Probabilistic Topic Models

David M. Blei

Department of Computer Science
Princeton University

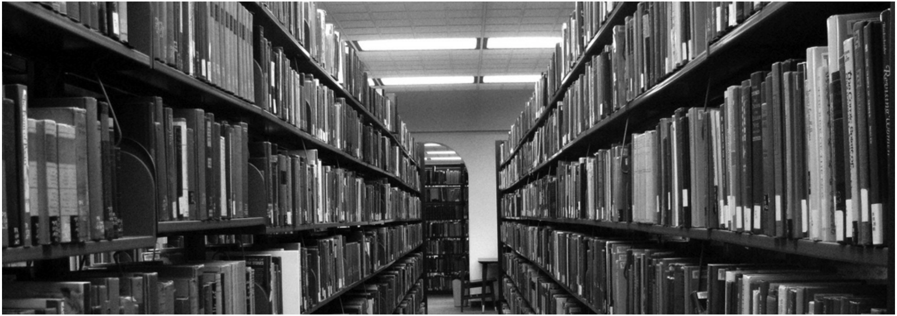
August 22, 2011

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

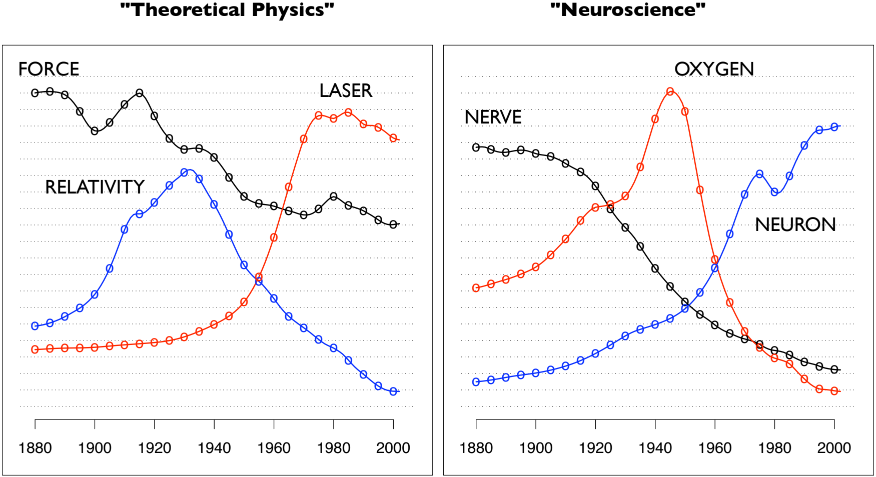
human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

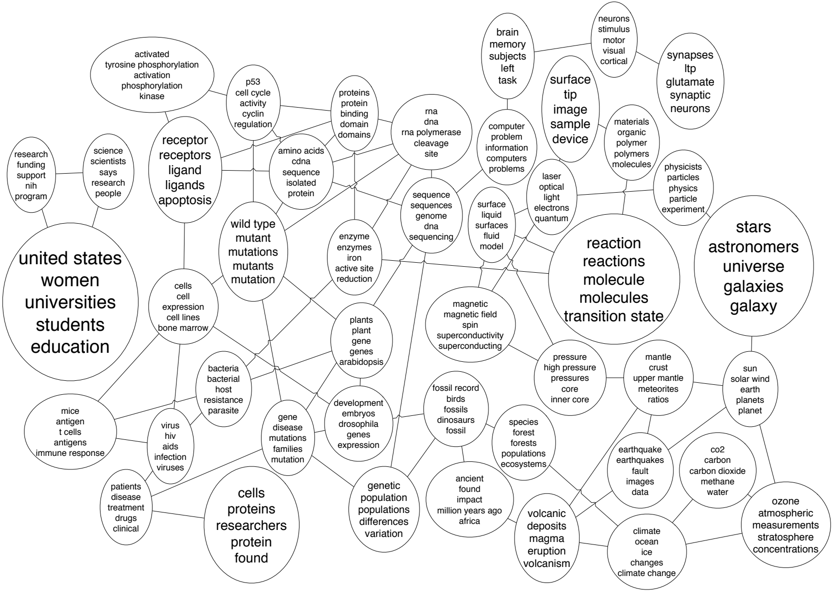
disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

Model the evolution of topics over time



Model connections between topics



brain
memory
subjects
left
task

neurons
stimulus
motor
visual
cortical

synapses
ltp
glutamate
synaptic
neurons

activated
tyrosine phosphorylation
activation
phosphorylation
kinase

p53
cell cycle
activity
cyclin
regulation

proteins
protein
binding
domain
domains

rna
dna
rna polymerase
cleavage
site

surface
tip
image
sample
device

materials
organic
polymer
polymers
molecules

computer
problem
information
computers
problems

research
funding
support
nih
program

science
scientists
says
research
people

receptor
receptors
ligand
ligands
apoptosis

amino acids
cdna
sequence
isolated
protein

physicists
particles
physics
particle
experiment

laser
optical
light
electrons
quantum

sequence
sequences
genome
dna
sequencing

surface
liquid
surfaces
fluid
model

wild type
mutant
mutations
mutants
mutation

enzyme
enzymes
iron
active site
reduction

stars
astronomers
universe
galaxies
galaxy

reaction
reactions
molecule
molecules
transition state

united states
women
universities
students
education

cells
cell
expression
cell lines
bone marrow

magnetic
magnetic field
spin
superconductivity
superconducting

plants
plant
gene
genes
arabidopsis

pressure
high pressure
pressures
core
inner core

mantle
crust
upper mantle
meteorites
ratios

sun
solar wind
earth
planets
planet

bacteria
bacterial
host
resistance
parasite

fossil record
birds
fossils
dinosaurs
fossil

development
embryos
drosophila
genes
expression

mice
antigen
t cells
antigens
immune response

virus
hiv
aids
infection
viruses

gene
disease
mutations
families
mutation

species
forest
forests
populations
ecosystems

earthquake
earthquakes
fault
images
data

co2
carbon
carbon dioxide
methane
water

patients
disease
treatment
drugs
clinical

cells
proteins
researchers
protein
found

genetic
population
populations
differences
variation

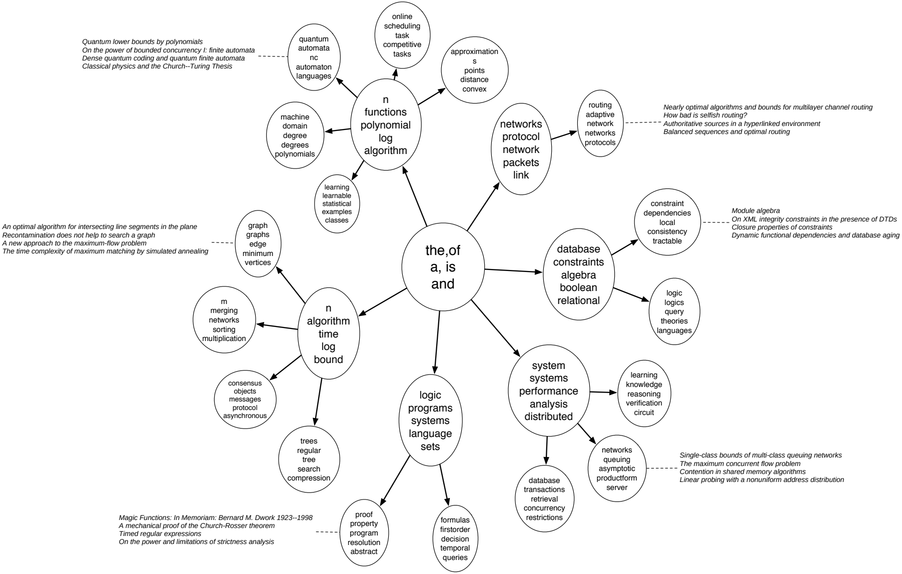
ancient
found
impact
million years ago
africa

volcanic
deposits
magma
eruption
volcanism

climate
ocean
ice
changes
climate change

ozone
atmospheric
measurements
stratosphere
concentrations

Find hierarchies of topics



the,of
a, is
and

n
functions
polynomial
log
algorithm

quantum
automata
nc
automaton
languages

Quantum lower bounds by polynomials
On the power of bounded concurrency I: finite automata
Dense quantum coding and quantum finite automata
Classical physics and the Church--Turing Thesis

online
scheduling
task
competitive
tasks

approximation
s
points
distance
convex

machine
domain
degree
degrees
polynomials

learning
learnable
statistical
examples
classes

networks
protocol
network
packets
link

routing
adaptive
network
networks
protocols

Nearly optimal algorithms and bounds for multilayer channel routing
How bad is selfish routing?
Authoritative sources in a hyperlinked environment
Balanced sequences and optimal routing

database
constraints
algebra
boolean
relational

constraint
dependencies
local
consistency
tractable

Module algebra
On XML integrity constraints in the presence of DTDs
Closure properties of constraints
Dynamic functional dependencies and database aging

logic
logics
query
theories
languages

n
algorithm
time
log
bound

graph
graphs
edge
minimum
vertices

An optimal algorithm for intersecting line segments in the plane
Recontamination does not help to search a graph
A new approach to the maximum-flow problem
The time complexity of maximum matching by simulated annealing

m
merging
networks
sorting
multiplication

consensus
objects
messages
protocol
asynchronous

trees
regular
tree
search
compression

logic
programs
systems
language
sets

proof
property
program
resolution
abstract

Magic Functions: In Memoriam: Bernard M. Dwork 1923--1998
A mechanical proof of the Church-Rosser theorem
Timed regular expressions
On the power and limitations of strictness analysis

formulas
firstorder
decision
temporal
queries

system
systems
performance
analysis
distributed

learning
knowledge
reasoning
verification
circuit

networks
queuing
asymptotic
productform
server

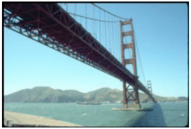
Single-class bounds of multi-class queuing networks
The maximum concurrent flow problem
Contention in shared memory algorithms
Linear probing with a nonuniform address distribution

database
transactions
retrieval
concurrency
restrictions

Annotate images

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE



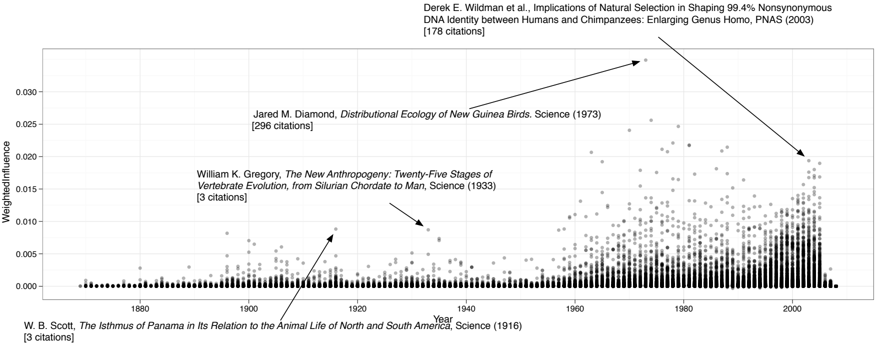
SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

Discover influential articles



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]

Jared M. Diamond, *Distributional Ecology of New Guinea Birds.* Science (1973)
[296 citations]

William K. Gregory, *The New Anthropogeny: Twenty-Five Stages of Vertebrate Evolution, from Silurian Chordate to Man*, Science (1933)
[3 citations]

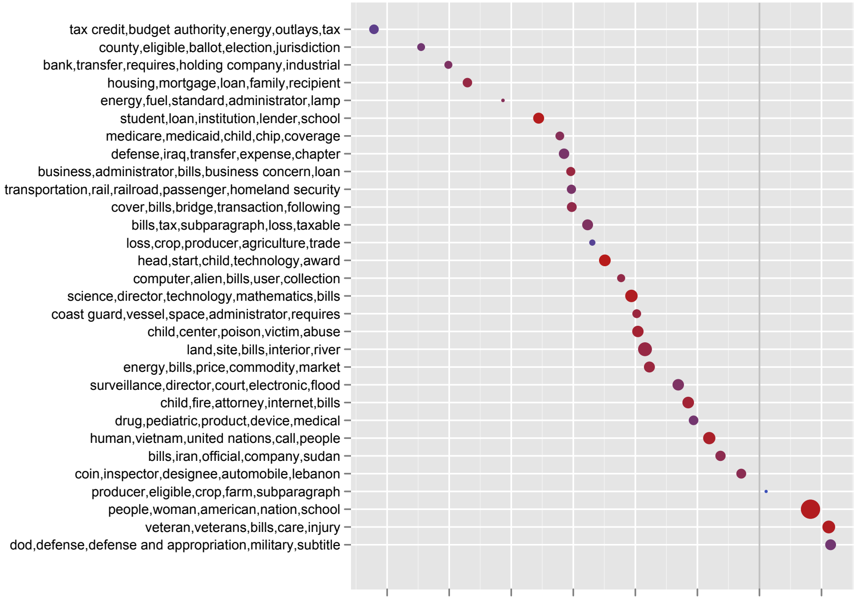
W. B. Scott, *The Isthmus of Panama in Its Relation to the Animal Life of North and South America*, Science (1916)
[3 citations]

Predict links between articles

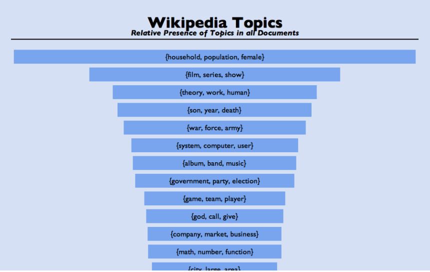
Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo	RTM (ψ_e)
Rates of convergence of the Hastings and Metropolis algorithms	
Possible biases induced by MCMC convergence diagnostics	
Bounding convergence time of the Gibbs sampler in Bayesian image restoration	
Self regenerative Markov chain Monte Carlo	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Rate of Convergence of the Gibbs Sampler by Gaussian Approximation	
Diagnosing convergence of Markov chain Monte Carlo algorithms	

Exact Bound for the Convergence of Metropolis Chains	LDA + Regression
Self regenerative Markov chain Monte Carlo	
Minorization conditions and convergence rates for Markov chain Monte Carlo	
Gibbs-markov models	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models	
Mediating instrumental variables	
A qualitative framework for probabilistic inference	
Adaptation for Self Regenerative MCMC	

Characterize political decisions



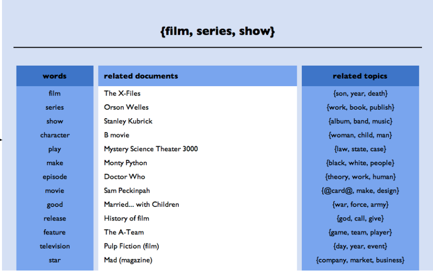
Organize and browse large corpora



Wikipedia Topics

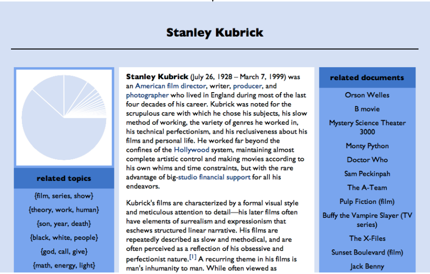
Relative Presence of Topics in all Documents

- {household, population, female}
- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {war, force, army}
- {system, computer, user}
- {album, band, music}
- {government, party, election}
- {game, team, player}
- {god, call, give}
- {company, market, business}
- {math, number, function}
- {city, large, area}



{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

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 method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far 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.

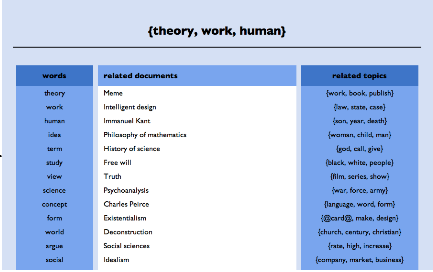
Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection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

related topics

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

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 Benny



{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}

This tutorial

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

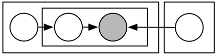
Related subjects

Topic modeling is a case study in modern machine learning with probabilistic models. 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

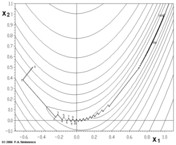
If you remember one picture...

Assumptions

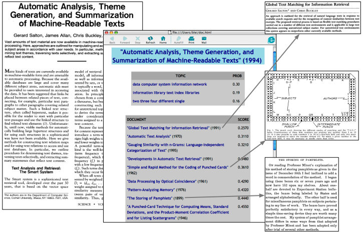


Data

Inference algorithm



Discovered structure



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

Gerard Salton, James Allan, Chris Buckley,

file:///Users/blei/doc.html

"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

THE STORING OF PAMPHLETS.

Organization

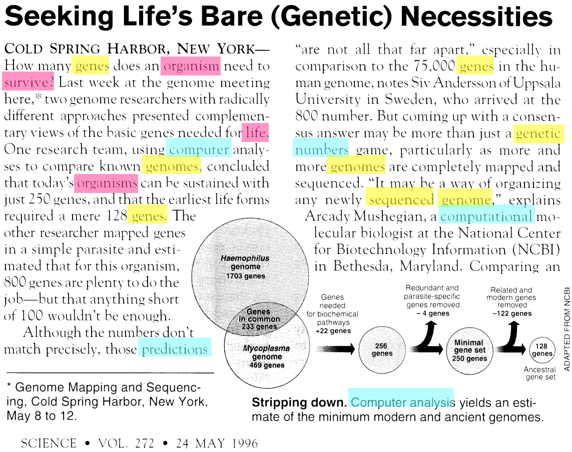
- **Introduction to topic modeling**
  - Latent Dirichlet allocation
  - Open source implementations and tools
- **Beyond latent Dirichlet allocation**
  - Modeling richer assumptions
  - Supervised topic modeling
  - Bayesian nonparametric topic modeling
- **Algorithms**
  - Gibbs sampling
  - Variational inference
  - Online variational inference
- **Discussion, open questions, and resources**

Introduction to Topic Modeling

Probabilistic modeling

1. Data are assumed to be observed from a generative probabilistic process that includes hidden variables.
   - *In text, the hidden variables are the thematic structure.*
2. Infer the hidden structure using posterior inference
   - *What are the topics that describe this collection?*
3. Situate new data into the estimated model.
   - *How does a new document fit into the topic structure?*

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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

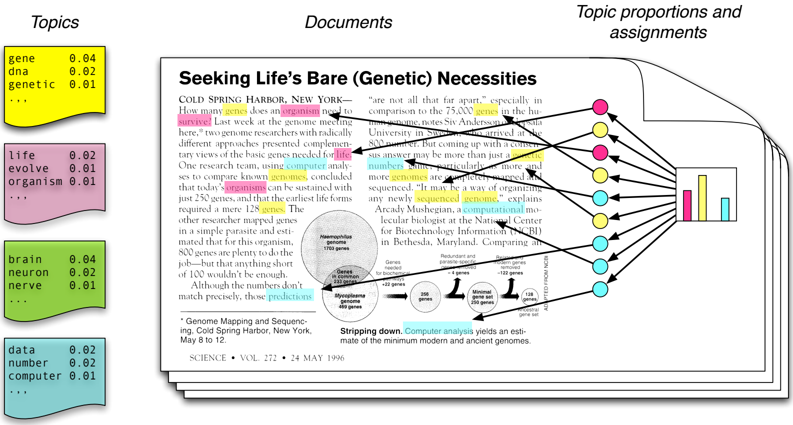
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



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
.,,	

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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

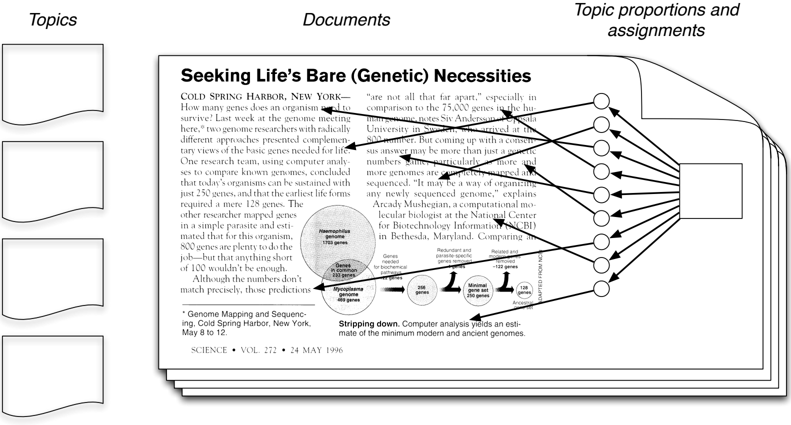
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- 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



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways 22 genes
256 genes
Redundant and parasite-specific genes removed –4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

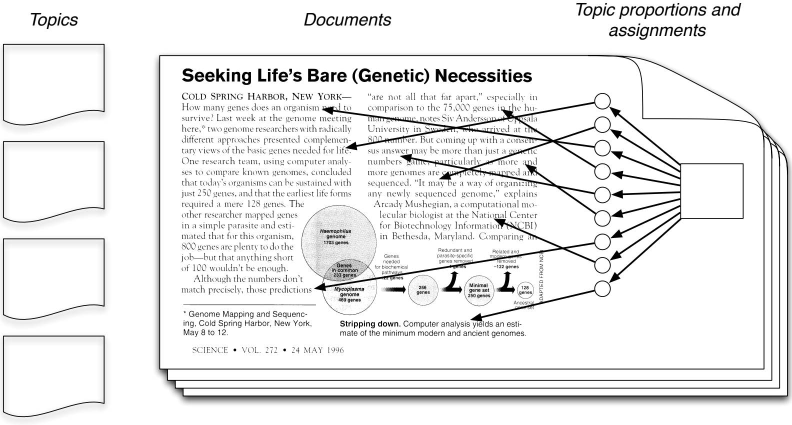
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

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

The posterior distribution


Topics
Documents
Topic proportions and assignments
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways
22 genes
256 genes
Redundant and parasite-specific genes removed
genes
Minimal gene set 250 genes
Related and modern genes removed
−122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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.

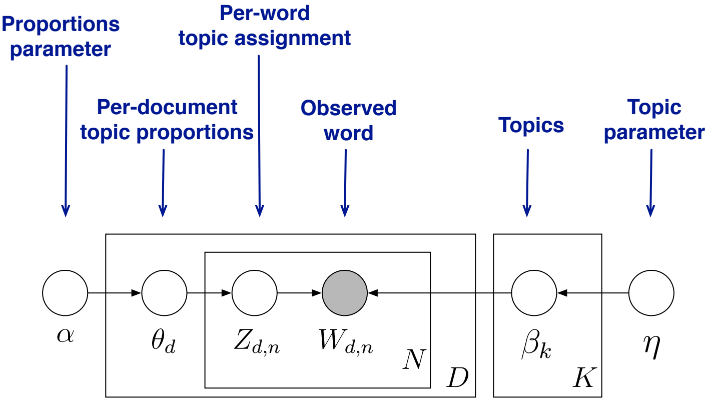
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

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

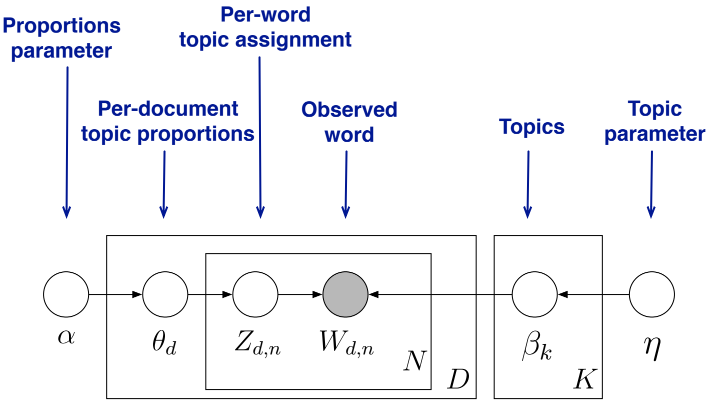
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

LDA as a graphical model



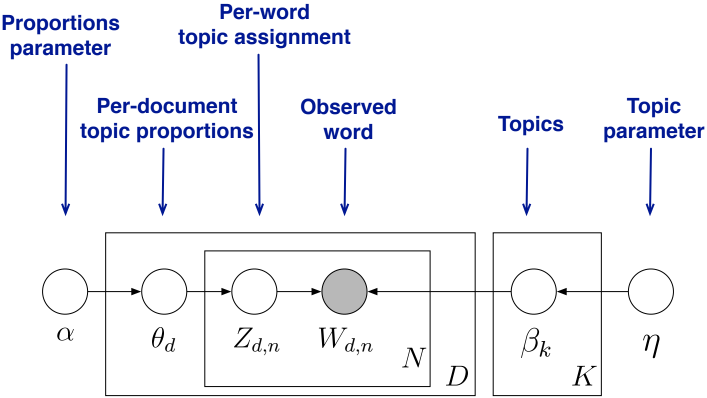
- Encodes our assumptions about the data
- Connects to algorithms for computing with data
- See *Pattern Recognition and Machine Learning* (Bishop, 2006).

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



Proportions parameter

Per-word topic assignment

Per-document topic proportions

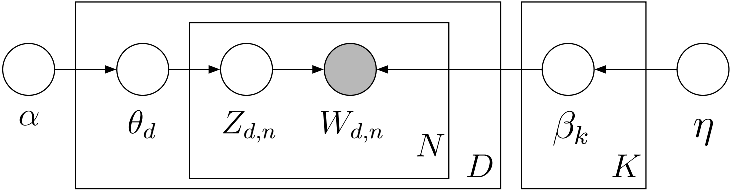
Observed word

Topics

Topic parameter

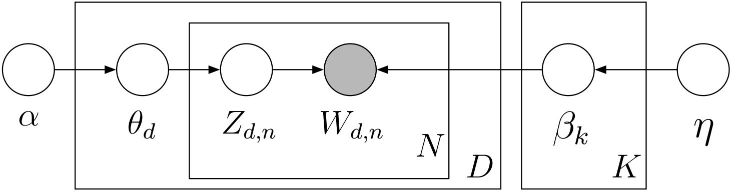
$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

LDA



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

LDA

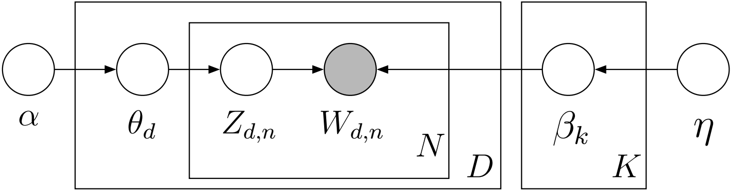


Approximate posterior inference algorithms

- 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).

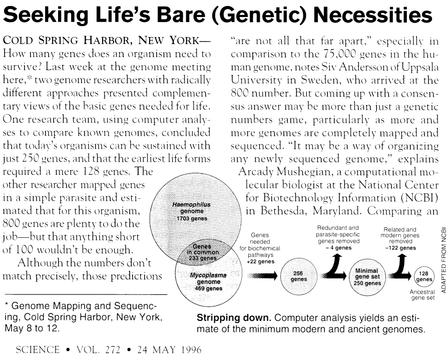
Example inference



- **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

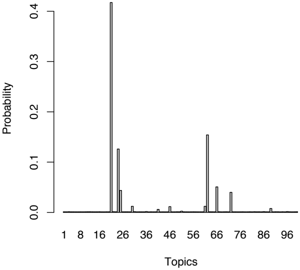
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

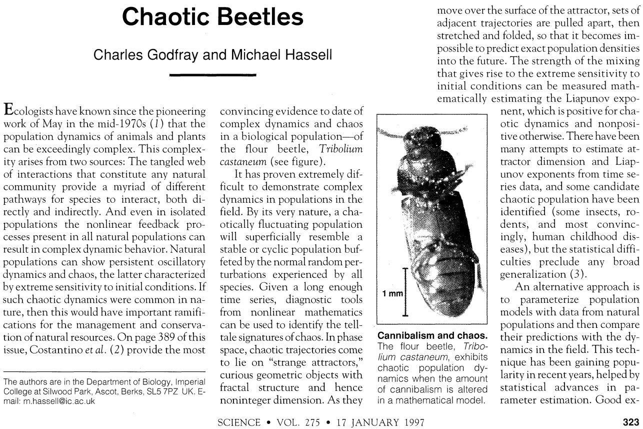
human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

Example inference (II)



Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).

Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

Example inference (II)

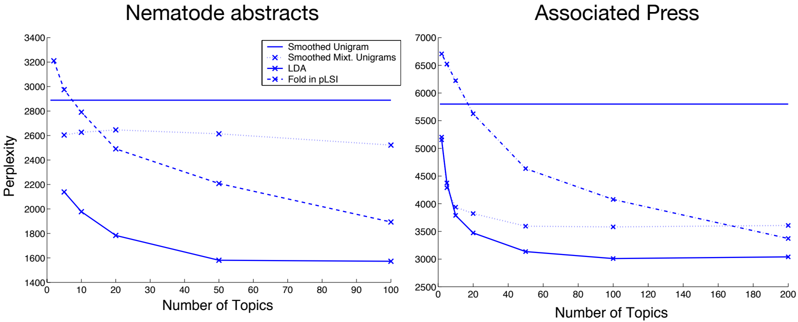
problem
problems
mathematical
number
new
mathematics
university
two
first
numbers
work
time
mathematicians
chaos
chaotic

model
rate
constant
distribution
time
number
size
values
value
average
rates
data
density
measured
models

selection
male
males
females
sex
species
female
evolution
populations
population
sexual
behavior
evolutionary
genetic
reproductive

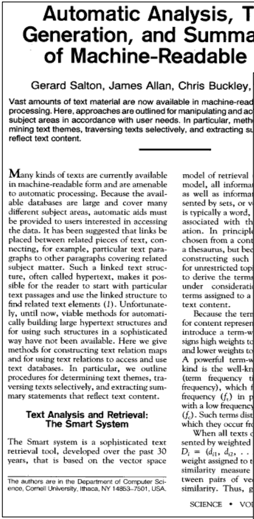
species
forest
ecology
fish
ecological
conservation
diversity
population
natural
ecosystems
populations
endangered
tropical
forests
ecosystem

Held out perplexity



$$\text{perplexity} = \exp\left\{\frac{-\sum_d \log p(\mathbf{w}_d)}{\sum_d N_d}\right\}$$

Used to explore and browse 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-read
processing. Here, approaches are outlined for manipulating and ac
subject areas in accordance with user needs. In particular, meth
mining text themes, traversing texts selectively, and extracting su
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

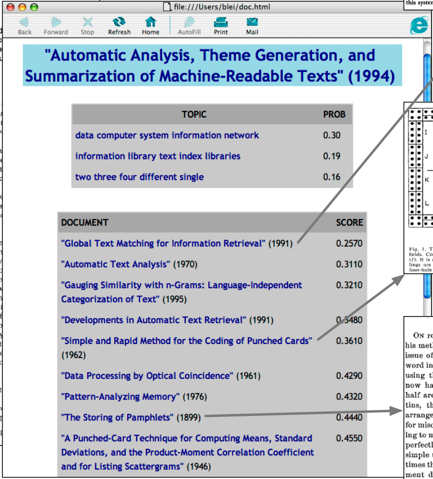
model of retrieval
model, all informa
as well as informat
sented by sets, or v
is typically a word,
associated with th
ation. In principle
chosen from a cont
a thesaurus, but bec
constructing such
for unrestricted top
to derive the terms
under consideratio
terms assigned to a
text content.

Because the term
for content represen
introduce a term-w
signs high weights to
and lower weights to
A powerful term-w
kind is the well-kn
(term frequency ti
frequency), which f
frequency (f_t) in p
with a low frequency
(f_c). Such terms dist
which they occur fr

When all texts c
sented by weighted
$D_i = (d_{i1}, d_{i2}, \ldots$
weight assigned to t
similarity measure
tween pairs of ve
similarity. Thus, g

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

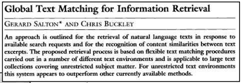
SCIENCE • VO



"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.

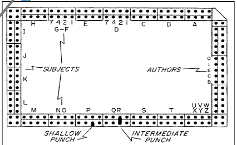
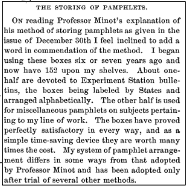


Fig. 1. The punch card, showing the different modes of punching and the "7-4-2-1" fields. Combinations of these four numbers can produce any number from 1 to 10 (1). It is also possible to code numbers 1 to 10 in a five-hole field and only two needlings are required to select the number desired (1). To select a given number in the four-hole field, it may be necessary to needle more than twice.



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.

Aside: The Dirichlet distribution

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

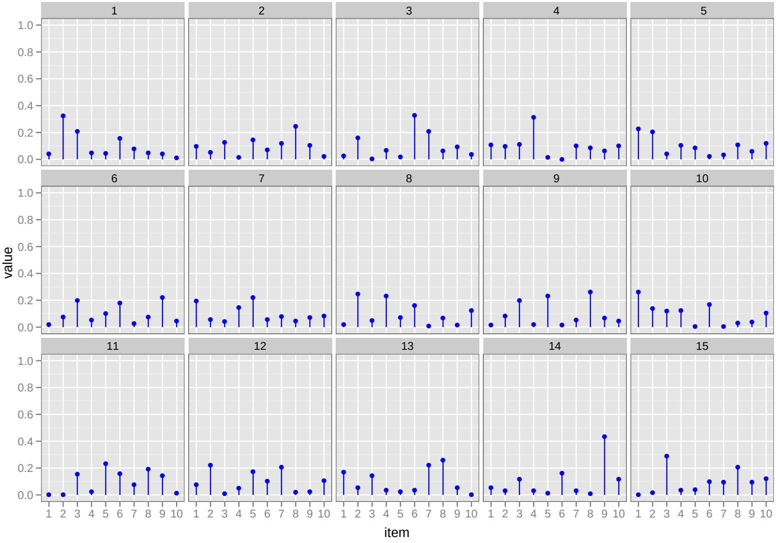
$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

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

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

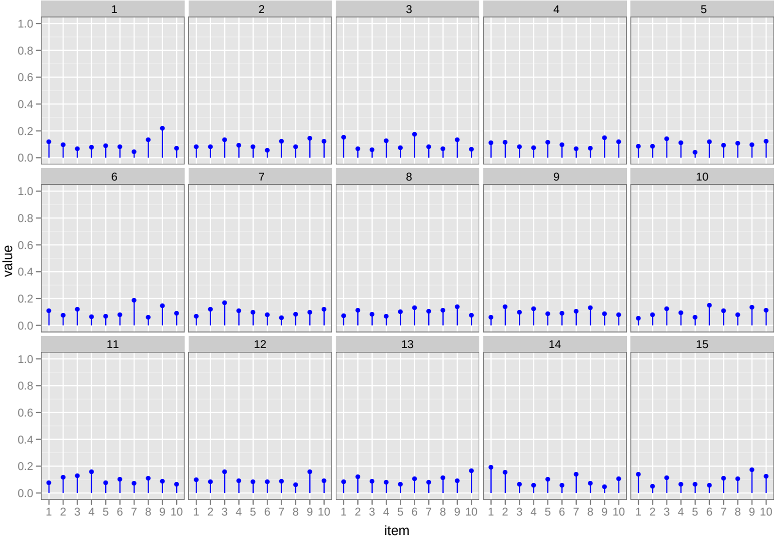
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

$\alpha = 1$

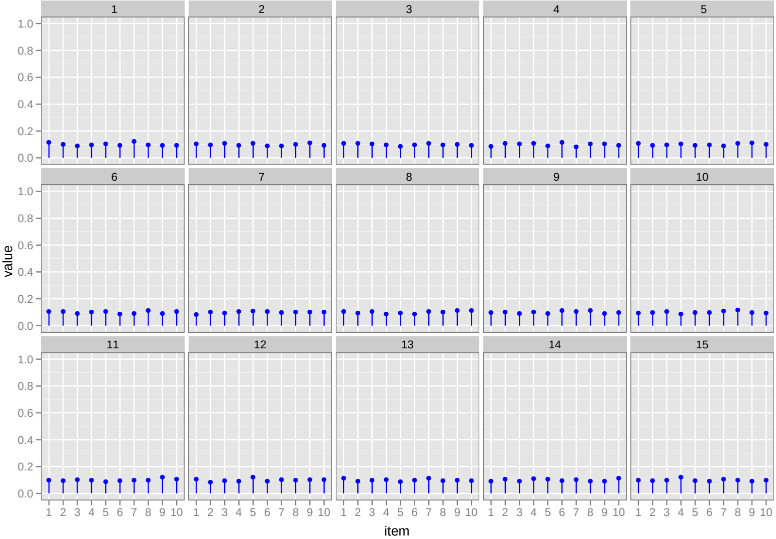


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
0.0
0.2
0.4
0.6
0.8
1.0
item
1 2 3 4 5 6 7 8 9 10

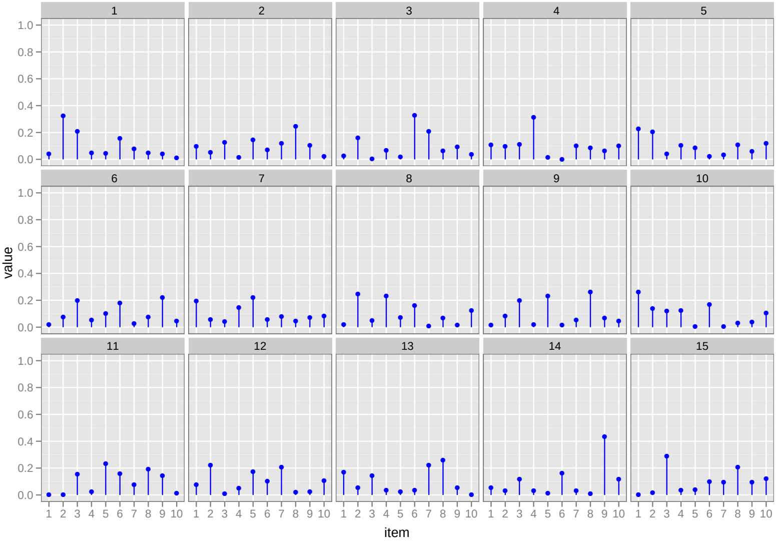
$\alpha = 10$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

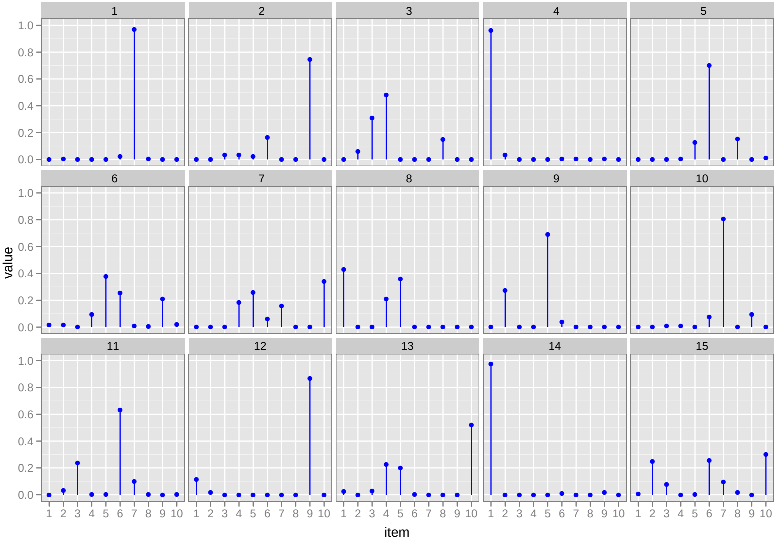
$\alpha = 100$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1.0
0.8
0.6
0.4
0.2
0.0
value
1 2 3 4 5 6 7 8 9 10
item

$\alpha = 1$

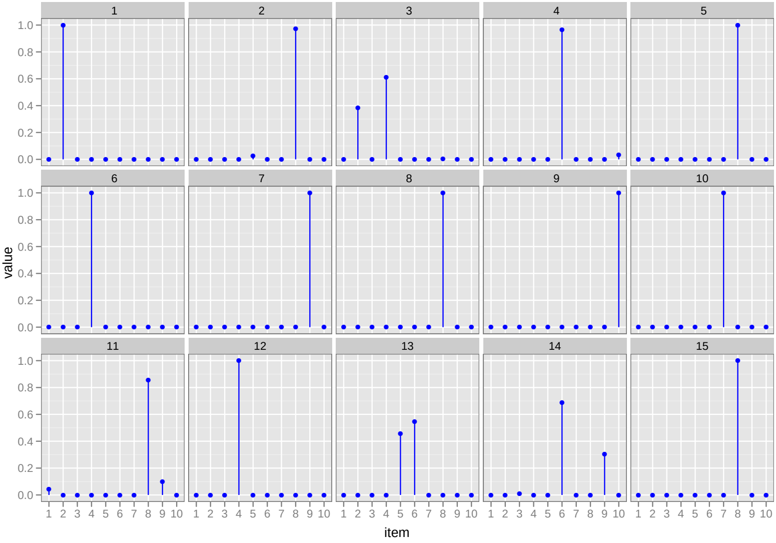

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1
2
3
4
5
6
7
8
9
10

$\alpha = 0.1$



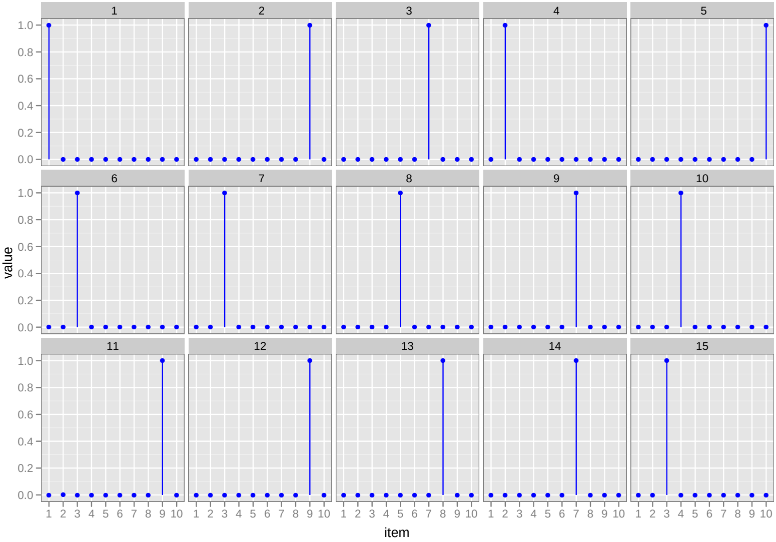
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

$\alpha = 0.01$



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1.0
0.8
0.6
0.4
0.2
0.0
value
1 2 3 4 5 6 7 8 9 10
item

$\alpha = 0.001$

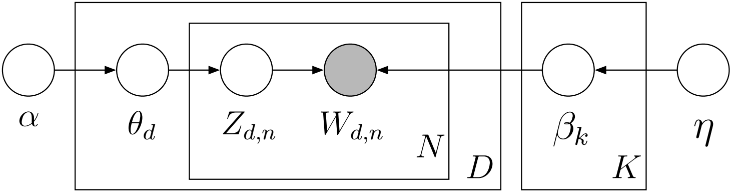

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

Why does LDA “work”?

Why does the LDA posterior put “topical” words together?

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

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 a *mixed membership model* (Erosheva, 2004).
  Relates to *multinomial PCA* (Jakulin and Buntine, 2002)

- Was independently invented for genetics (Pritchard et al., 2000)

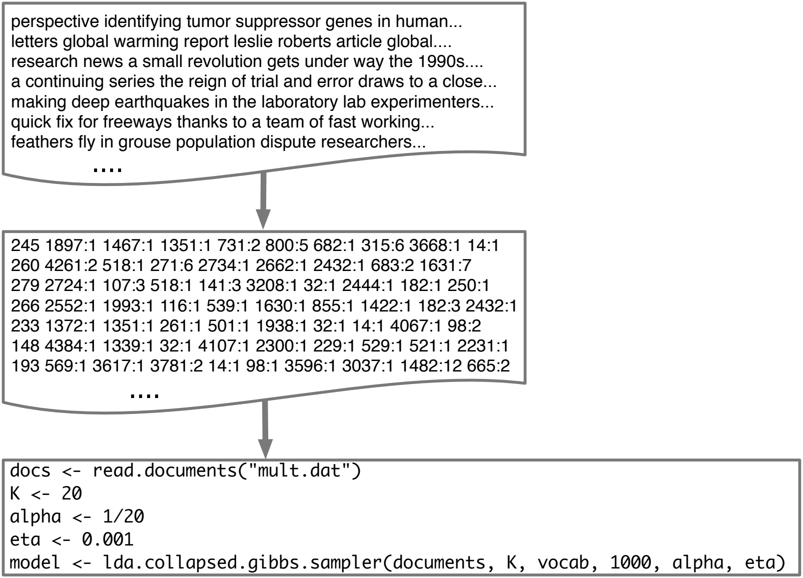
Implementations of LDA

There are many available implementations of topic modeling—

LDA-C*	A C implementation of LDA
HDP*	A C implementation of the HDP (“infinite LDA”)
Online LDA*	A python package for LDA on massive data
LDA in R*	Package in R for many topic models
LingPipe	Java toolkit for NLP and computational linguistics
Mallet	Java toolkit for statistical NLP
TMVE*	A python package to build browsers from topic models

* available at www.cs.princeton.edu/∼blei/

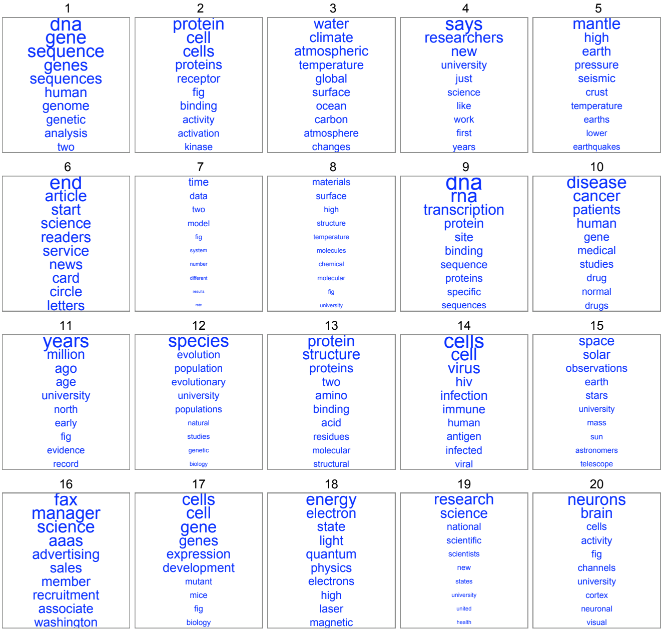
Example: LDA in R (Jonathan Chang)



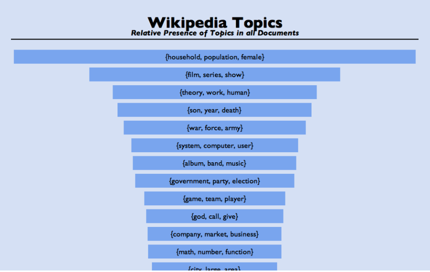
perspective identifying tumor suppressor genes in human...
letters global warming report leslie roberts article global....
research news a small revolution gets under way the 1990s....
a continuing series the reign of trial and error draws to a close...
making deep earthquakes in the laboratory lab experimenters...
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
rate
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

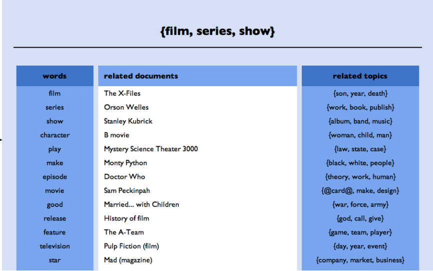
Open source document browser (with Allison Chaney)



Wikipedia Topics

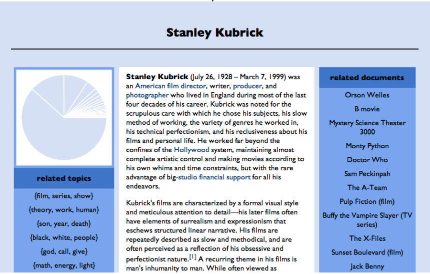
Relative Presence of Topics in all Documents

- {household, population, female}
- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {war, force, army}
- {system, computer, user}
- {album, band, music}
- {government, party, election}
- {game, team, player}
- {god, call, give}
- {company, market, business}
- {math, number, function}
- {city, large, area}



{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

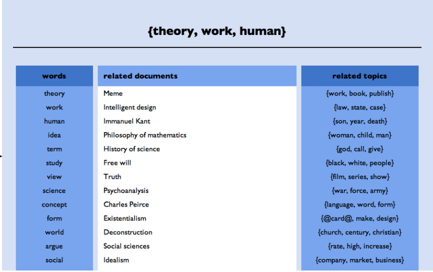
- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {black, white, people}
- {god, call, give}
- {math, energy, light}

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 method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far 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 eschews 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.[1] A recurring theme in his films is man's inhumanity to man. While often viewed as

related documents

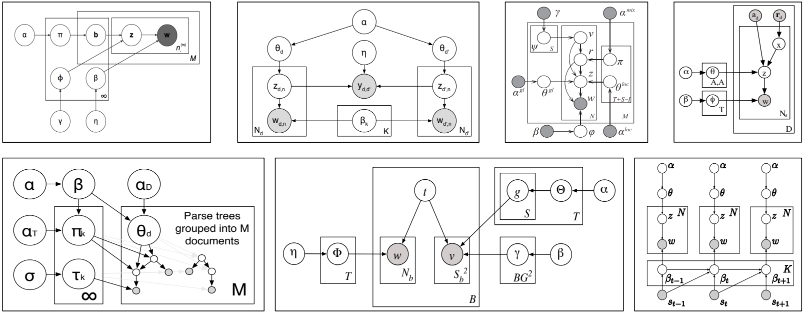
- Orson Welles
- B movie
- Mystery Science Theater 3000
- Monty Python
- Doctor Who
- Sam Peckinpah
- The A-Team
- Pulp Fiction (film)
- Buffy the Vampire Slayer (TV series)
- The X-Files
- Sunset Boulevard (film)
- Jack Benny



{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}

Why develop these kinds of models?



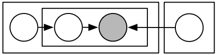
- Organizing and finding patterns in data has become important in the sciences, humanties, industry, and culture.

- LDA can be embedded in more complicated models that capture richer assumptions about the data.

- Algorithmic improvements let us fit models to massive data.

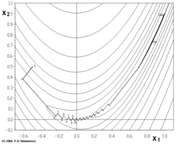
Bigger Picture: Probabilistic modeling

Assumptions

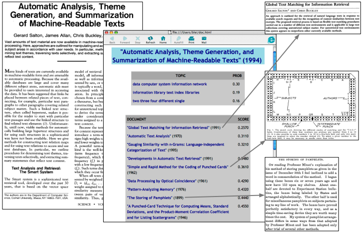


Data

Inference algorithm



Discovered structure



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-read processing. Here, approaches are outlined for manipulating and ac subject areas in accordance with user needs. In particular, meth mining text themes, traversing texts selectively, and extracting su 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

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

SCIENCE • VO

"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.

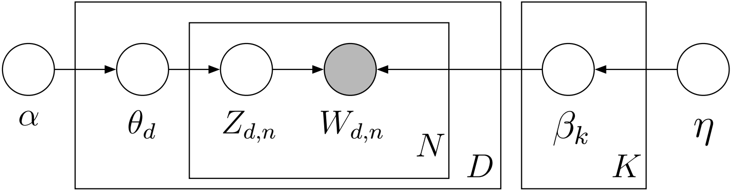
THE STORING OF PAMPHLETS.

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.

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

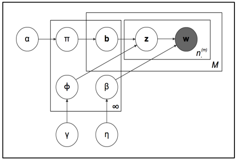
Beyond Latent Dirichlet Allocation

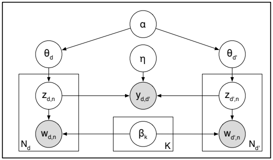
So far...

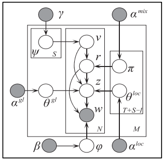


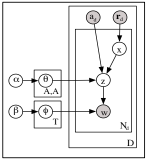
- LDA is a simple topic model
- Can be used to find topics that describe a corpus
- Each document exhibits multiple topics
- How can we build on this simple model of text?

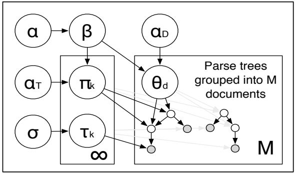
LDA is extendible

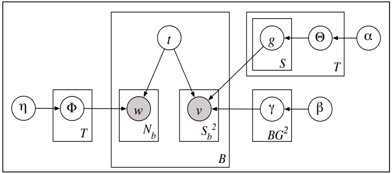


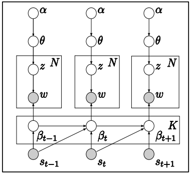






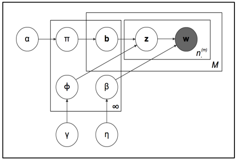


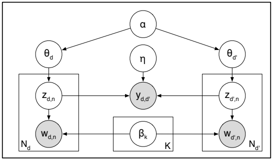


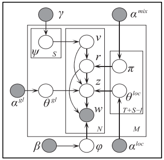


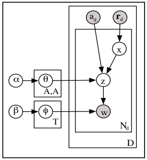
- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
- LDA models can include syntax, authorship, word sense, dynamics, correlation, hierarchies, ...

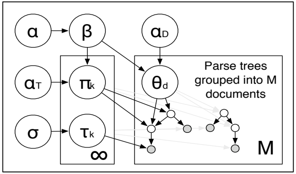
LDA is extendible

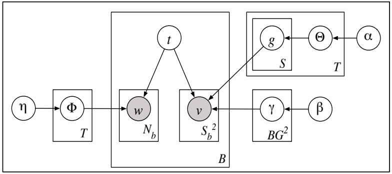


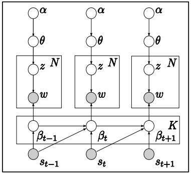






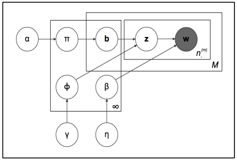


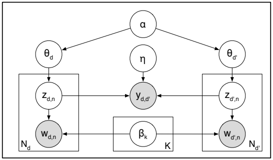


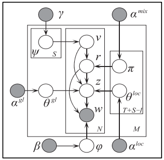


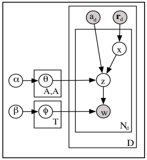
- The **data generating distribution** can be changed.
- LDA models can be built for images, social networks, music, purchase histories, computer code, genetic data, click-through-data, neural spike trains, ...

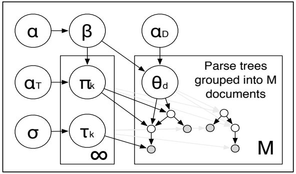
LDA is extendible

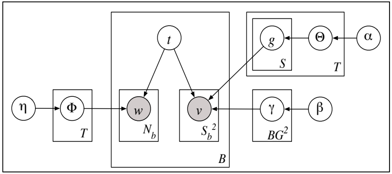


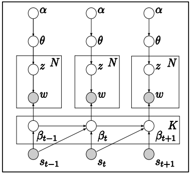












- The **LDA posterior** can be used in creative ways

- It can be used for information retrieval, collaborative filtering, document similarity, visualization, ...

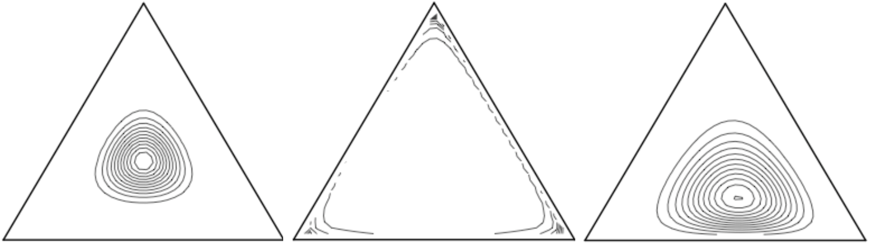
Beyond latent Dirichlet allocation

- Modeling richer assumptions
    - Correlated topic models
    - Dynamic topic models
    - Measuring scholarly impact
- Supervised topic models
    - Supervised LDA
    - Relational topic models
    - Ideal point topic models
- Bayesian nonparametric topic models

Modeling richer assumptions

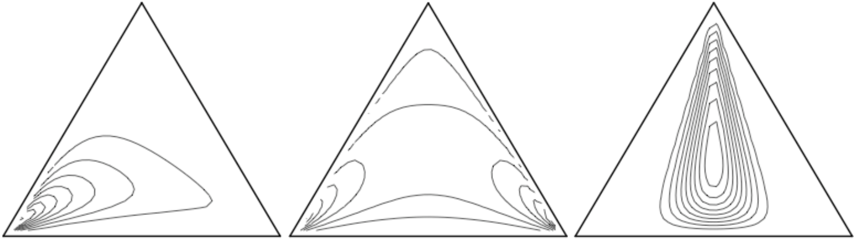
- Correlated topic models
- Dynamic topic models
- Measuring scholarly impact

The hidden assumptions of the Dirichlet



- The Dirichlet is an exponential family distribution on the *simplex*, positive vectors that sum to one.
- However, the near independence of components makes it a poor choice for modeling topic proportions.
- An article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

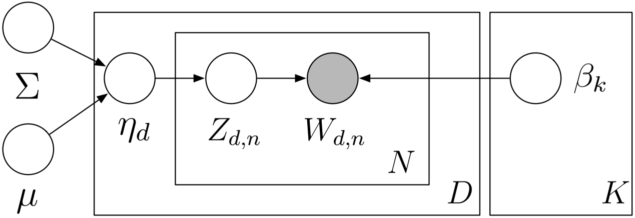
The logistic normal distribution



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

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

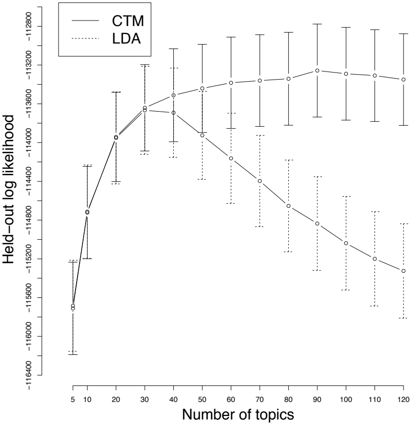
The correlated topic model (CTM) (Blei and Lafferty, 2007)

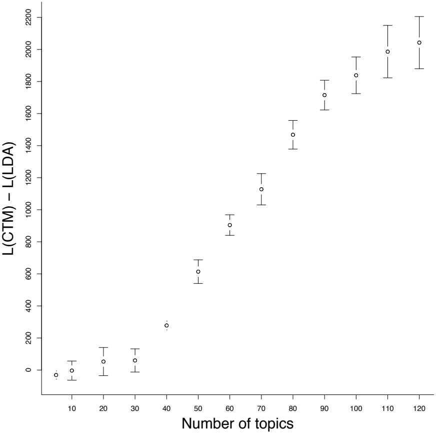


Noconjugate prior on topic proportions

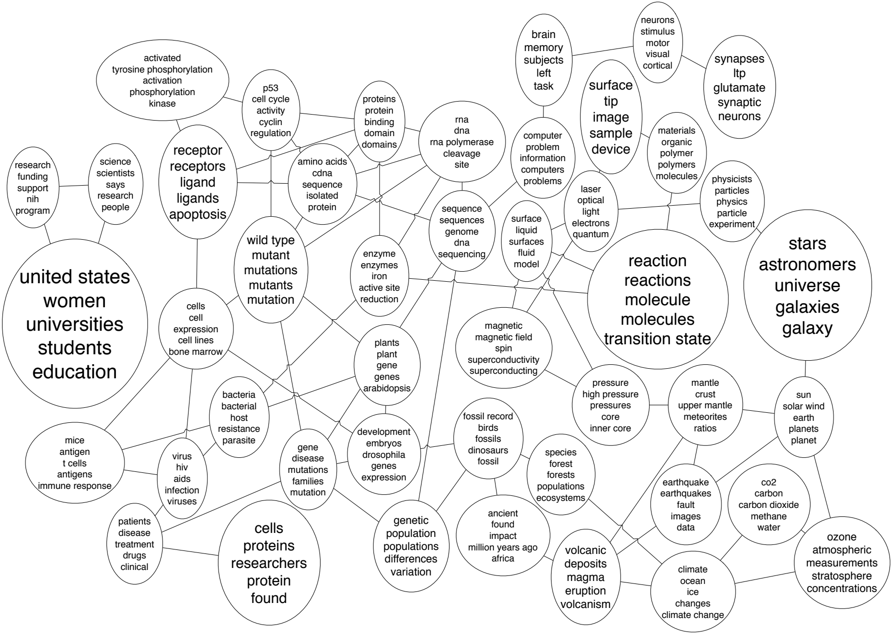
- Draw topic proportions from a logistic normal, where topic occurrences can exhibit correlation.
- Use for:
  - Providing a “map” of topics and how they are related
  - Better prediction via correlated topics

Held out log probability in a CTM





- Analyzed held-out log probability on *Science*, 1960.
- CTM supports more topics and provides a better fit than LDA.


activated
tyrosine phosphorylation
activation
phosphorylation
kinase
p53
cell cycle
activity
cyclin
regulation
proteins
protein
binding
domain
domains
rna
dna
rna polymerase
cleavage
site
brain
memory
subjects
left
task
neurons
stimulus
motor
visual
cortical
synapses
ltp
glutamate
synaptic
neurons
surface
tip
image
sample
device
materials
organic
polymer
polymers
molecules
computer
problem
information
computers
problems
research
funding
support
nih
program
science
scientists
says
research
people
receptor
receptors
ligand
ligands
apoptosis
amino acids
cdna
sequence
isolated
protein
sequence
sequences
genome
dna
sequencing
laser
optical
light
electrons
quantum
physicists
particles
physics
particle
experiment
surface
liquid
surfaces
fluid
model
wild type
mutant
mutations
mutants
mutation
enzyme
enzymes
iron
active site
reduction
reaction
reactions
molecule
molecules
transition state
stars
astronomers
universe
galaxies
galaxy
united states
women
universities
students
education
cells
cell
expression
cell lines
bone marrow
magnetic
magnetic field
spin
superconductivity
superconducting
plants
plant
gene
genes
arabidopsis
pressure
high pressure
pressures
core
inner core
mantle
crust
upper mantle
meteorites
ratios
sun
solar wind
earth
planets
planet
bacteria
bacterial
host
resistance
parasite
development
embryos
drosophila
genes
expression
fossil record
birds
fossils
dinosaurs
fossil
species
forest
forests
populations
ecosystems
mice
antigen
t cells
antigens
immune response
virus
hiv
aids
infection
viruses
gene
disease
mutations
families
mutation
earthquake
earthquakes
fault
images
data
co2
carbon
carbon dioxide
methane
water
ozone
atmospheric
measurements
stratosphere
concentrations
patients
disease
treatment
drugs
clinical
cells
proteins
researchers
protein
found
genetic
population
populations
differences
variation
ancient
found
impact
million years ago
africa
volcanic
deposits
magma
eruption
volcanism
climate
ocean
ice
changes
climate change

Dynamic topic models (Blei and Lafferty, 2006)

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...

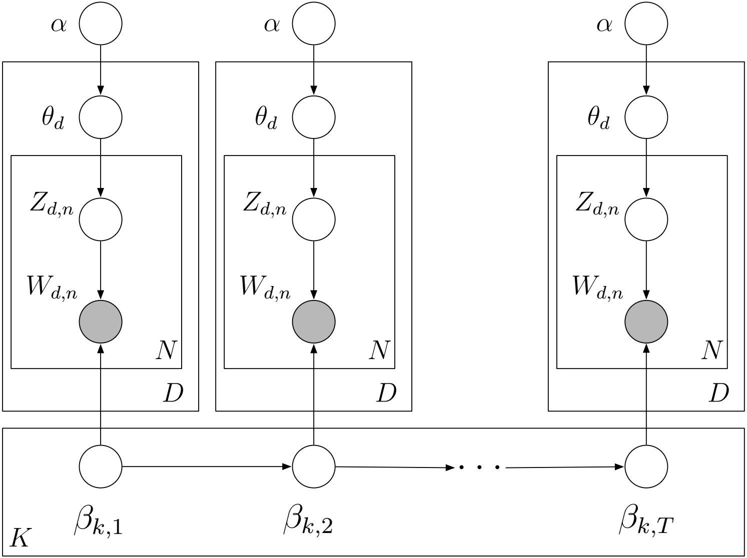
Inaugural addresses

2009

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...

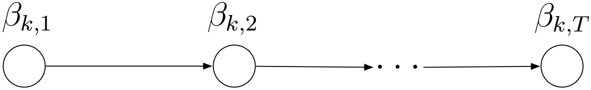
- LDA assumes that the order of documents does not matter.
- Not appropriate for corpora that span hundreds of years
- We may want to track how language changes over time.

Dynamic topic models



Topics drifting in time

Modeling evolving topics

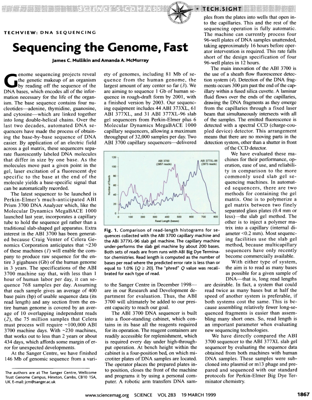


- Use a logistic normal distribution to model topics evolving over time (Aitchison, 1980)
- A state-space model on the natural parameter of the topic multinomial (West and Harrison, 1997)

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

Analyzing a document

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (1) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (2), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (3). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (4). Detection of the DNA fragments occurs 300 µm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

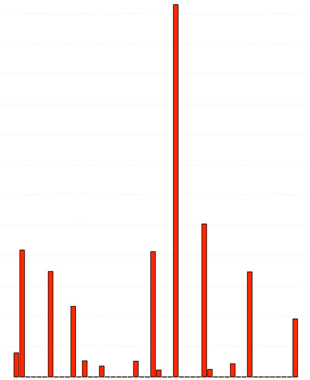
With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

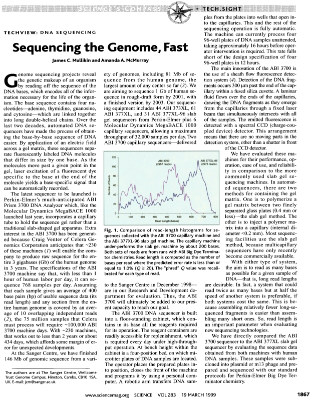
www.sciencemag.org SCIENCE VOL 283 19 MARCH 1999 1867

Topic proportions



Analyzing a document

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 µm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

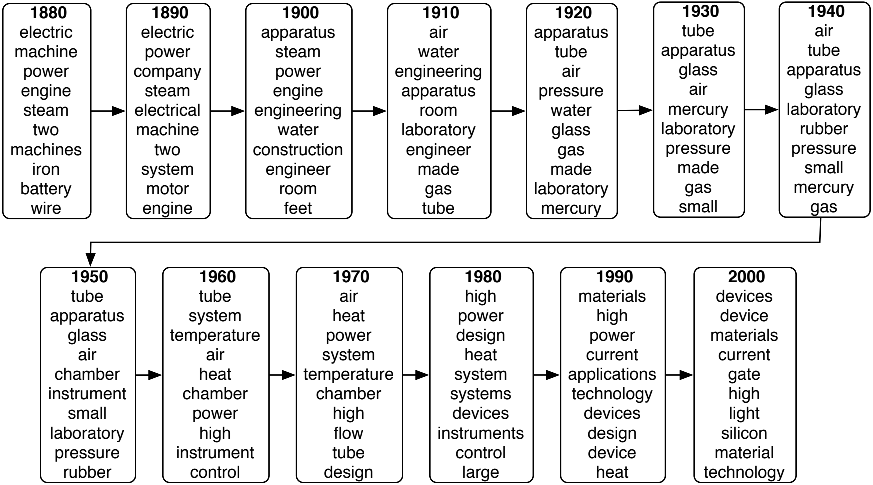
The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

www.sciencemag.org SCIENCE VOL 283 19 MARCH 1999 1867

Most likely words from top topics

sequence	devices	data
genome	device	information
genes	materials	network
sequences	current	web
human	high	computer
gene	gate	language
dna	light	networks
sequencing	silicon	time
chromosome	material	software
regions	technology	system
analysis	electrical	words
data	fiber	algorithm
genomic	power	number
number	based	internet

Analyzing a topic

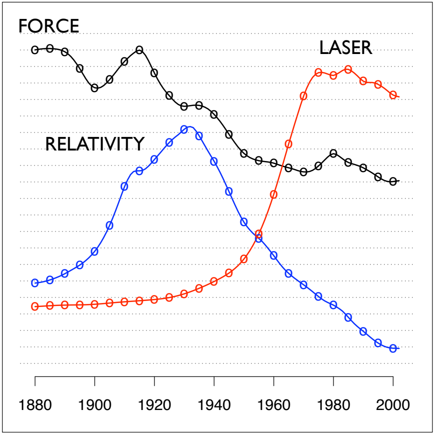


1880	1890	1900	1910	1920	1930	1940
electric	electric	apparatus	air	apparatus	tube	air
machine	power	steam	water	tube	apparatus	tube
power	company	power	engineering	air	glass	apparatus
engine	steam	engine	apparatus	pressure	air	glass
steam	electrical	engineering	room	water	mercury	laboratory
two	machine	water	laboratory	glass	laboratory	rubber
machines	two	construction	engineer	gas	pressure	pressure
iron	system	engineer	made	made	made	small
battery	motor	room	gas	laboratory	gas	mercury
wire	engine	feet	tube	mercury	small	gas

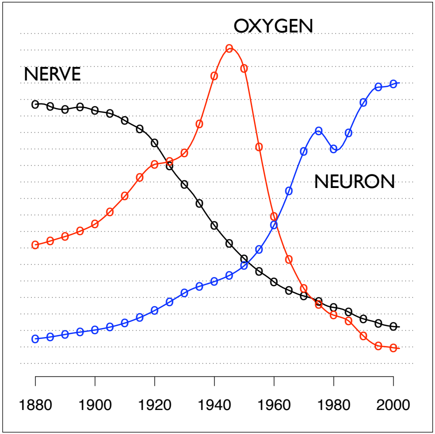
1950	1960	1970	1980	1990	2000
tube	tube	air	high	materials	devices
apparatus	system	heat	power	high	device
glass	temperature	power	design	power	materials
air	air	system	heat	current	current
chamber	heat	temperature	system	applications	gate
instrument	chamber	chamber	systems	technology	high
small	power	high	devices	devices	light
laboratory	high	flow	instruments	design	silicon
pressure	instrument	tube	control	device	material
rubber	control	design	large	heat	technology

Visualizing trends within a topic

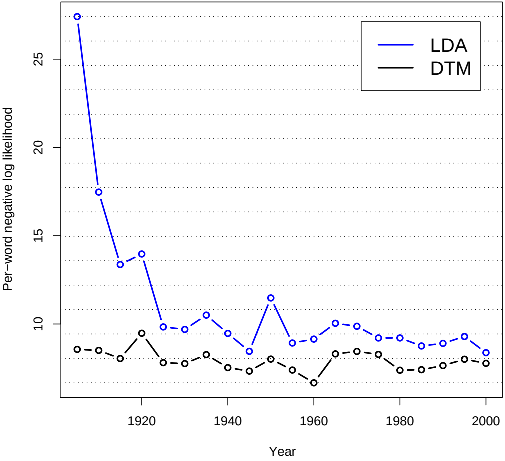
"Theoretical Physics"



"Neuroscience"



Evaluating the DTM on all of *Science*



See the browser at http://topics.cs.princeton.edu/Science/

Time-corrected document similarity

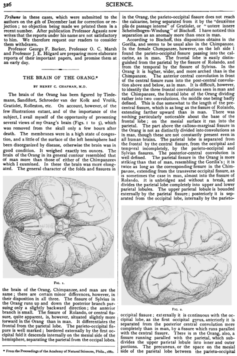
- Consider the expected Hellinger distance between the topic proportions of two documents,

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

- 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

Time-corrected document similarity

The Brain of the Orang (1880)



326 SCIENCE.

Tribune 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 recent number. After publication Professor Agassiz now writes that the reports under his name are not satisfactory to him. We therefore request our readers to consider them withdrawn.

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

THE BRAIN OF THE ORANG.*

BY HENRY C. CHAPMAN, M.D.

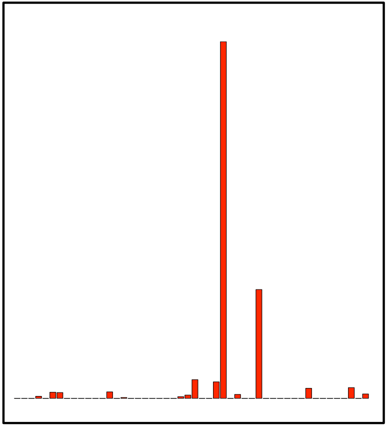
The brain of the Orang has been figured by Tiedemann, Sandifort, Schroeder van der Kolk and Vrolik, Gratiolet, Rolleston, etc. On account, however, of the few illustrations extant, and of the importance of the subject, I avail myself of the opportunity of presenting several views of my Orang's brain (Figs. 1 to 5), which was removed from the skull only a few hours after 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 general contour 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 Orang, Chimpanzee, and man are the same; there are certain minor differences, however, in their disposition in all three. The fissure of Sylvius in the Orang runs up and down the posterior branch pursuing only a slightly backward direction; the anterior branch is small. The fissure of Rolando, or central fissure, quite apparent, is, however, situated slightly more forward in the Orang than in man. It differentiates the frontal from the parietal lobe. The parieto-occipital fispure is well marked; bordered externally by the first occipital fold it descends internally on the mesial side of the hemisphere, separating the parietal from the occipal lobes. in the Orang, the parieto-occipital fissure does not reach the calcarine, being separated from it by the "deuxième plis de passage interne" of Gratiolet, or "untere innere Scheitelbogen-Windung" of Bischoff. I have noticed this separation as an anomaly more than once in man.

Fig. 1.

According to Bischoff, this disposition obtains in the Gorilla, and seems to be usual also in the Chimpanzee. In the female Chimpanzee, however, on the left side I found the parieto-occipital fissure passing into the calcarine, as in man. The frontal lobe is easily distinguished from the parietal by the fissure of Rolando, and from the temporal by the fissure of Sylvius. In the Orang it is higher, wider, and more arched than in the Chimpanzee. The anterior central convolution in front of the central fissure runs into the post-central convolution above and below, as in man. It is difficult, however, to identify the three frontal convolutions seen in man and the Chimpanzee, the frontal lobe of the Orang dividing rather into two convolutions, the middle one being badly defined. This is due somewhat to the length of the precentral fissure, which is as long as the fissure of Rolando, extending farther upward than in man. There was nothing particularly noticeable about the base of the frontal lobe; on the mesial surface it ran into the parietal. The part above the calloso-marginal fissure in the Orang is not as distinctly divided into convolutions as in man, though these are not constantly present even in all human brains. The parietal lobe is separated from the frontal by the central fissure, from the occipital and temporal incompletely, by the parieto-occipital and Sylvian fissures. The posterior-central convolution is well defined. The parietal fissure in the Orang is more striking than that of man, resembling the Gorilla's; it is twice as long as the corresponding fissure in the Chimpanzee, extending from the transverse occipital fissure, as is sometimes the case in man, almost into the fissure of Rolando. It is unbridged and without a break, and divides the parietal lobe completely into upper and lower parietal lobules. The upper parietal lobule is bounded externally by the parietal fissure; posteriorly it is separated from the occipital lobe, internally by the parieto-occipital fissure; externally it is continuous with the occipital lobe, as the first occipital gyrus, anteriorly it is separated from the posterior central convolution more completely than in man, by a fissure which runs paralled with the central fissure. There is in the Orang, also, a fissure running paralled with the parietal, which subdivides the upper parietal lobule into inner and outer portions. The precuneus, or the space on the mesial side of the parietal lobe between the parieto-occipital

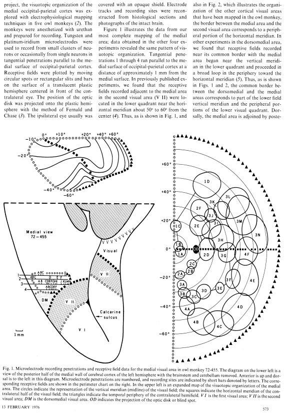
Fig. 2.

* From the Proceedings of the Academy of Natural Sciences, Phila., 1880.



Time-corrected document similarity

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



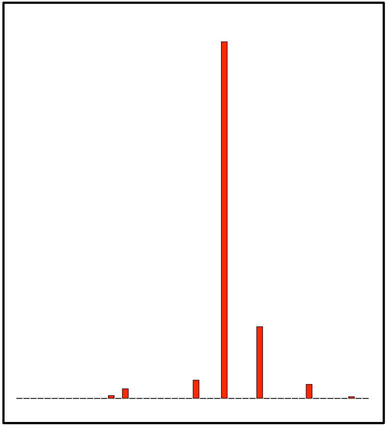
project, the visuotopic organization of the medial occipital-parietal cortex was explored with electrophysiological mapping techniques in five owl monkeys (*2*). The monkeys were anesthetized with urethan and prepared for recording. Tungsten and platinum-iridium microelectrodes were used to record from small clusters of neurons or occasionally from single neurons in tangential penetrations parallel to the medial surface of occipital-parietal cortex. Receptive fields were plotted by moving circular spots or rectangular slits and bars on the surface of a translucent plastic hemisphere centered in front of the contralateral eye. The position of the optic disk was projected onto the plastic hemisphere with the method of Fernald and Chase (*3*). The ipsilateral eye usually was covered with an opaque shield. Electrode tracks and recording sites were reconstructed from histological sections and photographs of the intact brain.

Figure 1 illustrates the data from our most complete mapping of the medial area; data obtained in the other four experiments revealed the same pattern of visuotopic organization. Tangential penetrations 1 through 4 ran parallel to the medial surface of occipital-parietal cortex at a distance of approximately 1 mm from the medial surface. In previously published experiments, we found that the receptive fields recorded adjacent to the medial area in the second visual area (V II) were located in the lower quadrant 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 that have been mapped in the owl monkey, the border between the medial area and the second visual area corresponds to a peripheral portion of the horizontal meridian. In other experiments in the dorsomedial area, we found that receptive fields recorded near its common border with the medial area began near the vertical meridian in the lower quadrant and proceeded in a broad loop in the periphery toward the horizontal meridian (*5*). Thus, as is shown in Figs. 1 and 2, the common border between the dorsomedial and the medial areas corresponds to part of the lower field vertical meridian and the peripheral portions of the lower visual quadrant. Dorsally, the medial area is adjoined by poste-

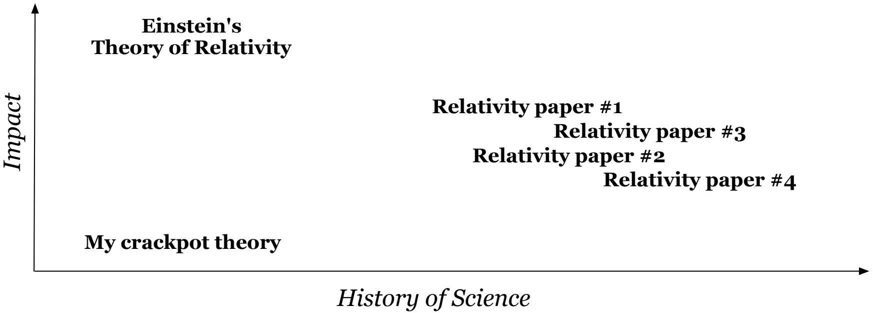
Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-455. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up and dorsal is to the left in this diagram. Microelectrode penetrations are numbered, and recording sites are indicated by short bars denoted by letters. The corresponding receptive fields are shown in the perimeter chart on the right. In the upper left is an expanded map of the visuotopic organization of the medial area. The circles indicate the representation of the vertical meridian (midline) of the visual field; the squares indicate the horizontal meridian of the contralateral half of the visual field; the triangles indicate the temporal periphery of the contralateral hemifield. *V I* is the first visual area; *V II* is the second visual area; *DM* is the dorsomedial visual area. *OD* indicates the projection of the optic disk or blind spot.

13 FEBRUARY 1976

573

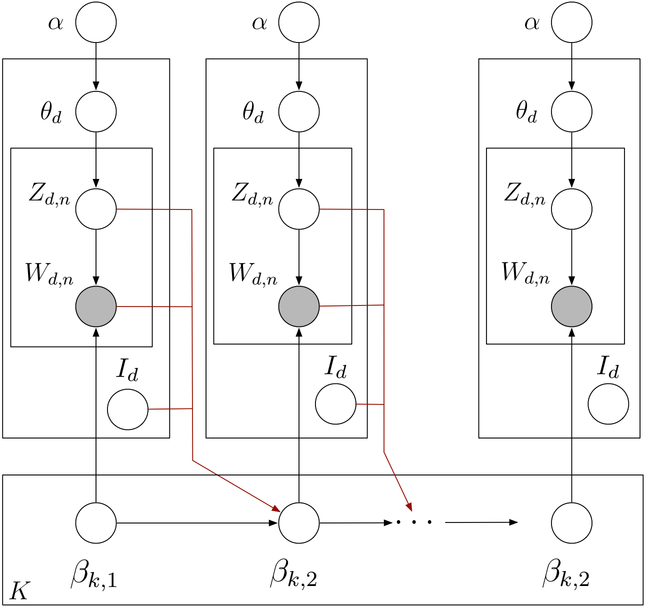


Measuring scholarly impact (Gerrish and Blei, 2009)

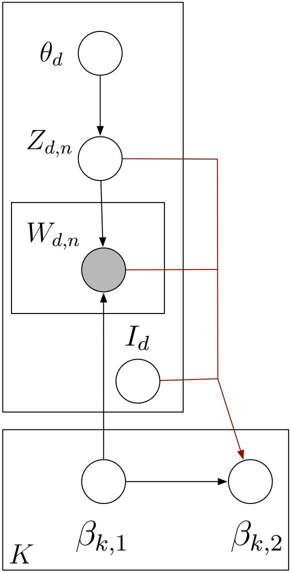


- Influential articles reflect future changes in language use.
- The “influence” of an article is a latent variable.
- Influential articles affect the drift of the topics that they discuss.
- The posterior gives a retrospective estimate of influential articles.

Measuring scholarly impact

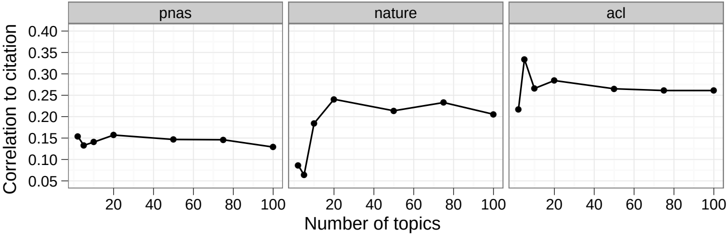


Measuring scholarly impact



- Each document has an influence score I_d.
- Each topic drifts in a way that is biased towards the documents with high influence.
- The posterior of $I_{1:D}$ can be examined to retrospectively find articles that best explain future changes in language.

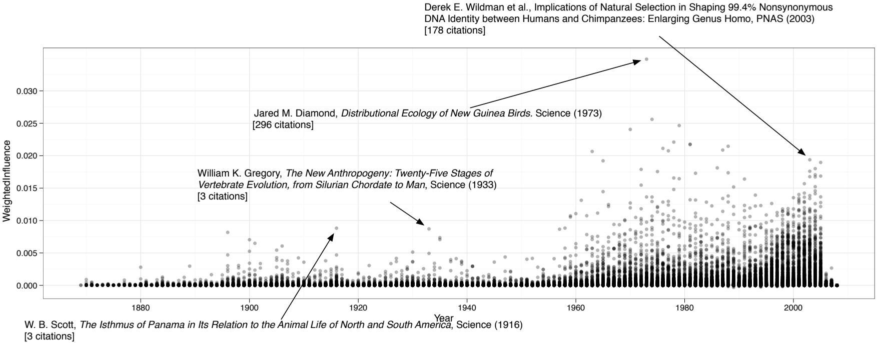
Measuring scholarly impact



- 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 at large scale

(with S. Gerrish, A. Chaney and D. Mimno)



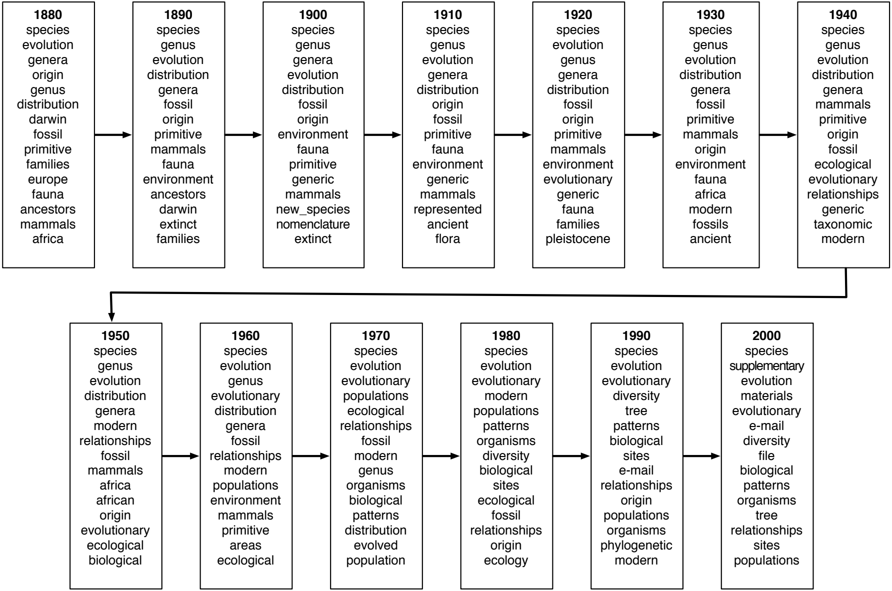
Derek E. Wildman et al., Implications of Natural Selection in Shaping 99.4% Nonsynonymous DNA Identity between Humans and Chimpanzees: Enlarging Genus Homo, PNAS (2003) [178 citations]

Jared M. Diamond, *Distributional Ecology of New Guinea Birds.* Science (1973) [296 citations]

William K. Gregory, *The New Anthropogeny: Twenty-Five Stages of Vertebrate Evolution, from Silurian Chordate to Man*, Science (1933) [3 citations]

W. B. Scott, *The Isthmus of Panama in Its Relation to the Animal Life of North and South America*, Science (1916) [3 citations]

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



1880
species
evolution
genera
origin
genus
distribution
darwin
fossil
primitive
families
europe
fauna
ancestors
mammals
africa

1890
species
genus
evolution
distribution
genera
fossil
origin
primitive
mammals
fauna
environment
ancestors
darwin
extinct
families

1900
species
genus
genera
evolution
distribution
fossil
origin
environment
fauna
primitive
generic
mammals
new_species
nomenclature
extinct

1910
species
genus
evolution
genera
distribution
origin
fossil
primitive
fauna
environment
generic
mammals
represented
ancient
flora

1920
species
evolution
genus
genera
distribution
fossil
origin
primitive
mammals
environment
evolutionary
generic
fauna
families
pleistocene

1930
species
genus
evolution
distribution
genera
fossil
primitive
mammals
origin
environment
fauna
africa
modern
fossils
ancient

1940
species
genus
evolution
distribution
genera
mammals
primitive
origin
fossil
ecological
evolutionary
relationships
generic
taxonomic
modern

1950
species
genus
evolution
distribution
genera
modern
relationships
fossil
mammals
africa
african
origin
evolutionary
ecological
biological

1960
species
evolution
genus
evolutionary
distribution
genera
fossil
relationships
modern
populations
environment
mammals
primitive
areas
ecological

1970
species
evolution
evolutionary
populations
ecological
relationships
fossil
modern
genus
organisms
biological
patterns
distribution
evolved
population

1980
species
evolution
evolutionary
modern
populations
patterns
organisms
diversity
biological
sites
ecological
fossil
relationships
origin
ecology

1990
species
evolution
evolutionary
diversity
tree
patterns
biological
sites
e-mail
relationships
origin
populations
organisms
phylogenetic
modern

2000
species
supplementary
evolution
materials
evolutionary
e-mail
diversity
file
biological
patterns
organisms
tree
relationships
sites
populations

Summary: Modeling richer assumptions

- 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.
  - See also Li and McCallum (2007) for another approach.
  - See http://www.cs.princeton.edu/$\sim$blei/ for code.

- The **dynamic topic model** uses a logistic normal in a linear dynamic model to capture how topics change over time.
  - Documents can exhibit sequential structure.
  - Opens the door to a citation-free model of scholarly impact.
  - See also Wang and Blei (2010) for a continuous time variant

- What's the catch? The Dirichlet is easier to compute with than the logistic normal. (Stay tuned.)

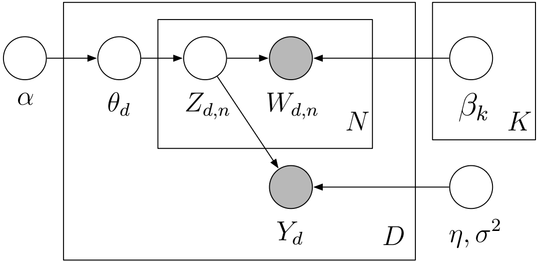
Supervised topic models

- Supervised LDA
- Relational topic models
- Ideal point topic models

Supervised LDA (Blei and McAuliffe, 2007)

- 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 topic models** are topic models of documents and responses, fit to find topics predictive of the response.

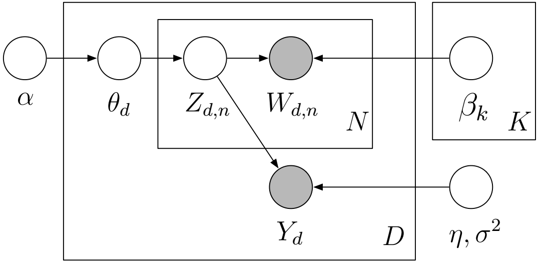
Supervised LDA



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

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

Supervised LDA



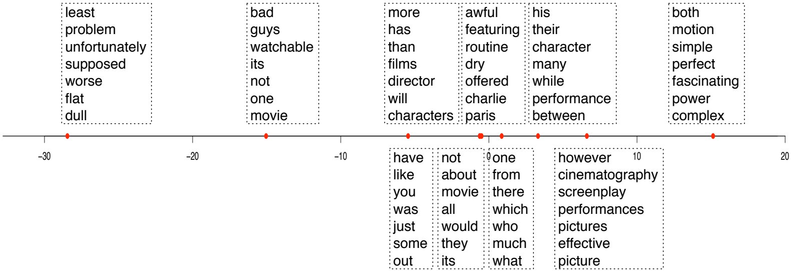
- The response variable *y* is drawn *after* the document because it depends on $z_{1:N}$, an assumption of **partial exchangeability.**
- Consequently, *y* is necessarily conditioned on the words.
- In a sense, this blends generative and discriminative modeling.

Prediction

- Fit sLDA parameters to documents and responses. This gives:
  - topics $\beta_{1:K}$
  - coefficients $\eta_{1:K}$
- We have a new document $w_{1:N}$ with unknown response value.
- We predict y using the SLDA expected value:

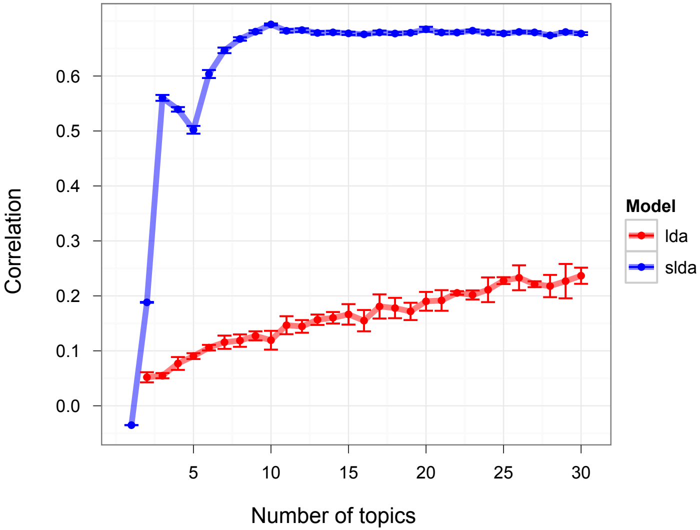
$$\mathrm{E}\left[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2\right] = \eta^\top \mathrm{E}\big[\bar{Z} \mid w_{1:N}\big]$$

Example: Movie reviews



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

Held out correlation



Diverse response types with GLMs

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

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

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

- Complicates inference, but allows for flexible modeling.

Image classification and annotation (Wang et al., 2009)

highway

car, sign, road

inside city

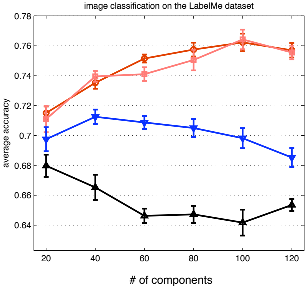
buildings, car, sidewalk

street

tree, car, sidewalk

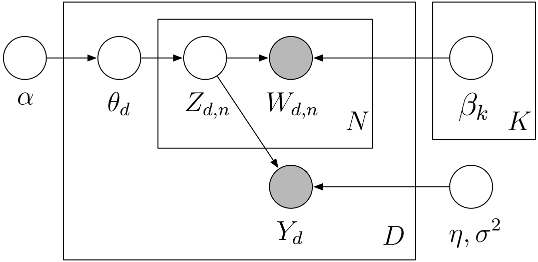
tall building

trees, buildings
occluded, window



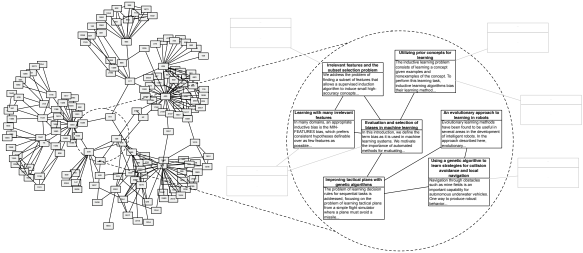
- Uses GLM sLDA for multiclass classification.
- Uses ideas from Blei and Jordan (2004) for annotation.

Supervised topic models



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

Relational topic models (Chang and Blei, 2010)



Irrelevant features and the subset selection problem
We address the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts...

Utilizing prior concepts for learning
The inductive learning problem consists of learning a concept given examples and nonexamples of the concept. To perform this learning task, inductive learning algorithms bias their learning method...

Learning with many irrelevant features
In many domains, an appropriate inductive bias is the MIN-FEATURES bias, which prefers consistent hypotheses definable over as few features as possible...

Evaluation and selection of biases in machine learning
In this introduction, we define the term bias as it is used in machine learning systems. We motivate the importance of automated methods for evaluating...

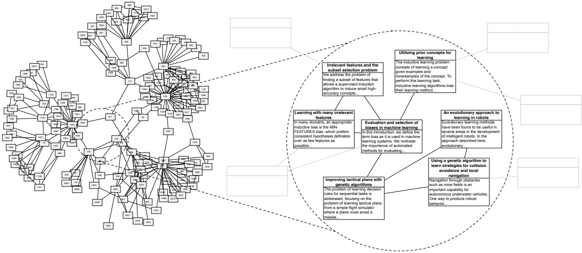
An evolutionary approach to learning in robots
Evolutionary learning methods have been found to be useful in several areas in the development of intelligent robots. In the approach described here, evolutionary...

Using a genetic algorithm to learn strategies for collision avoidance and local navigation
Navigation through obstacles such as mine fields is an important capability for autonomous underwater vehicles. One way to produce robust behavior...

Improving tactical plans with genetic algorithms
The problem of learning decision rules for sequential tasks is addressed, focusing on the problem of learning tactical plans from a simple flight simulator where a plane must avoid a missile...

- 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 (Chang and Blei, 2010)



Irrelevant features and the subset selection problem
We address the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts...

Utilizing prior concepts for learning
The inductive learning problem consists of learning a concept given examples and nonexamples of the concept. To perform this learning task, inductive learning algorithms bias their learning method...

Learning with many irrelevant features
In many domains, an appropriate inductive bias is the MIN-FEATURES bias, which prefers consistent hypotheses definable over as few features as possible...

Evaluation and selection of biases in machine learning
In this introduction, we define the term bias as it is used in machine learning systems. We motivate the importance of automated methods for evaluating...

An evolutionary approach to learning in robots
Evolutionary learning methods have been found to be useful in several areas in the development of intelligent robots. In the approach described here, evolutionary...

Using a genetic algorithm to learn strategies for collision avoidance and local navigation
Navigation through obstacles such as mine fields is an important capability for autonomous underwater vehicles. One way to produce robust behavior...

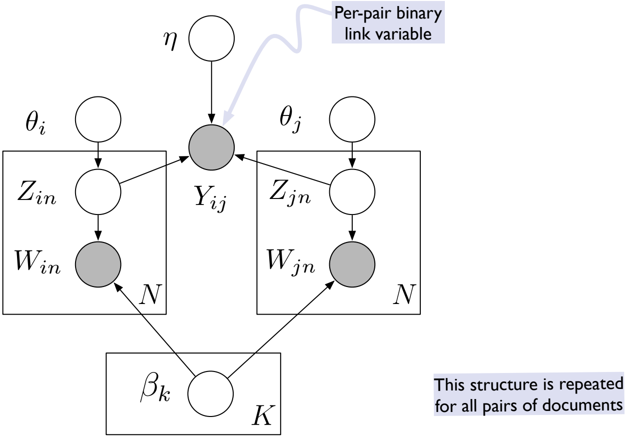
Improving tactical plans with genetic algorithms
The problem of learning decision rules for sequential tasks is addressed, focusing on the problem of learning tactical plans from a simple flight simulator where a plane must avoid a missile...

- Research has focused on finding communities and patterns in the link-structure of these networks.

- We adapt sLDA to pairwise response variables.
  This adaptation leads to a model of **content and connection**.

- RTMs find related hidden structure in both types of data.

Relational topic models



Per-pair binary link variable

This structure is repeated for all pairs of documents

- 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.

Predicting links from documents

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

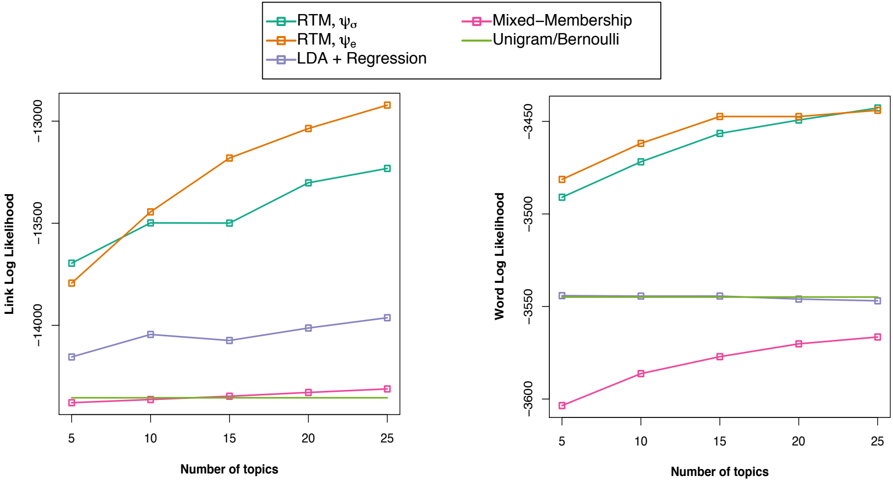
Given a new document, which documents is it likely to link to?

Predicting links from documents

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

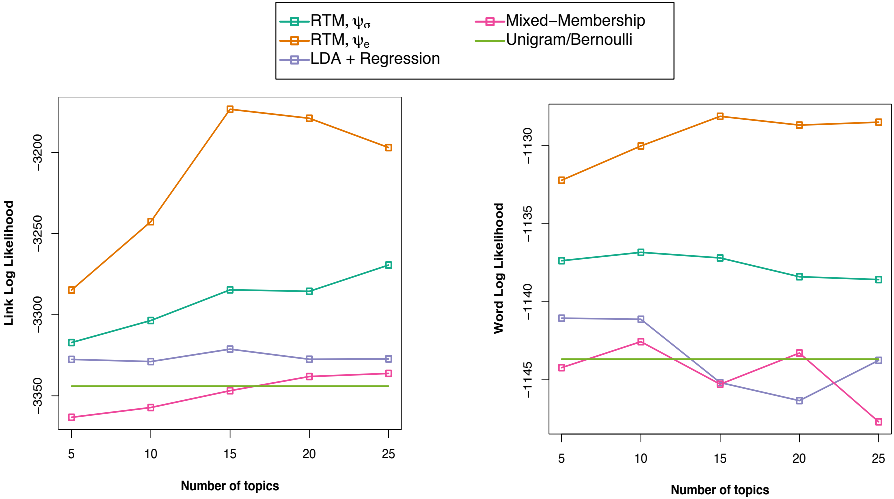
Given a new document, which documents is it likely to link to?

Predictive performance of each type



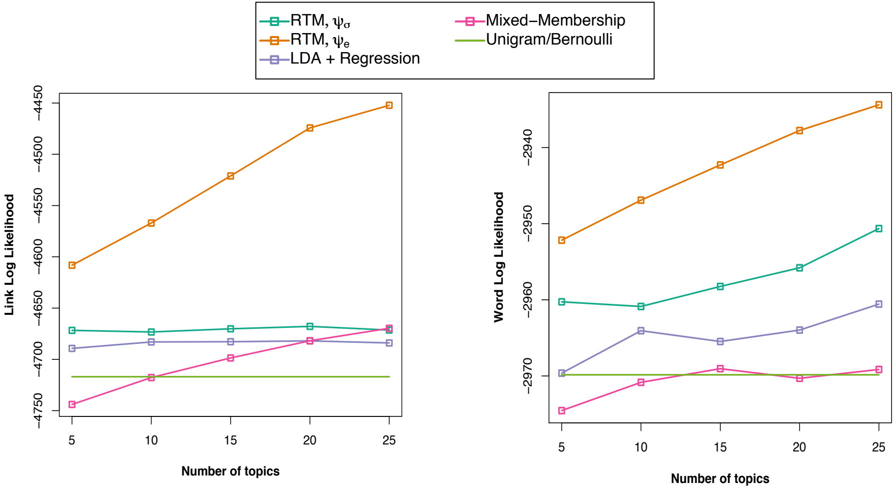
Cora corpus (McCallum et al., 2000)

Predictive performance of each type



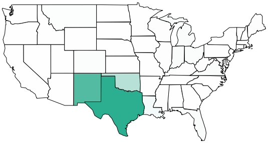
WebKB corpus (Craven et al., 1998)

Predictive performance of each type

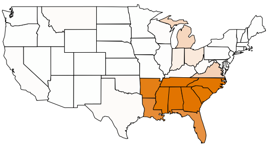


PNAS corpus (courtesy of JSTOR)

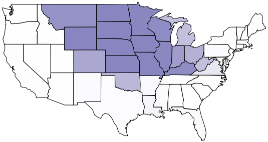
Spatially consistent topics



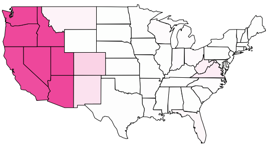
Topic 1



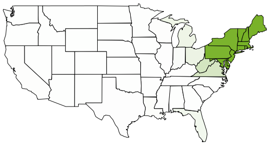
Topic 2



Topic 3



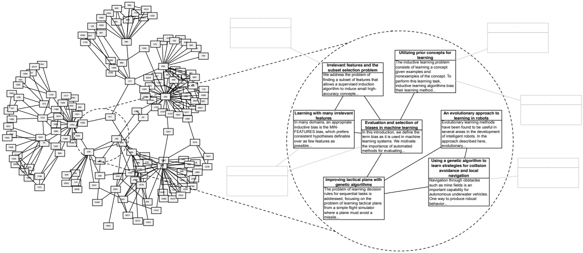
Topic 4



Topic 5

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

Relational topic models

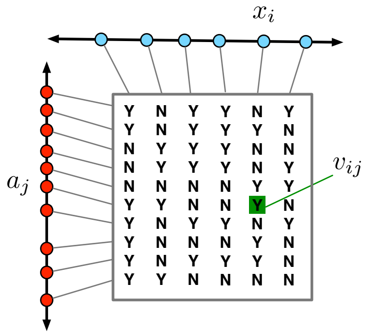

Irrelevant features and the subset selection problem
We address the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts...
Utilizing prior concepts for learning
The inductive learning problem consists of learning a concept given examples and nonexamples of the concept. To perform this learning task, inductive learning algorithms bias their learning method...
Learning with many irrelevant features
In many domains, an appropriate inductive bias is the MIN-FEATURES bias, which prefers consistent hypotheses definable over as few features as possible...
Evaluation and selection of biases in machine learning
In this introduction, we define the term bias as it is used in machine learning systems. We motivate the importance of automated methods for evaluating...
An evolutionary approach to learning in robots
Evolutionary learning methods have been found to be useful in several areas in the development of intelligent robots. In the approach described here, evolutionary...
Using a genetic algorithm to learn strategies for collision avoidance and local navigation
Navigation through obstacles such as mine fields is an important capability for autonomous underwater vehicles. One way to produce robust behavior...
Improving tactical plans with genetic algorithms
The problem of learning decision rules for sequential tasks is addressed, focusing on the problem of learning tactical plans from a simple flight simulator where a plane must avoid a missile...

- RTMs let us analyze connected documents, modeling both content and connections.

- Most network models cannot predict with new and unlinked data.

- RTMs allow for such predictions
  - links given the new words of a document
  - words given the links of a new document

The ideal point model

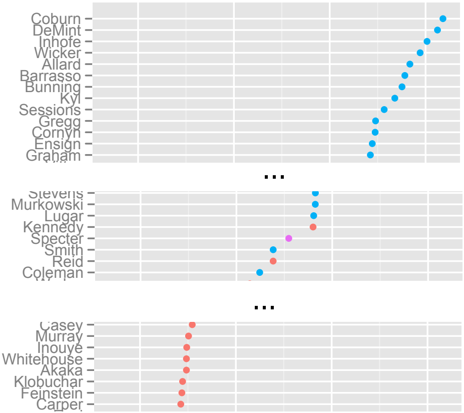


Y	N	Y	Y	N	Y
Y	N	Y	Y	Y	N
N	Y	Y	Y	N	N
N	Y	Y	Y	N	Y
N	N	N	N	Y	Y
Y	Y	N	N	Y	N
Y	N	Y	Y	N	Y
Y	N	N	N	Y	N
Y	N	Y	Y	Y	N
Y	Y	N	N	N	N

$$p(v_{ij}) = f(d(x_i, a_j))$$

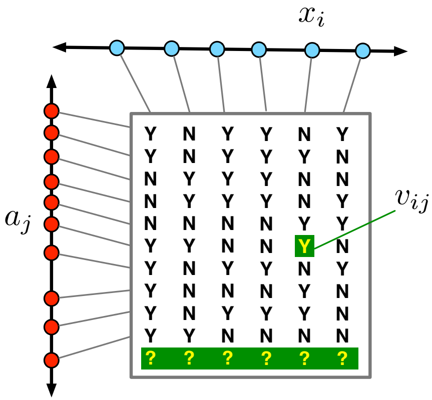
- A model devised to uncover voting patterns (Clinton et al., 2004).
- We observe roll call data v_{ij}.
- Bills attached to discrimination parameters a_j.
  Senators attached to ideal points x_i.

The ideal point model



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

The ideal point model is limited for prediction

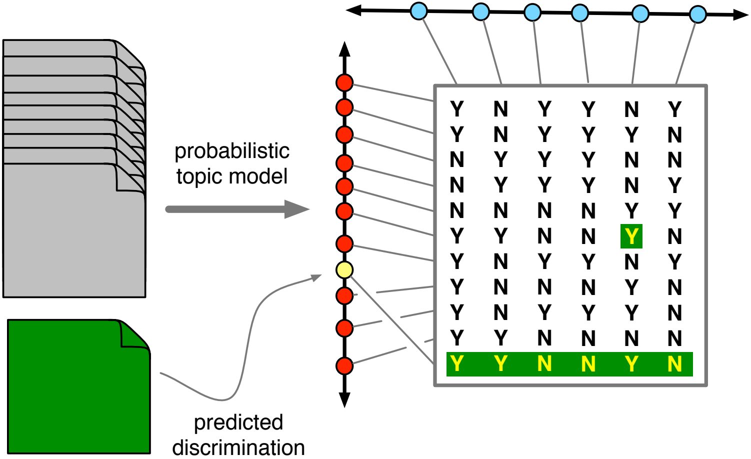


Y	N	Y	Y	N	Y
Y	N	Y	Y	Y	N
N	Y	Y	Y	N	N
N	Y	Y	Y	N	Y
N	N	N	N	Y	Y
Y	Y	N	N	Y	N
Y	N	Y	Y	N	Y
Y	N	N	N	Y	N
Y	N	Y	Y	Y	N
Y	Y	N	N	N	N
?	?	?	?	?	?

$$p(v_{ij}) = f(d(x_i, a_j))$$

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

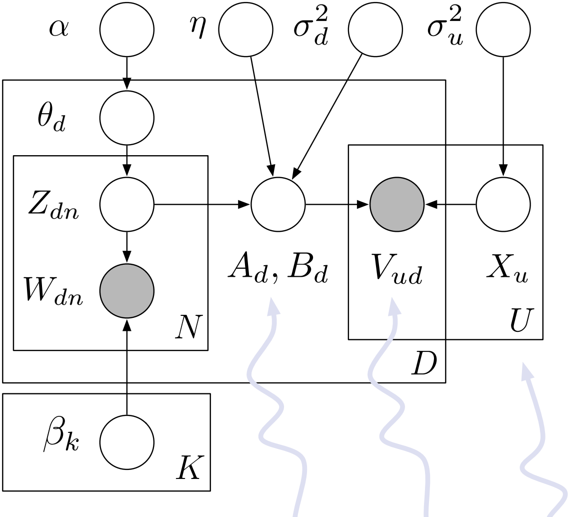
Ideal point topic models (Gerrish and Blei, 2010)



Y	N	Y	Y	N	Y
Y	N	Y	Y	Y	N
N	Y	Y	Y	N	N
N	Y	Y	Y	N	Y
N	N	N	N	Y	Y
Y	Y	N	N	Y	N
Y	N	Y	Y	N	Y
Y	N	N	N	Y	N
Y	N	Y	Y	Y	N
Y	Y	N	N	N	N
Y	Y	N	N	Y	N

Use supervised topic modeling assumptions as a predictive mechanism from bill texts to bill discrimination.

Ideal point topic models



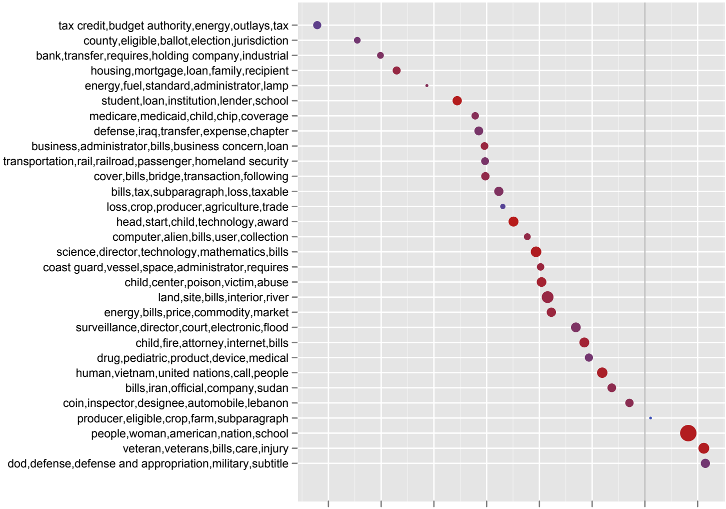
Bill content (topic model)

Bill sentiment variables

Observed votes

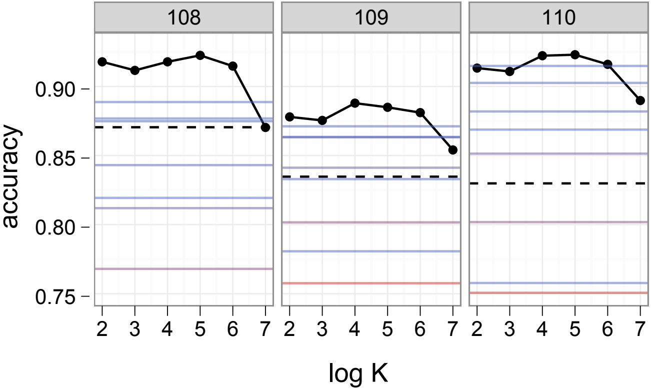
Legislator ideal points

Ideal point topics



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

Prediction on completely held-out votes



Versus the LASSO, the IPTM correctly predicted 126,000 more votes.

Summary: Supervised topic models

- Many of 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.
- Note that sLDA and the RTM (and others) are implemented in Jonathan Chang's excellent R package "lda."

Still other ways to build on LDA

New applications—

- Syntactic topic models (Boyd-Graber and Blei 2009)
- Topic models on images (Fei-fei and Perona 2005 and others)
- Topic models on social network data (Airoldi et al. 2008)
- Topic models on music data (Hoffman et al. 2008)
- **Topic models for user recommendation (Wang and Blei, 2011)**

Testing and relaxing assumptions—

- Spike and slab priors (Wang and Blei 2009 and Williamson et al. 2010)
- Models of word contagion (Elkan 2006)
- N-gram topic models (Wallach 2006)

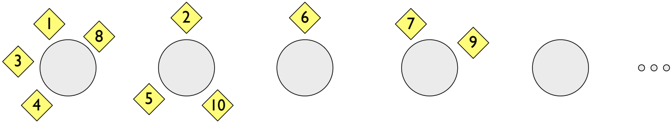
Bayesian nonparametric topic models

- Why Bayesian nonparametric models?
- The Chinese restaurant process
- Chinese restaurant process mixture models
- The Chinese restaurant franchise
- Bayesian nonparametric topic models

Why Bayesian nonparametric models?

- Topic models assume that the number of topics is fixed.
- It can be determined by cross validation and other model selection techniques.
- Bayesian nonparametric methods skirt model selection—
  - The data determine the number of topics during inference
  - Future data can exhibit new topics
- This is really a field unto itself, but it has found wide application in topic modeling.

The Chinese restaurant process (CRP)

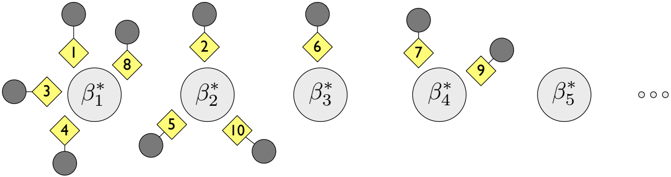


- N customers arrive to an infinite-table restaurant. Each sits down according to how many people are sitting at each table,

$$p(z_i = k \mid z_{1:(i-1)}, \alpha) \propto \begin{cases} n_k & \text{for} \quad k \leq K \\ \alpha & \text{for} \quad k = K + 1. \end{cases}$$

- The resulting seating plan provides a partition
- This distribution is **exchangeable**: Seating plan probabilities are the same regardless of the order of customers (Pitman, 2002).

CRP mixture models



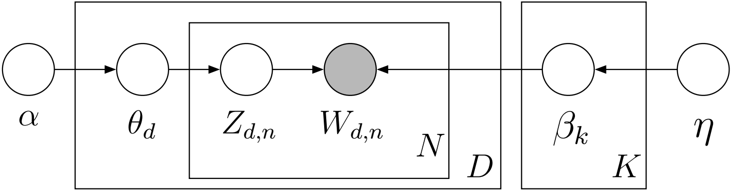
- Associate each table with a topic (β^*).
  Associate each customer with a data point (grey node).

- The number of clusters is infinite a priori; the data determines the number of clusters in the posterior.

- Further: the next data point might sit at new table.

- Exchangeability makes inference easy (see Neal, 2000).

The CRP is not a mixed-membership model

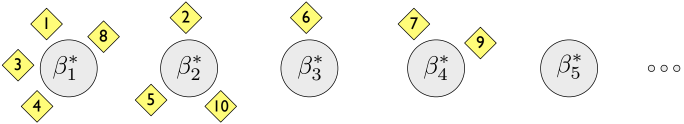


- Mixture models draw each data point from one component.
- The advantage of LDA is that it’s a **mixed membership model**.
- This is addressed by the **Chinese restaurant franchise**.

The Chinese restaurant franchise (Teh et al., 2006)

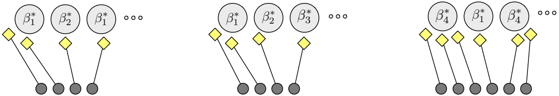
Corpus level restaurant

At the corpus level, topics are drawn from a prior.



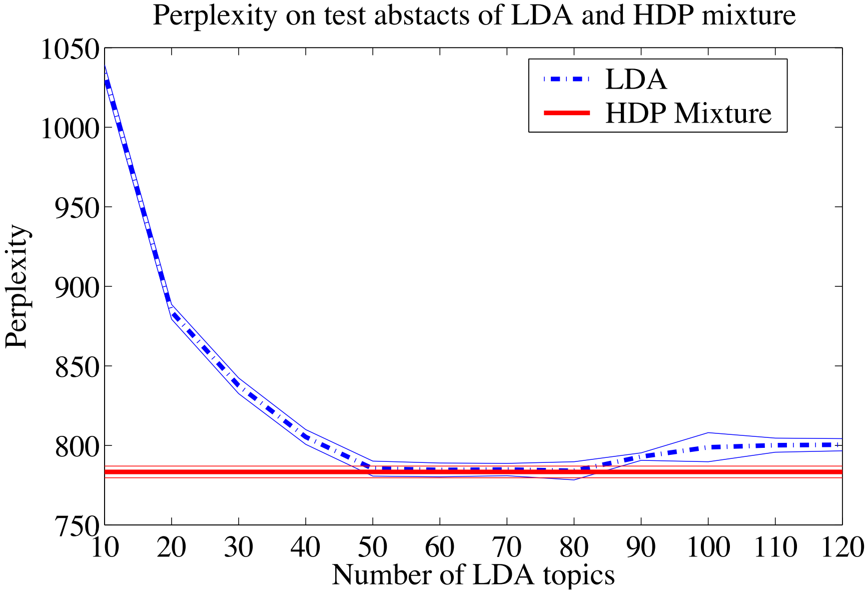
Document level restaurants

Each document-level table is associated with a customer at the corpus level restaurant.

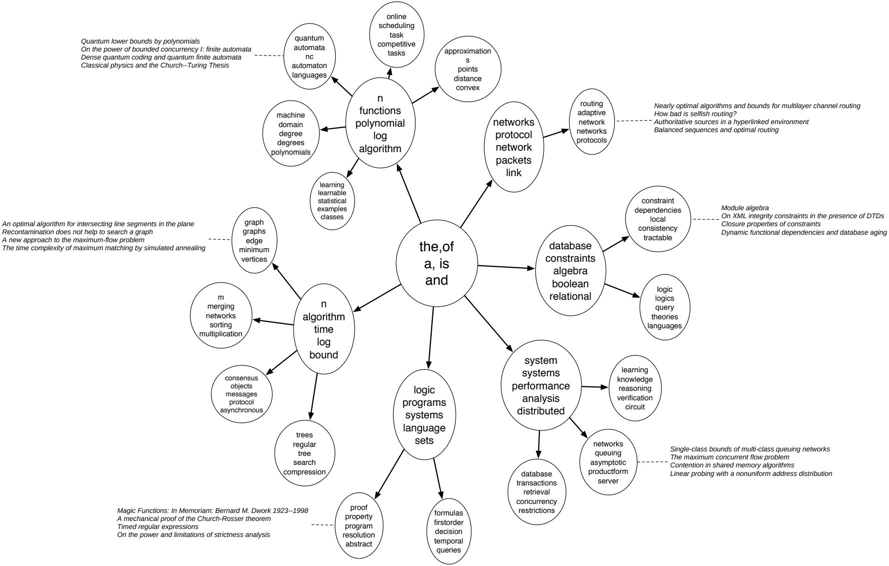


Each word is associated with a customer at the document's restuarant. It is drawn from the topic that it's table is associated with

The CRF selects the “right” number of topics



Extended to find hierarchies (Blei et al., 2010)



Quantum lower bounds by polynomials
On the power of bounded concurrency I: finite automata
Dense quantum coding and quantum finite automata
Classical physics and the Church--Turing Thesis

Nearly optimal algorithms and bounds for multilayer channel routing
How bad is selfish routing?
Authoritative sources in a hyperlinked environment
Balanced sequences and optimal routing

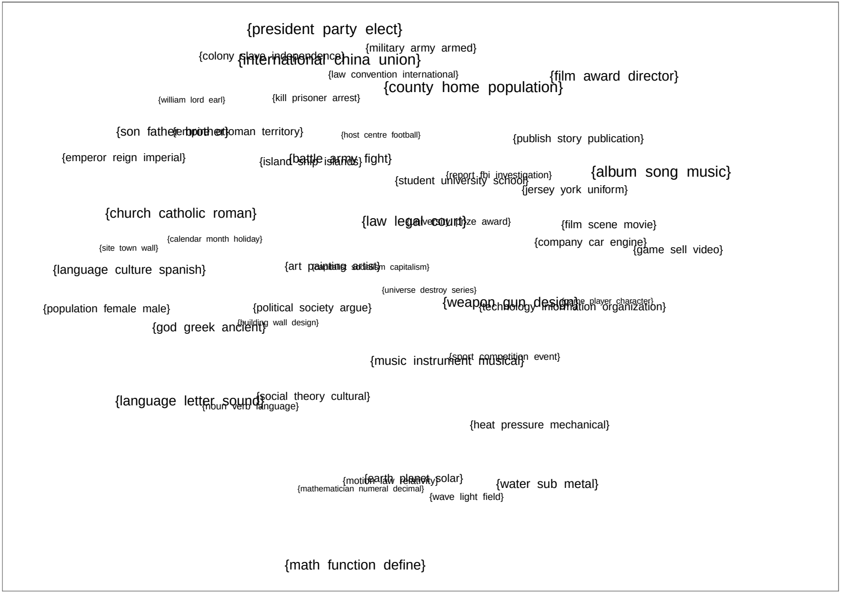
Module algebra
On XML integrity constraints in the presence of DTDs
Closure properties of constraints
Dynamic functional dependencies and database aging

An optimal algorithm for intersecting line segments in the plane
Recontamination does not help to search a graph
A new approach to the maximum-flow problem
The time complexity of maximum matching by simulated annealing

Single-class bounds of multi-class queuing networks
The maximum concurrent flow problem
Contention in shared memory algorithms
Linear probing with a nonuniform address distribution

Magic Functions: In Memoriam: Bernard M. Dwork 1923--1998
A mechanical proof of the Church-Rosser theorem
Timed regular expressions
On the power and limitations of strictness analysis

BNP correlated topic model (Paisley et al., 2011)



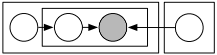
Summary: Bayesian nonparametrics

- **Bayesian nonparametrics** is a growing field (Hjort et al., 2011).
- BNP methods can define priors over combinatorial structures.
- In the posterior, the documents determine the particular form of the structure that is best for the corpus at hand.
- These models are also interpretable as **random distribution models**, such as the Dirichlet process (Fergusen 1973, Antoniak 1974).
- Recent innovations:
  - Improved inference methods (Blei and Jordan, 2005)
  - Dependent models, such as time series models (MacEachern 1999, Dunson 2010)
  - Models for predictions (Hannah et al. 2011)
  - Models for matrix factorization and other non-mixtures (Griffiths and Ghahramani, 2011)

Algorithms

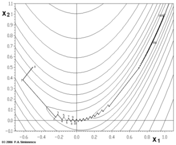
So far...

Assumptions

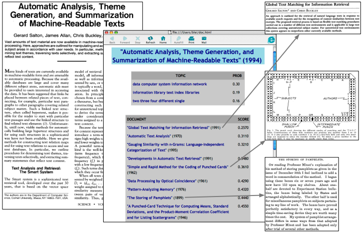


Data

Inference algorithm



Discovered structure


Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts
Gerard Salton, James Allan, Chris Buckley
Text Analysis and Retrieval: The Smart System
Global Text Matching for Information Retrieval
Gerard Salton and Chris Buckley
THE STORING OF PAMPHLETS.
file:///Users/blei/doc.html
Back
Forward
Stop
Refresh
Home
AutoFill
Print
Mail
SUBJECTS
AUTHORS
SHALLOW PUNCH
INTERMEDIATE PUNCH

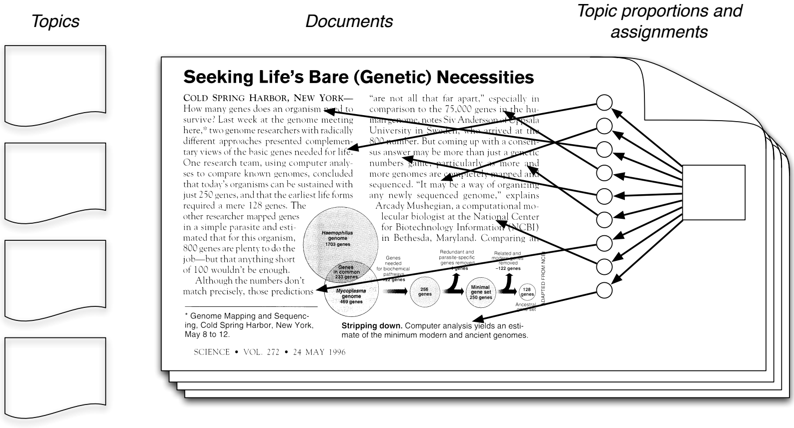
"Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts" (1994)

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

- We can express many kinds of assumptions about a corpus.
- Next: How can we analyze it under those assumptions?

Posterior 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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways −22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

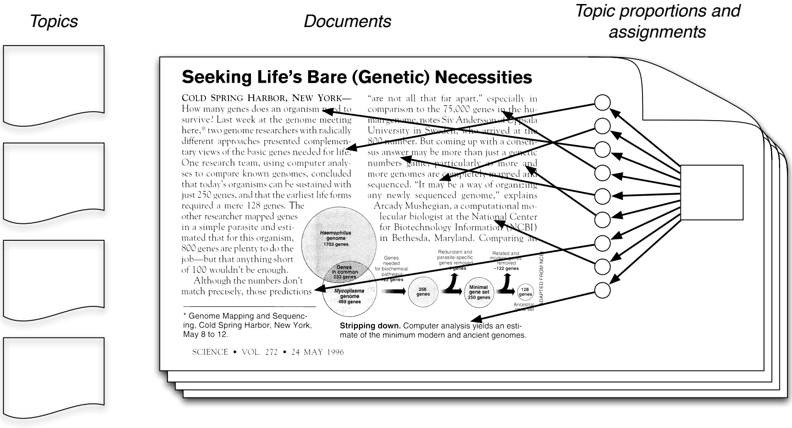
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- 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
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways –22 genes
256 genes
Redundant and parasite-specific genes removed –4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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.

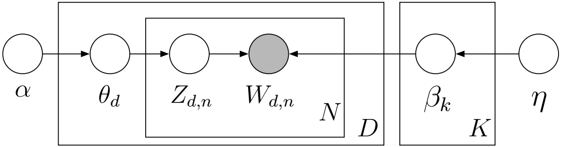
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- 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



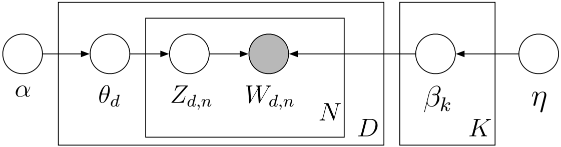
- The joint distribution of the latent variables and documents is

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

- The posterior of the latent variables given the documents is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N} \mid w_{1:D,1:N}).$$

Posterior inference for LDA

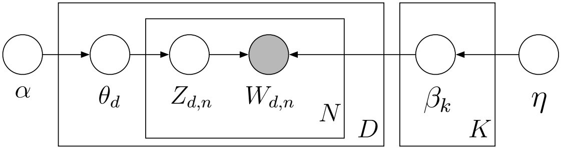


- This is equal to

$$\frac{p(\beta_{1:K}, \theta_{1:D}, \mathbf{z}_{1:D}, \mathbf{w}_{1:D})}{\int_{\beta_{1:K}} \int_{\theta_{1:D}} \sum_{\mathbf{z}_{1:D}} p(\beta_{1:K}, \theta_{1:D}, \mathbf{z}_{1:D}, \mathbf{w}_{1:D})}.$$

- We can't compute the denominator, the marginal $p(\mathbf{w}_{1:D})$.
- This is the crux of the inference problem.

Posterior inference for LDA

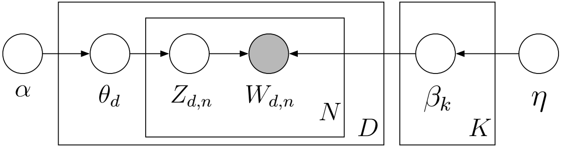


- There is a large literature on approximating the posterior.
- We will focus on
  - Gibbs sampling
  - Mean-field variational methods (batch and online)

Markov chain Monte Carlo

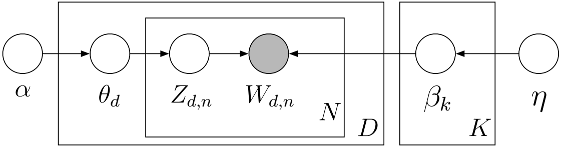
- Construct a **Markov chain** on the hidden variables, whose limiting distribution is the posterior.
- Collect **independent samples** from that distribution; approximate the posterior with them
- In **Gibbs sampling** the chain is defined by the conditional distribution of each hidden variable given observations and the current setting of the other hidden variables.

Local and global variables



- Local variables are local to each document
  - Topic proportions θ_d
  - Topic assignemnts $z_{d,n}$
- Global variables are shared by the corpus
  - Topics β_k

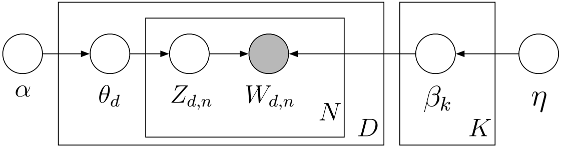
Local and global variables



- Assume the topics are fixed.
- Even "local inference" is intractable,

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

Local Gibbs sampling for LDA



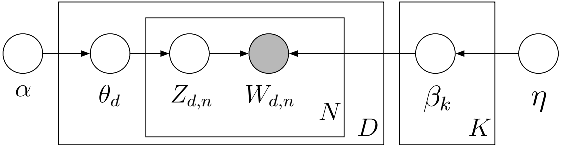
- We observe words $\mathbf{w} = w_{1:N}$. The Markov chain is defined on $\{\theta, z_{1:N}\}$, the topic proportions and topic assignments.

- Some notation—

$$\begin{aligned} n(z_{1:N}) &= \textstyle\sum_{n=1}^{N} z_n \\ m_k(\mathbf{z}_{1:D}, \mathbf{W}) &= \textstyle\sum_{d=1}^{D} \sum_{n=1}^{N} z_{d,n}^{k} w_{d,n}. \end{aligned}$$

- $n(z_{1:N})$ are topic counts;
  $m_k(z_{1:N}, \mathbf{W})$ are within-topic word counts.

Local Gibbs sampling for LDA



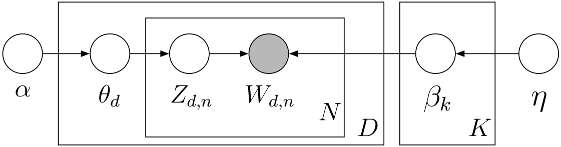
A simple Gibbs sampler is

$$\begin{aligned} \theta \,|\, \mathbf{w}, z_{1:N} &\sim \mathrm{Dir}(\gamma) \\ z_n \,|\, \theta, \mathbf{w} &\sim \mathrm{Mult}(\phi_n) \end{aligned}$$

where

$$\begin{aligned} \gamma &= \alpha + n(z_{1:N}) \\ \phi_n &\propto \theta \cdot p(w_n \,|\, \beta_{1:K}). \end{aligned}$$

Collapsed local Gibbs sampling



- The topic proportions θ can be integrated out,

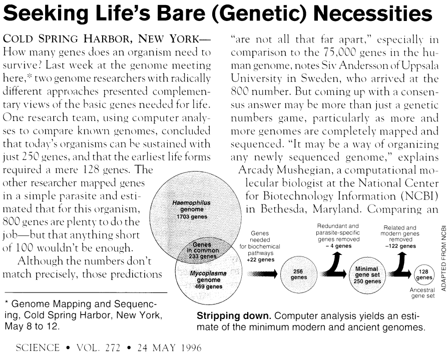
$$p(z_n \,|\, z_{-n}, \mathbf{w}) = p(w_n \,|\, \beta_{1:K}) \cdot \int_\theta p(z_n \,|\, \theta) p(\theta \,|\, z_{-n}) d\theta$$

- A collapsed Gibbs sampler constructs a chain on $z_{1:N}$,

$$z_n \,|\, z_{-n}, \mathbf{w} \sim \mathrm{Mult}(\phi_n),$$

where $\phi_n \propto p(w_n \,|\, \beta_{1:K})(n(z_{-n}) + \alpha)$.

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.

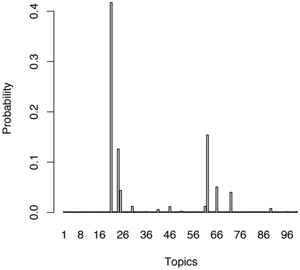
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

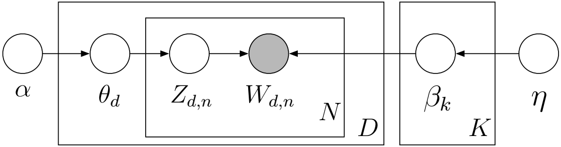
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996



Sampling the topics



- We observe the corpus $\mathbf{W} = \mathbf{w}_{1:D}$.
- We define the chain on $\{\mathbf{z}_{1:D}, \theta_{1:D}, \beta_{1:K}\}$.
- First, sample latent variables $(\mathbf{z}_d, \theta_d)$ for each document.
- Then, sample each topic from

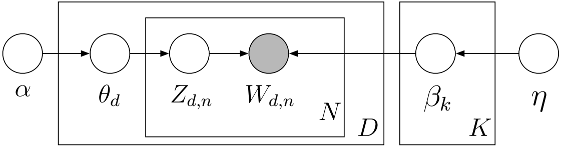
$$\beta_k \,|\, \mathbf{z}_{1:D}, \mathbf{W} \sim \mathrm{Dir}(\lambda_k),$$

where

$$\lambda_k := \eta + m_k(\mathbf{z}_{1:D}, \mathbf{W}).$$

Recall $m_k(\mathbf{z}_{1:D}, \mathbf{W})$ are words counts for topic k.

Collapsed Gibbs sampling with topics



- We can integrate out the topics $\beta_{1:K}$ too.
- The sampler is defined on the topic assigments $\mathbf{z}_{1:D}$,

$$p(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{W}) \propto \left(\frac{m_k(\mathbf{z}_{-(n,d)}, \mathbf{W}) + \eta}{\sum_v m_k^v(\mathbf{z}_{-(n,d)}) + V\eta} \right) (n_k(z_{-i}) + \alpha)$$

- This is an excellent Gibbs sampler for LDA. It was developed by Giffiths and Steyvers (2002) and is widely used.

Example topic inference

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

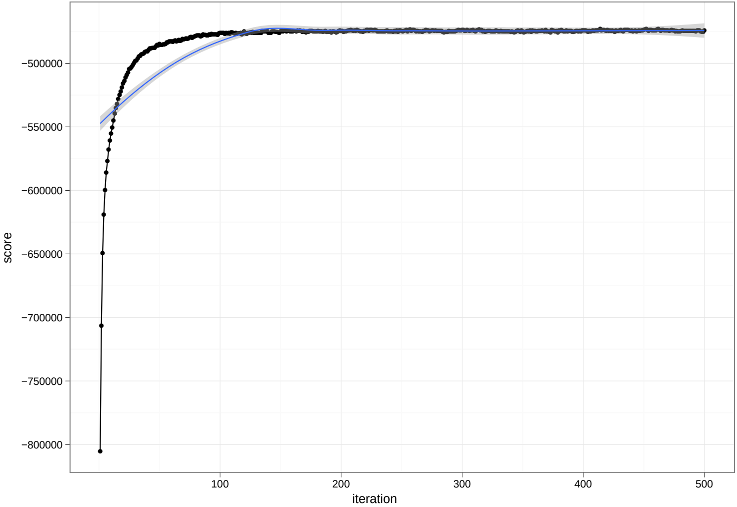
Gibbs sampling for LDA in practice

- In practice:
  1. Obtain a corpus of documents **W**
  2. Run the Gibbs sampler for some number of iterations.
  3. Store states at some lag, or store the MAP state.

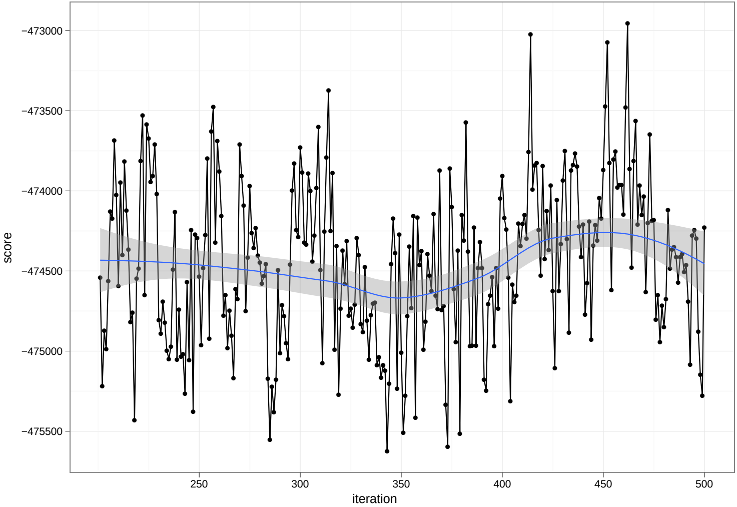
- Look at counts like $m_k(\mathbf{z}_{1:D}, \mathbf{W})$ to investigate the topics; look at $n(\mathbf{z}_d)$ to investigate how each document exhibits them.

- **A good habit: Assess the convergence of the chain.**
  - Monitor the log probability of the state & observations. (Its exponential is proportional to the posterior.)
  - Do something fancier, e.g., Raftery and Lewis (1992).

Assessing convergence example



Assessing convergence example



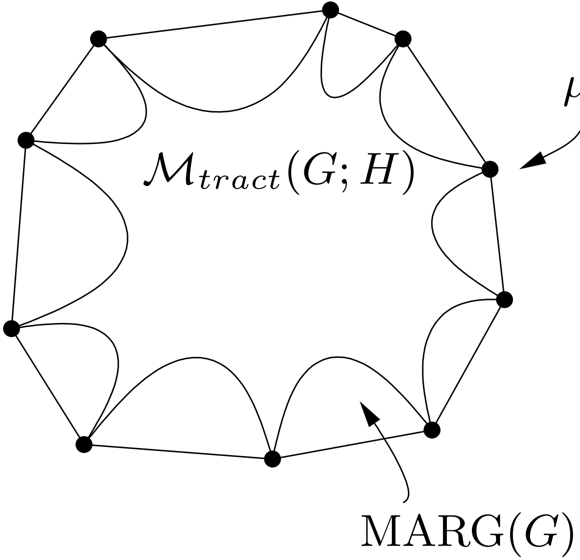
Gibbs sampling for LDA

- Simple algorithm for sampling from a complex distribution.
- Works well in practice. Is the best first algorithm to try.
- However
    - Can be slow for very large data sets
    - It is difficult to handle nonconjugacy; it is hard to generalize to the dynamic topic model and correlated topic model.

Variational inference

- Variational inference replaces sampling with **optimization.**
- The main idea—
  - Place a distribution over the hidden variables with free parameters, called **variational parameters.**
  - Optimize the variational parameters to make the distribution close (in KL divergence) to the true posterior
- In some settings, variational inference is faster than MCMC.
- It is easier to handle nonconjugate pairs of distributions with variational inference. (This is important in the CTM, DTM, etc.)

A useful picture (from Wainwright and Jordan, 2008)



Variational inference (in general)

- Let $x = x_{1:N}$ be observed variables;
  let $z = z_{1:M}$ be the latent variables.

- Our goal is to compute the posterior distribution

$$p(z \mid x) = \frac{p(z, x)}{\int p(z, x) dz}$$

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

Variational inference

- Introduce a distribution over the latent variables $q_\nu(z)$, parameterized by *variational parameters* ν.
- Use Jensen's inequality to bound the log probability of the observations, (Jordan et al., 1999)

$$\begin{aligned}\log p(x) &= \log \int p(z, x) dz \\ &= \log \int p(z, x) \frac{q_\nu(z)}{q_\nu(z)} dz \\ &\geq \mathrm{E}_{q_\nu}[\log p(Z, x)] - \mathrm{E}_{q_\nu}[\log q_\nu(Z)]\end{aligned}$$

(J. McAuliffe calls this the **evidence lower bound**, or ELBO.)

- Optimize the variational parameters to tighten this bound.
- This is the same as finding the member of the family q_ν that is closest in KL divergence to $p(z \mid x)$.

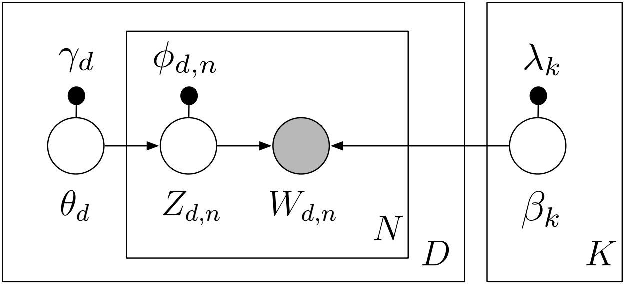
Mean-field variational inference

- Complexity is determined by the factorization of q_ν
- In *mean field variational inference* q_ν is fully factored

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

- Each latent variable is independently governed by its own variational parameter ν_m.
- In the true posterior they can exhibit dependence.
  (Often, this is what makes exact inference difficult.)

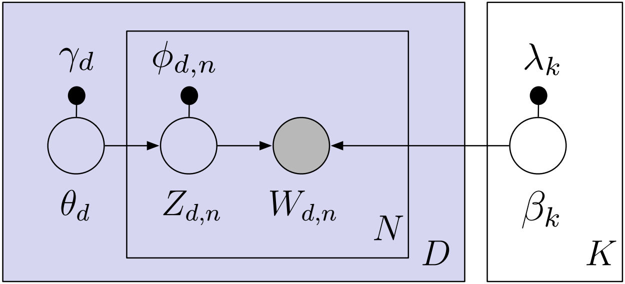
Variational inference for LDA



- 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.

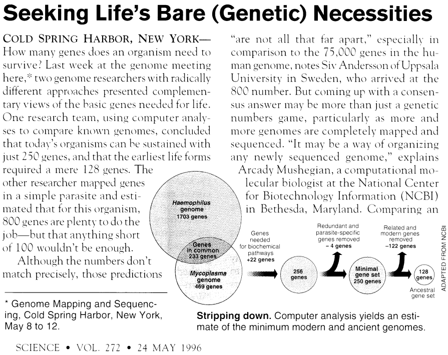
Variational inference for LDA



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

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

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.

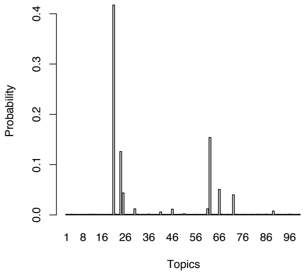
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

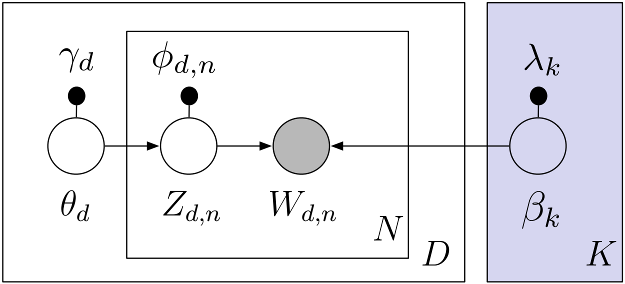
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996



Variational inference for LDA



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

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

Example topic inference

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

Variational inference for LDA (sketch)

```
Initialize topics randomly.
repeat
    for each document do
        repeat
            Update the topic assignment variational parameters.
            Update the topic proportions variational parameters.
        until document objective converges
    end for
    Update the topics from aggregated per-document parameters.
until corpus objective converges.
```

Variational inference for LDA

```
Initialize topics $\boldsymbol{\lambda}_{1:K}$ randomly.
while relative improvement in $\mathcal{L}(\mathbf{w}, \phi, \boldsymbol{\gamma}, \boldsymbol{\lambda}) > \epsilon$ do
    for $d = 1$ to $D$ do
        Initialize $\gamma_{d,k} = 1$.
        repeat
            Set $\phi_{d,n} \propto \exp\{\mathbb{E}_q[\log \theta_d] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
            Set $\gamma_d = \alpha + \sum_n \phi_{d,n}$
        until $\frac{1}{K} \sum_k |\text{change in } \gamma_{d,k}| < \epsilon$
    end for
    Set $\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}$
end while
```

“E step”

```
Initialize topics $\boldsymbol{\lambda}_{1:K}$ randomly.
while relative improvement in $\mathcal{L}(\mathbf{w}, \phi, \gamma, \boldsymbol{\lambda}) > \epsilon$ do
    for $d = 1$ to $D$ do
        Initialize $\gamma_{d,k} = 1$.
        repeat
            Set $\phi_{d,n} \propto \exp\{\mathbb{E}_q[\log \theta_d] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
            Set $\gamma_d = \alpha + \sum_n \phi_{d,n}$
        until $\frac{1}{K} \sum_k |\text{change in } \gamma_{d,k}| < \epsilon$
    end for
    Set $\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}$
end while
```

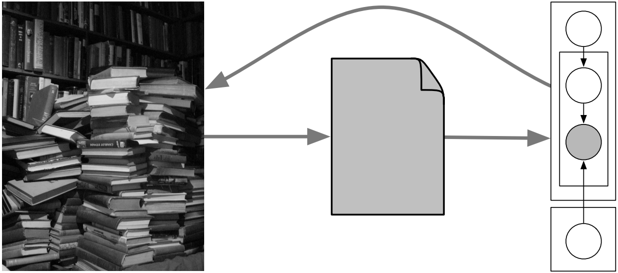
Do variational inference for each document.

“M step”

```
Initialize topics $\boldsymbol{\lambda}_{1:K}$ randomly.
while relative improvement in $\mathcal{L}(\mathbf{w}, \phi, \gamma, \boldsymbol{\lambda}) > \epsilon$ do
    for $d = 1$ to $D$ do
        Initialize $\gamma_{d,k} = 1$.
        repeat
            Set $\phi_{d,n} \propto \exp\{\mathbb{E}_q[\log \theta_d] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
            Set $\gamma_d = \alpha + \sum_n \phi_{d,n}$
        until $\frac{1}{K} \sum_k |\text{change in } \gamma_{d,k}| < \epsilon$
    end for
    Set $\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}$
end while
```

Update the posterior estimates of the topics based on the “E step.”

Online inference for LDA (with M. Hoffman and F. Bach)

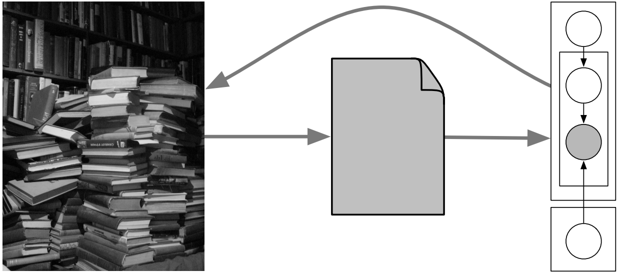


Sample one document ***Analyze it*** ***Update the model***

- Our goal is to use this (and related) models for analyzing massive collections of millions of documents.

- But, in the first step of batch inference we estimate the posterior for *every document* based on randomly initialized topics.

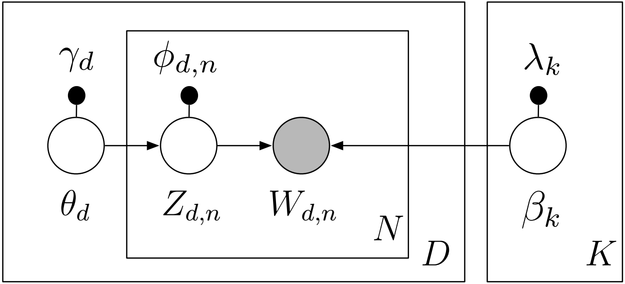
Online inference for LDA (with M. Hoffman and F. Bach)



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

- Online variational inference is much more efficient.
- It allows us to easily analyze millions of documents.
- It lets us develop topic models on streaming collections.

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.

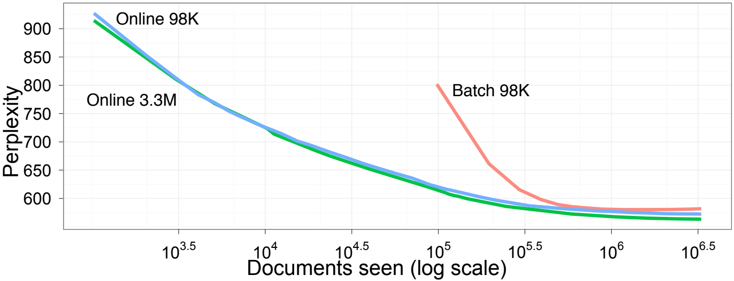
Online variational inference for LDA (sketch)

```
 1: Define an appropriate sequence of weights.
 2: Initialize topics randomly.
 3: for ever do
 4:     Choose a random document d.
 5:     repeat
 6:         Update the topic assignment variational parameters.
 7:         Update the topic proportions variational parameters.
 8:     until document objective converges
 9:     Compute topics as though d is the only document.
10:     Set the topics to a weighted average of the current topics and
        the topics from step 9.
11: end for
```

On-line variational inference for LDA

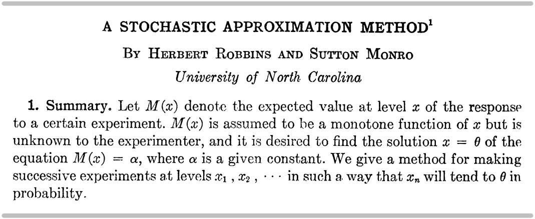
```
Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
**for** $t = 0$ to $\infty$ **do**
    Choose a random document $w_t$
    Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
    **repeat**
        Set $\phi_{t,n} \propto \exp\{\mathbb{E}_q[\log \theta_t] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
        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**
```

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[1]

BY HERBERT ROBBINS AND SUTTON MONRO

University of North Carolina

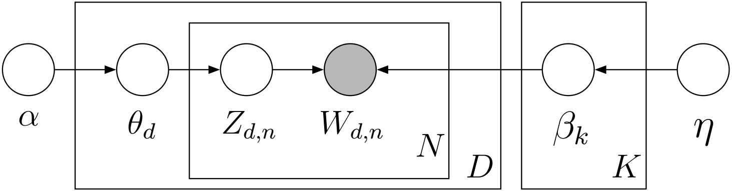
1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We 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).

Online inference is promising, in general

- Stochastic variational methods are a general way to approximate the posterior for massive/streaming data.
- No need to process the whole data set in advance; can easily link to web APIs and other data sources
- Powerful algorithm for topic modeling, and can be adapted hierarchical models for many types of data.
- Software and papers: www.cs.princeton.edu/$\sim$blei/

Latent Dirichlet allocation (flashback)

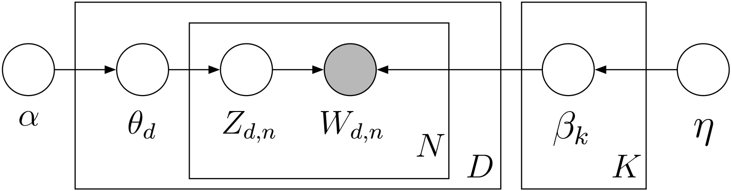


- This joint defines a posterior.

- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k

- Then use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, exploration, ...

Latent Dirichlet allocation (flashback)



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)

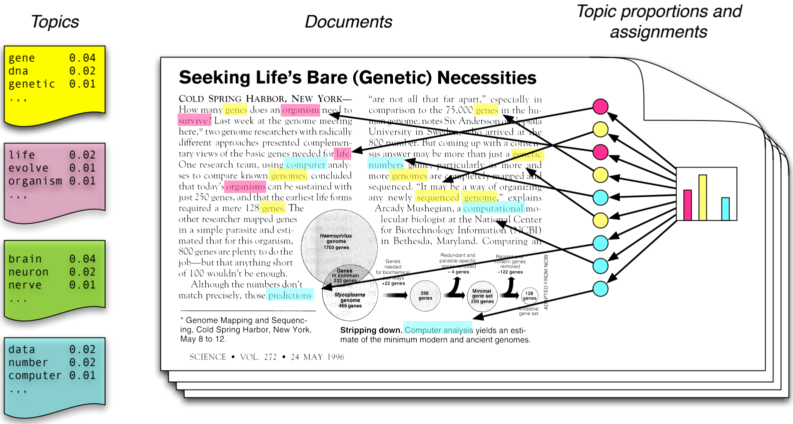
Also see Mukherjee and Blei (2009) and Asuncion et al. (2009).

Discussion

This tutorial

- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- What are some unsanswered 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

Topic proportions and assignments

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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

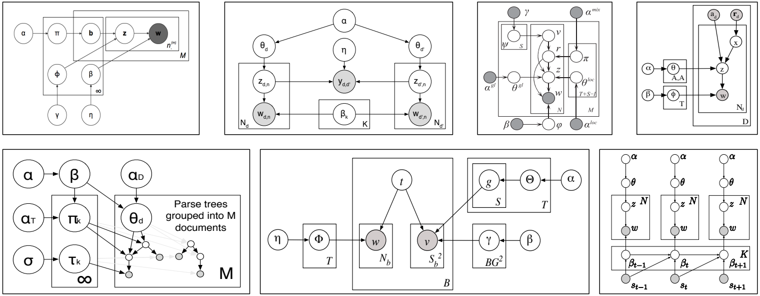
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- 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.

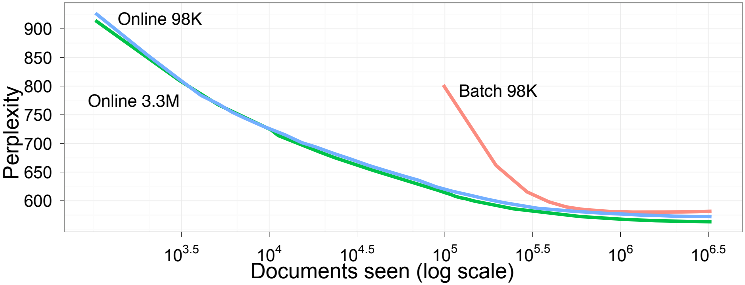
Extensions of LDA



- Topic models can be adapted to many settings

- Bayesian nonparametric topic models let the corpus determine the number of topics (or more complicated topic structure).

Posterior inference



- Posterior inference is the central computational problem.
- We discussed three algorithms
    - MCMC based on collapsed Gibbs sampling
    - Mean-field variational inference
    - Online variational inference

Some open issues

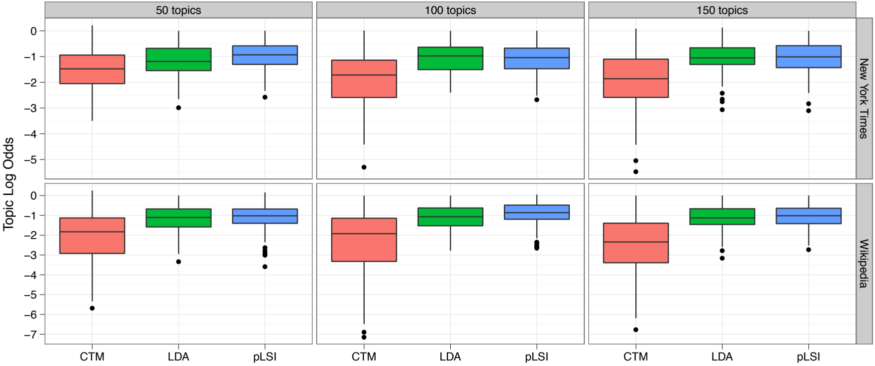
- **Model interpretation and model checking**
  Which model should I choose for which task?
  (Chang et al. 2009, Ramadge et al. 2009, Newman et al. 2010, Mimno and Blei 2011, Mimno et al. 2011)

- **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?

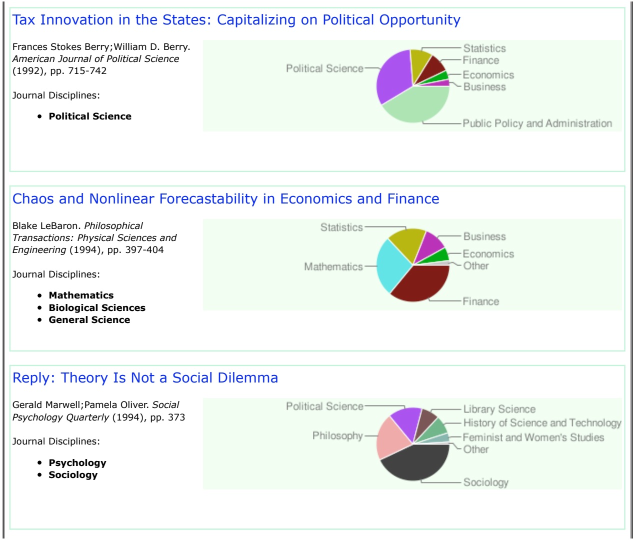
- **Theoretical understanding of approximate inference**
  What do we know about variational inference from either the statistical or learning perspective?

Interpretation I: Human studies of topic models



(see Chang et al. 2009 and Newman et al. 2010)

Interpretation II: Labelled LDA on JSTOR



Tax Innovation in the States: Capitalizing on Political Opportunity

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

Journal Disciplines:

- **Political Science**

Chaos and Nonlinear Forecastability in Economics and Finance

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

Journal Disciplines:

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

Reply: Theory Is Not a Social Dilemma

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

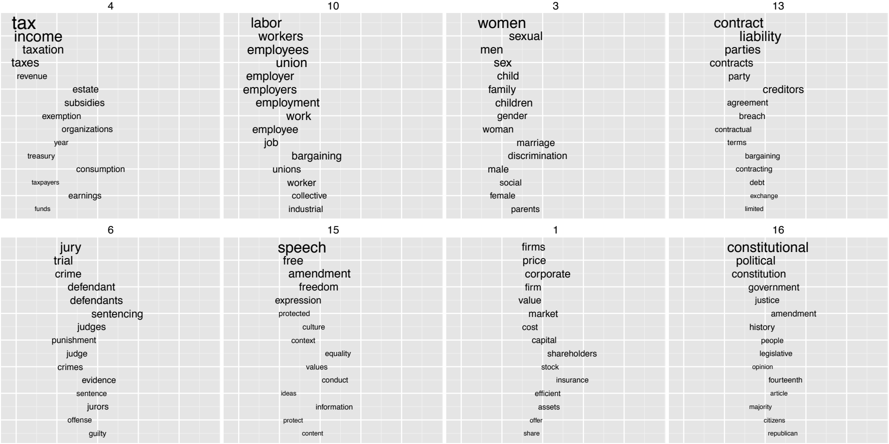
Journal Disciplines:

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

Political Science
Library Science
History of Science and Technology
Philosophy
Feminist and Women's Studies
Other
Sociology

(see Ramadge et al. 2009 and **Ramadge et al. 2011**)

Interptetation III: Mutual information discrepancy



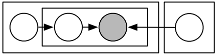
(see Mimno and Blei 2011)

Topic modeling resources

- The topic modeling mailing list is a good discussion group.
- Bibliography: http://www.cs.princeton.edi/∼mimno/
- Software and papers: http://www.cs.princeton.edu/∼blei/

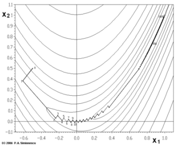
If you remember one picture...

Assumptions

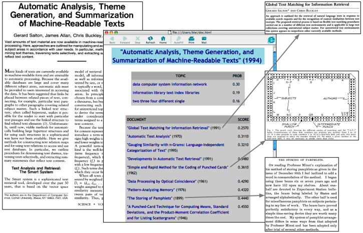


Data

Inference algorithm



Discovered structure



Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts
Gerard Salton, James Allan, Chris Buckley,
Text Analysis and Retrieval: The Smart System
Global Text Matching for Information Retrieval
Gerard Salton* and Chris Buckley
SUBJECTS
AUTHORS
SHALLOW PUNCH
INTERMEDIATE PUNCH
THE STORING OF PAMPHLETS.
file:///Users/blei/doc.html
Back
Forward
Stop
Refresh
Home
AutoFill
Print
Mail

"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

“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)