

# Probabilistic Topic Models

COS 424/524, SML 302: Fundamentals of Machine Learning  
Professor Engelhardt

COS424/524, SML302

Lecture 18

# Latent Dirichlet allocation and topic models

In previous lectures, we learned two ways to perform exploratory data analysis using dimension reduction:

- principal component analysis (PCA);
- factor analysis (FA).

Today we will learn a third canonical method for performing dimension reduction: latent Dirichlet allocation (LDA).

# Latent Dirichlet allocation and topic models

LDA is different from PCA and FA in two important ways:

- the observations are multinomial (or not necessarily Gaussian)
- each observation from a sample is assigned to one latent component

# What can we do with exploratory data analysis?



[www.betaversion.org/~stefano/linotype/news/26/](http://www.betaversion.org/~stefano/linotype/news/26/)

Exploratory data analysis helps us:

- **organize**
- **visualize**
- **summarize**
- **search**
- **predict**
- **understand**
- **... data.**

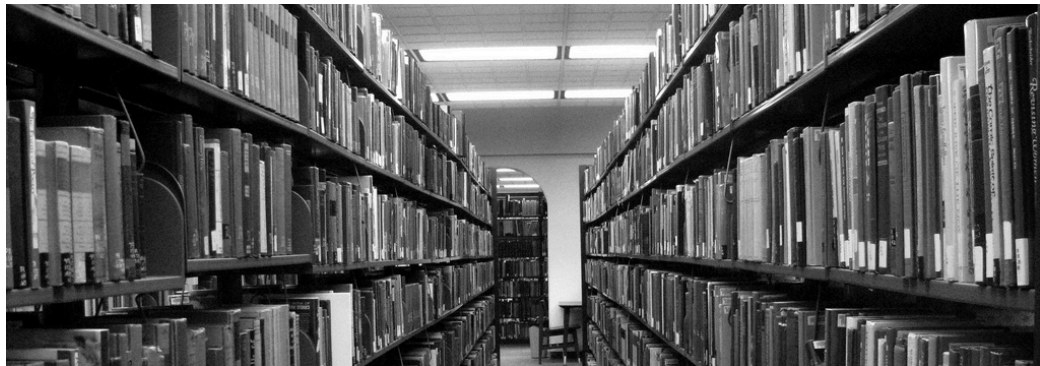
# Why do we need to perform dimension reduction on data?



Candida Hofer

- As more information becomes available, it becomes more difficult to search for something specific.
- We need tools to help us sift through these vast amounts of information.

# Probabilistic Topic Modeling



Topic models automatically organize, search, & summarize large electronic archives:

- Recover latent topic patterns across the collection of samples;
- Annotate samples according to these topics;
- Use annotations to organize, summarize, and search the samples.

# Examples of topic model applications

## Topic model applications

- Associated press news articles (topics: elections, sports, education)
- Digital humanities projects: million book Classical texts
- Scientific articles (topics: quantum mechanics, molecular evolution)
- Yale law journal (topics: judicial, employment, tax law)
- Machine learning papers (topics: neural networks, clustering)
- Recipe collections (topics: easy dinners, desserts)
- Political bloggers (topics: abortion, guns, free speech)
- Video game joystick moves (topics: sea world, jumping world)

# Example: topic model applied to a collection of documents

## Topic model

**Input:** An unorganized collection of documents

**Output:** A set of *topics* across the documents; for each document, the proportion of that document related to each topic.

In more detail:

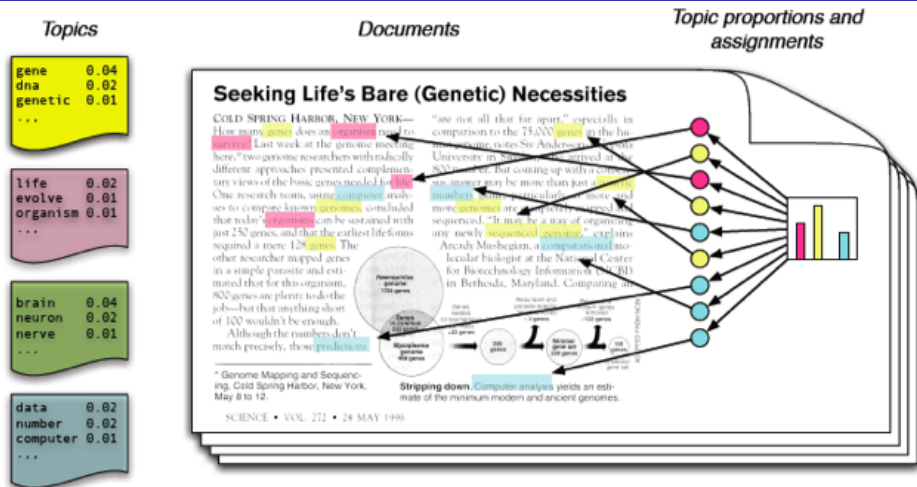
- We will represent each document as a *bag-of-words*;
- Each of  $K$  *topics* are multinomial distributions over the vocabulary;
- The *topic specific proportions* capture the proportion of each document generated from the  $k$ th topic.



# We will use a probabilistic model to address this problem

- We will treat each observation (word) from each sample (document) as a draw from an independent topic distribution with its own topic  $k$
- We will infer the latent structure using expectation maximization (or other methods)
- We can then classify, search, or predict new samples by determining how the new sample reflects the inferred topics.

# Admixture model: a document has words from many topics

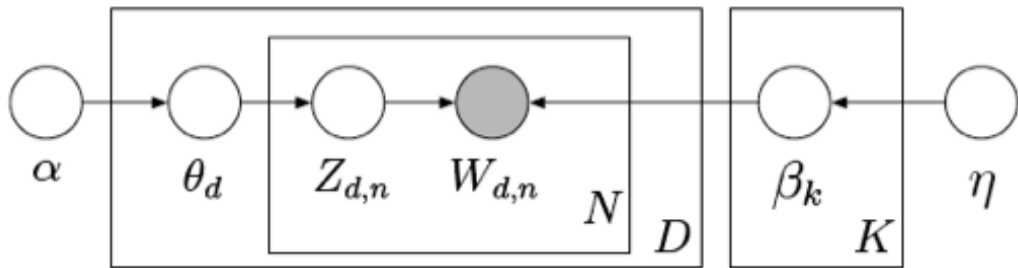


- Each document is an *admixture* of some corpus-wide topics
- Each word is drawn from exactly one of these topics

# We label a topic via the highest probability words

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

# Probabilistic model for LDA



For  $n = 1 : N$  words,  $d = 1 : D$  documents,  $k = 1 : K$  topics:

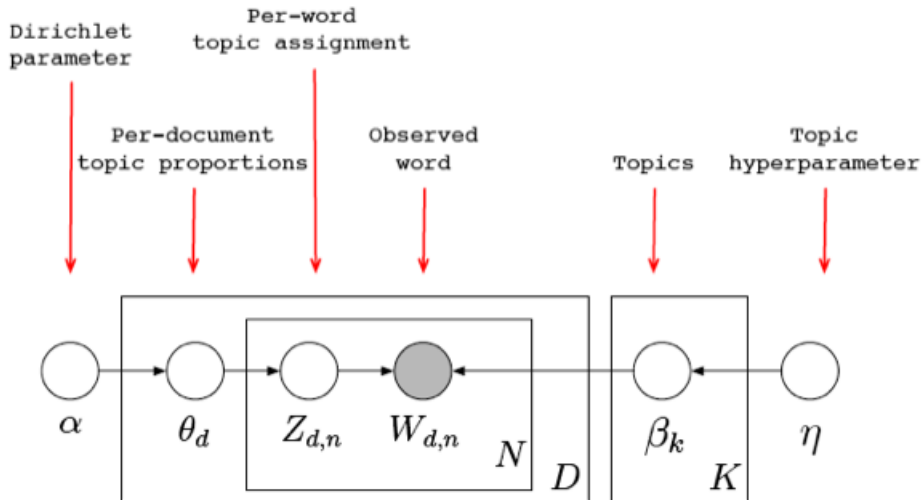
$$\theta_d | \alpha \sim \text{Dir}(\alpha)$$

$$\beta_k | \eta \sim \text{Dir}(\eta)$$

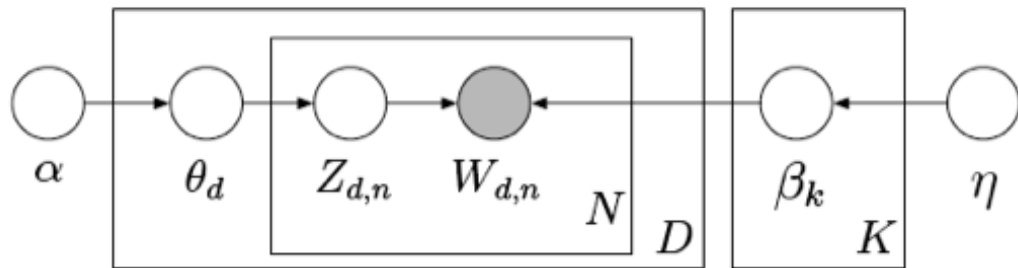
$$z_{d,n} | \theta_d \sim \text{Mult}(\theta_d)$$

$$w_{d,n} | \beta, z_{d,n} \sim \text{Mult}(\beta_{z_{d,n}}).$$

# Probabilistic model for LDA



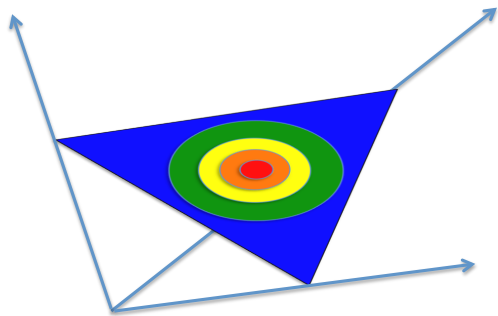
# Generative model for LDA



- 1 For  $k = 1 : K$ : draw  $\beta_z \sim \text{Dirichlet}(\eta)$ .
- 2 For  $d = 1 : D$ : draw  $\theta_d \sim \text{Dirichlet}(\alpha)$
- 3 For  $d = 1 : D, n = 1 : N_d$ : draw  $z_{d,n} \sim \text{Mult}(\theta_d)$
- 4 For  $d = 1 : D, n = 1 : N_d$ : draw  $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$

What are the latent variables? What are the parameters?

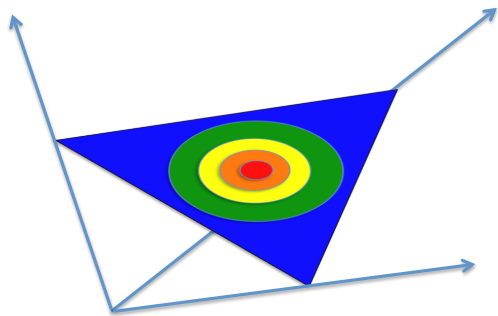
# Dirichlet distribution



- The Dirichlet distribution is an exponential family distribution over the *simplex*: non-negative vectors that sum to one

$$\theta_k \geq 0; \quad \sum_{k=1}^K \theta_k = 1$$

# Dirichlet distribution

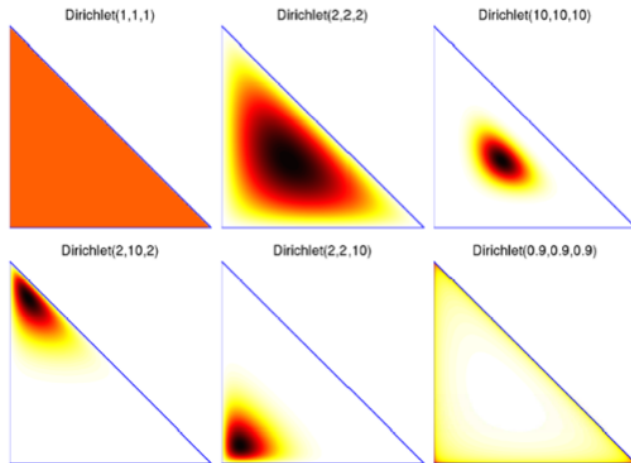


- The Dirichlet is conjugate to the multinomial; the posterior distribution is Dirichlet
- Parameter  $\alpha$  controls shape of the distribution on the simplex
- Topic proportions are a  $K$  dimensional Dirichlet; topics are a  $V$  dimensional Dirichlet.

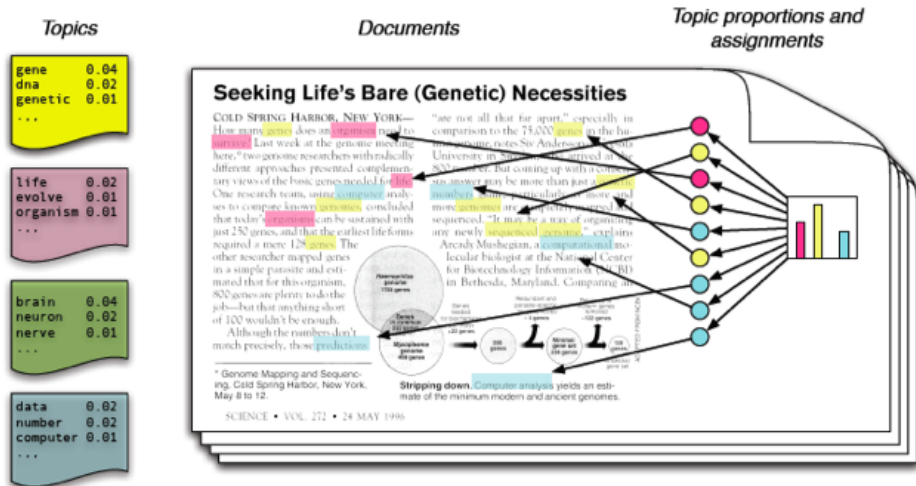


# Dirichlet distribution: examples

Plot three dimensional Dirichlet distributions in 2D (because third dimension  $\theta_3 = 1 - \theta_1 - \theta_2$ ):

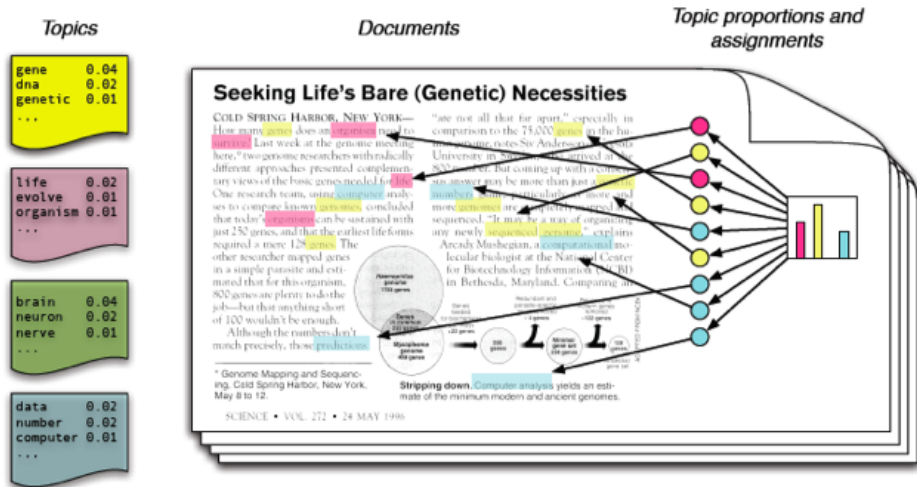


# Interpretation of LDA



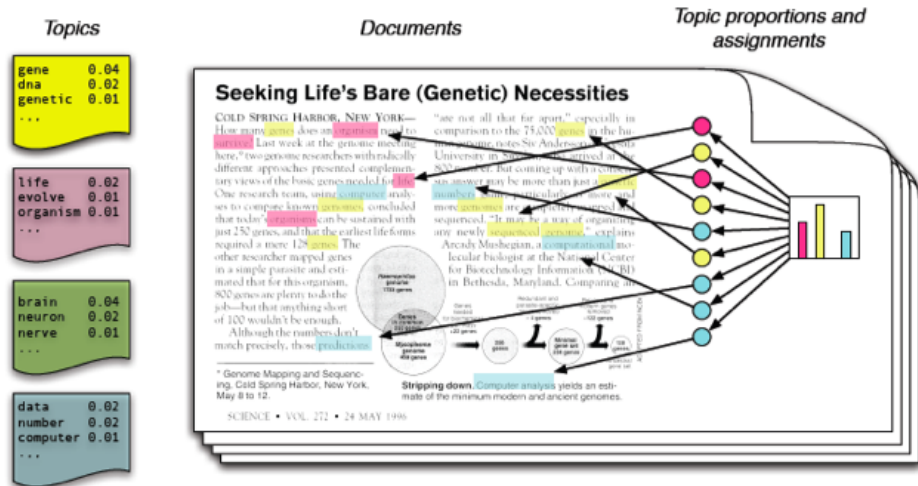
LDA assigns a word in a document to a topic according to document specific topic proportion, draws from word distribution of that topic

# Interpretation of LDA



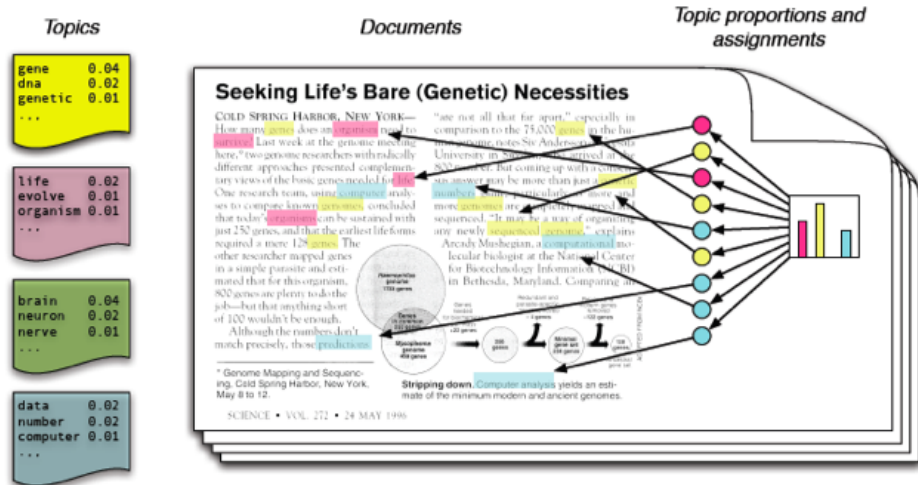
In latent variable  $z_{d,n}$ , each word in each document is assigned to a topic.

# Interpretation of LDA



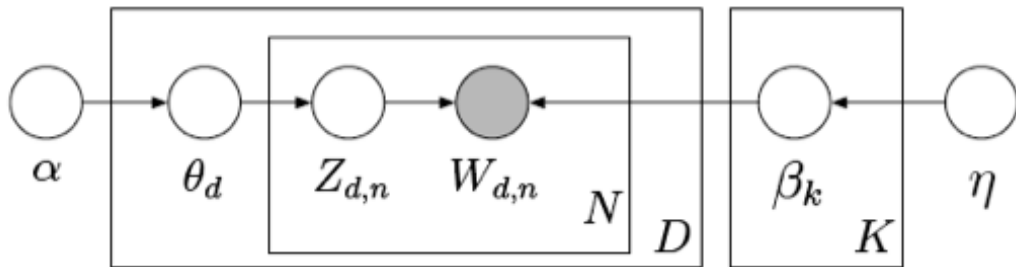
Latent variable  $\theta_d$  is the document-specific topic proportions.

# Interpretation of LDA



Parameter  $\beta_k$  is the topic-specific word proportions over the vocabulary.

# Interpretation of LDA

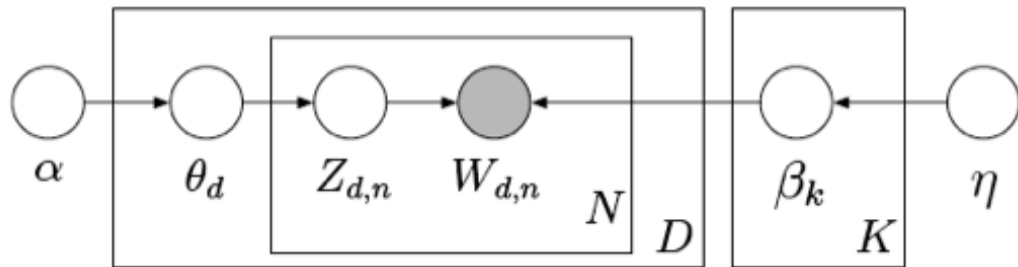


From a collection of documents, infer:

- Per-word topic assignment  $z_{d,n}$
- Per-document topic proportions  $\theta_d$
- Per-corpus topic distributions  $\beta_k$

Use posterior expectations to perform, e.g., topic-based search, document comparisons, document assignments, etc.

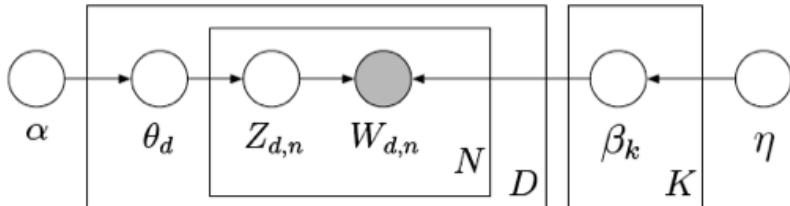
# Model of LDA



We can write out the per-document posterior distribution for LDA:

$$\begin{aligned} p(w_{d,1:N}, z_{d,1:N}, \theta_d \mid \alpha, \beta) &= p(\theta_d \mid \alpha) p(z_{d,1:N} \mid \theta_d) p(w_{d,1:N} \mid \beta, z_{d,1:N}) \\ &= p(\theta_d \mid \alpha) \prod_{n=1}^N p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}). \end{aligned}$$

# Why does LDA “work”?



## LDA trades off two goals

- 1 In each **document**, assign words to a **few high probability topics**.
- 2 In each **topic**, assign high probability to a **few terms**.

- We see this from the log posterior

$$\log p(x, z, \theta \mid \alpha, \beta) = \dots + \sum_d \sum_n \log p(z_{dn} \mid \theta_d) + \log p(w_{dn} \mid \beta_{z_{dn}}) + \dots$$

- Sparse proportions come from the 1st term.
- Sparse topics come from the 2nd term.



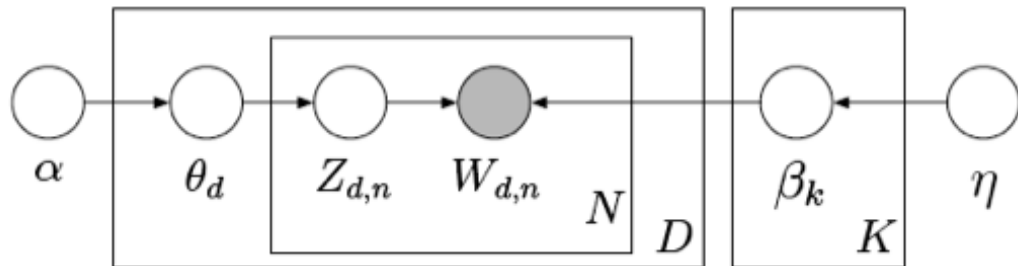
# Why does LDA “work”?

## LDA trades off two goals

- 1 In each **document**, assign words to a **few high probability topics**.
- 2 In each **topic**, assign high probability to a **few terms**.

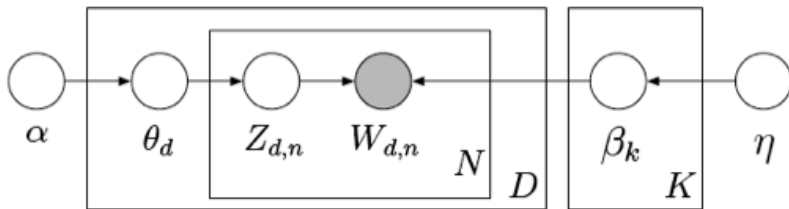
- These goals are at odds:
  - Putting a document in a single topic makes #2 hard.
  - Putting only a few words in each topic makes #1 hard.
- Trading off these goals results in recovering a few groups of commonly co-occurring words.

# Methods to fit LDA models



- This is a latent variable model, with latent parameters  $(z, \theta)$
- Will expectation-maximization work?

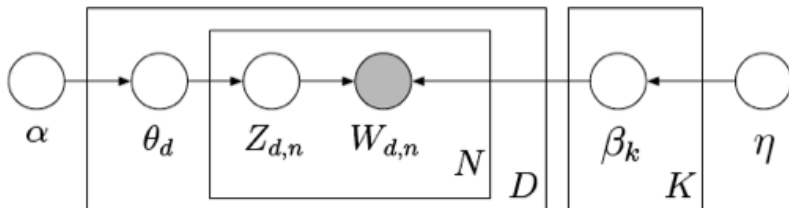
# Expectation-maximization for LDA models



- Write out complete log likelihood with respect to  $\Theta = \{\alpha, \beta\}$ :

$$\begin{aligned}\log p(w_d, z_d, \theta_d) &= \log p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta, z_{d,n}) \\ &= \log p(\theta_d | \alpha) + \sum_{n=1}^N \log p(z_{d,n} | \theta_d) + \log p(w_{d,n} | \beta, z_{d,n}) \\ &= \log p(\theta_d | \alpha) + \sum_{n=1}^N \sum_{k=1}^K z_{d,n}^k \log \theta_{d,k} + \sum_{v=1}^V \sum_{k=1}^K w_{d,n}^v z_{d,n}^k \log \beta_{v,k}\end{aligned}$$

# Expectation-maximization for LDA models

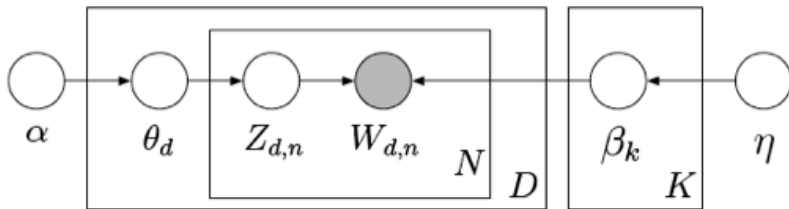


- Expected complete log likelihood:

$$\begin{aligned} \mathbb{E}[\log p(w_d, z_d \mid \Theta)] &= \mathbb{E}[\log p(\theta_d \mid \alpha)] + \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}[z_{d,n}^k \log \theta_{d,k}] \\ &\quad + \sum_{v=1}^V \sum_{k=1}^K w_{d,n}^v \mathbb{E}[z_{d,n}^k] \log \beta_{v,k} \end{aligned}$$

- The E-step expected sufficient statistics are not (generally) simple to compute. [Why not?](#)  
[Is there another way?](#)

# Posterior probability for LDA models



- Let's write out the posterior probability of the words, fixing  $\beta$ :

$$\begin{aligned}\log p(w_d | z_d, \theta_d, \Theta) &= \frac{p(z_d, \theta_d | \alpha) p(w_d | z_d, \theta_d, \beta)}{p(w_d | \beta, \alpha)} \\ &= \frac{p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta)}{\int_{\theta_d} p(\theta_d | \alpha) \prod_{n=1}^N \sum_{k=1}^K p(z_{d,n}^k | \theta_d) p(w_{d,n} | z_{d,n}^k, \beta_k) d\theta_d}.\end{aligned}$$

- The denominator is the sum of  $N^K$  tractable terms, which is intractable for most  $N$ .

# Approximate posterior inference algorithms to fit LDA

Two general approaches: sampling (MCMC) and variational inference.

- Mean field variational methods (Blei et al. 2003)
- Expectation propagation (Minka & Lafferty 2002)
- Collapsed Gibbs sampling (Griffiths & Steyvers 2004)
- Collapsed variation inference (Teh et al. 2006)
- Stochastic variational inference (Hoffman et al. 2013)
- Review and comparison: (Mukherjee & Blei 2009)
- *We will discuss these types of approaches in later lectures.*

# LDA as dimension reduction

How is LDA a probabilistic models for dimension reduction?

- First, integrate out latent variables  $Z$ ; not hard to do. Then:

$$w_d \mid \theta, \beta, \alpha, \eta \sim \text{Mult}(\theta_d^T \beta)$$

- In other words, each bag-of-words representation for document  $d$  is drawn from  $N$  rolls of a  $V$  dimensional die
- Each face of this die has a vocabulary word on it
- The probability of each face of the die is  $\theta_d^T \beta \in (0, 1)^V$ , which is on the simplex.

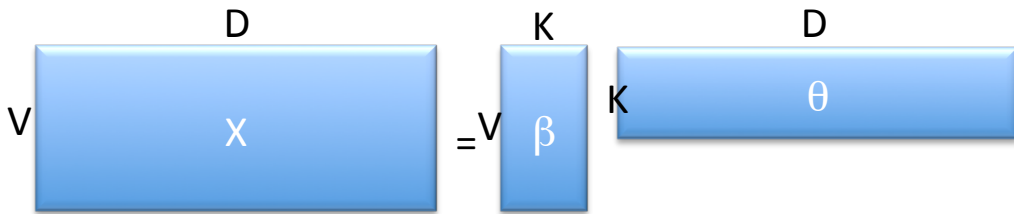
Exercise: prove this.

# LDA as dimension reduction

Marginal likelihood of data:

$$w_d \mid \theta, \beta, \alpha, \eta \sim \text{Mult}(\theta_d^T \beta)$$

- We can rewrite this likelihood in terms of a matrix factorization.
- Consider variables and parameters as matrices:  $W \in \mathcal{Z}^{D \times V}$ ,  $\theta \in (0, 1)^{D \times K}$ ,  $\beta \in (0, 1)^{K \times V}$





# Example of LDA: *Science* articles



- **Data:** The OCR'd collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model:** 100-topic LDA model using variational inference.

# Example of LDA: *Science* articles

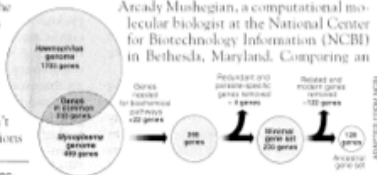
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

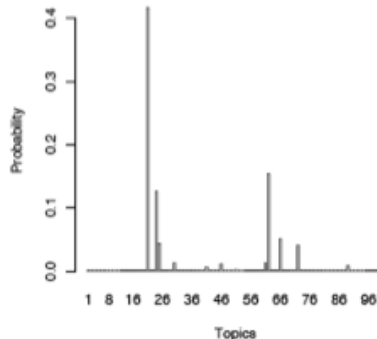
Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.



\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

We can look at the topics in an article to find the meaning of the article.

# Science article “Seeking Life’s Bare (Genetic) Necessities”

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

## Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (1) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (2) provide the most

convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the tell-tale signatures of chaos. In phase space, chaotic trajectories come to lie on “strange attractors,” curious geometric objects with fractal structure and hence noninteger dimension. As they

move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (3).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-



**Cannibalism and chaos.**

The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

SCIENCE • VOL. 275 • 17 JANUARY 1997

323

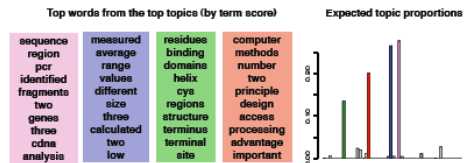
## Example of LDA: *Science* article “Chaotic Beetles”

problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

# Example of LDA: *Science* article

## Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel



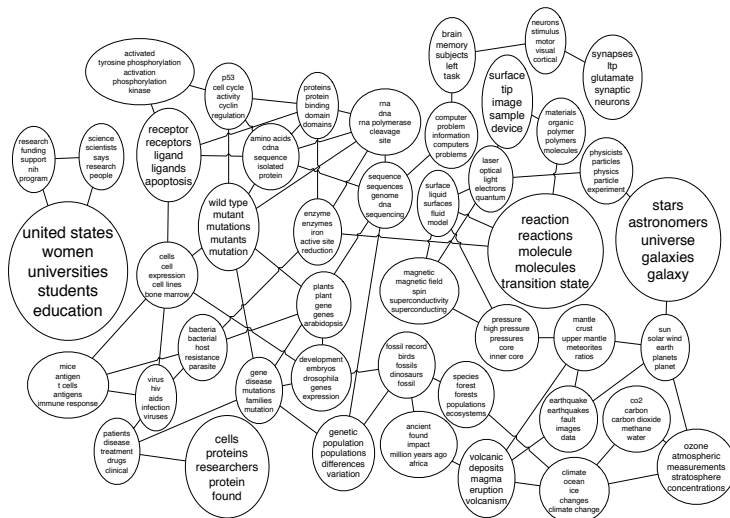
### Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

### Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database  
How Big Is the Universe of Exons?  
Counting and Discounting the Universe of Exons  
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment  
Ancient Conserved Regions in New Gene Sequences and the Protein Databases  
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure  
Testing the Exon Theory of Genes: The Evidence from Protein Structure  
Predicting Coiled Coils from Protein Sequences  
Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

# Example of LDA: *Science* topics



## Example of LDA: SciReader [scireader.org](http://scireader.org)

- Site: “Every month 90,000 new papers are published in BioMedicine. How do you find the essential science that you need to read?”
- “SciReader automatically sorts all new biomedical papers into related topics and allows you to view the recent papers that are most likely to become influential in each topic”
- “Like” or upload a set of PDFs to get personalized recommendations
- Underlying machinery: topic model with 20 high level topics; 140 lower level topics that partition the high level topics.



# Example of LDA: SciReader scireader.org

SciReader

Recommendations

Topics

Journals

About

Search Papers

Q

Sign In

Sign Up

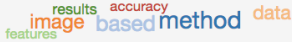
Topics Home

Topic List

My Topics

Topics

Computational Methods / Machine Learning



results accuracy data image based method features

Save this topic

**Balanced Sparse Model for Tight Frames in Compressed Sensing Magnetic Resonance Imaging.**  
Y Liu, JF Cai, ... Z Chen, X Qu.  
PLoS ONE. April 8, 2015. 10(4).

**Motor Imagery Classification via Combinatory Decomposition of ERP and ERSP using Sparse Nonnegative Matrix Factorization.**  
N Lu, T Yin.  
J. Neurosci. Methods. April 8, 2015.

**Multi-dimensional complete ensemble empirical mode decomposition with adaptive noise applied to laser speckle contrast images.**  
A Humeau-Heurtier, G Mahe, P Abraham.  
IEEE Trans Med Imaging. April 8, 2015.

**Restoration of Motion-Blurred Image Based on Border Deformation Detection: A Traffic Sign Restoration Model.**  
Y Zhang, H ...

Email to friends

Tweet this paper

Add to my library

Open in IEEE Trans L...

**Robust Two-Dimensional Principal Component Analysis: A Structured Sparsity Regularized Approach.**  
Yipeng Sun, Xiaoming Tao, Yang Li, Jianhua Lu.  
IEEE transactions on image processing : a publication of the IEEE Signal Processing Society. April 4, 2015. PMID: 25838521

**Abstract**

Principal component analysis (PCA) is widely used to extract features and reduce dimensionality in various computer vision and image/video processing tasks. Conventional approaches either lack robustness to outliers and corrupted data or are designed for one-dimensional signals. To address this problem, we propose a robust principal component analysis model for two-dimensional images incorporating structured sparse priors, referred to as structured sparse 2D-PCA. This robust model considers the prior of structured and grouped pixel values in two dimensions. As the proposed formulation is jointly non-convex and nonsmooth, which is difficult to tackle by joint optimization, we develop a two-stage alternating minimization approach to solve the problem. This approach iteratively learns the projection matrices by bi-directional decomposition and utilizes the proximal method to obtain the structured sparse outliers.

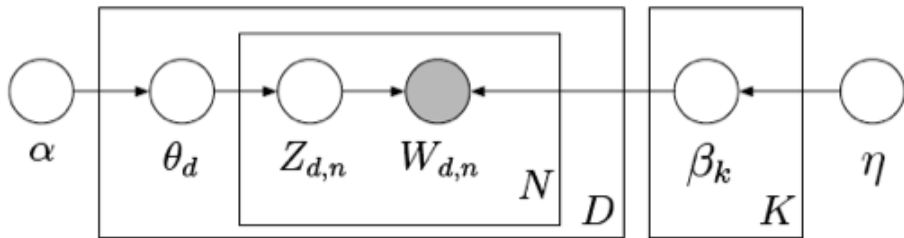
COS424/524, SML302

Probabilistic Topic Models

Lecture 18

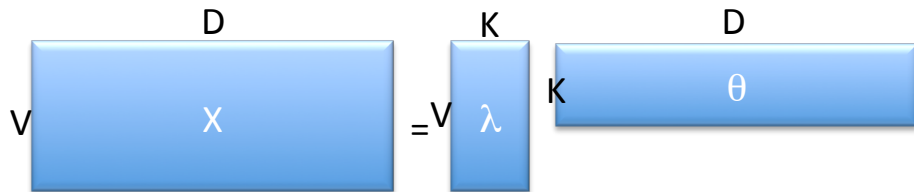
41 / 51

## LDA model: example of a mixed-membership model



- LDA is a *mixed membership model*, where each sample may have features from several topics simultaneously
- In MMMs: each sample is assigned to multiple topics via latent assignment  $z_{d,n}$  of each word to its own topic.
- This differs from a *mixture model*, where each sample has features from a single topic
- MMMs generalize LDA to any distribution (not just multinomial)

# Poisson matrix factorization

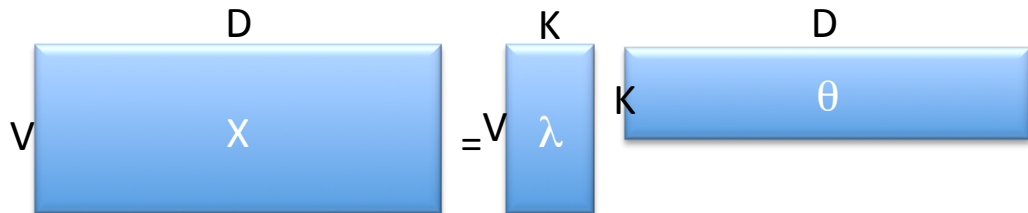


This model has been described for Poisson counts (and also models bag-of-words features):

- 1 For each term  $v$  and topic  $k$ : draw  $\beta_{k,v} \sim \text{Gamma}(a, b)$
- 2 For each document  $d$ :
  - For each topic  $k$ : draw  $\theta_{d,k} \sim \text{Gamma}(c, d)$ .
  - For each term  $v$ : draw  $w_{d,v} \sim \text{Pois}(\theta_d^\top \beta_v)$ .

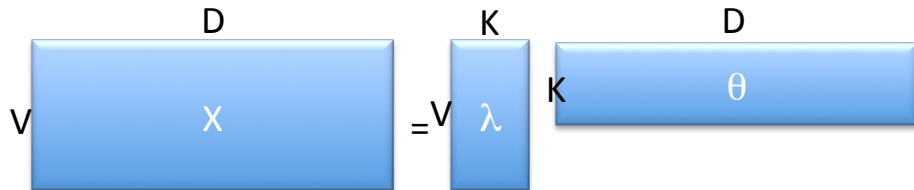
Related to non-negative matrix factorization (NMF) (Lee & Seung 1999)

# Poisson matrix factorization



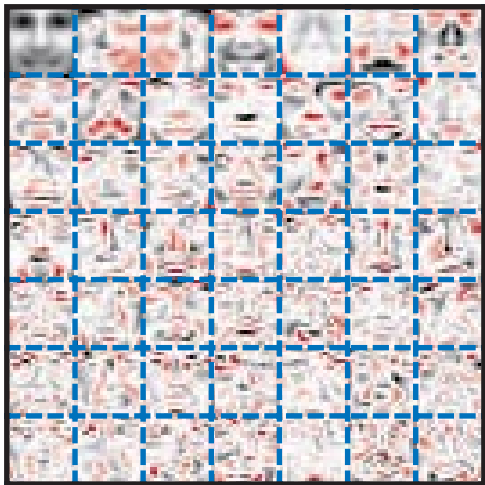
- Better prediction accuracy (according to specific metrics) than LDA. (*Canny, 2004*)
- Easy to fit with auxiliary variables
- Easy to extend the Poisson additive model on word counts
- Equivalent to LDA when we condition on document length (It is multinomial PCA.)
- Is a Bayesian form of NMF with “KL loss” (*Lee and Seung, 2000*)

# Poisson matrix factorization

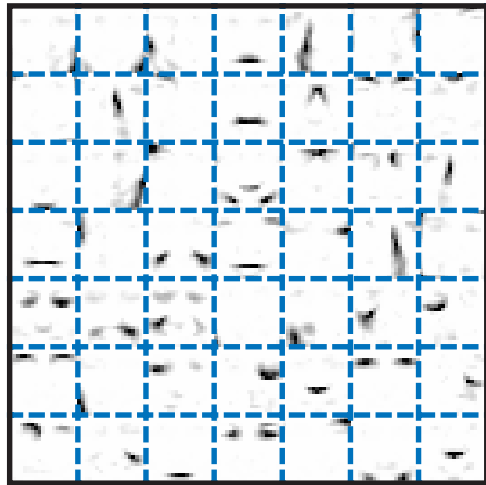


- Poisson matrix factorization works well in many settings
  - networks (*Ball et al., 2012*);
  - recommendation systems (*Gopalan et al., 2013*)
- Poisson model versus LDA
  - Poisson model explicitly models document length
  - Poisson model avoids difficult normalizations (Dirichlet)
  - LDA has normalized topic proportions: fractional information

# Eigenfaces: PCA versus NMF

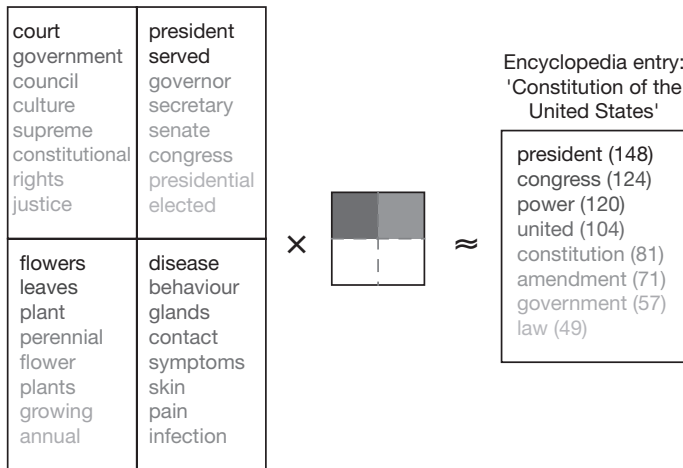


Low dimensional feature space using PCA



Low dimensional feature space using NMF

# Latent topics: NMF



NMF applied to encyclopedia articles. [Lee & Seung 1999]

- Dynamic topic models
- Correlated topic models
- Supervised Topic Models
- Relational topic models
- Topic models that include more language information (bigrams, syntax, etc.)
- Treed topic models
- Non-parametric topic models (Dirichlet process mixture models)



# LDA: Summary & Assumptions

- LDA assigns each word in a document to a topic according to the document-specific topic proportion, and draws from the word distribution of that topic
- LDA assumes that each word is independent, each topic is independent, and each document is independent
- LDA is a method to reduce a set of samples to a low dimensional latent space: a set of “topics”

# Topic models: a history

- Latent Semantic Analysis (LSA) (Deerwester 1990)
- probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999)
- Independently invented in population genetics (Pritchard et al. 2000)
- Interpretation as *multinomial PCA* (Buntine 2002)
- A *mixed-membership model* (Erosheva 2004)

## Additional Resources

- MLAPA: Chapter 26
- Description of LDA model in genomics: *[Pritchard, Stephens, Donnelly 2000]*
- Description of LDA model in ML context: *[Blei, Ng, Jordan 2003]*
- Application of LDA to scientific articles: *[Griffiths & Steyvers 2004]*
- Topic model review: *[Blei 2011]*
  
- (video) David M. Blei: *Topic models*
- *Metacademy*: Latent Dirichlet allocation
- Code: MALLET