Machine Learning 2

DS 4420 - Spring 2020

Topic Modeling 2

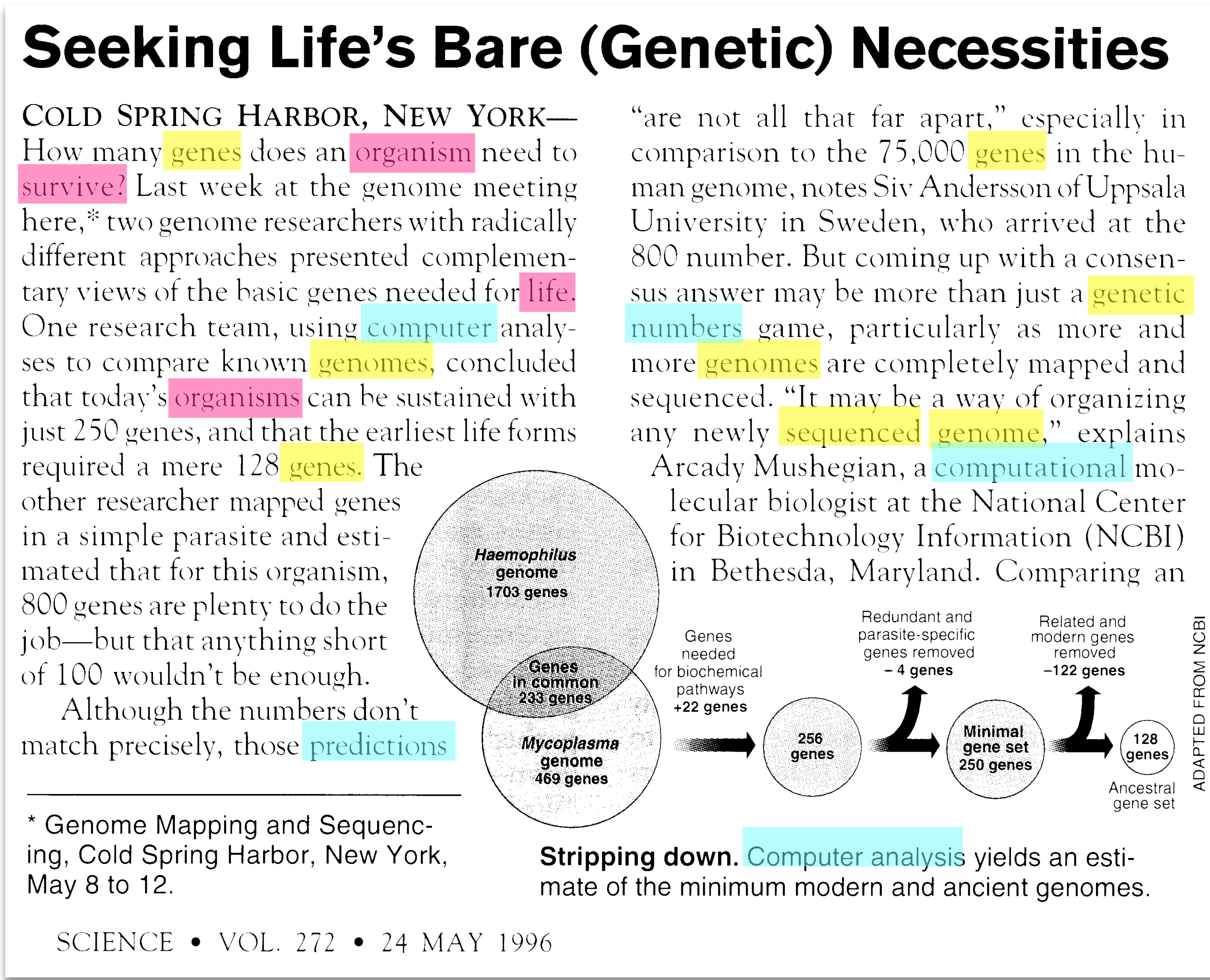
Byron C. Wallace

Last time:

Topic Modeling!

Word Mixtures

Idea: Model text as a mixture over words (ignore order)



Seeking Life's Bare (Genetic) Necessities

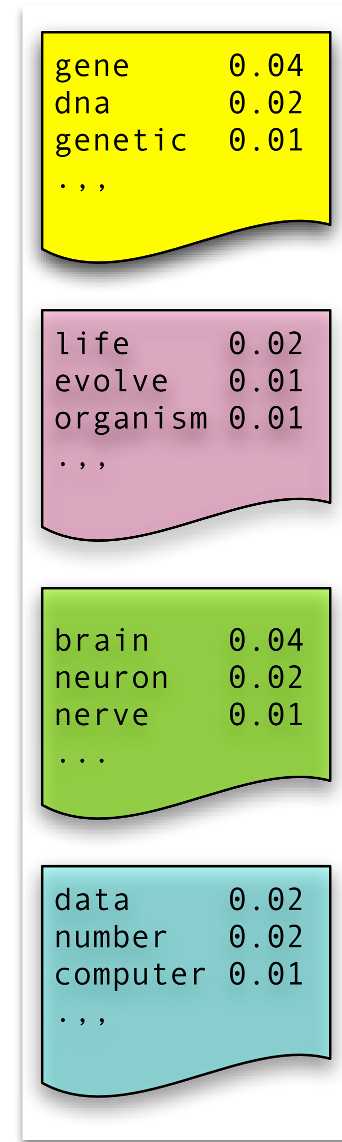
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

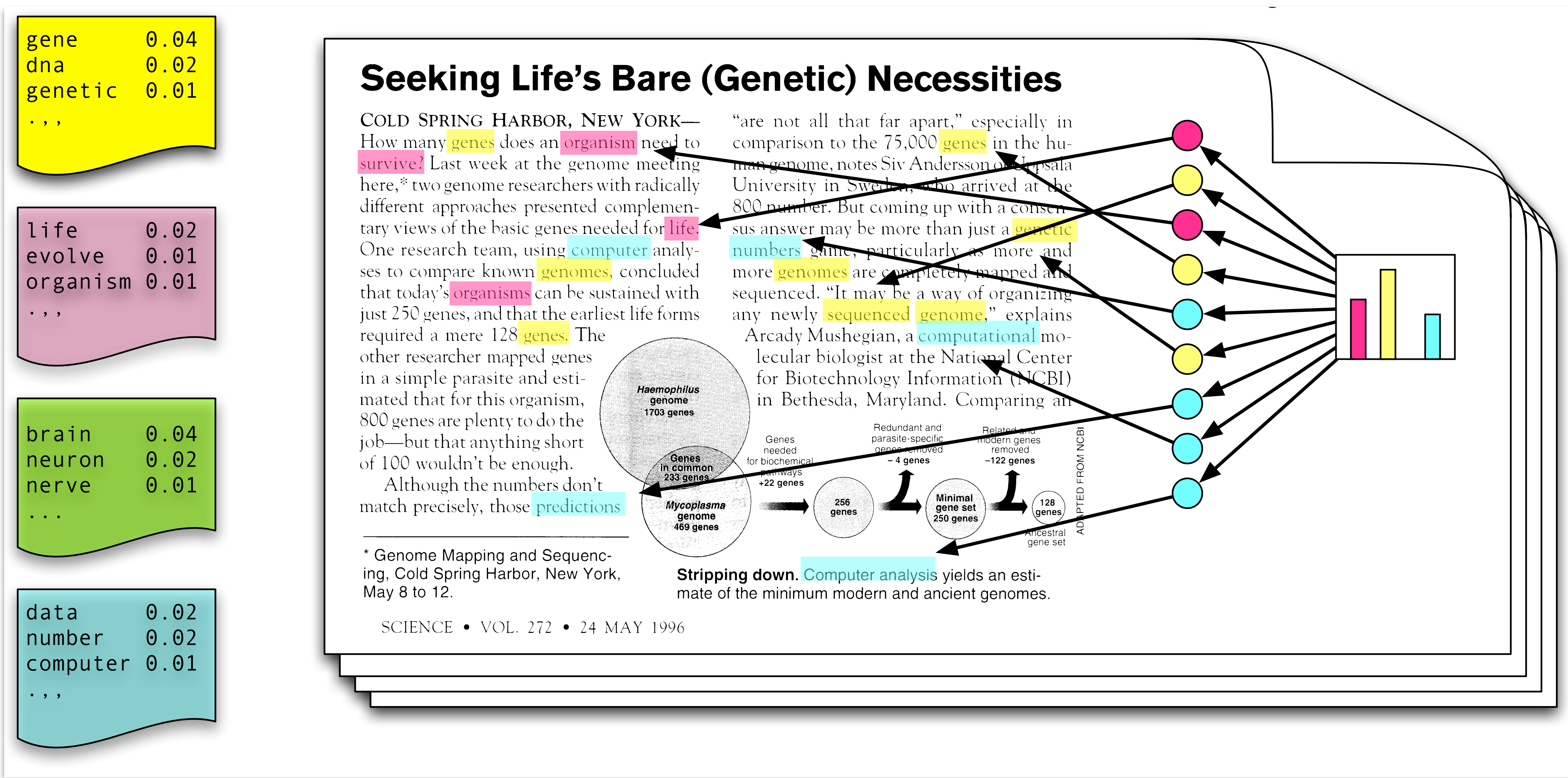
SCIENCE • VOL. 272 • 24 MAY 1996



Words: $x_n | z_n = k \sim \mathrm{Discrete}(\boldsymbol{\beta}_k)$ Topics: $z_n \sim \mathrm{Discrete}(\boldsymbol{\theta})$

Topic Modeling

Topics (shared) | **Words in Document** (mixture over topics) | **Topic Proportions** (document-specific)



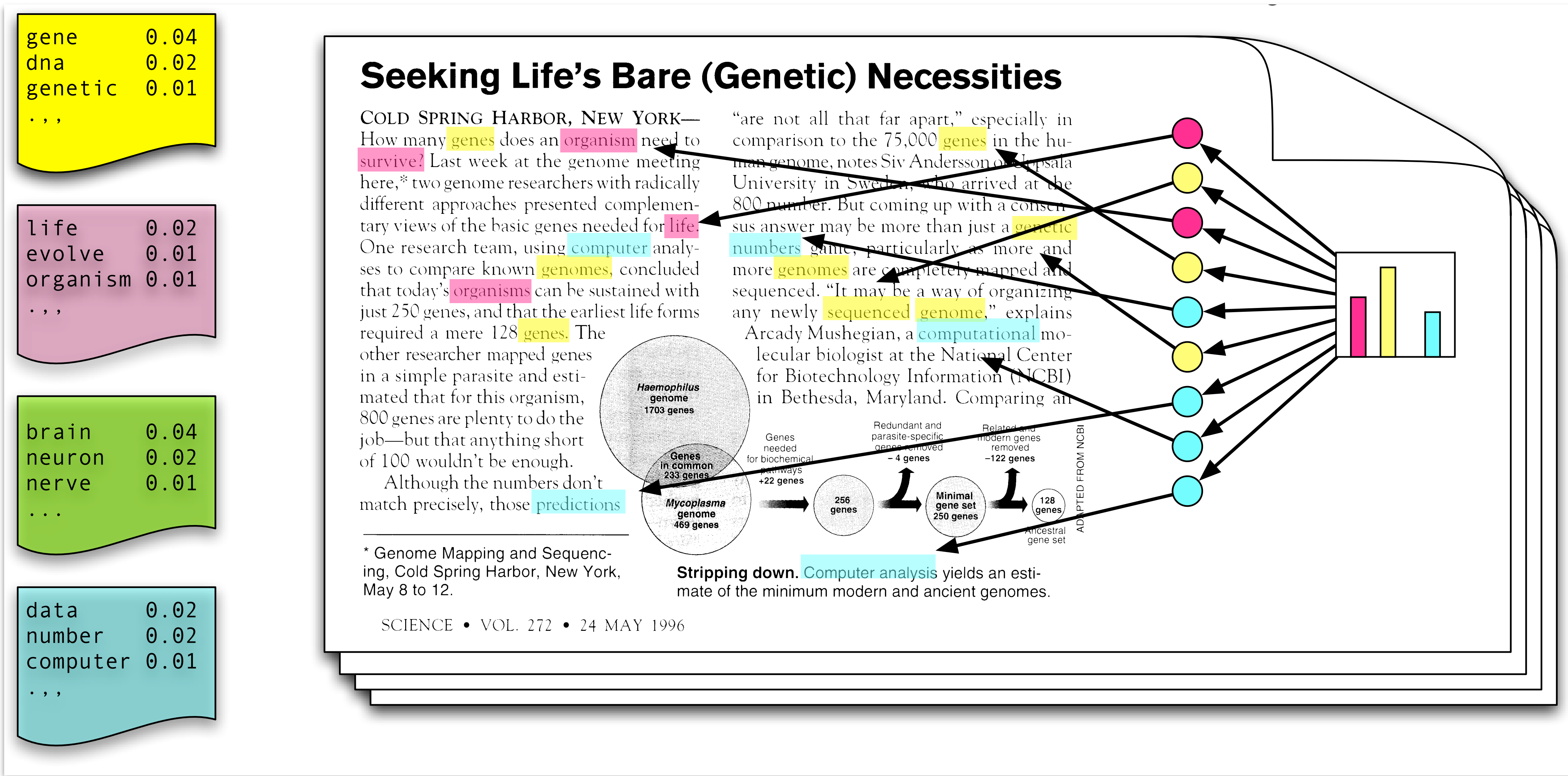
Idea: Model ***corpus*** of documents with ***shared*** topics

Topic Modeling

Topics
(shared)

Words in Document
(mixture over topics)

Topic Proportions
(document-specific)



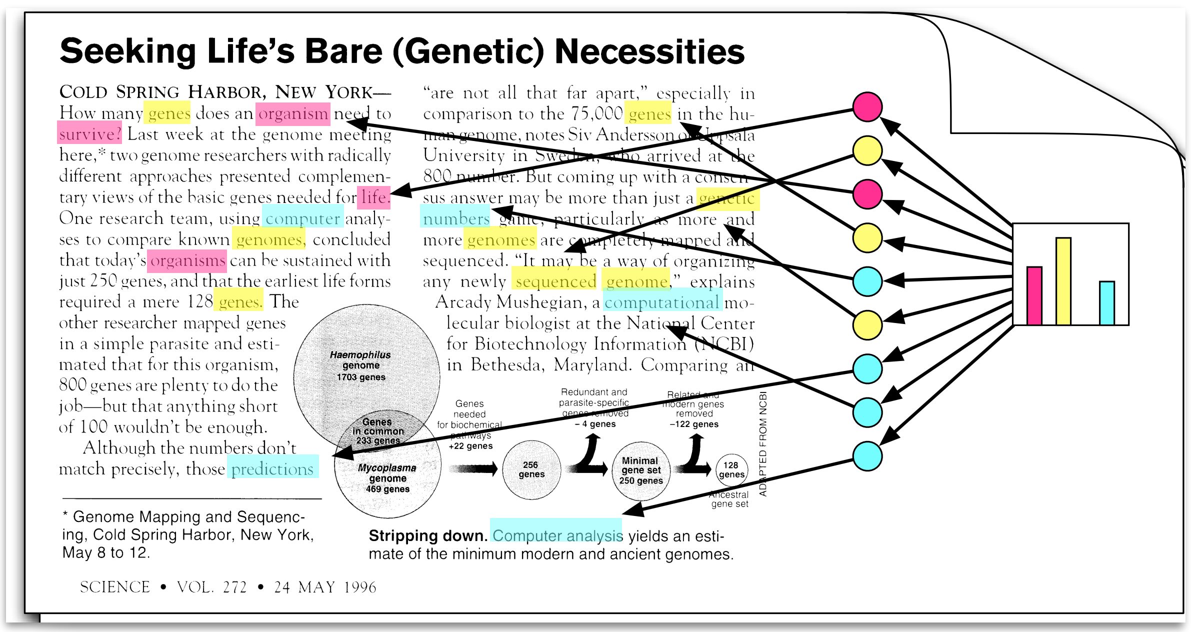
- Each **topic** is a distribution over words
- Each **document** is a mixture over topics
- Each **word** is drawn from one topic distribution

EM for Word Mixtures (PLSA)

Generative Model

$$z_n \sim \mathrm{Discrete}(\boldsymbol{\theta})$$

$$x_n | z_n = k \sim \mathrm{Discrete}(\boldsymbol{\beta}_k)$$



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

E-step: Update assignments

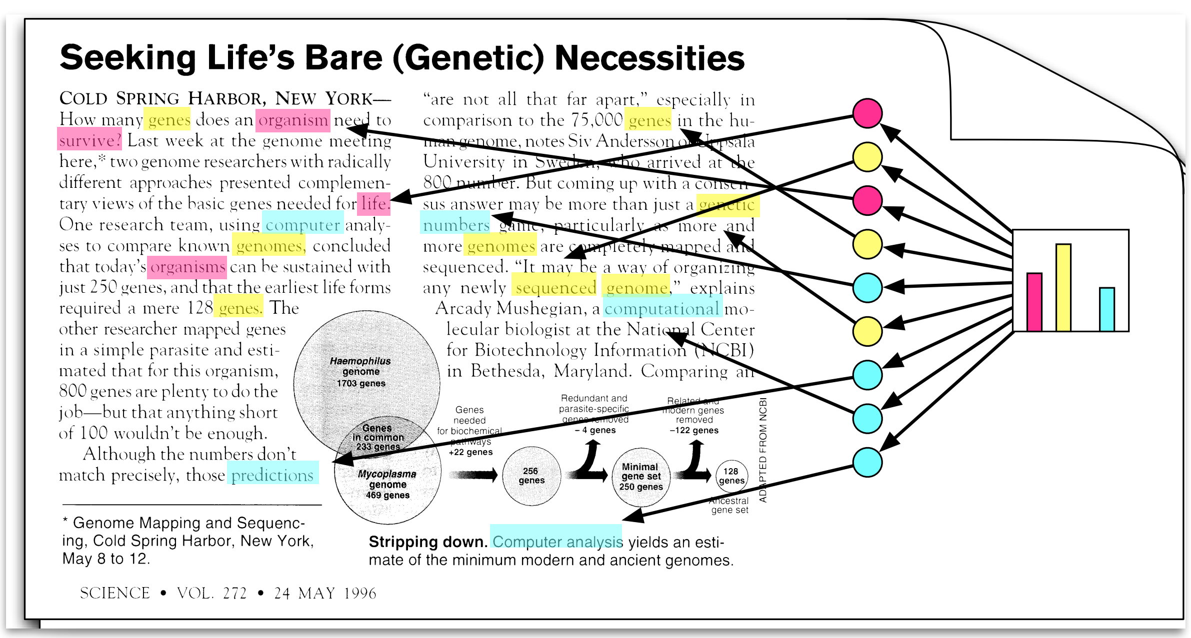
M-step: Update parameters

EM for Word Mixtures (PLSA)

Generative Model

$$z_n \sim \mathrm{Discrete}(\boldsymbol{\theta})$$

$$x_n | z_n = k \sim \mathrm{Discrete}(\boldsymbol{\beta}_k)$$



E-step: Update assignments

$$\phi_{nk} = \frac{\theta_k \beta_{kv}}{\sum_l \theta_l \beta_{lv}} \qquad x_v = v$$

M-step: Update parameters

EM for Word Mixtures (PLSA)

Generative Model

$$z_n \sim \text{Discrete}(\boldsymbol{\theta})$$

$$x_n | z_n = k \sim \text{Discrete}(\boldsymbol{\beta}_k)$$

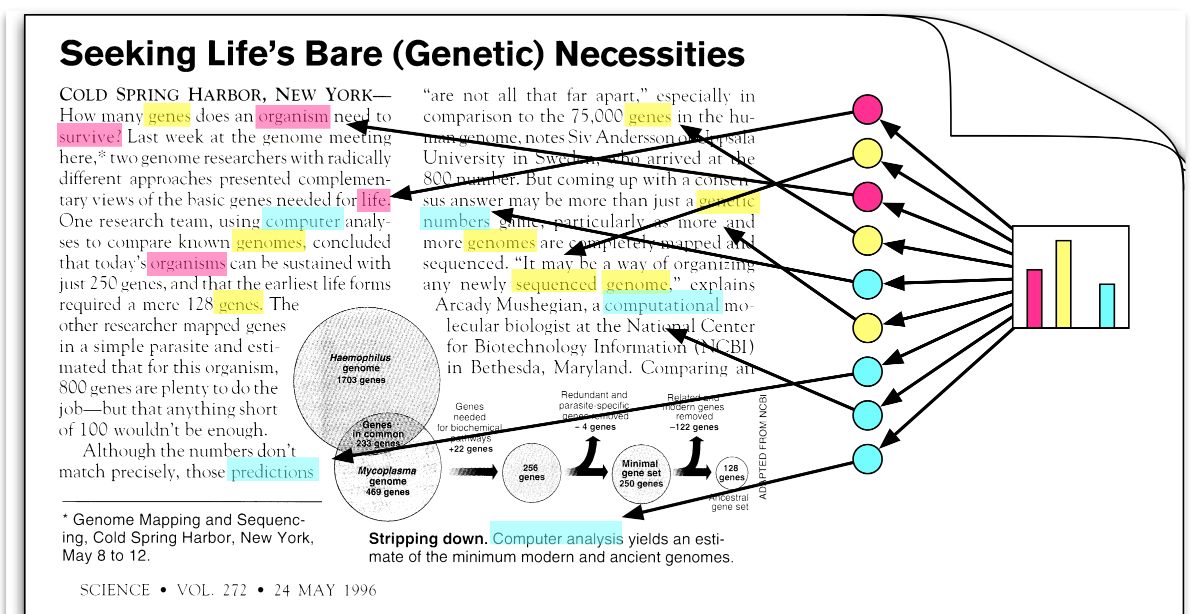
E-step: Update assignments

$$\phi_{nk} = \frac{\theta_k \beta_{kv}}{\sum_l \theta_l \beta_{lv}} \qquad x_v = v$$

M-step: Update parameters

$$\beta_{kv} = \frac{N_{kv}}{\sum_w N_{kw}} \qquad N_{kv} := \sum_{n=1}^{N} \phi_{nk} x_{nv}$$

$$\theta_k = \frac{N_k}{\sum_l N_l} \qquad N_k := \sum_{n=1}^{N} \phi_{nk}$$



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

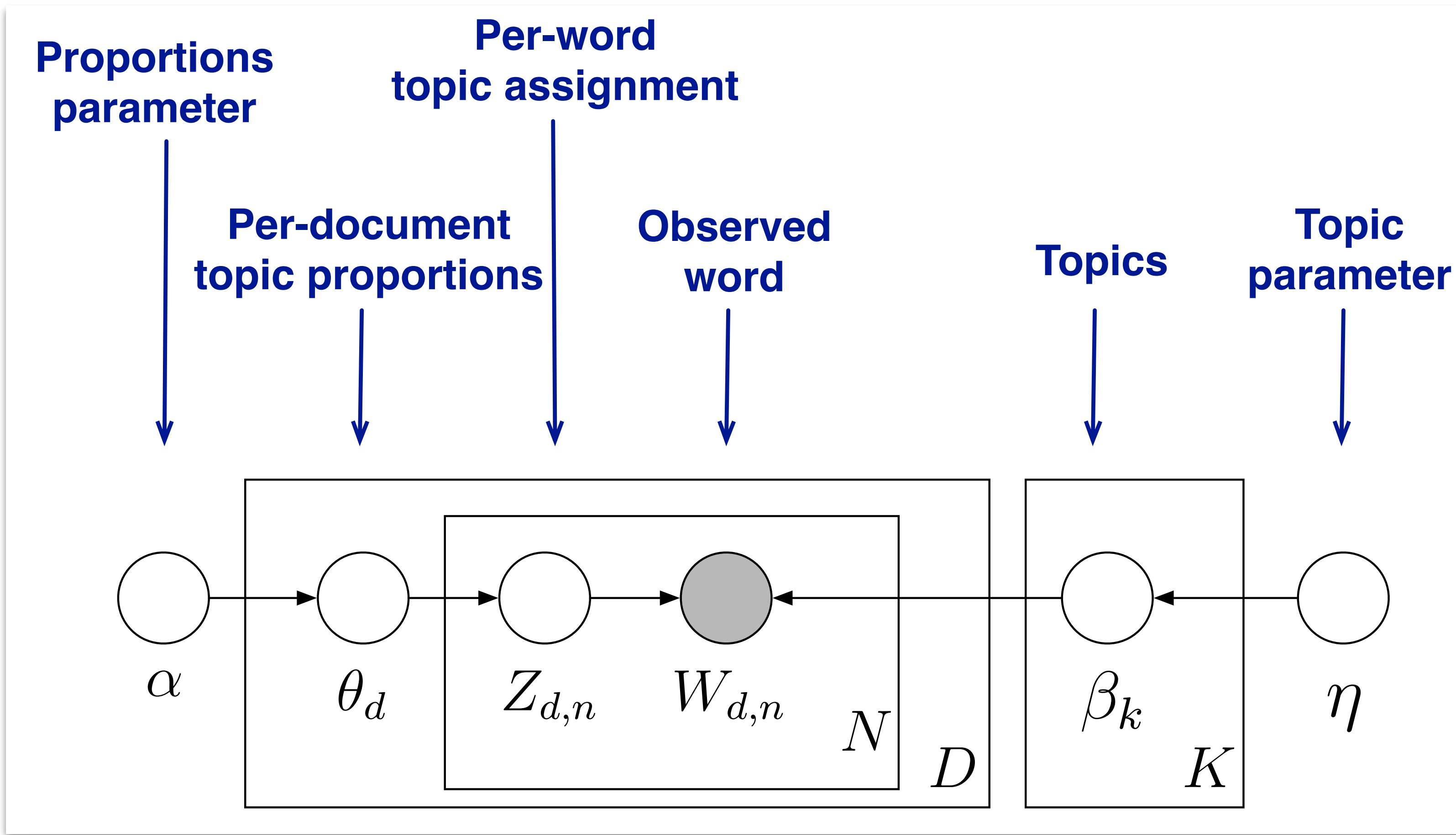
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Today: A Bayesian view — topic modeling with priors (or, LDA)

Latent Dirichlet Allocation

(a.k.a. PLSI/PLSA with priors)



$$\beta_k \sim \text{Dirichlet}(\eta) \quad k = 1, \ldots, K$$

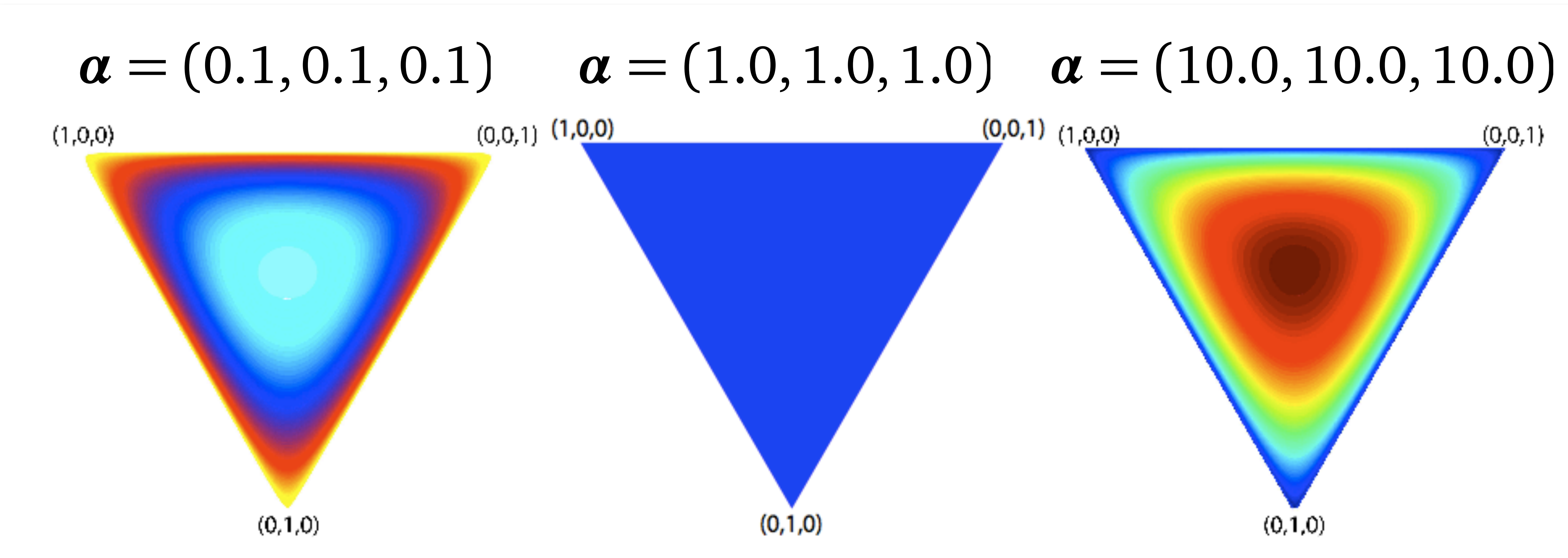
$$\theta_d \sim \text{Dirichlet}(\alpha) \quad d = 1, \ldots, D$$

$$Z_{d,n} \sim \text{Discrete}(\theta_d) \quad n = 1, \ldots, N_d$$

$$W_{d,n} | Z_{d,n} = k \sim \text{Discrete}(\beta_k) \quad n = 1, \ldots, N_d$$

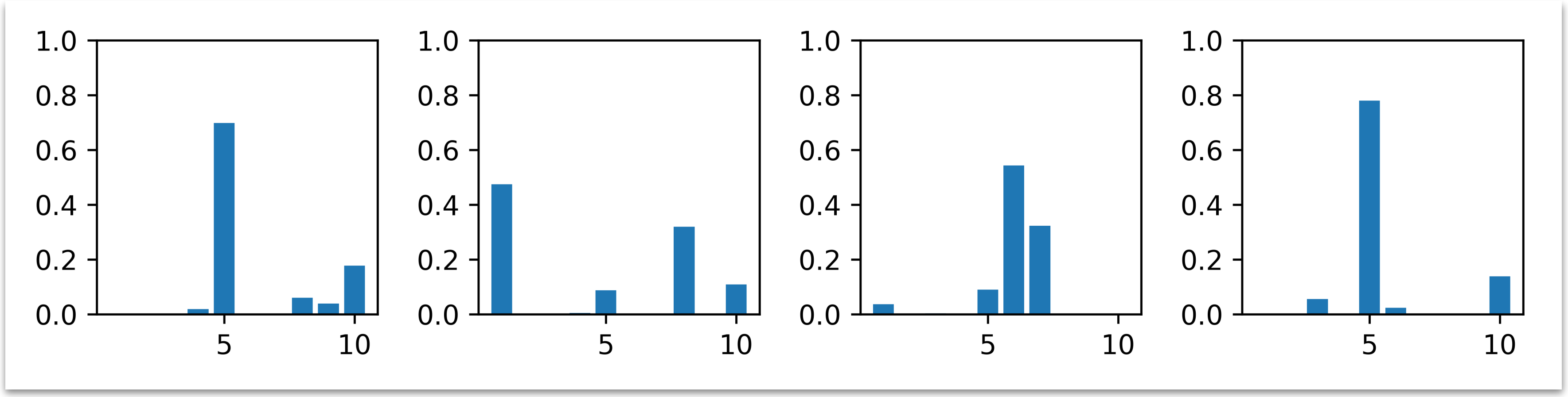
Dirichlet Distribution

$$p(\boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad B(\boldsymbol{\alpha}) := \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

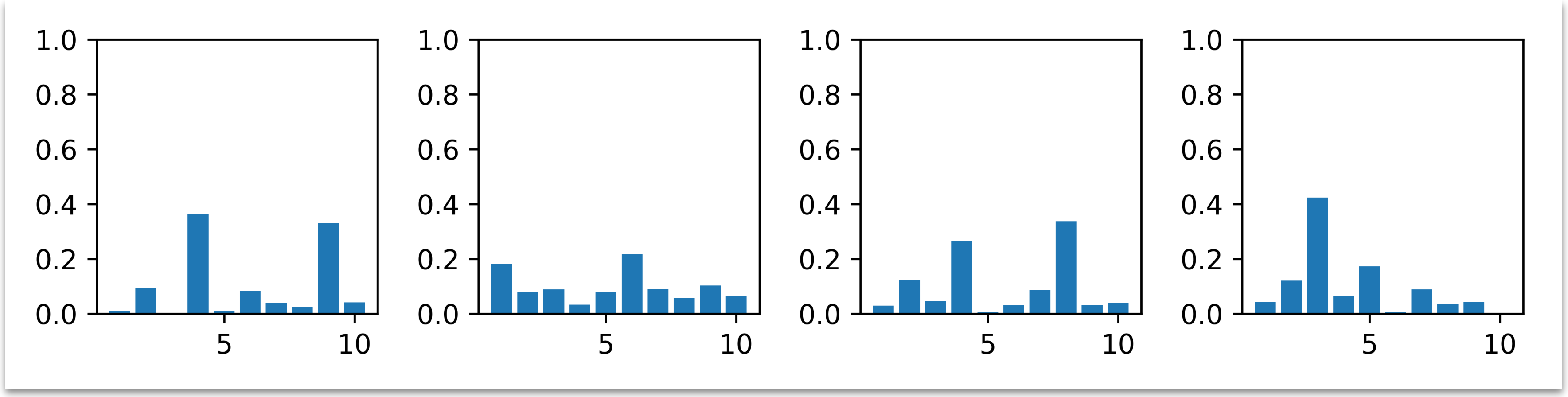


Dirichlet Distribution

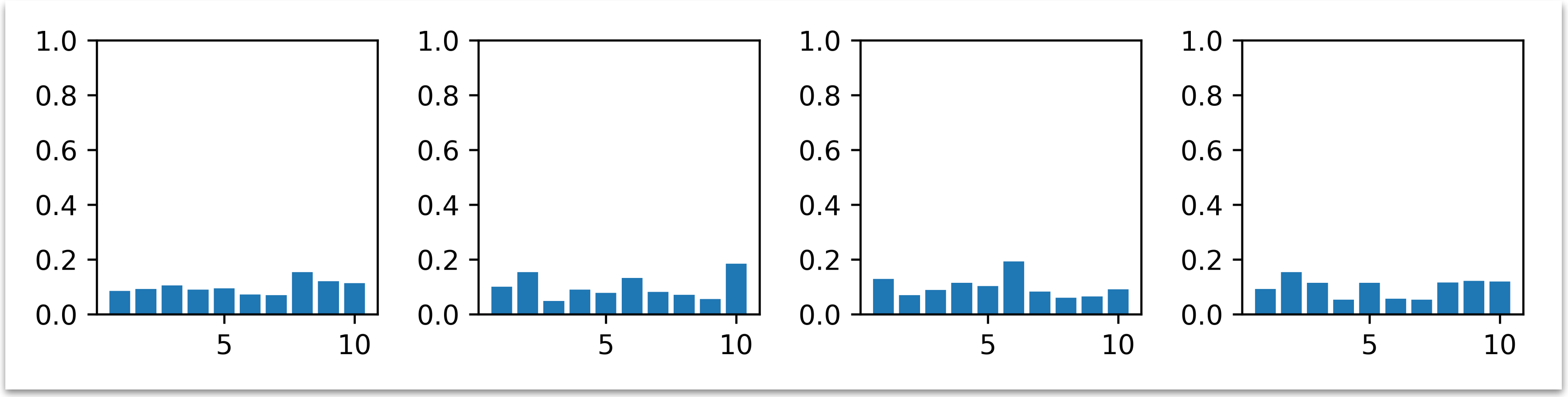
$\alpha_k = 0.1$



$\alpha_k = 1.0$



$\alpha_k = 10.0$



Common choice in LDA: $\alpha_k = 0.001$

Estimation via sampling (board)

Extensions of LDA

Latent dirichlet allocation
DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org
Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying ...
Cited by 15971 Related articles All 124 versions Cite Save

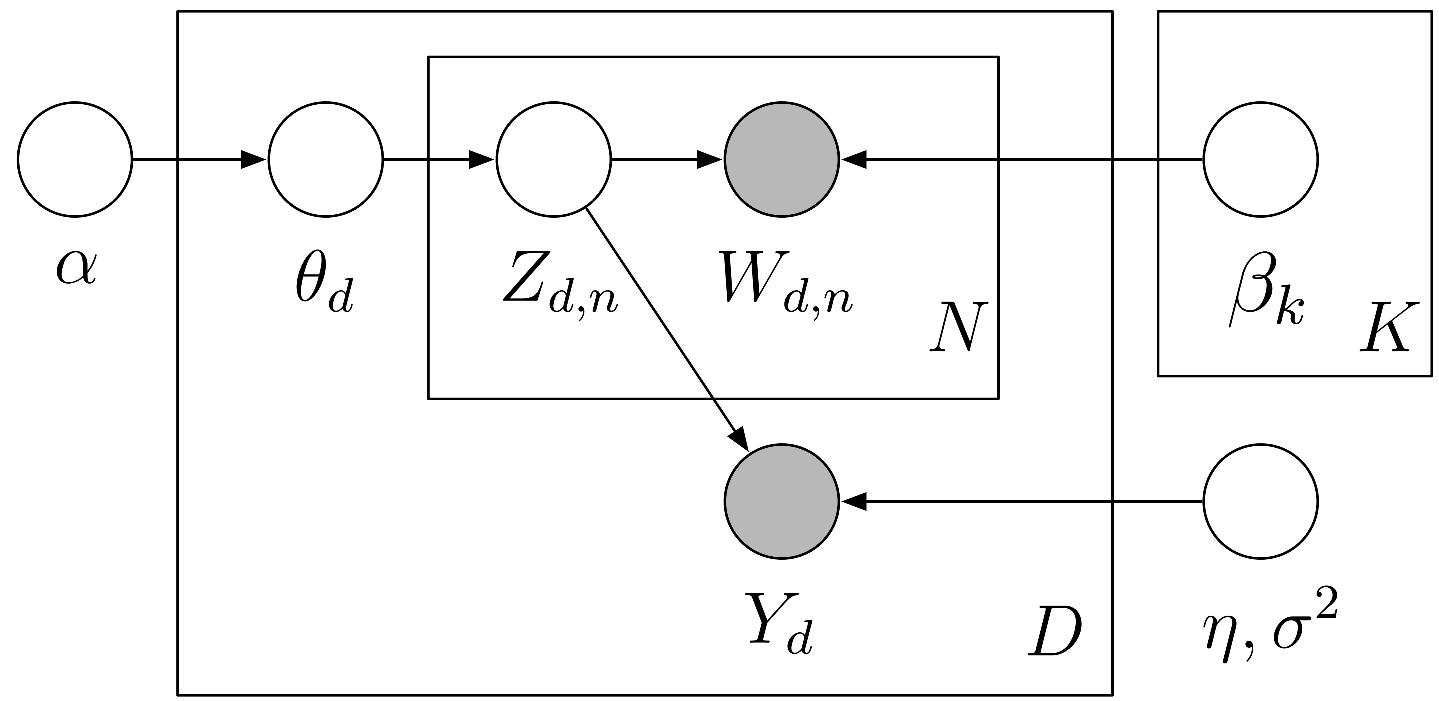
- EM inference (PLSA/PLSI) yields similar results to Variational inference or MAP inference (LDA) on most data

Extensions of LDA

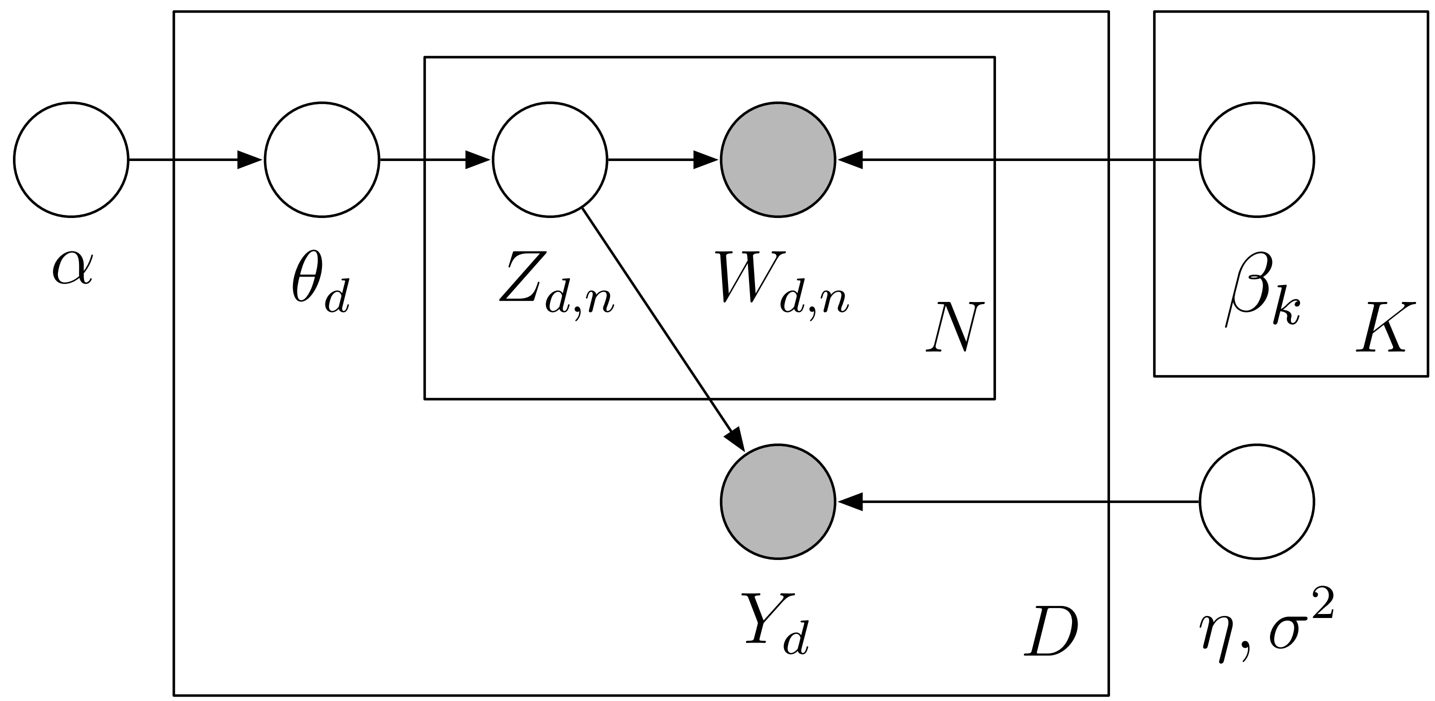
Latent dirichlet allocation
DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org
Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying ...
Cited by 15971 Related articles All 124 versions Cite Save

- EM inference (PLSA/PLSI) yields similar results to Variational inference or MAP inference (LDA) on most data
- Reason for popularity of LDA:
  can be embedded in more complicated models

Extensions: Supervised LDA

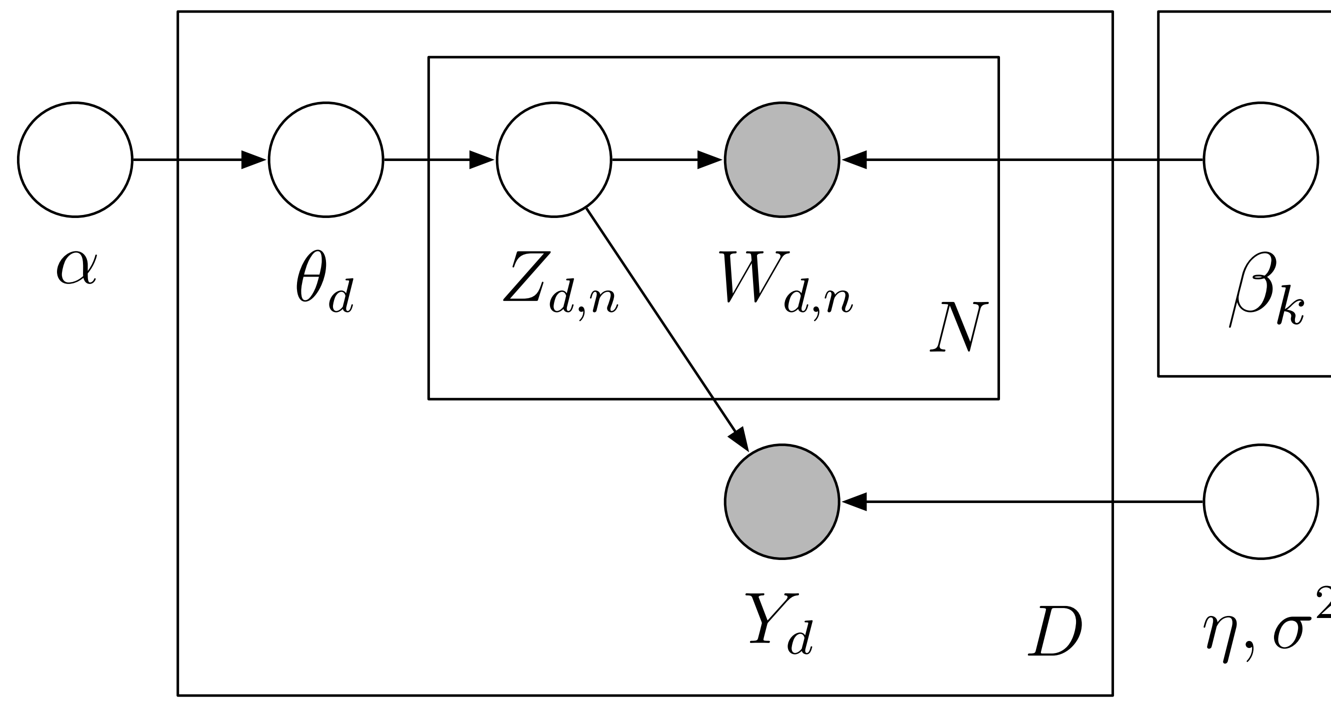


Extensions: Supervised LDA



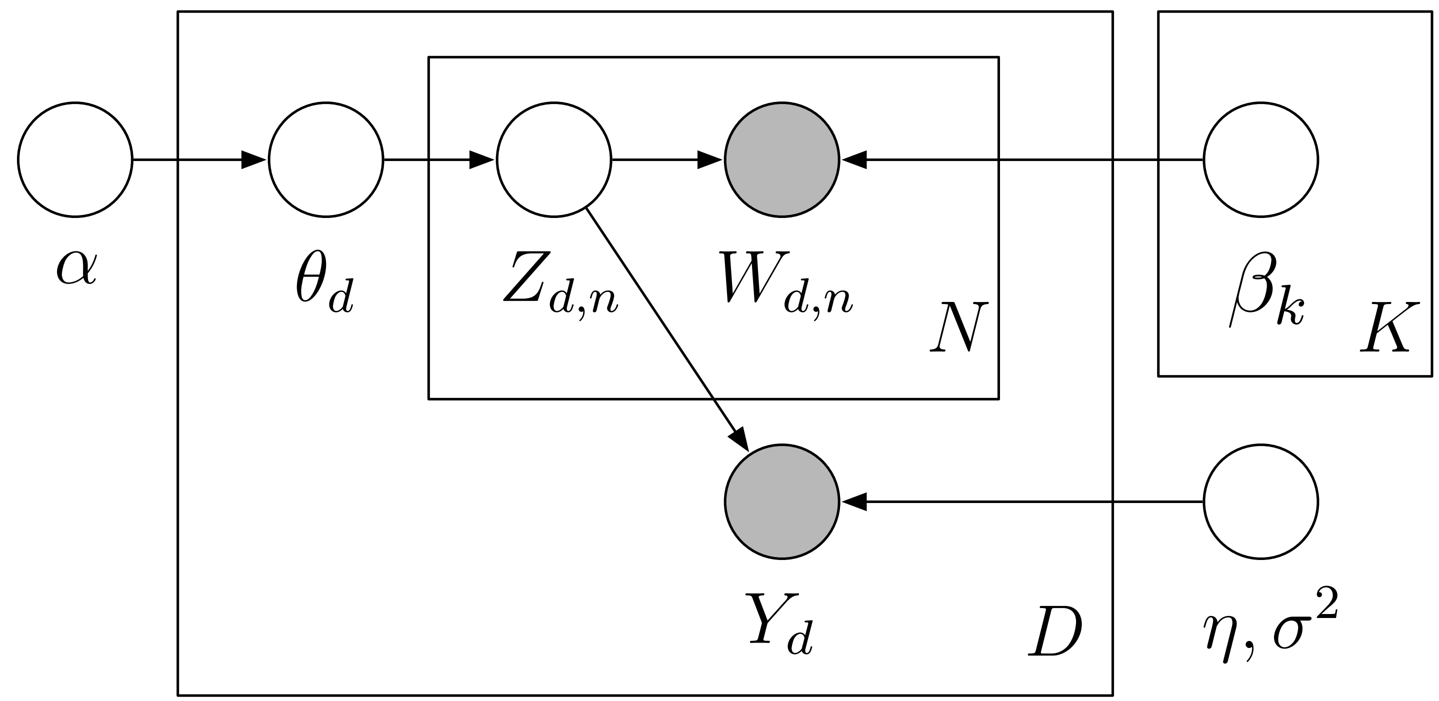
1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.

Extensions: Supervised LDA



1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.

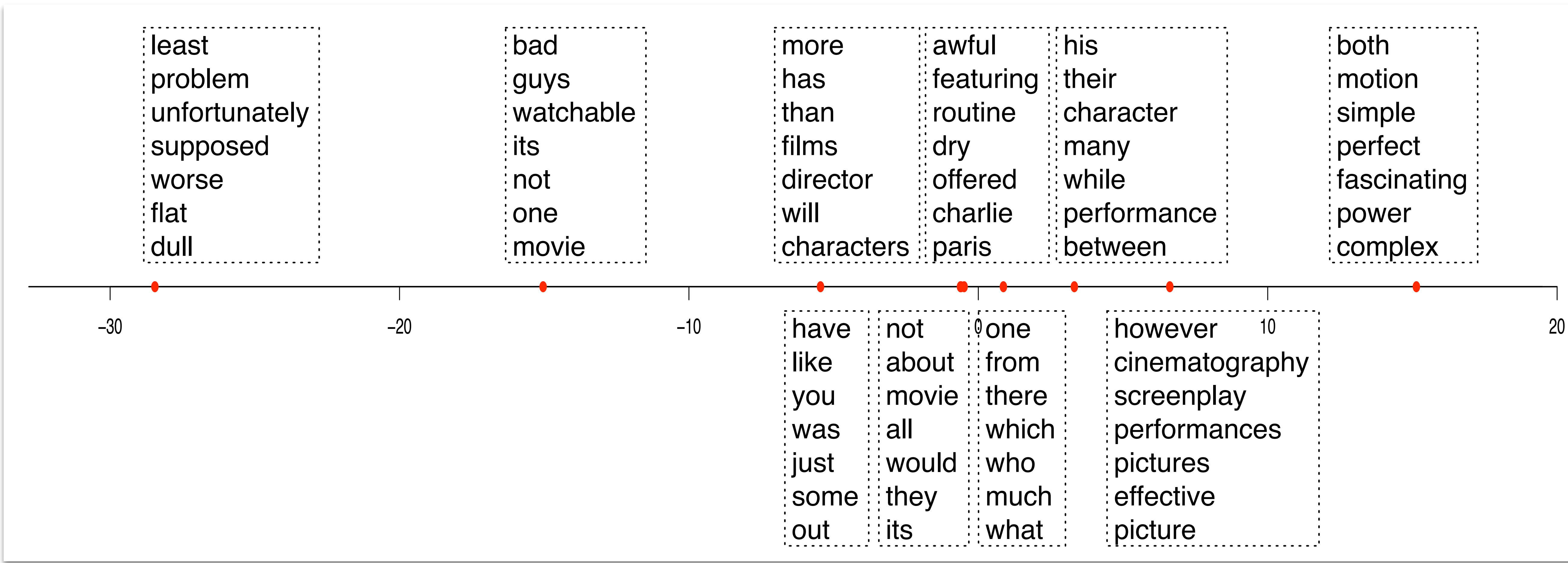
Extensions: Supervised LDA



1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\left(\eta^\top \bar{z}, \sigma^2\right)$, where

$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Extensions: Supervised LDA



Article | Cited By (3) | Tweetations (13) | Metrics

Original Paper

Characterizing the (Perceived) Newsworthiness of Health Science Articles: A Data-Driven Approach

Ye Zhang[1], MS ; Erin Willis[2], PhD ; Michael J Paul[3], PhD ; Noémie Elhadad[4], PhD ; Byron C Wallace[5], PhD

[1]Department of Computer Science, University of Texas at Austin, Austin, TX, United States

[2]College of Media, Communication and Information, University of Colorado Boulder, Boulder, CO, United States

[3]Department of Information Science, University of Colorado Boulder, Boulder, CO, United States

[4]Biomedical Informatics, Columbia University, New York, NY, United States

[5]College of Computer and Information Science, Northeastern University, Boston, MA, United States

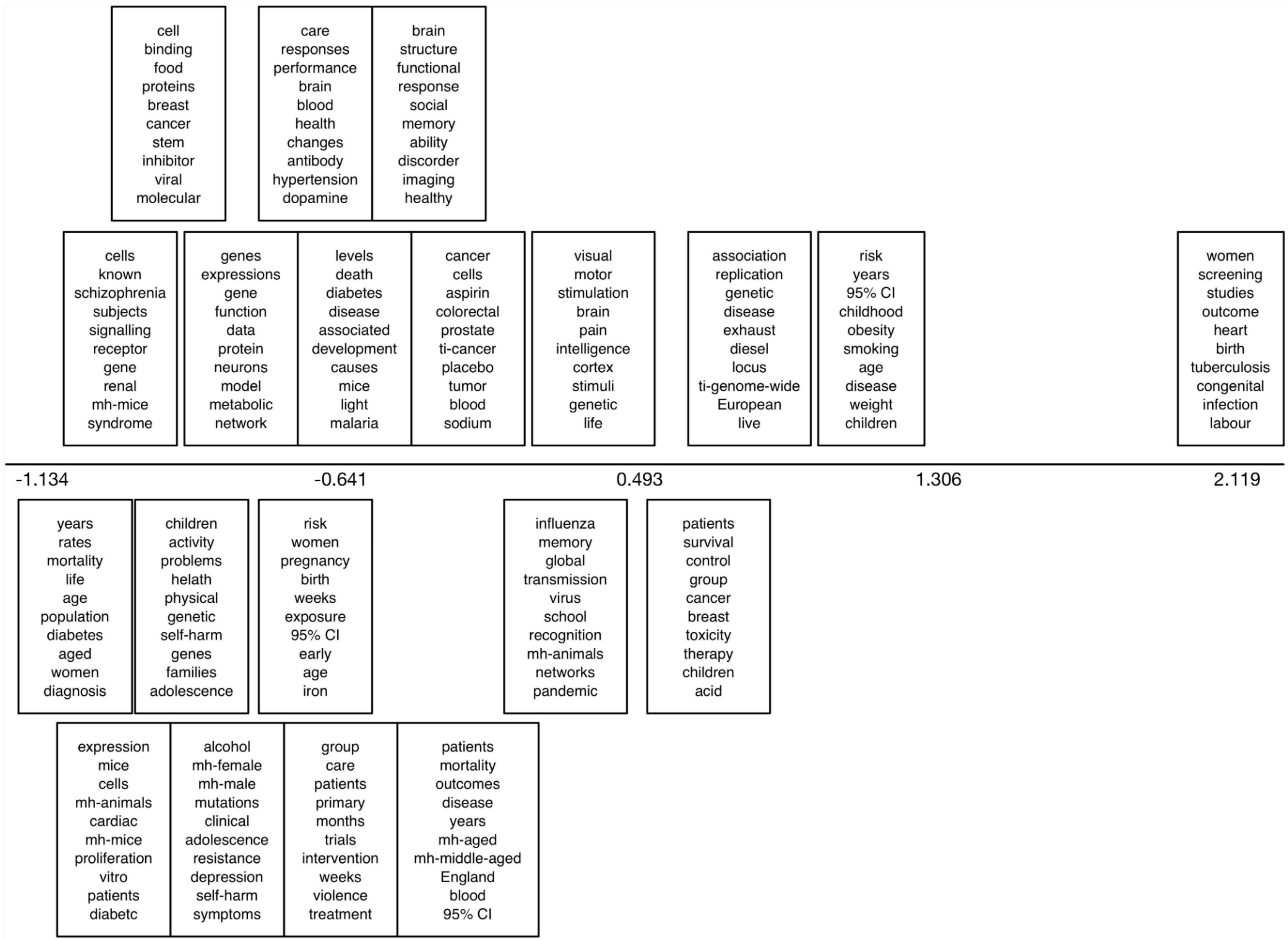


Figure 13. Top 10 most probable words in the topics uncovered by the supervised latent Dirichlet allocation model—again assuming 20 topics—fit to the Sumner news coverage dataset. mh: this prefix indicates a Medical Subject Headings (MeSH) term; ti: this prefix indicates a title term.

Extensions: Analyzing RateMDs ratings via “Factorial LDA”

**What Affects Patient (Dis)satisfaction?
Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model**

Michael J. Paul
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218
mpaul@cs.jhu.edu

Byron C. Wallace
Center for Evidence-based Medicine
Brown University
Providence, RI 02903
byron_wallace@brown.edu

Mark Dredze
Human Language Technology
Center of Excellence
Johns Hopkins University
Baltimore, MD 21211
mdredze@cs.jhu.edu

ratings			*review text*
5	5	5	Dr. *X* has a gentle and reassuring manner with the kids, her office staff is prompt, pleasant, responsive, and she seems very knowledgeable.
1	2	1	We were told outright that my wife, without question, did not have a uterine infection. She was discharged. 4 hours later she was very sick. We went back to triage and lo and behold, a uterine infection.

Table 1: A positive and negative review from our corpus. Ratings correspond to *helpfulness*, *staff* and *knowledgeability*, respectively; higher numbers convey positive sentiment.

Factors

Interpersonal manner		Technical competence		Systems issues	
positive	*negative*	*positive*	*negative*	*positive*	*negative*
shows empathy, professional, communicates well	poor listener, judgmental, racist	good decision maker, follows up on issues, knowledgeable	poor decision maker, prescribes the wrong medication, disorganized	friendly staff, short wait times, convenient location	difficult to park, rude staff, expensive

Table 2: Illustrative tags underneath the three main aspects identified in (López et al. 2012).

Factorial LDA

- We use f-LDA to model topic and sentiment
- Each (topic,sentiment) pair has a word distribution
- e.g. (Systems/Staff, Negative):

office
time
doctor
appointment
rude
staff
room
didn't
visit
wait

Factorial LDA

- We use f-LDA to model topic and sentiment
- Each (topic,sentiment) pair has a word distribution
- e.g. (Systems/Staff, Positive):

dr
time
staff
great
helpful
feel
questions
office
really
friendly

Factorial LDA

- We use f-LDA to model topic and sentiment
- Each (topic,sentiment) pair has a word distribution
- e.g. (Interpersonal, Positive):

dr
doctor
best
years
caring
care
patients
patient
recommend
family

- Why should the word distributions for pairs make any sense?

- Parameters are tied across the priors of each word distribution
  - The prior for (Systems,Negative) shares parameters with (Systems,Positive) which shares parameters with the prior for (Interpersonal,Positive)

	Systems		Positive			
exp(	staff	+	recommend	)	=	dr
	time		wonderful			time
	office		highly			staff
	questions		knowledgeable			great
	wait		professional			helpful
	helpful		kind			feel
	nice		great			doctor
	feel		dr			questions
	great		best			office
	appointment		helpful			friendly
	nurse		amazing			really

Systems

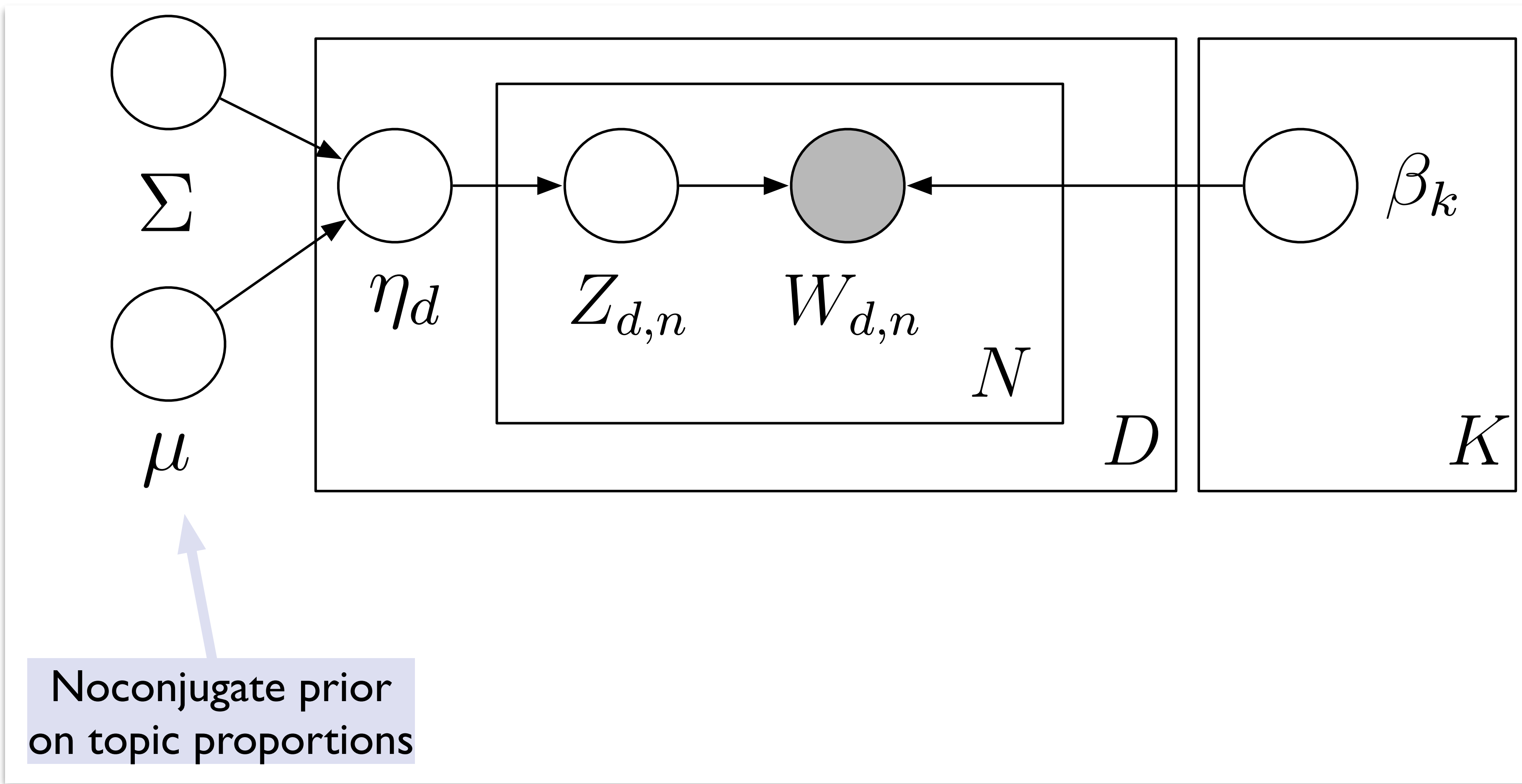
Positive

Positive		Prior
dr		dr
time	multinomial parameters sampled from Dirichlet	time
staff	←	staff
great		great
helpful		helpful
feel		feel
questions		doctor
office		questions
really		office
friendly		friendly
doctor		really

Table 2 Highest ranking (most probable) words for each aspect and polarity

Systems		Technical		Interpersonal	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Loves	Charged	Son	MRI	Excellent	Arrogant
Kids	Pharmacy	Gyn	Foot	Notch	Report
Awesome	Told	Delivered	Bleeding	Caring	Drug
Wonderful	Awful	Breast	Ray	Compassionate	Misdiagnosed
Love	Unprofessional	Thankful	Nerve	Highly	Reaction
Loved	Paying	Delivery	Hurt	Exceptional	Prescribed
Comfortable	Terrible	Ob	Bone	Best	License
Knowledgeable	Billed	Children	Antibiotic	Knowledgeable	Lack
Explains	Rude	Baby	Remove	Outstanding	Drugs

Extensions: Correlated Topic Model



Estimate a covariance matrix Σ that parameterizes correlations between topics in a document

Extensions: Dynamic Topic Models

1789

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

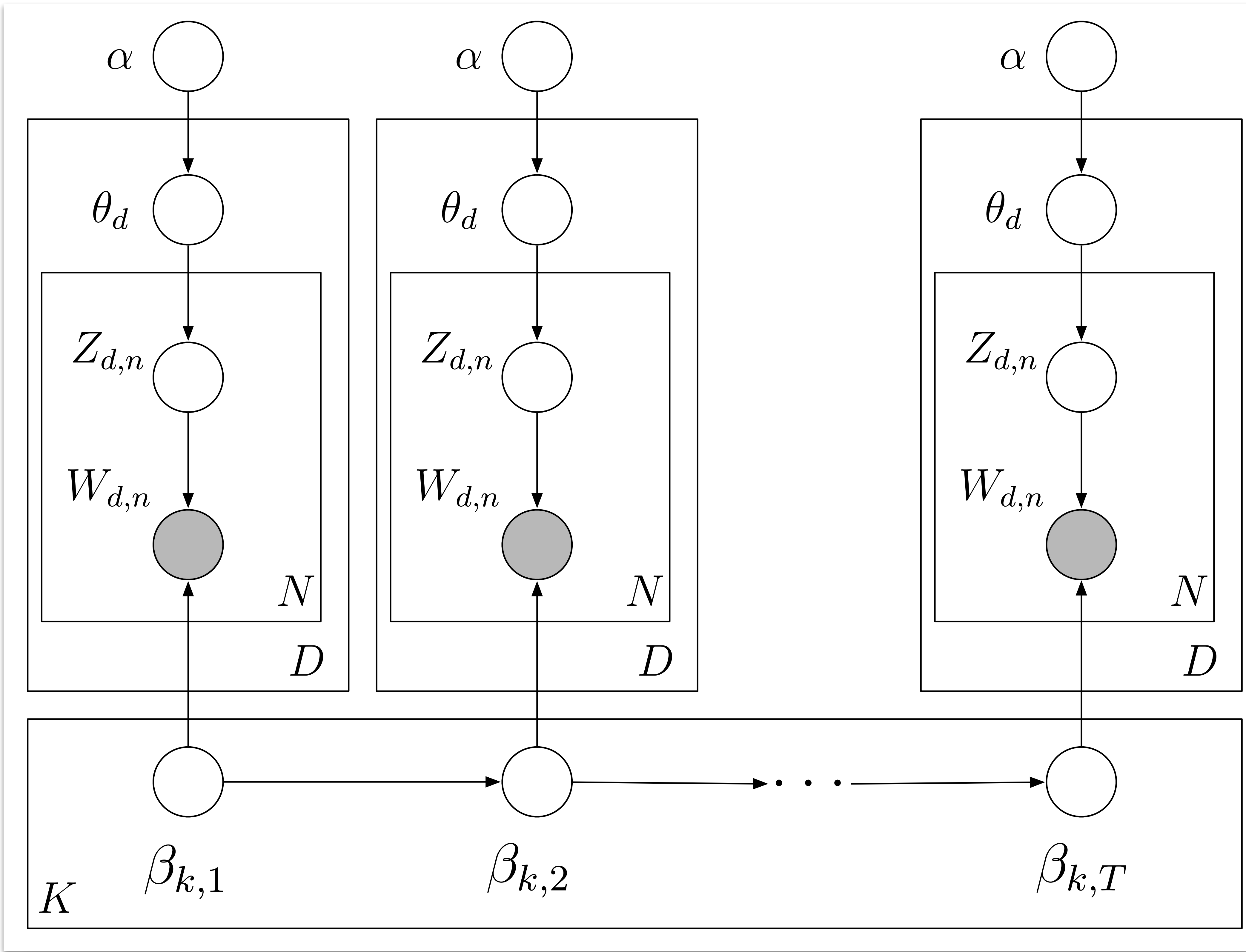
Inaugural addresses

2009

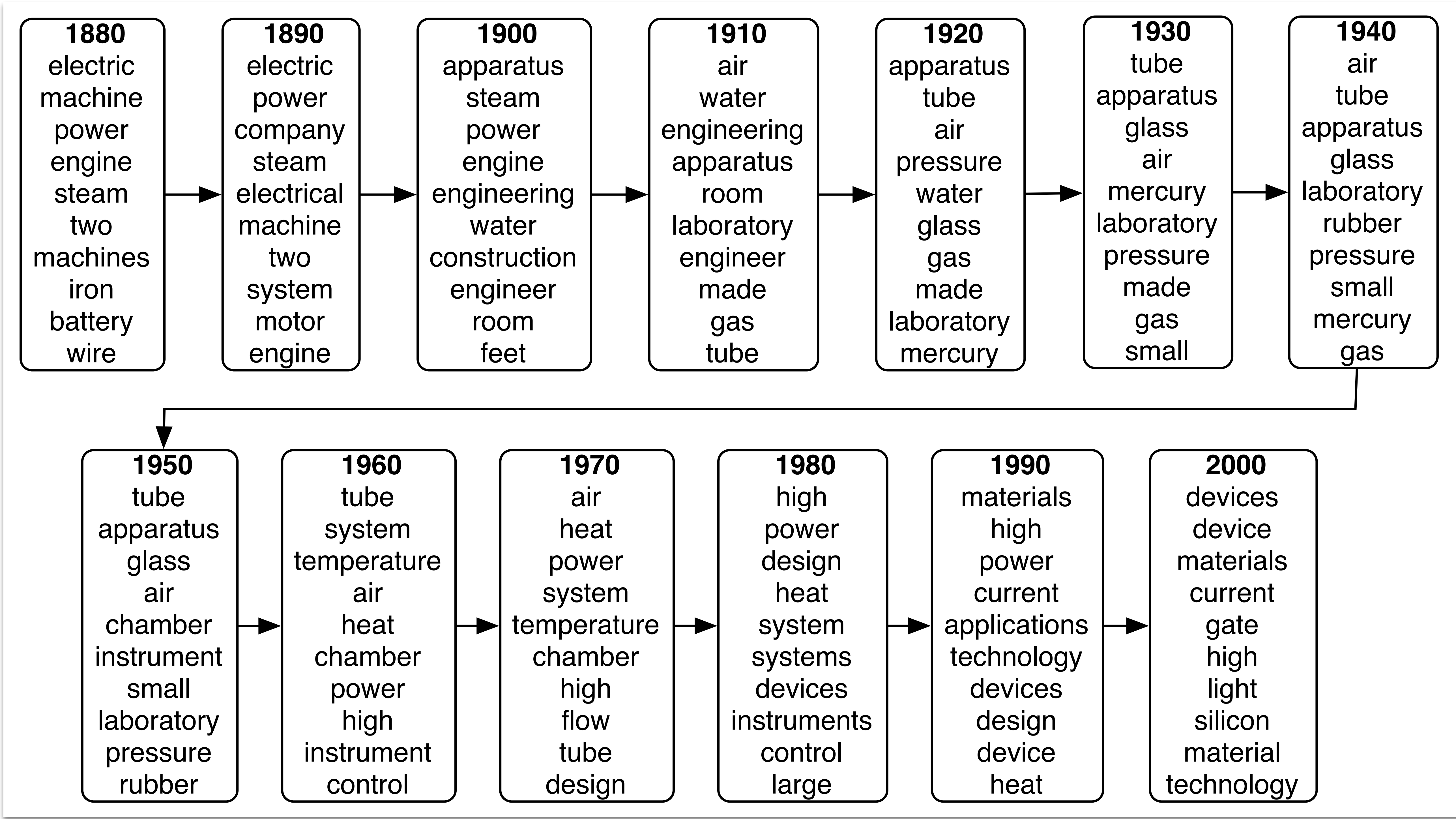
AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

Track changes in word distributions associated with a topic over time.

Extensions: Dynamic Topic Models



Extensions: Dynamic Topic Models



Summing up

- Latent Dirichlet Allocation (LDA) is a Bayesian topic model that is readily extensible

- To estimate parameters, we used a *sampling* based approach. General idea: draw samples of parameters and keep those that make the observed data likely

- *Gibbs* sampling is a particular variant of this approach, and draws individual parameters conditioned on all others