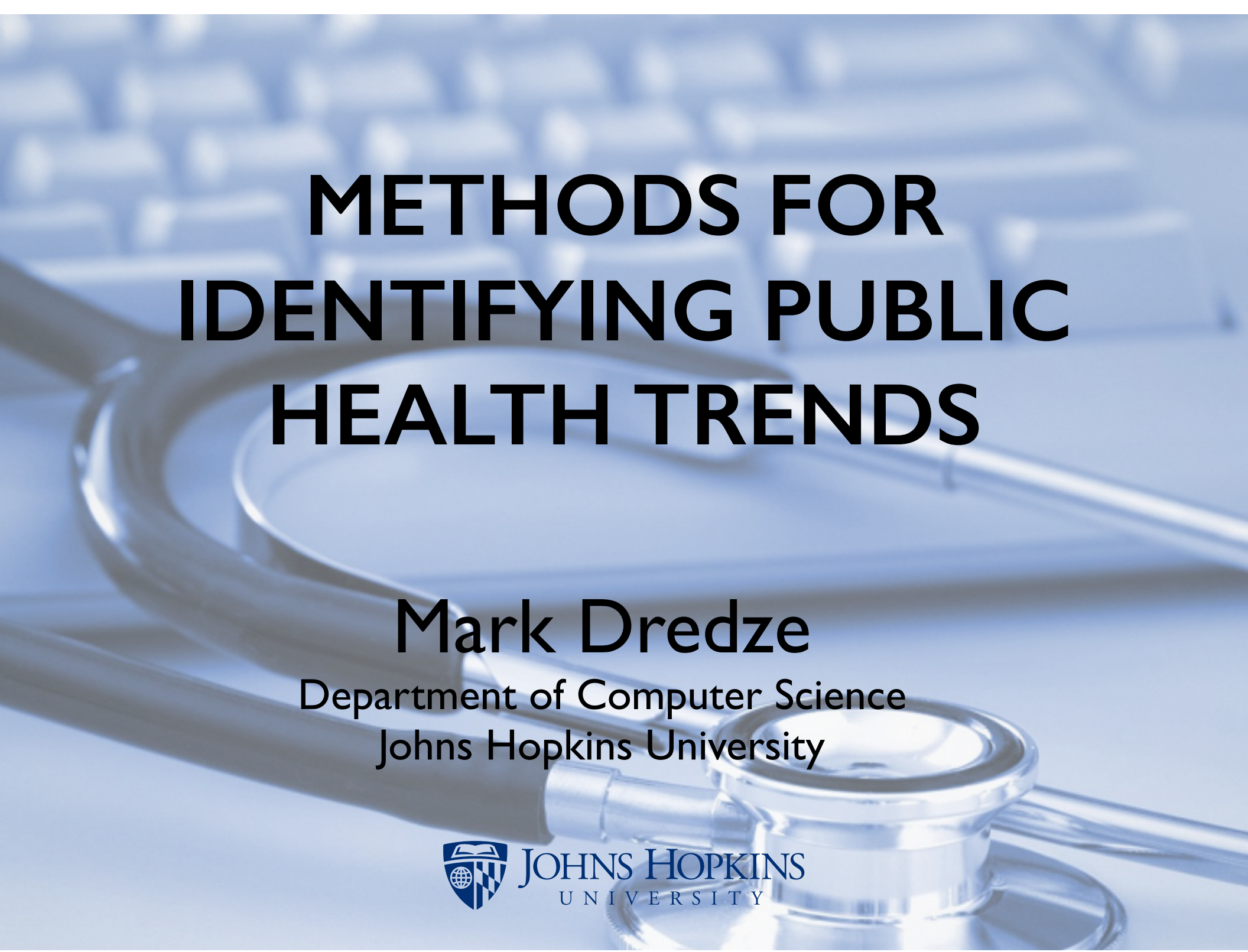


METHODS FOR IDENTIFYING PUBLIC HEALTH TRENDS

Mark Dredze
Department of Computer Science
Johns Hopkins University

JOHNS HOPKINS UNIVERSITY

disease surveillance

self medicating

vaccination

PUBLIC HEALTH

The prevention of disease, prolonging of life and promotion of health

education

tobacco use

drug use

PUBLIC HEALTH CYCLE

Surveillance

REQUIRES:
DATA ON THE POPULATION

Population

Doctors

Intervention

WEB DATA

PUBLIC HEALTH CYCLE

Surveillance

Population

Doctors

Intervention

PUBLIC HEALTH CYCLE

Surveillance

Find health trends

- Examples from public health
- Methods from computer science

PUBLIC AWARENESS

DO AWARENESS CAMPAIGNS WORK?


American Cancer Society
GREAT AMERICAN
SMOKEOUT


BE A
QUITTER
November 15, 2012
The Great American Smokeout®
Call us at 1-800-227-2345


JUST FOR 1 DAY,
WHAT IF
YOU STOPPED
SMOKING?


Quit Tobacco
We can Help!
QUIT
THE GREAT AMERICAN
Smoke Out

HOW MANY PEOPLE QUIT SMOKING?

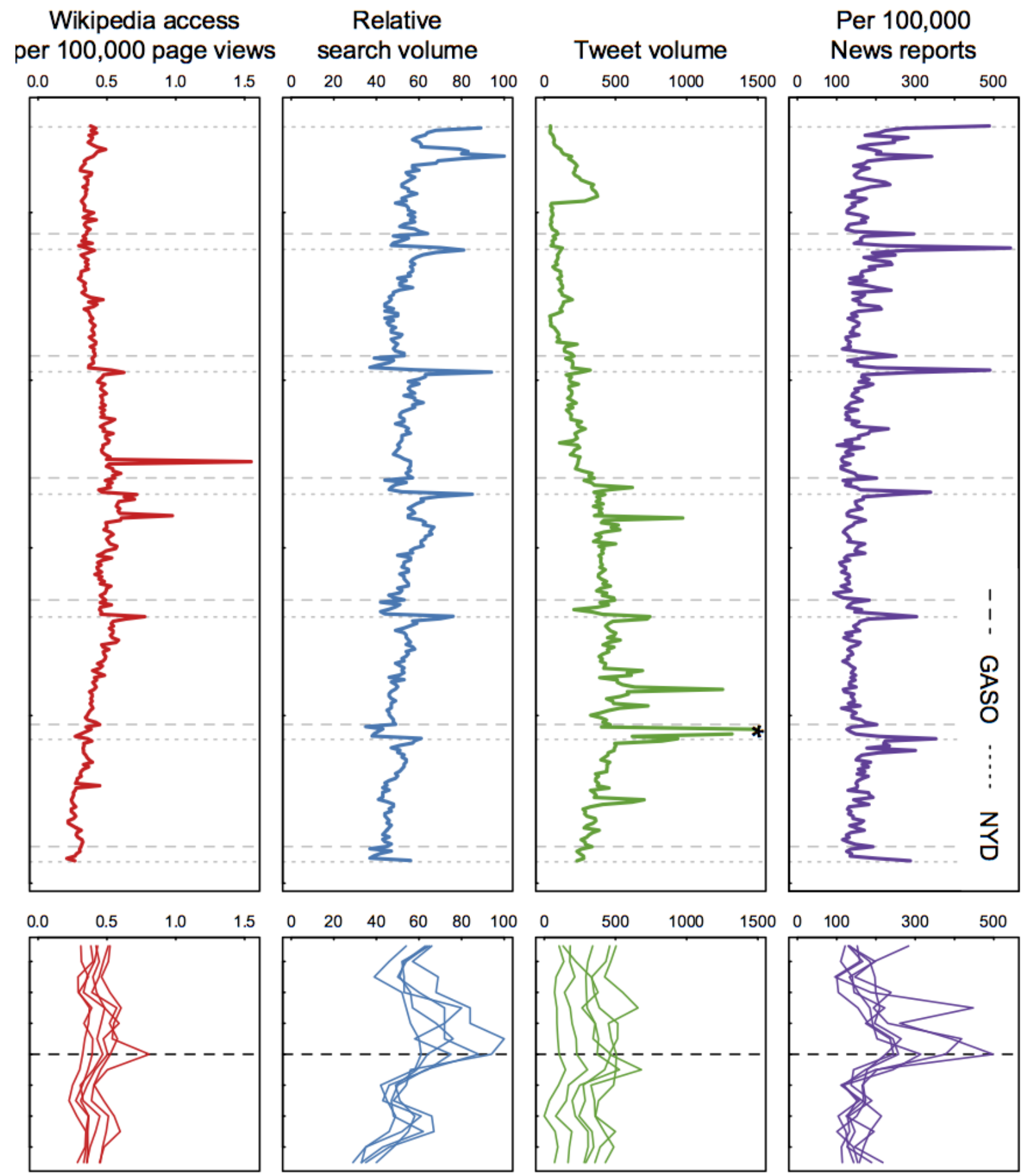
MEASURE ONLINE BEHAVIOR

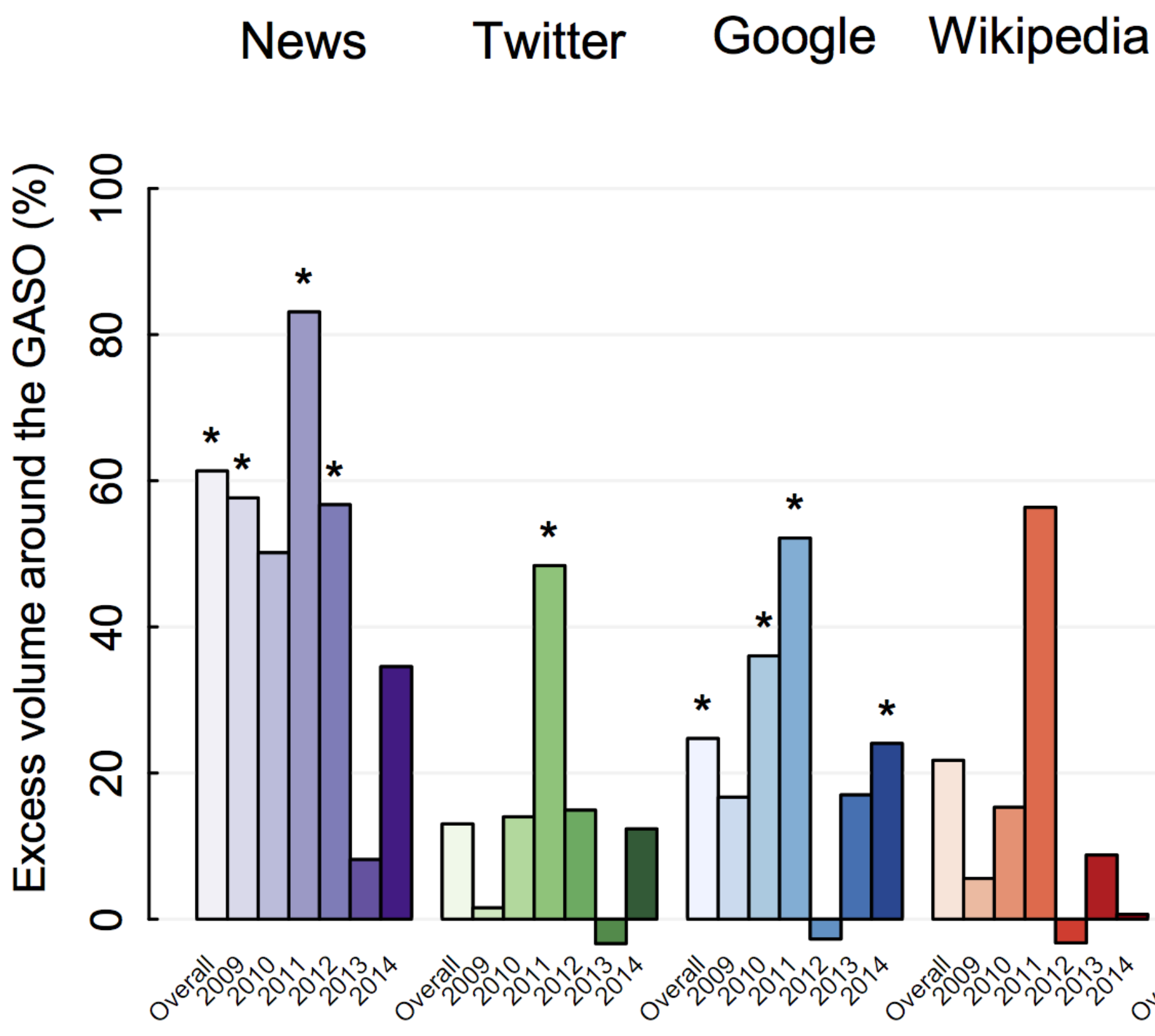
- Media covers awareness event
- People go online to find information about tobacco cessation
- Use these digital signals to quantify impact of GASO

ONLINE DATA SOURCES

- News articles
- Social media: Twitter
- Google searches
- Wikipedia page views

Google Trends


Wikipedia access
per 100,000 page views
0.0
0.5
1.0
1.5
Relative
search volume
0
20
40
60
80
100
Tweet volume
0
500
1000
1500
Per 100,000
News reports
0
100
300
500
GASO
NYD


News
Twitter
Google
Wikipedia
Excess volume around the GASO (%)
0
20
40
60
80
100
Overall
2009
2010
2011
2012
2013
2014

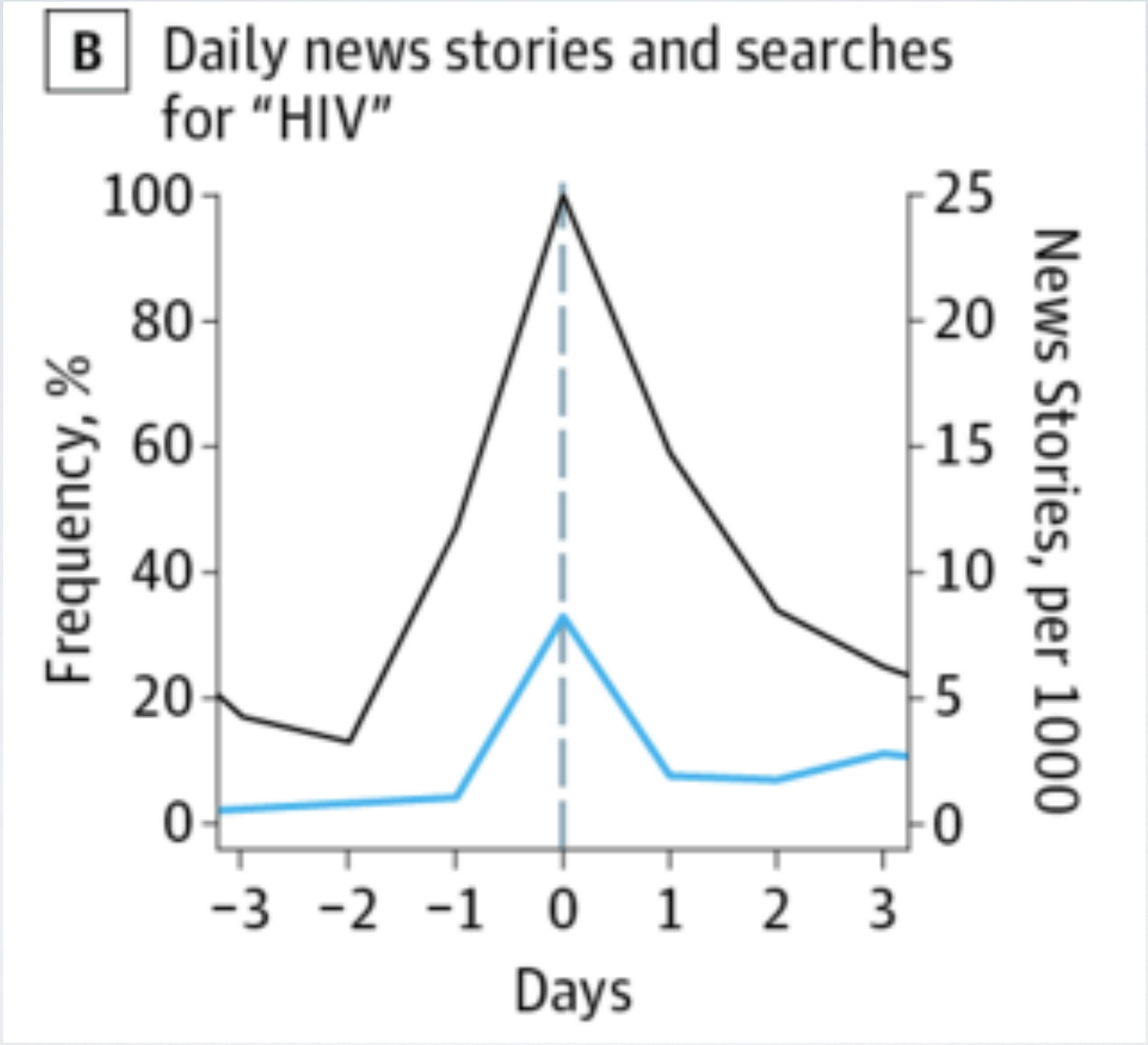
AWARENESS EVENTS

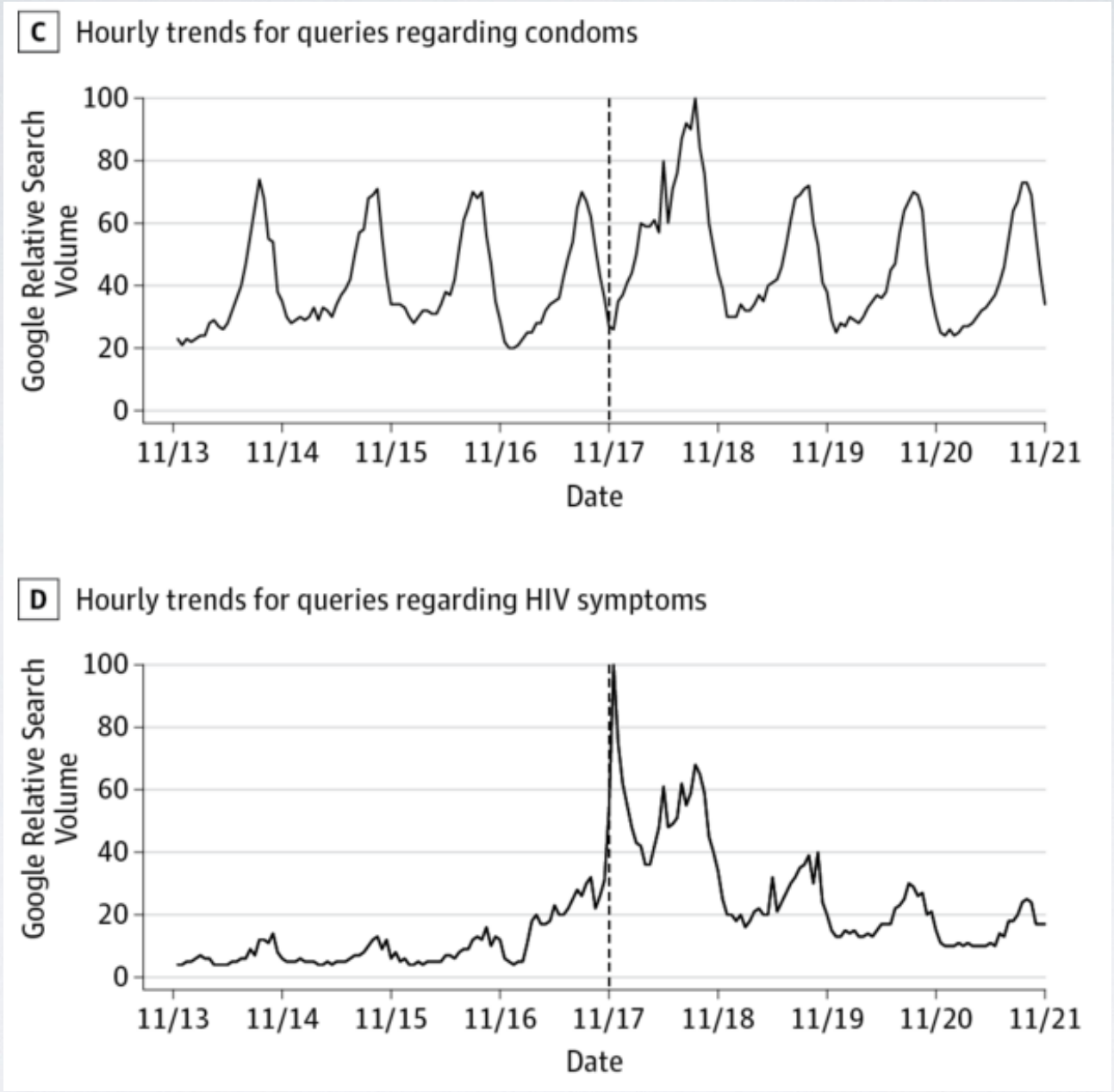
- Not all awareness events are planned

HIV PUBLIC HEALTH STRATEGY

ONLINE DATA SOURCES

- News articles (Bloomberg Terminal)
  - Sheen and HIV
- Google searches
  - Searches for HIV, condom, symptom, testing


B
Daily news stories and searches for "HIV"
Frequency, %
100
80
60
40
20
0
25
20
15
10
5
0
News Stories, per 1000
−3
−2
−1
0
1
2
3
Days


C
Hourly trends for queries regarding condoms
Google Relative Search Volume
100
80
60
40
20
0
11/13
11/14
11/15
11/16
11/17
11/18
11/19
11/20
11/21
Date
D
Hourly trends for queries regarding HIV symptoms
Google Relative Search Volume
100
80
60
40
20
0
11/13
11/14
11/15
11/16
11/17
11/18
11/19
11/20
11/21
Date

Charlie Sheen
@charliesheen

Follow

honored, humbled,
inspired,
hopeful...

bit.ly/1VzvcRO
x
©

News and Internet Searches About HIV After Celebrity Discl...
This study examines trends for information about human immunodeficiency virus in the mass media and on the Internet after celebrity disclosure of infection with the virus.
archinte.jamanetwork.com

RETWEETS 227 LIKES 876

1:53 PM - 22 Feb 2016


ZIKA VIRUS
James Gathany / CDC / MGN


ZIKA
Artralgias
Exantema maculopapular (rash)
Edema en articulaciones
Petequias
Neurológicos (especifique):

Brazilians not buying Zika excuse for babies with shrunken brains

January 25, 2016 theunhivedmind Leave a comment

Jan 25 2016

http://82.221.129.208/ifyouareinamericayouprobablycantseethisj9.html

Jim Stone

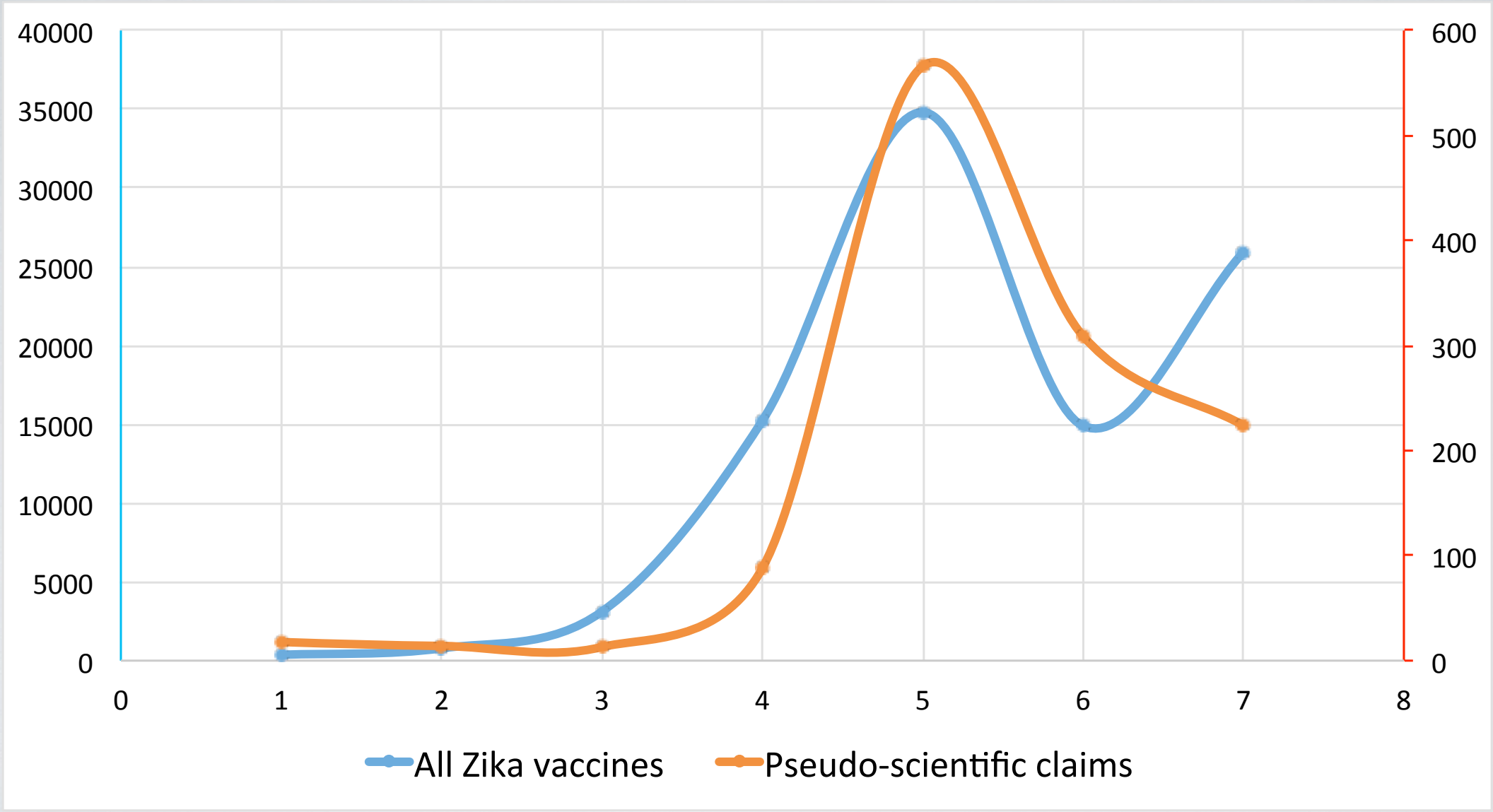
Brazilians not buying Zika excuse for babies with shrunken brains

Over 4,000 babies have now been born in Brazil with shrunken brains since November 1 2015. Brazil normally gets approximately 150 cases of this type of birth defect per year, which means that if this all happened in less than a three month time window, abnormal births of this type have increased by approximately 13,000 percent. HERE IS A KEY REPORT


2015
Micro Cephalic Babies Conceived
May DTaP Vaxx Begins
October Micro-cephaly

PSEUDO SCIENTIFIC BELIEFS

- "GMO Mosquitos are the cause of the zika virus."
- "#Zika may help accelerate Sterilization in the US, and with the use of GMO Mosquitoes sterility will be delivered to you, #Depopulation#NATO"
- 0.1 babies had zika, 100% had DTAP given to mother during pregancy? Wonder which caused this?
- Factors:Those pregnant women were #Vaxxed=dtap,GMO mozzies released,pesticides put in drinking water so blame #Zika


40000
35000
30000
25000
20000
15000
10000
5000
0
600
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8
All Zika vaccines
Pseudo-scientific claims

- Examples from public health
- Methods from computer science

TREND IDENTIFICATION

- Identify health trends in social media
  - Organize data into themes or topics
  - Identify patterns in topics over time, location, sub-population

APPROACH

- Topic modeling for Twitter
  - Probabilistic models of text, e.g. latent Dirichlet allocation
  - Identify major themes in corpus
  - Yields human interpretable topics
  - Identify/correlate topics with survey responses
  - Unsupervised: no tweet level supervision

SUPERVISION OF TOPIC MODELS

- Standard topic models: unsupervised
- Supervised LDA: Document labels
- What do we have?
  - Supervision on aggregated documents
  - e.g. 60% of people in this location think X

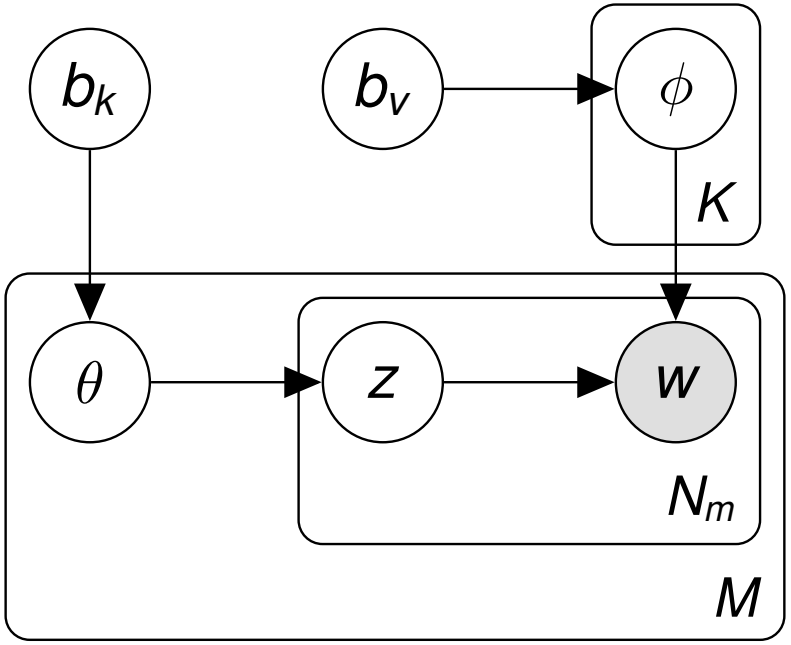
COLLECTIVE SUPERVISION

- Train models at the document level to make predictions at the population level
- Data
  - Telephone survey results
  - Prediction: estimating survey values for populations from social media features
  - Topic models can learn low-dimensional, generalizable features that can be used in predictive models
  - Analysis: Topic models are interpretable: we can better understand trends by viewing topics

APPROACH

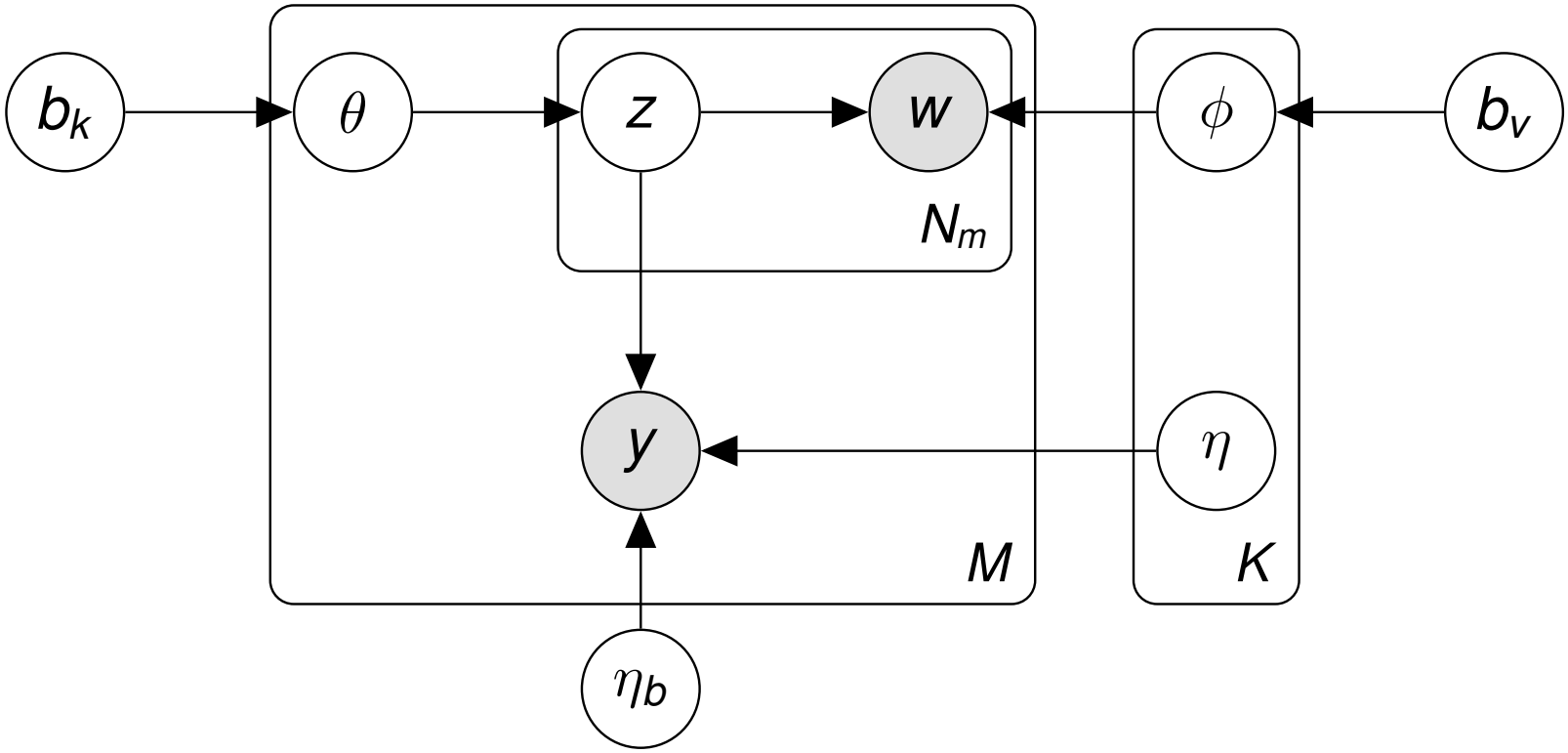
- Modify topic models to incorporate collective supervision
- Extend different types of topic models in different ways, and compare
- Evaluate effectiveness at predicting public health telephone surveys

Latent Dirichlet Allocation (LDA)



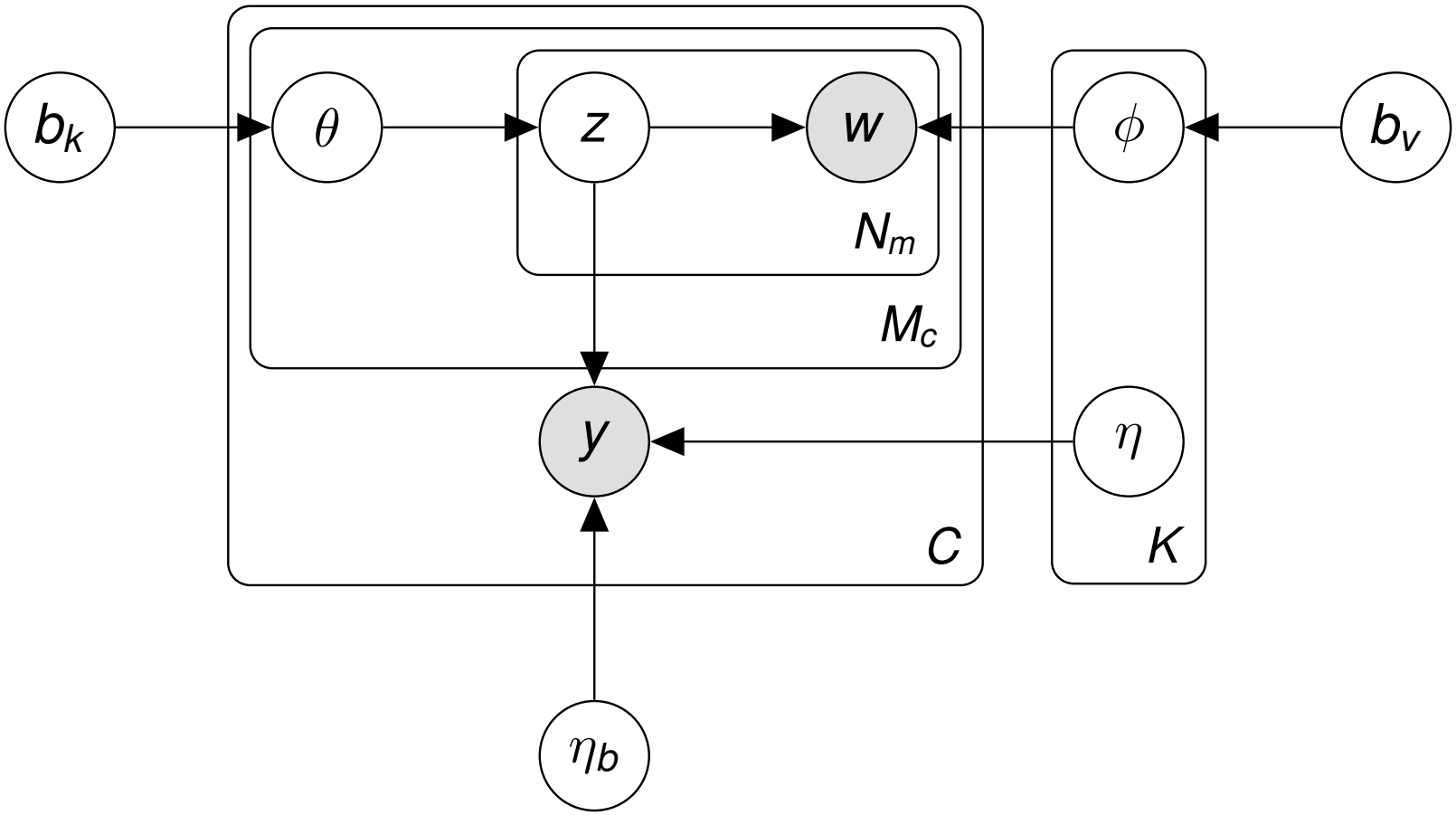
- $\tilde{\theta}_{mk} = \exp(b_k); \theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$
- $\tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \text{Dirichlet}(\tilde{\phi}_k)$
- $z_{mn} \sim \theta_m; w_{mn} \sim \phi_{z_{mn}}$

Supervised LDA (Downstream-sLDA)



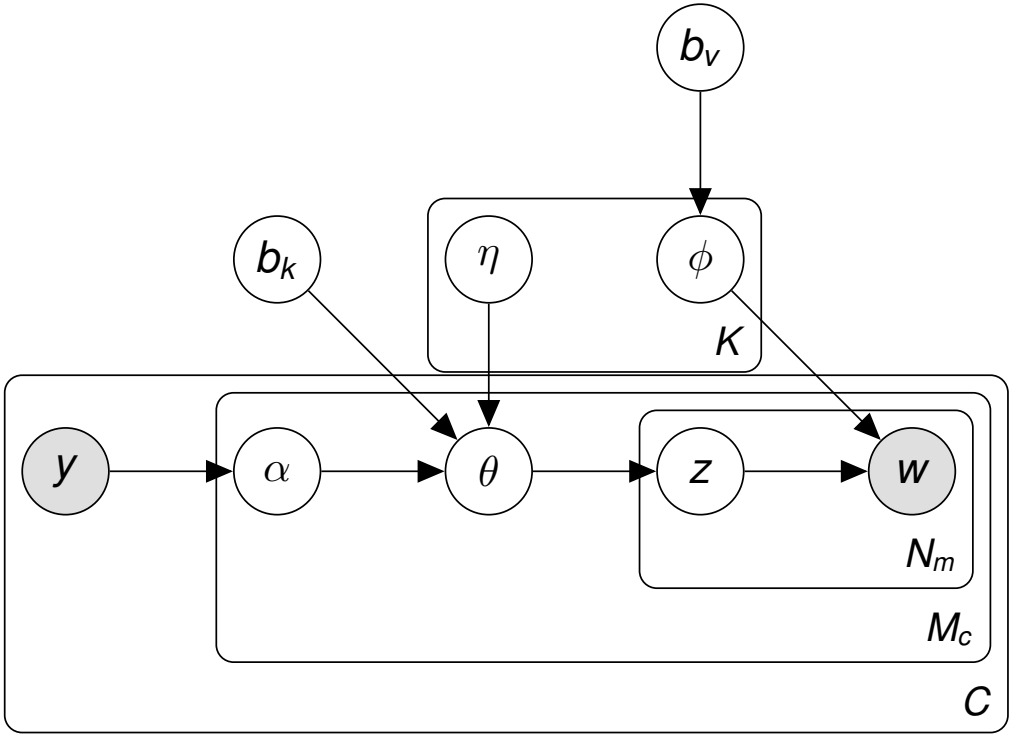
- Supervised LDA (sLDA) [2]
- $\overline{z}_{mk}$ is the average proportion of topic k in document m
- $y_m \sim \mathcal{N}(\eta_b + \eta^T \overline{z}_m, \sigma_y^2)$

Collectively Supervised LDA (Downstream-collective)



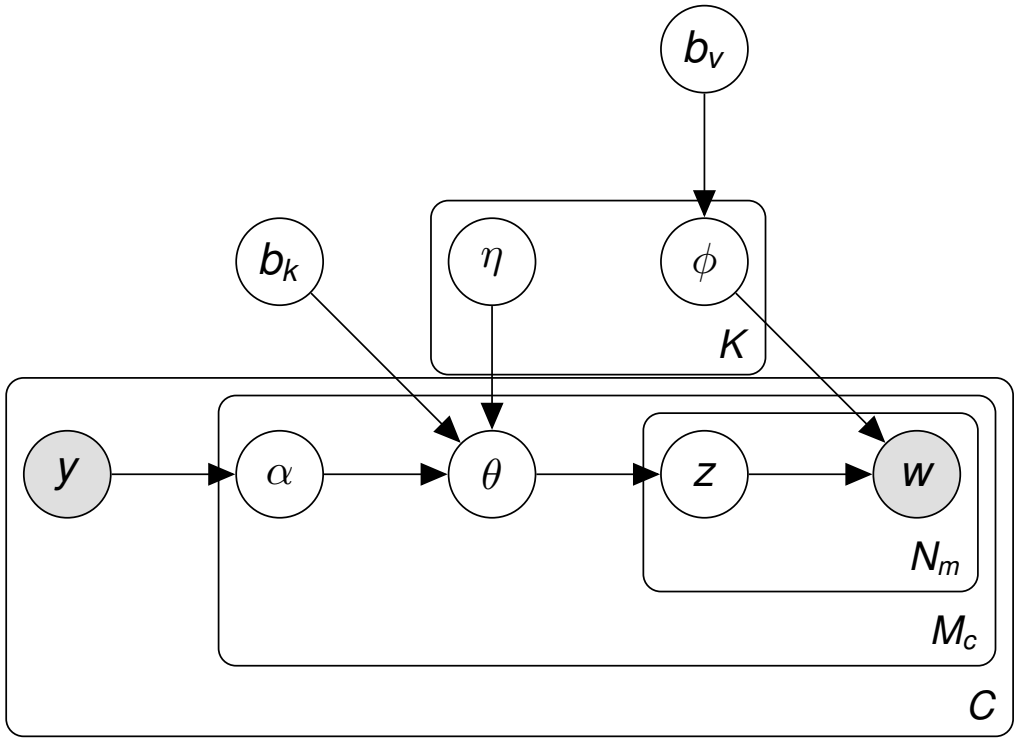
- Let $\bar{z}_{jk}$ be the average proportion of topic k in collection j
- $y_j \sim \mathcal{N}(\eta_b + \eta^T \bar{z}_j, \sigma_y^2)$
- Supervised LDA is a special case of this, where each document has its own unique collection ID

Dirichlet Multinomial Regression (Upstream)



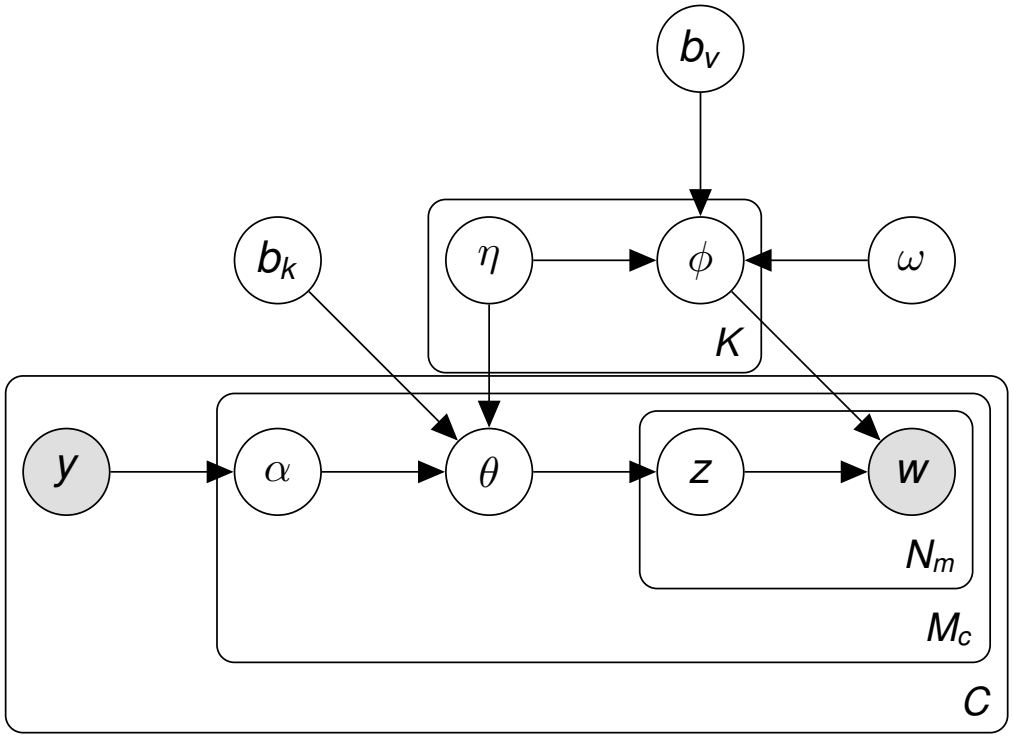
- Dirichlet-multinomial regression (DMR) [3]
- $\alpha_m = y_{c_m}$, feature value associated with document's collection c_m
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k); \theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$
- $\tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \text{Dirichlet}(\tilde{\phi}_k)$

DMR with adaptive supervision (Upstream-ada)



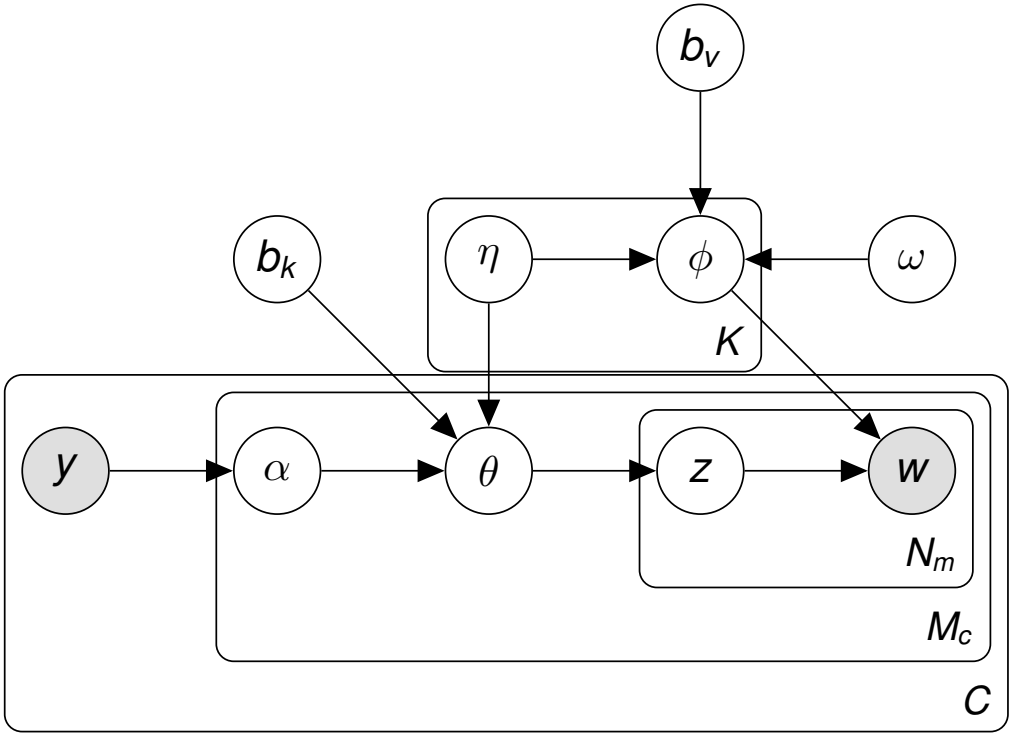
- $\alpha_m \sim \mathcal{N}(y_{c_m}, \sigma_\alpha^2)$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k)$
- $\tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \text{Dirichlet}(\tilde{\phi}_k)$
- Document value can deviate from given input – can help infer likely values when supervision is noisy or missing.

DMR with word priors (Upstream-words)



- $\alpha_m = y_{c_m}$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k)$
- $\tilde{\phi}_{kv} = \exp(b_v + \omega_v \eta_k)$
- Supervision affects priors over words. Extension to DMR known as SPRITE [7].

DMR + adaptive + word prior (Upstream-ada-words)



- Combined upstream model
- $\alpha_m \sim \mathcal{N}(y_{c_m}, \sigma_\alpha)$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k)$
- $\tilde{\phi}_{kv} = \exp(b_v + \omega_v \eta_k)$

Surveys

- Behavioral Risk Factor Surveillance System: annual survey by US federal government to learn about health/behavior of population.
- We selected three questions from BRFSS phone surveys:
  - **Guns**: Do you have a firearm in your house? (2001)
  - **Vaccines**: Have you had a flu shot in the past year? (2013)
  - **Smoking**: Are you a current smoker? (2013)
- Survey responses are aggregated at the level of US state.

Twitter Data

Dataset	Vocab	BRFSS
Guns	12,358	Owns firearm
Vaccines	13,451	Had flu shot
Smoking	13,394	Current smoker

- 100,000 tweets per dataset (filtered by relevant keywords)
  - collected between Dec. 2012 - Jan. 2015
- Identified as English using langid
  `https://github.com/saffsd/langid.py`
- Stopwords removed and low-frequency tokens excluded
- Location inferred using Carmen
  `https://github.com/mdredze/carmen-python`

Supervision

For each dataset:

- Each collection is defined as the set of tweets per US state
  - 50 collections
- Each collection's y_c value is the proportion respondents answering "Yes" to the BRFSS question

Predicting survey values:

- L2-regularized linear regression model
- Features: mean topic distributions θ per collection

Experiment Details

- Lots of hyperparameters – selected hyperparameters that maximized perplexity on heldout sample
- Optimized each model using Spearmint: `https://github.com/JasperSnoek/spearmint`
- Fit models using Gibbs sampling with AdaGrad for parameter (η) optimization
- Prediction task tuned with 5-fold cross validation: 80% train, 10% dev, 10% test.

Results

Features	Model	Guns		Vaccines		Smoking	
		RMSE	Perplexity	RMSE	Perplexity	RMSE	Perplexity
None	LDA	17.44	2313 ($\pm$52)	8.67	2524 ($\pm$20)	4.50	2118 ($\pm$5)
Survey	Upstream	15.37	1529 ($\pm$12)	6.54	1552 ($\pm$11)	3.41	1375 ($\pm$6)
	Upstream-words	11.50	**1429** ($\pm$22)	6.37	1511 ($\pm$57)	3.41	1374 ($\pm$2)
	Upstream-ada	11.48	1506 ($\pm$67)	5.82	**1493** ($\pm$49)	3.41	1348 ($\pm$6)
	Upstream-ada-words	**11.47**	1535 ($\pm$28)	7.20	1577 ($\pm$15)	**3.40**	1375 ($\pm$3)
	Downstream-sLDA	11.52	1561 ($\pm$22)	11.22	1684 ($\pm$7)	3.95	1412 ($\pm$3)
	Downstream-collective	12.81	1573 ($\pm$20)	9.17	1684 ($\pm$6)	4.35	1412 ($\pm$4)

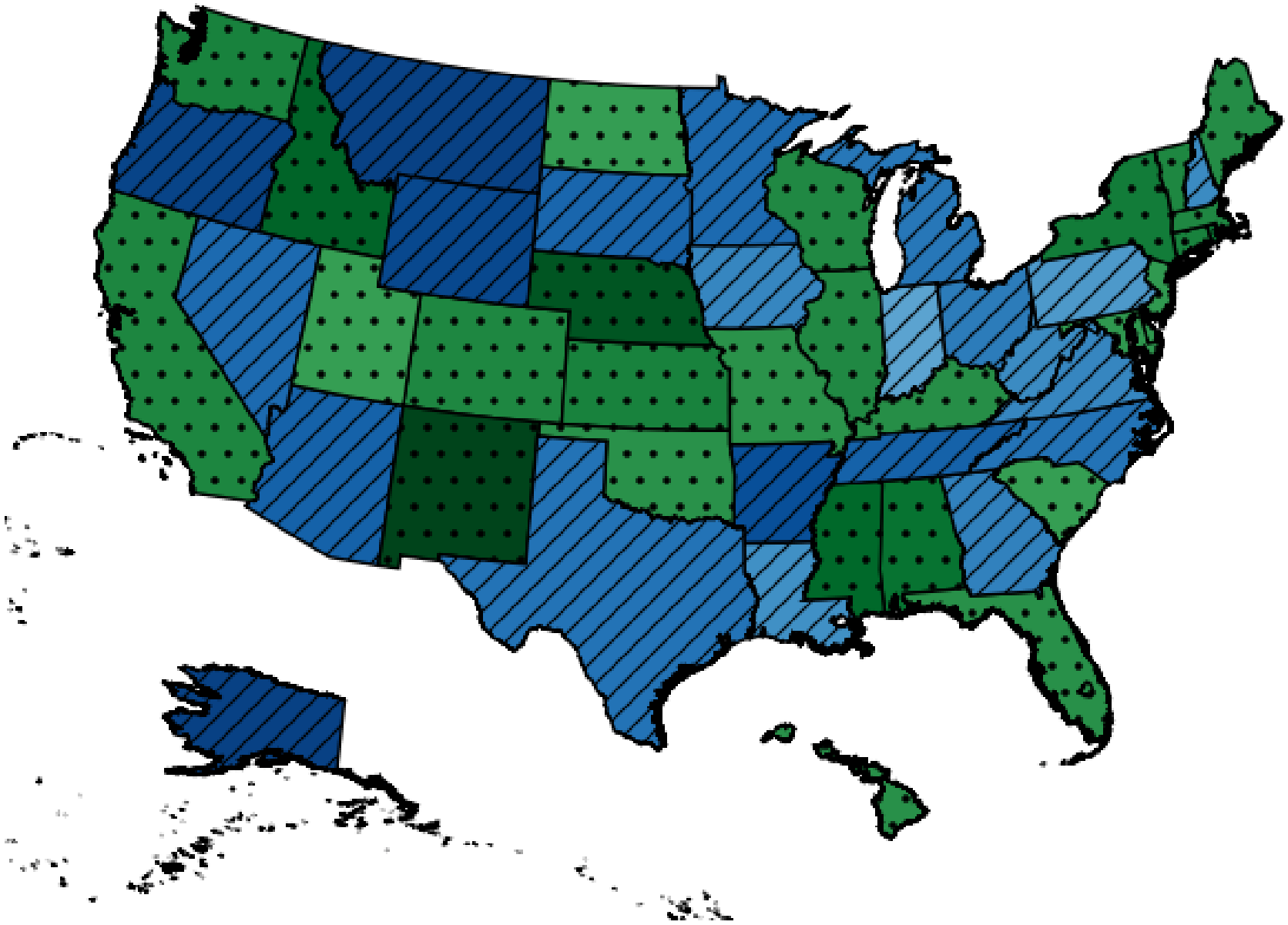
Use Case – Support for Universal Background Checks

- UBCs were a big US political issue in 2013, when national gun control legislation was floated
- We collected surveys on support for UBCs for 22 states from various polls (mostly Public Policy Polling)
- Baseline: use older 2001 survey of proportion households containing a firearm

Use Case – Support for Universal Background Checks

Features	Model	RMSE (2001 Y included)	RMSE (2001 Y omitted)
None	No model	7.26	7.59
	Bag of words	5.16	7.31
	LDA	6.40	7.59
Survey	Upstream-ada-words	**5.11**	**5.48**

Use Case – Support for Universal Background Checks



Thank You

http://socialmediahealthresearch.org

mdredze@cs.jhu.edu www.dredze.com

John Ayers
David Broniatowski
Michael Paul
Adrian Benton
Michael Smith
Glen Coppersmith
Craig Harman

Ben Althouse
Morgan Johnson
Jon-Patrick Allem
Matt Childers
Karen Hilyard
Angie Chen