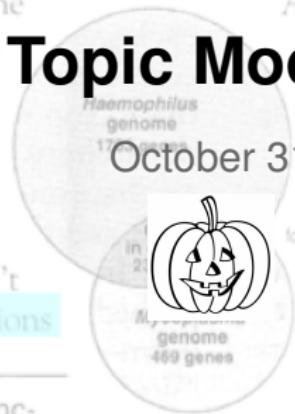


COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the *Escherichia coli* genome 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 a genome," explains Arcady Mushegian, a computational molecular biologist at the National Center of Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Topic Models



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?*

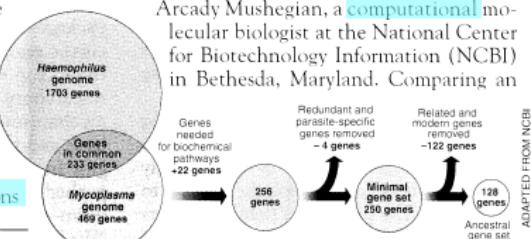
Latent Dirichlet allocation (LDA)

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



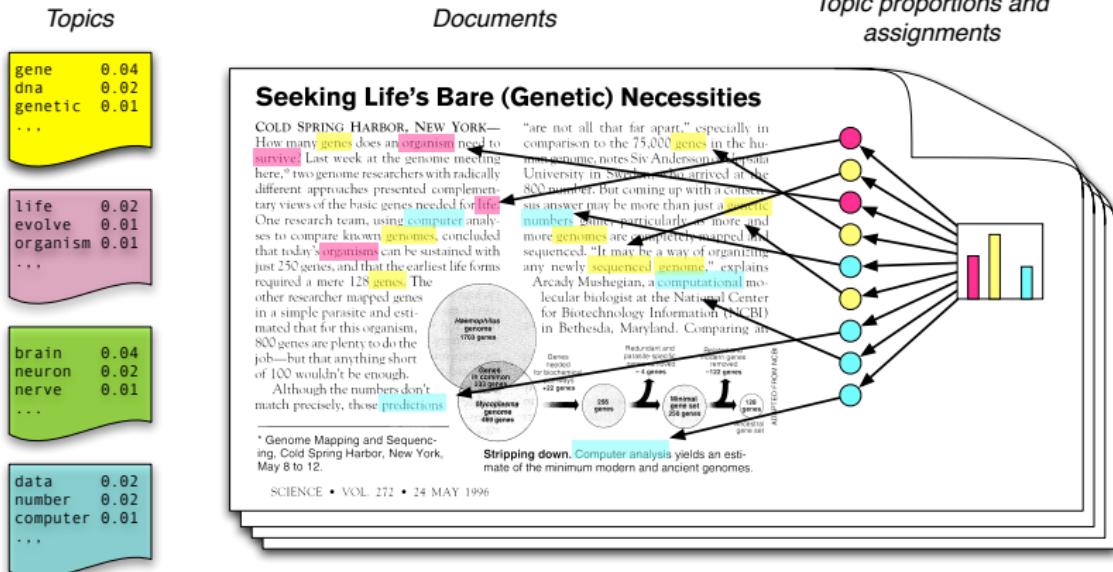
ADAPTED FROM NCBI

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

SCIENCE • VOL. 272 • 24 MAY 1996

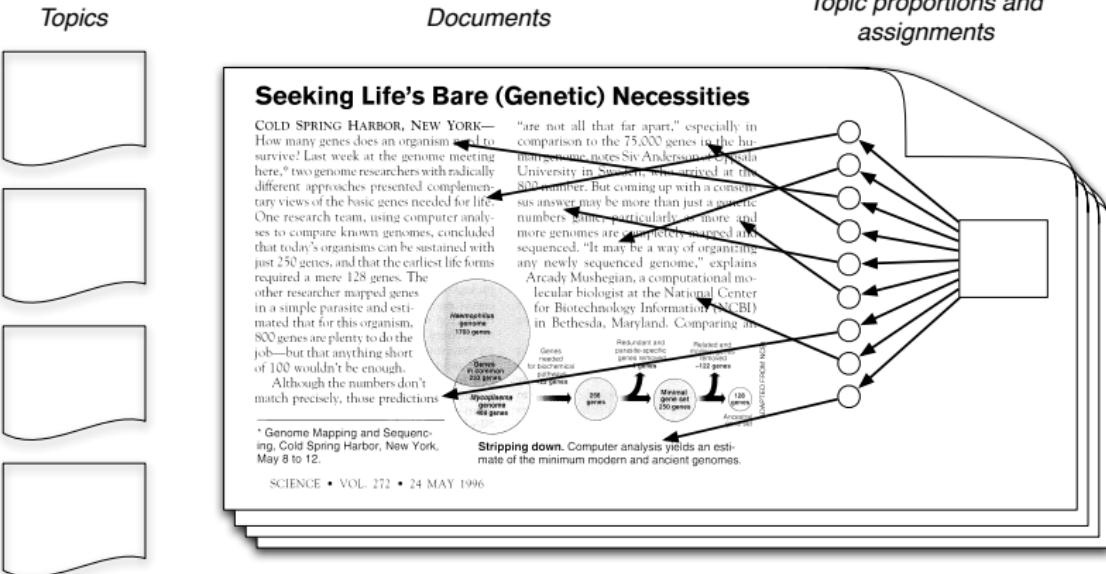
Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



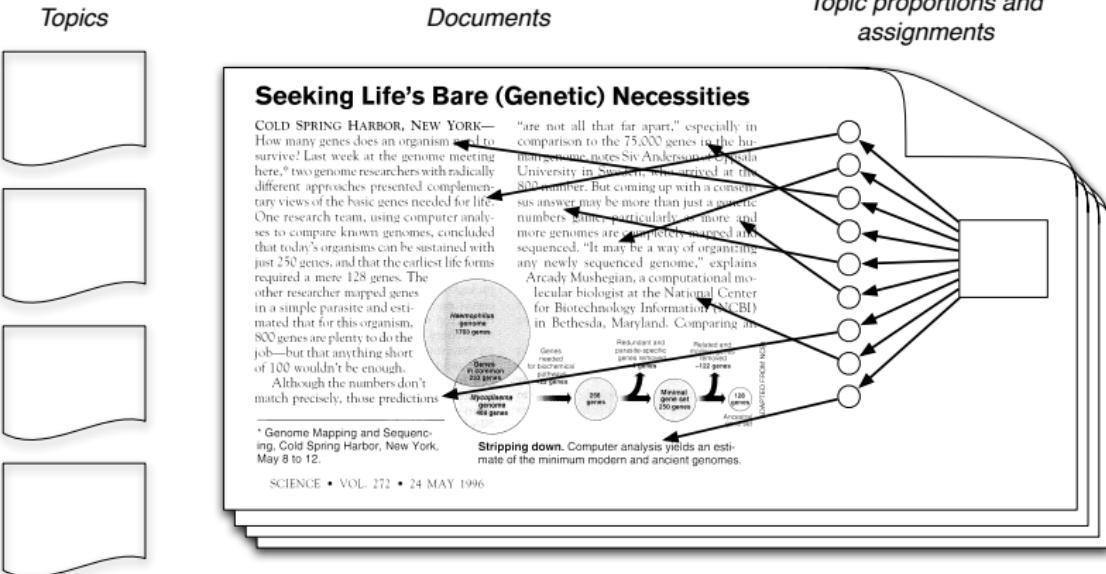
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



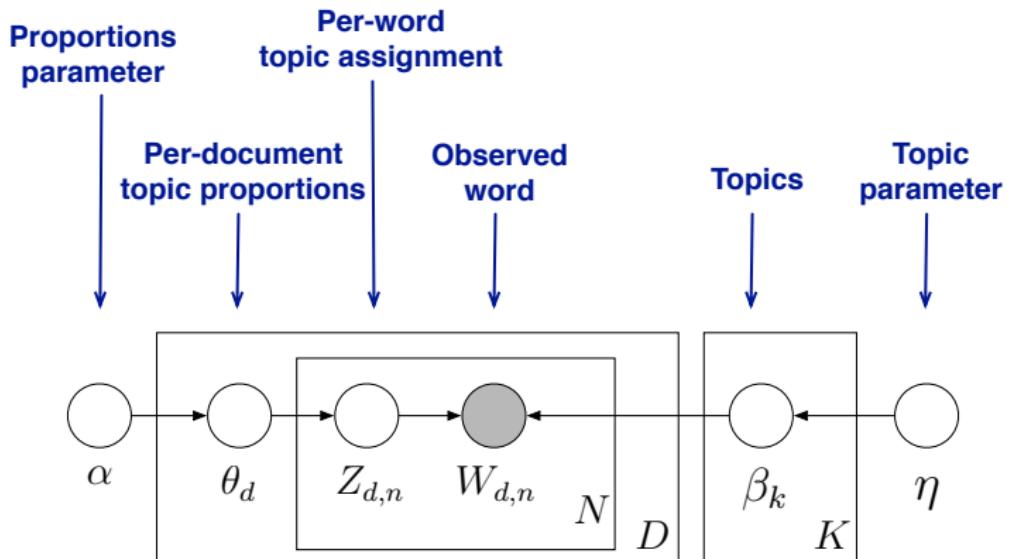
- In reality, we only observe the documents
- The other structure are **hidden variables**

The posterior distribution



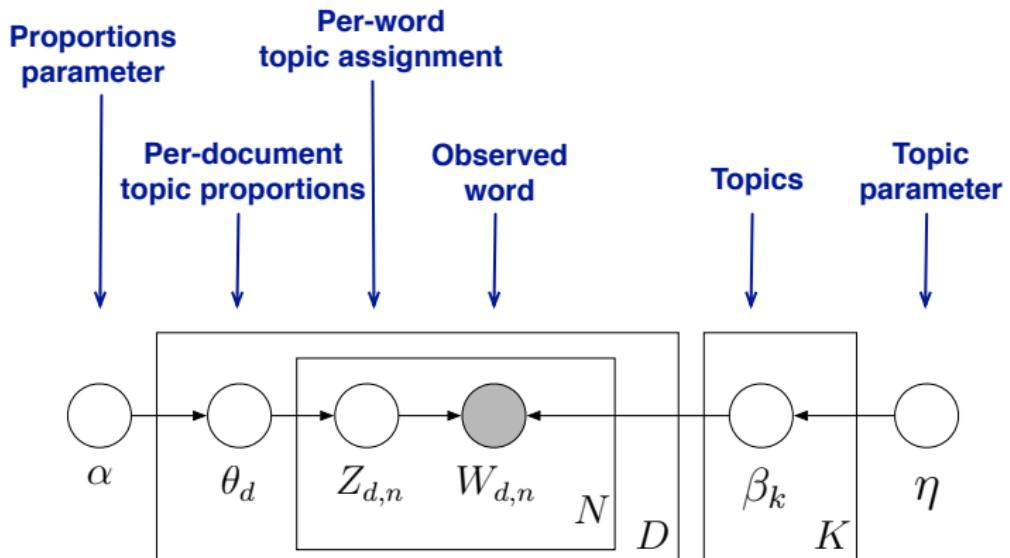
- Our goal is to **infer** the hidden variables
 - I.e., compute their distribution conditioned on the documents
- $$p(\text{topics, proportions, assignments} \mid \text{documents})$$

LDA as a graphical model



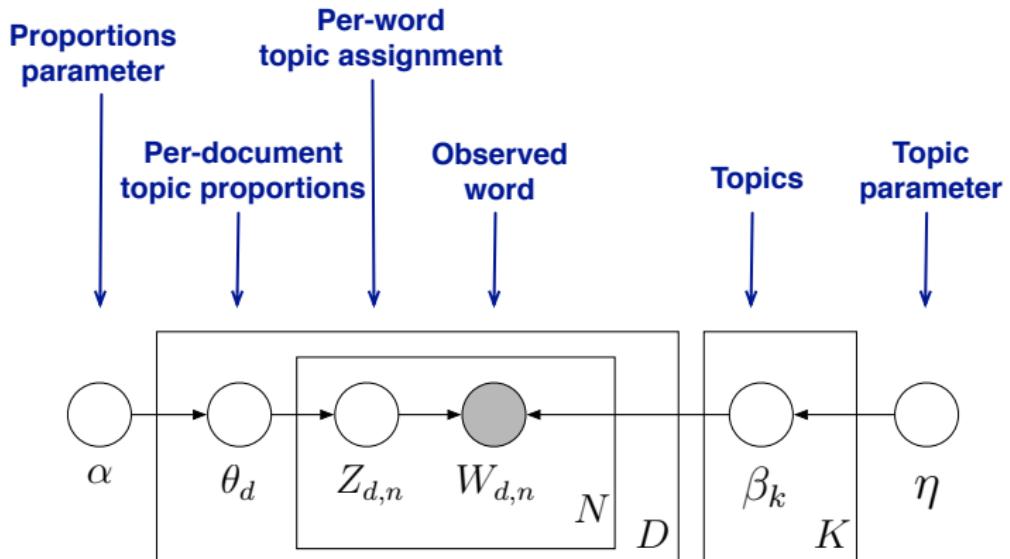
- Encodes our assumptions about the data
- Connects to algorithms for computing with data
- See *Pattern Recognition and Machine Learning* (Bishop, 2006).

LDA as a graphical model



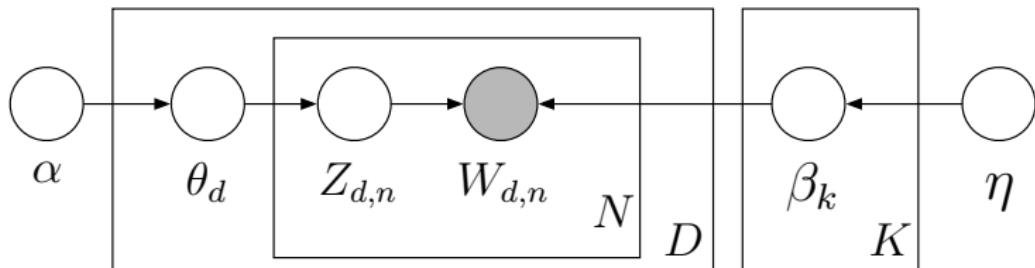
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

LDA as a graphical model



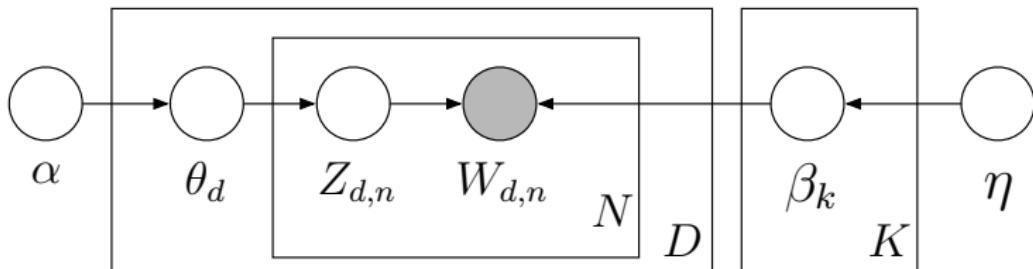
$$\prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

LDA



- This joint defines a posterior.
- From a collection of documents, infer
 - Per-word topic assignment $Z_{d,n}$
 - Per-document topic proportions 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, ...

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.

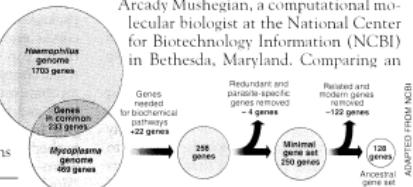
Example inference

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,⁶ 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.

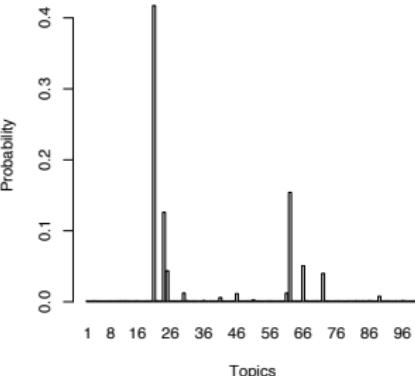
Although the numbers don't match precisely, those predictions

"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-



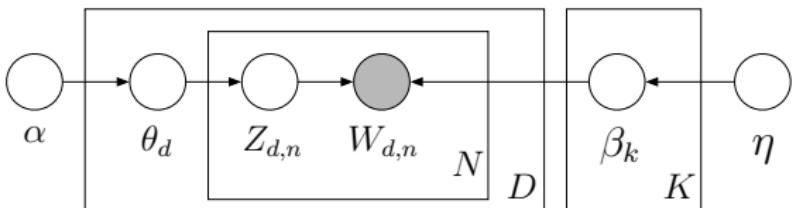
Cannibalism and chaos.
The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

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 θ_d (per document)
- And the topics β_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

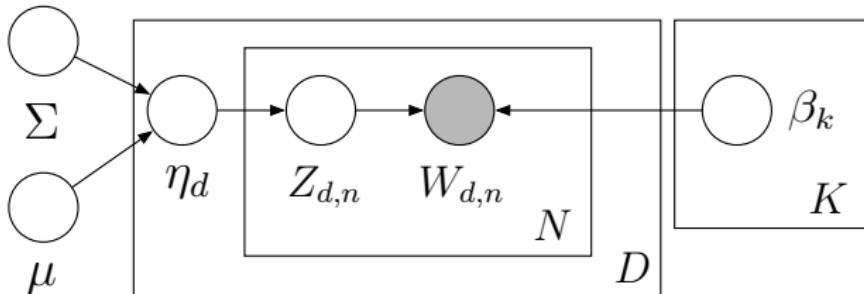
Shortcoming of the Dirichlet

- Dirichlet for topic proportions:

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

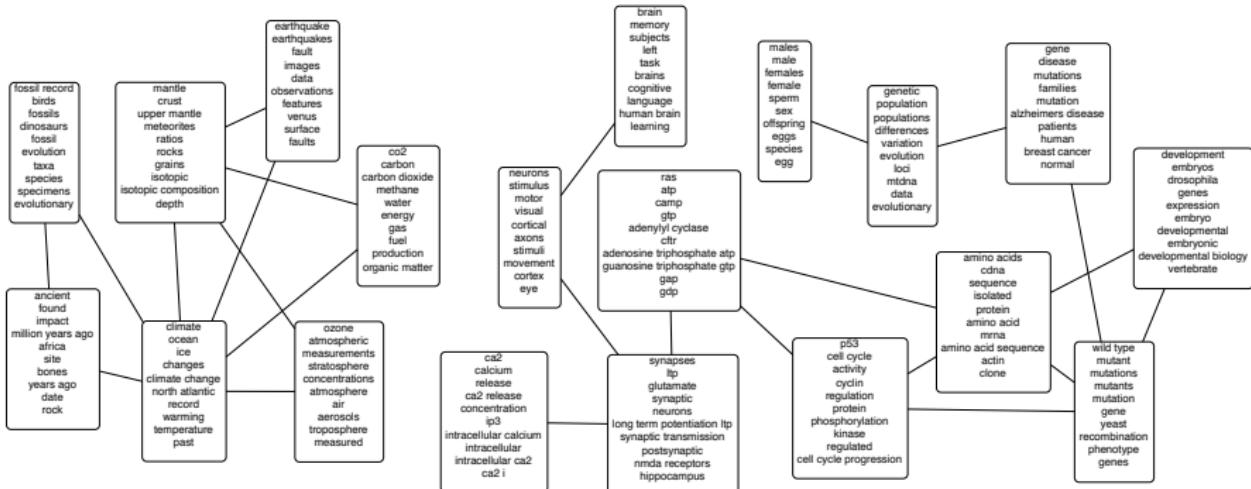
- Near independence of components makes it an unrealistic model
 - ▶ An article about *fossil fuels* is more likely to also be about *geology* than about *genetics*

Correlated topic model

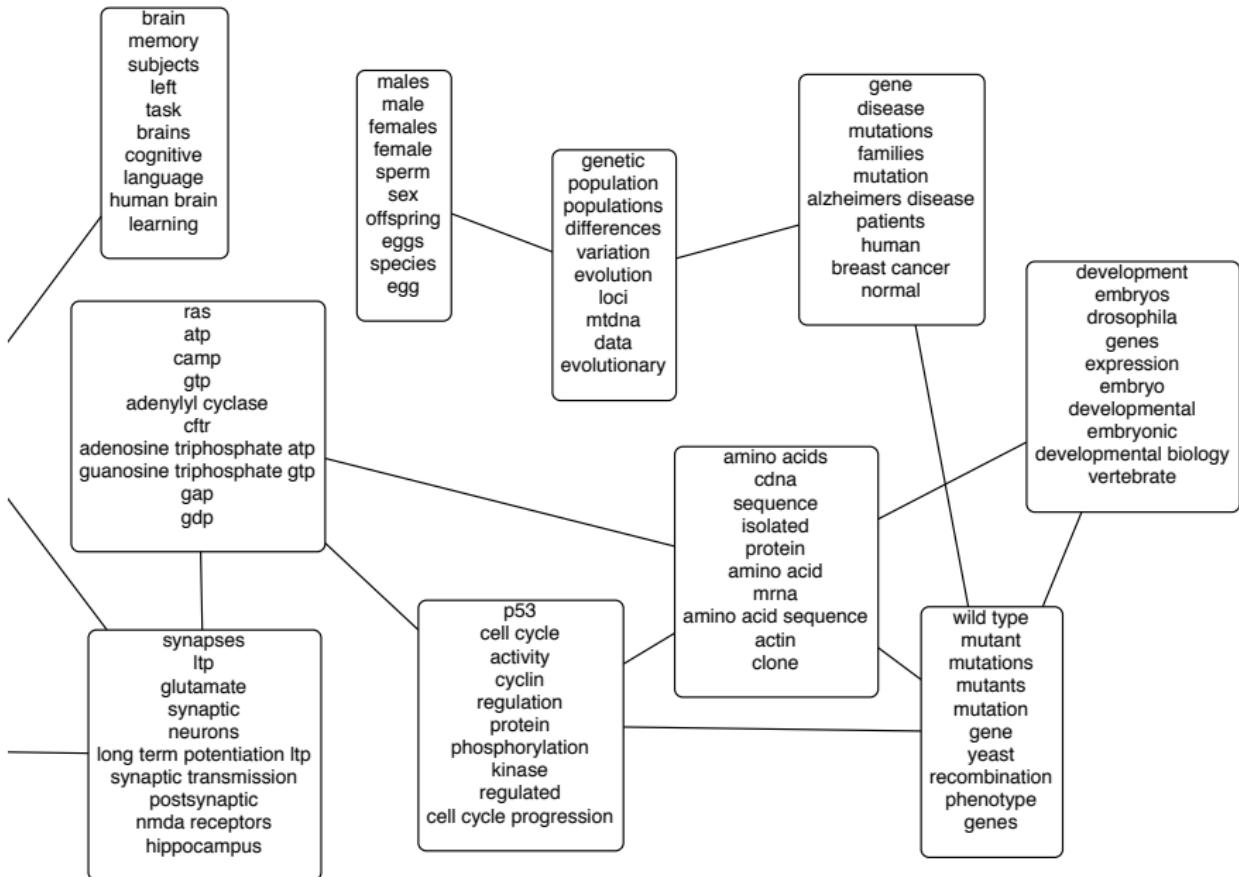


- Draw topic proportions from a logistic normal.
- Useful for:
 - ▶ providing a “map” of topics and how they are related;
 - ▶ better prediction via correlated topics.
- Sacrifice conjugacy: Posterior over θ does not have same form

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

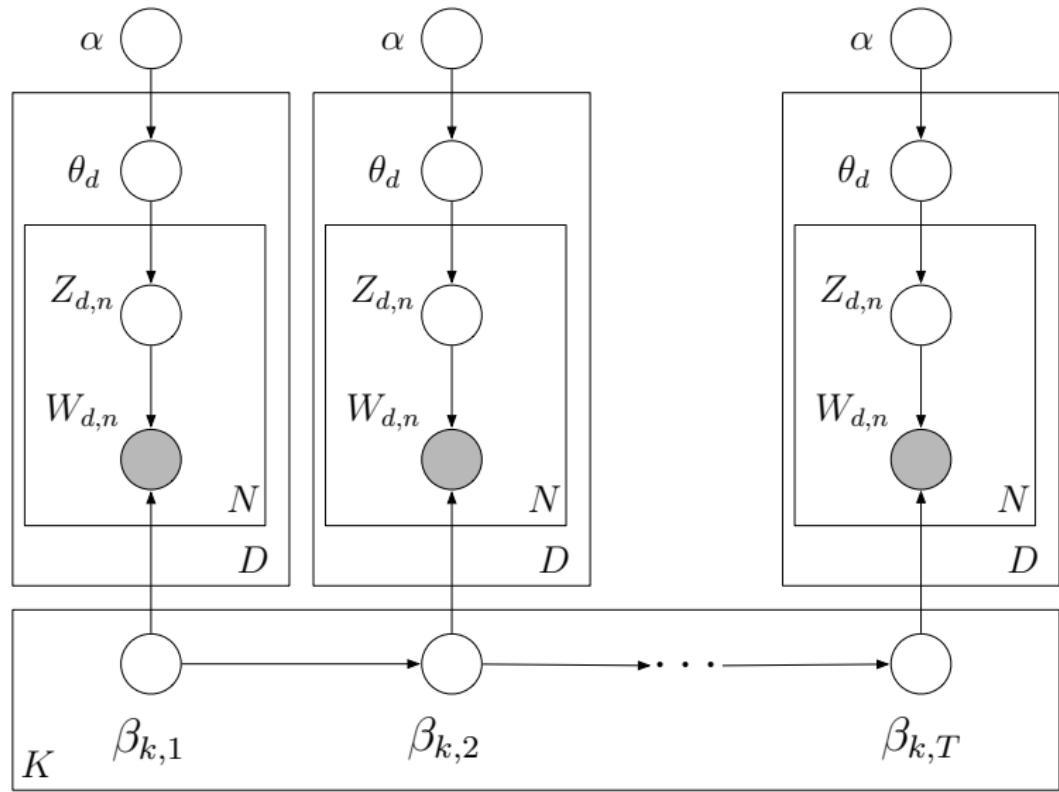
| Administration | Limnology | Astronomy | Psychology |
|----------------|-----------|--------------|------------|
| association | water | observatory | mind |
| meeting | lake | observations | nature |
| american | sea | stars | say |
| committee | waters | time | science |
| congress | lakes | made | 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)

200

EXTRACT

Truth in these cases, which were submitted to us at the end of December last for concurrence or rejection; no objection being made we printed them in a pamphlet and sent it to the author. In the pamphlet we wrote that the signs under his name are not satisfactory to him. We therefore request our readers to consider this.

George E. Pardee, Frederick D. C. Marsh and Professor J. H. Bigelow are preparing some elaborate and important papers, and promise them in 28, 29, and 30, 1857.

THE MEAN OF THE SAMPLE



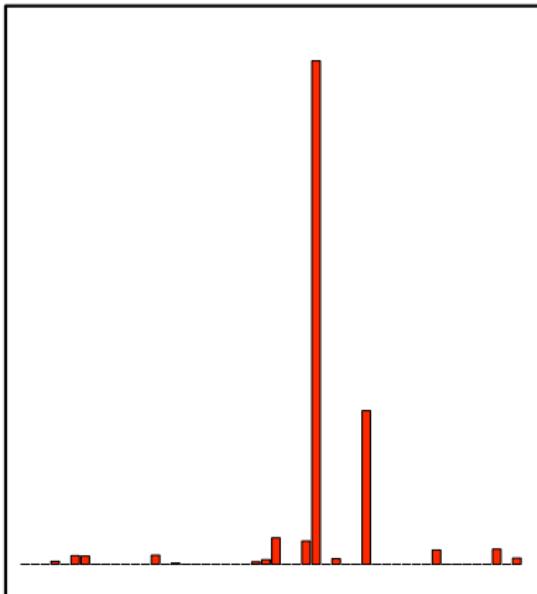
1



6

acipenser fissury, externally it is continuous with the occipital lobe, at the first acipenser synaps, anteriorly 4° separated from the posterior central concave margin conspicuously than is usual, by a fissure which runs parallel with the central fissure. There is in the Cetacean also a fissure running parallel with the parietal, which again divides the upper parietal lobule into lateral and anterior portions. The parietosyn., or the space on the mesial side of the parietal lobe between the parieto-occip-

• Florida Proceedings of the Academy of Natural Sciences, Phila., 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 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 cortex or occasionally from single resections it was possible to expose the parietal cortex. The surface of exposed parietal cortex, that is, the surface of the medial occipital cortex, was covered with a thin layer of a translucent plastic hemisphere centered in front of the contralateral eye. A circular white disk was projected onto the plane hemispherical screen with the method of Fernald and Chase (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 all experiments, this form of our most complete mapping of the visual field area, has obtained in the other four experiments, was used to map the visual field organization. Tungsten penetrations were made at an angle of approximately 45° to the surface of the exposed medial occipital cortex in a distance of approximately 1 mm from the center of the receptive field. In all experiments, we found that the receptive fields recorded adjacent to the medial area were more elongated and oriented near the lower quadrant near the vertical meridian about 30° to 40° from the center (6). Thus, as is shown in Fig. 1, and

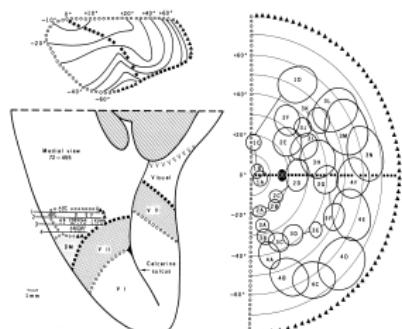
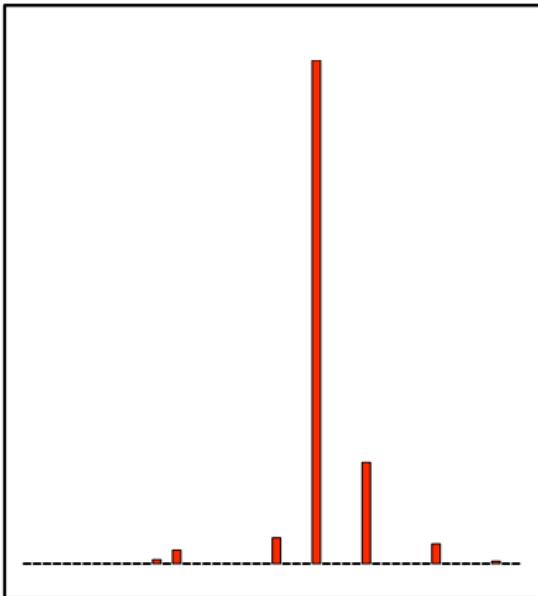


Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-415. The diagram on the lower left is a medial view of the left hemisphere of the left hemisphere with the brainstem and cerebellum removed. An arrow is used to denote the direction of depth in the diagram. Microelectrode tracks are shown in the medial wall of the medial visual area. The diagram on the right shows the upper-left areal representation of the visual field on the medial visual area. The diagram on the left shows the visual field in the visual field. The diagram shows the areal representation of the visual field on the medial visual area. PMA = dorsomedial visual area. DMS indicates the proportion of the optic disc at the visual field.

13 FEBRUARY 1985

575



Exploring the UN General Debates with Dynamic Topic Models



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



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

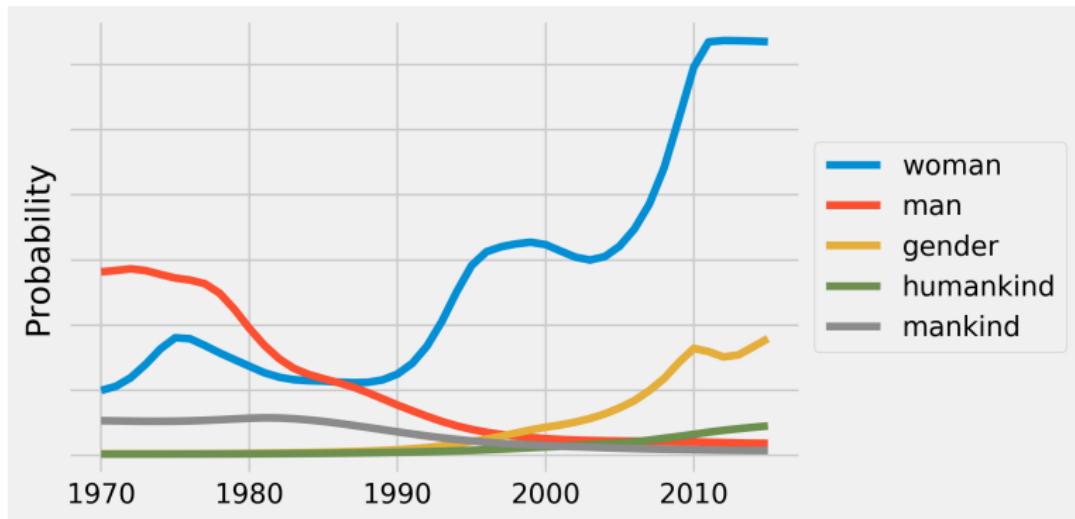
<https://towardsdatascience.com/>

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 |

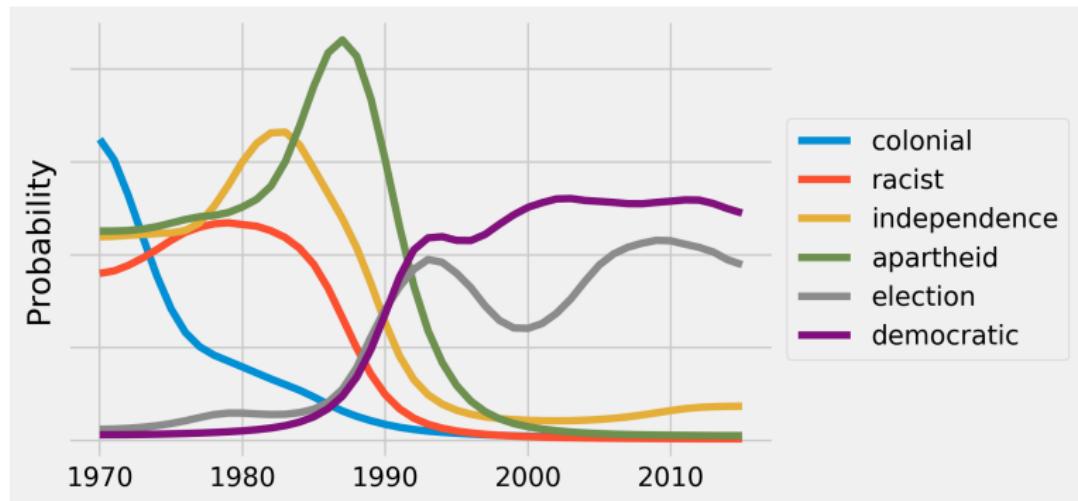
Human rights



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 |

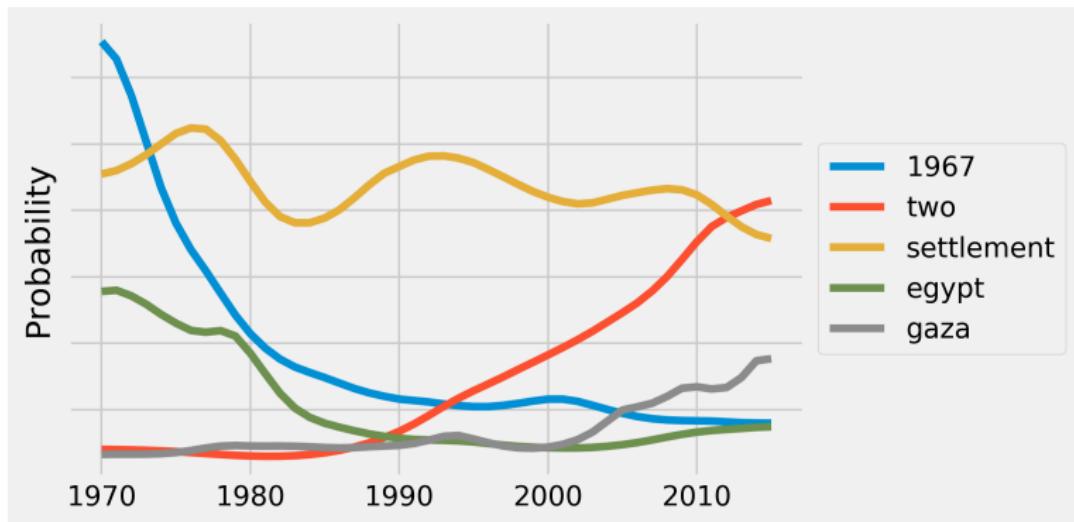
Apartheid



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 |

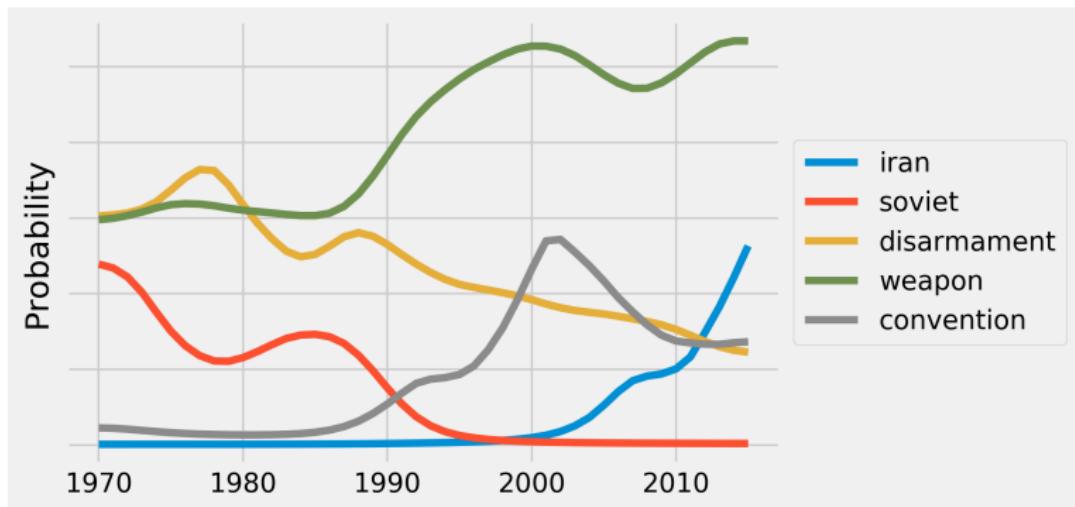
Arab-Israeli conflict



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