Machine Learning 2

DS 4420 - Spring 2020

Topic Modeling 1

Byron C. Wallace

Last time:

Clustering —> Mixture Models —> Expectation Maximization (EM)

Today:

Topic models

Mixture models

Data: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^M$

Slide credit: ***Matt Gormley and Eric Xing (CMU)***

Mixture models

Data: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^M$

Generative Story: $z \sim \text{Multinomial}(\boldsymbol{\phi})$

$\mathbf{x} \sim p_{\boldsymbol{\theta}}(\cdot|z)$

Slide credit: ***Matt Gormley and Eric Xing (CMU)***

Mixture models

Data: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^M$

Generative Story: $z \sim \text{Multinomial}(\boldsymbol{\phi})$

$\mathbf{x} \sim p_{\boldsymbol{\theta}}(\cdot|z)$

Model:

Joint: $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}, z) = p_{\boldsymbol{\theta}}(\mathbf{x}|z)p_{\boldsymbol{\phi}}(z)$

Marginal: $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \sum_{z=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}|z)p_{\boldsymbol{\phi}}(z)$

Slide credit: ***Matt Gormley and Eric Xing (CMU)***

Mixture models

Data: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^M$

Generative Story:

$$z \sim \text{Multinomial}(\boldsymbol{\phi})$$

$$\mathbf{x} \sim p_{\boldsymbol{\theta}}(\cdot|z)$$

Model:

Joint: $$p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}, z) = p_{\boldsymbol{\theta}}(\mathbf{x}|z)p_{\boldsymbol{\phi}}(z)$$

Marginal: $$p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \sum_{z=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}|z)p_{\boldsymbol{\phi}}(z)$$

(Marginal) Log-likelihood:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log \prod_{i=1}^{N} p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}^{(i)}) \\ &= \sum_{i=1}^{N} \log \sum_{z=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|z)p_{\boldsymbol{\phi}}(z) \end{aligned}$$

Slide credit: ***Matt Gormley and Eric Xing (CMU)***

Naive Bayes

The model

$$p(c|w_{1:N}, \pi, \theta) \propto p(c|\pi) \prod_{n=1}^{N} p(w_n|\theta_c)$$

$$p(\mathcal{D}|\theta_{1:C}, \pi) \quad = \quad \prod_{d=1}^{D} \left(p(c_d|\pi) \prod_{n=1}^{N} p(w_n|\theta_{c_d}) \right)$$

(Soft) EM

Initialize **parameters** **randomly**

while not converged

1. **E-Step:**
   Create one training example for each possible value of the **latent variables**
   Weight each example according to model's confidence
   Treat parameters as observed
2. **M-Step:**
   Set the **parameters** to the values that maximizes likelihood
   Treat pseudo-counts from above as observed

Slide credit: ***Matt Gormley and Eric Xing (CMU)***

And for NB

For "soft" EM

expected # of times
t occurs in c

$$P(t \mid c) = \frac{\sum_i P(z_i = c) \cdot \text{count}(t \text{ in } x_i)}{\sum_i P(z_i = c) \cdot |x_i|}$$

total token count in x_i

TOPIC MODELS

Some content borrowed from:
David Blei
(Columbia)

Topic Models: Motivation

- Suppose we have a giant dataset ("corpus") of text, e.g., all of the NYTimes or all emails from a company
  - Cannot read all documents
  - But want to get a sense of what they contain

Topic Models: Motivation

- Topic models are a way of uncovering, well, “topics” (themes) in a set of documents

Topic Models: Motivation

- Topic models are a way of uncovering, well, "topics" (themes) in a set of documents
- Topic models are *unsupervised*

Topic Models: Motivation

- Topic models are a way of uncovering, well, "topics" (themes) in a set of documents
- Topic models are *unsupervised*
- Can be viewed as a sort of soft clustering of documents into topics.

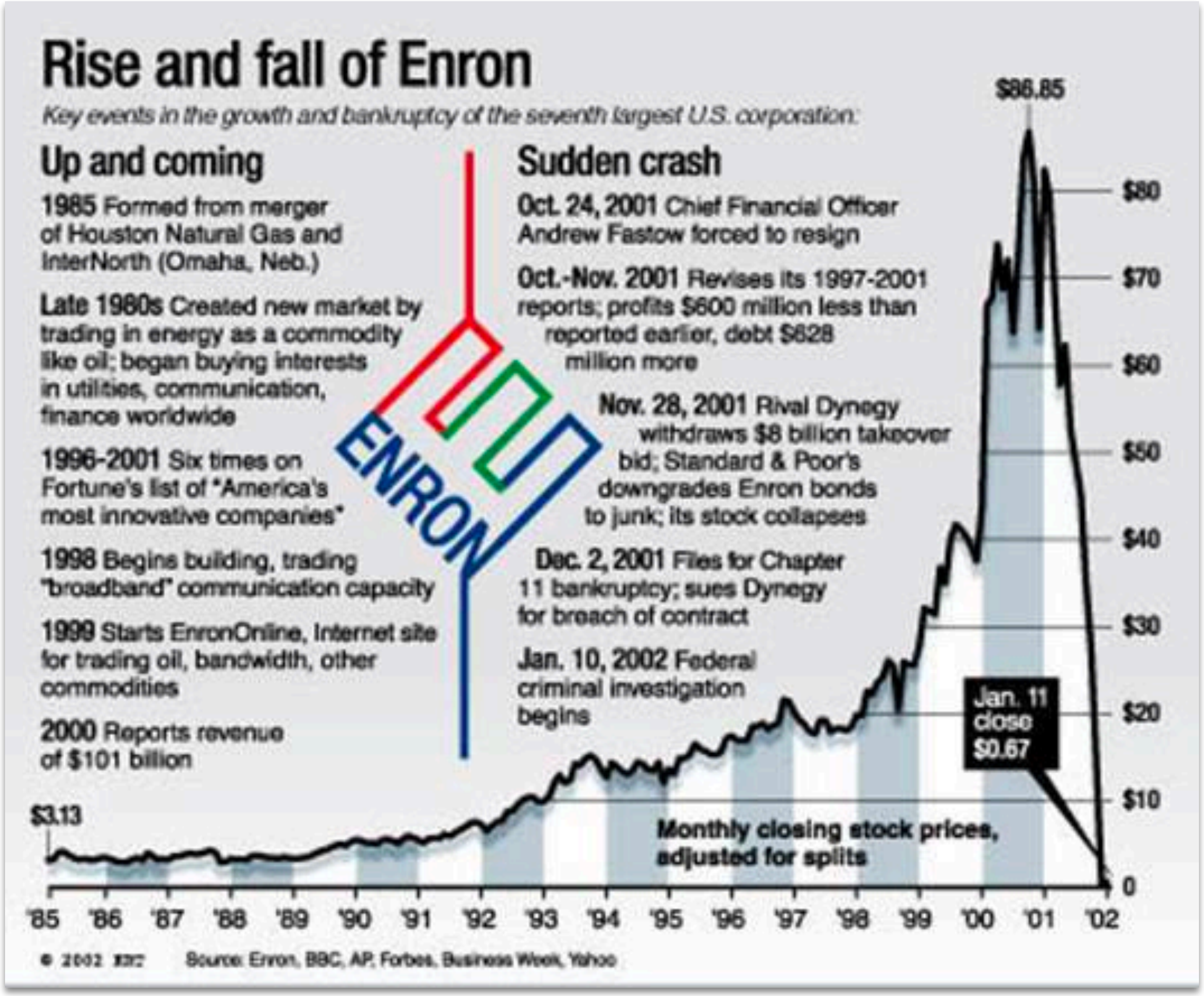
Topic 1	Topic 2	Topic 3	Topic 4
the	i	that	**easter**
"number"	is	**proteins**	**ishtar**
in	**satan**	the	a
to	the	of	the
espn	which	to	have
hockey	and	i	with
a	of	if	but
this	**metaphorical**	"number"	**english**
as	**evil**	you	and
run	there	**fact**	is

Example from Wallach, 2006

Key outputs

- ***Topics*** Distributions over words; we hope these are somehow thematically coherent
- ***Document-topics*** Probabilistic assignments of topics to documents

Example: Enron emails



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

https://www.cs.cmu.edu/~enron/

Example: 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

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

Example from Boyd-Graber, Hu and Mimno, 2017

Document-topic probabilities

Yesterday, SDG&E filed a motion for adoption of an electric procurement cost recovery mechanism and for an order shortening time for parties to file comments on the mechanism. The attached email from SDG&E contains the motion, an executive summary, and a detailed summary of their proposals and recommendations governing procurement of the net short energy requirements for SDG&E's customers. The utility requests a 15-day comment period, which means comments would have to be filed by September 10 (September 8 is a Saturday). Reply comments would be filed 10 days later.

Topic	Probability
9	0.42
11	0.05
8	0.05

Example from Boyd-Graber, Hu and Mimno, 2017

Topics as Matrix Factorization

- One can view topics as a kind of matrix factorization

$$\underbrace{\begin{bmatrix} M \times K \end{bmatrix}}_{\text{Topic Assignment}} \times \underbrace{\begin{bmatrix} K \times V \end{bmatrix}}_{\text{Topics}} \approx \underbrace{\begin{bmatrix} M \times V \end{bmatrix}}_{\text{Dataset}}$$

Figure from Boyd-Graber, Hu and Mimno, 2017

Topics as Matrix Factorization

- One can view topics as a kind of matrix factorization

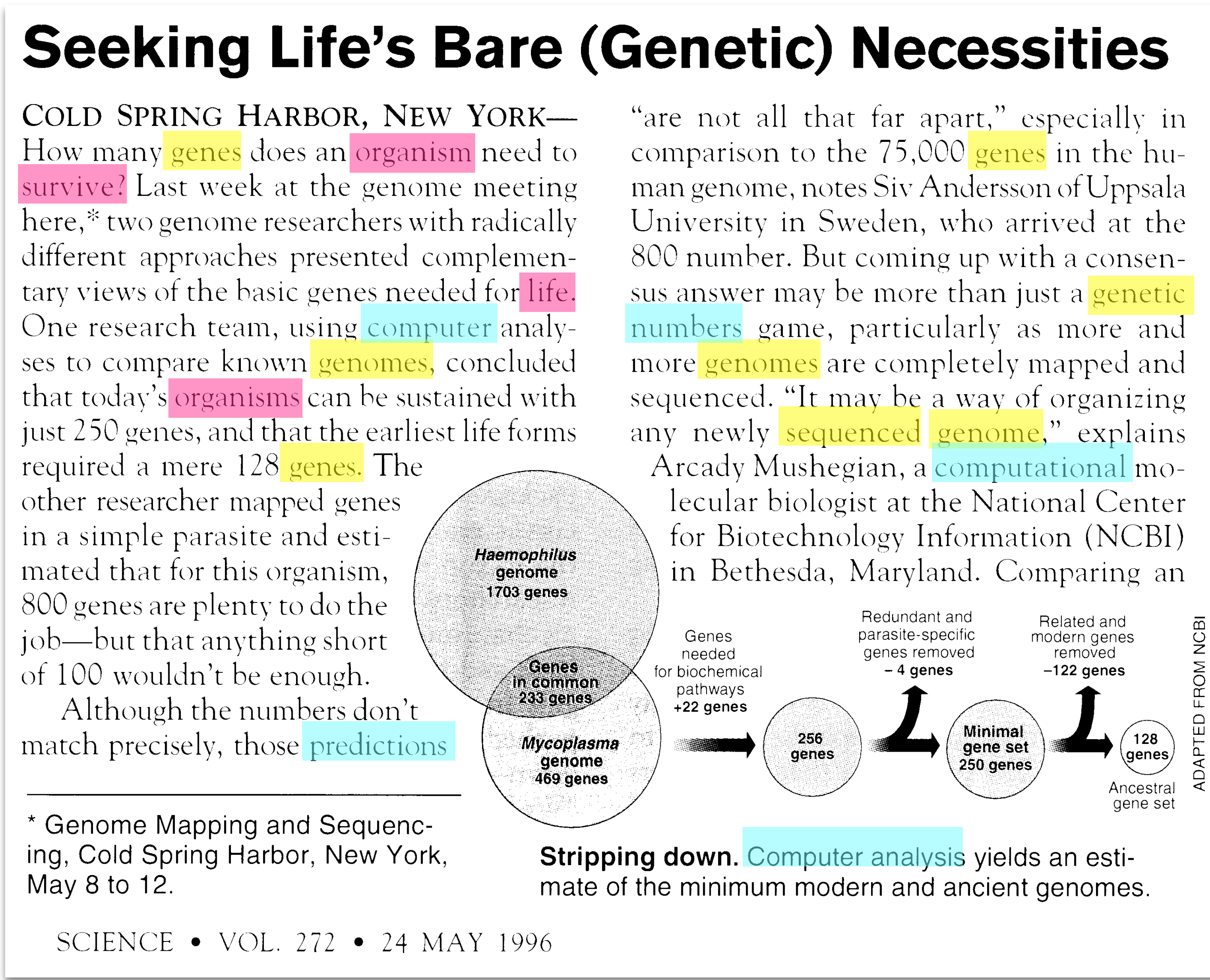
$$\underbrace{\begin{bmatrix} M \times K \end{bmatrix}}_{\text{Topic Assignment}} \times \underbrace{\begin{bmatrix} K \times V \end{bmatrix}}_{\text{Topics}} \approx \underbrace{\begin{bmatrix} M \times V \end{bmatrix}}_{\text{Dataset}}$$

- We will try and take a more probabilistic view, but useful to keep this in mind

Figure from Boyd-Graber, Hu and Mimno, 2017

Probabilistic Word Mixtures

Idea: Model text as a mixture over words (ignore order)



Seeking Life's Bare (Genetic) Necessities

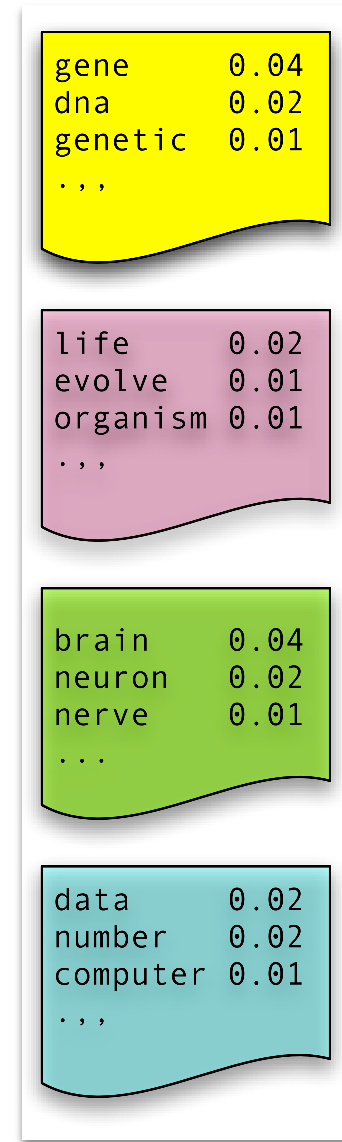
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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.

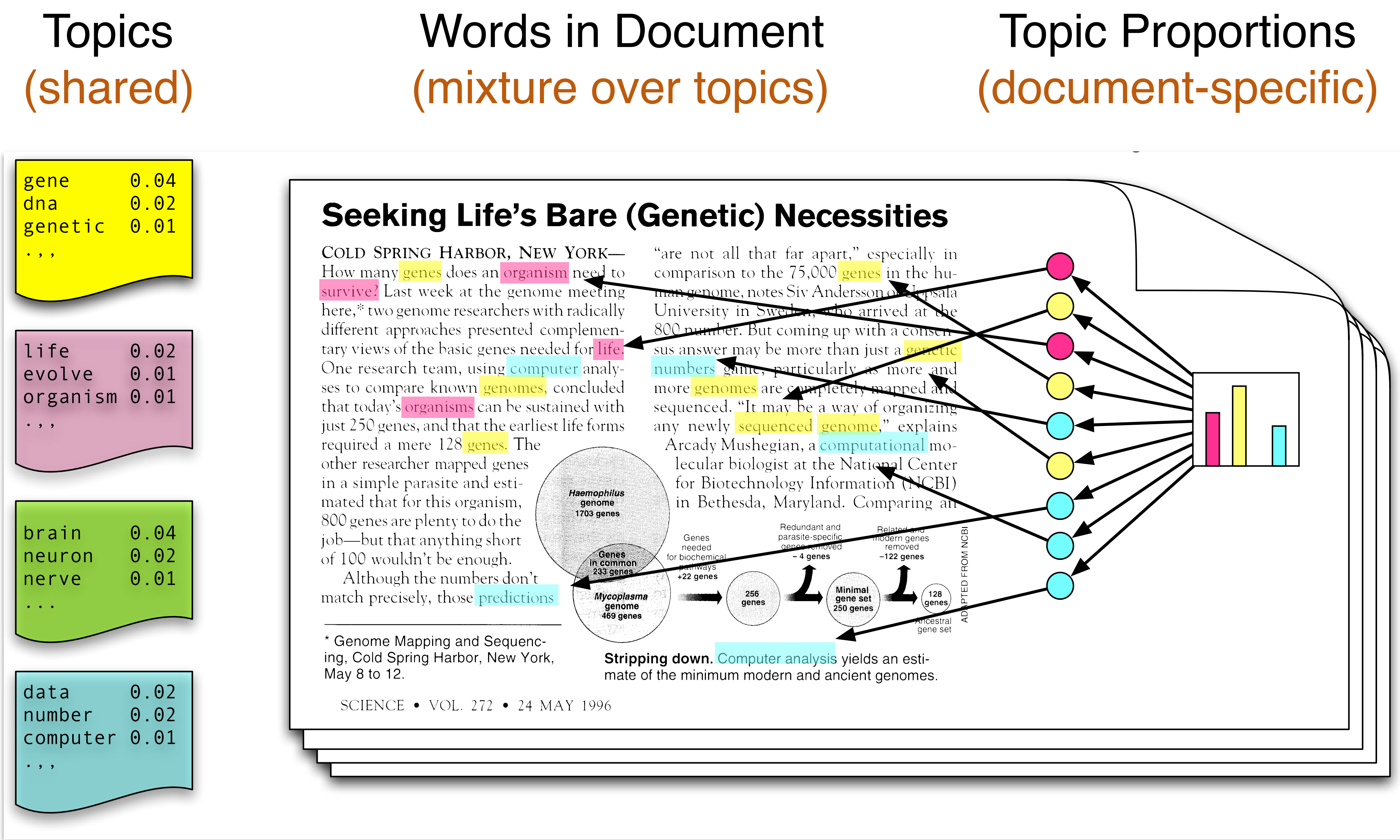
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



Words: $x_n | z_n = k \sim \mathrm{Discrete}(\boldsymbol{\beta}_k)$ Topics: $z_n \sim \mathrm{Discrete}(\boldsymbol{\theta})$

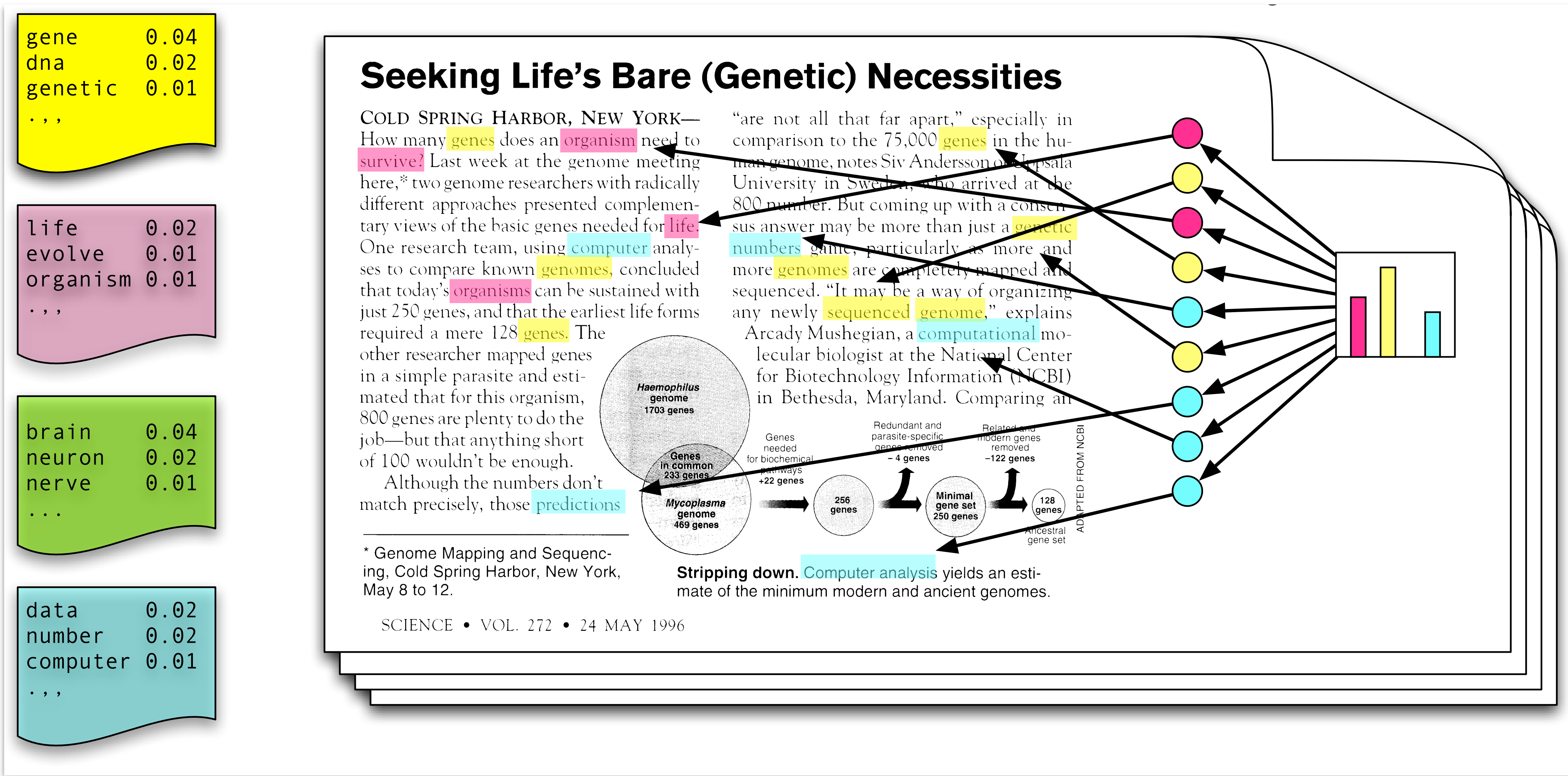
Topic Modeling



Idea: Model ***corpus*** of documents with ***shared*** topics

Topic Modeling

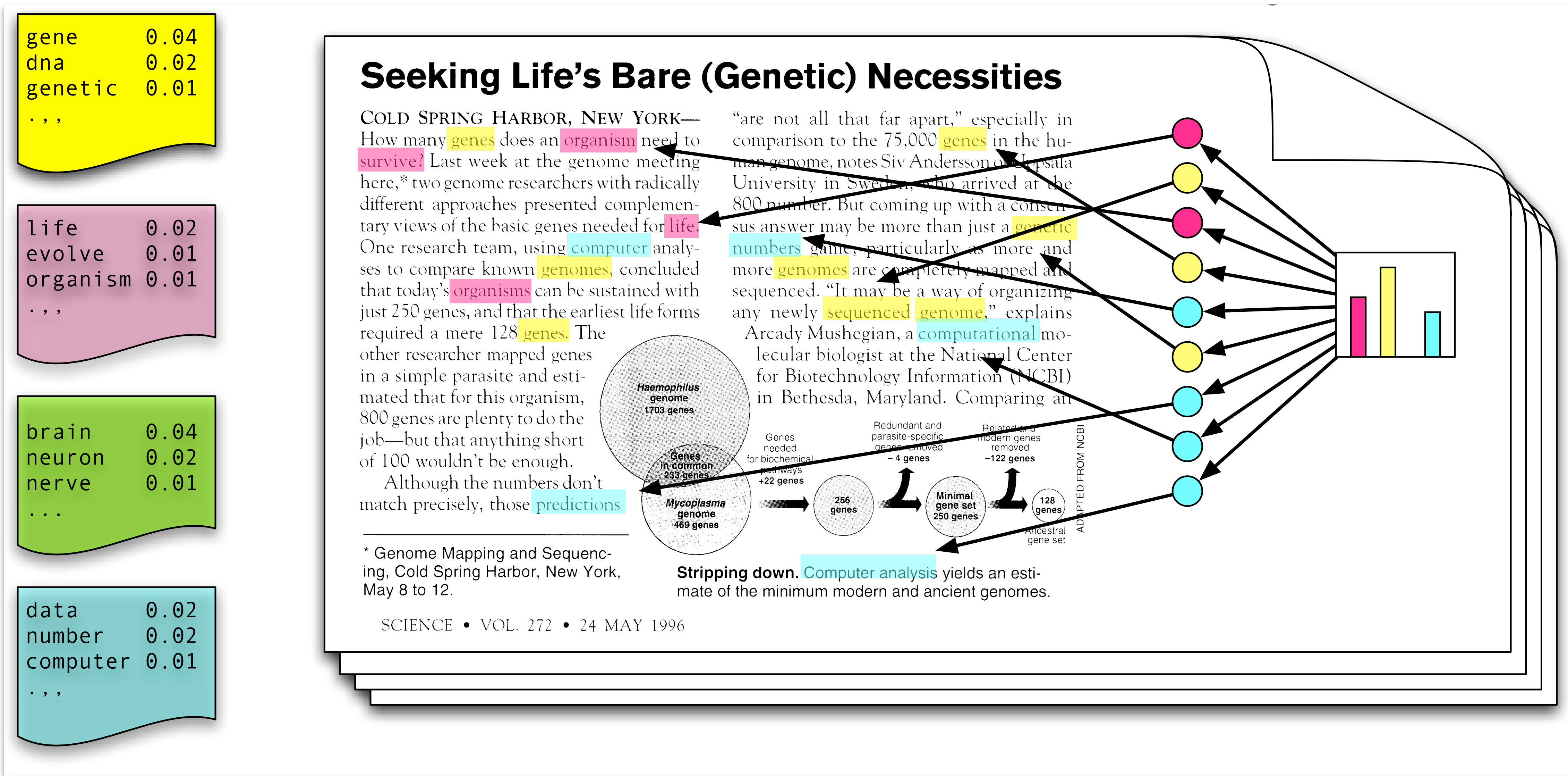
Topics	Words in Document	Topic Proportions
(shared)	(mixture over topics)	(document-specific)



- Each **topic** is a distribution over words

Topic Modeling

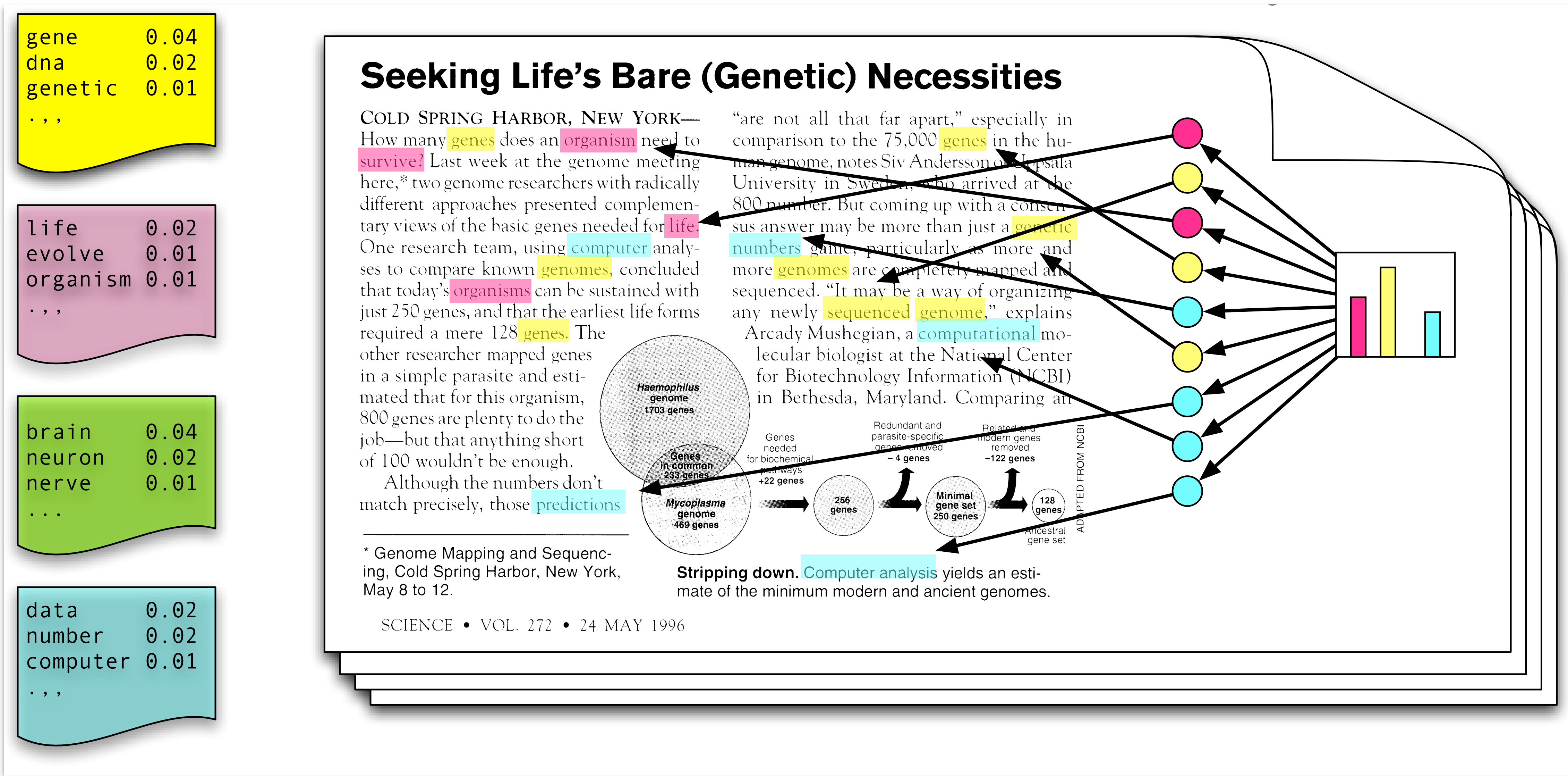
Topics	Words in Document	Topic Proportions
(shared)	(mixture over topics)	(document-specific)



- Each **topic** is a distribution over words
- Each **document** is a mixture over topics

Topic Modeling

Topics (shared) | Words in Document (mixture over topics) | Topic Proportions (document-specific)



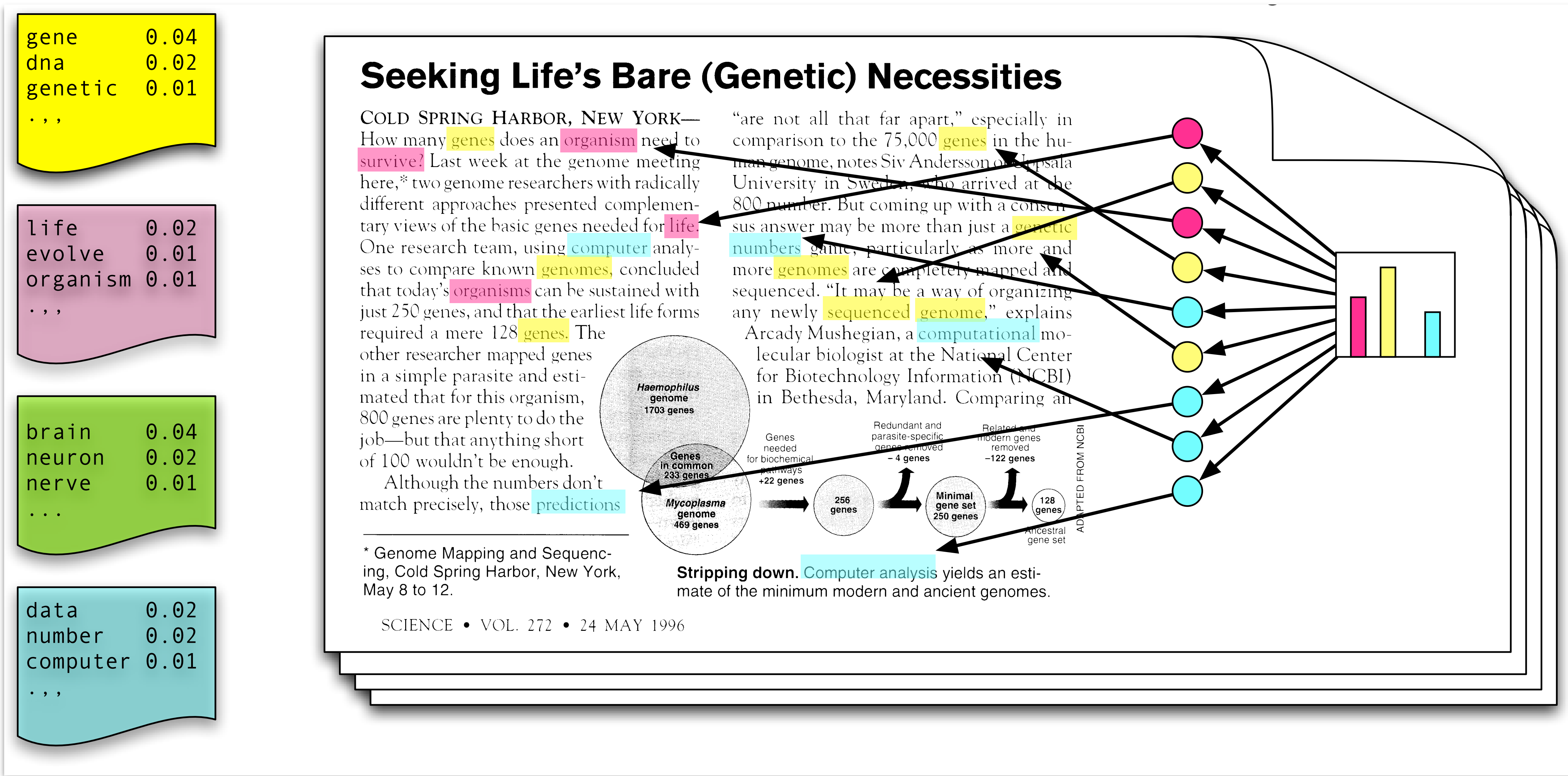
- Each **topic** is a distribution over words
- Each **document** is a mixture over topics
- Each **word** is drawn from one topic distribution

Topic Modeling

Topics (shared)

Words in Document (mixture over topics)

Topic Proportions (document-specific)



$$z_{dn} \sim \text{Discrete}(\theta_d)$$

$$x_{dn} \mid z_{dn} = k \sim \text{Discrete}(\boldsymbol{\beta}_k)$$

Each document has

Different topic proportions

LDA’s view of a document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

“Arts” **“Budgets”** **“Children”** **“Education”**

Slide credit: ***William Cohen (CMU)***

Example: Discovering scientific topics

Example Inference

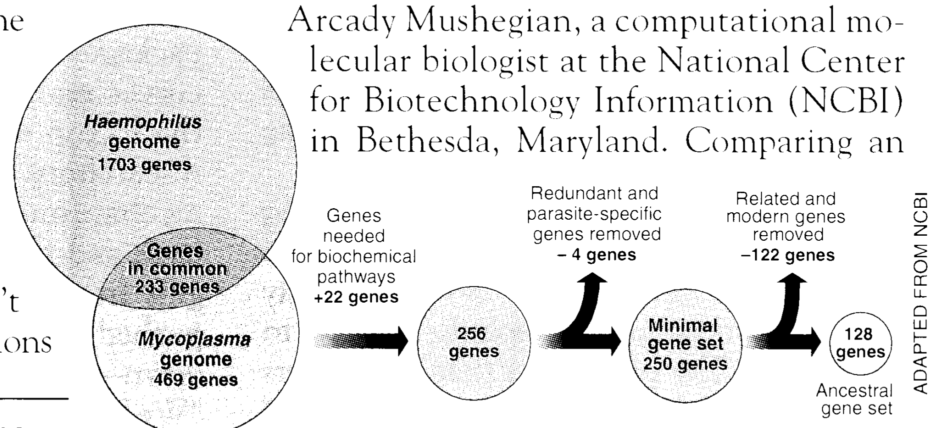
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Example Inference

Seeking Life's Bare (Genetic) Necessities

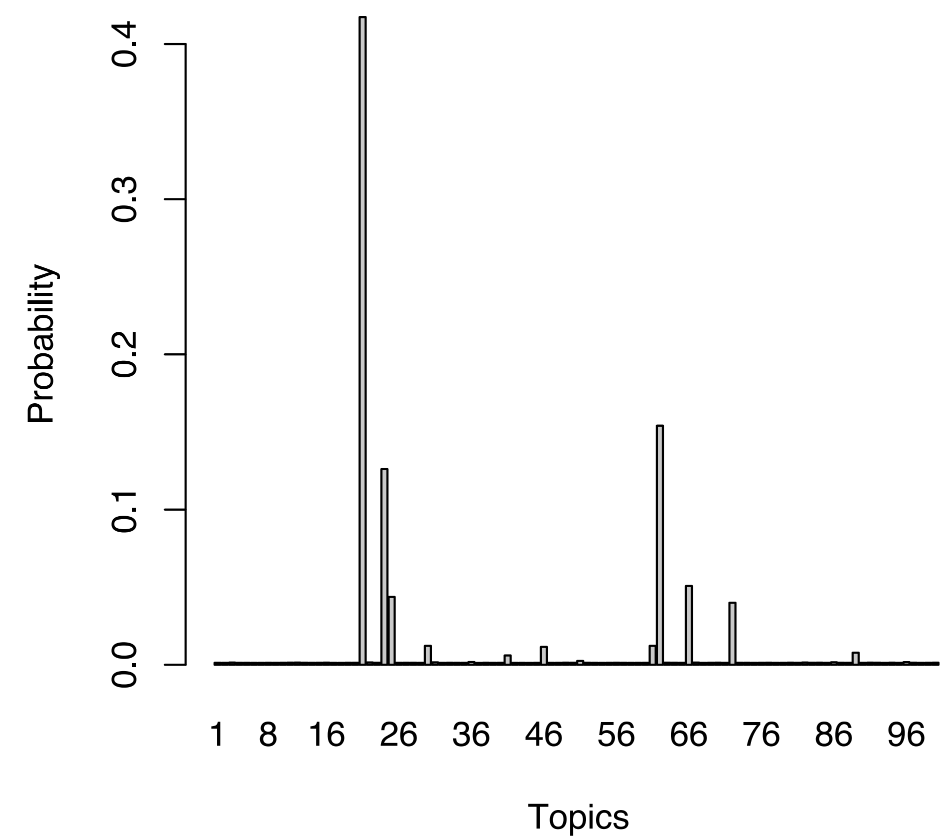
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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



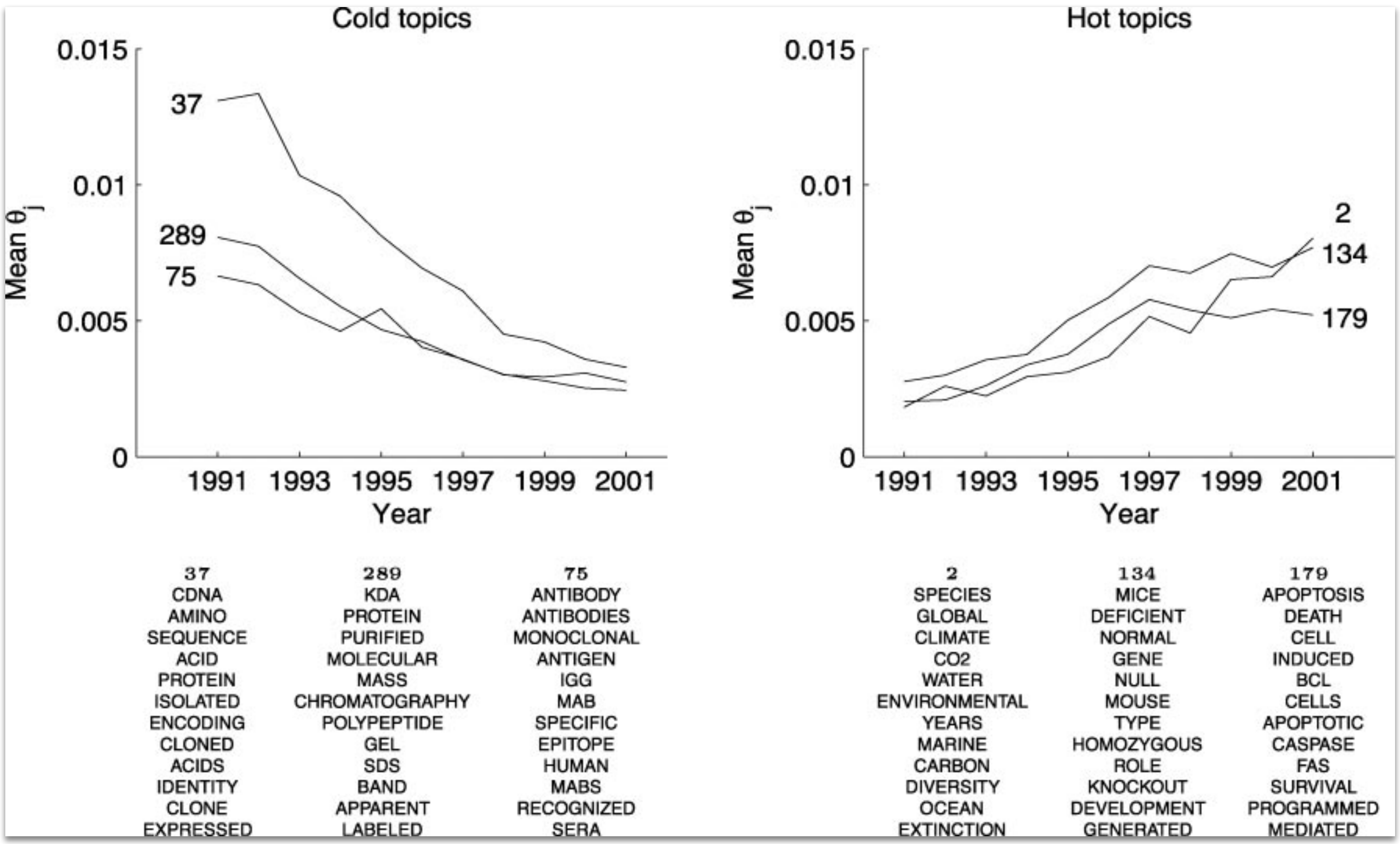
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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



Example Inference

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



From Griffiths and Steyvers, PNAS 2004

From Naive Bayes to Topic Models (*board*)

Likelihood

$$
\begin{aligned}
\log(p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)) &= \sum_n \log(p(\boldsymbol{x}_{dn} \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)) \\
&= \sum_n \log\left(\prod_v p(\boldsymbol{x}_{dn} = v \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)^{I[\boldsymbol{x}_{dn}=v]}\right) \\
&= \sum_{n,v} I[\boldsymbol{x}_{dn} = v] \log(p(\boldsymbol{x}_{dn} = v \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)) \\
&= \sum_{n,v} I[\boldsymbol{x}_{dn} = v] \log\left(\sum_k p(\boldsymbol{x}_{dn} = v, \boldsymbol{z}_{dn}=k \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)\right) \\
&= \sum_{n,v} I[\boldsymbol{x}_{dn} = v] \log\left(\sum_k p(\boldsymbol{z}_{dn}=k \mid \boldsymbol{\theta}_d)\, p(\boldsymbol{x}_{dn} = v \mid \boldsymbol{z}_{dn}=k, \boldsymbol{\beta})\right) \\
&= \sum_{n,v} I[\boldsymbol{x}_{dn} = v] \log\left(\sum_k \boldsymbol{\theta}_{d,k} \boldsymbol{\beta}_{k,v}\right) \\
&= \boldsymbol{X} \log \boldsymbol{\theta} \boldsymbol{\beta}
\end{aligned}
$$

How to estimate parameters in PLSA?

Let’s implement…

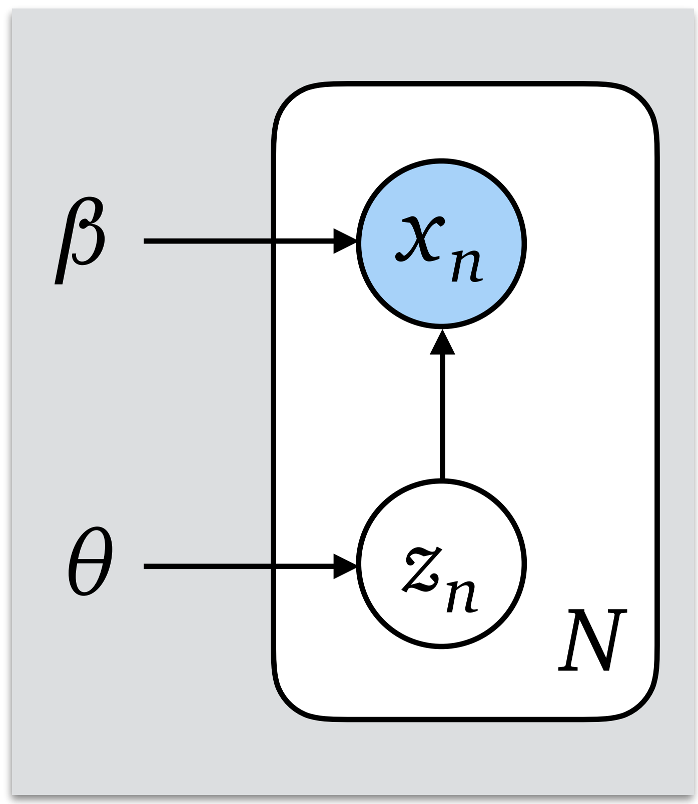
(*in class exercise*)

Evaluation:
Are these topics any good?

- As for clustering: a bit tricky. Thoughts on how we might evaluate topics?

Likelihood of held-out data

Observed data

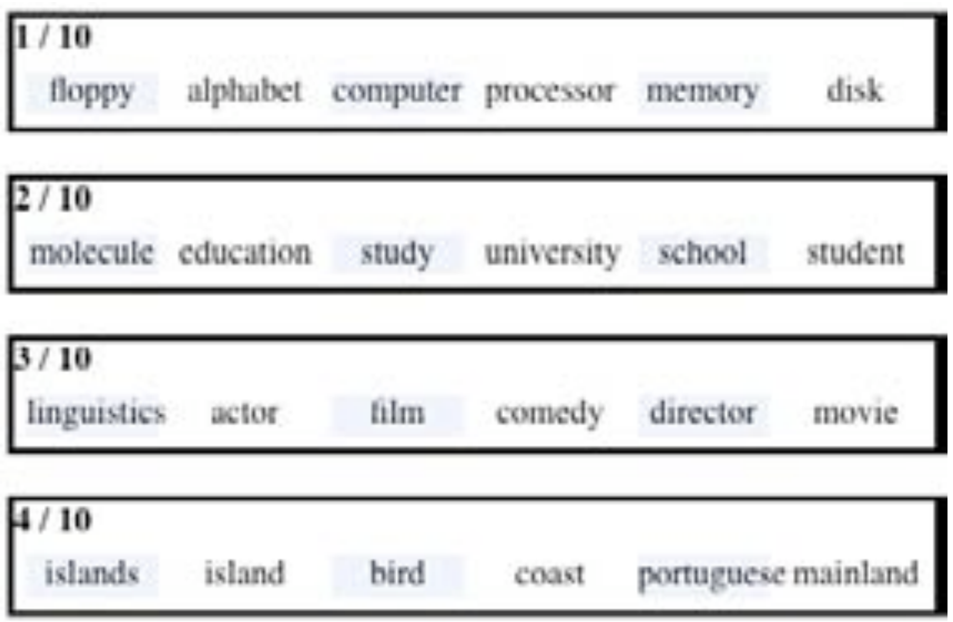


Fit model

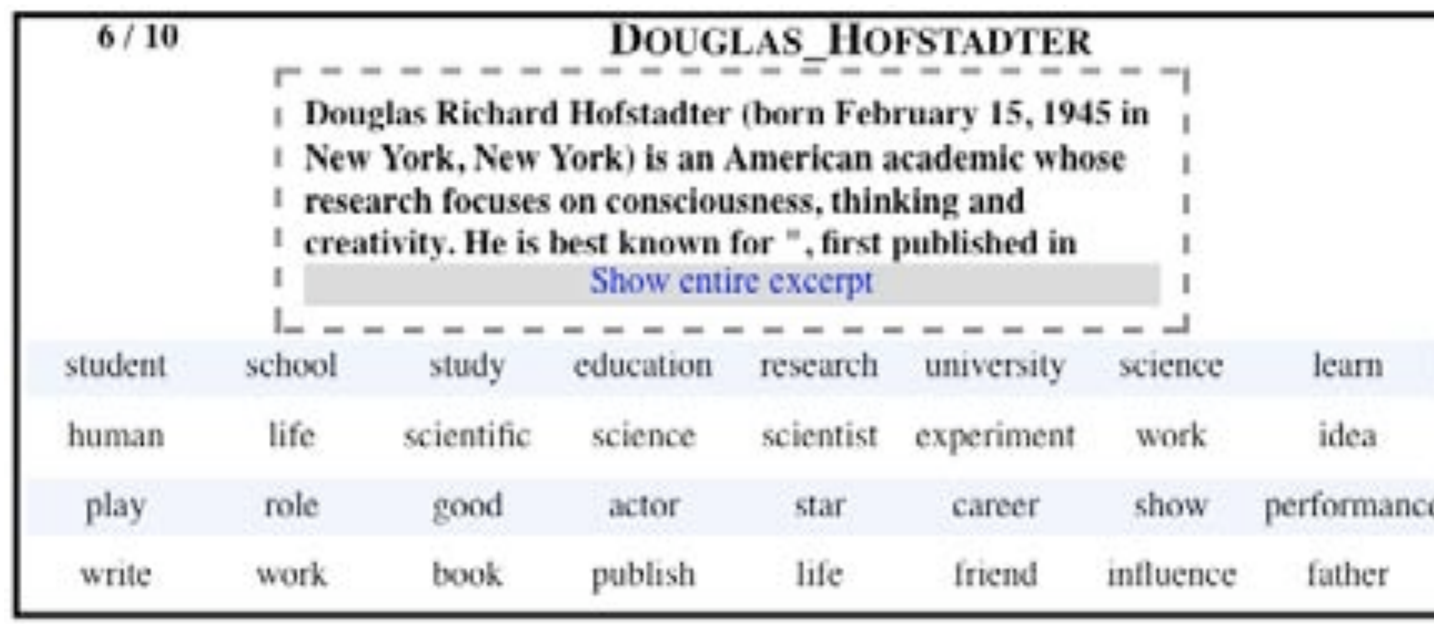
Held-out data

“Intrusion detection”

Word Intrusion



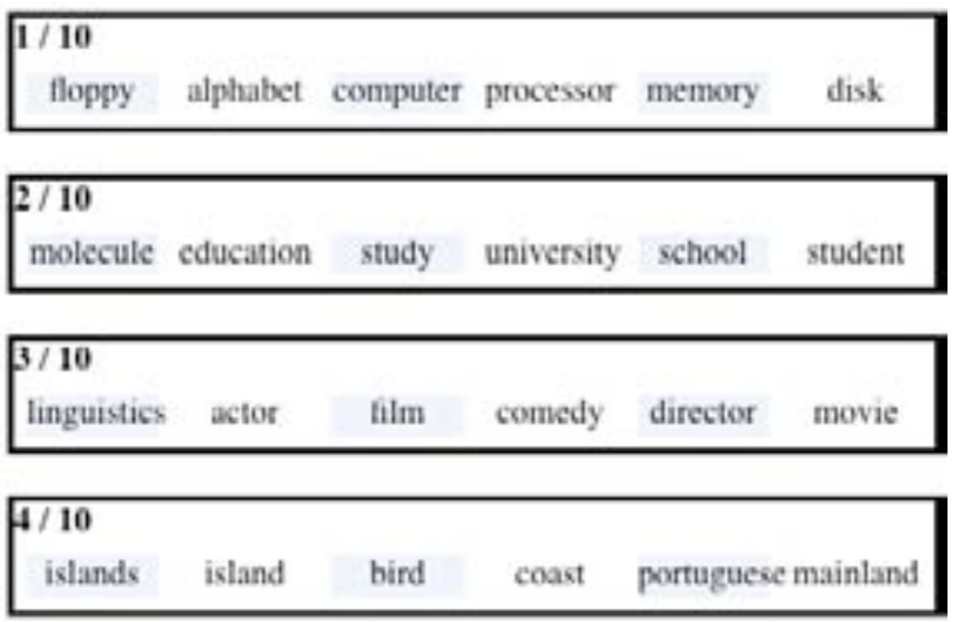
Topic Intrusion



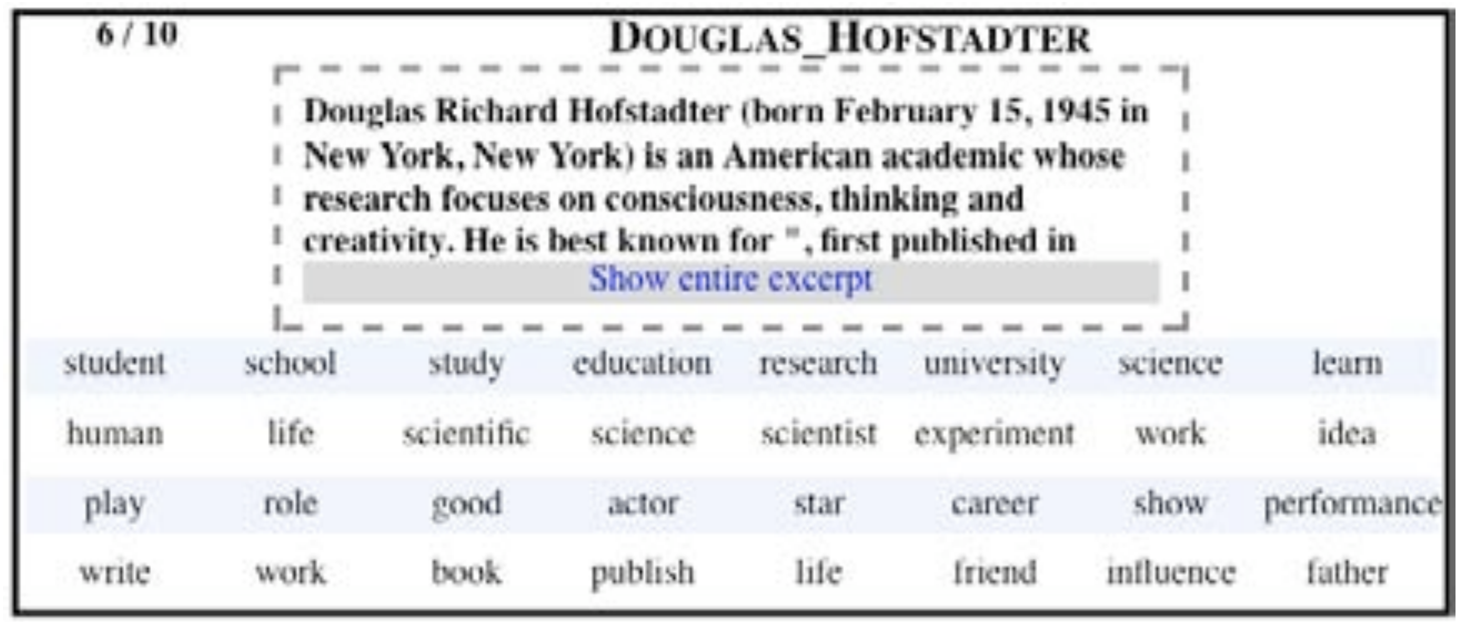
From Chang et al., 2009

“Intrusion detection”

Word Intrusion



Topic Intrusion

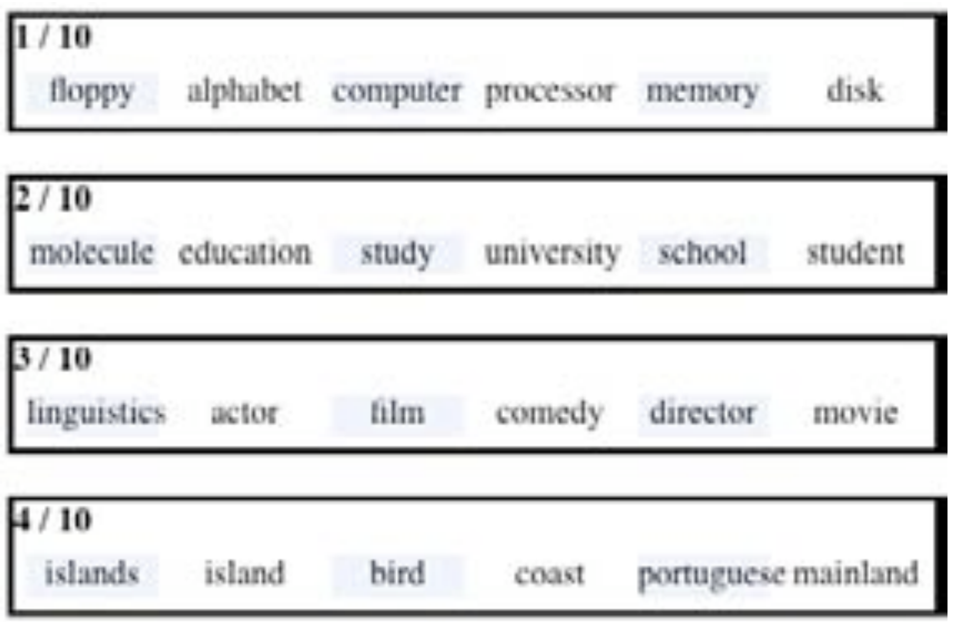


Which word doesn’t belong?

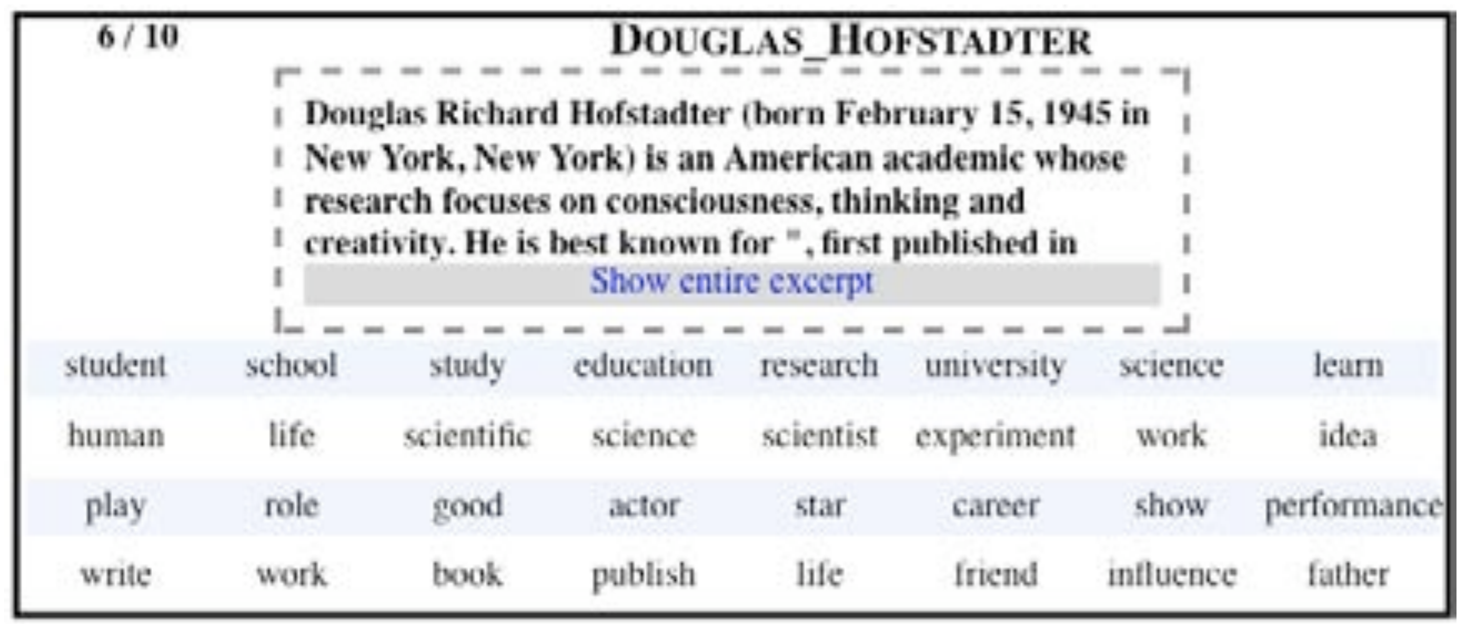
From Chang et al., 2009

“Intrusion detection”

Word Intrusion



Topic Intrusion



Which topic doesn’t belong?

From Chang et al., 2009

Summing up

- PLSA is a simple ad-mixture model that uncovers *topics* (distributions over words) and soft-assigns instances to these.

Summing up

- PLSA is a simple ad-mixture model that uncovers *topics* (distributions over words) and soft-assigns instances to these.

- We saw parameter estimation via Expectation-Maximization.

Summing up

- PLSA is a simple ad-mixture model that uncovers *topics* (distributions over words) and soft-assigns instances to these.

- We saw parameter estimation via Expectation-Maximization.

- Next time: Introducing priors into topic models — Latent Dirichlet Allocation (**LDA**).

  ★ This will motivate *sampling-based* estimation