Topic Models

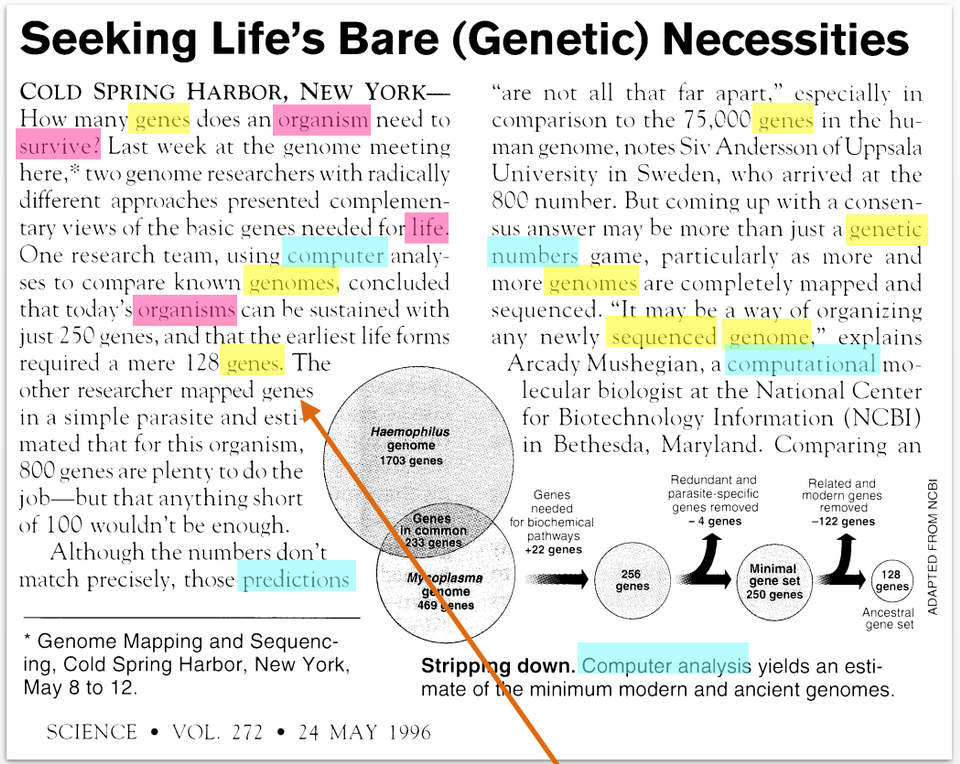
Shantanu Jain

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

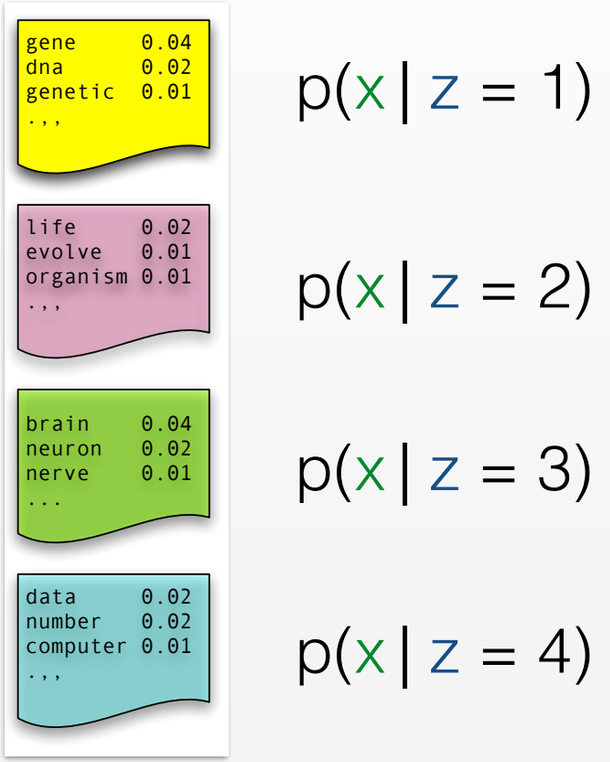
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

* 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



$p(x \mid z = 1)$

$p(x \mid z = 2)$

$p(x \mid z = 3)$

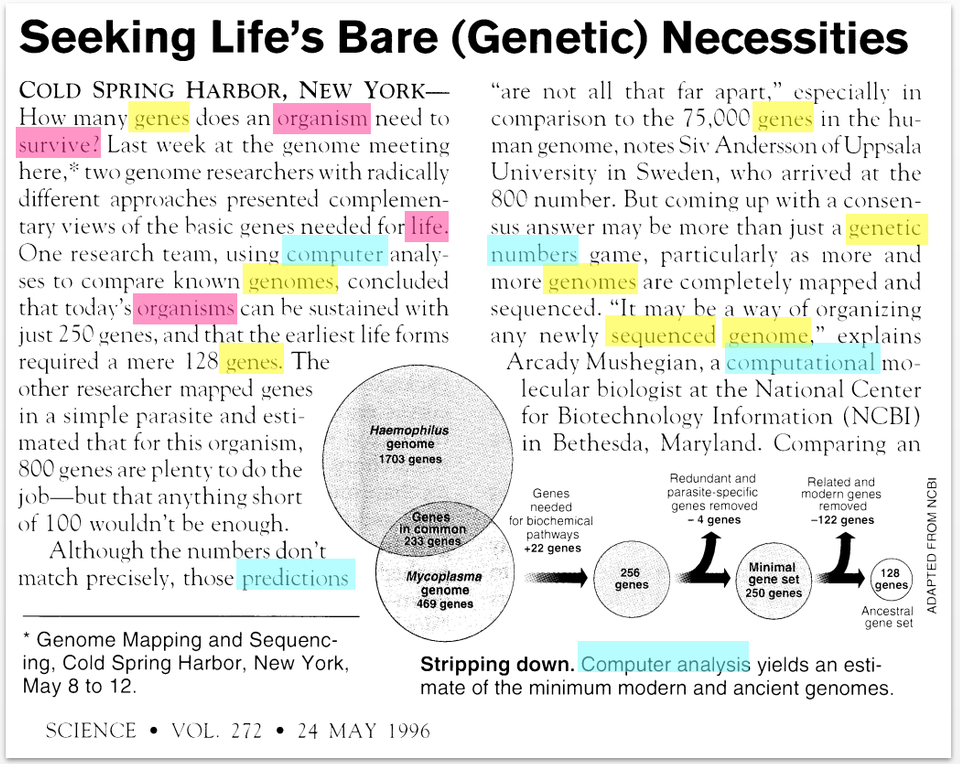
$p(x \mid z = 4)$

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 a position 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

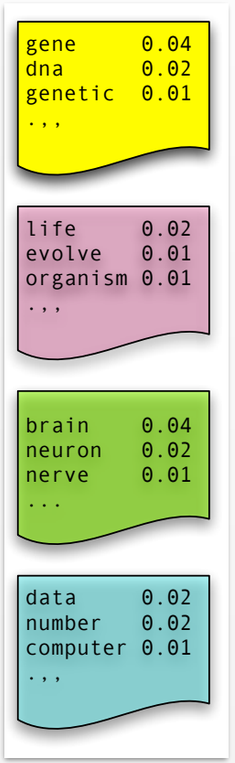
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

* 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



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

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

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

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

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

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

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

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

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

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

$$z_n \sim \text{Discrete}(\theta)$$

$$x_n \mid z_n = k \sim \text{Discrete}(\beta_k)$$

Pick a topic

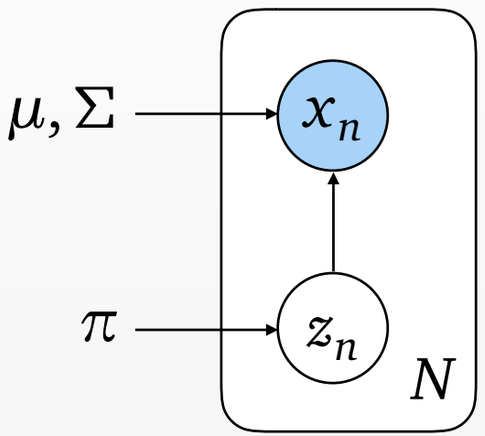
Pick a word given topic

By Discrete distribution, we mean Categorical distribution

Gaussian Mixtures vs Word Mixtures

Gaussian Mixture Model

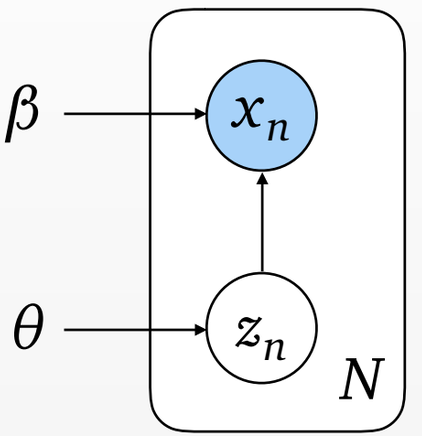
$\mu = \{\mu_k\}$
$\Sigma = \{\Sigma_k\}$



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

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

Discrete Mixture Model



$\beta = \{\beta_k\}$

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

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

Difference: Replace Gaussian with Discrete

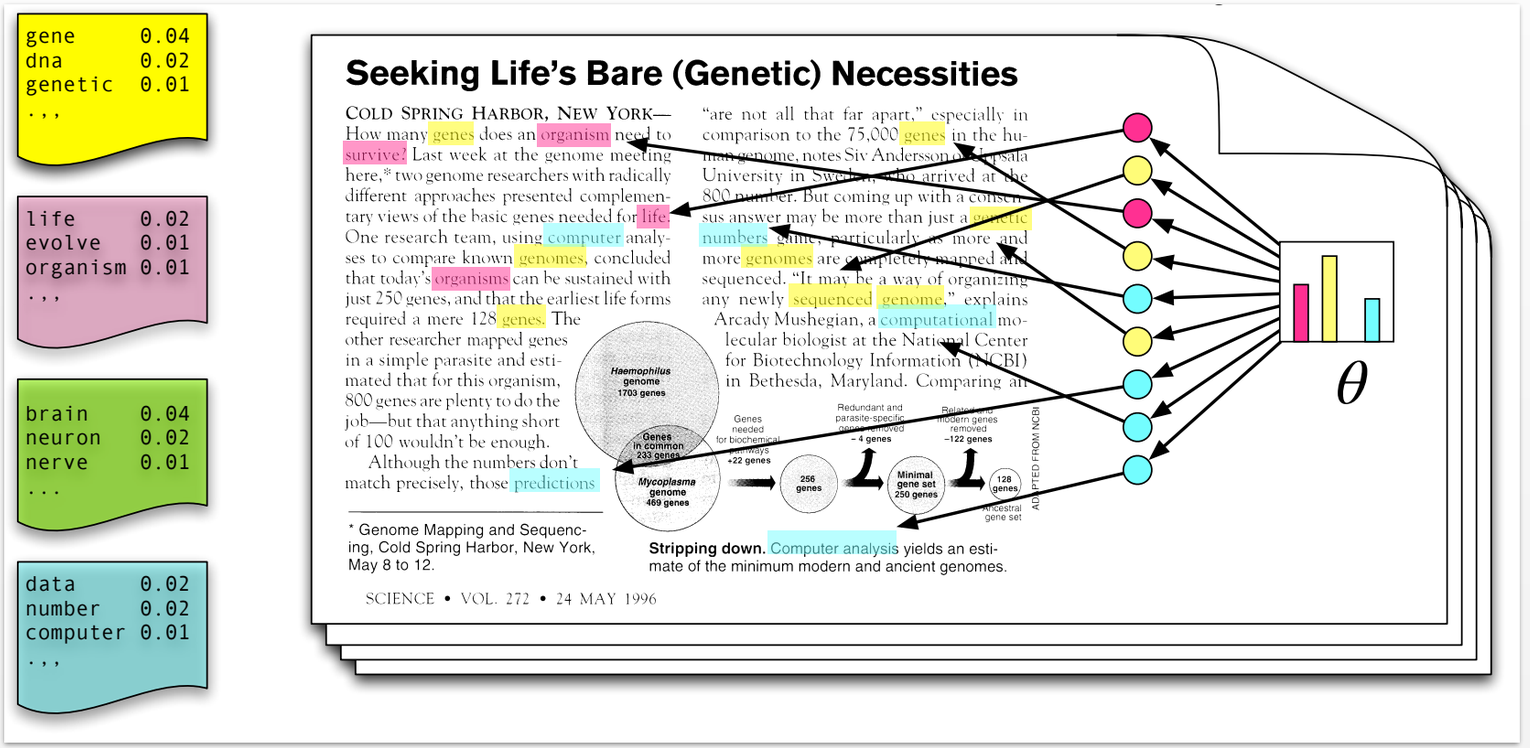
Using the term Discrete distribution to mean Categorical distribution

Topic Modeling

Topics (shared)

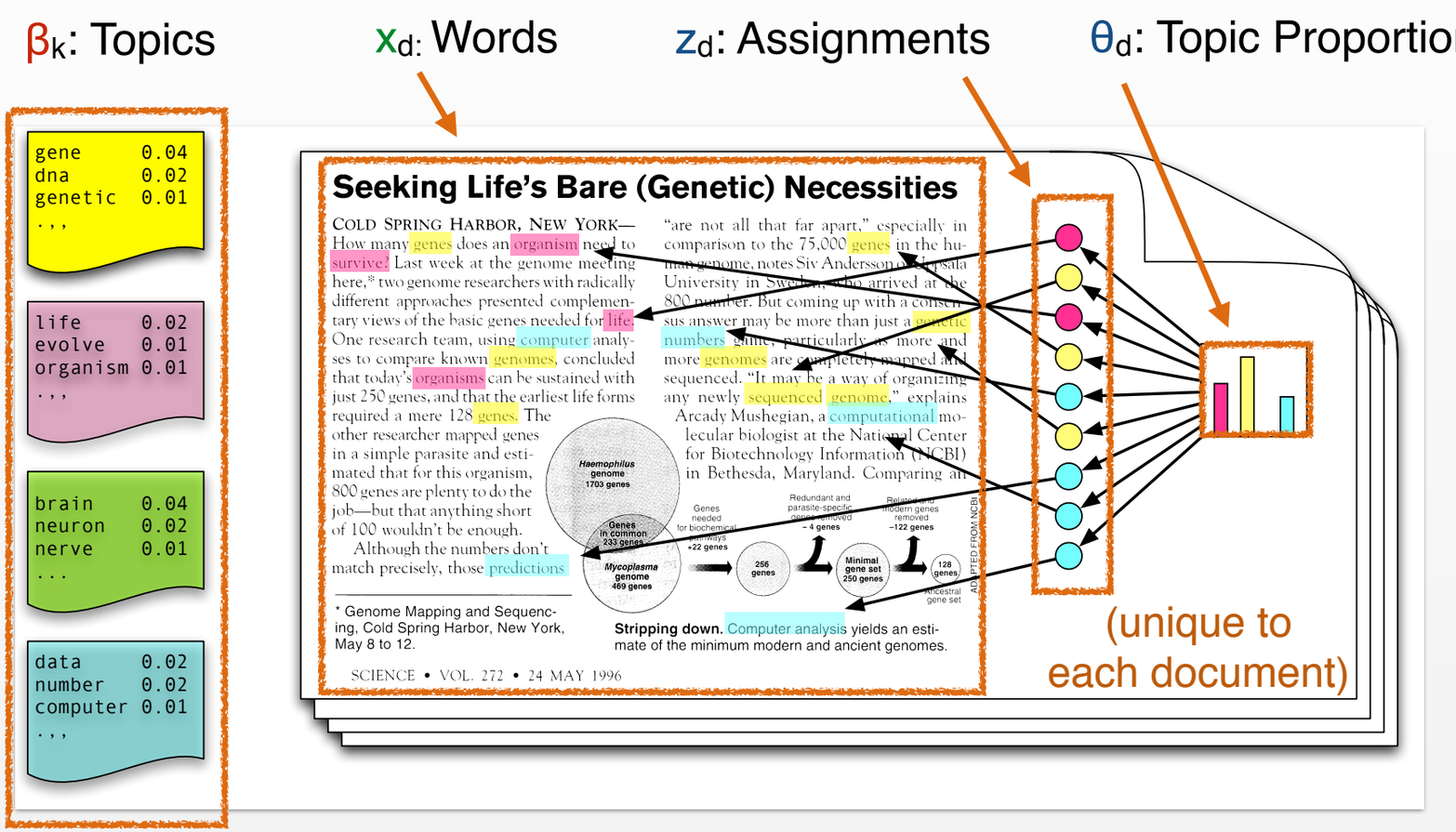
Words in Document (mixture over topics)

Topic Proportions (document-specific)



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

Topic Modeling



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

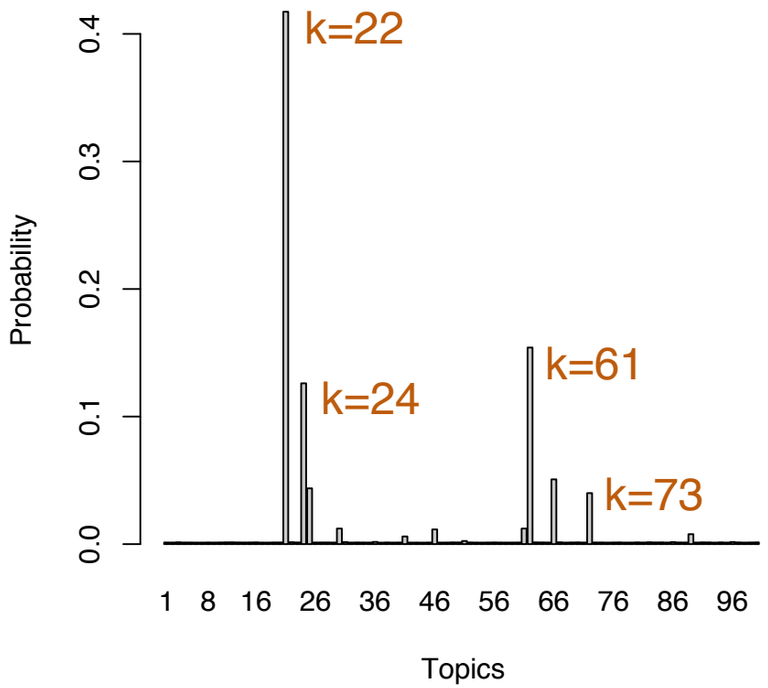
$$x_{d,n} \mid z_{d,n} = k \sim \text{Discrete}(\beta_k)$$

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

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



Next Slide: Frequent words in these topics

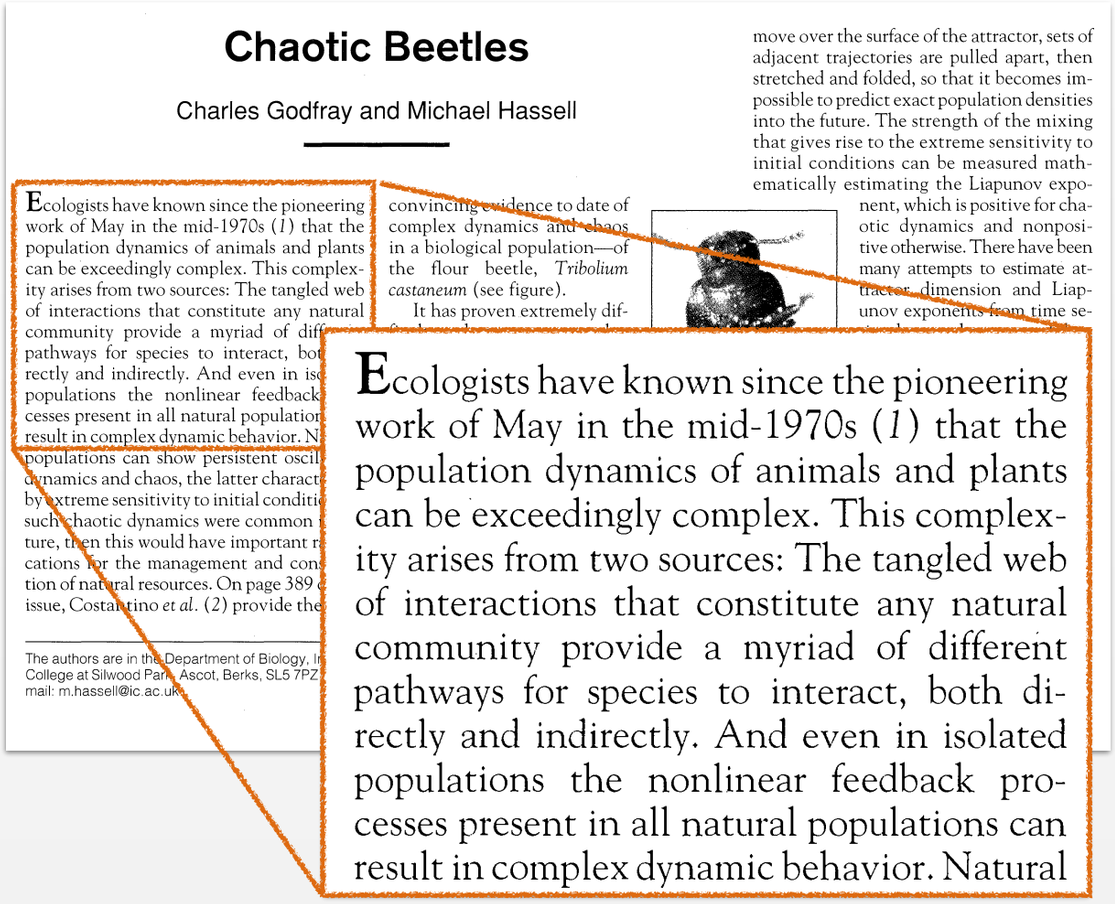
Most Probable Words in Topics

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

Most frequent (*within topic*)

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

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 isc populations the nonlinear feedback cesses present in all natural populatio 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

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

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

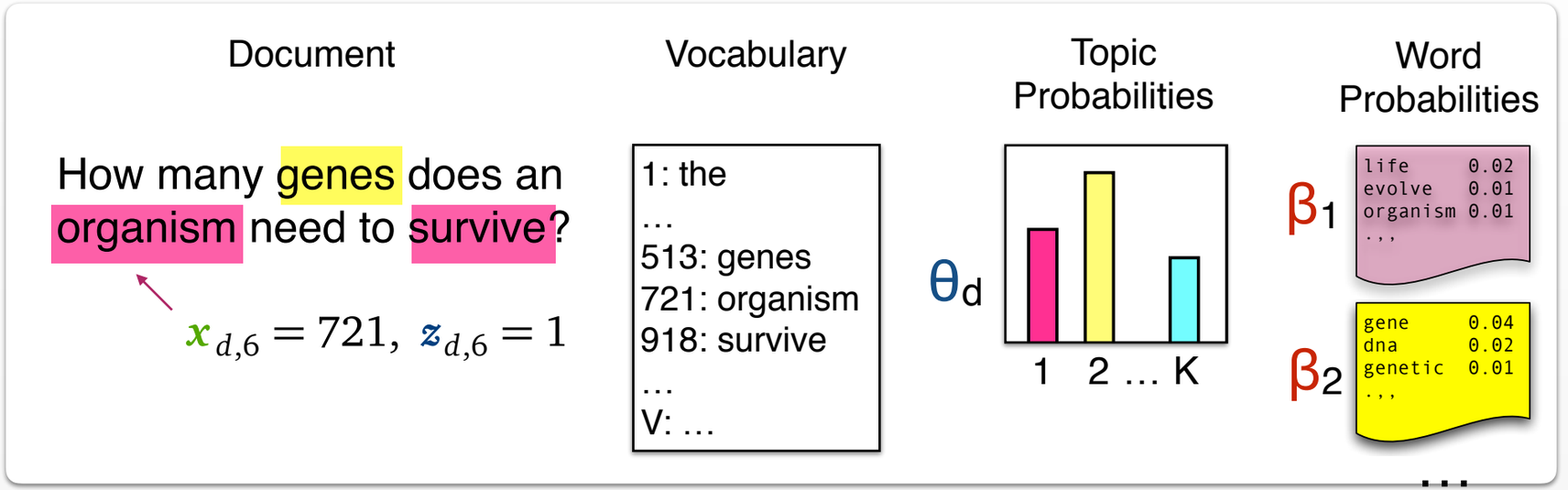
It has proven extremely dif-

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-

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	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

Estimating the Parameters



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

Document Collection

documents

d=1 $x_1 = [x_{11}, x_{12}, \ldots\ x_{1N_1}]$

d=2 $x_2 = [x_{21}, x_{22}, \ldots\ x_{2N_2}]$

d=3 $x_3 = [x_{D1}, x_{D2}, \ldots\ x_{DN_D}]$

positions

(not a matrix)

Topic proportions

documents

$$\begin{bmatrix} \theta_{11}, \theta_{12}, \ldots\ \theta_{1K} \\ \theta_{21}, \theta_{22}, \ldots\ \theta_{2K} \\ \vdots \\ \theta_{D1}, \theta_{D2}, \ldots\ \theta_{DK} \end{bmatrix}$$

topics

D x K

Topic word probabilities

topics

$$\begin{bmatrix} \beta_{11}, \beta_{12}, \ldots\ \beta_{1V} \\ \beta_{21}, \beta_{22}, \ldots\ \beta_{2V} \\ \vdots \\ \beta_{K1}, \beta_{K2}, \ldots\ \beta_{KV} \end{bmatrix}$$

words

K x V

Interpretation as Matrix Factorization

Bag of word Vector

$X_d = [X_{d1}, X_{d2}, \ \dots \ X_{dV}]$,

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

$$\mathbb{E}[X_{dv}] = N_d p(x_{dn} = v \mid \theta, \beta) = N_d \sum_{k=1}^{K} \theta_{dk}\beta_{kv}$$

$$\mathbb{E}[X] = N \times \theta \times \beta$$

Document word counts **Topic counts** **Topic word probabilities**

documents

$$\mathbb{E}\begin{bmatrix} X_{11}, X_{12}, \dots X_{1V} \\ X_{21}, X_{22}, \dots X_{2V} \\ \vdots \\ X_{D1}, X_{D2}, \dots X_{DV} \end{bmatrix} = \begin{bmatrix} N_1, 0, \dots 0 \\ 0, N_2, \dots 0 \\ \vdots \\ 0, 0, \dots N_D \end{bmatrix} \times \begin{bmatrix} \theta_{11}, \theta_{12}, \dots \theta_{1K} \\ \theta_{21}, \theta_{22}, \dots \theta_{2K} \\ \vdots \\ \theta_{D1}, \theta_{D2}, \dots \theta_{DK} \end{bmatrix} \times \begin{bmatrix} \beta_{11}, \beta_{12}, \dots \beta_{1V} \\ \beta_{21}, \beta_{22}, \dots \beta_{2V} \\ \vdots \\ \beta_{K1}, \beta_{K2}, \dots \beta_{KV} \end{bmatrix}$$

$$X \approx N \times \theta \times \beta$$

stocks chairman the wins game finance sports stocks game

$$\begin{bmatrix} 2 & 4 & 8 & \dots & 0 & 1 \\ \dots & & & & & \\ 0 & 1 & 7 & \dots & 2 & 3 \end{bmatrix} \simeq \begin{bmatrix} 112 \cdot 0.91 & \dots & 112 \cdot 0.01 \\ \dots & & \\ 234 \cdot 0.02 & \dots & 234 \cdot 0.86 \end{bmatrix} \begin{bmatrix} 0.0081 & \dots & 0.0002 \\ \dots & & \\ 0.0001 & \dots & 0.0072 \end{bmatrix}$$

Relationship to Latent Semantic Analysis

LSA: Factorize word counts (using PCA)

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

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

PCA as matrix factorization:

- $Z = U^T X$
- X is approximated as $X \approx UZ$

Minimize Reconstruction error

$$\min_{U,Z} \| X - UZ \|^2$$

Topic Models: Factorize word counts (using mixture model)

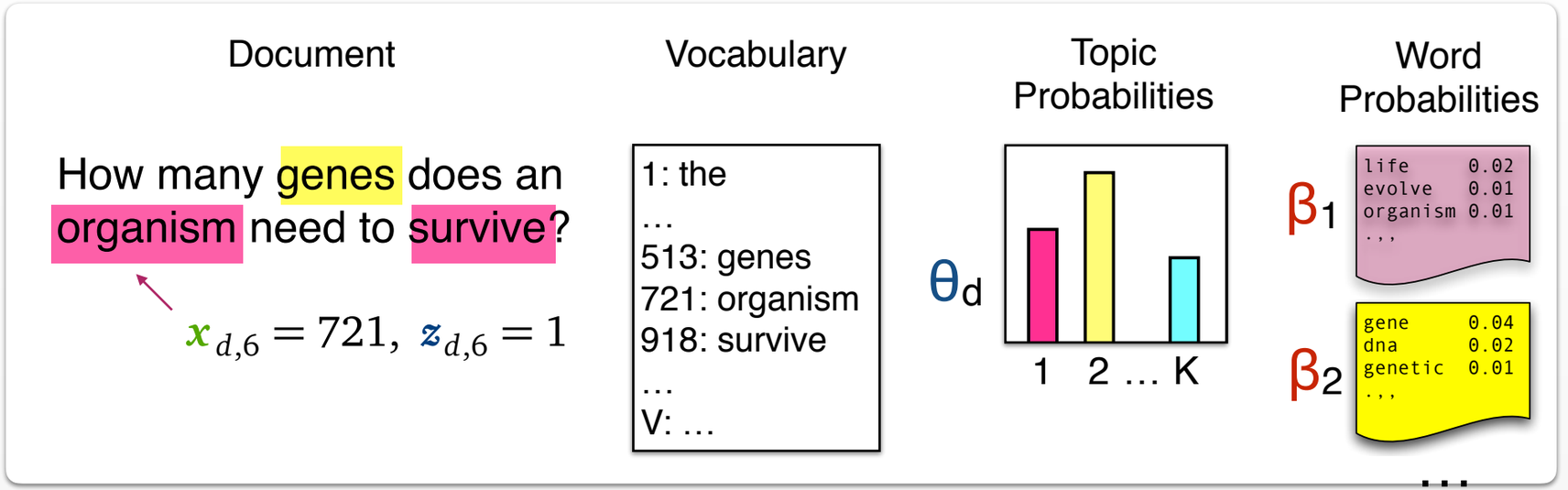
$X \approx N \times \theta \times \beta$

$N^{-1}X$(D x V) $\approx$ θ(D x K) β(K x V)

$X^T N^{-1}$(V x D) $\approx$ β^T(V x K) θ^T(K x D)

Topic models can also be tackled as matrix factorization problem, where rows of β, θ are constrained to be probability vectors

Estimating the Parameters



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

Document Collection

documents

d=1 $x_1 = [x_{11}, x_{12}, \ldots\ x_{1N_1}]$

d=2 $x_2 = [x_{21}, x_{22}, \ldots\ x_{2N_2}]$

d=3 $x_3 = [x_{D1}, x_{D2}, \ldots\ x_{DN_D}]$

positions

(not a matrix)

Topic proportions

documents

$$\begin{bmatrix} \theta_{11}, \theta_{12}, \ldots\ \theta_{1K} \\ \theta_{21}, \theta_{22}, \ldots\ \theta_{2K} \\ \vdots \\ \theta_{D1}, \theta_{D2}, \ldots\ \theta_{DK} \end{bmatrix}$$

topics

D x K

Topic word probabilities

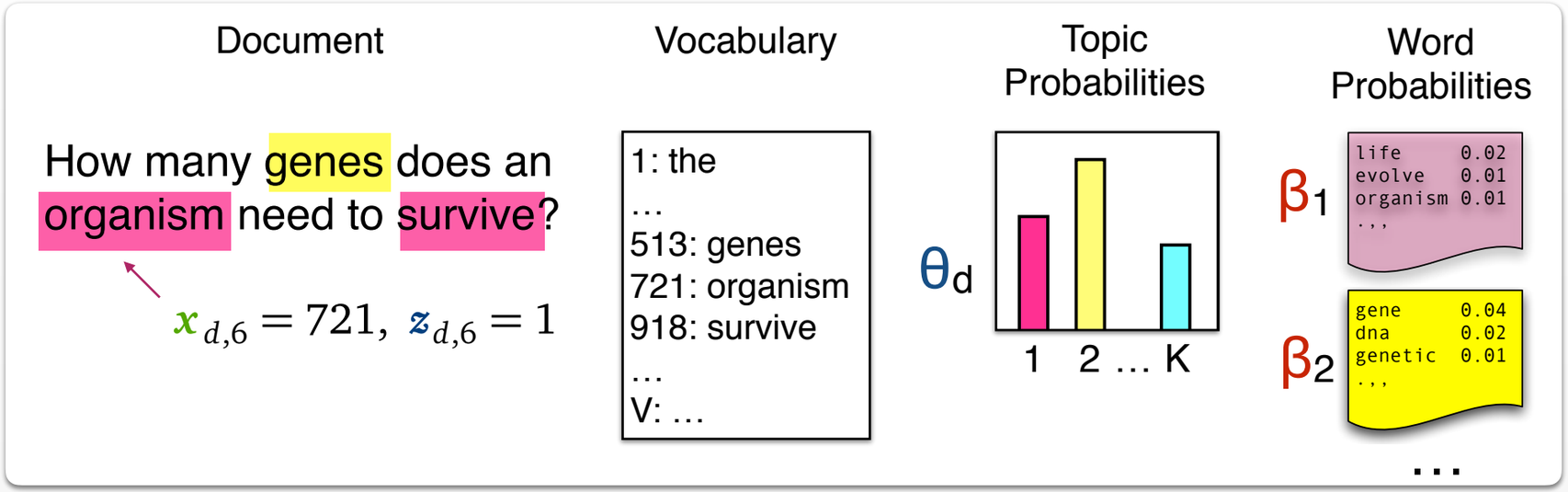
topics

$$\begin{bmatrix} \beta_{11}, \beta_{12}, \ldots\ \beta_{1V} \\ \beta_{21}, \beta_{22}, \ldots\ \beta_{2V} \\ \vdots \\ \beta_{K1}, \beta_{K2}, \ldots\ \beta_{KV} \end{bmatrix}$$

words

K x V

Calculating the Likelihood for each Word



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

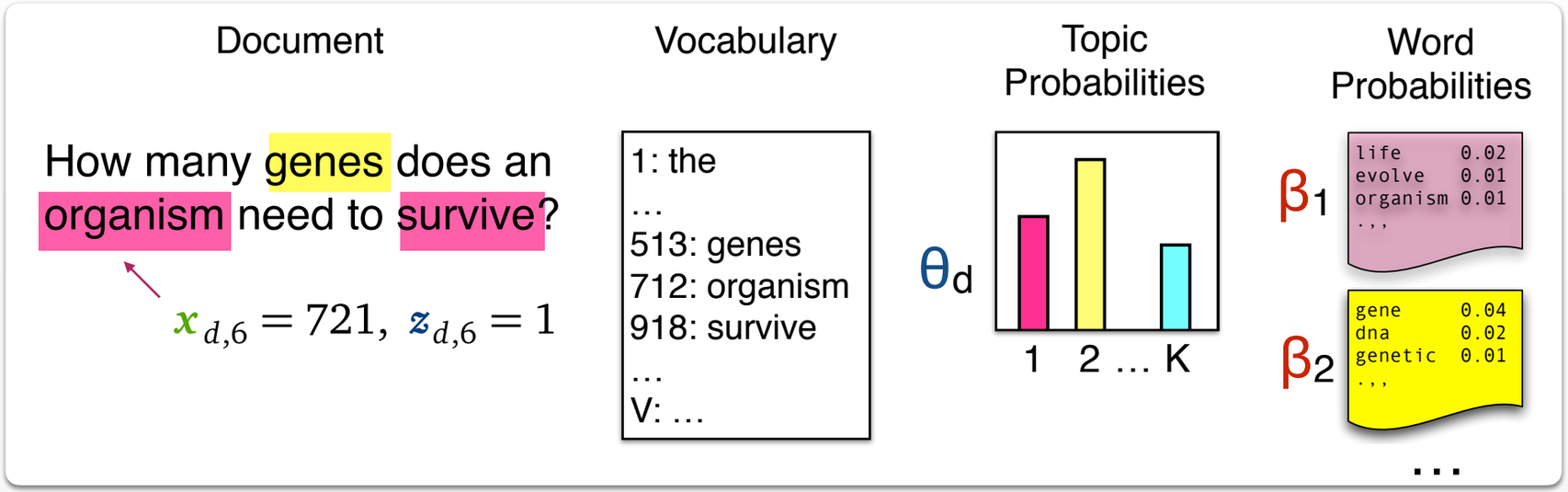
Probability of word *v* given topic *k*

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



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

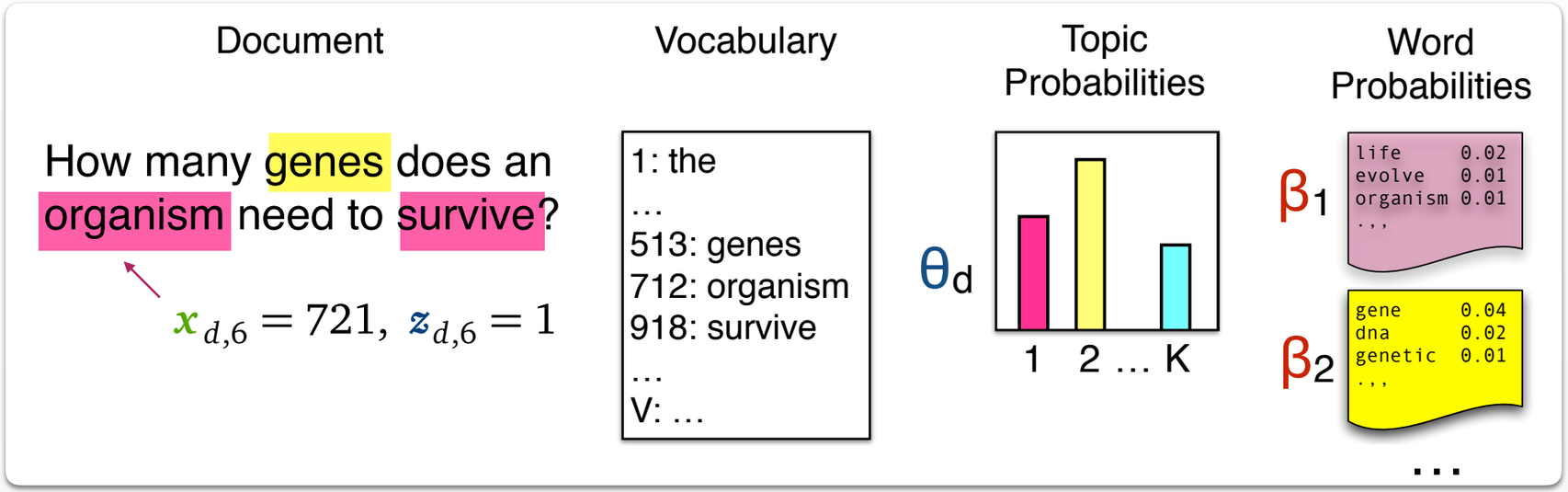
$$x_d = [x_{d1}, x_{d2}, \ \ldots \ x_{dN_d}]$$

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

take log probability, substitute result from previous slide

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

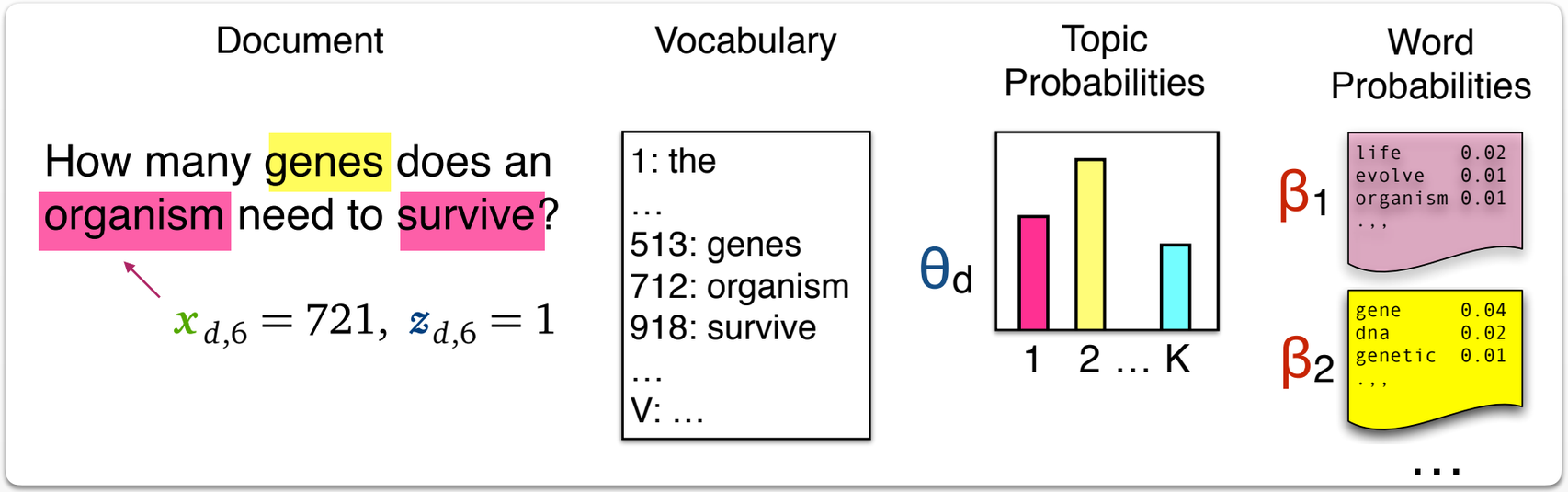
Calculating the Likelihood for all Words



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



log probability of all words in document *d*

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

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

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

bag-of-word vector

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

Interpretation as Matrix Factorization

Log likelihood

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

$$\log p(x \mid \beta, \theta) = \sum_{d=1}^{D} X_d \log(\theta_d \beta)^T$$

Bag of word Vector

$X_d = [X_{d1}, X_{d2}, \ \ldots \ X_{dV}]$,

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

Difficult to optimize with gradient ascent.

Need to use EM!

Sketch of EM derivation

Topic models

Observed variable

- $x = \{x_{nd}\}$: all words in the corpus.

Unobserved variable

- $z = \{z_{nd}\}$: the topics of all words in the corpus

Model Parameters

- $\beta = \{\beta_k\}$: word probabilities, one per topic
- $\theta = \{\theta_d\}$: topic probabilities, one per document

$\eta = \{\theta, \beta\}$

$$l(\eta) = \log p(x \mid \eta)$$

1) Initialize η as η_0

2) For $t = 0,1\ldots$, repeat until convergence

2a) E-Step:

$$\mathcal{Q}^t(\eta) = \mathbf{E}_{q^t(z)}[\log p(x, z \mid \eta)],$$

where $q^t(z) = p(z \mid x, \eta^t)$

2b) M-Step:

$$\eta^{t+1} \leftarrow \text{argmax}_\eta \mathcal{Q}^t(\eta)$$

Clustering: GMM

Observed variable

- $x = \{x_n\}$: all datapoint

Unobserved variable

- $z = \{z_n\}$: the cluster index of all datapoint

Model Parameters

- $\mu = \{\mu_k\}$: the means of all gaussian clusters.
- $\Sigma = \{\Sigma_k\}$: the covariances of all gaussian clusters
- $\alpha = \{\alpha_k\}$: the cluster proportions

$\theta = \{\alpha, \mu, \Sigma\}$

$$l(\theta) = \log p(x \mid \theta)$$

1) Initialize θ as θ_0

2) For $t = 0,1\ldots$, repeat until convergence

2a) E-Step:

$$\mathcal{Q}^t(\theta) = \mathbf{E}_{q^t(z)}[\log p(x, z \mid \theta)],$$

where $q^t(z) = p(z \mid x, \theta^t)$

2b) M-Step:

$$\theta^{t+1} \leftarrow \text{argmax}_\theta \mathcal{Q}^t(\theta)$$

PLSI/PLSA*: EM for Topic Models

Generative Model

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

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

E-step: Update assignments

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

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

Need to compute $q^t(z) = p(z \mid x, \theta^t, \beta^t)$ for the E-Step. This is a joint distribution over all the topic assignment variables, one per word in the corpus. It can be decomposed as a product of $\phi_{d,n,k}$.

M-step: Update parameters

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

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

PLSI/PLSA: E-step

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

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

(Apply Bayes' Rule)

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

(Substitute results from previous slides)

General Form, with One-hot Indexing Trick

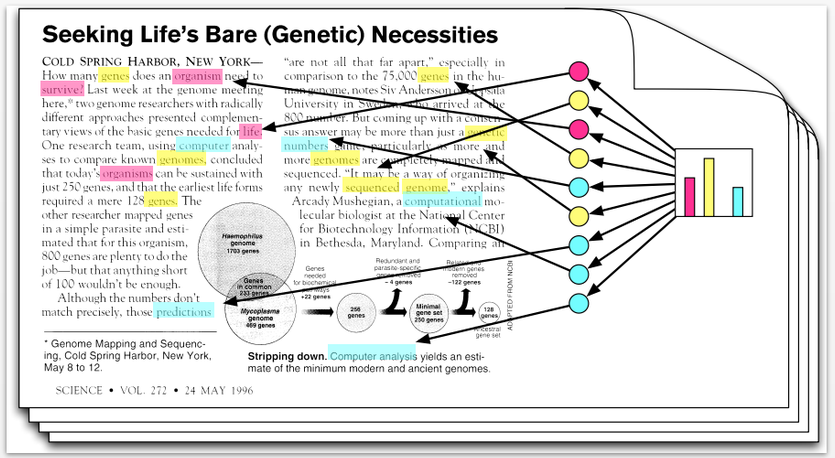
$$\phi_{d,n,k} = \frac{\theta_{d,k}\left(\sum_{v=1}^{V}\beta_{k,v}I[x_{d,n}=v]\right)}{\sum_{l=1}^{K}\theta_{d,l}\left(\sum_{v=1}^{V}\beta_{l,v}I[x_{d,n}=v]\right)}$$

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 completed, 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

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 as

$\theta_{d,k} = \frac{N^{\theta}_{d,k}}{N_d}$ Fraction of topic *k* in document *d*

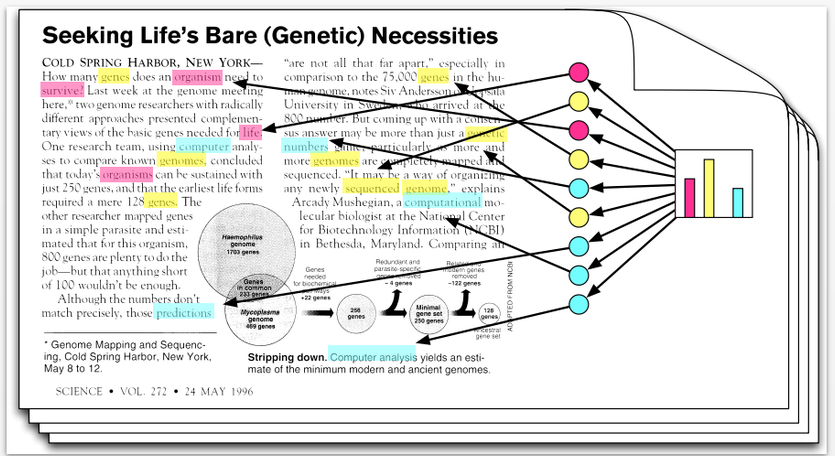
$\beta_{k,v} = \frac{N^{\beta}_{k,v}}{\sum_{d=1}^{D} N^{\theta}_{d,k}}$ Fraction of word *v* in topic *k*

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 matter of numbers, since, particularly as more and more genomes are completed, 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

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)

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

Latent Dirichlet Allocation

Topic Models with Dirichlet Priors

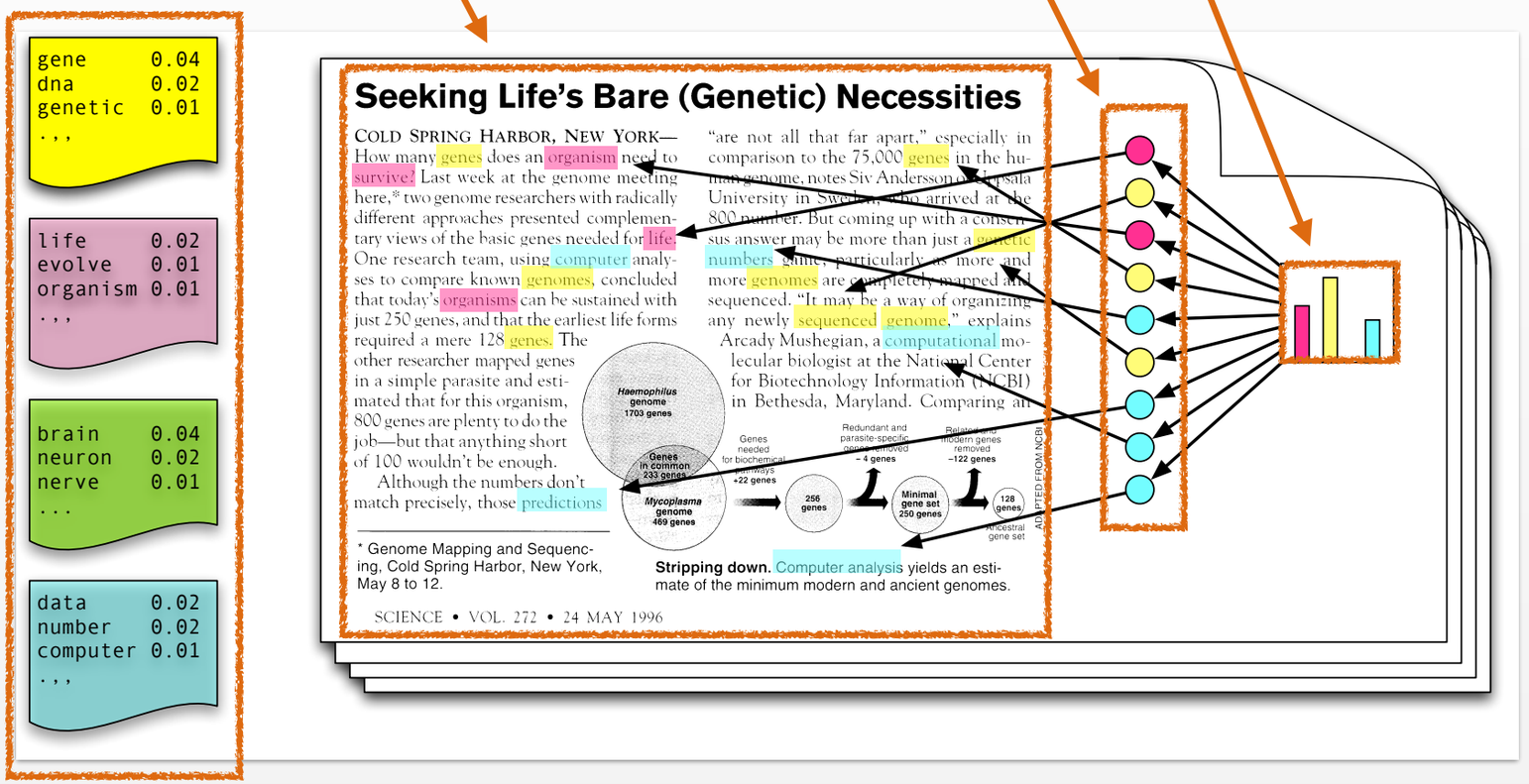
Review: Topic Modeling with PLSA/PLSI

β_k: Topics (shared)

x_d: Words

z_d: Assignments (document-specific)

θ_d: Topic Proportions

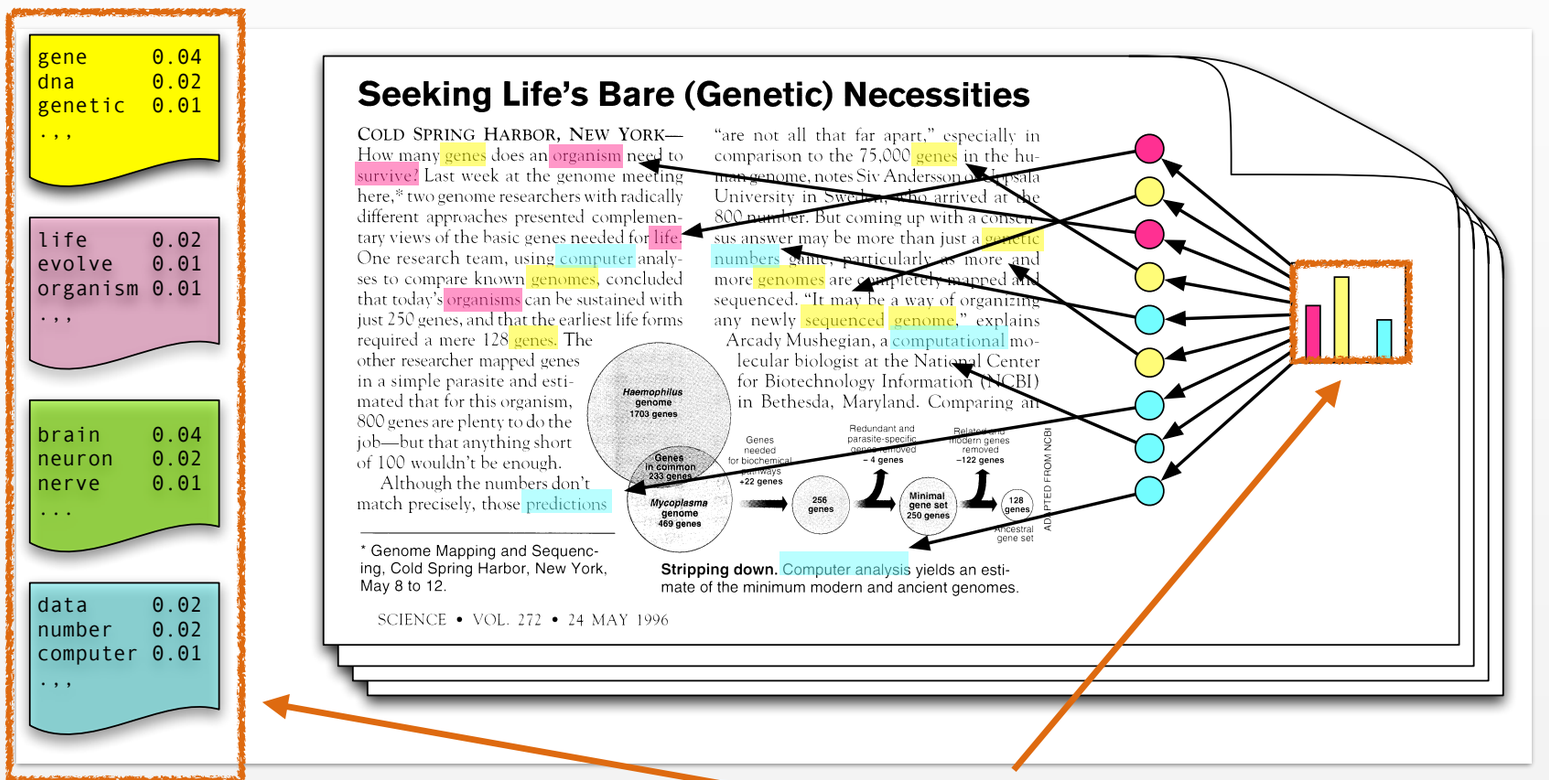


$$z_{d,n} \sim \text{Discrete}(\theta_d)$$

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

LDA: Add *Dirichlet* Priors

β_k: Topics (shared) | x_d: Words | z_d: Assignments (document-specific) | θ_d: Topic Proportions



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

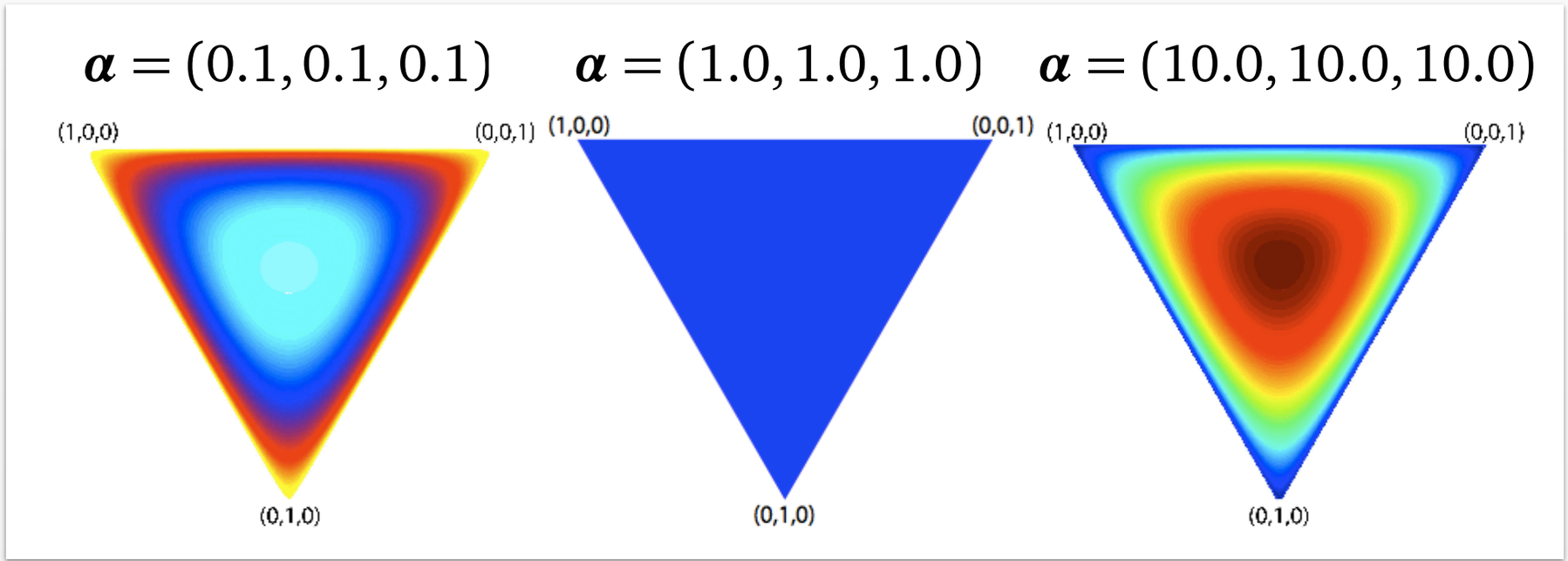
$$x_{d,n} \mid z_{d,n} = k \sim \text{Discrete}(\beta_k)$$

$$\theta_d \sim \text{Dirichlet}(\alpha)$$

$$\beta_k \sim \text{Dirichlet}(\eta_k)$$

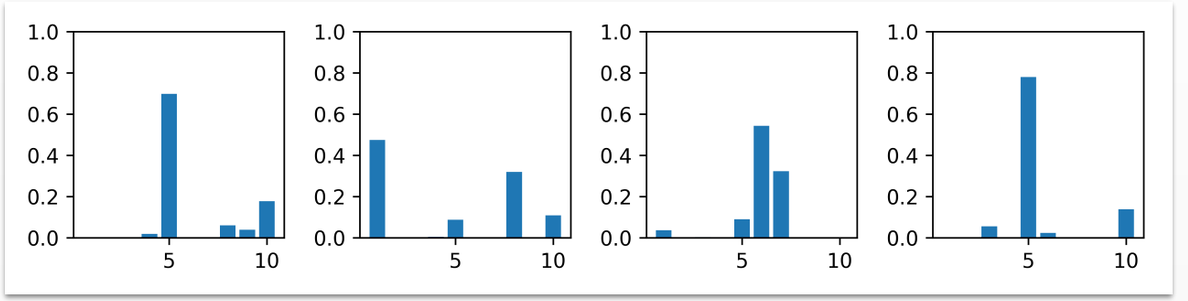
Review: Dirichlet Distribution

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

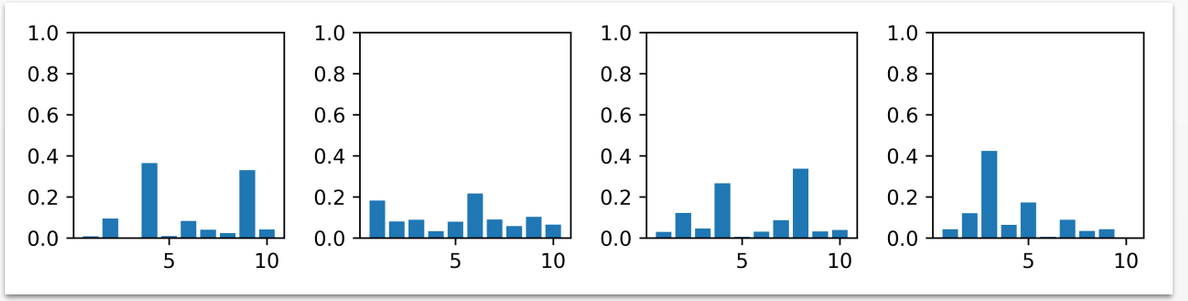


Review: Dirichlet Distribution

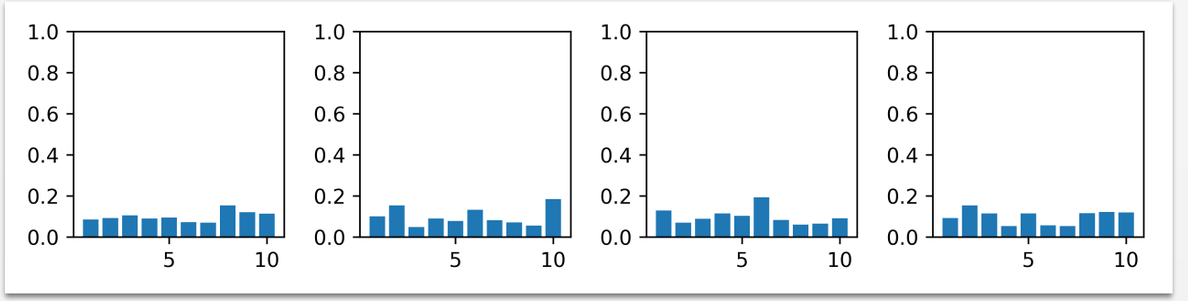
$\alpha_k = 0.1$



$\alpha_k = 1.0$



$\alpha_k = 10.0$



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

Maximum likelihood estimation

X: a dataset with N points:

$$X = \{x_n\}_{n=1}^{N}$$

$x_n \in \mathcal{X}$: the space where x_n takes values. Typically $\mathcal{X} = \mathbb{R}^D$

Model Assumption:

$$x_n \sim P_\theta$$

A distribution parametrized by θ (unknown) defined on the sample space the space $\mathcal{X}$

Maximum likelihood estimate of θ.

$$\theta^{ML} = \text{argmax}_\theta \; p(X|\theta)$$
$$= \text{argmax}_\theta \; \log p(X|\theta)$$

Does not depend on the prior distribution, $p(\theta)$.

Maximum likelihood is a frequents approach. It assumes that the parameters are unknown, but fixed, in the sense that they do not come from a distribution. In the frequentist world the prior distribution does not exist.

Maximum a posteriori (MAP) estimation

X: a dataset with N points:

$$X = \{x_n\}_{n=1}^{N}$$

MAP Estimate

Posterior

$$\theta^{MAP} = \text{argmax}_\theta \; p(\theta \mid X, \alpha)$$

$$= \text{argmax}_\theta \; \frac{p(X \mid \theta, \alpha) p(\theta \mid \alpha)}{P(X \mid \alpha)}$$

$$= \text{argmax}_\theta \; \frac{p(X \mid \theta) p(\theta \mid \alpha)}{P(X \mid \alpha)}$$

Not a function of θ

$$= \text{argmax}_\theta \; p(X \mid \theta) \times p(\theta \mid \alpha)$$

Likelihood — Prior

As the number of data points increase the contribution of the likelihood term increases and the prior term reduces

Model Assumption:

$x_n \sim P_\theta$ The data distribution, parametrized by θ

$\theta \sim P_\alpha$ The prior distribution on the parameter θ. In the Bayesian setting, the parameter itself is a random variable having a “prior” distribution, P_α. The prior distribution parameter, α, is called the hyper-parameter. Which is typically assumed to be know or searched via grid-search.

$$\theta^{ML} = \text{argmax}_\theta \; p(X \mid \theta)$$

$$= \text{argmax}_\theta \; \log p(X \mid \theta)$$

Maximum a posteriori (MAP) estimation

Likelihood

$$p(X \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

Prior

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

Posterior

$$p(\mu \mid X, a, b) \propto p(X \mid \mu) p(\mu \mid a, b)$$

$$\propto \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right) \frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)}$$

Not a function of μ

$$\propto \mu^{\left[\sum_{n=1}^{N} x_n\right]} (1-\mu)^{\left[N - \sum_{n=1}^{N} x_n\right]} \mu^{a-1}(1-\mu)^{b-1}$$

$$\propto \mu^{N_1+a-1}(1-\mu)^{N_0+b-1}$$

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

$$N_1 = \sum_{n=1}^{N} x_n$$

$$N_0 = N - N_1$$

When the prior and posterior distribution belong to the same family, the choice of the prior is called conjugate prior

MAP Estimate

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

$$X = \{x_n\}_{n=1}^{N} \qquad x_n \in \{0,1\}$$

$$x_n \sim \mathrm{Bernoulli}(\mu)$$

$$p(x_n \mid \mu) = \begin{cases} \mu & \text{if } x_n = 1 \\ 1-\mu & \text{if } x_n = 0 \end{cases}$$

$$p(x_n \mid \mu) = \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\mu \sim \mathrm{Beta}(a, b)$$

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

MAP generalization to Discrete distribution

Likelihood

$$p(X|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{I[x_n=k]}$$

Prior

$$p(\mu|\alpha) = \frac{\prod_{k=1}^{K}\mu_k^{\alpha_k}}{\mathrm{B}(\alpha)}$$

Posterior

$$p(\mu|X,a,b) \propto p(X|\mu)p(\mu|a,b)$$

$$\propto \left(\prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{I[x_n=k]}\right)\frac{\prod_{k=1}^{K}\mu_k^{\alpha_k}}{\mathrm{B}(\alpha)}$$

Not a function of μ

$$\propto \prod_{k=1}^{K}\mu_k^{\left[\sum_{n=1}^{N} I[x_n=k]\right]}\prod_{k=1}^{K}\mu_k^{\alpha_k}$$

$$\propto \prod_{k=1}^{K}\mu_k^{N_k+\alpha_k} \qquad N_k = \sum_{n=1}^{N} I[x_n=k]$$

$$= \mathrm{Dirichlet}(\mu|N_1+\alpha_2, N_2+\alpha_1, \ldots, N_K+\alpha_K)$$

MAP Estimate $$\mu_k^{MAP} = \frac{N_k+\alpha_k-1}{N+\sum_{k=1}^{K}\alpha_k - K}$$

$$X = \{x_n\}_{n=1}^{N} \quad x_n \in \{1,2,\ldots,K\}$$

$$x_n \sim \mathrm{Discrete}(\mu) \qquad \sum_{k=1}^{K}\mu_k = 1$$

$$p(x_n|\mu) = \begin{cases} \mu_1 & \text{if } x_n = 1 \\ \mu_2 & \text{if } x_n = 2 \\ \vdots & \\ \mu_K & \text{if } x_n = K \end{cases}$$

$$p(x_n|\mu) = \prod_{k=1}^{K}\mu_k^{I[x_n=k]}$$

$$\mu \sim \mathrm{Dirichlet}(\alpha)$$

$$p(\mu|\alpha) = \frac{\prod_{k=1}^{K}\mu_k^{\alpha_k}}{\mathrm{B}(\alpha)}$$

$$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K] \qquad \alpha_k > 0$$

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