

Probabilistic Graphical Models

Jiří Materna





@mlcollegecom



@mlcollegecom



About me

- Ph.D. in Natural Language Processing and Artificial Intelligence at Masaryk University
- 10 years at seznam.cz (last 8 years as Head Of Research)
- Founder and co-organizer of ML Prague
- Mentor at StartupYard
- ML Freelancer and consultant

www.mlguru.com

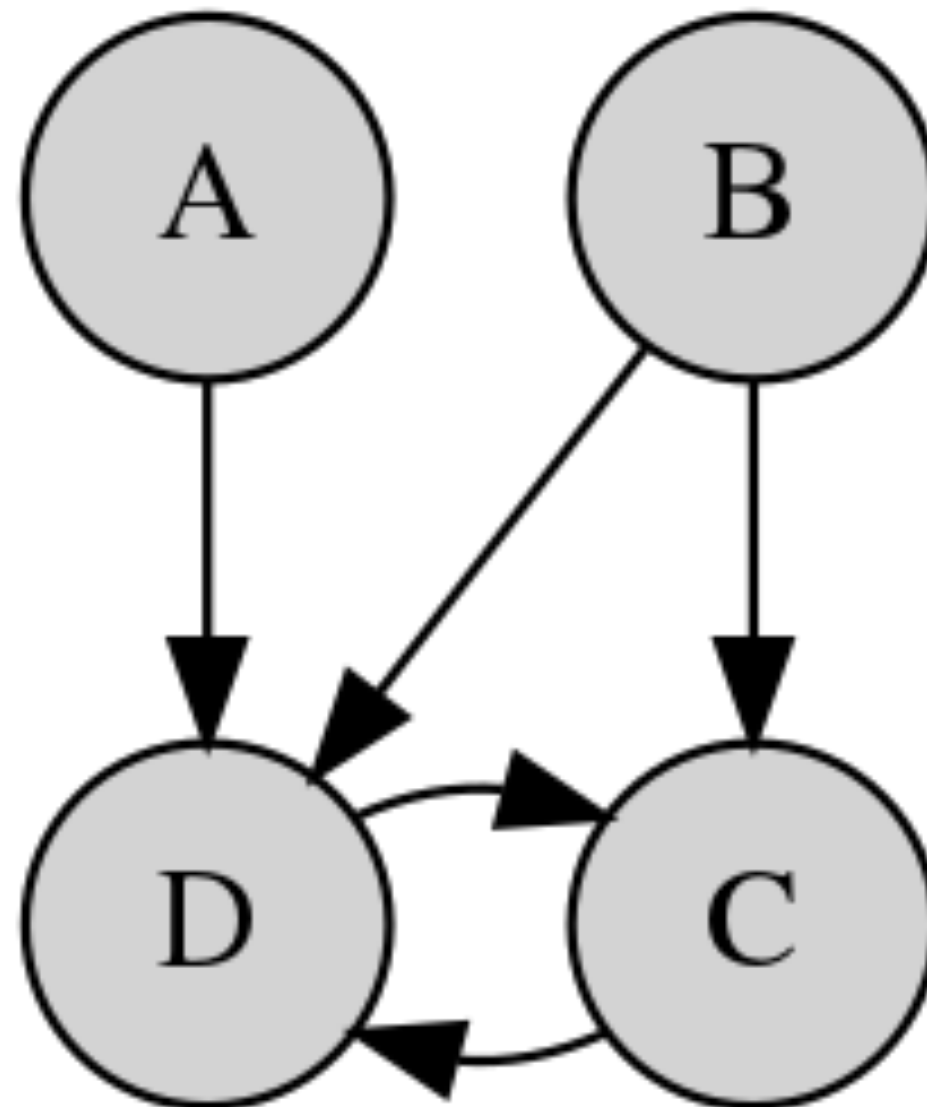
www.mlprague.com

www.mlcollege.com

Outline

- Topic modeling
- Basics of the Probability theory
- Probabilistic Graphical Models
- Inference in Bayesian Networks
- Gaussian Linear Regression
- Gaussian Mixtures for clustering
- (Probabilistic) Latent Semantic Analysis
- Latent Dirichlet Allocation

Probabilistic Graphical Models



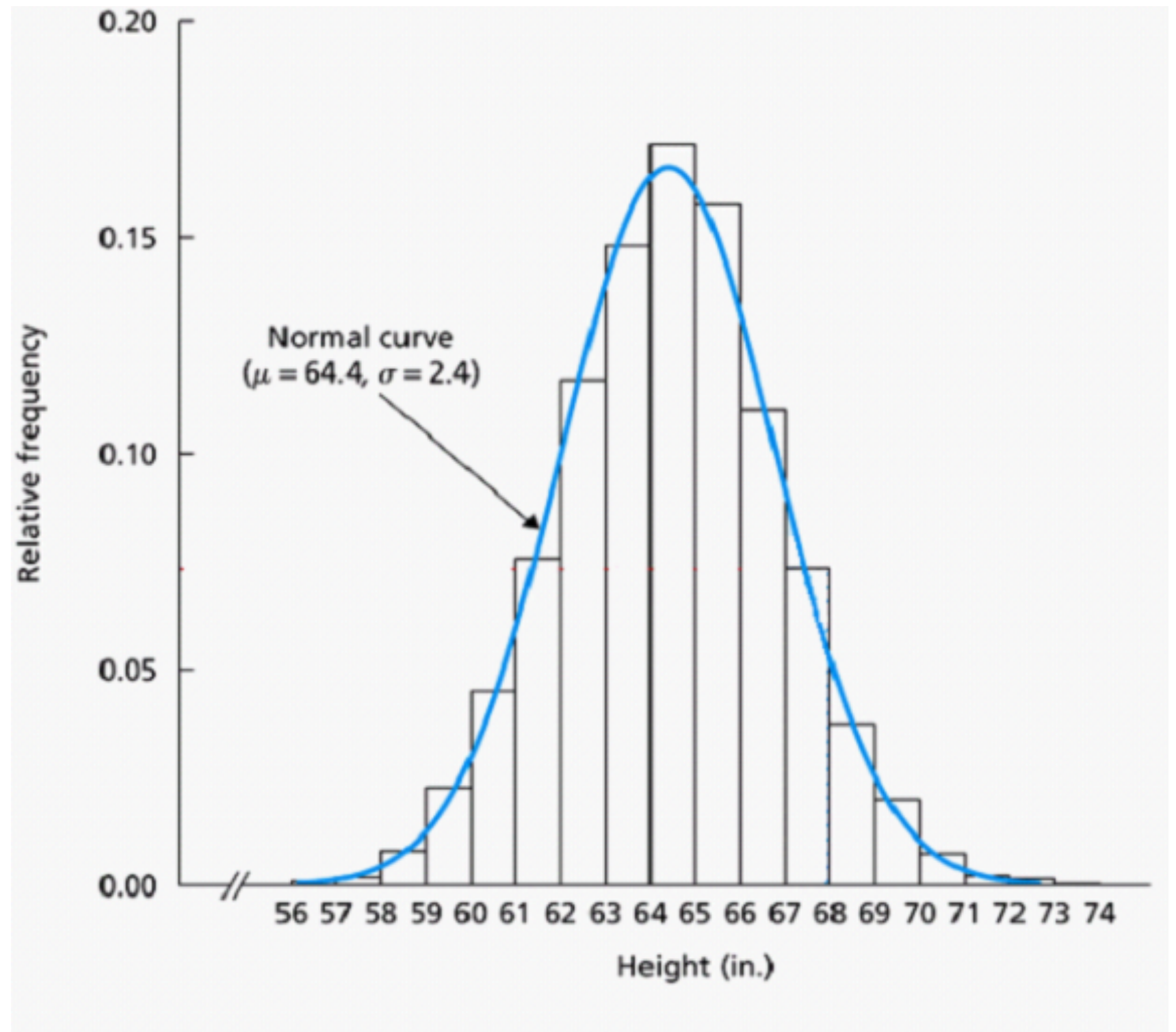
Conditional probability and independence

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

$$A \perp B \iff P(A \cap B) = P(A)P(B)$$

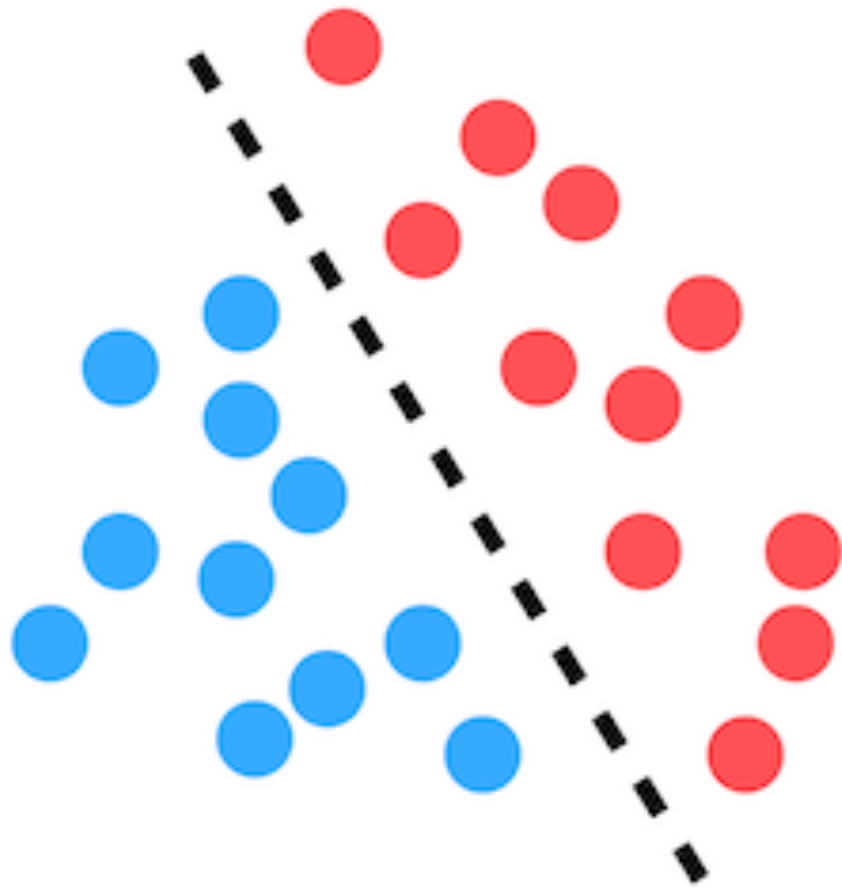
Probability distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

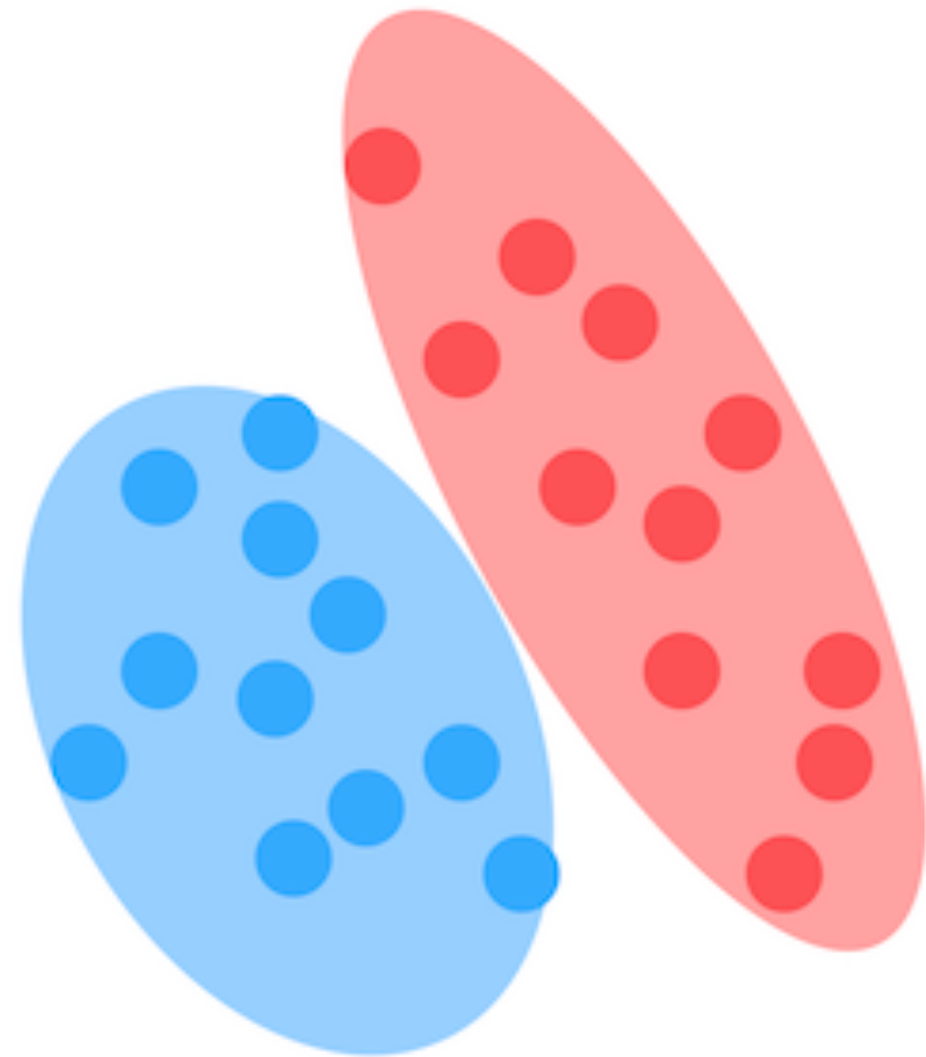


Discriminative vs. generative models

Discriminative



Generative



Topic Modeling

Topics

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
...

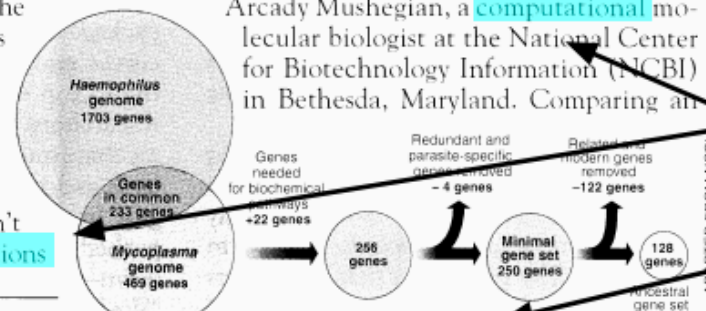
Documents

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

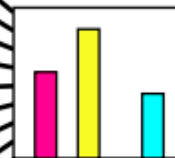


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions & assignments



Generative model of people's heights

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

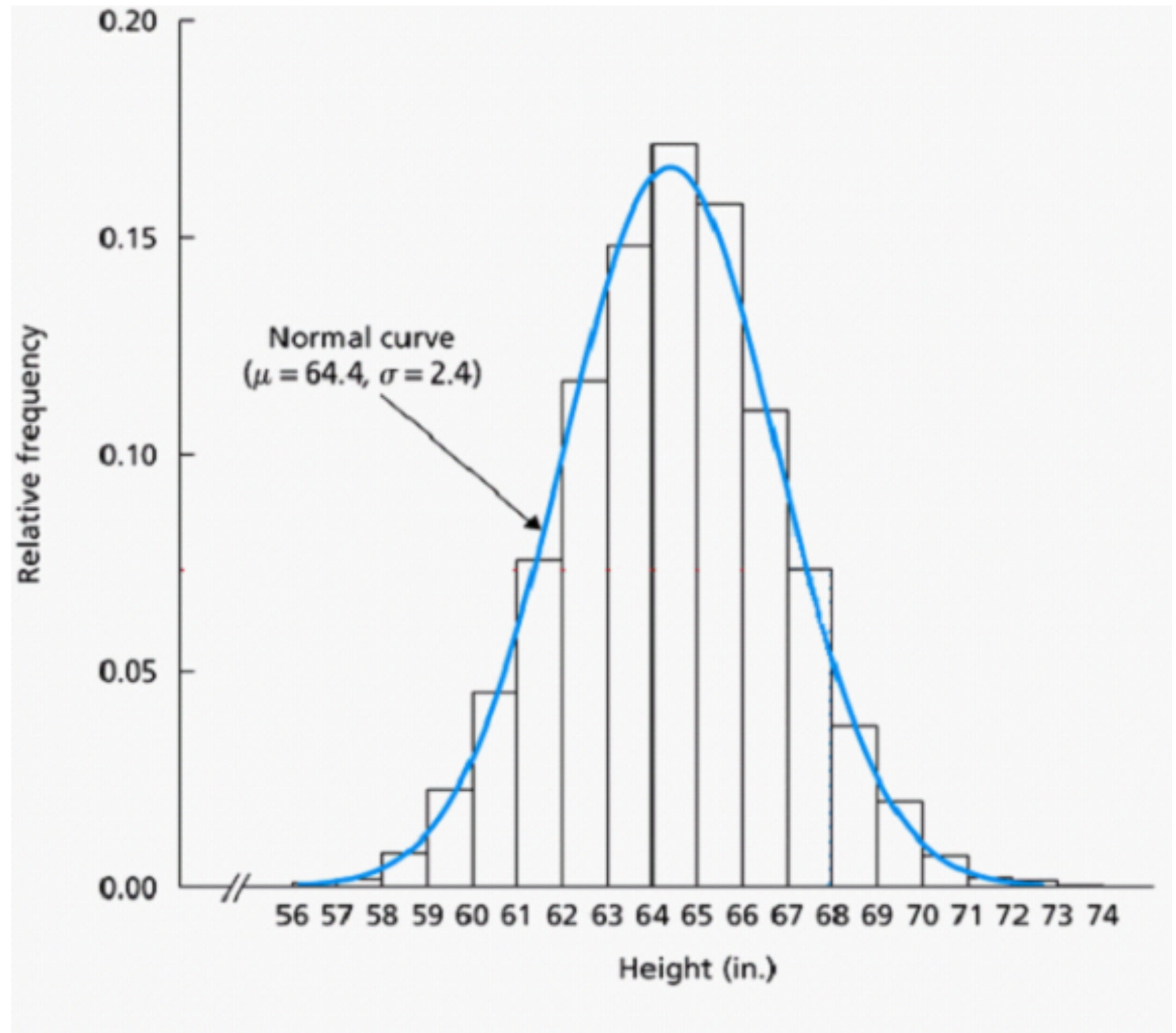
$$X = \{x_1, x_2 \dots x_n\}$$

$$X \sim N(\mu, \sigma^2), \alpha = (\mu, \sigma^2)$$

$$\bar{\alpha} = \arg \max_{\alpha} P(\alpha|X)$$

$$P(\alpha|X) = \frac{P(X|\alpha) \cdot P(\alpha)}{P(X)}$$

posterior
likelihood
prior

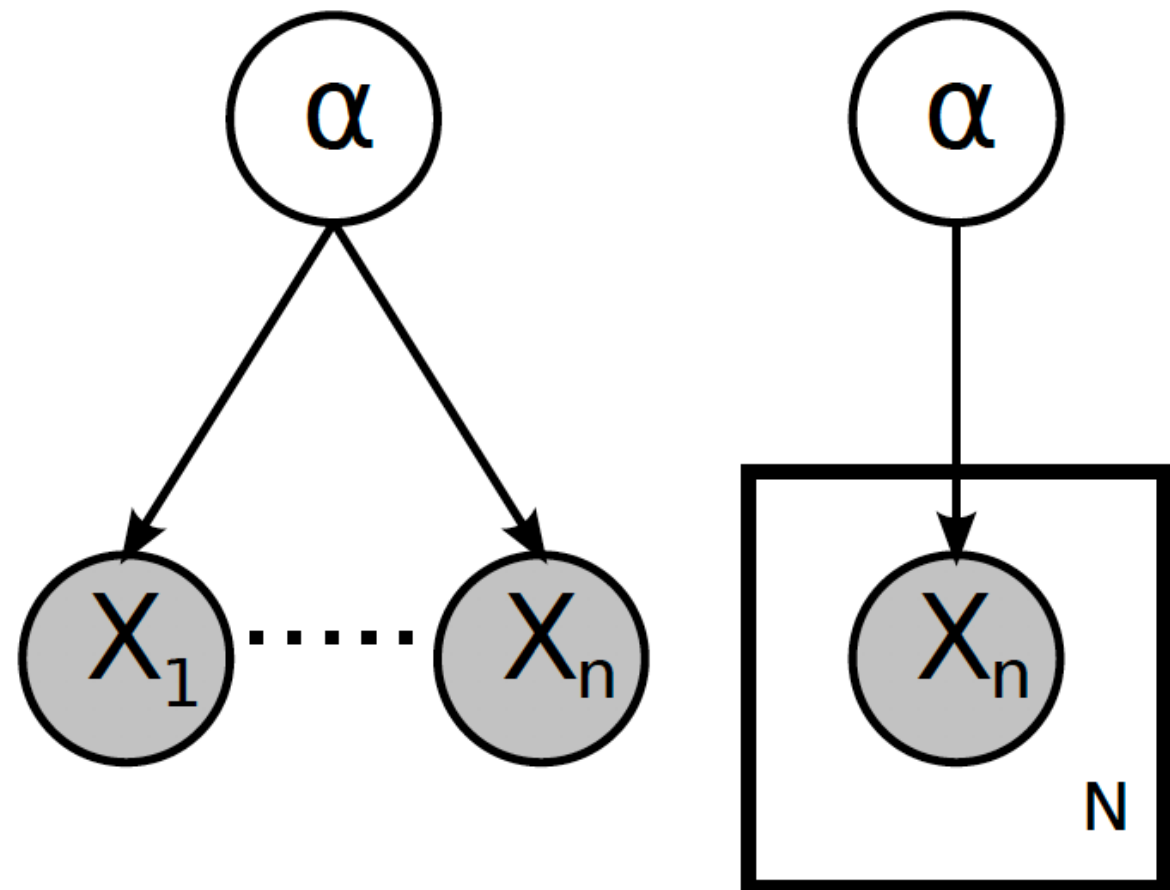


Probabilistic graphical models

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$X = \{x_1, x_2 \dots x_n\}$$

$$X \sim N(\mu, \sigma^2), \alpha = (\mu, \sigma^2)$$



Inference in graphical models

$$P(\alpha|X) = \frac{P(X|\alpha).P(\alpha)}{P(X)} \propto P(X|\alpha).P(\alpha) = \prod_{i=1}^n P(x_i|\alpha).P(\alpha)$$

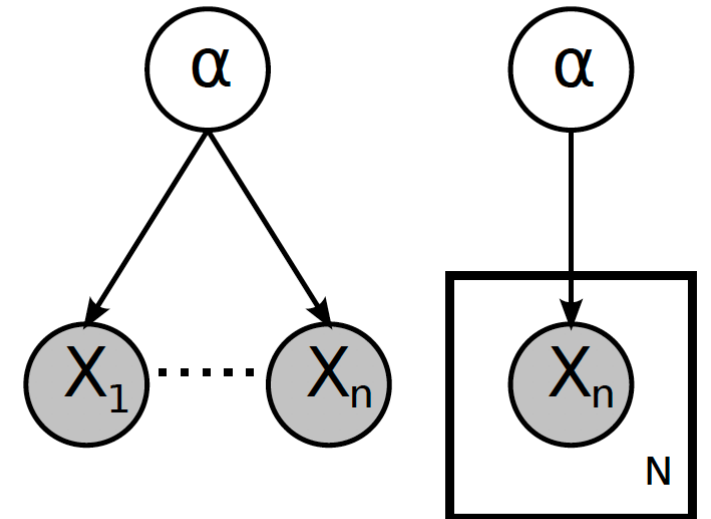
$$\bar{\alpha} = \arg \max_{\alpha} P(\alpha|X)$$

Variational inference

1. Approximate the posterior function with a simpler one
2. Compute the hidden variables by minimization of KL Divergence of the true and simpler distributions

Sampling (e.g. Gibbs sampling)

1. Draw samples from the true posterior
2. Compute mean of the samples



Generative model for linear regression

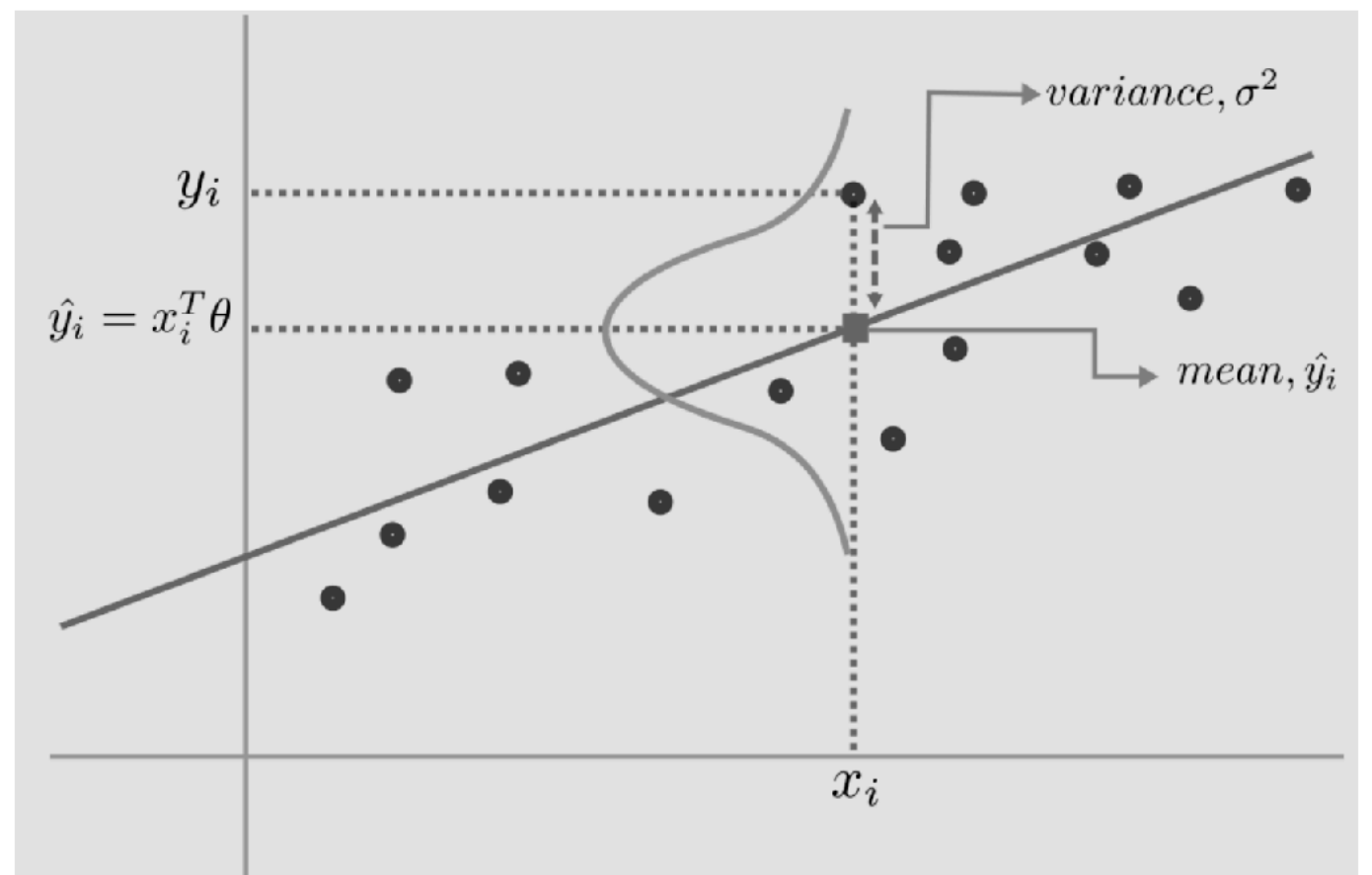
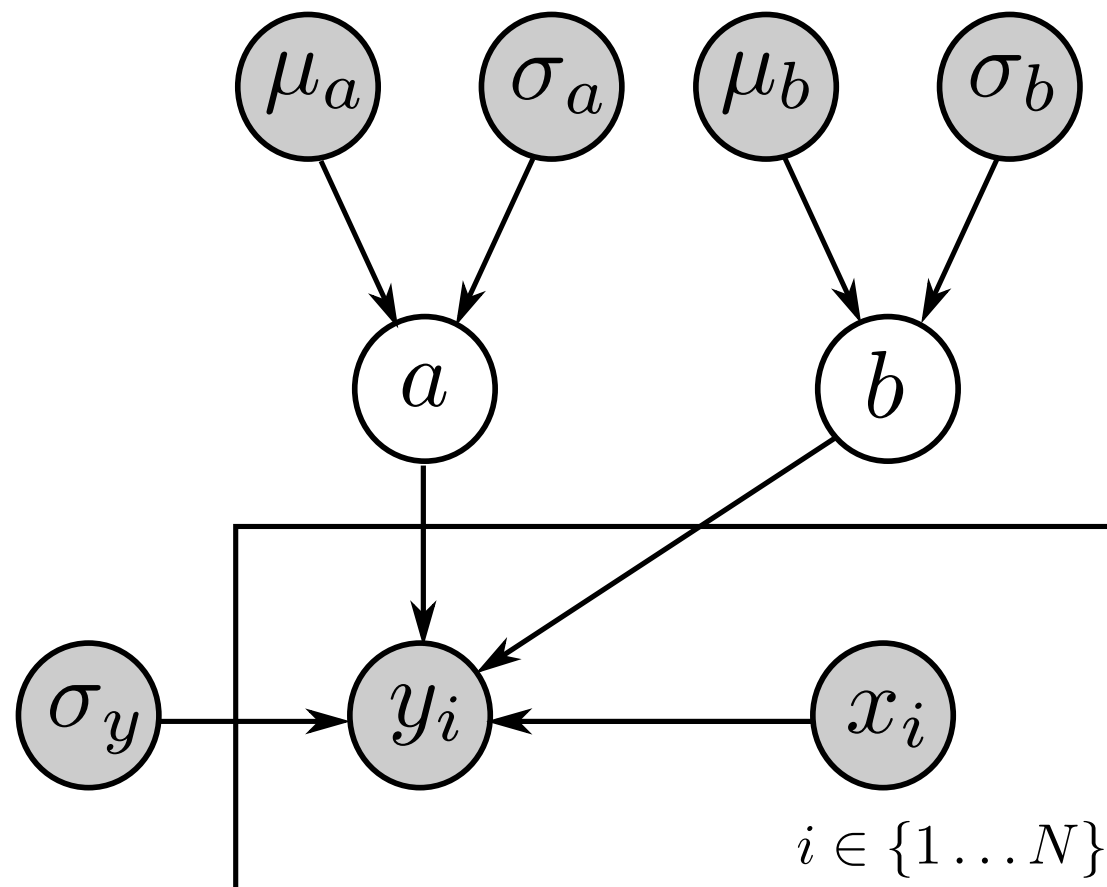
$$\mathbf{x} = \{x_1, x_2, \dots, x_N\}$$

$$f(\mathbf{x}) = a\mathbf{x} + b$$

$$\mathbf{y} \sim \mathcal{N}(a\mathbf{x} + b, \sigma_y)$$

$$a \sim \mathcal{N}(\mu_a, \sigma_a)$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b)$$



Generative model for linear regression in Edward

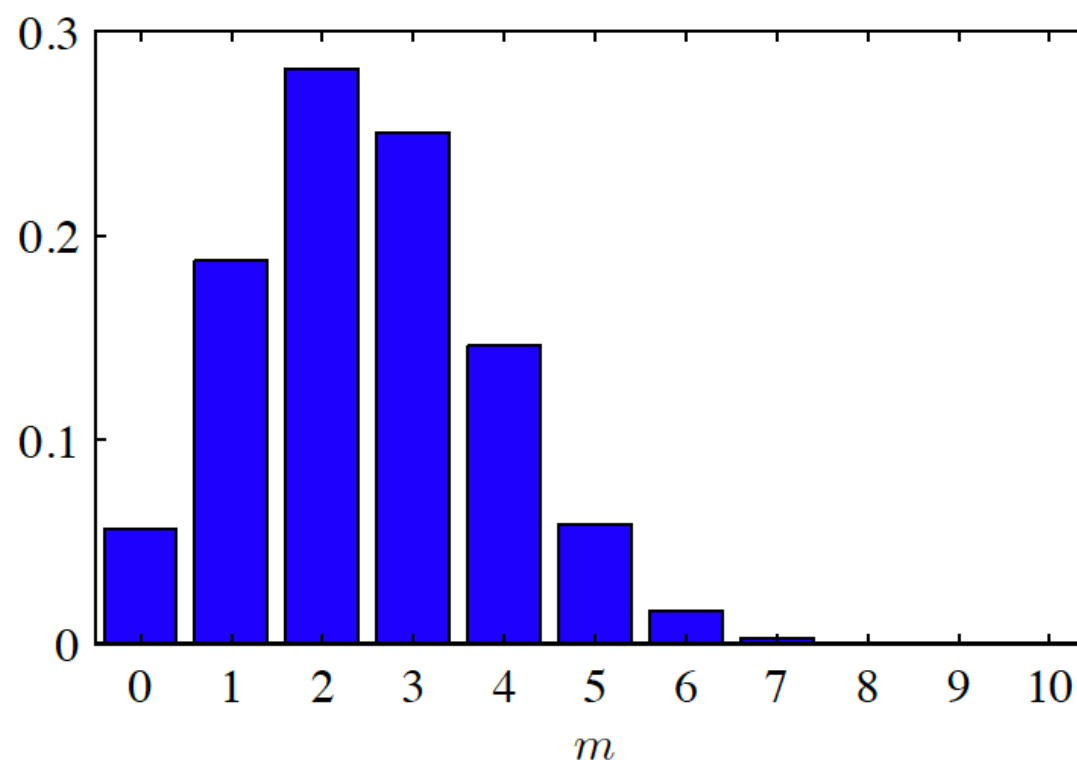
01-Generative-linear-regression-edward.ipynb

Binomial distribution

$$\text{Bin}(k|n, p) = \binom{n}{k} p^k \cdot (1 - p)^{n-k} =$$

$$\text{Bin}(x_1, x_2|p_1, p_2) = \frac{(x_1 + x_2)!}{x_1! x_2!} p_1^{x_1} \cdot p_2^{x_2}$$

$$p_1 + p_2 = 1$$



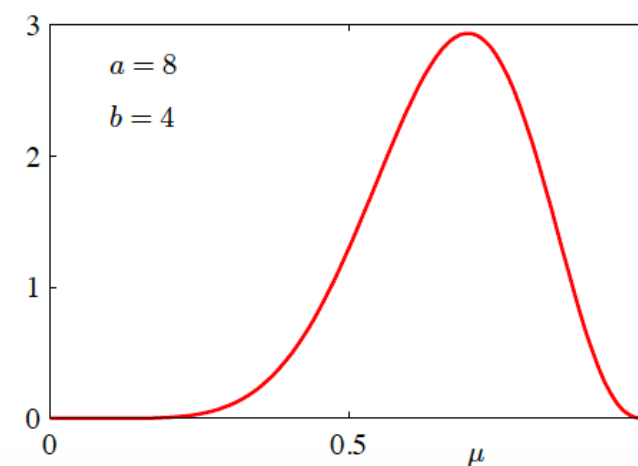
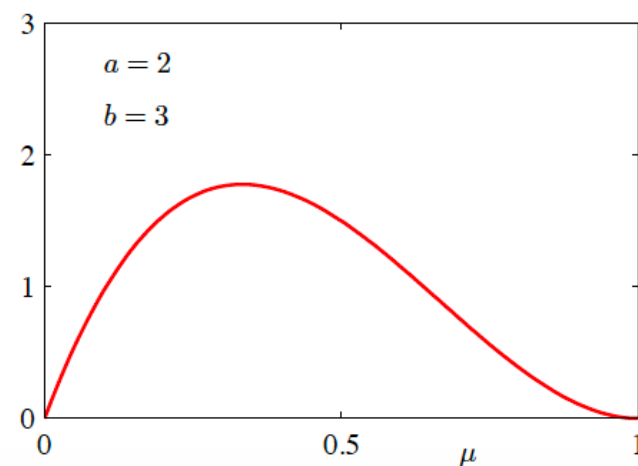
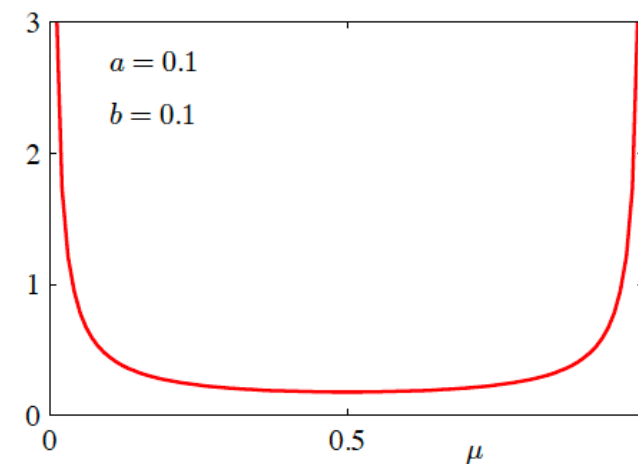
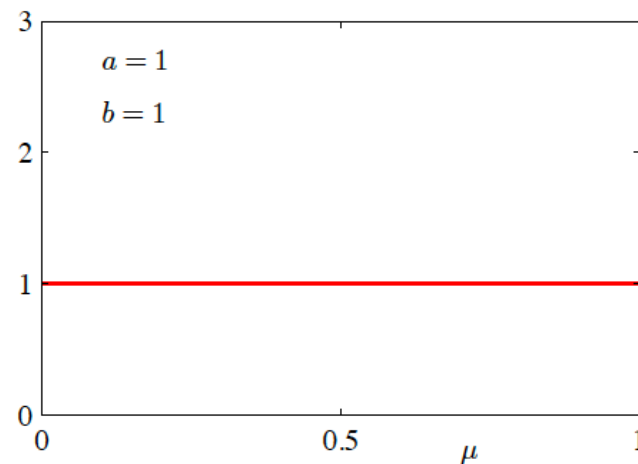
Example: $n = 10, p = 0.25$

Beta distribution

$$\text{Beta}(p_1, p_2 | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_1^{\alpha-1} \cdot p_2^{\beta-1}$$

$$p_1 + p_2 = 1$$

$$\Gamma(x) = (x-1)!$$



Multinomial and Dirichlet distributions

Multinomial

$$\text{Mult}(x_1 \dots x_n | p_1 \dots p_n) = \frac{(\sum x_i)!}{\prod x_i!} \prod_{i=1}^n p_i^{x_i}$$

Dirichlet

$$\text{Dir}(p_1 \dots p_n | \alpha_1 \dots \alpha_n) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod_{i=1}^n p_i^{\alpha_i - 1}$$

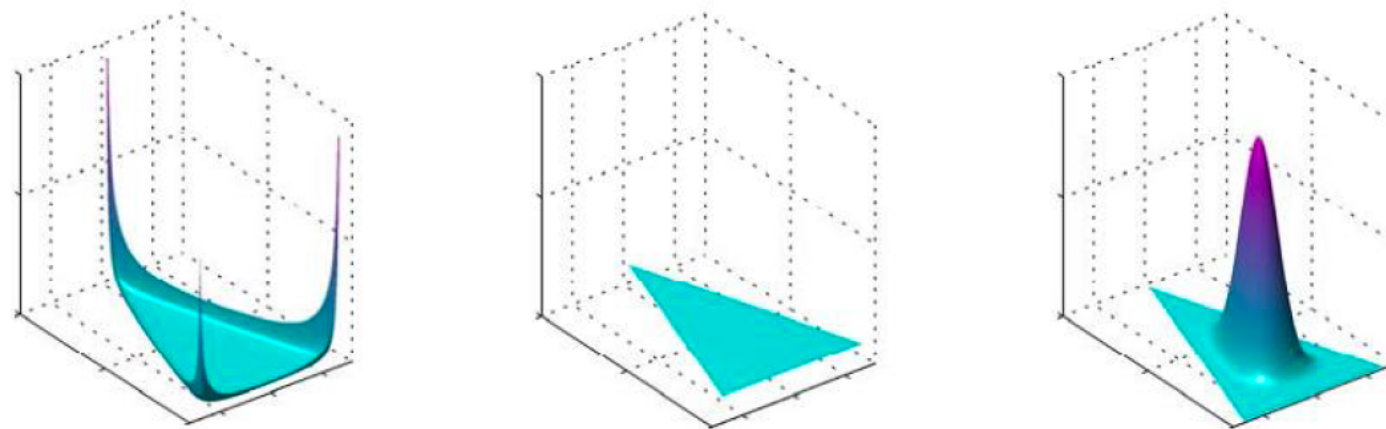
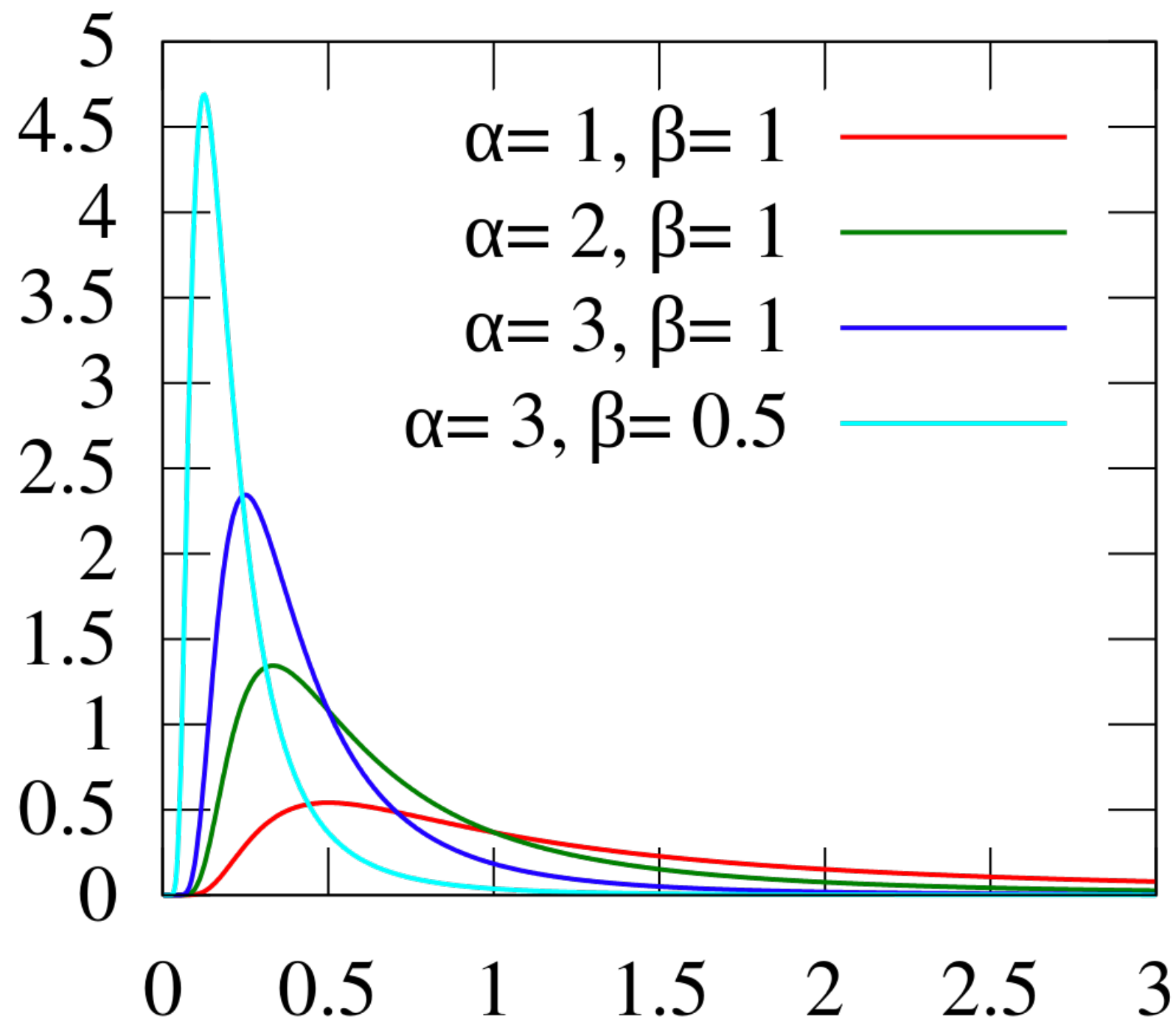
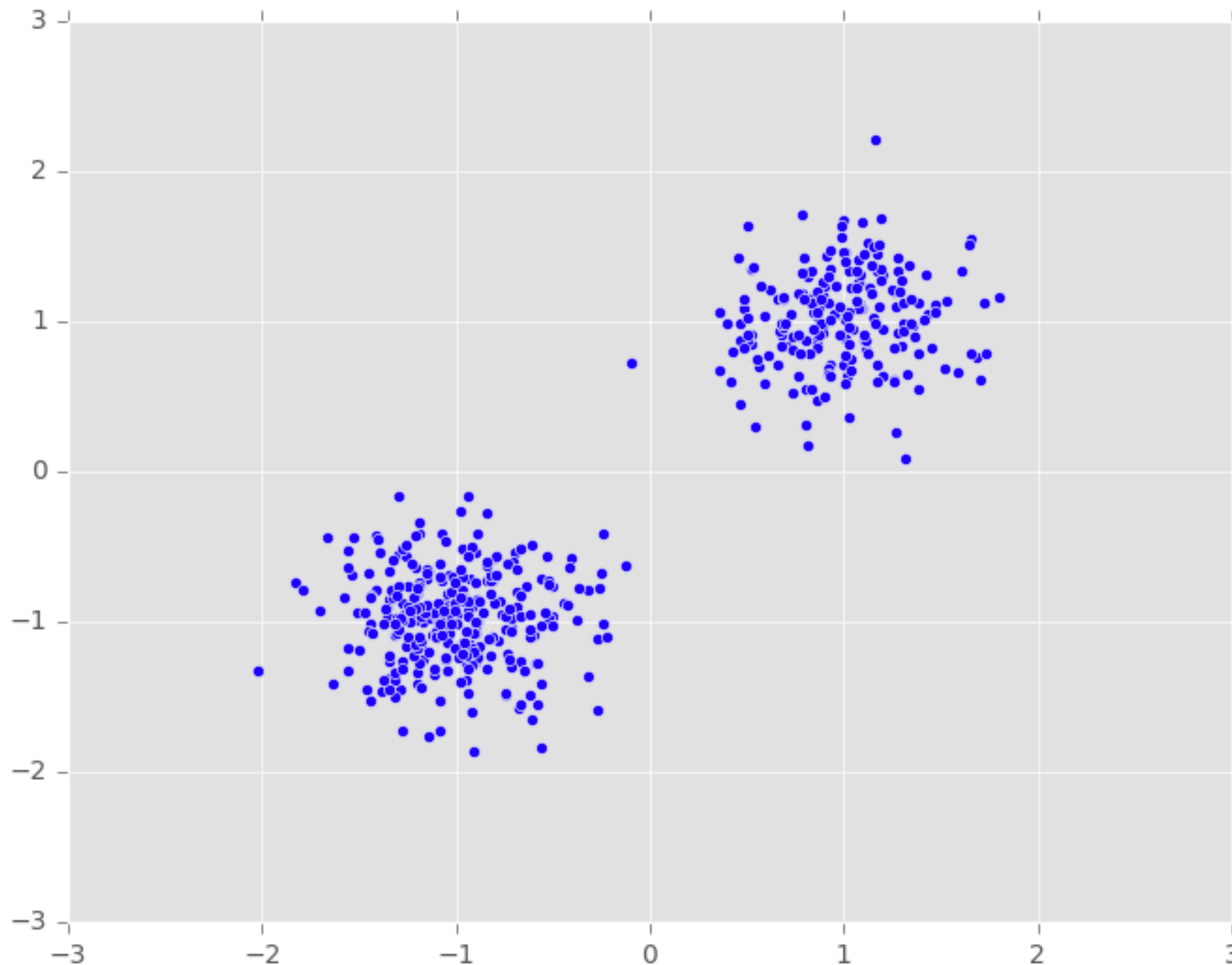


Figure 2.5 Plots of the Dirichlet distribution over three variables, where the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density. Here $\{\alpha_k\} = 0.1$ on the left plot, $\{\alpha_k\} = 1$ in the centre plot, and $\{\alpha_k\} = 10$ in the right plot.

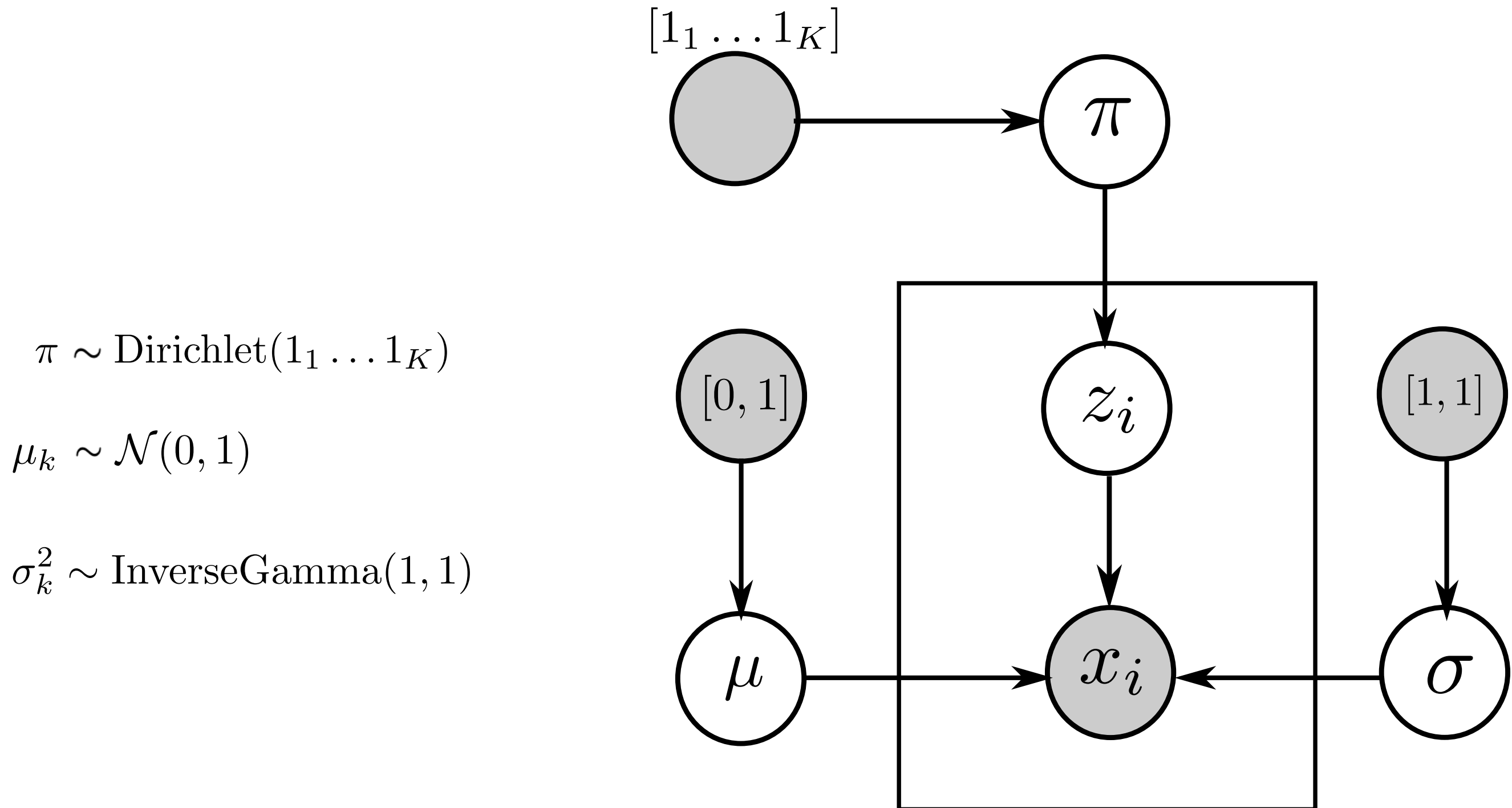
Inverse-gamma Distribution



Clustering as gaussian mixtures



Clustering as gaussian mixtures



Gibbs sampling

- 1 Initialize $z_i : i \in 1, \dots, M$
- 2 For $\tau \in 1, \dots, T$:
 - Sample $z_1^{(\tau+1)} \sim P(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \dots, z_M^{(\tau)})$
 - Sample $z_2^{(\tau+1)} \sim P(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \dots, z_M^{(\tau)})$
 - Sample $z_3^{(\tau+1)} \sim P(z_3 | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \dots, z_M^{(\tau)})$
 - ...
 - Sample $z_M^{(\tau+1)} \sim P(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \dots, z_{M-1}^{(\tau+1)})$

Clustering as gaussian mixtures

02-clustering-edward.ipynb

Topic Modeling

Topics

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
...	

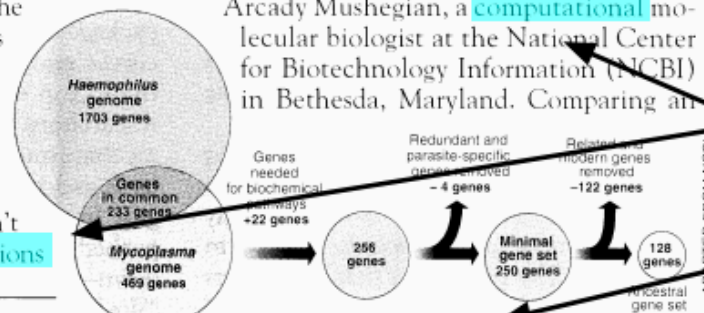
Documents

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

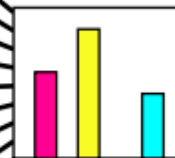


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions & assignments



Latent Semantic Analysis

Topic modeling using LSA example:

$$\textit{soccer} = 1.8 * \textit{'soccer'} + 0.4 * \textit{'ball'} + 0.2 * \textit{'FIFA'} - 0.4 * \textit{'tennis'}$$

$$\textbf{doc} = 2.3 * \textit{soccer} + 1.8 * \textit{sport} + 0.9 * \textit{Europe} + 0.8 * \textit{news}$$

Latent Semantic Analysis

Doc₁: *Machine learning helps people to understand data.*

Doc₂: *Data can be understood using machine learning.*

Doc₃: *People can use machine learning for data understanding.*

	Doc ₁	Doc ₂	Doc ₃										
be	0	1	0										
can	0	1	1										
data	1	1	1										
for	0	0	1										
helps	1	0	0										
learning	1	1	1										
machine	1	1	1										
people	1	0	1										
to	1	0	0										
understand	1	0	0										
understanding	0	0	1										
understood	0	1	0										
use	0	0	1										
using	0	1	0										

6x4

TOPICS

TERMS

X

TOP

0

0

0

0

IC

0

0

0

0

0

IMPO

0

0

0

0

RTAN

CE

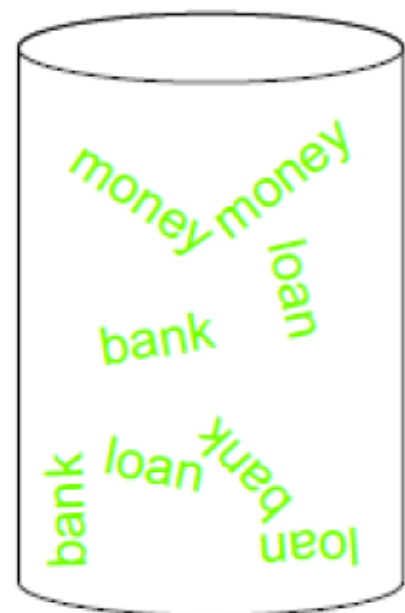
X

4x4

DOCUMENTS

TOPICS

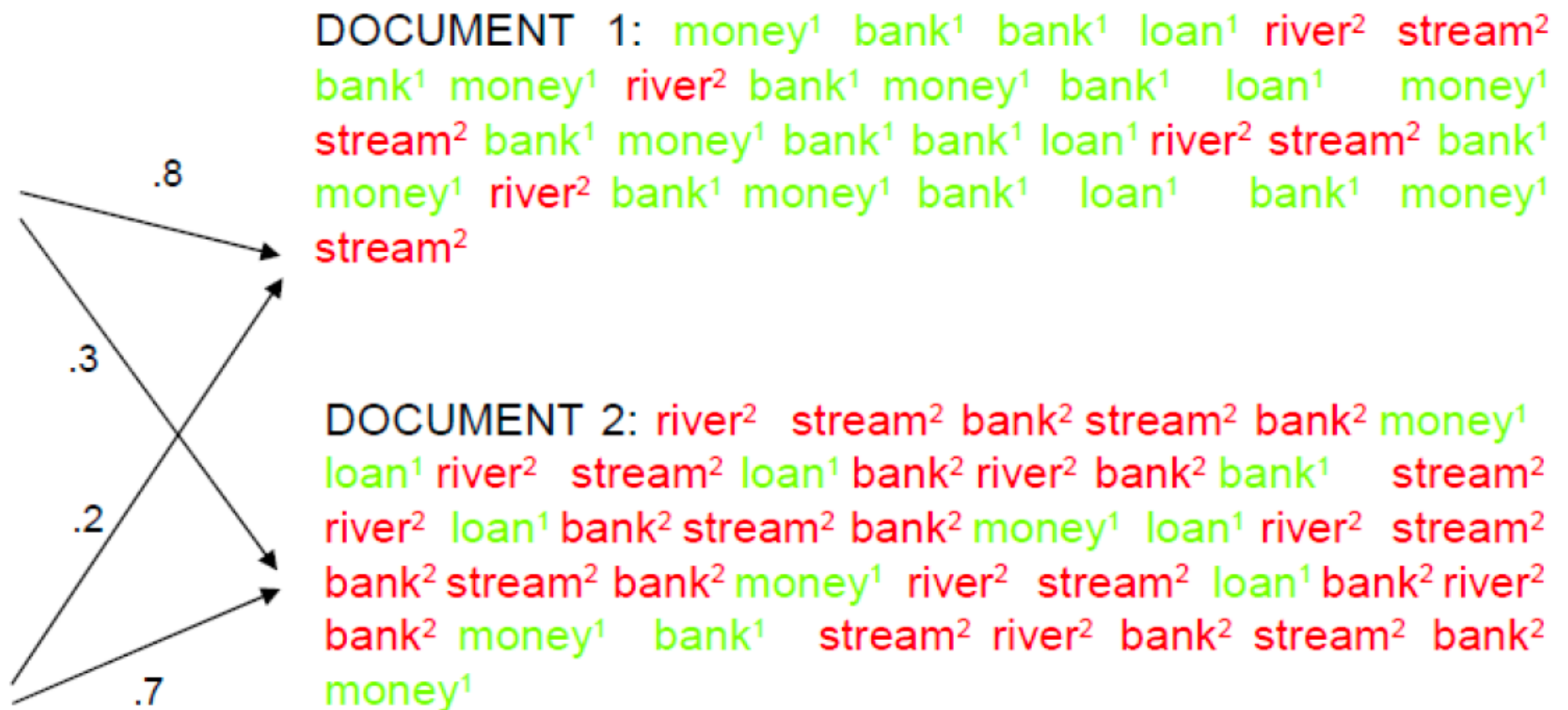
Probabilistic Latent Semantic Analysis



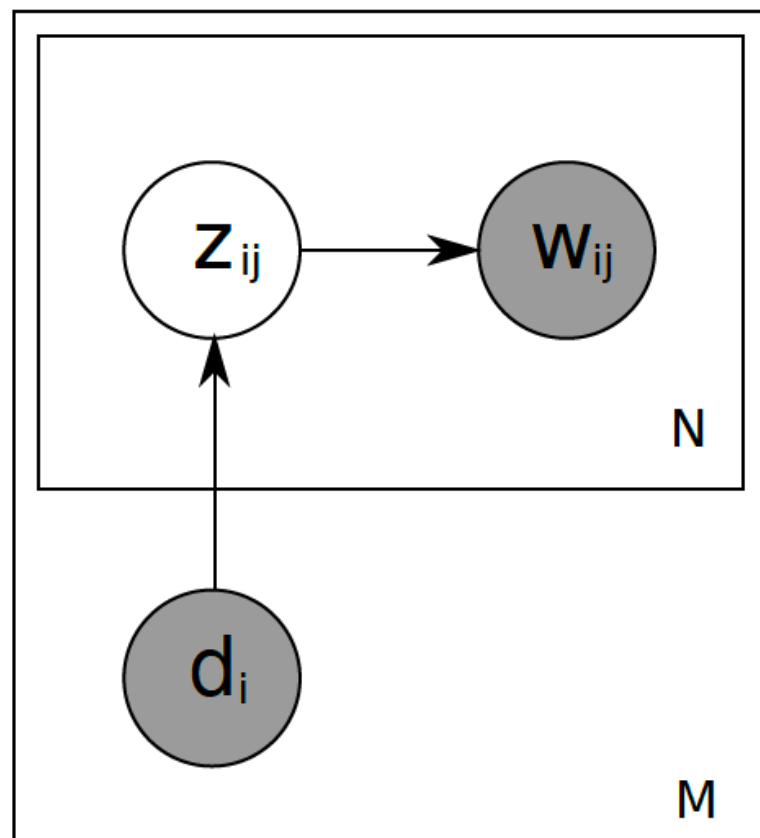
TOPIC 1



TOPIC 2



Model of Probabilistic Latent Semantic Analysis



```
1: for  $i \in \{1, 2, \dots, N\}$  do  
2:   for  $j \in \{1, 2, \dots, M\}$  do  
3:     Choose a latent topic  $z_{ij}$  with probability  $P(z_{ij}|d_i)$   
4:     Choose a word  $w_{ij}$  with probability  $P(w_{ij}|z_{ij})$   
5:   end for  
6: end for
```

Probabilities are computed from frequency analysis of words

Not a generative model (works for training data only)

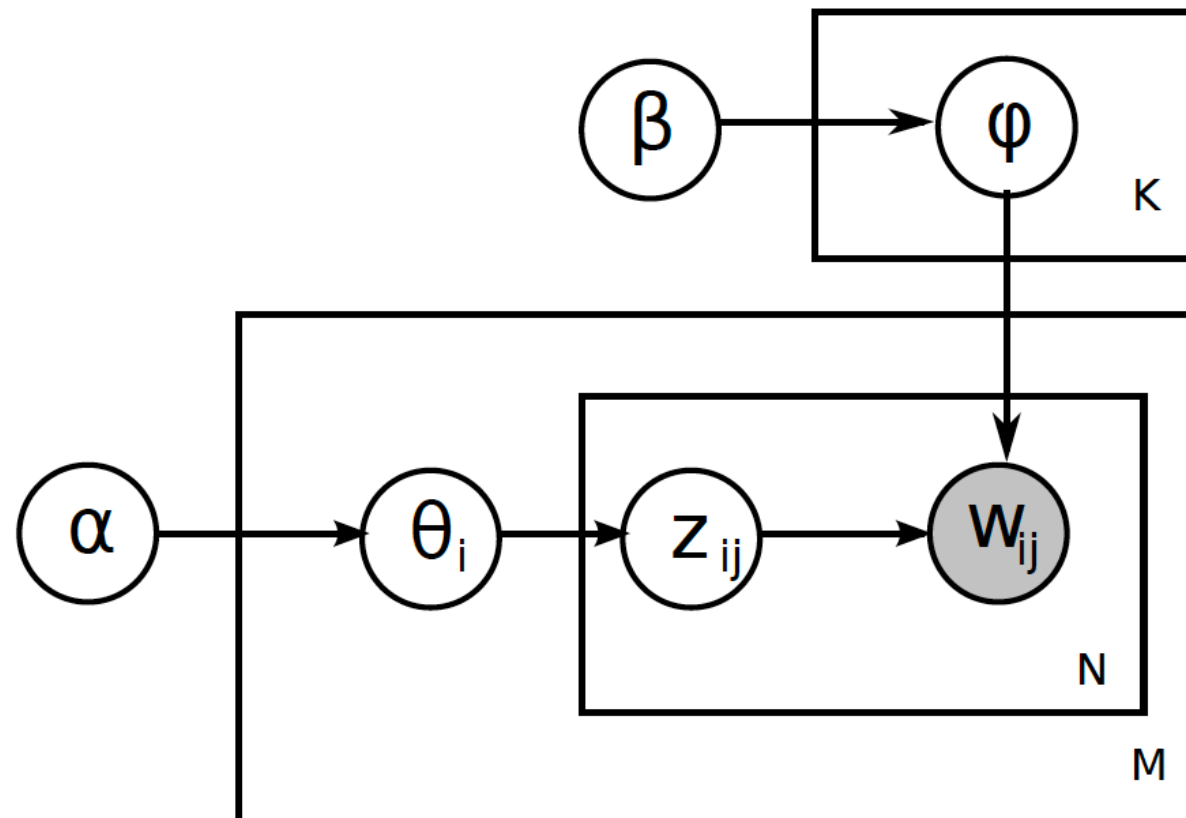
Latent Dirichlet Allocation

For each document $i \in 1 \dots M$ choose $\theta_i \sim \text{Dir}(\alpha)$

For each word position $j \in \dots N_i$ choose topic $z_{i,j} \in 1 \dots K$,

$z_{i,j} \sim \text{Mult}(\theta_i)$

For each word position j choose word $w_{i,j} \sim \text{Mult}(\varphi_{z_{i,j}})$



Latent Dirichlet Allocation

Topics

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
...

Documents

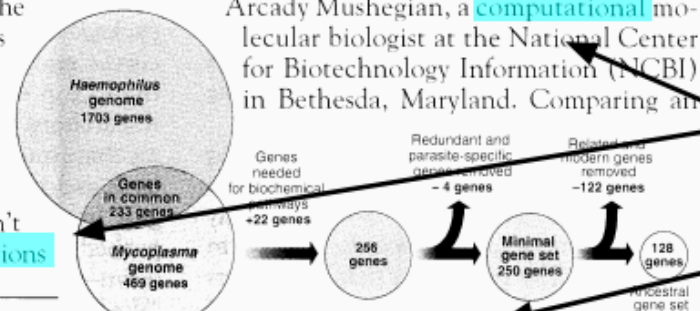
Seeking Life's Bare (Genetic) Necessities

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**

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

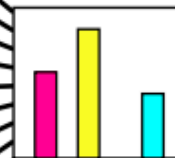
"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **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.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions & assignments



Topic modeling

03_Topic_modeling.ipynb

Thank you for your attention

e-mail: jiri@mlcollege.com

Web: www.mlcollege.com

Twitter: @JiriMaterna

Facebook: <https://www.facebook.com/maternajiri>

LinkedIn: <https://www.linkedin.com/in/jirimaterna/>