主题模型（Topic Models）

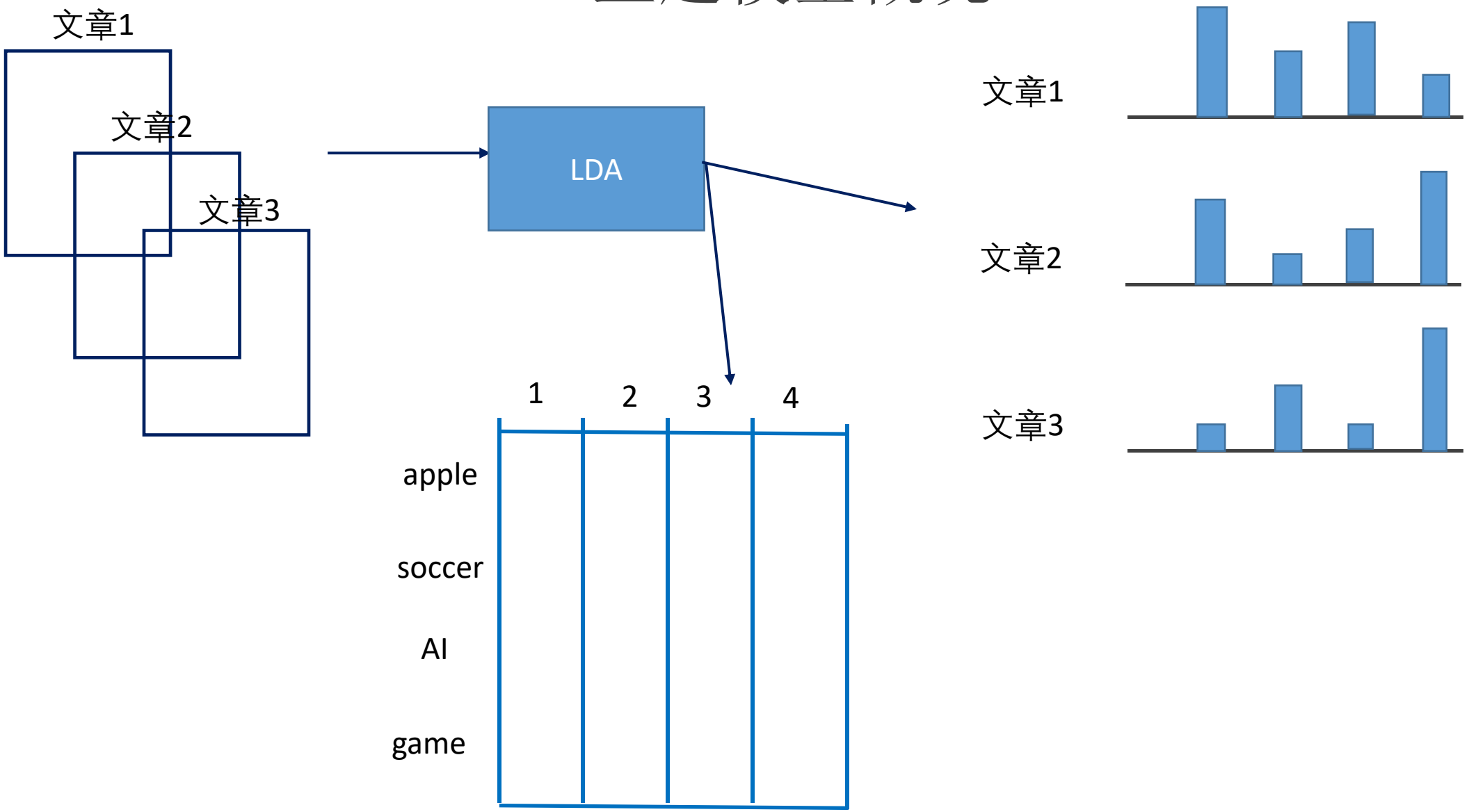
(Partially Completed)

2019.08.07

主题模型

- 主题模型属于贝叶斯模型(Fully Bayesian)
- 主题模型是生成模型，以无监督的方式来学习
- 主题模型属于Mixed Membership模型
- 对于主题模型的推导需要近似算法如MCMC

主题模型概况



主题模型介绍

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

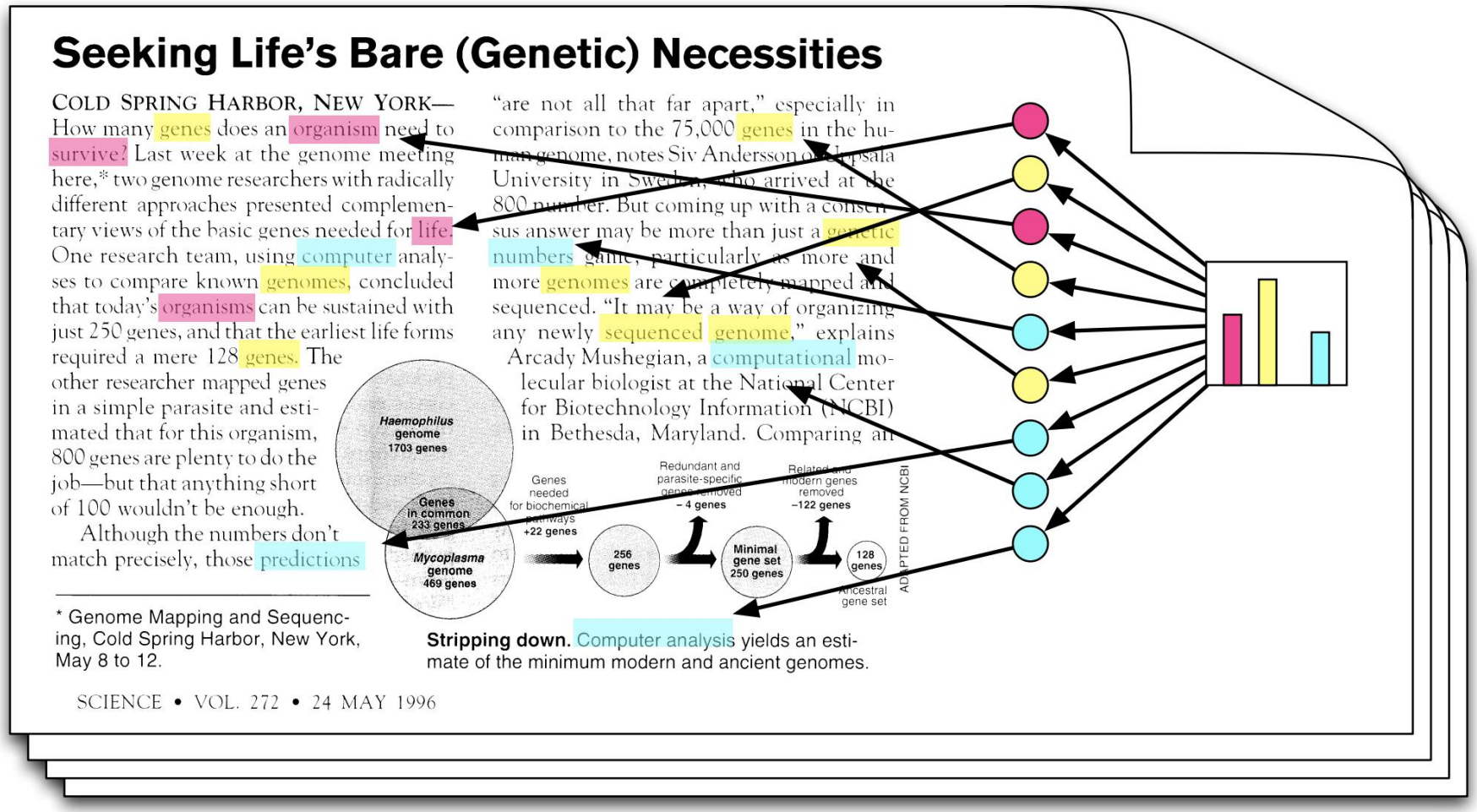
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and 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 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.

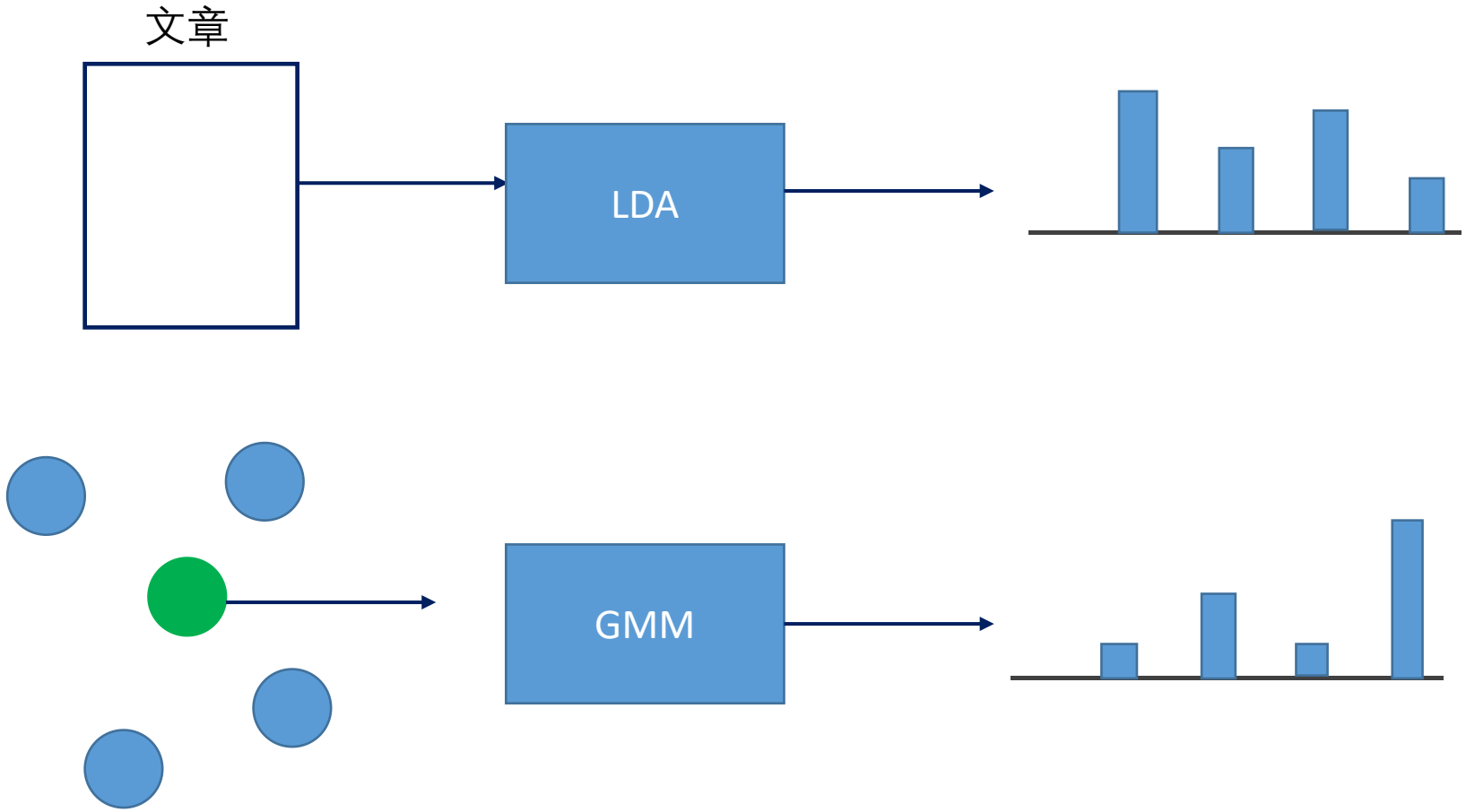
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

Mixed Membership Models



问题： 朴素贝叶斯是否属于这一类？

MLE vs MAP vs Bayesian

MLE MAP Bayesian

贝叶斯模型

贝叶斯模型也可以看作是集成模型，而且一般集成了无穷多的模型。

贝叶斯模型的预测

$$\int p(y'|x',\theta)p(\theta|D)d\theta$$

问题：如何计算$p(\theta|D)$？

$p(\theta|D)$=

蒙特卡洛采样

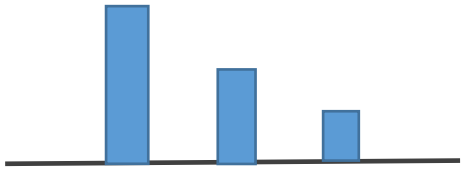
$$\int p(y'|x',\theta)p(\theta|D)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} p(y'|x',\theta^s) \qquad \theta^s \sim p(\theta|D)$$

1. Independent Sampling

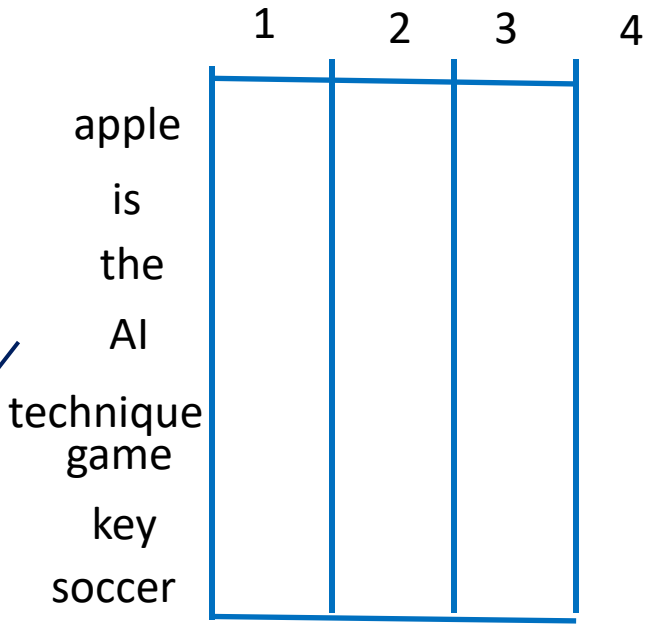
2. Sequential Sampling

从生成的角度来看LDA

步骤1： 指定文章的主题



步骤2： 根据主题采样单词



AI is the key techniques

文章

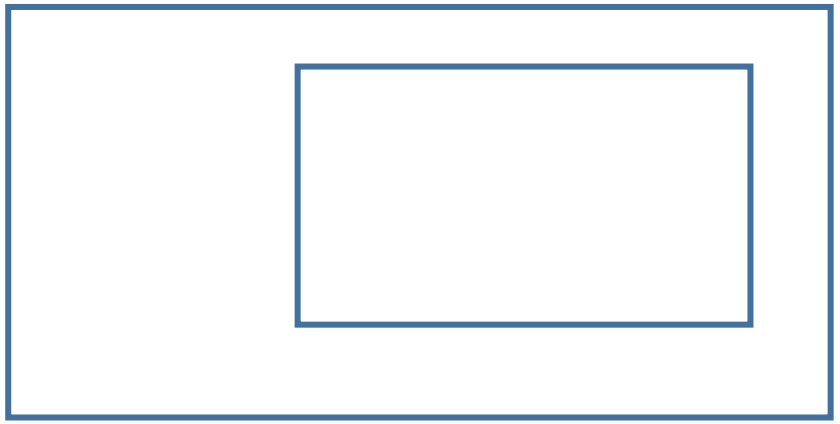
LDA的生成过程

K: 主题个数

N: 文档个数

N_i: 文档i中单词的个数

θ_i: 文档i的主题分布



生成过程:

近似算法

Gibbs Sampler

Metropolish Hasting

Collapsed Gibbs Sampler

Variational Inference

Langevin Dynamics

Importance Sampling

……