概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(Lesson 14)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

Tel: 62796193

Email: ozj@tsinghua.edu.cn

课程章节

- 第一章 引言（1）
- 第二章 图模型的表示理论（3）
  - DGM-UGM
  - Semantics
  - HMM-CRF
- 第三章 图模型的推理理论（6）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：maxlikelihoodEstimate，BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）

课程总结

Summary

- 概率建模，统计推理和学习
  - 机器智能
  - 处理不确定性是智能的一种重要表现
  - 不同领域的许多应用问题可归结为概率建模，统计推理和学习

    信号估计、滤波、跟踪

    模式识别（语音、图像、文本、网页、...）

- 概率图模型理论知识体系
  - 概率模型的表示（representation）
  - 基于概率模型的推理（inference）
  - 概率模型的学习（learning）

图模型已逐渐成为描述和应用概率模型的有力工具，
其核心是概率论和数理统计

Summary

- 概率建模，统计推理和学习
  - 机器智能
  - 处理不确定性是智能的一种重要表现

Uncertainty appears to be an inescapable aspect of most real-world applications. Koller & Friedman, p.2

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. —— Albert Einstein, 1956.

就认为数学的法则是源于现实而言，它们就不是确定性的；就它们是确定性的而言，它们就不是源于现实的。—— 爱因斯坦, 1956.

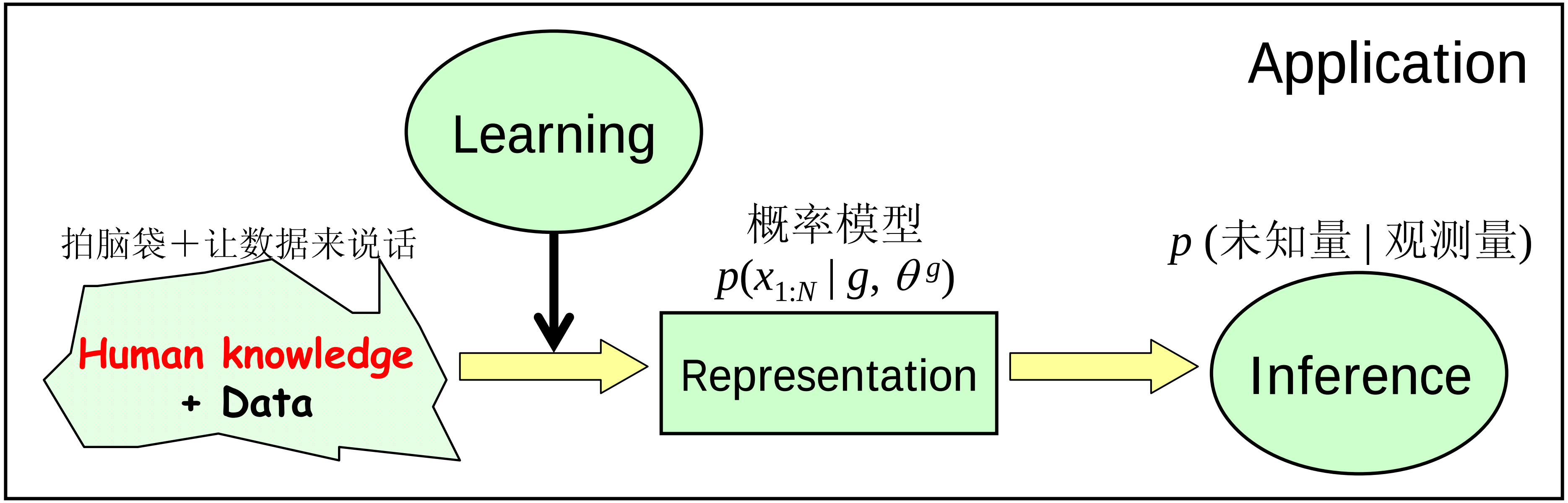
  - 概率模型的学习（learning）

图模型已逐渐成为描述和应用概率模型的有力工具，其核心是概率论和数理统计

Summary

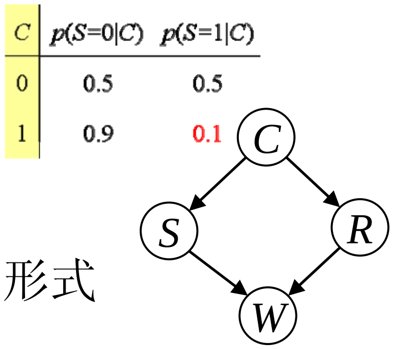
- 概率模型/联合分布表示成一个图（画图、读图）
- 学习：从样本归纳学习出模型
- 推理：利用模型进行演绎，推断出新样本的行为

Bayesian models of cognition
Bayesian mind



图模型建模的四要素

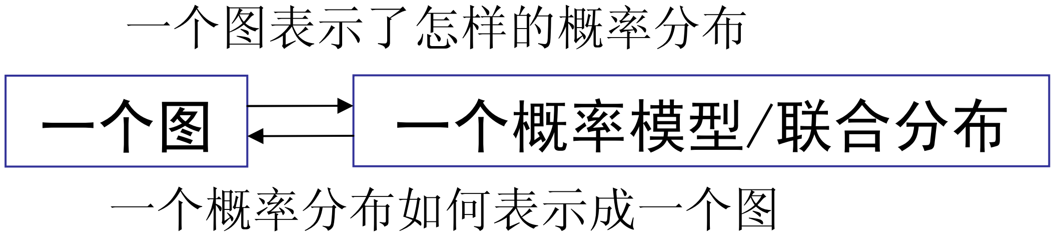
- 当使用图模型去表示一个具体的概率分布时（ to describe a particular distribution），需逐步明确以下四要素
- 语义（Semantics）
  - 定义了图和概率分布如何发生联系（有向图、无向图、因子图、链图、...）
- 结构（Structure）
  - 定义随机变量及之间联系（即明确图包含哪些结点以及边）
- 实现（Implementation）
  - 指定结点类型（ discrete/continuous ）以及局部函数的具体形式
- 参数（Parameter）
  - 凭经验指定或利用数据进行估计——局部函数的待定参数的取值



C	$p(S=0\|C)$	$p(S=1\|C)$
0	0.5	0.5
1	0.9	0.1

Representation (3课时)

- 图的语义（Semantics）



一个图表示了怎样的概率分布

一个图

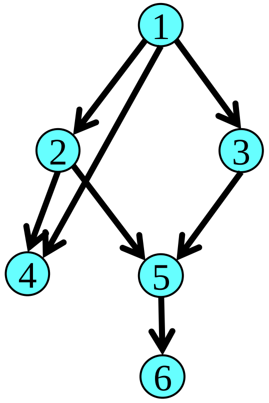
一个概率模型/联合分布

一个概率分布如何表示成一个图

性质诱导分布：

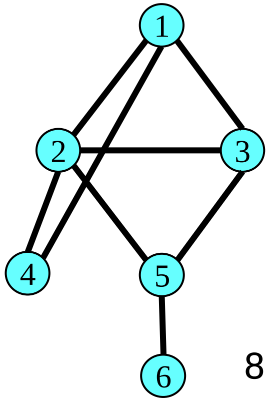
$$(DF) \Leftrightarrow (DG) \Leftrightarrow (DL) \Leftrightarrow (DO)$$

$$p(x_V) = \prod_{n=1}^{N} p(x_n \mid pa_n)$$

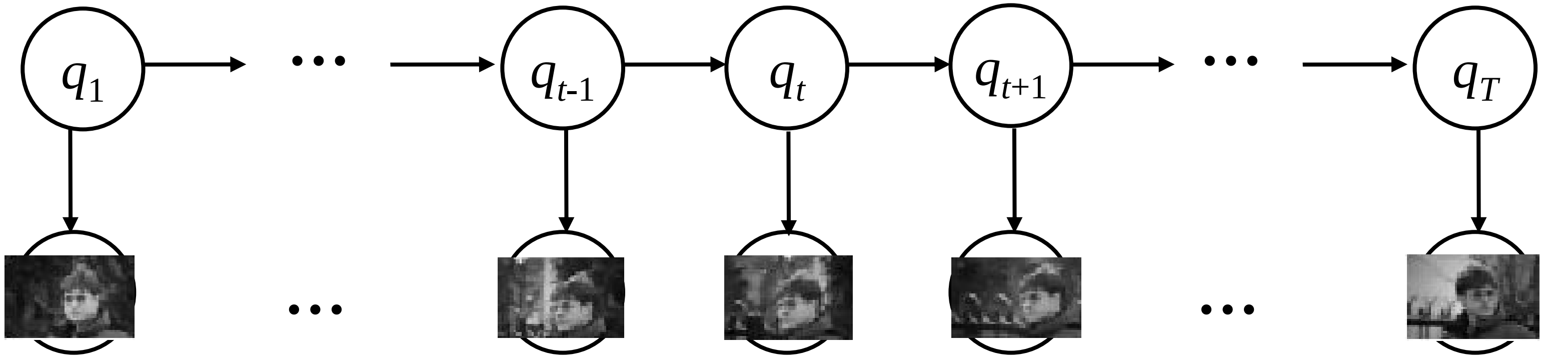


$$(F) \Rightarrow (G) \Rightarrow (L) \Rightarrow (P)$$

$$p(x_V) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$$



- 第三章 图模型的推理理论（6课时）
  - 精确推理：**variable-elimination，cluster-tree，triangulate**
  - 连续变量：**Kalman**
  - 采样近似：**sampling**
  - 变分近似：**variational**

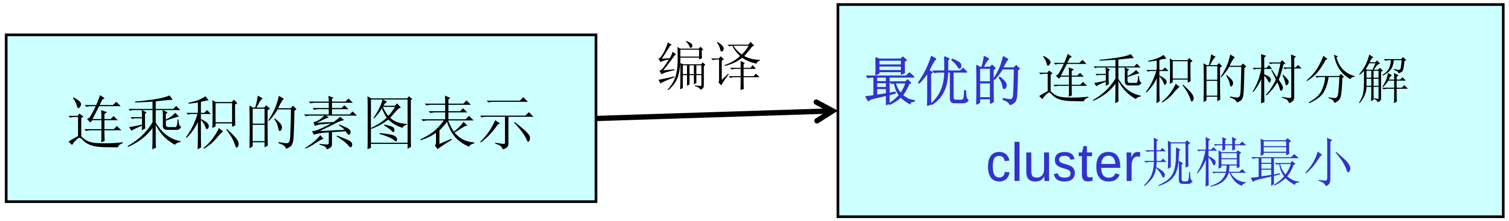


$$p\left(x_Q \mid x_E\right) \propto p\left(x_Q, x_E\right) = \sum_{x_{V\setminus(Q,E)}} p\left(x_Q, x_E, x_{V\setminus(Q,E)}\right)$$

精确推理：连乘积求和

精确推理 编译 与 计算 分离

连乘积 $\prod_i f_i$，如 $\phi(a,b,c)\phi(b,c,f)\phi(a,b,d)\phi(f,g)\delta(g=1)$

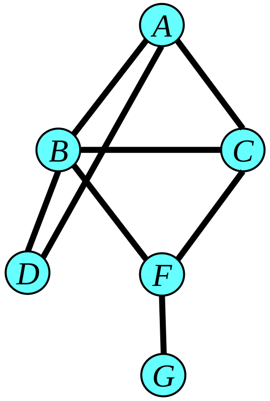


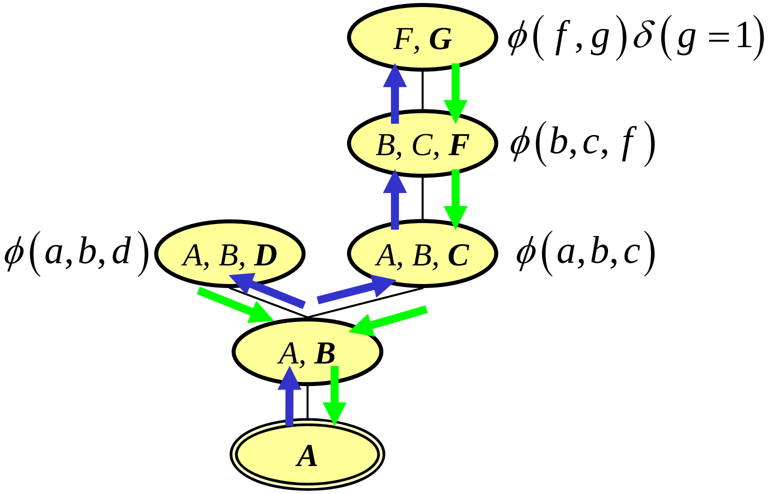
连乘积的素图表示

编译

最优的 连乘积的树分解

cluster规模最小





$$\lambda_{C_1\to C_2} = \Downarrow_{sep(C_1,C_2)} \left\{ C_1\text{里的函数}\cdot \prod_{Z\in C_1\text{ 的邻居}\backslash C_2} \lambda_{Z\to C_1} \right\}$$

$$\left(\prod_i f_i\right)\Downarrow_{\chi(C)} \propto C\text{里的函数}\cdot \prod_{Z\in C\text{的邻居}} \lambda_{Z\to C}$$

含有连续变量的图模型推理

树消除算法 (Cluster-tree elimination) 依然成立

算法本身并没有对联合分布中的变量是离散还是连续做限制

Linear-Gaussian

KalmanFilter

$$\phi(x\,|\,g,h,K) = \exp\left\{g + x^T h - \frac{1}{2}x^T K x\right\}$$

典范函数的运算法则

Linear-Gaussian CPD

Entering evidence

Product

Marginalization

使函数操作变得容易

Conditional Gaussian (CG)

Mixed discrete-Gaussian

典范表示：(g_i, h_i, K_i)

Exact inference is feasible, when

先消除连续变量，

再消除离散变量

Weak marginalization

Moment matching

Assumed density filtering

Non-linear non-Gaussian

Exact inference is usually not feasible !

Two classes of approximation:

Stochastic

Deterministic

Learning (3课时)

当使用图模型去表示一个具体的概率分布时（ to describe a particular distribution），需逐步明确五要素

语义？ 引入哪些变量？ 变量间关系？ 关系的表征函数形式？ 函数的表征参数？

- 基于独立同分布样本集 $D=(x[1],\ldots,x[M])$，建立具体的概率模型 $p(x\mid g,\ \theta^{g})$
  - 估计出 g ：结构学习
  - 固定 g 估计出 θ^{g} ：参数学习

	Known structure		Unknown structure
Complete data	ML	Bayesian	Learning tree Score-and-search for BN
Incomplete data	ML	Bayesian	Structural EM

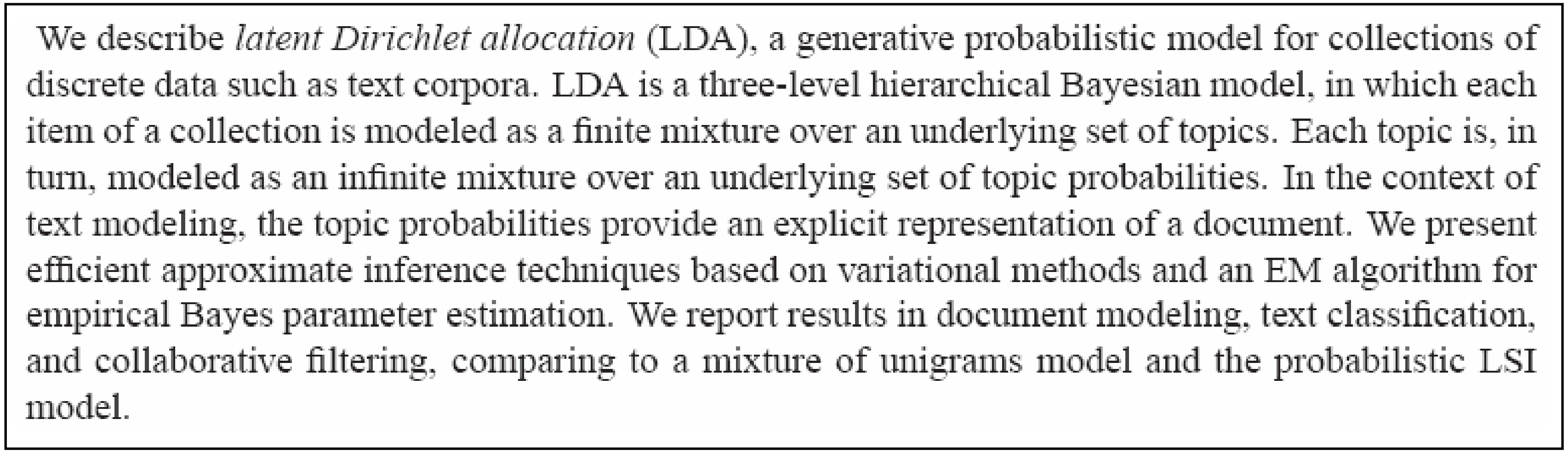
一个综合例子：文档建模

LDA (Latent Dirichlet Allocation)

for text classification, thematic structure discovery, ...

一些概念和符号

- 词(word)：取值在 1:V 的离散随机变量
- 文档(document)：词变量序列 $d = w_{1:N_d}$
  - Science杂志中某篇文章
- 文档集(corpus)：M个文档的集合 $D= d_{1:M}$
  - Science杂志集

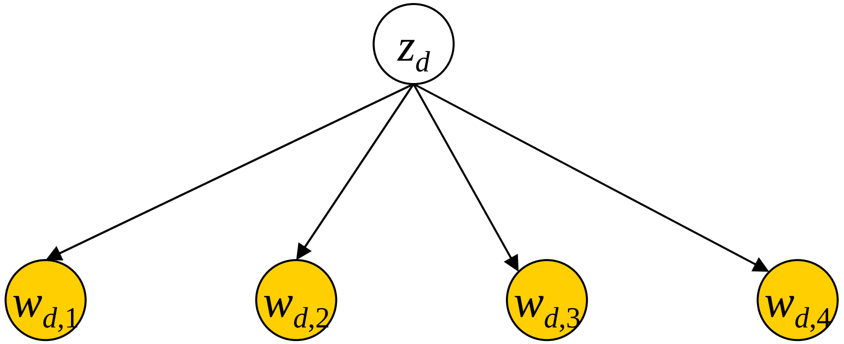


We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

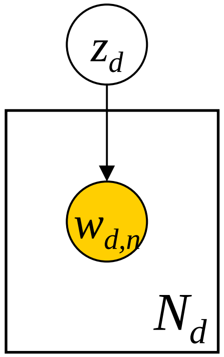
Document modeling

英语词汇出现的直方图

- 为文档建立概率模型：$p(w_{d,1:N_d}) \overset{unigram}{=} \prod_{n=1}^{N_d} p\left(w_{d,n}\right)$
  - e.g. 机器翻译：$\max_{C} p\left(中文文档C|英文文档W\right) \propto p\left(W\right)p\left(C|W\right)$



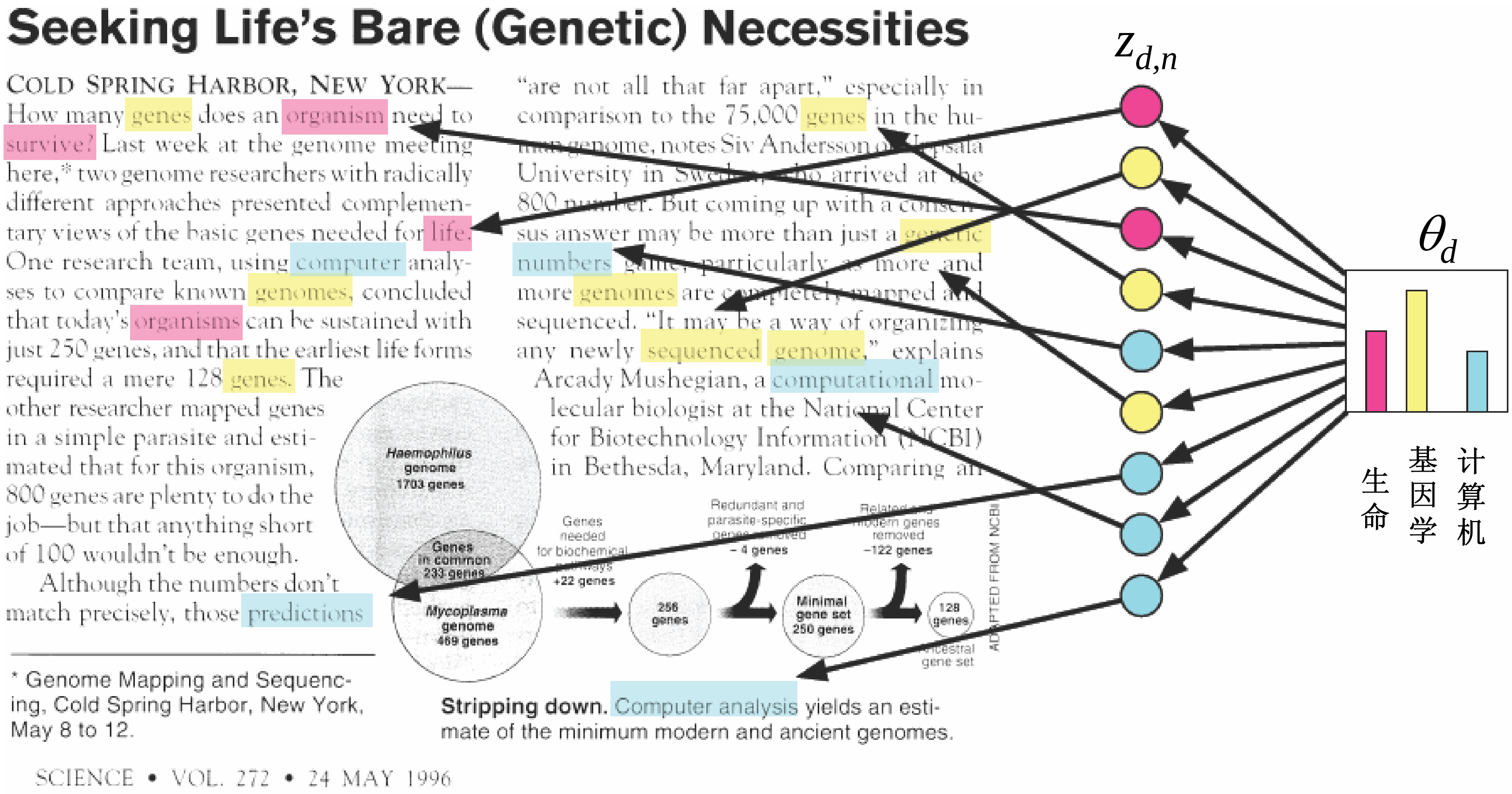
Mixture of Unigram Model



- 一个文档 d 背后有一个潜在的topic，抽象成一个随机变量 $z_d \in \{1:K\}$

第k个topic下英语词汇出现的直方图

$$p\left(w_{d,1:N_d}\right) = \sum_{z_d} p\left(z_d, w_{d,1:N_d}\right) = \sum_{k} p\left(z_d\right) \prod_{n=1}^{N_d} p\left(w_{d,n} \mid z_d = k\right)$$



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI
$z_{d,n}$
θ_d
生命
基因学
计算机

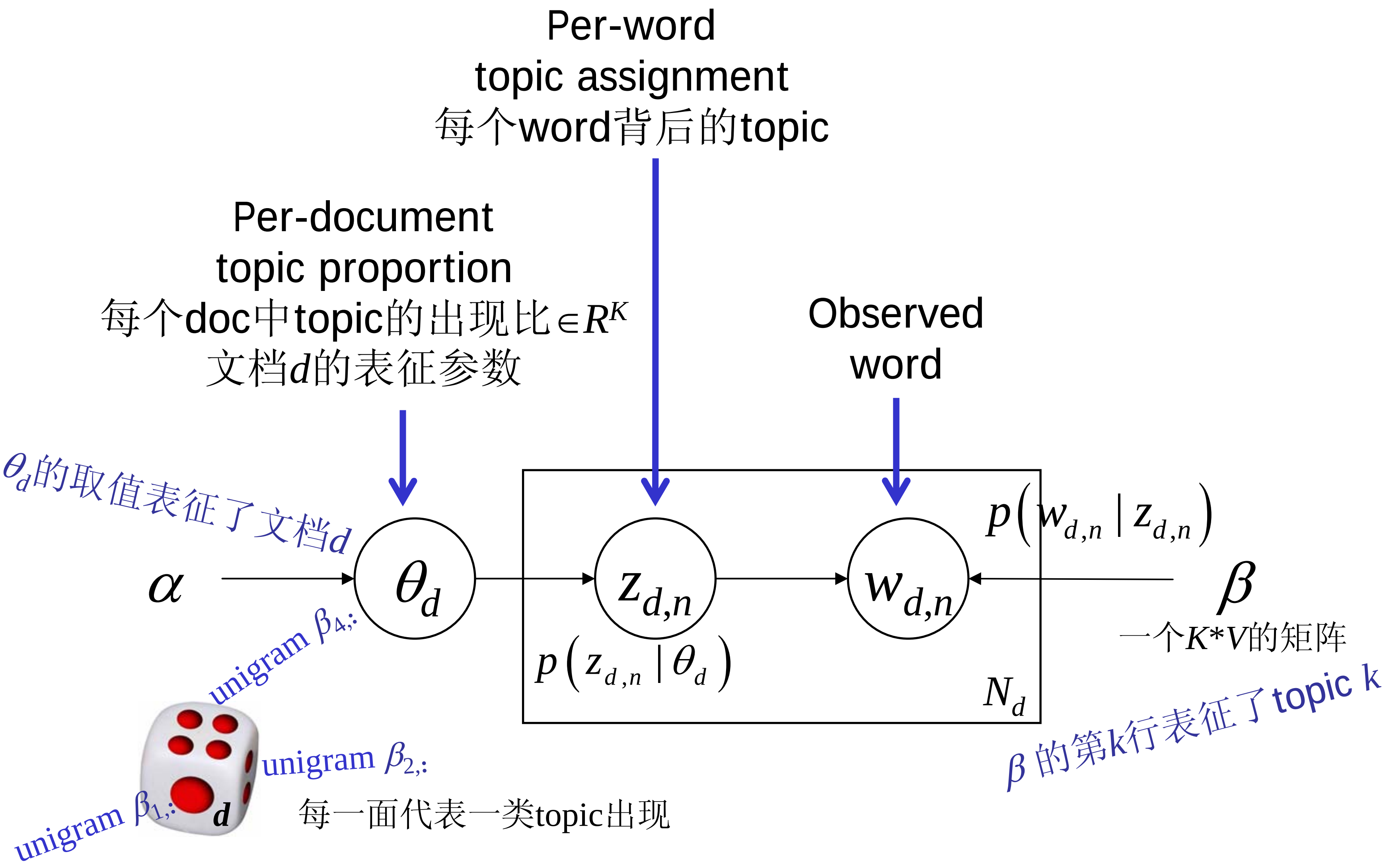
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- Intuition: 每个文档 d 呈现出多个topic，每个文档是 K 个 topic的一个混合
- 整个文档集有 K 类topic。
- 每个word $w_{d,n}$ 背后有一个潜在的topic，抽象成一个随机变量 $z_{d,n}\in\{1:K\}$
- 对于不同文档， topic的出现权重不同

文档的LDA模型 (Latent Dirichlet Allocation)



Per-word
topic assignment
每个word背后的topic

Per-document
topic proportion
每个doc中topic的出现比$\in R^K$
文档d的表征参数

Observed
word

θ_d的取值表征了文档d

α

θ_d

$z_{d,n}$

$w_{d,n}$

β

一个K*V的矩阵

$p\left(w_{d,n} \mid z_{d,n}\right)$

$p\left(z_{d,n} \mid \theta_d\right)$

N_d

β 的第k行表征了topic k

unigram $\beta_{4,:}$

unigram $\beta_{2,:}$

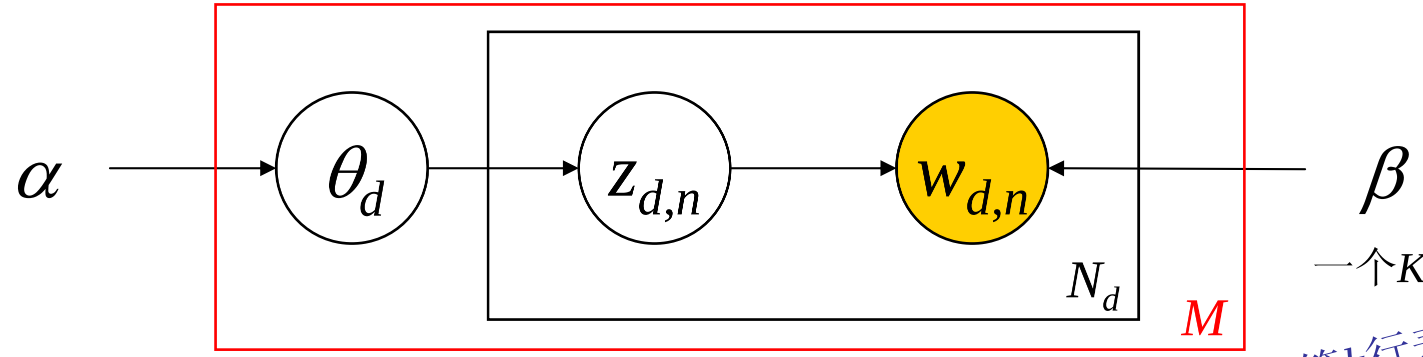
unigram $\beta_{1,:}$

d

每一面代表一类topic出现

LDA – Parameter estimation

- 总体分布 $p\left(\theta_d, z_{d,1:N}, w_{d,1:N} \mid \alpha, \beta\right)$
- 给定观测数据：M 个文档 $D=\{w_{d,1:N_d}\}_{d=1:M}$



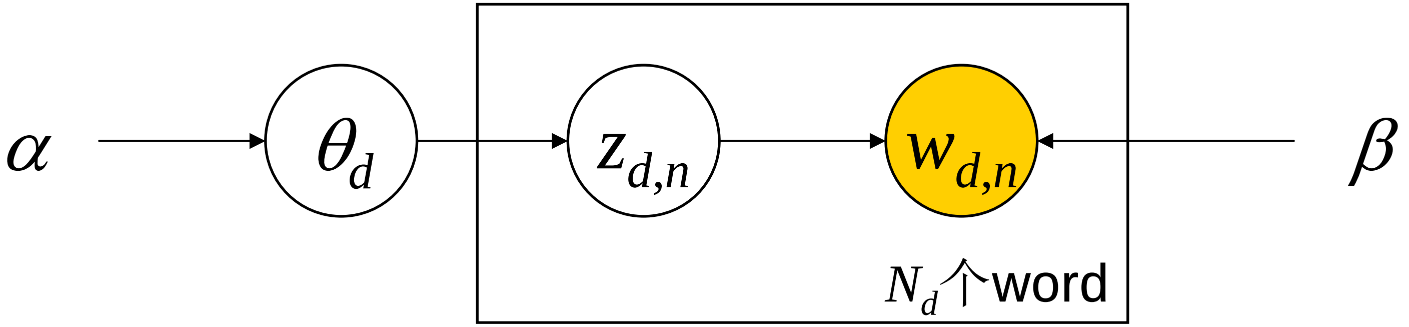
一个K*V的矩阵

β 的第k行表征了topic k

- Variational EM
  - 直接最大化 $log\ p(D|\alpha, \beta)$ 困难
  - 最大化一个辅助函数
  - E-step: 变分推理求出 老参数α, β下 后验分布 $p(\theta_d, z_{d,1:N} \mid w_{d,1:N}, \alpha, \beta)$, 即求出 $q(\theta_d \mid \gamma)$, $q(z_{d,1} \mid \phi_1)$,…, $q(z_{d,N} \mid \phi_N)$
  - M-step: 最大化 完备似然函数的条件期望，得到新的α, β

LDA – Inference

文档的LDA模型



$$p\left(\theta_d, z_{d,1:N_d}, w_{d,1:N_d} \mid \alpha, \beta\right) = p\left(\theta_d \mid \alpha\right) \prod_{n=1}^{N_d} p\left(z_{d,n} \mid \theta_d\right) p\left(w_{d,n} \mid z_{d,n}, \beta\right)$$

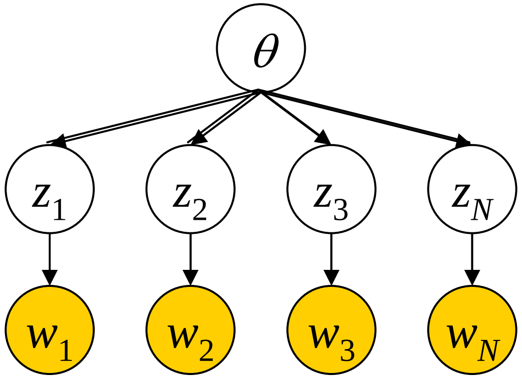
$$p\left(\theta, z_{1:N}, w_{1:N} \mid \alpha, \beta\right) = p\left(\theta \mid \alpha\right) \prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \beta\right)$$

- The inference problem in LDA is to compute the posterior of the hidden variables given a document and parameters α and β.

  That is, compute $p(\theta, z_{1:N} \mid w_{1:N}, \alpha, \beta)$, $p(\theta \mid w_{1:N}, \alpha, \beta)$, $p(z_n \mid w_{1:N}, \alpha, \beta)$

- Unfortunately, exact inference is intractable, so …

LDA – cluster-tree elimination

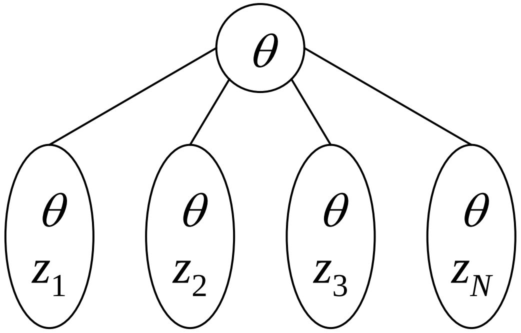


$$p(\theta, z_{1:N}, w_{1:N}) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta)$$

$$p(\theta \mid w_{1:N}) = ?$$

$$p(\theta \mid w_{1:N}) \propto p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z_n = i} p(\overset{=i}{z_n} \mid \theta)\, p(w_n \mid \overset{=i}{z_n}, \beta)$$

$p(\theta \mid \alpha)$

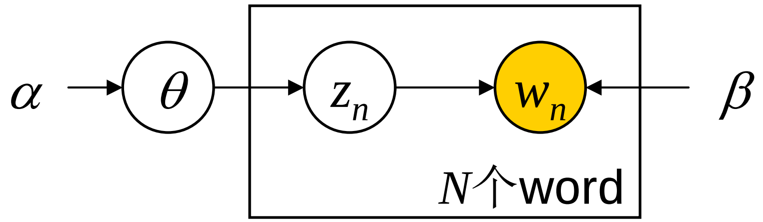


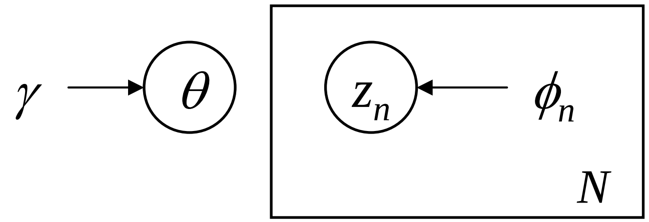
$p(z_n \mid \theta)\, p(w_n \mid z_n, \beta)$

$$\propto \left(\prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{K} \theta_i \beta_{i, w_n} \right)$$

$$\propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \left\{ \begin{array}{l} \left(\theta_1 \beta_{1, w_{n=1}} + \theta_2 \beta_{2, w_{n=1}} \right) \\ \left(\theta_1 \beta_{1, w_{n=2}} + \theta_2 \beta_{2, w_{n=2}} \right) \\ \left(\theta_1 \beta_{1, w_{n=3}} + \theta_2 \beta_{2, w_{n=3}} \right) \\ \left(\theta_1 \beta_{1, w_{n=4}} + \theta_2 \beta_{2, w_{n=4}} \right) \end{array} \right\}$$

LDA – Variational inference I





- 原概率联合分布 $p(\theta, z_{1:N}, w_{1:N}) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n,\beta)$
- 变分近似分布 $q(\theta, z_{1:N}) = q(\theta)\prod_{n=1}^{N} q(z_n)$

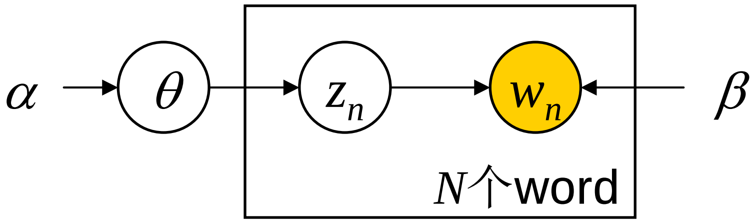
套公式 $\log q(x_k) = E_q\left[\log p(x_H, x_E) \mid x_k\right] + const$

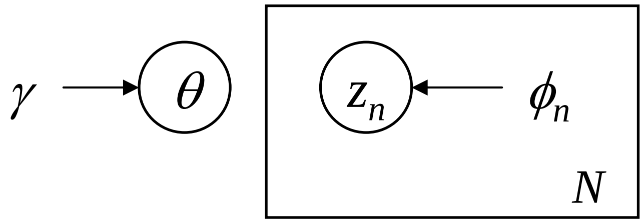
$$\log q(\theta) = \langle \log p(\theta|\alpha)\rangle_{q(\cdot|\theta)} + \sum_n \langle \log p(z_n|\theta)\rangle_{q(\cdot|\theta)} + const$$

$$= \log p(\theta|\alpha) + \sum_n \begin{array}{l} \sum_{z_n=k} q(z_n=k) \log p(z_n=k|\theta) \\ = \sum_k \phi_{n,k} \log \theta_k \end{array} = \sum_k \left(\sum_n \phi_{n,k}\right) \log \theta_k$$

$$p(\theta|w_{1:N}) \approx \quad q(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \times \prod_{k=1}^{K} \theta_k^{\sum_n \phi_{n,k}} = \prod_{k=1}^{K} \theta_k^{\alpha_k + \sum_n \phi_{n,k} - 1}$$

LDA – Variational inference II





- 原概率联合分布 $p(\theta, z_{1:N}, w_{1:N}) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta)$
- 变分近似分布 $q(\theta, z_{1:N}) = q(\theta) \prod_{n=1}^{N} q(z_n)$

套公式 $\log q(x_k) = E_q\left[\log p(x_H, x_E) \mid x_k\right] + const$

$$\log q(z_n) = \left\langle \log p(z_n \mid \theta) \right\rangle_{q(\,\cdot\,\mid z_n)} + \left\langle \log p(w_n \mid z_n, \beta) \right\rangle_{q(\,\cdot\,\mid z_n)} + const$$

$$\log q(z_n = k) = \left\langle \log \theta_k \right\rangle_{q(\theta)} + \log \beta_{z_n = k, w_n}$$

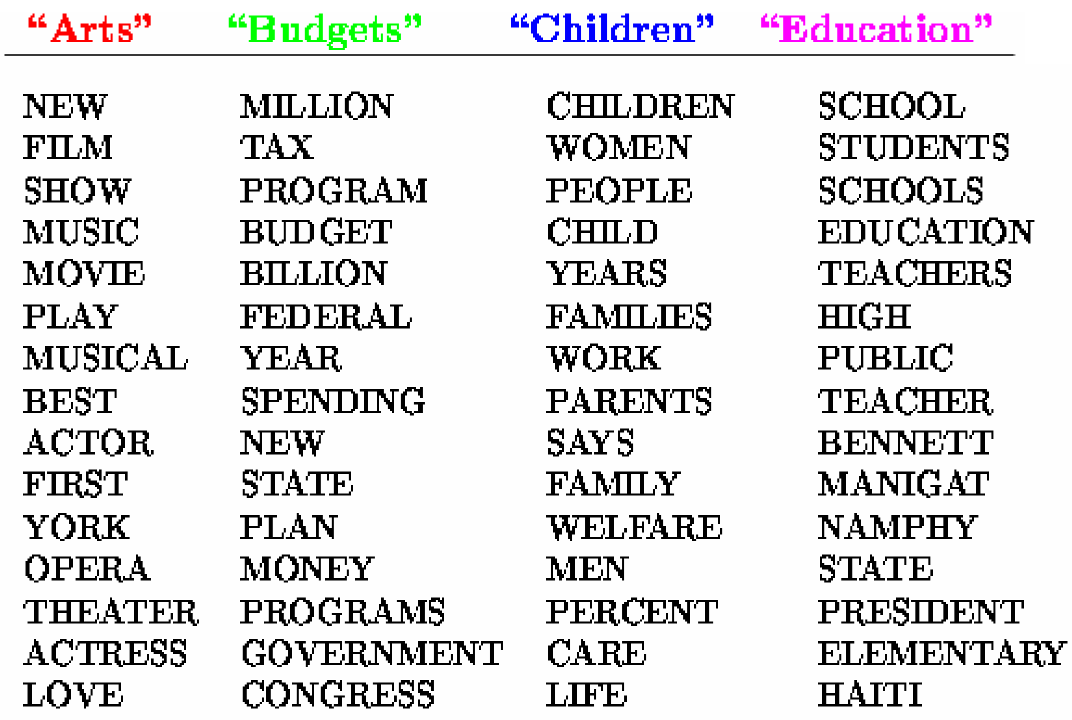
$$\phi_{n,k} \propto \beta_{k, w_n} \exp\left\{ \left\langle \log \theta_k \right\rangle_{q(\theta)} \right\}$$

Experimental result

- Data
  - A subset of TREC AP corpus (Newswire articles)
  - 16,333 documents (90% training, 10% held-out)
  - 23,075 unique terms
  - Removed 50 stop words, words appearing once
- Train a 100-topic LDA model
- Perform inference on a held-out document

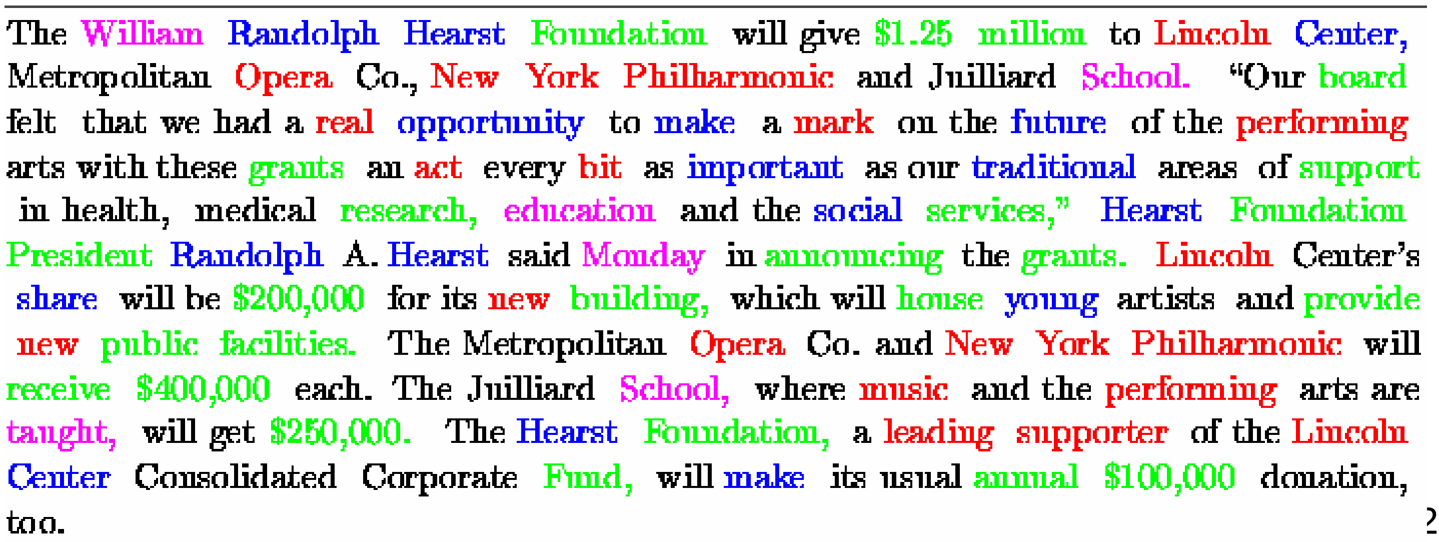
根据β的第k行一topic k下word分布，选出在topic k中出现概率较大的word (top-word)

据此，为学习到的topic取名



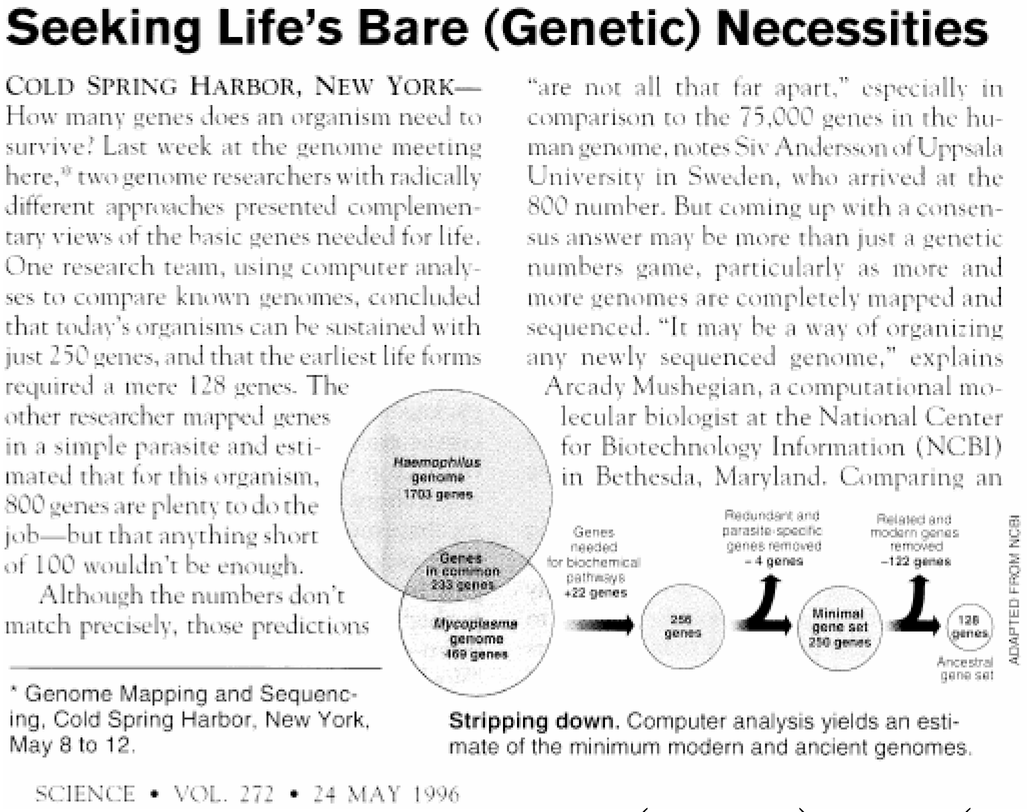
“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

第n个word w_n的颜色为k，如果 $p(z_n=k \mid w_{1:N}) \approx q(z_n=k)>0.9$



The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

LDA for document classification



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.

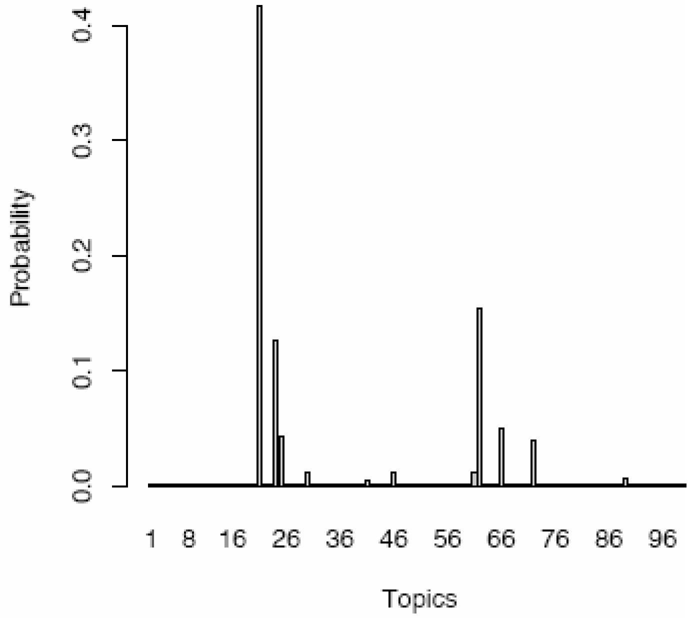
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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.

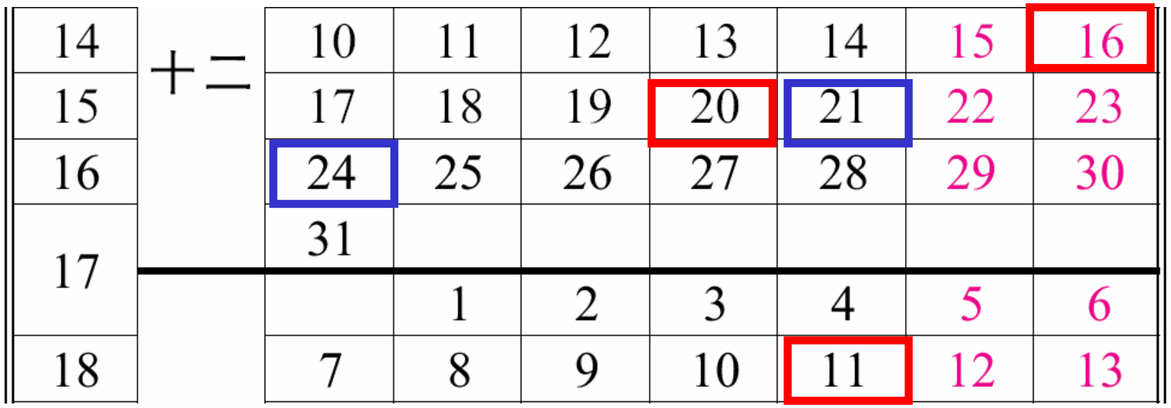
SCIENCE • VOL. 272 • 24 MAY 1996



$$p(\theta \mid w_{1:N}) \approx Dir(\theta \mid \gamma)$$

Reduce $w_{1:N}$ to $\overline{\gamma}(w_{1:N})$

通知 @14th week（严格时间点）



14	十二	10	11	12	13	14	15	16
15		17	18	19	20	21	22	23
16		24	25	26	27	28	29	30
17		31						
			1	2	3	4	5	6
18		7	8	9	10	11	12	13

- 第14周周末（12月16日）23:59前：每位同学递交评估版报告
  - 按要求的（中文）模板书写
- 第15周周四（12月20日）23:59前：每位同学返回互评表
  - 书写清晰, 新意及深入程度, 工作量及完善程度
- 第15周周五（12月21日）23:59前：网络学堂上公布选做口头报告的同学名单
- 第16周周一（12月24日）the last lesson：口头报告
- 课程大作业提交截止
  - 现场检查时间：1月11日，每人10分钟ppt汇报（含演示）

课程章节

- 第一章 引言（1）
- 第二章 图模型的表示理论（3）
  - DGM-UGM
  - Semantics
  - HMM

感谢对课程的支持！

  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：maxlikelihoodEstimate，BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）