Machine Learning 读书会第8期

主题模型简介

沈志勇
Data scientist @ IDL.Baidu

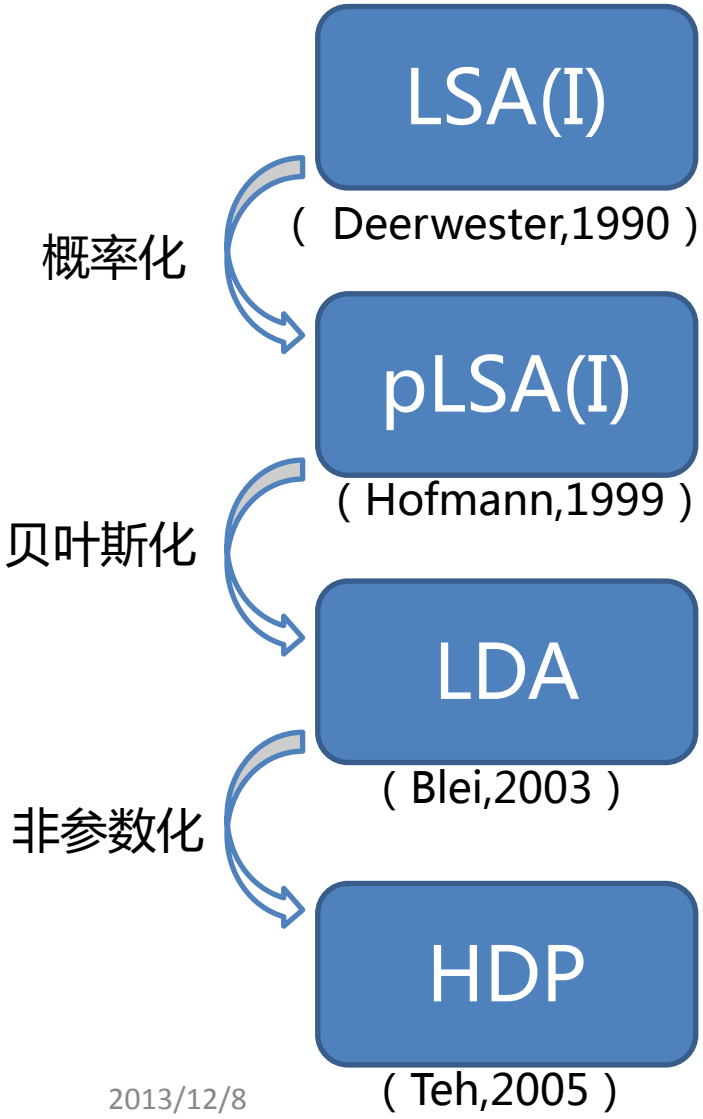
合办方：超级计算大脑研究部@自动化所

rakecas@gmail.com

outline

- What's topic model
  - 来龙去脉
  - 相关模型的比较
- Learning topic model
  - 来自模型的痛苦
  - 来自数据的痛苦
- Using topic model
  - 参数的使用
  - 模型层面的利用：表达能力、学习机制等

来龙去脉



全称：Latent Sematic Analysis (Indexing)
优势：刻画近义词，计算word和doc的距离

全称：probabilistic Latent Semantic Analysis (Index)
优势：更好刻画一词多义，用多项式分布描述词频向量

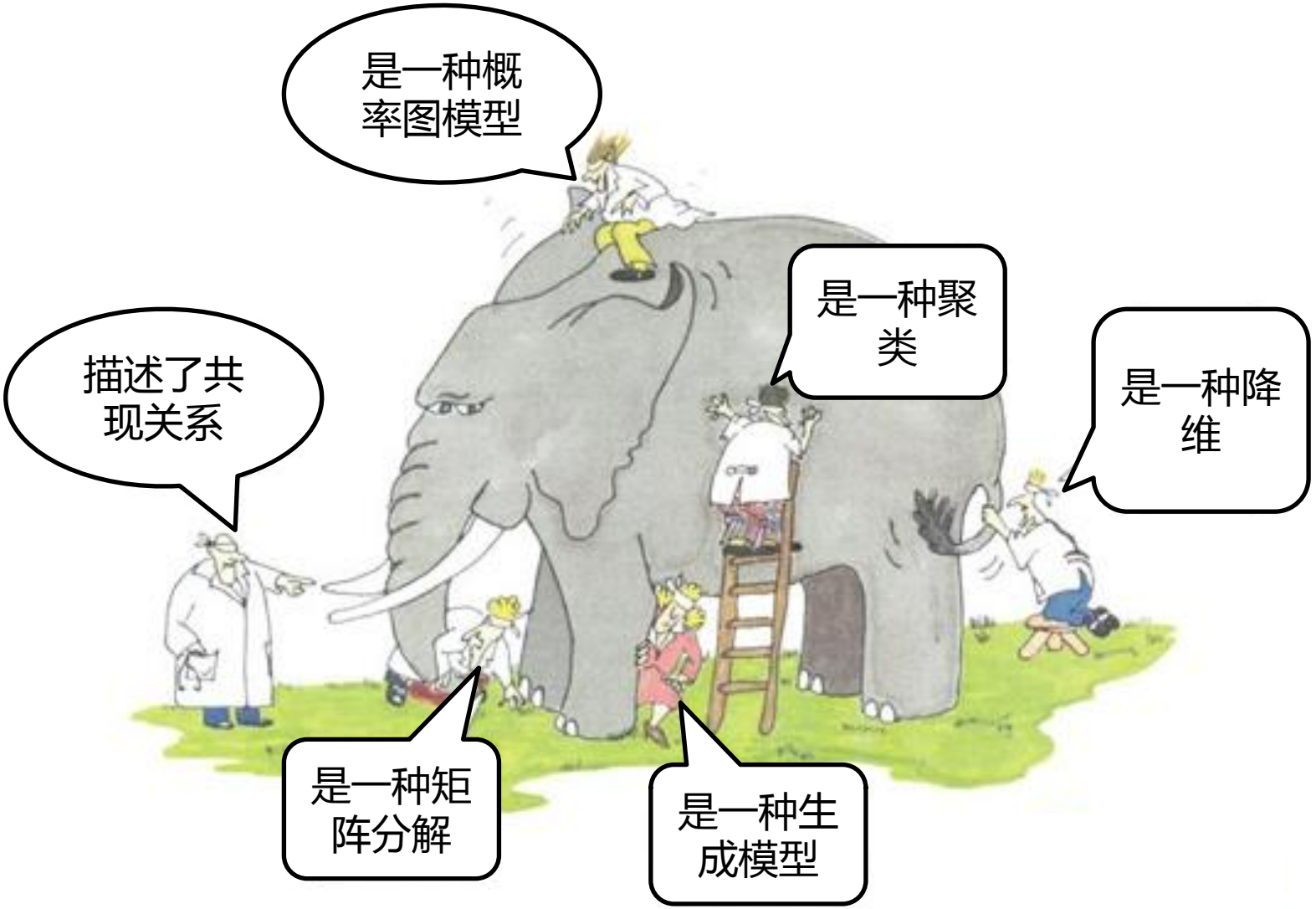
Topic model（狭义）

全称：Latent Dirichlet Allocation
优势：贝叶斯化带来的各种好处（后面细说）

全称： Hierarchical Dirichlet Process
优势：自动确定topic的个数

rakecas@gmail.com

什么是Topic Model?



rakecas@gmail.com

生成模型：汪老师写歌词

有一位网友统计了汪老师在大陆发行的9张专辑共117首歌曲的歌词

同一词语在一首歌出现只算一次。形容词，名次和动词的前十名分别是（词语后面的数字是出现的次数）：

	形容词		名词		动词
0	孤独：34	0	生命：50	0	爱：54
1	自由：17	1	路：37	1	碎：37
2	迷惘：16	2	夜：29	2	哭：35
3	坚强：13	3	天空：24	3	死：27
4	绝望：8	4	孩子：23	4	飞：26
5	青春：7	5	雨：21	5	梦想：14
6	迷茫：6	6	石头：9	6	祈祷：10
7	光明：6	7	鸟：9	7	离去：10
9	理想：6	8	瞬间：8	8	再见：9
9	荒谬：5	9	桥：5	9	埋：6

如果我们随便写一串数字，然后按数位，依次在形容词，名次和动词中取出个词，连在一起会怎样呢？

比如圆周率3.1415926，对应的词语就是：**坚强，路，飞，自由，雨，埋，迷惘。**

稍微连接和润色一下：

坚强的孩子，

依然前行在路上，

张开翅膀飞向自由，

让雨水埋葬他的迷惘。

再来一个，今年是2013对应的词语就是：**迷惘，生命，碎，坚强。**

润色一下：

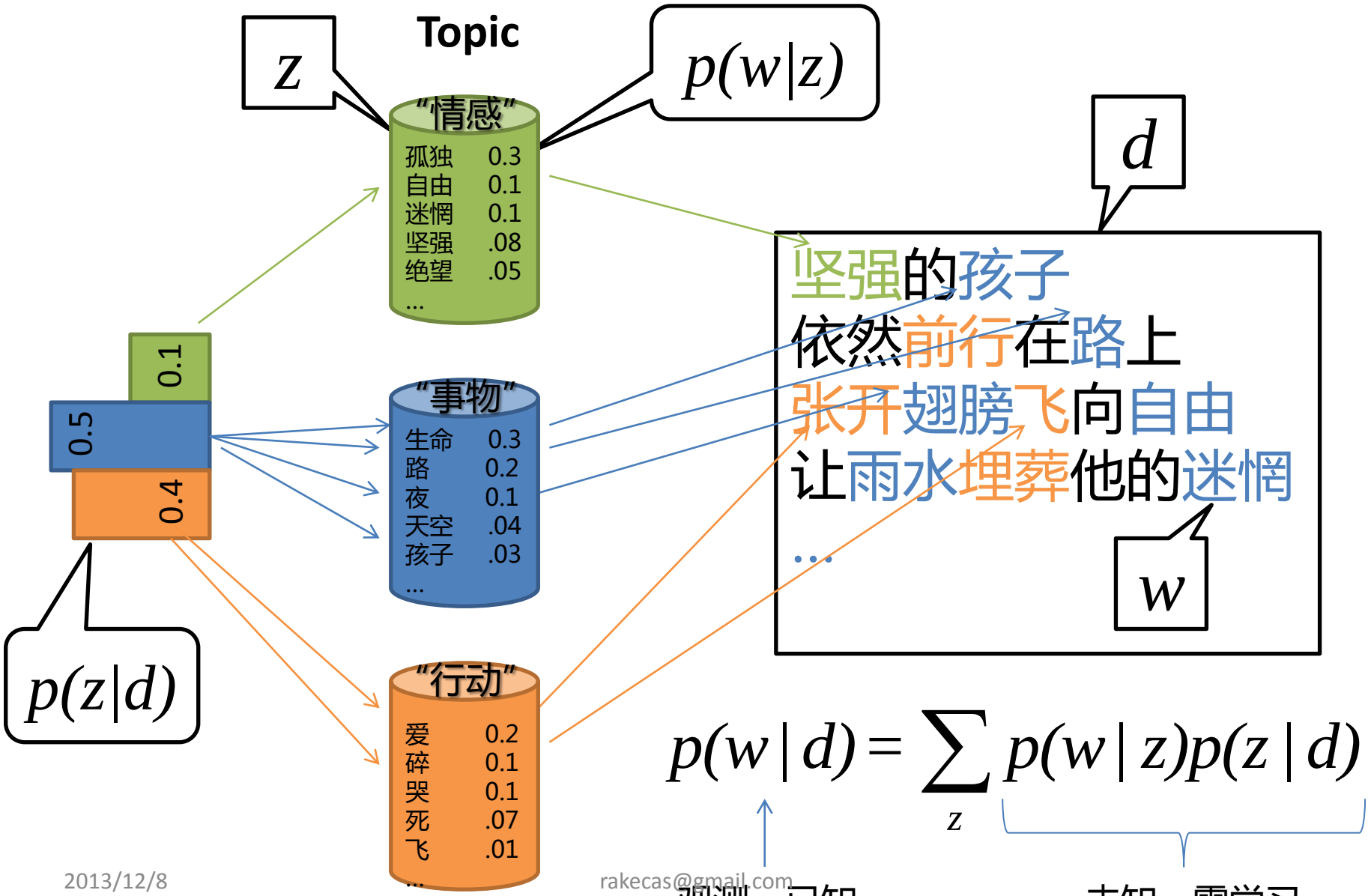
不再迷惘的生命，

被燃碎千万次，也依然坚强。

（来自 @于弦cos 微博图片）

rakecas@gmail.com

生成模型：PLSA

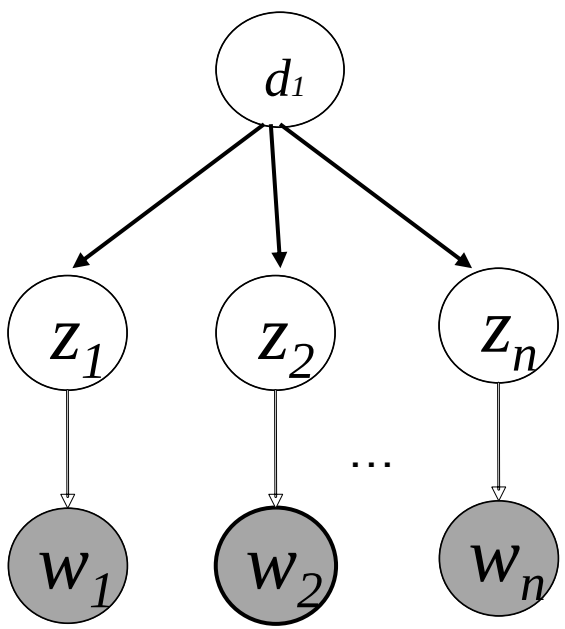


$$p(w \mid d) = \sum_{z} p(w \mid z) p(z \mid d)$$

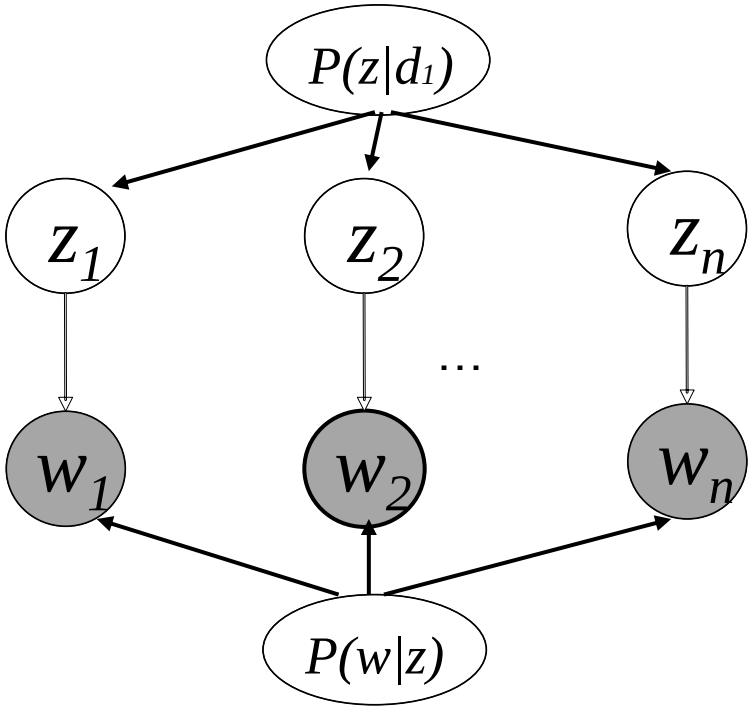
观测，已知　　未知，需学习

rakecas@gmail.com

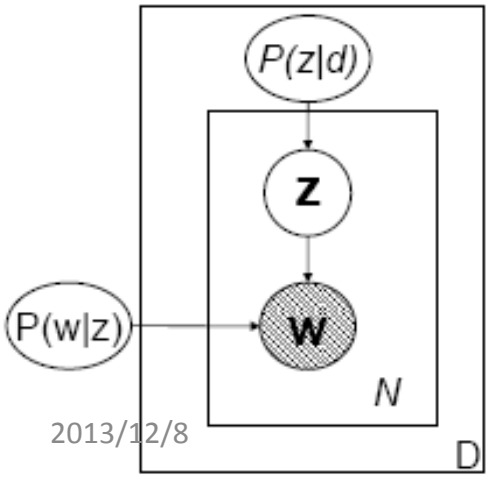
PLSA的图模型表达



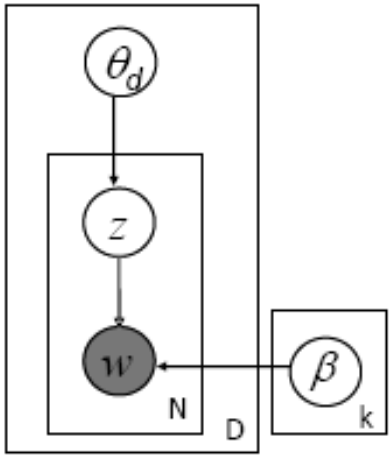
引入参数



盘子表达

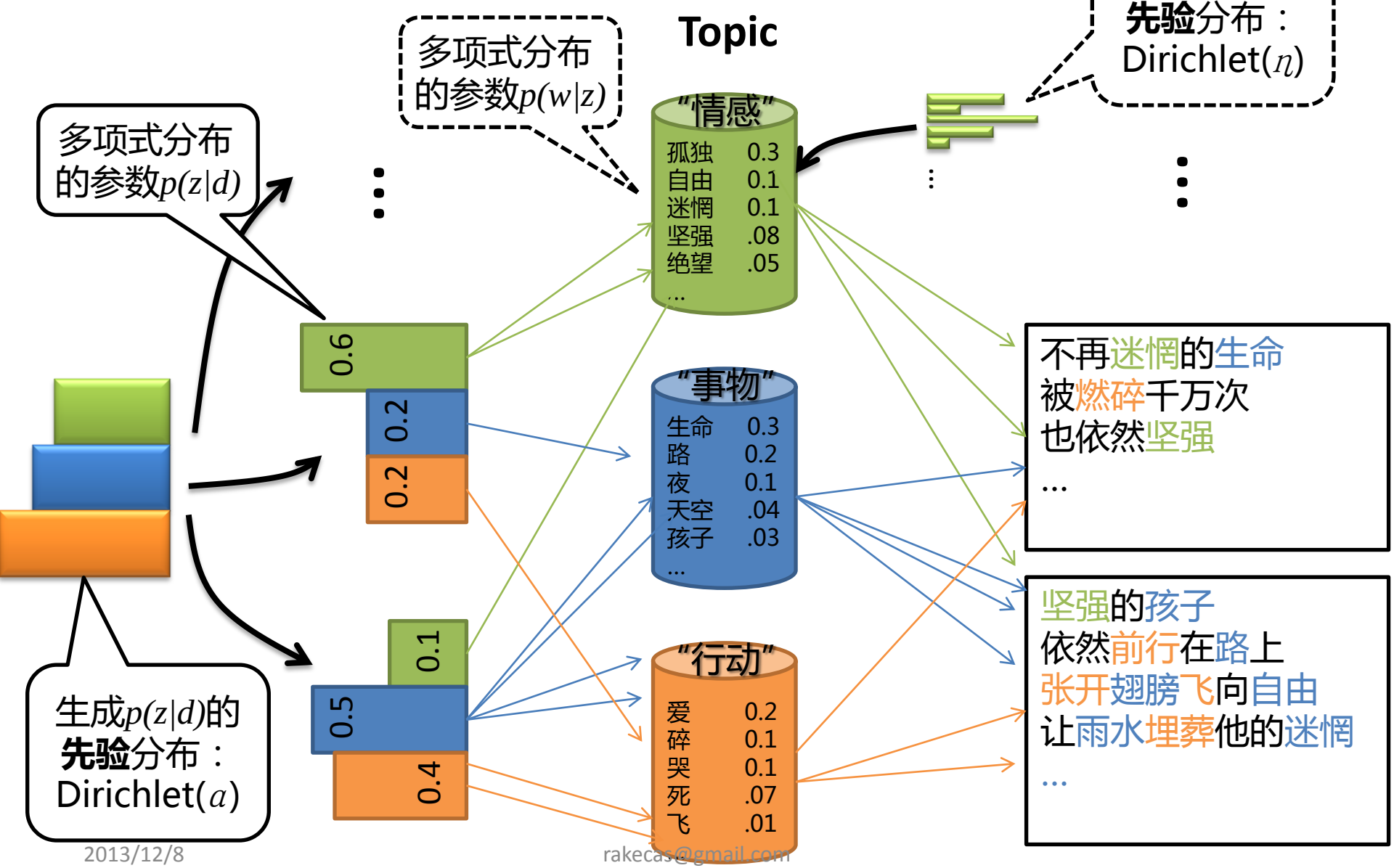


希腊字母表达



rakecas@gmail.com

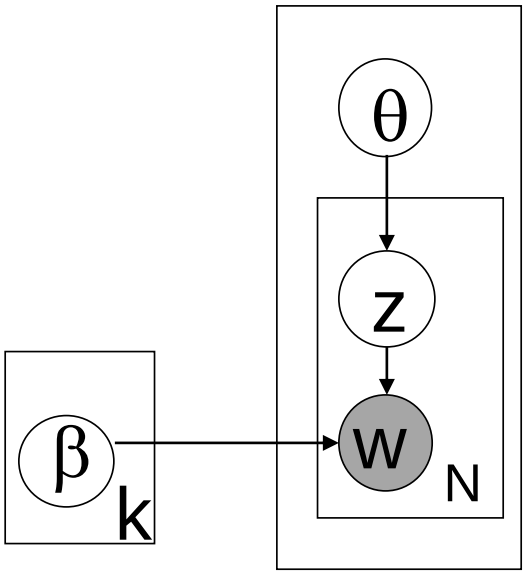
生成模型：LDA



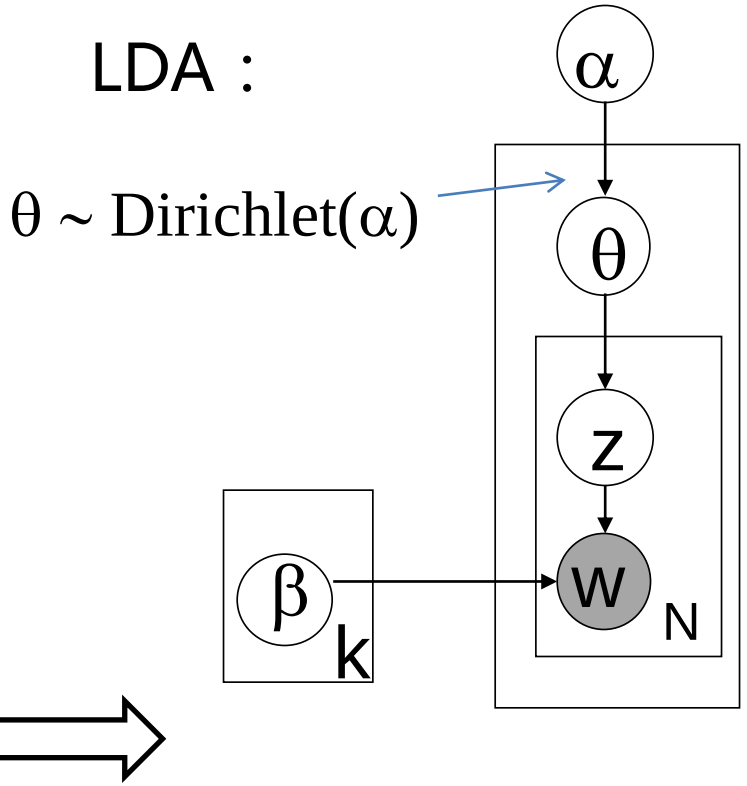
rakecas@gmail.com

LDA的图模型表达

PLSA：



LDA：



- 贝叶斯化
- 层次化（deep化）

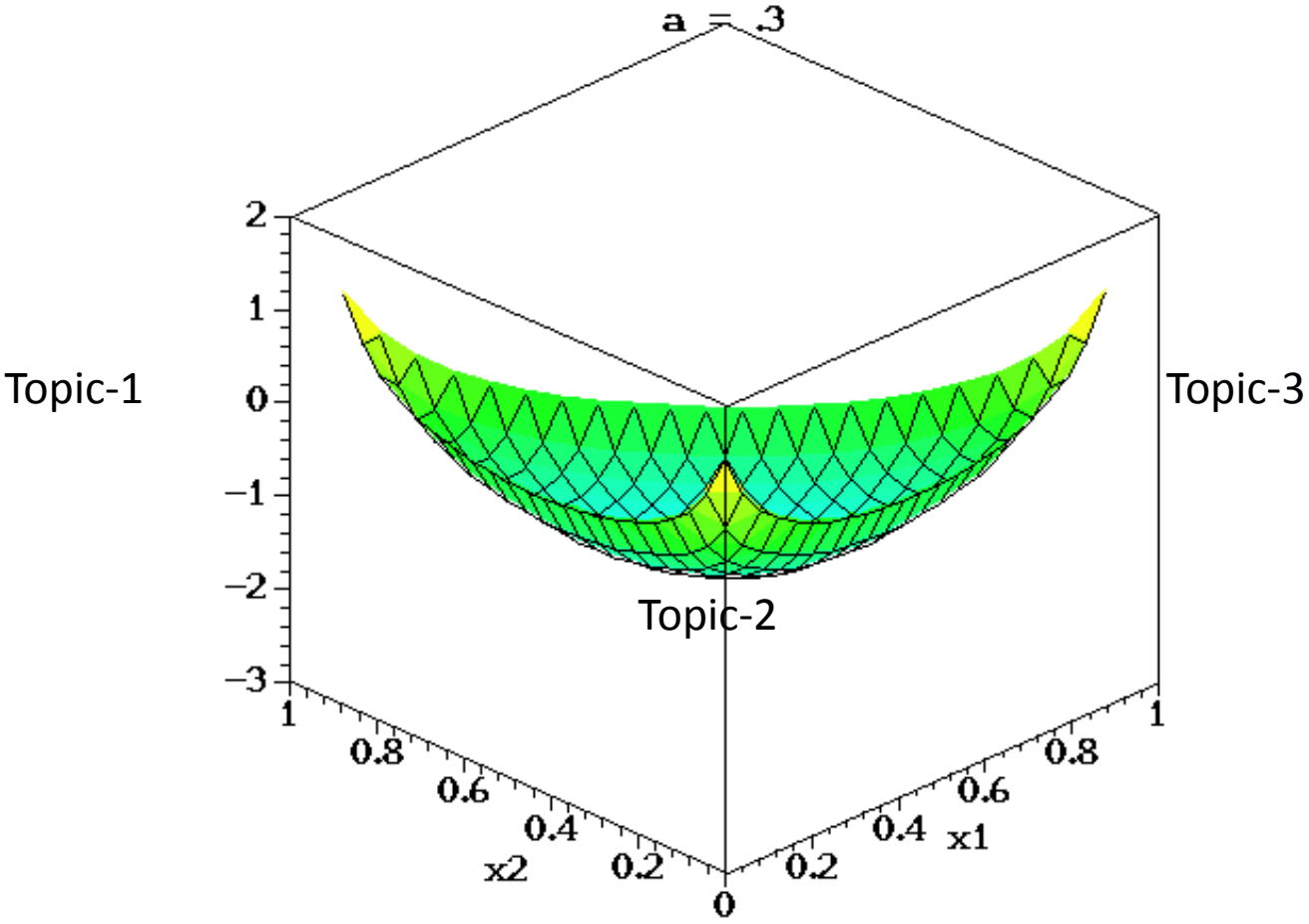
rakecas@gmail.com

从PLSA到LDA （续）

- 参数$p(z|d)$有了以α为参数的先验（贝叶斯化）
  - 若α未知、需学习
    - 模型可描述整个Corpus的生成
    - 不同的文档之间的$p(z|d)$不独立：短的doc可从长的doc得到帮助；“差”的doc不会带来太大的影响
  - 若α由人为设定
    - 注入领域知识
    - MAP，相当于正则——稀疏表达 + 平滑
- 为什么是Dirichlet?
  - 指数族分布
  - 多项式分布的共轭先验

学习的便利

rakecas@gmail.com



3维Dirichlet的参数向量 $\vec{\boldsymbol{\alpha}} = (\alpha, \alpha, \alpha)$

- 当 α 小于 1 时，生成的概率向量偏向于某几维，值越小，偏的越厉害 – **稀疏表达**
- 当 α 大于 1 时，生成的概率向量倾向于中间，值越大，趋中越厉害 – **平滑**

rakecas@gmail.com

TM是降维：对比PCA

- PCA ：

目标最小化下面的 X 重构误差：

$$J(\mathbf{W},\mathbf{Z}) = ||\mathbf{X} - \mathbf{W}\mathbf{Z}^T||_F^2$$

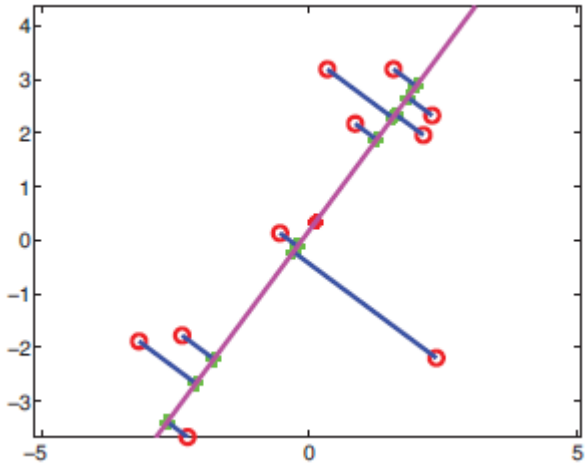
其中W正交,和X同维数，Z 为score（投影坐标）矩阵，降维

PCA → PLSA

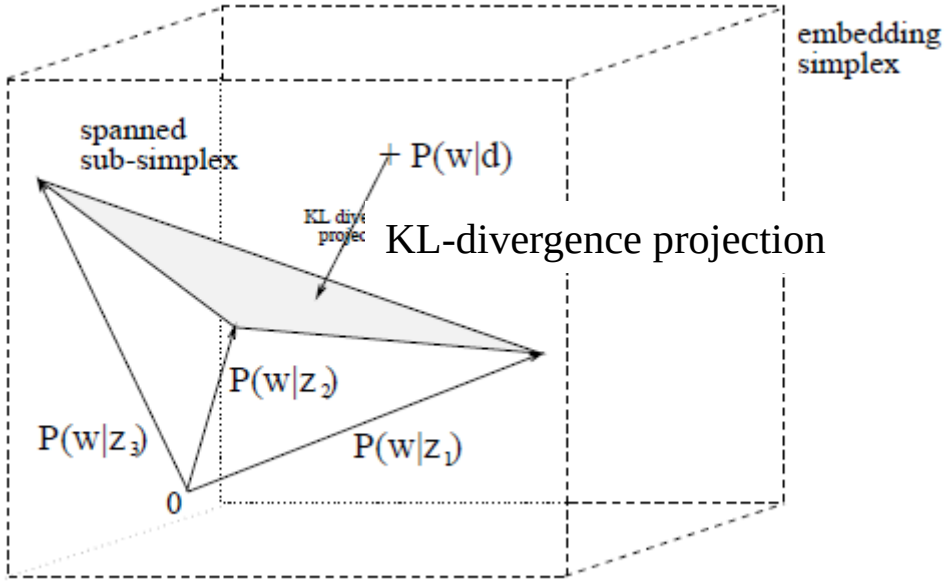
$X \rightarrow p(w|d)$

$W \rightarrow p(w|z)$

$Z \rightarrow p(z|d)$



（图片来自Murphy的MLAPP）

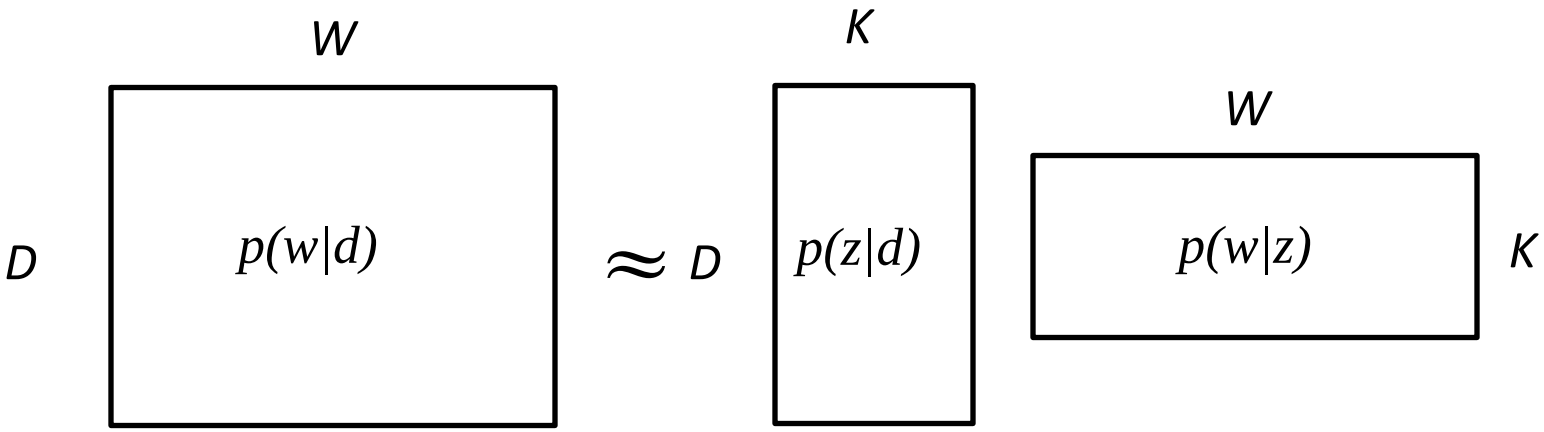


（图片来自Hofmann的PLSA论文）

rakecas@gmail.com

PLSA是矩阵分解：对比NMF

$$\vec{p}(w|d_i) = \sum_z \vec{p}(z|d_i)\vec{p}(w|z)$$

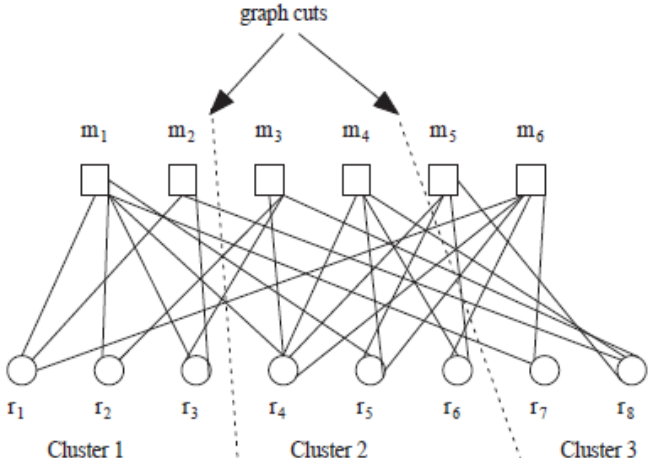


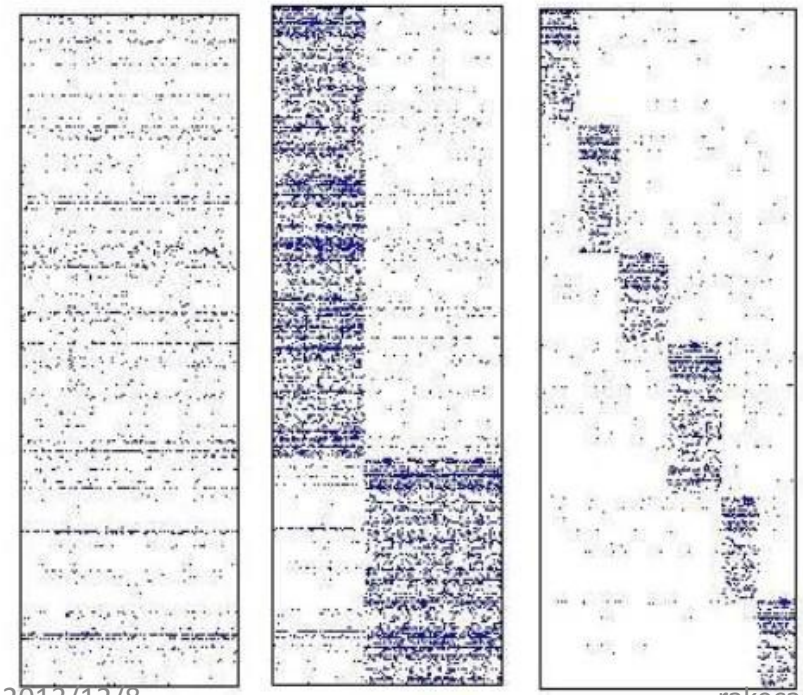
1. $K<<W, K<<D$
2. 所有矩阵行和为1，且所有元素大于0（概率）

（Chris Ding AAAI06 的论文中给出了一个等价性证明）

Takecas@gmail.com

TM是聚类：对比co-clustering





（图片来自Rege DMKD2008）

- TM 是一种soft co-clustering

$p(z|d)$: doc 类从属概率
$p(z|w)$: word 类从属概率

如何求$p(z|w)$？

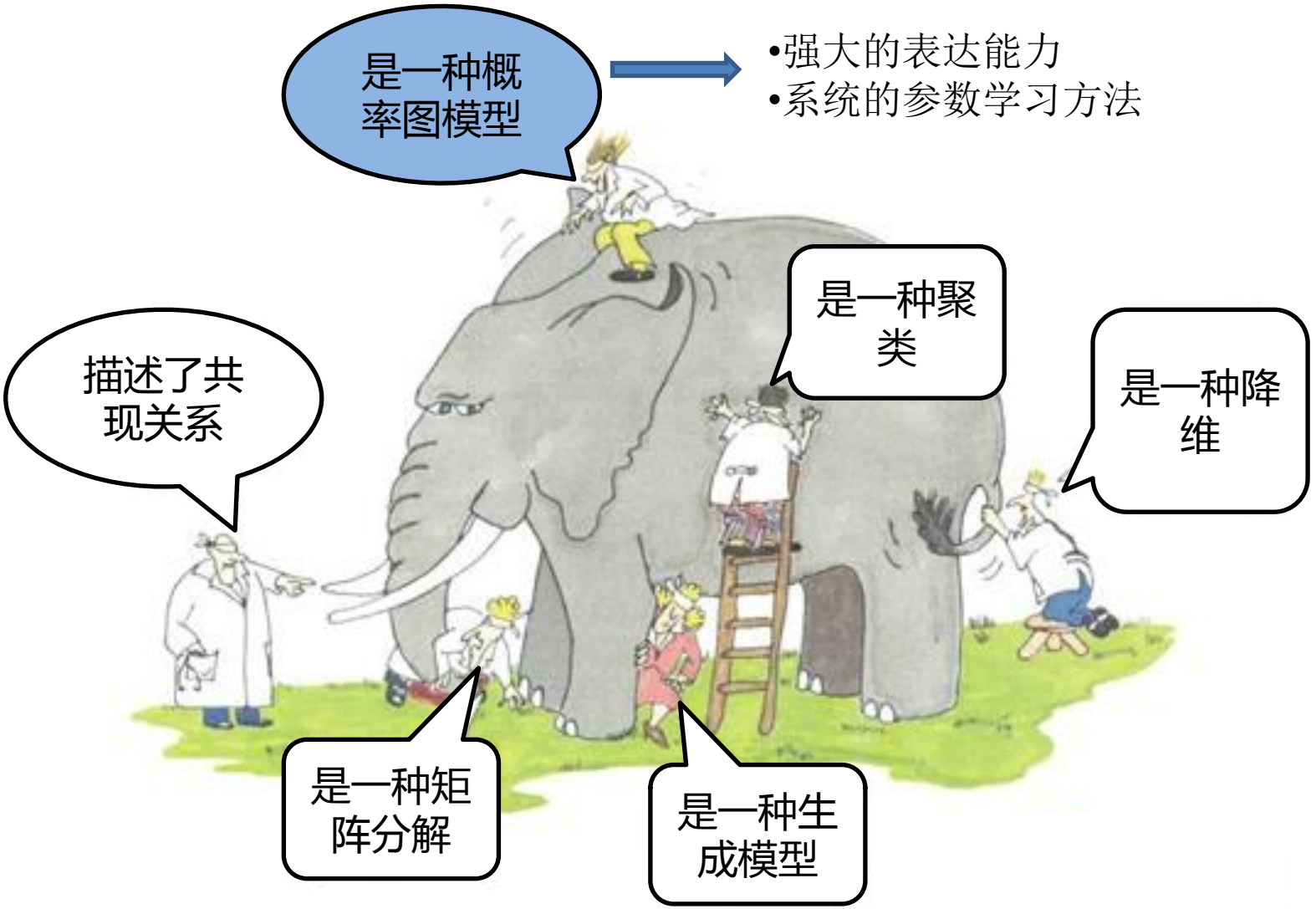
- 在TM的参数上很容易得到传统co-clustering的结果：

$$C_d = \arg\max_c \; p(z = c \mid d)$$

$$C_w = \arg\max_c \; p(z = c \mid w)$$

rakecas@gmail.com

什么是Topic Model-回顾



rakecas@gmail.com

outline

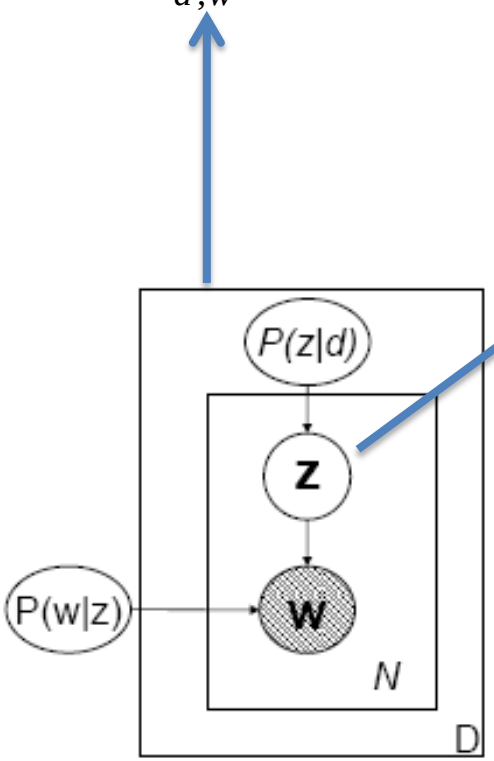
- What's topic model
  - 来龙去脉
  - 相关模型的比较
- **Learning topic model**
  - **来自模型的痛苦**
  - 来自数据的痛苦
- Using topic model
  - 参数的使用
  - 模型层面的利用：表达能力、学习机制等

TM参数学习：最大化似然

PLSA：

$$\sum_{d,w} \log P(d,w) = \sum_{d} \sum_{w} O(w,d) \log \sum_{z} P(w|z)P(z|d)$$

No close form resolution



- 似然表达式中:
  - 盘子带来log
  - 隐变量带来求和（积分）

- PLSA参数学习的痛苦来自于模型中，处在盘子里的隐变量z

rakecas@gmail.com

$$\mathcal{L} = \sum_{d,w} \log \sum_z \mathcal{Q}(z) \frac{p(w|z;\beta)p(z|d;\theta)}{\mathcal{Q}(z)}$$

(Jensen's inequality)

$$\geq \sum_{d,w} \sum_z \mathcal{Q}(z) \log \frac{p(w|z;\beta)p(z|d;\theta)}{\mathcal{Q}(z)}$$

where

$$\mathcal{Q}(z) = p(z|d, w; \beta^{\text{old}}, \theta^{\text{old}})$$

Jensen不等式可以把求和从log中提出来，得到原似然的一个下界

迭代起来就是——**EM算法**：

- E步： 更新Q(z)
- M步： 最大化下界

rakecas@gmail.com

来自数据的痛苦

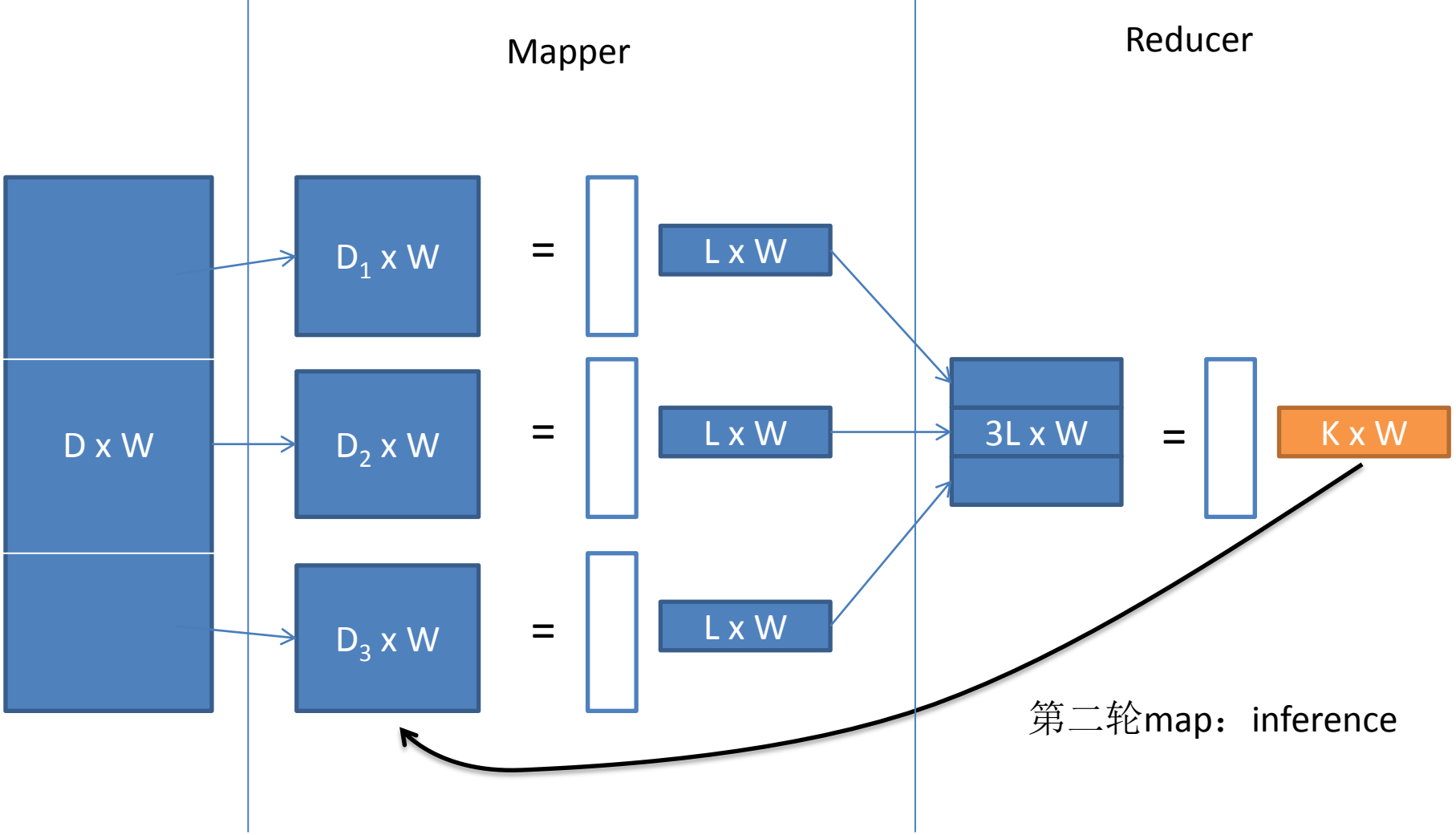
- 数据“少”
  - Word 出现频率的幂律分布
  - 部分doc较短
  - 所有doc都是短文本

Burn-in + voting

→ 将短文根据某种context拼成长文

- 数据多
  - Online 算法 → EM和bayes框架都有较好的online特性
  - 分布式算法
    - 以MPI为主
    - 本人提出过一种2个map+1个reduce的近似解法

rakecas@gmail.com



参见拙作：Topic Modeling Ensembles. ICDM 2010

rakecas@gmail.com

outline

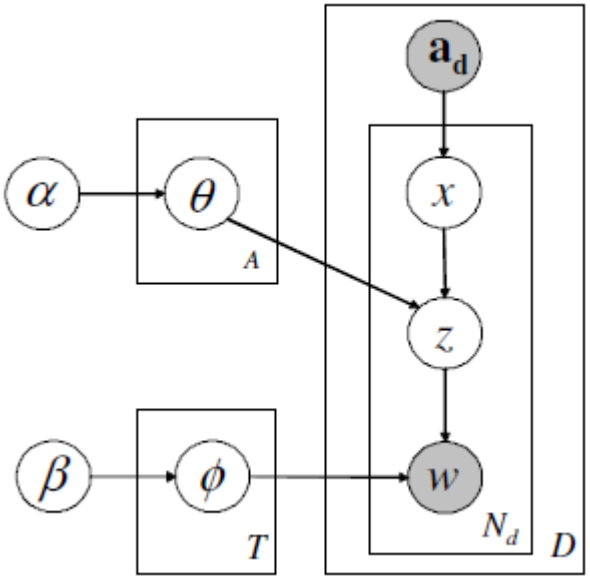
- What's topic model
  - 来龙去脉
  - 相关模型的比较
- Learning topic model
  - 来自模型的痛苦
  - 来自数据的痛苦
- **Using topic model**
  - **参数的使用**
  - **模型层面的利用：表达能力、推理机制等**

rakecas@gmail.com

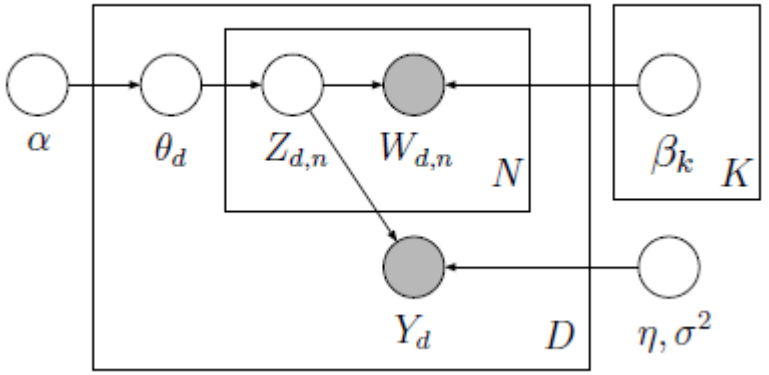
参数的使用

- 各种条件概率的使用
  - p(w|z) – top words主题相关的关键词表
  - p(z|d) – doc的soft clustering；doc的低维表达等
  - p(w'|w) – 关联关系，可用于关联推荐等
  - 可视化
- 超参数的使用 – 注入领域知识（supervision）
  - p(w|z)的超参数η – 某些topic已知的关键字提权
  - p(z|d)的超参数α –如果η有监督信息，也可以相应控制权重；一般情况，短文本适当取大一点
  - K的确定：肉眼看最靠谱；有一些系统评估P(D|K)的方法，没必要

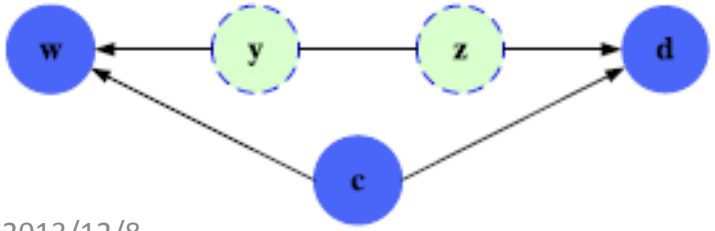
运用图模型强大的表达能力



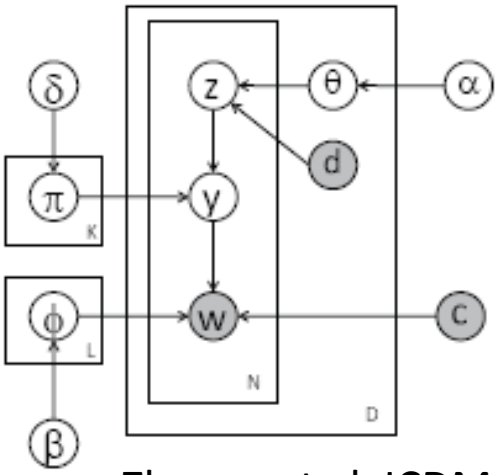
Rosen et al. UAI 2004



Blei et al. NIPS 2007



Zhuang et al. TKDE 2012



Zhuang et al. ICDM 2010

rakecas@gmail.com

利用模型的可推理机制

- 固定的$p(w|z)$, 动态变化的doc
  - 可以随时对p(z|d)做inference
  - p(w'|w_1,w_2,…)，高阶association
- 利用优质数据学习$p(w|z)$，对剩余doc上的$p(z|d)$只作inference
  - Inference过程是独立的
  - 也可以在一定程度上解决单机内存不足的问题

Thanks

PDF版略作删节，有任何问题请联系：

Weibo： @沈醉2011
Email: rakecas@gmail.com
QQ: 7997868