University of Colorado **Boulder**

Department of Computer Science
CSCI 5622: Machine Learning
Chris Ketelsen

# Lecture 24:

# Introduction to Topic Modeling

# Why Topic Models?

Suppose you want to learn things about a huge number of documents

- What's trending on Twitter?
- What fields are heavily features in Science?
- What types of proposals are being funded by NSF?

Why is this useful?

# Why Topic Models?

Suppose you want to learn things about a huge number of documents

- What's trending on Twitter?
- What fields are heavily features in Science?
- What types of proposals are being funded by NSF?

Why is this useful?

- Document Retrieval
- Feature Engineering for Document Classification
- Document clustering for Recommendation

# What is a Topic?

- Topic Modeling is a form of Unsupervised Learning
- Topics are a form of a Latent variable

So how do we represent / identify a topic?

# What is a Topic?

Topics are represented as groups of related terms

- Topics from Enron Emails

| Topic | Terms |
|---|---|
| 3 | trading financial trade product price |
| 6 | gas capacity deal pipeline contract |
| 9 | state california davis power utilities |
| 14 | ferc issue order party case |
| 22 | group meeting team process plan |

[Boyd-Graber]

# What is a Topic?

Topics are represented as groups of related terms

- Science Magazine

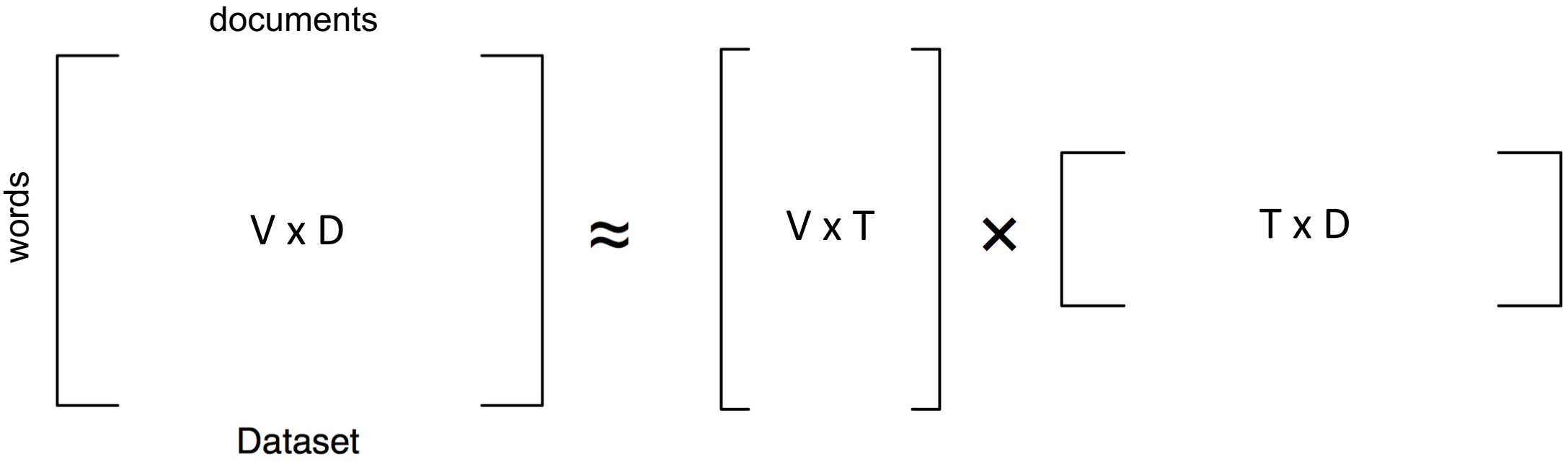
| | | | |
|---|---|---|---|
| human | evolution | disease | computer |
| genome | evolutionary | host | models |
| dna | species | bacteria | information |
| genetic | organisms | diseases | data |
| genes | life | resistance | computers |
| sequence | origin | bacterial | system |
| gene | biology | 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 |

[Blei]

# How Do We Find Topics in Data?

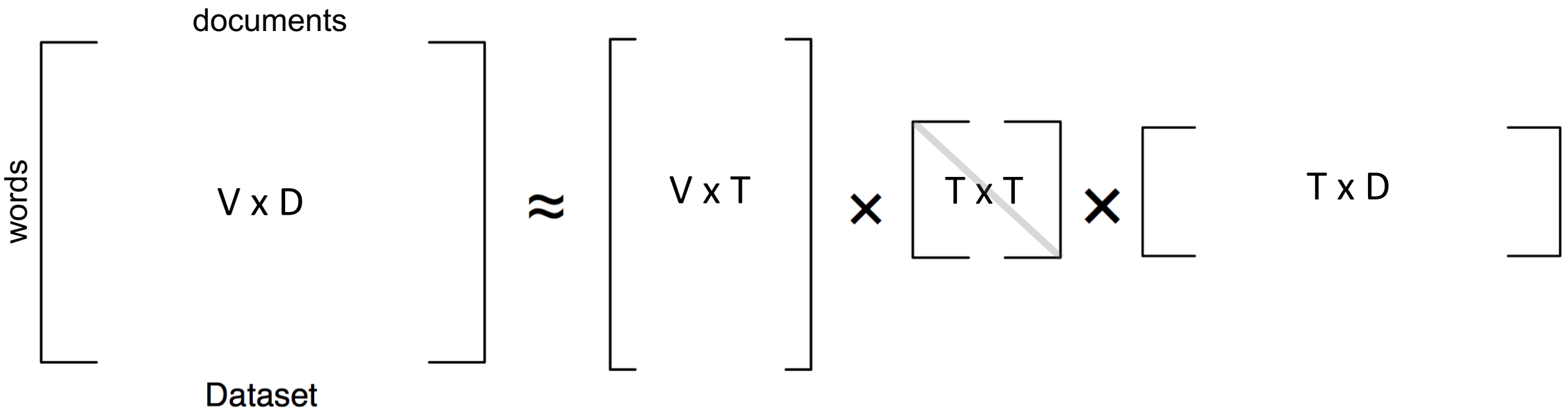
**Matrix Factorization Techniques**: Factor the Term-Document Matrix

- Suppose we have a large corpus of D documents with V distinct terms



# Latent Semantic Analysis

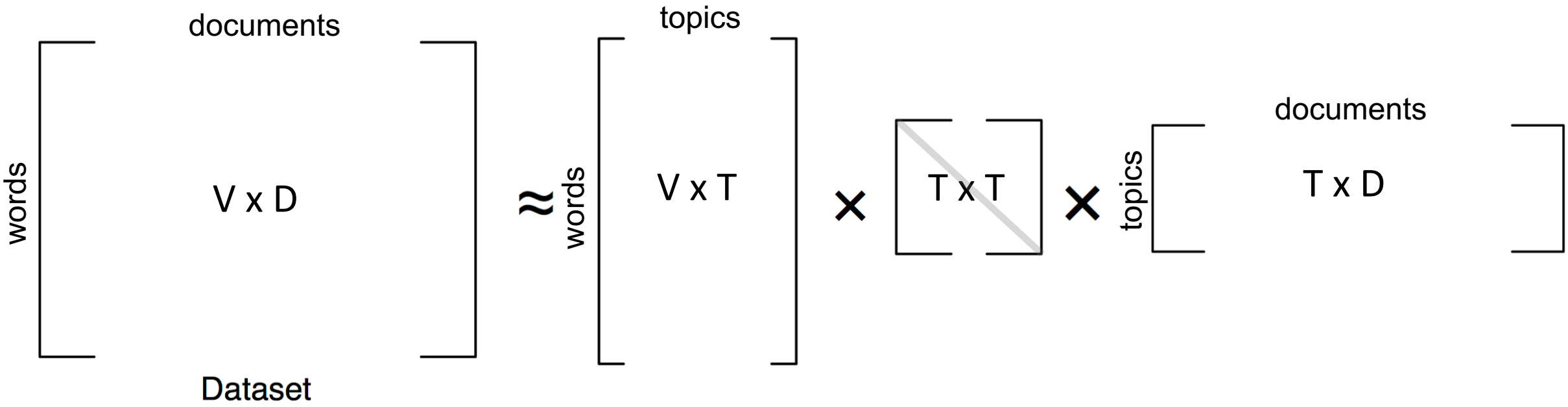
- Perform Singular Value Decomposition (SVD) on Term-Document Matrix



# Singular Value Decomposition

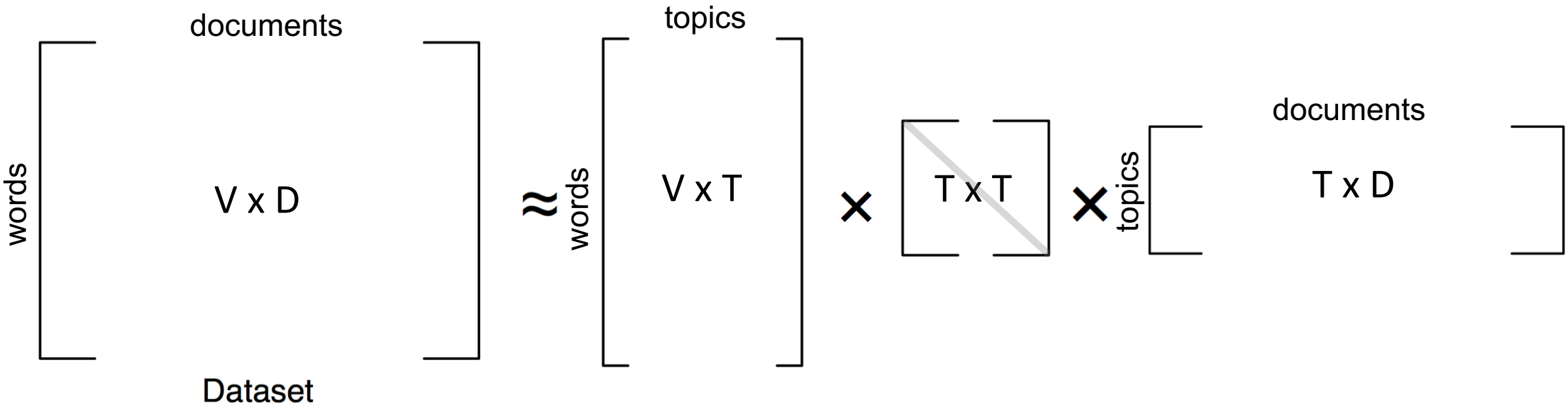
# Latent Semantic Analysis

- Use reduced basis to perform document similarity



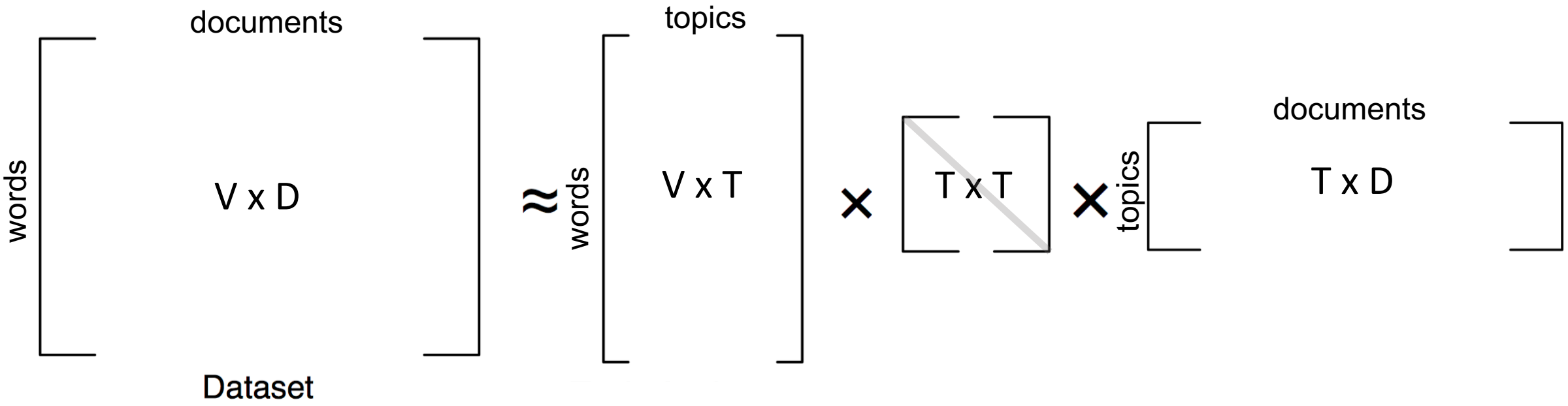
# Latent Semantic Analysis

- Use reduced basis to perform word similarity
- Does a good job of capturing synonymy but poor job of capturing polysymy



# Latent Semantic Analysis

- Use coordinates in reduce basis to cluster documents or words




- **Historical Note**: Pioneered by Thomas Landauer in CU Psychology Dept.

# Probabilistic Topic Modeling

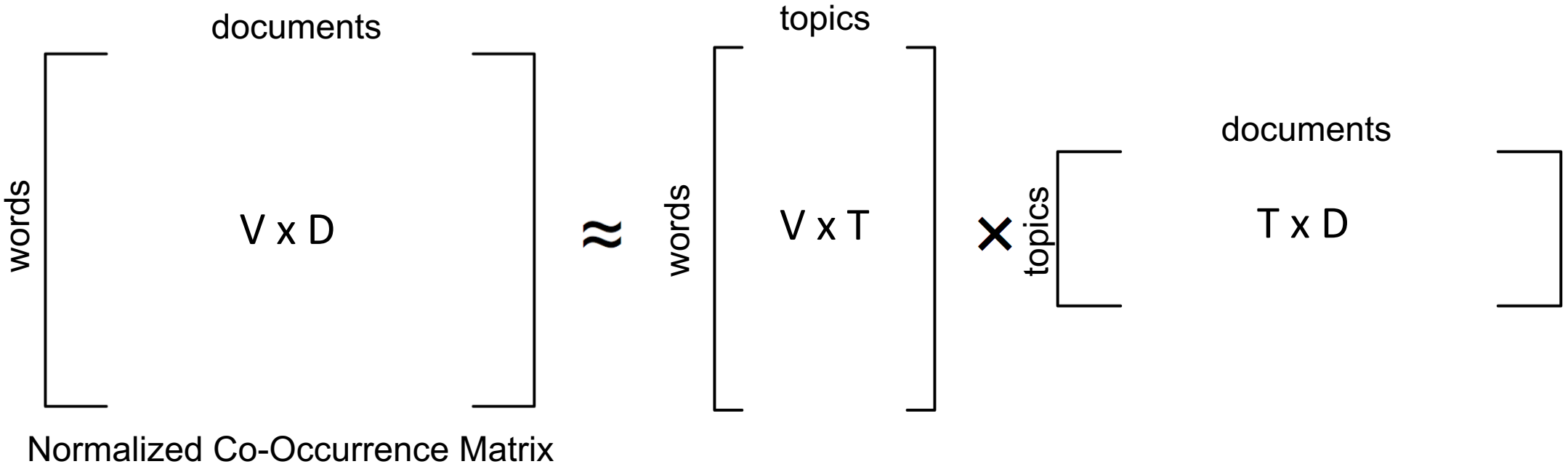
Assume that there is some underlying distribution of terms in a document

$$p(w \mid d) = \sum_{t=1}^{T} p(w, t \mid d) = \sum_{t=1}^{T} p(w \mid t) p(t \mid d)$$

# Probabilistic Topic Modeling

Assume that there is some underlying distribution of terms in a document

$$p(w \mid d) = \sum_{t=1}^{T} p(w, t \mid d) = \sum_{t=1}^{T} p(w \mid t) p(t \mid d)$$



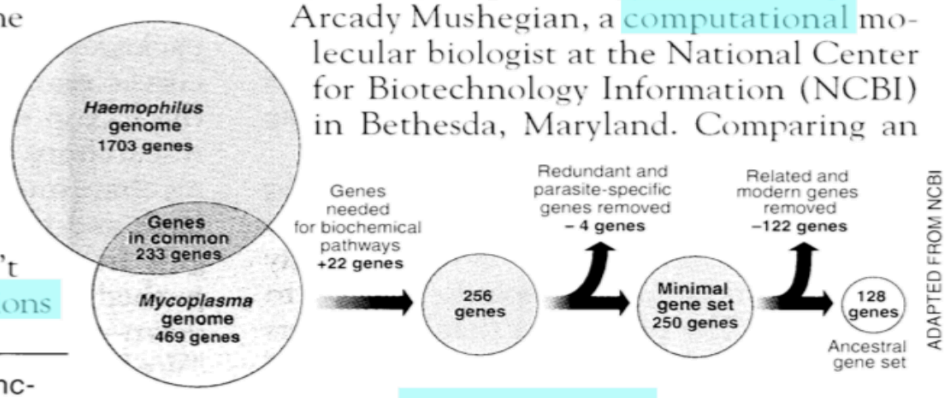

Normalized Co-Occurrence Matrix

# A Generative Story

## 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.

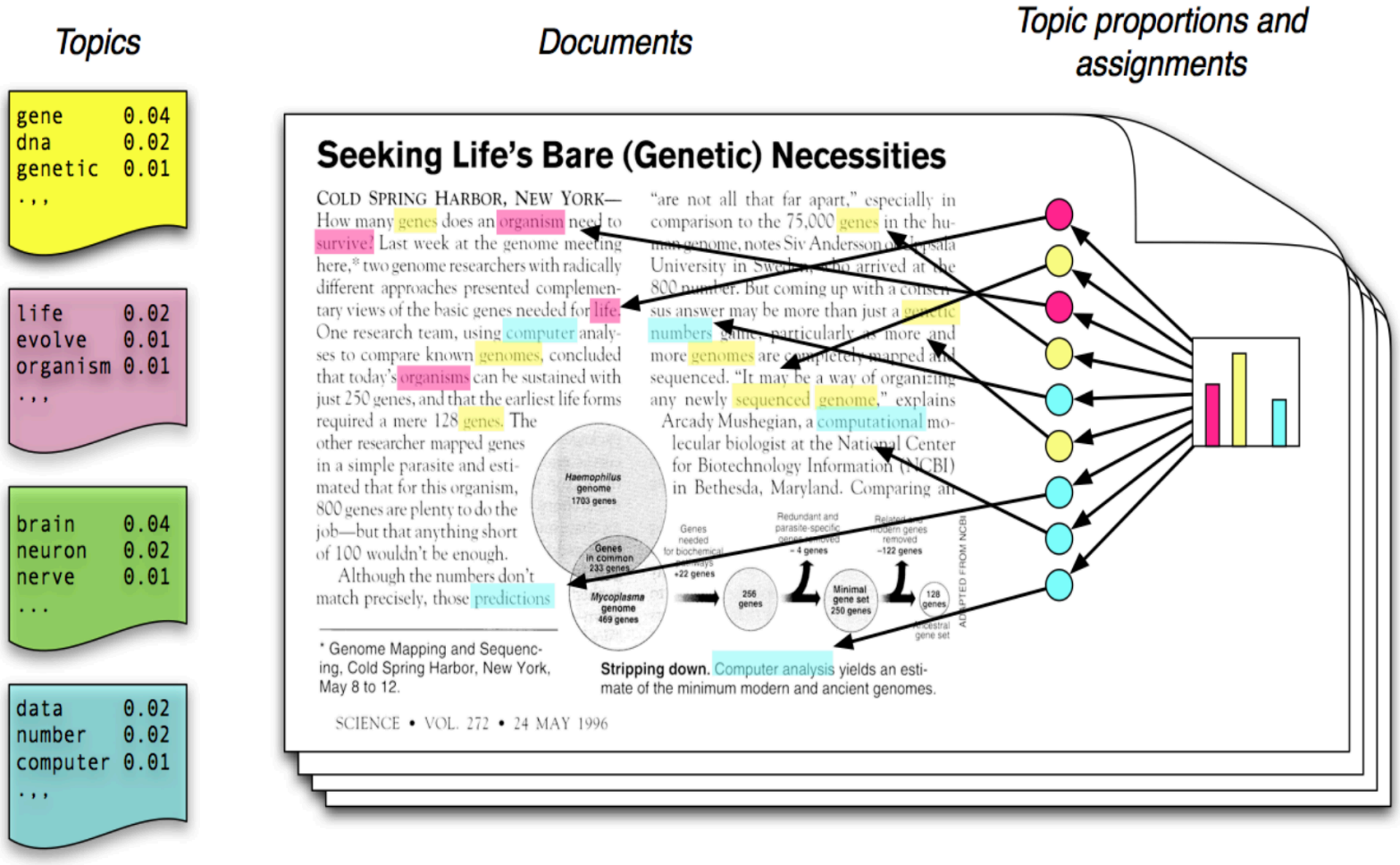
**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Documents exhibit multiple topics

[Blei]

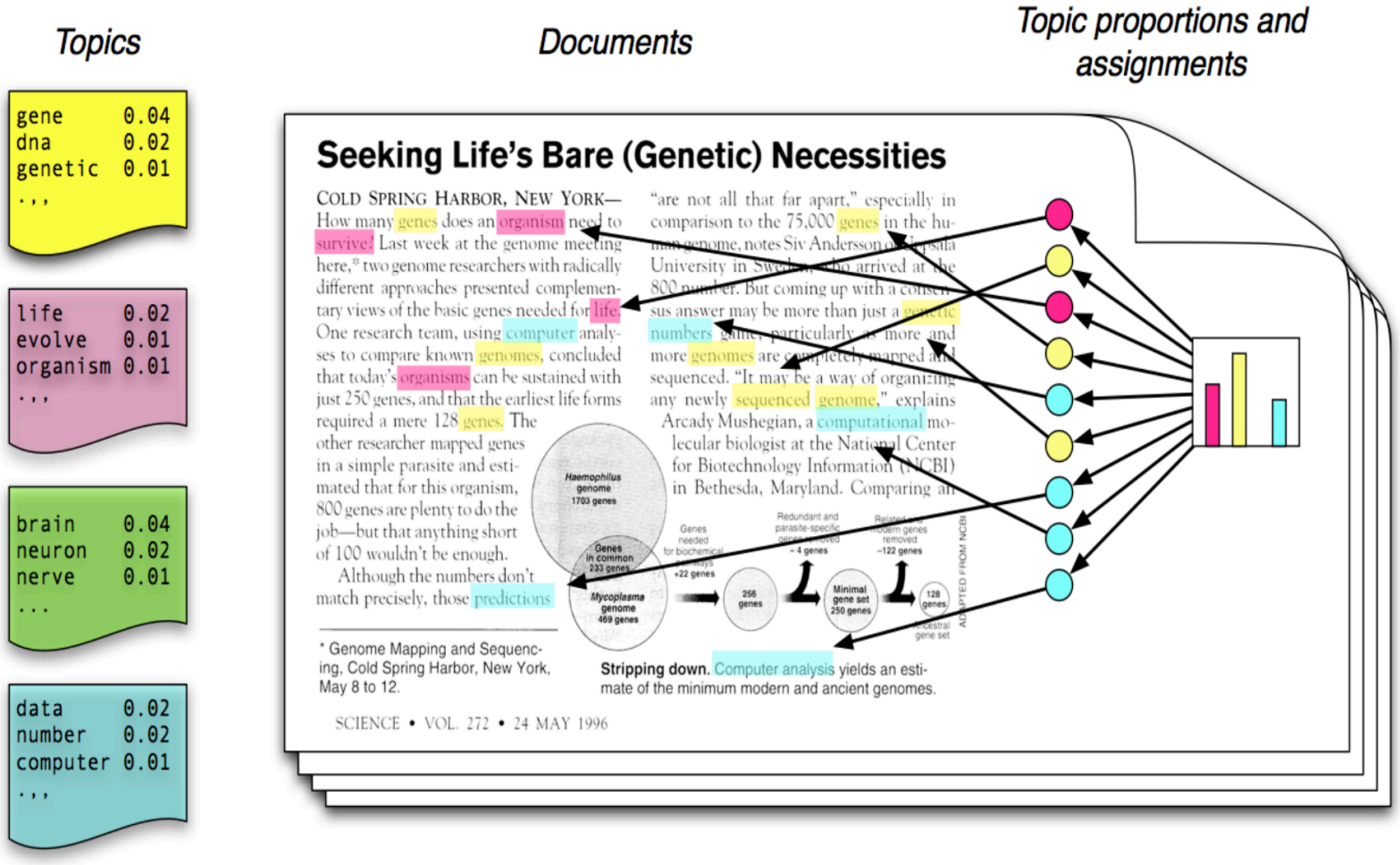
# A Generative Story




- Each document is a random mixture of topics

[Blei]

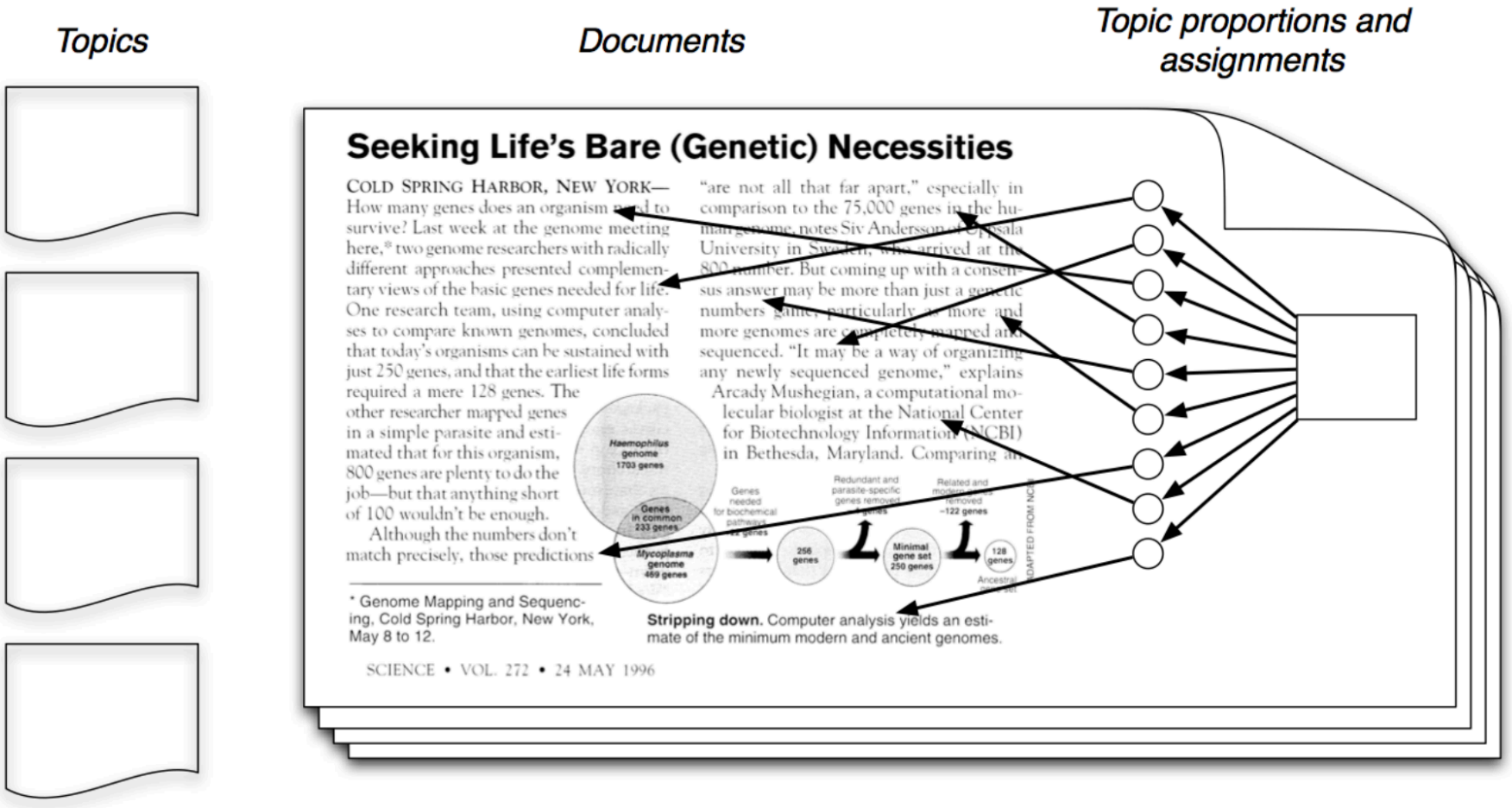
# A Generative Story




- Each word is drawn from one of those topics

[Blei]

# The Inference Problem




- We only have the data. We have to estimate these distributions [Blei]

# Latent Dirichlet Allocation

What are the distributions that we need to estimate from the data?

- **For each topic k, need a distribution on words**: $\phi_t \sim \mathrm{Dir}(\lambda)$

This is typically represented by a Categorical distribution

# Latent Dirichlet Allocation

What are the distributions that we need to estimate from the data?

- **For each topic k, need a distribution on words**: $\phi_t \sim \mathrm{Dir}(\lambda)$

- **For each document, need a distribution over topics**: $\theta_d \sim \mathrm{Dir}(\alpha)$

This is also typically defined as a Categorical distribution

# Latent Dirichlet Allocation

What are the distributions that we need to estimate from the data?

- **For each topic k, need a distribution on words**: $\phi_t \sim \mathrm{Dir}(\lambda)$
- **For each document, need a distribution over topics**: $\theta_d \sim \mathrm{Dir}(\alpha)$
- **For each word in each document, first generate topic**: $z_{n,d} \sim \mathrm{Cat}(\theta_d)$

# Latent Dirichlet Allocation

What are the distributions that we need to estimate from the data?

- **For each topic k, need a distribution on words**: $\phi_t \sim \mathrm{Dir}(\lambda)$
- **For each document, need a distribution over topics**: $\theta_d \sim \mathrm{Dir}(\alpha)$
- **For each word in each document, first generate topic**: $z_{n,d} \sim \mathrm{Cat}(\theta_d)$
- **Given the word's topic, generate the actual word**: $w_{n,d} \sim \mathrm{Cat}(\phi_{z_{n,d}})$

# Latent Dirichlet Allocation

What are the distributions that we need to estimate from the data?

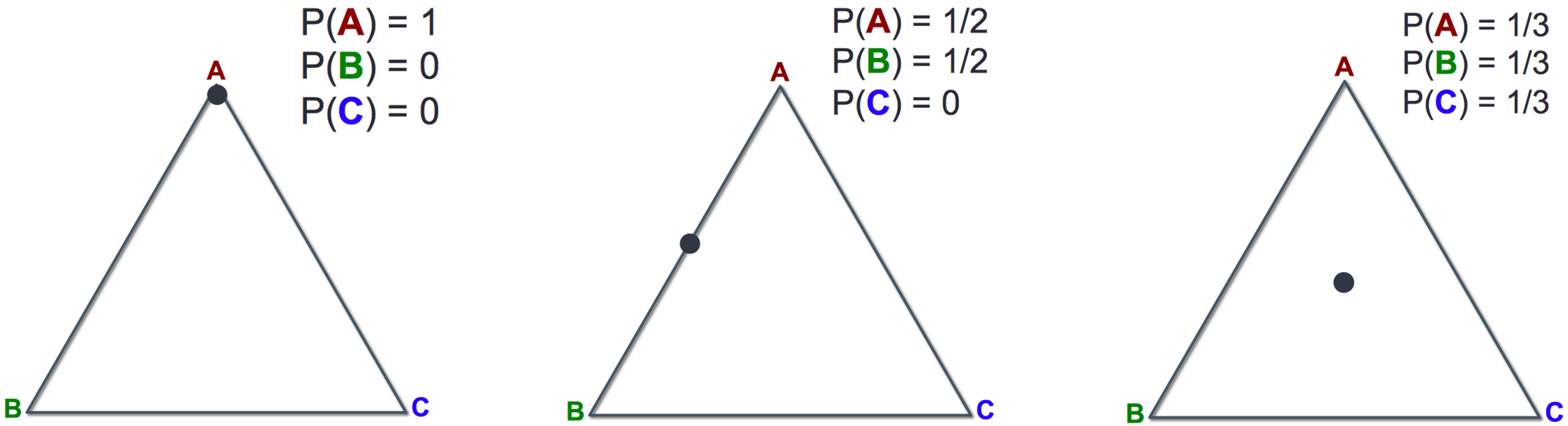
- **For each topic k, need a distribution on words**: $\phi_t \sim \mathrm{Dir}(\lambda)$
- **For each document, need a distribution over topics**: $\theta_d \sim \mathrm{Dir}(\alpha)$
- **For each word in each document, first generate topic**: $z_{n,d} \sim \mathrm{Cat}(\theta_d)$
- **Given the word's topic, generate the actual word**: $w_{n,d} \sim \mathrm{Cat}(\phi_{z_{n,d}})$

Does something seem weird about this?

# The Dirichlet Distribution

$\phi_k \sim \mathrm{Dir}(\lambda)$ and $\theta_d \sim \mathrm{Dir}(\alpha)$ are distributions sampled from a distribution

Consider the case where we have 3 topics

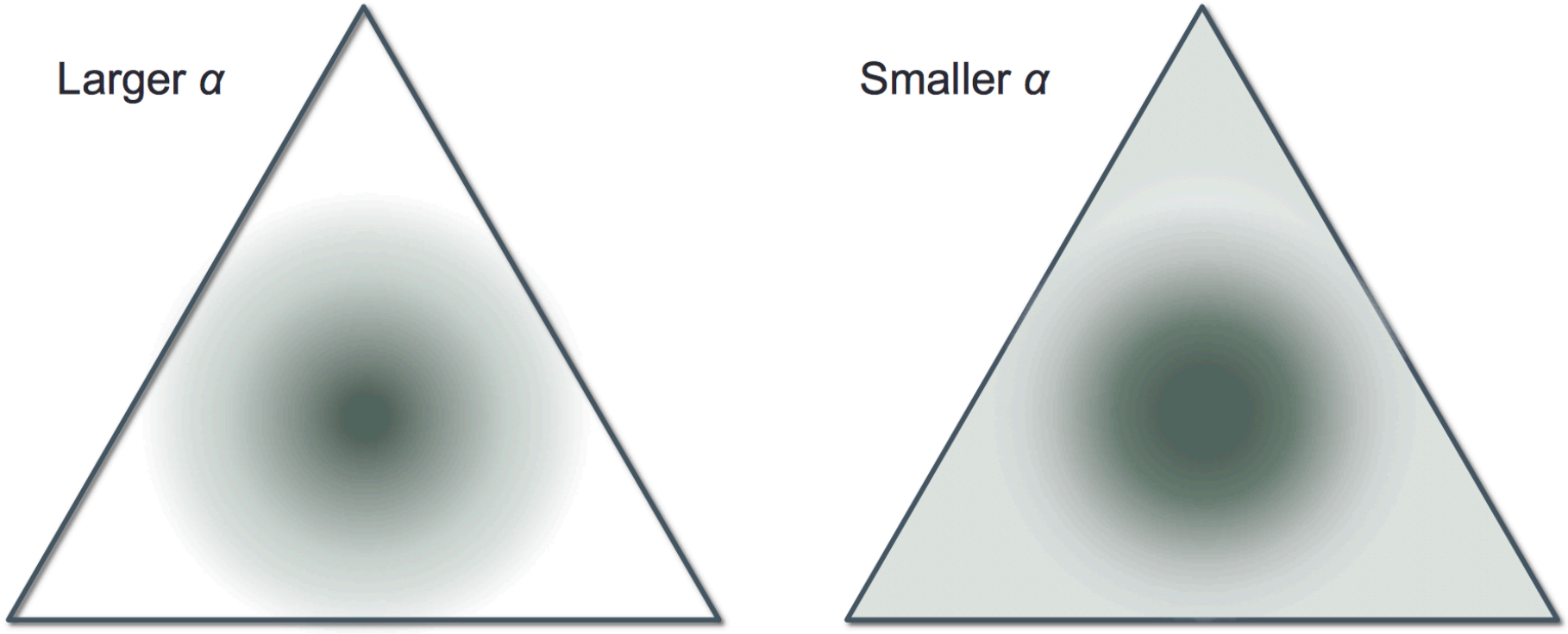



[Paul]

# The Dirichlet Distribution

$\phi_k \sim \mathrm{Dir}(\lambda)$ and $\theta_d \sim \mathrm{Dir}(\alpha)$ are distributions sampled from a distribution

A draw from a Dirichlet distribution returns a Categorical distribution




[Paul]

# Latent Dirichlet Allocation

So we need to estimate the parameters of these distributions from the data

$\phi_t \sim \mathrm{Dir}(\lambda)$

$\theta_d \sim \mathrm{Dir}(\alpha)$

If we're interested in the topics, then we're most interested in $\phi_t$


[Blei]

# Latent Dirichlet Allocation

So we need to estimate the parameters of these distributions from the data

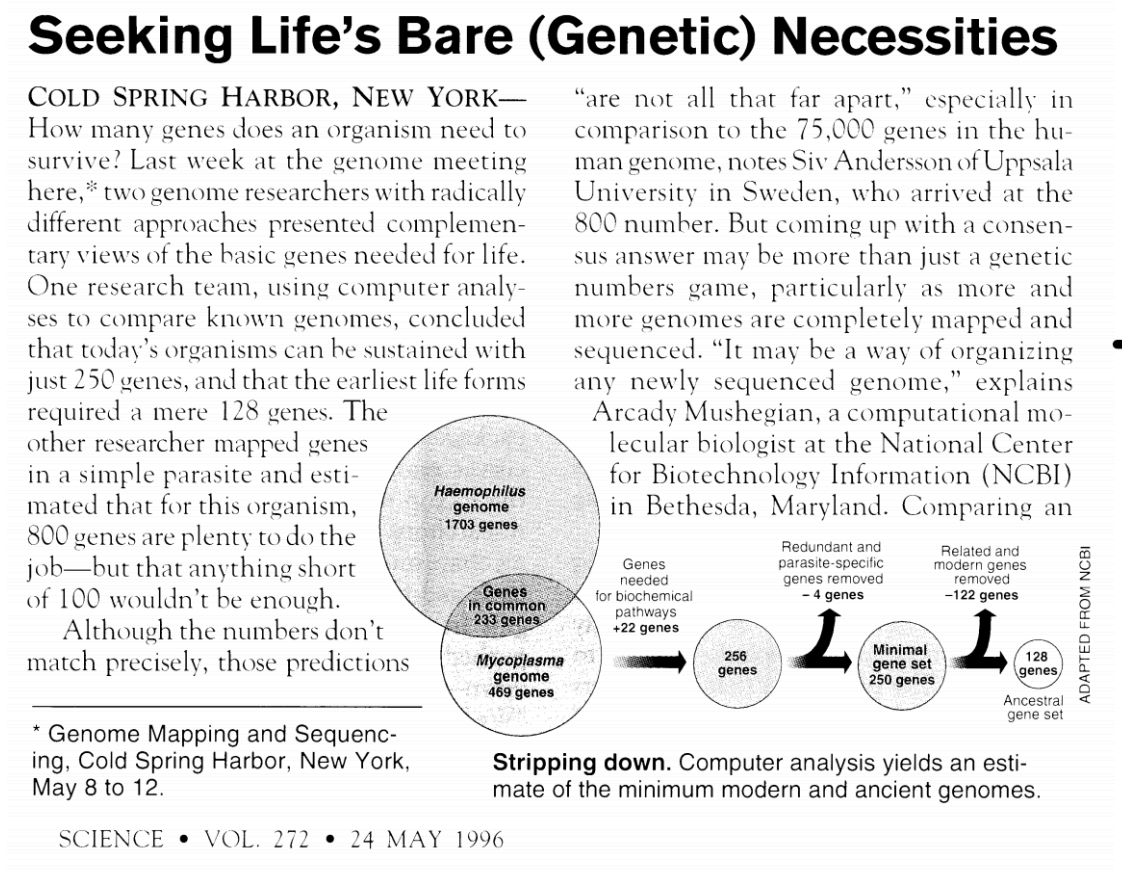
$\phi_t \sim \mathrm{Dir}(\lambda)$

$\theta_d \sim \mathrm{Dir}(\alpha)$

If we want to use the topics as latent variables to do other things, then $\theta_d$

# Latent Dirichlet Allocation

If we want to use the topics as latent variables to do other things, then $\theta_d$



## 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.

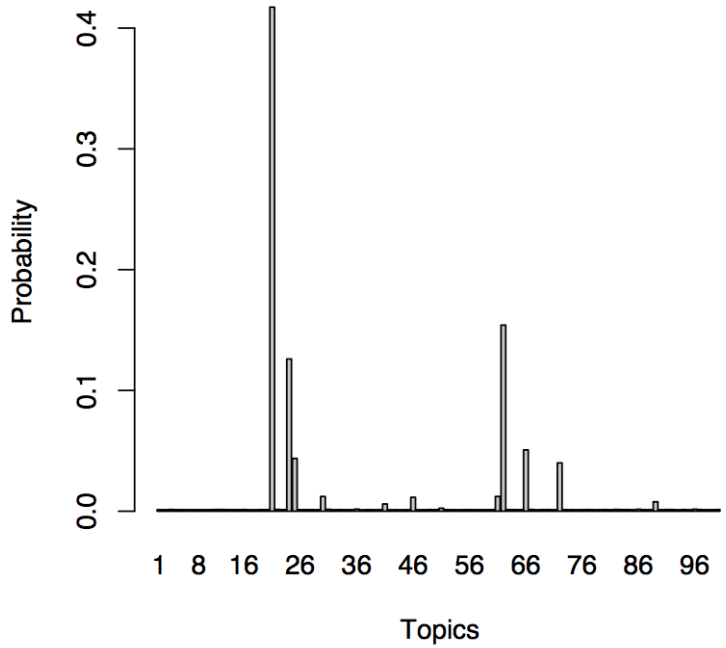
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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




[Blei]

# Latent Dirichlet Allocation

If we want to use the topics as latent variables to do other things, then $\theta_d$

Think of $\theta_d$ as a low-dimensional fingerprint of a document

- Use $\theta_d$ as features in document classification
- Use $\theta_d$ as a low-dimensional feature-embedding and perform clustering

# Latent Dirichlet Allocation

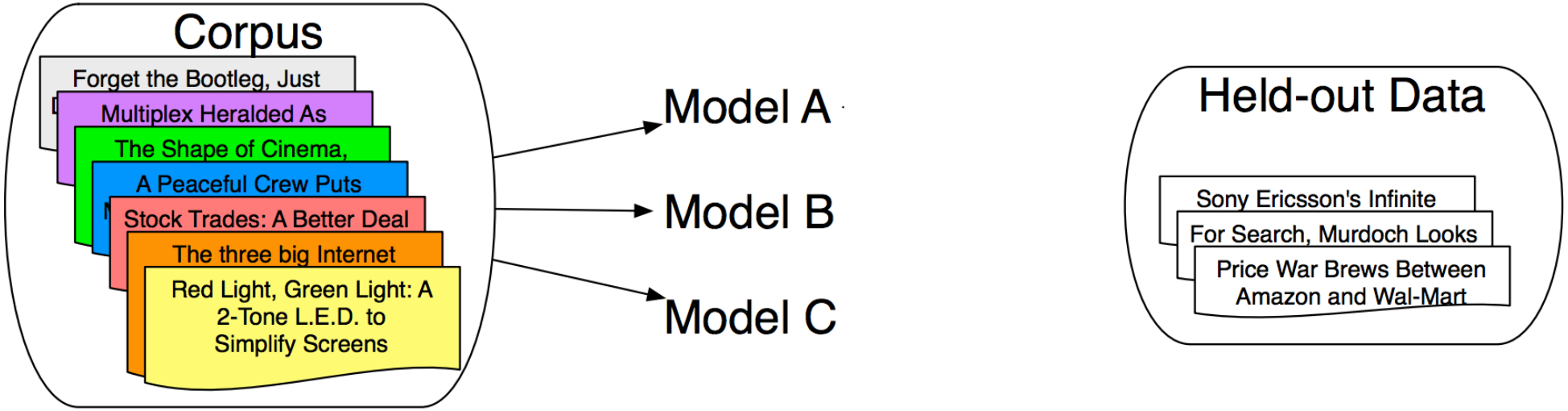
OK, so how do we actually learn these parameters?

Lots of methods:

- Expectation-Maximization
- Collapsed Gibbs Sampling
- Variational Inference

# Evaluation of Topic Models

How do we know if our model is good?

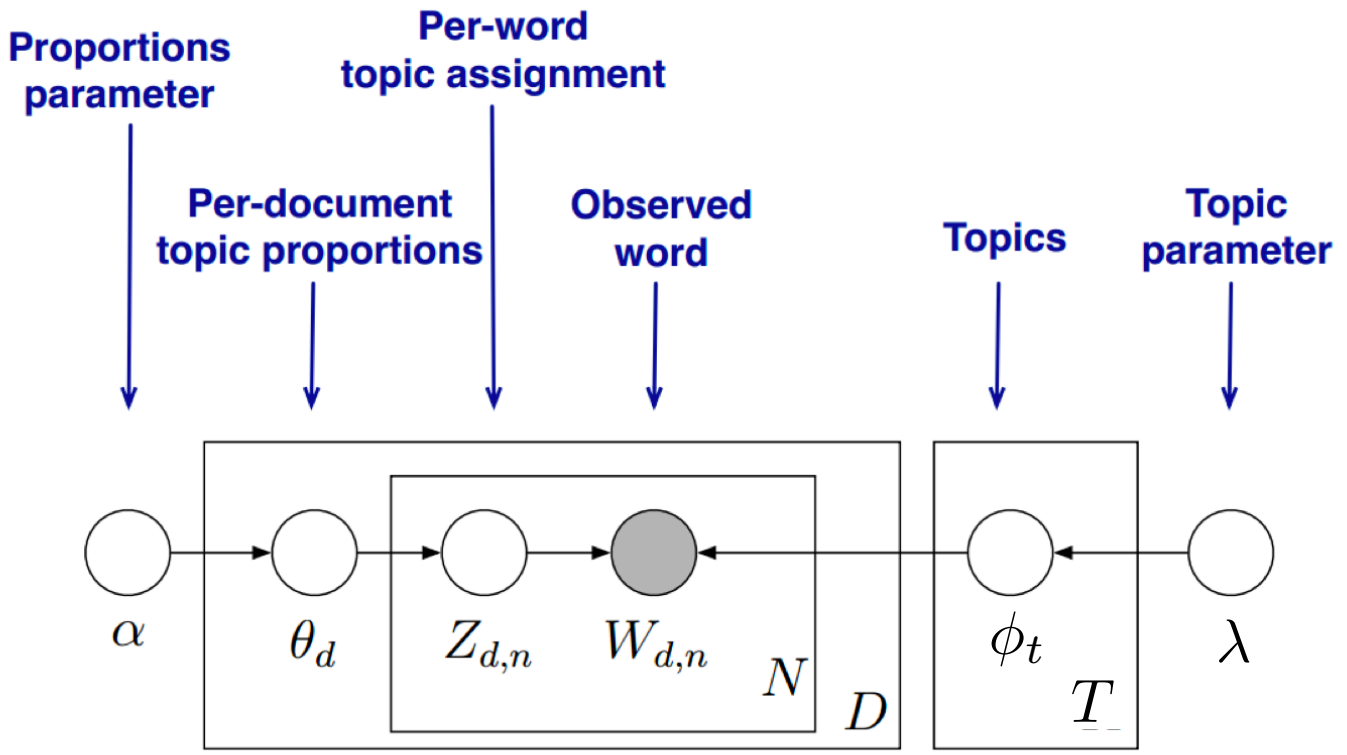



With LDA you can evaluate various models on a held-out data set

Performance metric is likelihood induced by the probabilistic model

[Boyd-Graber]

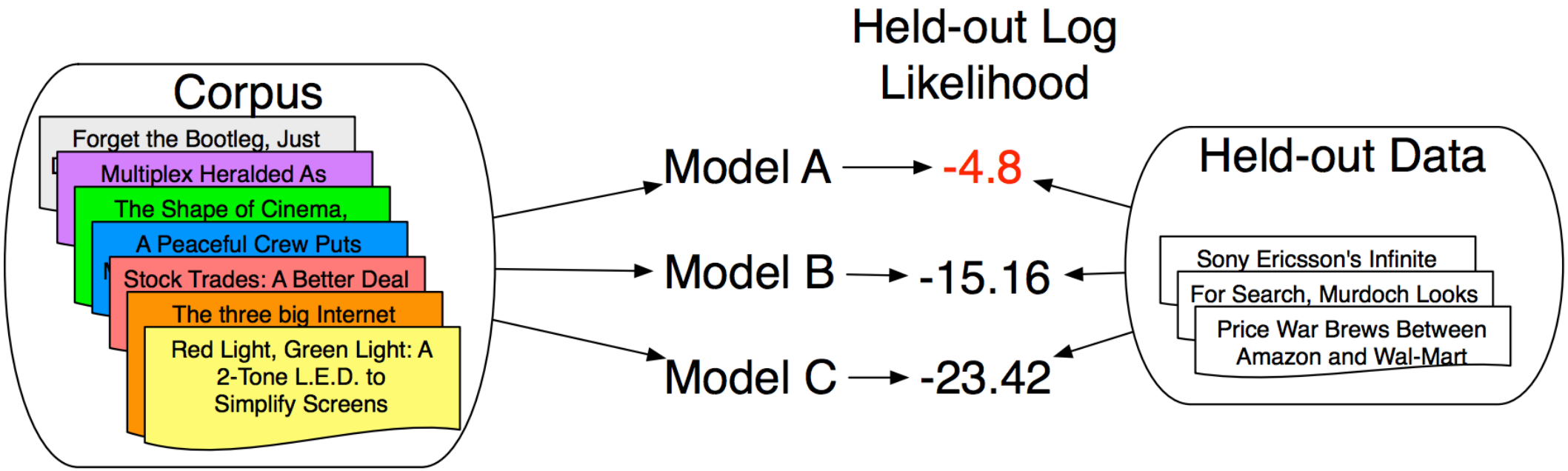
# Evaluation of Topic Models




$$\prod_{t=1}^{T} p_\lambda(\phi_t) \prod_{d=1}^{D} p_\alpha(\theta_d) \prod_{n=1}^{N_d} p(Z_{d,n} \mid \theta_d) p(W_{d,n} \mid \phi_{1:T}, Z_{d,n})$$

[Blei]

# Evaluation of Topic Models




With LDA you can evaluate various models on a held-out data set

Performance metric is likelihood induced by the probabilistic model

[Boyd-Graber]

# Evaluation of Topic Models

If we're more interested in understanding, use a more human-centric measure

**Word Intrusion**:

- Take the highest probability words from a topic
- Add a high-probability word from another topic
- Ask users to pick word that doesn't belong

| Original Topic |
| --- |
| dog, cat, horse, pig, cow |

| Topic with Intruder |
| --- |
| dog, cat, apple, horse, pig, cow |

**Hypothesis**: If topics are interpretable, user will usually pick the intruder

[Boyd-Graber]

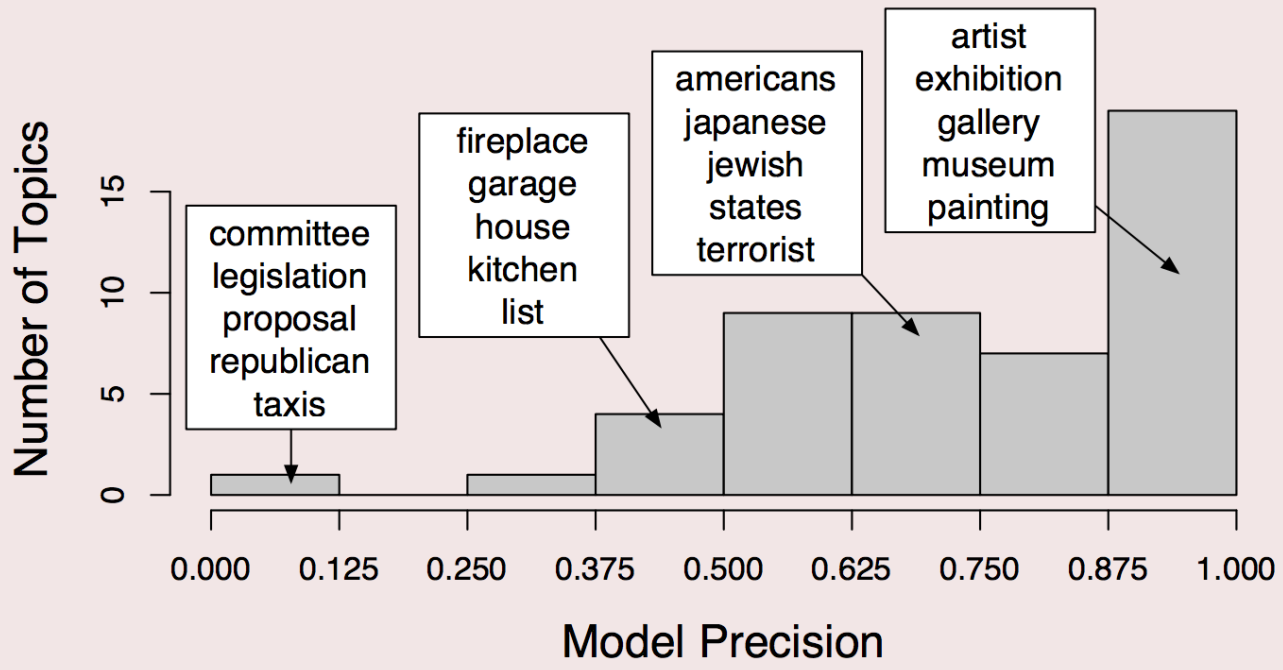
# Word Intrusion

| | | | | | |
|---|---|---|---|---|---|
| **1 / 10** | | | | | |
| crash | accident | board | agency | tibetan | safety |
| **2 / 10** | | | | | |
| commercial | network | television | advertising | viewer | layoff |
| **3 / 10** | | | | | |
| arrest | crime | inmate | pitcher | prison | death |
| **4 / 10** | | | | | |
| hospital | doctor | health | care | medical | tradition |

[Boyd-Graber]

# Word Intrusion

## New York Times, 50 LDA Topics




[Boyd-Graber]

# Evaluation of Topic Models

**Take-ways**:

- Measure what you care about
- If you care about prediction, likelihood is usually good
- If you care about interpretability, maybe word intrusion is better
- If you care about a particular task, measure that

[Boyd-Graber]

# Acknowledgments

Many images and entire slides were borrowed from:

- David Blei
- Jordan Boyd-Graber
- Michael Paul