Scalable Inference in Probabilistic Topic Models

(Online variational inference is great)

David M. Blei

Department of Computer Science
Princeton University

July 27, 2011

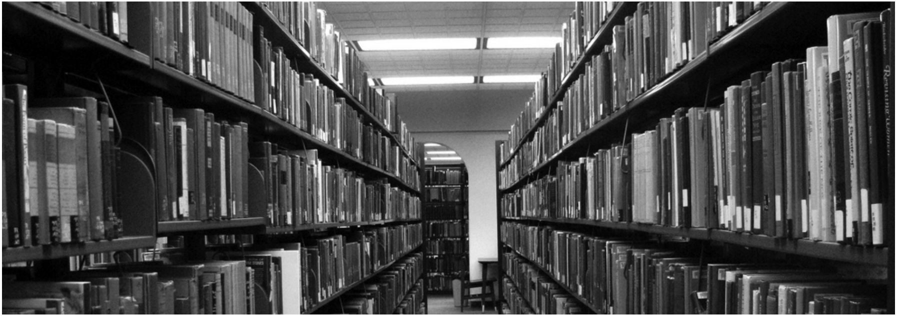
Supported by the AFOSR, ONR and NSF

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.

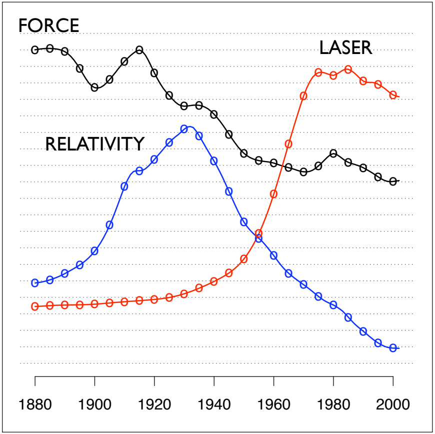
1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, and search the texts.

Discover topics from a corpus

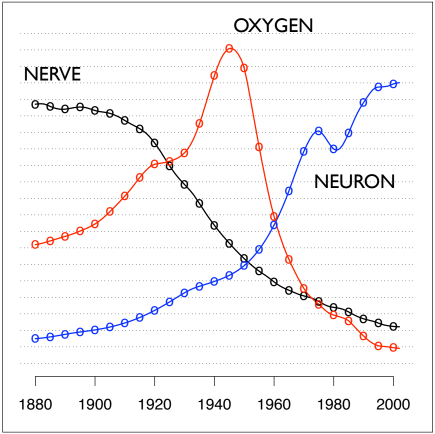
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	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

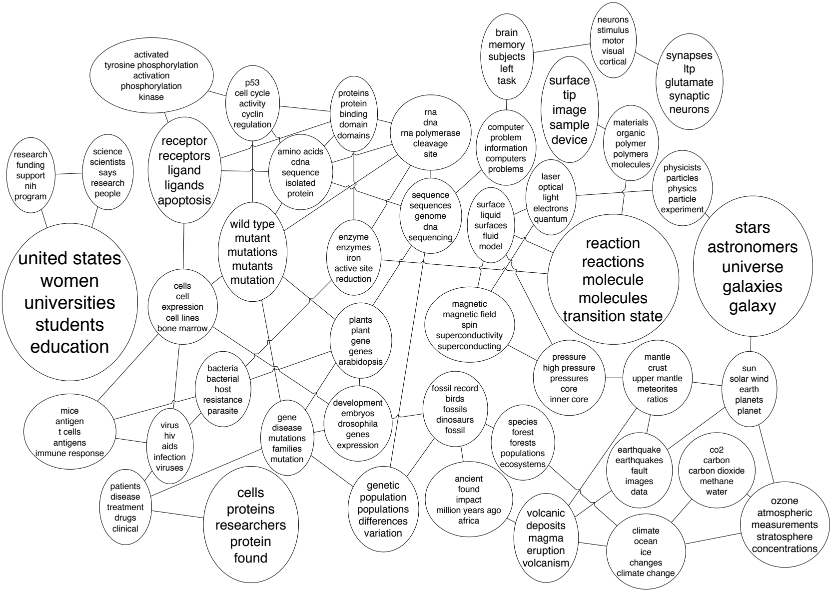
"Theoretical Physics"



"Neuroscience"

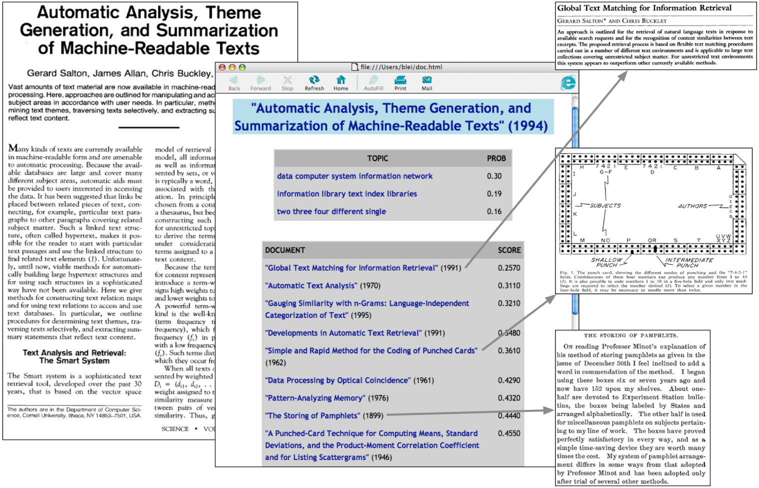


Model connections between topics

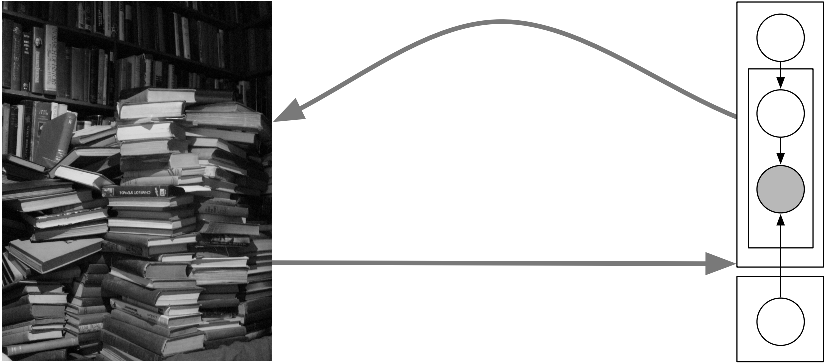


Browse and discover patterns in large data sets

Original article | Topic-based browser | Related articles



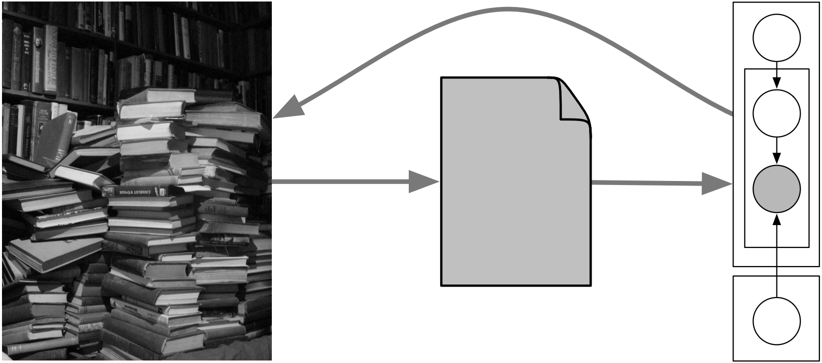
We need online inference



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

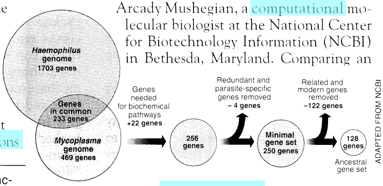
Introduction to topic modeling

Latent Dirichlet allocation

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

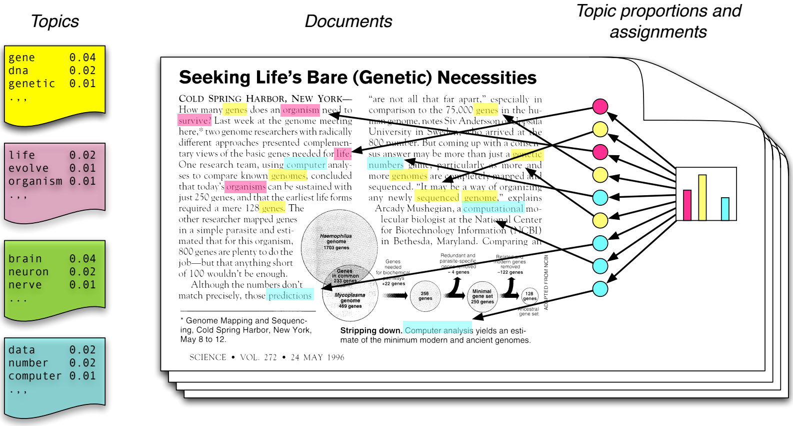


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

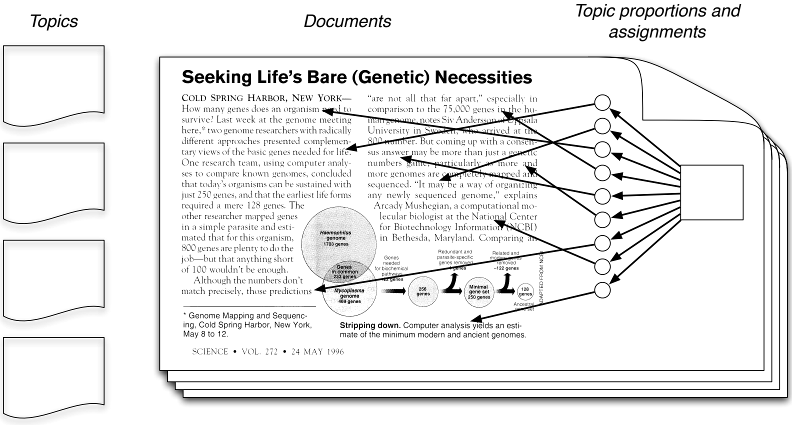
Simple intuition: Documents exhibit multiple topics.

Generative model for latent Dirichlet allocation (LDA)



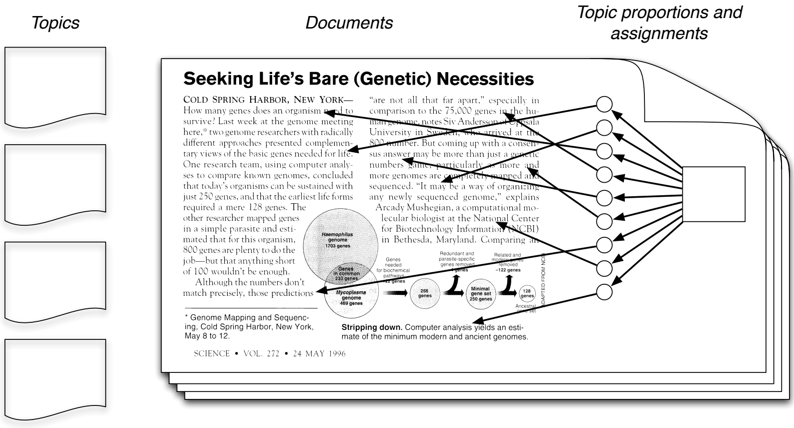
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



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

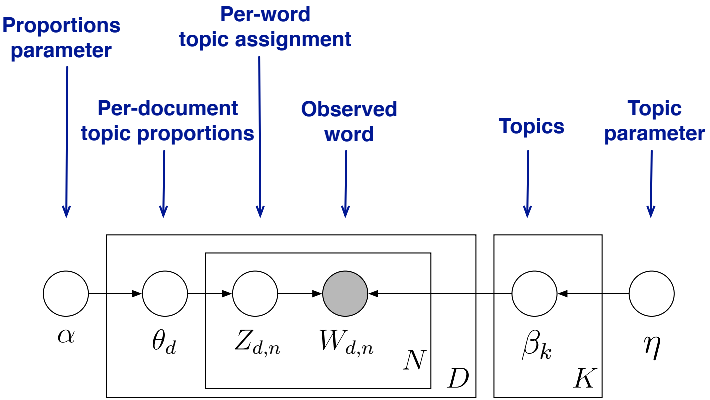
The posterior distribution



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

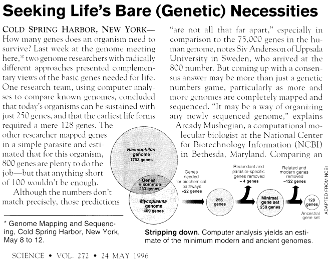
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

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

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)

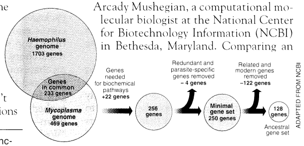
- **Model**: 100-topic LDA model using variational inference.

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.

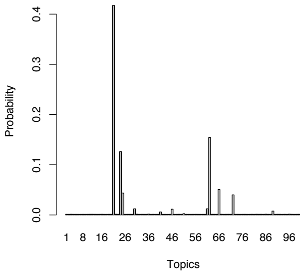
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

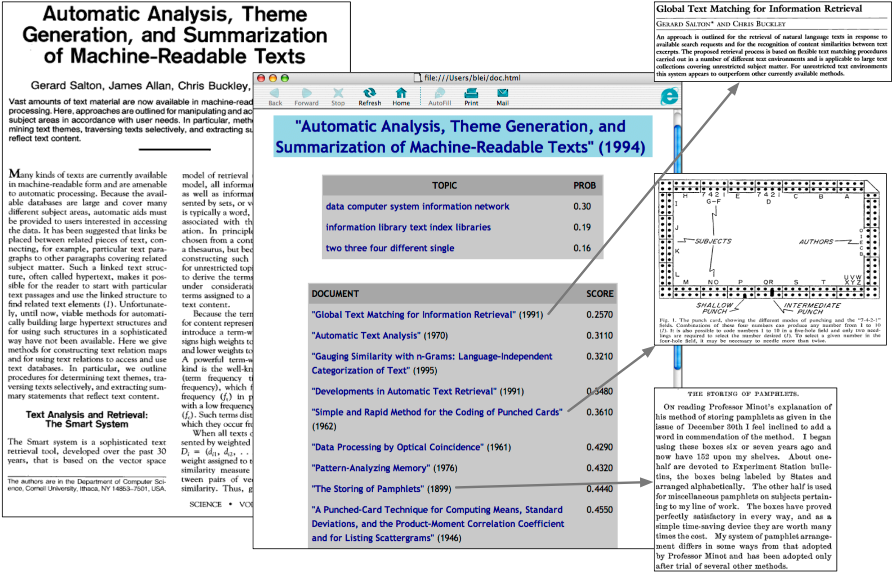
SCIENCE • VOL. 272 • 24 MAY 1996



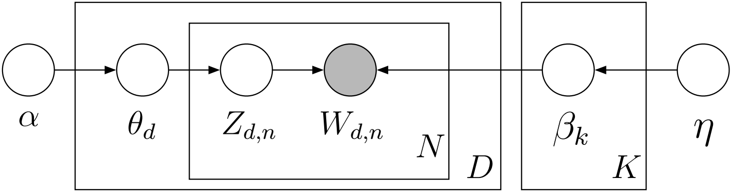
Example inference

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

Used in exploratory tools of document collections

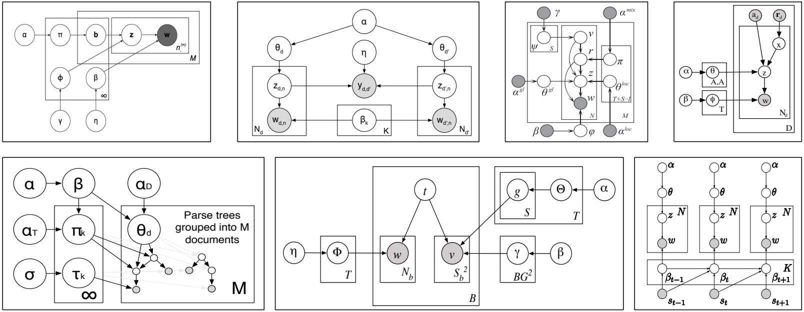


Summary of LDA



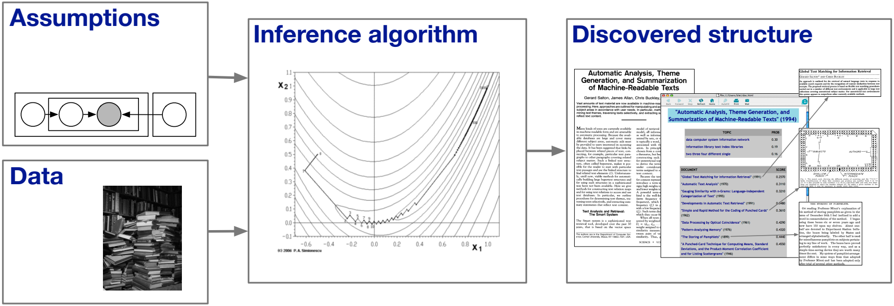
- LDA can
  - visualize the hidden thematic structure in large corpora
  - generalize new data to fit into that structure
- Builds on Deerwester et al. (1990) and Hofmann (1999)
  It is an example of a *mixed membership model* (Erosheva, 2004)
  Relates to *multinomial PCA* (Jakulin and Buntine, 2002).
- 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, humanties, industry, and culture.
- LDA can be embedded in more complicated models.
- Algorithmic improvements let us fit models to massive data.

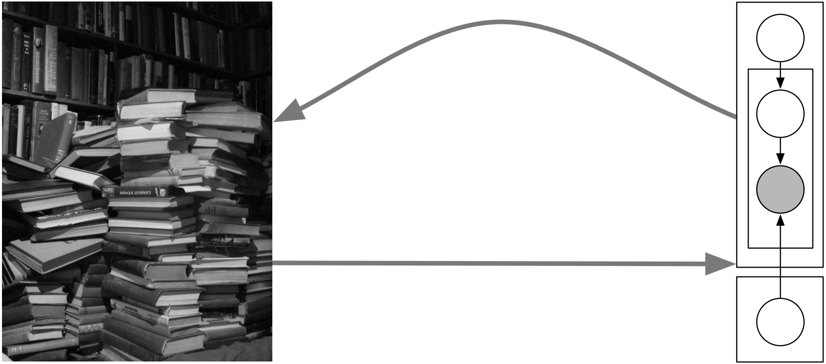
Bigger Picture: Probabilistic modeling



- 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

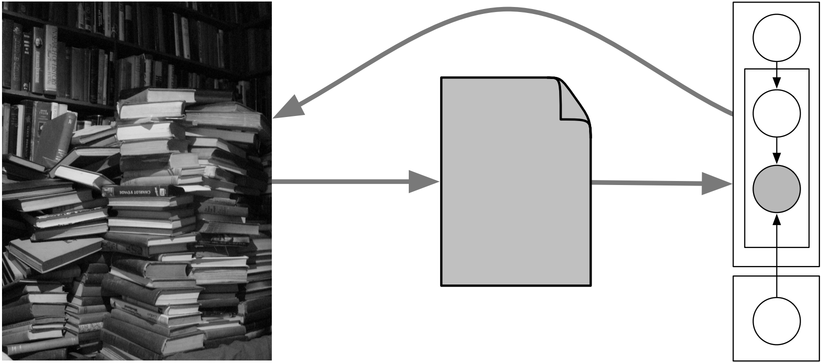


Analyze the collection

Update the model

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

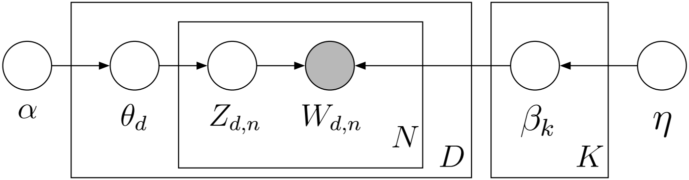
We need online inference



Sample one document ***Analyze it*** ***Update the model***

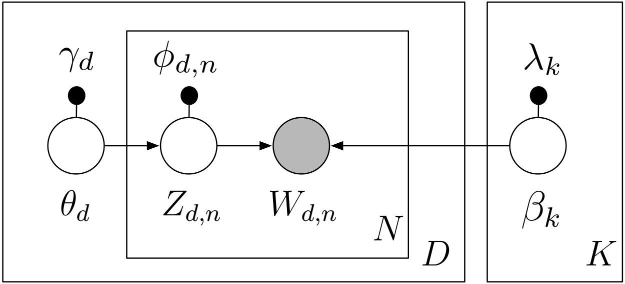
- Allows us to analyze millions of documents
- Lets us develop topic models on streaming collections

Computation with LDA



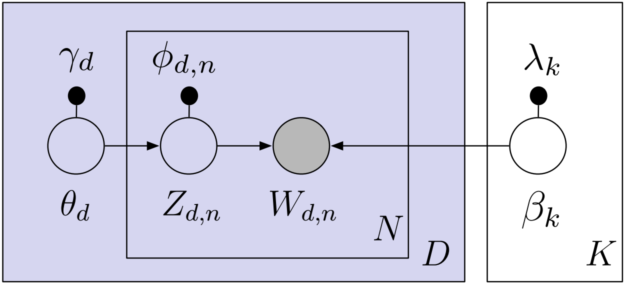
- Our goal is to compute the *posterior*, the conditional distribution of the hidden variables given the documents.
- We will build on *variational inference*.
  - Posit a parameterized distribution *q* over hidden variables.
  - Optimize to make *q* close (in KL) to the posterior.

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



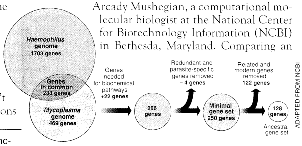
- In the "local step" we iteratively update the parameters for each document, holding the topic parameters fixed.

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.

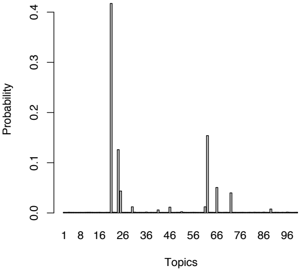
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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



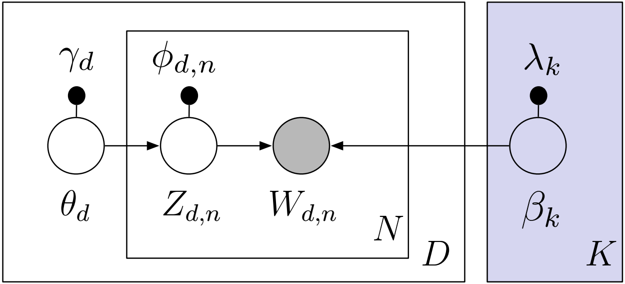
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996



Batch variational inference for LDA

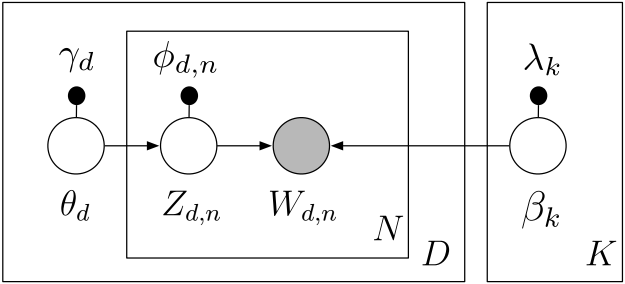


- In the "global step" we aggregate the parameters computed from the local step and update the parameters for the topics.

Example topic inference

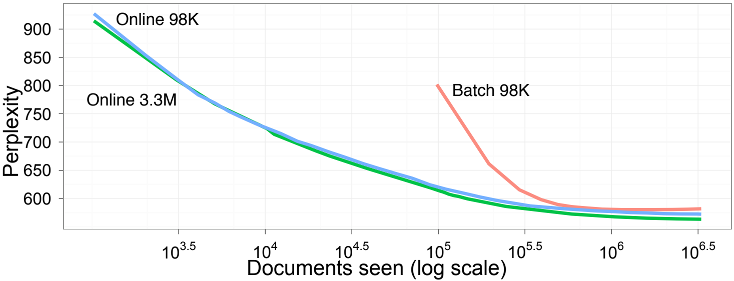
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	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	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

Why does this work?

A STOCHASTIC APPROXIMATION METHOD[1]

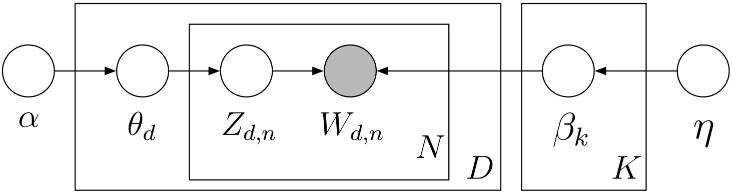
BY HERBERT ROBBINS AND SUTTON MONRO

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 α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

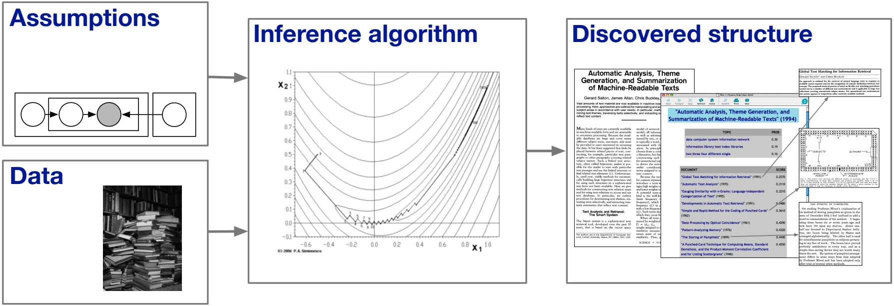
- Why waste time with the real gradient, when a cheaper noisy estimate of the gradient will do (Robbins and Monro, 1951)?
- Idea: Follow a noisy estimate of the gradient with a step-size.
- By decreasing the step-size according to a certain schedule, we guarantee convergence to a local optimum.
- See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

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

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