

# Lecture 7: Topic Modeling

Pilsung Kang

School of Industrial Management Engineering  
Korea University

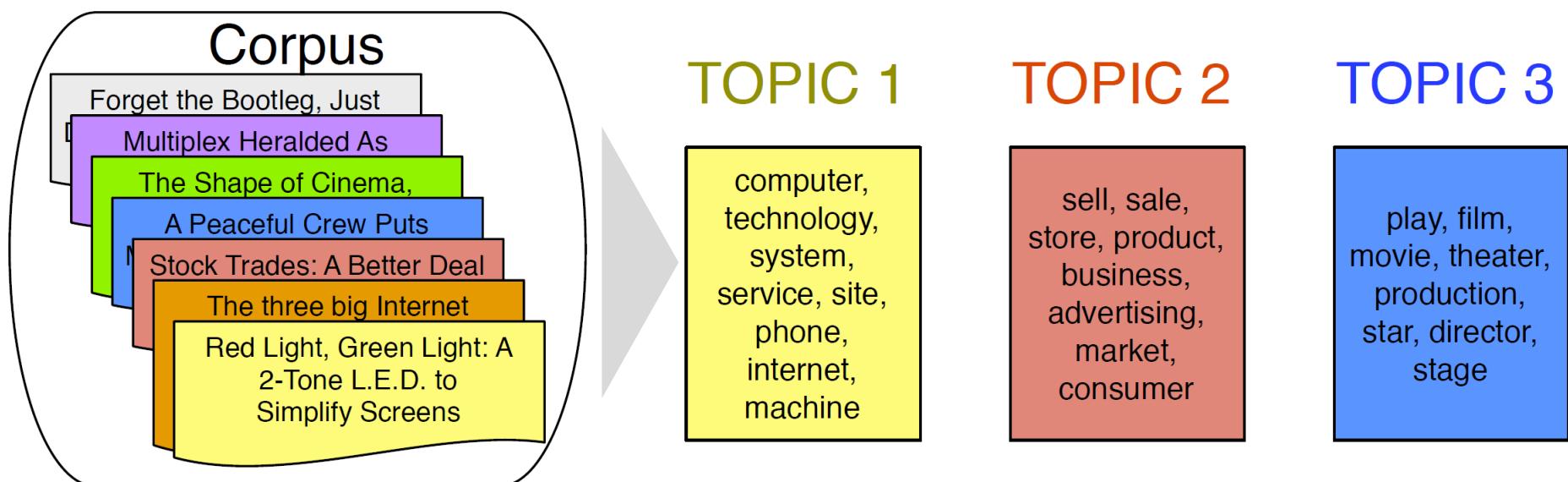
# AGENDA

- 01 Topic Modeling
- 02 Probabilistic Latent Semantic Analysis
- 03 LDA: Document Generation Process
- 04 LDA Inference: Gibbs Sampling
- 05 LDA Evaluation

# Topic Model: Conceptual Approach

- Topic Model

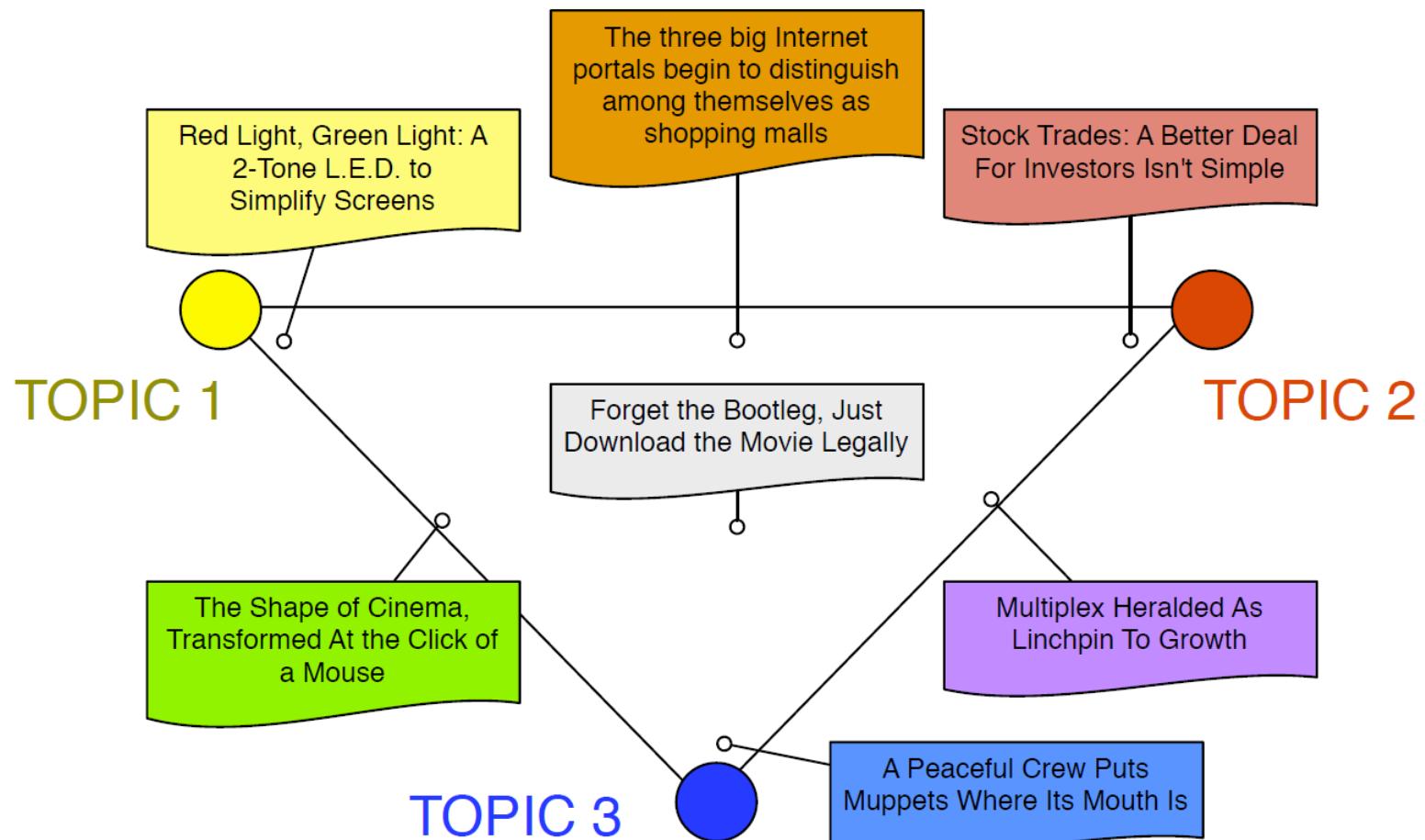
- ✓ From **an input corpus** and the number of topics K → words to topics
- ✓ From an input corpus and the number of topics K → **words to topics**



# Topic Model: Conceptual Approach

- Topic Model

- ✓ For each document, what topics are expressed by that document?

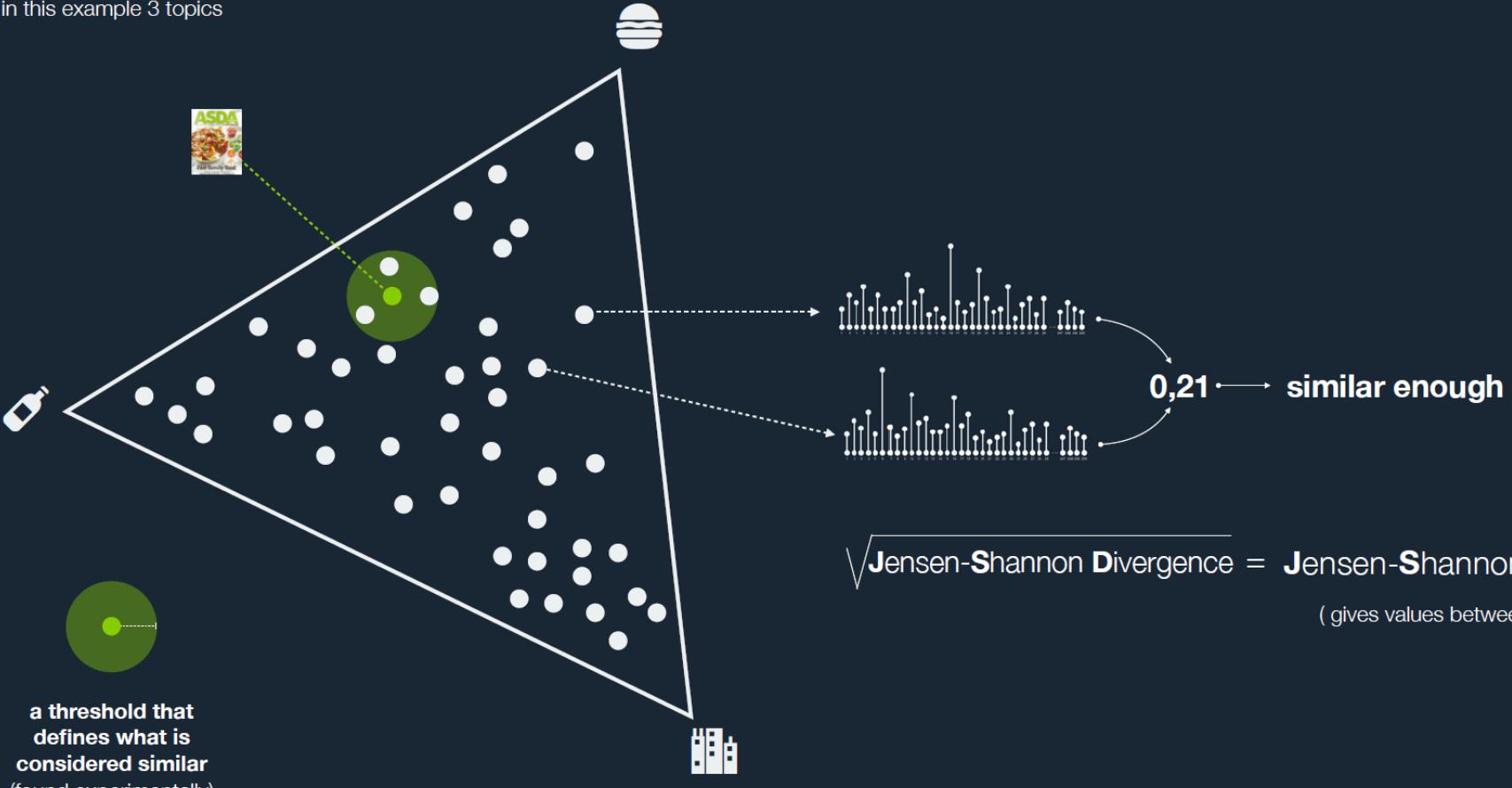


# Topic Model: Conceptual Approach

Knispelis (2015)

LDA space  
a simplex

in this example 3 topics



# Topic Models: Topic Extraction

Kim et al. (2016)

- Topic Extraction

✓ 30 Topics discovered for “Deep Learning”

Fault detection with DBN	Convolutional neural network	Network Learning	Representation learning	Face Recognition	Speech Recognition	Acoustic Modeling	Extreme Learning	Deep learning architecture	Image Segmentation
deep belief network dbn fault	neural convolutional pool convolution convnet	layer input output unit hide function	feature level extract learn extraction	face recognition estimation facial shape	speaker speech noise adaptation source	speech recognition acoustic hmm neural	deep learn algorithm structure extreme	deep architecture neural standard explore	image scene scale segmentation pixel
Long-short term memory	Predictive analytics	Signal processing	Classification models	Large-scale computing	Image quality assessment	Visual recognition	NLP	Detection using CNN	Action recognition
term recurrent long lstm network	data prediction technique information research	analysis filter signal component audio	classification classifier class vector support	application implementation efficient process power	domain state quality resolution relationship	pattern process compute visual field	word text language representation semantic	cnn detection convolutional neural detect	video human temporal action track
Image retrieval	Medical image diagnosis	Reinforcement learning	Parameter optimization	Auto encoder	RBM and variations	Learning with few labeled data	Fast learning complexity reduction	Applications for vehicles & robots	Character recognition
image visual retrieval descriptor attribute	image segmentation disease cell medical	learn question state answer reinforcement	train algorithm gradient sample optimization	representation learn sparse encode stack	machine boltzmann rbm restrict distribution	train data label few transfer	fast reduce parameter weight complexity	time real application drive Vehicle	recognition system character network neural

# Topic Models: Topic Extraction

- Topic Extraction

✓ 50 Topics discovered for “Ultrasound/Ultrasonography”

Vascular	Prostate	heart	CAD	MSK	nerve	tumor	OB	surgery	intervention
plaque	biopsy	artery	image	joint	block	case	ultrasound	surgery	guide
ivus	prostate	carotid	ultrasound	patient	nerve	lesion	fetal	patient	patient
coronary	cancer	patient	method	disease	ultrasound	diagnosis	infant	intraoperative	complication
intravascular	patient	stenosis	base	score	guide	ultrasound	abnormality	preoperative	treatment
stent	transrectal	plaque	propose	arthritis	patient	cyst	prenatal	surgical	percutaneous
patient	trus	ultrasound	feature	ultrasound	pain	mass	case	ultrasound	ultrasound
lesion	guide	cardiac	algorithm	clinical	anesthesia	tumor	fetus	localization	drainage
mm.	core	dus	segmentation	inflammatory	surgery	finding	anomaly	operative	month
ultrasound	ultrasound	stroke	analysis	activity	plexus	ultrasonography	diagnosis	resection	rate
area	rate	arterial	result	study	technique	present	congenital	surgeon	procedure

osteoporosis	cerebral	ER&ICU	cancer	Lab test	US general	vein	lymph node	lung	Healthcare
age	brain	patient	cancer	extraction	ultrasound	vein	node	lung	patient
ultrasound	dog	emergency	patient	assist	imaging	venous	lymph	chest	risk
child	fus	care	tumor	ultrasound	technique	patient	patient	ultrasound	ultrasound
bone	bbb	ultrasound	stage	method	clinical	internal	biopsy	patient	year
year	ultrasound	department	eus	liquid	review	ultrasound	metastasis	pulmonary	study
study	blood	bedside	gastric	sample	application	jugular	ultrasound	lus	follow
fat	study	perform	ovarian	time	diagnostic	thrombosis	cancer	pleural	clinical
qus	day	physician	endoscopic	solvent	disease	central	guide	line	factor
body	follicle	point	ultrasonography	determination	article	dvt	negative	radiography	month
measure	barrier	cardiac	invasion	extract	role	femoral	positive	diagnosis	age

# Topic Models: Topic Extraction

- Topic Extraction

✓ 10 Topics discovered for “Insider Threat”

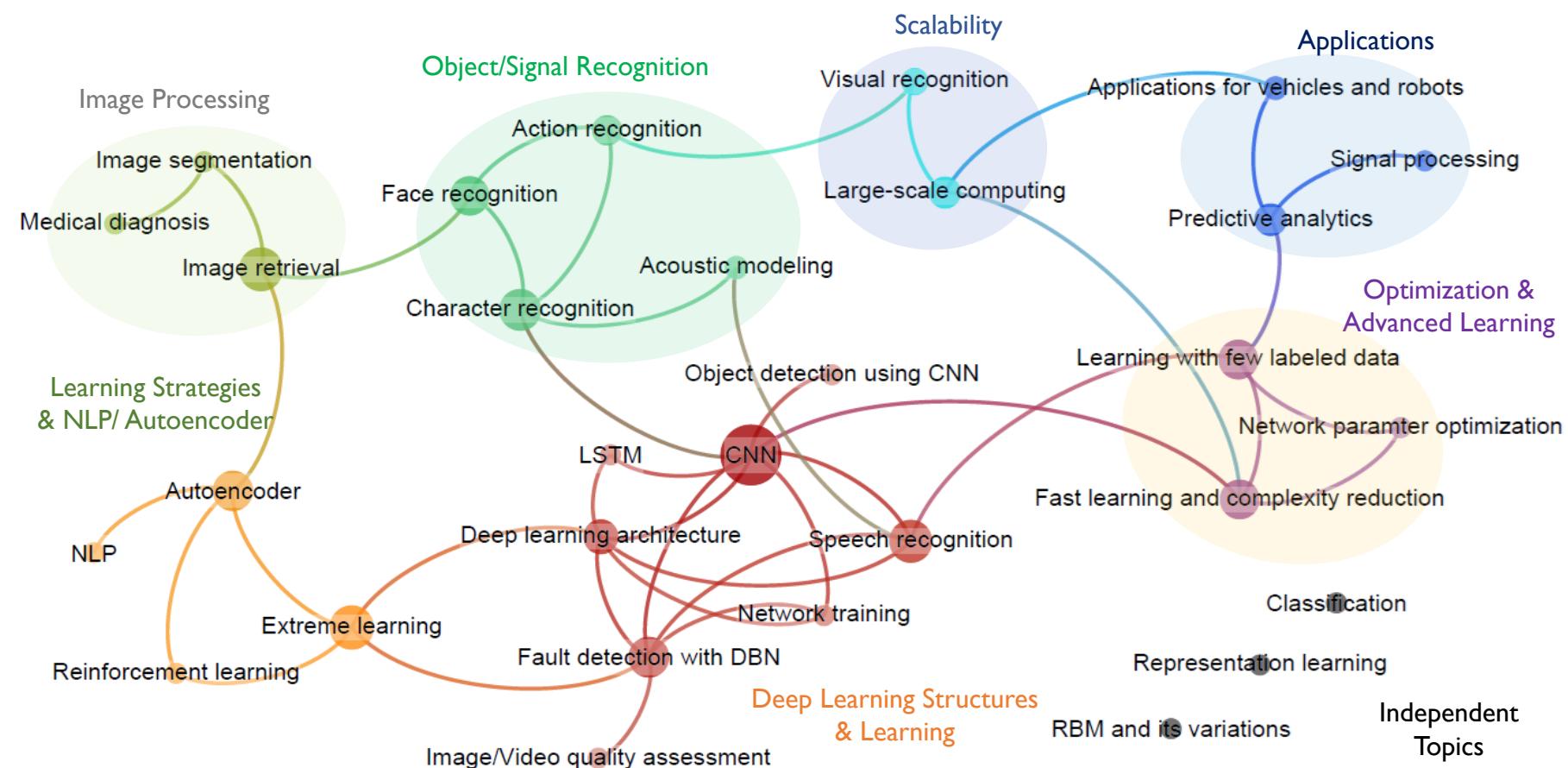
No.	Insider threat in relational database	Assessment of insider threat	Insider attacks on Communication protocol	Modeling and system framework for insider threat	Masquerade detection
1	data	measure	attack	insider	user
2	information	assess	agent	threat	behavior
3	database	security	scheme	social	detect
4	leakage	behavior	protocol	analysis	activity
5	access	analysis	monitor	framework	malicious
6	detect	management	mitigation	mitigate	masquerade
7	transaction	privacy	fraud	monitor	attack
8	confidential	policy	damage	factor	legitimate
9	document	risk	psychological	technical	abnormal
10	file	threat	financial	business	decoy

No.	Access control for insider threat mitigation	Network intrusion detection systems	Feature selection for intrusion detection	Miscellaneous	Malicious domain detection
1	insider	network	detection	software	attack
2	access	detection	algorithm	security	malicious
3	user	intrusion	feature	system	domain
4	control	malicious	classification	device	event
5	cloud	traffic	accuracy	server	scenario
6	misuse	log	dataset	malicious	human
7	trust	event	performance	protect	knowledge
8	risk	packet	pattern	web	ontology
9	abuse	internet	learning	architecture	represent
10	attacker	resource	random	electronic	generate

# Topic Models: Relation between Topics

Kim et al. (2016)

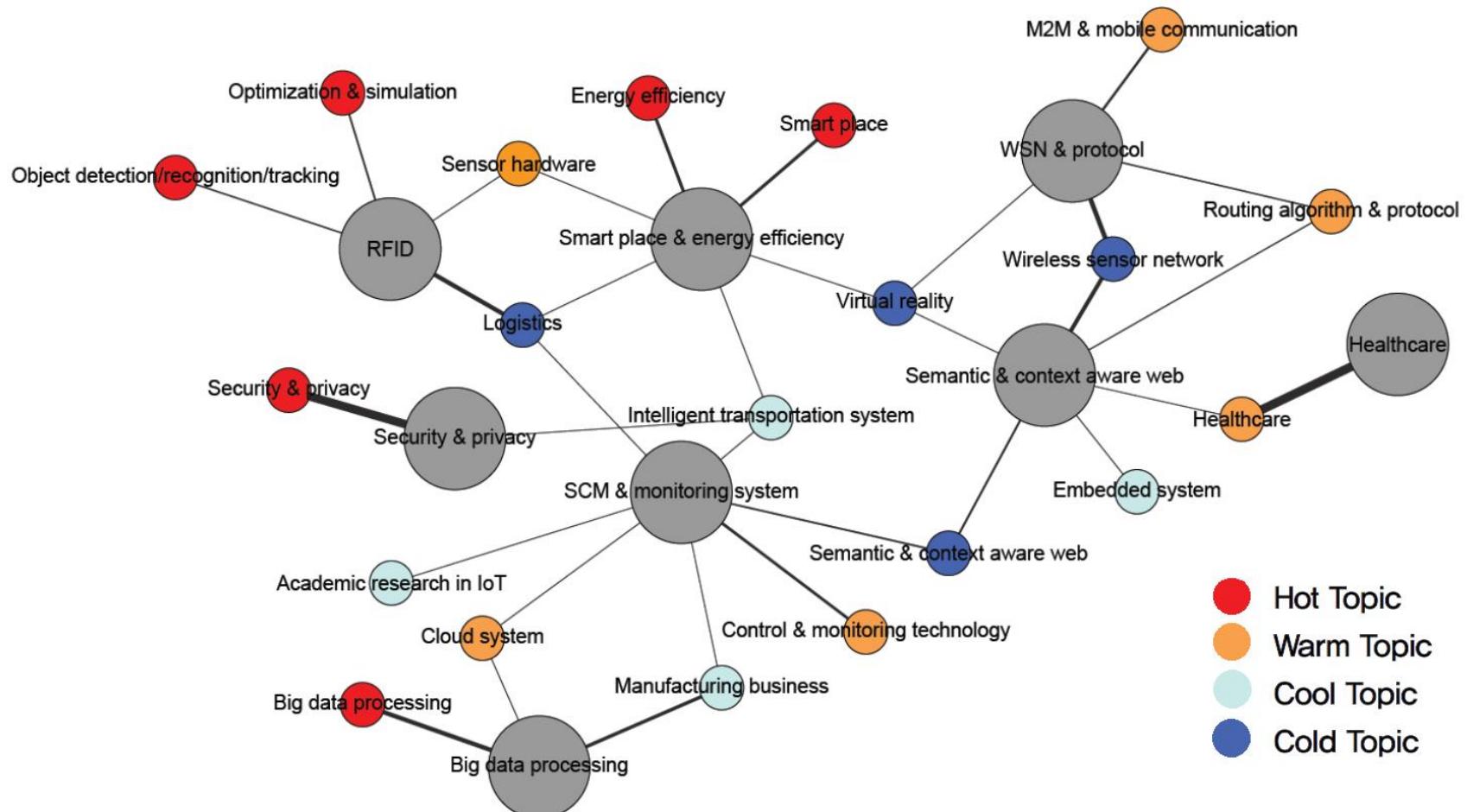
- Relation between Topics: Deep Learning



# Topic Models: Relation between Topics

Kim and Kang (2018+)

- Relation between Topics: Internet of Things

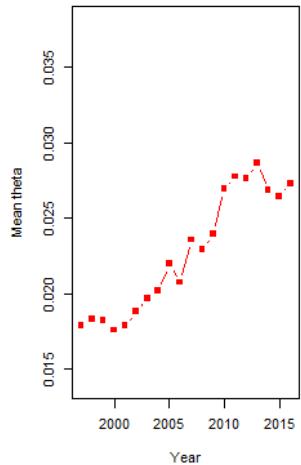


# Topic Models: Trend Analysis

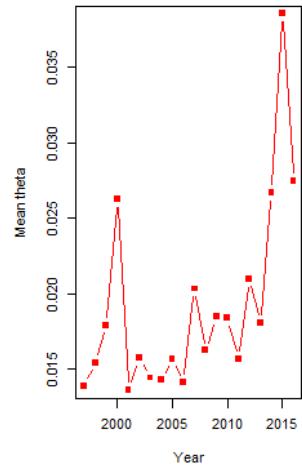
Lee and Kang (2017)

- Topic trends for “technology and innovation management”

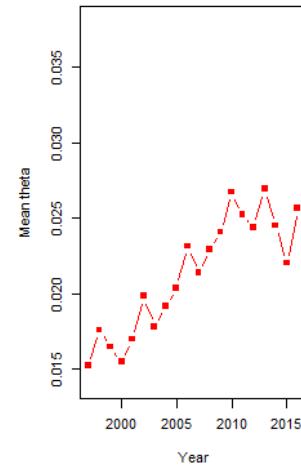
[T2] Empirical study on innovation and firm performance



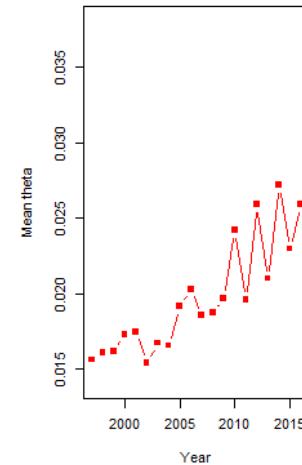
[T12] Energy policy and sustainability



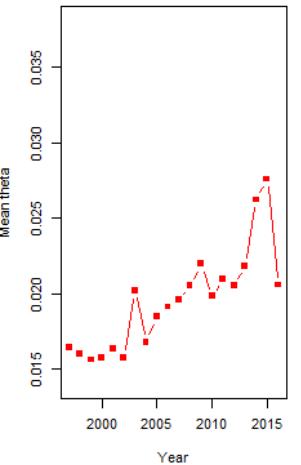
[T5] Open innovation



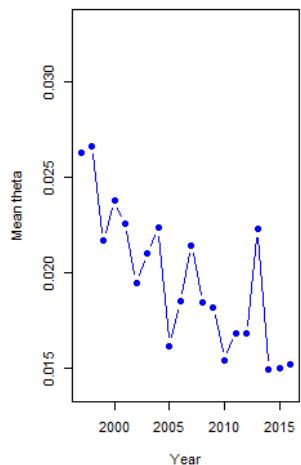
[T11] Technological transitions



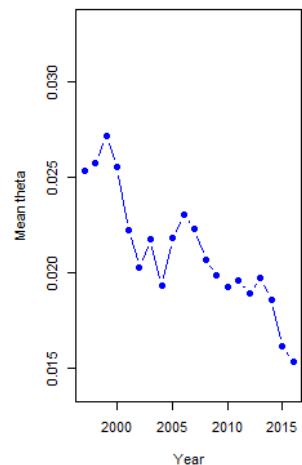
[T13] Consumer behavior and satisfaction



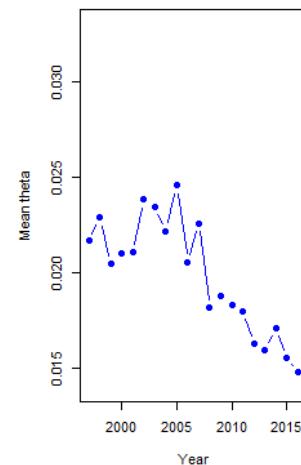
[T39] Total quality management and continuous improvement



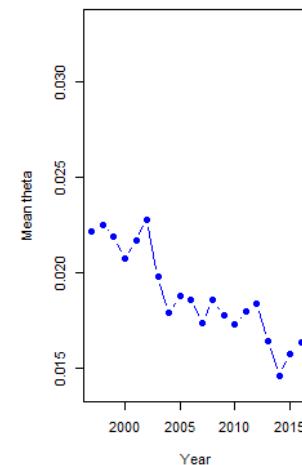
[T18] Organizational culture and organizational innovation



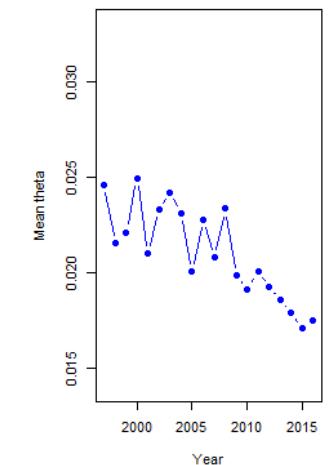
[T35] Innovation in high-tech industries



[T46] Innovation in SMEs

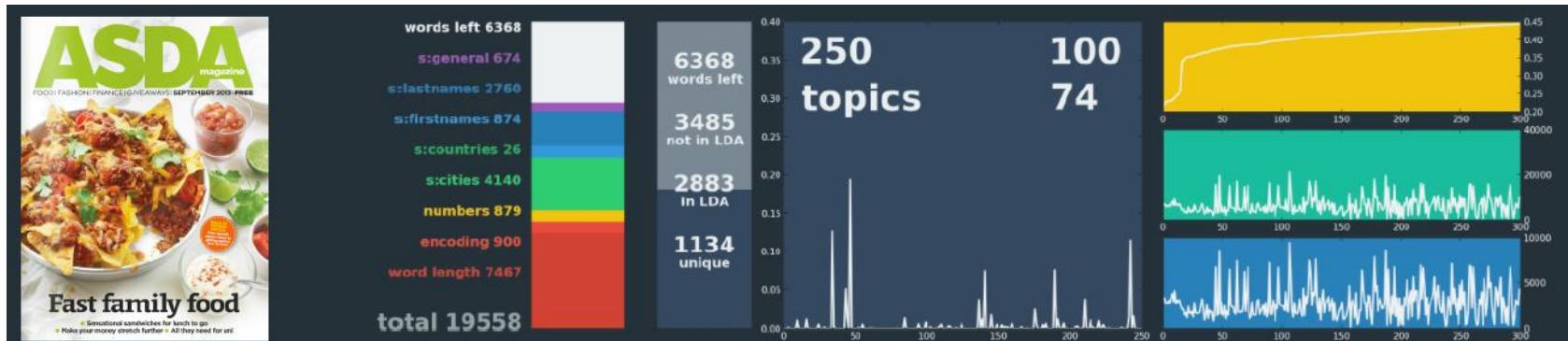


[T16] Organizational learning



# Topic Model: Document Retrieval

Knispelis (2015)

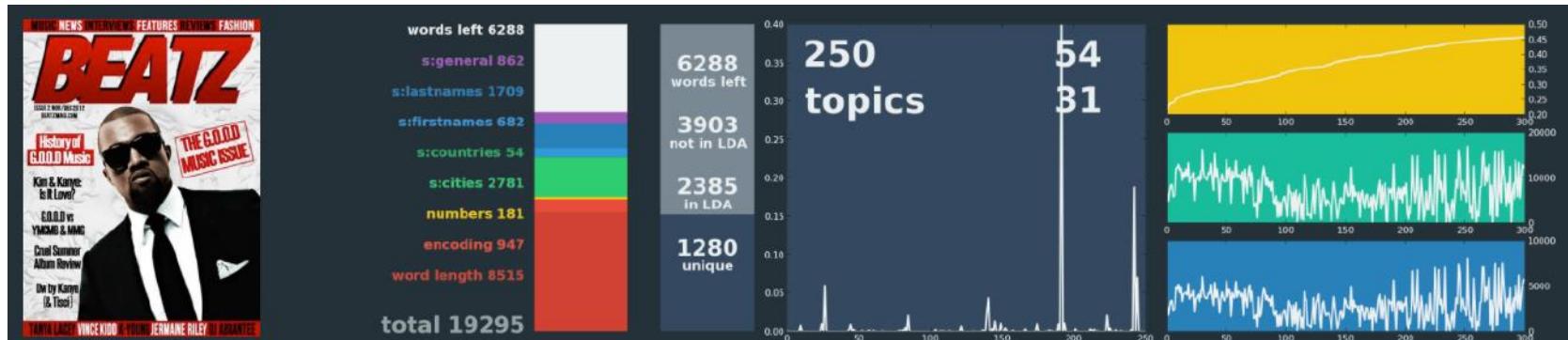


Related documents:



# Topic Model: Document Retrieval

Knispelis (2015)

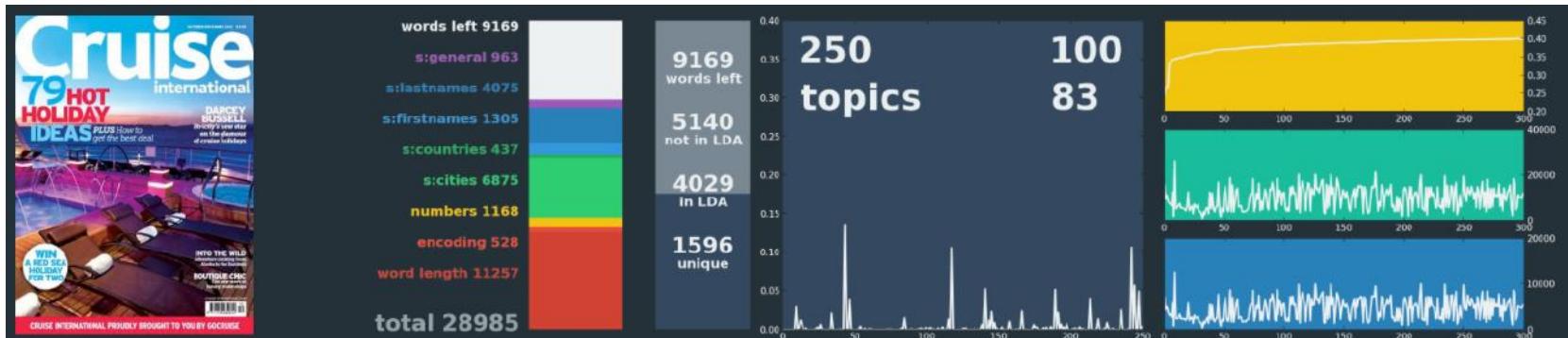


Related documents:



# Topic Model: Document Retrieval

Knispelis (2015)



Related documents:



# Topic Model

- Matrix Factorization Approach

$$\begin{bmatrix} M \times K \\ \text{Topic Assignment} \end{bmatrix} \times \begin{bmatrix} K \times V \\ \text{Topics} \end{bmatrix} \approx \begin{bmatrix} M \times V \\ \text{Dataset} \end{bmatrix}$$

✓ If we use singular value decomposition (SVD), it is called latent semantic analysis (LSA)

Reduced SVD                          Approximating

$$\begin{array}{c} \text{n Documents} \\ \text{m Terms} \\ \text{sparse} \end{array} = \begin{array}{c} r \\ m \\ \text{dense} \end{array} = \begin{array}{c} r \\ r \\ \Sigma \\ r \\ n \\ \text{dense} \end{array}$$
$$\begin{array}{c} : \\ m \\ \text{dense} \end{array} = \begin{array}{c} k \\ k \\ \Sigma_k \\ : \\ U_k \end{array} = \begin{array}{c} k \\ k \\ n \\ \text{dense} \end{array} = \begin{array}{c} V^t \\ : \\ V_k^t \\ = \\ A_k \end{array}$$

# Topic Model

Helic (2014)

- Disadvantage of LSA

- ✓ Statistical foundation is missing
- ✓ SVD assumes normally distributed data
- ✓ Term occurrence is not normally distributed
- ✓ Still, often it works remarkably good because matrix entries are weighted (e.g. tf-idf)  
and those weighted entries may be normally distributed

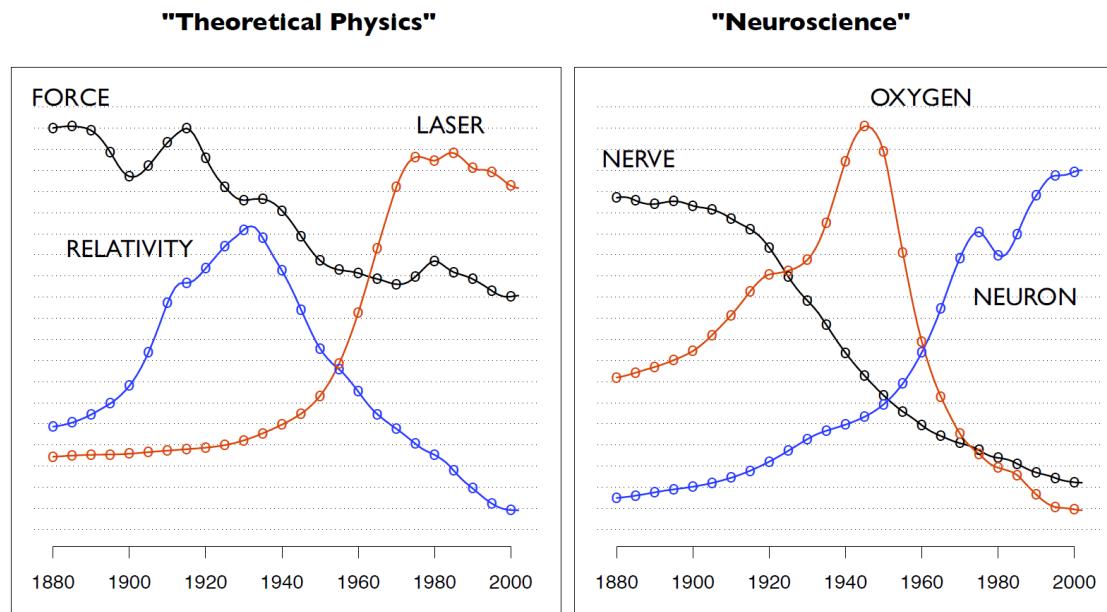
$$A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad U = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad S = \begin{bmatrix} 5.47 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad V = \begin{bmatrix} 0.40 & -0.91 \\ 0.91 & 0.40 \end{bmatrix}$$

# Topic Model

Helic (2014)

- Probabilistic Topic Model: Generative Approach

- ✓ Each document is a probability distribution over topics
- ✓ Distribution over topics represents the essence of a given document
- ✓ Each topic is a probability distribution over words
  - Topic “Education”: school, students, education, university, ...
  - Topic “Budget”: million, finance, tax, program, ...



# Topic Model: Generative Approach

Helic (2014)

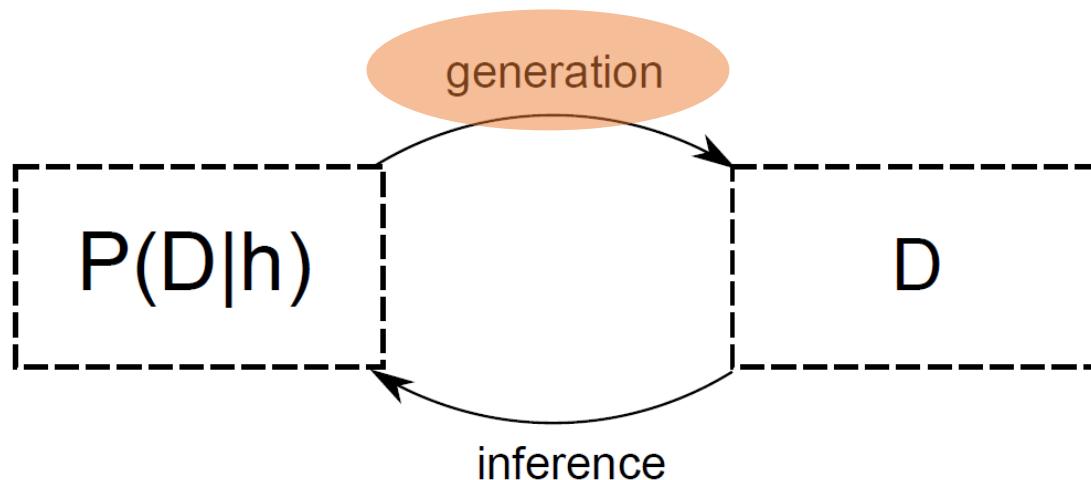
- Model-based methods
  - ✓ Statistical inference is based on fitting a probabilistic model of data
  - ✓ The idea is based on a probabilistic or generative model
  - ✓ Such models assign a probability for observing specific data examples
    - Observing words in a text document
  - ✓ Generative models are powerful method to encode specific assumptions of how unknown parameters interact to create data
- How it work?
  - ✓ It defines a conditional probability distribution over data given a hypothesis  $P(D|h)$
  - ✓ Given  $h$ , we generate data from the conditional distribution  $P(D|h)$
  - ✓ Has many advantages but the main disadvantage is that fitting the model can be more complicated than an algorithmic approach

# Topic Model: Generative Approach

Helic (2014)

- How it work?

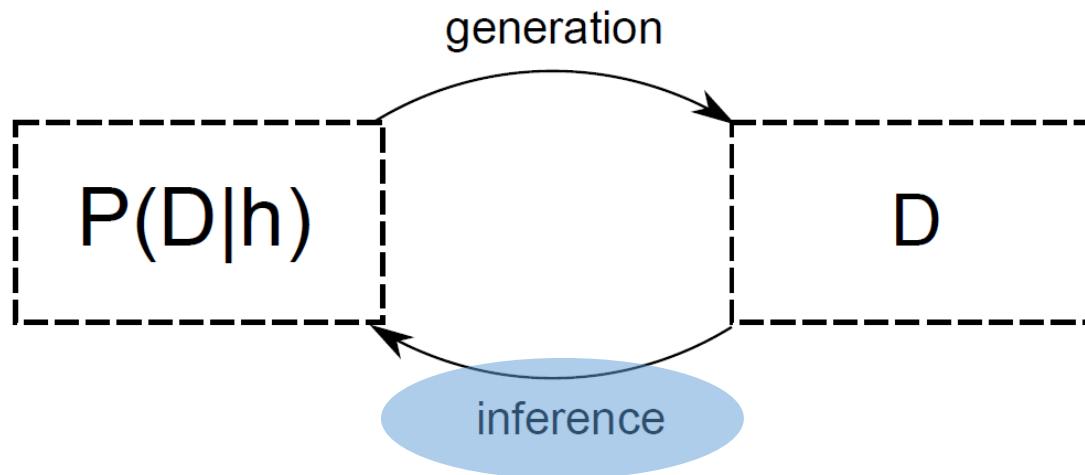
- ✓ It defines a conditional probability distribution over data given a hypothesis  $P(D|h)$
- ✓ Given  $h$ , we generate data from the conditional distribution  $P(D|h)$
- ✓ Has many advantages but the main disadvantage is that fitting the model can be more complicated than an algorithmic approach



# Topic Model: Generative Approach

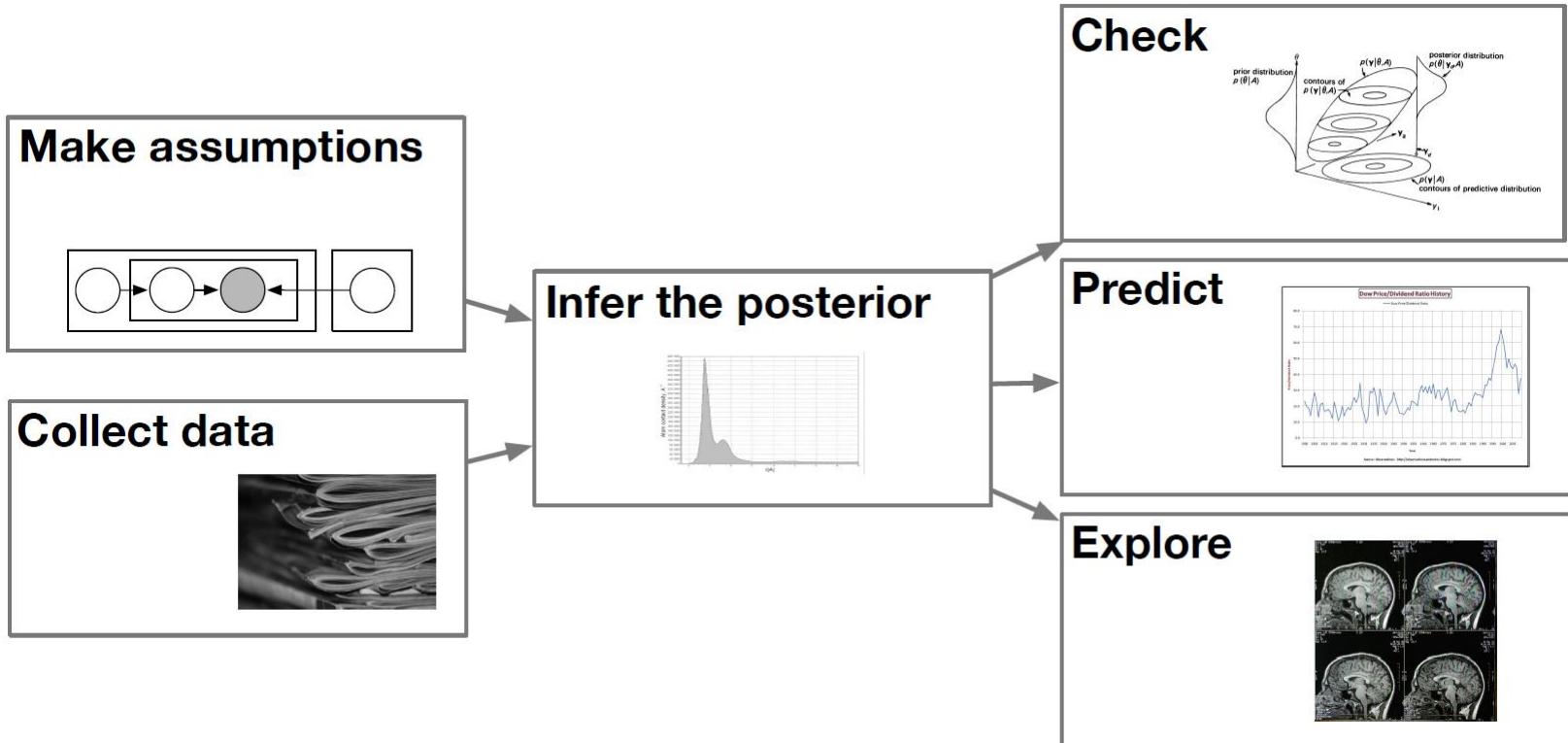
Helic (2014)

- (Statistical) inference is the reverse of the generation process
  - ✓ We are given some data  $D$ , e.g. a collection of documents
  - ✓ We want to estimate the model, or more precisely the parameters of the hypothesis  $h$  that are most likely to have generated data



# Topic Model: Generative Approach

- Process of generative model



# AGENDA

- 01 Topic Modeling
- 02 Probabilistic Latent Semantic Analysis
- 03 LDA: Document Generation Process
- 04 LDA Inference: Gibbs Sampling
- 05 LDA Evaluation

# Latent Structure

Hofmann (2005)

- Given a matrix that “encodes” data (e.g. term-document matrix), we have following potential problems
  - ✓ Too large
  - ✓ Too complicated
  - ✓ Lack of structure
  - ✓ Missing Entries
  - ✓ Noisy Entries, ...

$$\mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} & \dots & a_{im} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nm} \end{pmatrix}$$

- Questions
  - ✓ Is there a **simpler** way to **explain** entities?
  - ✓ There might be a **latent structure** underlying the data
  - ✓ How can we reveal or **discover** this structure?

# Matrix Decomposition

Hofmann (2005)

- Common approach: approximately factorize matrix

$$A \approx \hat{A} = L \cdot R$$

approximation      left factor      right factor

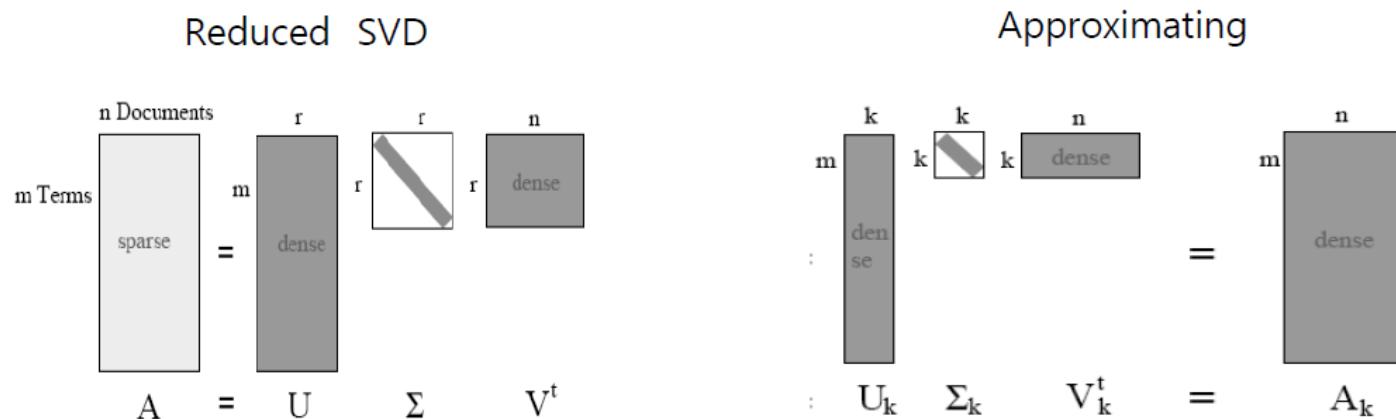
- Factors are typically constrained to be “thin”

$$\begin{array}{c|ccccc} & \overbrace{\hspace{1cm}}^m & & \overbrace{\hspace{1cm}}^m & \\ \hline | & \boxed{\text{A}} & \approx & | & \overbrace{\hspace{1cm}}^n & \overbrace{\hspace{1cm}}^q \\ & n & & | & \boxed{\text{L}} & \cdot \boxed{\text{R}} & q \\ & & & & \diagdown & & \\ & & & & n & & \end{array}$$

reduction  
 $n \cdot m \gg n \cdot q + m \cdot q$   
factors = latent structure

# LSA Decomposition (revisited)

- Reduce the dimensions using SVD



- ✓ Step 1) Construct the approximated matrix  $A_k$  from the original term-document matrix  $A$  using SVD

$$A \approx A_k = U_k \Sigma_k V_k^T$$

- ✓ Step 2) Multiply the transpose of  $U_k$  to obtain  $k (<<m)$  by  $n$  term-document matrix

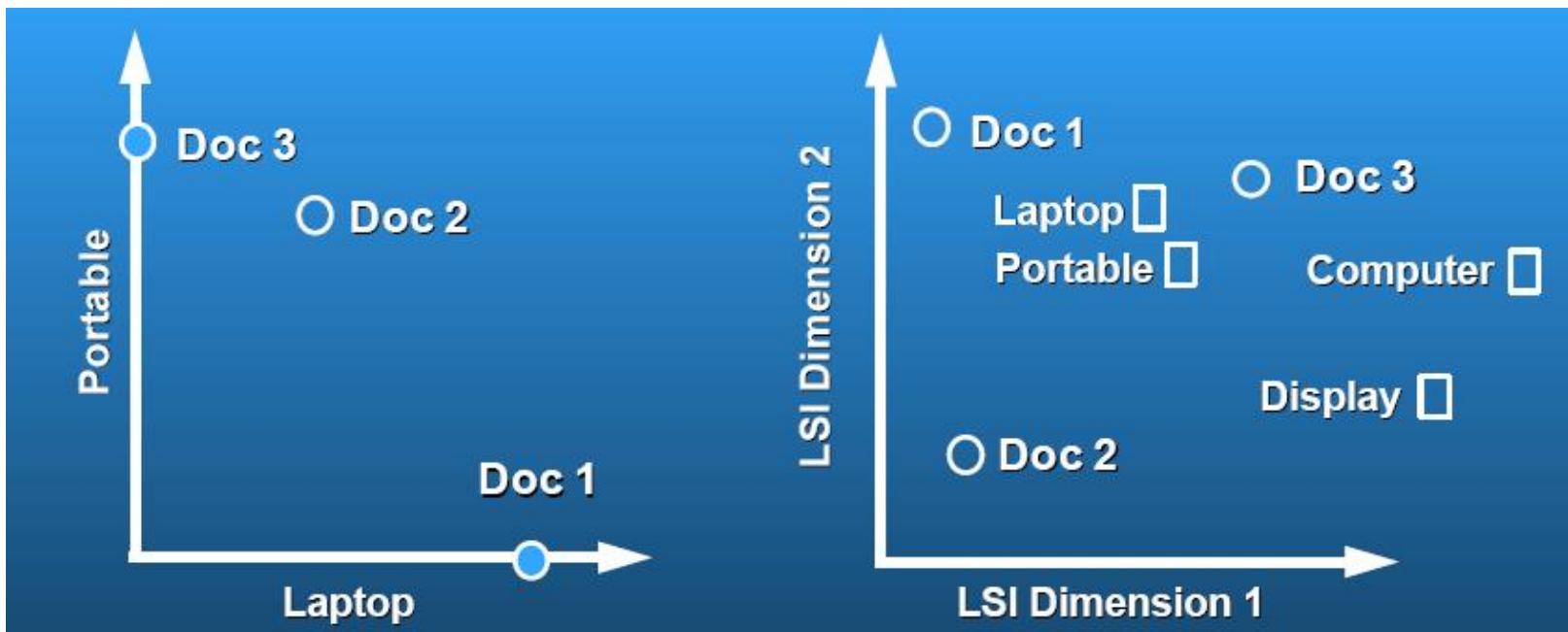
$$U_k^T A_k = \color{blue}{U_k^T U_k} \Sigma_k V_k^T = \color{blue}{I} \Sigma_k V_k^T = \Sigma_k V_k^T$$

- ✓ Step 3: Apply data mining algorithms

# LSA Decomposition

Hofmann (2005)

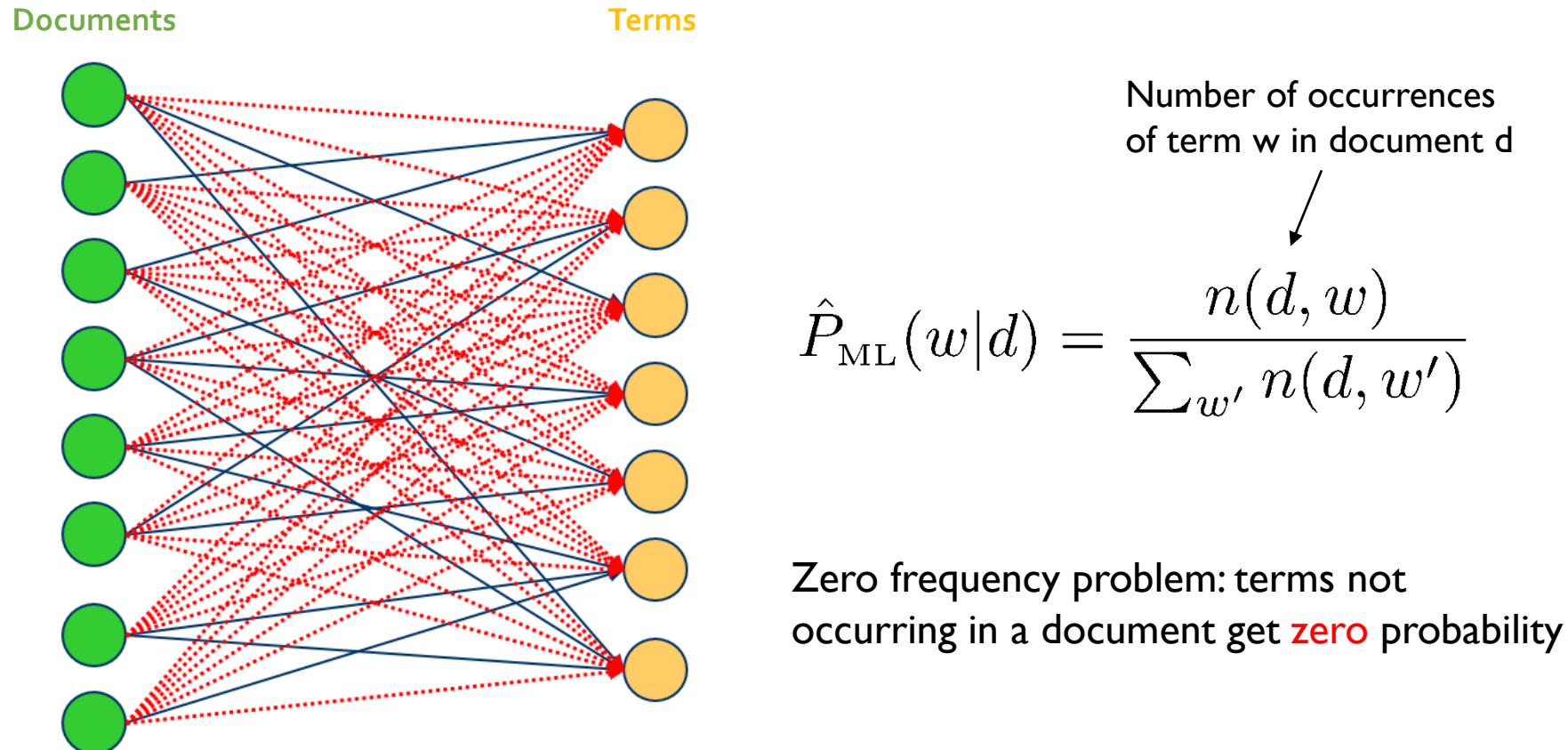
- Illustrative Example



# Language Model: Naïve Approach

Hofmann (2005)

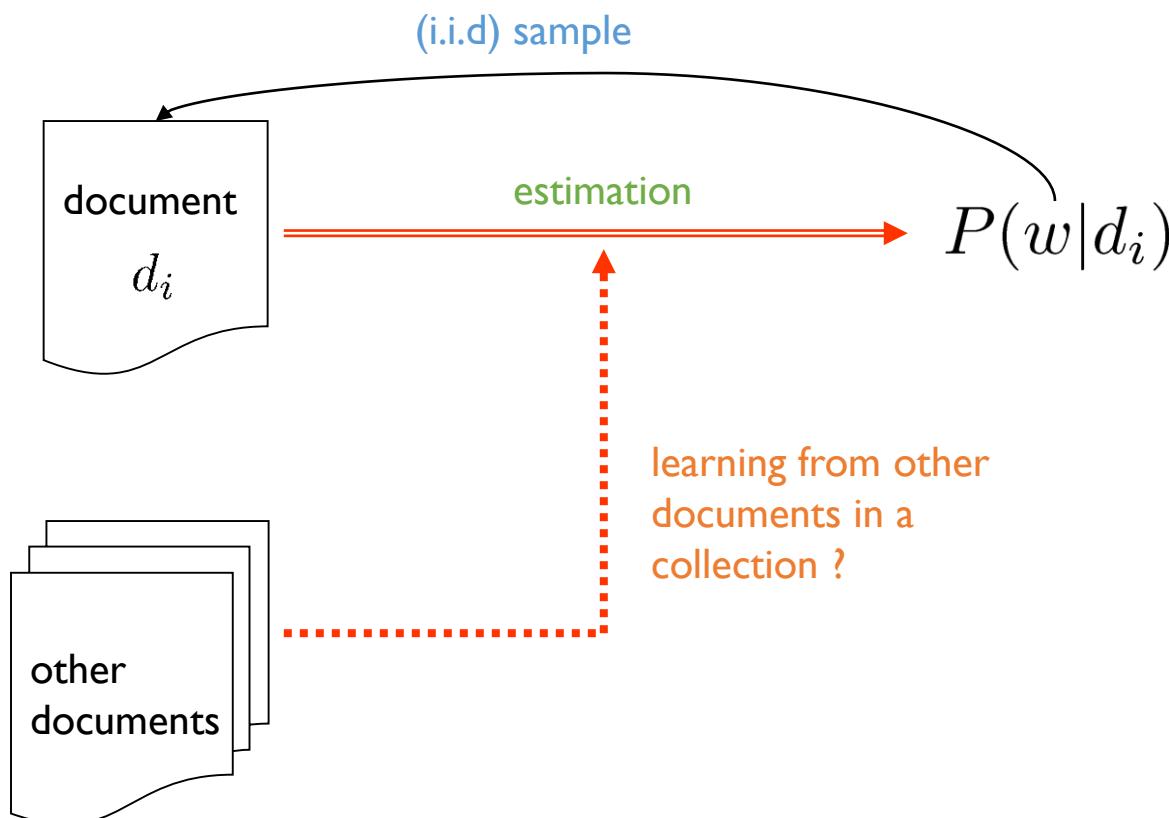
- Maximum likelihood estimation (MLE)



# Language Model: Estimation Problem

Hofmann (2005)

- Crucial question
  - ✓ In which way can the document collection be utilized to improve estimates?

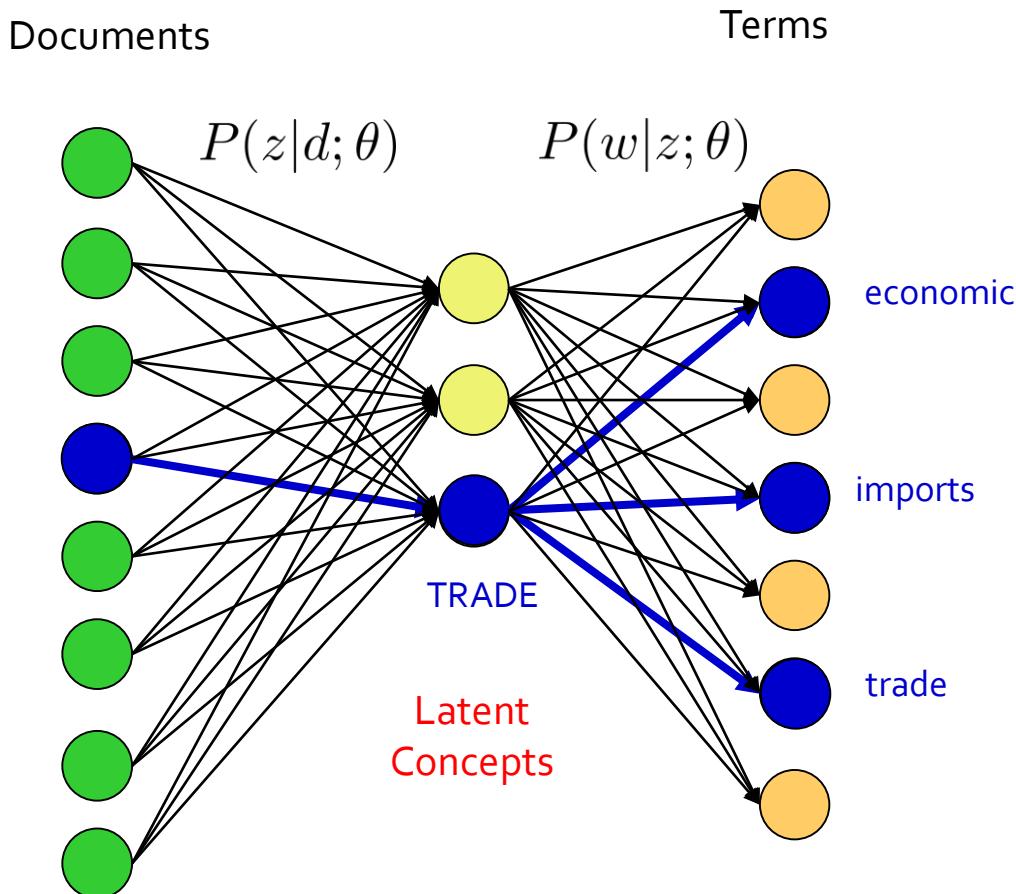


# Probabilistic Latent Semantic Analysis (pLSA)

Hofmann (2005)

- Concept expression probability

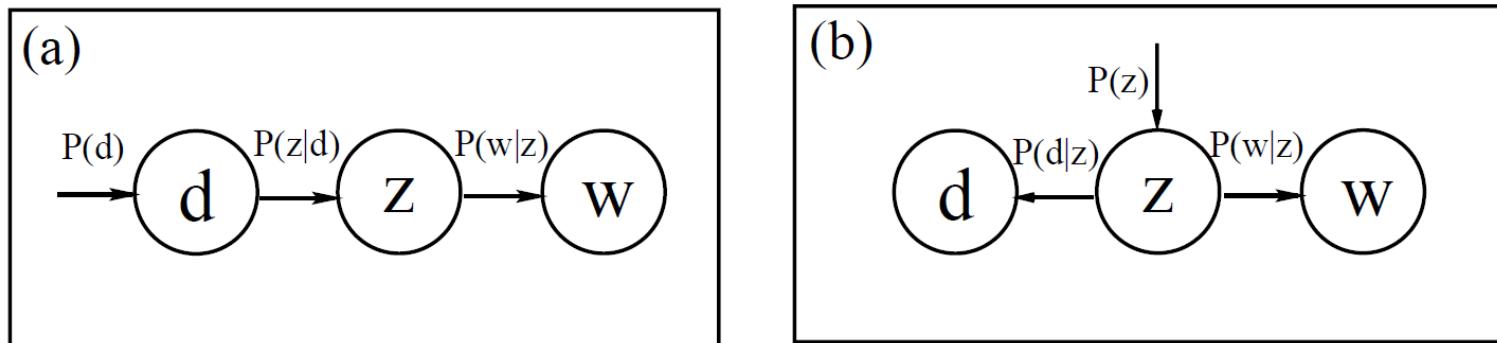
- ✓ Estimated based on all documents that are dealing with a concept
- ✓ “Unmixing” of superimposed concepts is achieved by statistical learning algorithm
- ✓ No prior knowledge about concepts required, context and term co-occurrences are exploited



# pLSA: Latent Variable Model

Hofmann (2005)

- Structural modeling assumption (mixture model)



$$\hat{P}_{\text{LSA}}(w|d) = \sum_z P(w|z; \theta) P(z|d; \pi)$$

Document language model

Latent concepts or topics

Document-specific mixture proportions

Concept expression probabilities

Model fitting

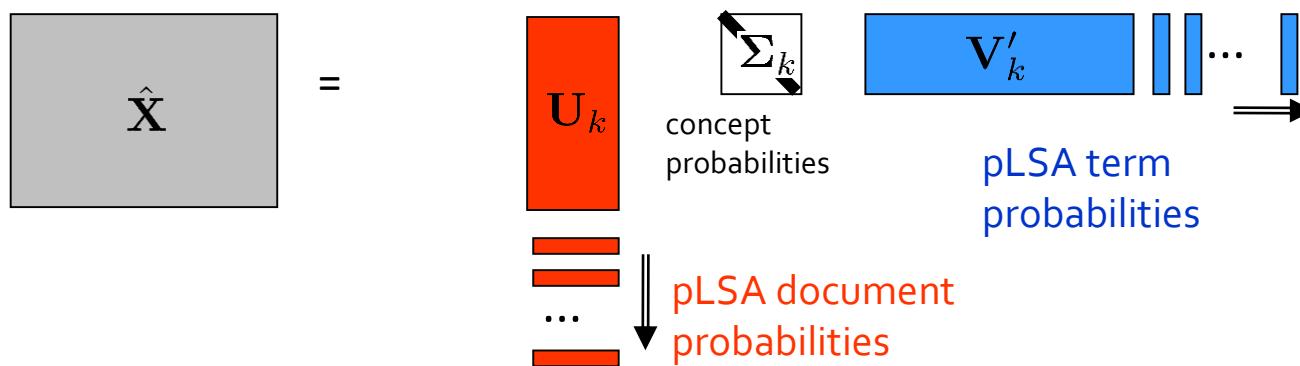
A red oval highlights the term  $P(w|z; \theta) P(z|d; \pi)$ . A red arrow points from the oval to the text "Model fitting".

# pLSA: Matrix Decomposition

Hofmann (2005)

- Mixture model can be written as a matrix factorization

$$\hat{P}_{\text{LSA}}(d, w) = \sum_z P(d|z) P(z) P(w|z) = P(d) \sum_z P(w|z) P(z|d)$$



- Contrast to LSA

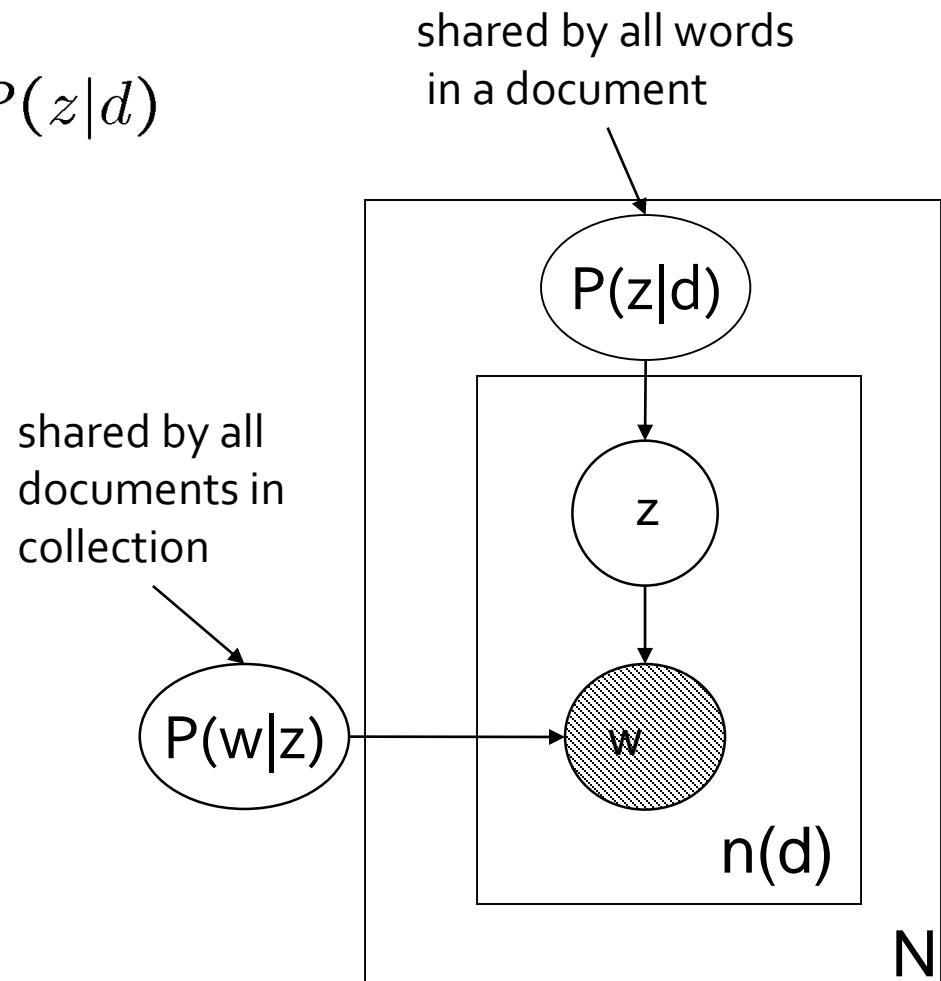
- ✓ **Non-negativity**: every element in  $U$  &  $V$  is non-negative
- ✓ **Normalization**: Each document vector in  $U$  and each term vector in  $V$  has sum 1

# pLSA: Graphical Model

Hofmann (2005)

- Graphical Representation

$$P(w|d) = \sum_z P(w|z)P(z|d)$$



# pLSA: Parameter Inference

Helic (2014)

- Parameter inference
  - ✓ We will infer parameters using Maximum Likelihood Estimator (MLE)
  - ✓ First, we need to write down the likelihood function
  - ✓ Let  $n(w_i, d_j)$  be the number of occurrences of word  $w_i$  in document  $d_j$
  - ✓  $p(w_i, d_j)$  is the probability of observing a single occurrence word  $w_i$  in document  $d_j$
  - ✓ Then, the probability of observing  $n(w_i, d_j)$  occurrence of word  $w_i$  in document  $d_j$  is given by:

$$p(w_i, d_j)^{n(w_i, d_j)}$$

# pLSA: Parameter Inference

Helic (2014)

- Parameter Inference

- ✓ The probability of observing the compete document collection is then given by the product of probabilities of observing every single word in every document with corresponding number of occurrences
- ✓ Then, the likelihood function becomes

$$L = \prod_{i=1}^m \prod_{j=1}^n p(w_i, d_j)^{n(w_i, d_j)}$$

- ✓ The log-likelihood function becomes

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^m \sum_{j=1}^n n(w_i, d_j) \log(p(w_i, d_j)) \\ &= \sum_{i=1}^m \sum_{j=1}^n n(w_i, d_j) \log\left(\sum_{l=1}^k p(w_i|z_l)p(z_l)p(d_j|z_l)\right)\end{aligned}$$

# pLSA: Parameter Inference

Helic (2014)

- Parameter Inference
  - ✓ We can not maximize the likelihood analytically because of the logarithm of the sum
  - ✓ A standard procedure is to use an algorithm called Expectation-Maximization (EM)
  - ✓ This is an iterative method to estimate parameters of the models with latent variables
  - ✓ Each iteration consists of two steps: expectation step (E) and maximization step (M)

# pLSA: EM Algorithm

- E-Step: Posterior probability of latent variables (concepts)

$$p(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')}$$

Probability that the occurrence of term w in document d can be “explained” by concept z

- M-Step: Parameter estimation based on “completed” statistics

$$P(w|z) = \frac{\sum_{d \in D} n(d, w) P(z|d, w)}{\sum_{d \in D, w' \in W} n(d, w') P(z|d, w')}$$

how often is term w associated with concept z ?

$$P(d|z) = \frac{\sum_{w \in W} n(d, w) P(z|d, w)}{\sum_{d' \in D, w \in W} n(d', w) P(z|d', w)}$$

how often is document d associated with concept z ?

$$P(z) = \frac{\sum_{d \in D, w \in W} n(d, w) P(z|d, w)}{\sum_{d \in D, w \in W} n(d, w)}$$

how prevalent is the concept z ?

# pLSA: A Simple Example

- Raw Data

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6
Baseball	1	2	0	0	0	0
Basketball	3	1	0	0	0	0
Boxing	2	0	0	0	0	0
Money	3	3	2	3	2	4
Interest	0	0	3	2	0	0
Rate	0	0	4	1	0	0
Democrat	0	0	0	0	4	3
Republican	0	0	0	0	2	1
Cucus	0	0	0	0	3	2
President	0	0	1	0	2	3

# pLSA: A Simple Example

- Parameter Initialization

$P(z)$

Topic 1	Topic 2	Topic 3
0.525	0.407	0.068

$P(d|z)$

	Topic 1	Topic 2	Topic 3
Doc 1	0.020	0.008	0.048
Doc 2	0.294	0.255	0.329
Doc 3	0.204	0.138	0.178
Doc 4	0.200	0.146	0.007
Doc 5	0.186	0.196	0.233
Doc 6	0.096	0.257	0.205

$P(w|z)$

	Topic 1	Topic 2	Topic 3
Term 1	0.022	0.016	0.010
Term 2	0.018	0.133	0.166
Term 3	0.242	0.058	0.133
Term 4	0.123	0.088	0.145
Term 5	0.016	0.030	0.044
Term 6	0.020	0.167	0.056
Term 7	0.147	0.129	0.201
Term 8	0.188	0.156	0.039
Term 9	0.146	0.114	0.008
Term 10	0.077	0.110	0.199

# pLSA: A Simple Example

- After 1 EM step

Initialization

Topic 1	Topic 2	Topic 3
0.525	0.407	0.068

After 1 EM step

Topic 1	Topic 2	Topic 3
0.459	0.430	0.111

	Topic 1	Topic 2	Topic 3
Doc 1	0.020	0.008	0.048
Doc 2	0.294	0.255	0.329
Doc 3	0.204	0.138	0.178
Doc 4	0.200	0.146	0.007
Doc 5	0.186	0.196	0.233
Doc 6	0.096	0.257	0.205

	Topic 1	Topic 2	Topic 3
Doc 1	0.180	0.077	0.382
Doc 2	0.124	0.089	0.091
Doc 3	0.147	0.213	0.149
Doc 4	0.125	0.110	0.004
Doc 5	0.266	0.204	0.167
Doc 6	0.158	0.308	0.207

# pLSA: A Simple Example

- After 1 EM step

Initialization

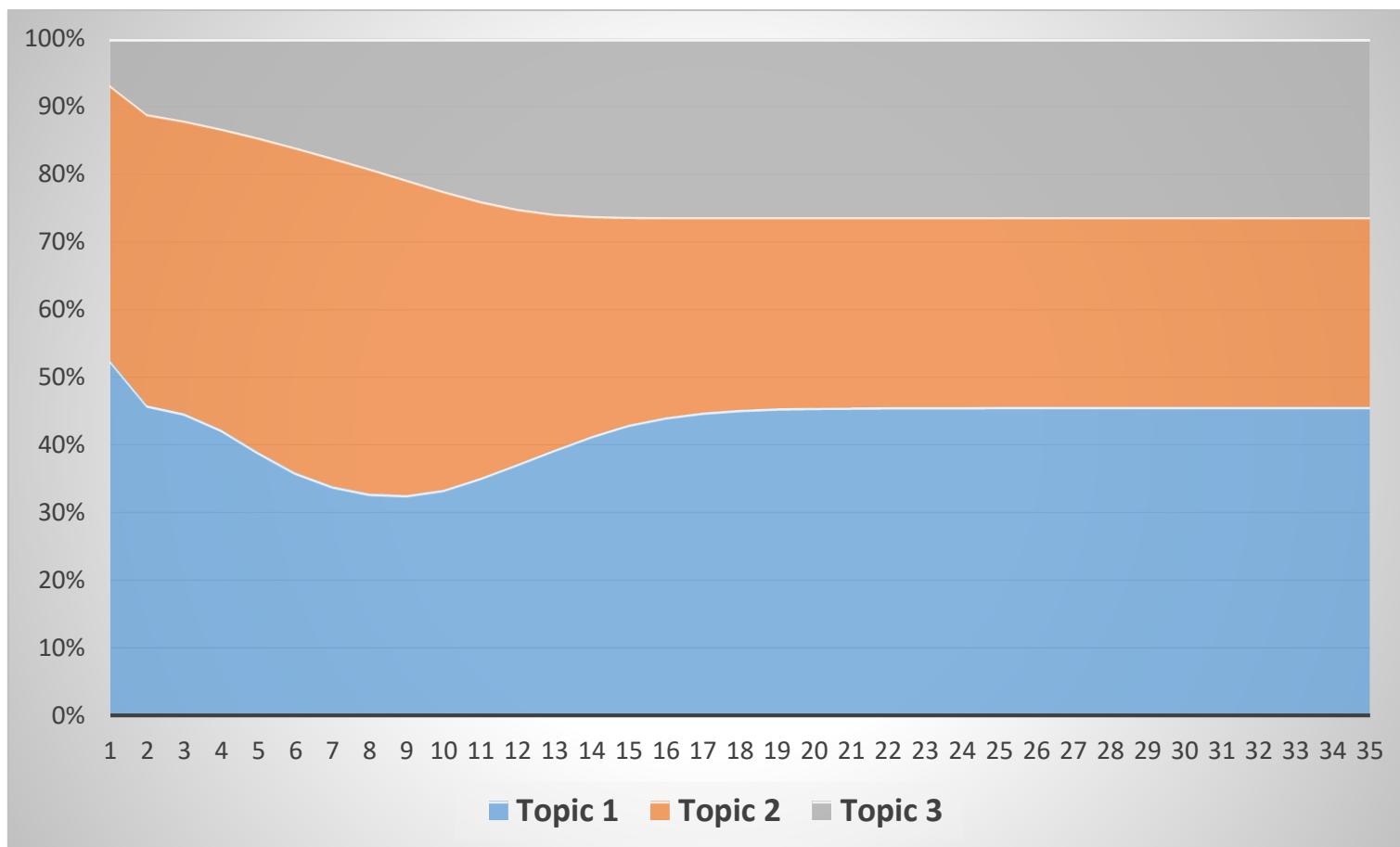
	Topic 1	Topic 2	Topic 3
Term 1	0.022	0.016	0.010
Term 2	0.018	0.133	0.166
Term 3	0.242	0.058	0.133
Term 4	0.123	0.088	0.145
Term 5	0.016	0.030	0.044
Term 6	0.020	0.167	0.056
Term 7	0.147	0.129	0.201
Term 8	0.188	0.156	0.039
Term 9	0.146	0.114	0.008
Term 10	0.077	0.110	0.199

After 1 EM step

	Topic 1	Topic 2	Topic 3
Term 1	0.077	0.033	0.028
Term 2	0.024	0.074	0.245
Term 3	0.061	0.005	0.043
Term 4	0.370	0.222	0.295
Term 5	0.088	0.093	0.065
Term 6	0.033	0.159	0.035
Term 7	0.115	0.129	0.129
Term 8	0.058	0.058	0.010
Term 9	0.099	0.098	0.004
Term 10	0.073	0.129	0.146

# pLSA: A Simple Example

- Topic Distribution
  - ✓ Topic distribution changes w.r.t. the EM iterations



# pLSA: A Simple Example

- Final result

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6
Baseball	1	2	0	0	0	0
Basketball	3	1	0	0	0	0
Boxing	2	0	0	0	0	0
Money	3	3	2	3	2	4
Interest	0	0	3	2	0	0
Rate	0	0	4	1	0	0
Democrat	0	0	0	0	4	3
Republican	0	0	0	0	2	1
Cucus	0	0	0	0	3	2
President	0	0	1	0	2	3

Topic 1	Topic 2	Topic 3
0.456	0.281	0.263

	Topic 1	Topic 2	Topic 3
Doc 1	0.000	0.000	0.600
Doc 2	0.000	0.000	0.400
Doc 3	0.000	0.625	0.000
Doc 4	0.000	0.375	0.000
Doc 5	0.500	0.000	0.000
Doc 6	0.500	0.000	0.000

	Topic 1	Topic 2	Topic 3
Baseball	0.000	0.000	0.200
Basketball	0.000	0.000	0.267
Boxing	0.000	0.000	0.133
Money	0.231	0.313	0.400
Interest	0.000	0.312	0.000
Rate	0.000	0.312	0.000
Democrat	0.269	0.000	0.000
Republican	0.115	0.000	0.000
Cucus	0.192	0.000	0.000
President	0.192	0.063	0.000

# pLSA: Example

- Concepts extracted from Science Magazine articles

$P(w|z)$

universe	0.0439
galaxies	0.0375
clusters	0.0279
matter	0.0233
galaxy	0.0232
cluster	0.0214
cosmic	0.0137
dark	0.0131
light	0.0109
density	0.01

drug	0.0672
patients	0.0493
drugs	0.0444
clinical	0.0346
treatment	0.028
trials	0.0277
therapy	0.0213
trial	0.0164
disease	0.0157
medical	0.00997

cells	0.0675
stem	0.0478
human	0.0421
cell	0.0309
gene	0.025
tissue	0.0185
cloning	0.0169
transfer	0.0155
blood	0.0113
embryos	0.0111

sequence	0.0818
sequences	0.0493
genome	0.033
dna	0.0257
sequencing	0.0172
map	0.0123
genes	0.0122
chromosome	0.0119
regions	0.0119
human	0.0111

years	0.156
million	0.0556
ago	0.045
time	0.0317
age	0.0243
year	0.024
record	0.0238
early	0.0233
billion	0.0177
history	0.0148

$P(w|z)$

bacteria	0.0983
bacterial	0.0561
resistance	0.0431
coli	0.0381
strains	0.025
microbiol	0.0214
microbial	0.0196
strain	0.0165
salmonella	0.0163
resistant	0.0145

male	0.0558
females	0.0541
female	0.0529
males	0.0477
sex	0.0339
reproductive	0.0172
offspring	0.0168
sexual	0.0166
reproduction	0.0143
eggs	0.0138

theory	0.0811
physics	0.0782
physicists	0.0146
einstein	0.0142
university	0.013
gravity	0.013
black	0.0127
theories	0.01
aps	0.00987
matter	0.00954

immune	0.0909
response	0.0375
system	0.0358
responses	0.0322
antigen	0.0263
antigens	0.0184
black	0.0176
immunity	0.0176
immunology	0.0145
antibody	0.014
autoimmune	0.0128

stars	0.0524
star	0.0458
astrophys	0.0237
mass	0.021
disk	0.0173
black	0.0161
gas	0.0149
stellar	0.0127
astron	0.0125
hole	0.00824

# pLSA: Example

- Example

✓ Polysemy: a word may have multiple senses and multiple types of usage in different context

“segment 1”	“segment 2”	“matrix 1”	“matrix 2”	“line 1”	“line 2”	“power 1”	power 2”
imag SEGMENT	speaker speech	robust MATRIX	manufactur cell	constraint LINE	alpha redshift	POWER spectrum	load memori
texture	recogni	eigenvalu	part	match	LINE	omega	vlsi
color	signal	uncertainti	MATRIX	locat	galaxi	mpc	POWER
tissue	train	plane	cellular	imag	quasar	hsup	systolic
brain	hmm	linear	famili	geometr	absorp	larg	input
slice	source	condition	design	impos	high	redshift	complex
cluster	speakerind.	perturb	machinepart	segment	ssup	galaxi	arra
mri	SEGMENT	root	format	fundament	densiti	standard	present
volume	sound	suffici	group	recogn	veloc	model	implement

Document 1,  $P\{z_k|d_1, w_j = \text{'segment'}\} = (0.951, 0.0001, \dots)$

$P\{w_j = \text{'segment'}|d_1\} = 0.06$

SEGMENT medic imag challeng problem field imag analysi diagnost base proper SEGMENT digit imag SEGMENT medic imag need applic involv estim boundari object classif tissu abnorm shape analysi contour detec textur SEGMENT despit exist techniqu SEGMENT specif medic imag remain crucial problem [...]

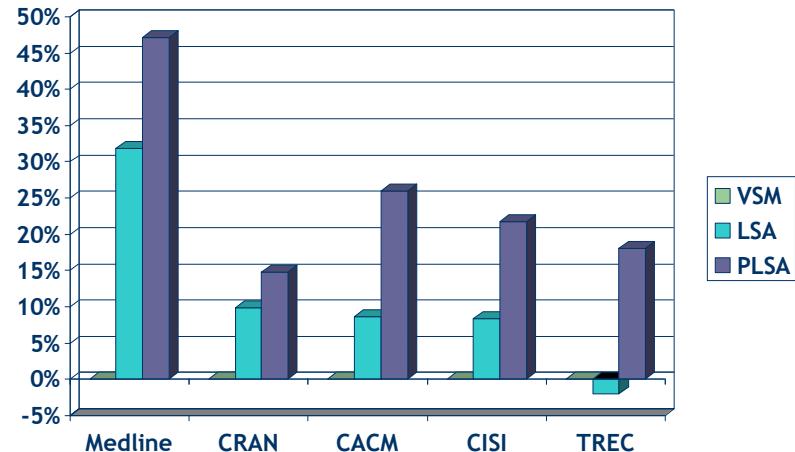
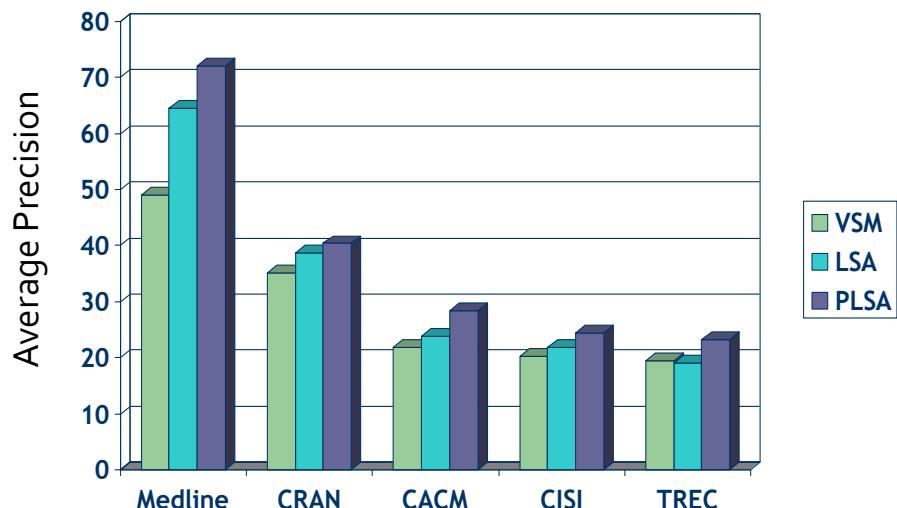
Document 2,  $P\{z_k|d_2, w_j = \text{'segment'}\} = (0.025, 0.867, \dots)$

$P\{w_j = \text{'segment'}|d_2\} = 0.010$

consid signal origin sequenc sourc specif problem SEGMENT signal relat SEGMENT sourc address issu wide applic field report describ resolu method ergod hidden markov model hmm hmm state correspond signal sourc signal sourc sequenc determin decod procedur viterbi algorithm forward algorithm observ sequenc baumwelch train estim hmm paramet train materi applic multipl signal sourc identif problem experi perform unknown speaker identif [...]

# pLSA: Example

- Experimental Evaluation



- ✓ Consistent improvements of retrieval accuracy
- ✓ Relative improvement of average precision: 15-45%

# AGENDA

- 01 Topic Modeling
- 02 Probabilistic Latent Semantic Analysis
- 03 LDA: Document Generation Process
- 04 LDA Inference: Gibbs Sampling
- 05 LDA Evaluation

# LDA: Intuition

Blei (2012)

- Documents exhibit multiple topics

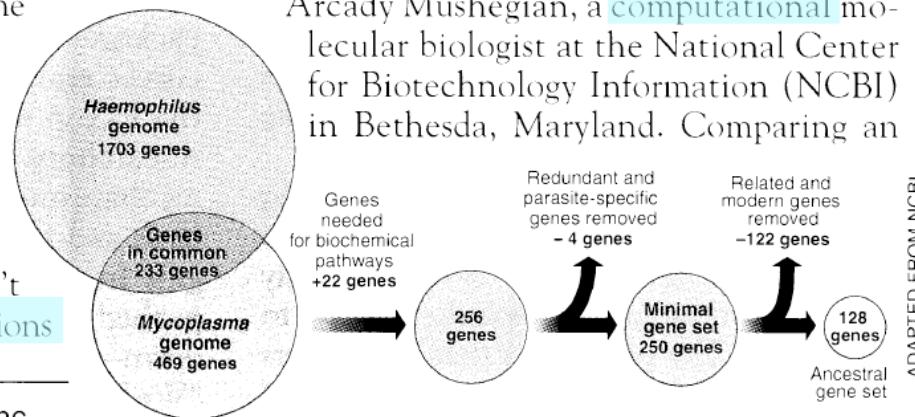
## Seeking Life's Bare (Genetic) Necessities

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



ADAPTED FROM NCBI

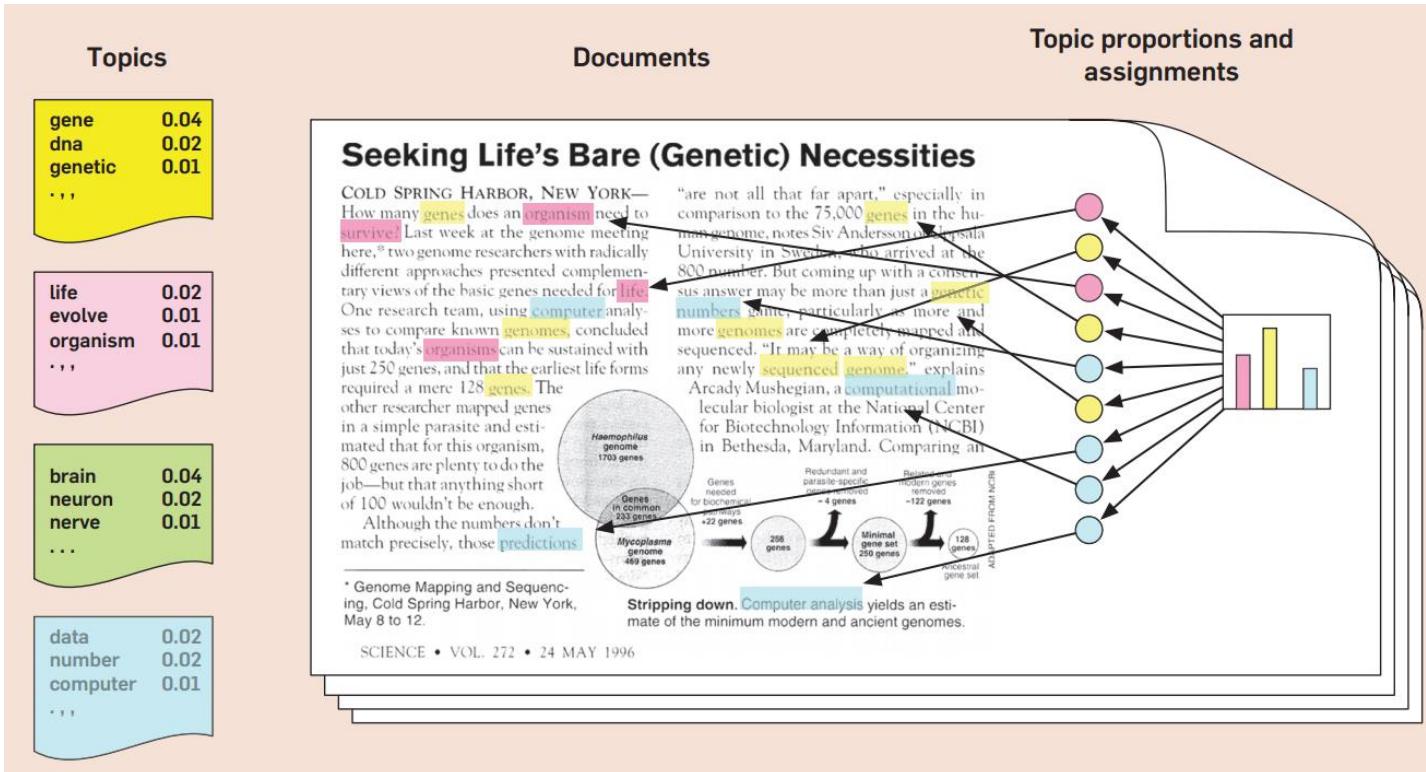
**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

# LDA: Intuition

Blei (2012)

- Documents exhibit multiple topics

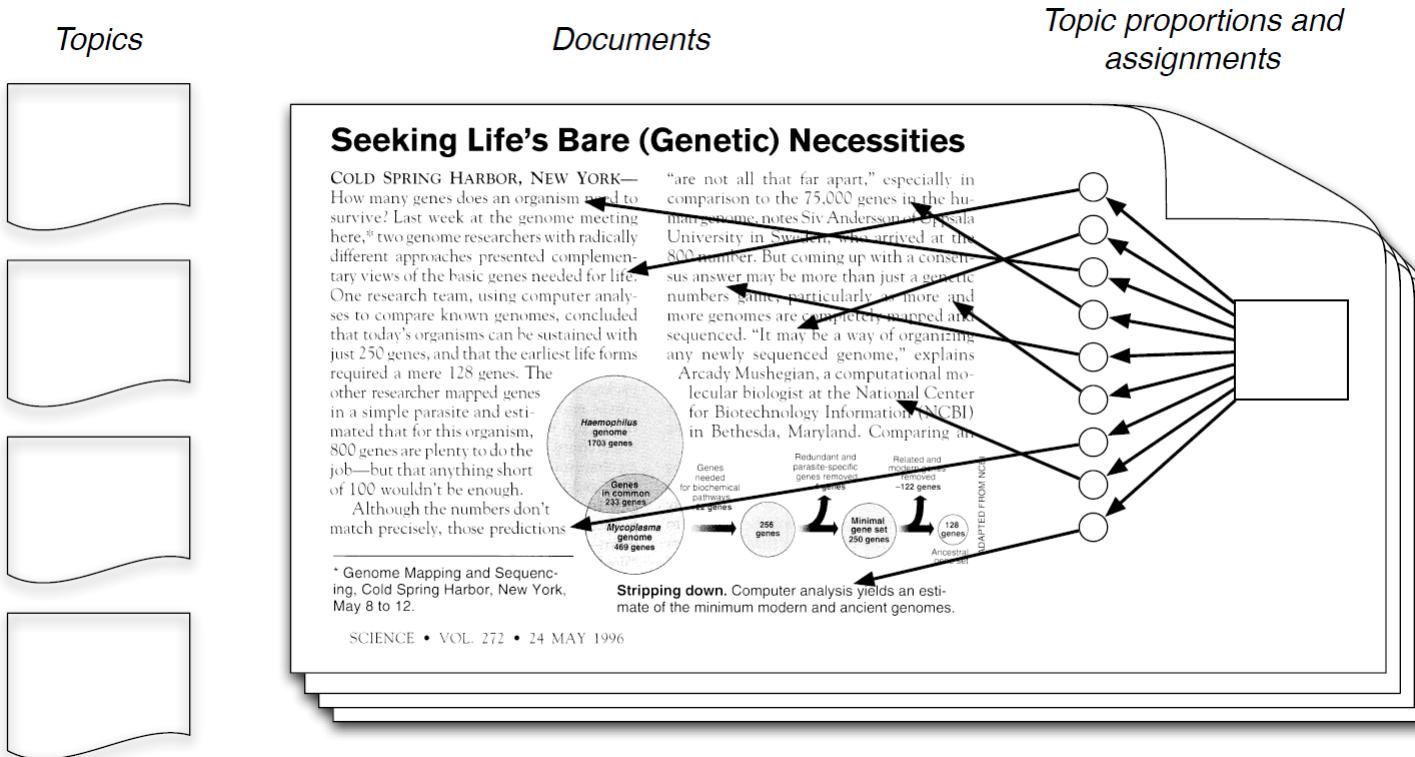


- ✓ Each **topic** is a distribution over words
- ✓ Each **document** is a mixture of corpus-wide topics
- ✓ Each **word** is drawn from one of those topics

# LDA: Intuition

Blei (2012)

- Documents exhibit multiple topics

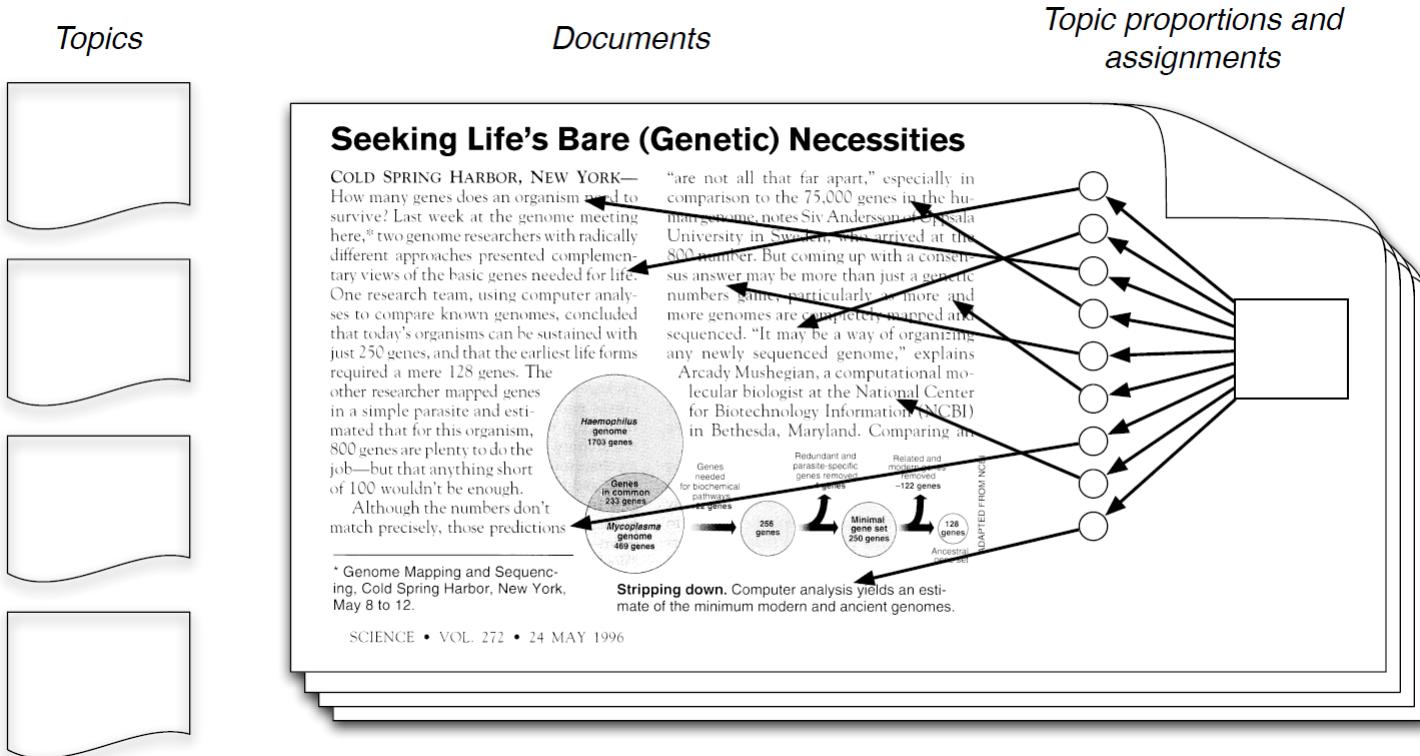


- ✓ In reality, we only observe the documents
- ✓ The other structure are hidden variables

# LDA: Intuition

Blei (2012)

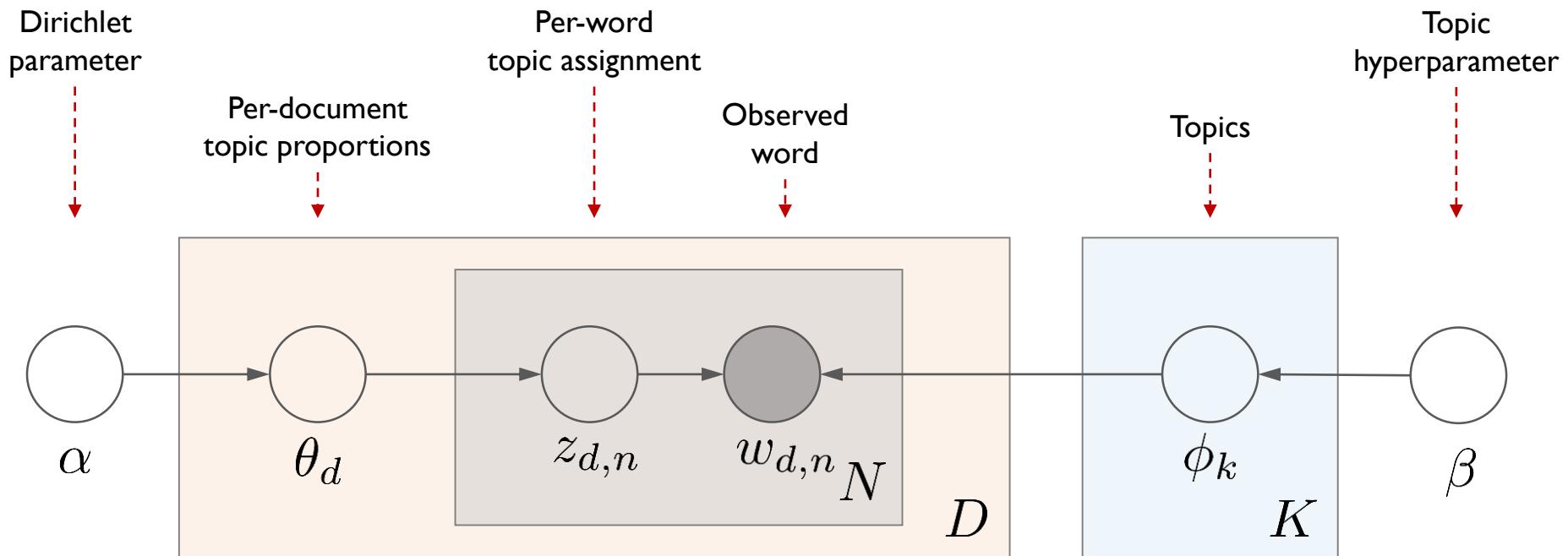
- Documents exhibit multiple topics



- ✓ The goal of LDA is to infer the hidden variables
- ✓ i.e. compute their distribution conditioned on the document
- ➔  $p(\text{topics, proportions, assignments} \mid \text{documents})$

# LDA Overview

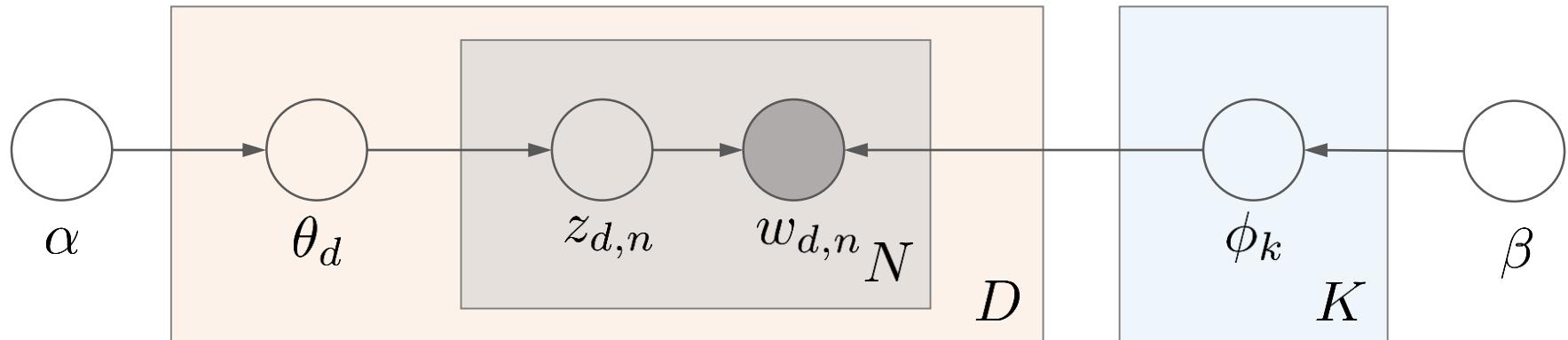
- Documents exhibit multiple topics



- ✓ Encode **assumptions**
- ✓ Define a **factorization** of the joint distribution
- ✓ Connect to **algorithms** for computing with data

# LDA Overview

- LDA structure



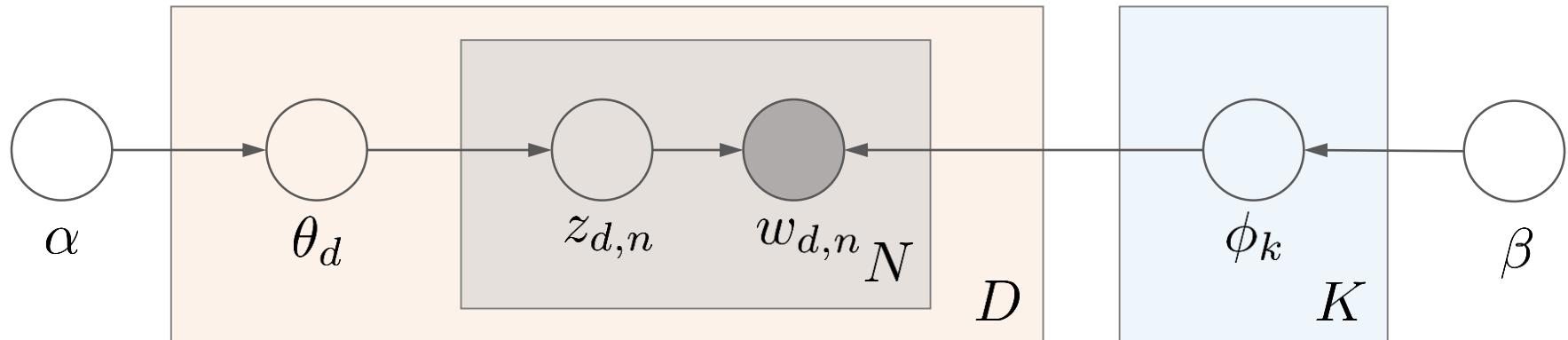
- ✓ Nodes are random variables while edges indicate dependence
- ✓ Shaded nodes are observed
- ✓ Plates indicate replicated variables

$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

$$\prod_{i=1}^K p(\phi_i | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \phi_{1:K}, z_{d,n}) \right)$$

# LDA: Document generation process

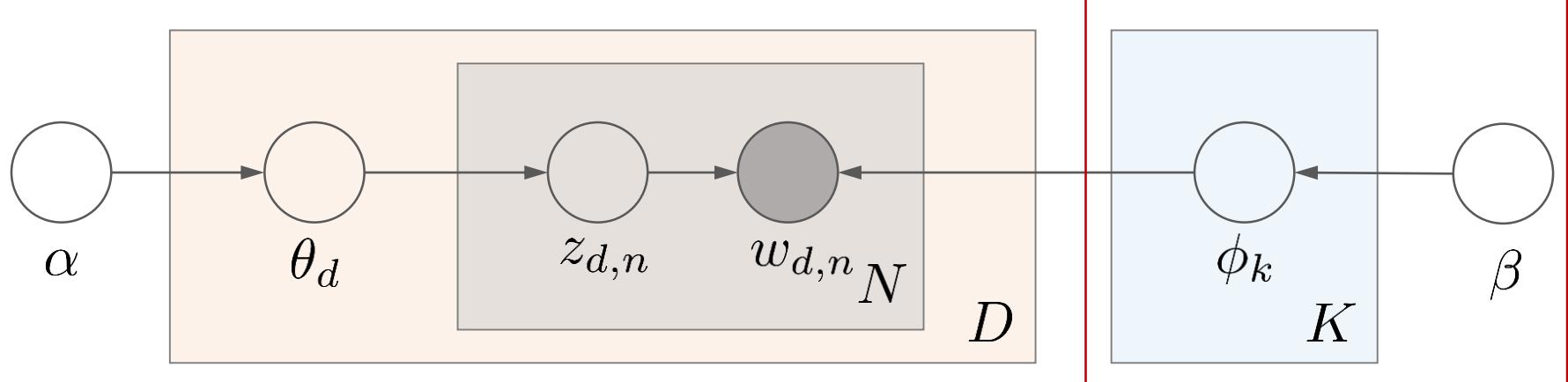
- Document generation process



- ✓ Draw each topic  $\phi_k \sim Dir(\beta)$  for  $i \in \{1, \dots, K\}$
- ✓ For each document
  - Draw topic proportions  $\theta_d \sim Dir(\alpha)$
  - For each word
    - Draw  $z_{d,n} \sim Multi(\theta_d)$
    - Draw  $w_{d,n} \sim Multi(\phi_{z_{d,n}, n})$

# LDA: Document generation process

- Document generation process



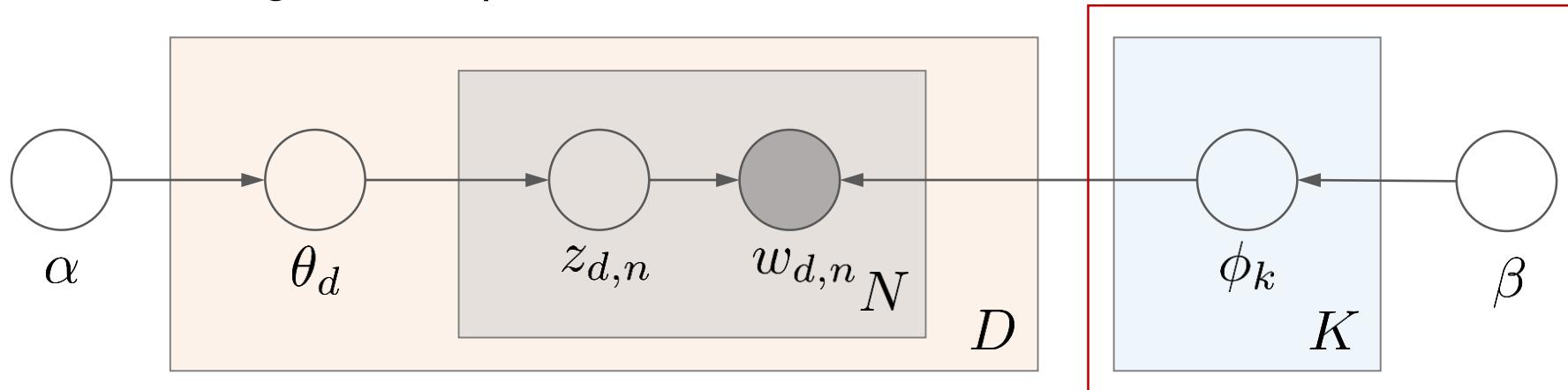
✓ Term distribution per topic

- Drawn from the Dirichlet distribution, given the Dirichlet parameter  $\beta$ , which is a  $V$ -vector with component  $\beta_v > 0$

$$p(\phi|\beta) = \prod_{k=1}^K \frac{\Gamma(\beta_{k,\cdot})}{\prod_{v=1}^V \Gamma(\beta_{k,v})} \prod_{v=1}^V \phi_{k,v}^{\beta_{k,v}-1}$$

# LDA: Document generation process

- Document generation process



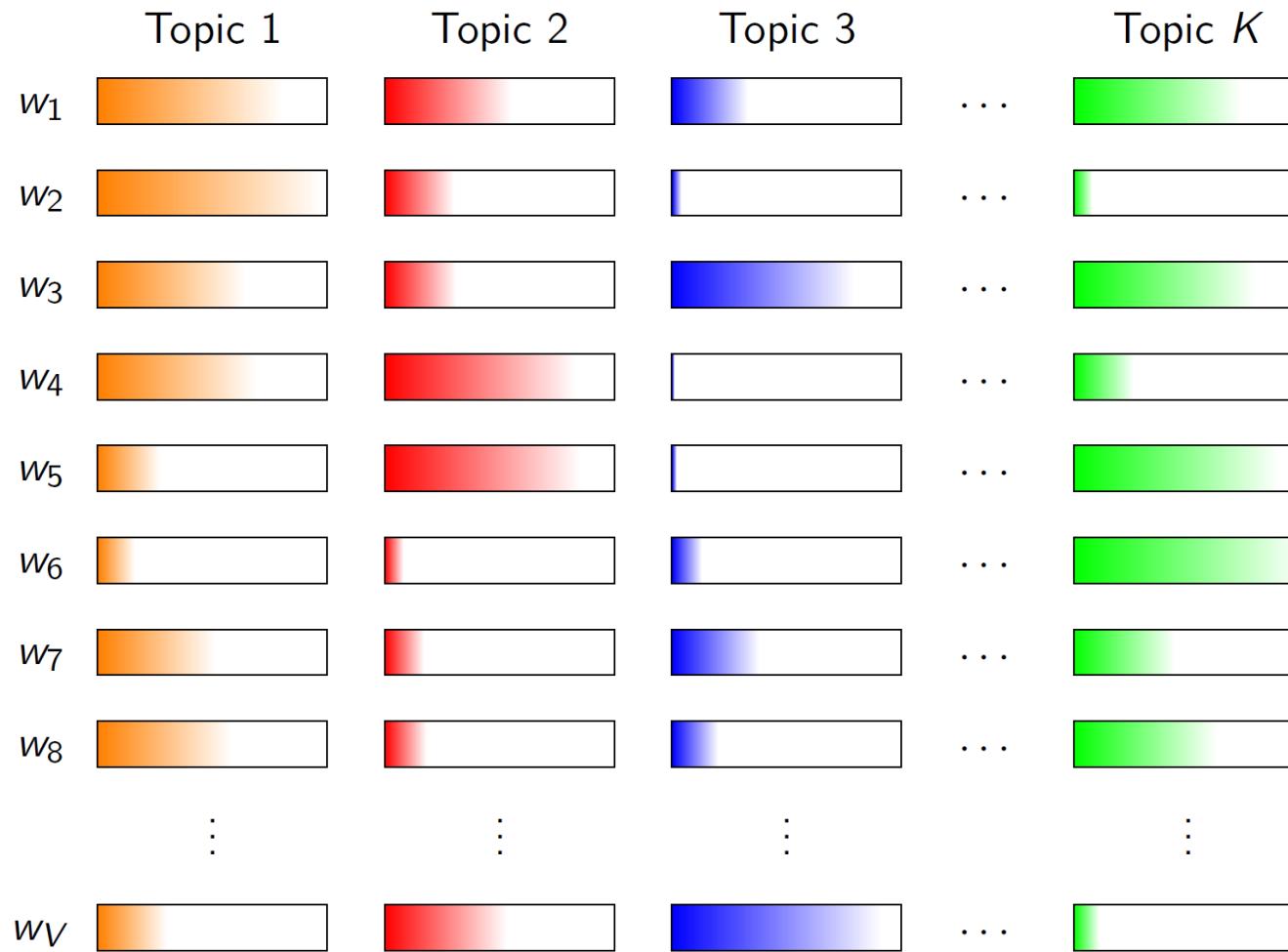
✓ Term distribution per topic

		Term 1	Term 2	Term 3	Term 4	Term 5=V
Topic 1	$\phi_{k=1}$	0.1	0.1	0	0.7	0.1
Topic 2	$\phi_{k=2}$	0.2	0.1	0.2	0.2	0.3
Topic 3	$\phi_{k=3}$	0.01	0.2	0.39	0.3	0.1
Topic 4	$\phi_{k=4=K}$	0.0	0.0	0.5	0.3	0.2

# LDA: Document generation process

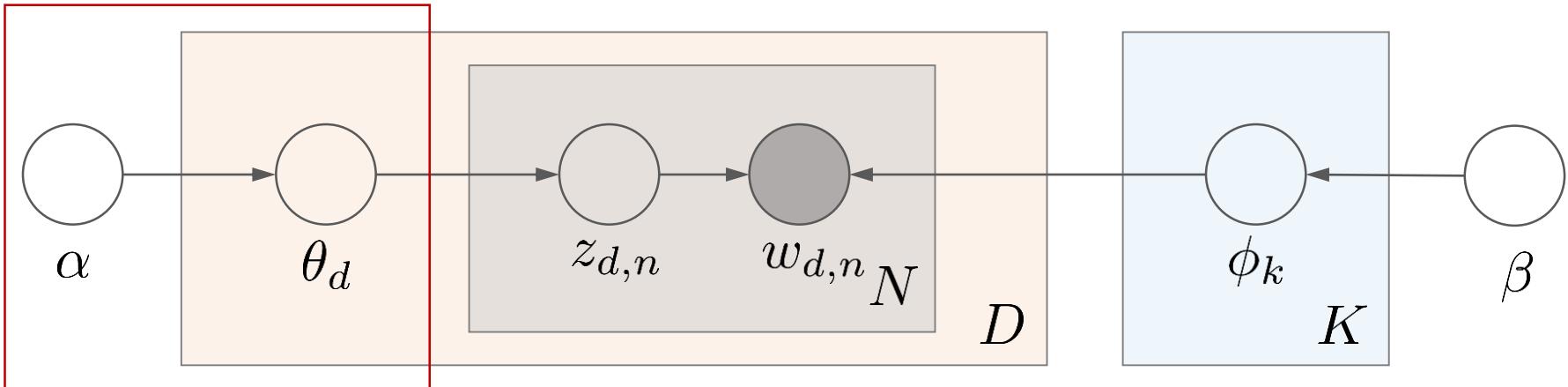
- Document generation process

- ✓ Term distribution per topic



# LDA: Document generation process

- Document generation process



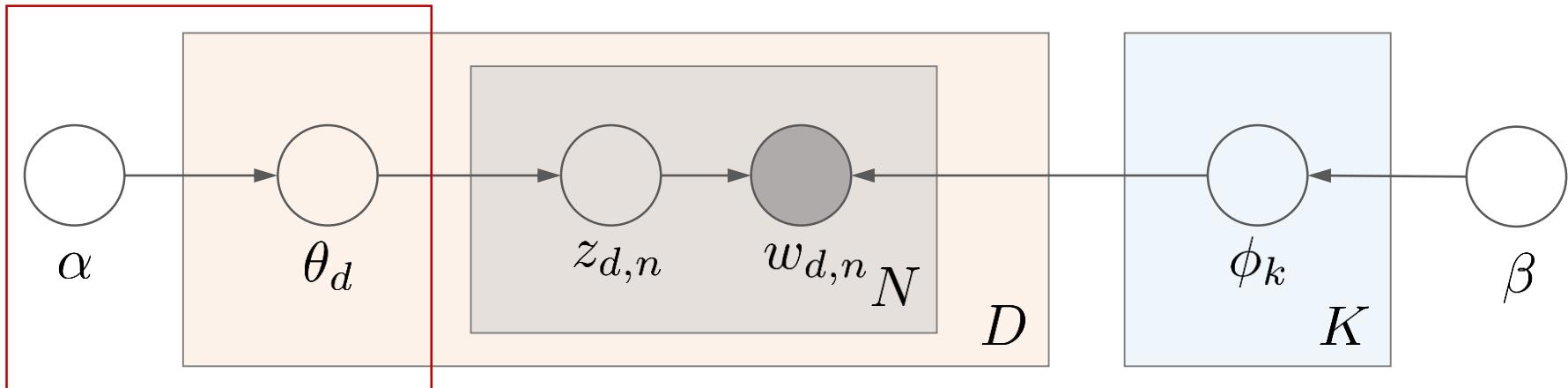
✓ Topic distribution per document

- Drawn from the Dirichlet distribution, given the Dirichlet parameter  $\alpha$ , which is a K-vector with components  $\alpha_k > 0$

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \theta_1^{\alpha_1-1} \dots \theta_K^{\alpha_K-1} = \frac{\Gamma(\alpha.)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k-1}$$

# LDA: Document generation process

- Document generation process

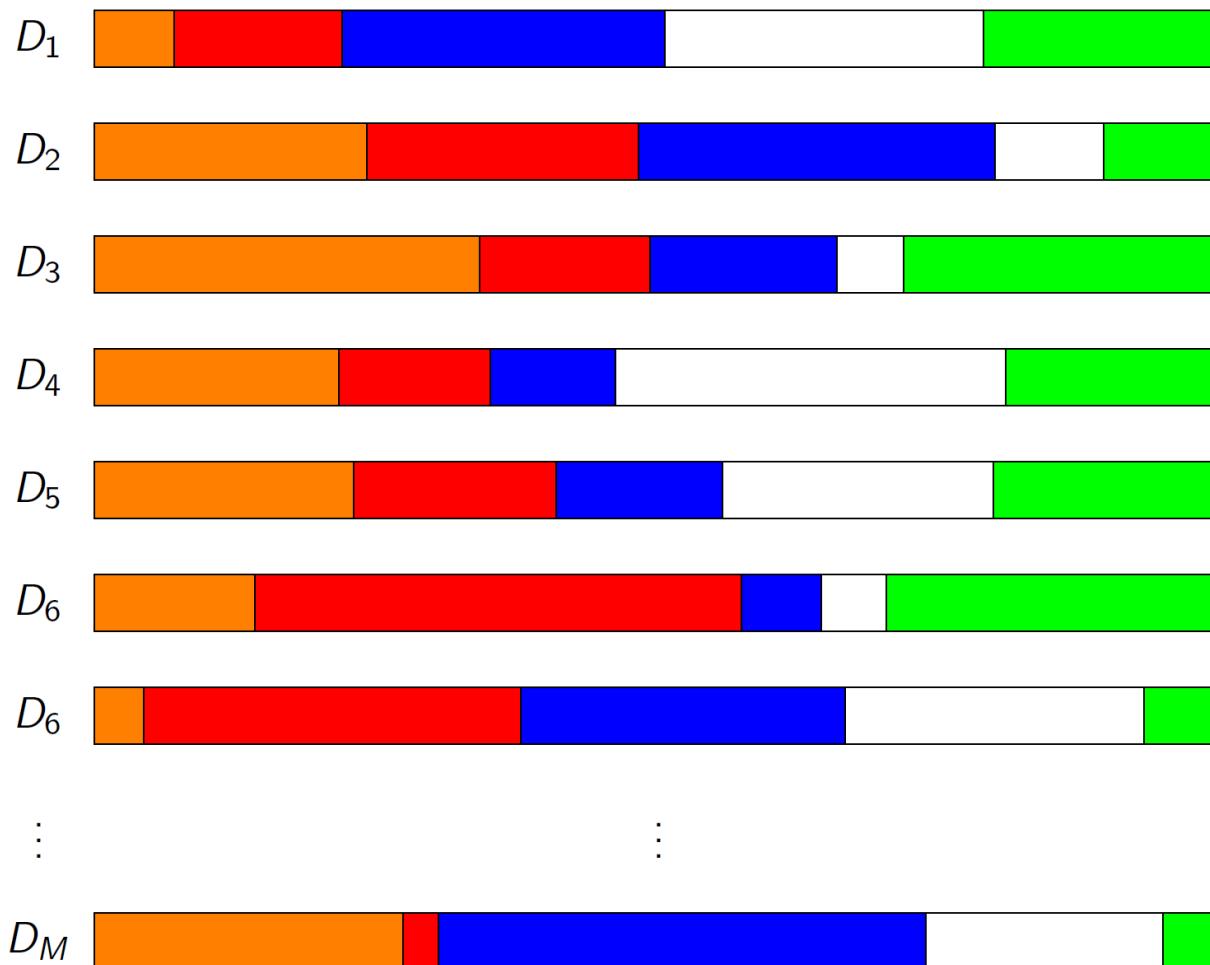


✓ Topic distribution per document

		Topic 1	Topic 2	Topic 3	Topic 4
Document 1	$\theta_{d=1}$	0.5	0.1	0.3	0.1
Document 2	$\theta_{d=2}$	0.0	0.9	0.1	0.0
Document 3	$\theta_{d=3=D}$	0.02	0.48	0.25	0.25

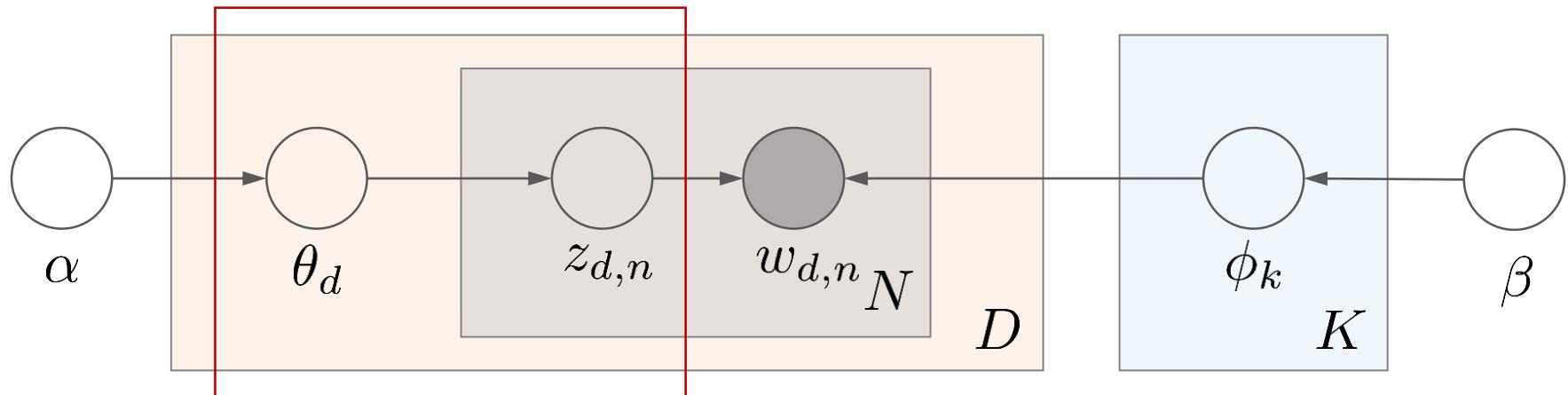
# LDA: Document generation process

- Document generation process
  - ✓ Topic distribution per document



# LDA: Document generation process

- Document generation process



✓ Topic to words assignments

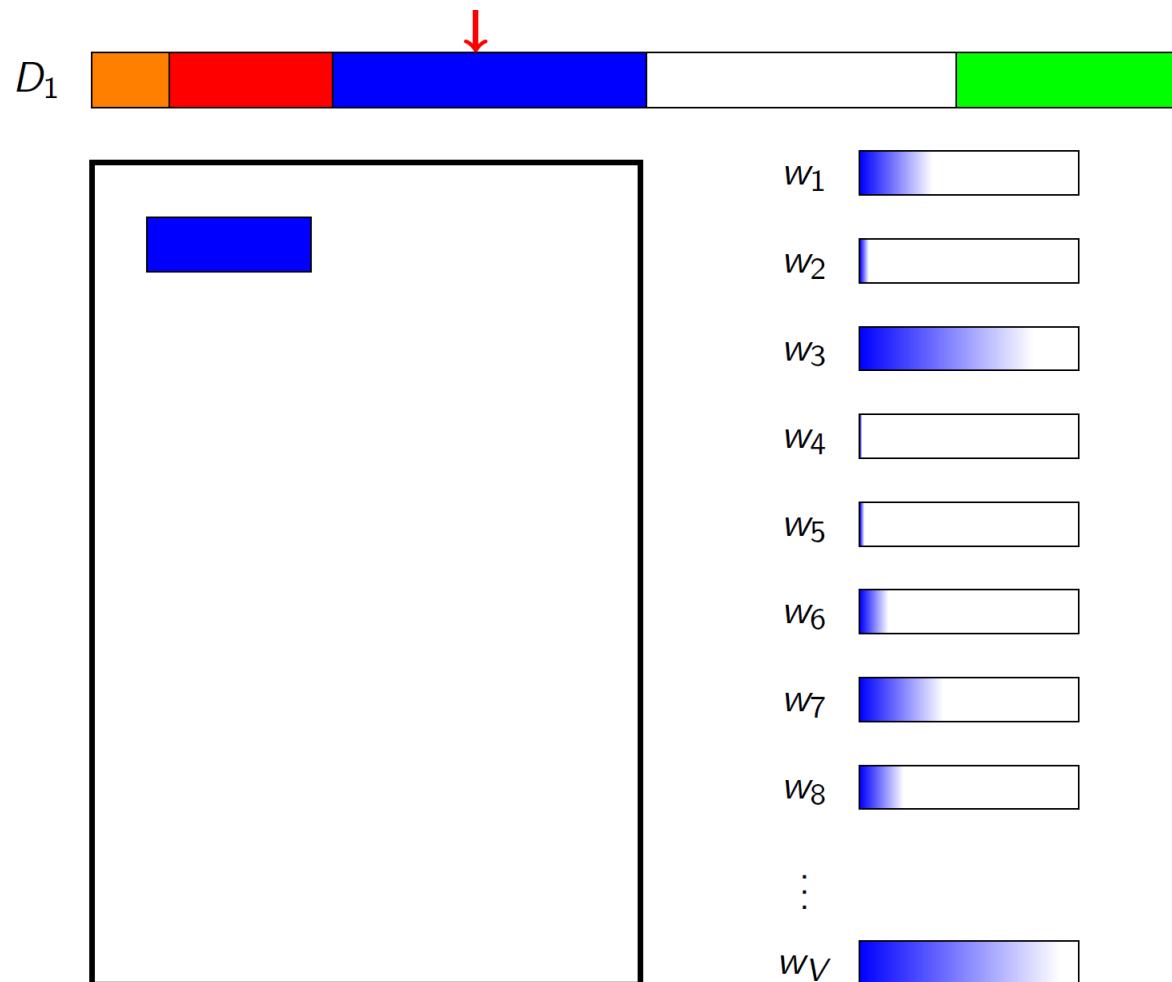
		Word $w_1$	Word $w_2$	Word $w_3$	Word $w_4$	Word $w_5$	Word $w_6$
Document 1	$z_{d=1}$	Topic k=2	Topic k=1	Topic k=1	Topic k=4	Topic k=3	Topic k=3
Document 2	$z_{d=2}$	Topic k=2	Topic k=3	Topic k=2	Topic k=2		
Document 3	$z_{d=3=D}$	Topic k=4	Topic k=2	Topic k=2	Topic k=4	Topic k=3	

$$p(z|\theta) = \prod_{d=1}^D \prod_{k=1}^K \theta_{d,k}^{n_{d,k},.}$$

# LDA: Document generation process

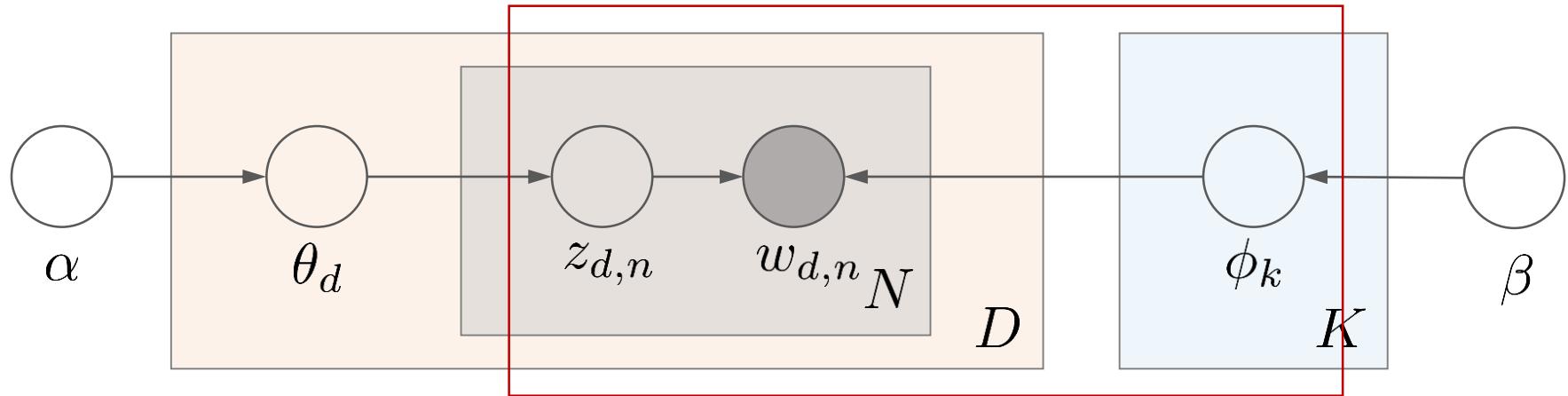
- Document generation process

- ✓ Topic to words assignments



# LDA: Document generation process

- Document generation process



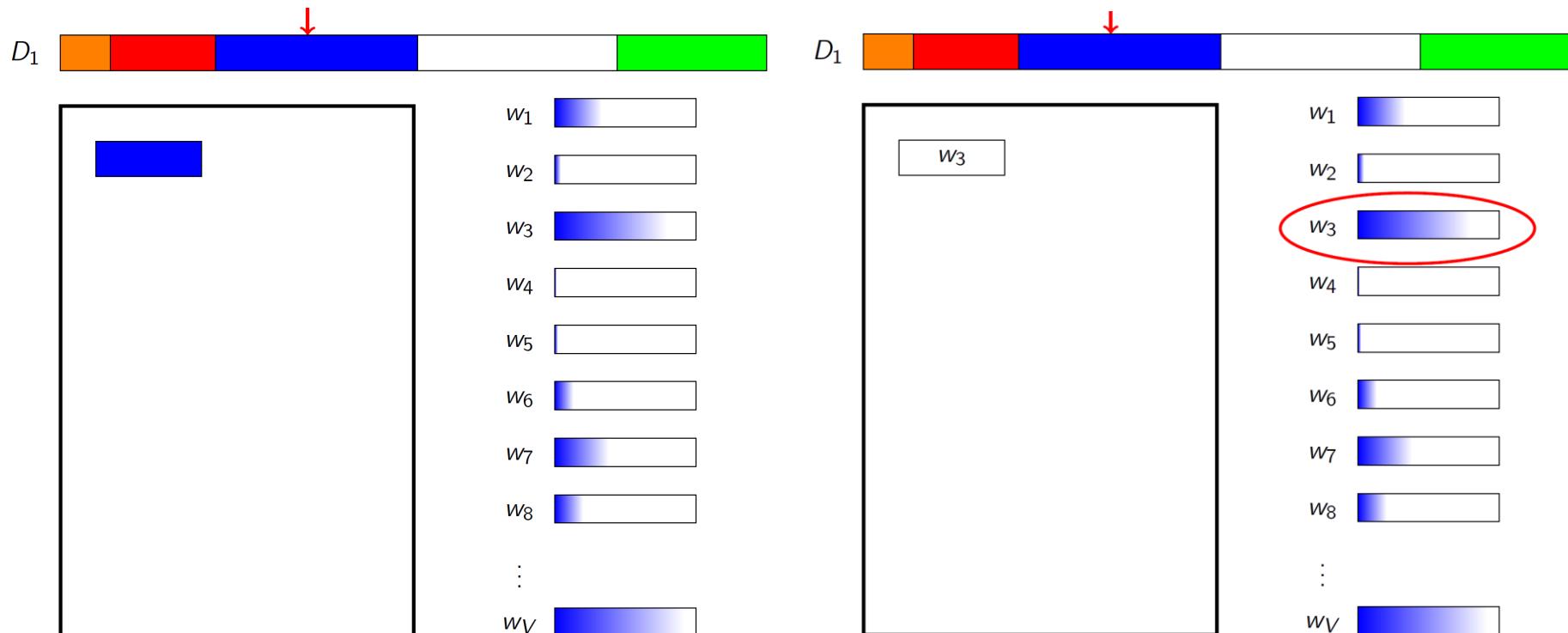
✓ Probability of a corpus

$$p(w|z, \phi) = \prod_{k=1}^K \prod_{v=1}^V \phi_{k,v}^{n_{.,k,v}}$$

# LDA: Document generation process

- Document generation process

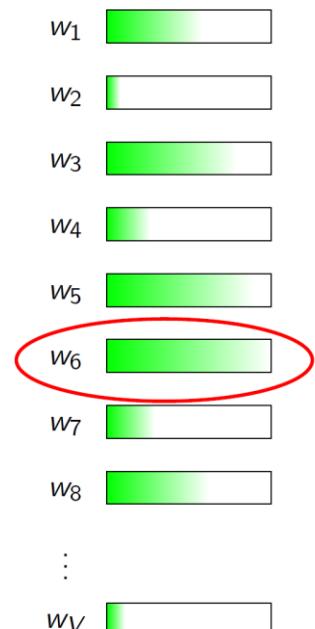
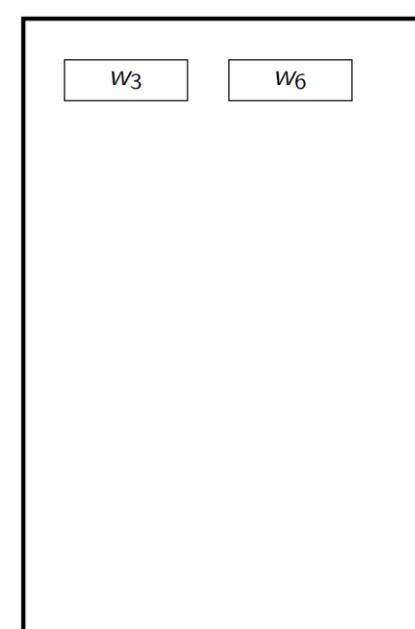
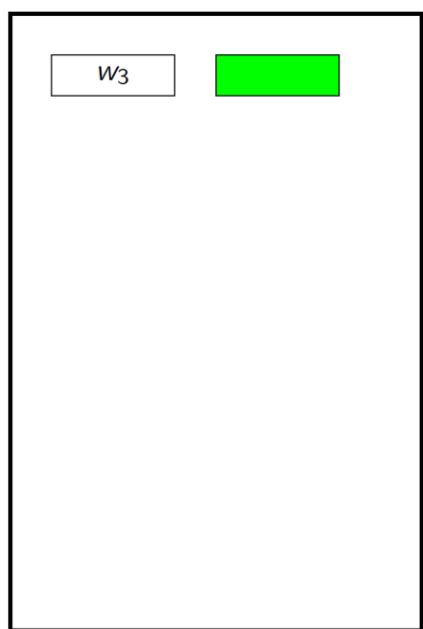
✓ Word selection



# LDA: Document generation process

- Document generation process

✓ Word selection

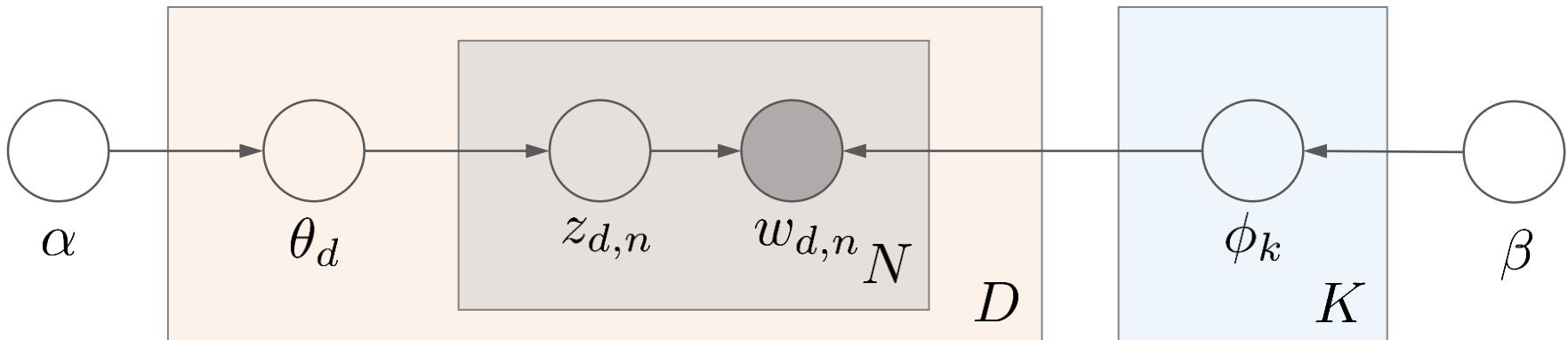


# AGENDA

- 01 Topic Modeling
- 02 Probabilistic Latent Semantic Analysis
- 03 LDA: Document Generation Process
- 04 LDA Inference: Gibbs Sampling
- 05 LDA Evaluation

# LDA Inference

- LDA structure



$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \\ \prod_{i=1}^K p(\phi_i | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \phi_{1:K}, z_{d,n}) \right)$$

✓ From a collection of documents, we infer

- Per-word topic assignment  $z_{d,n}$
- Per-document topic proportions  $\theta_d$
- Per-corpus topic distributions  $\phi_k$

# LDA Inference

- Inference

- ✓ The posterior of the latent variables given the document is

$$p(\phi, \theta, \mathbf{z} | \mathbf{w}) = \frac{p(\phi, \theta, \mathbf{z}, \mathbf{w})}{\int_{\phi} \int_{\theta} \sum_{\mathbf{z}} p(\phi, \theta, \mathbf{z}, \mathbf{w})}$$

- ✓ Computing the posterior is intractable (we cannot compute the denominator, the marginal  $p(\mathbf{w})$ )

- ✓ Approximate posterior inference algorithms

- Mean field variational methods
- Expectation propagation
- Collapsed Gibbs sampling
- Collapsed variational inference
- Online variational inference

# LDA: Dirichlet Distribution

- Binomial & Multinomial

✓ **Binomial distribution:** the number of successes in a sequence of independent yes/no experiments (Bernoulli trials)

$$p(X = x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

✓ **Multinomial distribution:** suppose that each experiment results in one of **k possible outcomes** with probabilities  $p_1, \dots, p_k$ , multinomial models the distribution of the histogram vector which indicates how many time each outcome was observed over N trials of experiments.

$$p(x_1, \dots, x_k | n, p_1, \dots, p_k) = \frac{N!}{\prod_{i=1}^k x_i!} p_i^{x_i}, \quad \sum_i x_i = N, \quad x_i \geq 0$$

# LDA: Dirichlet Distribution

- Beta distribution

$$p(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

- ✓  $p \in [0, 1]$ : considering  $p$  as the parameter of a Binomial distribution, we can think of Beta is a “distribution over distributions” (binomials)

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

$$\Gamma(x) = \int_0^\infty \frac{t^{x-1} dt}{\exp(t)}, \quad (x > 0) \quad \Gamma(n) = (n-1)!$$

# LDA: Dirichlet Distribution

- Dirichlet distribution

$$p(P = \{p_i\} | \alpha_i) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i p_i^{\alpha_i - 1}$$

- ✓  $\sum_i p_i = 1, \quad p_i \geq 0$
- ✓ Two parameters
  - the scale (or concentration):  $\sigma = \sum_i \alpha_i$
  - the base measure:  $(\alpha'_1, \dots, \alpha'_k)$ ,  $\alpha'_i = \alpha_i / \sigma$
- ✓ A generalization of Beta
  - Beta is a distribution over binomials (in an interval  $p \in [0, 1]$ )
  - Dirichlet is a distribution over Multinomials (in the so-called simplex  $\sum_i p_i = 1, \quad p_i \geq 0$ )
- ✓ Dirichlet is the conjugate prior of multinomial

# LDA: Dirichlet Distribution

- Important properties of Dirichlet distribution

- ✓ Posterior is also Dirichlet

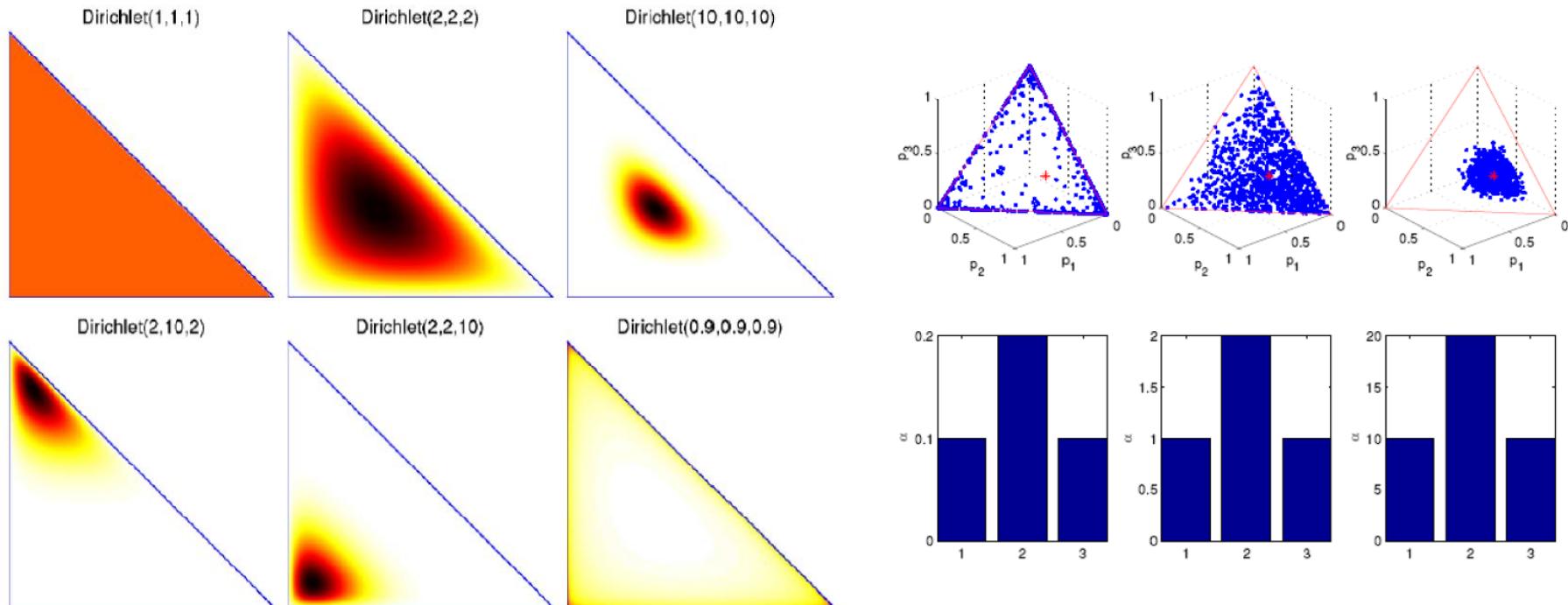
$$p(P = \{p_i\} | \alpha_i) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i p_i^{\alpha_i - 1}$$

$$p(x_1, \dots, x_k | n, p_1, \dots, p_k) = \frac{n!}{\prod_i^k x_i!} p_i^{x_i}$$

$$p(\{p_i\} | x_1, \dots, x_k) = \frac{\Gamma(N + \sum_i \alpha_i)}{\prod_i^k \Gamma(\alpha_i + x_i)} \prod_i p_i^{\alpha_i + x_i - 1}$$

# LDA: Dirichlet Distribution

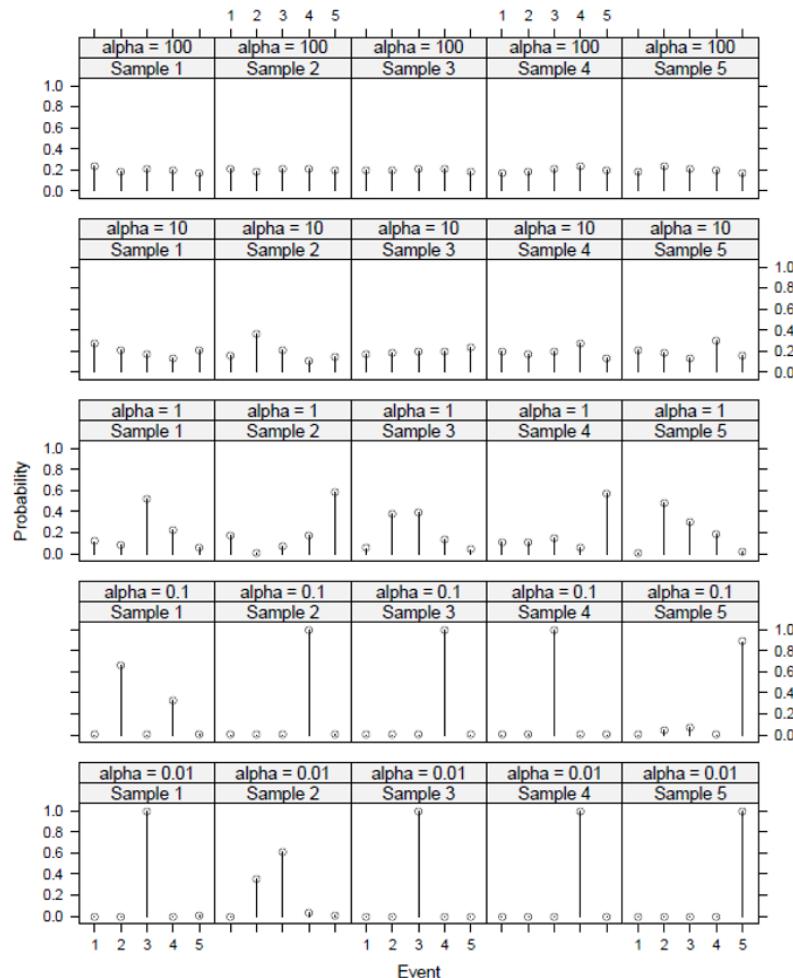
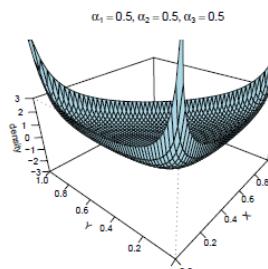
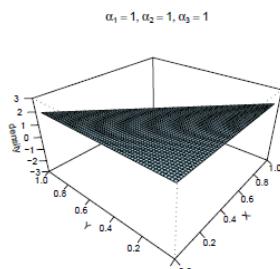
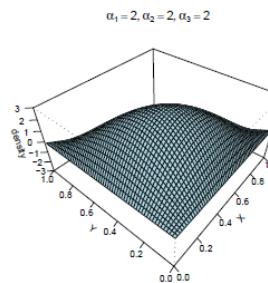
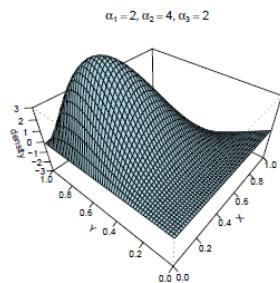
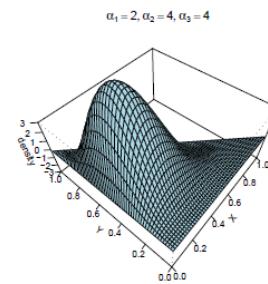
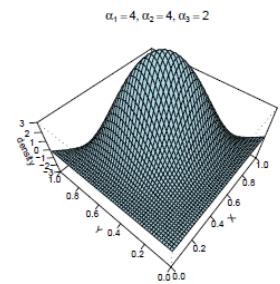
- Important properties of Dirichlet distribution
  - ✓ The parameter  $\alpha$  controls the mean shape and sparsity of  $\theta$
  - ✓ A Dirichlet with  $\alpha_i < 1$  favors extreme distribution



# LDA: Dirichlet Distribution

- Important properties of Dirichlet distribution

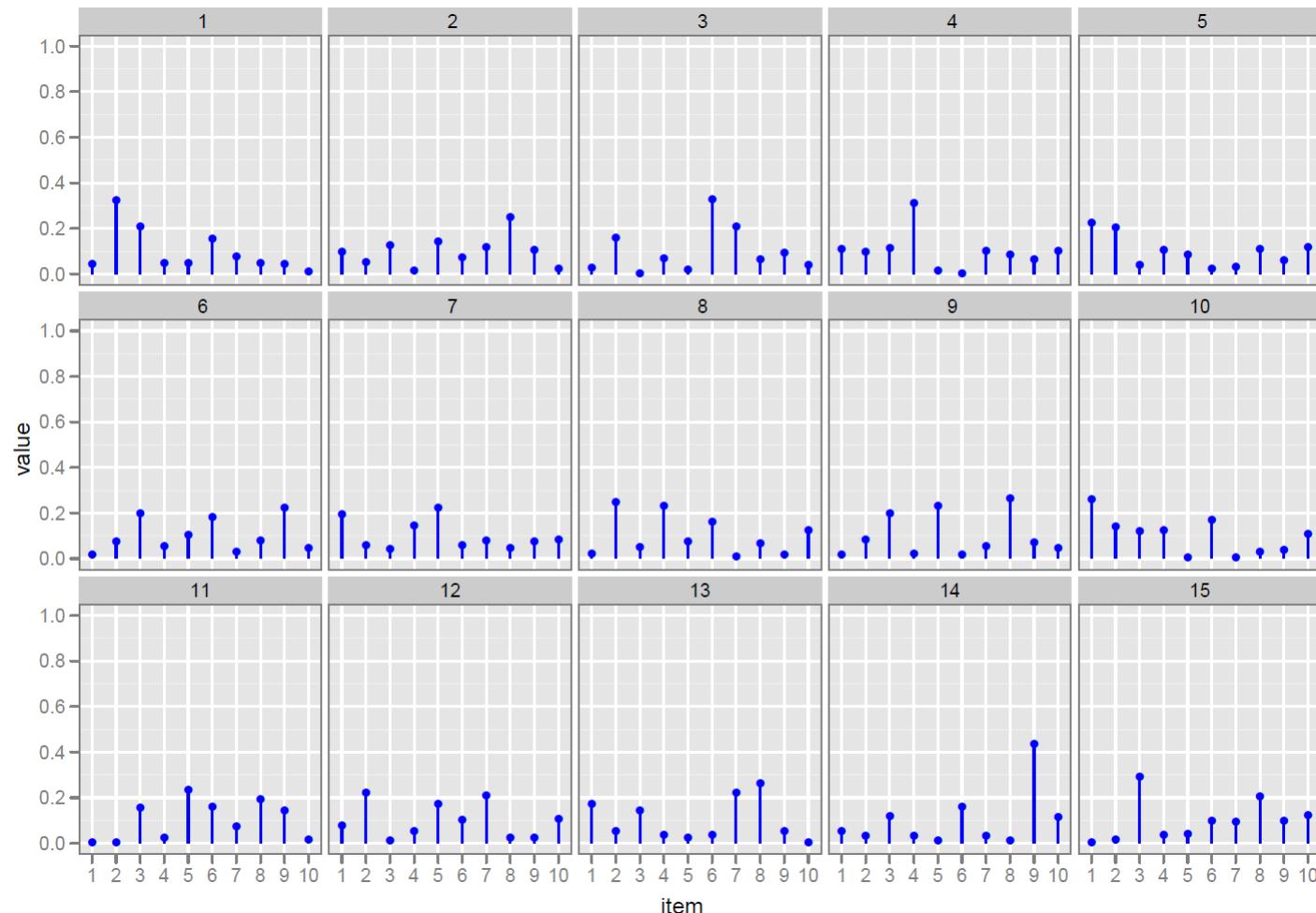
✓ The parameter  $\alpha$  controls the mean shape and sparsity of  $\theta$



# LDA: Dirichlet Distribution

- Sampling results from Dirichlet with different  $\alpha$

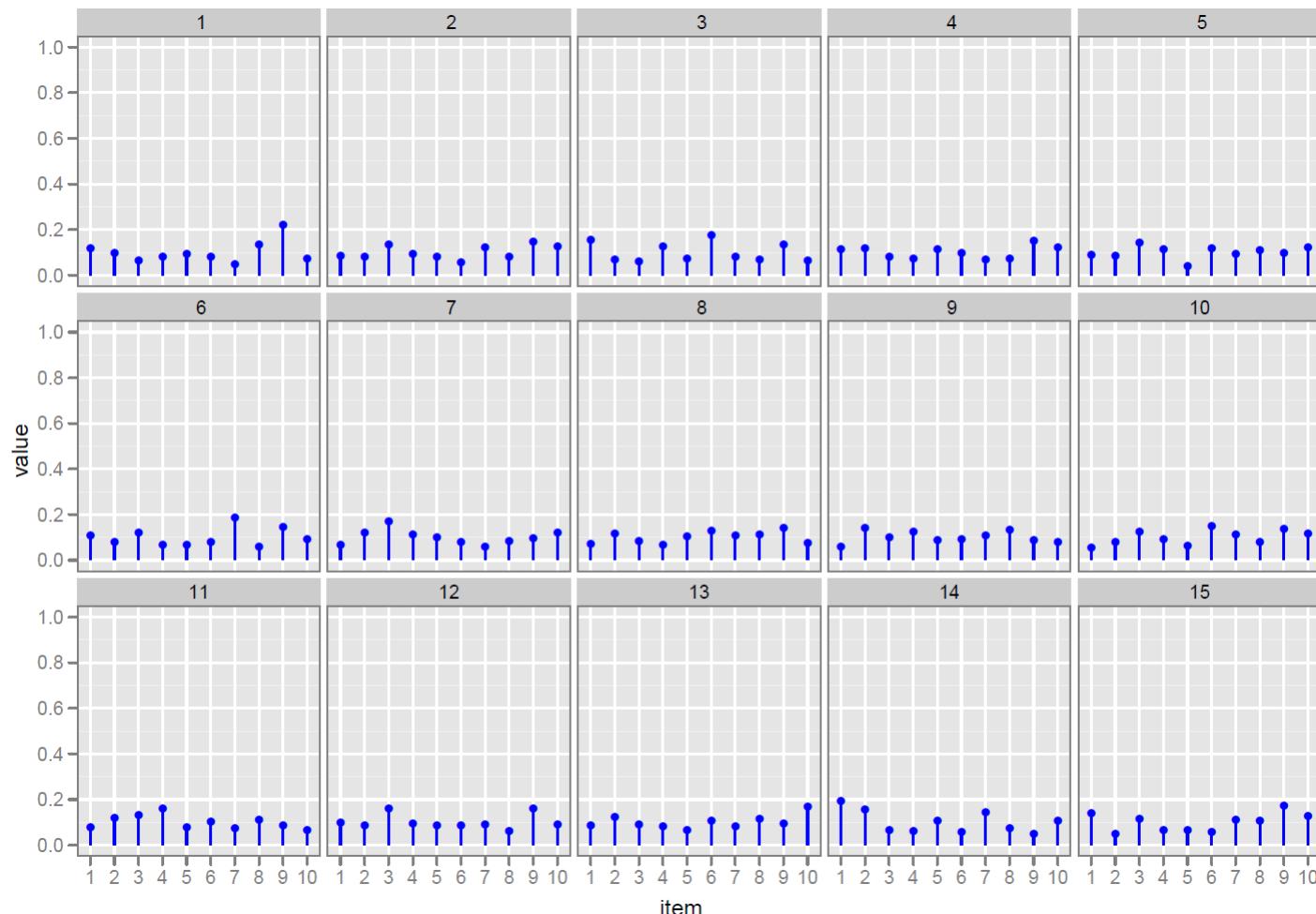
✓  $\alpha = 1$



# LDA: Dirichlet Distribution

- Sampling results from Dirichlet with different  $\alpha$

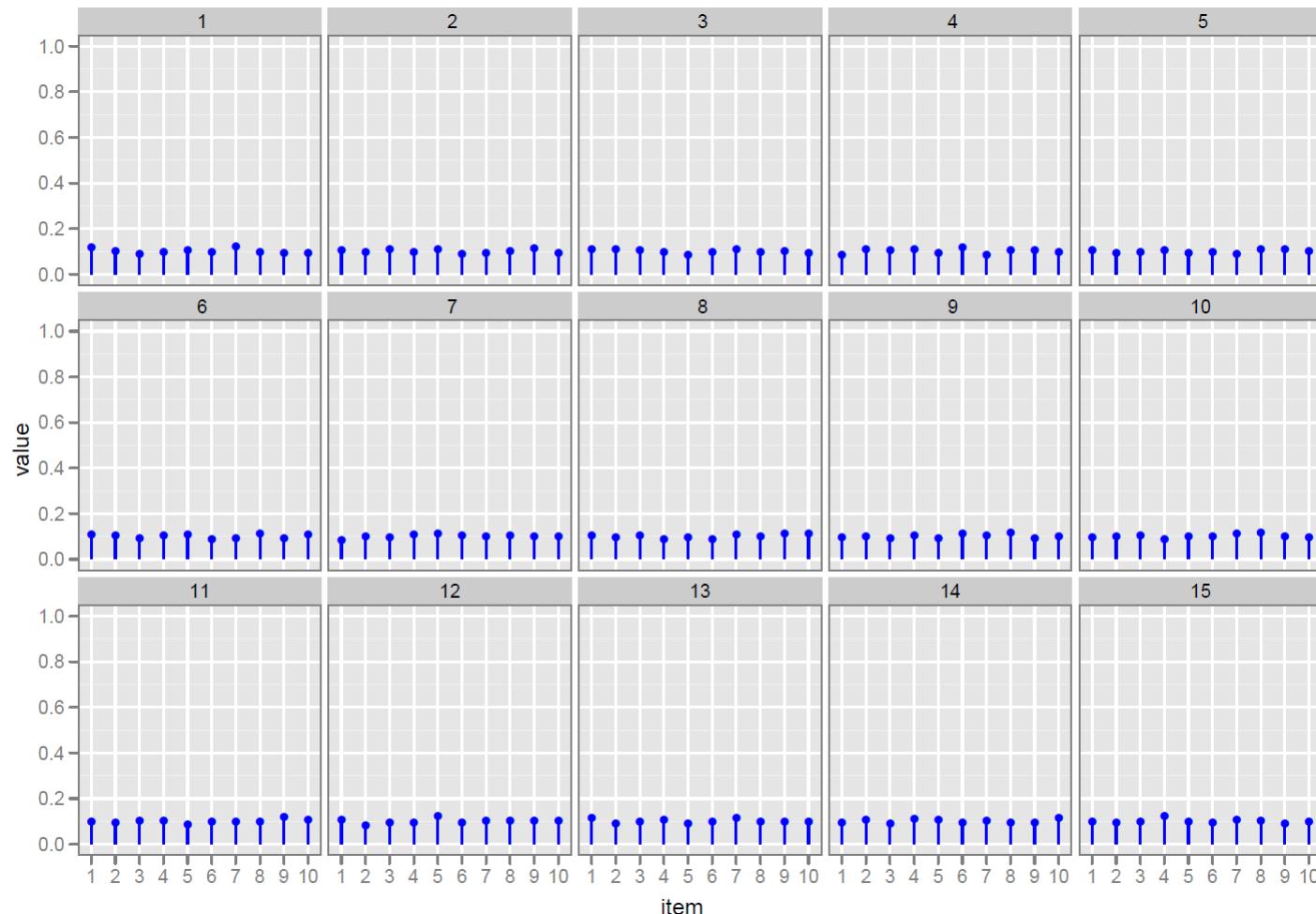
✓  $\alpha = 10$



# LDA: Dirichlet Distribution

- Sampling results from Dirichlet with different  $\alpha$

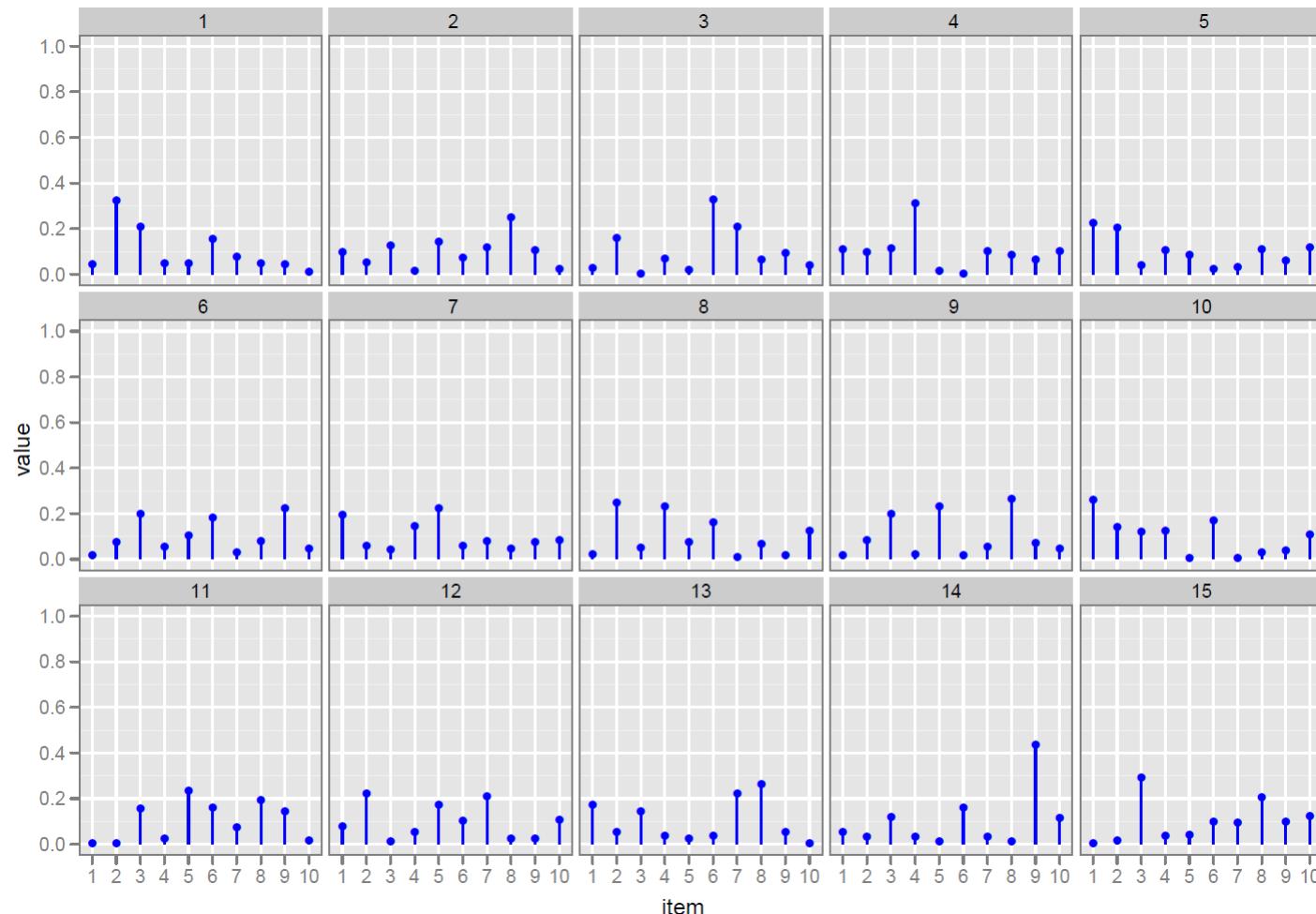
✓  $\alpha = 100$



# LDA: Dirichlet Distribution

- Sampling results from Dirichlet with different  $\alpha$

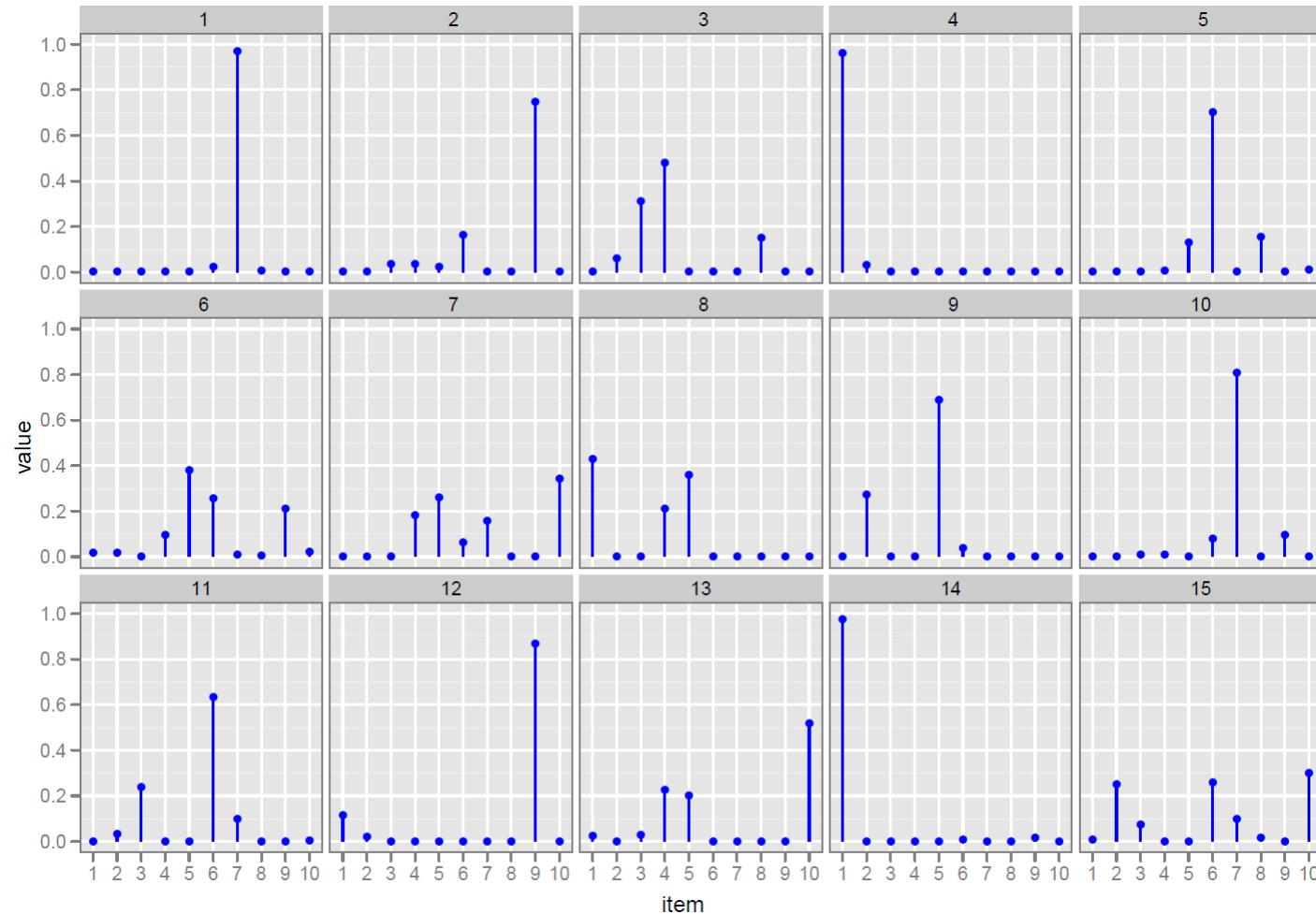
✓  $\alpha = 1$



# LDA: Dirichlet Distribution

- Sampling results from Dirichlet with different  $\alpha$

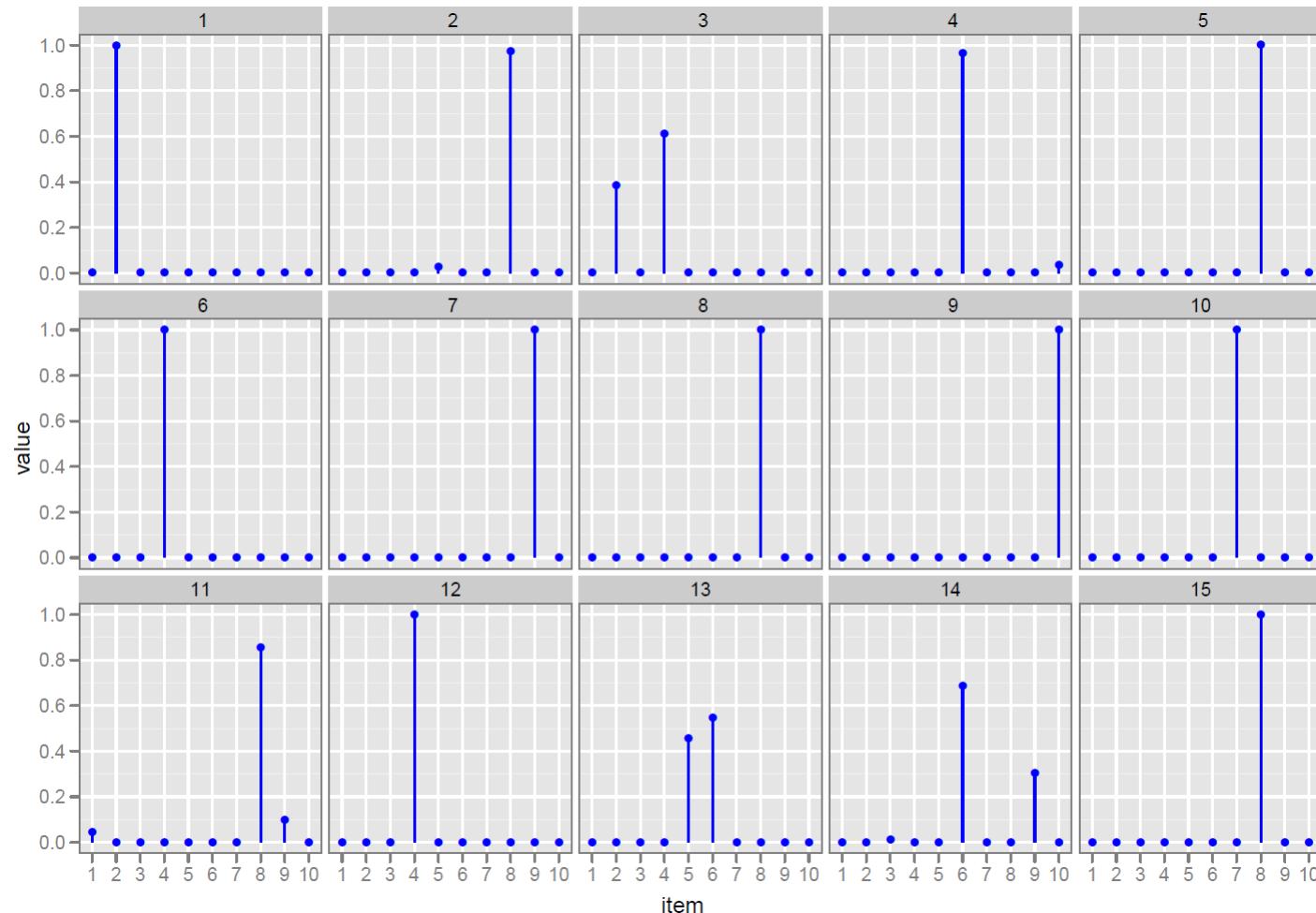
✓  $\alpha = 0.1$



# LDA: Dirichlet Distribution

- Sampling results from Dirichlet with different  $\alpha$

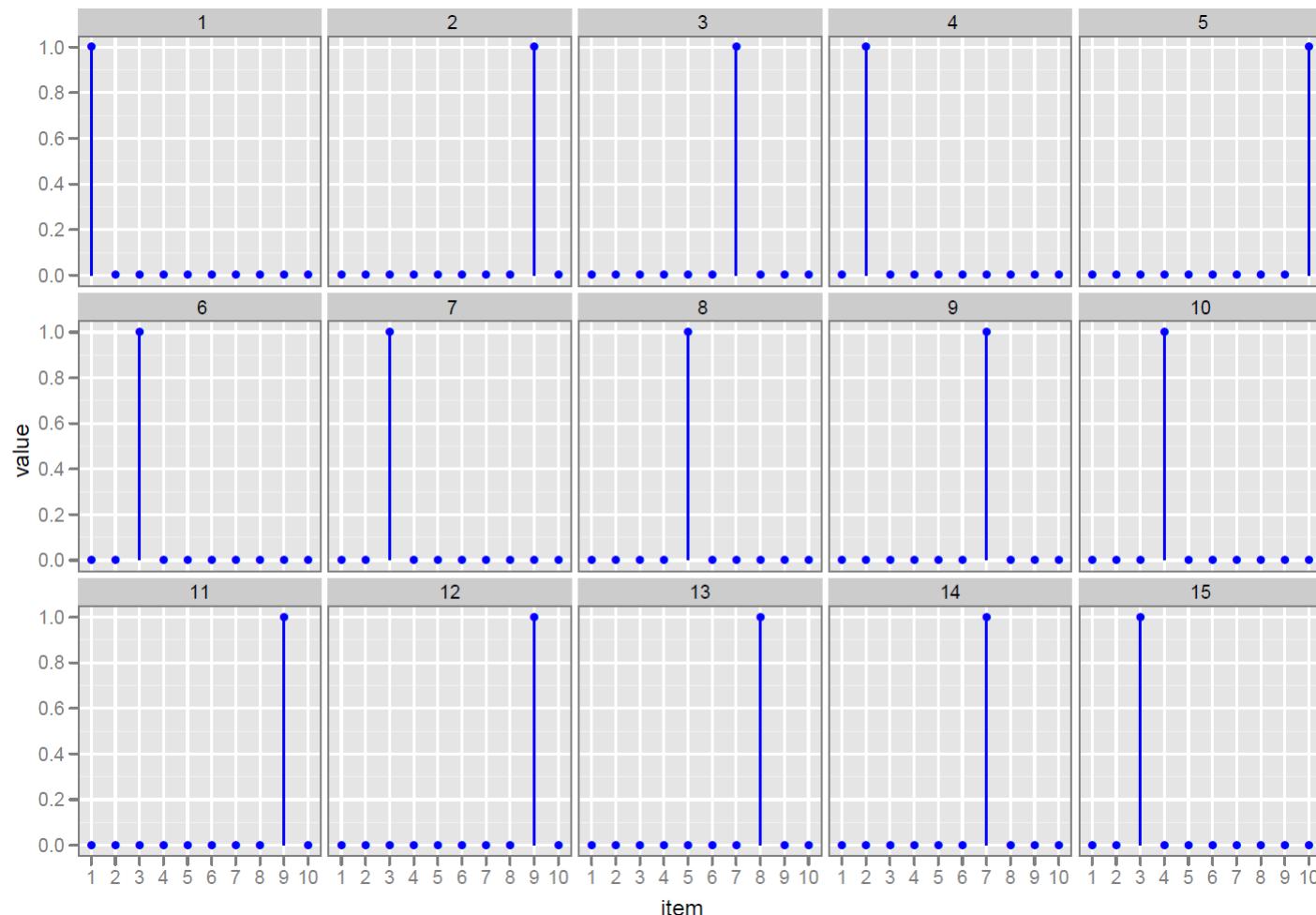
✓  $\alpha = 0.01$



# LDA: Dirichlet Distribution

- Sampling results from Dirichlet with different  $\alpha$

✓  $\alpha = 0.001$



# LDA Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta)$$

- ✓ Here, latent variables are topic assignments  $z$  and topics  $\theta$ .  $X$  is the words (divided into documents), and  $\Theta$  are hyper-parameters to Dirichlet distributions:  $\alpha$  for topic proportions,  $\beta$  for topics

$$p(\mathbf{z}, \phi, \theta | \mathbf{w}, \alpha, \beta)$$

$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

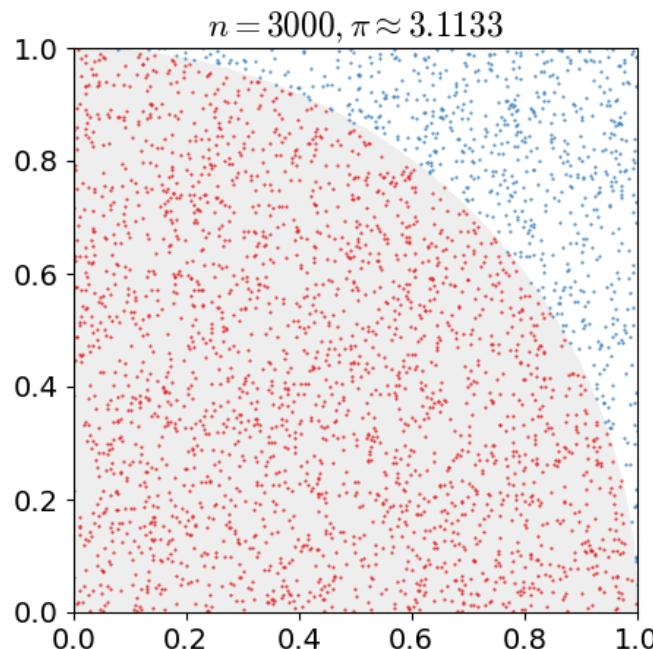
$$\prod_{i=1}^K p(\phi_i | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \phi_{1:K}, z_{d,n}) \right)$$

# LDA Inference

- Gibbs Sampling
  - ✓ A form of Markov Chain Monte Carlo
  - ✓ Chain is a sequence of random variable states
  - ✓ Given a state  $\{z_1, \dots, z_N\}$  given certain technical conditions, drawing  $z_k \sim p(z_1, \dots, z_{k-1}, z_{k+1}, \dots, z_N | X, \Theta)$  for all k (repeatedly) results in a Markov Chain whose stationary distribution is the posterior
  - ✓ For notational call  $\mathbf{z}$  with  $z_{d,n}$  removed  $\mathbf{z}_{-d,n}$

# LDA Inference

- Monte Carlo method
  - ✓ Computational algorithms that rely on repeated random sampling to obtain numerical results
  - ✓ Use randomness to solve problems that might be deterministic in principle
  - ✓ Example: approximating the value of  $\pi$



# LDA Inference

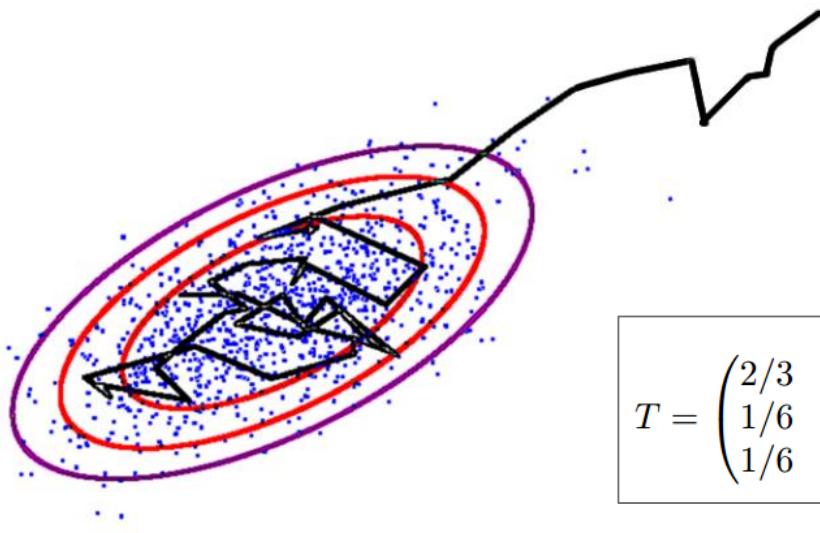
Murray (2009)

- Markov Chain Monte Carlo sampling

- ✓ Sampling from a probability distribution based on constructing a Markov Chain that has the desired distribution as its equilibrium distribution
- ✓ Use the local information rather than a complete randomness

**Construct a biased random walk that explores target dist  $P^*(x)$**

Markov steps,  $x_t \sim T(x_t \leftarrow x_{t-1})$



MCMC gives approximate, correlated samples from  $P^*(x)$

# LDA Inference

Murray (2009)

- Gibbs Sampling

A method with no rejections:

- Initialize  $\mathbf{x}$  to some value
- Pick each variable in turn or randomly and resample  $P(x_i|\mathbf{x}_{j \neq i})$

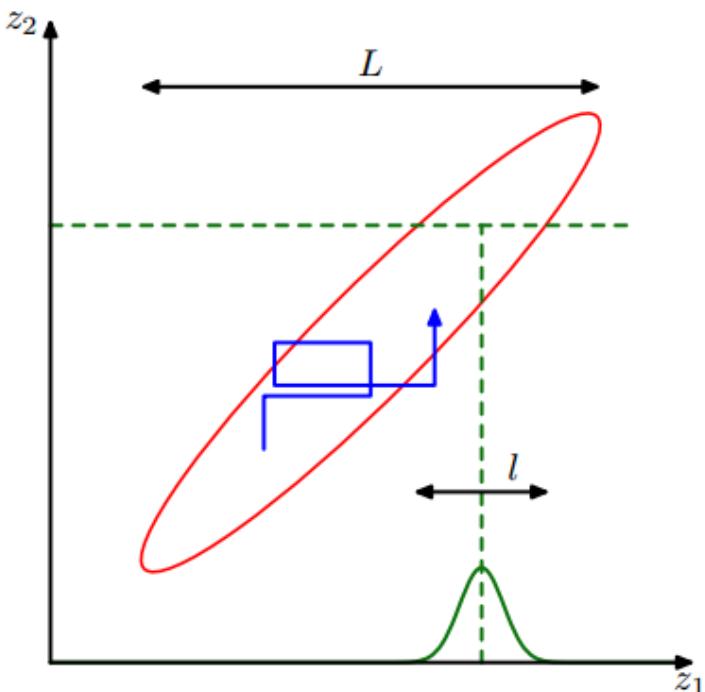
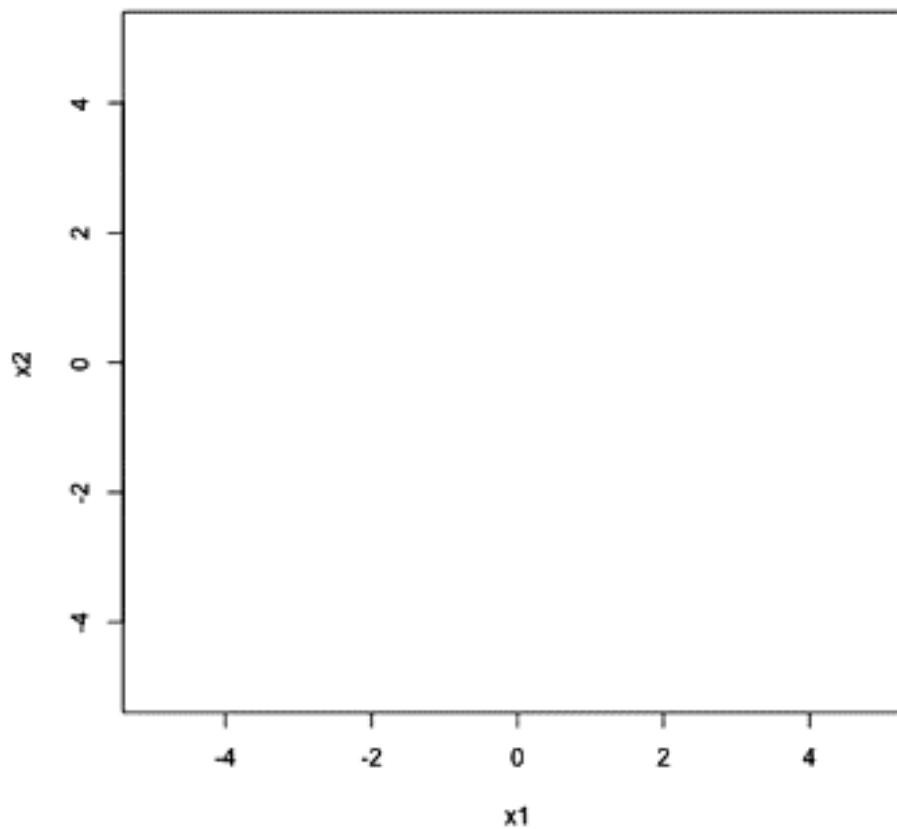


Figure from PRML, Bishop (2006)

# LDA Inference

- Gibbs Sampling



# LDA Inference

Tang (2008)

- Gibbs Variants

- ✓ Gibbs Sampling

- Draw a conditioned on b, c
    - Draw b conditioned on a, c
    - Draw c conditioned on a, b

- ✓ Block Gibbs Sampling

- Draw a, b conditioned on c
    - Draw c, conditioned on a, b

- ✓ Collapsed Gibbs Sampling

- Draw a conditioned on c
    - Draw c conditioned on a
    - b is collapsed out during the sampling process

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs sampling procedure boils down to estimate

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w})$$

- ✓  $\theta$  and  $\phi$  are integrated out

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs sampling procedure boils down to estimate

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w})$$

- ✓  $\theta$  and  $\phi$  are integrated out

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

# LDA Inference: Collapsed Gibbs Sampling

Tang (2008)

- Conditional posterior distribution for  $z_i$  is given by

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto p(z_i = j, \mathbf{z}_{-i}, \mathbf{w})$$

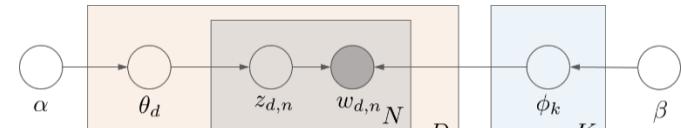
$$= p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}_{-i})$$

$$= p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) p(z_i = j | \mathbf{z}_{-i})$$

- The first term is the likelihood and the second term acts like a prior

$$\begin{aligned} p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) &= \int p(w_i | z_i = j, \phi^{(j)}) p(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)} \\ &= \int \phi_{w_i}^{(j)} p(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)} \end{aligned}$$

$$p(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto p(\mathbf{w}_{-i} | \phi^{(j)}, \mathbf{z}_{-i}) p(\phi^{(j)}) \sim Dirichlet(\beta + n_{-i,j}^{(w)})$$



# LDA Inference: Collapsed Gibbs Sampling

Tang (2008)

- Conditional posterior distribution for  $z_i$  is given by

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto p(z_i = j, \mathbf{z}_{-i}, \mathbf{w})$$

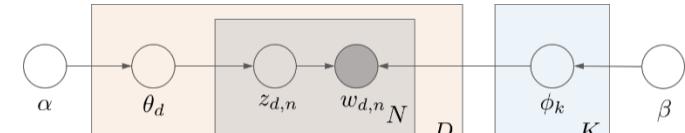
$$= p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}_{-i})$$

$$= p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) p(z_i = j | \mathbf{z}_{-i})$$

- Here,  $n_{-i,j}^{(w)}$  is the number of instances of word w assigned to topic j excluding the current one.
- Using the property of expectation of Dirichlet distribution, we have

$$p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta}$$

- where  $n_{-i,j}^{(\cdot)}$  is the total number of words assigned to topic j excluding the current one in the corpus



# LDA Inference: Collapsed Gibbs Sampling

Tang (2008)

- Conditional posterior distribution for  $z_i$  is given by

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto p(z_i = j, \mathbf{z}_{-i}, \mathbf{w})$$

$$= p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}_{-i})$$

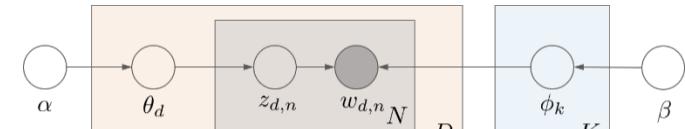
$$= p(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) p(z_i = j | \mathbf{z}_{-i})$$

$$p(z_i = j | \mathbf{z}_{-i}) = \int p(z_i = j | \theta^{(d)}) p(\theta^{(d)} | \mathbf{z}_{-i}) d\theta^{(d)}$$

$$p(\theta^{(d)} | \mathbf{z}_{-i}) \propto p(\mathbf{z}_{-i} | \theta^{(d)}) p(\theta^{(d)}) \sim Dirichlet(n_{-i,j}^{(d)} + \alpha)$$

- ✓ where  $n_{-i,j}^{(d)}$  is the number of words assigned to topic  $j$  excluding current one in the document  $d$ .

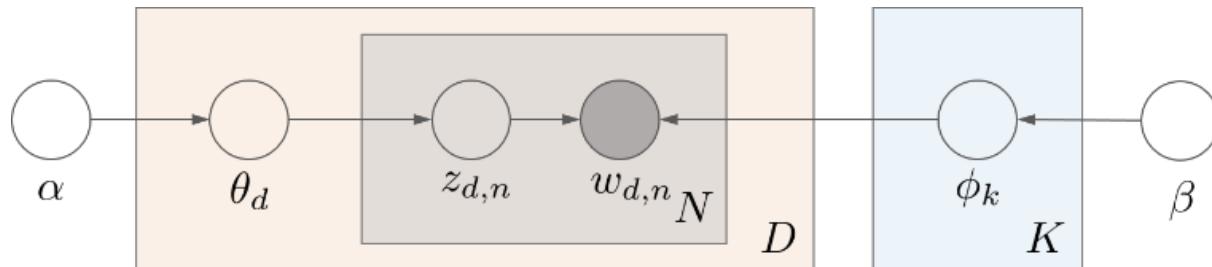
$$p(z_i = j | \mathbf{z}_{-i}) = \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,.}^{(d)} + K\alpha}$$



# LDA Inference: Collapsed Gibbs Sampling

Tang (2008)

- Gibbs Sampling Equation



$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$

✓ Need to record four count variables

- Document-topic count  $n_{-i,j}^{(d)}$
- Document-topic sum  $n_{-i,\cdot}^{(d)}$
- Topic-term count  $n_{-i,j}^{(w_i)}$
- Topic-term sum  $n_{-i,j}^{(\cdot)}$

# LDA Inference: Collapsed Gibbs Sampling

Tang (2008)

- Parameter Estimation
  - ✓ To obtain  $\theta$  and  $\phi$ , two ways are possible (draw one sample of  $z$  or draw multiple samples of  $z$  to calculate the average)

$$\phi_{j,w} = \frac{n_w^{(j)} + \beta}{\sum_{w=1}^V n_w^{(j)} + V\beta} \quad \theta_j^{(d)} = \frac{n_j^{(d)} + \alpha}{\sum_{z=1}^K n_z^{(d)} + K\alpha}$$

- ✓ where  $n_w^{(j)}$  is the frequency of word assigned to topic  $j$ , and  $n_z^{(d)}$  is the number of word assigned to topic  $z$  in the document  $d$

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs Sampling Equation: Another Form

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$
$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

- ✓ Number of times document d uses topic k
- ✓ Number of times topic k uses word type  $w_{d,n}$
- ✓ Dirichlet parameter for document topic distribution
- ✓ Dirichlet parameter for topic to word distribution
- ✓ How much this document likes topic k
- ✓ How much this topic likes word  $w_{d,n}$

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs Sampling Equation: Another Form

$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$
$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$

- ✓ Number of times document d uses topic k
- ✓ Number of times topic k uses word type  $w_{d,n}$
- ✓ Dirichlet parameter for document topic distribution
- ✓ Dirichlet parameter for topic to word distribution
- ✓ How much this document likes topic k
- ✓ How much this topic likes word  $w_{d,n}$

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs Sampling Equation: Another Form

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$
$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

- ✓ Number of times document d uses topic k
- ✓ Number of times topic k uses word type  $w_{d,n}$
- ✓ Dirichlet parameter for document topic distribution
- ✓ Dirichlet parameter for topic to word distribution
- ✓ How much this document likes topic k
- ✓ How much this topic likes word  $w_{d,n}$

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs Sampling Equation: Another Form

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$
$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

- ✓ Number of times document d uses topic k
- ✓ Number of times topic k uses word type  $w_{d,n}$
- ✓ Dirichlet parameter for document topic distribution
- ✓ Dirichlet parameter for topic to word distribution
- ✓ How much this document likes topic k
- ✓ How much this topic likes word  $w_{d,n}$

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs Sampling Equation: Another Form

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$
$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

- ✓ Number of times document d uses topic k
- ✓ Number of times topic k uses word type  $w_{d,n}$
- ✓ Dirichlet parameter for document topic distribution
- ✓ Dirichlet parameter for topic to word distribution
- ✓ How much this document likes topic k
- ✓ How much this topic likes word  $w_{d,n}$

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs Sampling Equation: Another Form

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$
$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

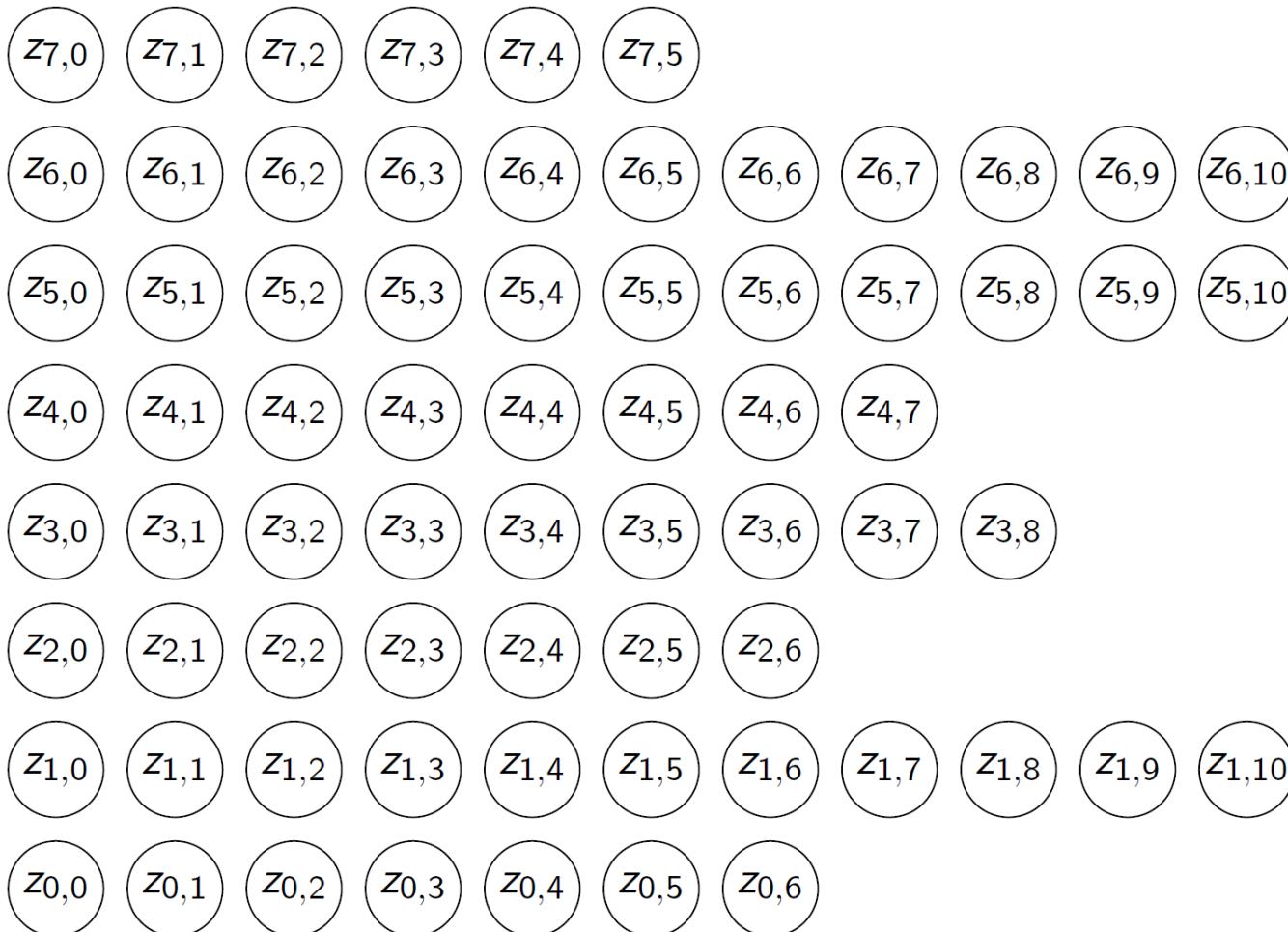
- ✓ Number of times document d uses topic k
- ✓ Number of times topic k uses word type  $w_{d,n}$
- ✓ Dirichlet parameter for document topic distribution
- ✓ Dirichlet parameter for topic to word distribution
- ✓ How much this document likes topic k
- ✓ How much this topic likes word  $w_{d,n}$

# LDA Inference: Collapsed Gibbs Sampling

Speh et al. (2013)

- Collapsed Gibbs Sampling

- ✓ Illustrative procedure

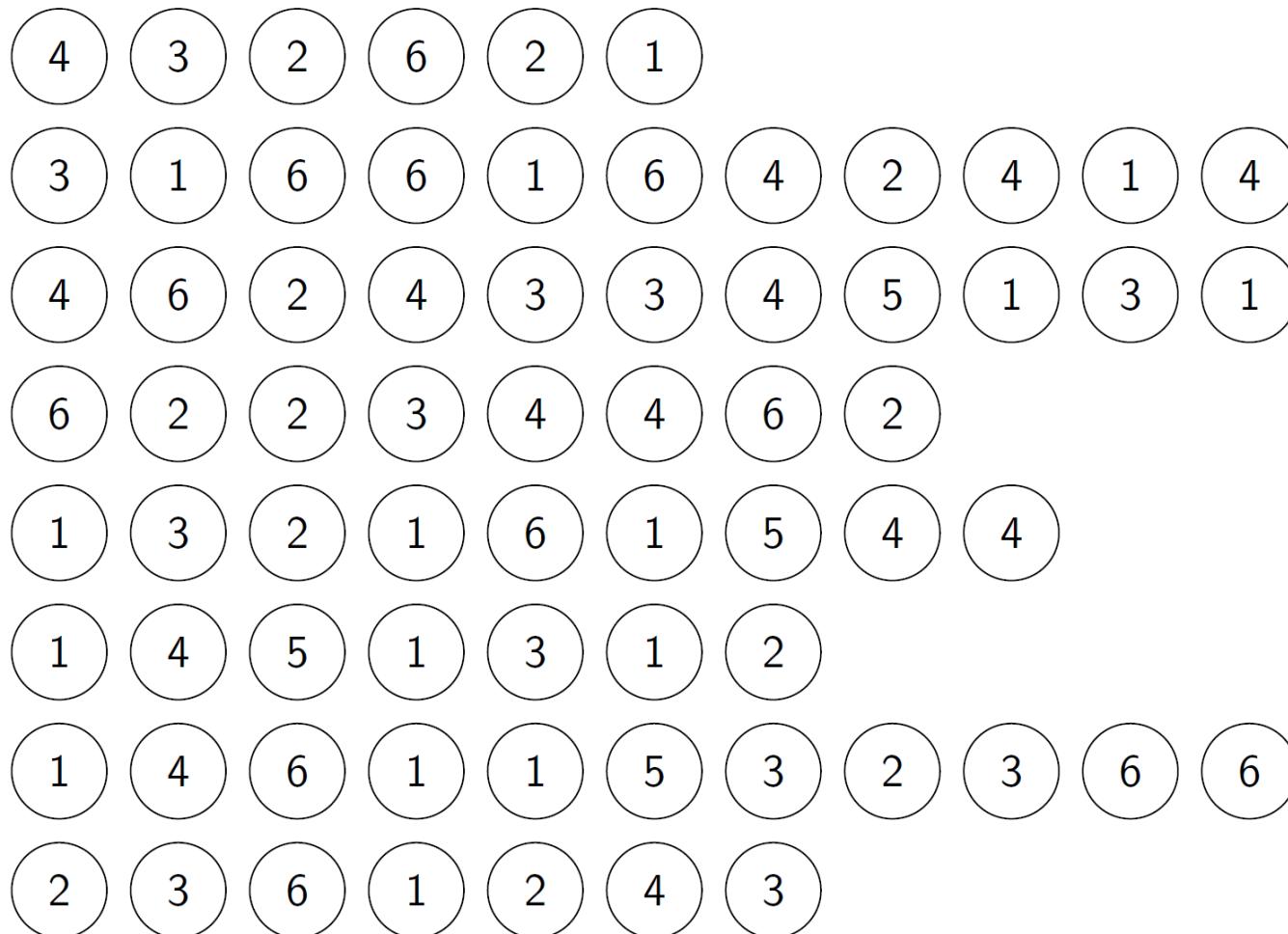


# LDA Inference: Collapsed Gibbs Sampling

Speh et al. (2013)

- Collapsed Gibbs Sampling

- ✓ Illustrative procedure

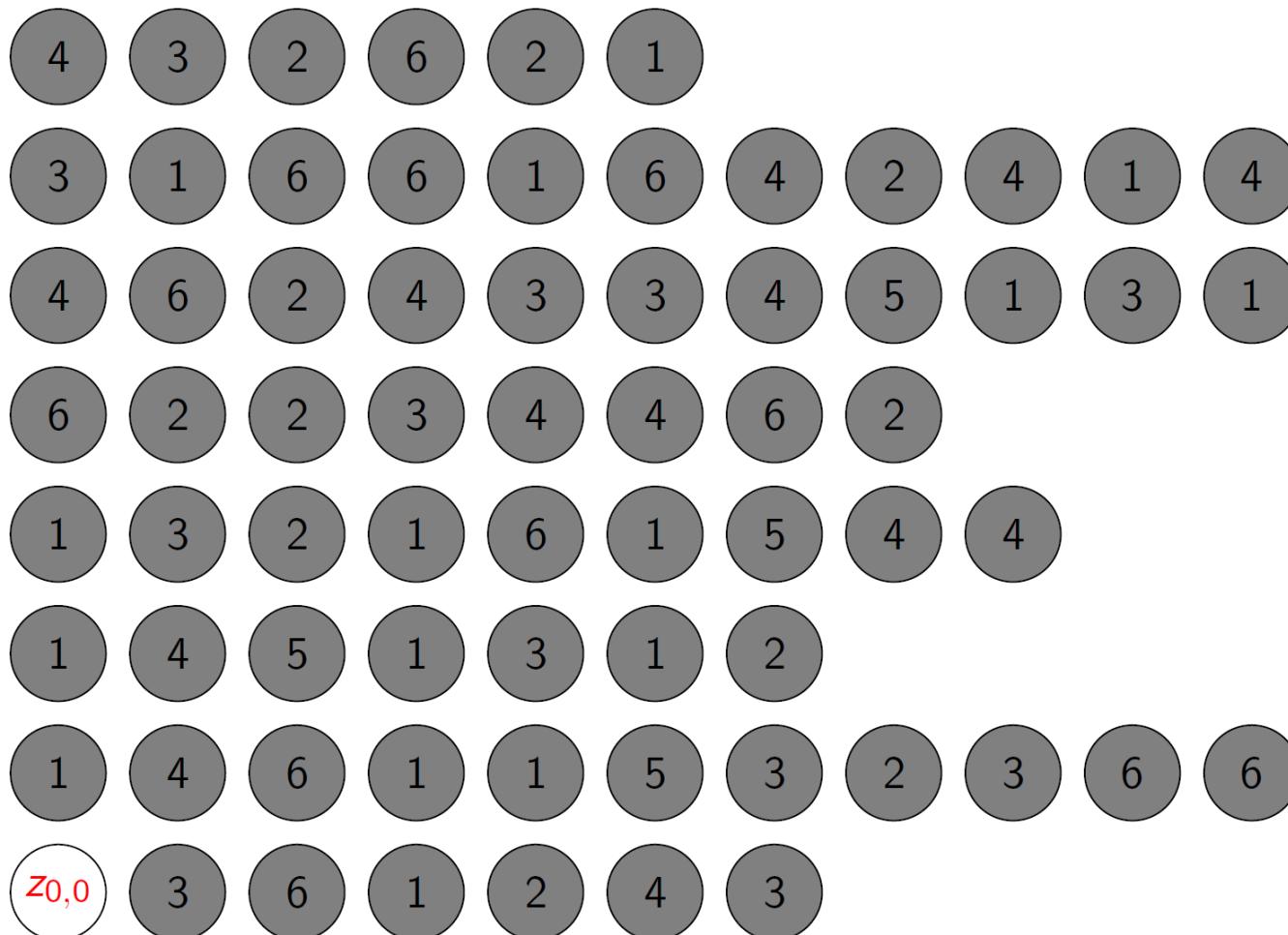


# LDA Inference: Collapsed Gibbs Sampling

Speh et al. (2013)

- Collapsed Gibbs Sampling

- ✓ Illustrative procedure

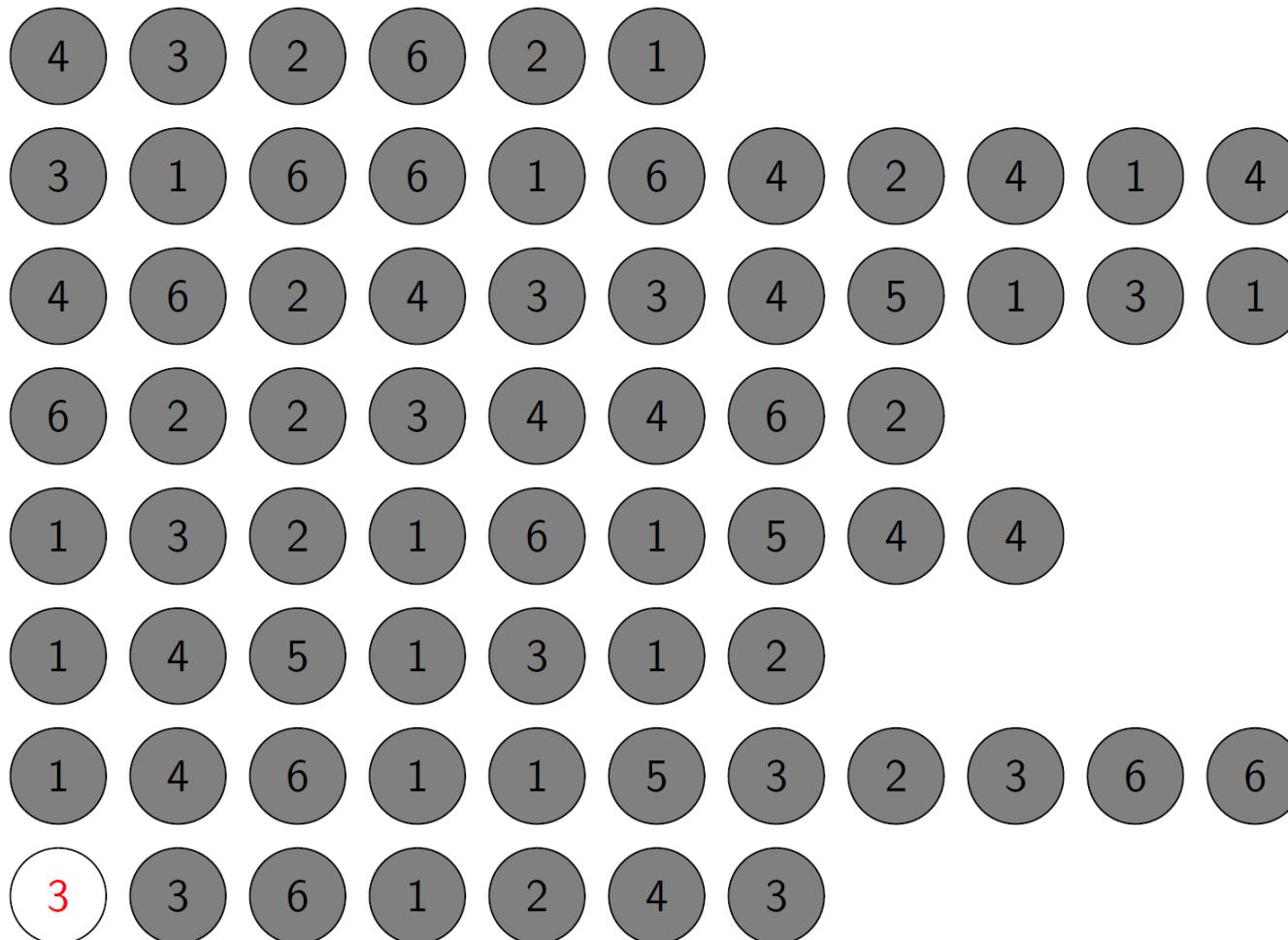


# LDA Inference: Collapsed Gibbs Sampling

Speh et al. (2013)

- Collapsed Gibbs Sampling

- ✓ Illustrative procedure

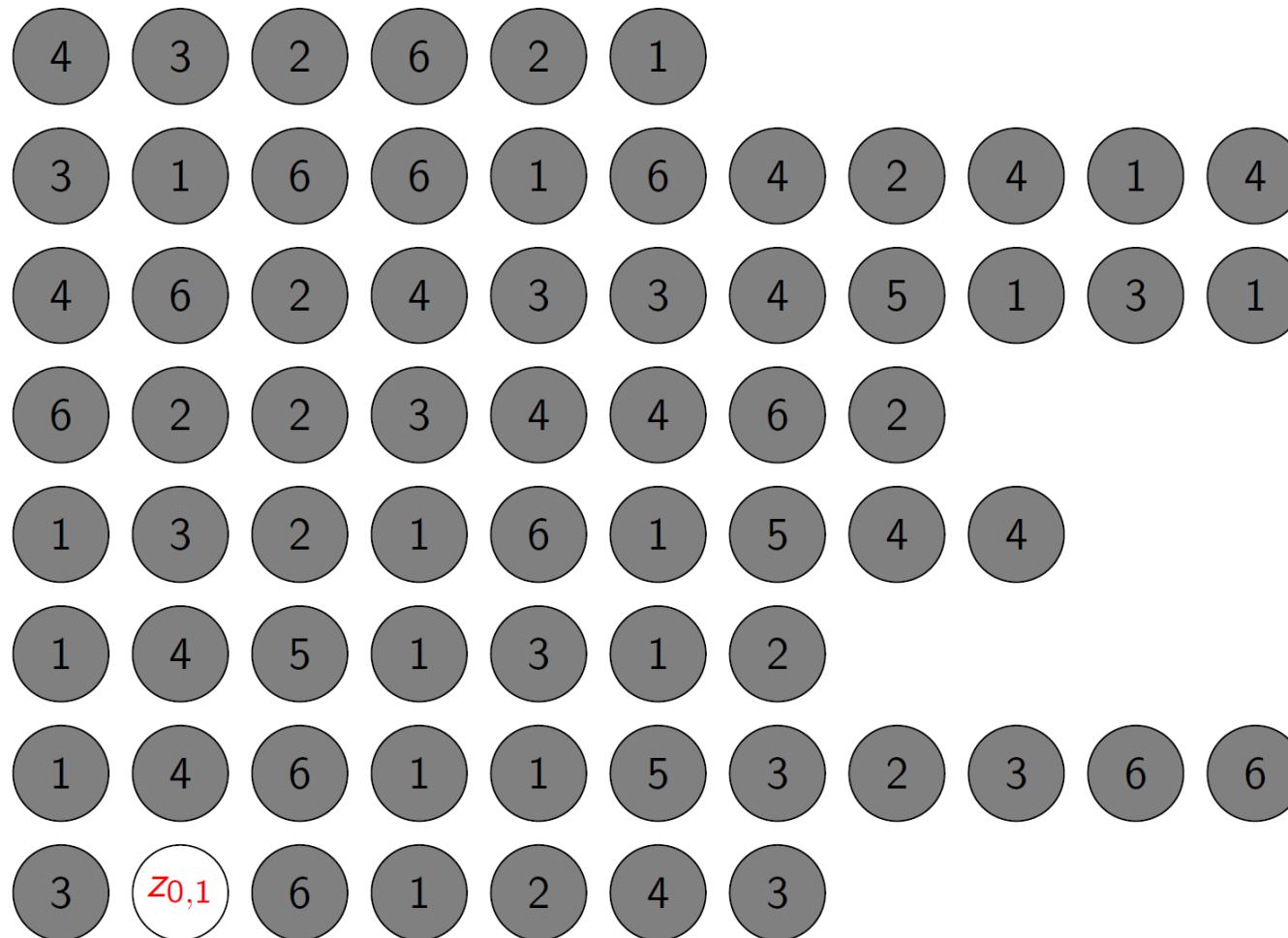


# LDA Inference: Collapsed Gibbs Sampling

Speh et al. (2013)

- Collapsed Gibbs Sampling

- ✓ Illustrative procedure

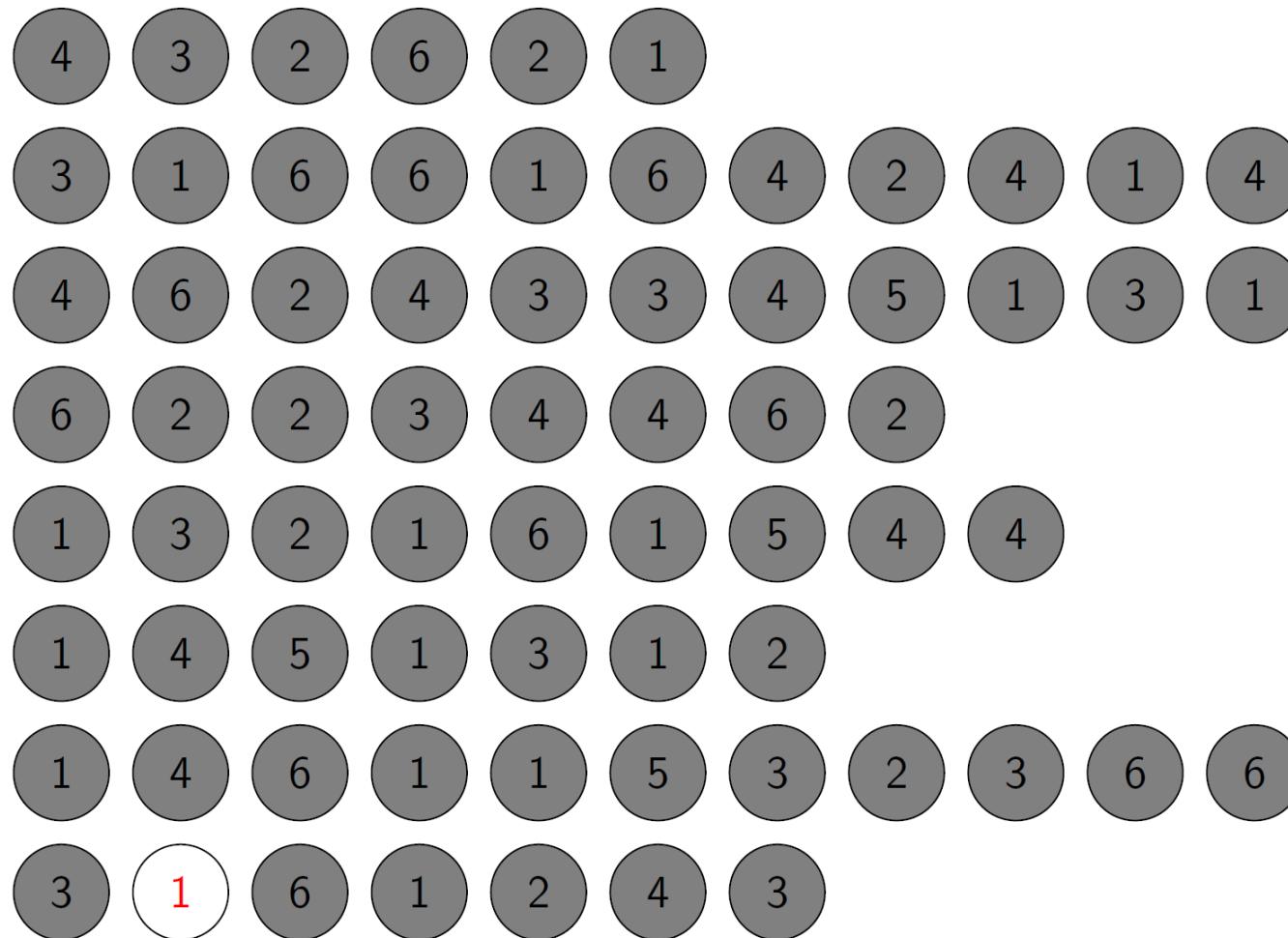


# LDA Inference: Collapsed Gibbs Sampling

Speh et al. (2013)

- Collapsed Gibbs Sampling

- ✓ Illustrative procedure

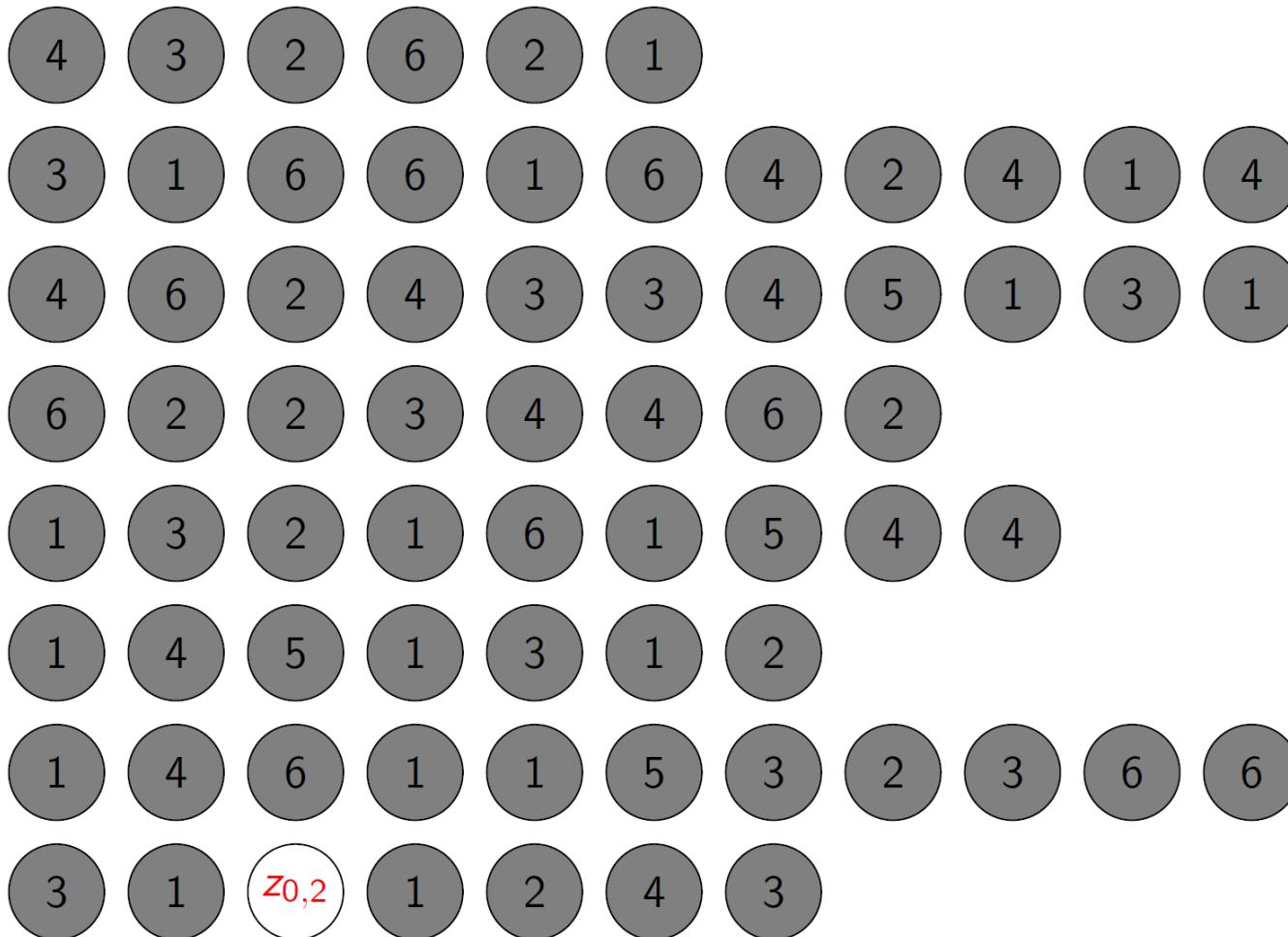


# LDA Inference: Collapsed Gibbs Sampling

Speh et al. (2013)

- Collapsed Gibbs Sampling

- ✓ Illustrative procedure

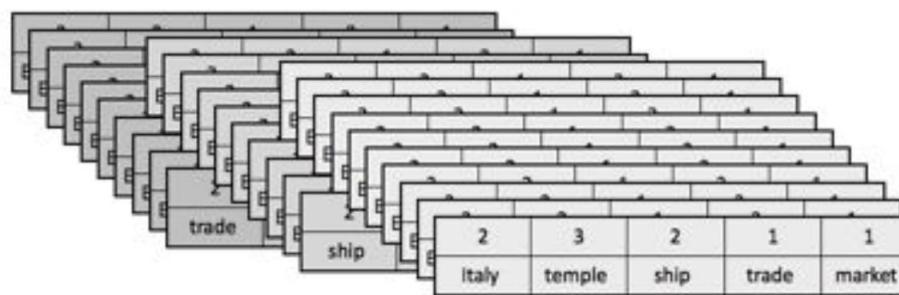
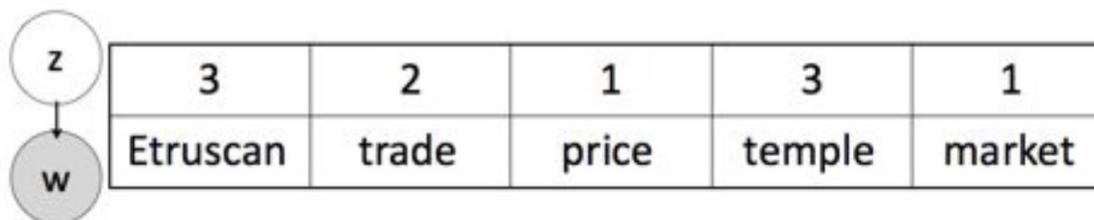


# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Randomly assign topics

Etruscan	trade	price	temple	market



# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Randomly assign topics

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts from all docs →

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

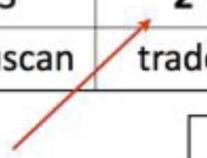
$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Sampling

3	2	1	3	1
Etruscan	trade	price	temple	market



	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
...			

- What is the conditional distribution for this topic
  - Part 1: How much does this document like each topic?
  - Part 2: How much does each topic like the word?

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- What is the conditional distribution for this topic
  - ✓ Part I: How much does this document like each topic?

$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

3	?	1	3	1
Etruscan	trade	price	temple	market



# LDA Inference: Collapsed Gibbs Sampling

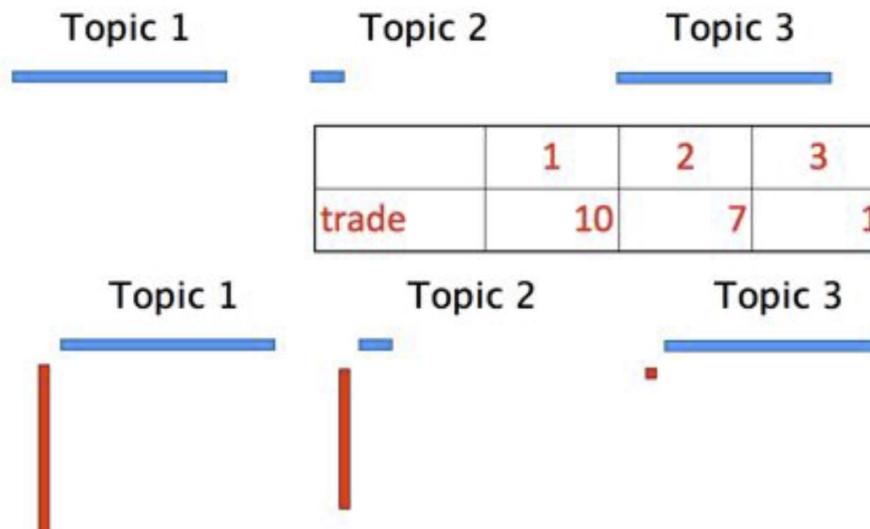
Boyd-Graber (2014)

- What is the conditional distribution for this topic

- ✓ Part 2: How much does each topic like the word?

$$\frac{n_{d,k} + \alpha_k}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

3	?	1	3	1
Etruscan	trade	price	temple	market

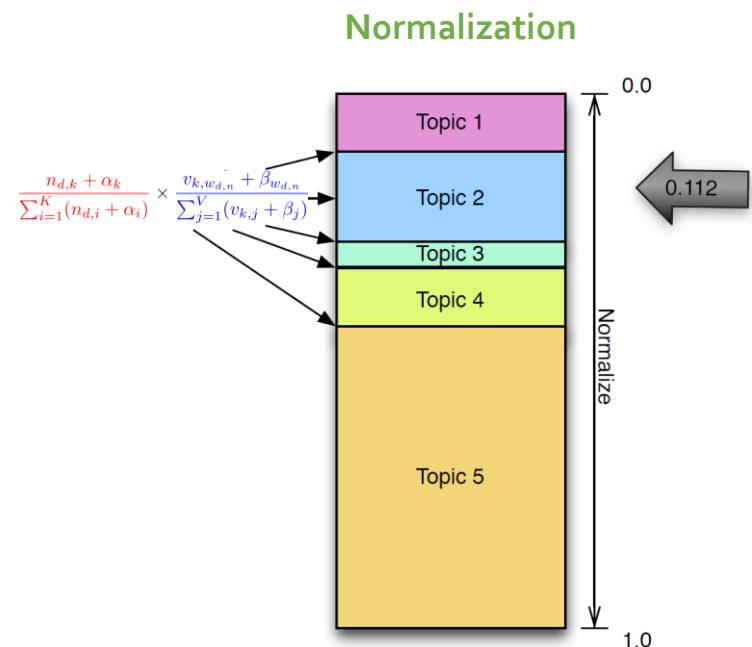
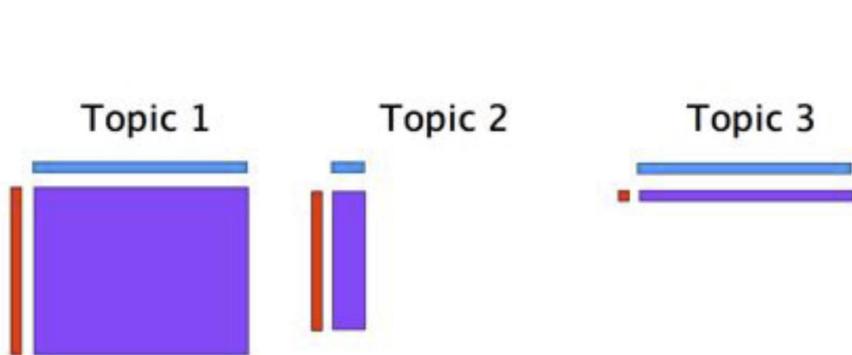


# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- What is the conditional distribution for this topic
- ✓ Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market



# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Update count

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
...			

3	1	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	11	7	1
...			

# LDA Inference: Collapsed Gibbs Sampling

Boyd-Graber (2014)

- Gibbs Sampling

AT A sawmill in the misty  
hills of Michoacán in  
central Mexico, loggers  
sporting damp sombreros  
and droopy moustaches  
are working through a  
drizzle, hauling pine logs.  
With iron spikes they  
lever them into  
  
of the last

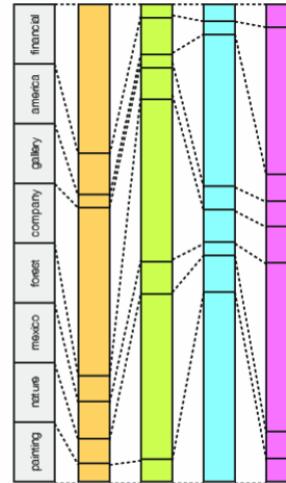
Sampling

AT A sawmill in the misty  
hills of Michoacán in  
central Mexico, loggers  
sporting damp sombreros  
and droopy moustaches  
are working through a  
drizzle, hauling pine logs.  
With iron spikes they  
lever them into  
  
of the last

We have no idea about topics

We now see how the model  
relates topics and words  
(What  $P(\mathbf{z}|\mathbf{d}, \alpha, \beta)$  looks like)

Estimation



We have each topic's  
word distribution

# AGENDA

- 01 Topic Modeling
- 02 Probabilistic Latent Semantic Analysis
- 03 LDA: Document Generation Process
- 04 LDA Inference: Gibbs Sampling
- 05 LDA Evaluation

# LDA Evaluation & Model Selection

Qui et al. (2014)

- How many topics are optimal?
  - ✓ Split the data into training and test data sets
    - Log-likelihood for Gibbs sampling

$$\log(p(w|z)) = k \log \left( \frac{\Gamma(V\delta)}{\Gamma(\delta)^V} \right) + \sum_{K=1}^k \left\{ \left[ \sum_{j=1}^V \log(\Gamma(n_K^{(j)} + \delta)) \right] - \log(\Gamma(n_K^{(.)} + V\delta)) \right\}$$

- Perplexity (the geometric mean per-word likelihood) is often used

$$\text{Perplexity}(w) = \exp \left\{ -\frac{\log(p(w))}{\sum_{d=1}^D \sum_{j=1}^V n^{(jd)}} \right\} \quad \log(p(w)) = \sum_{d=1}^D \sum_{j=1}^V n^{(jd)} \log \left[ \sum_{K=1}^k \theta_K^{(d)} \beta_K^{(j)} \right]$$

- Topic weights is determined for the new data (holdout data set) using Gibbs sampling
- Term distributions for topics are kept fixed from the training corpus

# LDA Evaluation & Model Selection

- Model Selection based on Perplexity

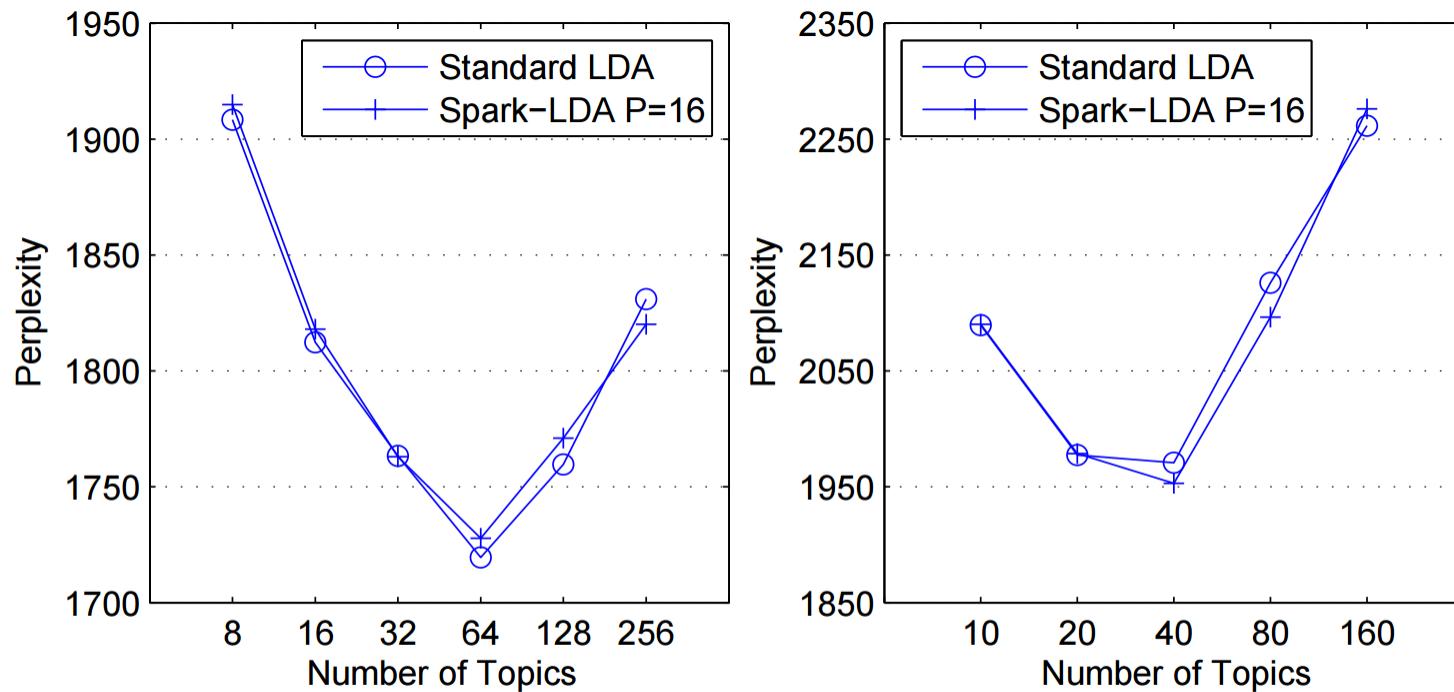


Figure 5: Test set perplexity versus number of topics  $K$  for KOS (left) and NIPS (right).

# LDA Visualization

- LDAviz

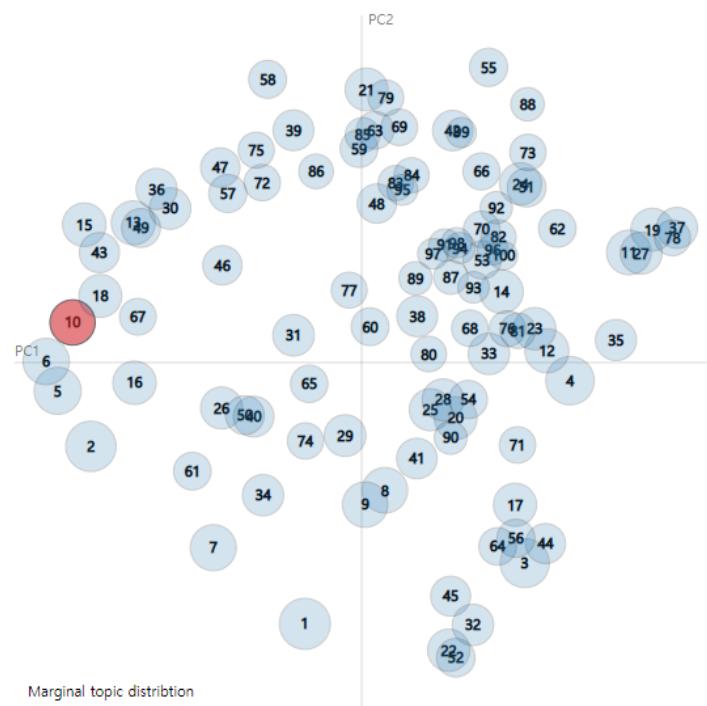
```
In [8]: pyLDAvis.graphlab.prepare(topic_model, bows)
```

Out [8] :

Selected Topic: 10    Previous Topic    Next Topic    Clear Topic

Slide to adjust relevance metric:<sup>(2)</sup>  
 $\lambda = 1$     0.0    0.2    0.4    0.6    0.8    1.0

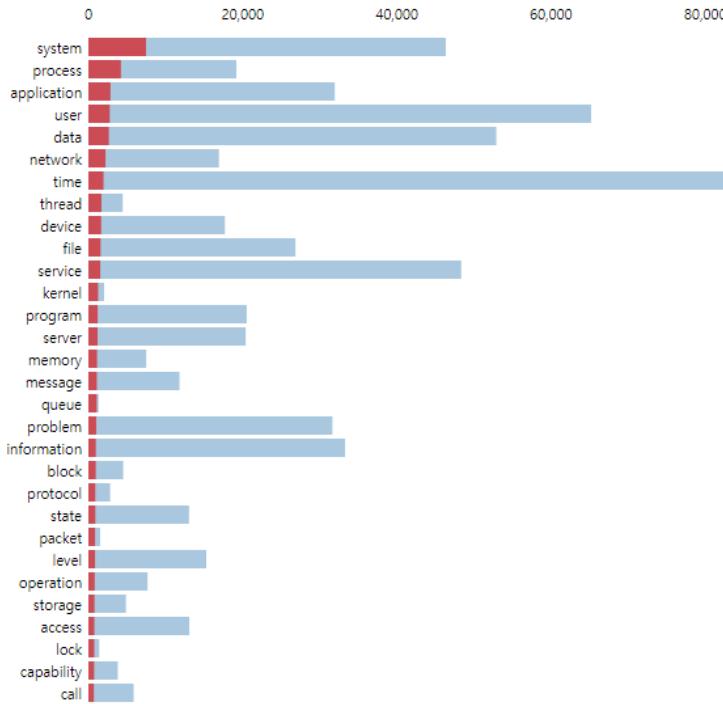
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 10 (1.3% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term, w) = frequency(w) \* [sum\_t p(t, w) \* log(p(t, w)/p(t))] for topics t; see Chuang et al (2012).

2. relevance(term, w | topic, t) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t) / p(w)$ ; see Sievert & Shirley (2014).



# References

## Research Papers

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). ACM.
- Kim, H., Park, M., & Kang, P. (2016). 토픽모델링과 사회연경망을 통한 딥러닝 연구동향 분석. 대한산업공학회 춘계공동학술대회, 제주.
- Kim, J. & Kang, P. (2018+). Analyzing International Collaboration and Identifying Core Topics for the “Internet of Things” based on Network Analysis and Topic Modeling, Under review
- Lee, H. & Kang, P. (2017+). Identifying core topics in technology and innovation management studies: A topic model approach, *Journal of Technology Transfer*, Accepted for Publication.
- Qiu, Z., Wu, B., Wang, B., Shi, C., Yu, L. (2014). [Collapsed Gibbs Sampling for Latent Dirichlet Allocation on Spark](#), *JMLR: Workshop and Conference Proceedings* 36: 17-28.

# References

## Other Materials

- Image on the first page: <http://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext>
- Blei, D.M. (2012). [Probabilistic Topic Models](#), ICML'12 Tutorial.
- Boyd-Graber, J. (2014). [Topic Models](#), Natural Language Processing Course, Dept. of Computer Science, University of Colorado Boulder. ([Video Lecture Link](#))
- Helic, D. (2014). Knowledge Discovery and Data Mining I: Probabilistic Latent Semantic Analysis.
- Hofmann, T. (2005). [Latent Semantic Variable Models](#), Workshop on Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimisation Perspectives, Bohinj 2005.
- Knispelis, A. (2015). LDA Topic Models. Slide: [https://issuu.com/andriusknispelis/docs/topic\\_models\\_-\\_video](https://issuu.com/andriusknispelis/docs/topic_models_-_video), Youtube video: <https://www.youtube.com/watch?v=3mHy4OSyRf0>
- Murray, I. (2009). Markov Chain Monte Carlo. Lectures on Machine Learning Summer School 2009: <http://homepages.inf.ed.ac.uk/imurray2/teaching/09mlss/slides.pdf>
- Speh, J., Muhic, A., and Rupnik, J. (2013). [Parameter Estimation for the Latent Dirichlet Allocation](#), SiKDD'13. ([Video Lecture Link](#))
- Tang, L. (2008). [Gibbs Sampling for LDA](#)