Analyzing online social networks for interdisciplinary science

Aron Culotta
Associate Professor
Department of Computer Science

joint work with Libby Hemphill, Jennifer Cutler, Sherry Emery, Virgile Landeiro, Ehsan Ardehaly, Nirmal Ravi, Zhao Wang, Xintian Li, Karthik Shivaram, and many others

Text analysis, circa 2000

...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....

Text analysis, circa 2000

...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....

Text analysis, today

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

Text analysis, today

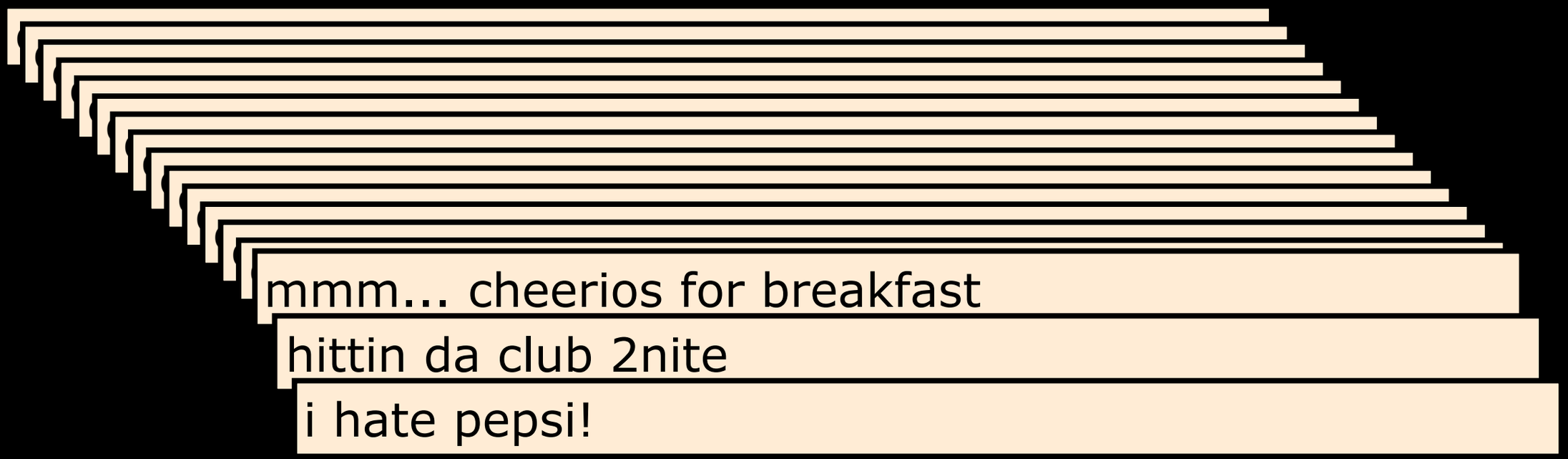
Central 2 DIS F4t wz PhRMA pres, CEO & el33t lobbyist Billy Tauzin, a Dem

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

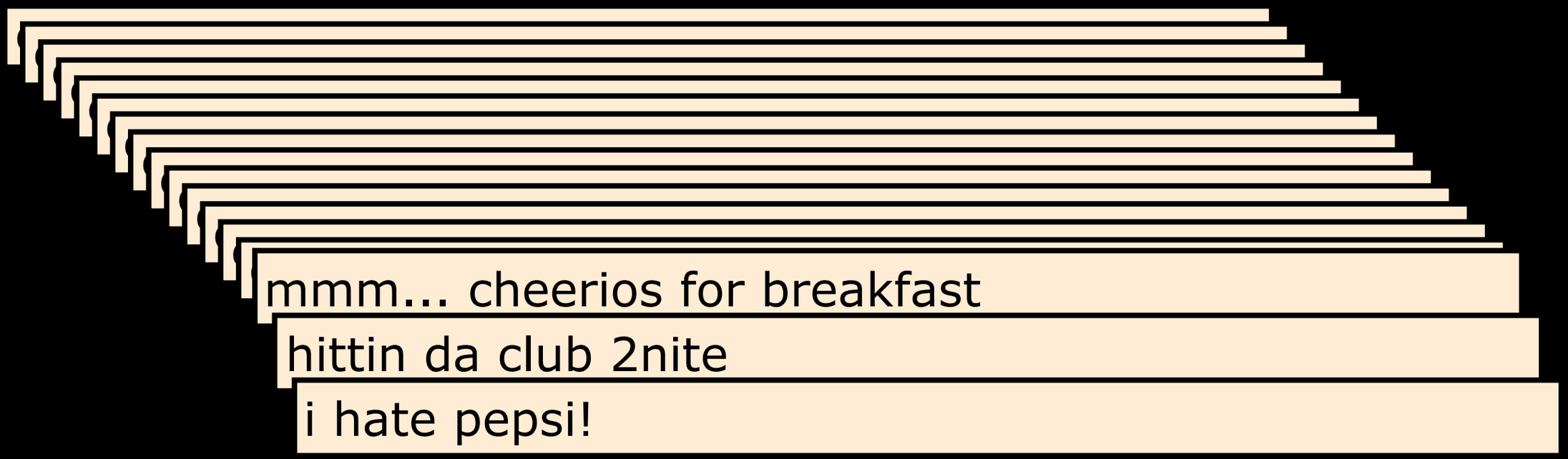
Text analysis, today

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

Text analysis, today



Text analysis, today

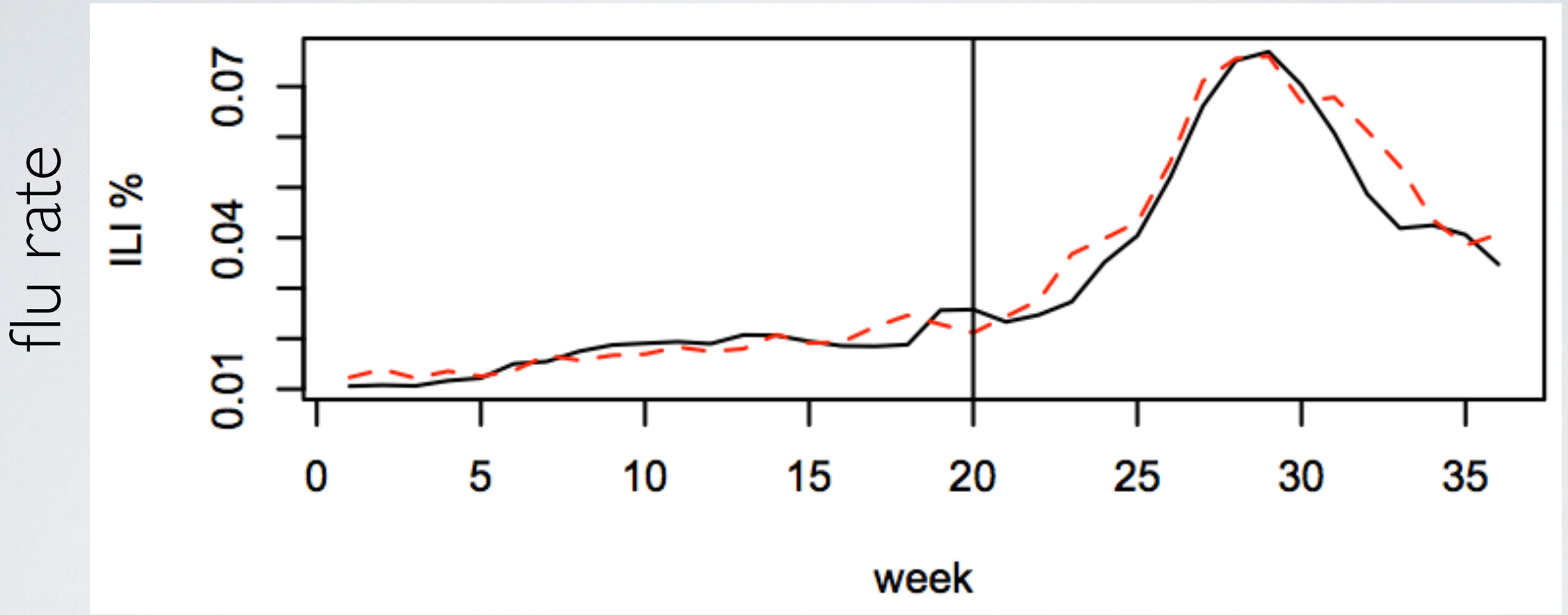


Pros: Big data, real-time, diverse topics

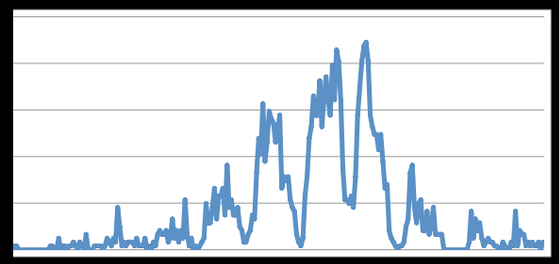
Cons: Noisy, unstructured, biased sample

Can we do anything good
with these data?

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



"Nowcasting" from Social Media



CDC historical statistics

Validation Time-Series

Regression

Document Sampling

Document Classification

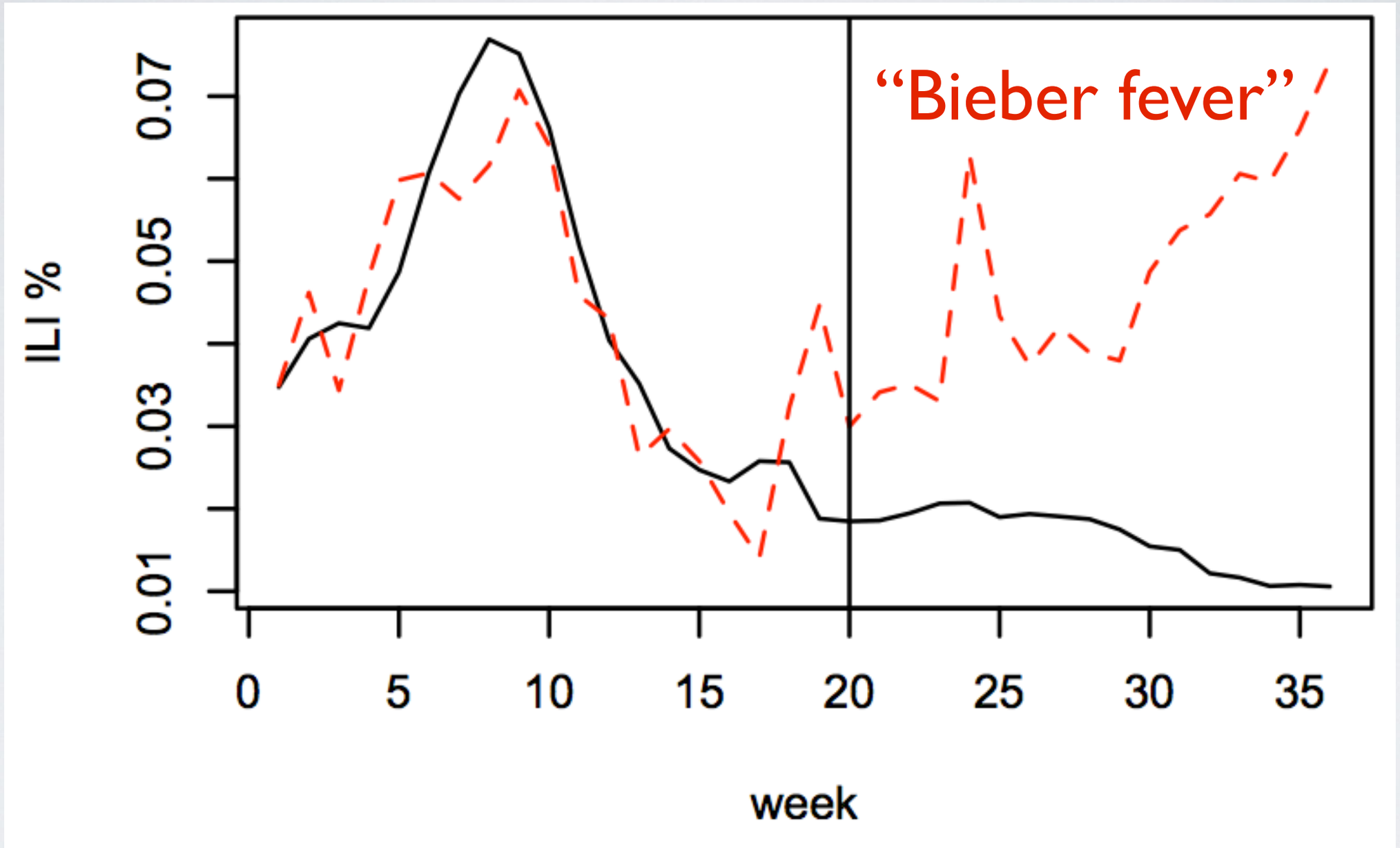
Aggregation

All tweets in week ending 3/7/15

flu related?

.05% positive

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



Nowcasting Rate of Influenza-like Illness (ILI)

Classification: simple keyword matching

logistic regression (~84% accuracy)

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

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

fraction of tweets matching ILI classifier

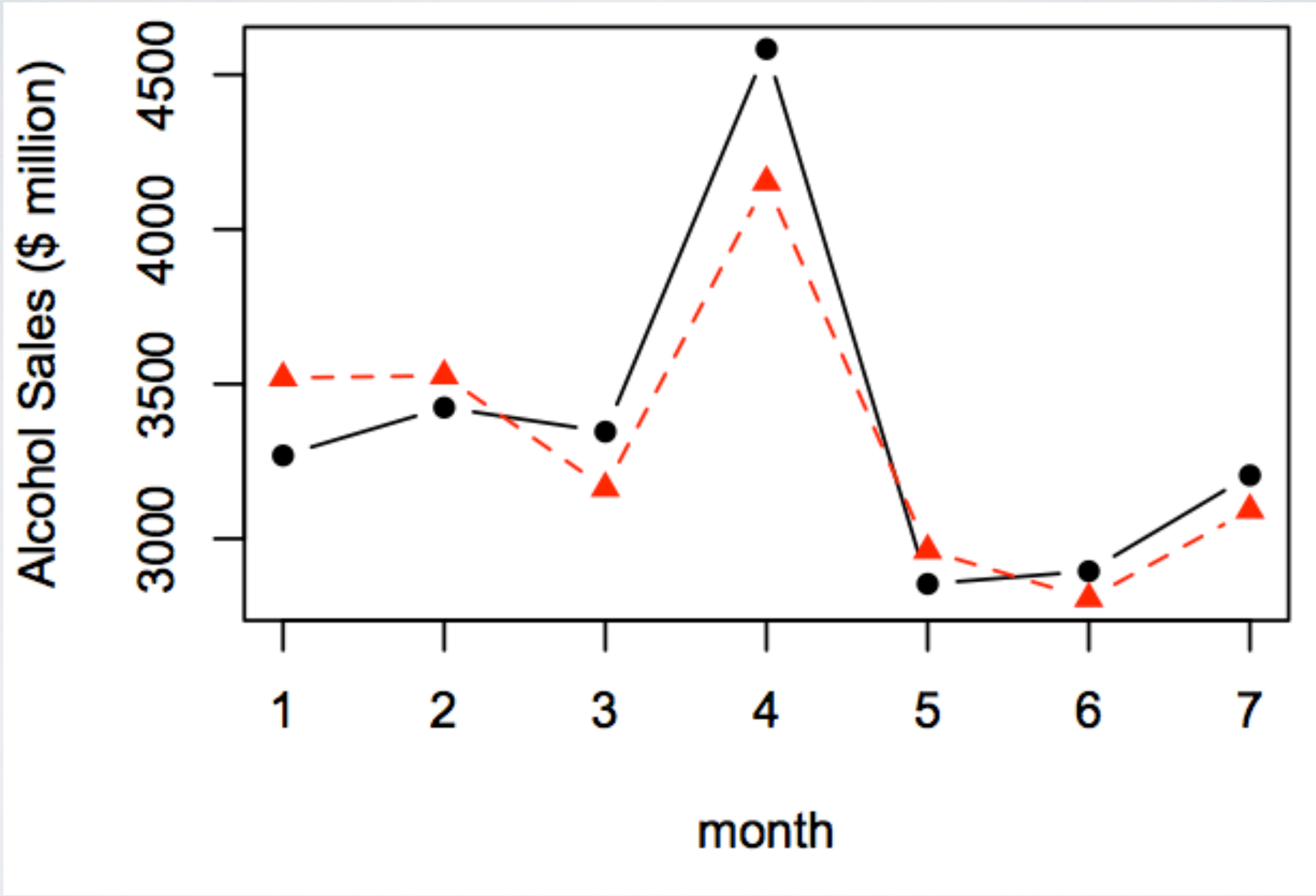
$$f_s(w, D) = \frac{\sum_{d_i} \Pr(y = 1|d_i) \mathbf{1}_w}{|D|}$$

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

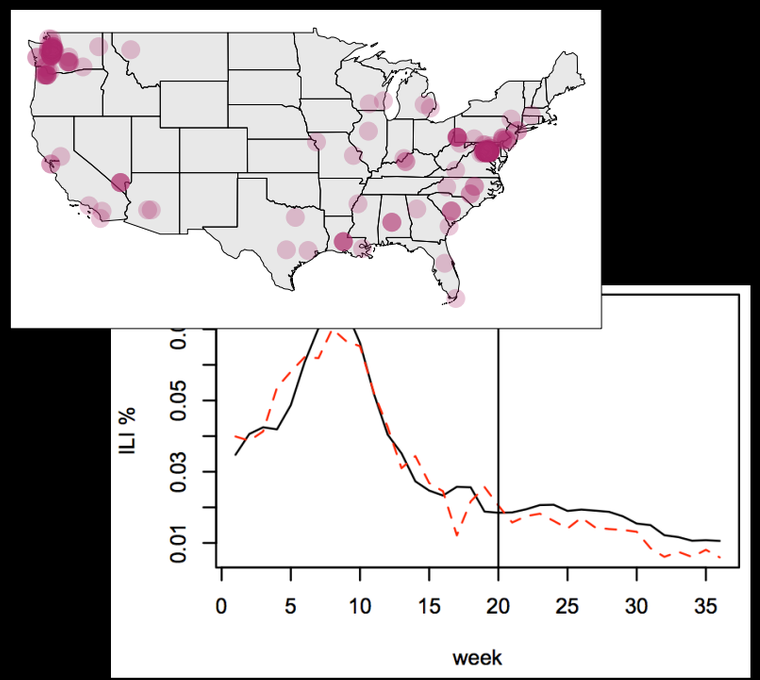
w : tweets matching word w

D : all tweets this week

r = 0.93



Generating knowledge from Social Media



descriptive knowledge

influenza	Lampos et al. ’10; **Culotta ’10**; Paul & Dredze ’11,…
food poisoning	Sadilek et al. ’13
chronic illness	Paul & Dredze ’11, Schwartz et al. ’13, **Culotta ’14**,…
drug use	Hanson et al. ’13, **Culotta et al. ’13** …
insomnia	Heaivilin et al. ’11
depression	De Choudhury et al ’13, ..
roadkill	Xu et al. ‘13

Generating knowledge from Social Media

causal knowledge?
risk factors?

exercise <—> mood
anxiety <—> physical health
drinking <—> hostility

stop smoking ads —> increase in cessation attempts

Outline

- Applications
    - Public Health
    - Marketing
    - Crisis Informatics
- Machine learning methods
    - Learning from label proportions
    - Deconfounded classification
    - Domain adaptation

Does population health vary with language?

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

27 health-related statistics

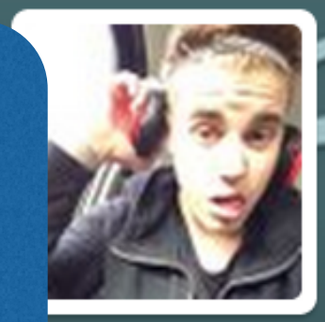
Outcomes
Poor Health
Unhealthy Days
Mentally Health
Low Birthweight
Diabetes
Obesity

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

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

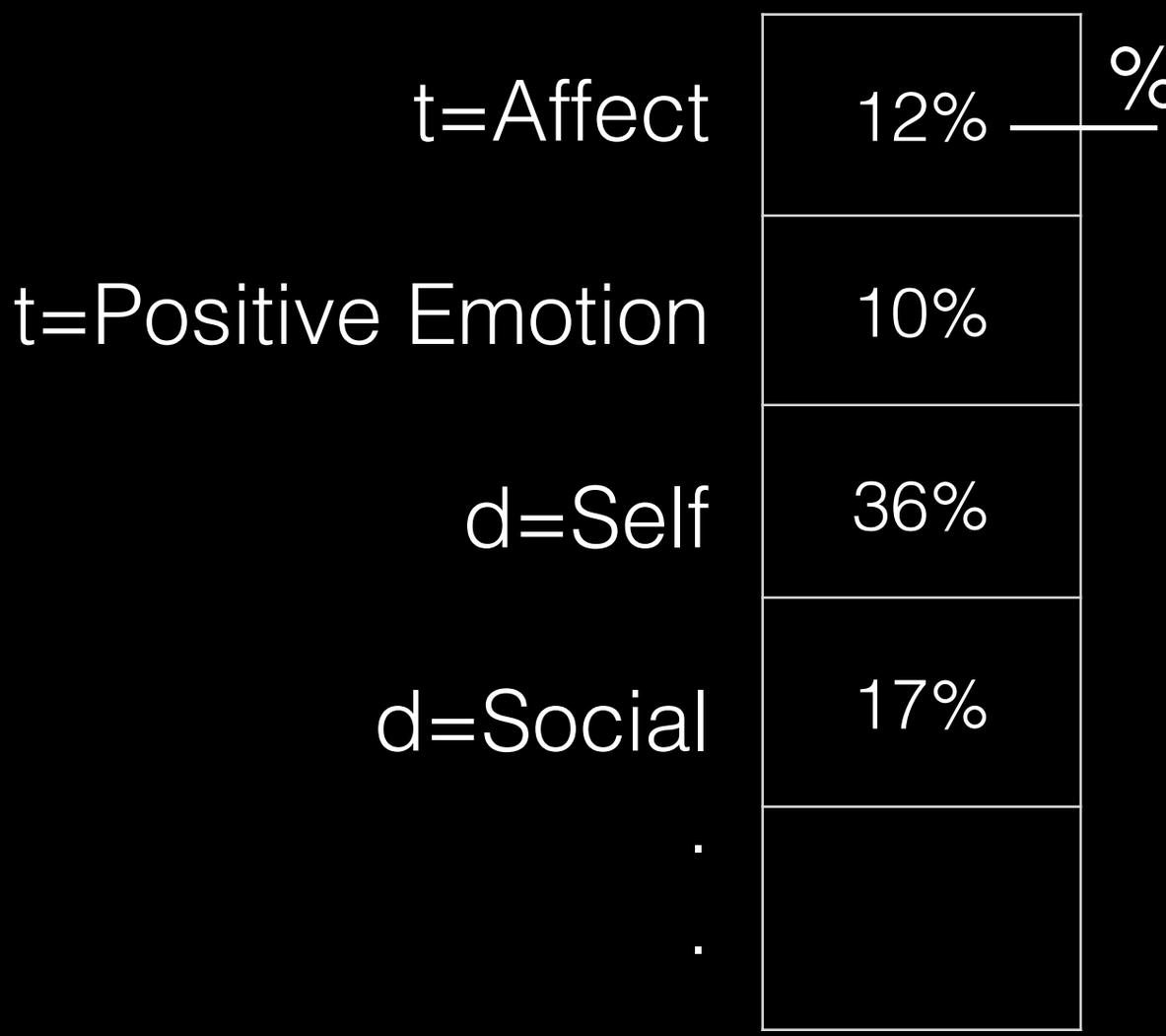
feeling happy and creative. alot of great things coming.

11:18 AM - 28 Apr 2014

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

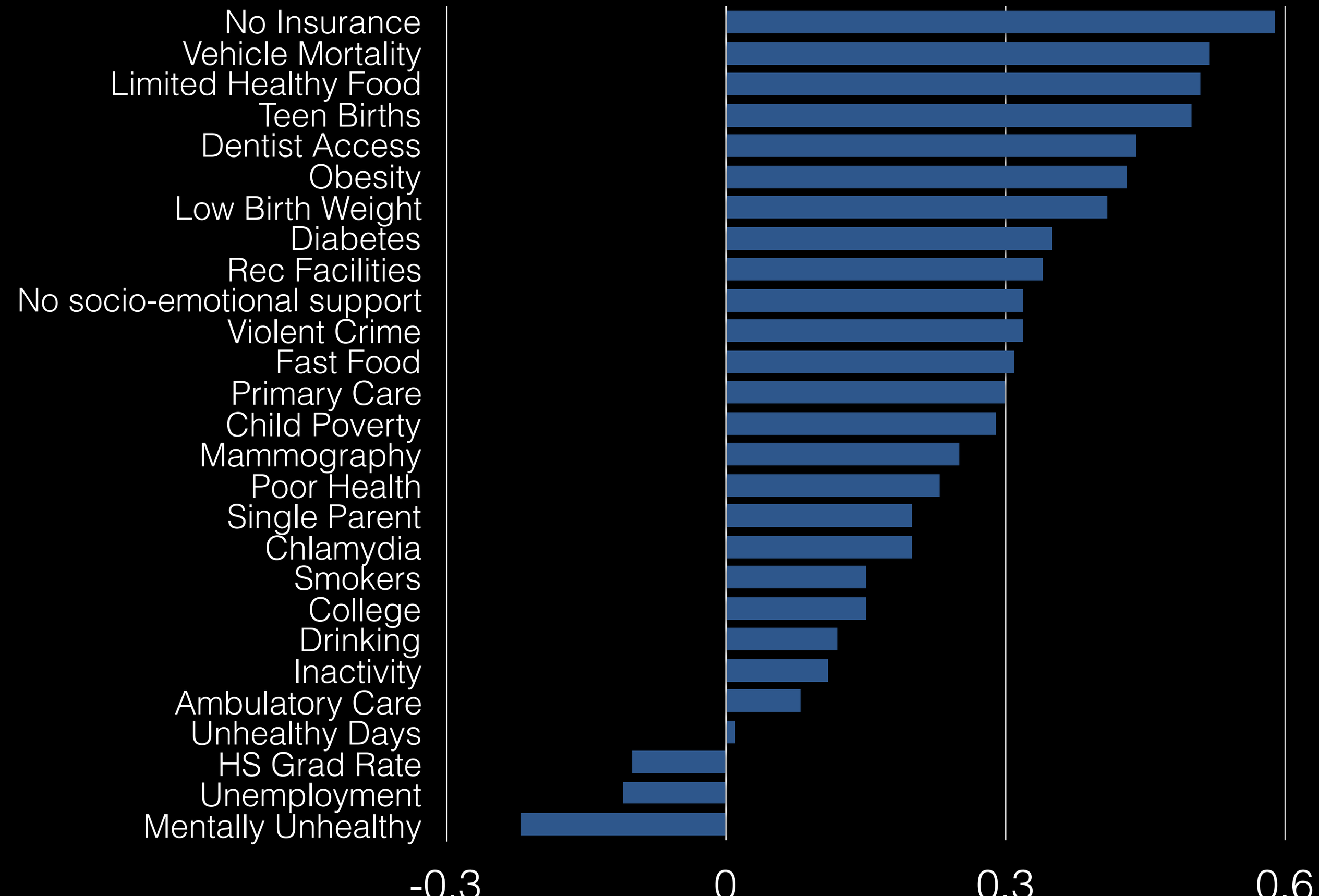
Linguistic Profile of each County

L.A. County . . . [100]



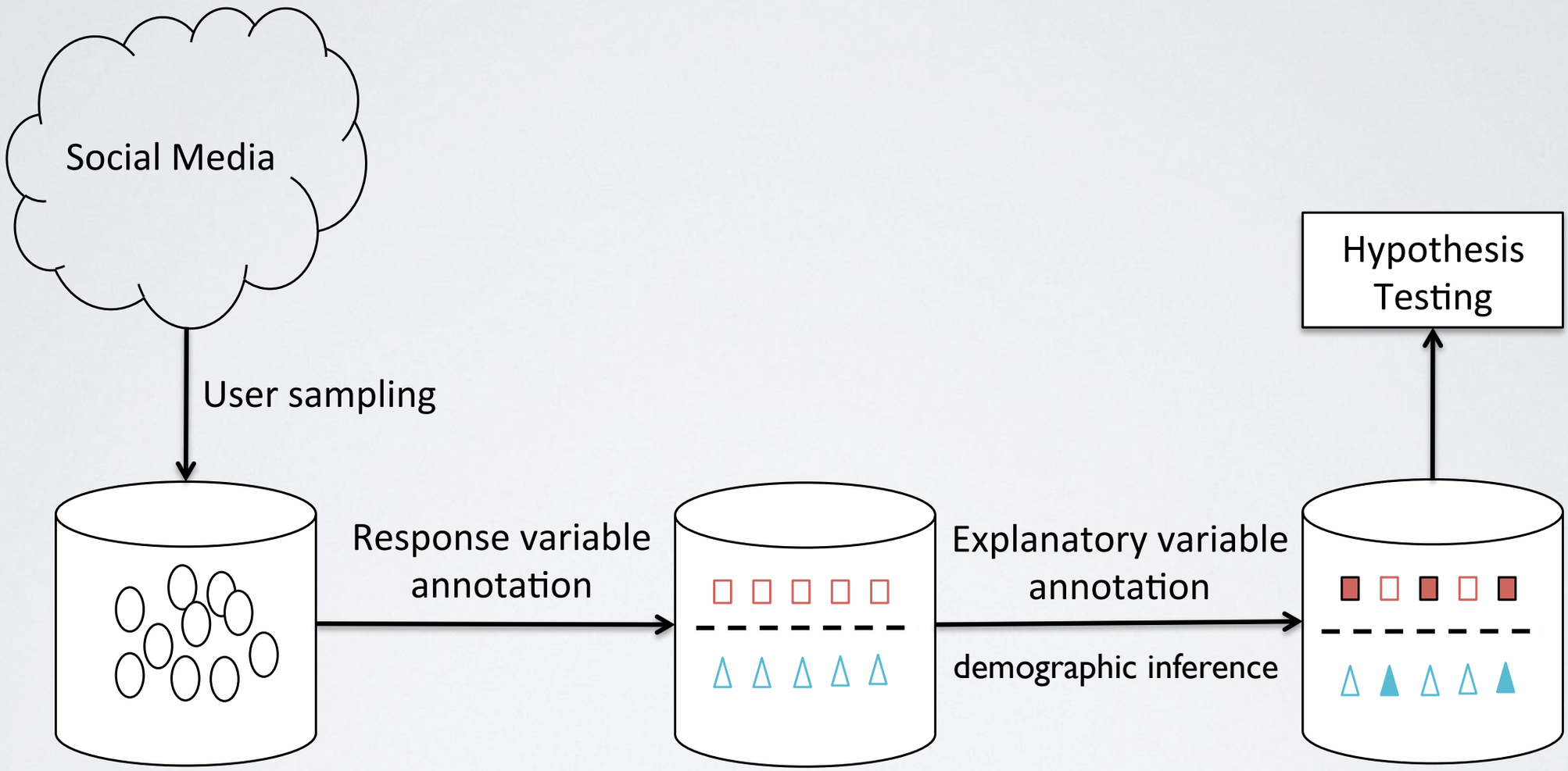
% of users who mention a word from this category

Held-out Correlation



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

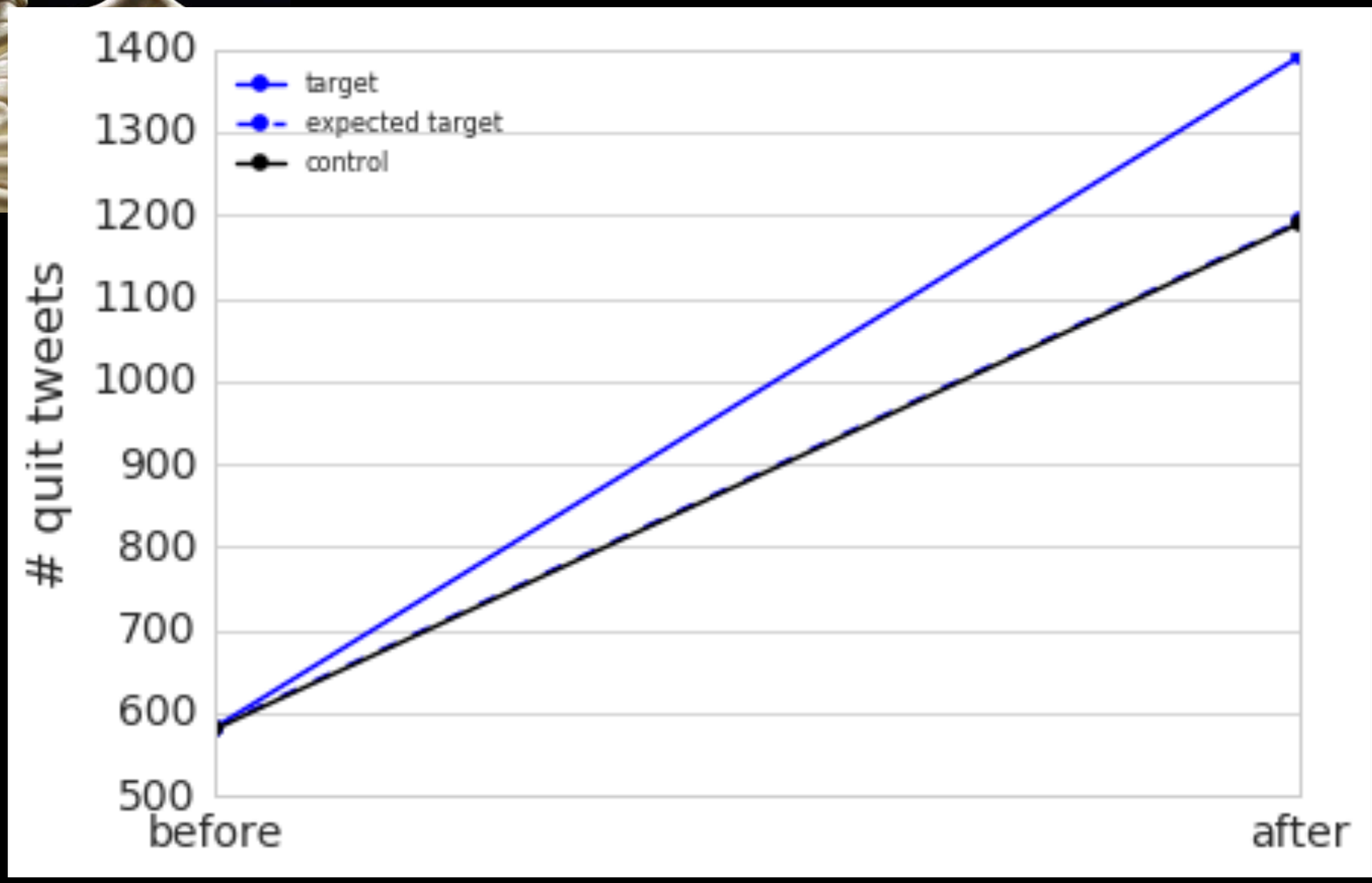
Web-scale Observational Studies



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

Do smoking cessation campaigns work?

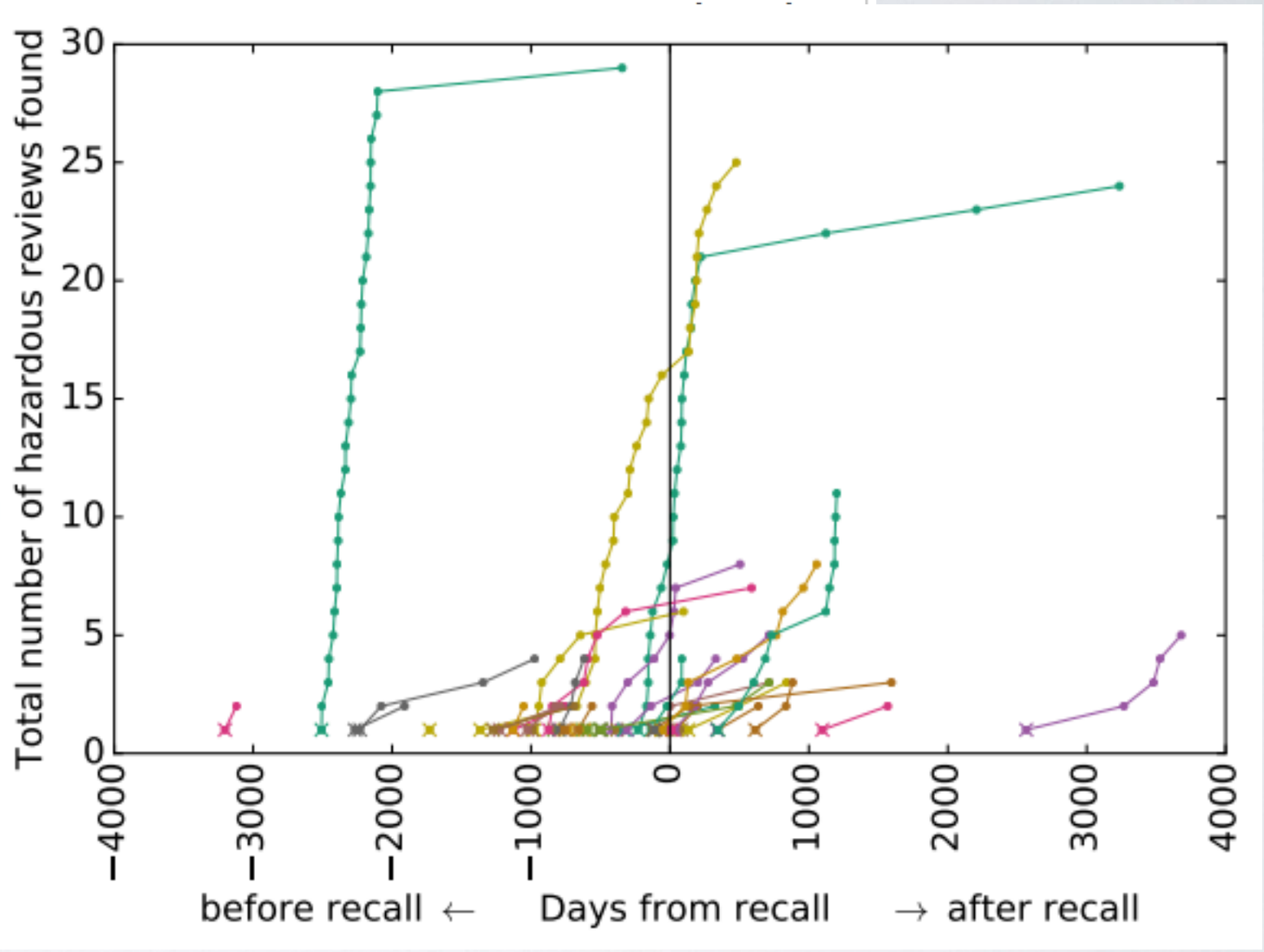
with Sherry Emery (NORC)



Forecasting Product Recalls

Search for products.. or Search for keywords within reviews..

ID	Product Name	Review Title	Product link	Review link	Probability of
1	Fisher-Price Starlight Cradle 'n Swing	FALL ALERT - My 6 month c though fastened by seat be			
2	Fisher-Price My Little Lamb Cradle 'n Swing	This Product Destroyed My NOT BUY THIS PRODUCT**			
3	Topeak BabySeat Child Bike Seat with Aluminum Rack (Disc Mount Version)	Todson Recalls Bicycle Child Laceration and Fingertip An			
4	Peg Perego Tatamia Stripes, Grey	Plastic, plastic, and more pl			
5	Fisher-Price Newborn Rock 'n Play Sleeper, Yellow (Discontinued	MOLD ISSUE?			



CONSUMER PRODUCT SAFETY COMMISSION
UNITED STATES

with Shreesh Kumara Bhat

Forecasting Cyberbullying

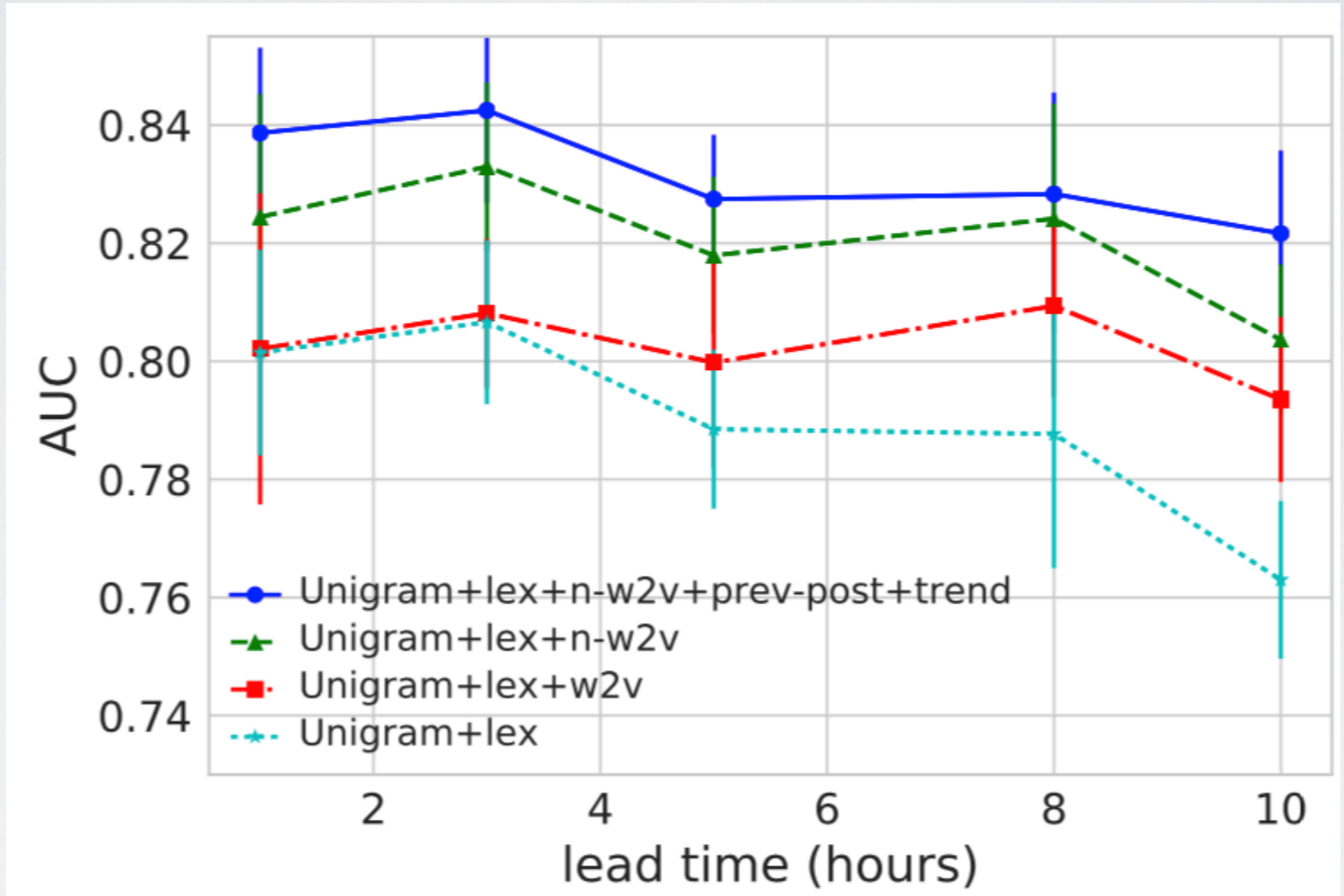
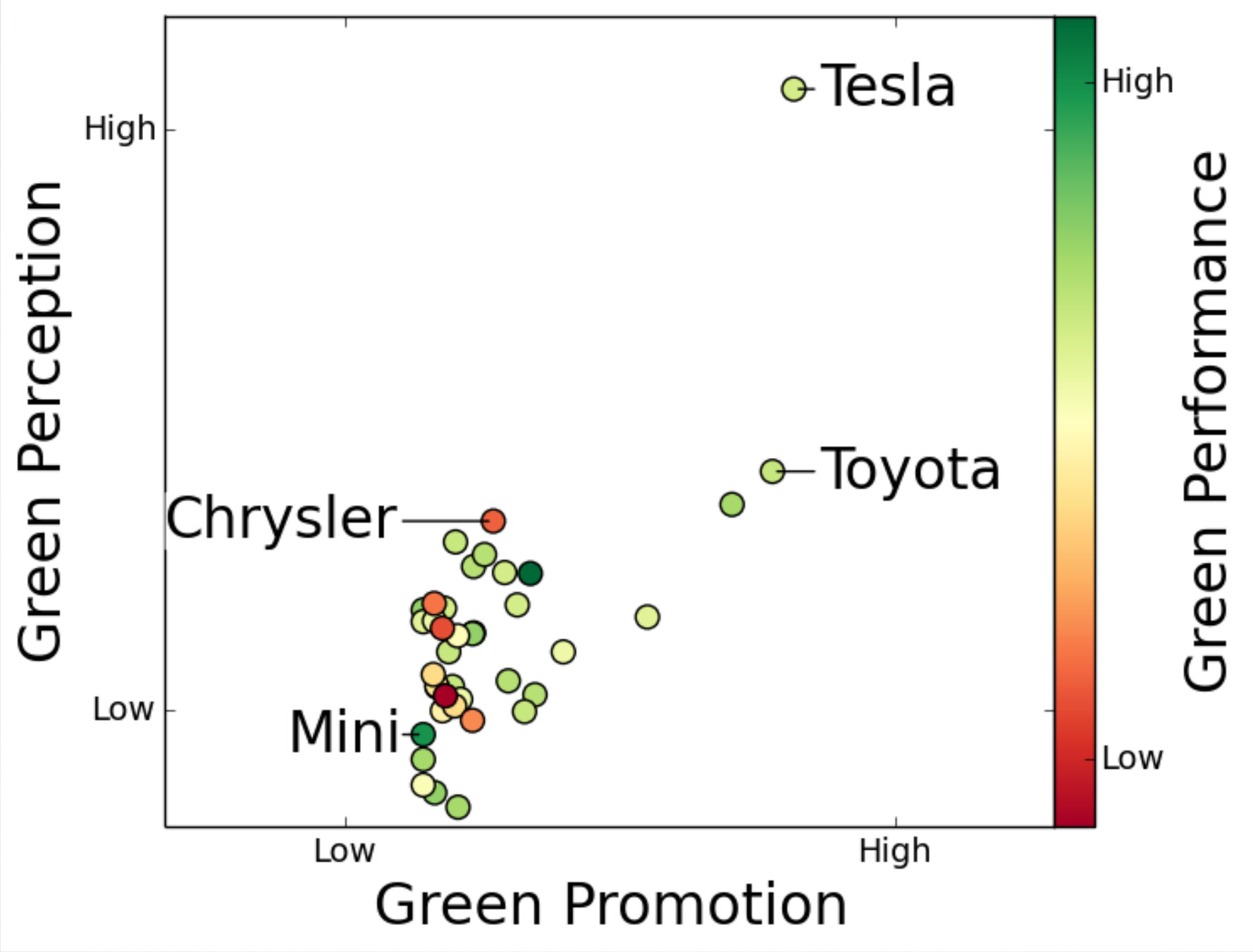


Figure 3: Hostility presence forecasting accuracy as lead time increases.

with Libby Hemphill (Michigan) and Ping Liu

NAYAR PRIZE
ILLINOIS INSTITUTE OF TECHNOLOGY

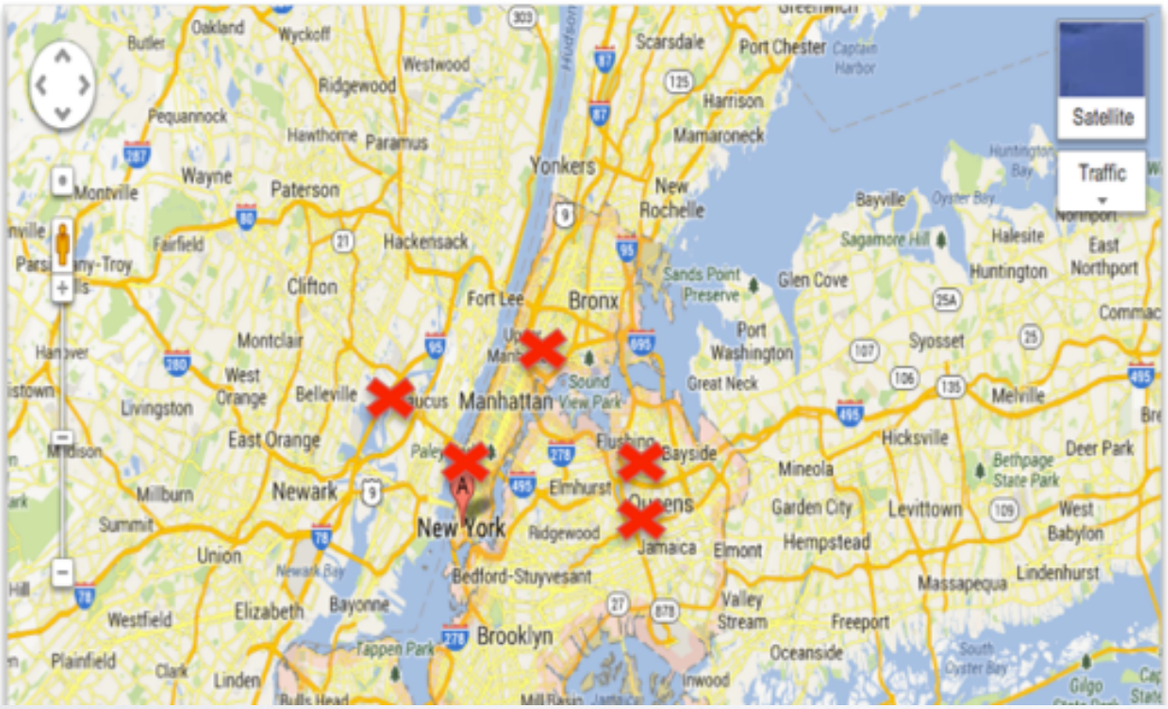
Eco-friendliness: truth vs. advertising vs. perception



with Jennifer Cutler (Northwestern) and Zhao Wang

Crisis 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)

Outline

- Applications
  - Public Health
  - Marketing
  - Crisis Informatics
- Machine learning methods
  - Learning from label proportions
  - Deconfounded classification
  - Domain adaptation

Learning from Label Proportions

- Supervised classification assumes training data like:

$$(\vec{x}, y) \quad y \in \{0, 1\}$$

But, this can be hard to get.
Instead, LLP assumes training data like:

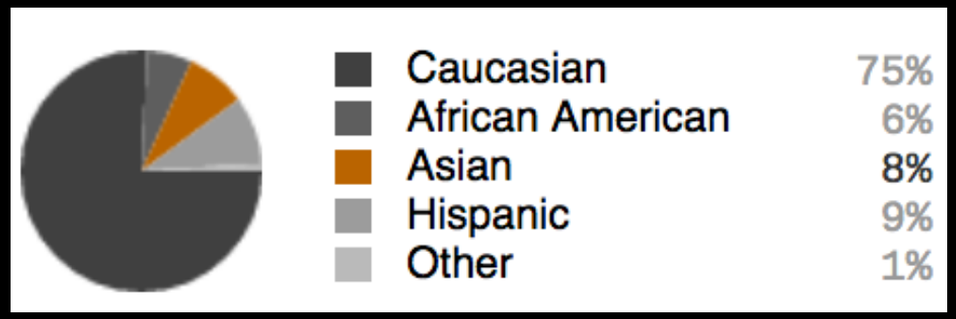
$$(\{\vec{x_j}\}_{j=1}^{n}, y) \;\; y \in [0, 1]$$

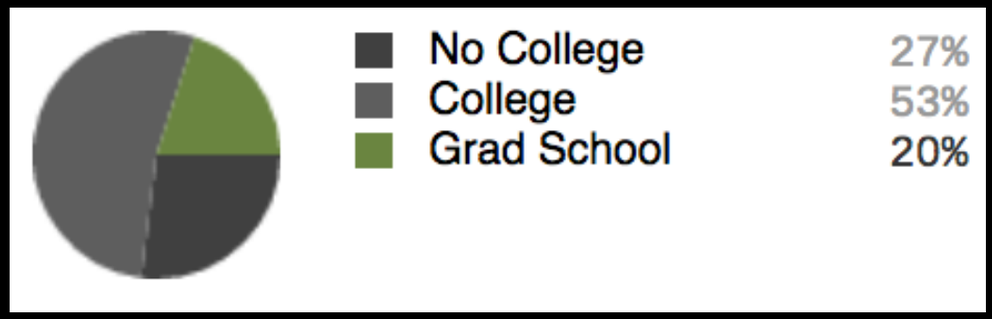
quαntcast

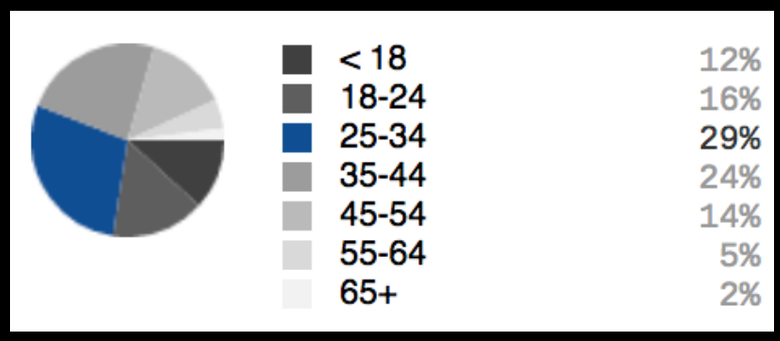
lifehacker.com

↳ 7 Subdomains

14.8M US 26.8M Global







Top-weighted Predictors

Text

Republican	**Democrat**
@foxnews	women
christmas	u
#tcot	ain't
football	equality
obama's	@nytimes

Top-weighted Predictors

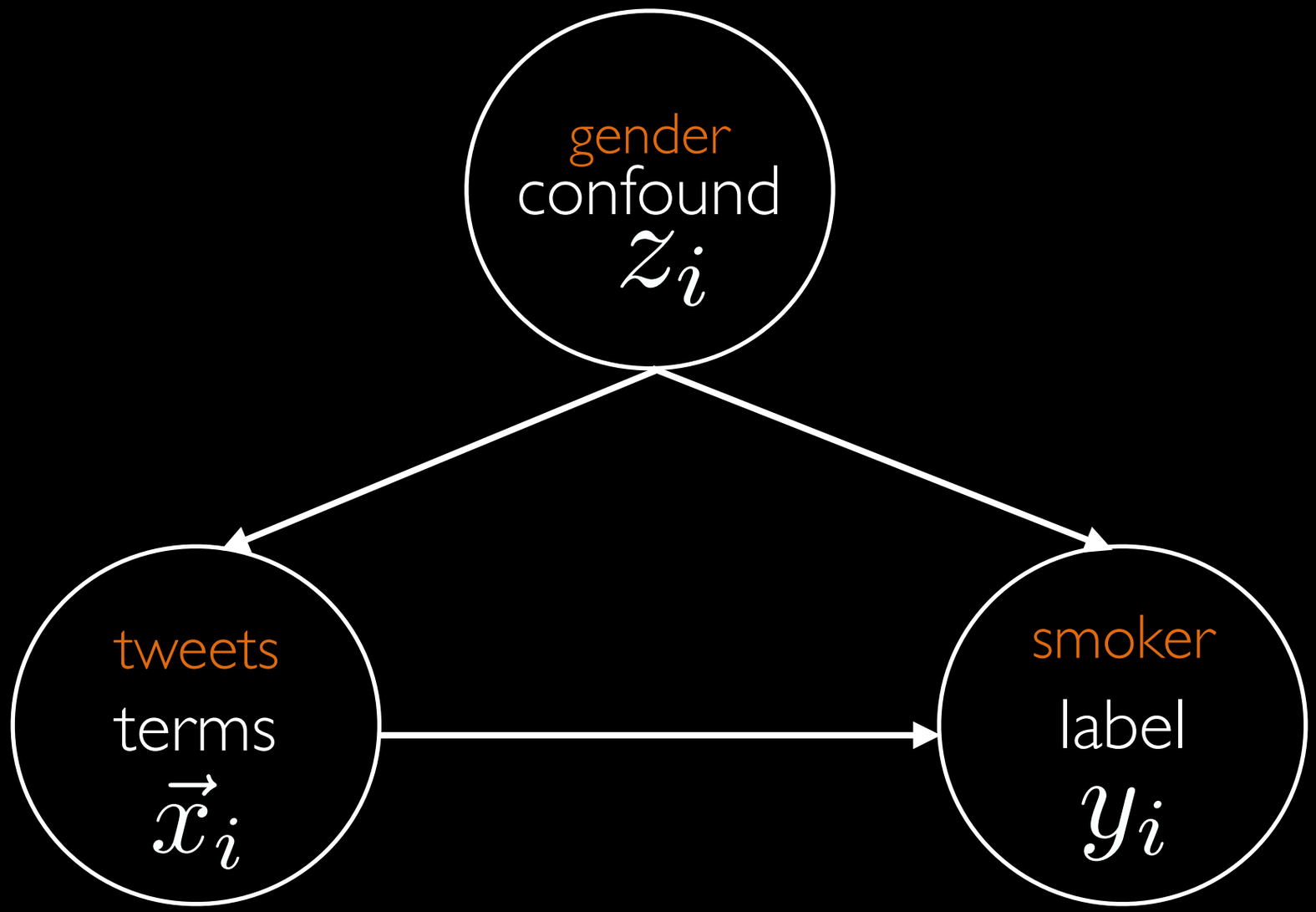
Friends

College	**Grad School**
ConanOBrien	NewYorker
LouisCK	nytimes
DanielTosh	TheEconomist
AzizAnsari	WSJ
Wired	WashingtonPost

Top-weighted Predictors
Text

18-24	25-34	35-44	45-54	55-64
haha	super	star	wow	vote
album	dc	fans	vote	**golf**
stream	**baby**	**kids**	american	country
wanna	**pregnancy**	tv	**boys**	**holiday**
im	**wedding**	**son**	**nice**	**smile**

Deconfounded classification



Build a classifier p(y|x) that is not influenced by z

Domain adaptation

- Machine learning assumes training and testing data come from same distribution.

- This is rarely true in practice.

$$(\vec{x}, y) \sim P_{\text{train}}(X, Y) \qquad (\vec{x}, y) \sim P_{\text{test}}(X, Y)$$

Domain adaptation uses unlabeled test data to improve accuracy.

Deep learning

We have developed deep learning algorithms for

- learning from label proportions
- deconfounded classification
- domain adaptation

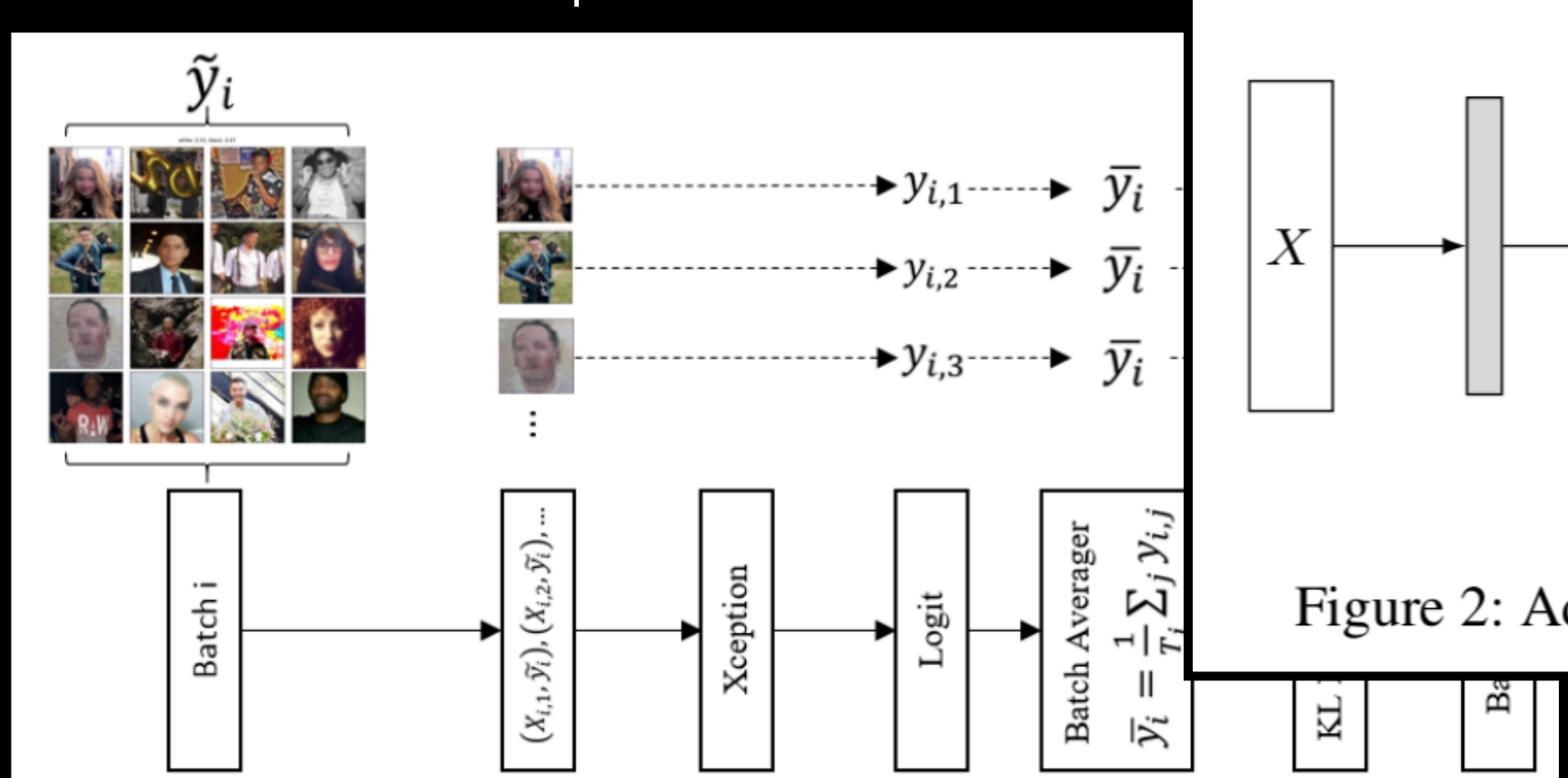


Fig. 1. The deep LLP framework applied to image classification.

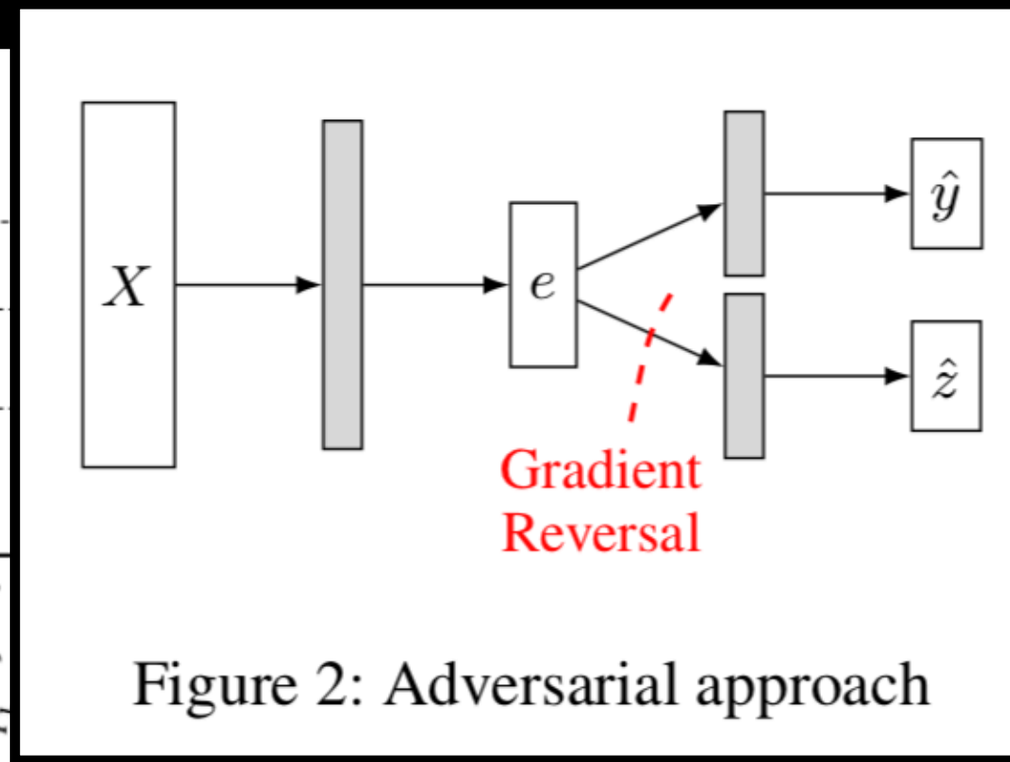


Figure 2: Adversarial approach

Ethical Issues in AI / Web Research

Privacy: What data can be collected? shared? with whom?

Consent: Scientists are bound by ethical guidelines when conducting research involving humans.

Fairness: When algorithms are used to make decisions about people, they must not discriminate based on protected classes of people (e.g., by gender, race/ethnicity, sexual orientation, etc.).

Politics

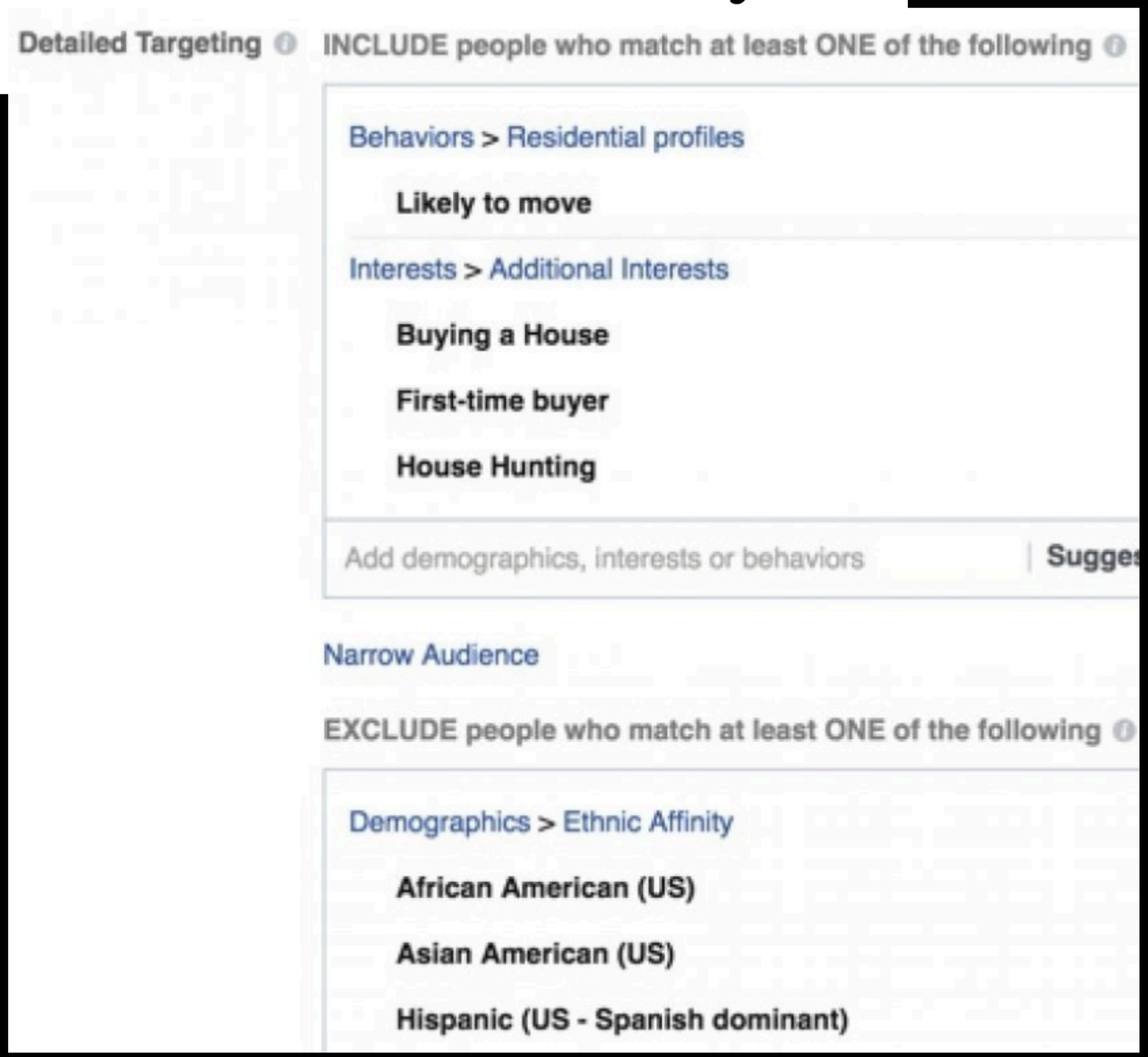
Bannon oversaw Cambridge Analytica's collection of Facebook data, according to former employee

July 2019 : The Federal Trade Commission has announced that Facebook will pay **$5 billion** to settle the charge that it broke a 2012 FTC order concerning the privacy of user data. And, as part of the settlement, Facebook has had to agree to a new management structure and new rules about how it manages user data.

PROPUBLICA

MACHINE BIAS

Facebook (Still) Letting Housing Advertisers Exclude Users by Race



Women Less Likely To Be Shown Ads for High-Paying Jobs

By Prachi Patel

Posted 8 Jul 2015 | 18:02 GMT

Experiment: create fake Google profiles, then search for jobs

- - The only difference between profiles was gender

The male profiles were much more likely to be shown ads for a career coaching service for executive positions paying over $200,000. The Google ad network showed this ad to the male users more than 1800 times, but only about 300 times to women.

AOL Proudly Releases Massive Amounts of Private Data

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

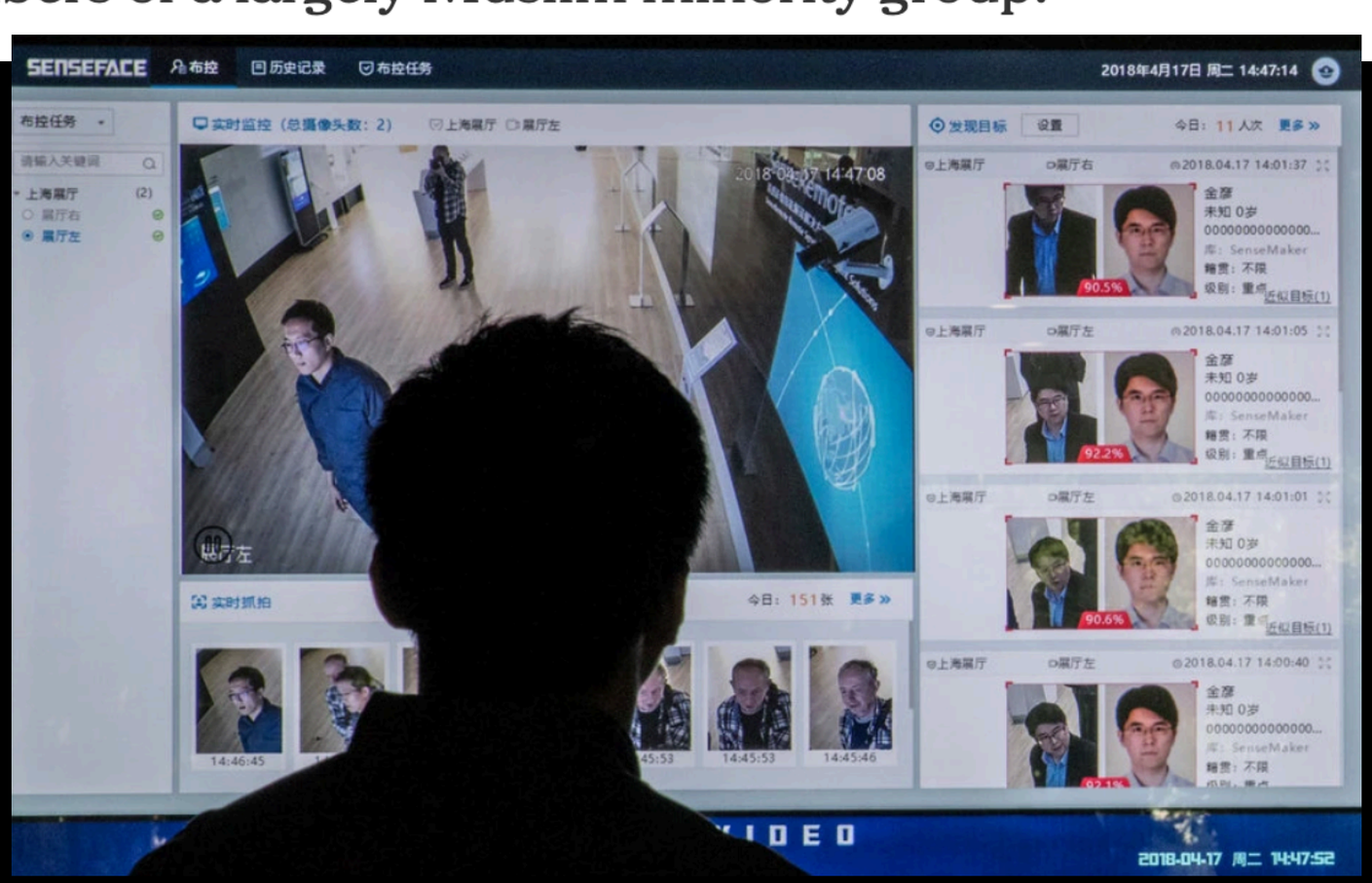
Next Story

AOL has released very private data about its users without their permission. While the AOL username has been changed to a random ID number, the abilitiy to analyze all searches by a single user will often lead people to easily determine who the user is, and what they are up to. The data includes personal names, addresses, social security numbers and everything else someone might type into a search box.

The New York Times

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority

In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group.



Ethics in AI Organizations

AI Ethics Lab	aiethicslab.com
AI4Good	ai4good.org
AINOW	annowinstitute.org
Data & Society	datasociety.net
IIT AI Ethics Working Group	ai.iit.edu/ethics

Thanks!

aculotta@iit.edu
tapilab.github.io