

Overview of Topic Models

February 20th

S&DS 171
YData: Text Data Science

Yale

Intro to Topic Modeling

Some of the following slides are from Dave Blei's tutorial on Topic Modeling

<http://www.cs.columbia.edu/~blei/topicmodeling.html>

A survey paper describing many of these ideas in more detail is here:

<http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf>

Topic modeling



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

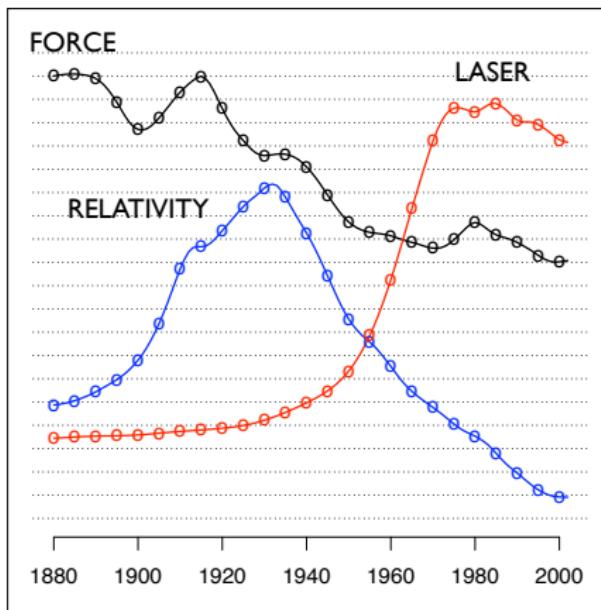
- ① Discover the hidden themes that pervade the collection.
- ② Annotate the documents according to those themes.
- ③ Use annotations to organize, summarize, and search the texts.

Discover topics from a corpus

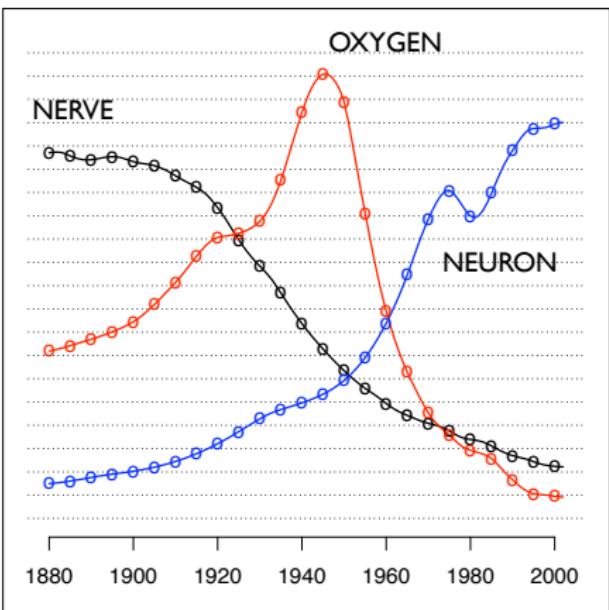
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

Model the evolution of topics over time

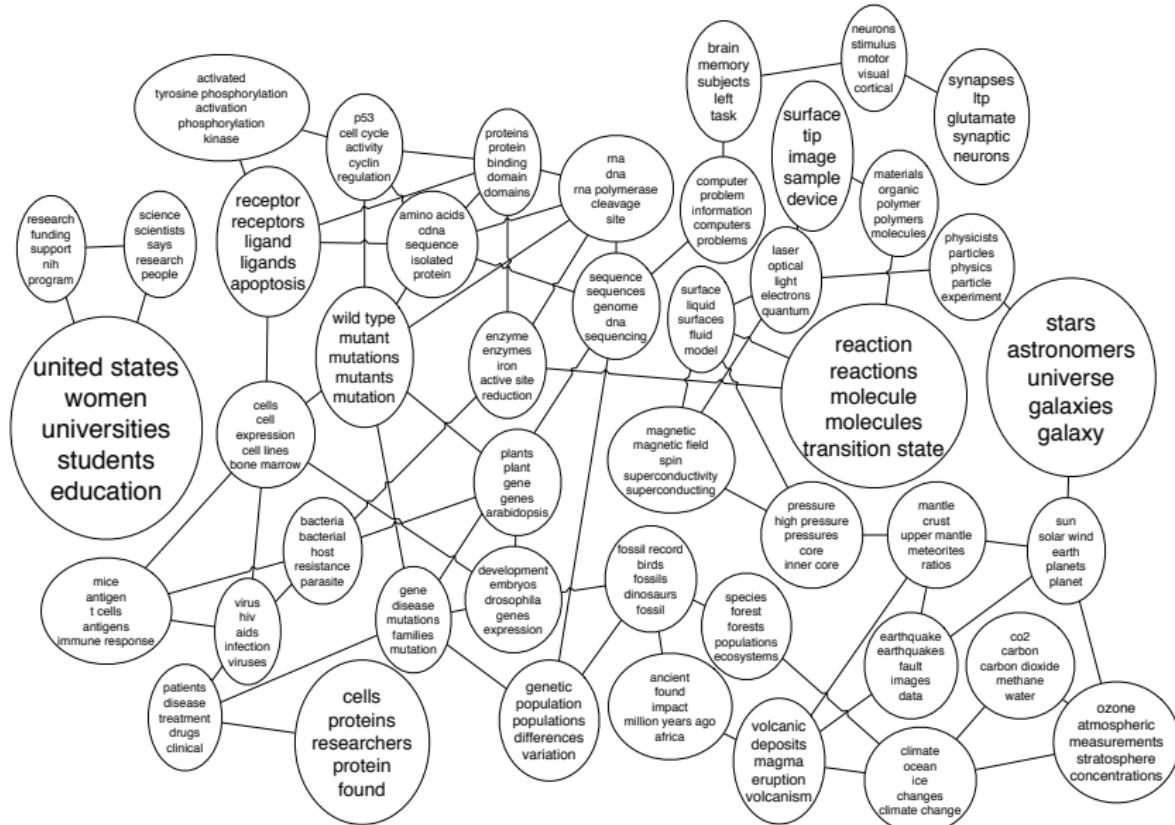
"Theoretical Physics"



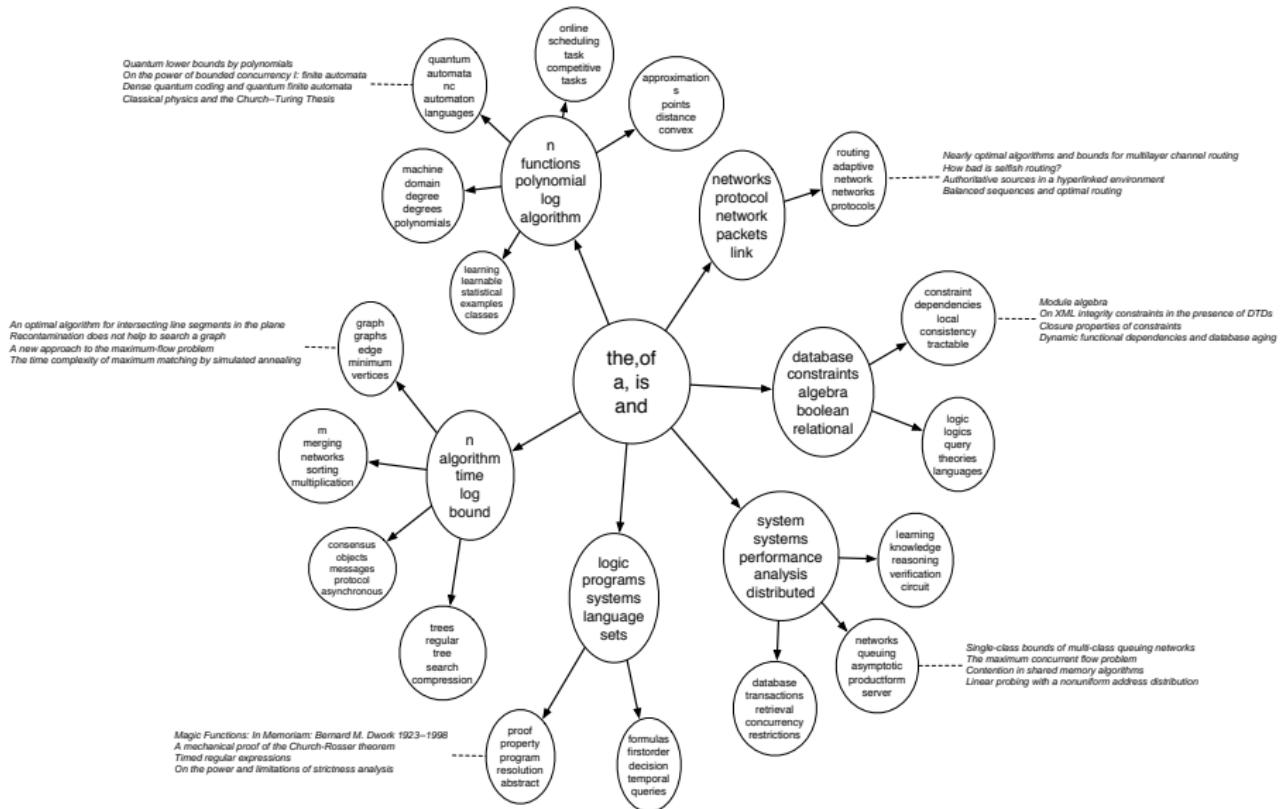
"Neuroscience"



Model connections between topics



Find hierarchies of topics



Annotate images



SKY WATER TREE
MOUNTAIN PEOPLE



SCOTLAND WATER
FLOWER HILLS TREE



SKY WATER BUILDING
PEOPLE WATER



FISH WATER OCEAN
TREE CORAL

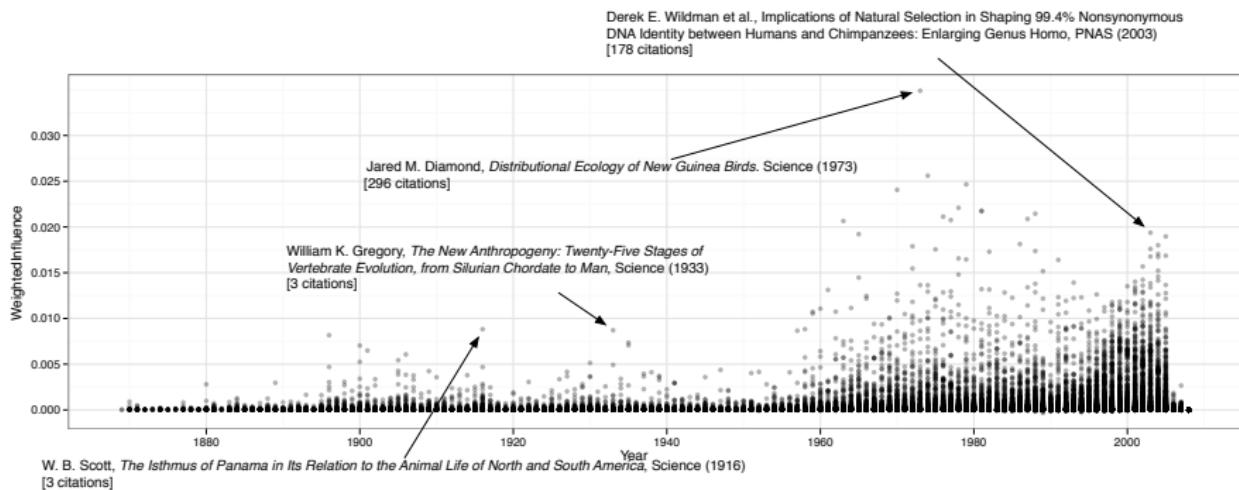


PEOPLE MARKET PATTERN
TEXTILE DISPLAY



BIRDS NEST TREE
BRANCH LEAVES

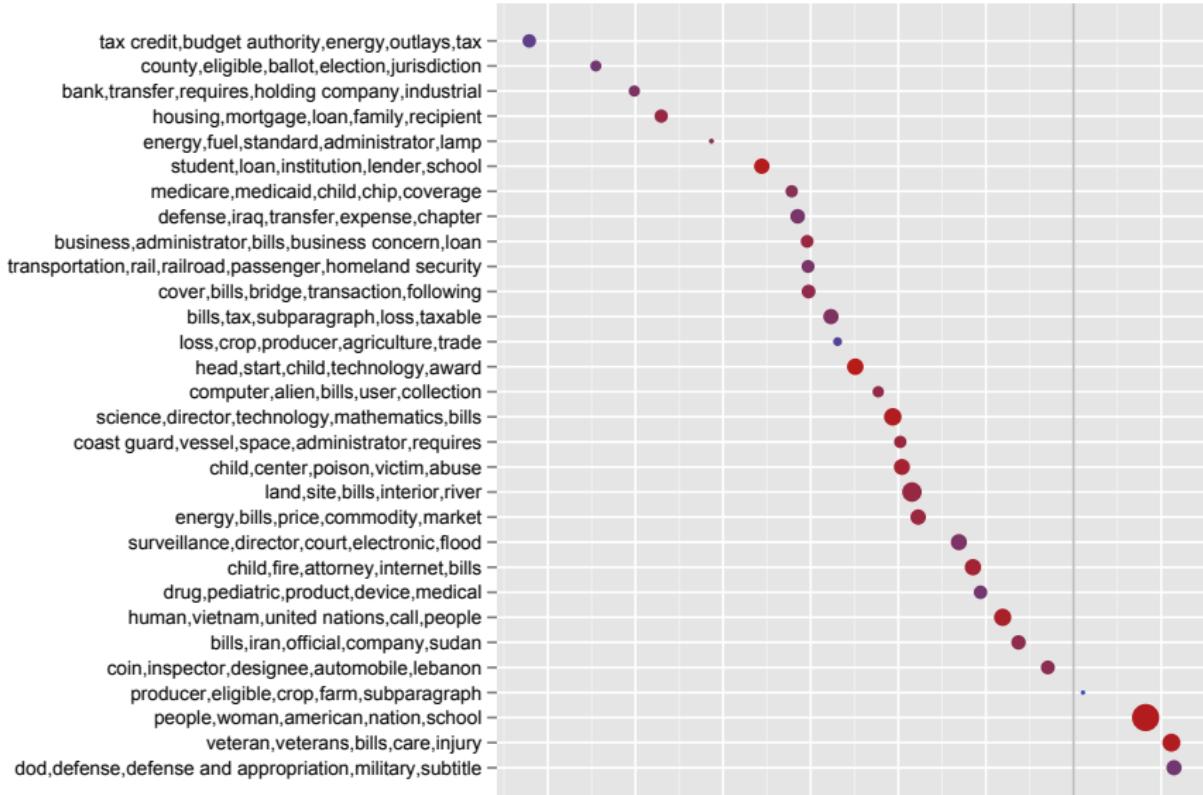
Discover influential articles



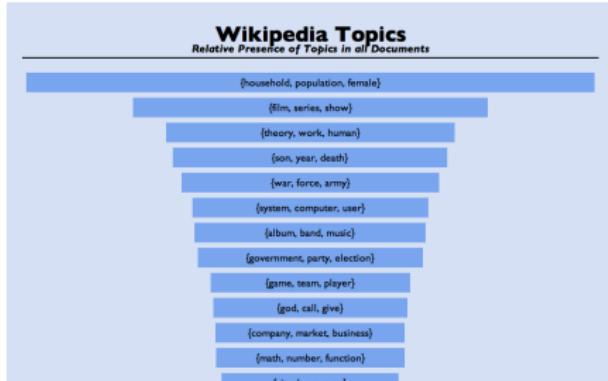
Predict links between articles

<p><i>Markov chain Monte Carlo convergence diagnostics: A comparative review</i></p> <p>Minorization conditions and convergence rates for Markov chain Monte Carlo</p> <ul style="list-style-type: none">Rates of convergence of the Hastings and Metropolis algorithms <p>Possible biases induced by MCMC convergence diagnostics</p> <ul style="list-style-type: none">Bounding convergence time of the Gibbs sampler in Bayesian image restorationSelf regenerative Markov chain Monte Carlo <p>Auxiliary variable methods for Markov chain Monte Carlo with applications</p> <p>Rate of Convergence of the Gibbs Sampler by Gaussian Approximation</p> <ul style="list-style-type: none">Diagnosing convergence of Markov chain Monte Carlo algorithms	<p>RTM (ψ_e)</p>
<p>Exact Bound for the Convergence of Metropolis Chains</p> <p>Self regenerative Markov chain Monte Carlo</p> <p>Minorization conditions and convergence rates for Markov chain Monte Carlo</p> <ul style="list-style-type: none">Gibbs-markov models <p>Auxiliary variable methods for Markov chain Monte Carlo with applications</p> <p>Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models</p> <ul style="list-style-type: none">Mediating instrumental variablesA qualitative framework for probabilistic inferenceAdaptation for Self Regenerative MCMC	<p>LDA + Regression</p>

Characterize political decisions



Organize and browse large corpora



{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

A pie chart illustrating the distribution of topics related to Stanley Kubrick. The largest segment is 'film, series, show', followed by 'theory, work, human' and 'son, year, death'.

Topic	Relative Presence (approx.)
film, series, show	45%
theory, work, human	25%
son, year, death	15%
black, white, people	5%
god, call, give	5%
math, energy, light	2%

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}
ideas	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}

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?*

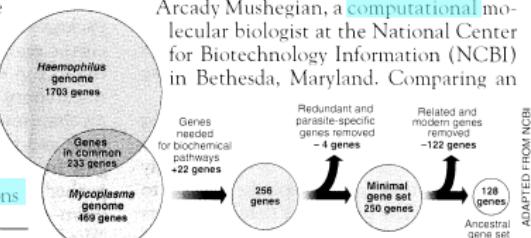
Latent Dirichlet allocation (LDA)

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

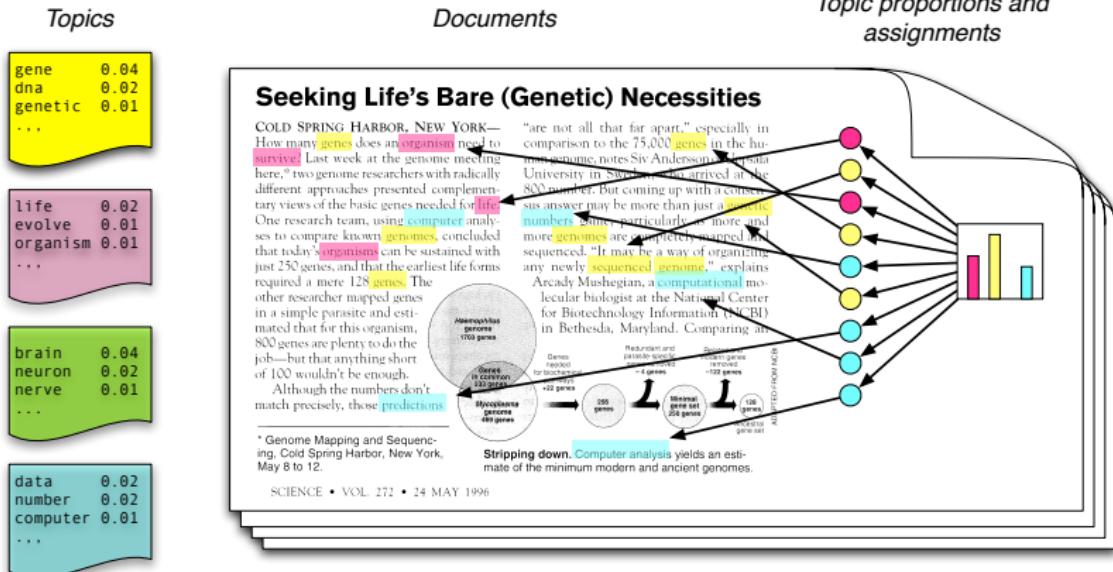


ADAPTED FROM NCBI

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

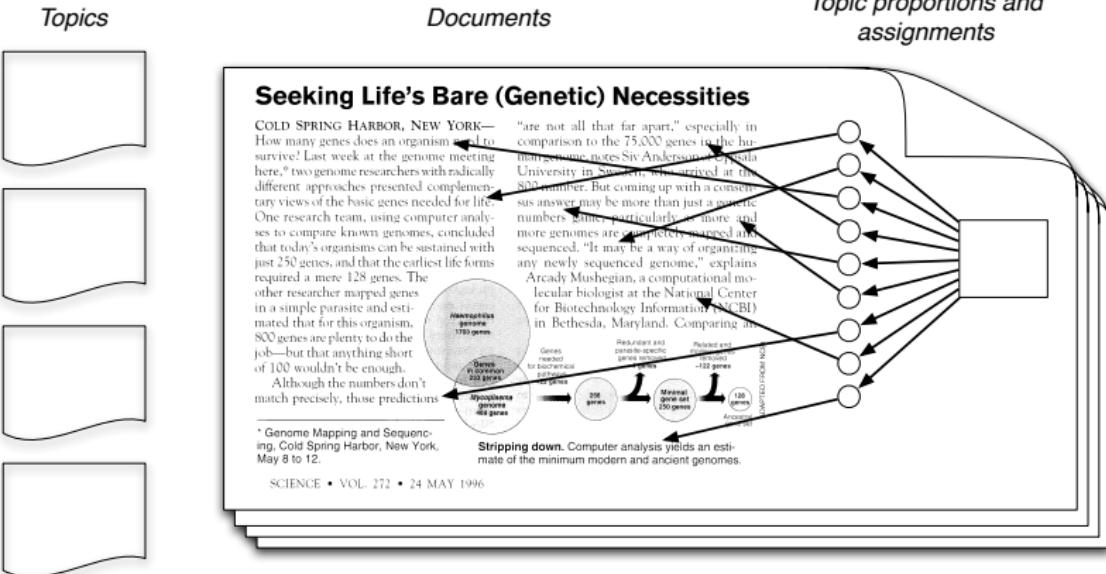
Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



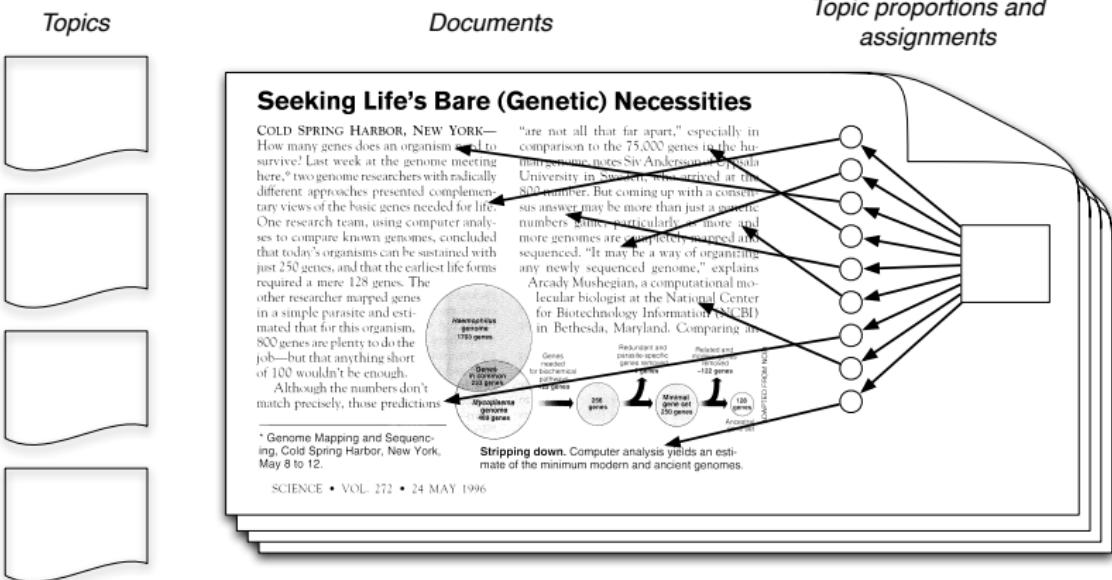
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



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

The posterior distribution



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

Summary

- Topic models automatically extract “semantic themes” from large document collections
- Based on latent variable models
- Can be useful for a wide variety of data