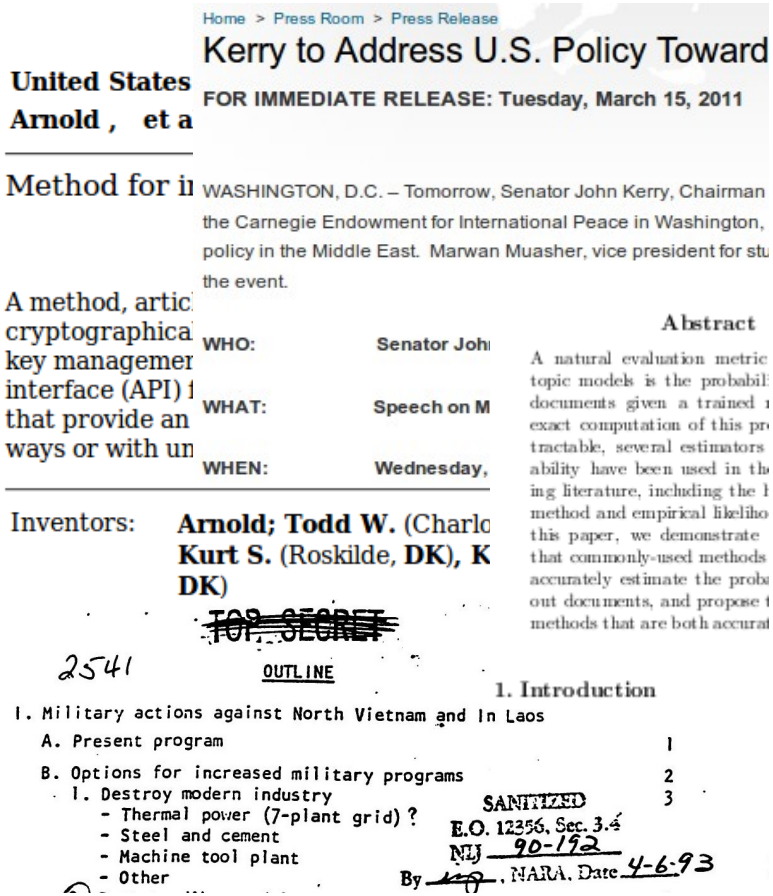
Statistical Topic Modeling

Hanna M. Wallach

University of Massachusetts Amherst

wallach@cs.umass.edu

Text as Data



Home > Press Room > Press Release

Kerry to Address U.S. Policy Toward

FOR IMMEDIATE RELEASE: Tuesday, March 15, 2011

WASHINGTON, D.C. – Tomorrow, Senator John Kerry, Chairman
the Carnegie Endowment for International Peace in Washington,
policy in the Middle East. Marwan Muasher, vice president for stu
the event.

WHO: Senator Joh

WHAT: Speech on M

WHEN: Wednesday,

United States
Arnold , et a

Method for i

A method, artic
cryptographica
key managemer
interface (API) f
that provide an
ways or with un

Inventors: **Arnold; Todd W.** (Charlo
Kurt S. (Roskilde, **DK**), **K**
DK)

Abstract

A natural evaluation metric
topic models is the probabil
documents given a trained
exact computation of this pr
tractable, several estimators
ability have been used in th
ing literature, including the
method and empirical likeliho
this paper, we demonstrate
that commonly-used methods
accurately estimate the prob
out documents, and propose
methods that are both accurat

~~TOP SECRET~~

2541

OUTLINE

I. Military actions against North Vietnam and In Laos
A. Present program 1
B. Options for increased military programs 2
1. Destroy modern industry 3
- Thermal power (7-plant grid) ?
- Steel and cement
- Machine tool plant
- Other

SANITIZED
E.O. 12356, Sec. 3.4
NLJ 90-192
By , NARA, Date 4-6-93

1. Introduction

- Structured and formal: e.g., publications, patents, press releases
- Messy and unstructured: e.g., chat logs, OCRed documents, transcripts

⇒ Large scale, robust methods for analyzing text

Topic Modeling

- Three fundamental assumptions:
  - Documents have latent semantic structure (“topics”)
  - We can infer topics from word–document co-occurrences
  - Can simulate this inference algorithmically
- Given a data set, the goal is to
  - Learn the composition of the topics that best represent it
  - Learn which topics are used in each document

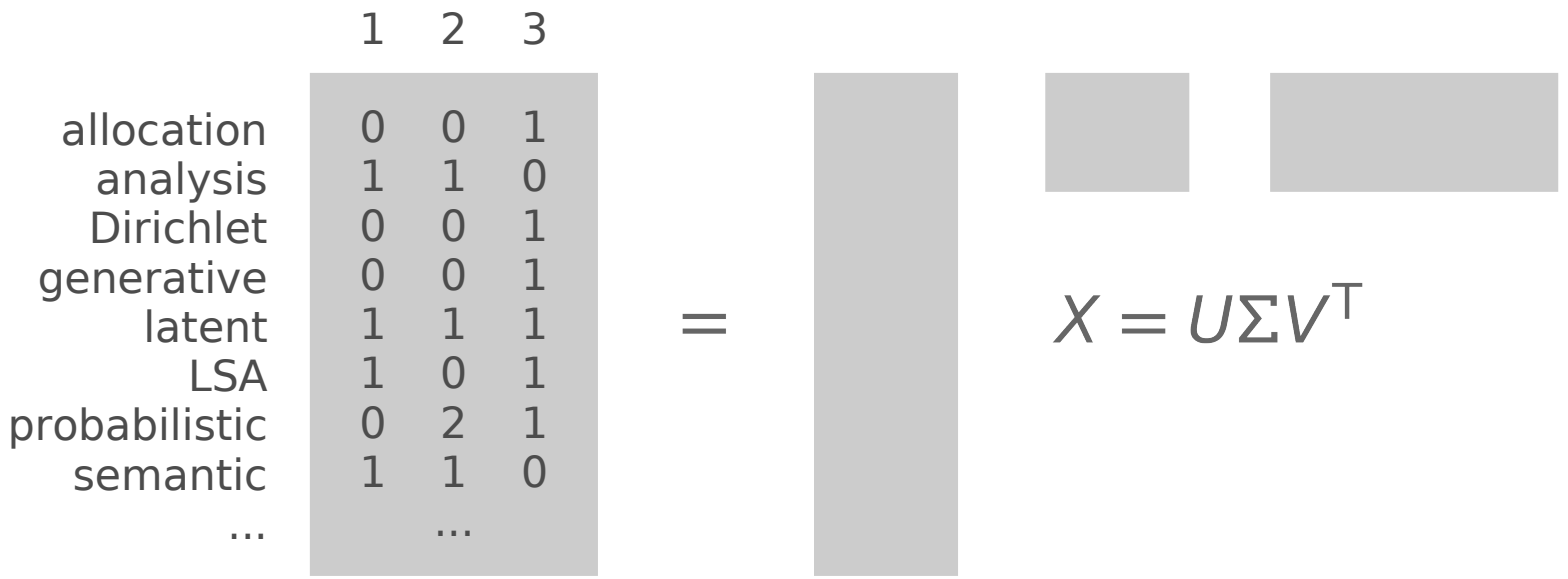
Latent Semantic Analysis (LSA)

(Deerwester et al., 1990)

- Based on ideas from linear algebra
- Form sparse term–document co-occurrence matrix X
  - Raw counts or (more likely) TF-IDF weights
- Use SVD to decompose X into 3 matrices:
  - U relates terms to “concepts”
  - V relates “concepts” to documents
  - Σ is a diagonal matrix of singular values

Singular Value Decomposition

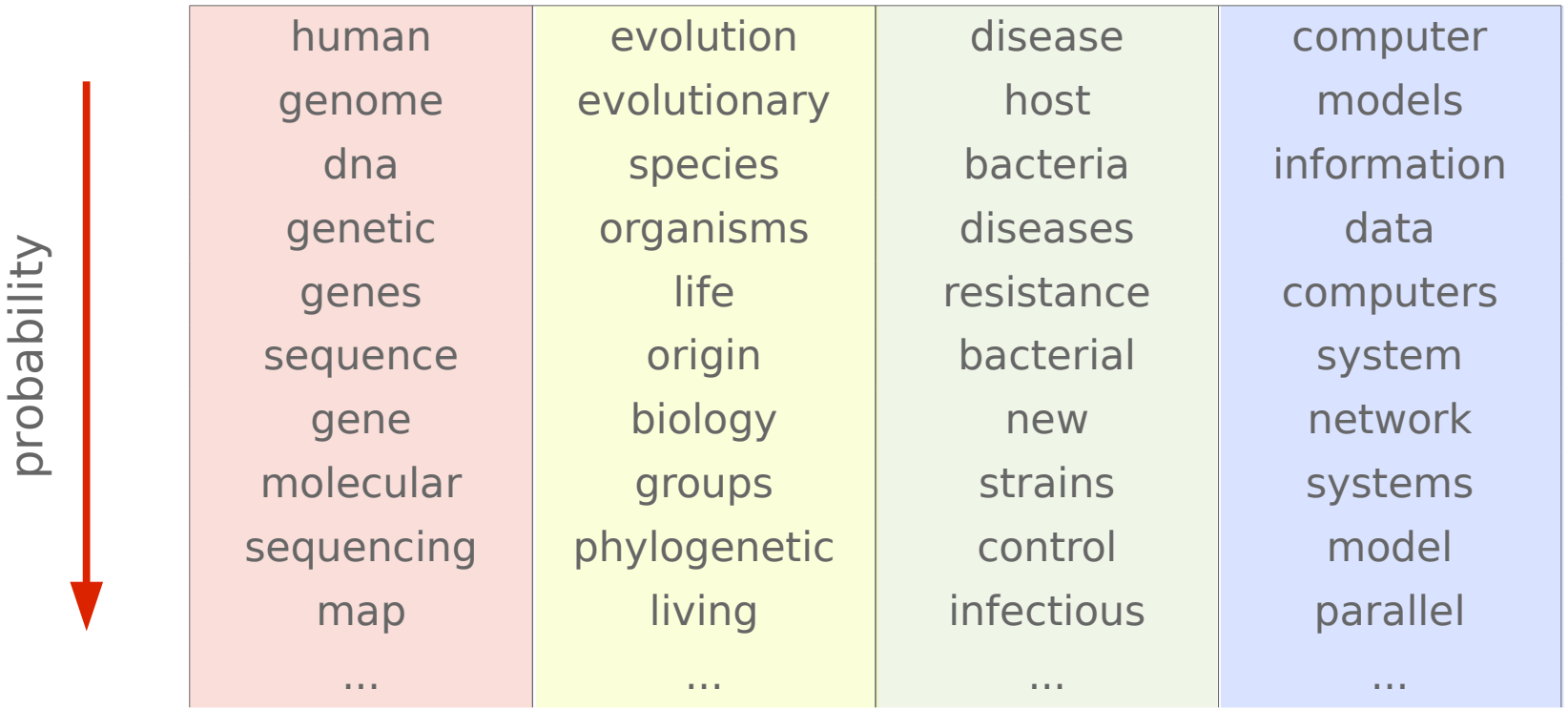
1. Latent semantic analysis (LSA) is a theory and method for ...
2. Probabilistic latent semantic analysis is a probabilistic ...
3. Latent Dirichlet allocation, a generative probabilistic model ...



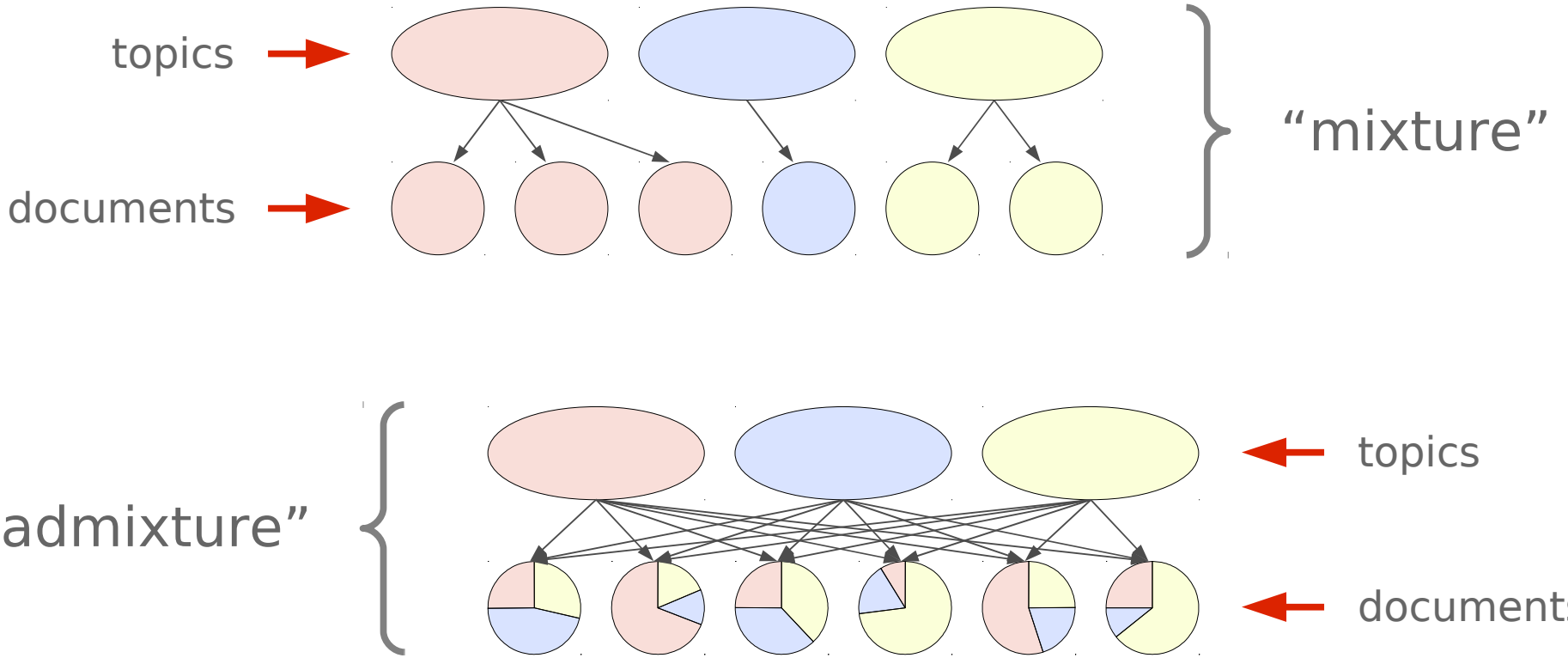
Generative Statistical Modeling

- Assume data was generated by a probabilistic model:
  - Model may have hidden structure (latent variables)
  - Model defines a joint distribution over all variables
  - Model parameters are unknown
- Infer hidden structure and model parameters from data
- Situate new data in estimated model

Topics and Words



Mixtures vs. Admixtures

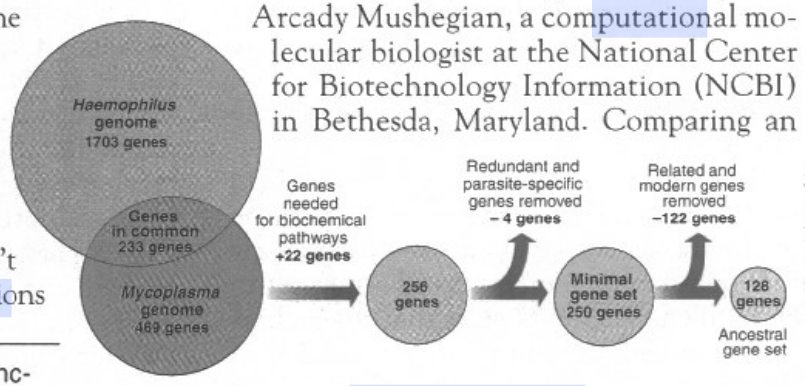


Documents and Topics

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.

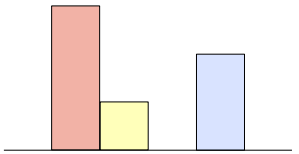
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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



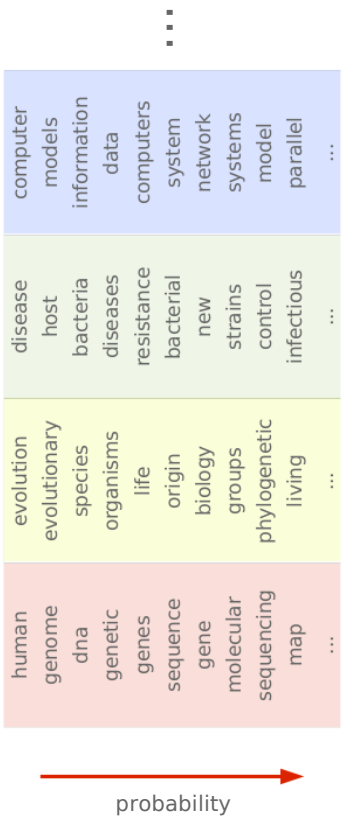
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

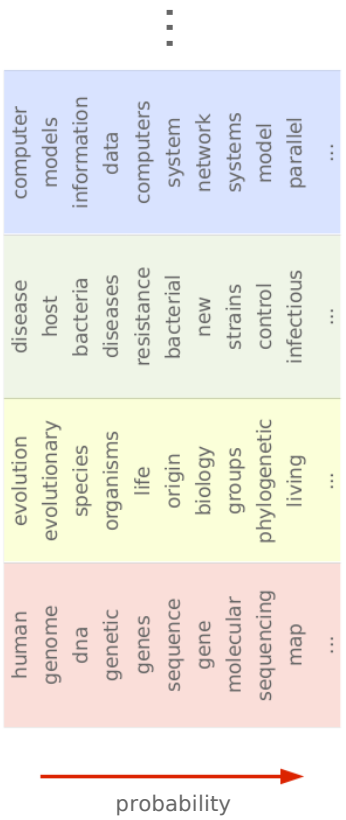
SCIENCE • VOL. 272 • 24 MAY 1996

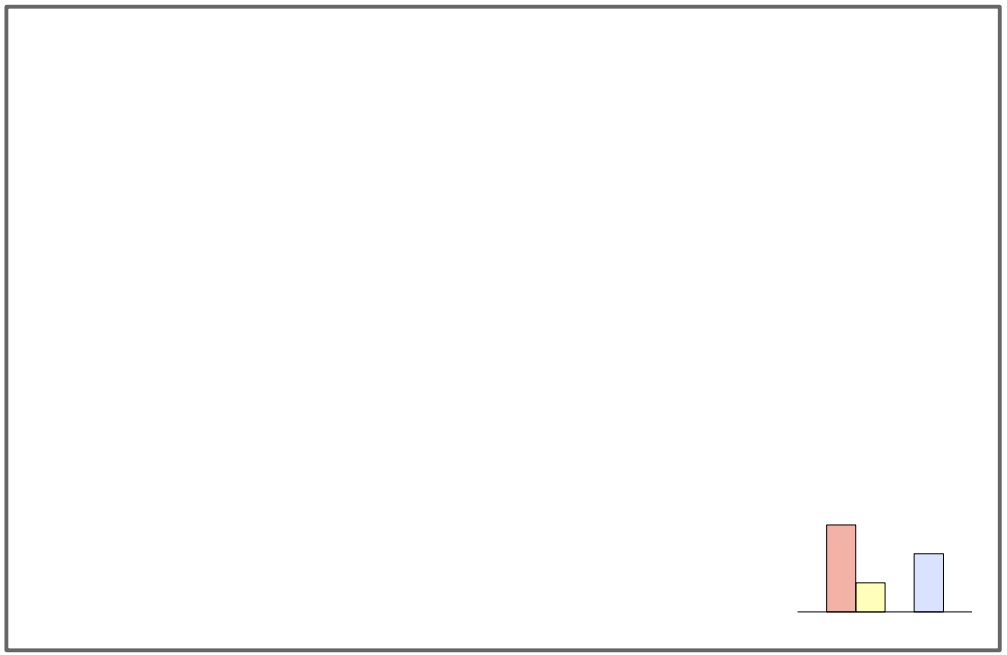


Generative Process

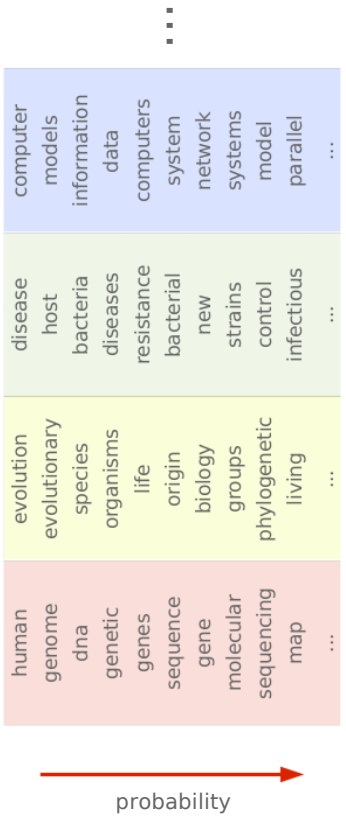


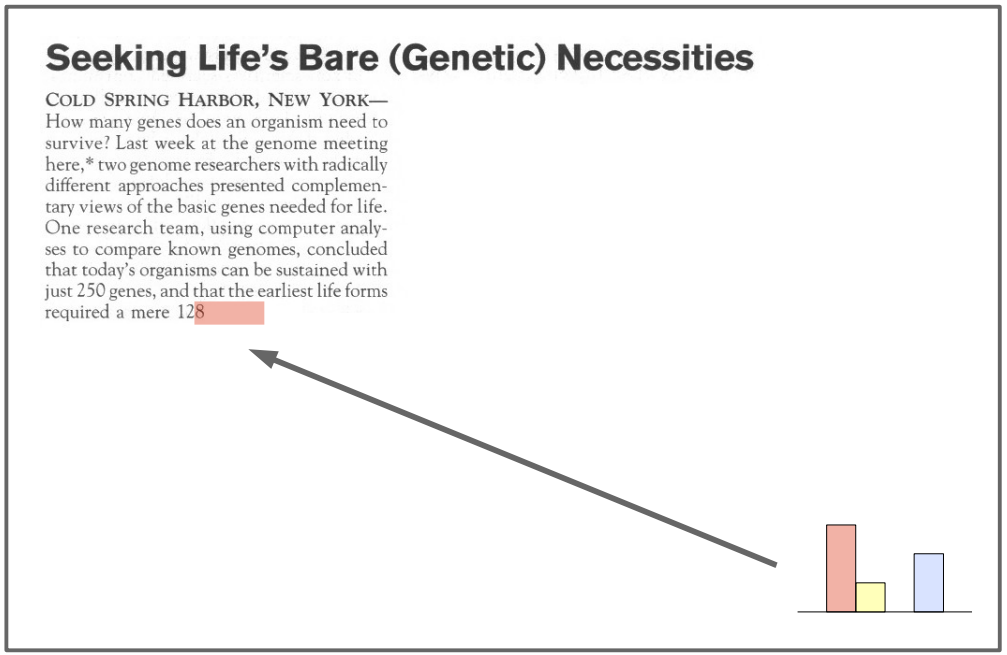
Choose a Distribution Over Topics





Choose a Topic

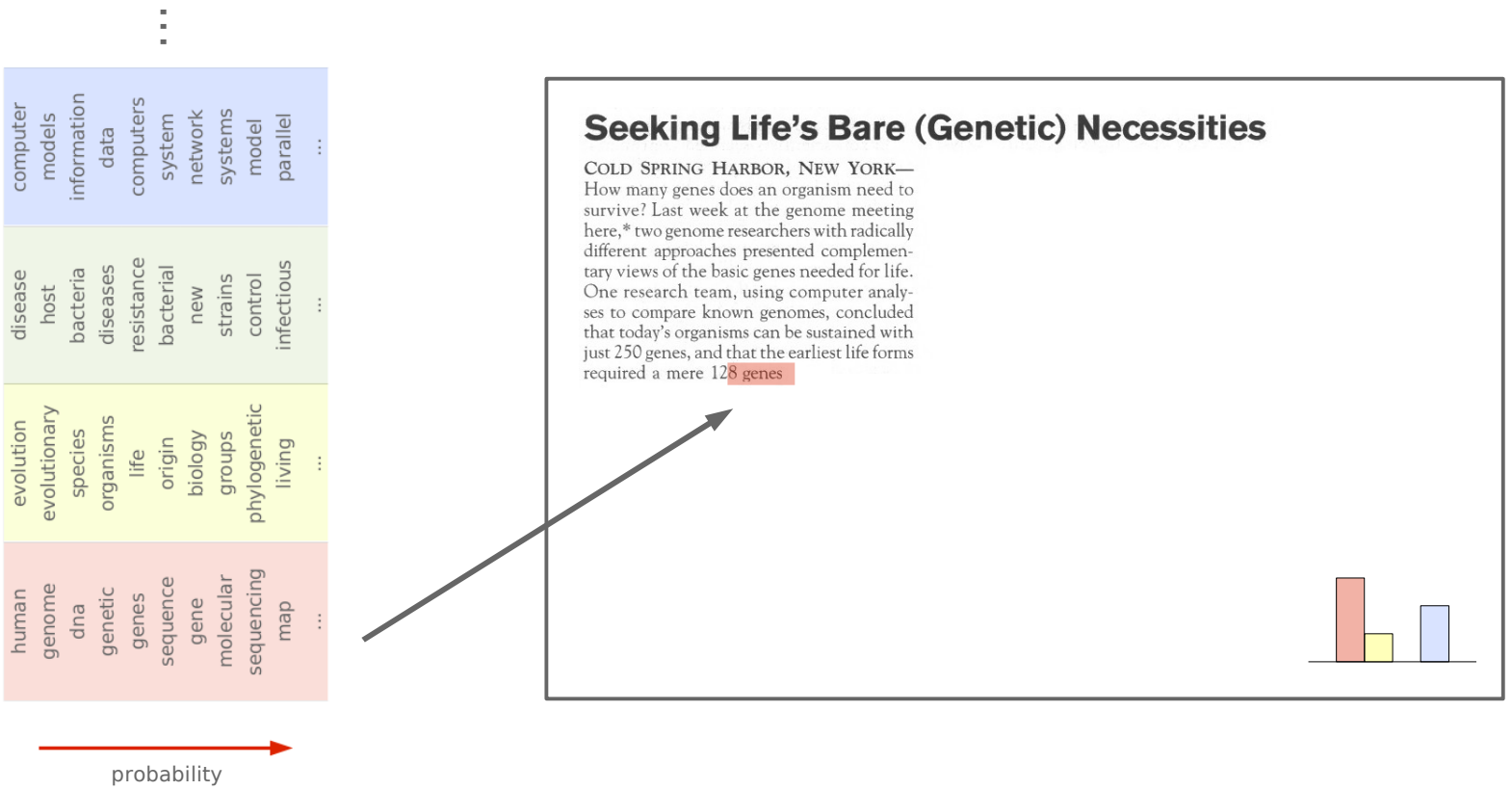




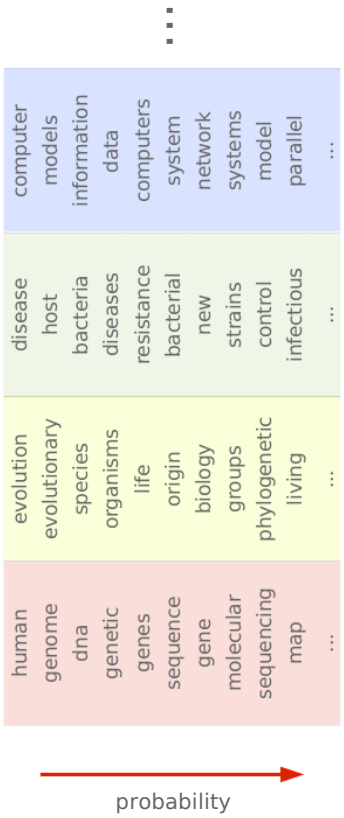
Seeking Life's Bare (Genetic) Necessities

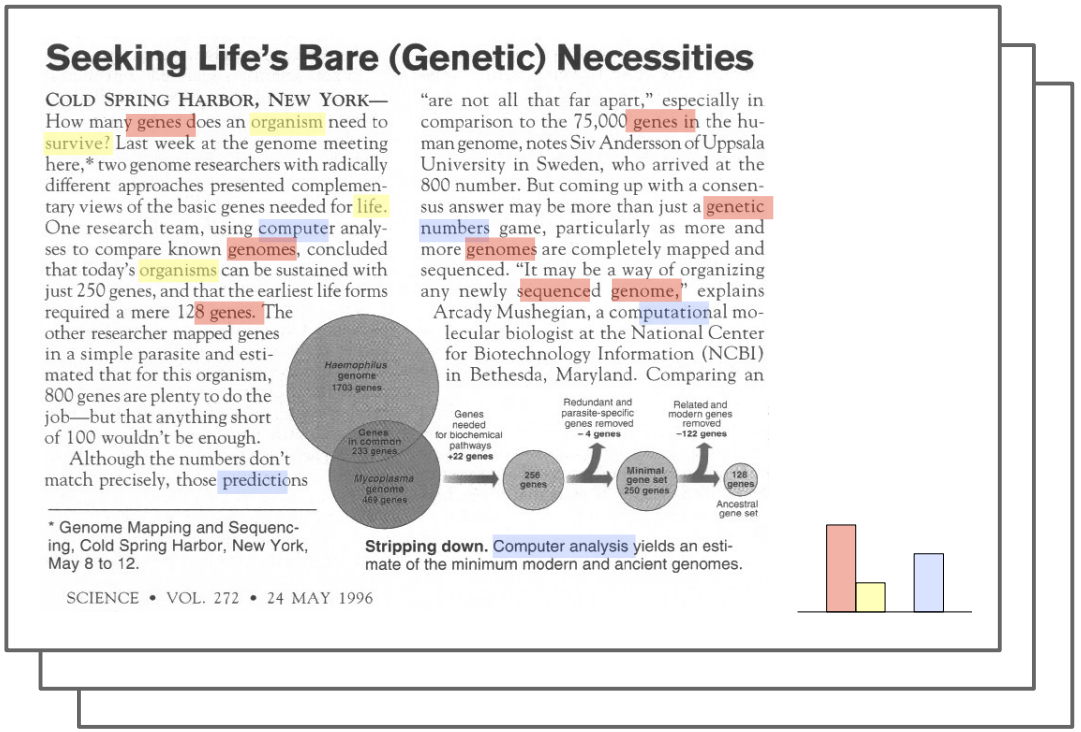
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128

Choose a Word



... And So On





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

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

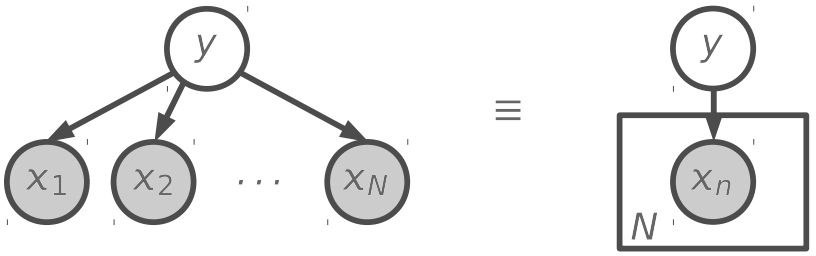
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Directed Graphical Models

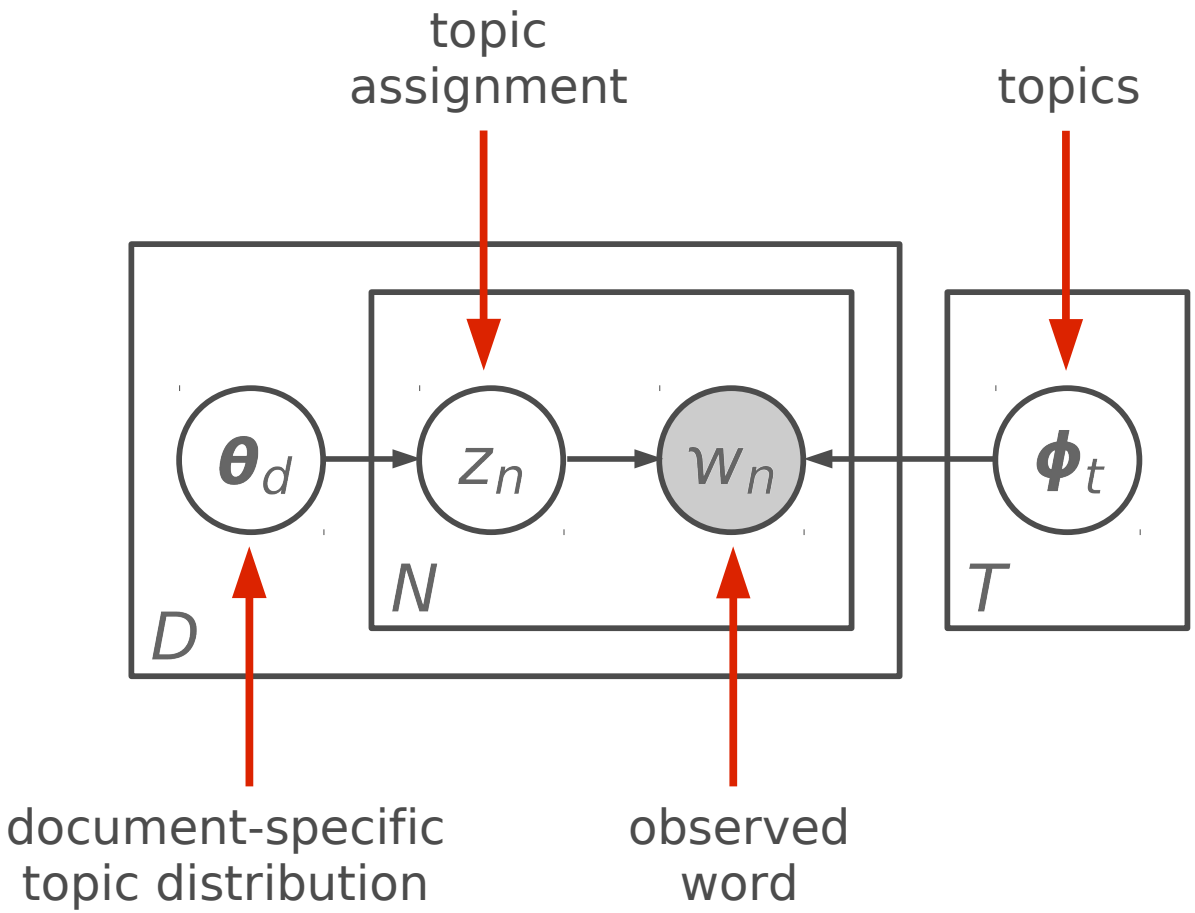
$$P(y, x_1, \ldots, x_N) = P(y) \prod_{n=1}^{N} P(x_n \mid y)$$

- Nodes: random variables (latent or observed)
- Edges: probabilistic dependencies between variables
- Plates: “macros” that allow subgraphs to be replicated



Probabilistic LSA

[Hofmann, '99]



Strengths and Weaknesses

✔ Probabilistic graphical model: can be extended and embedded into other more complicated models

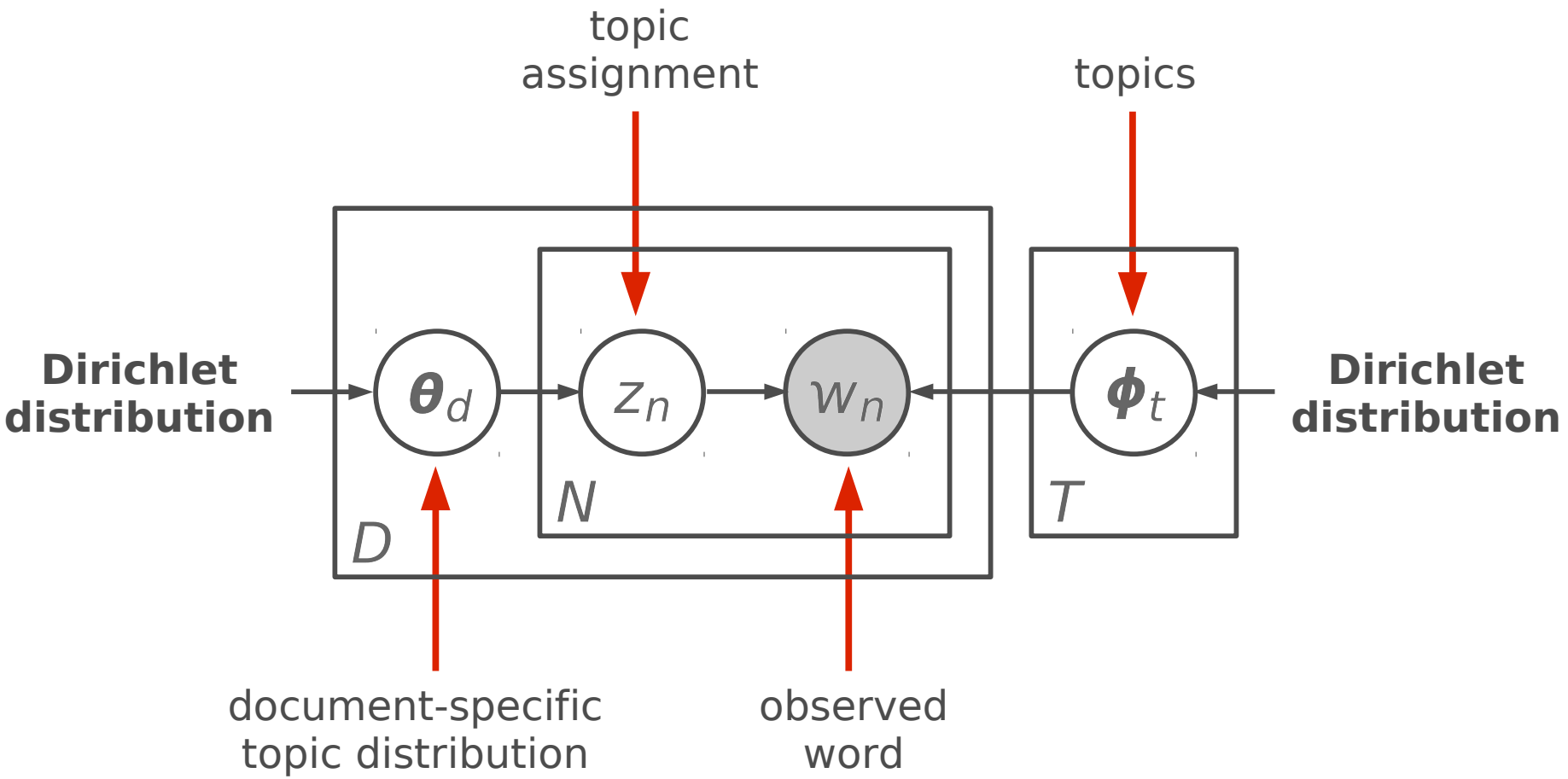
✗ Not a well-defined generative model: no way of generalizing to new, unseen documents

✗ Many free parameters: linear in # training documents

✗ Prone to overfitting: have to be careful when training

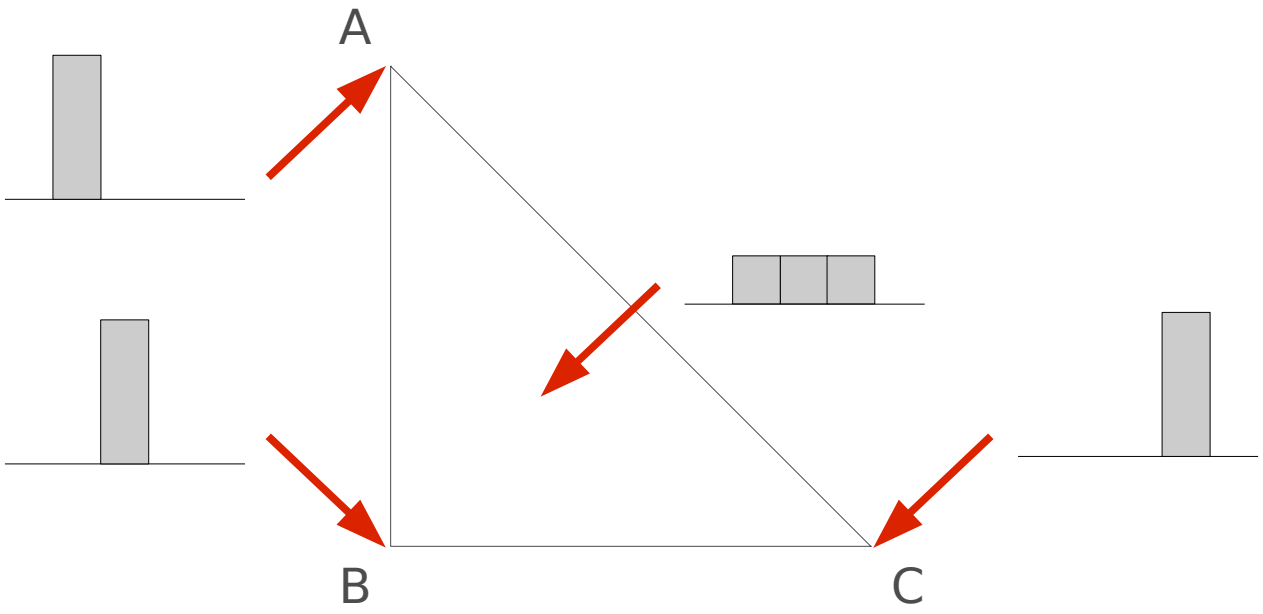
Latent Dirichlet Allocation (LDA)

[Blei, Ng & Jordan, '03]



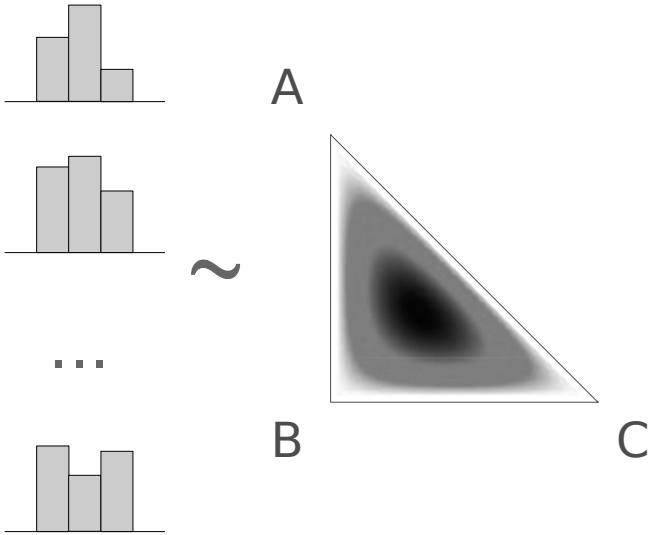
Discrete Probability Distributions

- 3-dimensional discrete probability distributions can be visually represented in 2-dimensional space:



Dirichlet Distribution

- Distribution over discrete probability distributions:

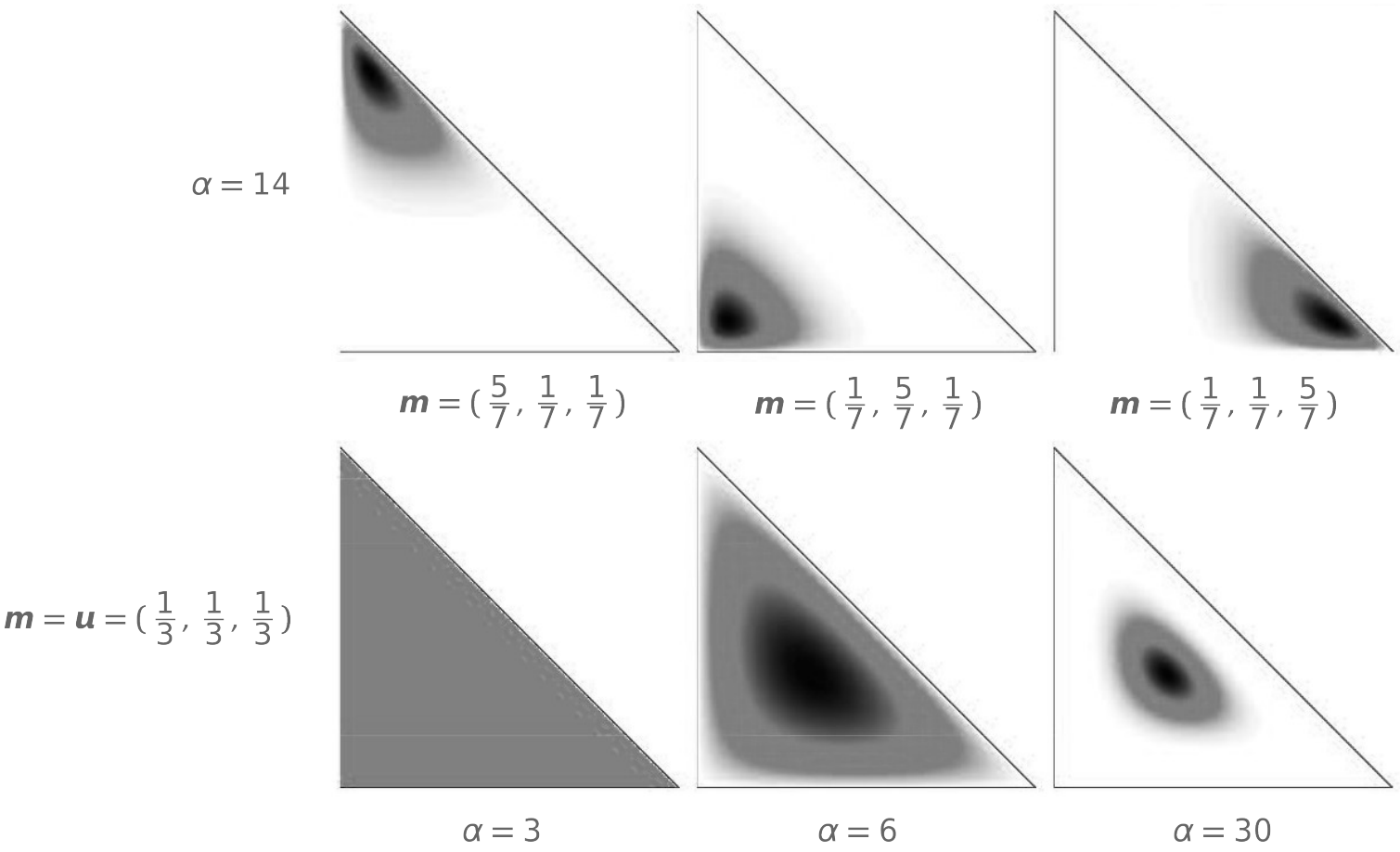


$\sim$ $\equiv$ $\boldsymbol{p} \sim \mathrm{Dir}(\alpha \boldsymbol{m})$

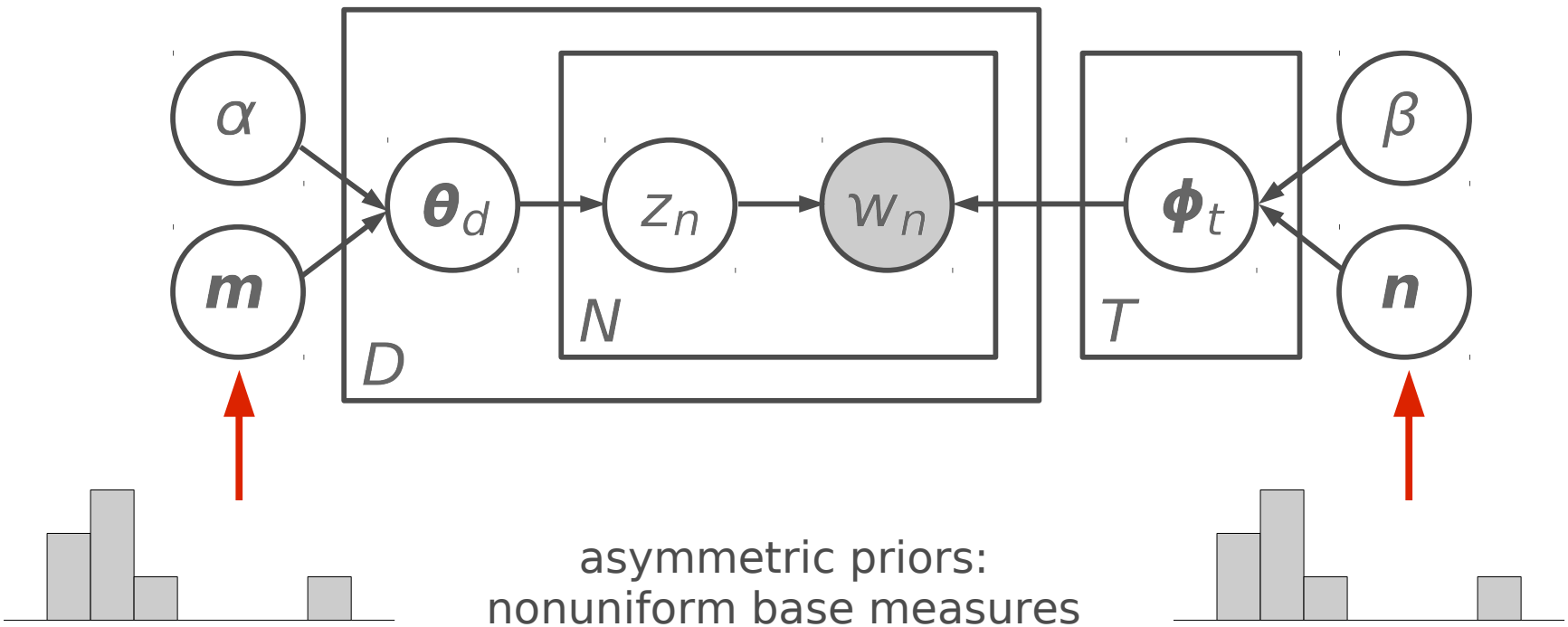
base measure (mean)

scalar concentration parameter

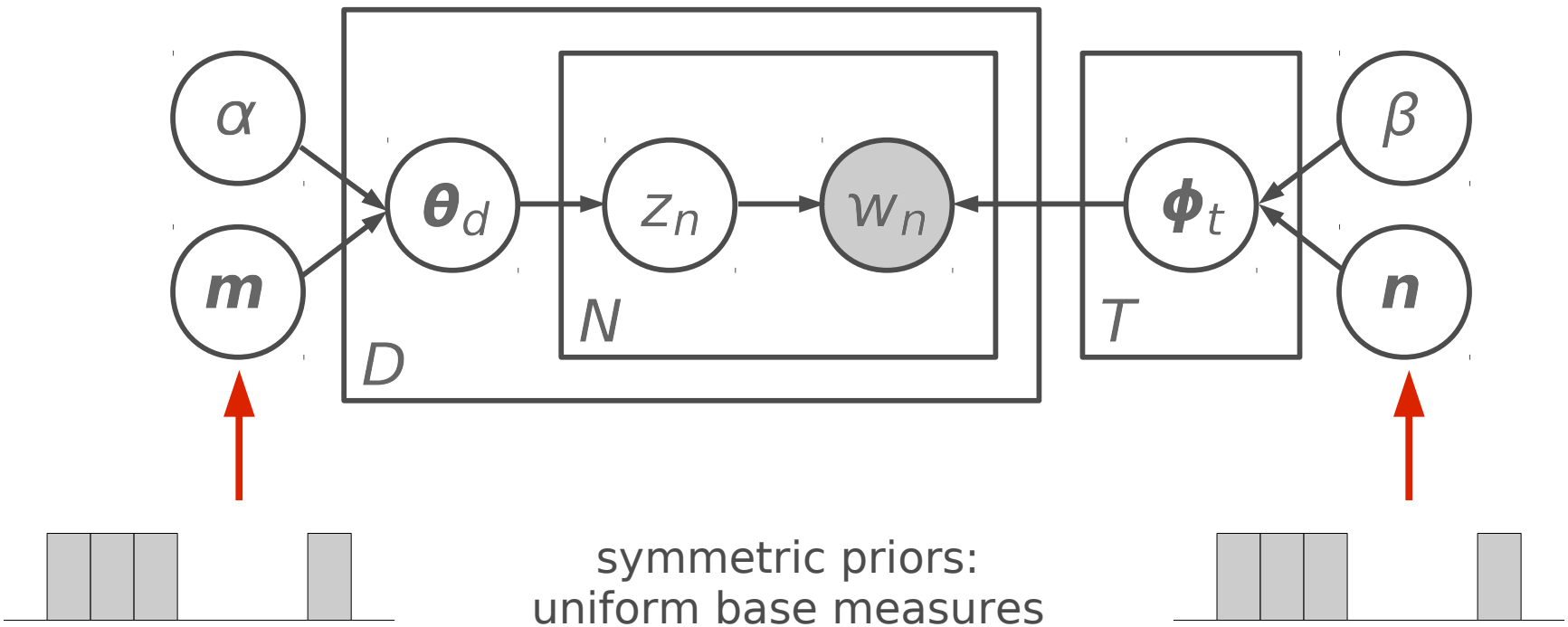
Dirichlet Parameters



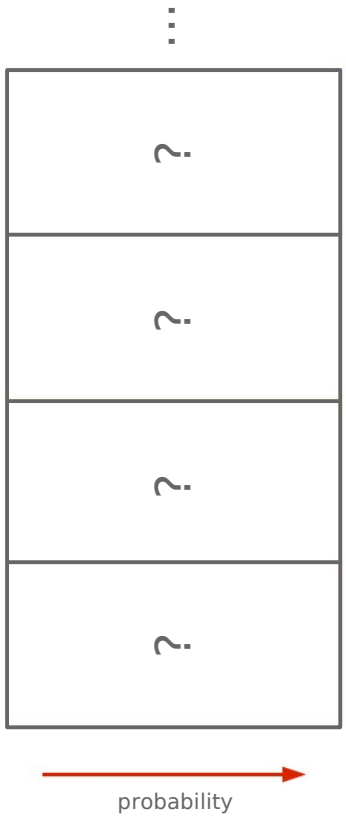
Dirichlet Priors for LDA

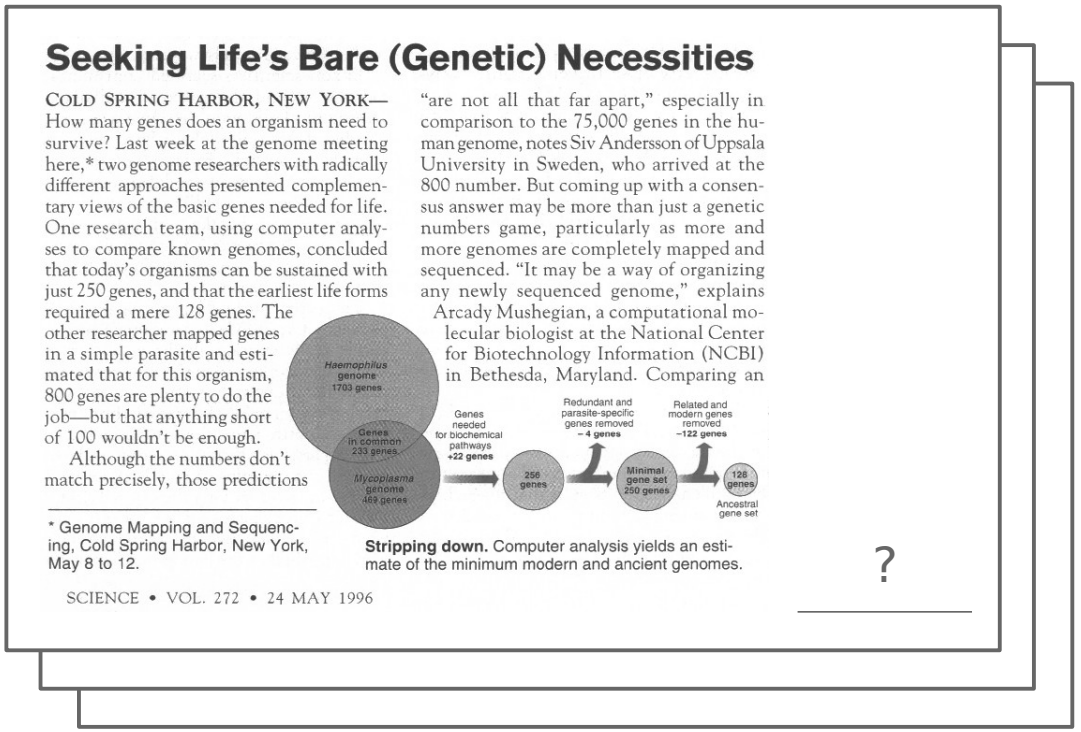


Dirichlet Priors for LDA



Real Data: Statistical 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,* 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

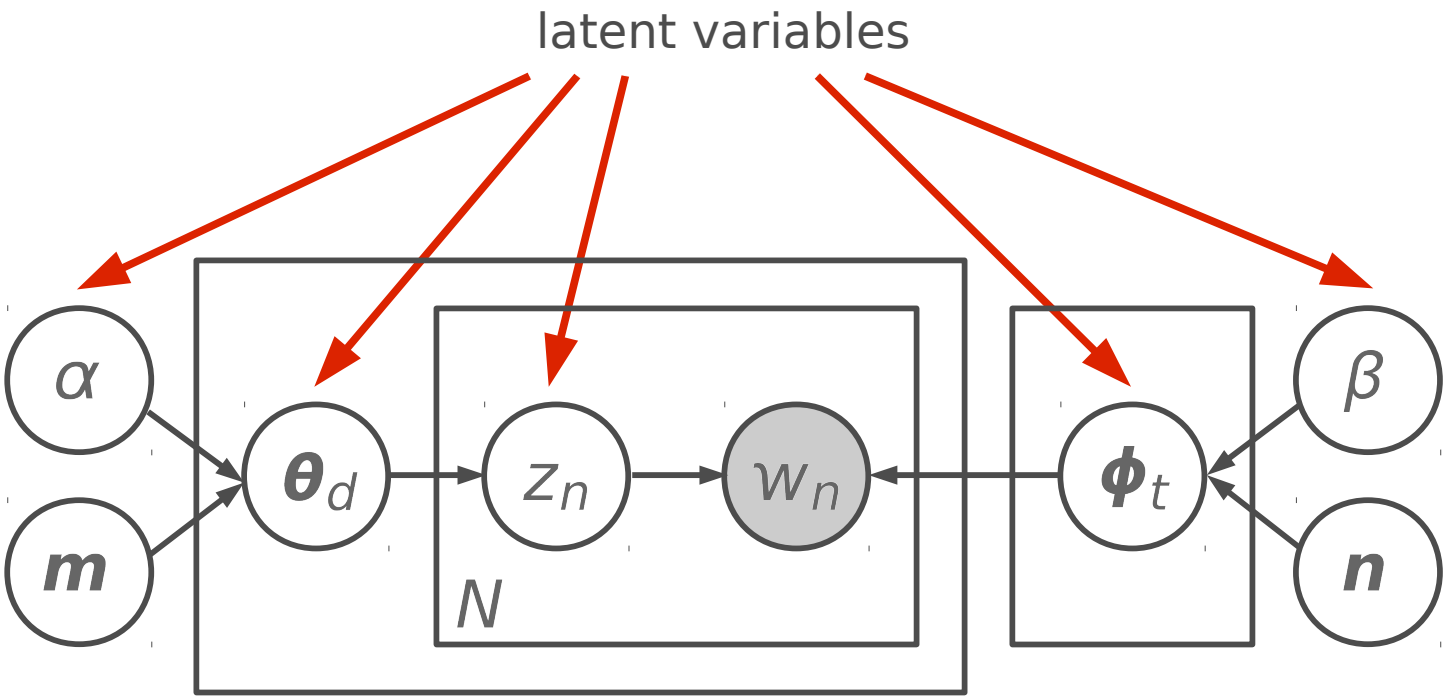
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

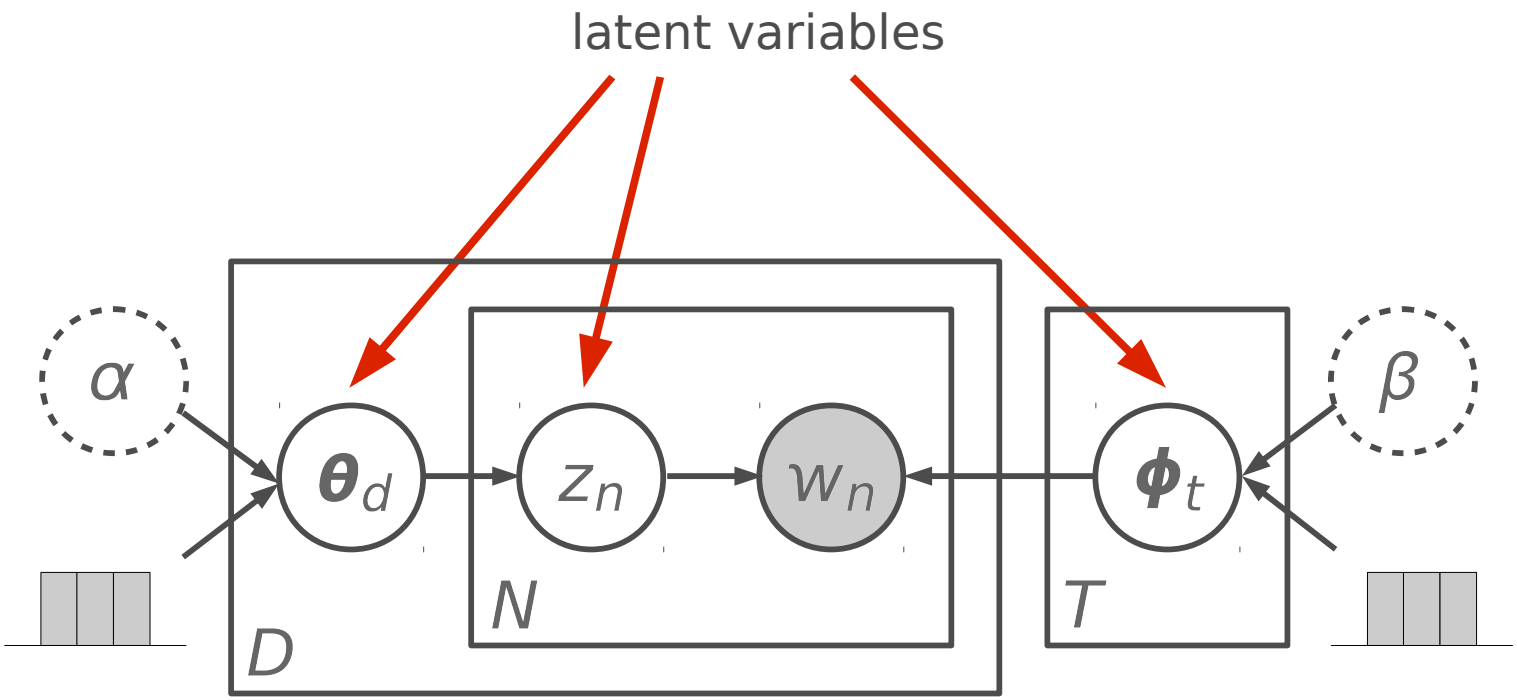
?

Posterior Inference



- Infer or integrate out all latent variables, given tokens

Posterior Inference

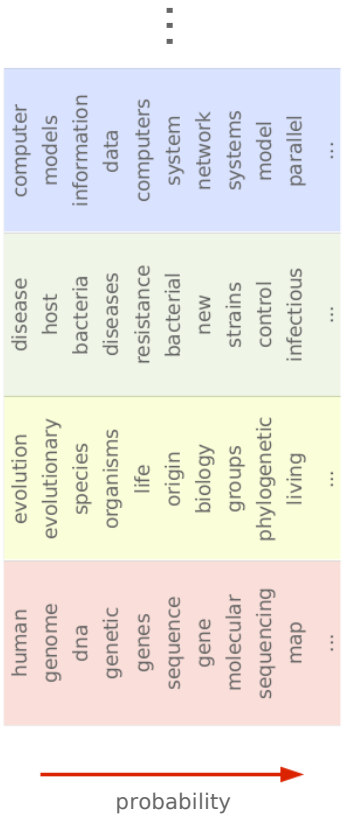


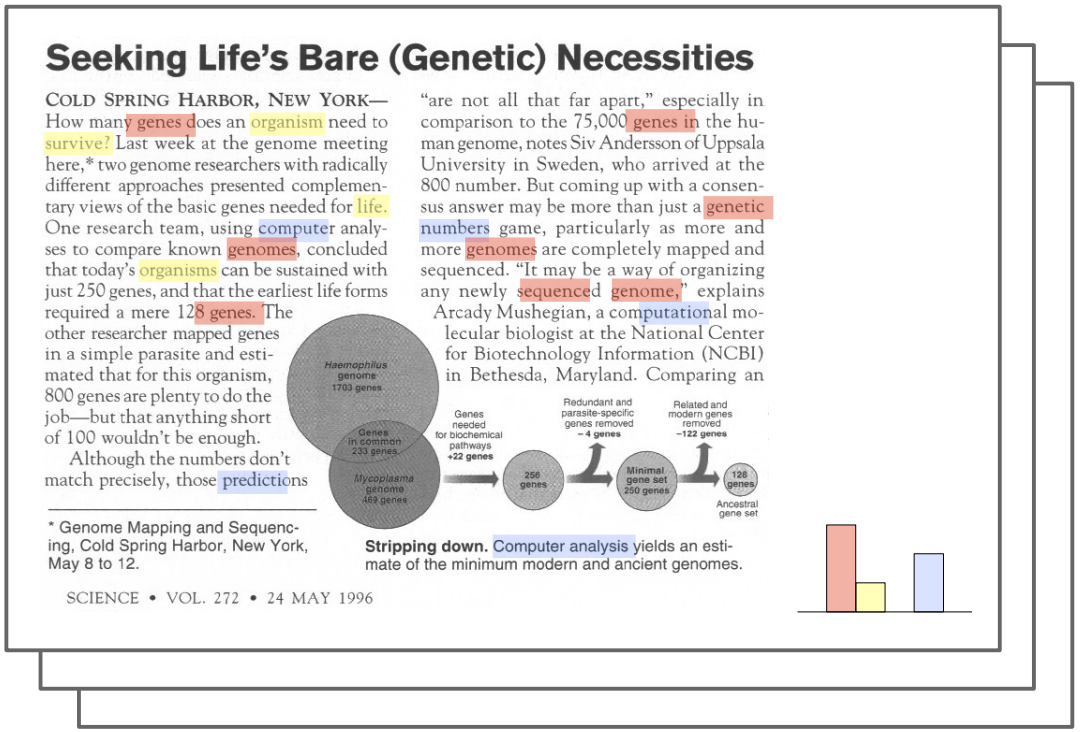
$$P(\Theta, \Phi, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta)$$

Inference Algorithms

- Exact inference in LDA is not tractable
- Approximate inference algorithms:
  - Mean field variational inference (Blei et al., 2001; 2003)
  - Expectation propagation (Minka & Lafferty, 2002)
  - Collapsed Gibbs sampling (Griffiths & Steyvers, 2002)
  - Collapsed variational inference (Teh et al., 2006)
- Each method has advantages and disadvantages

The End Result...





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

Evaluating LDA

- Unsupervised nature of LDA makes evaluation hard
- Compute probability of held-out documents:
  - Classic way of evaluating generative models
  - Often used to evaluate topic models
- Problem: have to approximate an intractable sum

$$P(\boldsymbol{w} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u}) =$$
$$\sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u})$$

Computing Log Probability

(Wallach et al., 2009)

- Simple importance sampling methods
- The “harmonic mean” method (Newton & Raftery, 1994)
  - Known to overestimate, used anyway
- Annealed importance sampling (Neal, 2001)
  - Prohibitively slow for large collections of documents
- Chib-style method (Murray & Salakhutdinov, 2009)
- “Left-to-Right” method (Wallach, 2008)

Thanks!

wallach@cs.umass.edu
http://www.cs.umass.edu/~wallach/