Introduction to Topic Models

Present by Shenglin

The problem with information

www.betaversion.org/~stefano/linotype/news/26/

As more information becomes available, it becomes more difficult to access what we are looking for.

We need new tools to help us organize, search, and understand these vast amounts of information.

Topic modeling

Candida Hofer

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Uncover the hidden topical patterns that pervade the collection.
2. Annotate the documents according to those topics.
3. Use the annotations to organize, summarize, and search the texts.

Discover topics from a corpus

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Topic modeling topics

From a machine learning perspective, topic modeling is a case study in applying hierarchical Bayesian models to grouped data, like documents or images. Topic modeling research touches on

- Directed graphical models
- Conjugate priors and nonconjugate priors
- Time series modeling
- Modeling with graphs
- Hierarchical Bayesian methods
- Fast approximate posterior inference (MCMC, variational methods)
- Exploratory data analysis
- Model selection and nonparametric Bayesian methods
- Mixed membership models

Outline

- Bayesian inference
- Latent Dirichlet allocation
- Application
- Latest progress

Inference problems

- Machine learning is to learn patterns, rules or regularities from data.
- Patterns or rules are represented by distributions.
- Two inference problems:
  - Learning: estimate distribution parameters λ to explain observations $\mathcal{O}$.
  - Prediction: calculate the probability of new observation $\tilde{o}$ given training observations, i.e., to find $P(\tilde{o}|\mathcal{O}) \approx P(\tilde{o}|\lambda)$.
-

$$P(\lambda|\mathcal{O}) = \frac{P(\mathcal{O}|\lambda) \cdot P(\lambda)}{P(\mathcal{O})}, \tag{1}$$

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}. \tag{2}$$

Maximum likelihood (ML)

- Given distribution parameters, the observations are assumed to be independently and identically distributed (i.i.d.).
- Learning:

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} \mathcal{L}(\lambda|\mathcal{O}) \approx \arg\max_{\lambda} \sum_{o \in \mathcal{O}} \log P(o|\lambda). \tag{3}$$

- The common way to estimate the parameter is to solve the system:

$$\frac{\partial \mathcal{L}(\lambda|\mathcal{O})}{\partial \lambda_k} = 0, \forall \lambda_k. \tag{4}$$

- Prediction:

$$\begin{aligned} P(\tilde{o}|\mathcal{O}) &= \int_{\lambda} P(\tilde{o}|\lambda) P(\lambda|\mathcal{O}) d\lambda \\ &\approx \int_{\lambda} P(\tilde{o}|\hat{\lambda}_{ML}) P(\lambda|\mathcal{O}) d\lambda \\ &= P(\tilde{o}|\hat{\lambda}_{ML}). \end{aligned} \tag{5}$$

Maximum a posterior (MAP)

- Learning:

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} \left\{ \sum_{o \in \mathcal{O}} \log P(o|\lambda) + \log P(\lambda) \right\}. \tag{6}$$

- We use parameterized priors $P(\lambda|\alpha)$ with hyperparameters α, in which the belief in the anticipated values of λ can be expressed within the framework of probability.
- *A hierarchy of parameters is created.*
- Prediction:

$$P(\tilde{o}|\mathcal{O}) \approx \int_{\lambda} P(\tilde{o}|\hat{\lambda}_{MAP}) P(\lambda|\mathcal{O}) d\lambda = P(\tilde{o}|\hat{\lambda}_{MAP}). \tag{7}$$

Bayesian inference

- Learning:

$$P(\lambda|\mathcal{O}) = \frac{P(\mathcal{O}|\lambda)P(\lambda)}{\int_{\lambda} P(\mathcal{O}|\lambda)P(\lambda)d\lambda}. \tag{8}$$

- Prediction:

$$\begin{aligned} P(\tilde{o}|\mathcal{O}) &= \int_{\lambda} P(\tilde{o}|\lambda)P(\lambda|\mathcal{O})d\lambda \\ &= \int_{\lambda} P(\tilde{o}|\lambda)\frac{P(\mathcal{O}|\lambda)P(\lambda)}{P(\mathcal{O})}d\lambda. \end{aligned} \tag{9}$$

- The summations or integrals of the marginal likelihood are intractable or there are unknown variables.

Conjugate priors

- Bayesian inference often uses conjugate prior for convenient computation.
- A conjugate prior $P(\lambda)$ of a likelihood function $P(o|\lambda)$ is a distribution that results in a posterior distribution $P(\lambda|o)$ with the same functional form as the prior with different parameters (e.g., exponential family).
- For example, the Dirichlet distribution is the conjugate prior for the multinomial distribution.

Topic modeling

- Large document and image networks require new statistical tools for deep analysis.
- Topic models have become a powerful unsupervised learning tool for large-scale document and image networks.
- Topic modeling introduces **latent topic variables** in text and image that reveal the underlying structure with **posterior inference**.
- Topic models can be used in text summarization, document classification, information retrieval, link prediction, and collaborative filtering.

What are topics?

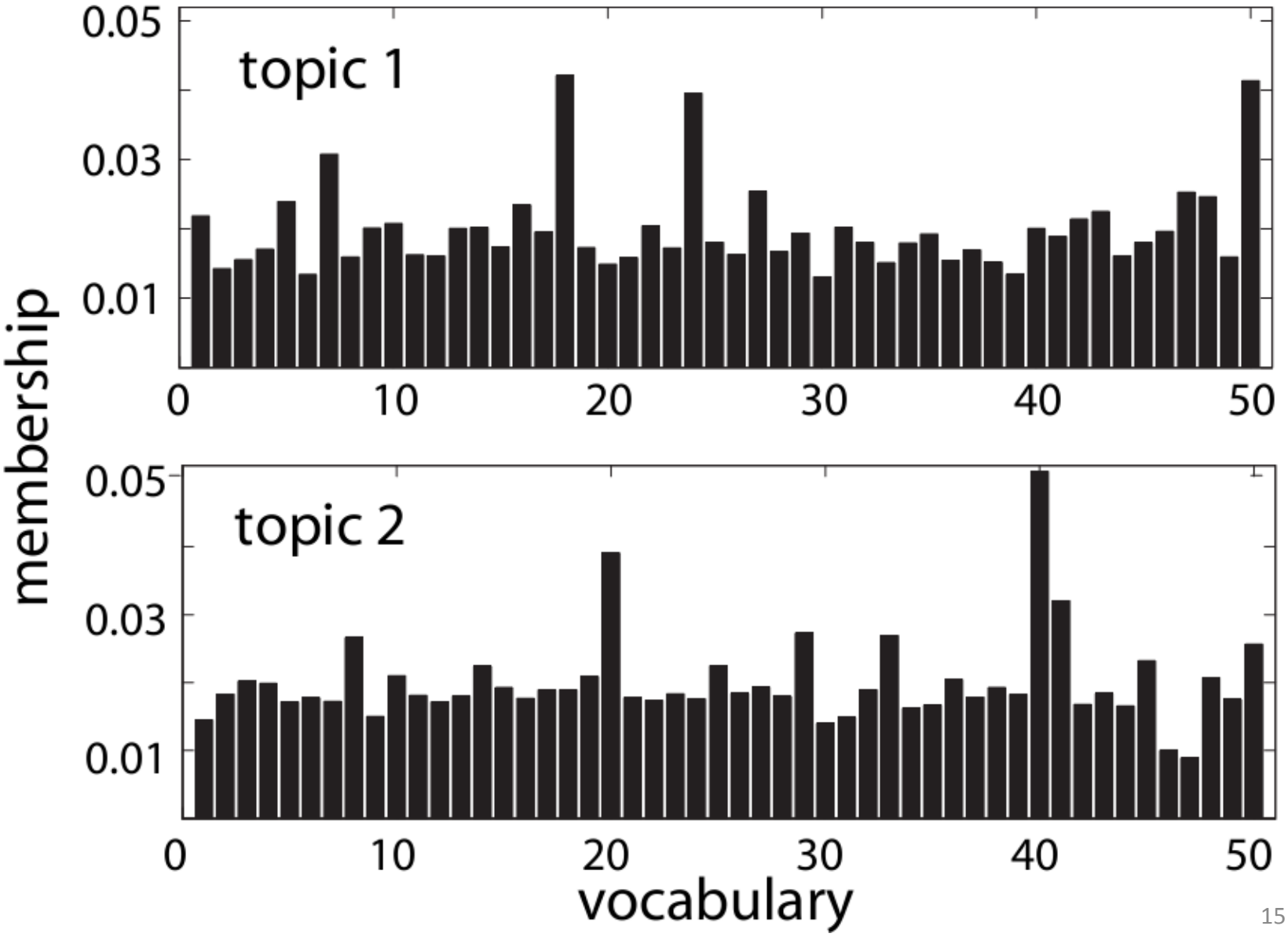
- Topics are a group of related words.

	Topics
1	learn kernel model reinforc algorithm machin classif
2	model learn network neural bayesian time visual
3	retriev inform base model text queri system
4	model imag motion track recognit object estim
5	imag base model recognit segment object detect
1	cell protein express gene activ mutat signal
2	pcr assai detect dna method probe specif
3	apo diseas allel alzheim associ onset gene
4	cell express gene tumor apoptosi protein cancer
5	mutat gene apc cancer famili protein diseas

Topic modeling = word clustering

- A topic is a **fuzzy set of words** with different membership grades based on word co-occurrences in documents.
- Fewer topics per document (homogeneous): Words co-occur within the same document tend to be clustered within the same topic (attractive force).
- Fewer words per topic (heterogeneous): Words tend to be clustered into their dominant topics with the highest membership grade (repulsive force).

Examples


topic 1
topic 2
membership
vocabulary
0.05
0.03
0.01
0
10
20
30
40
50

Outline

- Bayesian inference
- Latent Dirichlet allocation
- Application
- Latest progress

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

ABSTRACT

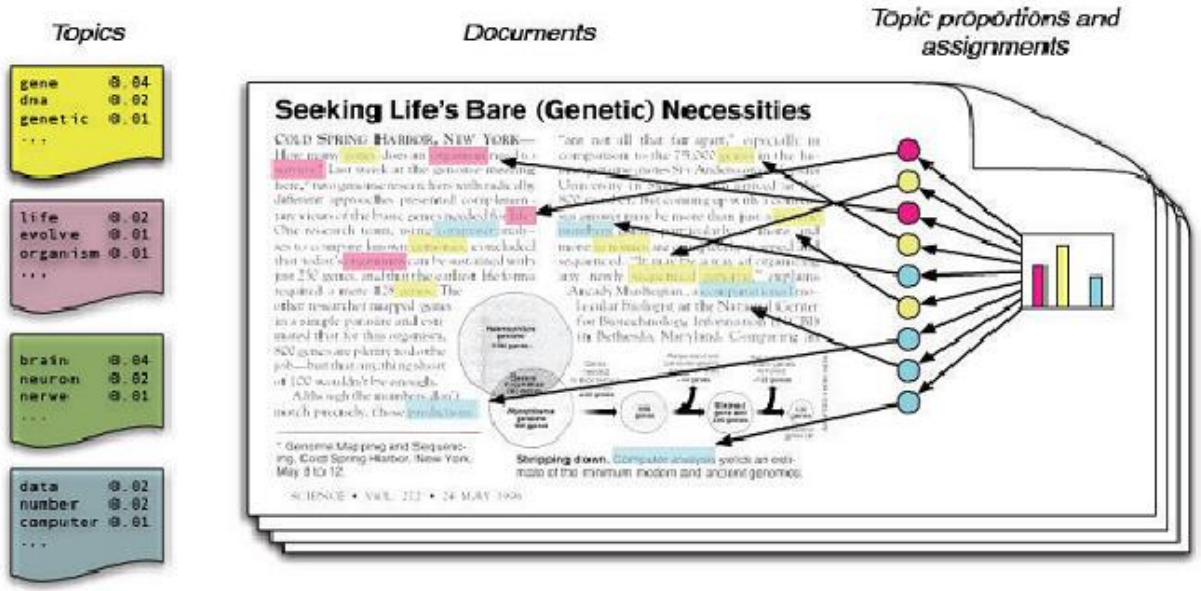
Motivation: Finding precise promoter positions is a topic of considerable interest in bioinformatics. After locating promoters, a further interesting question is to predict their tissue-specificity (TS), which plays an important role in understanding many complex diseases and cell-cell variability in gene expression. Identification of TS promoters often involves technically demanding and expensive microarray techniques or cap analysis of gene expression. Therefore, efficient TS promoter classification algorithms are needed without experimental support of ESTs and microarray data.

Results: Based on promoter word compositions, we propose latent Dirichlet allocation (LDA) to extract statistically significant groups of related words referred to as modules. Thus, each promoter can be represented as the low-dimensional LDA-based module proportions rather than the high-dimensional word vector. Subsequently, we use

- Simple intuition: a document exhibits multiple topics.

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

Probabilistic model

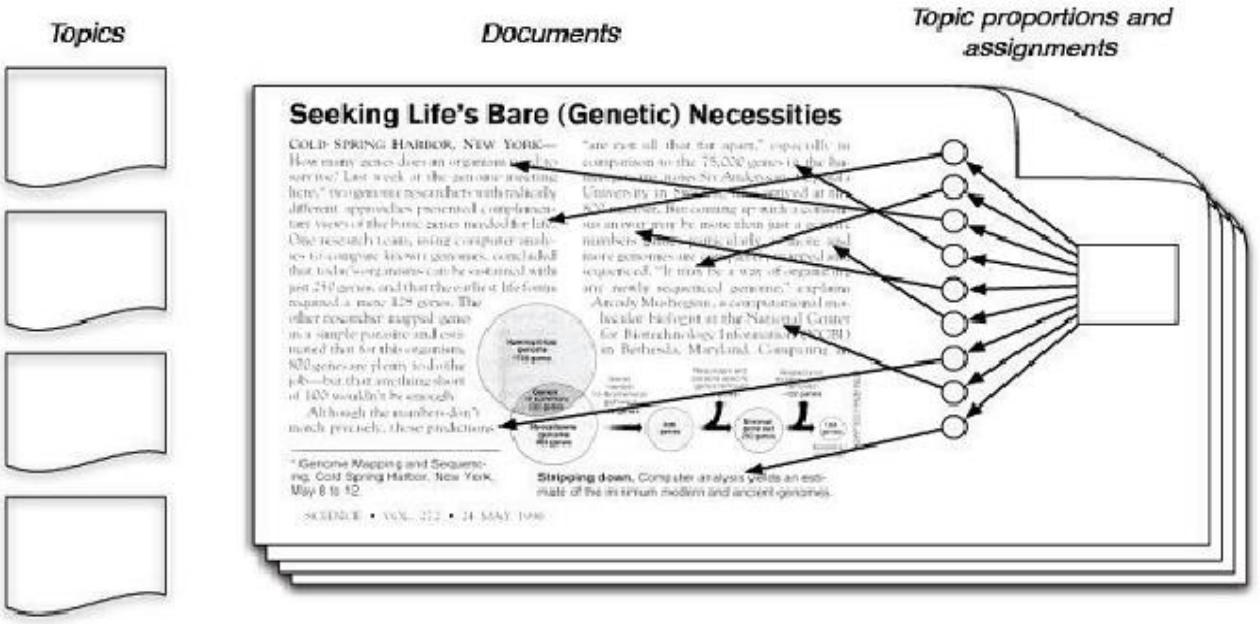


(from David Blei)

- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

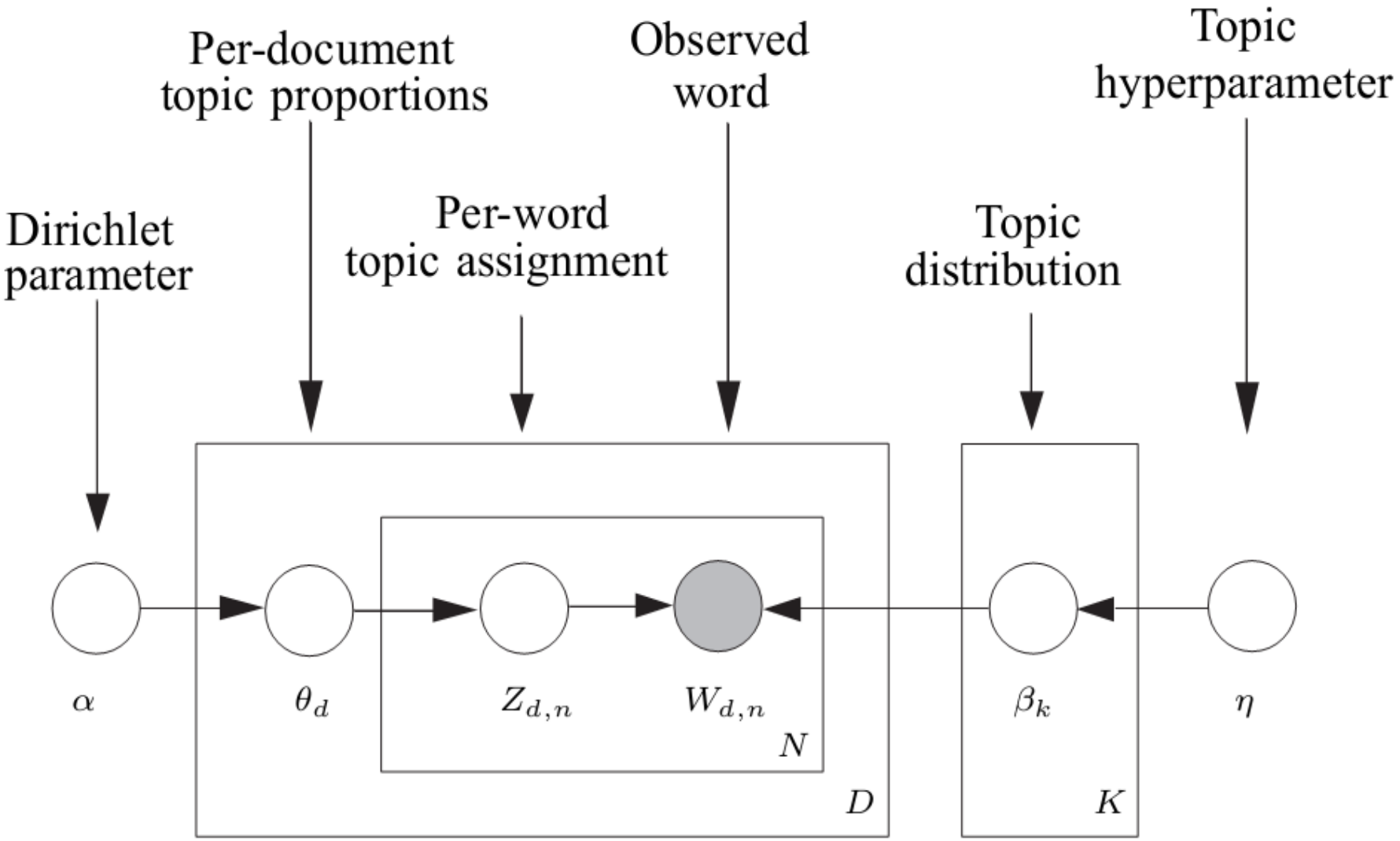
Probabilistic model (2)



(from David Blei)

- We only observe the documents
- Our goal is to **infer** the underlying topic structure

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

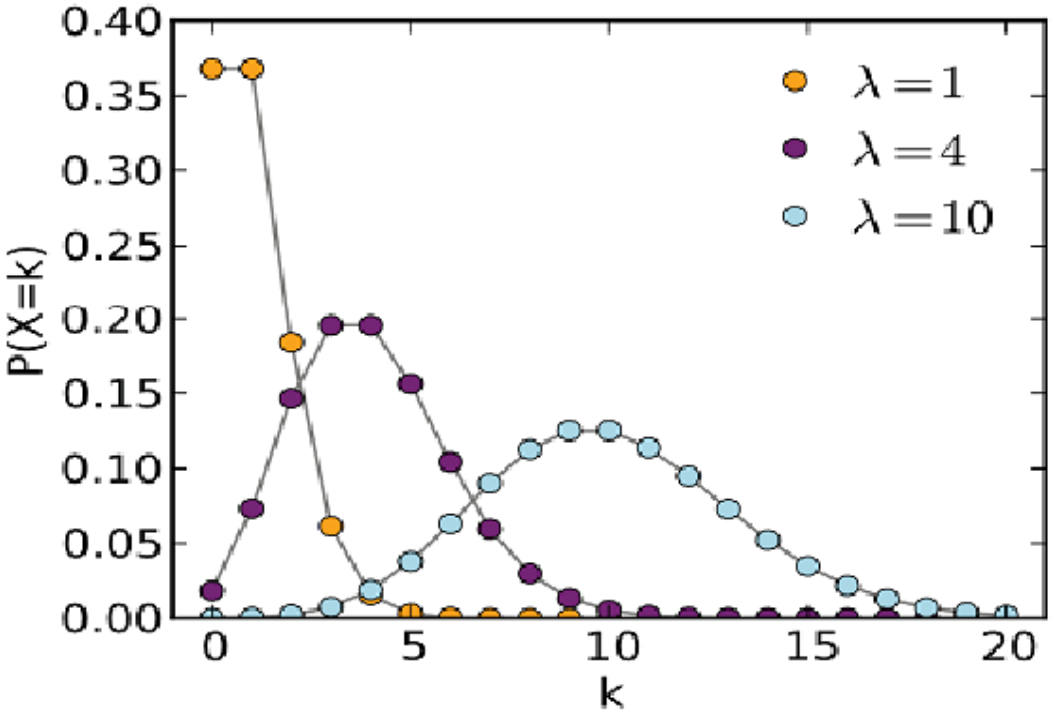


Latent Dirichlet allocation (LDA) [Blei et al., 2003]

LDA assumes the following generative process:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of N words w_n:
   (a) Choose topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose word $w_n \sim$ from $P(w_n|z_n, \beta)$

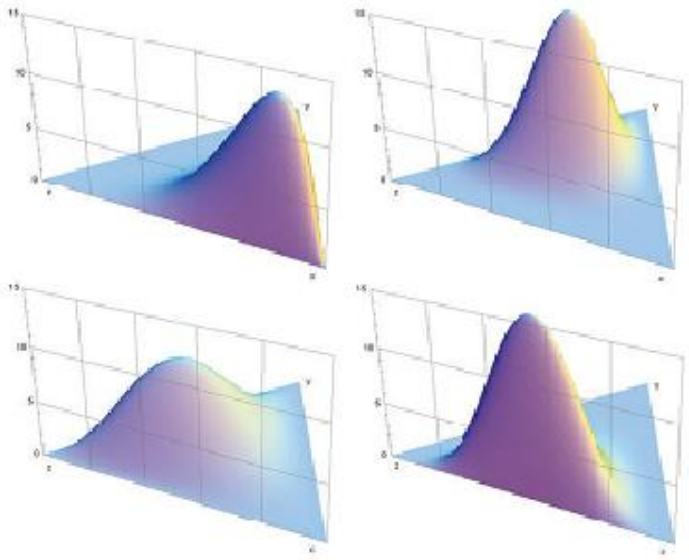
Recap on distributions: Poisson



(from Wikipedia)

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

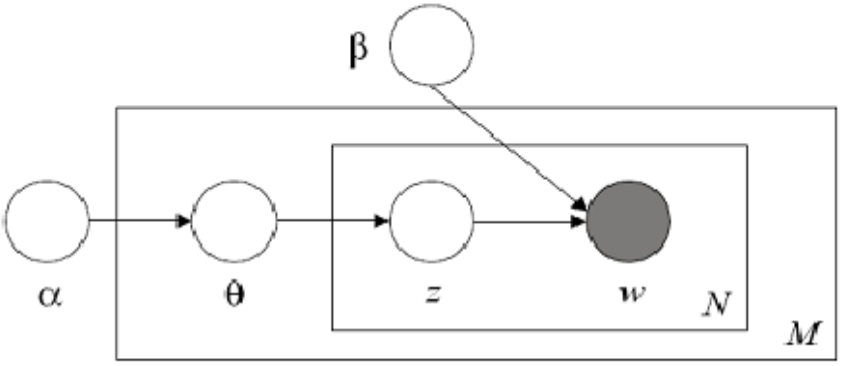
Recap on distributions: Dirichlet



Dirichlet distribution, K=3 for various parameter vectors α
Clockwise from top left:
$\alpha = (6, 2, 2), (3, 7, 5), (6, 2, 6), (2, 3, 4)$.
(from Wikipedia)

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

Inference



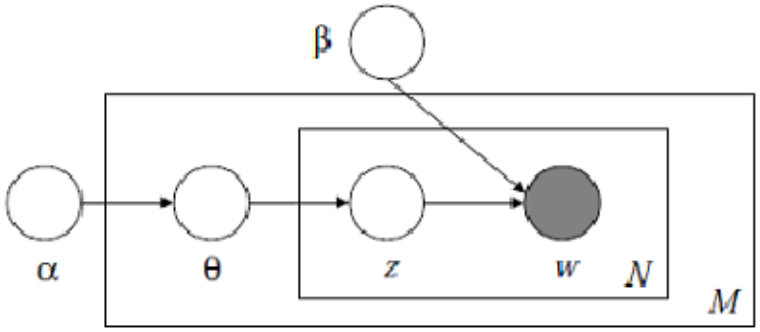
- Given corpus (w is observed), parameters (α, β), calculate p(θ,z| α, β, w)
- Intractable

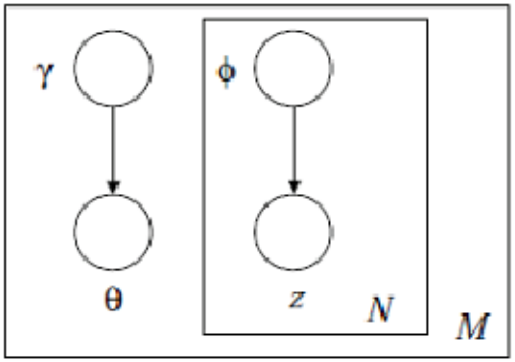
$$\frac{p(\theta \mid \alpha)\prod_{n=1}^{N} p(z_n \mid \theta)p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha)\prod_{n=1}^{N}\sum_{z=1}^{K} p(z_n \mid \theta)p(w_n \mid z_n, \beta_{1:K})}$$

  - Gibbs sampling
  - Variational inference

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

Variational Inference





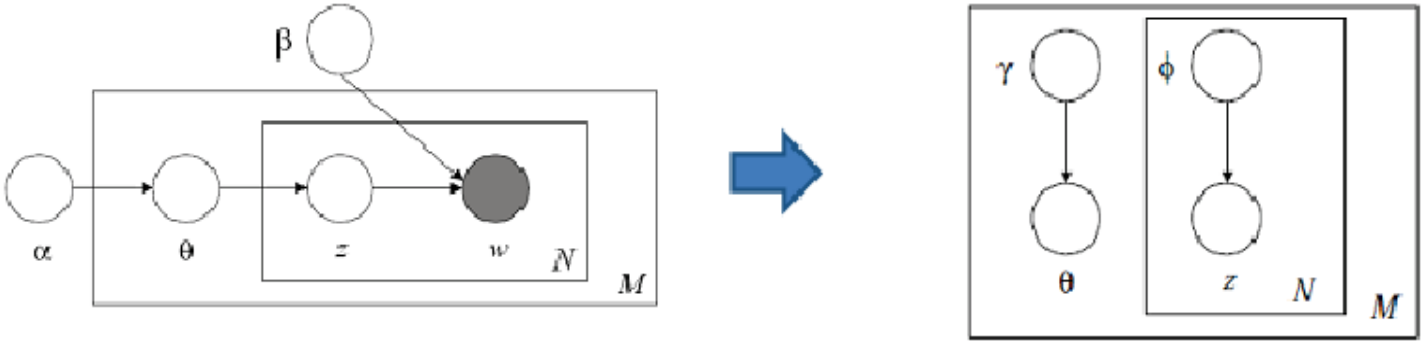
Choose γ, ϕ to approximate posterior distribution of θ,z

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} D(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)).$$

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right).$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}.$$

Variational Inference



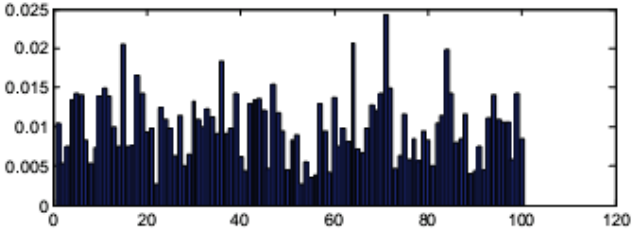
(1) initialize $\phi_{ni}^0 := 1/k$ for all i and n
(2) initialize $\gamma_i := \alpha_i + N/k$ for all i
(3) **repeat**
(4) **for** $n = 1$ **to** N
(5) **for** $i = 1$ **to** k
(6) $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
(7) normalize ϕ_n^{t+1} to sum to 1.
(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
(9) **until** convergence

Latent Dirichlet allocation (LDA) [Blei et al., 2003]

Parameter estimation

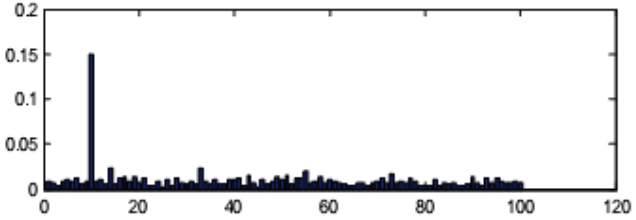
- α controls proportion distribution of topics in one document.

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ equally large



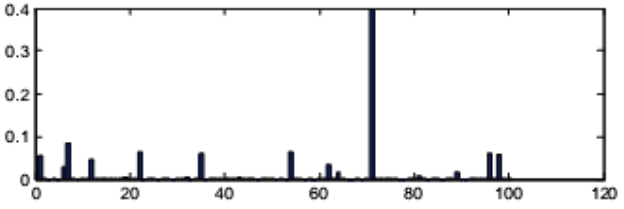
Topics are almost equally likely

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ equal, but α_{10} is larger



10th topic is more likely to appear

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ are equally small

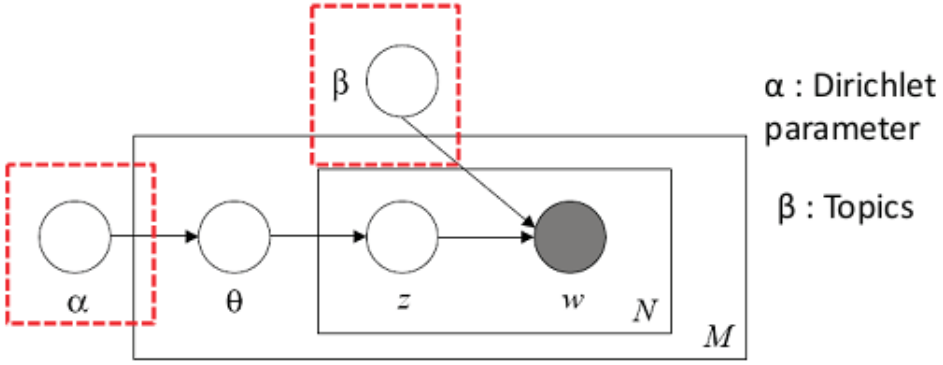


Topics distribution is sparse (few topics in one document), with one random peak

- β is the probability matrix of topics and words

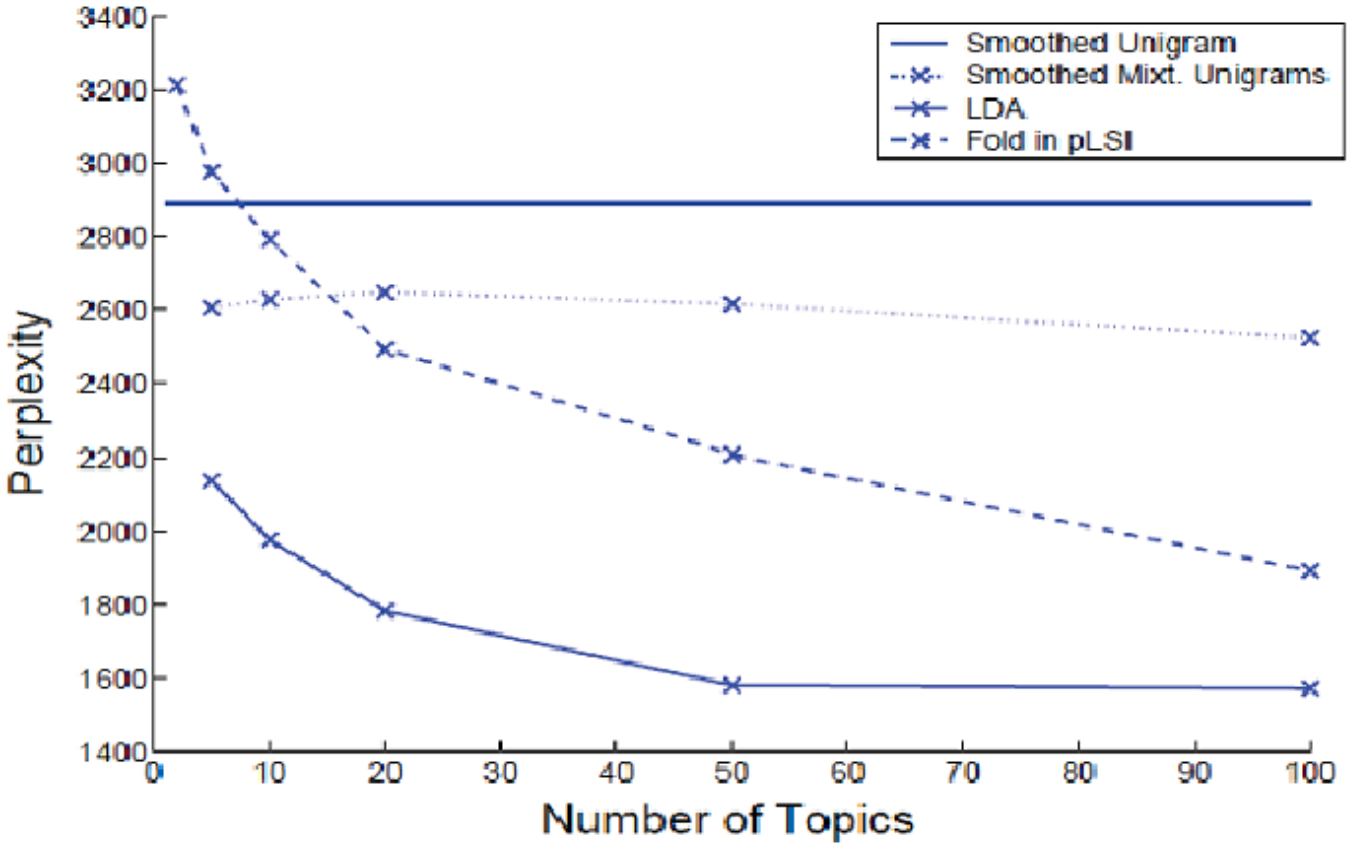
Latent Dirichlet allocation (LDA) [Blei et al., 2003]

Parameter Estimation



- Try to estimate parameters (α, β), given corpus {w}.
- EM algorithm:
  - E step: find the optimizing value of ϒ, ϕ
  - M step: maximize log likelihood w.r.t α and β.

Application/Empirical Results



$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}.$$

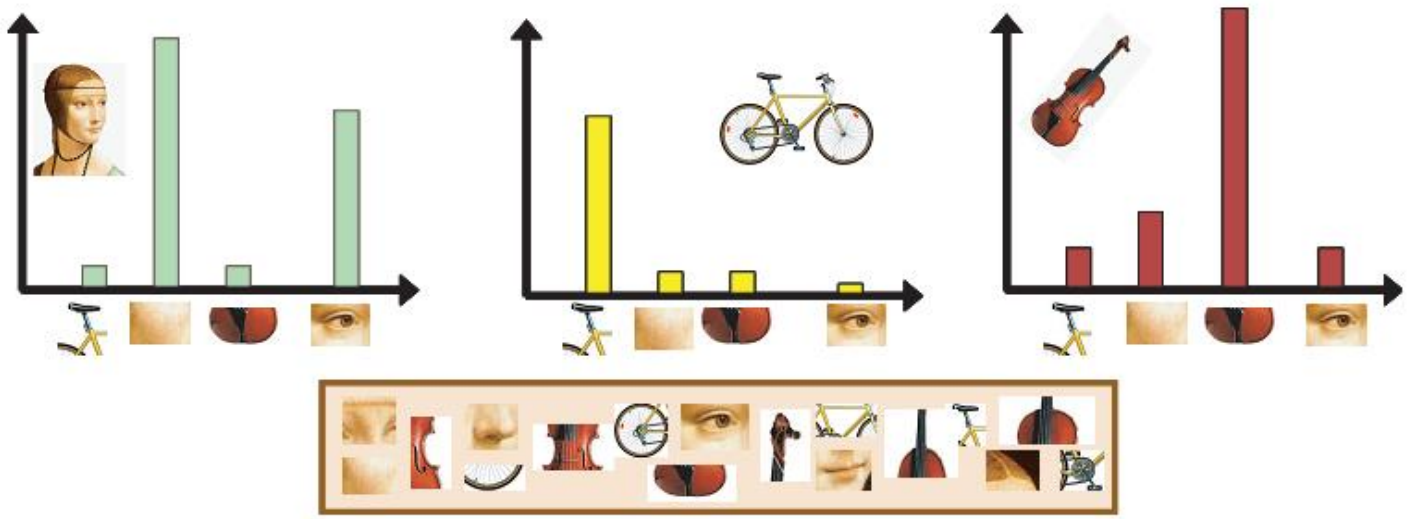
Outline

- Bayesian inference
- Latent Dirichlet allocation
- Application
- Latest progress

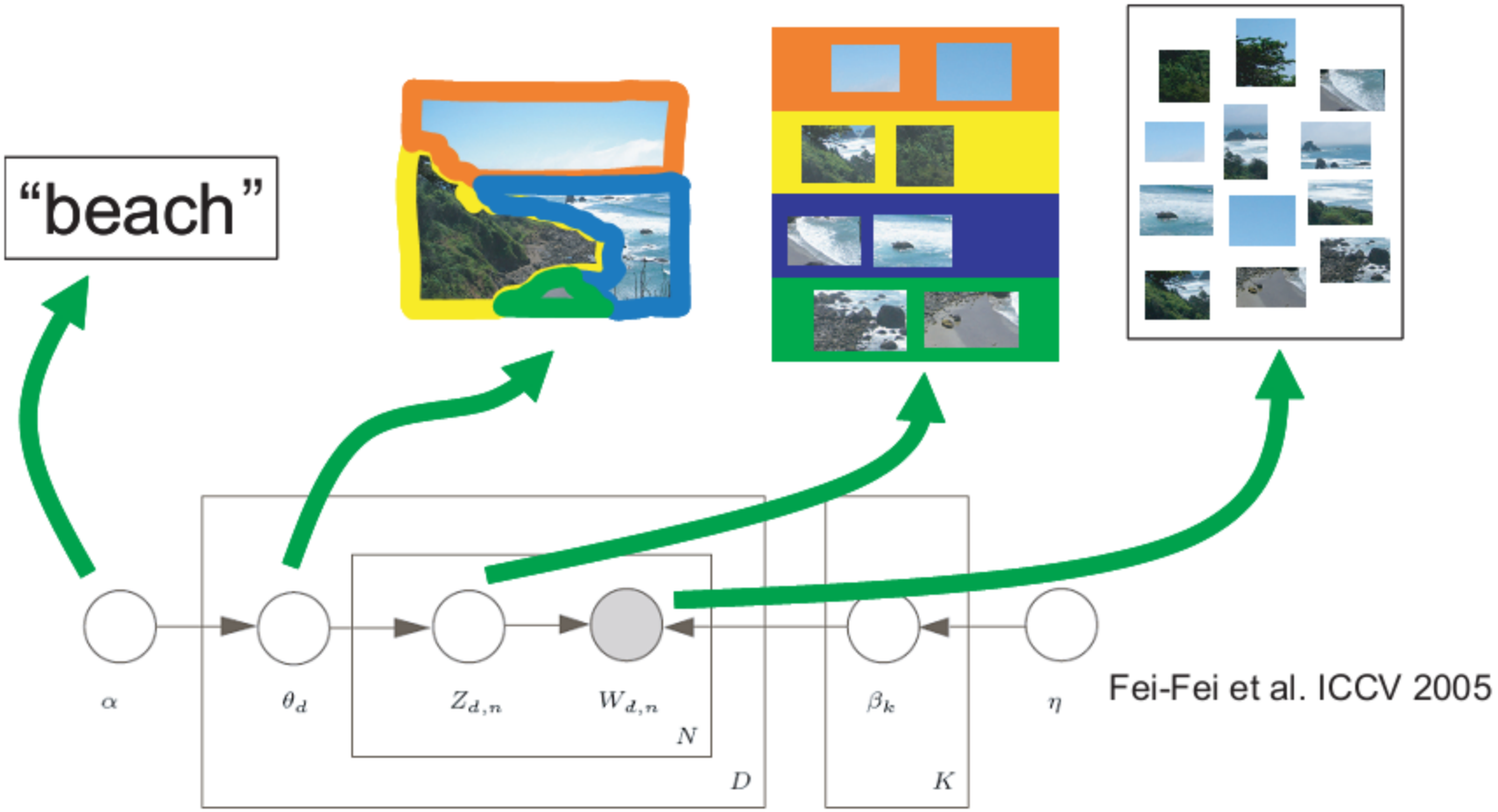
Natural scene categorization [Fei-Fei and Perona, 2005]

Object → **Bag of 'words'**

Bag of words (BOW)

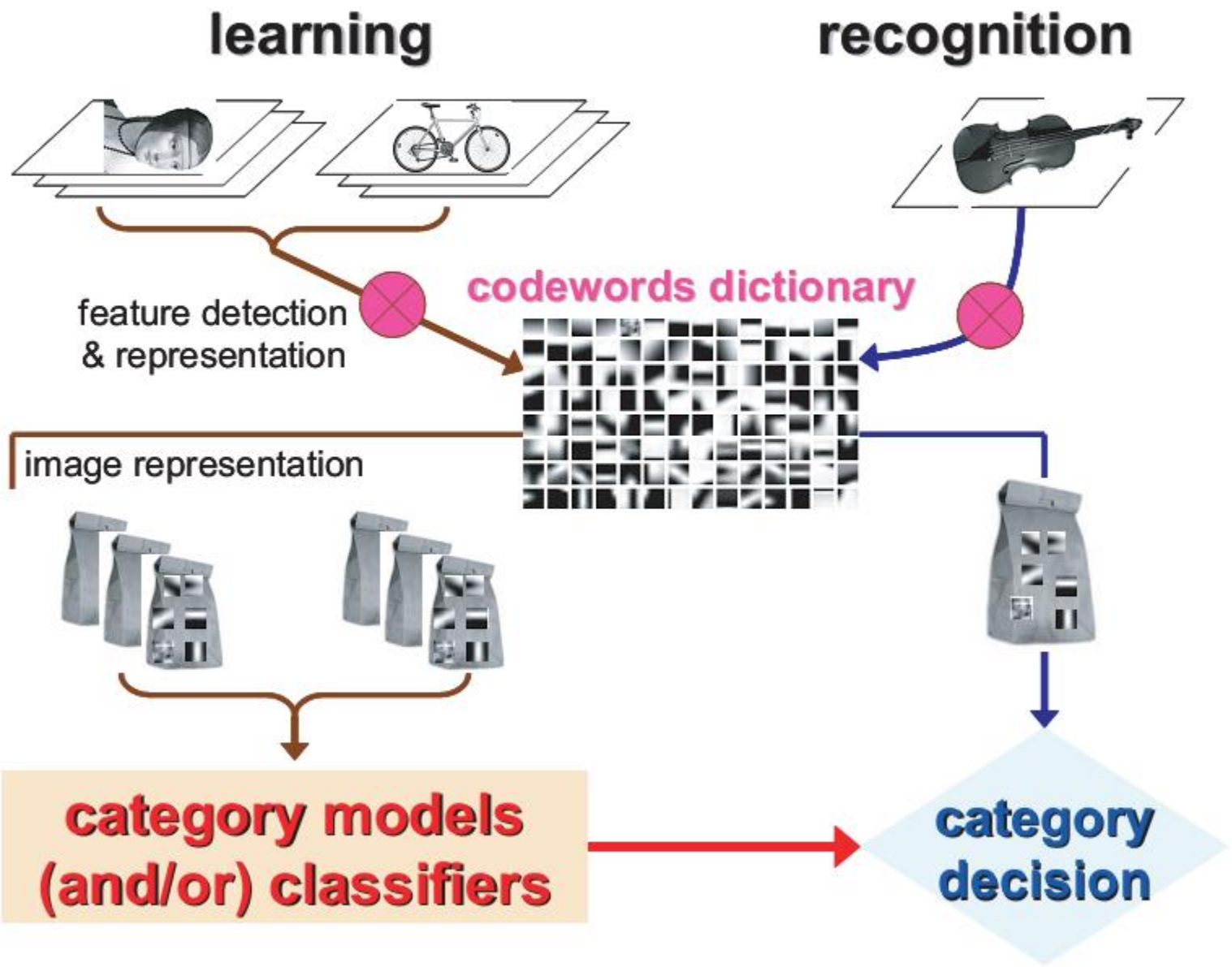


- Independent features
- Histogram representation



Latent Dirichlet Allocation (LDA)

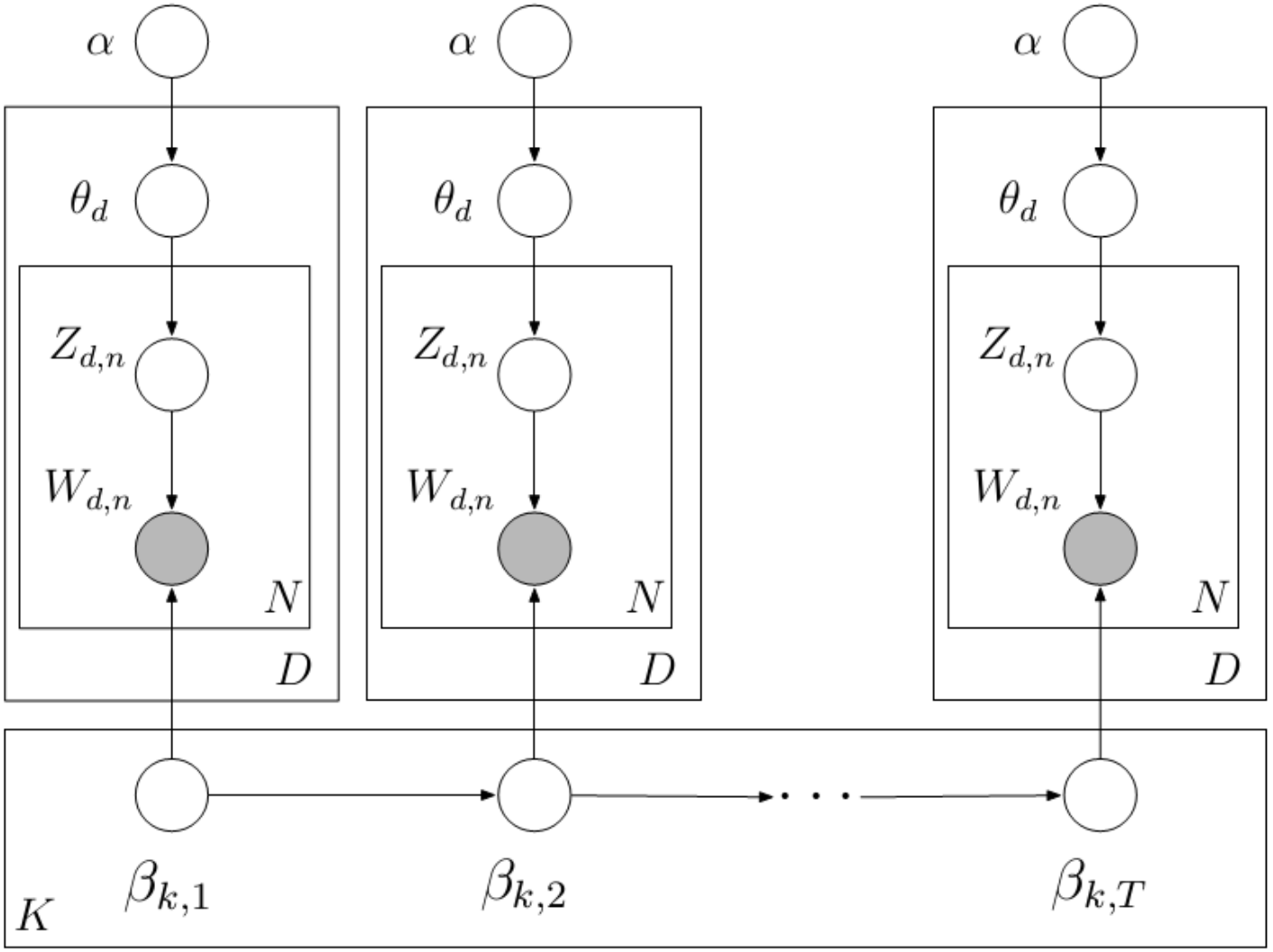
Categorization system



Outline

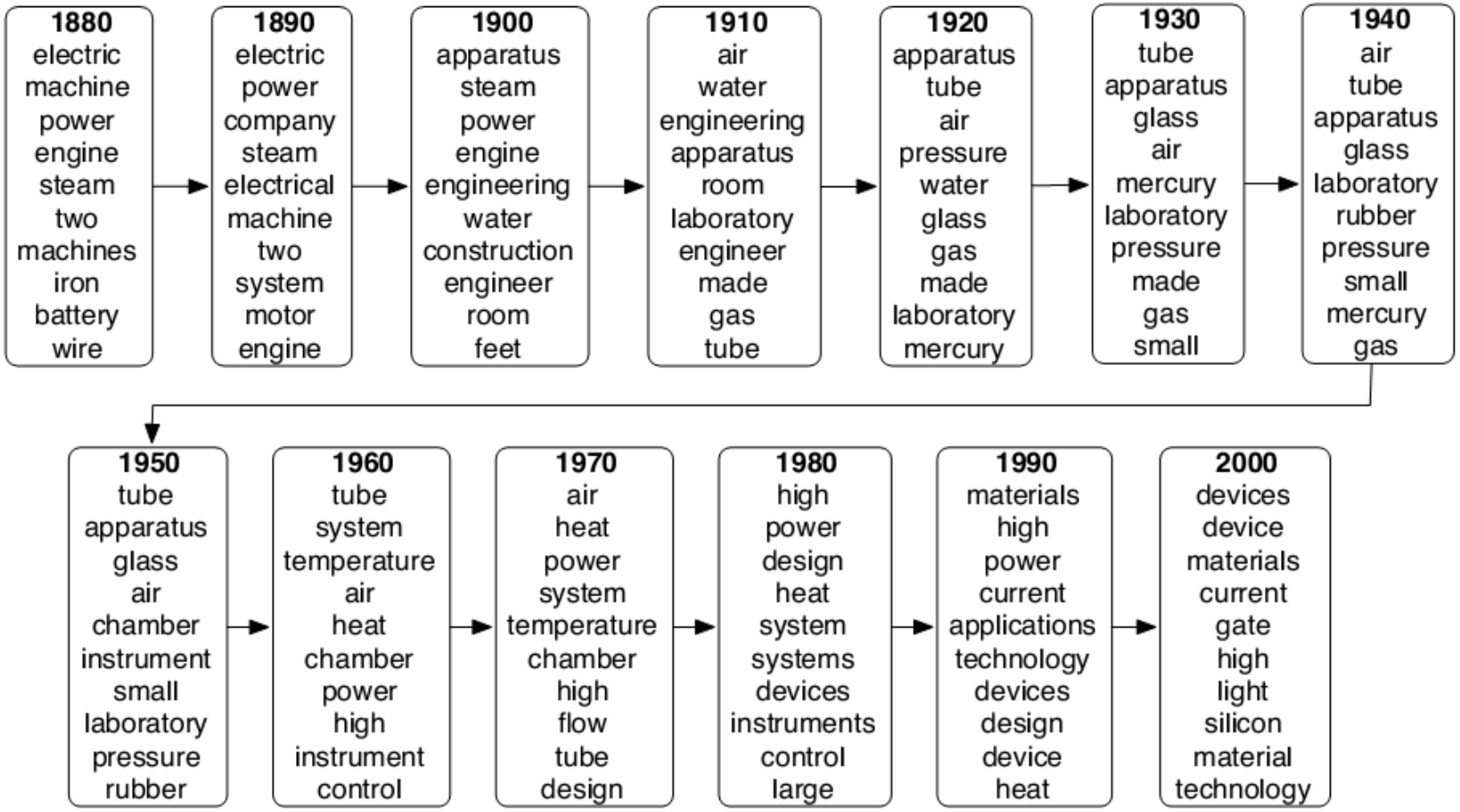
- Bayesian inference
- Latent Dirichlet allocation
- Application
- Latest progress

Dynamic topic models

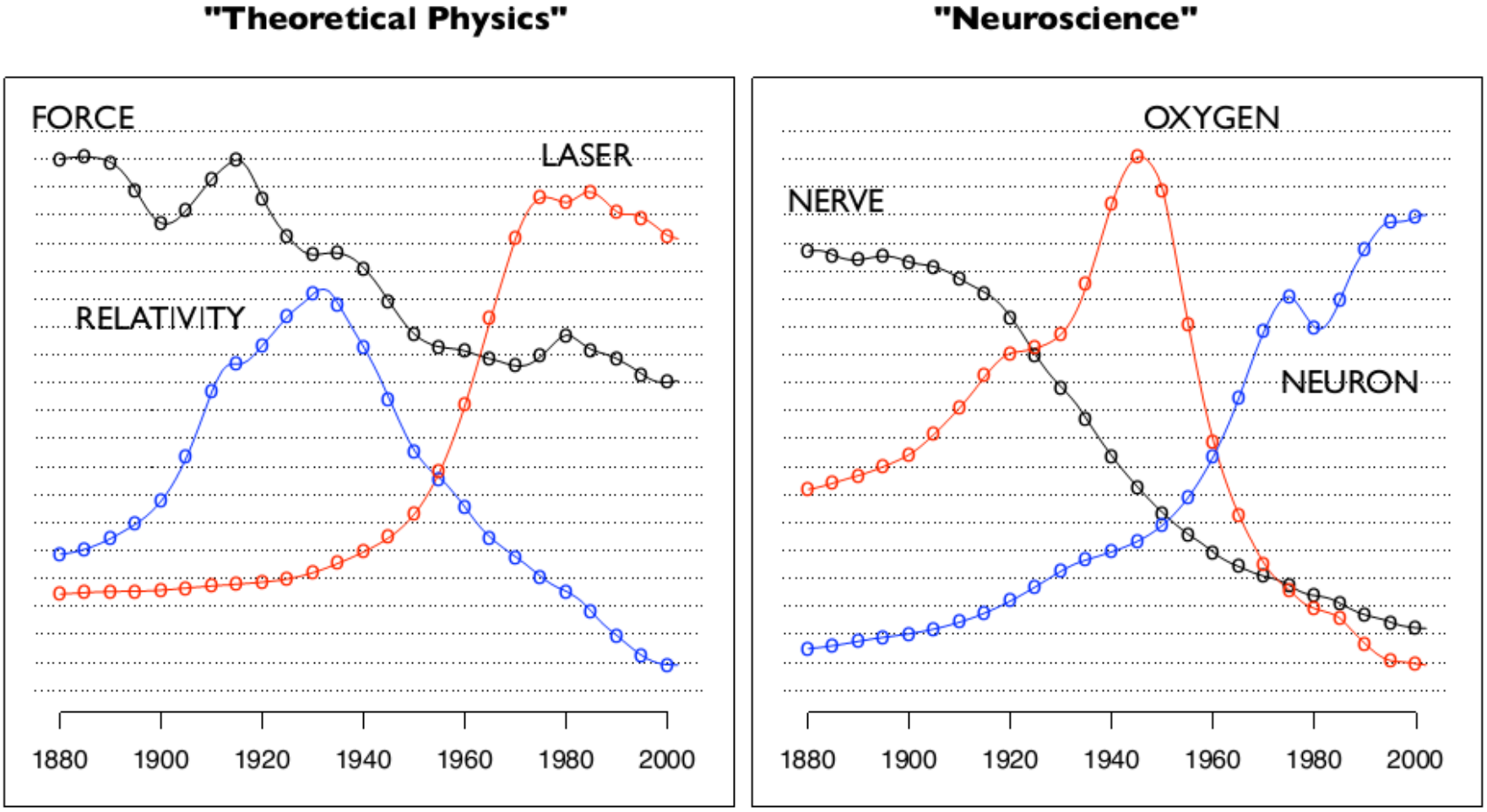


Topics drifting in time

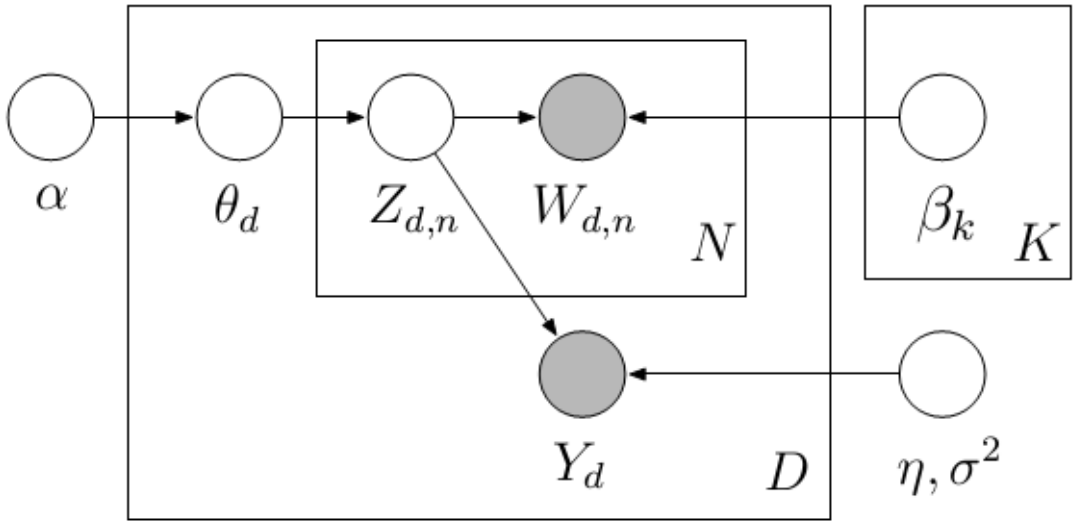
Dynamic topic models



Dynamic topic models

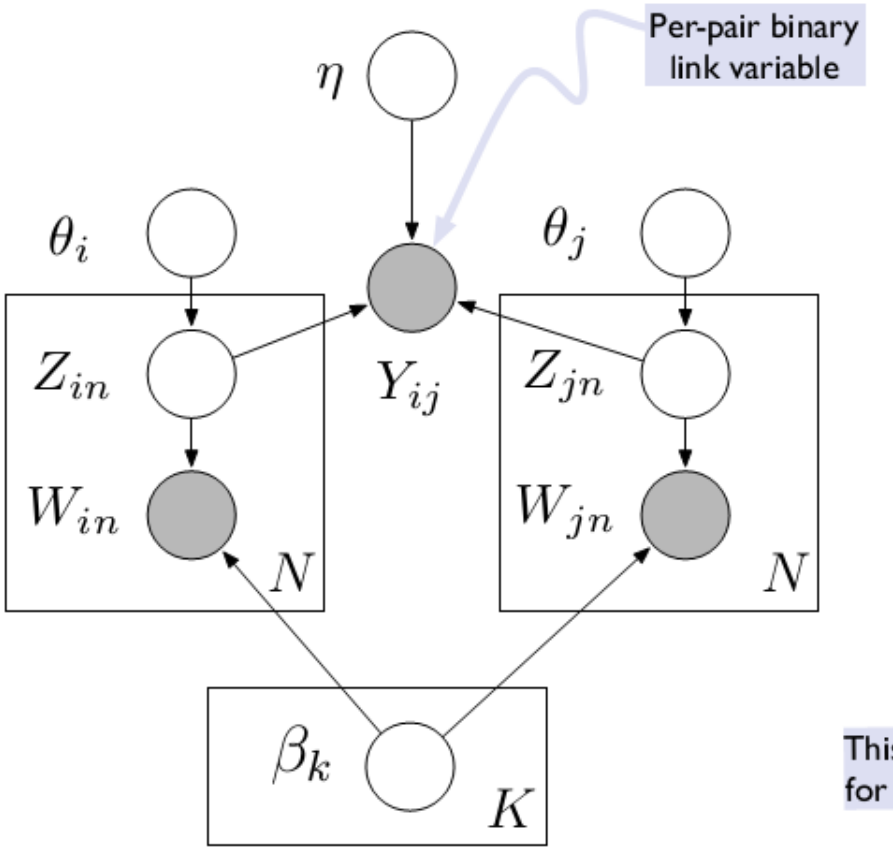


Supervised LDA



- SLDA enables model-based regression where the predictor is a document.
- It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).
- SLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.
- SLDA has been extended to generalized linear models, e.g., for image classification and other non-continuous responses.

Relational topic models



- Adapt fitting algorithm for sLDA with binary GLM response
- RTMs allow predictions about new and unlinked data.
- These predictions are out of reach for traditional network models.

Ideal point topic models