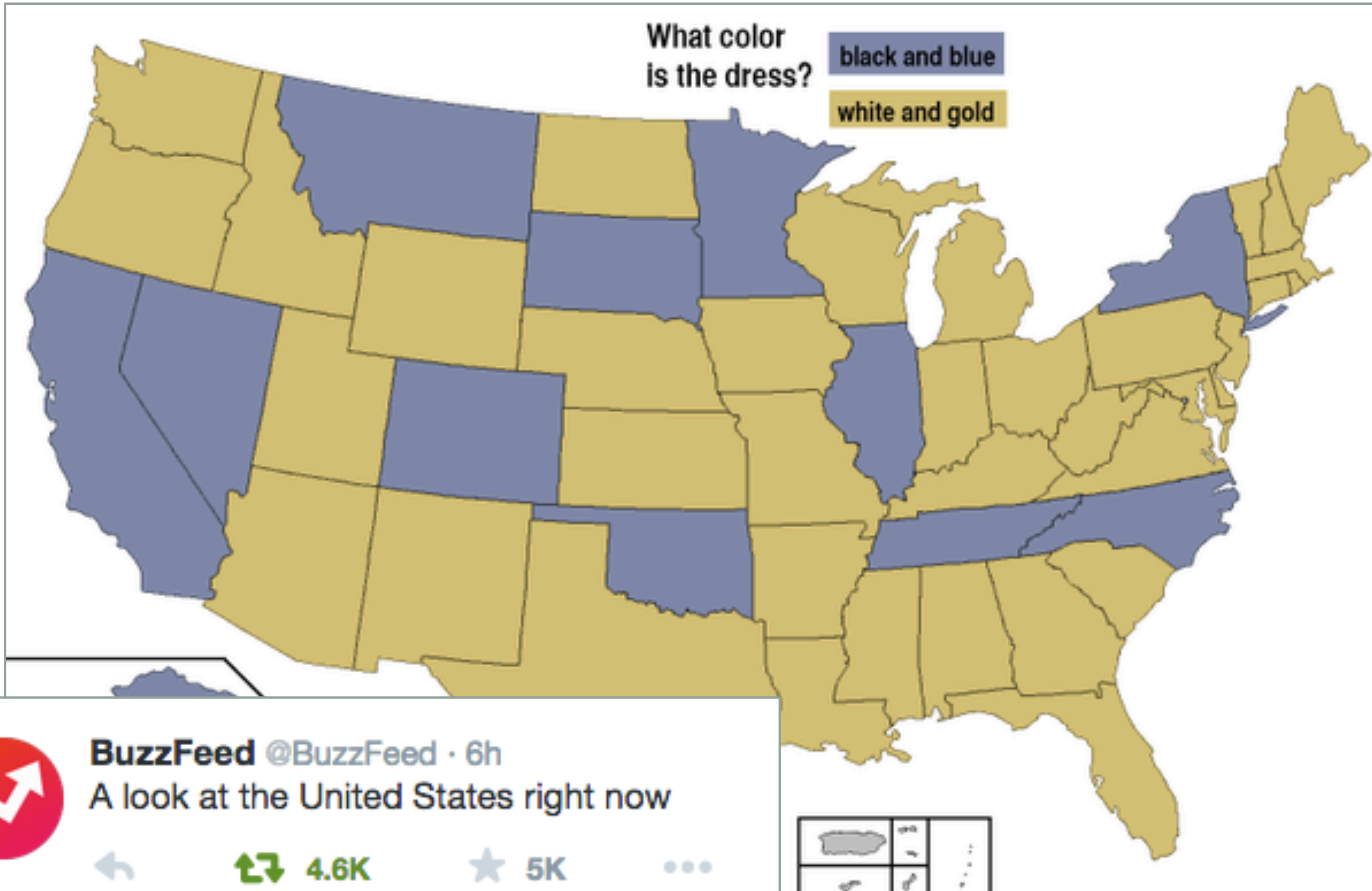
TOPIC MODELING WITH STRUCTURED PRIORS FOR TEXT-DRIVEN SCIENCE

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY

University of Colorado, Boulder | February 27, 2015

#whiteandgold

#blueandblack

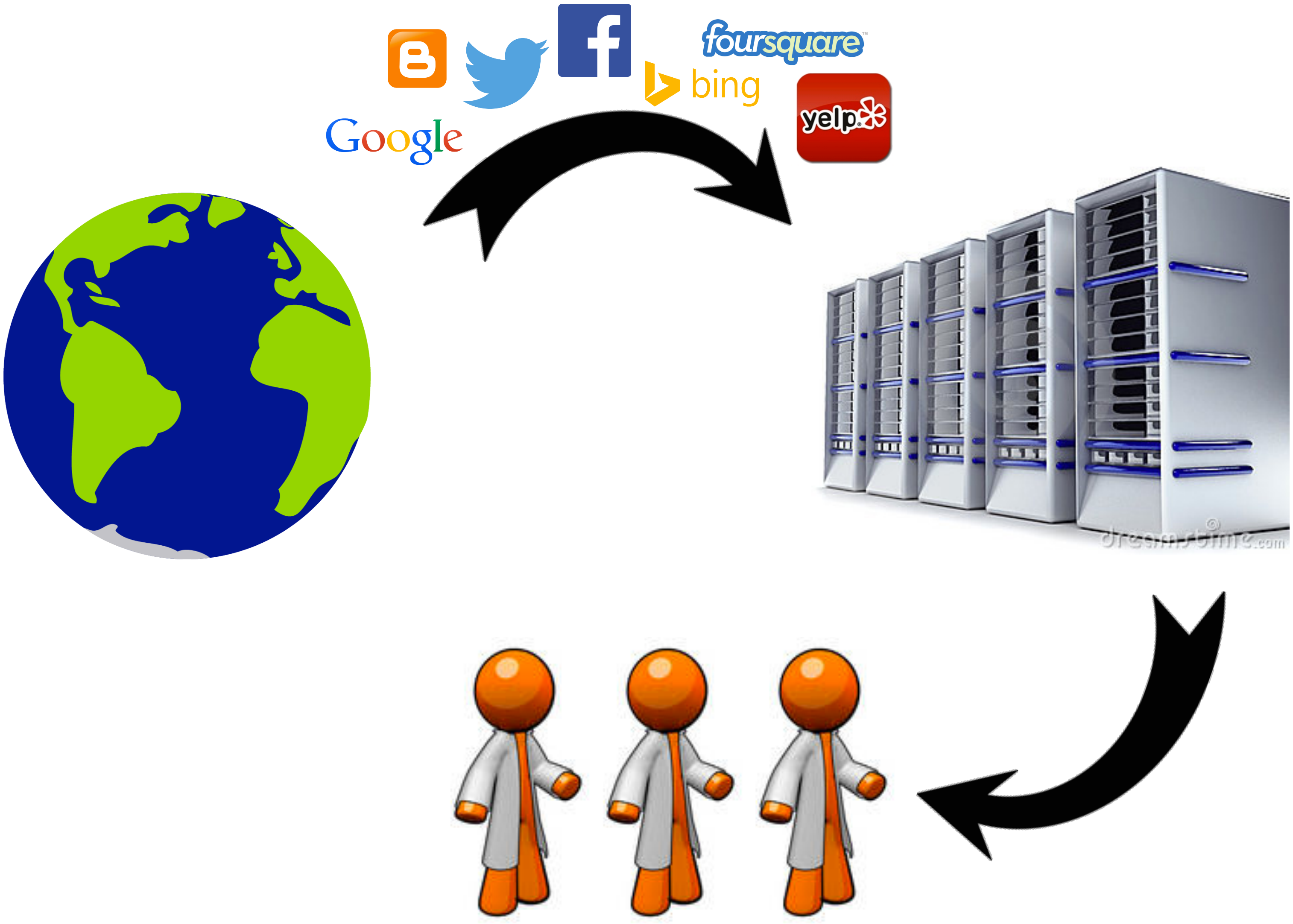


BuzzFeed @BuzzFeed · 6h
A look at the United States right now

4.6K 5K

TEXT AS DATA

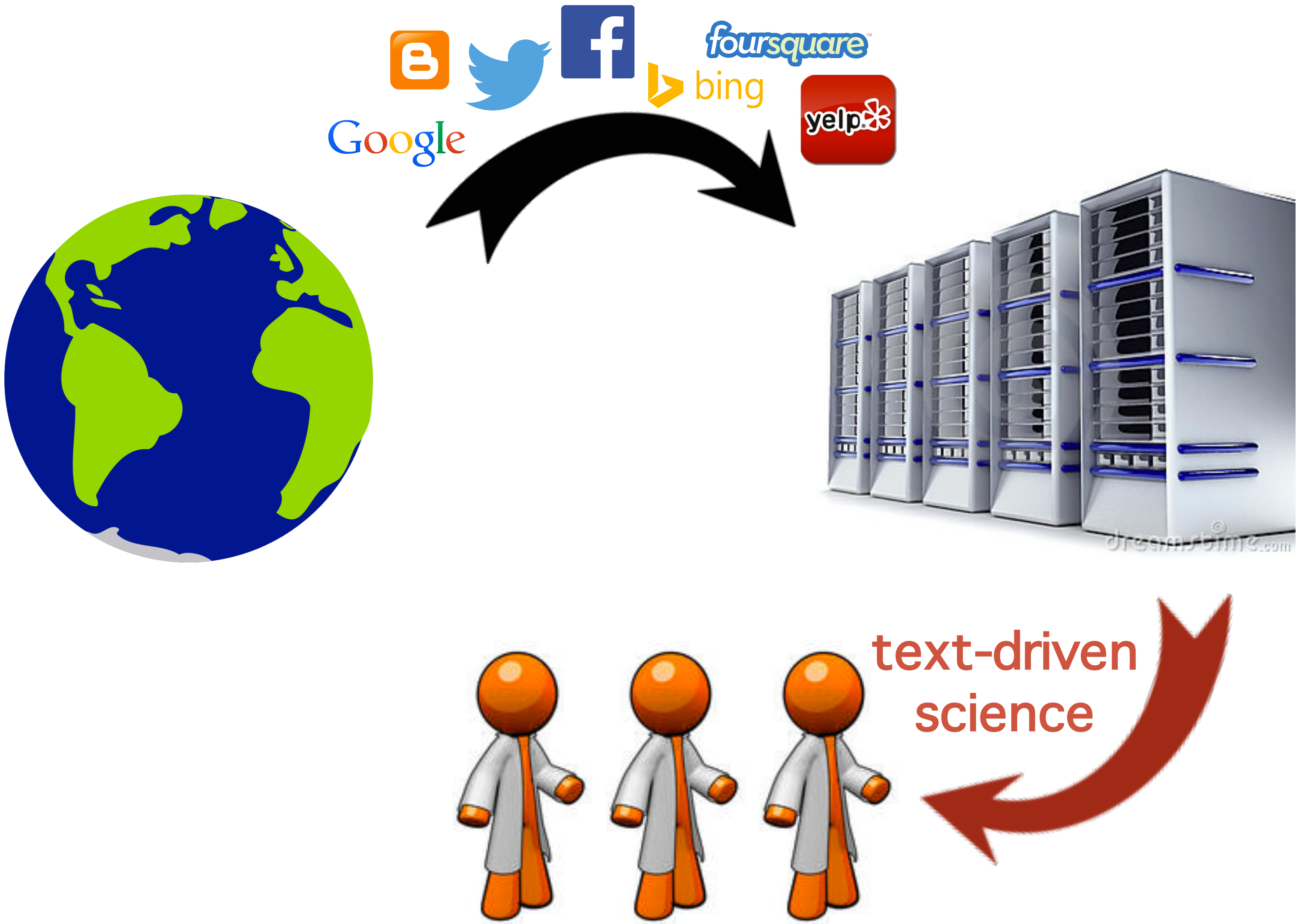
TEXT AS DATA



TEXT AS DATA


Google
bing
foursquare
yelp
natural language
processing (NLP)

TEXT AS DATA



TEXT AS DATA

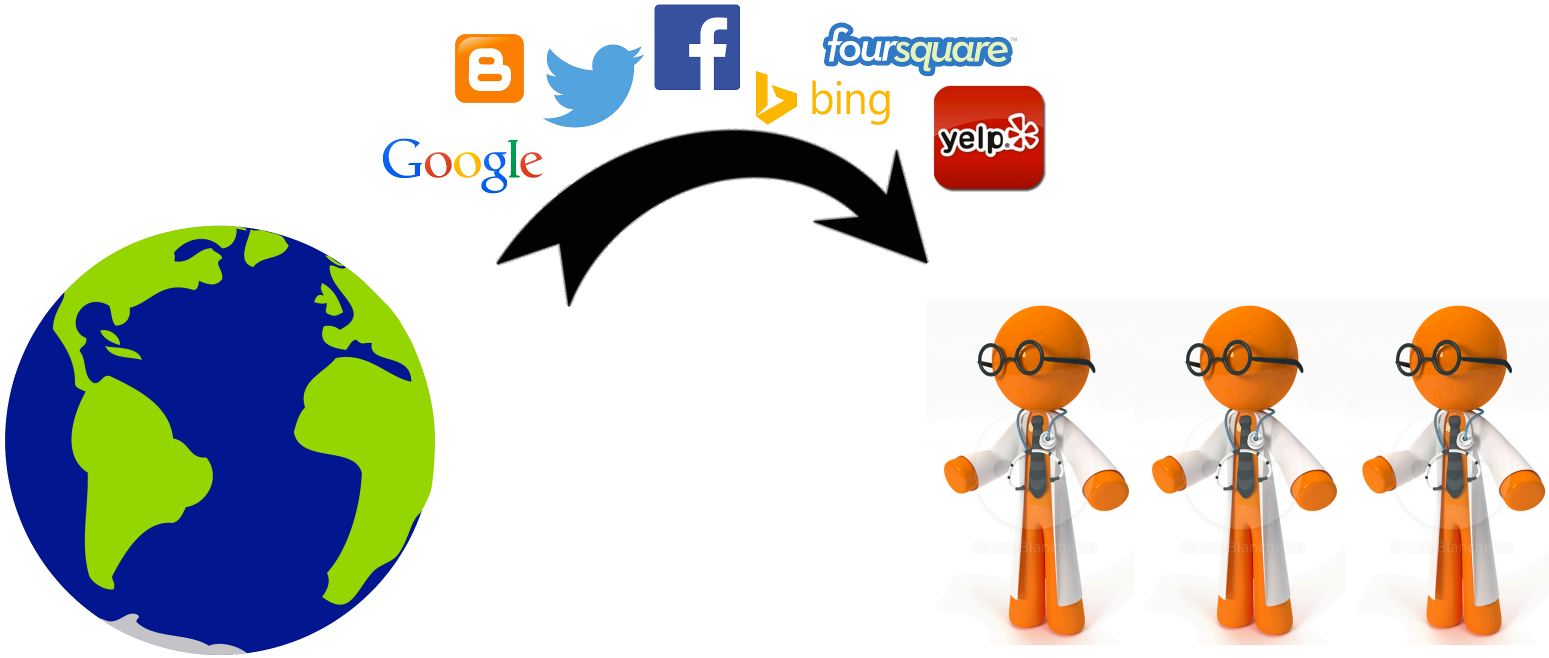
- Computational social science
- Computational journalism
- Crisis informatics
- Public health informatics
- Computational epidemiology

TEXT AS DATA

- Computational social science
- Computational journalism
- Crisis informatics
- Public health informatics
- Computational epidemiology

text-driven science

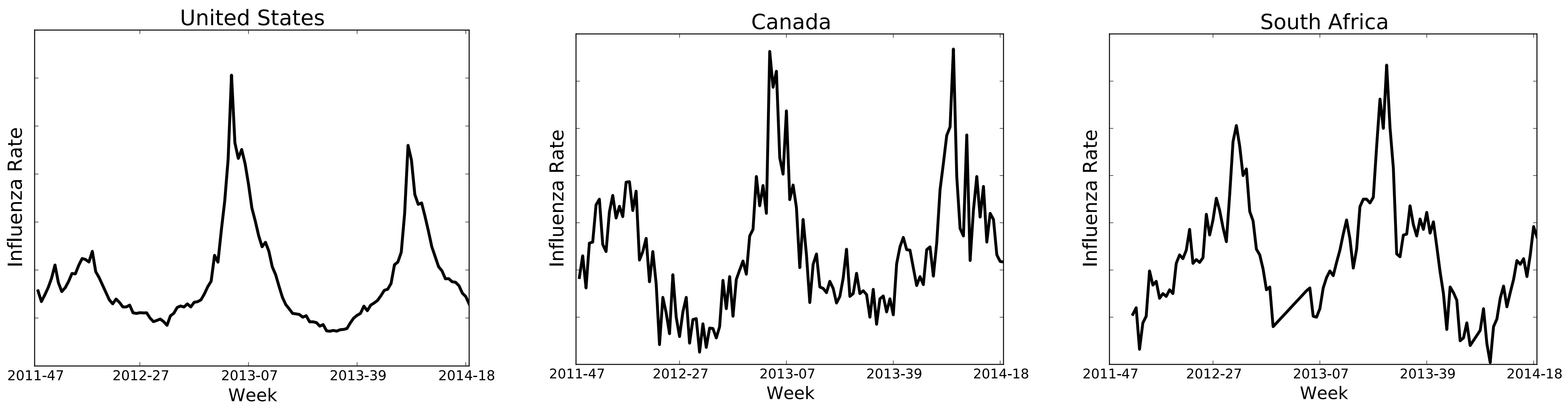
TEXT-DRIVEN PUBLIC HEALTH



Let’s look at some examples!

TEXT-DRIVEN PUBLIC HEALTH — FLU MONITORING

Important public health task: disease surveillance



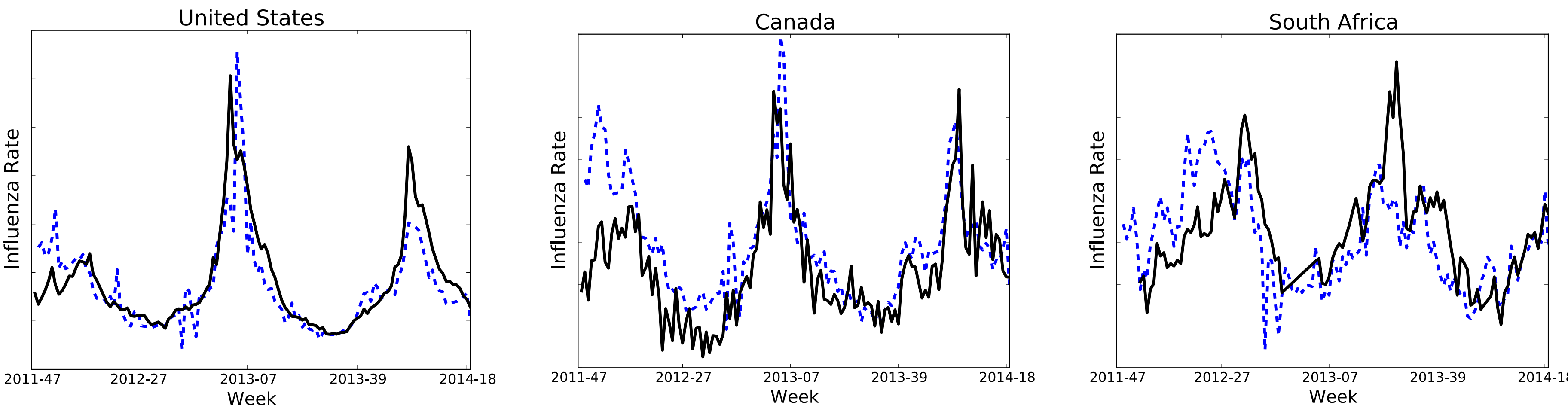
Many health agencies do flu monitoring

- But reports have a delay of ~2 weeks

TEXT-DRIVEN PUBLIC HEALTH

FLU MONITORING

Tracking the spread of influenza through tweets:



Paul, Dredze, Broniatowski (2014) **Twitter improves influenza forecasting.** *PLOS Currents: Outbreaks.*

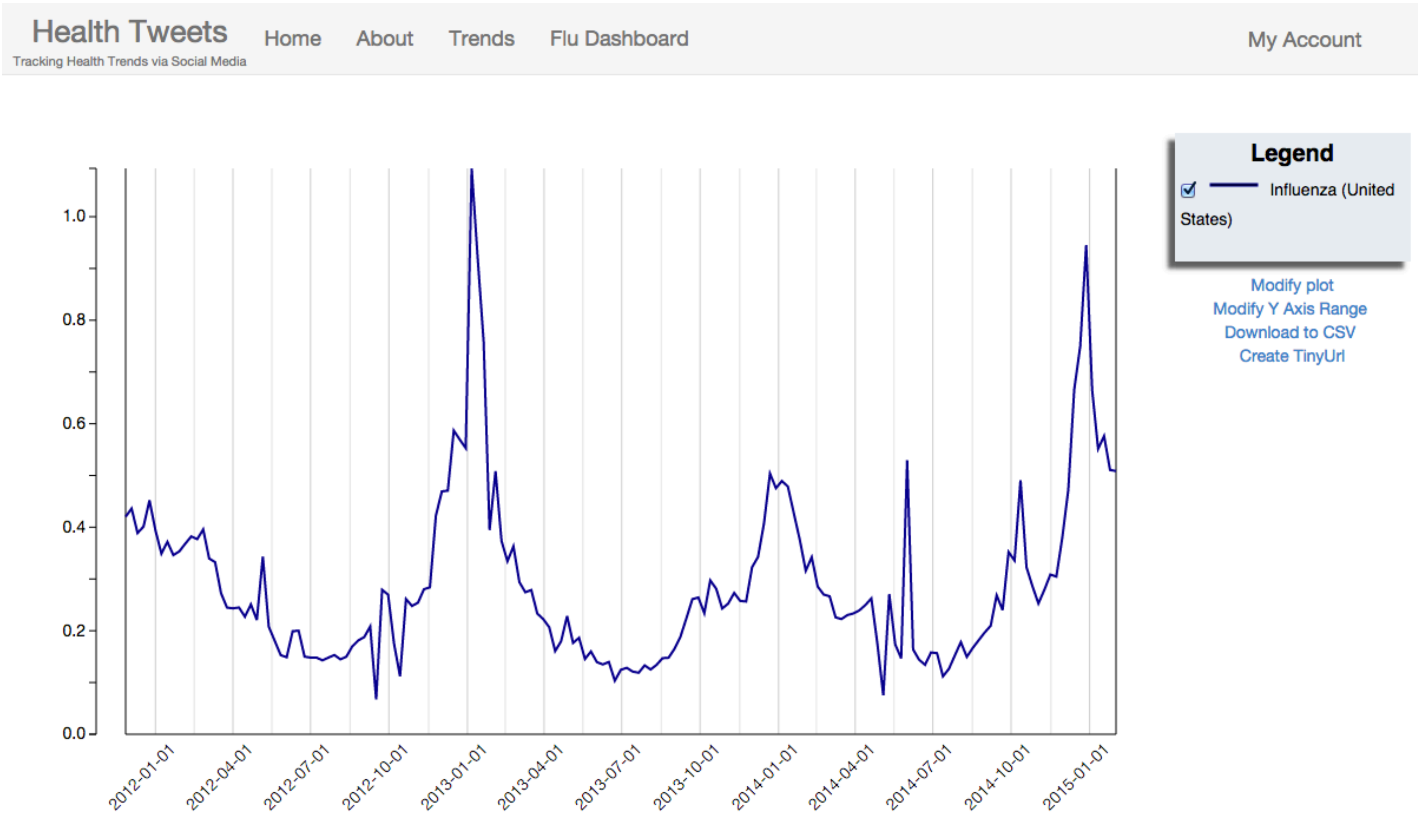
Paul, Dredze, Broniatowski, Generous (2015) **Worldwide influenza surveillance through Twitter.** *AAAI Workshop on the World Wide Web and Public Health Intelligence.*

Lamb, Paul, Dredze (2013) **Separating fact from fear: Tracking flu infections on Twitter.** *NAACL.*

Broniatowski, Paul, Dredze (2013) **National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic.** *PLOS ONE* 8(12): e83672.

TEXT-DRIVEN PUBLIC HEALTH

FLU MONITORING



TEXT-DRIVEN PUBLIC HEALTH

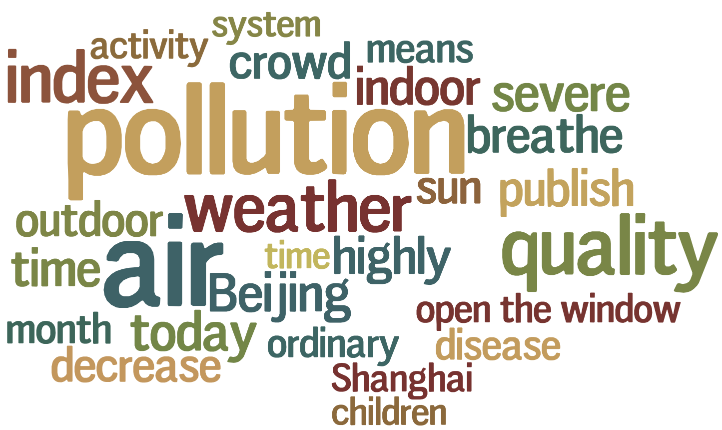
AIR QUALITY

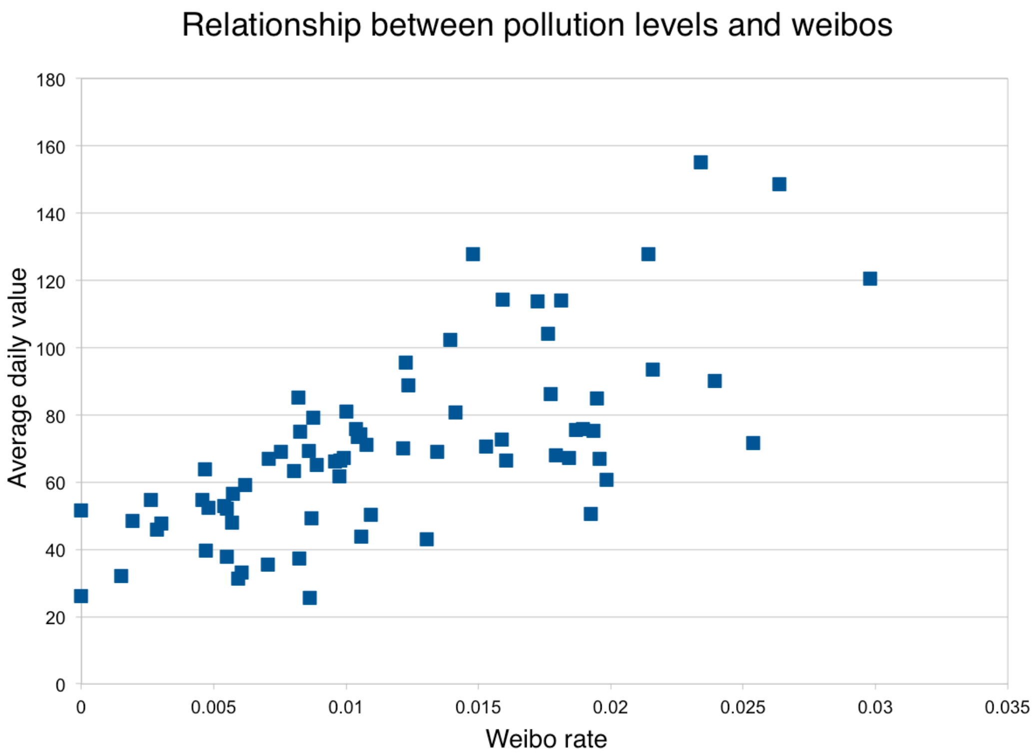
Air pollution in China

Public health tasks:

- Measure pollution levels
- Identify health effects
- Understand public response

Monitoring air pollution through social media:





Wang, Paul, Dredze (2015) Social media as a sensor of air quality and public response in China. *Journal of Medical Internet Research.*

TEXT-DRIVEN PUBLIC HEALTH

DRUG USE

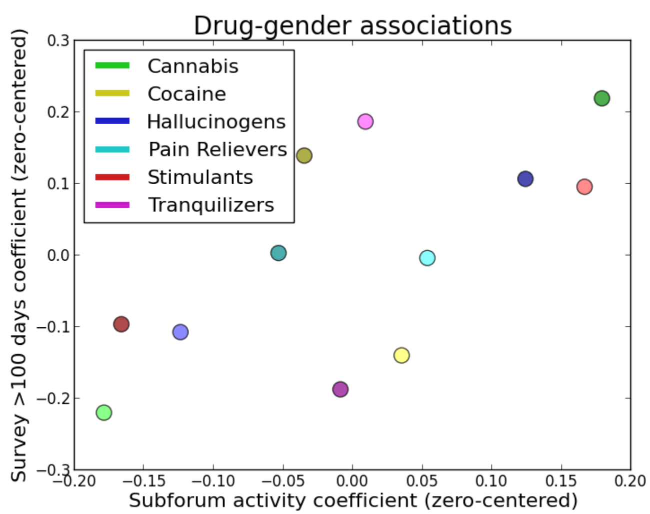
New trends in drug use

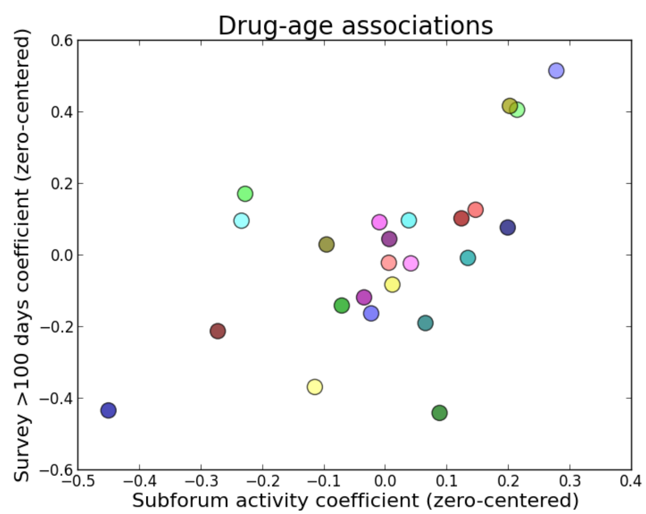
- Record numbers of new drugs recently
- Health officials can be years behind

TEXT-DRIVEN PUBLIC HEALTH

DRUG USE

Analyzing online forums:



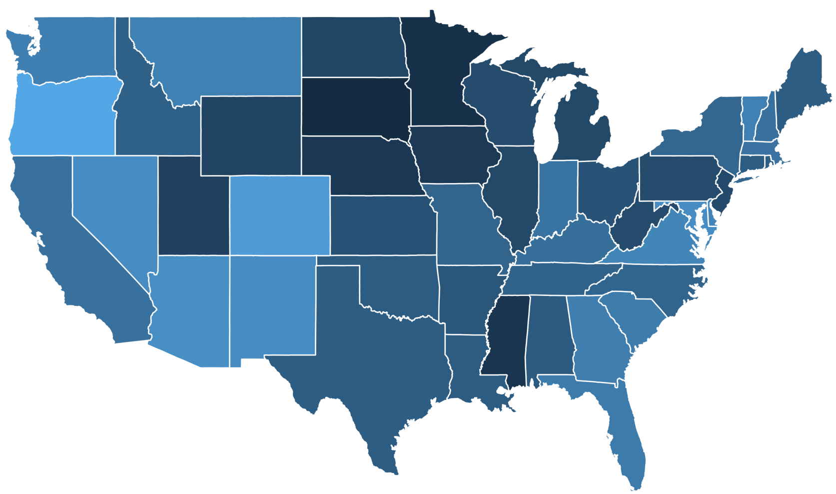


Paul, Dredze (2013) **Summarizing drug experiences with multi-dimensional topic models.** *North American ACL (NAACL).*

Paul, Chisolm, Johnson, Vandrey, Dredze (in preparation) **Who participates in online drug communities? A large-scale analysis of demographic and temporal trends.**

TEXT-DRIVEN PUBLIC HEALTH | HEALTHCARE QUALITY

Understanding healthcare quality from online reviews:



Text from reviews is significantly predictive of external measures of healthcare quality

Paul, Wallace, Dredze (2013) **Analyzing online doctor ratings with a joint topic-sentiment model.** *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI.* CS

Wallace, Paul, Sarkar, Trikalinos, Dredze (2014) **A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews.** *Journal of the American Medical Informatics Association* 21(6), 1098-1103.

TEXT-DRIVEN PUBLIC HEALTH

Many other applications:

- Air pollution in Chinese social media

Wang, Paul, Dredze (2015) **Social media as a sensor of air quality and public response in China.** *Journal of Medical Internet Research*.

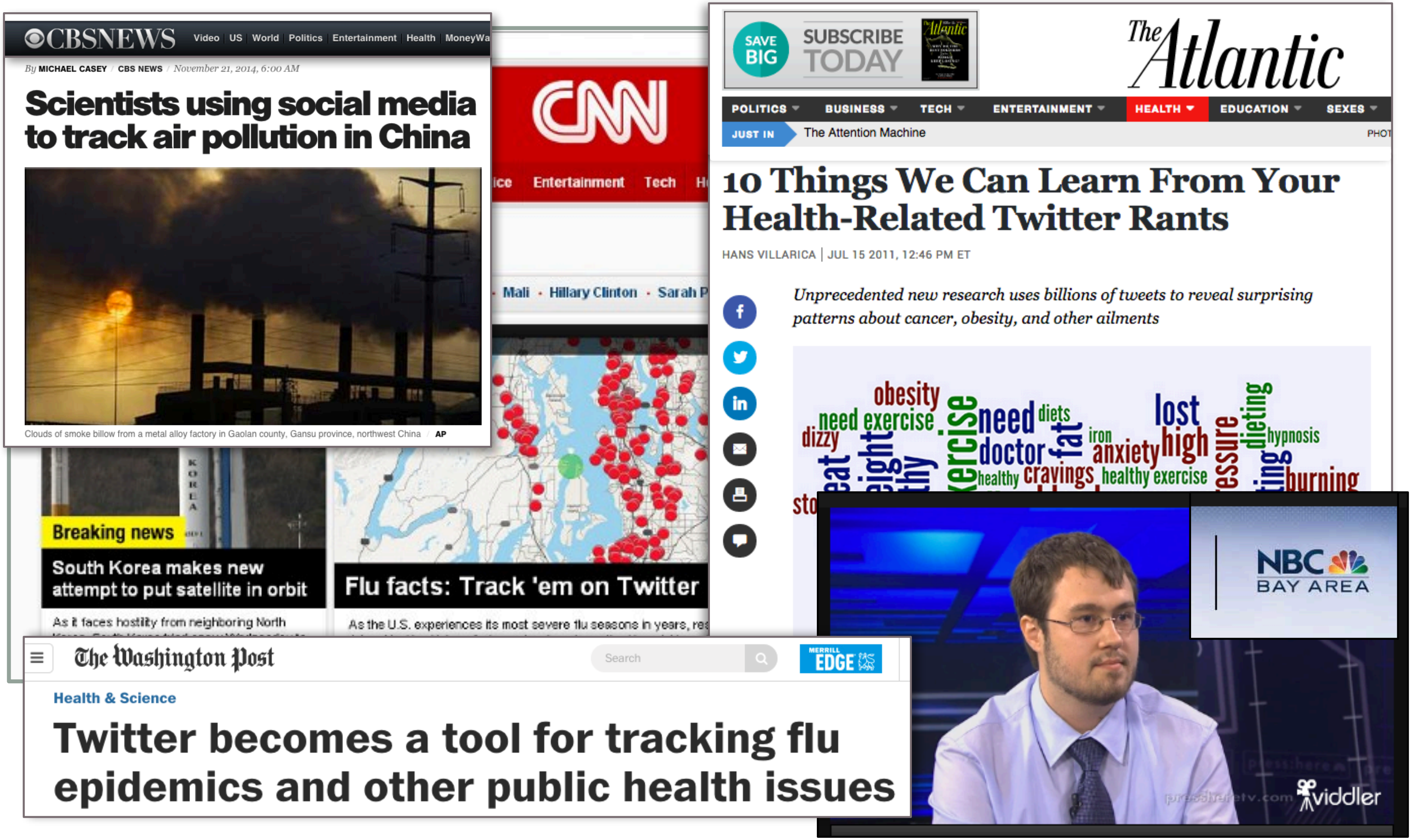
- Health decision-making in search logs

Paul, White, Horvitz (2015) **Web search as medical decision support for cancer.** *WWW*.

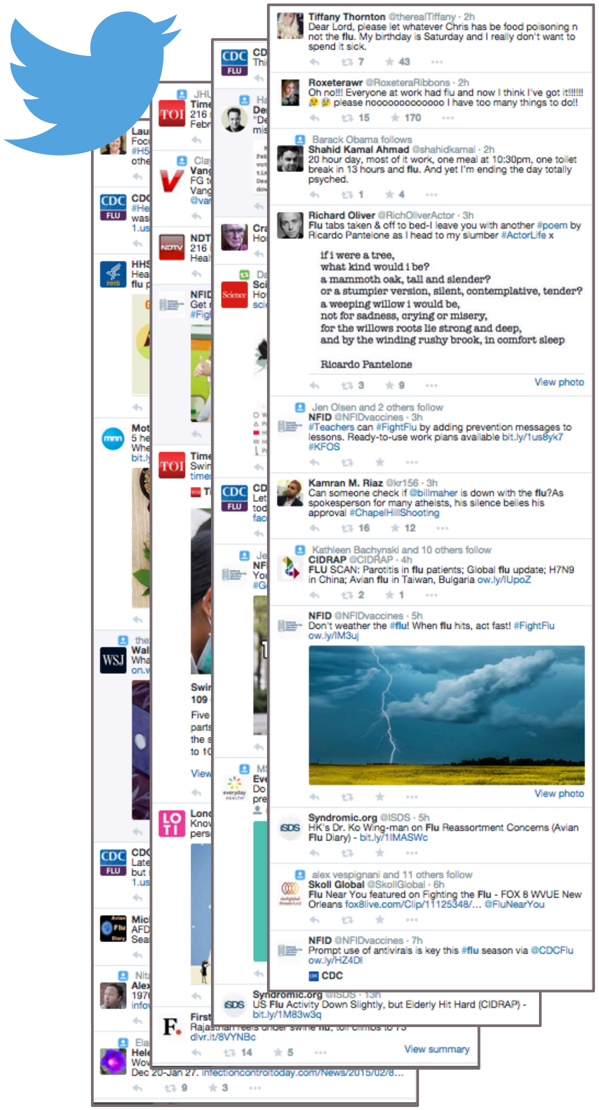
- Public opinion in Twitter on public health issues:
  - Gun control
  - Vaccination
  - Smoking

Benton, Paul, Hancock , Dredze (under review) **A joint model of topic and perspective in social media.**

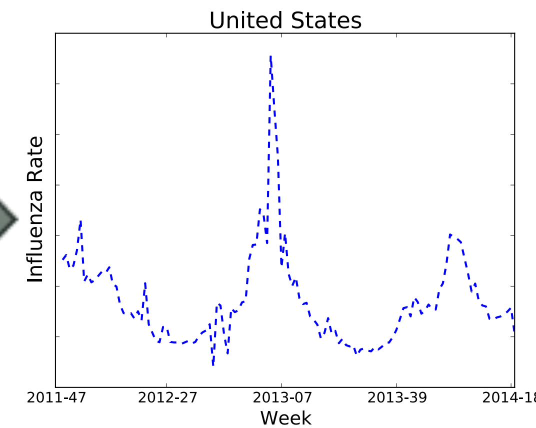
TEXT-DRIVEN PUBLIC HEALTH



TEXT AS DATA


Tiffany Thornton @therealTiffany · 2h
Dear Lord, please let whatever Chris has be food poisoning n not the flu. My birthday is Saturday and I really don't want to spend it sick.
Roxeterawr @RoxeteraRibbons · 2h
Oh no!!! Everyone at work had flu and now I think I've got it!!!!!! please noooooooooooooo I have too many things to do!!
Barack Obama follows
Shahid Kamal Ahmad @shahidkamal · 2h
20 hour day, most of it work, one meal at 10:30pm, one toilet break in 13 hours and flu. And yet I'm ending the day totally psyched.
Richard Oliver @RichOliverActor · 3h
Flu tabs taken & off to bed-I leave you with another #poem by Ricardo Pantelone as I head to my slumber #ActorLife x
if i were a tree,
what kind would i be?
a mammoth oak, tall and slender?
or a stumpier version, silent, contemplative, tender?
a weeping willow i would be,
not for sadness, crying or misery,
for the willows roots lie strong and deep,
and by the winding rushy brook, in comfort sleep
Ricardo Pantelone
View photo
Jen Olsen and 2 others follow
NFID @NFIDvaccines · 3h
#Teachers can #FightFlu by adding prevention messages to lessons. Ready-to-use work plans available bit.ly/1us8yk7 #KFOS
Kamran M. Riaz @kr156 · 3h
Can someone check if @billmaher is down with the flu?As spokesperson for many atheists, his silence belies his approval #ChapelHillShooting
Kathleen Bachynski and 10 others follow
CIDRAP @CIDRAP · 4h
FLU SCAN: Parotitis in flu patients; Global flu update; H7N9 in China; Avian flu in Taiwan, Bulgaria ow.ly/IUpoZ
NFID @NFIDvaccines · 5h
Don't weather the #flu! When flu hits, act fast! #FightFlu ow.ly/IM3uj
View photo
Syndromic.org @ISDS · 5h
HK's Dr. Ko Wing-man on Flu Reassortment Concerns (Avian Flu Diary) - bit.ly/1MASWc
alex vespignani and 11 others follow
Skoll Global @SkollGlobal · 6h
Flu Near You featured on Fighting the Flu - FOX 8 WVUE New Orleans fox8live.com/Clip/11125348/... @FluNearYou
NFID @NFIDvaccines · 7h
Prompt use of antivirals is key this #flu season via @CDCFlu ow.ly/HZ4Dl
CDC
Syndromic.org @ISDS · 13h
US Flu Activity Down Slightly, but Elderly Hit Hard (CIDRAP) - bit.ly/1M83w3q
View summary

...

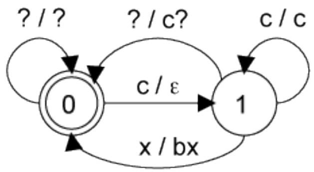

United States
Influenza Rate
2011-47
2012-27
2013-07
2013-39
2014-18
Week

TEXT AS DATA

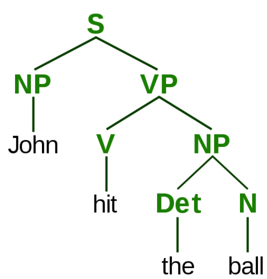
NATURAL LANGUAGE PROCESSING

Structure of language:

- Morphology/strings



- Syntax/grammar



- Discourse/speech acts
- Topics/concepts

Paul, Eisner (2012) **Implicitly intersecting weighted automata using dual decomposition.** *NAACL.* CS

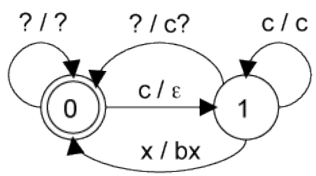
Darling, Paul, Song (2012) **Unsupervised part-of-speech tagging in noisy domains with a syntactic-semantic Bayesian HMM.** *EACL Workshop on Social Media.* CS

Paul (2012) **Mixed membership Markov models for unsupervised conversation modeling.** *EMNLP.* CS

TEXT AS DATA | NATURAL LANGUAGE PROCESSING

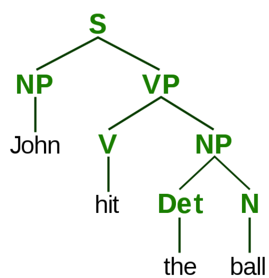
Structure of language:

- Morphology/strings



Paul, Eisner (2012) **Implicitly intersecting weighted automata using dual decomposition.** *NAACL.* CS

- Syntax/grammar



Darling, Paul, Song (2012) **Unsupervised part-of-speech tagging in noisy domains with a syntactic-semantic Bayesian HMM.** *EACL Workshop on Social Media.* CS

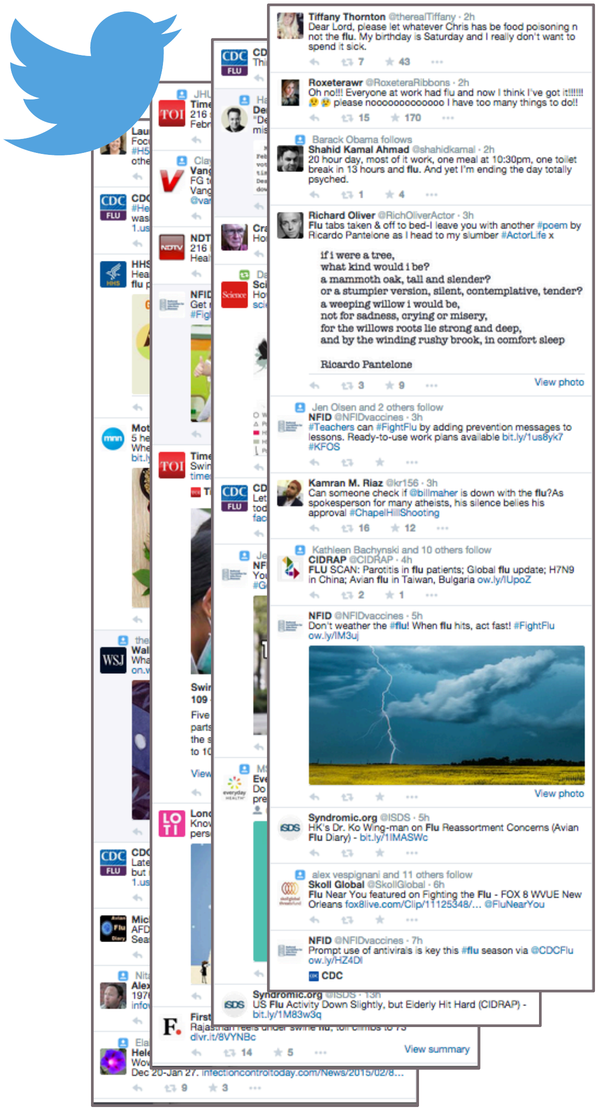
- Discourse/speech acts

Paul (2012) **Mixed membership Markov models for unsupervised conversation modeling.** *EMNLP.* CS

- Topics/concepts

TEXT AS DATA

NATURAL LANGUAGE PROCESSING



sick
sore
throat
feel
fever
flu

hurts
knee
ankle
hurt
neck
ouch

...

allergies
nose
eyes
allergy
allergic
sneezing

ow
teeth
tooth
wisdom
toothache
dentist

...

cancer
pray
surgery
hospital
pain
breast

body
pounds
gym
weight
lost
workout

...

Paul, Dredze (2011) **You are what you tweet: Analyzing Twitter for public health.** *5th International Conference on Weblogs and Social Media (ICWSM).* CS

Paul, Dredze (2014) **Discovering health topics in social media using topic models.** *PLOS ONE* 9(8).

TOPIC MODELING

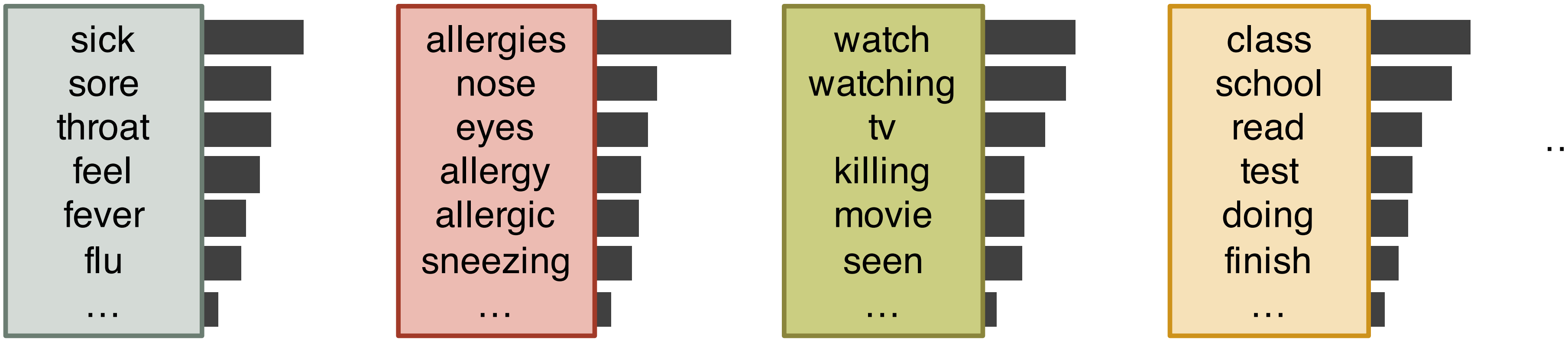
A topic model is a **statistical model** of text

- We pretend that our data (text) are the output of a probabilistic process that generates data

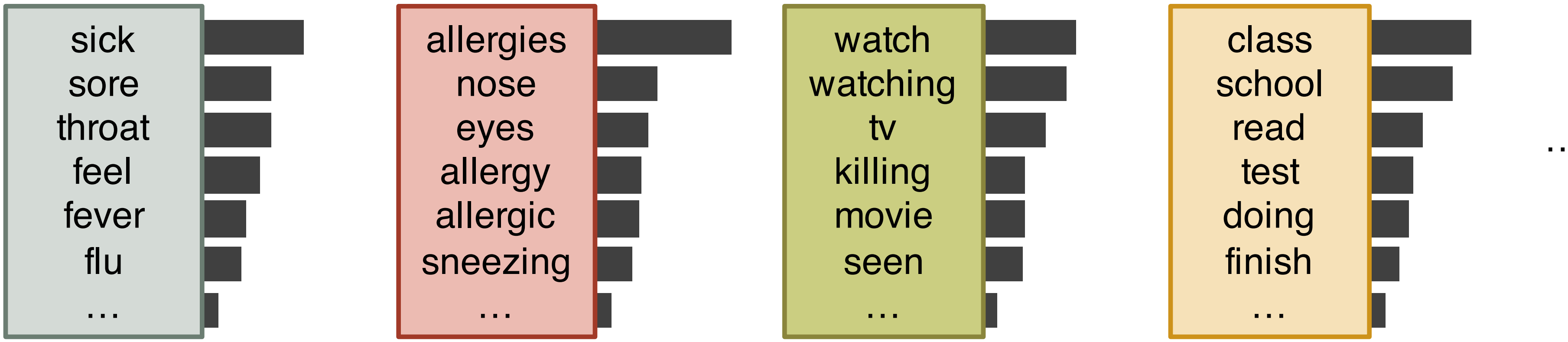
TOPIC MODELING

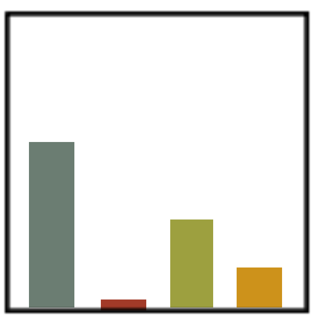
sick	allergies	watch	class	…
sore	nose	watching	school	
throat	eyes	tv	read	
feel	allergy	killing	test	
fever	allergic	movie	doing	
flu	sneezing	seen	finish	
…	…	…	…	

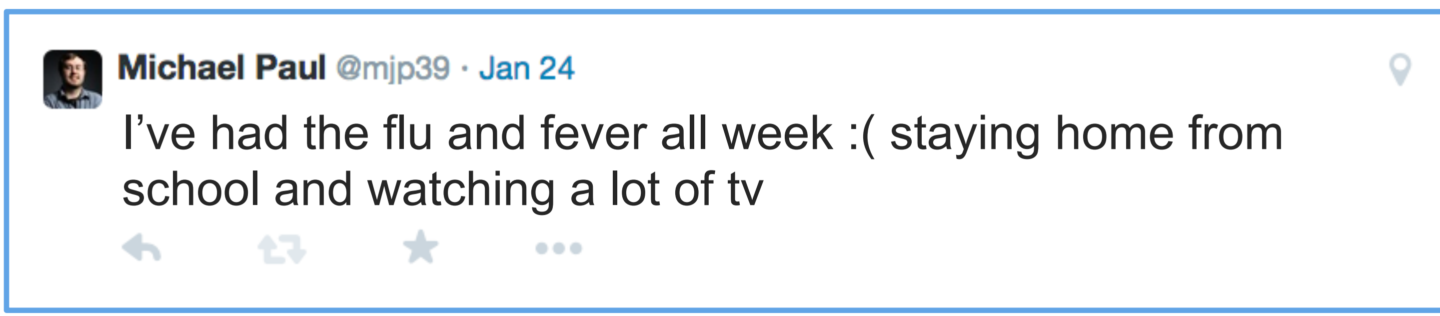
TOPIC MODELING



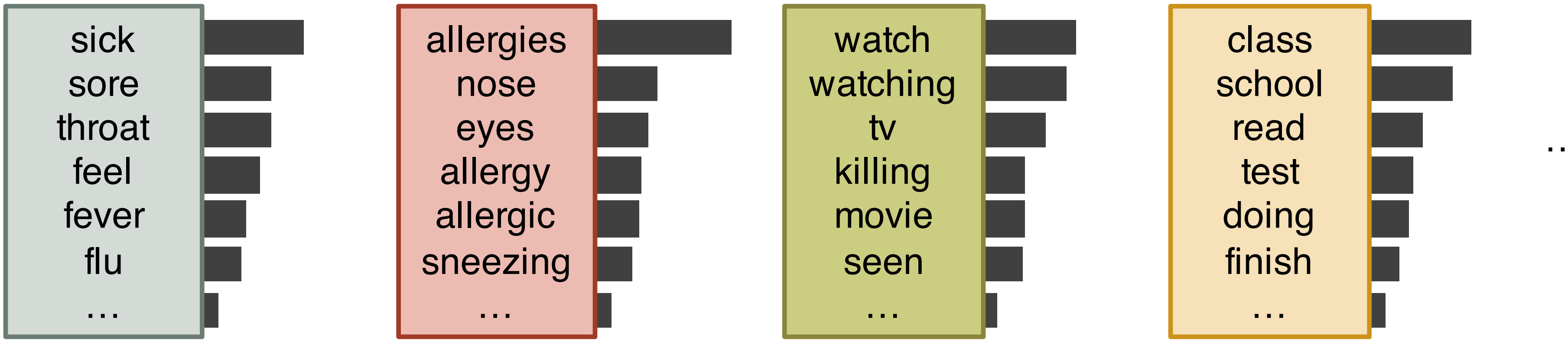
TOPIC MODELING

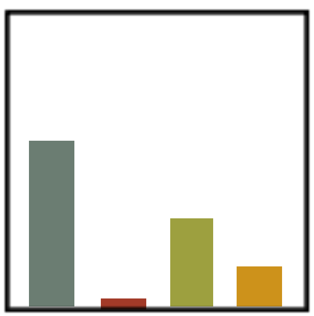


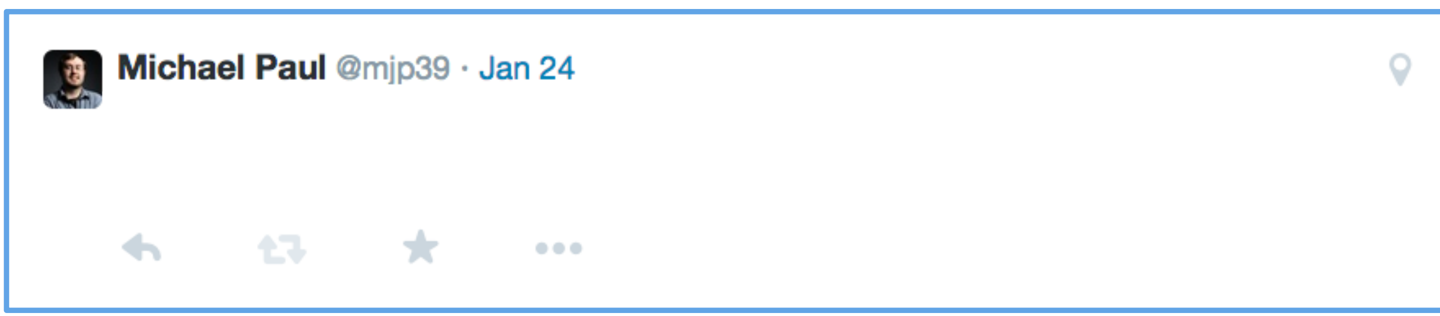




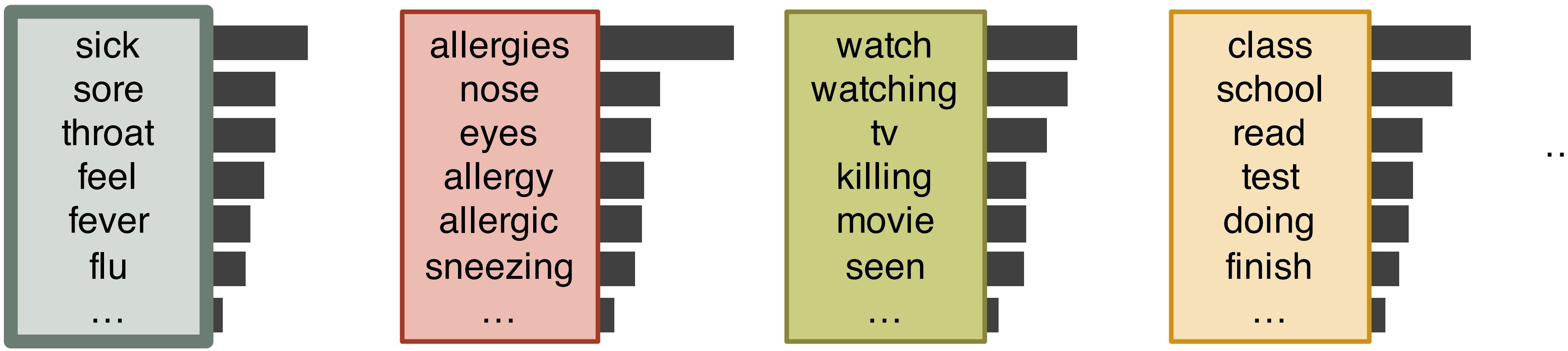
TOPIC MODELING

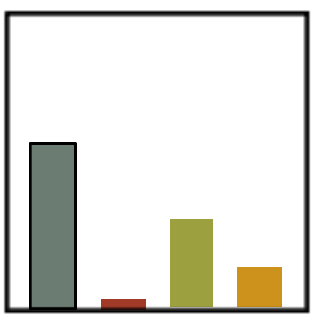


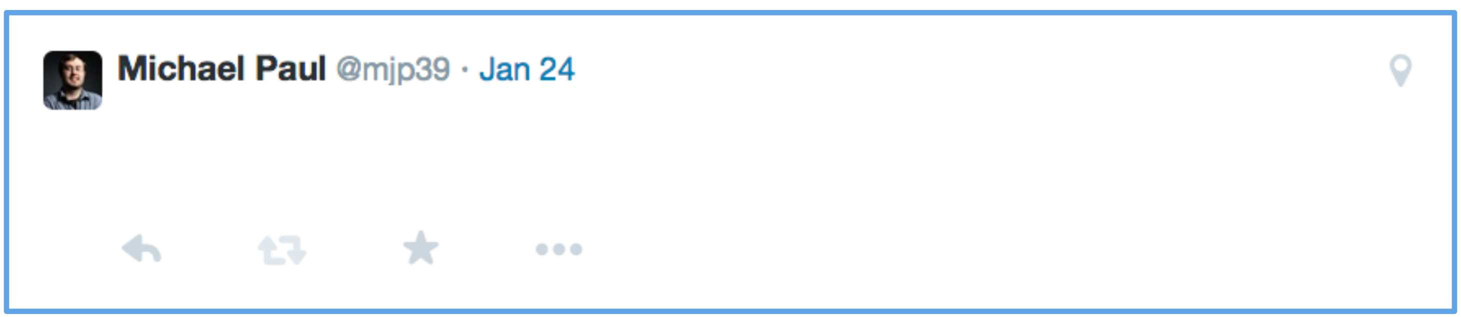




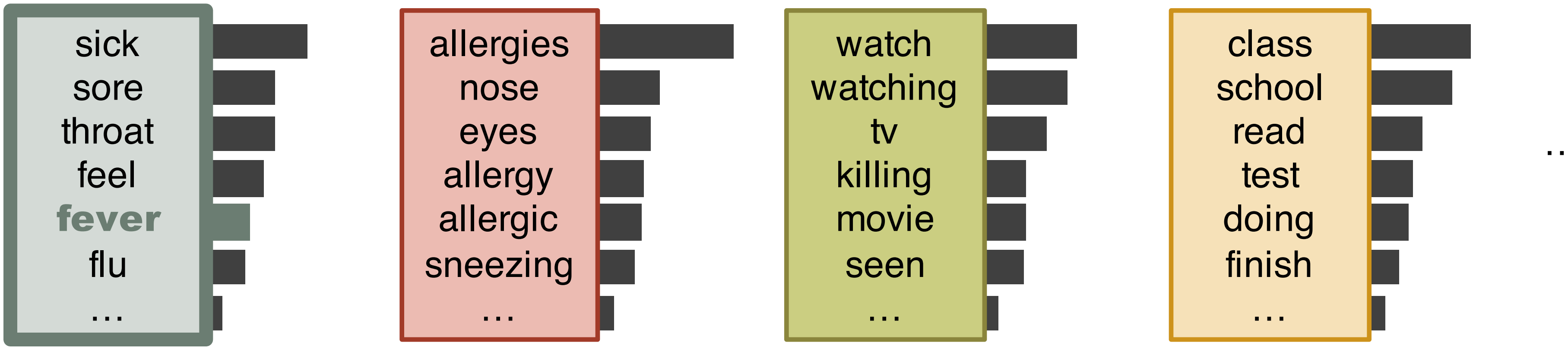
TOPIC MODELING

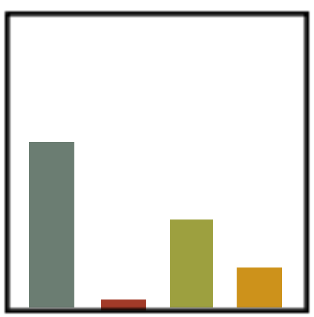


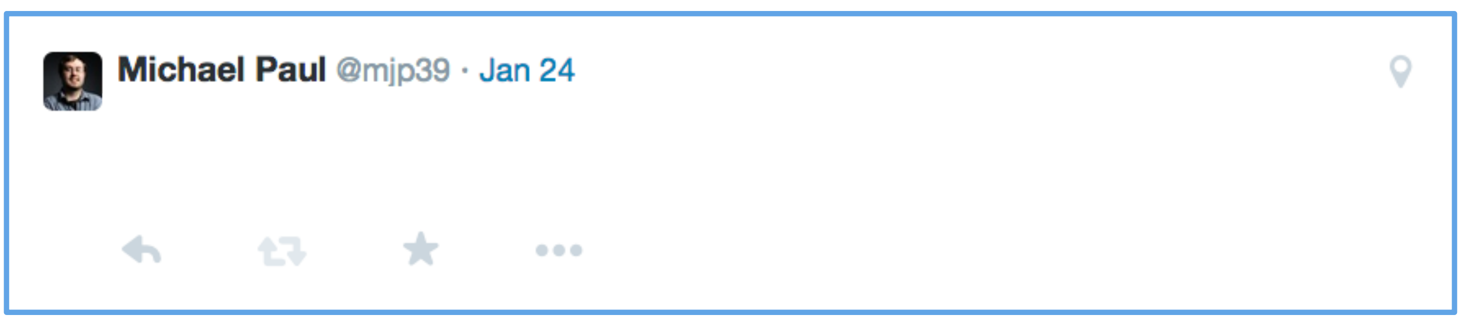




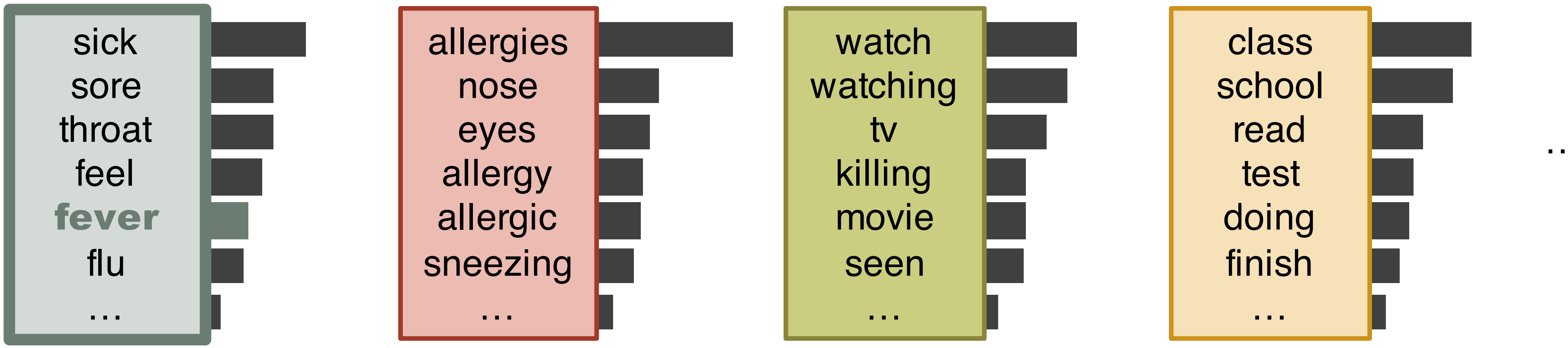
TOPIC MODELING

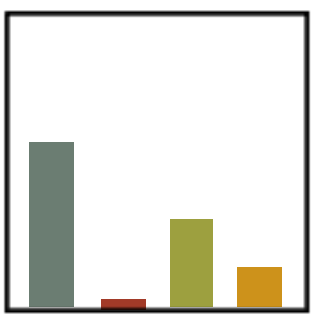


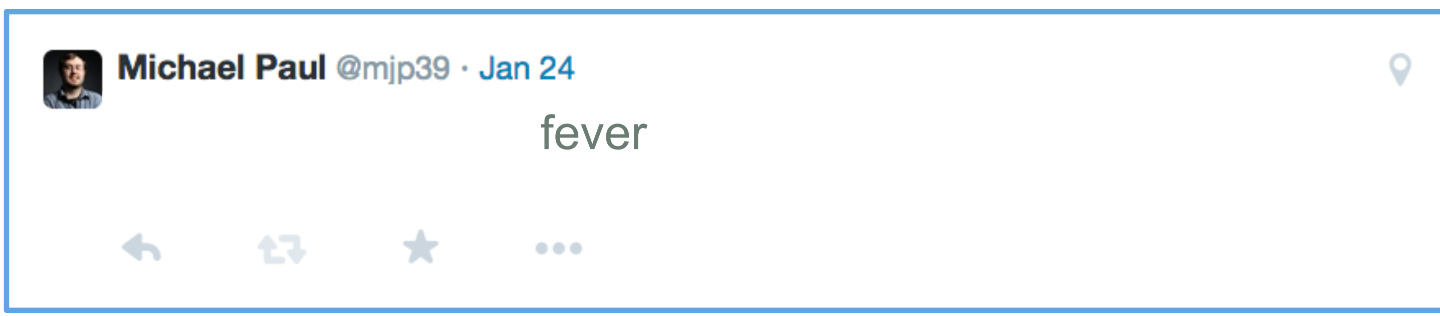




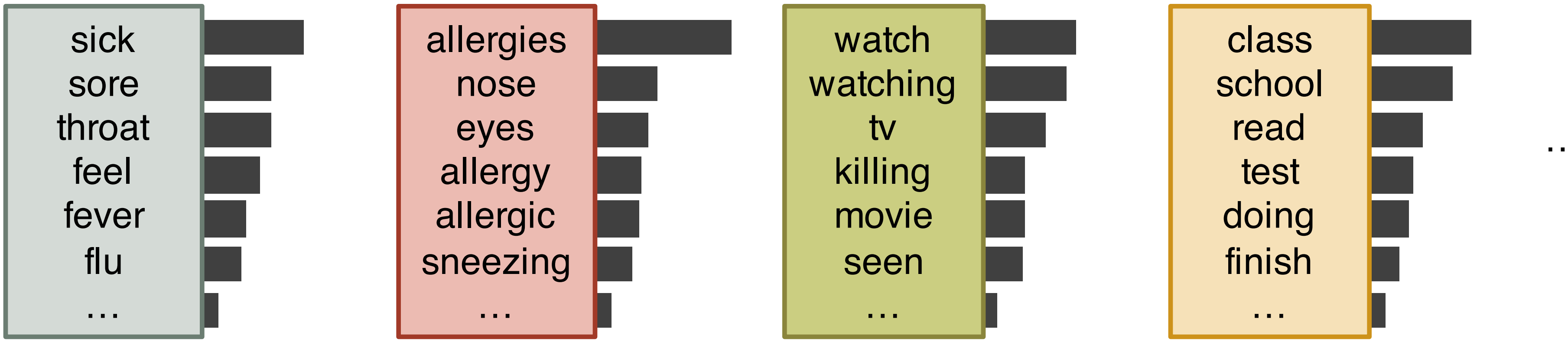
TOPIC MODELING

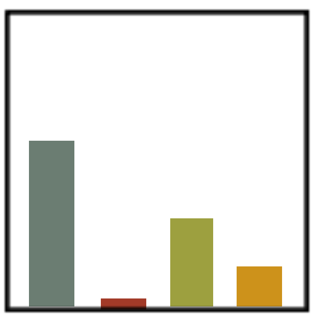


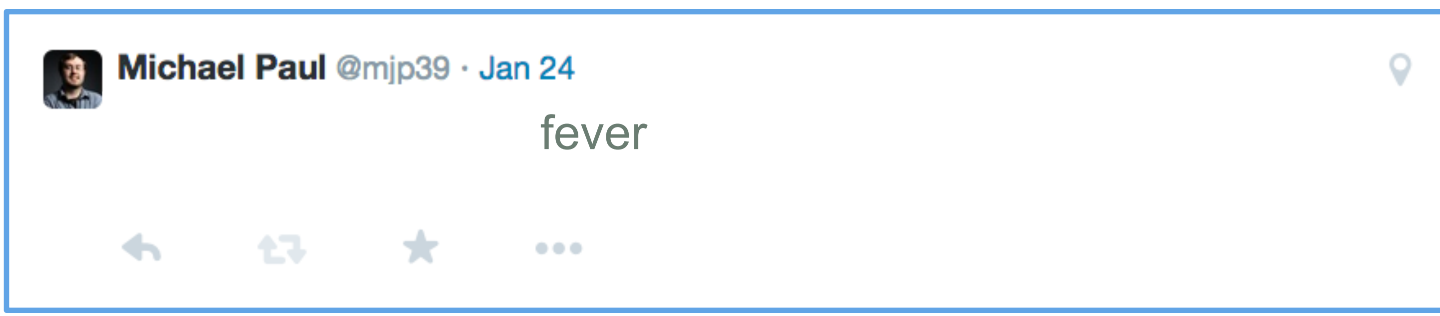




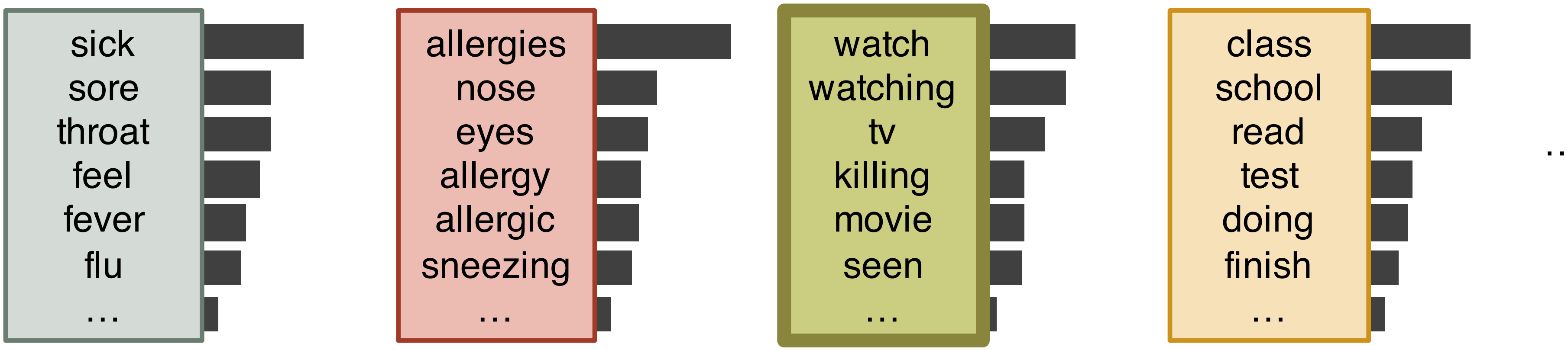
TOPIC MODELING

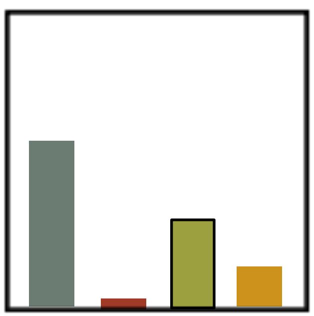


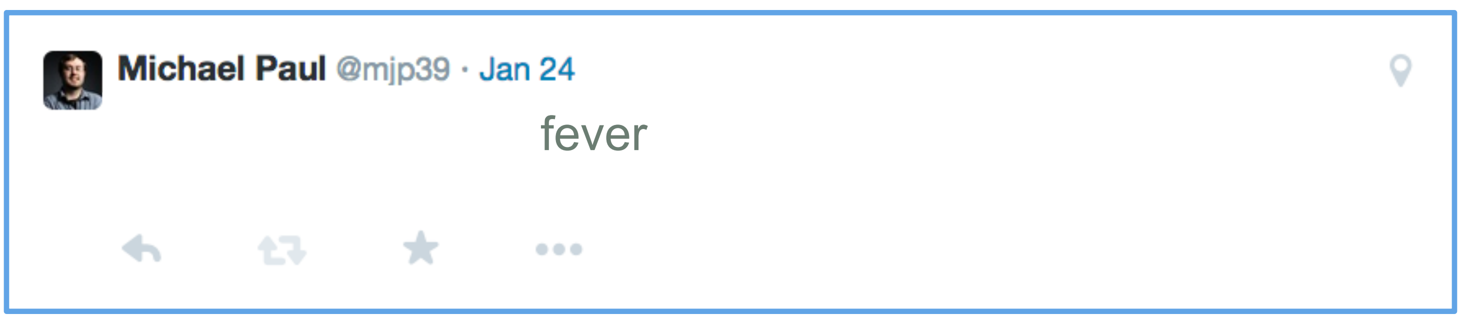




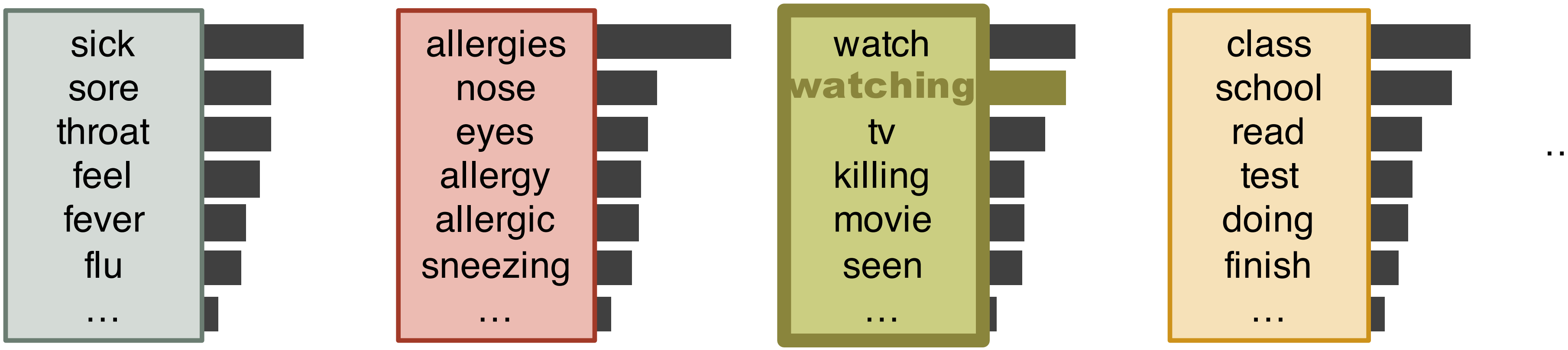
TOPIC MODELING

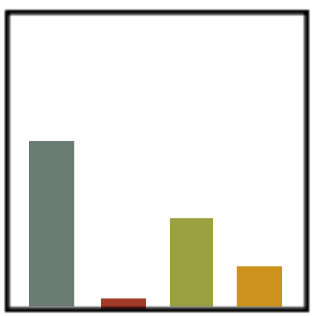


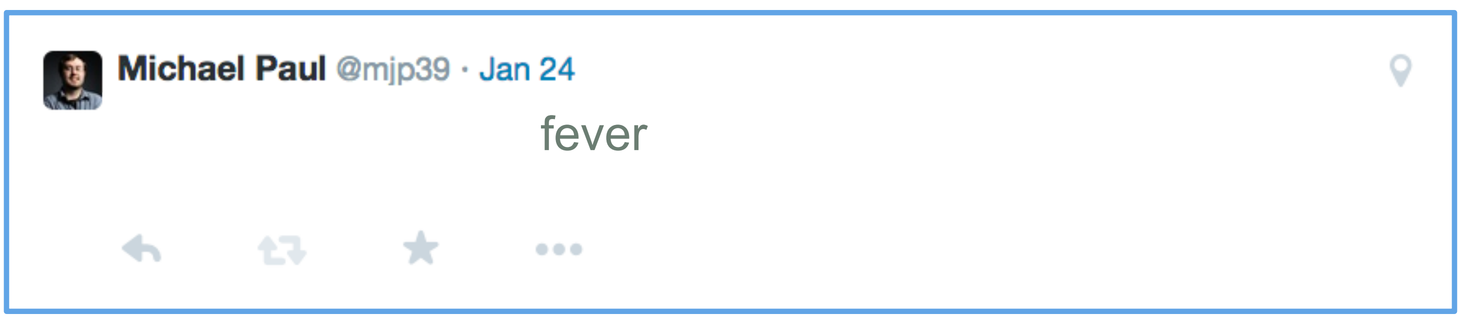




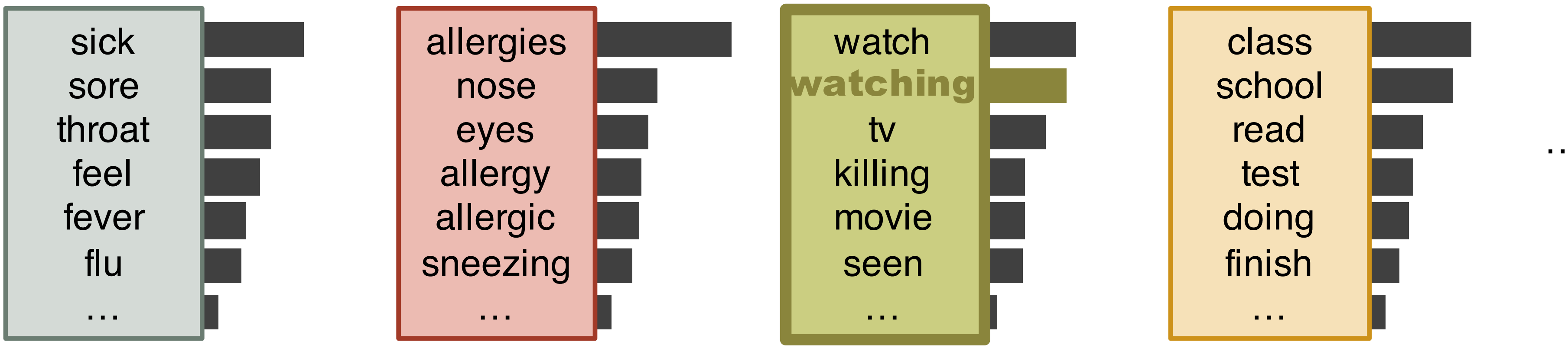
TOPIC MODELING

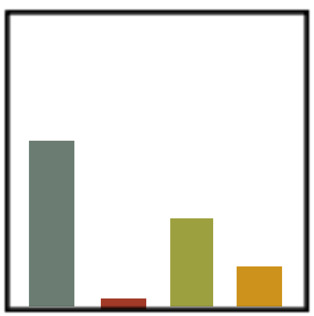


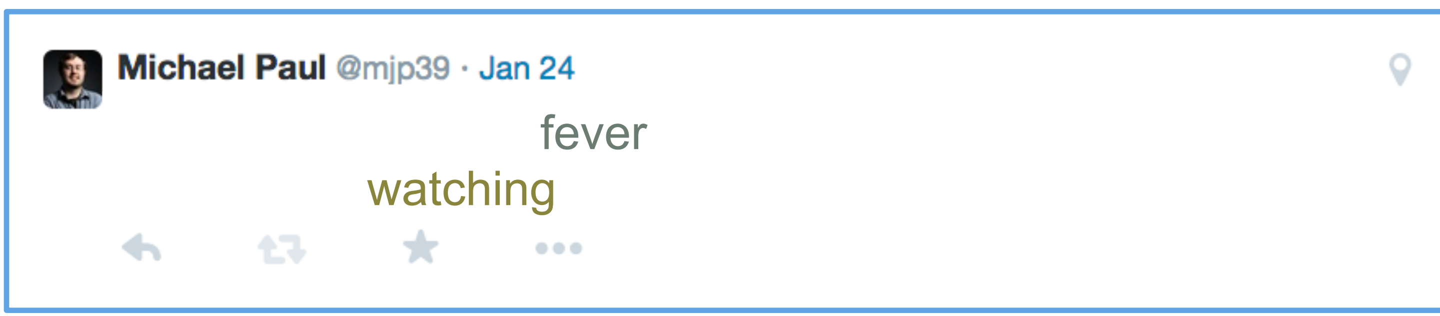




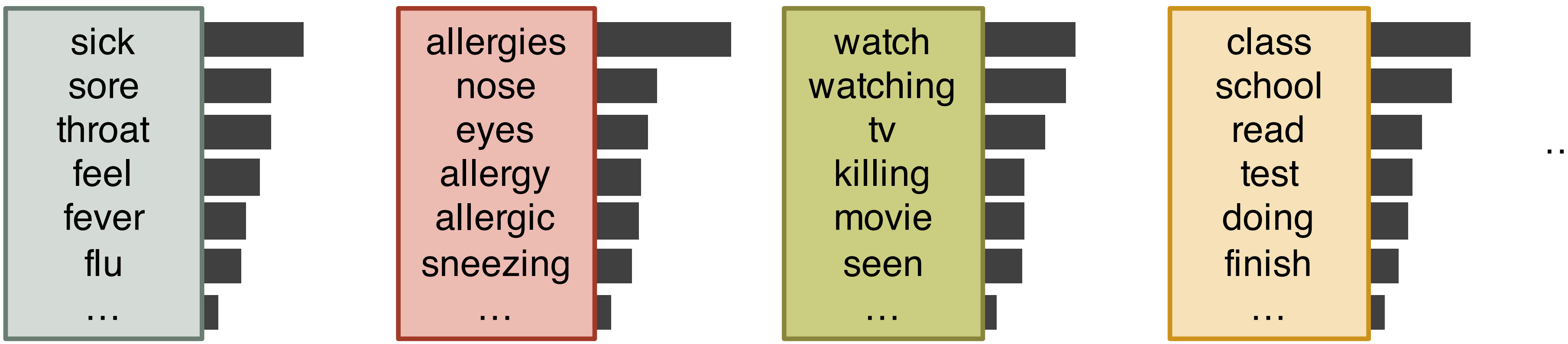
TOPIC MODELING

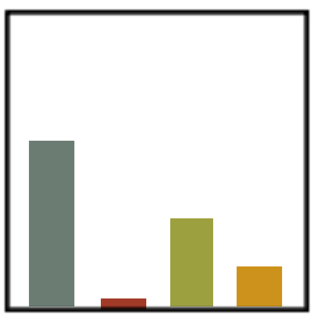


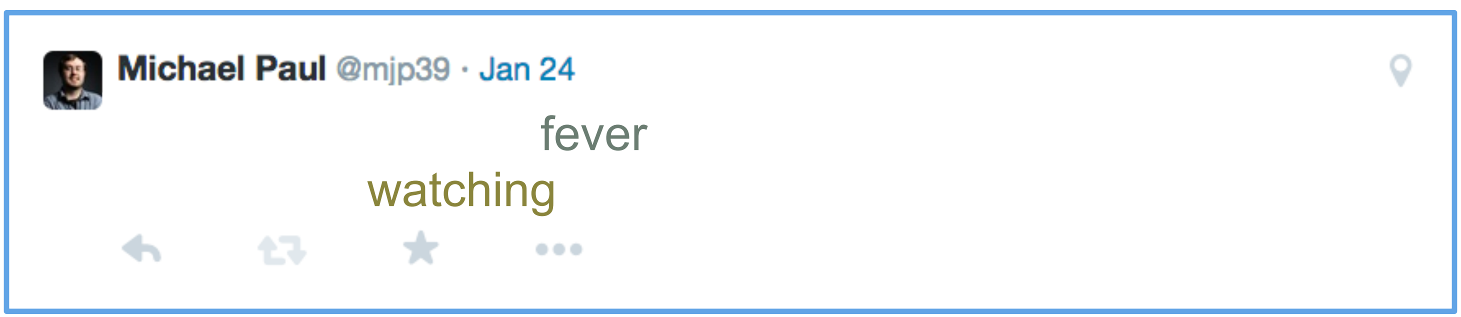




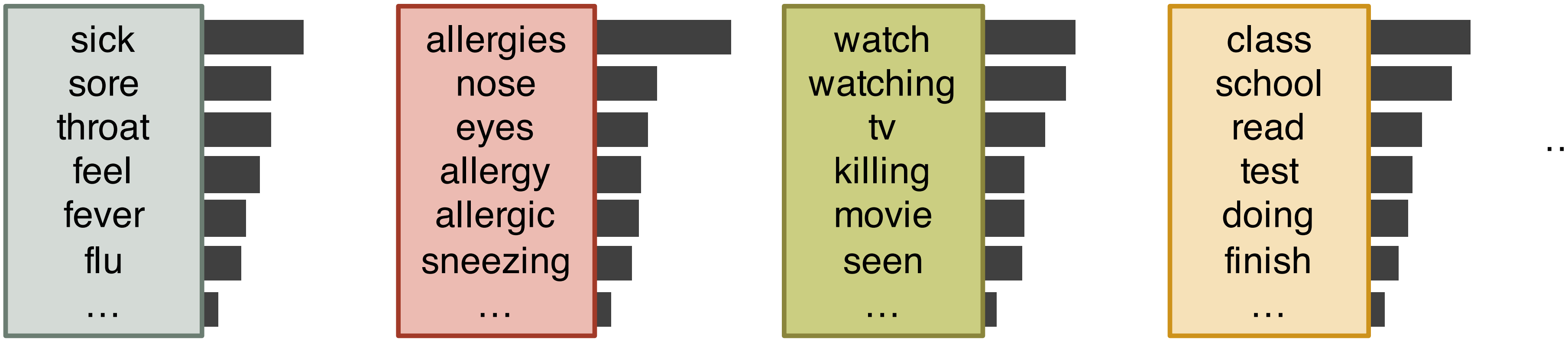
TOPIC MODELING

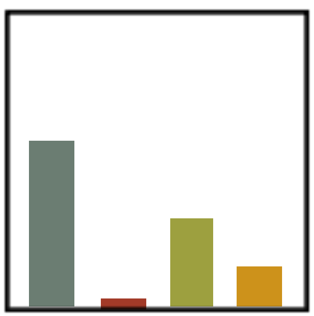


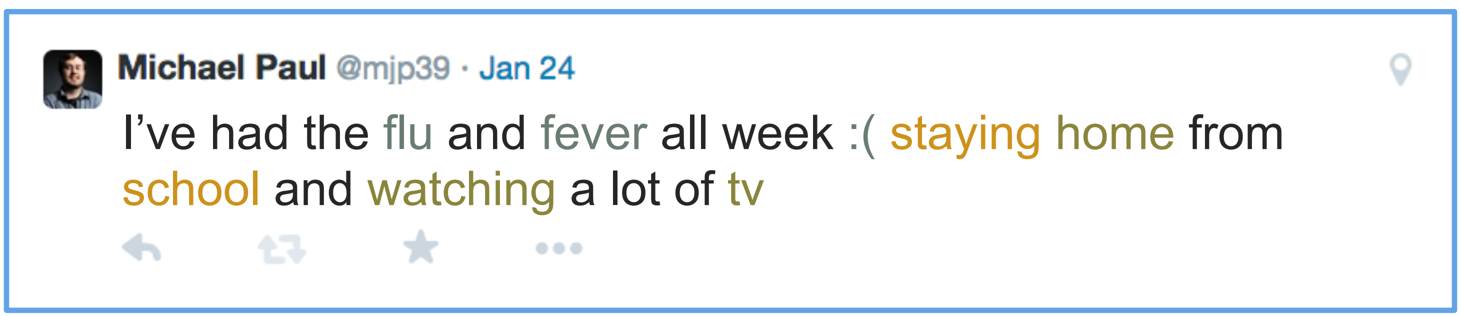




TOPIC MODELING

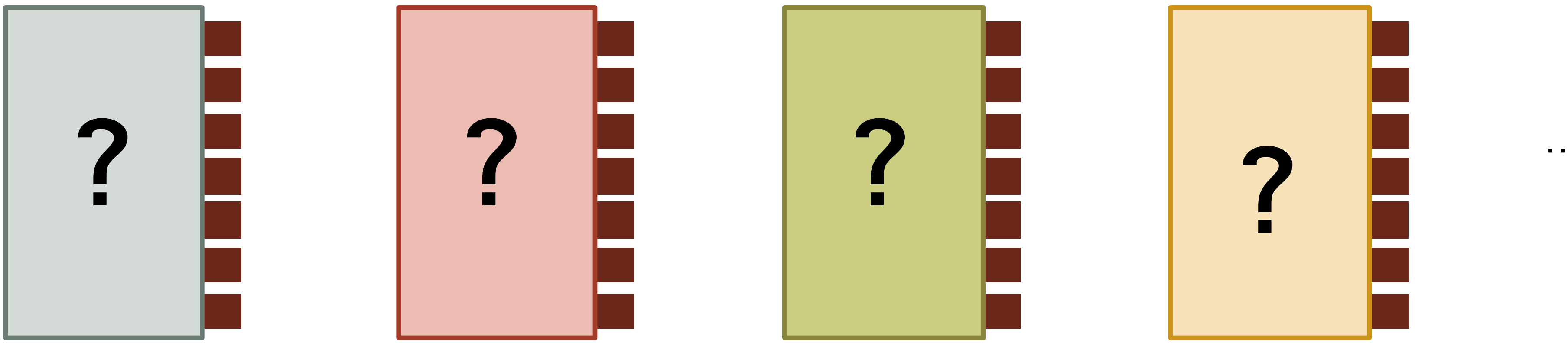


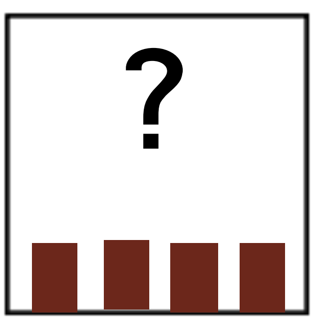


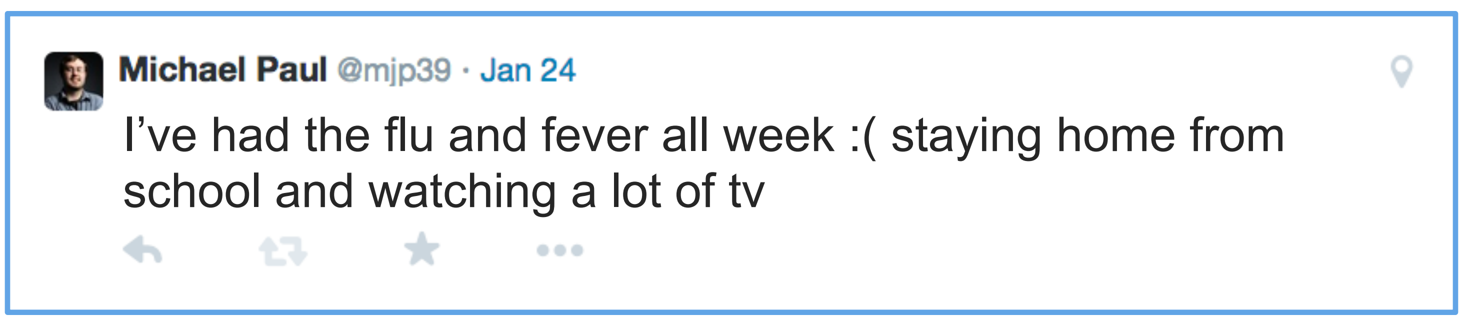


TOPIC MODELING

PARAMETER ESTIMATION

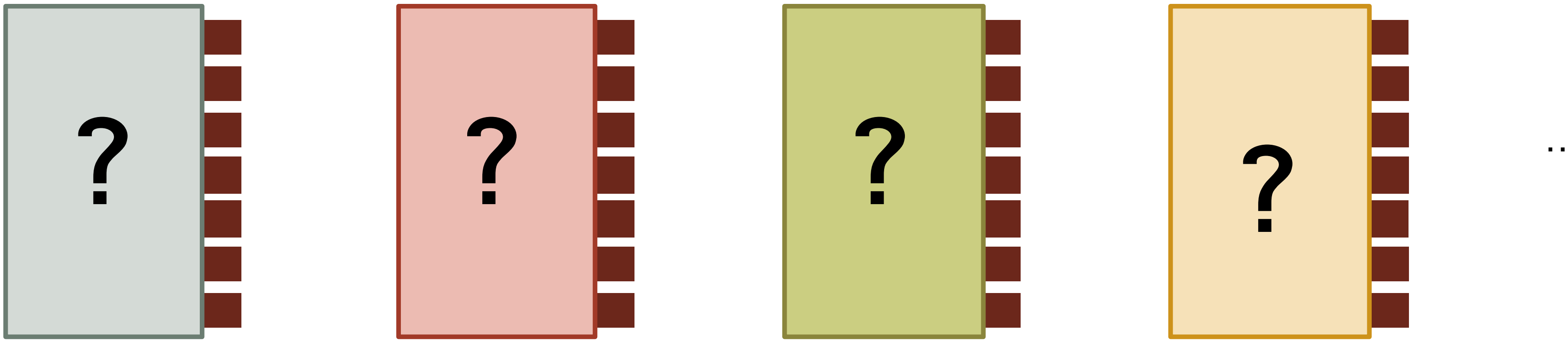


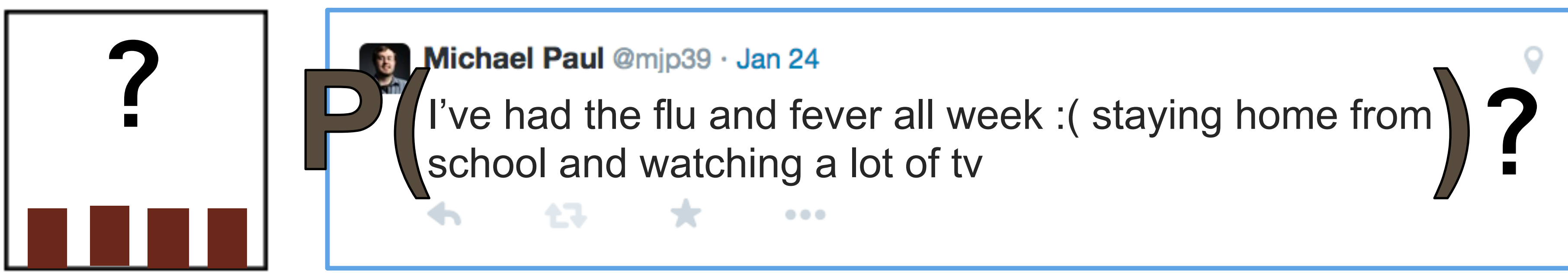




TOPIC MODELING

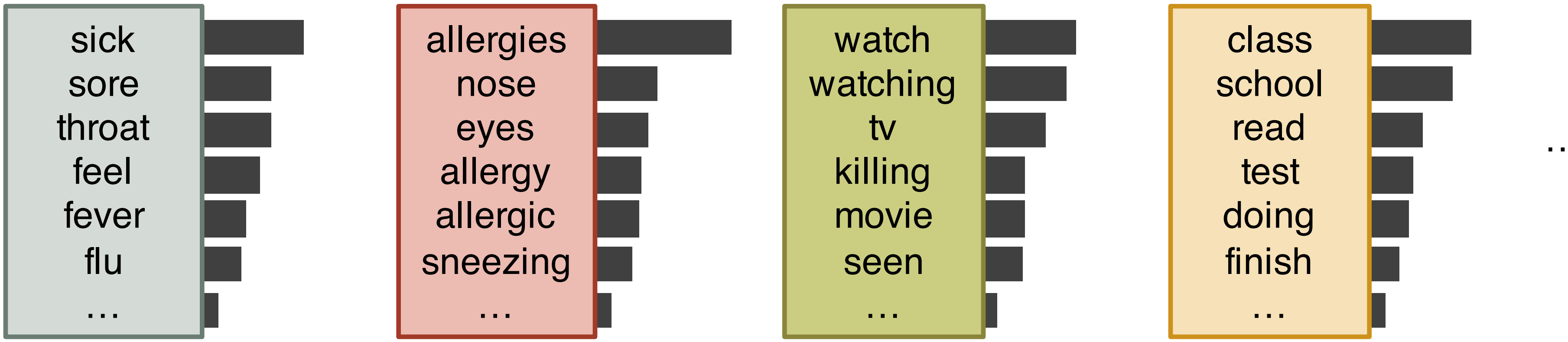
PARAMETER ESTIMATION

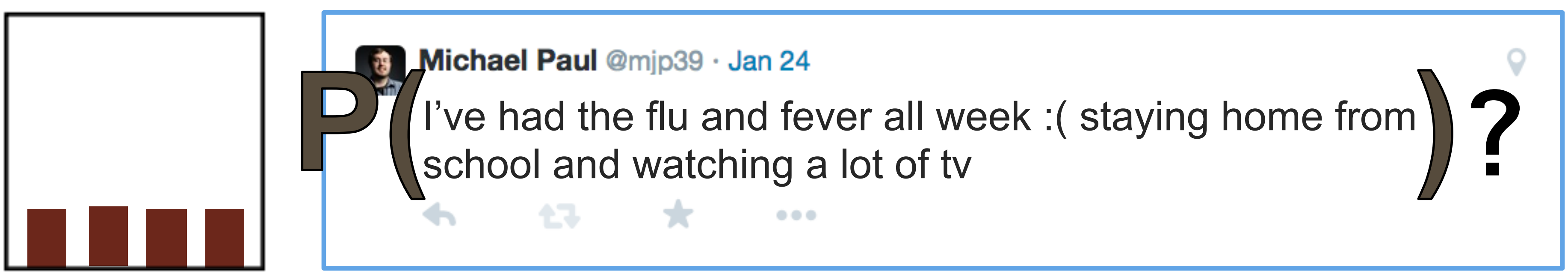




TOPIC MODELING

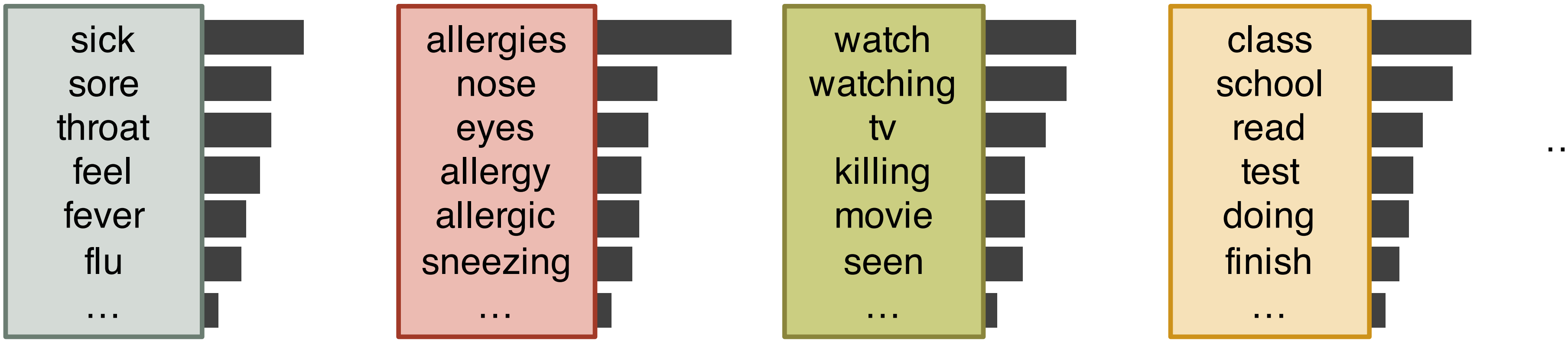
PARAMETER ESTIMATION

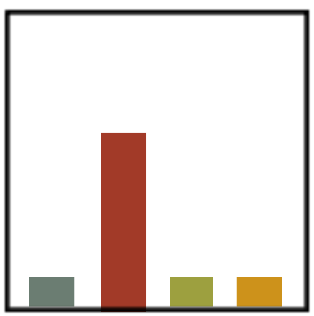


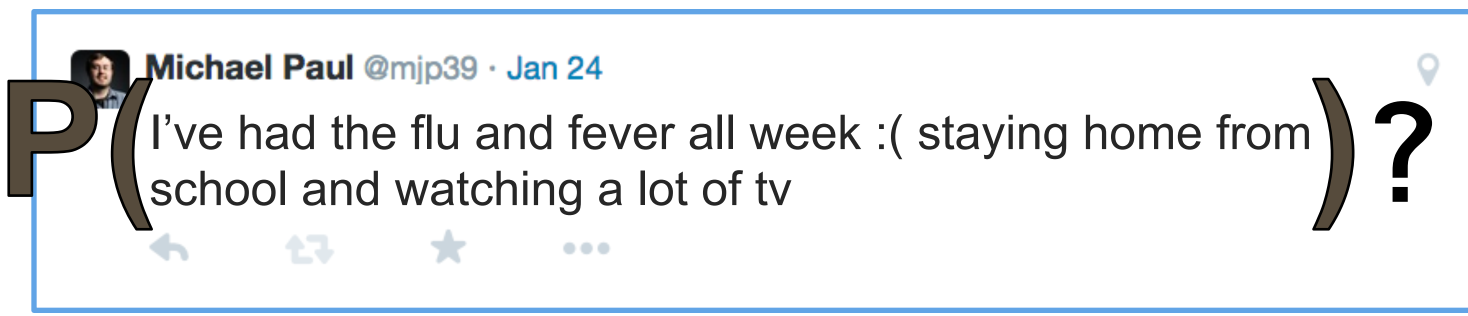


TOPIC MODELING

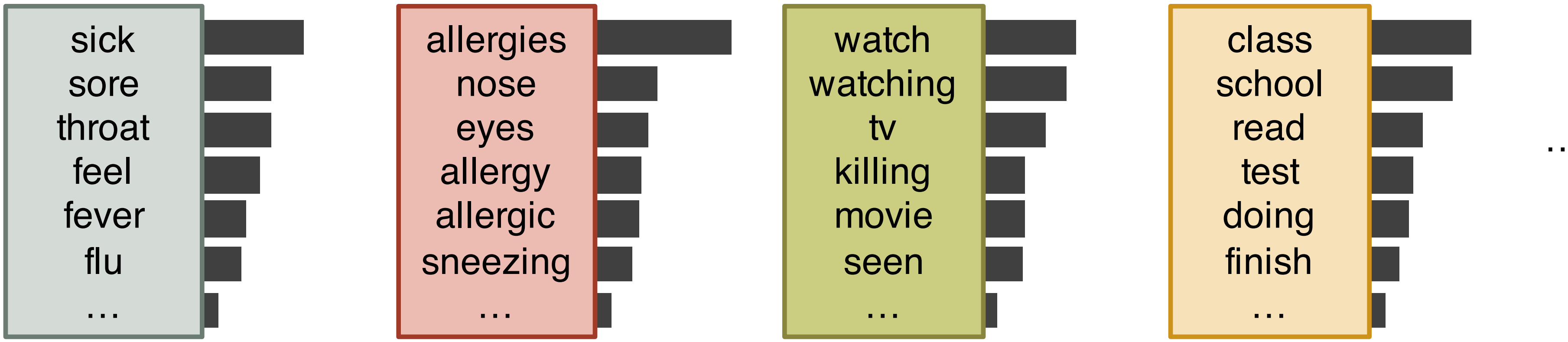
PARAMETER ESTIMATION

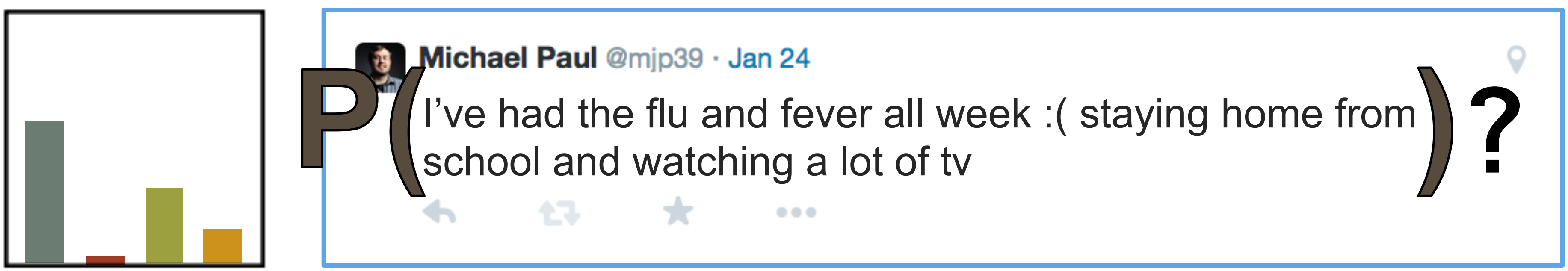




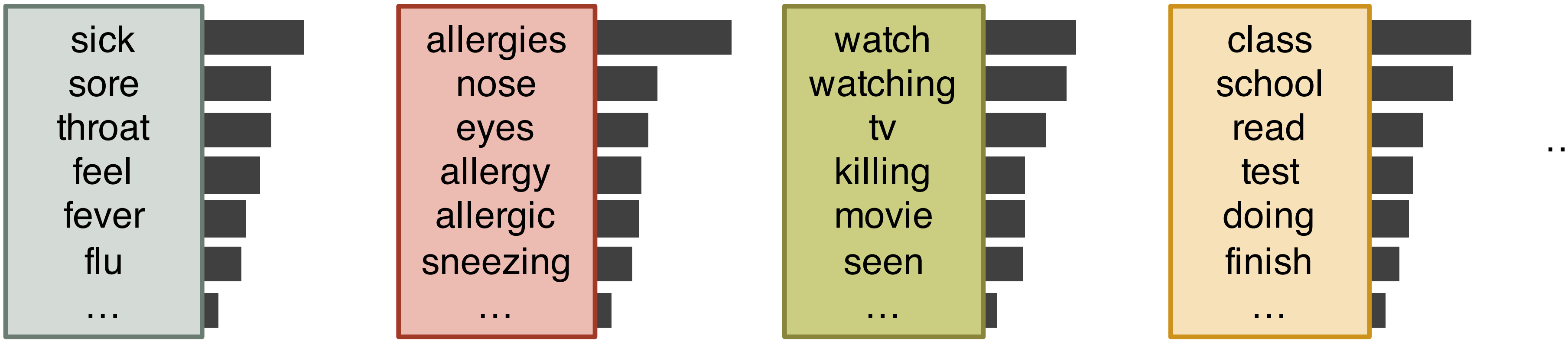


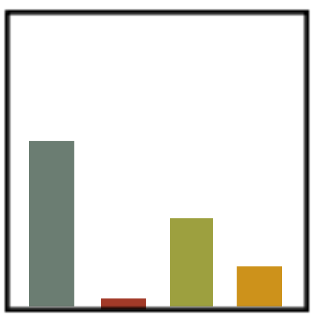
TOPIC MODELING | PARAMETER ESTIMATION





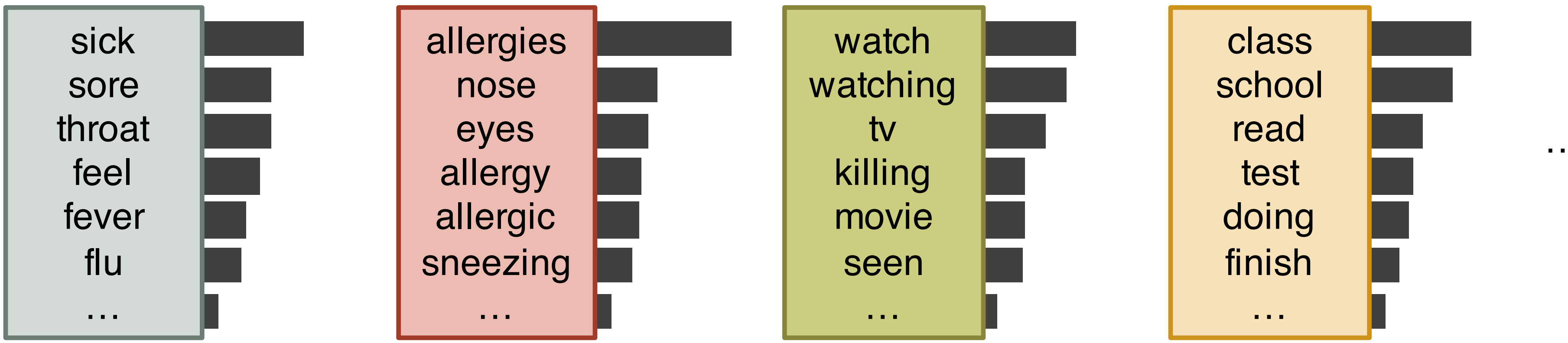
TOPIC MODELING PRIORS

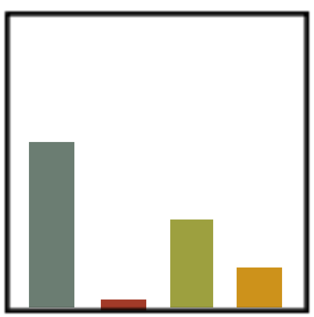




Our imaginary process also needs to generate all these distributions

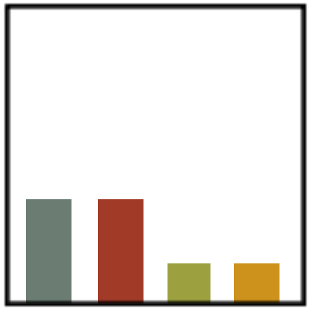
TOPIC MODELING PRIORS





Our imaginary process also needs to generate all these distributions

- We need a distribution over distributions
  - Called a **prior** distribution

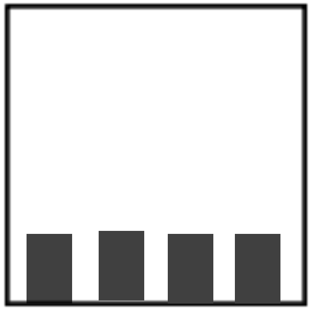
Dirichlet(ρ ✖  **)**

Precision

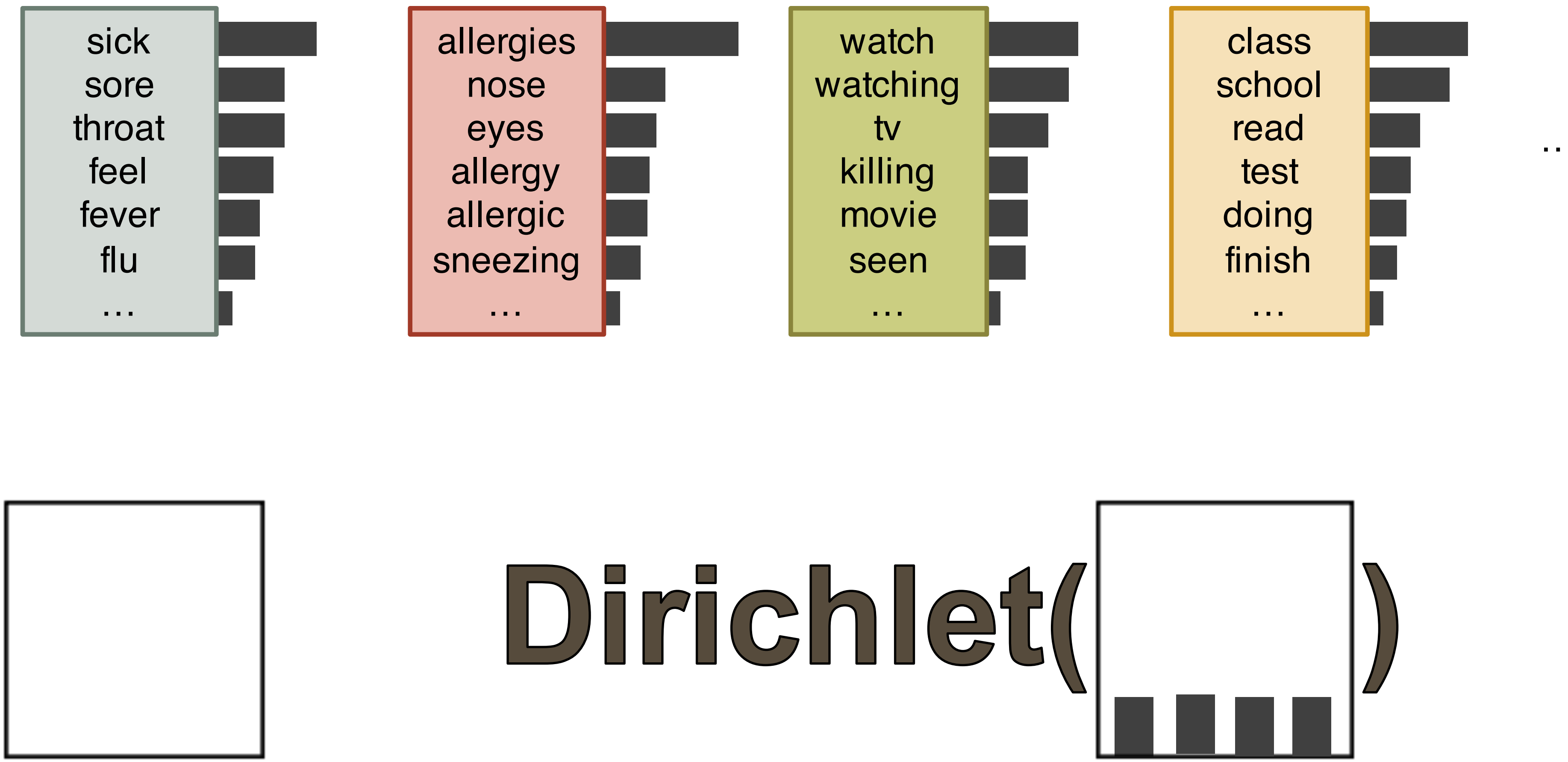
Mean

TOPIC MODELING | PRIORS

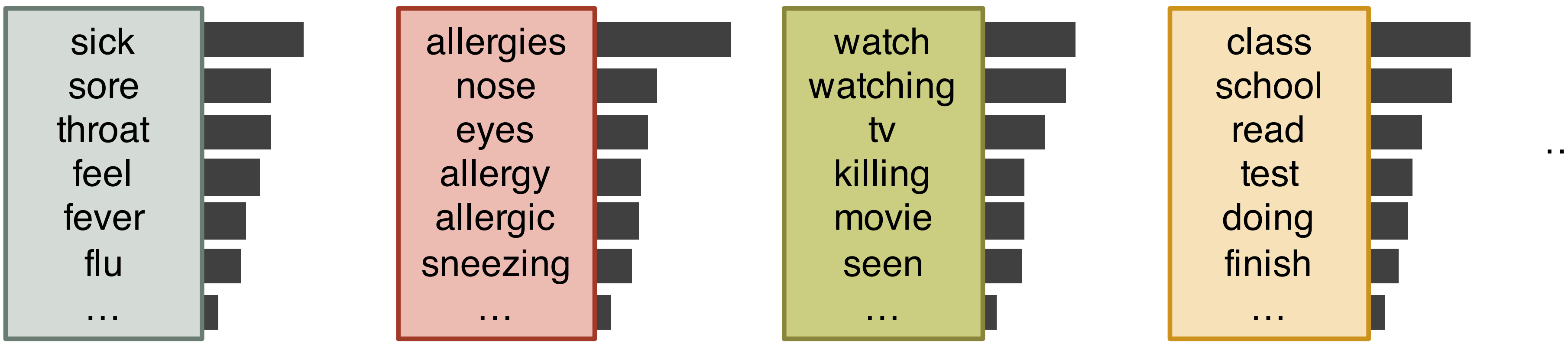
…

Dirichlet( **)**

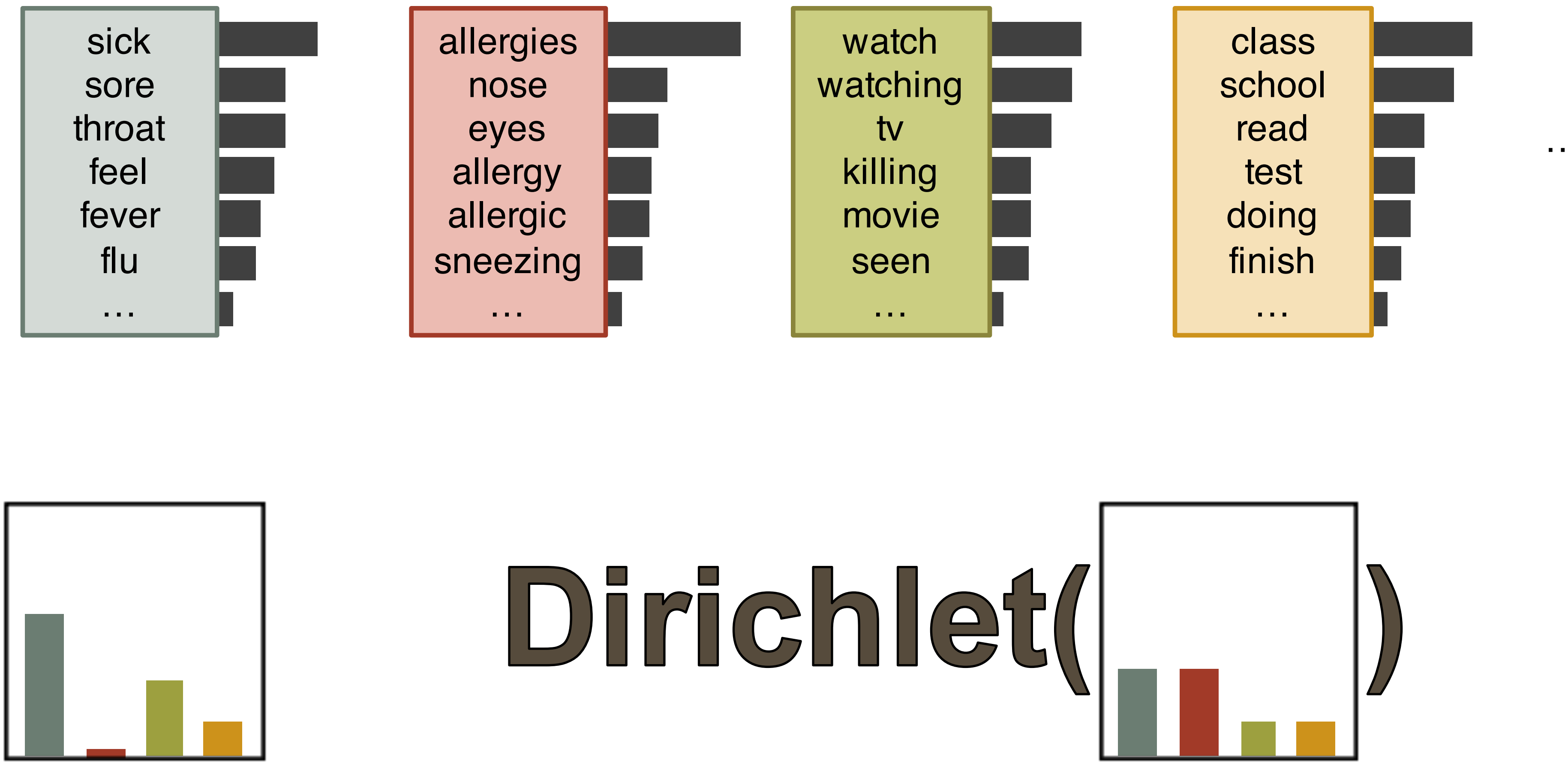
TOPIC MODELING PRIORS



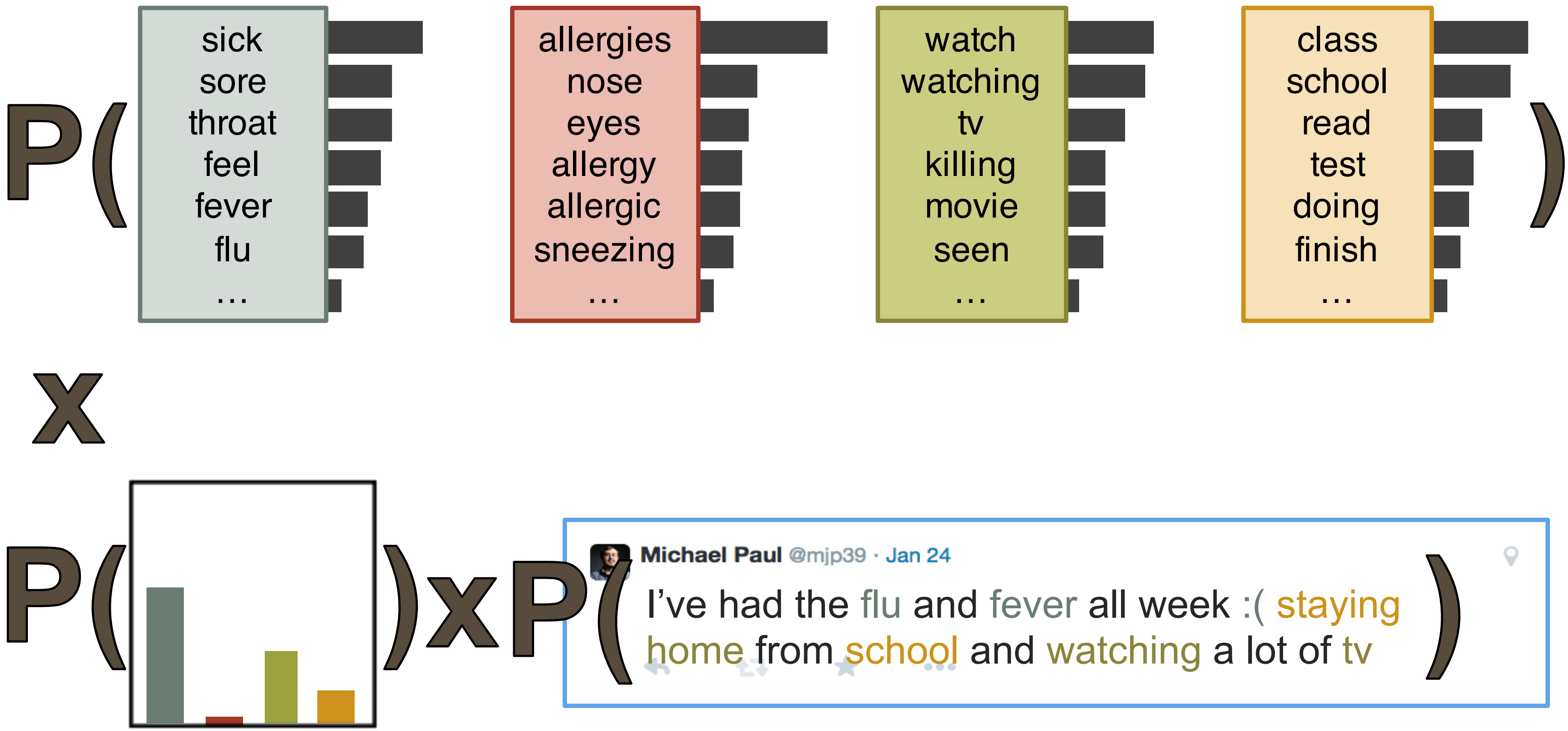
TOPIC MODELING | PRIORS



TOPIC MODELING PRIORS



TOPIC MODELING PRIORS



Latent Dirichlet Allocation (LDA)
Blei, Ng, Jordan 2003

The topic and word distributions have Dirichlet priors

PRIORS

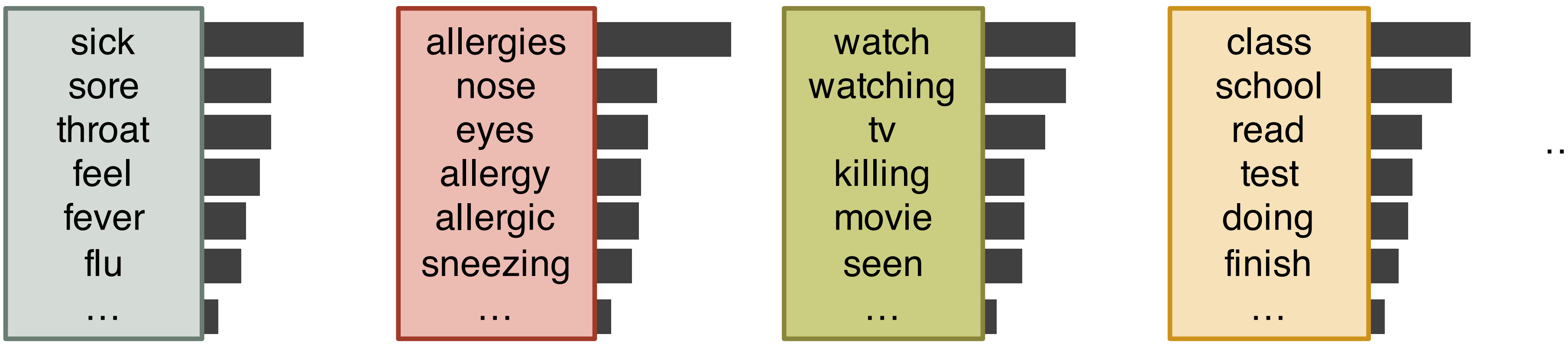
Latent Dirichlet Allocation (LDA)
Blei, Ng, Jordan 2003

The topic and word distributions have Dirichlet priors

Standard topic models are often insufficient for particular applications

- We need richer structure

TOPIC MODELING APPLICATIONS



Paul, Dredze (2011) **You are what you tweet: Analyzing Twitter for public health.** *5th International Conference on Weblogs and Social Media (ICWSM).* CS

Paul, Dredze (2014) **Discovering health topics in social media using topic models.** *PLOS ONE* 9(8): e103408.

TOPIC MODELING APPLICATIONS

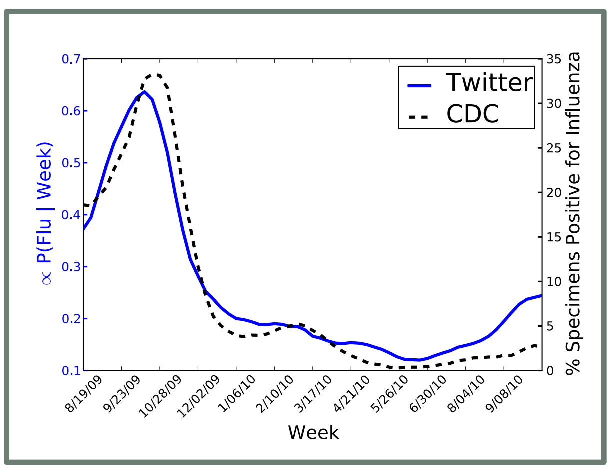
sick
sore
throat
feel
fever
flu
…

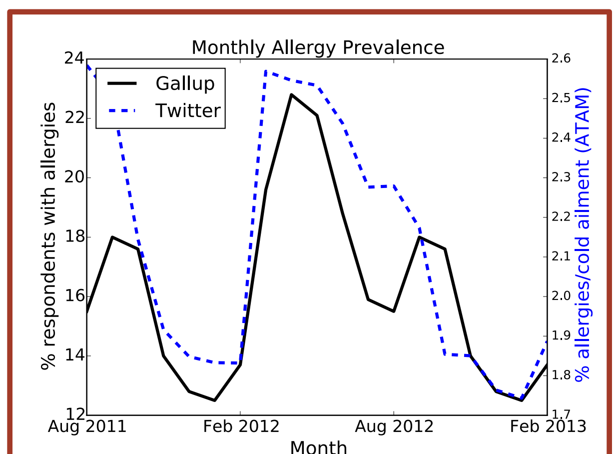
allergies
nose
eyes
allergy
allergic
sneezing
…

watch
watching
tv
killing
movie
seen
…

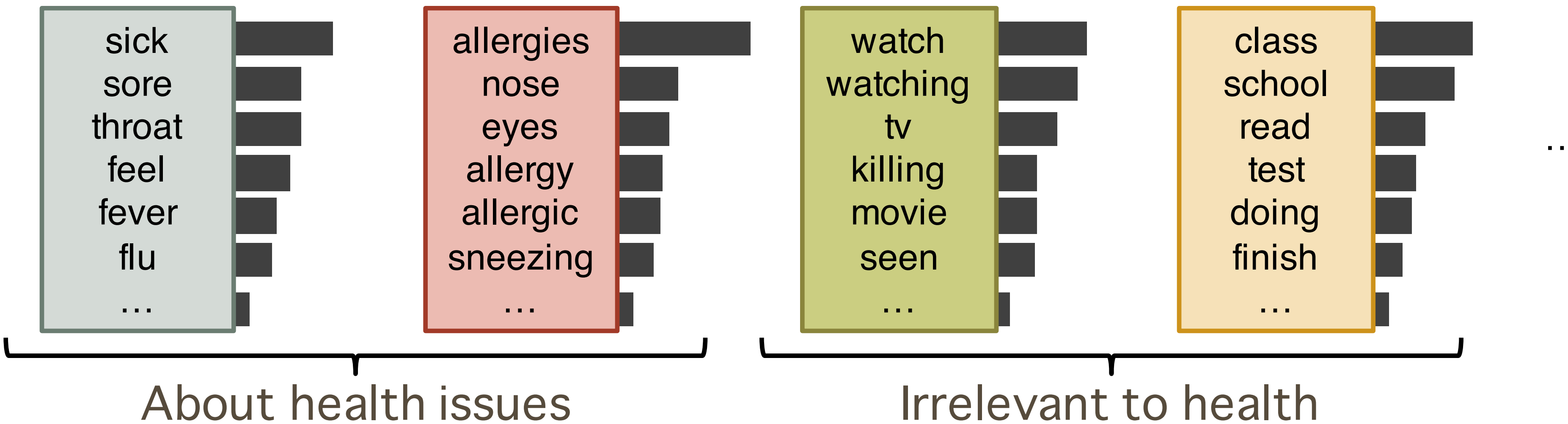
class
school
read
test
doing
finish
…

…

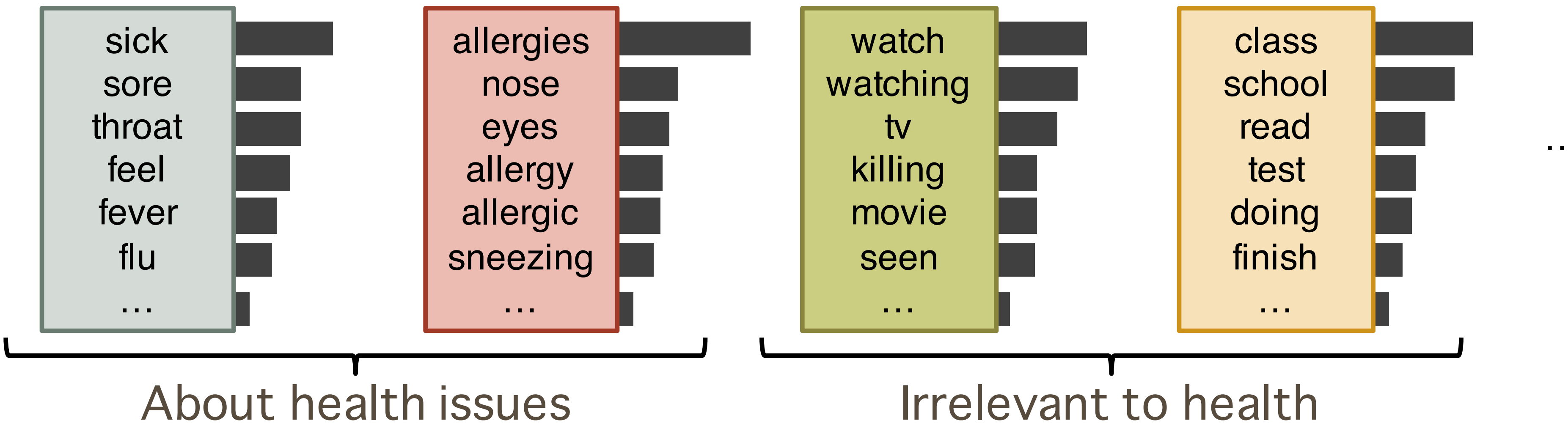




TOPIC MODELING ADDING STRUCTURE



TOPIC MODELING | ADDING STRUCTURE

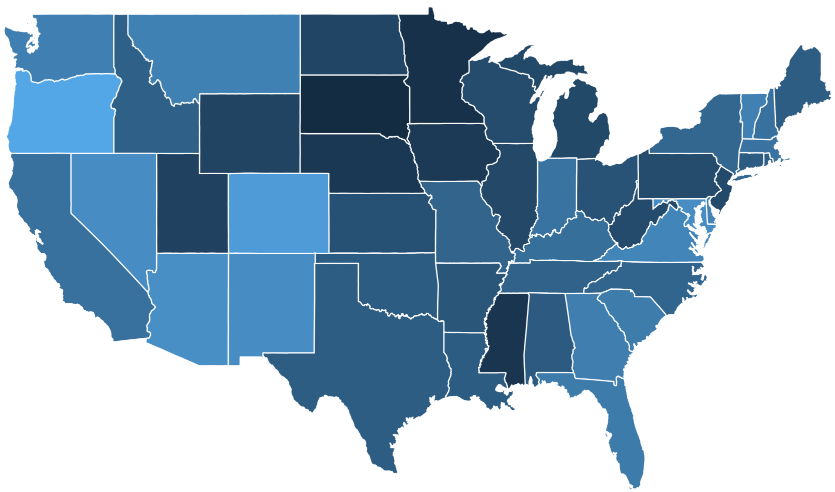


In general:

Topics can be organized in ways that are more interpretable to users

TOPIC MODELING ADDING STRUCTURE

Understanding healthcare quality from online reviews:



Topics in online doctor reviews:

best	time	office
years	staff	time
caring	great	appointment
care	helpful	rude
patients	feel	staff
patient	questions	room
recommend	office	didn’t
family	friendly	wait

TOPIC MODELING ADDING STRUCTURE

Topics in online doctor reviews:

RateMDs
DOCTORS YOU CAN TRUST

Both have positive sentiment

best	time	office
years	staff	time
caring	great	appointment
care	helpful	rude
patients	feel	staff
patient	questions	room
recommend	office	didn’t
family	friendly	wait

Both about staff/office issues

Topics in online doctor reviews:

RateMDs DOCTORS YOU CAN TRUST

	Staff/Office	Personality	Surgery
Positive	time staff great helpful feel questions office friendly	best years caring care patients patient recommend family	surgery first son life surgeon daughter recommend thank
Negativ e	office time appointment rude staff room didn’t wait	care medical patients doesn’t help know don’t problem	pain told went said surgery later didn’t months

FACTORIAL LDA

A multi-dimensional topic model

Word distributions are grouped into different concepts

- e.g. sentiment and aspect

Paul and Dredze (2012) **Factorial LDA: Sparse multi-dimensional text models.** Proceedings of *Advances in Neural Information Processing Systems (NIPS)*.

FACTORIAL LDA | DRUG DISCUSSIONS

Analyzing online drug forums:

FACTORIAL LDA | DRUG DISCUSSIONS

3-dimensional model:

- Drug type
- Route of administration (i.e. method of intake)
- Aspect

Drugs-Forum

Drug (22 total)	Route	Aspect
• Alcohol • Amphetamine • Cannabis • Cocaine • ... • Salvia • Tobacco	• Injection • Oral • Smoking • Snorting	• Chemistry • Culture • Effects • Health • Usage

FACTORIAL LDA | DRUG DISCUSSIONS

Joint model with **3** factors:

- Drug type
- Route of administration (i.e. method of intake)
- Aspect

Each “topic” is a triple such as:

(Cocaine,Snorting,Health)

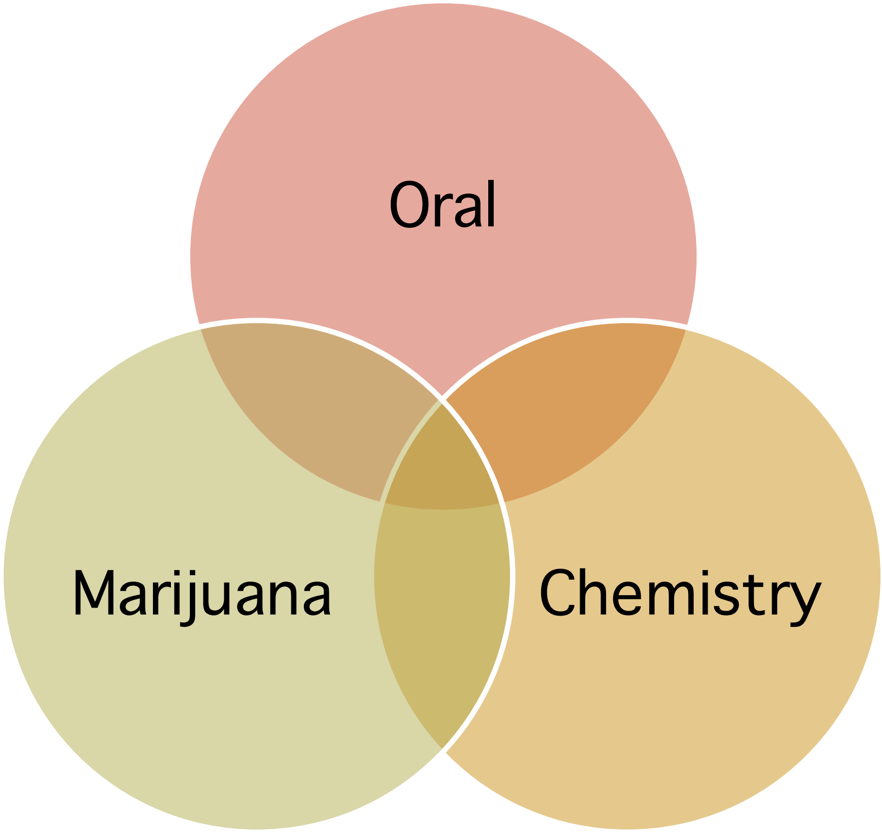
nose
pain
damage
blood
cocaine
problem

(Cocaine,Snorting,Usage)

coke
line
lines
nose
small
cut

FACTORIAL LDA | DRUG DISCUSSIONS

Suppose we want to model: (Marijuana,Oral,Chemistry)



FACTORIAL LDA DRUG DISCUSSIONS

Marijuana	Oral	Chemistry
weed	capsules	solvent
cannabis	consumes	extraction
thc	toast	evaporate
marijuana	stomach	evaporated
stoned	chewing	solvents
bowl	ambien	evaporation
bud	digestion	yield
joint	juice	chloride
blunt	absorbed	alkaloids
herb	ingestion	tek
bong	meal	compounds
pot	tiredness	evaporating
sativa	chew	atom
blaze	juices	aromatic
indica	gelatin	non-polar
smoking	yogurt	purified
blunts	fruit	jar
…	…	….

FACTORIAL LDA DRUG DISCUSSIONS

exp(

Marijuana		Oral		Chemistry
weed		capsules		solvent
cannabis		consumes		extraction
thc		toast		evaporate
marijuana		stomach		evaporated
stoned		chewing		solvents
bowl		ambien		evaporation
bud		digestion		yield
joint	+	juice	+	chloride
blunt		absorbed		alkaloids
herb		ingestion		tek
bong		meal		compounds
pot		tiredness		evaporating
sativa		chew		atom
blaze		juices		aromatic
indica		gelatin		non-polar
smoking		yogurt		purified
blunts		fruit		jar
…		…		….

)

FACTORIAL LDA

DRUG DISCUSSIONS

exp(Marijuana + Oral + Chemistry) =

Marijuana	Oral	Chemistry	
weed	capsules	solvent	thc
cannabis	consumes	extraction	method
thc	toast	evaporate	extraction
marijuana	stomach	evaporated	plant
stoned	chewing	solvents	material
bowl	ambien	evaporation	cannabis
bud	digestion	yield	simple
joint	juice	chloride	coffee
blunt	absorbed	alkaloids	oil
herb	ingestion	tek	contains
bong	meal	compounds	jar
pot	tiredness	evaporating	dried
sativa	chew	atom	process
blaze	juices	aromatic	dry
indica	gelatin	non-polar	water
smoking	yogurt	purified	extract
blunts	fruit	jar	results
…	…	….	…

Dirichlet(
thc
method
extraction
plant
material
cannabis
simple
coffee
oil
contains
jar
dried
process
dry
water
extract
results
…
)

FACTORIAL LDA — DRUG DISCUSSIONS

word distribution for the triple:

(Marijuana Oral Chemistry)

oil
water
butter
thc
weed
hash
cannabis
alcohol
make
milk
high
marijuana
add
…
mixture
hours
try
brownies

~ Dirichlet(

thc
method
extraction
plant
material
cannabis
simple
coffee
oil
contains
jar
dried
process
dry
water
extract
results
…

)

FACTORIAL LDA | DRUG DISCUSSIONS

oil
water
butter
thc
weed
hash
cannabis
alcohol
make
milk
high
marijuana
add
…
mixture
hours
try
brownies

~ Dirichlet(

thc
method
extraction
plant
material
cannabis
simple
coffee
oil
contains
jar
dried
process
dry
water
extract
results
…

)

FACTORIAL LDA | DRUG DISCUSSIONS

word distribution for the triple:

(Marijuana
Oral
Chemistry)

oil
water
butter
thc
weed
hash
cannabis
alcohol
make
milk
high
marijuana
add
…
mixture
hours
try
brownies

~ Dirichlet(

thc
method
extraction
plant
material
cannabis
simple
coffee
oil
contains
jar
dried
process
dry
water
extract
results
…

)

FACTORIAL LDA DRUG DISCUSSIONS

We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

Drugs-Forum

"What is the dosage when taking mephedrone orally?"

FACTORIAL LDA | DRUG DISCUSSIONS

We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

Drugs-Forum

"What is the dosage when taking mephedrone orally?"

$$\begin{pmatrix} \text{Mephedrone} \\ \text{Oral} \\ \text{Usage} \end{pmatrix}$$

FACTORIAL LDA — DRUG DISCUSSIONS

We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

Drugs-Forum

“What is the dosage when taking mephedrone orally?”

$\begin{pmatrix} \text{Mephedrone} \\ \text{Oral} \\ \text{Usage} \end{pmatrix}$

> If it is [someone who isn’t you]’s first time using Mephedrone [someone who isn’t me] recommends a 100mg oral dose on an empty stomach.

FACTORIAL LDA | DRUG DISCUSSIONS

We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

Drugs-Forum

"What is the dosage when taking mephedrone orally?"

(Mephedrone
Oral
Usage)

If it is [someone who isn't you]'s first time using Mephedrone [someone who isn't me] recommends a 100mg oral dose on an empty stomach.

Reference text:

It is recommended by users that Mephedrone be taken on an empty stomach. Doses usually vary between 100mg – 1g.

FACTORIAL LDA

SUMMARY

Word distributions are factored into multiple dimensions

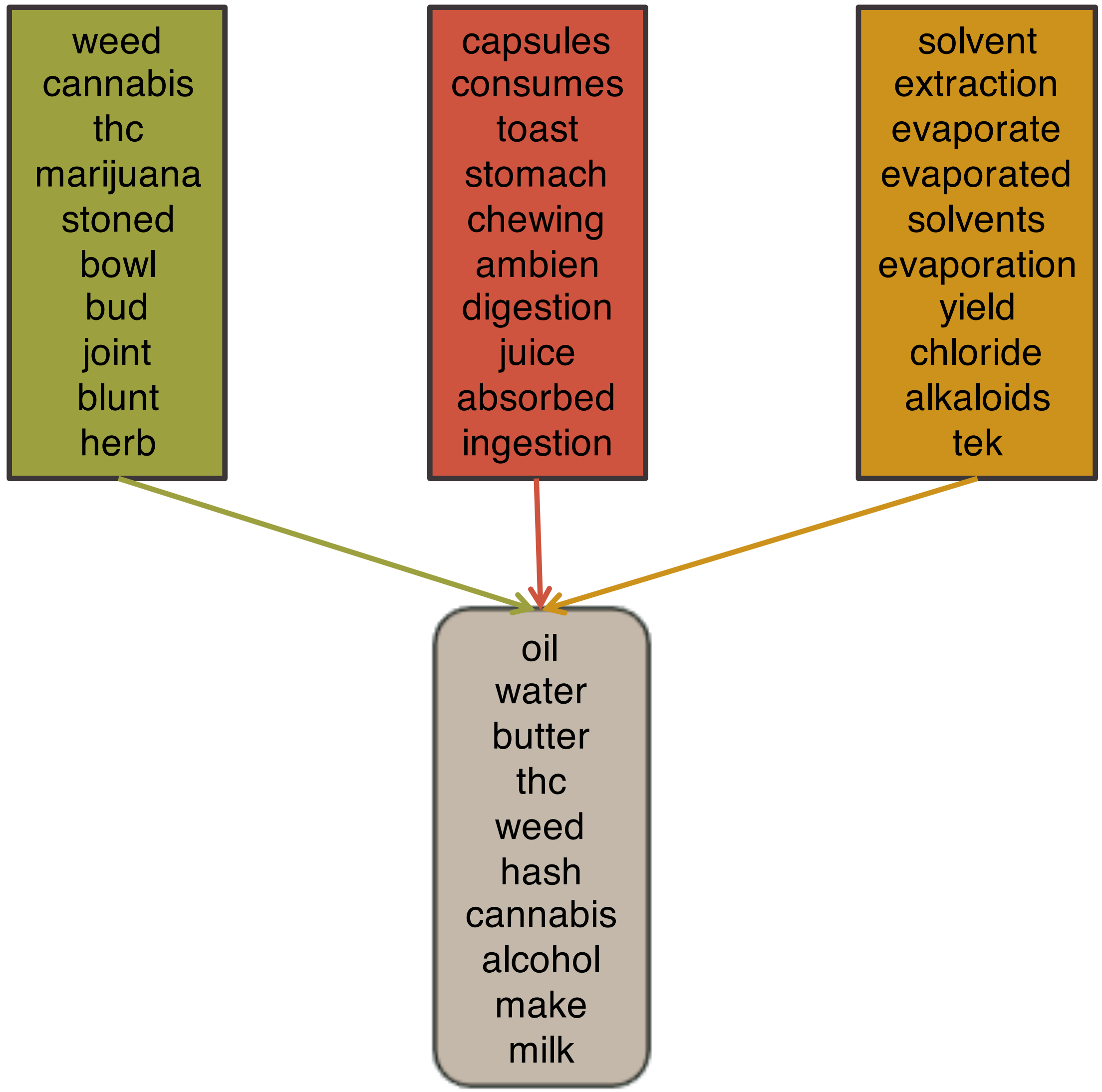
Each topic prior is informed by parameters for each dimension

	Staff/Office	Personality	Surgery
Positive	time staff great helpful feel questions office friendly	best years caring care patients patient recommend family	surgery first son life surgeon daughter recommend thank
Negative	office time appointment rude staff room didn’t wait	care medical patients doesn’t help know don’t problem	pain told went said surgery later didn’t months

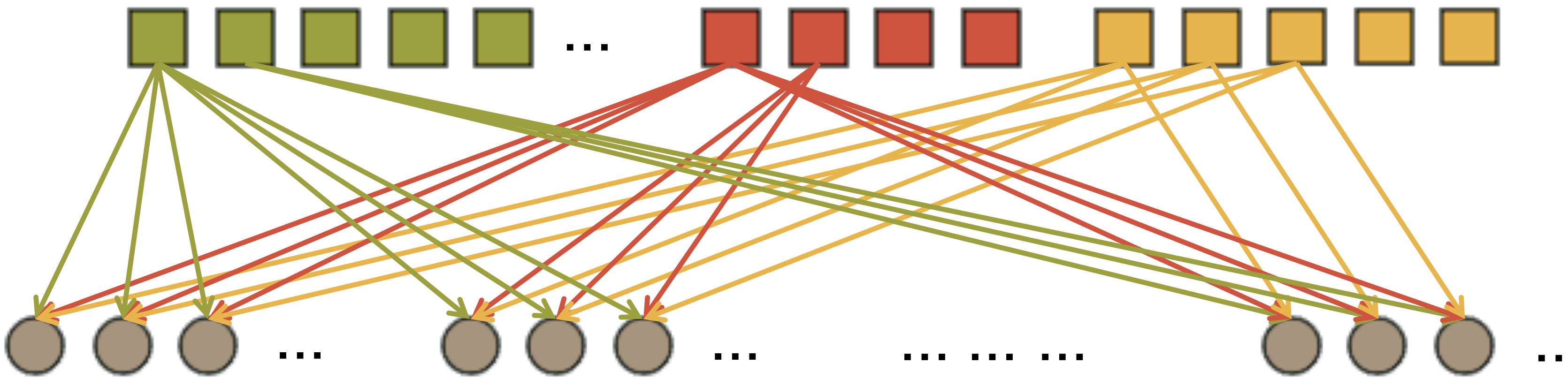
TOPIC STRUCTURES FACTORIZATION

Components:

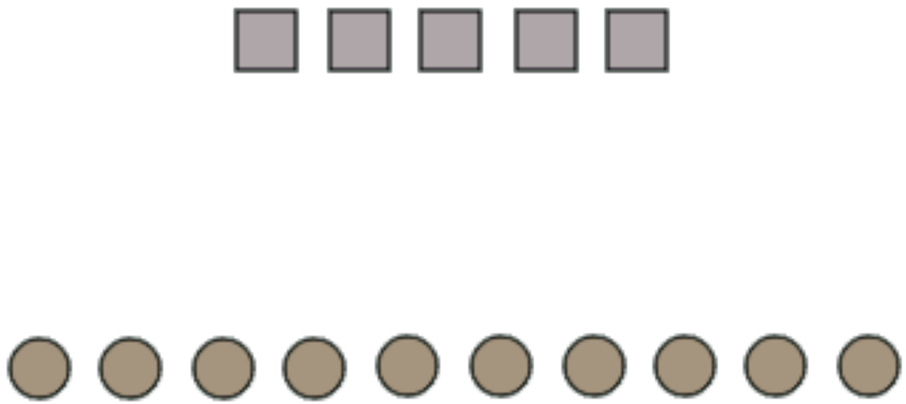
Topics:



TOPIC STRUCTURES | FACTORIZATION

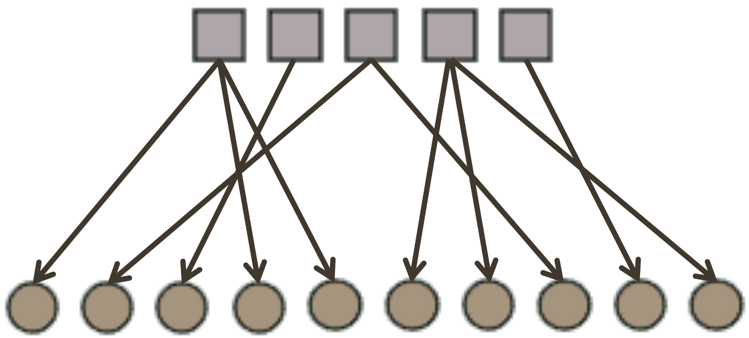


TOPIC STRUCTURES

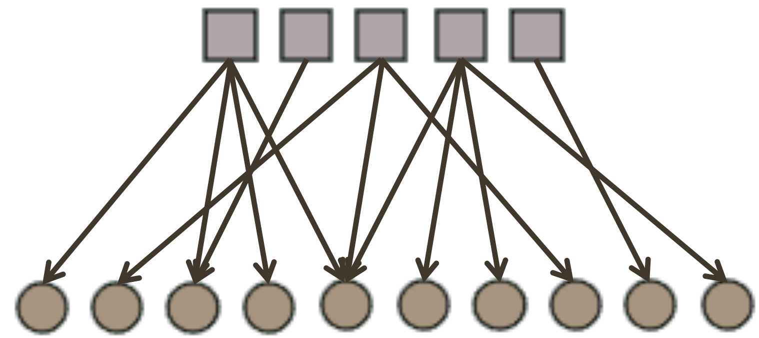


Generalization:
Components don’t need to be factored into groups.
They can feed into topics in many different ways!

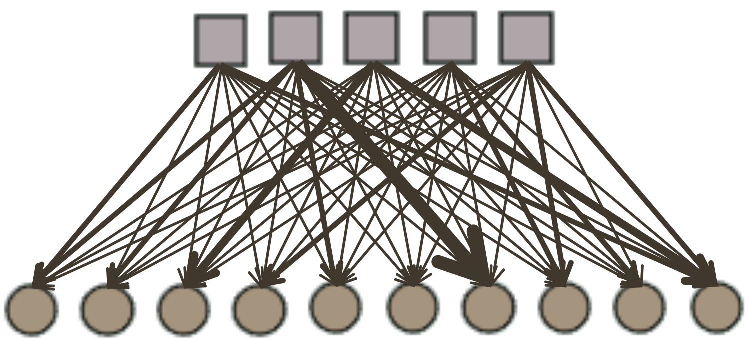
TOPIC STRUCTURES



Tree



Directed acyclic graph (DAG)



Weighted DAG

TOPIC STRUCTURES TREE

RateMDs
DOCTORS YOU CAN TRUST

"Surgery"

surgery
pain
went
dr
surgeon
told
procedure
months
performed
removed
left
fix
said
later
years

pain
surgery
dr
went
knee
foot
neck
mri
injury
shoulder
bone
months
told
surgeon
therapy

told
hospital
dr
blood
went
later
days
mother
said
er
cancer
weight
home
father
months

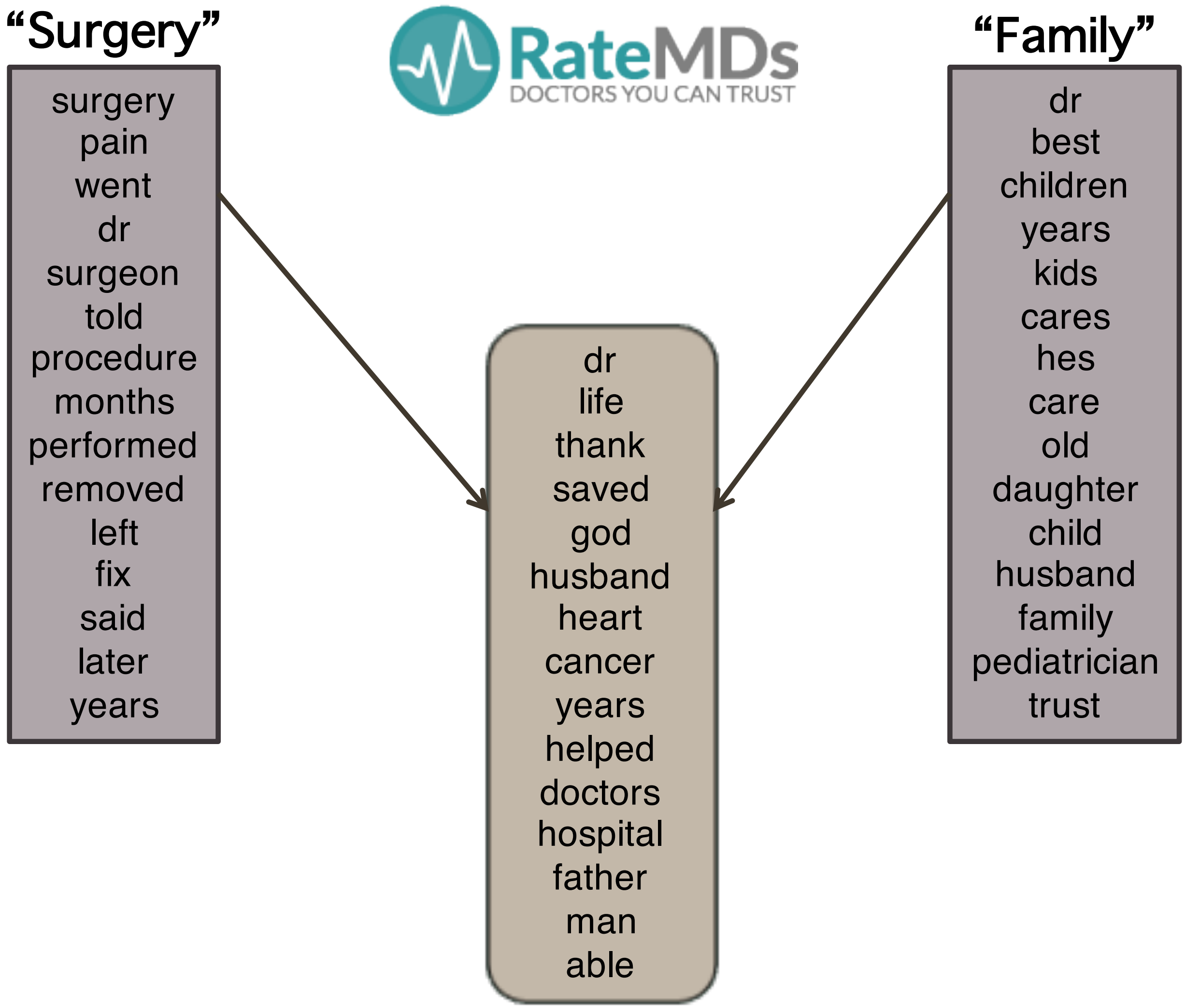
"Family"

dr
best
children
years
kids
cares
hes
care
old
daughter
child
husband
family
pediatrician
trust

dr
best
years
doctor
love
cares
ive
children
patients
hes
family
kids
seen
doctors
son

baby
son
pregnancy
dr
child
pregnant
ob
daughter
first
delivered
gyn
birth
delivery
section
hospital

TOPIC STRUCTURES (SPARSE) DAG



SPRITE

Structured-prior topic models

A family of topic models in which the Dirichlet priors are functions of underlying components

Paul and Dredze (2015) **SPRITE: Generalizing topic models with structured priors.** *Transactions of the Association for Computational Linguistics (TACL)* 3: 43-57.

SPRITE

SPRITE generalizes many existing topic models:

Model	Document priors	Topic priors
LDA	Single component	Single component
SCTM	Single component	Sparse binary β
SAGE	Single component	Sparse ω
FLDA	Binary δ is transpose of β	Factored binary β
PAM	α are supertopic weights	Single component
DMR	α are feature values	Single component

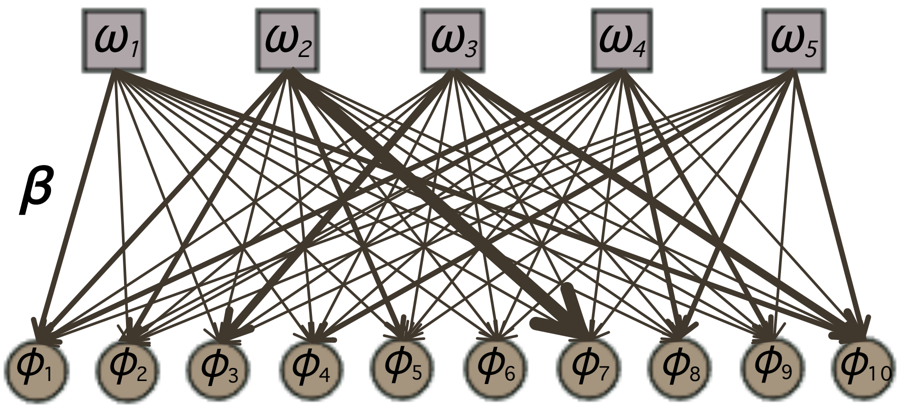
SPRITE | DEFINITION

The priors over word distributions are weighted combinations of components:

$$\tilde{\phi}_{iv} = \exp\left(\sum_{c=1}^{C(\phi)} \beta_{ic}\omega_{cv}\right)$$

$$\phi_i \sim \text{Dirichlet}(\tilde{\phi}_i)$$

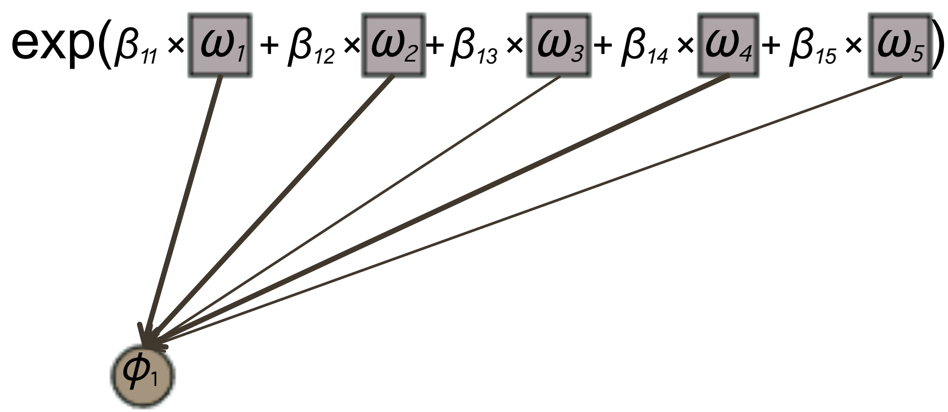
distribution over words in *i*th topic



SPRITE

DEFINITION

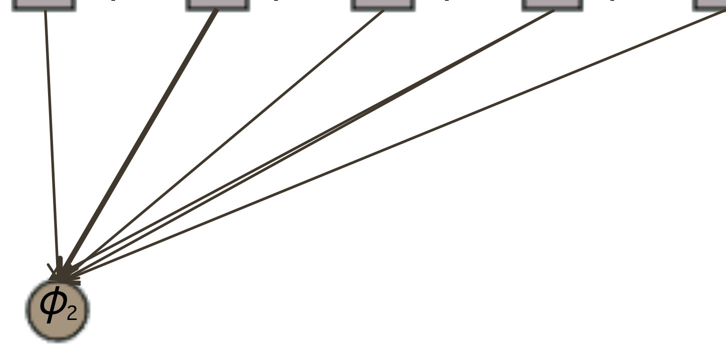
The priors over word distributions are weighted combinations of components:



SPRITE | DEFINITION

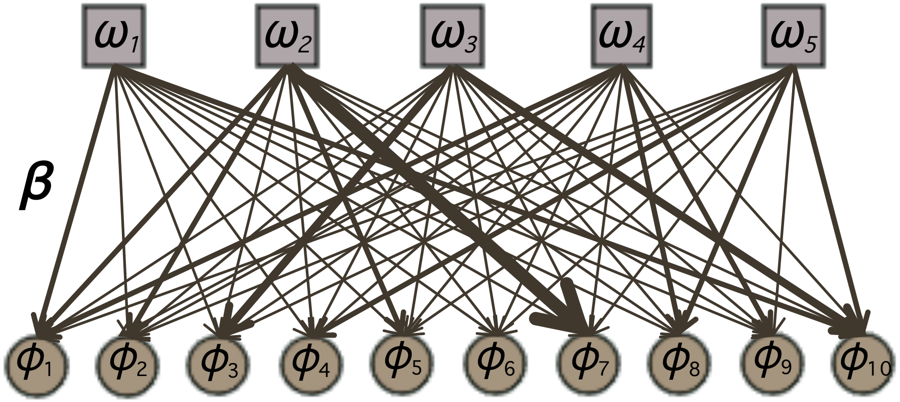
The priors over word distributions are weighted combinations of components:

$$\exp(\beta_{21} \times \boxed{\omega_1} + \beta_{22} \times \boxed{\omega_2} + \beta_{23} \times \boxed{\omega_3} + \beta_{24} \times \boxed{\omega_4} + \beta_{25} \times \boxed{\omega_5})$$



SPRITE DEFINITION

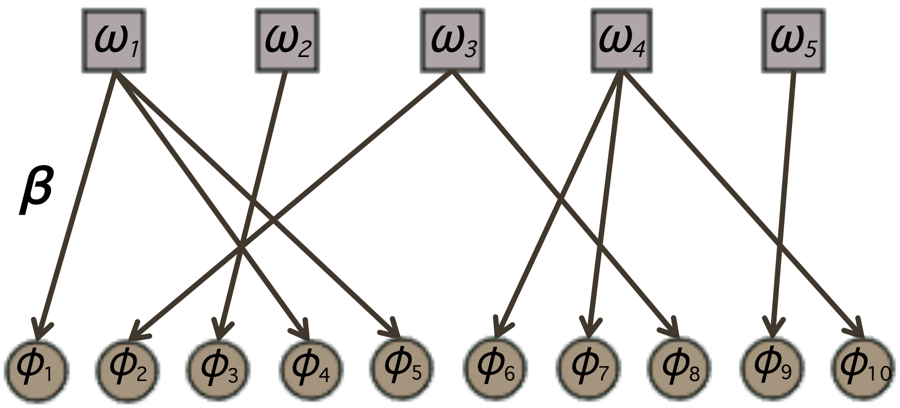
The priors over word distributions are weighted combinations of components:



We can induce different structures by constraining the values of β

SPRITE | DEFINITION

The priors over word distributions are weighted combinations of components:



SPRITE | DEFINITION

The priors over word distributions are weighted combinations of components:

$$\exp(1 \times \boxed{\omega_1} + 0 \times \boxed{\omega_2} + 0 \times \boxed{\omega_3} + 0 \times \boxed{\omega_4} + 0 \times \boxed{\omega_5})$$

$$\phi_1$$

SPRITE | DEFINITION

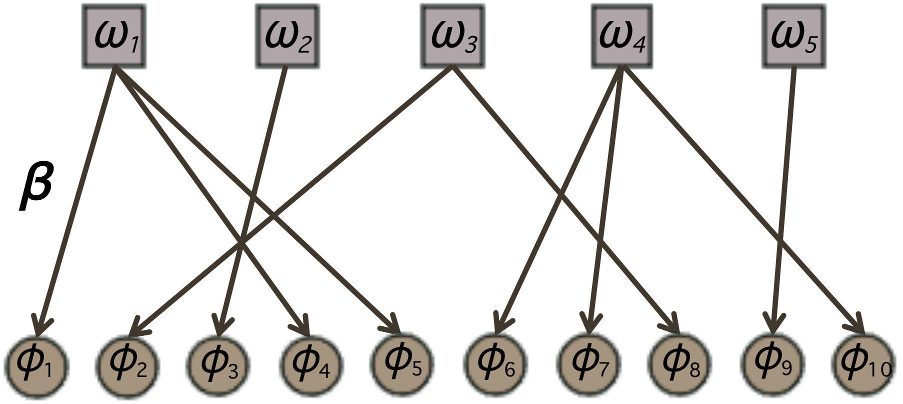
The priors over word distributions are weighted combinations of components:

$$\exp(0 \times \boxed{\omega_1} + 0 \times \boxed{\omega_2} + 1 \times \boxed{\omega_3} + 0 \times \boxed{\omega_4} + 0 \times \boxed{\omega_5})$$

ϕ_2

SPRITE | DEFINITION

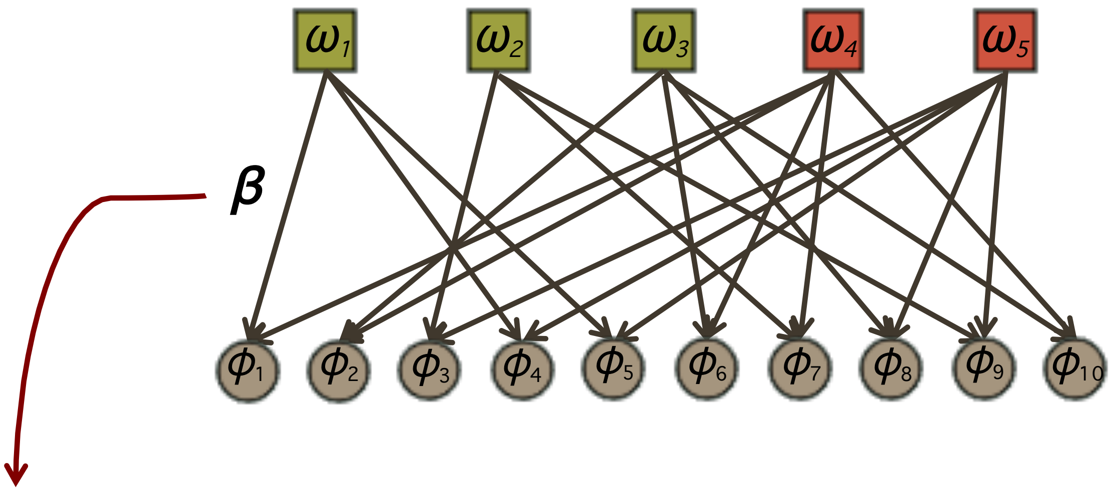
The priors over word distributions are weighted combinations of components:



Tree: Each topic's $\boldsymbol{\beta}$ vector is zero in all but one component

SPRITE	DEFINITION

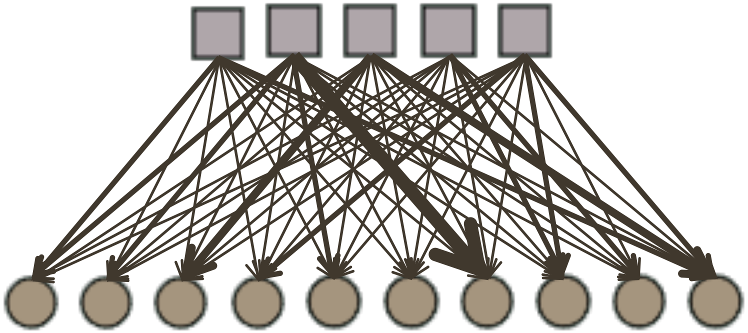
The priors over word distributions are weighted combinations of components:



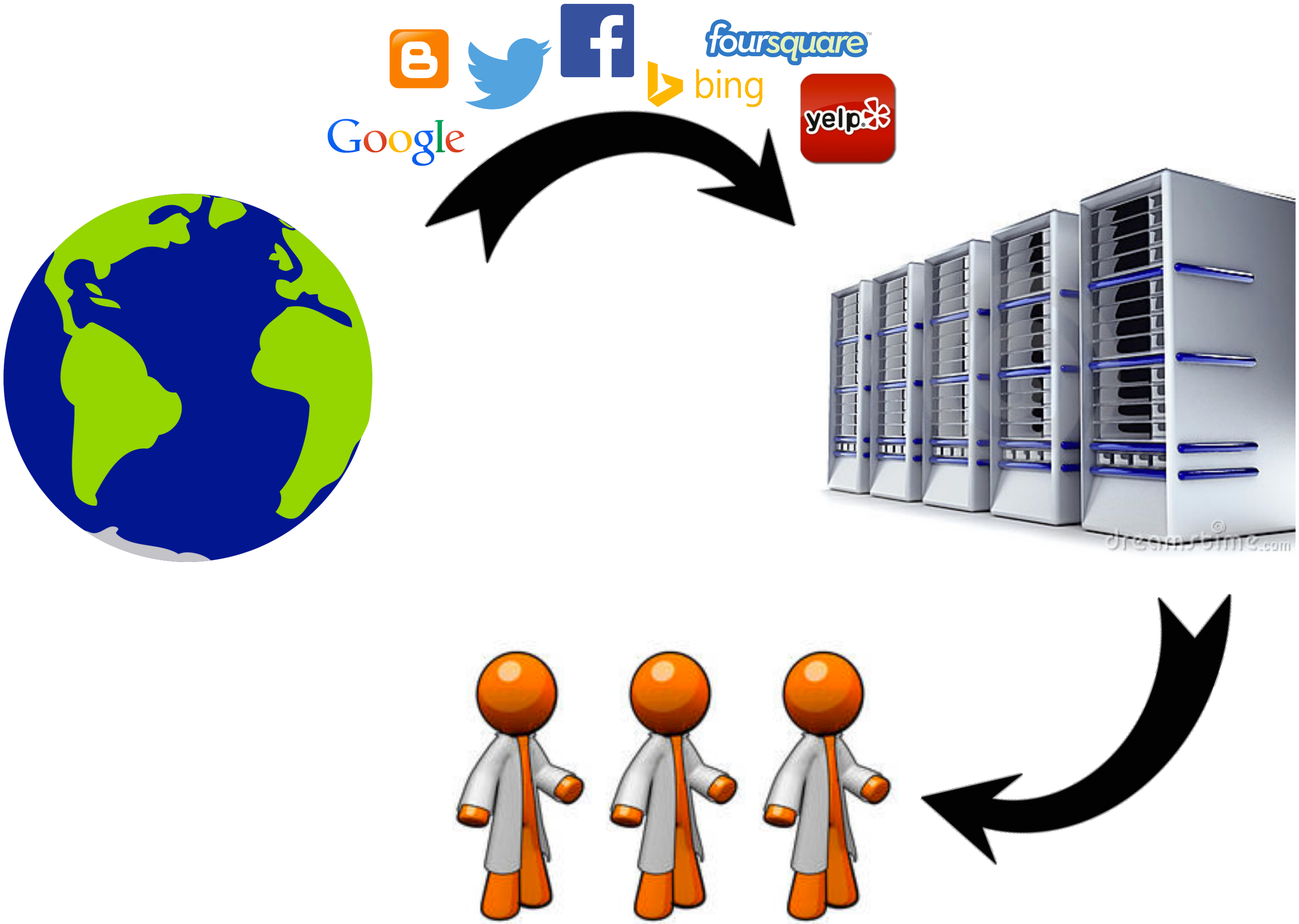
Factorization: Like a tree, but a nonzero component in each factor

SPRITE | SUMMARY

- Organizes topics in a variety of useful ways
  - Can be tailored toward different applications
- Generalizes many topic models
  - While opening up new possibilities
- Allows practitioners to make sense of big text data
  - Can drive new scientific research



MOVING FORWARD



MOVING FORWARD

MOVING FORWARD — BETTER MODELS

Text data $\xrightarrow[\text{Modeling}]{}$ NLP $\xrightarrow[\text{Validation}]{}$ External data

MOVING FORWARD | BETTER MODELS

$$\text{Text data} \xrightarrow[\text{Modeling}]{} \text{NLP} \xleftarrow[\text{Modeling}]{} \text{External data}$$

MOVING FORWARD

BETTER MODELS

DATA.GOV

GET STARTED

SEARCH OVER 138,198 DATASETS

Agriculture

Business

Climate

Consumer

Ecosystems

Education

Energy

Finance

Health

Local Government

Manufacturing

Ocean

Public Safety

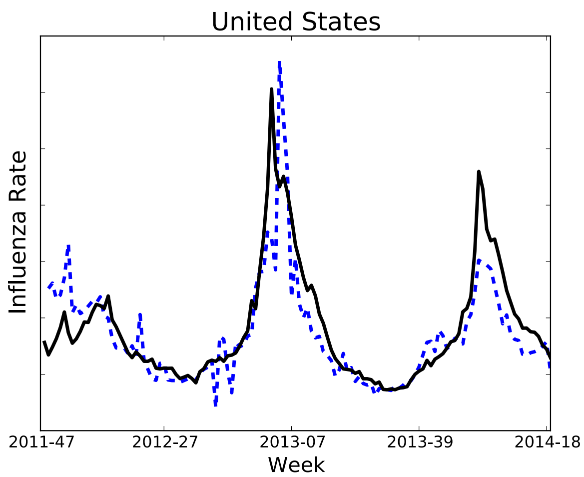
Science & Research

MOVING FORWARD BETTER MODELS

Challenges with big models:

- Spurious correlations between text and datasets

Modeling flu prevalence in Twitter:



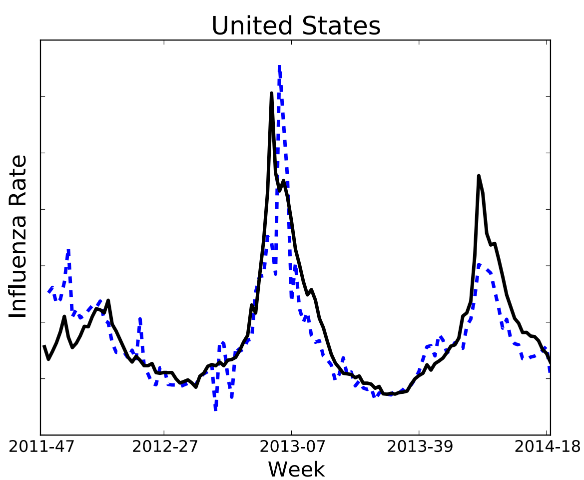
flu
sick
swine
shot
cancer
fever
h1n1
#beatcancer
better
getting
home
halloween
breast

MOVING FORWARD BETTER MODELS

Challenges with big models:

- Spurious correlations between **text** and **datasets**

Modeling flu prevalence in Twitter:



flu
sick
swine
shot
cancer
fever
h1n1
#beatcancer
better
getting
home
halloween
breast

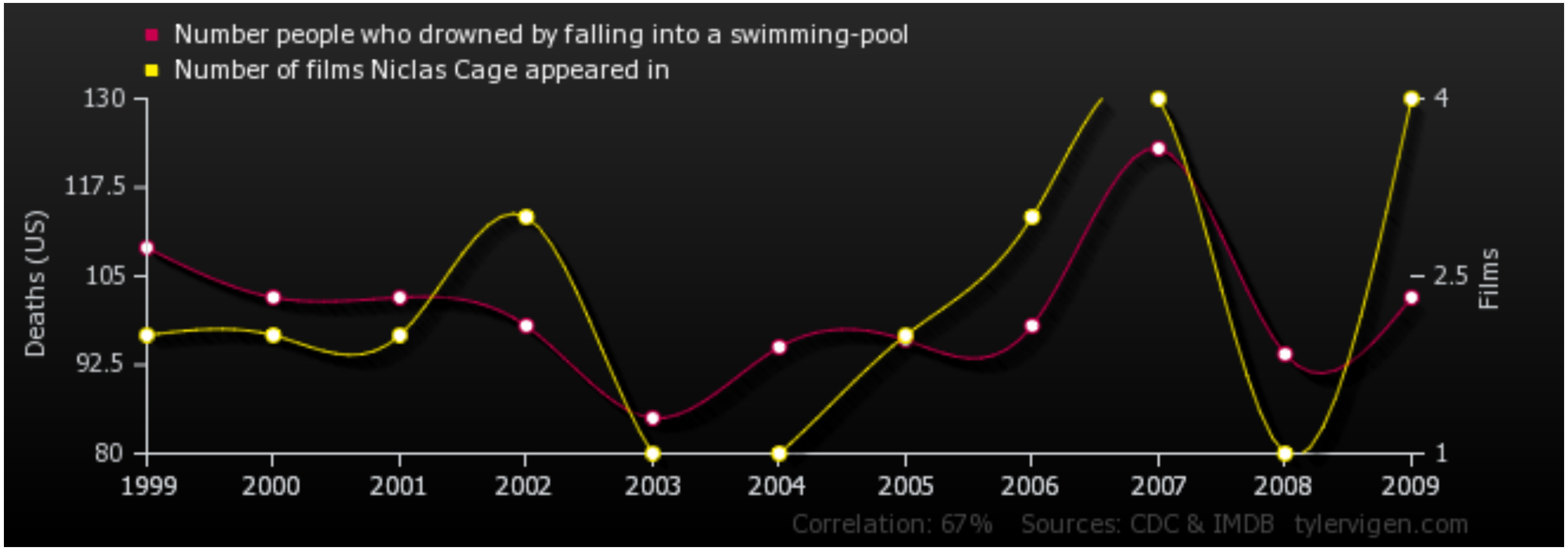
Covariance prior:

sick
flu
swine
better
shot
getting
cancer
home
hope
fever
feel
feeling
h1n1

MOVING FORWARD | BETTER MODELS

Challenges with big models:

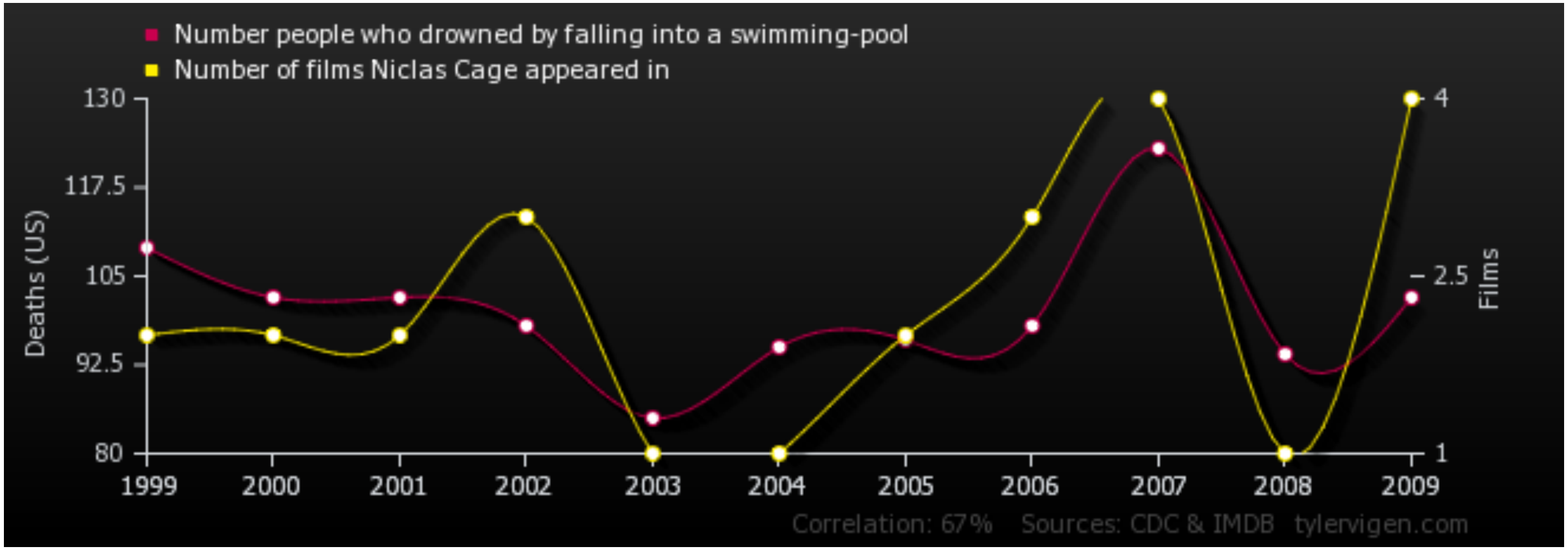
- Spurious correlations between different datasets



MOVING FORWARD — BETTER MODELS

Challenges with big models:

- Spurious correlations between different datasets



- Need models with domain expertise
  - Opportunities for interactive machine learning

BETTER MODELS

Challenges with big models:

- Solutions: language structure and human feedback

MOVING FORWARD

MOVING FORWARD | NEW QUESTIONS

What happens when this article gets published?

TIME

HEALTH CANCER

Most Types of Cancer Just 'Bad Luck,' Researchers Say

MOVING FORWARD | NEW QUESTIONS

What happens when this article gets published?

TIME

HEALTH CANCER

Most Types of Cancer Just 'Bad Luck,' Researchers Say

Most cancers are due to bad luck, not lifestyle, researchers say. time.com/3651785/cancer… ...Anyone fancy a smoke?

Most cancer just 'bad luck'? time.com/3651785/cancer… So I'm going for checkup w/ my magic eightball today. Just saved copay

MOVING FORWARD | NEW QUESTIONS

Variation in cancer risk among tissues can be explained by the number of stem cell divisions

Cristian Tomasetti[1*] and Bert Vogelstein[2*]

Some tissue types give rise to human cancers millions of times more often than other tissue types. Although this has been recognized for more than a century, it has never been explained. Here, we show that the lifetime risk of cancers of many different types is strongly correlated (0.81) with the total number of divisions of the normal self-renewing cells maintaining that tissue's homeostasis. These results suggest that only a third of the variation in cancer risk among tissues is attributable to environmental factors or inherited predispositions. The majority is due to "bad luck," that is, random mutations arising during DNA replication in normal, noncancerous stem cells. This is important not only for understanding the disease but also for designing strategies to limit the mortality it causes.

Extreme variation in cancer incidence across different tissues is well known; for example, the lifetime risk of being diagnosed with cancer is 6.9% for lung, 1.08% for thyroid, 0.6% for brain and the rest of the nervous system, 0.003% for pelvic bone and 0.00072% for laryngeal cartilage (*1–3*). Some of these differences are associated with well-known risk factors such as smoking, alcohol use, ultraviolet light, or human papilloma virus (HPV) (*4, 5*), but this applies only to specific populations exposed to potent mutagens or viruses. And such exposures cannot explain why cancer risk in tissues within the alimentary tract can differ by as much as a factor of 24 [esophagus (0.51%), large intestine (4.82%), small intestine (0.20%), and stomach (0.86%)] (*3*). Moreover, cancers of the small intestinal epithelium are three times less common than brain tumors (*3*), even though small intestinal epithelial cells are exposed to much higher levels of environmental mutagens than are cells within the brain, which are protected by the blood-brain barrier.

Another well-studied contributor to cancer is

Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows

--Statistical modeling links cancer risk with number of stem cell divisions

Release Date: January 1, 2015
Addendum to news release added Jan. 7, 2015

Johns Hopkins Medicine is gratified by the responses and discussion generated by Cristian Tomasetti and Bert Vogelstein's research paper, "Variation in cancer risk among tissues can be explained by the number of stem cell divisions," published in Science on Jan. 2, 2015, and a news release describing the work, "Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows." Cancer is driven by a number of factors and causes, and concepts related to calculating risk are complex and often the subject of debate. To facilitate the ongoing discussion, and to address the many thoughtful questions their research stimulated, the two scientists have provided the following answers to frequently asked questions.

Is there an analogy that can help put the results of your research in perspective?

Getting cancer could be compared to getting into a car accident. Our results would be equivalent to showing a high correlation between length of trip and getting into an

FOR THE ME
Contacts:
Vanessa Wasta
410-614-2916
wasta@jhmi.edu

Downl
Photo/video files
this story are ava
download.

We are providing thi understanding that i help illustrate the st corresponding news find appropriate cred image in the release else you need, pleas JHMedia@jhmi.edu

View Files

HEALTH CANCER

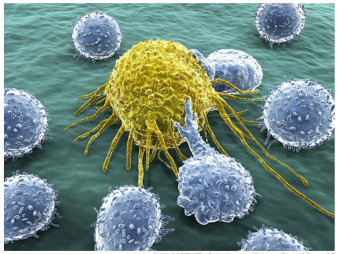
Most Types of Cancer Just 'Bad Luck,' Researchers Say

Helen Regan @hcregan | Jan. 2, 2015

Two thirds of cancers could be explained as biological misfortune

Researchers have found that bad luck plays a major role in determining most types of cancer, rather than genetics or risky lifestyle choices such as smoking.

The results, published in the journal *Science* on Thursday, found that random DNA mutations that amass in the body when stem cells divide into various tissues cause two thirds of cancers.



JUAN GARTNER–Getty Images/Science Photo Library RF
Lymphocytes and cancer cell

Next Avenue @NextAvenue · Feb 9
Surprising Research on #Cancer and Bad Luck: is.gd/oFfIBj
1

Food&Drink Fanatic @foodanddrinkfan · Feb 8
Nope, cancer is not "just bad luck": Red meat contains saturated fats and the amino acid carnitine. If you con... cur.lv/irej5

Cancer Daily @cancrdaily · Feb 8
Nope, cancer is not "just bad luck" - Arizona Daily Star bit.ly/1Iz7oer #cancer #health
1

Next Avenue @NextAvenue · Feb 8
If you're a #cancer survivor, you'll want to read this piece: is.gd/oFfIBj
1

Stem Cell News @mystemcellnews · Feb 8
Experts pan tying cancer to bad luck: JHU scientists Christian Tomasetti and Bert Vogelstein had analysed stem... bit.ly/1KxLdCT

MOVING FORWARD

NEW QUESTIONS

Variation in cancer risk among tissues can be explained by the number of stem cell divisions

Cristian Tomasetti[1*] and Bert Vogelstein[2*]

Some tissue types give rise to human cancers millions of times more often than other tissue types. Although this has been recognized for more than a century, it has never been explained. Here, we show that the lifetime risk of cancers of many different types is strongly correlated (0.81) with the total number of divisions of the normal self-renewing cells maintaining that tissue's homeostasis. These results suggest that only a third of the variation

Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows

--Statistical modeling links cancer risk with number of stem cell divisions

Release Date: January 1, 2015

Addendum to news release added Jan. 7, 2015

Johns Hopkins Medicine is gratified by the responses and discussion generated by

FOR THE ME

Contacts:

Vanessa Wasta

410-614-2916

wasta@jhmi.edu

View Files

supplementary materials). With the number of clusters set equal to two, the tumors were classified in an unsupervised manner into one cluster with high ERS (9 tumor types) and another with low ERS (22 tumor types) (Fig. 2).

Next Avenue @NextAvenue · Feb 9

Surprising Research on #Cancer and Bad Luck: is.gd/oFf8j

Food&Drink Fanatic @foodanddrinkfan · Feb 8

Nope, cancer is not "just bad luck": Red meat contains saturated fats and the amino acid carnitine. If you con... cur.lv/inej5

Cancer Daily @cancrdaily · Feb 8

Next Avenue @NextAvenue · Feb

Stem Cell News

HEALTH CANCER

Most Types of Cancer Just 'Bad Luck,' Researchers Say

Helen Regan @hcregan | Jan. 2, 2015

Two thirds of cancers could be explained as biological misfortune

After examining 31 cancer types, researchers found 22 were from mutations in stem cells that could not be prevented.

MOVING FORWARD | NEW QUESTIONS

Variation in cancer risk among tissues can be explained by the number of stem cell divisions

Cristian Tomasetti[1*] and Bert Vogelstein[2*]

Some tissue types give rise to human cancers millions of times more often than other tissue types. Although this has been recognized for more than a century, it has never been explained. Here, we show that the lifetime risk of cancers of many different types is strongly correlated (0.81) with the total number of divisions of the normal self-renewing cells maintaining that tissue's homeostasis. These results suggest that only a third of the variation in cancer risk among tissues is attributable to environmental factors or inherited predispositions. The majority is due to "bad luck," that is, random mutations arising during DNA replication in normal, noncancerous stem cells. This is important not only for understanding the disease but also for designing strategies to limit the mortality it causes.

Extreme variation in cancer incidence across different tissues is well known; for example, the lifetime risk of being diagnosed with cancer is 6.9% for lung, 1.08% for thyroid, 0.6% for brain and the rest of the nervous system, 0.003% for pelvic bone and 0.00072% for laryngeal cartilage (*1–3*). Some of these differences are associated with well-known risk factors such as smoking, alcohol use, ultraviolet light, or human papilloma virus (HPV) (*4, 5*), but this applies only to specific populations exposed to potent mutagens or viruses. And such exposures cannot explain why cancer risk in tissues within the alimentary tract can differ by as much as a factor of 24 [esophagus (0.51%), large intestine (4.82%), small intestine (0.20%), and stomach (0.86%)] (*3*). Moreover, cancers of the small intestinal epithelium are three times less common than brain tumors (*3*), even though small intestinal epithelial cells are exposed to much higher levels of environmental mutagens than are cells within the brain, which are protected by the blood-brain barrier.

Another well-studied contributor to cancer is

Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows

--Statistical modeling links cancer risk with number of stem cell divisions

Release Date: January 1, 2015
Addendum to news release added Jan. 7, 2015

Johns Hopkins Medicine is gratified by the responses and discussion generated by Cristian Tomasetti and Bert Vogelstein's research paper, "Variation in cancer risk among tissues can be explained by the number of stem cell divisions," published in Science on Jan. 2, 2015, and a news release describing the work, "Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows." Cancer is driven by a number of factors and causes, and concepts related to calculating risk are complex and often the subject of debate. To facilitate the ongoing discussion, and to address the many thoughtful questions their research stimulated, the two scientists have provided the following answers to frequently asked questions.

Is there an analogy that can help put the results of your research in perspective?

Getting cancer could be compared to getting into a car accident. Our results would be equivalent to showing a high correlation between length of trip and getting into an

FOR THE ME
Contacts:
Vanessa Wasta
410-614-2916
wasta@jhmi.edu

Downl
Photo/video files
this story are ava
download.

We are providing thi understanding that i help illustrate the st corresponding news find appropriate cred image in the release else you need, pleas JHMedia@jhmi.edu

View Files

Wallace, Paul, Elhadad (2015) **What predicts media coverage of health science articles?** *AAAI Workshop on the World Wide Web and Public Health Intelligence*.

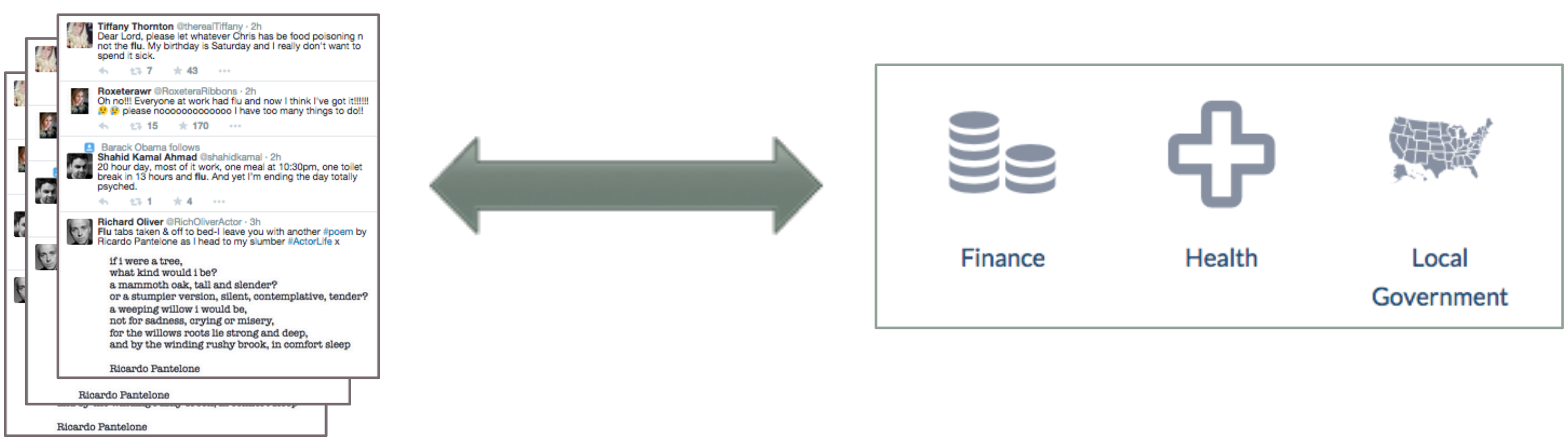
CS

MOVING FORWARD | SUMMARY

There's no end to exciting questions we can ask of **big, open data**

We need methods to link what people are saying on **the web** with **real-world trends**

This requires advancements at the intersection of **language processing** and **data science**



THANK YOU

WITH HELP FROM:

Advisors: Mark Dredze, Jason Eisner
Funding: Microsoft Research, NSF, JHU Dean's office

Flu:
- CS David Broniatowski
- Andrea Dugas
- Nicholas Generous
- CS Alex Lamb
- CS Michael Smith

Medical search:
- CS Eric Horvitz
- CS Ryen White
- CS Janice Tsai
- Sara Javid
- Luis Diaz

Air pollution:
- CS Shiliang Wang
- CS Angie Chen
- Brian Schwartz

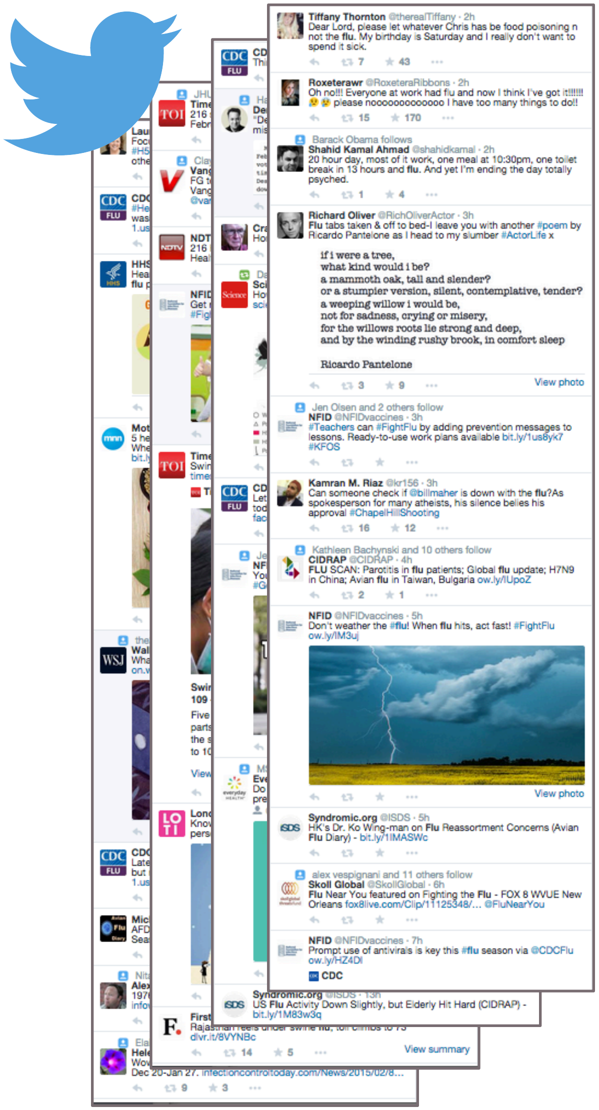
Doctor reviews:
- CS Byron Wallace
- Urmimala Sarkar
- Thomas Trikalinos

Drug forums:
- Meg Chisolm
- Matthew Johnson
- Ryan Vandrey

THANK YOU

QUESTIONS?

MOVING FORWARD | MORE DATA



…

"What opinions are people tweeting about gun rights?"

…

MOVING FORWARD

MORE DATA

Sprite can model this!

We created a structured prior that incorporates geographic data about gun ownership

- Certain topics have a higher/lower prior depending on whether a tweet is from a high/low gun ownership state

Alabama	2,623	1,294	51.7	1,329	48.3
Alaska	2,716	1,627	57.8	1,089	42.2
Arizona	3,066	989	31.1	2,077	68.9
Arkansas	2,780	1,431	55.3	1,349	44.7
California*	3,897	846	21.3	3,051	78.7
Colorado	1,947	629	34.7	1,318	65.3
Connecticut*	7,449	1,279	16.7	6,170	83.3
Delaware*	3,421	934	25.5	2,487	74.5
The District	1,859	66	3.8	1,793	96.2
Florida*	4,454	1,072	24.5	3,382	75.5
Georgia	4,277	1,745	40.3	2,532	59.7
Hawaii*	4,450	477	8.7	3,973	91.3
Idaho	4,430	2,394	55.3	2,036	44.7
Illinois*	2,103	396	20.2	1,707	79.8

Benton, Paul, Hancock, Dredze. **A structured model of topic and perspective in social media.** Submitted to *ICWSM*.

CS

Sprite can model this!

Low gun ownership		High gun ownership	
violence	children	nra	teachers
culture	kids	war	school
less	likely	even	armed
sense	times	keep	schools
problem	murdered	murder	carry
makes	giving	members	god
world	nothing	liberal	kids
country	stand	fear	teacher
common	sorry	government	protect
america's	insane	call	security

Associated with: Low gun ownership — High gun ownership

This topic is associated with high gun ownership:

guns
’merica
truck
shoot
deer
hunting
day
beer
season
friends

MOVING FORWARD — MORE DATA

This topic is associated with high gun ownership:

guns
’merica
truck
shoot
deer
hunting
day
beer
season
friends

Probably also associated with:

- Population density
- Political affiliation
- …

MOVING FORWARD

FACTORIAL LDA — DEFINITION

Prior for triple (*i*,*j*,*k*):

$$\tilde{\phi}_{(i,j,k)v} = \exp(\omega_{iv}^{(\text{drug})} + \omega_{jv}^{(\text{route})} + \omega_{kv}^{(\text{aspect})})$$

$$\phi_{(i,j,k)} \sim \text{Dirichlet}(\tilde{\phi}_{(i,j,k)})$$

distribution over words for this triple

In general, prior for tuple ***t***:

$$\tilde{\phi}_{\vec{t}v} = \exp(\sum_{k=1}^{K} \omega_{\vec{t}_k v}^{(k)})$$

$$\phi_{\vec{t}} \sim \text{Dirichlet}(\tilde{\phi}_{\vec{t}})$$

SPRITE | DEFINITION

β_i has one value of 1 in each factor; 0 elsewhere

Recall (FLDA):

$$\tilde{\phi}_{iv} = \exp(\sum_{c=1}^{C(\phi)} \beta_{ic}\omega_{cv}) \quad \Longleftrightarrow \quad \tilde{\phi}_{\vec{t}v} = \exp(\sum_{k=1}^{K} \omega_{\vec{t}_k v}^{(k)})$$

$$\tilde{\theta}_{mi} = \exp(\sum_{c=1}^{C(\theta)} \alpha_{mc}\delta_{ci}) \quad \Longleftrightarrow \quad \tilde{\theta}_{m\vec{t}} = \exp(\sum_{k=1}^{K} \alpha_{m\vec{t}_k}^{(k)})$$

δ_c is transpose of β_i

SPRITE	DEFINITION

β_i has one value of 1 in each factor; 0 elsewhere

Recall (FLDA):

$$\tilde{\phi}_{iv} = \exp(\sum_{c=1}^{C(\phi)} \beta_{ic}\omega_{cv}) \quad \Longleftrightarrow \quad \tilde{\phi}_{\vec{t}v} = \exp(\sum_{k=1}^{K} \omega_{\vec{t}_k v}^{(k)})$$

$$\tilde{\theta}_{mi} = \exp(\sum_{c=1}^{C(\theta)} \alpha_{mc}\delta_{ci}) \quad \Longleftrightarrow \quad \tilde{\theta}_{m\vec{t}} = \exp(\sum_{k=1}^{K} \alpha_{m\vec{t}_k}^{(k)})$$

δ_c is transpose of β_i

SPRITE DEFINITION

Suppose we want to model how **perspective** influences topics

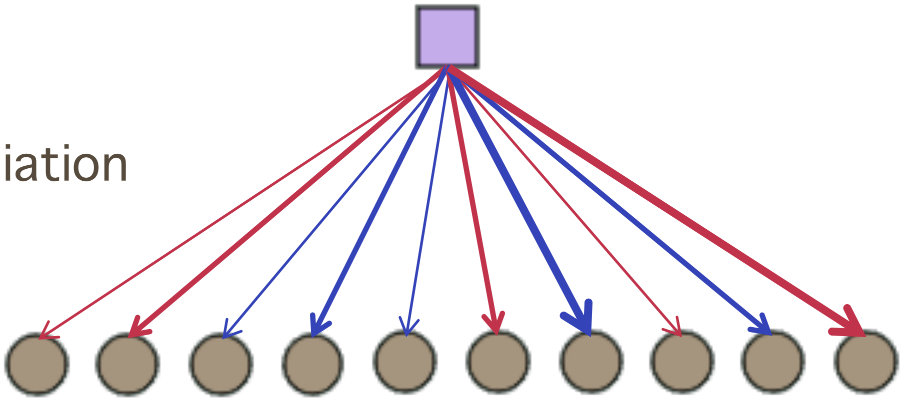
- e.g. certain topics are "pro" or "anti" gun control

A single-component SPRITE model:

$$\tilde{\phi}_{kv} = \exp(r_k \omega_v)$$

The vth word's perspective association

The kth topic's perspective association



SPRITE	DEFINITION

Suppose we want to model how **perspective** influences topics

- e.g. certain topics are "pro" or "anti" gun control

A single-component SPRITE model:

$$\tilde{\phi}_{kv} = \exp(r_k \omega_v)$$

The vth word's perspective association

The kth topic's perspective association

$$\tilde{\theta}_{mk} = \exp(\alpha_m r_k)$$

The mth document's perspective association

A positive r_k means:

- Words with positive ω_k are more likely in topic k
- Topic k is more likely in documents with positive α_m

SPRITE DEFINITION

Suppose we want to model how **perspective** influences topics

- e.g. certain topics are "pro" or "anti" gun control

A single-component SPRITE model:

Incorporating soft supervision: $\alpha_m \sim \mathcal{N}(s_m, \sigma^2)$

$$\tilde{\theta}_{mk} = \exp(\alpha_m r_k)$$

Supervision s_m is a function of:

- Survey data (% gun ownership in each state)
- Hashtags (#GunControlNow vs #NoGunControl)

MOVING FORWARD

BETTER DATA

Previous section: linking text to population data

Another idea: linking text to individual data

MOVING FORWARD | BETTER DATA

Previous section: linking text to population data

Another idea: linking text to individual data

Extraversion:

Introversion:

Schwartz et al. (2013) **Personality, gender, and age in the language of social media: the open-vocabulary approach.** *PLOS ONE*.

MOVING FORWARD | BETTER DATA

Previous section: linking text to population data

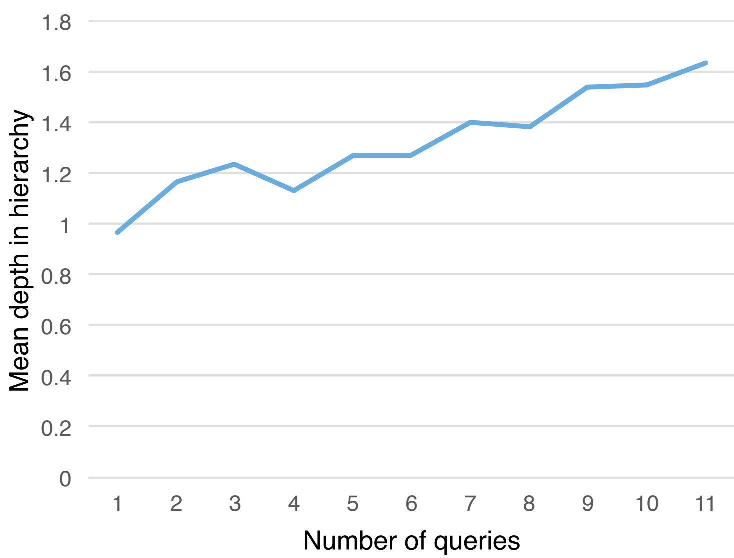
Another idea: linking text to individual data

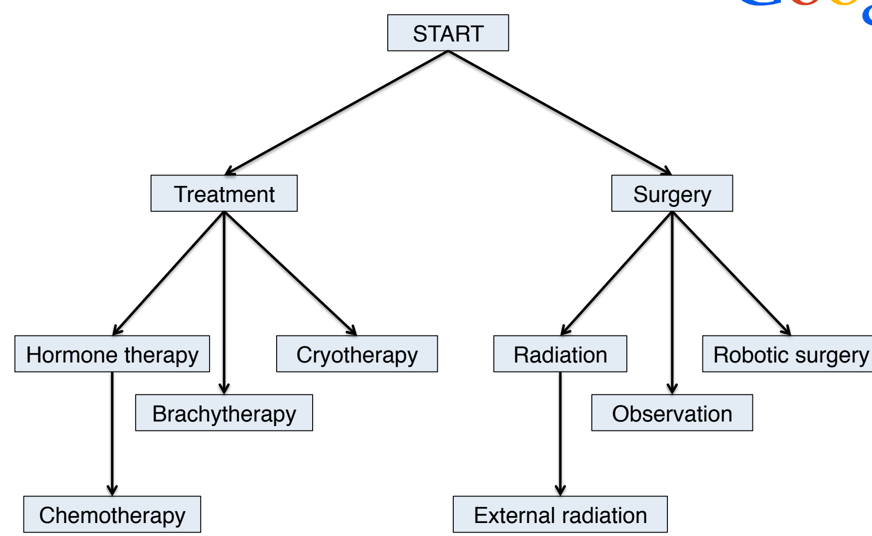
- New territory: clinical records

MOVING FORWARD BETTER DATA

How does the web influence medical decision-making?

bing Google





Paul, White, Horvitz (2015) **Web search as medical decision support for cancer.** *International World Wide Web Conference (WWW).*

CS