Latent Dirichlet Allocation
Introduction/Overview

David Meyer

http://www.1-4-5.net/~dmm/ml/lda_intro.pdf

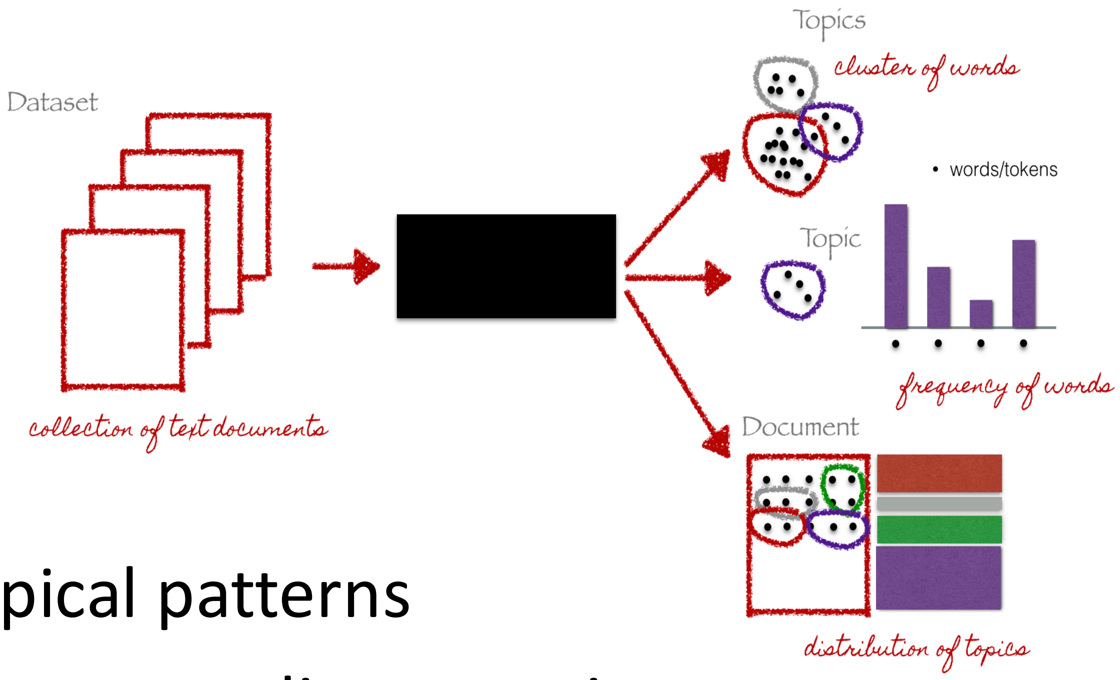
03.10.2016

Agenda

- What is Topic Modeling?
- Parametric vs. Non-Parametric Models
- Latent Dirichlet Allocation
- Probabilistic Graphical Models
- The Effect of the Dirichlet parameter α
- Dynamic LDA
- Q&A

Topic Modeling

- Methods for automatically organizing, understanding, searching and summarizing *categorical data*



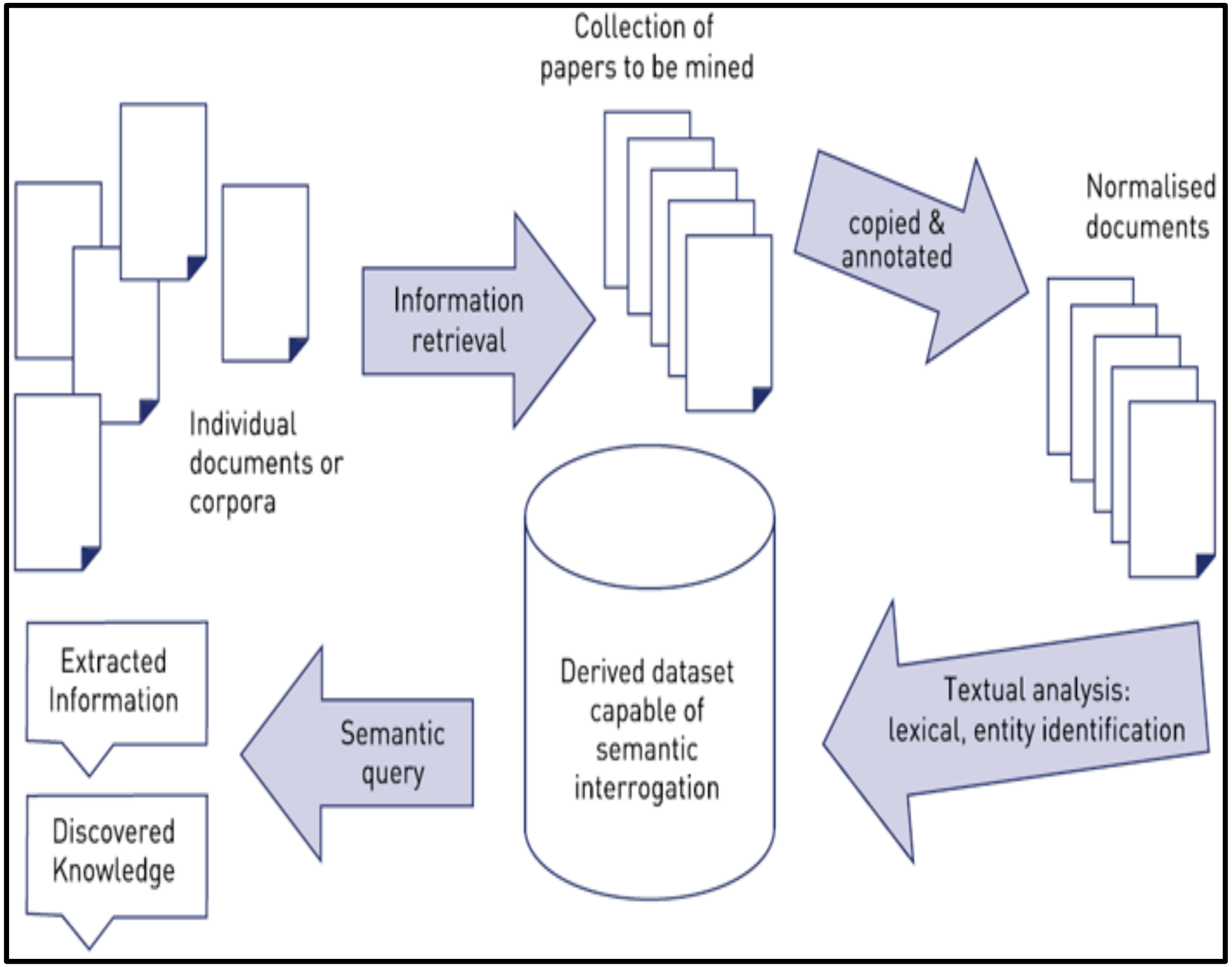
- Goals
  - Uncover hidden topical patterns
  - Annotate documents according to topics
  - Organize, Summarize, Search

A Bit of Information Retrieval (IR) Terminology

- **Corpus:** is a large and structured set of texts
- **Stop words:** words which are filtered out before or after processing of natural language data (text)
- **Unstructured text:** information that either does not have a pre-defined data model or is not organized in a pre-defined manner.
- **Tokenizing:** process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens (see also lexical analysis)
- **Natural language processing:** field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages
- **Term document (or document term) matrix:** is a mathematical matrix that describes the frequency of terms that occur in a collection of documents
- **Supervised learning:** s the machine learning task of inferring a function from labeled training data
- **Unsupervised learning:** find hidden structure in unlabeled data
- **Stemming:** the process for reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form

Graphic courtesy Wikipedia

IR Workflow

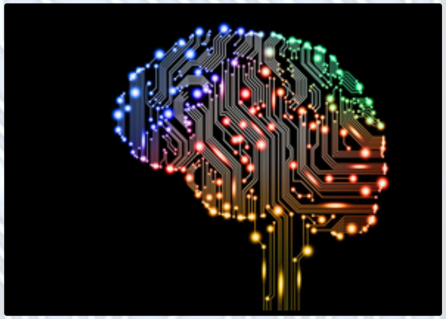

Collection of papers to be mined
Individual documents or corpora
Information retrieval
copied & annotated
Normalised documents
Textual analysis: lexical, entity identification
Derived dataset capable of semantic interrogation
Semantic query
Extracted Information
Discovered Knowledge

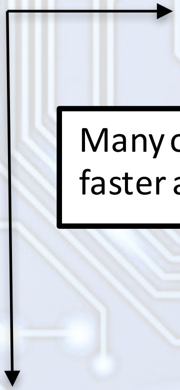
Graphic courtesy http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

Aside: We Know Machines Are Getting Smarter, But Where Does Knowledge Come From?

Evolution

Experience

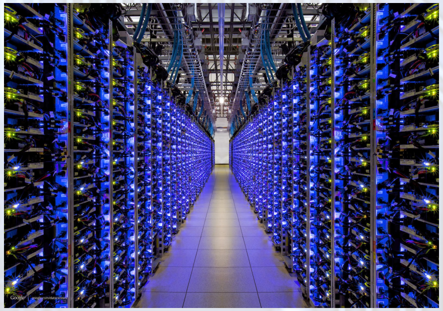




Many orders of magnitude faster and larger

Culture

Machines



Ok, But How Can Machines Discover New Knowledge?

- **Fill the gaps in existing knowledge**
  - Symbolists
  - Technology: Induction/Inverse Deduction
- **Emulate the brain**
  - Connectionists
  - Technology: Deep neural nets
- **Emulate evolution**
  - Evolutionaries
  - Technology: Genetic Algorithms
- **Systematically reduce uncertainty**
  - Bayesians
  - Technology: Bayesian Inference
- **- Notice similarities between old and new**
  - Analogizers
  - Technology: Kernel machines/Support Vector Machines

These correspond to the 5 major schools of thought in machine learning

Agenda

- ~~What is Topic Modeling?~~
- Parametric vs. Non-Parametric Models
- Latent Dirichlet Allocation
- Probabilistic Graphical Models
- The Effect of the Dirichlet parameter α
- Dynamic LDA
- Q&A

Parametric vs. Non-parametric Models

- Parametric models assume some finite set of parameters θ. This means that given parameters θ, future predictions **x** are *independent* of the data *D.* That is:
  - $p(\mathbf{x}|\theta, \mathrm{D}) = p(\mathbf{x}|\theta)$
- This implies that θ captures everything there is to know about the data
- Also means that the complexity of the model is bounded even if the amount of data isn't
- Taken together these factors make parametric models inflexible
- Information theoretic view
  - Information is constrained to flow from the prior through finite θ to the posterior
  - Remembering that the $posterior = \frac{likelihood \cdot prior}{evidence}$
  - $p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)} = \frac{p(x,\theta)}{p(x)} = \frac{p(x,\theta)}{\int p(x,\theta)d\theta}$
  - http://www.1-4-5.net/~dmm/ml/ps.pdf
- Latent Dirichlet Allocation is a parametric model (why?)

Non-Parametric Models

- Non-parametric models assume that the data distribution cannot be defined in terms of such a finite set of parameters

- Parameters can often be described using an infinite dimensional θ
  - Think of θ as a function

- The amount of information that θ can capture about the data D can grow as the amount of data grows

- This makes non-parametric models more flexible
  - Better predictive performance
  - More “realistic”
  - Most successful ML models are non-parametric

- Examples include
  - Kernel methods (SVMs, GPs)
  - DNNs
  - K-NNs, …

Agenda

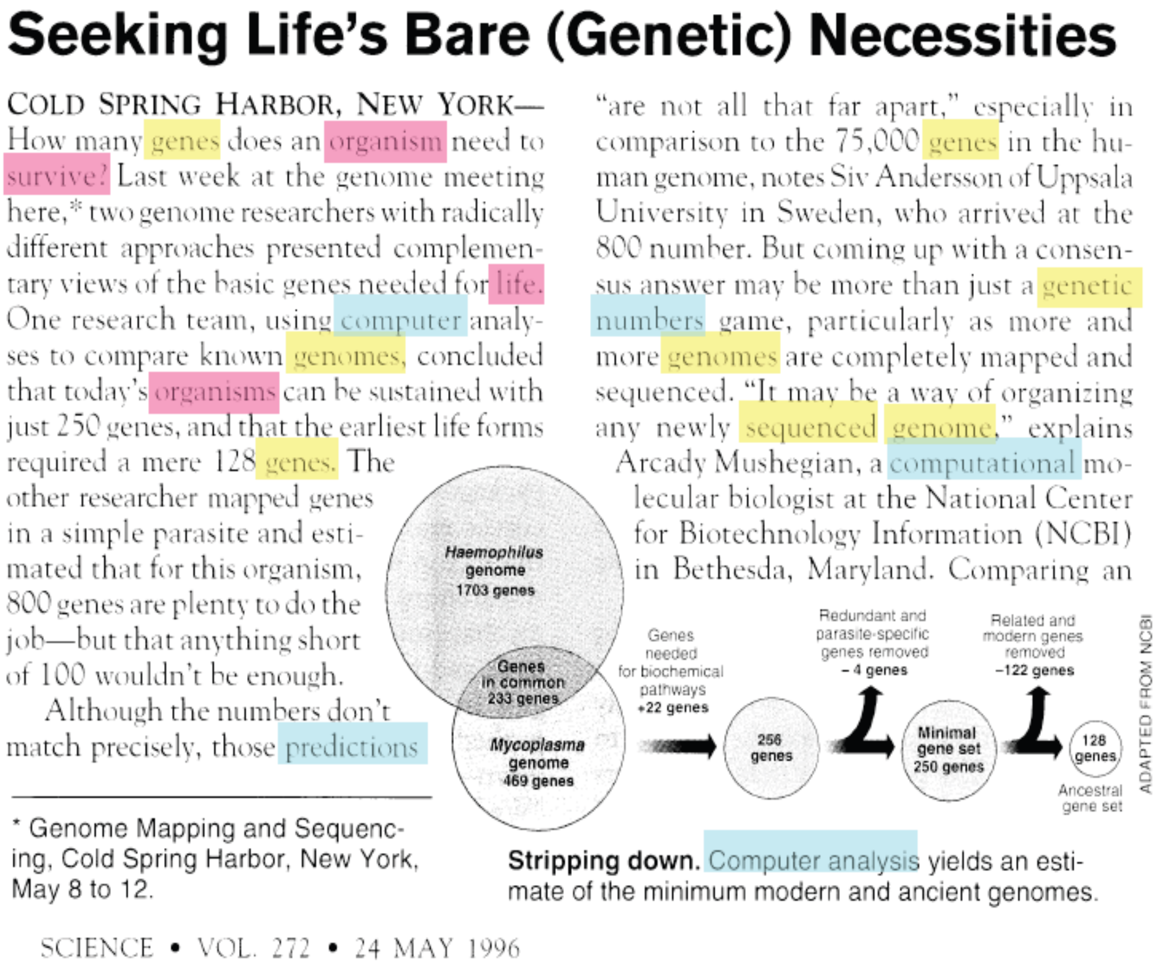
- ~~What is Topic Modeling?~~
- ~~Parametric vs. Non-Parametric Models~~
- Latent Dirichlet Allocation
- Probabilistic Graphical Models
- The Effect of the Dirichlet parameter α
- Dynamic LDA
- Q&A

Latent Dirichlet Allocation (LDA)

- Generative probabilistic model
  - ***Parametric*** Bayesian Probabilistic Graphical Model
  - Treats data as observations
  - Contains hidden variables
  - Hidden variables reflect thematic structure of the collection
  - Wealth of material here:
    - https://www.cs.princeton.edu/~blei/topicmodeling.html
- Approach: Infer hidden structure using *posterior inference*
  - Discovering topics in the collection using Bayesian inference
- Placing new data into the estimated model
  - Situating new documents into the estimated topic structure
- Other Approaches
  - Latent Semantic Indexing (LSI)
  - Probabilistic Latent Semantic Indexing (pLSI)
  - …

LDA Basic Intuition

http://ai.stanford.edu/~ang/papers/nips01-lda.pdf



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

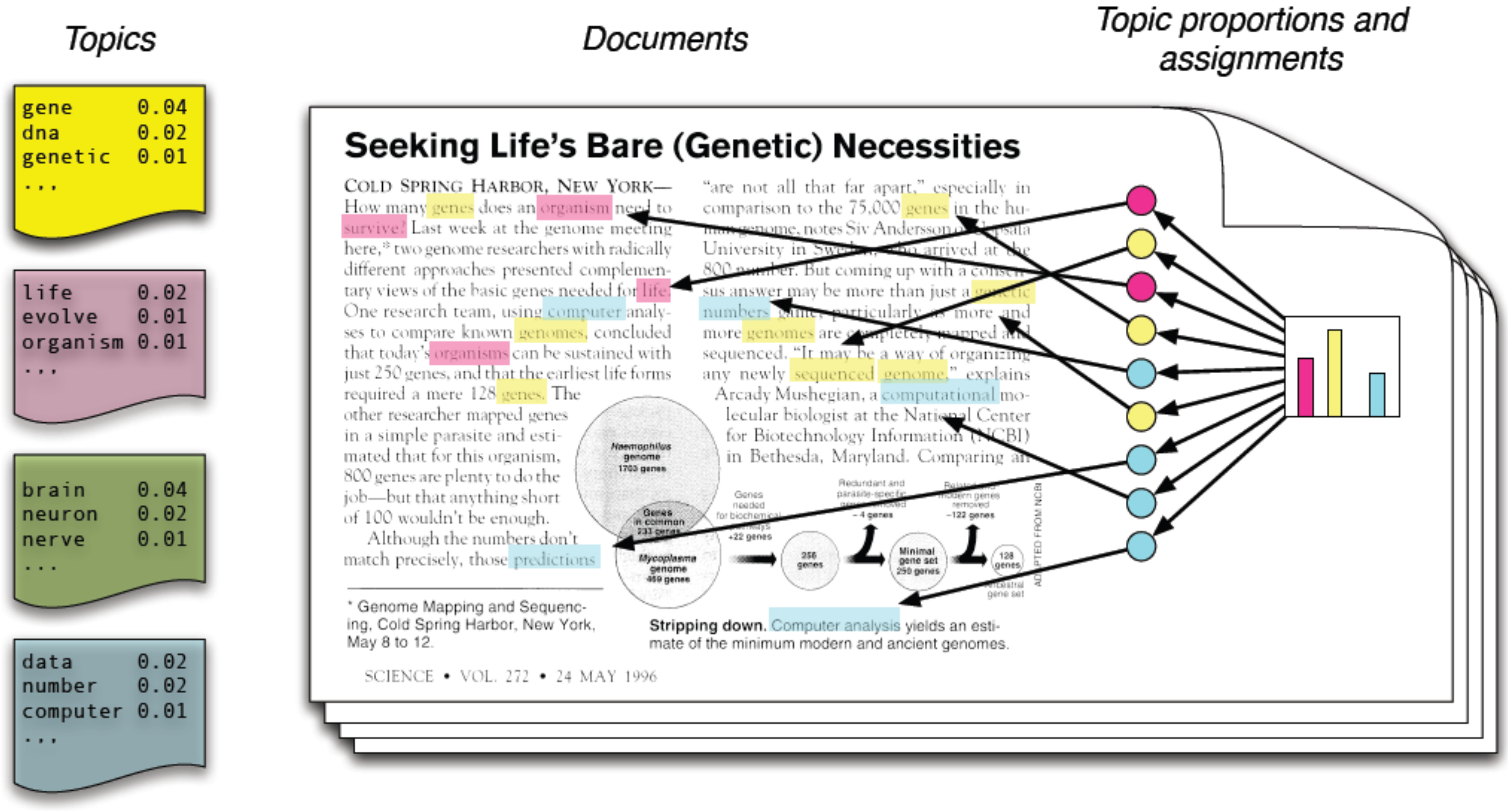
Obvious Questions

- What exactly is a topic?
- Where do topics come from?
- How many "topics"/document
- Simple Intuition:
  - Documents exhibit multiple topics
  - Contrast "mixture" models

Graphic courtesy David Blei https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf

LDA Generative Model

Hallucinate that the observed data were generated this way


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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.

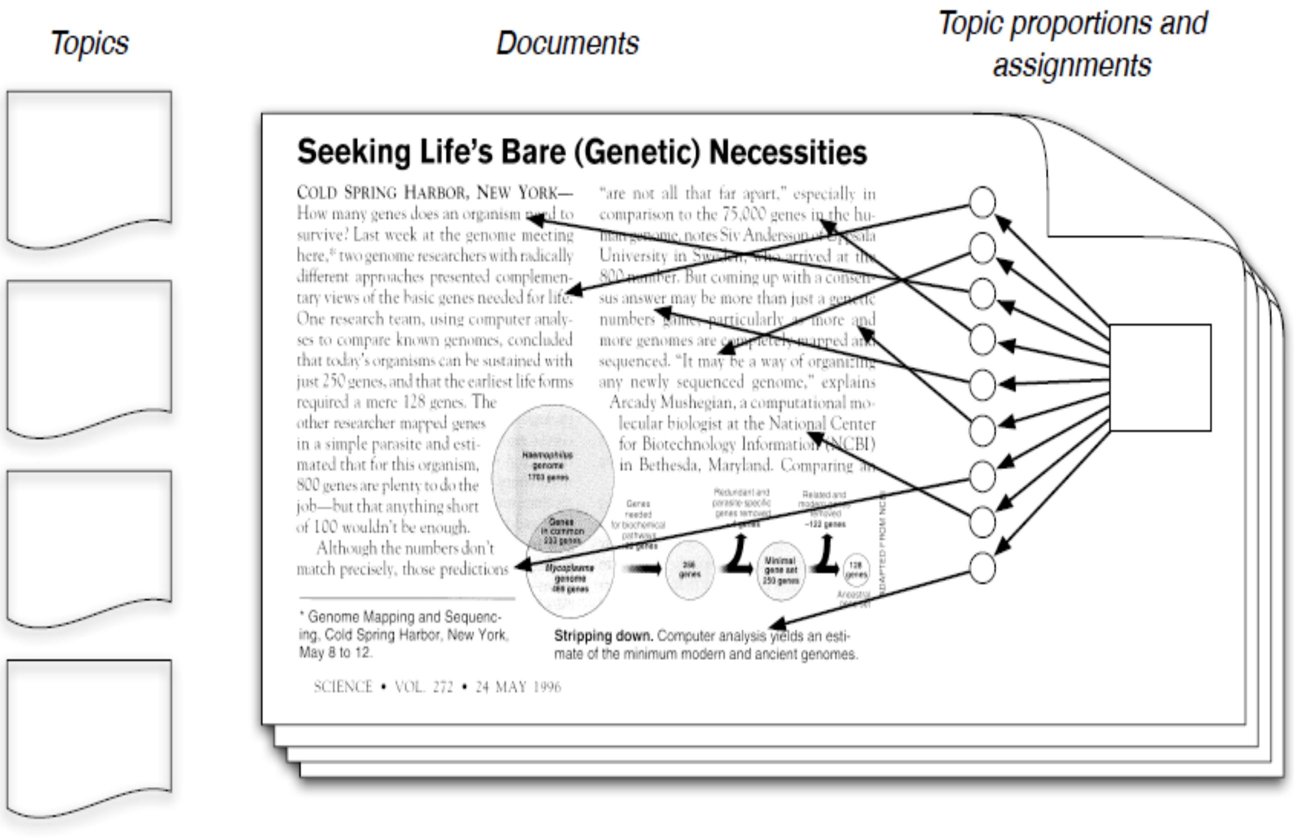
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Assume topics exists outside of the document collection
- Each topic is a *distribution* over fixed vocabulary
  - Each word is drawn from one of those topics
- Each document is a *random mixture* of corpus-wide topics

Graphic courtesy David Blei https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf

BTW, What Do We Actually Observe?


Topics
Documents
Topic proportions and assignments
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways
256 genes
Redundant and parasite-specific genes removed
Related and modern genes removed ~122 genes
Minimal gene set 250 genes
128 genes
Ancestral
ADAPTED FROM NCBI

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

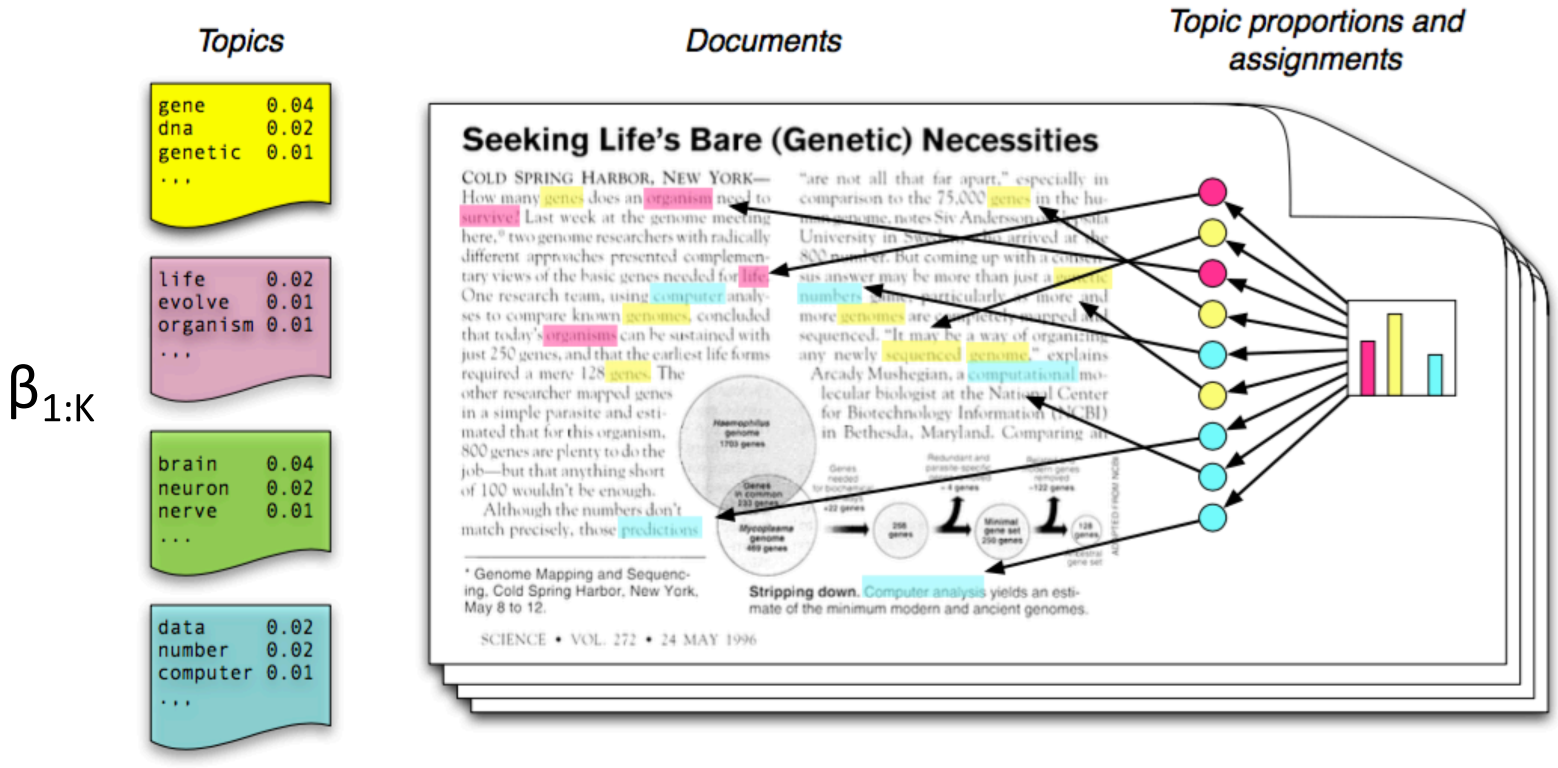
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- So our goal here is to **infer** the hidden (latent) variables
- i.e., compute their distribution conditioned on the documents: *p(*topics, proportions, assignments | documents*)*

Graphic courtesy David Blei https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf


Topics
Documents
Topic proportions and assignments
$\beta_{1:K}$
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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.

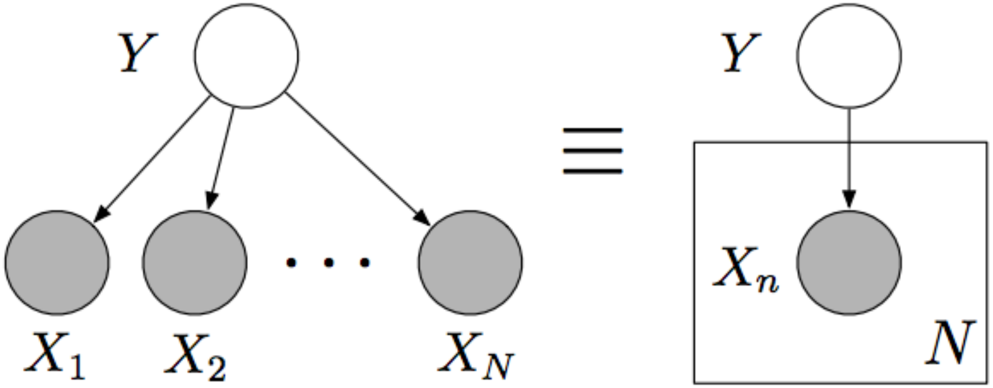
SCIENCE • VOL. 272 • 24 MAY 1996

```
Let K be the number of topics (distributions over words)          # $\beta_{1:K} \sim Dir(\eta)$
For each document d in corpus C:
    Draw the topic proportion distribution $\theta_d$ for document d    # $\theta_d \sim Dir(\alpha)$
    For each n = 1..$N_d$:                                           # $N_d$ = # words in d
        Draw topic index $z_{d,n}$ for the $n^{th}$ word of d from $\theta_d$   # $1 \leq z_{d,n} \leq K$
        Draw word $w_{d,n}$ from topic $\beta[z_{d,n}]$                  # $z_{d,n}$ indexes $\beta_{1:K}$
```

Agenda

- ~~What is Topic Modeling?~~
- ~~Parametric vs. Non-Parametric Models~~
- ~~Latent Dirichlet Allocation~~
- Probabilistic Graphical Models
- The Effect of the Dirichlet parameter α
- Dynamic LDA
- Q&A

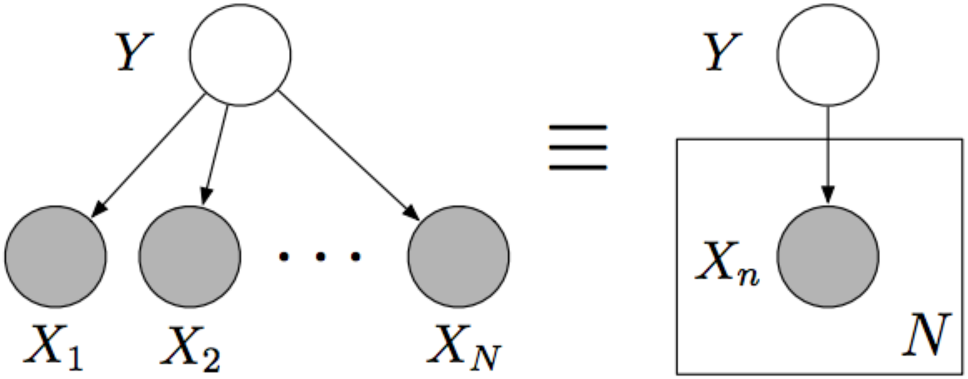
Probabilistic Graphical Models



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

Graphic courtesy https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf

Probabilistic Graphical Models, Cont



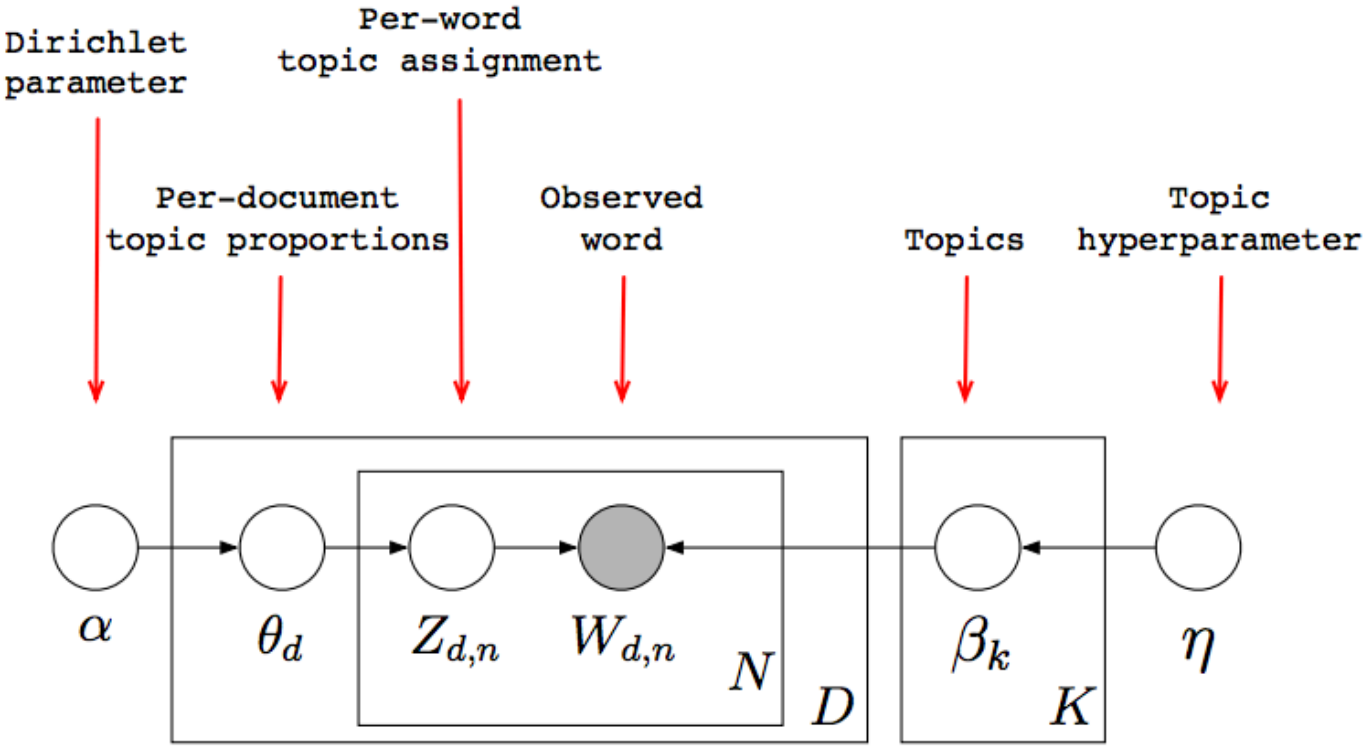
- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

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

Graphic courtesy https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf

LDA Graphical Model

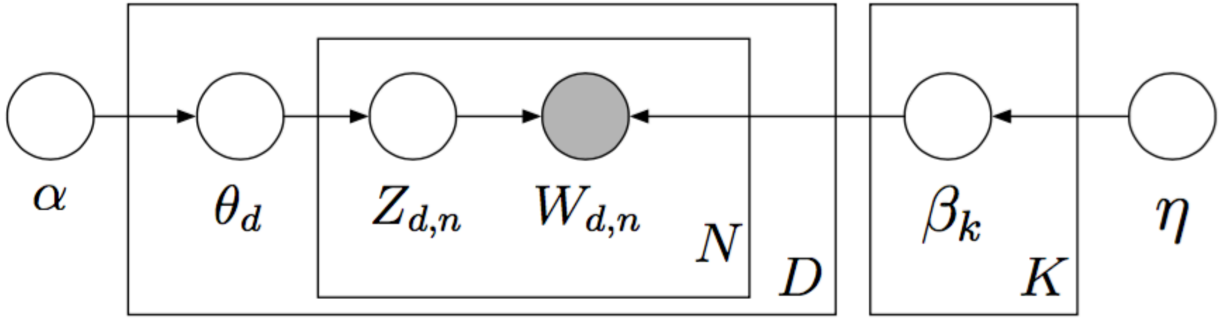
(based on the generative model)



Each piece of the structure is a random variable.

Graphic courtesy https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf

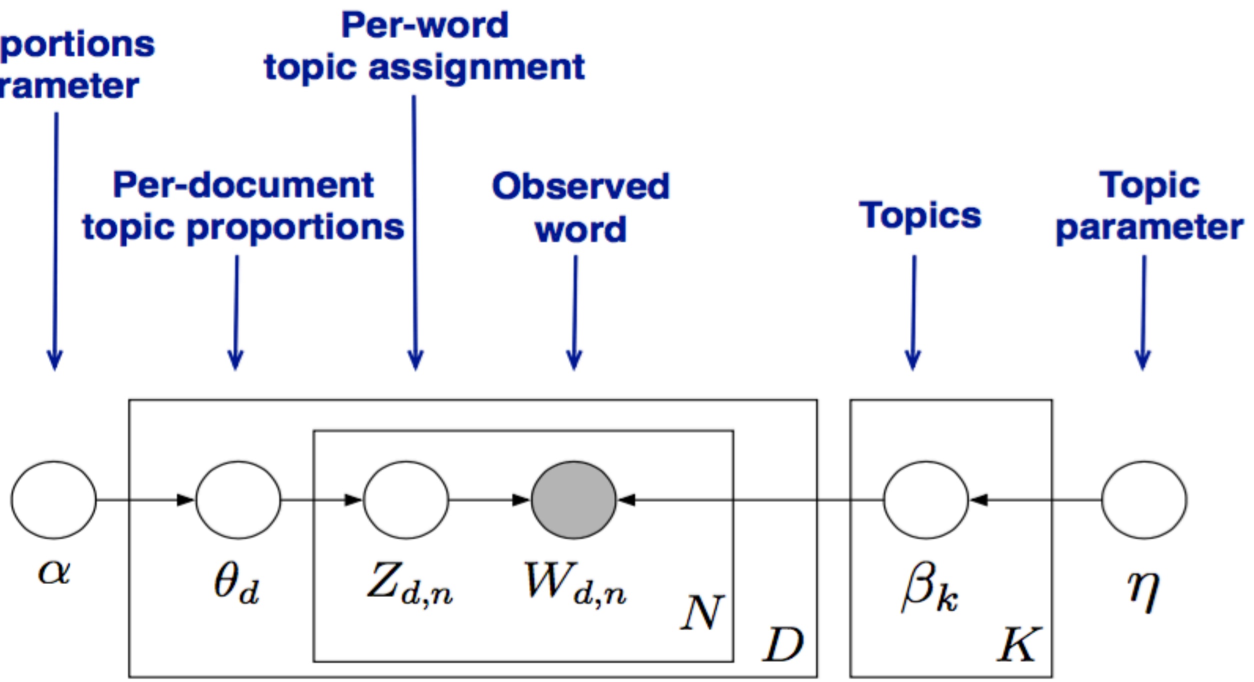
LDA Graphical Model Details



- 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

- Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.

What is the LDA Joint Distribution?

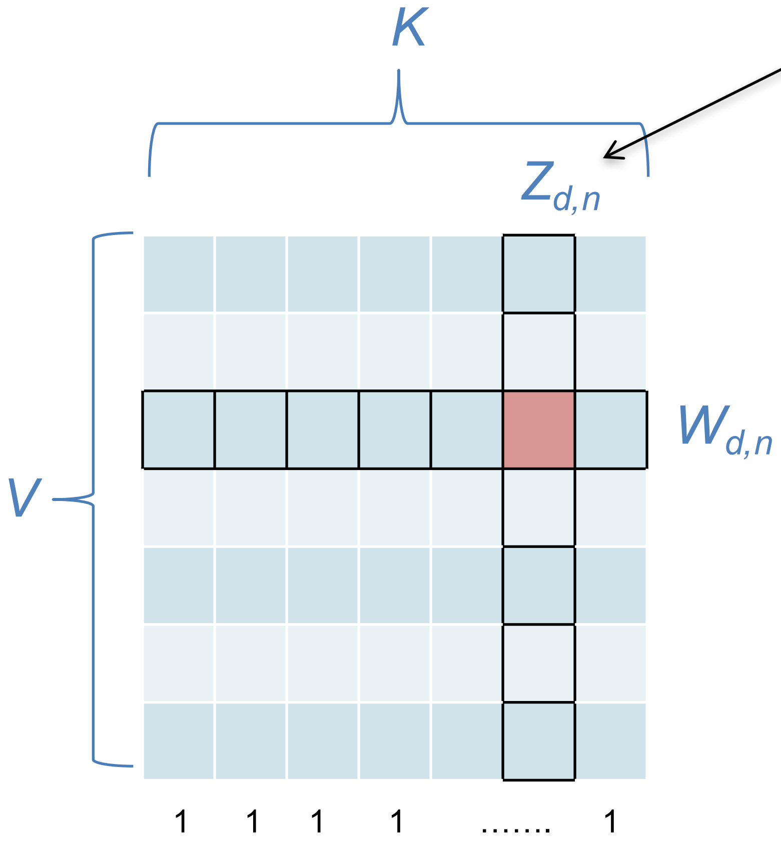


$$p(\beta,\theta,\mathbf{z},\mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta)\right)\left(\prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n})\right)$$

Why does $w_{d,n}$ depend on $z_{d,n}$ and β?

V x K Topic Matrix

Columns are the β_ks



K	-- number of topics
V	-- number of words in the vocabulary
TM_{VxK}	-- topic matrix
$Z_{d,n}$	-- index of topic which $W_{d,n}$ comes from
$W_{d,n}$	-- the nth word in the dth document
$TM[W_{d,n,}Z_{d,n}]$	-- the probability of $W_{d,n}$, $\beta_k[w_{d,n}]$ $(k = z_{d,n})$

Now, What Exactly is the Dirichlet Distribution

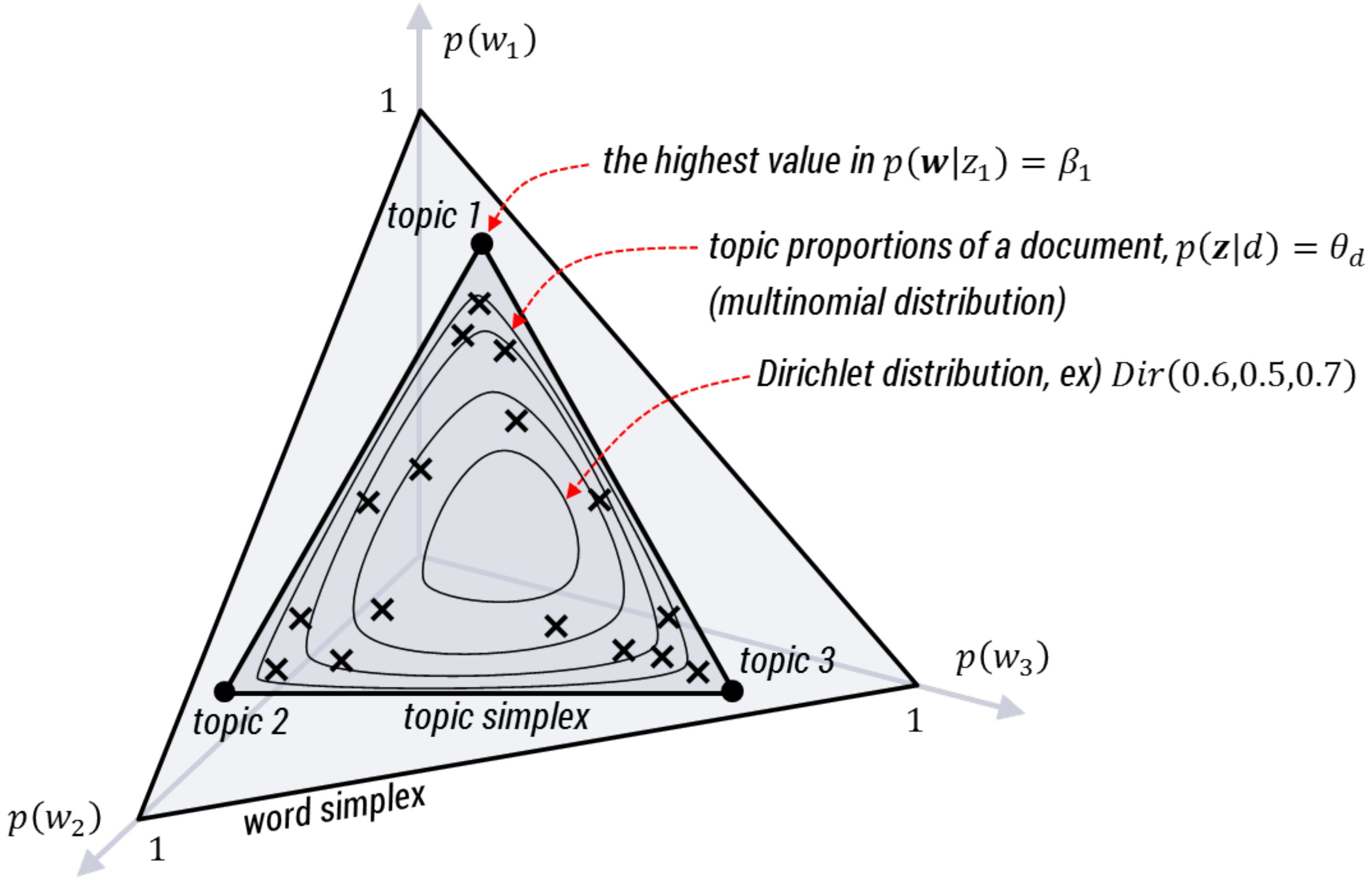
(and why are we using it?)

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

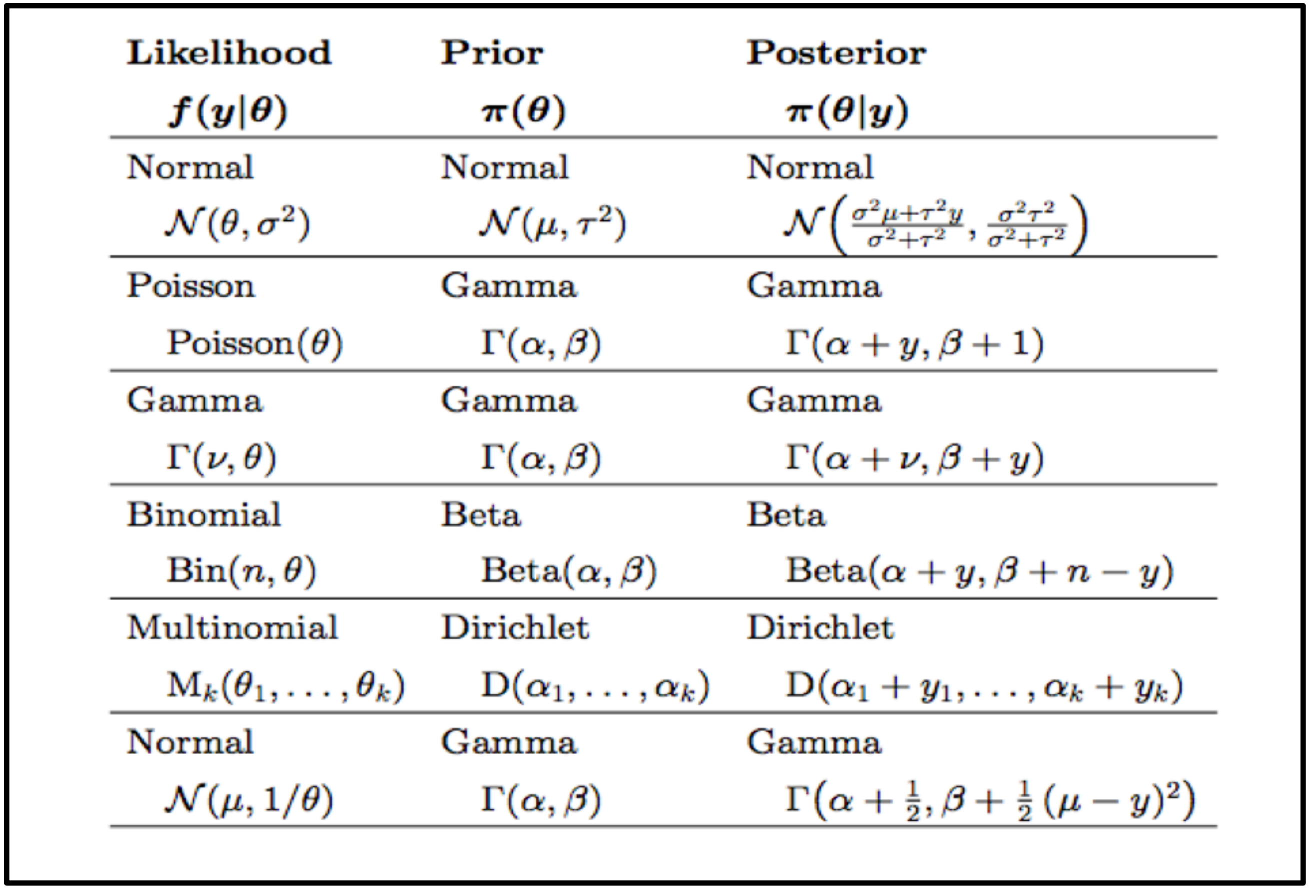
- The Dirichlet is a "dice factory"
  - Multivariate equivalent of the Beta distribution ("coin factory")
  - Parameters α determine the form of the prior
- The Dirichlet is defined over the (k-1) simplex
  - The k non-negative arguments which sum to one
- The Dirichlet is the ***conjugate prior*** to the multinomial distribution
  - If the likelihood has conjugate prior P then the posterior has the same form as P
    - If we have a conjugate prior we know the (closed) form of the posterior
    - So in this case the posterior is also a Dirichlet
- The parameter α controls the mean shape and sparsity of θ
- In LDA the topics are a V-dimensional Dirichlet and the topic proportions are a K-dimensional Dirichlet

Simplex?

(space of non-negative vectors which sum to one)


$p(w_1)$
1
the highest value in $p(\mathbf{w}|z_1) = \beta_1$
topic 1
topic proportions of a document, $p(\mathbf{z}|d) = \theta_d$
(multinomial distribution)
Dirichlet distribution, ex) $Dir(0.6,0.5,0.7)$
topic 3
$p(w_3)$
1
topic 2
topic simplex
word simplex
$p(w_2)$
1

Aside: Conjugate Priors



Likelihood	**Prior**	**Posterior**
$f(y\|\theta)$	$\pi(\theta)$	$\pi(\theta\|y)$
Normal $\mathcal{N}(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	Normal $\mathcal{N}\left(\frac{\sigma^2\mu+\tau^2 y}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$
Poisson $\text{Poisson}(\theta)$	Gamma $\Gamma(\alpha,\beta)$	Gamma $\Gamma(\alpha+y,\beta+1)$
Gamma $\Gamma(\nu,\theta)$	Gamma $\Gamma(\alpha,\beta)$	Gamma $\Gamma(\alpha+\nu,\beta+y)$
Binomial $\text{Bin}(n,\theta)$	Beta $\text{Beta}(\alpha,\beta)$	Beta $\text{Beta}(\alpha+y,\beta+n-y)$
Multinomial $\text{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\text{D}(\alpha_1,\ldots,\alpha_k)$	Dirichlet $\text{D}(\alpha_1+y_1,\ldots,\alpha_k+y_k)$
Normal $\mathcal{N}(\mu,1/\theta)$	Gamma $\Gamma(\alpha,\beta)$	Gamma $\Gamma(\alpha+\frac{1}{2},\beta+\frac{1}{2}(\mu-y)^2)$

Notes/Issues:

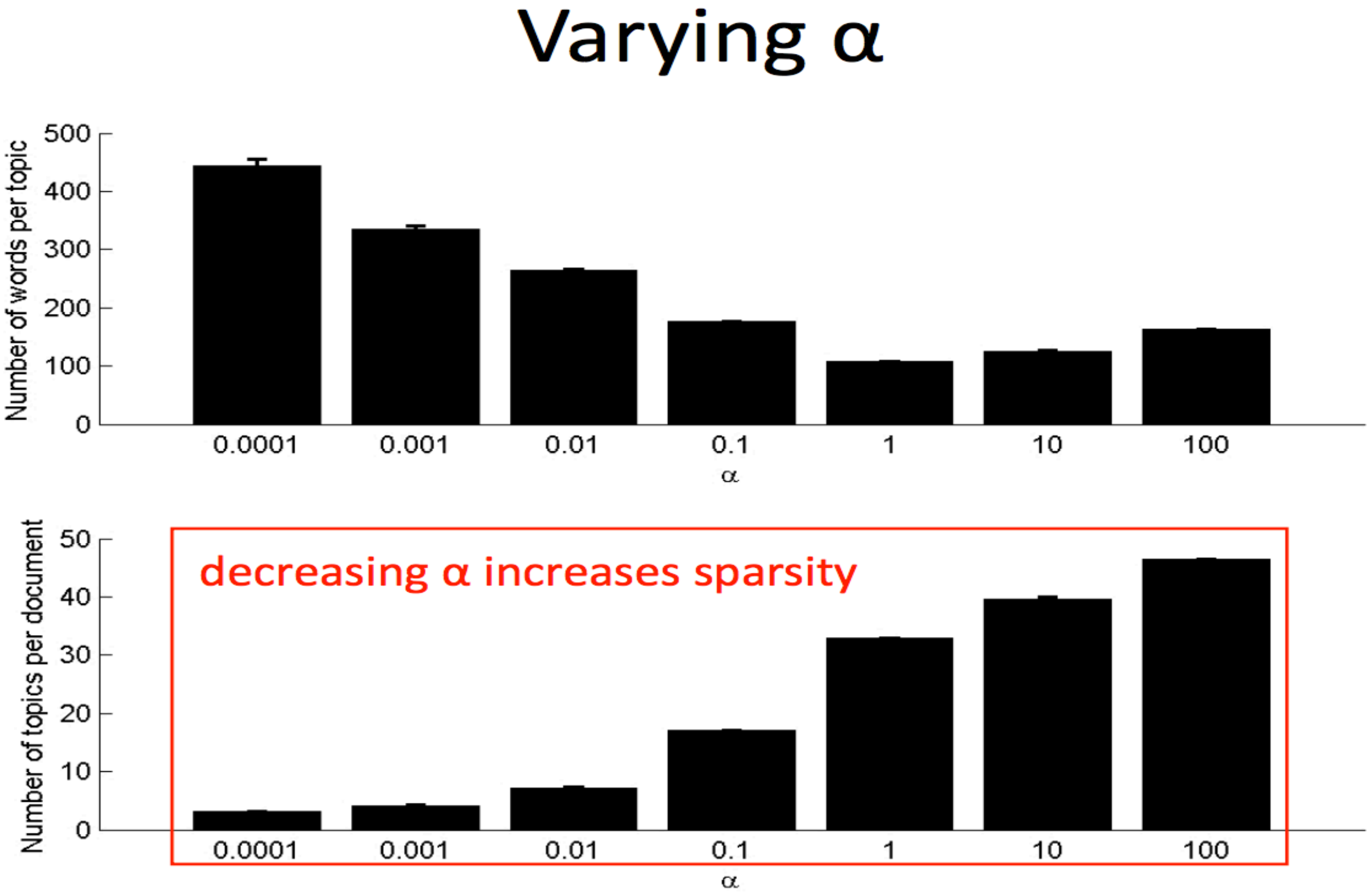
- Conjugate Prior: If the likelihood has conjugate prior P then the posterior has the same form as P
- The conjugate prior might not correctly reflect our uncertainty about θ
- Non-conjugate priors are typically not available in analytic (closed) form
- It is difficult to quantify our uncertainty about θ in the form of a particular distribution

Agenda

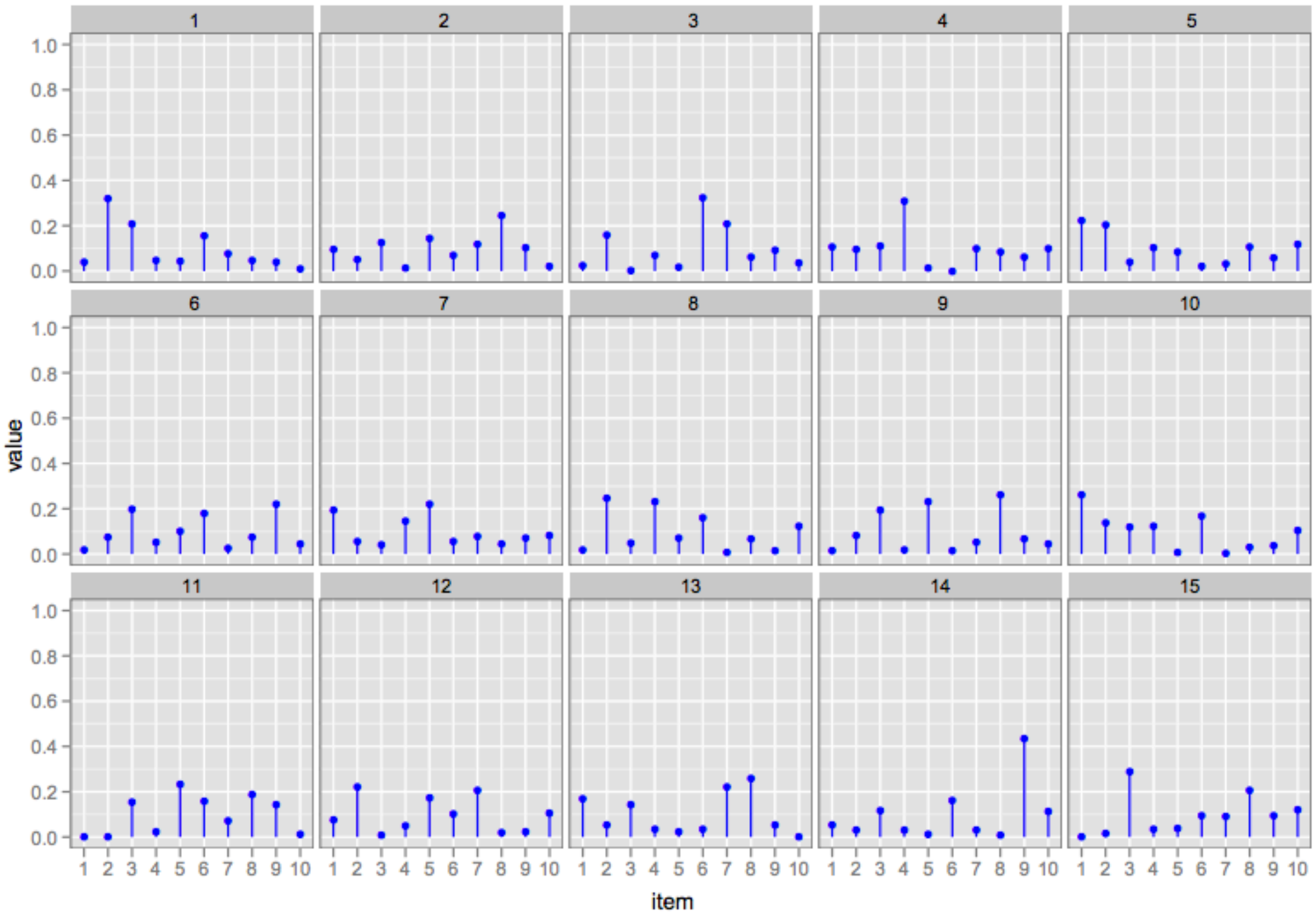
- ~~What is Topic Modeling?~~
- ~~Parametric vs. Non-Parametric Models~~
- ~~Latent Dirichlet Allocation~~
- ~~Probabilistic Graphical Models~~
- The Effect of the Dirichlet parameter α
- Dynamic LDA
- Q&A

The Effect of the Dirichlet parameter α

Generalizing the Idea of Co-Occurrence

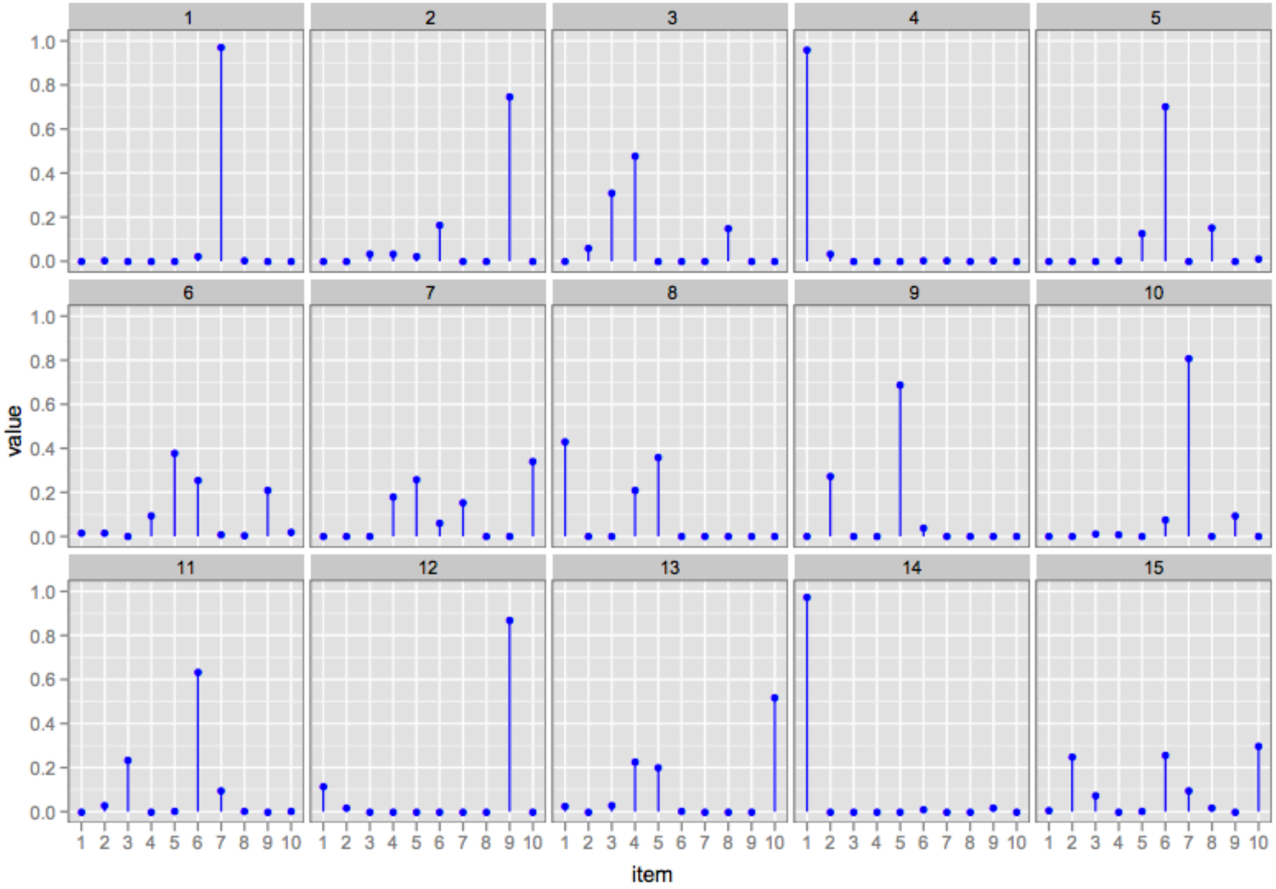


α = 1



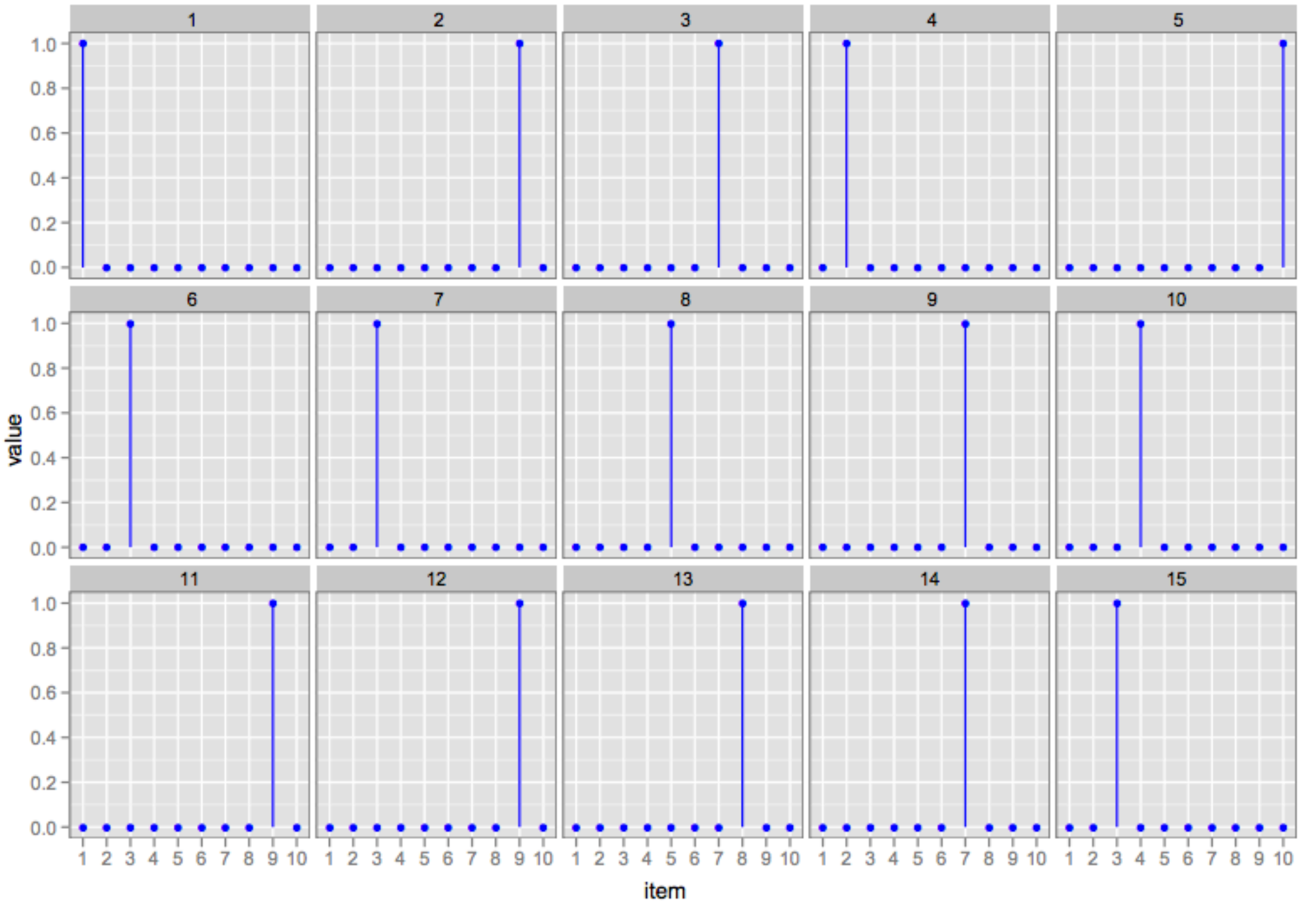
Graphic courtesy https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf

$\alpha = 0.1$



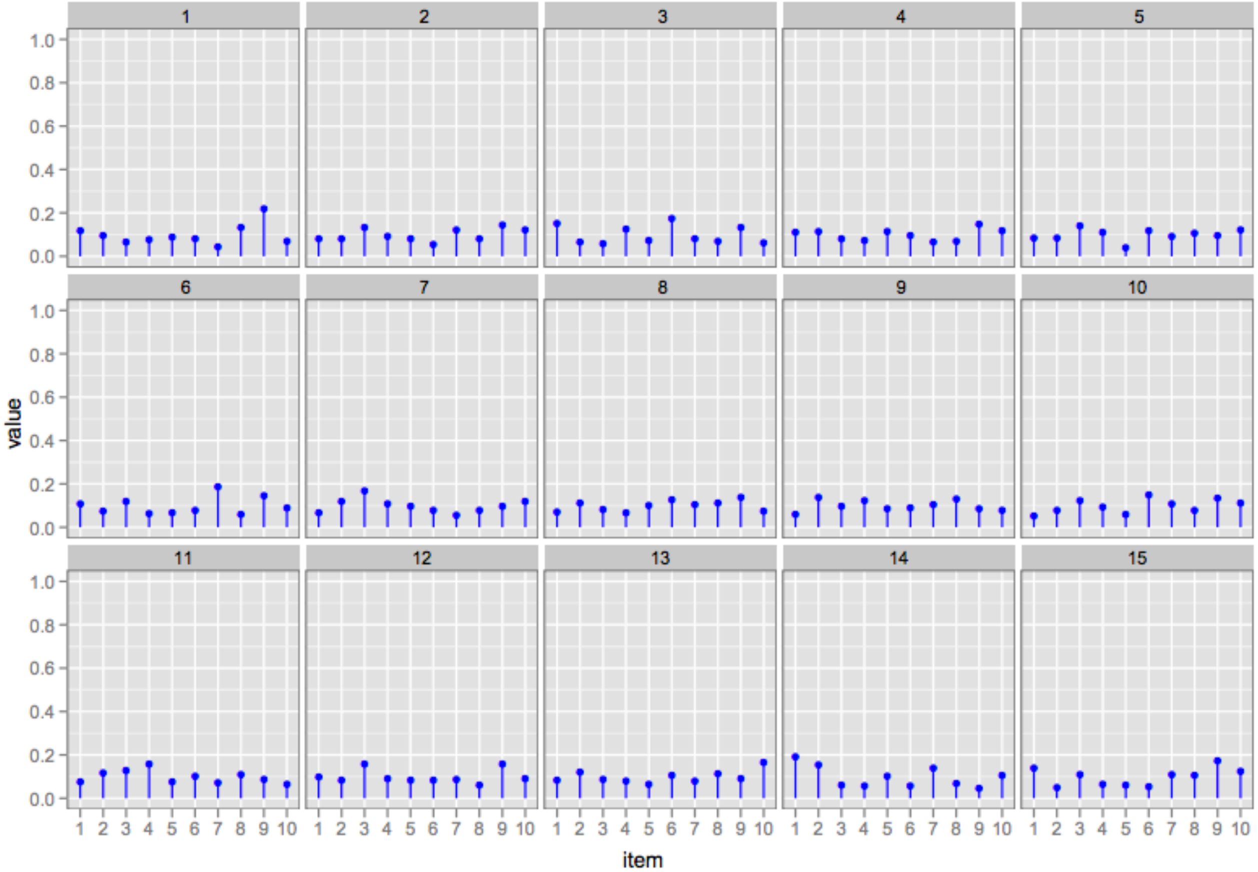
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

α = 0.001



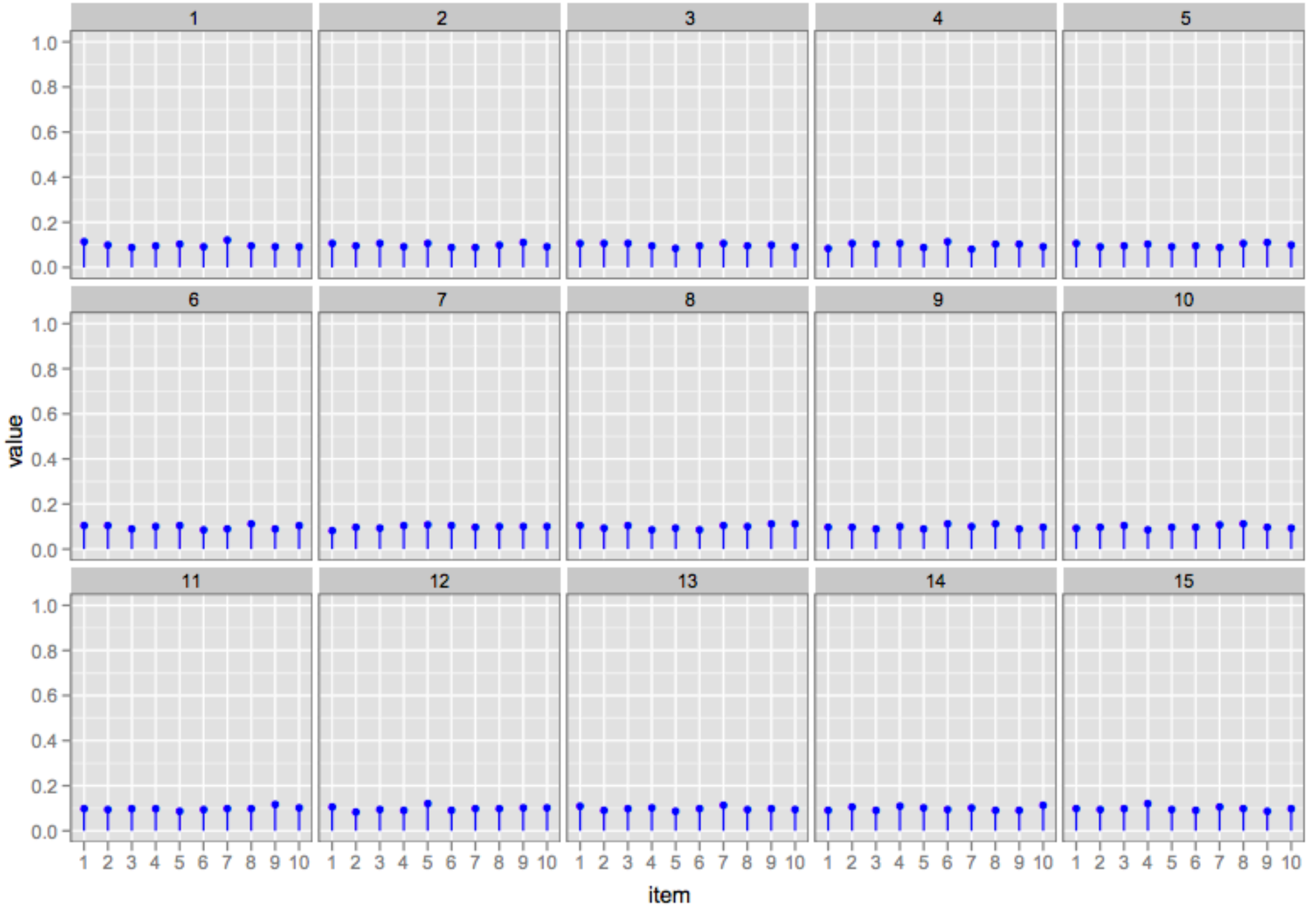
Essentially a mixture model (one topic/document)

α = 10



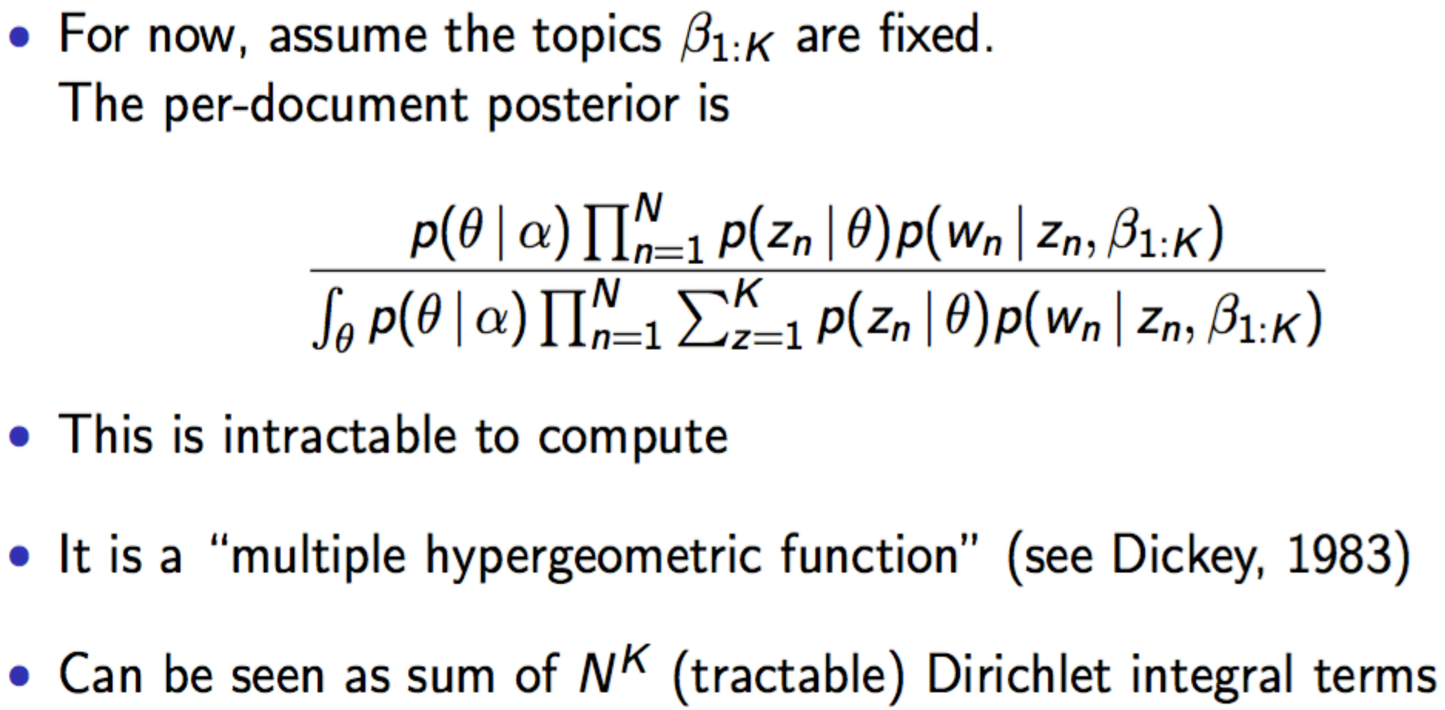
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

$\alpha = 100$



Larger α spreads out the mass

Ok, but we need the Posterior



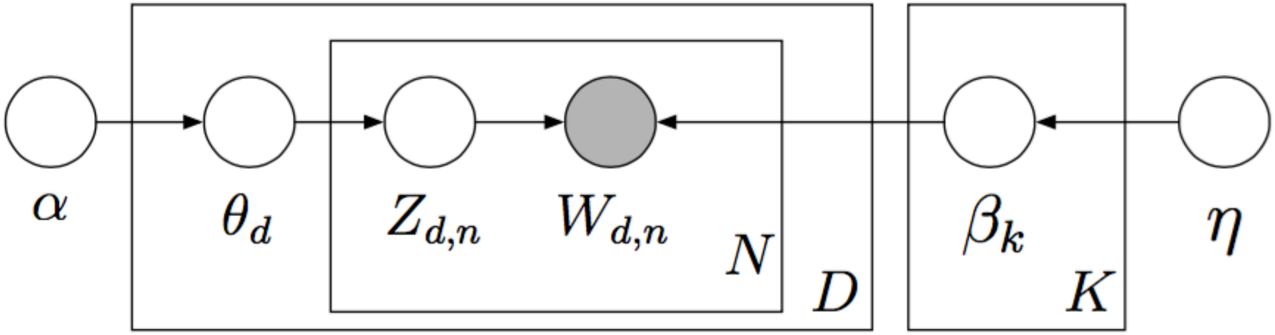
- For now, assume the topics $\beta_{1:K}$ are fixed.
  The per-document posterior is

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

- This is intractable to compute
- It is a "multiple hypergeometric function" (see Dickey, 1983)
- Can be seen as sum of N^K (tractable) Dirichlet integral terms

So we need to use approximate inference

Approximate Inference



We appeal to approximate posterior inference of the posterior,

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

- Gibbs sampling
- Variational methods
- Particle filtering

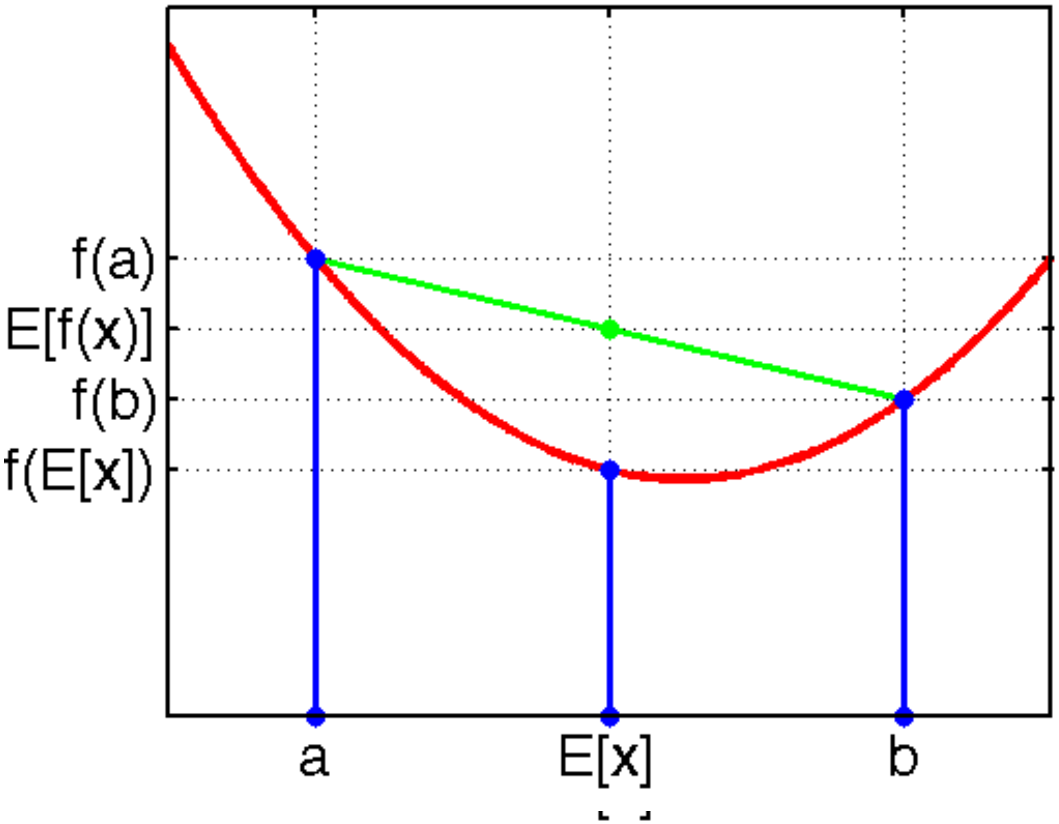
Variational Inference

- Variational methods are a deterministic alternative to MCMC.
- Let $x_{1:N}$ be observations and $z_{1:M}$ be latent variables
- Our goal is to compute the posterior distribution

$$p(z_{1:M} \,|\, x_{1:N}) = \frac{p(z_{1:M}, x_{1:N})}{\int p(z_{1:M}, x_{1:N}) dz_{1:M}}$$

- For many interesting distributions, the marginal likelihood of the observations is difficult to efficiently compute

Jensen’s Inequality



- Jensen's inequality generalizes the observation that the secant line of a convex function lies *above* the graph of the function

- In probability theory Jensen’s Inequality is generally stated in the following form: If X is a random variable and φ is a *convex* function, then φ(E[X]) ≤ E[φ(X)]

Jensen's Inequality, Bounds, and KL Divergence

- Use Jensen's inequality to bound the log prob of the observations:

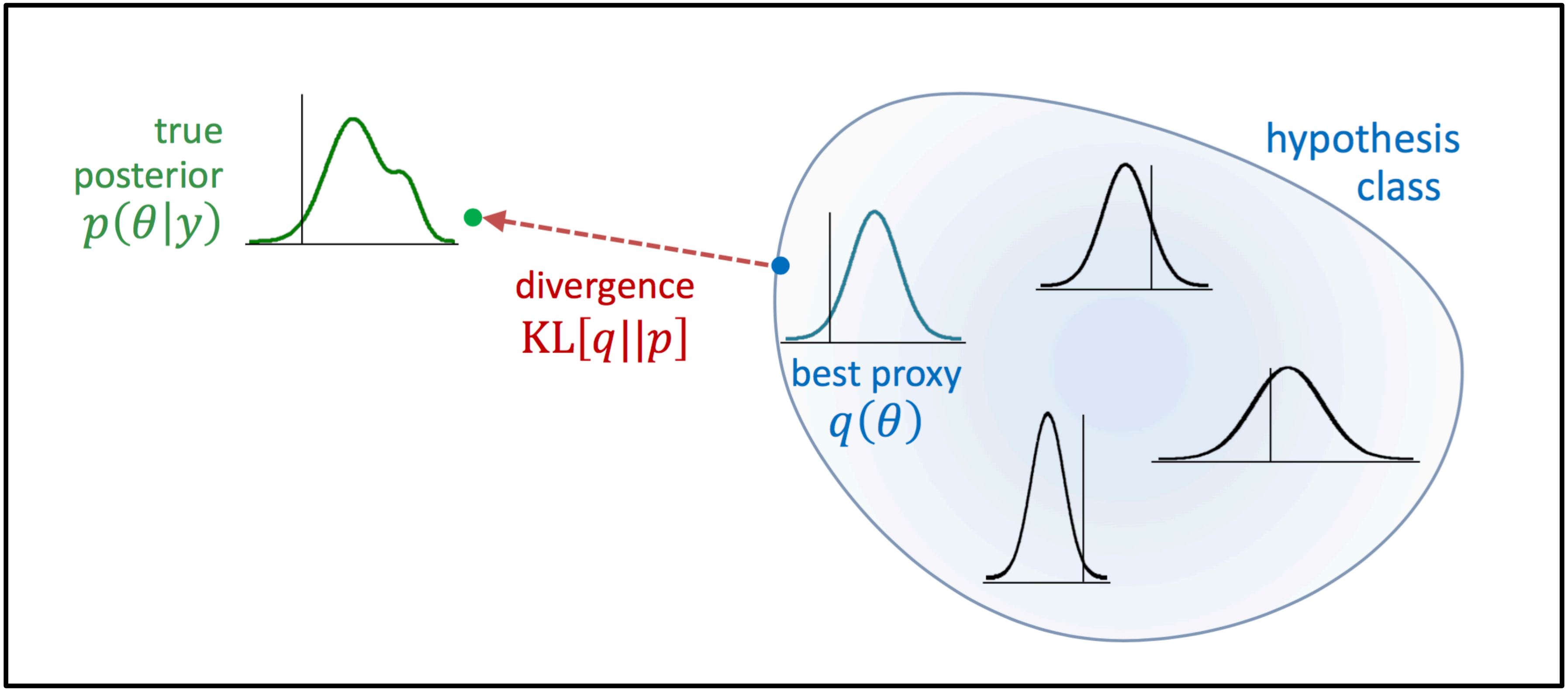
$$\begin{aligned}
\log p(x_{1:N}) &= \log \int p(z_{1:M}, x_{1:N}) dz_{1:M} \\
&= \log \int p(z_{1:M}, x_{1:N}) \frac{q_\nu(z_{1:M})}{q_\nu(z_{1:M})} dz_{1:M} \\
&\geq \mathrm{E}_{q_\nu}[\log p(z_{1:M}, x_{1:N})] - \mathrm{E}_{q_\nu}[\log q_\nu(z_{1:M})]
\end{aligned}$$

- We have introduced a distribution of the latent variables with free *variational parameters* ν.

- We optimize those parameters to tighten this bound.

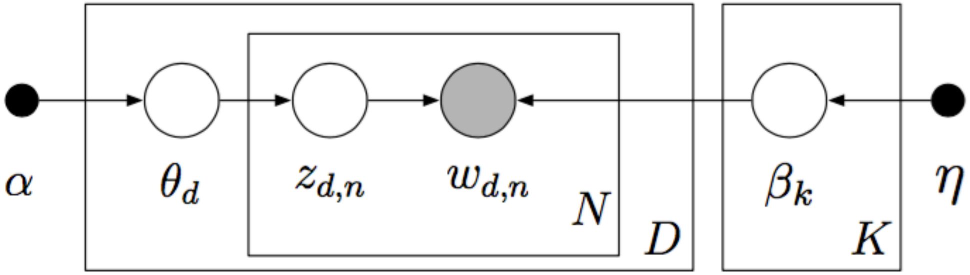
- This is the same as finding the member of the family q_ν that is closest in KL divergence to $p(z_{1:M} \mid x_{1:N})$.

KL-Divergence



$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}.$$

Why Does LDA Work?



- LDA trades off two goals

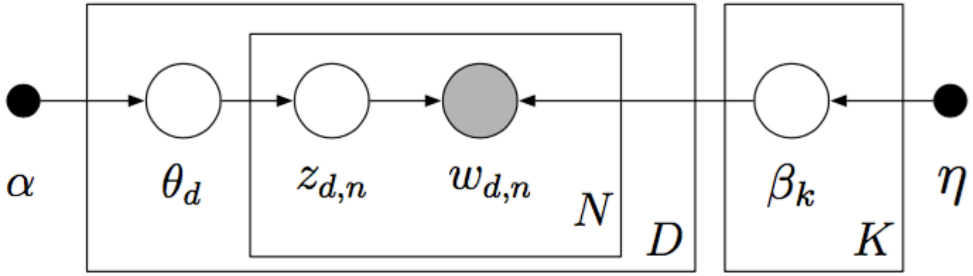
  1. In each **document**, allocate its words to **few topics**.
  2. In each **topic**, assign high probability to **few terms**.

- We see this from the joint

$$\log p(\cdot) = \ldots + \sum_d \sum_n \log p(z_{dn} \mid \theta_d) + \log p(w_{dn} \mid \beta_{z_{dn}}) + \ldots$$

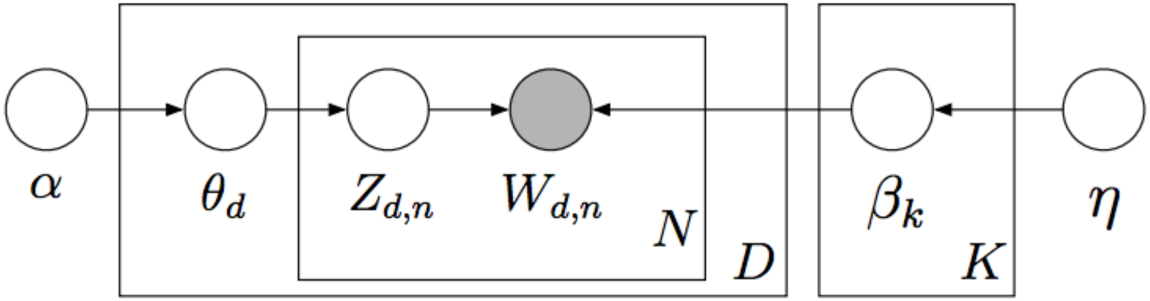
- Sparse proportions come from the 1st term.
  Sparse topics come from the 2nd term.

Why Does LDA Work?



- LDA trades off two goals
  1. In each **document**, allocate its words to **few topics**.
  2. In each **topic**, assign high probability to **few terms**.
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard.
  - Putting very few words in each topic makes #1 hard.
- Trading off these goals finds groups of tightly co-occurring words.

LDA Summary

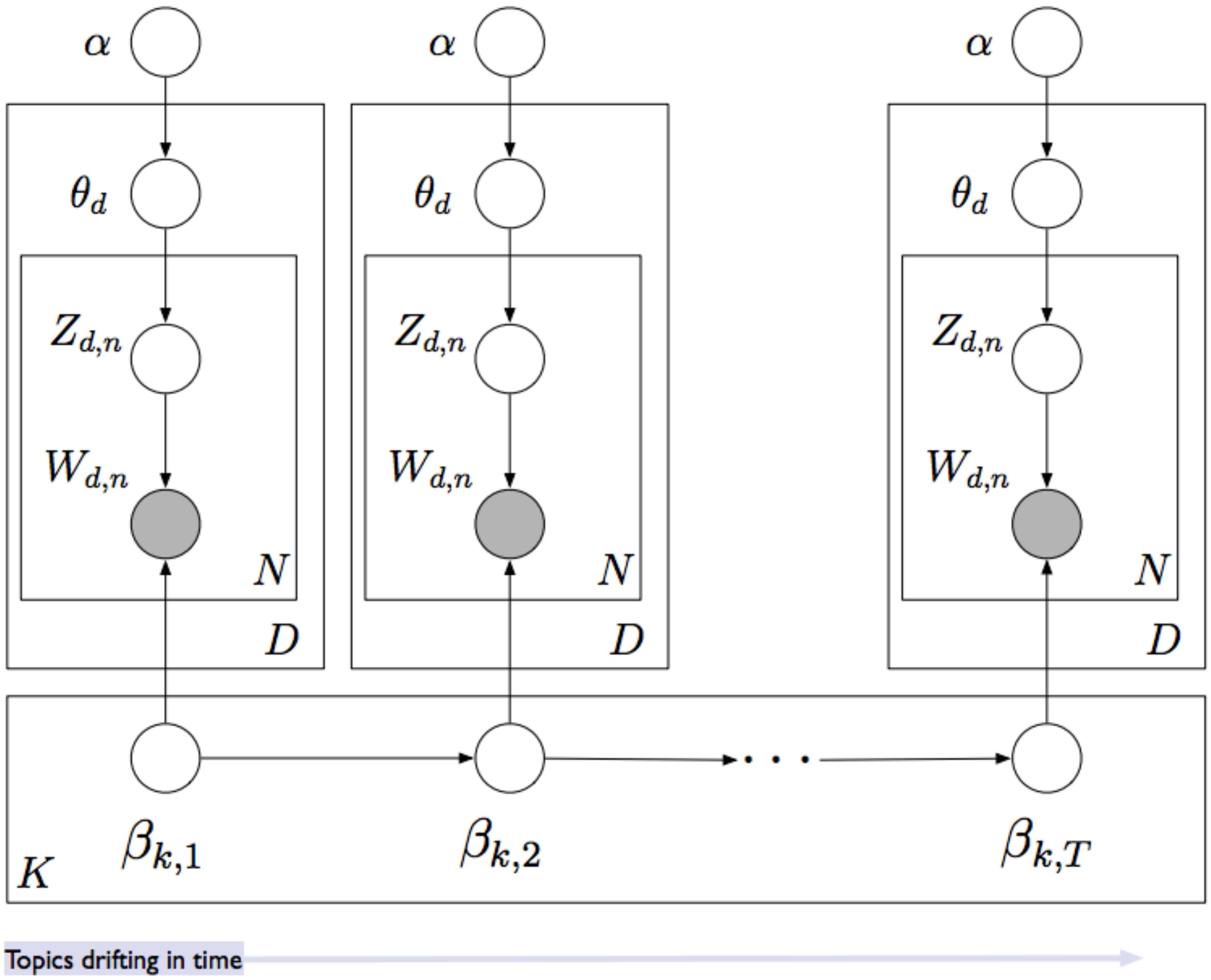


- LDA can
  - visualize the hidden thematic structure in large corpora
  - generalize new data to fit into that structure
- Builds on Deerwester et al. (1990) and Hofmann (1999)
  It is a *mixed membership model* (Erosheva, 2004).
  Relates to *multinomial PCA* (Jakulin and Buntine, 2002)
- Was independently invented for genetics (Pritchard et al., 2000)

Agenda

- ~~What is Topic Modeling?~~
- ~~Parametric vs. Non-Parametric Models~~
- ~~Latent Dirichlet Allocation~~
- ~~Probabilistic Graphical Models~~
- ~~The Effect of the Dirichlet parameter α~~
- Dynamic LDA
- Q&A

Beyond LDA: Dynamic Topic Models



Graphic courtesy https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf

Q&A

Thanks!