Probabilistic Graphical Models

Jiří Materna

Machine Learning College

@mlcollegecom

@mlcollegecom

slack

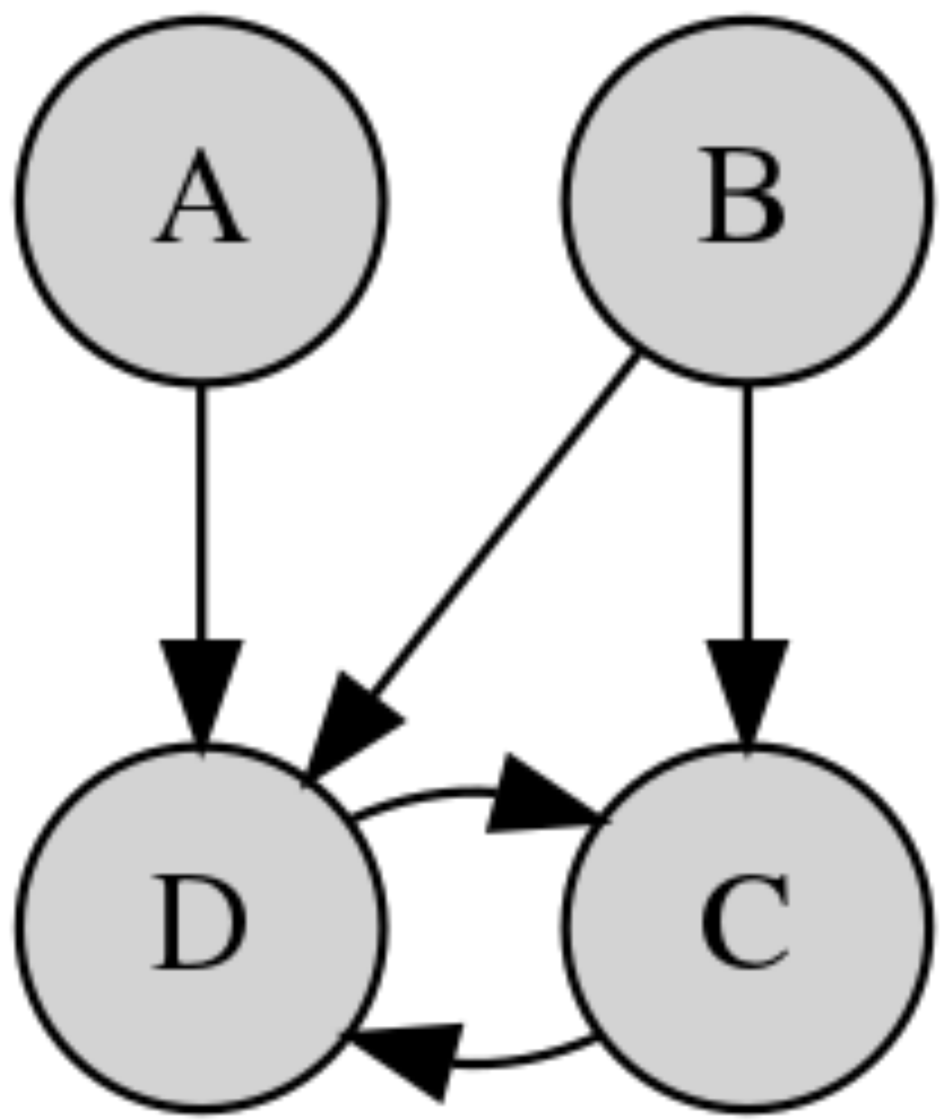
About me

- 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

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



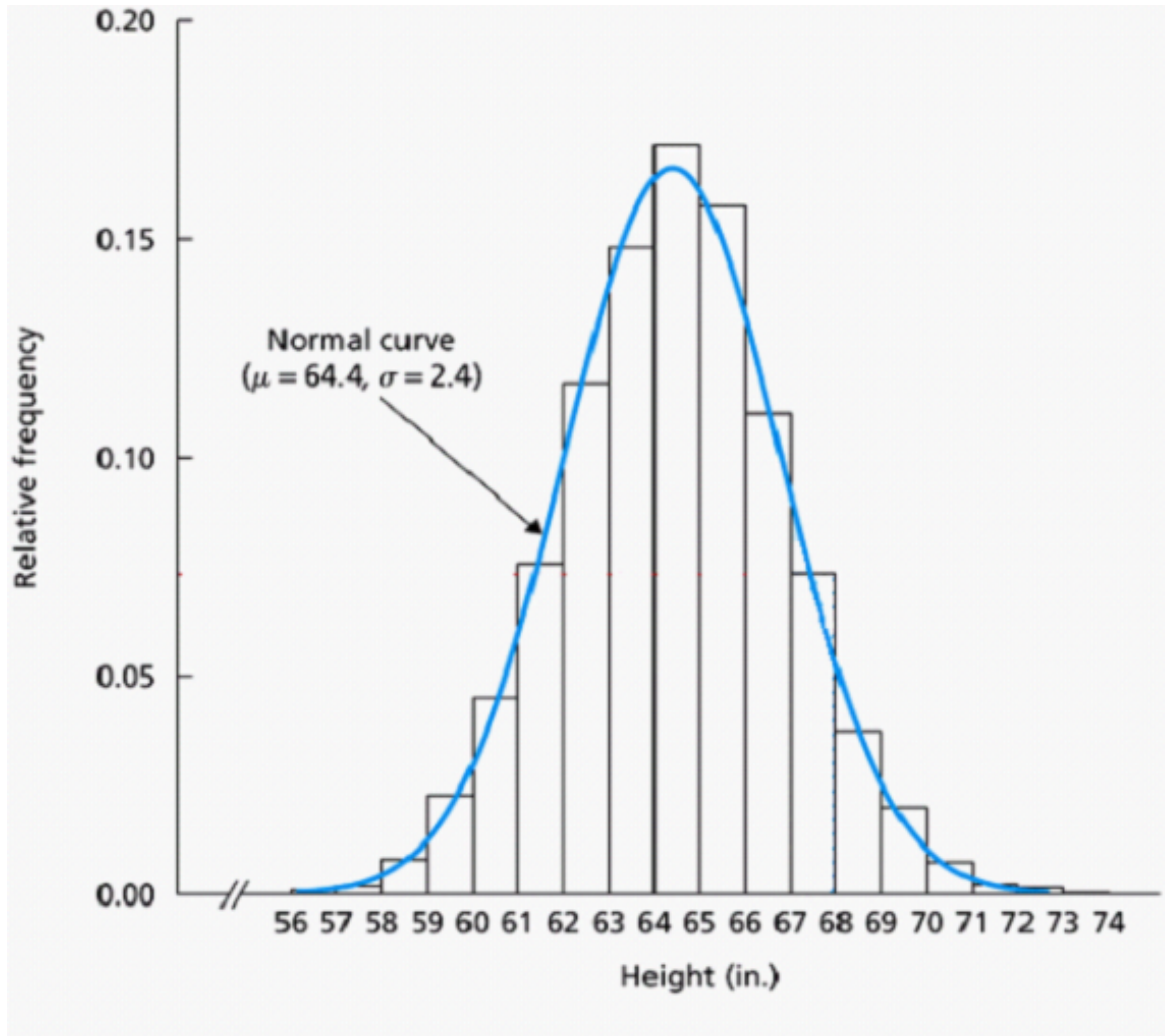
Conditional probability and independence

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

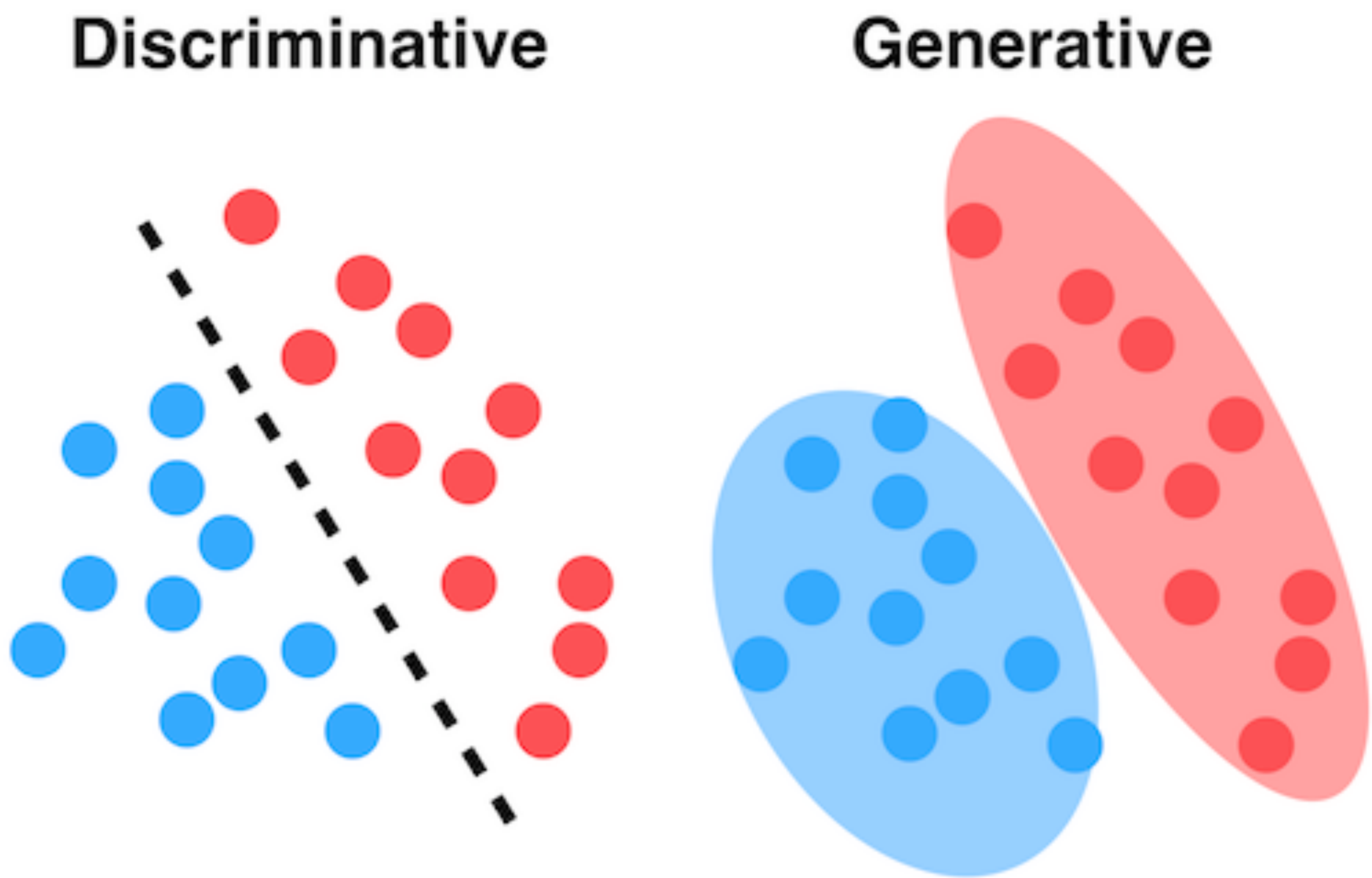
$$A \perp B \iff P(A \cap B) = P(A)P(B)$$

Probability distribution

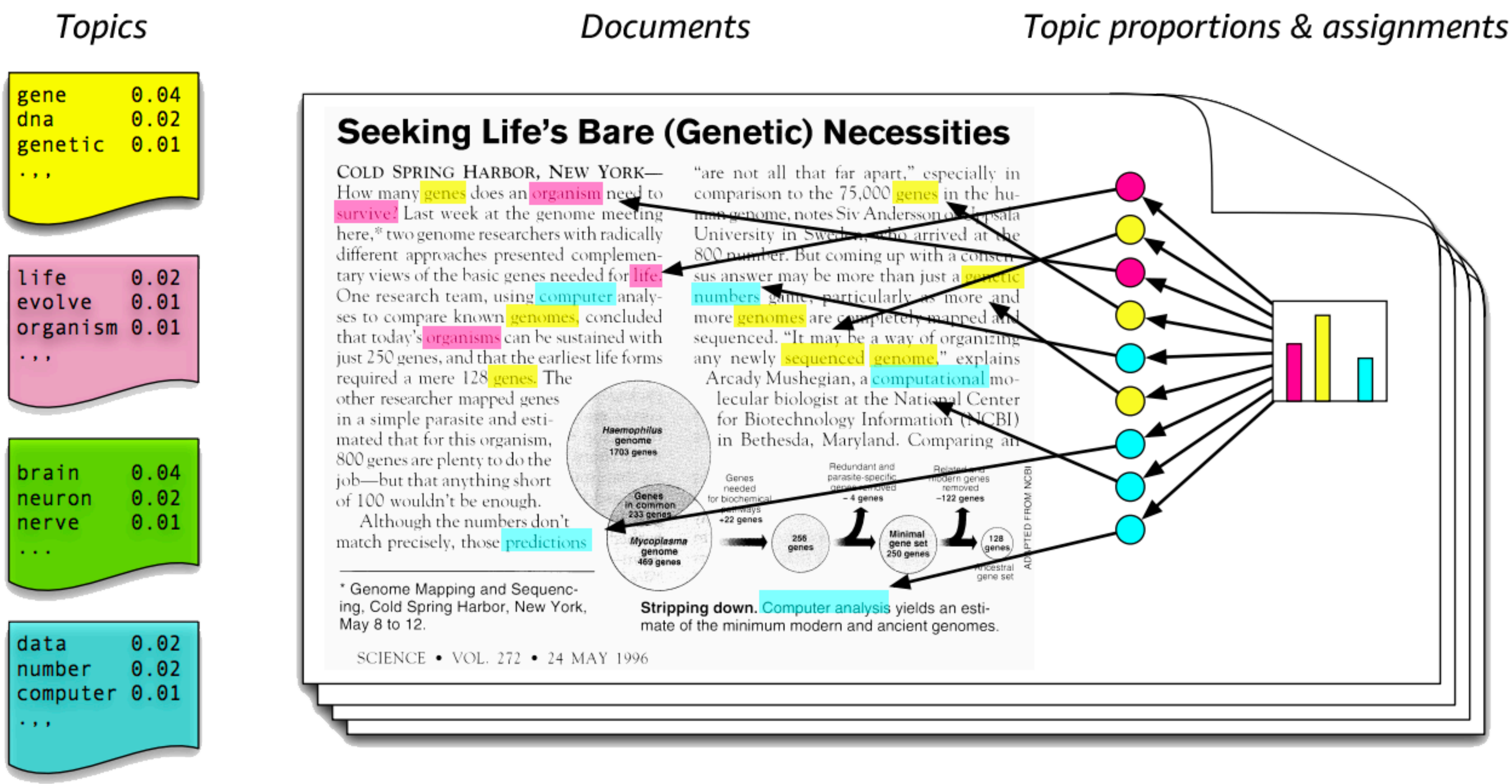
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Discriminative vs. generative models



Topic Modeling



https://kldavenport.com/topic-modeling-amazon-reviews/

Generative model of people's heights

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

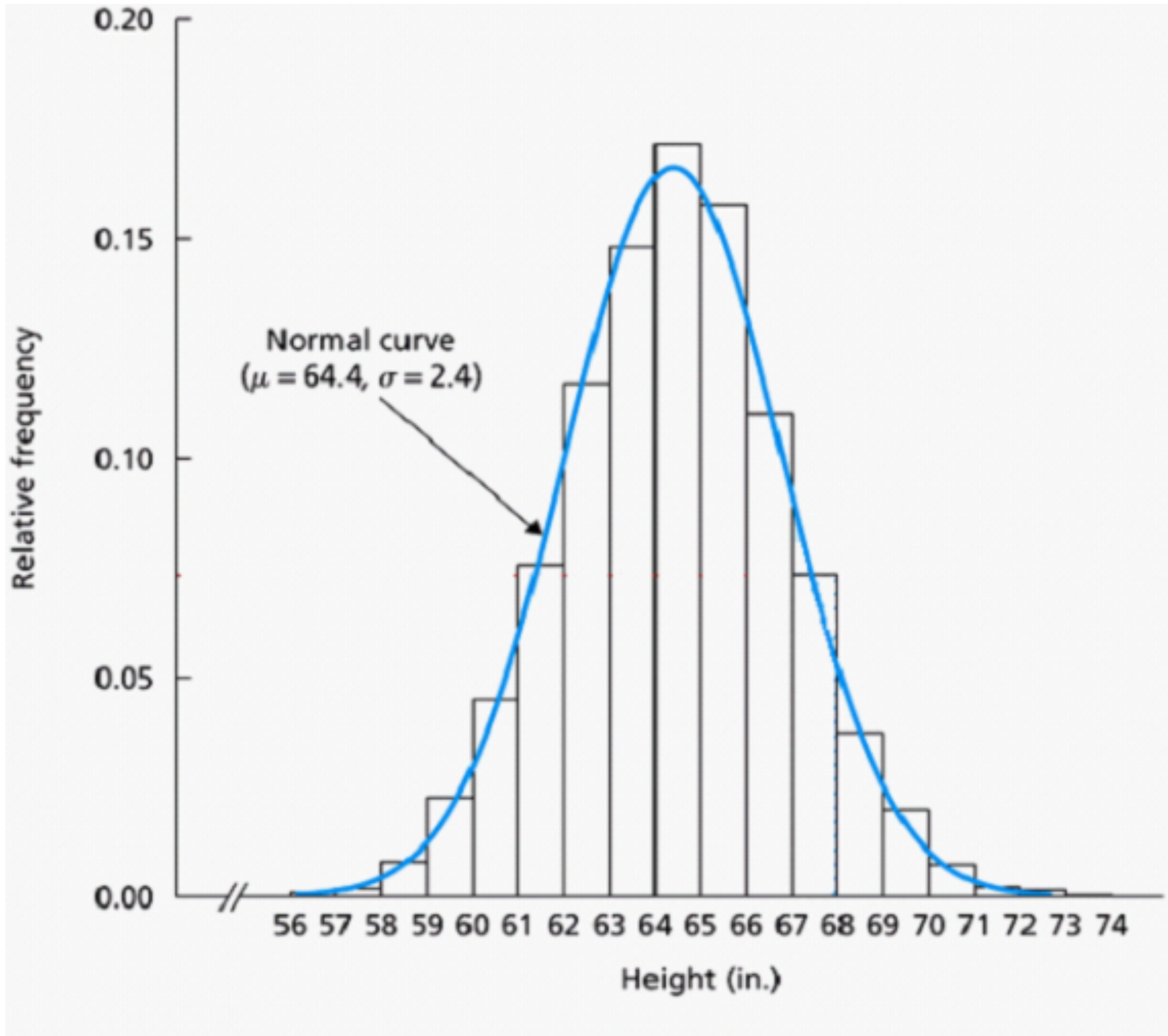
$$X = \{x_1, x_2 \ldots x_n\}$$

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

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

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

posterior
likelihood
prior

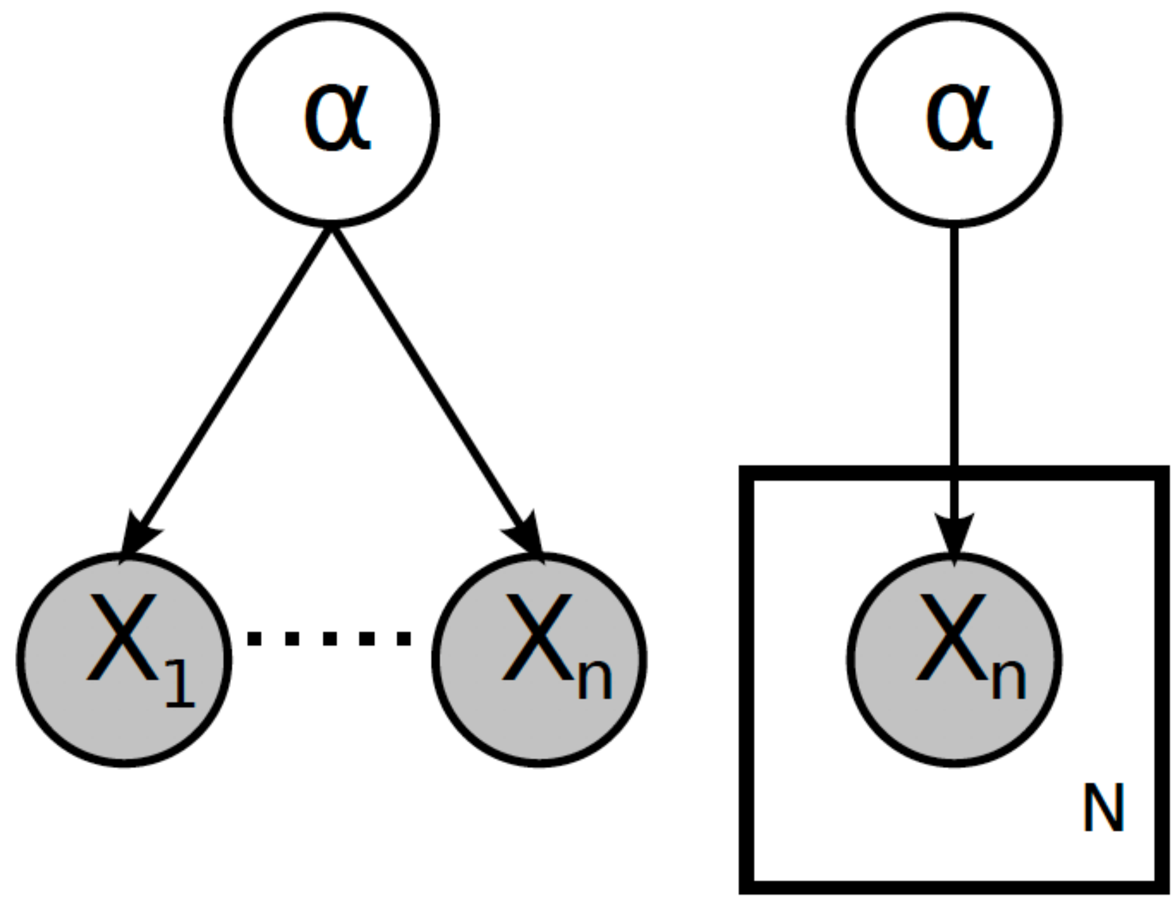


Probabilistic graphical models

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

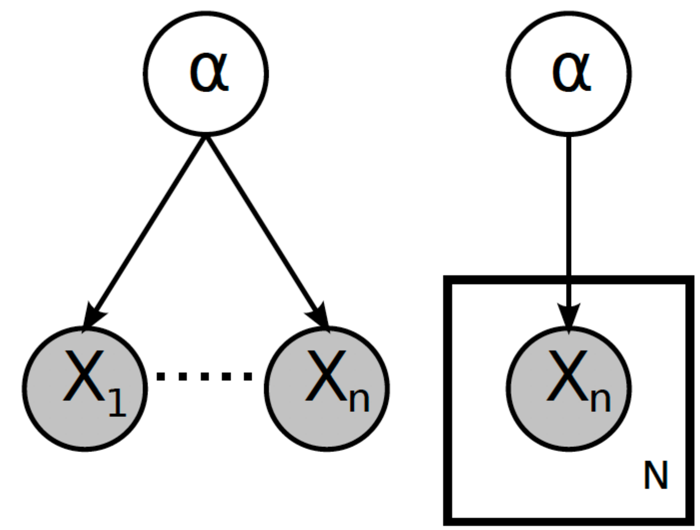
$$X = \{x_1, x_2 \ldots x_n\}$$

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



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

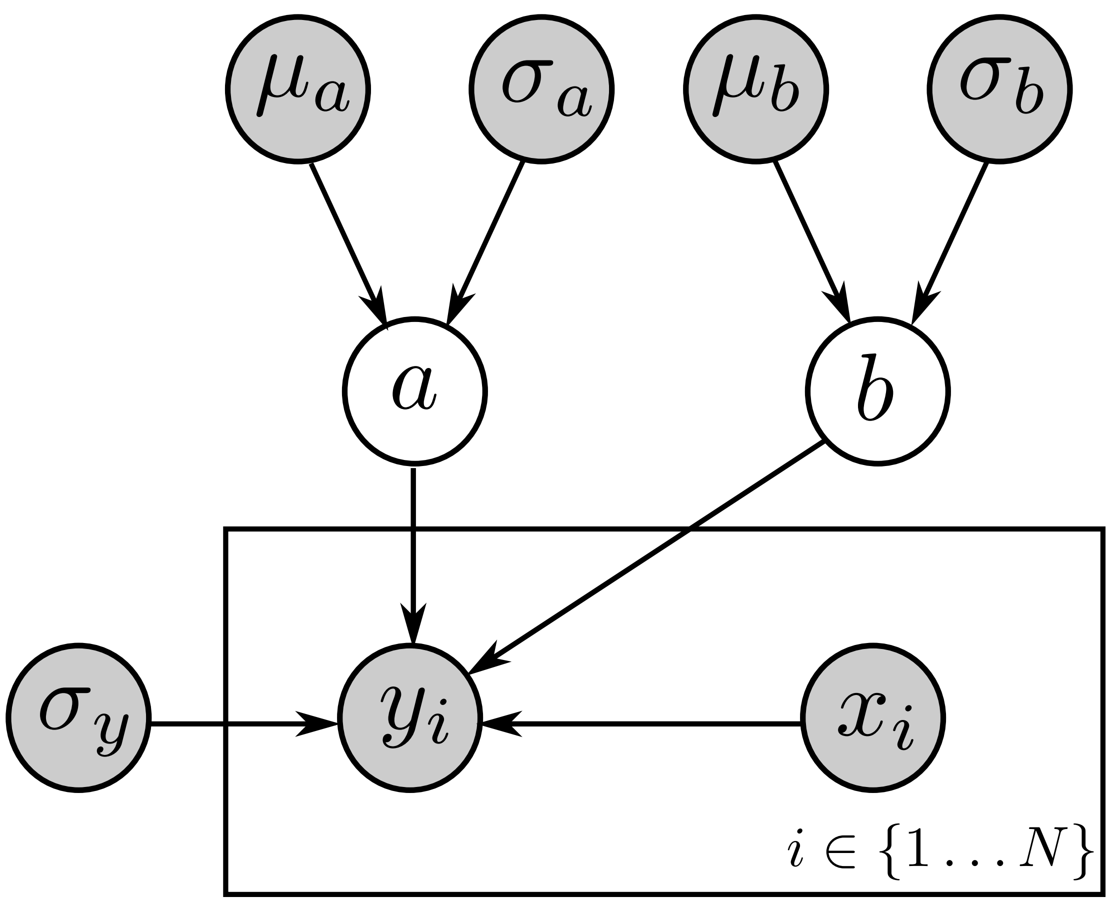
$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$$

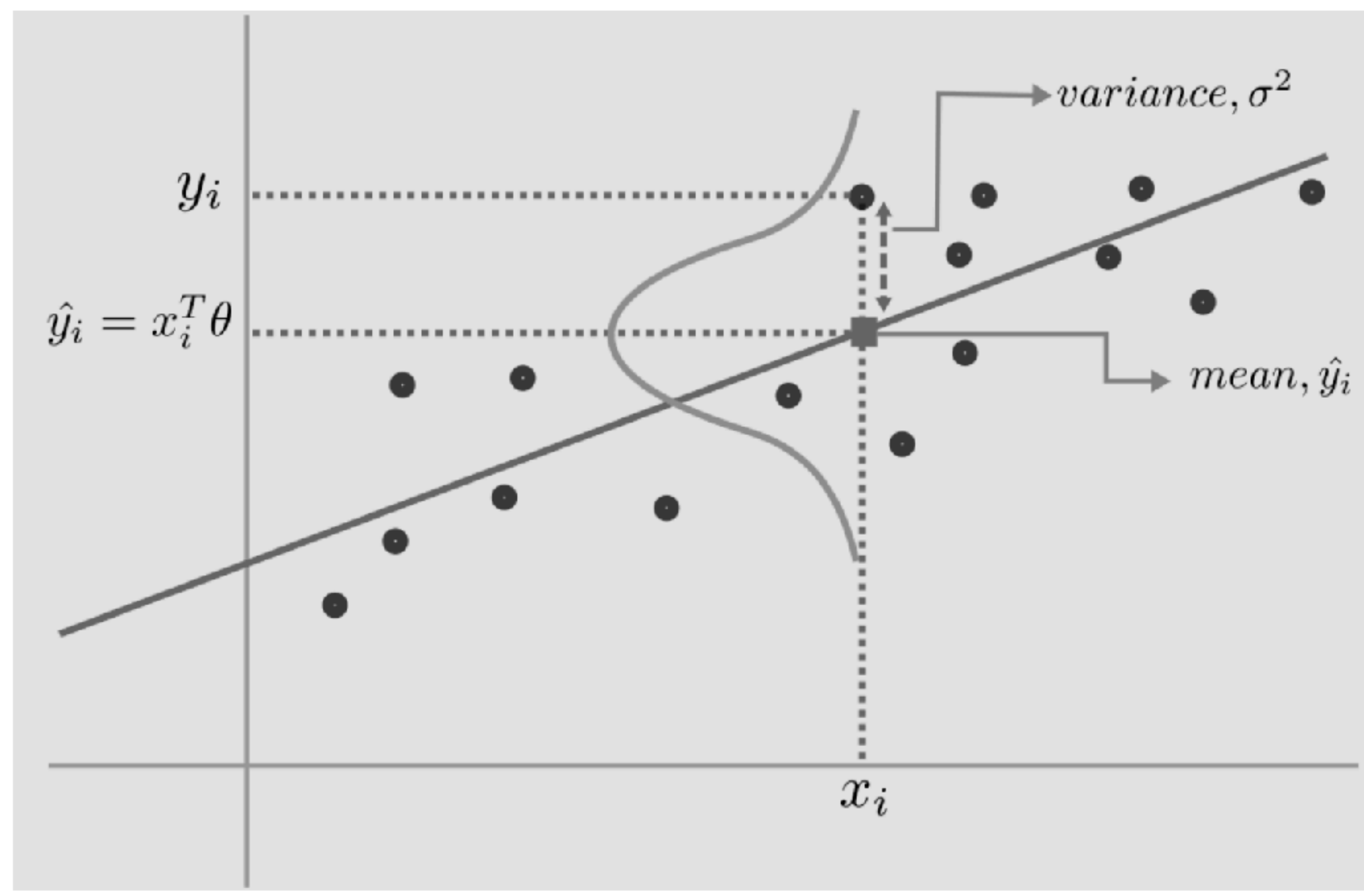
$$f(\boldsymbol{x}) = a\boldsymbol{x} + b$$

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

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

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





https://ethen8181.github.io/machine-learning/model_selection/collinearity.html

Generative model for linear regression in Edward

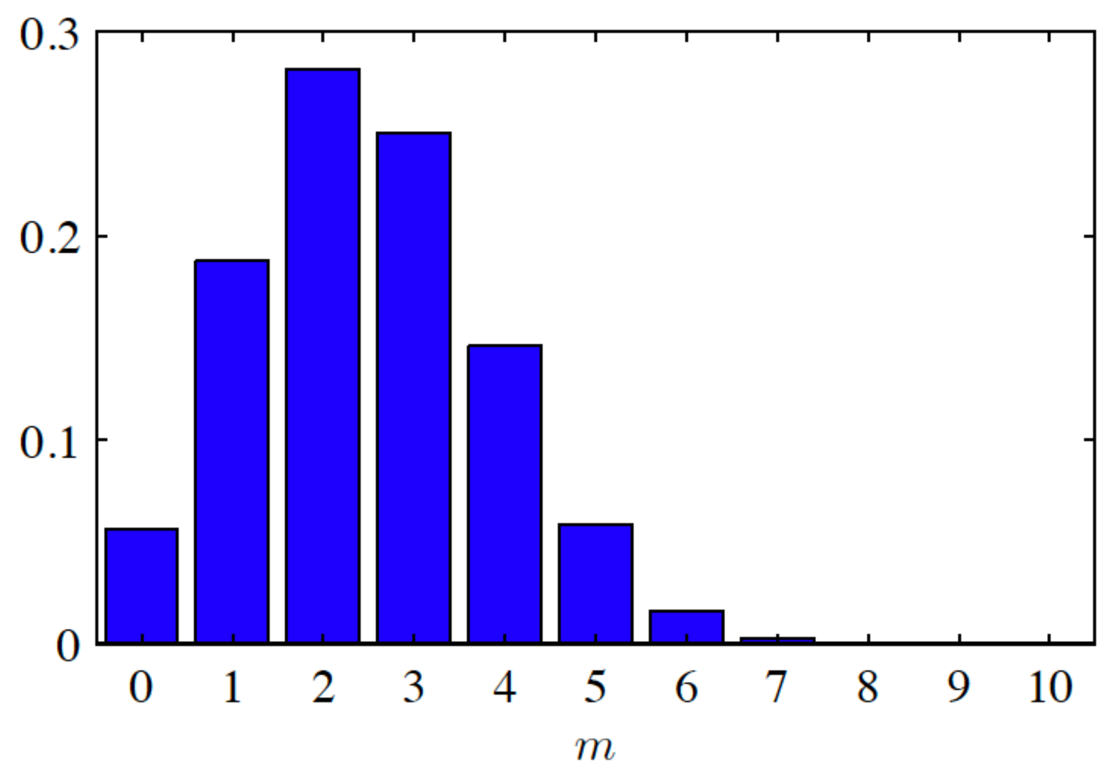
01-Generative-linear-regression-edward.ipynb

Binomial distribution

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

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

$$p_1 + p_2 = 1$$



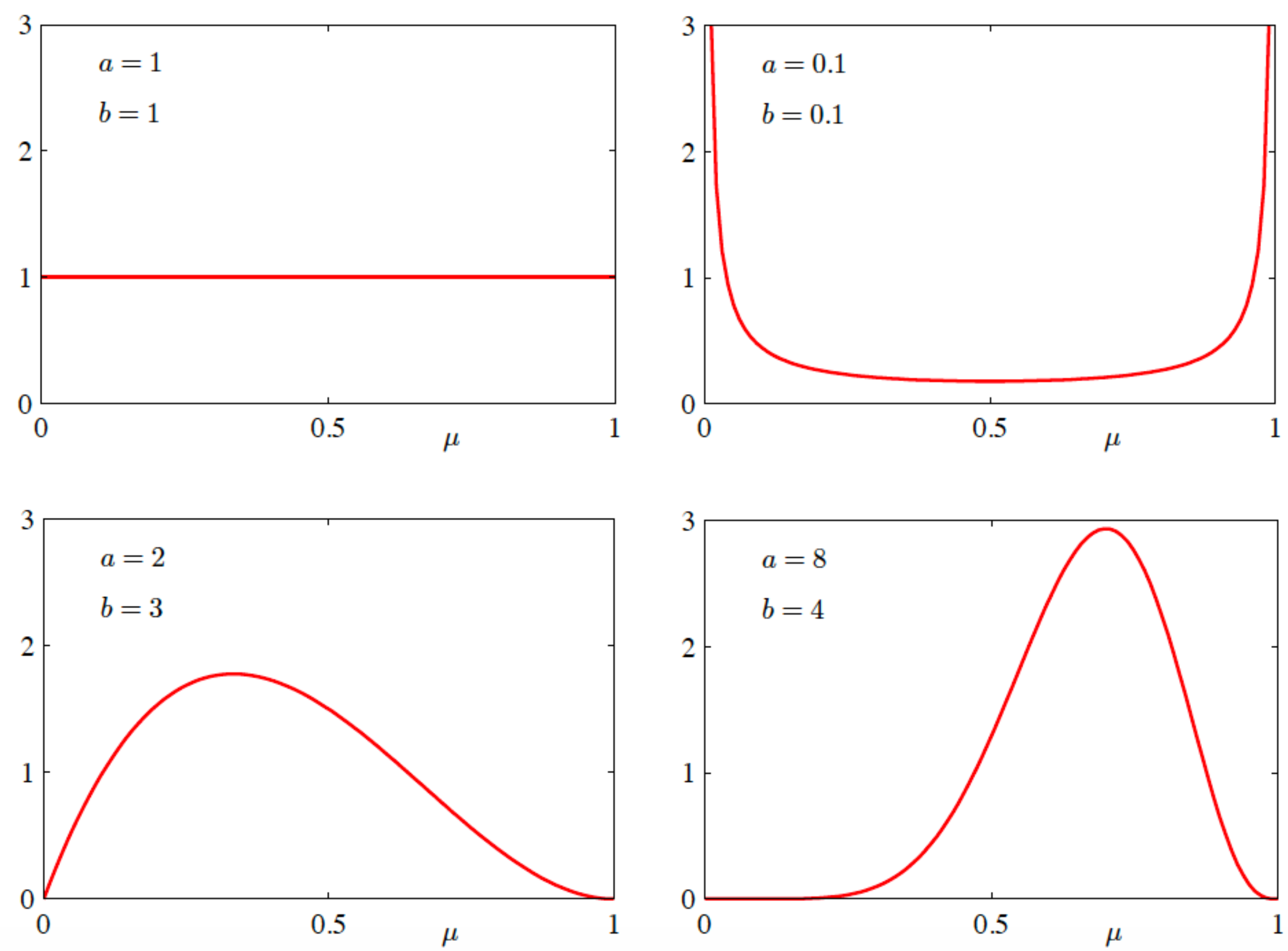
Example: $n = 10$, $p = 0.25$

Beta distribution

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

$$p_1 + p_2 = 1$$

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



Multinomial and Dirichlet distributions

Multinomial

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

Dirichlet

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

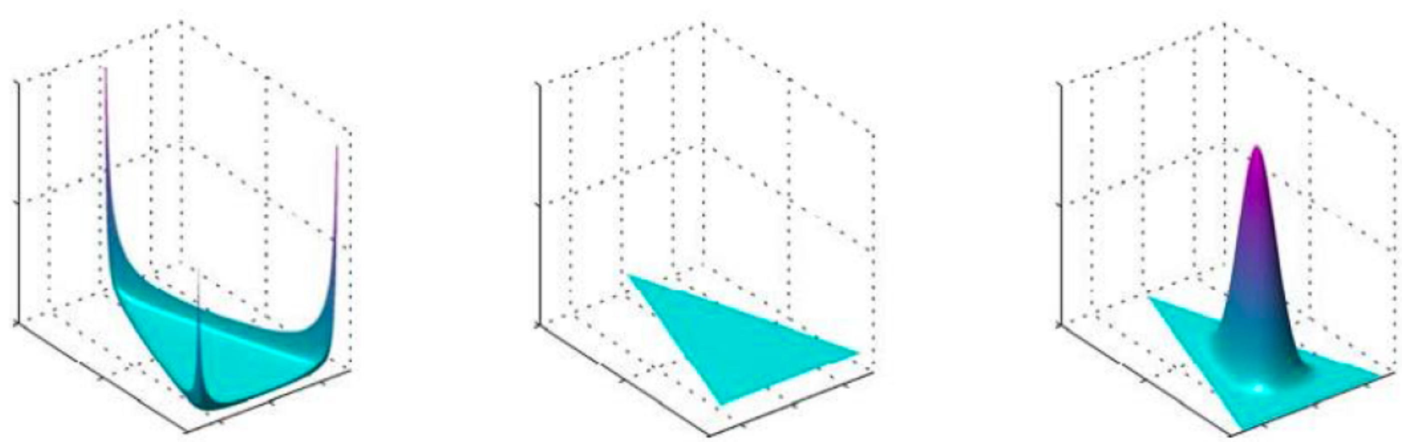
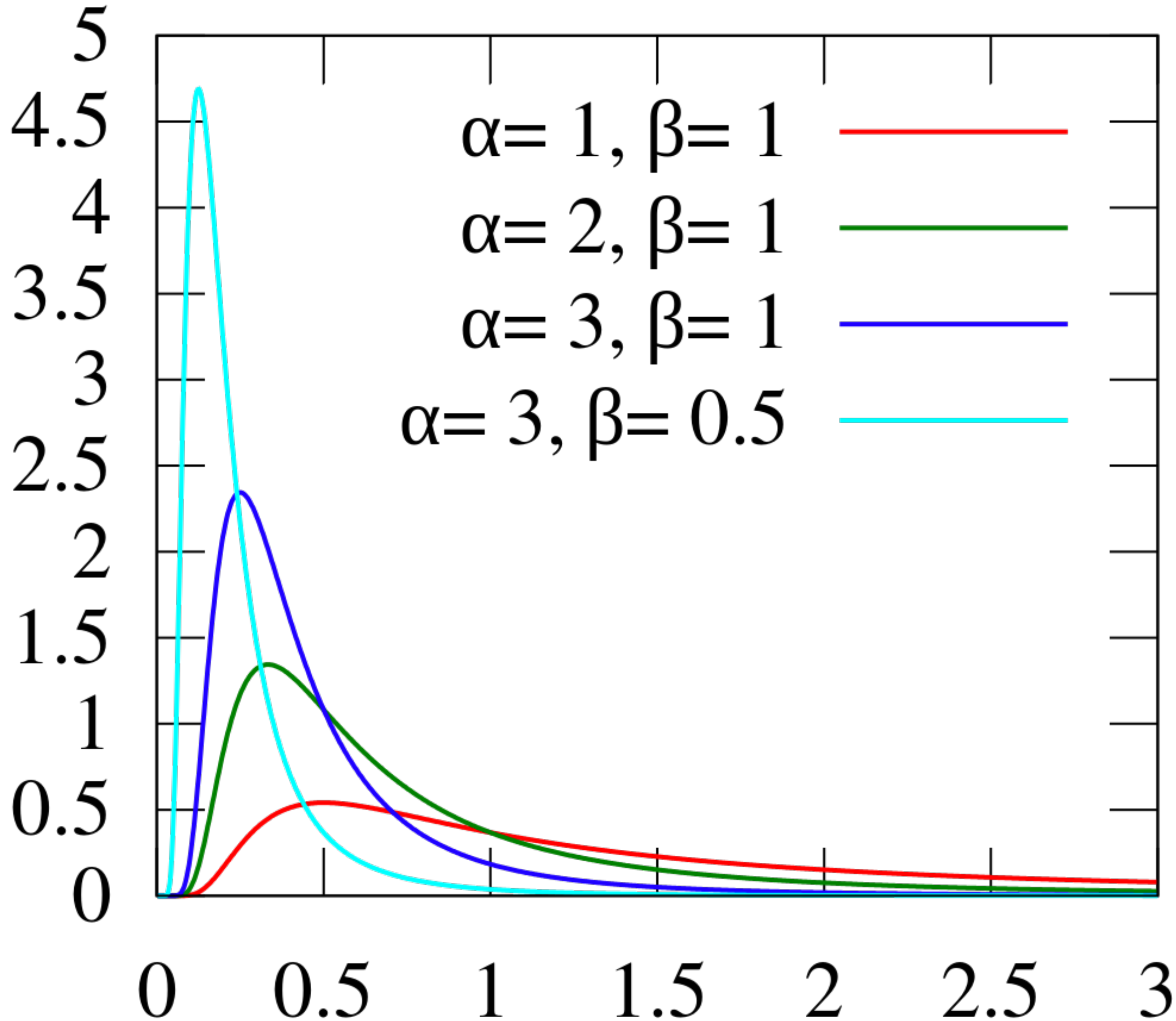
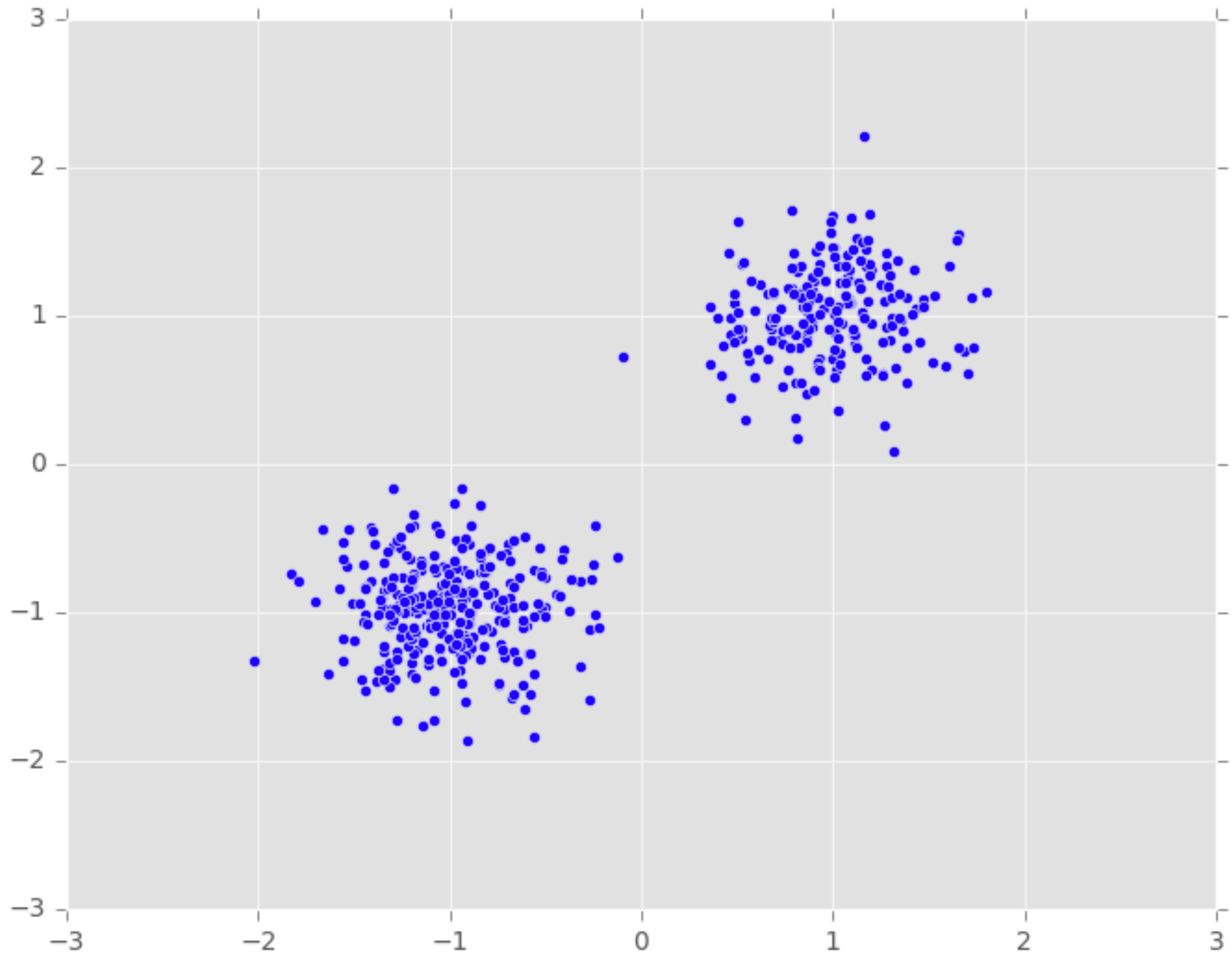


Figure 2.5 Plots of the Dirichlet distribution over three variables, where the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density. Here $\{\alpha_k\} = 0.1$ on the left plot, $\{\alpha_k\} = 1$ in the centre plot, and $\{\alpha_k\}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

Inverse-gamma Distribution



Clustering as gaussian mixtures

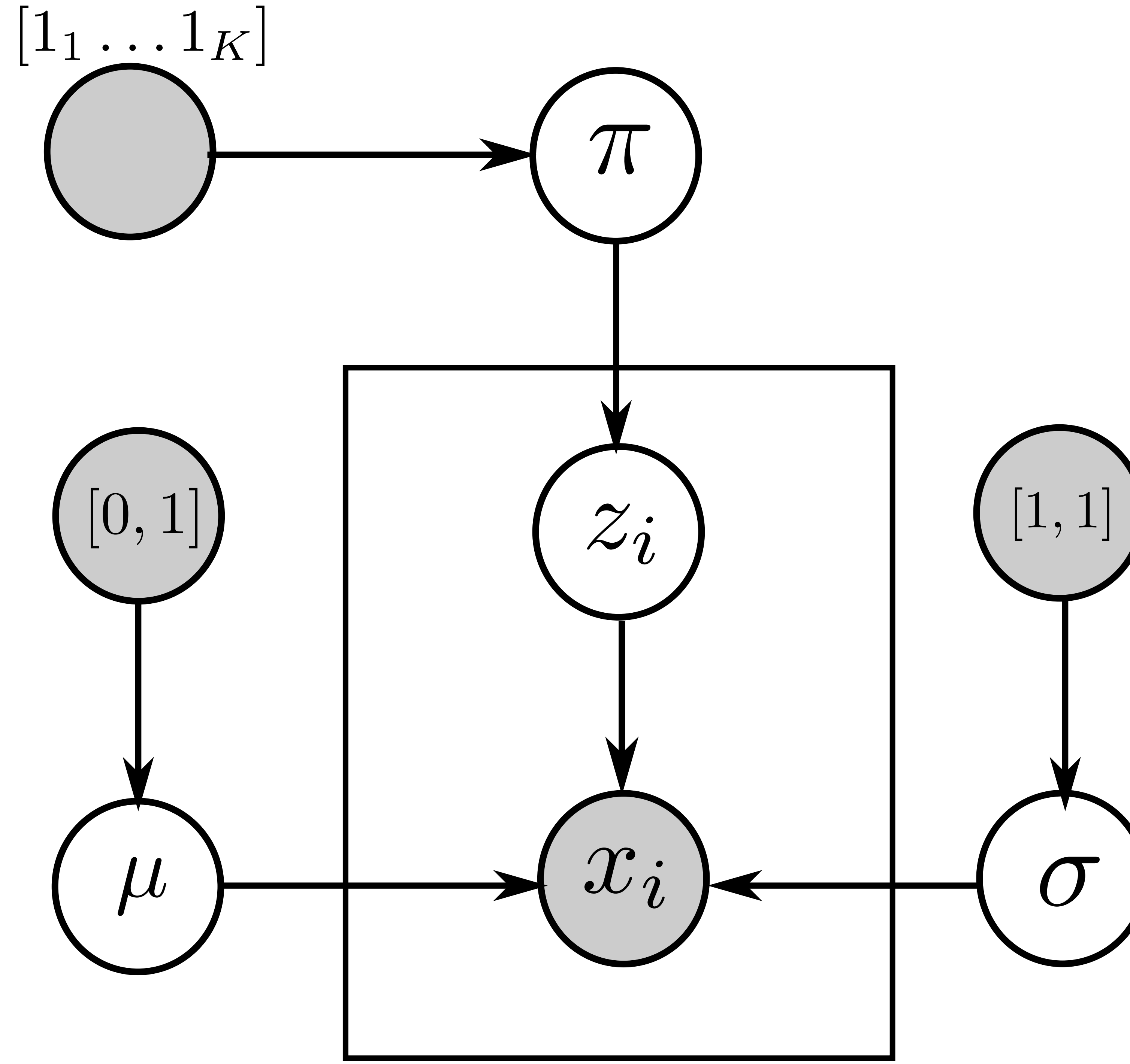


Clustering as gaussian mixtures

$$\pi \sim \text{Dirichlet}(1_1 \ldots 1_K)$$

$$\mu_k \sim \mathcal{N}(0,1)$$

$$\sigma_k^2 \sim \text{InverseGamma}(1,1)$$



Gibbs sampling

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

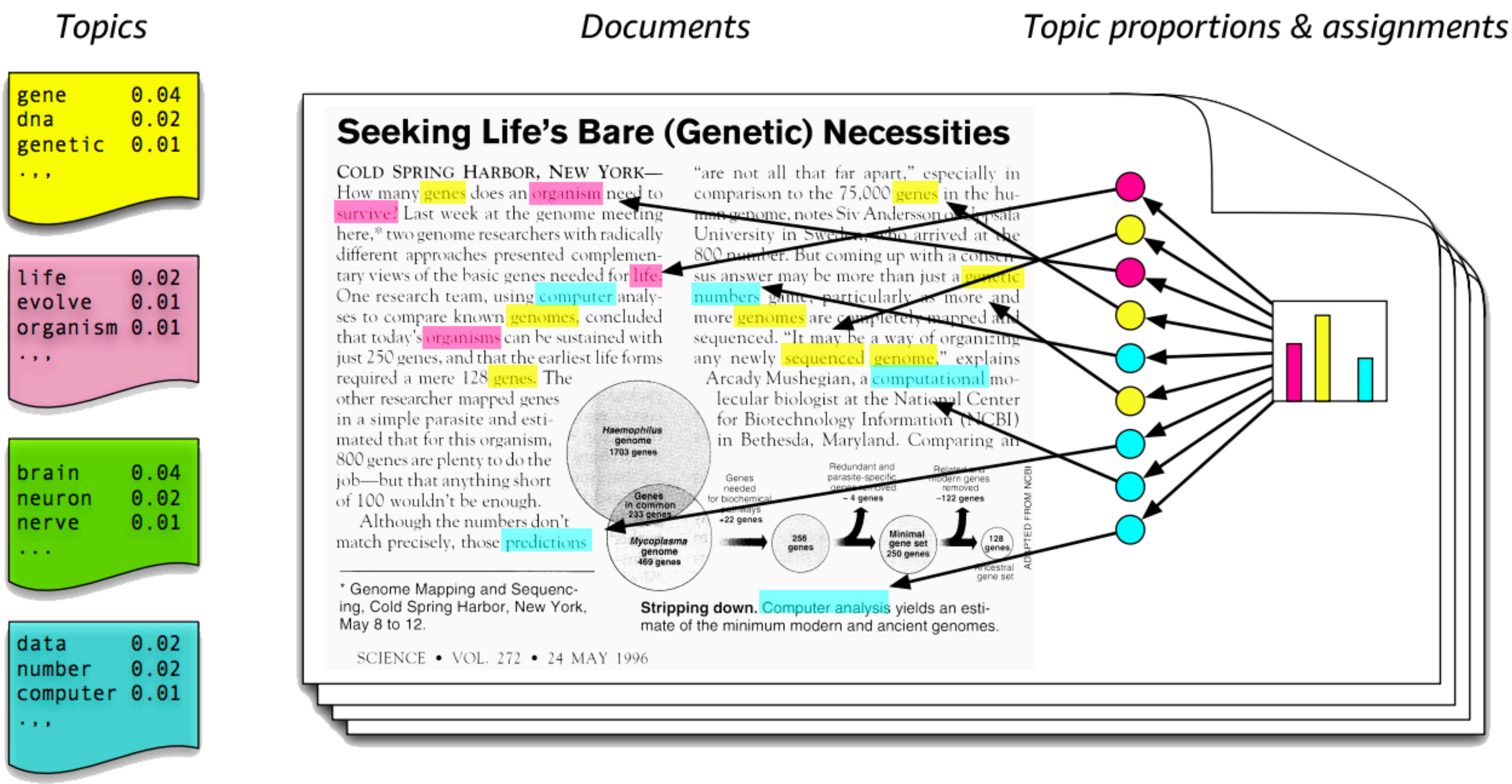
     $$\cdots$$

   - Sample $z_M^{(\tau+1)} \sim \mathrm{P}(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$

Clustering as gaussian mixtures

02-clustering-edward.ipynb

Topic Modeling



https://kldavenport.com/topic-modeling-amazon-reviews/

Latent Semantic Analysis

Topic modeling using LSA example:

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

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

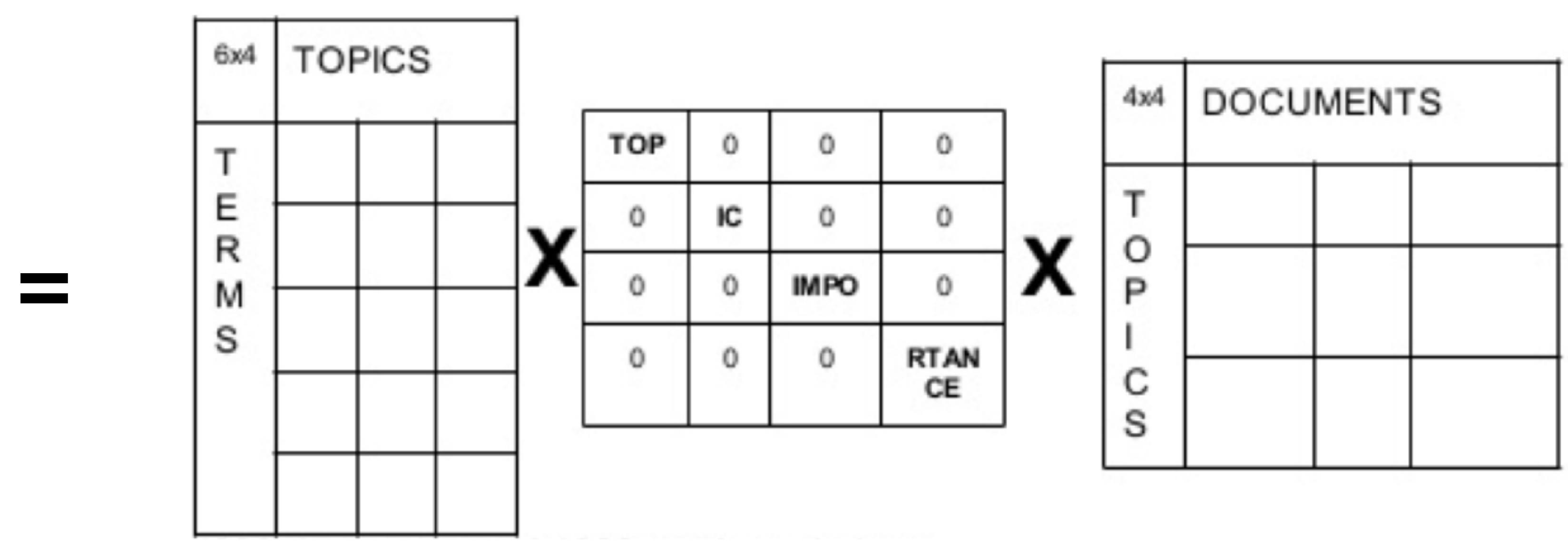
Latent Semantic Analysis

Doc$_1$: *Machine learning helps people to understand data.*

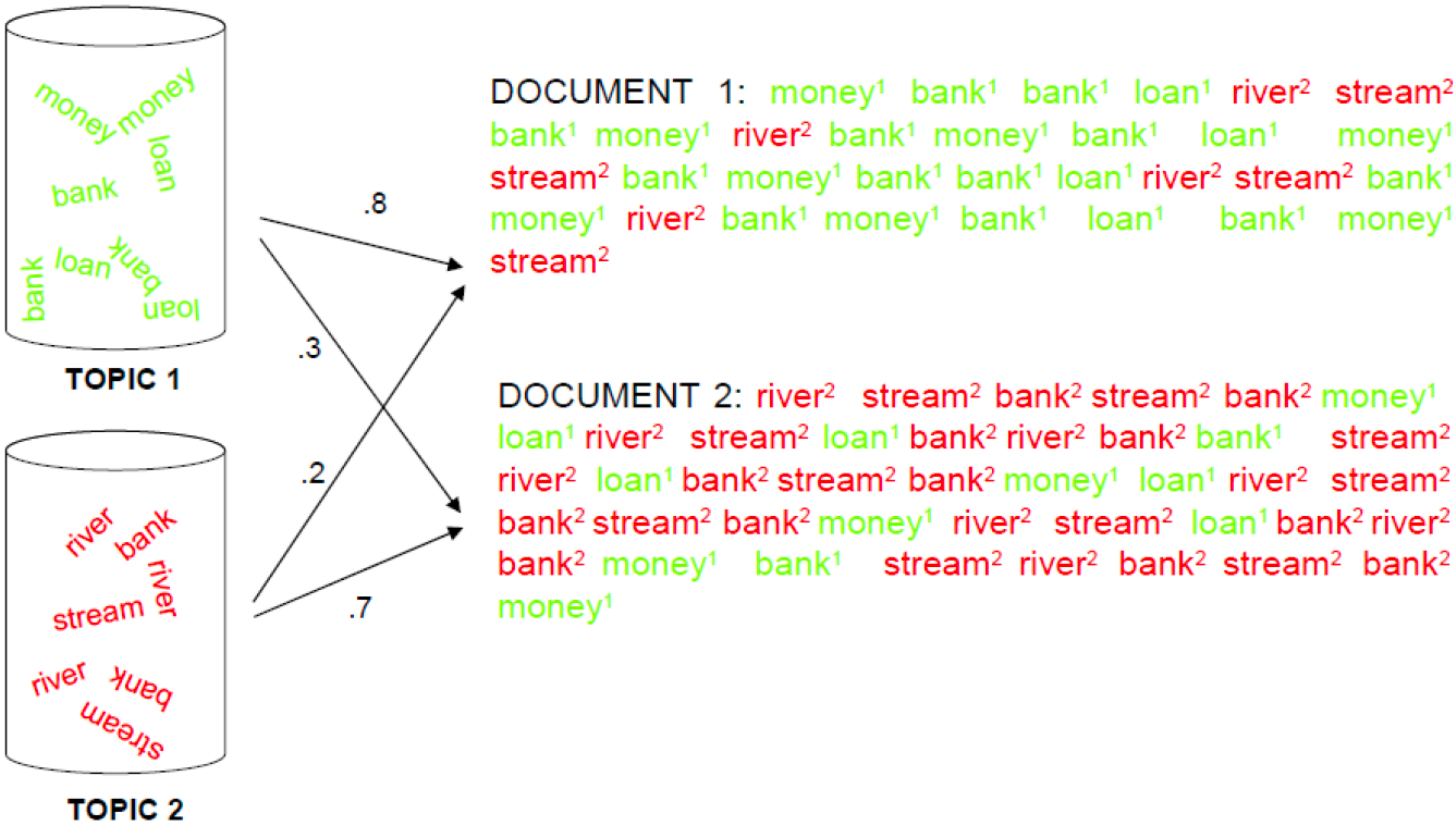
Doc$_2$: *Data can be understood using machine learning.*

Doc$_3$: *People can use machine learning for data understanding.*

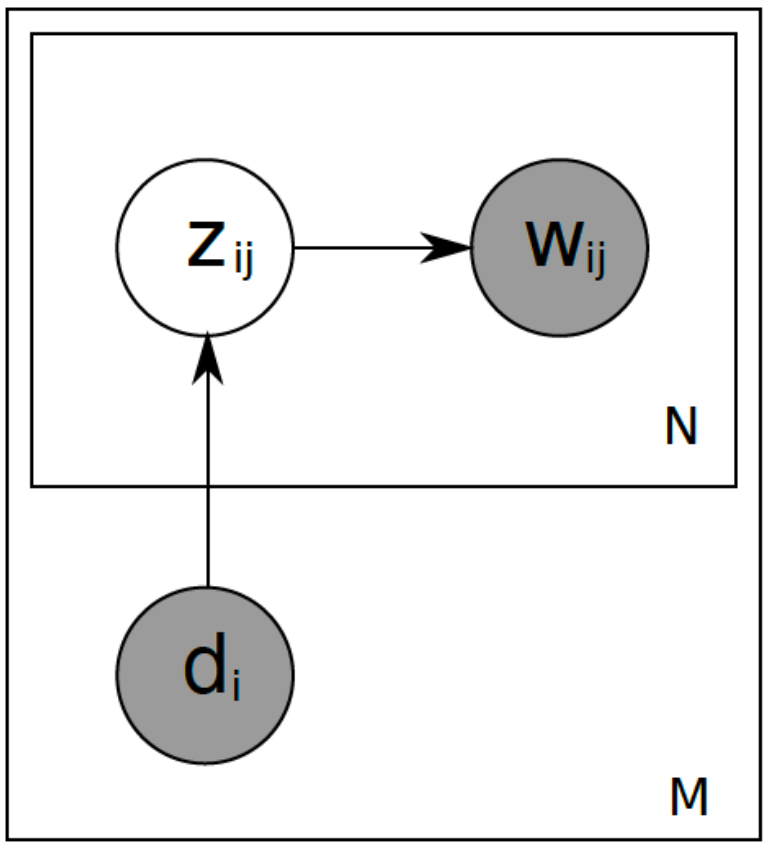
	Doc$_1$	**Doc$_2$**	**Doc$_3$**
be	0	1	0
can	0	1	1
data	1	1	1
for	0	0	1
helps	1	0	0
learning	1	1	1
machine	1	1	1
people	1	0	1
to	1	0	0
understand	1	0	0
understanding	0	0	1
understood	0	1	0
use	0	0	1
using	0	1	0



Probabilistic Latent Semantic Analysis



Model of Probabilistic Latent Semantic Analysis



for $i \in \{1, 2, \ldots, N\}$ **do**
  for $j \in \{1, 2, \ldots, M\}$ **do**
    Choose a latent topic z_{ij} with probability $P(z_{ij}|d_i)$
    Choose a word w_{ij} with probability $P(w_{ij}|z_{ij})$
  end for
end for

Probabilities are computed from frequency analysis of words

Not a generative model (works for training data only)

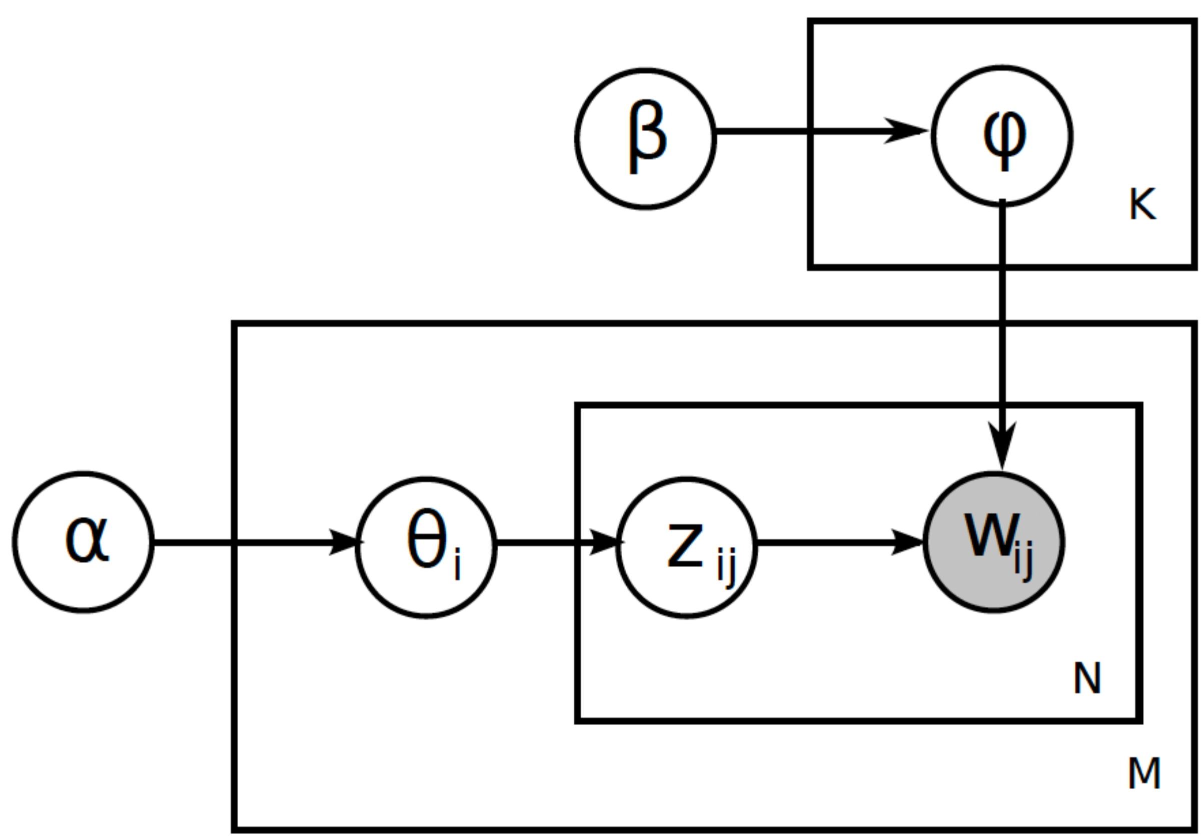
Latent Dirichlet Allocation

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

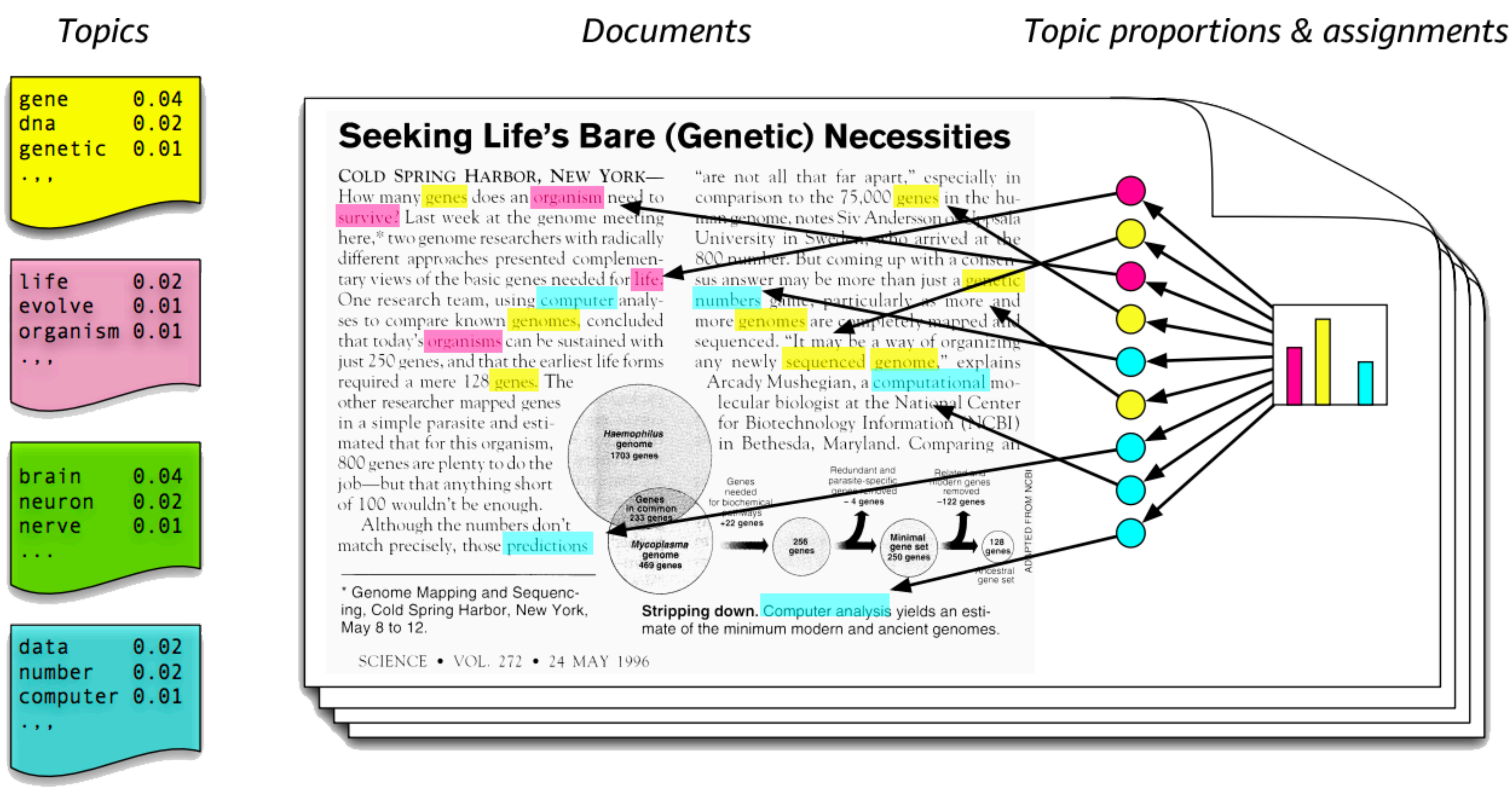
For each word position $j \in \ldots N_i$ choose topic $z_{i,j} \in 1 \ldots K$,

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

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



Latent Dirichlet Allocation



https://kldavenport.com/topic-modeling-amazon-reviews/

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/