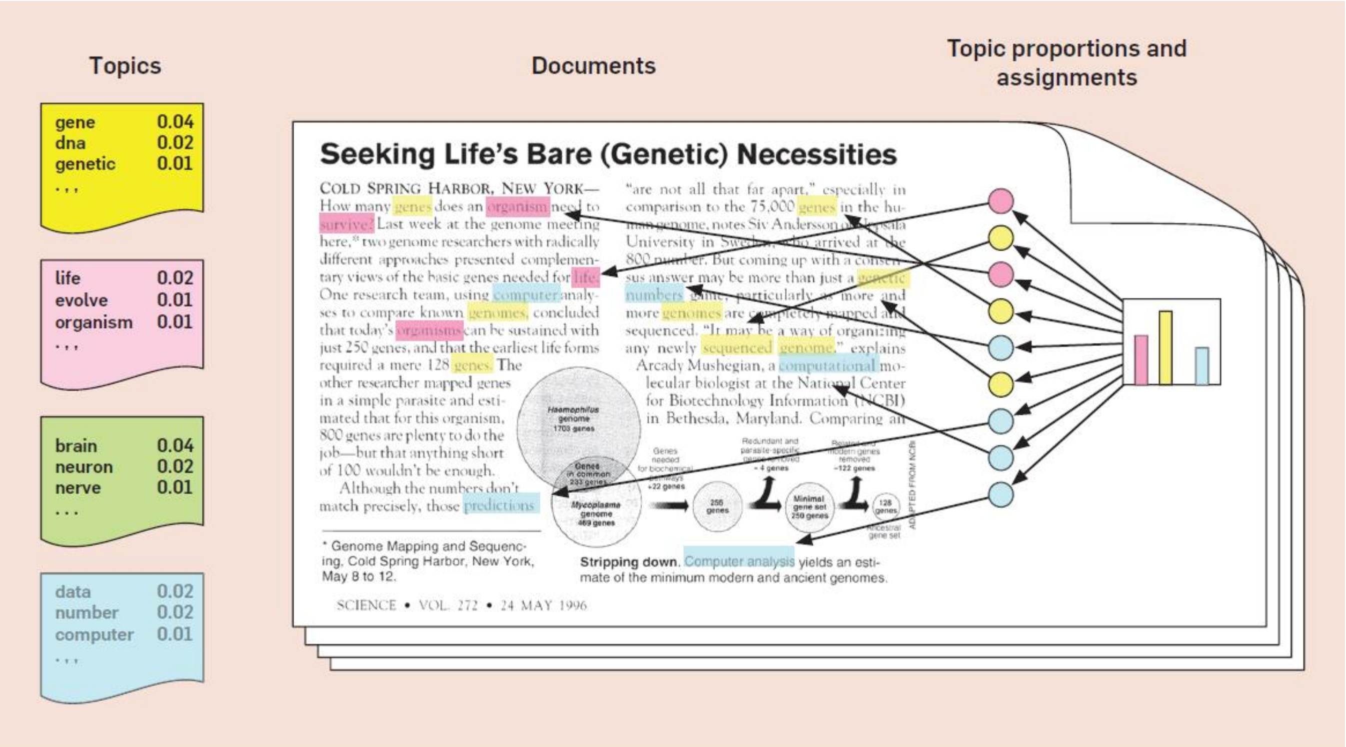


Lecture 7: Topic Modeling

Pilsung Kang

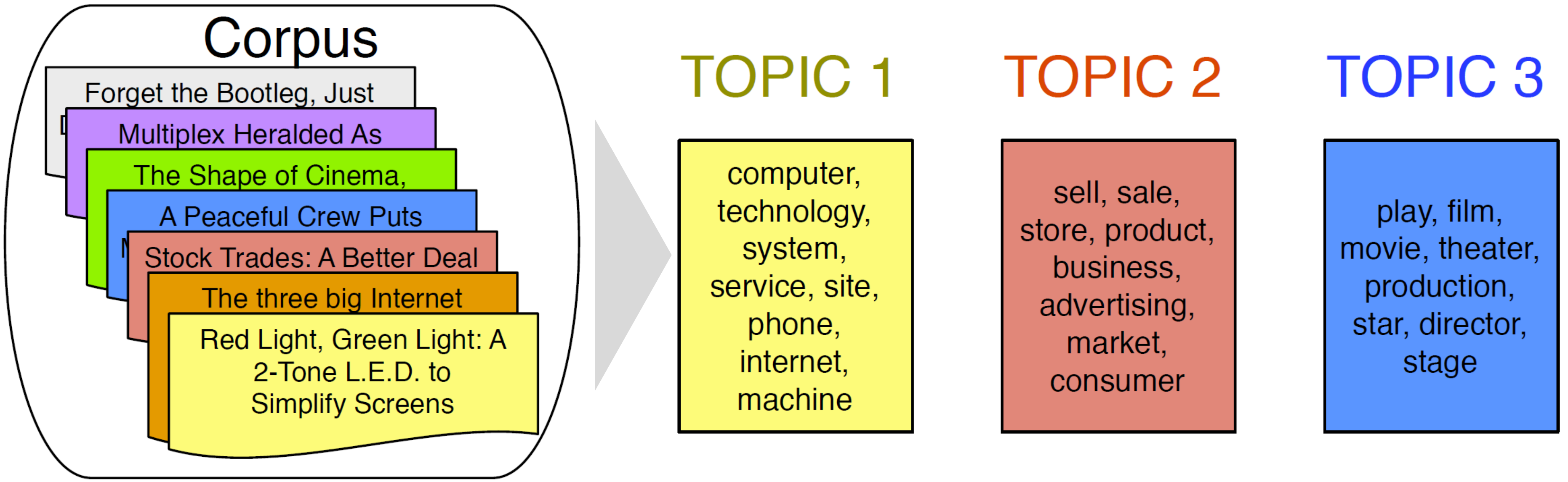
School of Industrial Management Engineering

Korea University

AGENDA

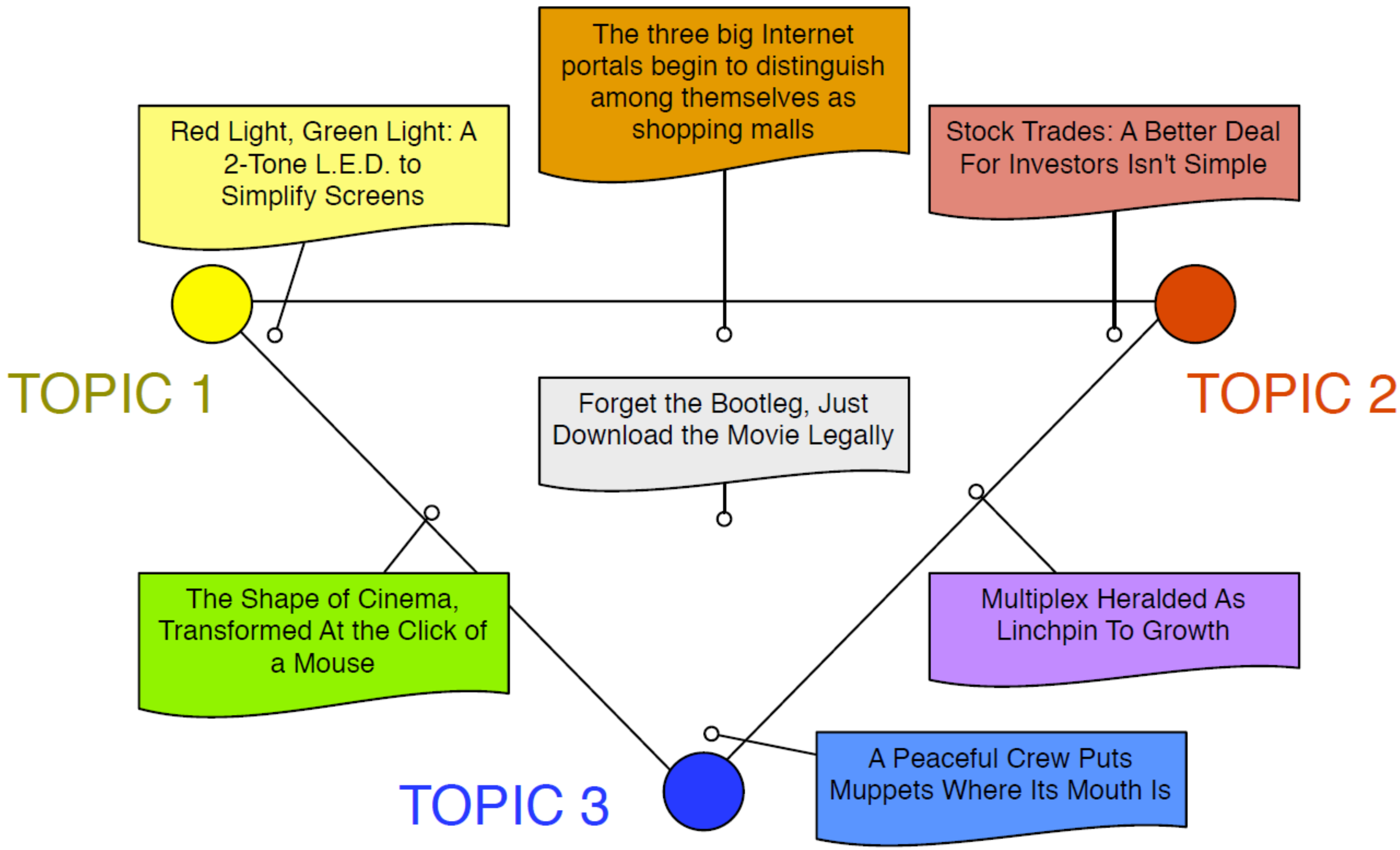
Topic Model: Conceptual Approach

- 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



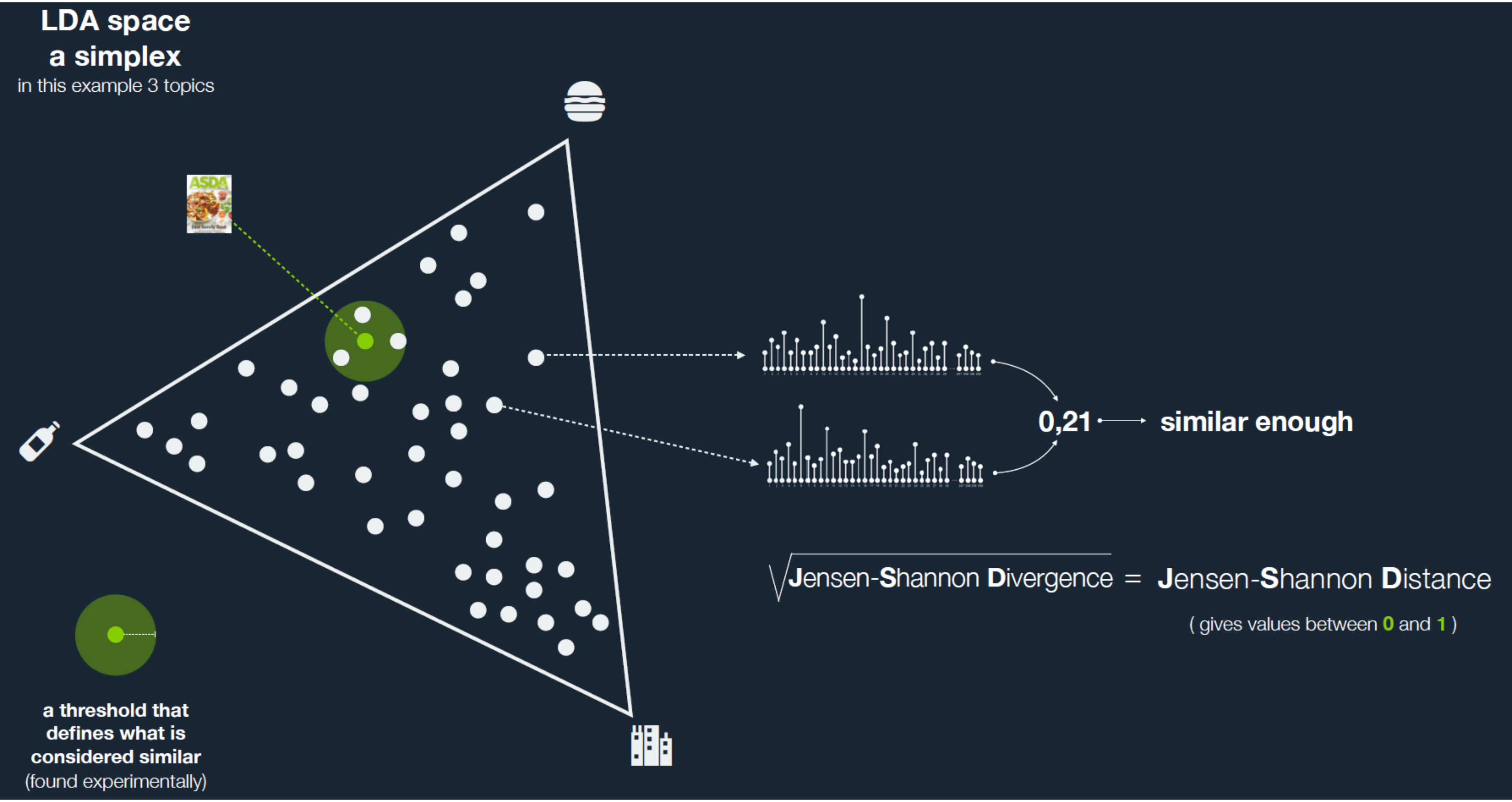
Topic Model: Conceptual Approach

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



Topic Model: Conceptual Approach

Knispelis (2015)



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"

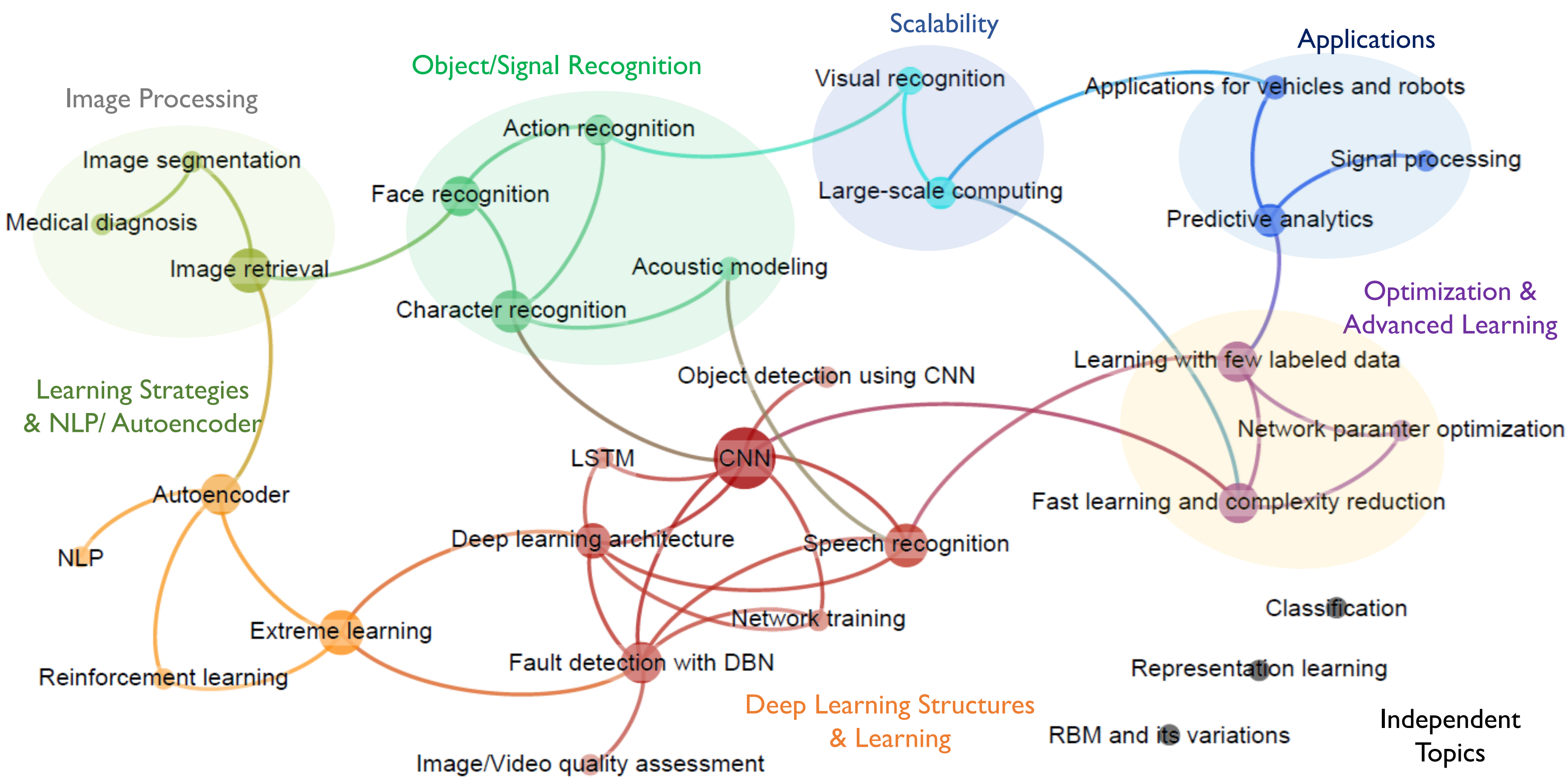
No.	Insider threat in relational database	Assessment of insider threat	Insider attacks on Communication protocol	Modeling and system framework for insider threat	Masquerade detection
1	data	measure	attack	insider	user
2	information	assess	agent	threat	behavior
3	database	security	scheme	social	detect
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)

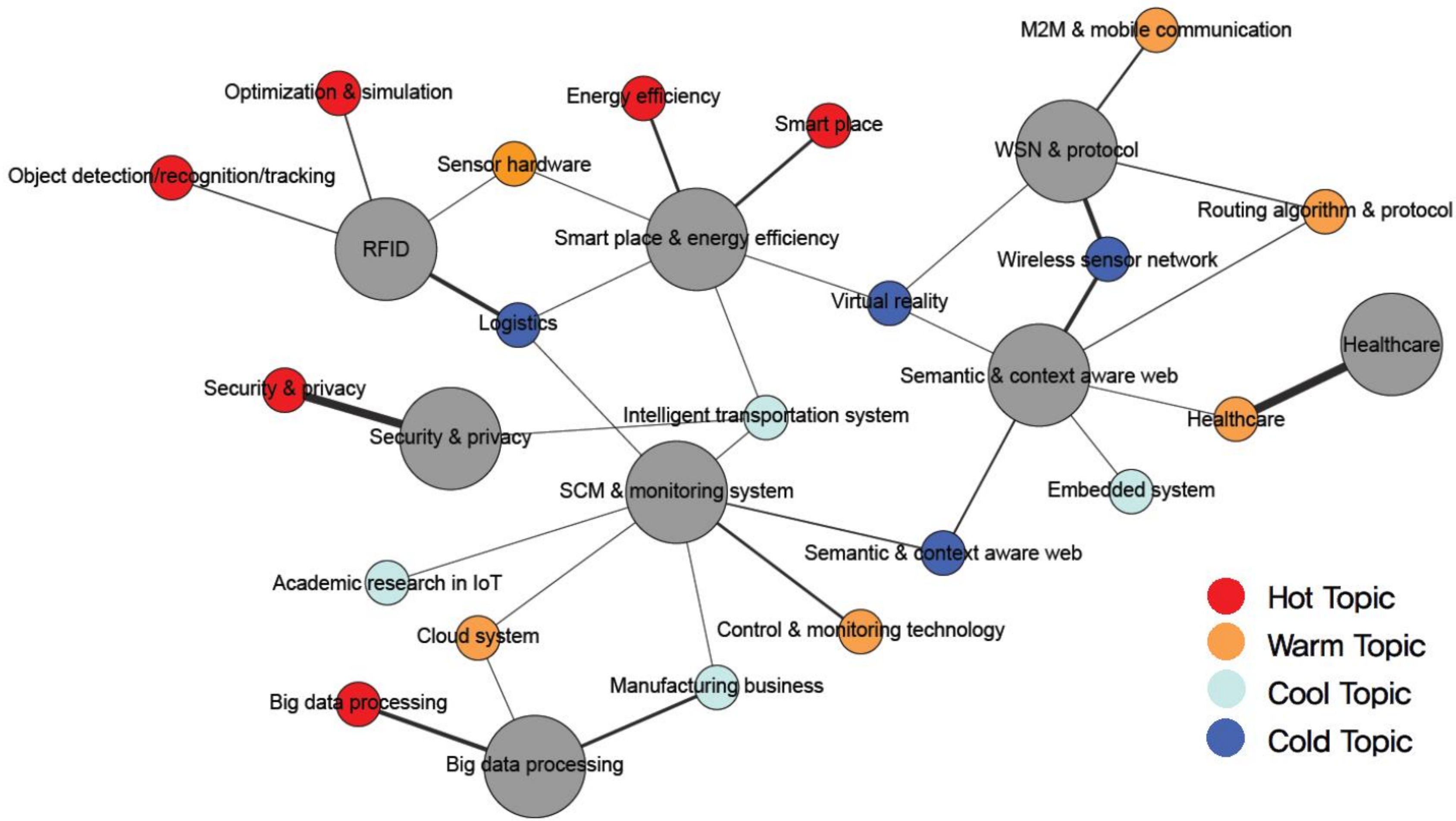
- Relation between Topics: Deep Learning



Topic Models: Relation between Topics

Kim and Kang (2018+)

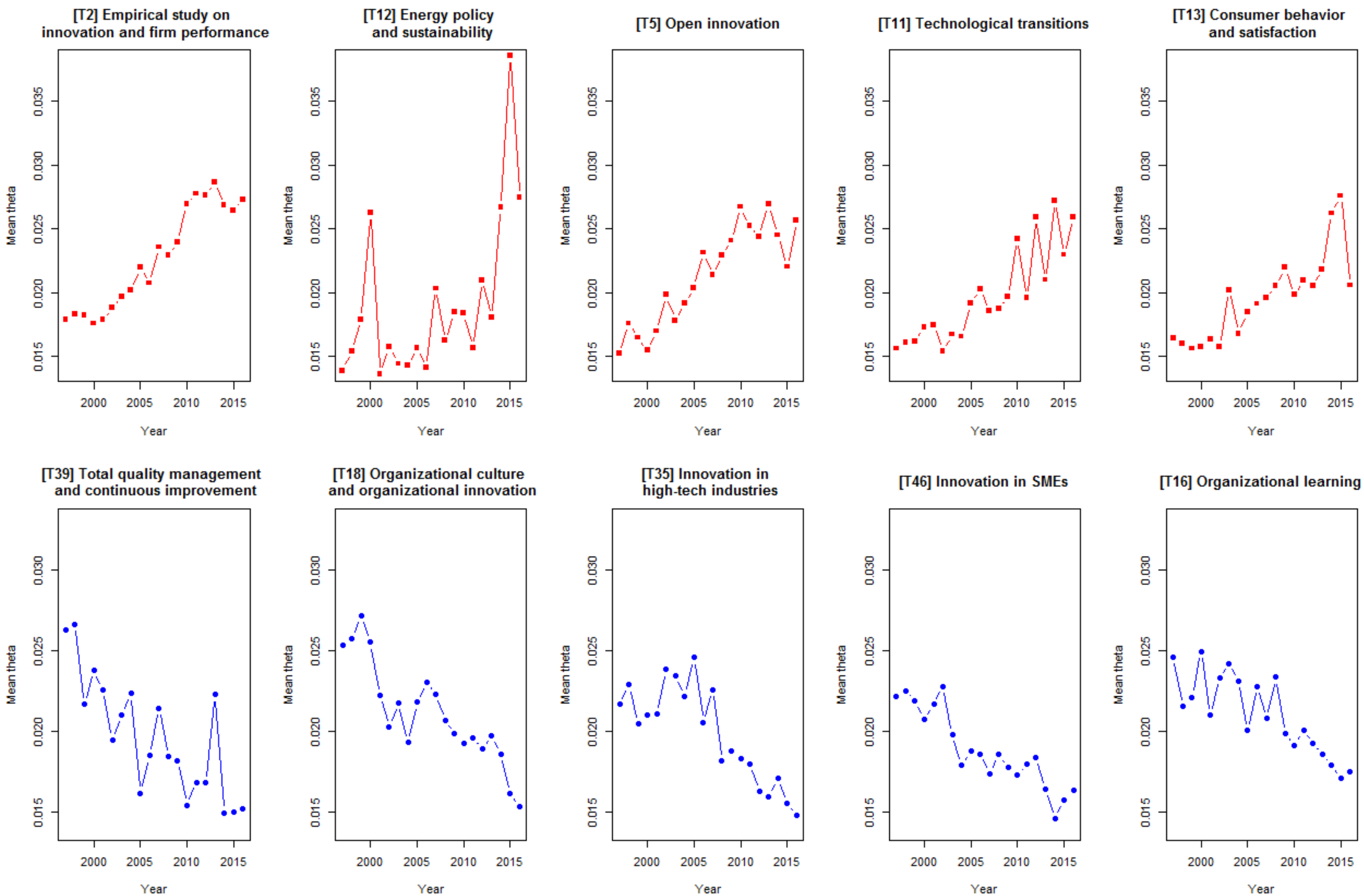
- Relation between Topics: Internet of Things



Topic Models: Trend Analysis

Lee and Kang (2017)

- Topic trends for "technology and innovation management"



Topic Model: Document Retrieval

Knispelis (2015)

Topic Model: Document Retrieval

Knispelis (2015)

Topic Model: Document Retrieval

Knispelis (2015)

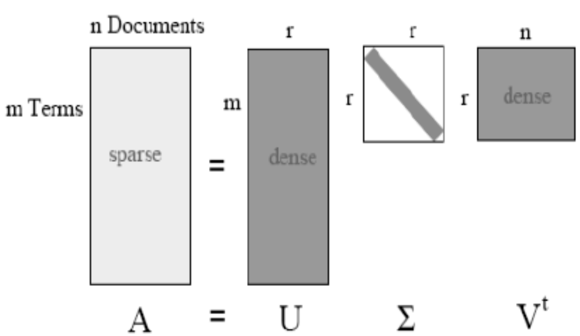
Topic Model

- Matrix Factorization Approach

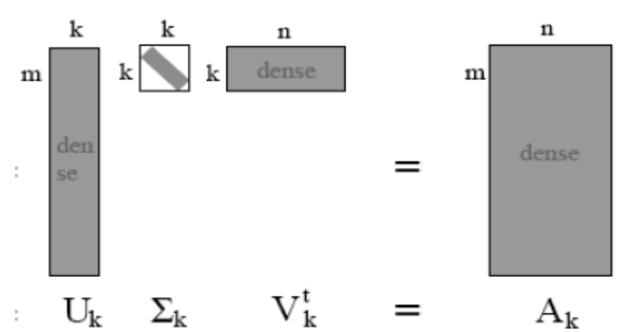
$$\underset{\text{Topic Assignment}}{\begin{bmatrix} M \times K \end{bmatrix}} \times \underset{\text{Topics}}{\begin{bmatrix} \quad K \times V \quad \end{bmatrix}} \approx \underset{\text{Dataset}}{\begin{bmatrix} \quad M \times V \quad \end{bmatrix}}$$

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

Reduced SVD



Approximating



Topic Model

Helic (2014)

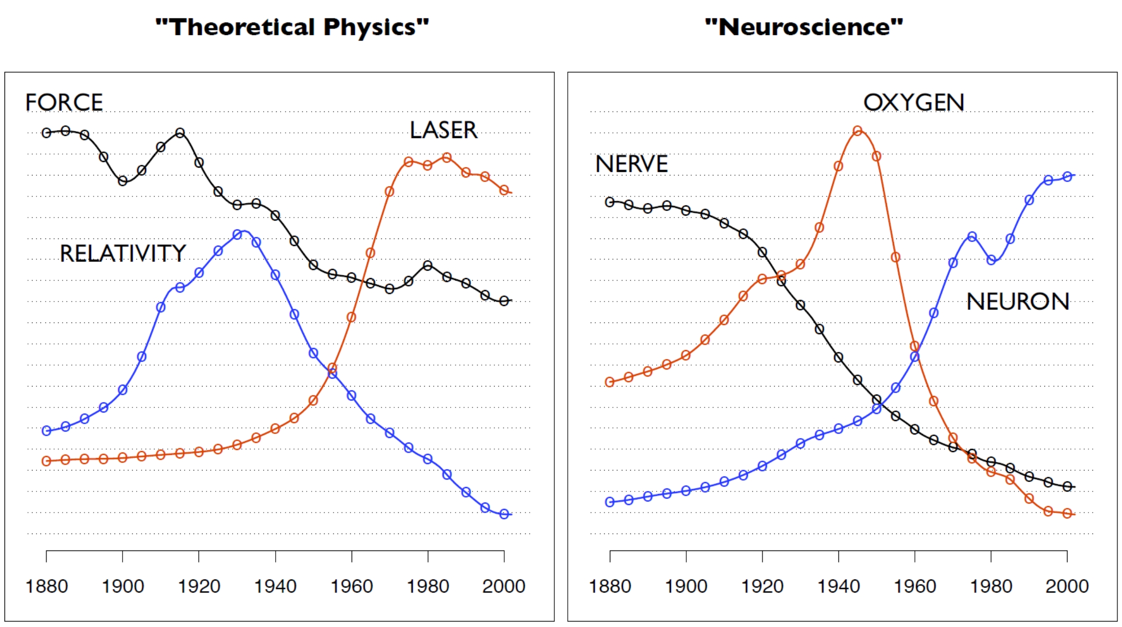
- Disadvantage of LSA
  - ✓ Statistical foundation is missing
  - ✓ SVD assumes normally distributed data
  - ✓ Term occurrence is not normally distributed
  - ✓ Still, often it works remarkably good because matrix entries are weighted (e.g. tf-idf) and those weighted entries may be normally distributed

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

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

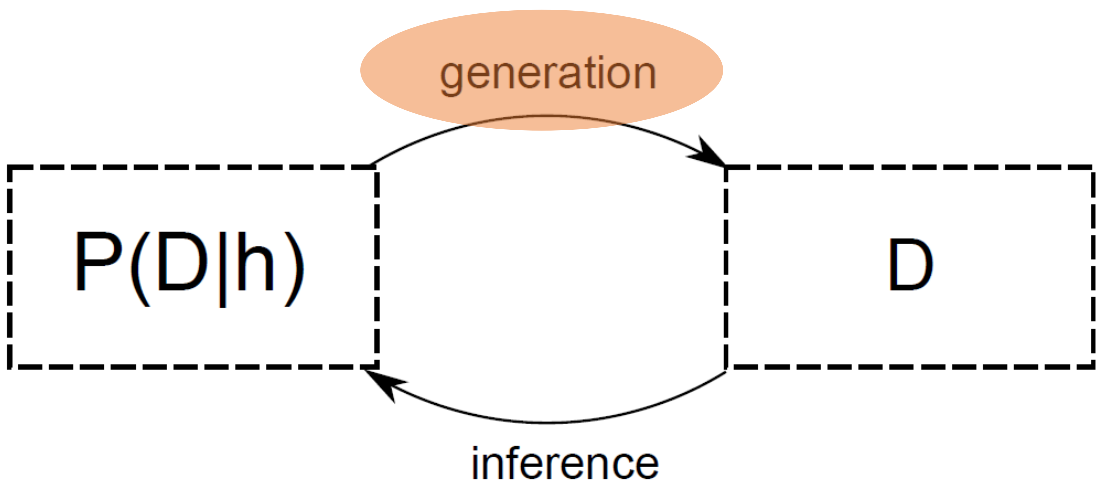
Helic (2014)

- 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)

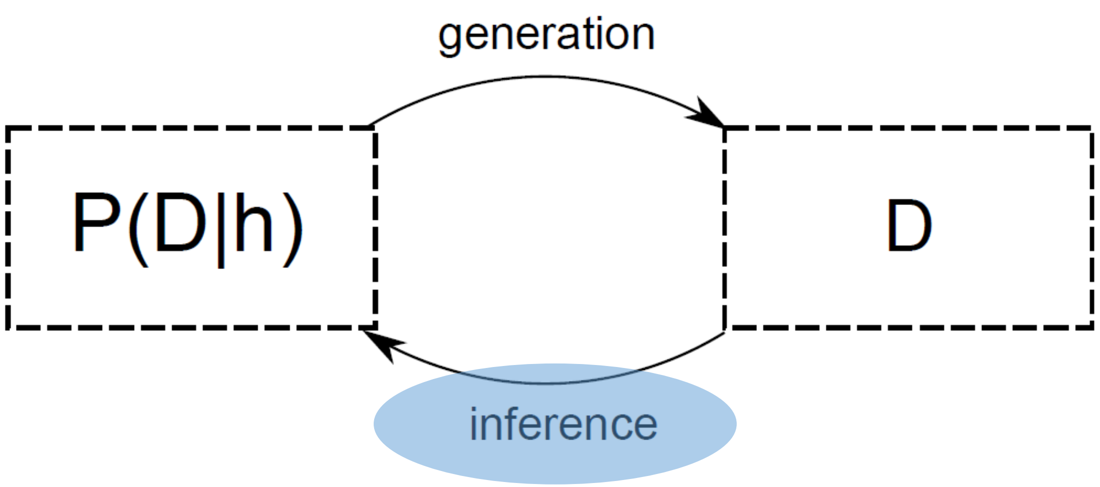
- How it work?
  - ✓ It defines a conditional probability distribution over data given a hypothesis P(D|h)
  - ✓ Given h, we generate data from the conditional distribution P(D|h)
  - ✓ Has many advantages but the main disadvantage is that fitting the model can be more complicated than an algorithmic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



Topic Model: Generative Approach

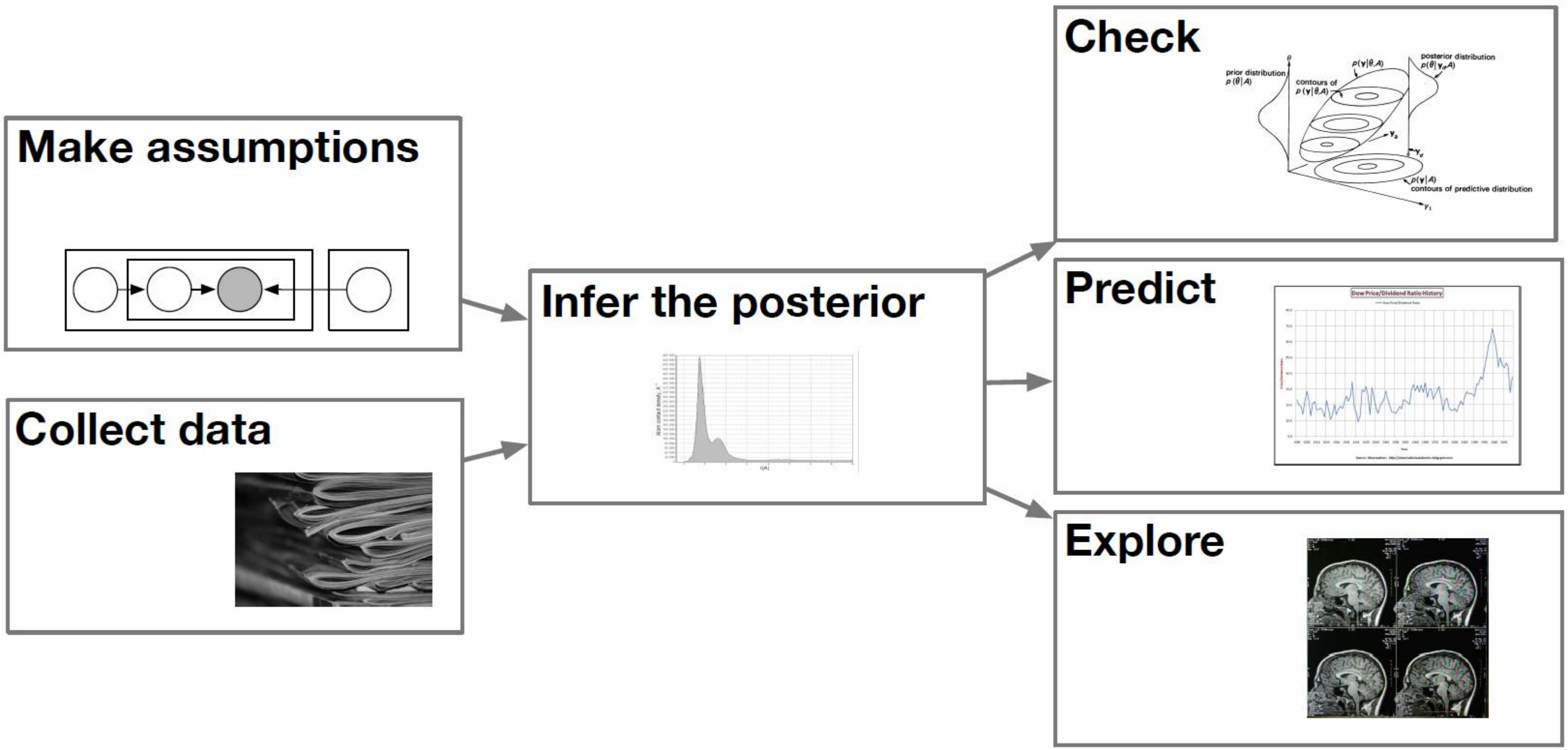
Helic (2014)

- (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

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, …

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

- 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

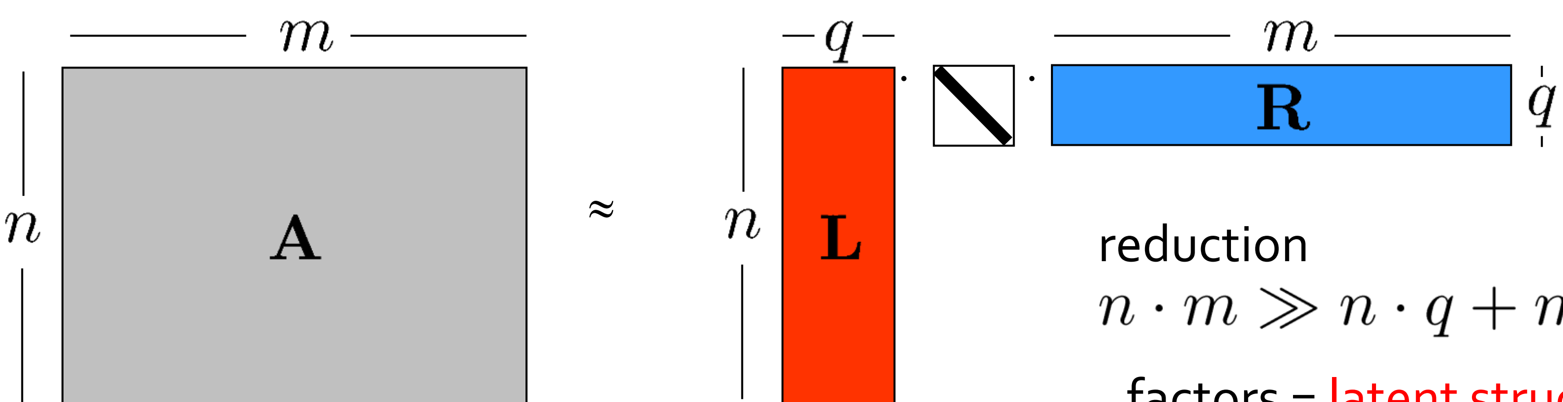
Hofmann (2005)

- Common approach: approximately factorize matrix

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

approximation left factor right factor

- Factors are typically constrained to be "thin"



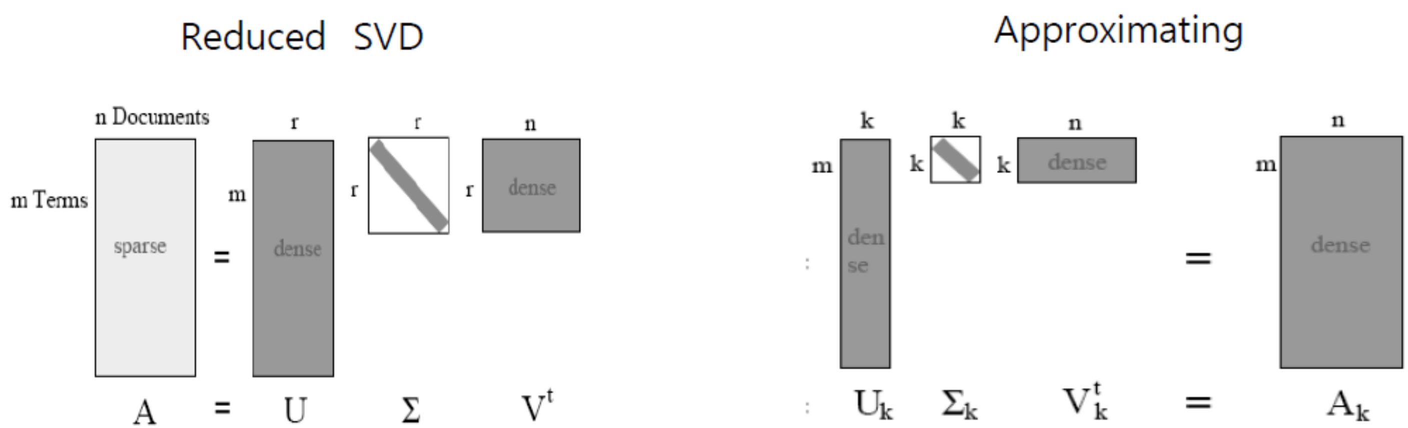
reduction

$$n \cdot m \gg n \cdot q + m \cdot q$$

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

$$\mathbf{A} \approx \mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$$

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

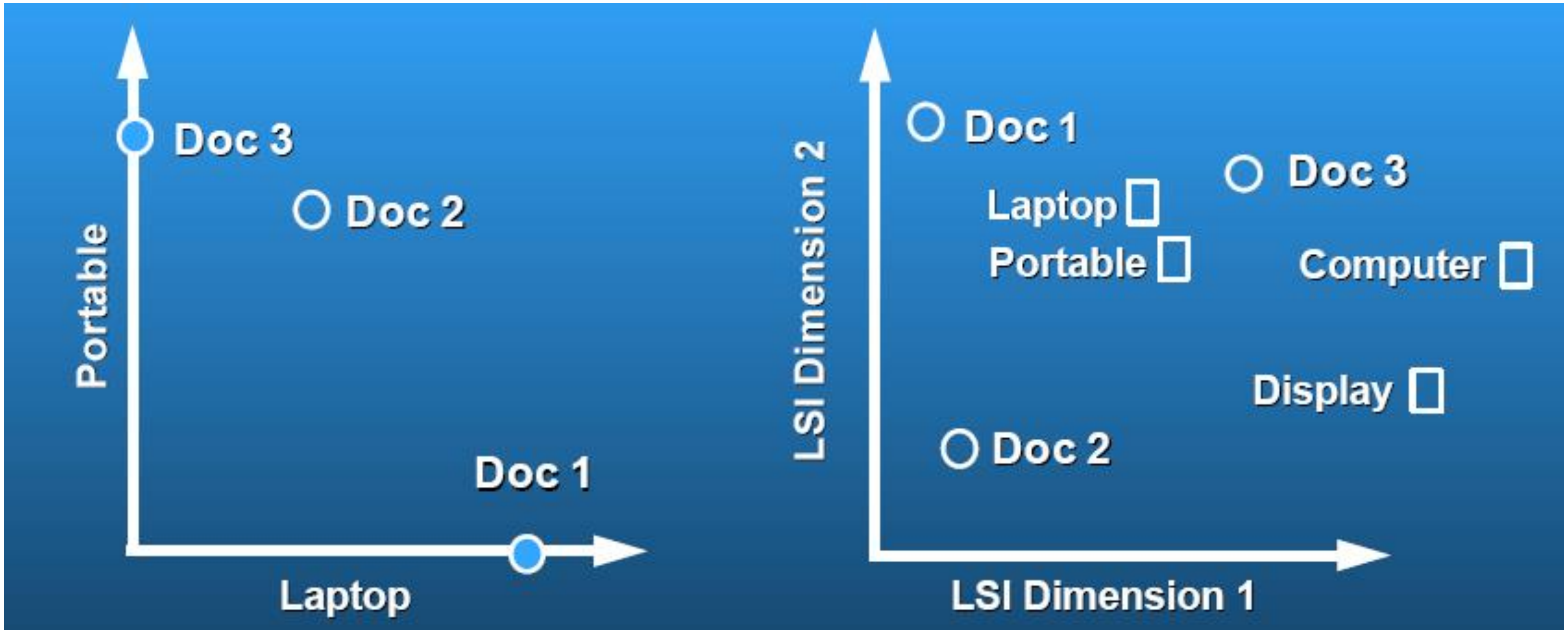
$$\mathbf{U}_k^T \mathbf{A}_k = \mathbf{U}_k^T \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T = \mathbf{I} \mathbf{\Sigma}_k \mathbf{V}_k^T = \mathbf{\Sigma}_k \mathbf{V}_k^T$$

✓ Step 3: Apply data mining algorithms

LSA Decomposition

Hofmann (2005)

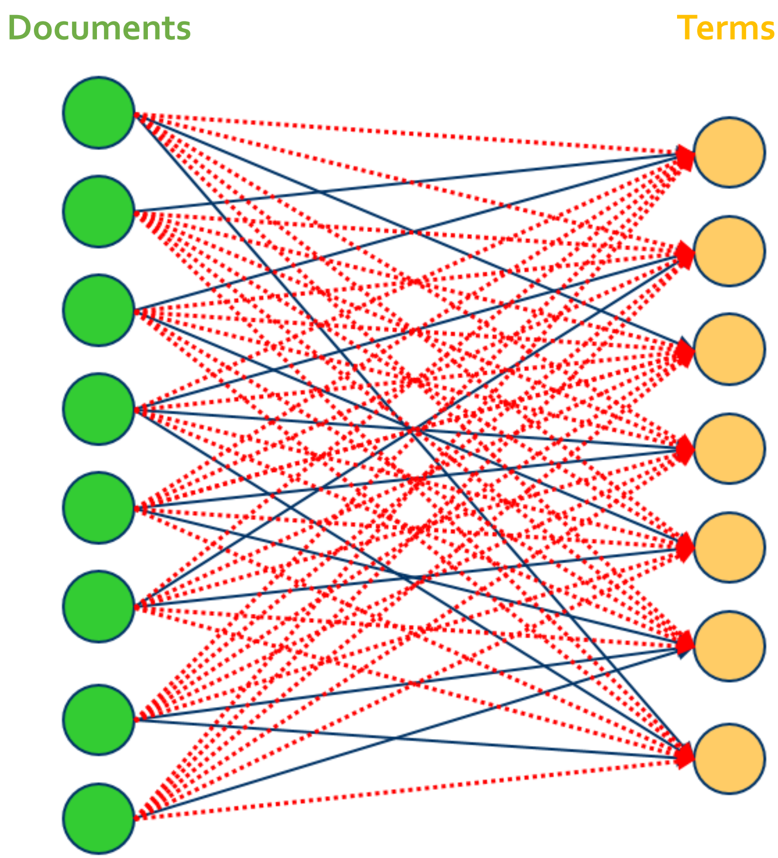
- Illustrative Example



Language Model: Naïve Approach

Hofmann (2005)

- Maximum likelihood estimation (MLE)



Number of occurrences of term w in document d

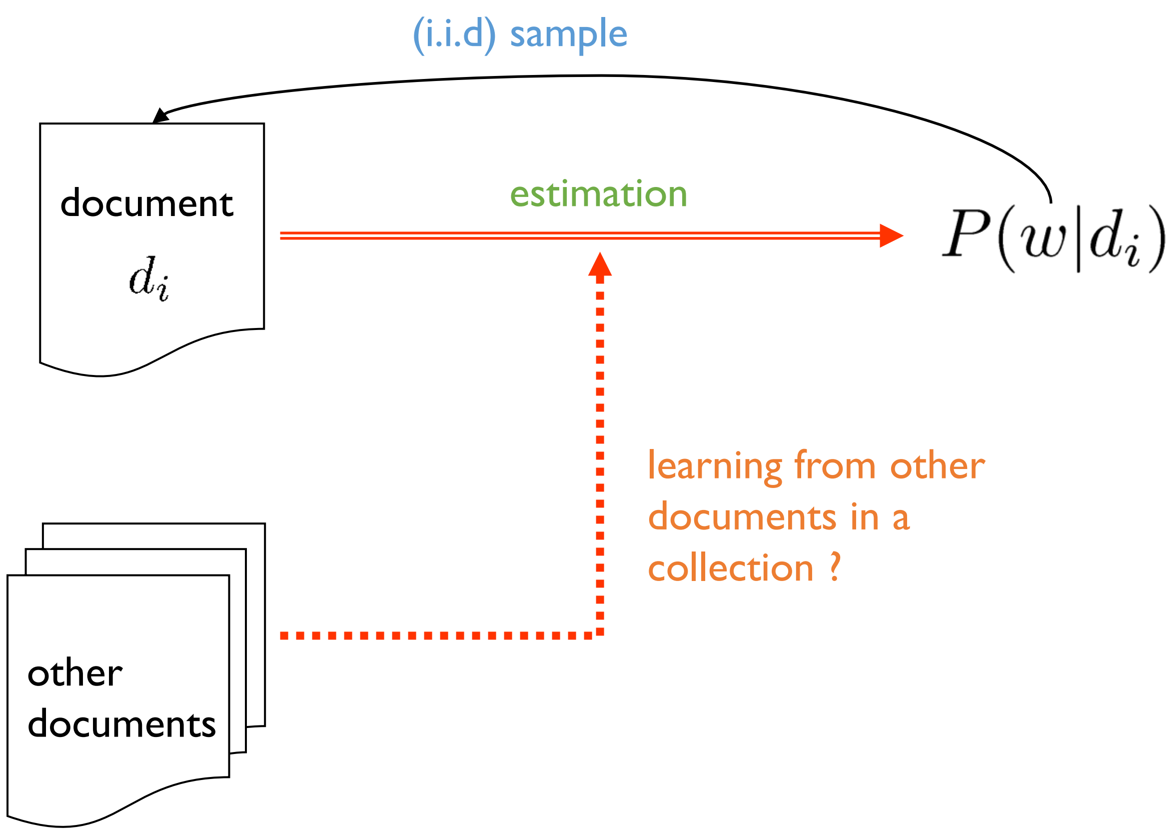
$$\hat{P}_{\mathrm{ML}}(w|d) = \frac{n(d,w)}{\sum_{w'} n(d,w')}$$

Zero frequency problem: terms not occurring in a document get zero probability

Language Model: Estimation Problem

Hofmann (2005)

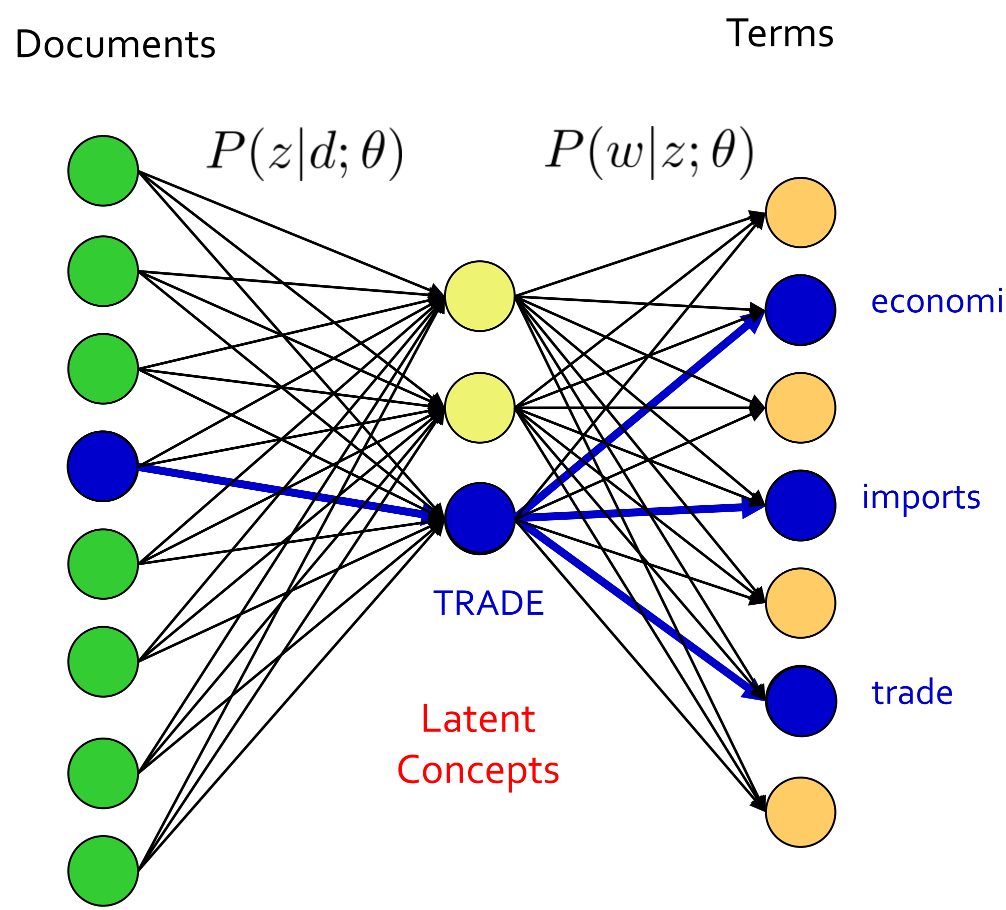
- Crucial question
  - ✓ In which way can the document collection be utilized to improve estimates?



Probabilistic Latent Semantic Analysis (pLSA)

Hofmann (2005)

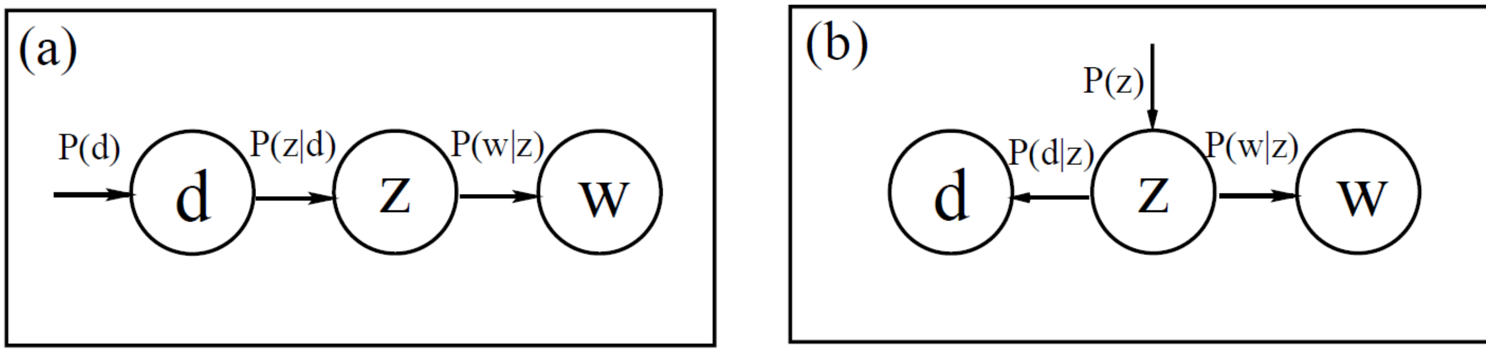
- Concept expression probability
  - ✓ Estimated based on all documents that are dealing with a concept
  - ✓ "Unmixing" of superimposed concepts is achieved by statistical learning algorithm
  - ✓ No prior knowledge about concepts required, context and term co-occurrences are exploited



pLSA: Latent Variable Model

Hofmann (2005)

- Structural modeling assumption (mixture model)



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

Document language model

Latent concepts or topics

Concept expression probabilities

Document-specific mixture proportions

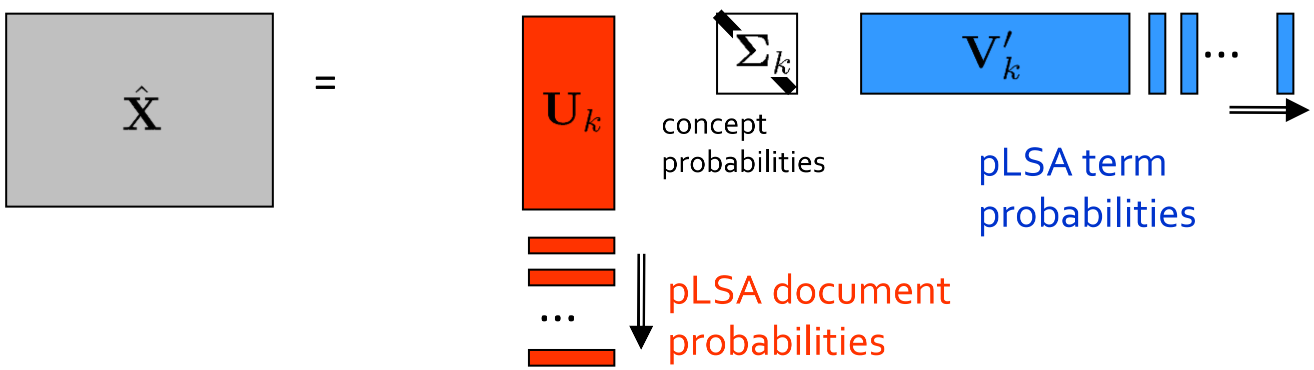
Model fitting

pLSA: Matrix Decomposition

Hofmann (2005)

- Mixture model can be written as a matrix factorization

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



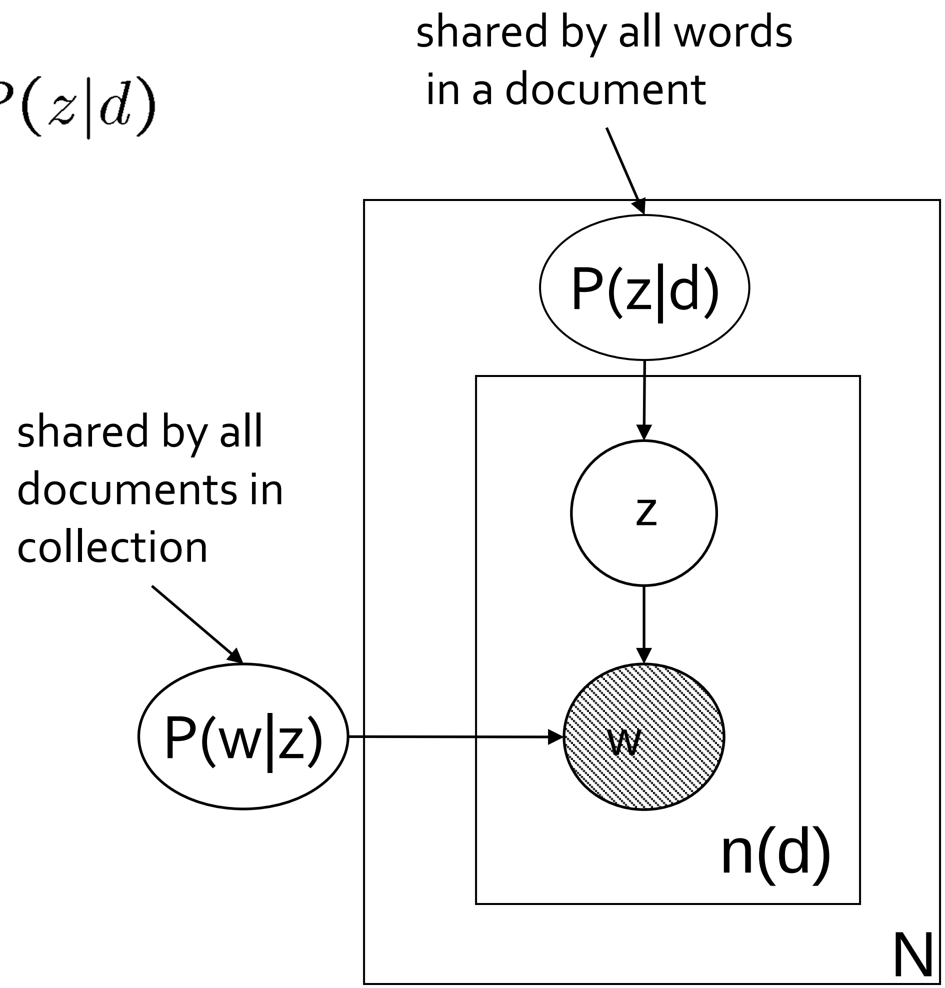
- Contrast to LSA
  - ✓ Non-negativity: every element in U & V is non-negative
  - ✓ Normalization: Each document vector in U and each term vector in V has sum 1

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 give 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

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

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 occurence 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
Cocus	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

	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

After 1 EM step

Topic 1	Topic 2	Topic 3
0.459	0.430	0.111

	Topic 1	Topic 2	Topic 3
Doc 1	0.180	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

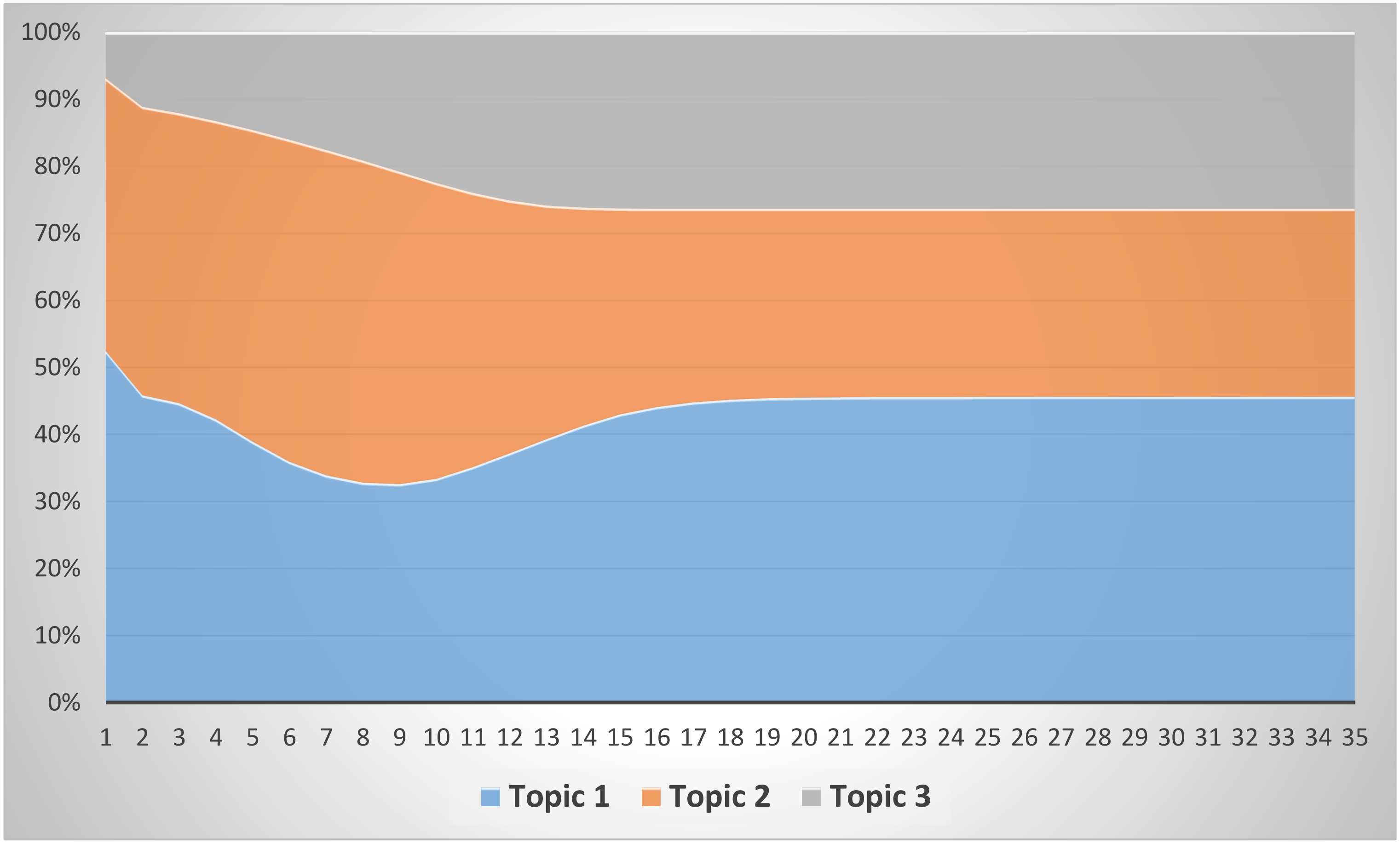
	Topic 1	Topic 2	Topic 3
Term 1	0.022	0.016	0.010
Term 2	0.018	0.133	0.166
Term 3	0.242	0.058	0.133
Term 4	0.123	0.088	0.145
Term 5	0.016	0.030	0.044
Term 6	0.020	0.167	0.056
Term 7	0.147	0.129	0.201
Term 8	0.188	0.156	0.039
Term 9	0.146	0.114	0.008
Term 10	0.077	0.110	0.199

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
Cocus	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
Cocus	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
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	speaker	robust	manufactur	constraint	alpha	POWER	load
SEGMENT	speech	MATRIX	cell	LINE	redshift	spectrum	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	arrai
mri	SEGMENT	root	format	fundament	densiti	standard	present
volume	sound	suffici	group	recogn	veloc	model	implement

Document 1, $P\{z_k|d_1, w_j = \text{`}segment\text{`}\} = (0.951, 0.0001, \ldots)$
$P\{w_j = \text{`}segment\text{`}|d_1\} = 0.06$

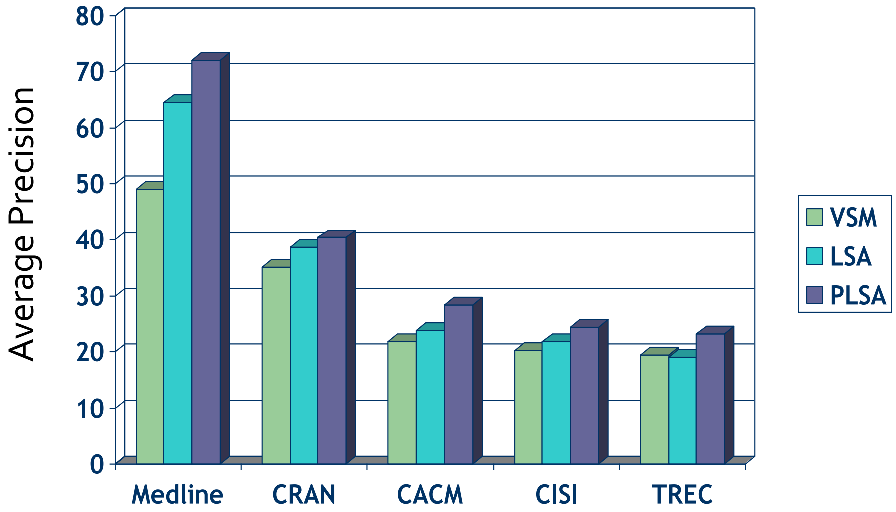
SEGMENT medic imag challeng problem field imag analysi diagnost base proper **SEGMENT** digit imag **SEGMENT** medic imag need applic involv estim boundari object classif tissu abnorm shape analysi contour detec textur **SEGMENT** despit exist techniqu **SEGMENT** specif medic imag remain crucial problem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

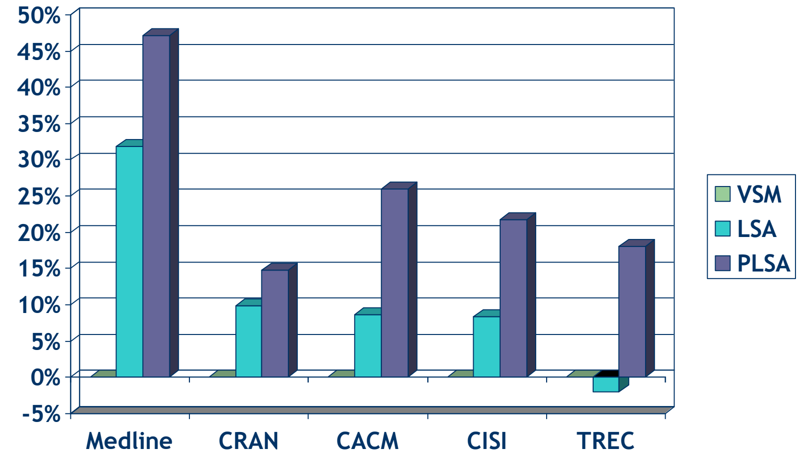
Document 2, $P\{z_k|d_2, w_j = \text{`}segment\text{`}\} = (0.025, 0.867, \ldots)$
$P\{w_j = \text{`}segment\text{`}|d_2\} = 0.010$

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%


ANY
questions?