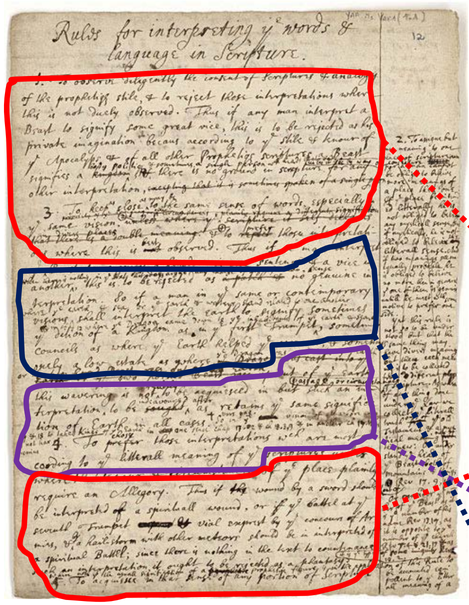
Probabilistic Topic Models

Hongning Wang
CS@UVa

Outline

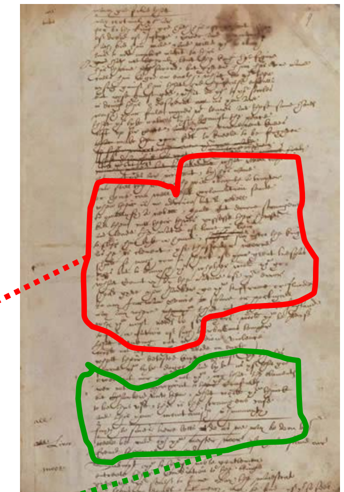
1. **General idea of topic models**
2. Basic topic models
   - Probabilistic Latent Semantic Analysis (pLSA)
   - Latent Dirichlet Allocation (LDA)
3. Variants of topic models
4. Summary

What is a “topic”?



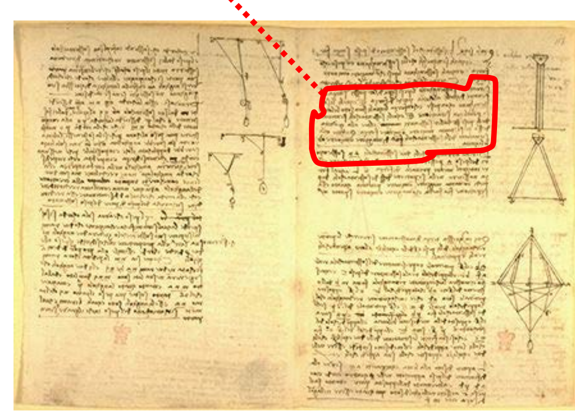
Representation: a probabilistic distribution over words.

retrieval	0.2
information	0.15
model	0.08
query	0.07
language	0.06
feedback	0.03
……	



Topic: A broad concept/theme, semantically coherent, which is *hidden* in documents

e.g., politics; sports; technology; entertainment; education etc.



Document as a mixture of topics

Topic θ_1

government 0.3
response 0.2
...

Topic θ_2

city 0.2
new 0.1
orleans 0.05
...

…

Topic θ_k

donate 0.1
relief 0.05
help 0.02
...

Background θ_k

is 0.05
the 0.04
a 0.03
...

[Criticism of government response to the hurricane primarily consisted of criticism of its response to the approach of the storm and its aftermath, specifically in the delayed response] to the [flooding of New Orleans. … 80% of the 1.3 million residents of the greater New Orleans metropolitan area evacuated] …[Over seventy countries pledged monetary donations or other assistance]. …

- How can we discover these topic-word distributions?
- Many applications would be enabled by discovering such topics
  - Summarize themes/aspects
  - Facilitate navigation/browsing
  - Retrieve documents
  - Segment documents
  - Many other text mining tasks

General idea of probabilistic topic models

- Topic: a multinomial distribution over words
- Document: a mixture of topics
  - A document is “generated” by first sampling topics from some prior distribution
  - Each time, sample a word from a corresponding topic
  - Many variations of how these topics are mixed
- Topic modeling
  - Fitting the probabilistic model to text
  - Answer topic-related questions by computing various kinds of posterior distributions
    - e.g., p(topic|time), p(sentiment|topic)

Simplest Case: 1 topic + 1 "background"

B

General Background English Text

the 0.03
a 0.02
is 0.015
we 0.01
...
food 0.003
computer 0.00001
…
text 0.000006
…

θ_B

Background Topic: $p(w|\theta_B)$

Assume words in d are from two topics:
1 topic + 1 background

How can we "get rid of" the common words from the topic to make it more discriminative?

d

Text mining paper

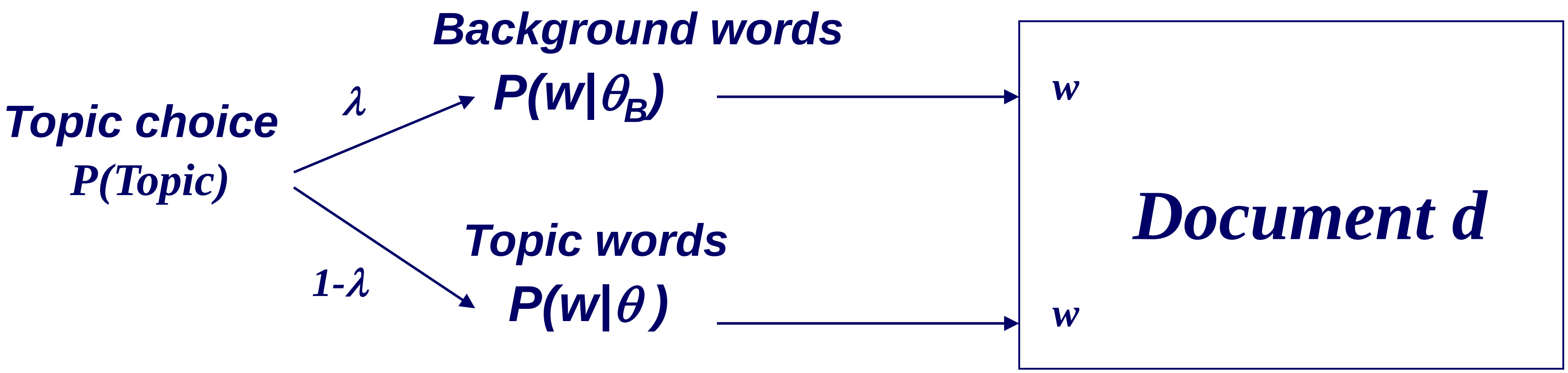
θ_d

the 0.031
a 0.018
…
text 0.04
mining 0.035
association 0.03
clustering 0.005
computer 0.0009
…
food 0.000001
…

Document Topic: $p(w|\theta_d)$

The Simplest Case:
One Topic + One Background Model

Assume $p(w|\theta_B)$ and λ are *known*
λ = mixing proportion of background topic in *d*



$$p(w) = \lambda p(w|\theta_B) + (1-\lambda) p(w|\theta)$$

$$\log p(d|\theta) = \sum_{w \in V} c(w,d) \log[\lambda p(w|\theta_B) + (1-\lambda) p(w|\theta)]$$

Expectation Maximization $\hat{\theta} = \arg\max_{\theta} \log p(d|\theta)$

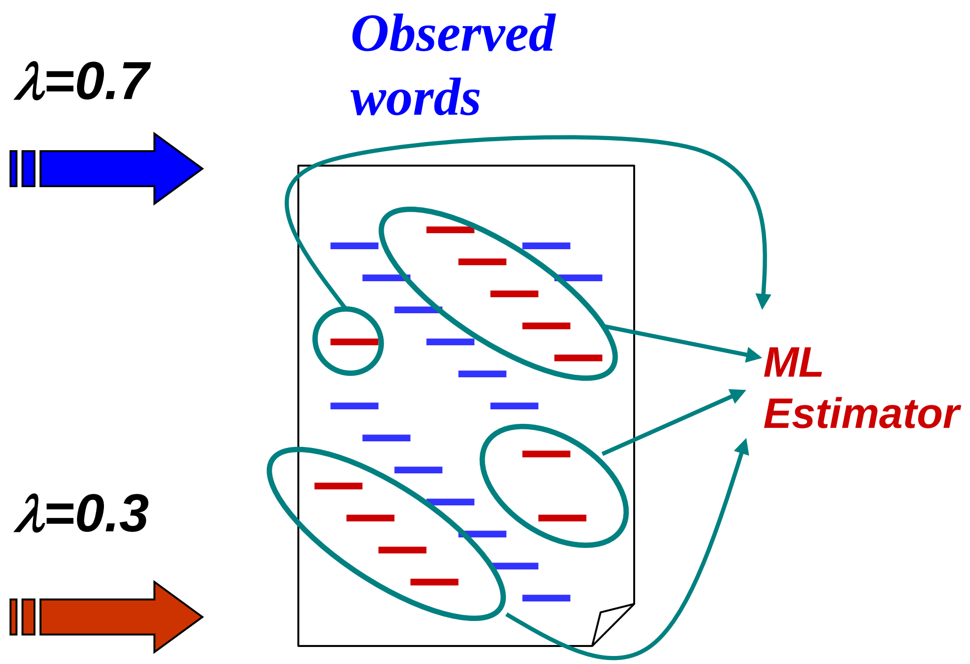
How to Estimate θ?

Known Background $p(w|\theta_B)$

the 0.2
a 0.1
we 0.01
to 0.02
…
text 0.0001
mining 0.00005
…

Unknown topic $p(w|\theta)$ for "Text mining"

…
text =?
mining =?
association =?
word =?
…



Suppose we know
the identity/label of each word ...

But we don't!

We guess the topic assignments

Assignment ("hidden") variable: $z_i \in \{1 \text{ (background)}, 0 \text{(topic)}\}$

	z_i
the	*1*
paper	*1*
presents	*1*
a	*1*
text	*0*
mining	*0*
algorithm	*0*
the	*1*
paper	*0*
...	*...*

Suppose the parameters are all known, what's a reasonable guess of z_i?

- ***depends on λ***
- ***depends on $p(w|\theta_B)$ and $p(w|\theta)$***

$$p(z_i=1|w_i)=\frac{p(z_i=1)\,p(w|z_i=1)}{p(z_i=1)\,p(w|z_i=1)+p(z_i=0)\,p(w|z_i=0)}$$

$$=\frac{\lambda p(w|\theta_B)}{\lambda p(w|\theta_B)+(1-\lambda)\,p^{current}(w|\theta)}$$

E-step

$$p^{new}(w_i|\theta)=\frac{c(w_i,d)(1-p(z_i=1|w_i))}{\sum_{w'\in V}c(w',d)(1-p(z_i=1|w'))}$$

M-step

θ_B and θ are competing for explaining words in document d!

Initially, set $p(w|\theta)$ to some random values, then iterate …

An example of EM computation

$$p^{(n)}(z_i=1|w_i)=\frac{\lambda p(w_i|\theta_B)}{\lambda p(w_i|\theta_B)+(1-\lambda)p^{(n)}(w_i|\theta)}$$

Expectation-Step:
Augmenting data by guessing hidden variables

$$p^{(n+1)}(w_i|\theta)=\frac{c(w_i,d)(1-p^{(n)}(z_i=1|w_i))}{\sum_{w_j\in vocabulary}c(w_j,d)(1-p^{(n)}(z_j=1|w_j))}$$

Maximization-Step
With the "augmented data", estimate parameters using maximum likelihood

Assume λ=0.5

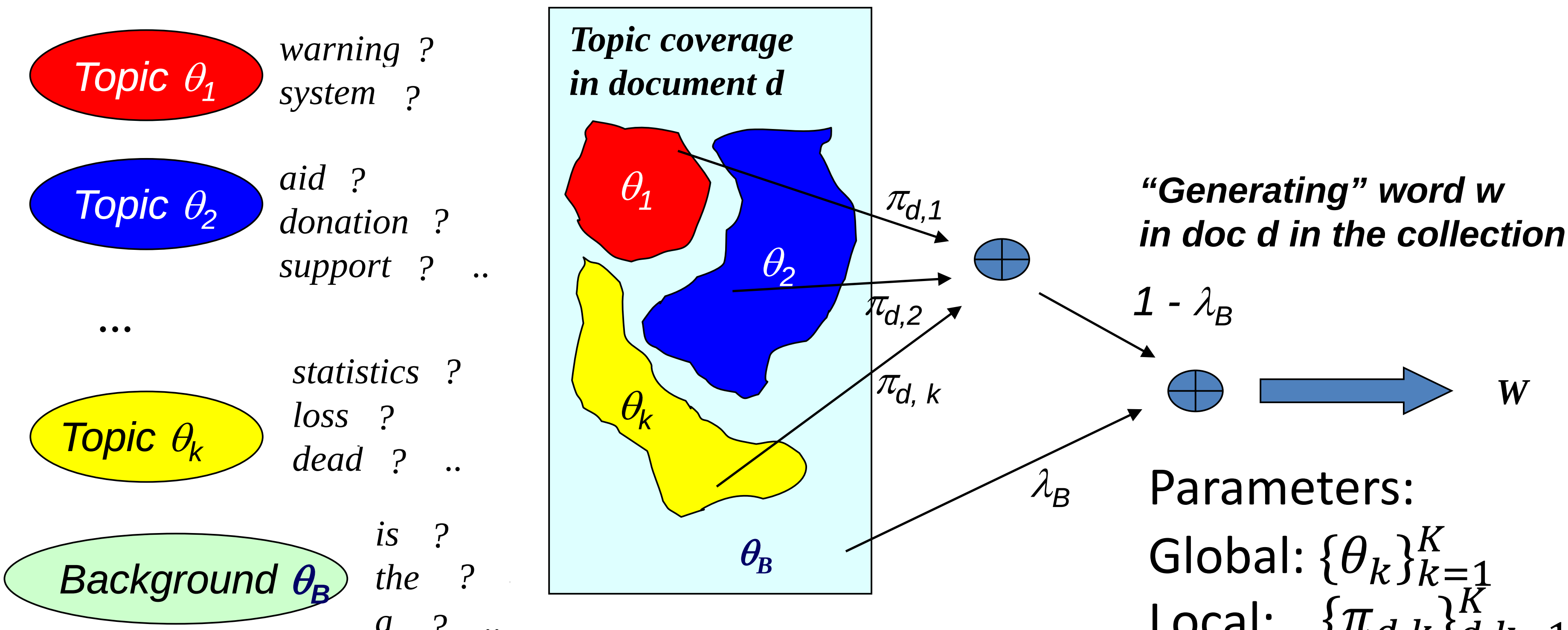
Word	#	$P(w\|\theta_B)$	Iteration 1		Iteration 2		Iteration 3	
			$P(w\|\theta)$	P(z=1)	$P(w\|\theta)$	P(z=1)	$P(w\|\theta)$	P(z=1)
The	4	0.5	**0.25**	0.67	**0.20**	0.71	**0.18**	0.74
Paper	2	0.3	**0.25**	0.55	**0.14**	0.68	**0.10**	0.75
Text	4	0.1	**0.25**	0.29	**0.44**	0.19	**0.50**	0.17
Mining	2	0.1	**0.25**	0.29	**0.22**	0.31	**0.22**	0.31
Log-Likelihood			-16.96		-16.13		-16.02	

Outline

1. General idea of topic models
2. **Basic topic models**
   - Probabilistic Latent Semantic Analysis (pLSA)
   - Latent Dirichlet Allocation (LDA)
3. Variants of topic models
4. Summary

Discover multiple topics in a collection

- Generalize the two topic mixture to *k* topics



Parameters:

Global: $\{\theta_k\}_{k=1}^{K}$

Local: $\{\pi_{d,k}\}_{d,k=1}^{K}$

Manual: λ_B

Probabilistic Latent Semantic Analysis
[Hofmann 99a, 99b]

- Topic: a multinomial distribution over words
- Document
  - Mixture of *k* topics
  - Mixing weights reflect the topic coverage
- Topic modeling
  - Word distribution under topic: $p(w|\theta)$
  - Topic coverage: $p(\pi|d)$

EM for estimating multiple topics

Known Background $p(w \mid \theta_B)$

the 0.2
a 0.1
we 0.01
to 0.02
…

Unknown topic model $p(w|\theta_1)$=?

"Text mining"

…
text =?
mining =?
association =?
word =?
…

…

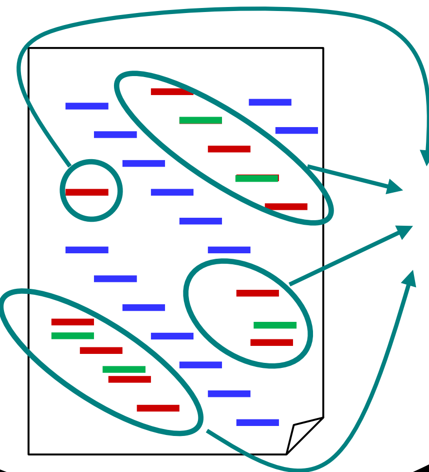
Unknown topic model $p(w|\theta_2)$=?

"information retrieval"

…
information =?
retrieval =?
query =?
document =?
…

E-Step:
Predict topic labels using Bayes Rule

Observed Words



M-Step:
ML Estimator based on "fractional counts"

Parameter estimation

E-Step:
Word w in doc d is generated
- ***from topic j***
- ***from background***

Posterior: application of Bayes rule

$$p(z_{d,w}=j)=\frac{\pi_{d,j}^{(n)}p^{(n)}(w|\theta_j)}{\sum_{j'=1}^{k}\pi_{d,j'}^{(n)}p^{(n)}(w|\theta_{j'})}$$

$$p(z_{d,w}=B)=\frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B)+(1-\lambda_B)\sum_{j=1}^{k}\pi_{d,j}^{(n)}p^{(n)}(w|\theta_j)}$$

M-Step:
Re-estimate
- ***mixing weights***
- ***word-topic distribution***

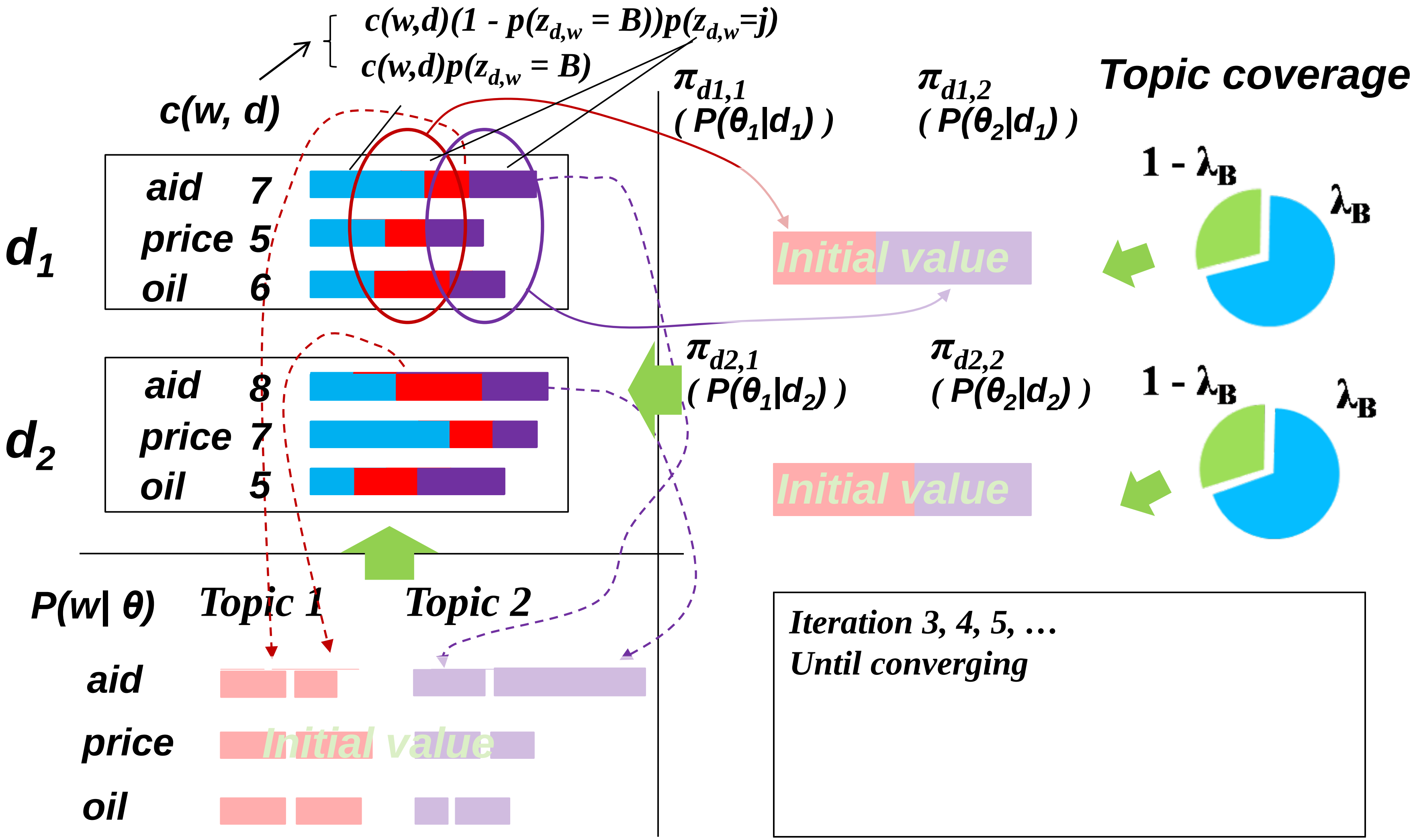
$$\pi_{d,j}^{(n+1)}=\frac{\sum_{w\in V}c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)}{\sum_{j'}\sum_{w\in V}c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j')}$$

$$p^{(n+1)}(w|\theta_j)=\frac{\sum_{d\in C}c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)}{\sum_{w'\in V}\sum_{d\in C}c(w',d)(1-p(z_{d,w'}=B))p(z_{d,w'}=j)}$$

Sum over all docs in the collection

Fractional counts contributing to
- ***using topic j in generating d***
- ***generating w from topic j***

How the algorithm works



Iteration 3, 4, 5, …
Until converging

Sample pLSA topics from TDT Corpus [Hofmann 99b]

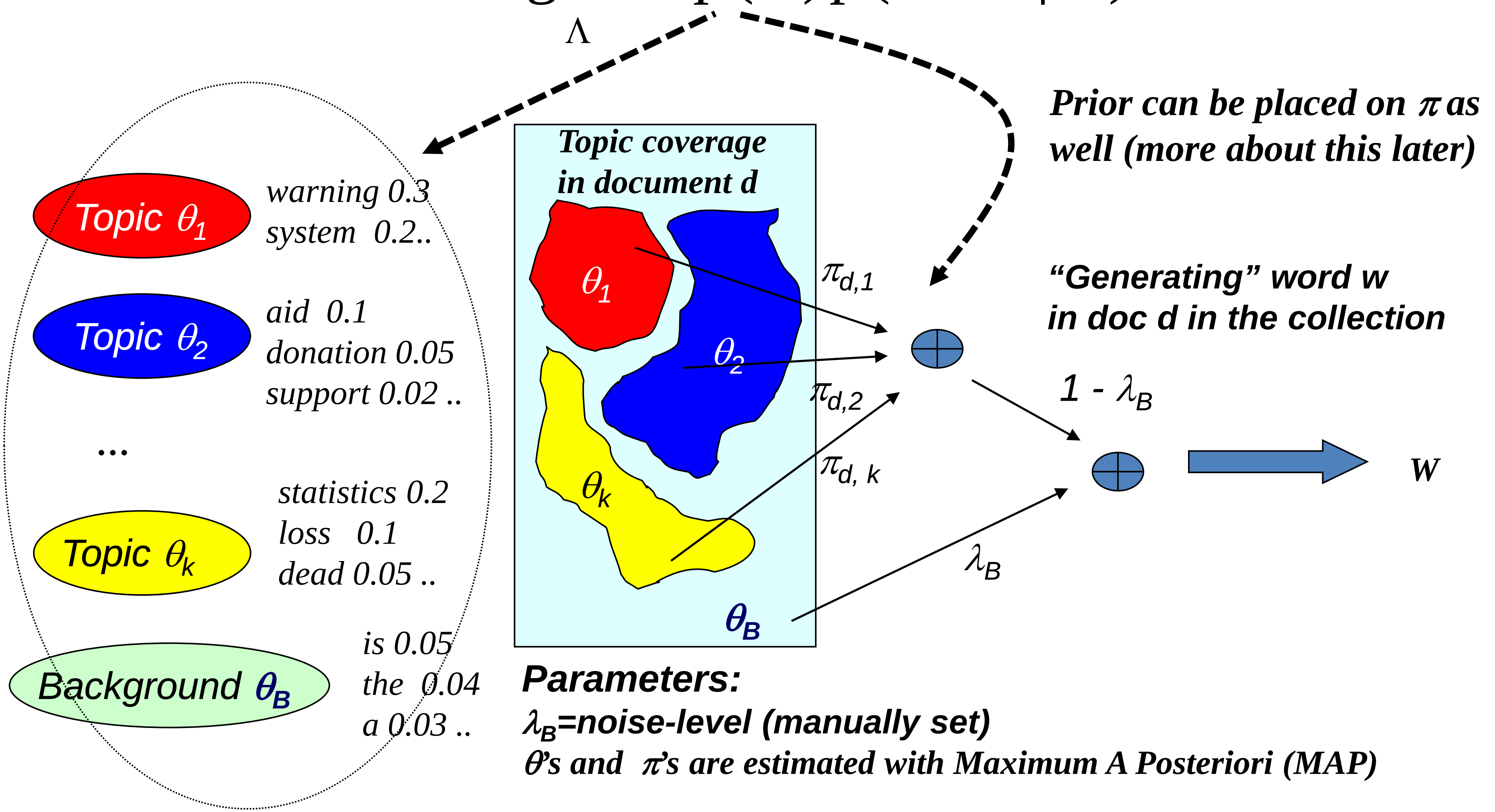
"plane"	"space shuttle"	"family"	"Hollywood"
plane	space	home	film
airport	shuttle	family	movie
crash	mission	like	music
flight	astronauts	love	new
safety	launch	kids	best
aircraft	station	mother	hollywood
air	crew	life	love
passenger	nasa	happy	actor
board	satellite	friends	entertainment
airline	earth	cnn	star

pLSA with prior knowledge

- What if we have some domain knowledge in mind
  - We want to see topics such as “battery” and “memory” for opinions about a laptop
  - We want words like “apple” and “orange” co-occur in a topic
  - One topic should be fixed to model background words (infinitely strong prior!)
- We can easily incorporate such knowledge as priors of pLSA model

Maximum a Posteriori (MAP) estimation

$$\Lambda^* = \arg\max_{\Lambda} p(\Lambda)\, p(Data \mid \Lambda)$$



Parameters:

λ_B=noise-level (manually set)

θ's and π's are estimated with Maximum A Posteriori (MAP)

MAP estimation

- Choosing conjugate priors
  - Dirichlet prior for multinomial distribution

Pseudo counts of w from prior θ'

$$p^{(n+1)}(w|\theta_j)=\frac{\sum_{d\in C}c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)+\mu p(w|\theta'_j)}{\sum_{w'\in V}\sum_{d\in C}c(w',d)(1-p(z_{d,w'}=B))p(z_{d,w'}=j)+\mu}$$

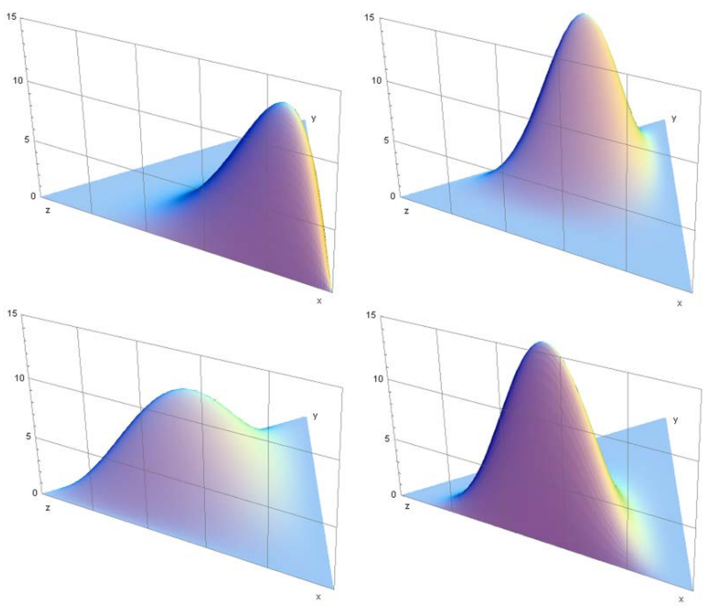
Sum of all pseudo counts

  - What if μ=0? What if μ=+∞?
  - A consequence of using conjugate prior is that the prior can be converted into "pseudo data" which can then be "merged" with the actual data for parameter estimation

Some background knowledge

- Conjugate prior
  - Posterior distribution in the same family as prior
- Dirichlet distribution
  - Continuous
  - Samples from it will be the parameters in a multinomial distribution

Gaussian -> Gaussian
Beta -> Binomial
Dirichlet -> Multinomial



Prior as pseudo counts

Known Background $p(w \mid B)$

the 0.2
a 0.1
we 0.01
to 0.02
…

Unknown topic model $p(w|\theta_1)$=?

"Text mining"

…
text =?
mining =?
association =?
word =?
…

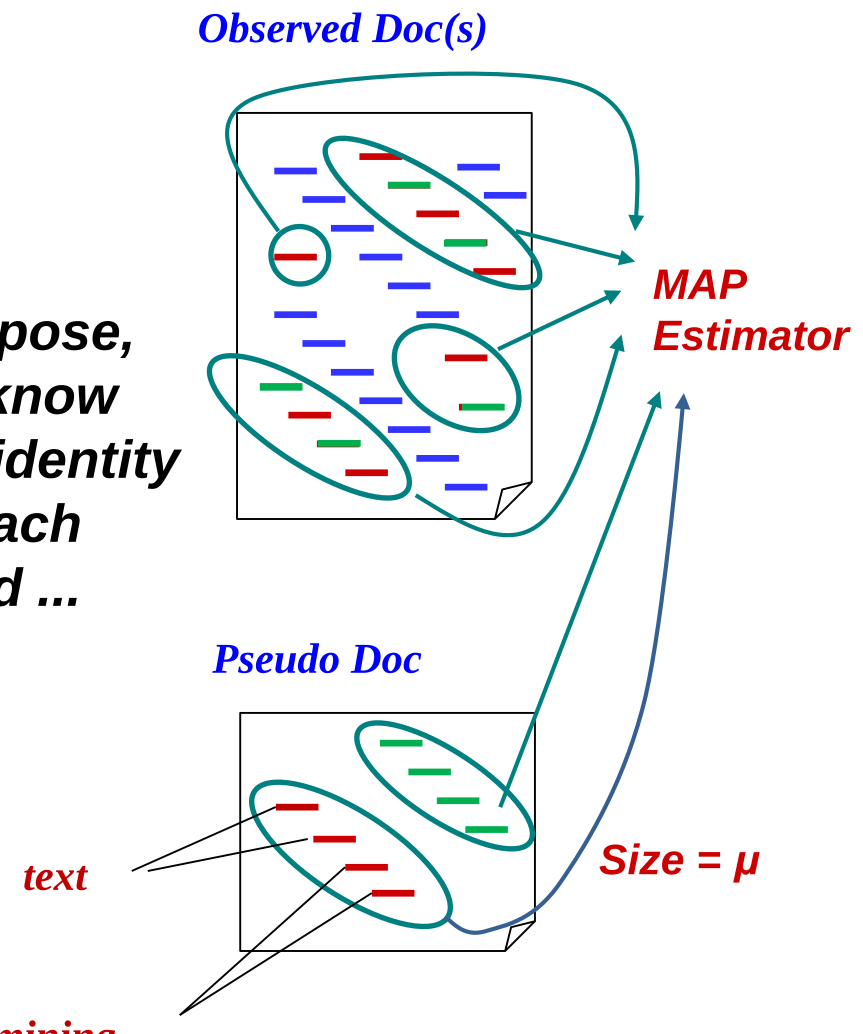
…

Unknown topic model $p(w|\theta_2)$=?

"information retrieval"

…
information =?
retrieval =?
query =?
document =?
…

Suppose, we know the identity of each word ...



Deficiency of pLSA

- Not a fully generative model
  - Can't compute probability of a new document
    - Topic coverage $p(\pi|d)$ is per-document estimated
  - Heuristic workaround is possible
- Many parameters ➔ high complexity of models
  - Many local maxima
  - Prone to overfitting

Latent Dirichlet Allocation [Blei et al. 02]

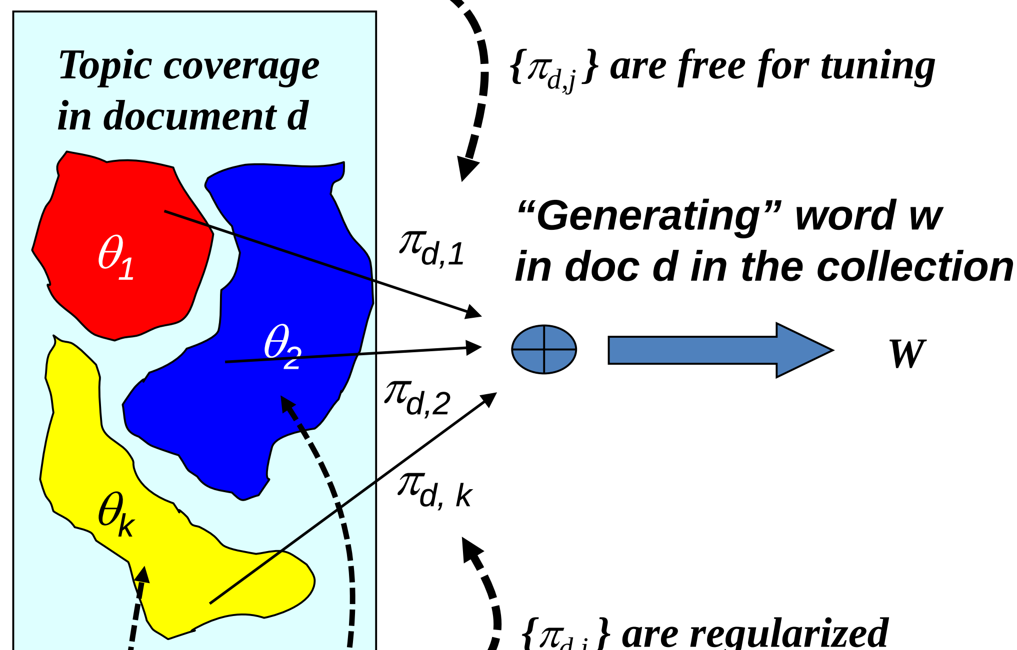
- Make pLSA a fully generative model by imposing Dirichlet priors
  - Dirichlet priors over $p(\pi|d)$
  - Dirichlet priors over $p(w|\theta)$
  - A Bayesian version of pLSA
- Provide mechanism to deal with new documents
  - Flexible to model many other observations in a document

LDA = Imposing Prior on PLSA

pLSA:

Topic coverage $\pi_{d,j}$ is specific to each "training document", thus can't be used to generate a new document

$\{\pi_{d,j}\}$ are free for tuning



$\{\pi_{d,j}\}$ are regularized

LDA:

Topic coverage distribution $\{\pi_{d,j}\}$ for any document is sampled from a Dirichlet distribution, allowing for generating a new doc

$$p(\vec{\pi}_d) = Dirichlet(\vec{\alpha})$$

In addition, the topic word distributions $\{\theta_j\}$ are also drawn from another Dirichlet prior

$$p(\vec{\theta}_i) = Dirichlet(\vec{\beta})$$

Magnitudes of α and β determine the variances of the prior, thus also the concentration of prior (larger α and β → stronger prior)

pLSA v.s. LDA

pLSA

$$\underline{p_d(w|\{\theta_j\},\{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j}\, p(w|\theta_j)}$$

$$\log p(d|\{\theta_j\},\{\pi_{d,j}\}) = \boxed{\sum_{w \in V} c(w,d) \log[\sum_{j=1}^{k} \pi_{d,j}\, p(w|\theta_j)]}$$

$$\log p(C|\{\theta_j\},\{\pi_{d,j}\}) = \sum_{d \in C} \log p(d|\{\theta_j\},\{\pi_{d,j}\})$$

Core assumption in all topic models

pLSA component

LDA

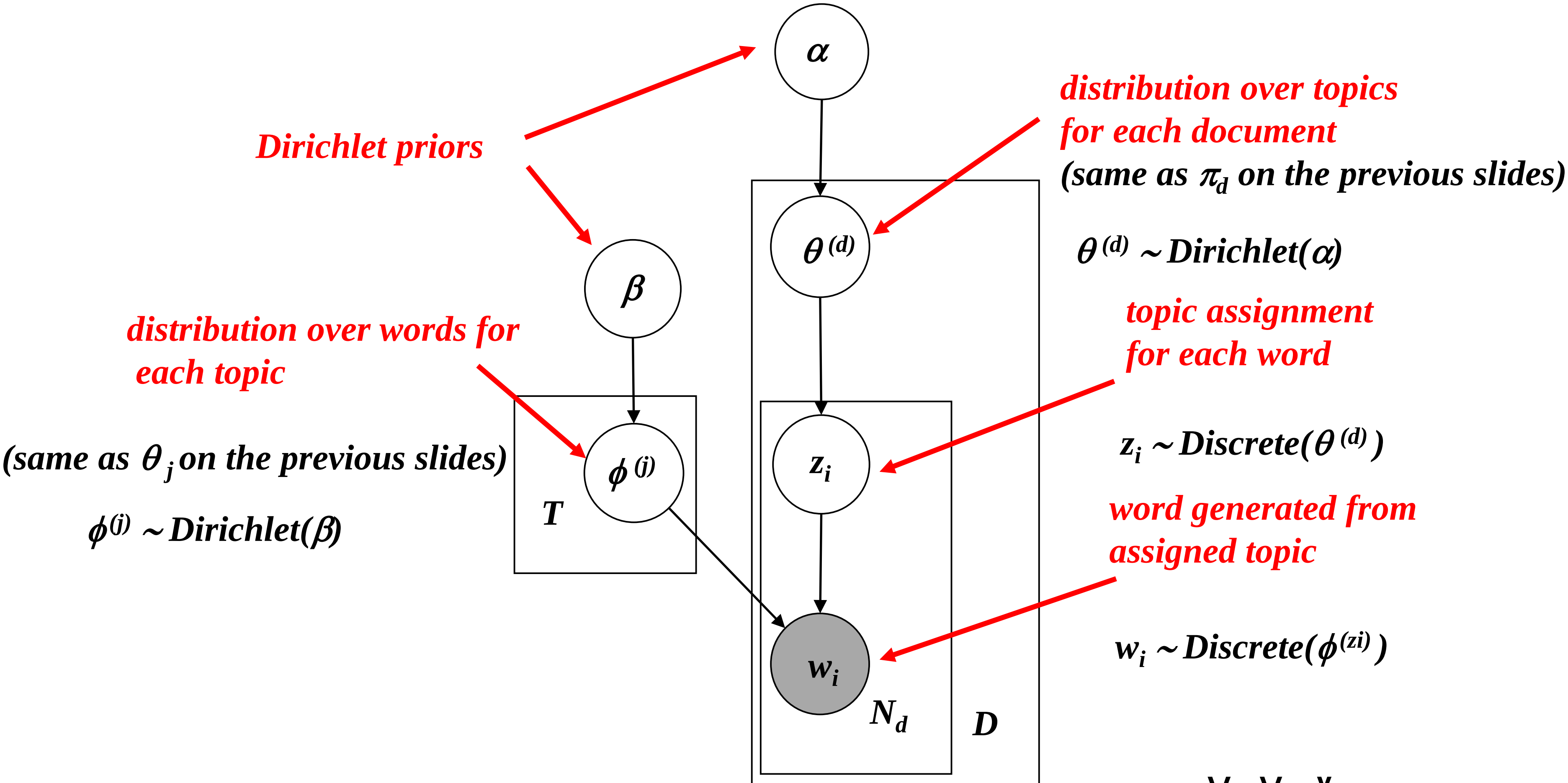
$$\underline{p_d(w|\{\theta_j\},\{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j}\, p(w|\theta_j)}$$

$$\log p(d|\vec{\alpha},\{\theta_j\}) = \boxed{\int \sum_{w \in V} c(w,d) \log[\sum_{j=1}^{k} \pi_{d,j}\, p(w|\theta_j)]}\, \underline{p(\vec{\pi}_d|\vec{\alpha})\, d\vec{\pi}_d}$$

$$\log p(C|\vec{\alpha},\vec{\beta}) = \int \sum_{d \in C} \log p(d|\vec{\alpha},\{\theta_j\}) \underline{\prod_{j=1}^{k} p(\theta_j|\vec{\beta})\, d\theta_1 \ldots d\theta_k}$$

Regularization added by LDA

LDA as a graphical model [Blei et al. 03a]



Most approximate inference algorithms aim to infer $p(z_i \mid \vec{w}, \vec{\alpha}, \vec{\beta})$ from which other interesting variables can be easily computed

Approximate inferences for LDA

- Deterministic approximation
  - Variational inference
  - Expectation propagation
- Markov chain Monte Carlo
  - Full Gibbs sampler
  - Collapsed Gibbs sampler

Most efficient, and quite popular, but can only work with conjugate prior

Collapsed Gibbs sampling [Griffiths & Steyvers 04]

- Using conjugacy between Dirichlet and multinomial distributions, integrate out continuous random variables

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

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

- Define a distribution on topic assignment **z**

With fixed assignment of **z**

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

Collapsed Gibbs sampling [Griffiths & Steyvers 04]

- Sample each z_i conditioned on $\mathbf{z}_{-i}$ ← *All the other words beside z_i*

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

Word-topic distribution **Topic proportion**

  - Implementation: counts can be cached in two sparse matrices; no special functions, simple arithmetic
  - Distributions on Φ and Θ can be analytic computed given z and w

Gibbs sampling in LDA

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

Gibbs sampling in LDA

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

Gibbs sampling in LDA

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

words in d_i assigned with topic j

Count of instances where w_i is assigned with topic j

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

Count of all words assigned with topic j

words in d_i assigned with any topic

Gibbs sampling in LDA

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

What's the most likely topic for w_i in d_i?

How likely would d_i choose topic j?

How likely would topic j generate word w_i ?

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

Gibbs sampling in LDA

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

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

Gibbs sampling in LDA

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

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

Gibbs sampling in LDA

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

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

Gibbs sampling in LDA

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

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

Gibbs sampling in LDA

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

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

Topics learned by LDA

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

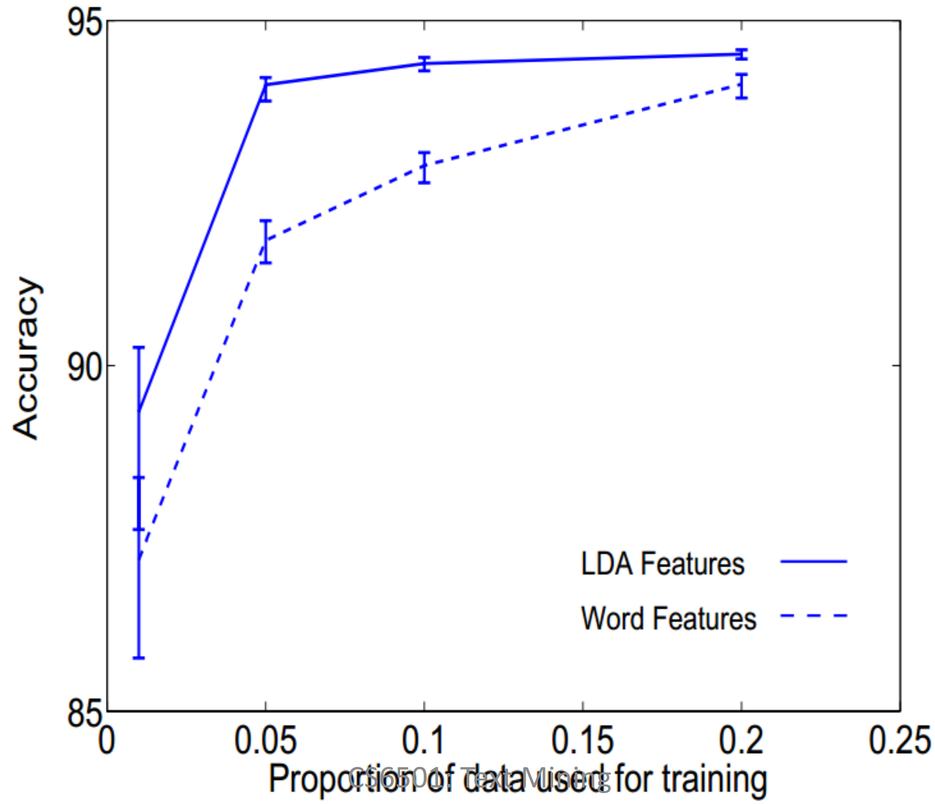
Topic assignments in document

- Based on the topics shown in last slide

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

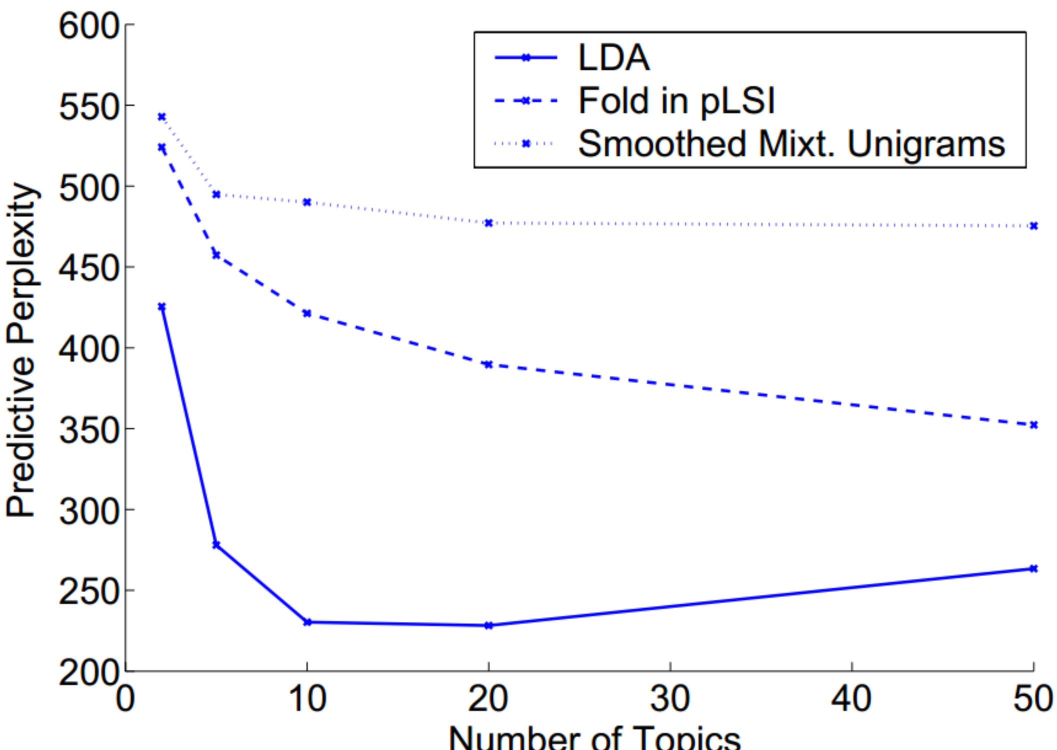
Application of learned topics

- Document classification
  – A new type of feature representation



Application of learned topics

- Collaborative filtering
  - A new type of user profile

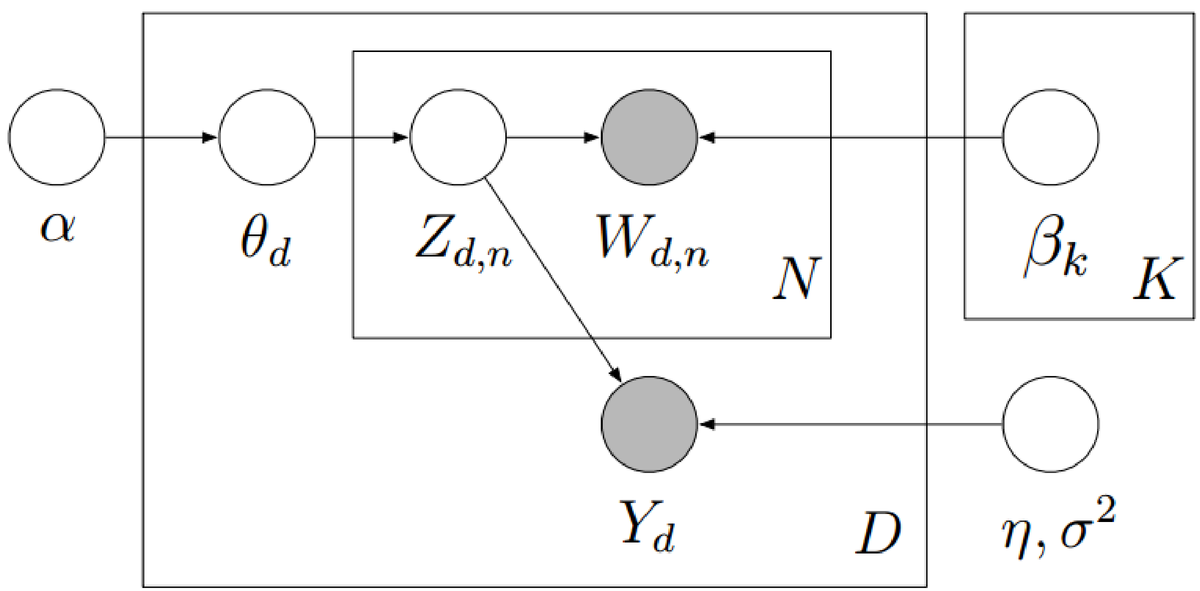


Outline

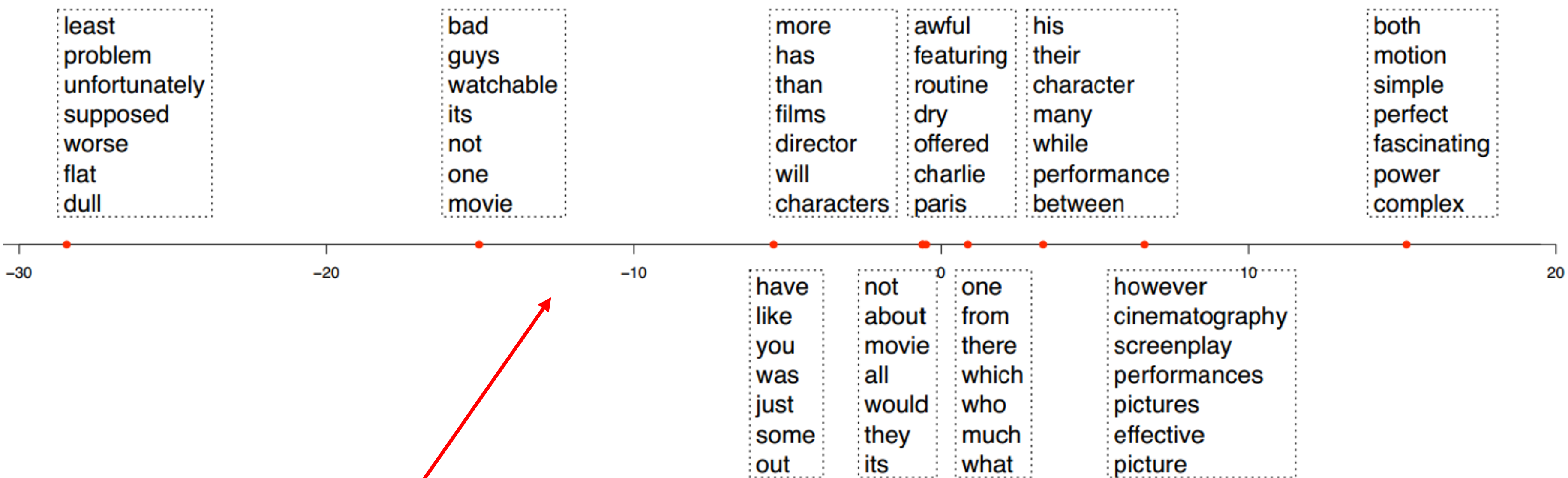
1. General idea of topic models
2. Basic topic models
   - Probabilistic Latent Semantic Analysis (pLSA)
   - Latent Dirichlet Allocation (LDA)
3. **Variants of topic models**
4. Summary

Supervised Topic Model [Blei & McAuliffe, NIPS'02]

- A generative model for classification
  - Topic generates both words and labels



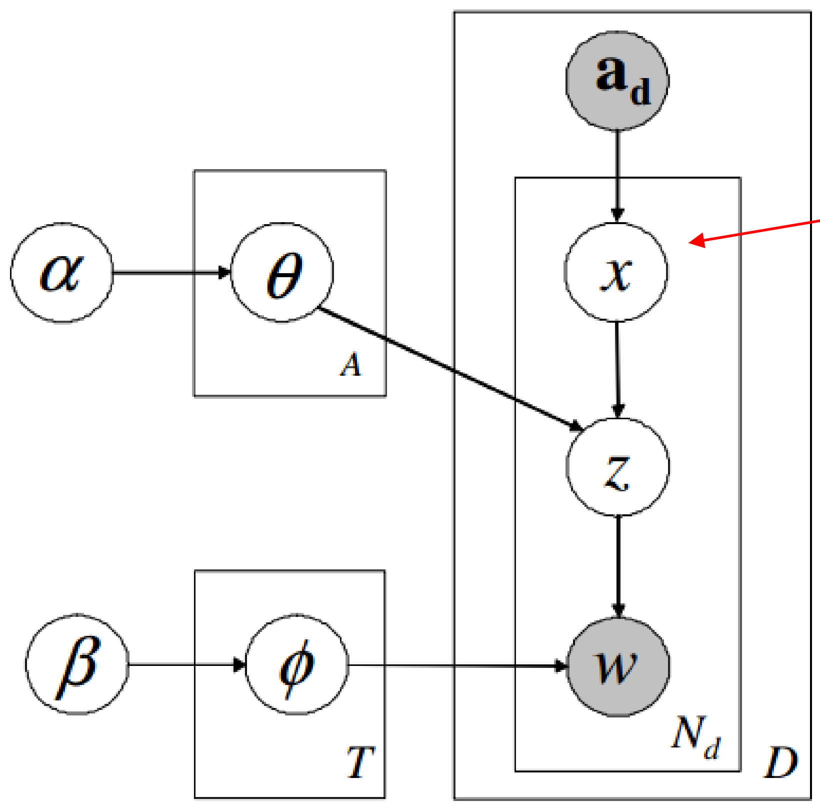
Sentiment polarity of topics



Sentiment polarity learned from classification model

Author Topic Model [Rosen-Zvi UAI'04]

- Authorship determines the topic mixture



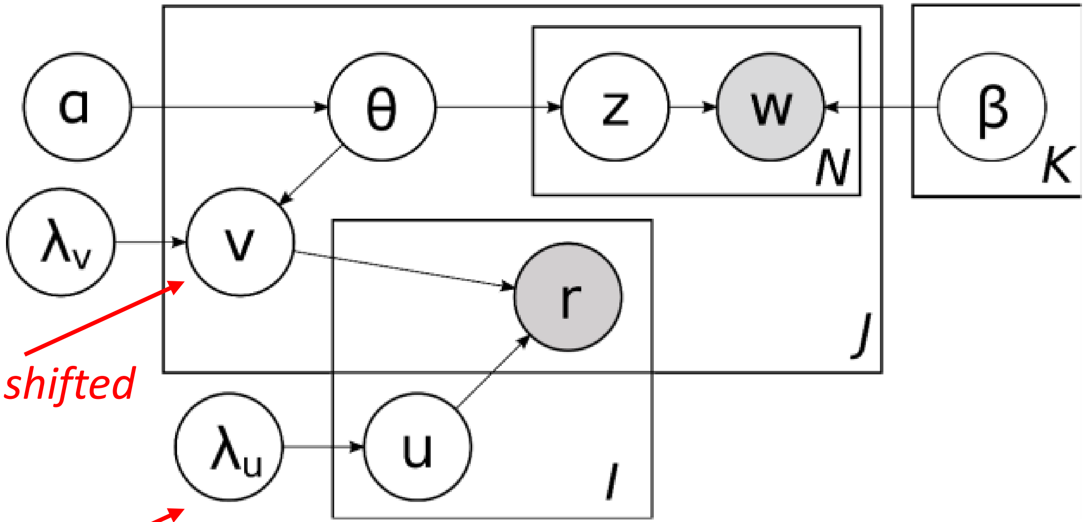
Each author chooses his/her topic to contribute in the document

Learned association between words and authors

TOPIC 19		TOPIC 24		TOPIC 29		TOPIC 87	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
LIKELIHOOD	0.0539	RECOGNITION	0.0400	REINFORCEMENT	0.0411	KERNEL	0.0683
MIXTURE	0.0509	CHARACTER	0.0336	POLICY	0.0371	SUPPORT	0.0377
EM	0.0470	CHARACTERS	0.0250	ACTION	0.0332	VECTOR	0.0257
DENSITY	0.0398	TANGENT	0.0241	OPTIMAL	0.0208	KERNELS	0.0217
GAUSSIAN	0.0349	HANDWRITTEN	0.0169	ACTIONS	0.0208	SET	0.0205
ESTIMATION	0.0314	DIGITS	0.0159	FUNCTION	0.0178	SVM	0.0204
LOG	0.0263	IMAGE	0.0157	REWARD	0.0165	SPACE	0.0188
MAXIMUM	0.0254	DISTANCE	0.0153	SUTTON	0.0164	MACHINES	0.0168
PARAMETERS	0.0209	DIGIT	0.0149	AGENT	0.0136	REGRESSION	0.0155
ESTIMATE	0.0204	HAND	0.0126	DECISION	0.0118	MARGIN	0.0151
AUTHOR	**PROB.**	**AUTHOR**	**PROB.**	**AUTHOR**	**PROB.**	**AUTHOR**	**PROB.**
Tresp_V	0.0333	Simard_P	0.0694	Singh_S	0.1412	Smola_A	0.1033
Singer_Y	0.0281	Martin_G	0.0394	Barto_A	0.0471	Scholkopf_B	0.0730
Jebara_T	0.0207	LeCun_Y	0.0359	Sutton_R	0.0430	Burges_C	0.0489
Ghahramani_Z	0.0196	Denker_J	0.0278	Dayan_P	0.0324	Vapnik_V	0.0431
Ueda_N	0.0170	Henderson_D	0.0256	Parr_R	0.0314	Chapelle_O	0.0210
Jordan_M	0.0150	Revow_M	0.0229	Dietterich_T	0.0231	Cristianini_N	0.0185
Roweis_S	0.0123	Platt_J	0.0226	Tsitsiklis_J	0.0194	Ratsch_G	0.0172
Schuster_M	0.0104	Keeler_J	0.0192	Randlov_J	0.0167	Laskov_P	0.0169
Xu_L	0.0098	Rashid_M	0.0182	Bradtke_S	0.0161	Tipping_M	0.0153
Saul_L	0.0094	Sackinger_E	0.0132	Schwartz_A	0.0142	Sollich_P	0.0141

Collaborative Topic Model [Wang & Blei,KDD'11]

- Collaborative filtering in topic space
  - User's preference over topics determines his/her rating for the item



Topics for the item, shifted with random noise

User profile over topical space

Topic-based recommendation

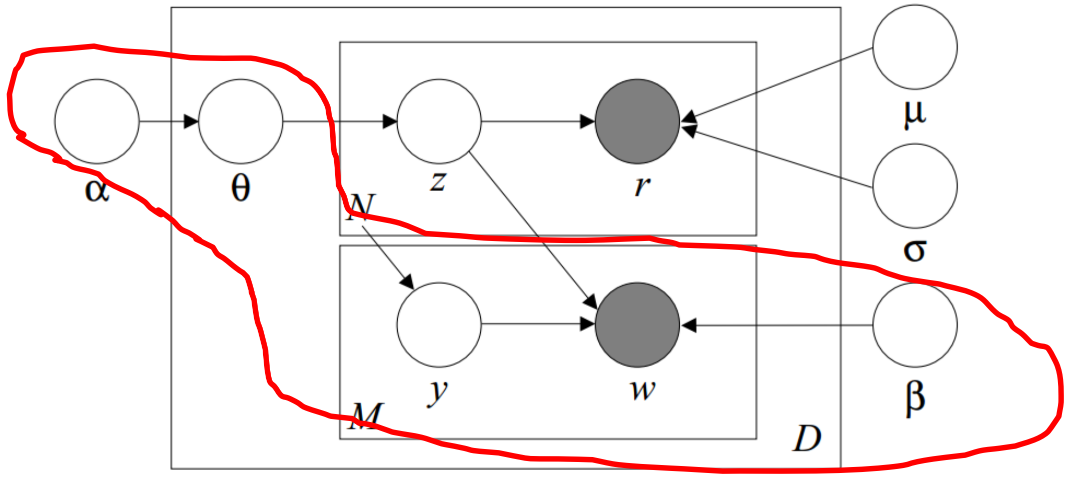
	user I	in user's lib?
top 3 topics	1. image, measure, measures, images, motion, matching, transformation, entropy, overlap, computed, match 2. learning, machine, training, vector, learn, machines, kernel, learned, classifiers, classifier, generalization 3. sets, objects, defined, categories, representations, universal, category, attributes, consisting, categorization	
top 10 articles	1. Information theory inference learning algorithms	✓
	2. Machine learning in automated text categorization	✓
	3. Artificial intelligence a modern approach	×
	4. Data xmining: practical machine learning tools and techniques	×
	5. Statistical learning theory	×
	6. Modern information retrieval	✓
	7. Pattern recognition and machine learning, information science and statistics	✓
	8. Recognition by components: a theory of human image understanding	×
	9. Data clustering a review	✓
	10. Indexing by latent semantic analysis	✓
	user II	in user's lib?
top 3 topics	1. users, user, interface, interfaces, needs, explicit, implicit, usability, preferences, interests, personalized 2. based, world, real, characteristics, actual, exploring, exploration, quite, navigation, possibilities, dealing 3. evaluation, collaborative, products, filtering, product, reviews, items, recommendations, recommender	
top 10 articles	1. Combining collaborative filtering with personal agents for better recommendations	×
	2. An adaptive system for the personalized access to news	✓
	3. Implicit interest indicators	×
	4. Footprints history-rich tools for information foraging	✓
	5. Using social tagging to improve social navigation	✓
	6. User models for adaptive hypermedia and adaptive educational systems	✓
	7. Collaborative filtering recommender systems	✓
	8. Knowledge tree: a distributed architecture for adaptive e-learning	✓
	9. Evaluating collaborative filtering recommender systems	✓
	10. Personalizing search via automated analysis of interests and activities	✓

Correspondence Topic Model [Blei SIGIR'03]

- Simultaneously modeling the generation of multiple types of observations
  - E.g., image and corresponding text annotations

Correspondence part (can be described with different distributions)

LDA part



Annotation results

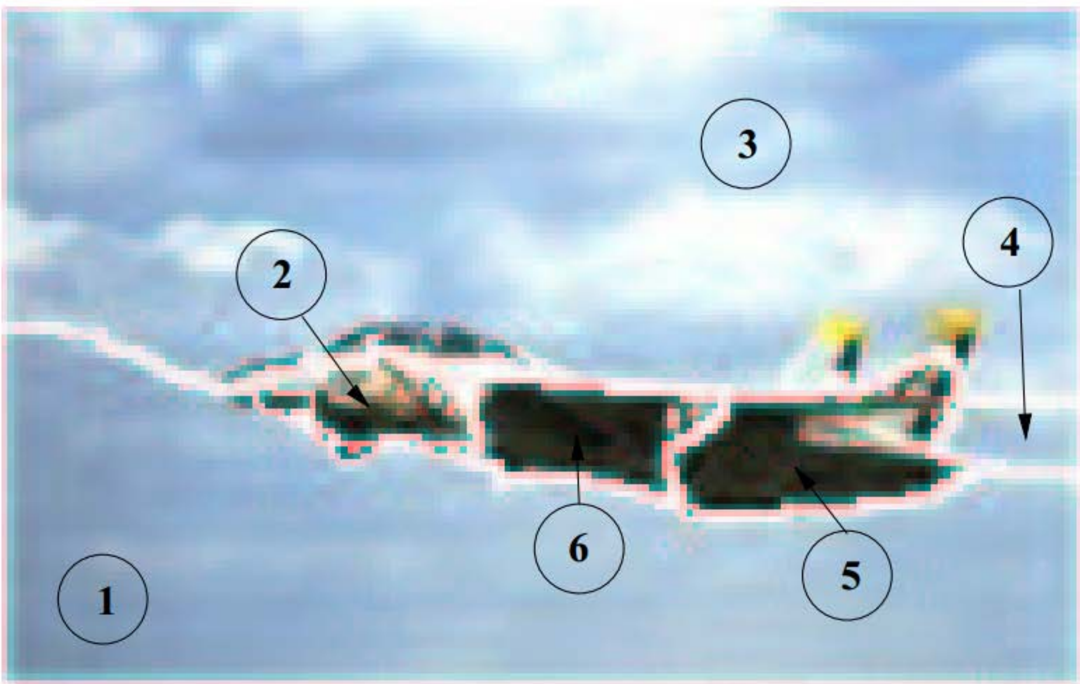
True caption
market people
Corr−LDA
people market pattern textile display

True caption
scotland water
Corr−LDA
scotland water flowers hills tree

True caption
birds tree
Corr−LDA
birds nest leaves branch tree

True caption
fish reefs water
Corr−LDA
fish water ocean tree coral

Annotation results

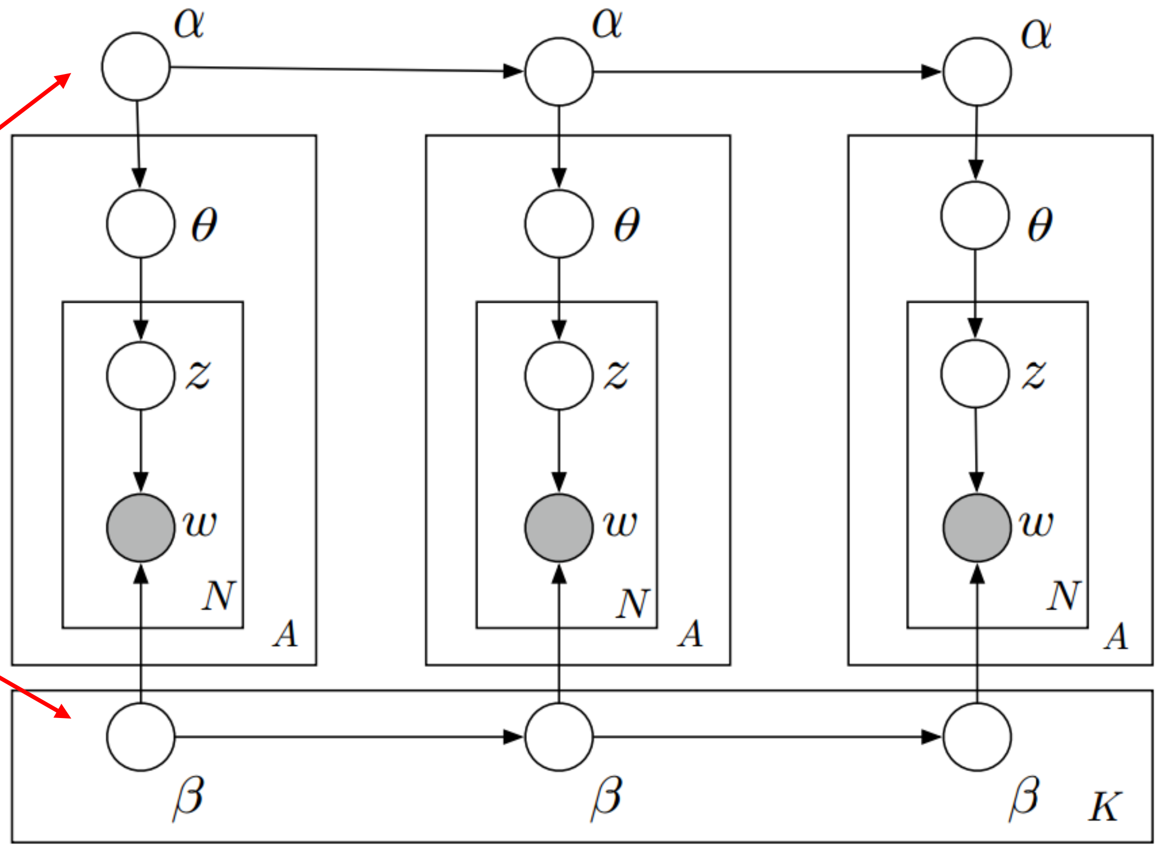


Corr−LDA:
1. PEOPLE, TREE
2. SKY, JET
3. SKY, CLOUDS
4. SKY, MOUNTAIN
5. PLANE, JET
6. PLANE, JET

Dynamic Topic Model [Blei ICML'06]

- Capture the evolving topics over time

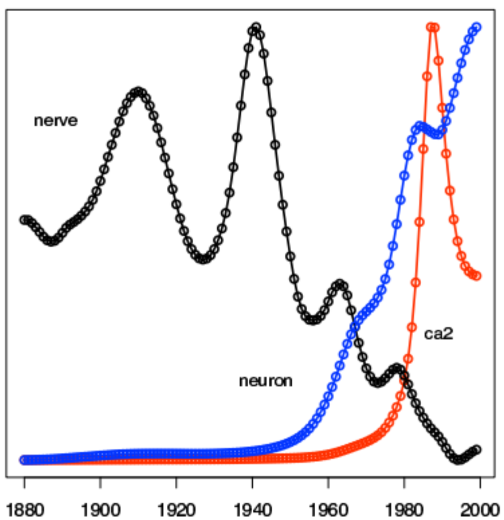
Markov assumption about the topic dynamics



Evolution of topics

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
brain	movement	brain	movement	movement	stimulate	record	respons	response	respons	cell	cell	neuron
movement	eye	eye	brain	sound	muscle	nerve	record	stimulate	cell	neuron	channel	active
action	right	movement	sound	muscle	sound	stimulate	stimulate	record	potential	response	neuron	brain
right	hand	right	nerve	active	movement	response	nerve	condition	stimul	active	ca2	cell
eye	brain	left	active	nerve	response	muscle	muscle	active	neuron	brain	active	fig
hand	left	hand	muscle	stimulate	nerve	electrode	active	potential	active	stimul	brain	response
left	action	nerve	left	fiber	frequency	active	frequency	stimulus	nerve	muscle	receptor	channel
muscle	muscle	vision	eye	reaction	fiber	brain	electrode	nerve	eye	system	muscle	receptor
nerve	sound	sound	right	brain	active	fiber	potential	subject	record	nerve	respons	synapse
sound	experiment	muscle	nervous	response	brain	potential	study	eye	abstract	receptor	current	signal

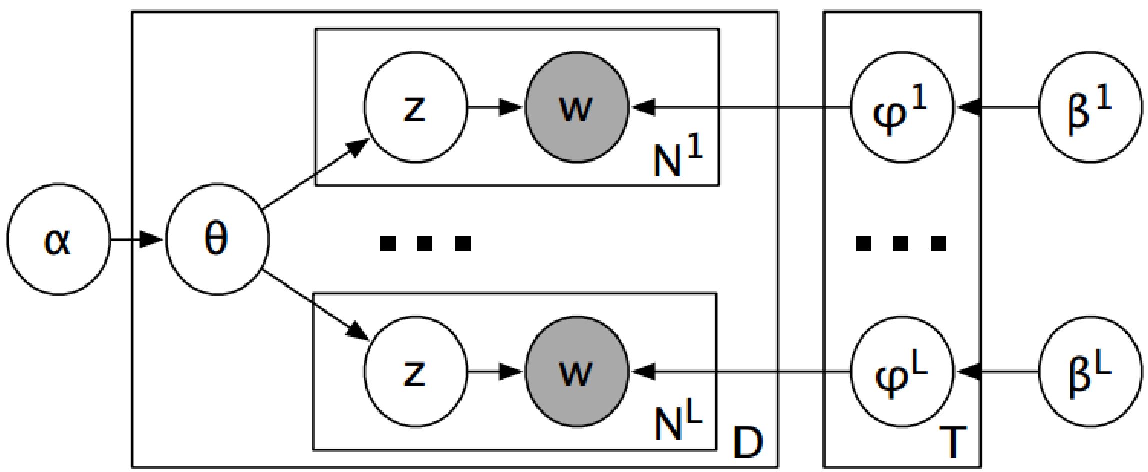
"Neuroscience"



1887 Mental Science
1900 Hemianopsia in Migraine
1912 A Defence of the ``New Phrenology"
1921 The Synchronal Flashing of Fireflies
1932 Myoesthesis and Imageless Thought
1943 Acetylcholine and the Physiology of the Nervous System
1952 Brain Waves and Unit Discharge in Cerebral Cortex
1963 Errorless Discrimination Learning in the Pigeon
1974 Temporal Summation of Light by a Vertebrate Visual Receptor
1983 Hysteresis in the Force-Calcium Relation in Muscle
1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

Polylingual Topic Models [Mimmo et al., EMNLP'09]

- Assumption: topics are universal over languages
  - Correspondence between documents are known
    - E.g., news report about the same event in different languages

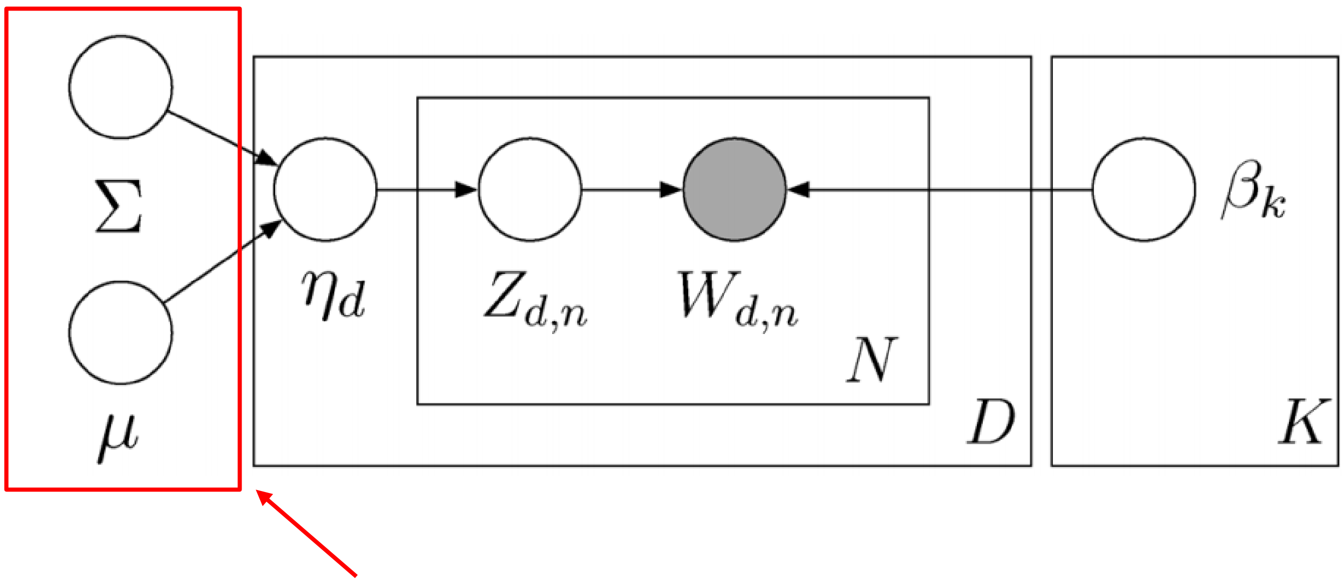


Topics learned in different languages

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

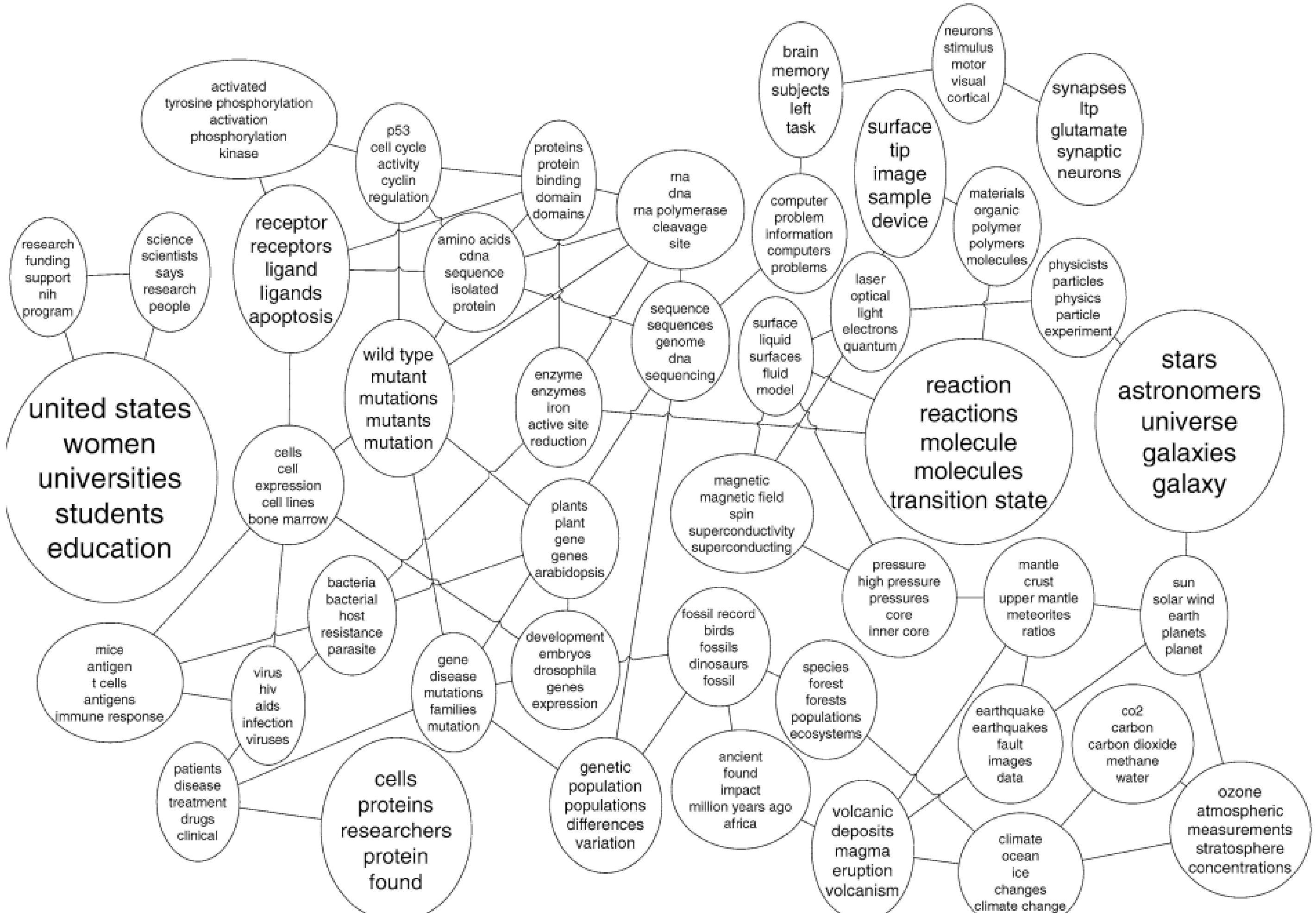
Correlated Topic Model [Blei & Lafferty, Annals of Applied Stat'07]

- Non-conjugate priors to capture correlation between topics



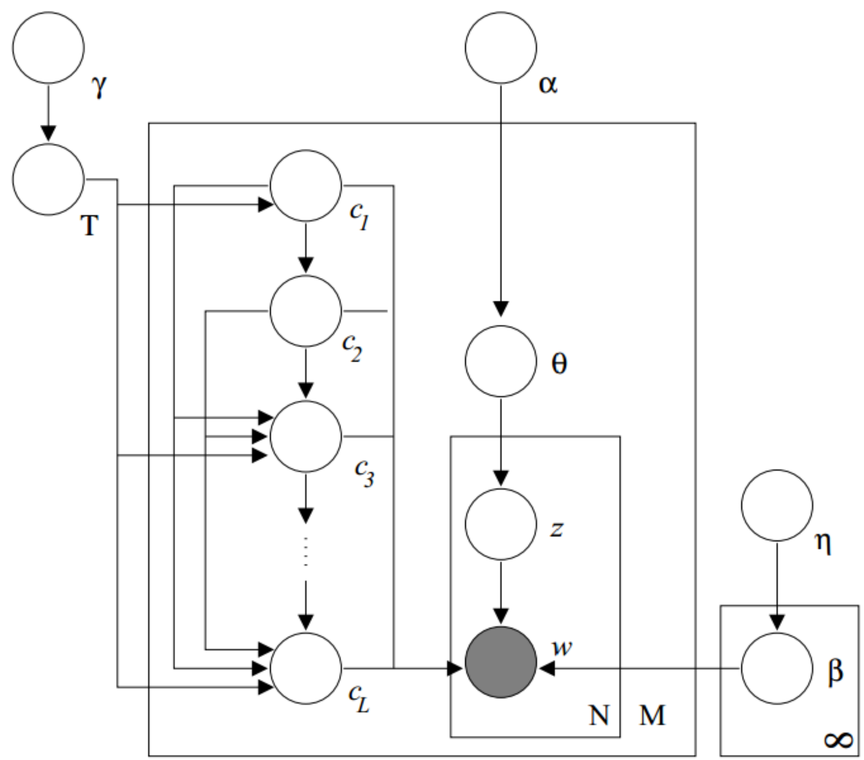
Gaussian as the prior for topic proportion (increase the computational complexity)

Learned structure of topics

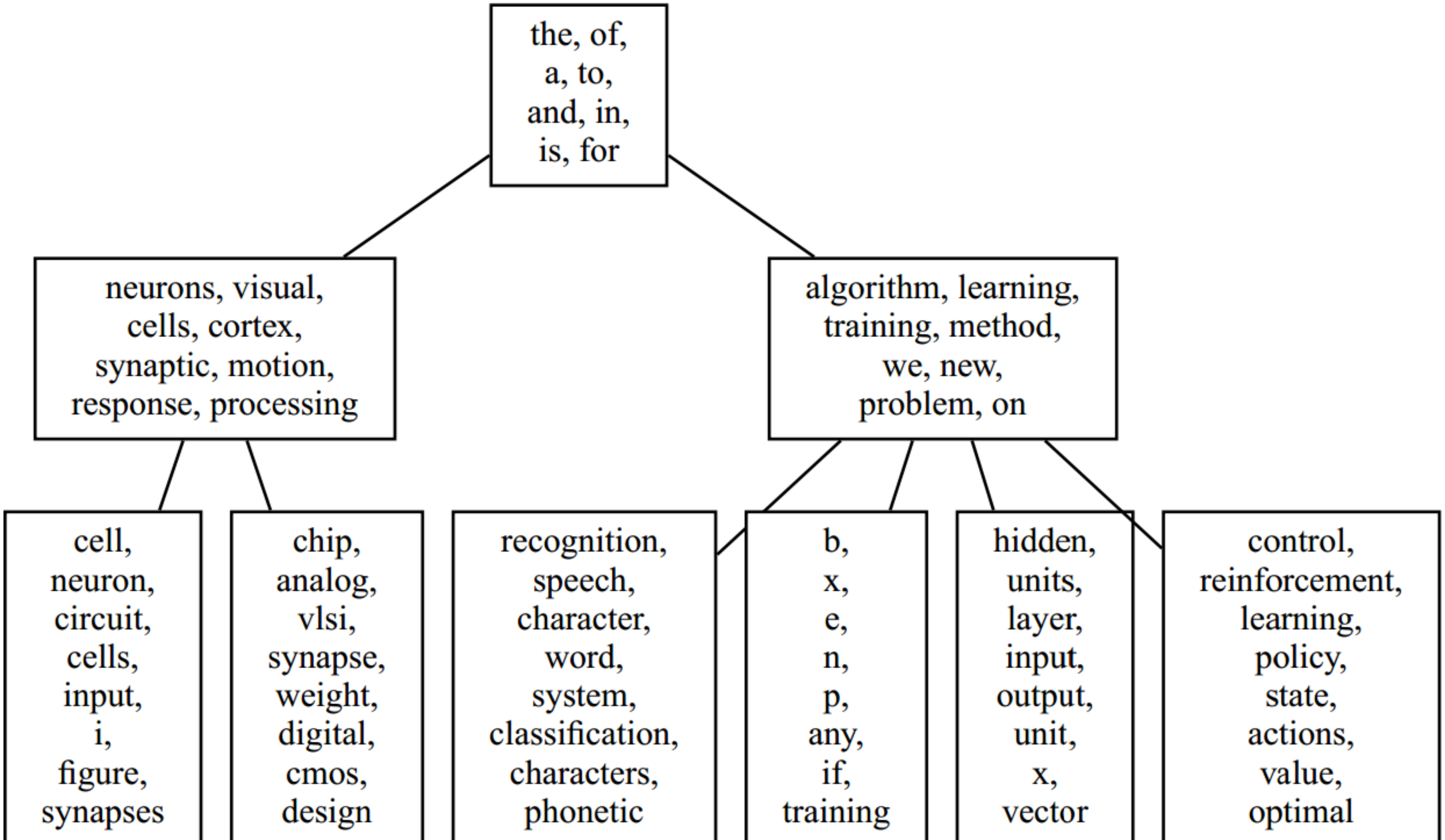


Hierarchical Topic Models [Blei et al. NIPS'04]

- Nested Chinese restaurant process as a prior for topic assignment



Hierarchical structure of topics



Outline

1. General idea of topic models
2. Basic topic models
   - Probabilistic Latent Semantic Analysis (pLSA)
   - Latent Dirichlet Allocation (LDA)
3. Variants of topic models
4. **Summary**

Summary

- Probabilistic Topic Models are a new family of document modeling approaches, especially useful for
  - Discovering latent topics in text
  - Analyzing latent structures and patterns of topics
  - Extensible for joint modeling and analysis of text and associated non-textual data
- pLSA & LDA are two basic topic models that tend to function similarly, with LDA better as a generative model
- Many different models have been proposed with probably many more to come
- Many demonstrated applications in multiple domains and many more to come

Summary

- However, all topic models suffer from the problem of multiple local maxima
  - Make it hard/impossible to reproduce research results
  - Make it hard/impossible to interpret results in real applications
- Complex models can't scale up to handle large amounts of text data
  - Collapsed Gibbs sampling is efficient, but only working for conjugate priors
  - Variational EM needs to be derived in a model-specific way
  - Parallel algorithms are promising
- Many challenges remain….