Supervised and Relational Topic Models

David M. Blei

Department of Computer Science
Princeton University

October 5, 2009

Joint work with Jonathan Chang and Jon McAuliffe

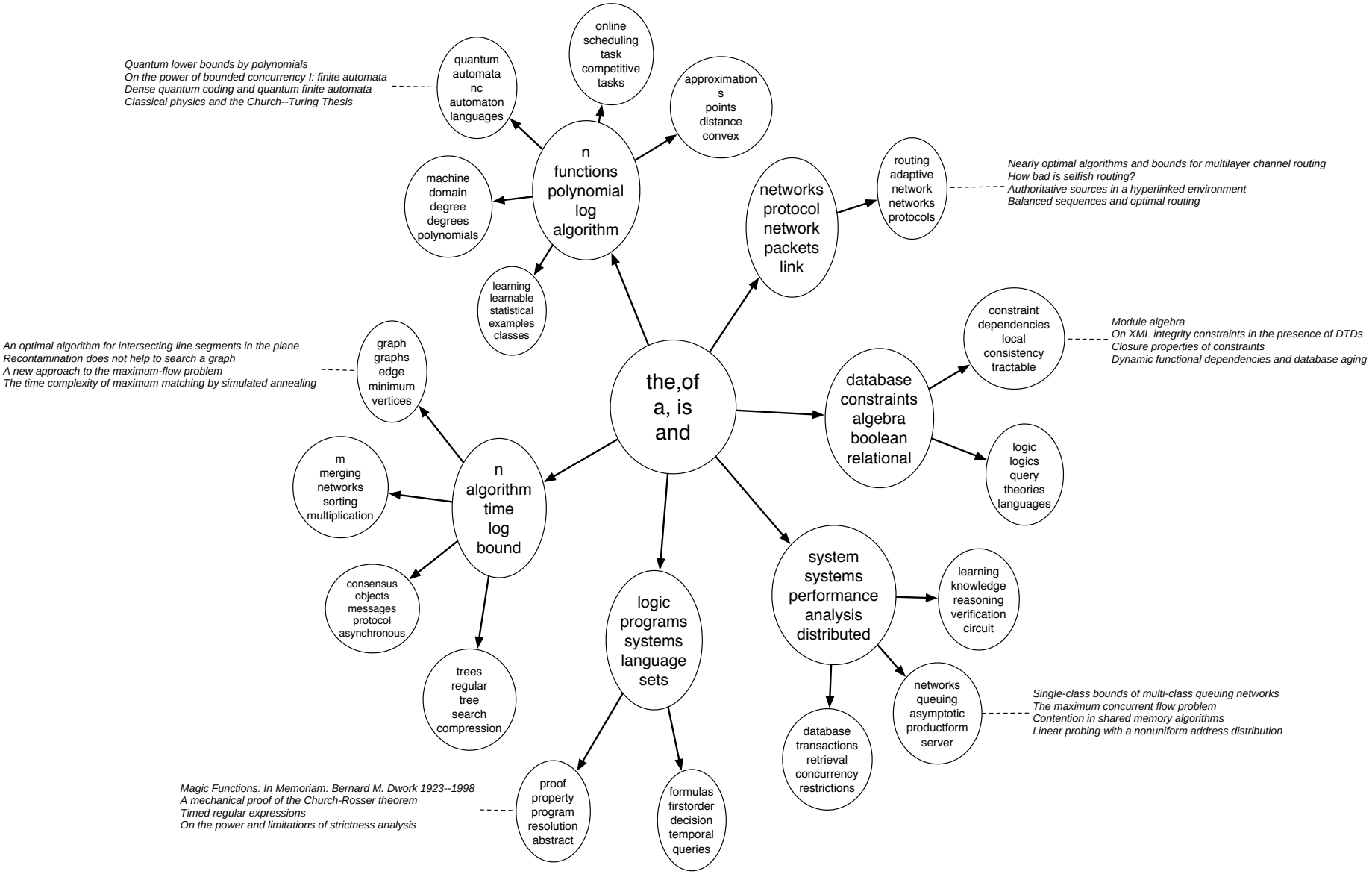
Topic modeling

- Large electronic archives of document collections require new statistical tools for analyzing text.
- **Topic models** have emerged as a powerful technique for unsupervised analysis of large document collections.
- Topic models posit **latent topics** in text using hidden random variables, and uncover that structure with **posterior inference**.
- Useful for tasks like browsing, search, information retrieval, etc.

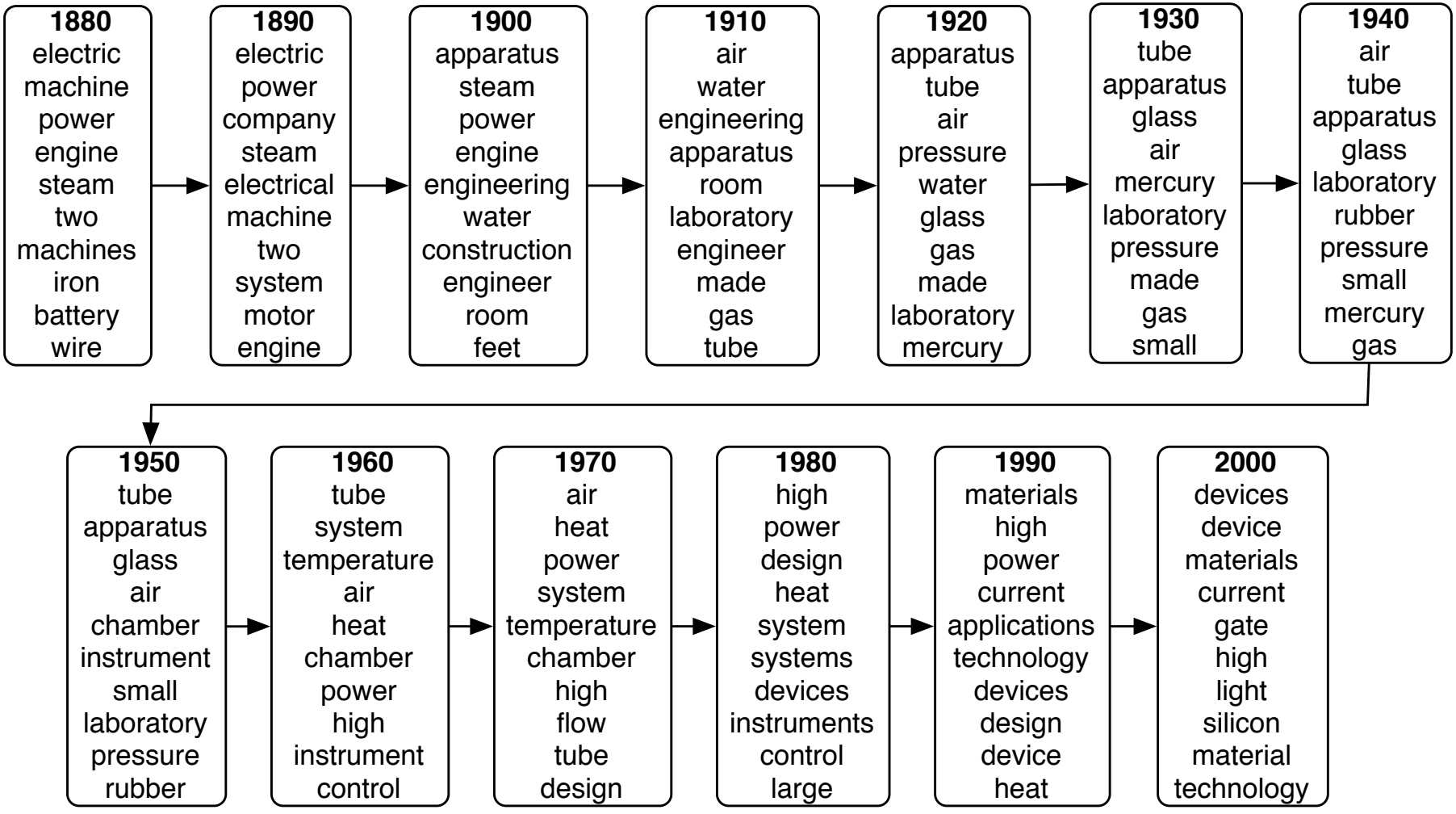
Examples of topic modeling

contractual	employment	female	markets	criminal
expectation	industrial	men	earnings	discretion
gain	local	women	investors	justice
promises	jobs	see	sec	civil
expectations	employees	sexual	research	process
breach	relations	note	structure	federal
enforcing	unfair	employer	managers	see
supra	agreement	discrimination	firm	officer
note	economic	harassment	risk	parole
perform	case	gender	large	inmates

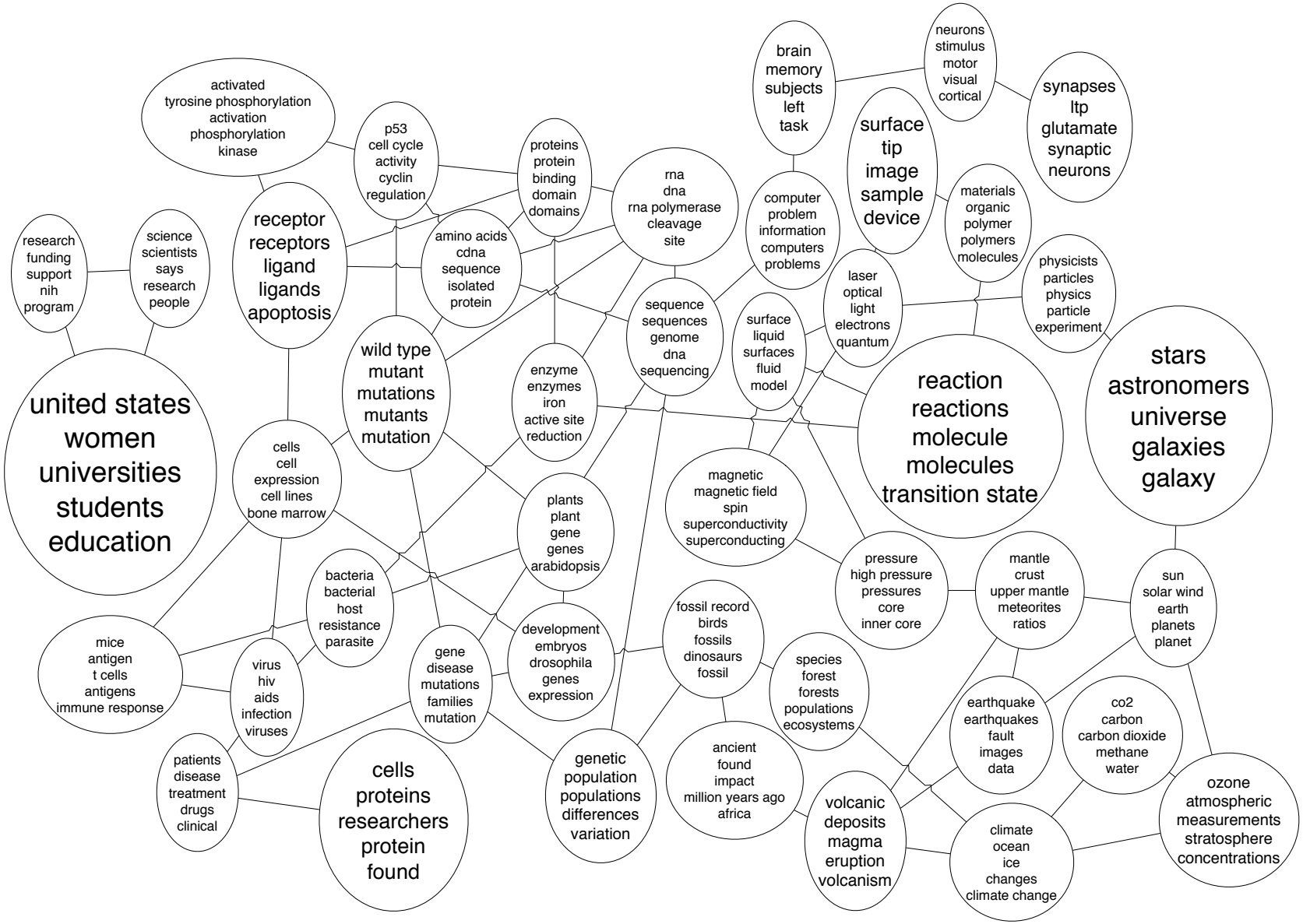
Examples of topic modeling



Examples of topic modeling



Examples of topic modeling



Supervised topic models

- These applications of topic modeling work in the same way.
- Fit a model using a likelihood criterion. Then, hope that the resulting model is useful for the task at hand.
- **Supervised topic models** and **relational topic models** fit topics explicitly to perform prediction.
- Useful for building topic models that can
  - Predict the rating of a review
  - Predict the category of an image
  - Predict the links emitted from a document

Outline

1. Unsupervised topic models
2. Supervised topic models
3. Relational topic models

Probabilistic modeling

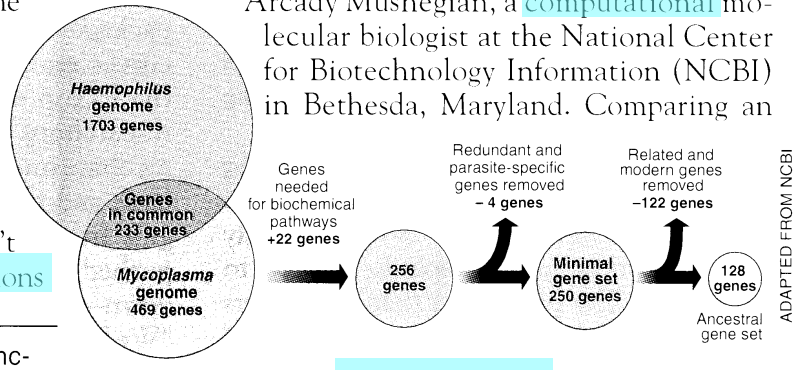
1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using *posterior inference*
   - What are the topics that describe this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



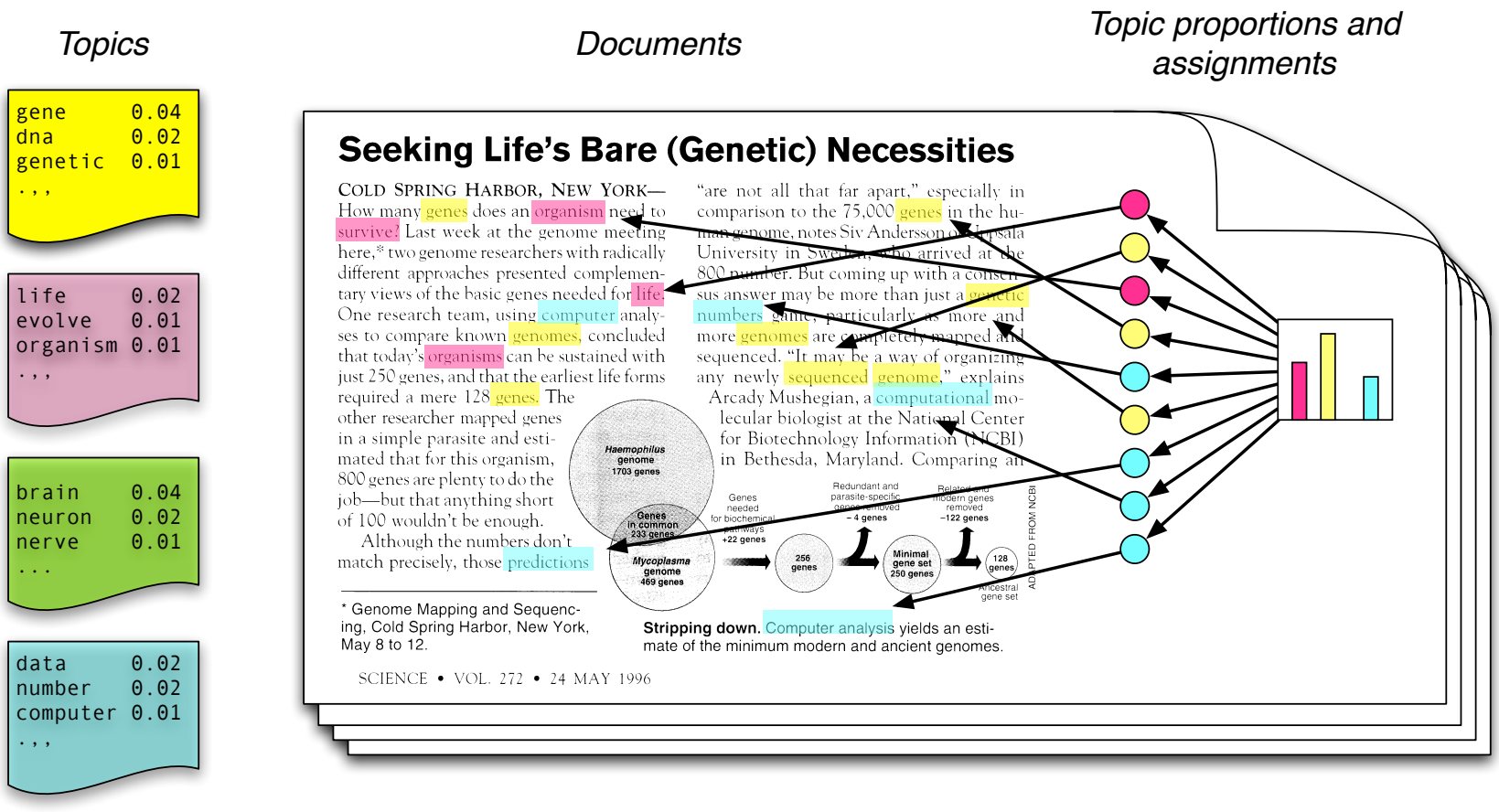
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

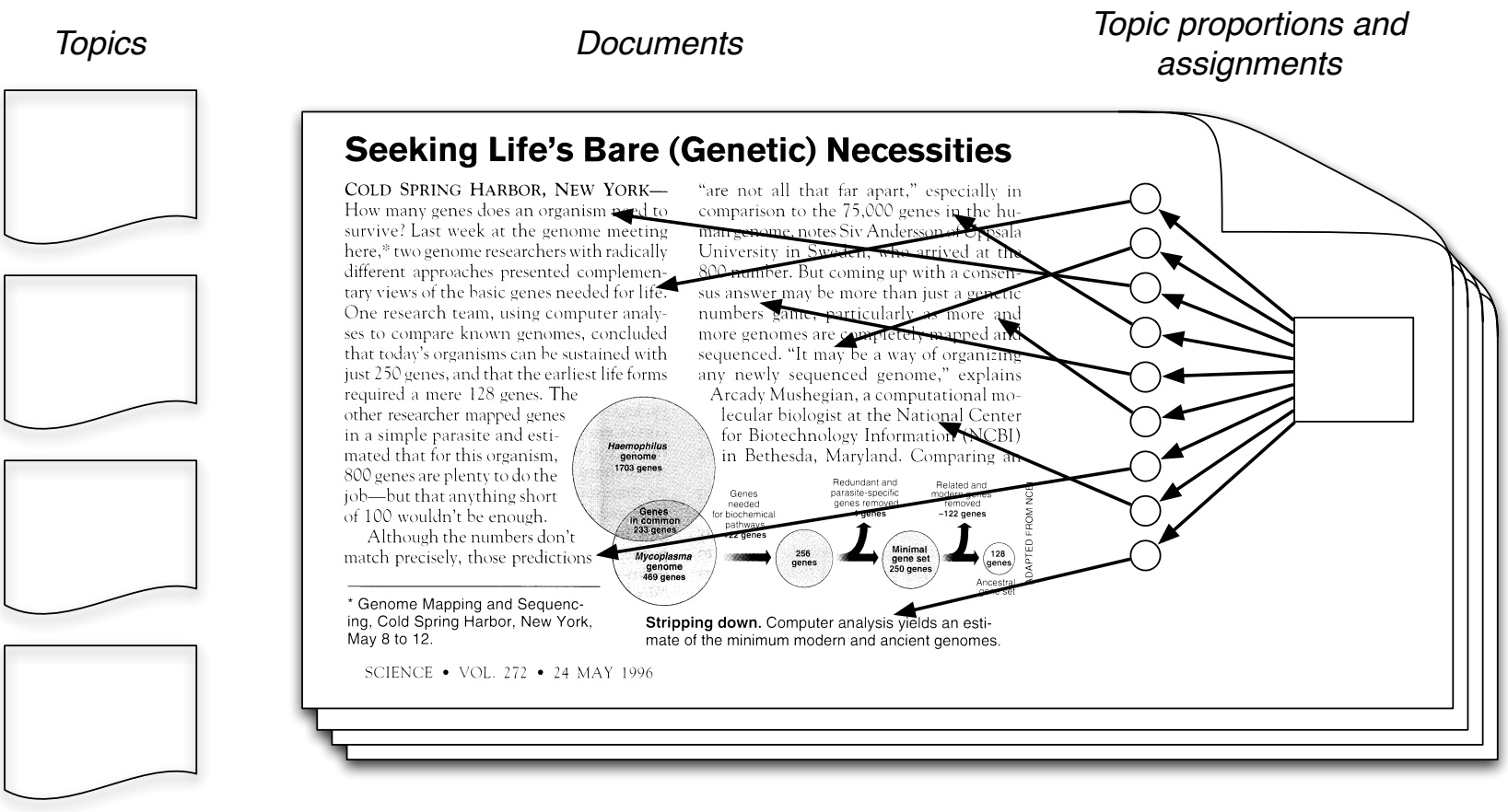
Simple intuition: Documents exhibit multiple topics.

Generative model



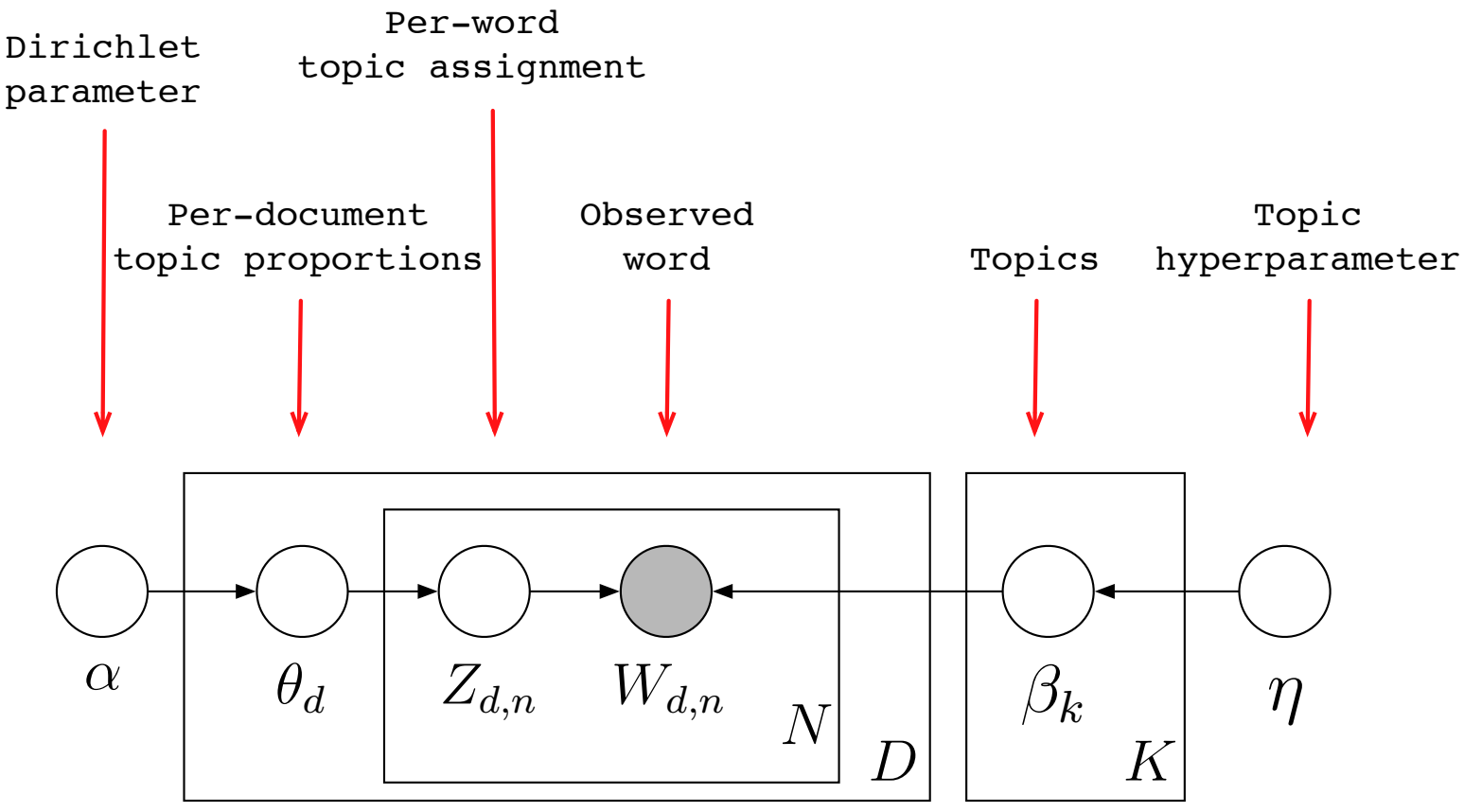
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

The posterior distribution



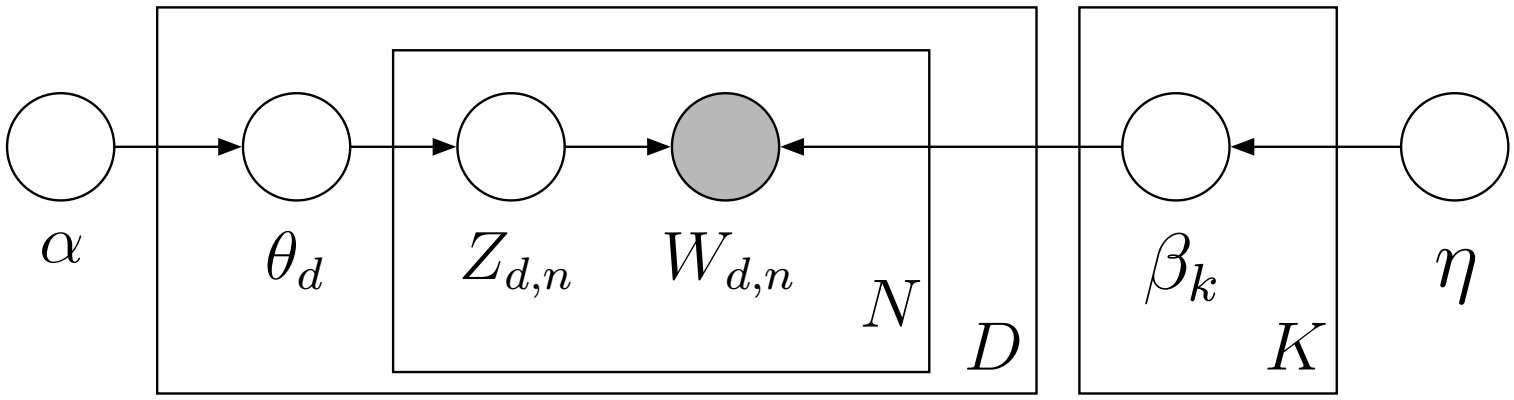
- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

Latent Dirichlet allocation



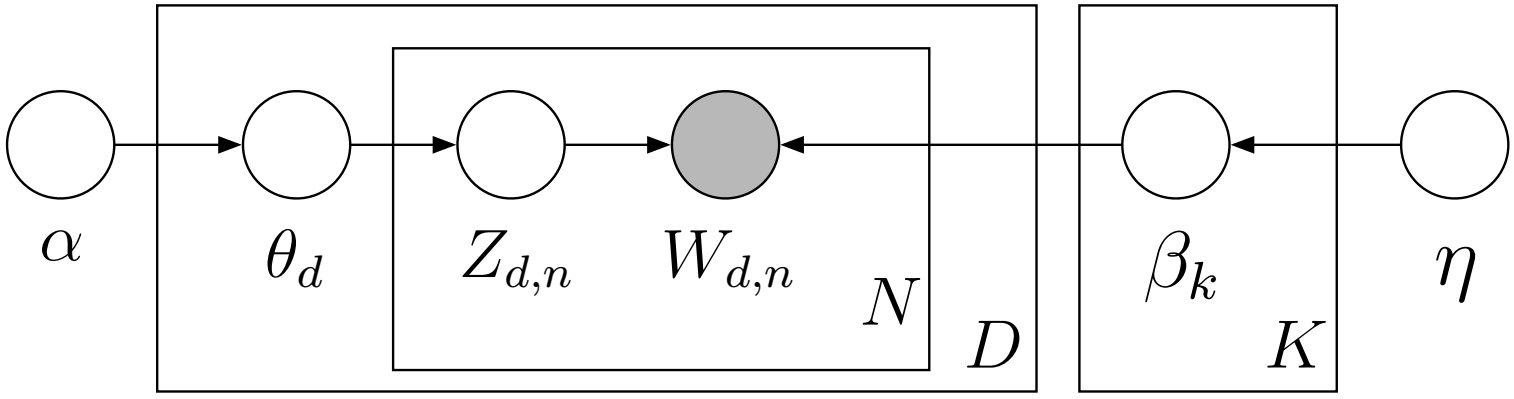
Each piece of the structure is a random variable.

Latent Dirichlet allocation



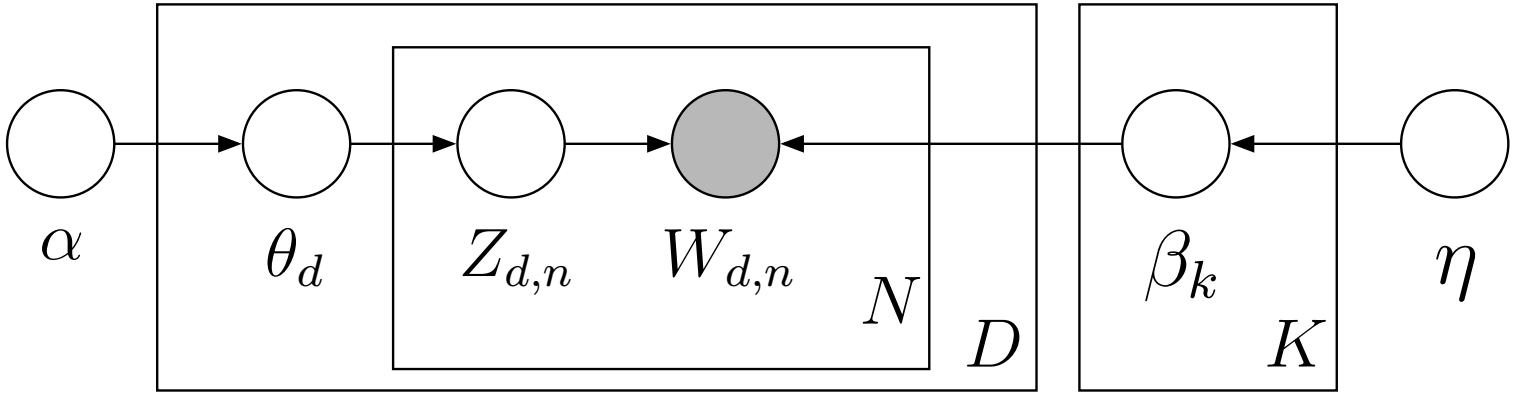
$$
\begin{aligned}
\beta_k &\sim \mathrm{Dir}(\eta) \quad k = 1 \ldots K \\
\theta_d &\sim \mathrm{Dir}(\alpha) \quad d = 1 \ldots D \\
Z_{d,n} \mid \theta_d &\sim \mathrm{Mult}(1, \theta_d) \quad d = 1 \ldots D, \quad n = 1 \ldots N \\
W_{d,n} \mid \theta_d, z_{d,n}, \beta_{1:K} &\sim \mathrm{Mult}(1, \beta_{z_{d,n}}) \quad d = 1 \ldots D, \quad n = 1 \ldots N
\end{aligned}
$$

Latent Dirichlet allocation



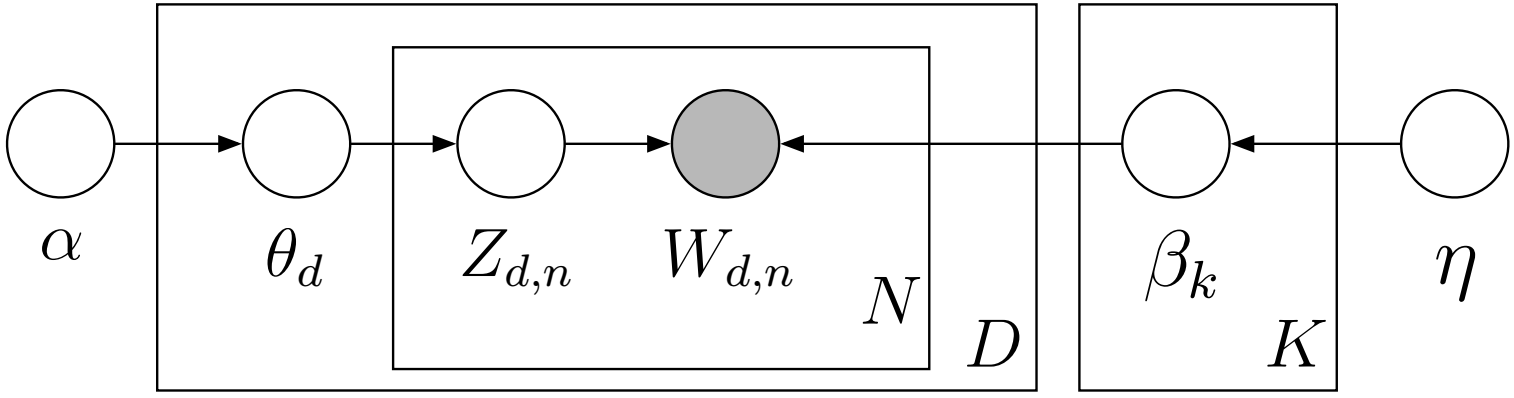
1. Draw each topic $\beta_k \sim \mathrm{Dir}(\eta)$, for $k \in \{1, \ldots, K\}$.
2. For each document:
   1. Draw topic proportions $\theta_d \sim \mathrm{Dir}(\alpha)$.
   2. For each word:
      1. Draw $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$.

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

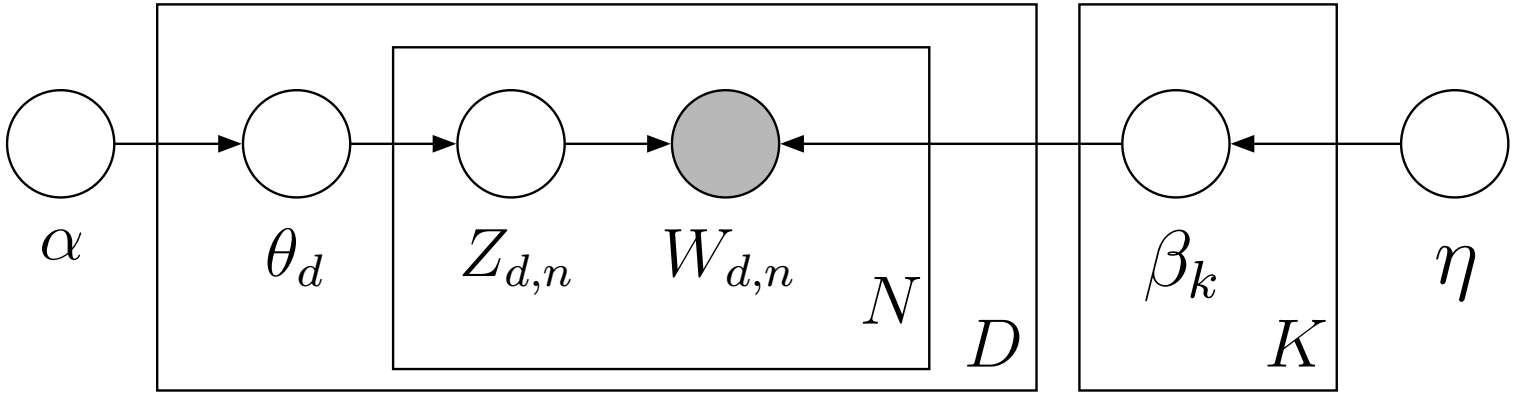


- Computing the posterior is intractable:

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

- Several approximation techniques have been developed.

Latent Dirichlet allocation



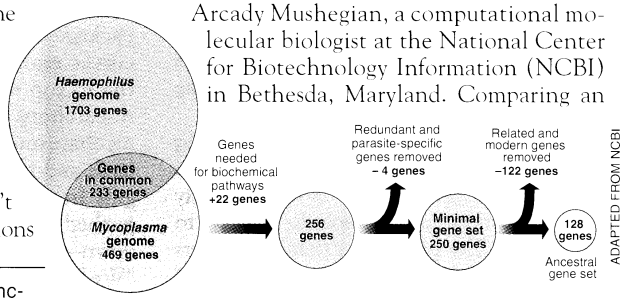
- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)

Example inference

Seeking Life's Bare (Genetic) Necessities

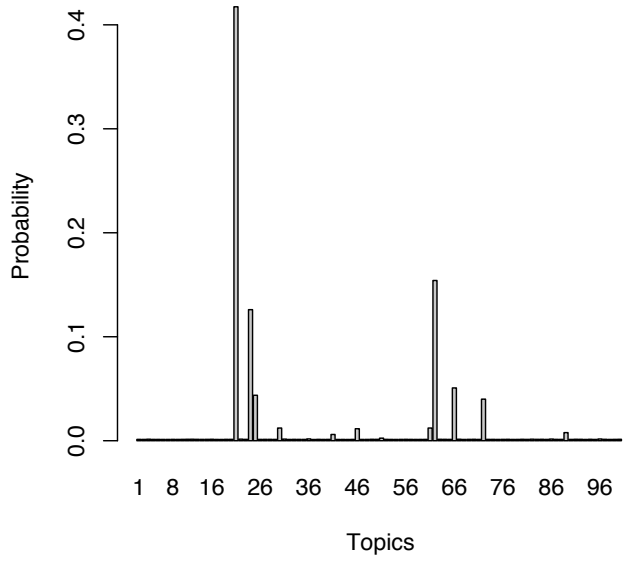
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

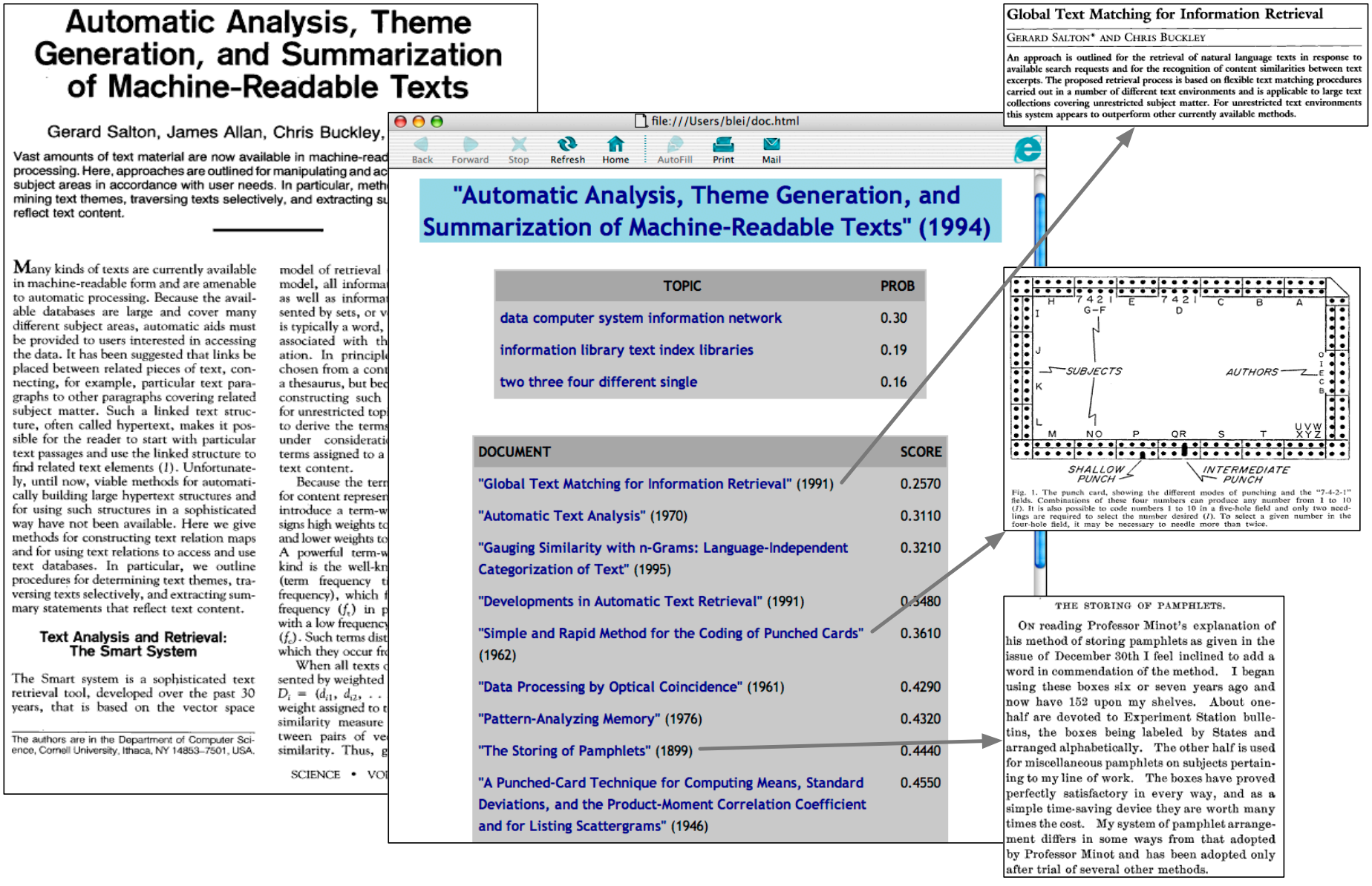
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.



Example topics

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

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).
- 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).

LDA summary

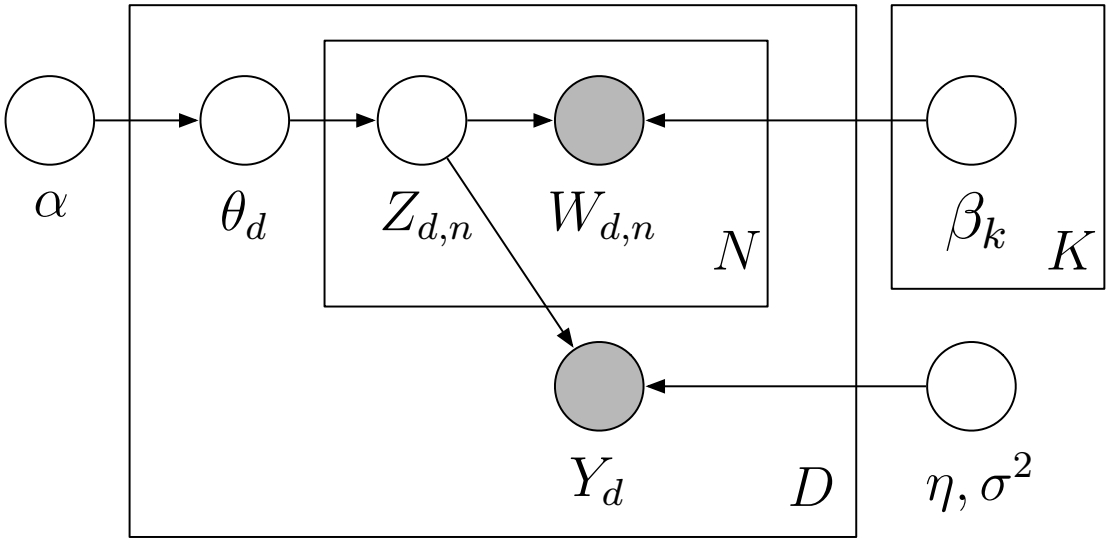
- *Modular*: It can be embedded in more complicated models.
- *General*: The data generating distribution can be changed.
- Variational inference is fast; lets us to analyze large data sets.

- See Blei et al., 2003 for details and a quantitative comparison. See my web-site for code and other papers.
- Jonathan Chang's excellent R package "lda" contains Gibbs sampling code for this model and many others.

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.

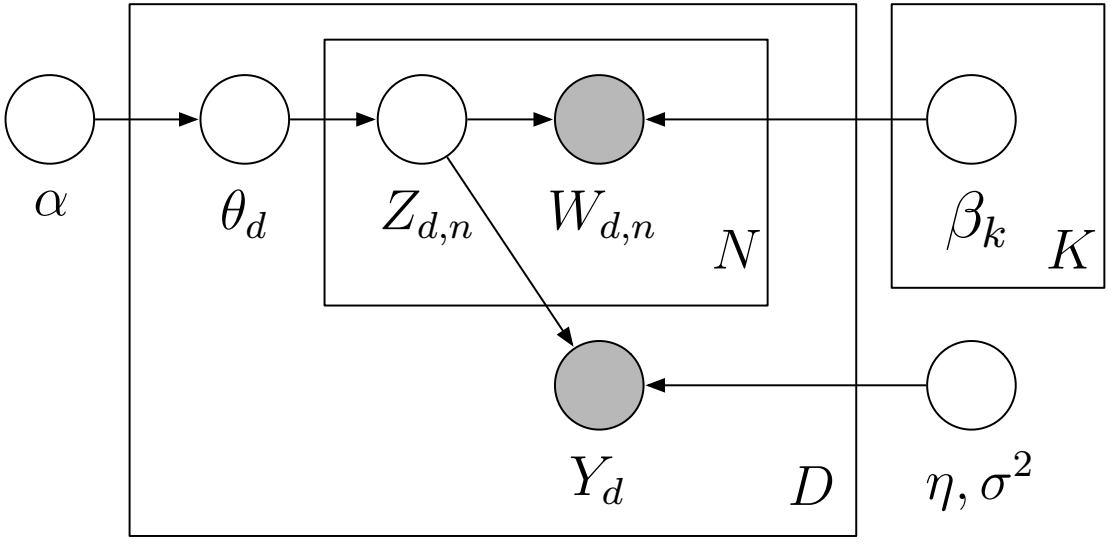
Supervised LDA



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

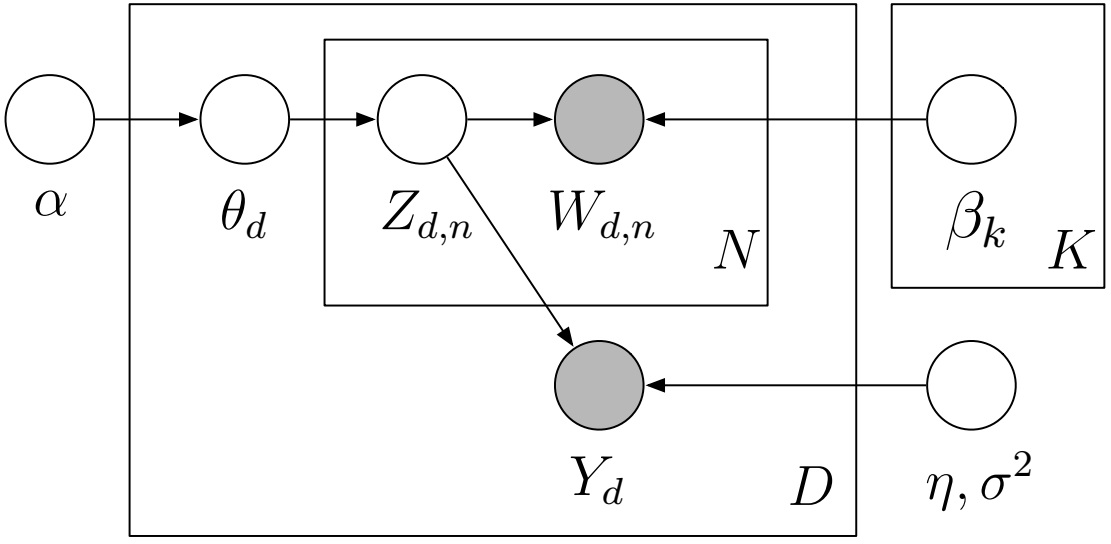
$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

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 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

- Factorization of q_ν determines complexity of optimization
- In *mean field variational inference* q_ν is 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 all other 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 as the corresponding conditional:

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

MFVI and conditional exponential families

- Variational inference is the following coordinate ascent algorithm

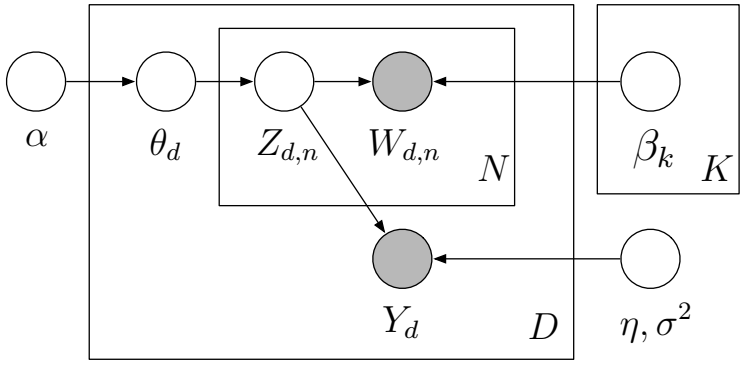
$$\nu_m = \mathrm{E}_{q_\nu}[g_m(\mathbf{Z}_{-m}, \mathbf{x})]$$

- Notice the relationship to Gibbs sampling.

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. **Many open issues!**
- Good papers: Beal's Ph.D. thesis, Wainwright and Jordan (2009)

Variational inference in sLDA



- In sLDA the variational bound is

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

- As in Blei, Ng, and Jordan (2003), 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 distinguishing term is

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

- The first expectation is

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

- The second expectation is

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

- Linear in ϕ_n, which 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

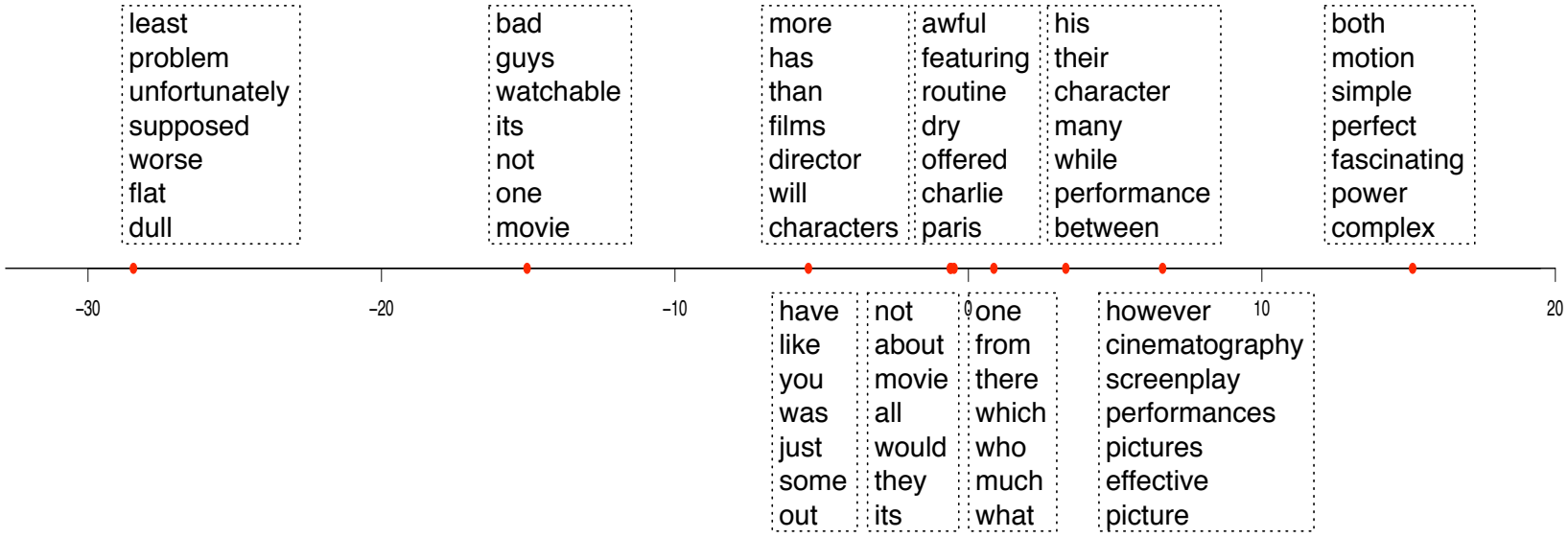
$$\hat{\sigma}^2_{\text{new}} \leftarrow (1/D)\{y^\top y - y^\top \mathrm{E}[A] \left(\mathrm{E}\left[A^\top A\right]\right)^{-1} \mathrm{E}[A]^\top y\}$$

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:

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

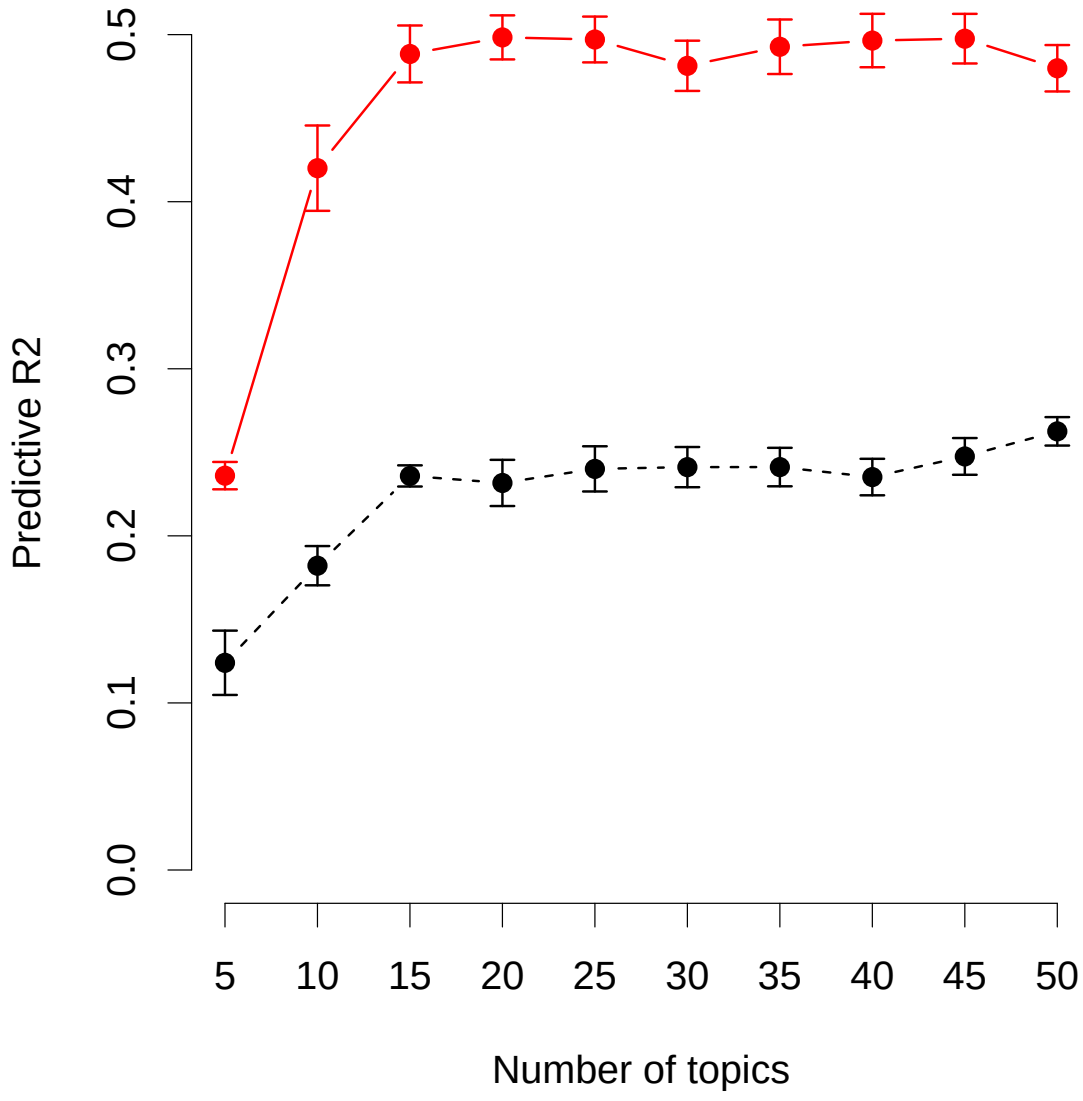
Example: Movie reviews



- 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.

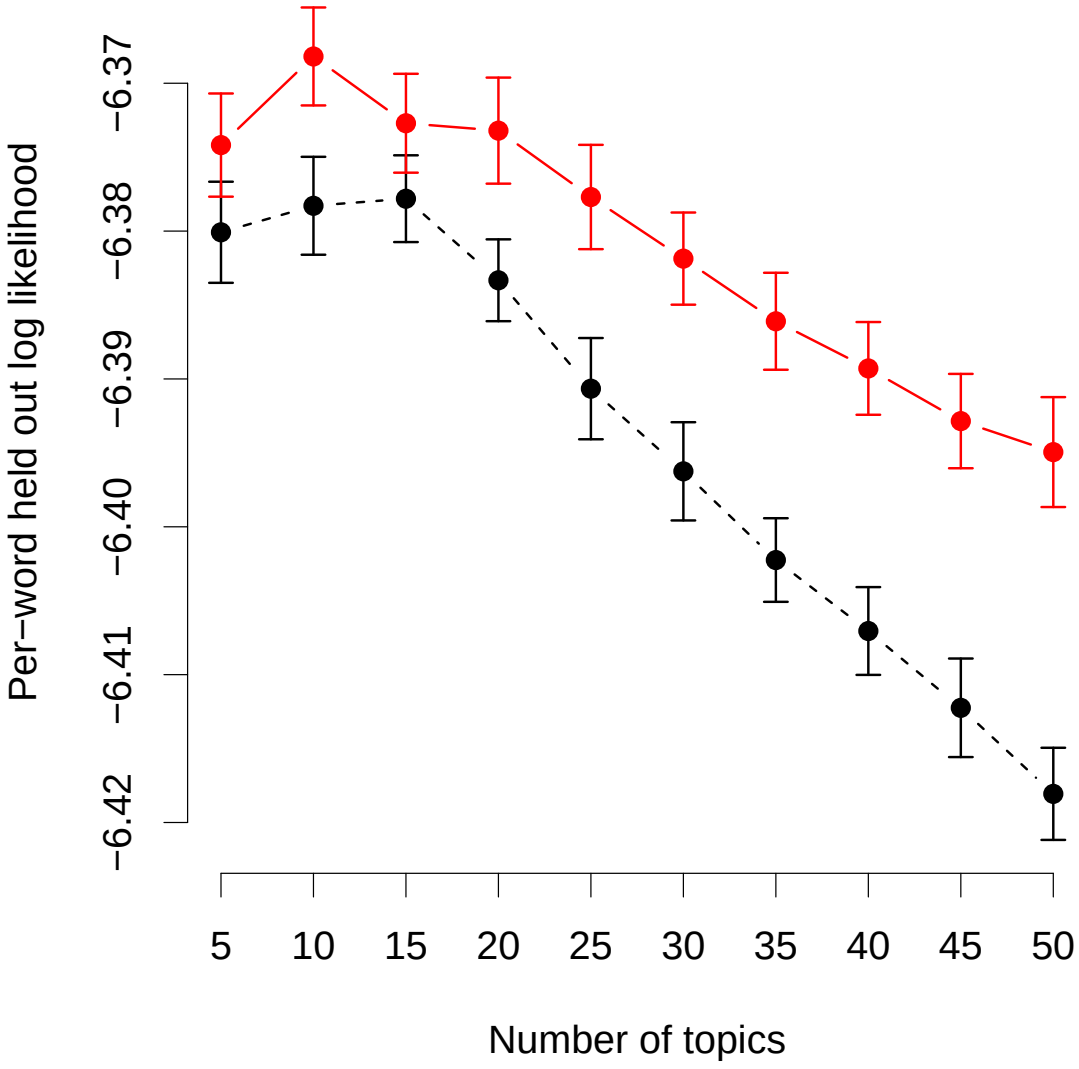
Predictive R2

(SLDA is red.)



Held out likelihood

(SLDA is red.)



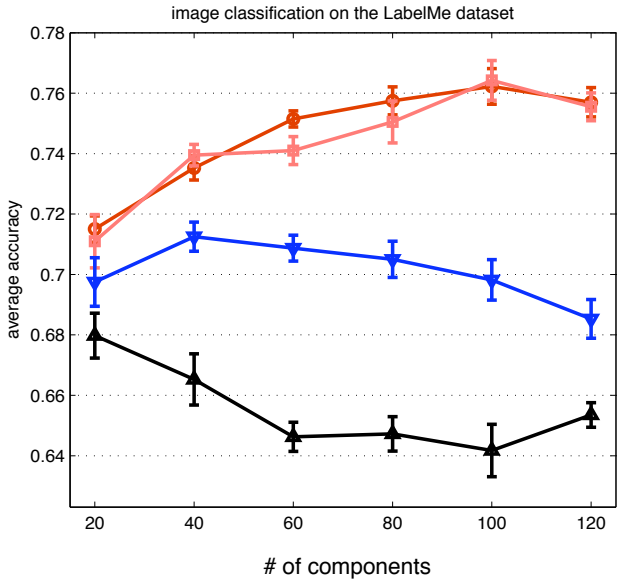
Diverse response types with GLMs

- 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

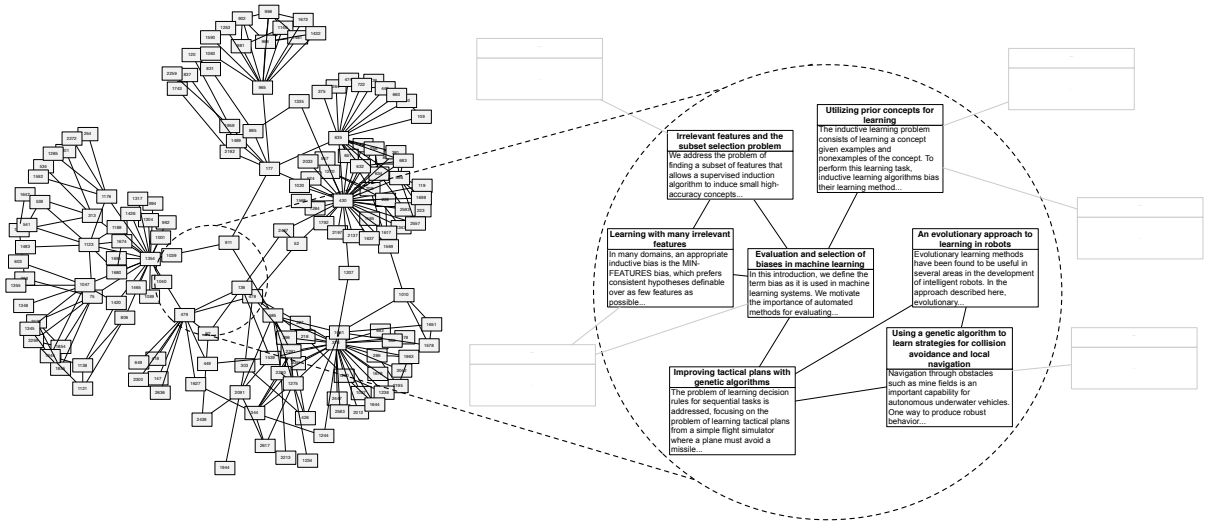
trees, buildings
occluded, window

(SLDA for image classification, with Chong Wang)

Supervised topic models

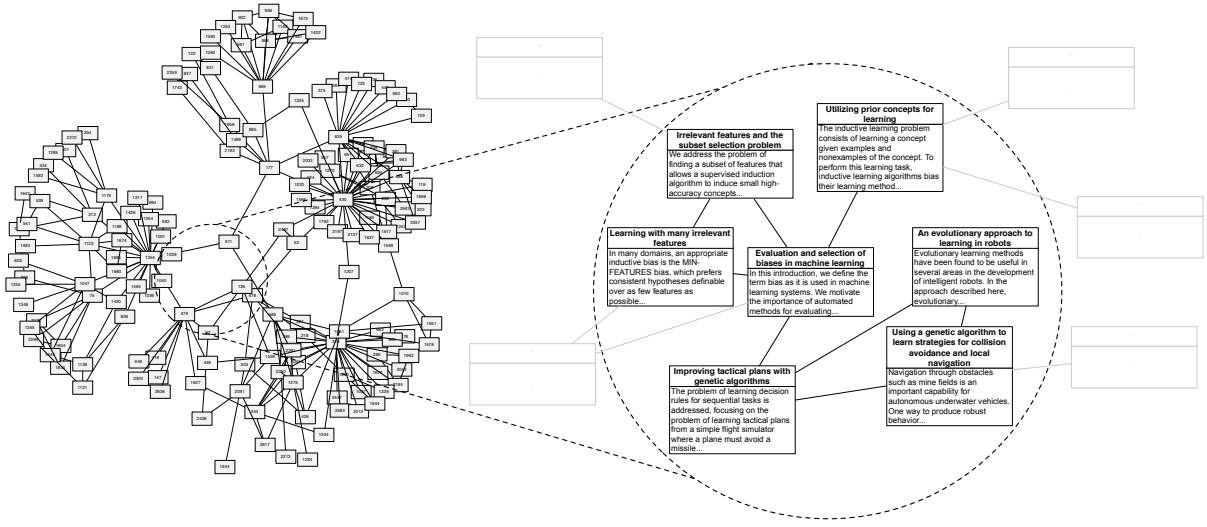
- SLDA enables model-based regression where the predictor "variable" is a text document.
- It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).
- SLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.
- LDA + regression compared to sLDA is like principal components regression compared to partial least squares.
- Paper: Blei and McAuliffe, NIPS 2007.

Relational topic models



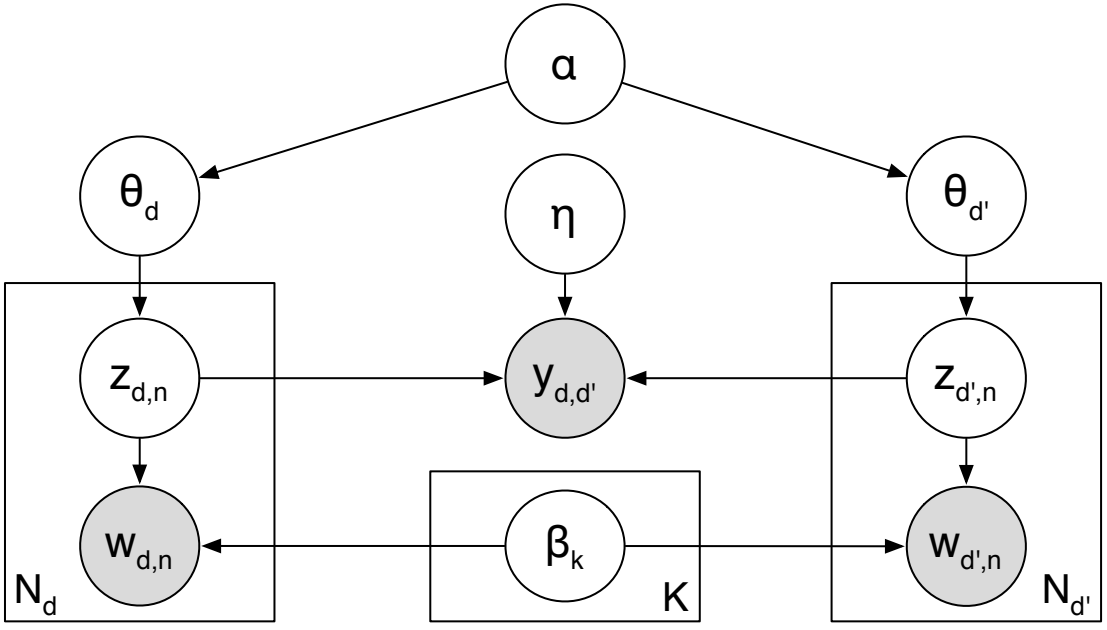
- Many data sets contain **connected observations**.
- For example:
  - Citation networks of documents
  - Hyperlinked networks of web-pages.
  - Friend-connected social network profiles

Relational topic models



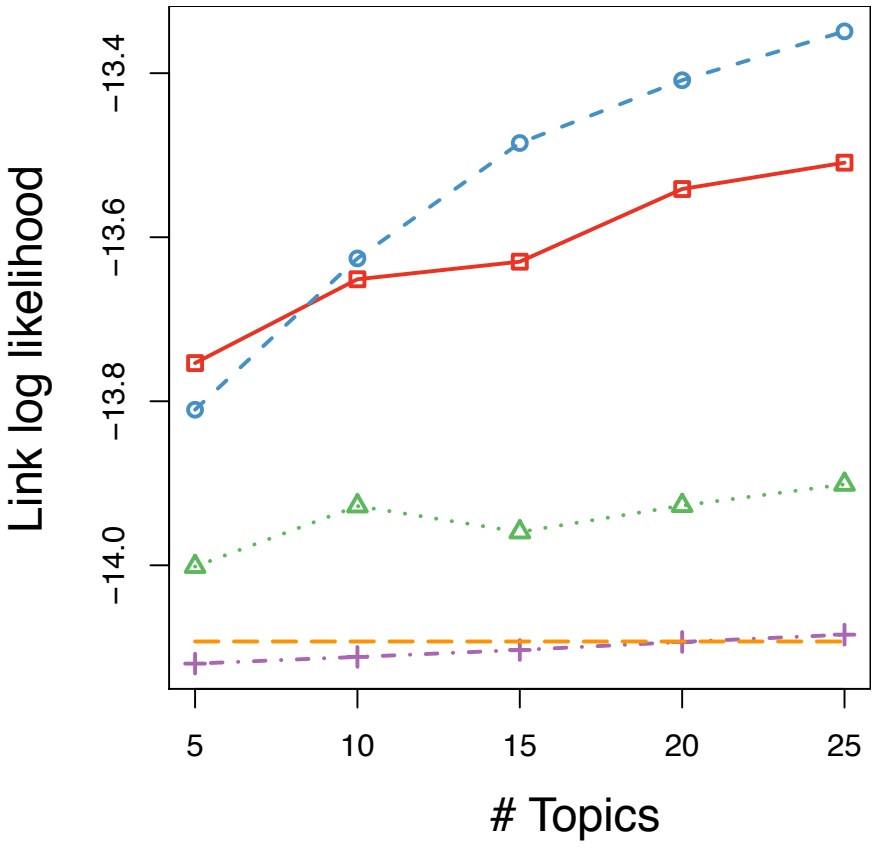
- Research has focused on finding communities and patterns in the link-structure of these networks (Kemp et al. 2004, Hoff et al., 2002, Hofman and Wiggins 2007, Airoldi et al. 2008).
- By adapting supervised topic modeling, we can build a good model of **content and structure**.
- RTMs find related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

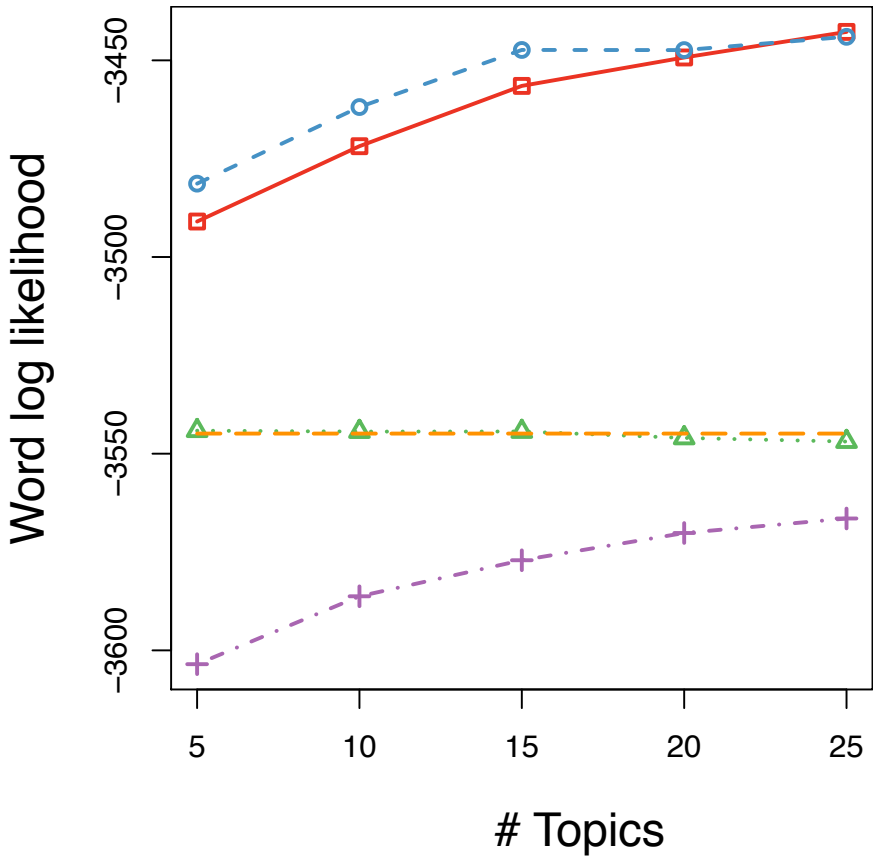
Relational topic models

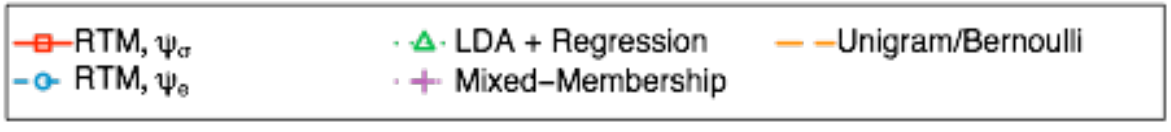


- Binary response variable with each pair of documents
- Adapt variational EM algorithm for sLDA with binary GLM response model (with different link probability functions).
- Allows predictions that are out of reach for traditional models.

Predictive performance of one type given the other



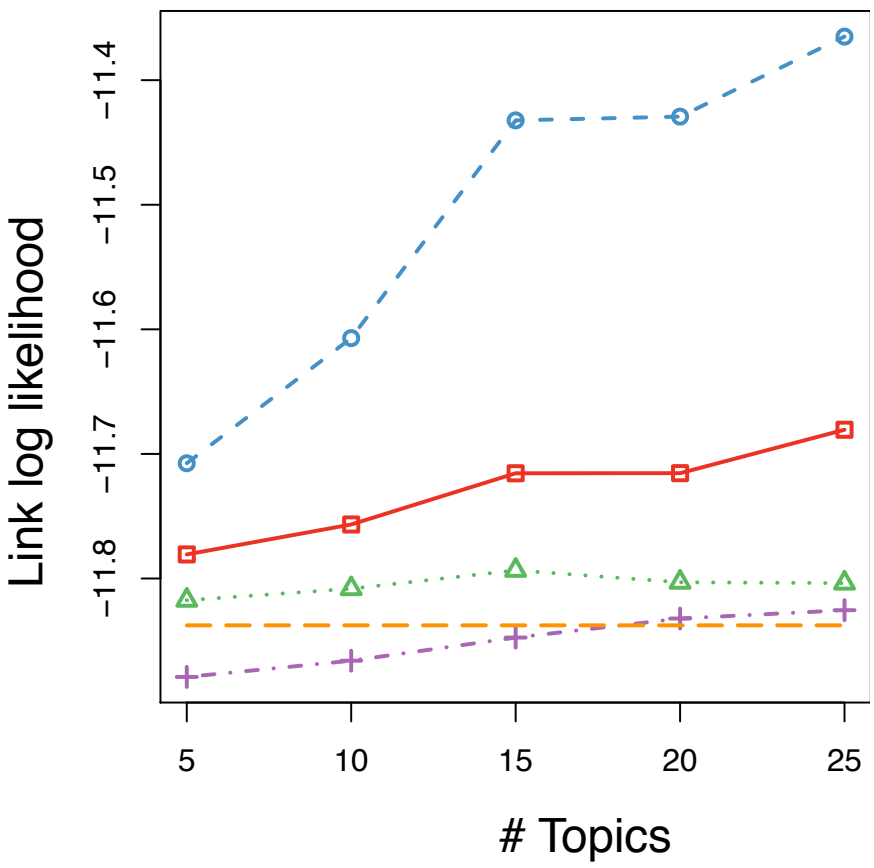


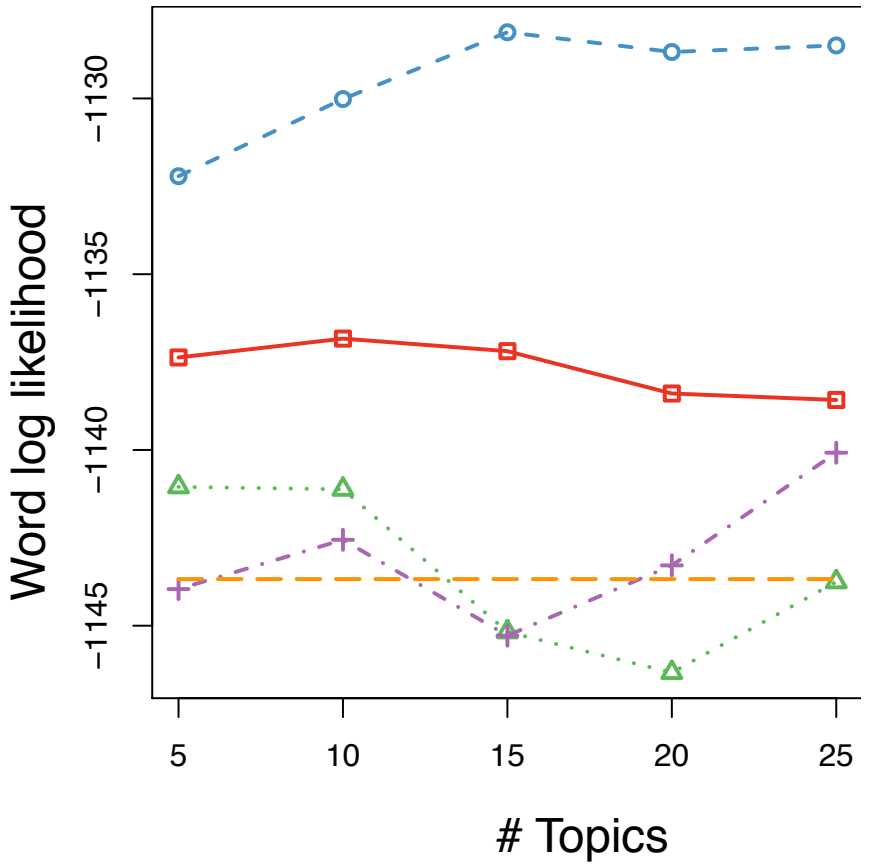


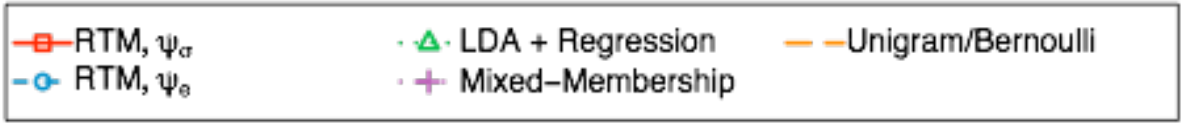
Cora corpus (McCallum et al., 2000)

Note: Traditional models of networks cannot perform these tasks.

Predictive performance of one type given the other

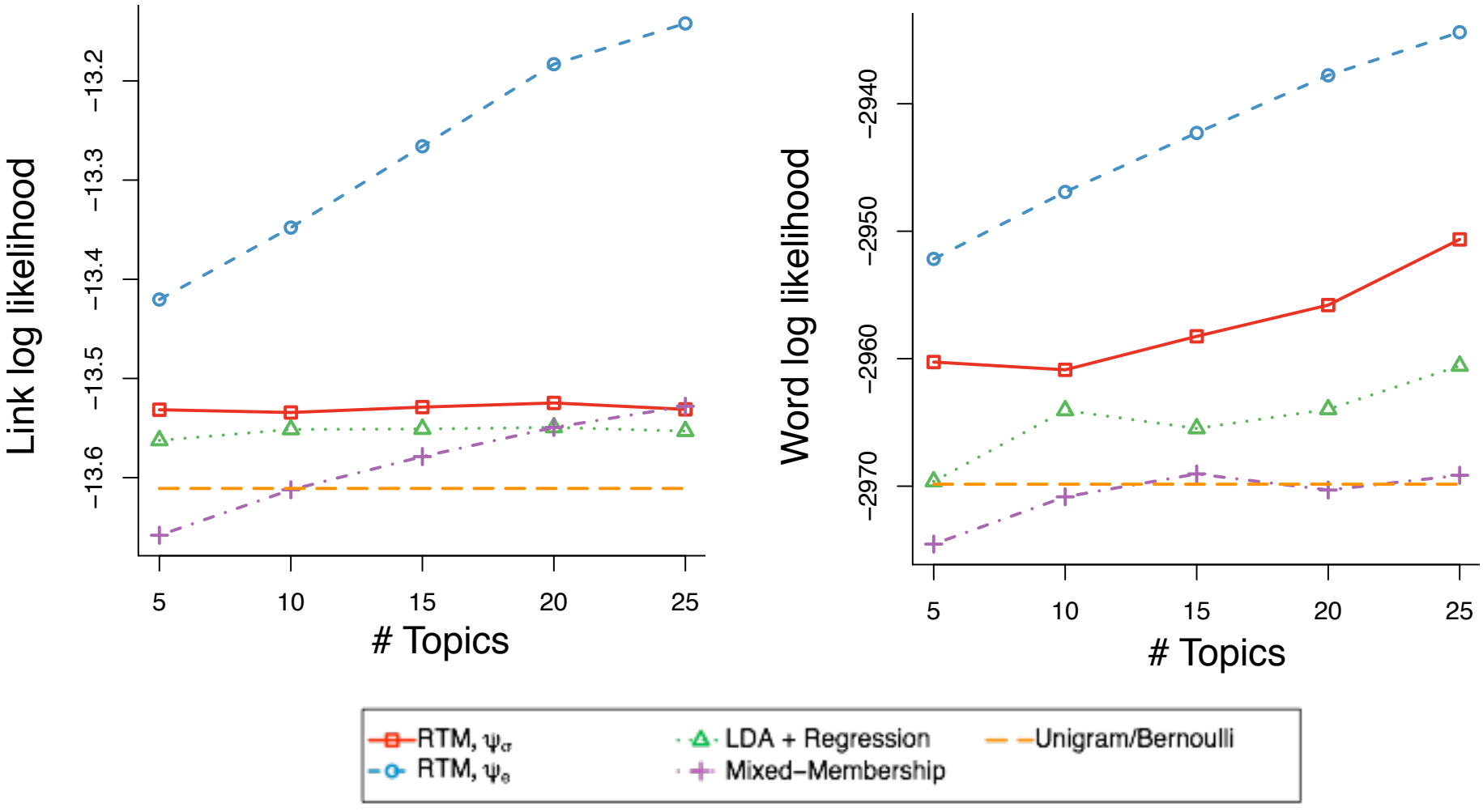






WebKB corpus (Craven et al., 1998)

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**

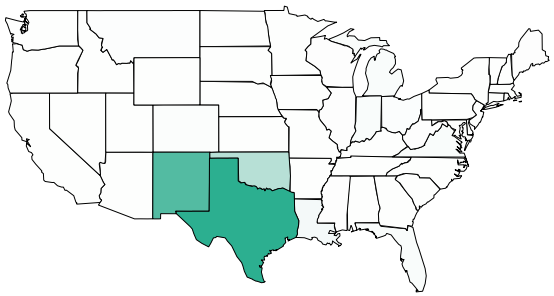
Given a new document, which documents is it likely to link to?

Predicting links from documents

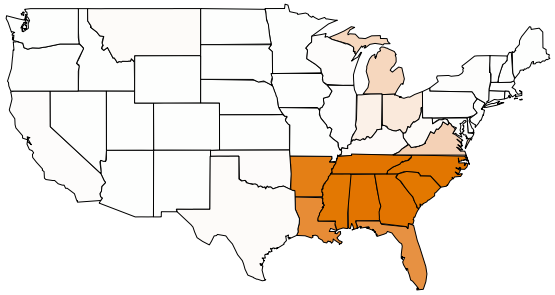
Competitive environments evolve better solutions for complex tasks	
Coevolving High Level Representations A Survey of Evolutionary Strategies **Genetic Algorithms in Search, Optimization and Machine Learning** **Strongly typed genetic programming in evolving cooperation strategies** Solving combinatorial 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?

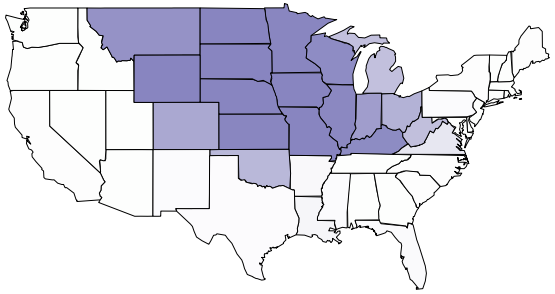
Spatially consistent topics



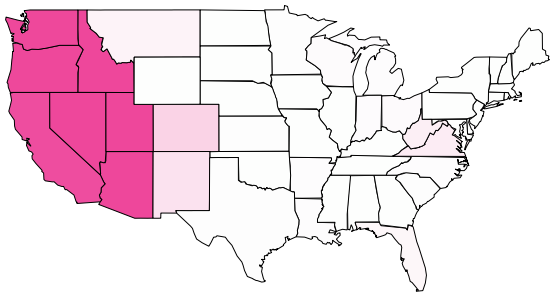
Topic 1



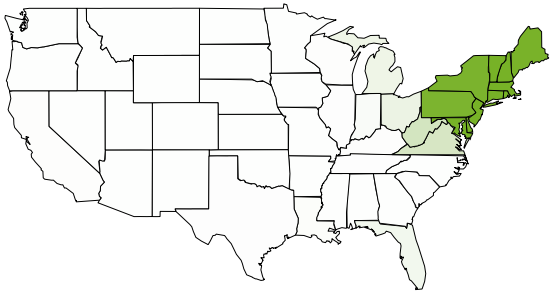
Topic 2



Topic 3



Topic 4



Topic 5

- Links are the adjacency matrix of states
- Documents are geographically-tagged news articles from Yahoo! Not exchangeable.
- RTM finds spatially consistent topics.

Summary

- 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
- Paper: Chang and Blei, AISTATS 2009.

JSTOR Discipline Analysis

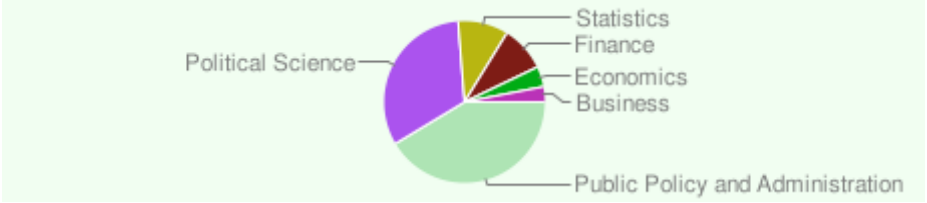
- Another kind of “supervised” topic model is where the topics have external meaning (Rammage and Manning, 2009)
- JSTOR attaches each journal to a discipline
- E.g., JASA is in Statistics; PNAS is in General Science
- We can how interdisciplinary the articles are
  - Compute a distribution over terms for each discipline
  - Find topic proportions for each article
- (Work done by Sean Gerrish)

Tax Innovation in the States: Capitalizing on Political Opportunity

Frances Stokes Berry;William D. Berry. *American Journal of Political Science* (1992), pp. 715-742

Journal Disciplines:

- **Political Science**

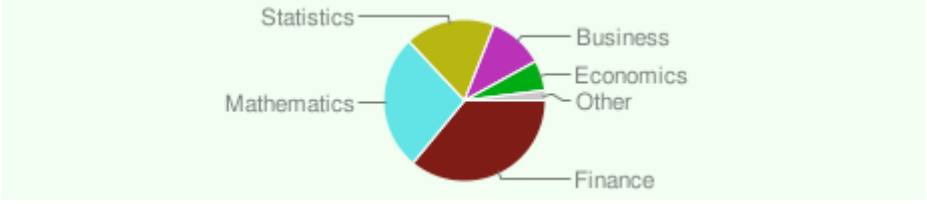


Chaos and Nonlinear Forecastability in Economics and Finance

Blake LeBaron. *Philosophical Transactions: Physical Sciences and Engineering* (1994), pp. 397-404

Journal Disciplines:

- **Mathematics**
- **Biological Sciences**
- **General Science**

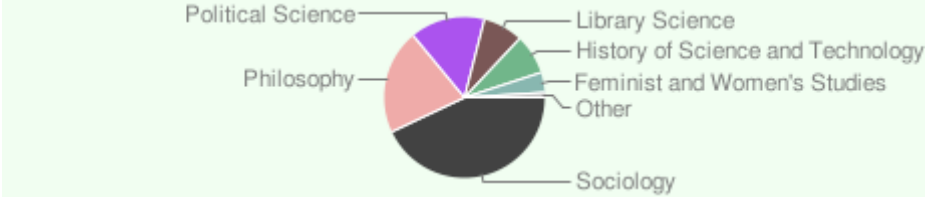


Reply: Theory Is Not a Social Dilemma

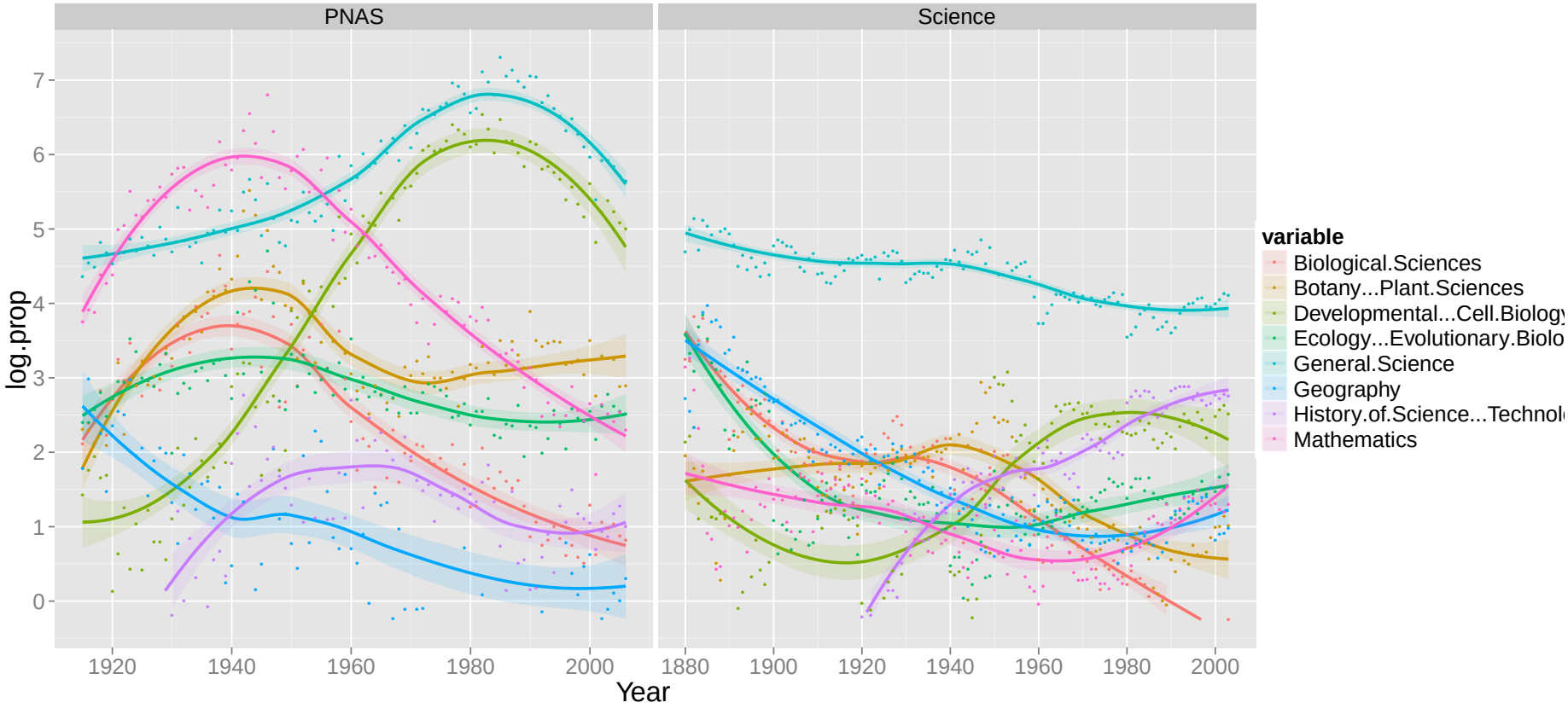
Gerald Marwell;Pamela Oliver. *Social Psychology Quarterly* (1994), pp. 373

Journal Disciplines:

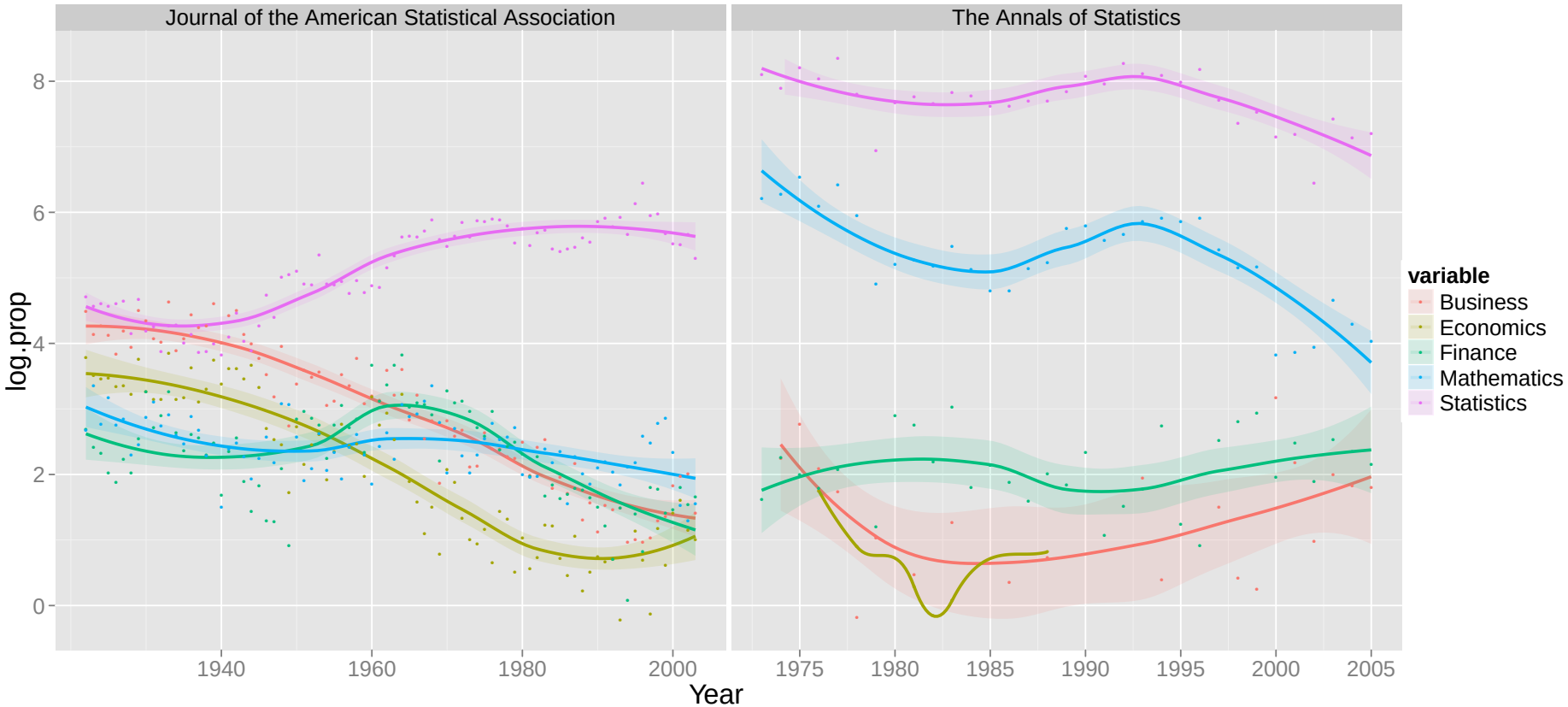
- **Psychology**
- **Sociology**



General Science over Time



Navel Gazing



“We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.”

(John Tukey, *The Future of Data Analysis*, 1962)