UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

BIG DATA
Analytics & Management

Lecture 10 (04/17, 04/19): Topic Modeling

Decisions, Operations & Information Technologies
Robert H. Smith School of Business
Spring, 2017

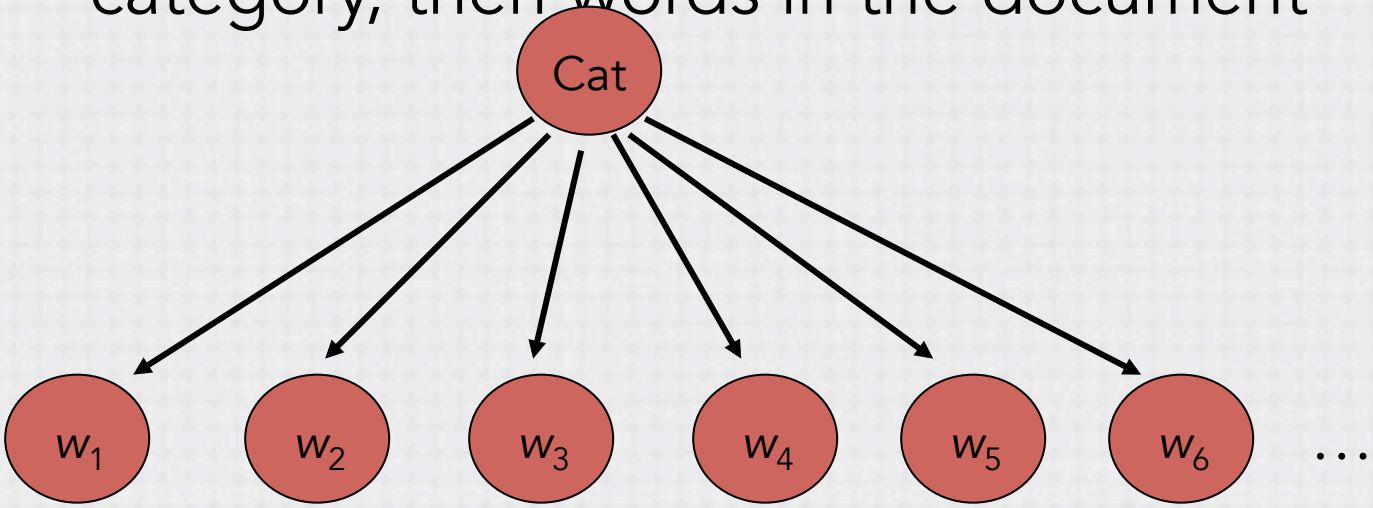
Outline

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Naïve Bayes model
- Topic model – Latent Dirichlet Allocation
  - ❑Graphical model
  - ❑Posterior inference (MCMC)
  - ❑Evaluation of LDA
  - ❑Implementation of LDA

Naïve Bayes model

- *Supervised* text categorization through Naïve Bayes
- Generative model: first "generate" a document category, then words in the document

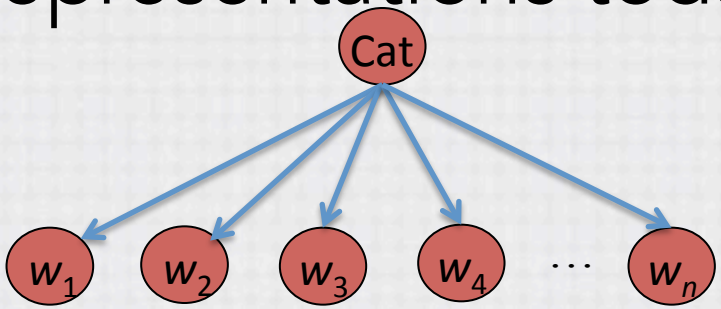


- Inference: obtain posterior over document categories using Bayes rule (argmax to choose the category)

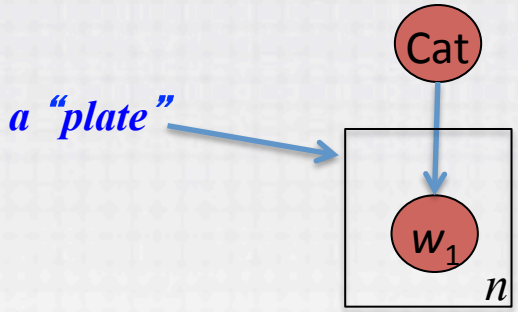
$$P(Cat \mid w_{1\ldots n}) = \frac{P(w_{1\ldots n} \mid Cat)P(Cat)}{P(w_{1\ldots n})}$$

Compact graphical model representations

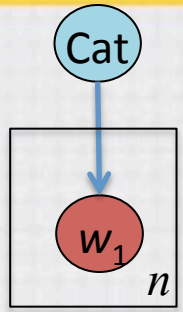
- We're going to lean heavily on graphical model representations today.



- We'll use a more compact notation:



"generate a word from Cat n times"



- Now suppose that Cat isn't observed
- We need to learn two distributions:
  - P(Cat)
  - P(*w*|Cat)
- How do we do this?
  - ❑We might use the method of maximum likelihood (MLE)

MLE example

- Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ

X	0	1	2	3
P(X)	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

MLE example

- The likelihood is:

$$L(\theta) = P(X=3)P(X=0)P(X=2)P(X=1)P(X=3)P(X=2)P(X=1)P(X=0)P(X=2)P(X=1)$$

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

Let us look at the log likelihood function

$$\begin{aligned} l(\theta) &= \log L(\theta) = \sum_{i=1}^{n} \log P(X_i|\theta) \\ &= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + 2\left(\log\frac{1}{3} + \log(1-\theta)\right) \\ &= C + 5\log\theta + 5\log(1-\theta) \end{aligned}$$

Let the derivative of $l(\theta)$ with respect to θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

Introduction to Topic Models

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

Latent Dirichlet Allocation

David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

What is the topic model?

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

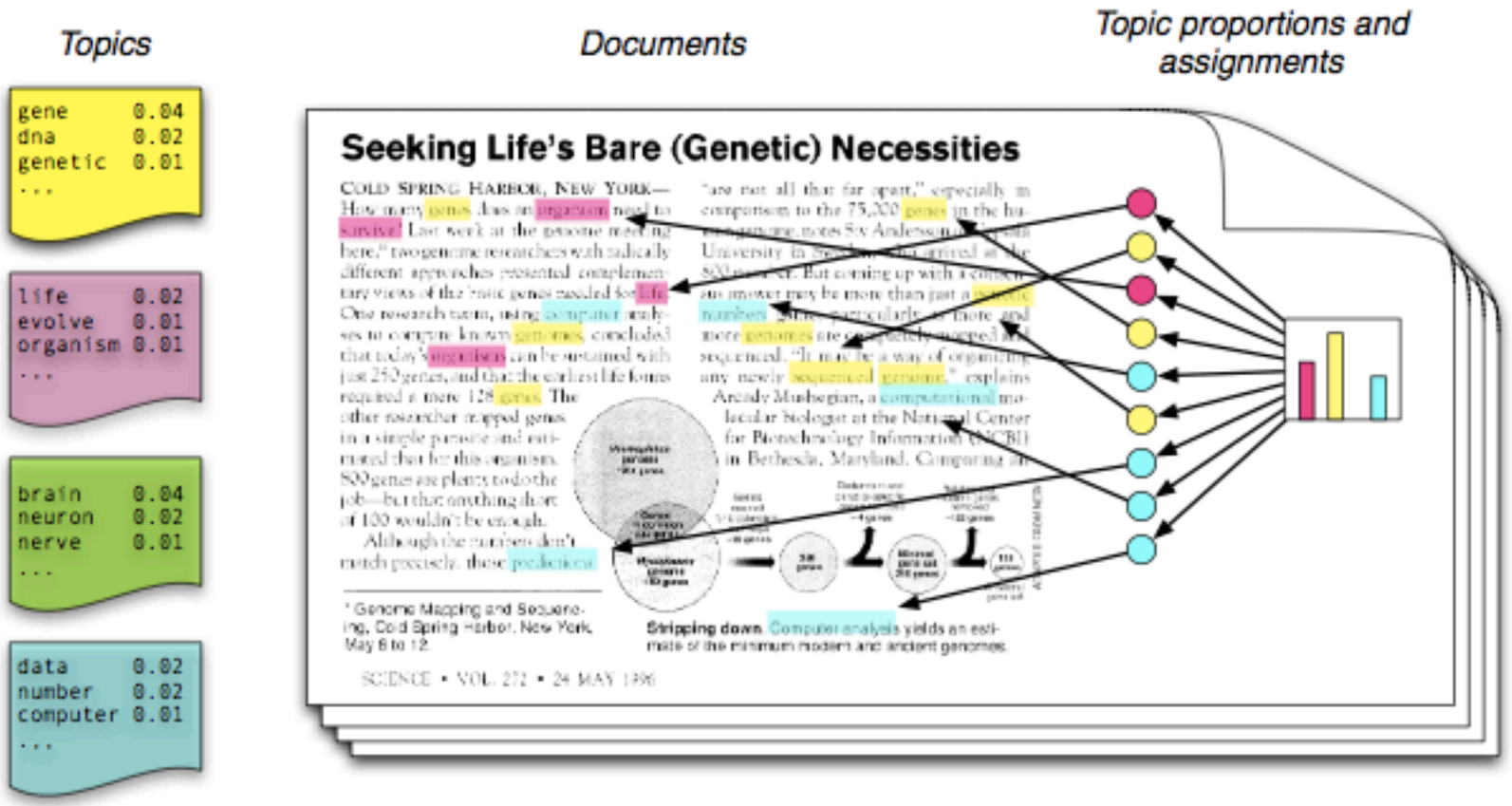
- The topic model is an algorithm that automatically learns topics (themes) from a collection of documents
  - ❑ It works by observing words that tend to co-appear in documents, for example gene and protein, or climate and warming
  - ❑ The topic model assumes each document exhibits multiple topics
  - ❑ The topic model learns topics directly from the text

Document exhibits multiple topics

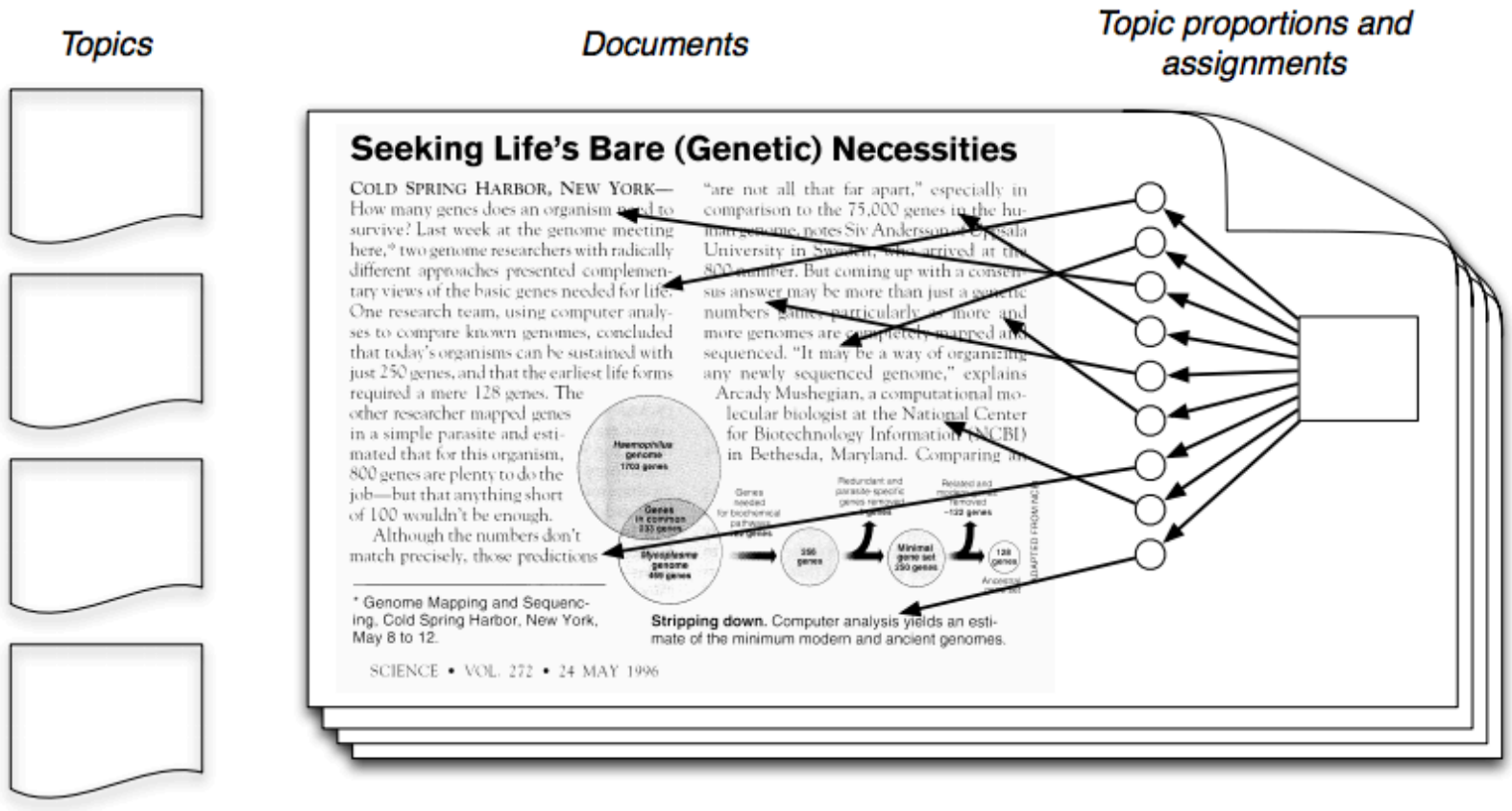
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

"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

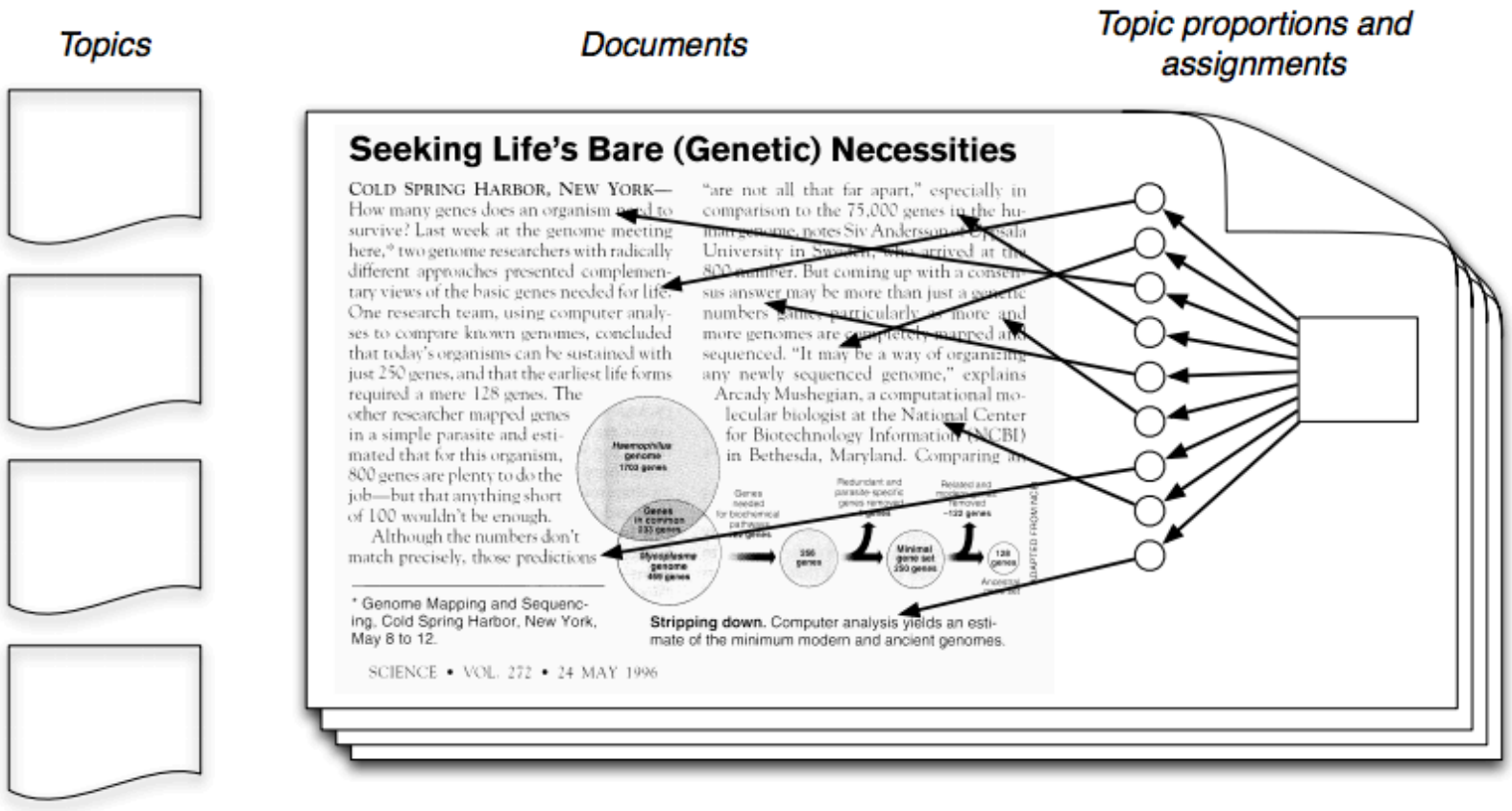
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics



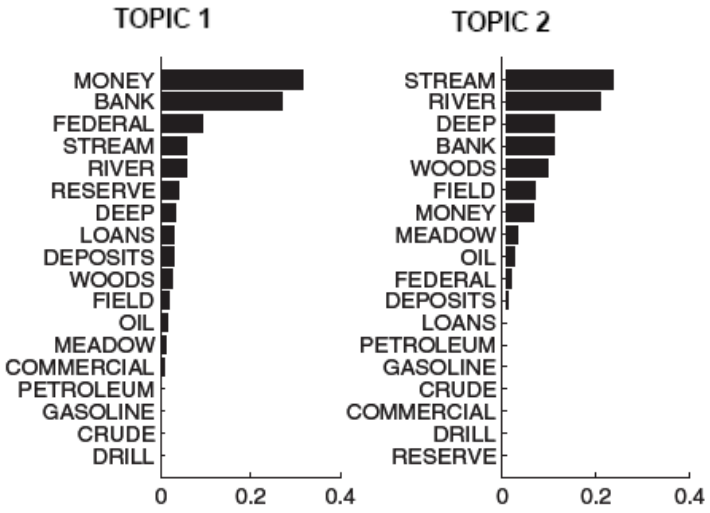
- In reality, we only observe the documents
- The other structure are **hidden variables**



- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Topic models

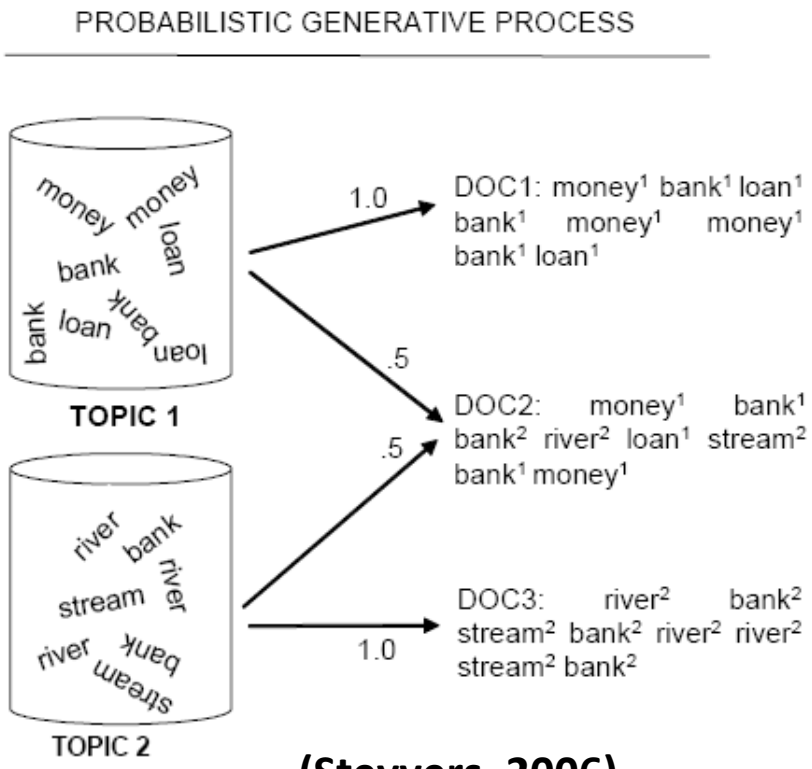


3 latent variables:

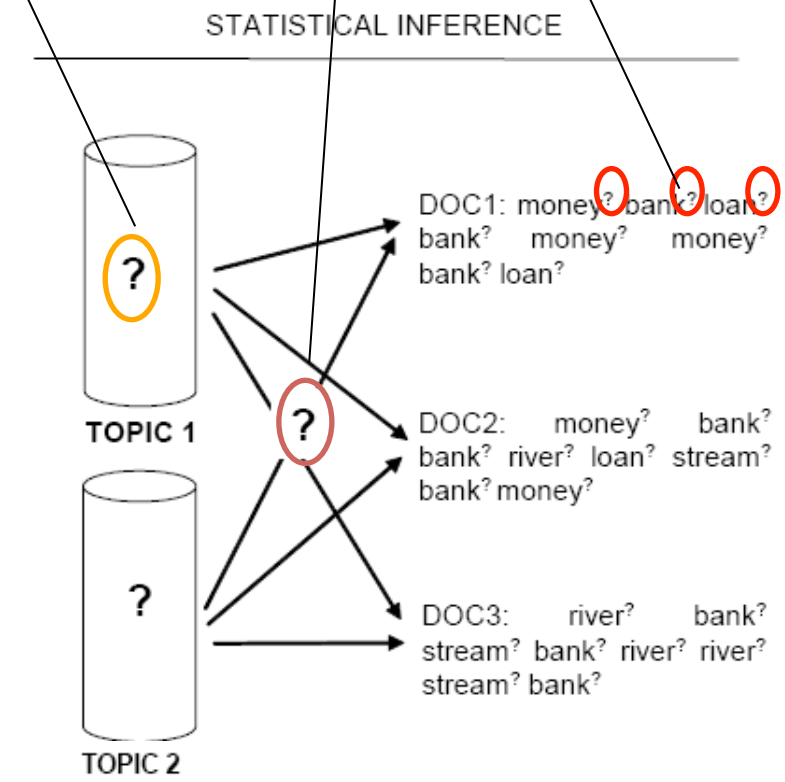
Word distribution per topic (word-topic-matrix)

Topic distribution per doc (topic-doc-matrix)

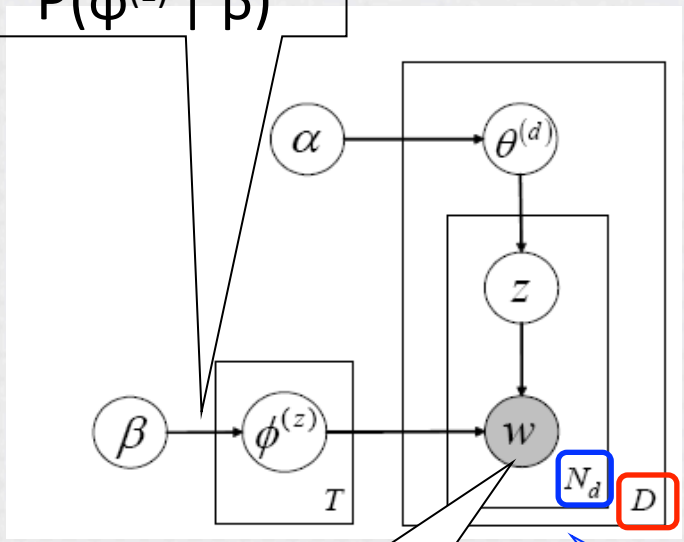
Topic word assignment



(Steyvers, 2006)



Latent Dirichlet Allocation (LDA) (Blei, 2003)



We learn topic distribution of a corpus → we can predict topic distribution of an unseen document of this corpus by observing its words

Hyper-parameters α and β are corpus-level parameters → are only sampled once

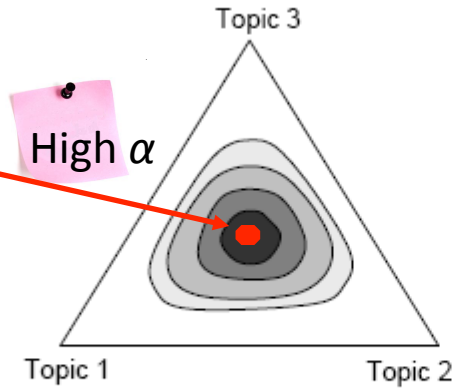
$$P(d,w) = P(d) * P(\theta^{(d)} \mid \alpha) * \sum_{z} P(\phi^{(z)} \mid \beta) * P(w \mid z, \phi^{(z)}) * P(z \mid \theta^{(d)})$$

Dirichlet Prior α

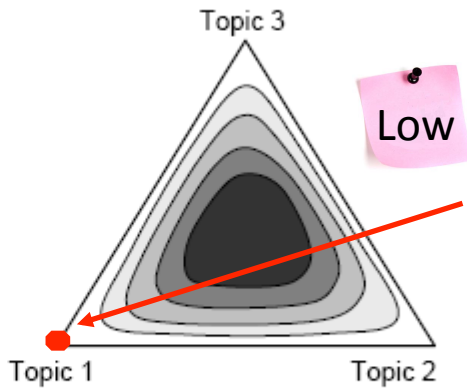
- α is a prior on the topic-distribution of documents (of a corpus)
- α is a corpus-level parameter (is chosen once)
- α is a force on the topic combinations
- Amount of smoothing determined by α
- Higher α → more smoothing → less distinct topics
- Low α → the pressure is to pick for each document a topic distribution favoring just a few topics
- Recommended value: $\alpha = 50/T$ (or less if T is very small)

Doc1

Topic-distribution of Doc1 = (1/3, 1/3, 1/3)



Each doc's topic distribution θ is a smooth mix of all topics



Doc2

Topic-distribution of Doc2 = (1, 0, 0)

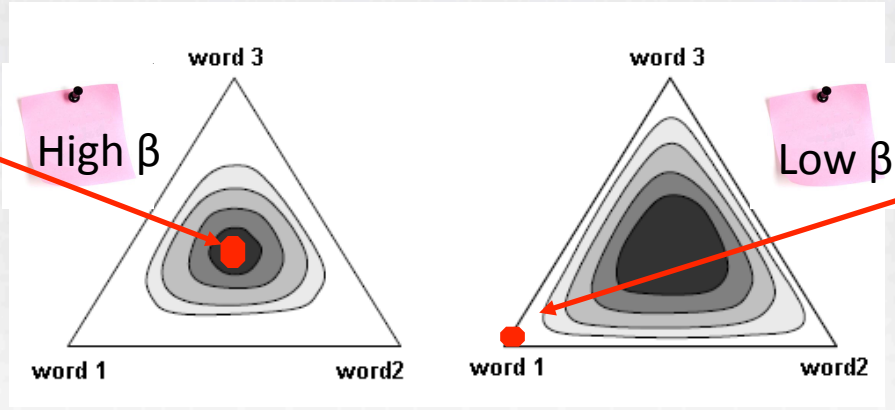
Each doc's topic distribution θ must favor few topics

Dirichlet Prior β

- β is a prior on the word-distribution
- β is a corpus-level parameter (is chosen once)
- β is a force on the word combinations
- Amount of smoothing determined by β
- Higher β → more smoothing
- Low β → the pressure is to pick for each topic w word distribution favoring just a few words
- Recommended values: β = 0.01

Topic1

Word-distribution of Topic1 = (1/3, 1/3, 1/3)



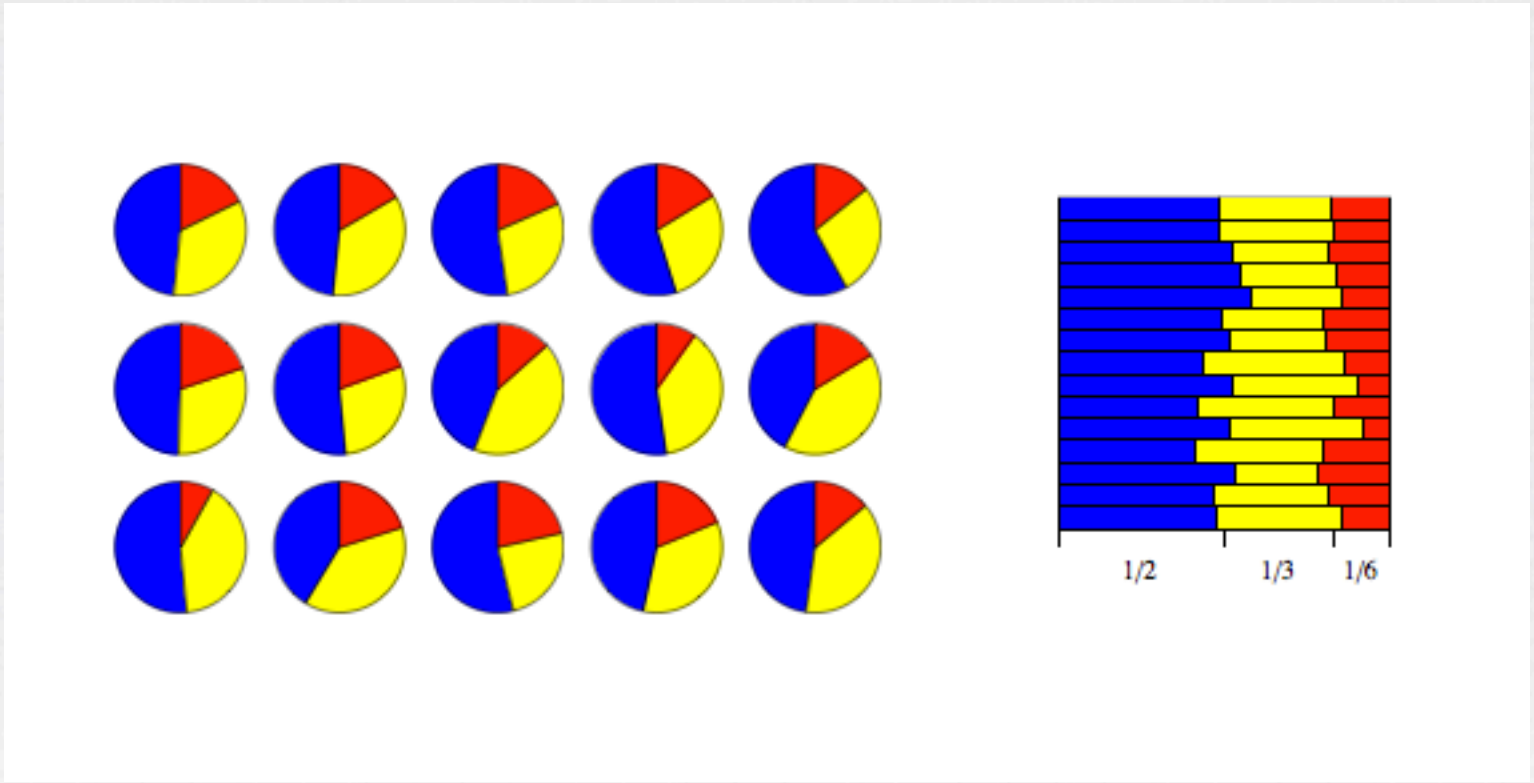
Topic2

Word-distribution of Topic2 = (1, 0, 0)

Dirichlet distribution

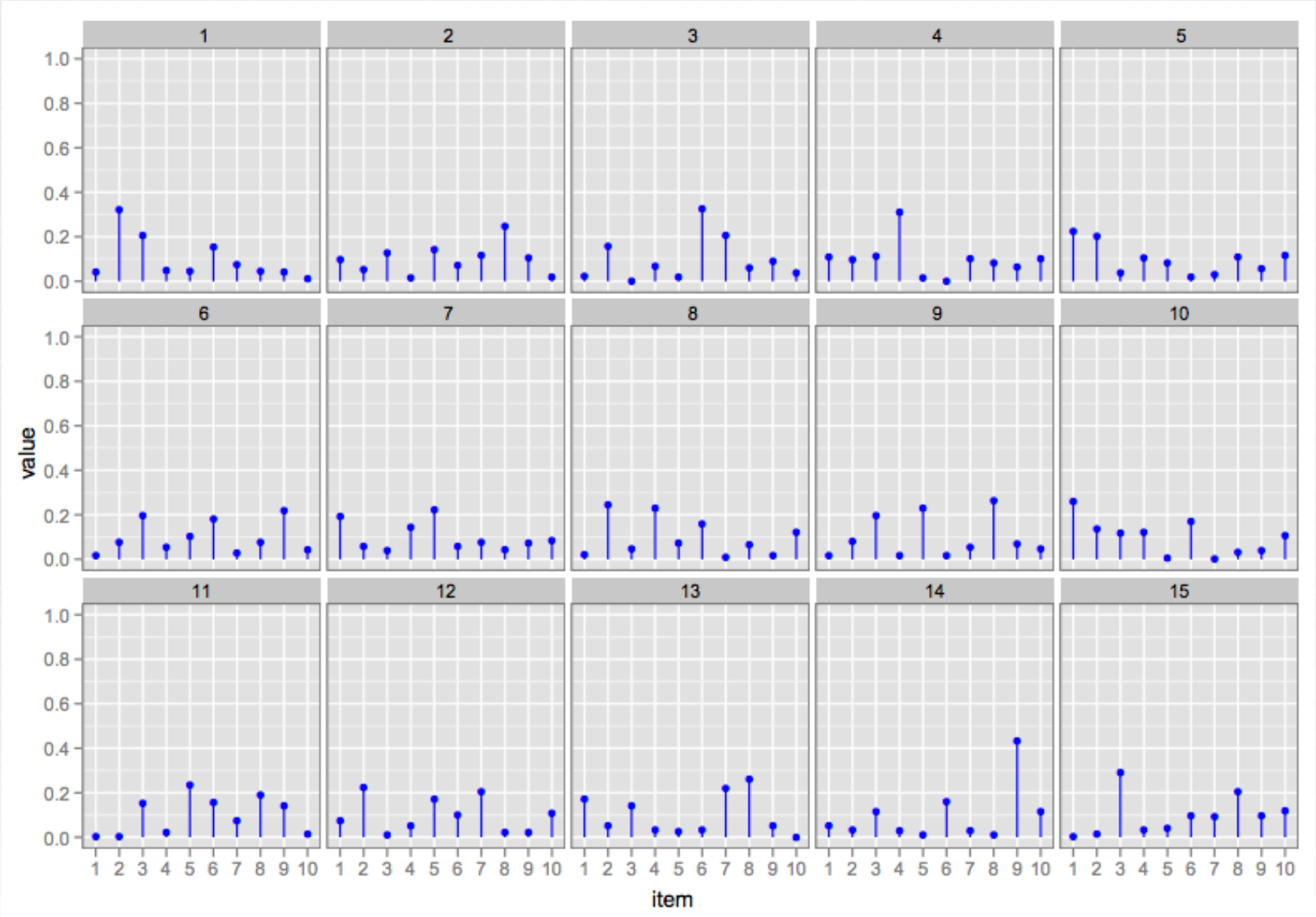
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Conjugate prior of multinomial distribution



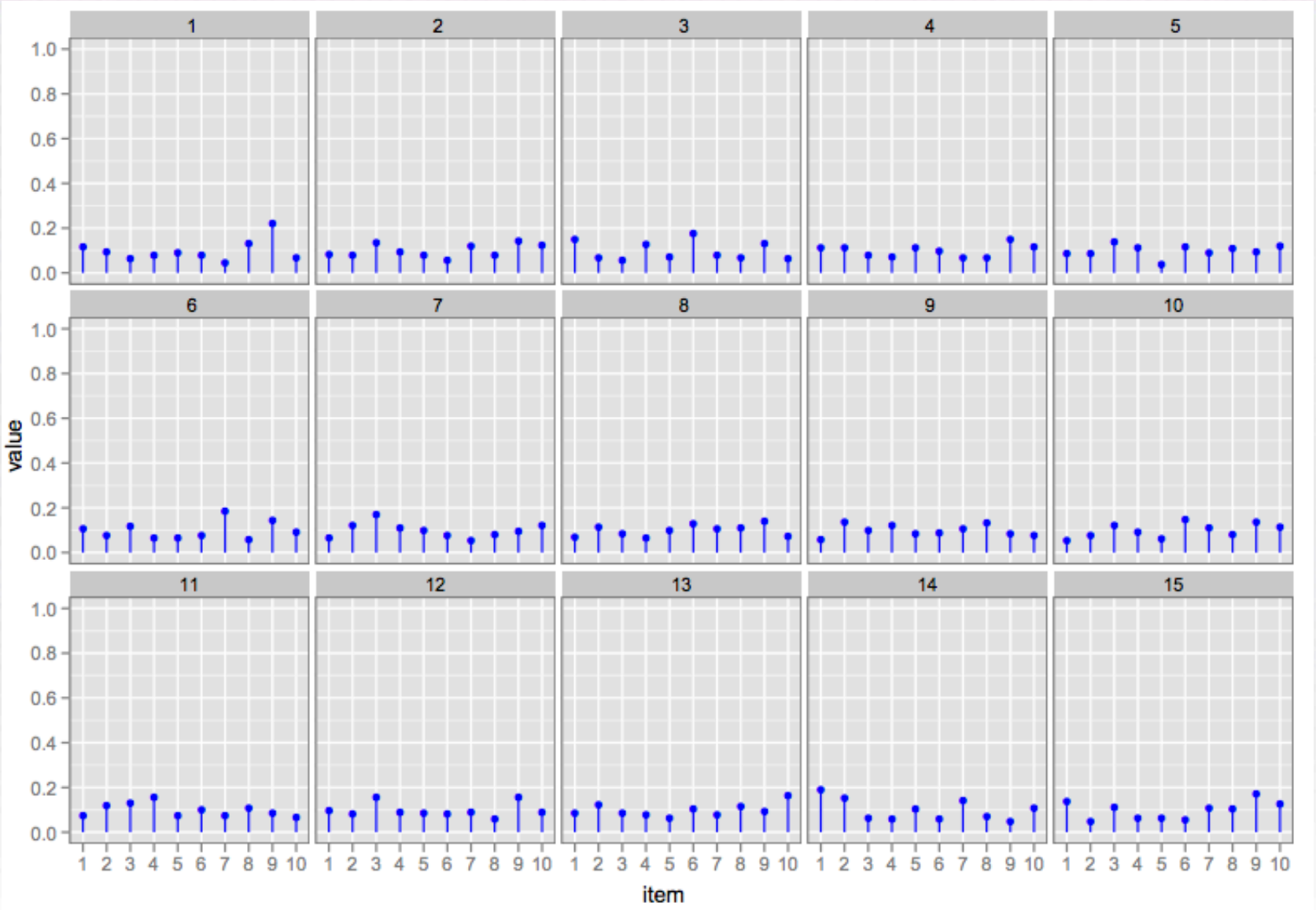
α=1

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



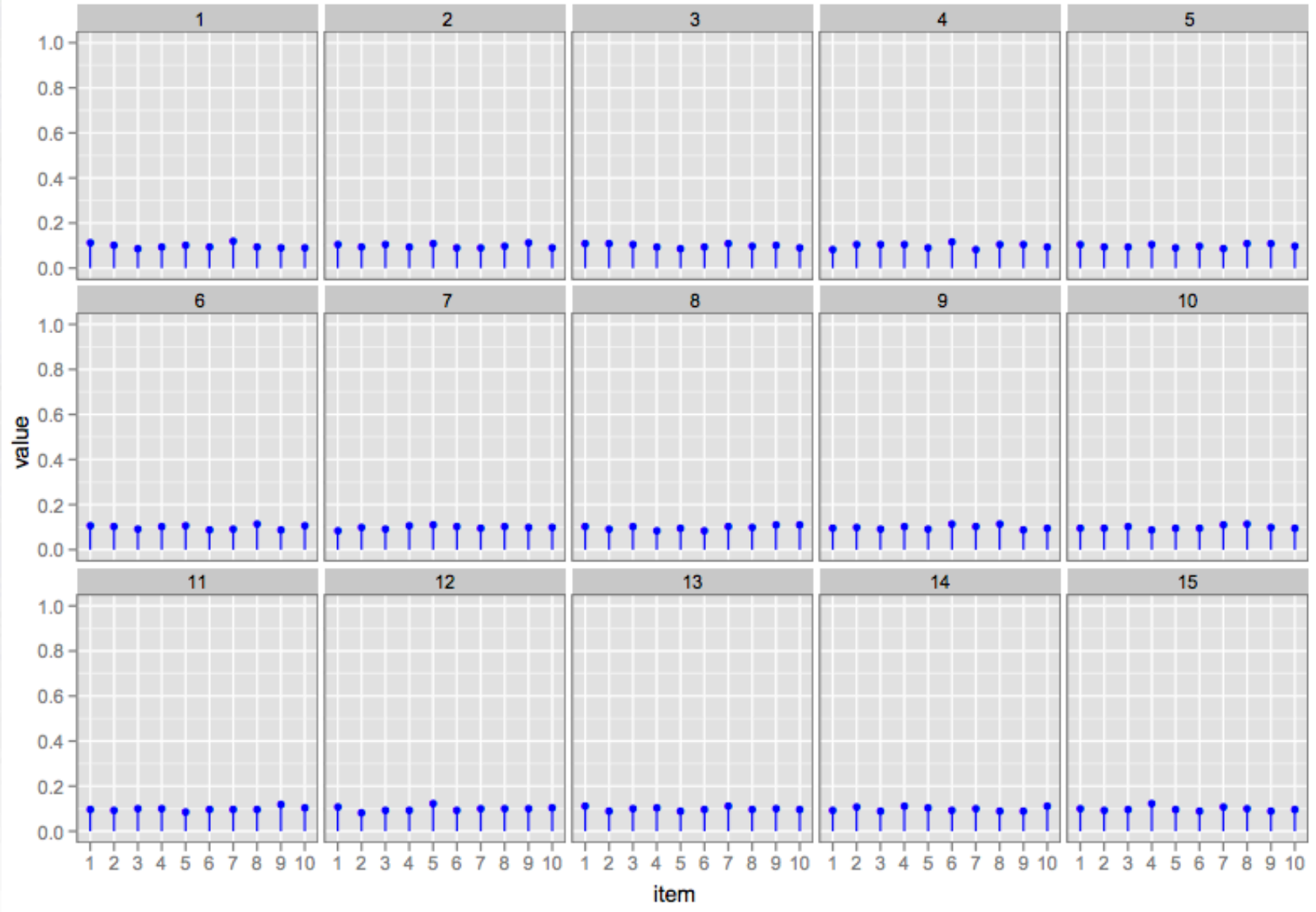
α=10

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



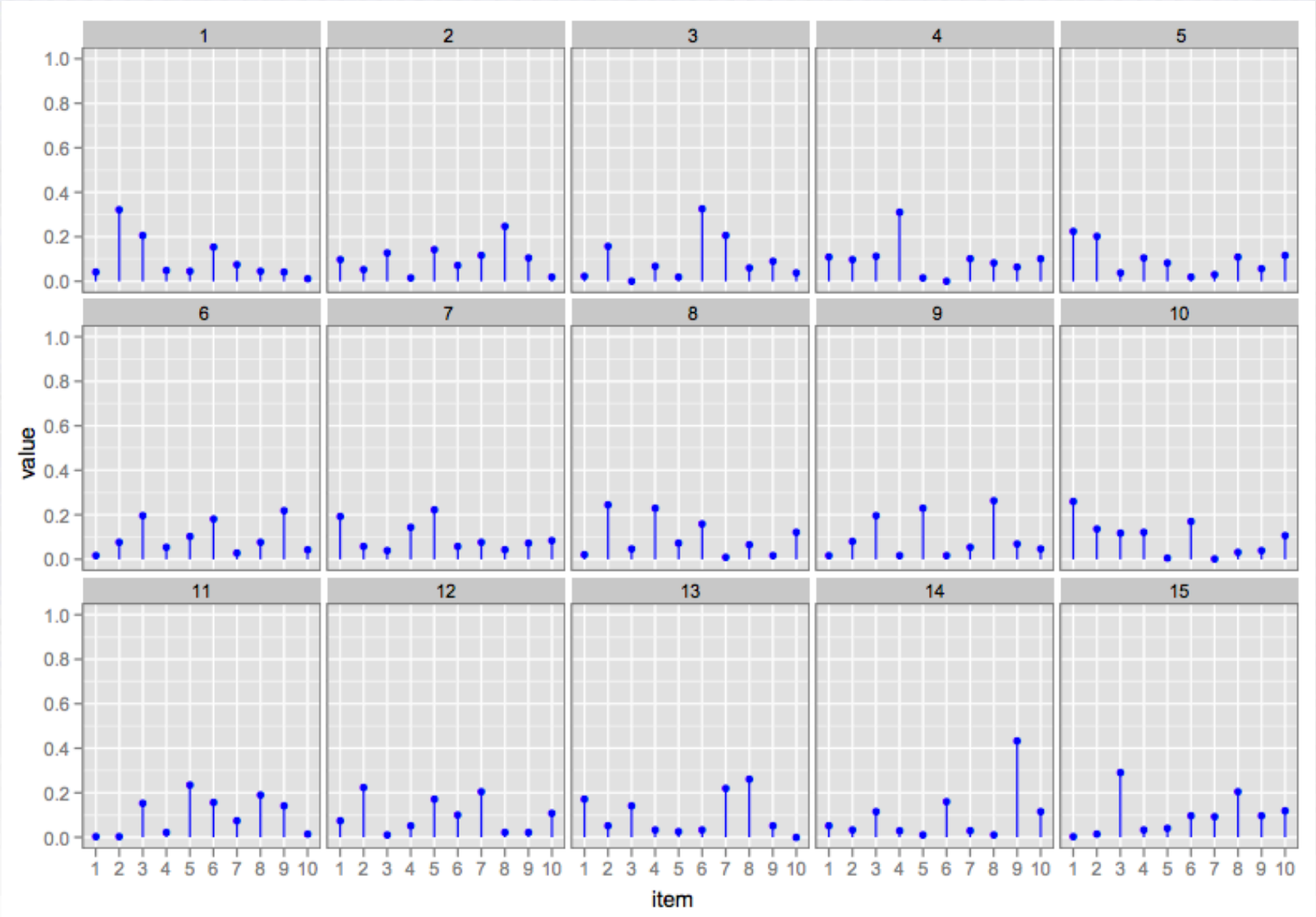
α=100

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



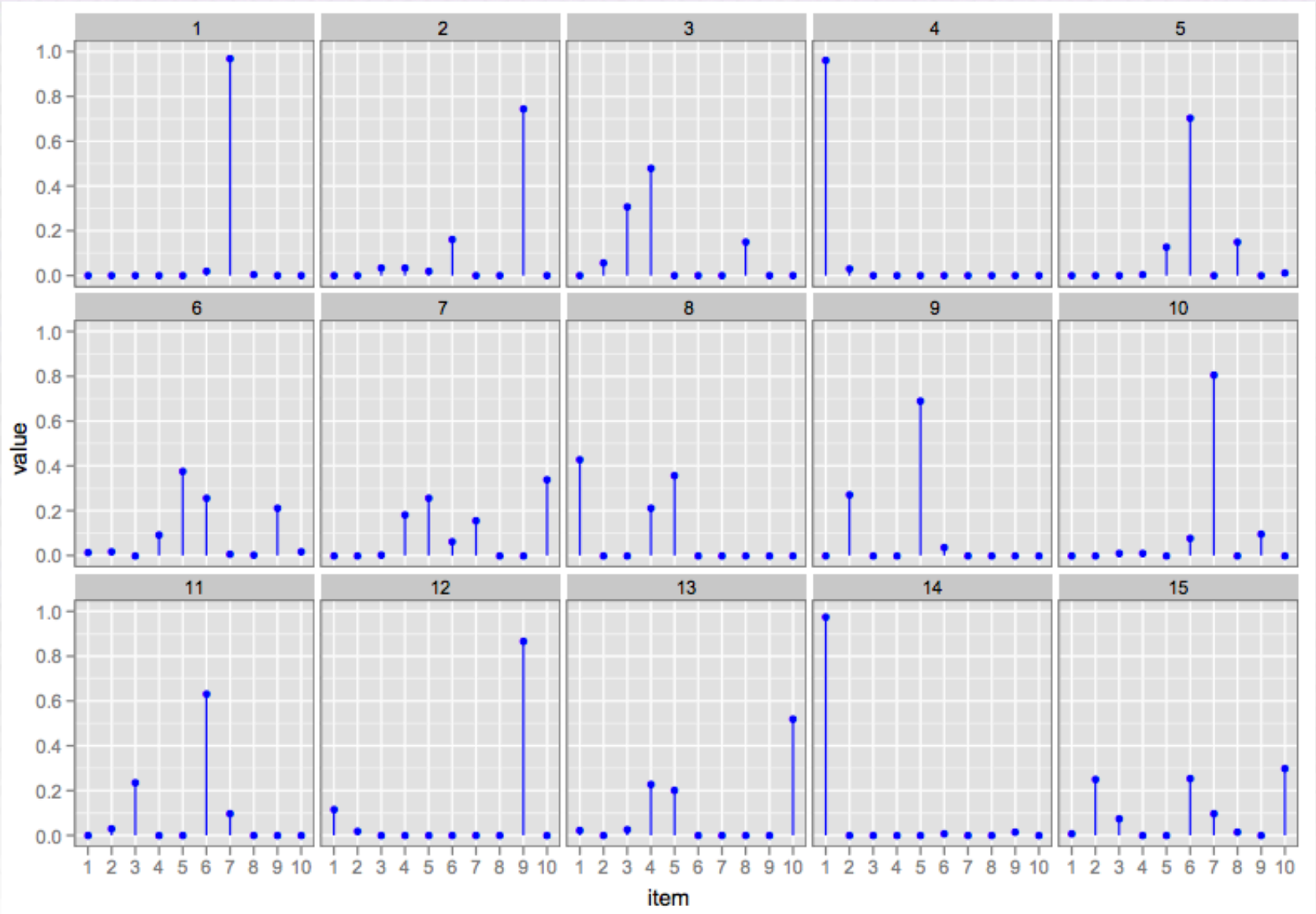
α=1

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



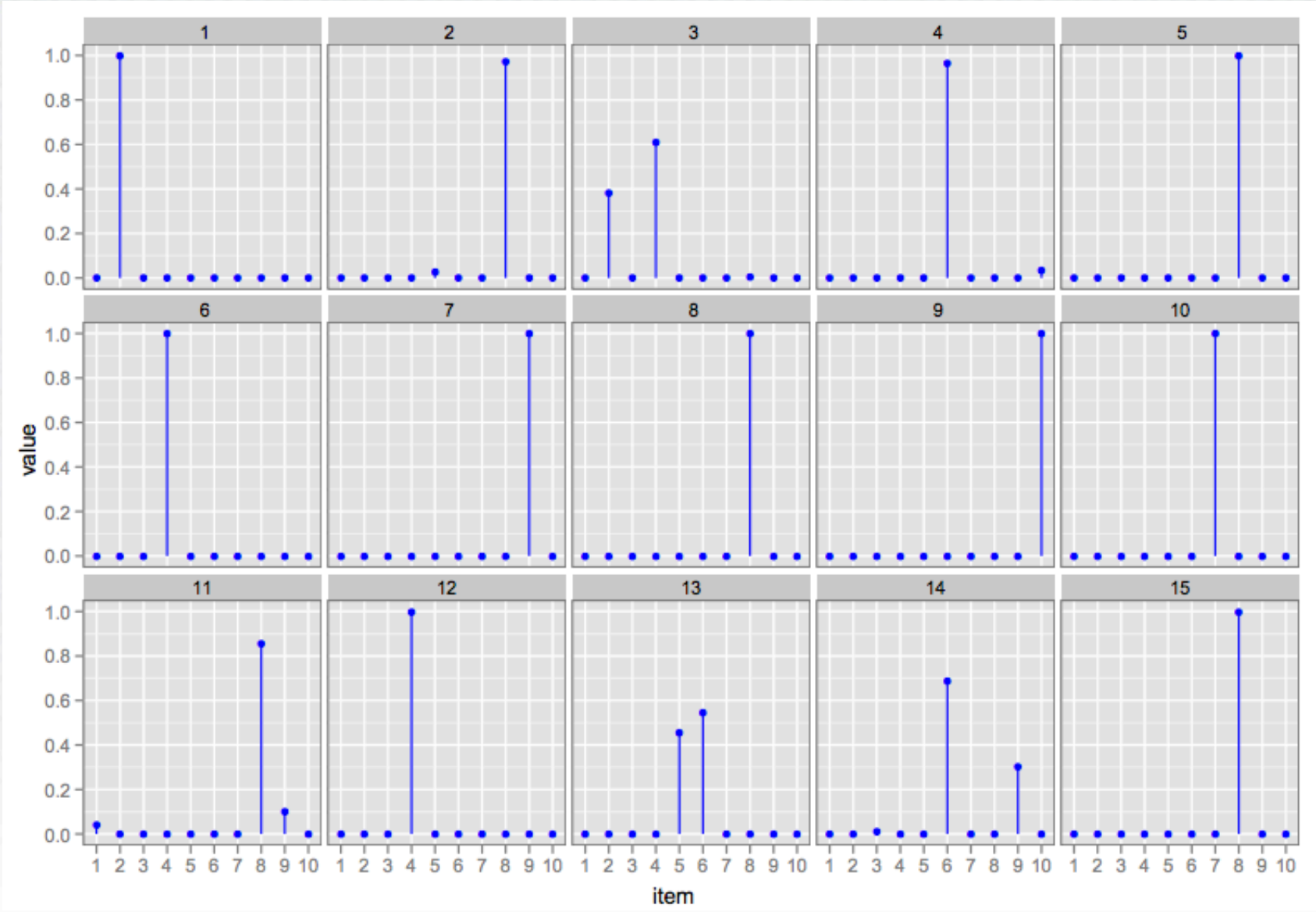
α=0.1

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



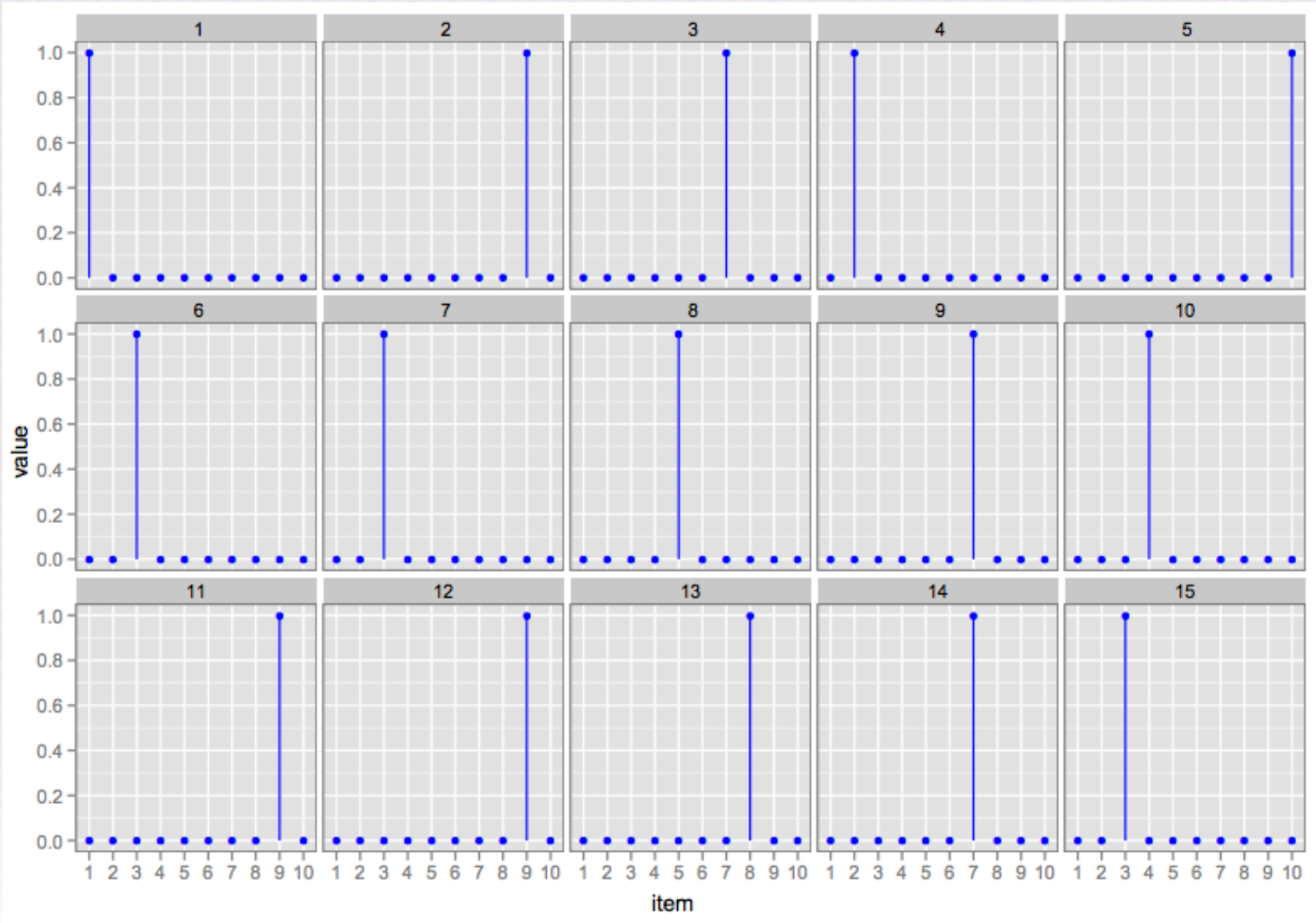
α=0.01

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



α=0.001

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



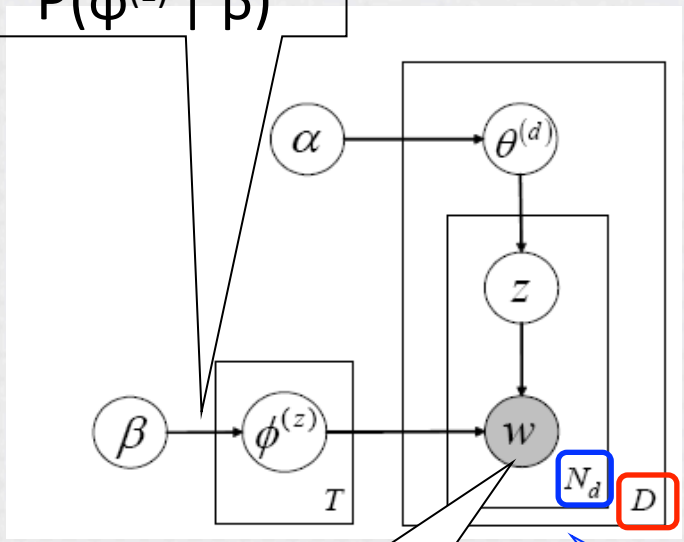
Why does "LDA" work?

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible.
  2. For each topic, assign high probability to as few terms as possible.
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

Posterior Inference

Latent Dirichlet Allocation (LDA) (Blei, 2003)



We learn topic distribution of a corpus → we can predict topic distribution of an unseen document of this corpus by observing its words

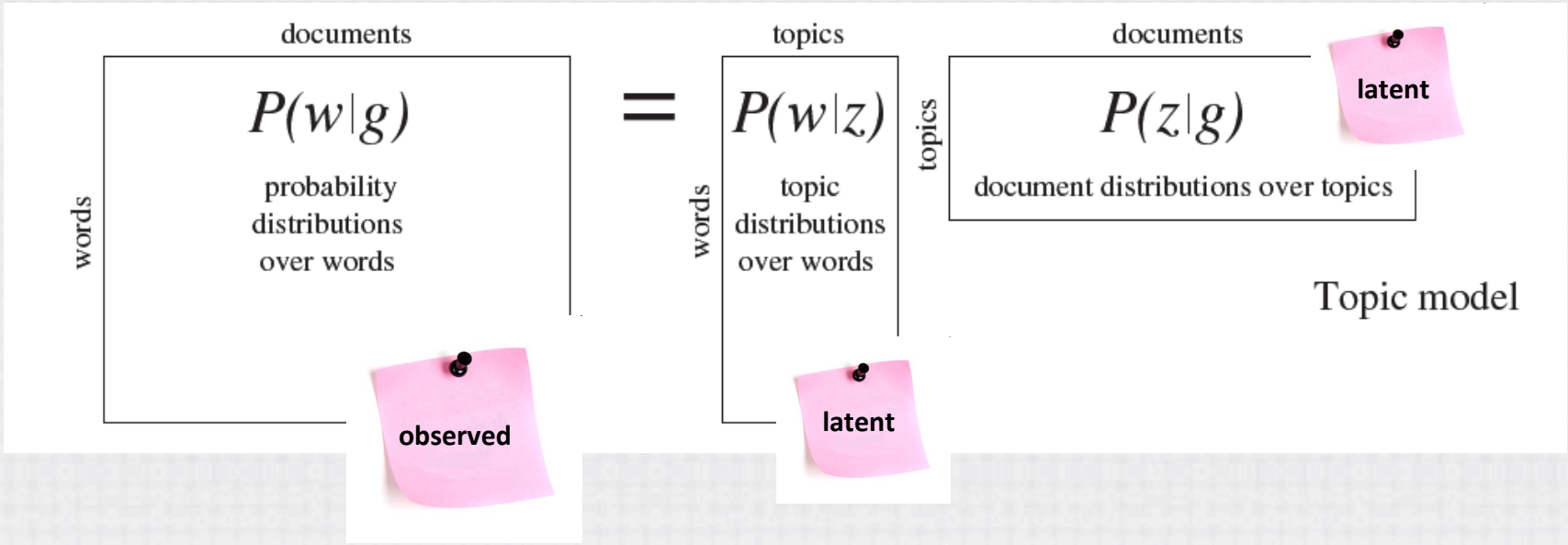
Hyper-parameters α and β are corpus-level parameters → are only sampled once

$$P(d,w) = P(d) * P(\theta^{(d)} \mid \alpha) * \sum_{z} P(\phi^{(z)} \mid \beta) * P(w \mid z, \phi^{(z)}) * P(z \mid \theta^{(d)})$$

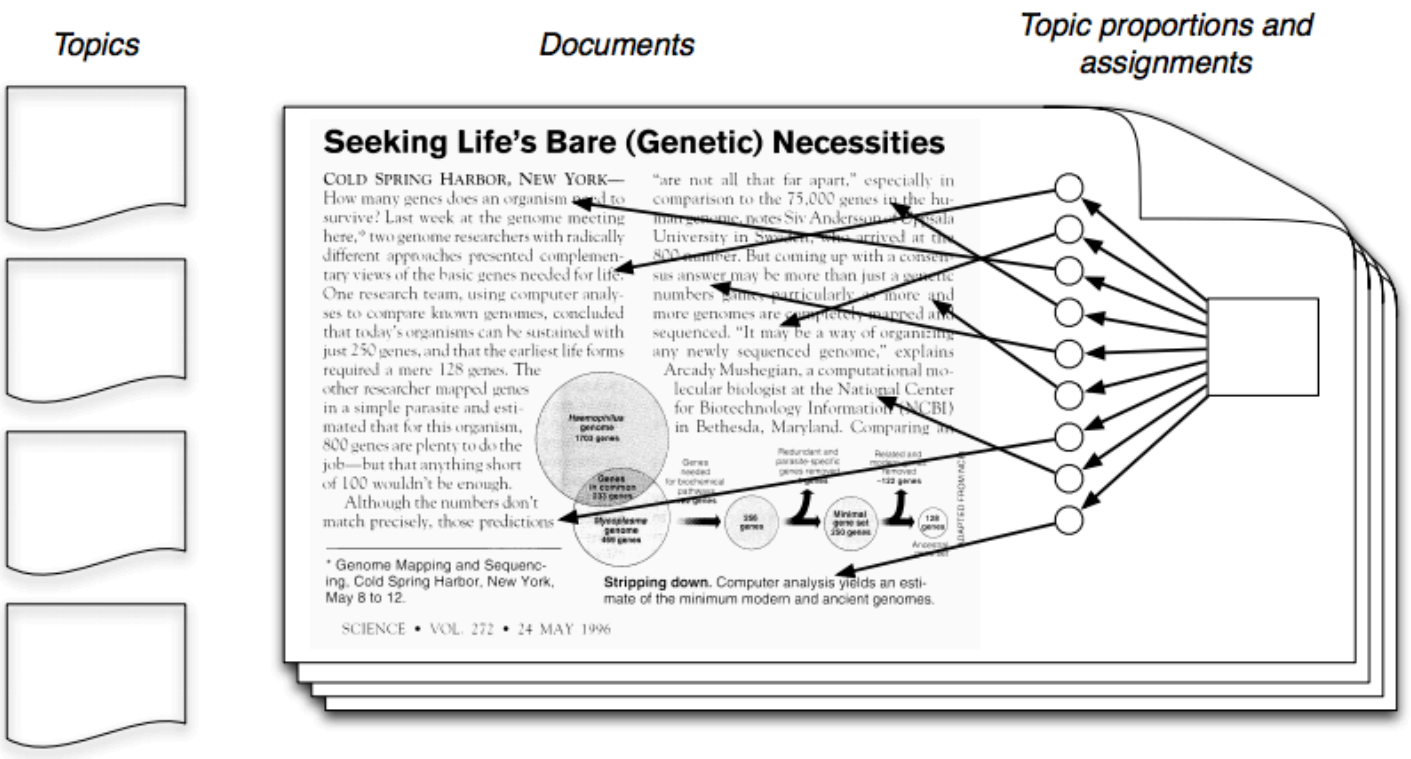
Matrix representation of LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

$\phi^{(z)}$ $\theta^{(d)}$



Posterior inference



- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Statistical inference and parameter estimation

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Key problem:

  Compute posterior distribution of the hidden variables given a document

$$p(\underbrace{\theta, \mathbf{z}}_{\text{Latent Variables}} \mid \underbrace{\mathbf{w}, \alpha, \beta}_{\text{Observed variables and Priors}}) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

(Blei, 2003)

- Posterior distribution is intractable for exact inference

Statistical inference and parameter estimation

- How can we estimate posterior distribution of hidden variables given a corpus of training-documents?
  - Direct (e.g. via expectation maximization, variational inference or expectation propagation algorithms)
  - Indirect → i.e. estimate the posterior distribution over *z* *(i.e. P(z))*
  - Gibbs sampling, a form of Markov chain Monte Carlo, is often used to estimate the posterior probability over a high-dimensional random variable ***z***

Gibbs sampling

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- generates a sequence of samples from the joint probability distribution of two or more random variables.

- Aim: compute posterior distribution over latent variable z
- Pre-request: we must know the conditional probability of z

$$P(z_i = j \mid z_{-i} , w_i , d_i , . .)$$

Why do we need to estimate P(z|w) via random walk?

z is a high-dimensional random variable

If num of topics T = 50 and num of words = 1000

We must visit 50^{1000} points and compute P(z) for all of them.

The collapsed Gibbs sampler

- Using conjugacy of Dirichlet and multinomial distributions, integrate out continuous parameters

$$P(\mathbf{z}) = \int_{\Delta_T^D} P(\mathbf{z} \mid \Theta) p(\Theta) d\Theta = \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(\alpha)^T} \frac{\Gamma(T\alpha)}{\Gamma(\sum_j n_j^{(d)} + \alpha)}$$

$$P(\mathbf{w} \mid \mathbf{z}) = \int_{\Delta_W^T} P(\mathbf{w} \mid \mathbf{z}, \Phi) p(\Phi) d\Phi = \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_w^{(j)} + \beta)}{\Gamma(\beta)^W} \frac{\Gamma(W\beta)}{\Gamma(\sum_w n_w^{(j)} + \beta)}$$

$$P(\mathbf{z} \mid \mathbf{w}) = \frac{P(\mathbf{w} \mid \mathbf{z}) P(\mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w} \mid \mathbf{z}) P(\mathbf{z})}$$

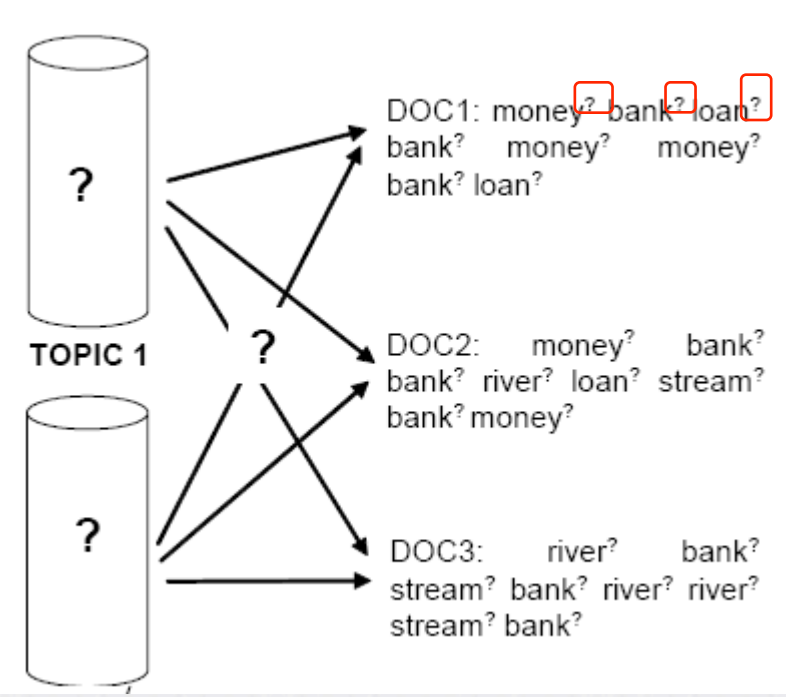
The collapsed Gibbs sampler

- Sample each z_i conditioned on $\mathbf{z}_{-i}$

$$P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_{\bullet}^{(z_i)} + W\beta} \frac{n_j^{(d_i)} + \alpha}{n_{\bullet}^{(d_i)} + T\alpha}$$

- This is nicer than your average Gibbs sampler:
  - memory: counts can be cached in two sparse matrices
  - optimization: no special functions, simple arithmetic
  - the distributions on Φ and Θ are analytic given $\mathbf{z}$ and $\mathbf{w}$, and can later be found for each sample

Gibbs sampling for LDA

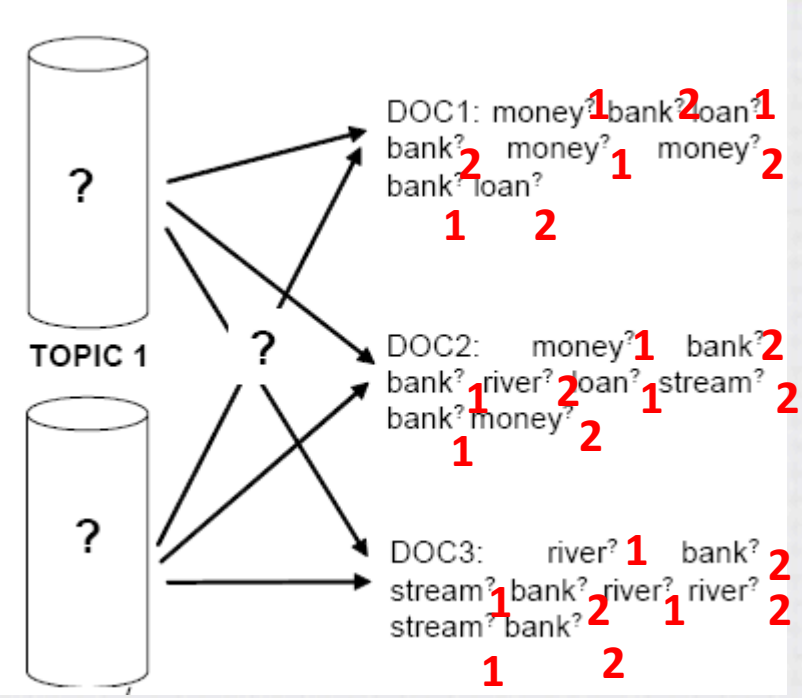


- Random start
- Iterative
- For each word we compute

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

- How dominante is a topic z in the doc d? How often was the topic z already used in doc d?
- How likely is a word for a topic z? How often was the word w already assigned to topic z?

Run Gibbs sampling example (1)



1. Random topic assignments
2. 2 count-matrices:

C^{WT} → Words per topic

C^{DT} → Topics per document

	topic1	topic2
money	3	2
bank	3	6
Loan	2	1
River	2	2
Stream	2	1

	doc1	doc2	doc3
topic1	4	4	4
topic2	4	4	4

Gibbs sampling for LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

Probability that topic ***j*** is chosen for word w_i, conditioned on all other assigned topics of words in this doc and all other observed vars.

Count number of times a word token w_i was assigned to a topic ***j*** across all docs

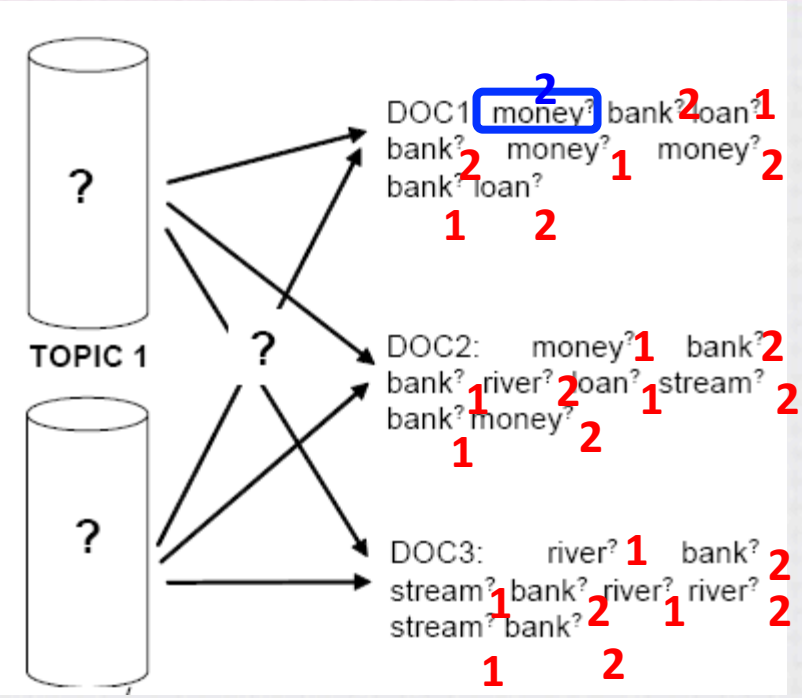
$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

unnormalized!

Count number of times a topic ***j*** was already assigned to some word token in doc d_i

=> divide the probability of assigning topic j to word w_i by the sum over all topics T

Run Gibbs sampling example (2)



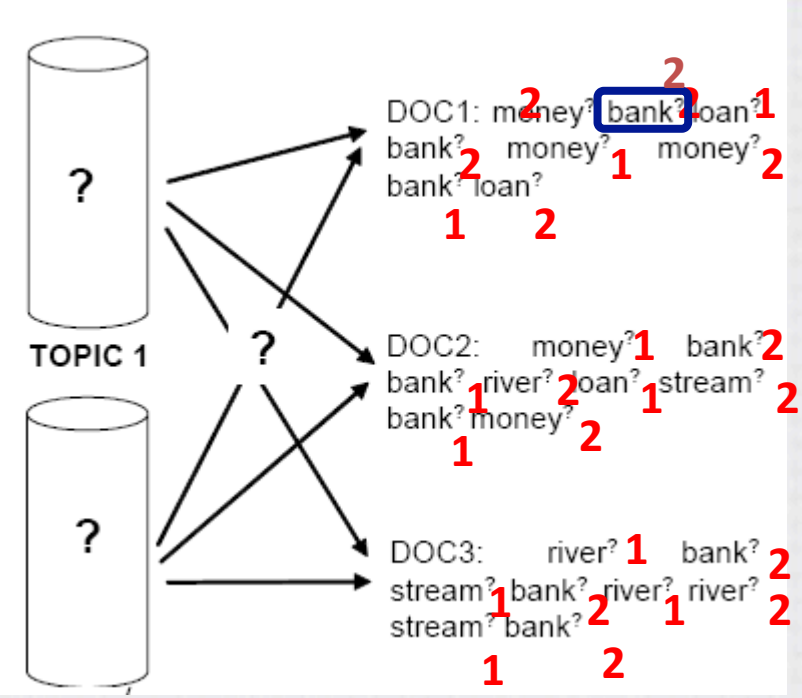
First Iteration:

- Sample new topic from the current topic-distribution of a doc

	topic1	topic2
money	3 2	2 3
bank	3	6
Loan	2	1
River	2	2
Stream	2	1

	doc1	doc2	doc3
topic1	4 3	4	4
topic2	4 5	4	4

Run Gibbs sampling example (2)



First Iteration:

- Sample new topic from the current topic-distribution of a doc

	topic1	topic2
money	2	3
bank	3	6 5 6
Loan	2	1
River	2	2
Stream	2	1

	doc1	doc2	doc3
topic1	3	4	4
topic2	5 4 5	4	4

Run Gibbs sampling example (3)

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

How often was topic j used in doc d_i

How often were all other topics used in doc d_i

α = 50/T = 25 and β = 0.01

$$P(z_i = topic1 \mid z_{-i}, bank, d_i, .) = \frac{3 + 0.01}{8 + 5 * 0.01} * \frac{3 + 25}{4 + 2 * 25} = 0.19$$

$$P(z_i = topic2 \mid z_{-i}, bank, d_i, .) = \frac{5 + 0.01}{7 + 5 * 0.01} * \frac{4 + 25}{3 + 2 * 25} = 0.39$$

"Bank" is assigned to Topic 2

Gibbs sampling convergence

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Random Start
- N iterations
- Each iteration updates count-matrices

Convergence:

- count-matrices stop changing

Gibbs samples start to approximate the target distribution (i.e., the posterior distribution over z)

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration 1
i	w_i	d_i	z_i
1	MATHEMATICS	1	2
2	KNOWLEDGE	1	2
3	RESEARCH	1	1
4	WORK	1	2
5	MATHEMATICS	1	1
6	RESEARCH	1	2
7	WORK	1	2
8	SCIENTIFIC	1	1
9	MATHEMATICS	1	2
10	WORK	1	1
11	SCIENTIFIC	2	1
12	KNOWLEDGE	2	1
.	.	.	.
.	.	.	.
.	.	.	.
50	JOY	5	2

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	?
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	?
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	?
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	2
5	MATHEMATICS	1	1	?
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

			iteration			
			1	2	…	1000
i	w_i	d_i	z_i	z_i		z_i
1	MATHEMATICS	1	2	2		2
2	KNOWLEDGE	1	2	1		2
3	RESEARCH	1	1	1		2
4	WORK	1	2	2		1
5	MATHEMATICS	1	1	2		2
6	RESEARCH	1	2	2		2
7	WORK	1	2	2		2
8	SCIENTIFIC	1	1	1	…	1
9	MATHEMATICS	1	2	2		2
10	WORK	1	1	2		2
11	SCIENTIFIC	2	1	1		2
12	KNOWLEDGE	2	1	2		2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
50	JOY	5	2	1		1

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Parameter Estimation

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}.$$

Effects of hyperparameters

- α and β control the relative sparsity of Φ and Θ
  - smaller α, fewer topics per document
  - smaller β, fewer words per topic

Evaluation

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Compute the *perplexity* of a held-out test to evaluate the models – lower perplexity score indicates better generalization.
- Perplexity is to calculate the likelihood. It is defined as the reciprocal geometric mean of the token likelihoods in the test corpus given the model:

Trained model

Number of docs in the test set

$$p(\vec{W}|M) = \exp - \frac{\sum_{m=1}^{N_{test}} \log p(\tilde{\vec{w}}_{\tilde{m}}|M)}{\sum_{m=1}^{N_{test}} N_m}$$

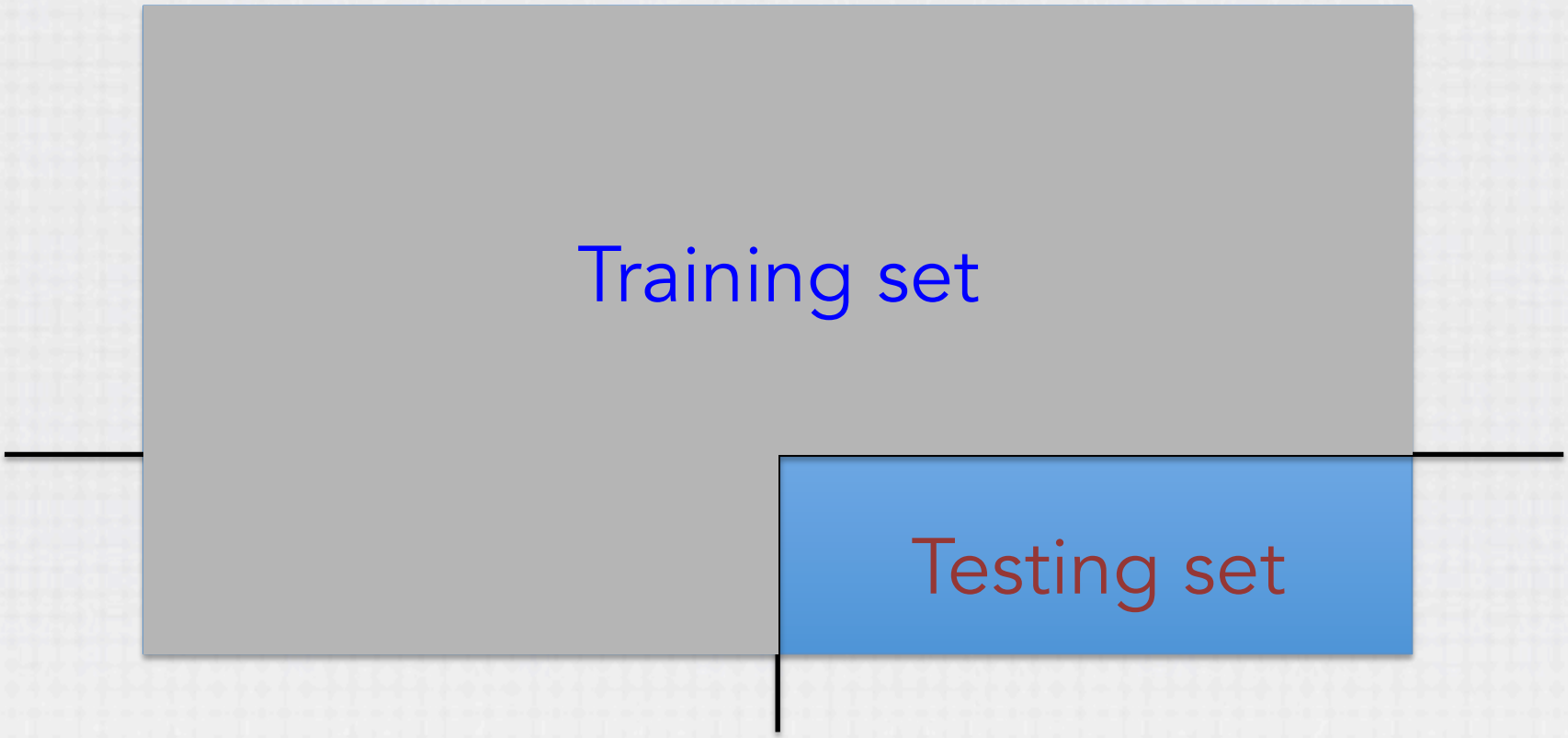
Number of words in the m^{th} doc

Log likelihood

$$\log p(\tilde{\vec{w}}_{\tilde{m}}|M) = \sum_{t=1}^{V} n_{\tilde{m}}^{(t)} \log(\sum_{k=1}^{K} \phi_{k,l} \cdot \theta_{\tilde{m},k})$$

- Where $n_{\tilde{m}}^{(t)}$ is the number of times word t occurs in document $\tilde{m}$.

Evaluation



Implementation of LDA

There are many available implementations of topic modeling.
Here is an incomplete list—

LDA-C*	A C implementation of LDA
HDP*	A C implementation of the HDP ("infinite LDA")
Online LDA*	A python package for LDA on massive data
LDA in R*	Package in R for many topic models
LingPipe	Java toolkit for NLP and computational linguistics
Mallet	Java toolkit for statistical NLP
TMVE*	A python package to build browsers from topic models

* available at www.cs.princeton.edu/~blei/