

# Seeking Life's Bare (Genetic) Necessities

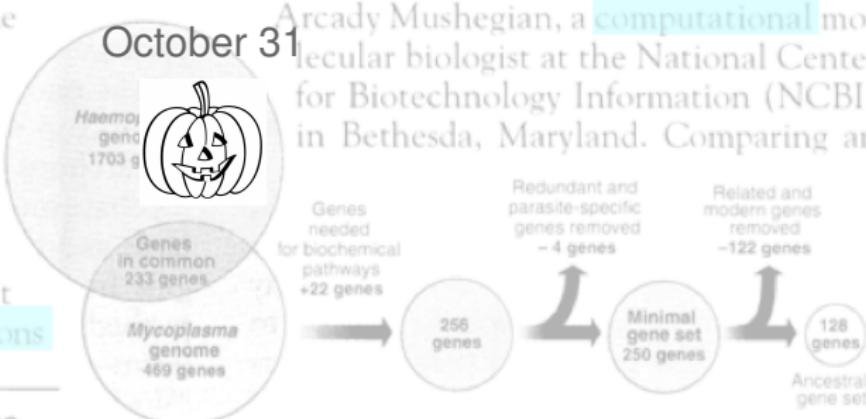
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be stripped down to 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 game, particularly as more and more genomes are completely mapped and analyzed. "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

## S&DS 265 / 565 Introductory Machine Learning Topic Models

October 31



**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.

# Checkpoint

- Assignment 4 (LMs and embeddings) due next Thursday
- Quiz 4 this Thursday
- Final exam: Friday, December 15, 2pm
- <https://registrar.yale.edu/general-information/final-exams>

# 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?*

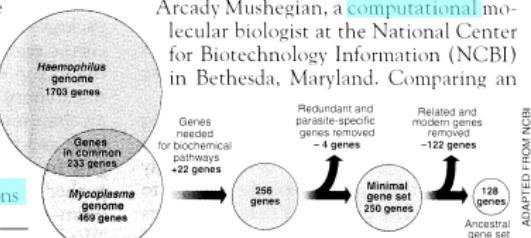
# Latent Dirichlet allocation (LDA)

## 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



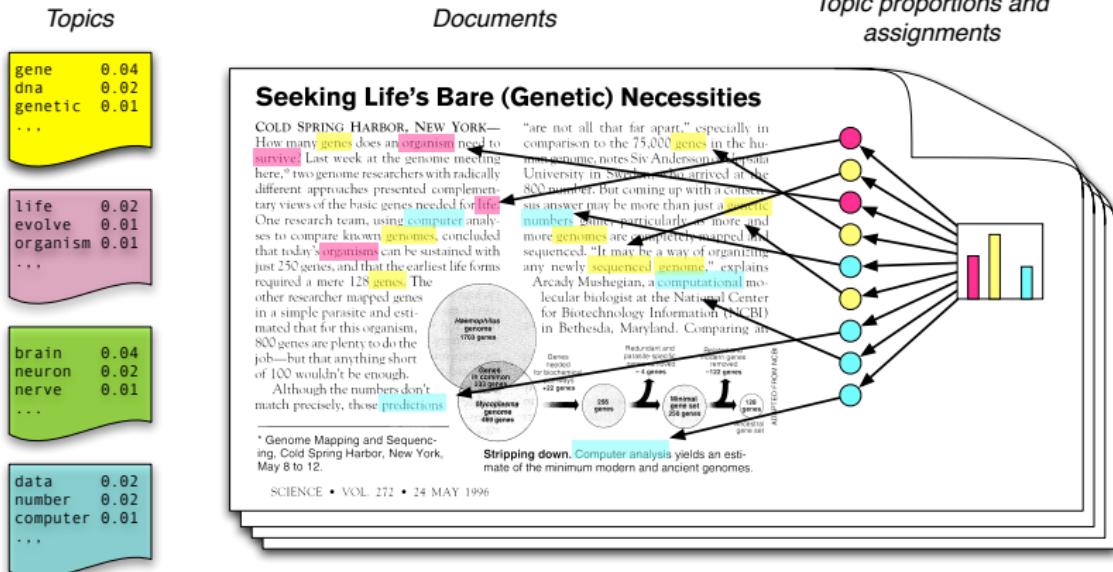
ADAPTED FROM NCBI

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

SCIENCE • VOL. 272 • 24 MAY 1996

**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

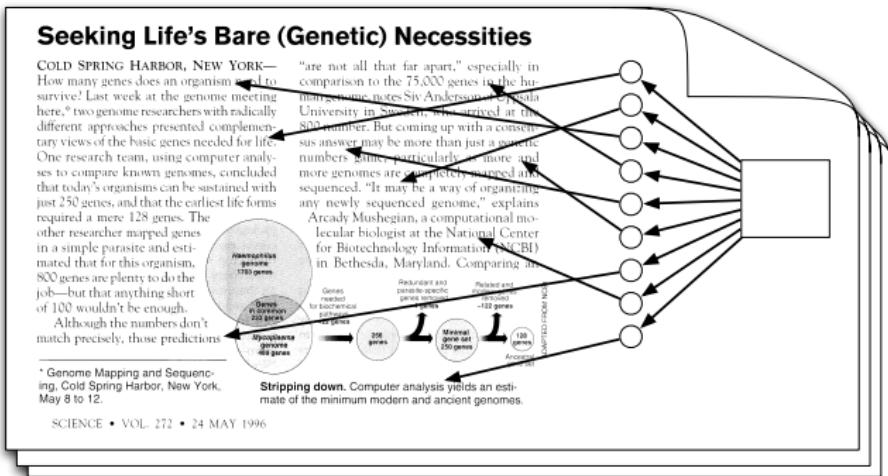
## *Topics*



## *Documents*

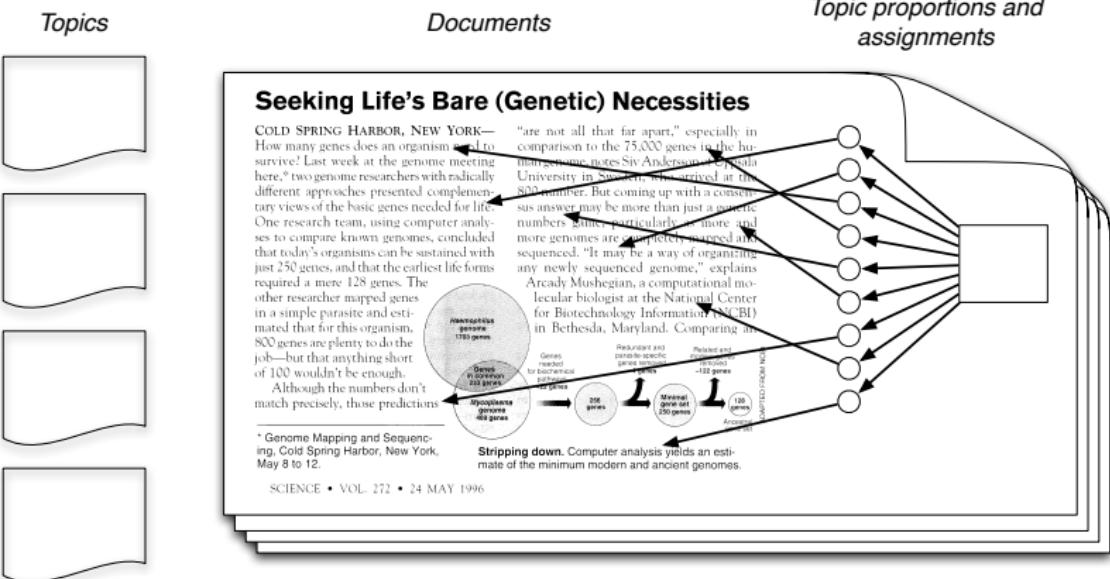


## *Topic proportions and assignments*



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

# A little dicey



Let's illustrate this generative process

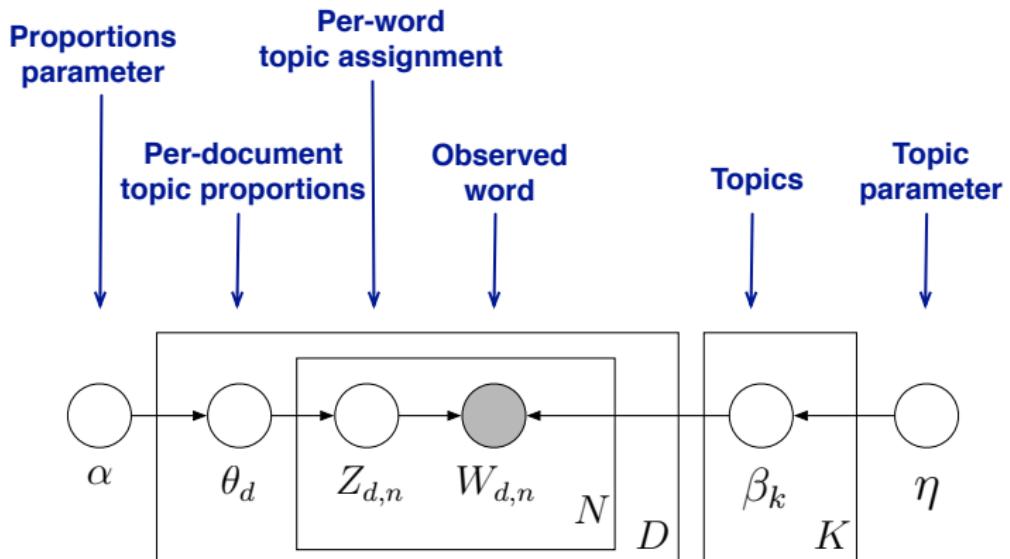
# A little dicey



Left: Four topics  $\beta_k$  drawn from the prior; used across documents

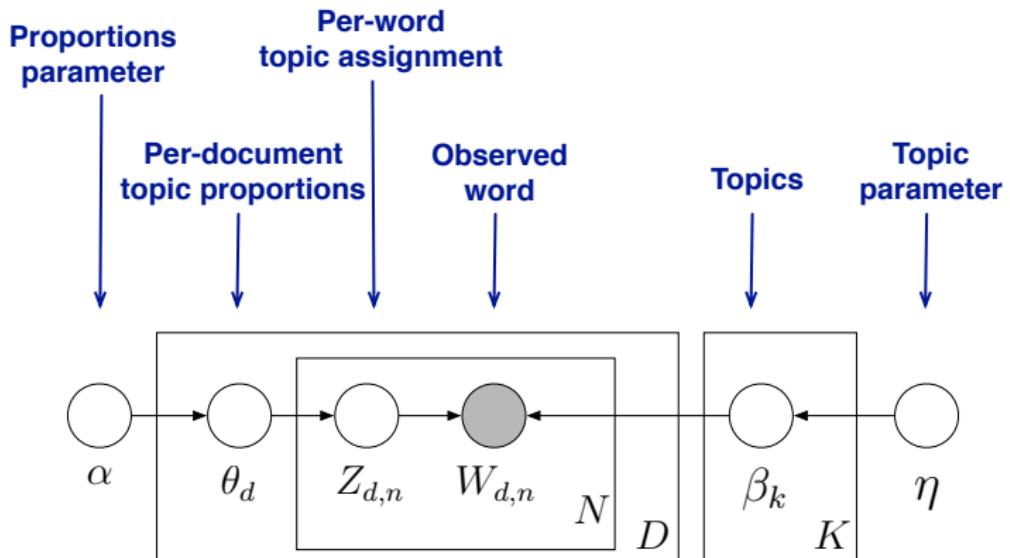
Right: A distribution over topics  $\theta$ ; used for a specific document

# LDA as a graphical model



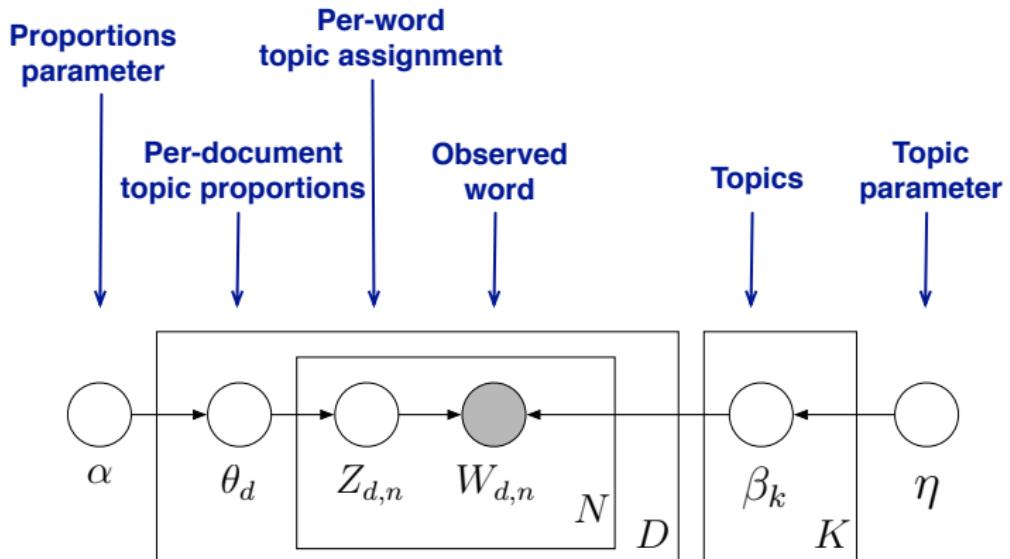
- Encodes our assumptions about the data
- Connects to algorithms for computing with data
- See *Pattern Recognition and Machine Learning* (Bishop, 2006).

# LDA as a graphical model



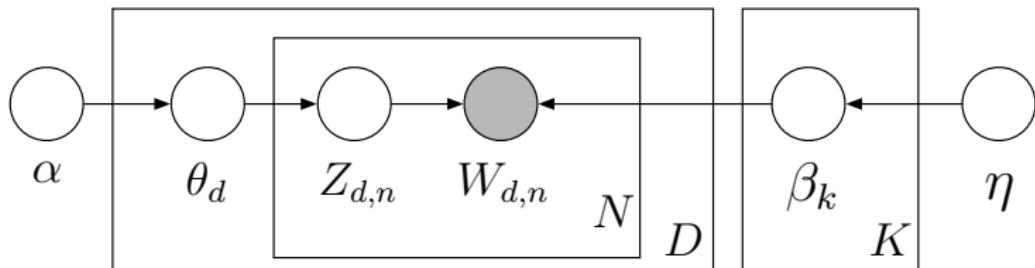
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

# LDA as a graphical model



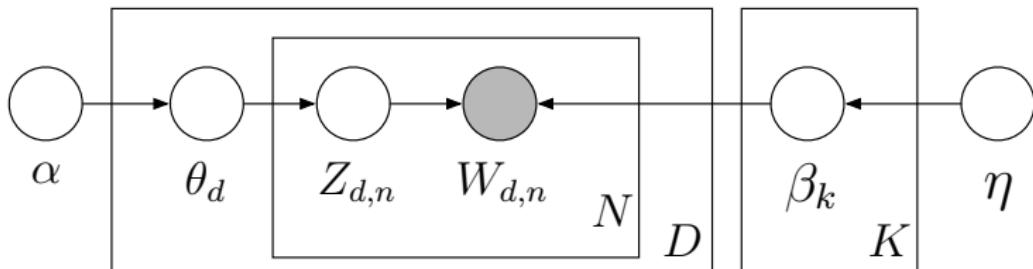
$$\prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

# LDA



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

# Example inference



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

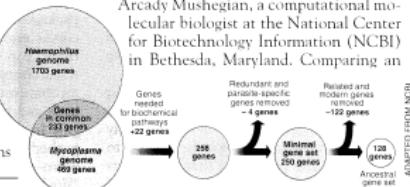
# Example inference

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>6</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

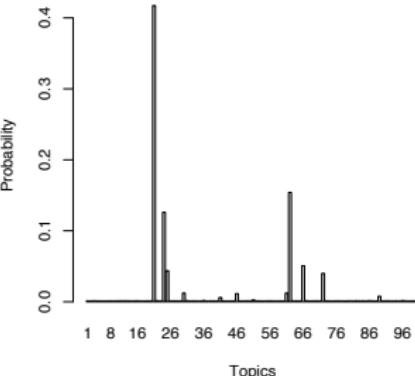
Although the numbers don't match precisely, those predictions

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



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

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



# Example inference

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

# 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).

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-



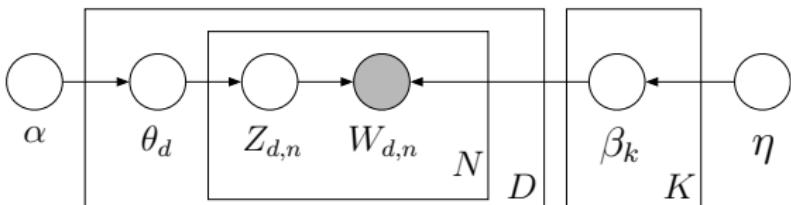
**Cannibalism and chaos.**  
The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

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	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

# 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.

# Approximate inference

- We'll talk a bit about Gibbs sampling
- Variational inference saved for Intermediate Machine Learning (S&DS 365, IML)

# Idea behind Gibbs sampling

- Only the assignments  $Z_{n,d}$  are needed
- From these we can infer the proportions  $\theta_d$  (per document)
- And the topics  $\beta_k$  (per corpus)
- The following slides indicate how in a toy example

# Idea behind Gibbs sampling

At each time step in the algorithm, we have an assignment  $Z_{n,d}$  of a topic to each word  $w_{n,d}$  in every document  $d$

Repeat forever:

- Select a word  $w_{n,d}$
- Holding all of the other assignments  $Z_{n',d'}$  fixed, calculate the probability distribution over  $Z_{n,d}$  for that word
- Sample from that distribution to get a (potentially new) assignment  $Z_{n,d}$

## Toy example: 3 topics, 3 docs

$w$	$z$	$w$	$z$	$w$	$z$
meth	2	drug	3	inning	1
father	1	baseball	2	mother	3
divorce	3	hit	1	son	1
drug	1	inning	2	hit	2
illegal	1	steroids	1	baseball	3

## Toy example: 3 topics, 3 docs

$w$	$z$	$w$	$z$	$w$	$z$
meth	1	drug	1	inning	1
father	2	baseball	2	mother	1
divorce	2	hit	2	son	2
drug	1	inning	3	hit	3
illegal	3	steroids	1	baseball	2

## Toy example: 3 topics, 3 docs

$w$	$z$	$w$	$z$	$w$	$z$
meth	1	drug	1	inning	2
father	3	baseball	2	mother	3
divorce	3	hit	2	son	3
drug	1	inning	2	hit	2
illegal	1	steroids	1	baseball	2

# **Extensions**

# Modeling richer assumptions

- Correlated topic model
- Dynamic topic model

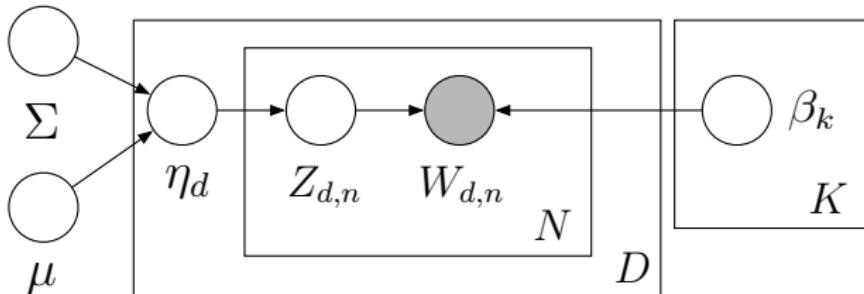
# Shortcoming of the Dirichlet

- Dirichlet for topic proportions:

$$p(\theta | \alpha) \propto \prod_{j=1}^k \theta_j^{\alpha_j - 1}$$

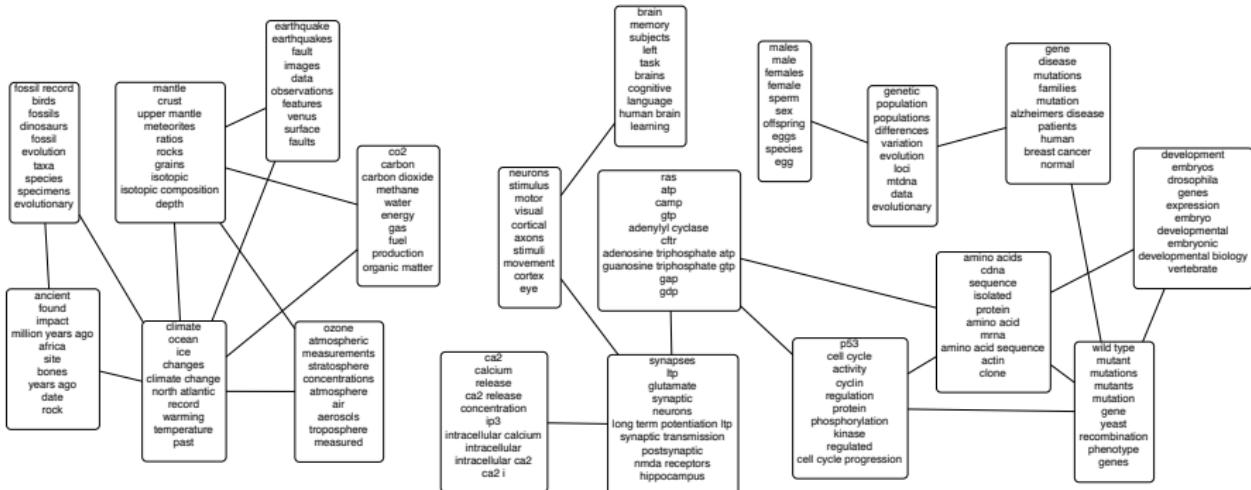
- Near independence of components makes it an unrealistic model
  - ▶ An article about *fossil fuels* is more likely to also be about *geology* than about *genetics*

# Correlated topic model

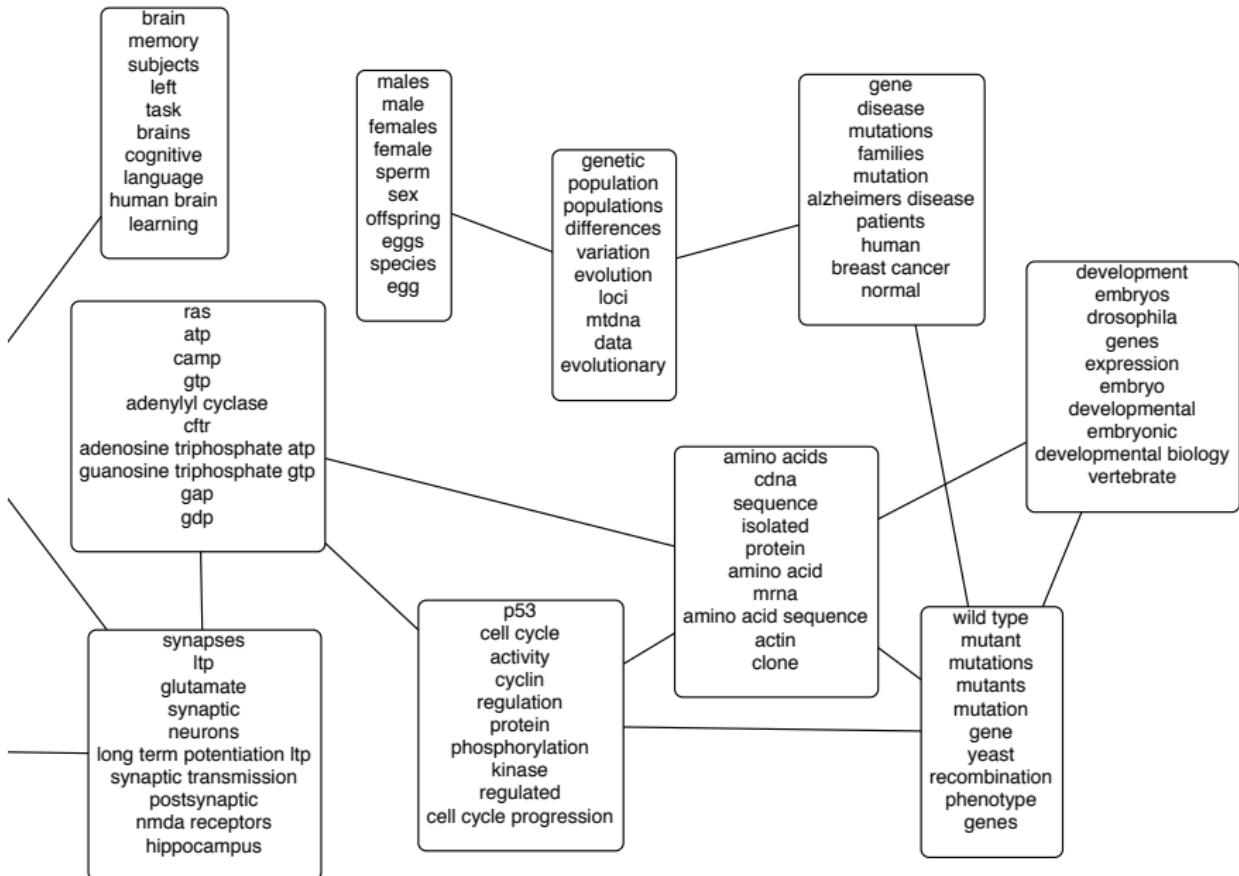


- Draw topic proportions from a logistic normal.
- Useful for:
  - ▶ providing a “map” of topics and how they are related;
  - ▶ better prediction via correlated topics.
- Sacrifice conjugacy: Posterior over  $\theta$  does not have same form

# Topic graphs



# Topic graphs



# Modeling Evolution of Topics

- In LDA, document order doesn't matter
- The topics should *evolve* over time
  - ▶ “Cleaning Birds” (1883)
  - ▶ “Interspecific Brood Parasitism in Blackbirds (Icterinae): A Phylogenetic Perspective” (1992)
- Many document collections have such dynamics: emails, query logs, news articles, etc.

# ***Science 1893 ⇒ Science 1976***

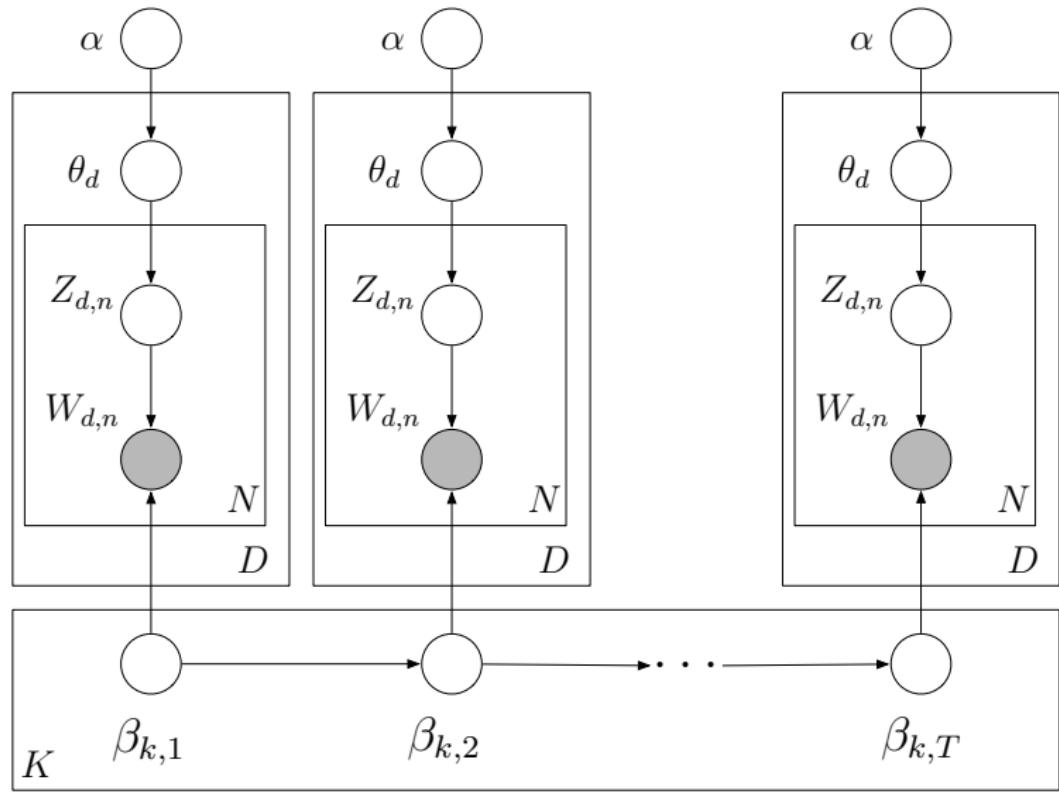
Administration	Limnology	Astronomy	Psychology
association	water	observatory	mind
meeting	lake	observations	nature
american	sea	stars	say
committee	waters	time	science
congress	lakes	made	psychology
members	gulf	astronomical	work
held	great	comet	knowledge
international	depth	star	truth
meetings	river	observed	religion
section	stream	telescope	human

Administration	Limnology	Astronomy	Psychology
house	water	mass	human
congress	concentrations	radio	attempts
science	mercury	objects	theory
bill	fish	astronomy	learning
nsf	samples	xray	ideas
president	soil	stars	new
budget	lake	astronomical	memory
office	ppm	sources	psychology
committee	concentration	observations	behavior
new	waters	observatory	complex

# Time series topic model

1. Allow topics evolve between time slices
3. For each document in the current time slice:
  - a. Select a distribution over topics;
  - b. Generate the words from the resulting topic mixture.

# Dynamic topic models



Topics drifting in time

## Time-corrected document similarity

## The Brain of the Orang (1880)

300

EXTRACT

Prithivie in their case, which were submitted to us in the *Journal of Diseases and its Remedies* or *Transactions*:—no objection being made by either of them to our doing so. We have, however, been informed by Mr. Prithivie that the remedies under his name are not satisfactory to him. We therefore request our readers to consider the whole subject.

George F. Barker, Frederick C. March, and Prithivie J. Blagden are preparing more elaborate and detailed descriptions of their respective papers, and promise them to us in due time.

In the Orient, the paroxysmal febrile disease does not always, being separated by us under the title "diaphoresis de passage intime" of *Goutoulas*, or "seizure de la fièvre de l'oreille" of *Champanier*, from the tertian ague, and is separated as an anomaly more than as a disease.

According to *Goutoulas*, this disposition obtains in Goutoulas' family, and in that of *Champanier*.

In the family *Champanier*, however, on the 1st of August, 1870, found the paroxysmal febrile disease passing into the tertian ague, and the tertian ague into the paroxysmal febrile disease, through the agency of *Ratson*.

#### THE MEAN OF THE CHAMPS

BY HENRY C. DUNHAM, M.D.  
  
The State of Oregon has signed by Tilden, Gresham, Schenck, and von Koln, the  
Chairmen of the Board of Education, the  
law (Illustration, page 1), and at the importance of  
the measure may be gauged from the  
various views of my Oregon's laws (Fig. 1) in which  
was reflected from the day or two hours ago  
that the bill had been introduced, and  
a little of the number of the 16 bills introduced  
and then discharged by the Legislature, and  
a single example of one so passed.



1

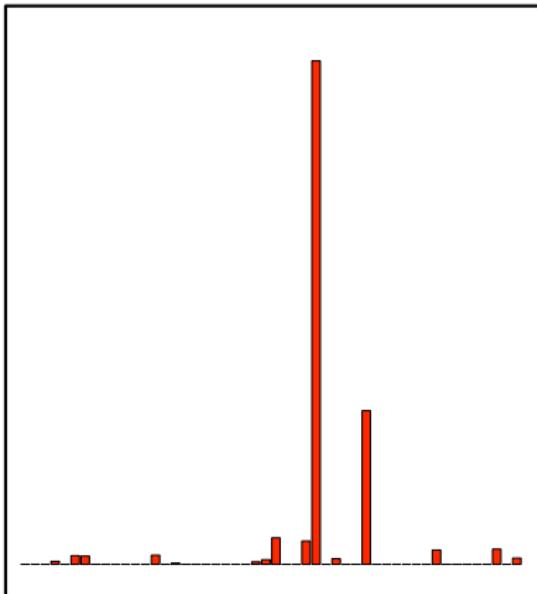
the brain of the orang, chimpanzee, and man are the same; they are certain racial differences, however, in their digestions in all three. The species of *Orang* in Sumatra and Borneo differ from each other in using only a slightly bimolar dentition; the molar tooth is small. The species of *Balado*, or central Sumatra, has a more complete dentition than the orang in the Orang than in man. It differentiates the orang from the orang's parent race. The pachymetopic part is well developed in the orang, but the mesopachymetopic parts decrease dorsally as the mental apertures of the upper and lower jaws increase.



6

**occipital fissure;** externally it is continuous with the occipital lobe, as the first occipital gyrus, anteriorly it is separated from the posterior central convolution by a sulcus which runs parallel to the cerebral fissure. There is in the Cerebellum a sulcus which separates the parahippocampal gyrus which divides the upper parital lobe into lateral and median positions. The parahippocampal gyrus is the space on the medial side of the parital lobe between the parahippocampal

\* *Transactions of the Academy of Natural Sciences, Phil., etc.*



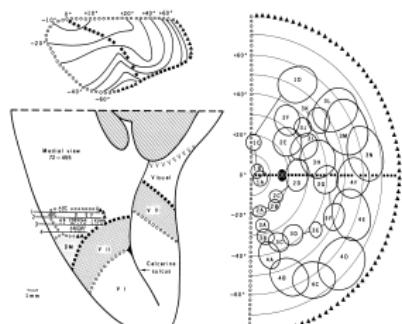
# Time-corrected document similarity

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

present by recording responses of the medial occipital-parietal cortex were plotted with描记technique using techniques in five owl monkeys (2). The monkeys were anesthetized, killed, and prepared for recording. Tissue and pterygomastoid incisions were used to expose the medial occipital-parietal cortex or occasionally from single resections it was exposed by resecting the adjacent temporal and/or lateral occipital-parietal cortices. Recording fields were plotted by moving microelectrodes across the exposed surface of a translucent plastic hemisphere centered in front of the contralateral eye. A translucent plastic disk was projected onto the plane hemispherical screen with the aid of Fernand and Chou (5). The preferred eye usually was

covered with a opaque shield. Electrode tracks and recording sites were extracted from histological sections and photographs of the intact brain.

In the present report we focus our most complete mapping of the visual field area, data obtained in the other four experiments are summarized. The visual field organization is demonstrated in Fig. 1, which illustrates the organization of the other cortical visual areas that have been mapped by the cat monkey, the border between the visual area and the auditory area, and the location of the upper portion of the horizontal meridian. In other experiments in the dorsomedial area of the medial occipital-parietal cortex near its caudal border with the medial temporal area, we found that the receptive fields recorded adjacent to the medial area were located in the lower quadrants near the vertical meridian and the peripheral portions of the lower visual quadrant. Dorsally, the medial area is adjacent by post-



# Exploring the UN General Debates with Dynamic Topic Models



Luke Lefebure [Follow](#)  
Oct 17, 2018 · 11 min read



Credit: [Vladislav Klapin](#) on [Unsplash](#)

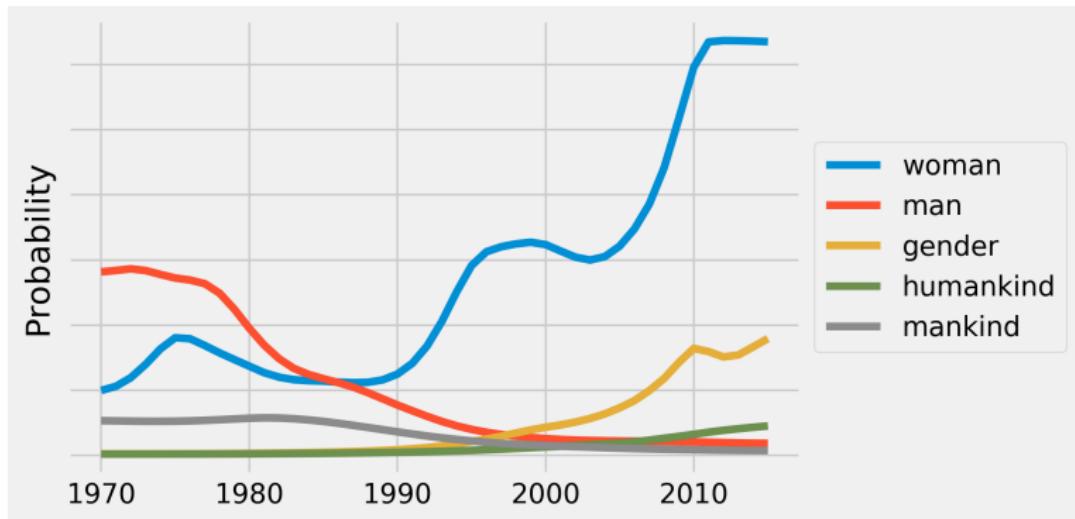
<https://towardsdatascience.com/>

exploring-the-un-general-debates-with-dynamic-topic-models-72dc0e307696

# Human rights

	1970	1980	1990	2000	2010
0	right	right	right	right	right
1	human	human	human	human	human
2	people	people	freedom	law	law
3	international	freedom	people	democracy	woman
4	principle	international	democracy	respect	freedom
5	justice	political	respect	international	respect
6	freedom	principle	law	people	people
7	law	respect	international	freedom	democracy
8	state	justice	principle	principle	rule
9	must	social	state	must	international

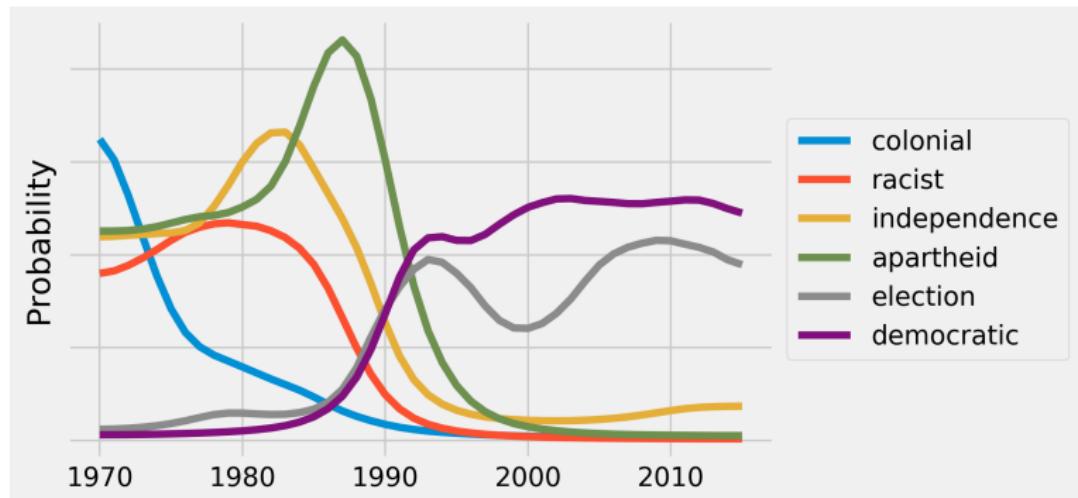
# Human rights



# Apartheid

	1970	1980	1990	2000	2010
0	africa	africa	africa	african	african
1	african	south	south	peace	country
2	south	african	african	africa	government
3	colonial	namibia	apartheid	country	africa
4	people	people	people	government	people
5	regime	regime	government	community	peace
6	southern	independence	country	democratic	political
7	government	apartheid	namibia	international	community
8	territory	racist	community	republic	democratic
9	apartheid	southern	process	effort	national

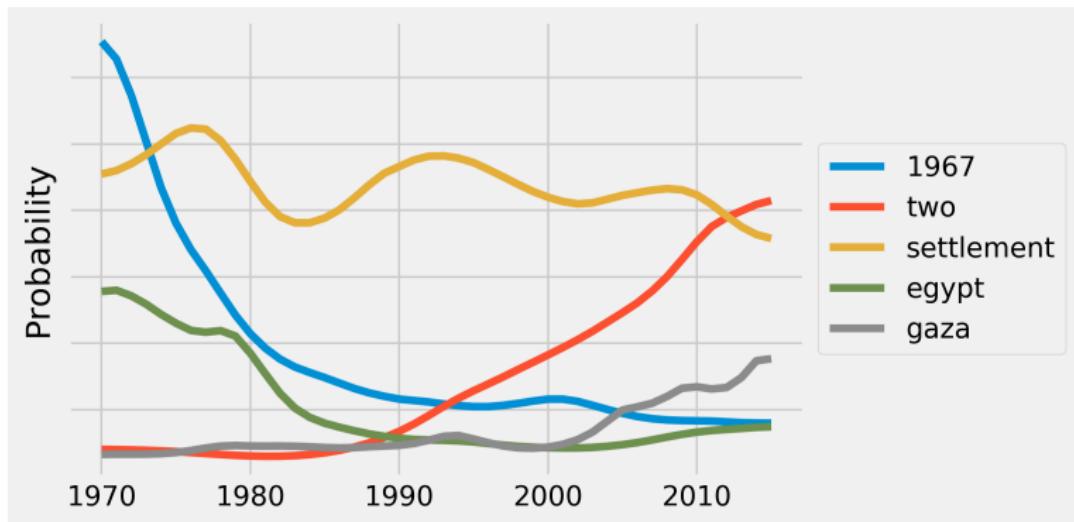
# Apartheid



# Arab-Israeli conflict

	1970	1980	1990	2000	2010
0	arab	palestinian	peace	peace	peace
1	israel	israel	east	east	state
2	east	right	middle	resolution	palestinian
3	middle	east	arab	palestinian	solution
4	peace	people	people	middle	east
5	territory	arab	palestinian	security	international
6	resolution	middle	international	people	israel
7	1967	peace	israel	israel	middle
8	palestinian	territory	kuwait	state	arab
9	israeli	state	right	international	people

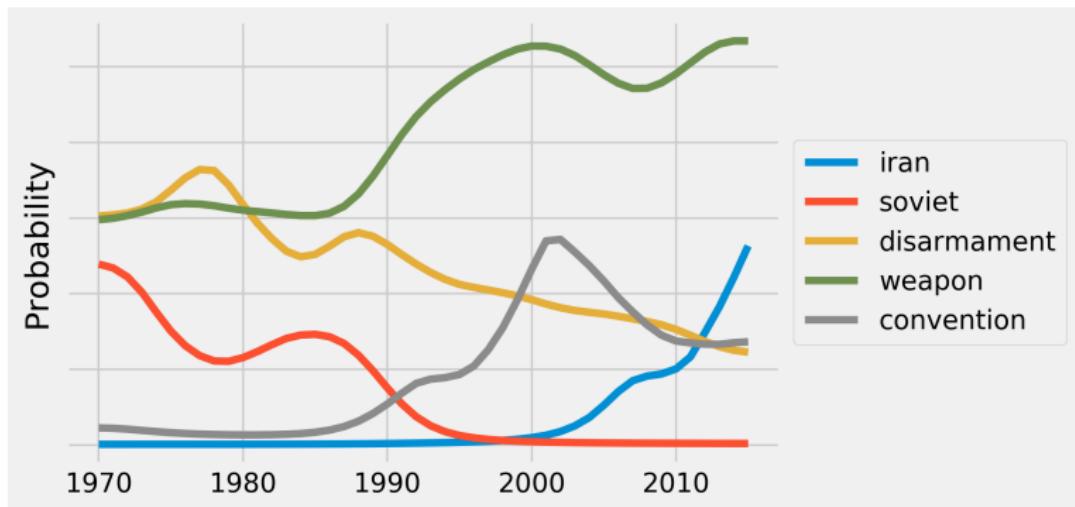
# Arab-Israeli conflict



# Nuclear arms

	1970	1980	1990	2000	2010
0	nuclear	nuclear	nuclear	weapon	nuclear
1	disarmament	disarmament	weapon	nuclear	weapon
2	weapon	weapon	disarmament	convention	non
3	soviet	arm	treaty	arm	proliferation
4	arm	state	arm	treaty	arm
5	treaty	race	state	disarmament	treaty
6	union	military	chemical	proliferation	international
7	agreement	treaty	agreement	international	disarmament
8	power	soviet	proliferation	non	convention
9	state	power	soviet	destruction	state

# Nuclear arms



# Tutorials

<https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d>

<https://towardsdatascience.com/>

[latent-dirichlet-allocation-intuition-math-implementation-and-visualisation-63ccb616e094](https://towardsdatascience.com/latent-dirichlet-allocation-intuition-math-implementation-and-visualisation-63ccb616e094)

# Summary

- Topic models automatically extract “semantic themes” from large document collections
- Use mixtures and latent variables
- Estimating Bayesian posterior done with Gibbs sampling
- Many extensions are possible