

# Scalable Inference in Probabilistic Topic Models

(Online variational inference is great)

David M. Blei

Department of Computer Science  
Princeton University

July 27, 2011

Supported by the AFOSR, ONR and NSF

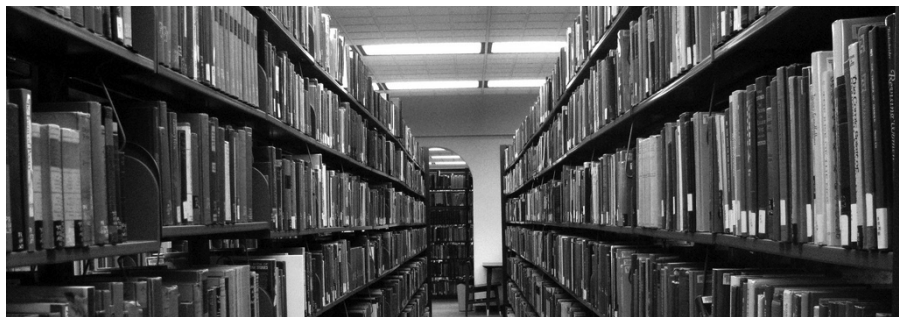
# Information overload



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.

# Topic modeling



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.

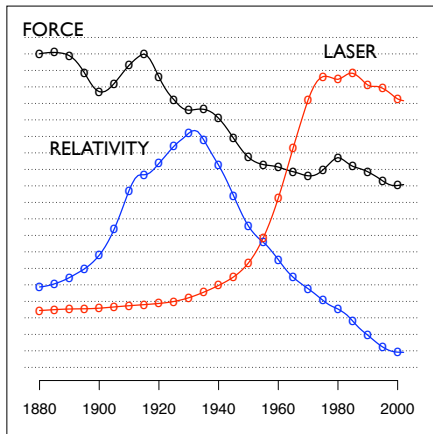
# Discover topics from a corpus

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

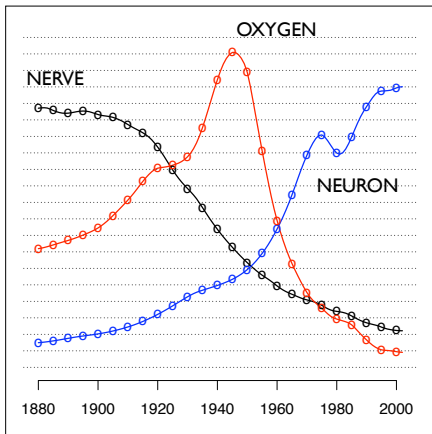


# Model the evolution of topics over time

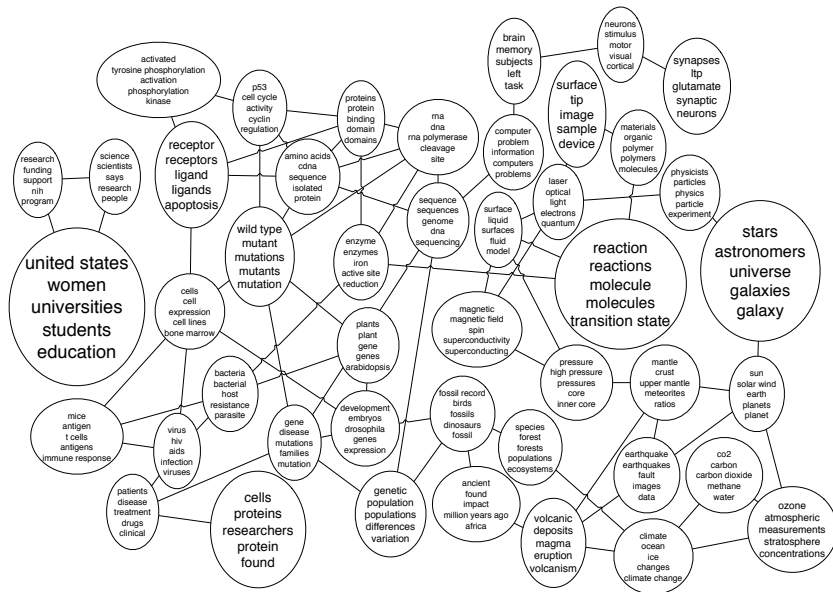
**"Theoretical Physics"**



**"Neuroscience"**



# Model connections between topics



# Browse and discover patterns in large data sets

Original article

Topic-based browser

Related articles

## Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts

Gerard Salton, James Allan, Chris Buckley,

Vast amounts of text material are now available in machine-readable processing. Here, approaches are outlined for manipulating and accessing subject areas in accordance with user needs. In particular, mining text themes, traversing texts selectively, and extracting summaries reflect text content.

Many kinds of texts are currently available in machine-readable form and are amenable to automatic processing. Because the available databases are large and cover many different subject areas, automatic aids must be provided to users interested in accessing the data. It has been suggested that links be placed between related pieces of text, connecting, for example, particular text paragraphs to other paragraphs covering related subject matter. Such a linked text structure, often called hypertext, makes it possible for the reader to start with particular text passages and use the linked structure to find related text elements (1). Unfortunately, until now, viable methods for automatically building large hypertext structures and for using such structures in a sophisticated way have not been available. Here we give methods for constructing text relation maps and for using text relations to access and use text databases. In particular, we outline procedures for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

### Text Analysis and Retrieval: The Smart System

The Smart system is a sophisticated text retrieval tool, developed over the past 30 years, that is based on the vector space

The authors are in the Department of Computer Science, Cornell University, Ithaca, NY 14853-7501, USA.

model of retrieval model, all information as well as information sent by text, or it is typically a word, associated with this action. In principle, chosen from a context a thesaurus, but by constructing such for unrestricted topics to derive the terms under consideration terms assigned to a text content.

Because the term for context represent introduce a term-weight high weights to and lower weights to. A powerful term-weight is the well-known term frequency (f), which is frequency (f) in p with a low frequency (f). Such terms that which they occur in

When all texts are sent by weighted  $D_i = (d_{i1}, d_{i2}, \dots)$  weight assigned to a similarity measure between pairs of vectors. Thus, g

SCIENCE • VOLUME

## "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts" (1994)

TOPIC	PROB
data computer system information network	0.30
information library text index libraries	0.19
two three four different single	0.16

DOCUMENT	SCORE
"Global Text Matching for Information Retrieval" (1991)	0.2570
"Automatic Text Analysis" (1970)	0.3110
"Gauging Similarity with n-Grams: Language-Independent Categorization of Text" (1995)	0.3210
"Developments in Automatic Text Retrieval" (1991)	0.5480
"Simple and Rapid Method for the Coding of Punched Cards" (1962)	0.3610
"Data Processing by Optical Coincidence" (1961)	0.4290
"Pattern-Analyzing Memory" (1976)	0.4320
"The Storing of Pamphlets" (1899)	0.4440
"A Punched-Card Technique for Computing Means, Standard Deviations, and the Product-Moment Correlation Coefficient and for Listing Scattergrams" (1946)	0.4550

## Global Text Matching for Information Retrieval

GERARD SALTON\* and CHRIS BUCKLEY

An approach is outlined for the retrieval of natural language texts in response to available search requests and for the recognition of content similarities between text corpora. The proposed retrieval process is based on flexible text matching procedures carried out in a number of different text environments and is applicable to large text collections covering unrestricted subject matter. For unrestricted text environments this system appears to outperform other currently available methods.

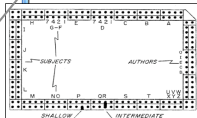
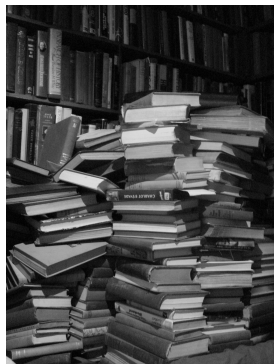


Fig. 1. The punch card, showing the different modes of searching and the "A-Z" table. Comparison of these two modes can produce any number from 1 to 26. It is also possible to code numbers 1 to 10 in a 10-hole field and only one mode. They are required to select the number stored (1). To select a given number in the first mode, it may be necessary to provide more than one value.

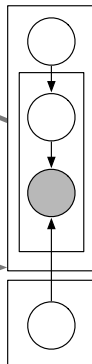
### THE STORING OF PAMPHLETS.

On reading Professor Minot's explanation of his method of storing pamphlets as given in the issue of December 30th I feel inclined to add a word in commendation of the method. I began using these boxes six or seven years ago and now have 150 upon my shelves. About one-half are devoted to Experiment Station bulletins, the boxes being labeled by States and arranged alphabetically. The other half is used for miscellaneous pamphlets on subjects pertaining to my line of work. The boxes have proved perfectly satisfactory in every way, and as a single time-saving device they are worth many times the cost. My system of pamphlet arrangement differs in some ways from that adopted by Professor Minot and has been adopted only after trial of several other methods.

# We need online inference



***Analyze the collection***



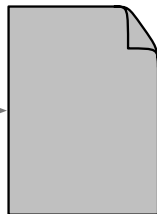
***Update the model***

- Existing algorithms process document collections in *batch*.
- This is inefficient for large collections.
- It cannot accommodate streaming collections.

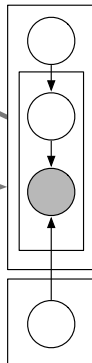
# We need online inference



***Sample one document***



***Analyze it***



***Update the model***

- Allows us to analyze millions of documents
- Lets us develop topic models on streaming collections

# This talk

- 1 Introduction to topic modeling
- 2 Online inference for topic models

# Introduction to topic modeling

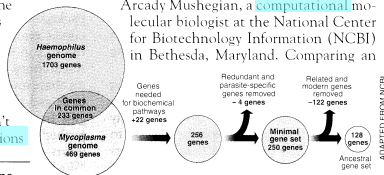
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

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

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



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

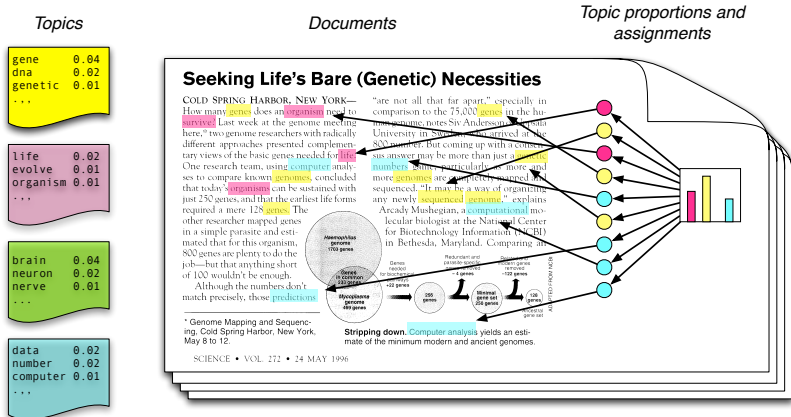
**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

**Simple intuition:** Documents exhibit multiple topics.

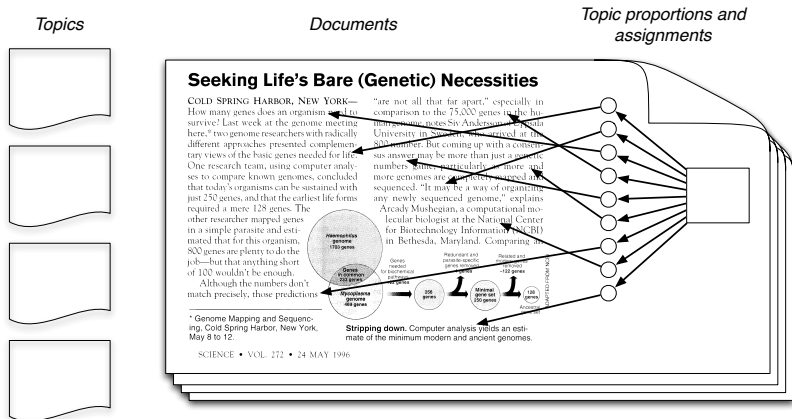


# Generative model for latent Dirichlet allocation (LDA)



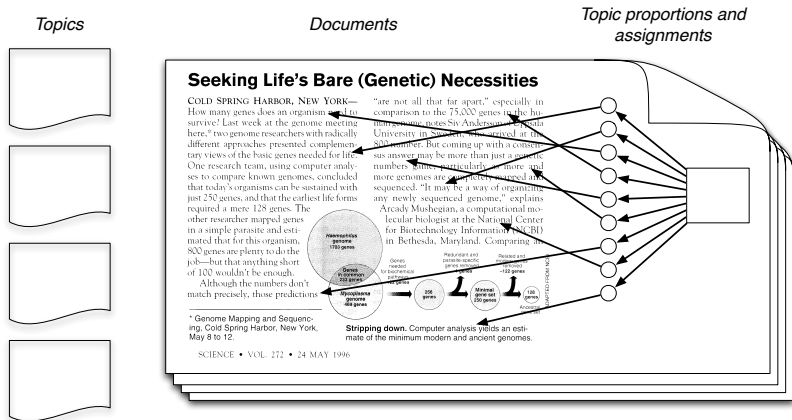
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

# The posterior distribution



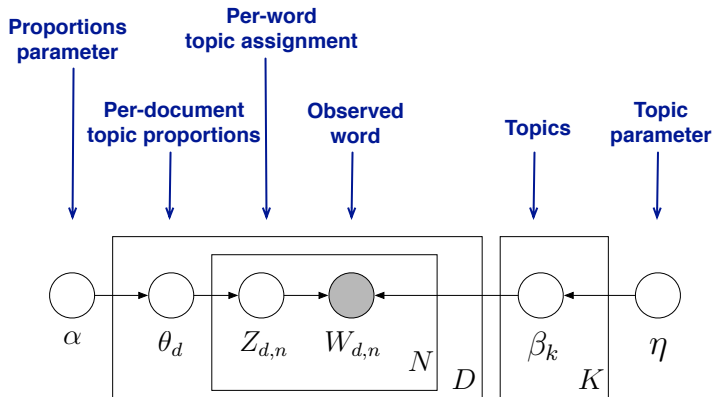
- In reality, we only observe the documents
- The other structure are **hidden variables**

# The posterior distribution



- Our goal is to **infer** the hidden variables
  - I.e., compute their distribution conditioned on the documents
- $p(\text{topics, proportions, assignments} \mid \text{documents})$

# LDA as a graphical model



- Encodes our assumptions about the data
- Helps us derive ways of computing with data
- Isolates independence assumptions, which are separate from other specific details of the model

# Example inference

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Anshu Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



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

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

SCIENCE • VOL. 272 • 24 MAY 1996

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

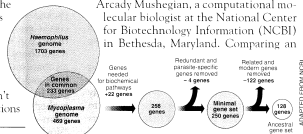
# Example inference

## Seeking Life's Bare (Genetic) Necessities

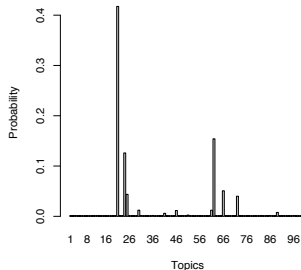
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



**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

# Used in exploratory tools of document collections

## Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts

Gerard Salton, James Allan, Chris Buckley,

Vast amounts of text material are now available in machine-readable form. Here, approaches are outlined for manipulating and accessing subject areas in accordance with user needs. In particular, methods for mining text themes, traversing texts selectively, and extracting summaries that reflect text content.

Many kinds of texts are currently available in machine-readable form and are amenable to automatic processing. Because the available databases are large and cover many different subject areas, automatic aids must be provided to users interested in accessing the data. It has been suggested that links be placed between related pieces of text, connecting, for example, particular text paragraphs to other paragraphs covering related subject matter. Such a linked text structure, often called hypertext, makes it possible for the reader to start with particular text passages and use the linked structure to find related text elements (1). Unfortunately, until now, viable methods for automatically building large hypertext structures and for using such structures in a sophisticated way have not been available. Here we give methods for constructing text relation maps and for using text relations to access and use text databases. In particular, we outline procedures for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

### Text Analysis and Retrieval: The Smart System

The Smart system is a sophisticated text retrieval tool, developed over the past 30 years, that is based on the vector space

model of retrieval. In this model, all information is represented by sets, or vectors, which are typically a word, associated with the data. In principle, documents are chosen from a corpus, such as a thesaurus, but because of the large number of terms, constructing such a structure for unrestricted text is difficult. To derive the terms, we use a list of terms under consideration, and assign terms to a text content.

Because the terms are represented by vectors, we introduce a term-weighting scheme that assigns high weights to terms that occur frequently and lower weights to terms that occur infrequently. A powerful term-weighting scheme is the well-known term frequency-inverse document frequency ( $f_i$ ), which is the frequency ( $f_i$ ) of a term in a document ( $d_i$ ) divided by the square root of the total number of documents ( $N$ ) in which the term occurs.

When all texts are represented by weighted vectors, the similarity between two texts is measured by the cosine of the angle between the two vectors. Thus, the

SCIENCE • VOL.

## "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts" (1994)

TOPIC	PROB
data computer system information network	0.30
information library text index libraries	0.19
two three four different single	0.16

DOCUMENT	SCORE
"Global Text Matching for Information Retrieval" (1991)	0.2570
"Automatic Text Analysis" (1970)	0.3110
"Gauging Similarity with n-Grams: Language-Independent Categorization of Text" (1995)	0.3210
"Developments in Automatic Text Retrieval" (1991)	0.3480
"Simple and Rapid Method for the Coding of Punched Cards" (1962)	0.3610
"Data Processing by Optical Coincidence" (1961)	0.4290
"Pattern-Analyzing Memory" (1976)	0.4320
"The Storing of Pamphlets" (1899)	0.4440
"A Punched-Card Technique for Computing Means, Standard Deviations, and the Product-Moment Correlation Coefficient and for Listing Scattergrams" (1946)	0.4550

## Global Text Matching for Information Retrieval

GERARD SALTON\* AND CHRIS BUCKLEY

An approach is outlined for the retrieval of natural language texts in response to available search requests and for the recognition of content similarities between text excerpts. The proposed retrieval process is based on flexible text matching procedures carried out in a number of different text environments and is applicable to large text collections covering unrestricted subject matter. For unrestricted text environments this system appears to outperform other currently available methods.

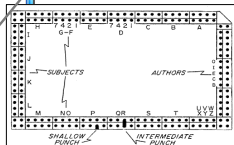


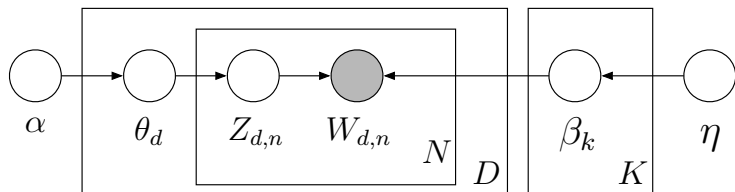
Fig. 1. The punch card showing the different number of punching and the "4-3-1" code. Contributions of these four numbers can produce any number from 1 to 10 (1). It is also possible to code numbers 1 to 10 in a five-bit field and only one mapping is required to select the number desired (1). To select a given number in the five-bit field, it may be necessary to encode more than five bits.

### THE STORING OF PAMPHLETS.

On reading Professor Minot's explanation of his method of storing pamphlets as given in the issue of December 30th I feel inclined to add a word in commendation of the method. I began using these boxes six or seven years ago and now have 152 upon my shelves. About one-half are devoted to Experiment Station bulletins, the boxes being labeled by States and arranged alphabetically. The other half is used for miscellaneous pamphlets on subjects pertaining to my line of work. The boxes have proved perfectly satisfactory in every way, and as a simple time-saving device they are worth many times the cost. My system of pamphlet arrangement differs in some ways from that adopted by Professor Minot and has been adopted only after trial of several other methods.

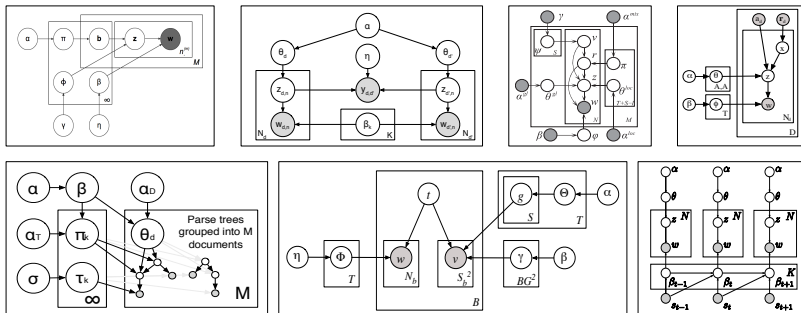


# Summary of LDA



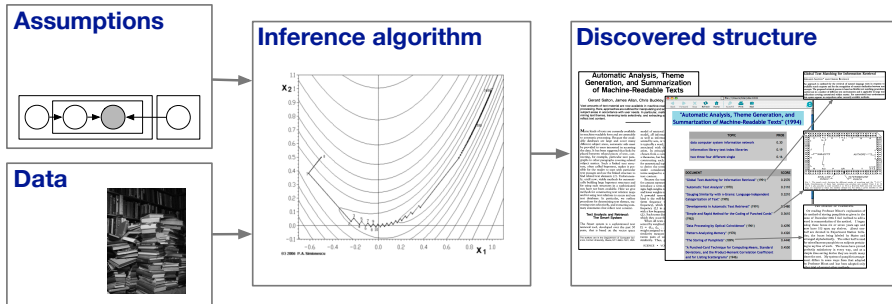
- LDA can
  - visualize the hidden thematic structure in large corpora
  - generalize new data to fit into that structure
- Builds on Deerwester et al. (1990) and Hofmann (1999)  
It is an example of a *mixed membership model* (Erosheva, 2004)  
Relates to *multinomial PCA* (Jakulin and Buntine, 2002).
- Was independently invented for genetics (Pritchard et al., 2000)

## Why develop these kinds of models?



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

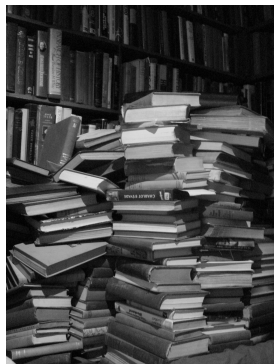
# Bigger Picture: Probabilistic modeling



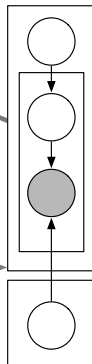
- Research in modeling separates these basic activities
- Though linked, we can work on each piece separately

# Online inference for topic models

# We need online inference



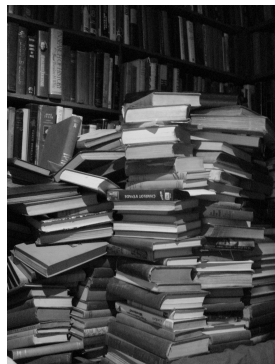
***Analyze the collection***



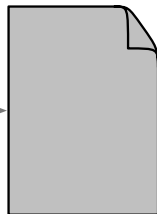
***Update the model***

- Existing algorithms process document collections in *batch*.
- This is inefficient for large collections.
- It cannot accommodate streaming collections.

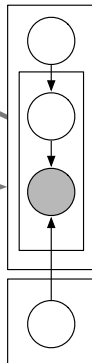
# We need online inference



***Sample one document***



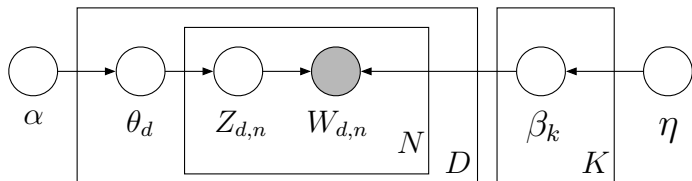
***Analyze it***



***Update the model***

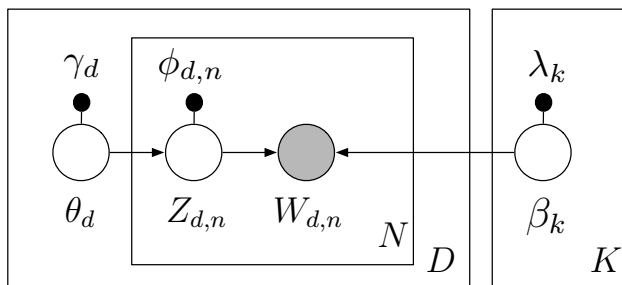
- Allows us to analyze millions of documents
- Lets us develop topic models on streaming collections

# Computation with LDA



- Our goal is to compute the *posterior*, the conditional distribution of the hidden variables given the documents.
- We will build on *variational inference*.
  - Posit a parameterized distribution  $q$  over hidden variables.
  - Optimize to make  $q$  close (in KL) to the posterior.

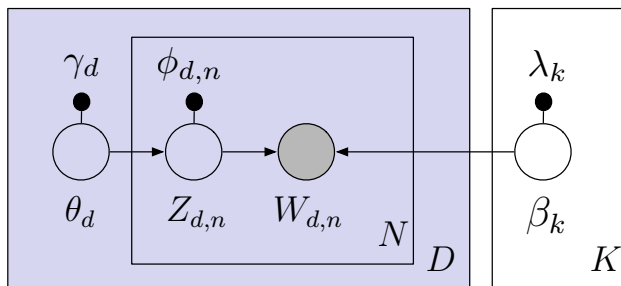
# Batch variational inference for LDA



- The *mean field distribution* places a variational parameter on each hidden variable.
- Optimize these with coordinate ascent, iteratively optimizing each parameter while holding the others fixed.



# Batch variational inference for LDA



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

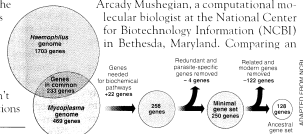
# Example inference (again)

## Seeking Life's Bare (Genetic) Necessities

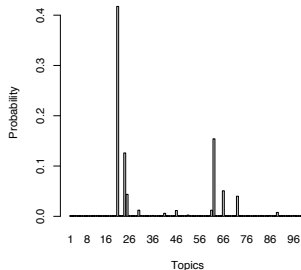
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

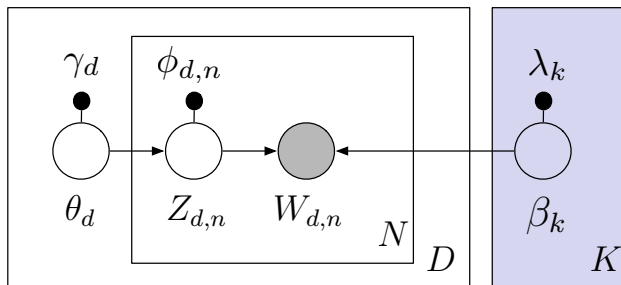


**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.



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

# Batch variational inference for LDA

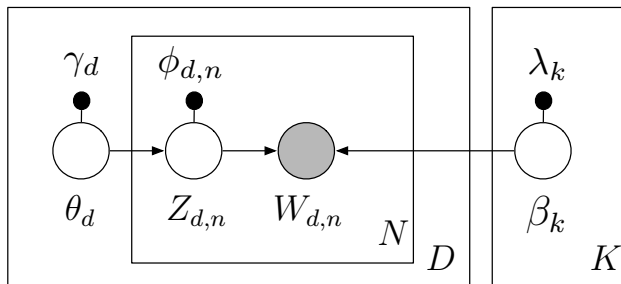


- In the “global step” we aggregate the parameters computed from the local step and update the parameters for the topics.

# Example topic inference

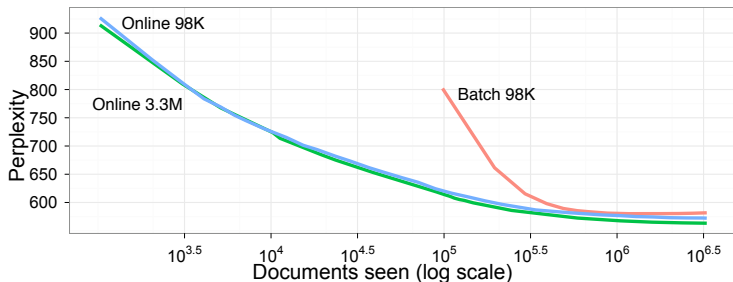
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

# Online inference for LDA



- 1 Randomly pick a document.
- 2 Perform local variational inference with the current topics.
- 3 Form “fake” topics, treating the sampled document as though it were the only document in the collection.
- 4 Update the topics to be a weighted average of the fake topics and current topics.

# Analyzing 3.3M articles from Wikipedia



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 industry market billion	business service companies industry company management systems services	business service companies industry services company management public	business industry service companies services company management public

# Why does this work?

---

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

By HERBERT ROBBINS AND SUTTON MONRO

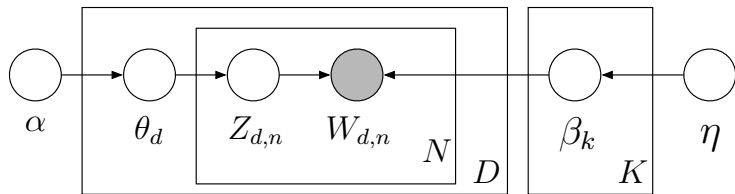
*University of North Carolina*

**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.
- See Hoffman et al. (2010) and Sato (2001).

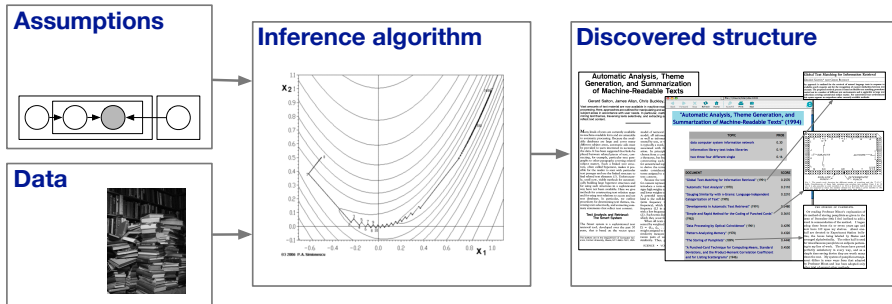
# Summary



- Hierarchical Bayesian models of text are a powerful way to explore, summarize and search large archives of documents.
- Algorithmic advances in approximate posterior inference let us apply complex models to large real-world data sets.



# Online inference is promising



- Stochastic variational methods are a general way to approximate a posterior with massive/streaming data.
- Powerful algorithm for topic modeling, and can be adapted hierarchical models for many types of data.
- Software and papers: [www.cs.princeton.edu/~blei/](http://www.cs.princeton.edu/~blei/)

# Open research directions

- **Model diagnostics and model checking**

Which model should I choose for which task? How does this problem change in the face of streaming data?

- **Interfaces and “downstream” applications of topic modeling**

What can I do with an annotated corpus? What can I do with a changing approximate posterior?

- **Theoretical understanding of approximate inference**

What do we know about variational inference from either the statistical or learning perspective?

“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)

# On-line variational inference for 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  $\tilde{\lambda}_k = \eta + D \sum_{n \sim w_t} \phi_{t,n}$
- 11:   Set  $\lambda_k = (1 - \rho_t)\lambda_k + \rho_t \tilde{\lambda}_k$ .
- 12: **end for**