NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS
FOUNDED AD 1898

Topic Models

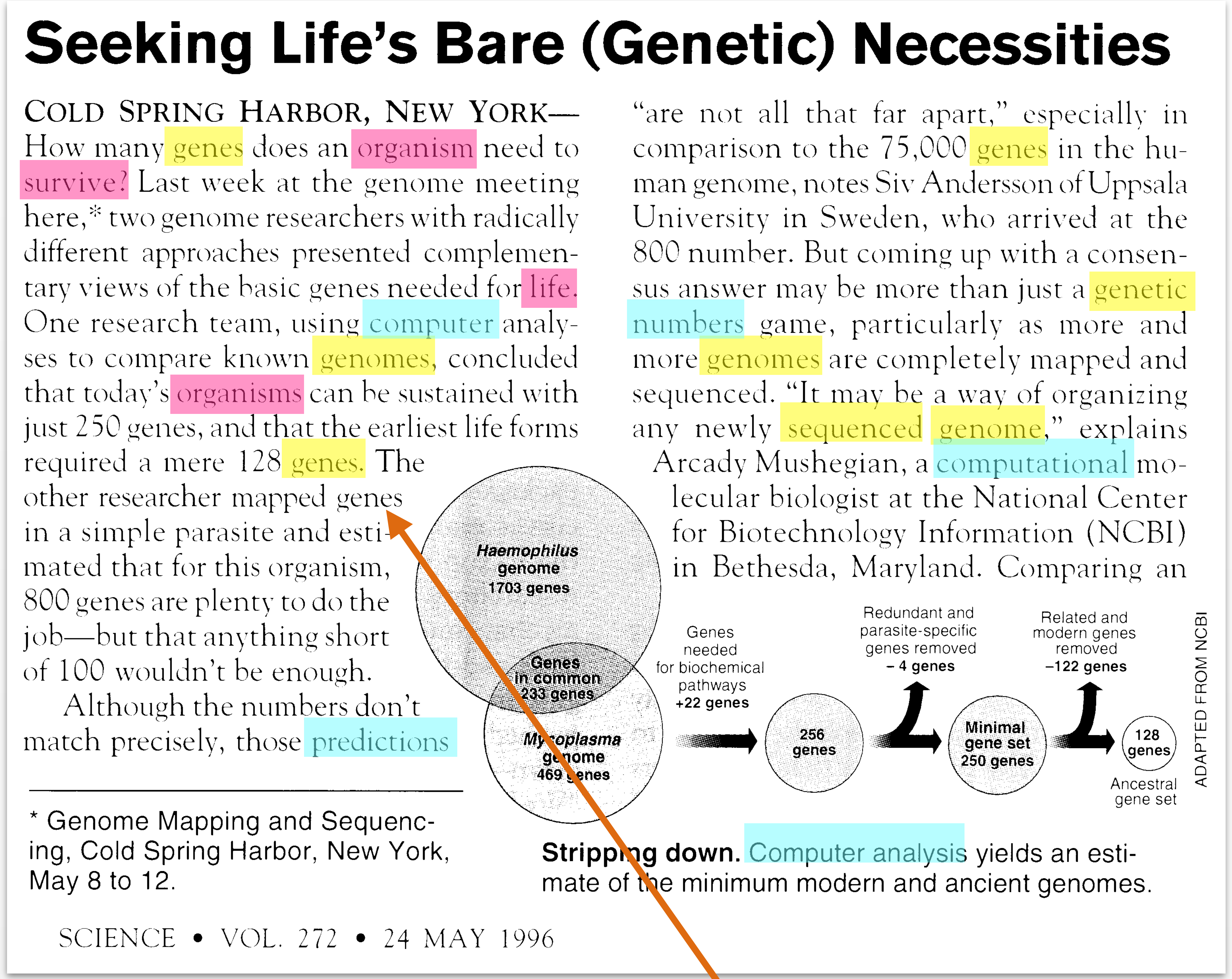
Shantanu Jain

Topic Modeling Basics

Borrowing from:
David Blei
(Columbia)

Word Mixtures

Idea: Model text as a “bag” of 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.

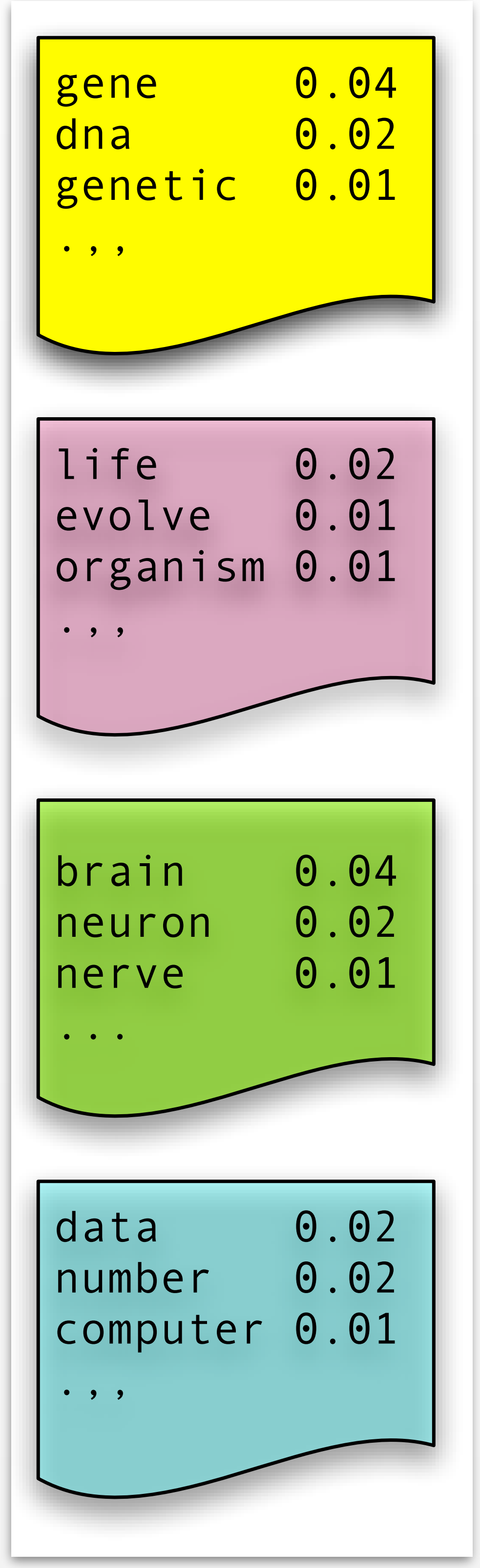
Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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



gene	0.04
dna	0.02
genetic	0.01
. . ,	

$p(x \mid z = 1)$

life	0.02
evolve	0.01
organism	0.01
. . ,	

$p(x \mid z = 2)$

brain	0.04
neuron	0.02
nerve	0.01
. . .	

$p(x \mid z = 3)$

data	0.02
number	0.02
computer	0.01
. . ,	

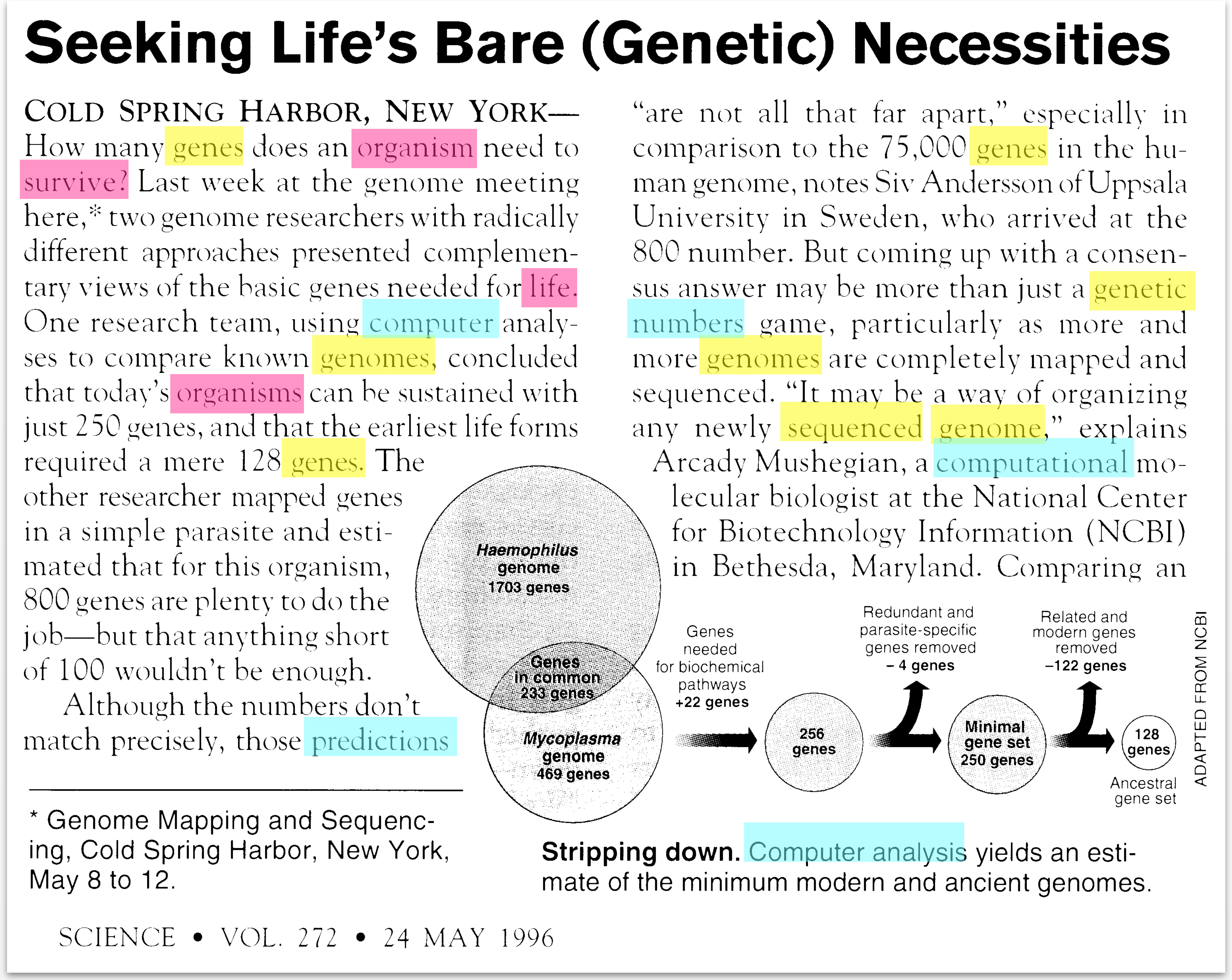
$p(x \mid z = 4)$

Word in vocabulary: $x_n \in \{1, \ldots, V\}$

Topic assignment: $z_n \in \{1, \ldots, K\}$

- Total N words in a document
- n denotes the index of the n^{th} word in the document.
- V is the number of words in the vocabulary.
- K is the number of topics.

Word Mixtures



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.

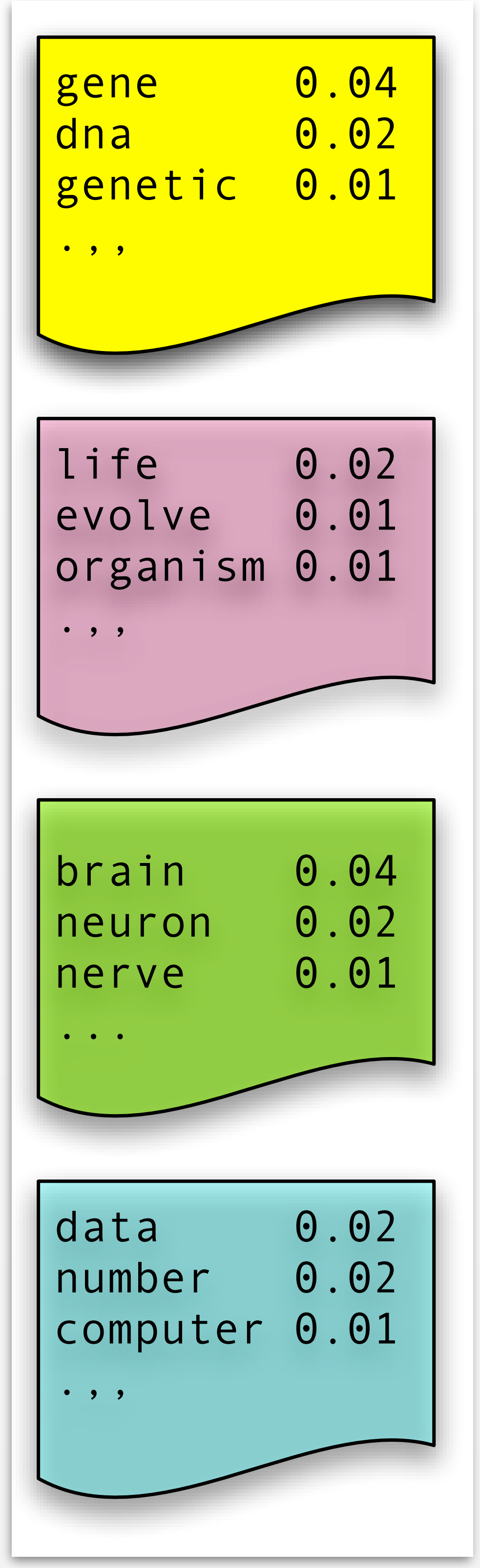
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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



gene	0.04
dna	0.02
genetic	0.01
...	

$p(x \mid z=1, \beta)$

life	0.02
evolve	0.01
organism	0.01
...	

$p(x \mid z=2, \beta)$

brain	0.04
neuron	0.02
nerve	0.01
...	

$p(x \mid z=3, \beta)$

data	0.02
number	0.02
computer	0.01
...	

$p(x \mid z=4, \beta)$

θ: topic proportions/probabilities, probability over the K topics

$$\theta = [\theta_1, \theta_2, \ldots \theta_K] \qquad \sum_k \theta_k = 1$$

$$p(z_n = k \mid \theta) = \theta_k$$

β_k: k^{th} topic's word probabilities over the vocabulary

$$\beta_k = [\beta_{k1}, \beta_{k2}, \ldots \beta_{kV}] \qquad \sum_i \beta_{ki} = 1$$

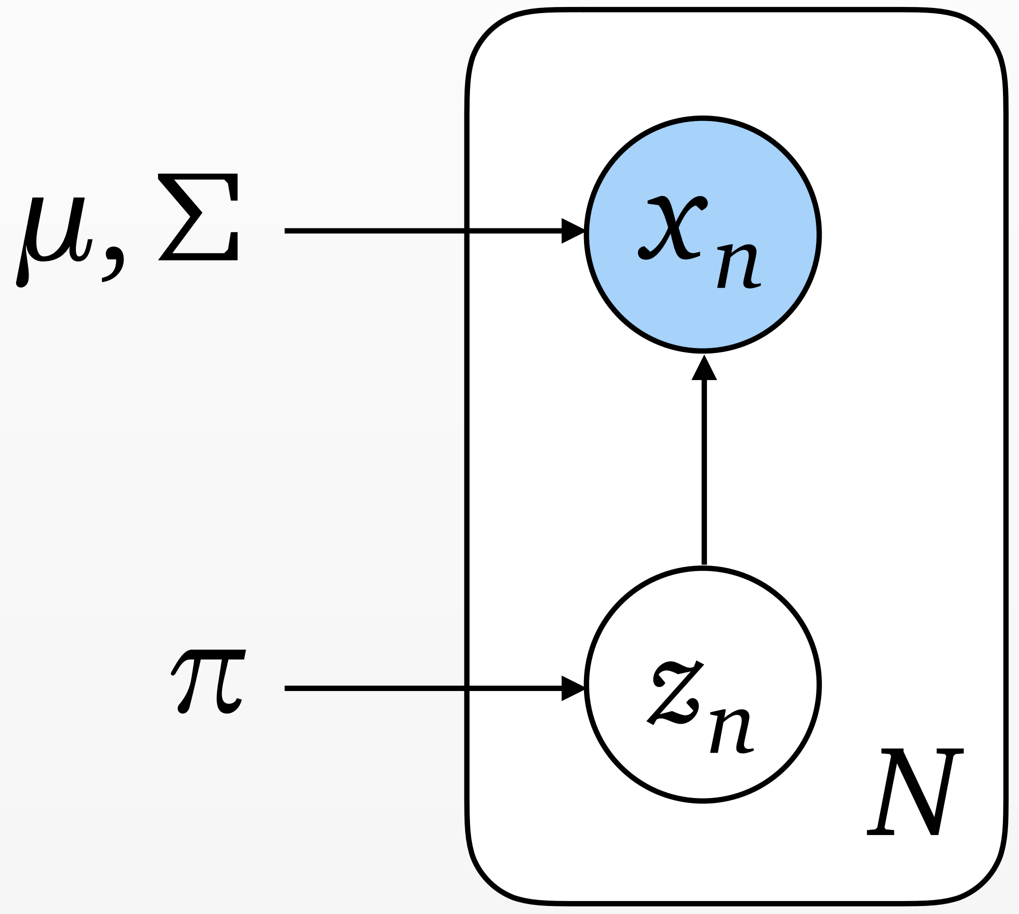
$$p(x_n = i \mid z_n = k, \beta) = \beta_{ki}$$

$z_n \sim \text{Discrete}(\theta)$ *Pick a topic*

$x_n \mid z_n = k \sim \text{Discrete}(\beta_k)$ *Pick a word given topic*

Gaussian Mixtures vs Word Mixtures

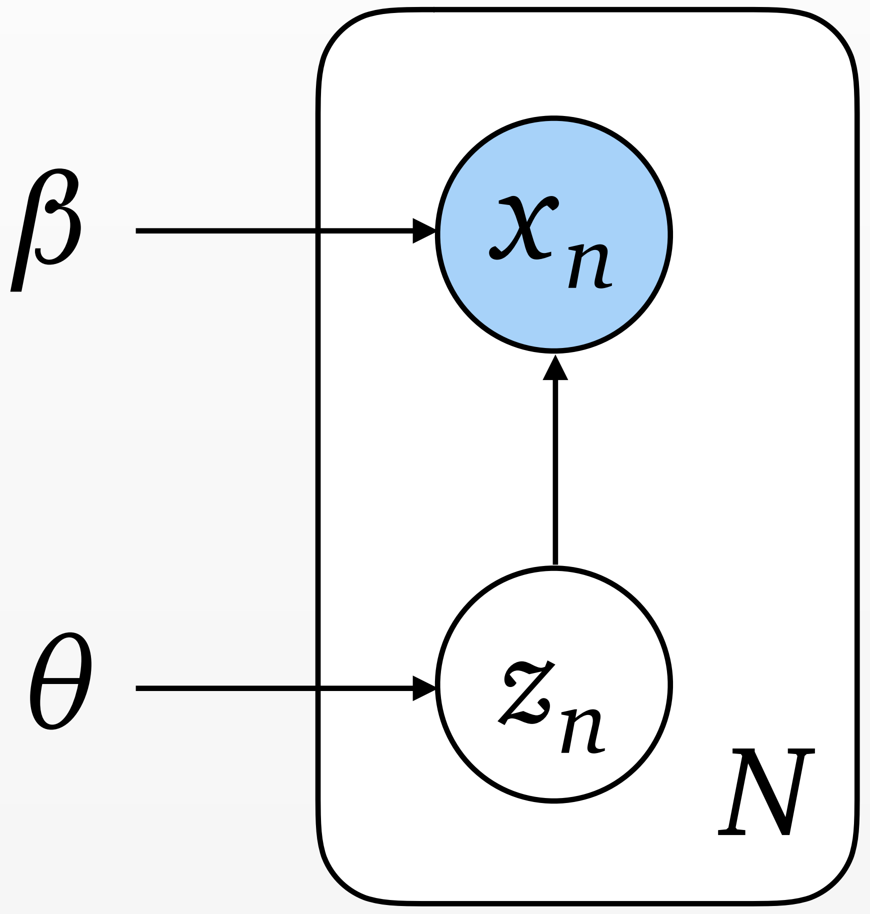
Gaussian Mixture Model



$$z_n \sim \text{Discrete}(\pi_1, \ldots, \pi_K)$$

$$x_n \mid z_n = k \sim \text{Normal}(\mu_k, \Sigma_k)$$

Discrete Mixture Model



$$z_n \sim \text{Discrete}(\theta_1, \ldots, \theta_K)$$

$$x_n \mid z_n = k \sim \text{Discrete}(\beta_{k,1}, \ldots, \beta_{k,V})$$

Difference: Replace Gaussian with Discrete

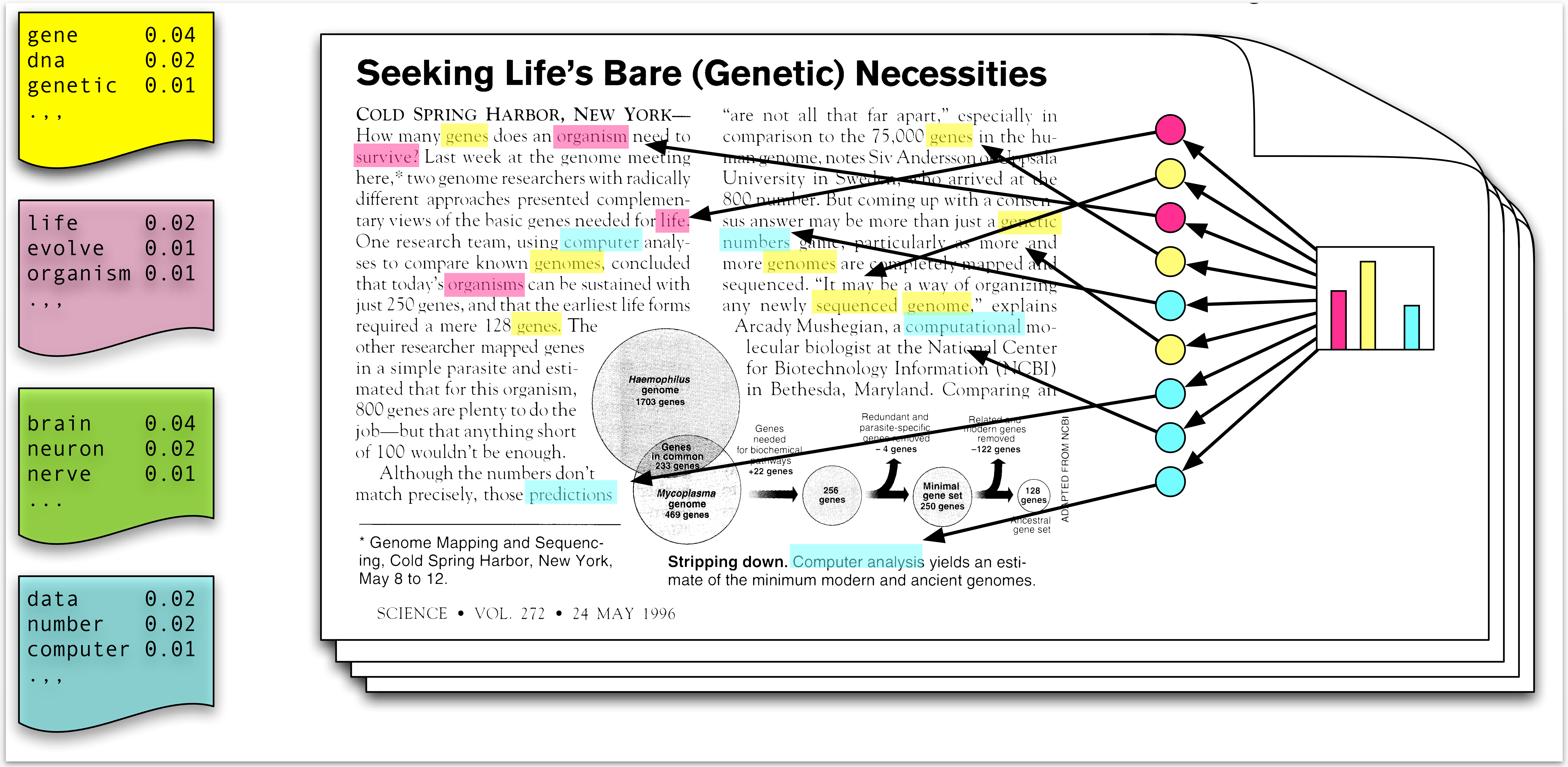
Using the term Discrete distribution to mean Categorical distribution

Topic Modeling

Topics
(shared)

Words in Document
(mixture over topics)

Topic Proportions
(document-specific)



gene 0.04
dna 0.02
genetic 0.01
. , ,

life 0.02
evolve 0.01
organism 0.01
. , ,

brain 0.04
neuron 0.02
nerve 0.01
. . .

data 0.02
number 0.02
computer 0.01
. , ,

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

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

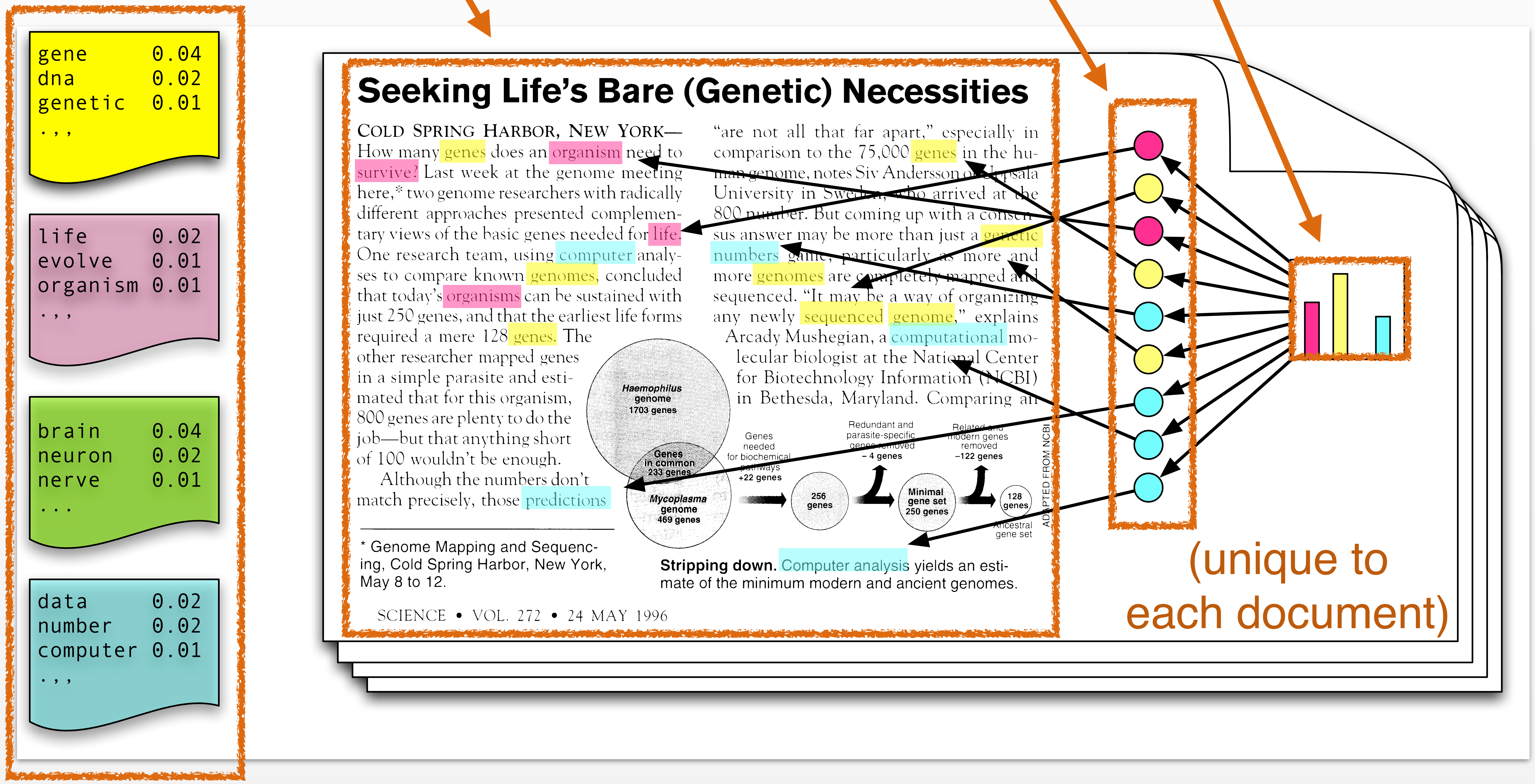
Topic Modeling

β_k: Topics

x_d: Words

z_d: Assignments

θ_d: Topic Proportions



gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

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

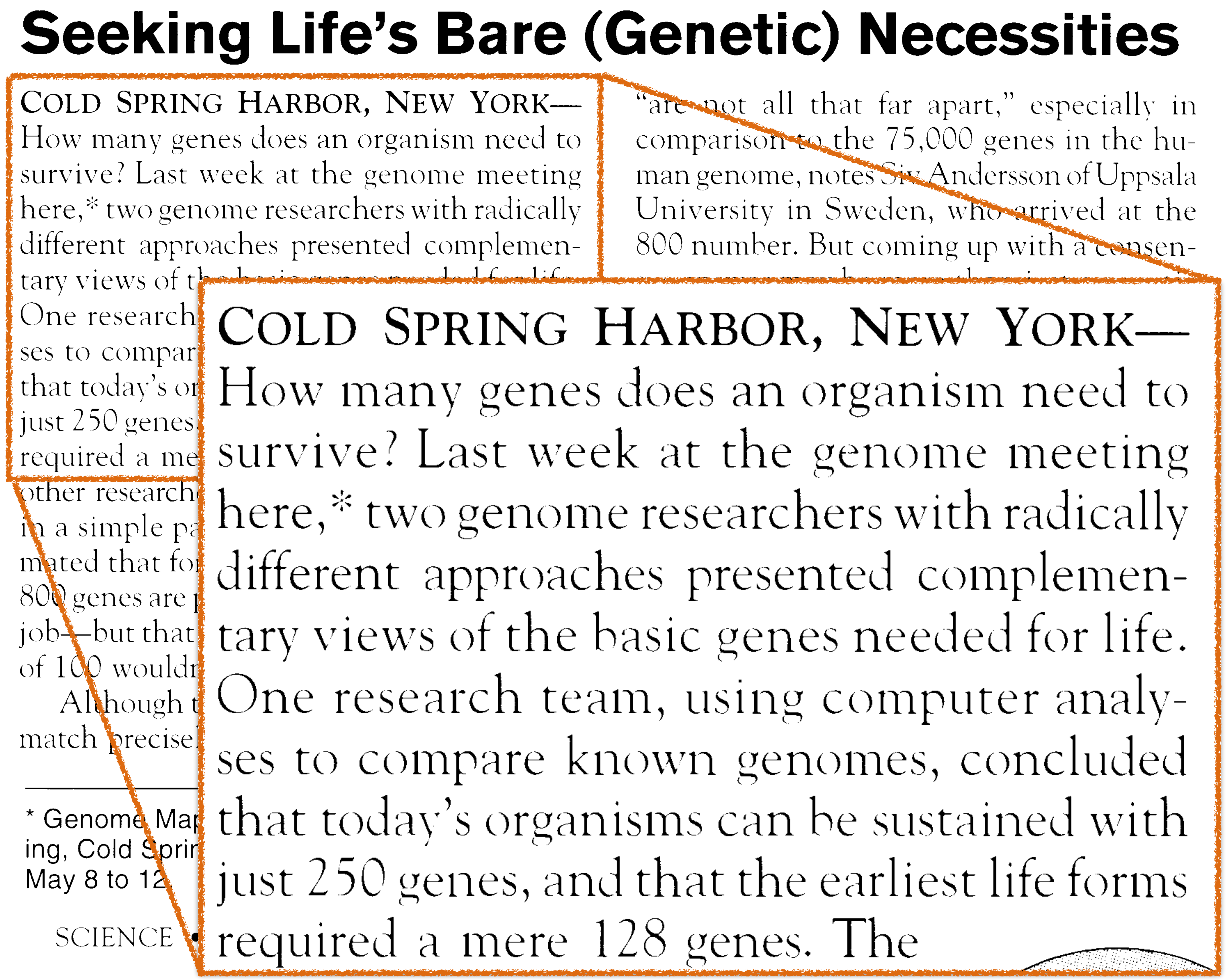
(unique to each document)

(shared across documents)

$$\boldsymbol{z}_{d,n} \sim \text{Discrete}(\boldsymbol{\theta}_d)$$

$$\boldsymbol{x}_{d,n} \mid \boldsymbol{z}_{d,n}=k \sim \text{Discrete}(\boldsymbol{\beta}_k)$$

Distribution over Topic Assignments

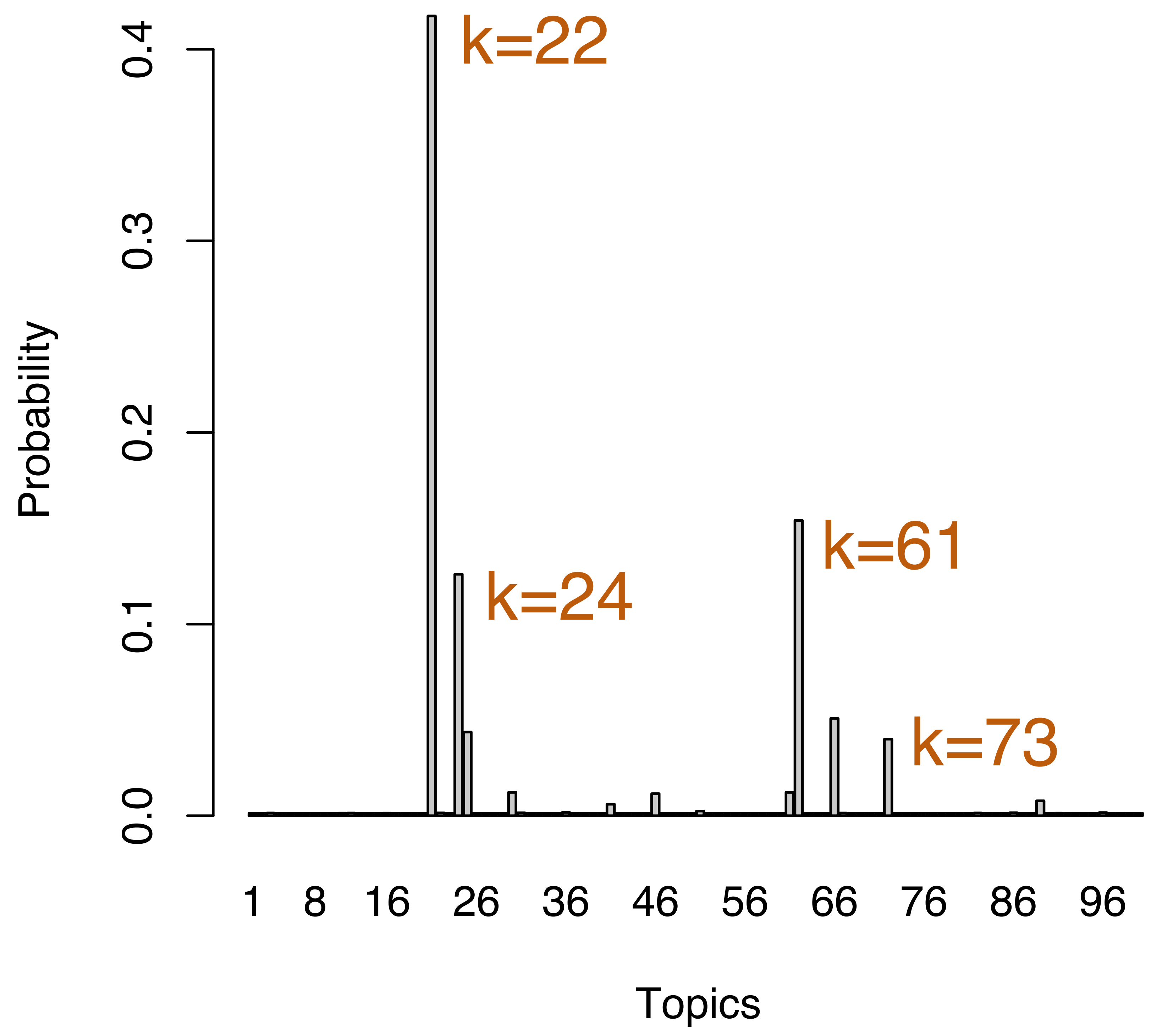


Seeking Life's Bare (Genetic) Necessities

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

"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 consen-

$$z_{d,n} \sim \text{Discrete}(\boldsymbol{\theta}_d)$$



Next Slide: Frequent words in these topics

Most Probable Words in Topics

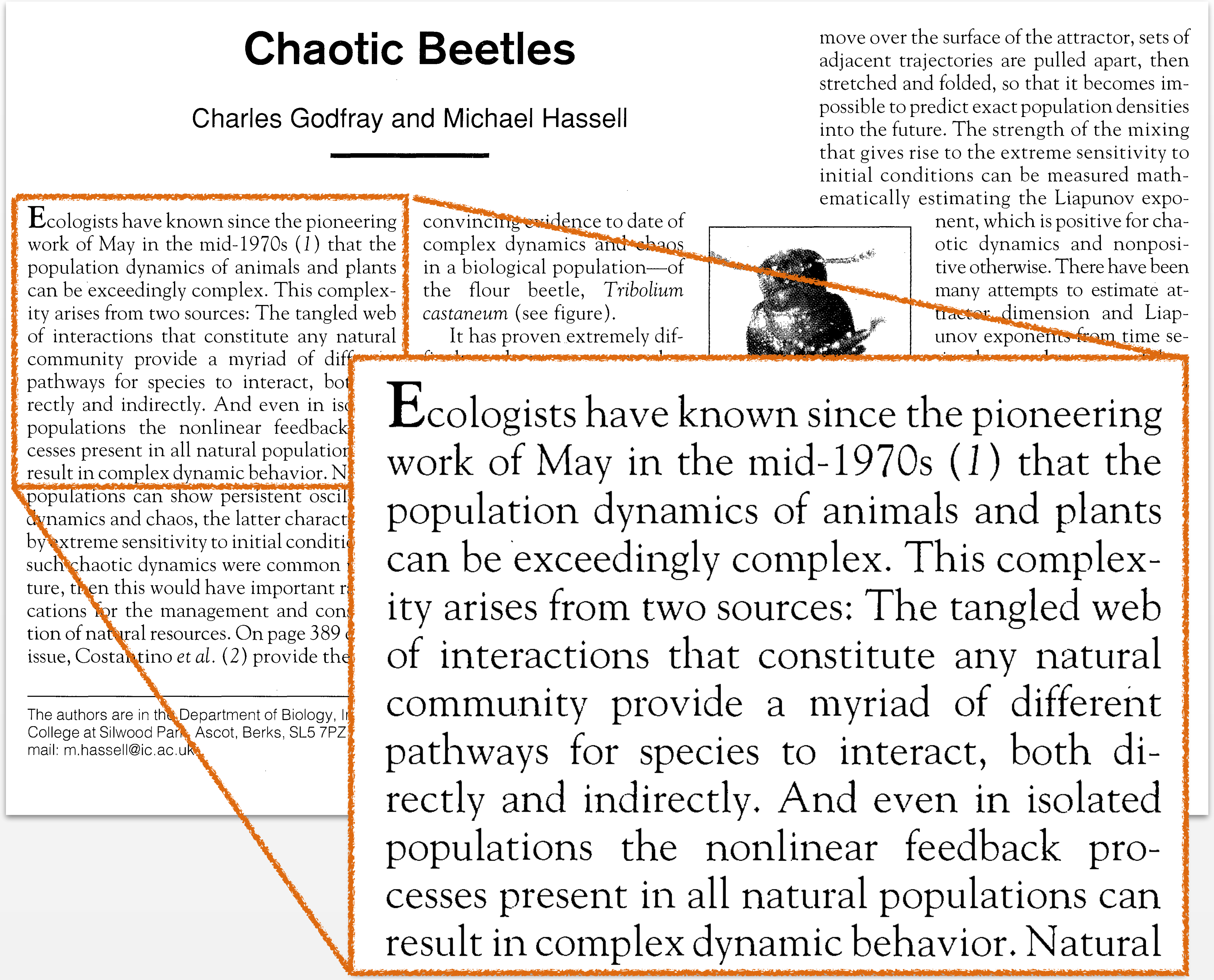
$$\boldsymbol{x}_{d,n} \mid \boldsymbol{z}_{d,n} = k \sim \text{Discrete}(\boldsymbol{\beta}_k)$$

Most frequent

(*within topic*)

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
k=22	k=24	k=61	k=73

Each Document has Different Topics



Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of diff
pathways for species to interact, bot
rectly and indirectly. And even in iso
populations the nonlinear feedback
cesses present in all natural population
result in complex dynamic behavior. N
populations can show persistent oscil
dynamics and chaos, the latter charact
by extreme sensitivity to initial conditi
such chaotic dynamics were common
ture, then this would have important r
cations for the management and con
tion of natural resources. On page 389
issue, Costantino *et al*. (2) provide the

convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely dif-

move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time se-

---

The authors are in the Department of Biology, I
College at Silwood Par Ascot, Berks, SL5 7PZ
mail: m.hassell@ic.ac.uk

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural

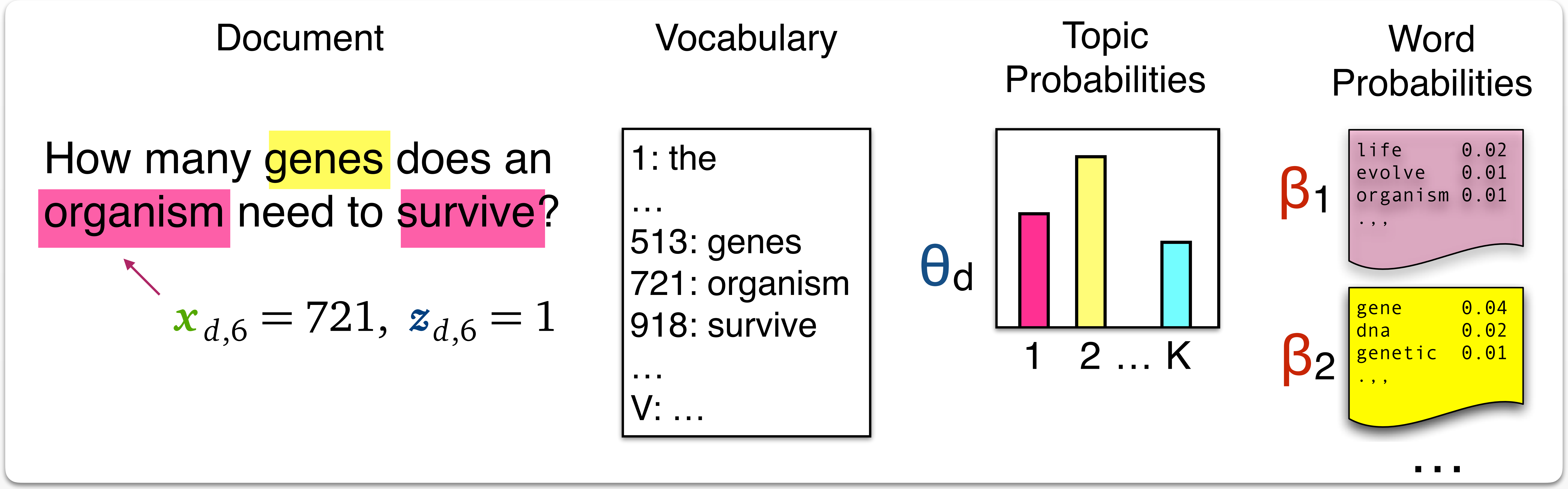
- problem
- problems
- mathematical
- number
- new
- mathematics
- university
- two
- first
- numbers
- work
- time
- mathematicians
- chaos
- chaotic

- model
- rate
- constant
- distribution
- time
- number
- size
- values
- value
- average
- rates
- data
- density
- measured
- models

- selection
- male
- males
- females
- sex
- species
- female
- evolution
- populations
- population
- sexual
- behavior
- evolutionary
- genetic
- reproductive

- species
- forest
- ecology
- fish
- ecological
- conservation
- diversity
- population
- natural
- ecosystems
- populations
- endangered
- tropical
- forests
- ecosystem

Estimating the Parameters



How many genes does an organism need to survive?

$x_{d,6} = 721, \; z_{d,6} = 1$

1: the
…
513: genes
721: organism
918: survive
…
V: …

life	0.02
evolve	0.01
organism	0.01
...	

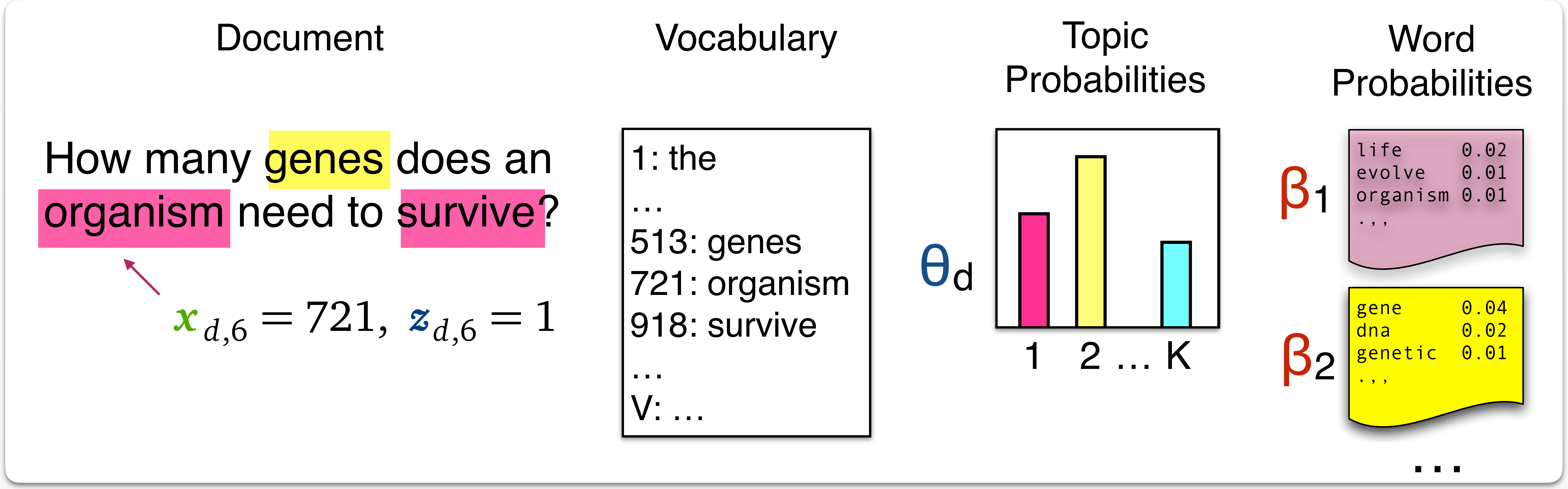
gene	0.04
dna	0.02
genetic	0.01
...	

Maximum Likelihood: $\max\limits_{\theta,\beta} \; \log p(x \mid \theta, \beta)$

$$\begin{array}{lccccr}
\text{d=1} & [x_{1,1}, x_{1,2}, \ldots, x_{1,N_1}] & & [\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{1,K}] & [\beta_{1,1}, \beta_{1,2}, \ldots, \beta_{1,V}] & \text{k=1} \\
\text{d=2} & [x_{2,1}, x_{2,2}, \ldots, x_{2,N_2}] & & [\theta_{2,1}, \theta_{2,2}, \ldots, \theta_{2,K}] & [\beta_{2,1}, \beta_{2,2}, \ldots, \beta_{2,V}] & \text{k=2} \\
 & \ldots & & \ldots & \ldots & \\
\text{d=D} & [x_{D,1}, x_{D,2}, \ldots, x_{D,N_D}] & & [\theta_{D,1}, \theta_{D,2}, \ldots, \theta_{D,K}] & [\beta_{K,1}, \beta_{K,2}, \ldots, \beta_{K,V}] & \text{k=K}
\end{array}$$

(not a matrix) D x K K x V

Calculating the Likelihood for each Word



Document

How many genes does an organism need to survive?

$x_{d,6} = 721, \ z_{d,6} = 1$

Vocabulary

1: the
…
513: genes
721: organism
918: survive
…
V: …

Topic Probabilities

Word Probabilities

β_1

```
life      0.02
evolve    0.01
organism  0.01
.,,
```

β_2

```
gene      0.04
dna       0.02
genetic   0.01
.,,
```

…

Probability that word *n* is entry *v* in the vocabulary

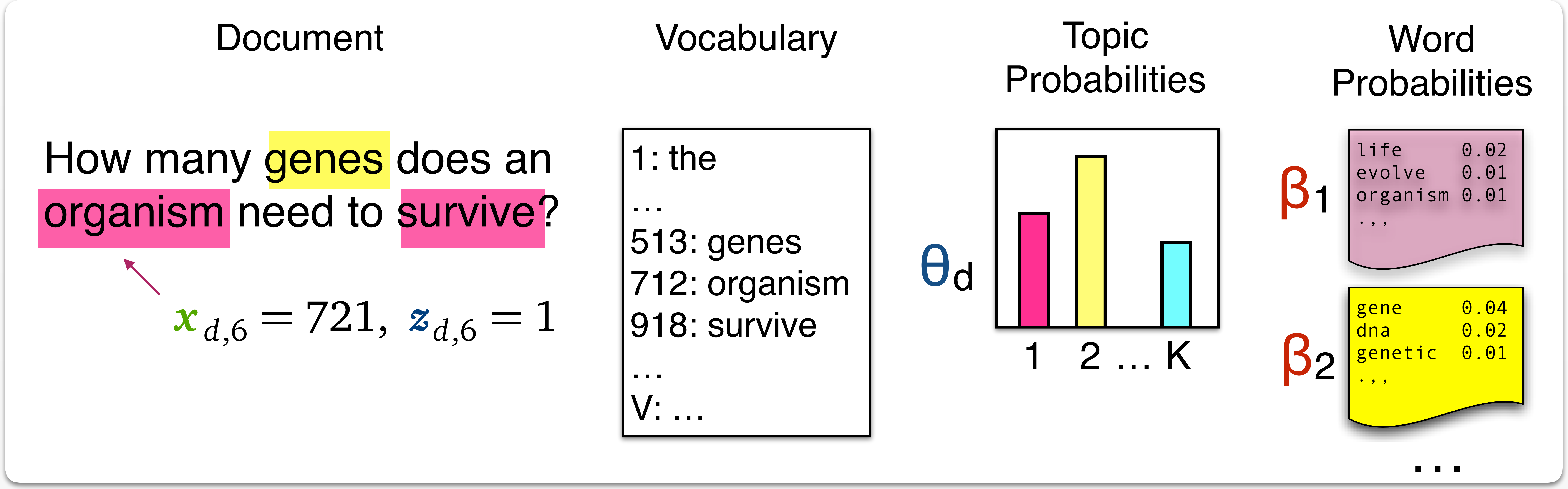
Probability of word *v* given topic *k*

Probability that word belongs to topic *k*

$$p(x_{d,n}=v \mid \beta, \theta_d) = \sum_{k=1}^{K} p(x_{d,n}=v \mid \beta, \ z_{d,n}=k) \, p(z_{d,n}=k \mid \theta_d)$$

$$= \sum_{k=1}^{K} \beta_{k,v} \theta_{d,k}$$

Computing the Likelihood



Document

How many genes does an organism need to survive?

$\boldsymbol{x}_{d,6} = 721, \ \boldsymbol{z}_{d,6} = 1$

Vocabulary

1: the
…
513: genes
712: organism
918: survive
…
V: …

Topic Probabilities

Word Probabilities

β_1

```
life      0.02
evolve    0.01
organism  0.01
.,,
```

β_2

```
gene      0.04
dna       0.02
genetic   0.01
.,,
```

…

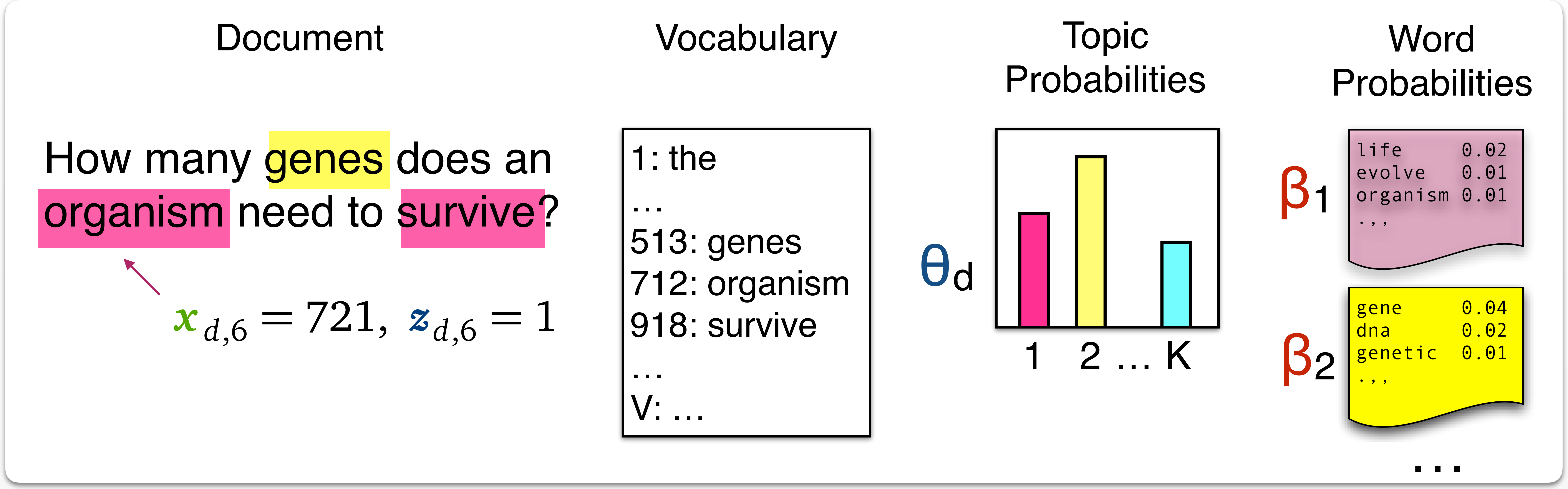
probability of all words $n = 1 \ldots N_d$ in document d (use one-hot trick)

$$p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \prod_{n=1}^{N_d} \prod_{v=1}^{V} p(\boldsymbol{x}_{d,n} = v \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)^{I[\boldsymbol{x}_{d,n} = v]}$$

take log probability, substitute result from previous slide

$$\log p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \sum_{n=1}^{N_d} \sum_{v=1}^{V} I[\boldsymbol{x}_{d,n} = v] \log \left(\sum_{k=1}^{K} \boldsymbol{\beta}_{k,v} \boldsymbol{\theta}_{d,k} \right)$$

Calculating the Likelihood for all Words



Document

How many genes does an organism need to survive?

$\boldsymbol{x}_{d,6} = 721, \ \boldsymbol{z}_{d,6} = 1$

Vocabulary

1: the
…
513: genes
712: organism
918: survive
…
V: …

Topic Probabilities

Word Probabilities

β_1

life	0.02
evolve	0.01
organism	0.01
...	

β_2

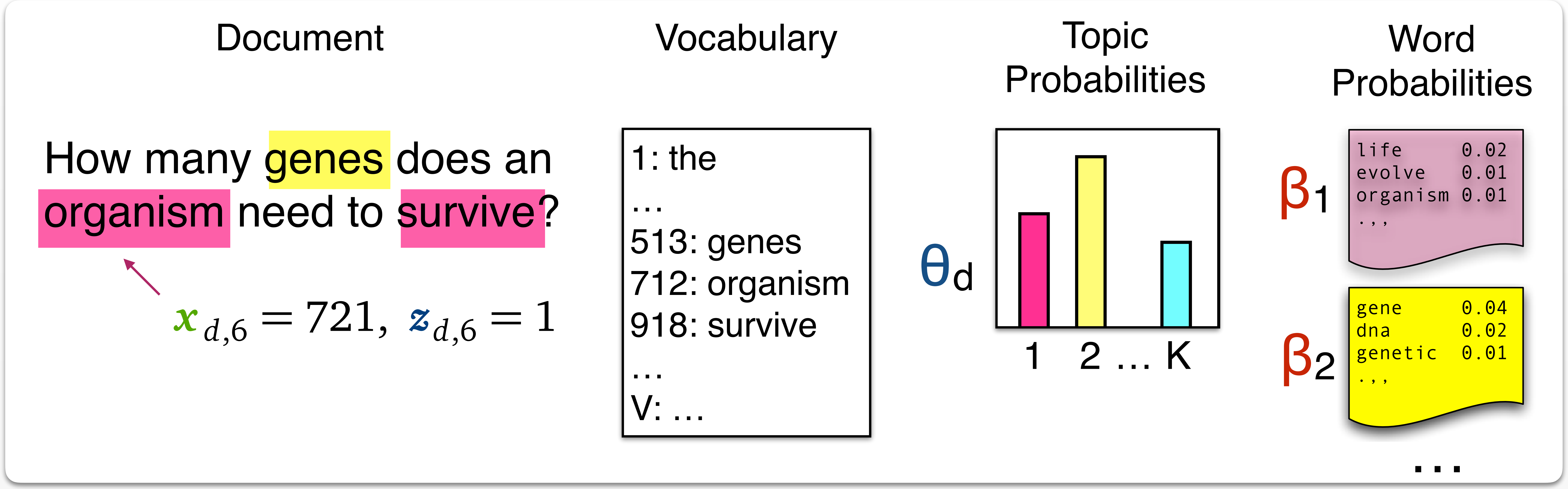
gene	0.04
dna	0.02
genetic	0.01
...	

…

log probability of all words in document *d*

$$\log p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \sum_{n=1}^{N_d} \sum_{v=1}^{V} I[\boldsymbol{x}_{d,n} = v] \log \left(\sum_{k=1}^{K} \boldsymbol{\beta}_{k,v} \boldsymbol{\theta}_{d,k} \right)$$

Calculating the Likelihood for all Words



How many genes does an organism need to survive?

$x_{d,6} = 721, \; z_{d,6} = 1$

1: the
…
513: genes
712: organism
918: survive
…
V: …

```
life     0.02
evolve   0.01
organism 0.01
...
```

```
gene     0.04
dna      0.02
genetic  0.01
...
```

log probability of all words in document *d*

$$\log p(\boldsymbol{x}_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d) = \sum_{v=1}^{V} \sum_{n=1}^{N_d} I[\boldsymbol{x}_{d,n} = v] \log\left(\sum_{k=1}^{K} \boldsymbol{\beta}_{k,v} \boldsymbol{\theta}_{d,k} \right)$$

$$= \boldsymbol{X}_d \log(\boldsymbol{\theta}_d \boldsymbol{\beta})^\top$$

bag-of-word vector

$$\boldsymbol{X}_{d,v} = \sum_{n=1}^{N_d} I[\boldsymbol{x}_{d,n} = v]$$

inner product between bag of word vector,
and log weighted average over topics

Interpretation as Matrix Factorization

Log likelihood

$$\log(p(X_d \mid \beta, \theta_d)) = X_d \log(\theta_d \beta)^\top$$

Bag of Word Vector

$$X_{d,v} = \sum_{n=1}^{N_d} I[x_{d,n} = v]$$

Word Counts

Topic Counts

Topic Word Probabilities

$$\mathbb{E}\left[\underset{\text{(D x V)}}{\begin{bmatrix} X_{1,1} & \dots & X_{1,V} \\ \dots & & \\ X_{D,1} & \dots & X_{D,V} \end{bmatrix}}\right] = \underset{\text{(D x K)}}{\begin{bmatrix} N_1\theta_{1,1} & \dots & N_1\theta_{1,K} \\ \dots & & \\ N_D\theta_{D,1} & \dots & N_D\theta_{D,K} \end{bmatrix}} \underset{\text{(K x V)}}{\begin{bmatrix} \beta_{1,1} & \dots & \beta_{1,V} \\ \dots & & \\ \beta_{K,1} & \dots & \beta_{K,V} \end{bmatrix}}$$

$$\left[\begin{array}{cccccc} \text{stocks} & \text{chairman} & \text{the} & & \text{wins} & \text{game} \\ 2 & 4 & 8 & \dots & 0 & 1 \\ \dots & & & & & \\ 0 & 1 & 7 & \dots & 2 & 3 \end{array}\right] \simeq \left[\begin{array}{ccc} \text{finance} & & \text{Sports} \\ 112 \cdot 0.91 & \dots & 112 \cdot 0.01 \\ \dots & & \\ 234 \cdot 0.02 & \dots & 234 \cdot 0.86 \end{array}\right] \left[\begin{array}{ccc} \text{stocks} & & \text{game} \\ 0.0081 & \dots & 0.0002 \\ \dots & & \\ 0.0001 & \dots & 0.0072 \end{array}\right]$$

Relationship to Latent Semantic Analysis

LSA: Factorize word counts (using PCA)

$$\mathbf{X}\ \text{(V x D)} \qquad \approx \qquad \mathbf{U}\ \text{(V x K)} \qquad \mathbf{Z}\ \text{(K x D)}$$

$$\left(\begin{array}{rccc} \text{stocks:} & 2 & \cdots\cdots & 0 \\ \text{chairman:} & 4 & \cdots\cdots & 1 \\ \text{the:} & 8 & \cdots\cdots & 7 \\ \cdots & \vdots & \cdots\cdots & \vdots \\ \text{wins:} & 0 & \cdots\cdots & 2 \\ \text{game:} & 1 & \cdots\cdots & 3 \end{array}\right) \approx \begin{pmatrix} 0.4 & \cdot\cdot & -0.001 \\ 0.8 & \cdot\cdot & 0.03 \\ 0.01 & \cdot\cdot & 0.04 \\ \vdots & \cdot\cdot & \vdots \\ 0.002 & \cdot\cdot & 2.3 \\ 0.003 & \cdot\cdot & 1.9 \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

Topic Models: Factorize word counts (using mixture model)

$$\mathbb{E}[X^\top]\ \text{(V x D)} \quad = \quad \boldsymbol{\beta}^\top\ \text{(V x K)} \quad \boldsymbol{\theta}^\top\ \text{(K x D)} \quad \boldsymbol{N}\,\boldsymbol{I}\ \text{(D x D)}$$

$$\mathbb{E}[X]\ \text{(D x V)} \quad = \quad \boldsymbol{N}\,\boldsymbol{I}\ \text{(D x D)} \quad \boldsymbol{\theta}\ \text{(D x K)} \quad \boldsymbol{\beta}\ \text{(K x V)}$$

Topic Models: *Summary so far*

Core Idea:

Model documents as ***mixtures*** over topics

Model Parameters:

θ_d Topic probabilities for each document

(K-dimensional vector)

β_k Word probabilities for each topic

(V-dimensional vector)

Relationship to Dimensionality Reduction:

Similar to LSA, but assumes Discrete mixture instead of Gaussian distribution on word counts

Topic Models

Shantanu Jain

NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS
FOUNDED AD 1898

Estimating Parameters

Maximum Likelihood with EM

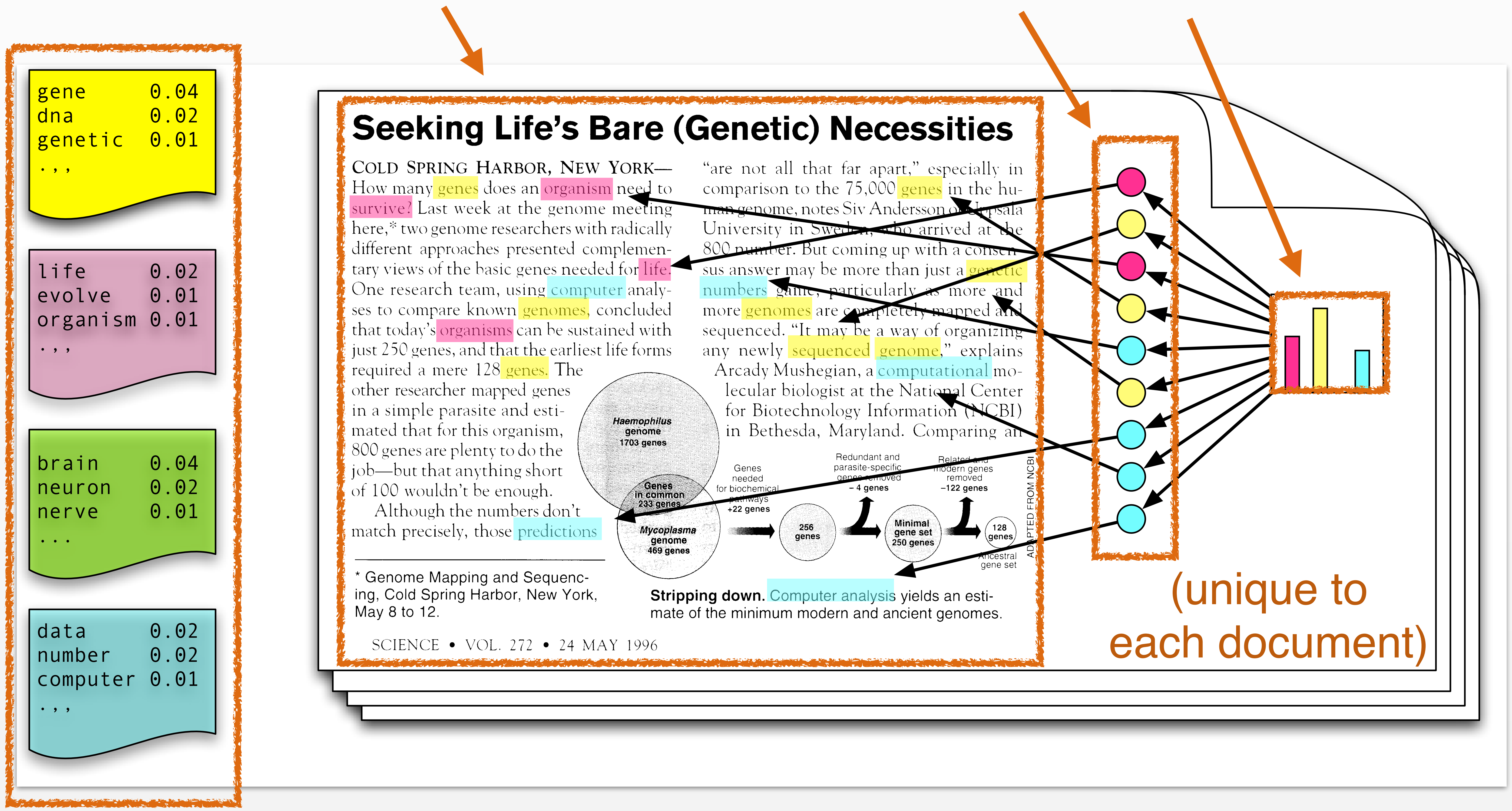
Review: Topic Modeling

β_k: Topics

x_d: Words

z_d: Assignments

θ_d: Topic Proportions



gene	0.04
dna	0.02
genetic	0.01
.,,	

life	0.02
evolve	0.01
organism	0.01
.,,	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
.,,	

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

(unique to each document)

(shared across documents)

$$z_{d,n} \sim \text{Discrete}(\boldsymbol{\theta}_d)$$

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

Review: Interpretation as Matrix Factorization

Log marginal likelihood

$$\log(p(X_d \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)) = X_d \log(\boldsymbol{\theta}_d \boldsymbol{\beta})^\top$$

Bag of Word Vector

$$X_{d,v} = \sum_{n=1}^{N_d} I[x_{d,n} = v]$$

Word Counts

Topic Counts

Topic Word Probabilities

$$\mathbb{E}\left[\underset{\text{(D x V)}}{\begin{bmatrix} X_{1,1} & \dots & X_{1,V} \\ \dots & & \\ X_{D,1} & \dots & X_{D,V} \end{bmatrix}}\right] = \underset{\text{(D x K)}}{\begin{bmatrix} N_1 \boldsymbol{\theta}_{1,1} & \dots & N_1 \boldsymbol{\theta}_{1,K} \\ \dots & & \\ N_D \boldsymbol{\theta}_{D,1} & \dots & N_D \boldsymbol{\theta}_{D,K} \end{bmatrix}} \underset{\text{(K x V)}}{\begin{bmatrix} \boldsymbol{\beta}_{1,1} & \dots & \boldsymbol{\beta}_{1,V} \\ \dots & & \\ \boldsymbol{\beta}_{K,1} & \dots & \boldsymbol{\beta}_{K,V} \end{bmatrix}}$$

$$\left[\begin{array}{cccccc} \text{stocks} & \text{chairman} & \text{the} & & \text{wins} & \text{game} \\ 2 & 4 & 8 & \dots & 0 & 1 \\ \dots & & & & & \\ 0 & 1 & 7 & \dots & 2 & 3 \end{array}\right] \simeq \left[\begin{array}{ccc} \text{“sports”} & & \text{“finance”} \\ 112 \cdot 0.91 & \dots & 112 \cdot 0.01 \\ \dots & & \\ 234 \cdot 0.02 & \dots & 234 \cdot 0.86 \end{array}\right] \left[\begin{array}{ccc} \text{stocks} & & \text{game} \\ 0.0081 & \dots & 0.0002 \\ \dots & & \\ 0.0001 & \dots & 0.0072 \end{array}\right]$$

Relationship to Latent Semantic Analysis

LSA: Factorize matrix of word counts (using PCA)

$$\mathbf{X}\ \text{(V x D)} \approxeq \mathbf{U}\ \text{(V x K)} \quad \mathbf{Z}\ \text{(K x D)}$$

$$\left(\begin{array}{rlcl} \text{stocks:} & 2 & \cdots & 0 \\ \text{chairman:} & 4 & \cdots & 1 \\ \text{the:} & 8 & \cdots & 7 \\ \cdots & \vdots & \cdots & \vdots \\ \text{wins:} & 0 & \cdots & 2 \\ \text{game:} & 1 & \cdots & 3 \end{array}\right) \approxeq \left(\begin{array}{rcl} 0.4 & \cdot\cdot & -0.001 \\ 0.8 & \cdot\cdot & 0.03 \\ 0.01 & \cdot\cdot & 0.04 \\ \vdots & \cdot\cdot & \vdots \\ 0.002 & \cdot\cdot & 2.3 \\ 0.003 & \cdot\cdot & 1.9 \end{array}\right) \left(\begin{array}{ccc} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{array}\right)$$

LSA: Assume Gaussian distribution

Topic Models: Assume mixture of Discrete distributions

Estimating Model Parameters

Question: How can we estimate β_k and θ_d?

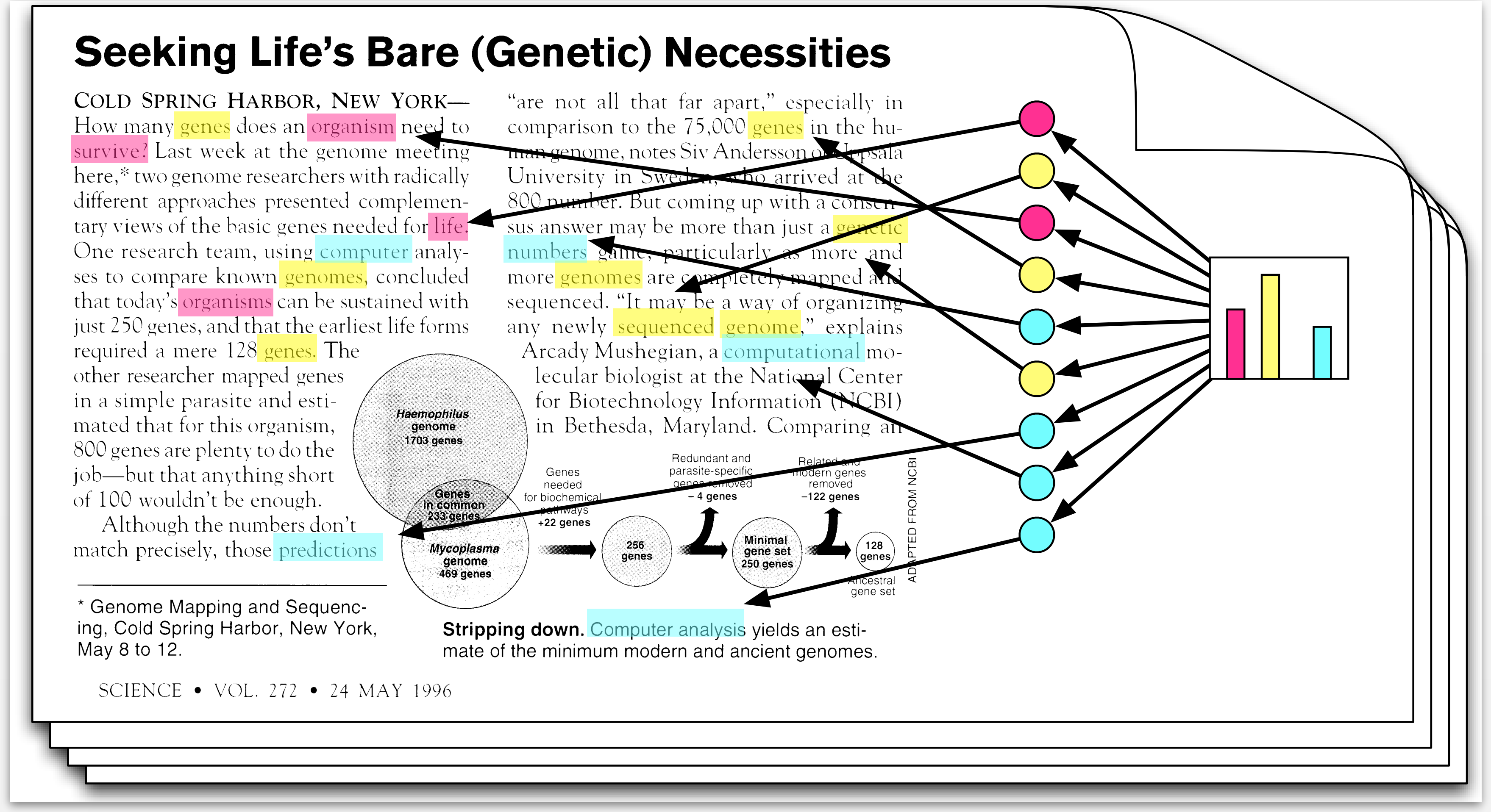
1. Expectation Maximization
   (this video)
2. Variational Inference
   (will discuss at a high level)
3. Gibbs Sampling
   (not in this module)

PLSI/PLSA*: EM for Topic Models

Generative Model

$$z_{d,n} \sim \text{Discrete}(\theta_d)$$

$$x_{d,n} \mid z_{d,n}=k \sim \text{Discrete}(\beta_k)$$



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

E-step: Update assignments

Calculate probability that word *n* in document *d* belongs to topic *k*

$$\phi_{d,n,k} = p(z_{d,n}=k \mid x_{d,n}, \beta, \theta_d)$$

M-step: Update parameters

Use assignment probabilities ϕ_d to update topics assignment probabilities θ_d and topic word probabilities β_k

*(Probabilistic Latent Semantic Indexing, a.k.a. Probabilistic Latent Semantic Analysis)

PLSI/PLSA: E-step

$$\boldsymbol{\phi}_{d,n,k} = p(\boldsymbol{z}_{d,n}=k \mid \boldsymbol{x}_{d,n}=v, \boldsymbol{\beta}, \boldsymbol{\theta}_d)$$

$$= \frac{p(\boldsymbol{x}_{d,n}=v,\ \boldsymbol{z}_{d,n}=k \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)}{p(\boldsymbol{x}_{d,n}=v \mid \boldsymbol{\beta}, \boldsymbol{\theta}_d)}$$

(Apply Bayes' Rule)

$$= \frac{\boldsymbol{\theta}_{d,k}\boldsymbol{\beta}_{k,v}}{\sum_{l=1}^{K} \boldsymbol{\theta}_{d,l}\boldsymbol{\beta}_{l,v}}$$

(Substitute results from previous slides)

General Form, with One-hot Indexing Trick

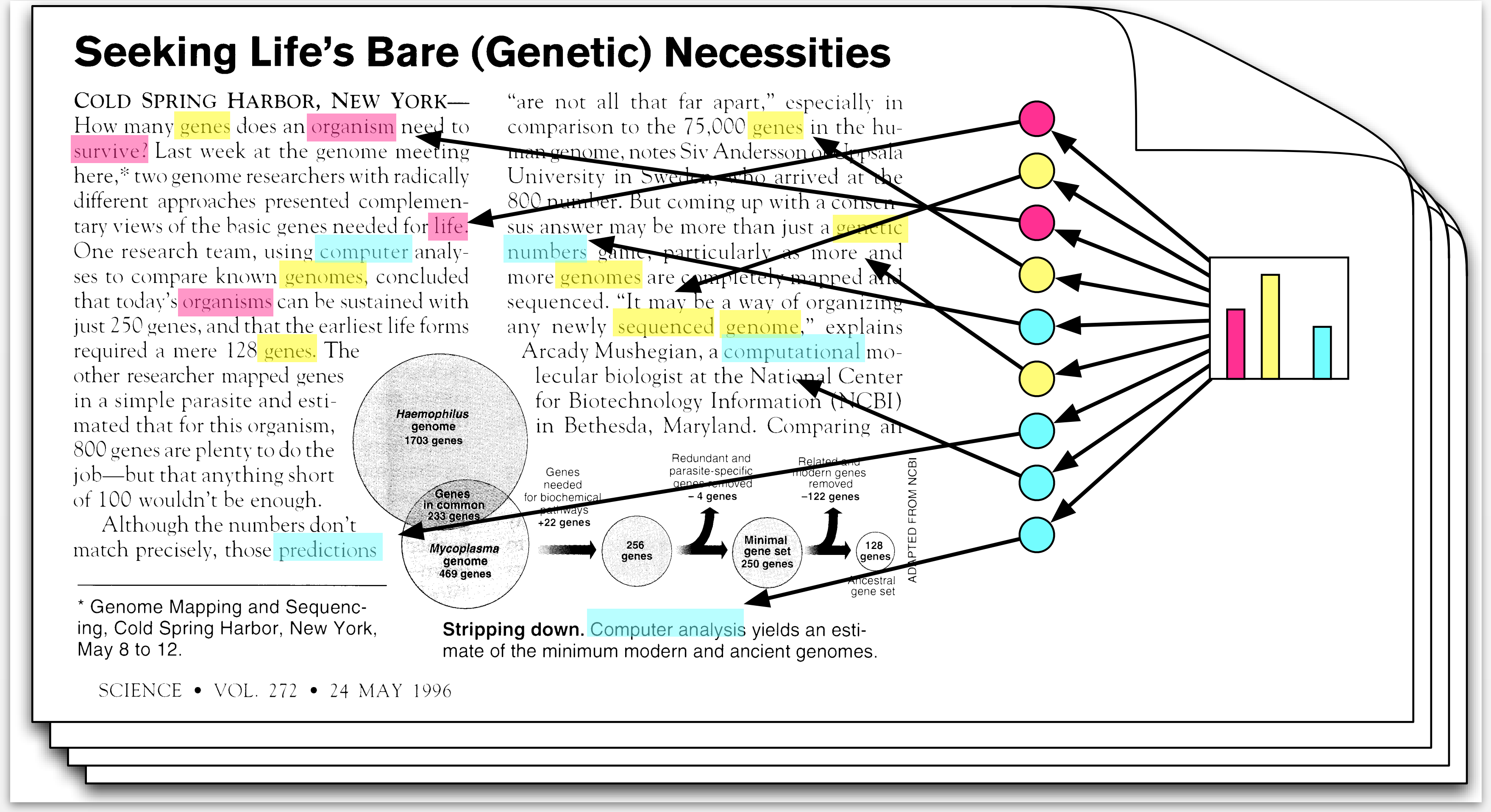
$$\boldsymbol{\phi}_{d,n,k} = \frac{\boldsymbol{\theta}_{d,k}\left(\sum_{v=1}^{V} \boldsymbol{\beta}_{k,v} I[\boldsymbol{x}_{d,n} = v]\right)}{\sum_{l=1}^{K} \boldsymbol{\theta}_{d,l}\left(\sum_{v=1}^{V} \boldsymbol{\beta}_{l,v} I[\boldsymbol{x}_{d,n} = v]\right)}$$

PLSI/PLSA*: EM for Topic Models

Generative Model

$$z_{d,n} \sim \text{Discrete}(\theta_d)$$

$$x_{d,n} \mid z_{d,n}=k \sim \text{Discrete}(\beta_k)$$



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

E-step: Update assignments

$$\phi_{d,n,k} = p(z_{d,n}=k \mid x_{d,n}=v, \beta, \theta_d)$$

$$= \frac{\theta_{d,k}\left(\sum_{v=1}^{V} \beta_{k,v} I[x_{d,n}=v]\right)}{\sum_{l=1}^{K} \theta_{d,l}\left(\sum_{v=1}^{V} \beta_{l,v} I[x_{d,n}=v]\right)}$$

M-step: Update parameters

Use assignment probabilities ϕ_d to update topics assignment probabilities θ_d and topic word probabilities β_k

*(Probabilistic Latent Semantic Indexing, a.k.a. Probabilistic Latent Semantic Analysis)

PLSI/PLSA: M-Step

Idea: Compute (expected) sufficient statistics

$\phi_{d,n,k}$	Probability that word *n* in document *d* belongs to topic *k*
$N^{\theta}_{d,k} = \sum_{n=1}^{N_d} \phi_{d,n,k}$	Number of words in document *d* that belong to topic *k*
$N^{\beta}_{k,v} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,k} I[x_{d,n} = v]$	Number of times word *v* appears in topic *k* (across *all* documents in corpus)

M-Step: Update parameters using sufficient statistics

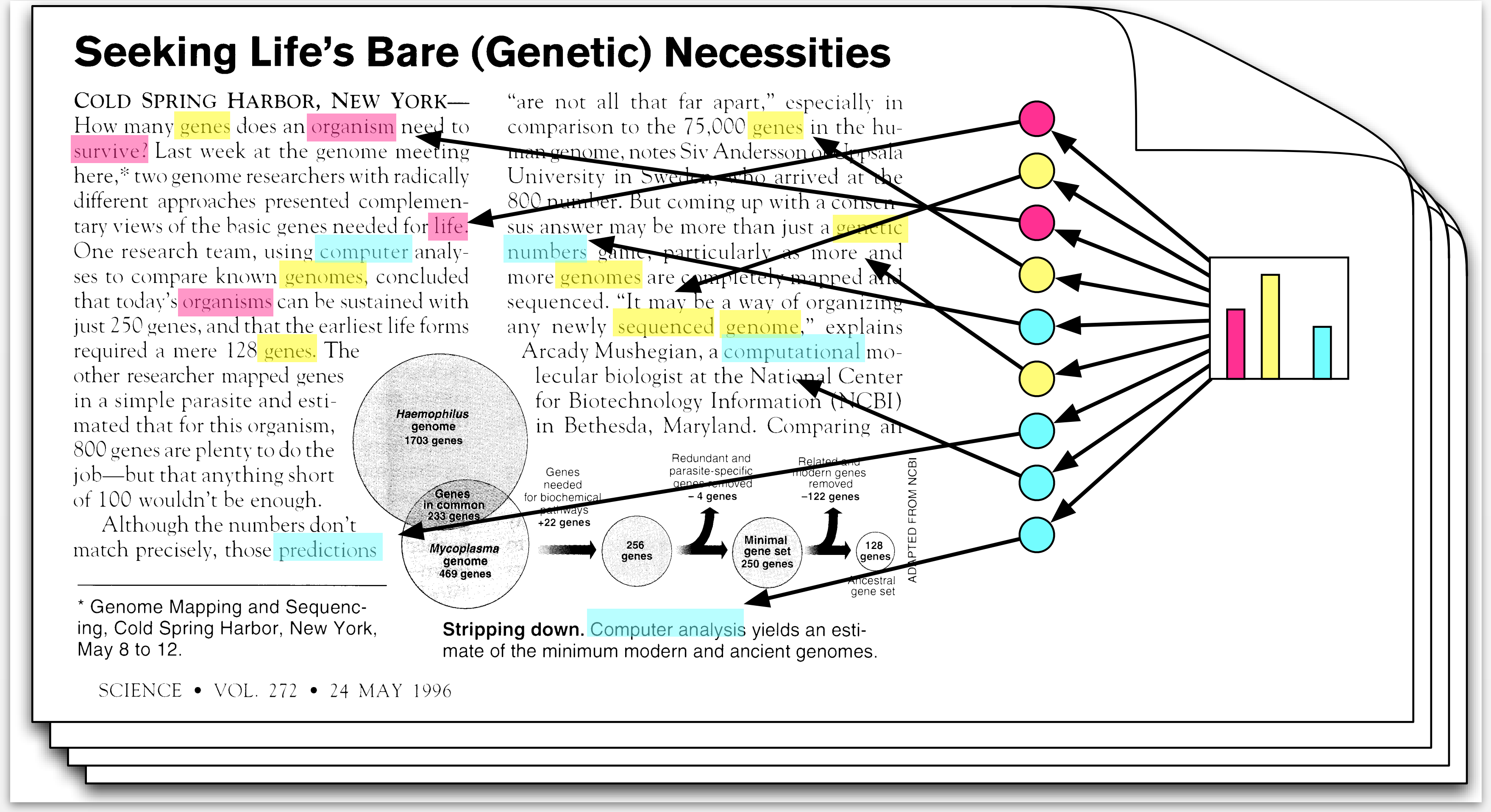
$\theta_{d,k} = \frac{N^{\theta}_{d,k}}{N_d}$	Fraction of topic *k* in document *d*
$\beta_{k,v} = \frac{N^{\beta}_{k,v}}{\sum_{d=1}^{D} N^{\theta}_{d,k}}$	Fraction of word *v* in topic *k*

PLSI/PLSA*: EM for Topic Models

Generative Model

$$z_{d,n} \sim \text{Discrete}(\theta_d)$$

$$x_{d,n} \mid z_{d,n} = k \sim \text{Discrete}(\beta_k)$$



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

E-step: Update assignments

$$\phi_{d,n,k} = p(z_{d,n} = k \mid x_{d,n} = v, \beta, \theta_d)$$

$$= \frac{\theta_{d,k}\left(\sum_{v=1}^{V} \beta_{k,v} I[x_{d,n} = v]\right)}{\sum_{l=1}^{K} \theta_{d,l}\left(\sum_{v=1}^{V} \beta_{l,v} I[x_{d,n} = v]\right)}$$

M-step: Update parameters

$$\beta_{k,v} = \frac{N^{\beta}_{k,v}}{\sum_{d=1}^{D} N^{\theta}_{d,k}} \qquad N^{\beta}_{k,v} = \sum_{d=1}^{D}\sum_{n=1}^{N_d} \phi_{d,n,k} I[x_{d,n} = v]$$

$$\theta_{d,k} = \frac{N^{\theta}_{d,k}}{N_d} \qquad N^{\theta}_{d,k} = \sum_{n=1}^{N_d} \phi_{d,n,k}$$

*(Probabilistic Latent Semantic Indexing, a.k.a. Probabilistic Latent Semantic Analysis)

NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS
FOUNDED AD 1898

Topic Models

Jan-Willem van de Meent

NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS
FOUNDED AD 1898

Latent Dirichlet Allocation

Topic Models with Dirichlet Priors

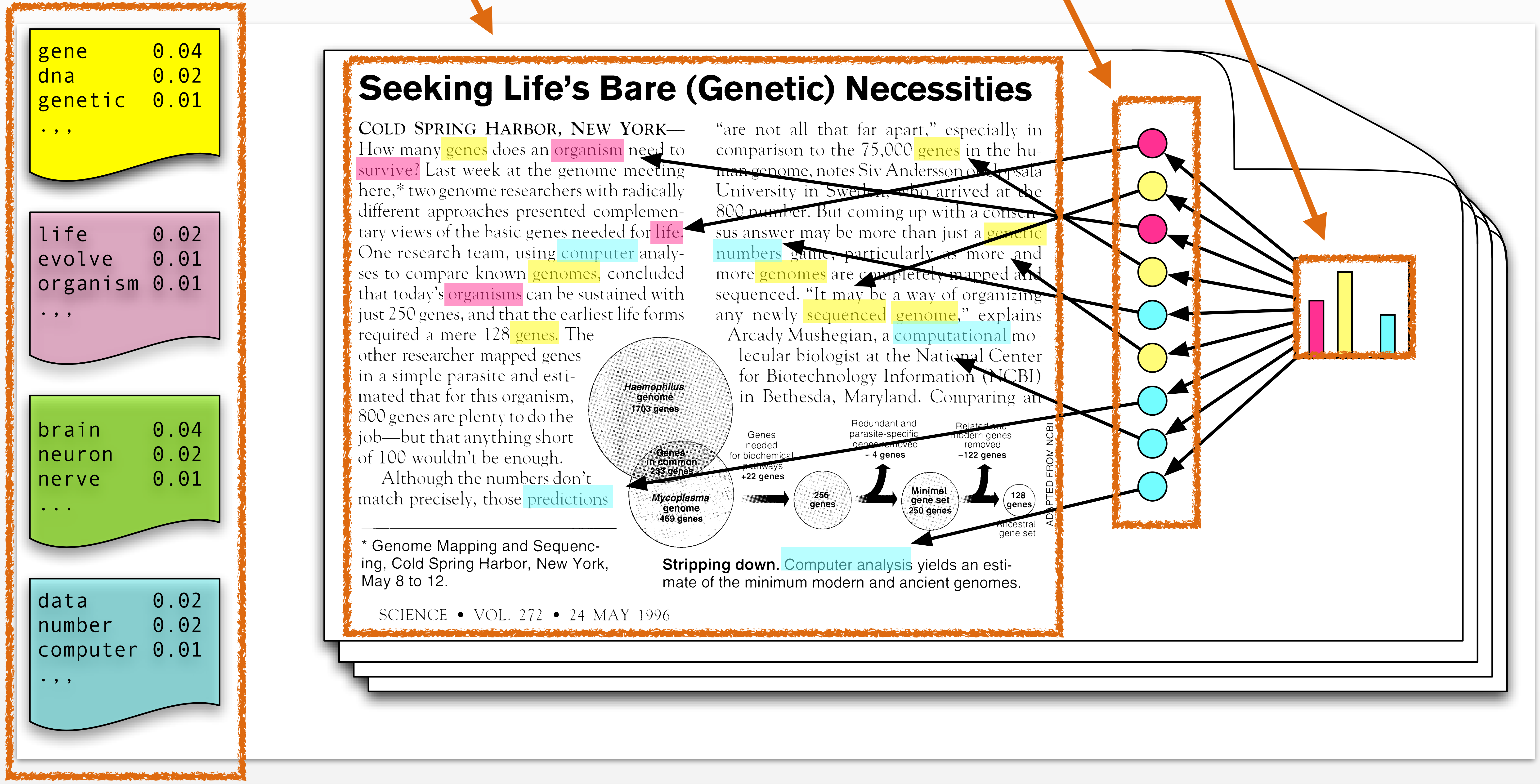
Review: Topic Modeling with PLSA/PLSI

β_k: Topics (shared)

x_d: Words

z_d: Assignments (document-specific)

θ_d: Topic Proportions



gene	0.04
dna	0.02
genetic	0.01
.,,	

life	0.02
evolve	0.01
organism	0.01
.,,	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
.,,	

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

$$z_{d,n} \sim \text{Discrete}(\theta_d)$$

$$x_{d,n} \mid z_{d,n}=k \sim \text{Discrete}(\beta_k)$$

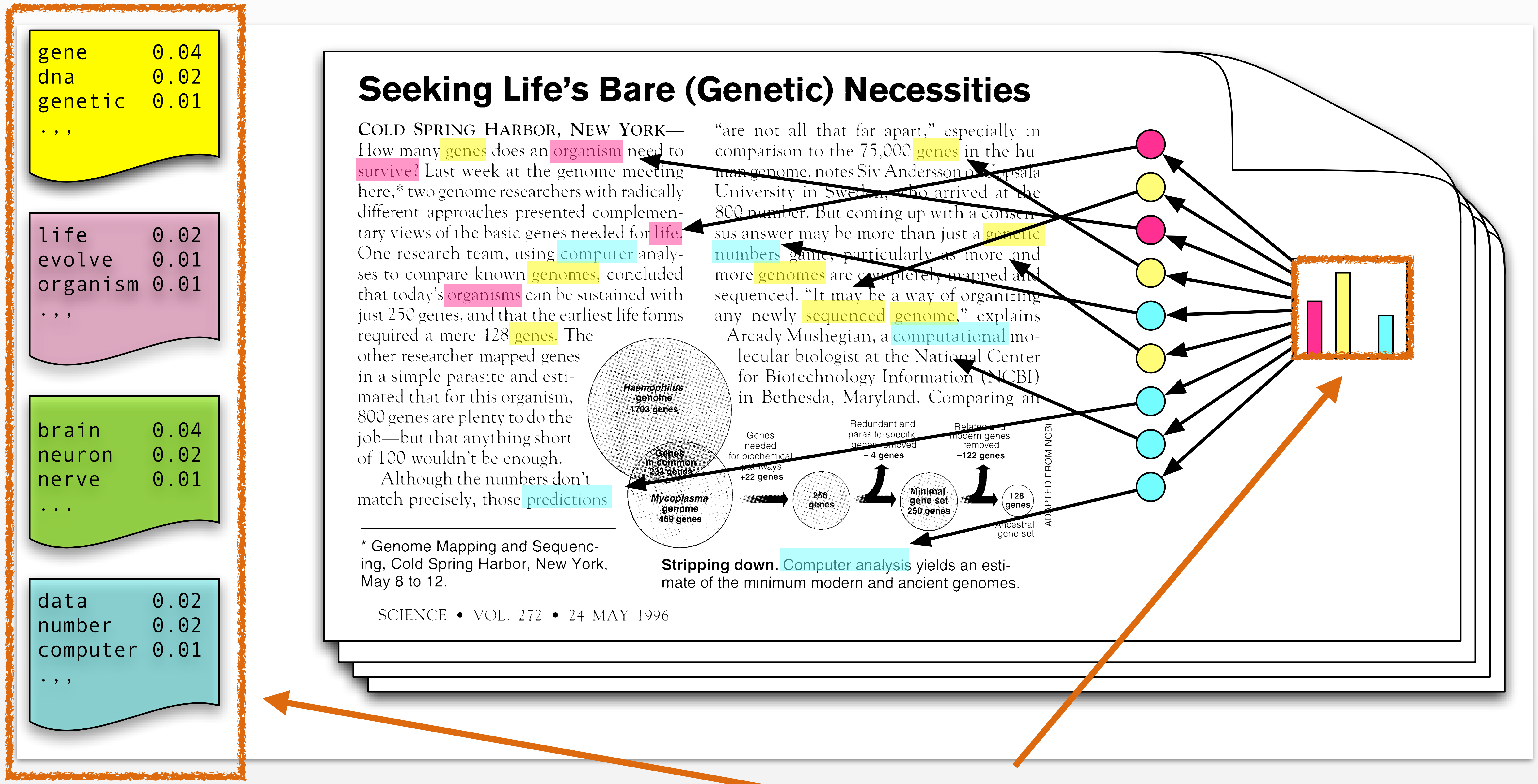
LDA: Add *Dirichlet* Priors

β_k: Topics
(shared)

x_d: Words

z_d: Assignments
(document-specific)

θ_d: Topic Proportions



gene	0.04
dna	0.02
genetic	0.01
.,,	

life	0.02
evolve	0.01
organism	0.01
.,,	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
.,,	

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

$$z_{d,n} \sim \text{Discrete}(\boldsymbol{\theta}_d)$$

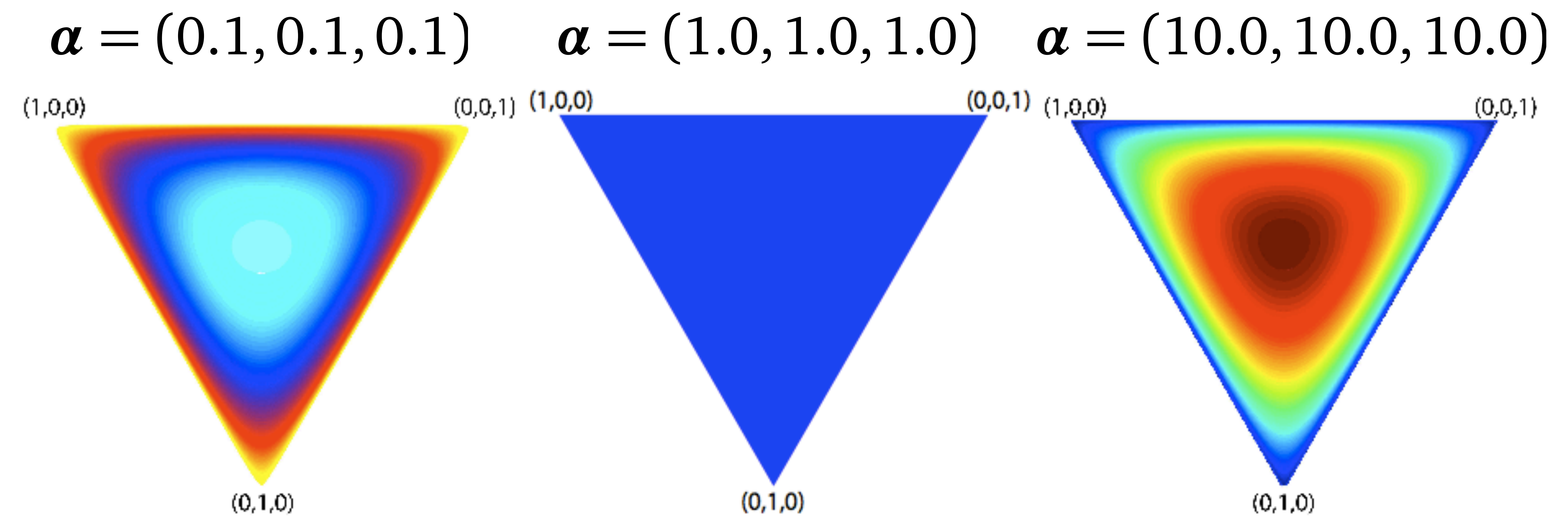
$$x_{d,n} \mid z_{d,n} = k \sim \text{Discrete}(\boldsymbol{\beta}_k)$$

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}(\boldsymbol{\eta}_k)$$

Review: Dirichlet Distribution

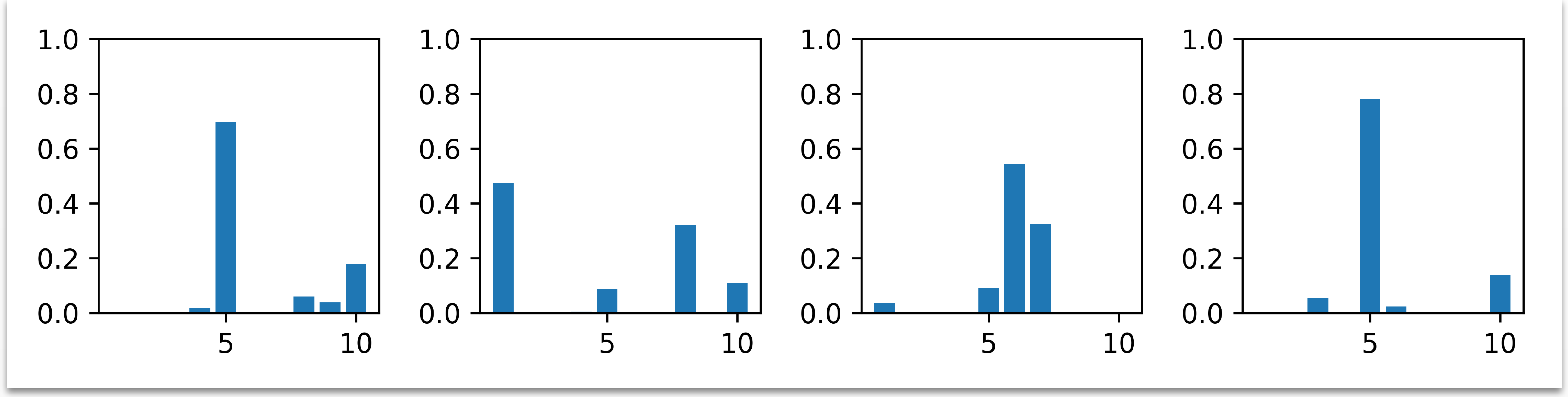
$$p(\boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad B(\boldsymbol{\alpha}) := \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$



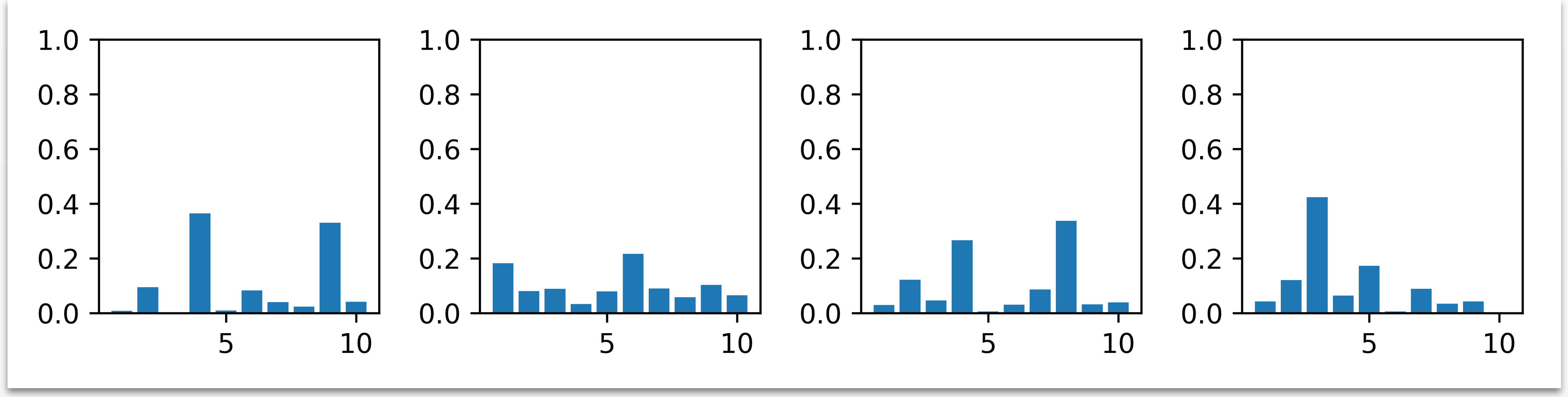
$\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$ $\boldsymbol{\alpha} = (1.0, 1.0, 1.0)$ $\boldsymbol{\alpha} = (10.0, 10.0, 10.0)$

Review: Dirichlet Distribution

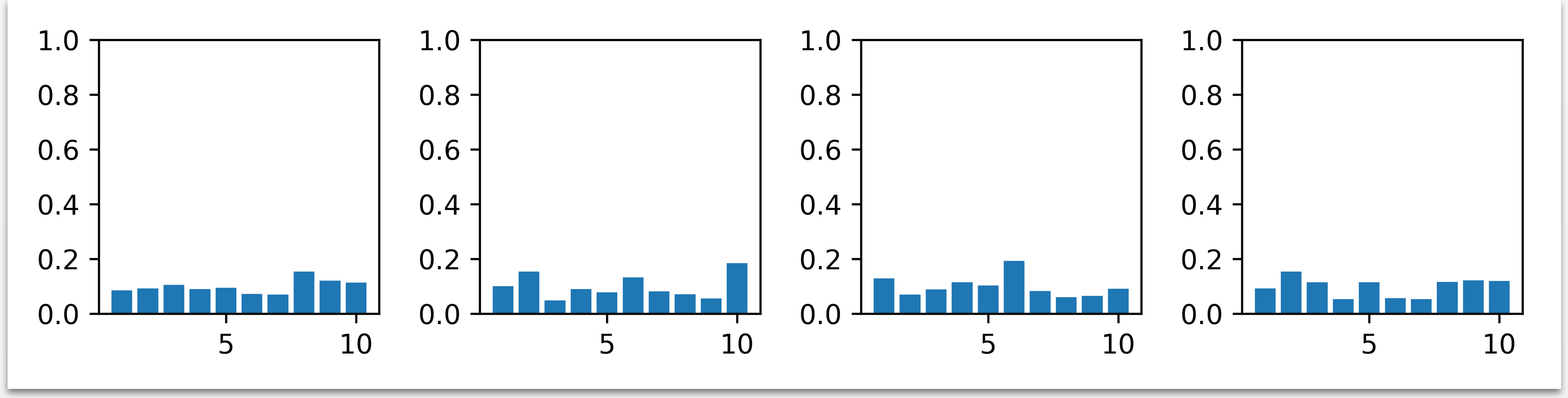
$\alpha_k = 0.1$



$\alpha_k = 1.0$



$\alpha_k = 10.0$



LDA: $\alpha_k = 0.001$ – Enforces Sparsity of Topic Weights θ_d

LDA: *Summary so far*

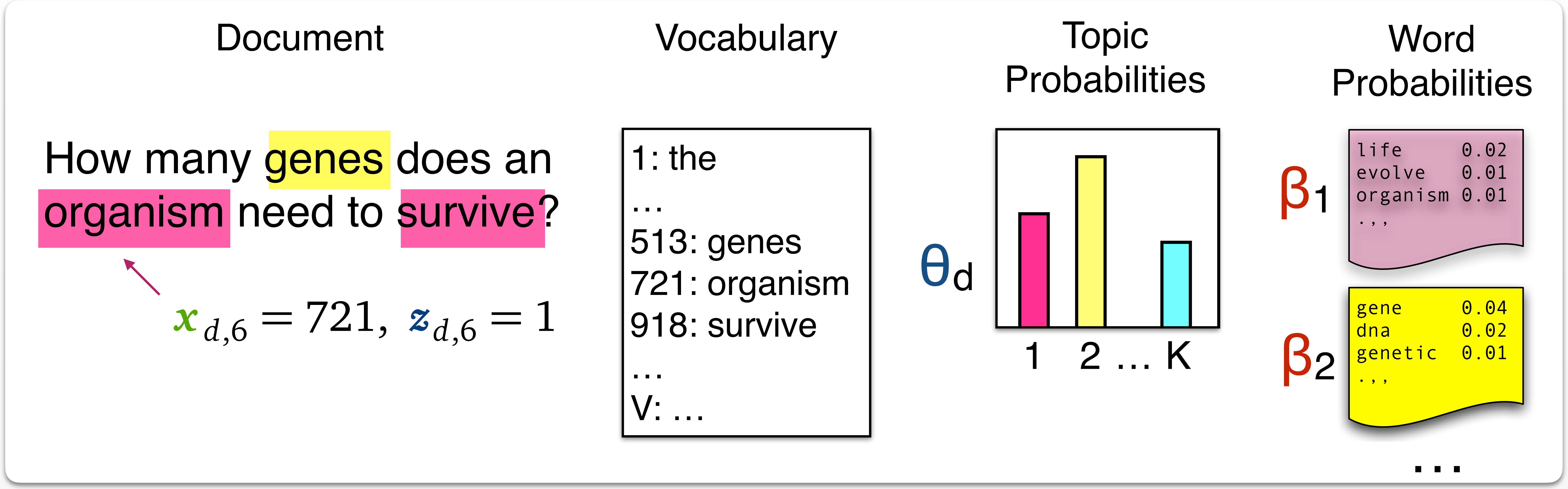
- **Idea:** Model documents as ***mixtures*** over topics
- **Model parameters:**
  θ_d Topic probabilities for each document
  (K-dimensional vector for each document)
  β_k Word probabilities for each topic
  (V-dimensional vector for each topic)
- **Interpretation Dimensionality Reduction:**
  Similar to LSA, but assumes Discrete mixture instead of Gaussian distribution on word counts
- **Dirichlet Priors:** Enforce sparsity, associate a small number of topics which each document

Estimating Model Parameters

Question: How can we estimate β_k and θ_d?

1. Expectation Maximization
   (previous video)
2. Variational Inference
   (high level)
3. Gibbs Sampling
   (not in this module)

Estimating the Parameters



How many genes does an organism need to survive?

$x_{d,6} = 721, \ z_{d,6} = 1$

1: the
…
513: genes
721: organism
918: survive
…
V: …

β_1

```
life     0.02
evolve   0.01
organism 0.01
.,,
```

β_2

```
gene     0.04
dna      0.02
genetic  0.01
.,,
```

Maximum Likelihood: $\max_{\theta,\beta} \ \log p(x \mid \theta, \beta)$

Maximum a Posteriori: $\max_{\theta,\beta} \ \log p(\theta, \beta \mid x)$

Review: Conjugate Priors for Coin Flips

Likelihood

$$\mathrm{Bern}(x \mid \mu) = \mu^{N_1}(1-\mu)^{N_0}$$

Sufficient Statistics

$$N_1 = \sum_{n=1}^{N} x_n, \quad N_0 = N - N_1$$

Conjugate Prior

$$\mathrm{Beta}(\mu \mid a, b) = \frac{1}{B(a,b)}\mu^{a-1}(1-\mu)^{b-1}$$

Posterior

$$\mathrm{Beta}(\mu \mid N_1 + a, N_0 + b)$$

MAP Estimate

$$\mu^* = \frac{N_1 + a - 1}{N + a + b - 2}$$

Generalization: Dirichlet and Discrete

Likelihood

$$p(z \mid \theta) = \prod_n p(z_n \mid \theta) = \prod_k \theta_k^{N_k}$$

Sufficient Statistics

$$N_k = \sum_n I[z_n = k]$$

Conjugate Prior

$$\mathrm{Dir}(\theta \mid \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

Posterior

$$\mathrm{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

MAP Estimate

$$\theta_k^* = \frac{N_k + \alpha_k - 1}{\sum_k N_k + \sum_k \alpha_k - K}$$

MAP estimation for LDA with EM

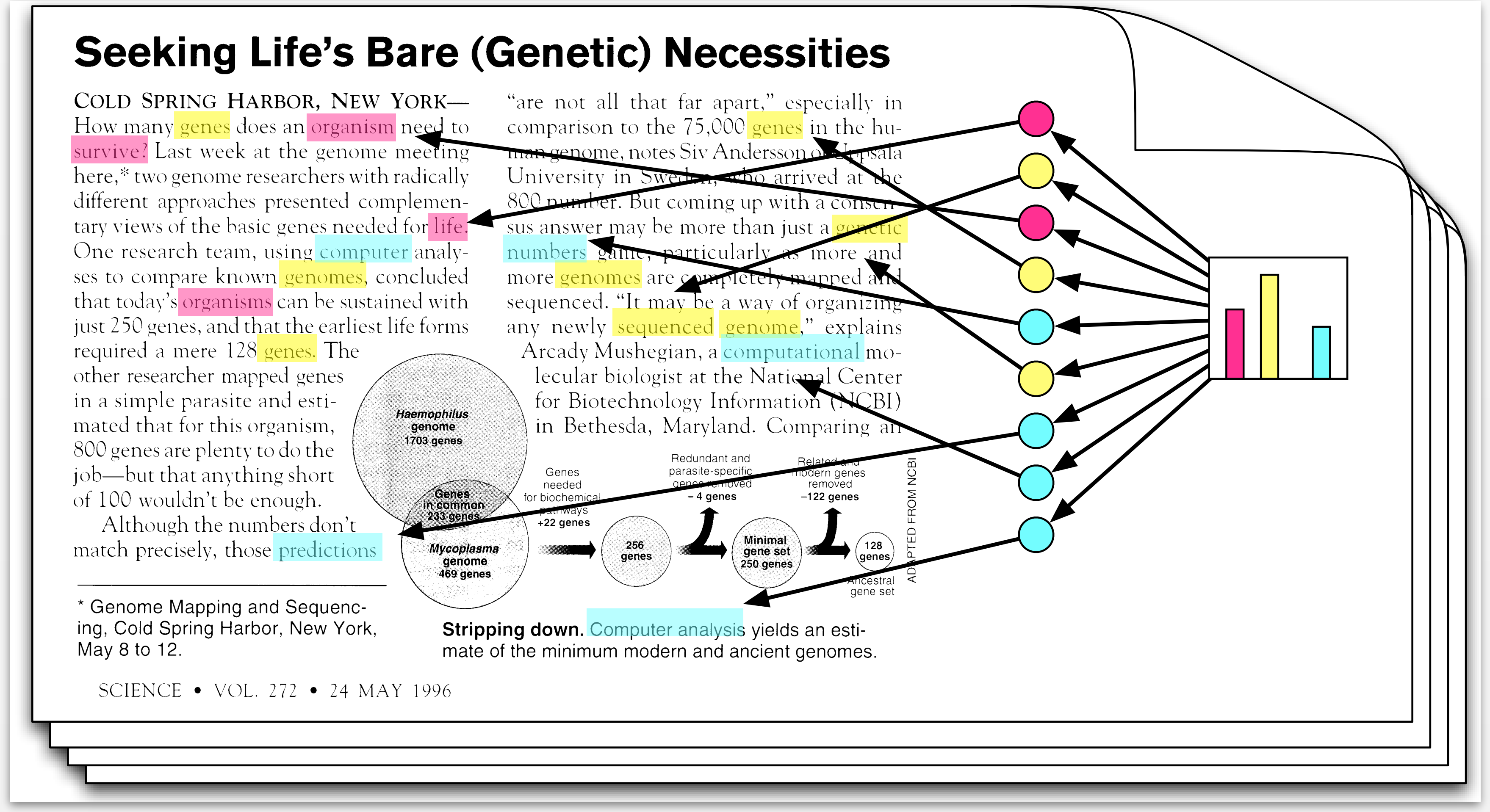
Generative Model

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}(\boldsymbol{\eta}_k)$$

$$z_{d,n} \sim \text{Discrete}(\boldsymbol{\theta}_d)$$

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



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

E-step: Update assignments

$$\phi_{d,n,k} = p(z_{d,n} = k \mid x_{d,n} = v, \boldsymbol{\beta}, \boldsymbol{\theta}_d)$$

$$= \frac{\theta_{d,k}\left(\sum_{v=1}^{V} \beta_{k,v} I[x_{d,n} = v]\right)}{\sum_{l=1}^{K} \theta_{d,l}\left(\sum_{v=1}^{V} \beta_{l,v} I[x_{d,n} = v]\right)}$$

M-step: Update parameters

$$\beta_{k,v} = \frac{N^{\beta}_{k,v} + \eta_{k,v} - 1}{\sum_{d=1}^{D} N^{\theta}_{d,k} + \sum_{v} \eta_{k,v} - V}$$

$$\theta_{d,k} = \frac{N^{\theta}_{d,k} + \alpha_k - 1}{N_d + \sum_k \alpha_k - K}$$

(not used in practice; requires $\alpha_k > 1$ and $\eta_{kv} > 1$)

Variational Expectation Maximization (high-level)

Idea: Approximate $p(z, \theta, \beta \mid x)$ with $q(z, \theta, \beta)$

$$\phi, \gamma, \lambda = \underset{\phi, \gamma, \lambda}{\operatorname{argmin}} \ \mathrm{KL}(q(z, \theta, \beta) \,||\, p(z, \theta, \beta \mid x))$$

$$q(z, \theta, \beta) = \underbrace{q(z; \phi)}_{\text{Discrete}} \, \underbrace{q(\theta; \gamma)}_{\text{Dirichlet}} \, \underbrace{q(\beta; \lambda)}_{\text{Dirichlet}}$$

Variational E-step: Update ϕ

$$\phi_{d,n,k} = \exp\left(\mathbb{E}_q \left[\log \theta_{d,k} + \sum_{v=1}^{V} I[x_{d,n} = v] \log \beta_{k,v} \right] \right)$$

(won't derive this – but can be computed in closed form)

Variational M-step: Update γ and λ

$$\gamma_{d,k} = \alpha_k + N^{\theta}_{d,k} \qquad \lambda_{k,v} = \eta_{k,v} + N^{\beta}_{k,v}$$

(analogous to MAP estimation – need to know this)

EM vs. Variational EM

EM $$\boldsymbol{\theta}, \boldsymbol{\beta} = \underset{\boldsymbol{\theta}, \boldsymbol{\beta}}{\operatorname{argmax}} \; \log p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\beta})$$

E-step: $$\boldsymbol{\phi}_{d,n,k} \propto \boldsymbol{\theta}_{d,k} \left(\sum_{v=1}^{V} \boldsymbol{\beta}_{k,v} I[\boldsymbol{x}_{d,n} = v] \right)$$

M-step: $$\boldsymbol{\theta}_{d,k} = \frac{N^{\theta}_{d,k}}{N_d} \qquad \boldsymbol{\beta}_{k,v} = \frac{N^{\beta}_{k,v}}{\sum_{d=1}^{D} N^{\theta}_{d,k}}$$

Variational EM $$\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda} = \underset{\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}}{\operatorname{argmin}} \; \mathrm{KL}(q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) \,||\, p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{x}))$$

E-step: $$\boldsymbol{\phi}_{d,n,k} = \exp\left(\mathbb{E}_q \left[\log \boldsymbol{\theta}_{d,k} + \sum_{v=1}^{V} I[x_{d,n} = v] \log \boldsymbol{\beta}_{k,v} \right] \right)$$

M-step: $$\boldsymbol{\gamma}_{d,k} = \boldsymbol{\alpha}_k + N^{\theta}_{d,k} \qquad \boldsymbol{\lambda}_{k,v} = \boldsymbol{\eta}_{k,v} + N^{\beta}_{k,v}$$

EM vs. Variational EM

Commonalities: Both compute sufficient statistics

$\boldsymbol{\phi}_{d,n,k}$	Probability that word *n* in document *d* belongs to topic *k*
$N^{\theta}_{d,k}$	Number of words in document *d* that belong to topic *k*
$N^{\beta}_{k,v}$	Number of times word *v* appears in topic *k* (across *all* documents in corpus)

Differences: Point estimates vs Distributions

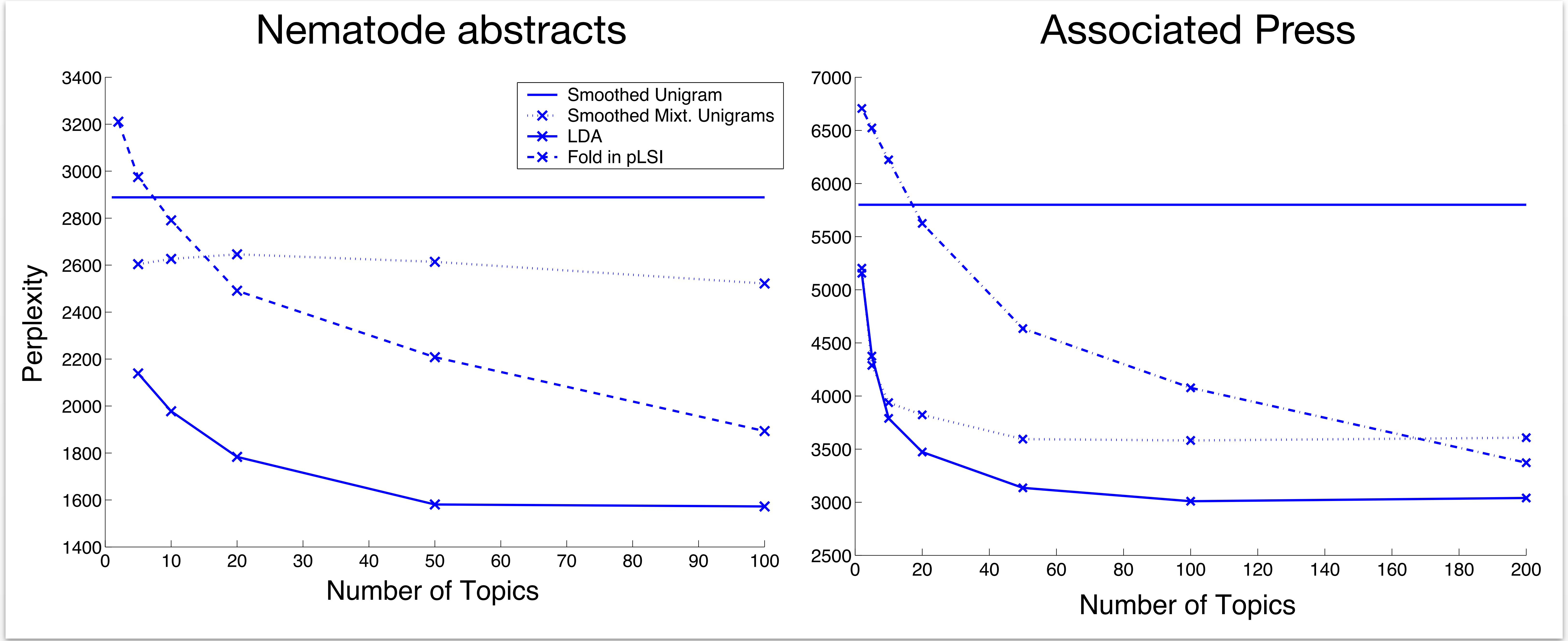
EM: Computes most likely values for parameters

$\boldsymbol{\theta}_{d,k}$	Fraction of words in document *d* for topic *k*
$\boldsymbol{\beta}_{k,v}$	Fraction of words in topic *k* for vocabulary entry *v*

Variational EM: Estimate *Posterior* over Parameters

$q(\boldsymbol{\theta}_d;\boldsymbol{\gamma}_d)$	Approximation of topic distribution for document *d*
$q(\boldsymbol{\beta}_k;\boldsymbol{\lambda}_k)$	Approximation of word distribution for topic *k*

Performance Metric: Perplexity



$$\text{Perplexity} = \exp\left[-\frac{1}{D'}\sum_{d=1}^{D'}\frac{1}{N_d}\log(x'_d \mid x, \alpha, \eta)\right] \qquad \{x'_1, \ldots, x'_{D'}\}$$

Exponent of per-word log predictive probability

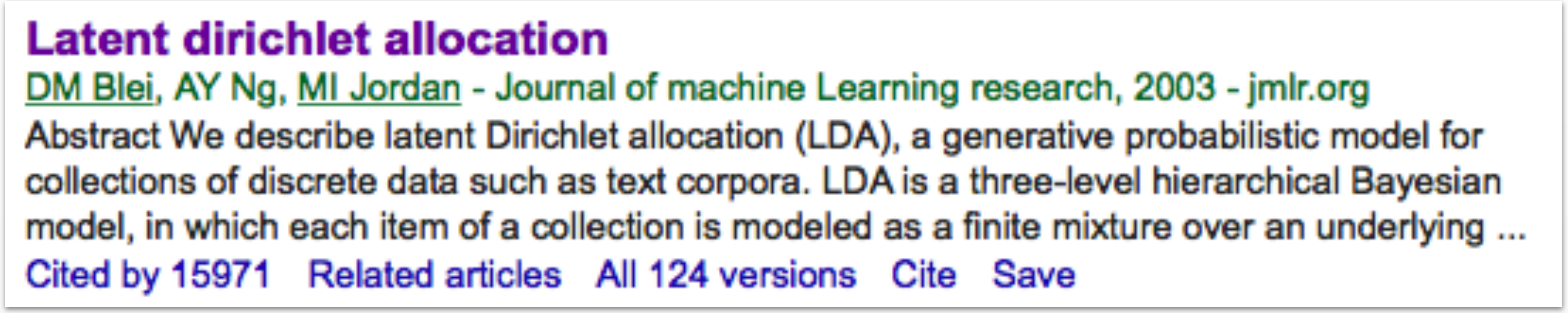
Topic Models

Jan-Willem van de Meent

Extensions of LDA

Borrowing from:
David Blei
(Columbia)

Extensions of LDA



Latent dirichlet allocation

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

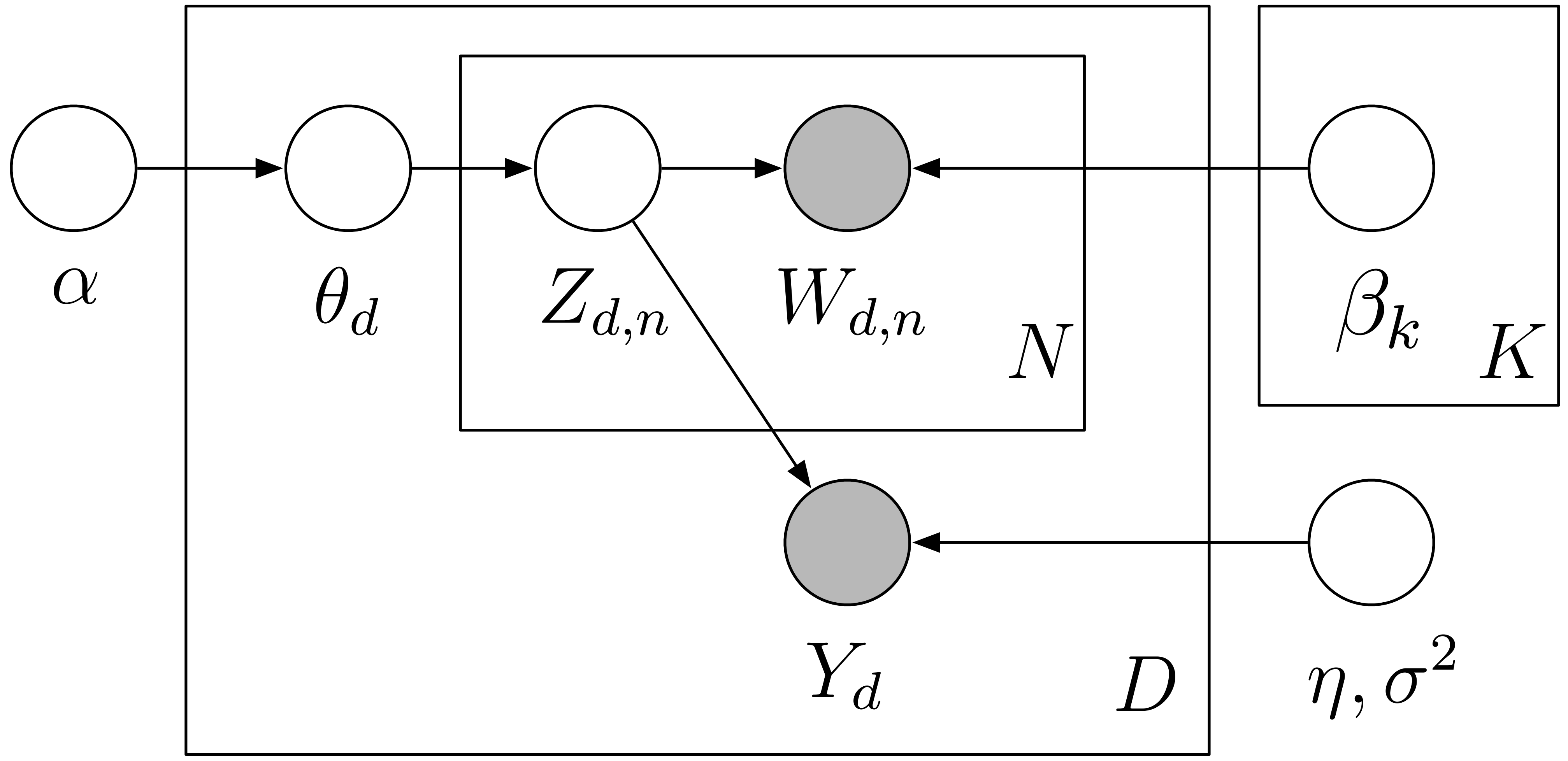
Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying ...

Cited by 15971 Related articles All 124 versions Cite Save

Reasons for popularity of LDA:

- Dirichlet prior gives sparser vectors θ_d
- LDA be extended to more sophisticated models

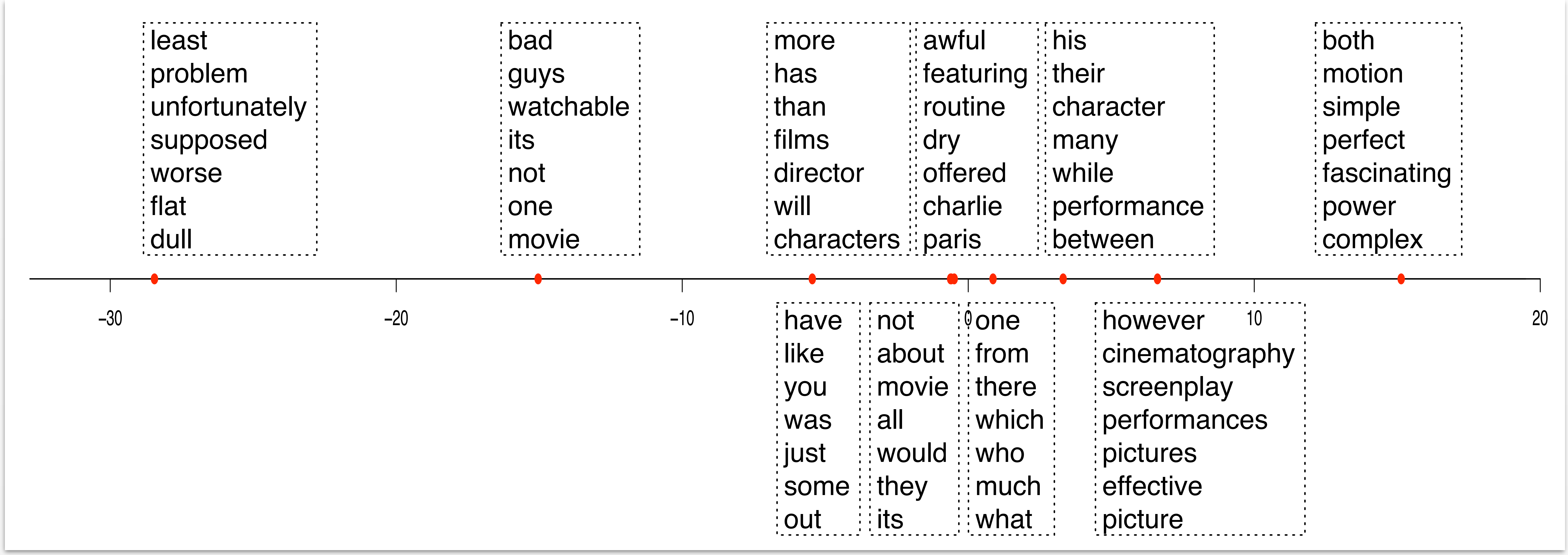
Extensions: Supervised LDA



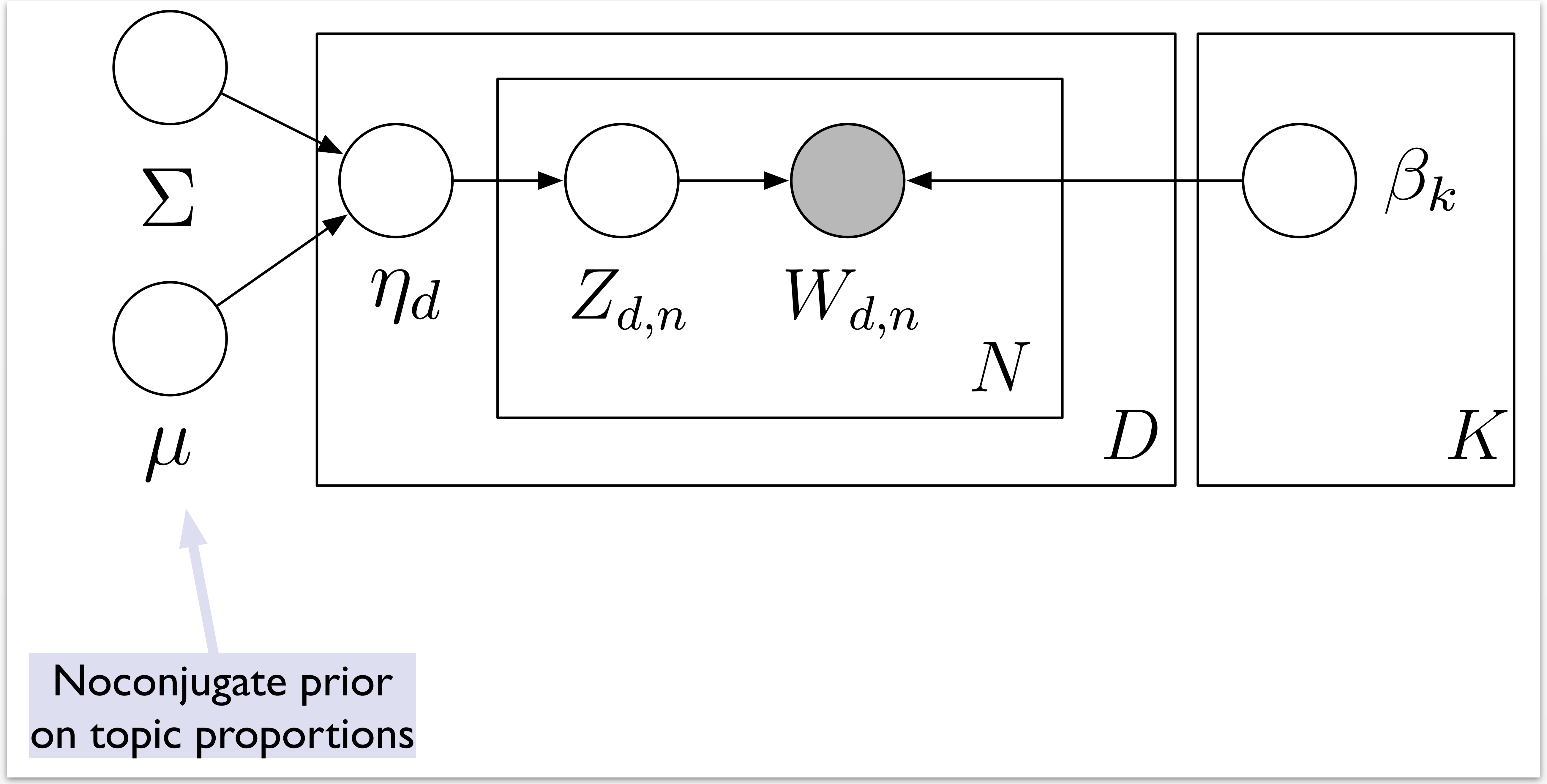
1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
    - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
    - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\big(\eta^\top \bar{z}, \sigma^2\big)$, where

$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Extensions: Supervised LDA

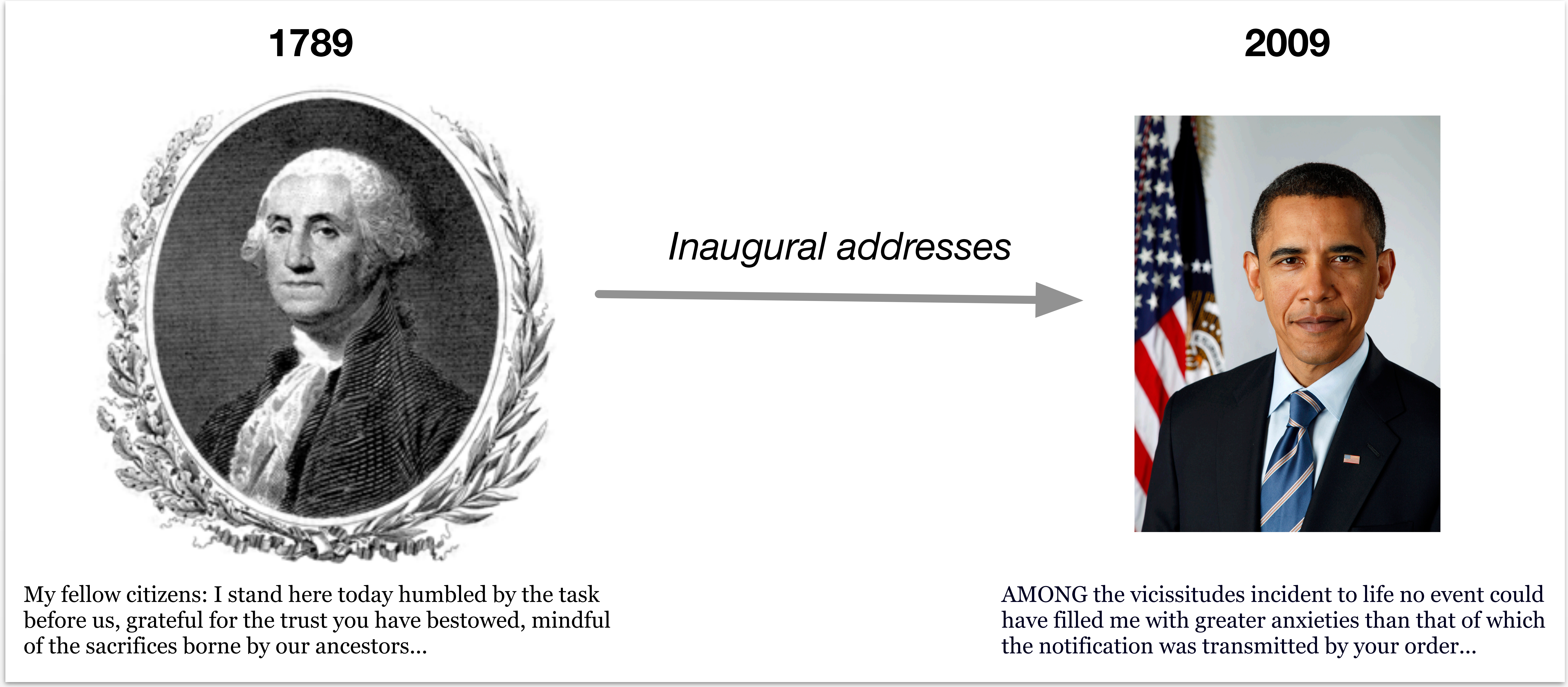


Extensions: Correlated Topic Model



Estimate a covariance matrix *Σ* that parameterizes correlations between topics in a document

Extensions: Dynamic Topic Models

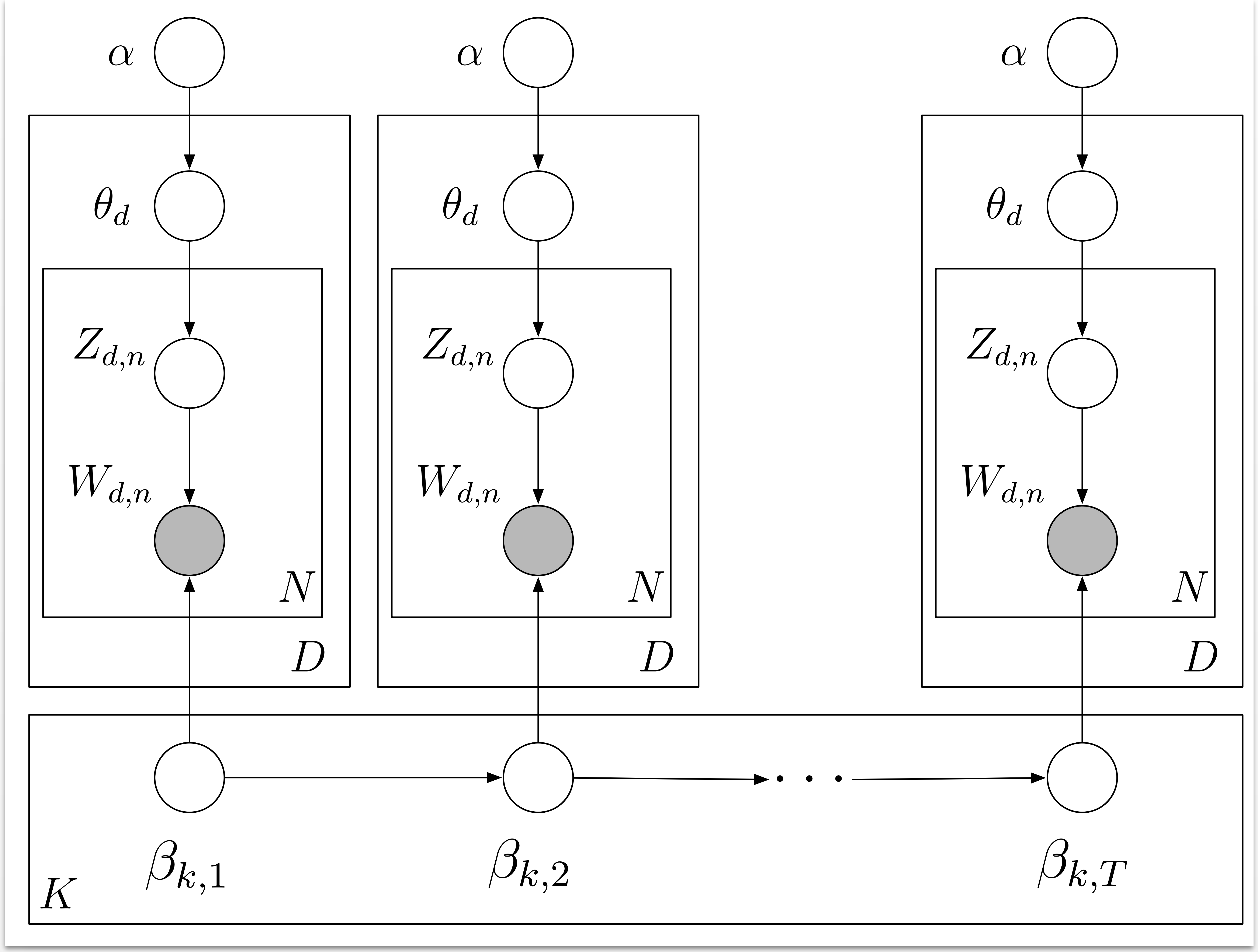


My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

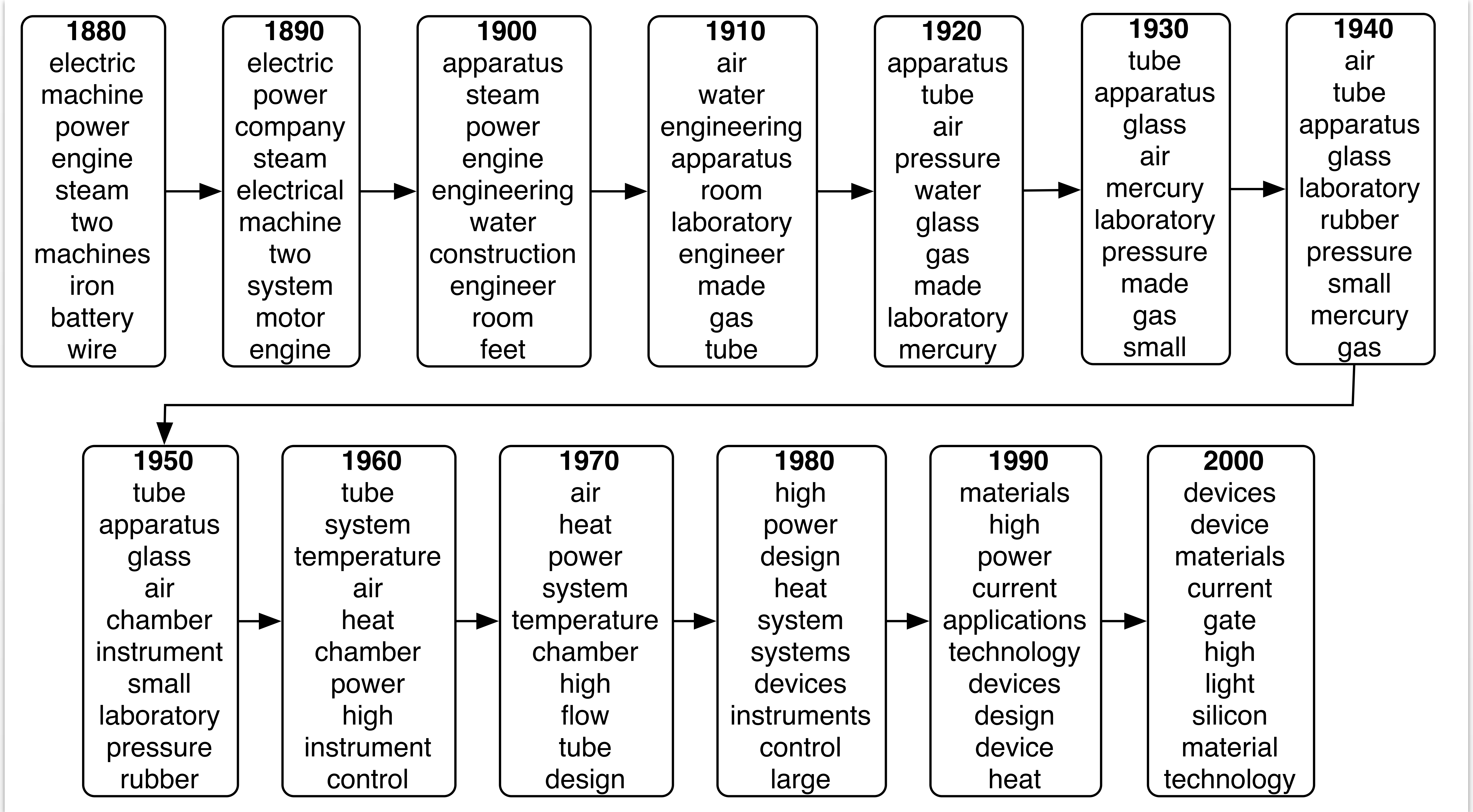
AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

Track changes in word distributions associated with a topic over time.

Extensions: Dynamic Topic Models



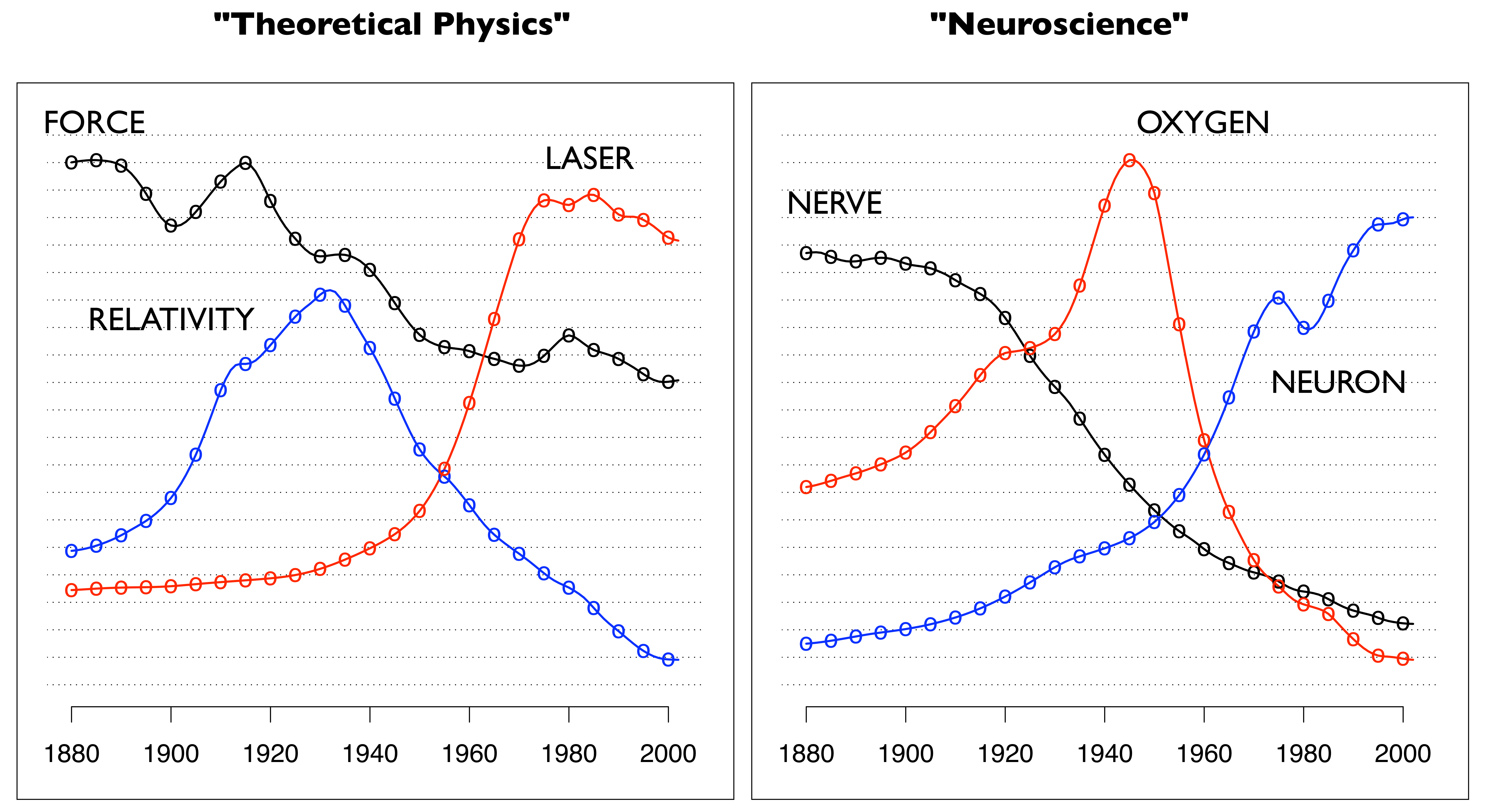
Extensions: Dynamic Topic Models



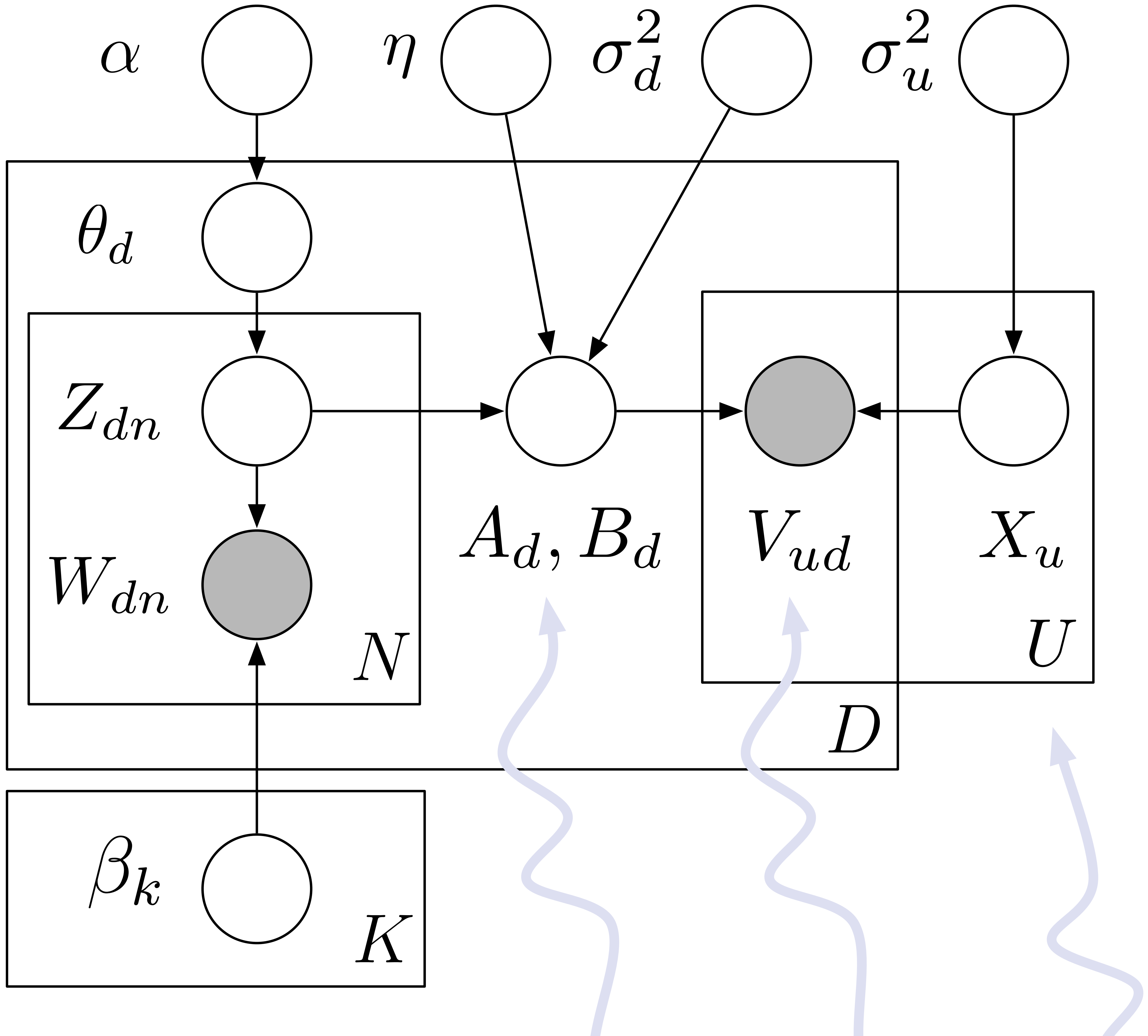
1880	1890	1900	1910	1920	1930	1940
electric	electric	apparatus	air	apparatus	tube	air
machine	power	steam	water	tube	apparatus	tube
power	company	power	engineering	air	glass	apparatus
engine	steam	engine	apparatus	pressure	air	glass
steam	electrical	engineering	room	water	mercury	laboratory
two	machine	water	laboratory	glass	laboratory	rubber
machines	two	construction	engineer	gas	pressure	pressure
iron	system	engineer	made	made	made	small
battery	motor	room	gas	laboratory	gas	mercury
wire	engine	feet	tube	mercury	small	gas

1950	1960	1970	1980	1990	2000
tube	tube	air	high	materials	devices
apparatus	system	heat	power	high	device
glass	temperature	power	design	power	materials
air	air	system	heat	current	current
chamber	heat	temperature	system	applications	gate
instrument	chamber	chamber	systems	technology	high
small	power	high	devices	devices	light
laboratory	high	flow	instruments	design	silicon
pressure	instrument	tube	control	device	material
rubber	control	design	large	heat	technology

Extensions: Dynamic Topic Models



Extensions: Ideal Point Topic Models



Bill content (topic model)

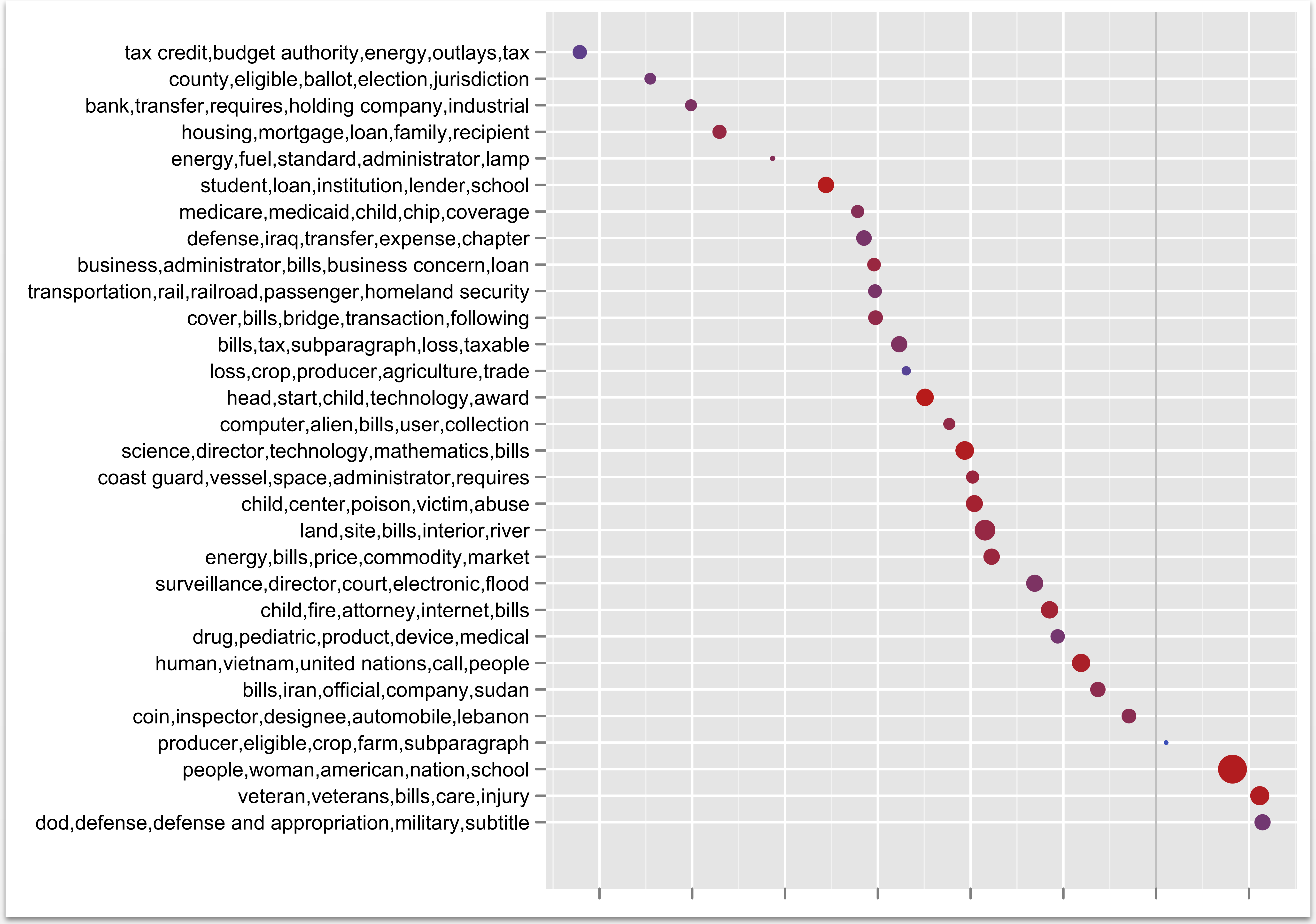
Bill sentiment variables

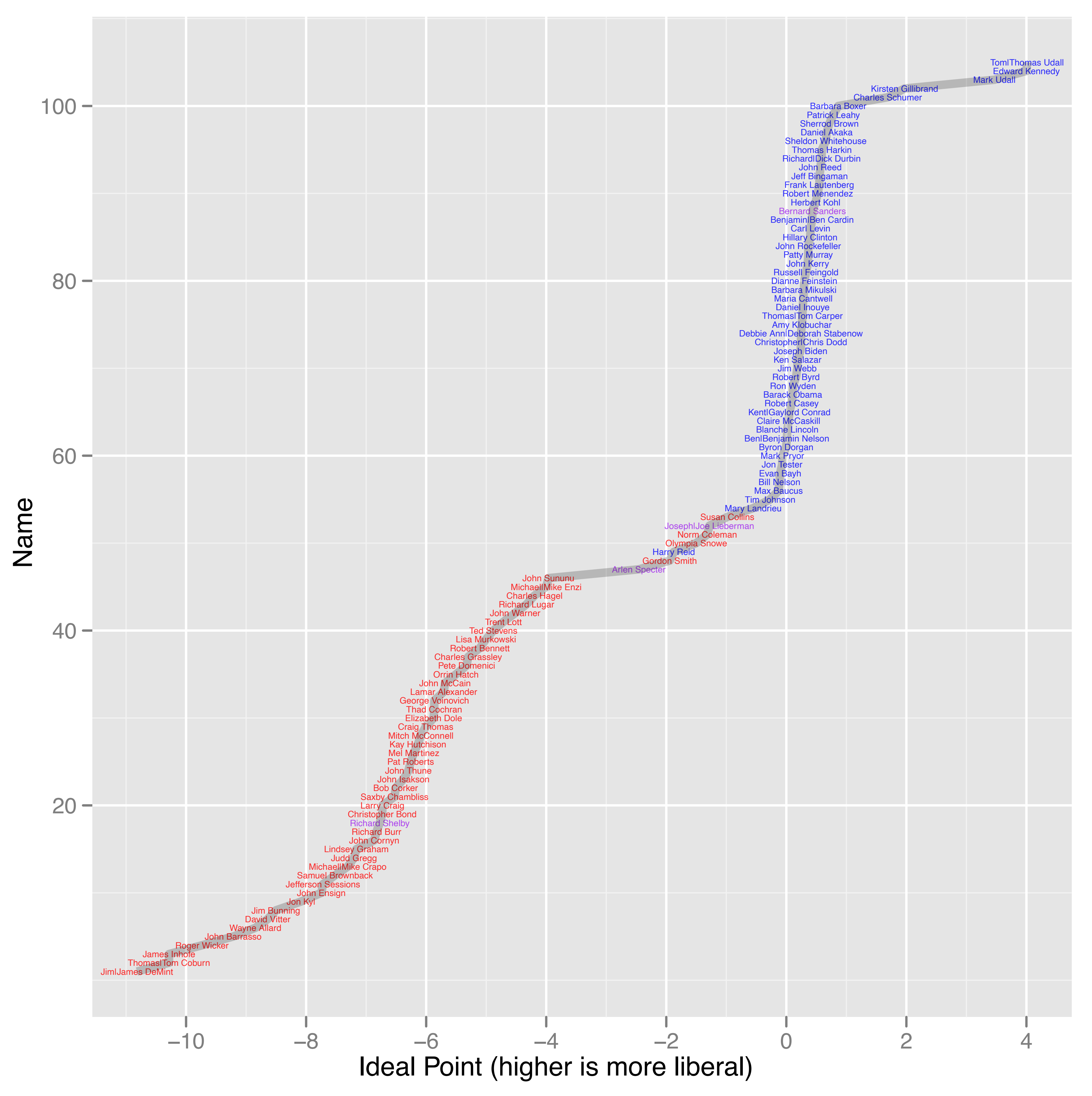
Observed votes

Legislator ideal points

$$p(V_{ud} = \text{"yea"} \mid X_u, A_d, B_d) = \frac{\exp(X_u A_d + B_d)}{\exp(X_u A_d + B_d) + 1}$$

Extensions: Ideal Point Topic Models





$$p(V_{ud} = \text{"yea"} \mid X_u, A_d, B_d) = \frac{\exp(X_u A_d + B_d)}{\exp(X_u A_d + B_d) + 1}$$

LDA: *Summary*

- **Idea:** Model documents as ***mixtures*** over topics
- **Model parameters (estimate with VBEM)**
  θ_d Topic probabilities for each document
  (K-dimensional vector for each document)
  β_k Word probabilities for each topic
  (V-dimensional vector for each topic)
- **Dirichlet Priors:** Enforce sparsity, associate a small number of topics which each document
- **Extensions:** Can design graphical models that build on LDA for a variety of modeling tasks