Latent Dirichlet Allocation

Generative and Deep Learning (GDL)

Davide Bacciu (davide.bacciu@unipi.it)

UNIVERSITÀ DI PISA

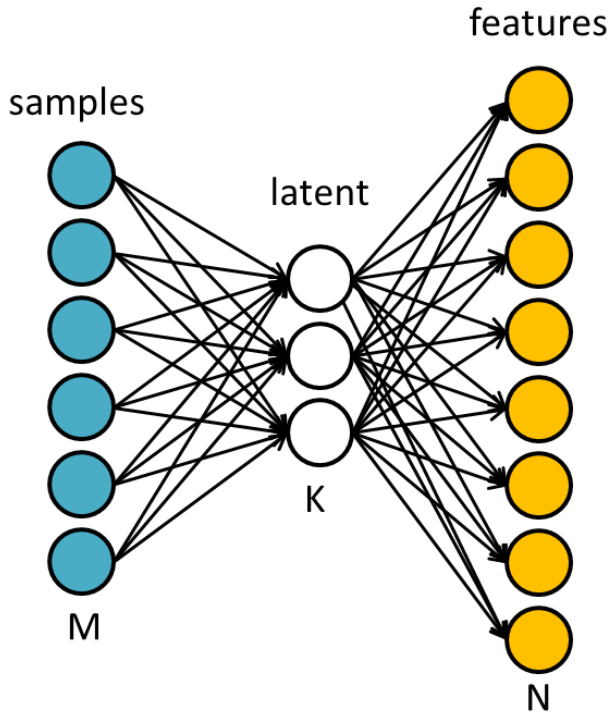
Lecture Outline

- Latent topic models
  - Hidden variable models where latents have a specialized semantics
- Bayesian latent topic models
  - A class of generative models for which variational or approximated methods are needed
- Latent Dirichlet Allocation
  - Possibly the simplest Bayesian latent variable model
  - Many applications in unsupervised text analytics, machine vision, …

Latent topic models

Latent Variable Models

Define a latent space where high-dimensional data can be represented

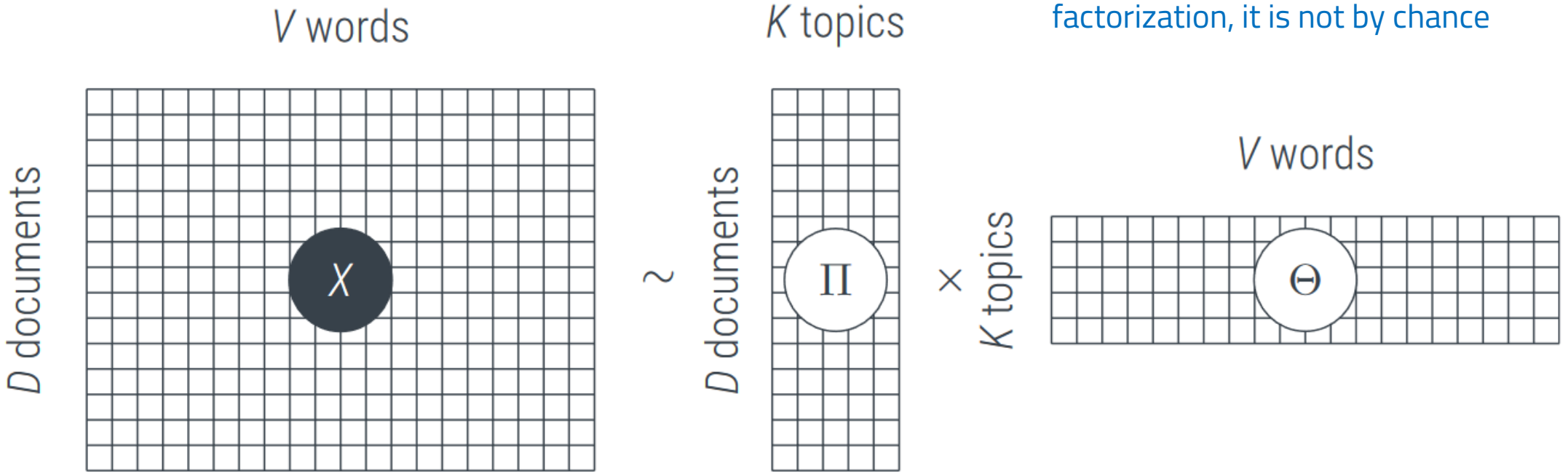


Assumption

Latent variables conditional and marginal distributions are more tractable than the joint distribution $P(\mathcal{X})$ (e.g $K \ll N$)

Topic Models

If you see similarities with matrix factorization, it is not by chance



Representing large-dimensional sparse documents (in word representation) in a compressed latent space of (sparse) topics

Adding a probabilistic perspective

A **Bag of Words** (BOW) representation of a document is the classical example of **multinomial data** (for text, images, graphs,...)

A BOW dataset (corpora) is the $N \times M$ term-document matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Ni} & \cdots & x_{NM} \end{bmatrix}$$

- N : number of vocabulary items w_j
- M : number of documents d_i
- $x_{ij} = n(w_j, d_i)$: number of occurrences of w_j in d_i

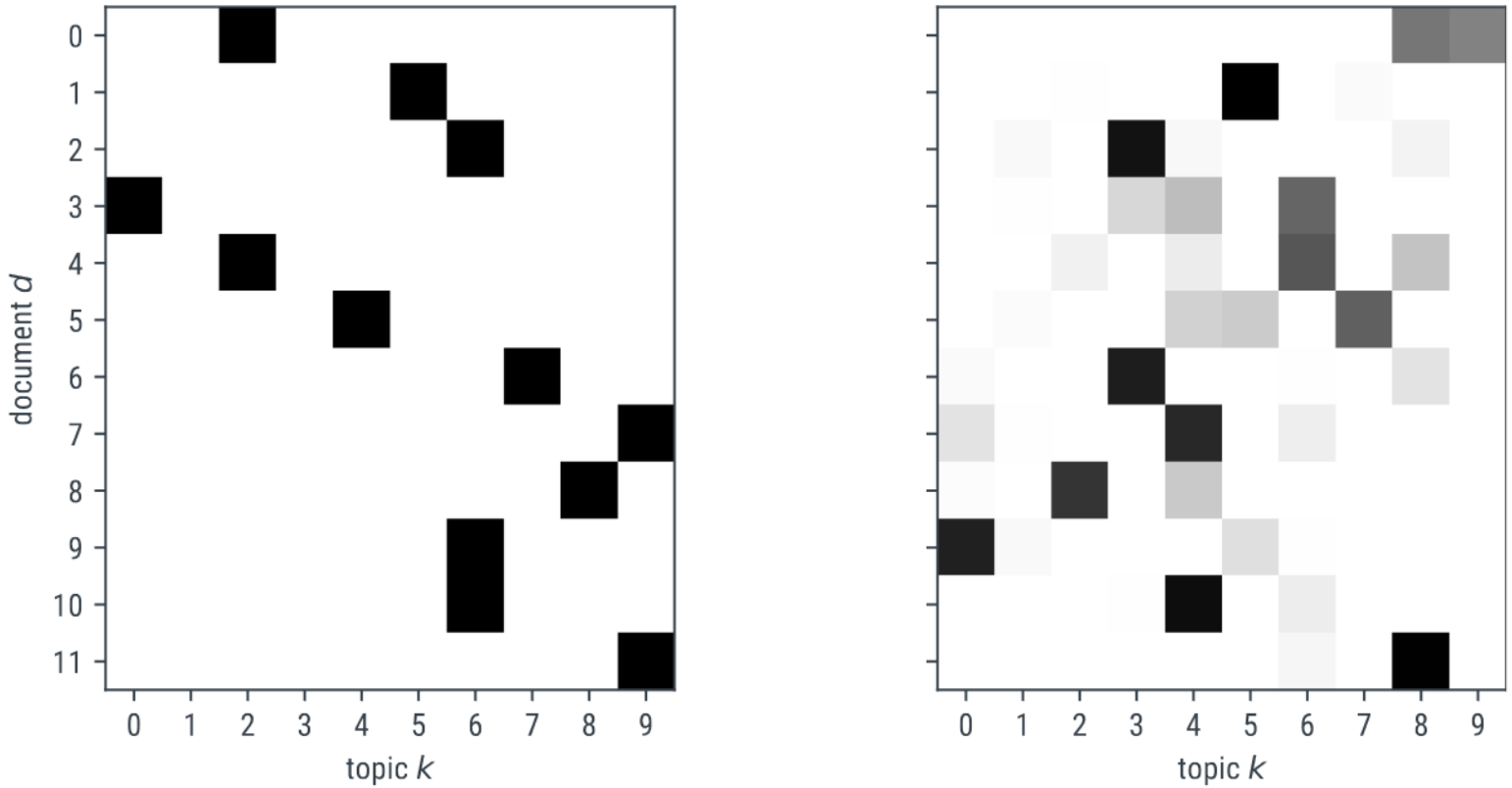
Documents as Mixtures of Latent Topics

Latent topic models are probabilistic models considering documents (containers of items w_j) as a mixture of topics

- A topic identifies a pattern in the co-occurrence of multinomial items w_j within the documents
- Mixture of topics ⇒ Associate an interpretation (topic) to each item in a document, whose interpretation is then a mixture of the items' topics

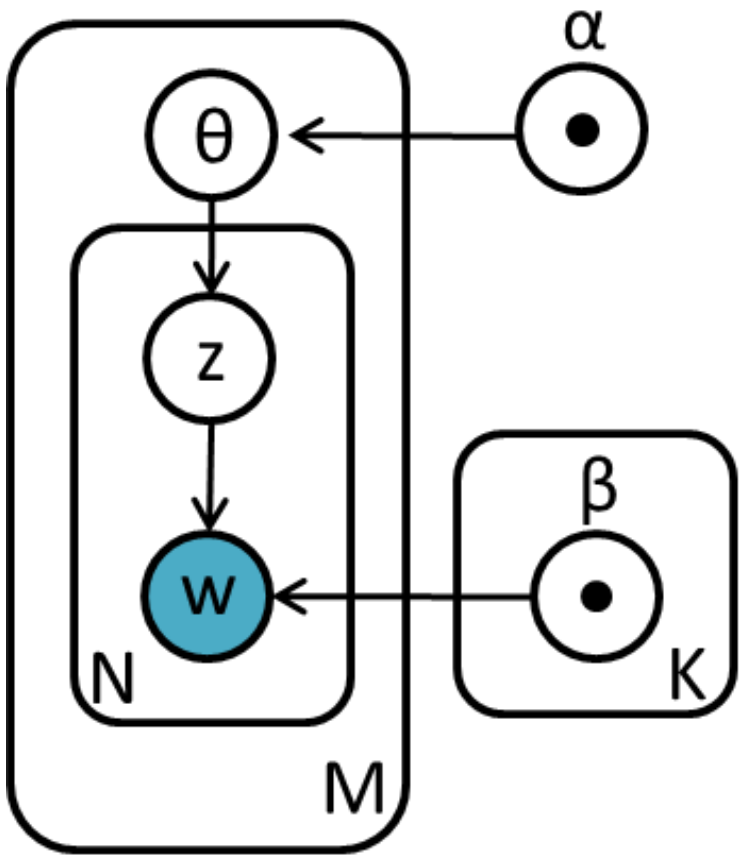
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Ni} & \cdots & x_{NM} \end{bmatrix}$$

Single Vs Mixed Topic-Document Membership



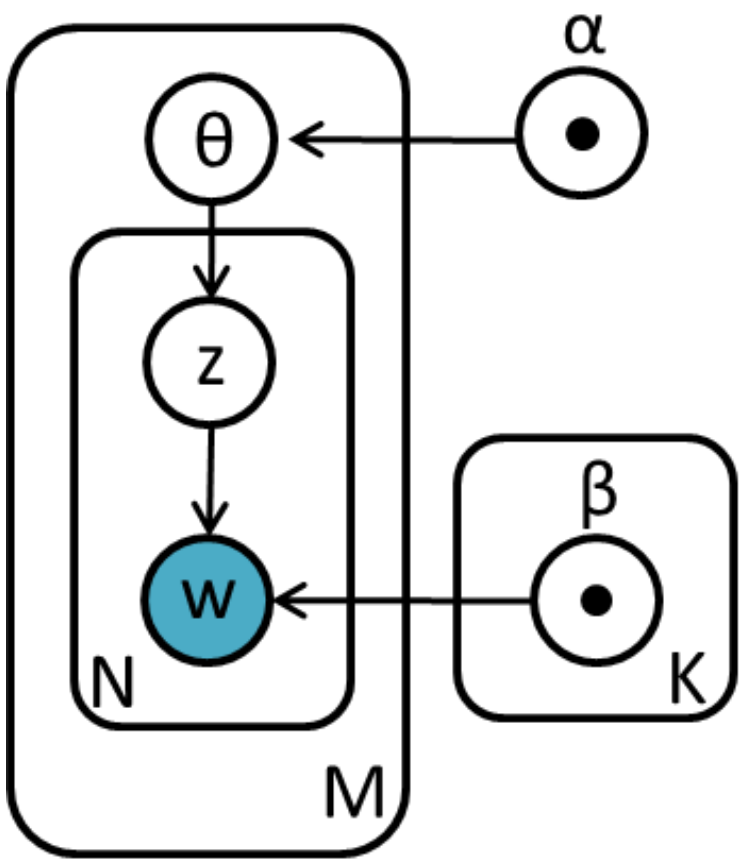
Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA)



- LDA models a document as a mixture of topics z
  - Assigning one topic z to each item w with probability $P(w|z,\beta)$
  - Pick one topic for the the whole document with probability $P(z|\theta)$
- Key point - Each document has its personal topic proportion θ sampled from a distribution
  - θ defines a multinomial distribution but it is a random variable as well

LDA Distributions



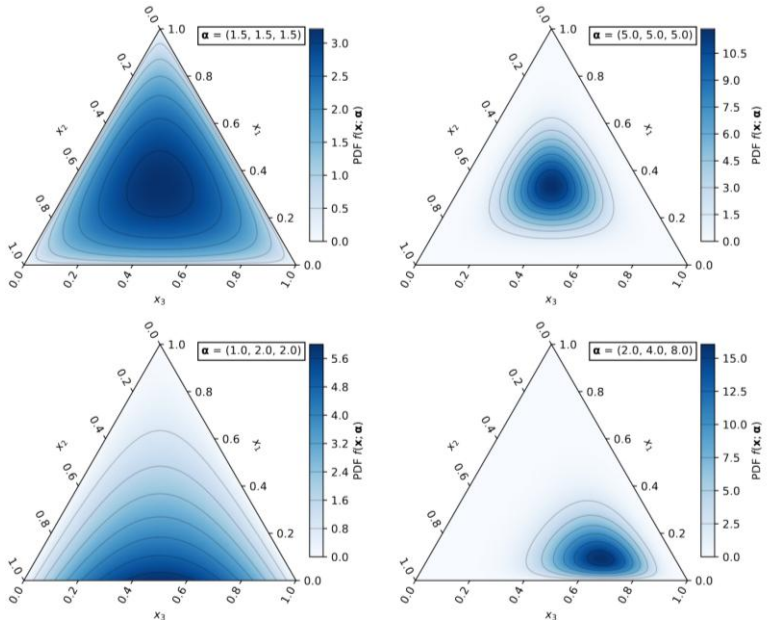
- $P(w|z,\beta)$ multinomial item-topic distribution
- $P(z|\theta)$ multinomial topic distribution with document-specific parameter θ
- $P(\theta|\alpha)$ Dirichlet distribution with hyperparameter α
  - A distribution for vectors that sum to 1 (simplex)
  - The elements of a multinomial are vector that sum to 1!

Dirichlet Distribution

- Why a Dirichlet distribution?
  - Conjugate prior to multinomial distribution
- Dirichlet distribution (Prior)

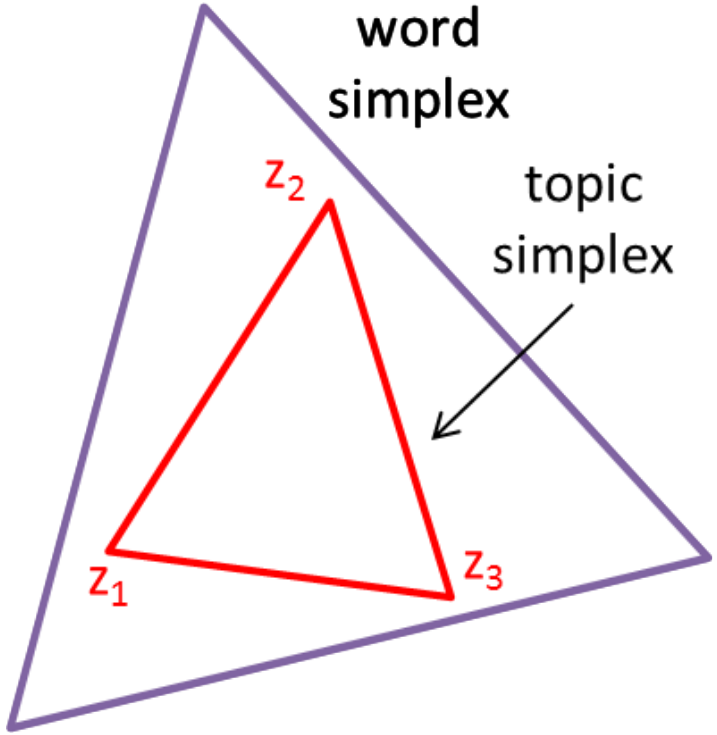
$$P(\theta|\alpha) = D(\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

  - Dirichlet parameter α_k is a prior count of the k-th topic
  - It controls the mean shape and sparsity of multinomial parameters θ



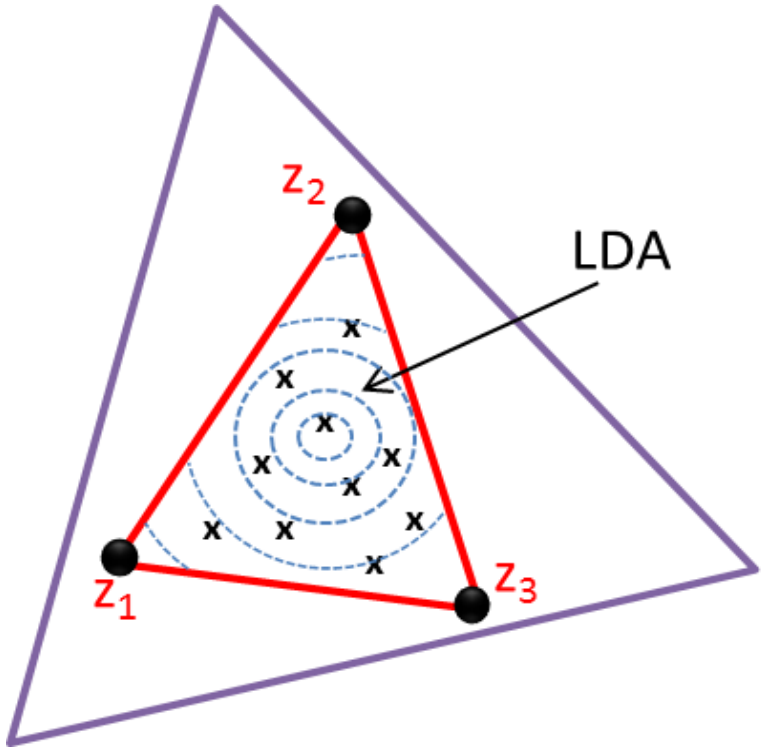
If the likelihood is multinomial with a Dirichlet prior $P(\theta|\alpha) = D(\alpha)$ then posterior is Dirichlet $P(\theta|\alpha, \boldsymbol{w}) = D(\alpha + n)$, where n is a vector of K entries counting how many samples have category k

Geometric Interpretation



LDA finds a set of K projection functions on the K-dimensional topic simplex

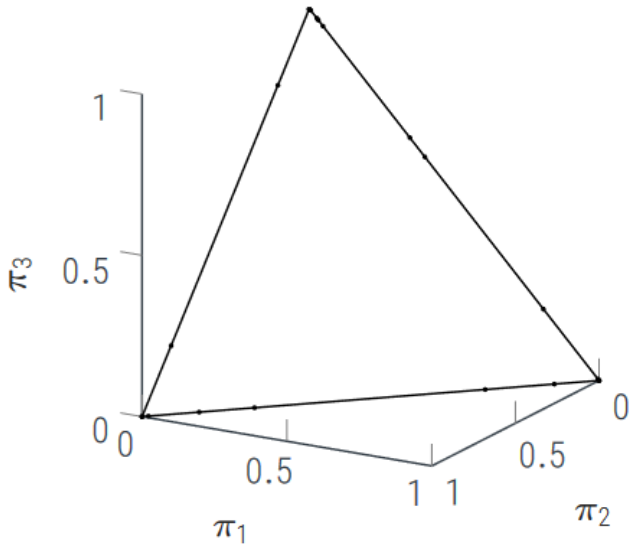
Geometric Interpretation

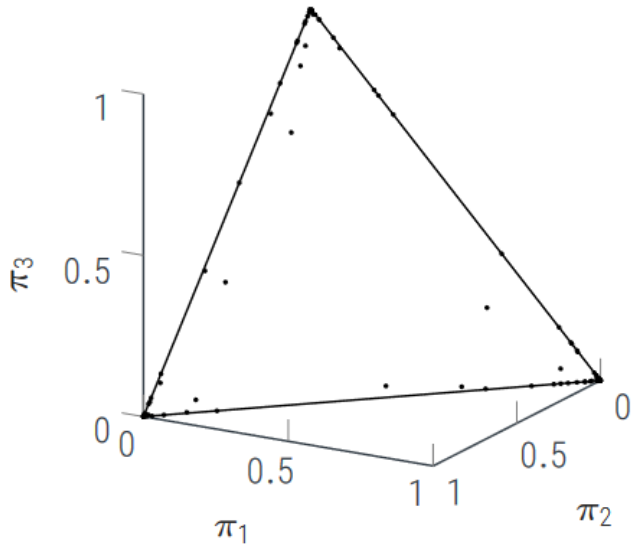


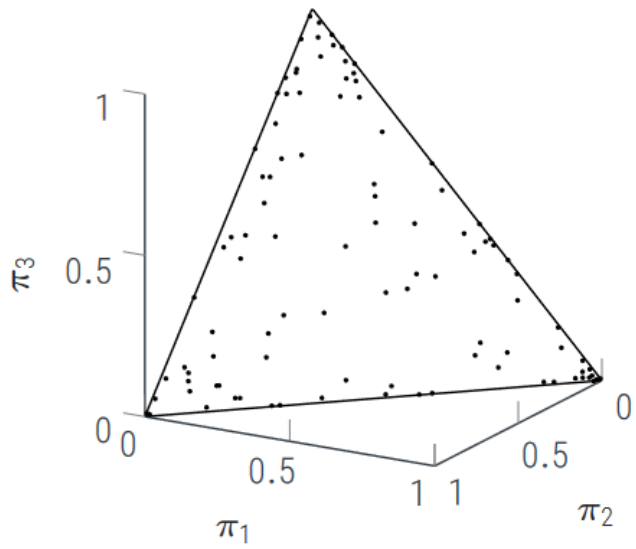
LDA finds a set of K projection functions on the K-dimensional topic simplex

Dirichlet can encode sparsity!

For $\alpha \sim 0.01, 0.1, 0.5$ (100 samples each)

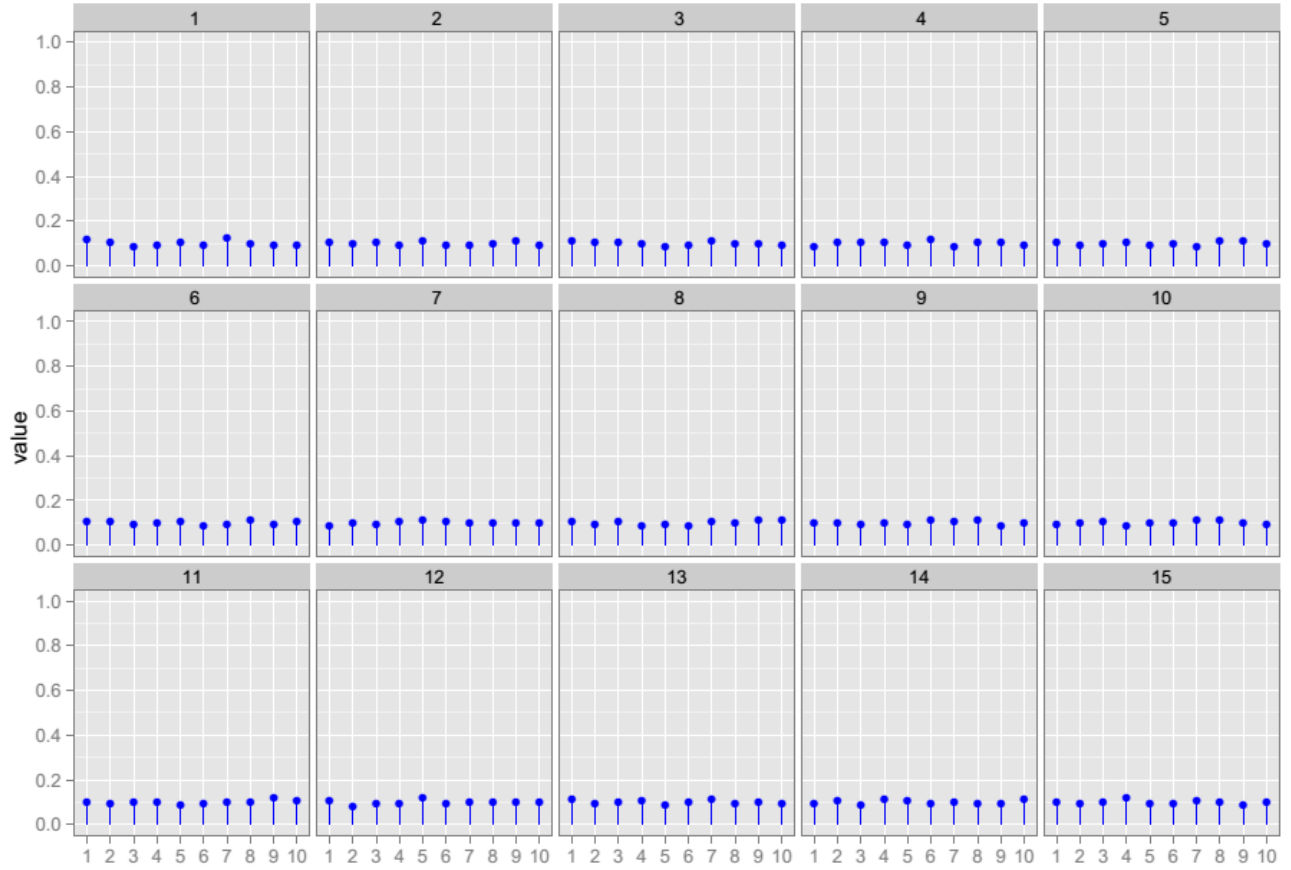






Effect of the α parameter

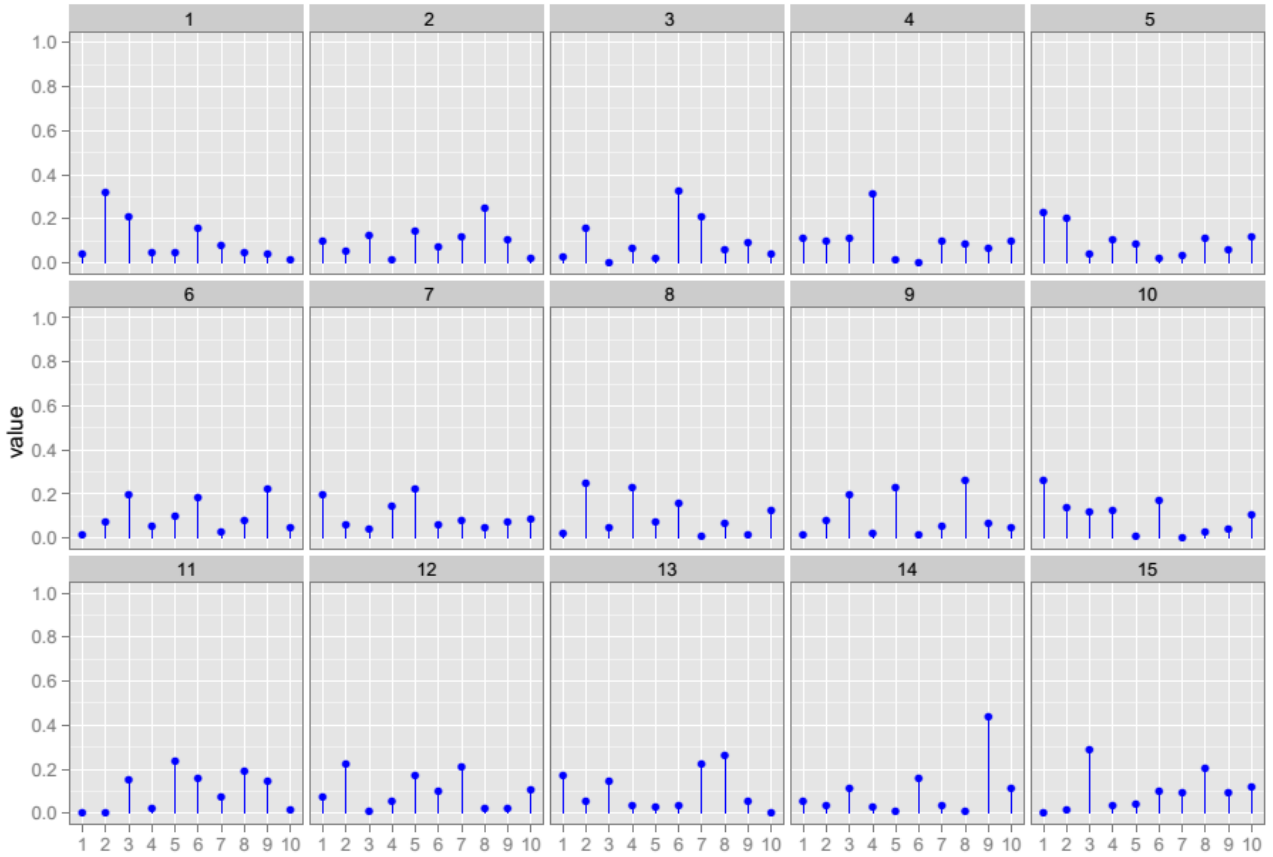
$\alpha = 100$



Slide Credit - Blei at KDD 2011 Tutorial

Effect of the α parameter

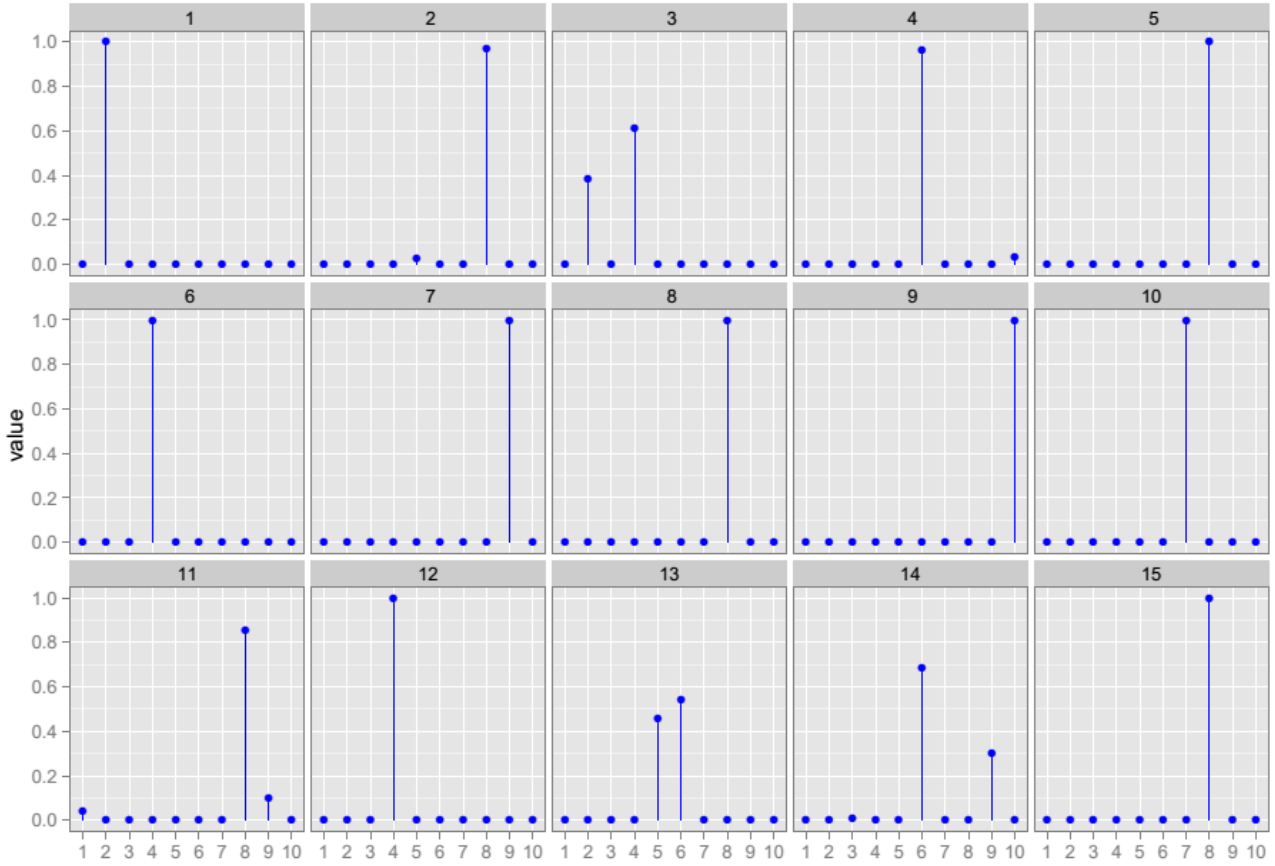
$\alpha = 1$



Slide Credit - Blei at KDD 2011 Tutorial

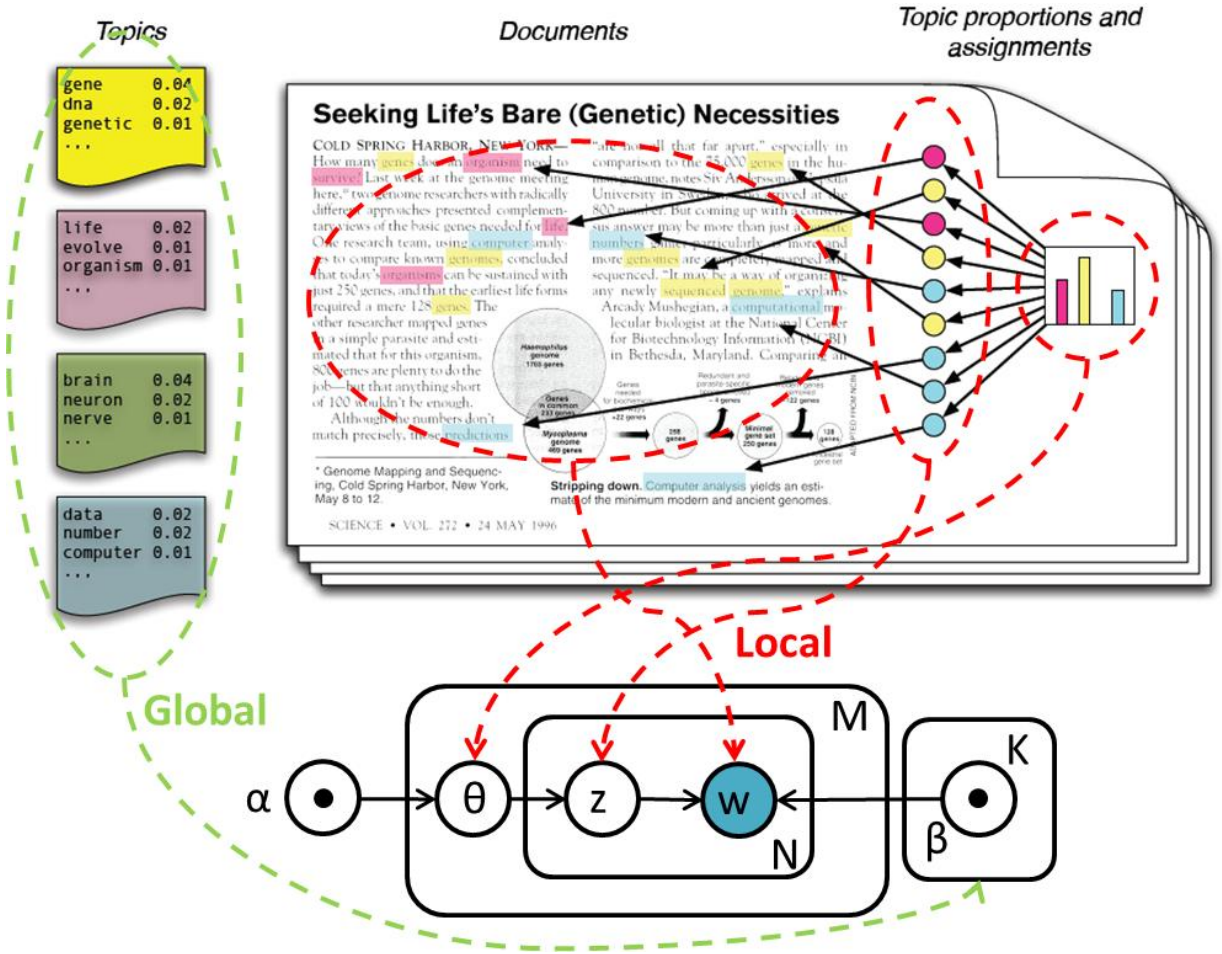
Effect of the α parameter

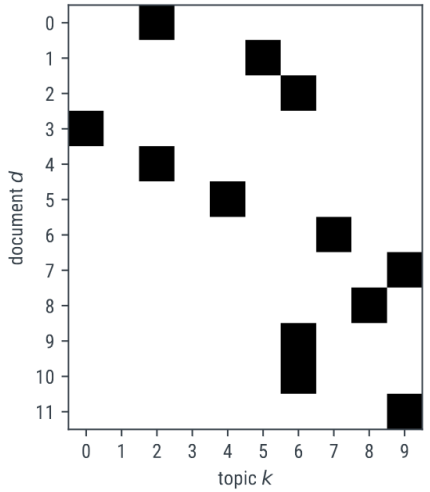
$\alpha = 0.01$



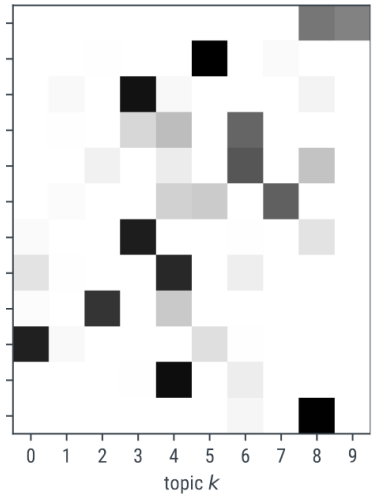
Slide Credit - Blei at KDD 2011 Tutorial

LDA and Text Analysis



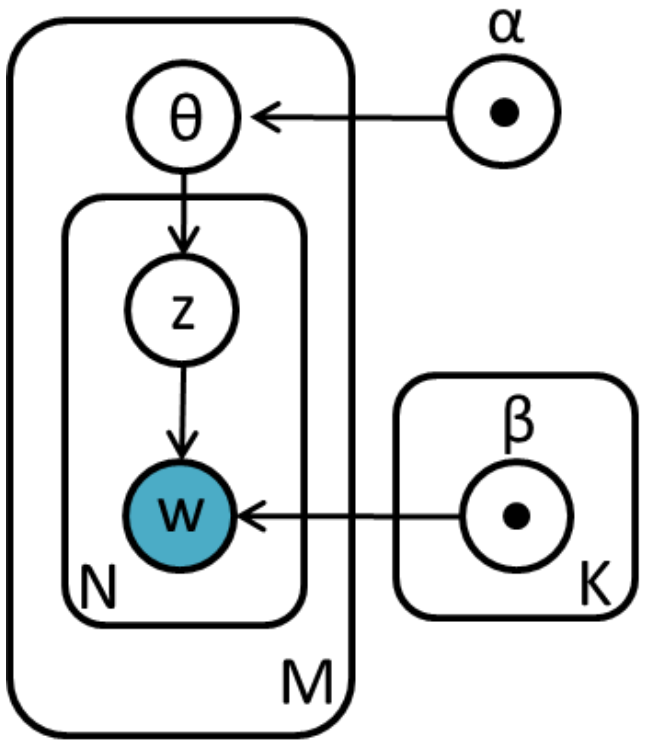


Standard mixture model



LDA

LDA Generative Process



For each of the M documents

- Choose $\theta \sim$ Dirichlet(α)
- For each of the N items
  - Choose a topic $z \sim$ Multinomial(θ)
  - Pick an item w_j with multinomial probability $P(w_j|z, \beta)$

Multinomial topic-item parameter matrix $[\beta]_{K \times N}$

$$\beta_{kj} = P(w_j = 1|z_k = 1) \text{ or } P(w_j = 1|z = k)$$

$$P(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = P(\theta|\alpha) \prod_{j=1}^{N} P(z_j|\theta) P(w_j\ |z_j\ , \beta)$$

LDA Variational Learning

Learning in LDA

Marginal distribution (a.k.a. likelihood) of a document $d = \boldsymbol{w}$

$$P(\boldsymbol{w}|\alpha,\beta) = \int \sum_{\boldsymbol{z}} P(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha,\beta)d\theta = \int P(\theta|\alpha) \prod_{j=1}^{N} \sum_{z_j=1}^{k} P(z_j|\theta)P(w_j|z_j,\beta)d\theta$$

Given $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$, find α, β (or only β, with α hyperparameter) maximizing

$$\mathcal{L}(\alpha,\beta) = \log \prod_{i=1}^{M} P(\boldsymbol{w}_i|\alpha,\beta)$$

Learning with hidden variables $\Longrightarrow$ Expectation-Maximization

Key problem is inferring latent variables posterior

Posterior Inference

- Optimal ELBO is achieved when Q(z) is equal to the latent variable posterior

$$P(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{P(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{P(\boldsymbol{w}|\alpha, \beta)}$$

- Key problem is that computation of the posterior is not tractable due to the denominator having couplings between β and θ in the summation over topics

$$P(\boldsymbol{w}|\alpha, \beta) = \int P(\theta|\alpha) \prod_{j=1}^{N} \sum_{k=1}^{K} P(z_j = k|\theta) P(w_j\,|z_j = k, \beta)\, d\theta = \cdots$$

$$\cdots = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \left(\prod_{j=1}^{N} \sum_{k=1}^{K} \prod_{v=1}^{V} (\theta_k \beta_{kv})^{w_j^v} \right) d\theta$$

Approximating Parameter Inference in LDA

The posterior would be much easier to compute if we could separate θ and β, i.e. compute one knowing the other (rings a bell?)

Variational Expectation Maximization

- Maximize the variational bound without using the optimal posterior solution
  - Write a $Q(\boldsymbol{z}|\phi)$ function that is sufficiently similar to the posterior but tractable
  - $Q(\boldsymbol{z}|\phi)$ should be such that β and θ are no longer coupled
  - Fit ϕ parameter to either maximize the bound or $Q(\boldsymbol{z}|\phi)$ closeness to posterior according to KL
- Variational LDA: Blei, Ng and Jordan, 2003

Variational Inference

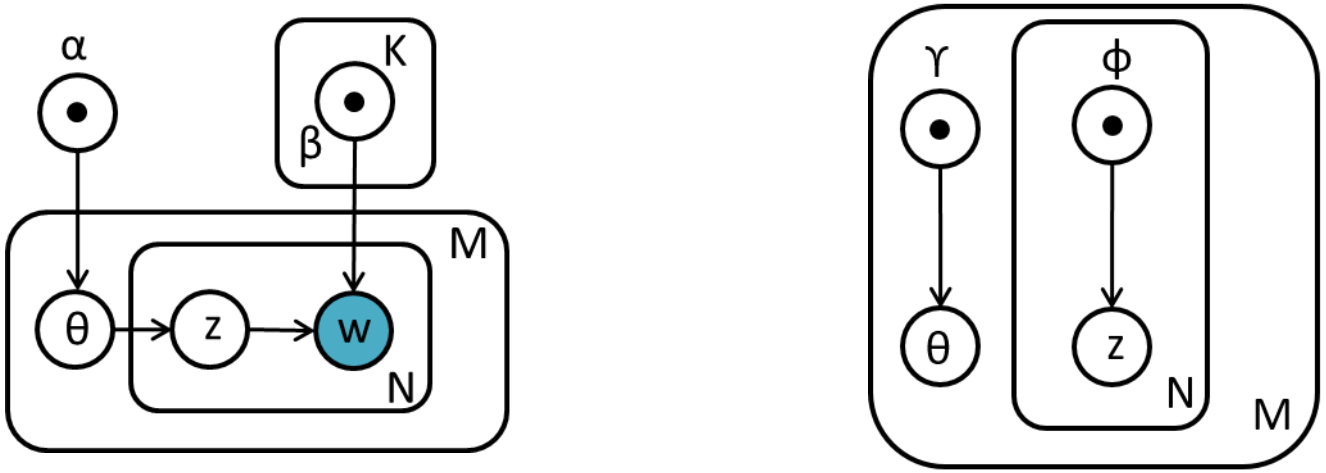
Key Idea

Assume that our distribution $Q(z|\phi)$ factorizes (it is tractable) $\rightarrow$ mean-field approximation

$$Q(\boldsymbol{z}|\phi) = Q(z_1, \ldots, z_N|\phi) = \prod_{j=1}^{N} Q(z_j\ |\phi_j)$$

- Can be made more general by factorizing on groups of latent variables
- Does not contain the true posterior because hidden variables are dependent
- Variational inference
  - Optimize ELBO using $Q(\boldsymbol{z}|\phi)$ factorized distribution
  - Coordinate ascent inference - Iteratively optimize each variational distribution holding the others fixed

Variational LDA Distribution



Given $\Phi = \{\gamma, \phi\}$ as variational approximation parameters

$$Q(\theta, \boldsymbol{z}|\Phi) = Q(\theta|\gamma) \prod_{j=1}^{N} Q\left(z_j|\phi_j\right)$$

Then we have the model parameters β of sample distribution $P(\theta, \boldsymbol{z}, \boldsymbol{w}|\ \beta)$

Variational Expectation-Maximization

Find the Φ, Ψ that maximize the ELBO

$$\mathcal{L}(\boldsymbol{w}, \Phi, \beta) = \mathbb{E}_Q[\log P(\theta, \boldsymbol{z}, \boldsymbol{w}|\beta)] - \mathbb{E}_Q[\log Q(\theta, \boldsymbol{z}|\Phi)]$$

by alternate maximization

1. **repeat**
2. Fix β: update variational parameters Φ^* (E-STEP)
3. Fix $\Phi = \Phi^*$: update model parameters β^* (M-STEP)
4. **until** little likelihood improvement

ELBO simplifies substantially thanks to mean field assumption

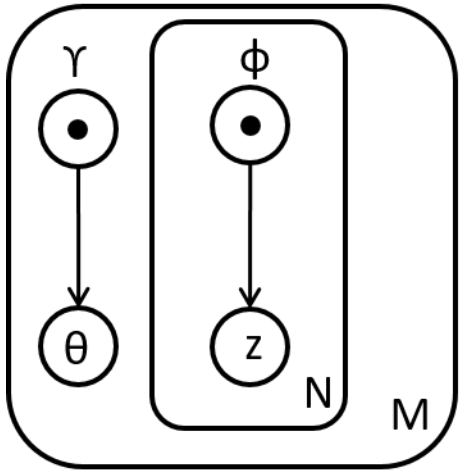
$$\mathcal{L}(\boldsymbol{w}, \Phi, \beta) = \mathbb{E}_{\theta \sim Q(\theta|\gamma)}\left[\log \frac{P(\theta, \boldsymbol{z}, \boldsymbol{w}|\beta)}{Q(\theta|\gamma)}\right] + \sum_{j=1}^{N} \mathbb{E}_{z_j \sim Q(z_j|\phi_j)}\left[\log \frac{P(\theta, \boldsymbol{z}, \boldsymbol{w}|\beta)}{Q(z_j|\phi_j)}\right]$$

The LDA Learning Algorithm

(A scary bunch of slides)

...not part of the oral exam though!

Preliminaries



Notation

- K: topics, N: vocabulary size, M: documents
- Document d: word counts n_{dv} for term $v \in \{1, \dots, N\}$
- Model parameters: multinomial topic–word distributions β_{kv}
  - As hyperparameters: $\alpha \in \mathbb{R}_+^K$ (Dirichlet prior on topic proportions)
- Variational parameters (per doc d):
  - $\gamma_d \in \mathbb{R}_+^K$ for $q(\theta_d) = \mathcal{D}ir(\gamma_d)$
  - $\phi_{dv} \in \Delta^{K-1}$ for $q(z_{d,v}) = Cat(\phi_{dv})$ (topic assignment for word-type v, weighted by counts)

Optimization problems

- E-STEP - $(\gamma^*, \phi^*) = \arg\min KL(Q(\theta, \boldsymbol{z}|\gamma, \phi)||P(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta))$ (other side of ELBO)
- M-STEP - $\beta^* = \arg\max_{\beta} \mathcal{L}(\boldsymbol{w}, \Phi, \beta)$ (classic ELBO)

E-STEP

Solving the $\arg\min KL()$ optimization problem by iterative fixed point yields

- Topic proportions (in log) update:

$$\mathbb{E}_q[\log \theta_{dk} | \gamma] = \Psi(\gamma_{dk}) - \Psi\left(\sum_{j=1}^{K} \gamma_{dj}\right)$$

Ψ is the first derivative of the log Γ function which is computable via Taylor approximations

- Topic responsibilities update:

$$\phi_{dvk} \propto \beta_{kv} \exp\left(\mathbb{E}_q[\log \theta_{dk} | \gamma]\right)$$

Normalize to sum-to-1 w.r.t k

- Dirichlet parameters update:

$P(w_d = v | z_d = k)$ $P(z=k)$

Variational Dirichlet posterior $P(z_d = k | w_d = v)$

$$\gamma_{dk} = \alpha_k + \sum_{v=1}^{V} n_{dv} \phi_{dvk}$$

Dirichlet update

Iterate until convergence (or for a small fixed number of iterations)

M-STEP

Maximize the ELBO (below, per single document)

$$\mathbb{E}_Q[\log P(\theta_d|\alpha)] + \sum_n \mathbb{E}_Q[\log P(z_{dn}|\theta_d)] + \sum_n \mathbb{E}_Q[\log P(w_{dn}|z_{dn},\beta)] - \log\Gamma\left(\sum_k \gamma_{dk}\right) + \sum_k \log\Gamma(\gamma_{dk}) - \sum_k (\gamma_{dk}-1)\,\mathbb{E}_Q[\log\theta_{dk}|\gamma]$$

- Expected topic–word counts:

$$\hat{n}_{kv} = \sum_{d=1}^{D} n_{dv}\ \phi_{dvk}$$

- Update topics (multinomial parameters):

$$\beta_{kv} = \frac{\hat{n}_{kv}}{\sum_{v'=1}^{V} \hat{n}_{kv'}}$$

- Also α can be updated, but optimization is a bit trickier and needs Newton-Rhapson

LDA Training Pseudocode

Input: corpus counts $\{n_{dv}\}$, topics K, init α, init β
repeat until *likelihood_convergence* {
 # E-step: per-document variational inference
 foreach document d = 1..D:
  initialize $\gamma_d = \alpha + (N_d/K) \cdot 1$, initialize φ_{dv} uniform over K
  repeat until variational_convergence:
   foreach word-type v with $n_{dv} > 0$:
   {
    $\varphi_{dvk} \propto \beta_{kv} \cdot \exp(\psi(\gamma_{dk}) - \psi(\sum_j \gamma_{dj}))$ for k=1..K
    normalize φ_{dv} over k
   for k=1..K:
    $\gamma_{dk} = \alpha_k + \sum_v n_{dv} \cdot \varphi_{dvk}$
   }
...

LDA Training Pseudocode

...

M-step: update topics β using expected counts under q

For k = 1..K, v = 1..V:

$$\hat{n}_{kv} = \sum_d n_{dv} \cdot \varphi_{dvk}$$

For k = 1..K:

$$\beta_{kv} = \frac{\hat{n}_{kv}}{\sum_{v'} \hat{n}_{kv'}}$$

Hyperparameter update (optional but in the paper)

Update α via Newton–Raphson using gradient g and Hessian H (digamma/trigamma)

}

Output: α, β and variational params $\{\gamma_d, \varphi_d\}$

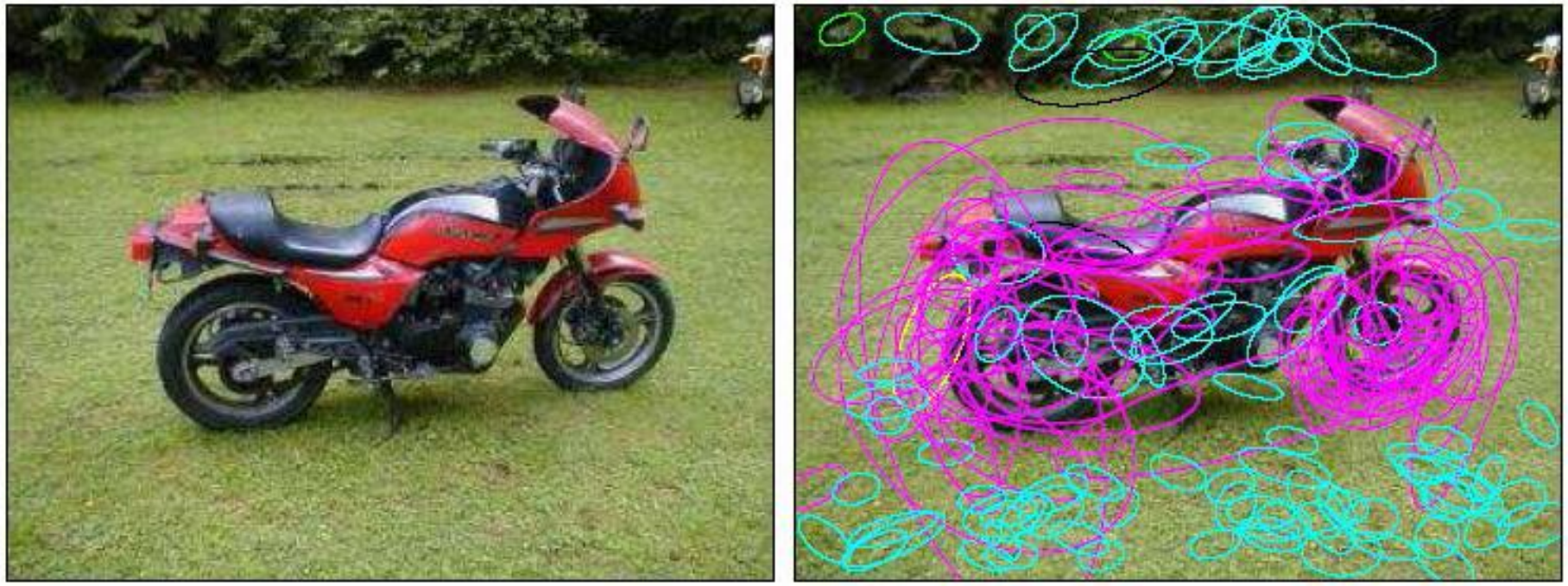
Wrap-up

LDA Applications

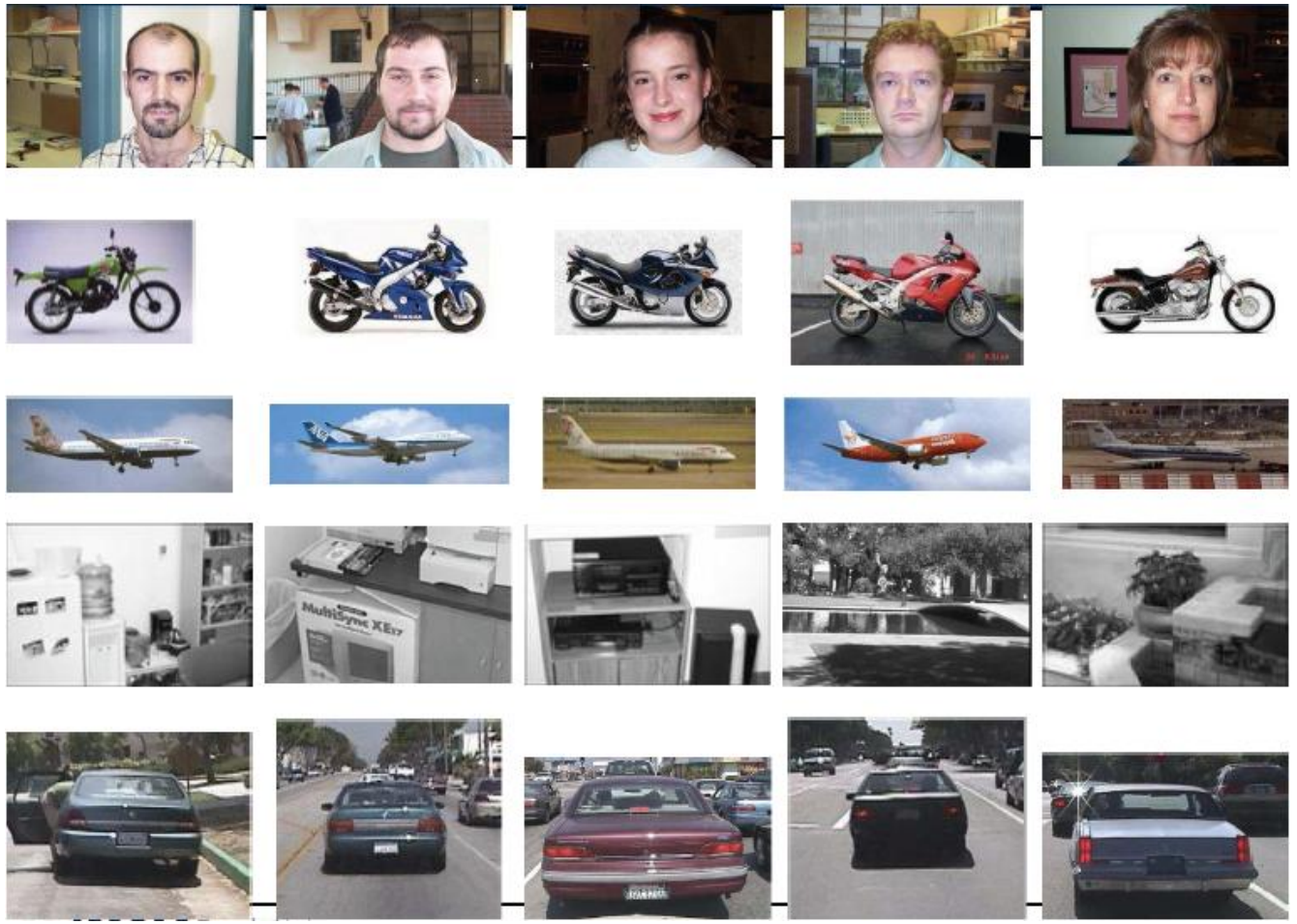
- Why using latent topic models?
- Organize large collections of documents by identifying shared topics
- Understanding the documents semantics (unsupervised)
- Documents are of different nature
  - Text
  - Images
  - Video
  - Relational data (graphs, time-series, etc..)
- In short: a model for collections of high-dimensional vectors whose attributes are multinomial distributions

LDA for Image Understanding

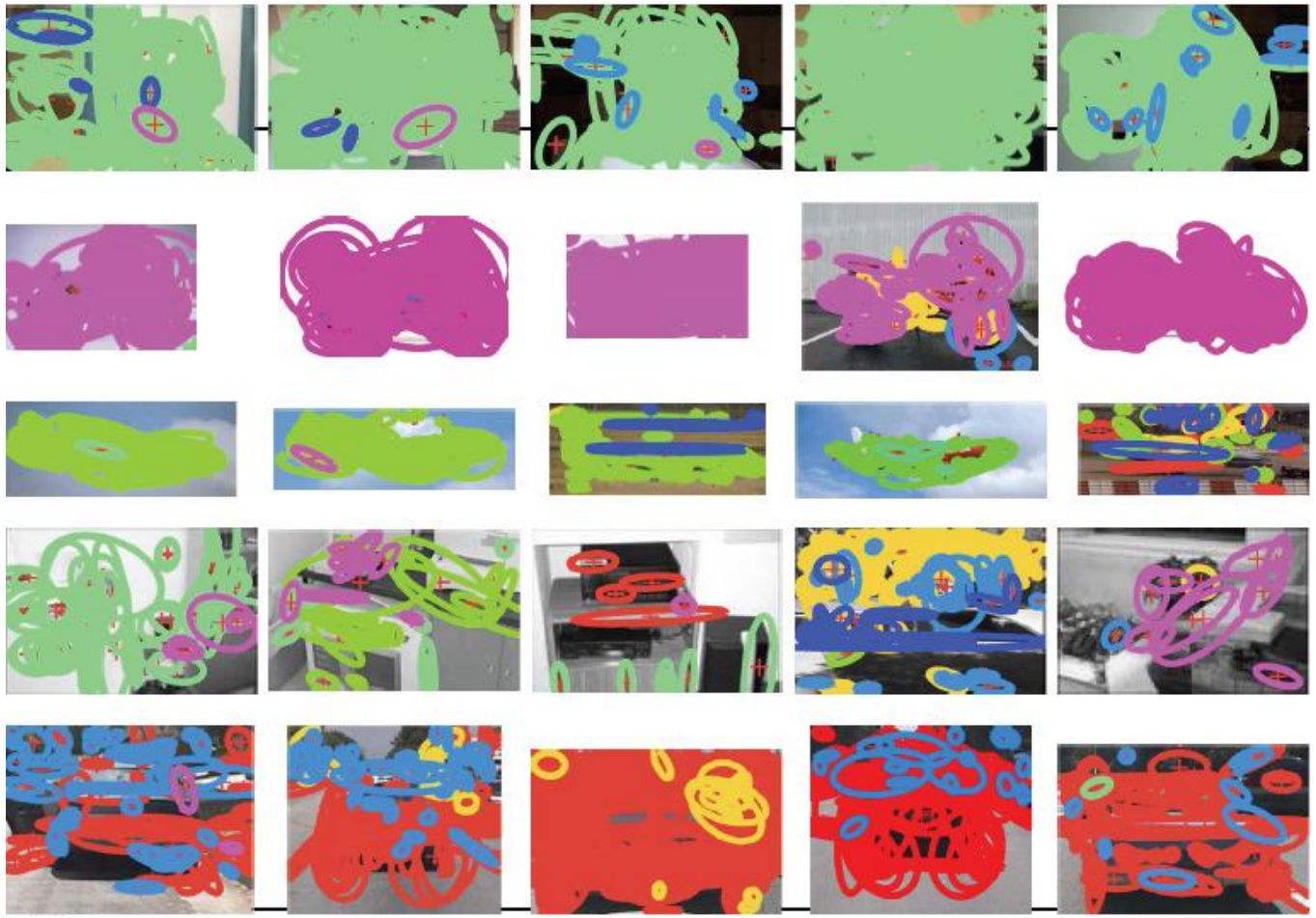
Assigning a topic to visual words (pixel patches extracted through some visual detector+descriptor pipeline)



LDA for Image Understanding



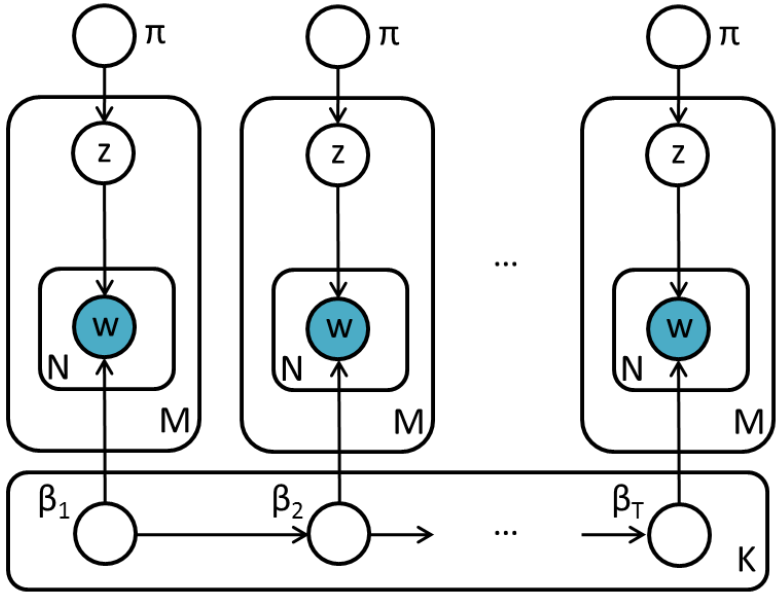
LDA for Image Understanding



Dynamical Topic Models

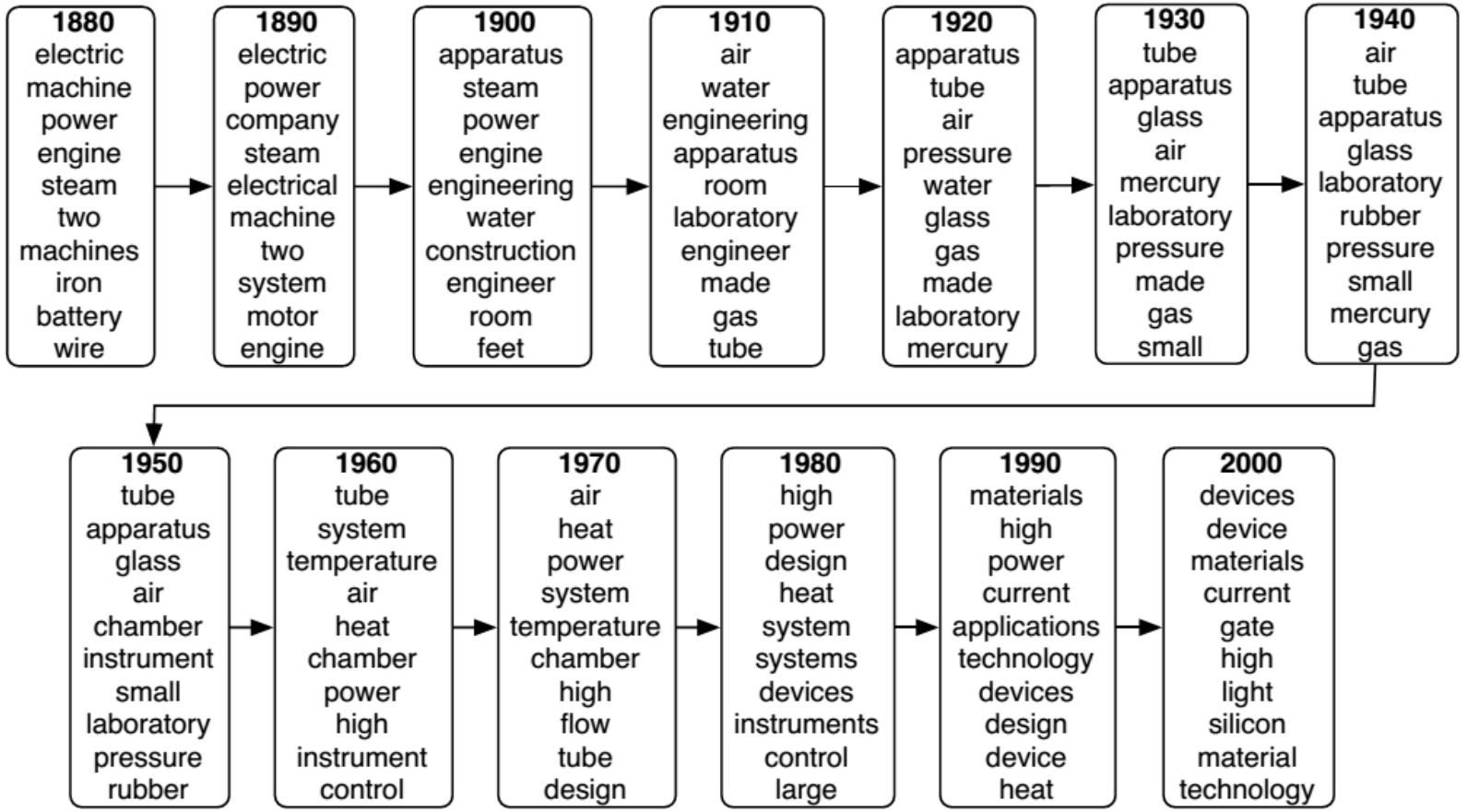
LDA assumes that the document order does not count

- What if we want to track topic evolution over time?
- Tracking how language changes over time
- Videos are image documents over time



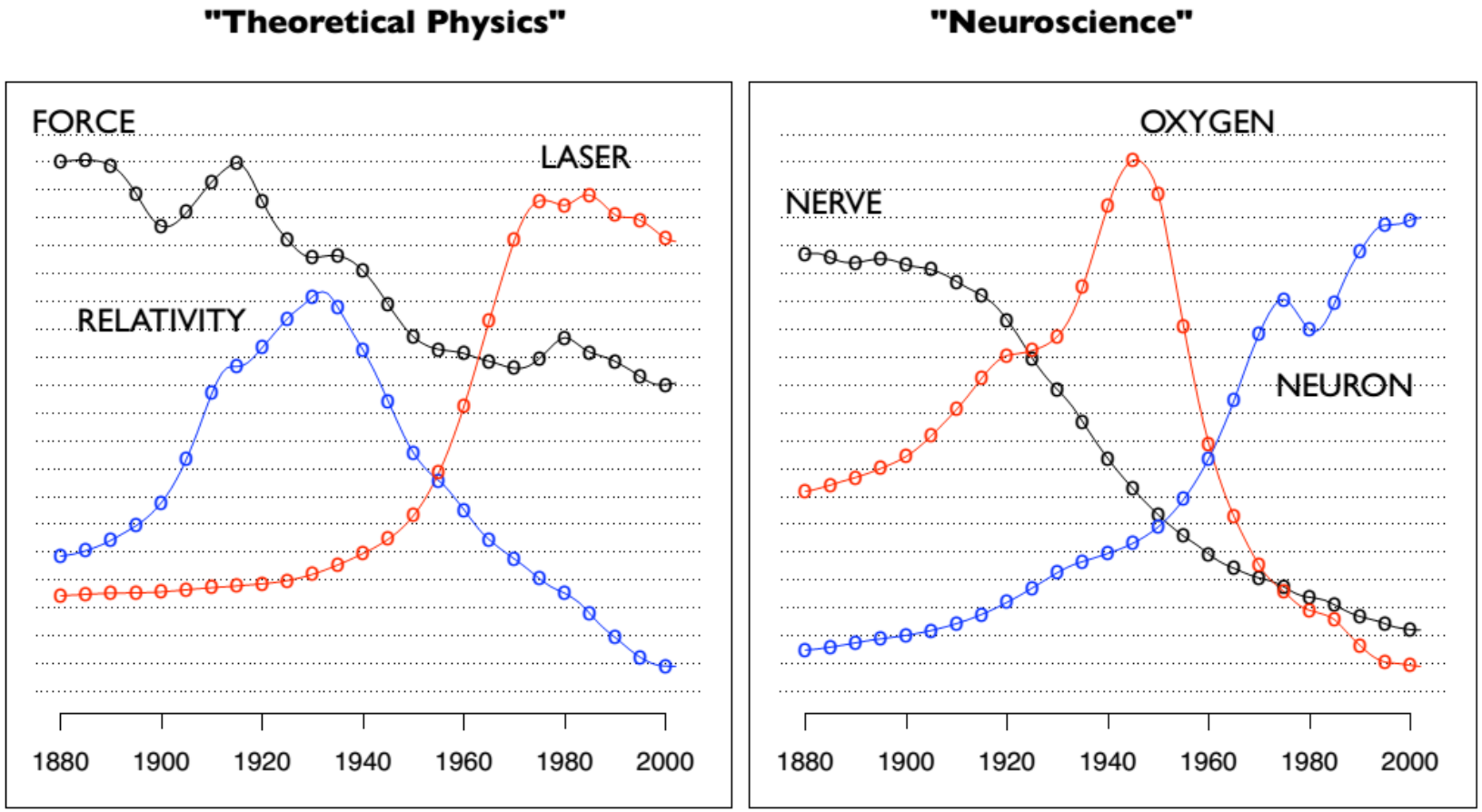
Blei and Lafferty. Dynamic topic models, ICML 2006

Topic Evolution over Time



https://github.com/blei-lab/dtm

Topic Trends



https://github.com/blei-lab/dtm

Variational Learning in Code

- PyMC3 - Python library with particular focus on variational algorithms (not PyMC!)
- Edward - Python library with lots of variational inference from the father of LDA
- Bayespy - Variational Bayesian inference for conjugate-exponential family only
- LDA is implemented in many Python libraries: scikit-learn, pypi, gensim (efficient topic models)

Take Home Messages

- Latent topic models give a specific semantics to the hidden variables
  - Topics capture meaning in documents made of collections of items (words, visterms, nodes)
  - When unsupervised, these are again clusters
- Latent Dirichlet Allocation
  - Latent topic model to organize collections of multinomial data
  - Show an example of probabilities becoming random variables (i.e. drawn from another distribution)
  - Learning by variational EM, using KL() minimization
- Bayesian latent topic models allow hierarchical topics, infinite topics, …

Next Lecture

Sampling Methods

- Introduction to sampling methods
- Ancestral sampling
- Gibbs Sampling
- Approximated parameter learning via sampling
- Worked out example on LDA