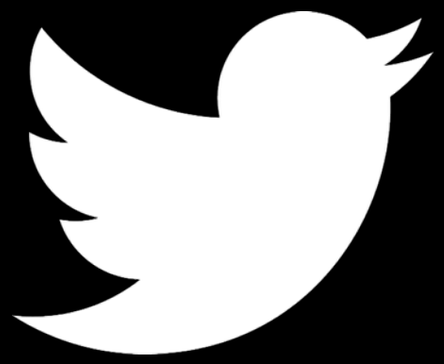
Environment
Education
Graduation Rate
Unemployment
Child Poverty
Social Support
Single Parent
Violent Crime
Rec. Facilities
Healthy Foods
Fast Food

Behaviors
Smoking
Inactivity
Drinking
Driving Deaths
STIs
Teen Birth Rate

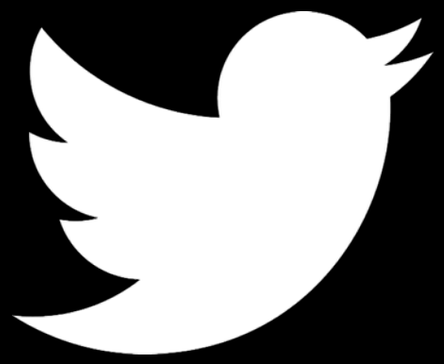
100 most populous counties in U.S.





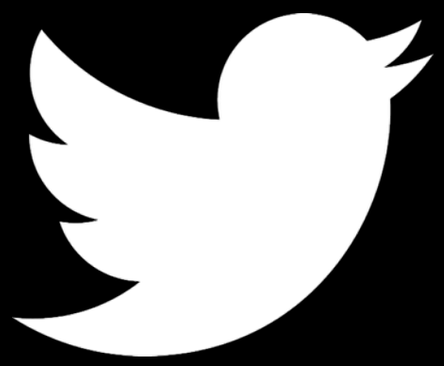


4.3M tweets



4.3M tweets

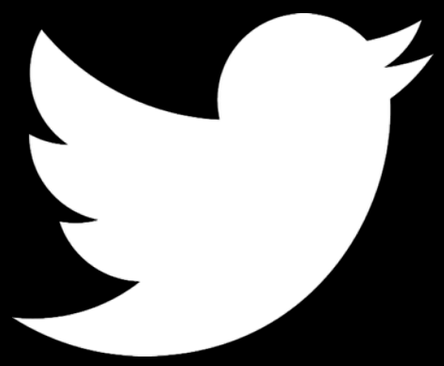
1.4M users



4.3M tweets

1.4M users

December, 2012 - September, 2013



4.3M tweets

1.4M users

December, 2012 - September, 2013

https://stream.twitter.com/1.1/statuses/filter.json?locations=-122.75,36.8,-121.75,37.8

Linguistic Representation

Justin Bieber

@justinbieber

Let's make the world better. Get @shots and spread the love and positivity. ♛

youtube.com/justinbieber

Justin Bieber

@justinbieber

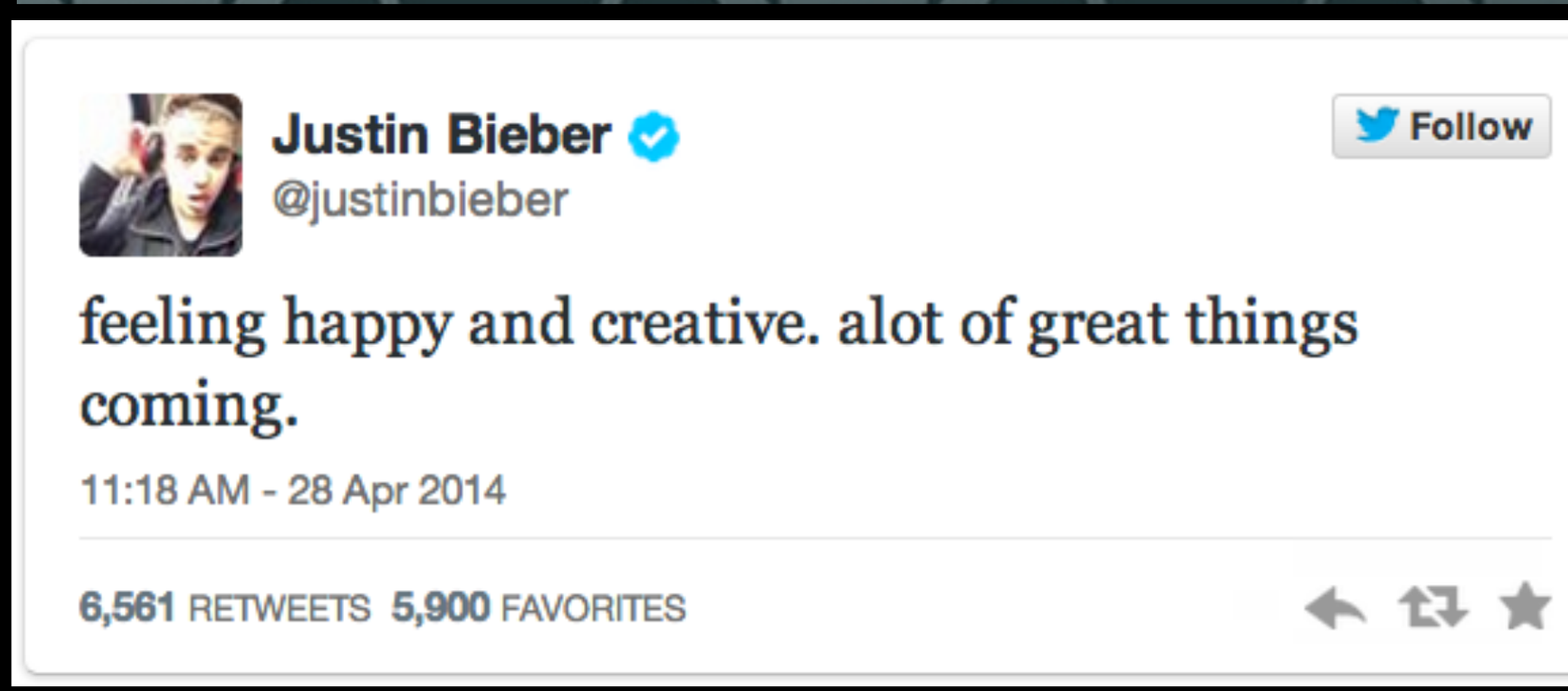
Follow

feeling happy and creative. alot of great things coming.

11:18 AM - 28 Apr 2014

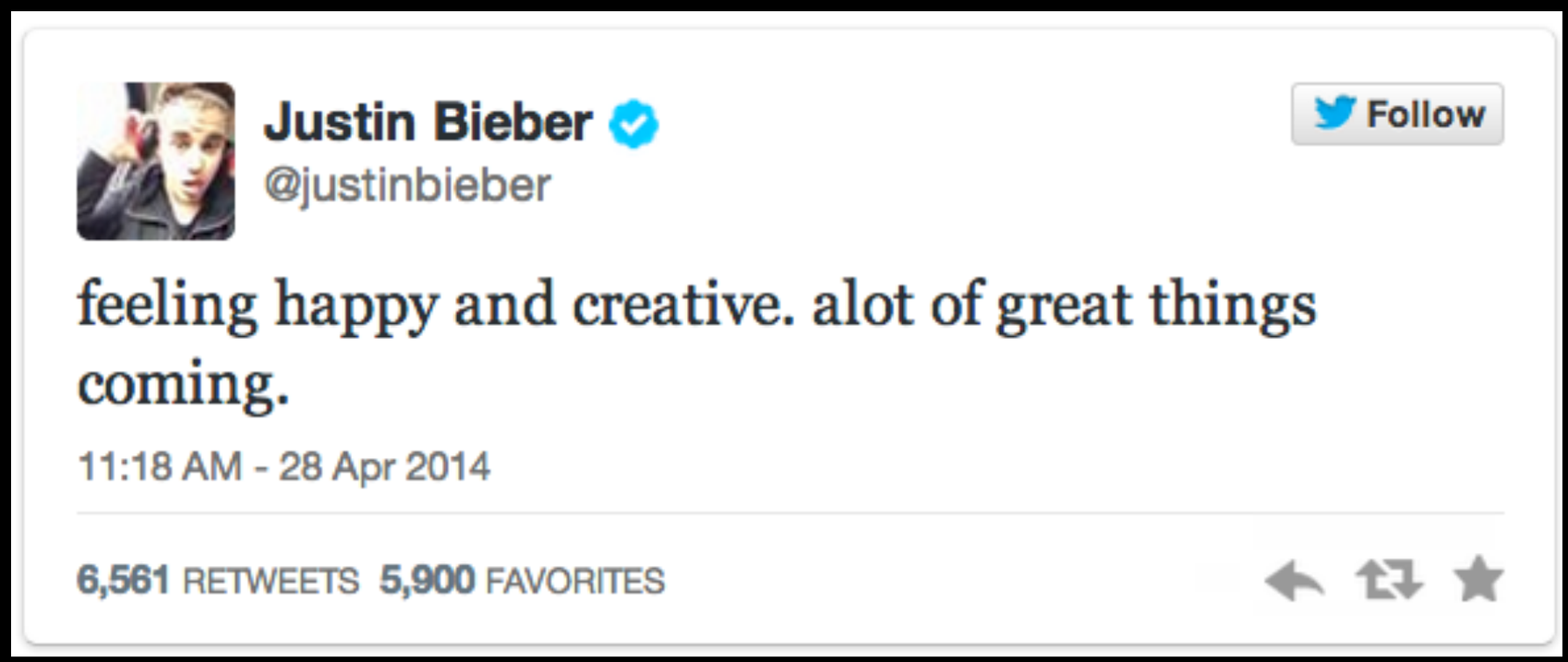
6,561 RETWEETS **5,900** FAVORITES


Justin Bieber
@justinbieber
Let's make the world better. Get @shots and spread the love and positivity.
youtube.com/justinbieber


Justin Bieber
@justinbieber
Follow
feeling happy and creative. alot of great things coming.
11:18 AM - 28 Apr 2014
6,561 RETWEETS 5,900 FAVORITES

description

text



LIWC: Linguistic Inquiry and Word Count Lexicon
[Pennebaker et al. 2001]
70 categories; 2.3K word patterns

PERMA: [Seligman 2011]
10 categories; 1.5K words

160 categories (80 description, 80 text)

LIWC: Linguistic Inquiry and Word Count Lexicon
[Pennebaker et al. 2001]
70 categories; 2.3K word patterns

PERMA: [Seligman 2011]
10 categories; 1.5K words

160 categories (80 description, 80 text)

Prior work:
- life satisfaction [Schwartz et al. 2013]
- personality [Qui et al. 2012]
- depression [De Choudhury 2013]
- obesity [Paul & Dredze 2011]

Justin Bieber

@justinbieber

Let's make the world better. Get @shots and spread the love and positivity. ♛

youtube.com/justinbieber

Justin Bieber

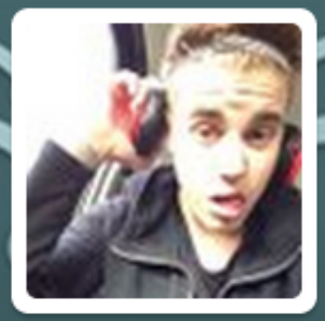
@justinbieber

Follow

feeling happy and creative. alot of great things coming.

11:18 AM - 28 Apr 2014

6,561 RETWEETS **5,900** FAVORITES



Justin Bieber

@justinbieber

Let's make the world better. Get @shots and spread the love and positivity. ♛

e.com/justinbieber

Affect
Positive Emotion
Positive Feeling

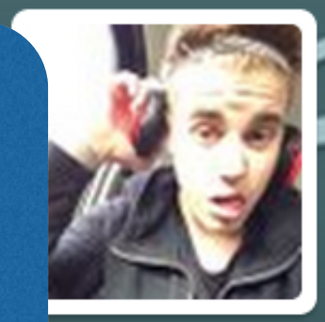
Follow

feeling happy and creative. alot of great things coming.

11:18 AM - 28 Apr 2014

6,561 RETWEETS **5,900** FAVORITES

We
Self
Social
Other-reference



n Bieber

@justinbieber

Let's make the world better. Get @shots and spread the love and positivity. ♛

e.com/justinbieber

Affect
Positive Emotion
Positive Feeling

Follow

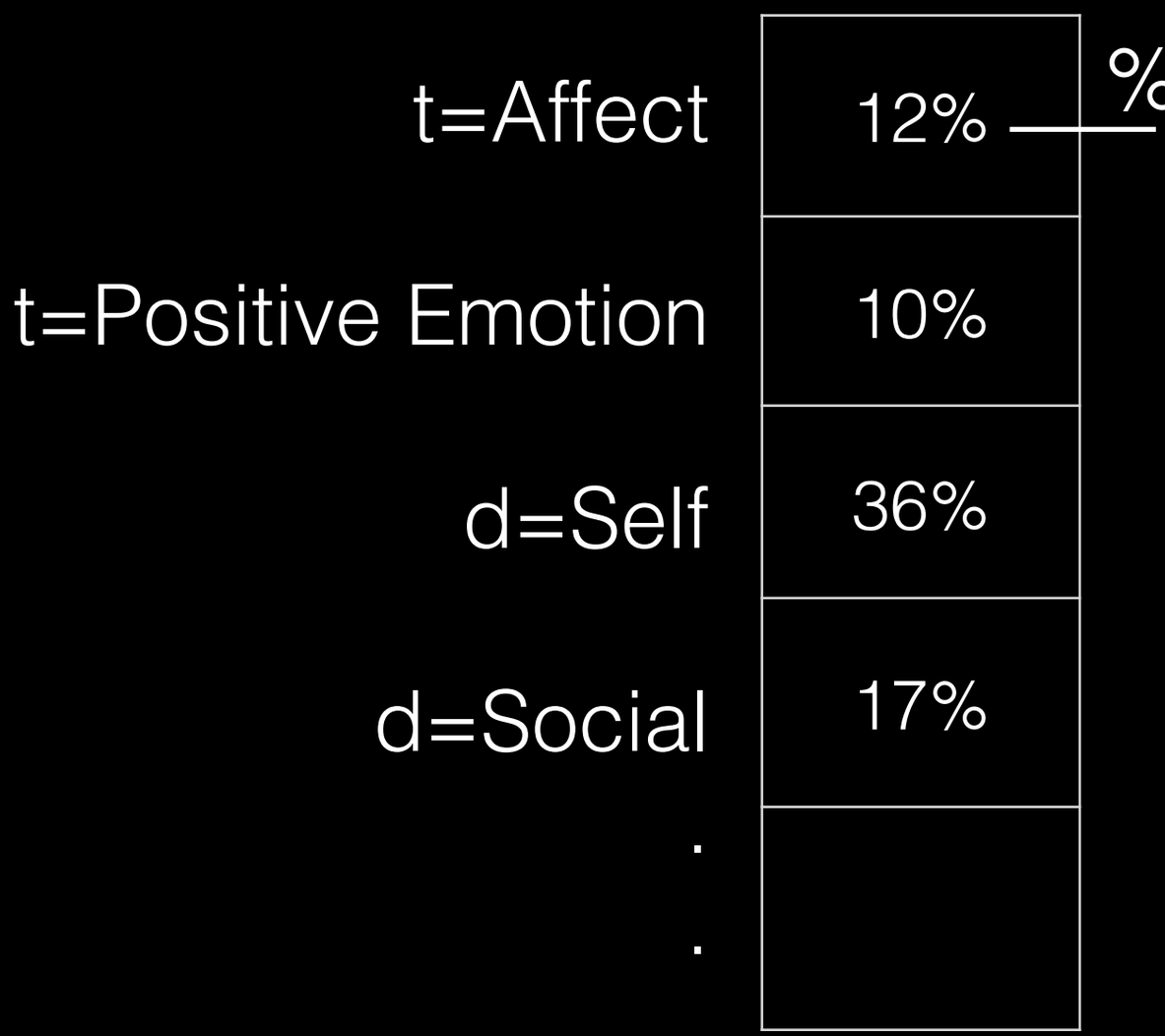
feeling happy and creative. alot of great things coming.

11:18 AM - 28 Apr 2014

6,561 RETWEETS **5,900** FAVORITES

Linguistic Profile of each County

L.A. County . . . [100]



% of users who mention a word from this category

27 Ridge Regression Models

L.A. County

t=Affect	12%	β_1^{obesity}	$\beta_1^{\text{insurance}}$
t=Positive Emotion	10%	β_2^{obesity}	$\beta_2^{\text{insurance}}$
d=Self	36%	β_3^{obesity}	$\beta_3^{\text{insurance}}$
d=Social	17%	β_4^{obesity}	$\beta_4^{\text{insurance}}$
.			
.			
.			

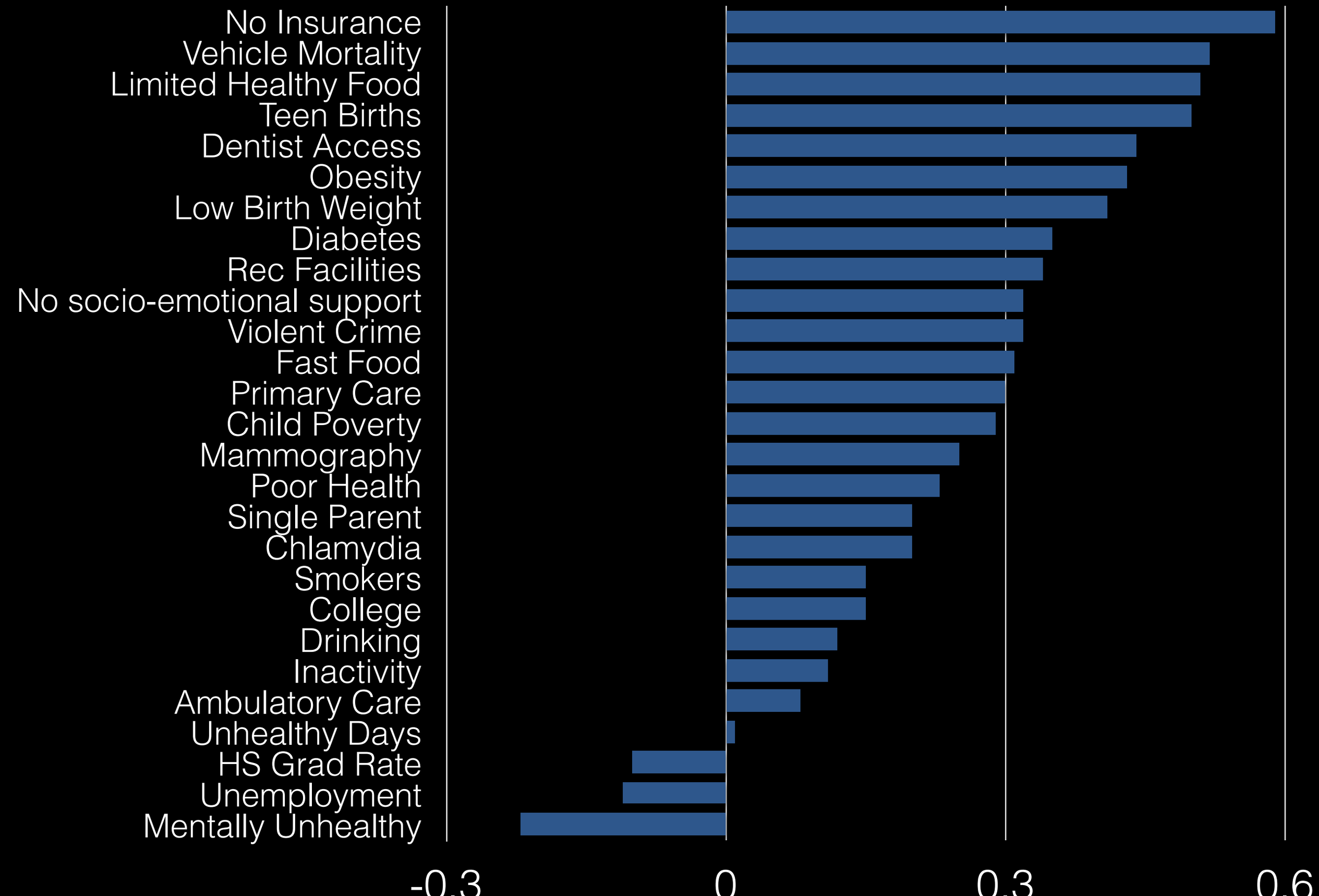
. . .

Evaluation

- 5-fold cross-validation (states grouped together)
- Held-out correlation (*r*)
- Symmetric Mean Absolute Percentage Error (SMAPE)

$$\frac{\sum_i |y_i - \hat{y}_i|}{\sum_i y_i + \hat{y}_i}$$

Held-out Correlation



Held-out Correlation

Compared to what?

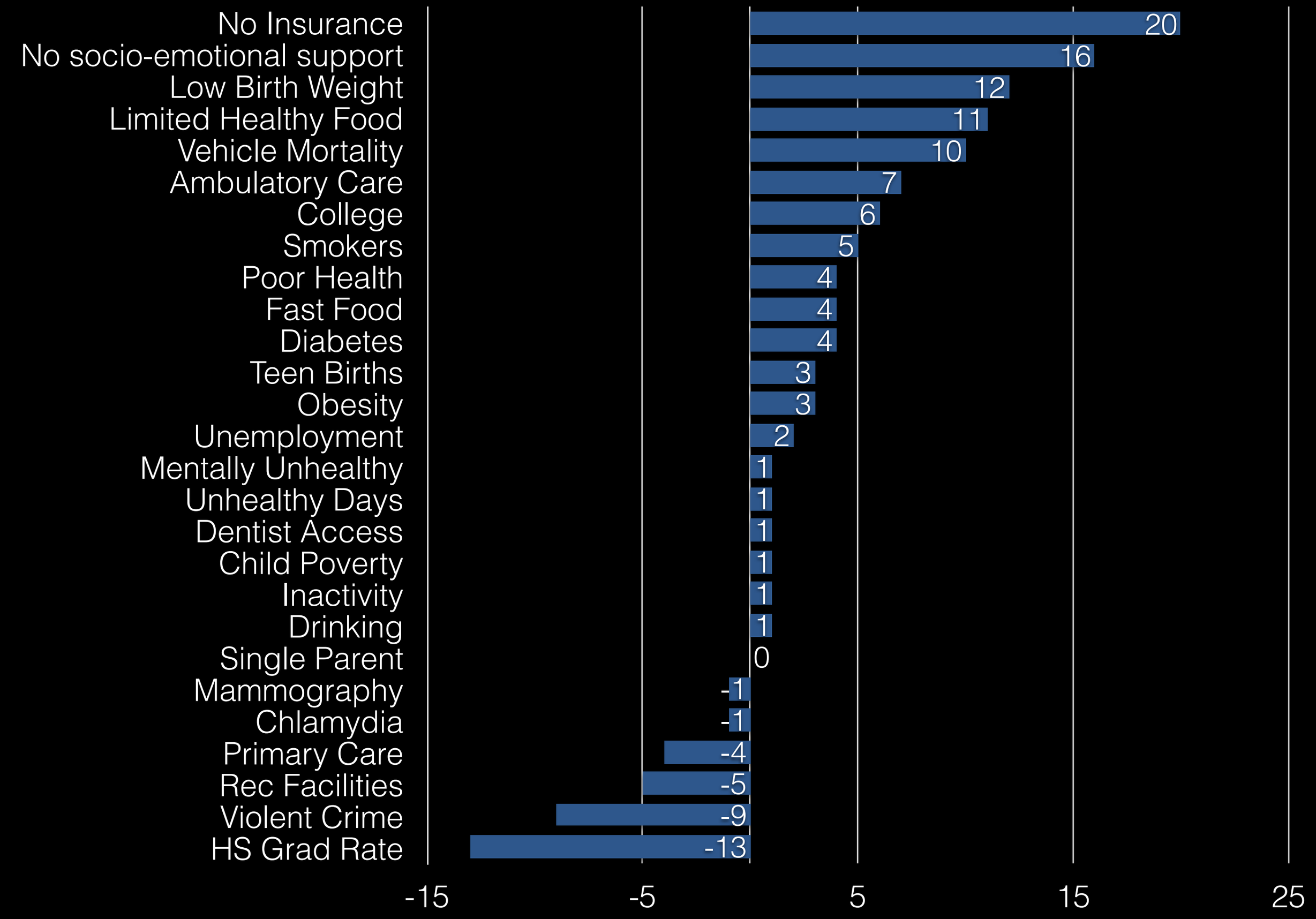
Compared to what?

Demographic variable model
< 18
65 and over
Female
Afro-Hispanic
Median Income

- Baseline in previous work [Schwartz et al. 2013]
- Core demographics in CDC data collection

Combine with Twitter variables and compare

% SMAPE Improvement over Baseline



Which word categories matter most?

Which word categories matter most?

$$y_i = \beta_0 + \beta_1 \text{Category1} + \beta_2 \text{18yo} + \beta_3 \text{65yo} + \beta_4 \text{Female} + \beta_5 \text{Afro-Hisp} + \beta_6 \text{Income} + \epsilon$$

Spatial two stage least squares regression [Anselin1988] to account for spatial auto-correlation.

More socio-emotional support

Religious references: *god, jesus*
e.g. "I love god and family."

References to others: *you, we, they*

Positive feelings: *love, happy, smile*
e.g. "Seeing people smile makes me happy."

Less socio-emotional support

TV: *show, TV, movies*

More teen births

family/love: *mom, son, daddy*

positive relationship: *love, friend*

<u>Less unemployment</u>

Inhibition: *stop, waiting, hold*
e.g. "set your goals high and don't stop until you get there"

Jobs: *work.* e.g.,*"do I have to go to work today?"*

	Population A	Population B
Obesity Rate	**?**	**?**
% Afro-Hispanic	**45.3%**	**51.8%**
Median Income	**$39K**	**$42K**
“tired”, “bored”	**7%**	**3%**
profanity	**12%**	**6%**

	Population A	Population B
Obesity Rate	**34%**	**25%**
% Afro-Hispanic	**45.3%**	**51.8%**
Median Income	**$39K**	**$42K**
“tired”, “bored”	**7%**	**3%**
profanity	**12%**	**6%**

	Wayne Cty	Kings Cty
Obesity Rate	**34%**	**25%**
% Afro-Hispanic	**45.3%**	**51.8%**
Median Income	**$39K**	**$42K**
“tired”, “bored”	**7%**	**3%**
profanity	**12%**	**6%**

Selection Bias

Selection Bias

Twitter is not a representative sample of the population.

Geolocatable Twitter users are not a representative sample of Twitter users.

Reducing Selection Bias

1. Infer demographics of Twitter users.
2. Compare to known demographics.
3. Reweight users proportional to mismatch.

Reducing Selection Bias

1. Infer demographics of Twitter users.
2. Compare to known demographics.
3. Reweight users proportional to mismatch.

San Diego:
37% Afro-Hispanic
31% Twitter users Afro-Hispanic
Weight = .37 / .31

Reducing Selection Bias

1. Infer demographics of Twitter users

Gender:

Name	% Females
Mary	2.629
Patricia	1.073
Linda	1.035
…	…

Name	% Males
James	3.318
John	3.271
Robert	3.143
…	…

Reducing Selection Bias

1. Infer demographics of Twitter users

Gender:

United States Census Bureau

Name		Name	Males
Mary			.318
Patricia	1.073	John	3.271
Linda	1.035	Robert	3.143
…	…	…	…

Reducing Selection Bias

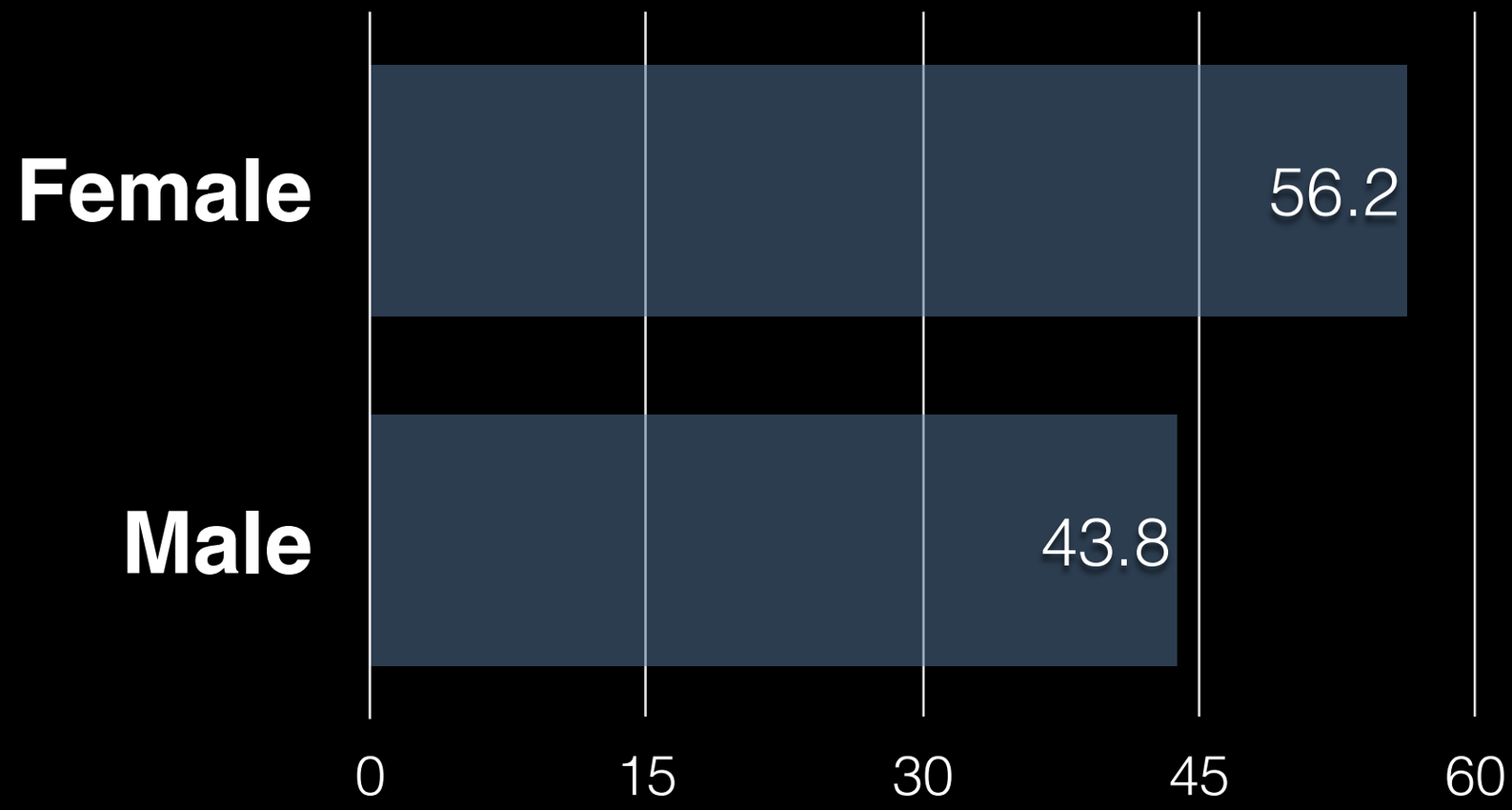
1. Infer demographics of Twitter users

Gender:

$$\%\ \text{Female (est.)} = \frac{\text{female\_names}}{(\text{female\_names} + \text{male\_names})}$$

Reducing Selection Bias

1. Infer demographics of Twitter users



48% of all users were assigned a gender

Reducing Selection Bias

1. **Infer demographics of Twitter users**

Race/ethnicity:

Reducing Selection Bias

1. Infer demographics of Twitter users

Race/ethnicity:

Train a classifier based on terms in profile description.
See [Mohammady & Culotta 2014]

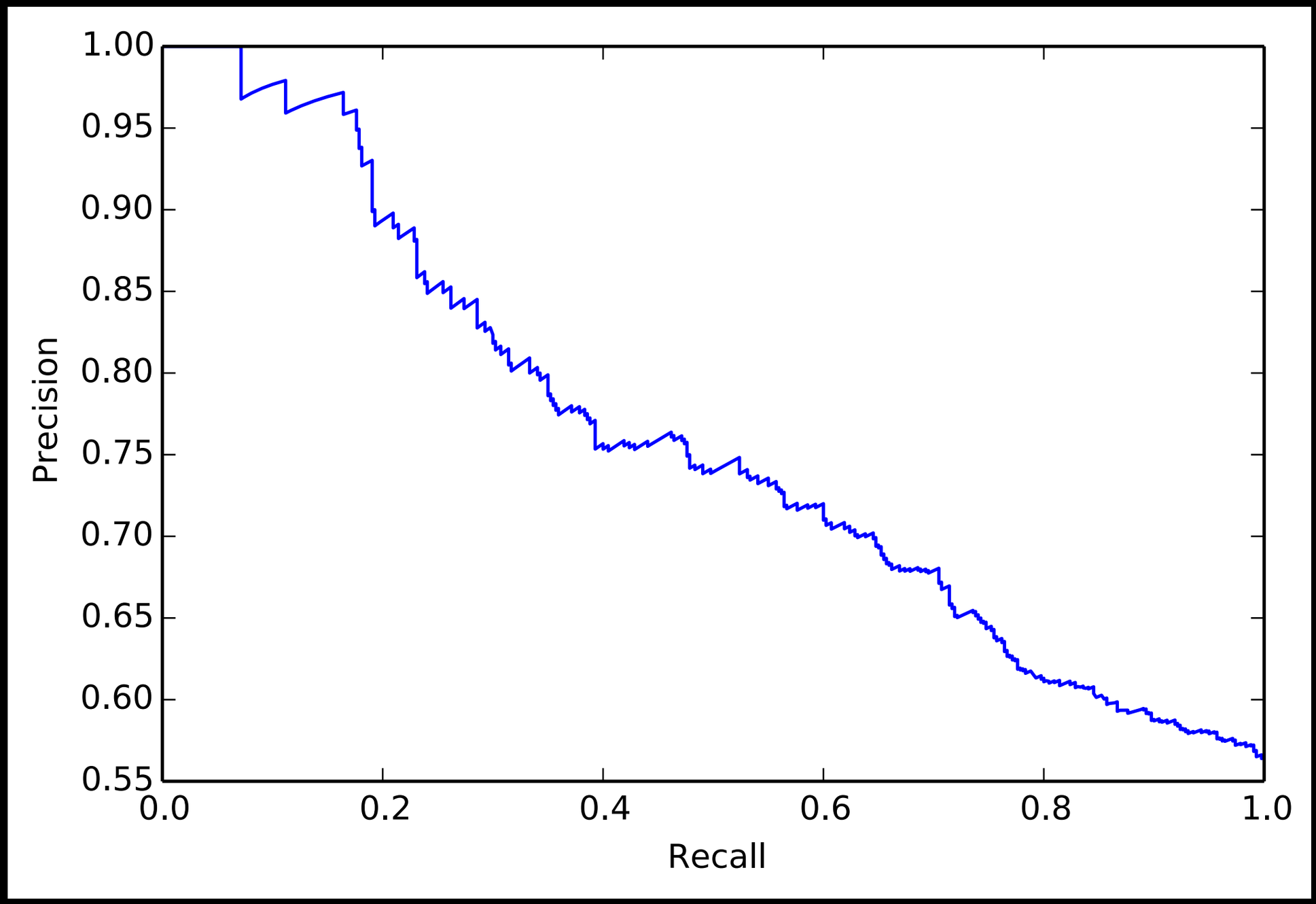
Classify as African-American, Hispanic, or White

Logistic regression trained on 770 profiles

Reducing Selection Bias

1. Infer demographics of Twitter users

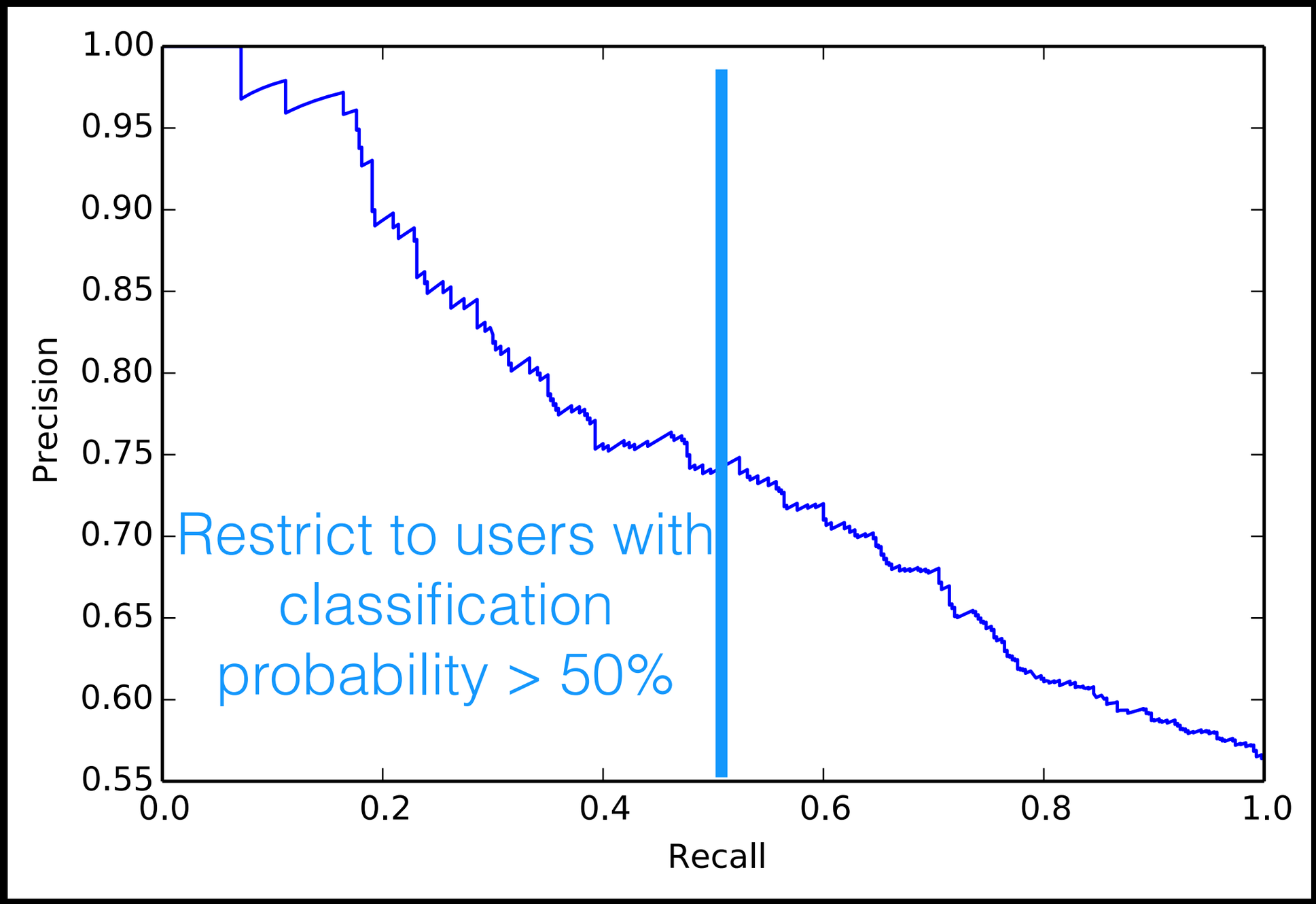
Race/ethnicity:



Reducing Selection Bias

1. Infer demographics of Twitter users

Race/ethnicity:



Reducing Selection Bias

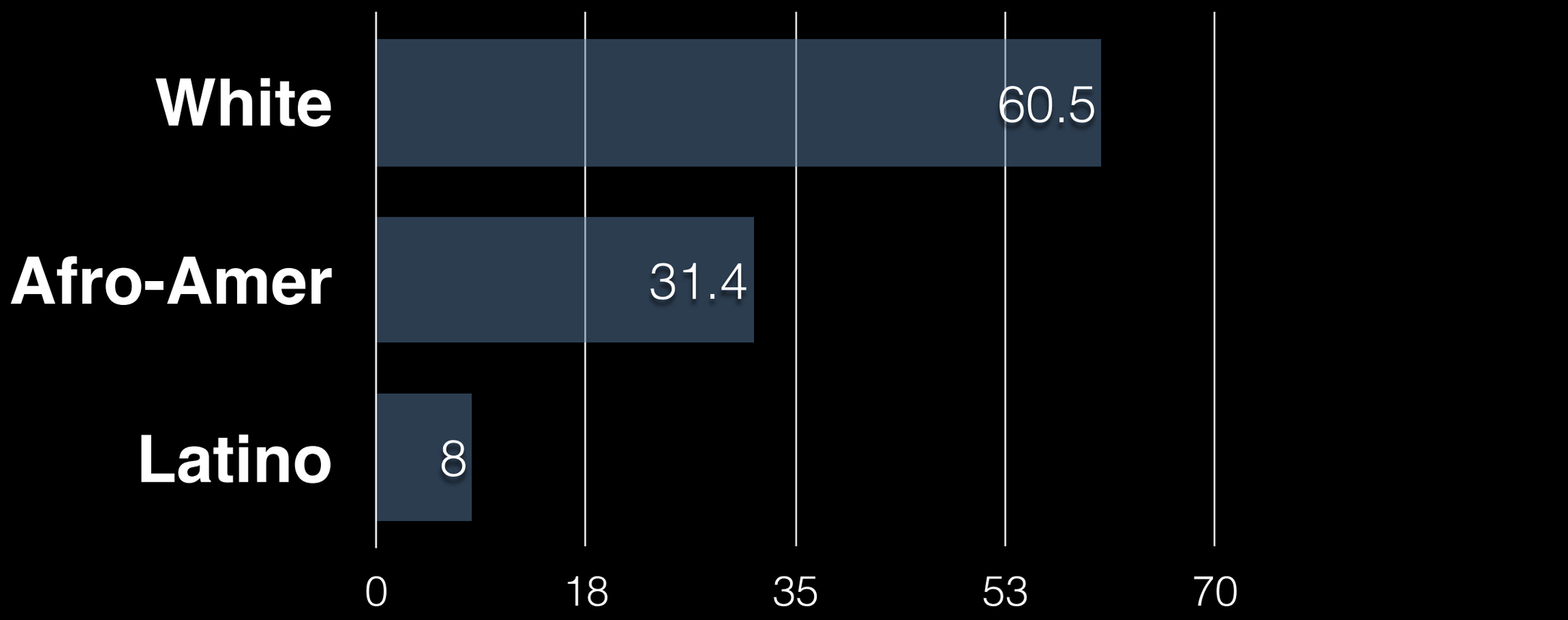
1. Infer demographics of Twitter users

Race/ethnicity:

$$\%\ \text{Afro-Hispanic (est.)} = \frac{\text{afro\_hispanic}}{(\text{afro\_hispanic} + \text{white})}$$

Reducing Selection Bias

1. Infer demographics of Twitter users



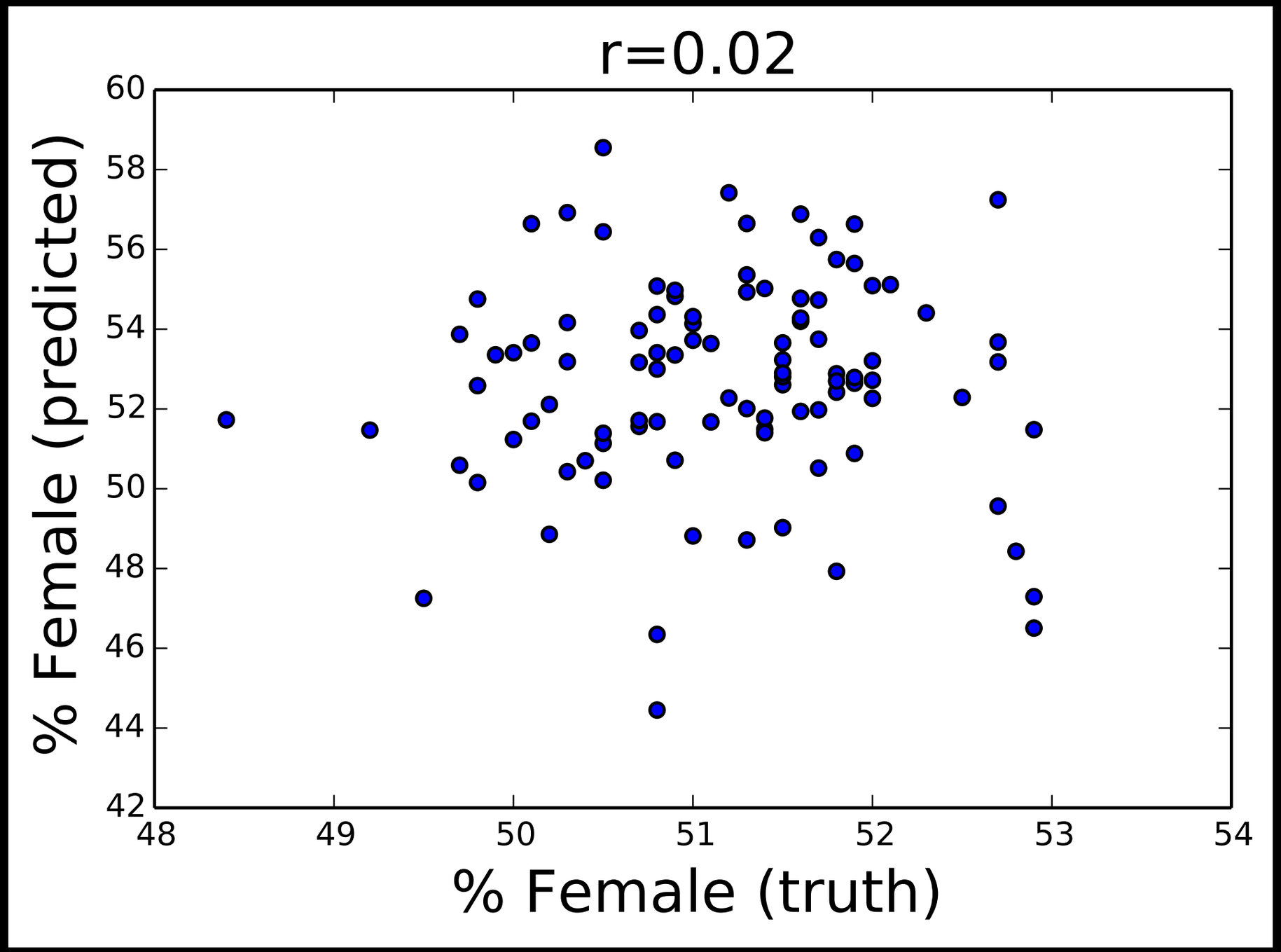
32% of all users were assigned a race/ethnicity

Reducing Selection Bias

2. Compare to known demographics

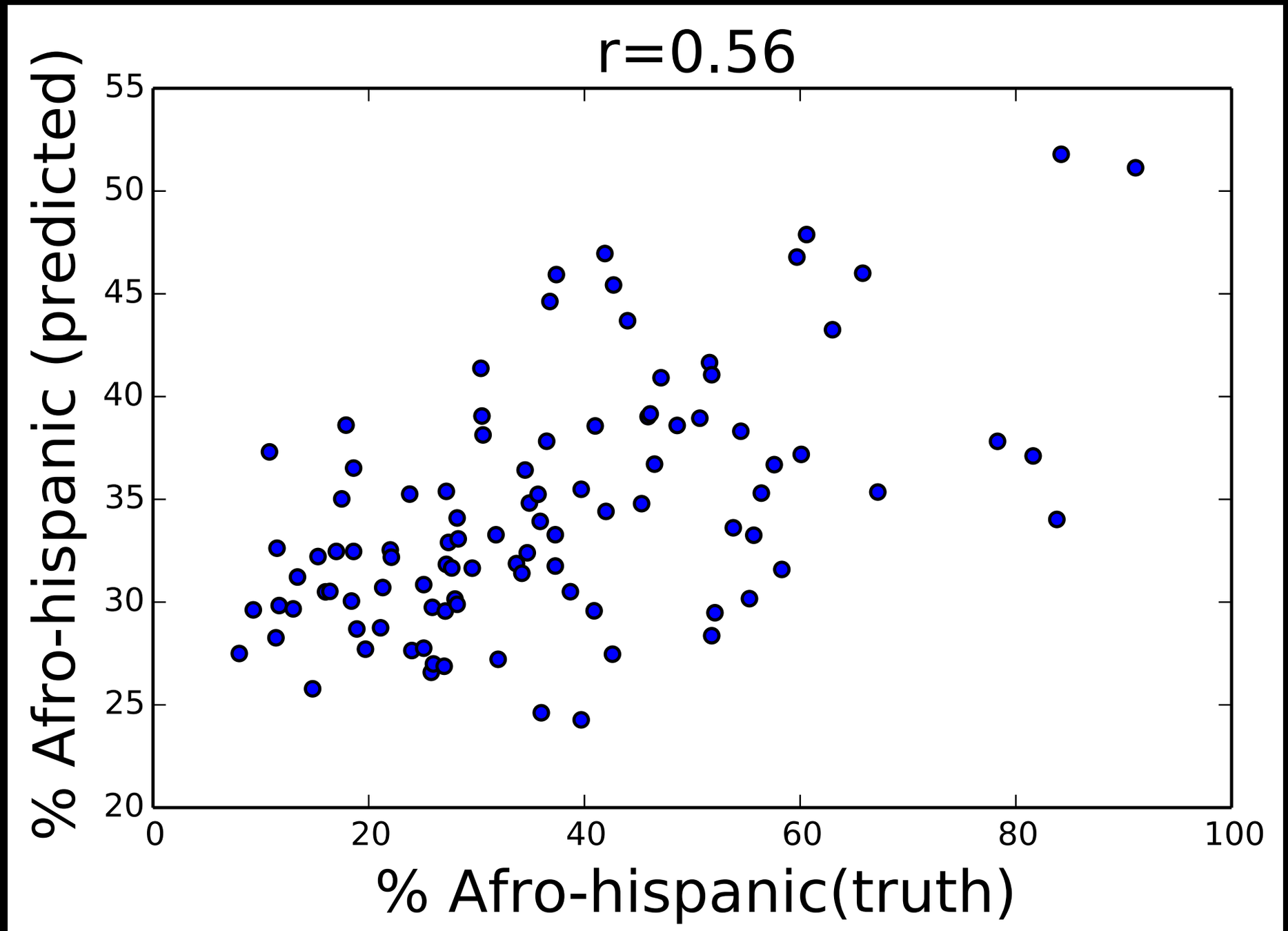
Reducing Selection Bias

2. Compare to known demographics



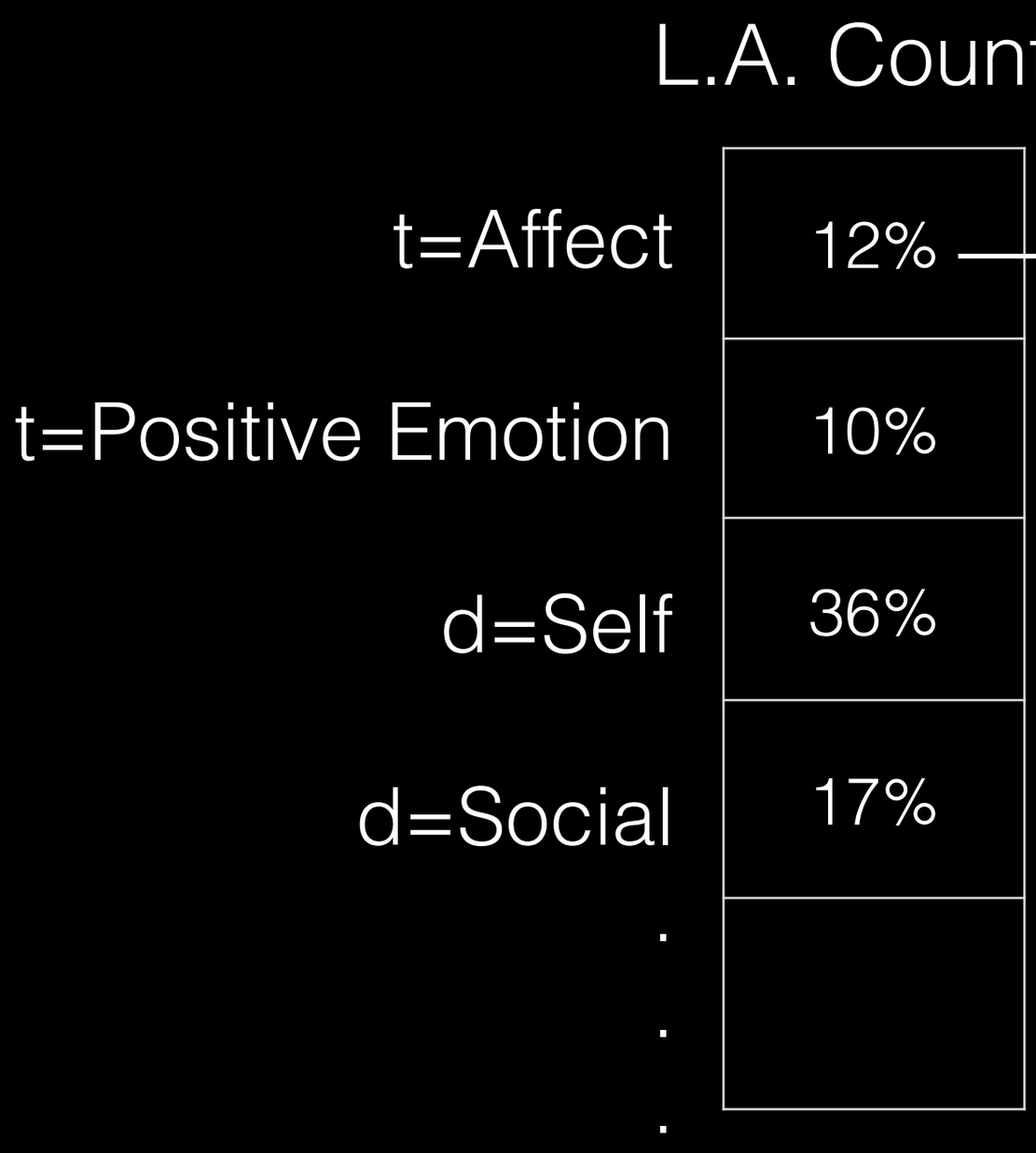
Reducing Selection Bias

2. Compare to known demographics



Reducing Selection Bias

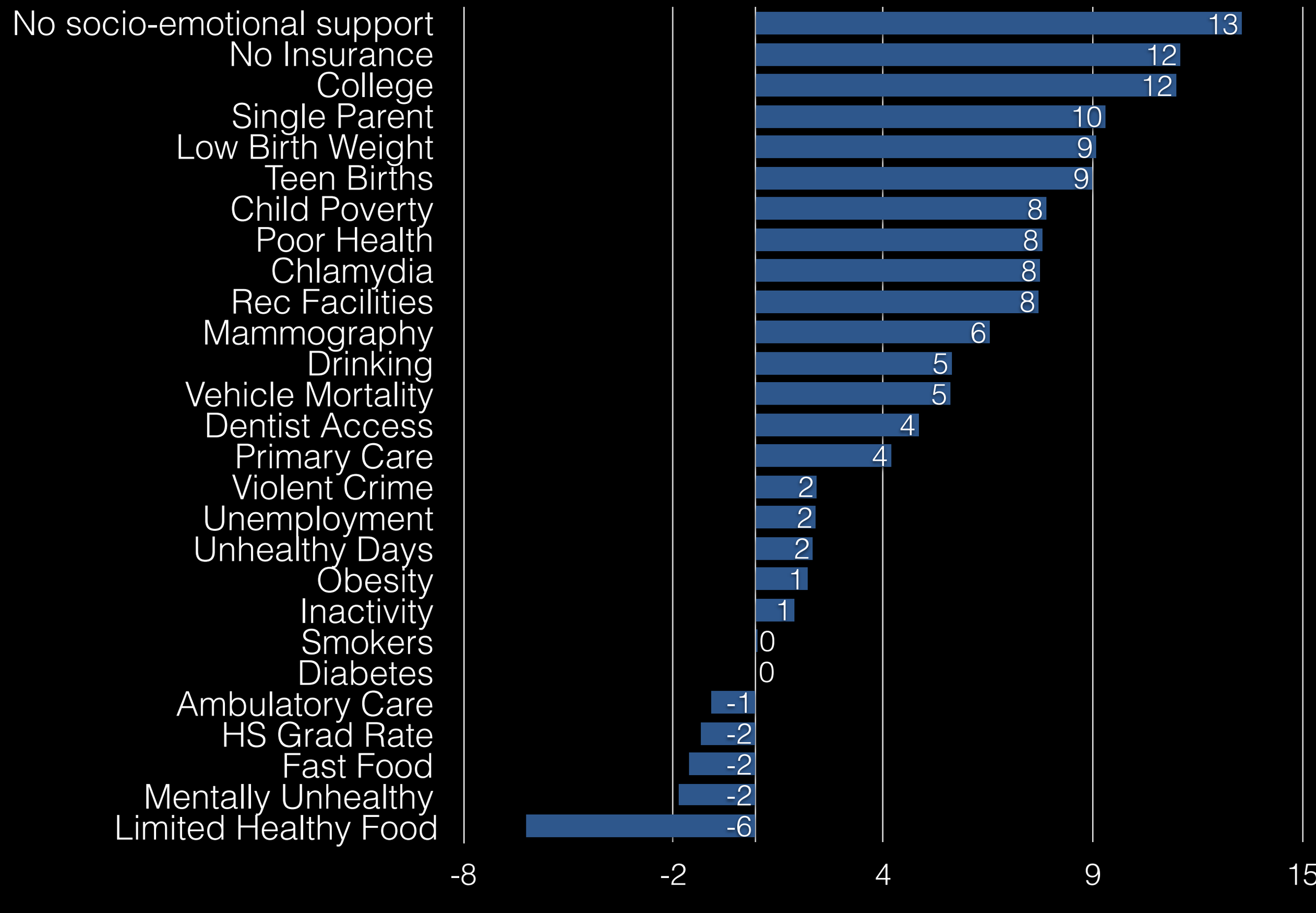
3. Reweight users proportional to mismatch



% of users who mention a word from this category

$$\frac{\sum_i w_i x_i}{\sum_i w_i}$$

% SMAPE Improvement after reweighting



Conclusions / Future Work

- Twitter can provide a more nuanced way to characterize a population than common demographics.
- Stat. sig. held-out correlation for 6 health statistics
- Improves a simple demographic model for 20 health statistics
- Adjusting for selection bias improves results, even with noise
- Code: https://github.com/tapilab/twcounty

Outline

1. Tracking influenza-like illness rate
2. Tracking alcohol sales
3. Tracking community health
4. Inferring the origin of social media messages

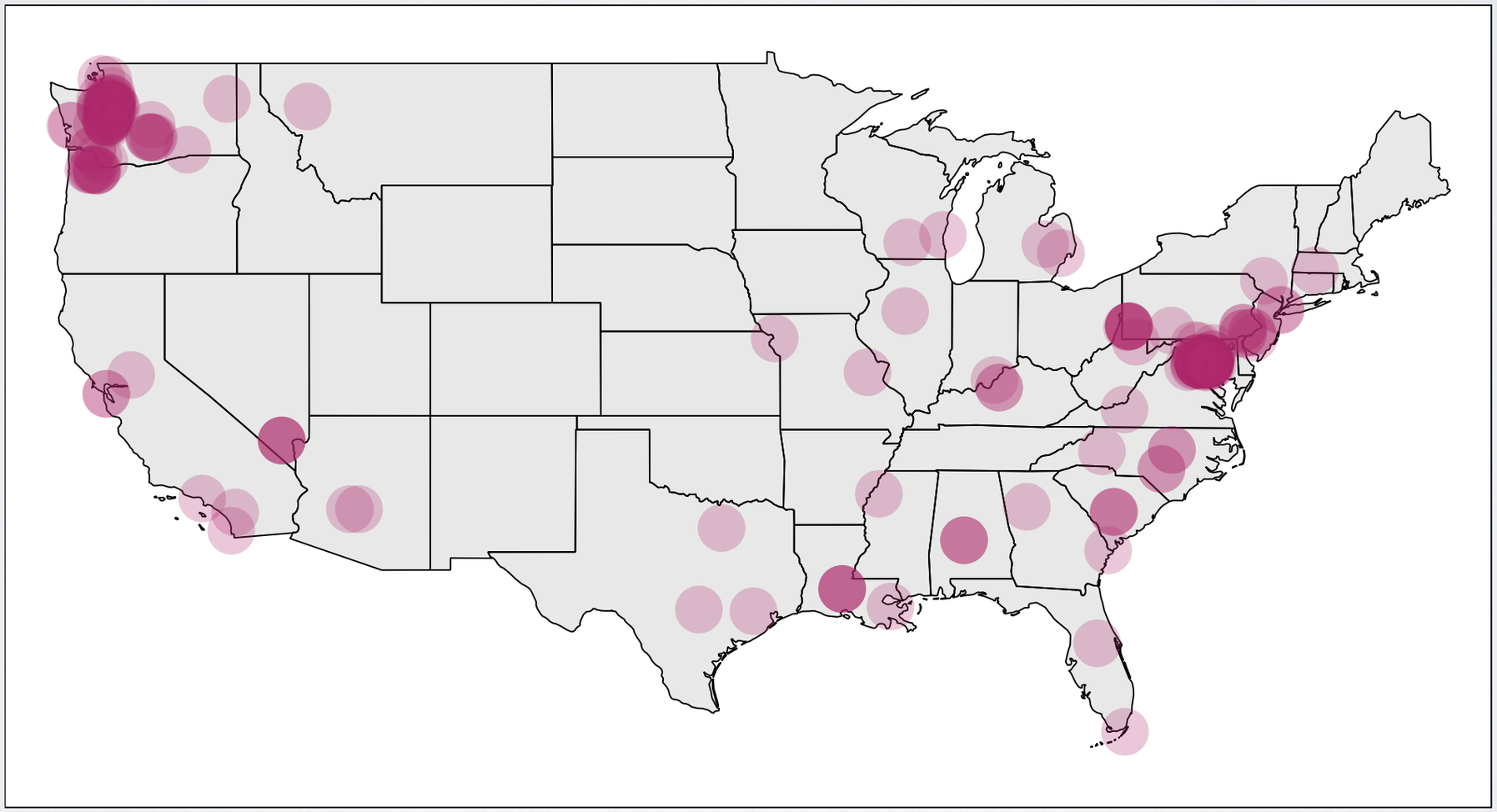
Probabilistic Estimates of Tweet Origins

[Priedhorsky & Culotta 2014]

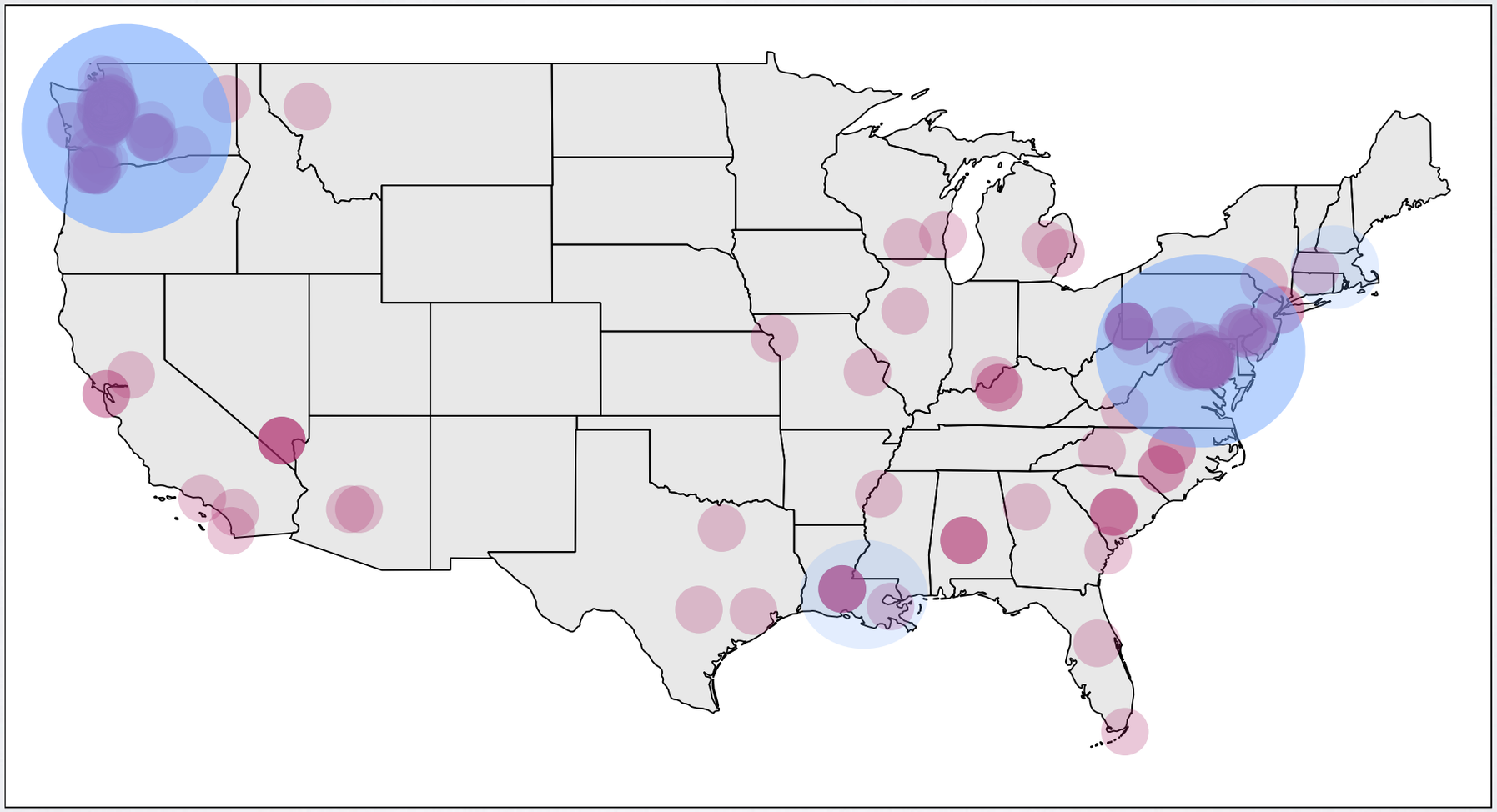
Given a tweet, infer a distribution over its origin.

Why?

Tweets mentioning washington



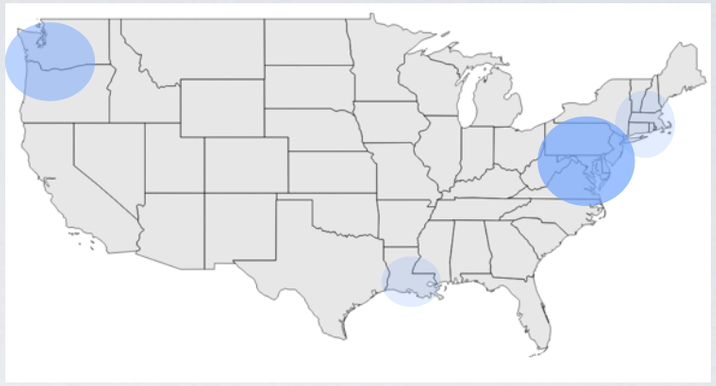
Tweets mentioning washington



Given many geo-tagged tweets, how would you learn a model to predict location of a new tweet?

washington

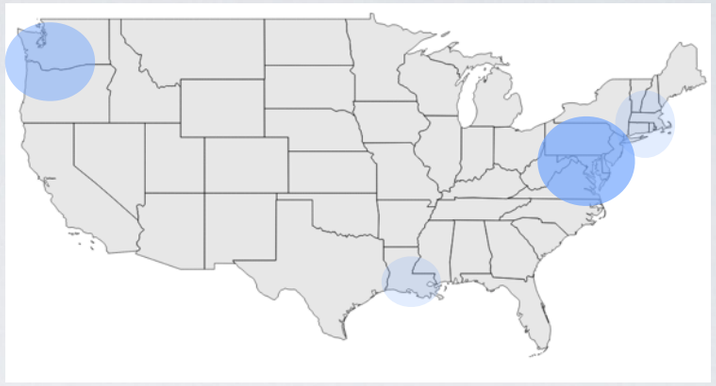
+

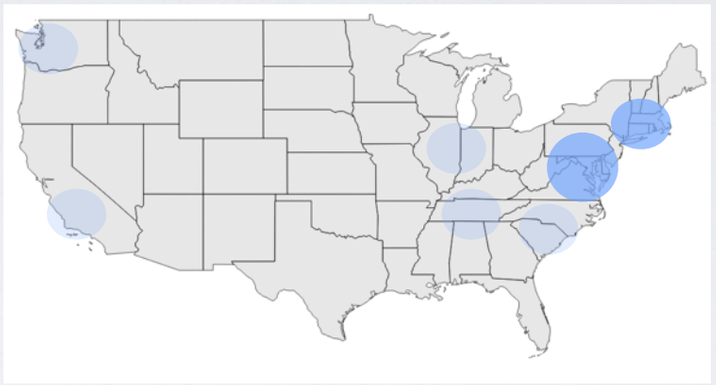


washington

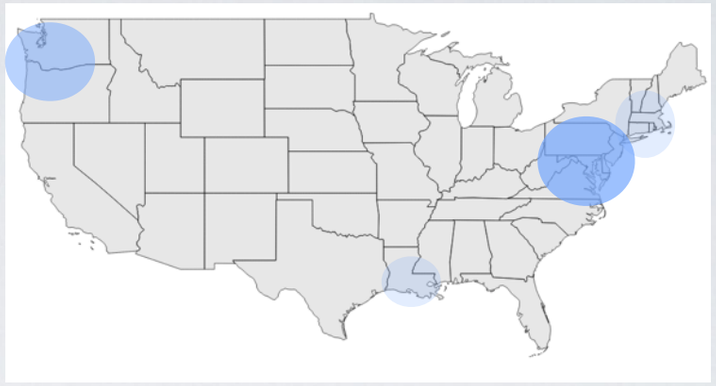
+

wicked



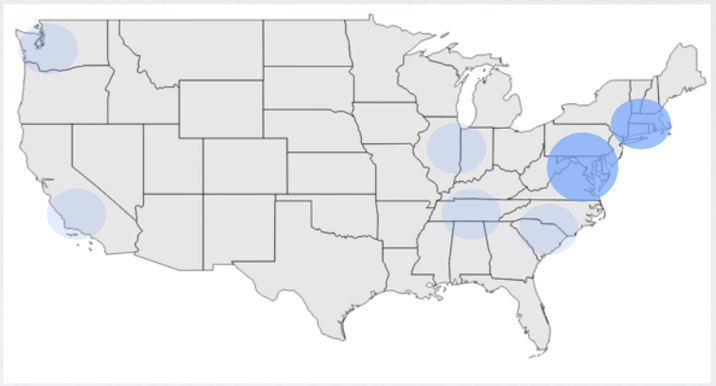


washington



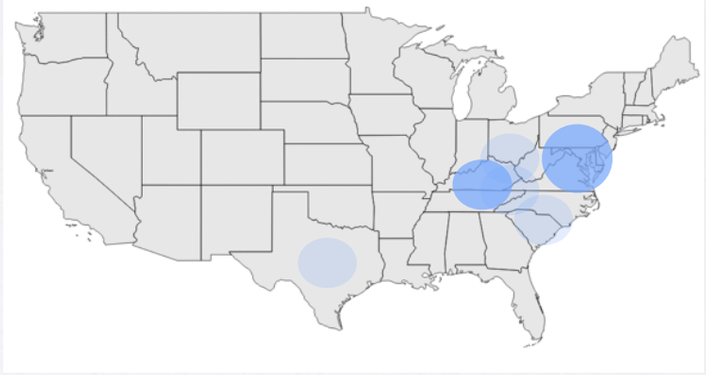
+

wicked

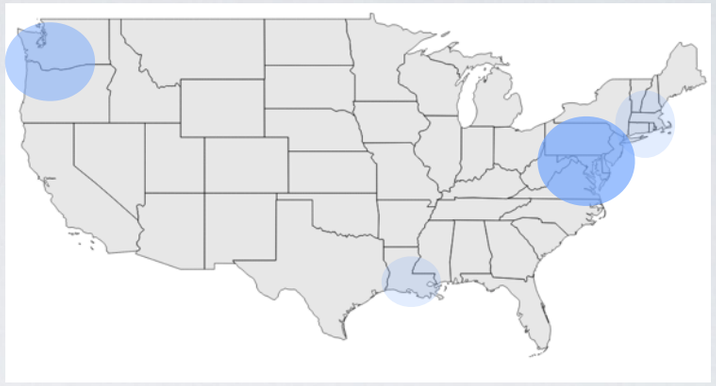


+

Howard

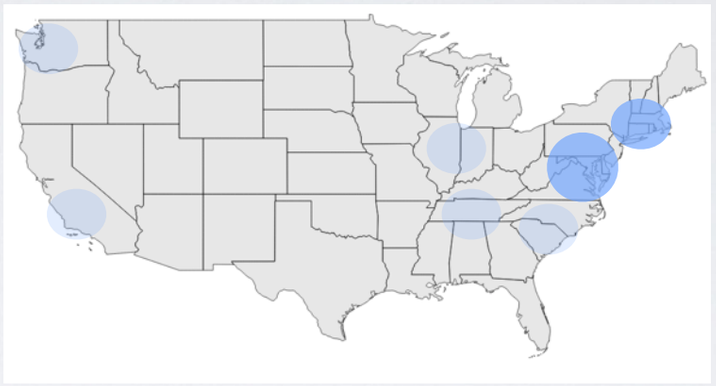


washington



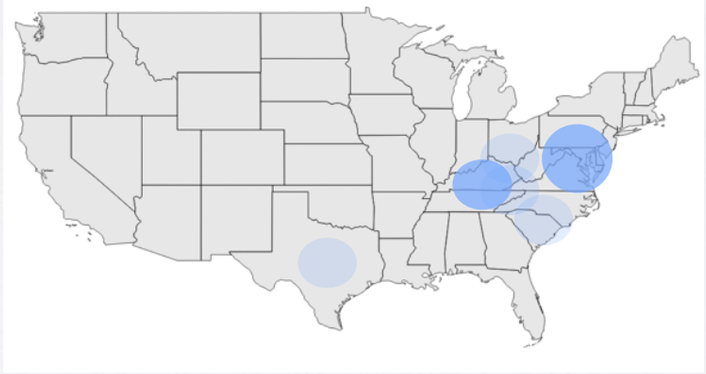
+

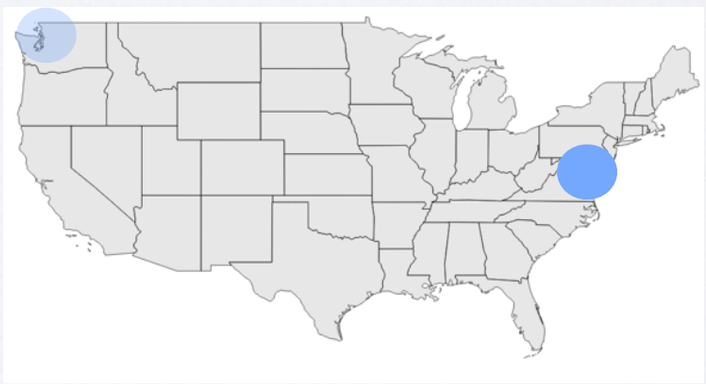
wicked



+

Howard





Geographical Mixtures of Gaussians

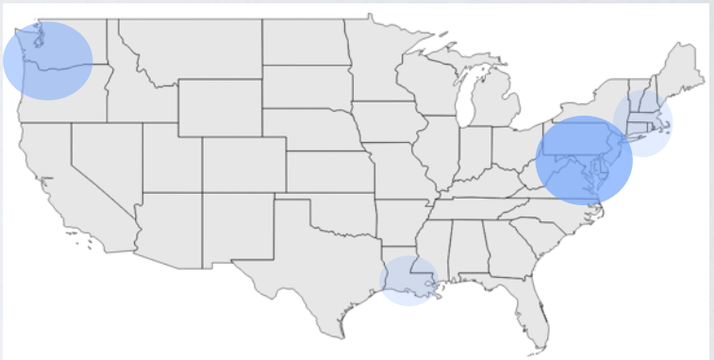
ngram mixture

$$g(y|w_j) = \sum_{k=1}^{r} \pi_k^j \mathcal{N}(y|\mu_k^j, \sigma_k^j)$$

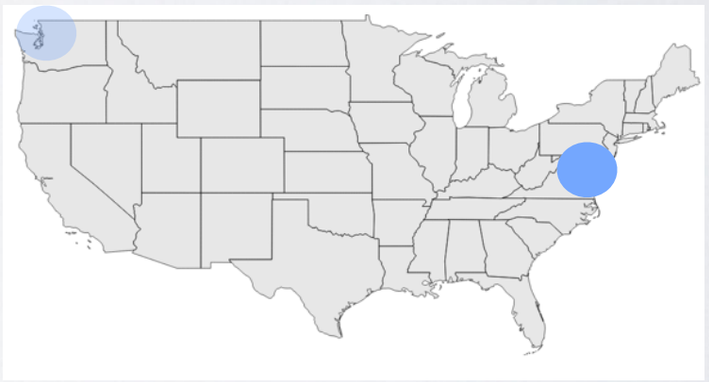
tweet mixture of mixtures

$$f(y|m) = \sum_{w_j \in m} \delta_j g(y|w_j)$$

washington



washington + wicked + Howard

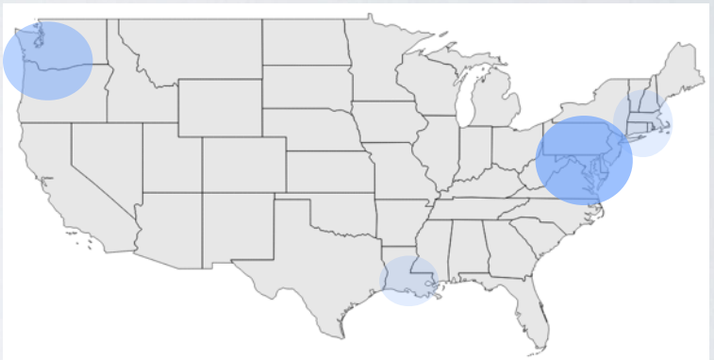


Geographical Mixtures of Gaussians

ngram mixture

$$g(y|w_j) = \sum_{k=1}^{r} \pi_k^j \mathcal{N}(y|\mu_k^j, \sigma_k^j)$$

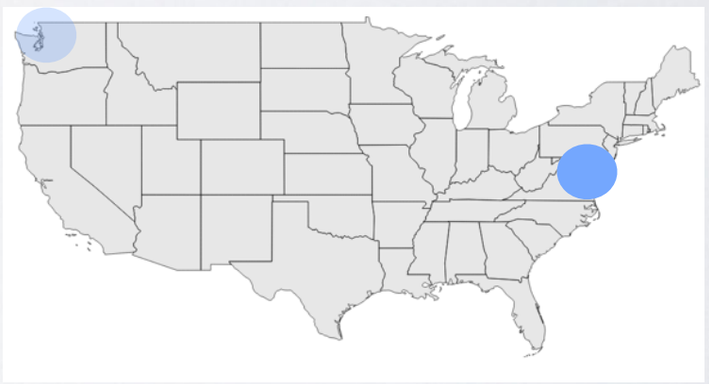
washington



tweet mixture of mixtures

$$f(y|m) = \sum_{w_j \in m} \delta_j g(y|w_j)$$

washington + wicked + Howard



How informative is this ngram?
"wicked" versus "washington"

Optimizing ngram weights δ_j

$$\delta_j^\theta = \frac{1}{1 + e^{-\theta_j}}$$

$$E(D, \theta) = \sum_{i=1}^{|D|} \frac{\sum_{w_j \in m_i} e_{ij} \delta_j^\theta}{\sum_{w_j \in m_i} \delta_j^\theta} + \frac{\lambda}{2} \|\theta\|^2$$

$$\theta^* \leftarrow \text{argmin}_\theta E(D, \theta)$$

Optimizing ngram weights δ_j

$$\delta_j^\theta = \frac{1}{1 + e^{-\theta_j}}$$

distance from ngram mixture j to true origin

$$E(D, \theta) = \sum_{i=1}^{|D|} \frac{\sum_{w_j \in m_i} e_{ij} \delta_j^\theta}{\sum_{w_j \in m_i} \delta_j^\theta} + \frac{\lambda}{2} \|\theta\|^2$$

$$\theta^* \leftarrow \text{argmin}_\theta E(D, \theta)$$

Optimizing ngram weights δ_j

$$\delta_j^\theta = \frac{1}{1 + e^{-\theta_j}}$$

distance from ngram mixture j to true origin

$$E(D, \theta) = \sum_{i=1}^{|D|} \frac{\sum_{w_j \in m_i} e_{ij} \delta_j^\theta}{\sum_{w_j \in m_i} \delta_j^\theta} + \frac{\lambda}{2} \|\theta\|^2$$

$$\theta^* \leftarrow \text{argmin}_\theta E(D, \theta)$$

gradient descent with L-BFGS

Optimizing ngram weights δ_j

$$\delta_j \propto \frac{1}{\left(\frac{1}{N_j}\sum_i e_{ij}\right)^{\alpha}}$$

Optimizing ngram weights δ_j

$$\delta_j \propto \frac{1}{\left(\frac{1}{N_j}\sum_i e_{ij}\right)^{\alpha}}$$

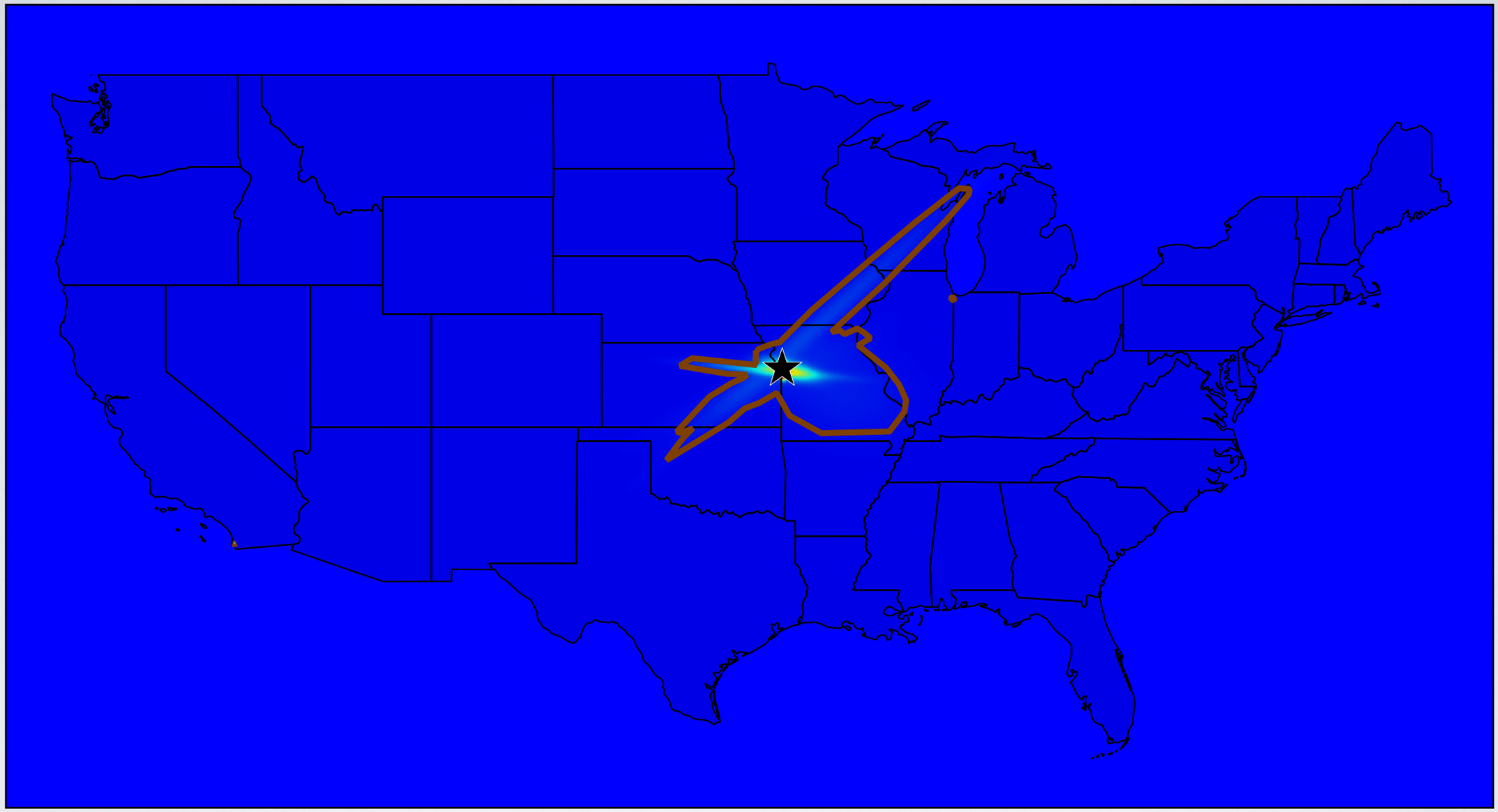
average error for ngram j

Optimizing ngram weights δ_j

$$\delta_j \propto \frac{1}{\left(\frac{1}{N_j}\sum_i e_{ij}\right)^{\alpha}}$$

tunable parameter

average error for ngram j



text	Ummmm i really love my *[lady]* @Kansas2Socal but she gonna have to ride to Stl SOLO Cuz im not missin @RichTheFactor10 for ANYONE! #TEAMRICH
language	en
location	Kansas City Mo

How do we know if it works?

Evaluating Location Estimates

Accuracy: How much mass is near the true origin?

Precision: How big is the prediction region?

Calibration: How reliable are confidence values?

Evaluating Location Estimates

Accuracy: Comprehensive Accuracy Error

Precision: How big is the prediction region?

Calibration: How reliable are confidence values?

Evaluating Location Estimates

Accuracy: Comprehensive Accuracy Error

$$\mathrm{CAE} = \int_y d(y, y^*) f(y|m) \mathrm{d}y = \mathrm{E}_f[d(y, y^*)]$$

Precision: How big is the prediction region?

Calibration: How reliable are confidence values?

Evaluating Location Estimates

Accuracy: Comprehensive Accuracy Error

$$\mathrm{CAE} = \int_y d(y, y^*) f(y|m) \mathrm{d}y = \mathrm{E}_f[d(y, y^*)]$$

Precision: Prediction Region Area at confidence β

Calibration: How reliable are confidence values?

Evaluating Location Estimates

Accuracy: Comprehensive Accuracy Error

$$\mathrm{CAE} = \int_y d(y, y^*) f(y|m) \mathrm{d}y = \mathrm{E}_f[d(y, y^*)]$$

Precision: Prediction Region Area at confidence β

$$\mathrm{PRA}_\beta = \int_{R_{f,\beta}} \mathrm{d}y$$

Calibration: How reliable are confidence values?

Evaluating Location Estimates

Accuracy: Comprehensive Accuracy Error

$$\mathrm{CAE} = \int_y d(y, y^*) f(y|m) \mathrm{d}y = \mathrm{E}_f[d(y, y^*)]$$

Precision: Prediction Region Area at confidence β

$$\mathrm{PRA}_\beta = \int_{R_{f,\beta}} \mathrm{d}y$$

Calibration: Observed Coverage at confidence β

Evaluating Location Estimates

Accuracy: Comprehensive Accuracy Error

$$\mathrm{CAE} = \int_y d(y, y^*) f(y|m) \mathrm{d}y = \mathrm{E}_f[d(y, y^*)]$$

Precision: Prediction Region Area at confidence β

$$\mathrm{PRA}_\beta = \int_{R_{f,\beta}} \mathrm{d}y$$

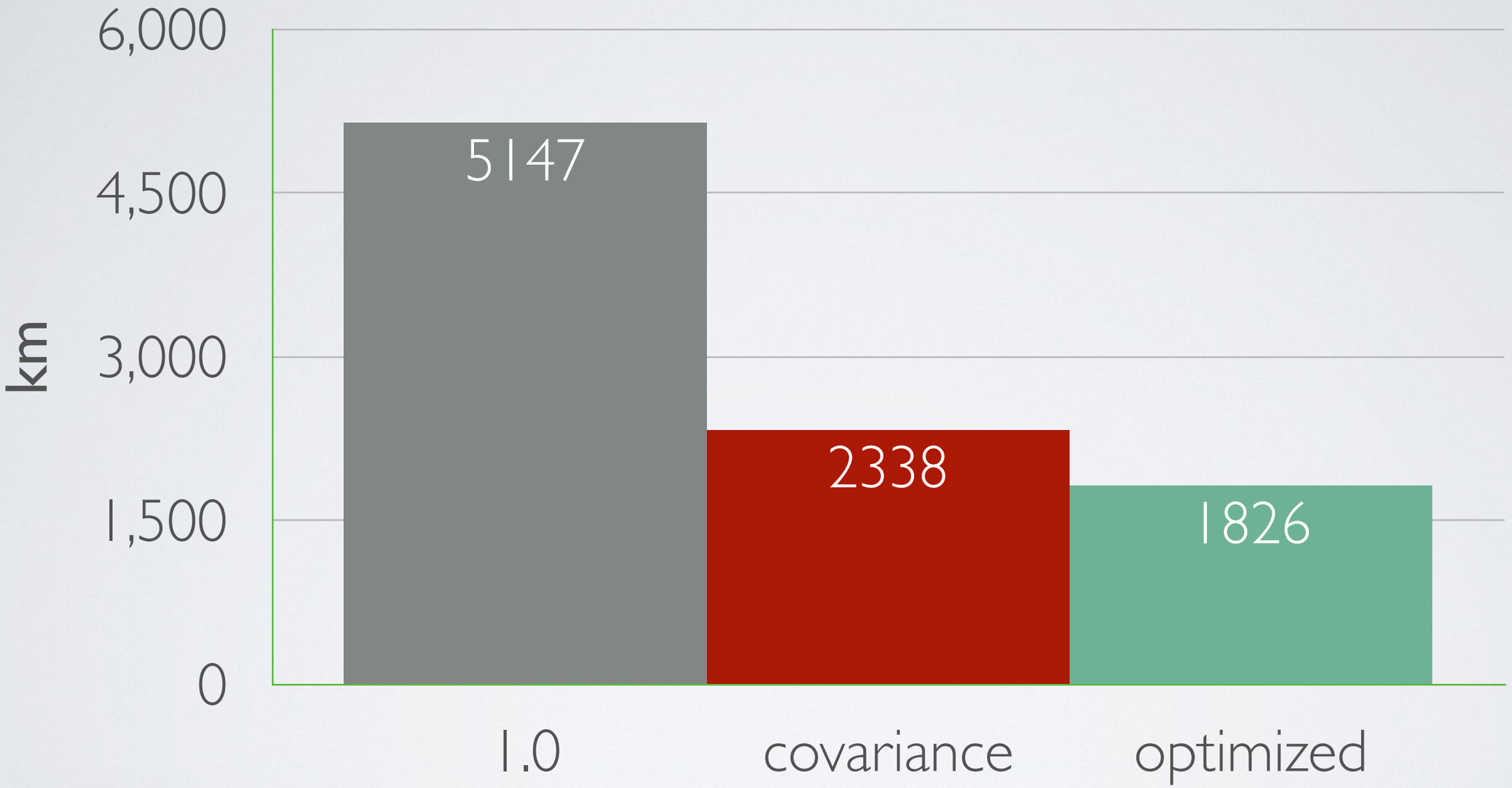
Calibration: Observed Coverage at confidence β

$$\mathrm{OC}_\beta = \frac{1}{n} \sum_{i=1}^{n} 1[y_i^* \in R_{f,\beta}^i]$$

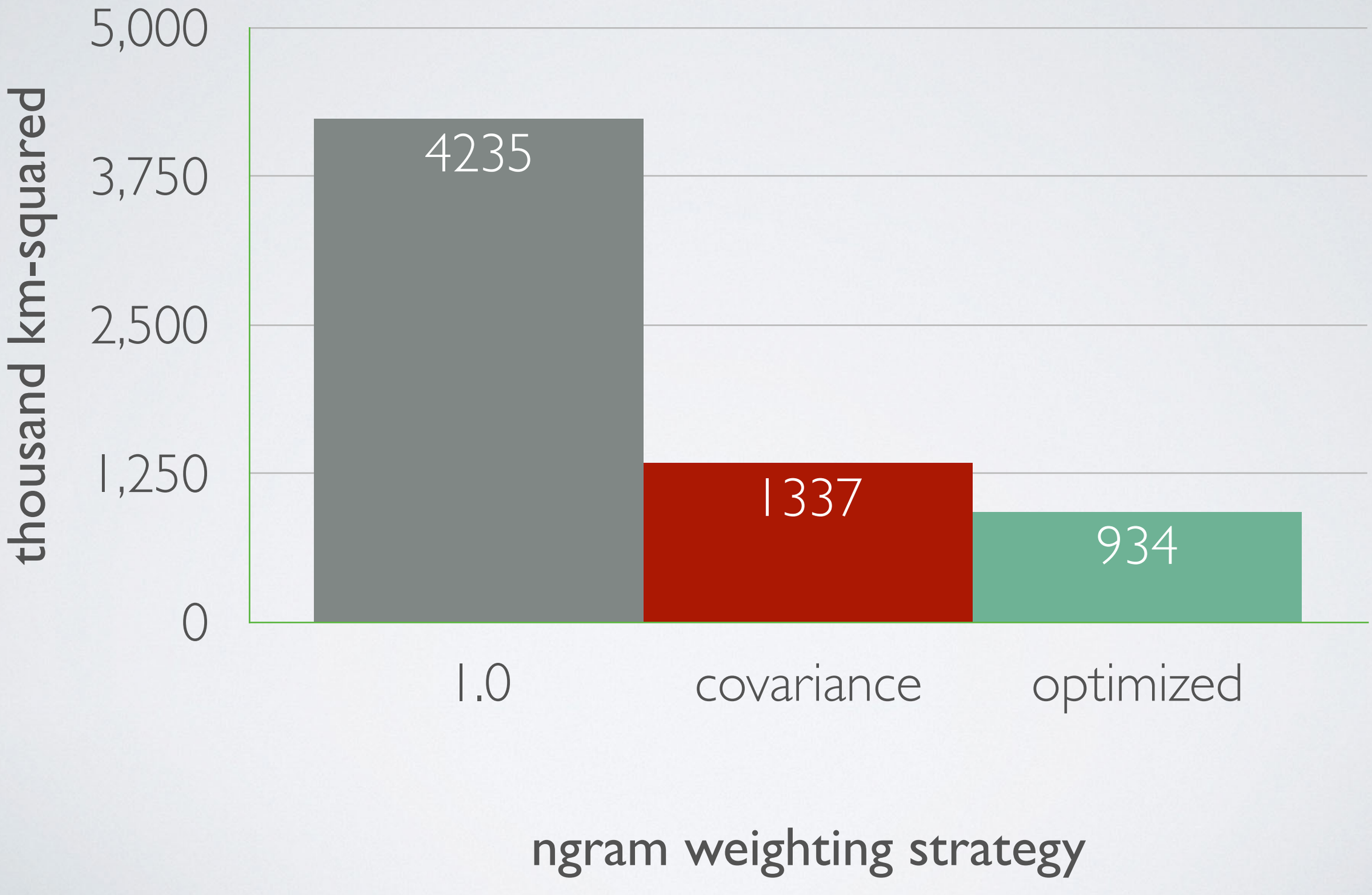
Geolocation Experiments

- 9 months of geotagged tweets (8 million)
- From anywhere on the globe
- All languages except Chinese, Thai, Lao, Cambodian, Burmese

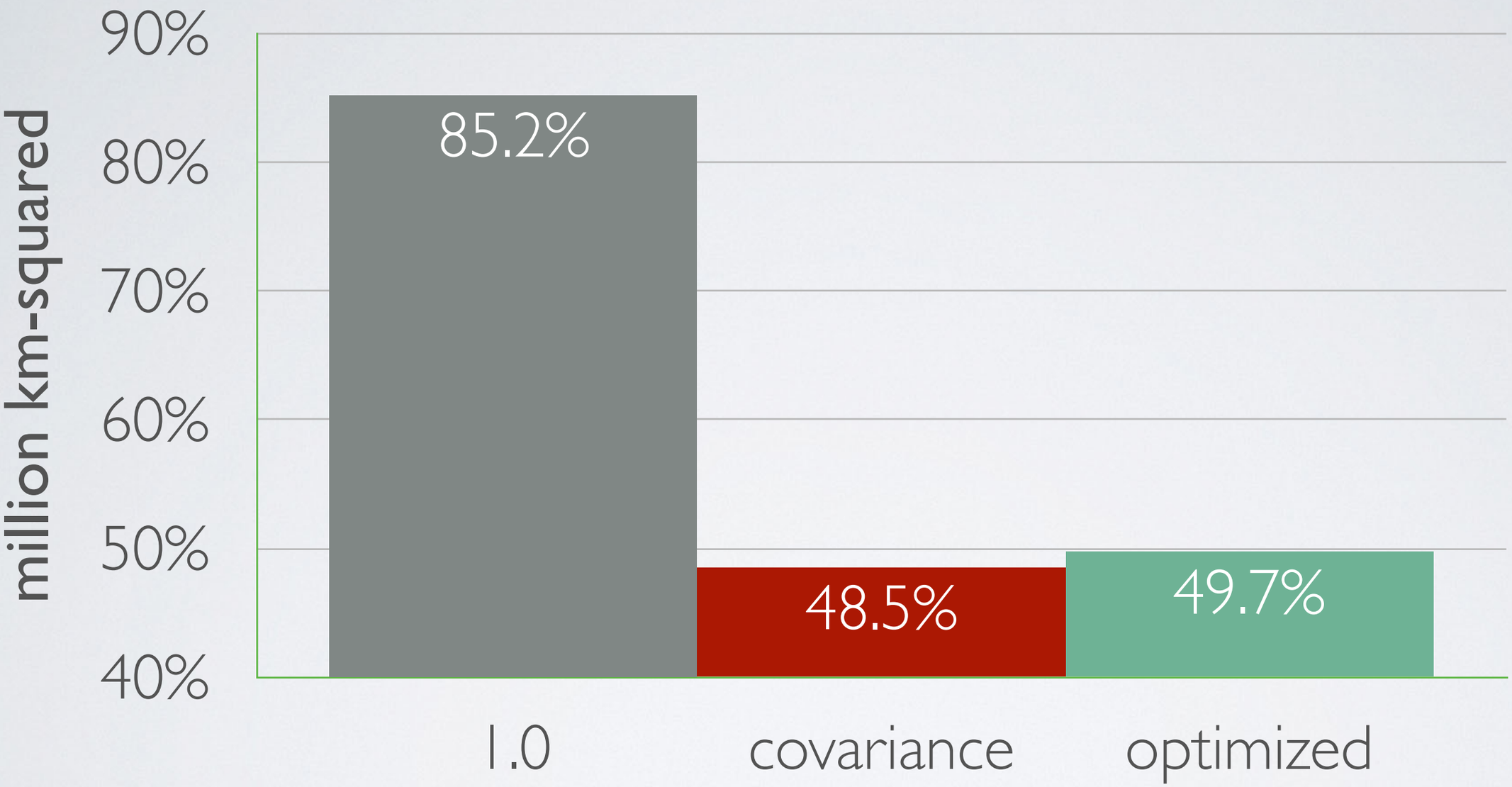
Accuracy: CAE



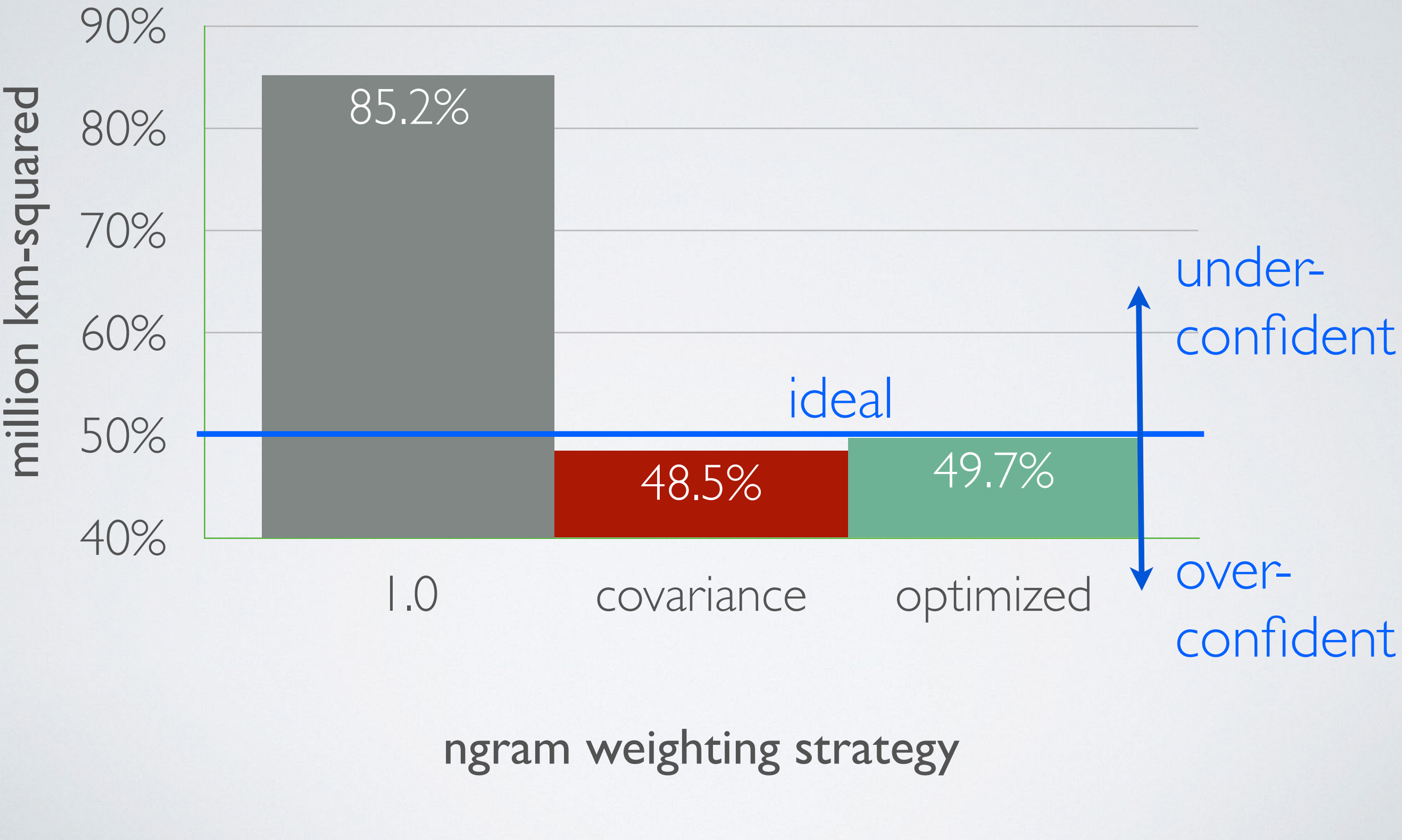
Precision: $MPRA_{50}$



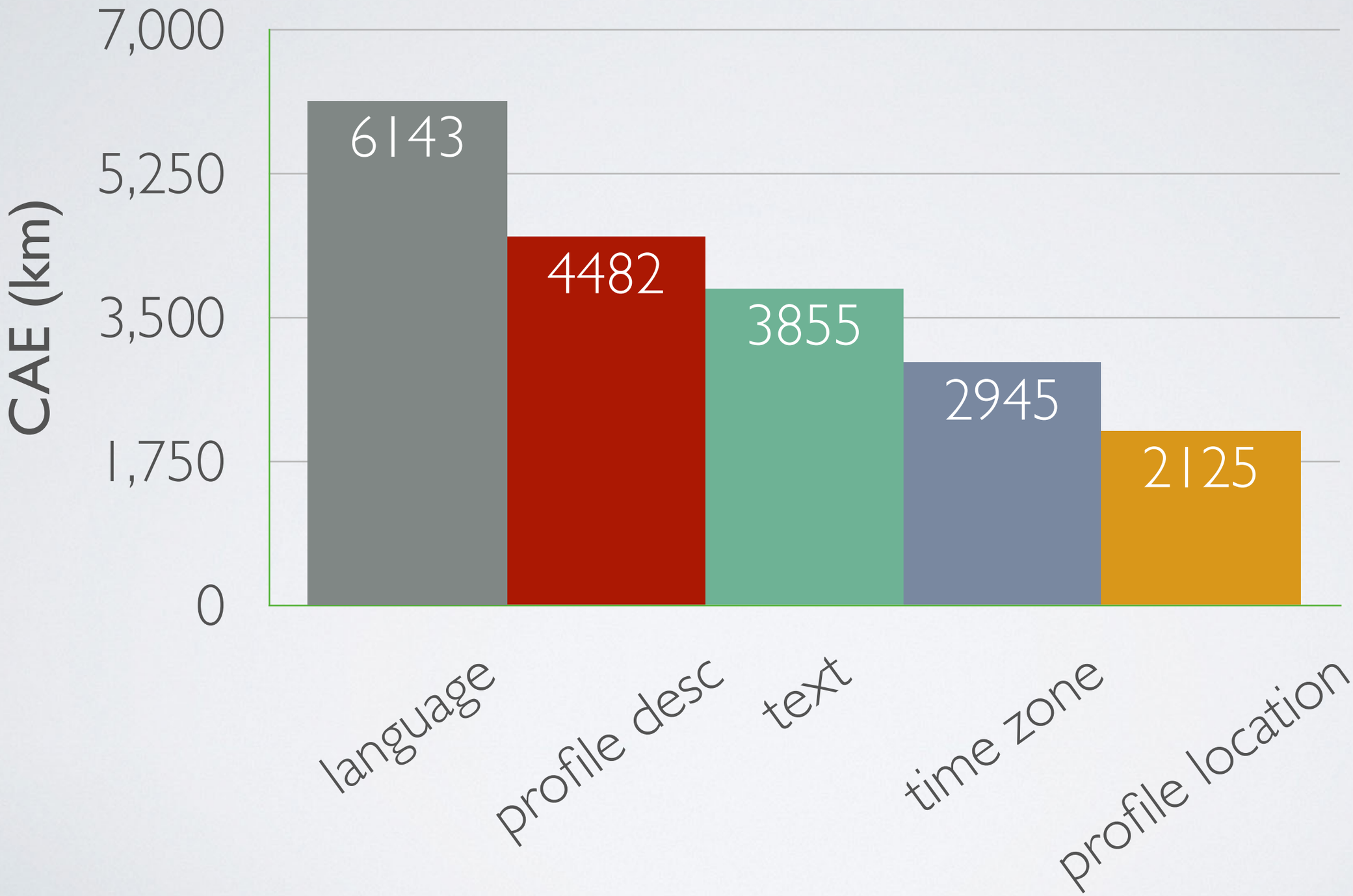
Calibration: OC_{90}



Calibration: OC_{90}



Which fields matter most?



Conclusion

- Social media offers a noisy but promising new source of data for public health
    - Positive results across many different domains
- Core issues include data collection, filtering, aggregation, regression, geolocation

Future Work

- How can we model social media to
    - detect an environmental hazard?
    - determine the source of a disease?
    - determine how social interactions affect spread of disease?

Future work

Demographic Inference

- Extend approach to gender, age, race
  - Student: Ehsan Mohammady
- Use probabilistically downstream

Time-series models

- Nowcasting --> Forecasting
- Joint models of classification and regression

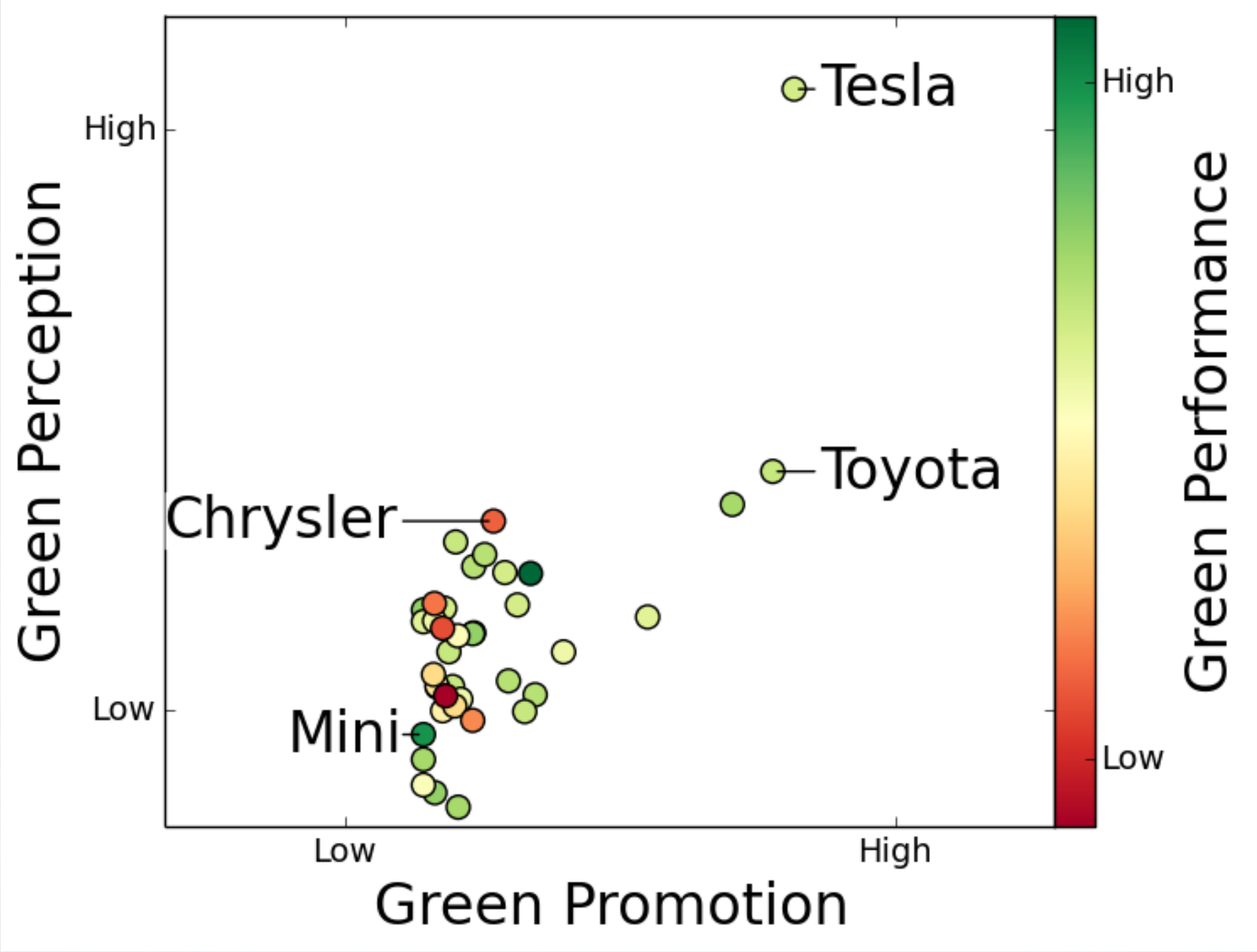
Validation

- Compare predictions with surveys of users

Hypothesis generator for social science and public health research

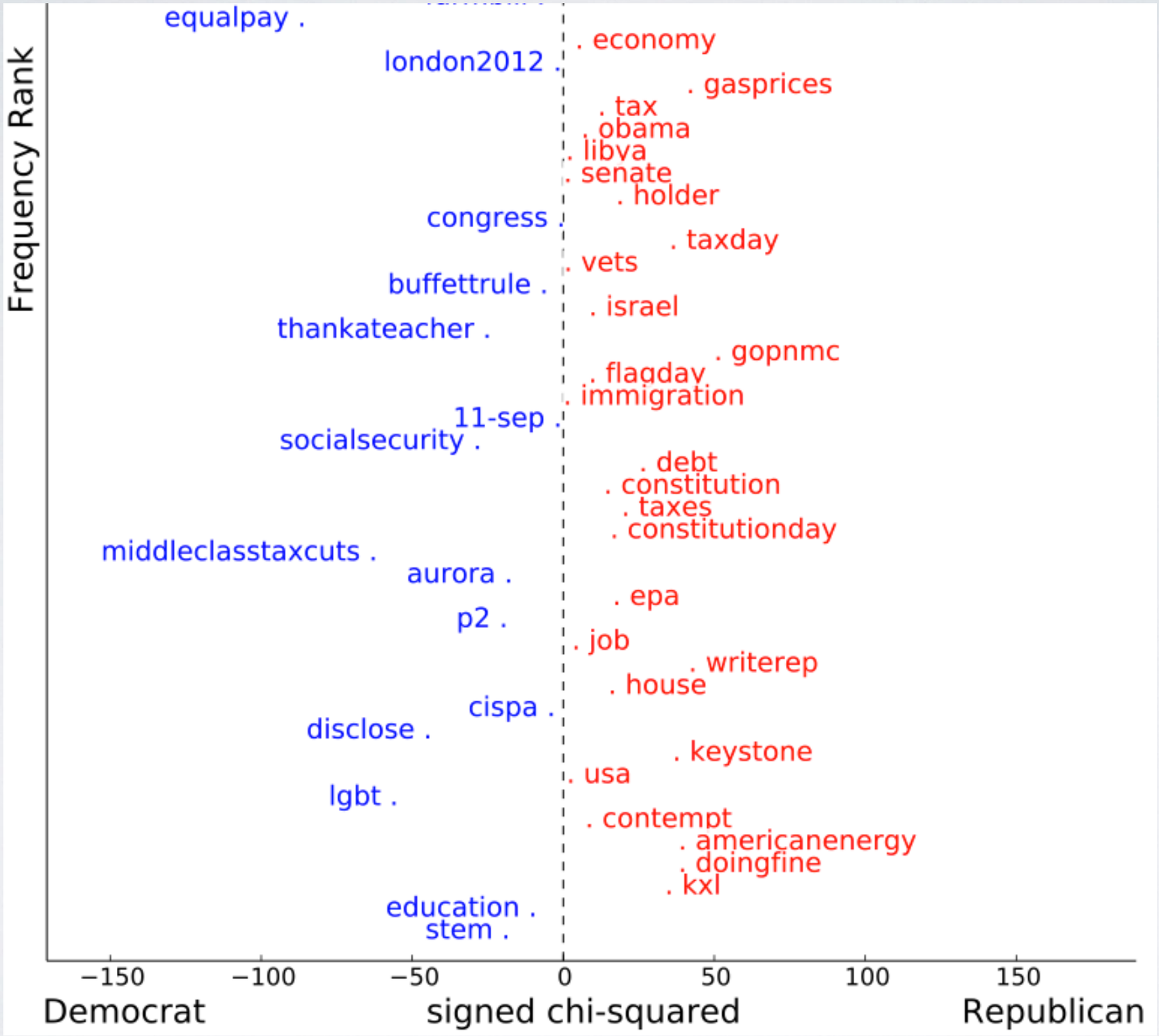
Other work

Eco-friendliness: truth vs. advertising vs. perception



with **Jennifer Cutler** (Stuart School of Business)

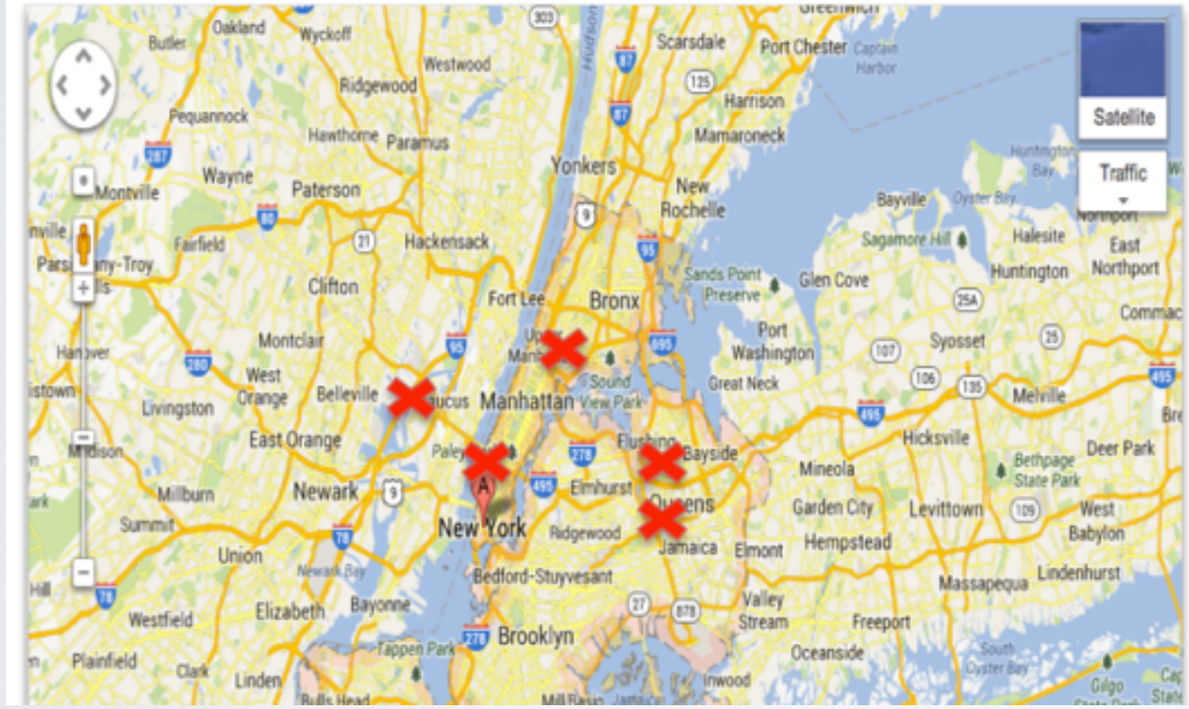
How partisan is a hashtag?



with **Libby Hemphill** & **Matthew Heston** (Comm. & Info Studies)

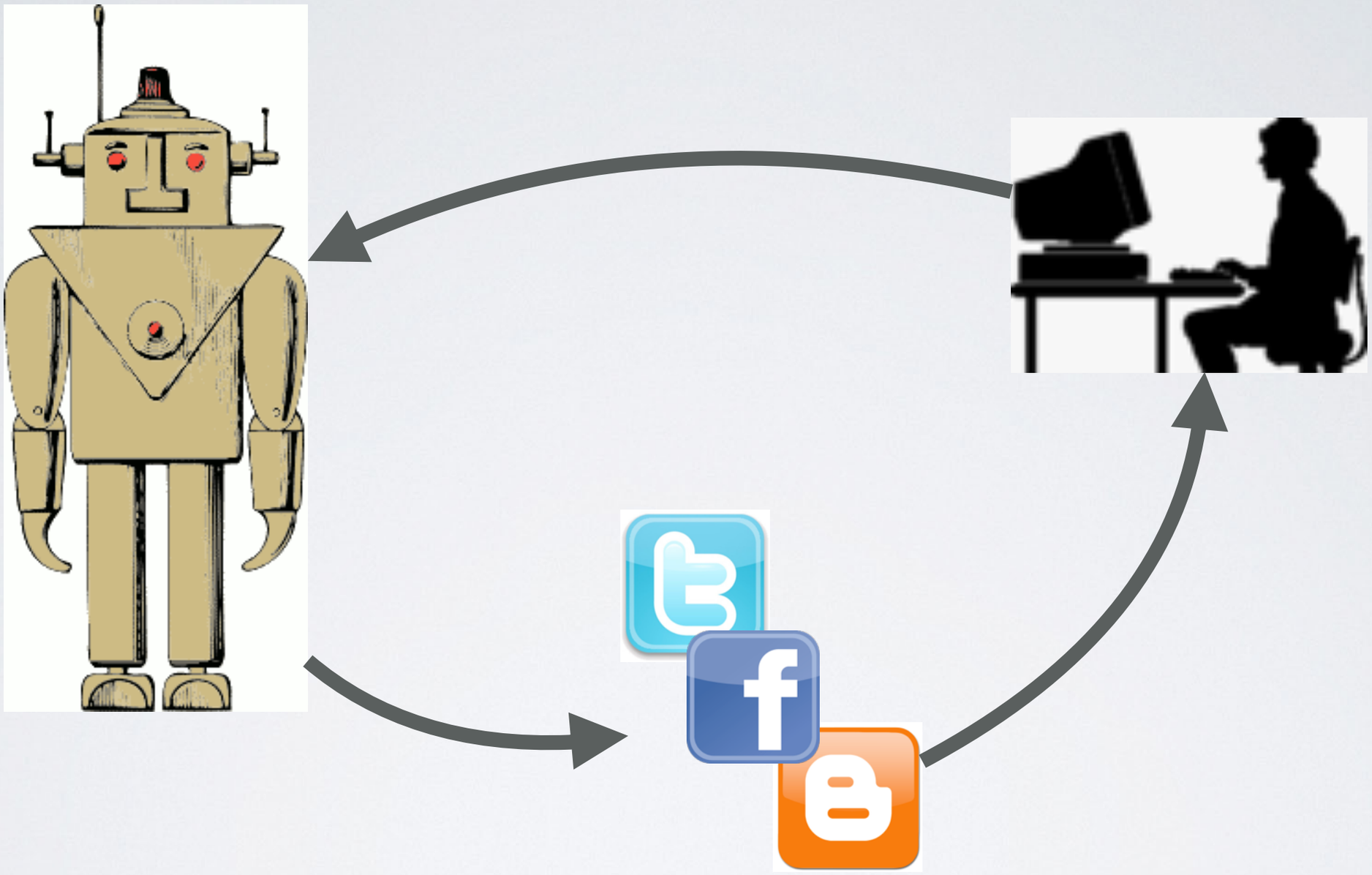
Disaster Informatics

Severity (1-5)	Entity Effected	Name of Entity	Type of Damage	Geo-location	Twitter Source Ids	Date and Time Reported	Credibility of Source(s) (1-10)
5	Bridge	Williamsburg Bridge	Fire/Burning	-79.8047, 22.5126	234, 12, 900234, 12, 900	6/19/13 12:30 am GMT	9
2	Intersection	45th & 52nd Street, New York	Flooding	163.5645, -48.6910	1093, 2768, 9330	6/18/13 11:45 pm GMT	8
3	Building	Long Beach Memorial Medical Center	No Electricity	-98.2095, 34.8838	8974, 7649,	6/15/13 03:30 am GMT	7
5	Neighborhood	Dumbo, Brooklyn	Flooding	-103.0025, 33.6158	2045, 13342, 9103, 2855, 934, 102, 945, 1332, 9054	6/10/13 04:50 am GMT	9



with **Zahra Ashktorab** (UMD), **Christopher Brown** (UT-Austin),
Jit Nandi (CMU)

Active Learning



with **Maria Ramirez-loaiza** & **Mustafa Bilgic** (IIT-CS)

Questions?

culotta@cs.iit.edu
http://cs.iit.edu/~culotta
http://tapilab.github.io

Hurricane Anxiety

[Mandel, Culotta, Boulahanis et al 2012]

Does concern toward impending hurricane vary...

by location?

by gender?

Hurricane Irene Experiment

- Train "concerned" classifier (~86% accuracy)

praying tht this hurricane goes back out to sea

Im actually scared for this hurricane

This hurricane is freaking me out

Literally don't care about the hurricane

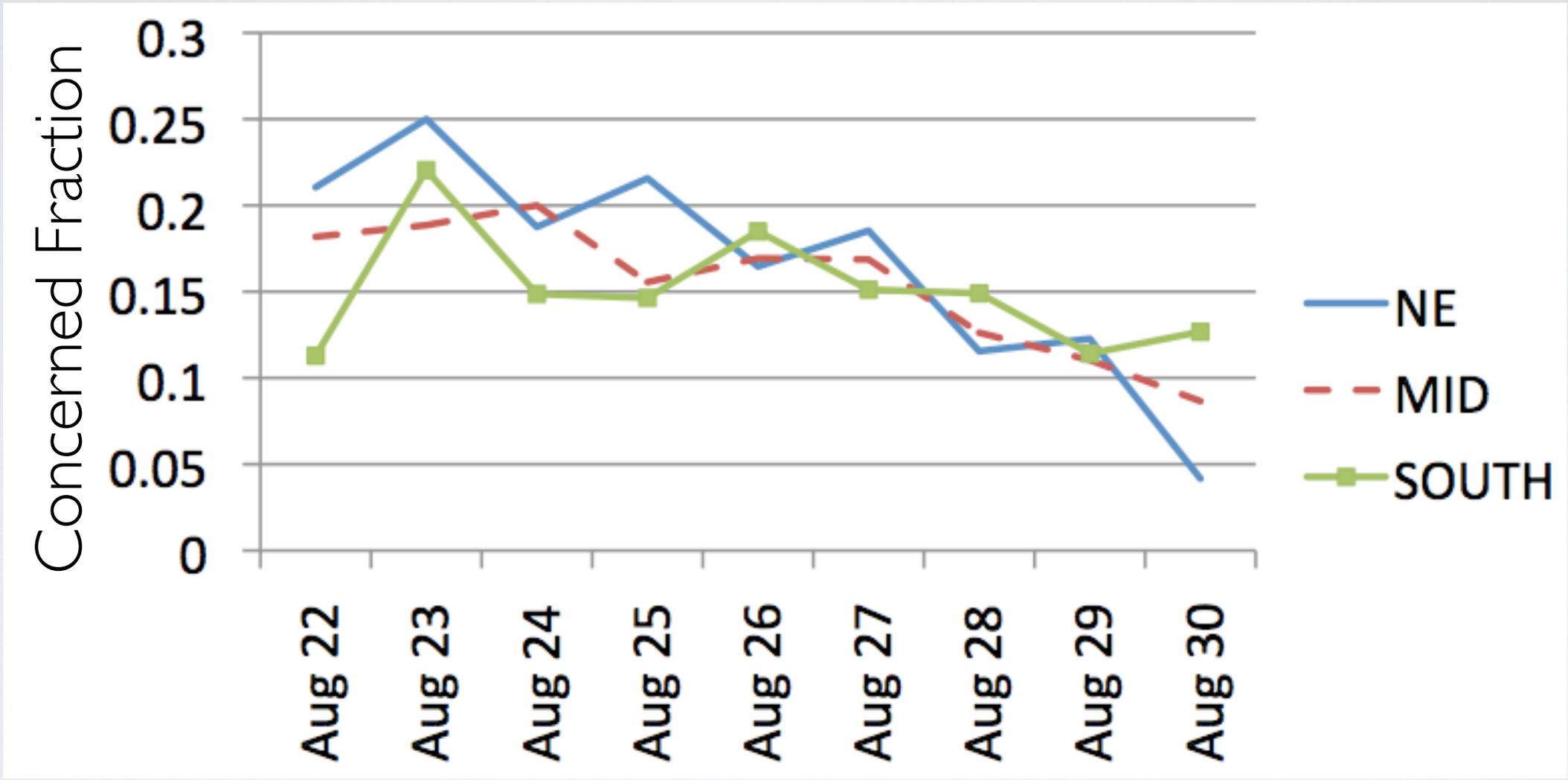
I honesty don't think the hurricane is going to be a big deal

for the very latest on hurricane irene like our fb page

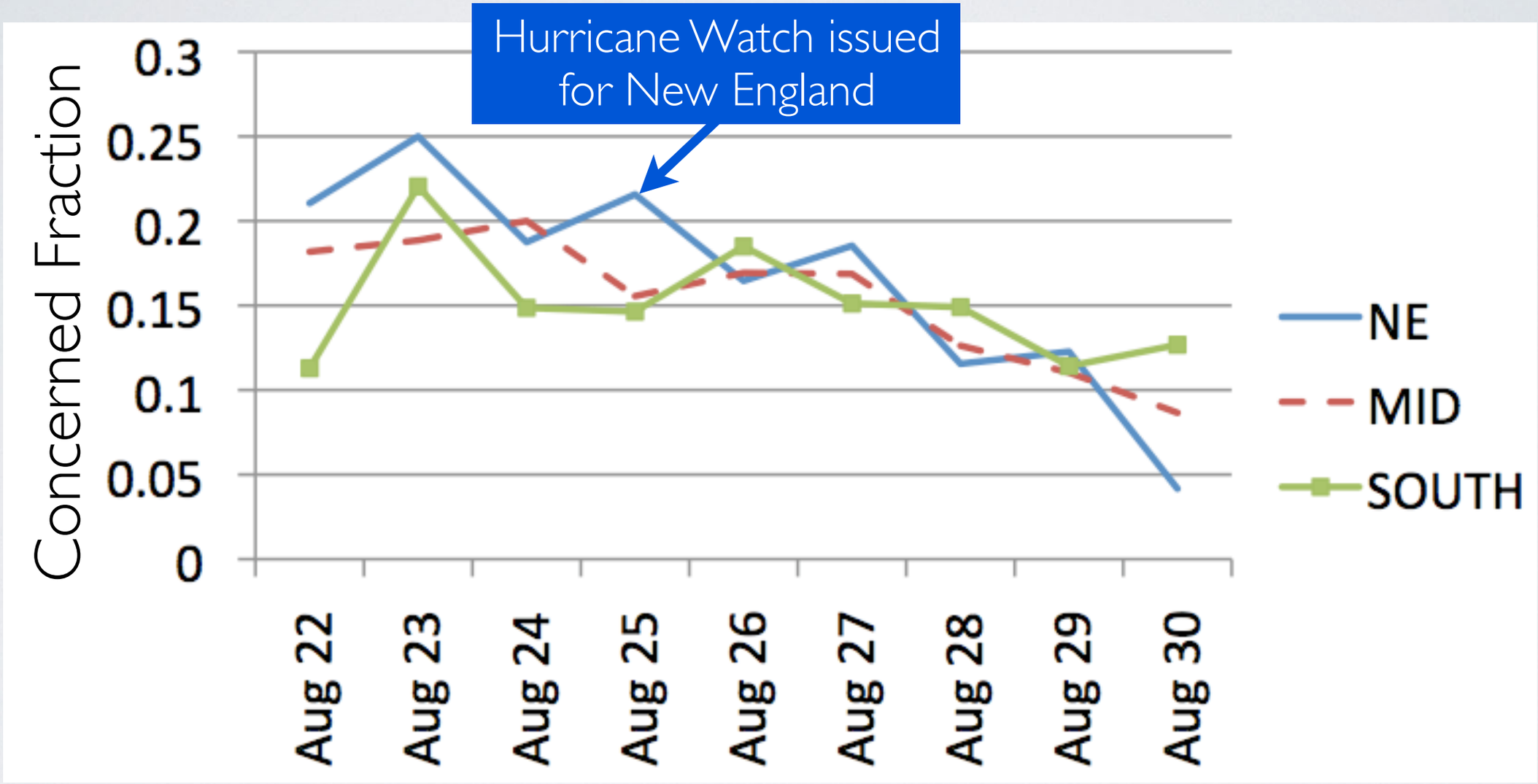
Hurricane Irene Experiment

- Demographic inference
  - Gender: lookup first names in census list of male/female names
    - ~46% of users assigned a gender
- Location: String match location field in profile
    - ~25% of tweets assigned a location

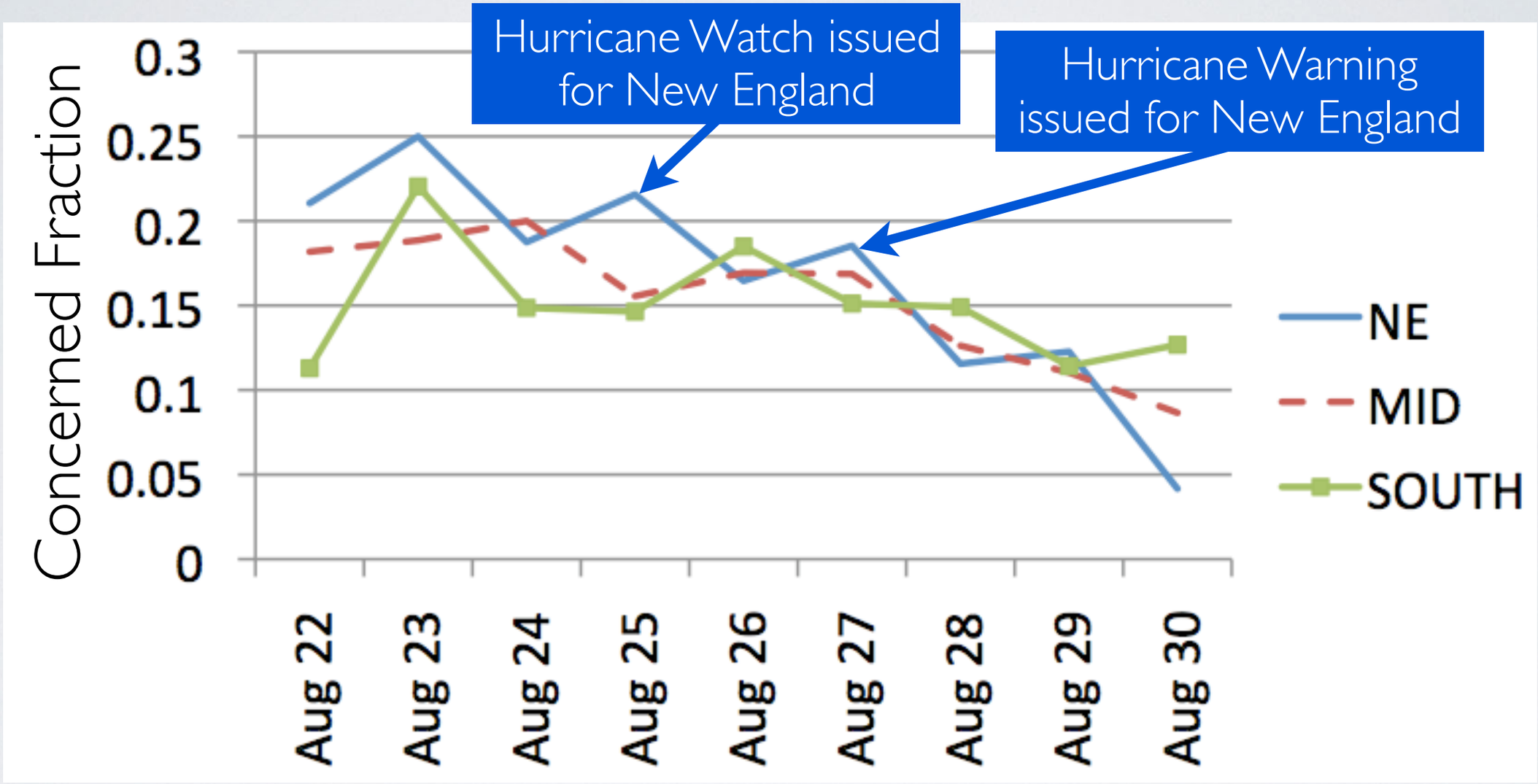
Concerned Fraction By Region



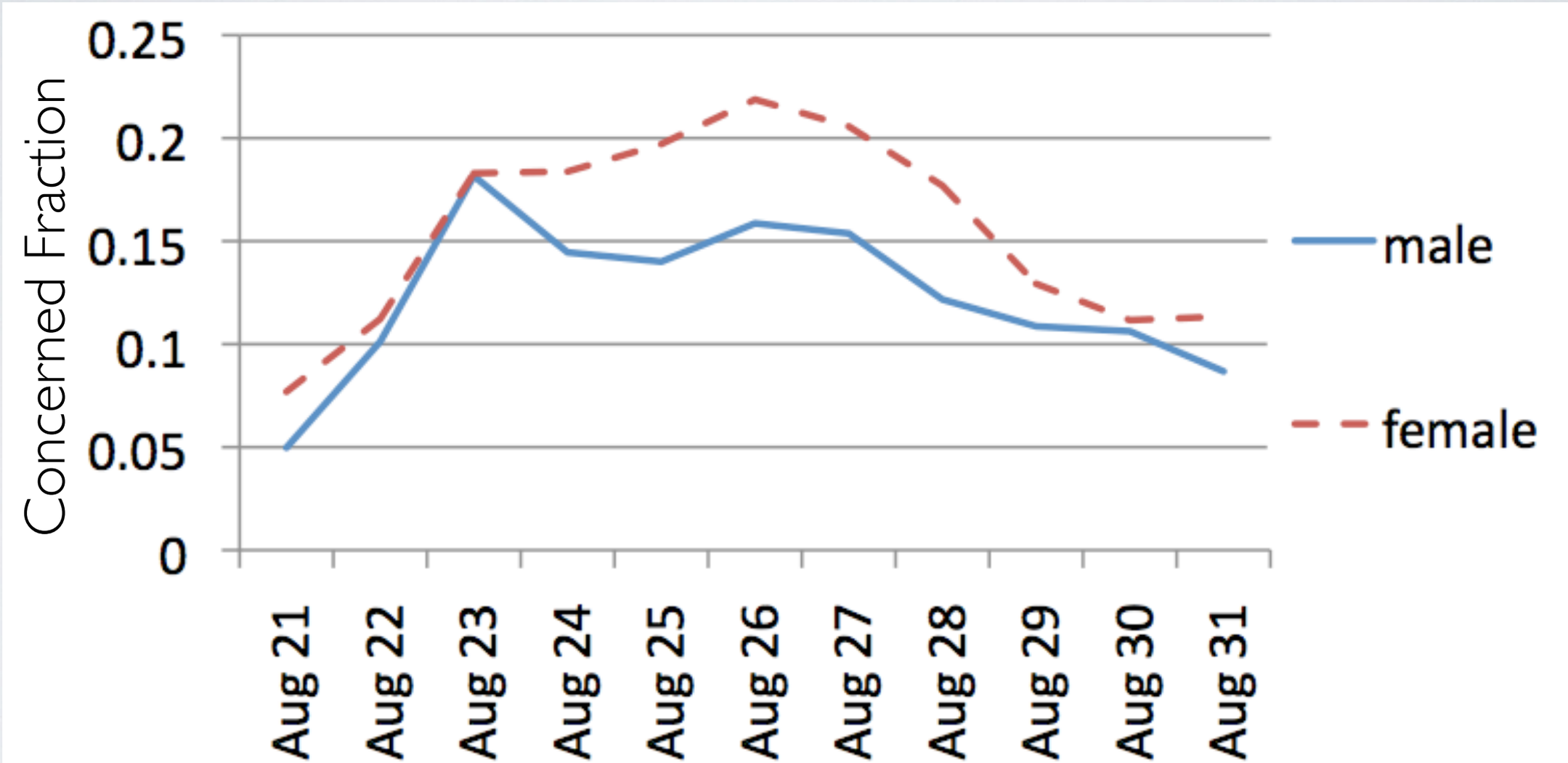
Concerned Fraction By Region



Concerned Fraction By Region



Concerned Fraction By Gender



Words by Gender

Information Gain Measure

Female	Male
I	http …
my	hurac …
safe	obama …
praying	blames …
this	(climate) change …
everyone	becomes …
died	dolphin

Hurricane Irene Experiment

- Demographic inference
  - Gender: lookup first names in census list of male/female names
    - ~46% users assigned a gender
- Location: String match location field in profile
    - ~25% tweets assigned a location

Hurricane Irene Experiment

- Demographic inference
  - Gender: lookup first names in census list of male/female names
    - ~46% users assigned a gender
  - Location: String match location field in profile
    - ~25% tweets assigned a location