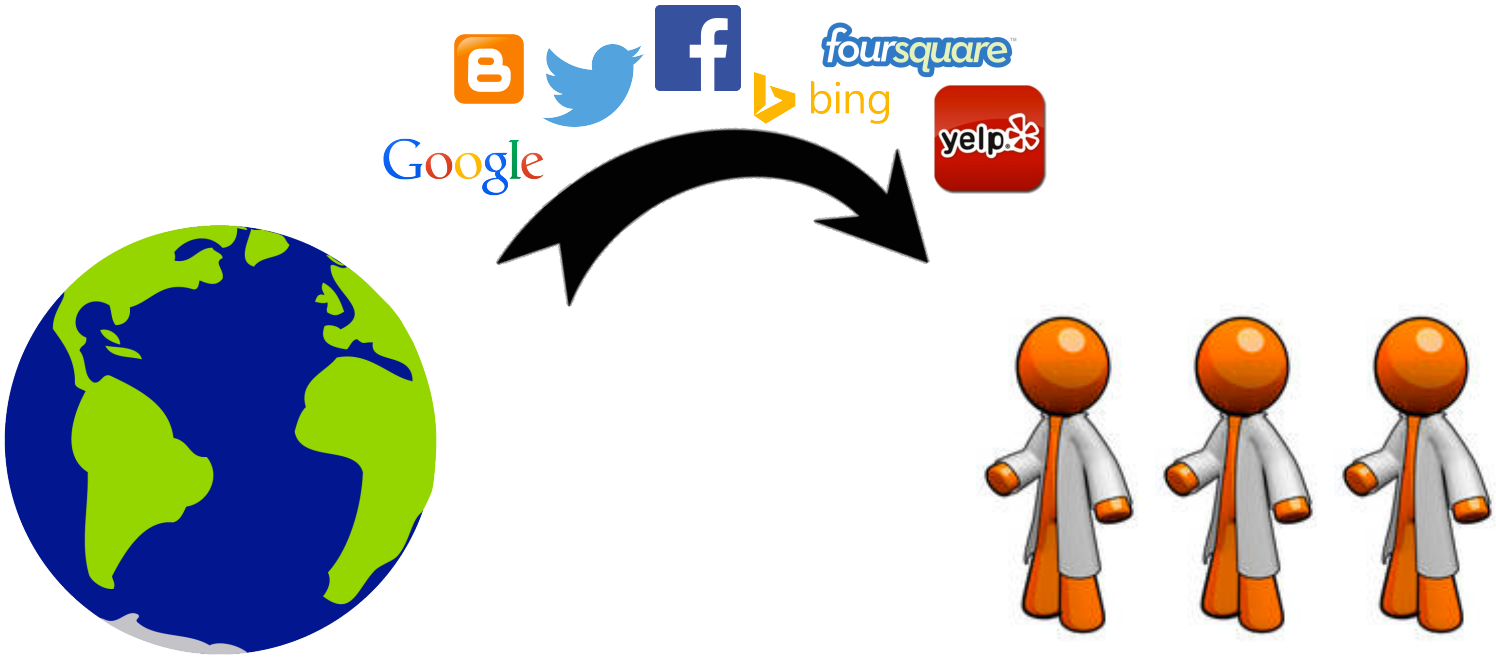
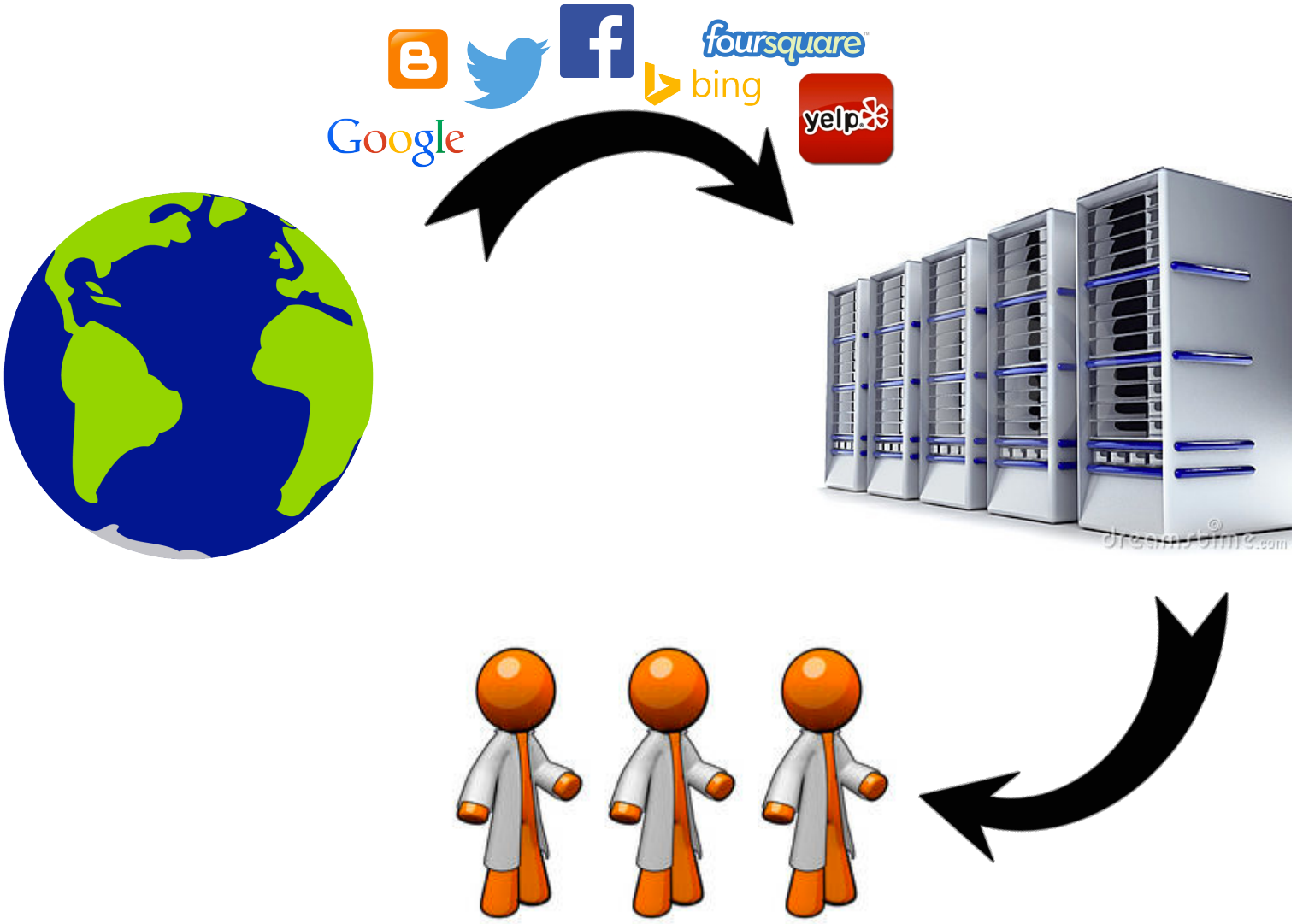
TOPIC MODELING WITH STRUCTURED PRIORS FOR TEXT-DRIVEN SCIENCE

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY

TEXT AS DATA



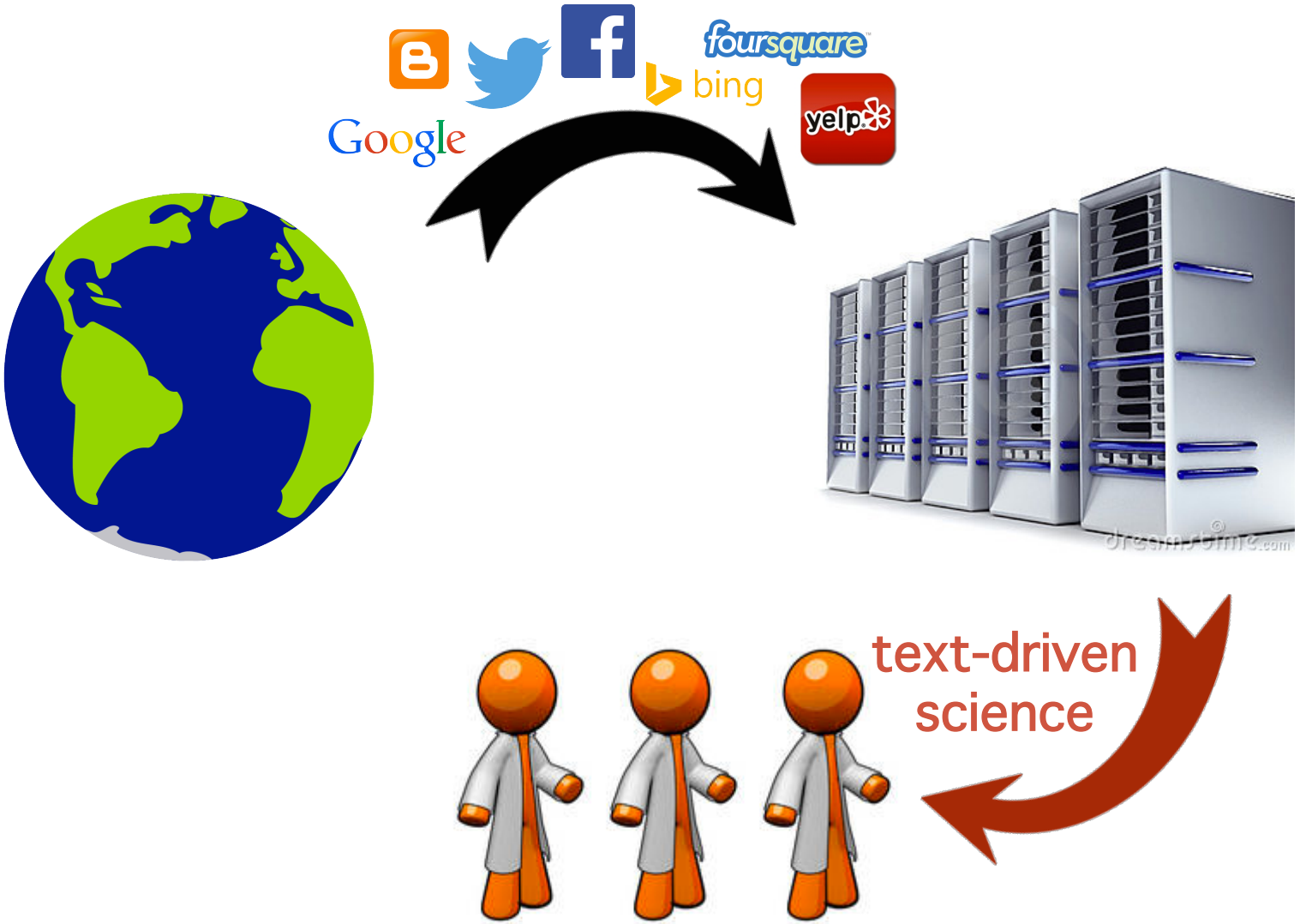
TEXT AS DATA



TEXT AS DATA


foursquare
bing
Google
yelp
natural language
processing (NLP)

TEXT AS DATA



TEXT AS DATA

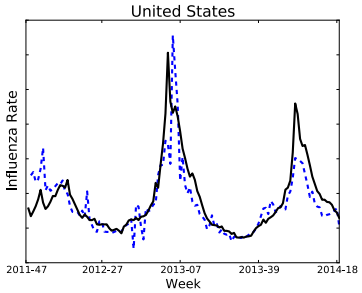
- Computational social science
- Computational journalism
- Crisis informatics
- Public health informatics
- Computational epidemiology

text-driven science

TEXT AS DATA

EXAMPLES

- Disease monitoring in Twitter



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.

- Air pollution in Chinese social media

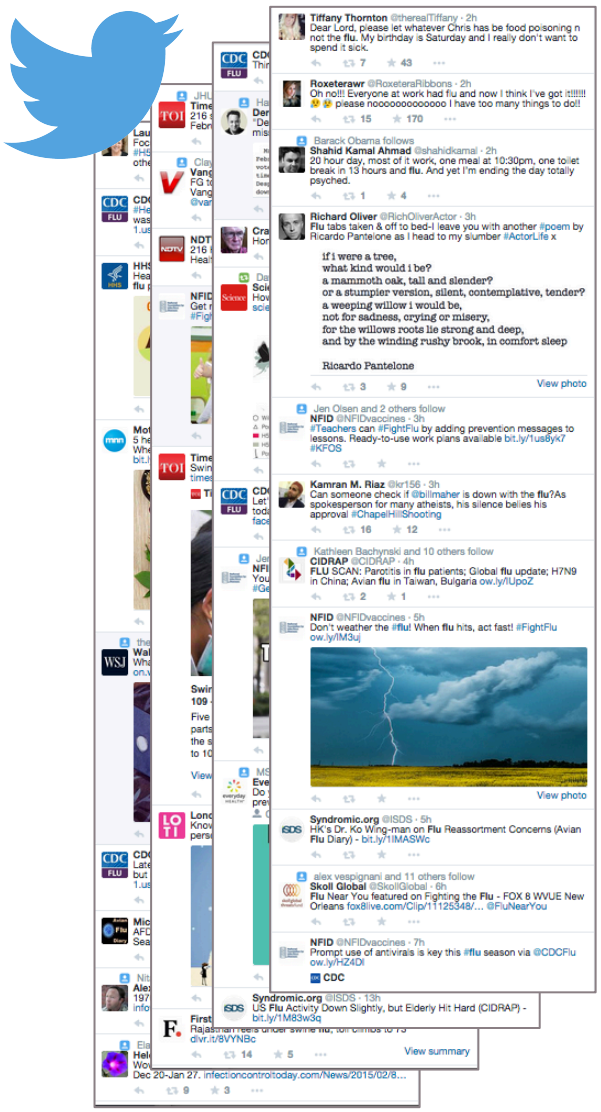
Wang, Paul, Dredze (2015) **Social media as a sensor of air quality and public response in China.** *Journal of Medical Internet Research*.

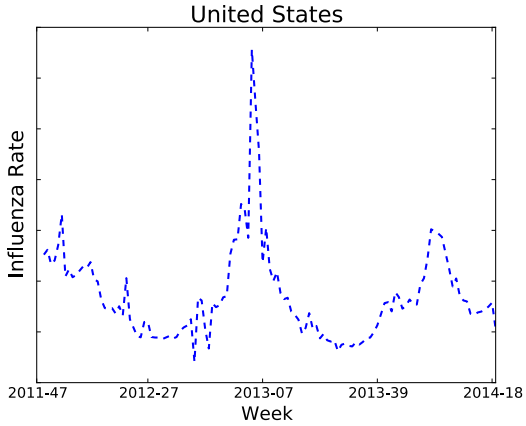
- Measuring healthcare quality from online reviews

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 AS DATA

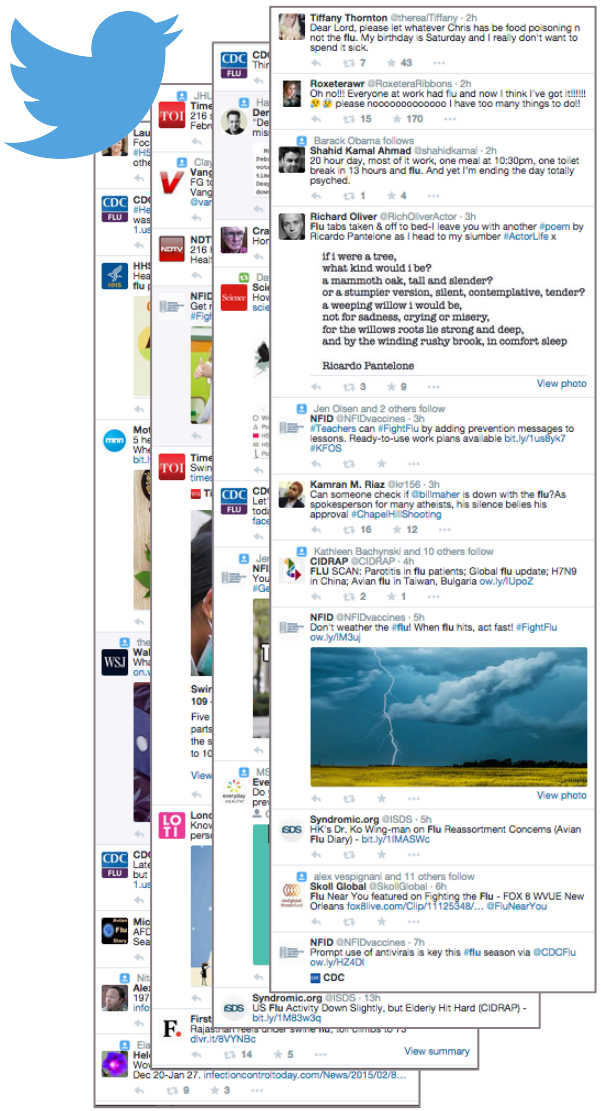
MAKING SENSE





TEXT AS DATA

MAKING SENSE



…

Topic modeling

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	…

TOPIC MODELING

A topic model is a probabilistic model of text

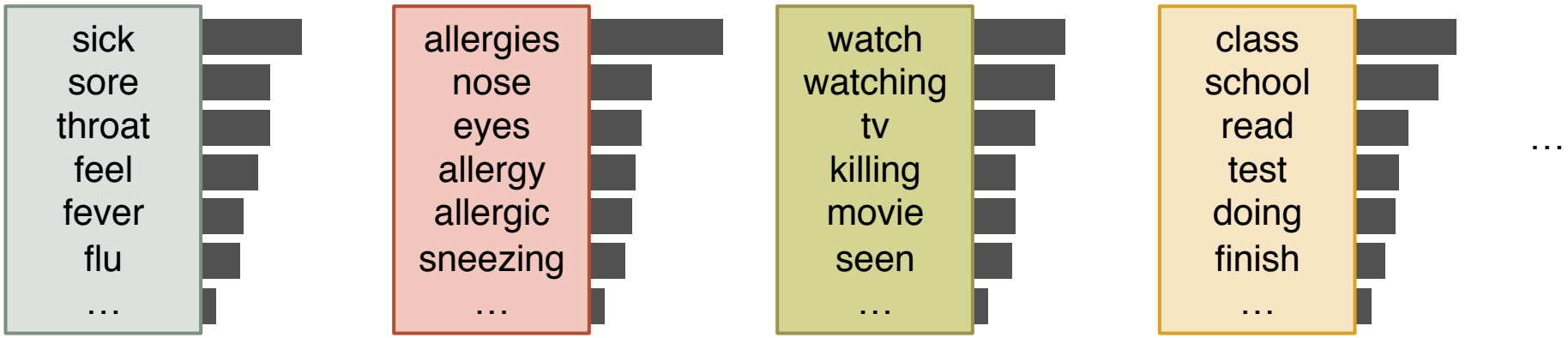
- We pretend that our data (text) are the output of a probabilistic process that generates data

TOPIC MODELING

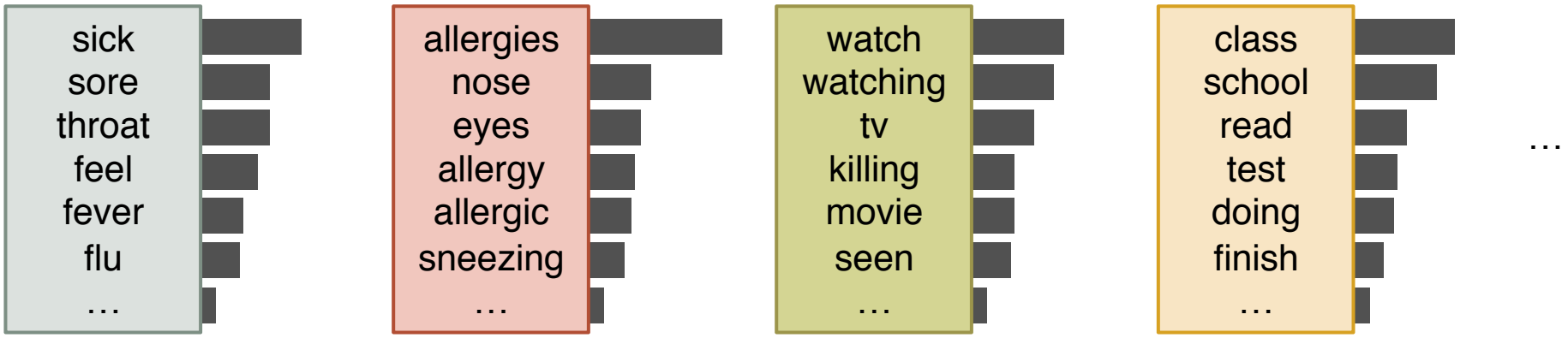
sick	allergies	watch	class
sore	nose	watching	school
throat	eyes	tv	read
feel	allergy	killing	test
fever	allergic	movie	doing
flu	sneezing	seen	finish
…	…	…	…

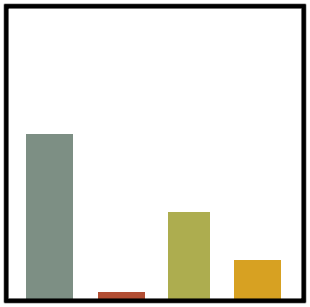
…

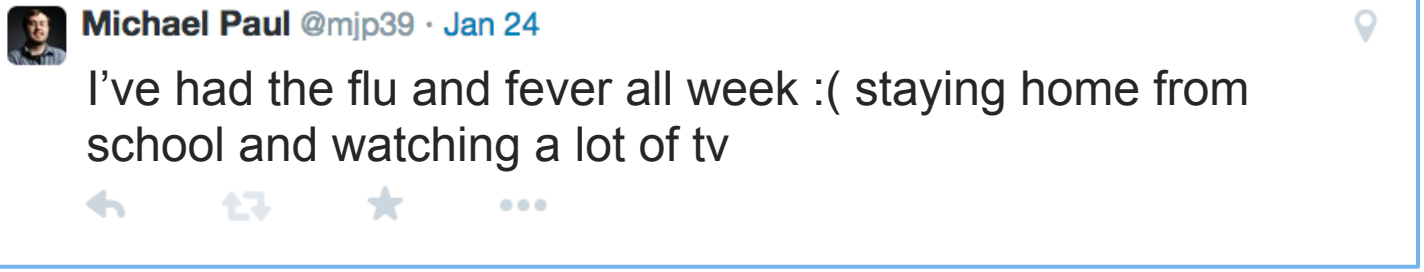
TOPIC MODELING



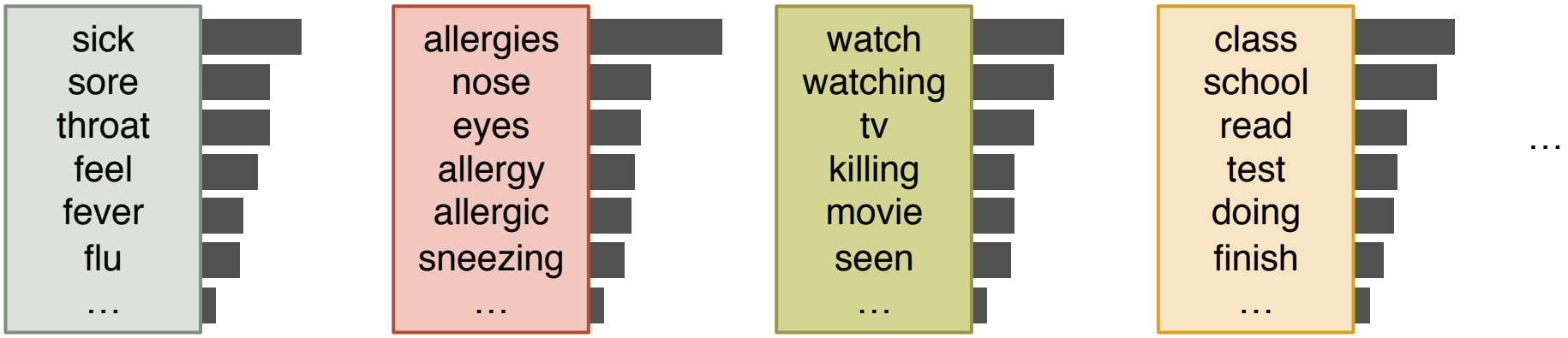
TOPIC MODELING

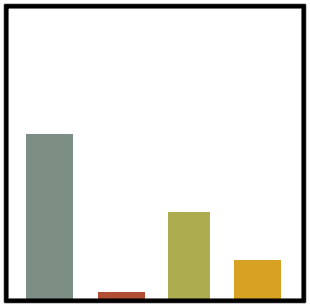


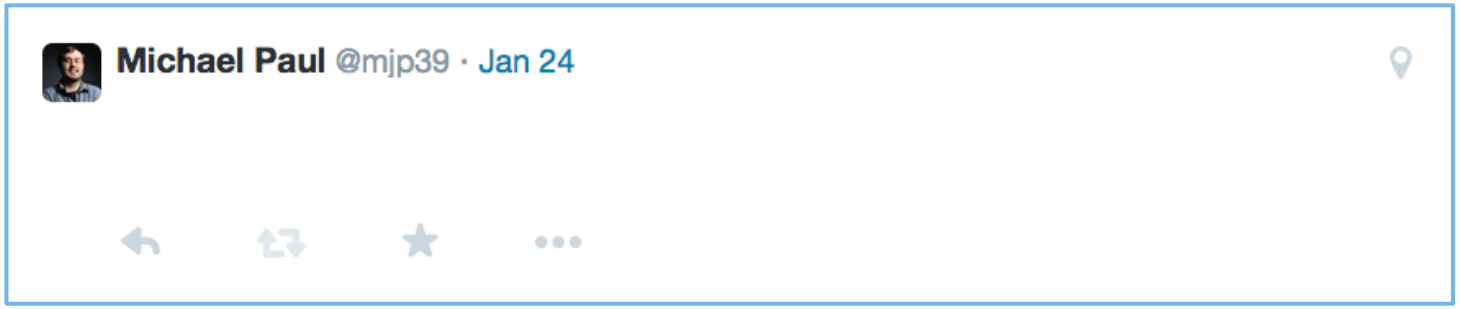




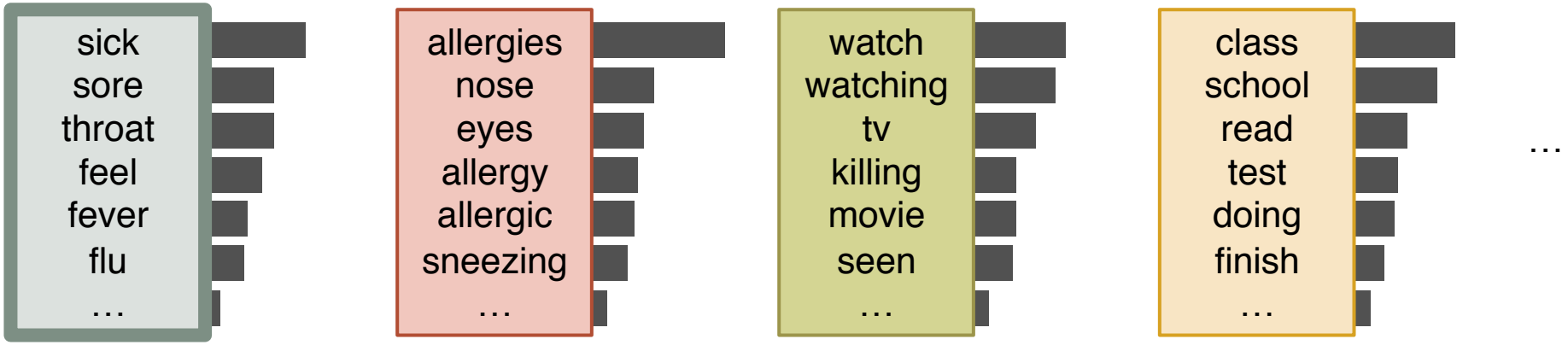
TOPIC MODELING

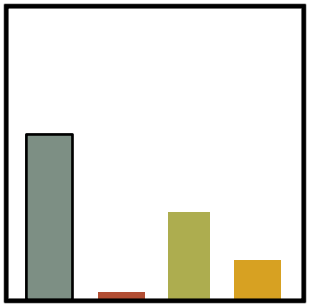


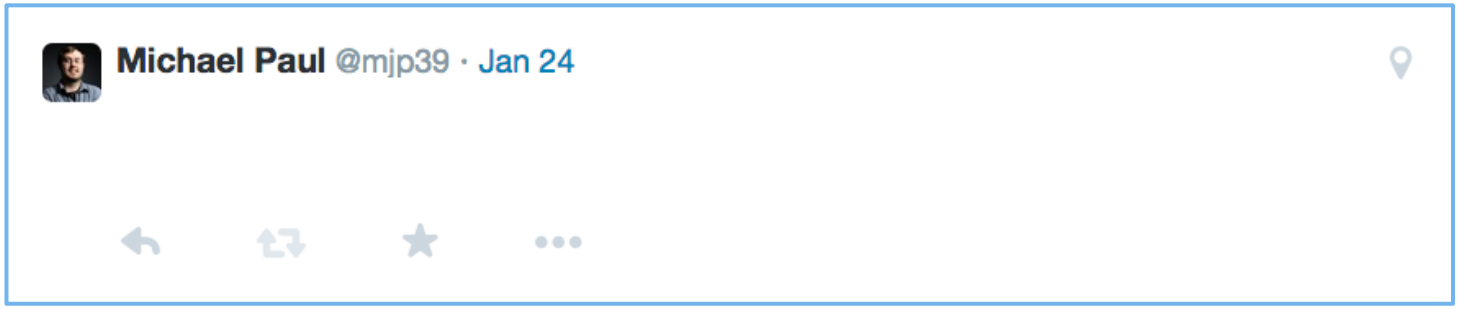




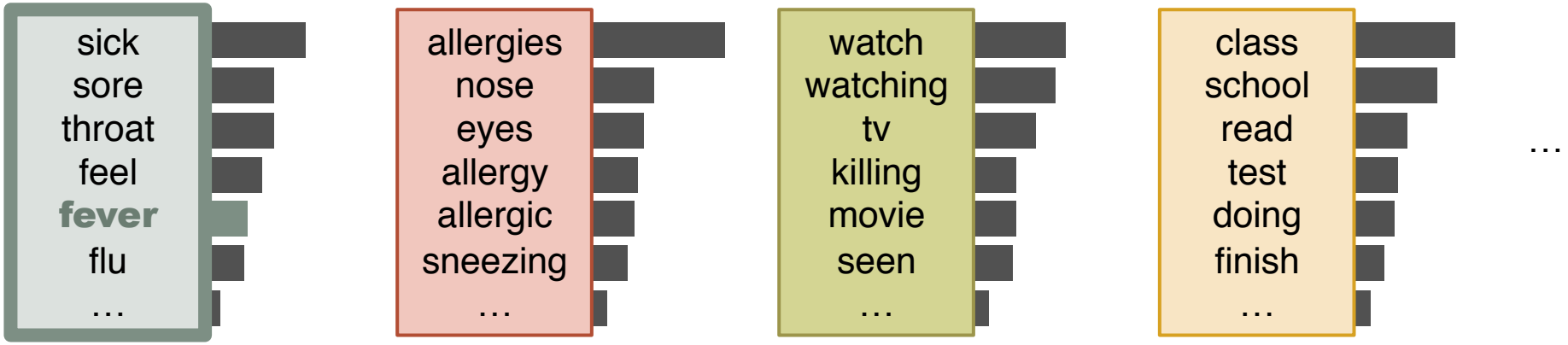
TOPIC MODELING

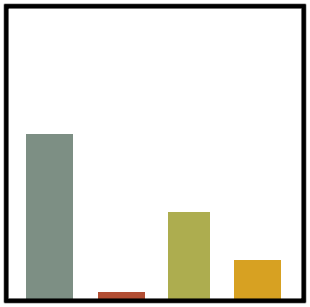


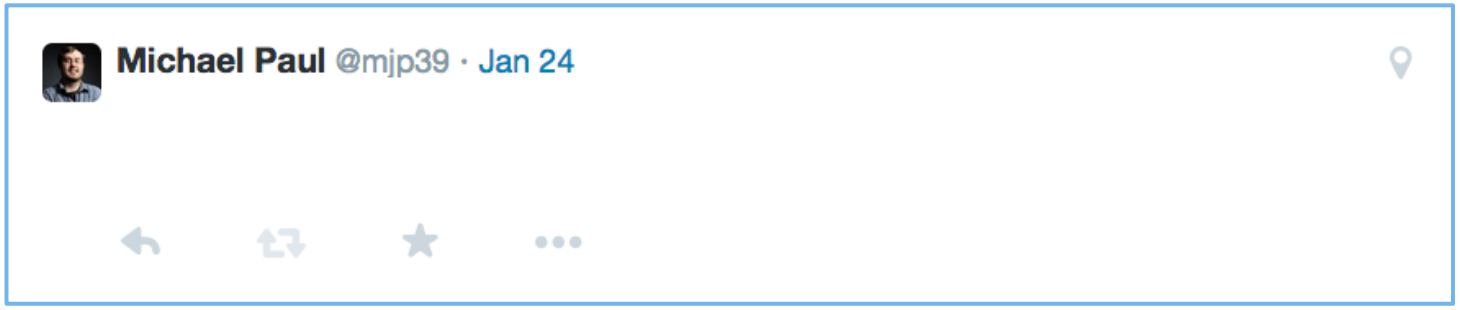




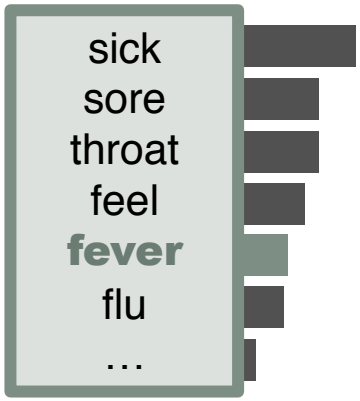
TOPIC MODELING

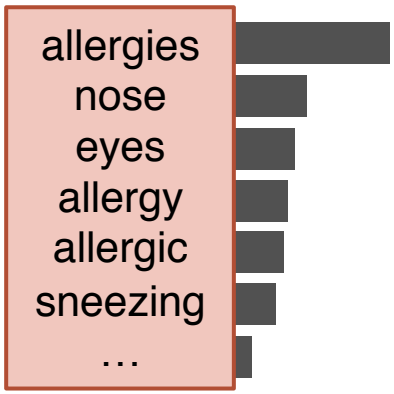


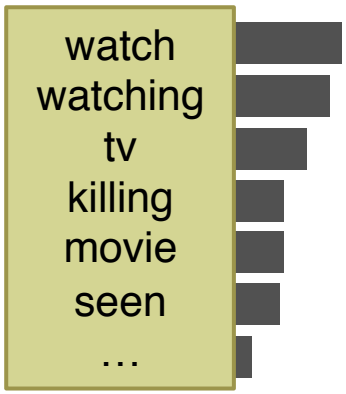


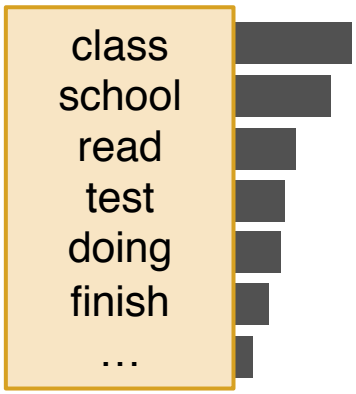


TOPIC MODELING

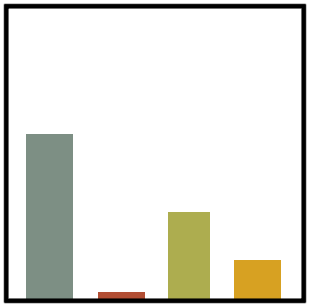


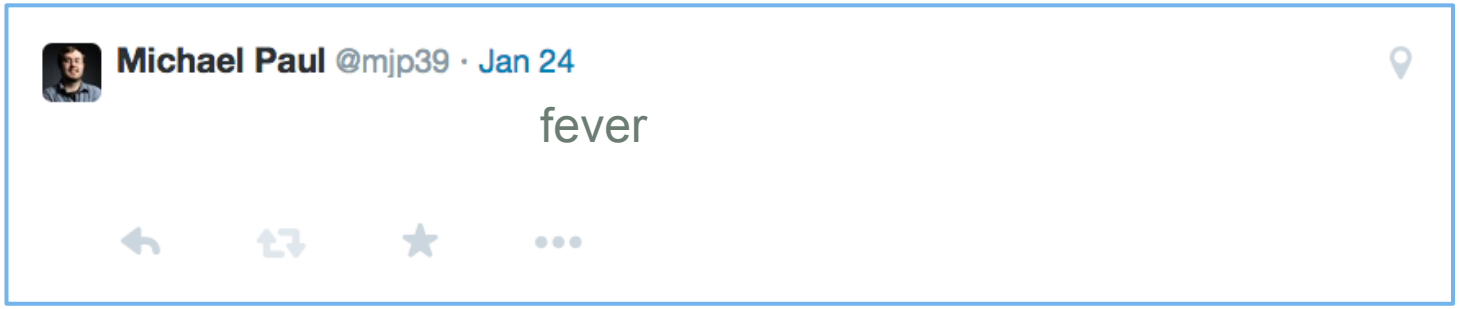




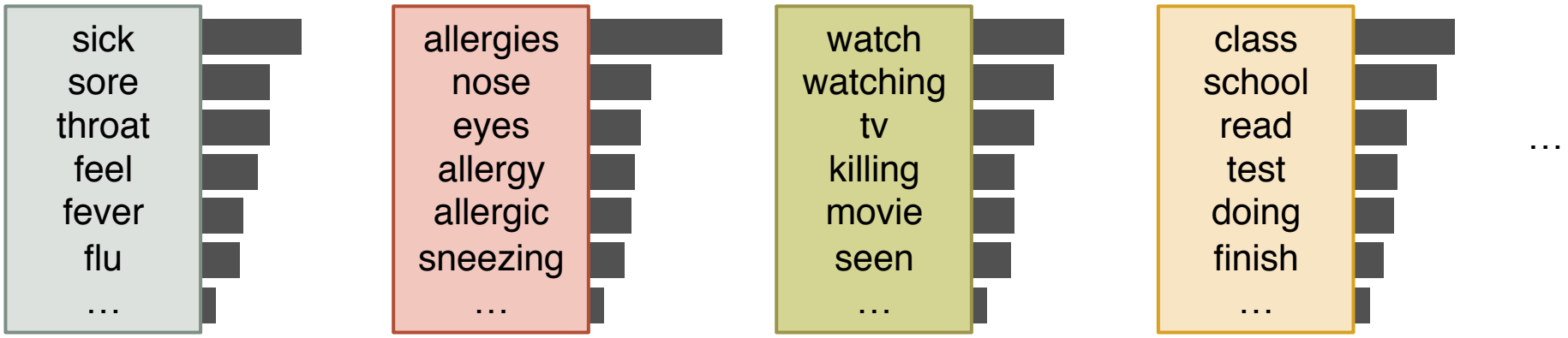


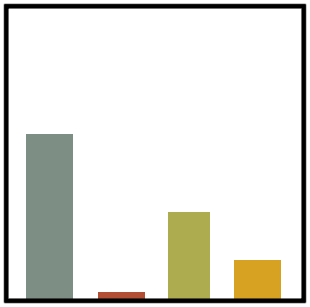
…

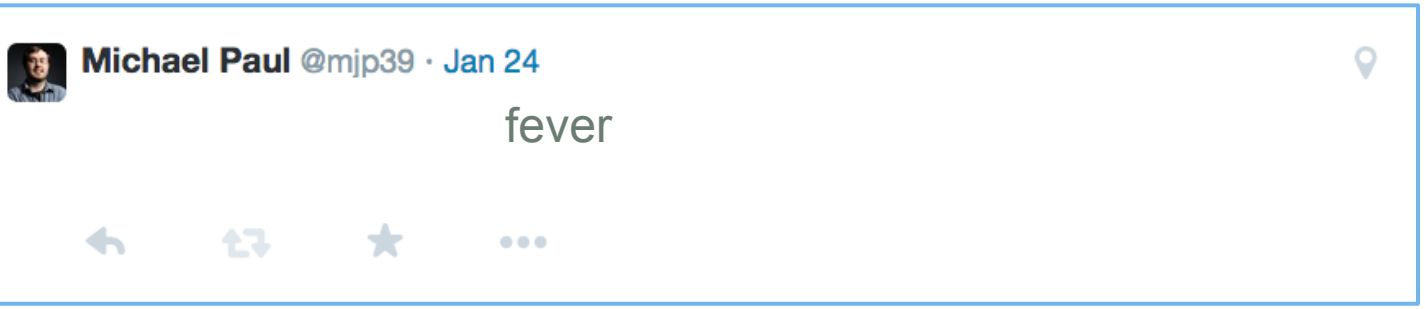




TOPIC MODELING



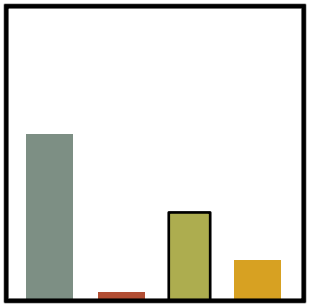




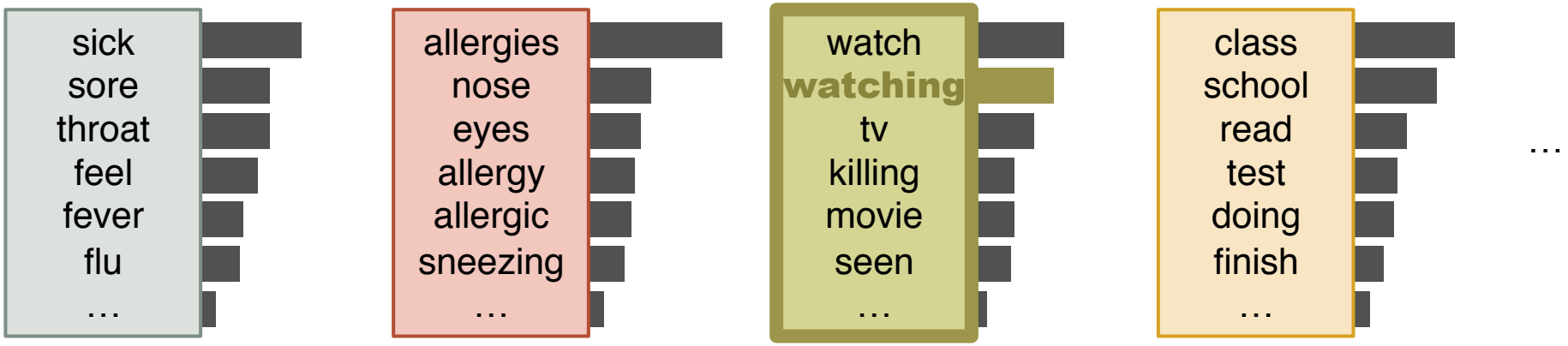
TOPIC MODELING

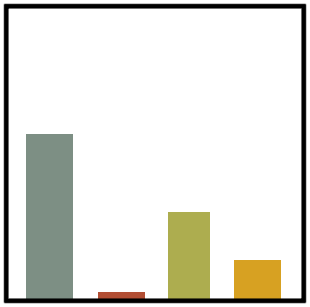
sick	allergies	watch	class
sore	nose	watching	school
throat	eyes	tv	read
feel	allergy	killing	test
fever	allergic	movie	doing
flu	sneezing	seen	finish
…	…	…	…

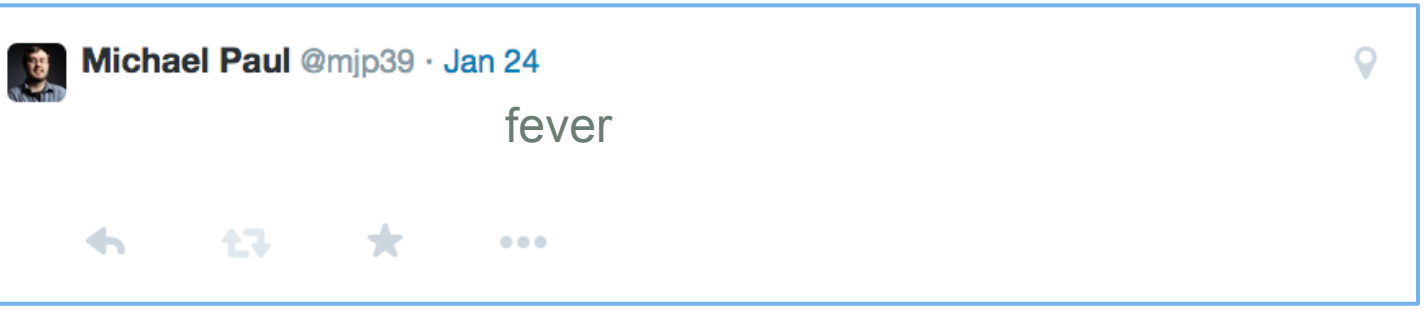
…



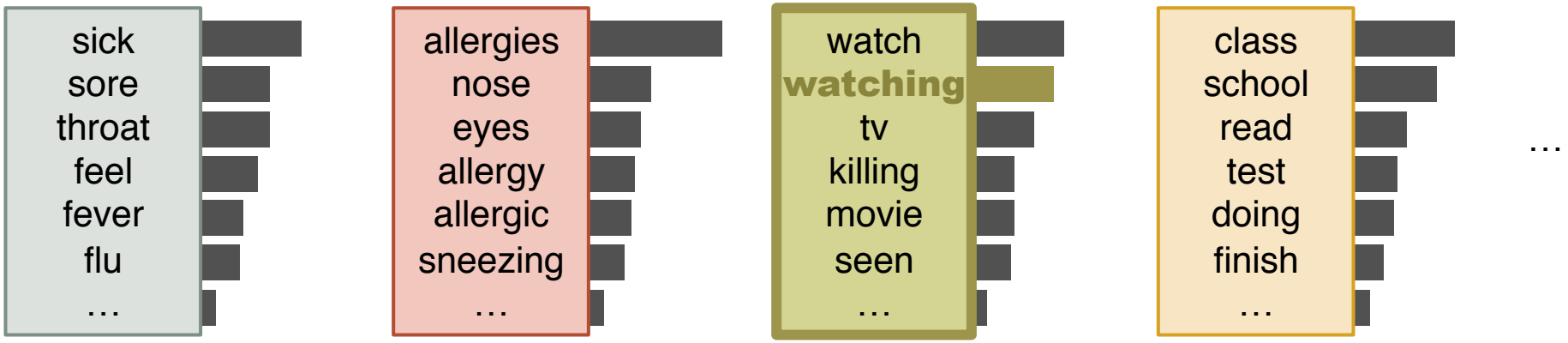
TOPIC MODELING

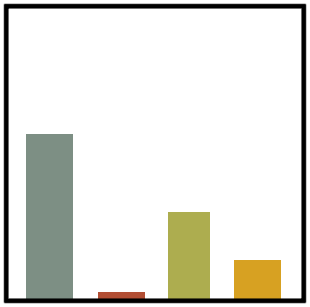


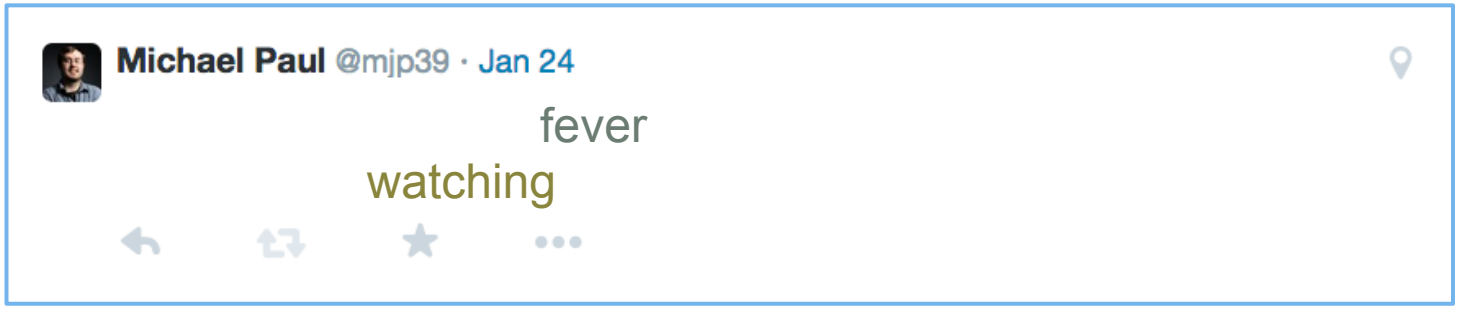




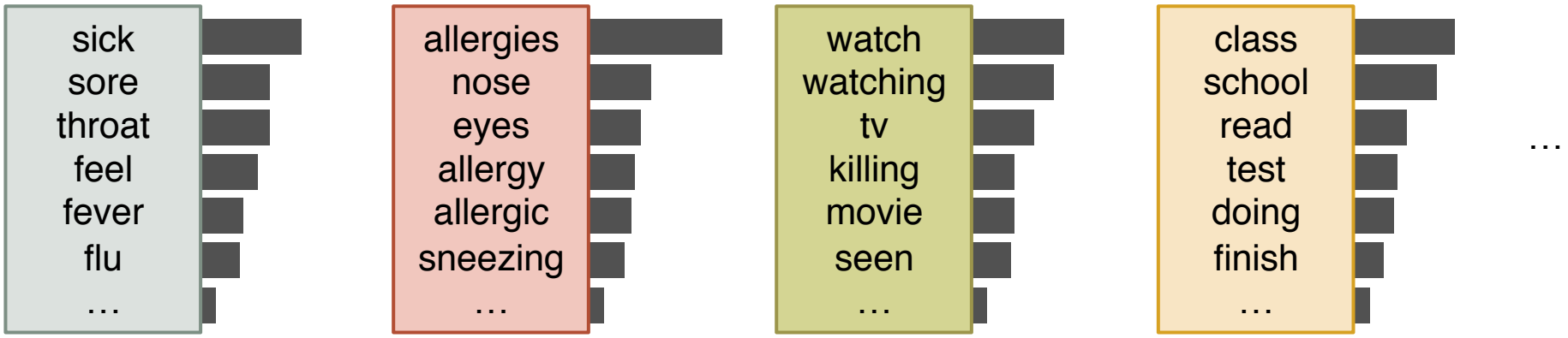
TOPIC MODELING

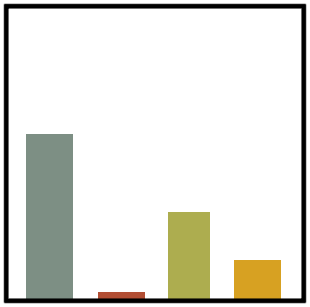


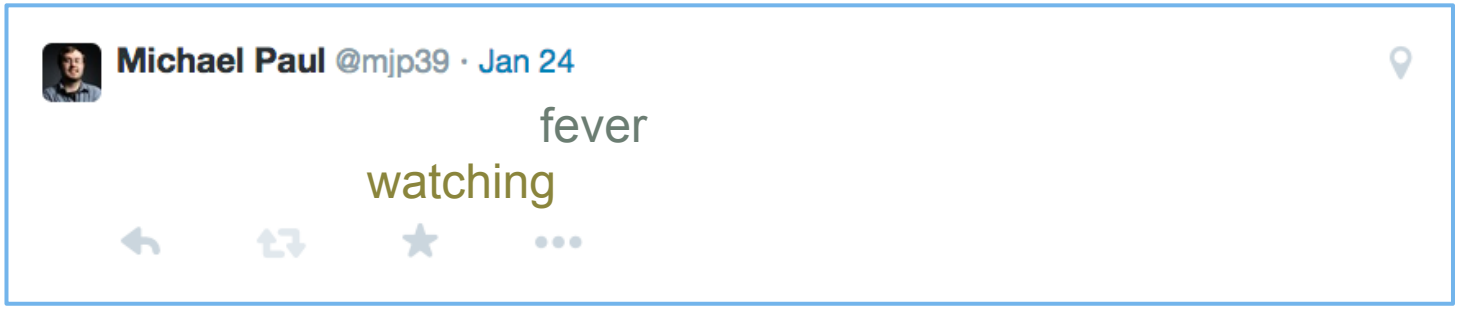




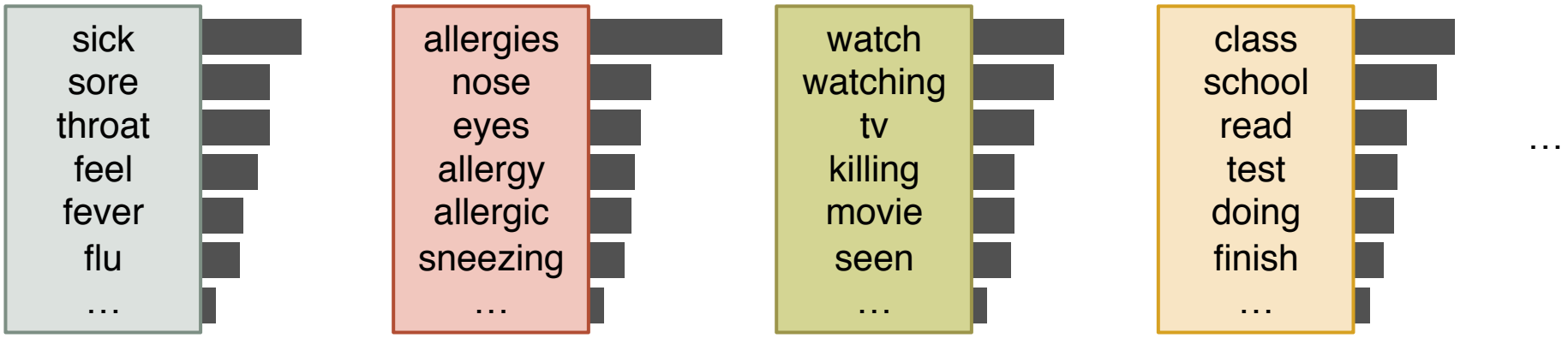
TOPIC MODELING

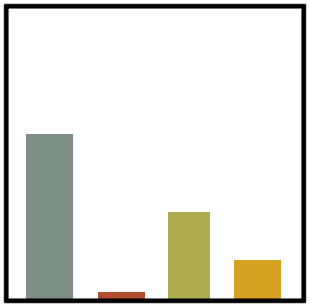


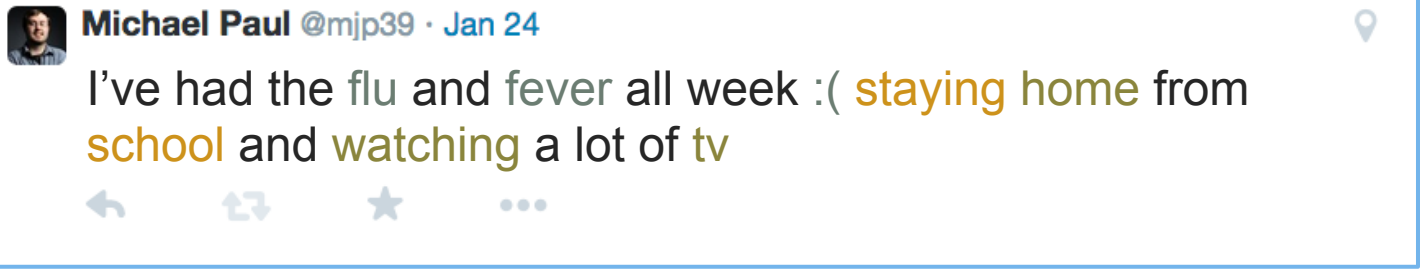




TOPIC MODELING

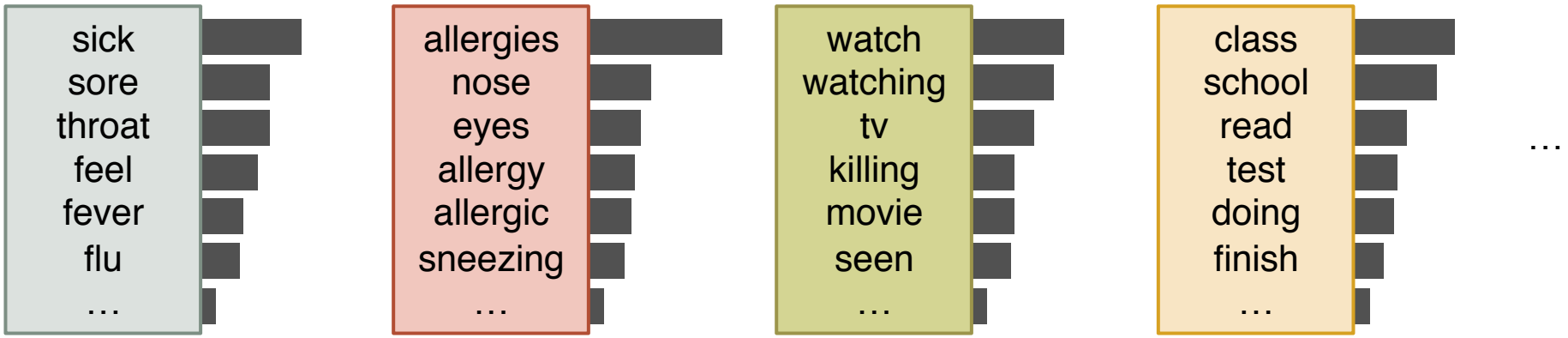


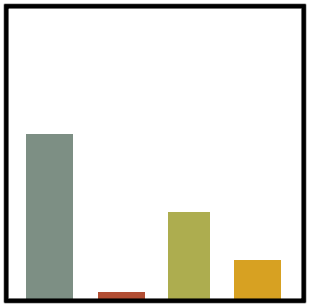




TOPIC MODELING

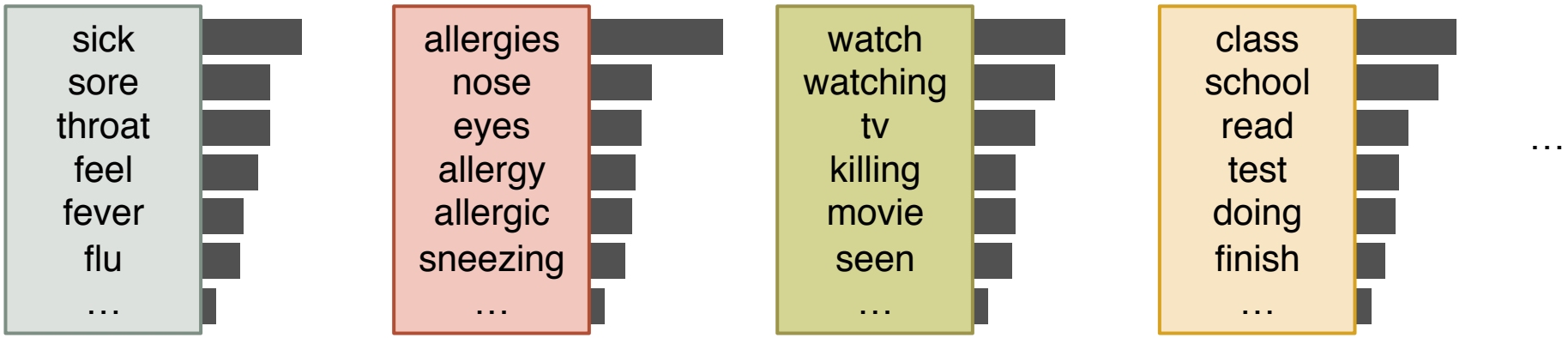
PRIORS

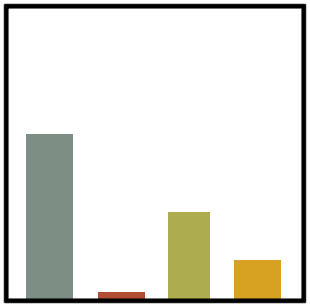




Our imaginary process also needs to generate all these distributions

TOPIC MODELING PRIORS

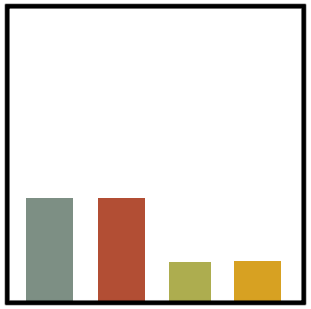




Our imaginary process also needs to generate all these distributions

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

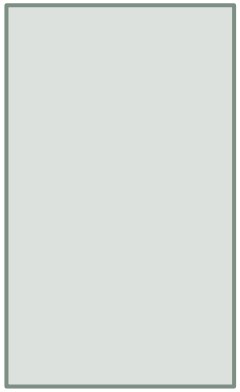
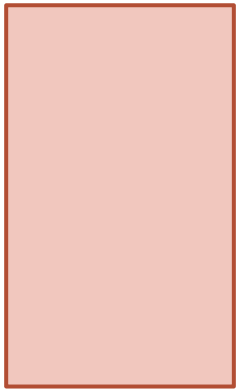
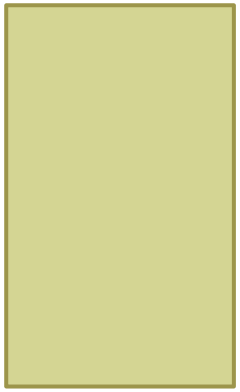
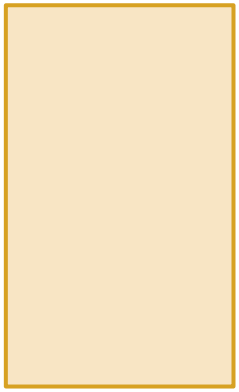
PRIORS

Dirichlet(ρ ×  **)**

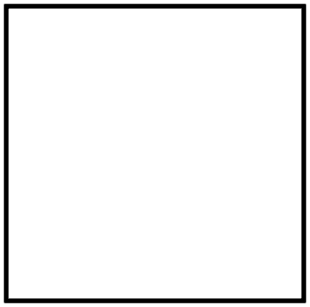
Precision Mean

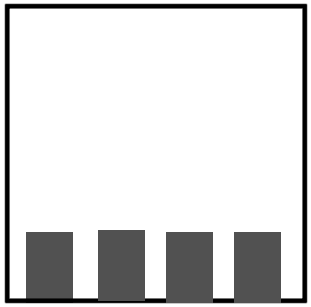
TOPIC MODELING

PRIORS

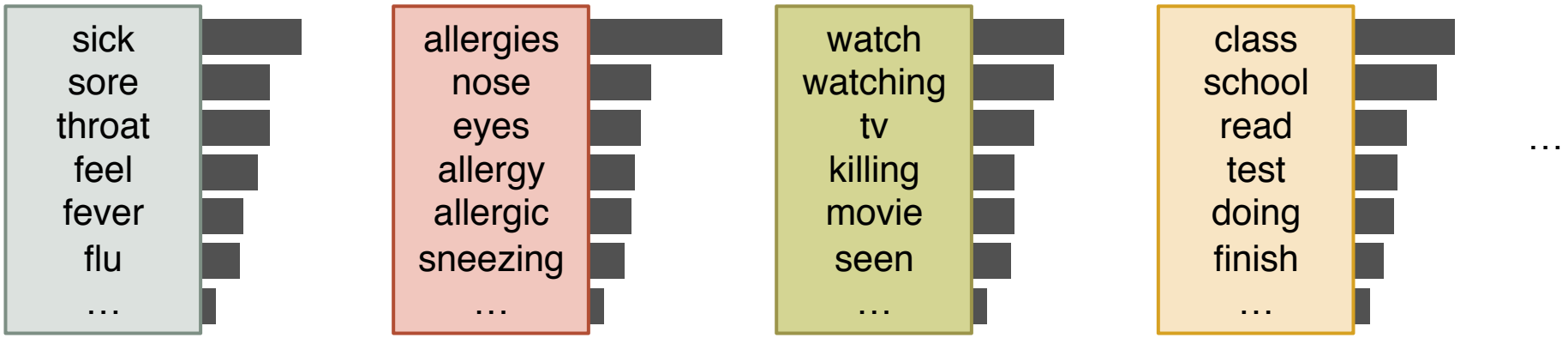
...

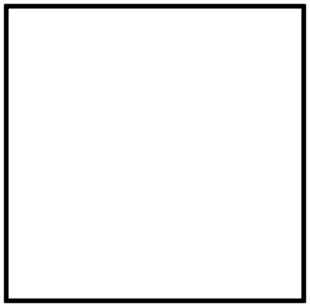


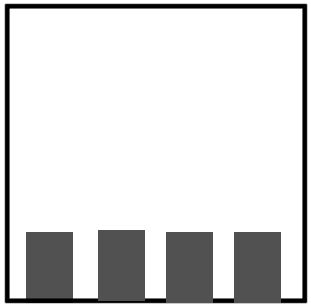
Dirichlet( **)**

TOPIC MODELING

PRIORS

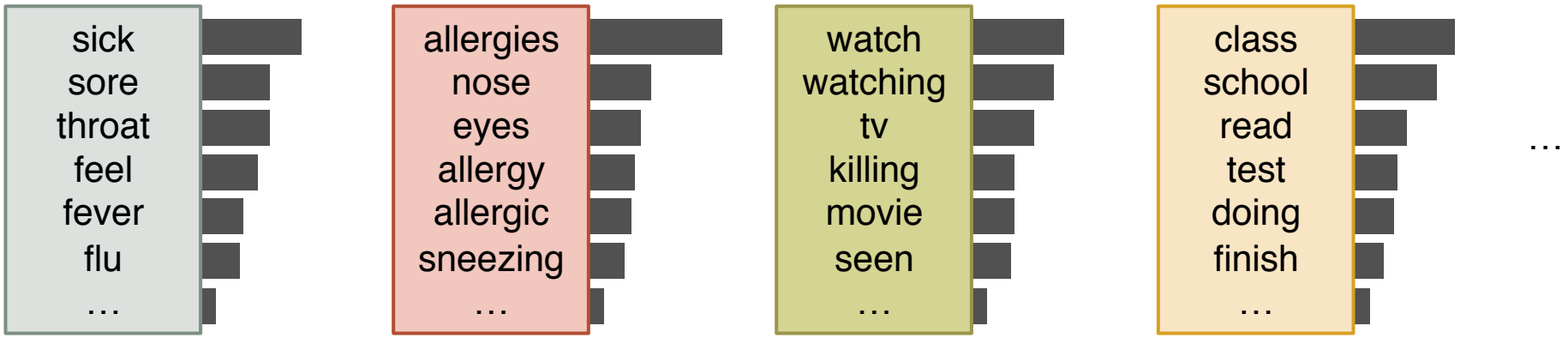


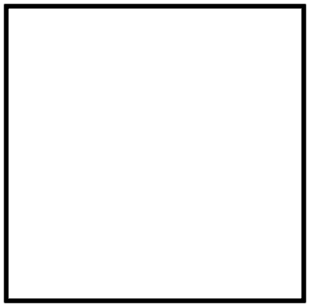


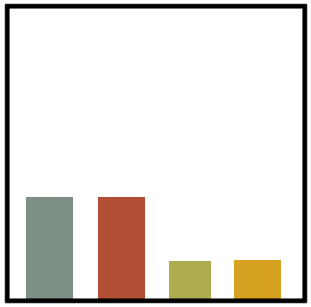
Dirichlet( **)**

TOPIC MODELING

PRIORS

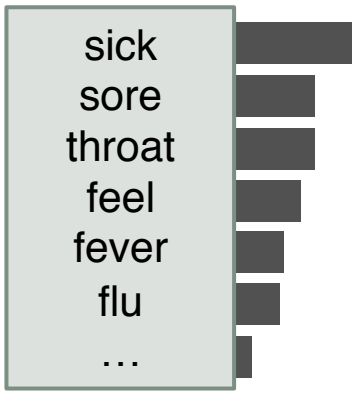


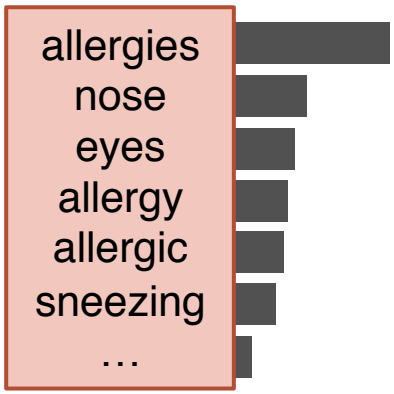


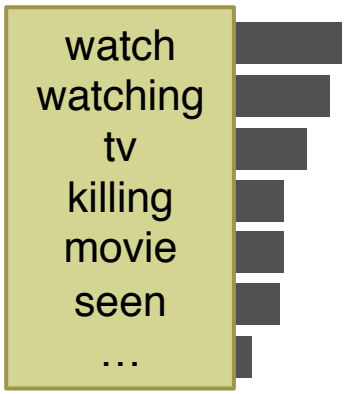
Dirichlet( **)**

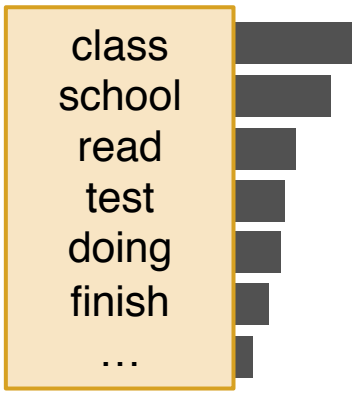
TOPIC MODELING

PRIORS

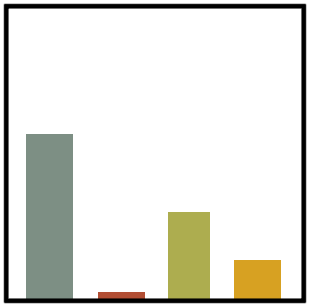




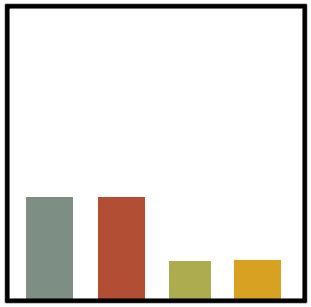




…

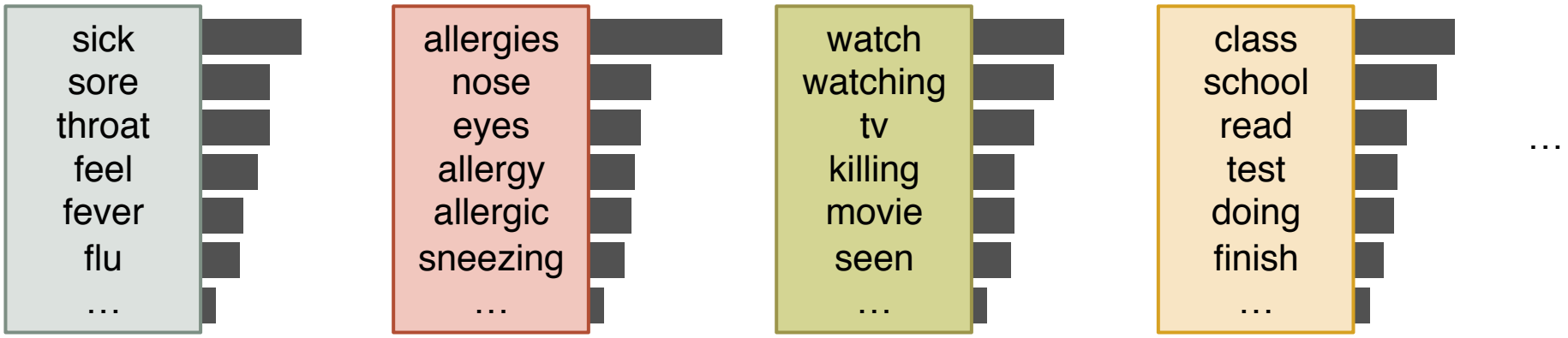


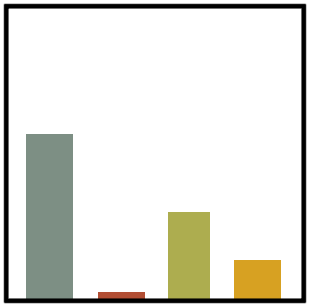
Dirichlet(

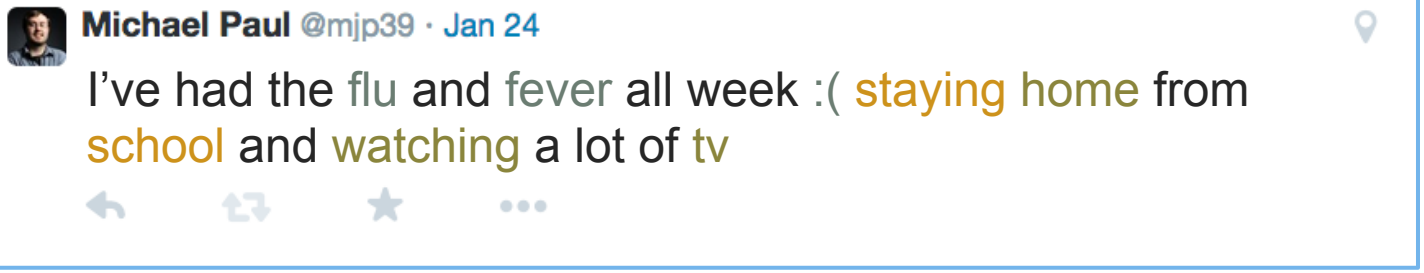


)

TOPIC MODELING







Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan 2003

The topic and word distributions have Dirichlet priors:

$$\phi_t \sim \text{Dirichlet}(\tilde{\phi})$$

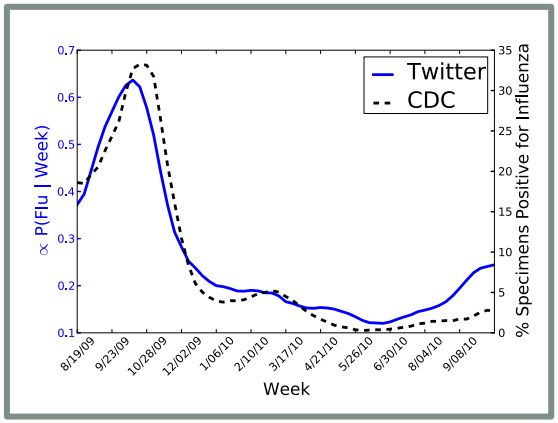
$$\theta_m \sim \text{Dirichlet}(\tilde{\theta})$$

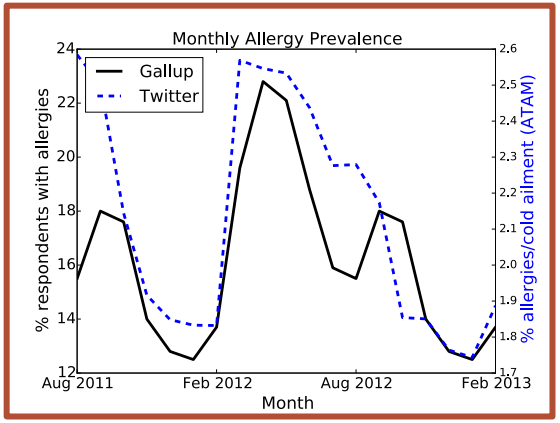
TOPIC MODELING

APPLICATIONS

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

…

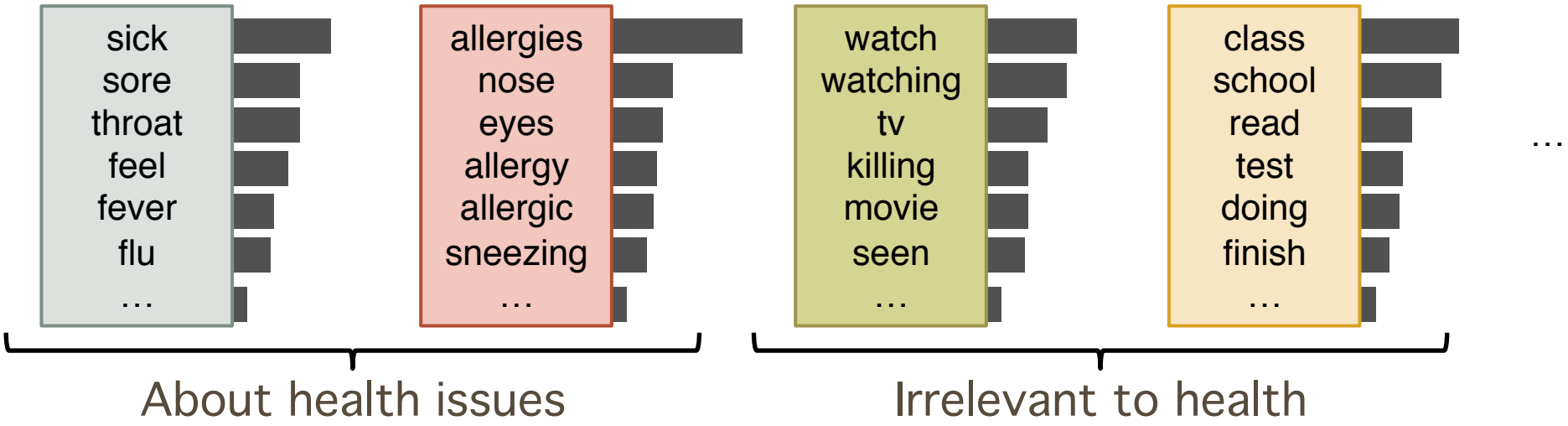




Paul, Dredze (2011) **You are what you tweet: Analyzing Twitter for public health.** *5th International Conference on Weblogs and Social Media (ICWSM).*

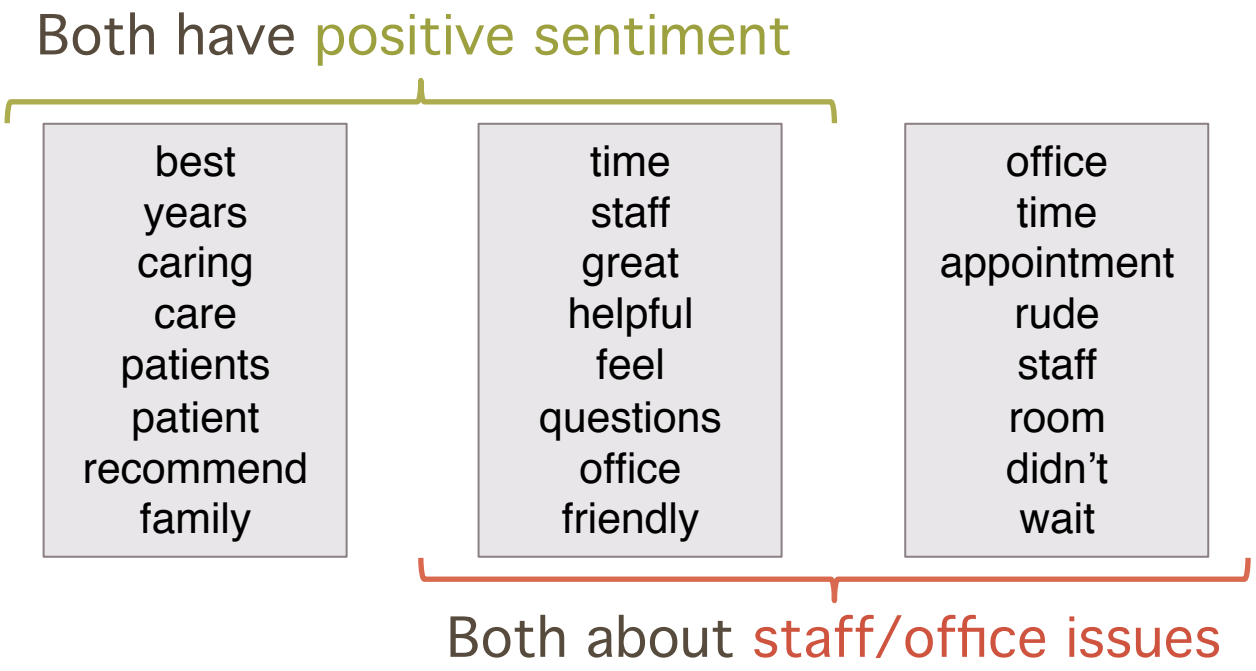
Paul, Dredze (2014) **Discovering health topics in social media using topic models.** *PLOS ONE* 9(8).

TOPIC MODELING ADDING STRUCTURE



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

Topics in online doctor reviews:



TOPIC MODELING ADDING STRUCTURE

Topics in online doctor reviews:

	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

FACTORIAL LDA

A multi-dimensional topic model

Word distributions are grouped into different concepts

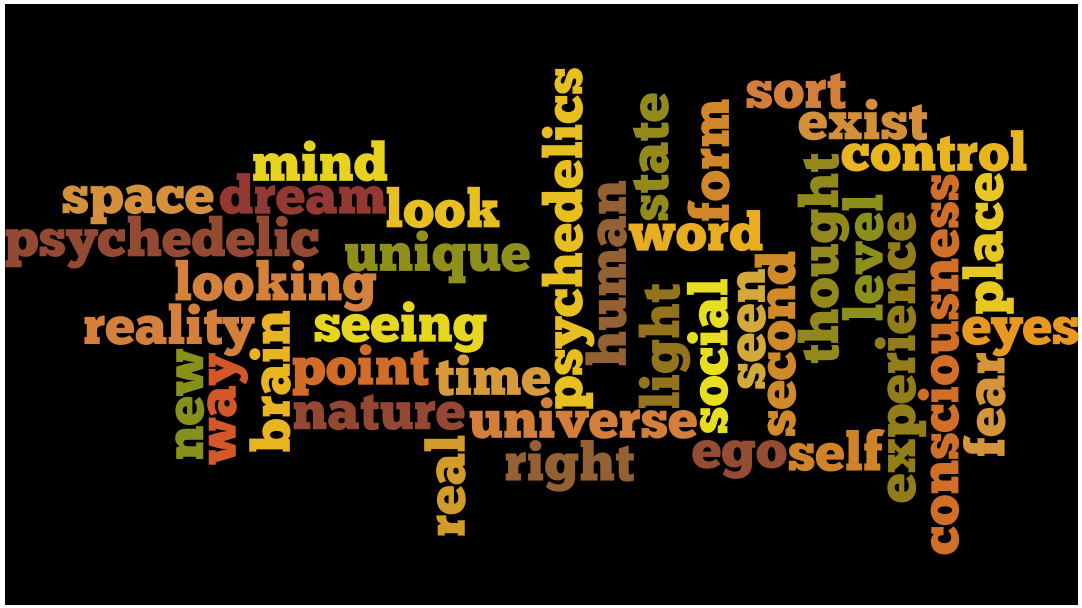
- e.g. sentiment and aspect

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:



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.**

FACTORIAL LDA

DRUG DISCUSSIONS

3-dimensional model:

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

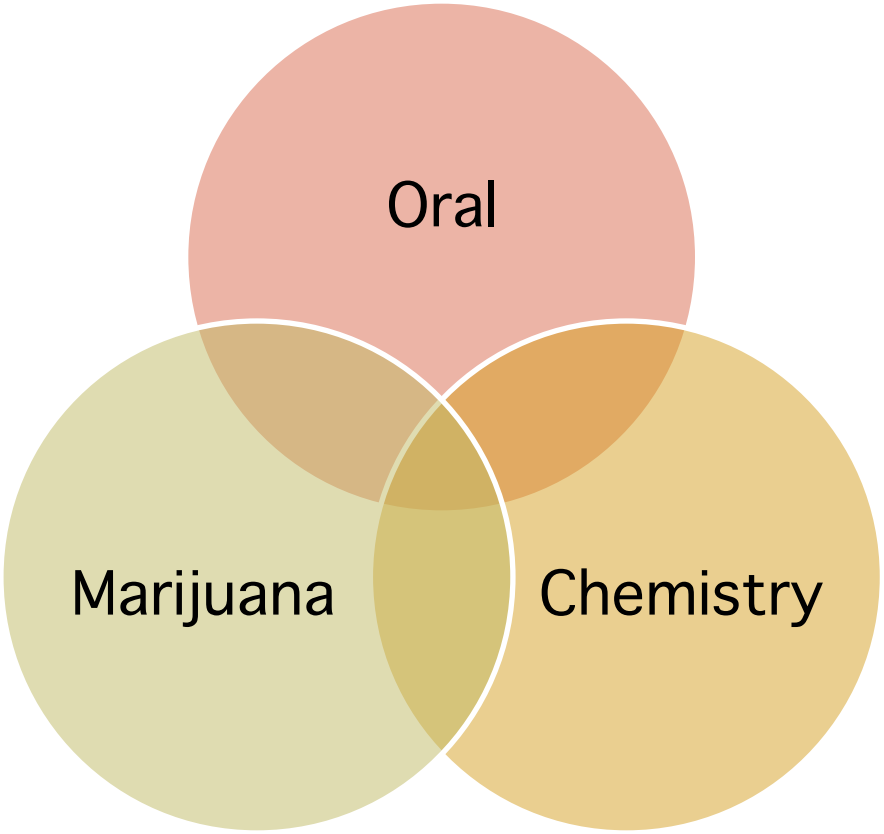
Drugs-Forum

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

FACTORIAL LDA

DRUG DISCUSSIONS

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



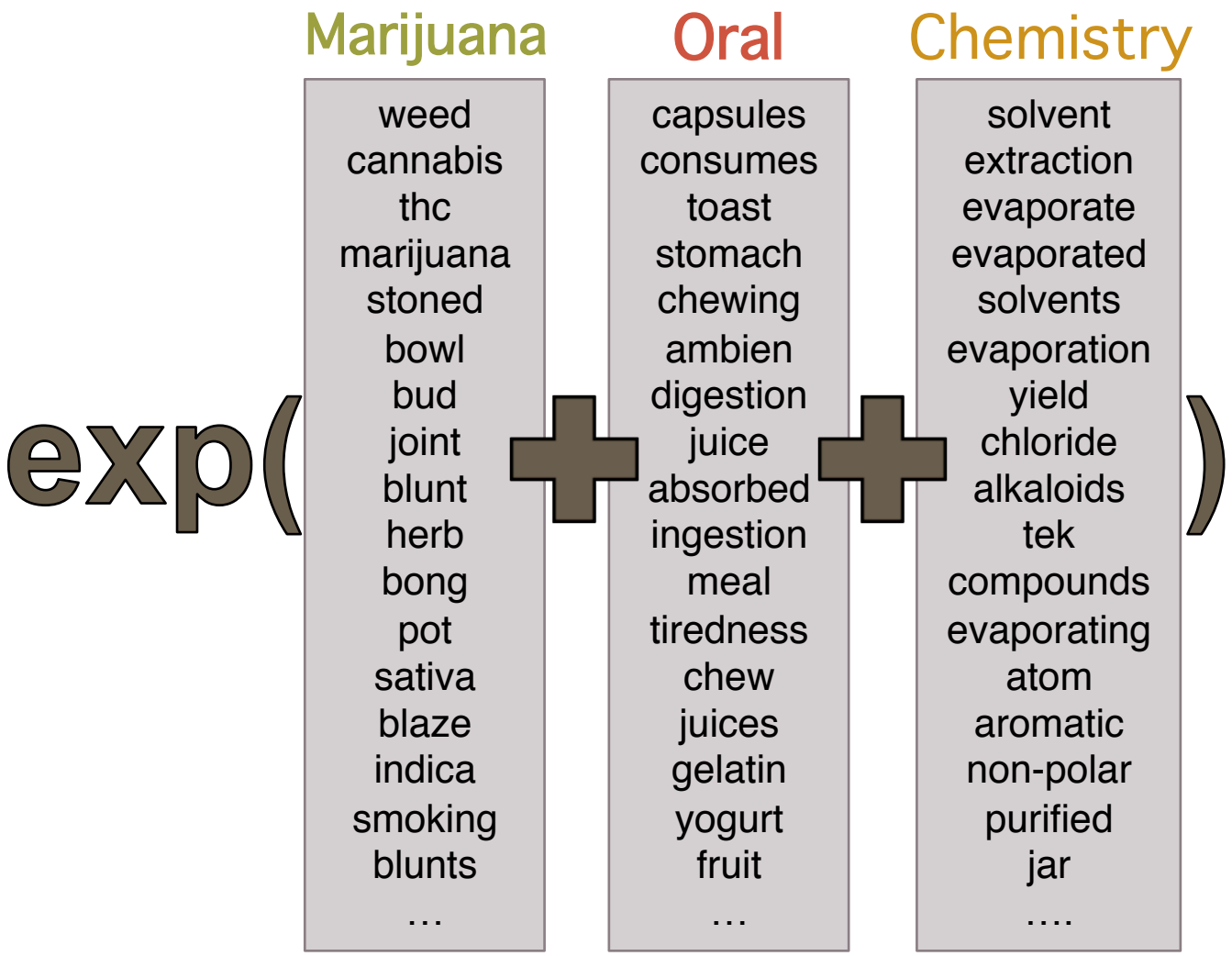
FACTORIAL LDA

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
…	…	….

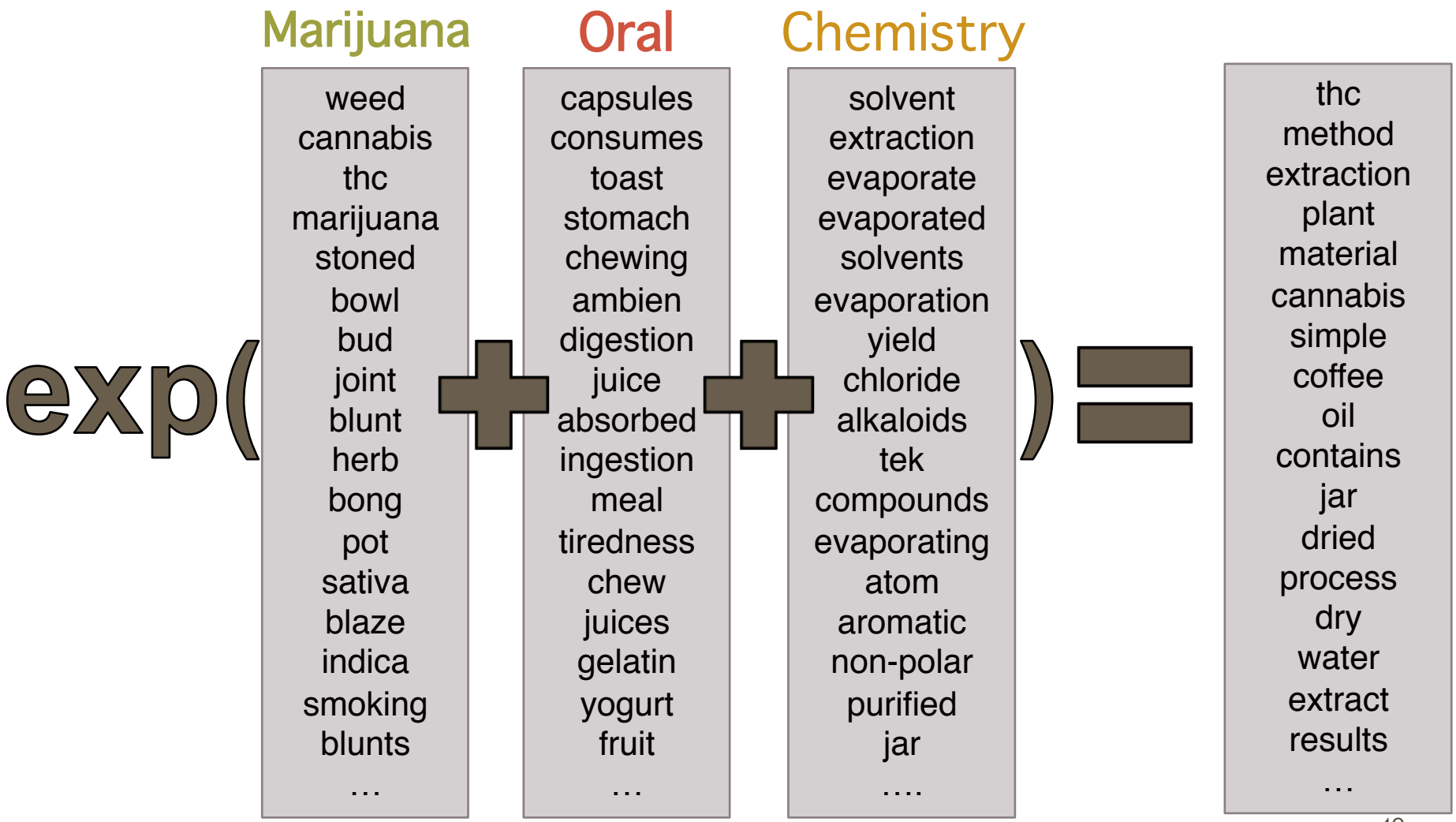
FACTORIAL LDA

DRUG DISCUSSIONS



FACTORIAL LDA

DRUG DISCUSSIONS



FACTORIAL LDA

DRUG DISCUSSIONS

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

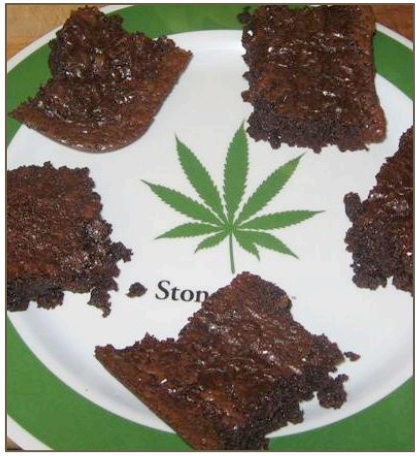
~ Dirichlet(

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

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 $\boldsymbol{t}$:

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

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

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 $\boldsymbol{t}$:

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

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

Document priors:

$$\tilde{\theta}_{m\vec{t}} = \exp(\sum_{k=1}^{K} \alpha_{m\vec{t}_k}^{(k)})$$

$$\theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$$

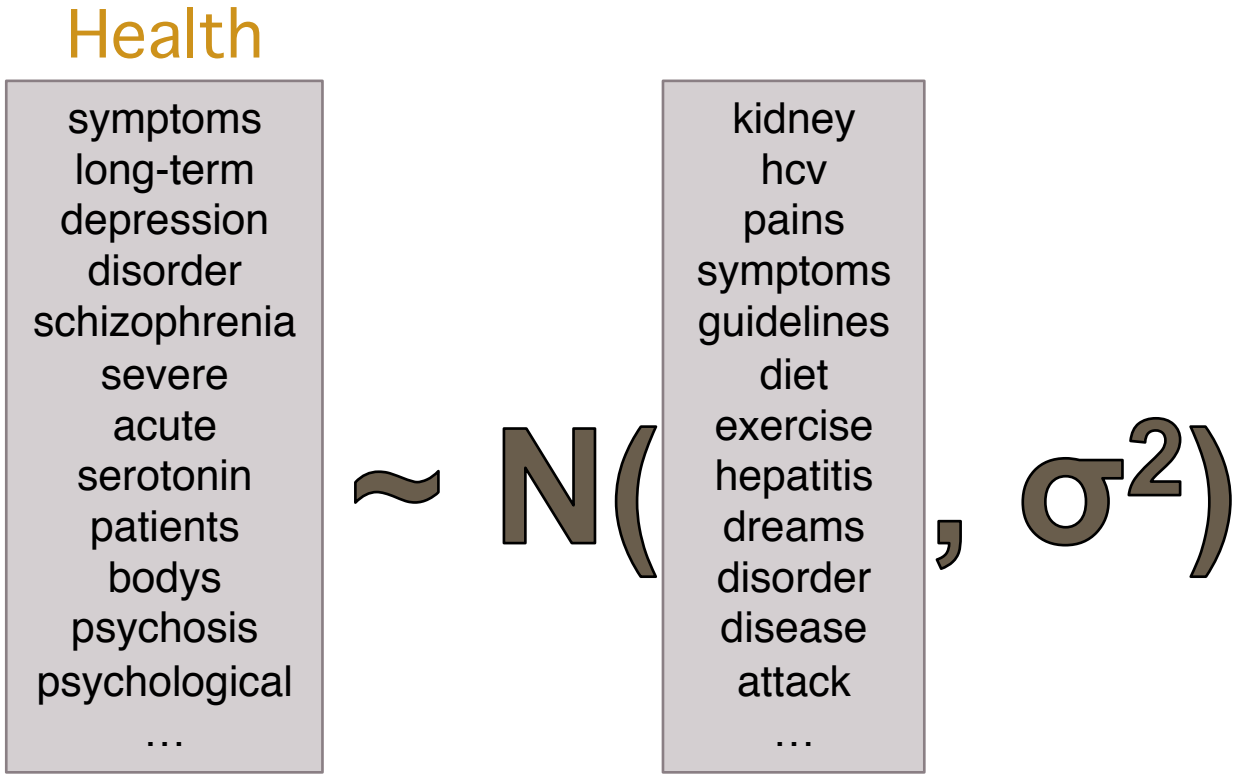
FACTORIAL LDA

SEMI-SUPERVISION

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
…	…	….

Where did these vectors come from?

Weights learned from a supervised model are then used to create a **Gaussian prior** over the FLDA weights:



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?"

(Mephedrone
Oral
Usage)

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.

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.

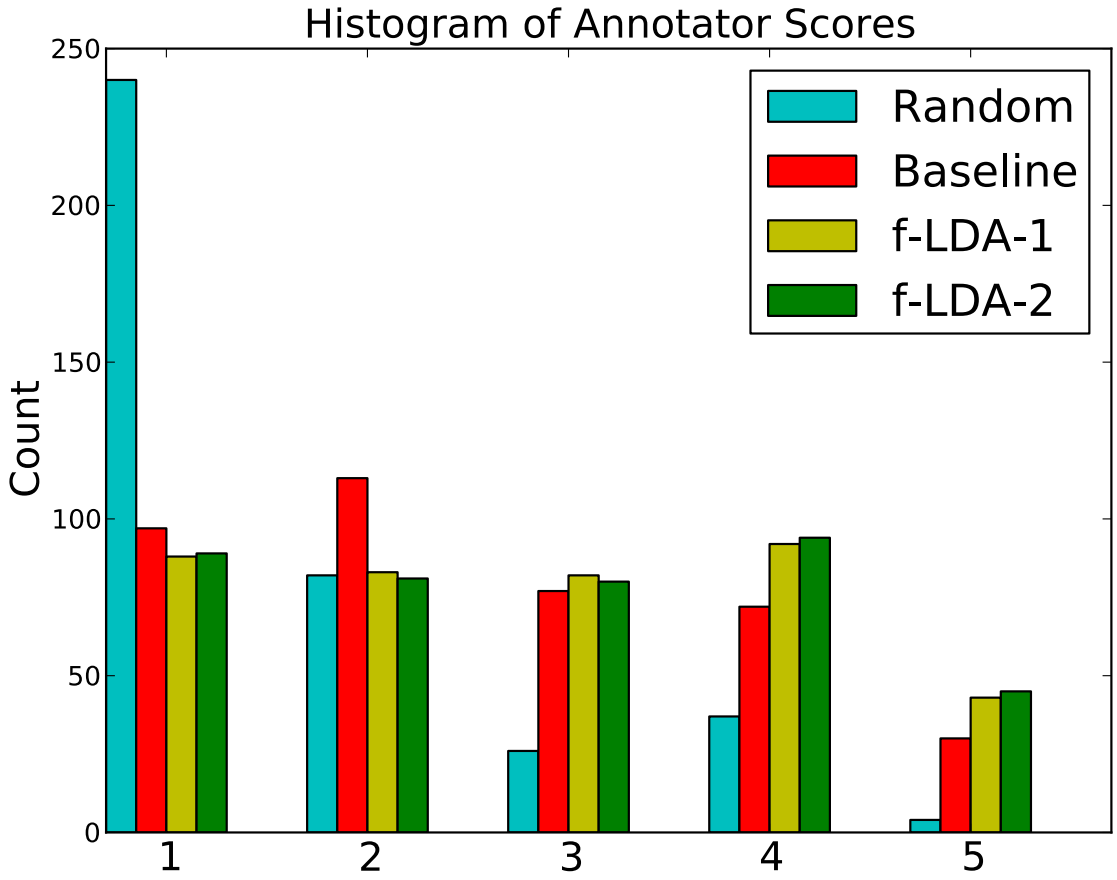
Reference text:

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

FACTORIAL LDA

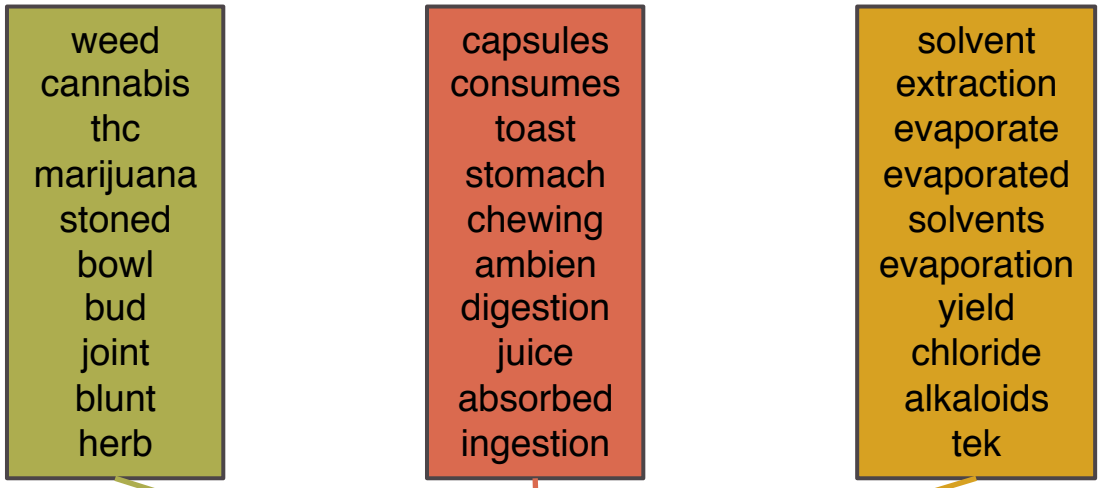
DRUG DISCUSSIONS

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

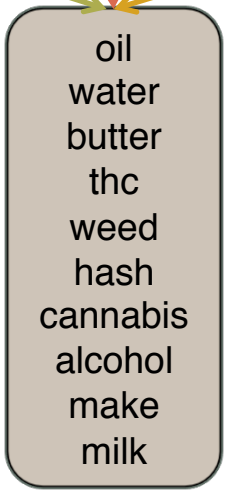


FACTORIAL LDA

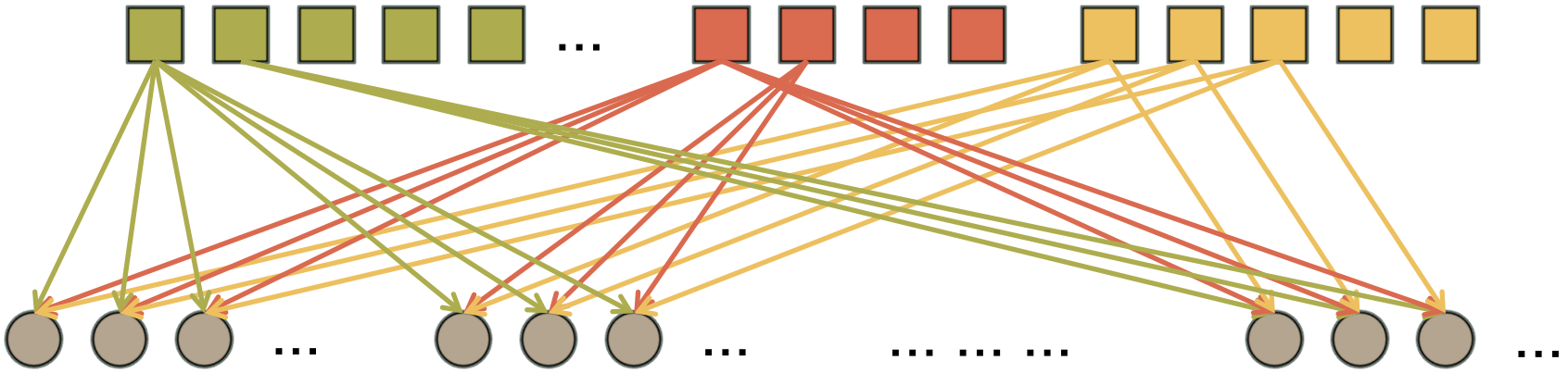
Components:



Topics:

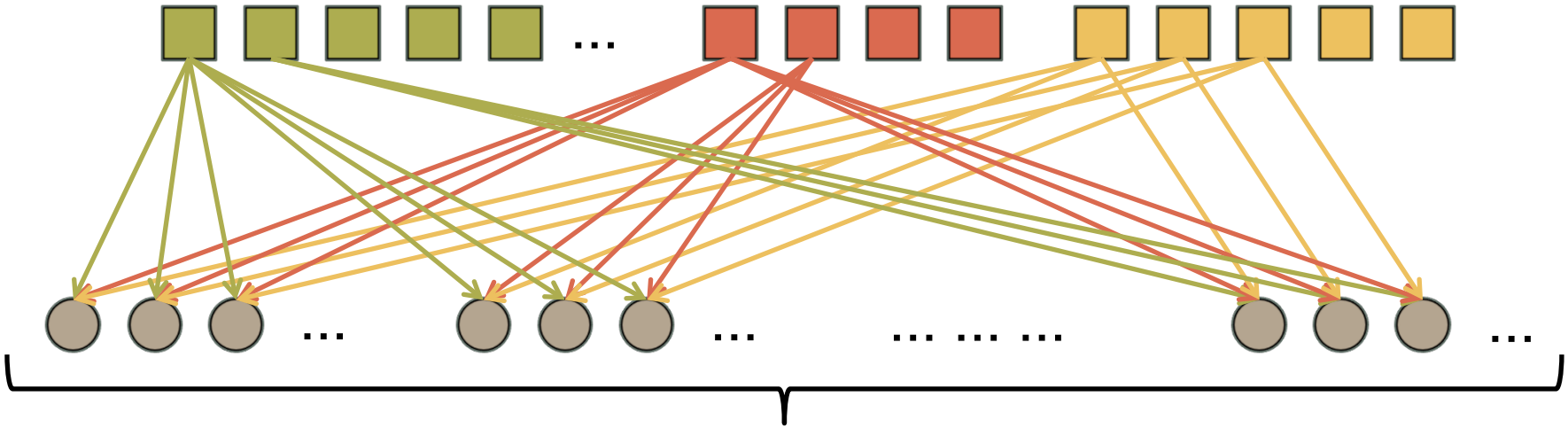


FACTORIAL LDA



FACTORIAL LDA

SPARSITY

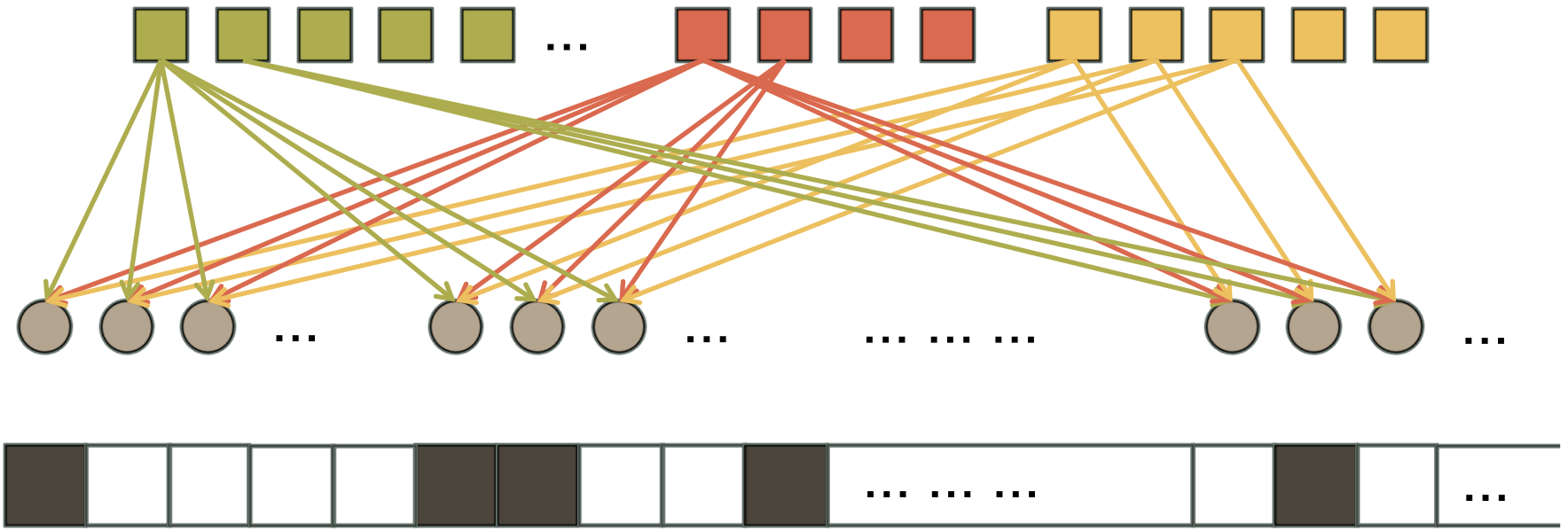


This Cartesian product can be huge!

- And not all triples make sense...

FACTORIAL LDA

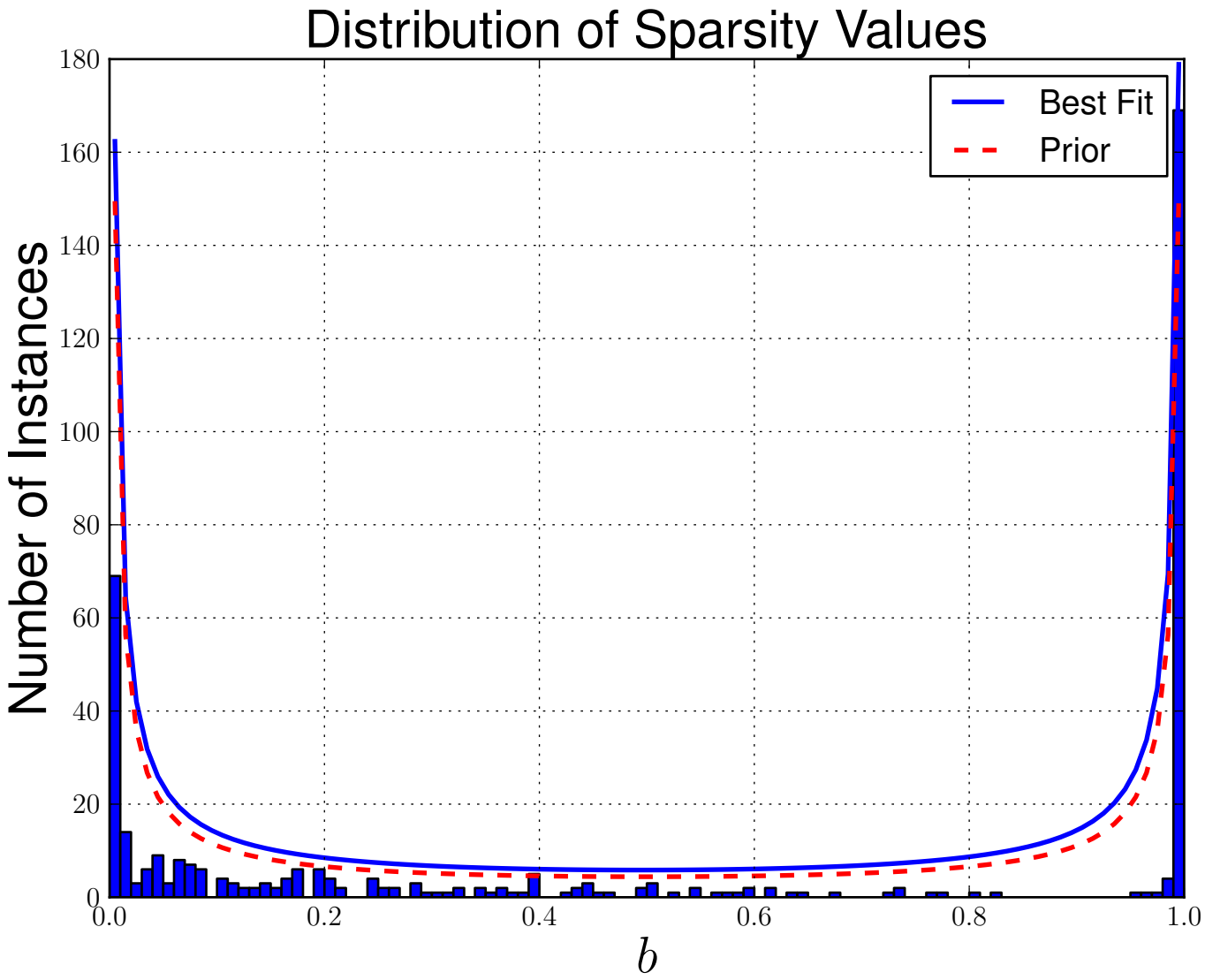
SPARSITY



- Proposed solution: learn a sparsity pattern

$$\tilde{\theta}_{m\vec{t}} = b_{\vec{t}} \exp\left(\sum_{k=1}^{K} \alpha^{(k)}_{m\vec{t}_k}\right) \qquad b_{\vec{t}} \in (0,1);\ b_{\vec{t}} \sim \text{Beta}(\rho < 1)$$

SPARSITY


Distribution of Sparsity Values
Best Fit
Prior
Number of Instances
0
20
40
60
80
100
120
140
160
180
0.0
0.2
0.4
0.6
0.8
1.0
b

FACTORIAL LDA

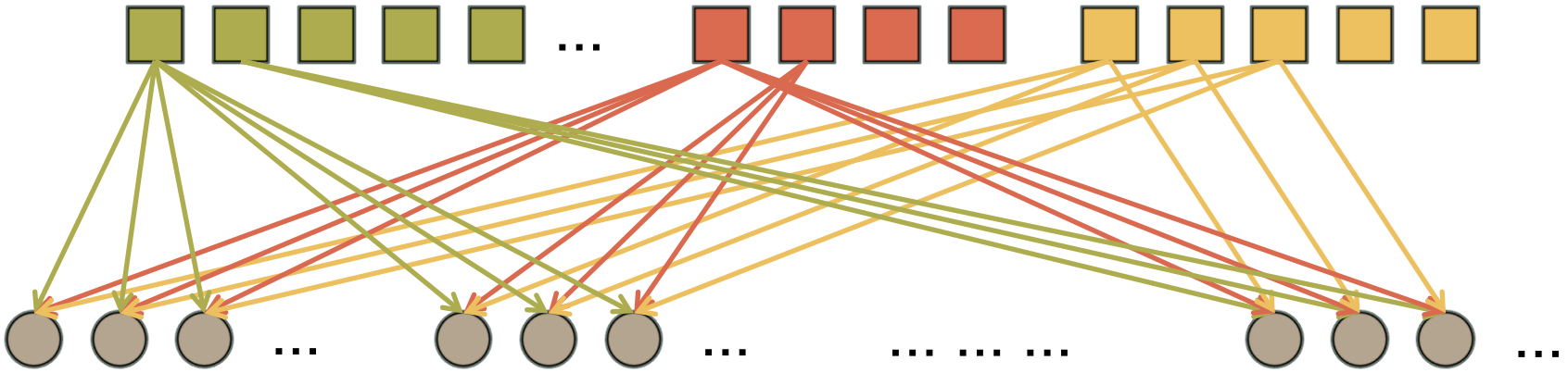
SPARSITY

"Topic"				"Approach"		"Focus"	
"SPEECH"	"I.R."	"M.T."	. . .	"EMPIRICAL"	"THEORETICAL"	"METHODS"	"APPLICATIONS"
speech	document	translation	. . .	task	theory	word	user
spoken	retrieval	machine	. . .	tasks	description	algorithm	research
recognition	documents	source	. . .	performance	formal	method	project
state	question	mt	. . .	improve	forms	accuracy	technology
vocabulary	web	parallel	. . .	accuracy	treatment	best	processing
recognizer	answering	french	. . .	learning	linguistics	sentence	science
utterances	query	bilingual	. . .	demonstrate	syntax	statistical	natural
synthesis	answer	transfer	. . .	using	ed	previously	development

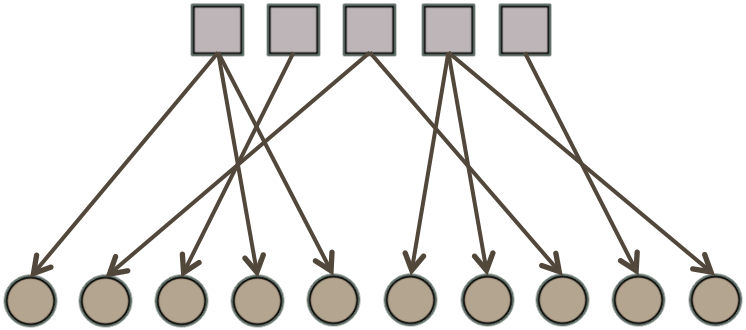
Topic		SPEECH		DATA		MODELING		GRAMMAR	
Focus		METHODS	APPL.	METHODS	APPL.	METHODS	APPL.	METHODS	APPL.
Approach	EMPIRICAL	**(b=0.20)**	**(b=1.00)** dialogue spoken speech dialogues understanding task recognition	**(b=1.00)** corpus data training model tagging annotated test	**(b=1.00)** data corpus annotation annotated corpora collection xml	**(b=1.00)** models model approach shown error errors statistical	**(b=0.50)**	**(b=1.00)** parsing parser syntactic tree parse dependency treebank	**(b=0.57)** grammar parsing based robust component processing linguistic
	THEORETICAL	**(b=0.99)** speech words recognition prosodic written phonological spoken	**(b=0.00)**	**(b=0.07)**	**(b=0.02)**	**(b=1.00)** rules rule model shown models right left	**(b=0.01)**	**(b=1.00)** grammar parsing grammars structures paper formalism based	**(b=1.00)** grammar grammars formalism parsing based efficient unification

TOPIC STRUCTURES

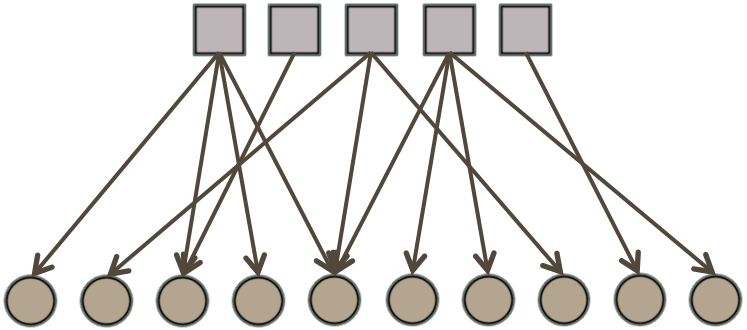
FACTORIZATION



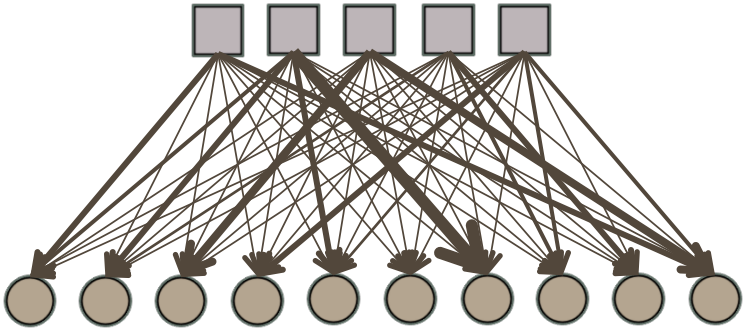
TOPIC STRUCTURES



Tree



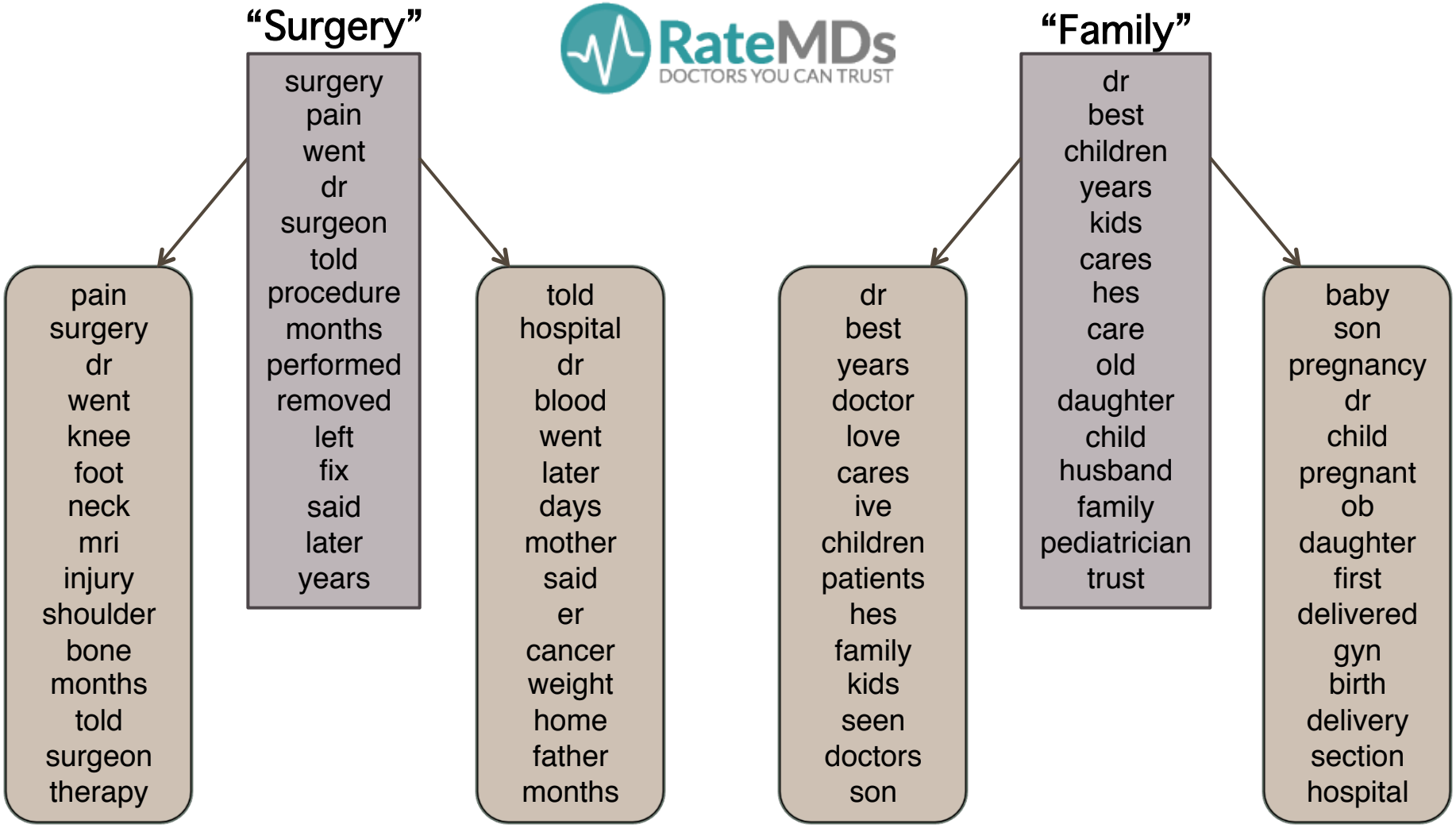
Directed acyclic graph (DAG)



Weighted DAG

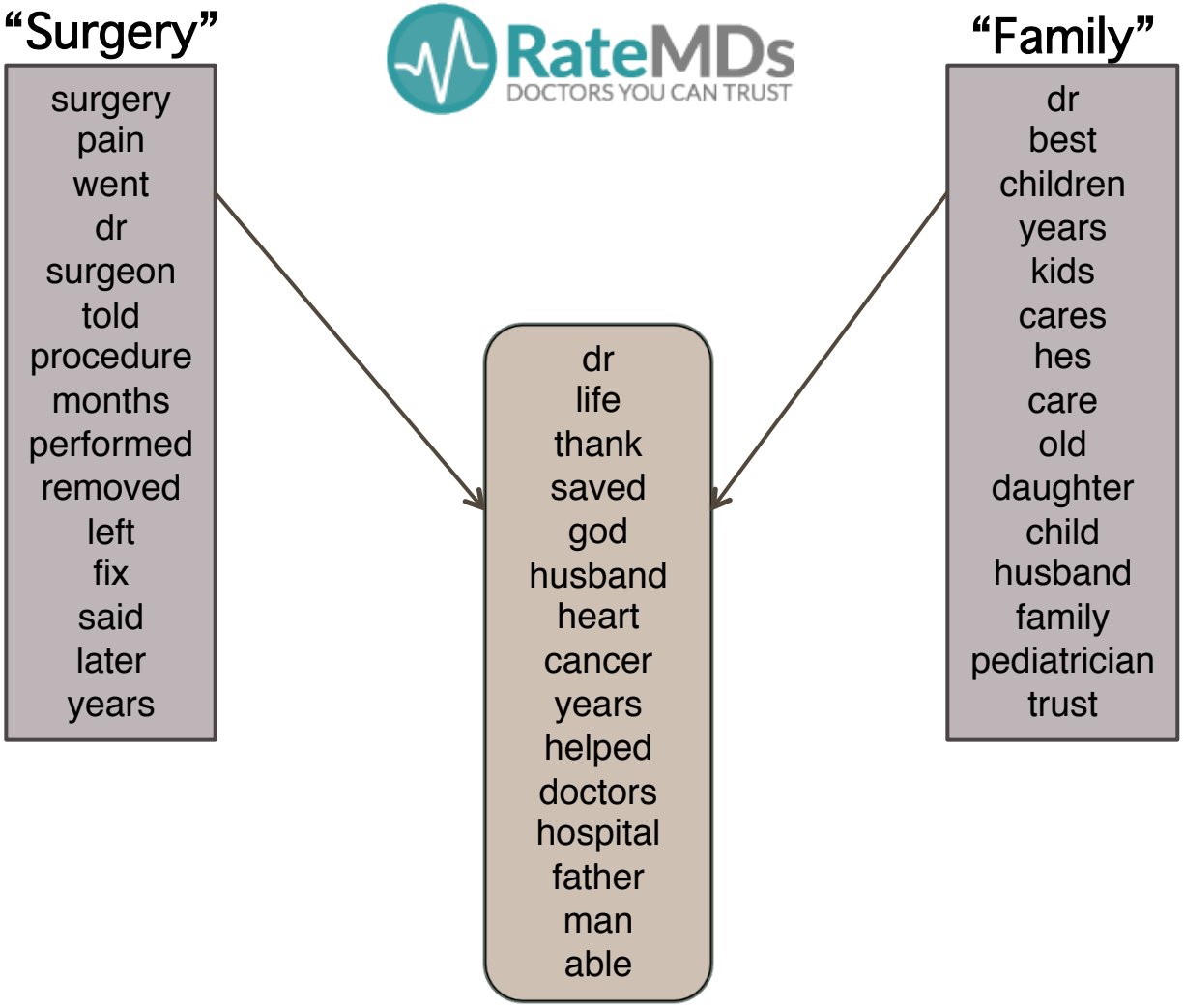
TOPIC STRUCTURES

TREE



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.

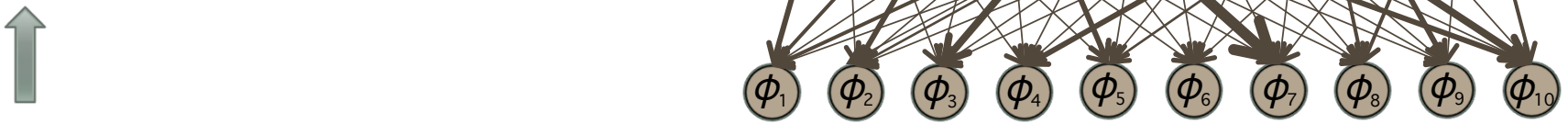
SPRITE

DEFINITION

The priors over distributions are weighted combinations of components:

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

$$\phi_t \sim \text{Dirichlet}(\tilde{\phi}_t)$$



distribution over words in *i*th topic

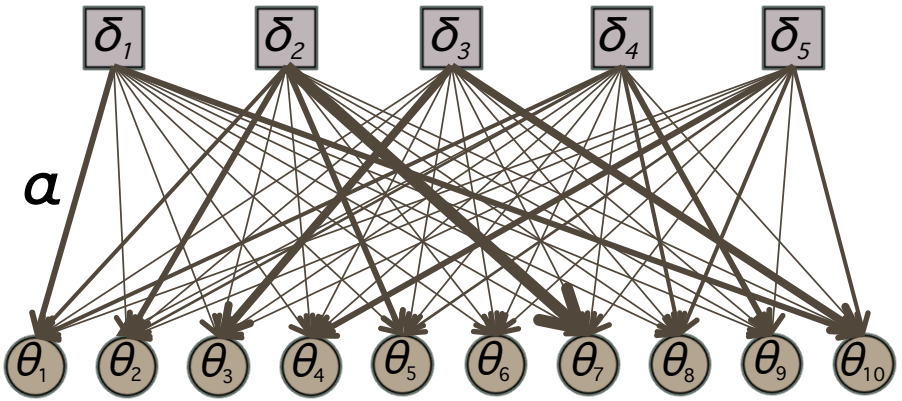
SPRITE

DEFINITION

The priors over distributions are weighted combinations of components:

$$\tilde{\theta}_{mt} = \exp(\sum_{c=1}^{C(\theta)} \alpha_{mc}\delta_{ct})$$

$$\theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$$

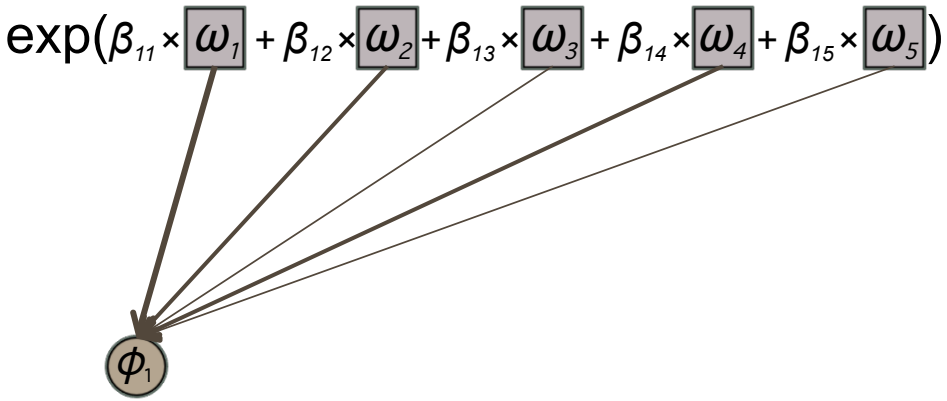


distribution over topics in *m*th document

SPRITE

DEFINITION

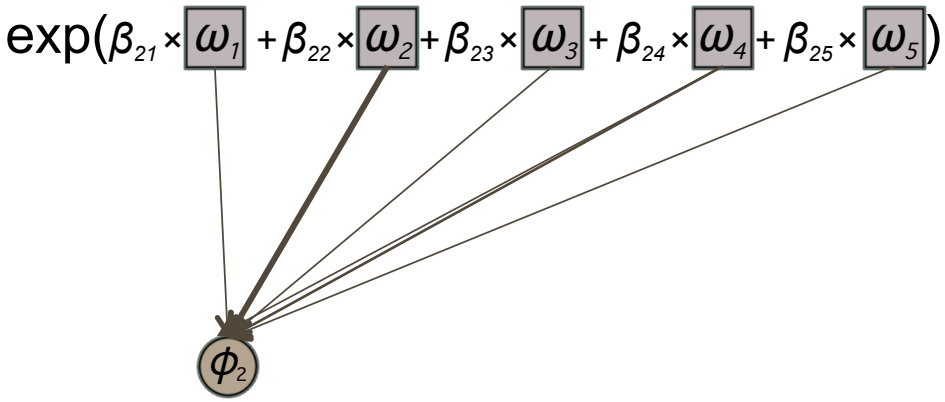
The priors over distributions are weighted combinations of components:



SPRITE

DEFINITION

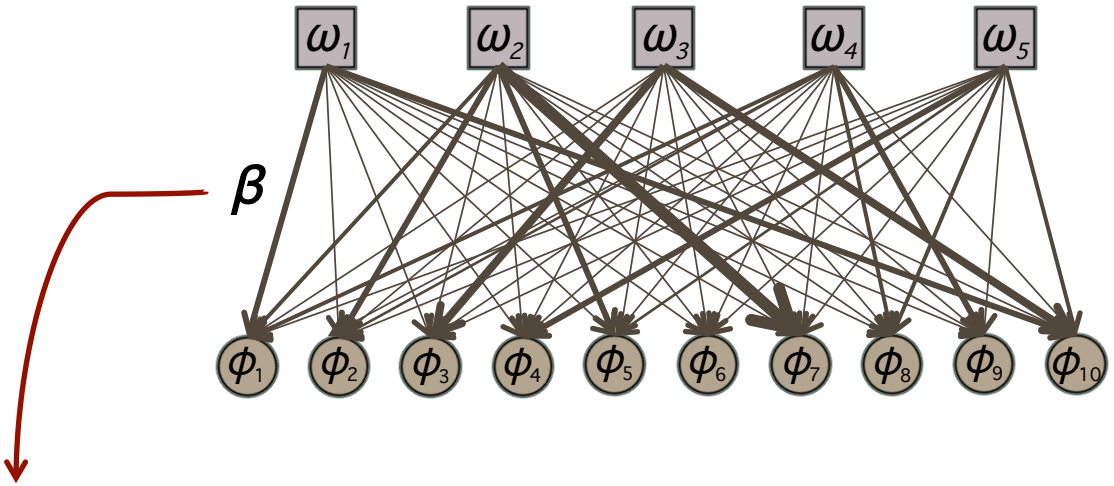
The priors over distributions are weighted combinations of components:



SPRITE

DEFINITION

The priors over distributions are weighted combinations of components:

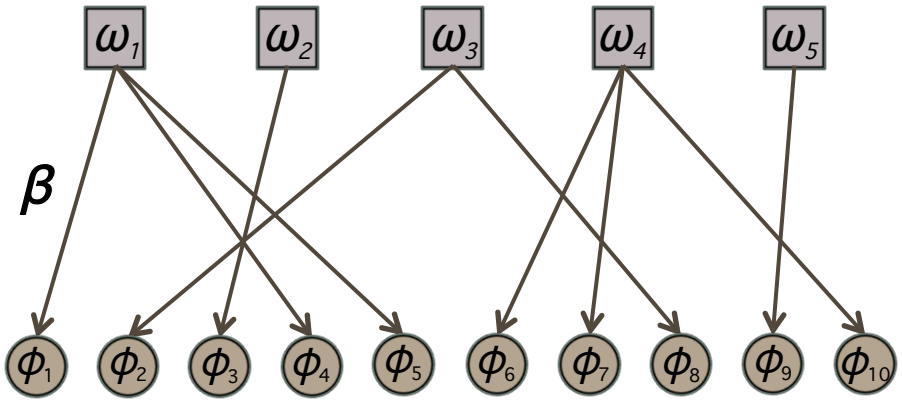


We can induce different structures by constraining the values of β

SPRITE

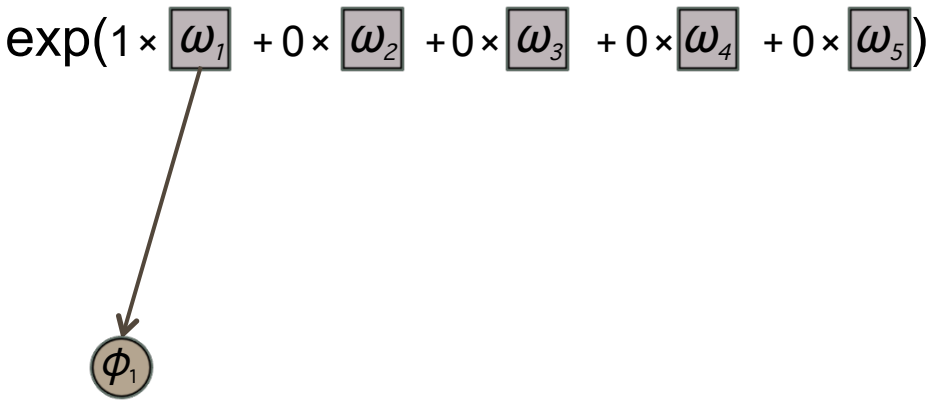
DEFINITION

The priors over distributions are weighted combinations of components:



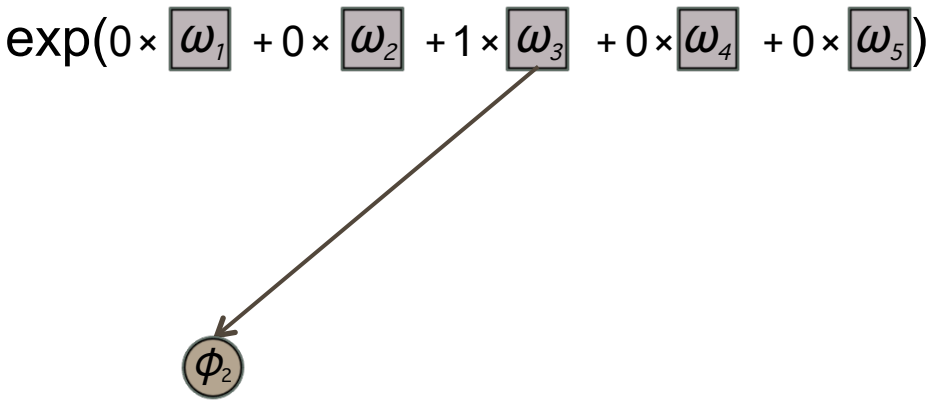
SPRITE DEFINITION

The priors over distributions are weighted combinations of components:



SPRITE DEFINITION

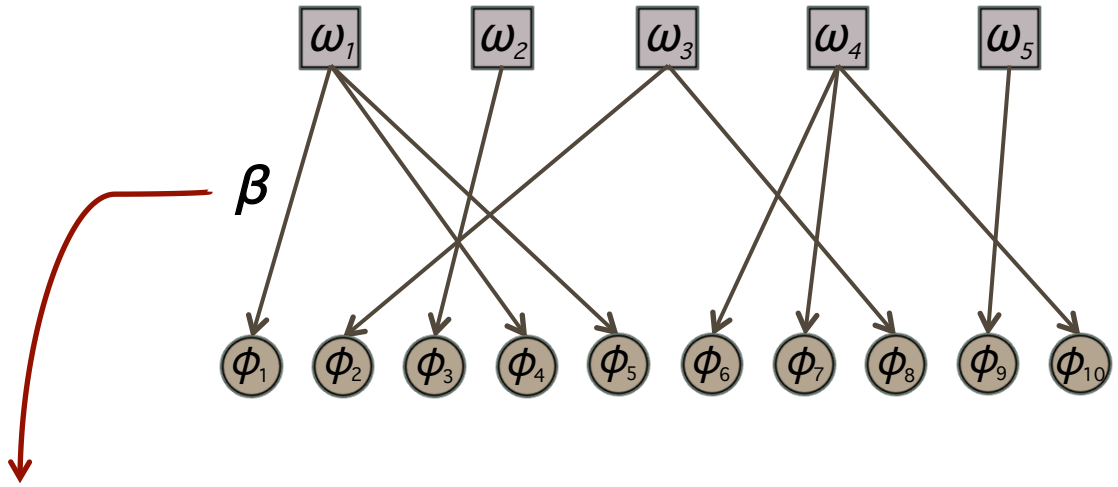
The priors over distributions are weighted combinations of components:



SPRITE

DEFINITION

The priors over distributions are weighted combinations of components:

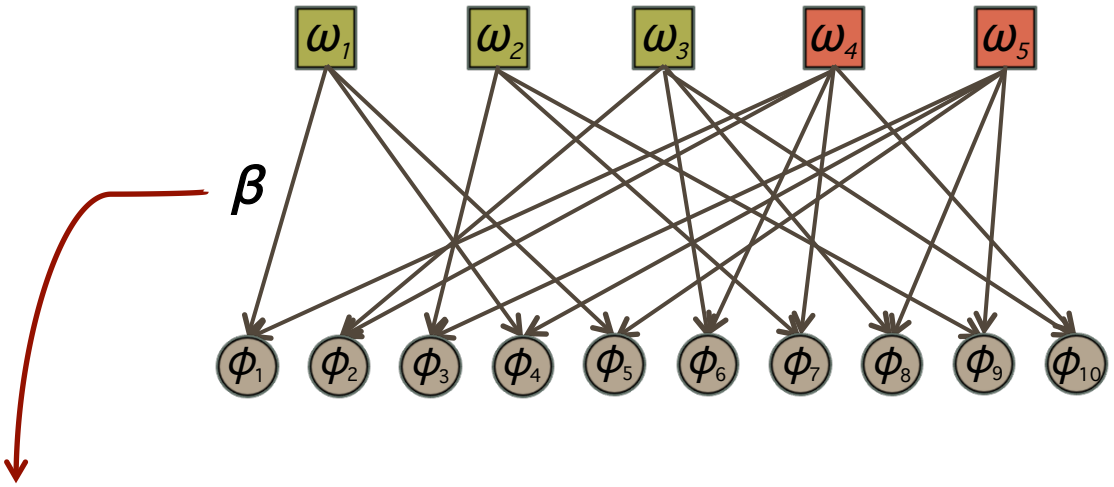


Tree: Each topic's β vector is zero in all but one component

SPRITE

DEFINITION

The priors over distributions are weighted combinations of components:



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

SPRITE

SPECIAL CASE: FLDA

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

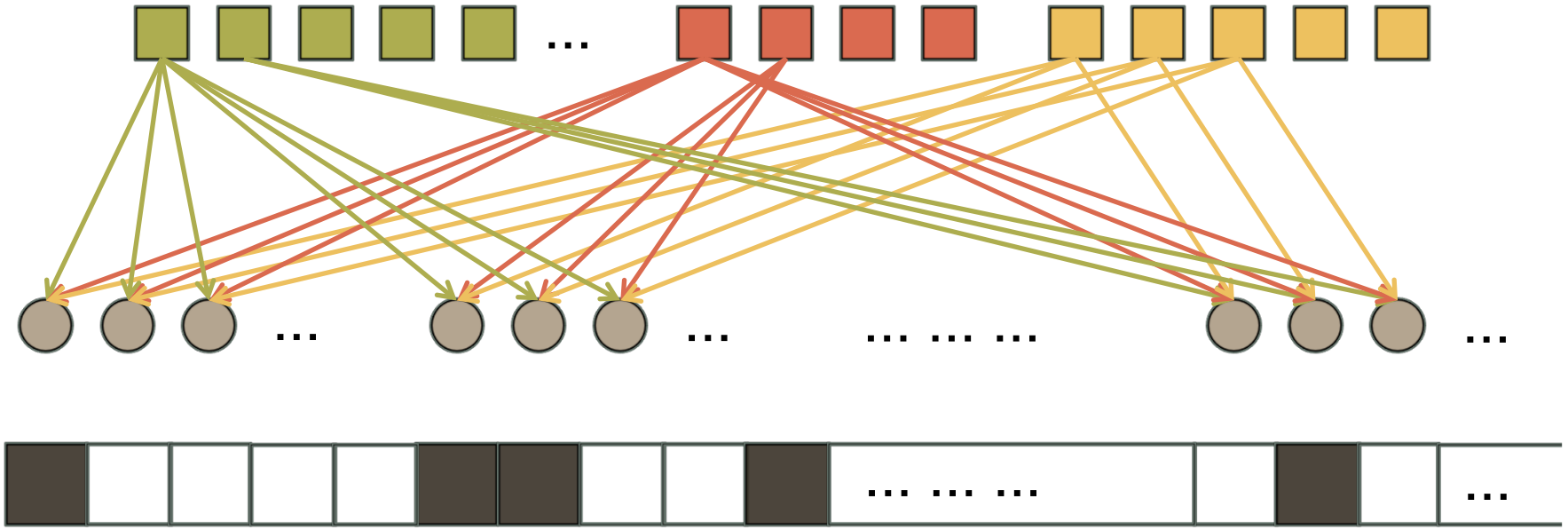
Factorial LDA:

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

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

δ_c is transpose of β_t

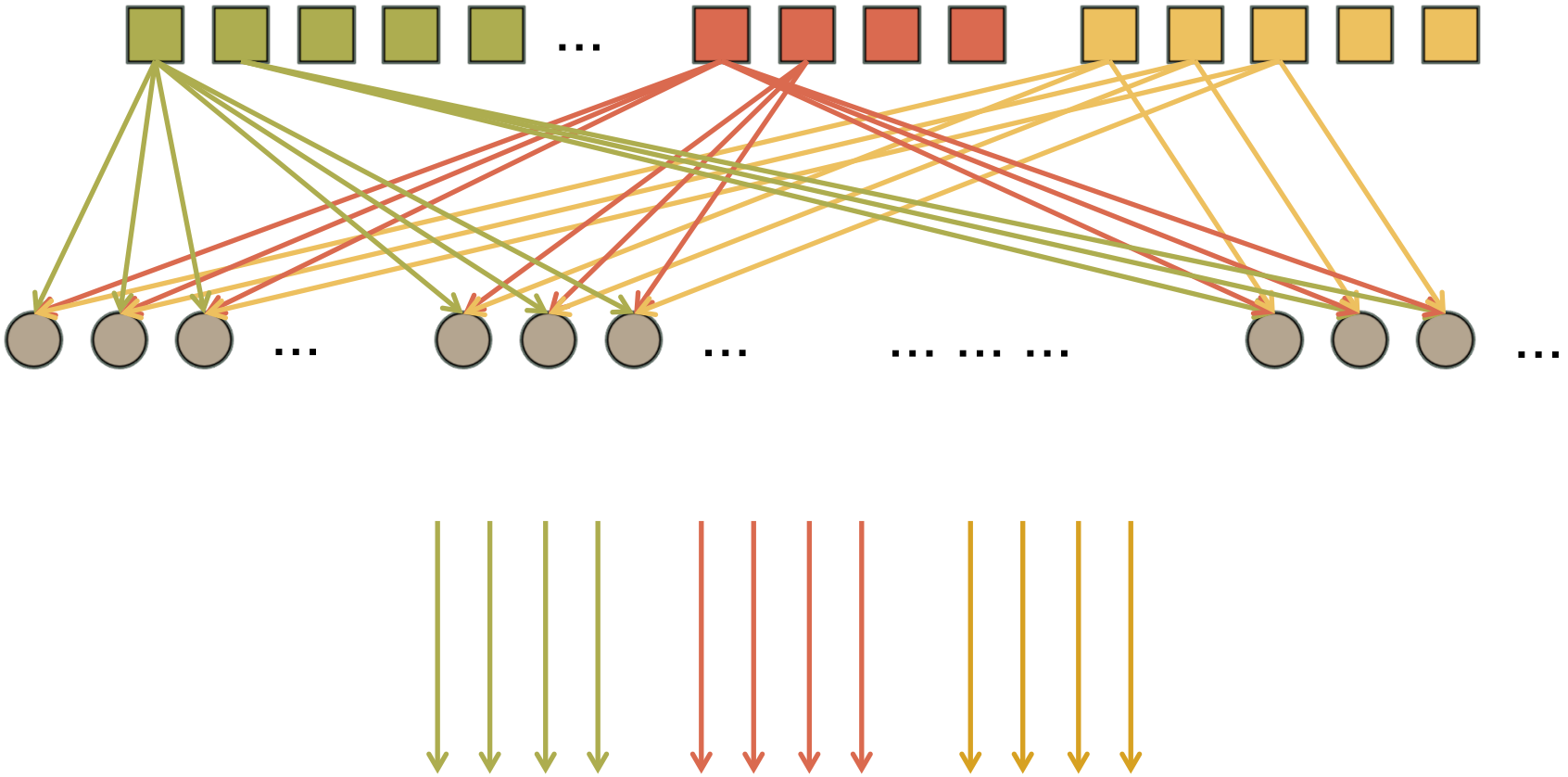
SPECIAL CASE: FLDA



- Initial solution: learn a sparsity pattern

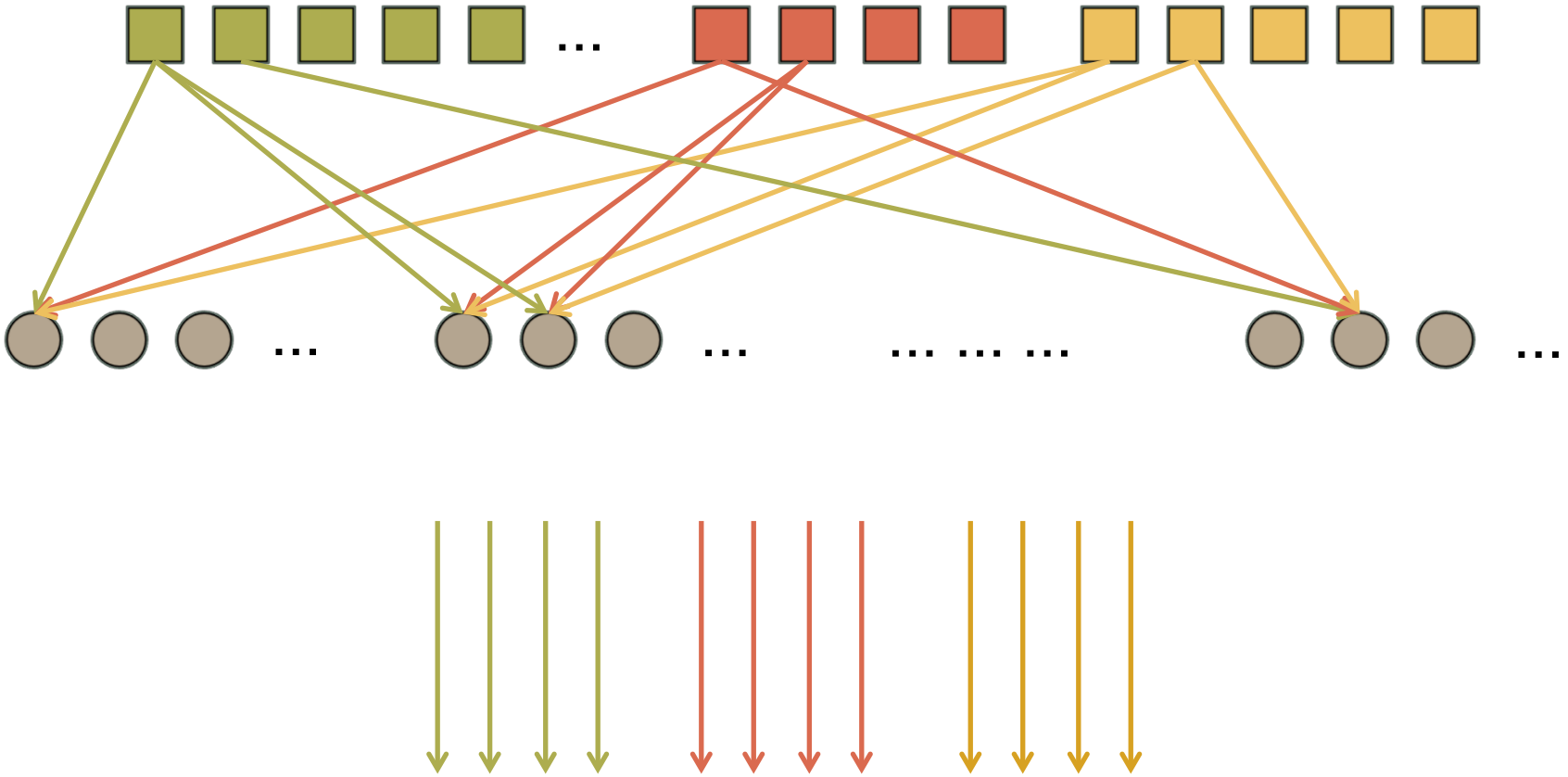
SPRITE

SPECIAL CASE: FLDA



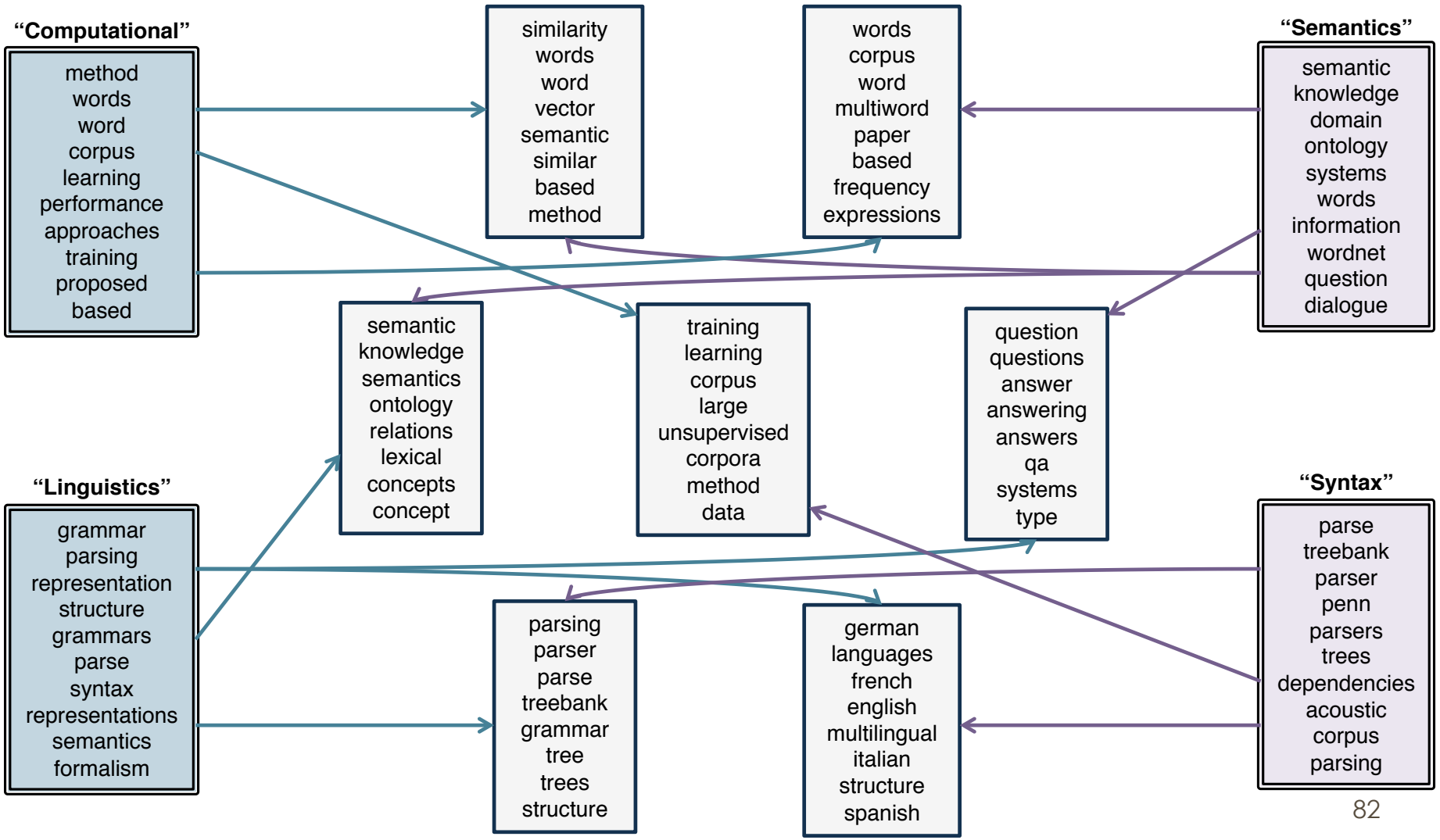
SPRITE

SPECIAL CASE: FLDA



SPRITE

SPECIAL CASE: FLDA



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

PARAMETER ESTIMATION

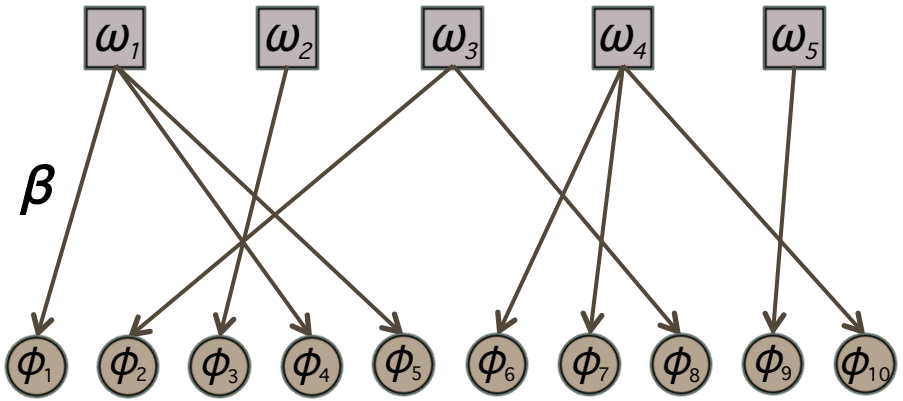
Need to estimate values for the parameters:

- Word and topic distributions
  - Collapsed Gibbs sampling
- Component parameters (i.e. β, ω)
  - Gradient ascent

PARAMETER ESTIMATION

What if β has constraints?

$$\beta_{tc} \in \{0,1\}, \forall c$$
$$\sum_c \beta_{tc} = 1$$



PARAMETER ESTIMATION

Relaxing the constraints

$$\beta_{tc} \in \{0,1\}, \forall c \qquad \sum_c \beta_{tc} = 1$$

$$\beta_{tc} \in (0,1), \forall c \qquad \sum_c \beta_{tc} = 1$$

PARAMETER ESTIMATION

Relaxing the constraints

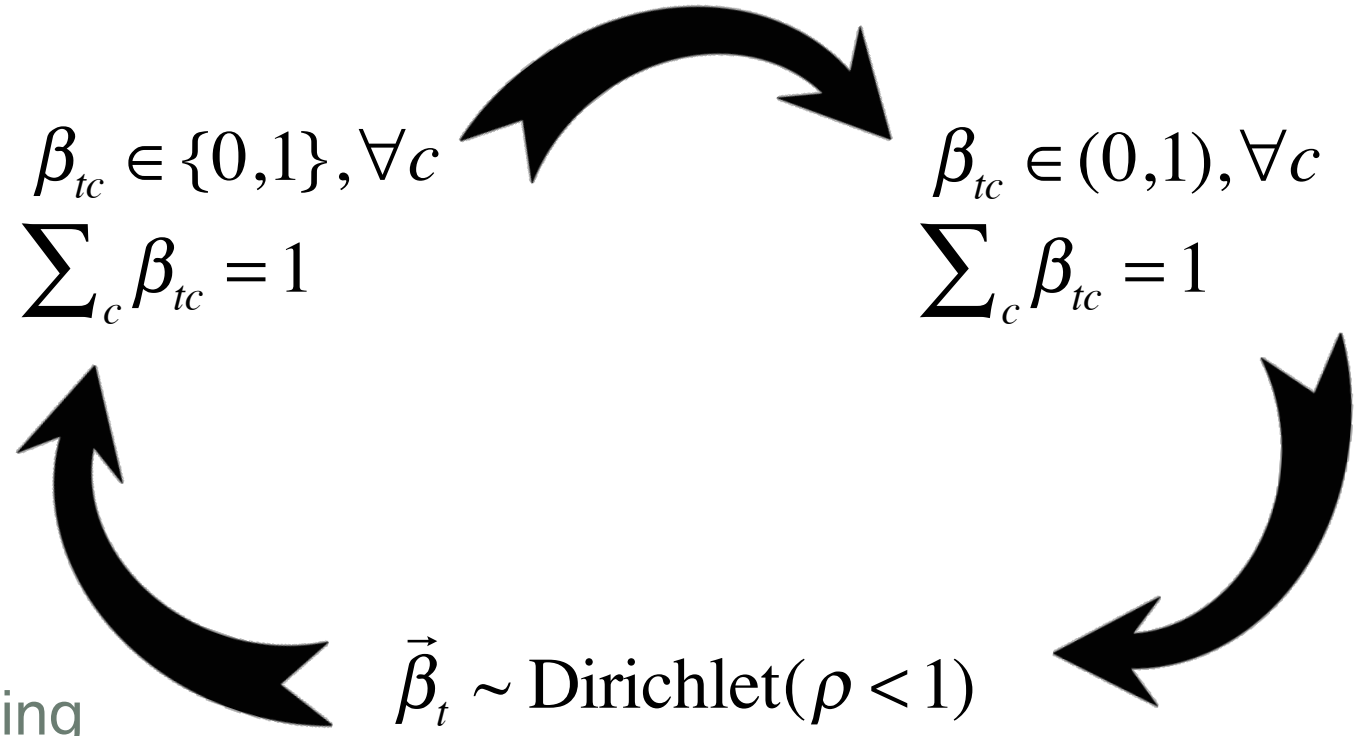
$$\beta_{tc} \in \{0,1\}, \forall c \qquad \sum_c \beta_{tc} = 1$$

$$\beta_{tc} \in (0,1), \forall c \qquad \sum_c \beta_{tc} = 1$$

$$\vec{\beta}_t \sim \text{Dirichlet}(\rho < 1)$$

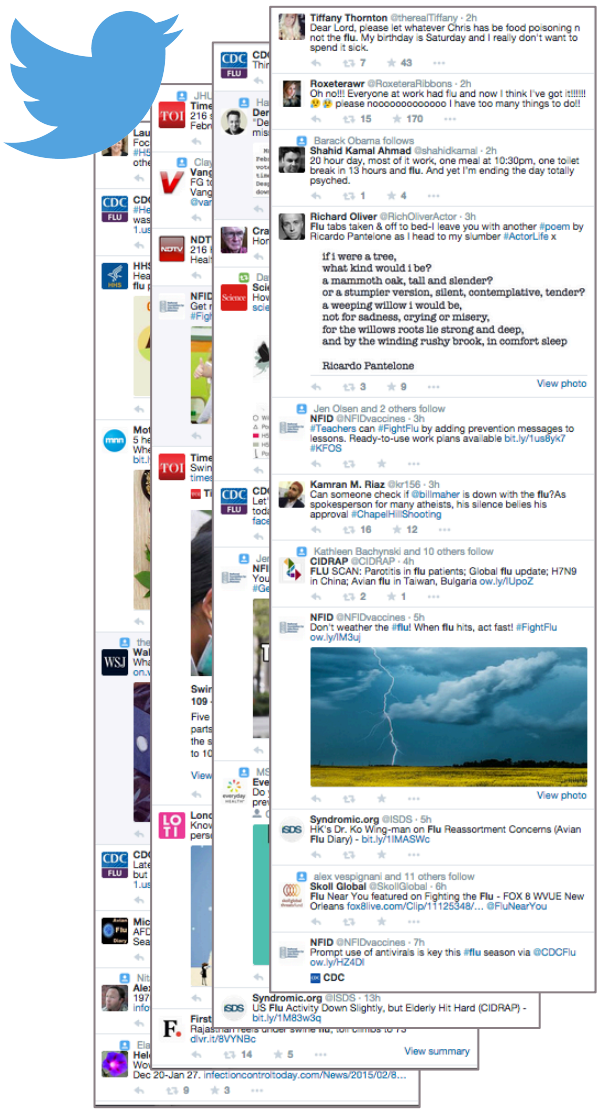
Sparsity-encouraging prior distribution

PARAMETER ESTIMATION



SPRITE

EXAMPLE 1



Modeling perspective in text:

"What opinions are people tweeting about gun control?"

Benton, Paul, Hancock, Dredze. **A structured model of topic and perspective in social media.** In preparation.

SPRITE

EXAMPLE 1

Suppose we want to model how perspective influences topics

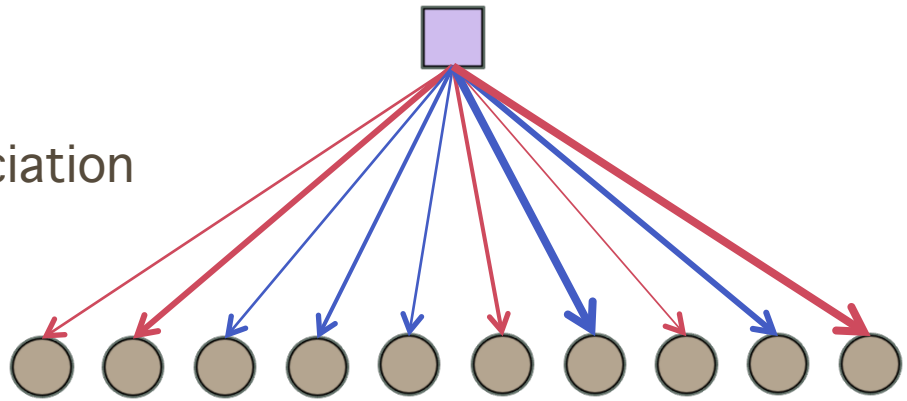
- e.g. certain topics are "pro" or "anti" gun control

A single-component SPRITE model:

$$\tilde{\phi}_{tv} = \exp(r_t \omega_v)$$

The vth word's perspective association

The tth topic's perspective association



SPRITE

EXAMPLE 1

Suppose we want to model how perspective influences topics

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

A single-component SPRITE model:

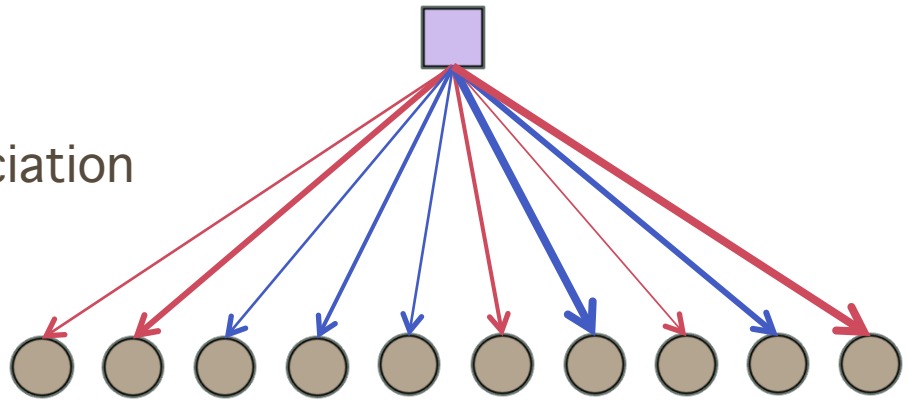
$$\tilde{\phi}_{tv} = \exp(r_t \omega_v)$$

The *v*th word's perspective association

$$\tilde{\theta}_{mt} = \exp(\alpha_m r_t)$$

The *t*th topic's perspective association

The *m*th tweet's perspective association



SPRITE

EXAMPLE 1

Suppose we want to model how perspective influences topics

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

A single-component SPRITE model:

$$\tilde{\phi}_{tv} = \exp(r_t \omega_v)$$

The *v*th word's perspective association

The *t*th topic's perspective association

$$\tilde{\theta}_{mt} = \exp(\alpha_m r_t)$$

Prior is a function of:

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

SPRITE

EXAMPLE 1

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: Pro gun control | Anti gun control

The model can help estimate the % gun ownership in US states:

- Mean error: 8.4
- Baselines: 12.7 – 16.4

EXAMPLE 2

Suppose we want to model perspective
and we want to organize topics in a hierarchy

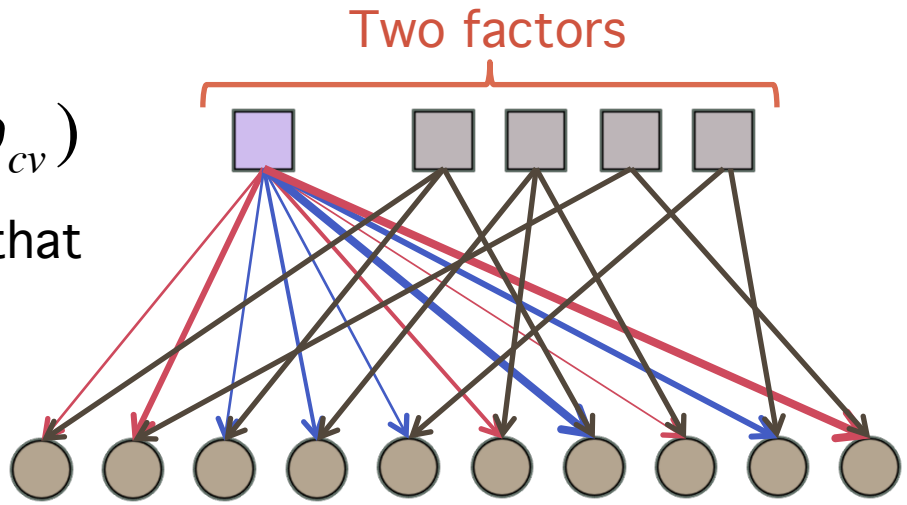
SPRITE

EXAMPLE 2

Suppose we want to model perspective
and we want to organize topics in a hierarchy

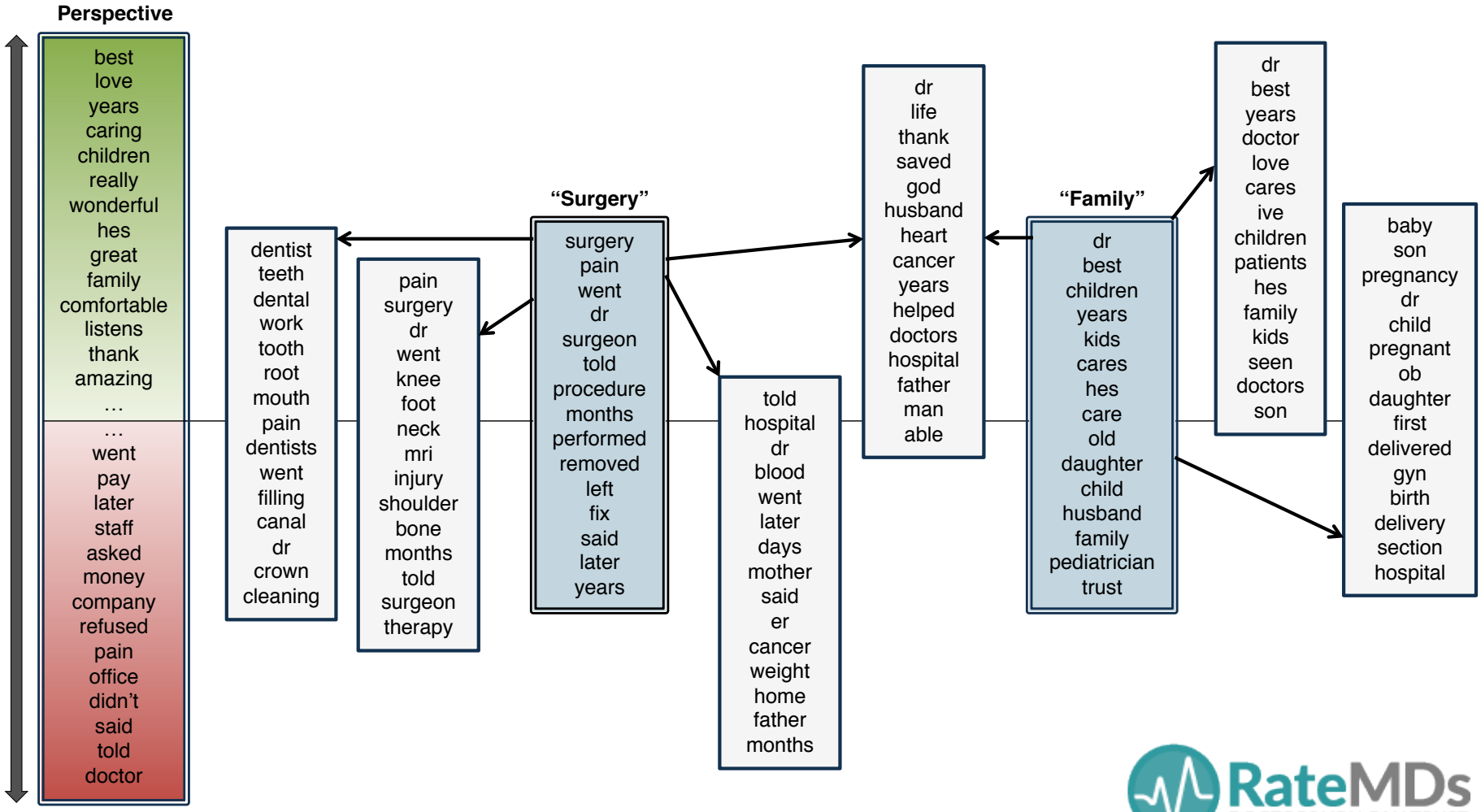
$$\tilde{\phi}_{tv} = \exp(r_t \omega_{0v} + \sum_{c=1}^{C(\phi)} \beta_{tc} \omega_{cv})$$

with (soft) constraints that
β_i is an indicator vector



SPRITE

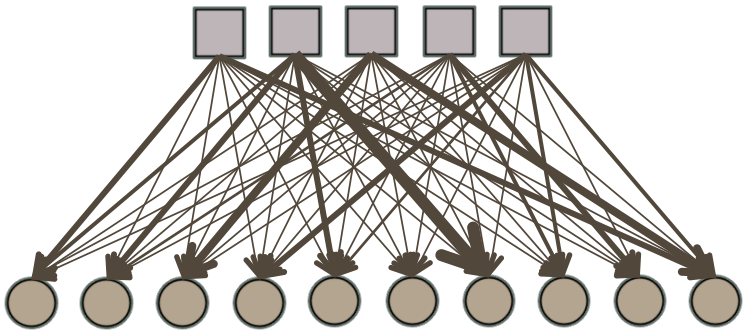
EXAMPLE 2



RateMDs
DOCTORS YOU CAN TRUST

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

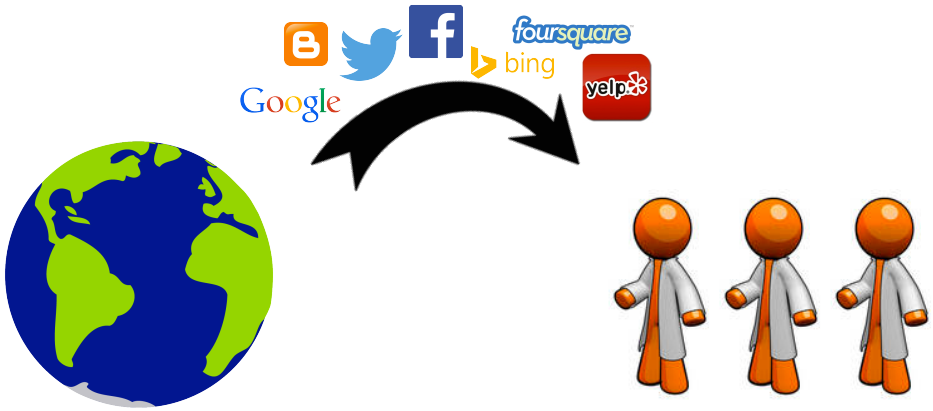


CONCLUSION

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

We need methods to understand what people are saying on the web and learn meaningful trends

This requires models that can discover patterns automatically, while accommodating user expectations



THANK YOU

WITH HELP FROM:

Committee: Mark Dredze, Jason Eisner, Eric Horvitz, Hanna Wallach
Funding: Microsoft Research, NSF, JHU Dean's office

Public opinion:
- Adrian Benton
- Braden Hancock

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

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

Microsoft Research

human language technology
center of excellence

THANK YOU

QUESTIONS?