

# Probabilistic Topic Models

David M. Blei

Department of Computer Science  
Princeton University

June 26, 2012

# Probabilistic topic models



As more information becomes available, it becomes more difficult to find and discover what we need.

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

# Probabilistic topic models



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

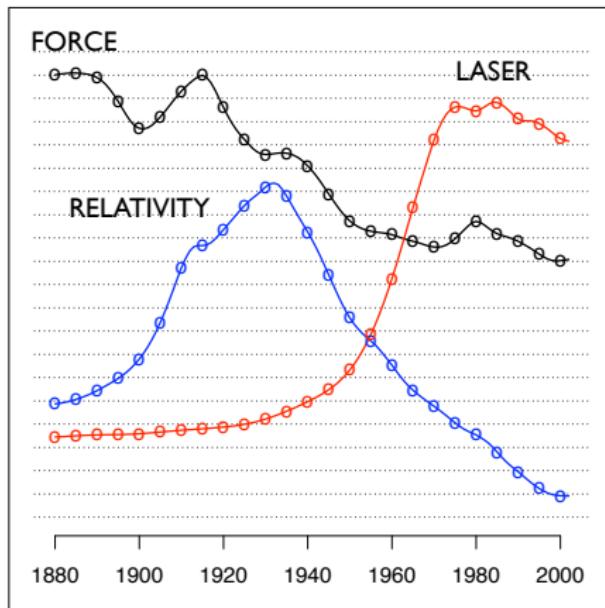
- ① Discover the hidden themes that pervade the collection.
- ② Annotate the documents according to those themes.
- ③ Use annotations to organize, summarize, and search the texts.

# Probabilistic topic models

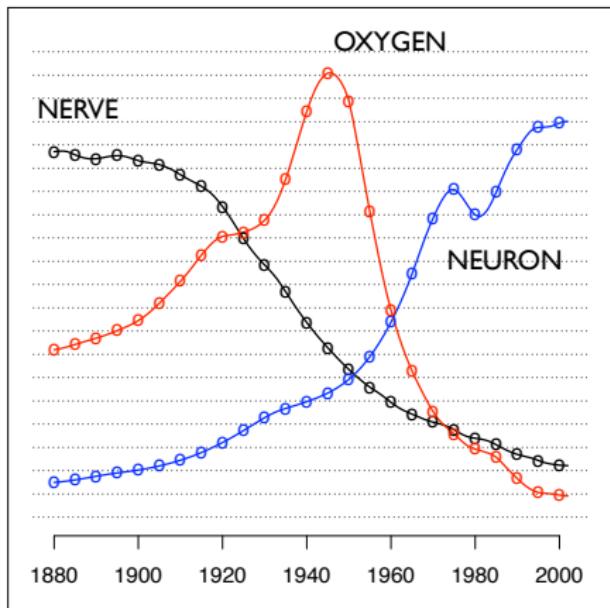
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

# Probabilistic topic models

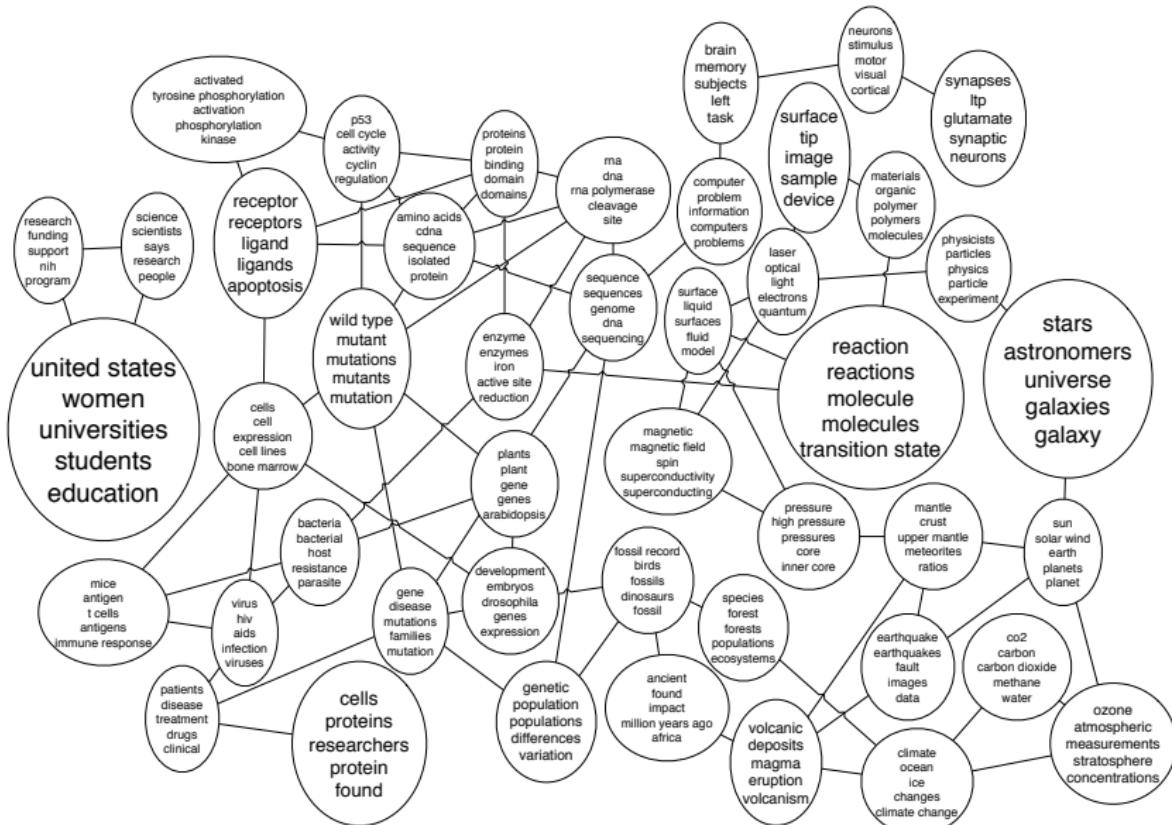
"Theoretical Physics"



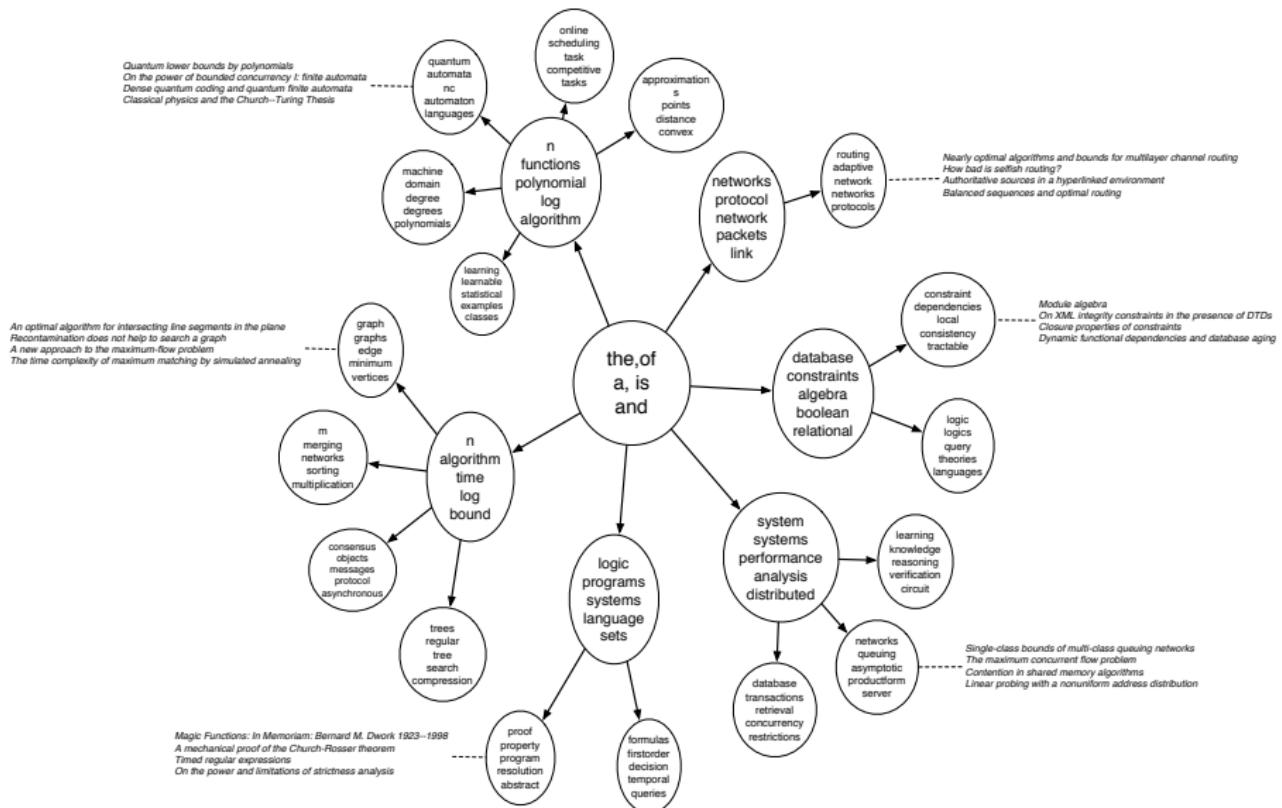
"Neuroscience"



# Probabilistic topic models



# Probabilistic topic models



# Probabilistic topic models



SKY WATER TREE  
MOUNTAIN PEOPLE



SCOTLAND WATER  
FLOWER HILLS TREE



SKY WATER BUILDING  
PEOPLE WATER



FISH WATER OCEAN  
TREE CORAL



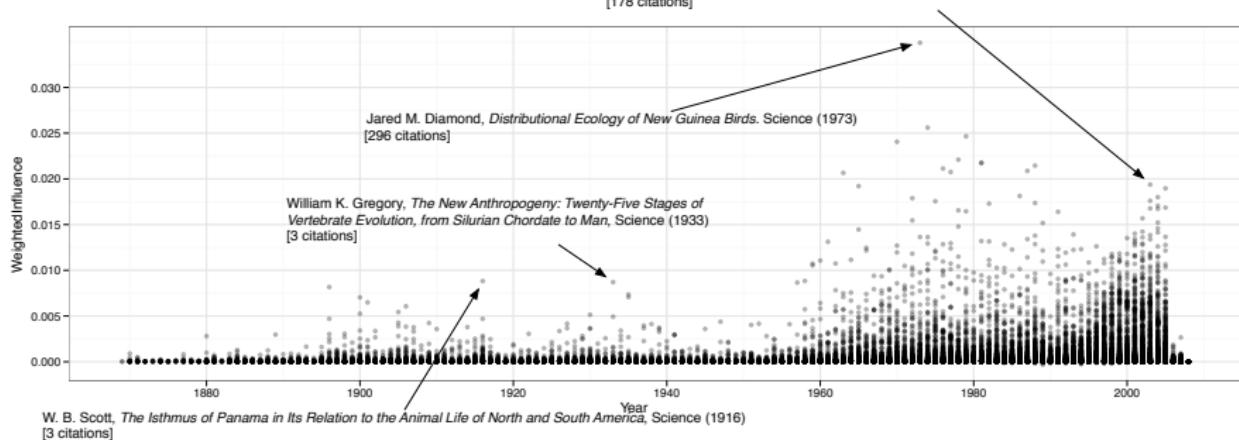
PEOPLE MARKET PATTERN  
TEXTILE DISPLAY



BIRDS NEST TREE  
BRANCH LEAVES

# Probabilistic topic models

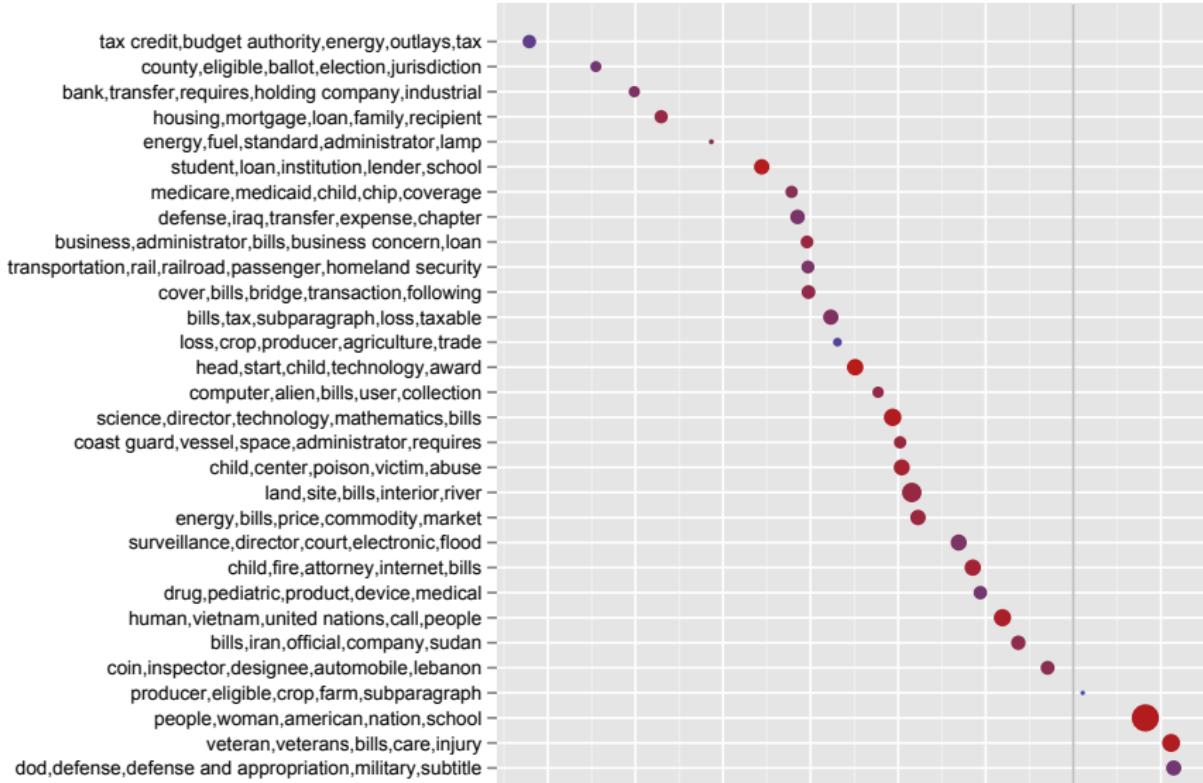
Derek E. Wildman et al., Implications of Natural Selection in Shaping 99.4% Nonsynonymous DNA Identity between Humans and Chimpanzees: Enlarging Genus Homo, PNAS (2003) [178 citations]



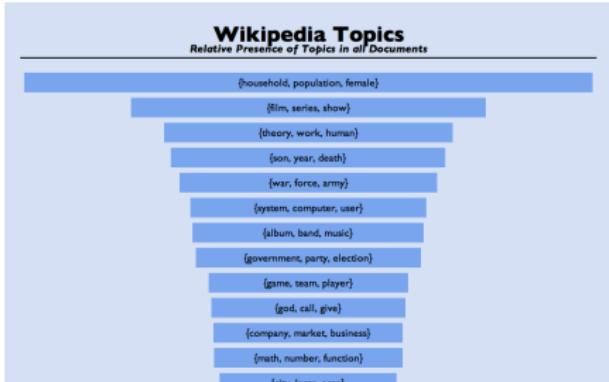
# Probabilistic topic models

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

# Probabilistic topic models

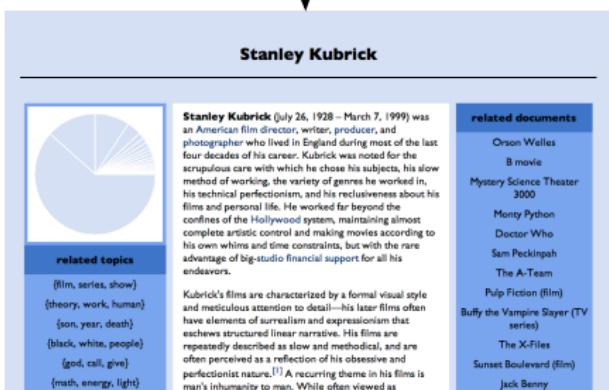


# Probabilistic topic models



**{film, series, show}**

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}



**{theory, work, human}**

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

# Probabilistic topic models

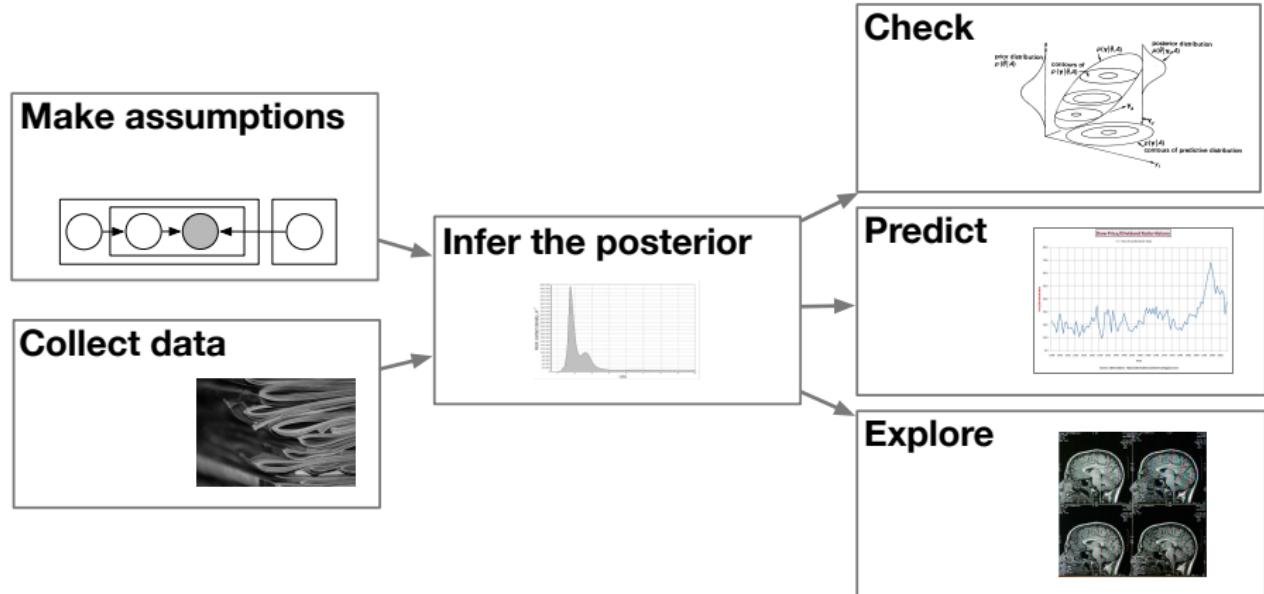
- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- How do I evaluate and check a topic model?
- What are some unanswered questions in this field?
- How can I learn more?

# Probabilistic topic models

Topic modeling is a case study in probabilistic modeling. It touches on

- Directed graphical models
- Conjugate priors and nonconjugate priors
- Time series modeling
- Modeling with graphs
- Hierarchical Bayesian methods
- Approximate posterior inference (MCMC, variational methods)
- Exploratory and descriptive data analysis
- Model selection and Bayesian nonparametric methods
- Mixed membership models
- Prediction from sparse and noisy inputs

**If you remember one picture...**



# Organization

- Introduction to topic modeling: Latent Dirichlet allocation
- Beyond latent Dirichlet allocation
- Posterior computation with scalable variational inference
- Model diagnostics with posterior predictive checks
- Discussion, open questions, and resources

## Some caveats

- This is a curated view of the field—we skip a lot of important ideas.
  - Gibbs sampling
  - Bayesian nonparametrics
- We focus on examples from our research group.
- To declutter, most references appear at the end. (Except, not yet.)

# **Introduction to Topic Modeling**

# Latent Dirichlet allocation (LDA)

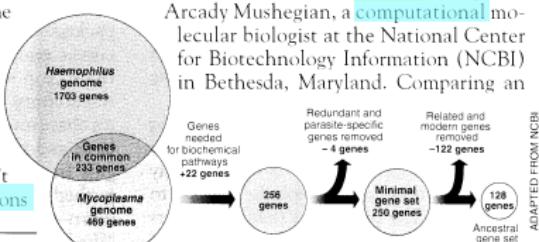
## Seeking Life's Bare (Genetic) Necessities

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

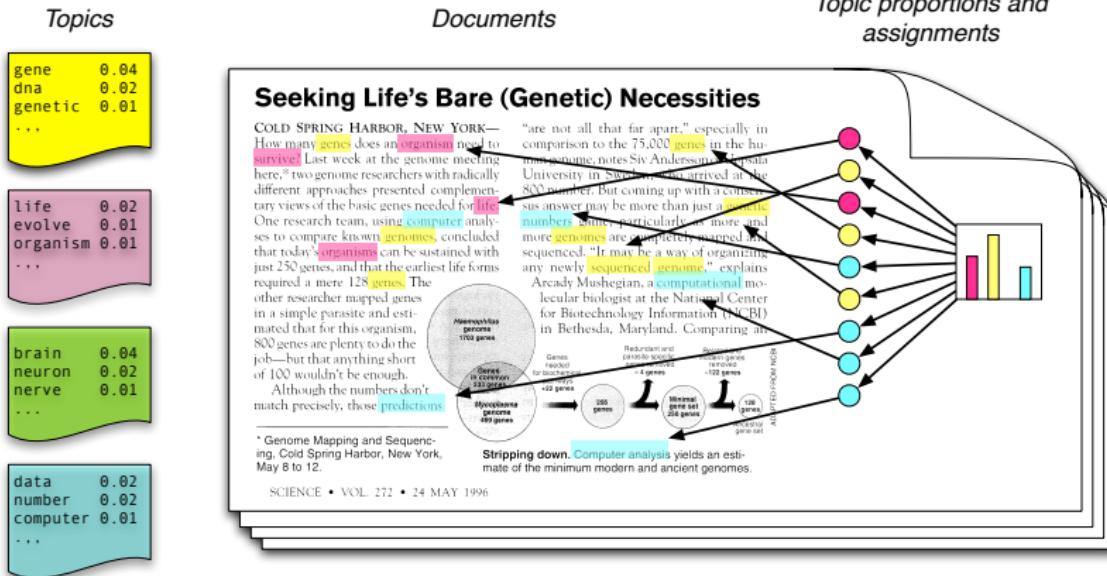
Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



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

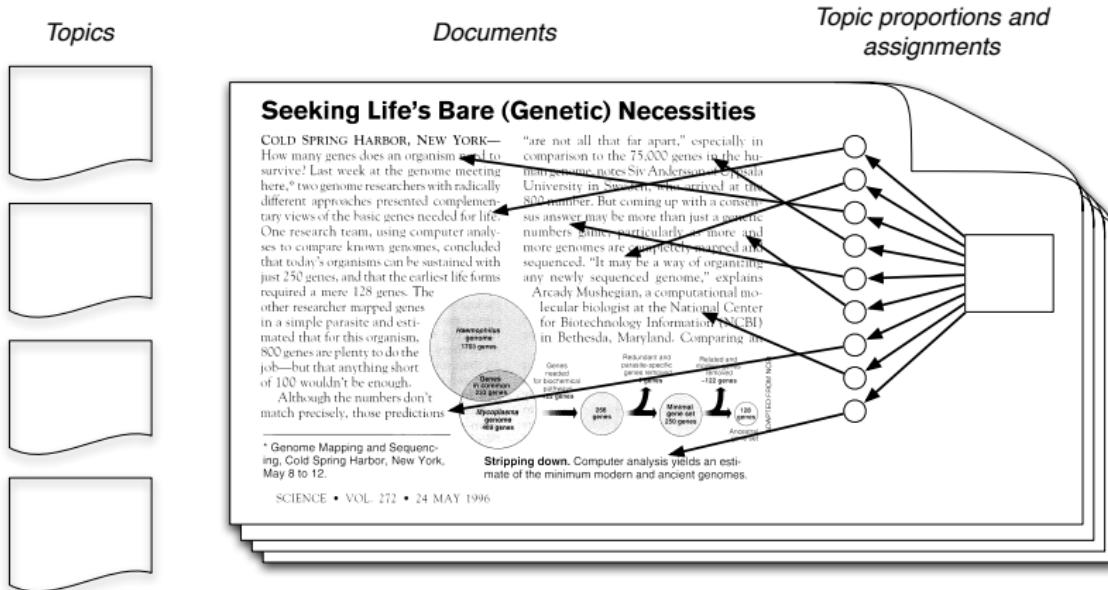
\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

# Latent Dirichlet allocation (LDA)



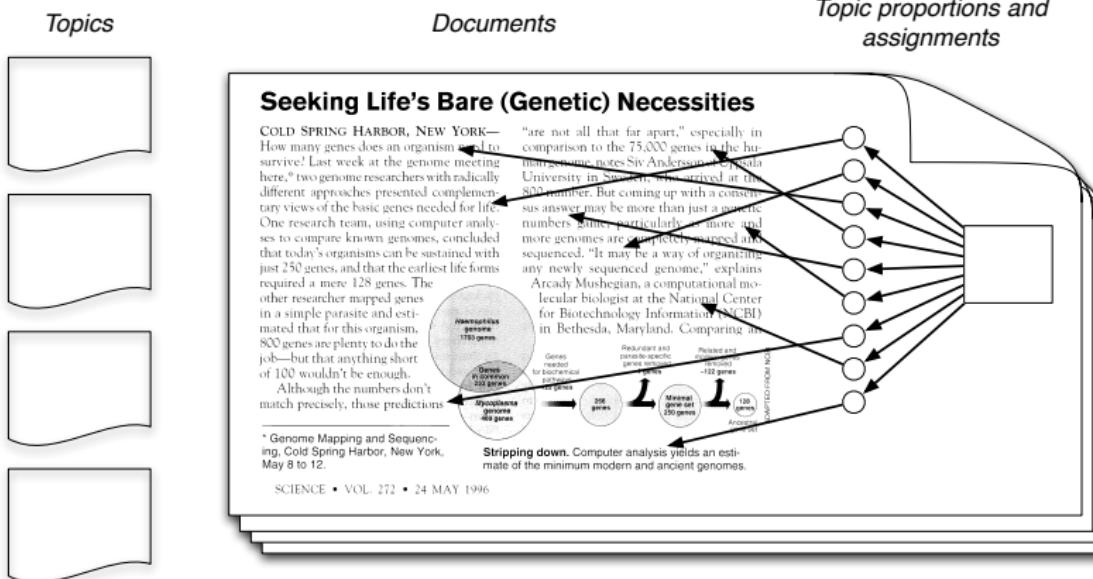
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

## Latent Dirichlet allocation (LDA)



- In reality, we only observe the documents
  - The other structure are **hidden variables**

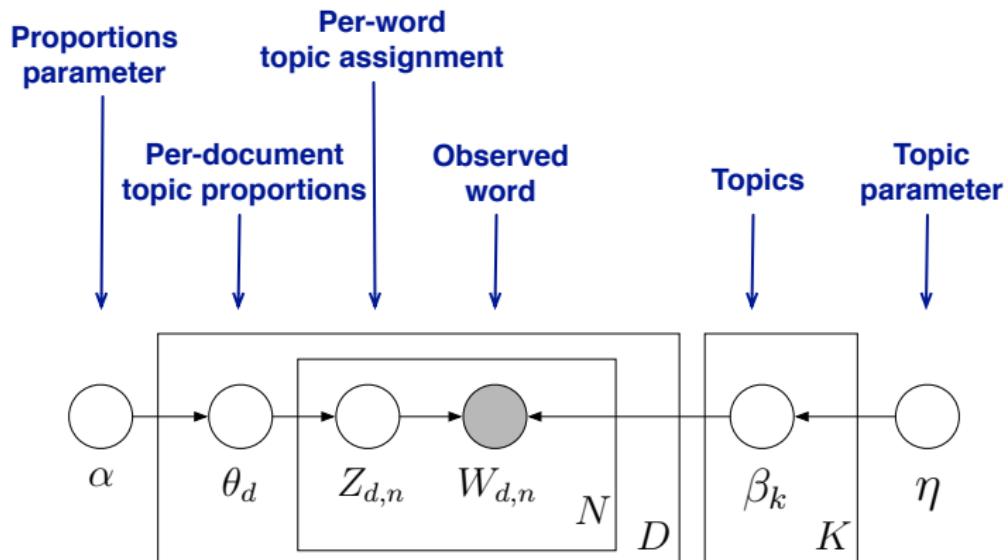
# Latent Dirichlet allocation (LDA)



- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

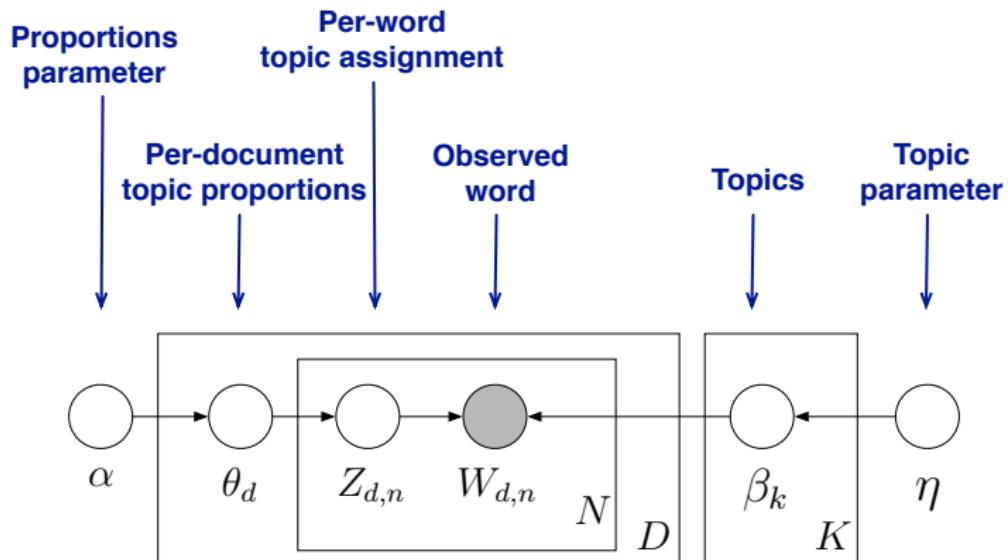
$$p(\text{topics, proportions, assignments} | \text{documents})$$

# LDA as a graphical model



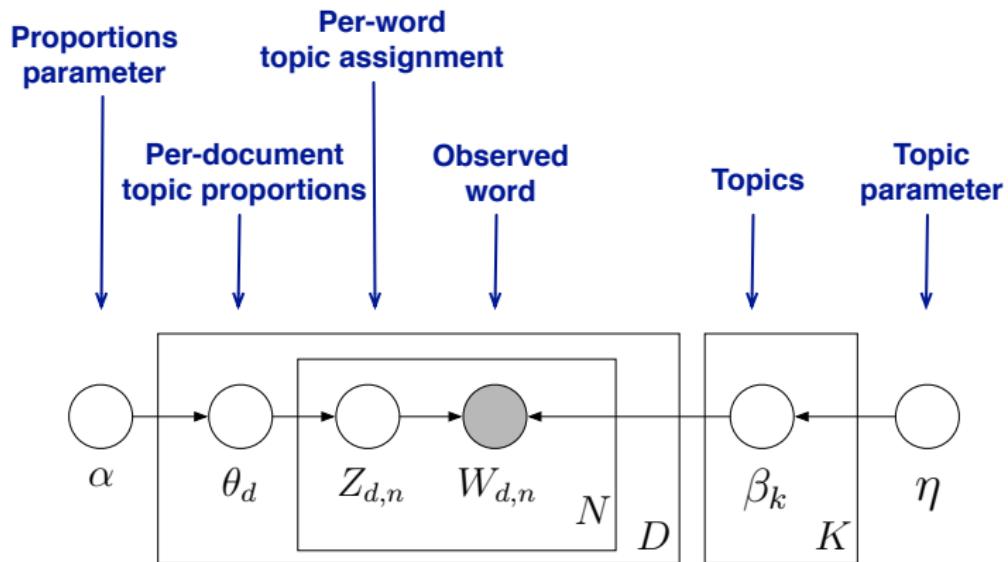
- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

# LDA as a graphical model



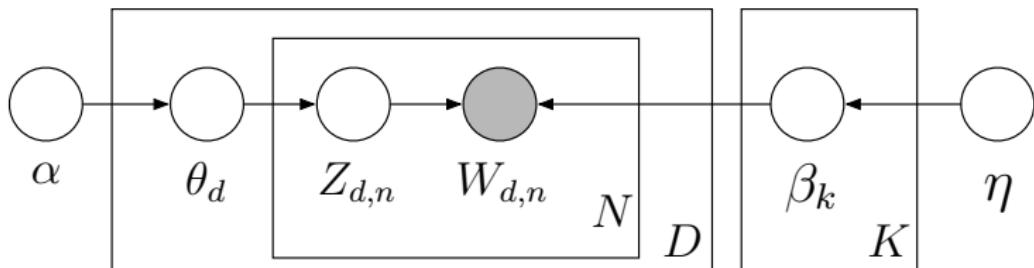
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

# LDA as a graphical model



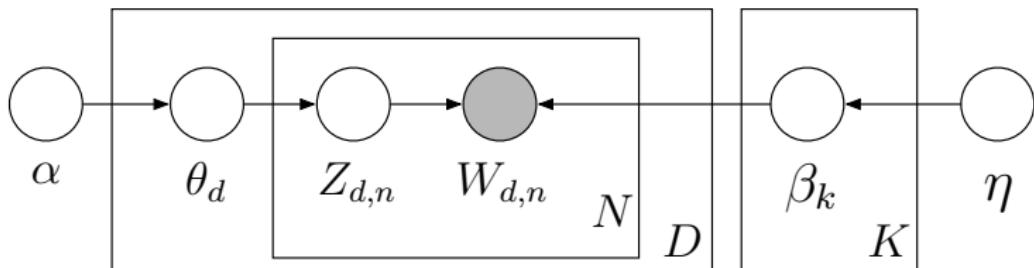
$$\prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

# LDA as a graphical model



- This joint defines a posterior.
- From a collection of documents, infer
  - Per-word topic assignment  $z_{d,n}$
  - Per-document topic proportions  $\theta_d$
  - Per-corpus topic distributions  $\beta_k$
- Then use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, exploration, ...

# LDA as a graphical model



## 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)
- Online variational inference (Hoffman et al., 2010)

Also see Mukherjee and Blei (2009) and Asuncion et al. (2009).

## Example inference



- **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

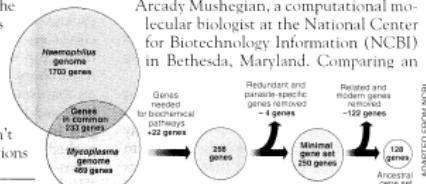
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>6</sup> 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

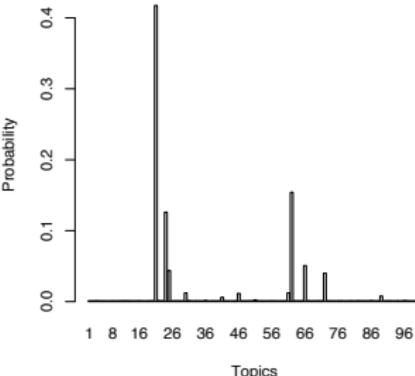
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Aracady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



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

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



# Example inference

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

1 dna gene sequence genes sequences human genome genetic analysis two	2 protein cell cells proteins receptor fig binding activity activation kinase	3 water climate atmospheric temperature global surface ocean carbon atmosphere changes	4 says researchers new university just science like work first years	5 mantle high earth pressure seismic crust temperature earths lower earthquakes
6 end article start science readers service news card circle letters	7 time data two model fig system number different results etc	8 materials surface high structure temperature molecules chemical molecular fig university	9 dna rna transcription protein site binding sequence proteins specific sequences	10 disease cancer patients human gene medical studies drug normal drugs
11 years million ago age university north early fig evidence record	12 species evolution population evolutionary university populations natural studies genetic biology	13 protein structure proteins two amino binding acid residues molecular structural	14 cells cell virus hiv infection immune human antigen infected viral	15 space solar observations earth stars university mass sun astronomers telescope
16 fax manager science aaas advertising sales member recruitment associate washington	17 cells cell gene genes expression development mutant mice fig biology	18 energy electron state light quantum physics electrons high laser magnetic	19 research science national scientific scientists new states university united health	20 neurons brain cells activity fig channels university cortex neuronal visual

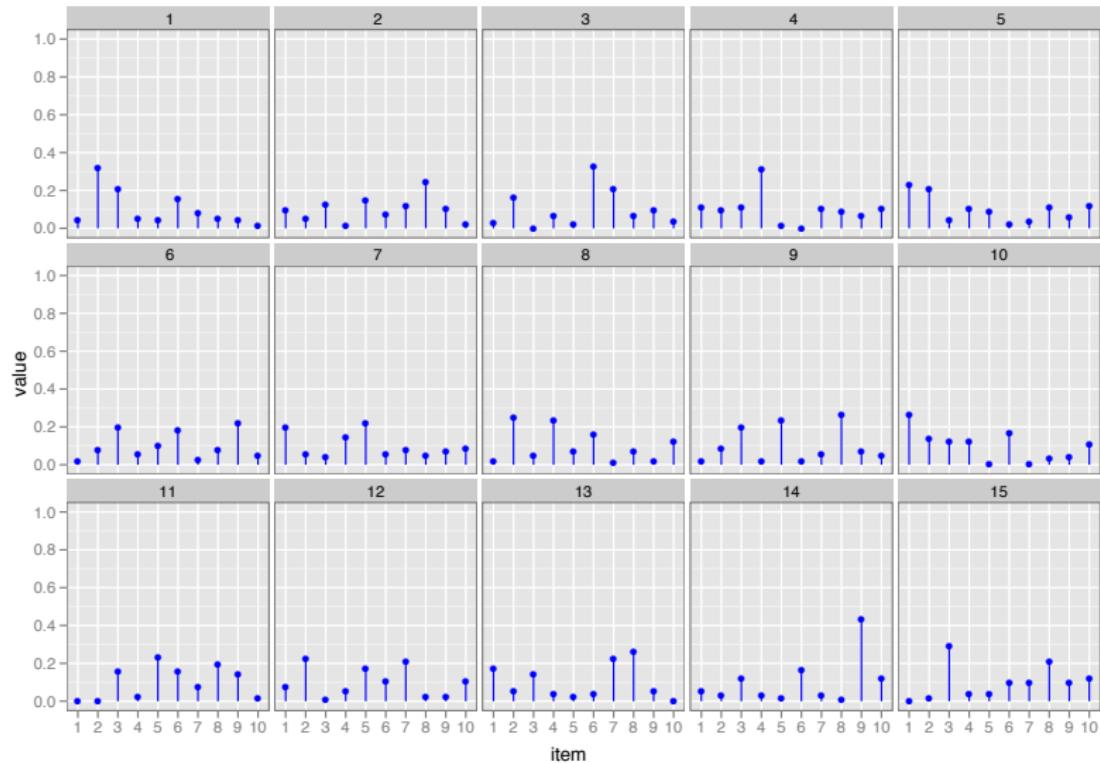
## Aside: The Dirichlet distribution

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

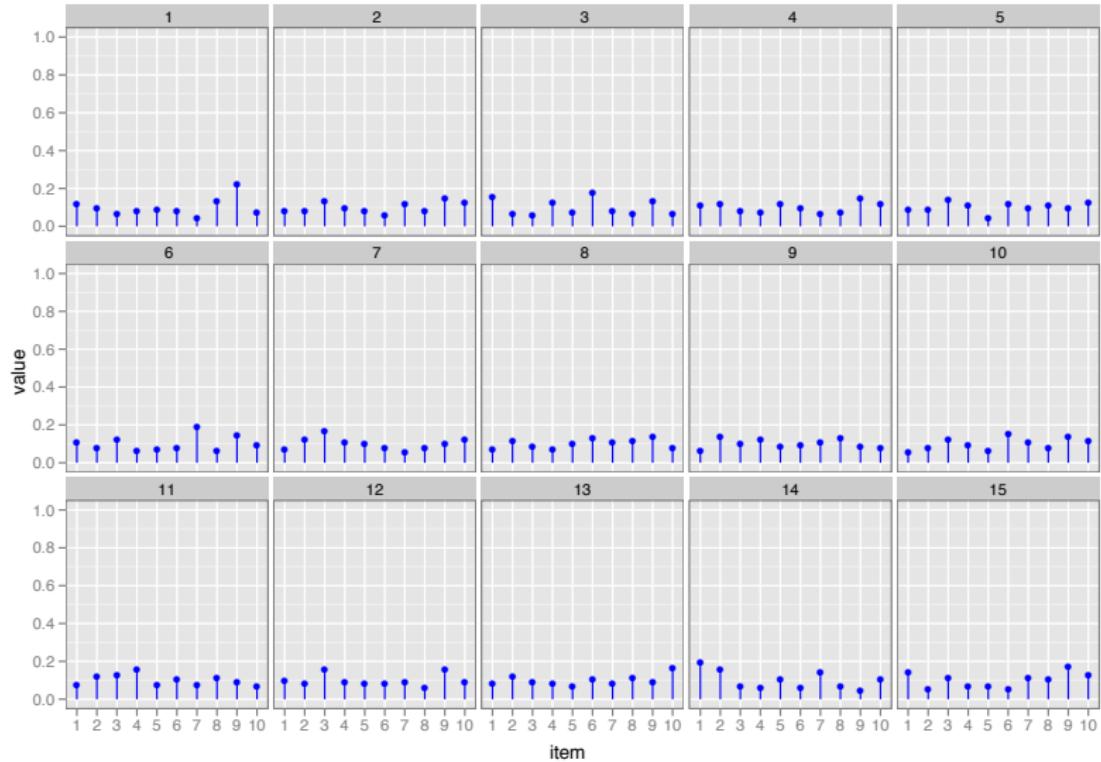
$$p(\theta | \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- It is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of  $\theta$  is a Dirichlet.
- The parameter  $\alpha$  controls the mean shape and sparsity of  $\theta$ .
- The topic proportions are a  $K$  dimensional Dirichlet.  
The topics are a  $V$  dimensional Dirichlet.

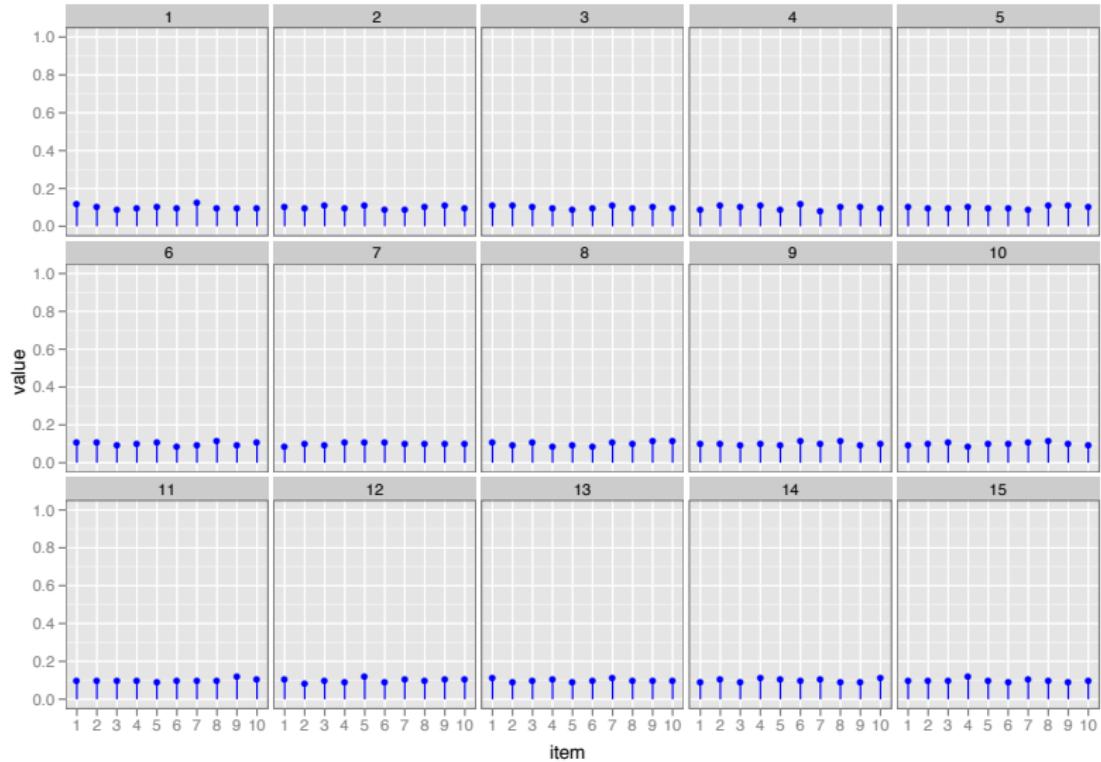
$\alpha = 1$



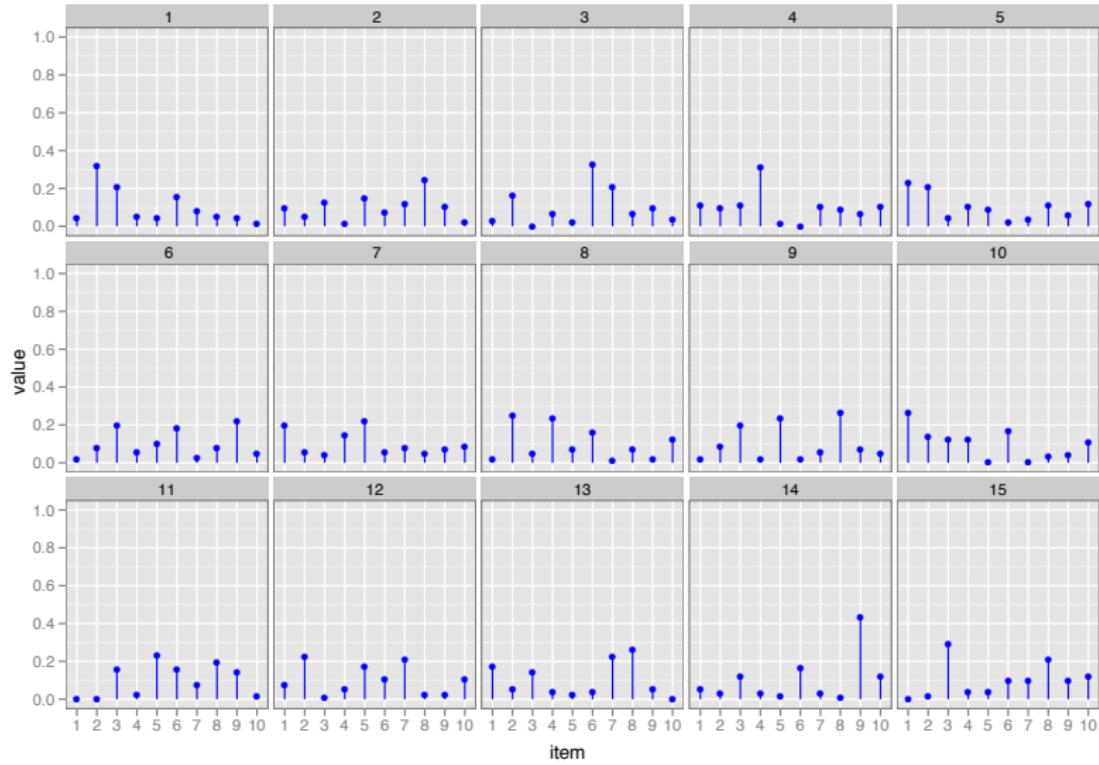
$\alpha = 10$



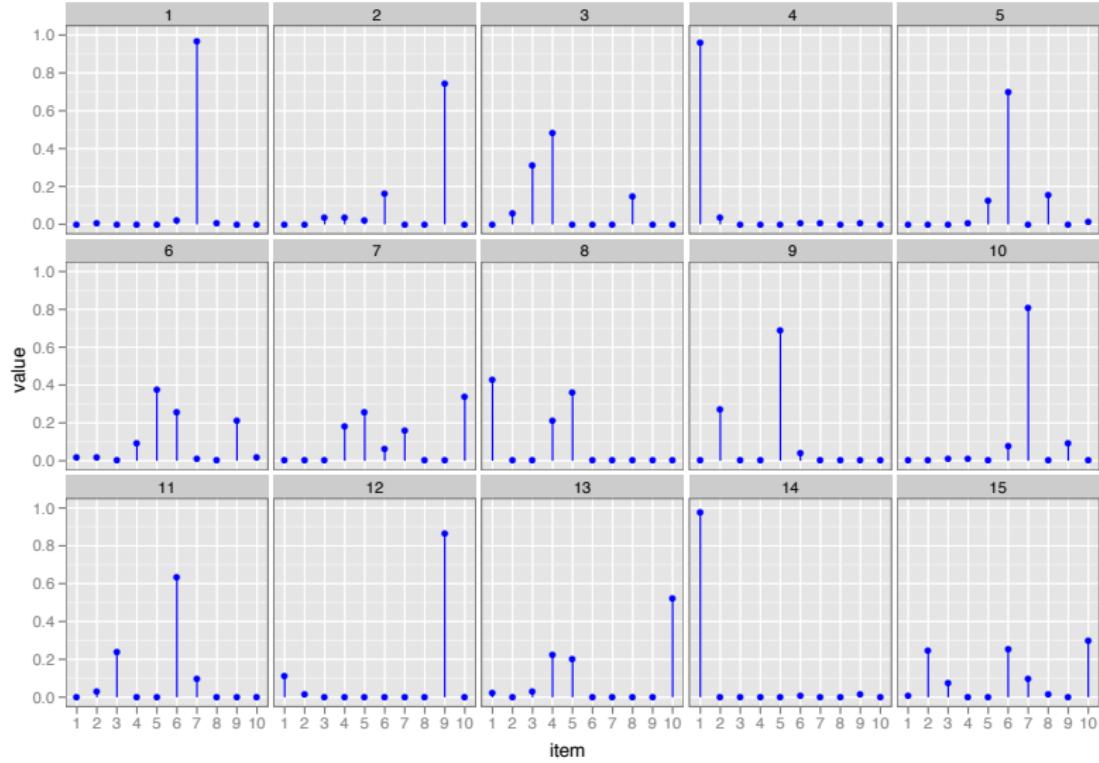
$\alpha = 100$



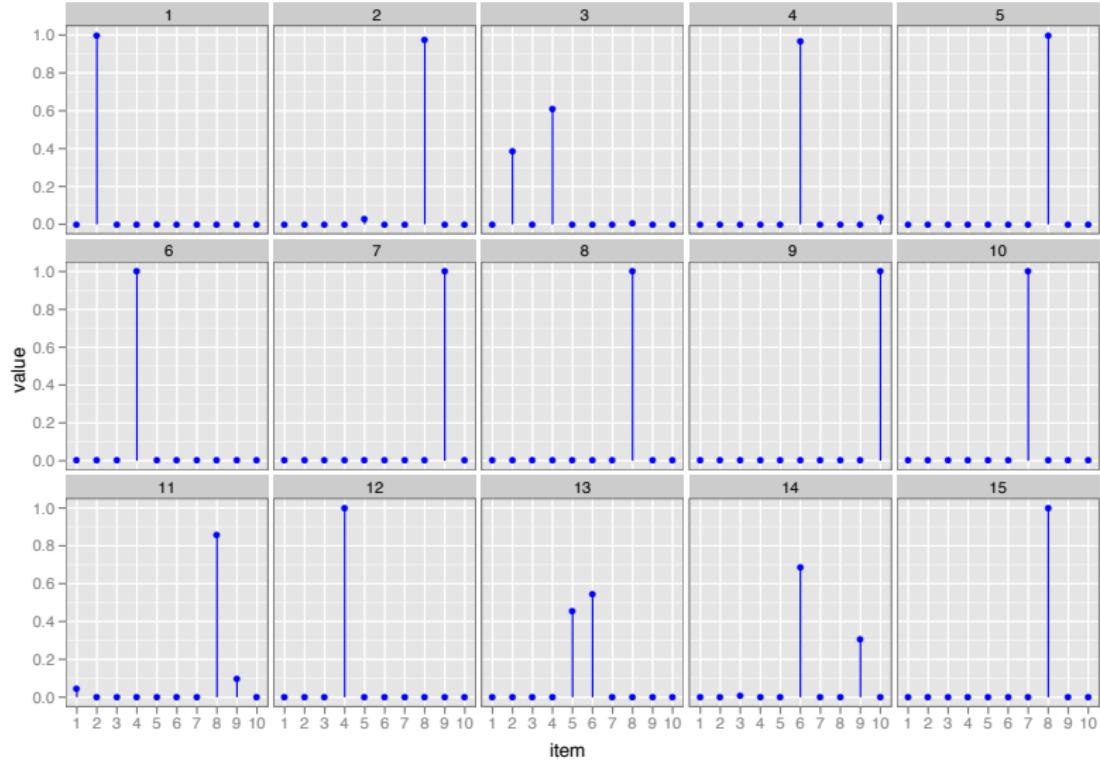
$$\alpha = 1$$



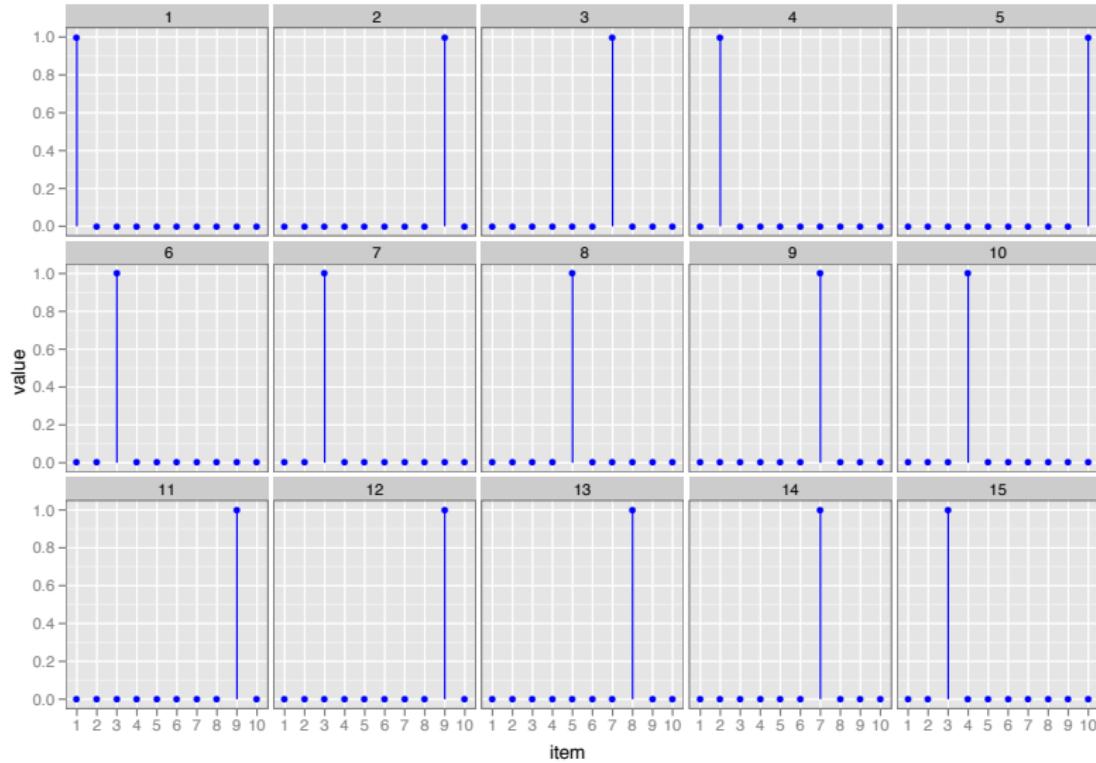
$$\alpha = 0.1$$



$\alpha = 0.01$



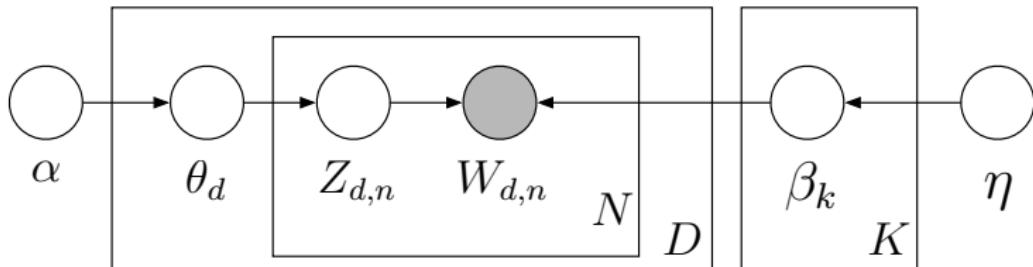
$\alpha = 0.001$



## Why does LDA “work”?

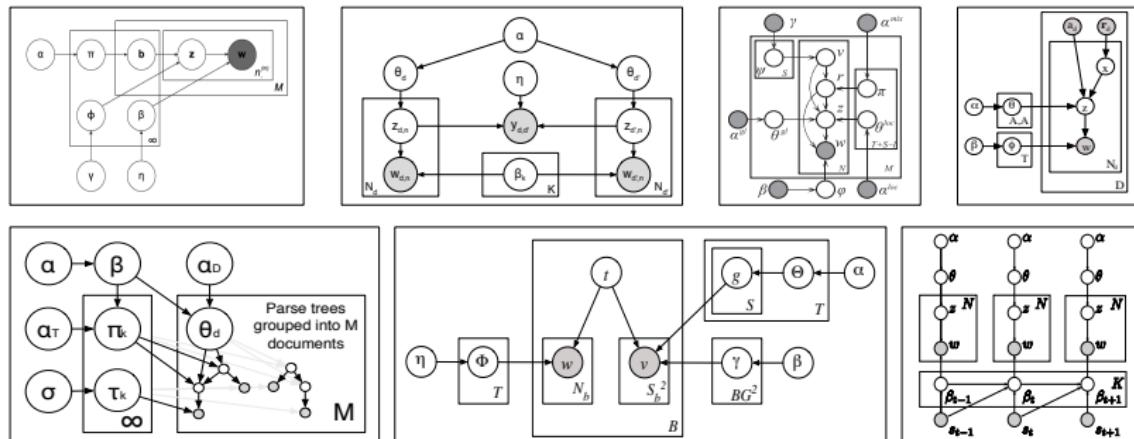
- Word probabilities are maximized by dividing the words among the topics.  
(More terms means more mass to be spread around.)
- In a mixture, this is enough to find clusters of co-occurring words.
- In LDA, the Dirichlet on the topic proportions can encourage sparsity, i.e., a document is penalized 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 summary



- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a posterior inference problem.
- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.
- Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999)  
It is mixed membership model (Erosheva, 2004).  
It relates to PCA and matrix factorization (Jakulin and Buntine, 2002)  
Was independently invented for genetics (Pritchard et al., 2000)

# LDA summary



- Organizing and finding patterns in data has become important in the sciences, humanities, industry, and culture.
- LDA can be embedded in more complicated models that capture richer assumptions about the data.
- Algorithmic improvements let us fit models to massive data.

## Example: LDA in R (Jonathan Chang)

perspective identifying tumor suppressor genes in human...  
letters global warming report leslie roberts article global....  
research news a small revolution gets under way the 1990s....  
a continuing series the reign of trial and error draws to a close...  
making deep earthquakes in the laboratory lab experimenters...  
quick fix for freeways thanks to a team of fast working...  
feathers fly in grouse population dispute researchers...

....



245 1897:1 1467:1 1351:1 731:2 800:5 682:1 315:6 3668:1 14:1  
260 4261:2 518:1 271:6 2734:1 2662:1 2432:1 683:2 1631:7  
279 2724:1 107:3 518:1 141:3 3208:1 32:1 2444:1 182:1 250:1  
266 2552:1 1993:1 116:1 539:1 1630:1 855:1 1422:1 182:3 2432:1  
233 1372:1 1351:1 261:1 501:1 1938:1 32:1 14:1 4067:1 98:2  
148 4384:1 1339:1 32:1 4107:1 2300:1 229:1 529:1 521:1 2231:1  
193 569:1 3617:1 3781:2 14:1 98:1 3596:1 3037:1 1482:12 665:2

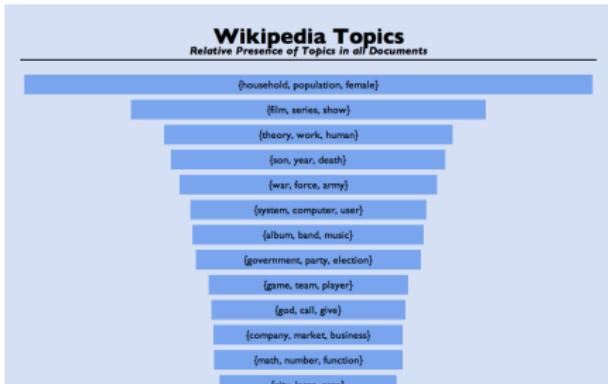
....



```
docs <- read.documents("mult.dat")
K <- 20
alpha <- 1/20
eta <- 0.001
model <- lda.collapsed.gibbs.sampler(documents, K, vocab, 1000, alpha, eta)
```

1 dna gene sequence genes sequences human genome genetic analysis two	2 protein cell cells proteins receptor fig binding activity activation kinase	3 water climate atmospheric temperature global surface ocean carbon atmosphere changes	4 says researchers new university just science like work first years	5 mantle high earth pressure seismic crust temperature earths lower earthquakes
6 end article start science readers service news card circle letters	7 time data two model fig system number different results etc	8 materials surface high structure temperature molecules chemical molecular fig university	9 dna rna transcription protein site binding sequence proteins specific sequences	10 disease cancer patients human gene medical studies drug normal drugs
11 years million ago age university north early fig evidence record	12 species evolution population evolutionary university populations natural studies genetic biology	13 protein structure proteins two amino binding acid residues molecular structural	14 cells cell virus hiv infection immune human antigen infected viral	15 space solar observations earth stars university mass sun astronomers telescope
16 fax manager science aaas advertising sales member recruitment associate washington	17 cells cell gene genes expression development mutant mice fig biology	18 electron state light quantum physics electrons high laser magnetic	19 research science national scientific scientists new states university united health	20 neurons brain cells activity fig channels university cortex neuronal visual

# Open source document browser (with Allison Chaney)



### {film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}

### Stanley Kubrick

related topics

- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {black, white, people}
- {god, call, give}
- {math, energy, light}

**Stanley Kubrick** (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow and methodical approach to his craft, his uncompromising standards, his technical perfectionism, and his reticulateness about his films and personal life. He worked for beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschew structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.<sup>[1]</sup> A recurring theme in his films is man's inhumanity to man. While often viewed as

related documents

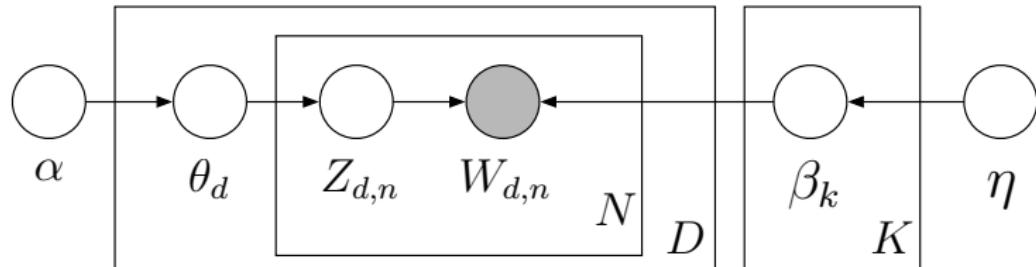
- Orson Welles
- B movie
- Mystery Science Theater 3000
- Monty Python
- Doctor Who
- Sam Peckinpah
- The A-Team
- Pulp Fiction (film)
- Buffy the Vampire Slayer (TV series)
- The X-Files
- Sunset Boulevard (film)
- Jack Palance

### {theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

# **Beyond Latent Dirichlet Allocation**

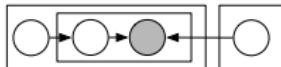
# Extending LDA



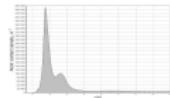
- LDA is a simple topic model
- Can be used to find topics that describe a corpus
- Each document exhibits multiple topics
- How can we build on this simple model of text?

# Extending LDA

**Make assumptions**



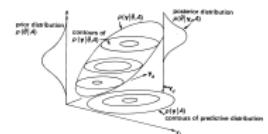
**Infer the posterior**



**Collect data**



**Check**



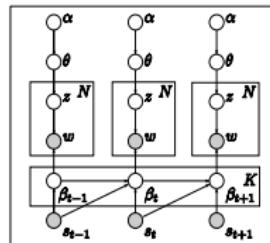
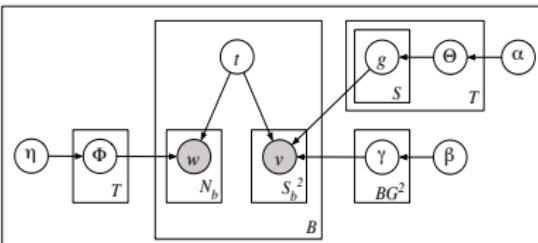
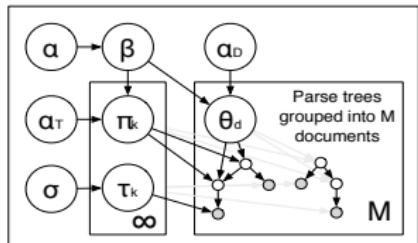
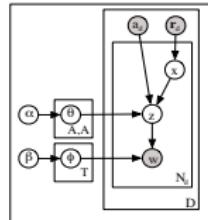
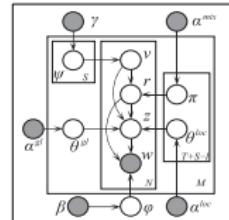
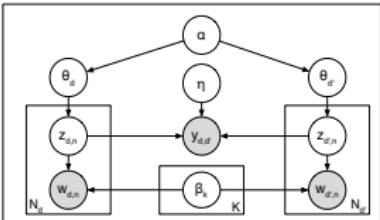
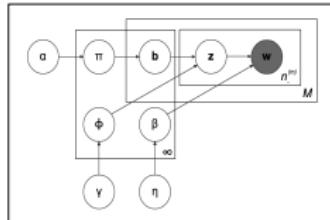
**Predict**



**Explore**

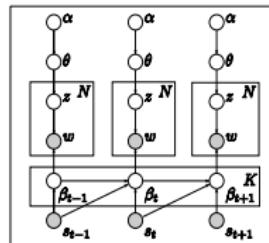
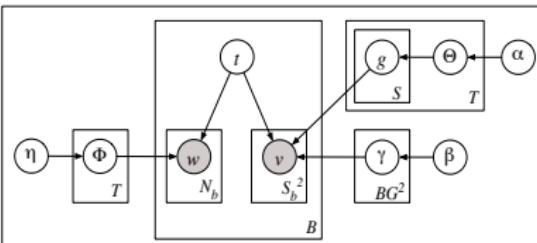
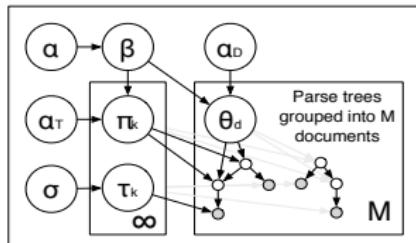
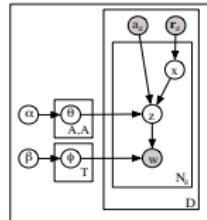
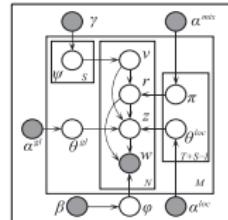
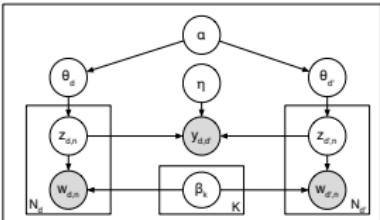
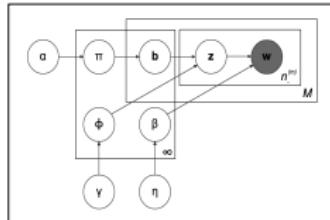


# Extending LDA



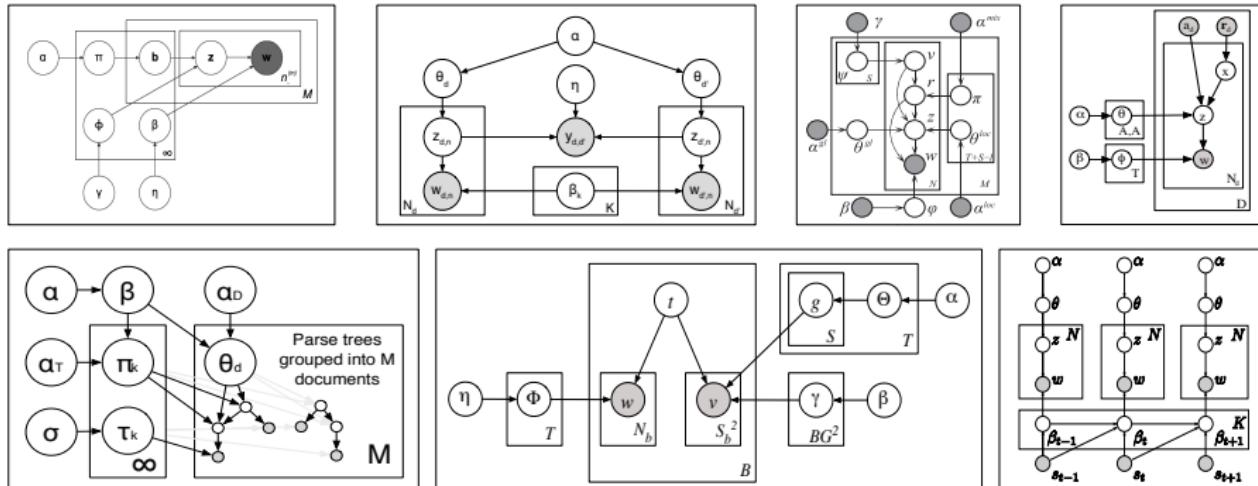
- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
- E.g., used in models that also account for syntax, authorship, word sense, dynamics, correlation, hierarchies, ...

# Extending LDA



- The **data generating distribution** can be changed, allowing us to apply mixed-membership assumptions to many kinds of data.
- E.g., can be adapted to images, social networks, music, purchase histories, computer code, genetic data, click-through-data, neural spike trains, ...

# Extending LDA

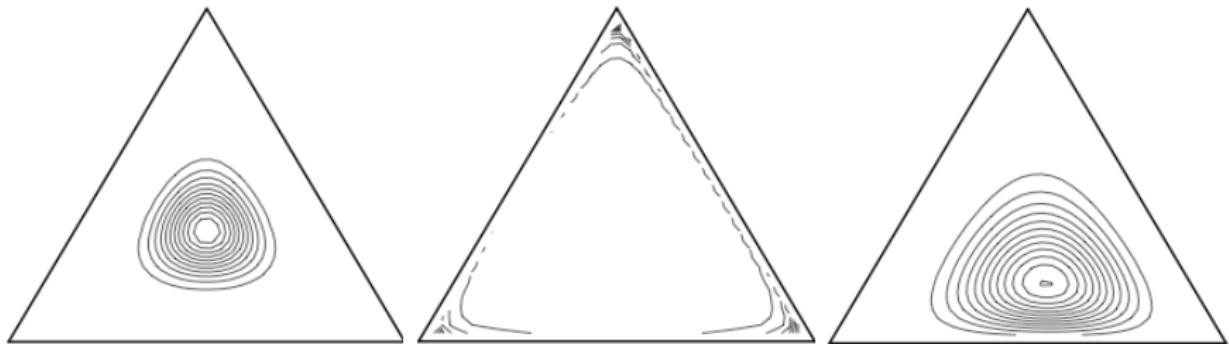


- The **posterior** can be used in creative ways.
- E.g., for IR, recommendation, document similarity, visualization, ...
- (For now, we will assume that we can compute the posterior.)

# Extending LDA

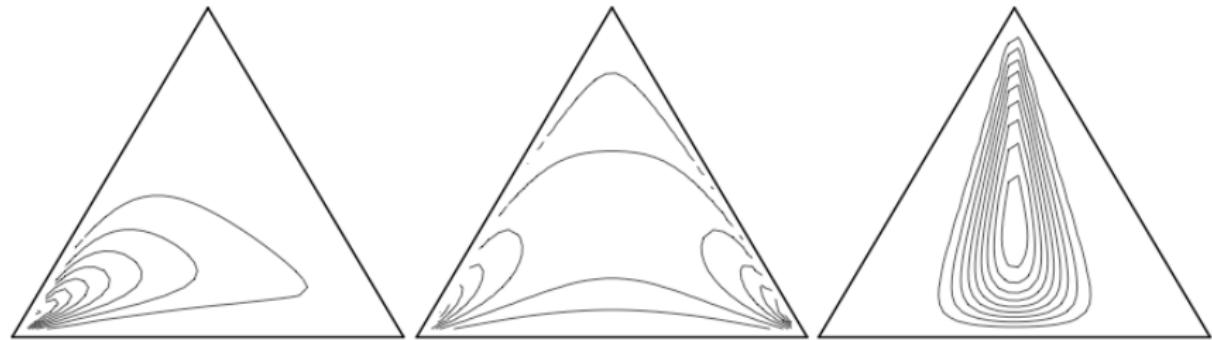
- These different kinds of extensions can be combined.
- (Really, these ways of extending LDA are a big advantage of using **probabilistic modeling** to analyze data.)
- To give a sense of how LDA can be extended, I'll describe several examples of extensions that my group has worked on.
- In this section we will discuss
  - **Correlated topic models**
  - **Dynamic topic models & measuring scholarly impact**
  - **Supervised topic models**
  - **Relational topic models**
  - **Ideal point topic models**

## Correlated topic models



- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

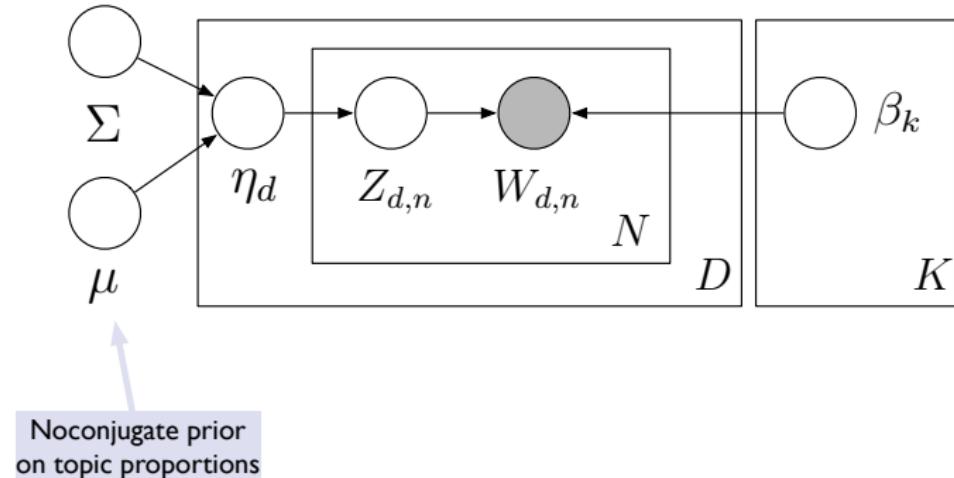
# Correlated topic models



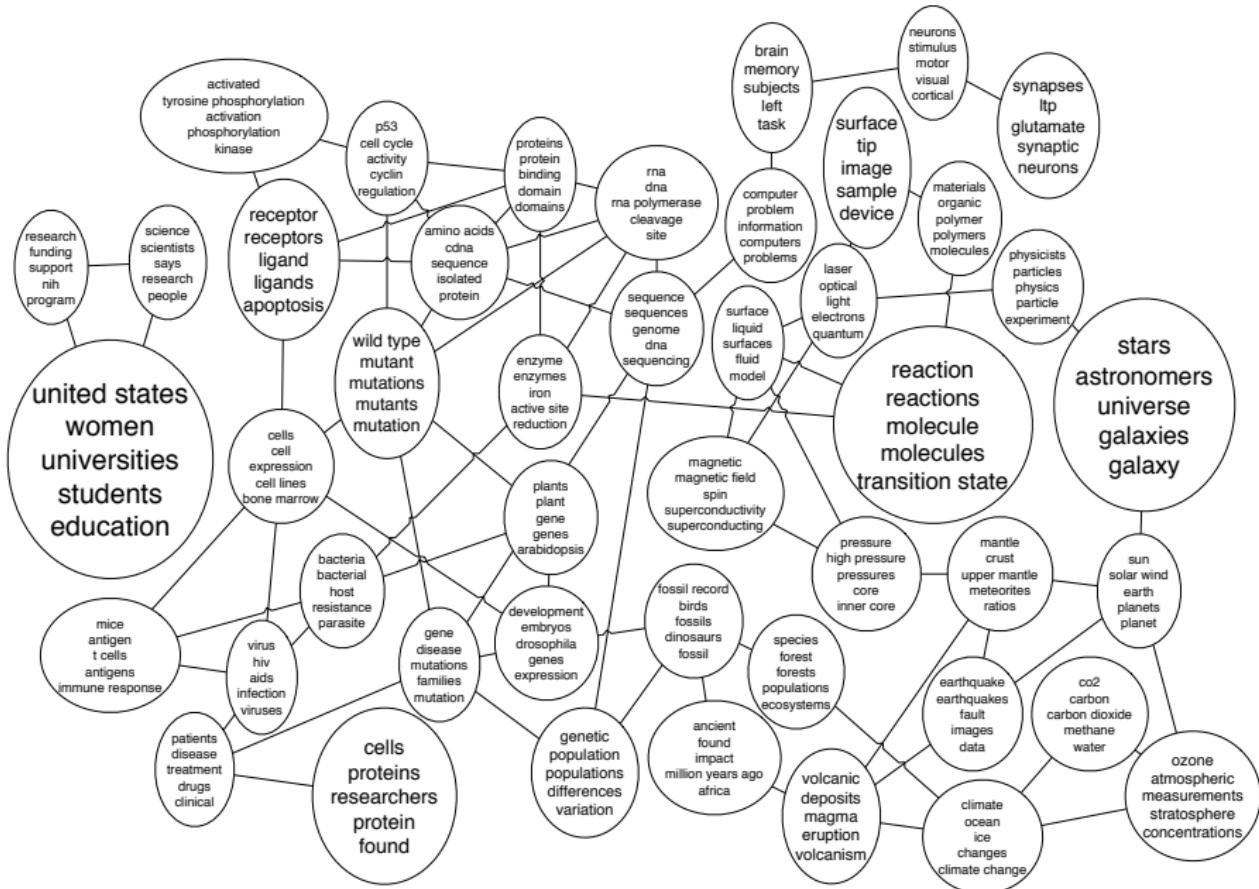
- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$\begin{aligned} X &\sim \mathcal{N}_{K-1}(\mu, \Sigma) \\ \theta_i &\propto \exp\{x_i\}. \end{aligned}$$

# Correlated topic models



- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a “map” of topics and how they are related
- Provides a better fit to text data, but is more complex to compute with



# Dynamic topic models

1789



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

2009

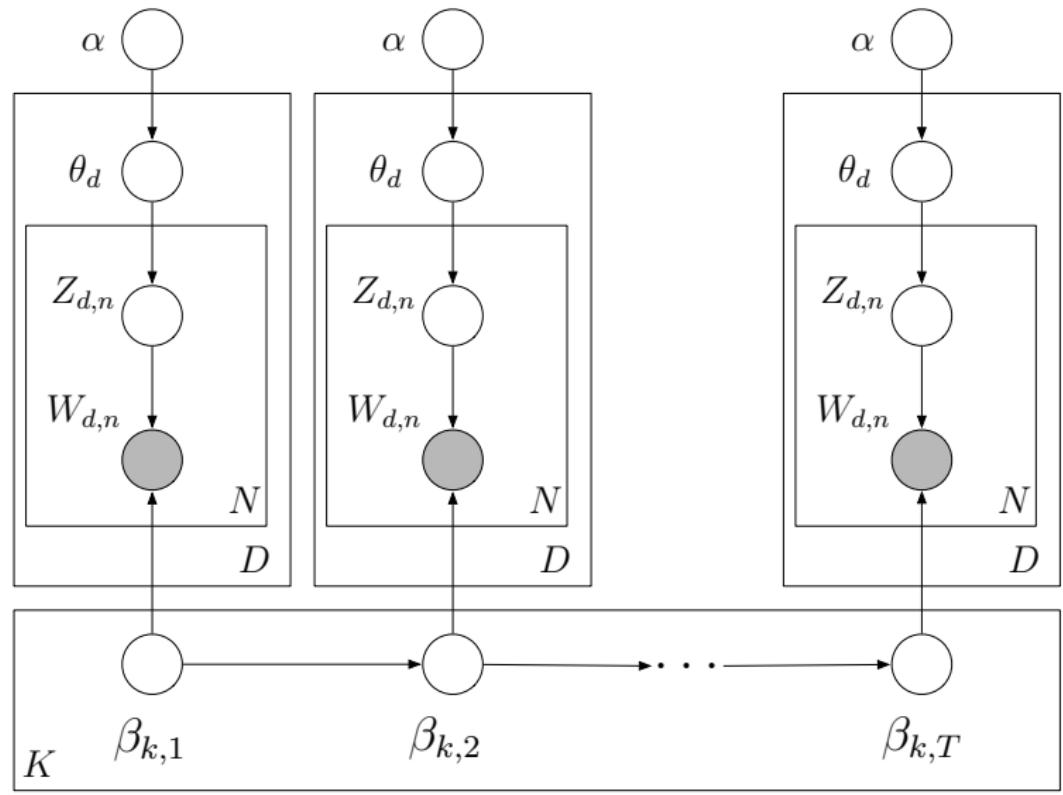


*Inaugural addresses*

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

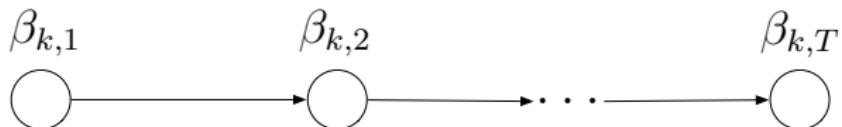
- LDA assumes that the order of documents does not matter.
- Not appropriate for corpora that span hundreds of years
- We may want to track how language changes over time.

# Dynamic topic models



Topics drifting in time →

# Dynamic topic models



- Use a logistic normal distribution to model topics evolving over time.
- Embed it in a state-space model on the log of the topic distribution

$$\begin{aligned}\beta_{t,k} | \beta_{t-1,k} &\sim \mathcal{N}(\beta_{t-1,k}, l\sigma^2) \\ p(w | \beta_{t,k}) &\propto \exp\{\beta_{t,k}\}\end{aligned}$$

- As for CTMs, this makes computation more complex. But it lets us make inferences about sequences of documents.

# Dynamic topic models

## Original article

## Topic proportions



TECHVIEW: DNA SEQUENCING

### Sequencing the Genome, Fast

James C. Pohlenz and Amanda A. McIver

Genomic sequencing projects reveal the sequence of the human genome. The reading off of the sequence of the DNA bases, which encode all of the information in our genome, is a complex process. The base sequence contains four nucleotides—adenine, thymine, guanine, and cytosine—which are linked together along the double helical chain. Over the last two decades, the cost of sequencing has gone down exponentially. The cost of sequencing has made the process of obtaining the sequence of the human genome easier. By application of an electric field across a gel, linear extraction of a fluoresced dye from the gel, and detection of the signal, the molecule yields a base-specific signal that can be automatically recorded.

The latest technology to be launched is Parkin-Elmer's much-anticipated ABI 3700 automated sequencer. The ABI 3700 is a Molecular Dynamics MagiCycle 1000 automated sequencer. It is designed to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Each sequencing reaction is contained in its own tube, which is inserted into a cartridge of because Craig Venter of Celera Genomics and his team have developed one of these cartridges (1) will hold the cartridge to produce one sequence for the computer to read. The cost of the cartridge is \$3. The specifications of the ABI 3700 are as follows: 96 samples per hour of human fiber per day; 3.3 nm sequence length per cycle; 100 fmol of each sample gives an average of 400 base pairs (bp) of double sequence data over 100 cycles. The average sequence for each human genome is covered by an average of 100,000 samples. The ABI 3700, with 17,750 reaction samples, cost Celera about \$10 million. The ABI 3700 takes 3700 machine days. With ~20 machines, the work out to less than 2 years or about 1000 human genomes. This is a major development for unanticipated developments.

At the same time, the Wellcome Trust Sanger Centre in December 1998—also in our Research and Development department for evaluation—launched the ABI 3700 and has had it added to its total capacity to reach our goal. The ABI 3700 is a very large piece of equipment in a floor-standing cabinet, which contains a sequencing gel chamber, which contains the sequencing gel, and a robotic arm for its operation. The reagent containers are not accessible for replacement, which makes the ABI 3700 a very high cost piece of equipment. At bench height within the ABI 3700 is a robotic arm that holds the carrier plates of DNA samples are located. The sequencing plates are prepared and sequenced by the use of reagents and programs it by using a personal computer. A robotic arm transfers DNA sam-

ple from the plates into wells that open into the capillaries. This is the rest of the sequencing process. The reaction can currently process four wells at a time, which takes approximately 10 hours before sequencing is required. This rate fully utilizes the ABI 3700, which has a total of 96 well plates at 12 hours.

The ABI 3700 is the first of the ABI 3700s in the use of a single flow fluorescence detection system. The ABI 3700 is the largest amount of memory so far (1). We are aiming to sequence 1 Gb of human DNA in 3 years. The sequencing gel is 96 wells with a finished version by 2005. Our sequencing gel is 96 wells with a finished version ABI 3700XL, and 31 ABI 3730-96 slab gel sequencer from Perkin Elmer (2). The ABI 3700 is a single flow fluorescence detector. The ABI 3700 uses a CTD (capillary tube detector) that simultaneously scans with all of the samples. The emitted fluorescence is detected by a CTD, which is a single flow detector. This arrangement allows for a single flow fluorescence detection system, other than a shelter in these boxes.

We have evaluated these machines for their performance, especially in comparison to the many other sequencing machines. In addition, we have evaluated the methods for sequencing the gel samples. Due to its polymerase chain reaction (PCR) and sequencing gel, the ABI 3700-96 slab gel machine. The capillary sequencing gel is a standard sequencing gel. The ABI 3700-96 slab gel machine has two sets of wells and four sets with All Big Dye Sequencing Reagents. The ABI 3700-96 slab gel machine has a higher Q value than the ABI 3700 (Q = 20), the "plate." Q value was calculated to be 1.00. The ABI 3700 has a Q value equal to 1.00 (Q = 20), the "plate." Q value was calculated to be 1.00.

With either type of system, the aim is to read as many bases as possible. The ABI 3700 is able to read twice as many bases but at half the cost of the ABI 3700-96 slab gel machine. Both systems cost the same. This is because sequencing relatively fewer longer sequences is more efficient than sequencing many short ones. So, total length is not the best measure of efficiency of new sequencing technologies.

We have recently completed the ABI 3700-96 slab gel and ABI 3730-96 slab gel sequencing by evaluating the sequence data from the ABI 3700-96 slab gel and ABI 3730-96 slab gel sequencing. These samples were submitted to the Wellcome Trust Sanger Centre. The sequencing plates were prepared and sequenced with the standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

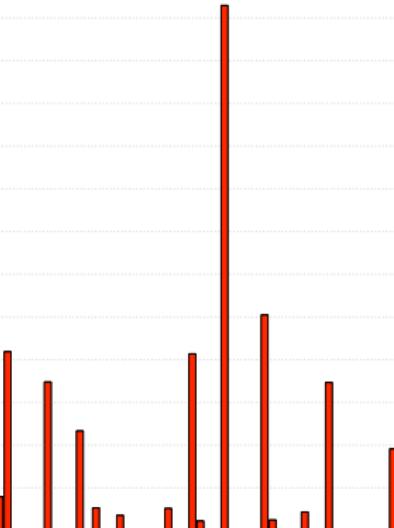


Fig. 3. Comparison of read length histograms for sequencing on the ABI 3700-96 slab gel machine and the ABI 3730-96 slab gel machine. The capillary sequencing gel is a standard sequencing gel. The ABI 3700-96 slab gel machine has two sets of wells and four sets with All Big Dye Sequencing Reagents. The ABI 3700-96 slab gel machine has a higher Q value than the ABI 3700 (Q = 20), the "plate." Q value was calculated to be 1.00. The ABI 3700 has a Q value equal to 1.00 (Q = 20), the "plate." Q value was calculated to be 1.00.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1HG, U.K. E-mail: jcpohlenz@tig.ac.uk

www.sciencemag.org SCIENCE VOL 283 9 MARCH 1999

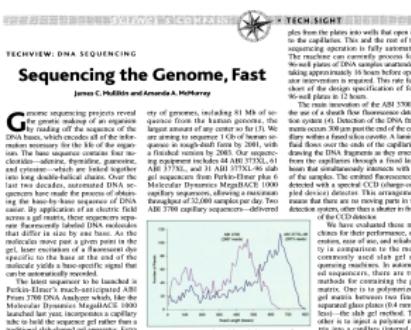
1857



# Dynamic topic models

## Original article

## Most likely words from top topics



**Sequencing the Genome, Fast**  
James C. Pohle and Amanda A. McRae  
**G**enome sequencing projects reveal the sequence of the segments of the DNA base, which encodes all of the information contained in the genome. The base sequence contains four nucleotides—adenine, thymine, guanine, and cytosine—which are linked together along the phosphate backbone. Over the last two decades, the cost of genome sequencing has made the process of obtaining a genome sequence more feasible. For example, its application of an electric field sequencing technology, which uses fluorescently labeled DNA molecules that differ at one base point, the method yields a base-specific signal that can be easily read.

The latest project to be launched is PerkinElmer's much-awaited ABI 3770, which is based on the Molecular Dynamics Magenta CE 1000 laboratory system. The ABI 3770 is designed to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Each sequencing reaction is contained within a cartridge, which is inserted into the ABI 3770. Because Cross-Vent of Celera sequencing is the standard for most of these cartridges (7), it will enable the company to produce new sequences for the customer by the end of the year, within 3 years. The specifications of the ABI 3770 are impressive: 100 samples per hour of human liver per day, 3 nm resolution, and a sequencing rate of 100 bp for each sample giving an average of 400 base pairs (bp) of usable sequence data from each sample. The cost of sequencing a human genome is covered by an average of 100 samples per run, which is equivalent to 12,750 reactions. The ABI 3770 can sequence up to 75 million samples. But Celera also claims that the ABI 3770 can sequence 16 Mb of genomic sequence in just 2700 machine days. With >20 machines, the work out to less than 2 years or about \$100 million per genome. This is a remarkable achievement for a company that started from scratch.

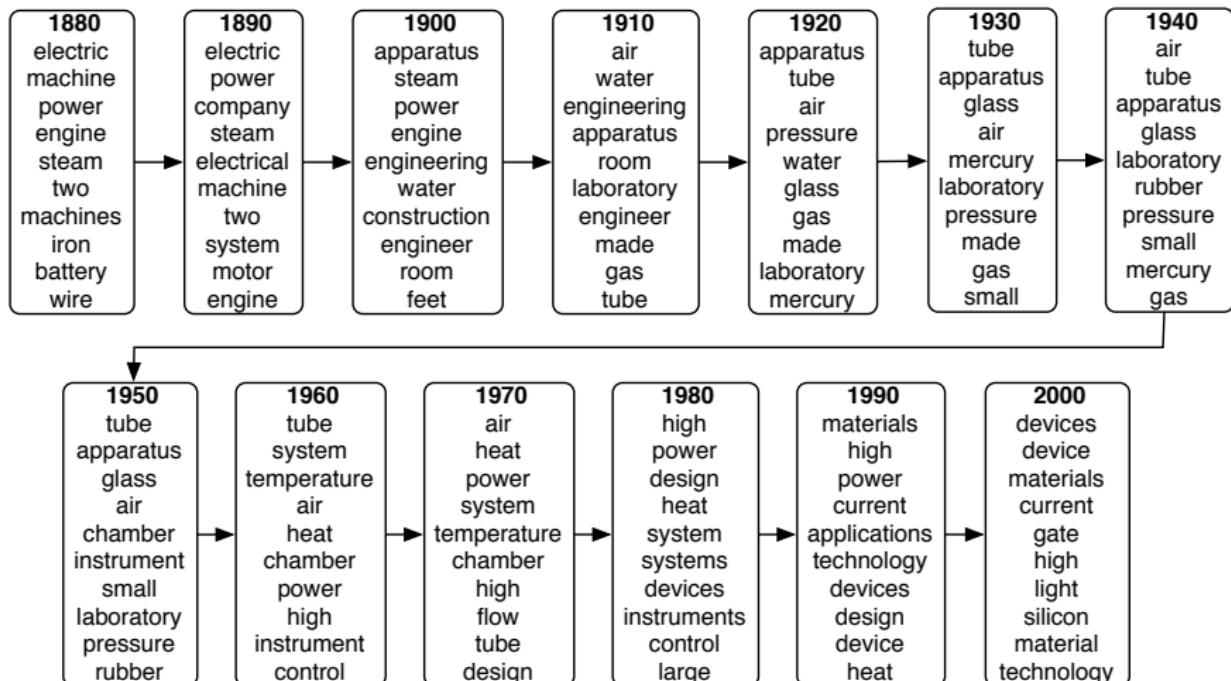
The authors thank The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1HG, U.K. E-mail: jcpohle@perkinelmer.com

sequence  
genome  
genes  
sequences  
human  
gene  
dna  
sequencing  
chromosome  
regions  
analysis  
data  
genomic  
number

devices  
device  
materials  
current  
high  
gate  
light  
silicon  
material  
technology  
electrical  
fiber  
power  
based

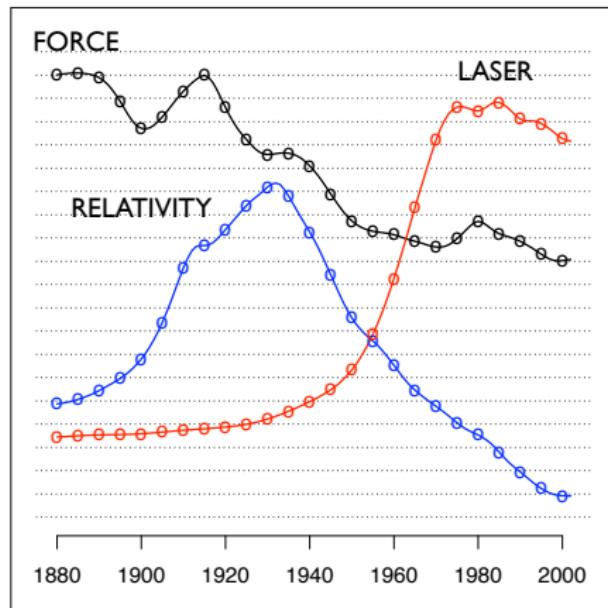
data  
information  
network  
web  
computer  
language  
networks  
time  
software  
system  
words  
algorithm  
number  
internet

# Dynamic topic models

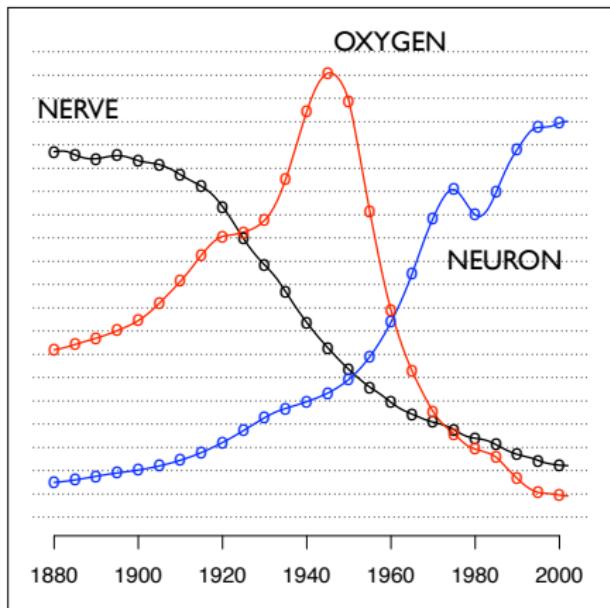


# Dynamic topic models

"Theoretical Physics"



"Neuroscience"



# Dynamic topic models

- **Time-corrected similarity** shows a new way of using the posterior.
- Consider the expected Hellinger distance between the topic proportions of two documents,

$$d_{ij} = E \left[ \sum_{k=1}^K (\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 | \mathbf{w}_i, \mathbf{w}_j \right]$$

- Uses the latent structure to define similarity
- Time has been factored out because the topics associated to the components are different from year to year.
- Similarity based only on topic proportions

## Dynamic topic models

## The Brain of the Orang (1880)

20

SEARCH

*Triflow* in these cases, which were submitted to the authors on the 4th of December last for correction or rejection: no objection being made we printed them in a second number. After publication Professor Agassiz now writes that the reports under his name are not satisfactory

Professor George F. Barker, Professor D. C. Marsh and Professor J. E. Blizard are preparing more elaborate reports of their important papers, and promise them at an early date.

THE MEAN OF THE SAMPLE

The brain of the orang has been figured by Tiedemann, Sandflit, Schroeder van der Kolk and Vossius, Gessels, Kellgren, etc. On account, however, of the few illustrations existing, and of the importance of the subject, I avail myself of the opportunity of presenting several views of my orang's brain (Figs. 1 to 5), which was removed from the skull only a few hours ago.

death. The membranes were in a high state of congestion, and a little of the surface of the left hemisphere had been disorganized by disease; otherwise the brain was in good condition. It weighed exactly ten ounces. The brain of the orang in its present condition resembled that of man more than those of either of the chimpanzees which I examined. In these the brain was more elongated. The general character of the folds and fissures in



the brain of the Grangier, Chalopin, and man are the same; there are certain slight differences, however, in their disposition in all three. The fissure of Sylvius in the Grangier runs up and down the posterior branch passing through the anterior limb of the latter, the anterior branch is small. The fissure of Roland, or central fissure, quite appears, is, however, situated slightly more toward the Grangier than in man. It disseminates the anterior limb of the posterior branch. The anterior pole is well marked; bounded externally by the first insular fillet it extends internally on the mental side of the bridging, separating the parietal from the occipital lobes.

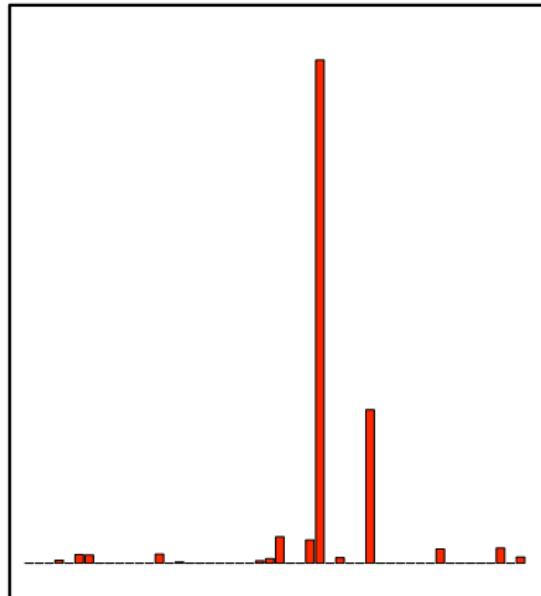
In the Ossang, the postnasalopatinal fissure does not in the caecilians, being separated from it by the "deglipidis" or passage between the "deglipidis" or "anterior" and "posterior" openings of the nostrils. There is no separation as in amphisbaenians than once in was.

According to Sheldan, this dispersion is observed in the *Homaloptera*, whereas, in the *Aeolidae* and *Platynoididae* it is limited to the *Homaloptera*. The latter, as in man, is the result of a failure of the heart to pump the blood from the peripheral veins. On the other hand, the heart is made to do its work by the pressure of the blood vessels. Once it is higher, wider, and more active than the normal heart, it can pump the blood from the central veins into the peripheral veins. In this way the heart can pump the blood from the veins into the arteries. The heart will be able to do this because the peripheral veins are thicker than the central veins. This is the reason that the heart is stronger in man than in other animals. The heart is also thicker in man than in other animals. The heart is also thicker in man than in other animals.



**occipital lobe;** externally it is continuous with the capital lobe, as the first occipital gyrus, anteriorly separated from the posterior central convolution or completely than is, by a fissure which runs parallel with the central fissure. There is in the Ong-ang a fissure running parallel with the parietal, which divides the upper parietal lobule into lower and upper portions. The parietal, or the square on the mesial side of the marginal lobe, between the parietal and

\* *Franziskanische Proceedings of the Academy of Natural Sciences*, Phil., 1816.



# Dynamic topic models

## Representation of the Visual Field on the Medial Wall of Occipital-Parietal Cortex in the Owl Monkey (1976)

rent, the cytoarchitectonic organization of the medial occipital-parietal cortex was explored by microelectrode recordings and techniques in five owl monkeys (2). The monkeys were anaesthetized with ketalar and propofol, and the head was held in a stereotaxic frame. Microelectrodes were used to record from the visual cortex. Recordings were made from the visual areas of each monkey from eight regions in a tangential parasagittal parallel to the medial wall of the hemisphere.

Recording fields were plotted by moving electric spots at rectangular sites and have an area of approximately 1 mm<sup>2</sup>.

The recording sites were distributed

uniformly over the medial wall of the hemisphere in front of the optic chiasm. The recording sites in the optic chiasm were also the posterior optic sphinx with the method of Fendt and Chase (3).

The ipsilateral eye usually was

covered with an opaque shield. Electrode tracks and recording sites were represented by small dots in the diagrams and photographs of the intact brain.

Figure 1 illustrates the data that we obtained from the medial wall of the hemisphere in front of the optic chiasm. In this figure, data obtained in the other four experiments are included. The results of the other experiments in the dorsomedial area, however, are not shown. The recording sites in the dorsal medial wall of the hemisphere in front of the optic chiasm are arranged in a tangential projection. Tangential projections 1 through 4 run parallel to the medial wall of the hemisphere and are located at a distance of approximately 1 mm from the medial surface. In previously published experiments, the recording sites in the visual fields recorded adjacent to the medial area of the dorsomedial area (H1) have been plotted in the lower quadrants near the horizontal meridian about 50° to 60° from the center (4). Thus, as is shown in Fig. 1, and

also in Fig. 2, which illustrates the organization of the other cortical visual areas in the owl monkey, the recording sites in the border between the medial area and the second visual area correspond to a peripheral projection of the upper visual field.

In

other experiments in the dorsomedial area, the recording sites in the medial area of the hemisphere in front of the optic chiasm were located near the horizontal meridian in the periphery toward the horizontal meridian (5). Thus, as is shown in Fig. 1, the recording sites in the border between the dorsomedial and the medial areas correspond to part of the lower field of view, and the recording sites in the periphery of the lower visual quadrant. Obviously, the medial area is adjusted by pos-

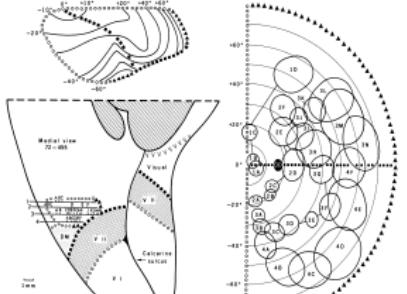
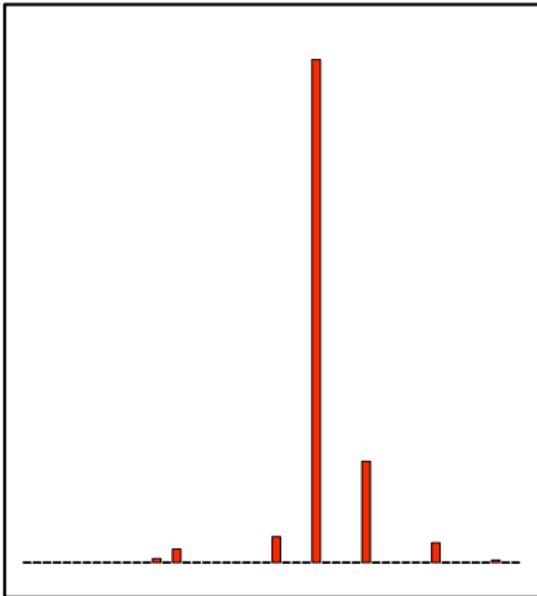
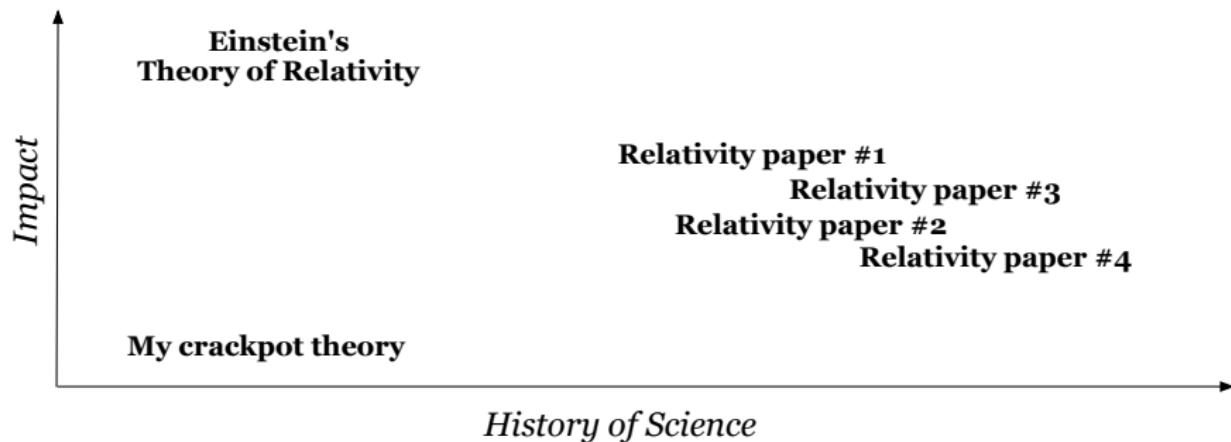


Fig. 1. Microelectrode recording sites and receptive field for the medial visual area in owl monkey 72-451. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up and dorsal is to the right. The numbers refer to the recording sites in the medial visual area. The numbers in parentheses refer to the recording sites in the corresponding receptive fields as shown in the postlateral area on the right. In the upper left is a regional map of the visual representation of the medial area of the hemisphere in front of the optic chiasm. The numbers refer to the recording sites in the medial area of the hemisphere in front of the optic chiasm. The numbers in parentheses refer to the recording sites in the second visual area. The numbers in parentheses refer to the dorsomedial visual area. DR indicates the projection of the optic disk as blind spot.

13 JUNE 1976 • J. Neurosci., 13(12):3523–3535

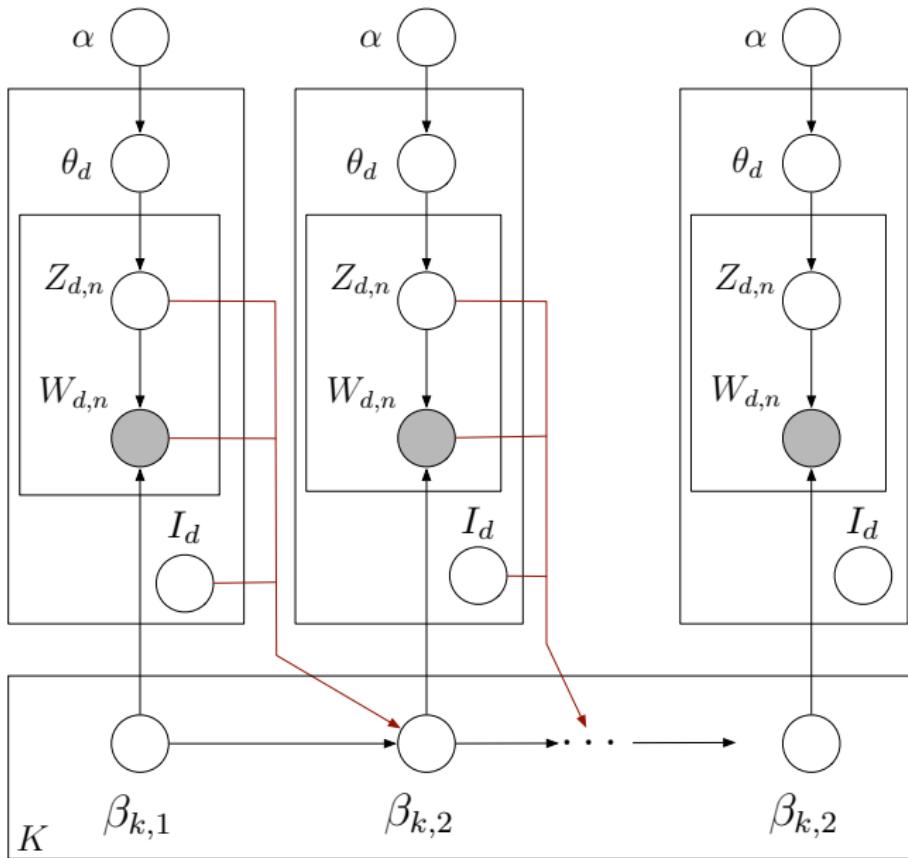


# Measuring scholarly impact

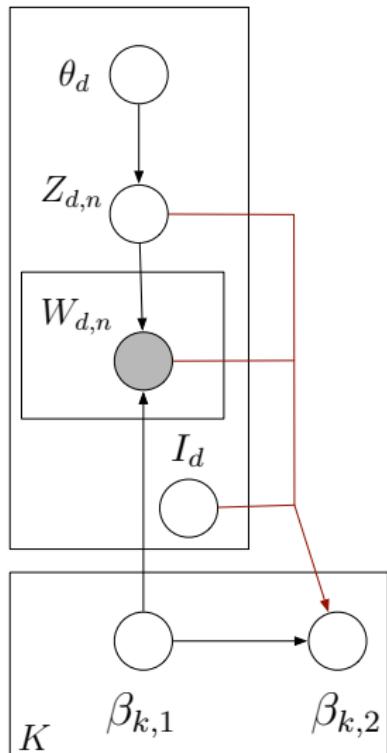


- We built on the DTM to measure **scholarly impact** with sequences of text.
- Influential articles reflect future changes in language use.
- The “influence” of an article is a latent variable.
- Influential articles affect the drift of the topics that they discuss.
- The posterior gives a retrospective estimate of influential articles.

# Measuring scholarly impact

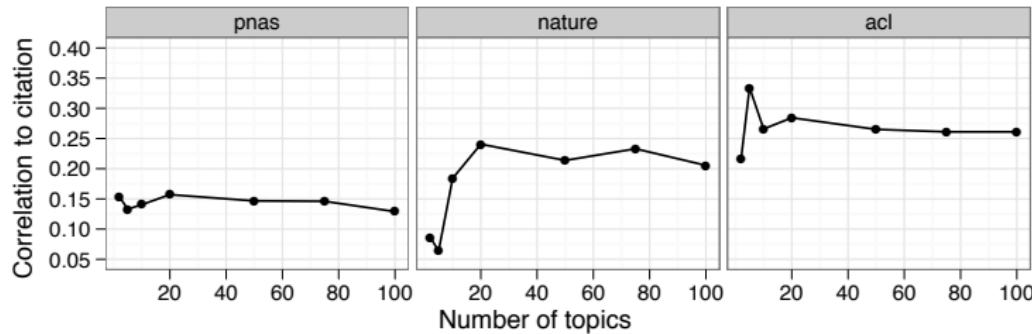


# Measuring scholarly impact



- Each document has an influence score  $I_d$ .
- Each topic drifts in a way that is biased towards the documents with high influence.
- We can examine the posterior of the influence scores to retrospectively find articles that best explain the changes in language.

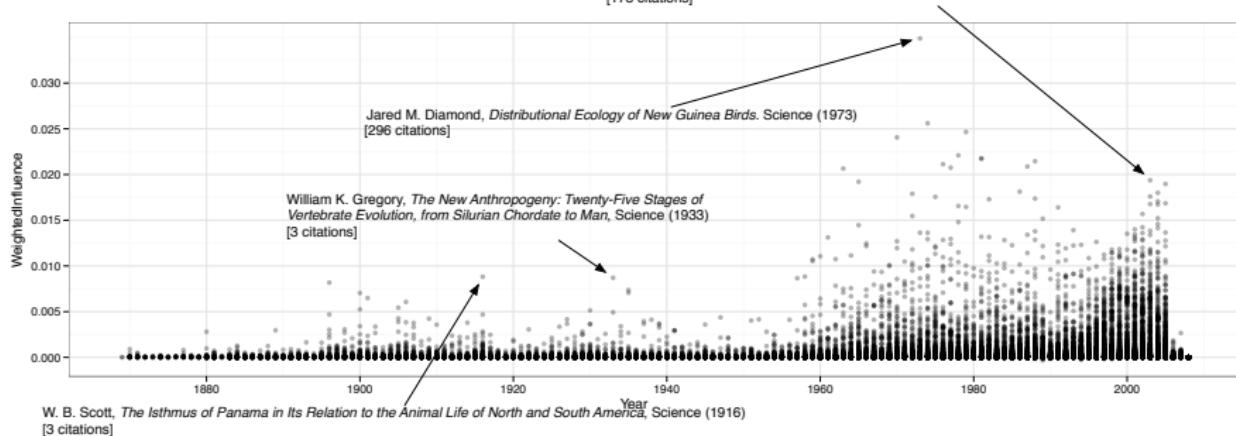
# Measuring scholarly impact



- This measure of impact only uses the words of the documents.  
It correlates strongly with citation counts.
- High impact, high citation: “The Mathematics of Statistical Machine Translation: Parameter Estimation” (Brown et al., 1993)
- “Low” impact, high citation: “Building a large annotated corpus of English: the Penn Treebank” (Marcus et al., 1993)

# Measuring scholarly impact

Derek E. Wildman et al., Implications of Natural Selection in Shaping 99.4% Nonsynonymous DNA Identity between Humans and Chimpanzees: Enlarging Genus *Homo*, PNAS (2003) [178 citations]



- PNAS, *Science*, and *Nature* from 1880–2005
- 350,000 Articles
- 163M observations
- Year-corrected correlation is 0.166

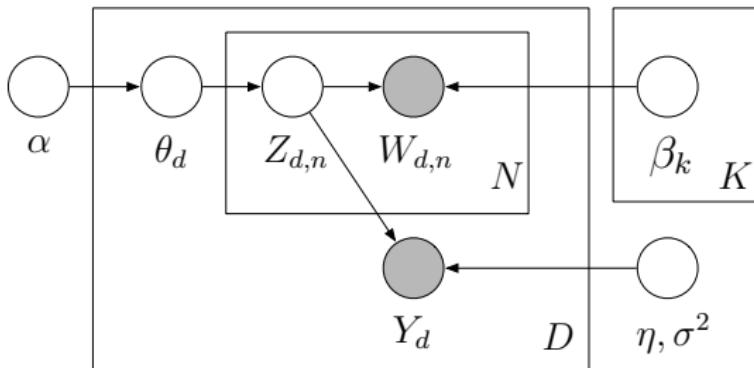
## Summary: Correlated and dynamic topic models

- The Dirichlet assumptions on topics and topic proportions makes strong conditional independence assumptions about the data.
- The **correlated topic model** uses a logistic normal on the topic proportions to find patterns in how topics tend to co-occur.
- The **dynamic topic model** uses a logistic normal in a linear dynamic model to capture how topics change over time.
- What's the catch? These models are harder to compute with. (Stay tuned.)

## Supervised LDA

- 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 “likes”
  - Documents paired with links to other documents
  - Images paired with a category
- **Supervised LDA** are topic models of documents and responses, fit to find topics predictive of the response.

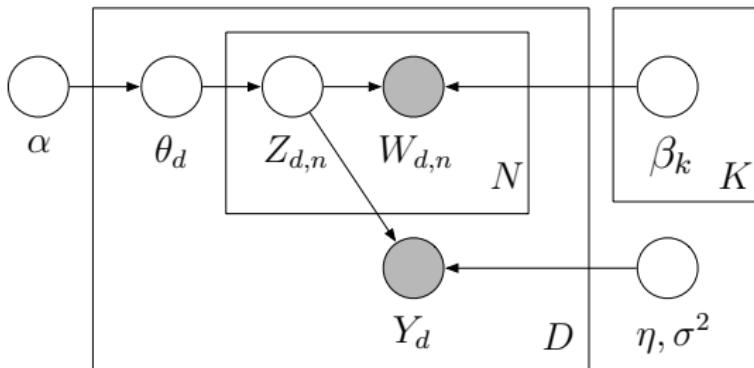
# Supervised LDA



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

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

# Supervised LDA

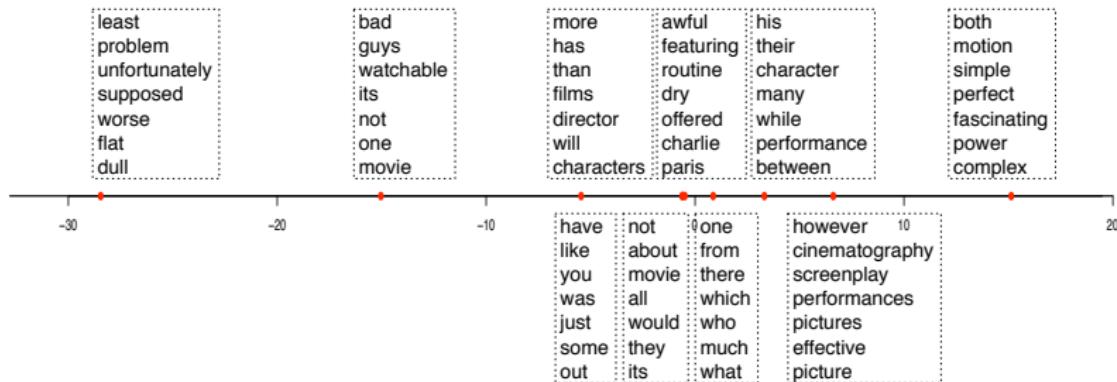


- Fit sLDA parameters to documents and responses.  
This gives: topics  $\beta_{1:K}$  and coefficients  $\eta_{1:K}$ .
- Given a new document, predict its response using the expected value:

$$E[Y|w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] = \eta^\top E[\bar{Z}|w_{1:N}]$$

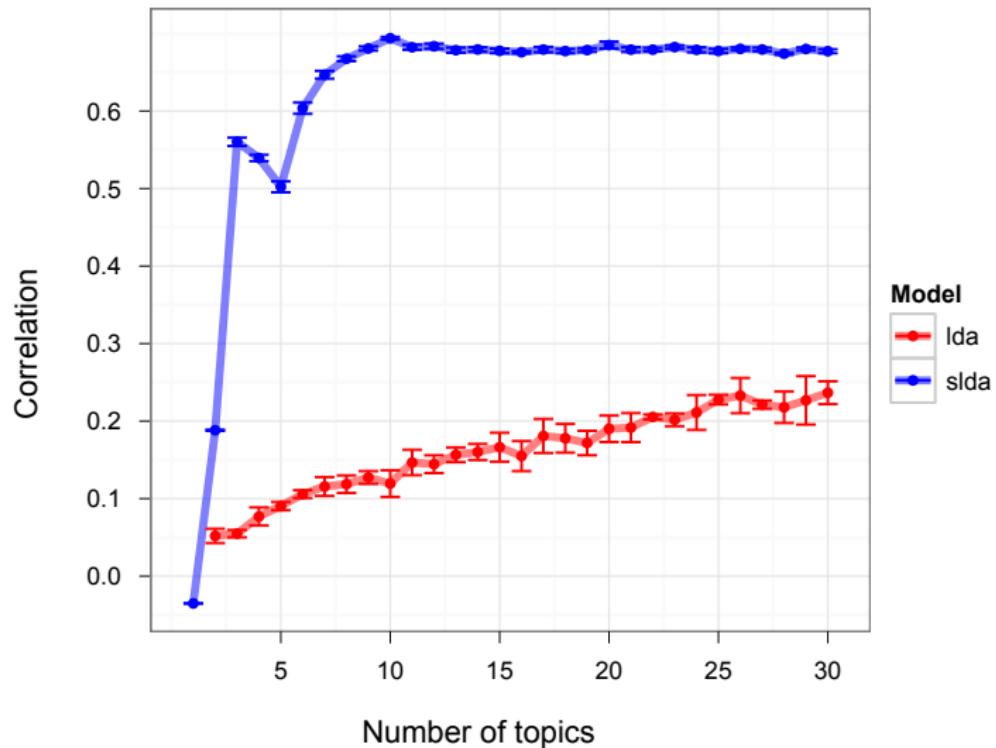
- This blends generative and discriminative modeling.

# Supervised LDA

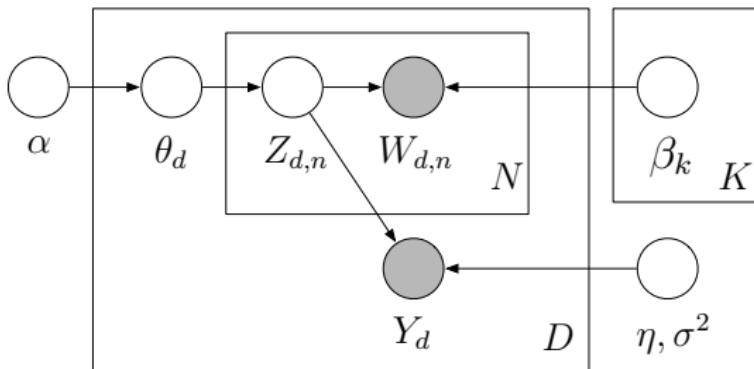


- 10-topic sLDA model on movie reviews (Pang and Lee, 2005).
- Response: number of stars associated with each review
- Each component of coefficient vector  $\eta$  is associated with a topic.

# Supervised LDA

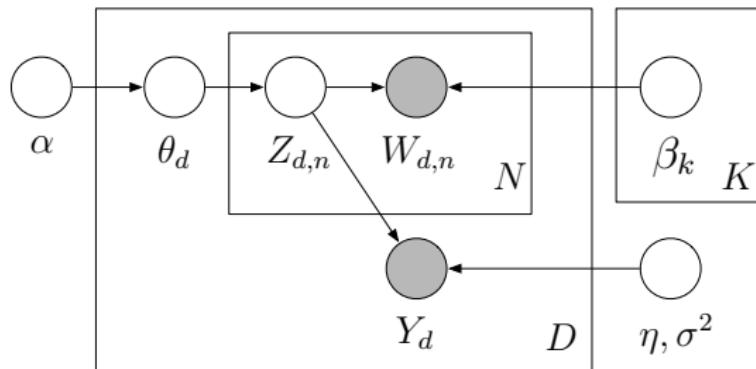


# Supervised LDA



- SLDA enables model-based regression where the predictor is a 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.
- SLDA has been extended to generalized linear models, e.g., for image classification and other non-continuous responses.

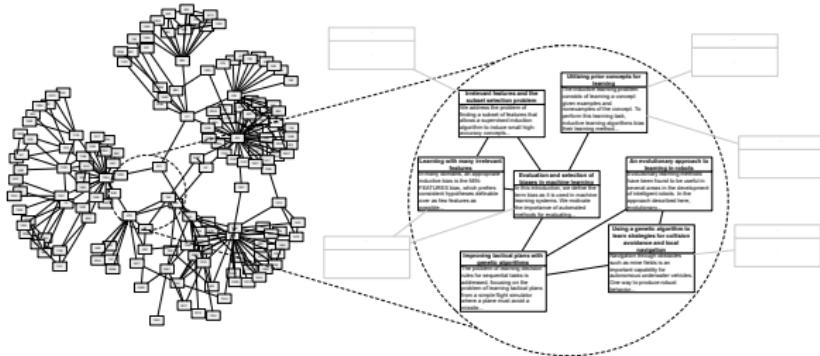
# Supervised LDA



We will discuss two extensions of sLDA

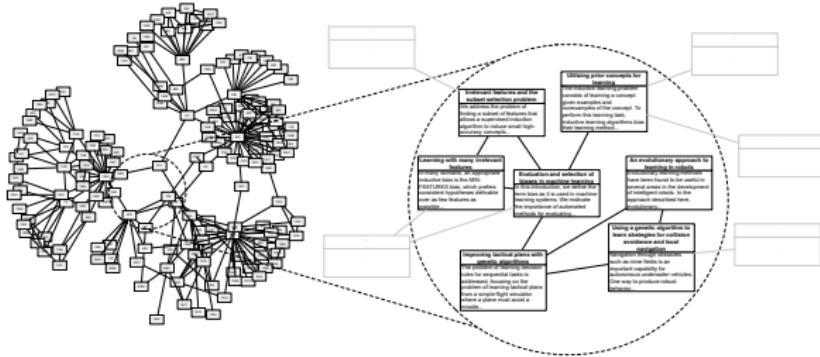
- Relational topic models
- Ideal point topic models

# Relational topic models



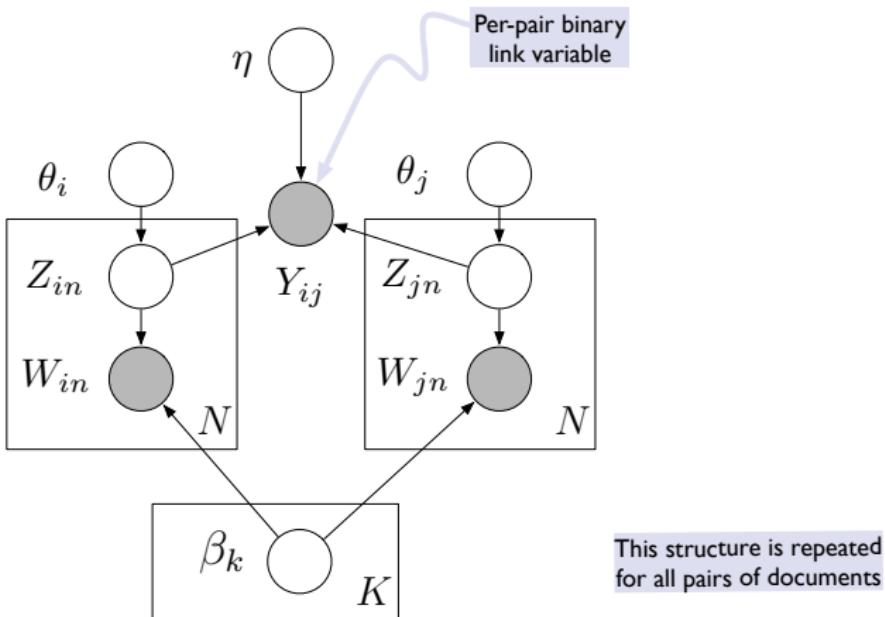
- Many data sets contain **connected observations**.
- For example:
  - Citation networks of documents
  - Hyperlinked networks of web-pages.
  - Friend-connected social network profiles

# Relational topic models



- Research has focused on finding communities and patterns in the link-structure of these networks.
- We adapted sLDA to pairwise response variables.  
This leads to a model of **content and connection**.
- Relational topic models find related hidden structure in both types of data.

# Relational topic models



- Adapt fitting algorithm for sLDA with binary GLM response
- RTMs allow predictions about new and unlinked data.
- These predictions are out of reach for traditional network models.

# Relational topic models

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

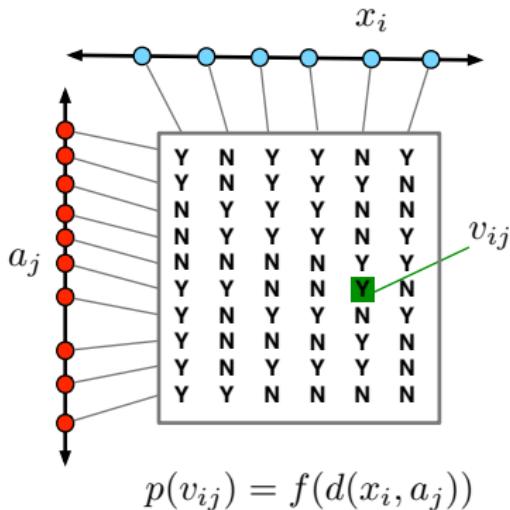
Given a new document, which documents is it likely to link to?

# Relational topic models

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

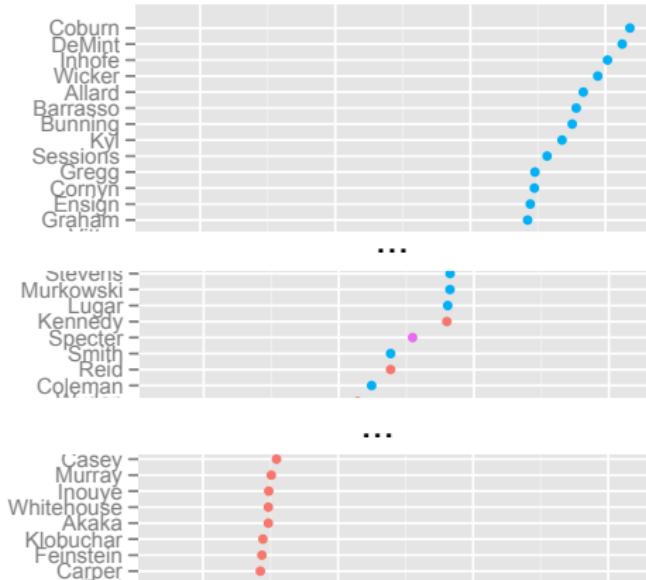
Given a new document, which documents is it likely to link to?

# Ideal point topic models



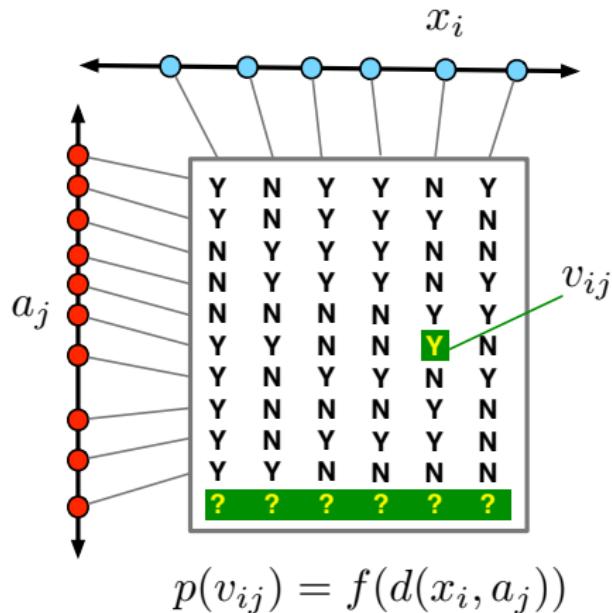
- The **ideal point model** uncovers voting patterns in legislative data
- We observe roll call data  $v_{ij}$ .
- Bills attached to discrimination parameters  $a_j$ .  
Senators attached to ideal points  $x_i$ .

# Ideal point topic models



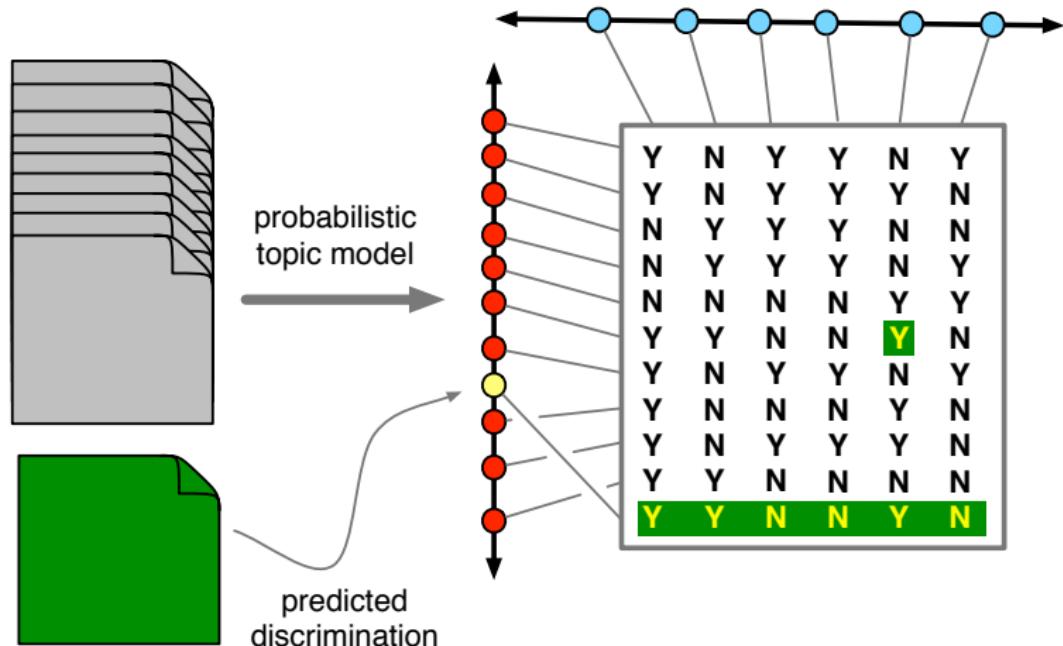
- Posterior inference reveals the political spectrum of senators
- Widely used in quantitative political science.

# Ideal point topic models



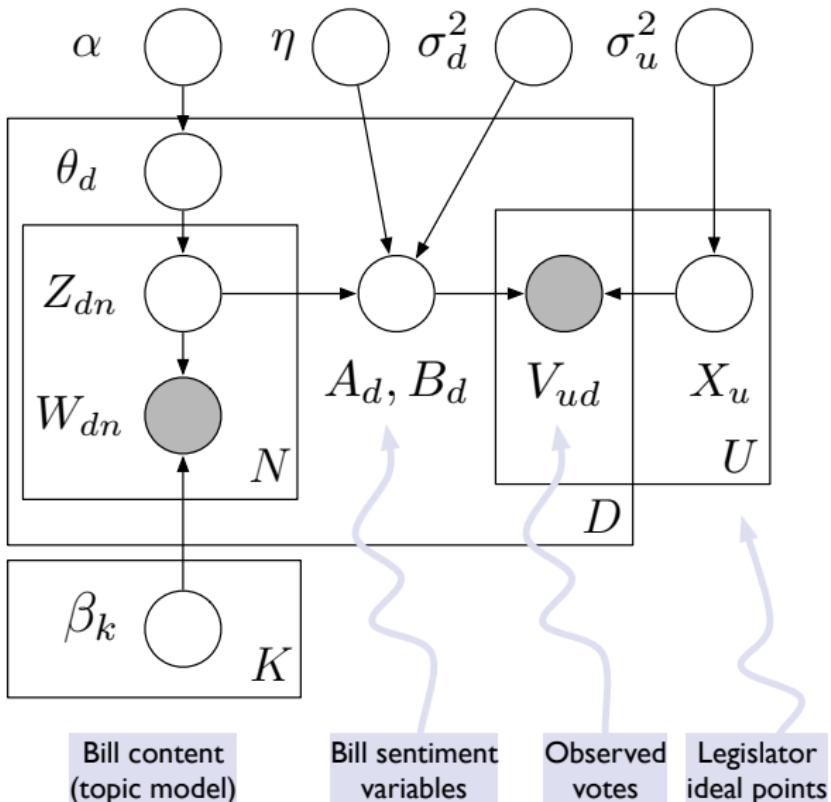
- We can predict a missing vote.
- But we cannot predict all the missing votes from a bill.
- Cf. the limitations of collaborative filtering

# Ideal point topic models

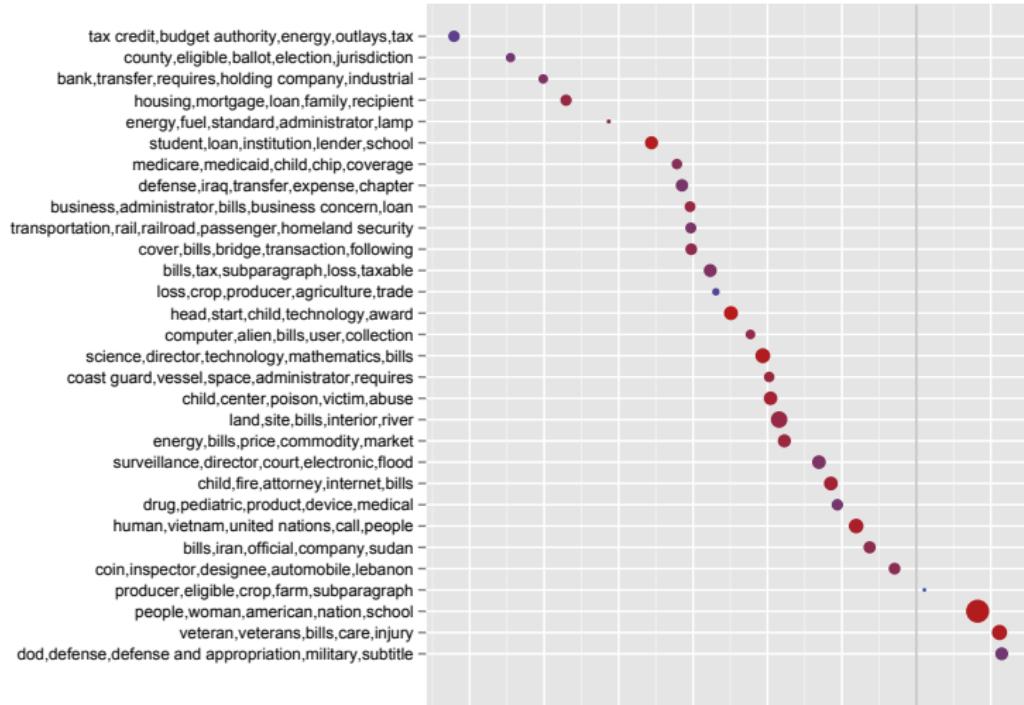


- Use supervised LDA to predict bill discrimination from bill text.
- But this is a **latent response**.

# Ideal point topic models



# Ideal point topic models



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

## Summary: Supervised topic models

- Many documents are associated with response variables.
- **Supervised LDA** embeds LDA in a generalized linear model that is conditioned on the latent topic assignments.
- **Relational topic models** use sLDA assumptions with pair-wise responses to model networks of documents.
- **Ideal point topic models** demonstrates how the response variables can themselves be latent variables. In this case, they are used downstream in a model of legislative behavior.
- (SLDA, the RTM, and others are implemented in the R package “lda.”)

# Still other ways to build on LDA

## New applications—

- Syntactic topic models
- Topic models on images
- Topic models on social network data
- Topic models on music data
- Topic models for recommendation systems

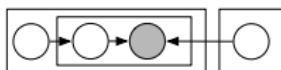
## Testing and relaxing assumptions—

- Spike and slab priors
- Models of word contagion
- N-gram topic models

# **Posterior Inference**

# Posterior inference

## Make assumptions



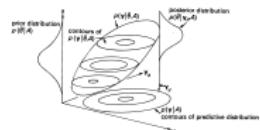
## Collect data



## Infer the posterior



## Check



## Predict



## Explore



- We can express many kinds of assumptions.
- How can we analyze the collection under those assumptions?

# Posterior inference

Topics



Documents

## 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, these predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Svante Paabo of the 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 Aracely Musigian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an-

alytical results from different researchers

and different databases

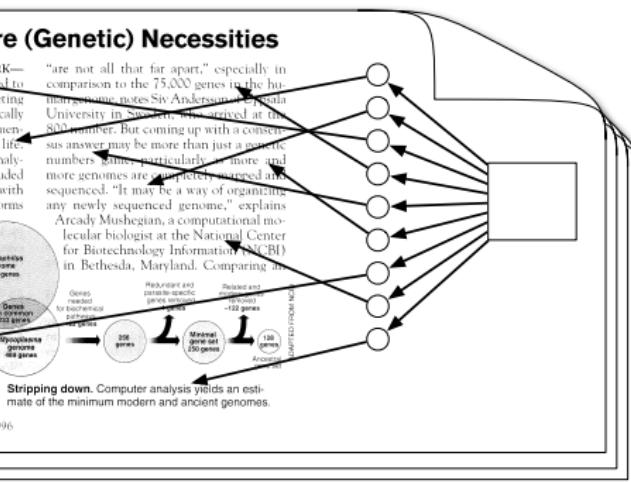
can help identify genes that are essential

for life, says Paabo.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



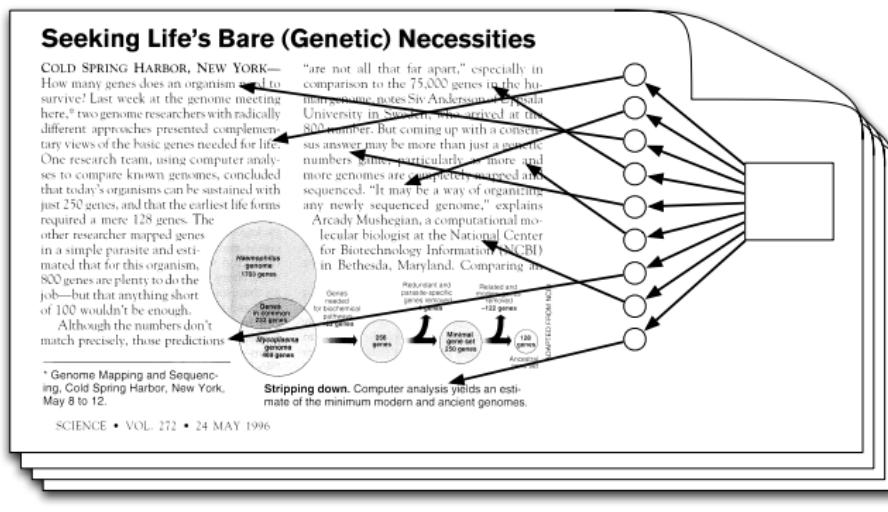
- Posterior inference is the main computational problem.
- Inference links observed data to statistical assumptions.
- Inference on large data is crucial for topic modeling applications.

# Posterior inference

## *Topics*



## *Documents*

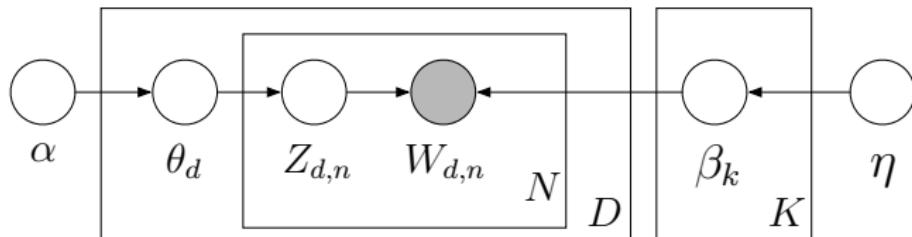


## *Topic proportions and assignments*

- Our goal is to compute the distribution of the hidden variables conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

# Posterior inference for LDA



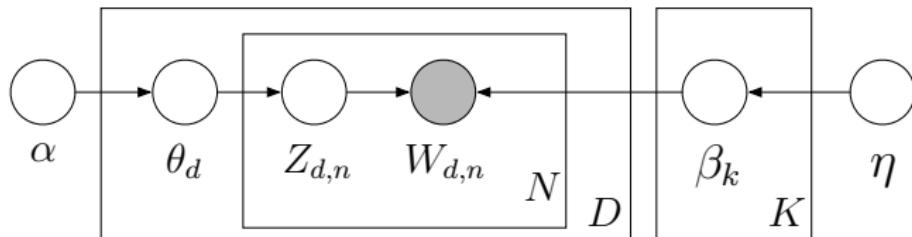
- The joint distribution of the latent variables and documents is

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

- The posterior of the latent variables given the documents is

$$p(\beta, \theta, z | w).$$

# Posterior inference for LDA

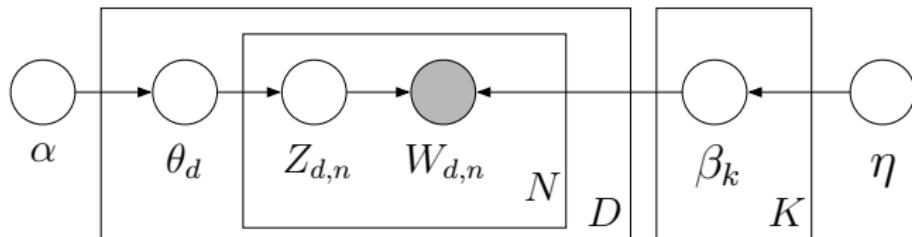


- This is equal to

$$\frac{p(\beta, \theta, \mathbf{z}, \mathbf{w})}{\int_{\beta} \int_{\theta} \sum_{\mathbf{z}} p(\beta, \theta, \mathbf{z}, \mathbf{w})}.$$

- We can't compute the denominator, the marginal  $p(\mathbf{w})$ .
- This is the crux of the inference problem.

# Posterior inference for LDA



- There is a large literature on approximating the posterior, both within topic modeling and Bayesian statistics in general.
- We will focus on **mean-field variational methods**.
- We will derive **stochastic variational inference**, a generic approximate inference method for very large data sets.

## Stochastic variational inference

- We want to condition on large data sets and approximate the posterior.
- In **variational inference**, we optimize over a family of distributions to find the member closest in KL divergence to the posterior.
- Variational inference usually results in an algorithm like this:
  - Infer local variables for each data point.
  - Based on these local inferences, re-infer global variables.
  - Repeat.

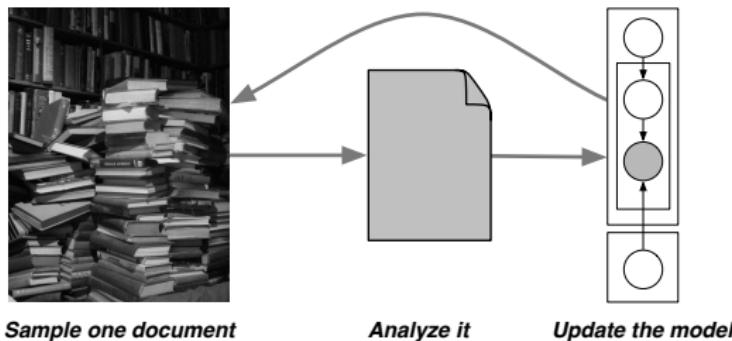
## Stochastic variational inference

- This is inefficient. We should know something about the global structure after seeing part of the data.
- And, it assumes a finite amount of data. We want algorithms that can handle **data sources**, information arriving in a constant stream.
- With **stochastic variational inference**, we can condition on large data and approximate the posterior of complex models.

# Stochastic variational inference

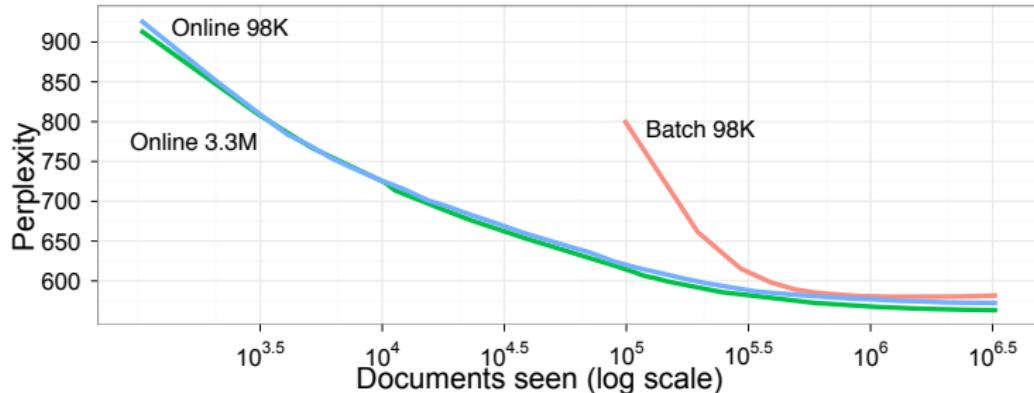
- The structure of the algorithm is:
  - Subsample the data—one data point or a small batch.
  - Infer local variables for the subsample.
  - Update the current estimate of the posterior of the global variables.
  - Repeat.
- This is efficient—we need only process one data point at a time.
- We will show: Just as easy as “classical” variational inference

# Stochastic variational inference for LDA



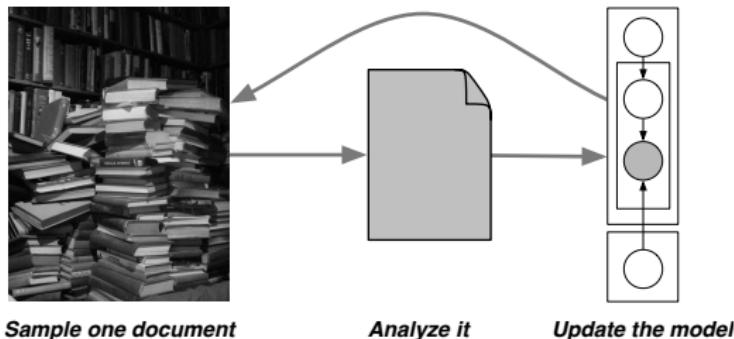
- ① Sample a document  $w_d$  from the collection
- ② Infer how  $w_d$  exhibits the current topics
- ③ Create “fake” topics, formed as though the  $w_d$  is the only document
- ④ Adjust the current topics according to the fake topics.
- ⑤ Repeat.

# Stochastic variational inference for LDA



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

# Stochastic variational inference for LDA



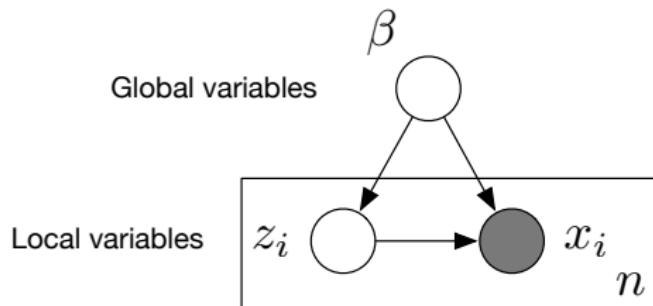
We have developed stochastic variational inference algorithms for

- Latent Dirichlet allocation
- The hierarchical Dirichlet process
- The discrete infinite logistic normal
- Mixed-membership stochastic blockmodels
- Bayesian nonparametric factor analysis
- Recommendation models and legislative models

# Organization

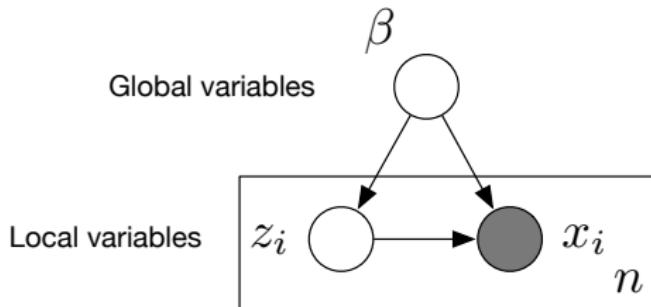
- Describe a generic class of models
- Derive mean-field variational inference in this class
- Derive natural gradients for the variational objective
- Review stochastic optimization
- Derive stochastic variational inference

# Organization



- We consider a **generic model**.
  - Hidden variables are local or global.
- We use **variational inference**.
  - Optimize a simple proxy distribution to be close to the posterior
  - Closeness is measured with Kullback-Leibler divergence
- Solve the optimization problem with **stochastic optimization**.
  - Stochastic gradients are formed by subsampling from the data.

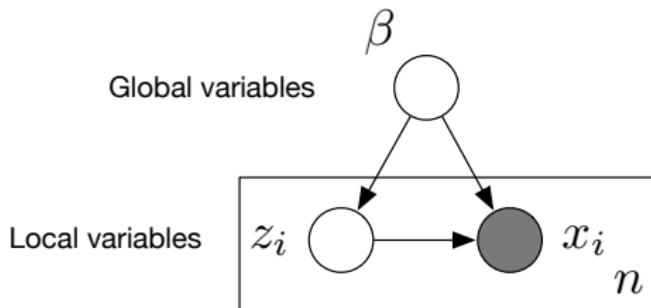
# Generic model



$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^n p(z_i | \beta) p(x_i | z_i, \beta)$$

- The observations are  $x = x_{1:n}$ .
- The **local** variables are  $z = z_{1:n}$ .
- The **global** variables are  $\beta$ .
- The  $i$ th data point  $x_i$  only depends on  $z_i$  and  $\beta$ .
- Our goal is to compute  $p(\beta, z | x)$ .

## Generic model

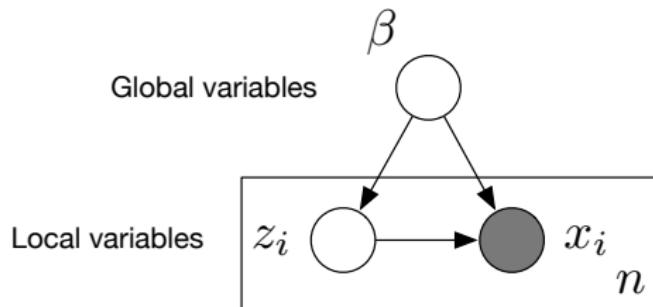


$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^n p(z_i | \beta) p(x_i | z_i, \beta)$$

- A **complete conditional** is the conditional of a latent variable given the observations and other latent variable.
- Assume each complete conditional is in the exponential family,

$$\begin{aligned} p(z_i | \beta, x_i) &= h(z_i) \exp\{\eta_\ell(\beta, x_i)^\top z_i - a(\eta_\ell(\beta, x_i))\} \\ p(\beta | z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\}. \end{aligned}$$

# Generic model



$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^n p(z_i | \beta) p(x_i | z_i, \beta)$$

- Bayesian mixture models
- Time series models  
(variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization  
(e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression  
(linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models  
(LDA and some variants)

# Mean-field variational inference

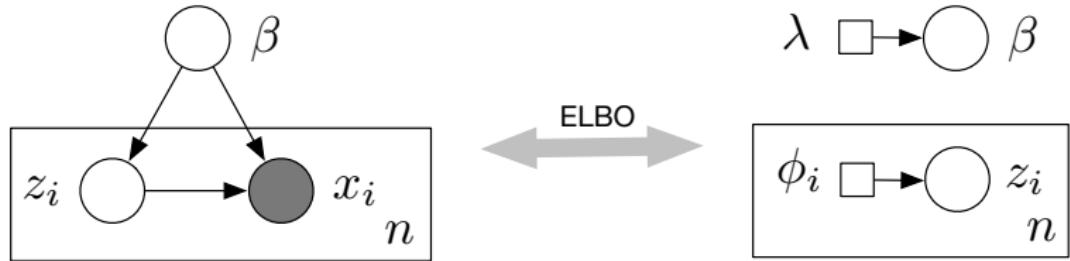


- Introduce a **variational distribution** over the latent variables  $q(\beta, z)$ .
- We optimize the **evidence lower bound** (ELBO) with respect to  $q$ ,

$$\log p(x) \geq E_q[\log p(\beta, Z, x)] - E_q[\log q(\beta, Z)].$$

- Up to a constant, this is the negative KL between  $q$  and the posterior.

# Mean-field variational inference



- We specify  $q(\beta, z)$  to be a fully factored variational distribution,

$$q(\beta, z) = q(\beta | \lambda) \prod_{i=1}^n q(z_i | \phi_i).$$

- Each instance of each variable has its own distribution.
- Each component is in the same family as the model conditional,

$$\begin{aligned} p(\beta | z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\} \\ q(\beta | \lambda) &= h(\beta) \exp\{\lambda^\top \beta - a(\lambda)\} \end{aligned}$$

(And, same for the local variational parameters.)

# Mean-field variational inference



- We optimize the ELBO with respect to these parameters,

$$\mathcal{L}(\lambda, \phi_{1:n}) = \text{E}_q[\log p(\beta, Z, x)] - \text{E}_q[\log q(\beta, Z)].$$

- Same as finding the  $q(\beta, z)$  that is closest in KL divergence to  $p(\beta, z|x)$
- The ELBO links the observations/model to the variational distribution.

# Mean-field variational inference



- Coordinate ascent: Iteratively update each parameter, holding others fixed.
- With respect to the global parameter, the gradient is

$$\nabla_\lambda \mathcal{L} = a''(\lambda)(\text{E}_\phi[\eta_g(Z, x)] - \lambda).$$

This leads to a simple coordinate update

$$\lambda^* = \text{E}_\phi [\eta_g(Z, x)].$$

- The local parameter is analogous.

# Mean-field variational inference

Initialize  $\lambda$  randomly.

Repeat until the ELBO converges

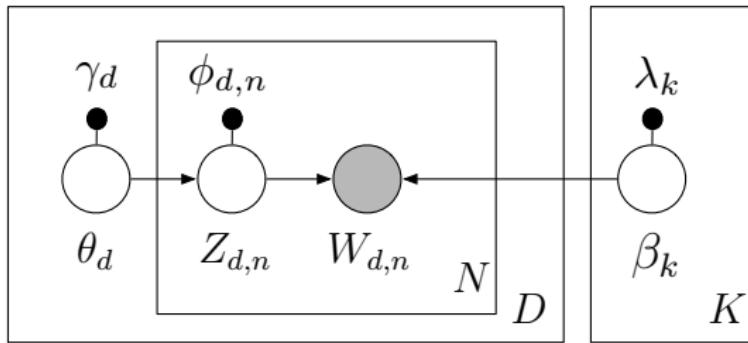
- ➊ For each data point, update the local variational parameters:

$$\phi_i^{(t)} = \mathbb{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \dots, n\}.$$

- ➋ Update the global variational parameters:

$$\lambda^{(t)} = \mathbb{E}_{\phi^{(t)}}[\eta_g(\mathcal{Z}_{1:n}, x_{1:n})].$$

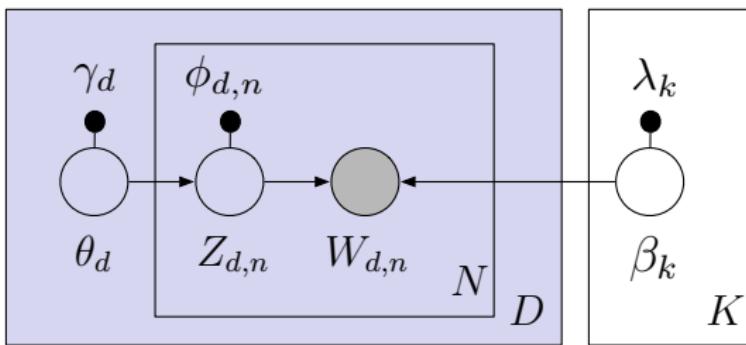
# Mean-field variational inference for LDA



- Document variables: Topic proportions  $\theta$  and topic assignments  $z_{1:N}$ .
- Corpus variables: Topics  $\beta_{1:K}$
- The variational distribution is

$$q(\beta, \theta, z) = \prod_{k=1}^K q(\beta_k | \lambda_k) \prod_{d=1}^D q(\theta_d | \gamma_d) \prod_{n=1}^N q(z_{d,n} | \phi_{d,n})$$

## Mean-field variational inference for LDA



- In the “local step” we iteratively update the parameters for each document, holding the topic parameters fixed.

$$\begin{aligned}\gamma^{(t+1)} &= \alpha + \sum_{n=1}^N \phi_n^{(t)} \\ \phi_n^{(t+1)} &\propto \exp\{\mathbb{E}_q[\log \theta] + \mathbb{E}_q[\log \beta_{\cdot, w_n}]\}.\end{aligned}$$

# Mean-field variational inference for LDA

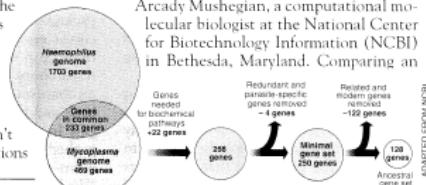
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>\*</sup> 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

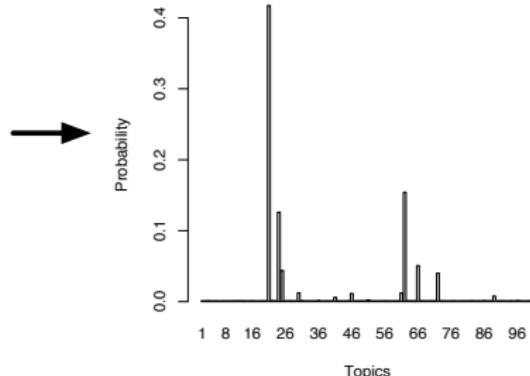
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Aracady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

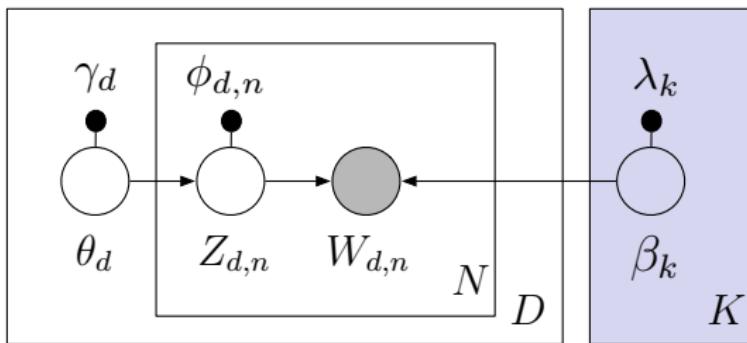


**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

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



# Mean-field variational inference for LDA



- In the “global step” we aggregate the parameters computed from the local step and update the parameters for the topics,

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

# Mean-field variational inference for LDA

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

# Mean-field variational inference for LDA

```
1: Initialize topics randomly.  
2: repeat  
3:   for each document do  
4:     repeat  
5:       Update the topic assignment variational parameters.  
6:       Update the topic proportions variational parameters.  
7:     until document objective converges  
8:   end for  
9:   Update the topics from aggregated per-document parameters.  
10:  until corpus objective converges.
```

# Mean-field variational inference

Initialize  $\lambda$  randomly.

Repeat until the ELBO converges

- ① Update the local variational parameters for each data point,

$$\phi_i^{(t)} = \mathbb{E}_{\lambda^{(t-1)}} [\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \dots, n\}.$$

- ② Update the global variational parameters,

$$\lambda^{(t)} = \mathbb{E}_{\phi^{(t)}} [\eta_g(Z_{1:n}, x_{1:n})].$$

- Note the relationship to existing algorithms like EM and Gibbs sampling.
- But we must analyze the whole data set before completing one iteration.

# Mean-field variational inference

Initialize  $\lambda$  randomly.

Repeat until the ELBO converges

- ① Update the local variational parameters for each data point,

$$\phi_i^{(t)} = \text{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \dots, n\}.$$

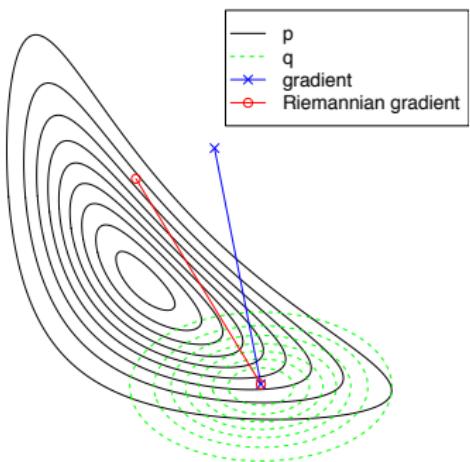
- ② Update the global variational parameters,

$$\lambda^{(t)} = \text{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$$

To make this more efficient, we need two ideas:

- Natural gradients
- Stochastic optimization

# The natural gradient



(from Honkela et al., 2010)

- In natural gradient ascent, we premultiply the gradient by the inverse of a Riemannian metric. Amari (1998) showed this is the steepest direction.
- For distributions, the Riemannian metric is the Fisher information.

# The natural gradient



- In the exponential family, the Fisher information is the second derivative of the log normalizer,  
$$G = a''(\lambda).$$
- So, the natural gradient of the ELBO is  
$$\hat{\nabla}_\lambda \mathcal{L} = \mathbb{E}_\phi [\eta_g(Z, x)] - \lambda.$$
- We can compute the natural gradient by computing the coordinate updates in parallel and subtracting the current variational parameters.

# Stochastic optimization

---

## A STOCHASTIC APPROXIMATION METHOD<sup>1</sup>

By HERBERT ROBBINS AND SUTTON MONRO

*University of North Carolina*

**1. Summary.** Let  $M(x)$  denote the expected value at level  $x$  of the response to a certain experiment.  $M(x)$  is assumed to be a monotone function of  $x$  but is unknown to the experimenter, and it is desired to find the solution  $x = \theta$  of the equation  $M(x) = \alpha$ , where  $\alpha$  is a given constant. We give a method for making successive experiments at levels  $x_1, x_2, \dots$  in such a way that  $x_n$  will tend to  $\theta$  in probability.

---

- 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



- We will use stochastic optimization for global variables.
- Let  $\nabla_{\lambda} \mathcal{L}_t$  be a realization of a random variable whose expectation is  $\nabla_{\lambda} \mathcal{L}$ .
- Iteratively set  $\lambda^{(t)} = \lambda^{(t-1)} + \epsilon_t \nabla_{\lambda} \mathcal{L}_t$
- This leads to a local optimum when

$$\begin{aligned}\sum_{t=1}^{\infty} \epsilon_t &= \infty \\ \sum_{t=1}^{\infty} \epsilon_t^2 &< \infty\end{aligned}$$

- Next step: Form a noisy gradient.

# A noisy natural gradient



- We need to look more closely at the conditional distribution of the global hidden variable given the local hidden variables and observations.
- The form of the local joint distribution is

$$p(z_i, x_i | \beta) = h(z_i, x_i) \exp\{\beta^\top f(z_i, x_i) - a(\beta)\}.$$

This means the conditional parameter of  $\beta$  is

$$\eta_g(z_{1:n}, x_{1:n}) = \langle \alpha_1 + \sum_{i=1}^n f(z_i, x_i), \alpha_2 + n \rangle.$$

- See the discussion of conjugacy in Bernardo and Smith (1994).

# A noisy natural gradient

- With local and global variables, we decompose the ELBO

$$\mathcal{L} = \mathbb{E}[\log p(\beta)] - \mathbb{E}[\log q(\beta)] + \sum_{i=1}^n \mathbb{E}[\log p(z_i, x_i | \beta)] - \mathbb{E}[\log q(z_i)]$$

- Sample a single data point  $t$  uniformly from the data and define

$$\mathcal{L}_t = \mathbb{E}[\log p(\beta)] - \mathbb{E}[\log q(\beta)] + n(\mathbb{E}[\log p(z_t, x_t | \beta)] - \mathbb{E}[\log q(z_t)]).$$

- The ELBO is the expectation of  $\mathcal{L}_t$  with respect to the sample.
- The gradient of the  $t$ -ELBO is a noisy gradient of the ELBO.
- The  $t$ -ELBO is like an ELBO where we saw  $x_t$  repeatedly.

## A noisy natural gradient

- Define the conditional as though our whole data set is  $n$  replications of  $x_t$ ,

$$\eta_t(z_t, x_t) = \langle \alpha_1 + n \cdot f(z_t, x_t), \alpha_2 + n \rangle$$

- The noisy natural gradient of the ELBO is

$$\nabla_{\lambda} \hat{\mathcal{L}}_t = E_{\phi_t}[\eta_t(Z_t, x_t)] - \lambda.$$

- This only requires the local variational parameters of one data point.
- In contrast, the full natural gradient requires all local parameters.

# Stochastic variational inference

Initialize global parameters  $\lambda$  randomly.

Set the step-size schedule  $\epsilon_t$  appropriately.

Repeat forever

- ① Sample a data point uniformly,

$$x_t \sim \text{Uniform}(x_1, \dots, x_n).$$

- ② Compute its local variational parameter,

$$\phi = E_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_t)].$$

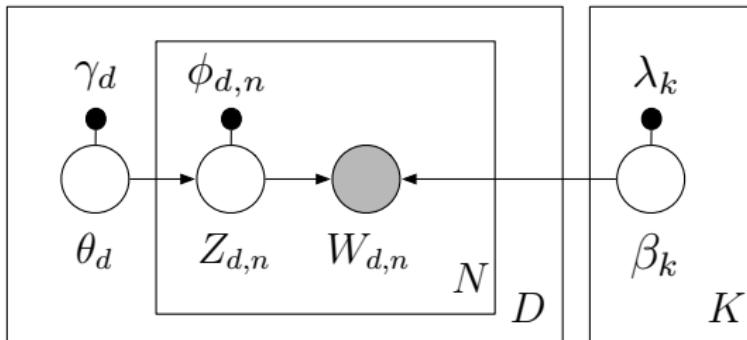
- ③ Pretend its the only data point in the data set,

$$\hat{\lambda} = E_\phi[\eta_t(Z_t, x_t)].$$

- ④ Update the current global variational parameter,

$$\lambda^{(t)} = (1 - \epsilon_t)\lambda^{(t-1)} + \epsilon_t \hat{\lambda}.$$

# Stochastic variational inference in LDA

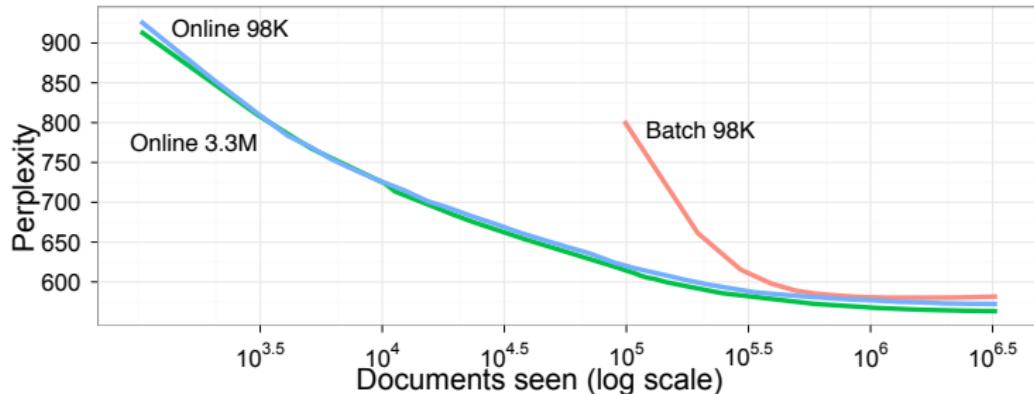


- ① Sample a document
- ② Estimate the local variational parameters using the current topics
- ③ Form “fake topics” from those local parameters
- ④ Update the topics to be a weighted average of “fake” and current topics

# Stochastic variational inference in LDA

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

# Stochastic variational inference in LDA



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

# Stochastic variational inference



We defined a generic algorithm for scalable variational inference.

- Bayesian mixture models
- Time series models  
(variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization  
(e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression  
(linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models  
(LDA and some variants)

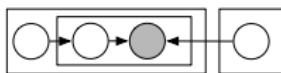
# Stochastic variational inference



- See Hoffman et al. (2010) for LDA (and code).
- See Wang et al. (2010) for Bayesian nonparametric models (and code).
- See Sato (2001) for the original stochastic variational inference.
- See Honkela et al. (2010) for natural gradients and variational inference.
- Many open issues, e.g., how to handle nonconjugacy (CTM, DTM)?
- This conference
  - *Sparse Stochastic Inference for Latent Dirichlet Allocation* (Mimno, Hoffman, Blei)
  - *Nonparametric Variational Inference* (Gershman, Hoffman, Blei)

# Stochastic variational inference

## Make assumptions



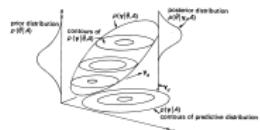
## Infer the posterior



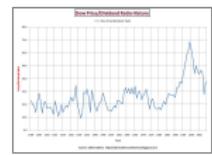
## Collect data



## Check



## Predict



## Explore

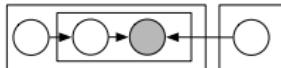


- Many applications posit a model, condition on data, and use the posterior.
- We can now apply this kind of data analysis to very large data sets.

# **Using and Checking Topic Models**

# Evaluating topic models

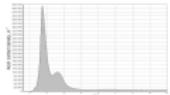
## Make assumptions



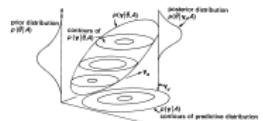
## Collect data



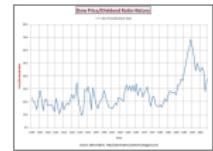
## Infer the posterior



## Check



## Predict



## Explore



- How do we check, predict, and explore?

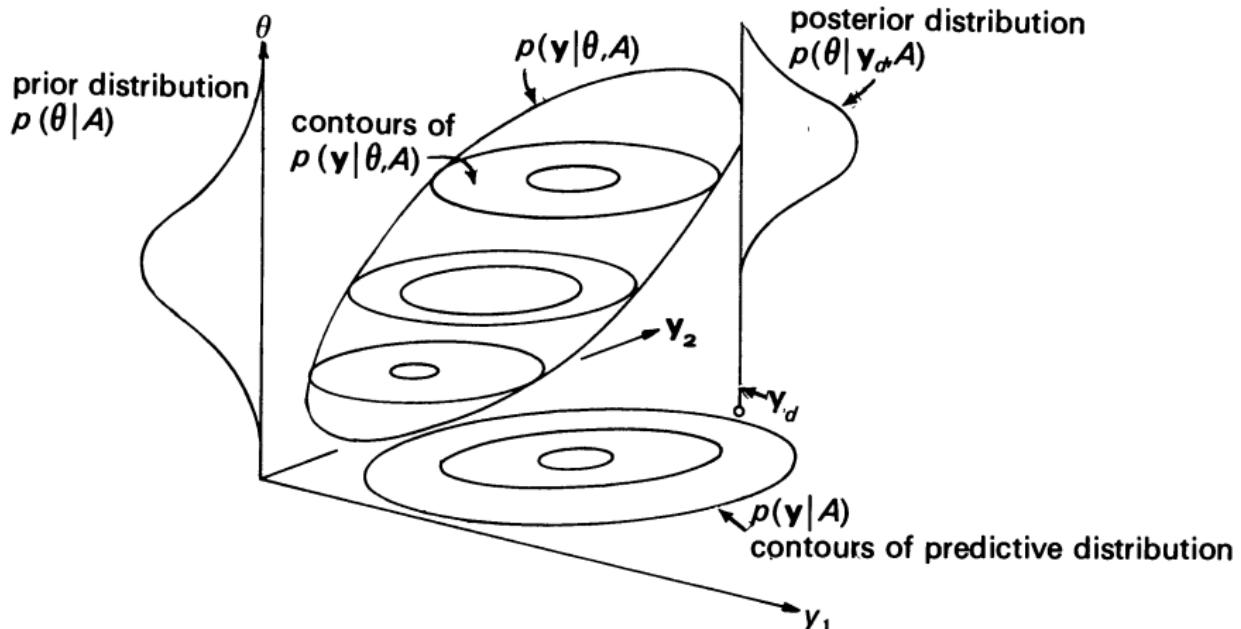
# Evaluating topic models

- Questions we should ask in evaluation:
  - Does my model work? Is it better than another model?
  - Which topic model should I choose? Should I make a new one?
- These questions are tied up in the application at hand.
- Sometimes evaluation is easy, especially in prediction tasks.
- But a promise of topic models is that they give good exploratory tools.  
Evaluation is complicated, e.g., is this a good navigator of my collection?
- And this leads to more questions:
  - How do I interpret a topic model?
  - What quantities help me understand what it says about the data?

# Evaluating topic models

- How to interpret and evaluate topic models is an active area of research.
  - Visualizing topic models
  - Naming topics
  - Matching topic models to human judgements
  - Matching topic models to external ontologies
  - Computing held out likelihoods in different ways
- We will discuss **posterior predictive checks** for topic modeling.

# Posterior predictive checks



This is a **predictive check** from Box (1980).

# Posterior predictive checks

- Three stages to model building: estimation, criticism, and revision.
- In **criticism**, the model “confronts” our data.
- Suppose we observe a data set  $\mathbf{y}$ . The predictive distribution is the distribution of data *if the model is true*:

$$p(\mathbf{y} | M) = \int_{\theta} p(\mathbf{y} | \theta) p(\theta)$$

- Locating  $\mathbf{y}$  in the predictive distribution indicates if we can “trust” the model.
- Or, locating a **discrepancy function**  $g(\mathbf{y})$  in its predictive distribution indicates if what is important to us is captured in the model.

## Posterior predictive checks

- Rubin (1984) located the data  $\mathbf{y}$  in the **posterior**  $p(y|\mathbf{y}, M)$ .
- Gelman, Meng, Stern (1996) expanded this idea to “realized discrepancies” that include **hidden variables**  $g(\mathbf{y}, \mathbf{z})$ .
- We might make modeling decisions based on a variety of simplifying considerations (e.g., algorithmic). But we can design the realized discrepancy function to capture what we really care about.
- Further, realized discrepancies let us consider which **parts of the model** fit well and which parts don’t. This is apt in exploratory tasks.

## Posterior predictive checks in topic models

- Consider a decomposition of a corpus into topics, i.e.,  $\{w_{d,n}, z_{d,n}\}$ . Note that  $z_{d,n}$  is a latent variable.
- For all the observations assigned to a topic, consider the variable  $\{w_{d,n}, d\}$ . This is the observed word and the document it appeared in.
- One measure of how well a topic model fits the LDA assumptions is to look at the **per-topic mutual information** between  $w$  and  $d$ .
- If the words from the topic are independently generated then we expect lower mutual information.
- What is “low”? To answer that, we can shuffle the words and recompute. This gives values of the MI when the words are independent.

# Posterior predictive checks in topic models

4	10	3	13
tax income taxation taxes revenue  estate subsidies exemption organizations year treasury  consumption taxpayers earnings funds	labor workers employees union employer employers employment work  employee job  bargaining unions  worker collective  industrial	women sexual  men sex child family children gender woman  marriage discrimination  male social female  parents	contract liability parties contracts party  creditors agreement breach contractual terms bargaining contracting debt exchange limited
6	15	1	16
jury trial crime defendant defendants sentencing judges  punishment judge crimes  evidence sentence jurors offense guilty	speech free  amendment freedom expression protected culture context  equality values conduct  ideas protect content	firms price corporate firm value market cost  capital shareholders stock  insurance efficient assets offer share	constitutional political constitution government justice amendment  history people legislative opinion  fourteenth article majority citizens republican

- This realized discrepancy measures model fitness
- Can use it to measure model fitness **per topic**.
- Helps us explore parts of the model that fit well.

# **Discussion**

# This tutorial

- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- How do I check and evaluate a topic model?
- What are some unanswered questions in this field?
- How can I learn more?

# Introduction to topic modeling

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

## Documents

### 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

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

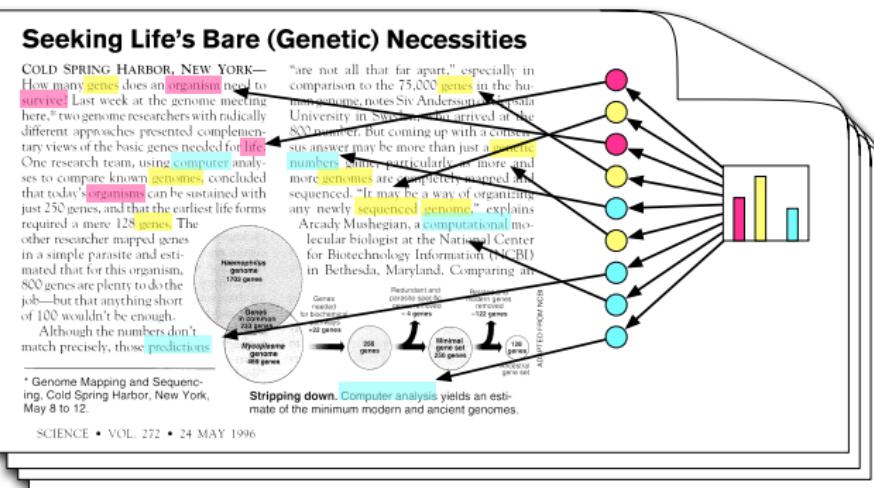
SCIENCE • VOL. 272 • 24 MAY 1996

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Stéphane Auclair, a University of Swabia scientist arrived at the SXS meeting. But coming up with a consensus answer may be more than just a matter of numbers. Some, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

the genomes, he says, "we can find the minimum set of genes that are common to all." Redundant and冗余的 genes are removed until only 128 genes remain. This is the minimum number of genes required for life.

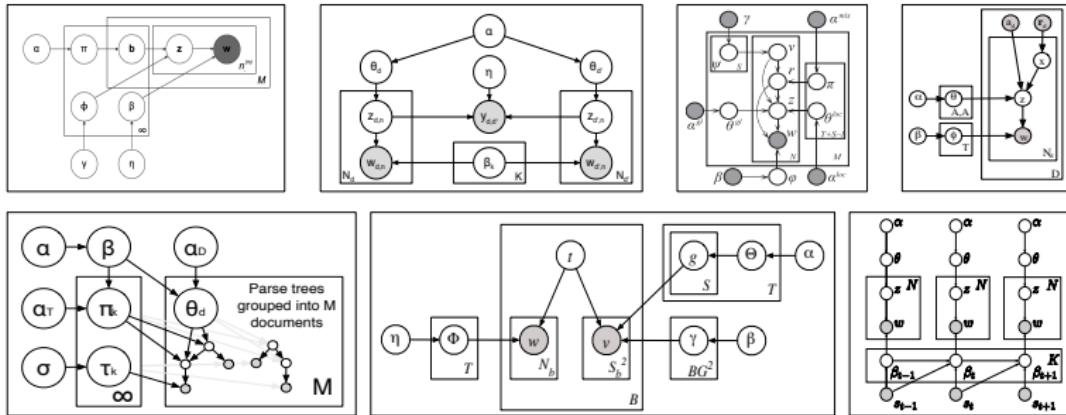
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

## Topic proportions and assignments



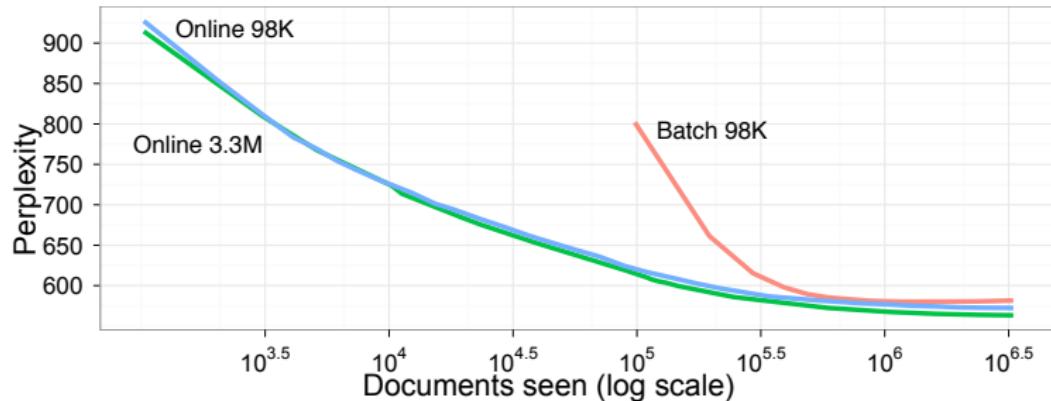
- LDA assumes that there are  $K$  topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.

# Extensions of LDA



- Topic models can be adapted to many settings
- We can relax assumptions, combine models, or model more complex data.

# Posterior inference



- Posterior inference is the central computational problem.
- Stochastic variational inference is a scalable algorithm.
- (Note: There are many types of inference we didn't discuss.)

# Posterior predictive checks

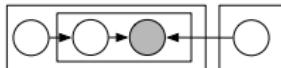
4	10	3	13
tax income taxation taxes revenue  estate subsidies exemption organizations year treasury  consumption taxpayers earnings funds	labor workers employees union employer employers employment  work employee job  bargaining unions worker collective industrial	women sexual men sex child family children gender woman  marriage discrimination male social female parents	contract liability parties contracts party  creditors agreement breach contractual terms bargaining contracting debt exchange limited
6	15	1	16
jury trial crime defendant defendants sentencing judges punishment judge crimes evidence sentence jurors offense guilty	speech free amendment freedom expression protected culture context  equality values conduct ideas information protect content	firms price corporate firm value market cost capital shareholders stock insurance efficient assets offer share	constitutional political constitution government justice amendment history people legislative opinion fourteenth article majority citizens republican

# Some open issues

- **Model interpretation and model checking**  
Which model should I choose for which task?
- **Incorporating corpus, discourse, or linguistic structure**  
How can our knowledge of language help us build and use exploratory models of text?
- **Interfaces and “downstream” applications of topic modeling**  
What can I do with an annotated corpus? How can I incorporate latent variables into a user interface?
- **Theoretical understanding of approximate inference**  
What do we know about variational inference? Can we analyze it from either the statistical or learning perspective?

# If you remember one picture...

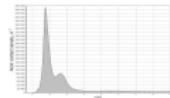
## Make assumptions



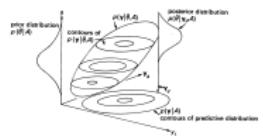
## Collect data



## Infer the posterior



## Check



## Predict



## Explore



“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.”

(J. Tukey, *The Future of Data Analysis*, 1962)