

Probabilistic graphical models and topic models

Sources:

- “Topic models”, David Blei, MLSS ’09, MLSS’12
http://videolectures.net/mlss09uk_blei_tm/?q=david%20blei
- Parts of “Probabilistic graphical models”, Christopher Bishop, MLSS’13
<https://www.youtube.com/watch?v=ju1Grt2hdko>
- Parts of “Machine learning: Graphical models”, Alex Smola,
<http://alex.smola.org/teaching/10-701-15/graphical.html>

Supervised methods
for text classification:

Graphical model of
Naïve Bayes classifier

Bayes' Rule applied to documents and classes

Imagine that we try to infer what is the class of a document d , where c stands for some given class.

A diagram illustrating Bayes' theorem. The equation $P(c | d) = \frac{P(d | c)P(c)}{P(d)}$ is centered. Three blue lines point to parts of the equation: one from the word 'likelihood' to $P(d | c)$, one from the word 'prior' to $P(c)$, and one from the word 'posterior' to $P(c | d)$.

$$P(c | d) = \frac{P(d | c)P(c)}{P(d)}$$

likelihood

prior

posterior

Bayes theorem

Bayes' Rule applied to documents and classes

Imagine that we try to infer what is the class of a document d , where c stands for some given class.

$$P(d \mid c) = P(\underbrace{x_1, x_2, \dots, x_n}_{\text{features}} \mid c)$$

Document d represented as features x_1, \dots, x_n

- word presence (binary value)
- word counts
- word frequencies (tf)
- tf-idf

Independence Assumptions

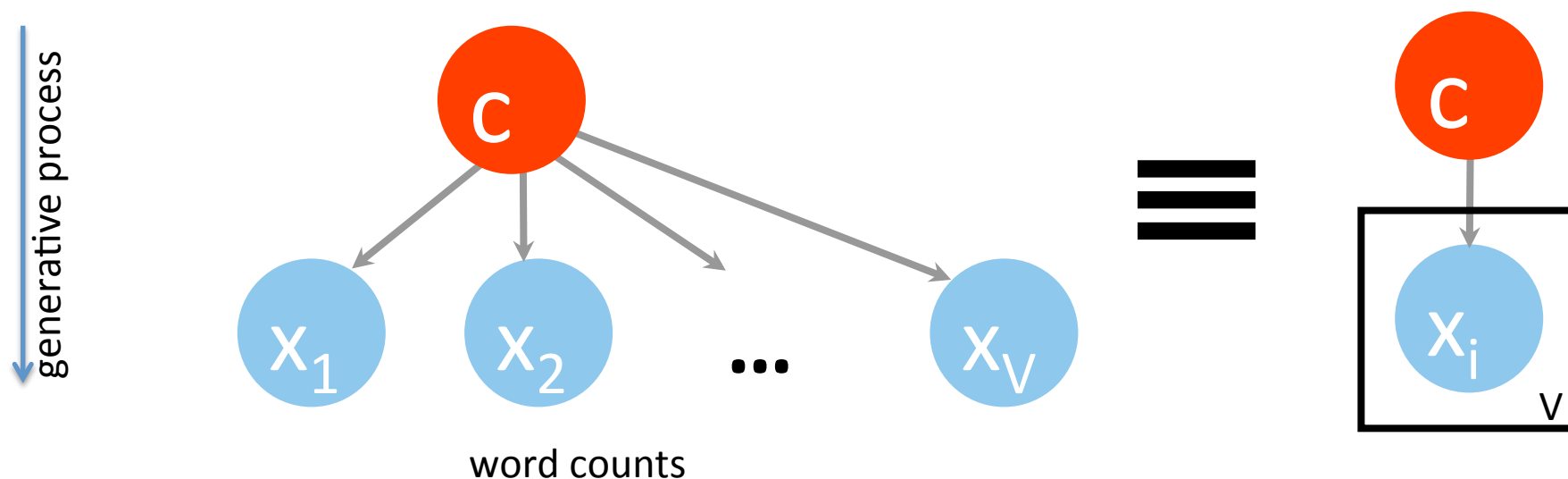
- **Bag of words assumption:** Assume that position of words doesn't matter (the **exchangeability** of random variables)

$$P(x_1, x_2, \dots, x_n | c) = P(x_{\delta(1)}, x_{\delta(2)}, \dots, x_{\delta(n)} | c)$$

- **Conditional Independence:** Assume the feature probabilities $P(x_i | c_j)$ are independent given the class c .

$$P(x_1, \dots, x_n | c) = P(x_1 | c) \cdot P(x_2 | c) \cdot P(x_3 | c) \cdot \dots \cdot P(x_n | c) = \prod_{i \in V} P(x_i | c)$$

Graphical model for Naïve Bayes (factorization describing generative process)

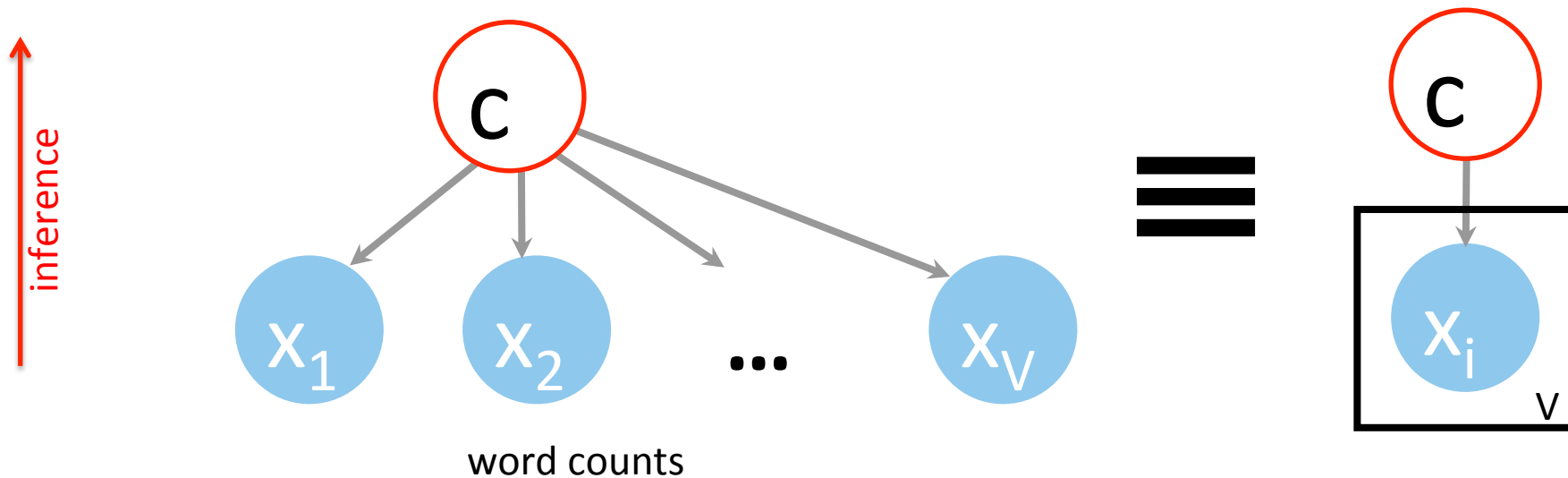


Factorization of
the joint probability:

$$P(x_1, \dots, x_n, c) = P(c) \prod_{i \in V} P(x_i | c)$$

vocabulary size

Graphical model for Naïve Bayes (inference)



Inference by finding
maximum a posteriori:

$$P(c|d) = P(d|c)P(c) = P(c) \prod_{i \in V} P(x_i|c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{i \in V} P(x_i|c)$$

Typical choices of the likelihood $P(\mathbf{x} | c)$ (Assumptions of the generative model)

Multinomial Naïve Bayes classifier:

$$P(\mathbf{x} | c) = \text{Multinomial}(\mathbf{x}, \boldsymbol{\theta})$$

$$\propto \prod_{i \in V} (\theta_{ic})^{x_i}$$

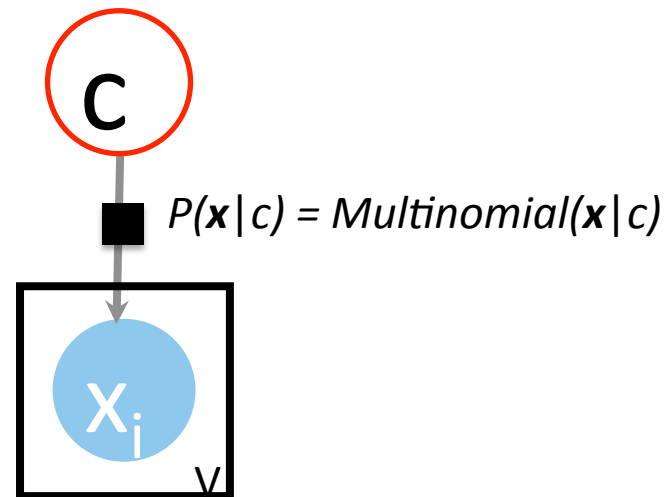
Bernoulli Naïve Bayes classifier:

$$P(\mathbf{x} | c) = \text{MultivariateBernoulli}(\mathbf{x}, \boldsymbol{\theta})$$

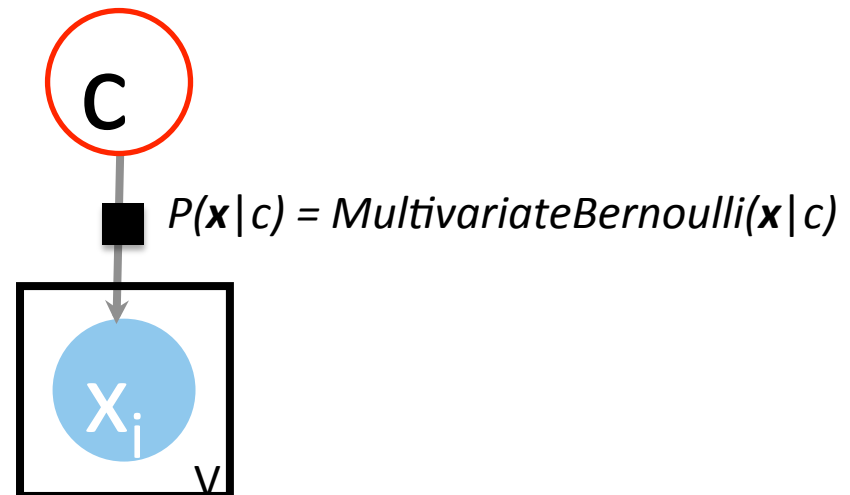
$$= \prod_{i \in V} (\theta_{ic})^{x_i} (1 - \theta_{ic})^{1-x_i}$$

Typical choices of the likelihood $P(\mathbf{x} | c)$ (Assumptions of the generative model)

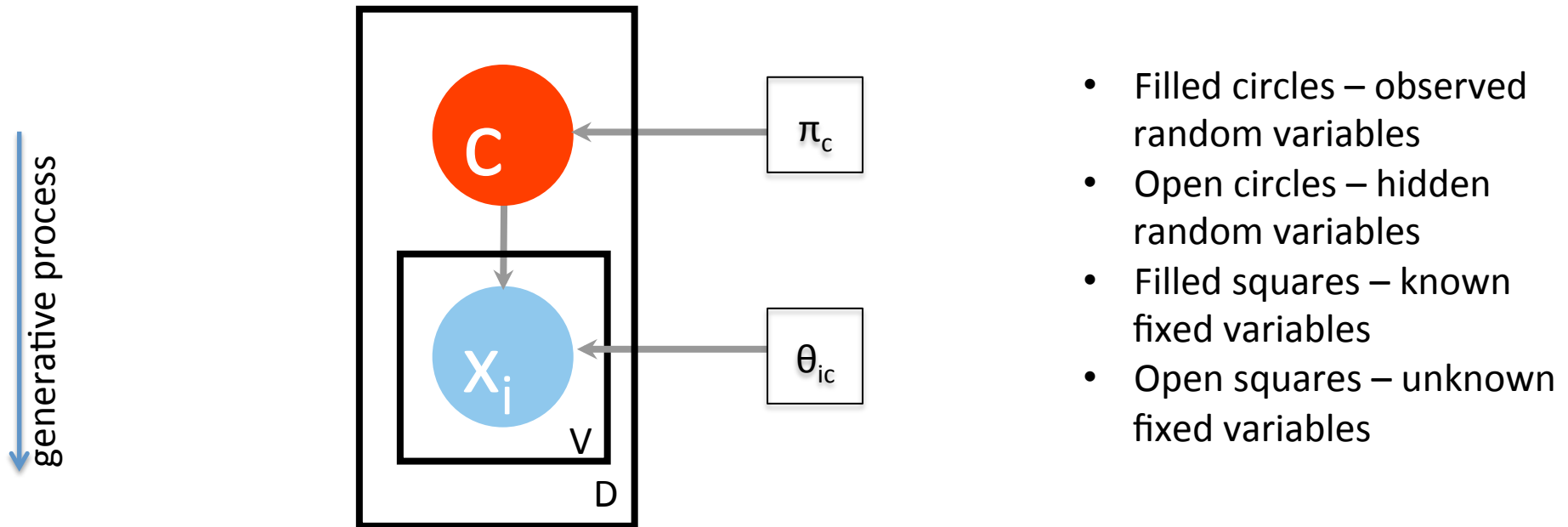
Multinomial Naïve Bayes classifier:



Bernoulli Naïve Bayes classifier:



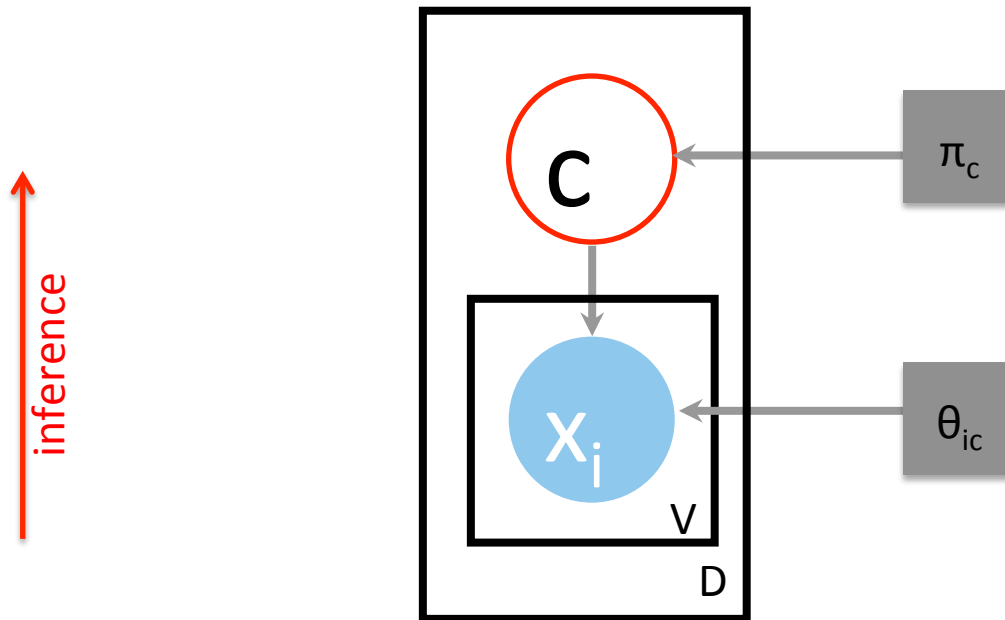
Full graphical model for Naïve Bayes (training)



Factorization of
the joint probability:

$$P(\vec{x}, \vec{c}) = \prod_{d=1}^D \overset{\pi_c}{P(c_d)} \underbrace{\prod_{i=1}^V P(x_{id} | c_d)}_{P(\mathbf{x} | c) = \text{Multinomial}(\mathbf{x}, \boldsymbol{\theta})}$$

Full graphical model for Naïve Bayes (inference)



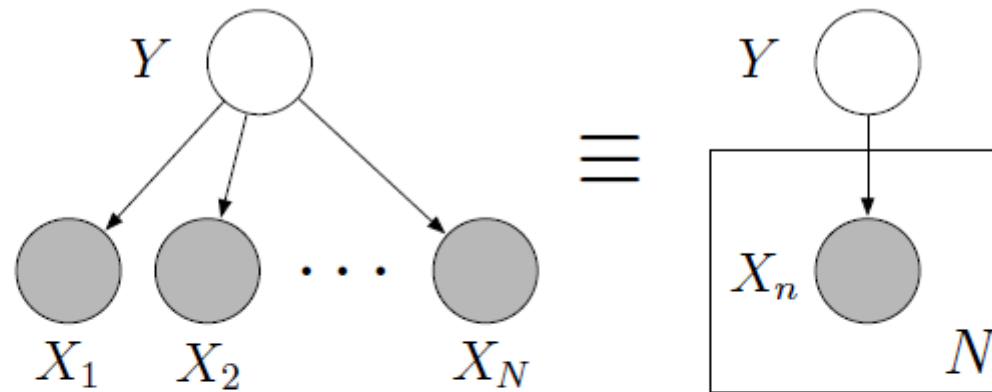
- Filled circles – observed random variables
- Open circles – hidden random variables
- Filled squares – known fixed variables
- Open squares – unknown fixed variables

For which class c the posterior has max value for the given document d ?

$$P(c | \vec{x}) \propto \underbrace{P(c_d)}_{\pi_c} \prod_{i=1}^N \underbrace{P(x_{id} | c_d)}_{P(\mathbf{x} | c) = \text{Multinomial}(\mathbf{x}, \boldsymbol{\theta})}$$

Graphical models

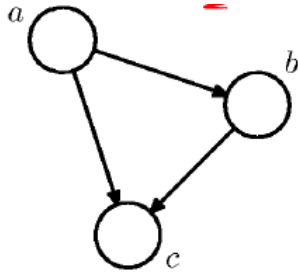
Graphical models (summary)



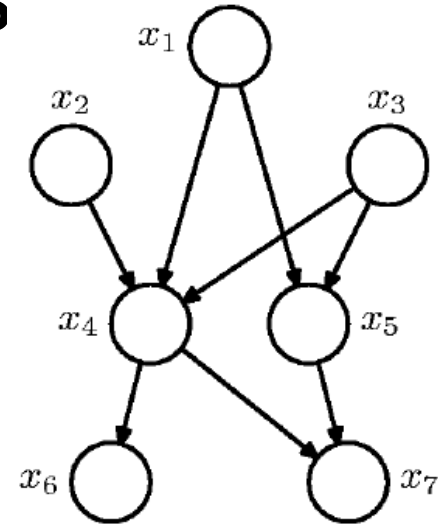
- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure
- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

$$p(y, x_1, \dots, x_N) = p(y) \prod_{n=1}^N p(x_n | y)$$

Graphical models



$$P(a,b,c) = P(a)P(b|c)P(c|a,b)$$

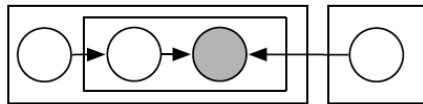


$$P(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \dots$$

- No cycles
- Full graph describes the most generic joint distribution
- Links missing from the full graph specify the joint distribution by making assumptions about conditional dependences between variables
- A directed link defines conditional dependence and may imply a causal relation
- Empty circles are hidden (latent) variables
- Filled circles are observed variables

Probabilistic graphical models

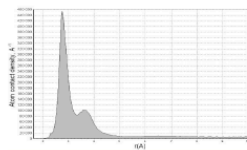
Make assumptions



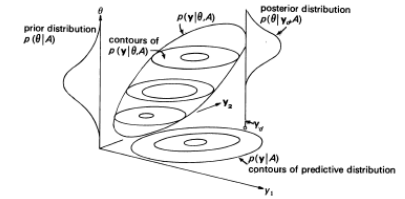
Collect data



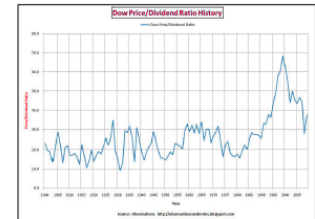
Infer the posterior



Check



Predict



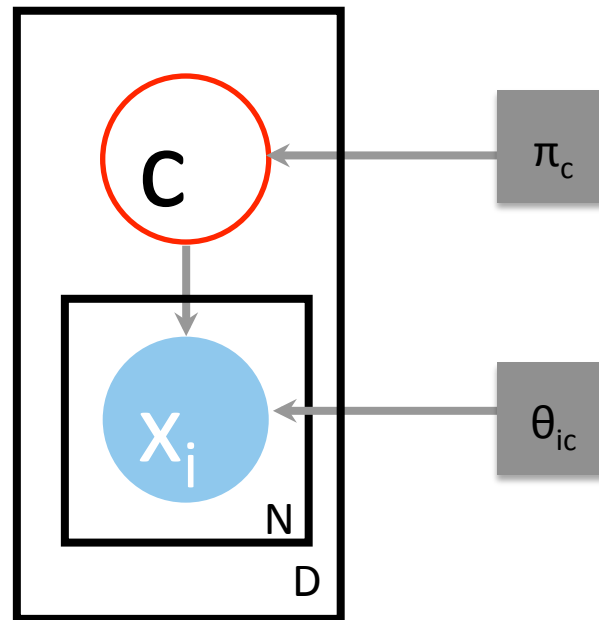
Explore



Unsupervised models

(Mixture models)

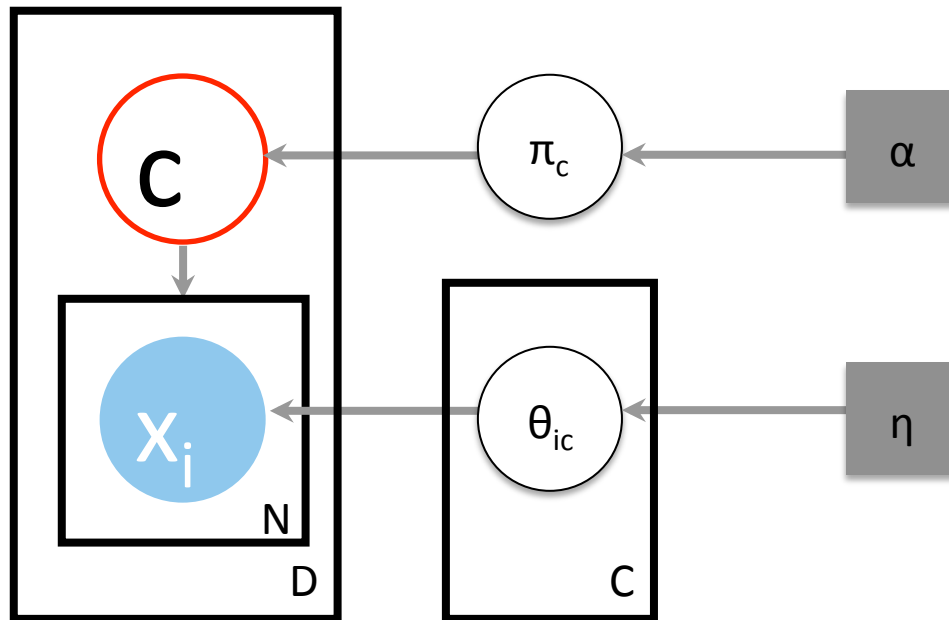
Full graphical model for trained Naïve Bayes classifier



1. Draw class z from a $\text{Categorical}(c, \boldsymbol{\pi})$
2. Draw word x from a $\text{Categorical}(x, \boldsymbol{\theta}_{\cdot c})$

- Filled circles – observed random variables
- Open circles – hidden random variables
- Filled squares – known fixed variables
- Open squares – unknown fixed variables

Bayesian categorical mixture model (unsupervised)



1. Draw class proportions from $\text{SymmetricDirichlet}(\boldsymbol{\pi}, \alpha)$
2. Draw class proportions from $\text{SymmetricDirichlet}(\boldsymbol{\theta}, \alpha)$
3. Draw class z from $\text{Categorical}(c, \boldsymbol{\pi})$
4. Draw word x from $\text{Categorical}(x, \boldsymbol{\theta}_{\cdot c})$

- Filled circles – observed random variables
- Open circles – hidden random variables
- Filled squares – known fixed variables
- Open squares – unknown fixed variables

Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

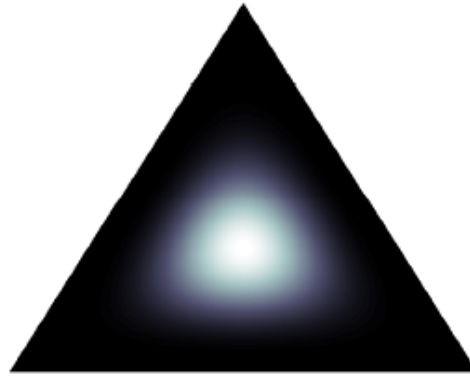
$$p(\theta | \vec{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- The Dirichlet is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ .
- The topic proportions are a K dimensional Dirichlet.
The topics are a V dimensional Dirichlet.

Dirichlet Examples



$$\alpha = (2, 2, 2)$$



$$\alpha = (5, 5, 5)$$



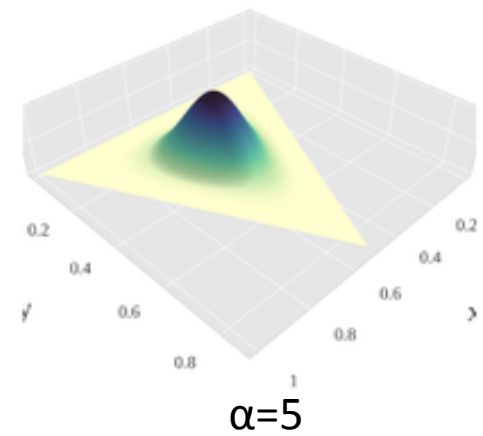
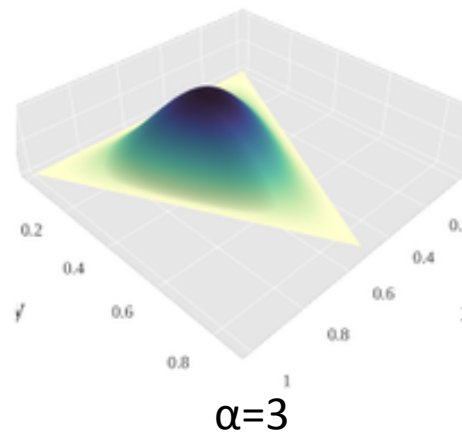
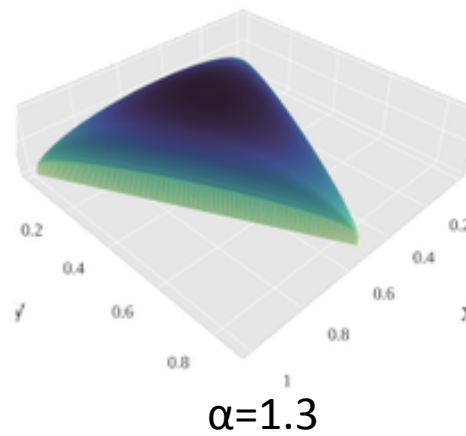
$$\alpha = (2, 2, 25)$$

Darker implies lower magnitude

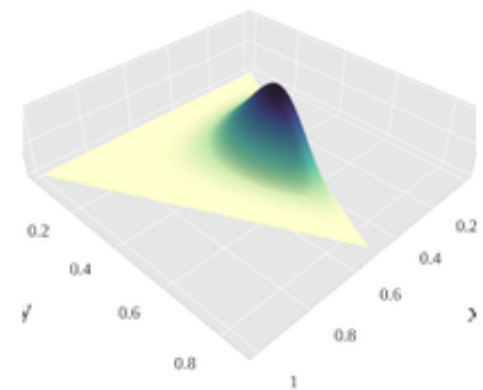
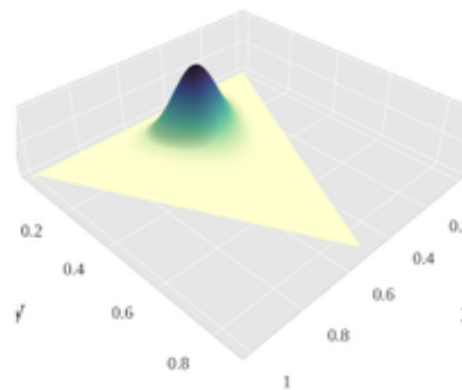
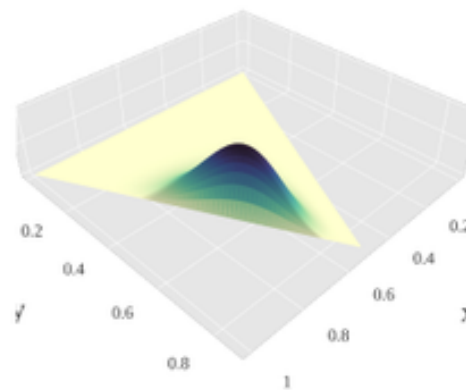
$\alpha < 1$ leads to sparser topics

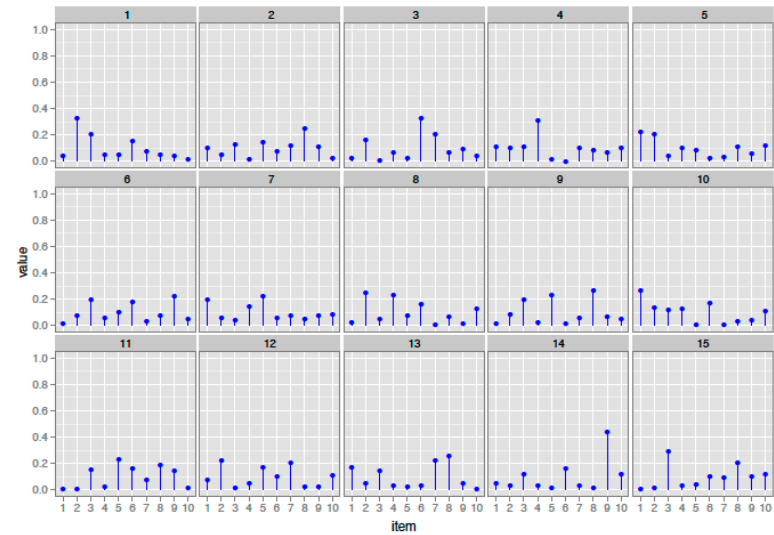
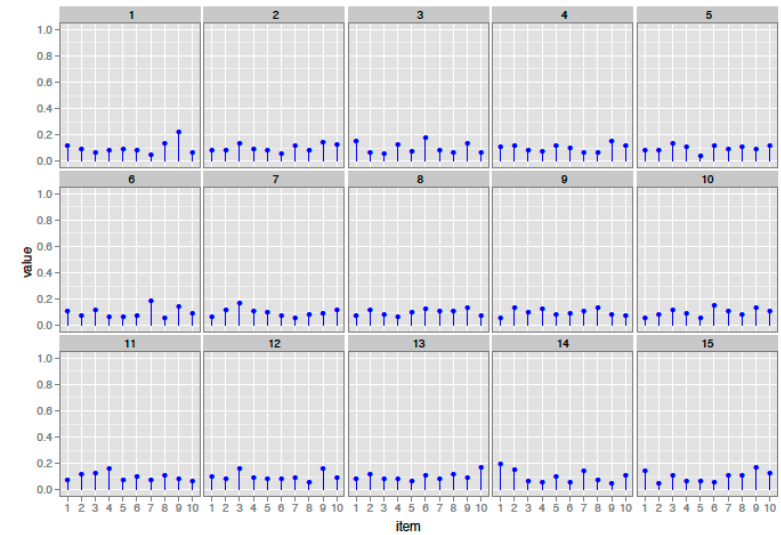
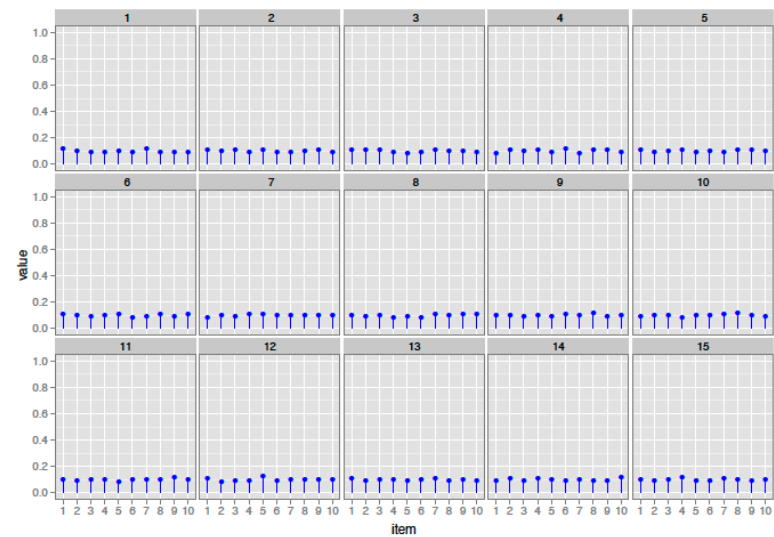
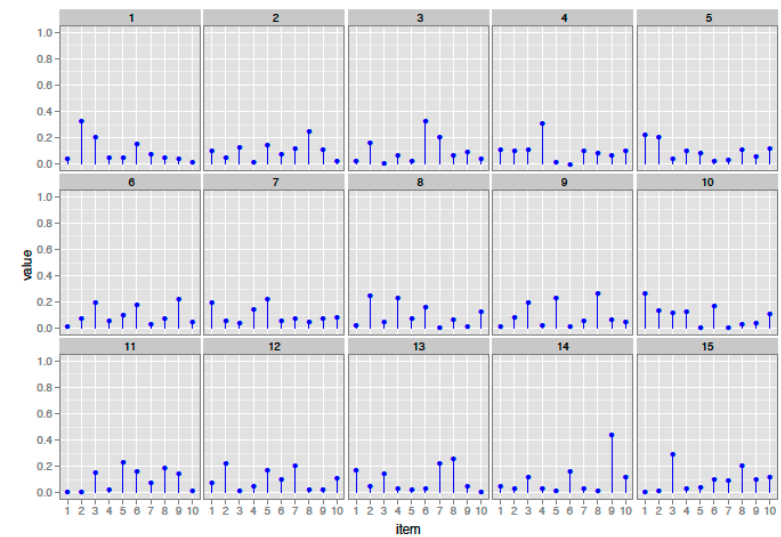
Dirichlet Examples

Symmetric
($\alpha_i = \alpha_j = \alpha$)

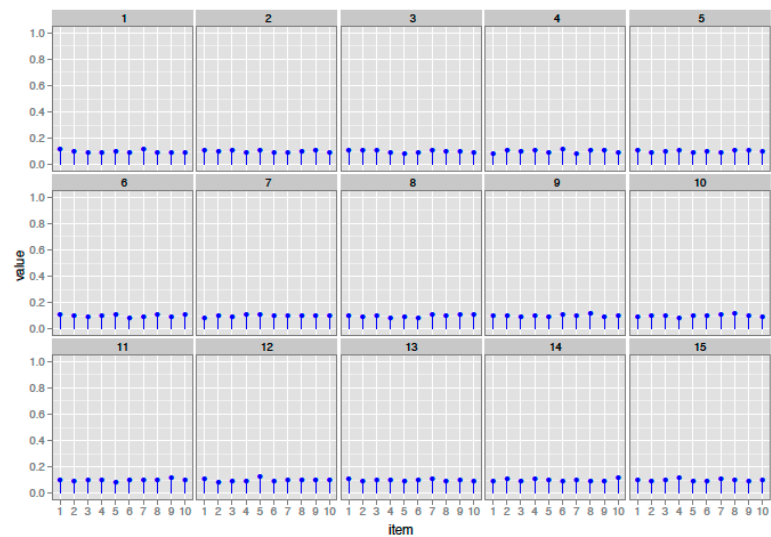


Asymmetric
($\alpha_i \neq \alpha_j$)

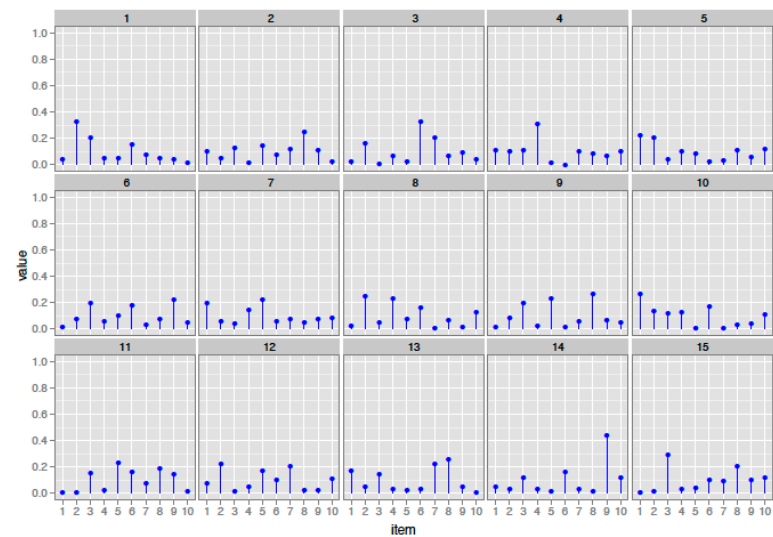


$\alpha = 1$  $\alpha = 10$  $\alpha = 100$  $\alpha = 1$ 

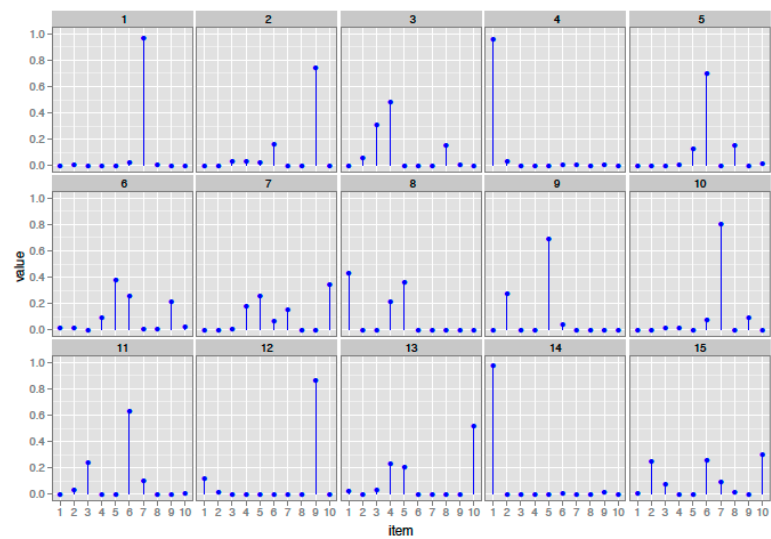
$\alpha = 100$



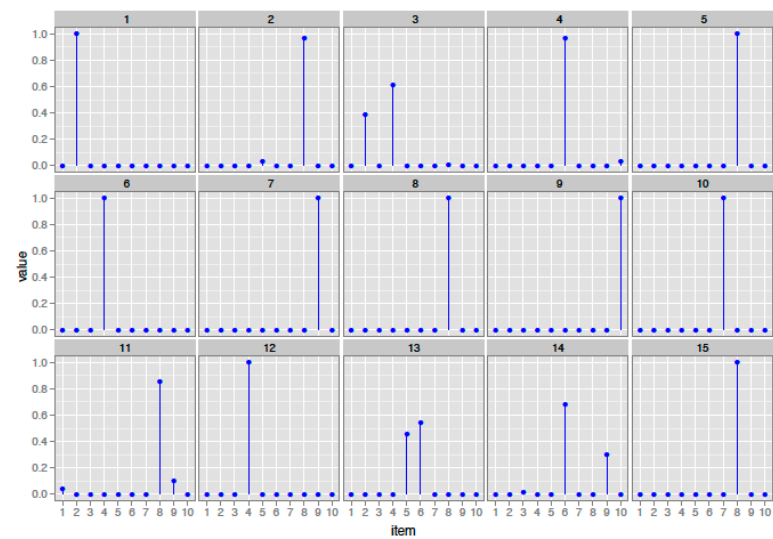
$\alpha = 1$



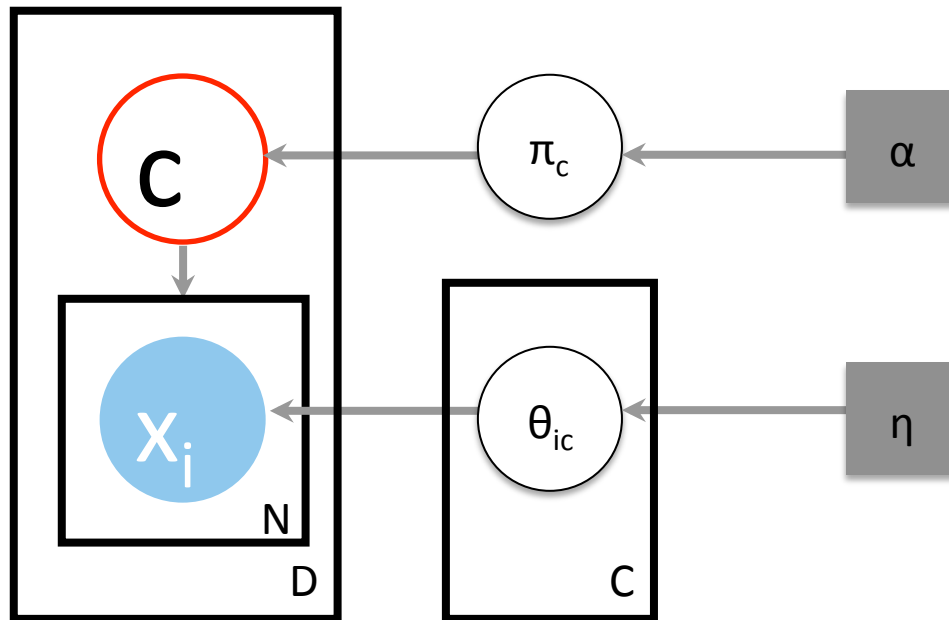
$\alpha = 0.1$



$\alpha = 0.01$



Bayesian categorical mixture model (how to learn all hidden parameters?)



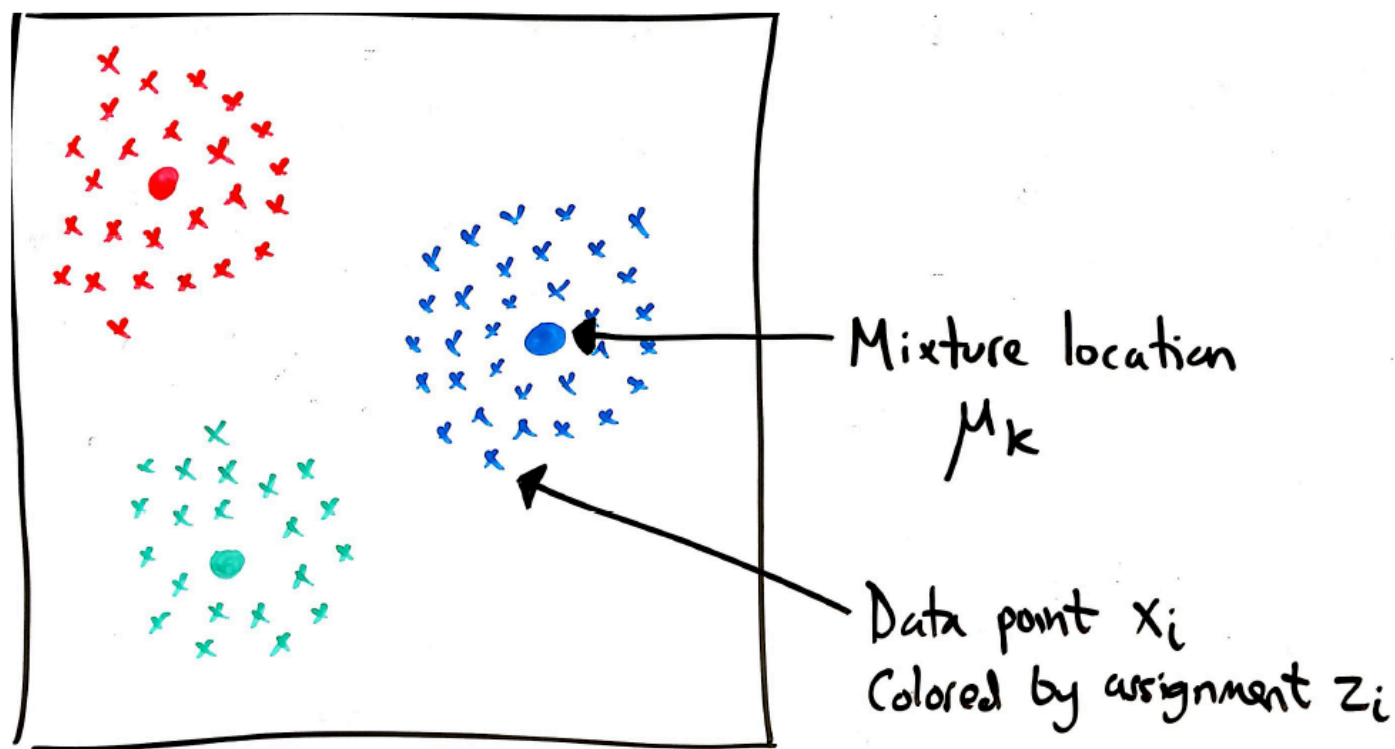
1. Draw class proportions from $\text{SymmetricDirichlet}(\boldsymbol{\pi}, \alpha)$
2. Draw class proportions from $\text{SymmetricDirichlet}(\boldsymbol{\theta}, \alpha)$
3. Draw class z from $\text{Categorical}(c, \boldsymbol{\pi})$
4. Draw word x from $\text{Categorical}(x, \boldsymbol{\theta}_{\cdot c})$

- Filled circles – observed random variables
- Open circles – hidden random variables
- Filled squares – known fixed variables
- Open squares – unknown fixed variables

Gaussian mixture model (unsupervised)

Source: David Blei's lecture

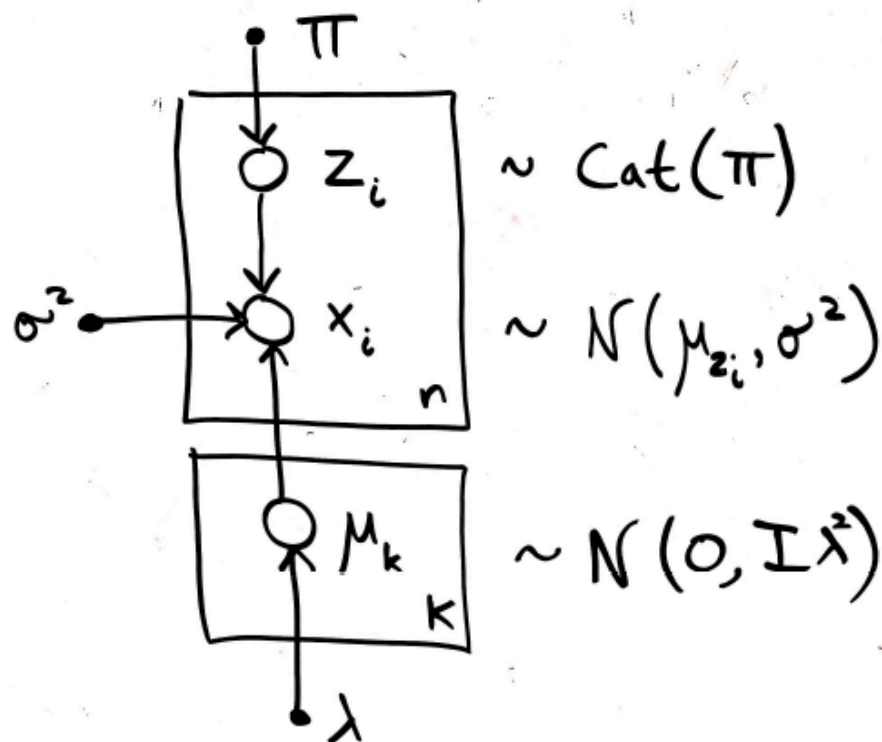
<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf>



Gaussian mixture model (unsupervised)

Source: David Blei's lecture

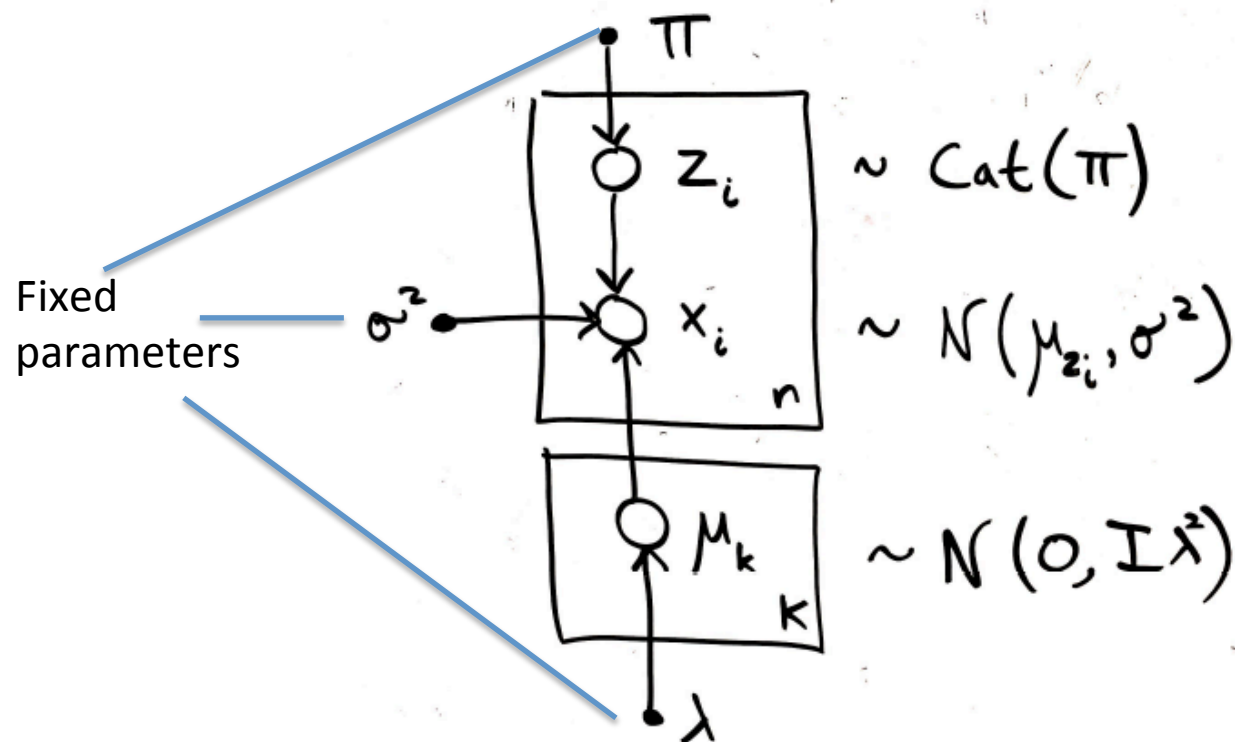
<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf>



Gaussian mixture model (unsupervised)

Source: David Blei's lecture

<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf>



Inference in Gaussian mixture model

Source: David Blei's lecture

<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf>

The Gibbs sampler

¶ The main idea behind Gibbs sampling (and all of MCMC) is to approximate a distribution with a set of samples. For example, in the mixture model,

Input: Data \mathbf{x} and a number of components K

Initialize: Mixture locations $\boldsymbol{\mu}$.

Maintain mixture locations $\boldsymbol{\mu}$ and mixture assignments \mathbf{z} .

Repeat:

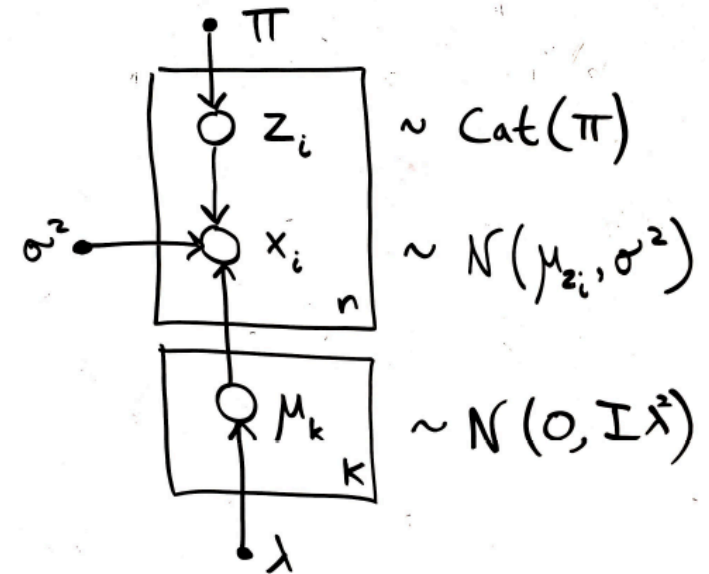
1. For each $i \in \{1, \dots, n\}$:
Sample $z_i \mid \{\boldsymbol{\mu}, \mathbf{z}_{-i}, \mathbf{x}\}$ from $p(z_i \mid \boldsymbol{\mu}, x_i)$
2. For each $k \in \{1, \dots, K\}$:
Sample $\mu_k \mid \{\boldsymbol{\mu}_{-k}, \mathbf{z}, \mathbf{x}\}$ from $p(\mu_k \mid \mathbf{z}, \mathbf{x})$

Figure 1: The Gibbs sampler for mixture of Gaussians.

Inference in Gaussian mixture model

Complete (full) conditional probability for variable z :

$$\begin{aligned} p(z_i | \mu, x_i) &\propto p(z_i) p(x_i | \mu_{z_i}) \\ &= \pi_{z_i} \phi(x_i; \mu_{z_i}, \sigma^2) \end{aligned}$$



Inference in Gaussian mixture model

Complete (full) conditional probability
for variable μ :

$$p(\mu_k | \mathbf{z}, \mathbf{x})$$

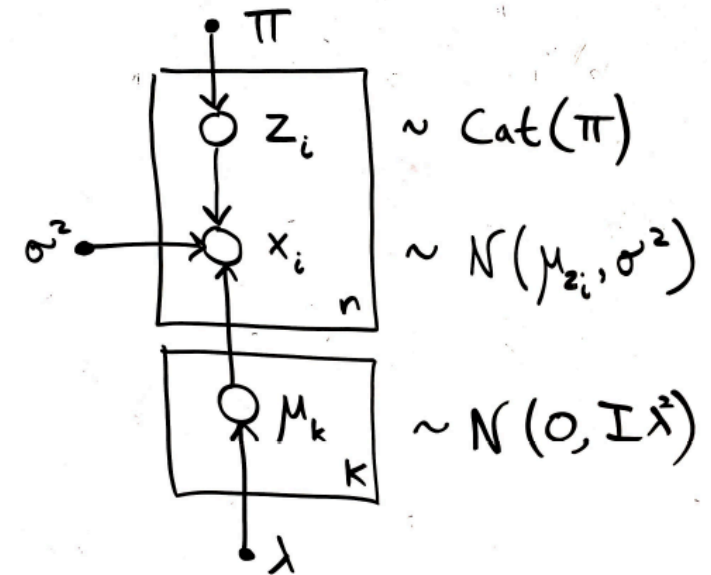
$$\mu_k | \mathbf{z}, \mathbf{x} \sim \mathcal{N}(\hat{\mu}_k, \hat{\lambda}_k)$$

$$\hat{\mu}_k = \left(\frac{n_k / \sigma^2}{n_k / \sigma^2 + 1 / \lambda^2} \right) \bar{x}_k$$

$$\hat{\lambda} = (n_k / \sigma^2 + 1 / \lambda^2)^{-1},$$

$$n_k = \sum_{i=1}^n z_i^k$$

$$\bar{x}_k = \frac{\sum_{i=1}^n z_i^k x_i}{n_k}.$$



The prior normal distribution is
the conjugate prior to normal
posterior wrt normal likelihood

https://en.wikipedia.org/wiki/Conjugate_prior

Inference in Gaussian mixture model

Source: David Blei's lecture

<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf>

The Gibbs sampler

¶ The main idea behind Gibbs sampling (and all of MCMC) is to approximate a distribution with a set of samples. For example, in the mixture model,

Input: Data \mathbf{x} and a number of components K

Initialize: Mixture locations $\boldsymbol{\mu}$.

Maintain mixture locations $\boldsymbol{\mu}$ and mixture assignments \mathbf{z} .

Repeat:

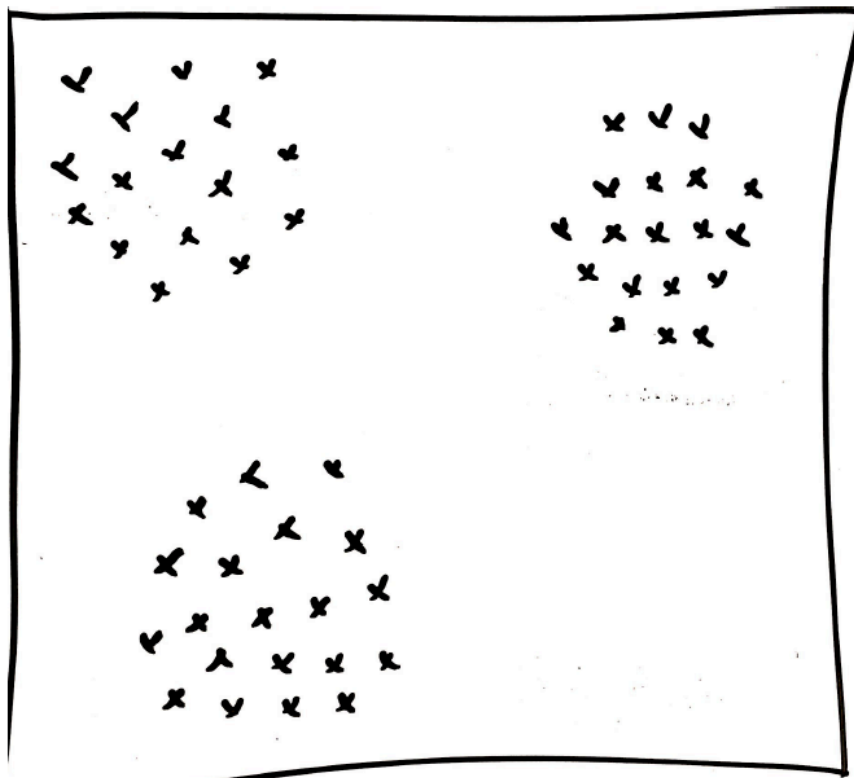
1. For each $i \in \{1, \dots, n\}$:
Sample $z_i \mid \{\boldsymbol{\mu}, \mathbf{z}_{-i}, \mathbf{x}\}$ from $p(z_i \mid \boldsymbol{\mu}, x_i)$
2. For each $k \in \{1, \dots, K\}$:
Sample $\mu_k \mid \{\boldsymbol{\mu}_{-k}, \mathbf{z}, \mathbf{x}\}$ from $p(\mu_k \mid \mathbf{z}, \mathbf{x})$

Figure 1: The Gibbs sampler for mixture of Gaussians.

Gaussian mixture model (learning)

Source: David Blei's lecture

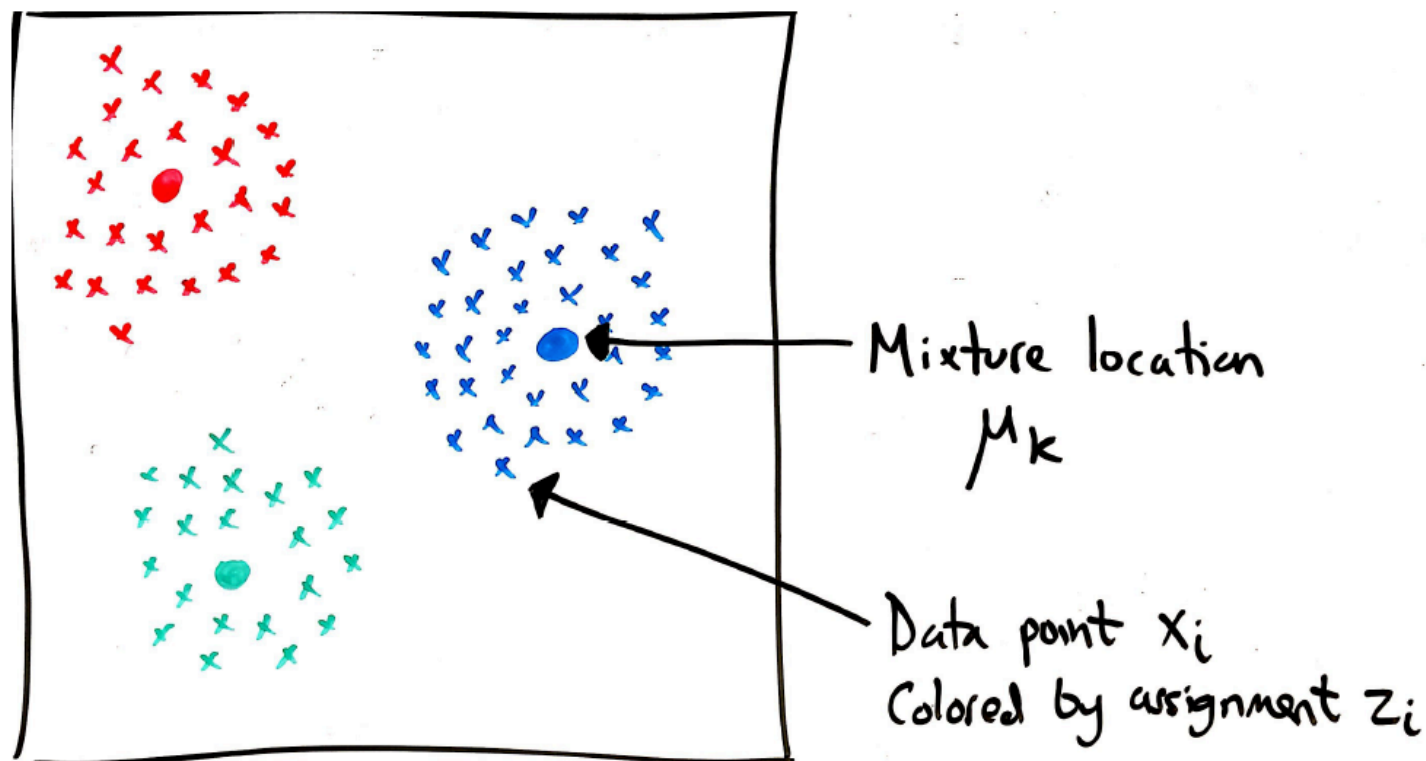
<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf>



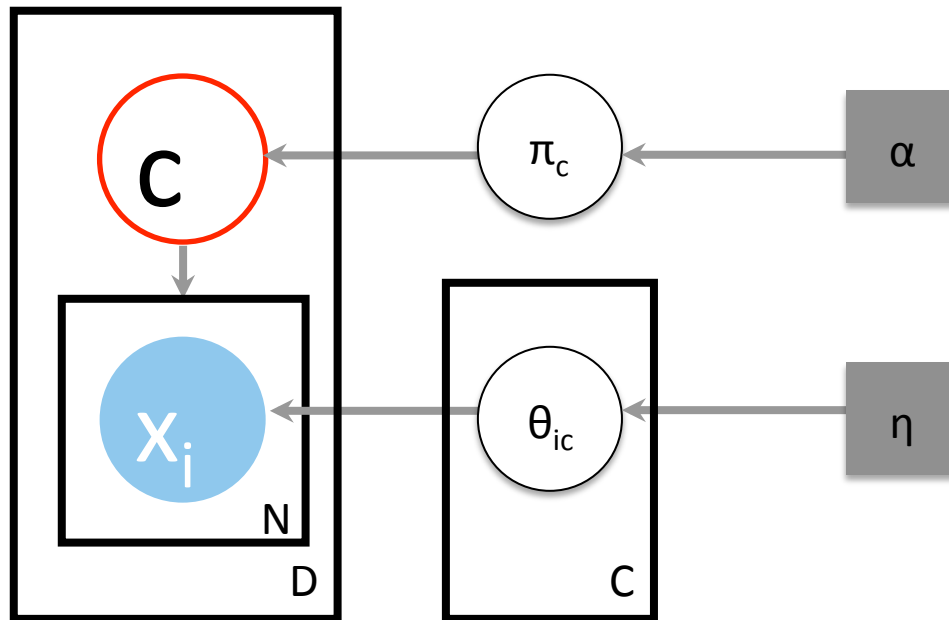
Gaussian mixture model (learning)

Source: David Blei's lecture

<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf>



Bayesian categorical mixture model (how to learn all hidden parameters?)



1. Draw class proportions from $\text{SymmetricDirichlet}(\boldsymbol{\pi}, \alpha)$
2. Draw class proportions from $\text{SymmetricDirichlet}(\boldsymbol{\theta}, \alpha)$
3. Draw class z from $\text{Categorical}(c, \boldsymbol{\pi})$
4. Draw word x from $\text{Categorical}(x, \boldsymbol{\theta}_{\cdot c})$

- Filled circles – observed random variables
- Open circles – hidden random variables
- Filled squares – known fixed variables
- Open squares – unknown fixed variables

Unsupervised text classification and topic modeling

Unsupervised learning



As more information becomes available, it becomes more difficult to find and discover what we need.

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

Often data is not labeled/annotated, so supervised methods are not possible or expensive.

Unsupervised learning

- Clustering (networks, vectors)
- Principal component analysis (PCA)
- Non-negative matrix factorization
- Mixture models
- Mixed membership models, topic models (e.g., LDA)

Applications?

A word cloud of application areas, with the words arranged in a circular pattern. The words are: spammers, ads, users, products, urls, abuse, mails, text, news, locations, queries, and events. The words are in a sans-serif font, with varying sizes and orientations.

spammers

ads

users

products

urls

abuse

mails

text

news

locations

queries

events

Topic modeling - Motivation

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.

Applications

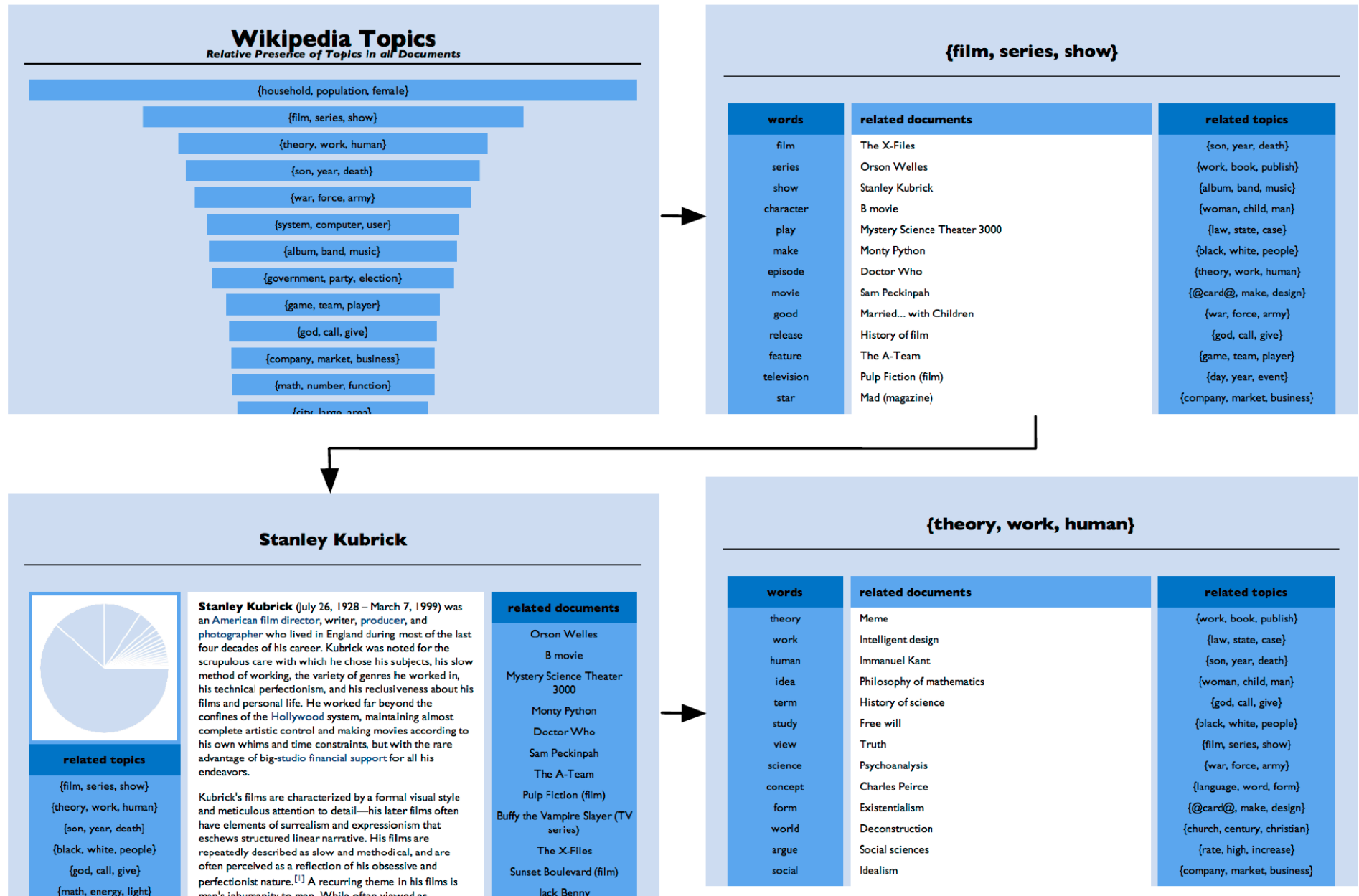


Image annotation



SKY WATER TREE
MOUNTAIN PEOPLE



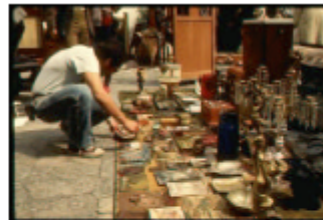
SCOTLAND WATER
FLOWER HILLS TREE



SKY WATER BUILDING
PEOPLE WATER



FISH WATER OCEAN
TREE CORAL



PEOPLE MARKET PATTERN
TEXTILE DISPLAY



BIRDS NEST TREE
BRANCH LEAVES

Topic-modeling timeline

1901 – PCA

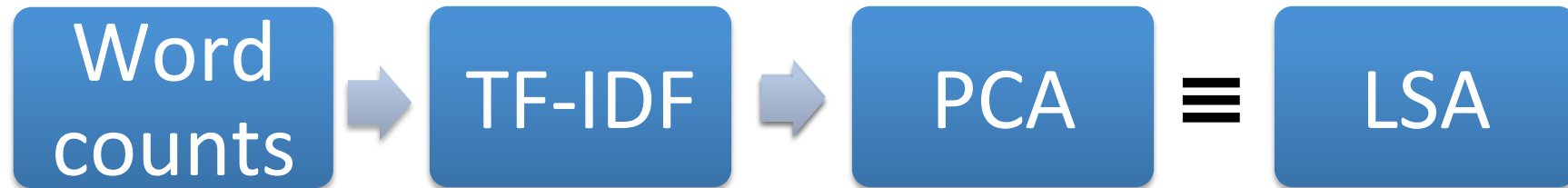
1988 – Latent semantic analysis (LSA)

1999 – probabilistic LSA (pLSA)

2003 – Latent Dirichlet Allocation (LDA)

2006 – non-parametric Bayesian topic models

Latent semantic analysis/indexing (LSA/LSI)



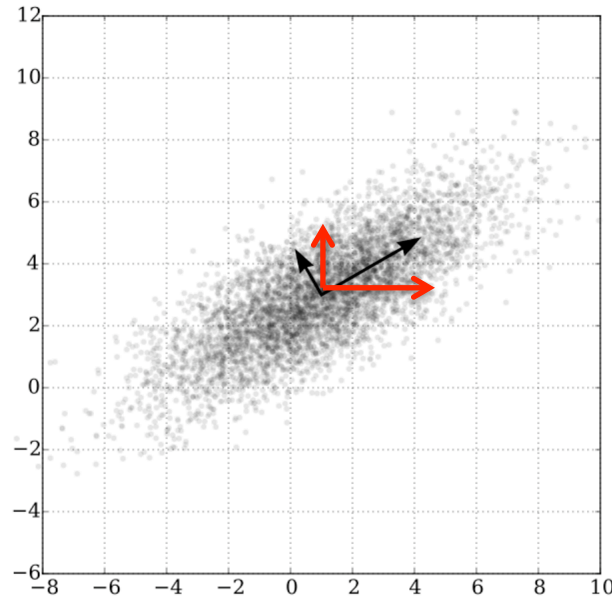
Word count
matrix:

Co-appearing
words

		Documents						
Words		0	0	0	0	1	0	0
		0	1	0	0	0	0	0
		0	0	0	0	0	0	0
		0	0	0	0	0	0	0
		0	0	1	0	0	0	0
	politics	2	0	2	0	0	1	1
	politician	1	0	4	0	0	0	1
	govern	1	0	3	0	0	1	0
	parliament	1	0	2	0	0	3	1
		0	0	0	1	0	0	0
		0	0	0	0	1	0	0
	president	1	0	1	0	0	2	1

Could these co-appearing words
could be represented by one
dimensional latent variable?

Principal component analysis (PCA)



- correlated variables in old coordinate system
- linearly uncorrelated variables in new (transformed) coordinate system

PCA is related to spectral clustering:

Principal component analysis is the eigen-decomposition of a covariance matrix, while spectral clustering is related to eigen-decomposition of a Laplacian (or related) matrix.

Principal component analysis (PCA)

In matrix form, the empirical covariance matrix for the original variables can be written

$$\mathbf{Q} \propto \mathbf{X}^T \mathbf{X} = \mathbf{W} \mathbf{\Lambda} \mathbf{W}^T$$

The empirical covariance matrix between the principal components becomes

$$\mathbf{W}^T \mathbf{Q} \mathbf{W} \propto \mathbf{W}^T \mathbf{W} \mathbf{\Lambda} \mathbf{W}^T \mathbf{W} = \mathbf{\Lambda}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues $\lambda_{(k)}$ of $\mathbf{X}^T \mathbf{X}$

- After decomposition, we typically choose the principal components of the new coordinate system that store the most information about the observed data
- Each consecutive principal component captures the consecutive strongest correlations across dimensions

Principal component analysis (PCA)

- After decomposition, we typically choose the principal components of the new coordinate system that store the most information about the observed data
- Each consecutive principal component captures the consecutive strongest correlations across dimensions

Latent Dirichlet Allocation (LDA)

LDA

Latent Dirichlet allocation (LDA)

- ① Introduction to LDA
- ② The posterior distribution for LDA

Approximate posterior inference

- ① Gibbs sampling
- ② Variational inference
- ③ Comparison/Theory/Advice

Other topic models

- ① Topic models for prediction: Relational and supervised topic models
- ② The logistic normal: Dynamic and correlated topic models
- ③ “Infinite” topic models, i.e., the hierarchical Dirichlet process

Interpreting and evaluating topic models

Probabilistic modeling

- ① Treat data as observations that arise from a generative probabilistic process that includes hidden variables
 - For documents, the hidden variables reflect the thematic structure of the collection.
- ② Infer the hidden structure using *posterior inference*
 - What are the topics that describe this collection?
- ③ Situate new data into the estimated model.
 - How does this query or new document fit into the estimated topic structure?

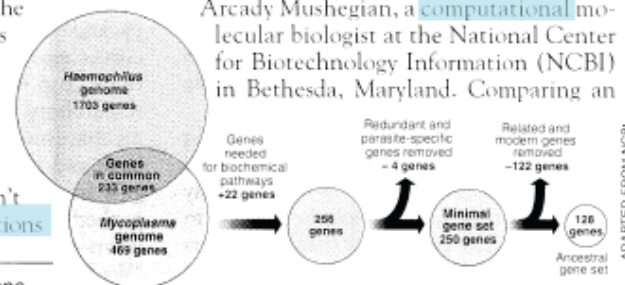
Intuition behind LDA

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



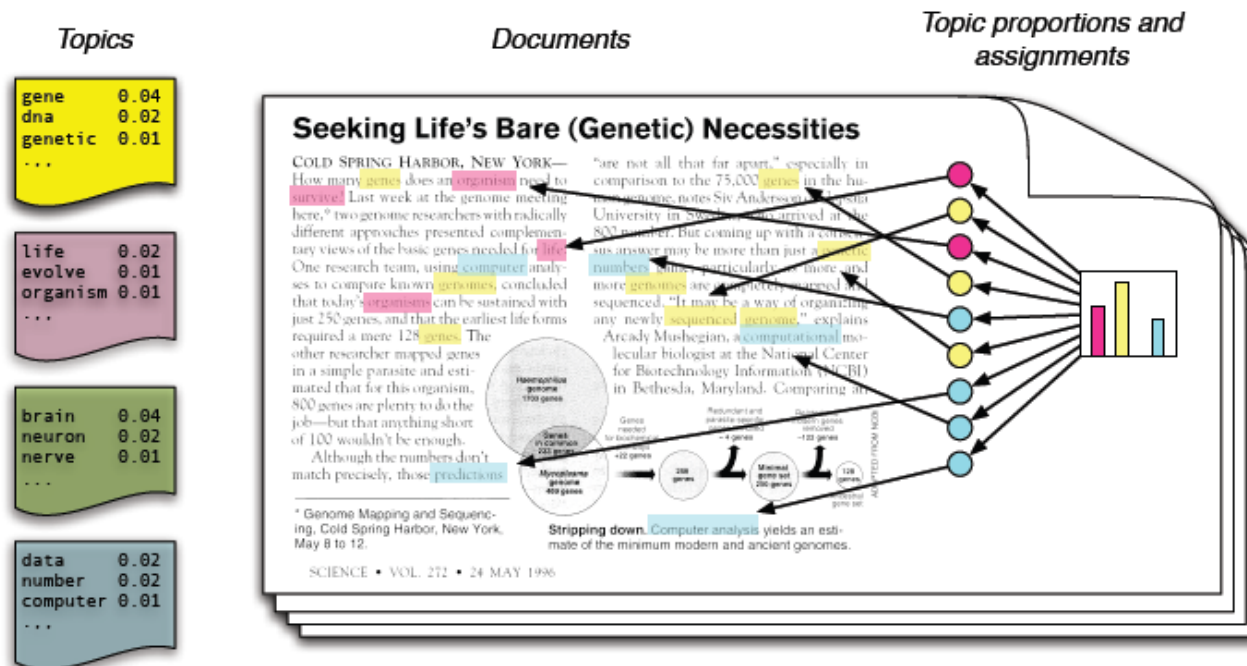
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

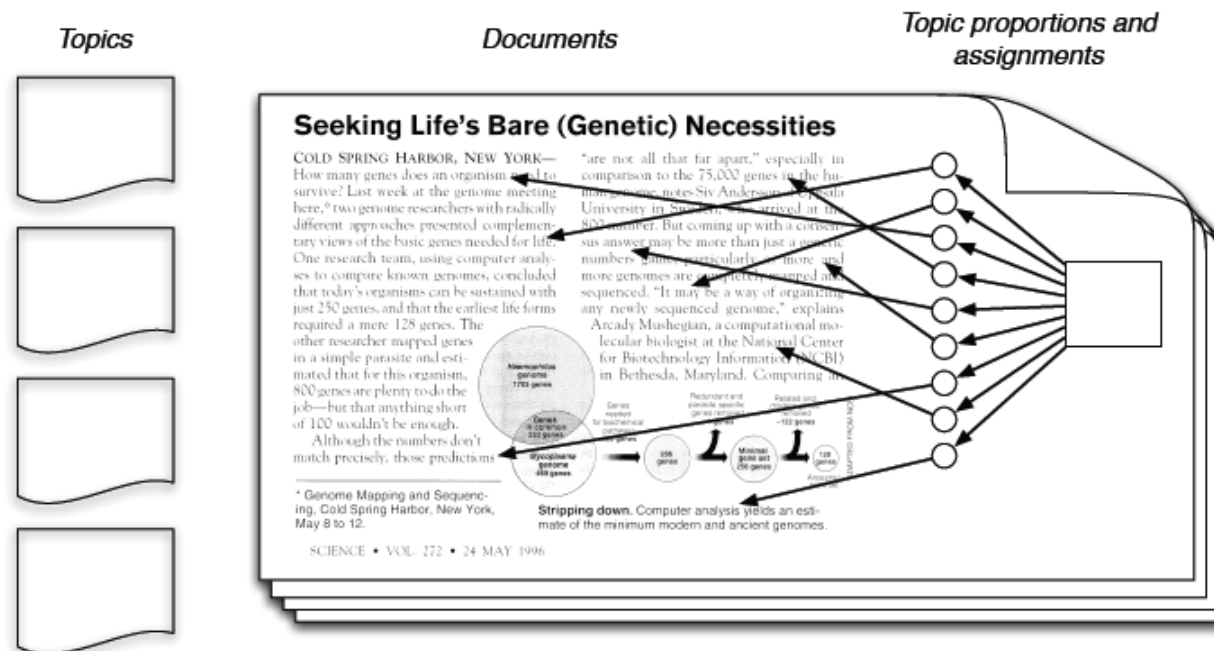
Simple intuition: Documents exhibit multiple topics.

Generative model



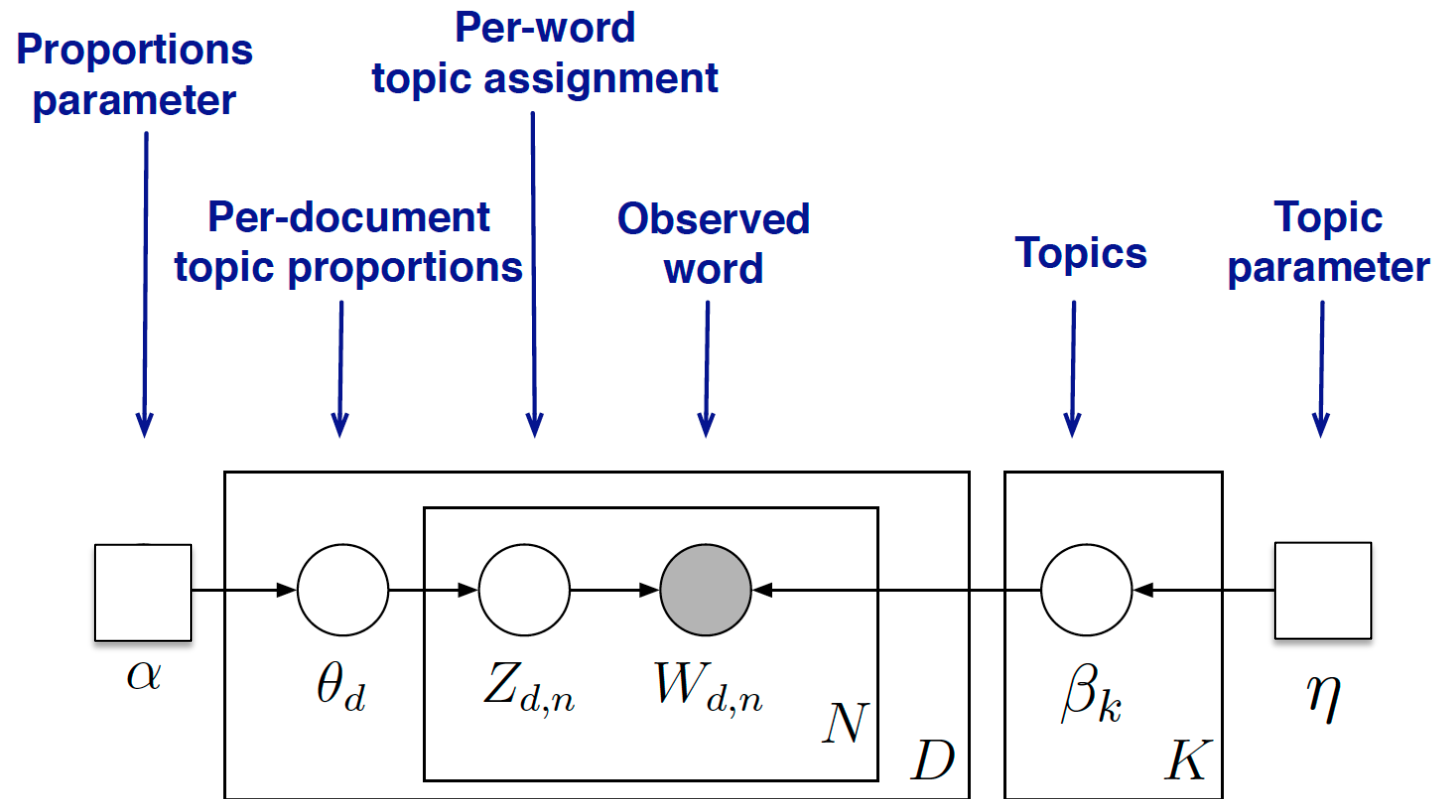
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

The posterior distribution



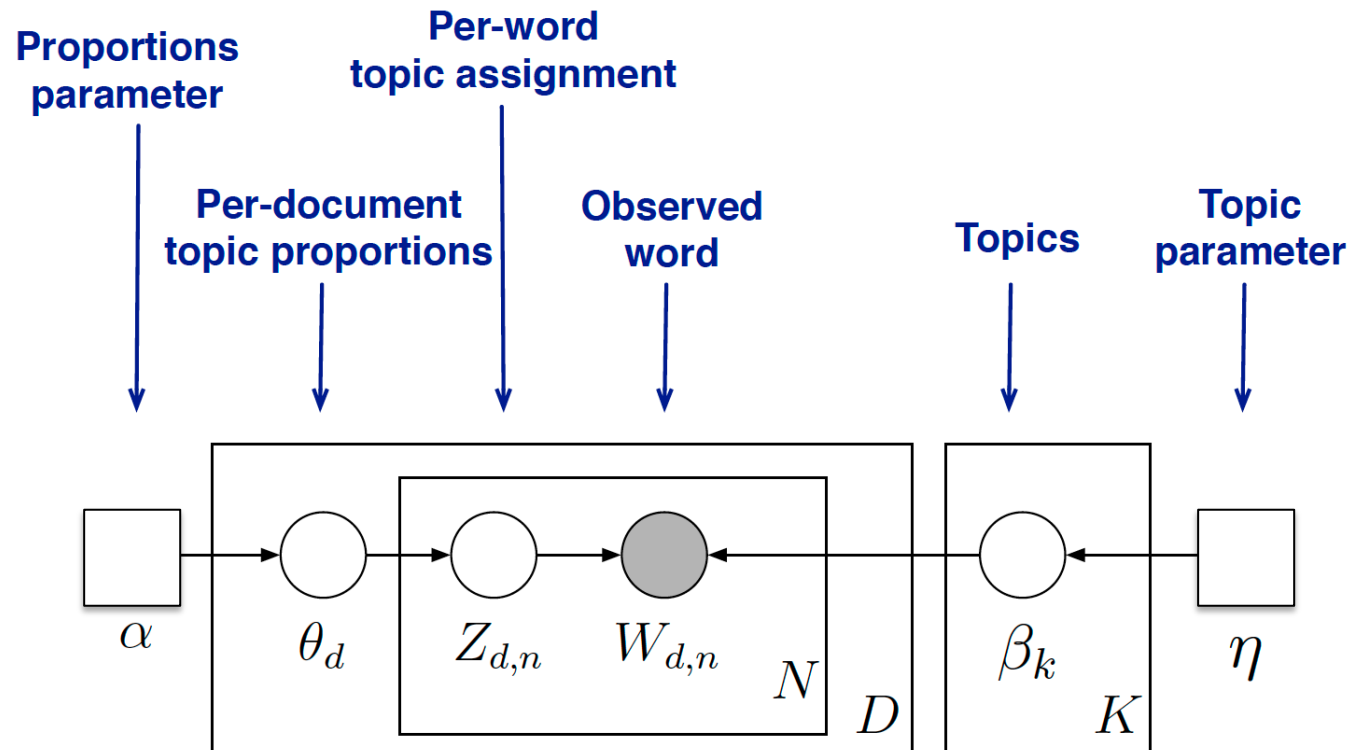
- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

LDA generative model



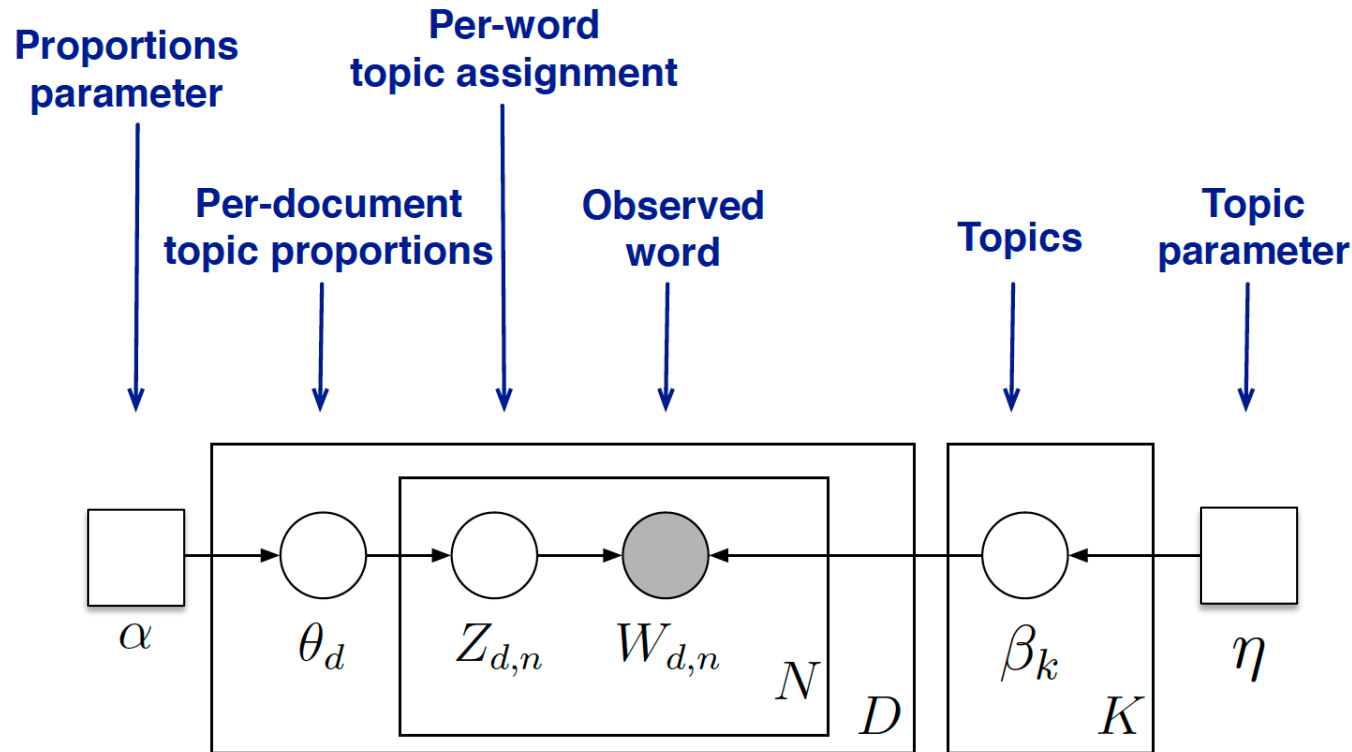
- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

LDA generative model



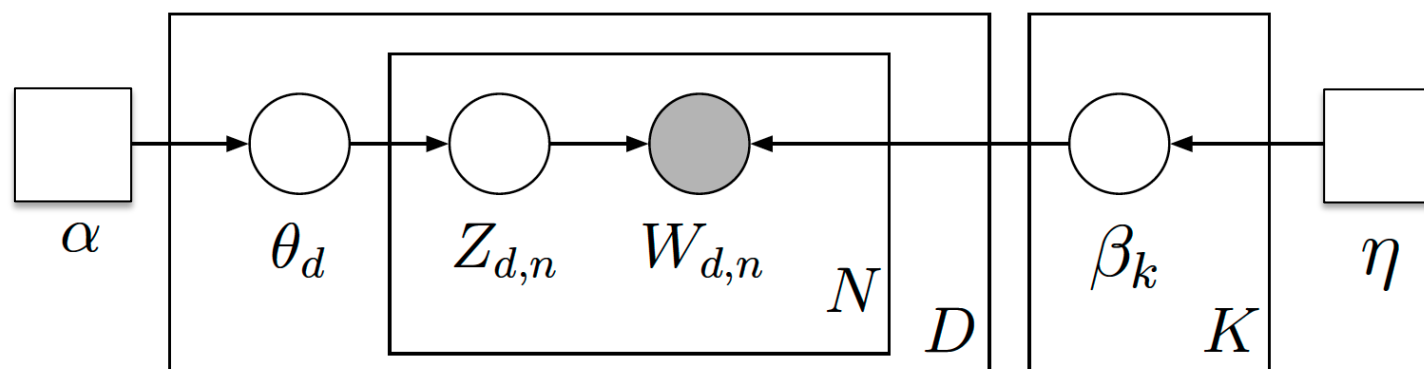
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

LDA generative model



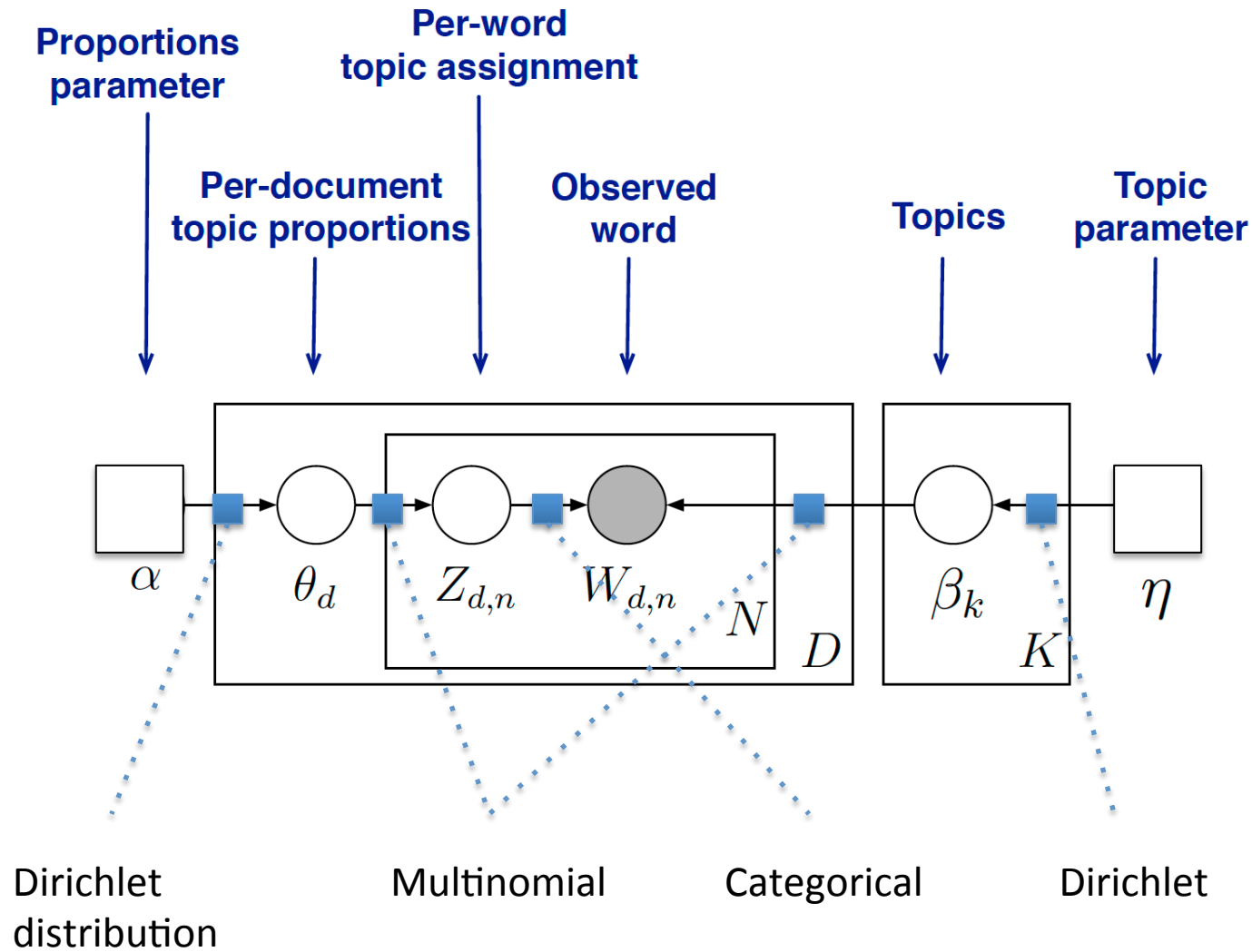
$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^K p(\beta_i | \eta) \right) \left(\prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

LDA

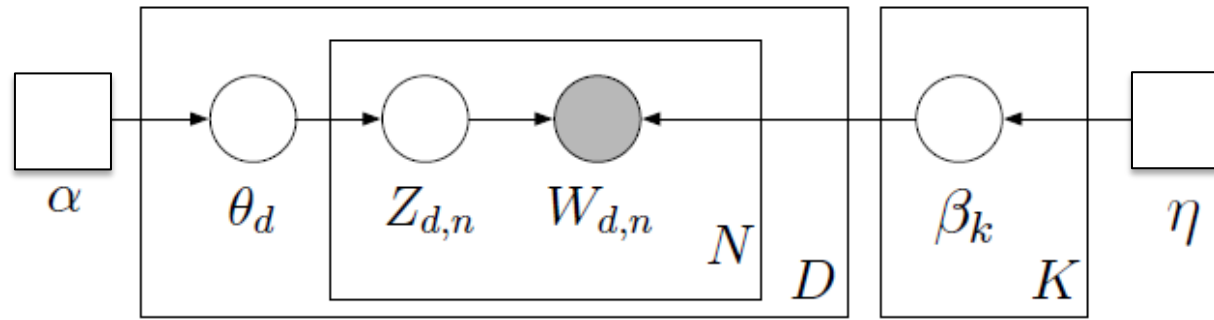


- This joint defines a posterior, $p(\theta, z, \beta | w)$.
- From a collection of documents, infer
 - Per-word topic assignment $z_{d,n}$
 - Per-document topic proportions θ_d
 - Per-corpus topic distributions β_k**Main output**
- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.

LDA generative model



Inference in LDA



Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)

For comparison, see Mukherjee and Blei (2009) and Asuncion et al. (2009).

Inference comparison

- Conventional wisdom says that:
 - Gibbs is easiest to implement
 - Variational can be faster, especially when dealing with nonconjugate priors (more on that later)
- There are other options:
 - Collapsed variational inference
 - Parallelized inference for large corpora
 - Particle filters for on-line inference
- An ICML paper examining these issues is Asuncion et al. (2009).

Libraries

- Mallet (Java)
- Stanford TMT (Java/Scala)
- Gensim (Python)
- scikit-learn (Python)
- ...

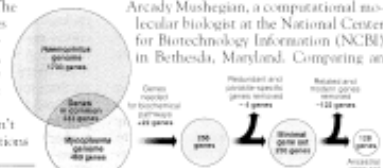
Example inference

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 number 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



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

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

SCIENCE • VOL. 272 • 24 MAY 1996

- **Data:** The OCR'ed 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 inference

Seeking Life's Bare (Genetic) Necessities

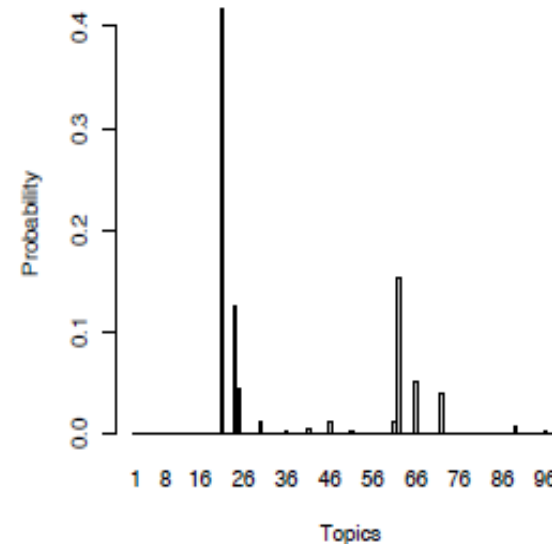
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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.

Topics vs 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

Explore and browse document collections

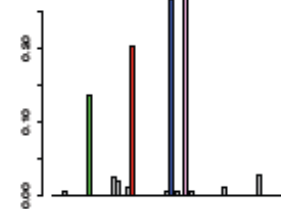
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



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

Why does LDA “work” ?

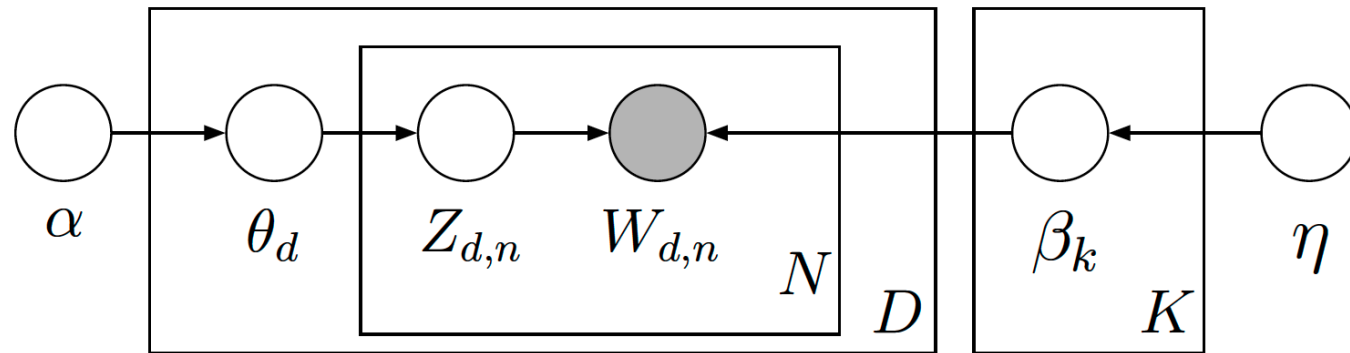
Why does the LDA posterior put “topical” words together?

- Word probabilities are maximized by dividing the words among the topics. (More terms means more mass to be spread around.)
- In a mixture, this is enough to find clusters of co-occurring words.
- In LDA, the Dirichlet on the topic proportions can encourage sparsity, i.e., a document is penalized for using many topics.
- Loosely, this can be thought of as softening the strict definition of “co-occurrence” in a mixture model.
- This flexibility leads to sets of terms that more tightly co-occur.

Why does LDA “work” ?

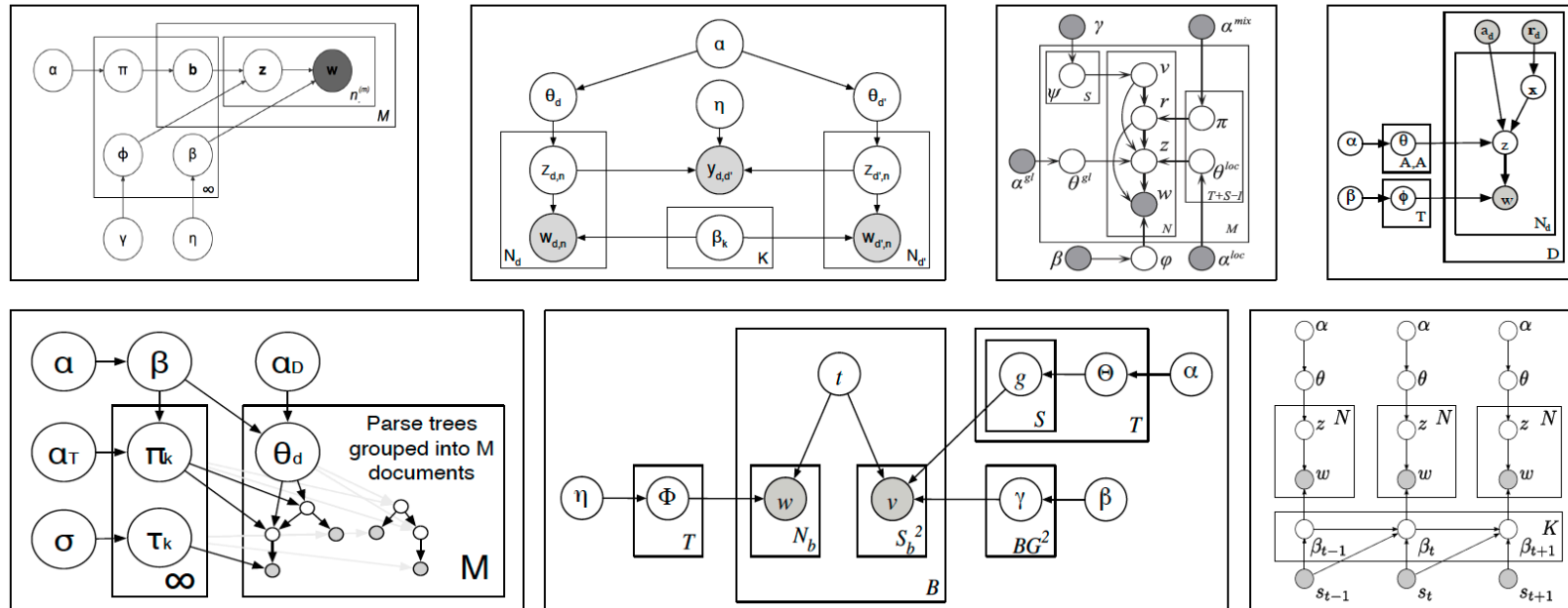
- LDA trades off two goals.
 - ① For each document, allocate its words to as few topics as possible.
 - ② For each topic, assign high probability to as few terms as possible.
- These goals are at odds.
 - Putting a document in a single topic makes #2 hard:
All of its words must have probability under that topic.
 - Putting very few words in each topic makes #1 hard:
To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

LDA summary



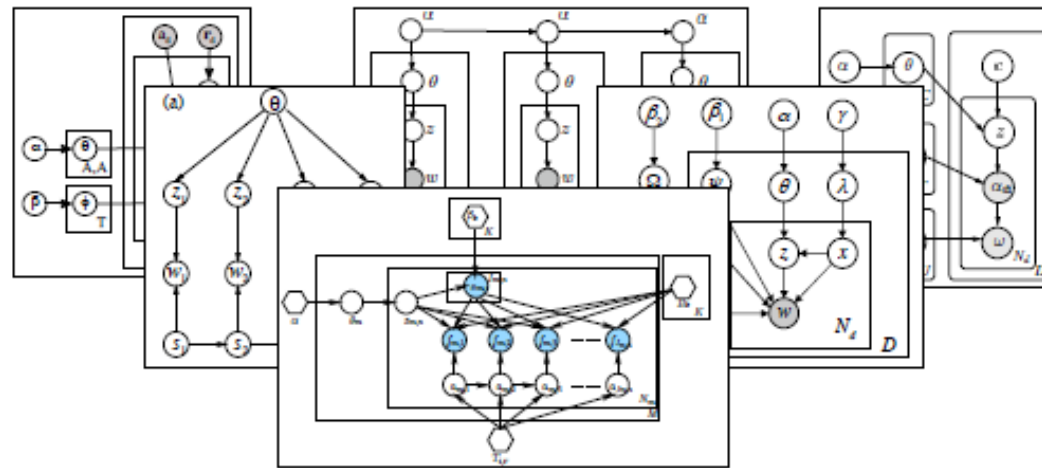
- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a posterior inference problem.
- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.
- Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999)
It is a mixed-membership model (Erosheva, 2004).
It relates to PCA and matrix factorization (Jakulin and Buntine, 2002).
Was independently invented for genetics (Pritchard et al., 2000)

LDA summary



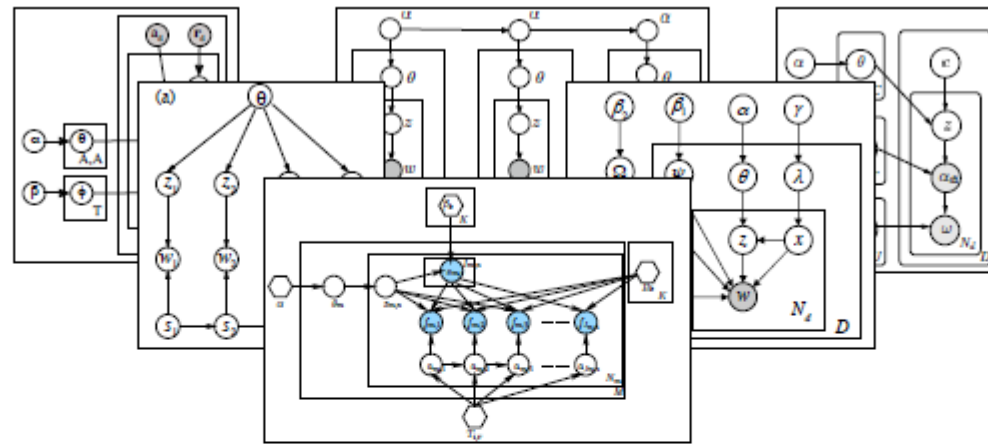
- LDA is a simple building block that enables many applications.
- It is popular because organizing and finding patterns in data has become important in the sciences, humanities, industry, and culture.
- Further, algorithmic improvements let us fit models to massive data.

LDA is modular, general, useful



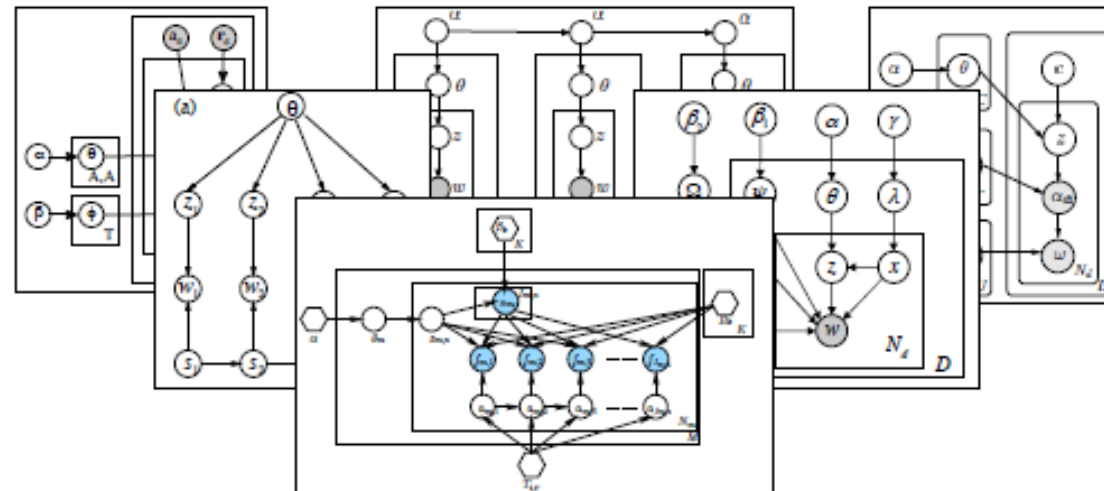
- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
- E.g., syntax; authorship; word sense; dynamics; correlation; hierarchies; nonparametric Bayes

LDA is modular, general, useful



- The **data generating distribution** can be changed.
- E.g., images, social networks, music, purchase histories, computer code, genetic data, click-through data; ...

LDA is modular, general, useful



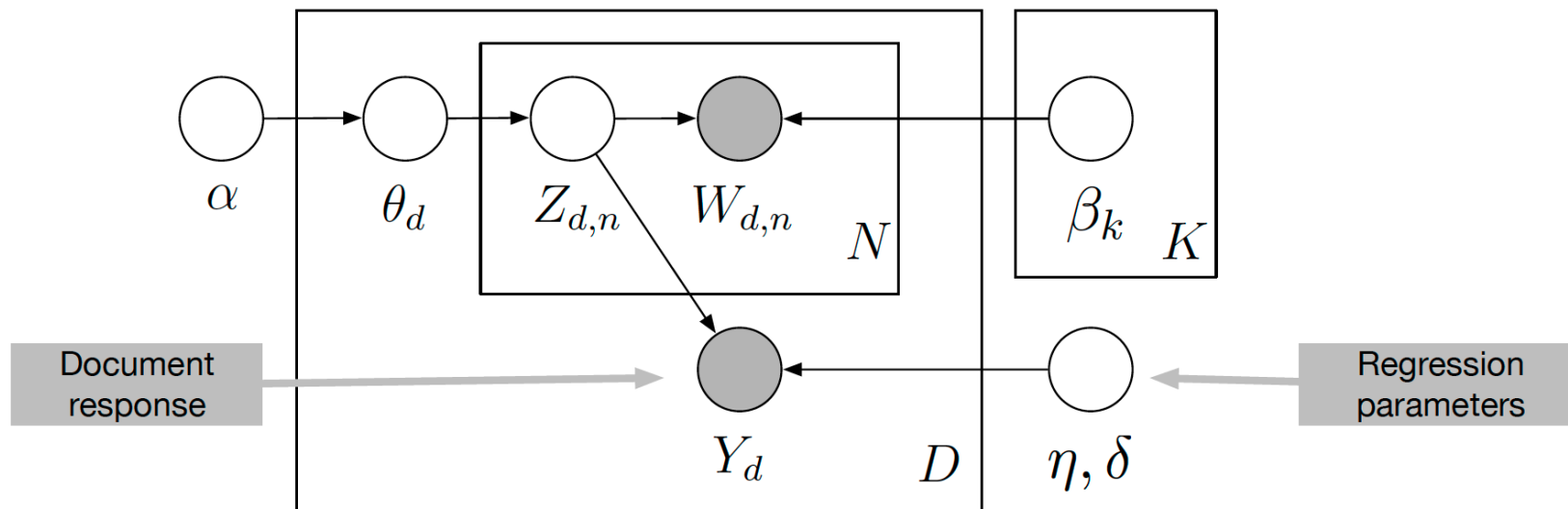
- The **posterior** can be used in creative ways
- E.g., IR, collaborative filtering, document similarity, visualizing interdisciplinary documents

Other LDA-like models

Supervised LDA

- LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with **response variables**.
 - User reviews paired with a number of stars
 - Web pages paired with a number of “likes”
 - Documents paired with links to other documents
 - Images paired with a category
- **Supervised LDA** are topic models of documents and responses. They are fit to find topics predictive of the response.

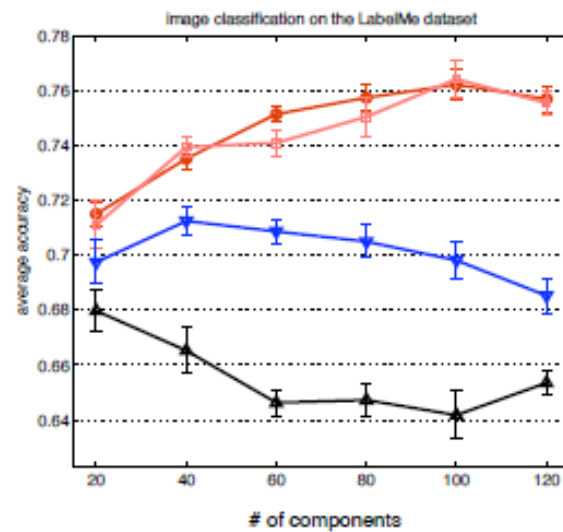
Supervised LDA



- 1 Draw topic proportions $\theta \mid \alpha \sim \text{Dir}(\alpha)$.
- 2 For each word
 - Draw topic assignment $z_n \mid \theta \sim \text{Mult}(\theta)$.
 - Draw word $w_n \mid z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
- 3 Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \text{N}(\eta^\top \bar{z}, \sigma^2)$, where

$$\bar{z} = (1/N) \sum_{n=1}^N z_n.$$

Example: Multi class classification



highway
car, sign, road



inside city
buildings, car, sidewalk



street
tree, car, sidewalk



tall building
trees, buildings
occluded, window

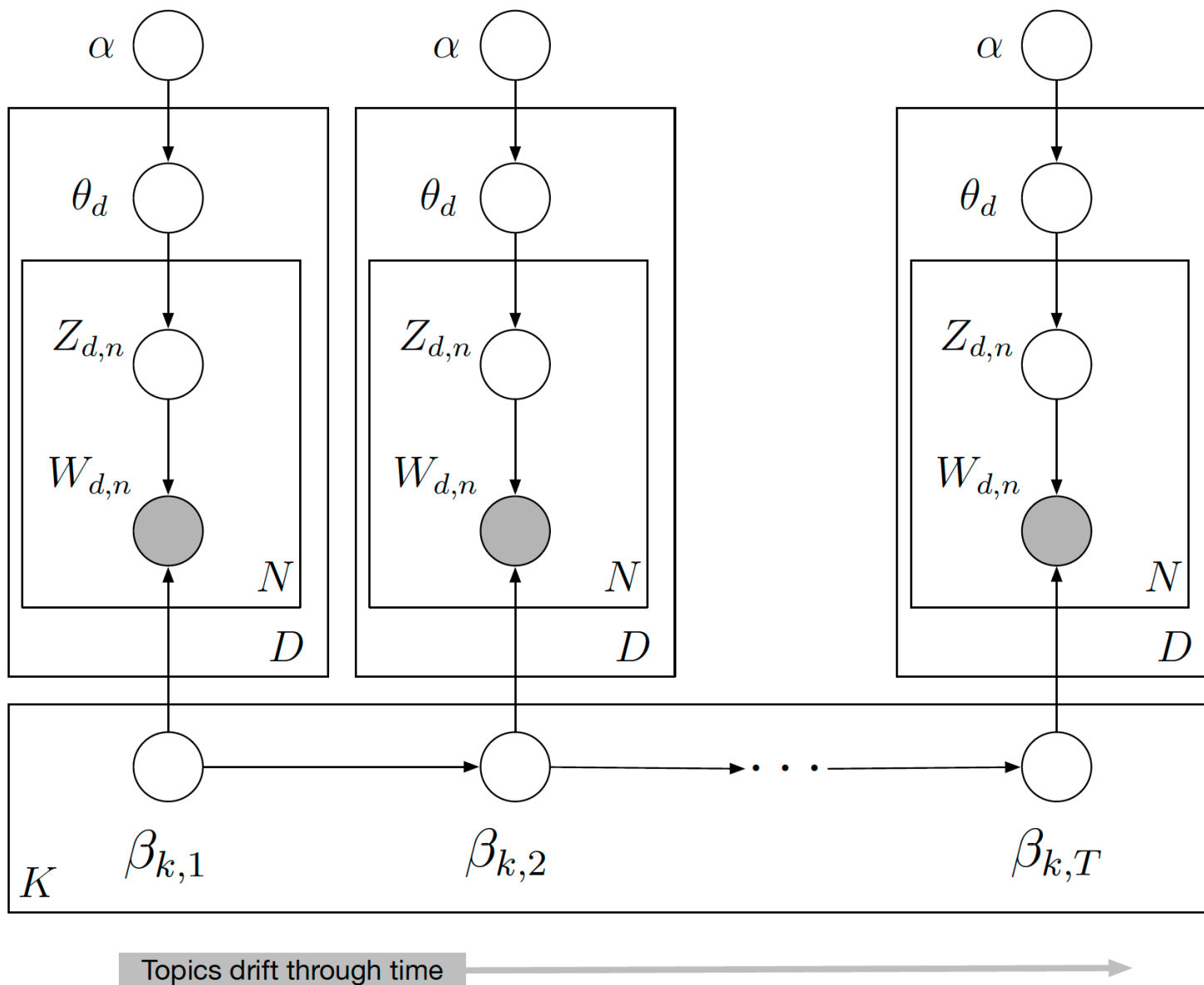


SLDA for image classification (with Chong Wang, CVPR 2009)

Supervised topic models

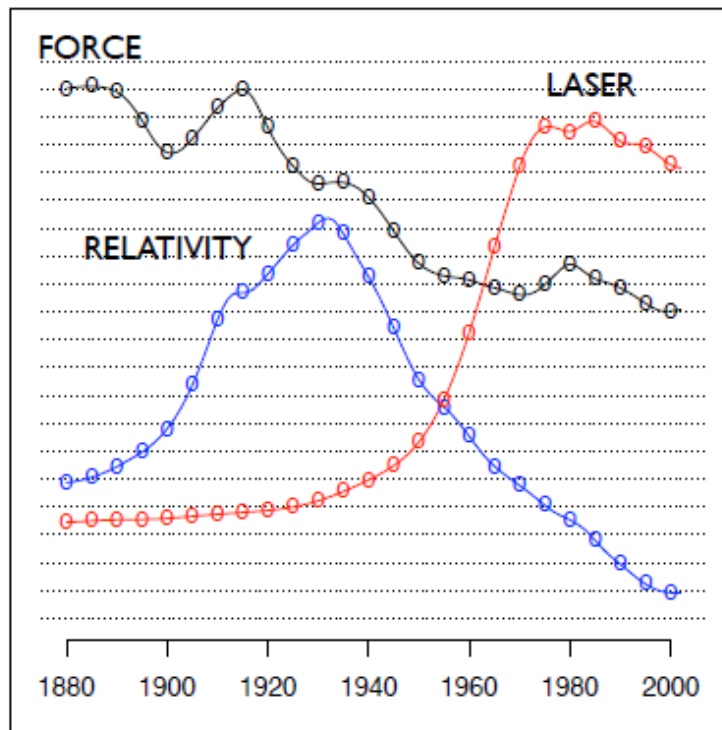
- SLDA enables model-based regression where the predictor “variable” is a text document.
- It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).
- SLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.
- LDA + regression compared to sLDA is like principal components regression compared to partial least squares.

Dynamic topic models

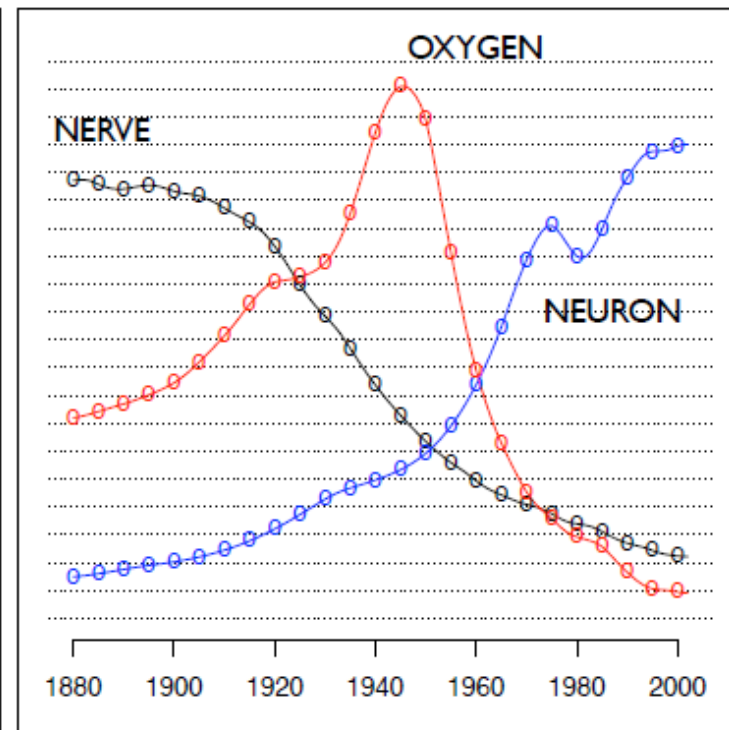


Dynamic topic models

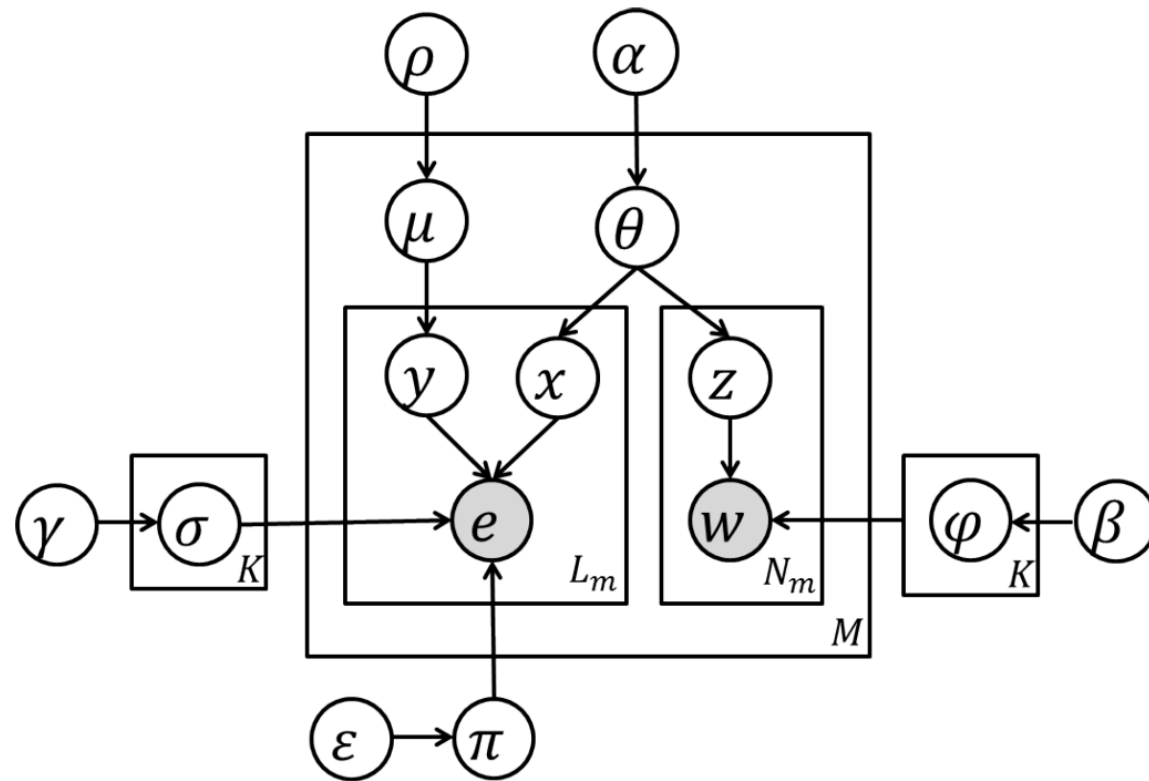
"Theoretical Physics"



"Neuroscience"

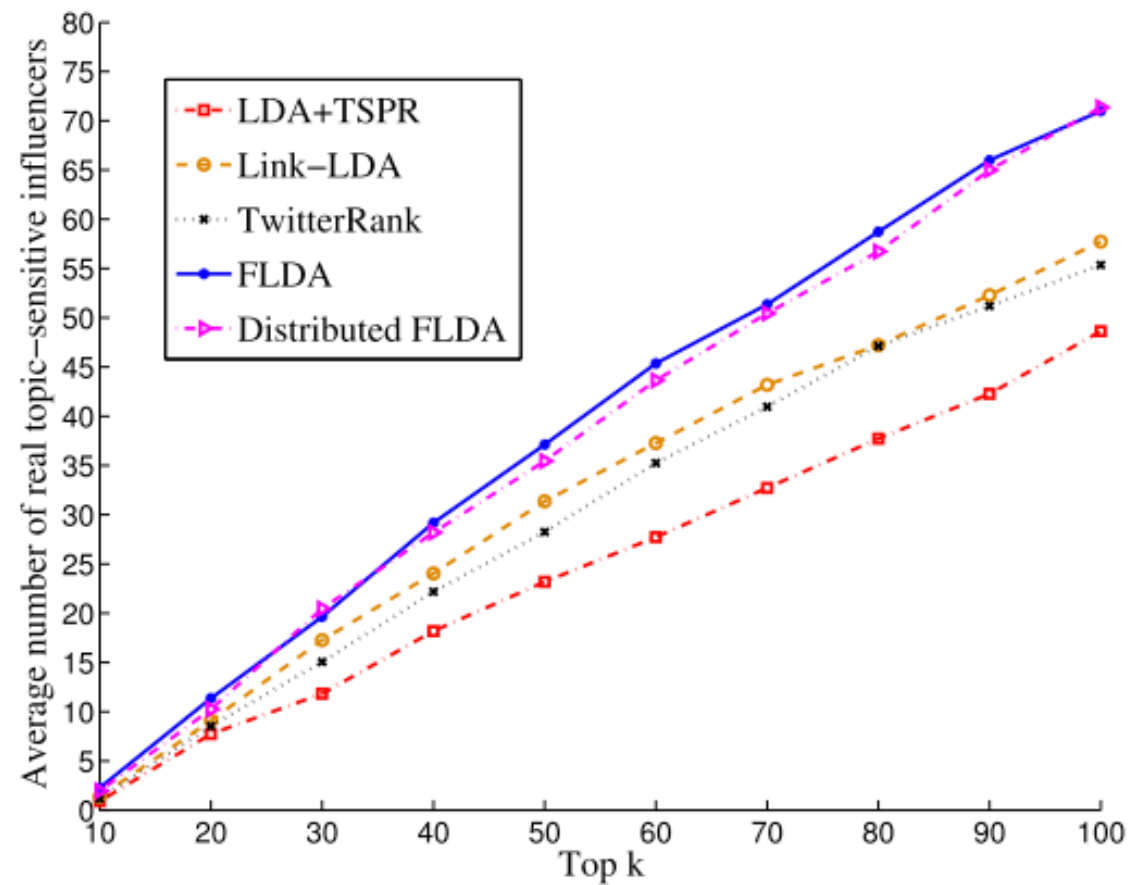


In social media: Followship LDA



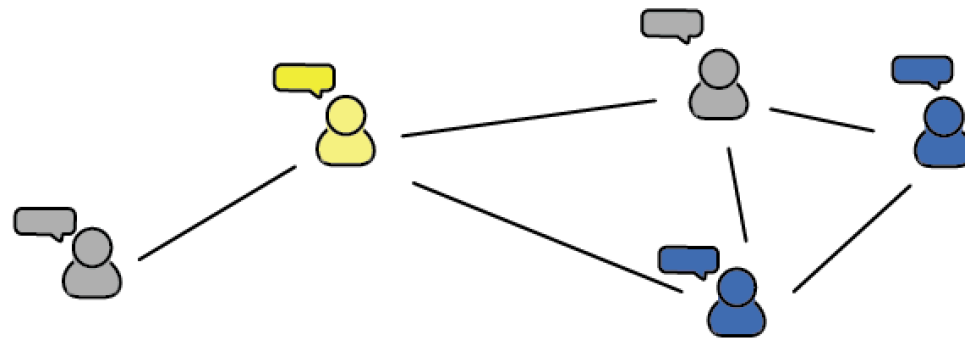
Bi, B., Tian, Y., Sismanis, Y., Balmin, A., & Cho, J. (2014). Scalable topic-specific influence analysis on microblogs.






In social media: Followship LDA



In social media: impact on information diffusion

People spread the information that matches their **topical interests**



some ( ) have strong topical interests, others () don't
some memes () have strong topical meaning, other () don't

In social media: impact on information diffusion

The **most topical** hashtags

Hashtag	Topic	Examples of tweets hand-picked from 30 chronological tweets from the beginning of June 2009
#ashraf	Iran election	<ol style="list-style-type: none"> 1. canadian friends of a democratic iran urge us president to ensure ashraf residents treated according international law #iran #ashraf 2. suppressive forces heavily present in streets of iran capital to prevent protests #iran #iranelection #tehran #neda #rajavi #ashraf 3. how u can help iranians http://tinyurl.com/m8fun4 #iranelection #tehran #azadi #ashraf #rajavi #neda #basij #iran #gr88 rt rt rt
#electionstudio	British election	<ol style="list-style-type: none"> 1. bnp is a left wing party. so were the german national socialists. #electionstudio 2. #electionstudio this bnp guy is a lunatic frankly. indigenous population? if it weren't for immigration britain would not exist at all. 3. #electionstudio i got very excited last night when purnell resigned but now can't see a leadership challenge coming from anywhere
#vene2ia	Well-being	<ol style="list-style-type: none"> 1. without freedom of choice there is no creativity kirk, star trek #vene2ia 2. it wasn't until quite late in life that i discovered how easy it is to say "i don't know!" somerset maugham #vene2ia (via @vene2ia) 3. rt @vene2ia: children are the living messages we send to a time we will not see john w. whitehead #vene2ia
#smum	Viral marketing	<ol style="list-style-type: none"> 1. learn how to blog your way to success at social media university milwaukee http://ow.ly/g83u (last day for early bird pricing) #smu 2. lrt @philgerb: rt @dnewman:great article on benefits of audience tweeting during your presentation http://bit.ly/rdgsl good 2 know 3. l@backseatgaffer glad you found it. you can sign up even if you're not going to #smum. details: http://bit.ly/3qyhot reg: http://bit.ly/3qyhot
#mytouch	Smart devices	<ol style="list-style-type: none"> 1. mytouch3g ordering starts tomorrow, very excited #tmobile #android #mytouch3g 2. im thinking of getting the tmobile #mytouch phone, i want to know, what do you all think of the phone? should i get it? 3. if you're using a #bold, #curve, #g1, #mytouch or an #iphone you should download @loopt an amazing lbs networking app

politics, innovation, ...

In social media: impact on information diffusion

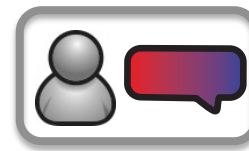
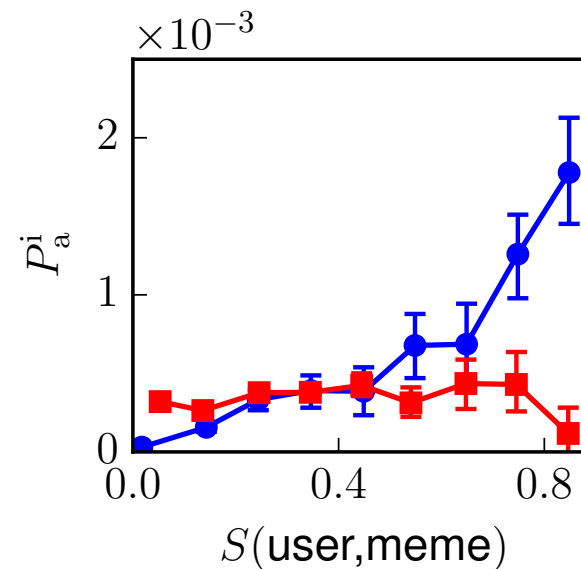
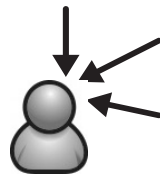
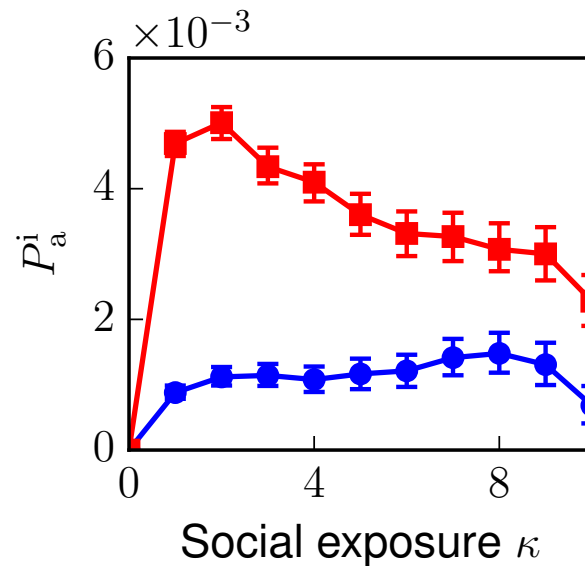
The **least topical** hashtags

Hashtag	Topic	Examples of tweets hand-picked from 30 chronological tweets from the beginning of June 2009
#uknowyouugly	-	1. #uknowyouugly when you're at the zoo and the monkeys run from you 2. #uknowyouugly when u gotta get fresh to go to the grocery store 3. #uknowyouugly when they ban you from the mac store
#unala2009	-	1. getting ready for my first unconference... the first one might as well be a big one! very excited. #ala2009 #unala2009 2. got new business cards designed by my son. unconference tag #unala2009 per conference wiki. ala facebook meetup: http:// 3. just for clarification, friday's unconference is at the chicago hilton, not the palmer house hilton. hth #ala2009 #unala2009
#skiphop	-	1. i just entered to #win a #skiphop bag with a #beautyfix kit in the #99daysofsummer #contest at #modernmom! http://bit.ly 2. modernmom i just entered to #win a #skiphop bag with a #beautyfix kit in the #99daysofsummer #contest at #modernmom! 3. i just entered to #win a #skiphop bag with a #beautyfix kit in the #99daysofsummer #contest at #modernmom! http://bit.ly
#sexisbetter	-	1. #sexisbetter wen your drunk 2. #sexisbetter with a scorpio! (guess what sign i am) ;-) 3. #sexisbetter when its anywhere but a bed
#besthead	-	1. #besthead when dey don't expect it baq.. lol 2. #besthead is when u black out and just pass out now dats da best 3. #besthead was in the morning right after she brushed her teeth,she went to work!! u heard me (master p voice)

idioms, neologisms, ...

Impact on information diffusion

- The probability that a user **κ -exposed** to a meme will adopt it
- The probability that a user **topically aligned** to a meme with **$S(\text{user}, \text{meme})$** will adopt this meme

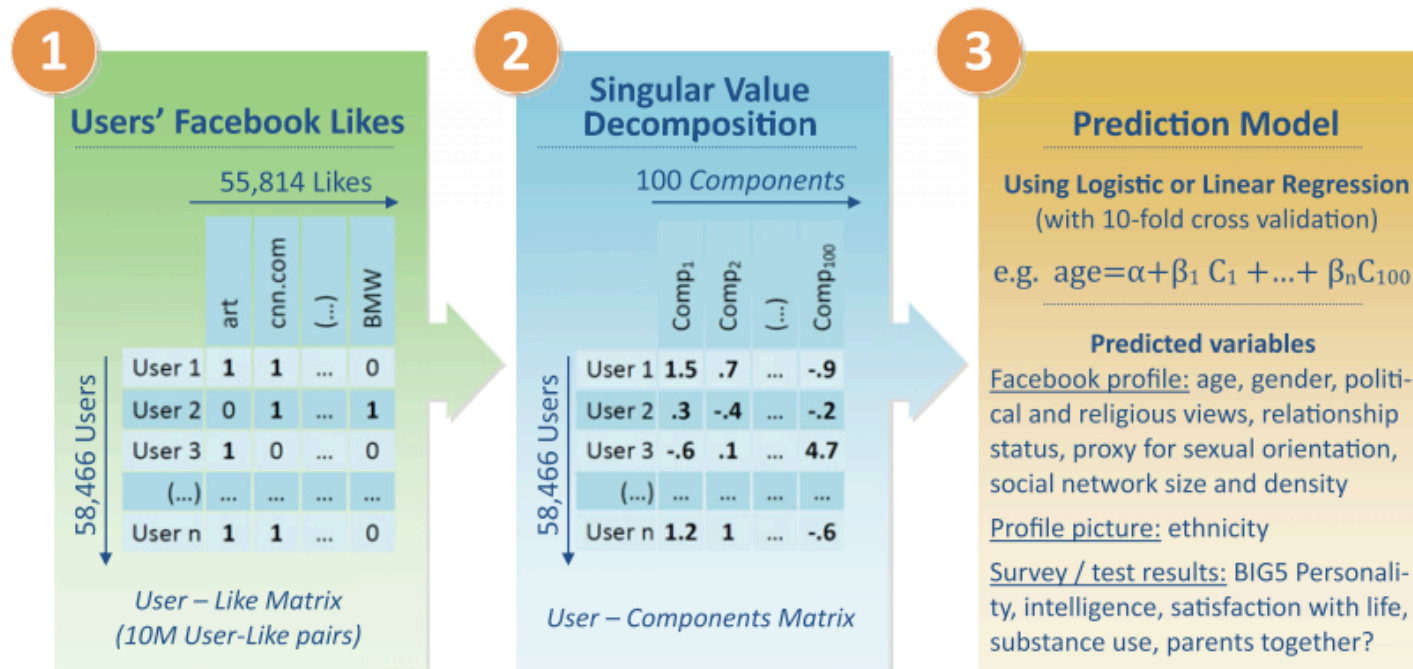


**Topical
memes**
(top 25%)

**Non-topical
memes**
(bottom 25%)

Grabowicz, P., Ganguly, N., & Gummadi, K. P. (2016). Distinguishing between Topical and Non-topical Information Diffusion Mechanisms in Social Media.

Other unsupervised methods: SVD (PCA) (in combination with supervised)



Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 2–5. doi:10.1073/pnas.1218772110

SVD (PCA) + supervised classifier

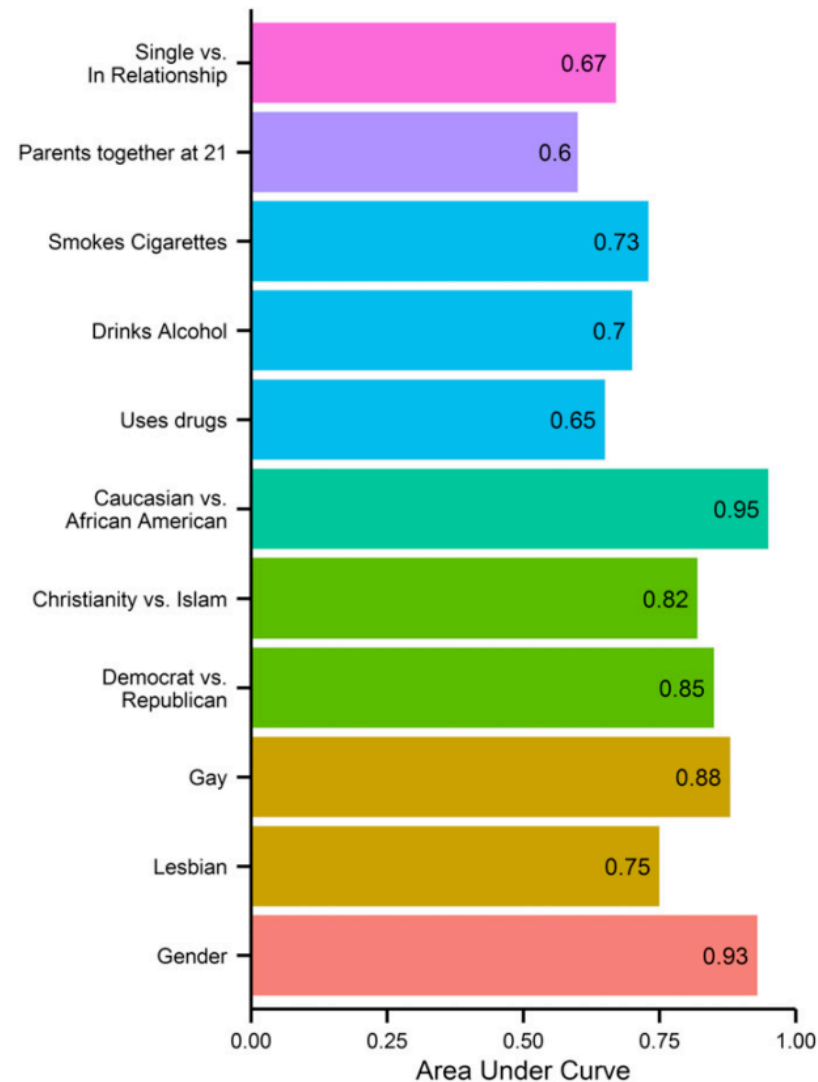


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

Connection to ML research

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

Probabilistic graphical models and topic models

Sources:

- “Topic models”, David Blei, MLSS ’09
http://videolectures.net/mlss09uk_blei_tm/?q=david%20blei
- Parts of “Probabilistic graphical models”, Christopher Bishop, MLSS’13
<https://www.youtube.com/watch?v=ju1Grt2hdko>
- Parts of “Machine learning: Graphical models”, Alan Smola,
<http://alex.smola.org/teaching/10-701-15/graphical.html>