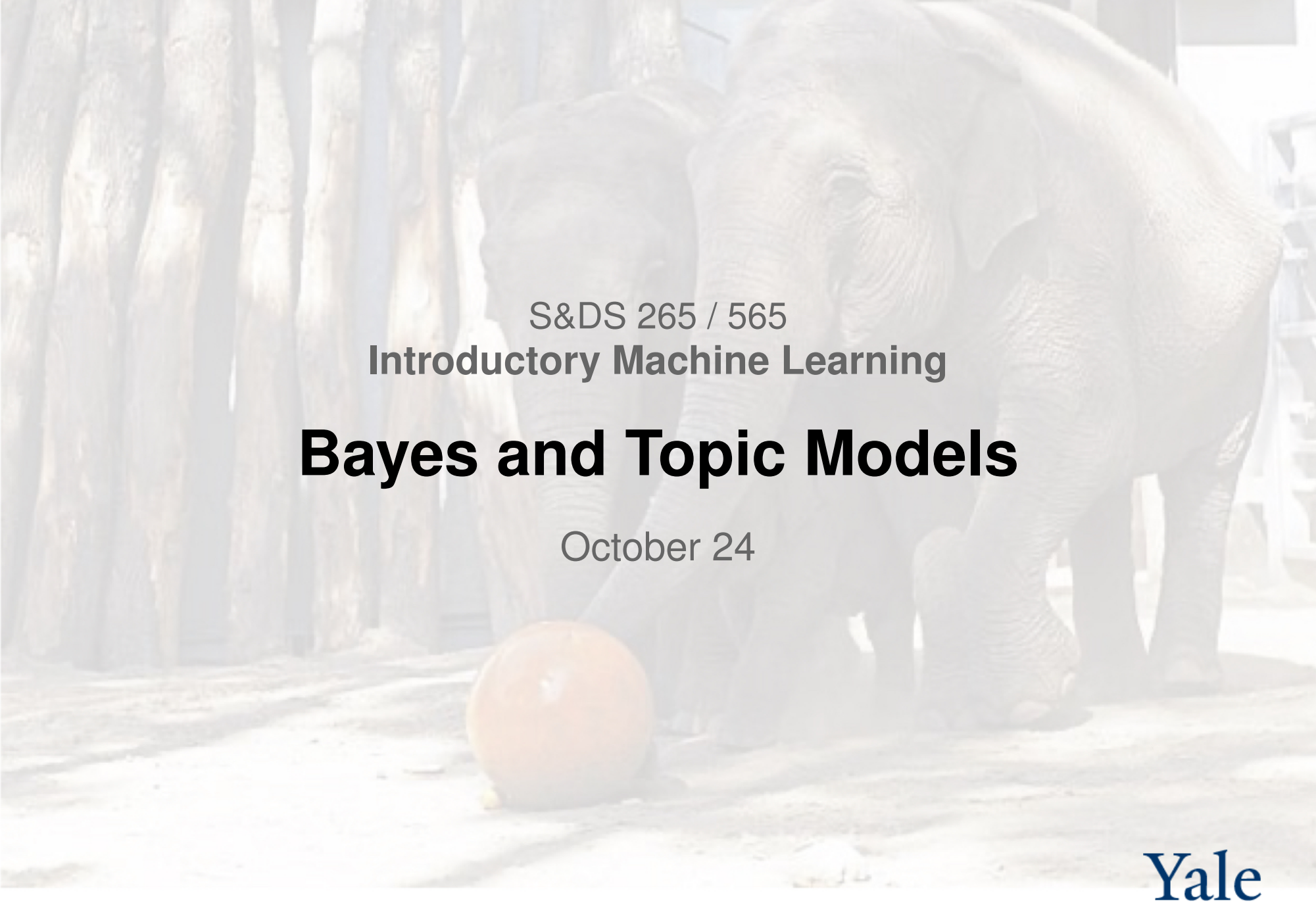


S&DS 265 / 565

Introductory Machine Learning

Bayes and Topic Models

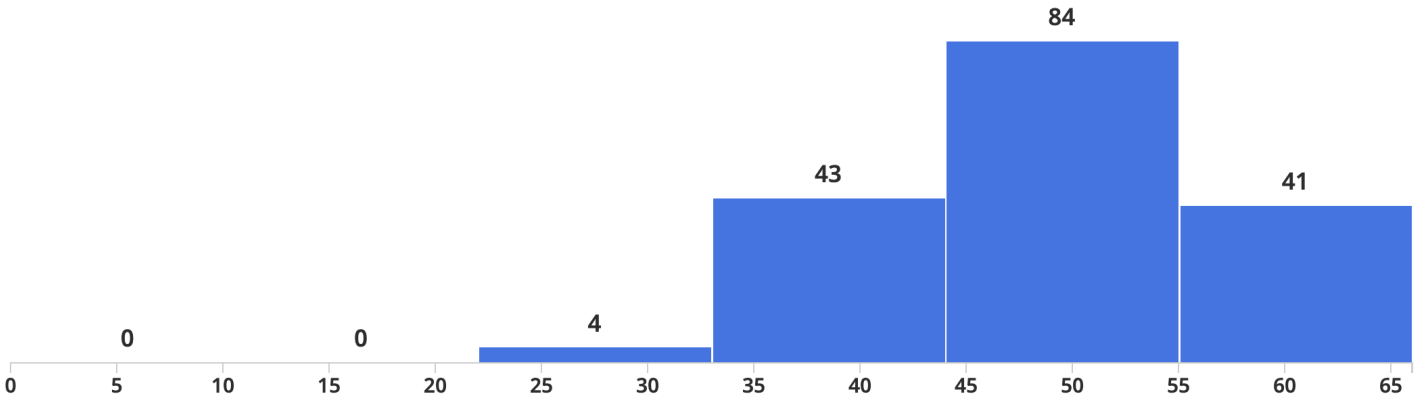
October 24

Yale

Quick notes

- Assignment 3 due tonight
- Two more to go! Assn4 is posted (language models, embeddings)
- Midterm scores and mid-semester totals posted
- Announcement this morning about mid-semester letter grades
- Quiz 4 next week

Midterm



Minimum	Median	Maximum	Mean	Std Dev
24.5	50.0	63.0	48.89	7.63

Reminder

- If you need to email me about any issues with your work for the course, please email sds265@yale.edu
- Otherwise, your email may be missed

For Today

- Mixtures
- Introduction to Bayesian analysis
- Overview of topic models

Where are we going?

- So far: Mostly “surface representations” & explicit features
  - Listing information for CA housing
  - Supernovæ classification — measured light

Where are we going?

- So far: Mostly “surface representations” & explicit features
  - Listing information for CA housing
  - Supernovæ classification — measured light
- Next: “Hidden” representations and latent variables

George Washington

1789

Among the many interesting objects which will engage your attention that of providing for the common defense will merit particular regard. To be prepared for war is one of the most effectual means of preserving peace.

Franklin Roosevelt

1941

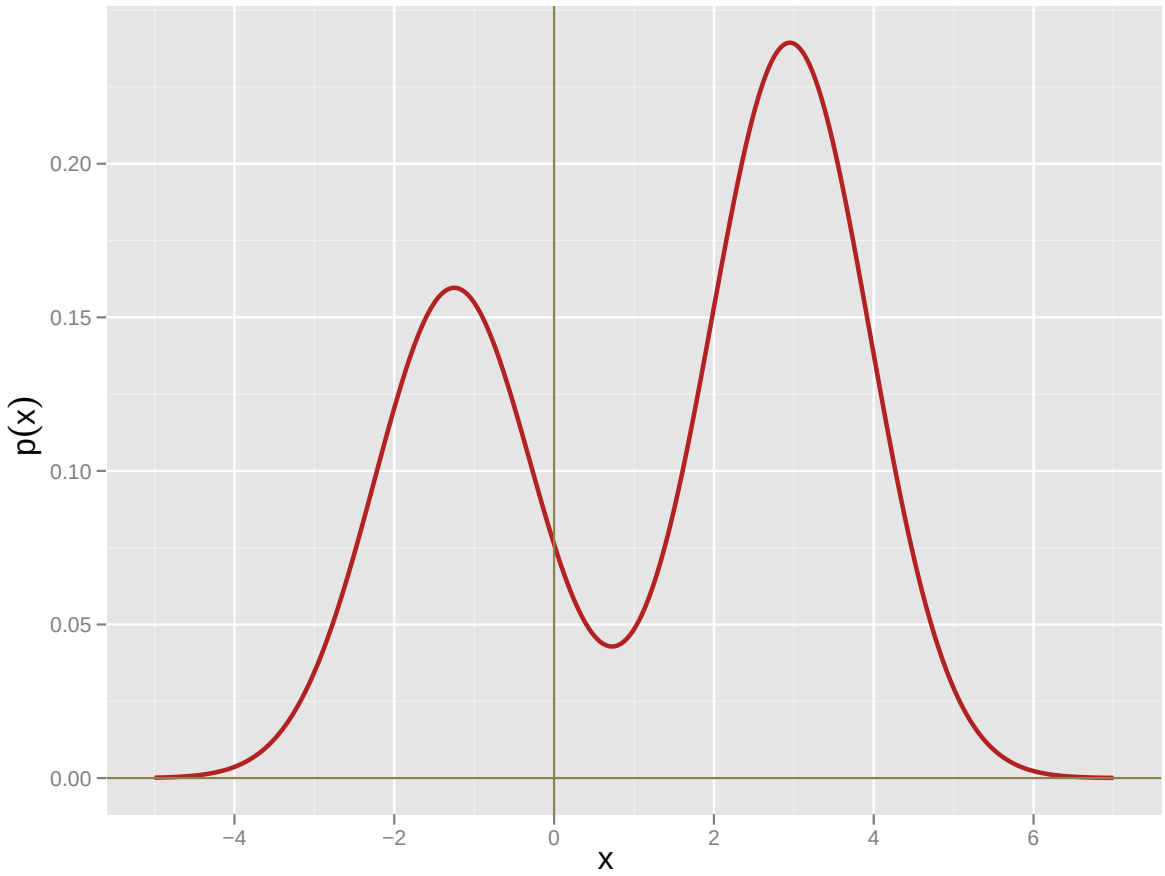
Such a peace would bring no security for us or for our neighbors. As a nation, we may take pride in the fact that we are softhearted; but we cannot afford to be soft-headed.

The elephants in the room

Mixtures

- Key technique: Mixture models
- Mixtures have latent variables
- Flexible tool
- Simple and difficult at the same time

Gaussian Mixture



Mixture of two normals, one with mean $-\frac{5}{2}$, another with mean 3.

Mixtures

- For each data point, we have a hidden (latent) variable: Which Gaussian generated the data point?
- $Z = 1$ it was generated from one model; $Z = 0$ it was generated from the other model
- We don't *observe* Z, it could be either value

Mixtures

- *Mixture of f and g*:

$$p(x) = \theta f(x) + (1 - \theta) g(x)$$

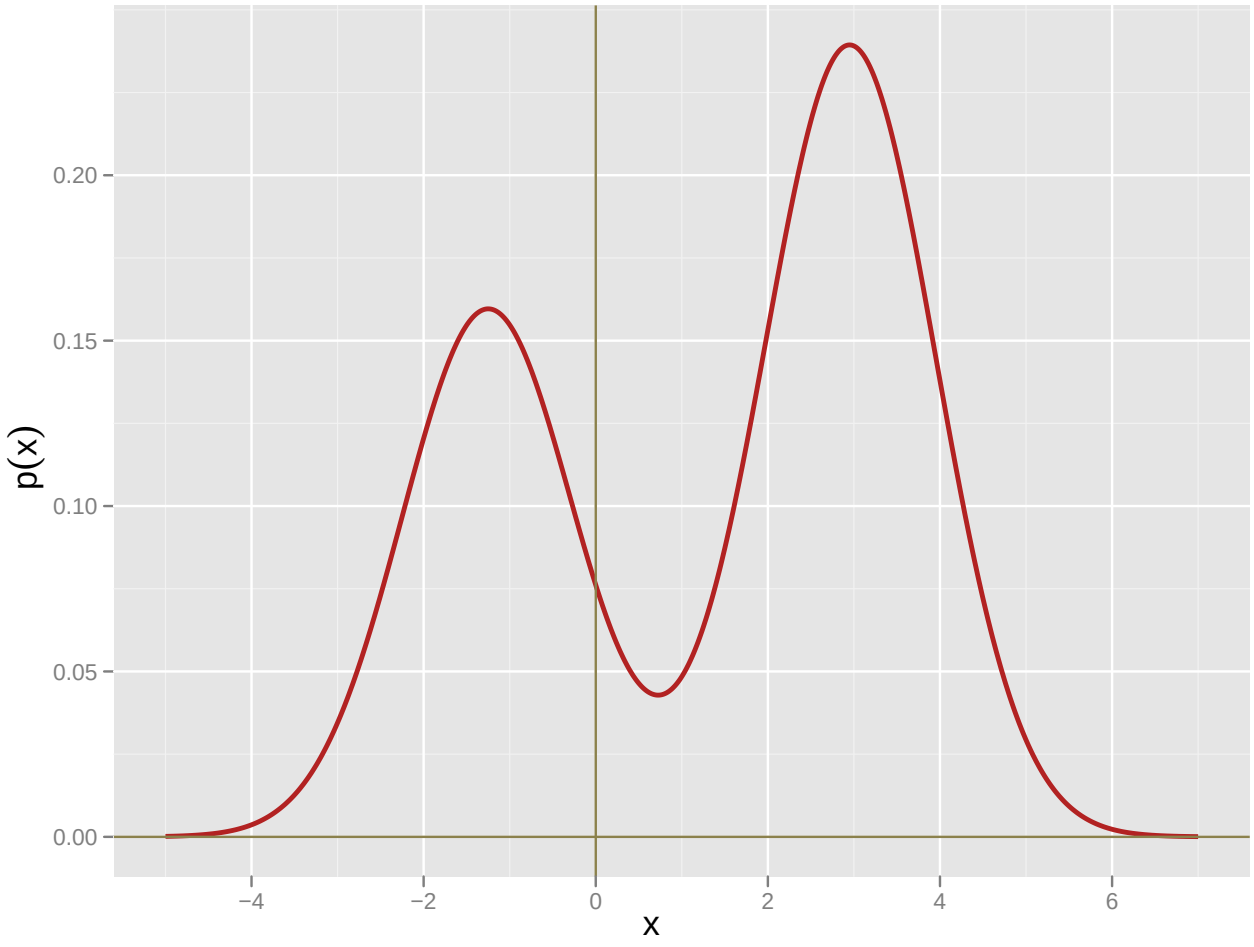
  Simplest, most common kind of latent variable model

- *Hidden variable represention*: Define $Z \sim \text{Bernoulli}(\theta)$ and

$$p(x) = \sum_{z=0,1} p(x \mid z)\, p(z)$$

  with $p(x \mid 1) = f(x)$, $p(x \mid 0) = g(x)$, $p(z) = \theta^z (1 - \theta)^{(1-z)}$.

Gaussian Mixture: All the Key Concepts



Bayesian Inference

The parameter θ of a model is viewed as a random variable. Inference usually carried out as follows:

- Choose a *generative model* $p(x \mid \theta)$ for the data.
- Choose a *prior distribution* $\pi(\theta)$ that expresses beliefs about the parameter before seeing any data.
- After observing data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, update beliefs and calculate the *posterior distribution* $p(\theta \mid \mathcal{D}_n)$.

Bayes' Theorem

A simple consequence of conditional probability:

$$\begin{aligned}\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}\end{aligned}$$

Bayes' Theorem

The posterior distribution can be written as

$$p(\theta \mid x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{p(x_1, \ldots, x_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta)$$

where $\mathcal{L}_n(\theta)$ is the *likelihood function* and

$$c_n = p(x_1, \ldots, x_n) = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$$

is the normalizing constant, which is also called *evidence*.

Important Example

Take model $X \sim \text{Bernoulli}(\theta)$.

This is a "coin flip": $X = 1$ means "heads" and $X = 0$ means "tails."

Natural prior is Beta(α, β) distribution

$$\pi_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Important Example

Take model $X \sim \text{Bernoulli}(\theta)$.

Natural prior is Beta(α, β) distribution

$$\pi_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The scaling constant is scary looking:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

where $\Gamma(\cdot)$ is the "Gamma function"

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

---

https://en.wikipedia.org/wiki/Gamma_function

Important Example

$X \sim$ Bernoulli(θ) with data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$. Prior Beta($\alpha, \beta$) distribution

$$\pi_{\alpha,\beta}(\theta) \;\propto\; \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Let $s = \sum_{i=1}^{n} x_i$ be the number of "heads"

Posterior distribution $\theta \,|\, \mathcal{D}_n$ is another beta distribution!

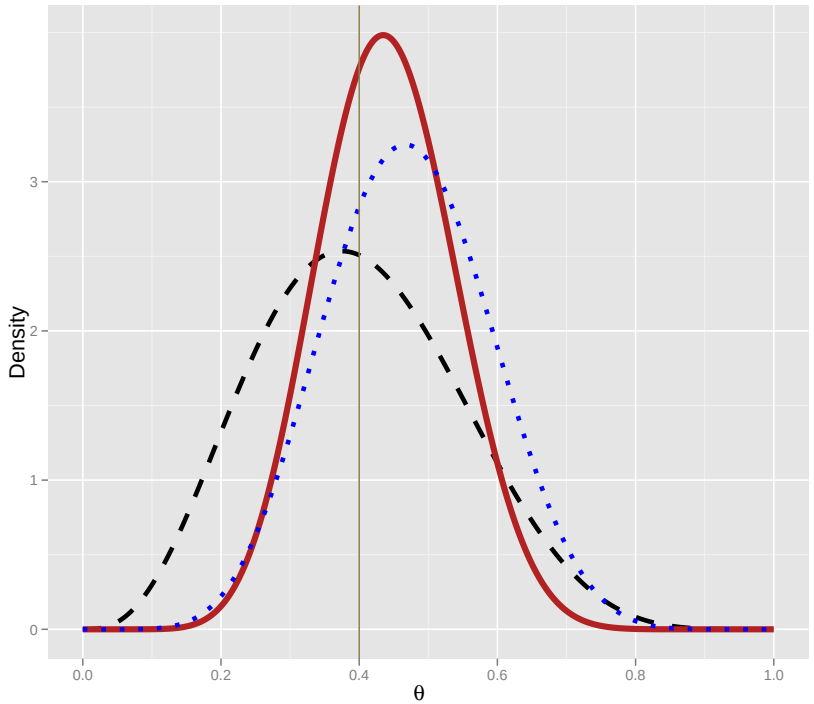
Specifically, with

$$\widetilde{\alpha} = \alpha + \text{number of heads} = \alpha + s$$
$$\widetilde{\beta} = \beta + \text{number of tails} = \beta + n - s$$

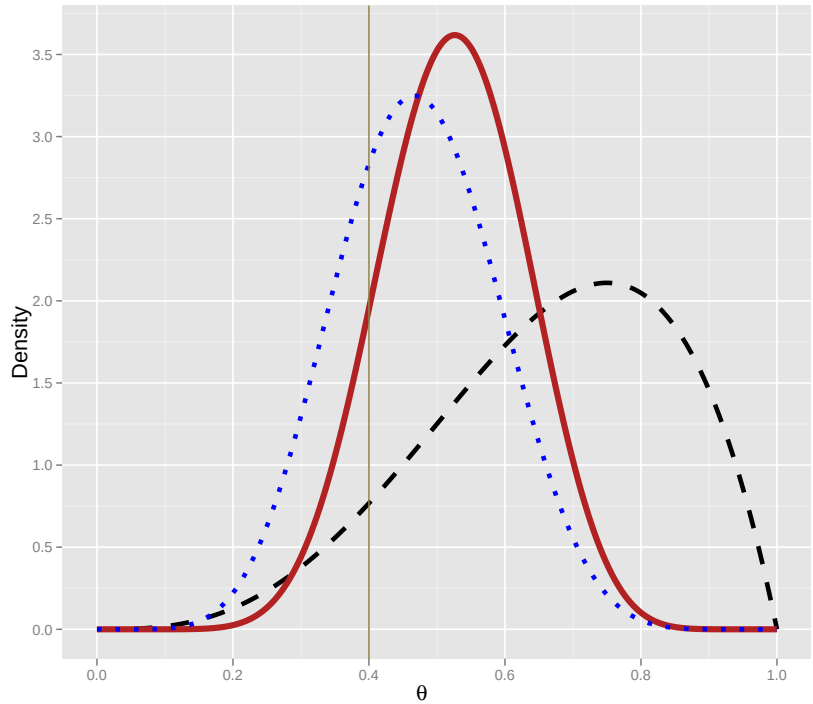
Showing this just uses the simple fact that $\theta^{\alpha-1}\theta^{x} = \theta^{x+\alpha-1}$

Example

$n = 15$ points sampled as $X \sim \text{Bernoulli}(\theta = 0.4)$, with $s = 7$ heads.



prior A



prior B

Prior distribution (black-dashed), likelihood function (blue-dotted), posterior distribution (red-solid).

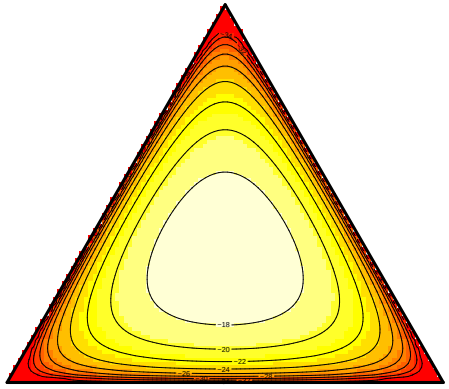
Dirichlet

Multinomial model with Dirichlet prior is generalization of the Bernoulli/Beta model.

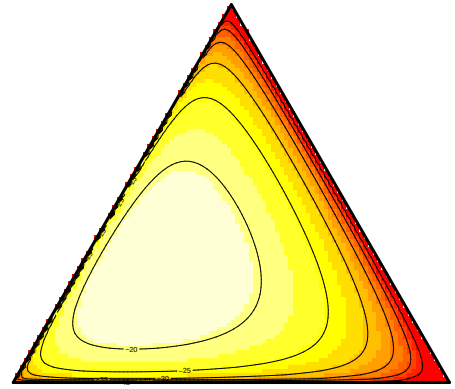
$$\text{Dirichlet}_\alpha(\theta) \;\propto\; \theta_1^{\alpha_1 - 1}\, \theta_2^{\alpha_2 - 1} \;\cdots \theta_K^{\alpha_K - 1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$ is a non-negative vector.

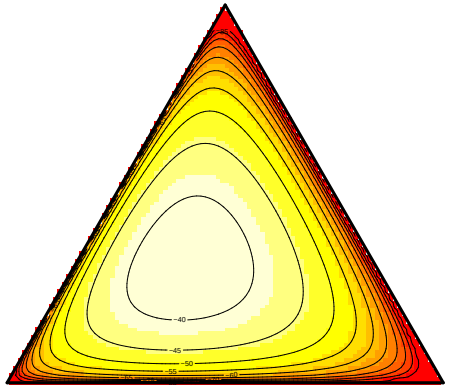
Example



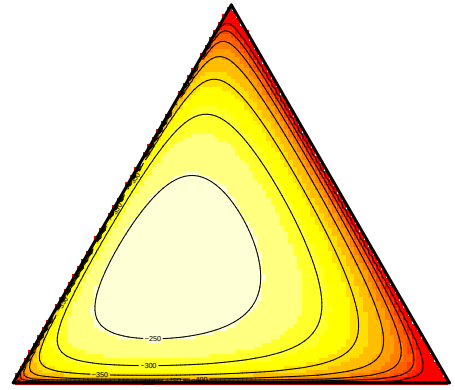
prior with Dirichlet(6,6,6)



likelihood function with $n = 20$



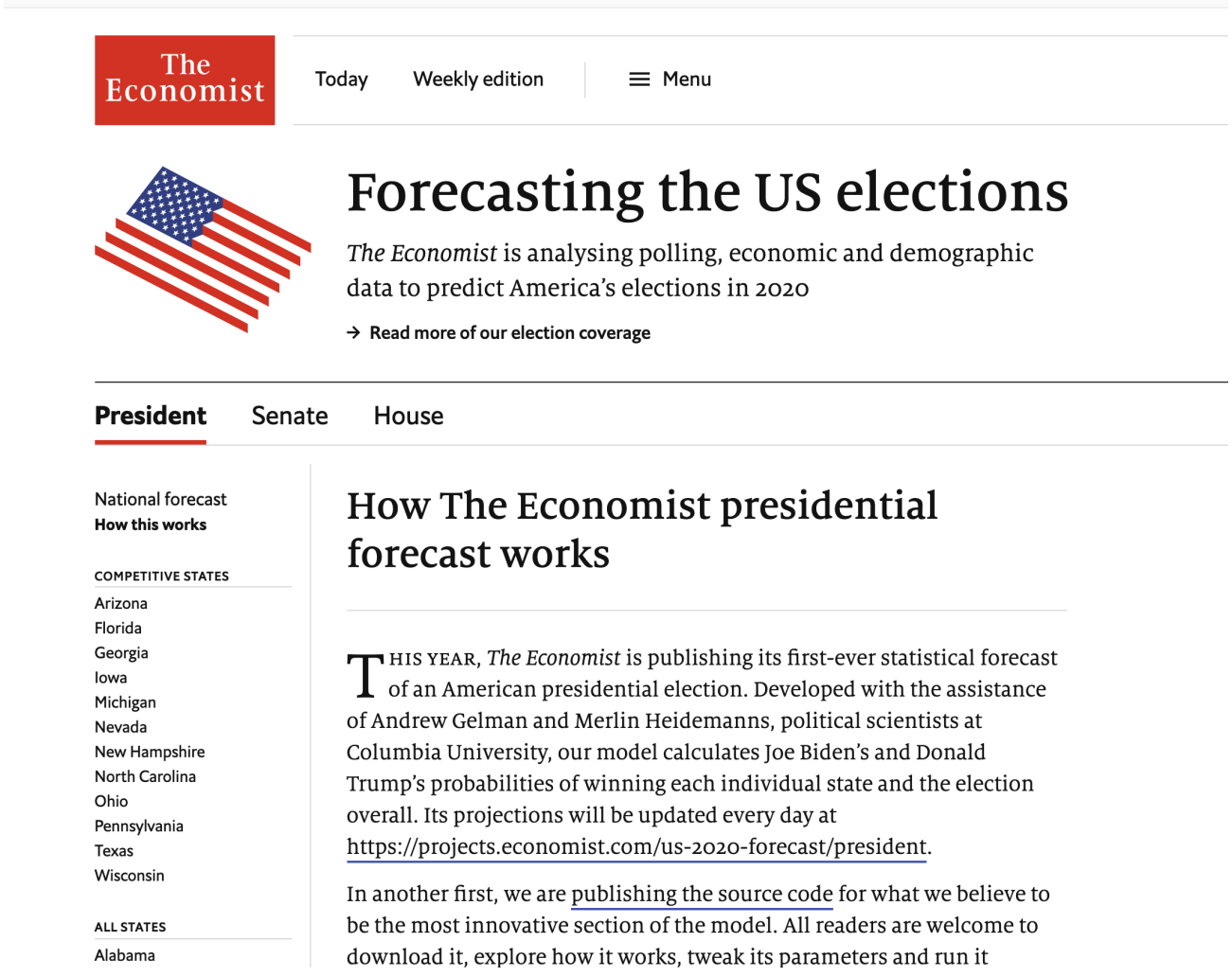
posterior distribution with $n = 20$



posterior distribution with $n = 200$

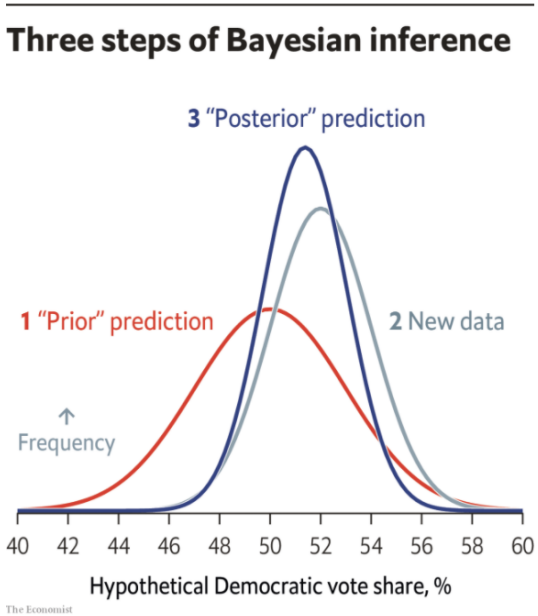
Example: Election Forecasting

https://projects.economist.com/us-2020-forecast/president/how-this-works



Example: Election Forecasting

https://projects.economist.com/us-2020-forecast/president/how-this-works



Back to Bayes-ics

Readers acquainted with the workings of similar forecasting models may be surprised that the phrase "state polls" has not yet entered the equation. This exclusion is by design. Our model follows a logical structure first developed by Thomas Bayes, an 18th-century reverend whose ideas have shaped a large and growing family of statistical techniques. His approach works in two stages. First, before conducting a study, researchers explicitly state what they believe to be true, and how confident they are in that belief. This is called a "prior". Next, after acquiring data, they update this prior to reflect the new information—gaining more confidence if it confirms the prior, and generally becoming more uncertain if it refutes the prior (though not if the new numbers are so definitive that leave little room for doubt). In this framework, the expected distribution of potential vote shares in each state derived above is the prior, and state polls that trickle in during the course of the campaign are the new data. The result—a "posterior", in Bayesian lingo—is our forecast.

Let’s go to the notebook!

Demo code for Bayesian analysis

In this notebook we illustrate some of the basic models and priors for Bayesian inference. These concepts will be important for our discussions about "topic models."

First, we illustrate the situation where the parameter θ that we are modeling is a Bernoulli parameter. This can be thought of as the probability that flipping a certain coin comes up heads. The most commonly used prior distribution for this model is a beta distribution. Under a beta prior, the posterior distribution is again a beta distribution.

This is illustrated in the following simula ior becomes concentrated on the true parameter. But as the variance of t

In [5]:
```
import os, gzip
import numpy as np
import matplotlib.pyplot as p
```

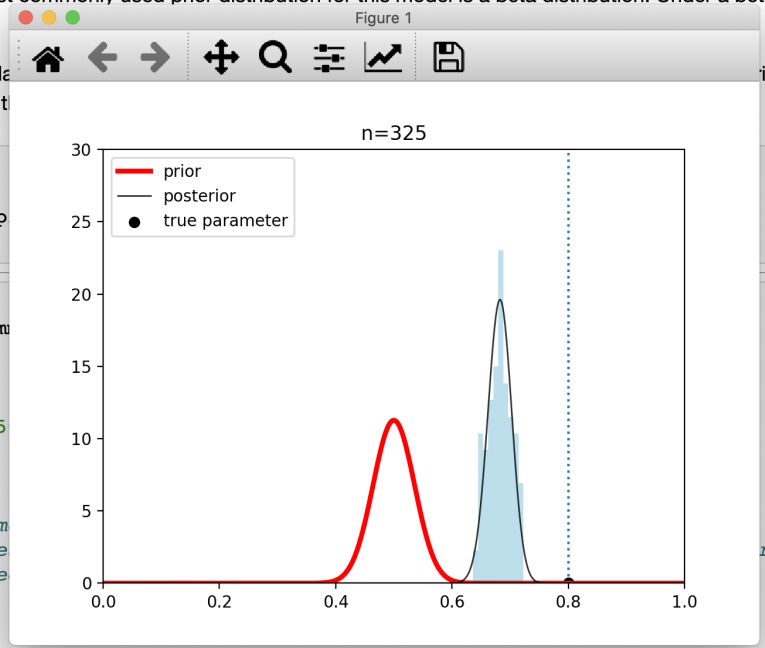
In [*]:
```
%matplotlib qt
from scipy.special import gam
from scipy import random
from scipy.stats import beta

theta = np.linspace(0,1,num=5
fig = plt.figure(1)
plt.ion()

# The following are the param                    the
# variance of the prior decre                    required for
# the posterior to be centere

scale = 100
a0 = scale*1
b0 = scale*1

sample_size = 100
```



Reading

- See notes on Bayesian inference on iML site (`bayes-notes.pdf`)
- Not necessary to understand everything—but you should be able to do some of the basic “coin flipping” calculations
- We’ll build on this when discussing topic models

Summary

- Mixtures are latent variable models
- The mixing weight encodes a hidden variable
- Computing with mixtures uses basic probabilistic reasoning
- Bayesian inference is popular in ML
- In a Bayesian approach, the parameters are random, and the data are fixed.
- Three ingredients: Prior, likelihood, posterior

Next up: Topic models

- High level intro to topic models
- Use of latent variables, mixtures
- More details and examples next week

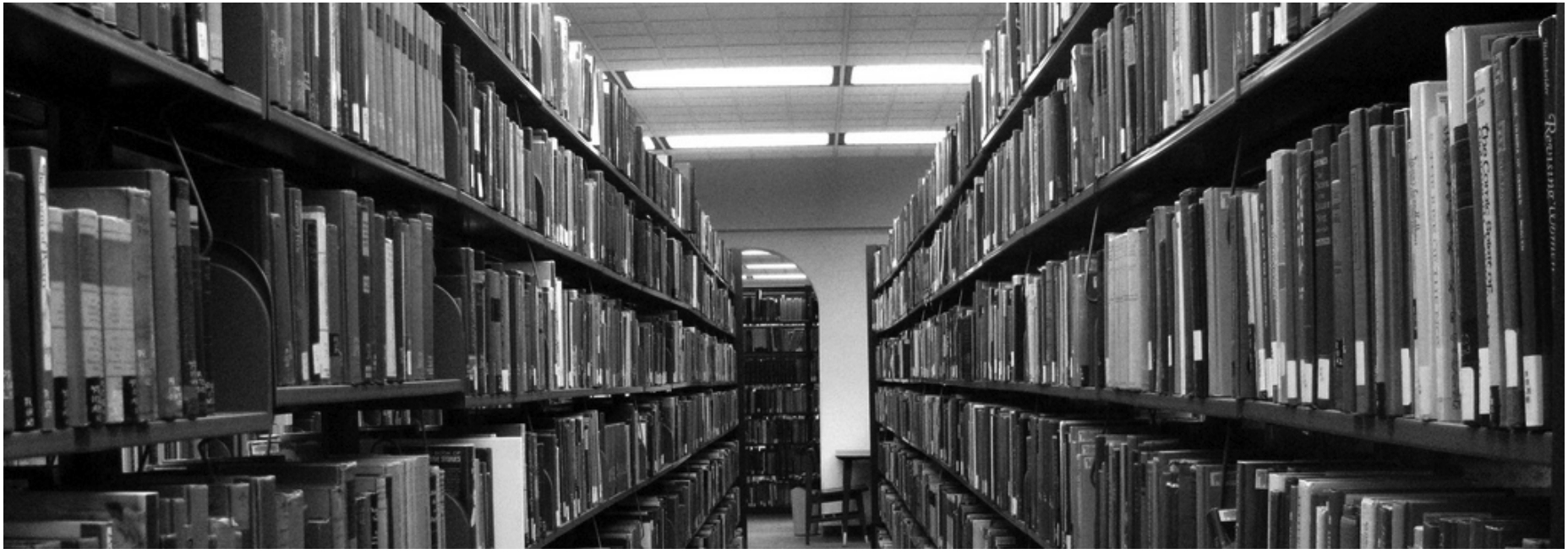
Readings:

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

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

`https://www.cs.columbia.edu/~blei/papers/Blei2012.pdf`

Topic modeling



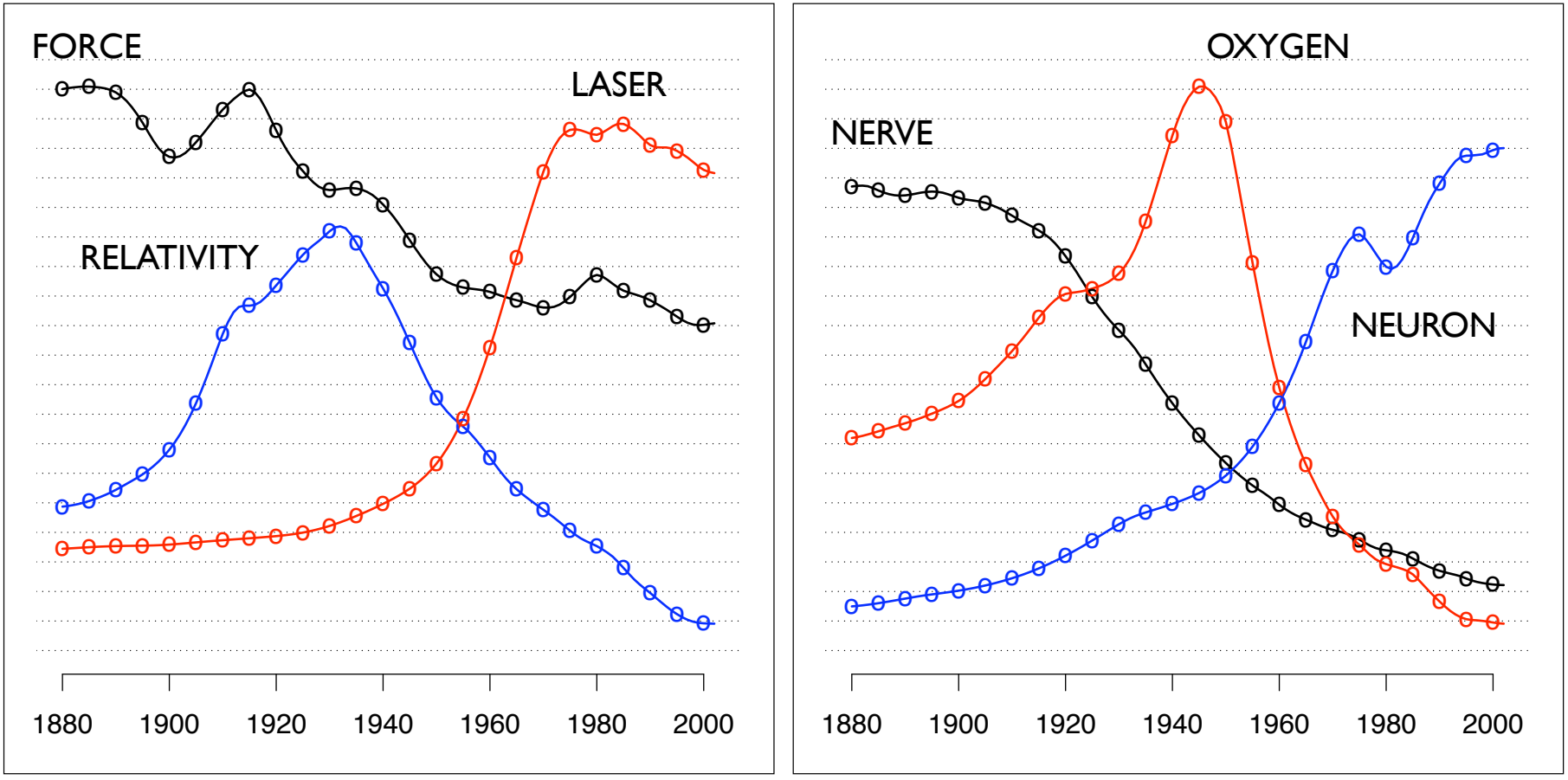
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

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.

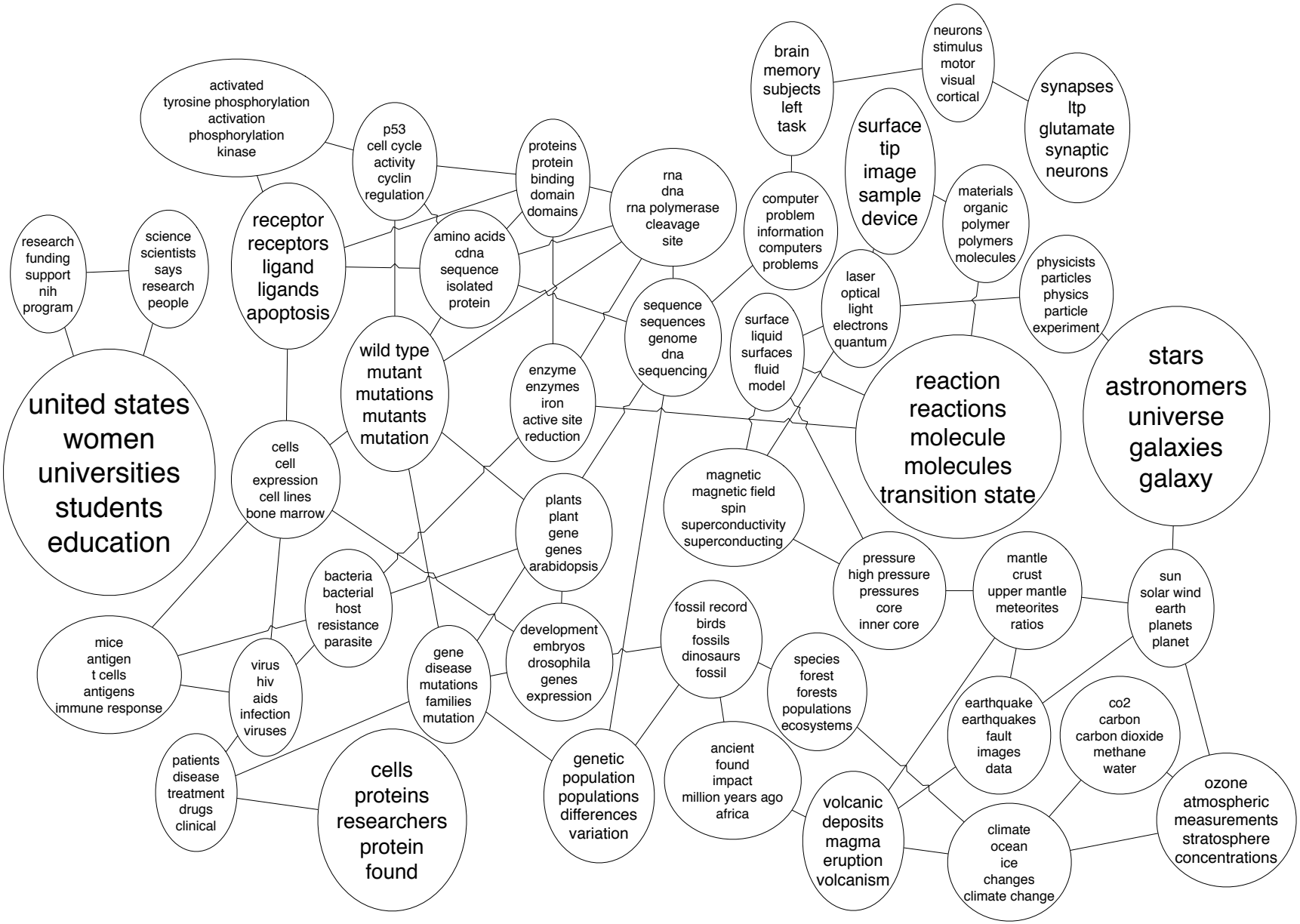
Discover topics from a corpus

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

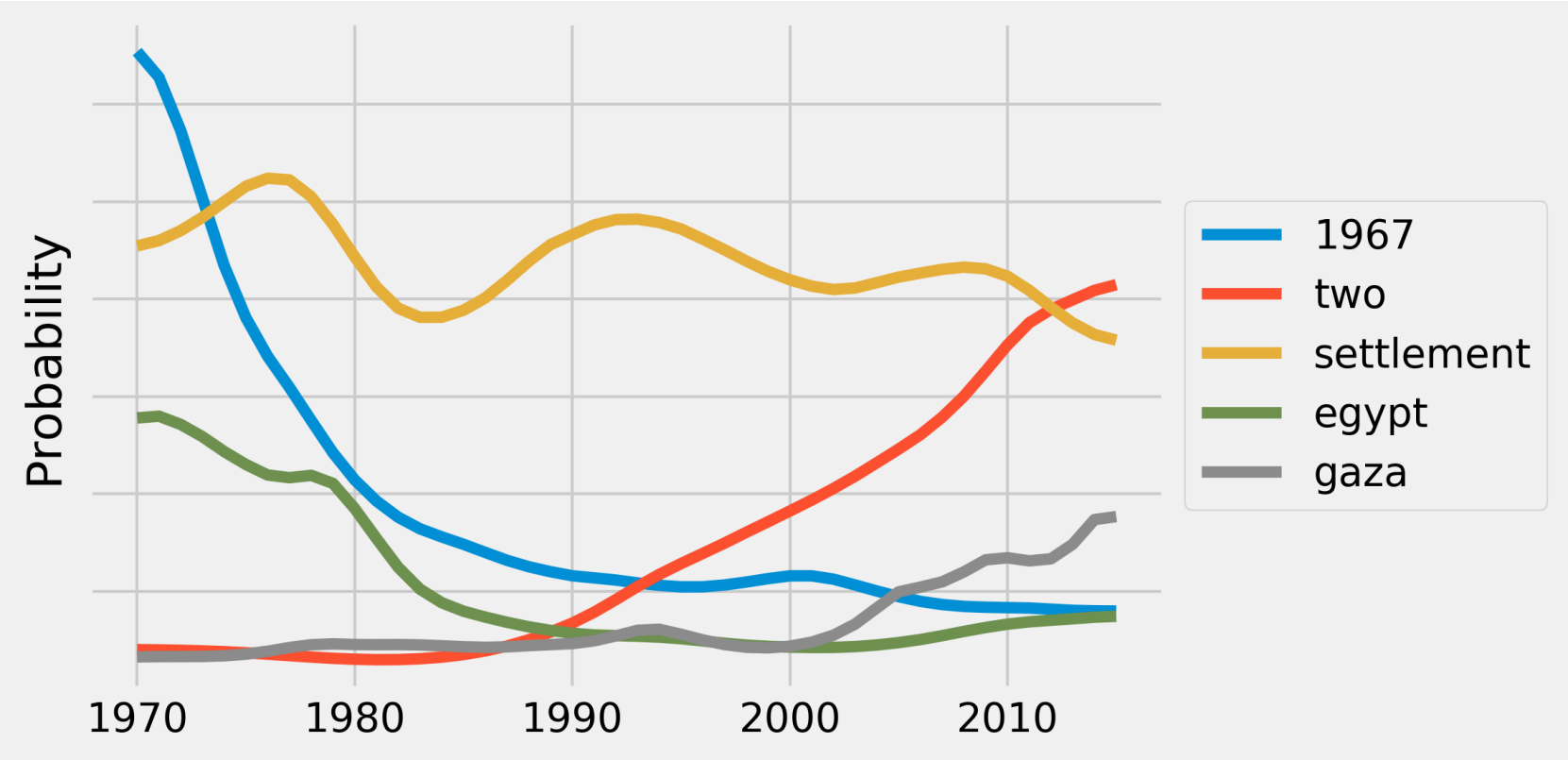
Model the evolution of topics over time



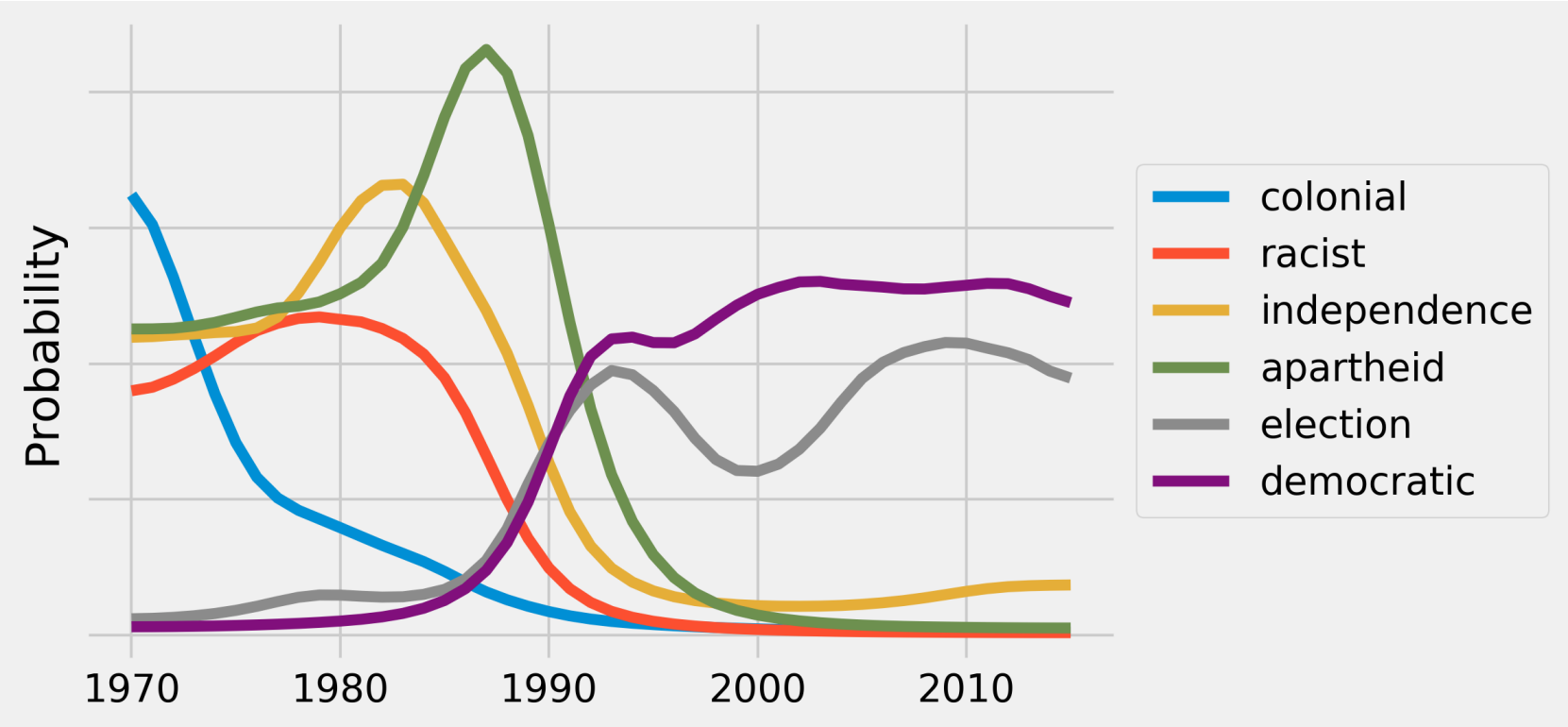
Model connections between topics



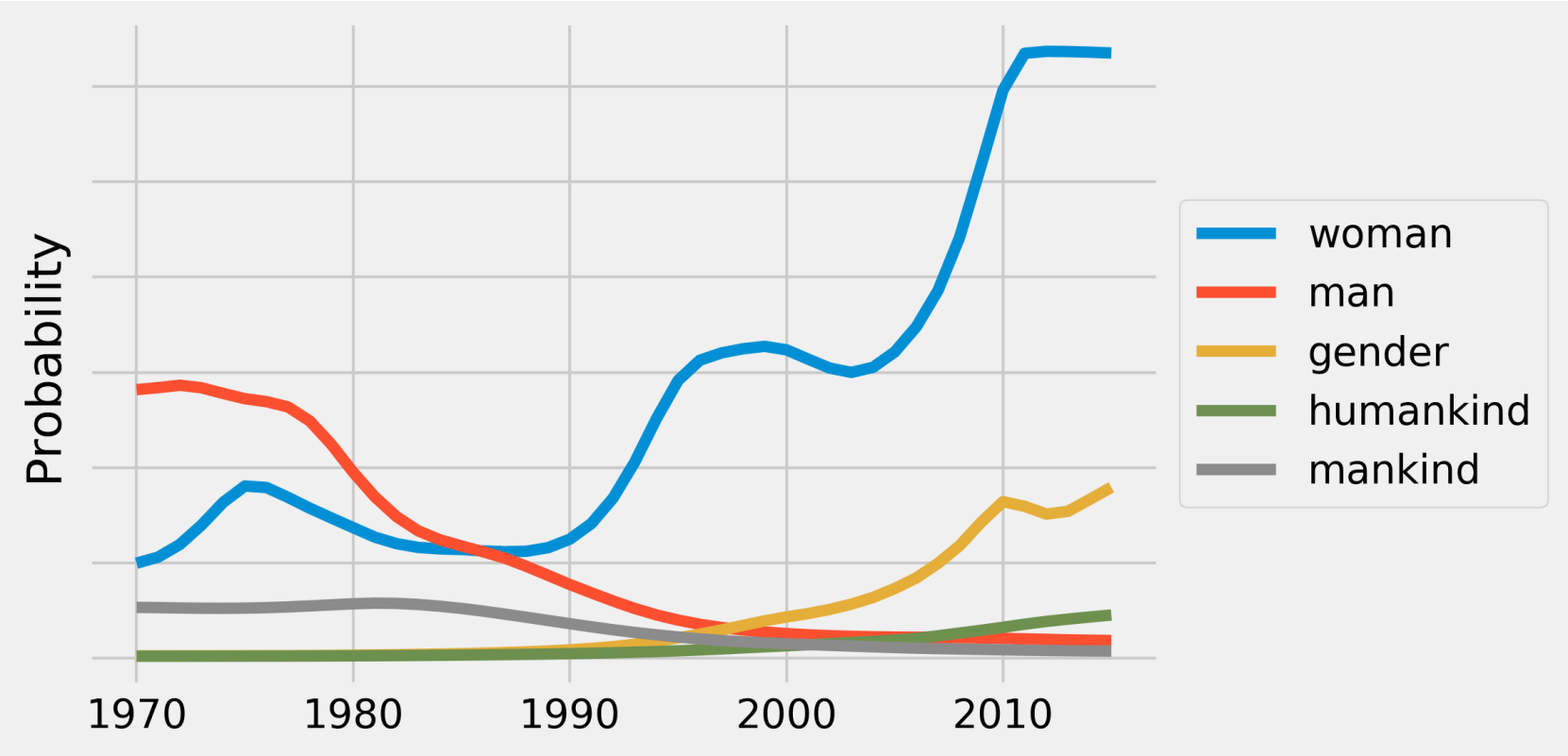
Proceedings of the United Nations



Proceedings of the United Nations



Proceedings of the United Nations



Topics in scientific texts

Quantum physics	spin energy field electron magnetic state states hamiltonian
Particle physics	higgs neutrino coupling decay scale masses mixing quark
Astrophysics	mass gas star stellar galaxies disk halo radius luminosity
Relativity	black metric hole schwarzschild gravity holes einstein
Number theory	prime integer numbers conjecture integers degree modulo
Graph theory	graph vertex vertices edges node edge number set tree
Linear algebra	matrix matrices vector basis vectors diagonal rank linear
Optimization	problem optimization algorithm function solution gradient
Probability	random probability distribution process measure time
Machine learning	layer word image feature sentence model cnn lstm training

In all of these examples, the topics are automatically inferred (unsupervised) through Bayesian inference

Topic modeling for equations

Topic	Generated Equations
Quantum physics	• $E = \hbar \frac{\partial^2 S}{\partial t^2} (\frac{\partial \varphi}{\partial c}) - \frac{k}{\hbar^2} \frac{\partial B}{\partial t} (t + \partial_t \delta)$. • $\Psi_{\text{pr}} = \sum_{\mathbf{l}} (\psi_{\mathbf{r}+\uparrow} - \psi^{\dagger}_{\mathbf{r}\downarrow}) + \sum_{\mathbf{r}'} (\psi^{\dagger}_{\mathbf{r}\downarrow,\uparrow} - \psi_{\mathbf{r}\downarrow} \sigma^{\dagger})$.
Particle physics	• $\mathcal{H} = \frac{2}{4} (\partial_\mu \phi)^2 + 2m\phi_\nu(\phi) + \frac{1}{2} m^2 (\phi)(1 - \phi^2)^2$. • $m_{\text{eff}}(M) = 1.4 \cdot 10^{-13}\,\text{GeV}$.
Relativity	• $\mathcal{M} = \frac{2}{2} g^{\mu\nu} (f_{\mu\nu,\mu} - g_{\mu\nu,\nu} + g_{\nu\nu,b} f_{\mu,\nu}) + \frac{1}{2} g^{\mu\nu}$. • $T_{\mu\nu} = \int_0^\infty ds_{\mu\nu} ds^2 + a_\mu^2 dr^2 + r^2 d\Omega^2$.

https://seas.yale.edu/news-events/news/michihiro-yasunaga-wins-cra-outstanding-undergraduate-researcher-award

Uber Topics

From a former student:

> *Hi Prof. Lafferty,*
>
> *I took your course last spring. I just wanted to let you know that I'm currently topic modeling at my current job at Uber!*
>
> *We're using topic models at Uber to discover topics in rider feedback – when riders write comments about their driver after the trip. We're trying to find topics such as 'unprofessional driver', 'driver no-show', 'sexual harassment', etc.*

Probabilistic modeling

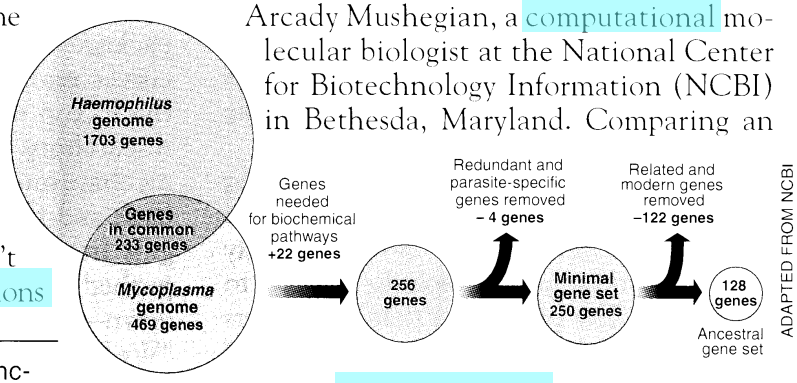
1. Data are assumed to be observed from a generative probabilistic process that includes hidden variables.
   - *In text, the hidden variables are the thematic structure.*
2. Infer the hidden structure using posterior inference
   - *What are the topics that describe this collection?*
3. Situate new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



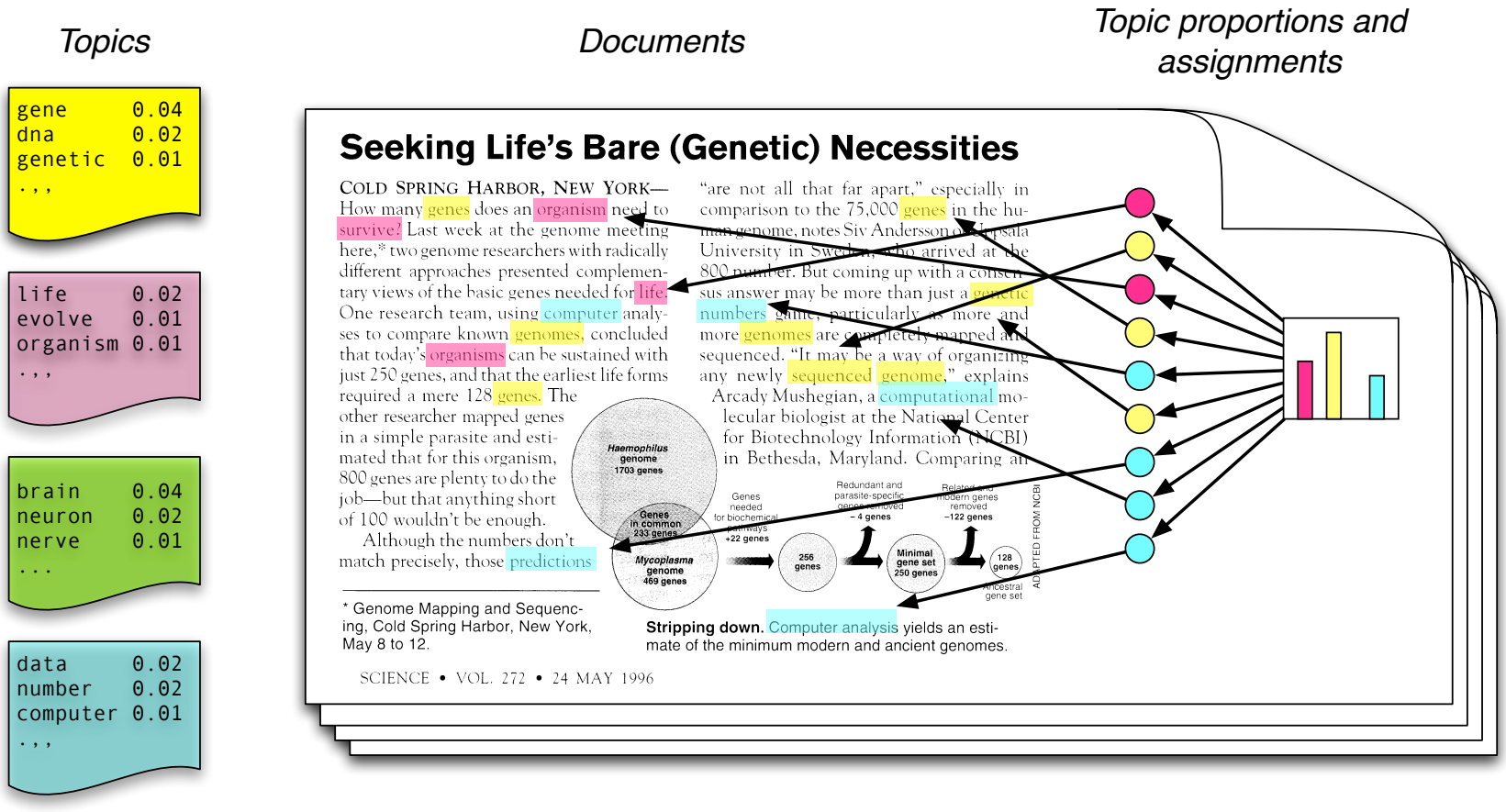
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

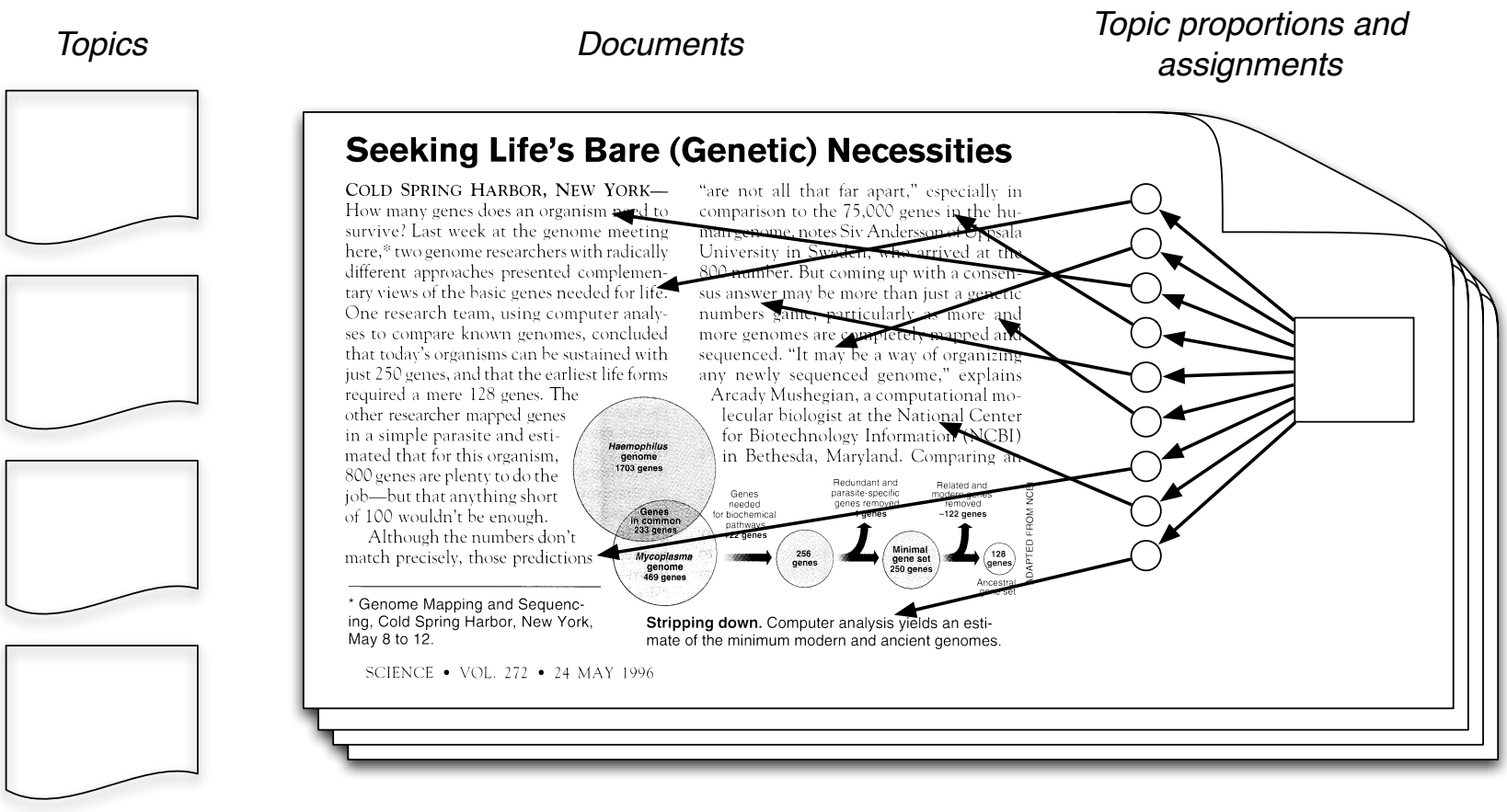
Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



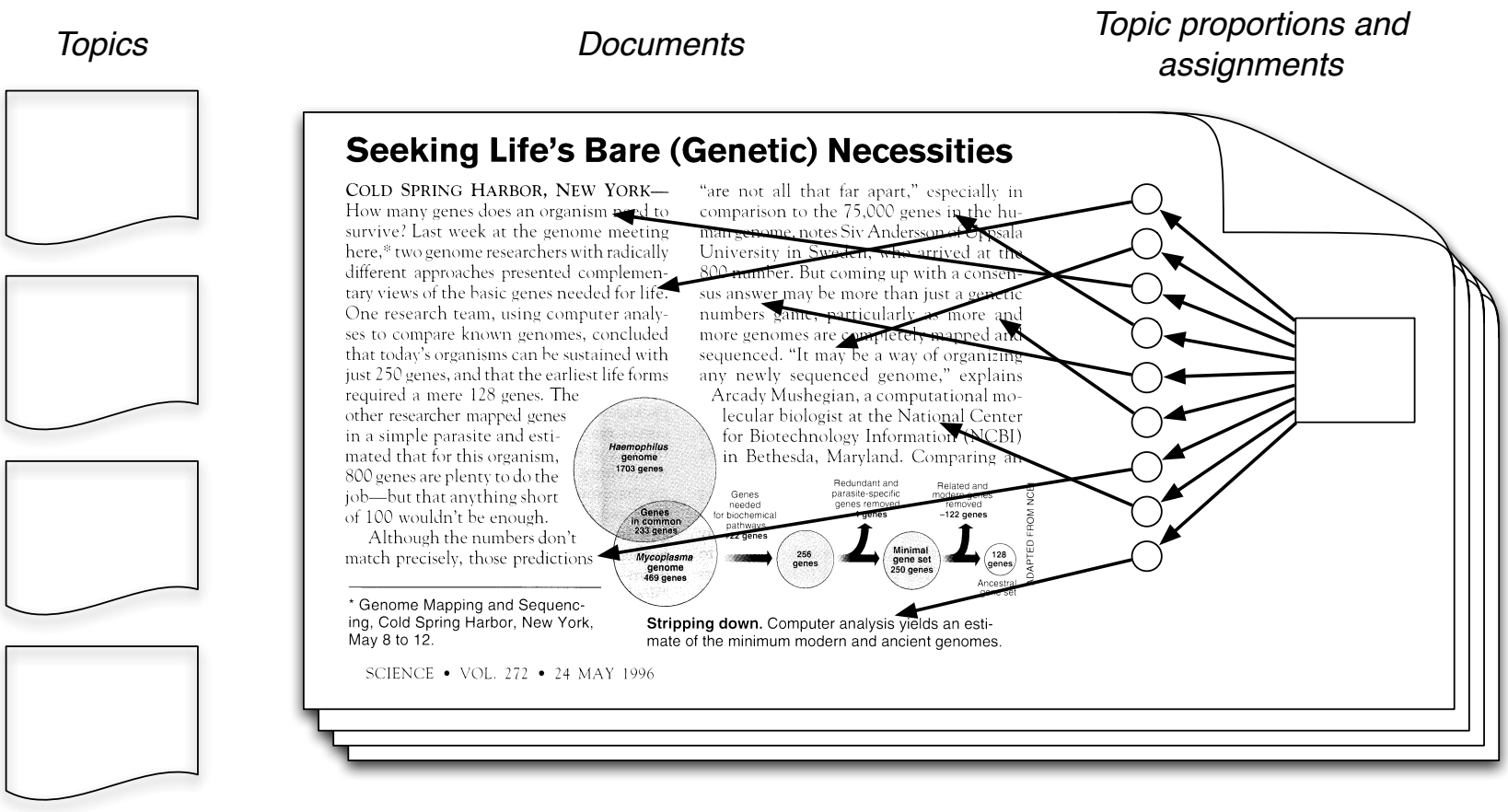
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

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

- Topic models automatically extract “semantic themes” from large document collections
- Based on latent variables, mixtures, and Bayesian inference
- Can be useful for a wide variety of data