Topic Models

David M. Blei

Department of Computer Science
Princeton University

September 1, 2009

The problem with information

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

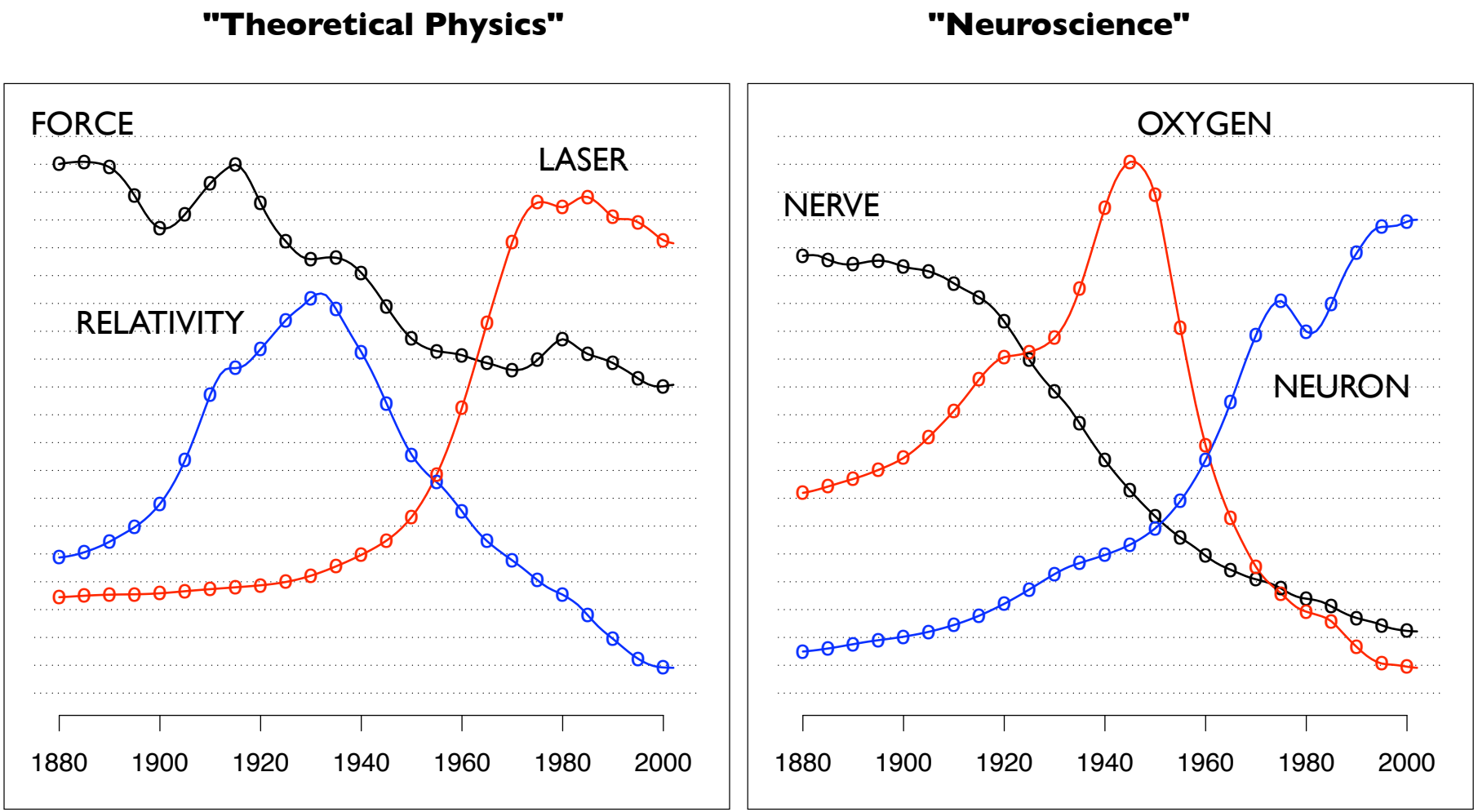
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

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

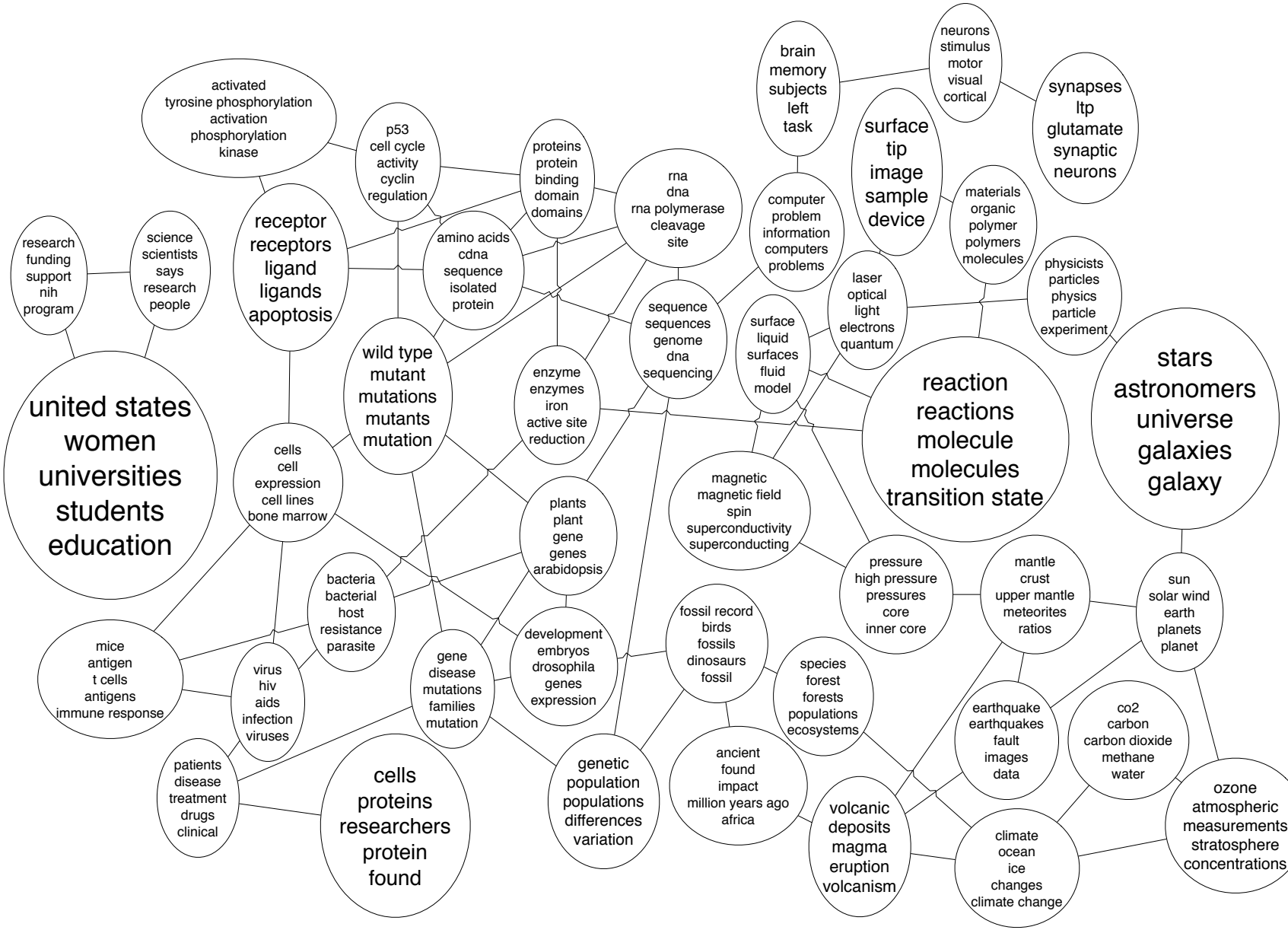
Discover topics from a corpus

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



Annotate images

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

Topic modeling topics

From a machine learning perspective, topic modeling is a case study in applying hierarchical Bayesian models to grouped data, like documents or images. Topic modeling research touches on

- Directed graphical models
- Conjugate priors and nonconjugate priors
- Time series modeling
- Modeling with graphs
- Hierarchical Bayesian methods
- Fast approximate posterior inference (MCMC, variational methods)
- Exploratory data analysis
- Model selection and nonparametric Bayesian methods
- Mixed membership models

Latent Dirichlet allocation (LDA)

1. Introduction to LDA
2. The posterior distribution for LDA

Approximate posterior inference

1. Gibbs sampling
2. Variational inference
3. Comparison/Theory/Advice

Other topic models

1. Topic models for prediction: Relational and supervised topic models
2. The logistic normal: Dynamic and correlated topic models
3. "Infinite" topic models, i.e., the hierarchical Dirichlet process

Interpreting and evaluating topic models

Latent Dirichlet Allocation

Probabilistic modeling

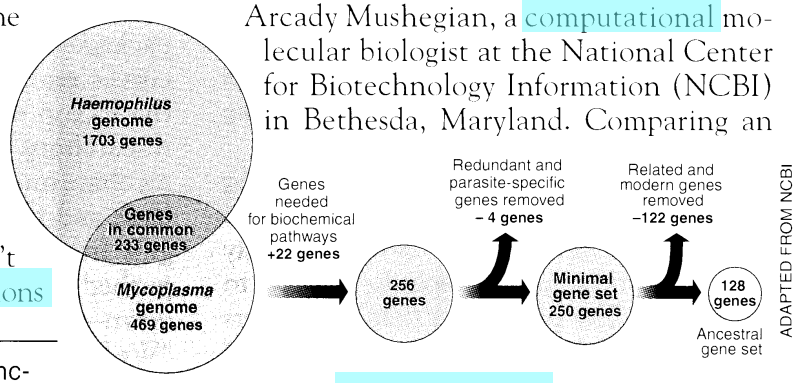
1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using *posterior inference*
   - What are the topics that describe this 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



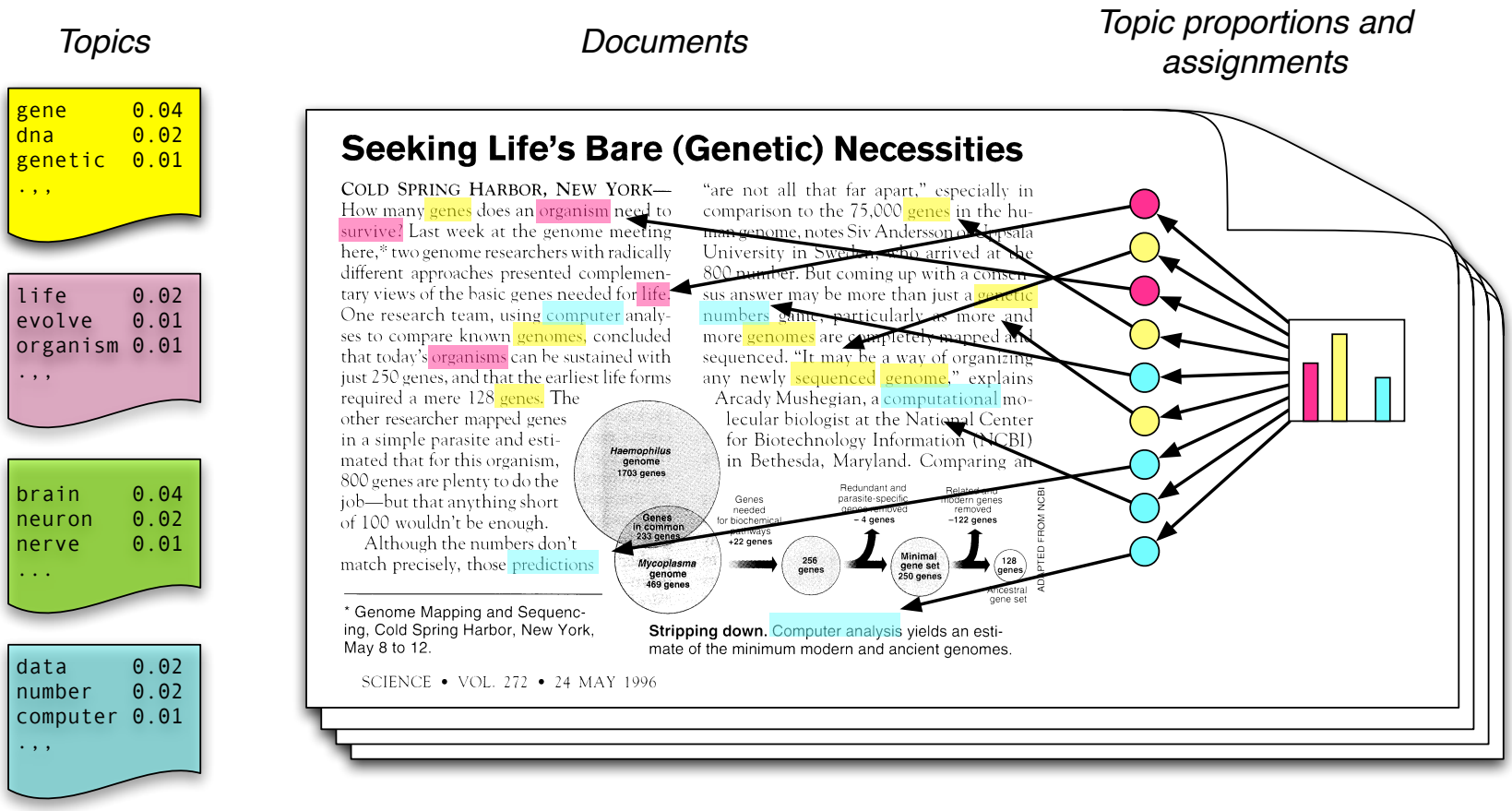
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

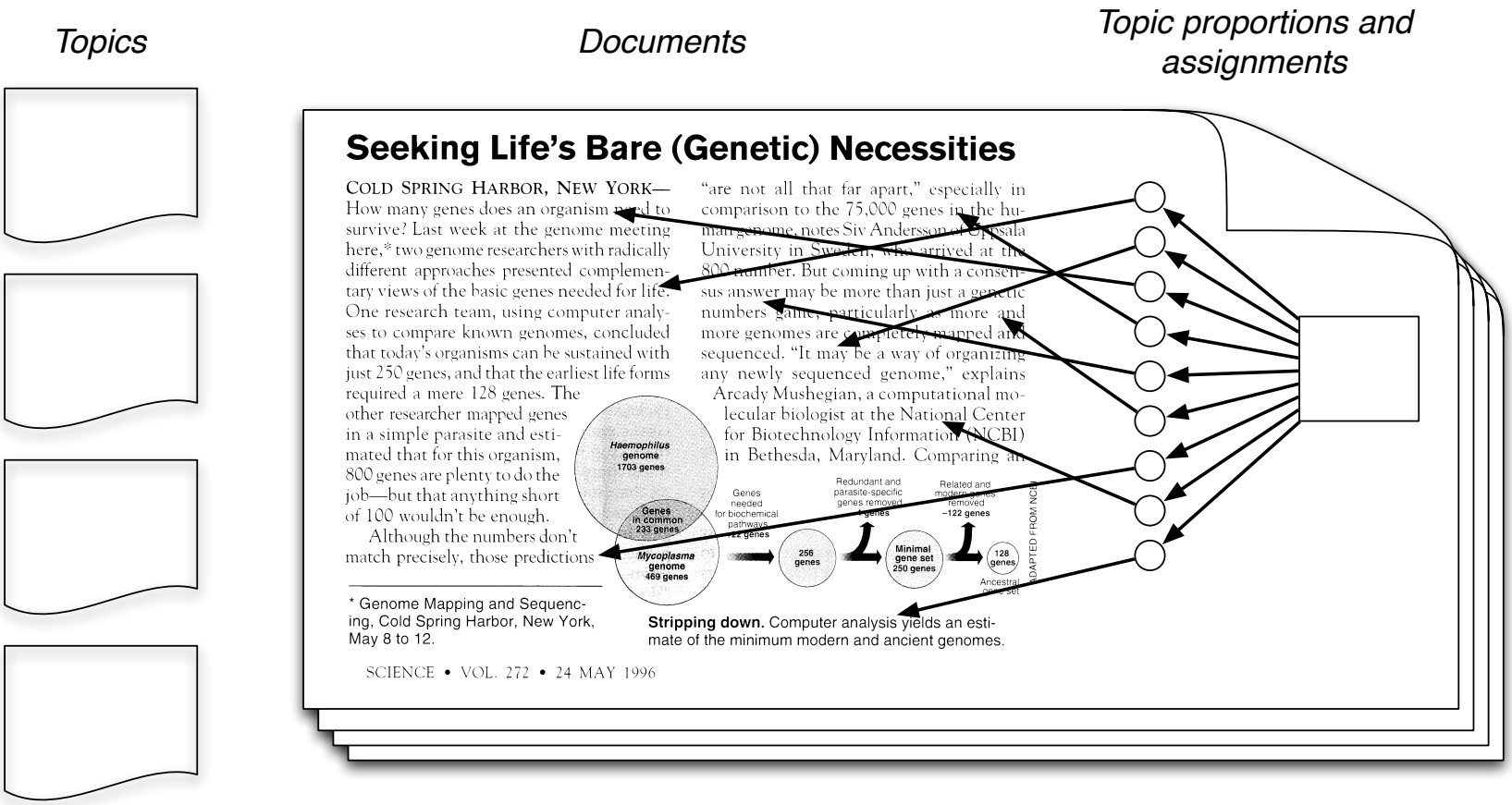
Simple intuition: Documents exhibit multiple topics.

Generative model



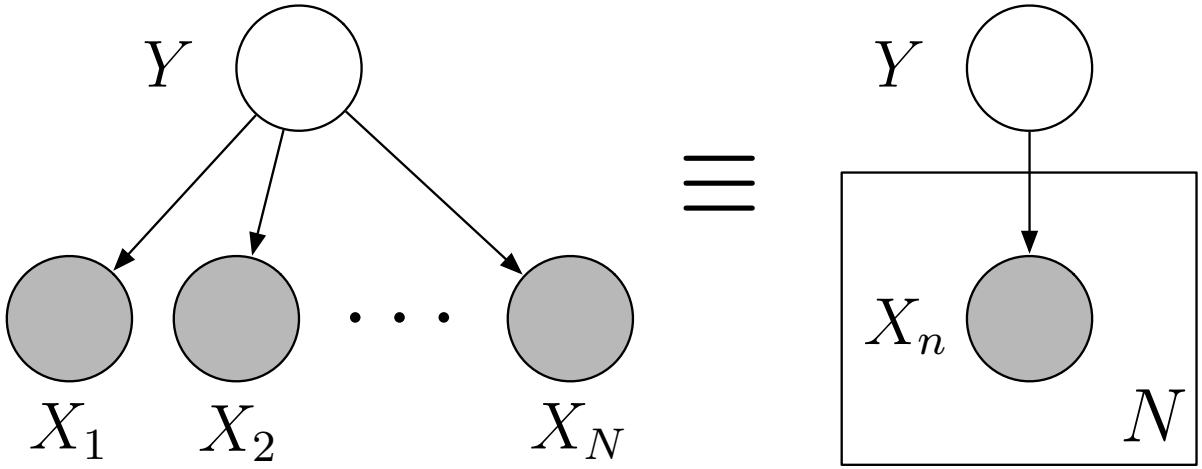
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

The posterior distribution



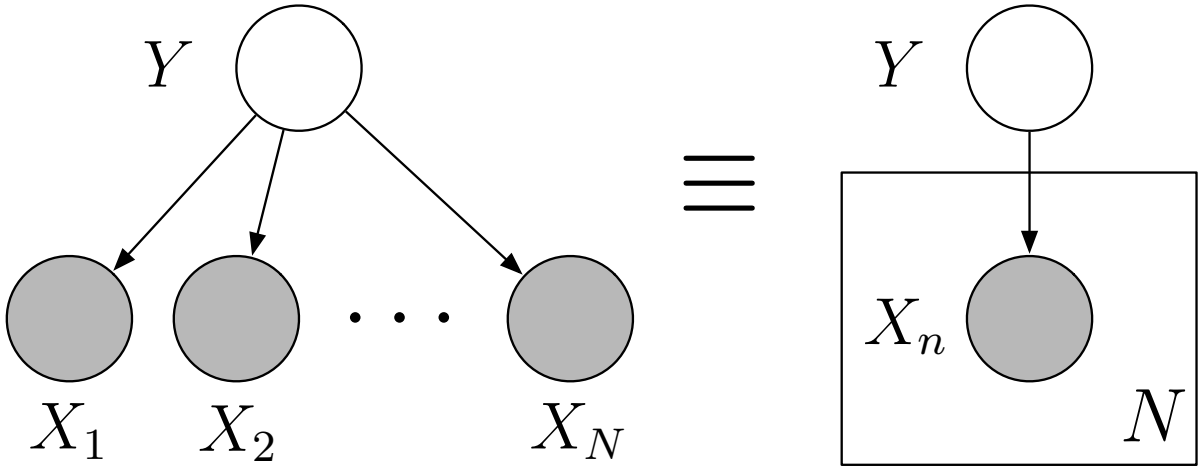
- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

Graphical models (Aside)



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

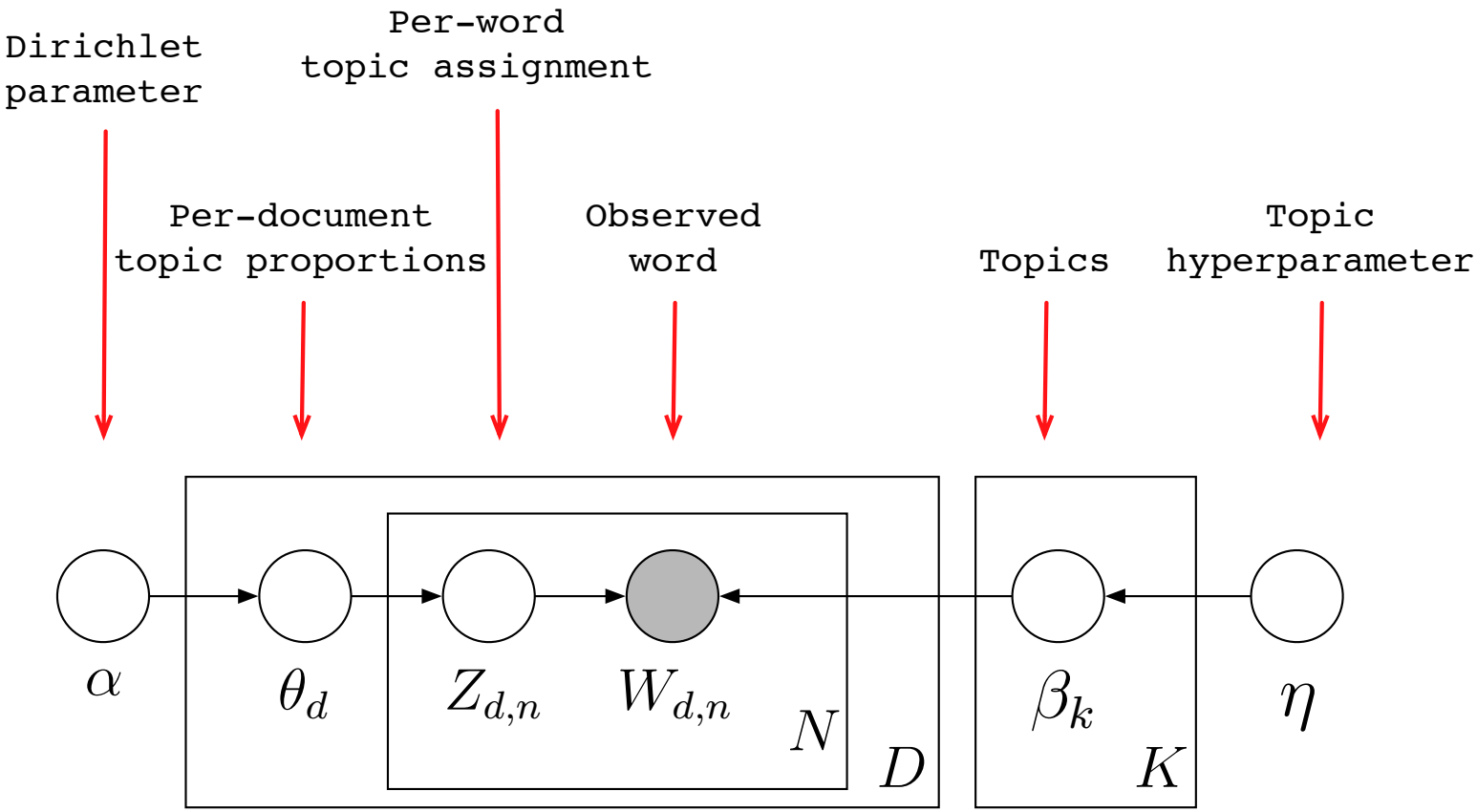
Graphical models (Aside)



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

Latent Dirichlet allocation



Each piece of the structure is a random variable.

The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

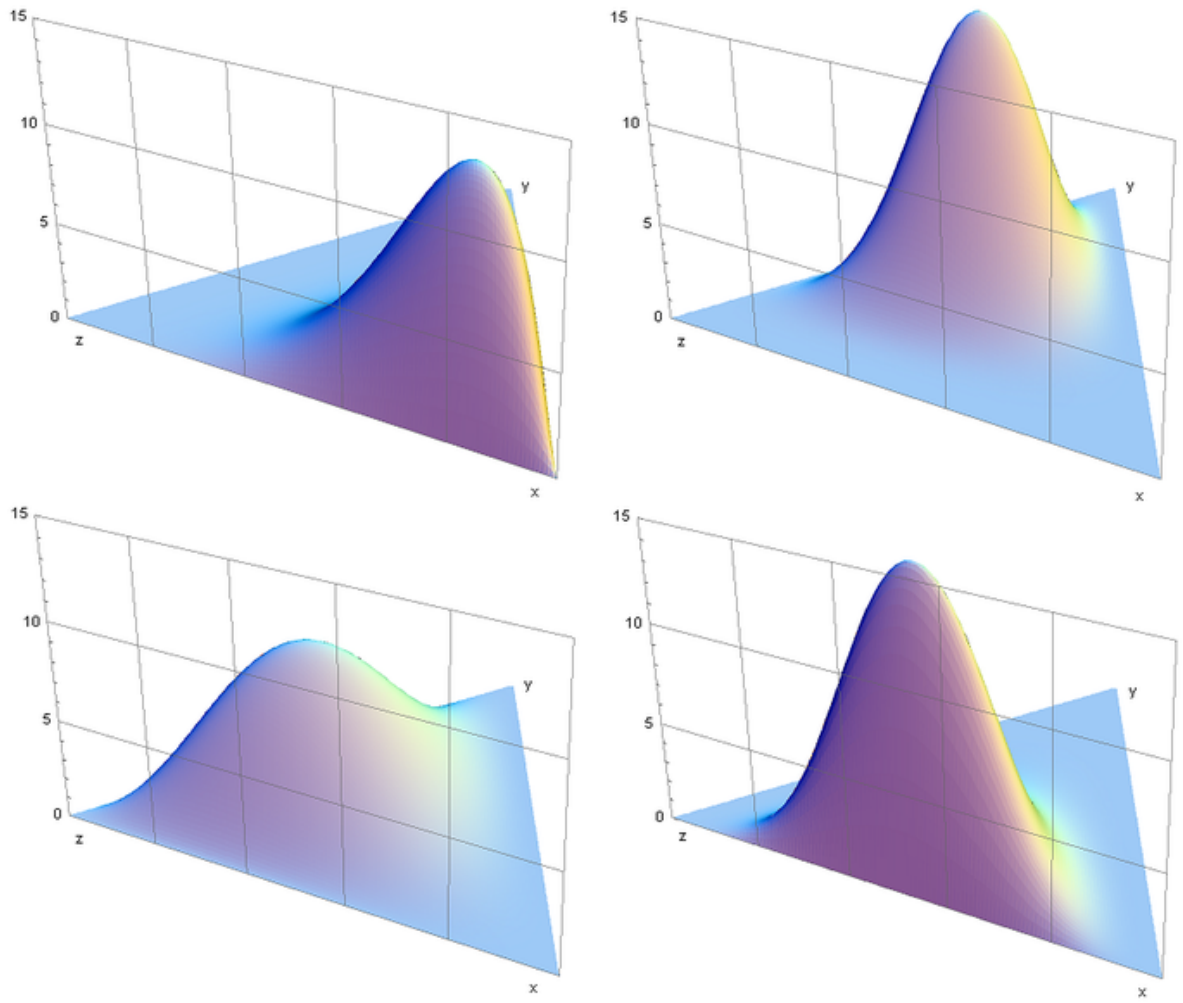
$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- The Dirichlet is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.

- The parameter α controls the mean shape and sparsity of θ.

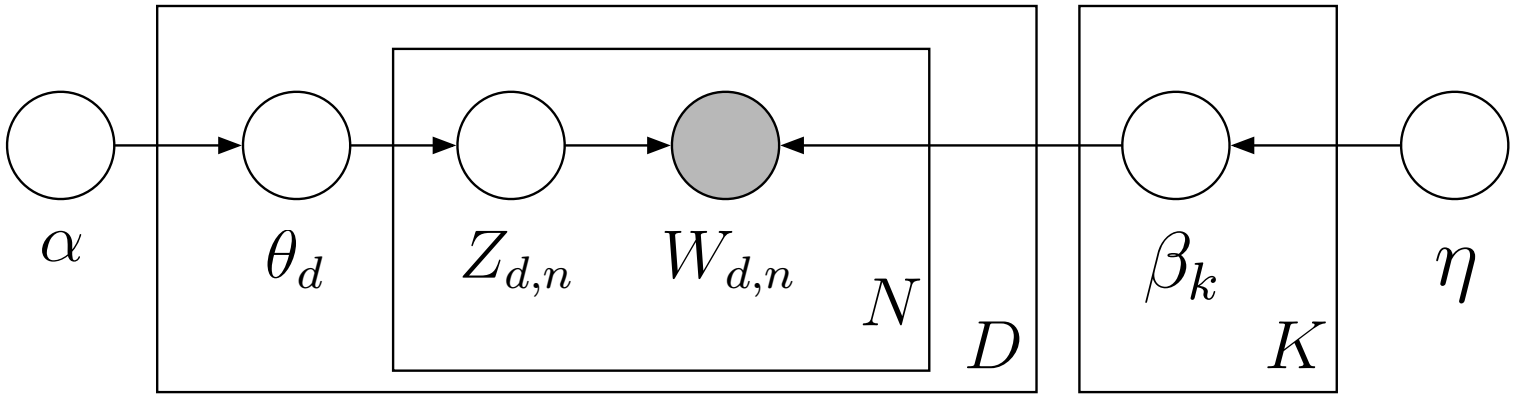
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

The Dirichlet distribution



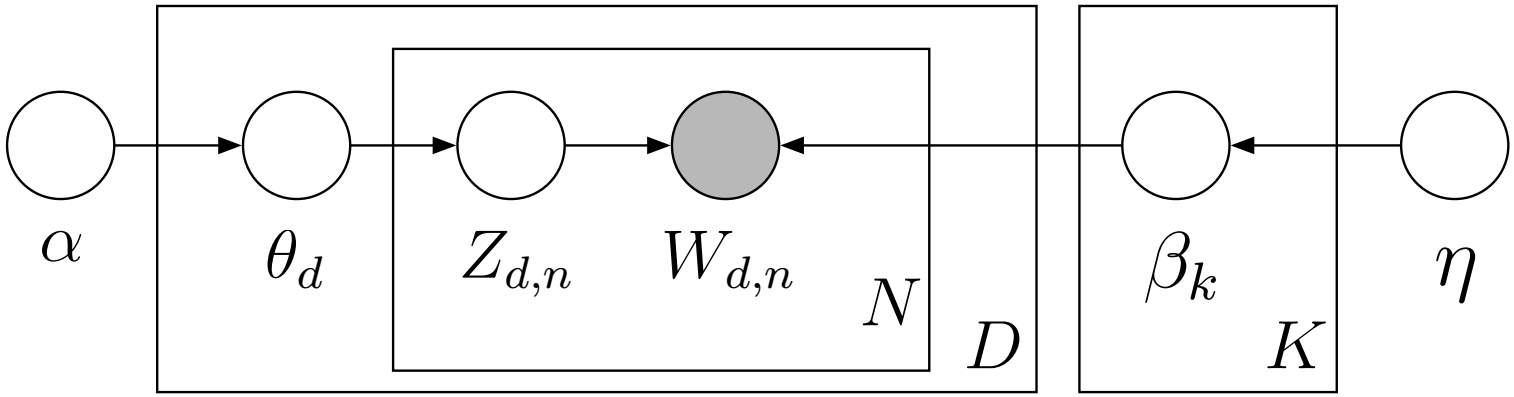
(From Wikipedia)

Latent Dirichlet allocation



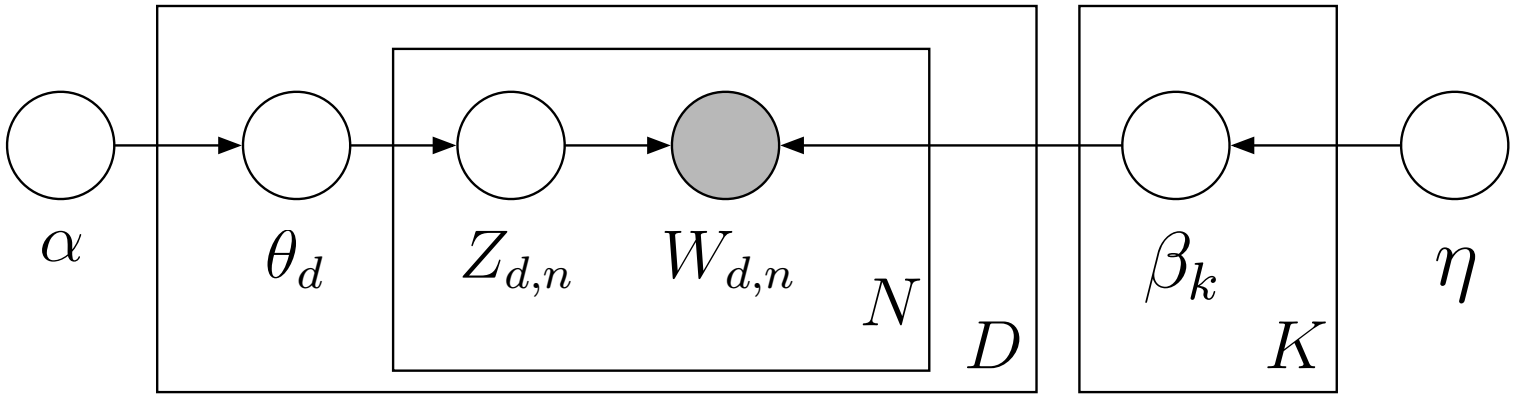
- LDA is a mixed membership model (Erosheva, 2004) that builds on the work of Deerwester et al. (1990) and Hofmann (1999).
- For document collections and other grouped data, this might be more appropriate than a simple finite mixture.
- The same model was independently invented for population genetics analysis (Pritchard et al., 2000).

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

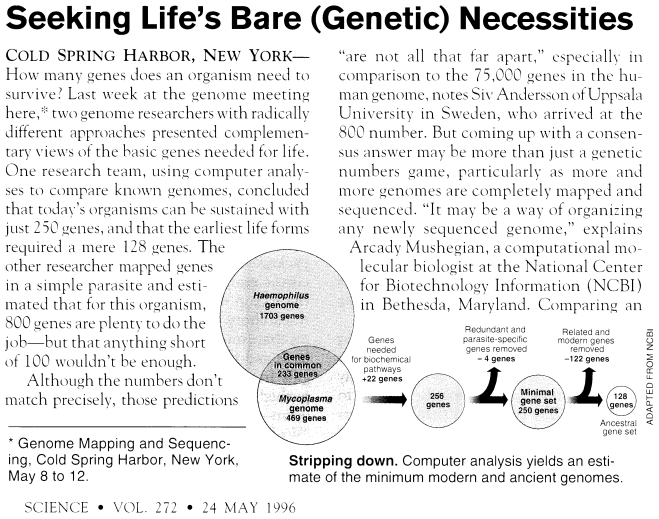


Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)

For comparison, see Mukherjee and Blei (2009) and Asuncion et al. (2009).

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

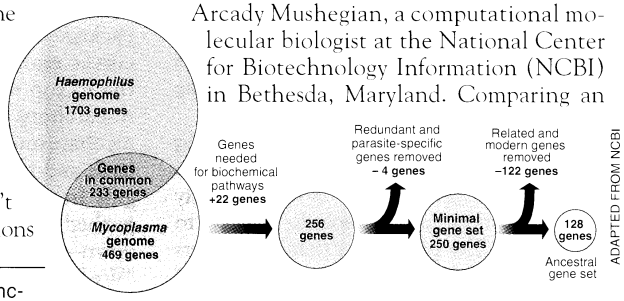
- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

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.

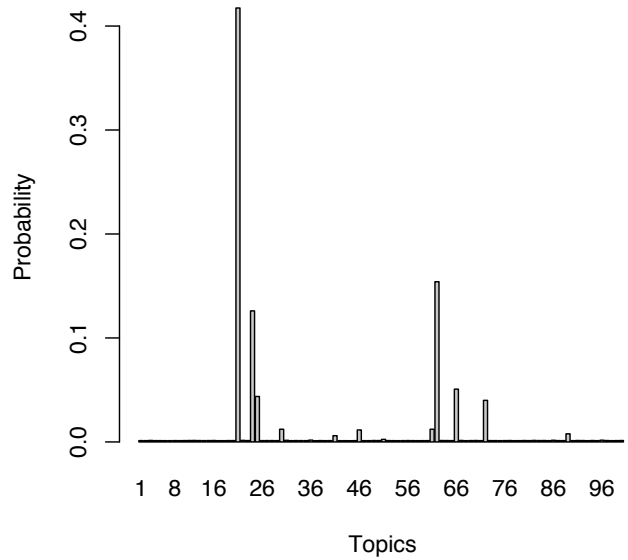
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

SCIENCE • VOL. 272 • 24 MAY 1996



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)

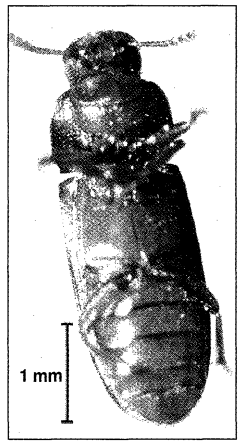
Chaotic Beetles

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 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*).



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 to explore and browse document collections

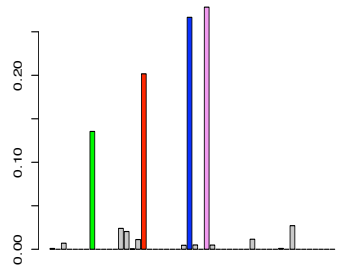
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

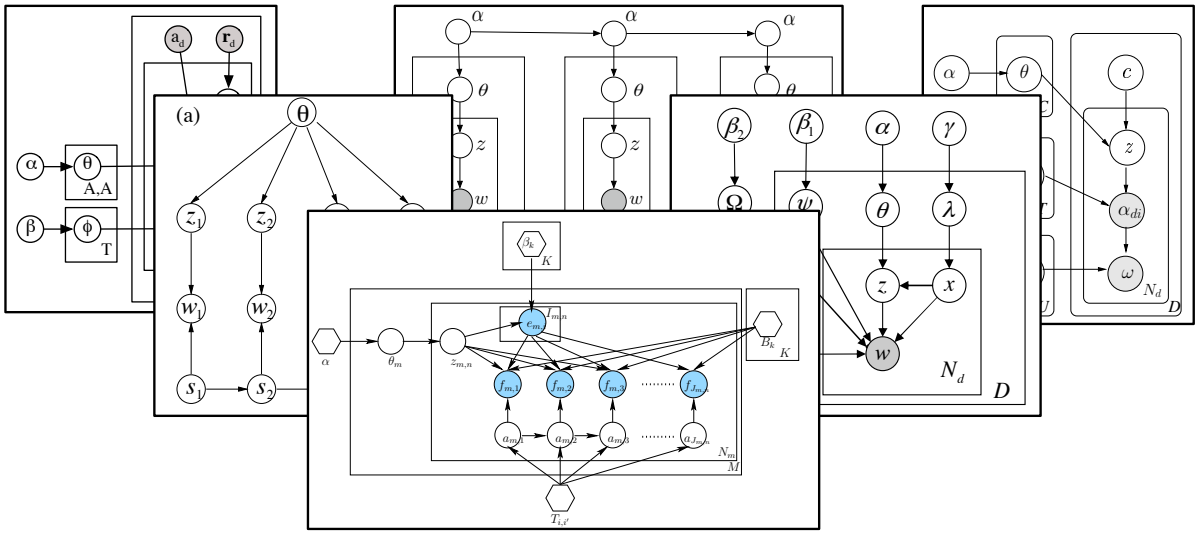
Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Why does LDA "work"?

Why does the LDA posterior put "topical" words together?

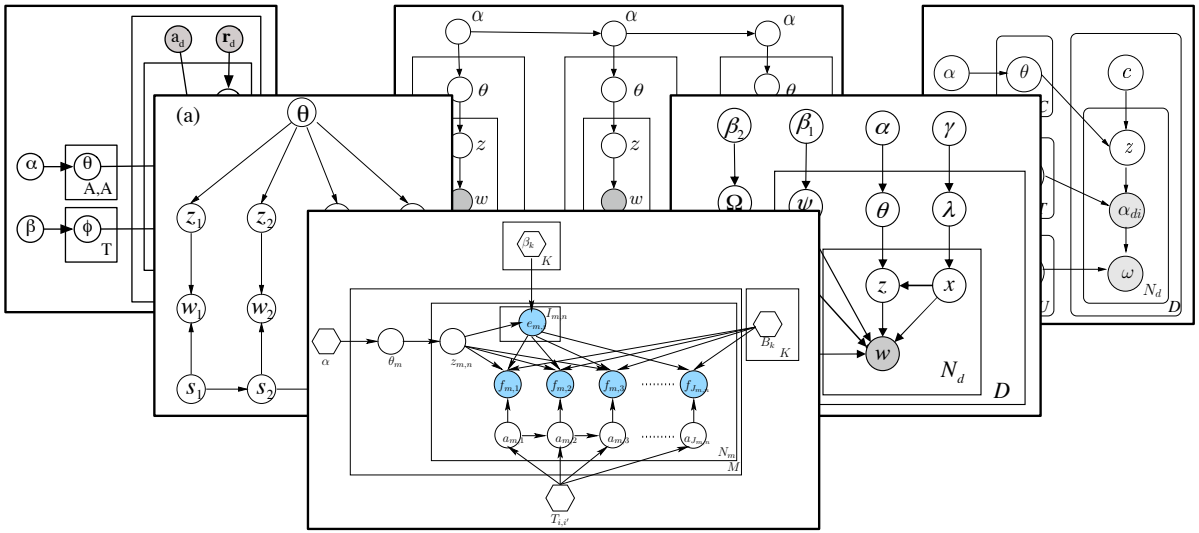
- Word probabilities are maximized by dividing the words among the topics. (More terms means more mass to be spread around.)
- In a mixture, this is enough to find clusters of co-occurring words.
- In LDA, the Dirichlet on the topic proportions can encourage sparsity, i.e., a document is penalized for using many topics.
- Loosely, this can be thought of as softening the strict definition of "co-occurrence" in a mixture model.
- This flexibility leads to sets of terms that more tightly co-occur.

LDA is modular, general, useful



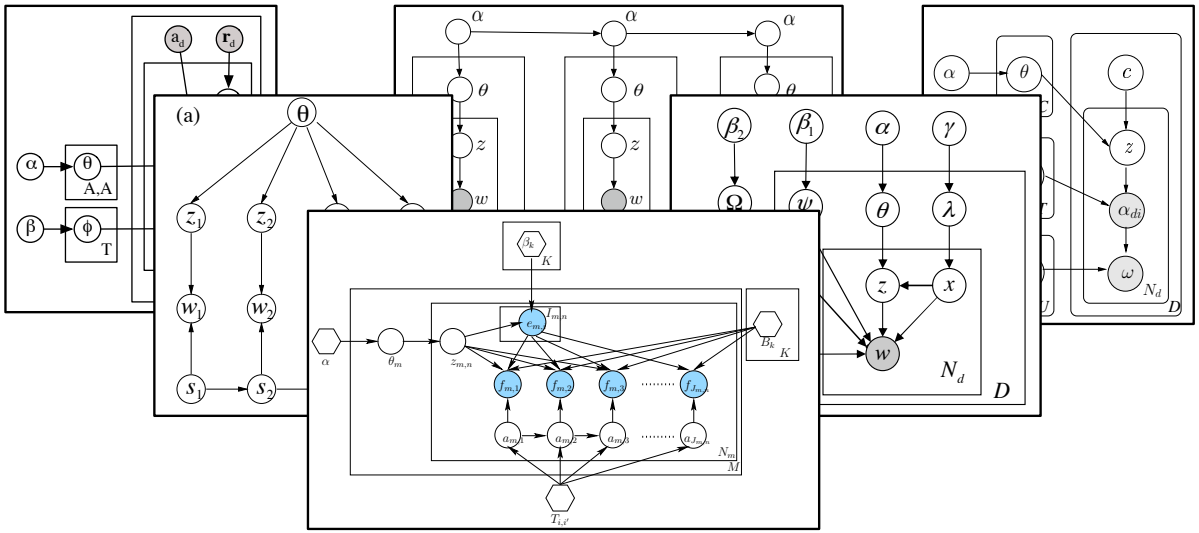
- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
- E.g., syntax; authorship; word sense; dynamics; correlation; hierarchies; nonparametric Bayes

LDA is modular, general, useful



- The **data generating distribution** can be changed.
- E.g., images, social networks, music, purchase histories, computer code, genetic data, click-through data; ...

LDA is modular, general, useful



- The **posterior** can be used in creative ways
- E.g., IR, collaborative filtering, document similarity, visualizing interdisciplinary documents

Approximate posterior inference

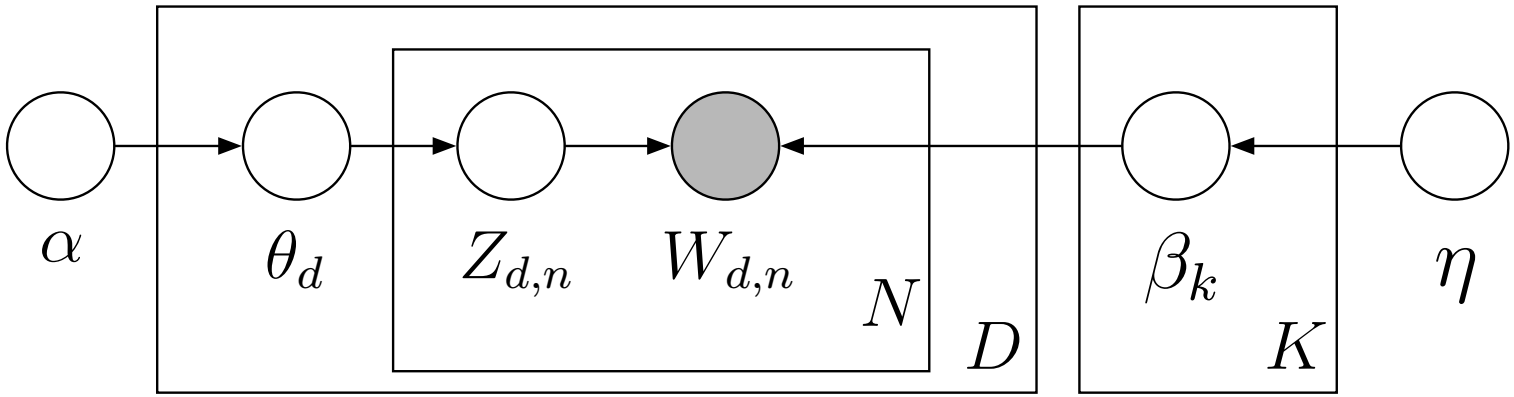
Posterior distribution for LDA

- For now, assume the topics $\beta_{1:K}$ are fixed.
  The per-document posterior is

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

- This is intractable to compute
- It is a "multiple hypergeometric function" (see Dickey, 1983)
- Can be seen as sum of N^K (tractable) Dirichlet integral terms

Posterior distribution for LDA



We appeal to approximate posterior inference of the posterior,

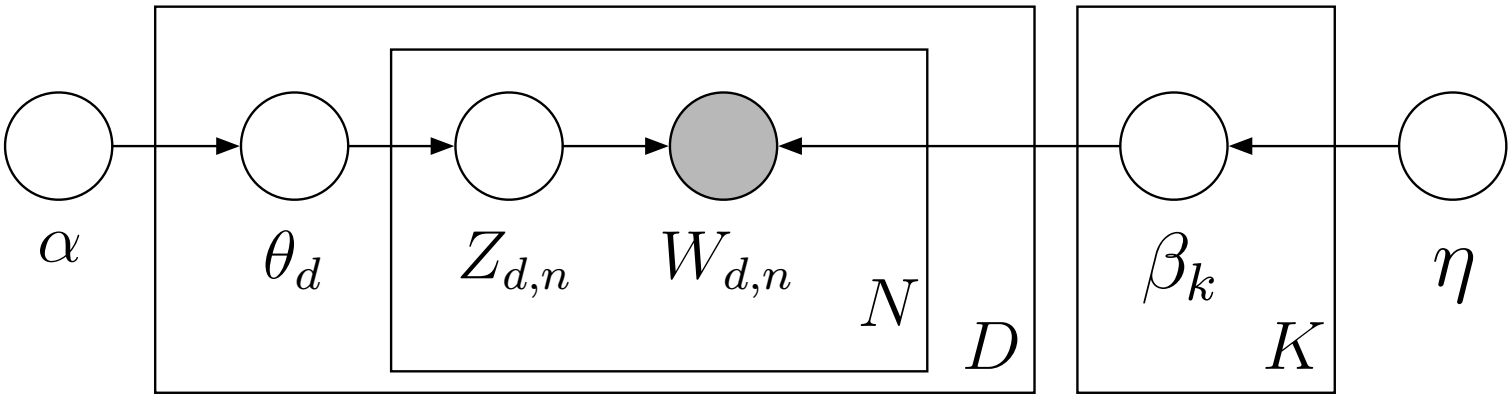
$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

- Gibbs sampling
- Variational methods
- Particle filtering

Gibbs sampling

- Define a **Markov chain** whose stationary distribution is the posterior of interest
- Collect **independent samples** from that stationary distribution; approximate the posterior with them
- In **Gibbs sampling**, the space of the MC is the space of possible configurations of the hidden variables.
- The chain is run by iteratively sampling from the conditional distribution of each hidden variable given observations and the current state of the other hidden variables
- Once a chain has "burned in," collect samples at a lag to approximate the posterior.

Gibbs sampling for LDA



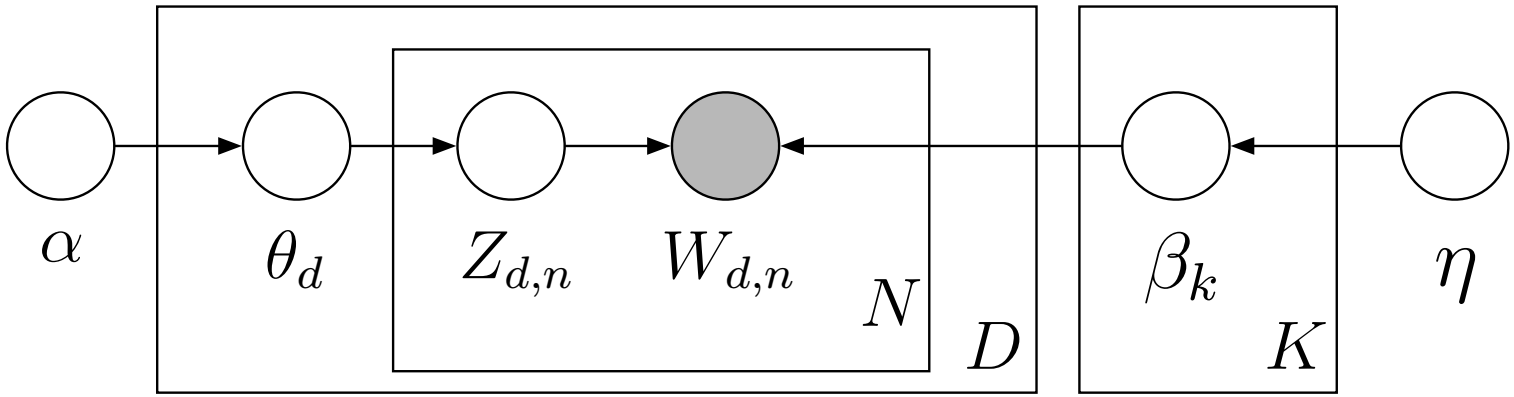
Define $n(z_{1:N})$ to be the counts vector. A simple Gibbs sampler is

$$\begin{aligned} \theta \mid w_{1:N}, z_{1:N} &\sim \mathrm{Dir}(\alpha + n(z_{1:N})) \\ z_i \mid z_{-i}, w_{1:N} &\sim \mathrm{Mult}(\pi(z_{-i}, w_i)) \end{aligned}$$

where

$$\pi(z_{-i}, w_i) \propto (\alpha + n(z_{1:N}))p(w_i \mid \beta_{1:K})$$

Gibbs sampling for LDA



- The topic proportions θ can be integrated out.
- A **collapsed Gibbs sampler** draws from

$$p(z_i \mid z_{-i}, w_{1:N}) \propto p(w_i \mid \beta_{1:K}) \prod_{k=1}^{K} \Gamma(n_k(z_{-i})),$$

  where $n_k(z_{-i})$ is the number of times we've seen topic k in the collection of topic assignments z_{-i}.
- Integrating out variables leads to a faster mixing chain.

Variational inference (in general)

- Variational methods are a deterministic alternative to MCMC.
- Let $x_{1:N}$ be observations and $z_{1:M}$ be latent variables
- Our goal is to compute the posterior distribution

$$p(z_{1:M} \mid x_{1:N}) = \frac{p(z_{1:M}, x_{1:N})}{\int p(z_{1:M}, x_{1:N}) dz_{1:M}}$$

- For many interesting distributions, the marginal likelihood of the observations is difficult to efficiently compute

Variational inference

- Use Jensen's inequality to bound the log prob of the observations:

$$\begin{aligned}
\log p(x_{1:N}) &= \log \int p(z_{1:M}, x_{1:N}) dz_{1:M} \\
&= \log \int p(z_{1:M}, x_{1:N}) \frac{q_\nu(z_{1:M})}{q_\nu(z_{1:M})} dz_{1:M} \\
&\geq \mathrm{E}_{q_\nu}[\log p(z_{1:M}, x_{1:N})] - \mathrm{E}_{q_\nu}[\log q_\nu(z_{1:M})]
\end{aligned}$$

- We have introduced a distribution of the latent variables with free *variational parameters* ν.
- We optimize those parameters to tighten this bound.
- This is the same as finding the member of the family q_ν that is closest in KL divergence to $p(z_{1:M} \mid x_{1:N})$.

Mean-field variational inference

- Complexity of optimization is determined by the factorization of q_ν
- In *mean field variational inference* we choose q_ν to be fully factored

$$q_\nu(z_{1:M}) = \prod_{m=1}^{M} q_{\nu_m}(z_m).$$

- The latent variables are independent.
  - Each is governed by its own variational parameter ν_m.
- In the true posterior they can exhibit dependence (often, this is what makes exact inference difficult).

MFVI and conditional exponential families

- Suppose the distribution of each latent variable conditional on the observations and other latent variables is in the exponential family:

$$p(z_m \mid \mathbf{z}_{-m}, \mathbf{x}) = h_m(z_m) \exp\{g_m(\mathbf{z}_{-m}, \mathbf{x})^T z_m - a_m(g_i(\mathbf{z}_{-m}, \mathbf{x}))\}$$

- Assume q_ν is fully factorized, and each factor is in the same exponential family:

$$q_{\nu_m}(z_m) = h_m(z_m) \exp\{\nu_m^T z_m - a_m(\nu_m)\}$$

MFVI and conditional exponential families

- Variational inference is the following coordinate ascent algorithm

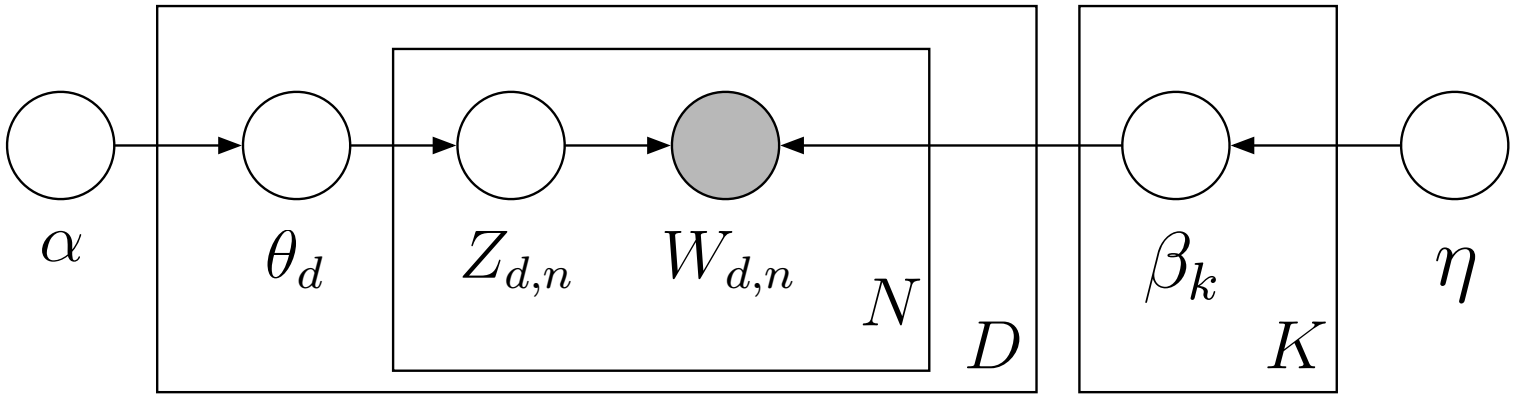
$$\nu_m = \mathrm{E}_{q_\nu}[g_m(\mathbf{Z}_{-m}, \mathbf{x})]$$

- Notice the relationship to Gibbs sampling.
- (You will hear much more about this from Minka and Winn.)

Variational inference

- Alternative to MCMC; replace sampling with optimization.
- Deterministic approximation to posterior distribution.
- Uses established optimization methods
  (block coordinate ascent; Newton-Raphson; interior-point).
- Faster, more scalable than MCMC for large problems.
- Biased, whereas MCMC is not.
- Emerging as a useful framework for fully Bayesian and empirical Bayesian inference problems.

Variational Inference for LDA

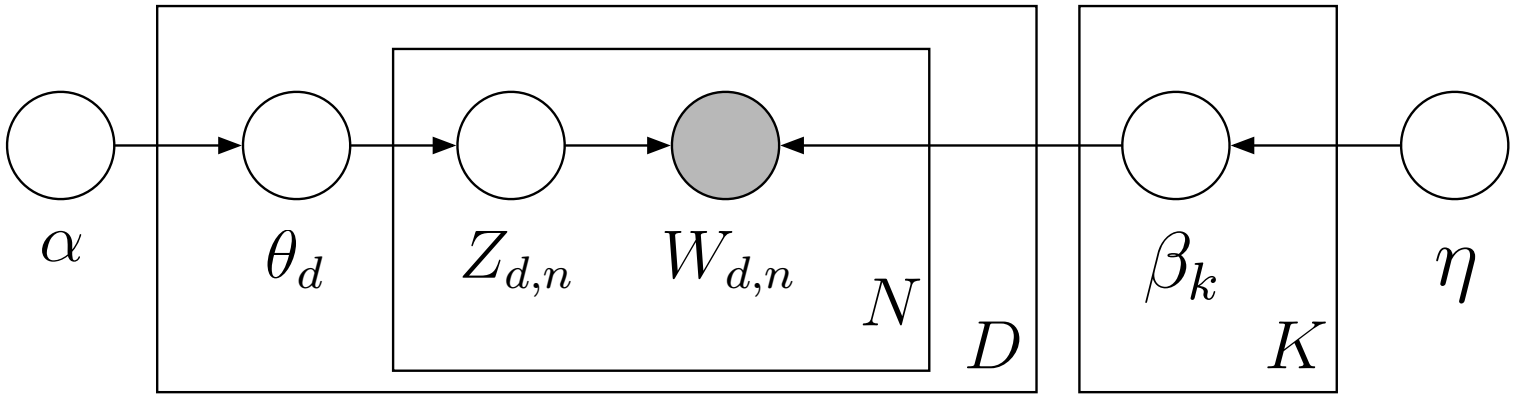


- The mean field variational distribution is

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

- This is a family of distributions over the latent variables, where all variables are independent and governed by their own parameters.
- In the true posterior, the latent variables are **not** independent.

Variational Inference for LDA

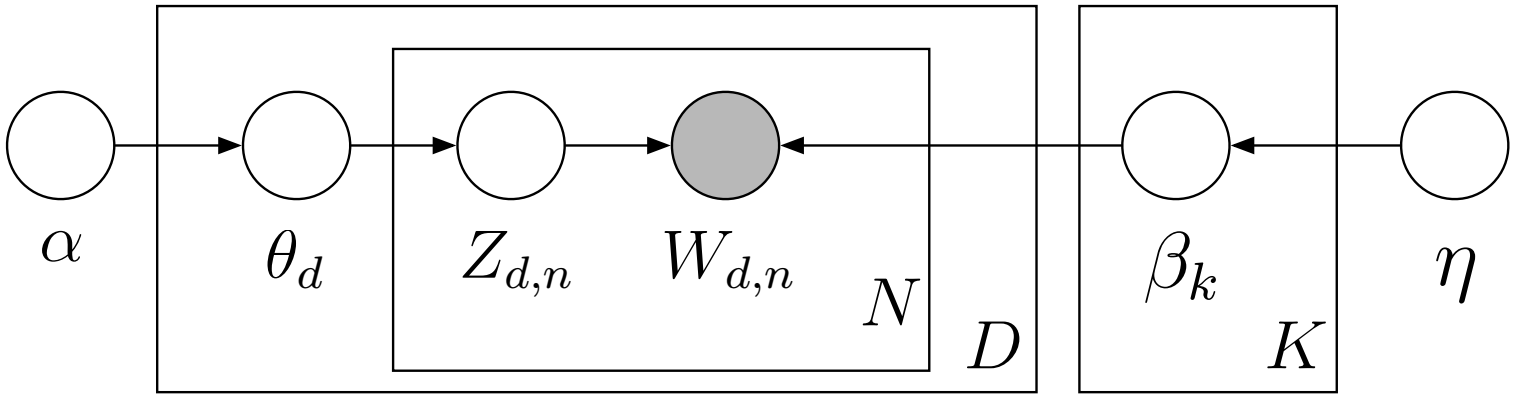


The variational paramters are:

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

There is a separate variational Dirichlet distribution for each document; there is a separate multinomial distribution for each word in each document. (Contrast this to the model.)

Variational Inference for LDA



Coordinate ascent on the variational objective,

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

where

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

Estimating the topics

Maximum likelihood: Expectation-Maximization

- E-step: Use variational or MCMC to approximate the per-document posterior
- M-step: Find MLE of $\beta_{1:K}$ from expected counts

Bayesian topics

- Put a Dirichlet prior on the topics (usually exchangeable)
  Note/Warning: This controls the sparsity of the topics
- Collapsed Gibbs sampling is still possible—we only need to keep track of the topic assignments.
- Variational: Use a variational Dirichlet for each topic

Inference comparison

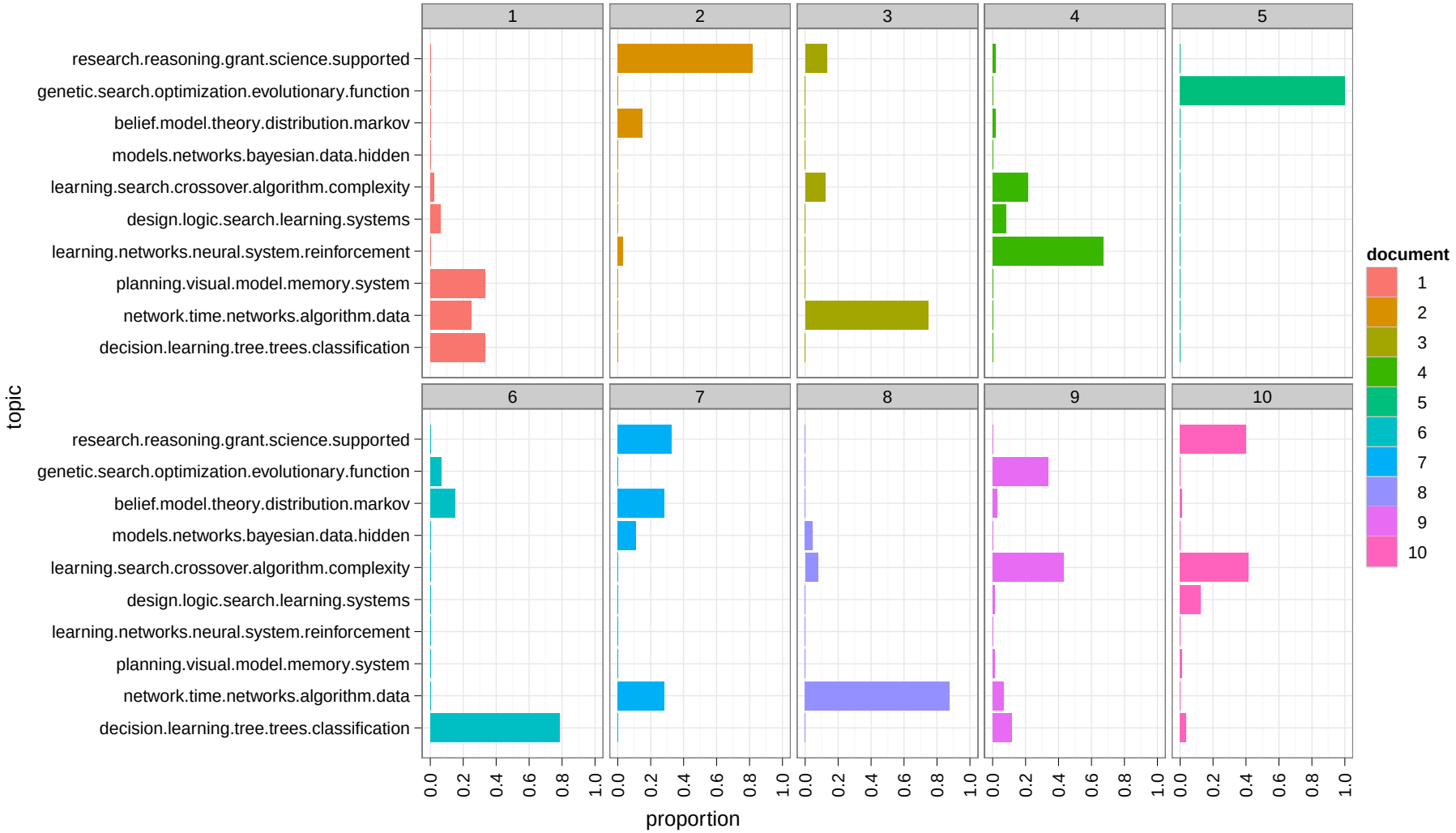
- Conventional wisdom says that:
  - Gibbs is easiest to implement
  - Variational can be faster, especially when dealing with nonconjugate priors (more on that later)
- There are other options:
  - Collapsed variational inference
  - Parallelized inference for large corpora
  - Particle filters for on-line inference
- An ICML paper examining these issues is Asuncion et al. (2009).

Jonathan Chang's R implementation

```
result <-
    lda.collapsed.gibbs.sampler(cora.documents,
                                K,  ## Num clusters
                                cora.vocab, ## vocabulary
                                100,  ## num iterations
                                0.1, ## topic dirichlet
                                0.1) ## prop dirichlet
```

See http://www.pleasescoopme.com/

Jonathan Chang's R implementation

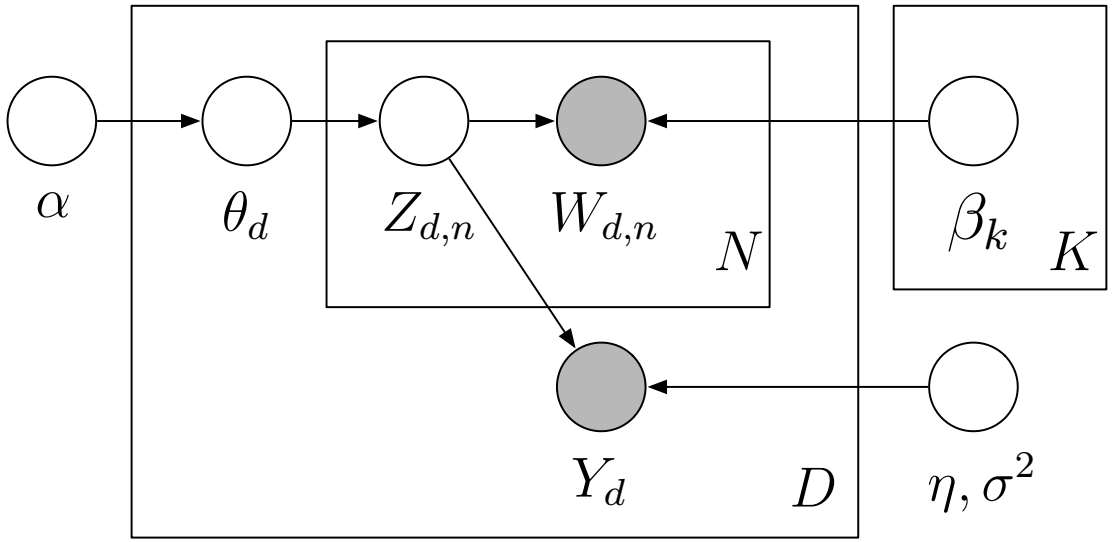


Supervised and relational topic models

Supervised topic models

- But LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with **response variables**.
  - User reviews paired with a number of stars
  - Web pages paired with a number of "diggs"
  - Documents paired with links to other documents
  - Images paired with a category
- **Supervised topic models** are topic models of documents and responses, fit to find topics predictive of the response.

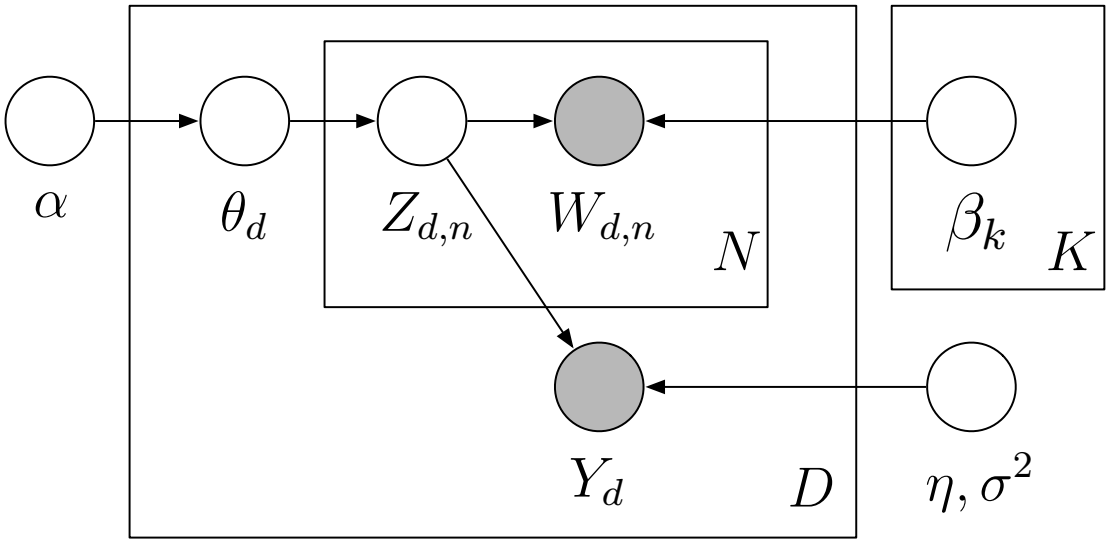
Supervised LDA



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

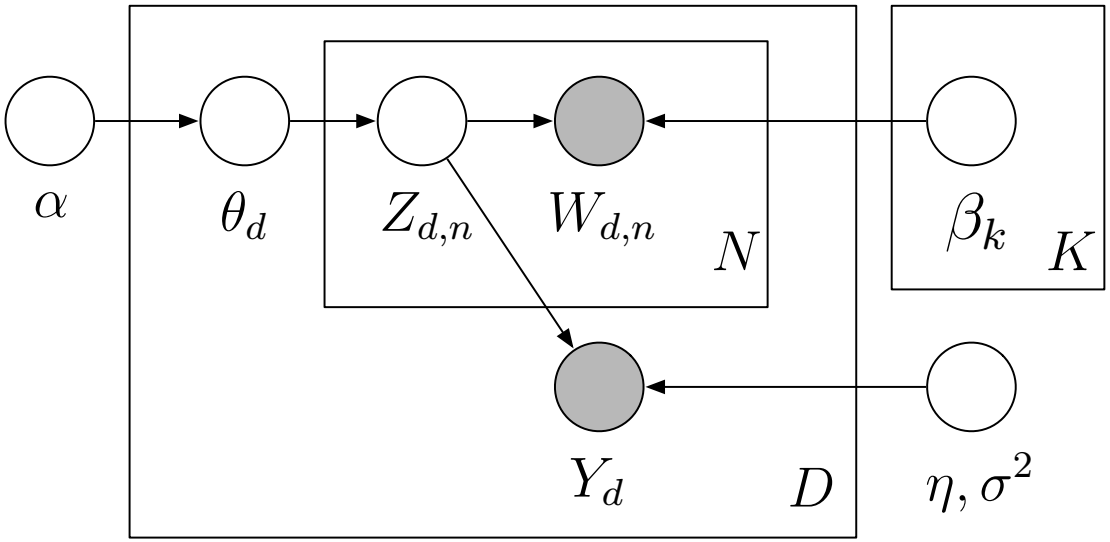
$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Supervised LDA



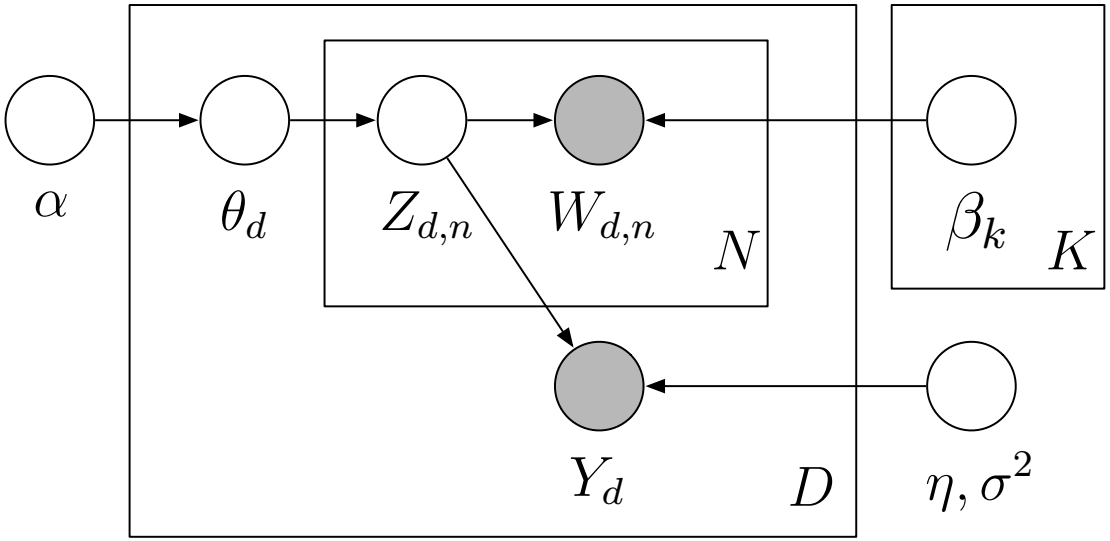
- The response variable y is drawn *after* the document because it depends on $z_{1:N}$, an assumption of **partial exchangeability**.
- Consequently, y is necessarily conditioned on the words.
- In a sense, this blends generative and discriminative modeling.

Supervised LDA



- Given a set of document-response pairs, fit the model parameters by **maximum likelihood**.
- Given a new document, compute a **prediction** of its response.
- Both of these activities hinge on **variational inference**.

Variational inference in sLDA



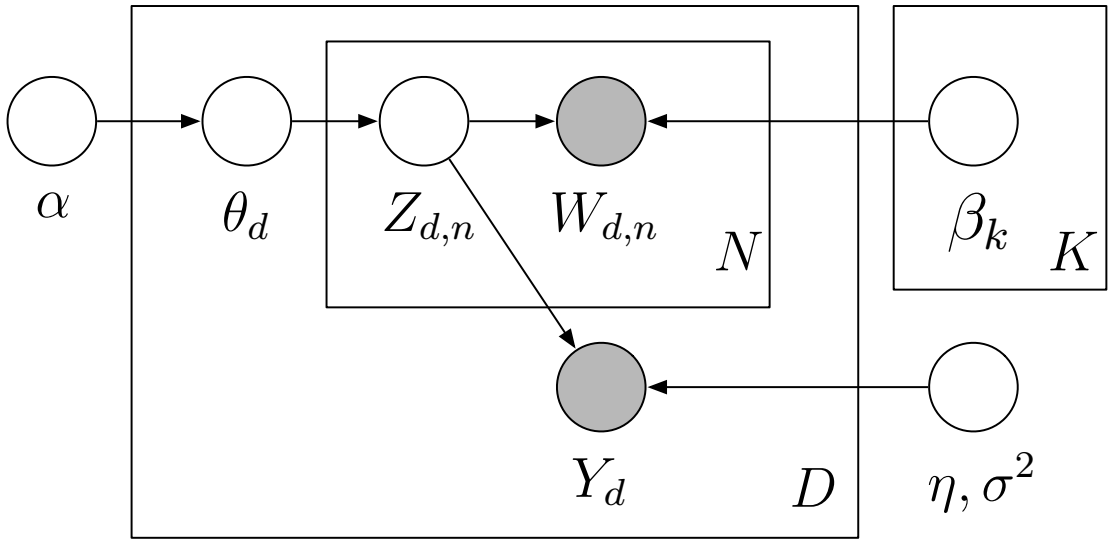
- Our goal is to compute the posterior distribution

$$p(\theta, z_{1:N} \mid w_{1:N}) = \frac{p(\theta, z_{1:N}, w_{1:N})}{\sum_{z_{1:N}} \int_\theta p(\theta, z_{1:N}, w_{1:N})}$$

- We approximate by minimizing the KL divergence to a simpler family of distributions,

$$q^*_\nu = \arg\min_{q \in \mathcal{Q}} \mathrm{K}L(q||p)$$

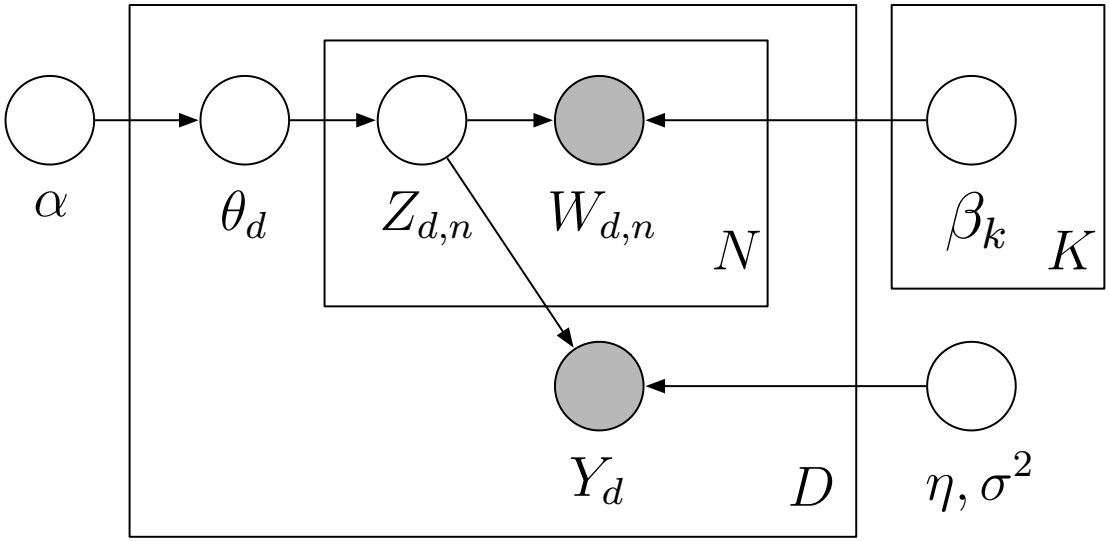
Variational inference in sLDA



Equivalently, maximize the Jensen's bound

$$\log p(w_{1:N}, y) \geq$$
$$\mathrm{E}[\log p(\theta \mid \alpha)] + \sum_{n=1}^{N} \mathrm{E}[\log p(Z_n \mid \theta)] + \sum_{n=1}^{N} \mathrm{E}[\log p(w_n \mid Z_n, \beta_{1:K})]$$
$$+\mathrm{E}[\log p(y \mid Z_{1:N}, \eta, \sigma^2)] + \mathrm{H}(q)$$

Variational inference in sLDA



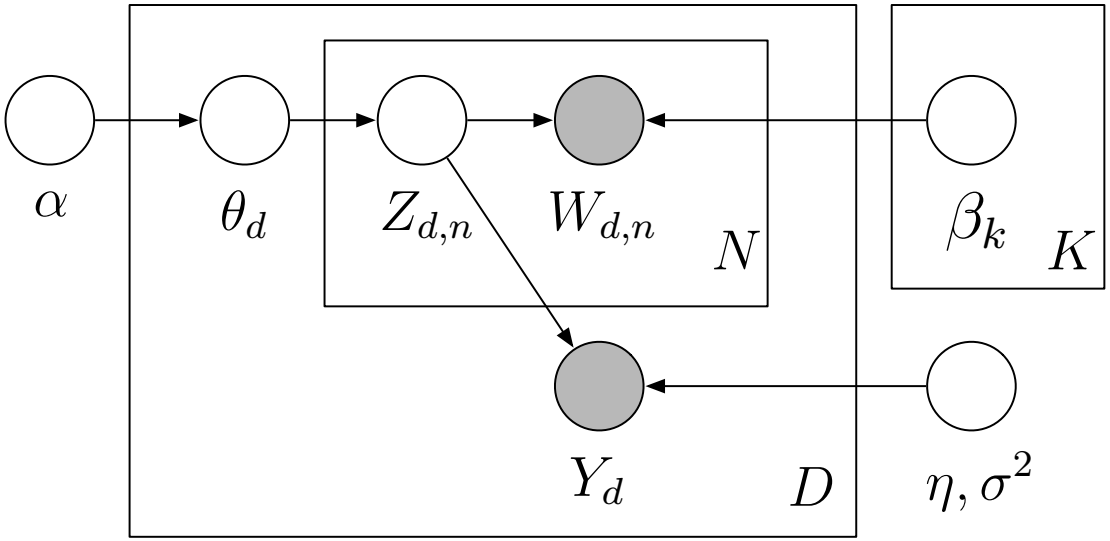
The distinguishing term is

$$\mathrm{E}[\log p(y \mid Z_{1:N})] = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{y^2 - 2y\eta^\top \mathrm{E}\left[\bar{Z}\right] + \eta^\top \mathrm{E}\left[\bar{Z}\bar{Z}^\top\right]\eta}{2\sigma^2}$$

We use the fully-factorized variational distribution

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

Variational inference in sLDA



- The expectations are

$$\mathrm{E}\left[\bar{Z}\right] = \bar{\phi} := \frac{1}{N}\sum_{n=1}^{N}\phi_n$$

$$\mathrm{E}\left[\bar{Z}\bar{Z}^\top\right] = \frac{1}{N^2}\left(\sum_{n=1}^{N}\sum_{m\neq n}\phi_n\phi_m^\top + \sum_{n=1}^{N}\mathrm{diag}\{\phi_n\}\right).$$

- Leads to an easy coordinate ascent algorithm.

Maximum likelihood estimation

- The M-step is an MLE under expected sufficient statistics.
- Define
  - $y = y_{1:D}$ is the response vector
  - A is the $D \times K$ matrix whose rows are $\bar{Z}_d^\top$.
- MLE of the coefficients solve the expected normal equations

$$\mathrm{E}\left[A^\top A\right]\eta = \mathrm{E}[A]^\top y \qquad \Rightarrow \qquad \hat{\eta}_{\text{new}} \leftarrow \left(\mathrm{E}\left[A^\top A\right]\right)^{-1} \mathrm{E}[A]^\top y$$

- The MLE of the variance is

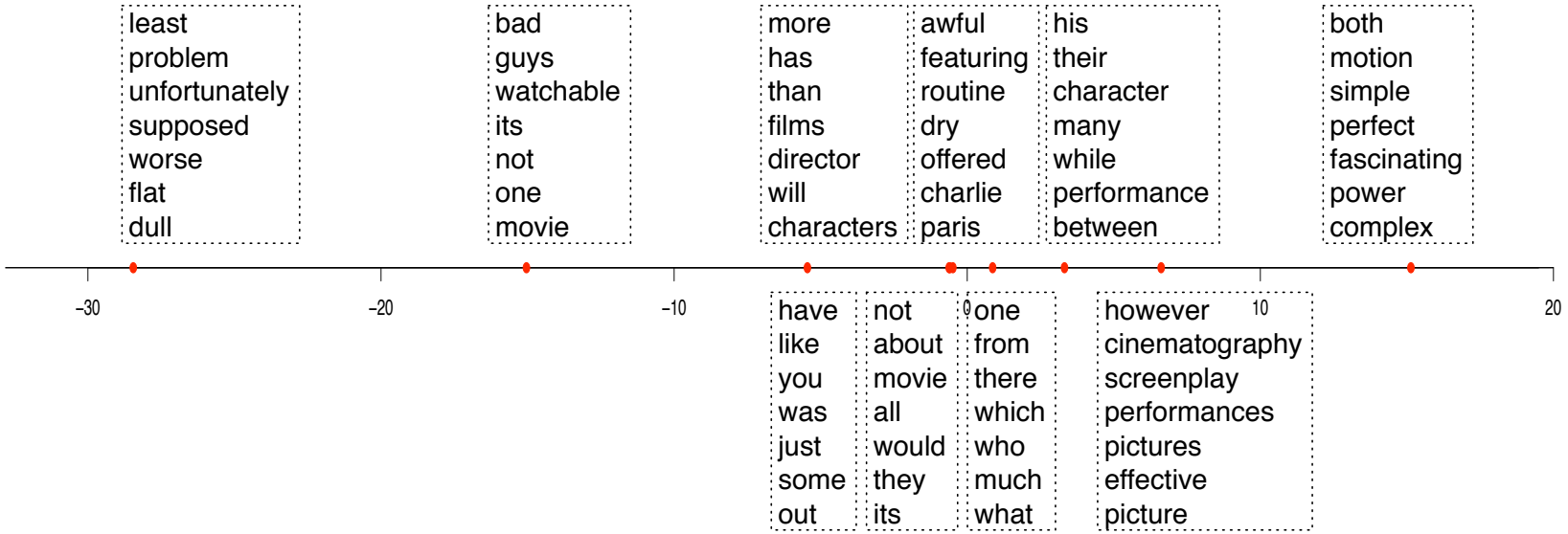
$$\hat{\sigma}^2_{\text{new}} \leftarrow (1/D)\{y^\top y - y^\top \mathrm{E}[A] \left(\mathrm{E}\left[A^\top A\right]\right)^{-1} \mathrm{E}[A]^\top y\}$$

Prediction

- We have fit SLDA parameters to a corpus, using variational EM.
- We have a new document $w_{1:N}$ with unknown response value.
- First, run variational inference in the unsupervised LDA model, to obtain γ and $\phi_{1:N}$ for the new document.
  (LDA $\Leftrightarrow$ integrating unobserved Y out of SLDA.)
- Predict y using SLDA expected value:

$$\mathrm{E}\left[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2\right] \approx \eta^\top \mathrm{E}_q\left[\bar{Z}\right] = \eta^\top \bar{\phi}.$$

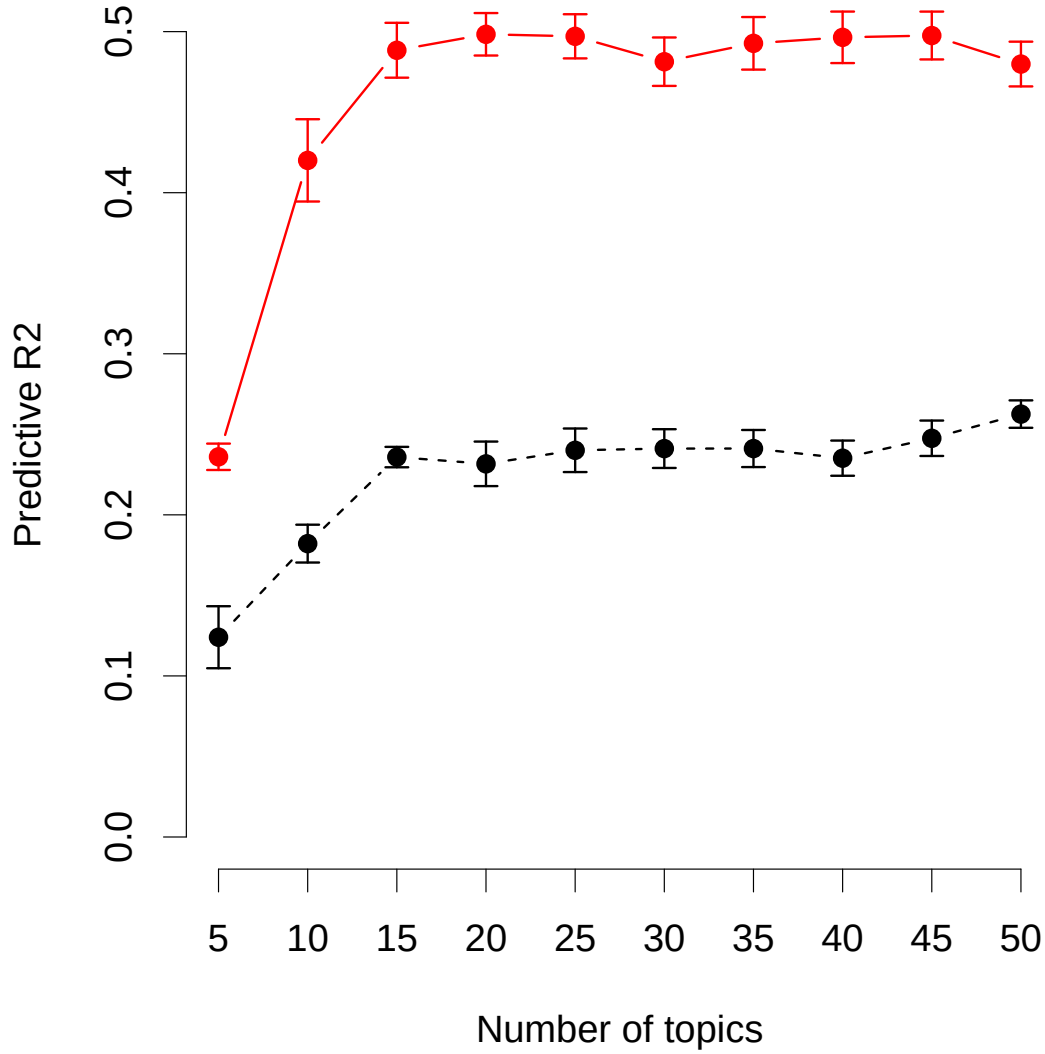
Example: Movie reviews



- 10-topic sLDA model on movie reviews (Pang and Lee, 2005).
- Response: number of stars associated with each review
- Each component of coefficient vector η is associated with a topic.

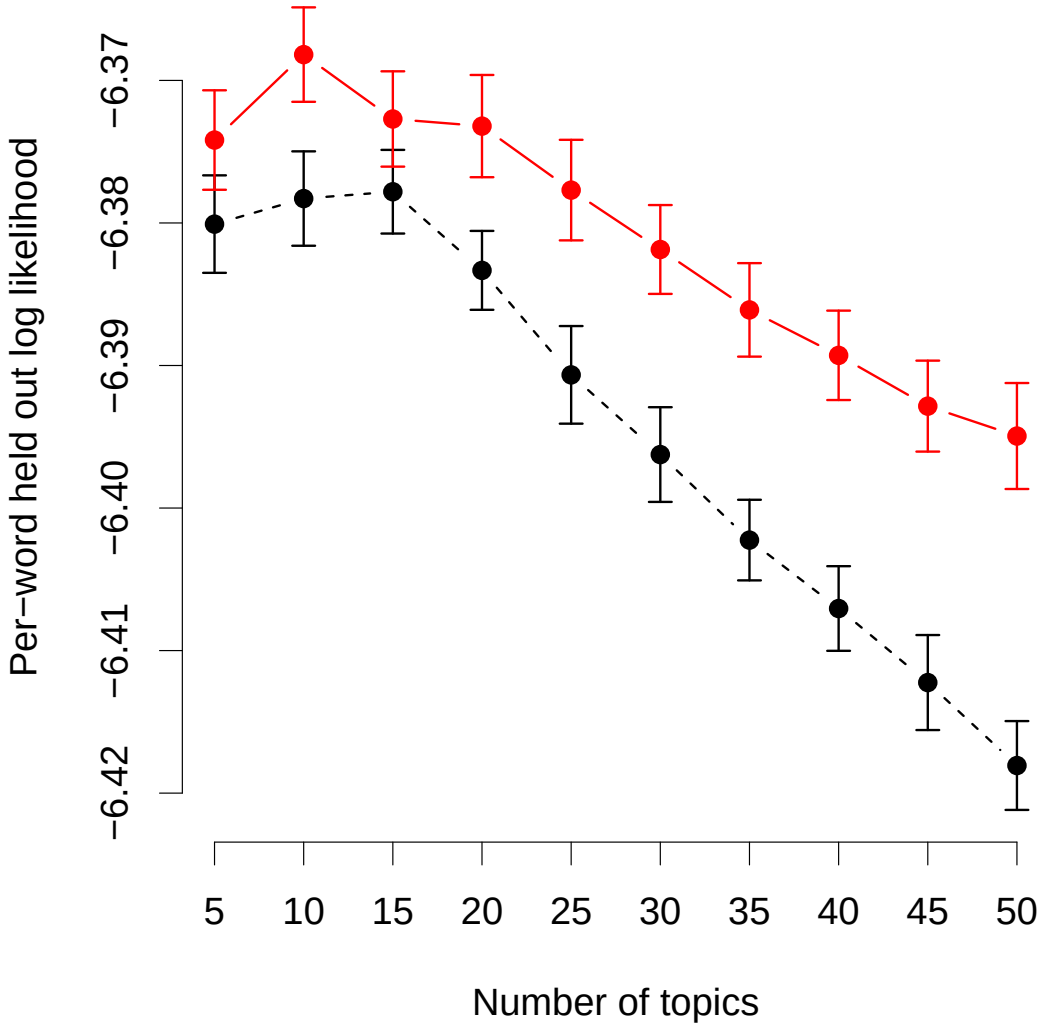
Predictive R2

(SLDA is red.)



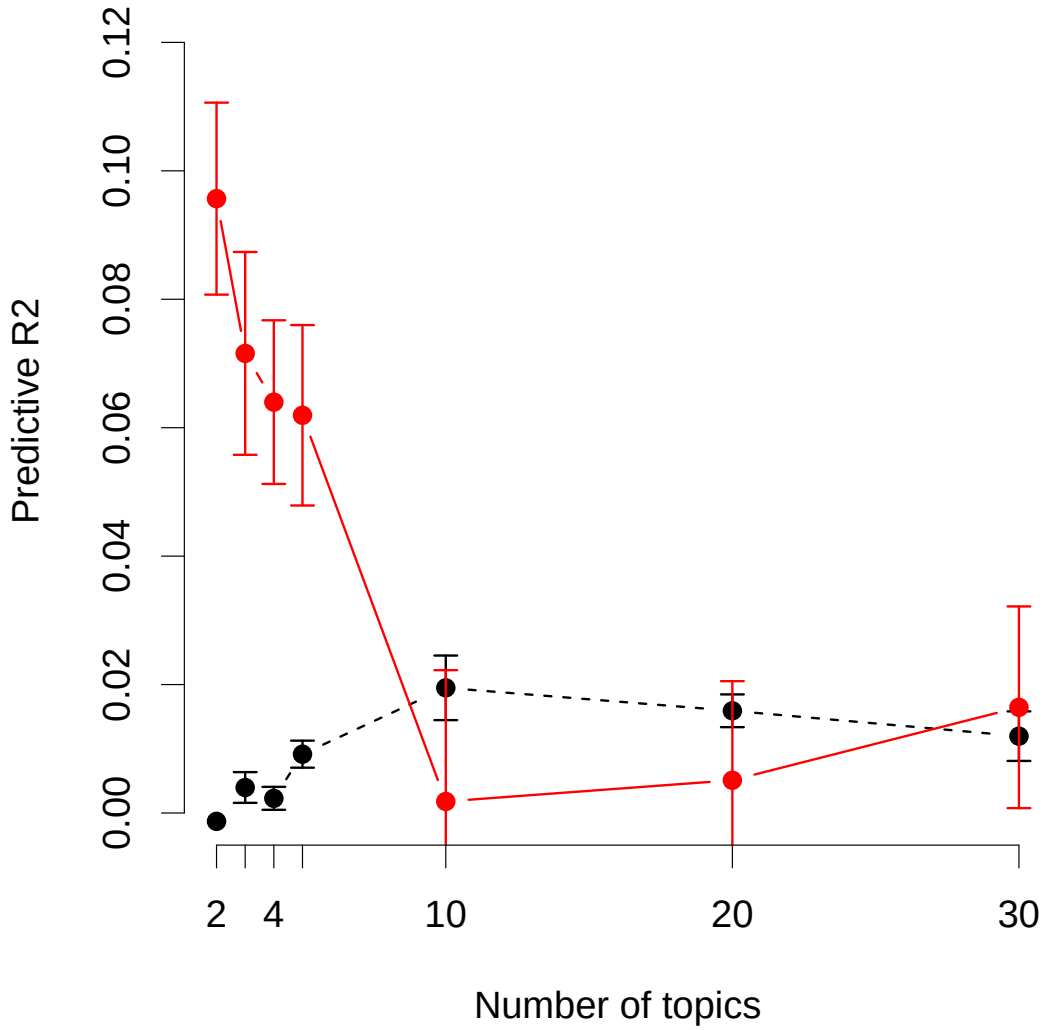
Held out likelihood

(SLDA is red.)



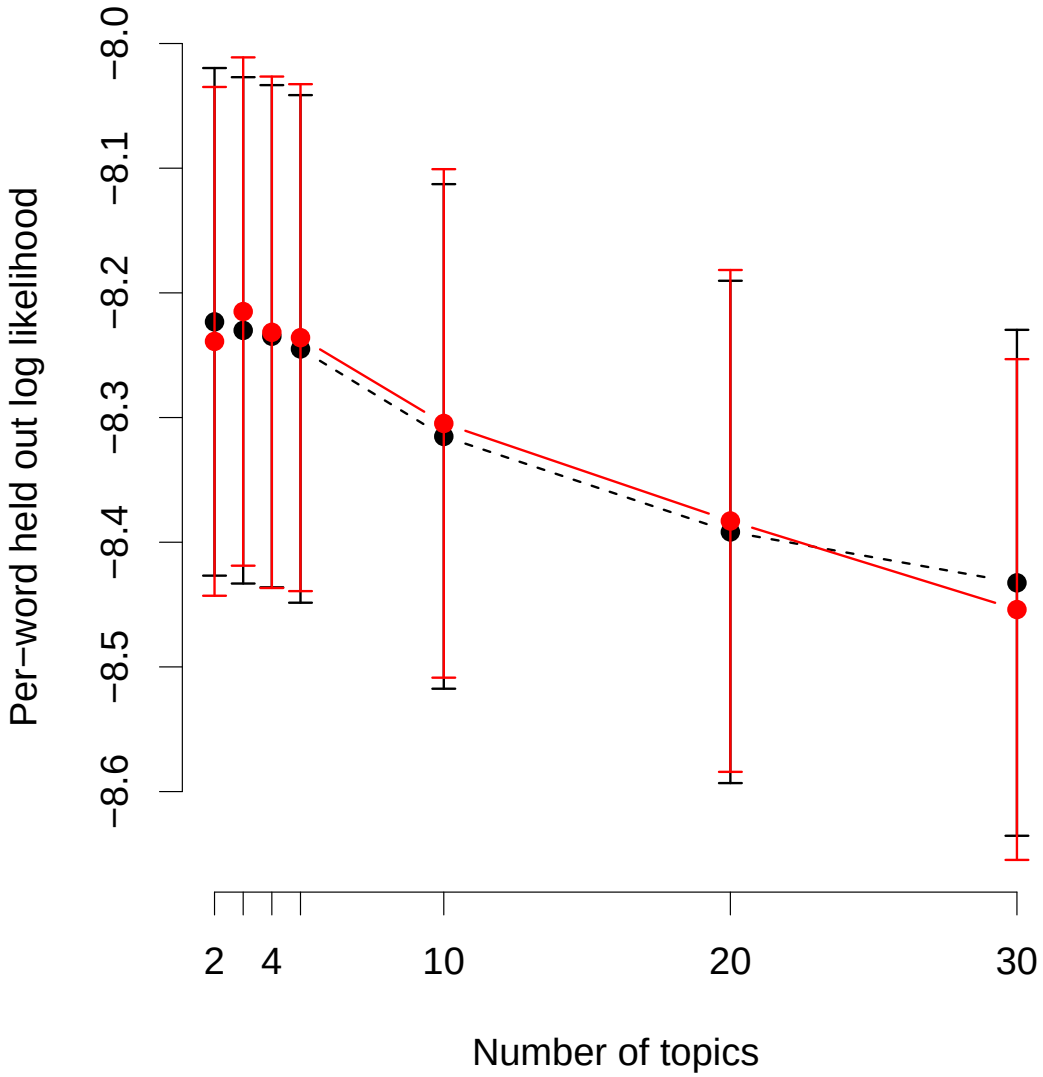
Predictive R2 on Digg

(SLDA is red.)



Held out likelihood on Digg

(SLDA is red.)



Diverse response types with GLMs

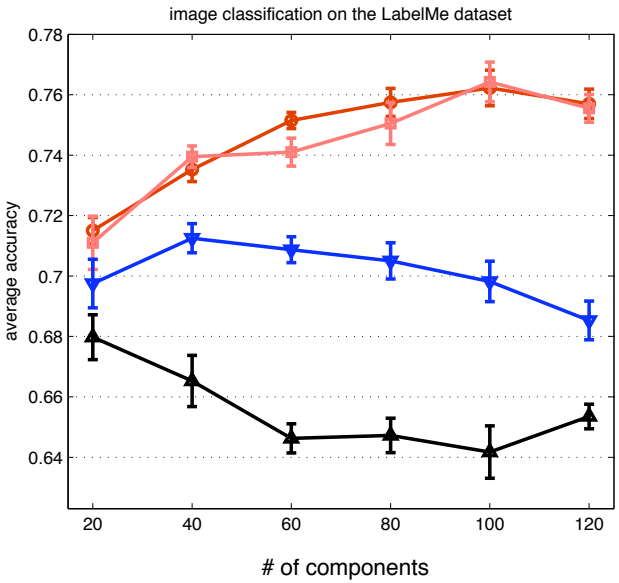
- Want to work with response variables that don't live in the reals.
  - binary / multiclass classification
  - count data
  - waiting time
- Model the response response with a generalized linear model

$$p(y \mid \zeta, \delta) = h(y, \delta) \exp\left\{ \frac{\zeta y - A(\zeta)}{\delta} \right\} ,$$

  where $\zeta = \eta^\top \bar{z}$.

- Complicates inference, but allows for flexible modeling.

Example: Multi-class classification



highway

car, sign, road

inside city

buildings, car, sidewalk

street

tree, car, sidewalk

tall building

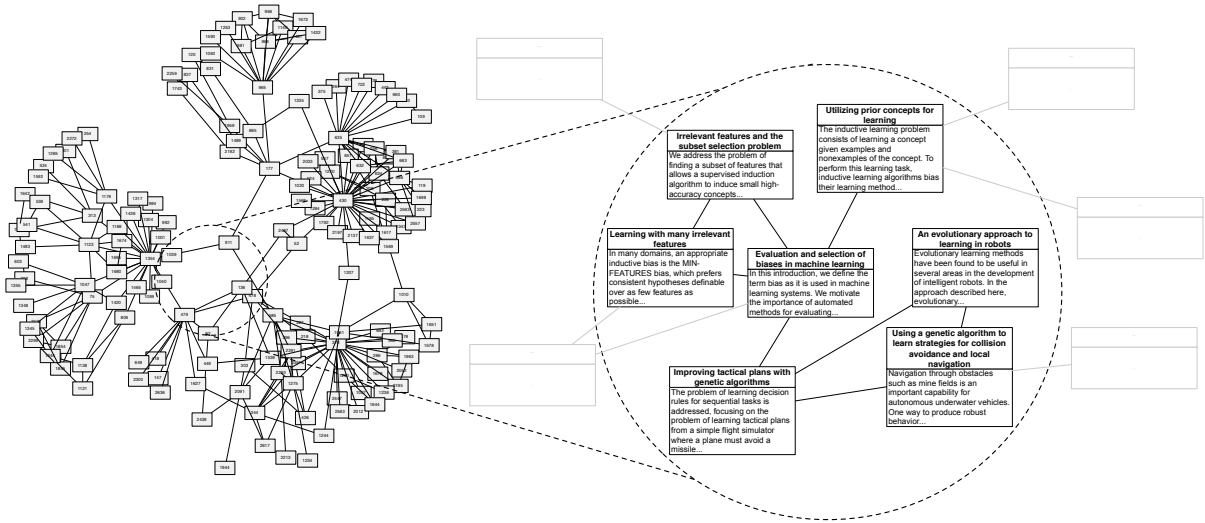
trees, buildings
occluded, window

SLDA for image classification (with Chong Wang, CVPR 2009)

Supervised topic models

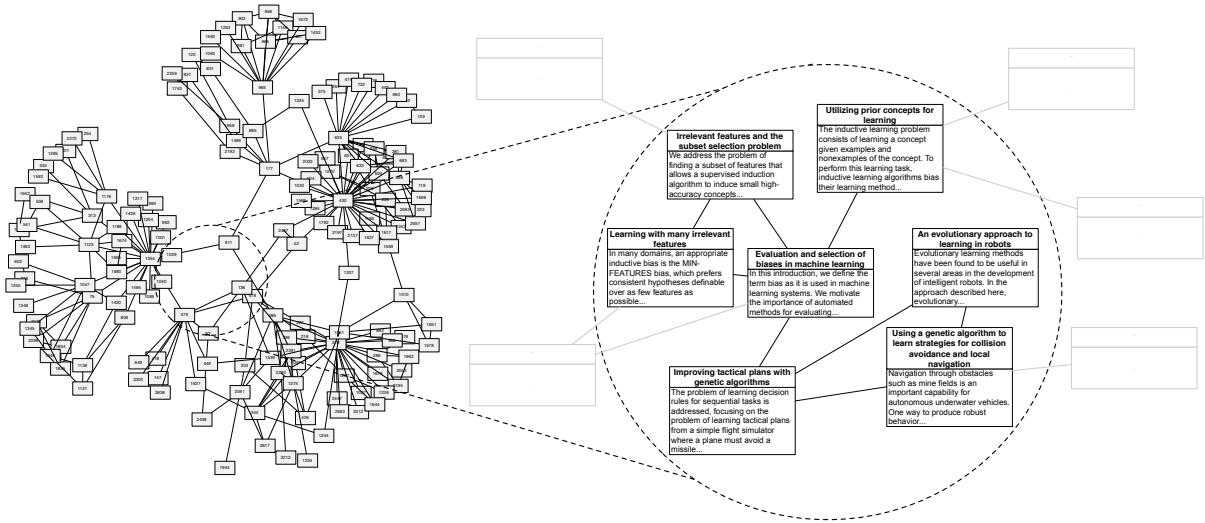
- SLDA enables model-based regression where the predictor "variable" is a text document.
- It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).
- SLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.
- LDA + regression compared to sLDA is like principal components regression compared to partial least squares.

Relational topic models



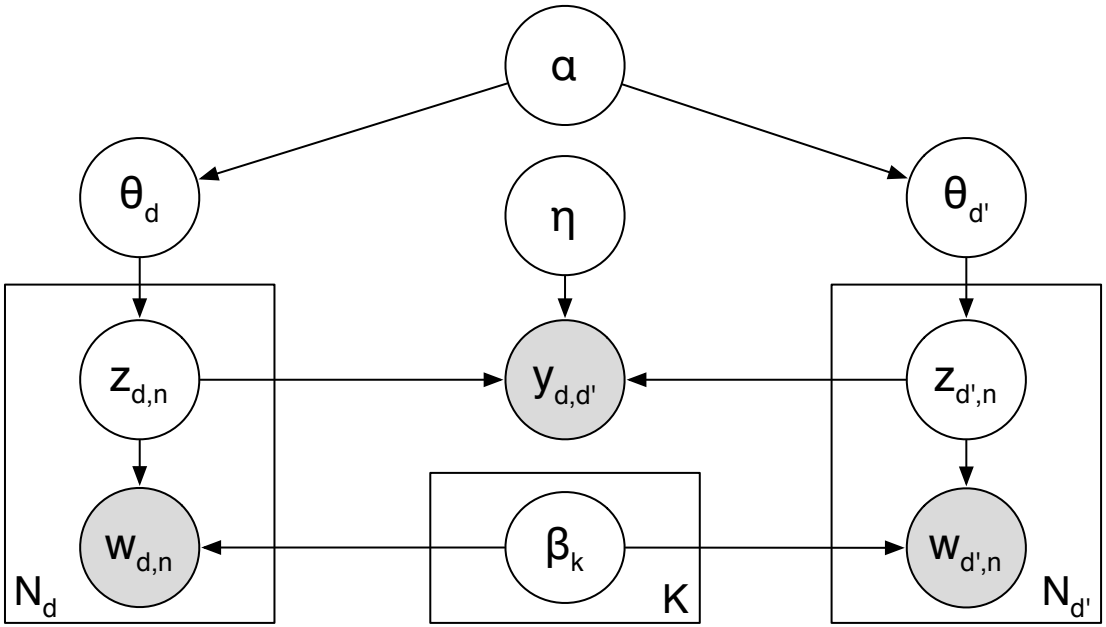
- Many data sets contain **connected observations**.
- For example:
  - Citation networks of documents
  - Hyperlinked networks of web-pages.
  - Friend-connected social network profiles

Relational topic models



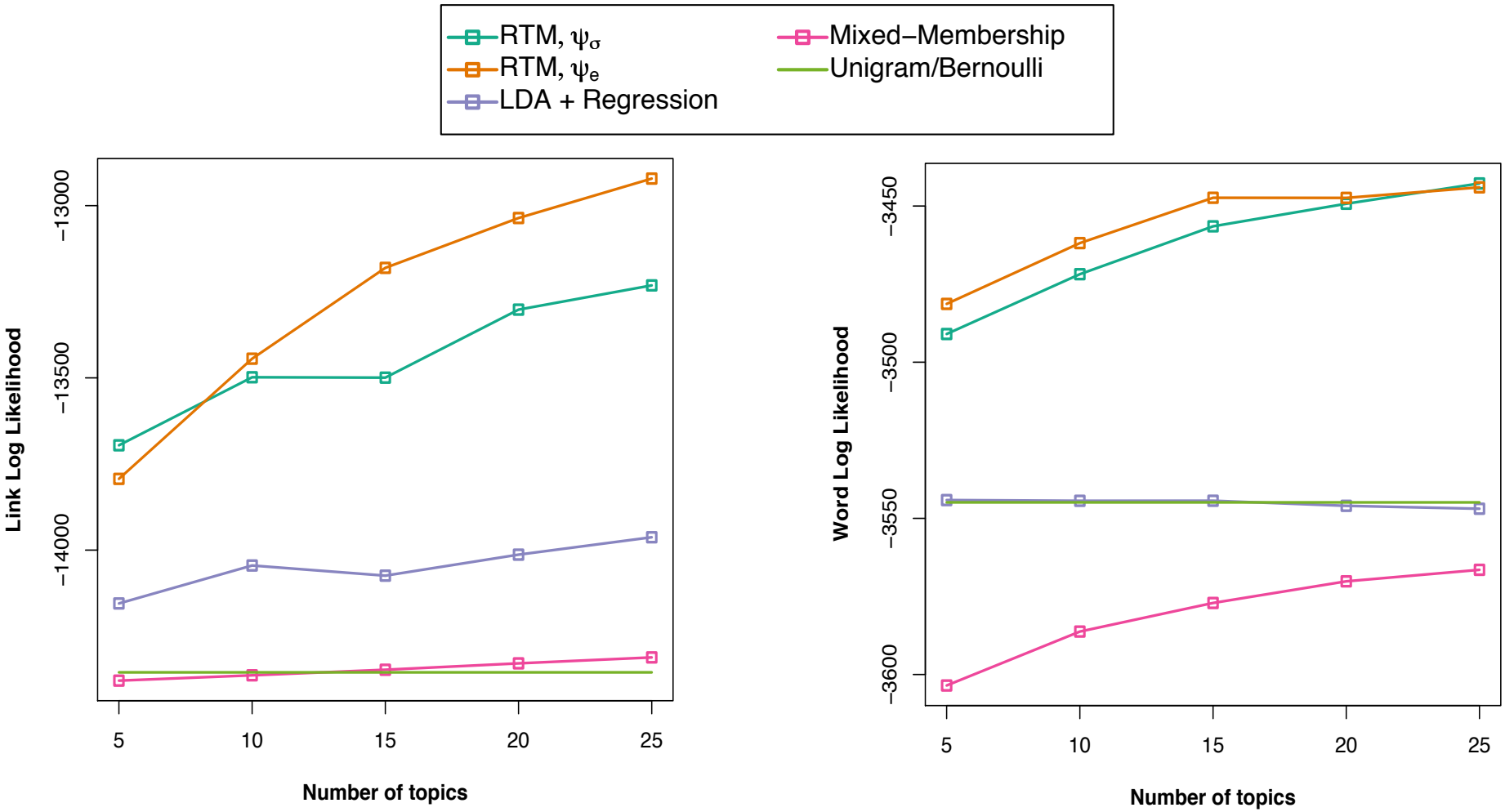
- Research has focused on finding communities and patterns in the link-structure of these networks (Kemp et al. 2004, Hoff et al., 2002, Hofman and Wiggins 2007, Airoldi et al. 2008).
- By adapting supervised topic modeling, we can build a good model of **content and structure**.
- RTMs find related hidden structure in both types of data.

Relational topic models



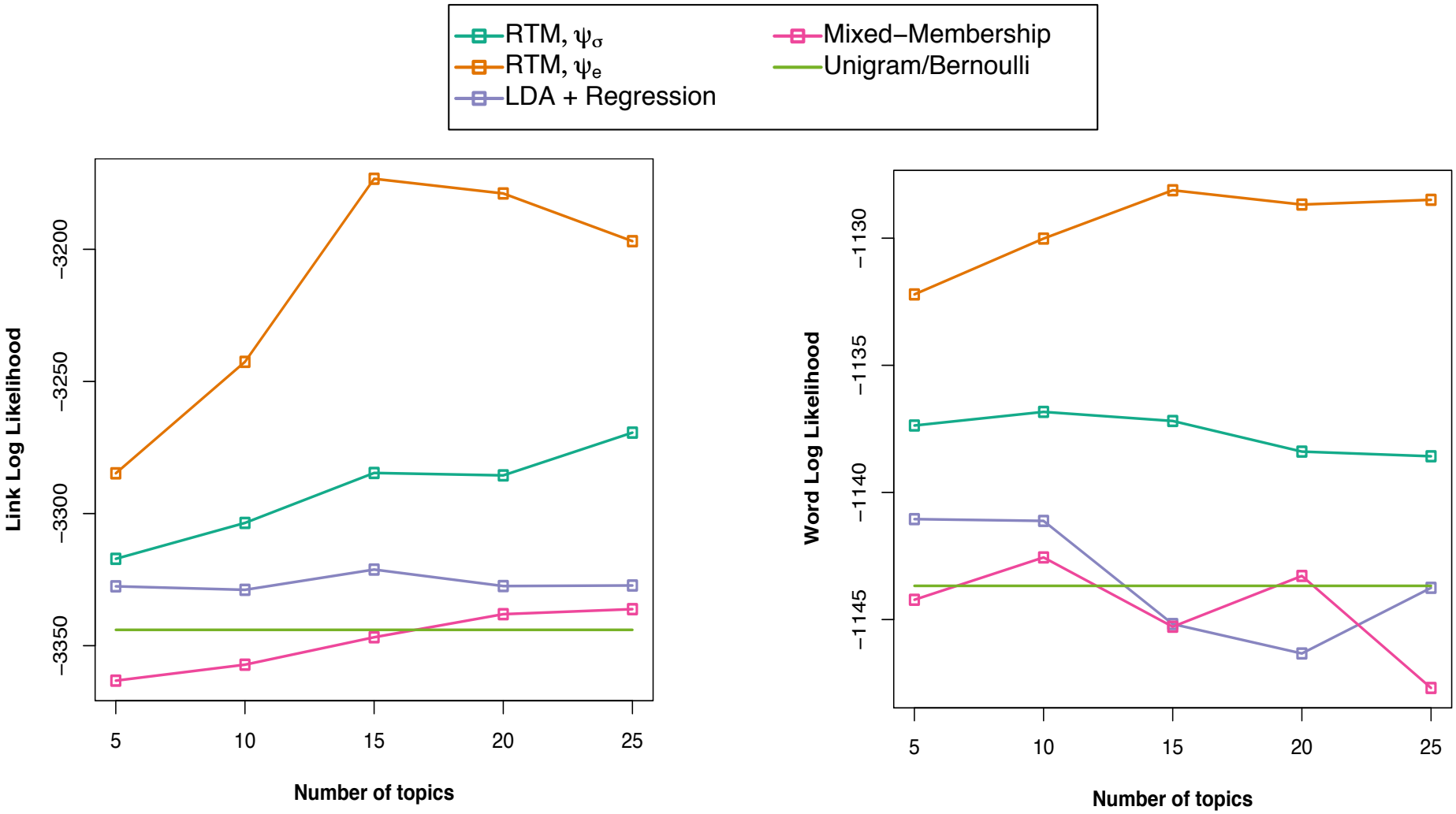
- Binary response variable with each pair of documents
- Adapt variational EM algorithm for sLDA with binary GLM response model (with different link probability functions).
- Allows predictions that are out of reach for traditional models.

Predictive performance of one type given the other



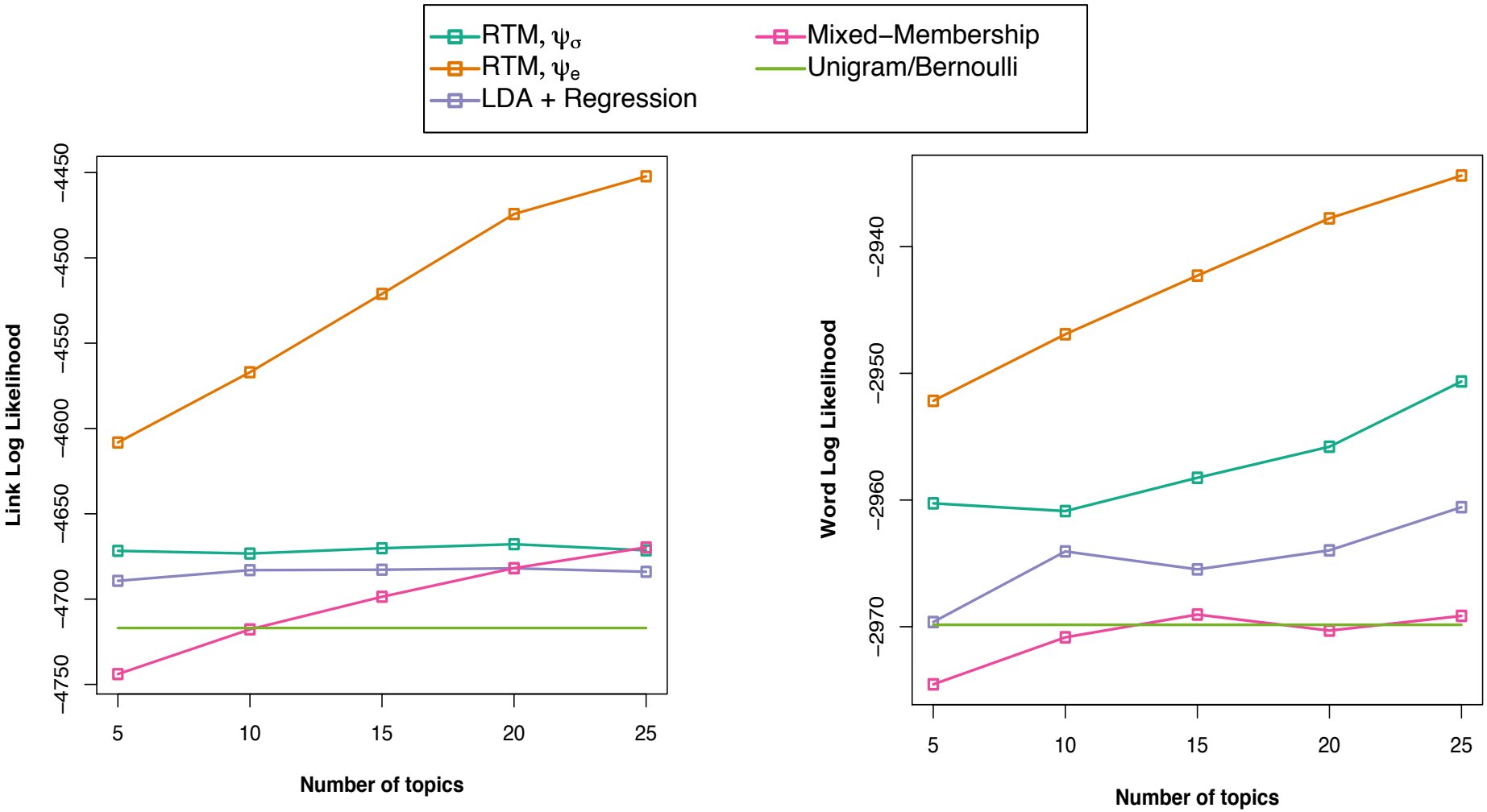
Cora corpus (McCallum et al., 2000)

Predictive performance of one type given the other



WebKB corpus (Craven et al., 1998)

Predictive performance of one type given the other



PNAS corpus (courtesy of JSTOR)

Predicting links from documents

Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo Rates of convergence of the Hastings and Metropolis algorithms **Possible biases induced by MCMC convergence diagnostics** Bounding convergence time of the Gibbs sampler in Bayesian image restoration Self regenerative Markov chain Monte Carlo Auxiliary variable methods for Markov chain Monte Carlo with applications **Rate of Convergence of the Gibbs Sampler by Gaussian Approximation** Diagnosing convergence of Markov chain Monte Carlo algorithms	**RTM (ψ_e)**
Exact Bound for the Convergence of Metropolis Chains Self regenerative Markov chain Monte Carlo **Minorization conditions and convergence rates for Markov chain Monte Carlo** Gibbs-markov models Auxiliary variable methods for Markov chain Monte Carlo with applications Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models Mediating instrumental variables A qualitative framework for probabilistic inference Adaptation for Self Regenerative MCMC	**LDA + Regression**

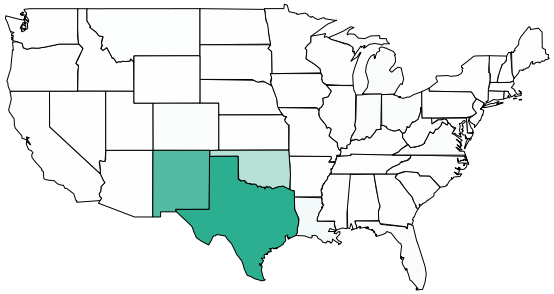
Given a new document, which documents is it likely to link to?

Predicting links from documents

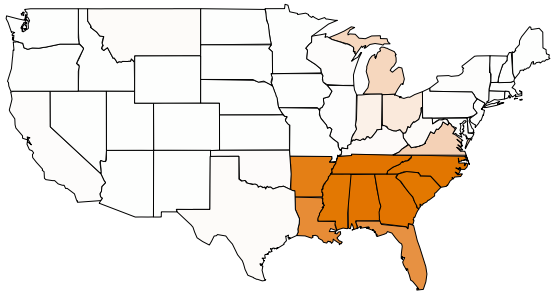
Competitive environments evolve better solutions for complex tasks	
Coevolving High Level Representations A Survey of Evolutionary Strategies **Genetic Algorithms in Search, Optimization and Machine Learning** **Strongly typed genetic programming in evolving cooperation strategies** Solving combinatorial problems using evolutionary algorithms A promising genetic algorithm approach to job-shop scheduling... Evolutionary Module Acquisition An Empirical Investigation of Multi-Parent Recombination Operators...	**RTM (ψ_e)**
A New Algorithm for DNA Sequence Assembly Identification of protein coding regions in genomic DNA Solving combinatorial problems using evolutionary algorithms A promising genetic algorithm approach to job-shop scheduling... A genetic algorithm for passive management The Performance of a Genetic Algorithm on a Chaotic Objective Function Adaptive global optimization with local search Mutation rates as adaptations	**LDA + Regression**

Given a new document, which documents is it likely to link to?

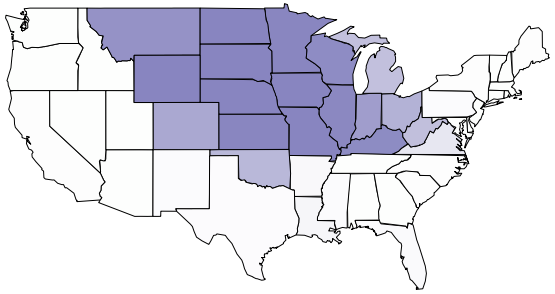
Spatially consistent topics



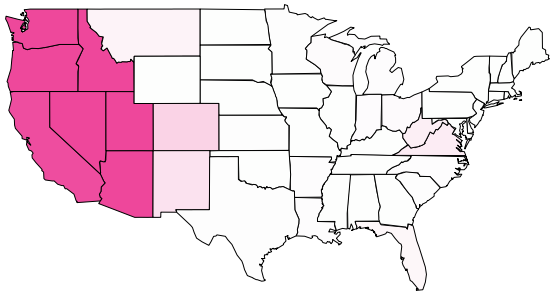
Topic 1



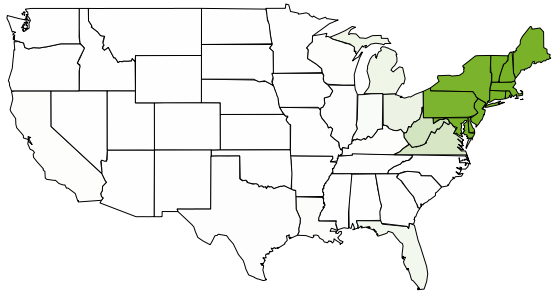
Topic 2



Topic 3



Topic 4



Topic 5

- For exploratory tasks, RTMs can be used to "guide" the topics
- Documents are geographically-tagged news articles from Yahoo! Links are the adjacency matrix of states
- RTM finds **spatially consistent** topics.

Relational Topic Models

- Relational topic modeling allows us to analyze connected documents, or other data for which the mixed-membership assumptions are appropriate.
- Traditional models cannot predict with new and unlinked data.
- RTMs allow for such predictions
  - links given the new words of a document
  - words given the links of a new document

Used in exploratory tools of document collections