

# Graphical models and topic modeling

Ho Tu Bao

Japan Advance Institute of Science and Technology

John von Neumann Institute, VNU-HCM

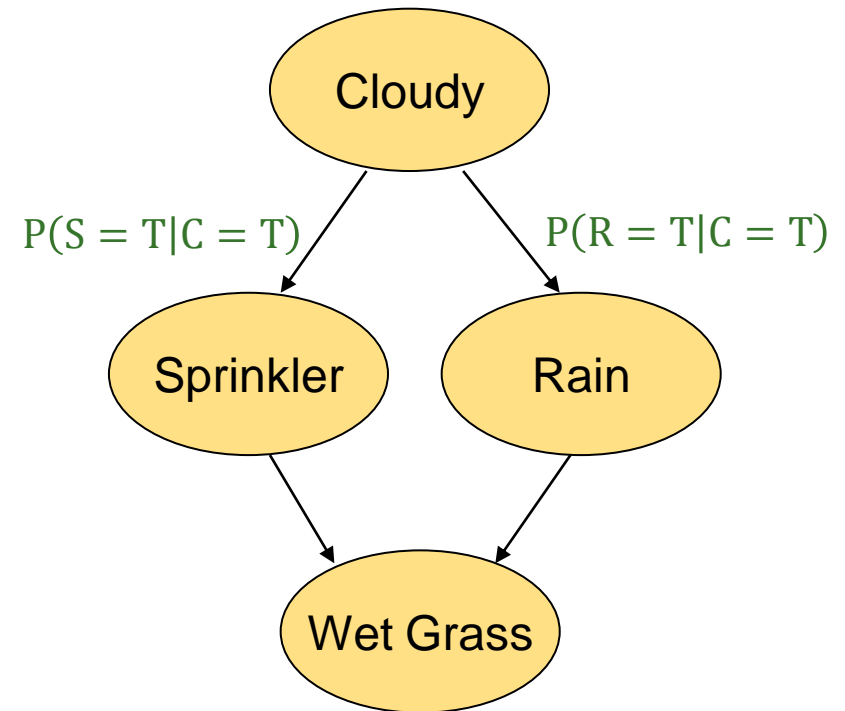
# Content

- Brief overview of graphical models
- Introduction to topic models
- Fully sparse topic model
- Conditional random fields in NLP

# Graphical models

## *What causes grass wet?*

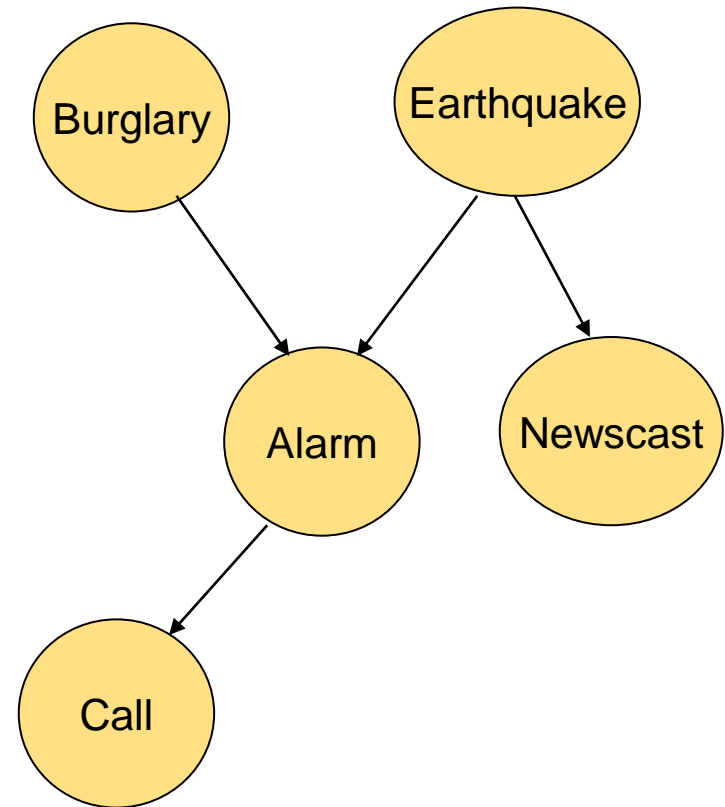
- Mr. Holmes leaves his house
  - ❑ The grass is wet in front of his house.
  - ❑ Two reasons are possible: either it rained or the sprinkler of Holmes has been on during the night.
- Then, Mr. Holmes looks at the sky and finds it is cloudy
  - ❑ Since when it is cloudy, usually the sprinkler is off and it is more possible it rained.
  - ❑ He concludes it is more likely that rain causes grass wet.



# Graphical models

## *Earthquake or burglary?*

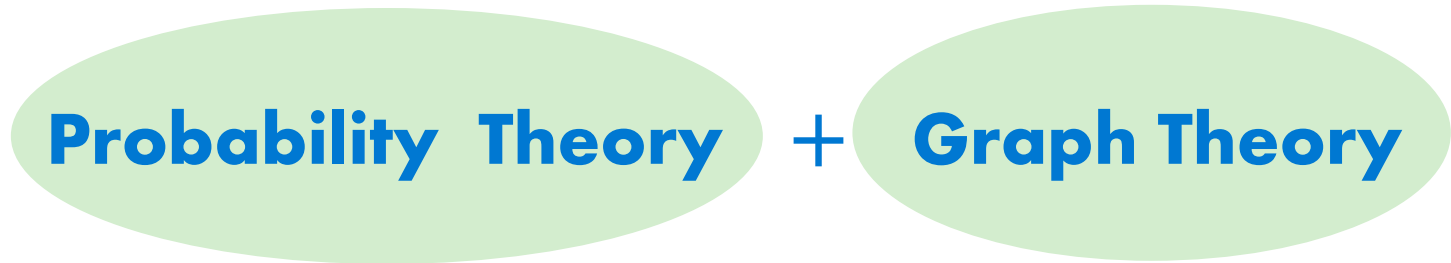
- Mr. Holmes is in his office
  - He receives a call from his neighbor that the alarm of his house went off.
  - He thinks that somebody broke into his house.
- Afterwards he hears an announcement from radio that a small earthquake just happened
  - Since the alarm has been going off during an earthquake.
  - He concludes it is more likely that earthquake causes the alarm.



# Graphical Models

## *An overview*

- Graphical models (probabilistic graphical models) are results from the marriage between graph theory and probability theory



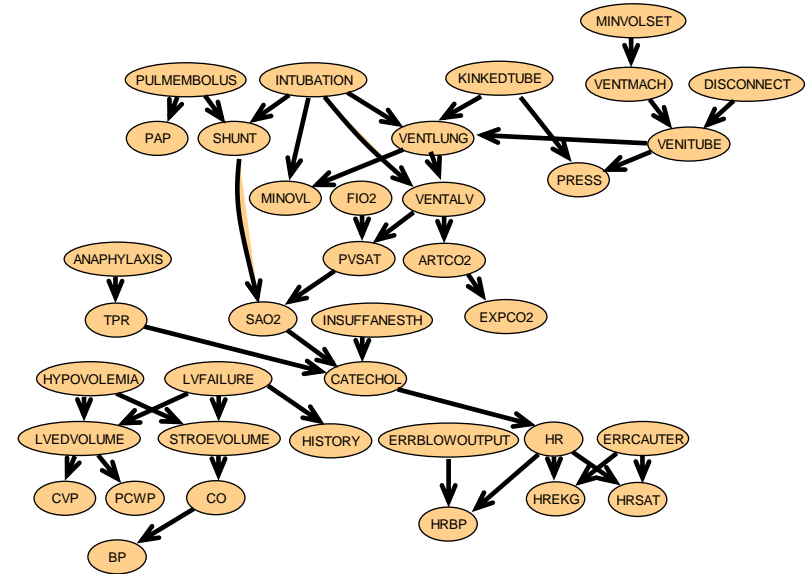
- Provides a powerful tool for modeling and solving problems related to

**Uncertainty** and **Complexity**

# Graphical Models

## *An overview*

- **Probability theory:** ensures consistency, provides interface models to data.
- **Graph theory:** intuitively appealing interface for humans.  
“The graphical language allows us to encode in practice: the property that variables tend to interact *directly* only with very few others”. (Koller’s book).
- **Modularity:** a complex system is built by combining simpler parts.

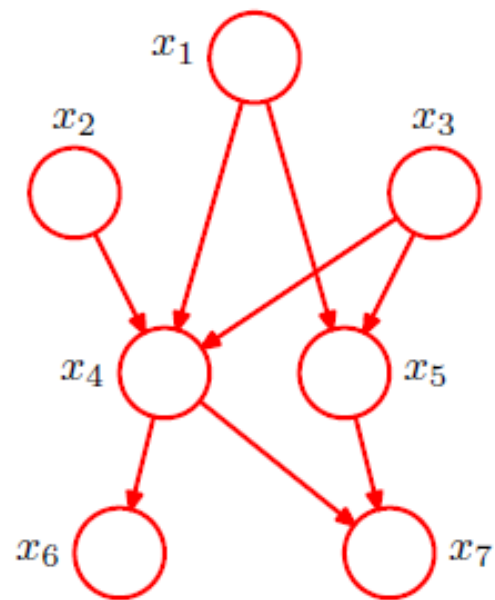


- **Issues:**
  - ❑ Representation
  - ❑ Learning
  - ❑ Inference
  - ❑ Applications

# Graphical Models

## *Useful properties*

- They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.
- Insights into the properties of the model can be obtained by inspection of the graph.
- Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly.



$$P(\mathbf{x}) = \prod_{i=1}^K P(x_i | pa_i)$$

$$P(x_1)P(x_2)P(x_3) \\ P(x_4 | x_1, x_2, x_3) P(x_5 | x_1, x_3) \\ P(x_6 | x_4) P(x_7 | x_4, x_5)$$

# Graphical models

## *Representation*

- **Graphical models** are composed by two parts:
  1. A set  $\mathbf{X} = \{X_1, \dots, X_p\}$  of **random variables** describing the quantities of interest (observed variables: training data; latent variables).
  2. A **graph**  $\mathcal{G} = (V, E)$  in which each **vertex** (node)  $v \in V$  is associated with one of the random variables, and **edges** (link)  $e \in E$  express the **dependence structure** of the data (the set of dependence relationships among subsets of the variables in  $\mathbf{X}$ ) with different semantics for
    - **undirected graphs** (Markov random field or Markov networks), and
    - **directed acyclic graphs** (Bayesian networks).
- The link between the dependence structure of the data and its graphical representation is expressed in terms of **conditional independence** (denoted with  $\perp_P$ ) and **graphical separation** (denoted with  $\perp_G$ ).



# Graphical models

## Representation

- A graph  $\mathcal{G}$  is a **dependency map** (or **D-map**, completeness) of the probabilistic dependence structure  $P$  of  $\mathbf{X}$  if there is a one-to-one correspondence between the random variables in  $\mathbf{X}$  and the nodes  $V$  of  $\mathcal{G}$ , such that for all disjoint subsets  $A, B, C$  of  $\mathbf{X}$  we have

$$A \perp_P B|C \Rightarrow A \perp_{\mathcal{G}} B|C$$

Similarly,  $\mathcal{G}$  is an **independency map** (or **I-map**, soundness) of  $P$  if

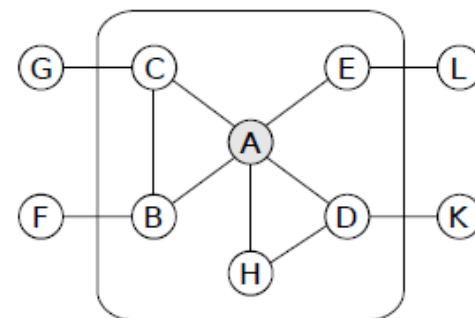
$$A \perp_{\mathcal{G}} B|C \Rightarrow A \perp_P B|C$$

$\mathcal{G}$  is a **perfect map** of  $P$  if it is both a D-map and an I-map, that is

$$A \perp_P B|C \Leftrightarrow A \perp_{\mathcal{G}} B|C$$

and in this case  $P$  is said to be **isomorphic** to  $\mathcal{G}$ .

- The key concept of **separation**
  - u-separation in undirected graphical models
  - d-separation in directed graphical models.



# Graphical models

## Factorization

- A fundamental result descending from the definitions of u-separation and d-separation is the **Markov property** (or **Markov condition**), which defines the decomposition of the global distribution of the data into a set of local distributions.

- For Bayesian networks

$$P(\mathbf{X}) = \prod_{i=1}^p P(X_i | \Pi_{X_i})$$

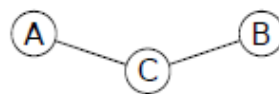
$\Pi_{X_i}$  is parents of  $X_i$ .

- For Markov networks

$$P(\mathbf{X}) = \prod_{i=1}^p \phi_i(C_i),$$

$\phi_i$  is factor potentials  
(representing the relative mass of probability of each clique  $C_i$ )

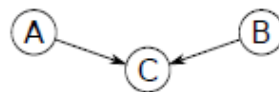
separation (undirected graphs)



$$A \perp\!\!\!\perp B \mid C$$

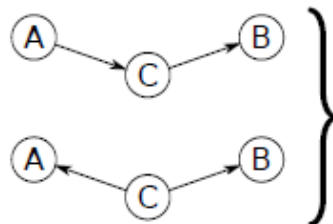
$$P(A, B, C) = P(A \mid C) P(B \mid C) P(C)$$

d-separation (directed acyclic graphs)



$$A \not\perp\!\!\!\perp B \mid C$$

$$P(A, B, C) = P(C \mid A, B) P(A) P(B)$$



$$A \perp\!\!\!\perp B \mid C$$

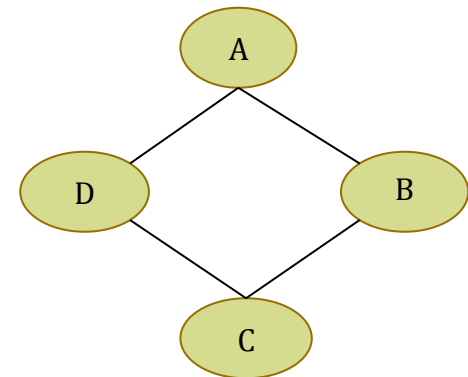
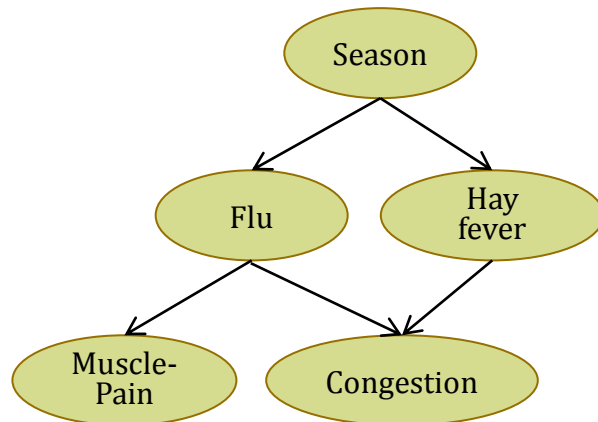
$$P(A, B, C) =$$

$$= P(B \mid C) P(C \mid A) P(A)$$

$$= P(A \mid C) P(B \mid C) P(C)$$

# Graphical models

## Examples



## Independence

$$\begin{aligned}(F \perp H \mid S) \\ (C \perp S \mid F, H) \\ (M \perp H, C \mid F) \\ (M \perp C \mid F)\end{aligned}$$

$$\begin{aligned}(A \perp C \mid B, D) \\ (B \perp D \mid A, C)\end{aligned}$$

## Factorization

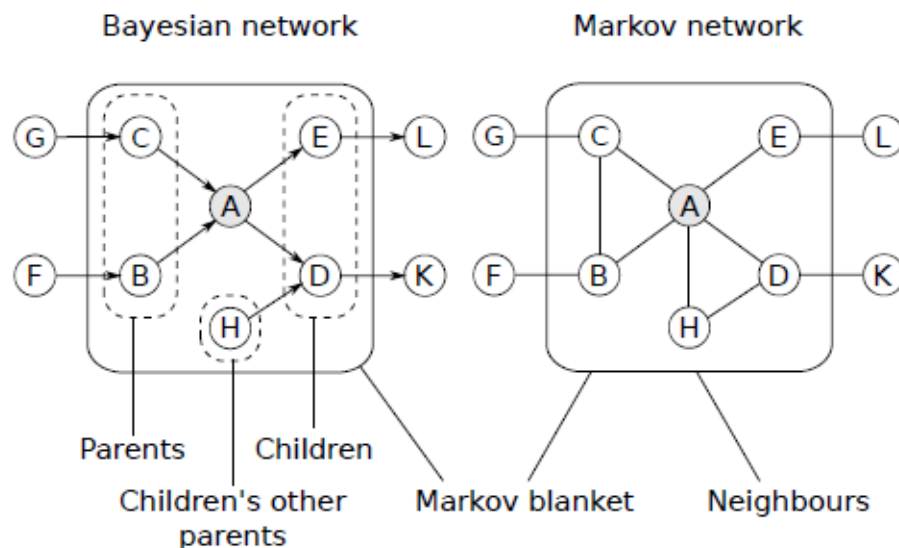
$$\begin{aligned}P(S, F, H, C, M) &= P(S)P(F|S) \\ &\quad P(H|S)P(C|F, H)P(M|F)\end{aligned}$$

$$\begin{aligned}P(A, B, C, D) &= \frac{1}{Z} \phi_1(A, B) \\ &\quad \phi_2(B, C)\phi_3(C, D)\phi_4(A, D)\end{aligned}$$

# Graphical models

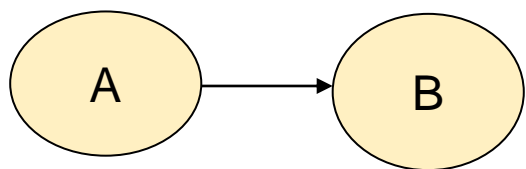
## Markov blanket

- Another fundamental result: **Markov blanket** (Pearl 1988) of a node  $X_i$ , the set that completely separates  $X_i$  from the rest of the graph.
- Markov blanket is the set of nodes that includes all the knowledge needed to do inference on  $X_i$  because all the other nodes are conditionally independent from  $X_i$  given its Markov blanket.
- In Markov networks the Markov blanket contains nodes that are connected to  $X_i$  by an edge. In Bayesian networks it is the union of the children of  $X_i$ , its parents, and its children's other parents.



# Graphical models

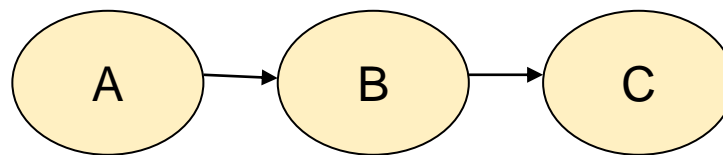
## Simple case and serial connection



$$P(\mathbf{X}) = \prod_{i=1}^n P(X_i | pa(X_i))$$

- Dependency is described by the **conditional probability**  $P(B|A)$
- Knowledge about A: priori probability  $P(A)$
- Calculate the joint probability of the A and B

$$P(A, B) = P(B|A)P(A)$$



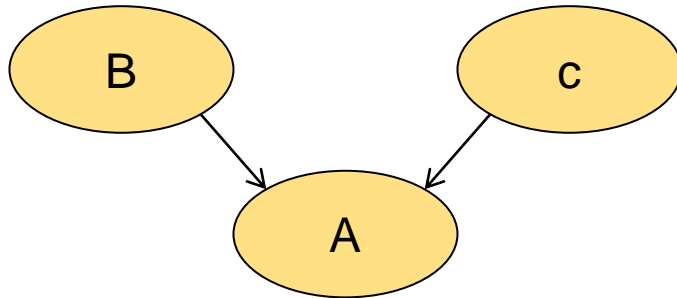
$$P(\mathbf{X}) = \prod_{i=1}^n P(X_i | pa(X_i))$$

- Calculate as before:
$$P(A, B) = P(B|A)P(A)$$
$$P(A, B, C) = P(C|A, B)P(A, B)$$
$$= P(C|B)P(B|A)P(A)$$
- $I(C, A|B)$

# Graphical models

## *Converging connection and diverging connection*

$$P(\mathbf{X}) = \prod_{i=1}^n P(X_i | pa(X_i))$$

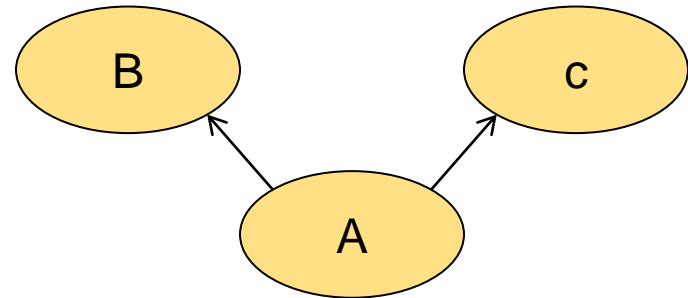


- Value of  $A$  depends on  $B$  and  $C$

$$P(A|B, C)$$

- $P(A, B, C) = P(A|B, C)P(B)P(C)$

$$P(\mathbf{X}) = \prod_{i=1}^n P(X_i | pa(X_i))$$



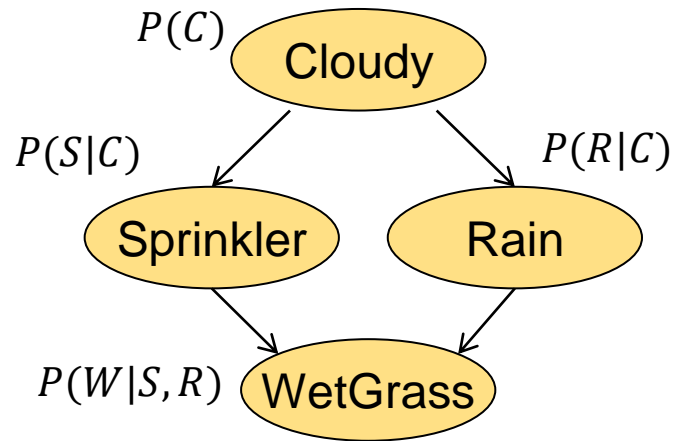
- $B$  and  $C$  depend on  $A$ :

$$P(B|A) \text{ and } P(C|A)$$

- $P(A, B, C) = P(B|A)P(C|A)P(A)$
- $I(B, C|A)$

# Graphical models

## Wet grass



$P(C = F)$	$P(C = T)$
0.5	0.5

C	$P(S = F)$	$P(S = T)$
F	0.5	0.5
T	0.9	0.1

C	$P(R = F)$	$P(R = T)$
F	0.8	0.2
T	0.2	0.8

$$P(C, S, R, W) = P(W|S, R)P(R|C)P(S|C)P(C)$$

Versus

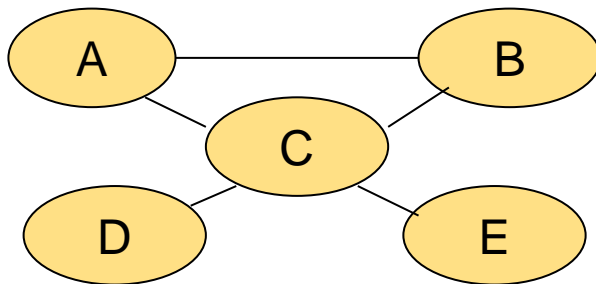
$$P(C, S, R, W) = P(W|\cancel{C}, S, R)P(R|\cancel{C}, \cancel{S})P(S|C)P(C)$$

S	R	$P(W = F)$	$P(W = T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

# Graphical models

## *Markov random fields*

- Links represent symmetrical probabilistic dependencies
- Direct link between  $A$  and  $B$ : conditional dependency.
- Weakness of MRF: inability to represent induced dependencies.



- **Global Markov property**:  $X$  is independent of  $Y$  given  $Z$  iff all paths between  $X$  and  $Y$  are blocked by  $Z$  (here:  $A$  is independent of  $E$ , given  $C$ )
- **Local Markov property**:  $X$  is independent of all other nodes given its neighbors (here:  $A$  is independent of  $D$  and  $E$ , given  $C$  and  $B$ ).

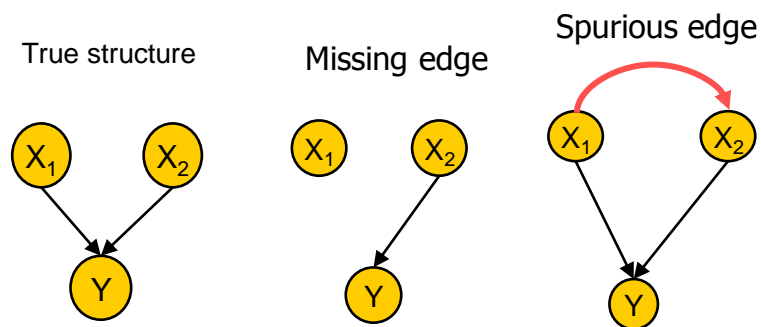
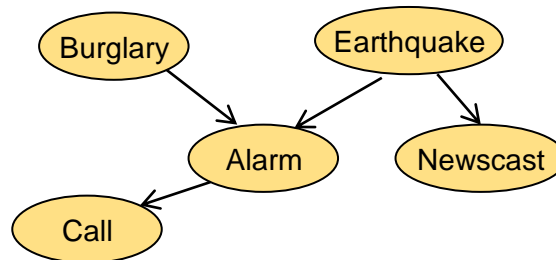


# Graphical models

## Learning

- Form the input of fully or partially observable data cases?
- The learning steps:
  - **Structure learning:** Qualitative dependencies between variables (edge between any two nodes?)
  - **Parameter learning:** Quantitative dependencies between variables are parameterized conditional distributions. Parameters of the functions are parameters of the graph.

B	E	A	C	N
$\bar{b}$	$e$	$a$	$c$	$\bar{n}$
$b$	$\bar{e}$	$\bar{a}$	$\bar{c}$	$n$



# Graphical models

## *Approaches to learning of graphical model structure*

### ■ Constraint-based approaches

- ❑ Identify a set of conditional independence properties
- ❑ Identify the network structure that best satisfies these constraints
- ❑ *Limitation:* sensitive to errors in single dependencies

### ■ Search-and-Score based approaches

- ❑ Define a scoring function specifying how well the model fits the data
- ❑ Search possible structures for one that has optimal scoring function.
- ❑ *Limitation:* intractable to evaluate → heuristic, greedy, sub-optimal

### ■ Regression-based approaches

- ❑ Gaining popularity in recent years
- ❑ Are essentially optimization problems which guarantees global optimum for the objective function, and have better scalability.

# Graphical models

## *Approaches to learning of graphical model parameters*

- Learning parameters from **complete data**: Using maximum likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{d}|\boldsymbol{\theta}) = \sum_{i=1}^n \log p(x^i|\boldsymbol{\theta})$$

- Learning parameters with **hidden variables**: EM algorithm

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_y \log p(x, y|\boldsymbol{\theta}) \geq \mathcal{F}(q, \boldsymbol{\theta})$$

$$E \text{ step: } q_{[k+1]} \leftarrow \arg \max_q \mathcal{F}(q, \theta_{[k]}), \quad M \text{ step: } \theta_{[k+1]} \leftarrow \max_{\theta} \mathcal{F}(q_{[k+1]}, \boldsymbol{\theta})$$

- Parameter learning in **undirected graphical models**
- **Bayesian learning** of parameter

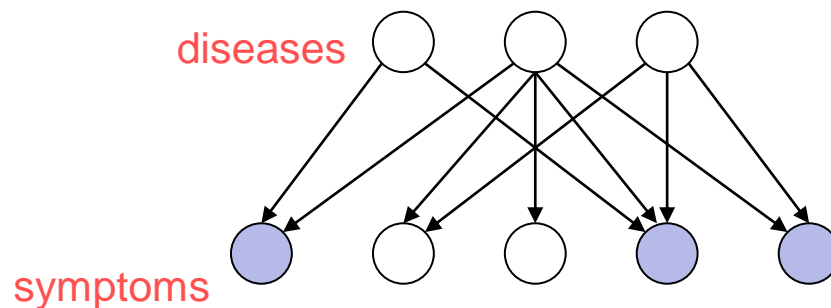
$$P(\boldsymbol{\theta}|\mathbf{m}, \mathbf{d}) = \frac{P(\mathbf{d}|\boldsymbol{\theta}, \mathbf{m})P(\boldsymbol{\theta}|\mathbf{m})}{P(\mathbf{d}|\mathbf{m})}$$

# Graphical models

## *Inference*

- Computational inference problems
  1. Computing the likelihood of observed data.
  2. Computing the marginal distribution  $P(x_A)$  over a particular subset  $A \subset V$  of nodes.
  3. Computing the posterior distribution of latent variables.
  4. Computing a mode of the density (i.e., an element  $\hat{x}$  in the set  $\arg \max_{x \in \mathcal{X}^m} P(x)$ )

- Example:  
What is the most probable disease?



# Graphical models

## *Inference*

### **Exact inference**

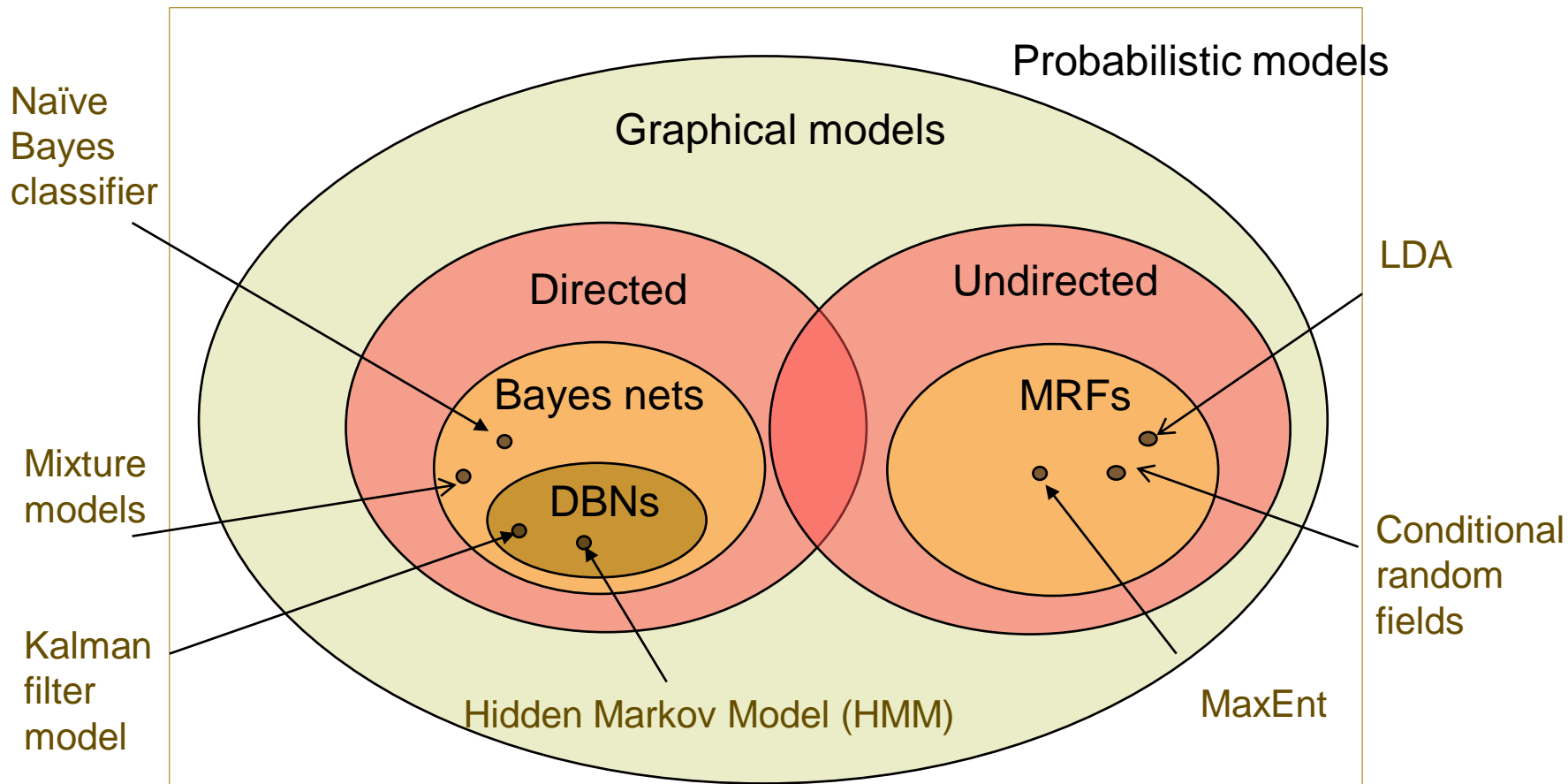
- Can exploit structure (conditional independence) to efficiently perform exact inference in many practical situations
- **Variable elimination** (remove irrelevant variables for the query)
- **Junction trees** and message passing, sum-product and max-product algorithms.
- Lack of tractability.

### **Approximate inference**

- **Sampling inference** (stochastic methods): Markov Chain Monte Carlo, yield their results in the form of a set of samples drawn from the posterior distribution.
- **Variational inference** (deterministic methods) seek the optimal member of a defined family of approximating distributions by minimizing a suitable criterion which measures the dissimilarity between the approximate distribution and the exact posterior distribution.

# Graphical models

## *Instances of graphical models*

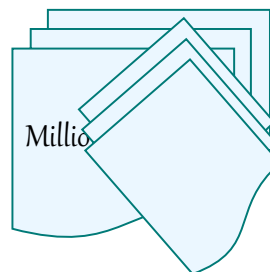


# Content

- Brief overview of graphical models
- Introduction to topic models
- Fully sparse topic model
- Conditional random fields in NLP

# Introduction to topic modeling

- The main way of automatic capturing the meaning of documents.
- *Topic*: the subject that we talk/write about
- *Topic of an image*: a cat, a dog, airplane, ...
- Topic in TM:
  - *a set of words which are semantically related* [Landauer and Dumais, 1997];
  - *a distribution over words* [Blei, Ng, and Jordan, 2003].



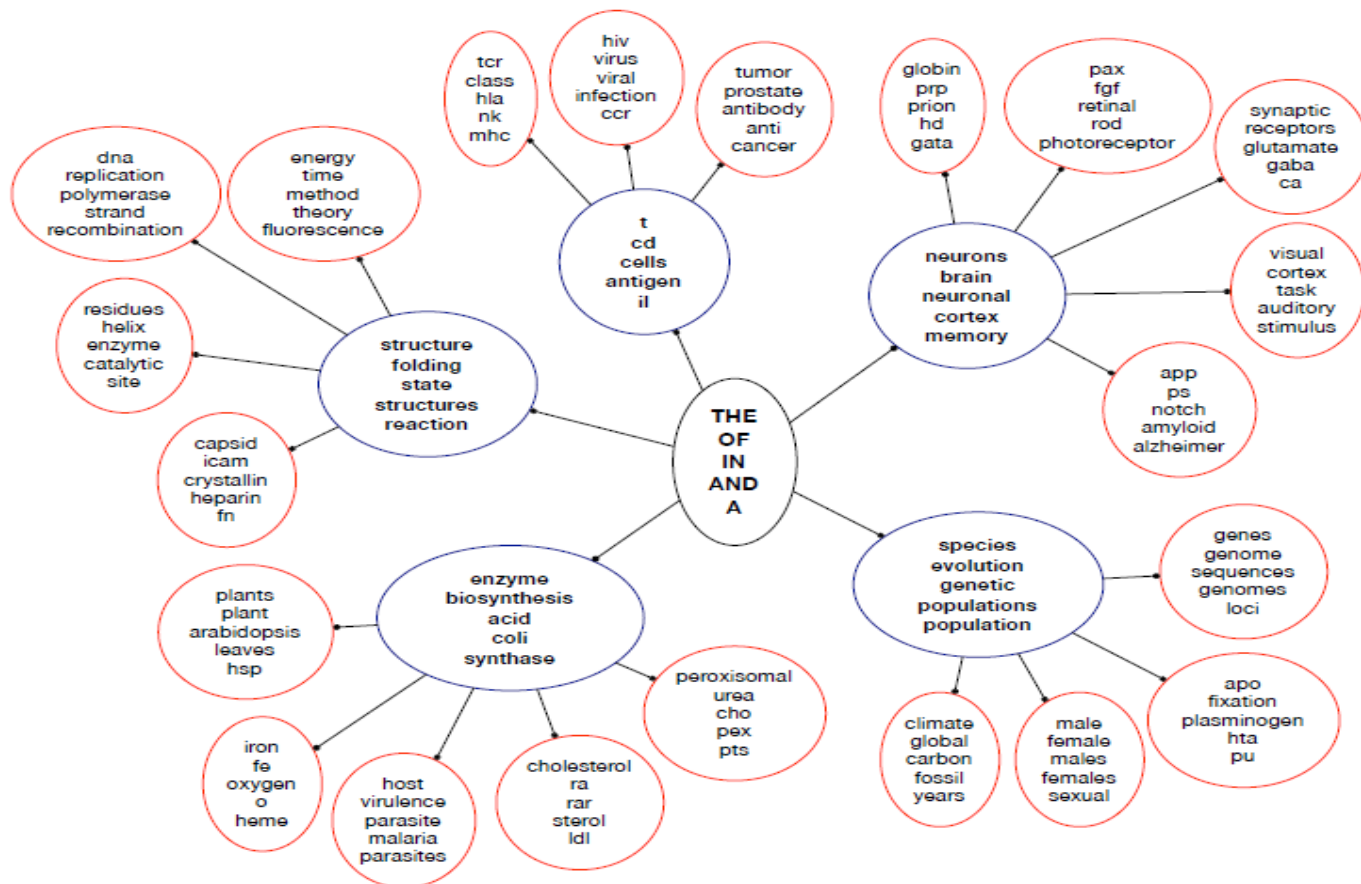
Film, movie, show, play, actor,  
cinema

Million, tax, program, budget,  
spending, money



# Introduction to topic modeling

Topic model: *a model about topics hidden in data.*



# Introduction to topic modeling

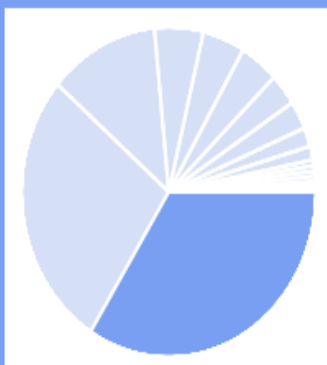
Two key problems in TM.

## ■ Learning

estimate a topic model from a given collection of documents.

## ■ Inference

we are asked to find the topics of a new document.

	<p><b>Elvis Aaron Presley</b><sup>a</sup> (January 8, 1935 – August 16, 1977) was one of the most popular American singers of the 20th century. A cultural icon, he is widely known by the single name <b>Elvis</b>. He is often referred to as the "King of Rock and Roll" or simply "the King".</p> <p>Born in Tupelo, Mississippi, Presley moved to Memphis, Tennessee, with his family at the age of 13. He began his career there in 1954 when Sun Records owner Sam Phillips, eager to bring the sound of African American music to a wider audience, saw in Presley the means to realize his ambition. Accompanied by guitarist Scotty Moore and bassist Bill Black, Presley was one of the originators of rockabilly, an uptempo, backbeat-driven fusion of country and rhythm and blues. RCA Victor acquired his contract in a deal arranged by Colonel Tom Parker, who</p>	<b>related documents</b>
<b>related topics</b>		Cher Wu-Tang Clan Musical theatre Alice Cooper Music video Whitney Houston Christina Aguilera Patsy Cline Tom Waits Kylie Minogue MTV Phil Collins Kate Bush Bohemian Rhapsody

# Topic models

## *Notation and terminology*

- A **word** is the basic unit of discrete data, from vocabulary indexed by  $V = \{1, \dots, V\}$ . The  $v$ th word is represented by a  $V$ -vector  $w$  such that  $w^v = 1$  and  $w^u = 0$  for  $u \neq v$ .

- A **document** is a sequence of  $N$  words denote by

$$d = (w_1, w_2, \dots, w_N)$$

- A **corpus** is a collection of  $M$  documents denoted by

$$D = \{d_1, d_2, \dots, d_M\}$$

# Topic models

## *Exchangeability and bag of word assumption*

- Random variables  $\{x_1, \dots, x_N\}$  are **exchangeable** if the joint distribution is invariant to permutation. If  $\pi$  is a permutation of the integers from 1 to N

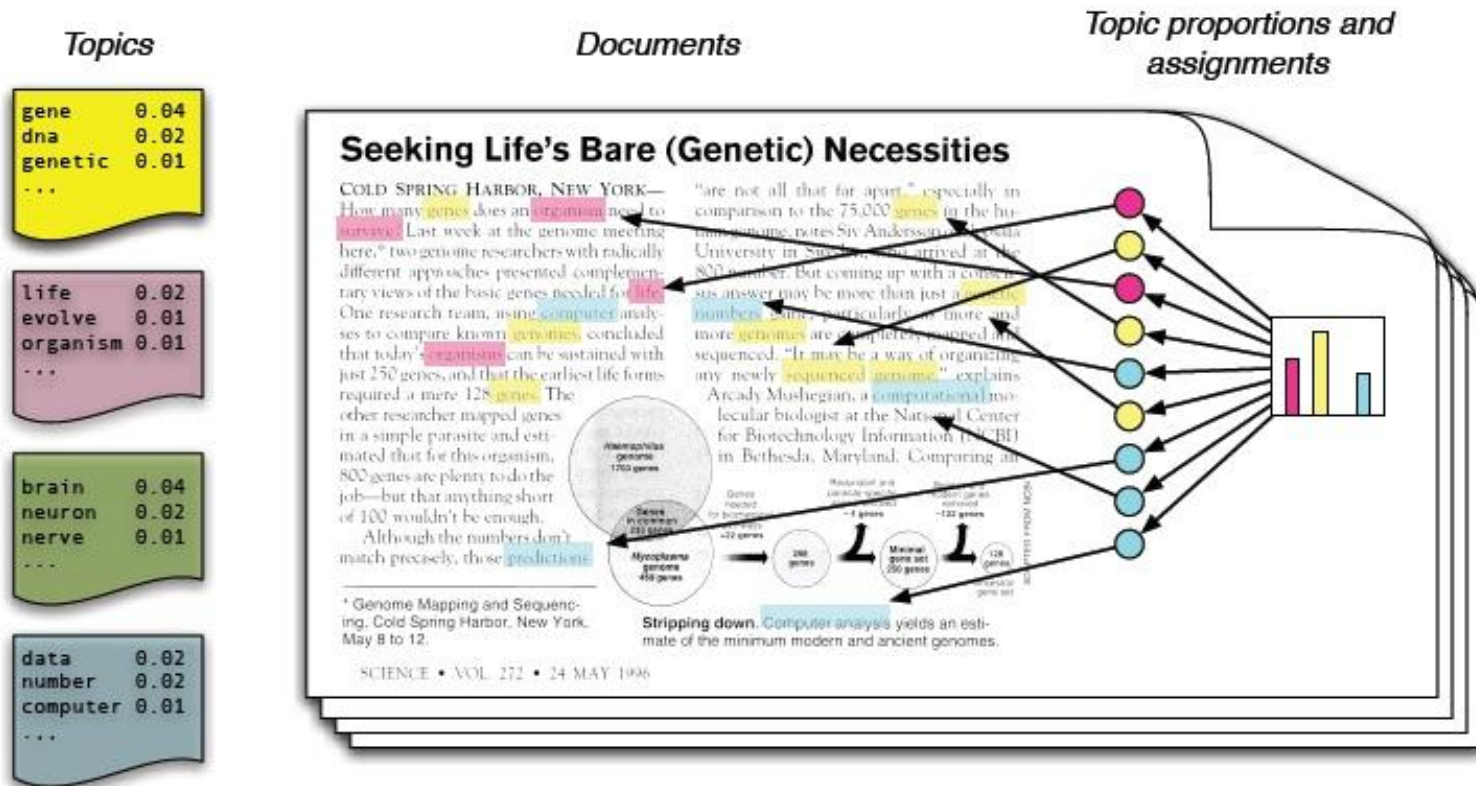
$$P(x_1, \dots, x_N) = P(x_{\pi(1)}, \dots, x_{\pi(N)})$$

- An infinite sequence of random is **infinitely exchangeable** if every finite subsequence is exchangeable.
- Word order is ignored  $\rightarrow$  “bag-of-words” – exchangeability, not i.i.d
- *Theorem (De Finetti, 1935)*: if  $\{x_1, \dots, x_N\}$  are infinitely exchangeable, then the joint probability has a representation as a mixture  $P(x_1, \dots, x_N)$  for some random variable  $\theta$

$$P(x_1, \dots, x_N) = \int d\theta p(\theta) \prod_{i=1}^N P(x_i|\theta)$$

# Topic models

## The intuitions behind topic models



Documents are mixtures of latent topics, where a topic is a probability distribution over words.

# Topic models

## *Probabilistic modeling*

- Topic models are part of the larger field of *probabilistic graphical modeling*.
- In generative probabilistic modeling, we treat our data as arising from a generative process that includes *hidden variables*. This generative process defines a *joint probability distribution* over both the observed and hidden random variables.
- We perform data analysis by using that joint distribution to compute the *conditional distribution* of the hidden variables given the observed variables. This conditional distribution is also called the *posterior distribution*.

# Topic models

## *Multinomial models for documents*

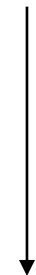
- Example: 50,000 possible words in our vocabulary
- Simple memoryless model
  - 50,000-sided die
    - a non-uniform die: each side/word has its own probability
    - to generate N words we toss the die N times
  - This is a simple probability model:
    - $P(\text{document} | \phi) = \prod_i P(\text{word } i | \phi)$
    - to “learn” the model we just count frequencies
    - $P(\text{word } i) = \text{number of occurrences of } i / \text{total number}$
  - Typically interested in conditional multinomials, e.g.,
    - $p(\text{words} | \text{spam})$  versus  $p(\text{words} | \text{non-spam})$

# Topic models

## *Real examples of word multinomials*

TOPIC 209	
WORD	PROB.
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
...	...

TOPIC 289	
WORD	PROB.
RETRIEVAL	0.1179
TEXT	0.0853
DOCUMENTS	0.0527
INFORMATION	0.0504
DOCUMENT	0.0441
CONTENT	0.0242
INDEXING	0.0205
RELEVANCE	0.0159
COLLECTION	0.0146
RELEVANT	0.0136
...	...



$P(w|z)$

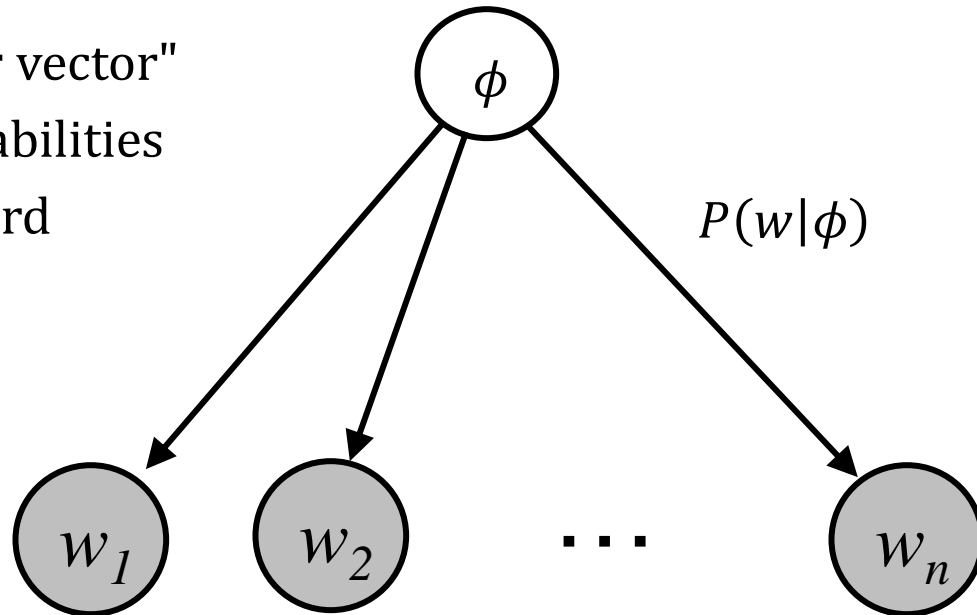


# Topic models

*A graphical model for multinomials*

$$P(doc | \phi) = \prod P(w_i | \phi)$$

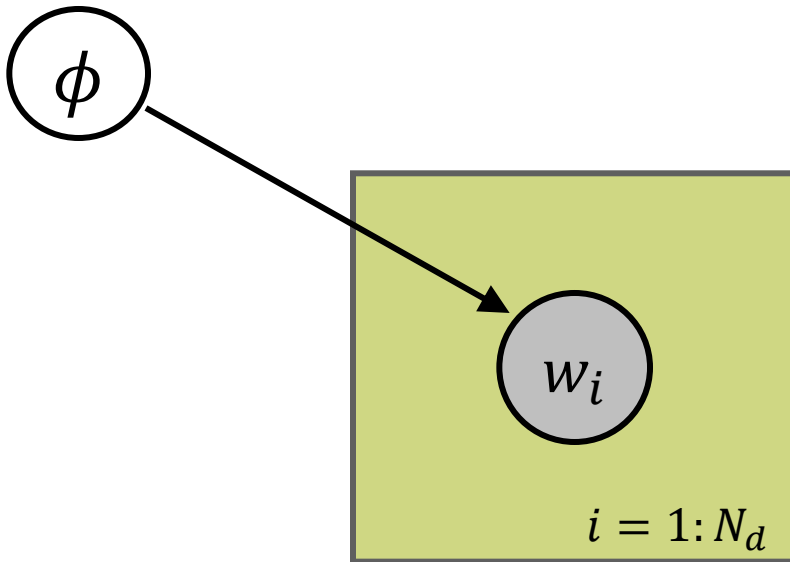
$\phi$  = "parameter vector"  
= set of probabilities  
one per word



# Topic models

## *Another view*

$$P(doc | \phi) = \prod P(w_i | \phi)$$



This is “plate notation”

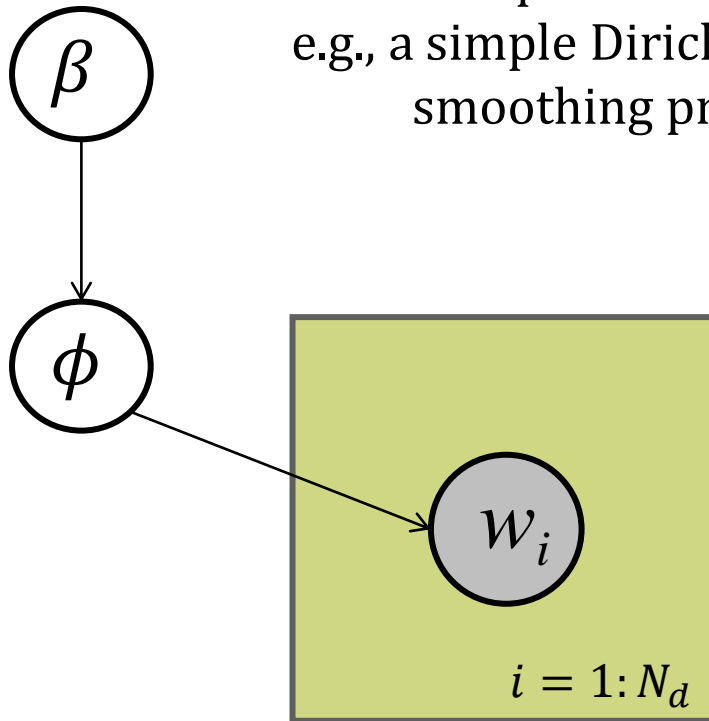
Items inside the plate are conditionally independent given the variable outside the plate.

There are “ $N_d$ ” conditionally independent replicates represented by the plate

# Topic models

*Being Bayesian...*

This is a prior on our multinomial parameters, e.g., a simple Dirichlet smoothing prior.

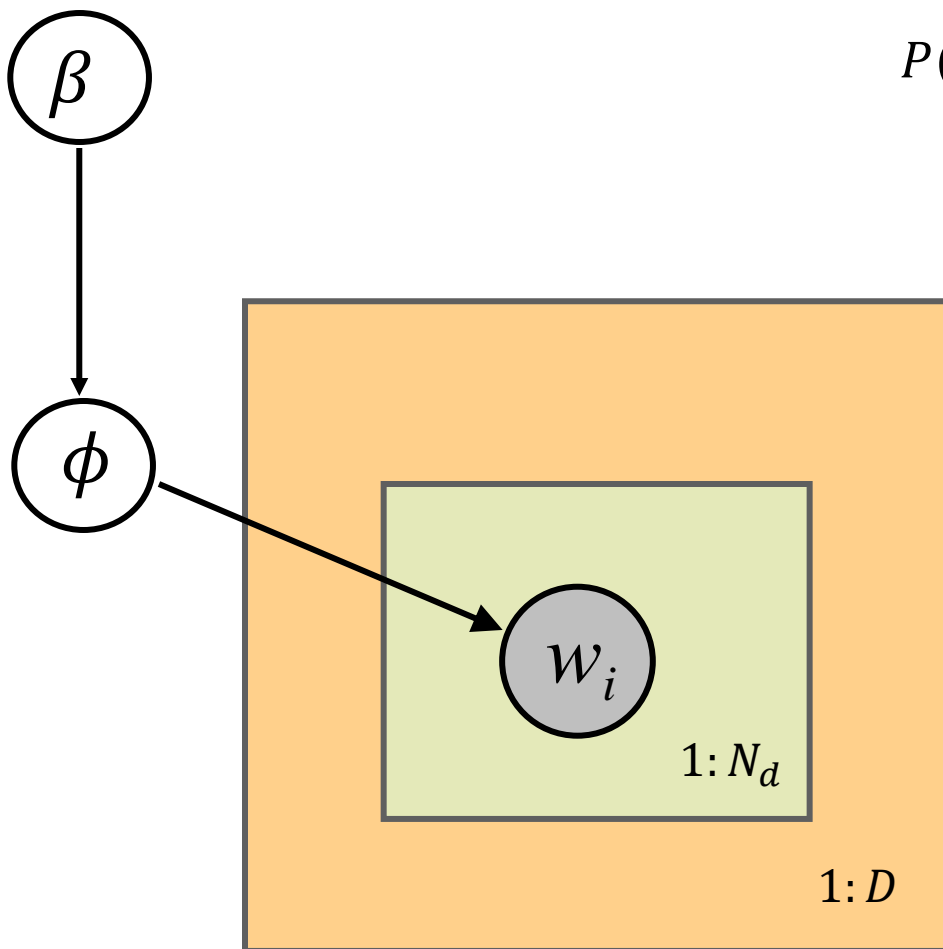


Learning: infer  $P(\phi | words, \beta)$   
proportional to  
 $P(words | \phi) P(\phi | \beta)$

# Topic models

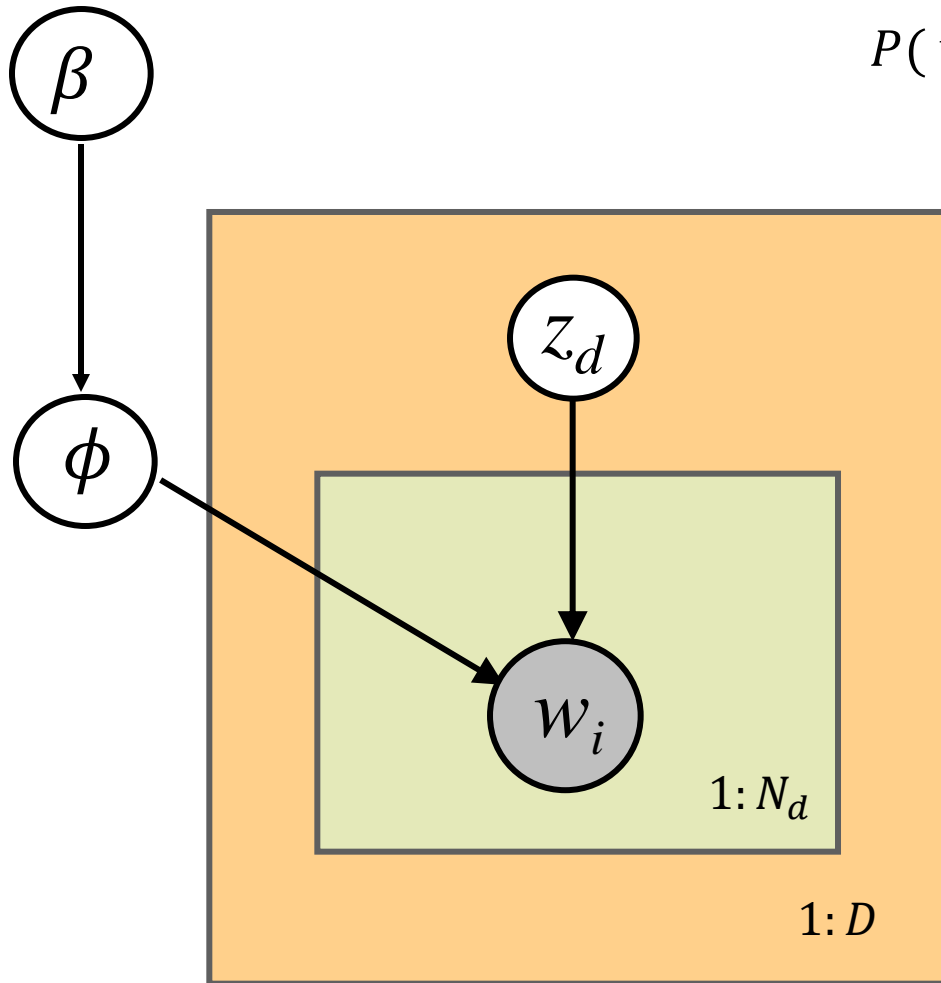
*Multiple Documents*

$$P(\text{corpus} | \phi) = \prod P(\text{doc} | \phi)$$



# Topic models

## *Different document types*



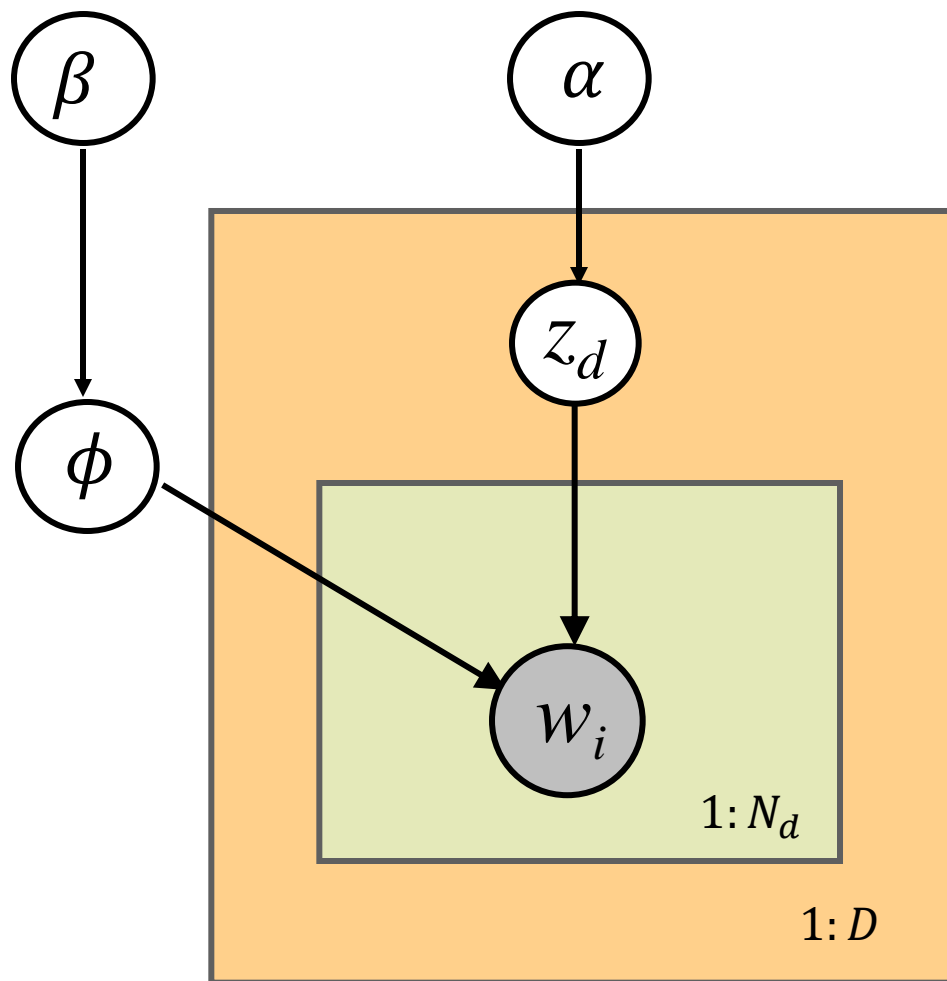
$P(w | \phi)$  is a multinomial over words

$P(w | \phi, z_d)$

- is a multinomial over words
- $z_d$  is the “label” for each doc
- Different multinomials, depending on the value of  $z_d$  (discrete)

# Topic models

## *Unknown document types*



Now the values of  $z$  for each document are unknown - hopeless?

Not hopeless :-)

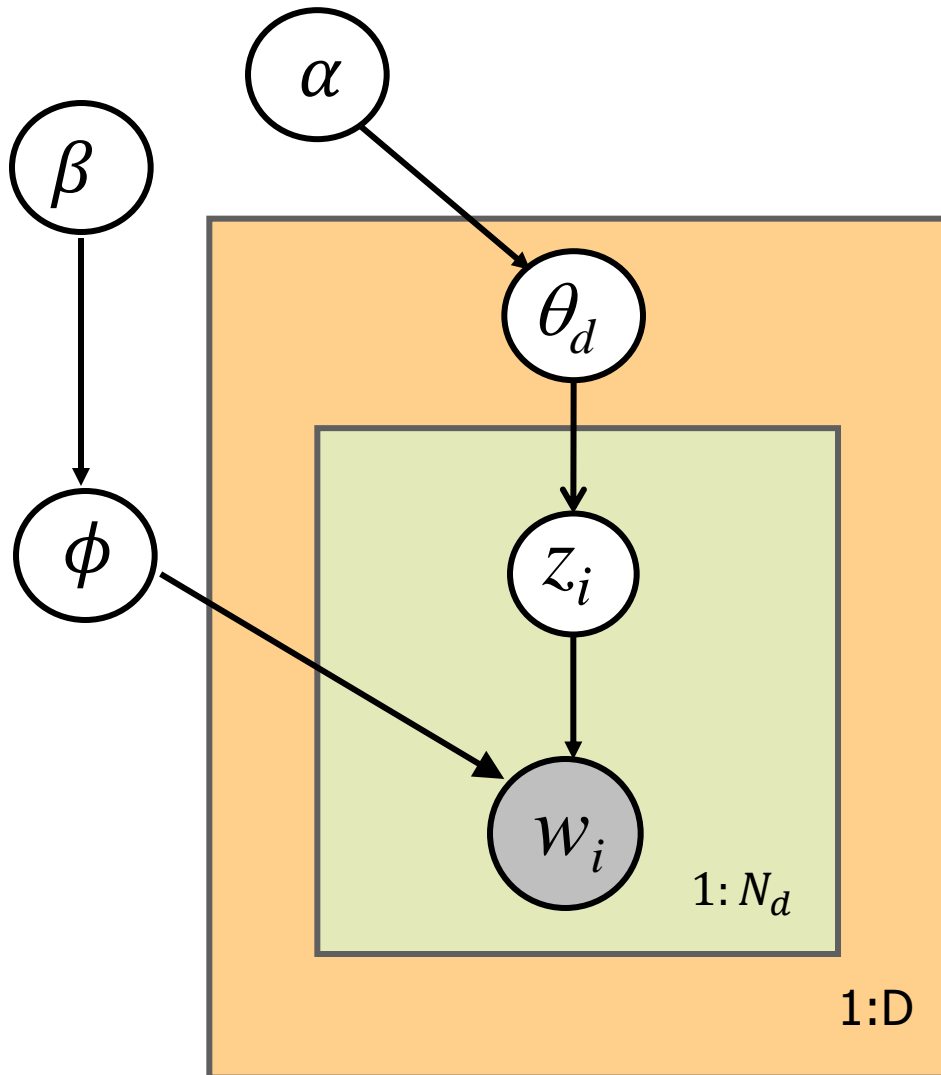
Can learn about both  $z$  and  $\theta$ , e.g., EM algorithm

This gives

$$P(w \mid z = k, \theta)$$

is the  $k$ th multinomial over words

# Topic Models



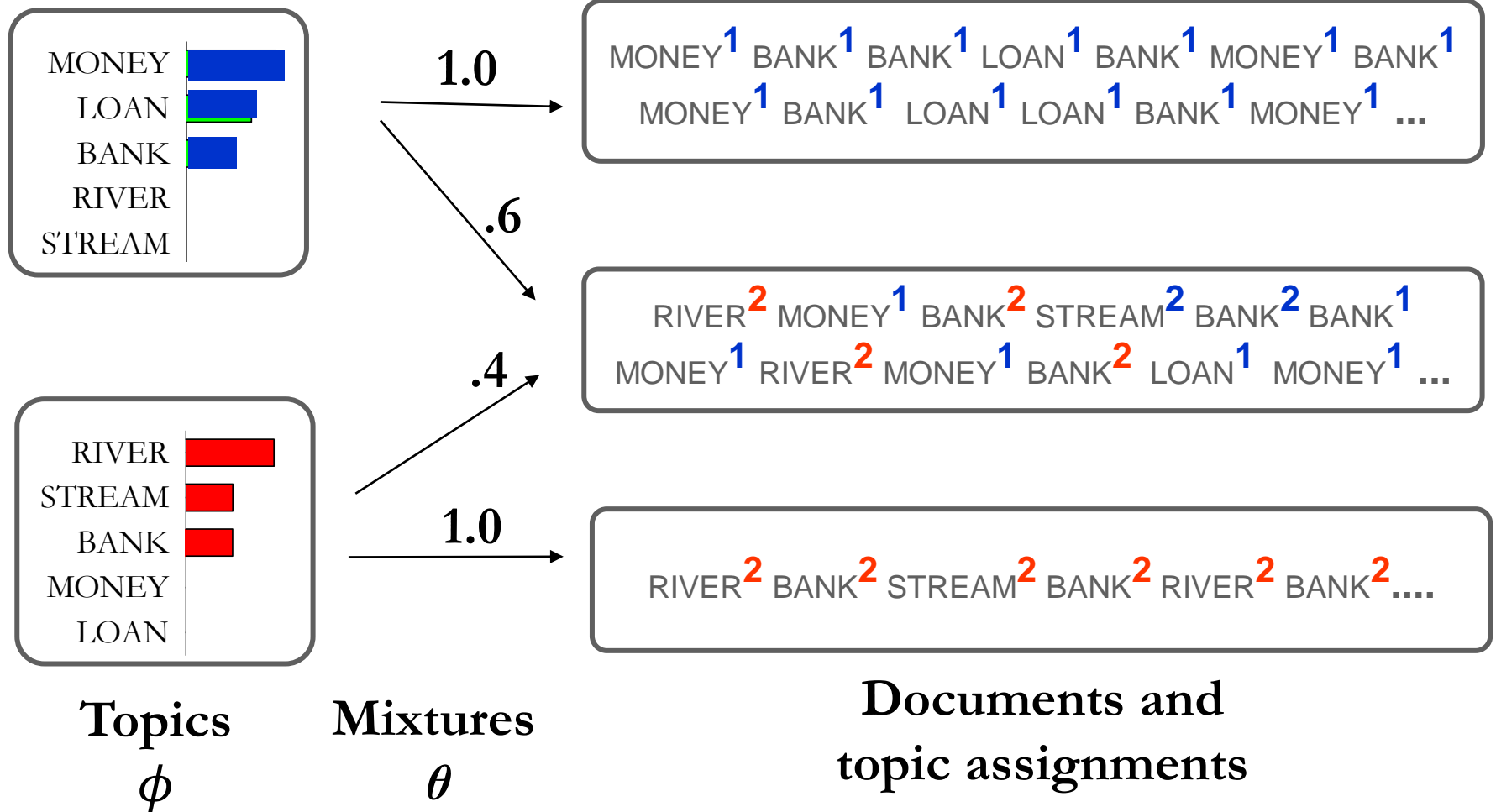
$z_i$  is a "label" for each *word*

$\theta$ :  $P(z_i | \theta_d) =$   
distribution over topics  
of a document specific

$\phi$ :  $P(w | \phi, z_i = k)$   
= multinomial over words  
= a "topic"

# Topic models

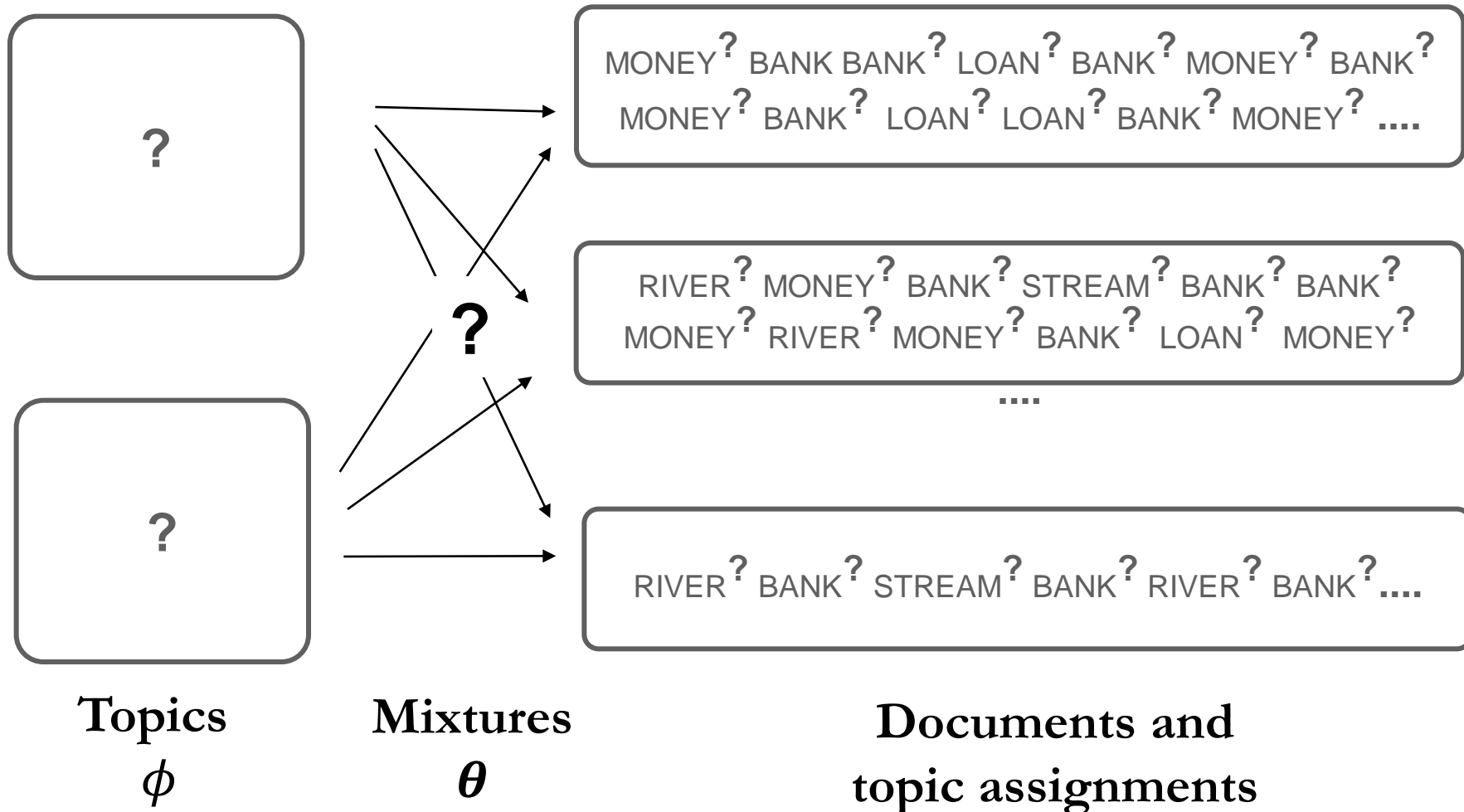
## Example of generating words





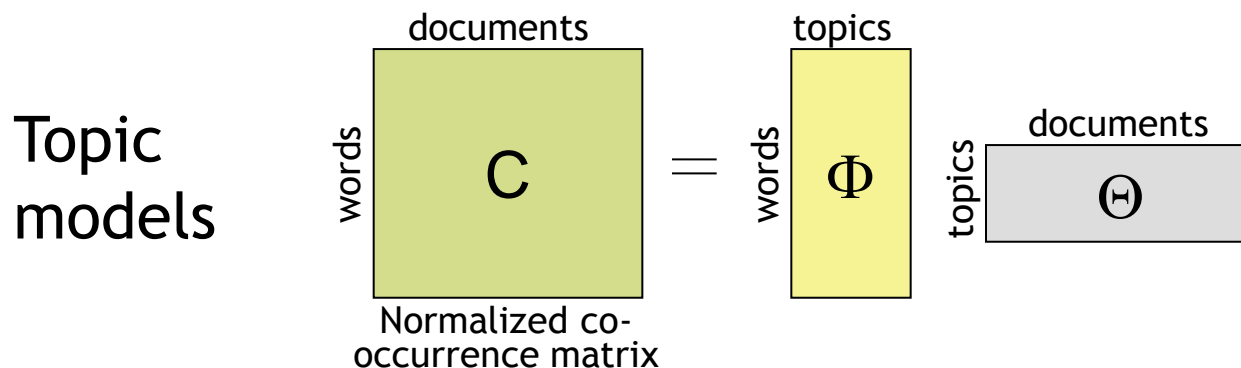
# Topic models

## Learning



# Topic models

## *The key ideas*

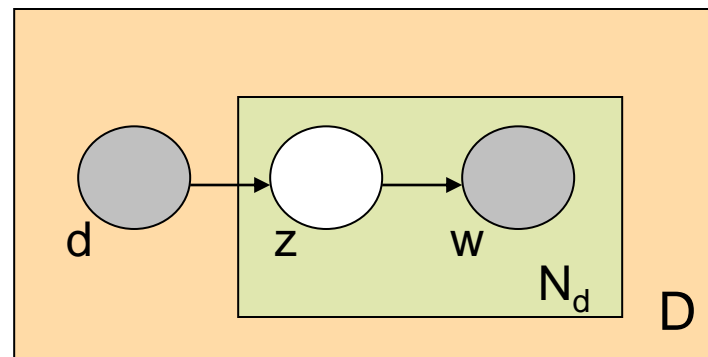


- **Key idea:** documents are mixtures of latent topics, where a topic is a probability distribution over words.
- Hidden variables, generative processes, and statistical inference are the foundation of probabilistic modeling of topics.

# Topic models

## *Probabilistic latent semantic indexing (Hofmann, 1999)*

- pLSI: Each word is generated from a single topic, different words in the document may be generated from different topics.
- Each document is represented as a list of mixing proportions for the mixture topics.
- Generative process:
  - Choose a document  $d_m$  with  $P(d)$
  - For each word  $w_n$  in the  $d_m$ 
    - Choose a  $z_n$  from a multinomial conditioned on  $d_m$ , i.e., from  $P(z|d_m)$
    - Choose a  $w_n$  from a multinomial conditioned on  $z_n$ , i.e., from  $P(w|z_n)$ .



$$P(d, w_n) = P(d) \sum_z P(w_n | z) P(z | d)$$

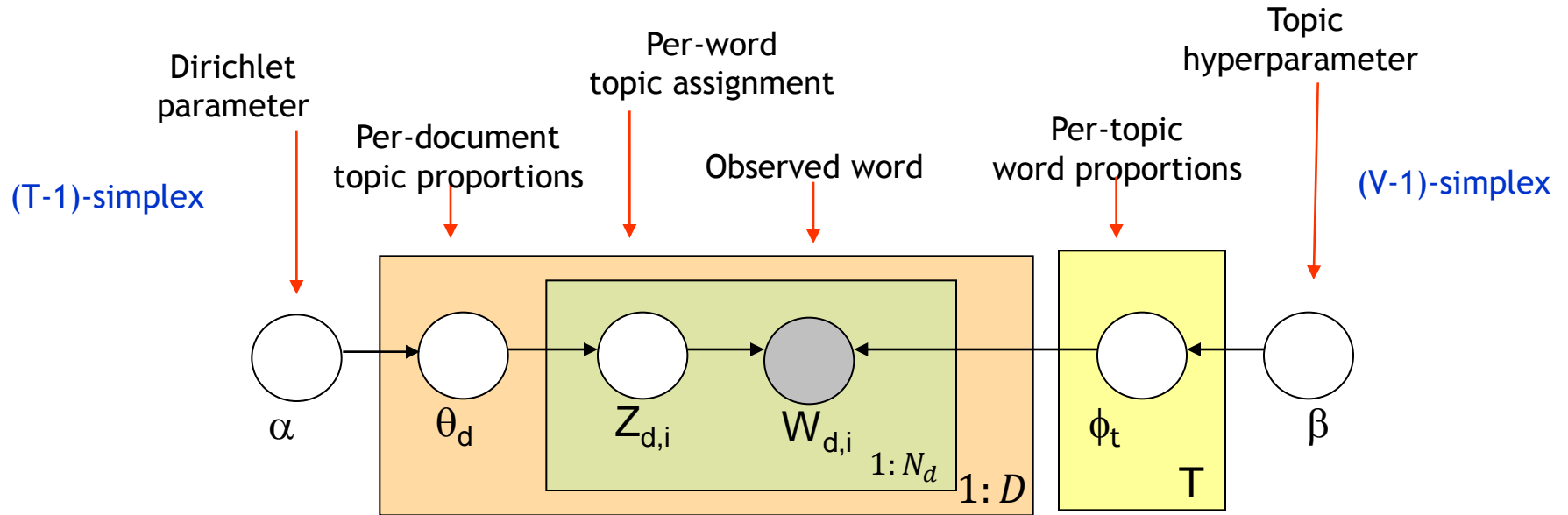
# Topic models

## *pLSI limitations*

- The model allows multiple topics in each document, but
  - the possible topic proportions have to be learned from the document collection
  - pLSI does not make any assumptions about how the mixture weights  $\theta$  are generated, making it difficult to test the generalizability of the model to new documents.
- Topic distribution must be learned for each document in the collection  
→ # parameters grows with the number of documents (billion documents?).
- Blei, Ng, and Jordan (2003) extended this model by introducing a *Dirichlet prior on  $\theta$* , calling **Latent Dirichlet Allocation** (LDA).

# Topic models

## *Latent Dirichlet allocation*



1. Draw each topic  $\phi_t \sim \text{Dir}(\beta)$ ,  $t=1, \dots, T$
2. For each document:
  1. Draw topic proportions  $\theta_d \sim \text{Dir}(\alpha)$
  2. For each word:
    1. Draw  $z_{d,i} \sim \text{Mult}(\theta_d)$
    2. Draw  $w_{d,i} \sim \text{Mult}(\phi_{z_{d,i}})$

1. From collection of documents, infer
  - per-word topic assignment  $z_{d,i}$
  - per-document topic proportions  $\theta_d$
  - per-topic word distribution  $\phi_t$
2. Use posterior expectations to perform the tasks: IR, similarity, ...

Choose  $N_d$  from a Poisson distribution with parameter  $\xi$

# Topic models

## LDA model

Dirichlet prior on the document-topic distributions

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

Joint distribution of topic mixture  $\theta$ , a set of  $N$  topic  $\mathbf{z}$ , a set of  $N$  words  $\mathbf{w}$

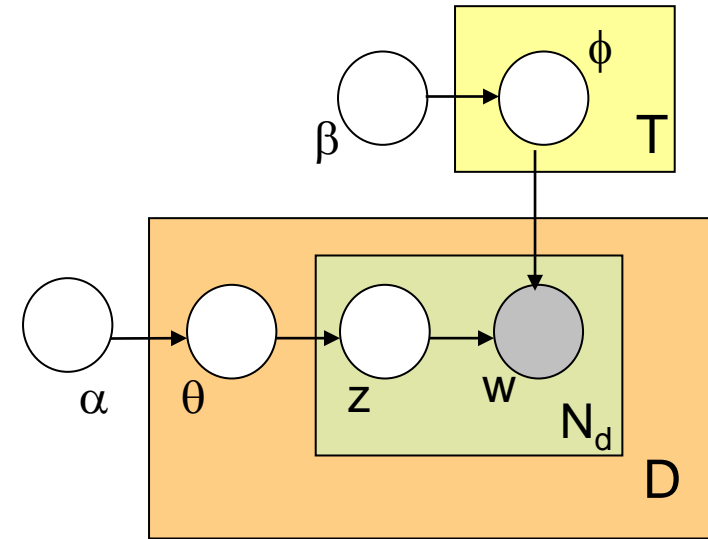
$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^N p(z_n|\theta) p(w_n|z_n, \beta)$$

Marginal distribution of a document by integrating over  $\theta$  and summing over  $\mathbf{z}$

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left( \prod_{n=1}^N \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d^k \theta$$

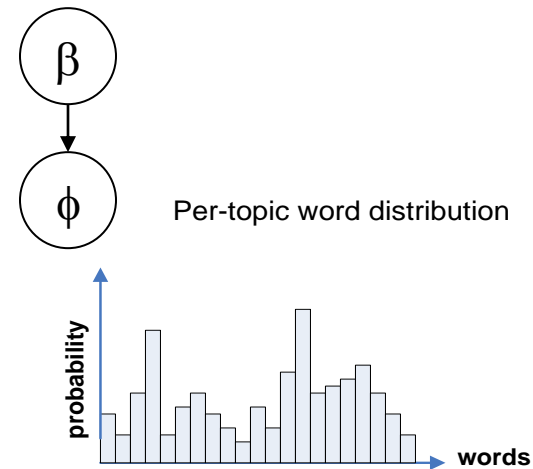
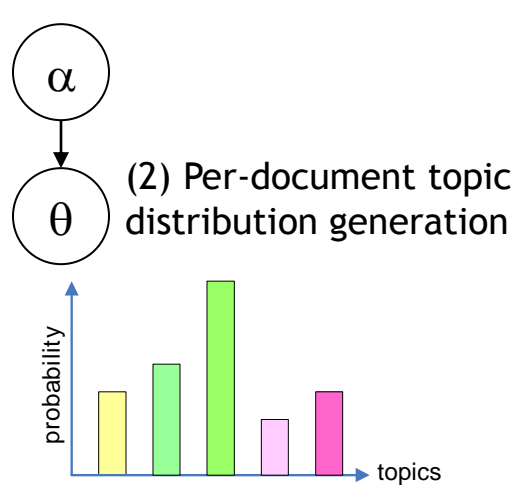
Probability of collection by product of marginal probabilities of single documents

$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d^k \theta_d$$

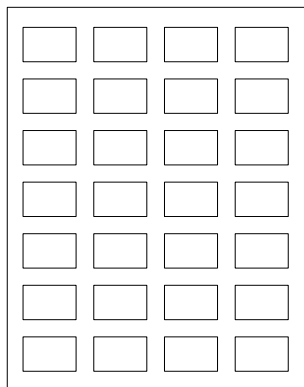


# Topic model

## Generative process

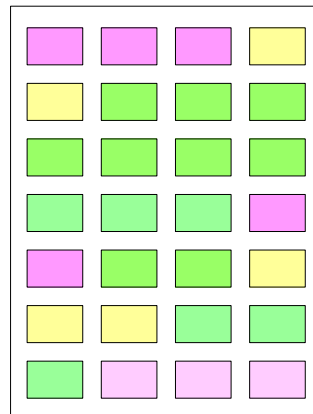


(1) Empty document



 word placeholder

(3) Topic sampling for word placeholders



(4) Real word generation

For decades, German software giant SAP ([SAP](#)) has been steadfast in its commitment to organic growth. During the last three years, SAP has spent a relatively modest \$1 billion or so on acquisitions. During the same period, rival Oracle ([ORCL](#)) has announced \$25 billion worth of deals, according to research analysts at Citigroup ([C](#)). But all of that changed on Oct. 7 when SAP said it would make [its largest acquisition ever](#) (BusinessWeek.com, 10/8/07) and pay \$6.8 billion for Business Objects ([BOB](#)), a business intelligence and data mining company based in France.

It's a sign of how mergers and acquisitions will reshape the software sector in the months and years ahead. Growth in software is slowing, and private equity firms are struggling to raise financing for big acquisitions in the rocky credit markets. That opens the door to strategic buyers—from SAP and Oracle to IBM ([IBM](#)) and Hewlett-Packard ([HPQ](#))—to seek out more deals. "The credit crunch has made business more difficult for private equity firms, and software companies now feel they have a free hand to do deals," says Bill Whyman, an analyst with researcher ISI [Group](#).

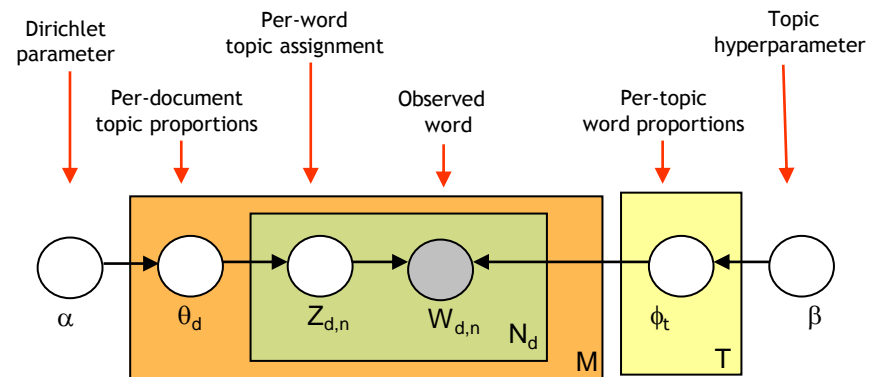
# Topic models

## *Inference in LDA*

- The posterior is

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

- The numerator: joint distribution of all the random variables, which can be computed for any setting of the hidden variables.
- The denominator: the marginal probability of the observations.
- In theory, it can be computed. However, is exponentially large and is intractable to compute.
- A central research goal of modern probabilistic graphical modeling is to develop efficient methods for approximating it.





# Topic models

## *Two categories of inference algorithms*

### **Sampling based algorithms**

- Attempt to collect samples from the posterior to approximate it with an empirical distribution.
- The most commonly used sampling algorithm for topic modeling is Gibbs sampling, where we construct a Markov chain— a sequence of random variables, each dependent on the previous— whose limiting distribution is the posterior.

### **Variational methods**

- Posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior.
- The inference problem is converted to an optimization problem.
- Variational methods open the door for innovations in optimization to have practical impact in probabilistic modeling

# Topic models

## Example

- From 16000 documents of AP corpus → 100-topic LDA model.
- An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated

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

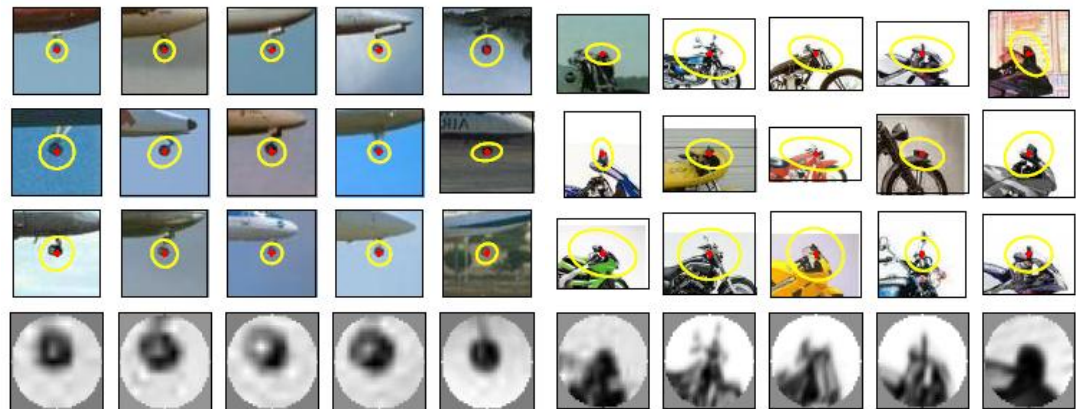
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as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

# Topic models

## Visual words

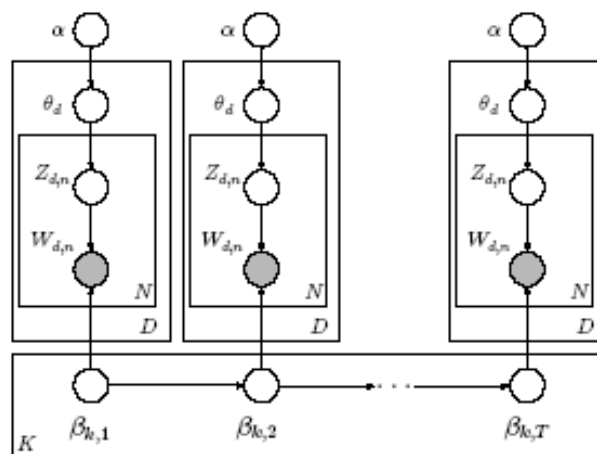
- Idea: Given a collection of images,
  - Think of each image as a document.
  - Think of feature patches of each image as words.
  - Apply the LDA model to extract topics.
- J. Sivic et al., Discovering object categories in image collections. *MIT AI Lab Memo AIM-2005-005*, Feb. 2005

Examples of 'visual words'



# Topic models

## *Applications in scientific trends*

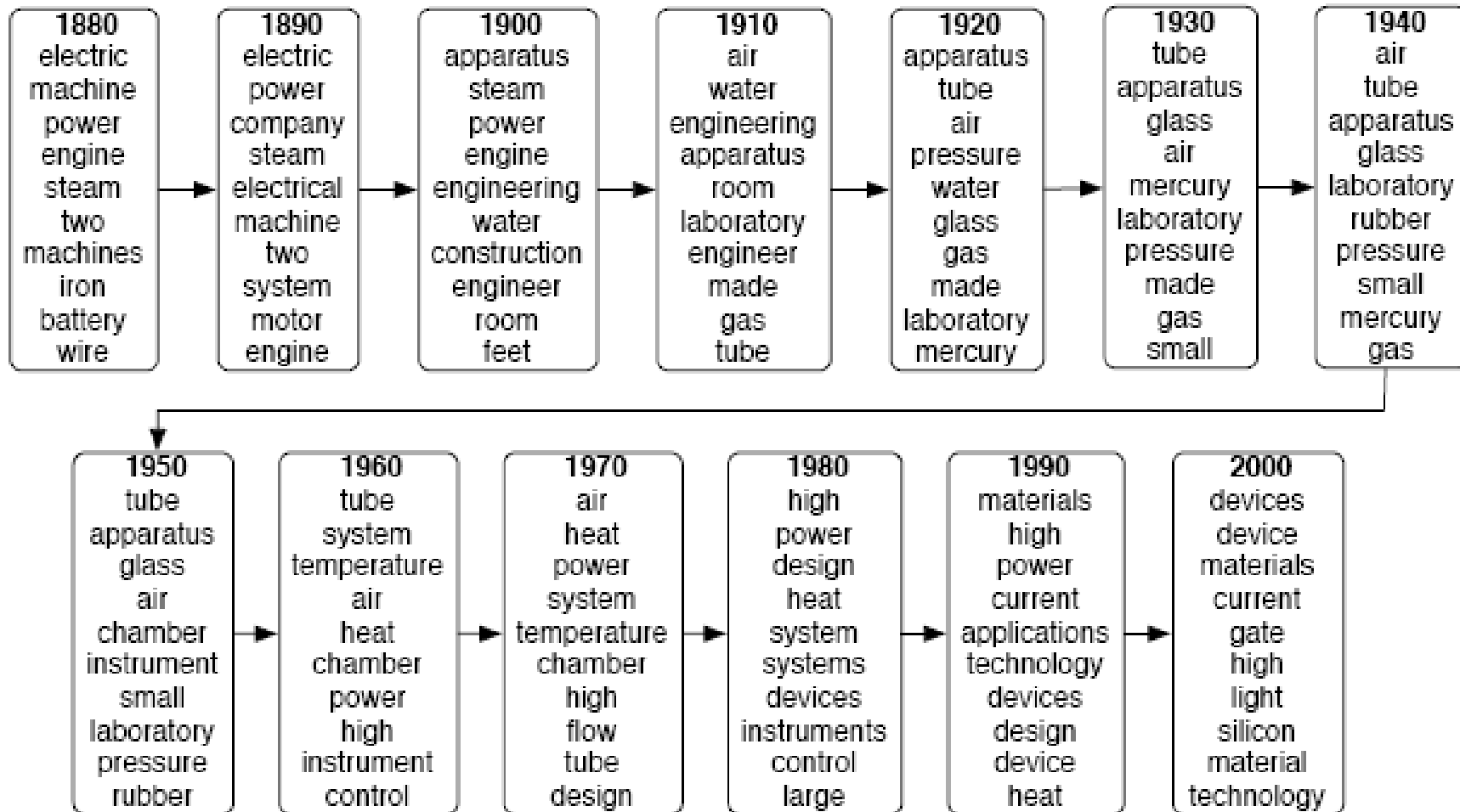


### Analyzed Data:

- JSTOR ([www.jstor.org](http://www.jstor.org)) scanned and ran optical character recognition on *Science* from 1880-2002.
- No reliable punctuation, meta-data, or references
- Restrict to 30K terms that occur more than ten times
- The data are 76M words in 130K documents

# Topic models

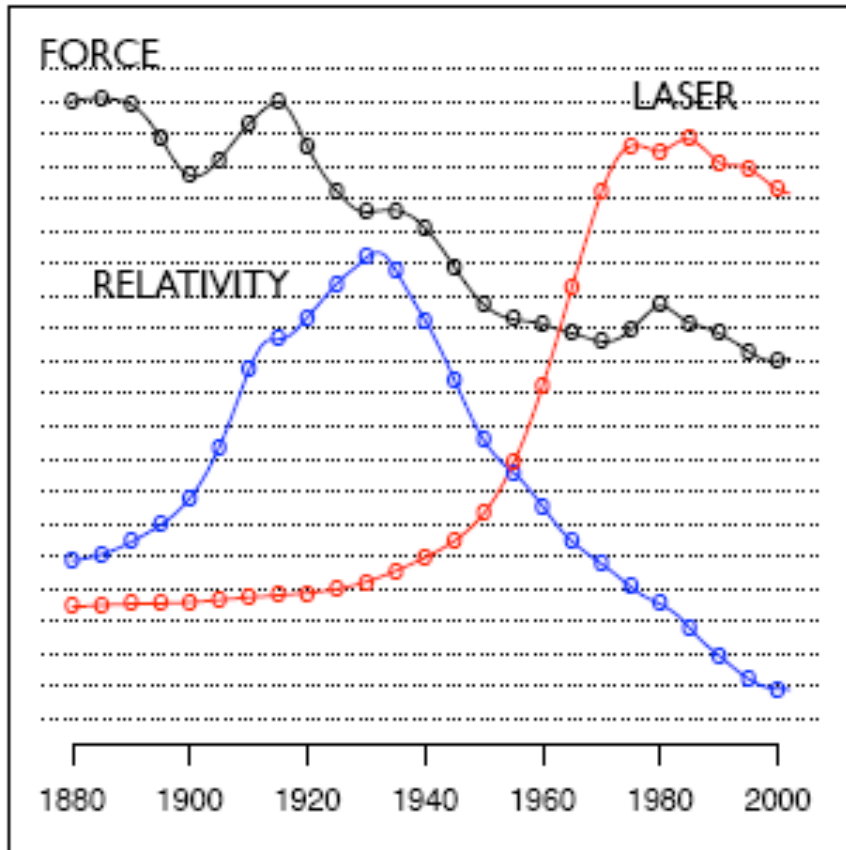
## *Analyzing a topic*



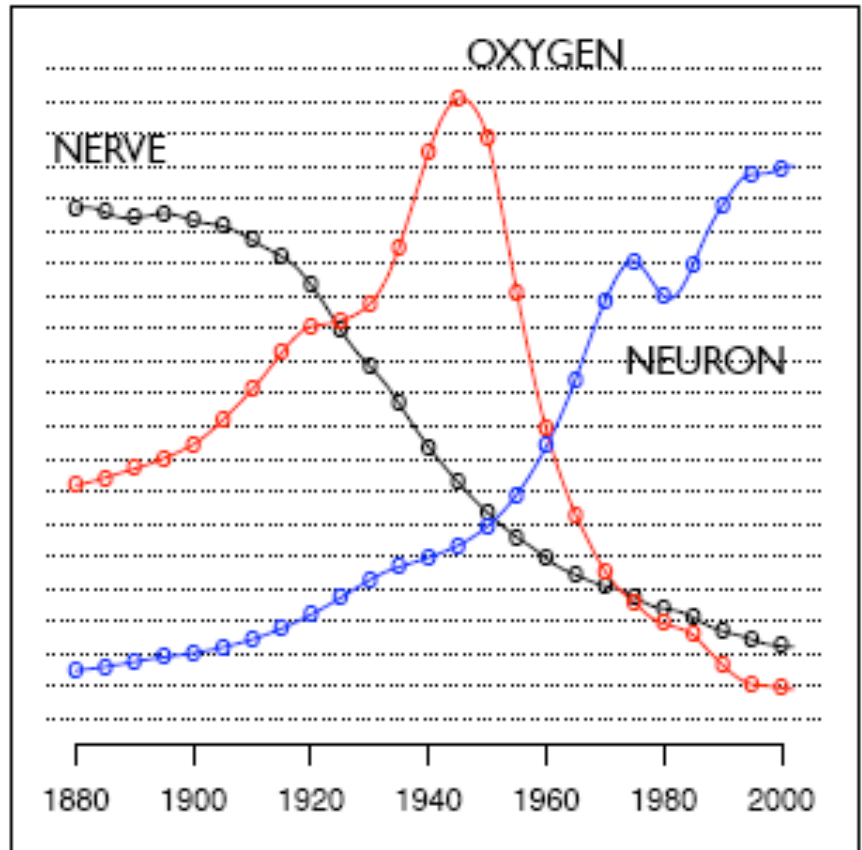
# Topic models

*Visualizing trends within a topic*

**"Theoretical Physics"**



**"Neuroscience"**



# Summary

- LSA and topic models are roads to text meaning.
- Can be viewed as a dimensionality reduction technique.
- Exact inference is intractable, we can approximate instead.
- Various applications and fundamentals for digitalized era.
- Exploiting latent information depends on applications, the fields, researcher backgrounds, ...

# Key references

- S Deerwester, et al. (1990). Indexing by latent semantic analysis. Journal American Society for Information Science (citation 6842).
- Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. Uncertainty in AI (citation 1959).
- Nigam et al. (2000). Text classification from labeled and unlabeled documents using EM, Machine learning (citation 1702).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. J. of Machine Learning Research (citation 3847).



# Some other references

- Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, seventh international conference on World Wide Web 7, p.107-117, April 1998, Brisbane, Australia.
- Taher H. Haveliwala, Topic-sensitive PageRank, 11th international conference on World Wide Web, May 07-11, 2002, Honolulu, Hawaii, USA.
- M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in PageRank. NIPS 14. MIT Press, 2002.
- Lan Nie , Brian D. Davison , Xiaoguang Qi, Topical link analysis for web search, 29th ACM SIGIR conference on Research and development in information retrieval, August 06-11, 2006, Seattle, Washington, USA.