Probabilistic graphical models and topic models

Sources:

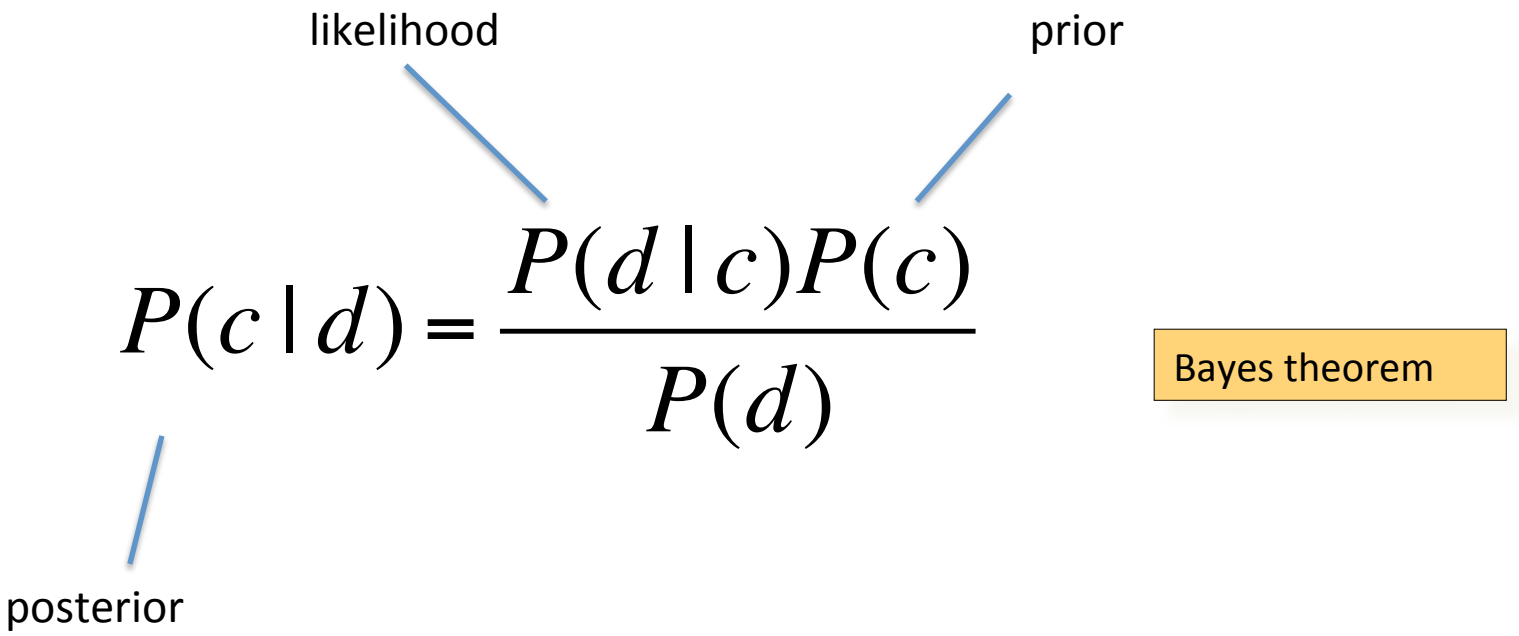
- "Topic models", David Blei, MLSS '09
  http://videolectures.net/mlss09uk_blei_tm/?q=david%20blei
- Parts of "Probabilistic graphical models", Christopher Bishop, MLSS'13
  https://www.youtube.com/watch?v=ju1Grt2hdko
- Parts of "Machine learning: Graphical models", Alex Smola,
  http://alex.smola.org/teaching/10-701-15/graphical.html

Supervised methods for text classification:

Naïve Bayes

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.



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.

$$P(d \mid c) = P(x_1, x_2, \ldots, x_n \mid c)$$

Document *d* represented as features $x_1,\ldots,x_n$

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

Independence Assumptions

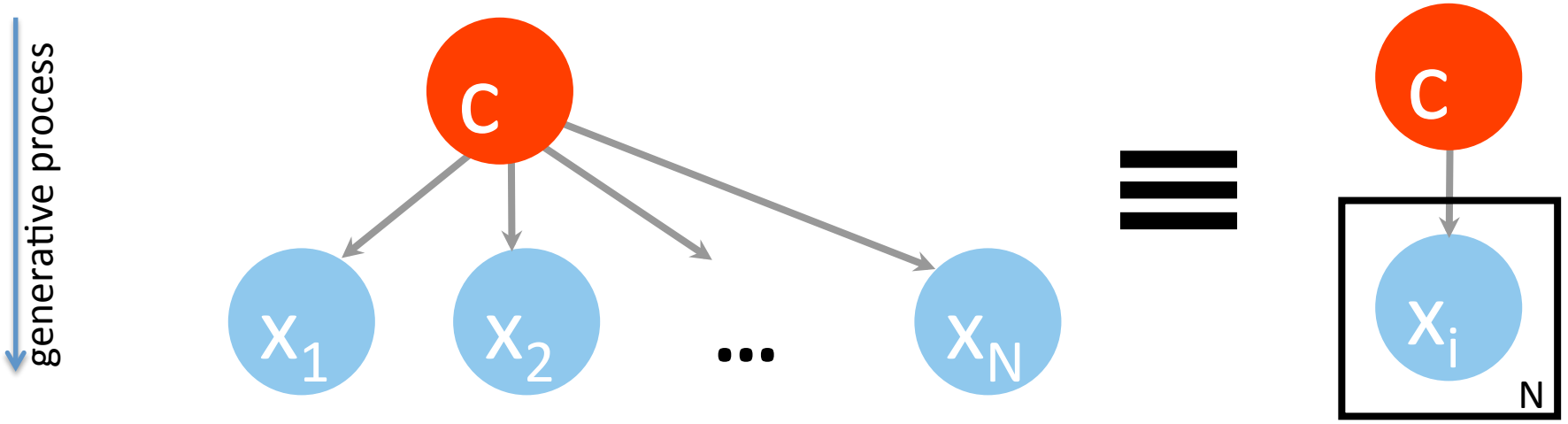
- **Bag of words assumption**: Assume that position of words doesn't matter (the **exchangeability** of random variables)

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

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

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c) = \prod_{i \in V} P(x_i \mid c)$$

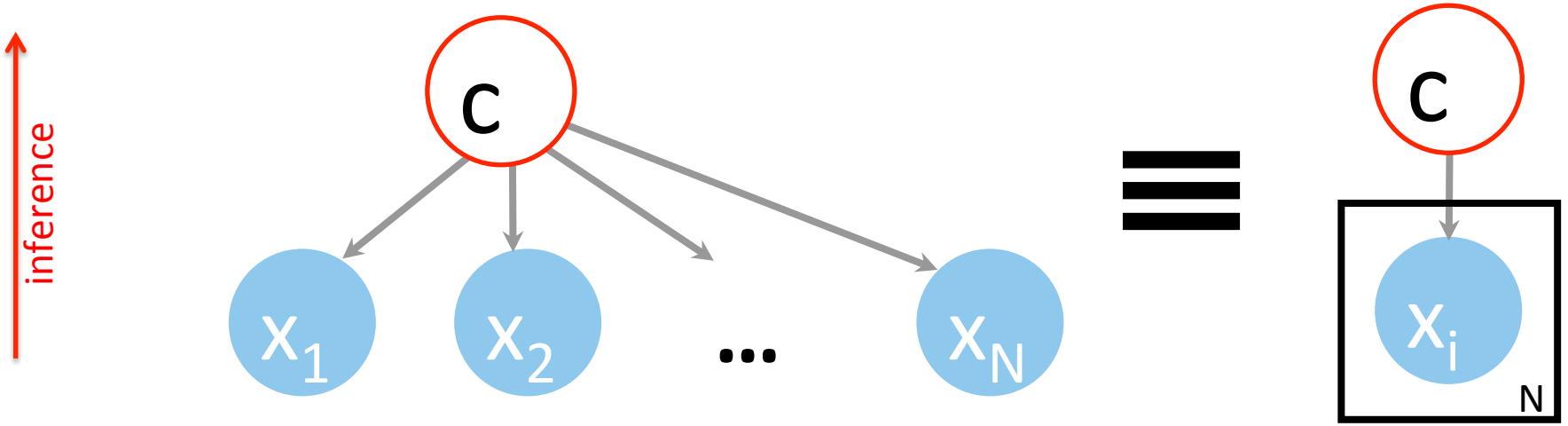
Graphical model for Naïve Bayes (factorization describing generative process)



Factorization of the joint probability:

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

Graphical model for Naïve Bayes (inference)



Inference by finding maximum a posteriori:

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

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}}\, P(c)\prod_{i \in V} P(x_i \mid c)$$

Typical choices of the likelihood P(x|c) (Assumptions of the generative model)

Multinomial Naïve Bayes classifier:
*P(**x**|c) = Multinomial(**x**|c)*

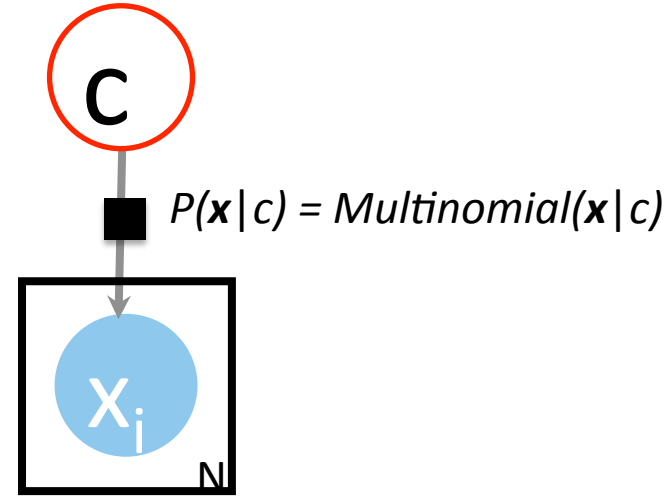
$$\propto \prod_{i \in V} \left(p_{ic} \right)^{x_i}$$

Bernoulli Naïve Bayes classifier:
*P(**x**|c) = MultivariateBernoulli(**x**|c)*

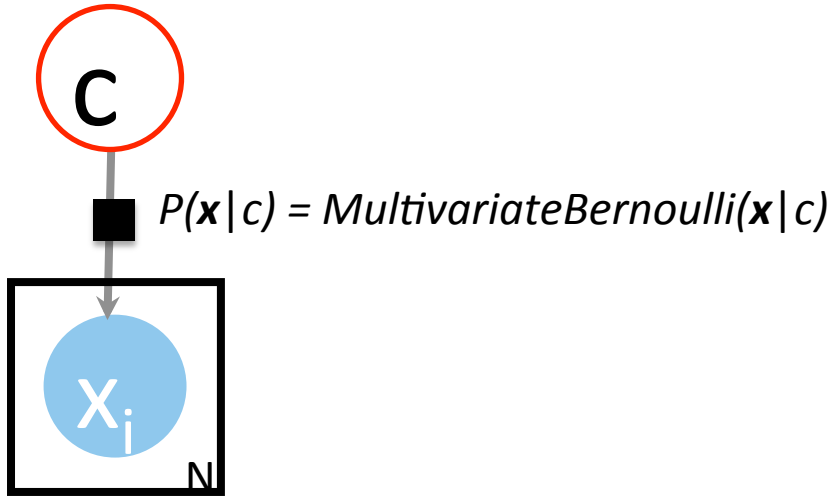
$$= \prod_{i \in V} \left(p_{ic} \right)^{x_i} \left(1 - p_{ic} \right)^{1 - x_i}$$

Typical choices of the likelihood P(x|c) (Assumptions of the generative model)

Multinomial Naïve Bayes classifier:

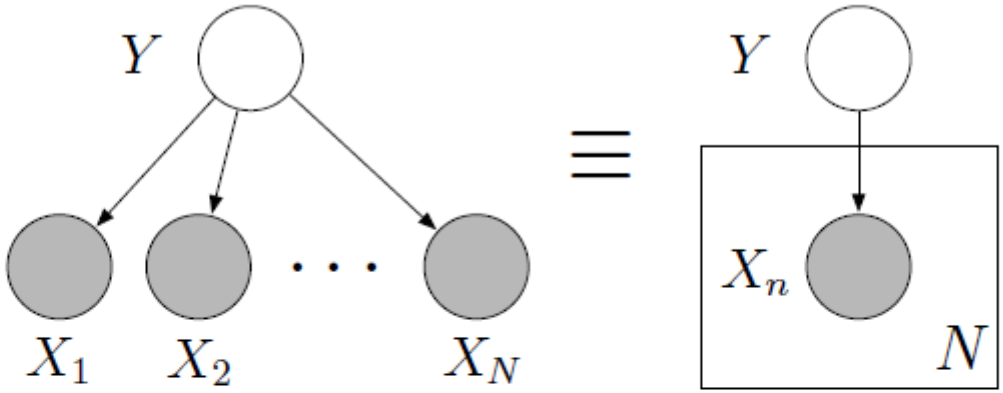


Bernoulli Naïve Bayes classifier:



Graphical models

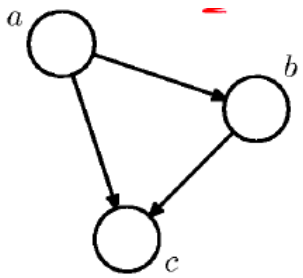
Graphical models (summary)



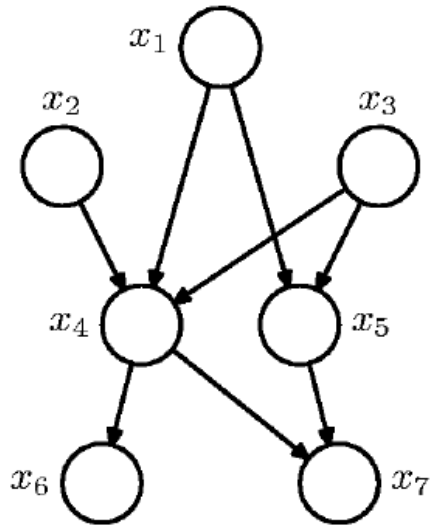
- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure
- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

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

Graphical models



$$P(a,b,c) = P(a)P(b \mid c)P(c \mid a,b)$$



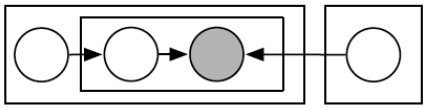
$$P(x_1,x_2,x_3,x_4,x_5,x_6,x_7) = \ldots$$

- 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

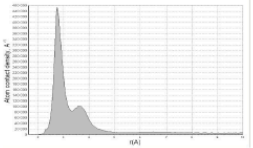
Probabilistic graphical models

Make assumptions

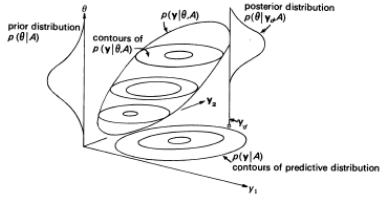


Collect data

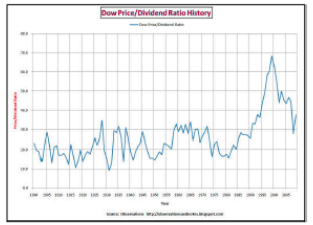
Infer the posterior



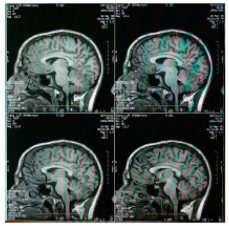
Check



Predict

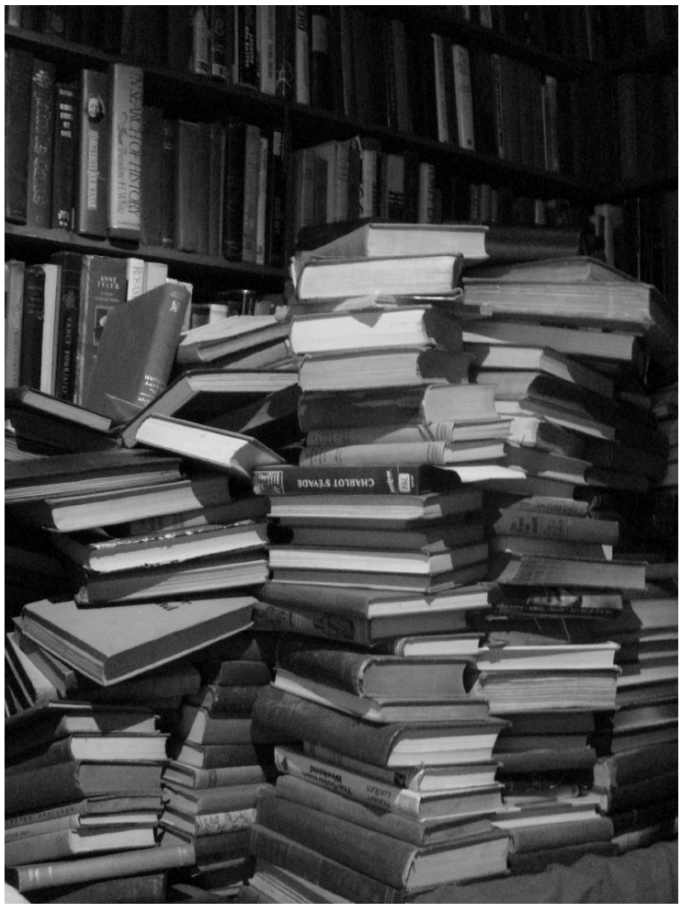


Explore



Unsupervised 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?

ads

spammers

users

products

abuse

urls

mails

text

news

locations

queries

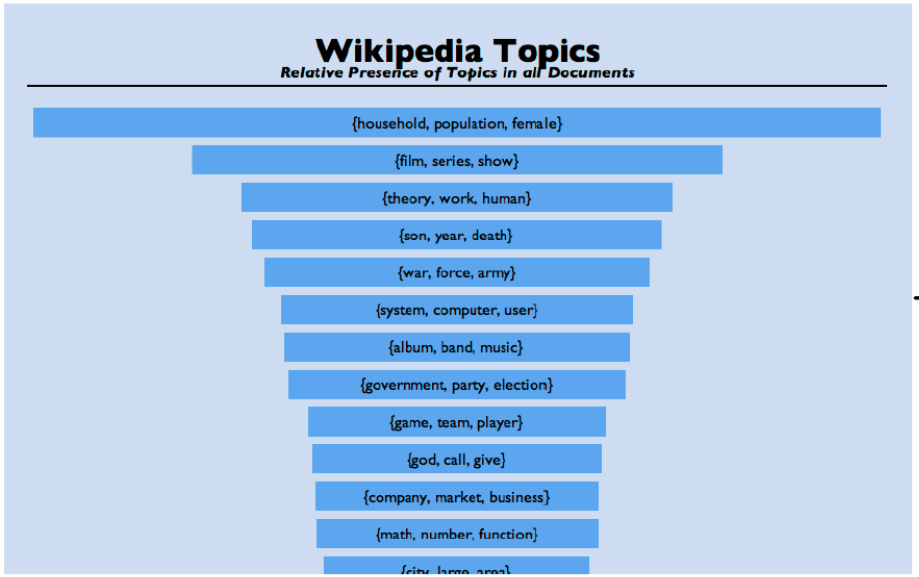
events

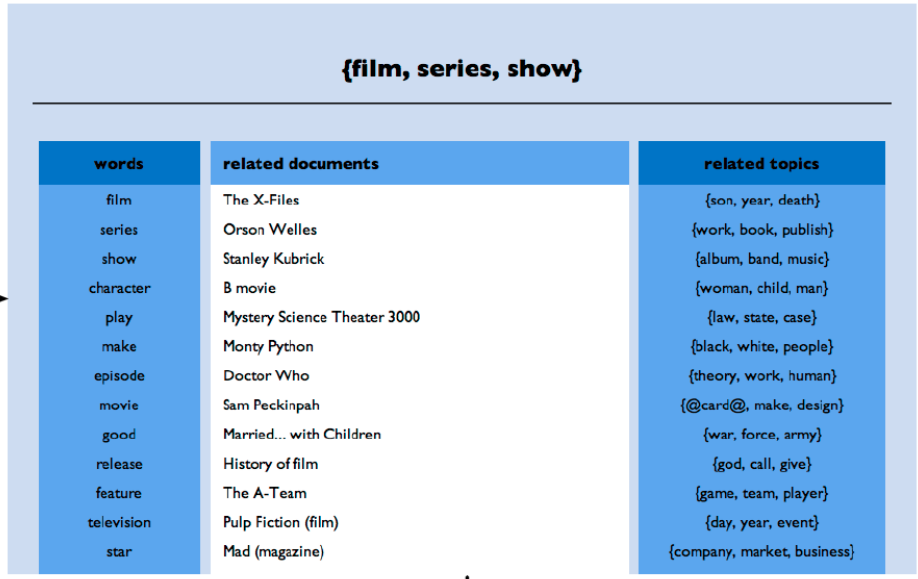
Topic modeling - Motivation

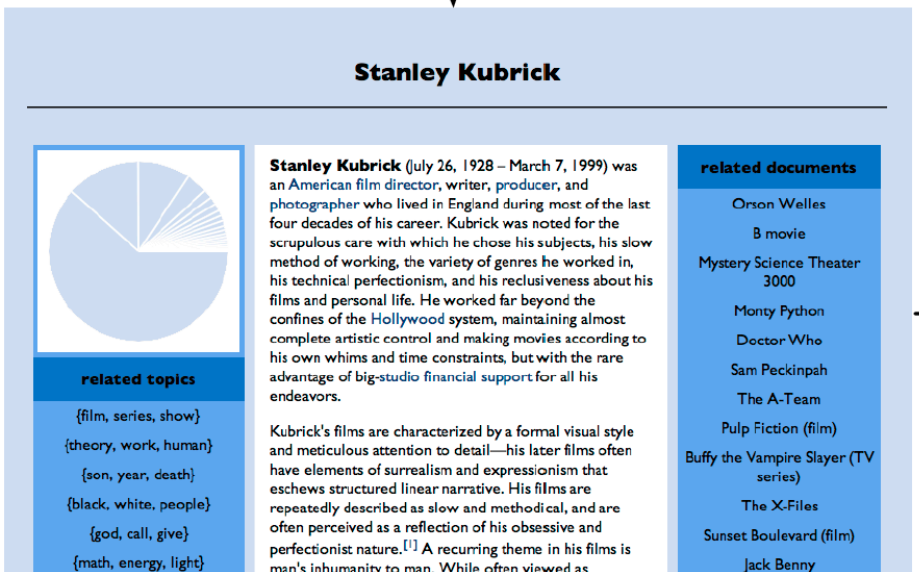
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

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







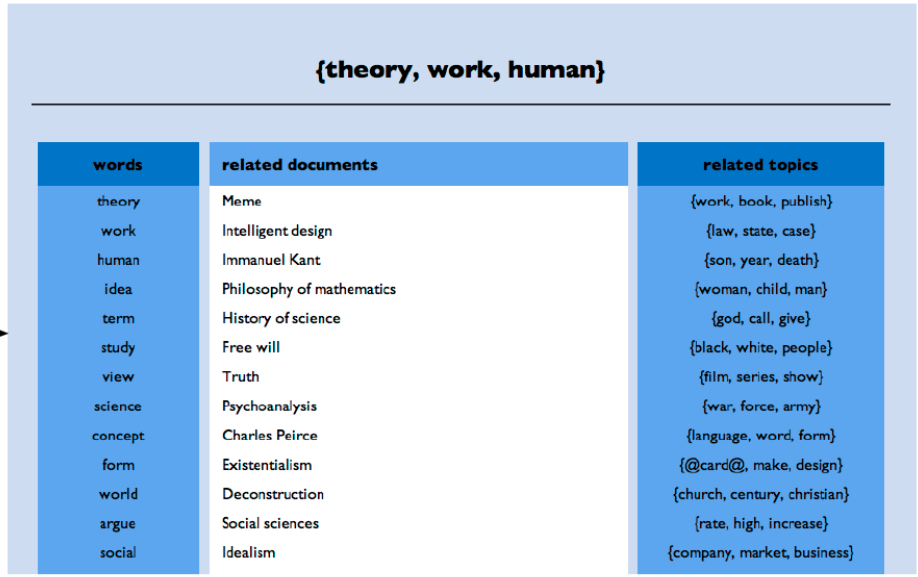


Image annotation

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

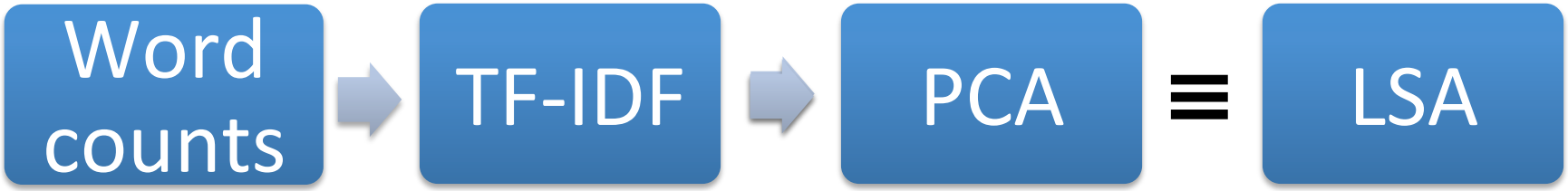
PEOPLE MARKET PATTERN
TEXTILE DISPLAY

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:

Documents

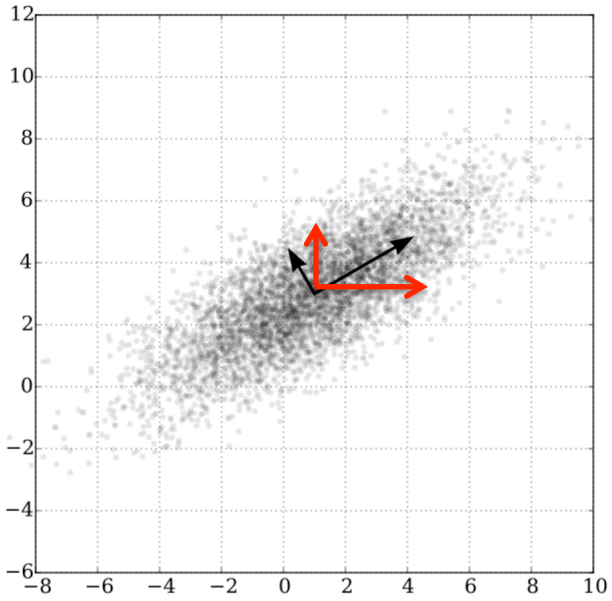
Words

	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

Co-appearing words

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

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

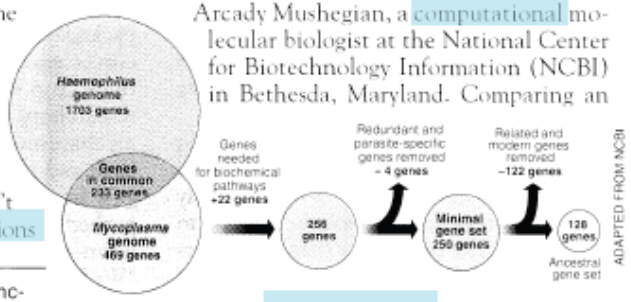
1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using *posterior inference*
   - What are the topics that describe this 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



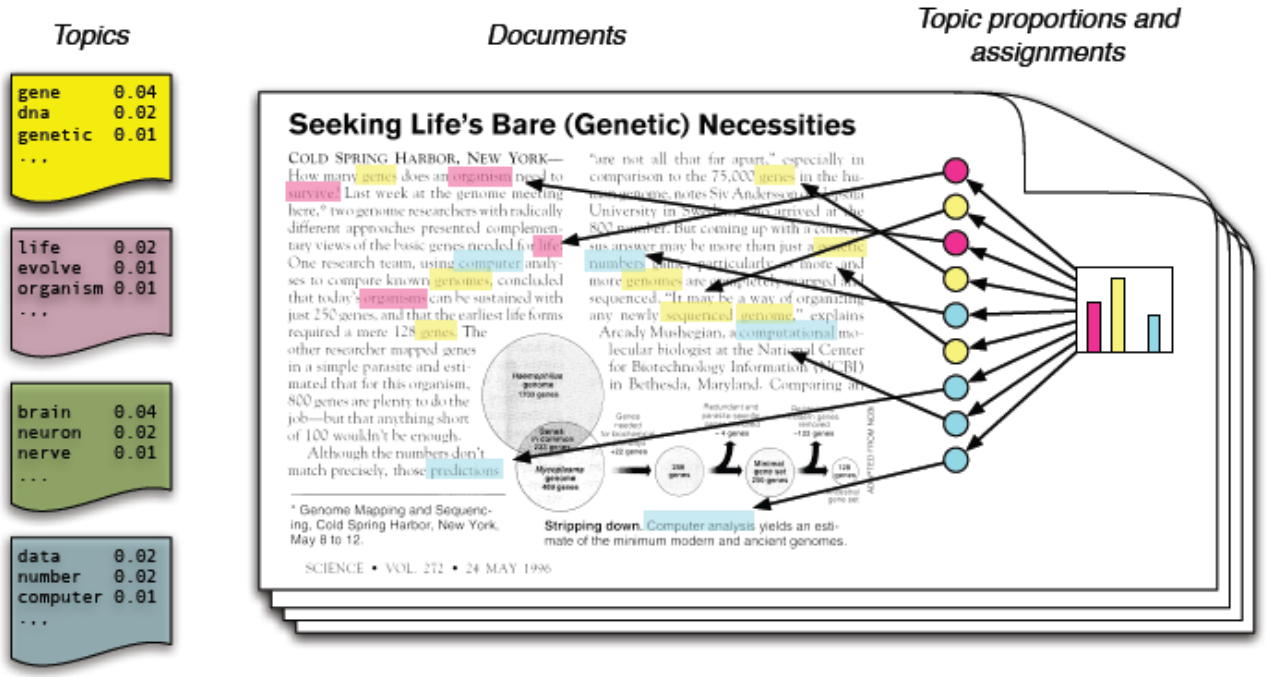
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

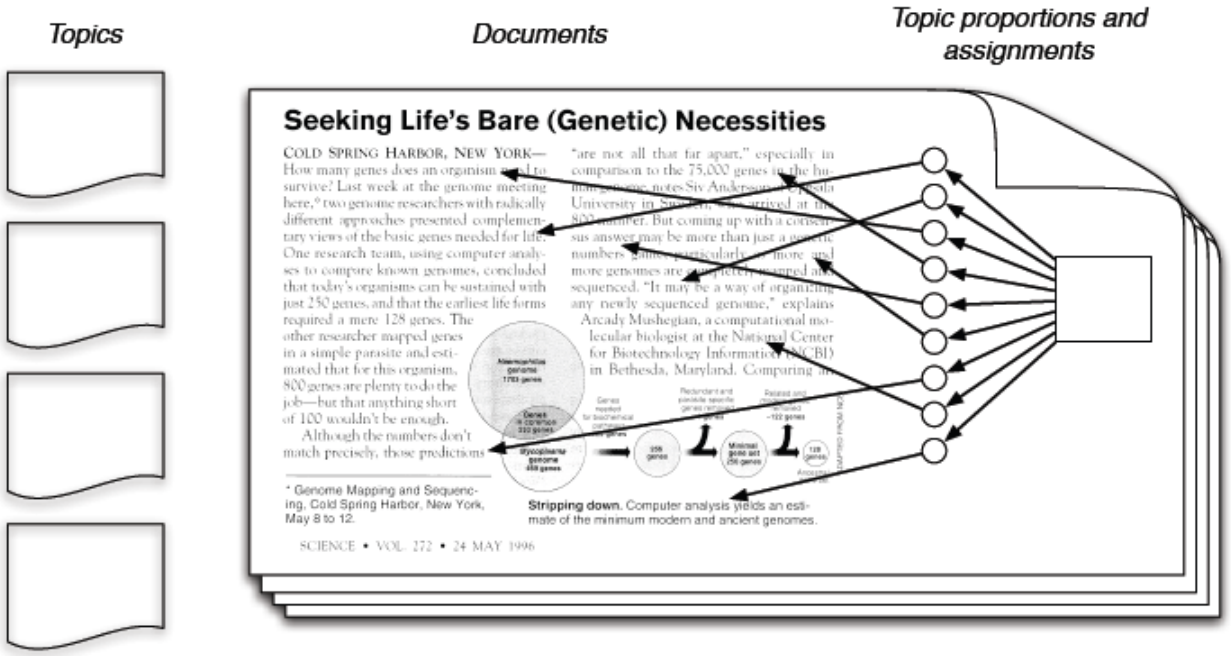
Simple intuition: Documents exhibit multiple topics.

Generative model



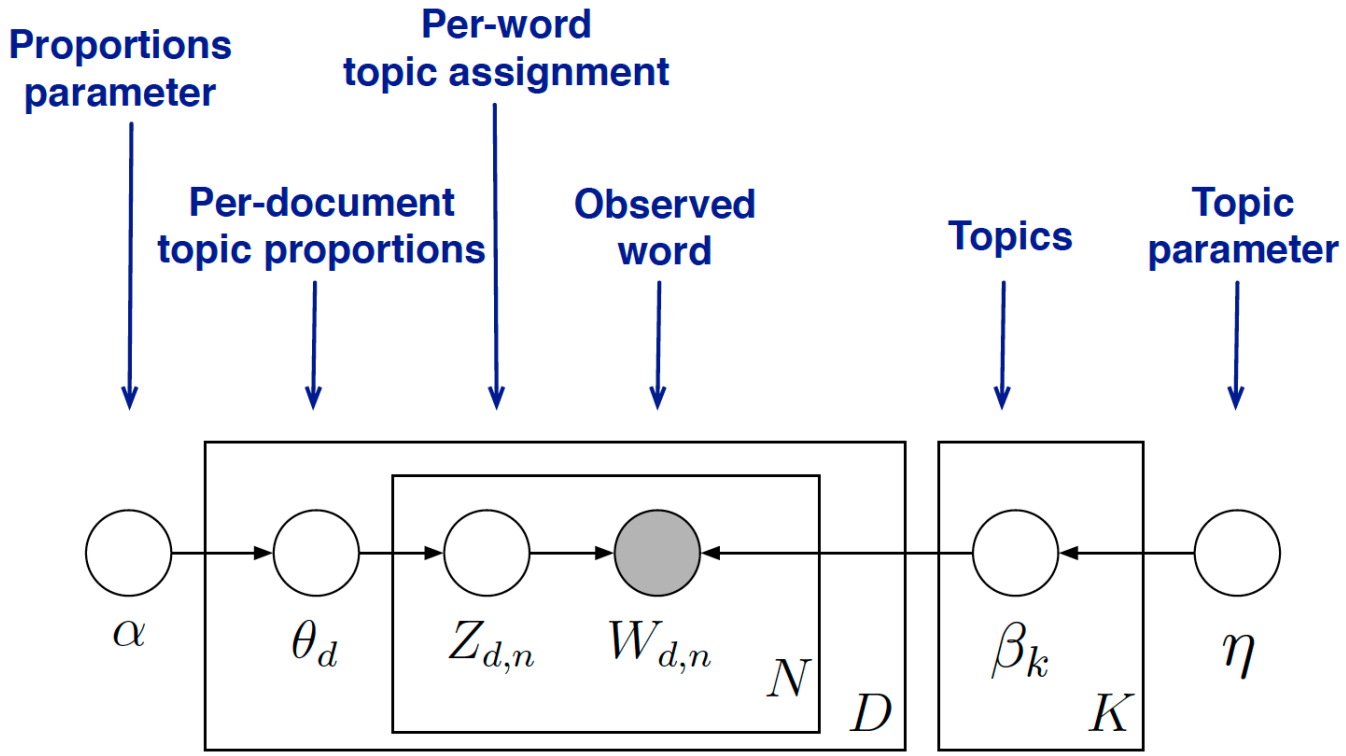
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

The posterior distribution



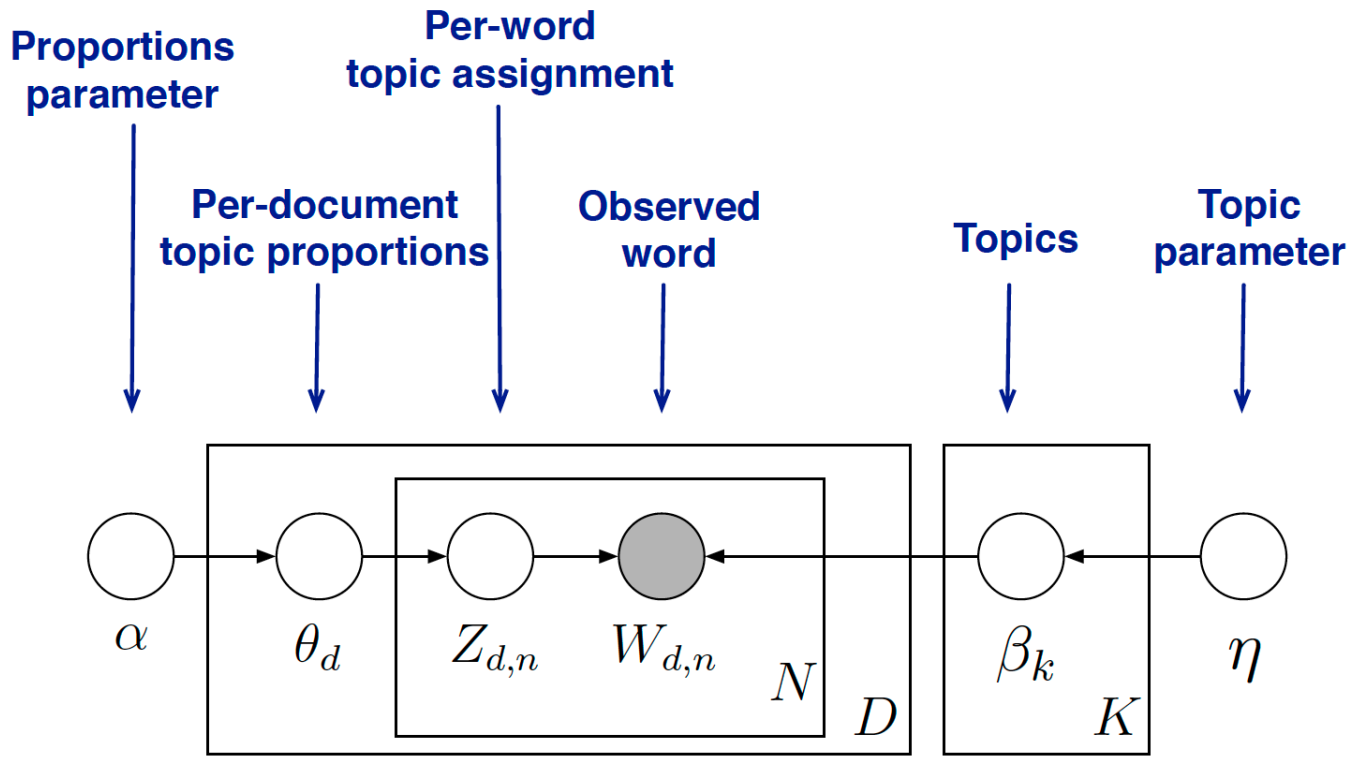
- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

LDA generative model



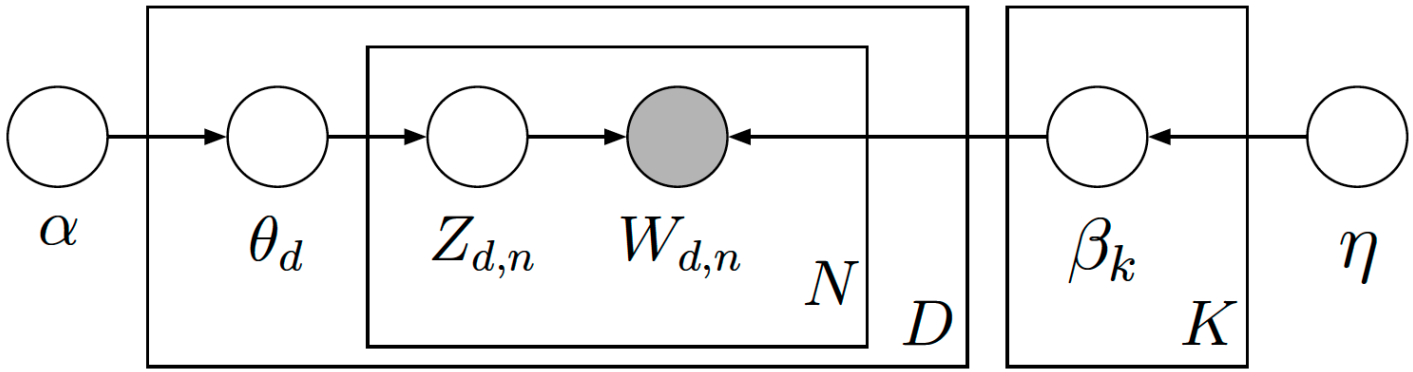
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

LDA generative model



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

LDA

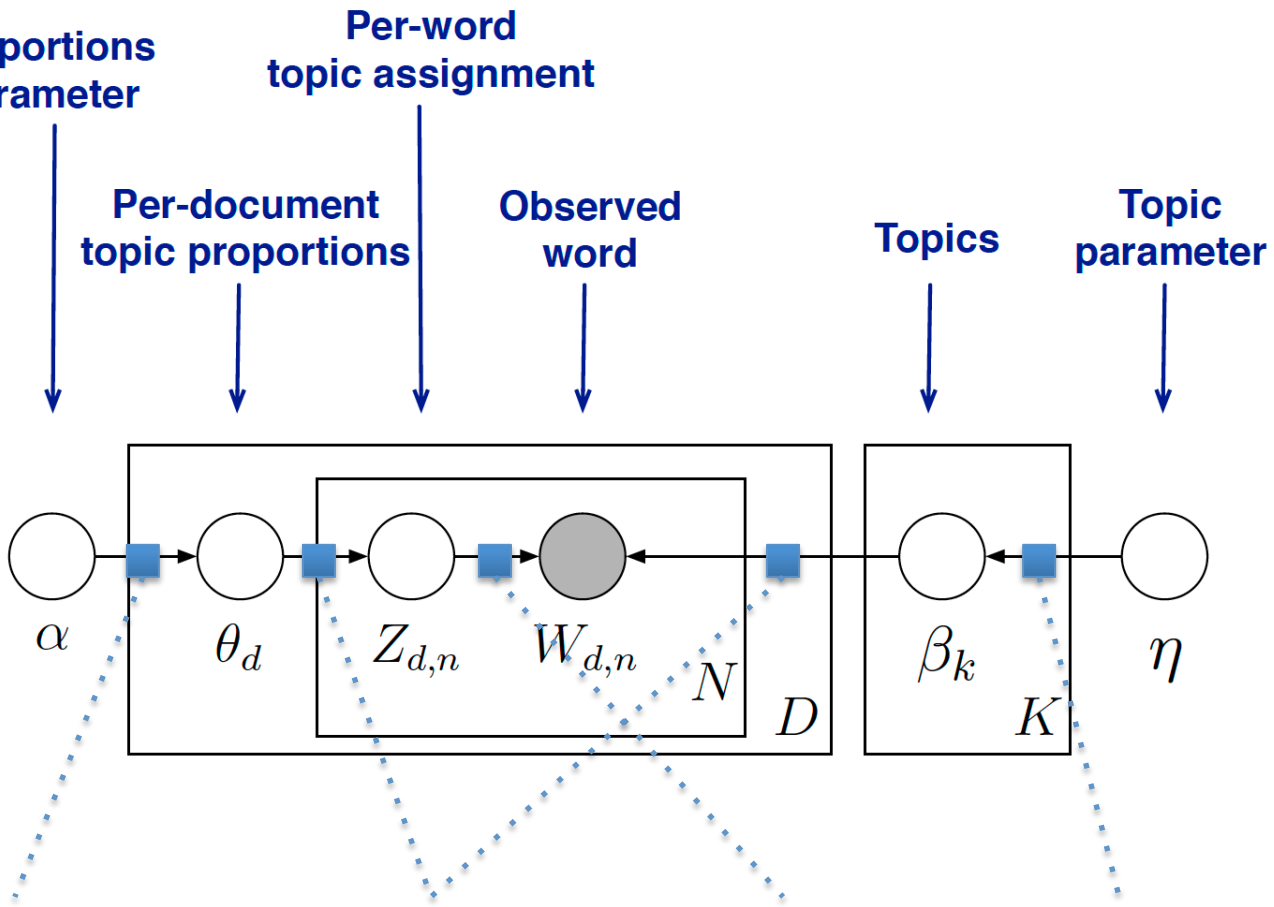


- This joint defines a posterior, $p(\theta, z, \beta \mid w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k

  } **Main output**

- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.

LDA generative model



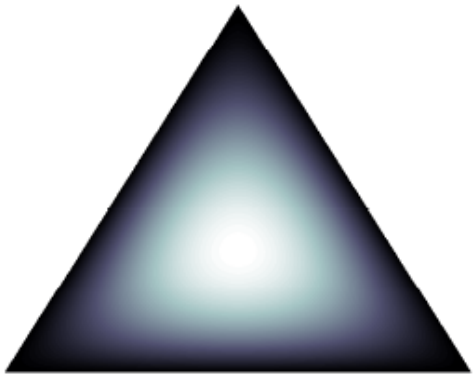
Dirichlet distribution

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

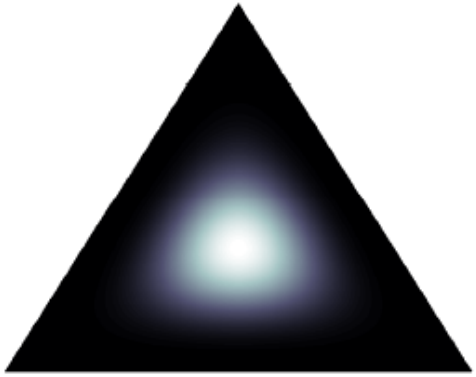
$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- 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



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



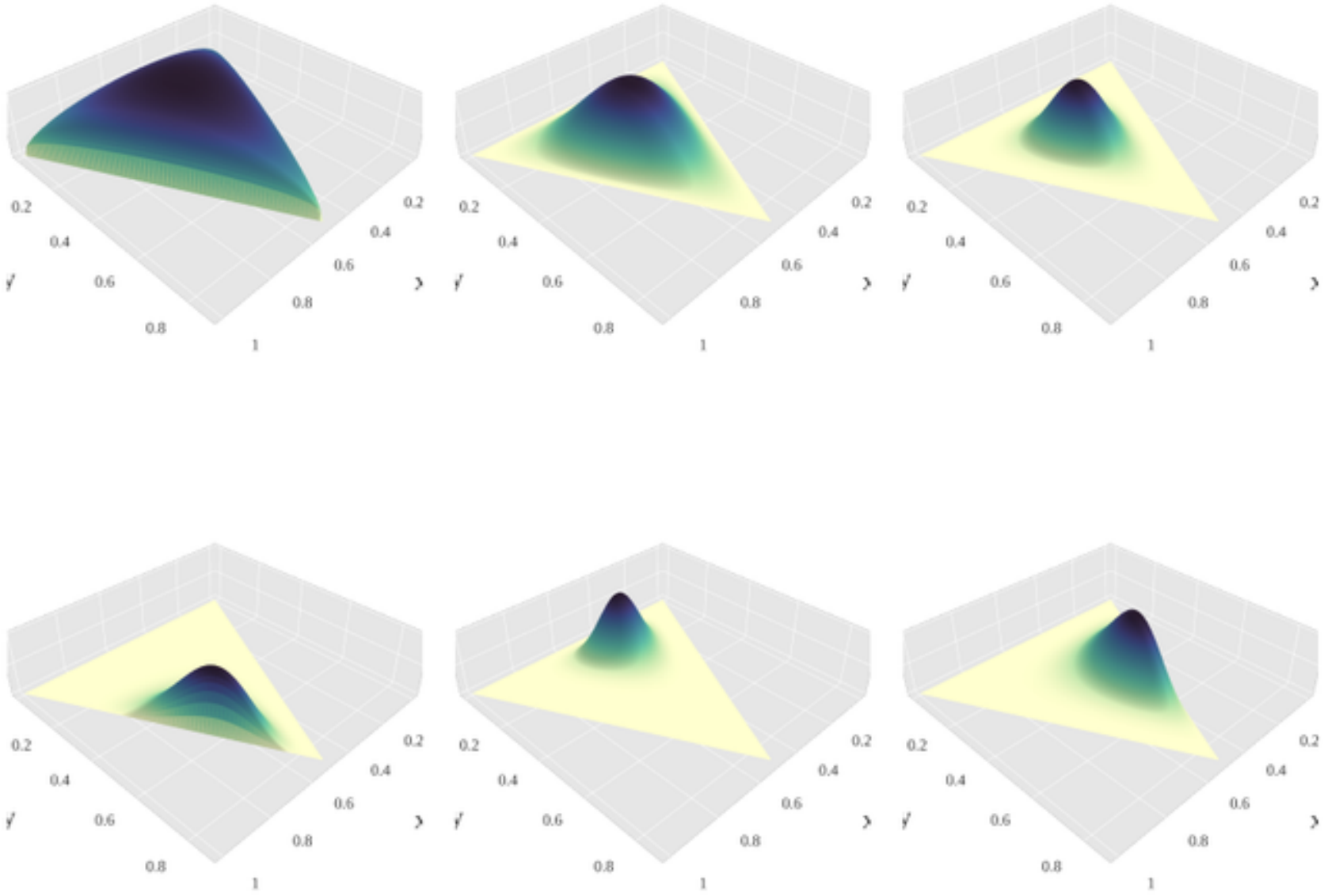
$\alpha = (5, 5, 5)$

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

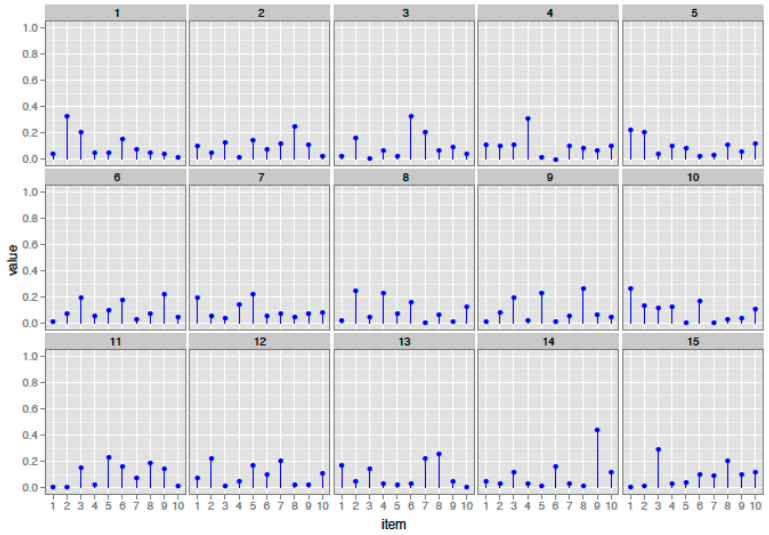
Darker implies lower magnitude

alpha < 1 leads to sparser topics

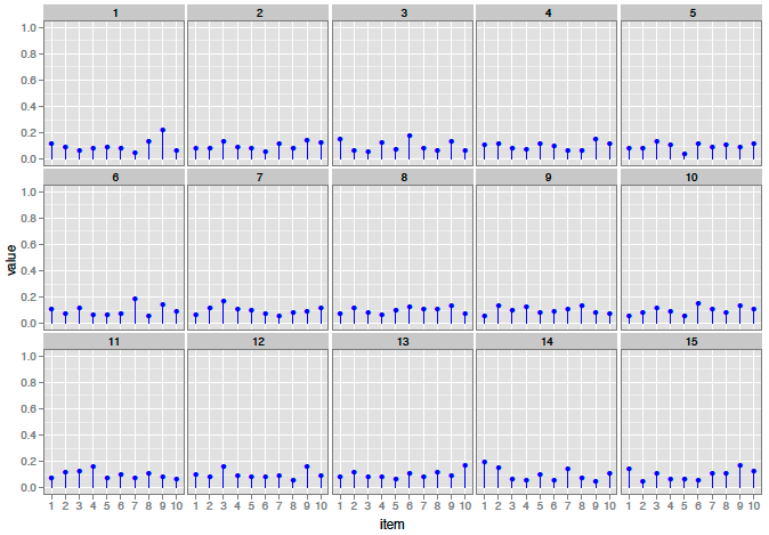
Dirichlet Examples



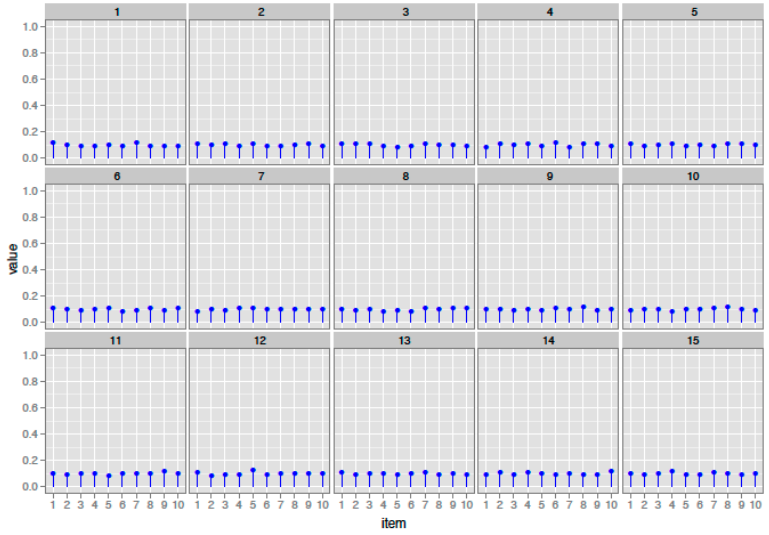
$\alpha = 1$



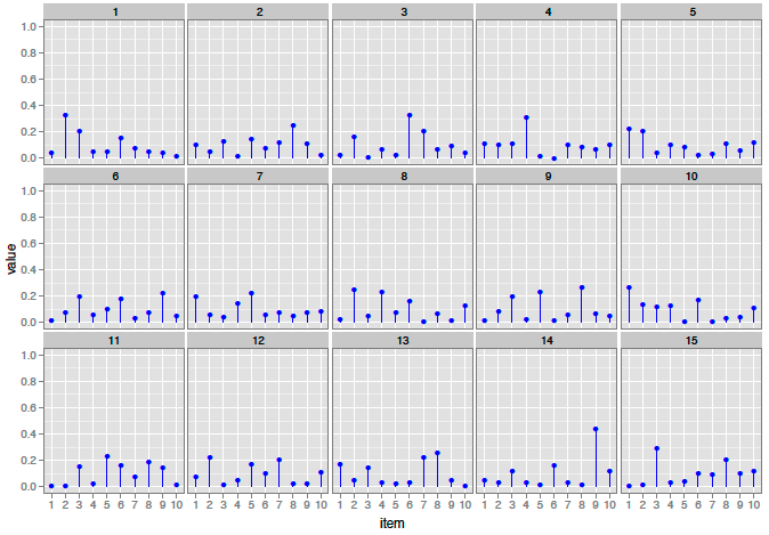
$\alpha = 10$



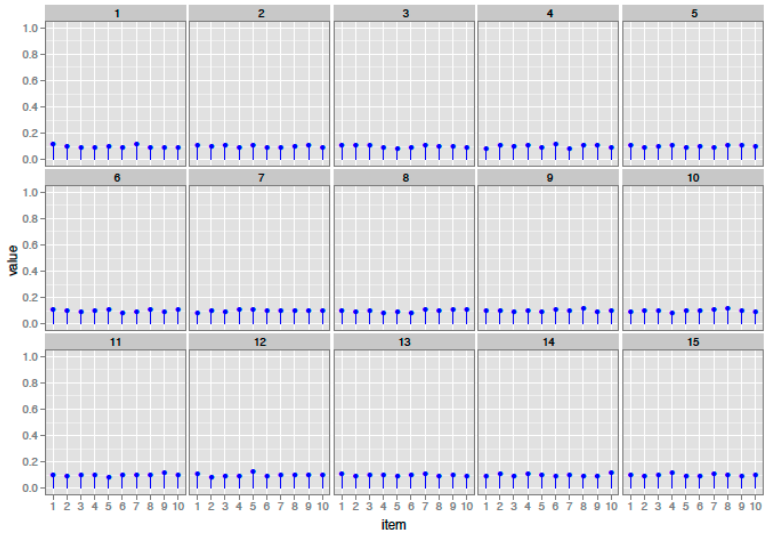
$\alpha = 100$



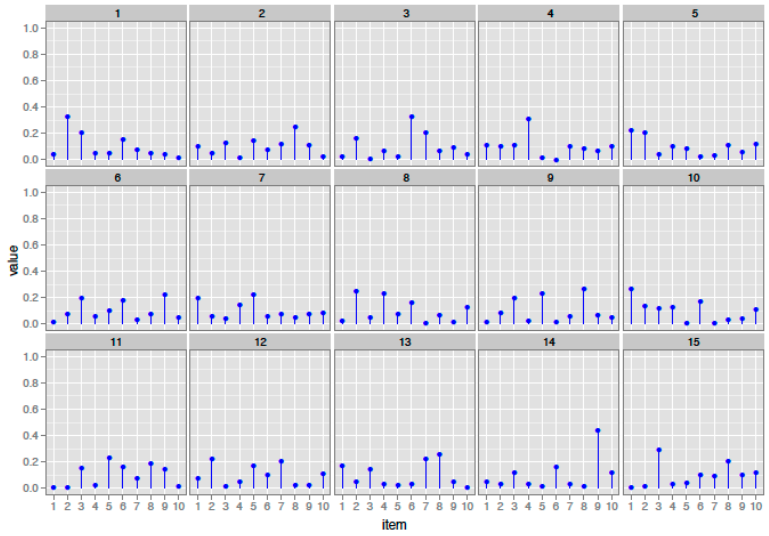
$\alpha = 1$



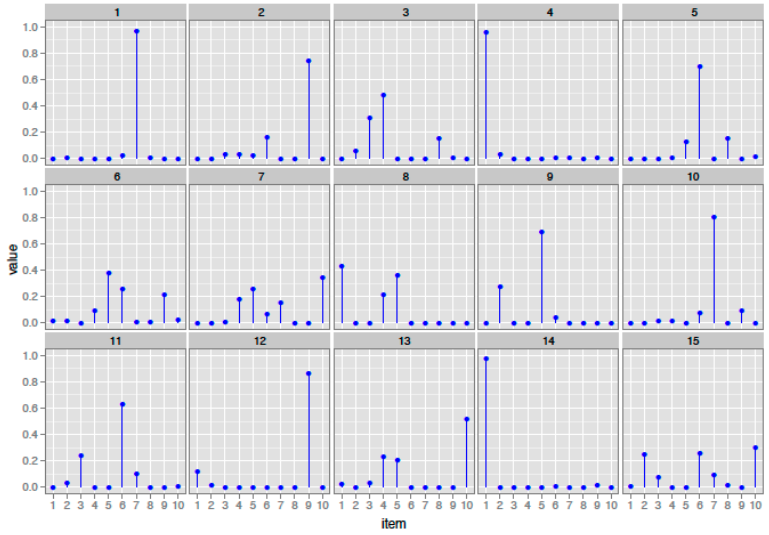
$\alpha = 100$



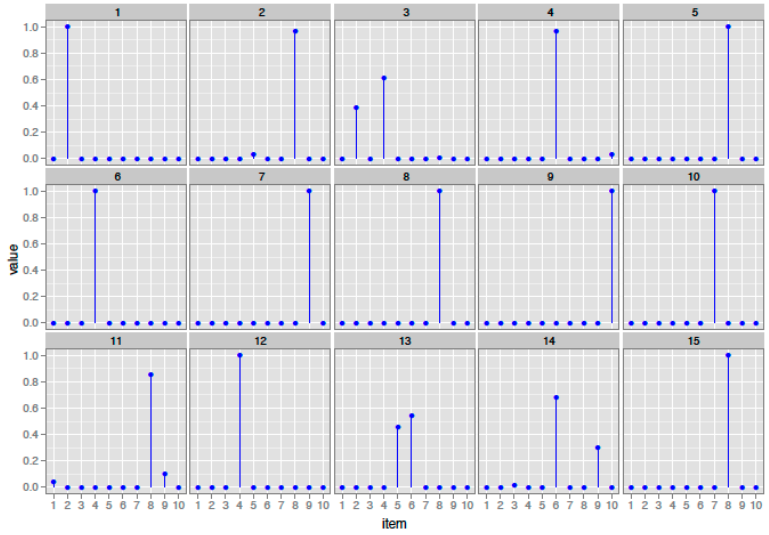
$\alpha = 1$



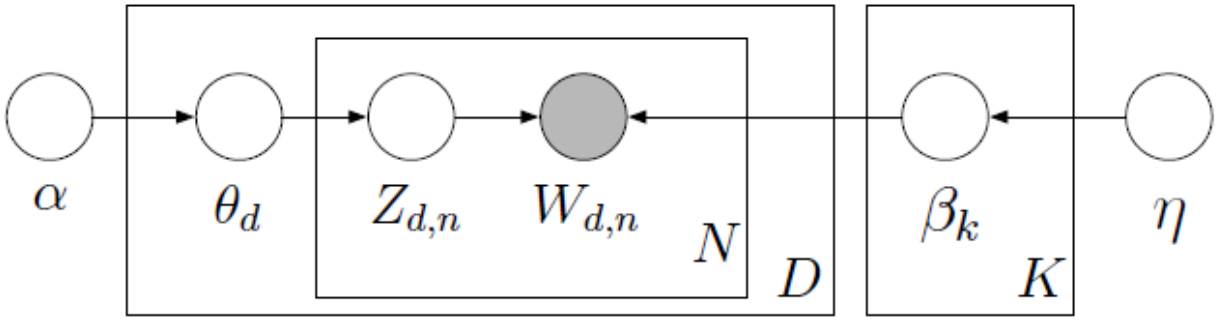
$\alpha = 0.1$



$\alpha = 0.01$



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).

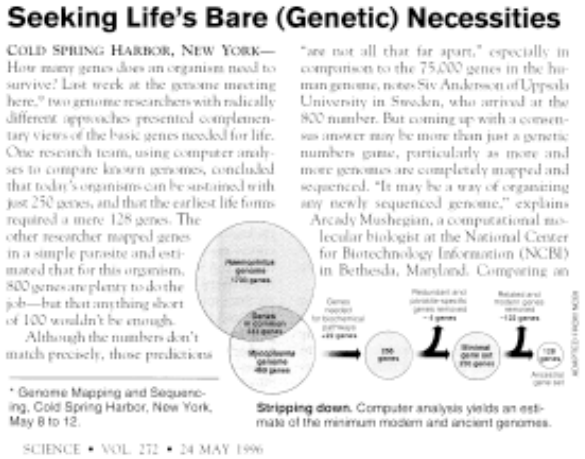
Inference comparison

- 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 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

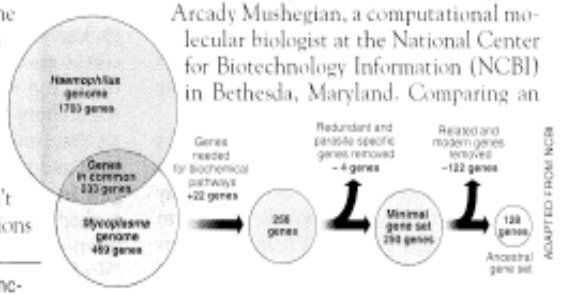
- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

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.

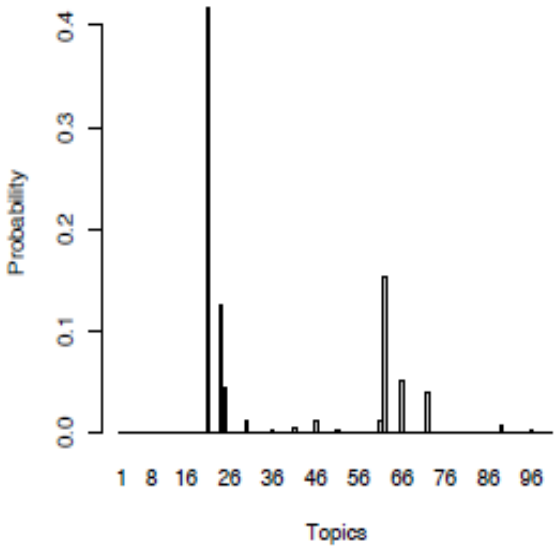
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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



* 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



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

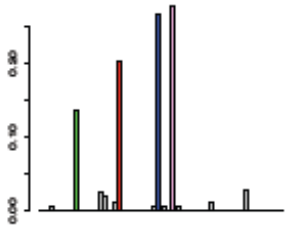
Chance and Statistical Significance in Protein and DNA Sequence Analysis

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.

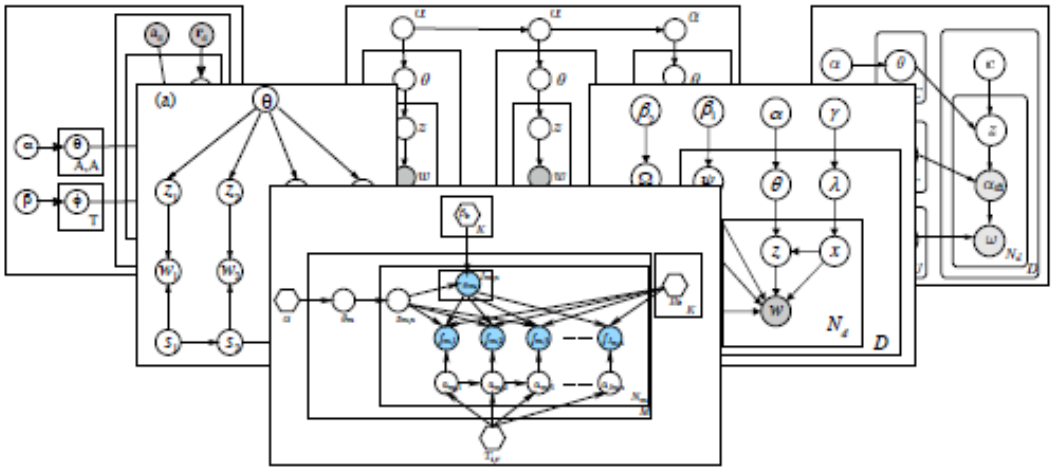
Top Ten Similar Documents

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” ?

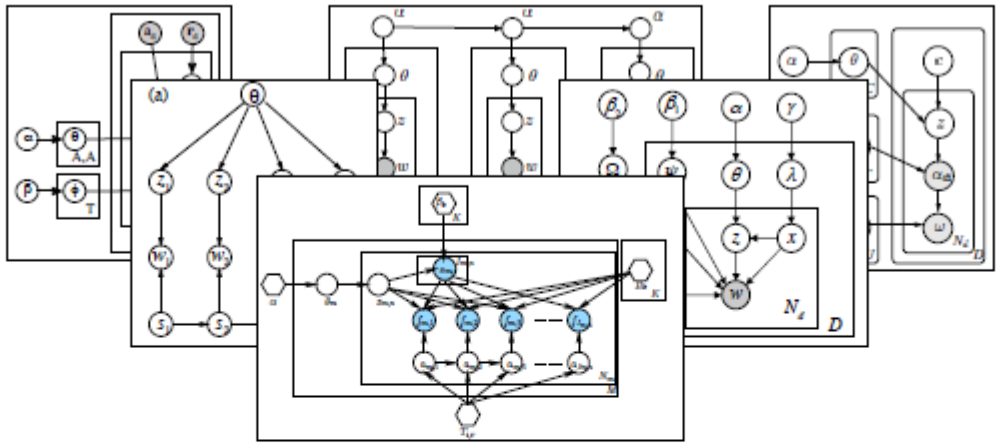
- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible.
  2. For each topic, assign high probability to as few terms as possible.
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability 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 is modular, general, useful



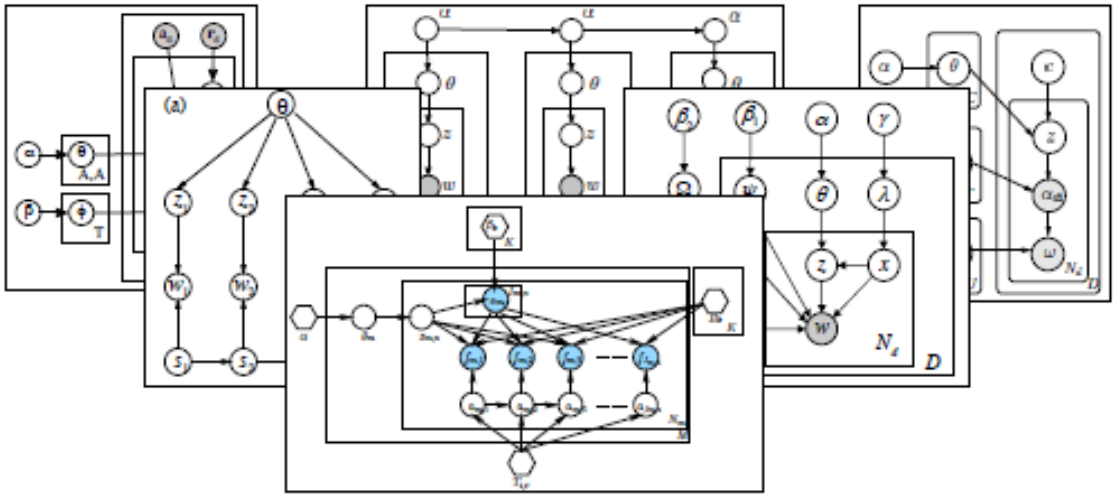
- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
- E.g., syntax; authorship; word sense; dynamics; correlation; hierarchies; nonparametric Bayes

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



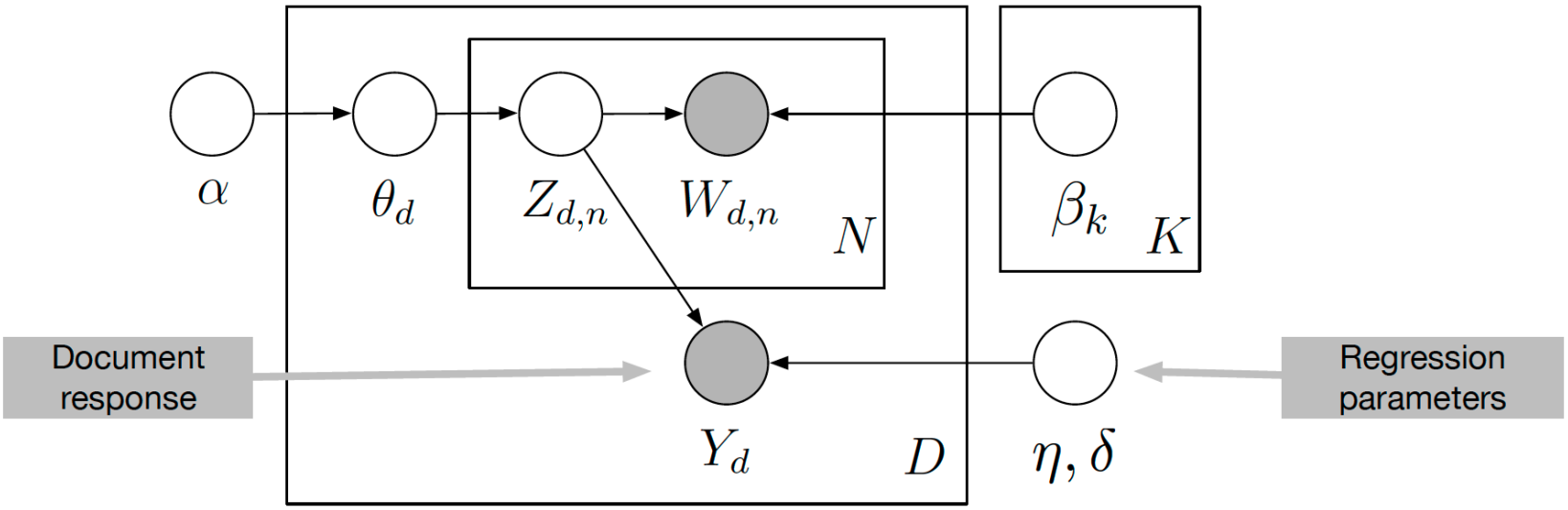
- The **posterior** can be used in creative ways
- E.g., IR, collaborative filtering, document similarity, visualizing interdisciplinary documents

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.

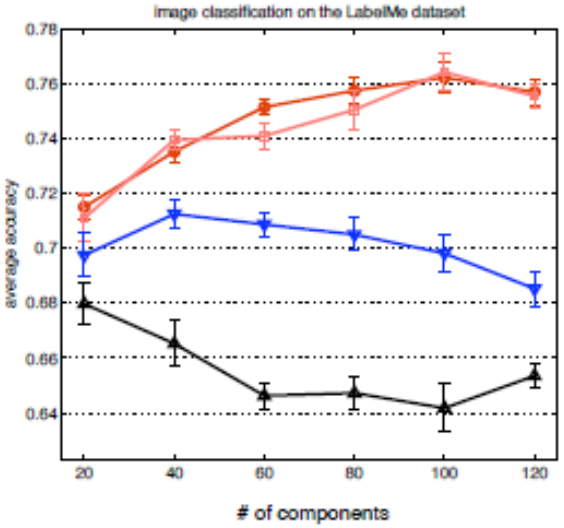
Supervised LDA

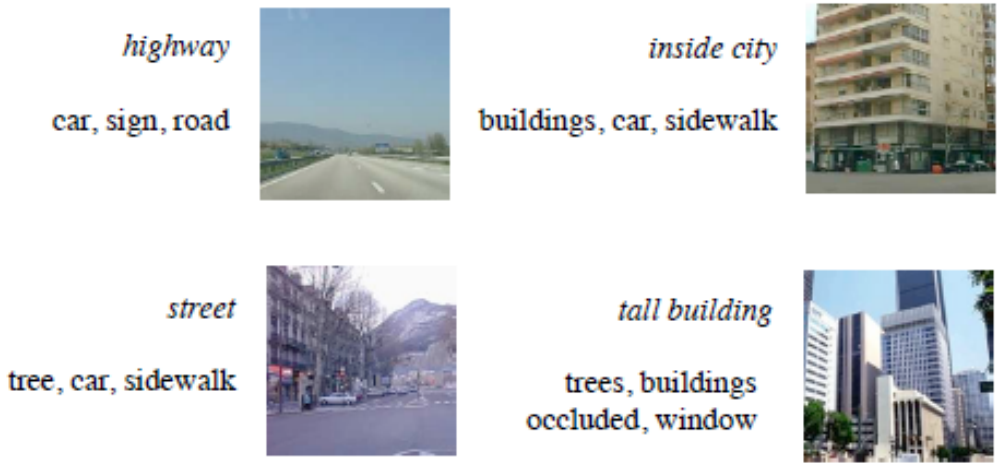


1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim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

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

Example: Multi class classification



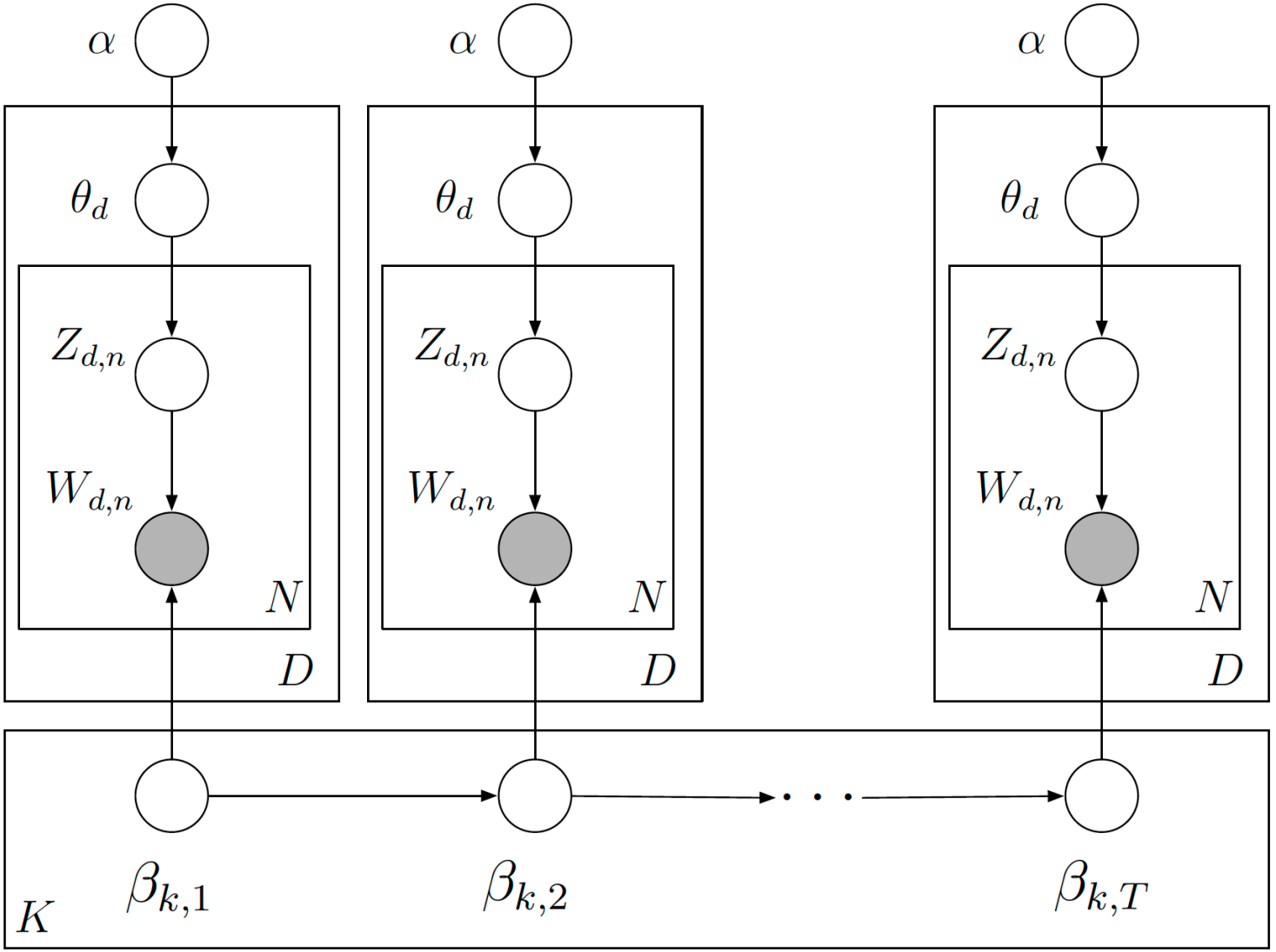


SLDA for image classification (with Chong Wang, CVPR 2009)

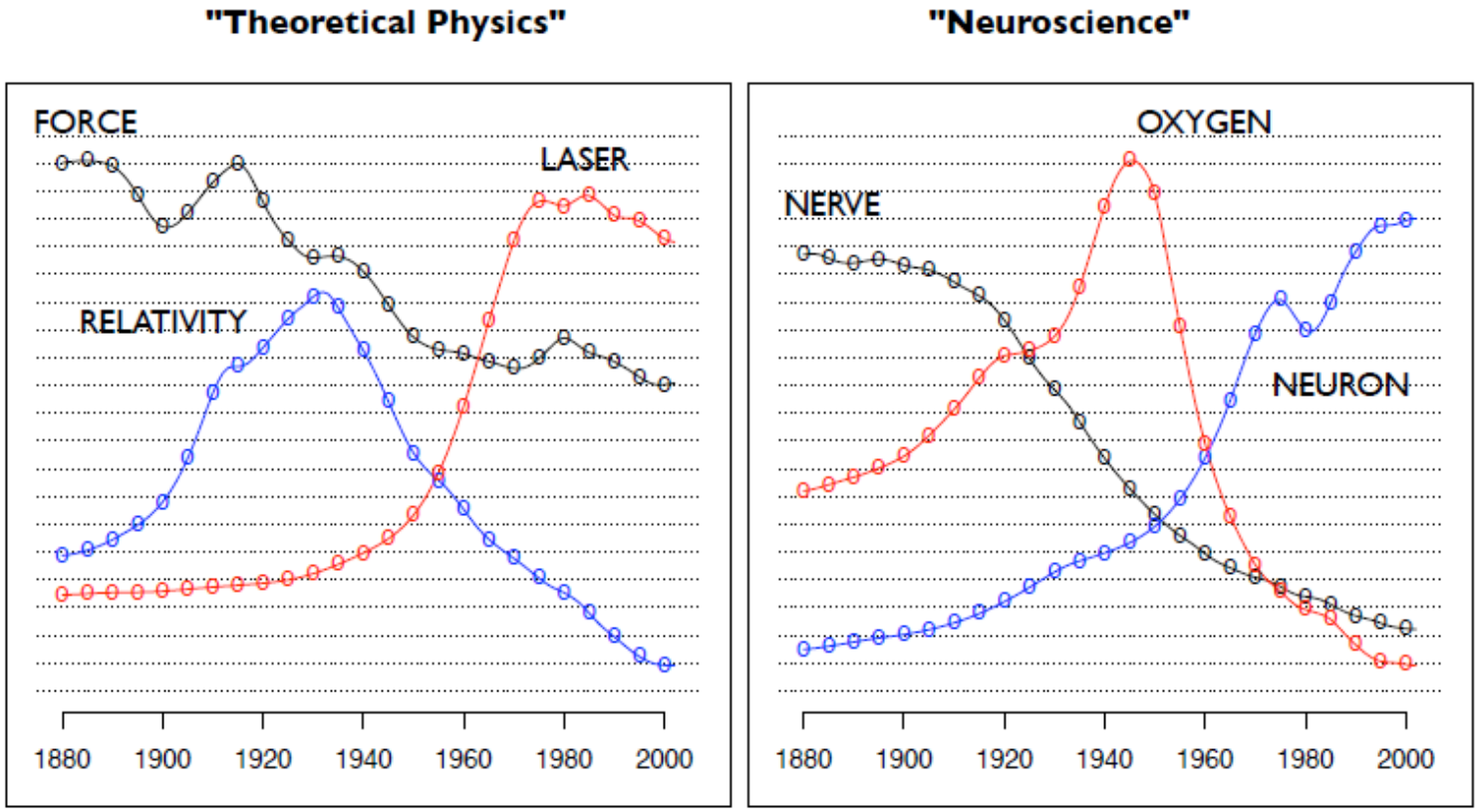
Supervised topic models

- 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



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