Scalable Inference for Probabilistic Topic Models

David M. Blei

Department of Computer Science
Princeton University

October 19, 2010

Joint work with Sean Gerrish, Matthew Hoffman and Francis Bach

Information overload

www.betaversion.org/~stefano/linotype/news/26/

As more information becomes available, it becomes more difficult to access what we are looking for.

We need new tools to help us organize, search, and understand these vast amounts of information.

Topic modeling

Candida Hofer

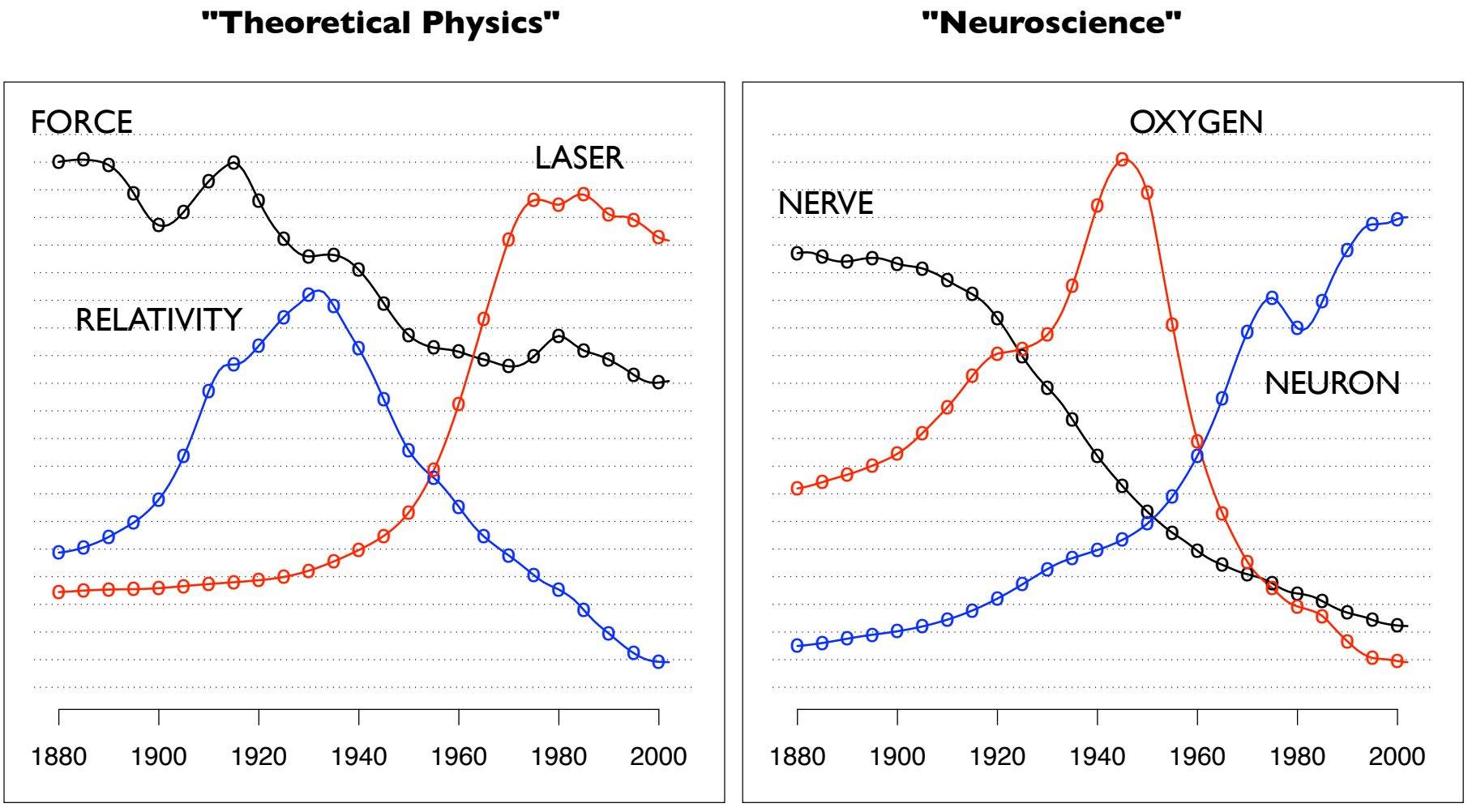
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Uncover the hidden topical patterns that pervade the collection.
2. Annotate the documents according to those topics.
3. Use annotations to organize, summarize, and search the texts.

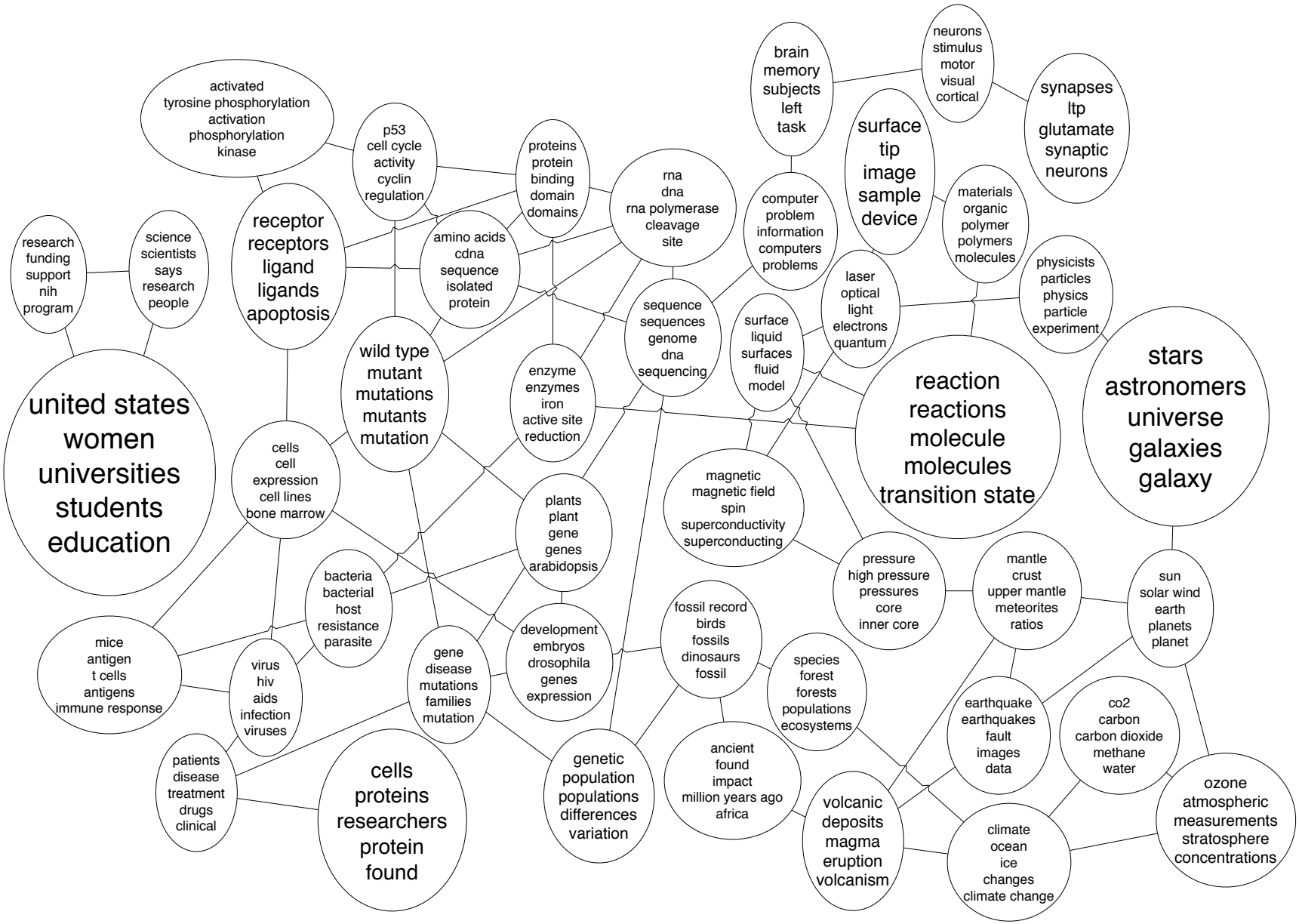
Discover topics from a corpus

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

Model the evolution of topics over time



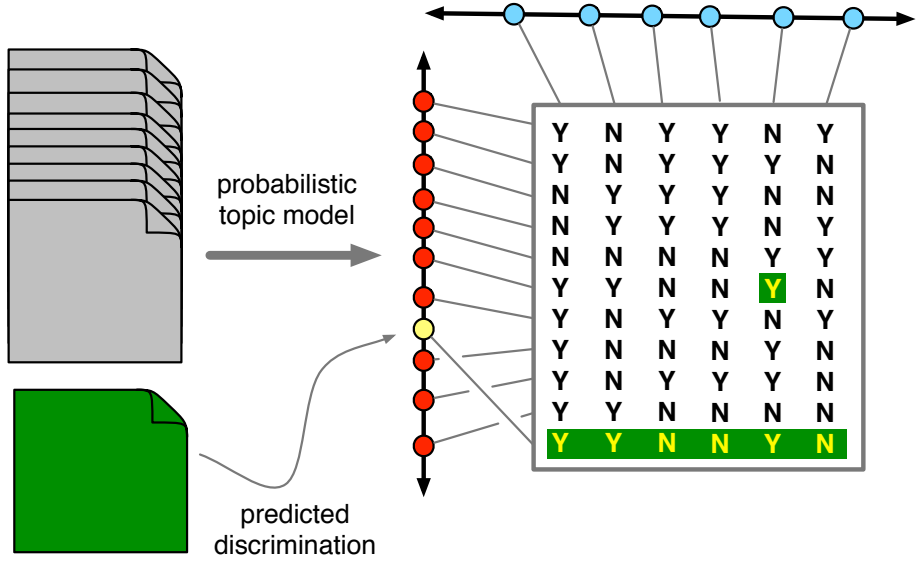
Model connections between topics



This talk

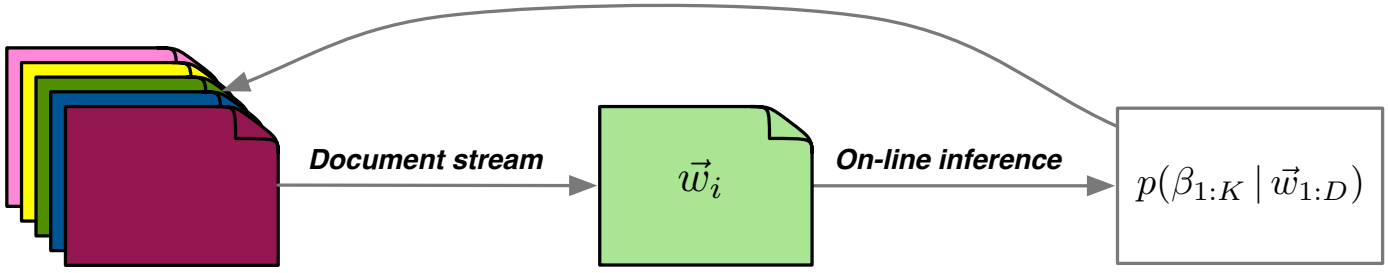
1. Introduction to topic modeling
2. A flurry of topic models
3. Two new ideas—
   - The ideal-point topic model
   - Scalable inference in topic models

Ideal point topic model



- Existing methods for collaborative filtering rely on other users preferences to fill in a new preference.
- Ideal point topic models let us predict user preferences from entirely new items.
- Illustrates how to build new topic models and use them to solve real-world textual problems

Scalable inference for probabilistic topic models



- Existing topic modeling algorithms process document collections in **batch**—iteratively examining each document
- Many applications of topic modeling could benefit from processing documents in a **stream**
  - Linking topic models to web APIs and databases
  - Handling millions and billions of documents
  - Refining topic models on the fly, e.g., for user interfaces

Latent Dirichlet allocation

Probabilistic modeling

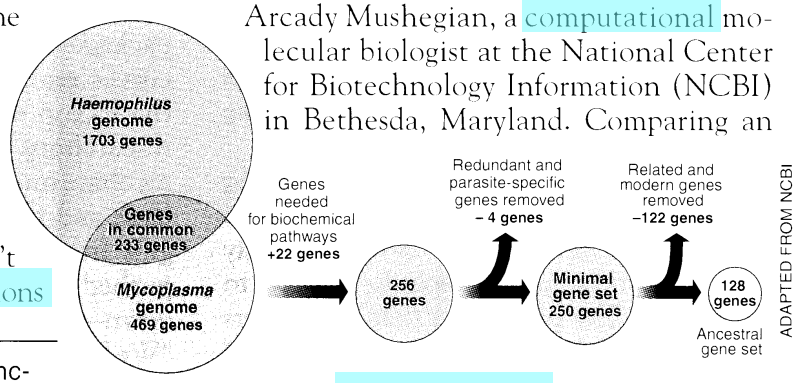
1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using *posterior inference*
   - What are the topics that describe this collection?
3. Situate new data into the estimated model.
   - How does this query or new document fit into the estimated topic structure?

Intuition behind LDA

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



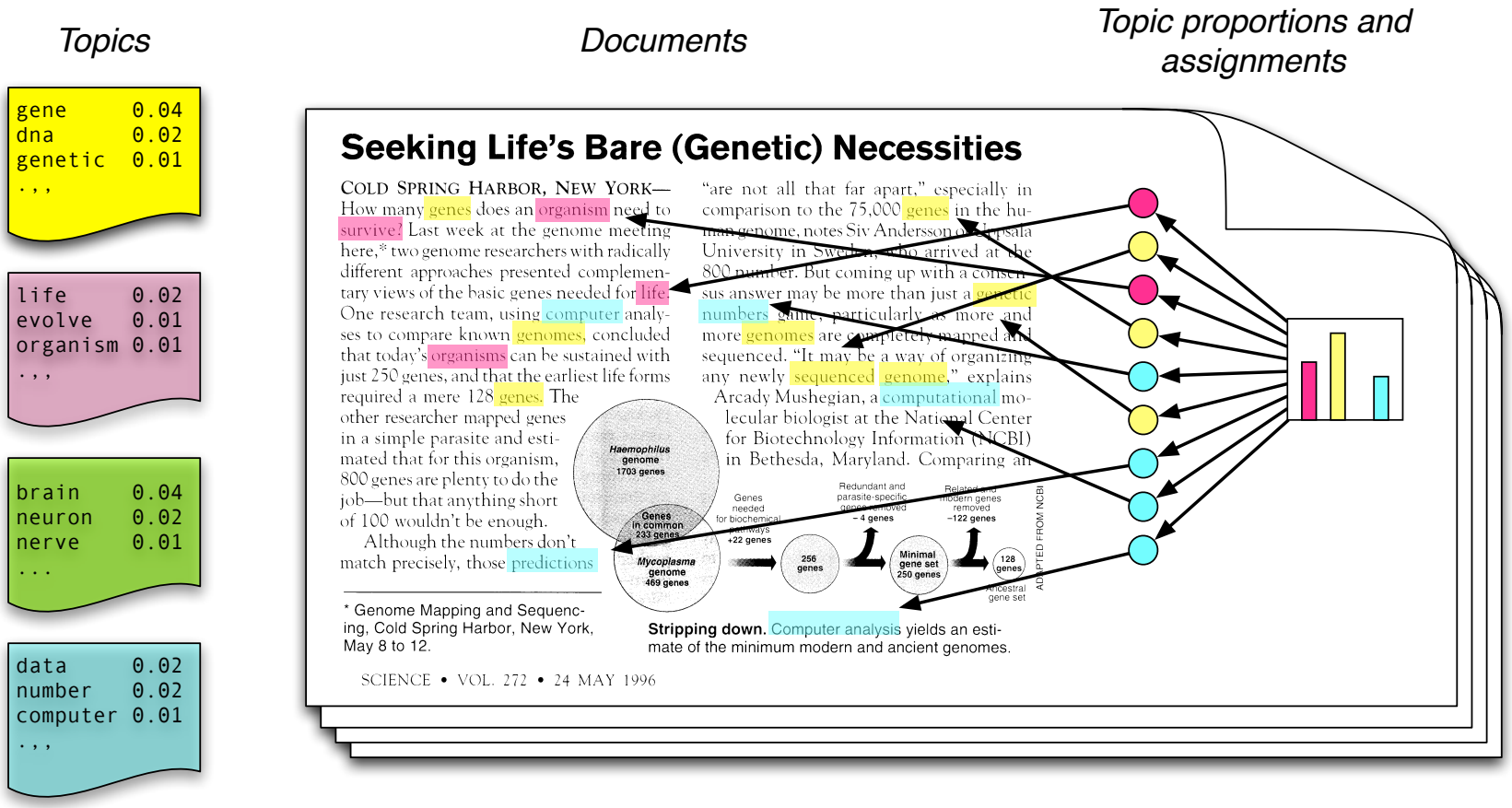
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

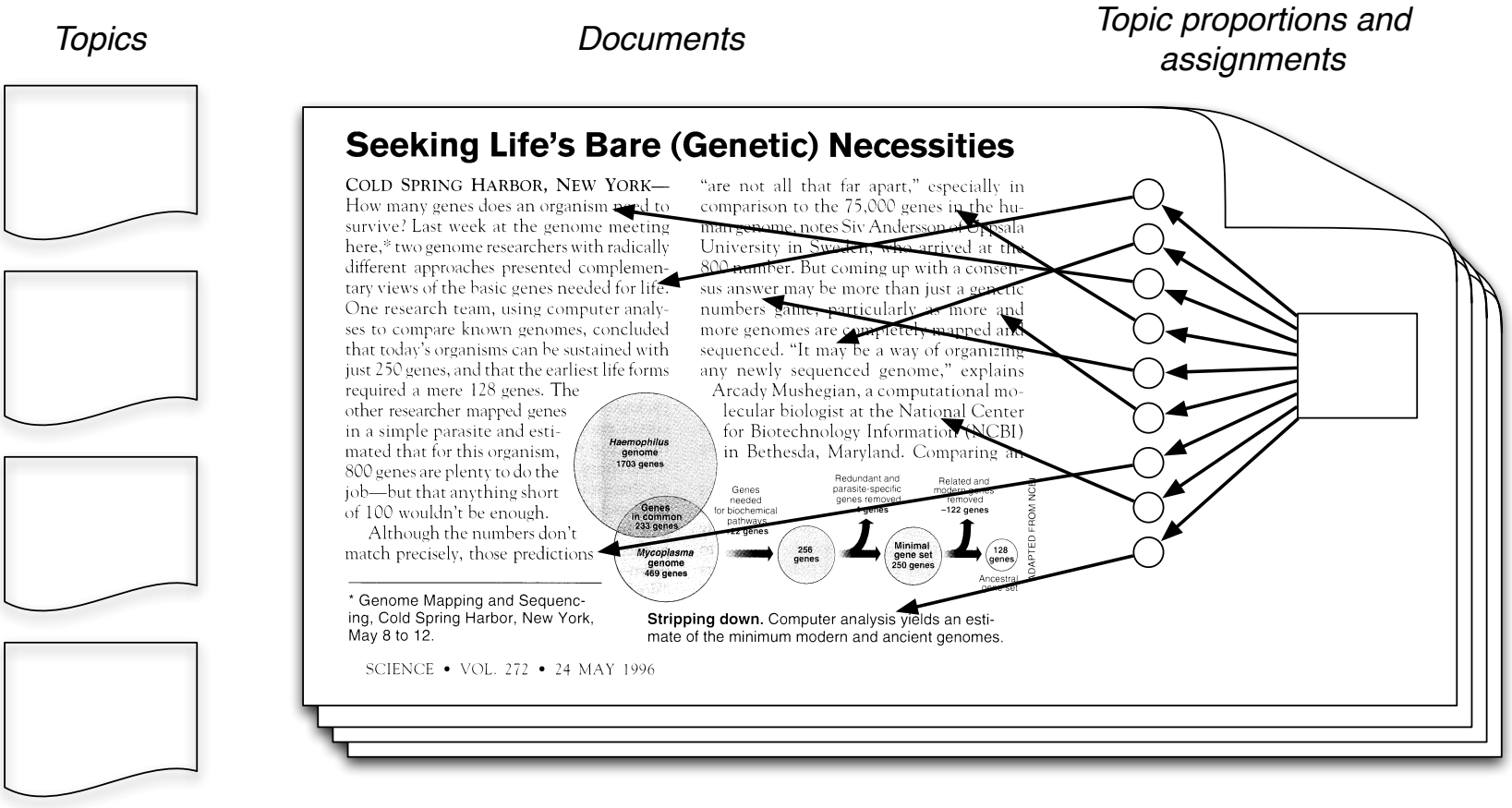
Simple intuition: Documents exhibit multiple topics.

Generative model



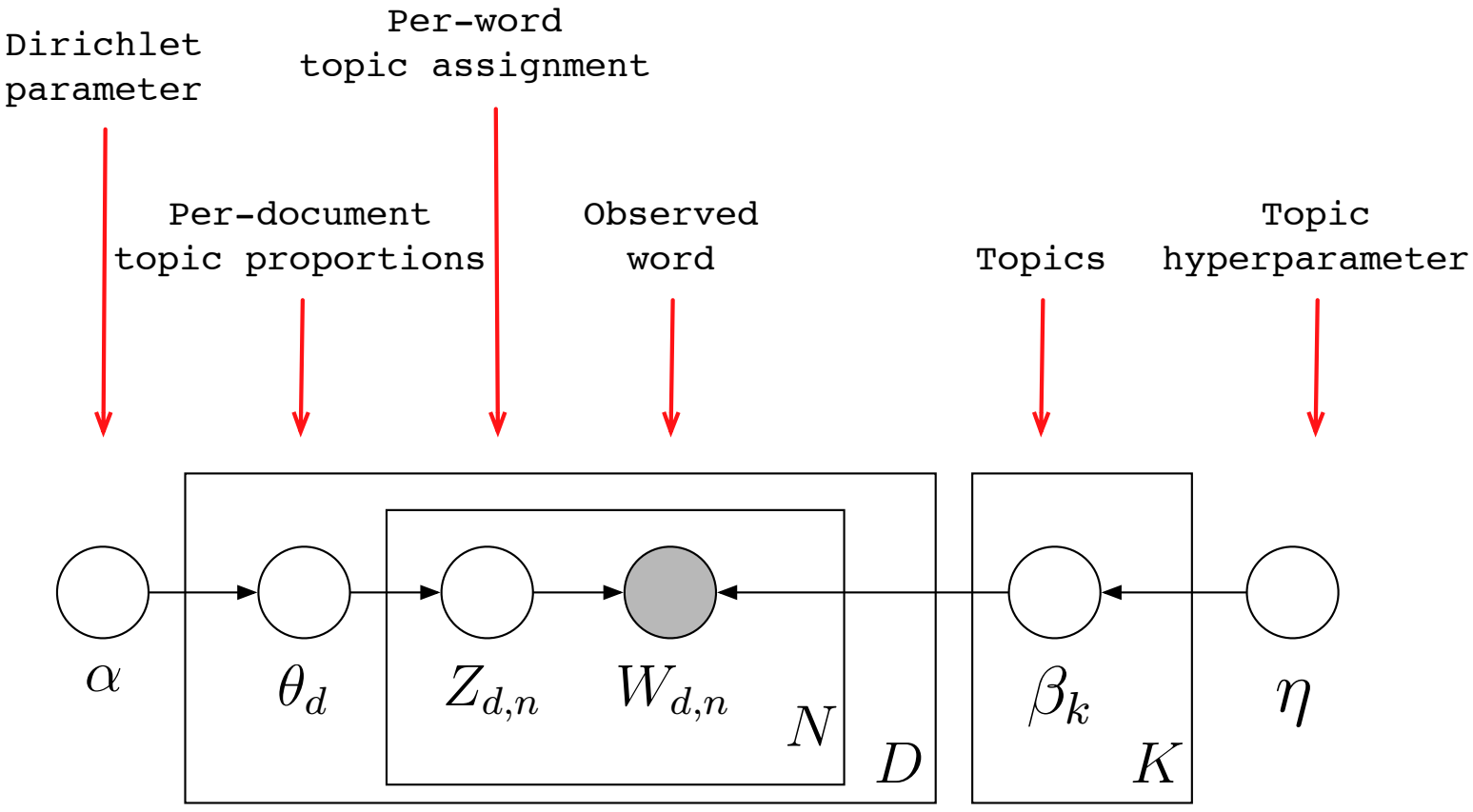
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

The posterior distribution



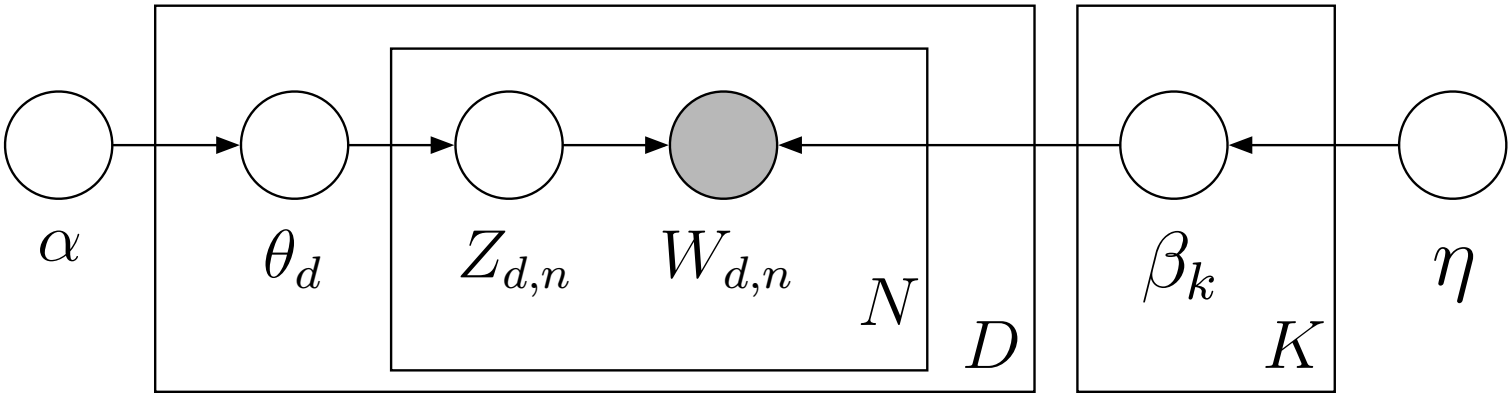
- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

Latent Dirichlet allocation



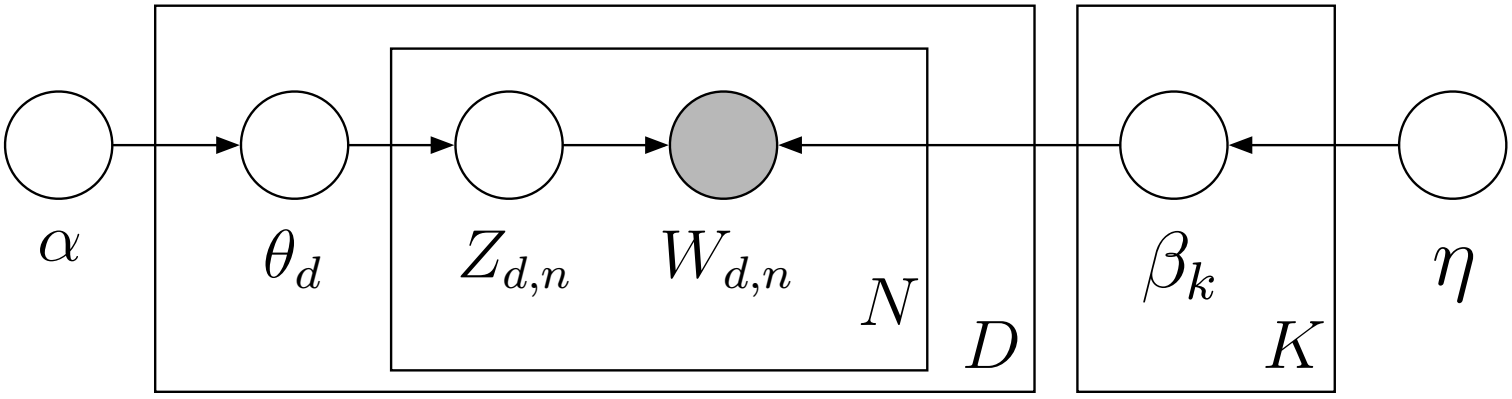
$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Latent Dirichlet allocation



- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.

Latent Dirichlet allocation

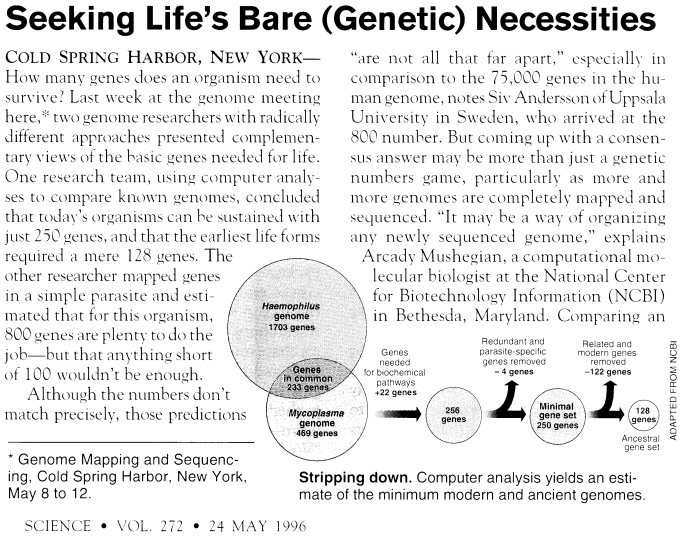


- Computing the posterior is intractable:

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

- Stay tuned for the second half of this talk...

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

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

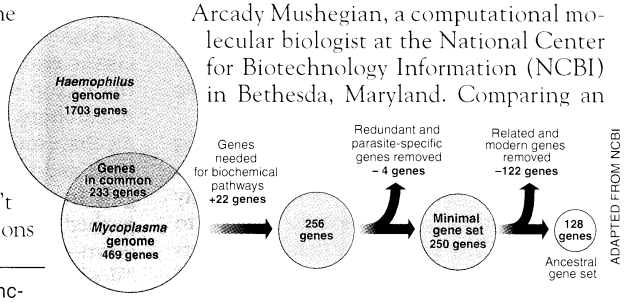
- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

Example inference

Seeking Life's Bare (Genetic) Necessities

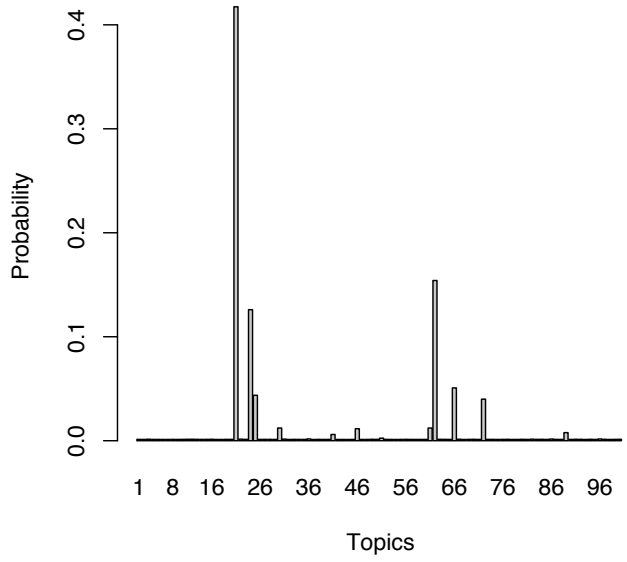
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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



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

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



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 (II)

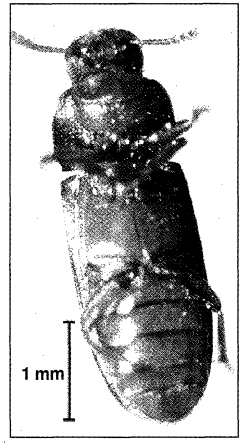
Chaotic Beetles

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).

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

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 bufeted 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*).



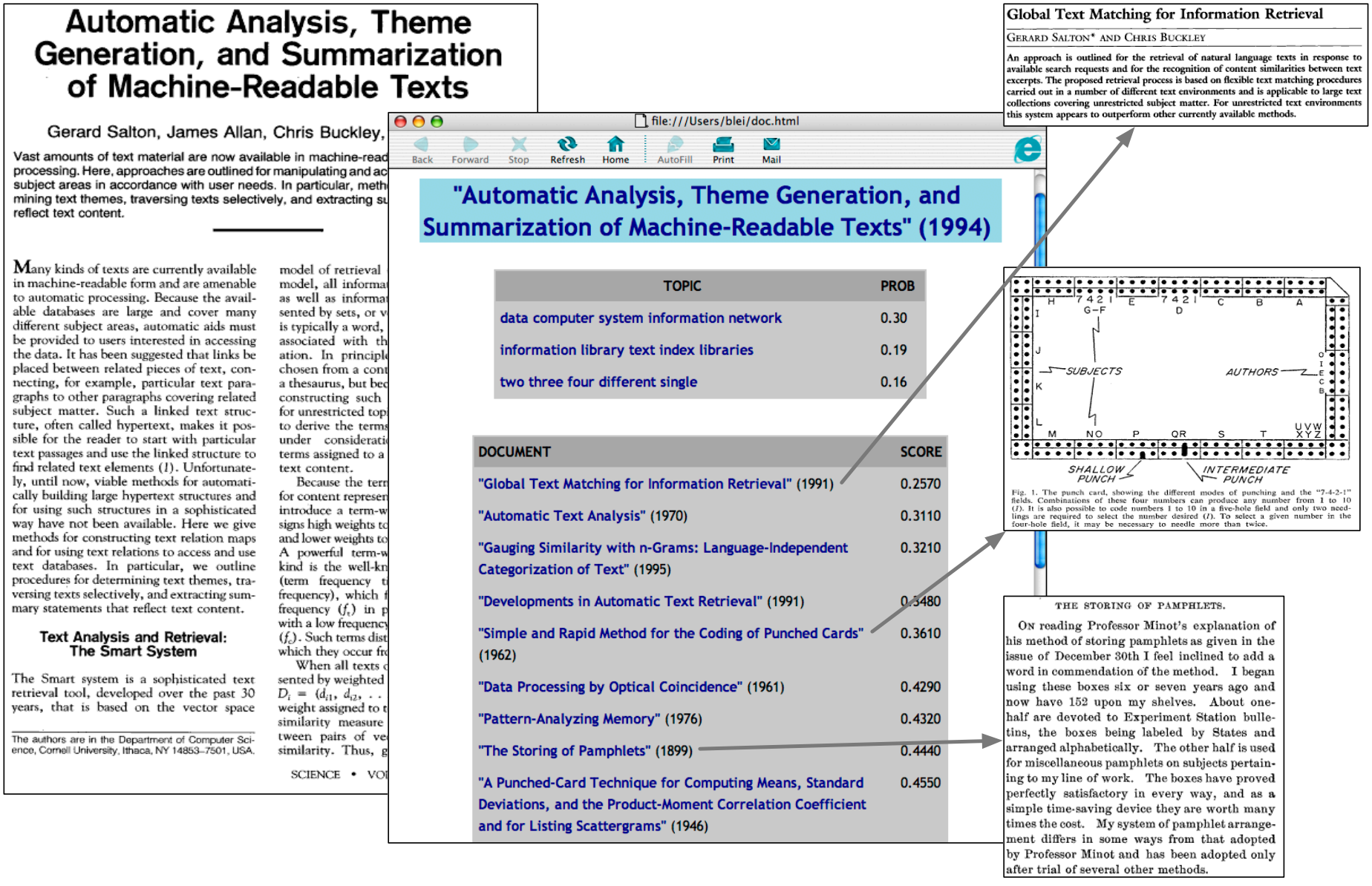
Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

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-

Example inference (II)

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

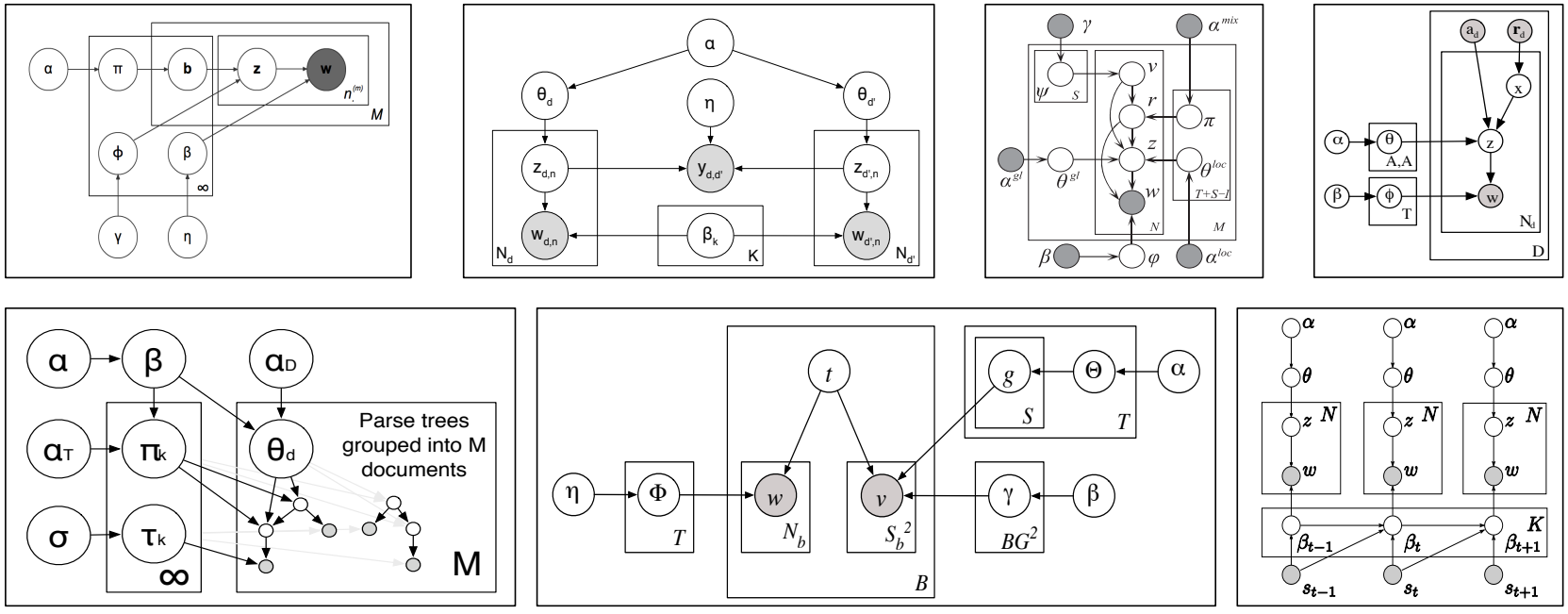
Used in exploratory tools of document collections



LDA summary

- LDA is a powerful model for
  - Visualizing the hidden thematic structure in large corpora
  - Generalizing new data to fit into that structure
- LDA is a mixed membership model (Erosheva, 2004) that builds on the work of Deerwester et al. (1990) and Hofmann (1999).
- See Blei et al. (2003) for details and a quantitative comparison. See my web-site for code and other papers.
- The same model was independently invented for population genetics analysis (Pritchard et al., 2000).

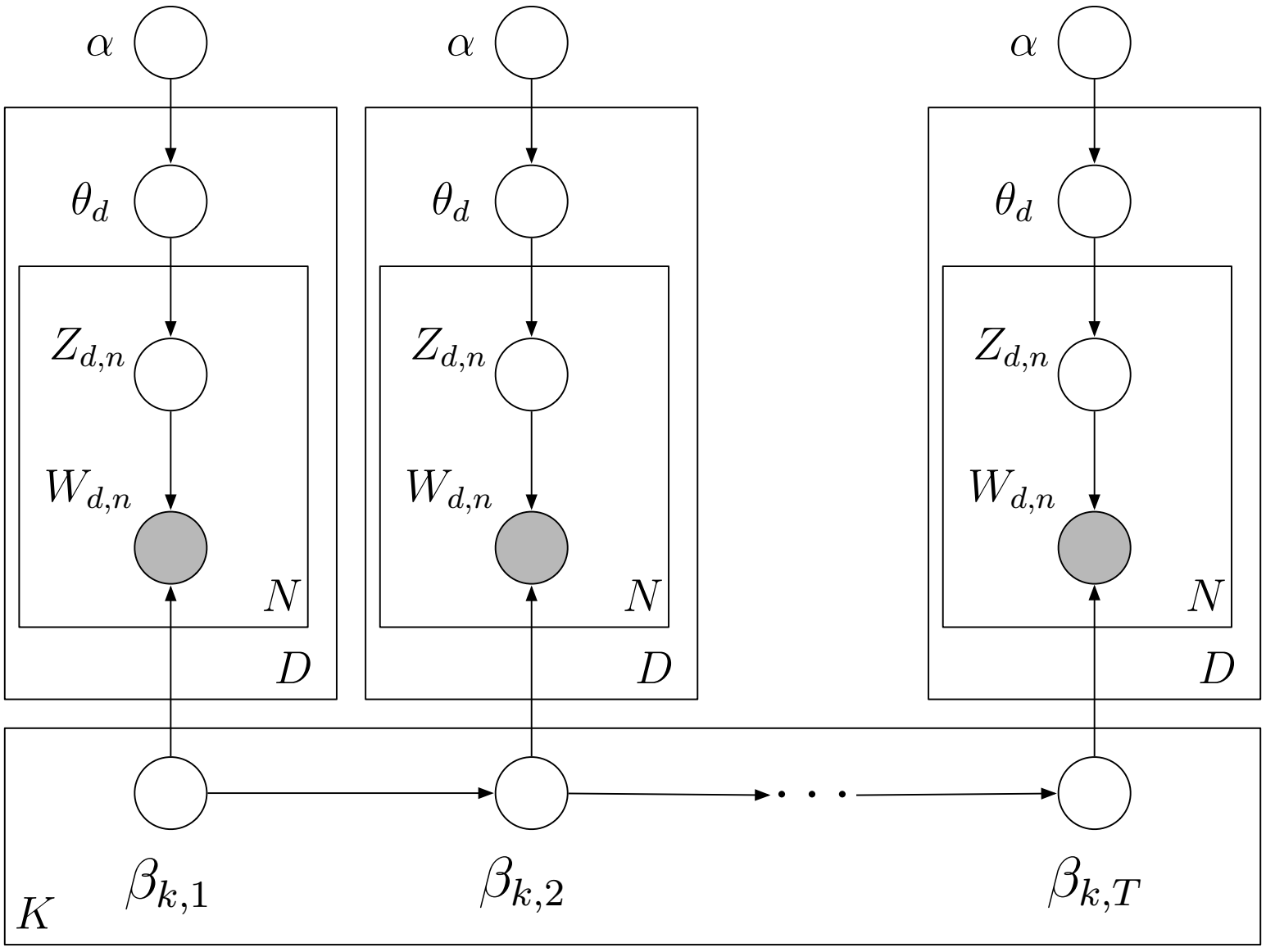
LDA summary



- There is a growing research literature on topic models using LDA
- This has been a success story for—
  - Describing assumptions with hierarchical Bayesian models
  - Computation on large data with approximate inference
  - Exploratory analysis with unsupervised learning

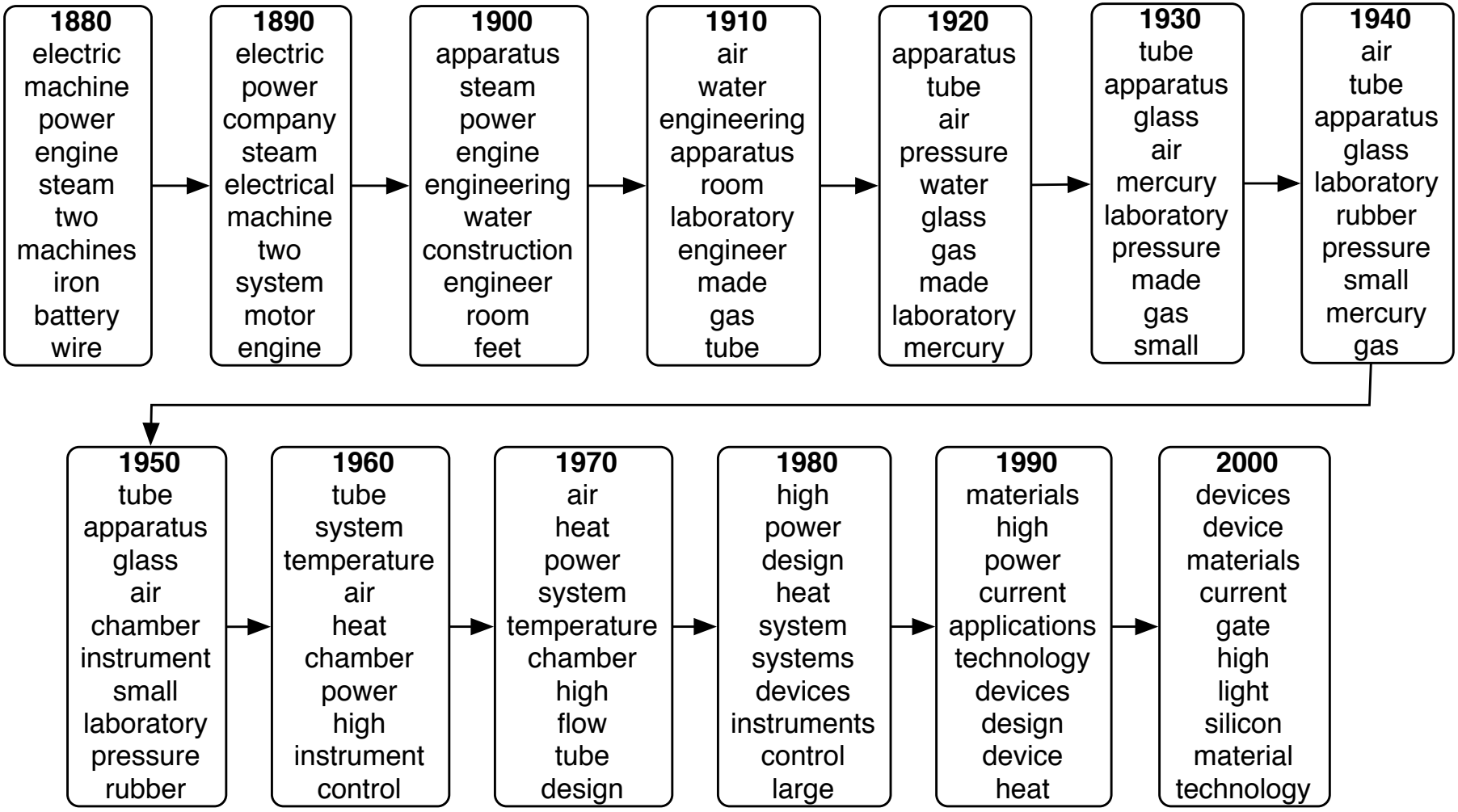
A flurry of models that extend LDA

Dynamic topic models



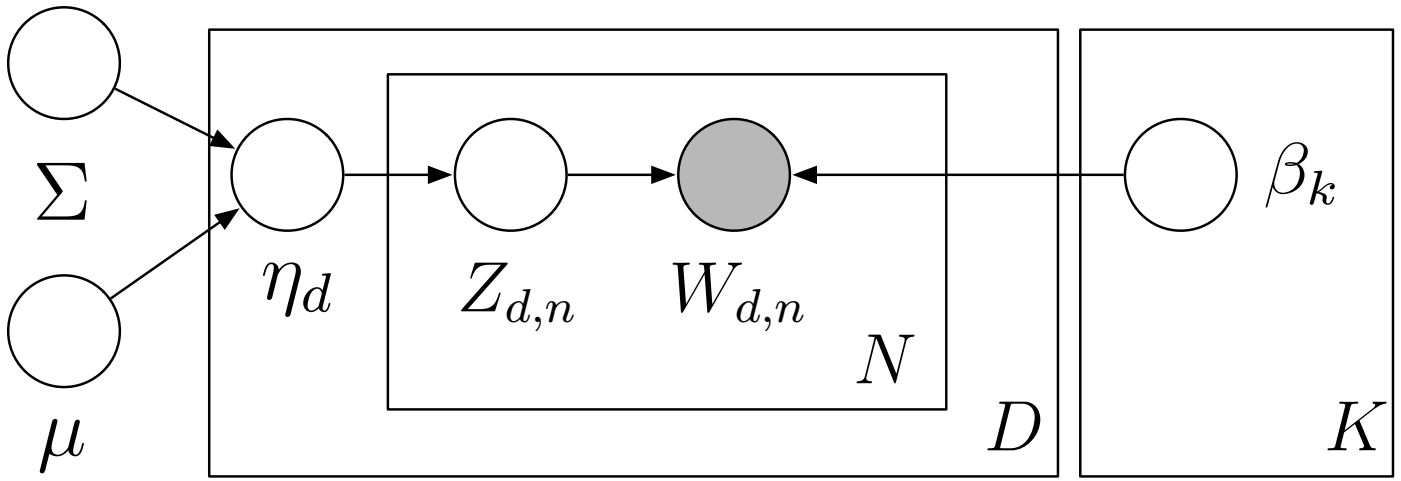
Topics drifting in time

Dynamic topic models



(See Blei and Lafferty, 2007)

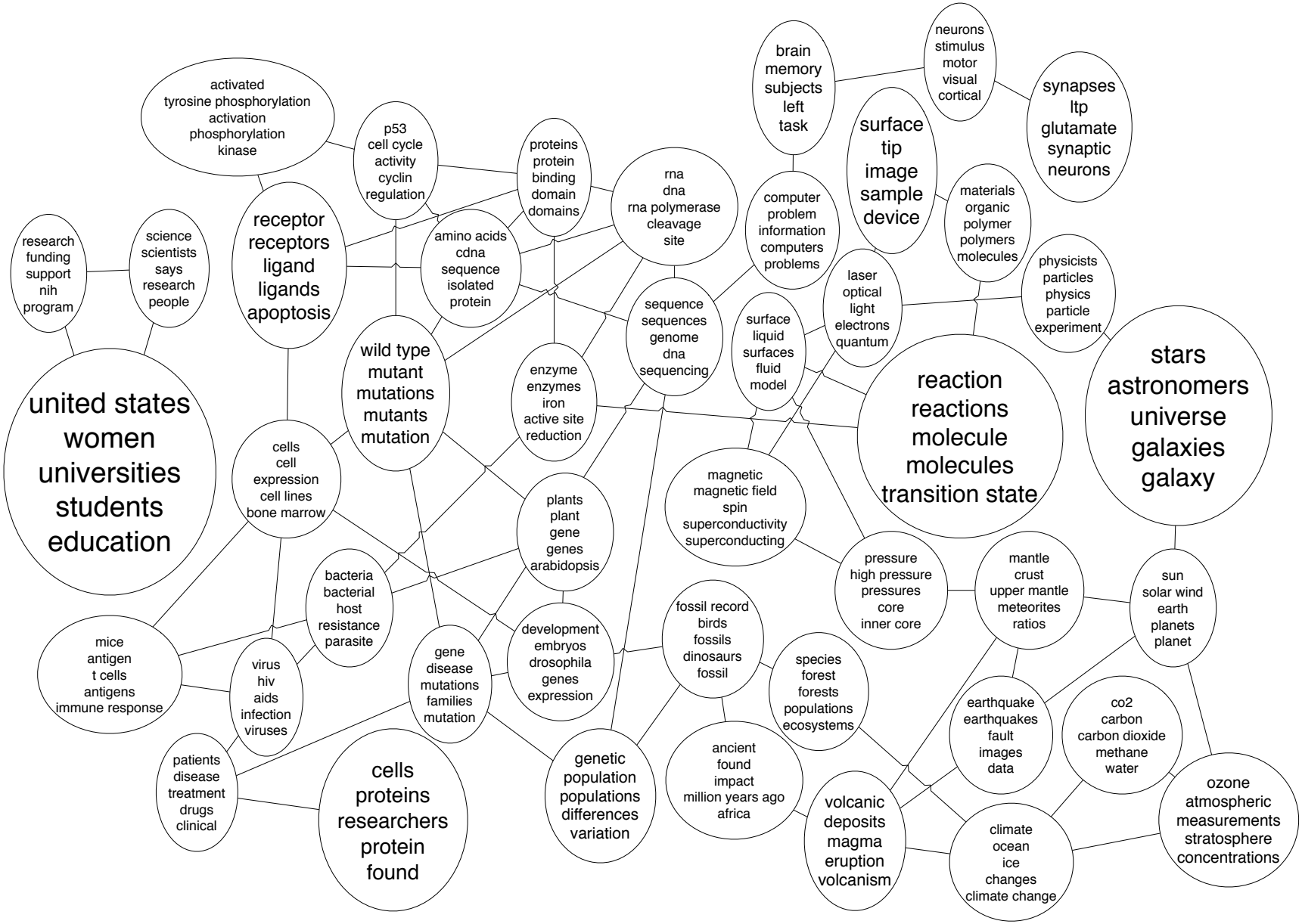
Correlated topic models



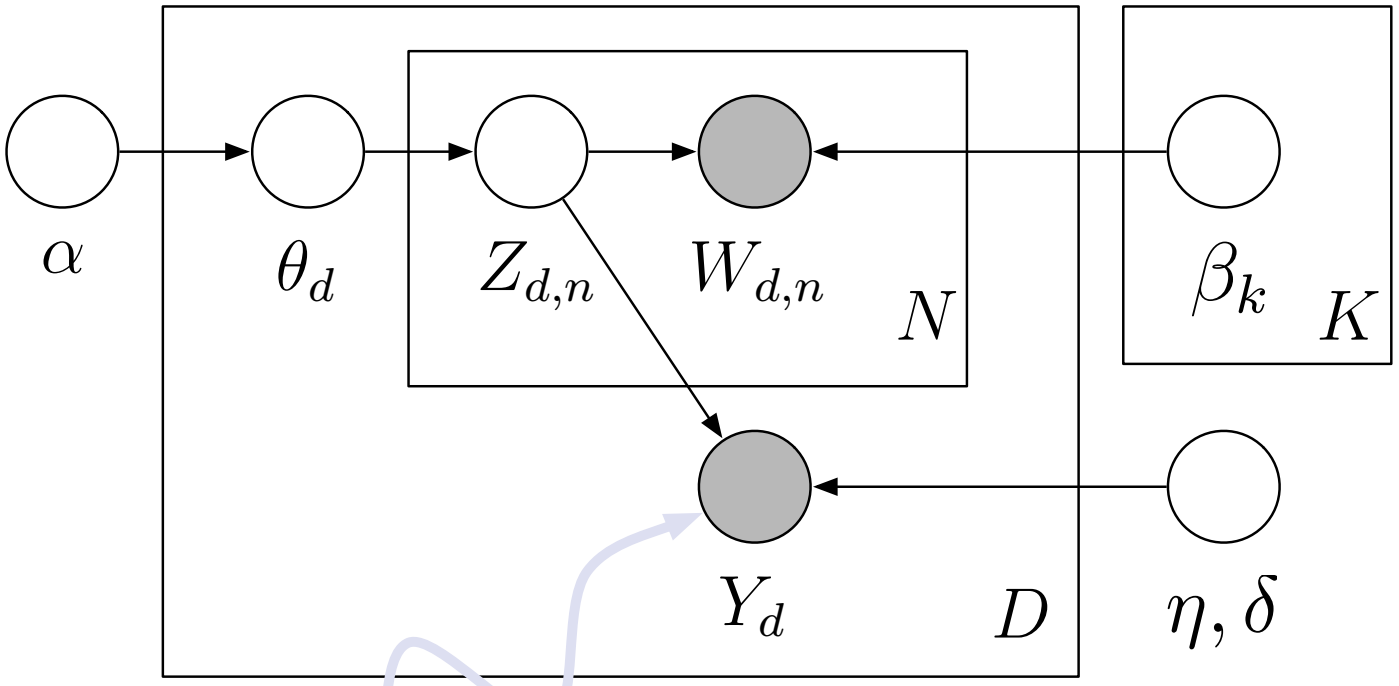
Noconjugate prior
on topic proportions

(See Blei and Lafferty, 2007b)

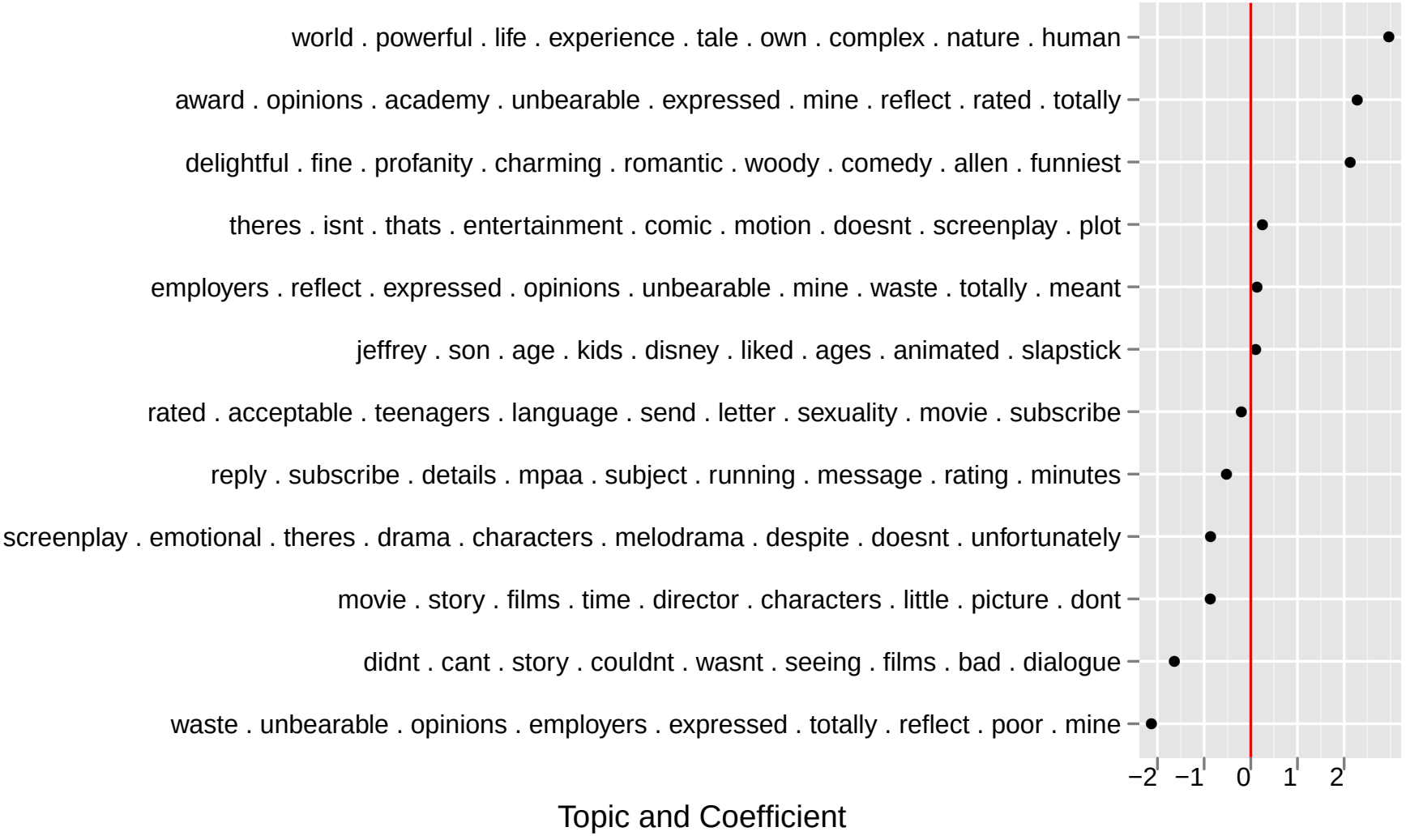
Correlated topic models



Supervised topic models

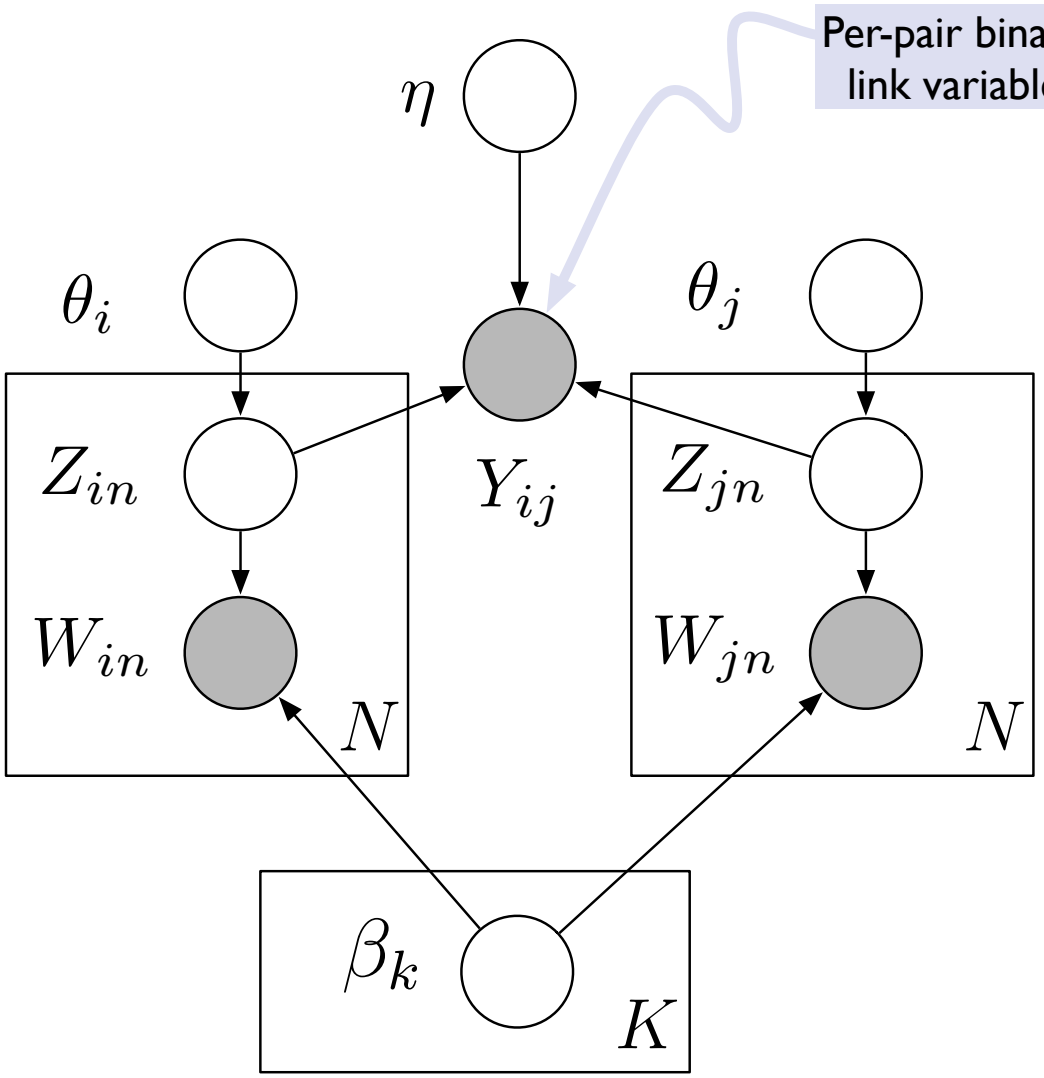


Supervised topic models



(See Blei and McAuliffe, 2007)

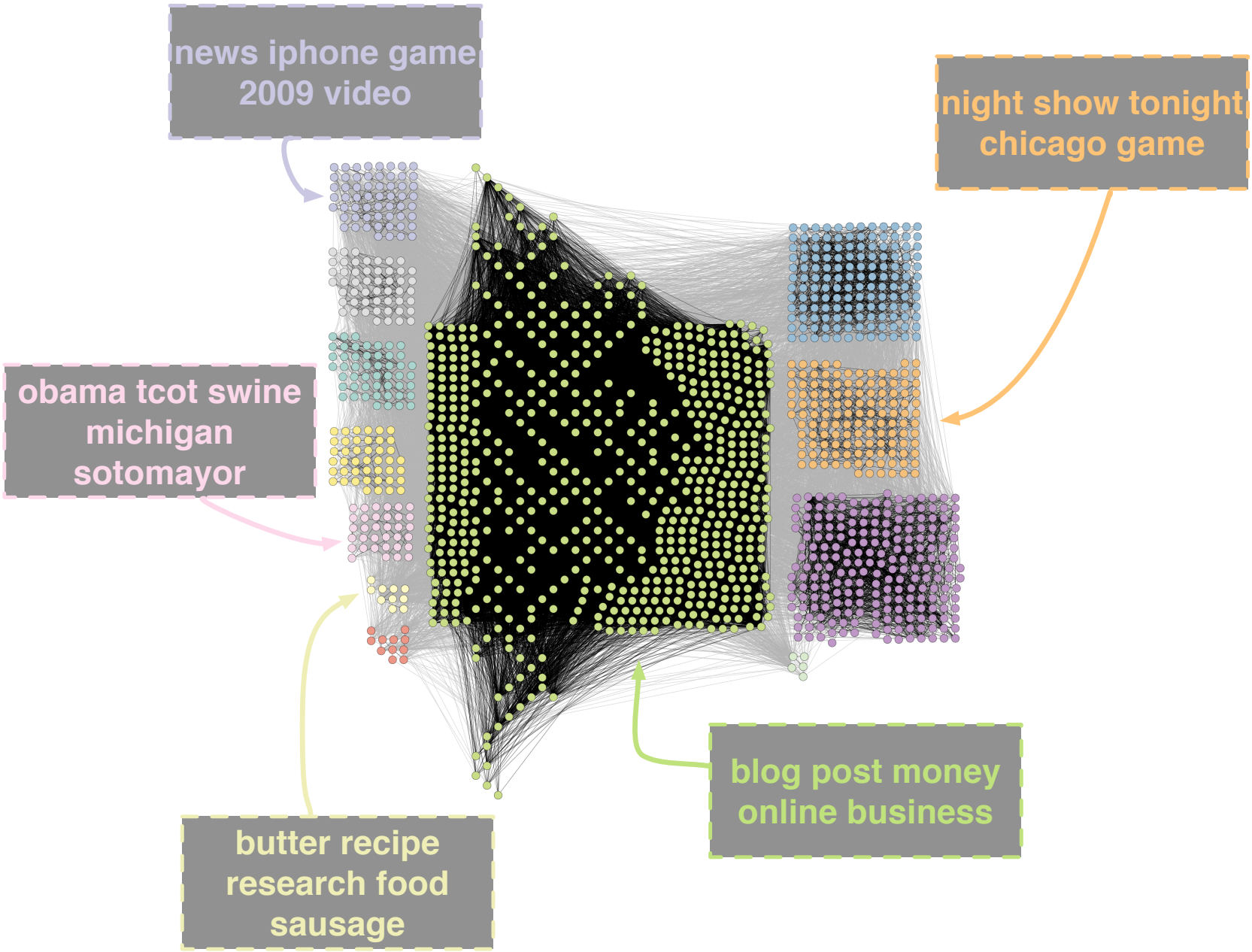
Relational topic models



This structure is repeated for all pairs of documents

(See Chang and Blei, 2010)

Relational topic models

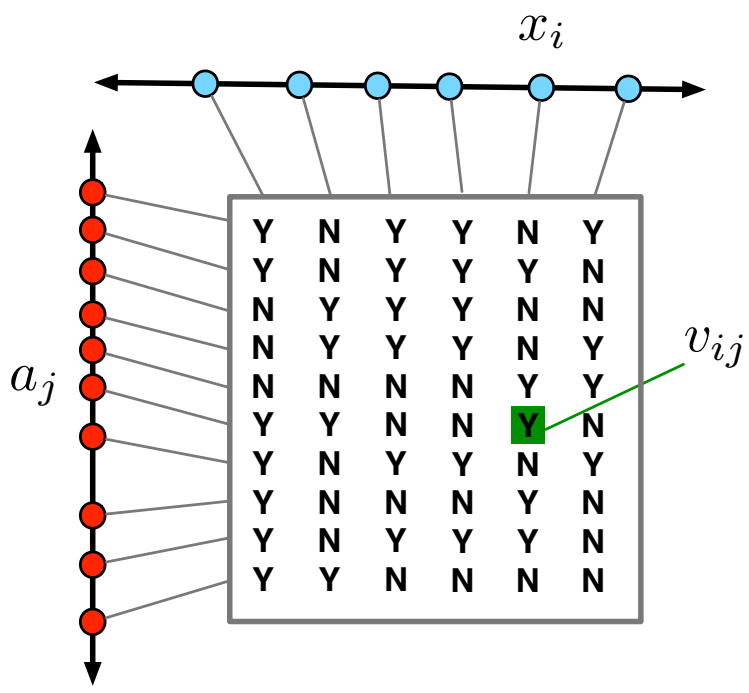


Other extensions

- Bayesian nonparametric topic models (Teh et al., 2006)
- Syntactic topic models (Boyd-Graber and Blei, 2009)
- Topic models on images (Fei-fei and Perona, 2005 and others)
- Topic models on social network data (Airoldi et al., 2008)
- Topic models on music data (Hoffman et al., 2008)
- Modeling scientific impact (Gerrish and Blei, 2010)

Ideal point topic models

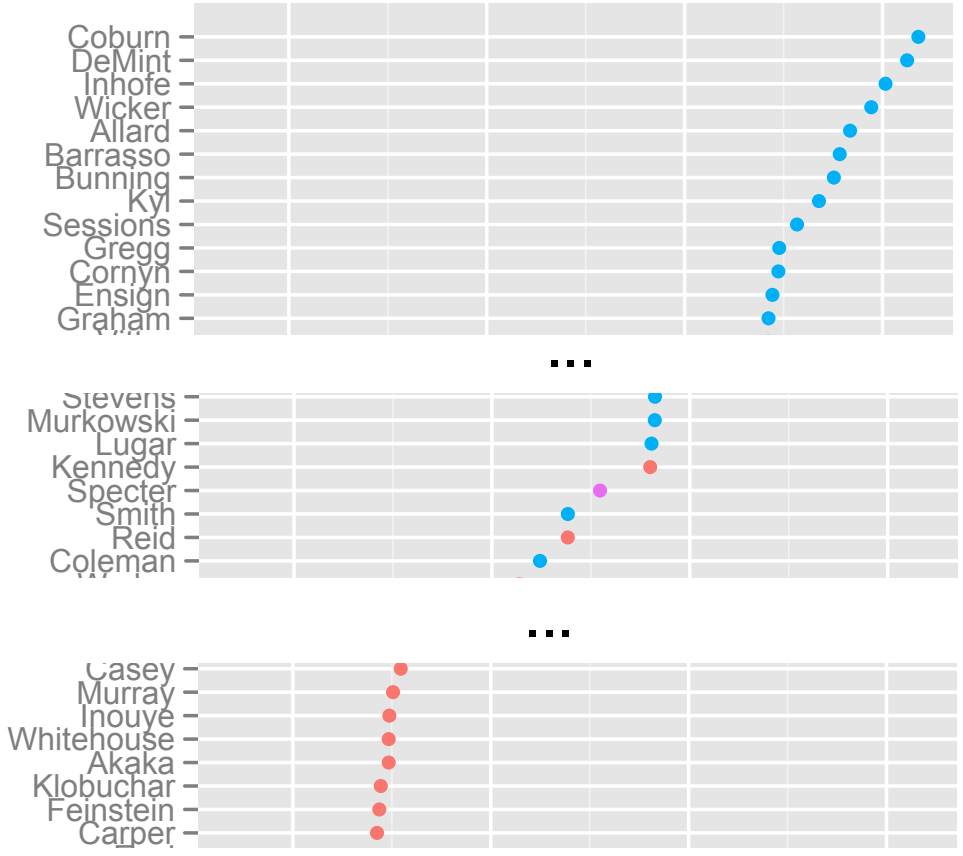
The ideal point model



$$p(v_{ij}) = f(d(x_i, a_j))$$

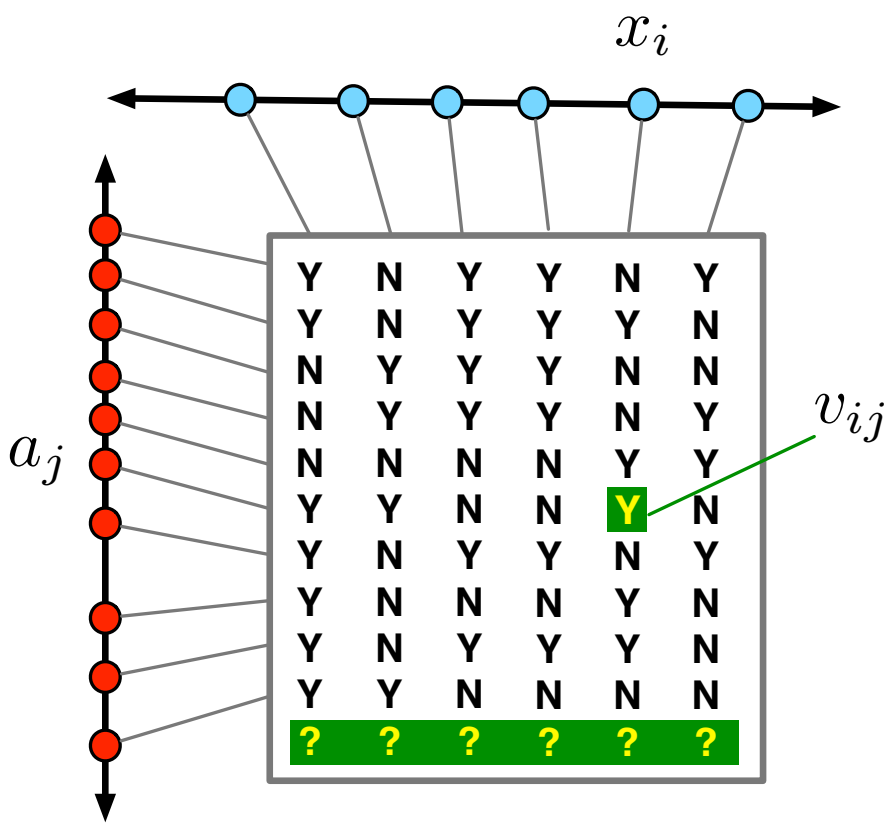
- A model devised to uncover voting patterns (Clinton et al., 2004).
- We observe roll call data v_{ij}.
- Bills attached to discrimination parameters a_j.
  Senators attached to ideal points x_i.

The ideal point model



- Posterior inference reveals the political spectrum of senators
- Widely used in quantitative political science.

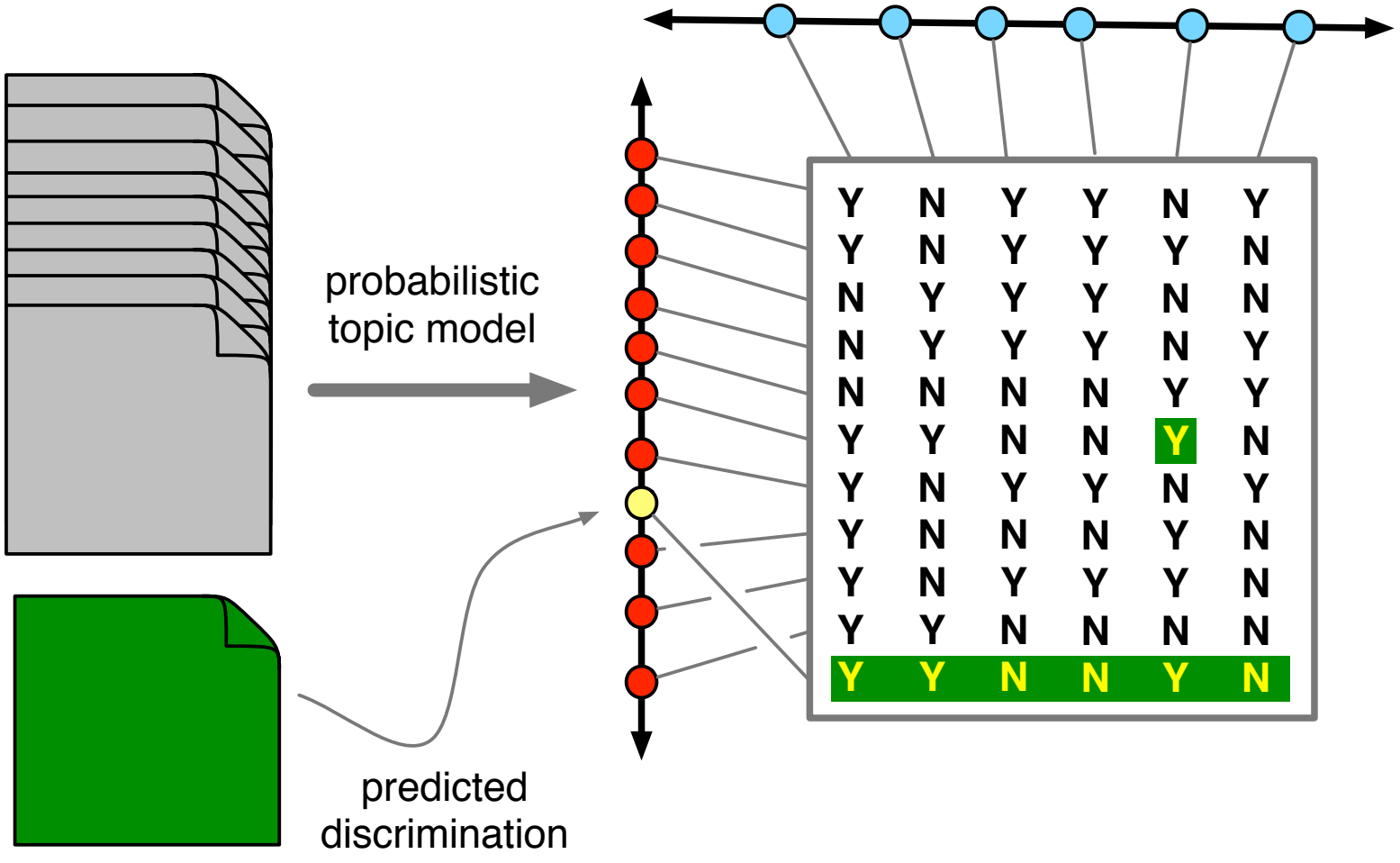
The ideal point model is limited for prediction



$$p(v_{ij}) = f(d(x_i, a_j))$$

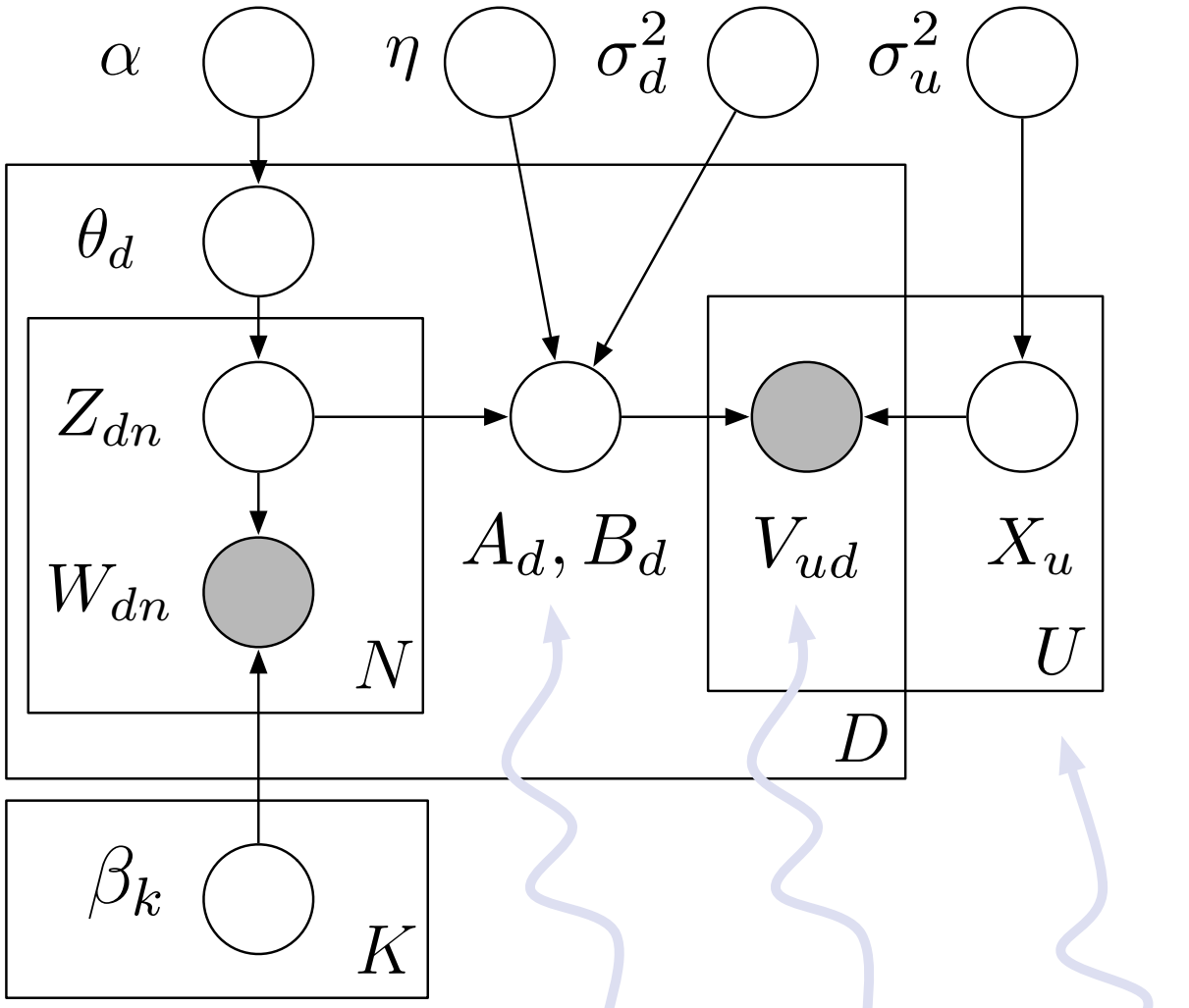
- We can predict a missing vote.
- But we cannot predict all the missing votes from a bill.
- Cf. the limitations of collaborative filtering

Ideal point topic models

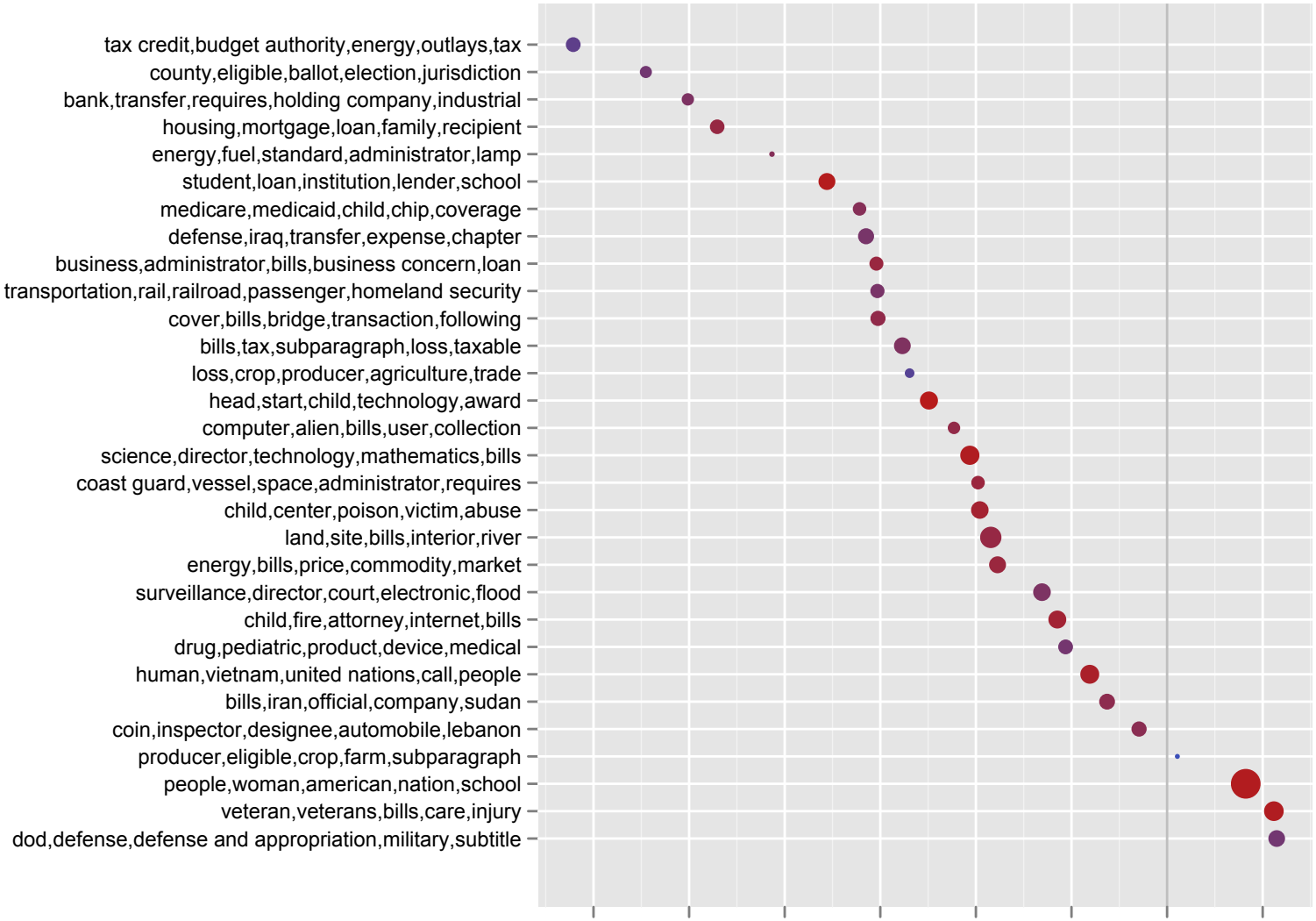


Use supervised topic modeling assumptions as a predictive mechanism from bill texts to bill discrimination.

Ideal point topic models

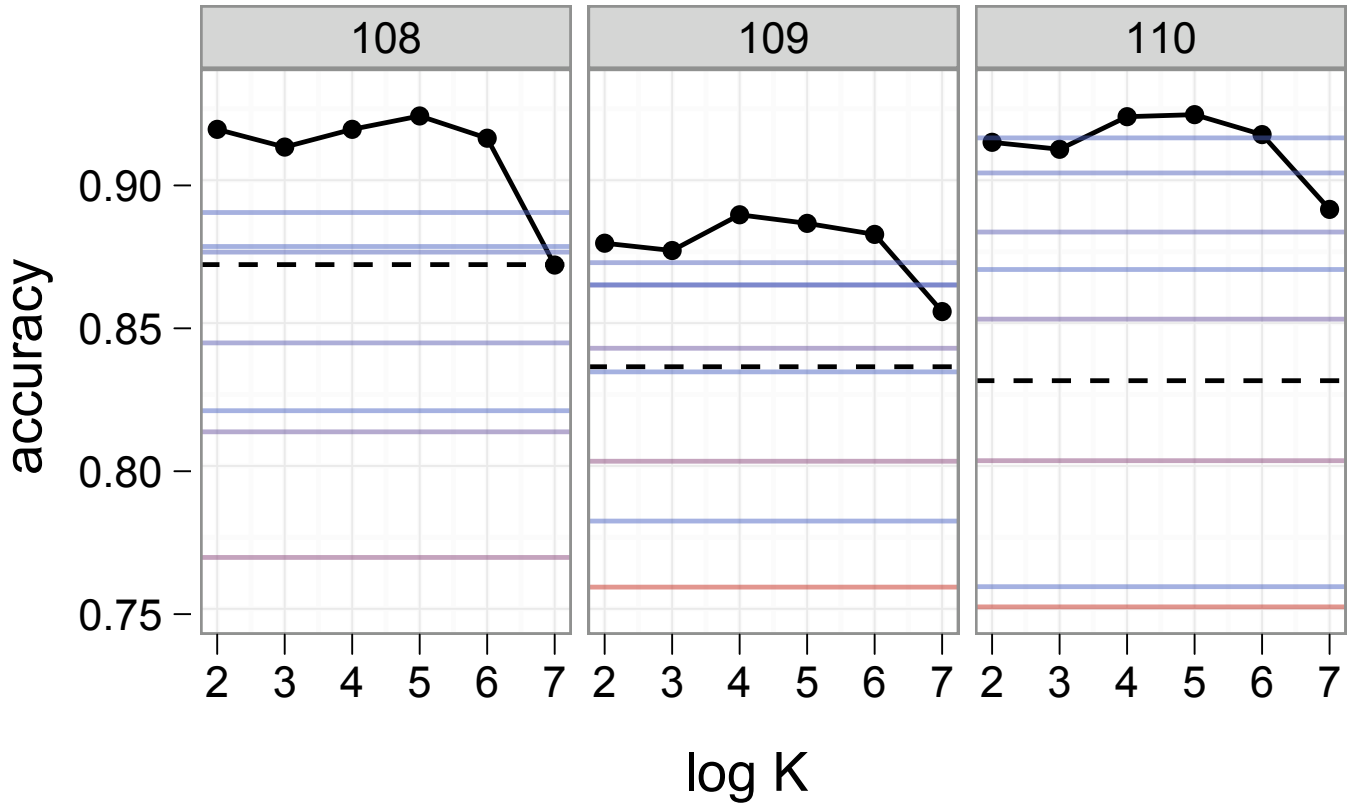


Ideal point topics



In addition to senators and bills, IPTM places **topics** on the spectrum.

Prediction on completely held-out votes

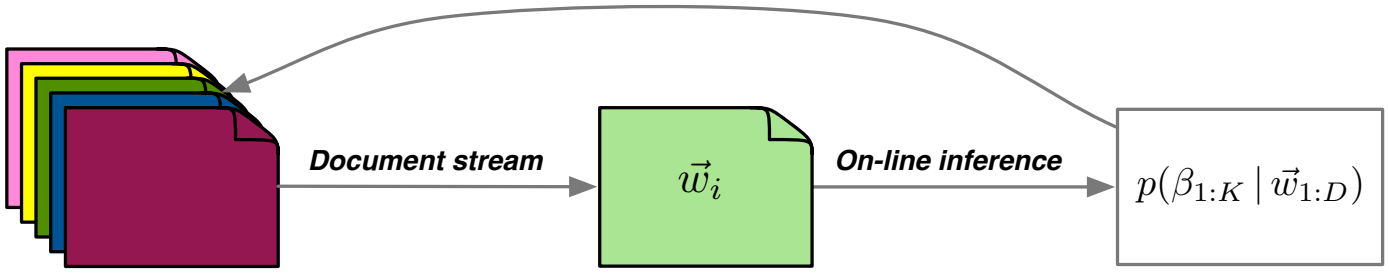


Ideal point topic models

- Ideal point topic model illustrates
  - Topic modeling embedded in a complex model
  - Topic modeling used to solve a real-world problem with text
  - Variational methods allow us to analyze larger data sets
- More generally, consider
  - Senators are *users*.
  - Bills are *items*.
- Existing collaborative filtering is akin to classical ideal point.
- Our model lets us predict preferences on *completely new items*.

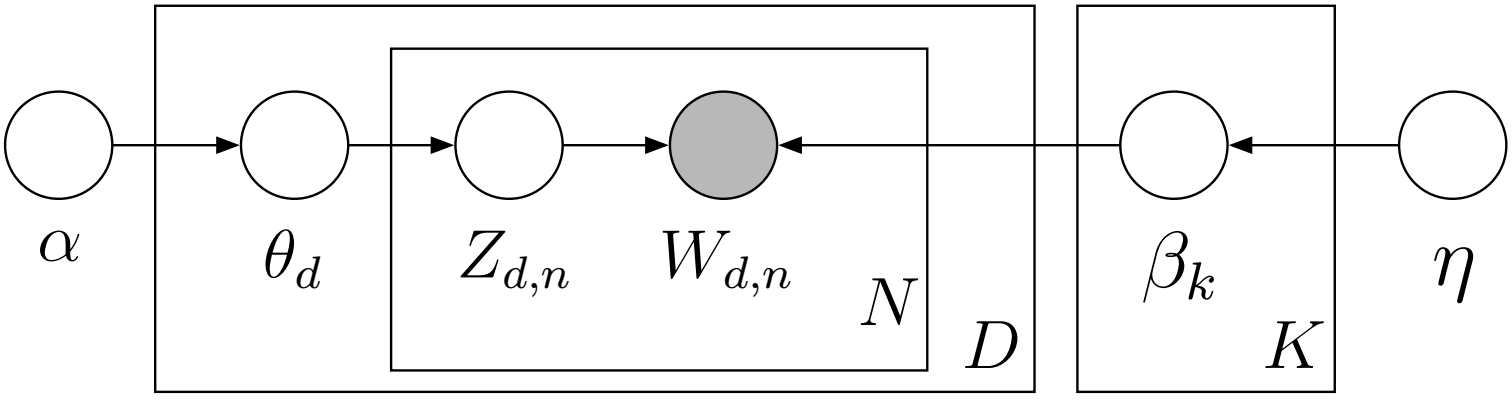
On-line posterior inference

The need for streaming inference



- These topic models work in the same way—
  - Posit a generative model
  - Cast the task at hand as a posterior computation
  - Approximate the posterior
- To approximate the posterior, existing topic modeling algorithms process document collections in **batch**.
- Many applications of topic modeling could benefit from processing documents in a **stream**.

Return to LDA

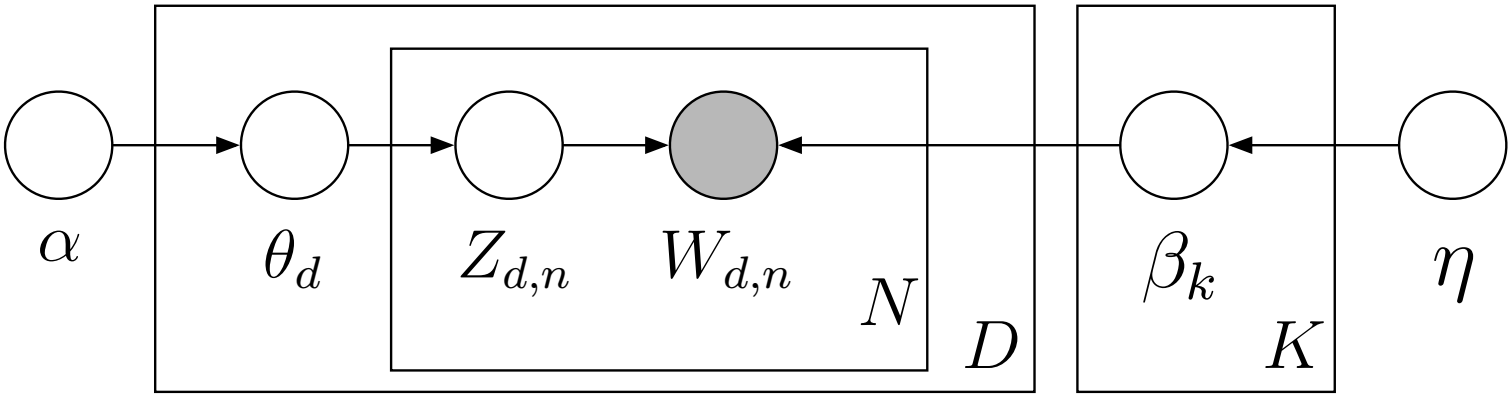


- Computing the posterior is intractable:

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

- Several approximation techniques have been developed.

Return to LDA



- We focus on variational methods
- Alternative to MCMC; replace sampling with optimization
- Often faster than MCMC
  (Blei and Jordan 2005, Braun and McAuliffe 2010)
- Provides the right ingredients for a streaming inference algorithm

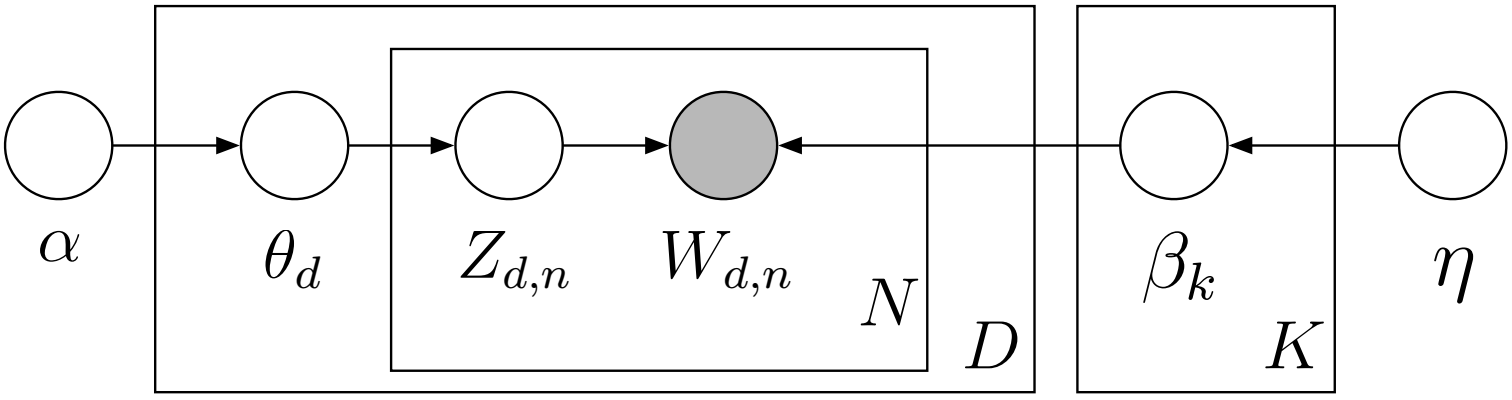
Variational inference

- Introduce a distribution over the latent variables $q(\theta, z)$, parameterized by *variational parameters*
- Use Jensen's inequality to bound the log probability of a document $w = w_{1:N}$.

$$\log p(w) \geq \mathrm{E}_q[\log p(\theta, Z, w)] - \mathrm{E}_q[\log q(\theta, Z)]$$

- We optimize the variational parameters to tighten this bound.
- This is the same as finding the member of the family q that is closest in KL divergence to $p(\theta, z \mid w)$.

Variational Inference for LDA

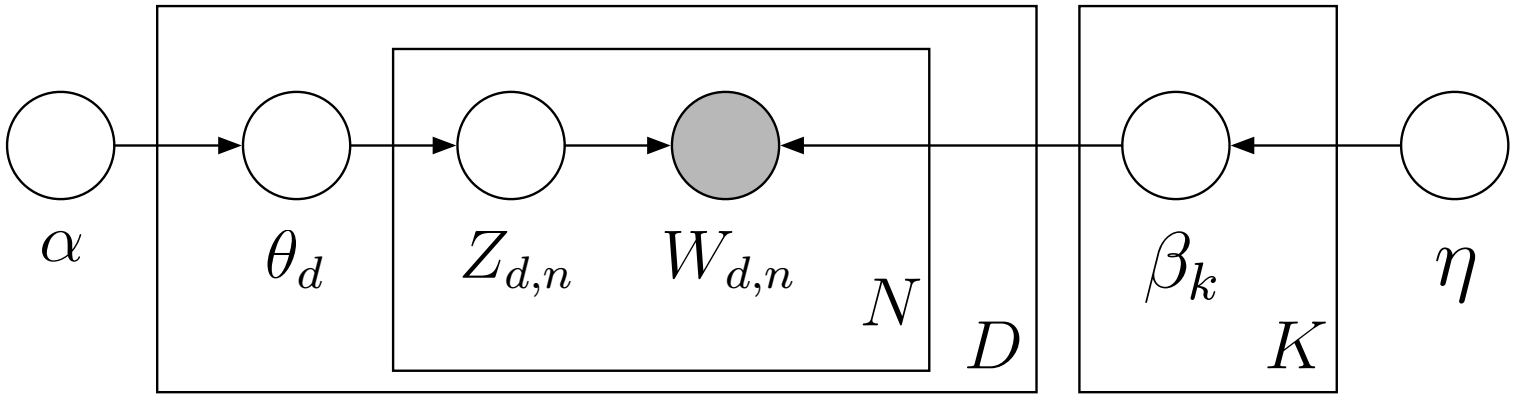


- The mean field variational distribution is

$$q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

- In the posterior, the latent variables are **not** independent.

Variational Inference for LDA



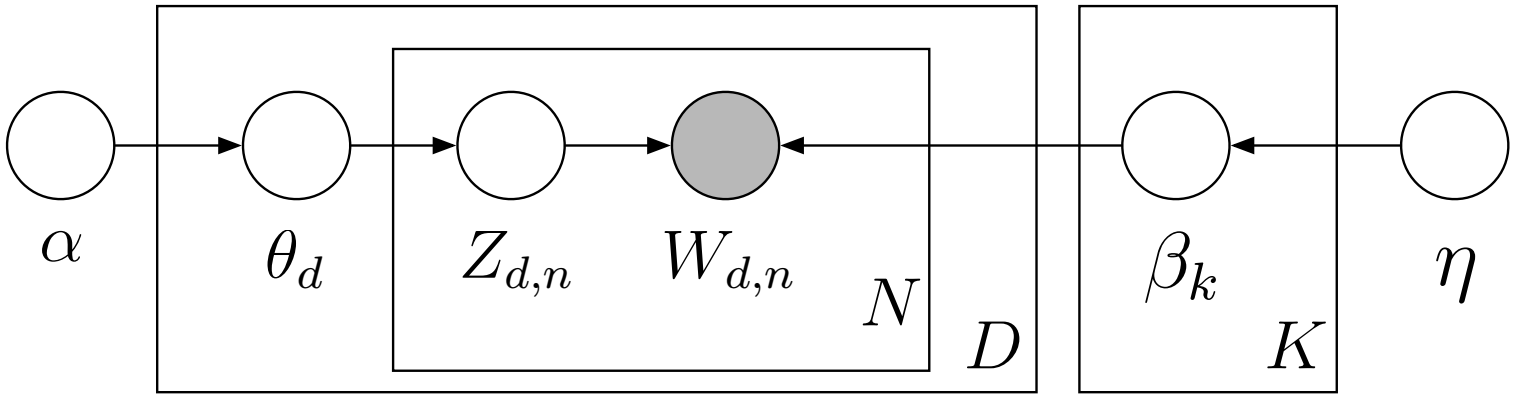
The variational parameters are

γ	Dirichlet parameters
$\phi_{1:N}$	Multinomial parameters for K-dim variables

Contrast this to the model.

- An individual Dirichlet distribution for each document
- An individual multinomial for each word in each document

Variational Inference for LDA



Given topics $\beta_{1:K}$ and words $w_{1:N}$,
optimize the ELBO with coordinate ascent—

$$\begin{aligned} \gamma &= \alpha + \textstyle\sum_{n=1}^{N} \phi_n \\ \phi_n &\propto \exp\{\mathsf{E}[\log \theta] + \log \beta_{\cdot, w_n}\}, \end{aligned}$$

where

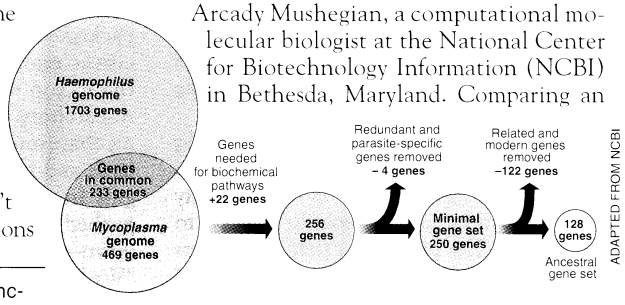
$$\mathsf{E}[\log \theta_i] = \Psi(\gamma_i) - \Psi(\textstyle\sum_j \gamma_j).$$

Example inference (again)

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.

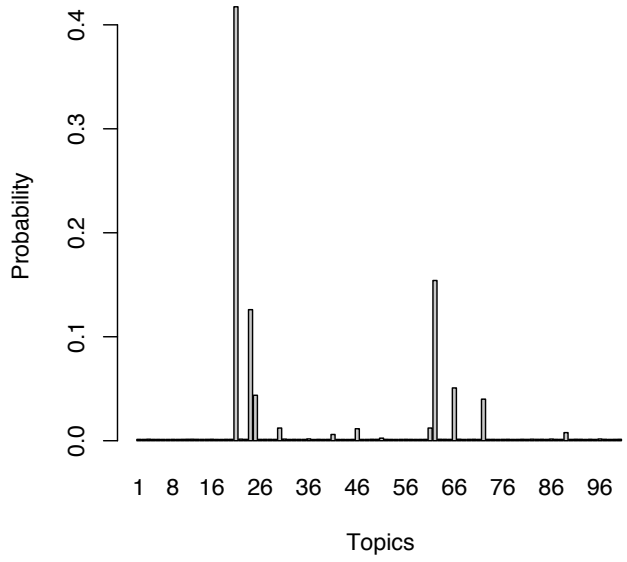
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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



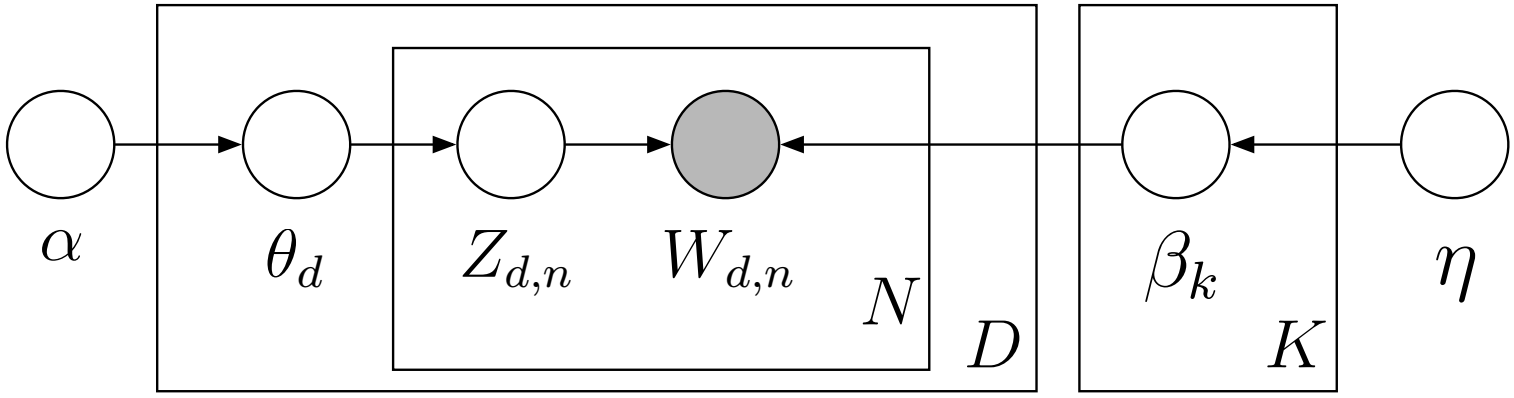
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996



Estimating the topics



- Use a variational Dirichlet for each topic, $q(\beta_k \mid \lambda_k)$.
- After doing per-document inference on each document,

$$\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}$$

- Notice: this is like the true posterior with $\mathrm{E}_q[Z_{d,n}]$.

Batch Variational Inference for LDA

1: Initialize topics $\boldsymbol{\lambda}_{1:K}$ randomly.
2: **while** relative improvement in $\mathcal{L}(\boldsymbol{w}, \phi, \gamma, \boldsymbol{\lambda}) > \epsilon$ **do**
3: **for** $d = 1$ to D **do**
4: Initialize $\gamma_{d,k} = 1$.
5: **repeat**
6: Set $\phi_{d,n} \propto \exp\{\mathbb{E}_q[\log \theta_d] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
7: Set $\gamma_d = \alpha + \sum_n \phi_{d,n}$
8: **until** $\frac{1}{K} \sum_k |\text{change in } \gamma_{d,k}| < \epsilon$
9: **end for**
10: Set $\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}$
11: **end while**

“E step”

1: Initialize topics $\lambda_{1:K}$ randomly.
2: **while** relative improvement in $\mathcal{L}(\boldsymbol{w}, \phi, \gamma, \lambda) > \epsilon$ **do**
3: **for** $d = 1$ to D **do**
4: Initialize $\gamma_{d,k} = 1$.
5: **repeat**
6: Set $\phi_{d,n} \propto \exp\{\mathbb{E}_q[\log \theta_d] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
7: Set $\gamma_d = \alpha + \sum_n \phi_{d,n}$
8: **until** $\frac{1}{K} \sum_k |\text{change in } \gamma_{d,k}| < \epsilon$
9: **end for**
10: Set $\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}$
11: **end while**

Do variational inference for each document.

“M step”

1: Initialize topics $\lambda_{1:K}$ randomly.
2: **while** relative improvement in $\mathcal{L}(\boldsymbol{w}, \phi, \gamma, \lambda) > \epsilon$ **do**
3: **for** $d = 1$ to D **do**
4: Initialize $\gamma_{d,k} = 1$.
5: **repeat**
6: Set $\phi_{d,n} \propto \exp\{\mathbb{E}_q[\log \theta_d] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
7: Set $\gamma_d = \alpha + \sum_n \phi_{d,n}$
8: **until** $\frac{1}{K} \sum_k |$ change in $\gamma_{d,k}| < \epsilon$
9: **end for**
10: Set $\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}$
11: **end while**

Update the posterior estimates of the topics based on the “E step.”

Example topic 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

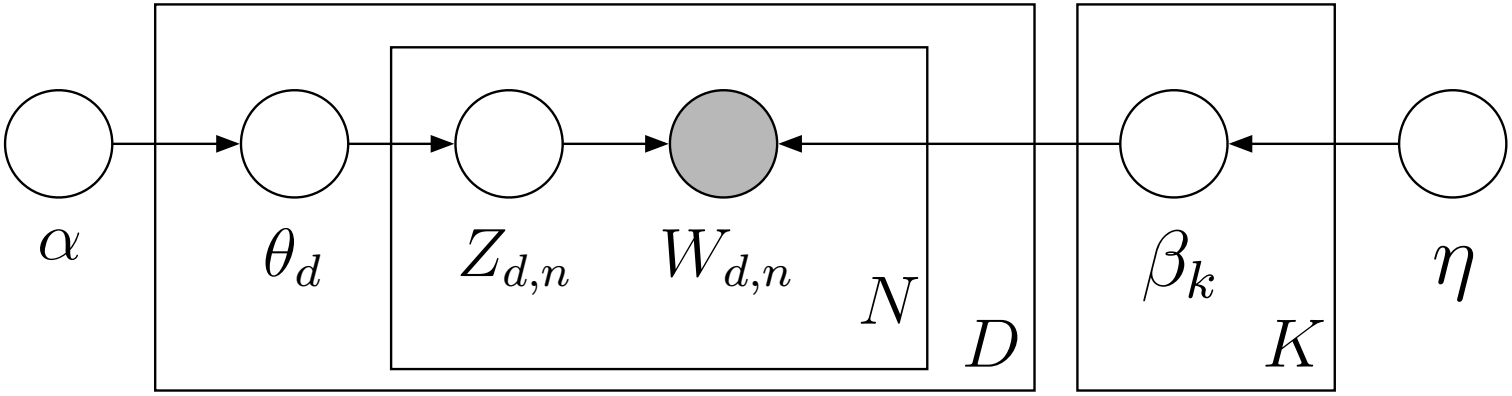
The need for on-line inference

- Our goal is to use this (and related) models for analyzing massive collections of millions of documents.
- But, in the first step of batch inference we estimate the posterior for *every document* based on randomly initialized topics.
- Rather, we use **stochastic natural gradient ascent**.
- The basic procedure
  - Sample a document from a source
  - Process that document and update the model
  - Repeat

A brief review of stochastic optimization

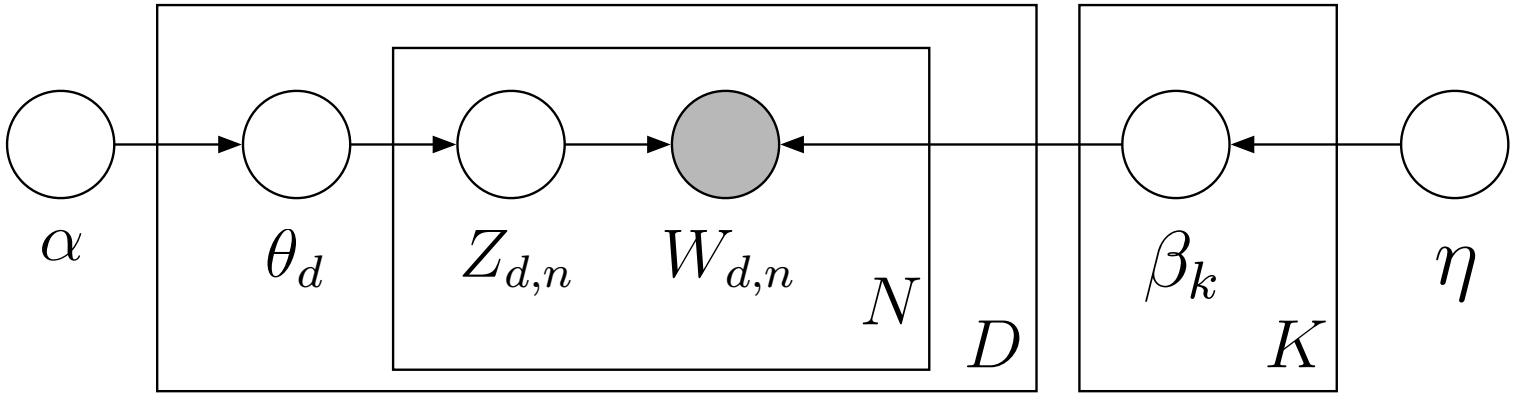
- Why waste time with the real gradient, when a cheaper noisy estimate of the gradient will do (Robbins and Monro, 1951)?
- Idea: Follow a noisy estimate of the gradient with a step-size.
- By decreasing the step-size according to a certain schedule, we guarantee convergence to a local optimum.

Stochastic optimization of the ELBO



- Many models contain "local" and "global" variables.
  - Local variables are drawn for each data point.
  - Global variables are drawn once for the whole data set.
- Idea—
  - Subsample a small subset of the data
  - Do variational inference for the local parameters
  - Do stochastic optimization for the global parameters

Stochastic optimization of the ELBO



- The procedure is to—
  - Draw an index *i* at random
  - Perform inference on document *i* and the current topics
  - Update the (global) topics with stochastic optimization
- No need to process the whole corpus before updating the model.
- Further, no need to keep the corpus around on disk!

On-line variational inference for LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
for $t = 0$ to ∞ **do**
  Choose a random document w_t
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{t,n} \propto \exp\{\mathbb{E}_q[\log \theta_t] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
    Set $\gamma_t = \alpha + \sum_n \phi_{t,n}$
  until $\frac{1}{K}\sum_k |\text{change in } \gamma_{t,k}| < \epsilon$
  Compute $\tilde{\lambda}_k = \eta + D\sum_n w_{t,n}\phi_{t,n}$
  Set $\lambda_k = (1-\rho_t)\lambda_k + \rho_t\tilde{\lambda}_k$.
end for

On-line variational inference for LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
for $t = 0$ to ∞ **do**
  Choose a random document w_t
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{t,n} \propto \exp\{\mathbb{E}_q[\log \theta_t] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
    Set $\gamma_t = \alpha + \sum_n \phi_{t,n}$
  until $\frac{1}{K}\sum_k |\text{change in } \gamma_{t,k}| < \epsilon$
  Compute $\tilde{\lambda}_k = \eta + D\sum_n w_{t,n}\phi_{t,n}$
  Set $\lambda_k = (1-\rho_t)\lambda_k + \rho_t\tilde{\lambda}_k$.
end for

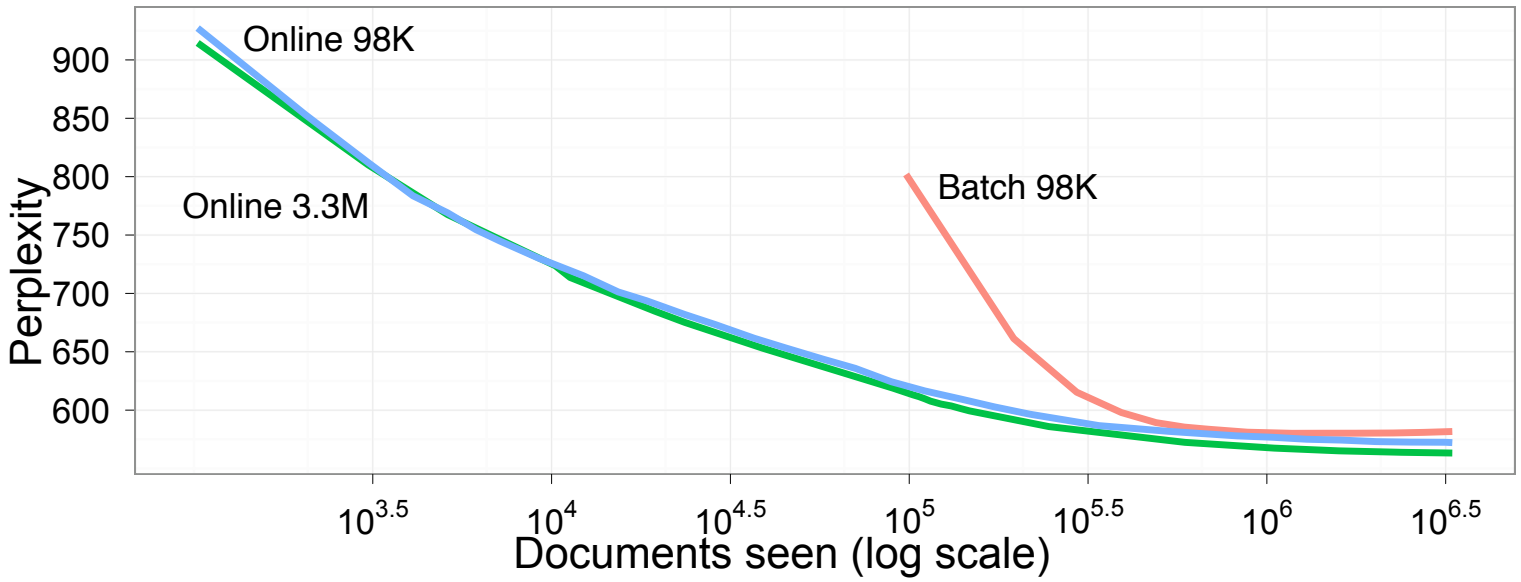
The E-step only processes a single document.

On-line variational inference for LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
for $t = 0$ to ∞ **do**
  Choose a random document w_t
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{t,n} \propto \exp\{\mathbb{E}_q[\log \theta_t] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
    Set $\gamma_t = \alpha + \sum_n \phi_{t,n}$
  until $\frac{1}{K}\sum_k |\text{change in } \gamma_{t,k}| < \epsilon$
  Compute $\tilde{\lambda}_k = \eta + D\sum_n w_{t,n}\phi_{t,n}$
  Set $\lambda_k = (1-\rho_t)\lambda_k + \rho_t\tilde{\lambda}_k$.
end for

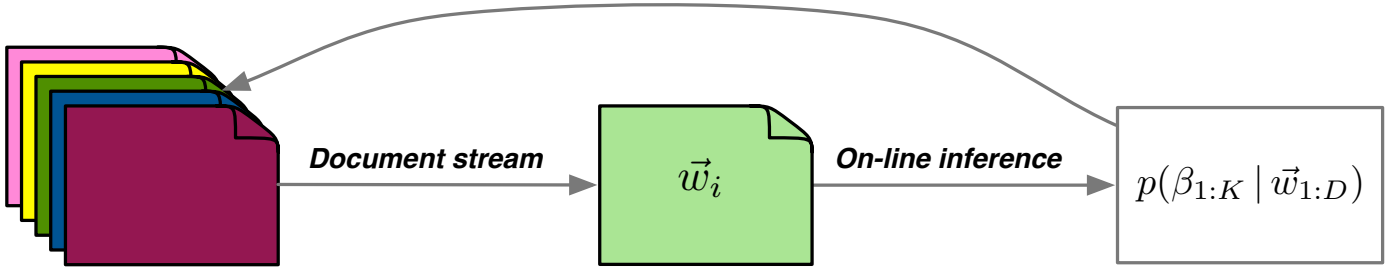
The M-step treats that document as the whole corpus.
But, it only slightly adjusts the topics based on it.

Analyzing 3.3M articles from Wikipedia



Documents analyzed	2048	4096	8192	12288	16384	32768	49152	65536
Top eight words	systems	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

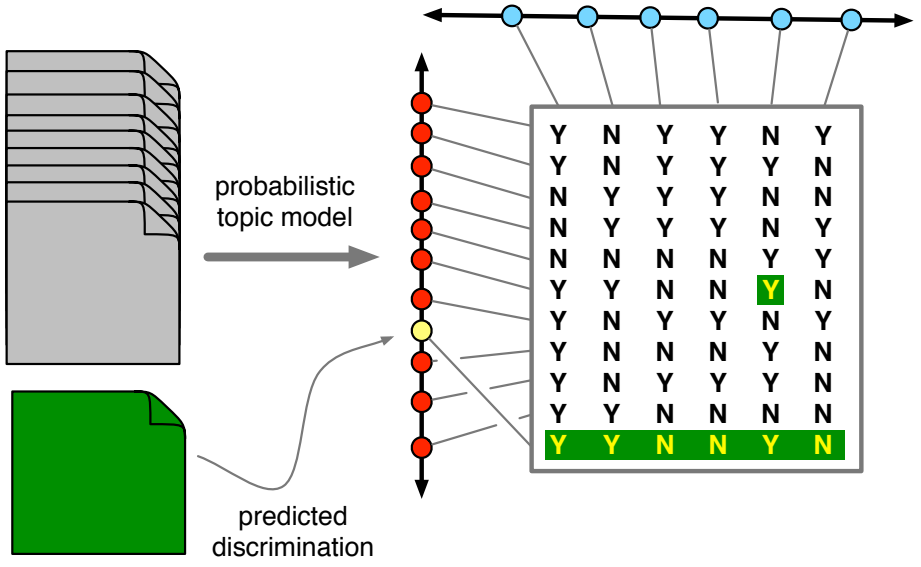
On-line variational inference for LDA



- We can build topic models from streams of documents.
- Some issues
  - How quickly to decrease the learning rate?
  - What is the "batch size"?
  - How to interpret *D* when there is a stream of documents?
- See the paper by Hoffman et al. (NIPS, 2010) and the foundational related work of Sato (Neural Computation, 2001).

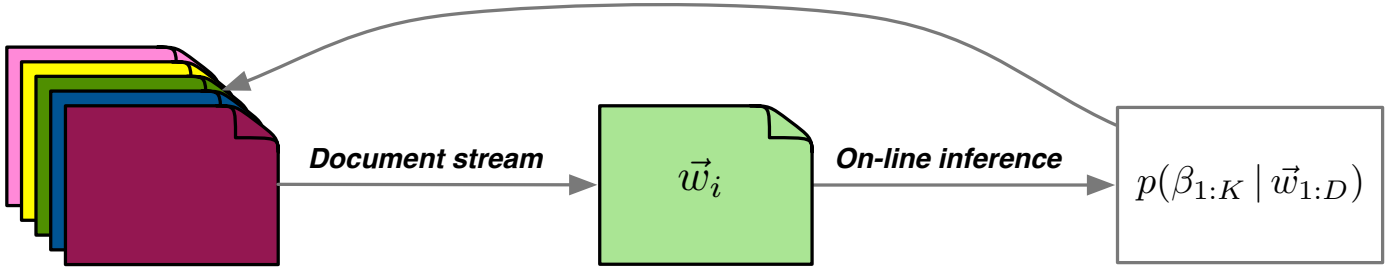
Summary

Ideal point topic model



- Ideal point topic models illustrate how probabilistic topic modeling can be used to solve a real-world problem
- Our current research—
  - On-line inference for this model
  - Modeling changing tastes and preferences
  - A scalable recommendation system for scientific articles

On-line inference for LDA



- Stochastic optimization and variational methods provide a way to approximate the posterior for massive and streaming data sets.
- This combination is very powerful for topic modeling.
  It can be adapted to hierarchical models for many data
  (e.g., biological data, natural images, network data)
- Our current research—
  - On-line methods for Bayesian nonparametric models
  - Working with non-conjugate priors and natural gradients
  - Guiding the stream towards points that are poorly modeled