Data Mining Techniques

CS 6220 - Section 3 - Fall 2016

Lecture 11

Jan-Willem van de Meent
(credit: Yijun Zhao, Dave Blei)

PROJECT GUIDELINES

(*updated*)

Project Goals

- Select a dataset / prediction problem
- Perform exploratory analysis
  and preprocesssing
- Apply one or more algorithms
- Critically evaluate results
- Submit a report and present project

Proposals

- Due: 28 October
- Presentation:10+5 mins
- Proposal: 1-2 pages
- Describe
    - Dataset
    - Prediction task
    - Proposed methods

Presentation and Report

- Due: 2 December
- Presentation
  - 20 mins + 10 discussion
- Report
  - 8-10 pages, 11 pts
- Code

Presentation and Report

- Due: 2 December
- Presentation
    - 20 mins + 10 discussion
- Report
    - 8-10 pages, 11 pts
- Code

Grading

- Proposal: 15%
- Problem and Results: 20%
- Data and Code: 15%
- Report: 35%
- Presentation: 15%

Grading

- Problem and Results: 20%
  - Novelty of task
  - Own dataset vs UCI dataset
  - Number of algorithms tested
  - Novelty of algorithms

Grading

- Data and Code: 15%
  - Documentation and Readability
  - TAs should be able to run code
  - Reproducibility
    (can figures and tables be generated by running code?)

Grading

- Report: 35%
  - Exploratory analysis of data
    - Explain how properties of data relate to choice of algorithm
  - Description of algorithms and methodology
  - Discussion of results
    - Which methods work well, which do not, and why?
    - Comparison to state of art?

Example: Minimum Viable Project

- Get 2-3 datasets from UCI repository
- Figure out what pre-processing (if any) is needed
- Run every applicable algorithm in scikit learn
- Explain which algorithms work well on which datasets and why

Example: More Ambitious Projects

- Find a new dataset or define a novel task (*i.e.* not classification or clustering)
- Attack a problem from a Kaggle competition
- Implement a recently published method (talk to me for suggestions)

Homework Updates

- HW3 now due on 2 November (after midterm and proposals)
- Removed HW5 to give more time to work on projects

MIDTERM REVIEW

List of Topics for Midterm

http://www.ccs.neu.edu/course/cs6220f16/sec3/midterm-topics.html

Northeastern University
College of Computer and Information Science

CS6220 - Fall 2016 - Section 3 - Data Mining Techniques

MIDTERM TOPIC LIST

LINEAR REGRESSION

- Problem definition
- Ordinary Least Squares, Pseudo-inverse
  - Implementation
  - Computational complexity

- Everything up until last Friday
  (expect final to emphasize later topics)
- Open book, focus on understanding

BINOMIAL MIXTURES

Mixture of Binomials

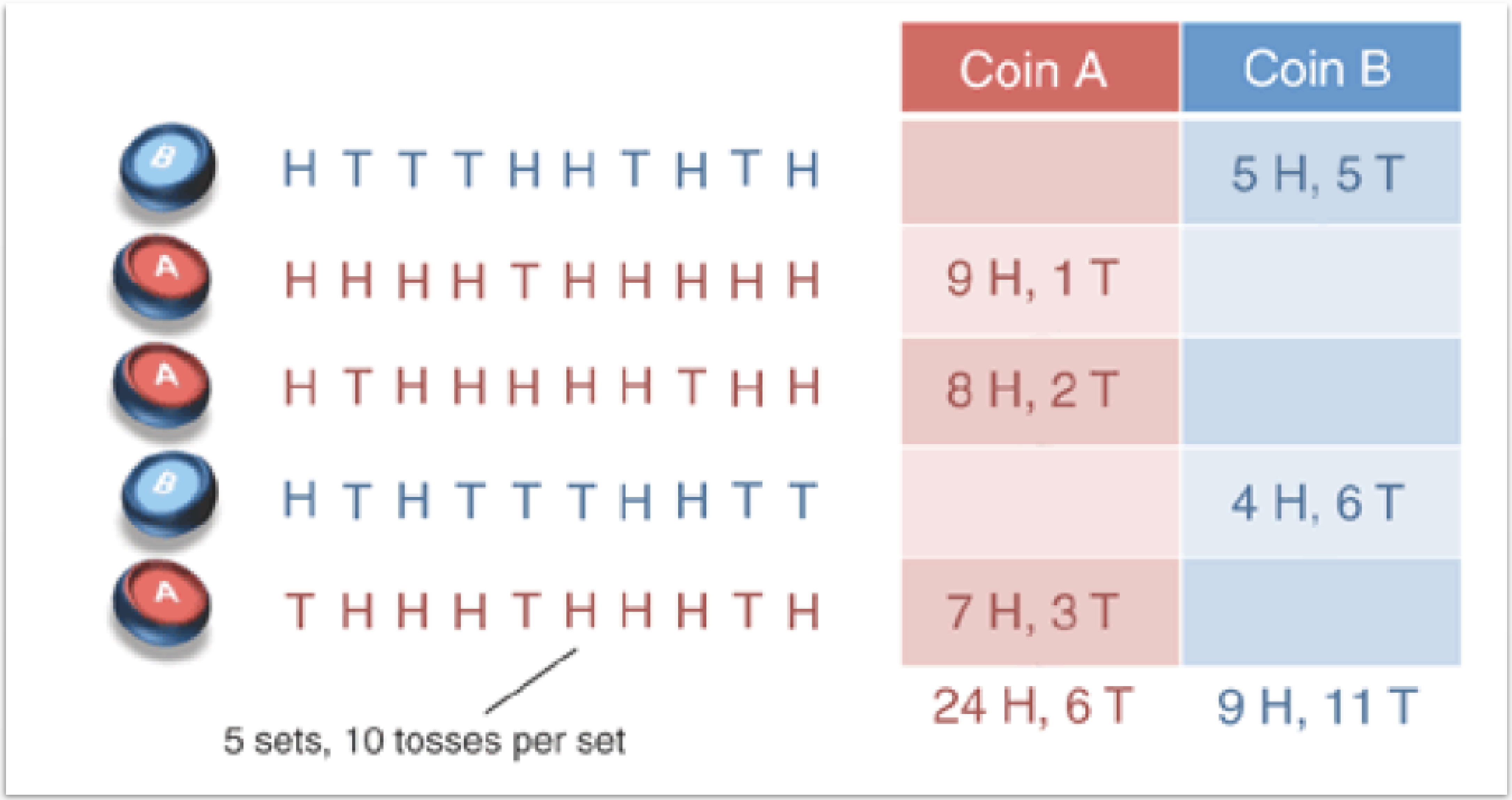
Suppose we have two coins A and B (weighted).
We want to estimate the bias of the two coins. i.e.

$$p_A(head) = \mu_A$$

$$p_B(head) = \mu_B$$

- Pick a coin at random
  (simplified version, a equal mixture)
- Flip 10 times and record 'H' and 'T'
- repeat the process until we have a good size of training data

Mixture of Binomials

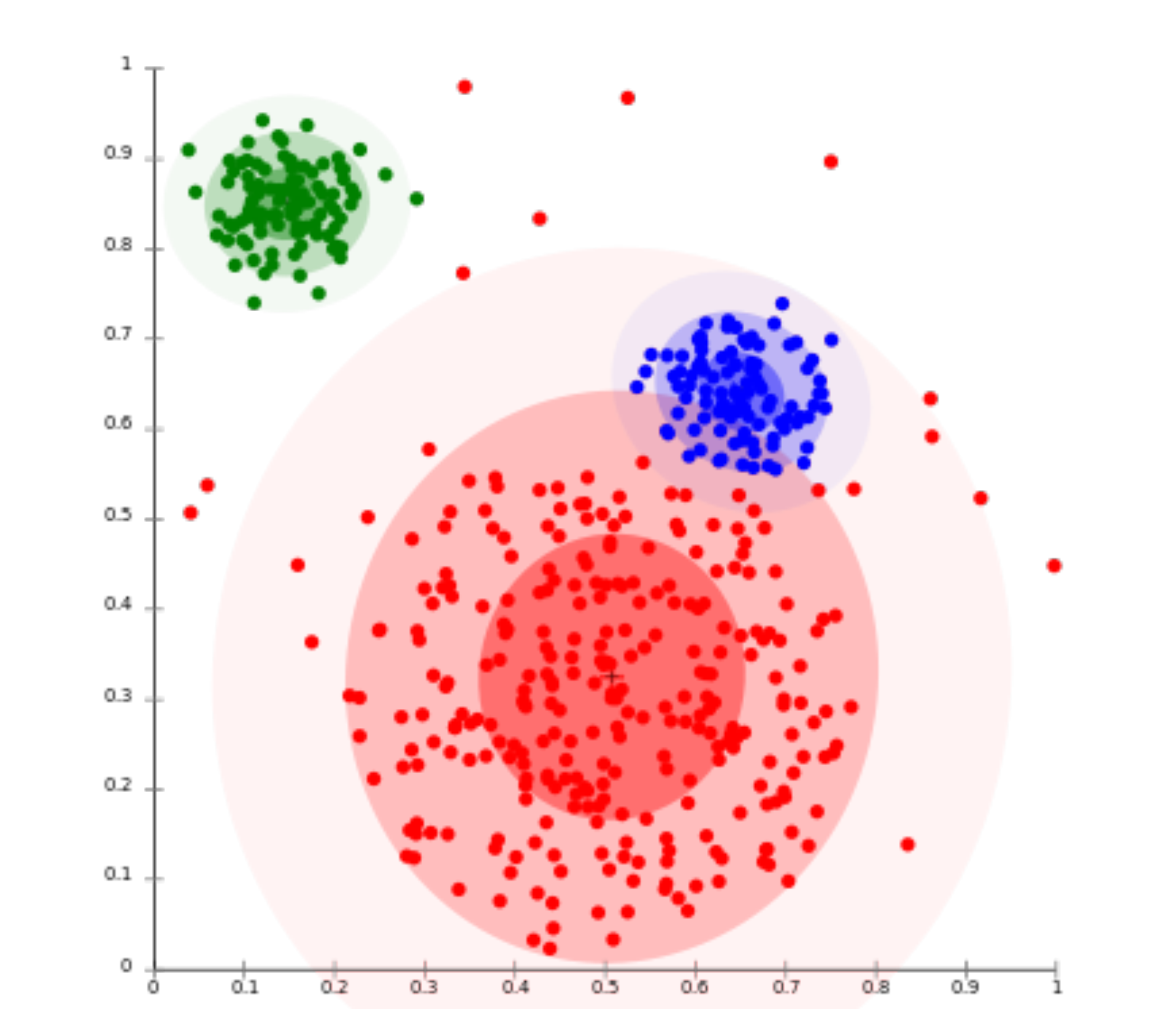


$$\text{Binomial}(m\,|, M, \mu) = \binom{M}{m}\mu^m(1-\mu)^{M-m}$$

Gaussian Mixture Model

Generative Model

$$z_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Expectation Maximization

Initialize $\boldsymbol{\theta}$

Repeat until convergence

1. Expectation Step

$$q^i(\boldsymbol{z}) = \underset{q(\boldsymbol{z})}{\text{argmax}}\, \mathcal{L}(q(\boldsymbol{z}), \boldsymbol{\theta}^{i-1})$$

2. Maximization Step

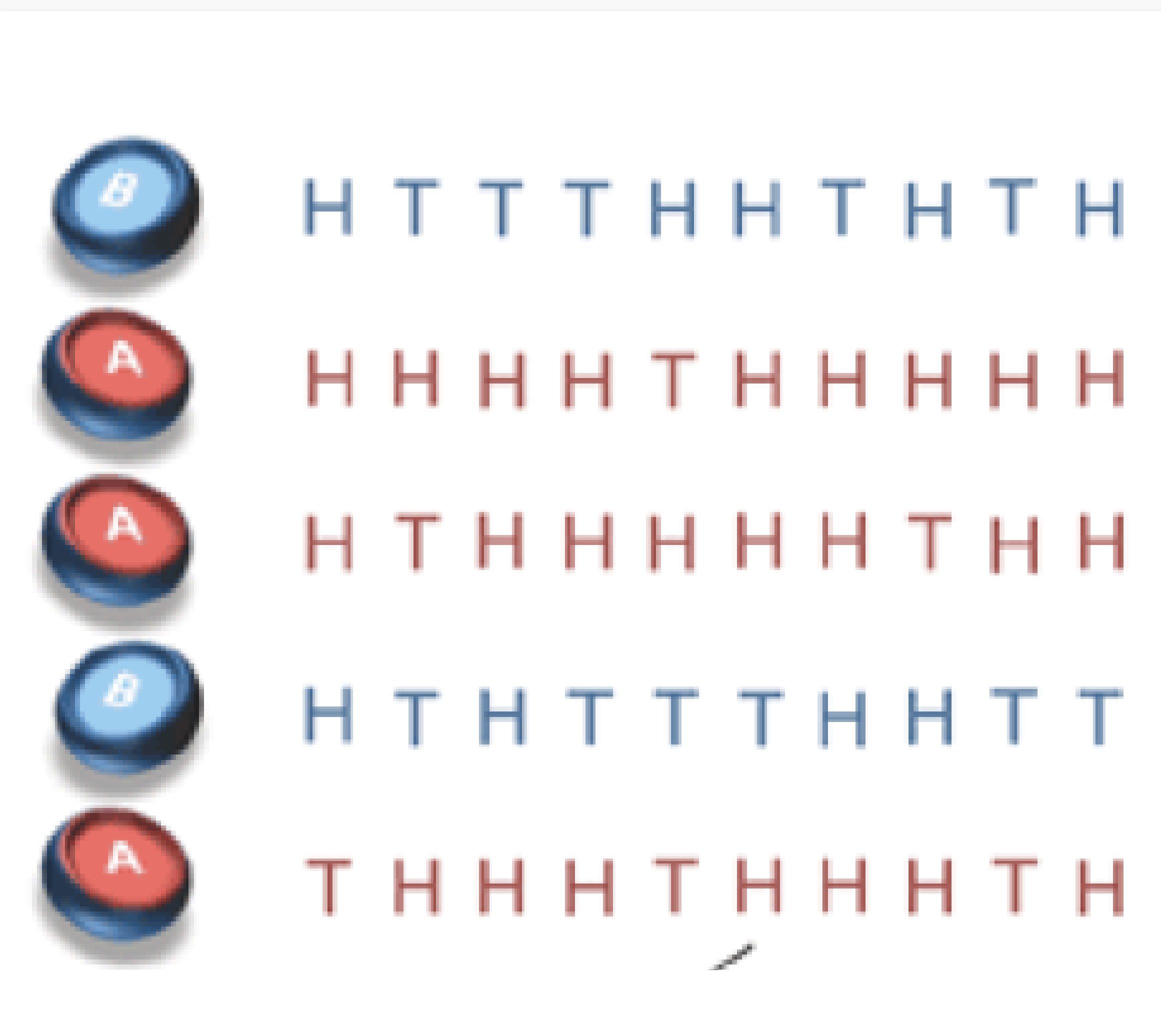
$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \mathcal{L}(q^i(\boldsymbol{z}), \boldsymbol{\theta})$$

$$\mathcal{L}(q(\boldsymbol{z}), \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{z} \,|\, \boldsymbol{\theta})}{q(\boldsymbol{z})}$$

Binomial Mixture Model

Generative Model

$$z_n \sim \text{Discrete}(\pi)$$
$$x_n \,|\, z_n = k \sim \text{Binomial}(\mu_k, M)$$



Expectation Maximization

Initialize $\boldsymbol{\theta}$

Repeat until convergence

1. Expectation Step

$$q^i(\boldsymbol{z}) = \underset{q(\boldsymbol{z})}{\text{argmax}}\, \mathscr{L}(q(\boldsymbol{z}), \boldsymbol{\theta}^{i-1})$$

2. Maximization Step

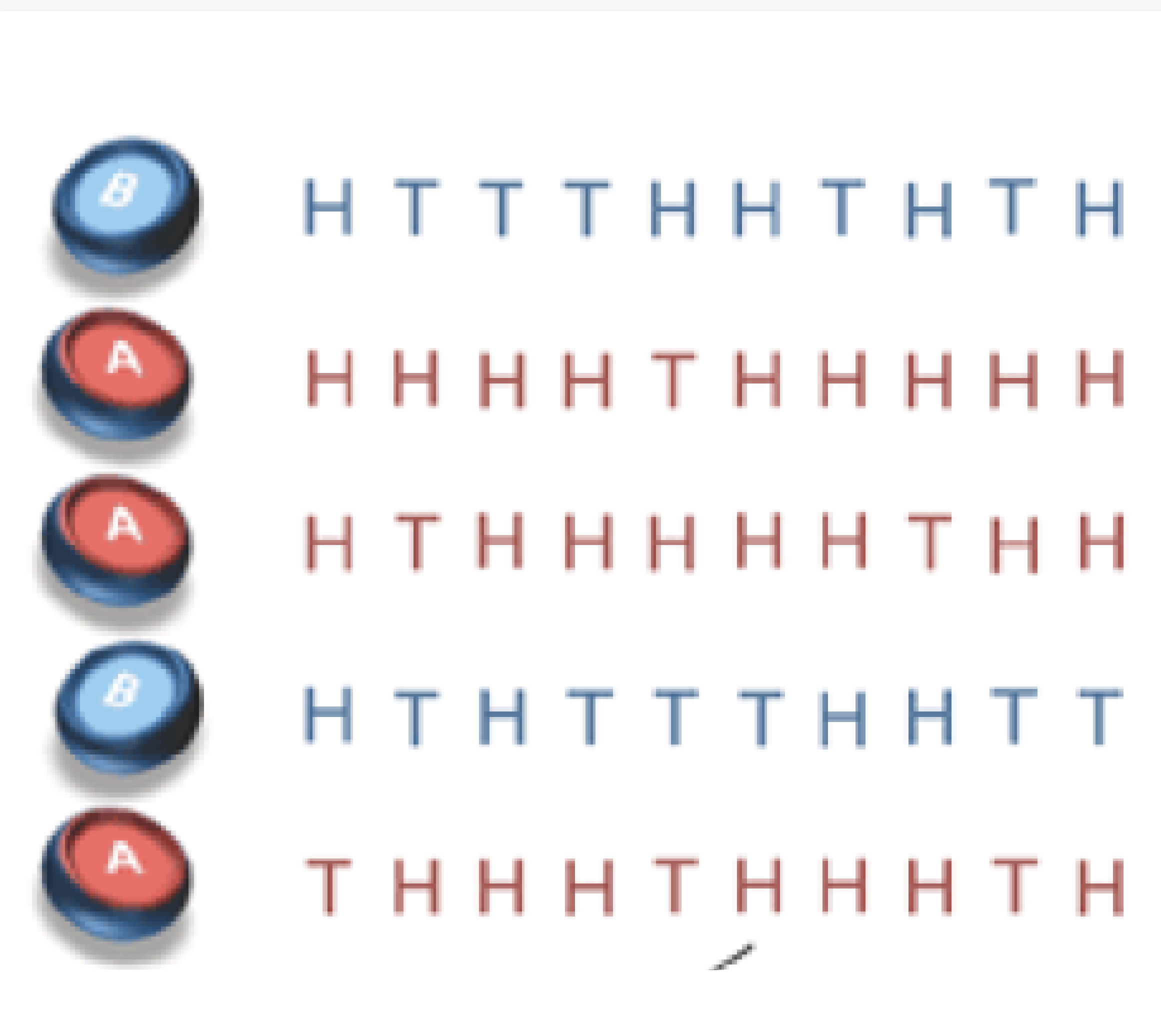
$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \mathscr{L}(q^i(\boldsymbol{z}), \boldsymbol{\theta})$$

$$\mathscr{L}(q(\boldsymbol{z}), \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{z} \,|\, \boldsymbol{\theta})}{q(\boldsymbol{z})}$$

Binomial Mixture Model

Generative Model

$$z_n \sim \text{Discrete}(\pi)$$
$$x_n \,|\, z_n = k \sim \text{Binomial}(\mu_k, M)$$



Expectation Maximization

Initialize $\boldsymbol{\theta}$

Repeat until convergence

1. Expectation Step

$$\gamma^i_{nk} = p(z_n = k \,|\, \boldsymbol{\mu}^{i-1}, \boldsymbol{\pi}^{i-1})$$

2. Maximization Step

$$\mu^i_k = \frac{1}{N^i_k} \sum_{n=1}^{N} \gamma^i_{nk} \frac{x_n}{M}$$
$$\pi^i_k = N^i_k / N$$
$$N^i_k = \sum_{n=1}^{N} \gamma^i_{nk}$$

TOPIC MODELS

Borrowing from:
David Blei
(Columbia)

Review: Naive Bayes

Features: Words in E-mail

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Labels: Spam or not Spam

$$y_n \in \{0, 1\}$$

Generative Model

$$y_n \sim \text{Bernoulli}(\mu)$$

$$\boldsymbol{x}_{nd} \mid y_n = k \sim \text{Bernoulli}(\phi_{kd})$$

Maximum Likelihood

$$\mu = \frac{1}{N} \sum_{n=1}^{N} I[y_n = 1]$$

$$\phi_{kd} = \frac{1}{N_k} \sum_{n:y_n=k} I[x_{nd} = 1]$$

Review: Naive Bayes

Features: Words in E-mail

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{array}$$

Labels: Spam or not Spam

$$y_n \in \{0, 1\}$$

Generative Model (with prior)

$$\begin{aligned} \mu &\sim \text{Beta}(1,1) \\ \phi_{kd} &\sim \text{Beta}(1,1) \\ y_n &\sim \text{Bernoulli}(\mu) \\ x_{nd} \,|\, y_n = k &\sim \text{Bernoulli}(\phi_{kd}) \end{aligned}$$

Posterior Mean for Parameters

$$\mu^*, \boldsymbol{\phi}^* = \mathbb{E}_{p(\mu, \boldsymbol{\phi} \,|\, \boldsymbol{x}_{1:N}, y_{1:N})}[\mu, \boldsymbol{\phi}]$$

$$\mu^* = \frac{N_1 + 1}{N + 2} \qquad \phi^*_{kd} = \frac{N_{kd} + 1}{N_k + 2}$$

Mixtures of Documents

Observations: Bag of Words

$$
\boldsymbol{x}_d = \begin{bmatrix} \mathbf{24} \\ \mathbf{1} \\ 0 \\ \vdots \\ \mathbf{4} \\ \vdots \\ 0 \end{bmatrix} \quad \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}
$$

Clusters: Types of Documents

$$
z_d \in \{1, \ldots, K\} \quad d = 1, \ldots, D
$$

Mixtures of Documents

Observations: Bag of Words

$$\boldsymbol{x}_d = \begin{bmatrix} \mathbf{24} \\ \mathbf{1} \\ 0 \\ \vdots \\ \mathbf{4} \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Clusters: Types of Documents

$$z_d \in \{1, \ldots, K\} \quad d = 1, \ldots, D$$

Generative Model (with prior)

$$\begin{aligned} \mu &\sim \text{Beta}(1,1) \\ \phi_{kd} &\sim \text{Beta}(1,1) \\ y_n &\sim \text{Bernoulli}(\mu) \\ x_{nd} \mid y_n = k &\sim \text{Bernoulli}(\phi_{kd}) \end{aligned}$$

How should we modify the generative model?

Mixtures of Documents

Observations: Bag of Words

$$\boldsymbol{x}_d = \begin{bmatrix} \mathbf{24} \\ \mathbf{1} \\ 0 \\ \vdots \\ \mathbf{4} \\ \vdots \\ 0 \end{bmatrix} \quad \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

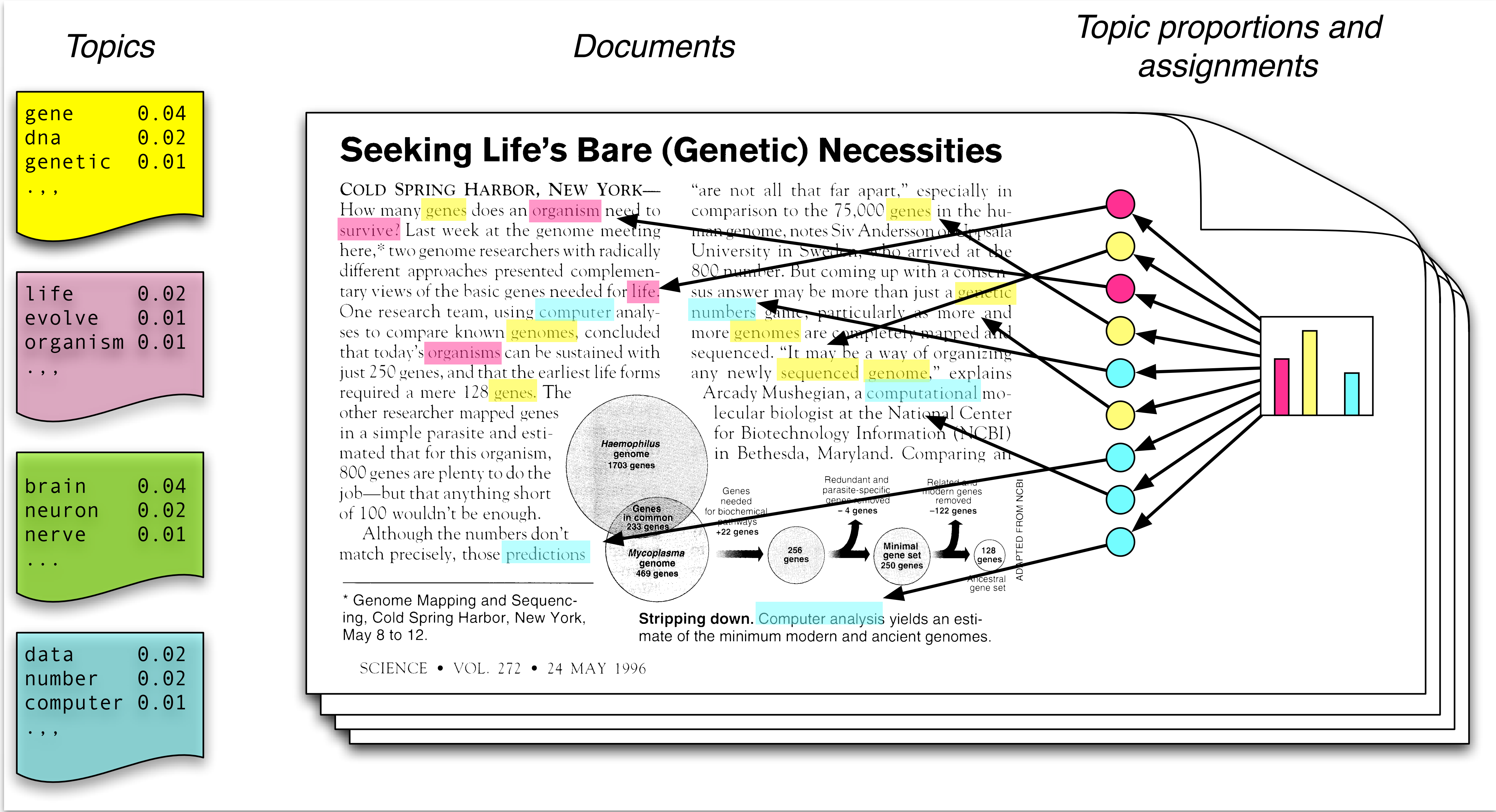
Generative Model (with prior)

$$\begin{aligned} \boldsymbol{\theta} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{\beta}_k &\sim \text{Dirichlet}(\boldsymbol{\eta}) \\ z_d &\sim \text{Discrete}(\boldsymbol{\theta}) \\ \boldsymbol{x}_d \,|\, z_d = k &\sim \text{Mult}(\boldsymbol{\beta}_k, N_d) \end{aligned}$$

Clusters: Types of Documents

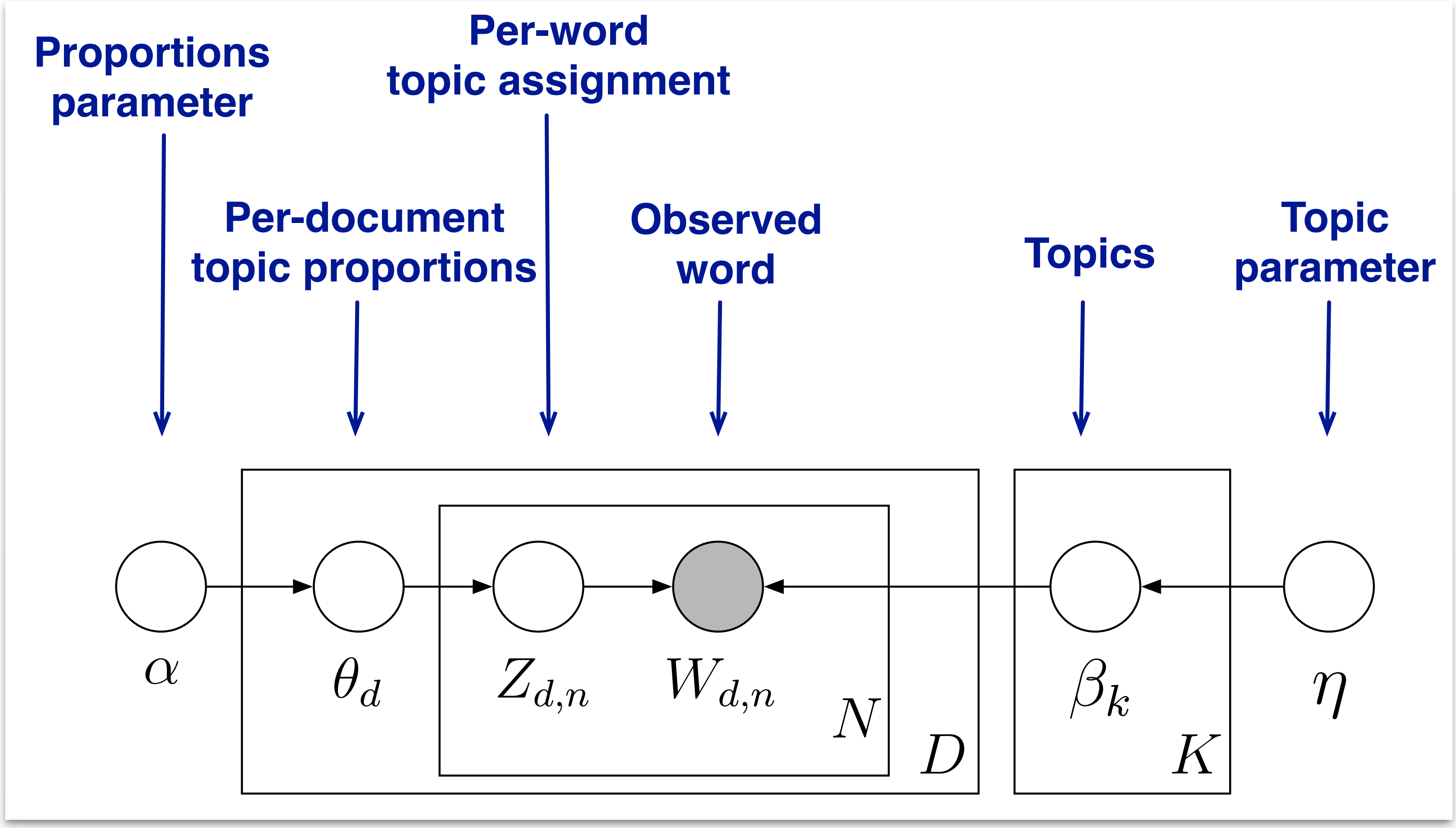
$$z_d \in \{1, \ldots, K\} \quad d = 1, \ldots, D$$

Topic Modeling



- *Naive Bayes*: Documents belong a class
- *Topic Models*: Words belong to a class

Latent Dirichlet Allocation



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

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

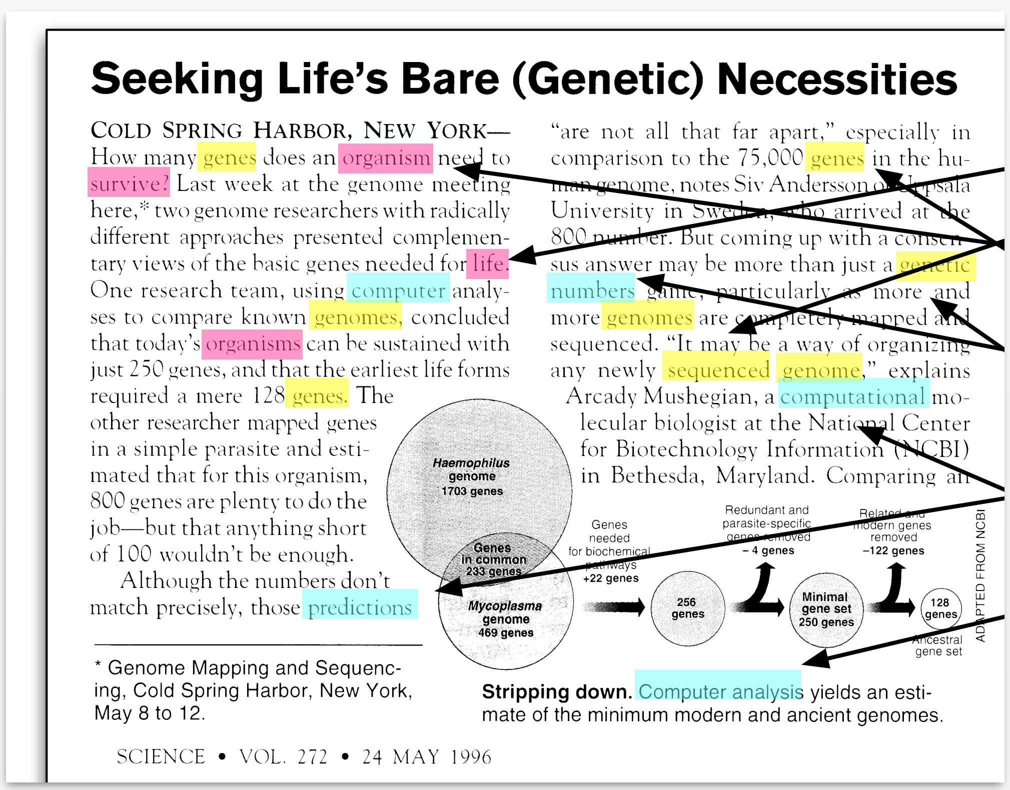
$$Z_{d,n} \sim \text{Discrete}(\theta_d) \quad n = 1, \ldots, N_d$$

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

PLSI/PLSA: EM for LDA

Generative Model (no priors)

$$Z_{d,n} \sim \text{Discrete}(\theta_d)$$
$$W_{d,n} | Z_{d,n} = k \sim \text{Discrete}(\beta_k)$$



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 all

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

Expectation Step

$$\gamma^i_{d,n,k} = p(Z_{d,n} = k \,|\, \boldsymbol{\theta}^{i-1}, \boldsymbol{\beta}^{i-1})$$

Maximization Step

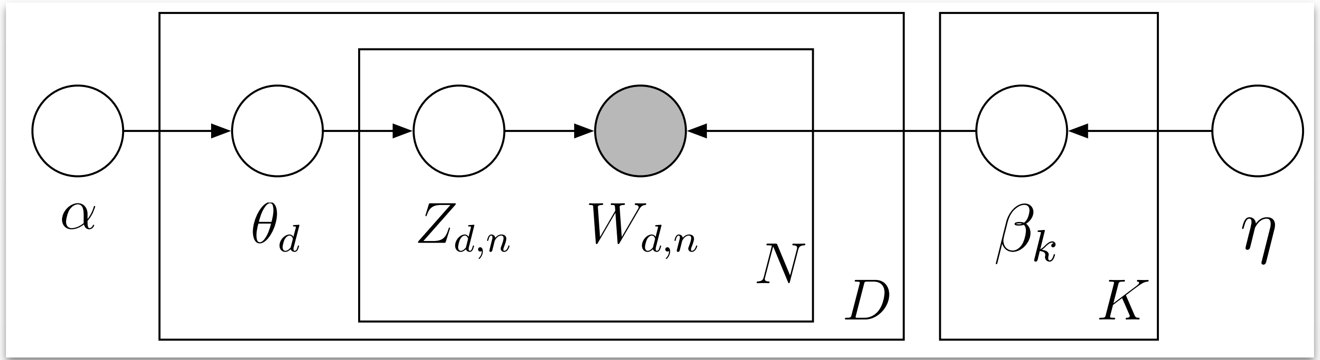
$$\boldsymbol{\beta}^i_{k,w} = \frac{1}{N_k} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \gamma^i_{d,n,k} I[W_{d,n} = w]$$

$$\boldsymbol{\theta}^i_{d,k} = \frac{1}{N_d} \sum_{n=1}^{N_d} \gamma_{d,n,k}$$

Variational Inference for LDA (sketch)

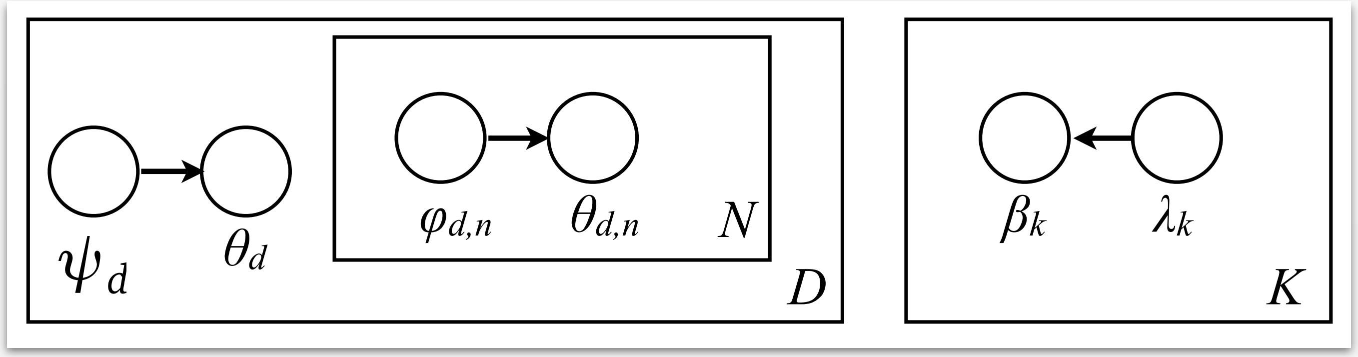
Generative Model

$$
\begin{aligned}
\beta_k &\sim \text{Dirichlet}(\eta) \\
\theta_d &\sim \text{Dirichlet}(\alpha) \\
Z_{d,n} &\sim \text{Discrete}(\theta_d) \\
W_{d,n} | Z_{d,n} = k &\sim \text{Discrete}(\beta_k)
\end{aligned}
$$



Variational Approximation

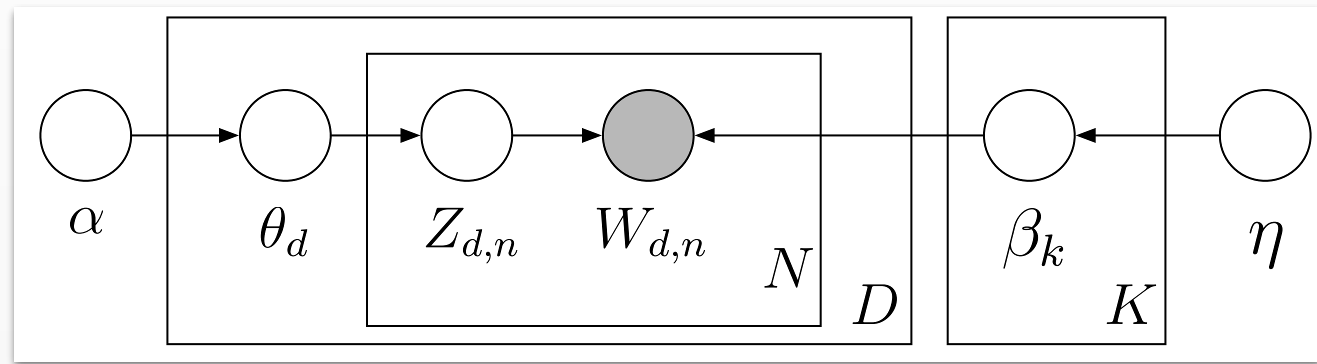
$$
\begin{aligned}
\beta_k &\sim \text{Dirichlet}(\lambda_k) \\
\theta_{d,n} &\sim \text{Dirichlet}(\phi_{d,n}) \\
\theta_d &\sim \text{Dirichlet}(\psi_d)
\end{aligned}
$$



Variational Inference for LDA (sketch)

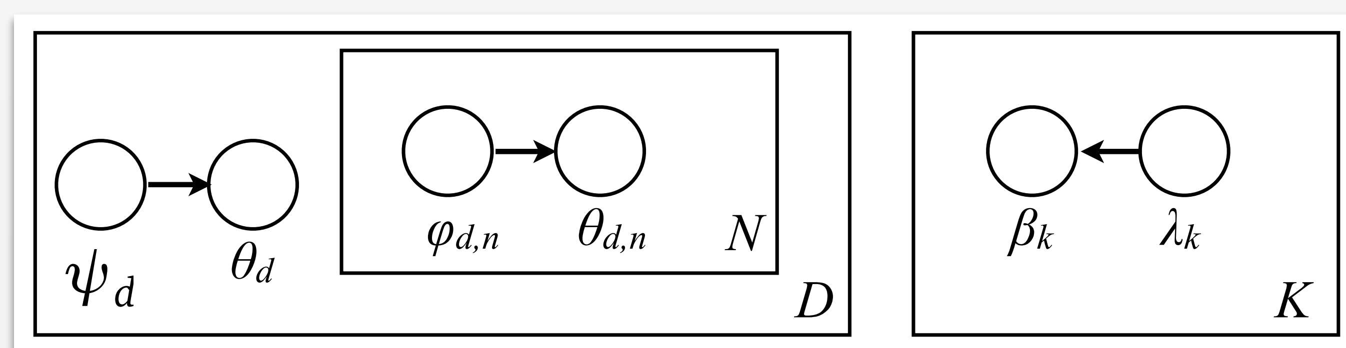
Generative Model

$$\beta_k \sim \text{Dirichlet}(\eta)$$
$$\theta_d \sim \text{Dirichlet}(\alpha)$$
$$Z_{d,n} \sim \text{Discrete}(\theta_d)$$
$$W_{d,n} | Z_{d,n} = k \sim \text{Discrete}(\beta_k)$$



Variational Approximation

$$\beta_k \sim \text{Dirichlet}(\lambda_k)$$
$$\theta_{d,n} \sim \text{Dirichlet}(\phi_{d,n})$$
$$\theta_d \sim \text{Dirichlet}(\psi_d)$$



Variational Inference for LDA (sketch)

One iteration of mean field variational inference for LDA

(1) For each topic k and term v:

$$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)}.$$

(2) For each document d:

(a) Update ψ_d

$$\psi_{d,k}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} \phi_{d,n,k}^{(t)}.$$

(b) For each word n, update $\vec{\phi}_{d,n}$:

)
$$\phi_{d,n,k}^{(t+1)} \propto \exp\left\{\Psi(\psi_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\textstyle\sum_{v=1}^{V} \lambda_{k,v}^{(t+1)})\right\},$$

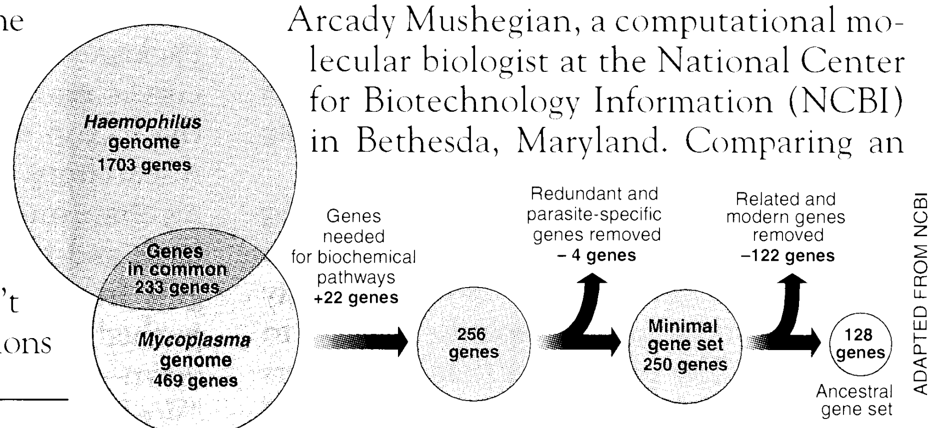
where Ψ is the digamma function, the first derivative of the $\log \Gamma$ function.

Example Inference

Seeking Life's Bare (Genetic) Necessities

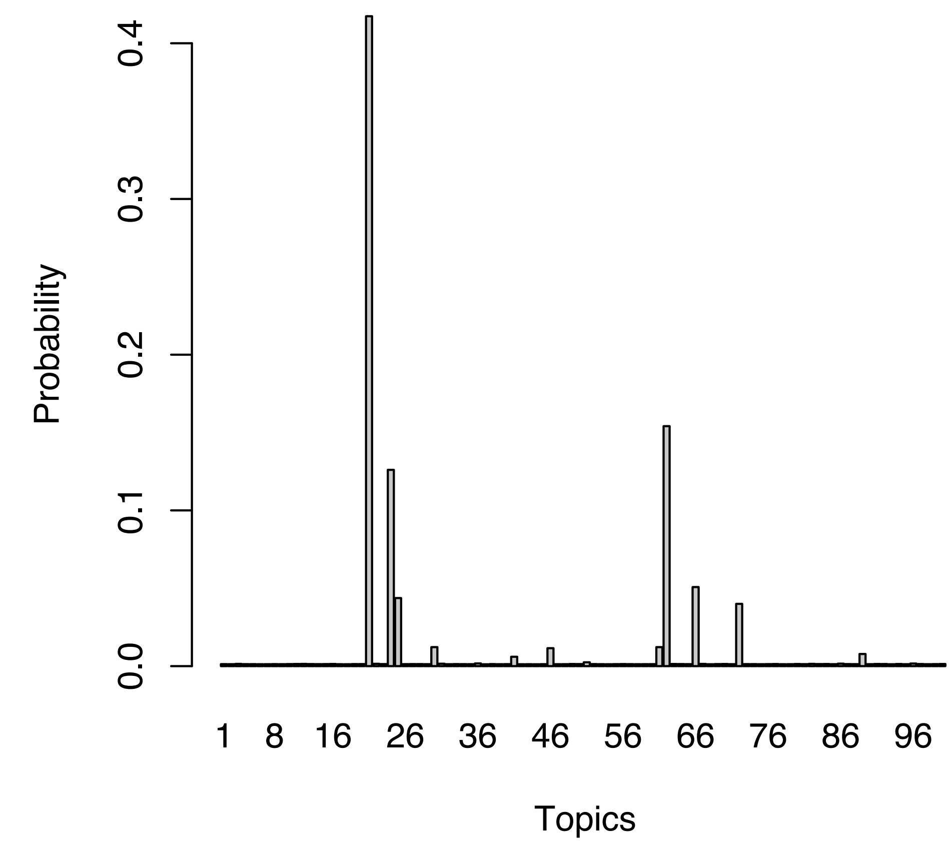
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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.



Example Inference

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Example Inference

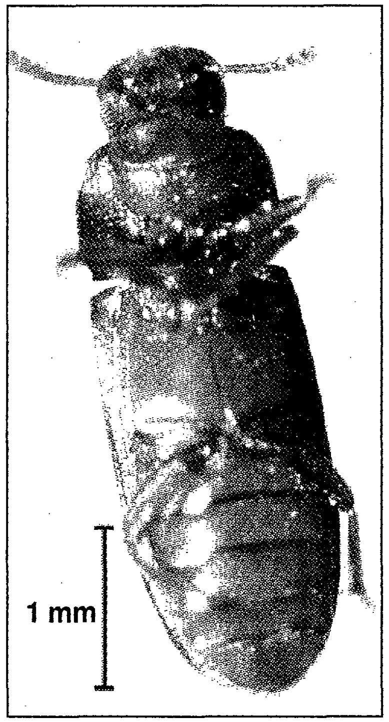
Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-



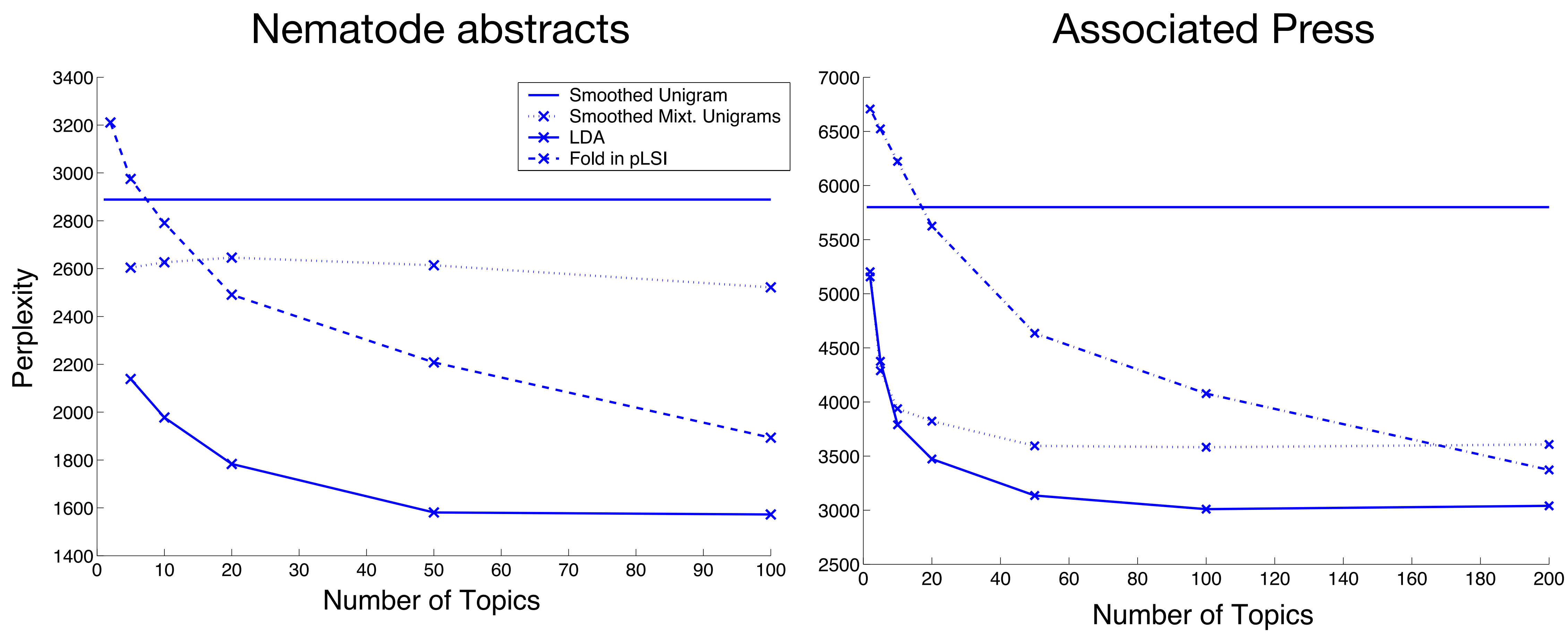
Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

Example Inference

problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

Performance Metric: Perplexity



$$\text{perplexity} = \exp\left\{\frac{-\sum_d \log p(\mathbf{w}_d)}{\sum_d N_d}\right\}$$

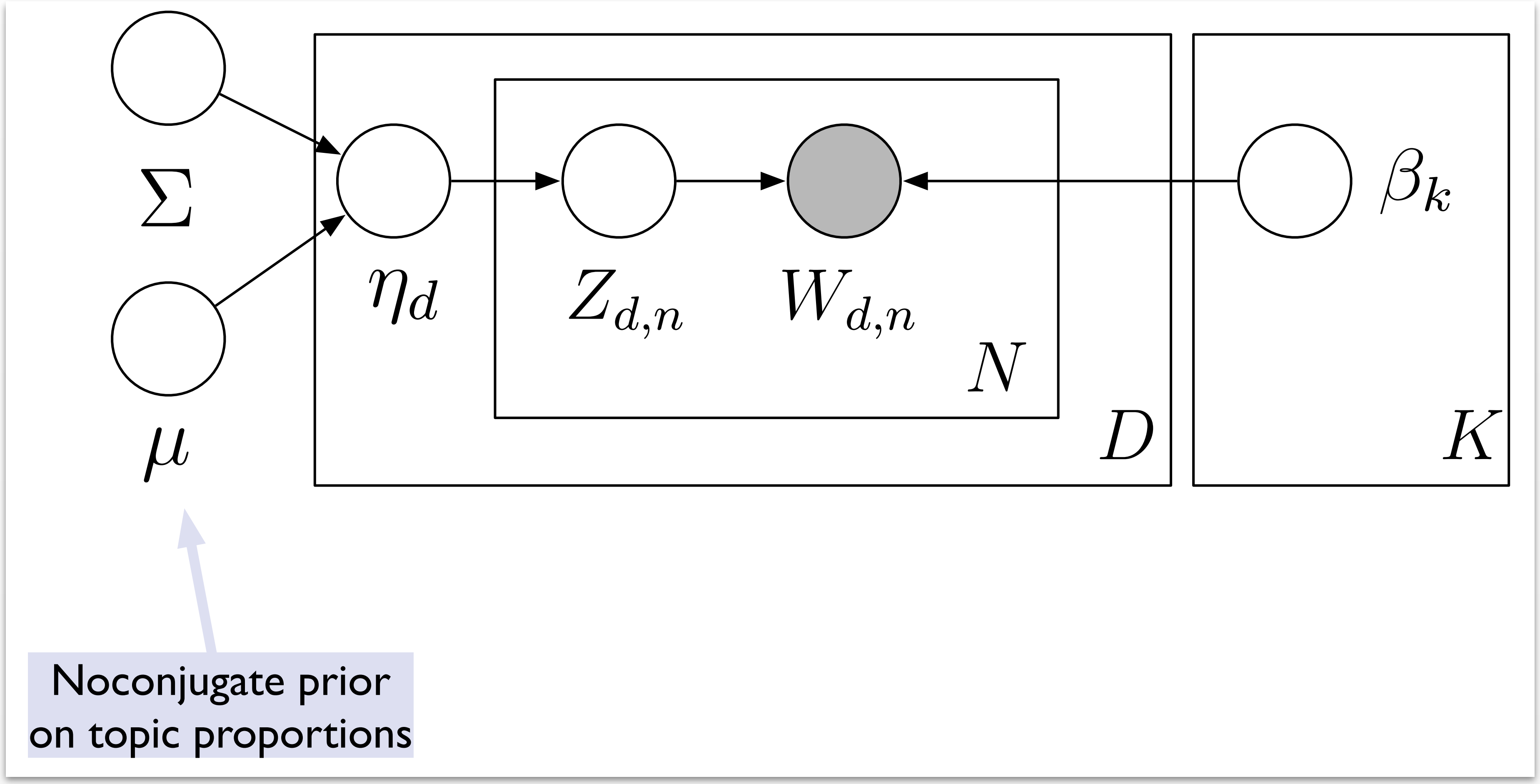
Marginal likelihood (evidence) of held out documents

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 (LDA) on most data
- Reason for popularity of LDA:
  can be embedded in more complicated models

Extensions: Correlated Topic Model



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

Extensions: Dynamic Topic Models

1789

Inaugural addresses

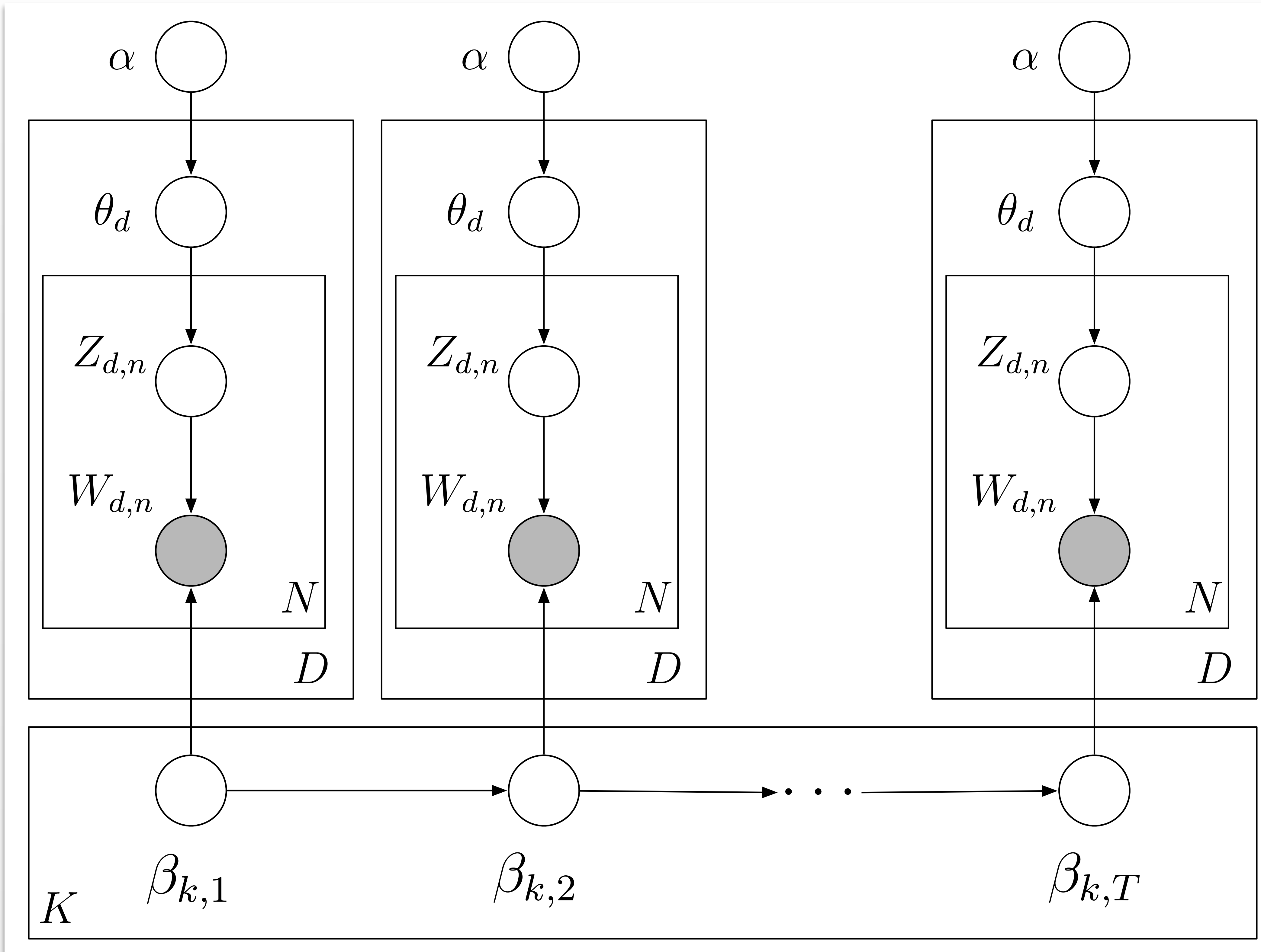
2009

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...

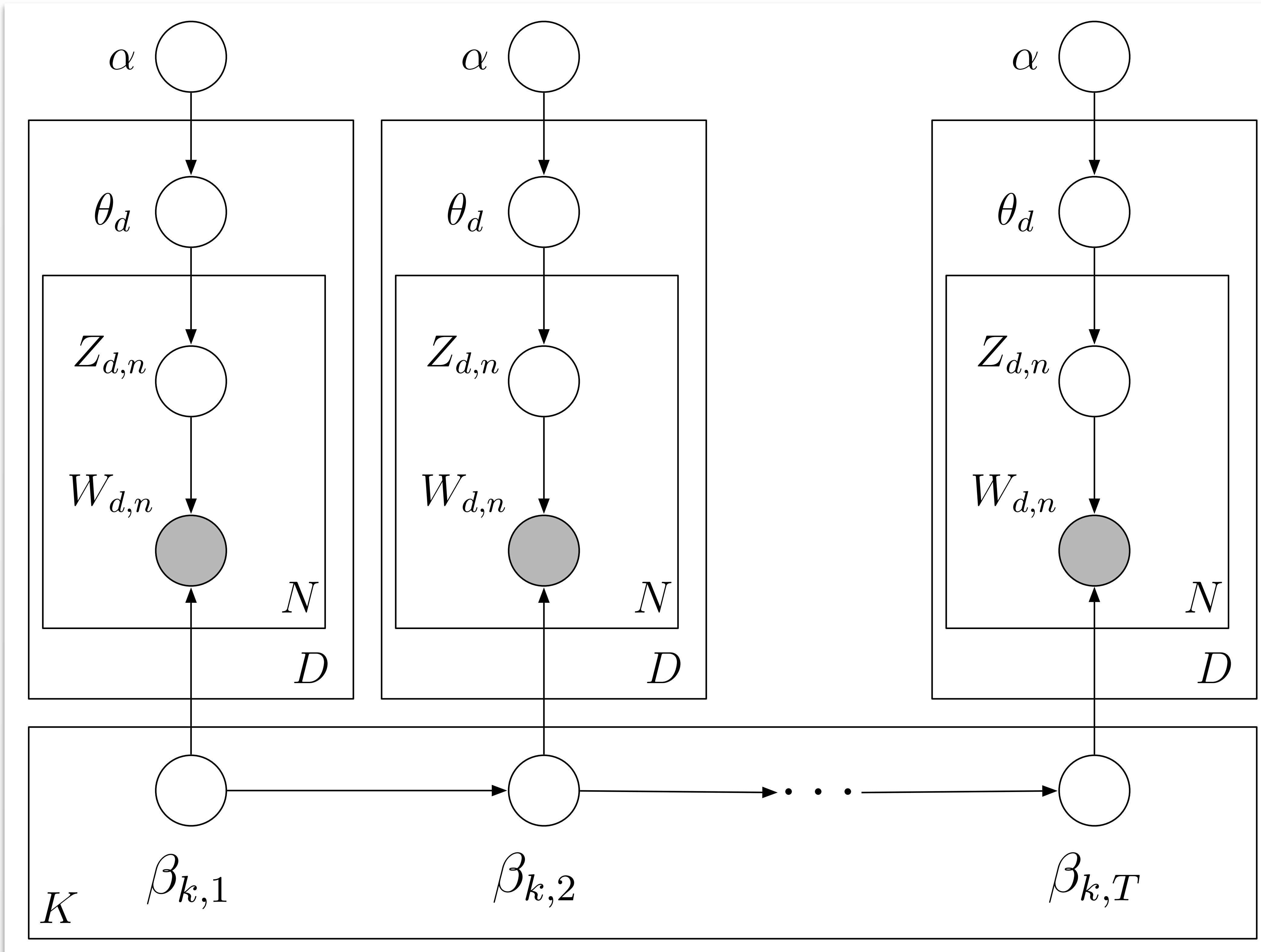
AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

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

Extensions: Dynamic Topic Models



Extensions: Dynamic Topic Models

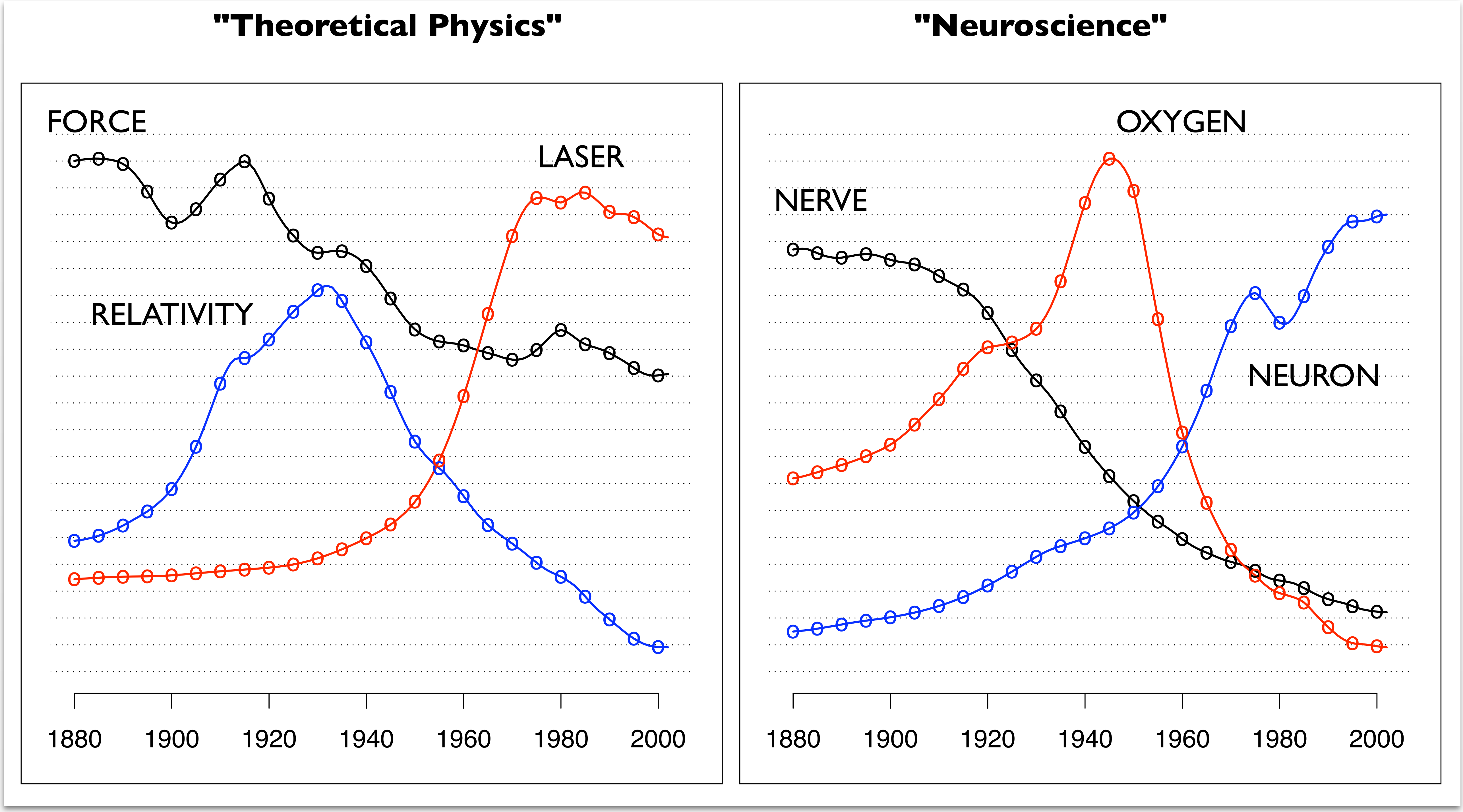


Extensions: Dynamic Topic Models

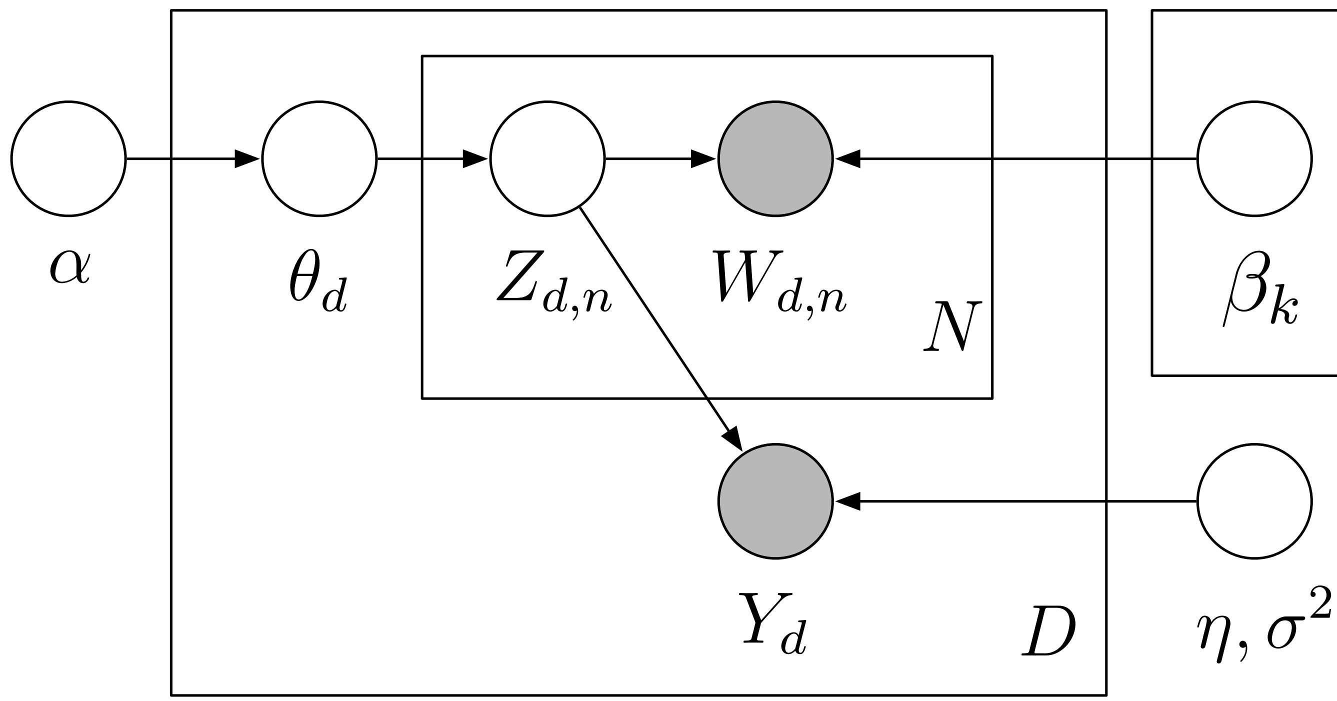
1880	1890	1900	1910	1920	1930	1940
electric	electric	apparatus	air	apparatus	tube	air
machine	power	steam	water	tube	apparatus	tube
power	company	power	engineering	air	glass	apparatus
engine	steam	engine	apparatus	pressure	air	glass
steam	electrical	engineering	room	water	mercury	laboratory
two	machine	water	laboratory	glass	laboratory	rubber
machines	two	construction	engineer	gas	pressure	pressure
iron	system	engineer	made	made	made	small
battery	motor	room	gas	laboratory	gas	mercury
wire	engine	feet	tube	mercury	small	gas

1950	1960	1970	1980	1990	2000
tube	tube	air	high	materials	devices
apparatus	system	heat	power	high	device
glass	temperature	power	design	power	materials
air	air	system	heat	current	current
chamber	heat	temperature	system	applications	gate
instrument	chamber	chamber	systems	technology	high
small	power	high	devices	devices	light
laboratory	high	flow	instruments	design	silicon
pressure	instrument	tube	control	device	material
rubber	control	design	large	heat	technology

Extensions: Dynamic Topic Models



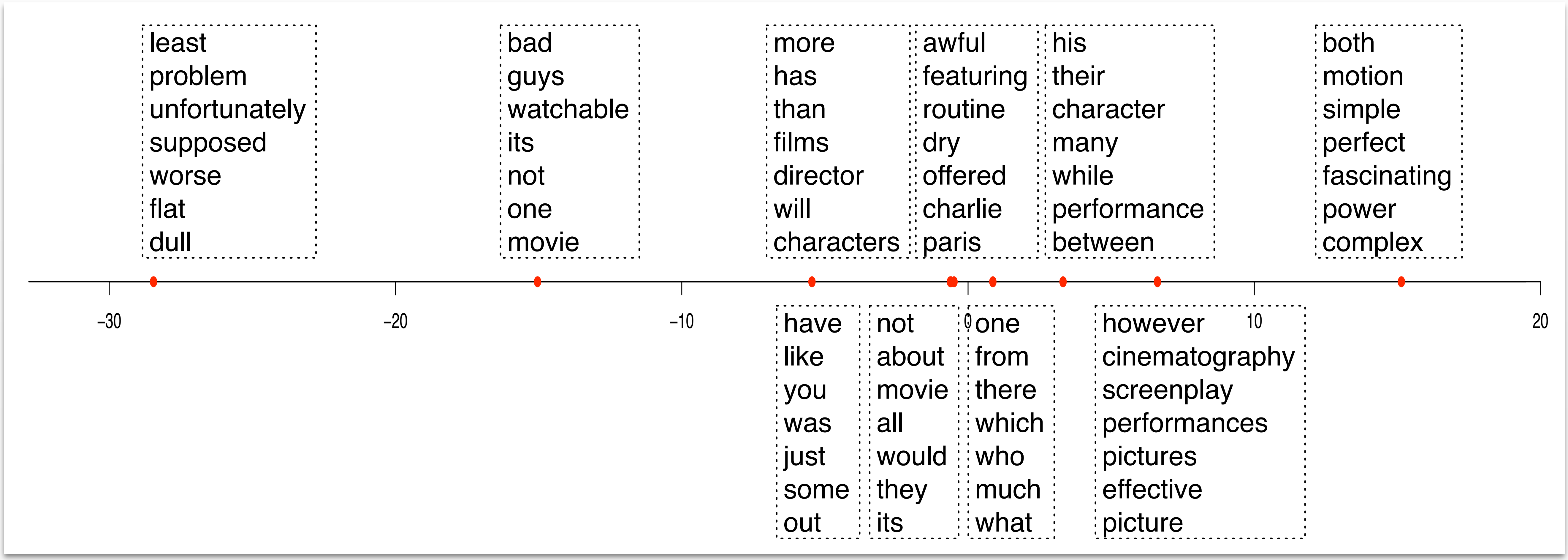
Extensions: Supervised LDA



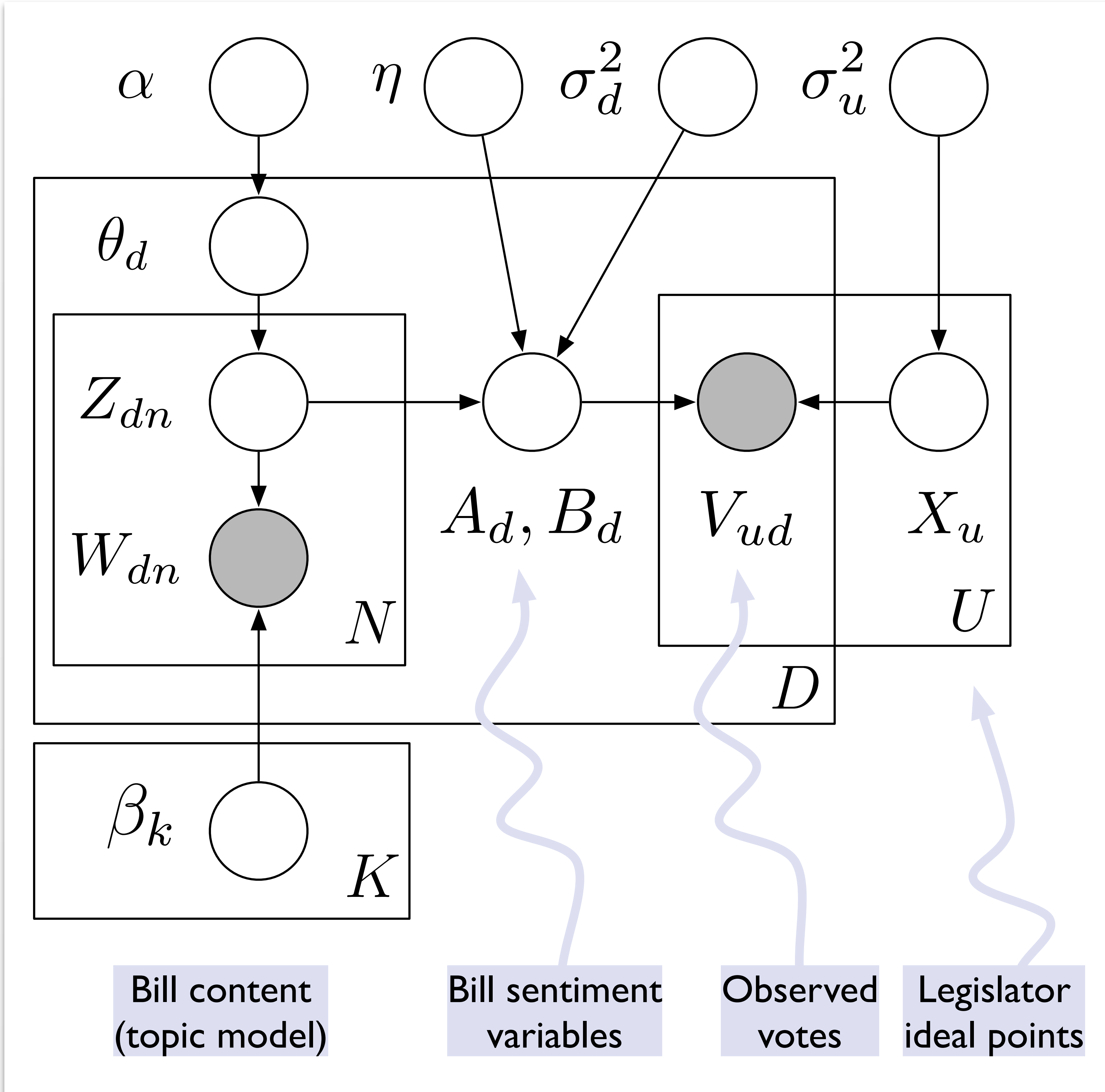
1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim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

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

Extensions: Supervised LDA



Extensions: Ideal Point Topic Models



Extensions: Ideal Point Topic Models