For docs:

① generate a distribution of topics in a doc: { 70% politics, 30% sports

② pick a topic eg. sport

③ sport has a list of words, same distribution across all docs { 10% word A, 5% word B

④ pick a word

S&DS 365 / 565

Data Mining and Machine Learning

Topic Modelling

Yale

Intro to Topic Modeling

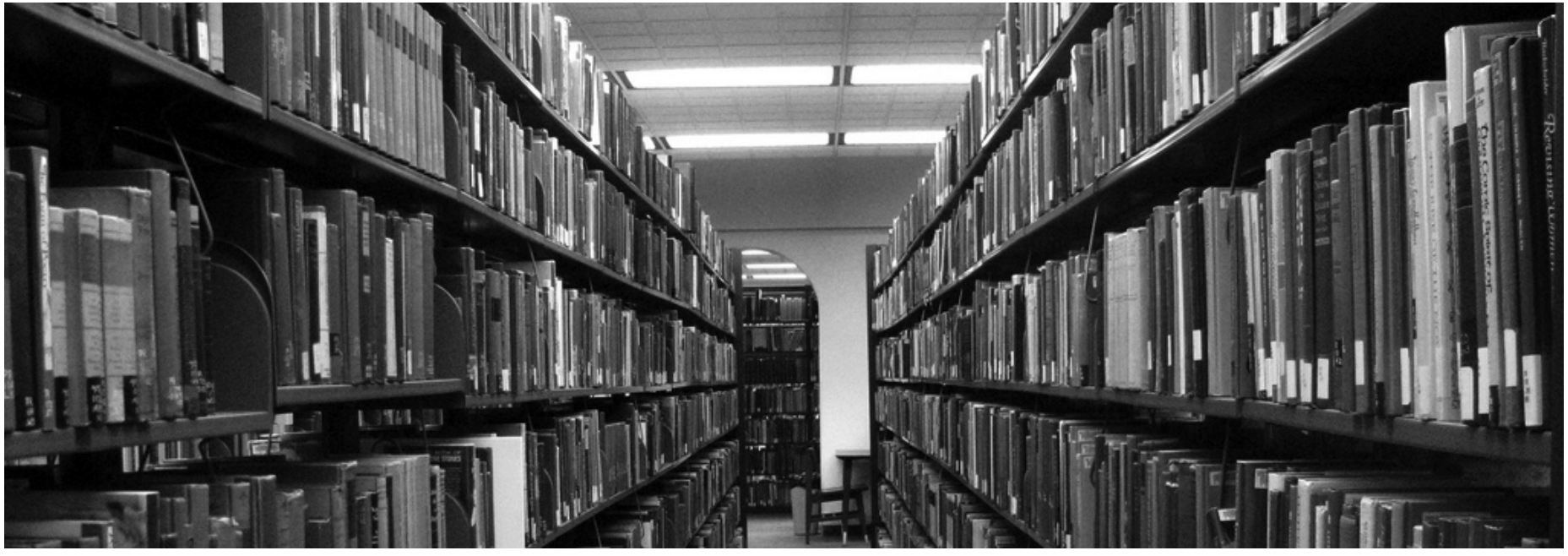
Some of the following slides are from Dave Blei's tutorial on Topic Modeling

`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



Aim : create index for books

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

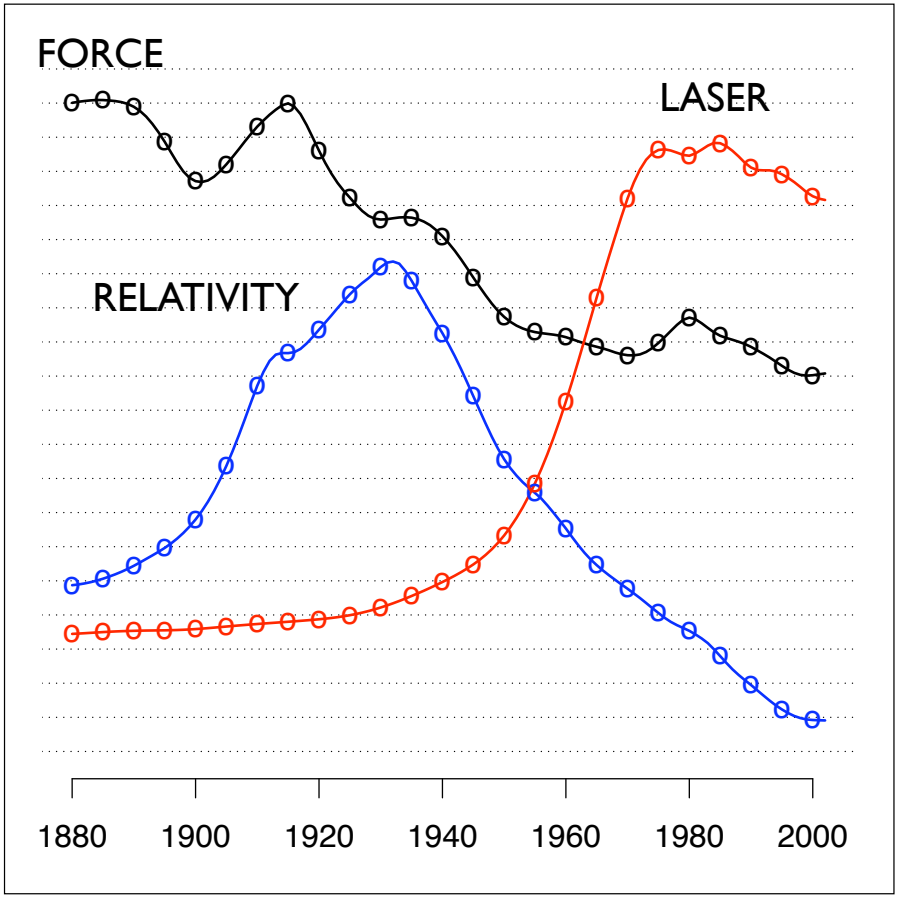
1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, and search the texts.

Discover topics from a corpus 语料库

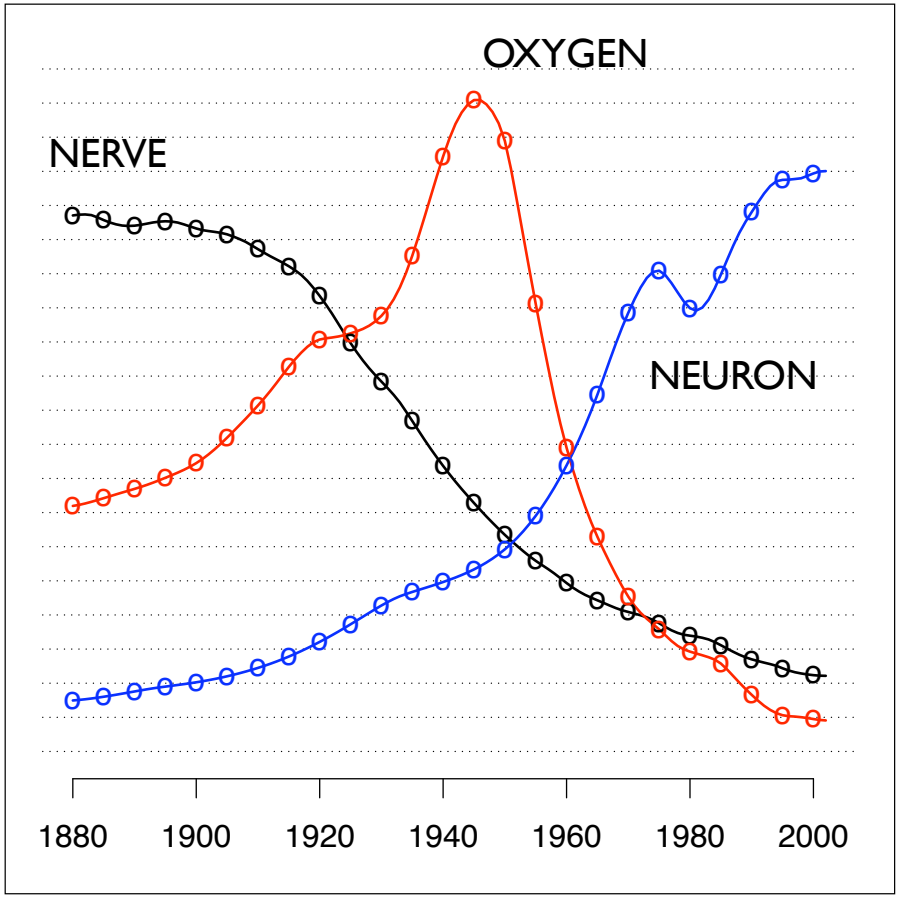
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	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

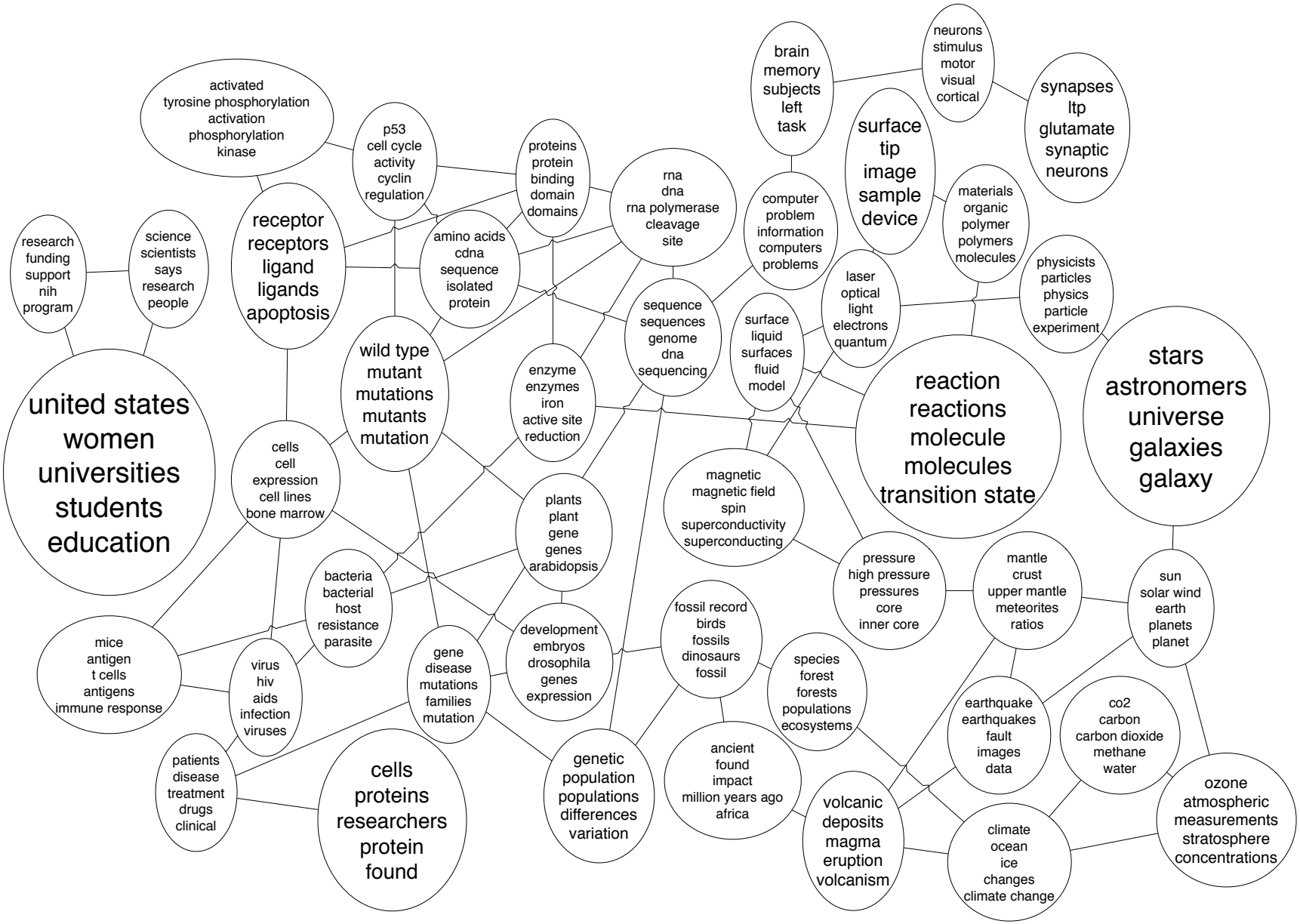
"Theoretical Physics"



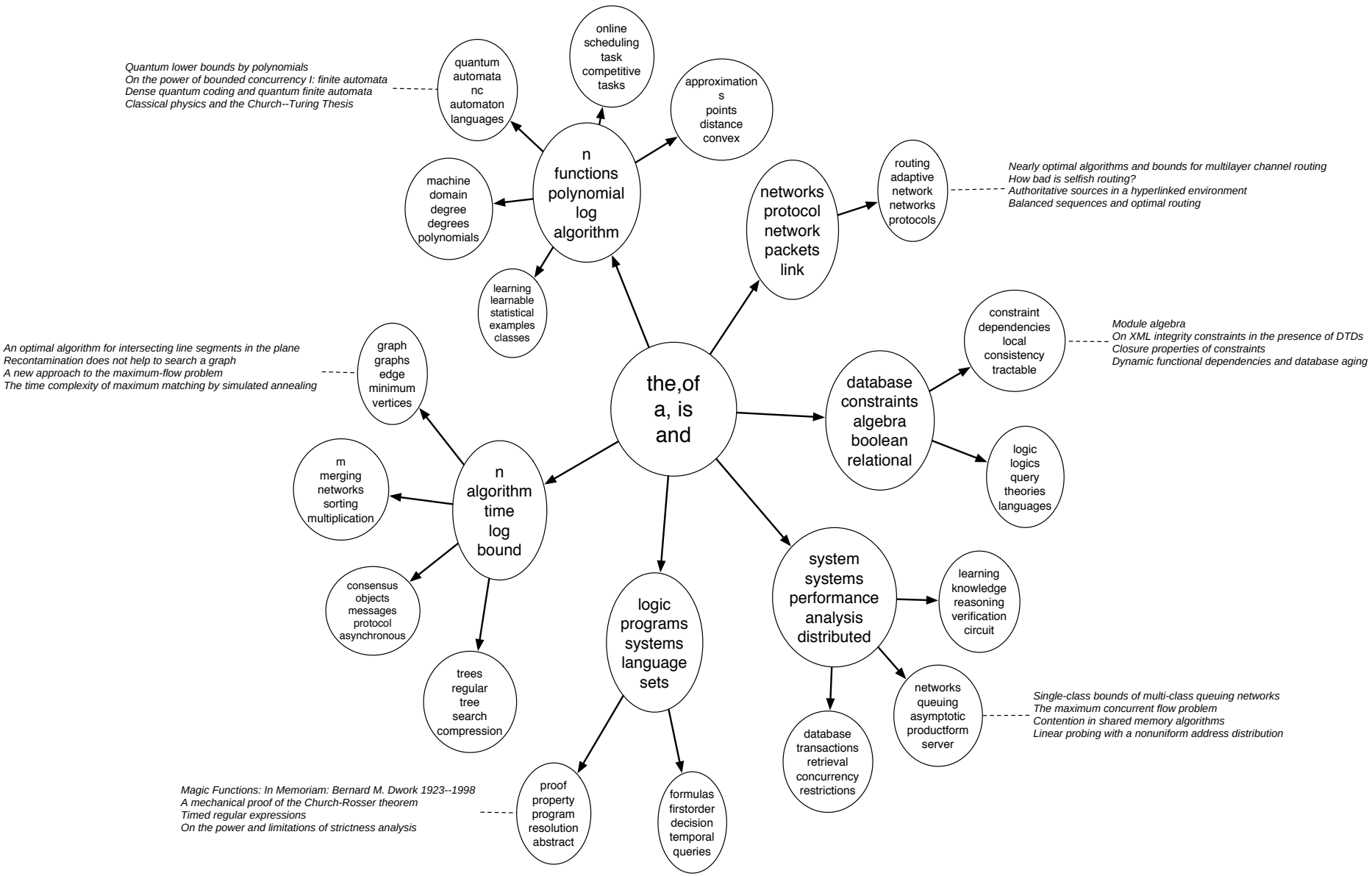
"Neuroscience"



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

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?*

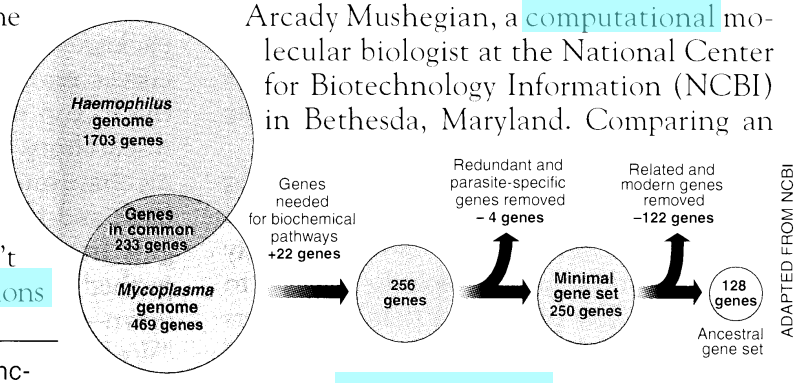
Latent Dirichlet allocation (LDA)

topics { yellow : genetics
pink : biology
blue : cs

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



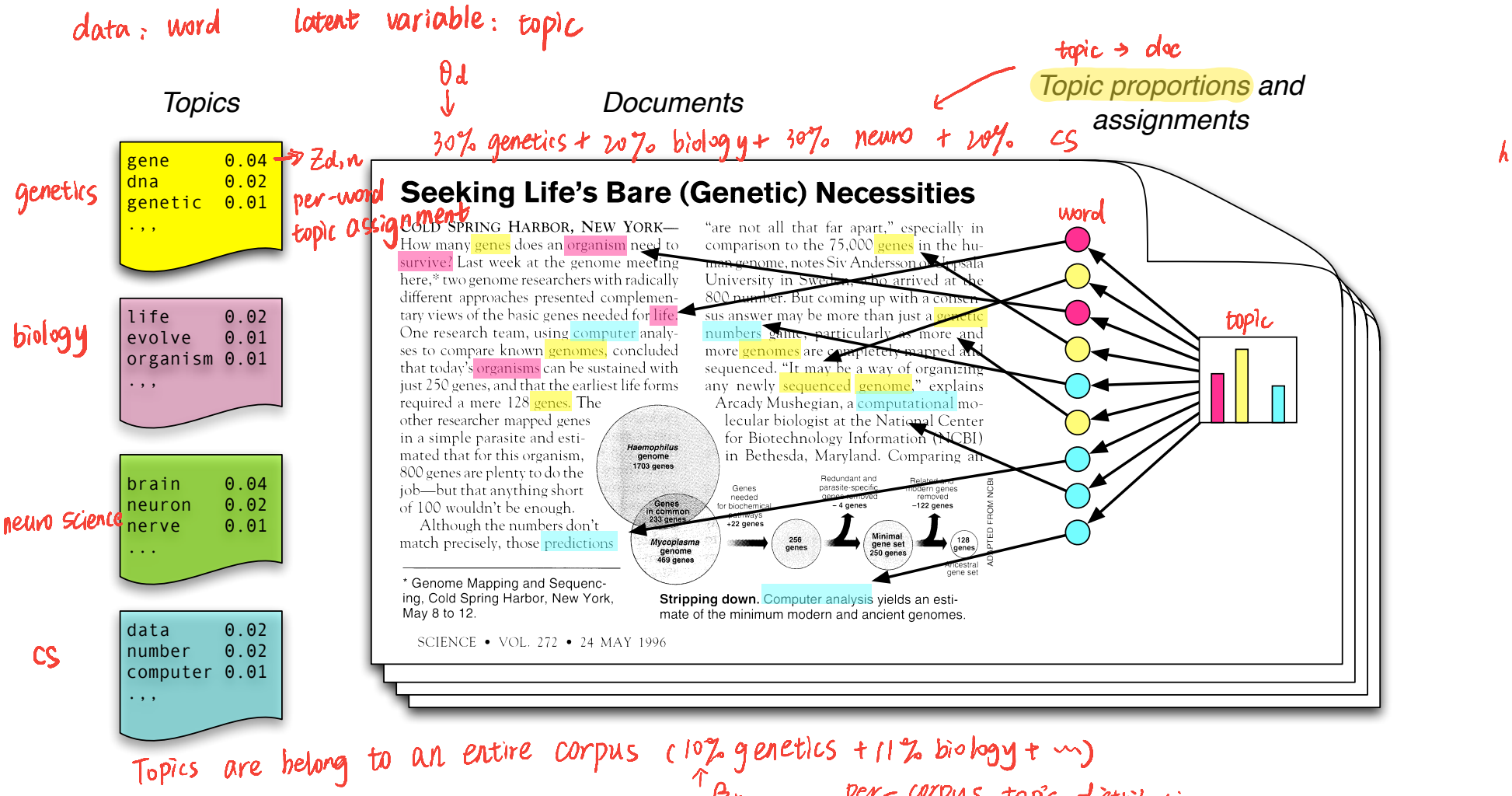
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

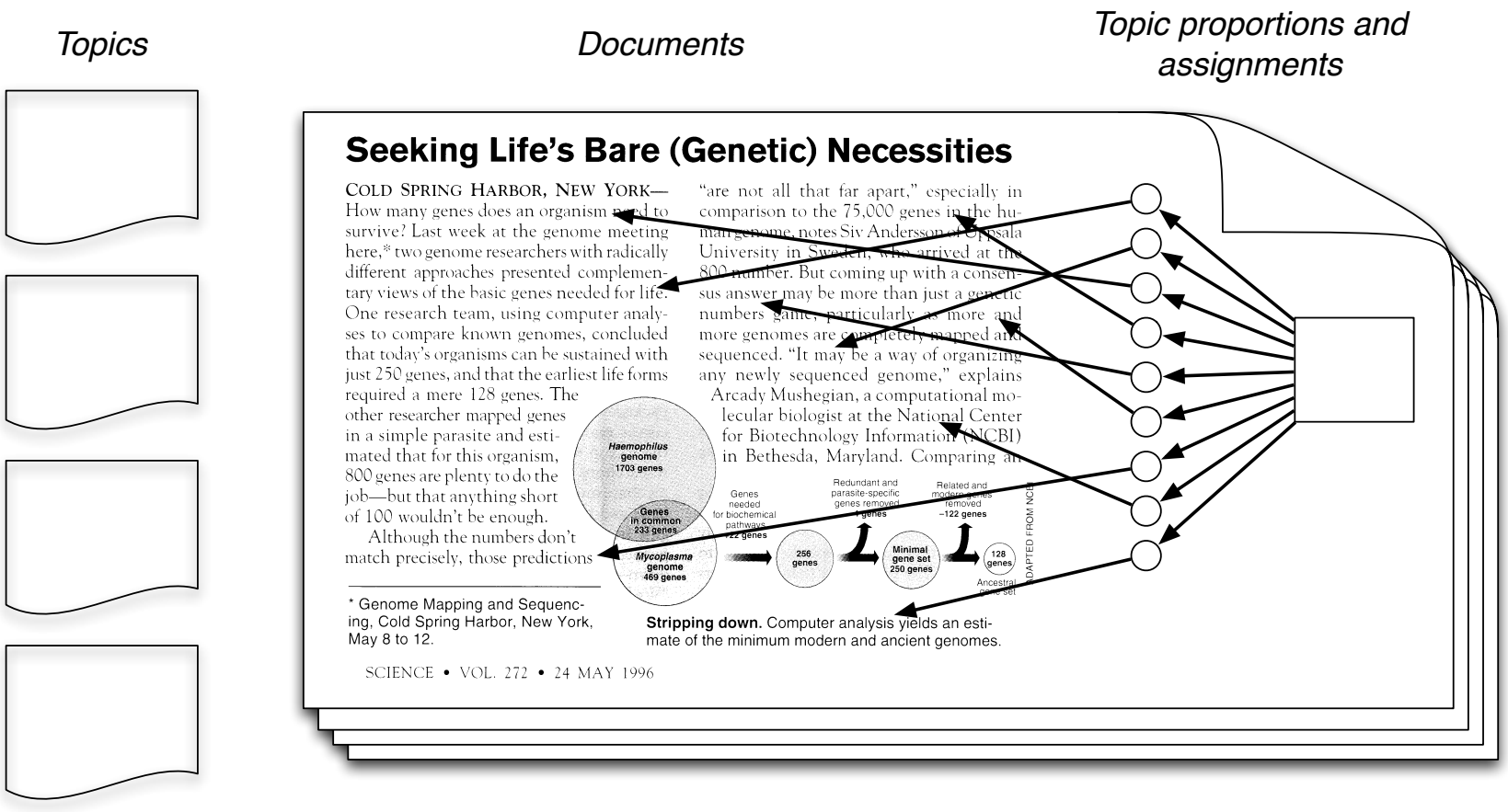
Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



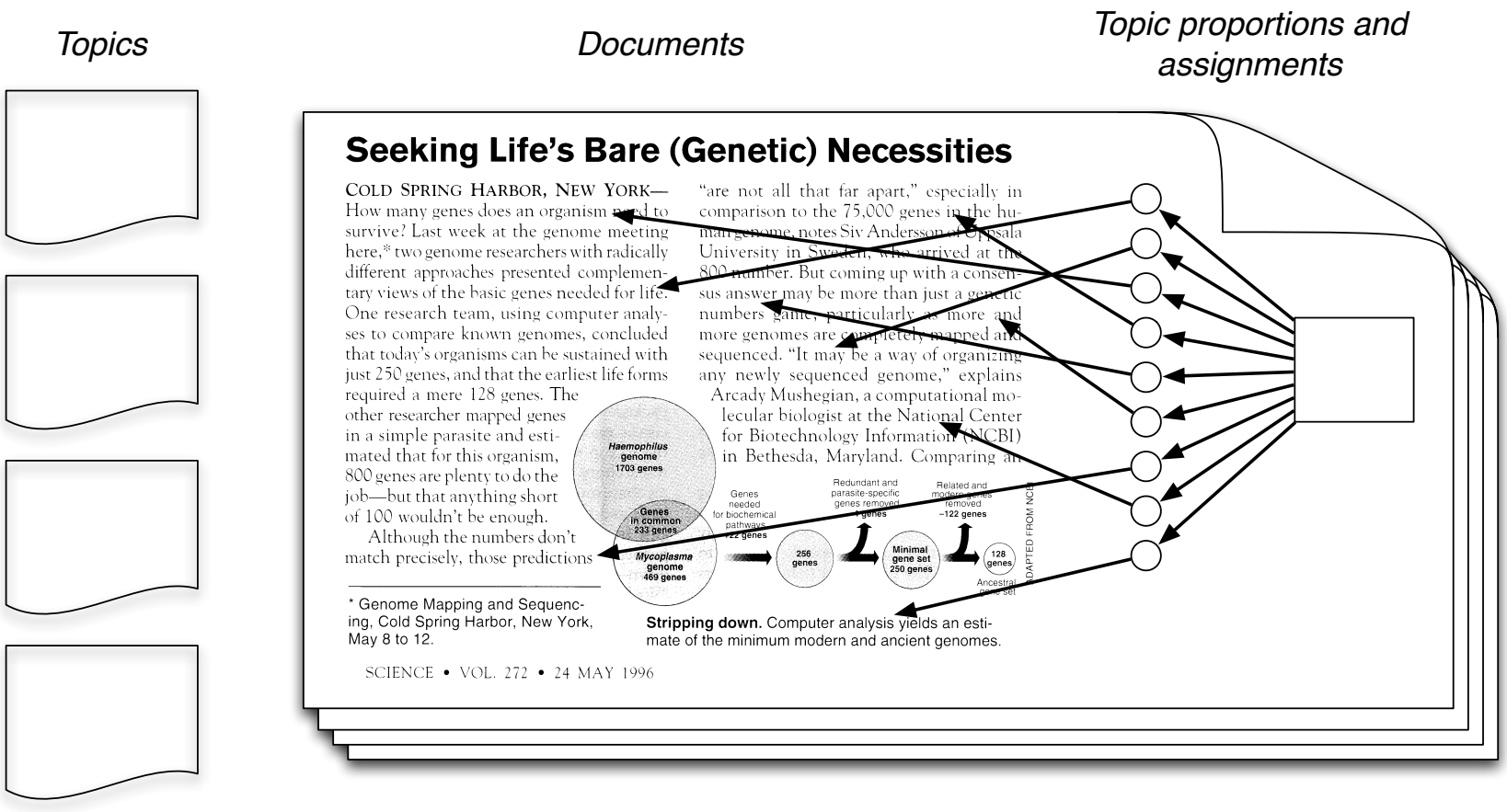
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



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

The posterior distribution



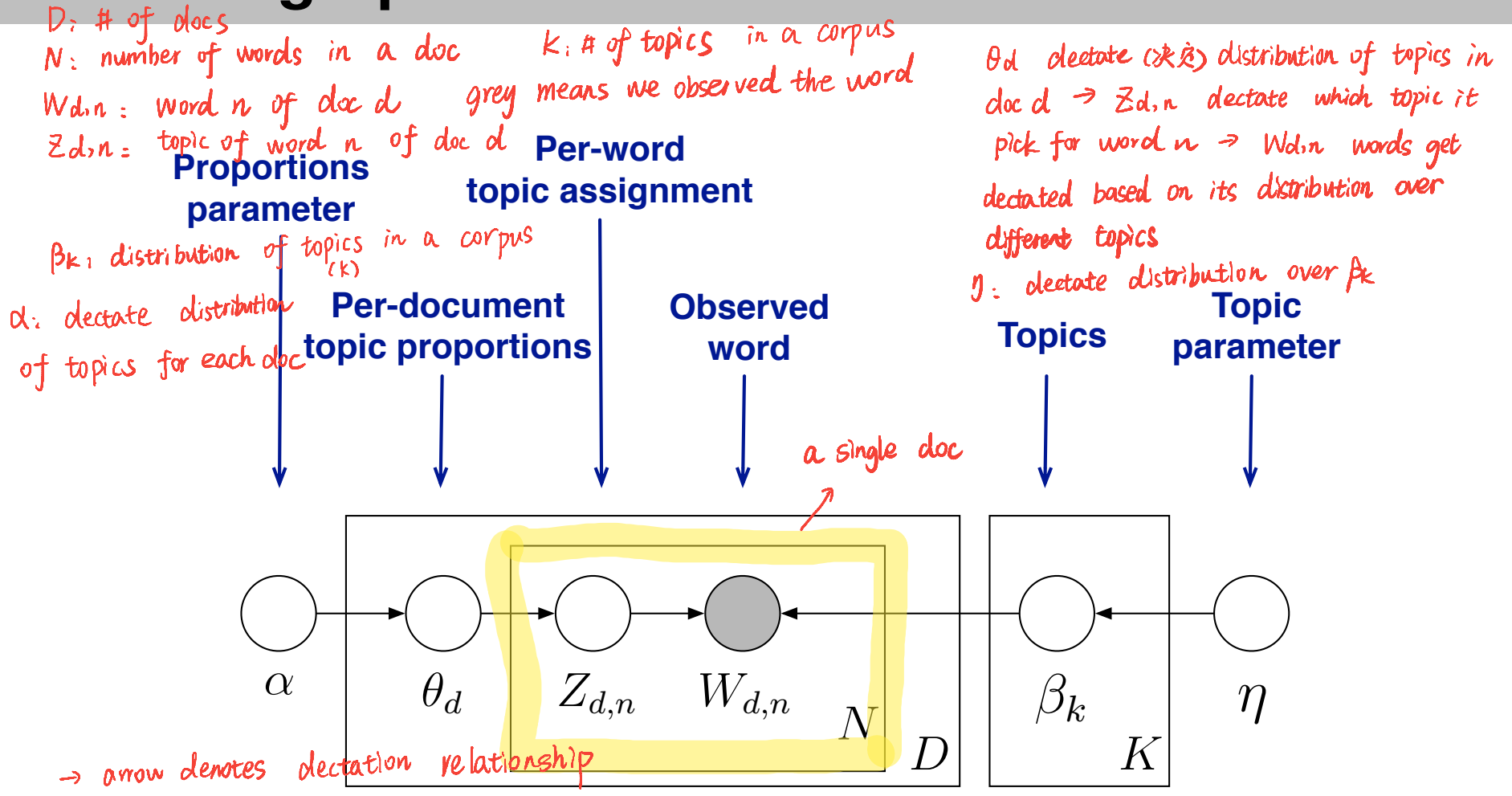
- Our goal is to **infer** the hidden variables
- I.e., compute their (posterior) distribution conditioned on the documents

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

Summary

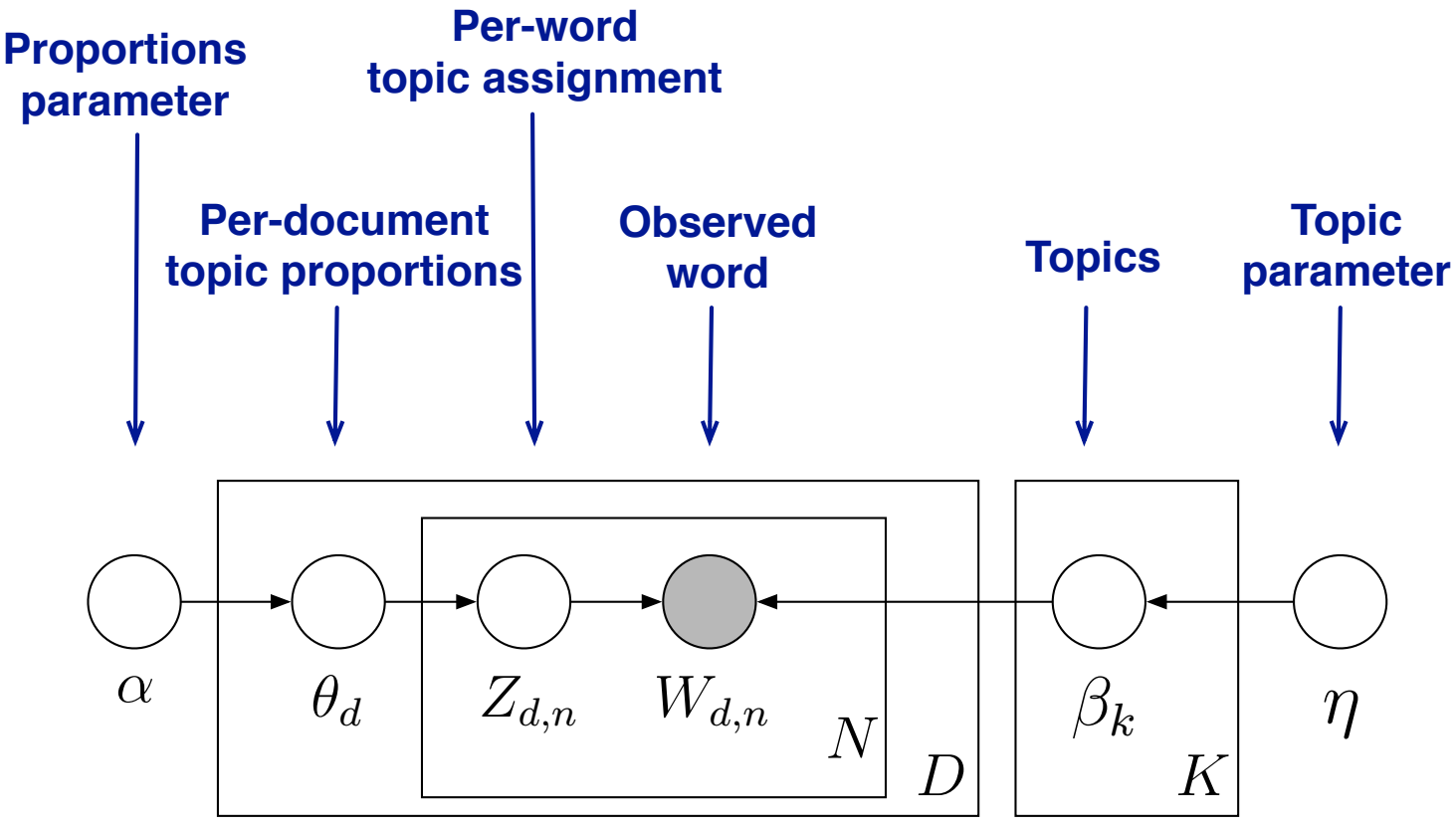
- Topic models automatically extract “semantic themes” from large document collections
- Based on latent variable (mixture) models
- “Dice rolling” generative models of words
- The topics come from computing the posterior

LDA as a graphical model



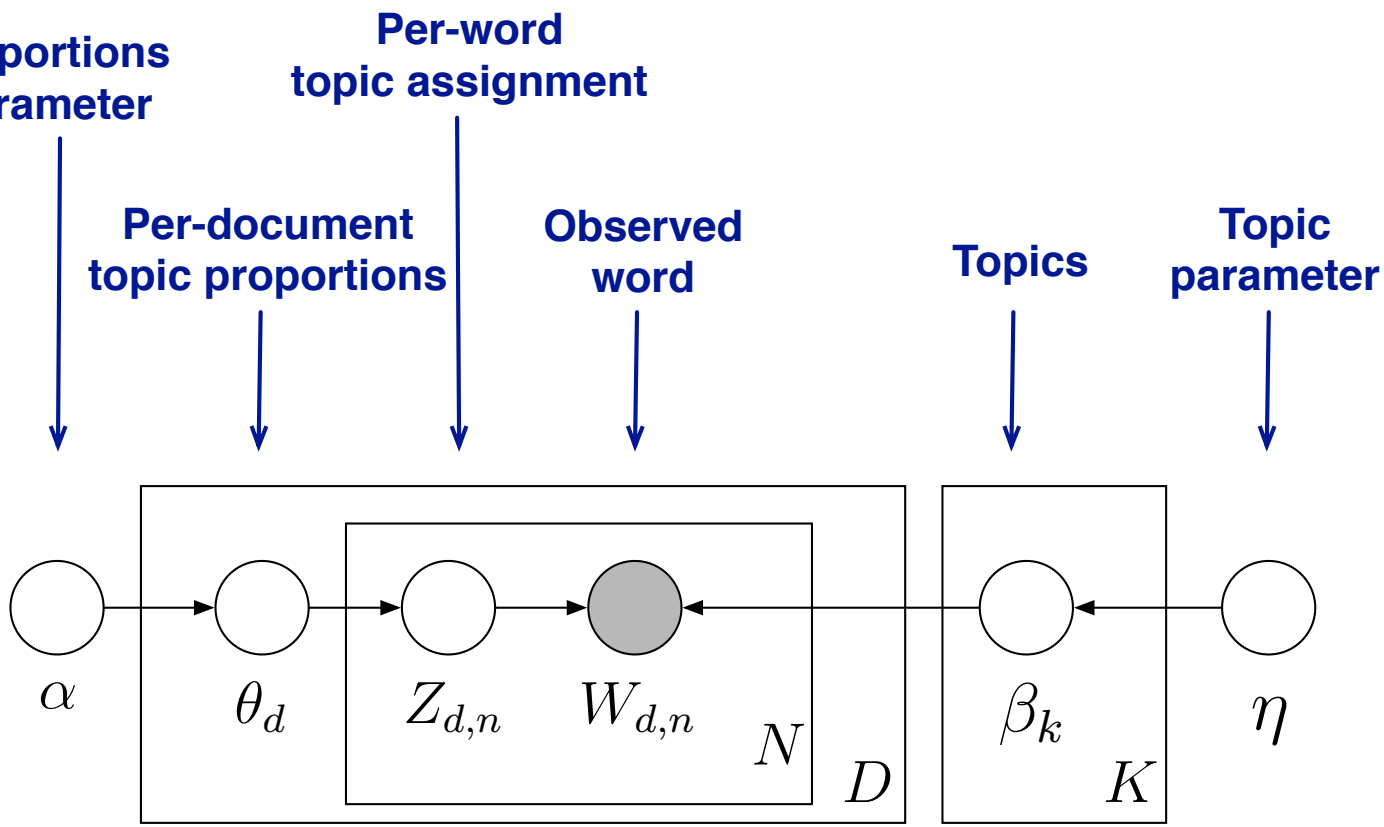
- Encodes our assumptions about the data
- Connects to algorithms for computing with data
- See *Pattern Recognition and Machine Learning* (Bishop, 2006).

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



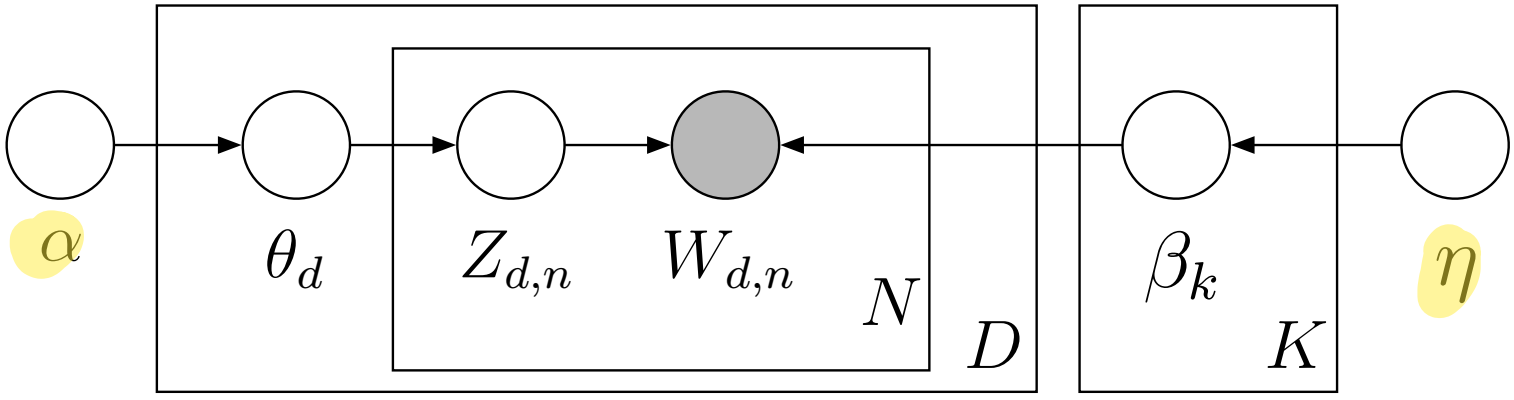
$\beta_i =$ [0.2, 0.1, 0.3, ⋮] → word

for 每个 i

$\theta_d =$ [0.1, 0.5, ⋮] → topic

$$p(W_{d,n}) = \underbrace{\prod_{i=1}^{K} p(\beta_i \mid \eta)}_{\text{topic}} \underbrace{\prod_{d=1}^{D} p(\theta_d \mid \alpha)}_{\text{doc}} \underbrace{\left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)}_{\text{pick topic for word} \times \text{pick word} \;=\; \text{word}}$$

LDA



- This joint defines a posterior. 后验概率 params: α, η
- 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, 在手头
  e.g., information retrieval, document similarity, exploration, ...

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.

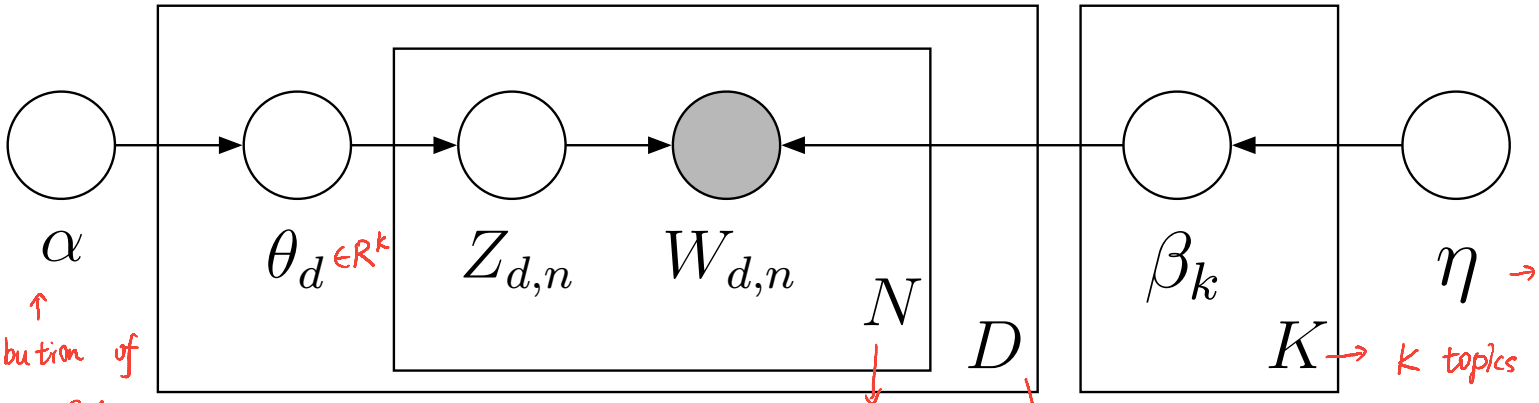
Why does LDA “work”?

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

Model choice: how to pick α, η will determine how spiky θ_d, β_k is



α: prior distribution of structure of θ_d
- one extreme: a doc only has a topic
- another extreme: a doc is a uniformly flat mixed of all topics

in practical: mixed topics in a doc

N: N words in a single doc

D: D docs

β_k: K topics across doc

η: prior distribution of structure of β_k

- 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)