School of Computer Science
Carnegie Mellon

10-701 Introduction to Machine Learning

Topic Modeling

Readings:
Blei, Ng, & Jordan (2003)
Griffiths & Steyvers (2004)

Matt Gormley
Lecture 20
November 16, 2016

Reminders

- Homework 4
  - deadline extended to Wed, Nov. 16th
  - 10 extra points for submitting by Mon, Nov. 14th
- Poster Sessions
  - two sessions on Fri, Dec. 2nd
  - session 1: 8 - 11:30 am
  - session 2: 2 - 6 pm

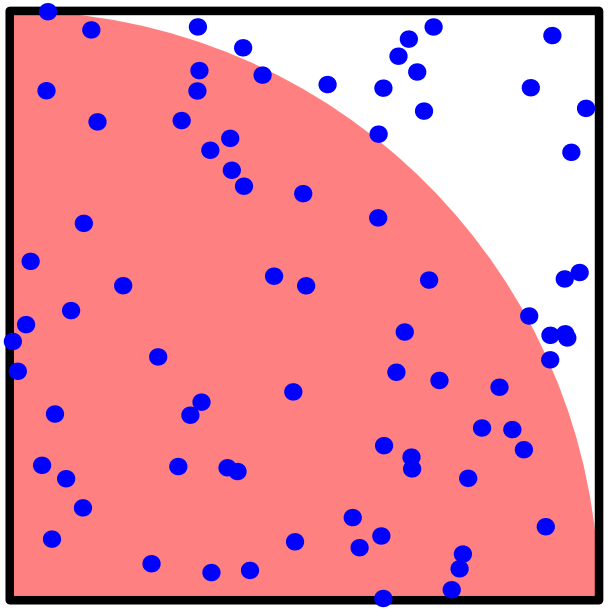
Outline

- **MCMC**
  - Monte Carlo, Rejection Sampling
  - Gibbs Sampling
- **Applications of Topic Modeling**
- **Latent Dirichlet Allocation (LDA)**
  1. Beta-Bernoulli
  2. Dirichlet-Multinomial
  3. Dirichlet-Multinomial Mixture Model
  4. LDA
- **Contrast of methods for Inference / Learning**
  - Exact inference
  - EM
  - Monte Carlo EM
  - Gibbs sampler
  - Collapsed Gibbs sampler
- **Extensions of LDA**
  - Correlated topic models
  - Dynamic topic models
  - Polylingual topic models
  - Supervised LDA

Monte Carlo, Rejection Sampling

SAMPLING BASICS

A dumb approximation of π



$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2)<1\right) P(x,y)\, \mathrm{d}x\, \mathrm{d}y$$

```
octave:1> S=12; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.3333
octave:2> S=1e7; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.1418
```

Aside: don't always sample!

"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."

— Alan Sokal, 1996

Example: numerical solutions to (nice) 1D integrals are fast

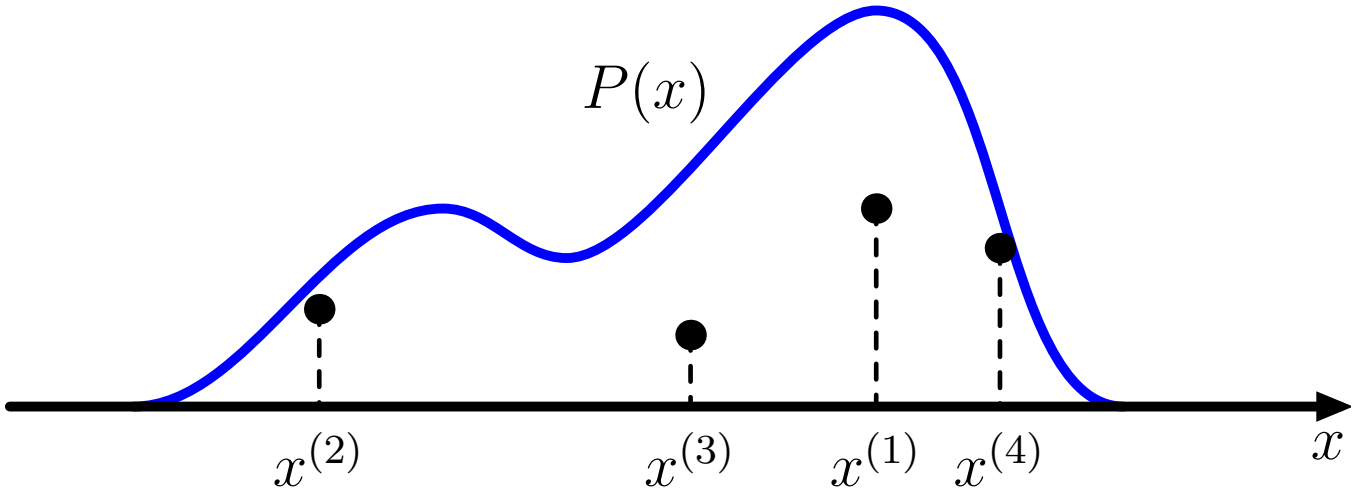
```
octave:1> 4 * quadl(@(x) sqrt(1-x.^2), 0, 1, tolerance)
```

Gives π to 6 dp's in 108 evaluations, machine precision in 2598.

(NB Matlab's `quadl` fails at zero tolerance)

Sampling from distributions

Draw points uniformly under the curve:



Probability mass to left of point $\sim$ Uniform[0,1]

Sampling from distributions

How to convert samples from a Uniform[0,1] generator:

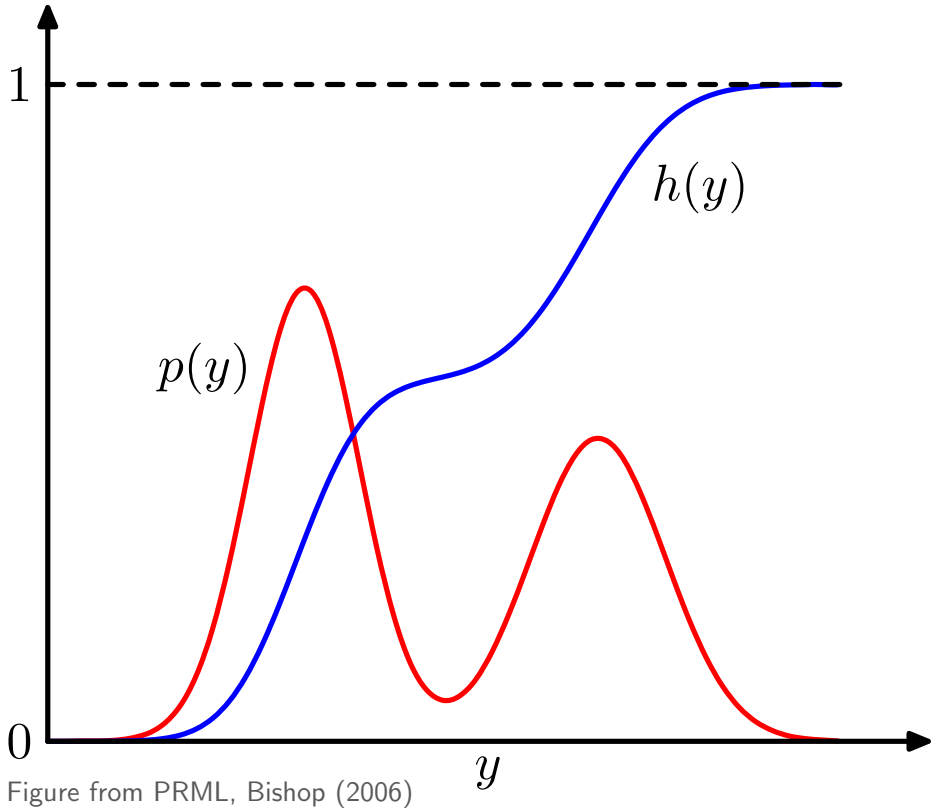


Figure from PRML, Bishop (2006)

$$h(y) = \int_{-\infty}^{y} p(y')\,\mathrm{d}y'$$

Draw mass to left of point:

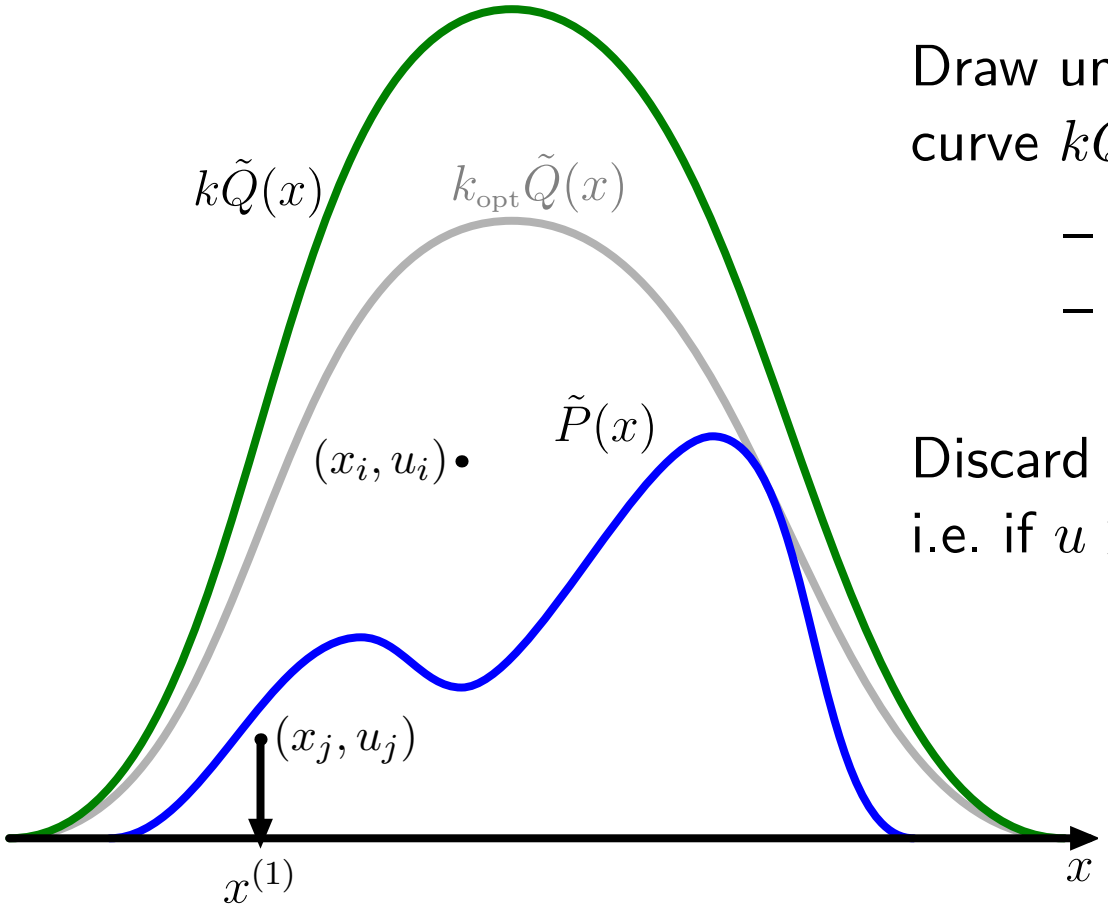
$u \sim$ Uniform[0,1]

Sample, $y(u) = h^{-1}(u)$

Although we can't always compute and invert $h(y)$

Rejection sampling

Sampling underneath a $\tilde{P}(x) \propto P(x)$ curve is also valid



Draw underneath a simple curve $k\tilde{Q}(x) \geq \tilde{P}(x)$:

- Draw $x \sim Q(x)$
- height $u \sim \text{Uniform}[0, k\tilde{Q}(x)]$

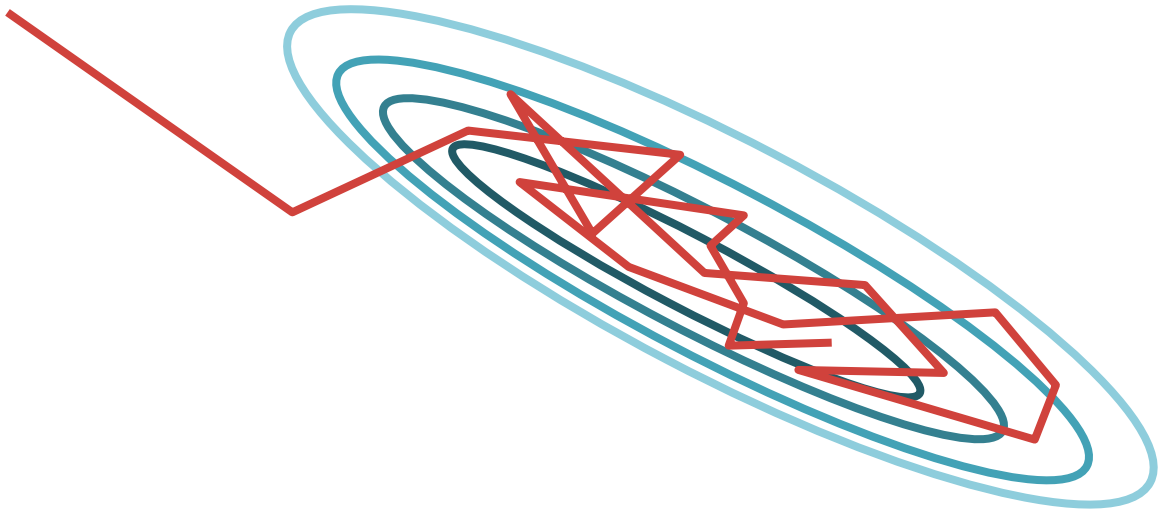
Discard the point if above $\tilde{P}$, i.e. if $u > \tilde{P}(x)$

Markov Chain Monte Carlo (MCMC)

GIBBS SAMPLING

MCMC

- **Goal:** Draw approximate, correlated samples from a target distribution p(x)
- **MCMC:** Performs a biased random walk to explore the distribution



Simulations of MCMC

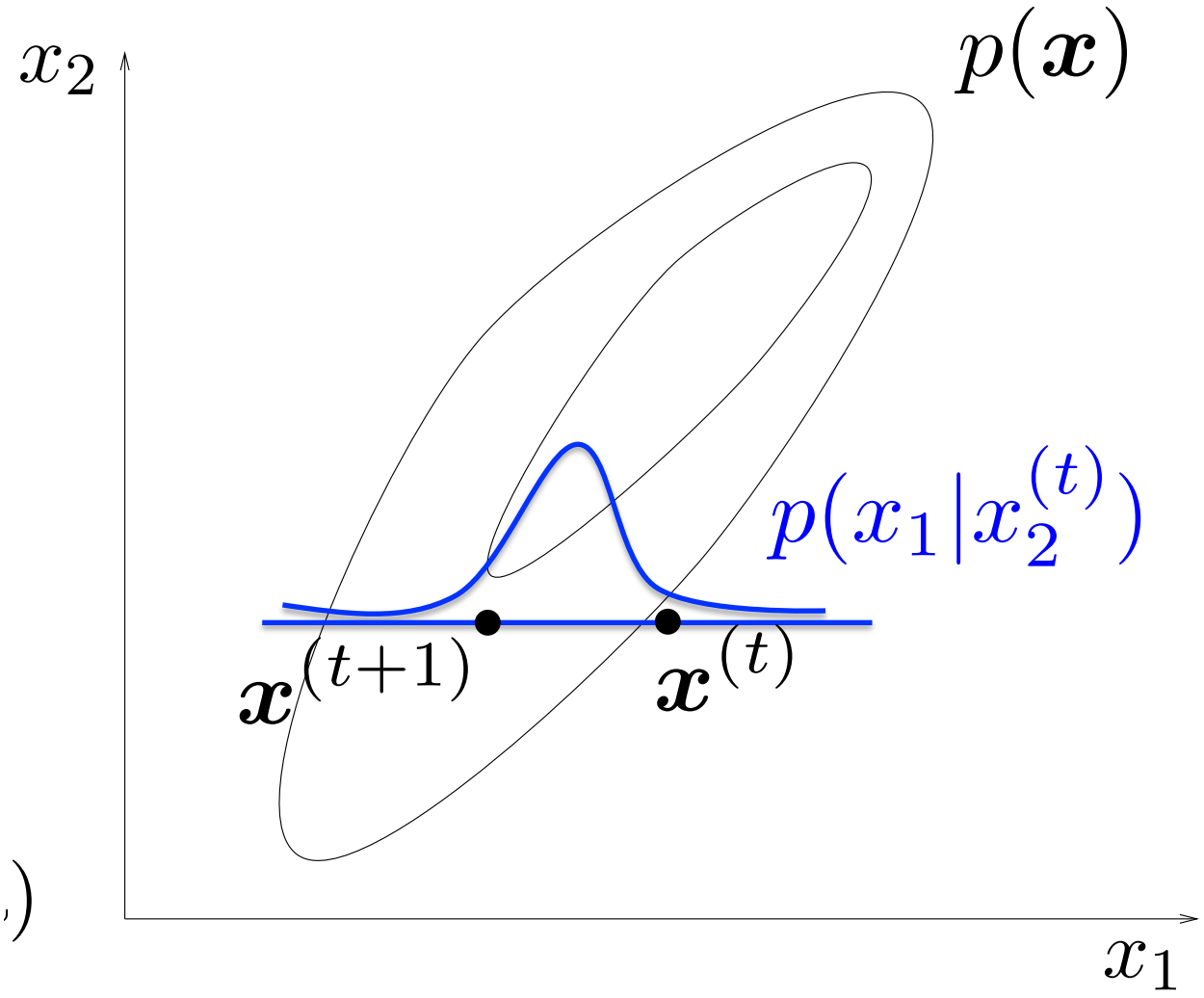
Visualization of Metroplis-Hastings, Gibbs Sampling, and Hamiltonian MCMC:

http://twiecki.github.io/blog/2014/01/02/visualizing-mcmc/

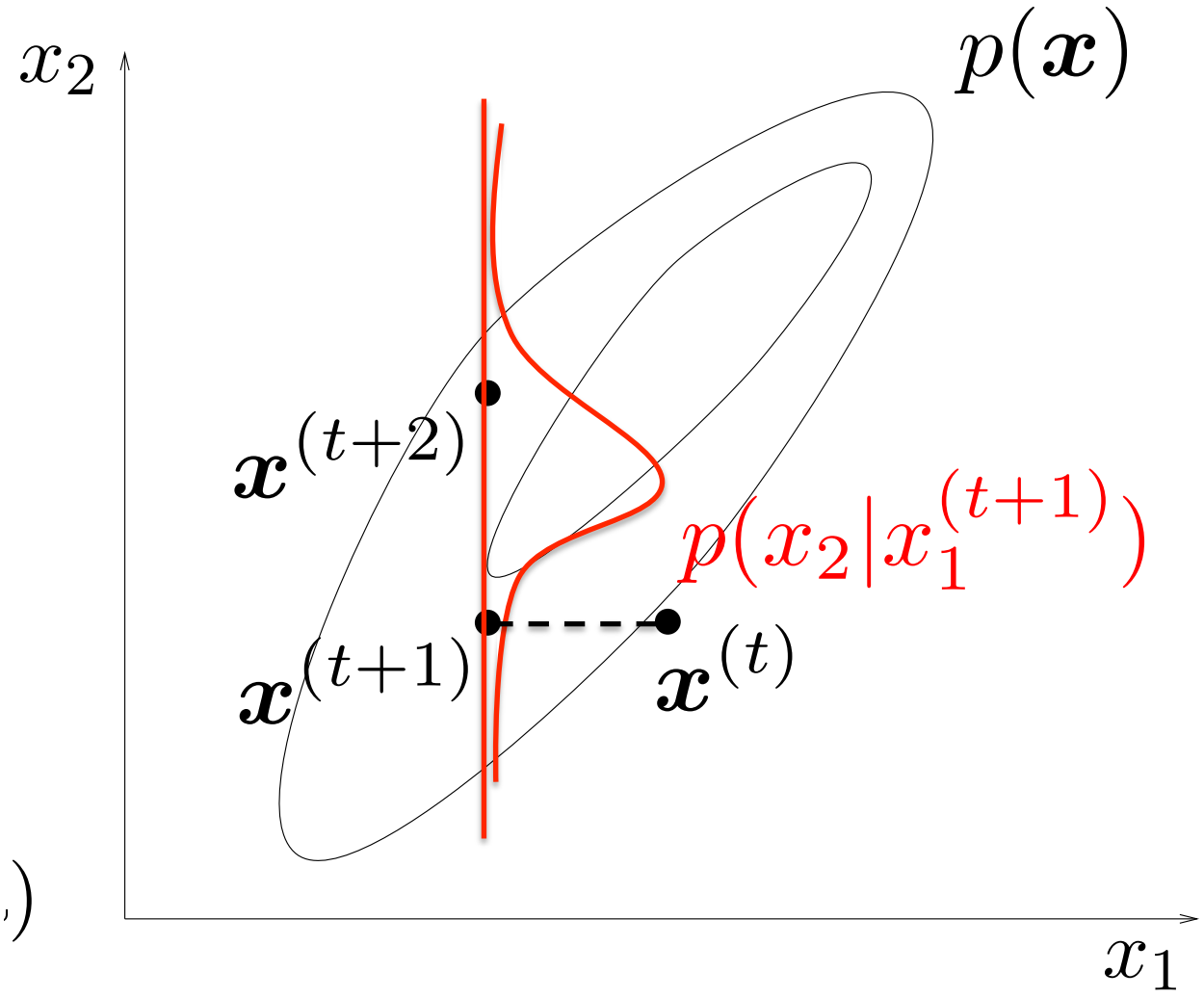
Whiteboard

- Gibbs Sampling psuedocode

Gibbs Sampling



Gibbs Sampling



Gibbs Sampling

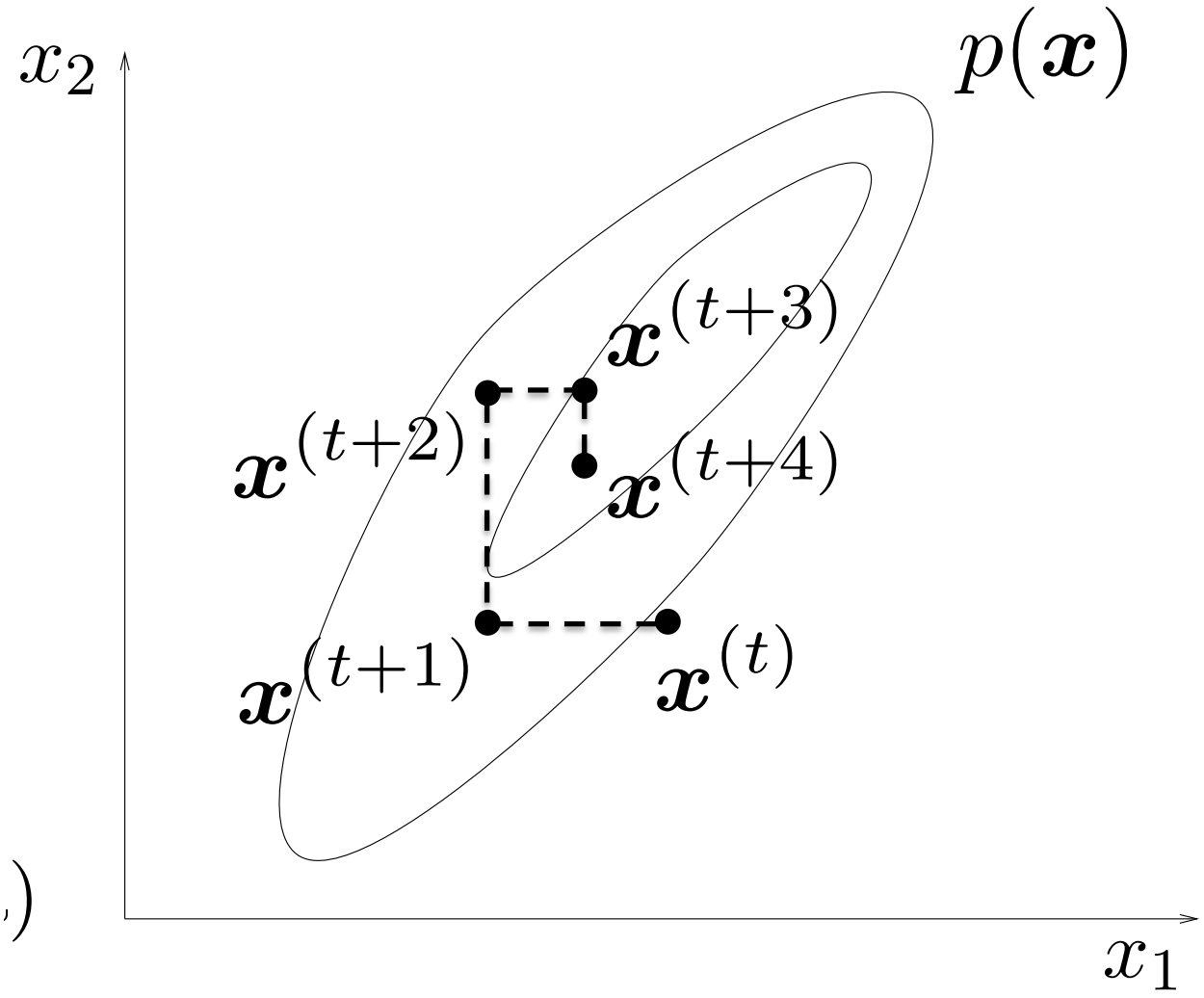
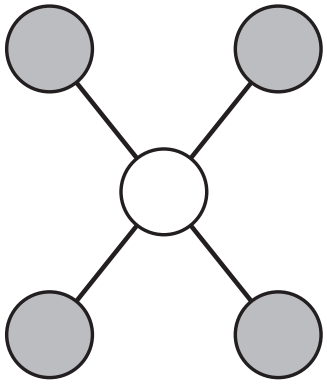


Figure from Jordan Ch. 21

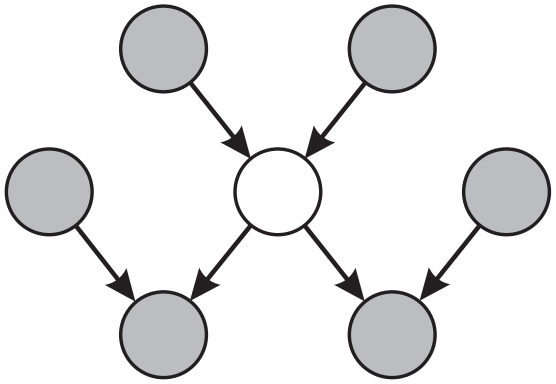
Gibbs Sampling

Full conditionals only need to condition on the Markov Blanket

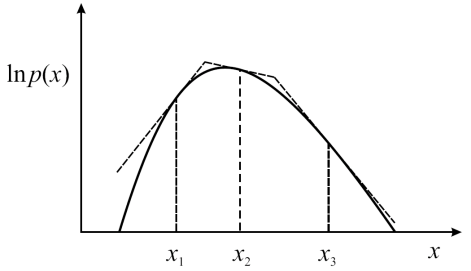
MRF



Bayes Net



- Must be "easy" to sample from conditionals
- Many conditionals are log-concave and are amenable to adaptive rejection sampling

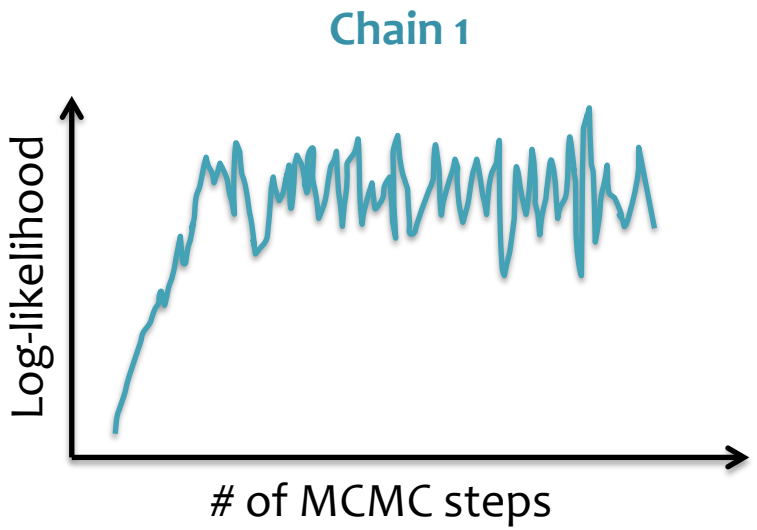


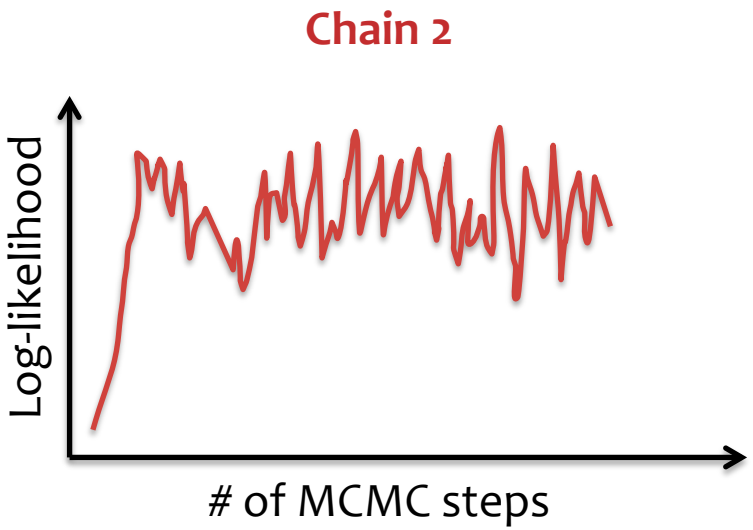
Why does Gibbs sampling work?

- Metropolis-Hastings
  - Markov chains
  - Stationary distribution
  - MH Algorithm
    - Constructs a Markov chain whose stationary distribution is the desired distribution
  - Proof that samples will be from desired distribution:
    - Sufficient conditions for constructing a markov chain with desired stationary distribution:
      - ergodicity
      - detailed balance (stronger, than what we need, but easier for the proof)
- Gibbs Sampling is a special case of Metropolis-Hastings
  - a special proposal distribution, which ensures the hastings ratio is always 1.0

Practical Issues

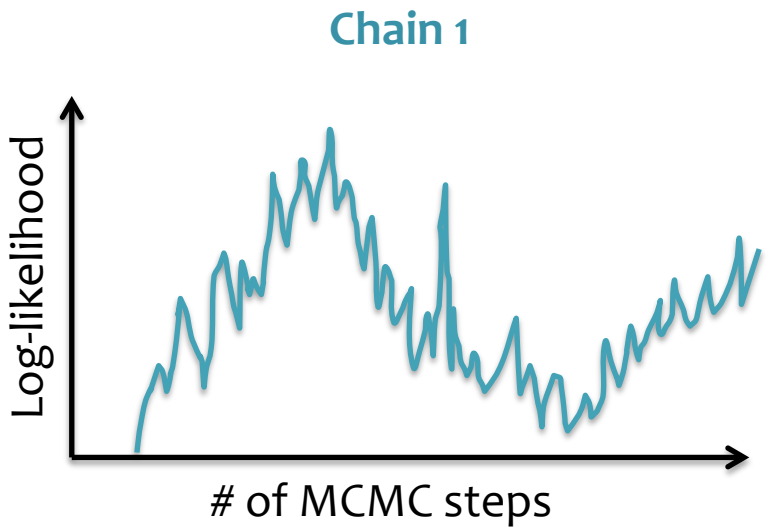
- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
  - Compare statistics of multiple independent chains
  - Ex: Compare log-likelihoods

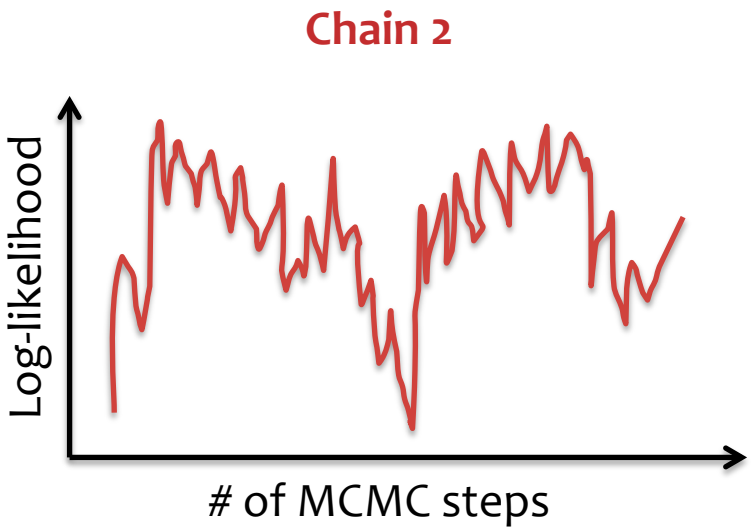




Practical Issues

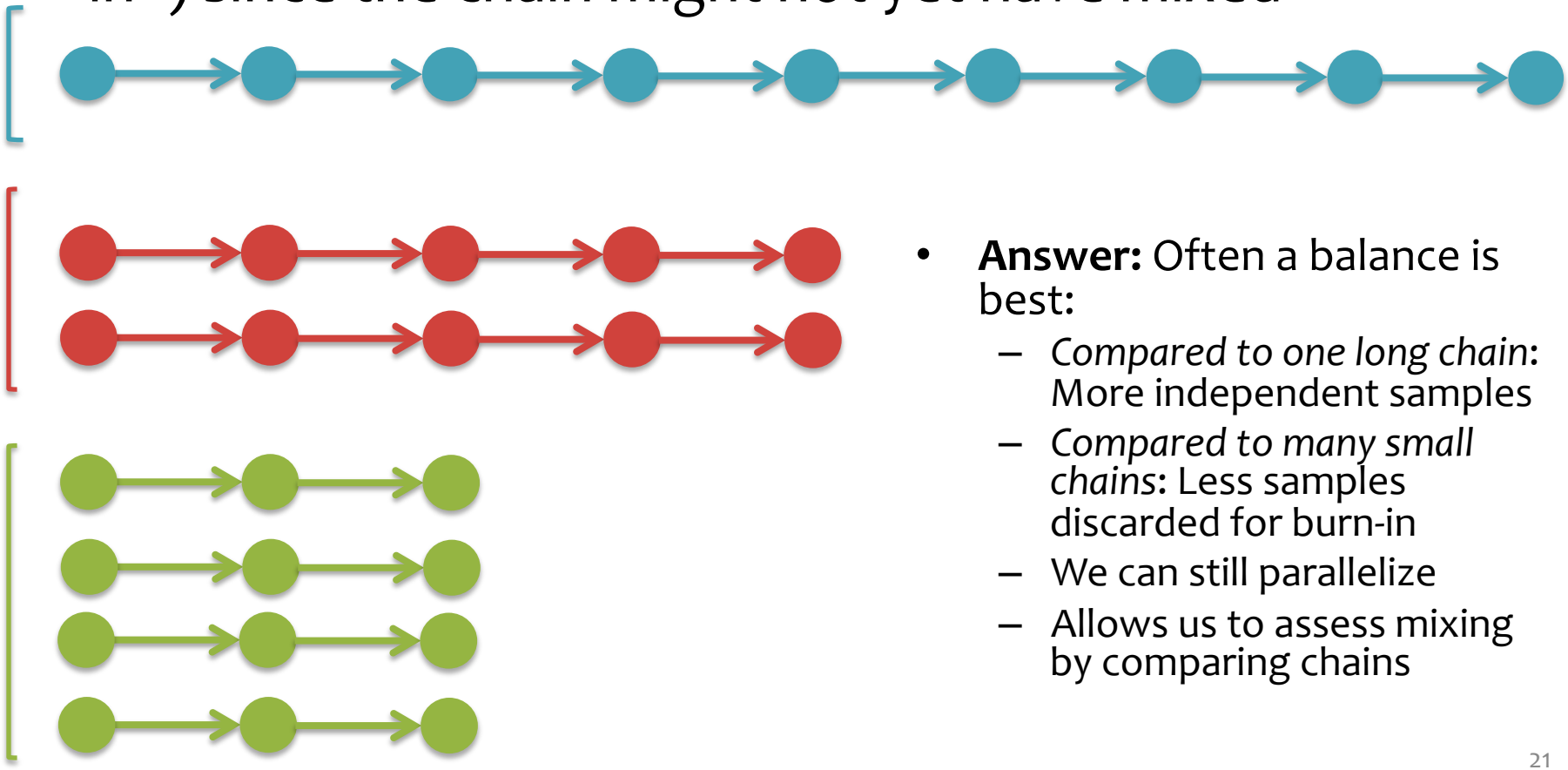
- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
  - Compare statistics of multiple independent chains
  - Ex: Compare log-likelihoods





Practical Issues

- **Question:** Is one long Markov chain better than many short ones?
- **Note:** typical to discard initial samples (aka. "burn-in") since the chain might not yet have mixed



- **Answer:** Often a balance is best:
  - *Compared to one long chain*: More independent samples
  - *Compared to many small chains*: Less samples discarded for burn-in
  - We can still parallelize
  - Allows us to assess mixing by comparing chains

TOPIC MODELING

Topic Modeling

Motivation:

Suppose you're given a massive corpora and asked to carry out the following tasks

- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content

Topic Modeling

Motivation:

Suppose you're given a massive corpora and asked to carry out the following tasks

- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content

Topic Modeling:

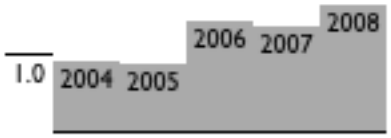
A method of (usually unsupervised) discovery of latent or hidden structure in a corpus

- Applied primarily to text corpora, but **techniques are more general**
- Provides a **modeling toolbox**
- Has prompted the exploration of a variety of new **inference methods** to accommodate **large-scale datasets**

Topic Modeling

Dirichlet-multinomial regression (DMR) topic model on ICML
(Mimno & McCallum, 2008)

Topic 0 [0.152]

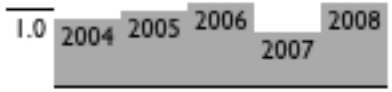


problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions

Topic 54 [0.051]

decision trees, trees, tree, decision tree, decision, tree ensemble, junction tree, decision tree learners, leaf nodes, arithmetic circuits, ensembles modts, skewing, ensembles, anytime induction decision trees, trees trees, random forests, objective decision trees, tree learners, trees grove, candidate split

Topic 99 [0.066]



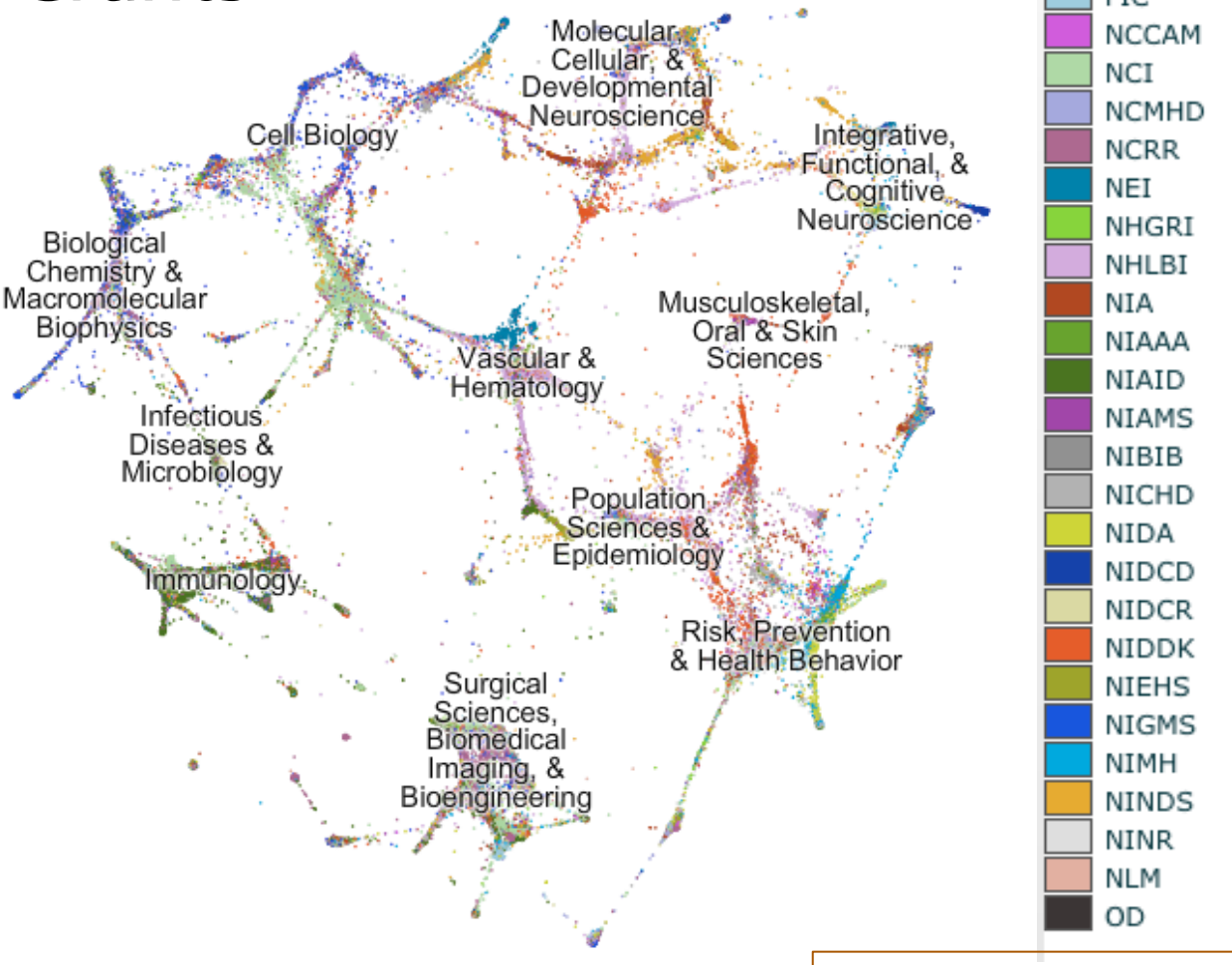
inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient

http:// www.cs.umass.edu/~mimno/icml100.html

Topic Modeling

- Map of NIH Grants

(Talley et al., 2011)

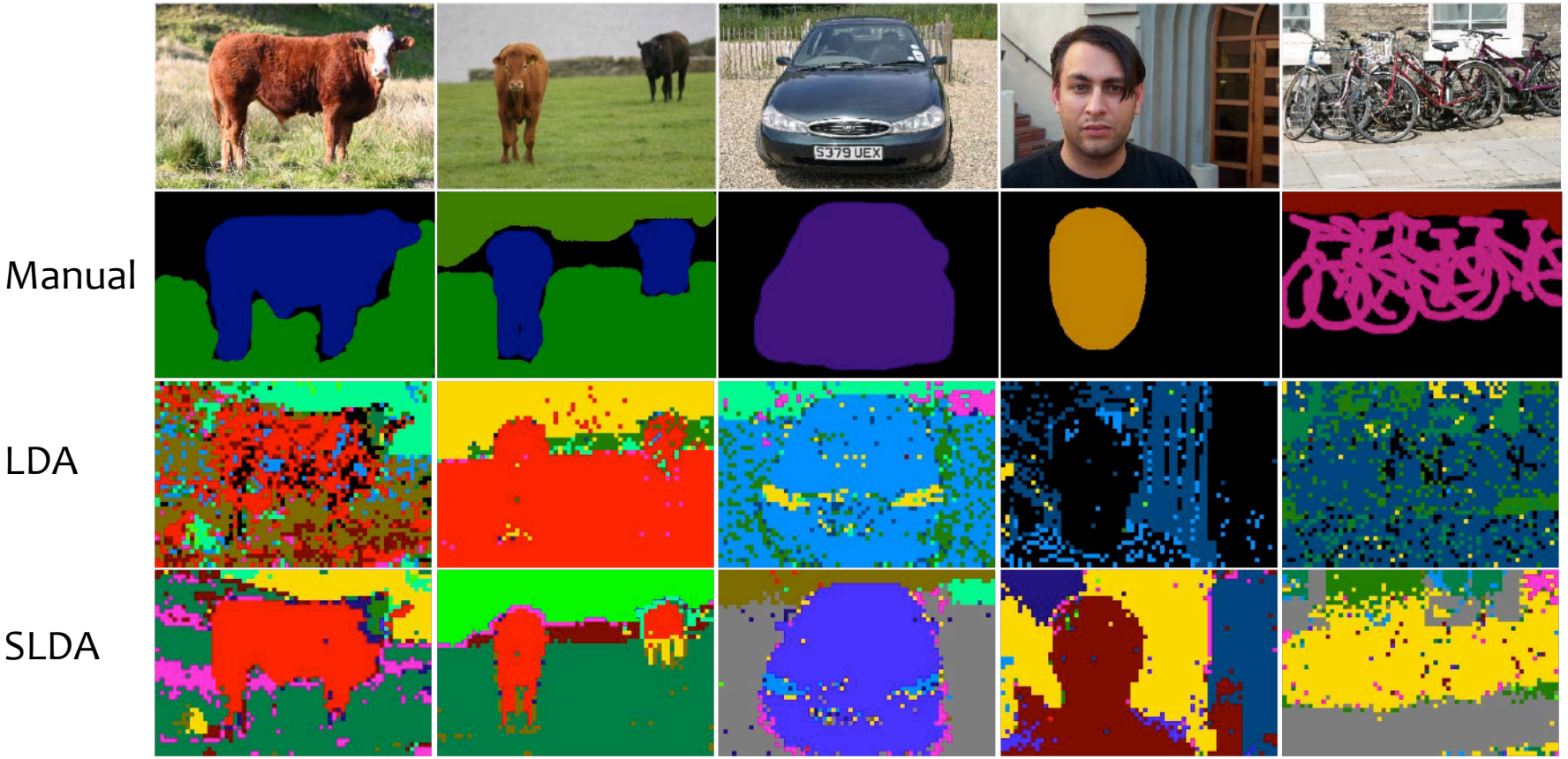


https://app.nihmaps.org/

Other Applications of Topic Models

- Spacial LDA

(Wang & Grimson, 2007)



Other Applications of Topic Models

- Word Sense Induction

(Brody & Lapata, 2009)

Senses of *drug* (WSJ)
1. U.S., administration, federal, against, war, dealer
2. patient, people, problem, doctor, company, abuse
3. company, million, sale, maker, stock, inc.
4. administration, food, company, approval, FDA

Senses of *drug* (BNC)
1. patient, treatment, effect, anti-inflammatory
2. alcohol, treatment, patient, therapy, addiction
3. patient, new, find, effect, choice, study
4. test, alcohol, patient, abuse, people, crime
5. trafficking, trafficker, charge, use, problem
6. abuse, against, problem, treatment, alcohol
7. people, wonder, find, prescription, drink, addict
8. company, dealer, police, enforcement, patient

- Selectional Preference

(Ritter et al., 2010)

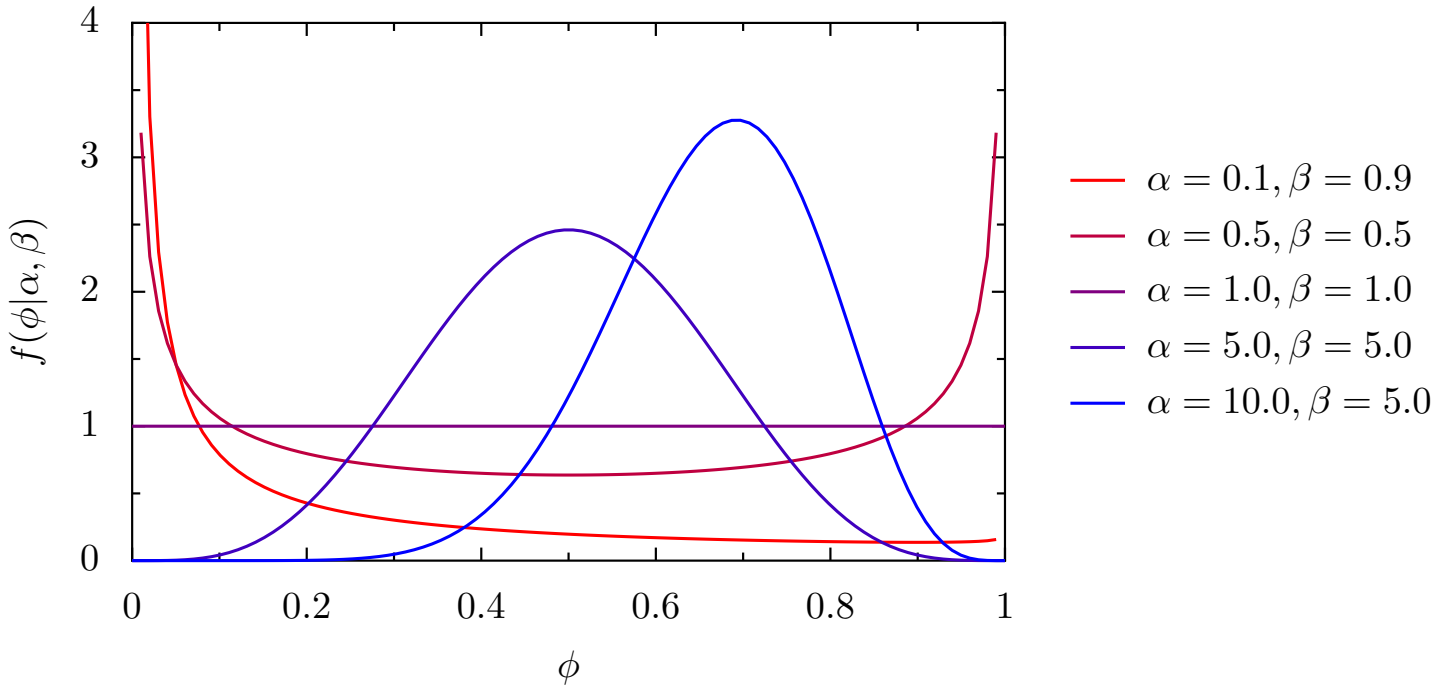
Topic t	Arg1	Relations which assign highest probability to t	Arg2
18	The residue - The mixture - The reaction mixture - The solution - the mixture - the reaction mixture - the residue - The reaction - the solution - The filtrate - the reaction - The product - The crude product - The pellet - The organic layer - Thereto - This solution - The resulting solution - Next - The organic phase - The resulting mixture - C.)	was treated with, is treated with, was poured into, was extracted with, was purified by, was diluted with, was filtered through, is disolved in, is washed with	EtOAc - CH2Cl2 - H2O - CH.sub.2Cl.sub.2 - H.sub.2O - water - MeOH - NaHCO3 - Et2O - NHCl - CHCl.sub.3 - NHCl - dropwise - CH2Cl.sub.2 - Celite - Et.sub.2O - Cl.sub.2 - NaOH - AcOEt - CH2C12 - the mixture - saturated NaHCO3 - SiO2 - H2O - N hydrochloric acid - NHCl - preparative HPLC - to0 C

LATENT DIRICHLET ALLOCATION

Beta-Bernoulli Model

- Beta Distribution

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$



Beta-Bernoulli Model

- Generative Process

$\phi \sim \text{Beta}(\alpha, \beta)$ [*draw distribution over words*]
For each word $n \in \{1, \ldots, N\}$
$x_n \sim \text{Bernoulli}(\phi)$ [*draw word*]

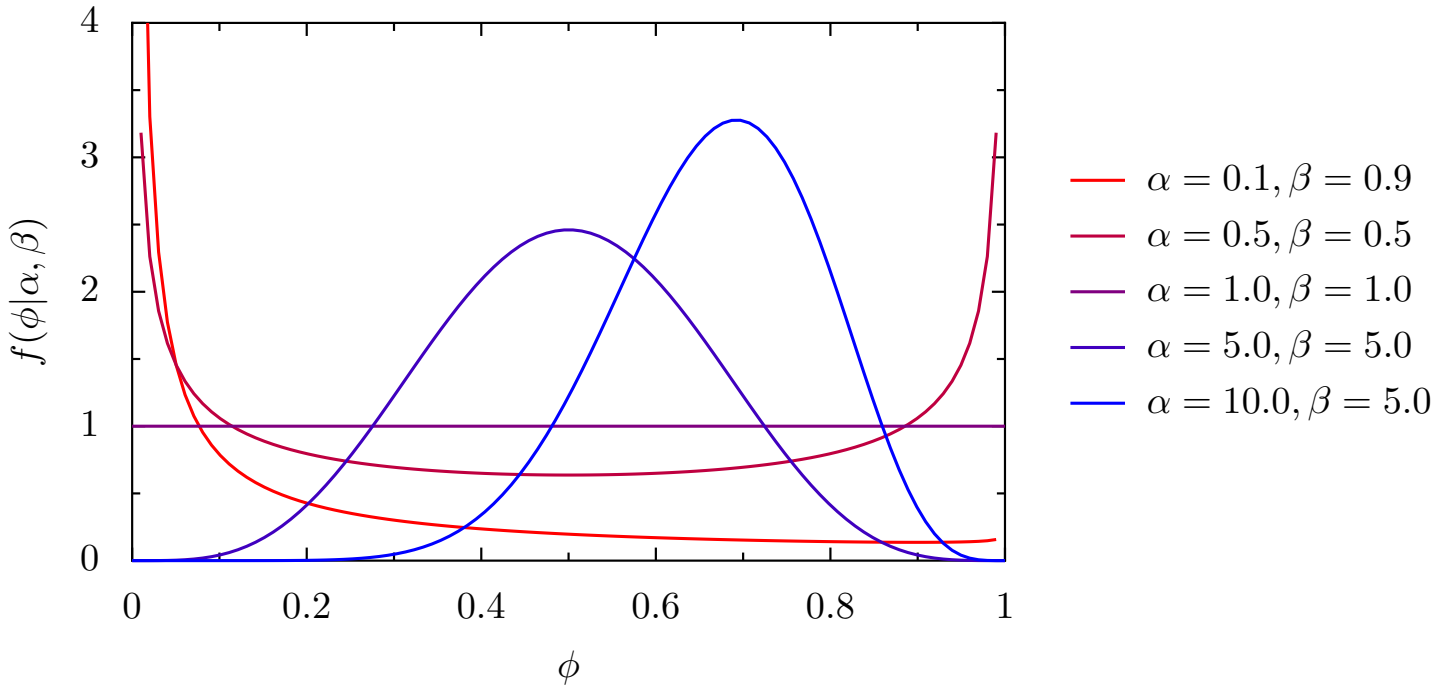
- Example corpus (heads/tails)

H	T	T	H	H	T	T	H	H	H
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

Dirichlet-Multinomial Model

- Dirichlet Distribution

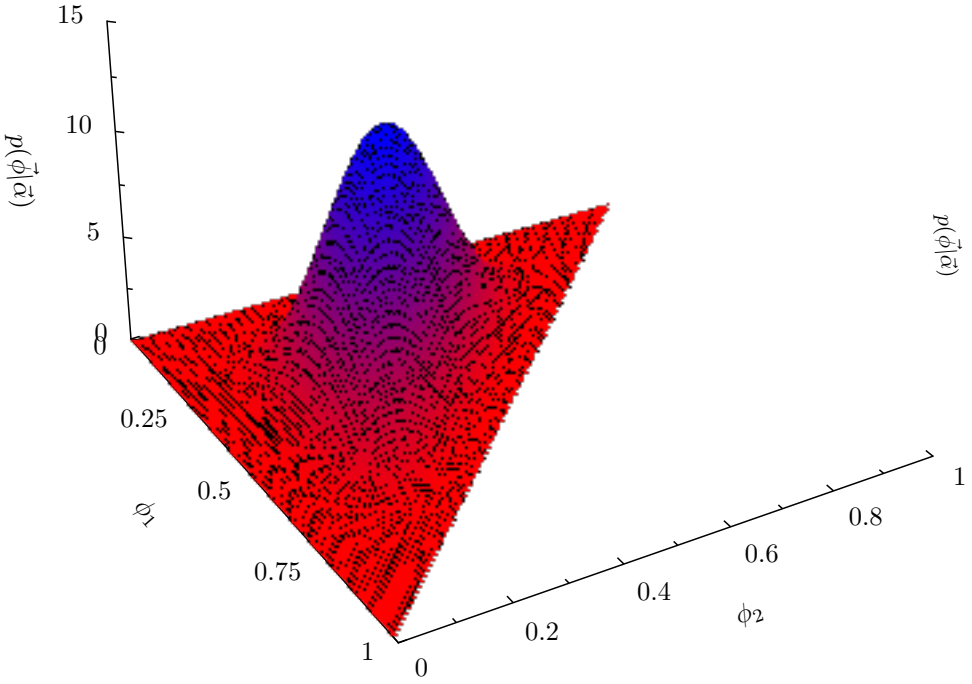
$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

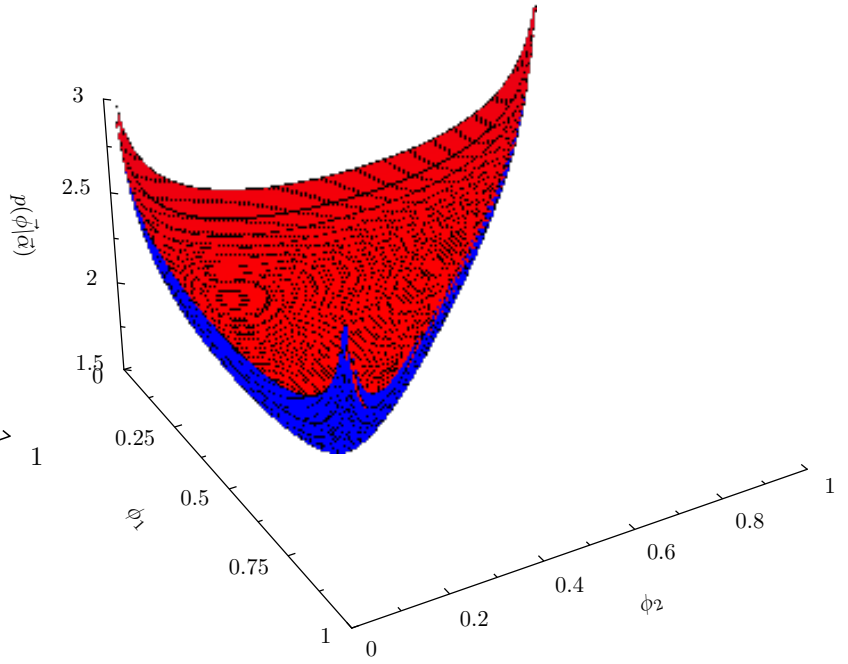


Dirichlet-Multinomial Model

- Dirichlet Distribution

$$p(\vec{\phi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \quad \text{where } B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$





Dirichlet-Multinomial Model

- Generative Process

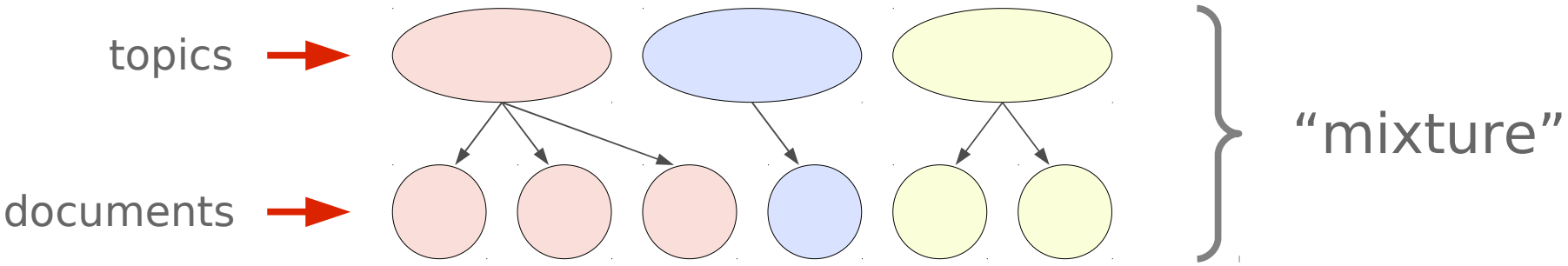
$\phi \sim \text{Dir}(\beta)$ [*draw distribution over words*]
For each word $n \in \{1, \ldots, N\}$
$x_n \sim \text{Mult}(1, \phi)$ [*draw word*]

- Example corpus

the	he	is	the	and	the	she	she	is	is
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

Dirichlet-Multinomial Mixture Model

- Generative Process



- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

Dirichlet-Multinomial Mixture Model

- Generative Process

For each topic $k \in \{1, \ldots, K\}$:
  $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\beta})$ [*draw distribution over words*]
$\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ [*draw distribution over topics*]
For each document $m \in \{1, \ldots, M\}$
  $z_m \sim \text{Mult}(1, \boldsymbol{\theta})$ [*draw topic assignment*]
  For each word $n \in \{1, \ldots, N_m\}$
    $x_{mn} \sim \text{Mult}(1, \boldsymbol{\phi}_{z_m})$ [*draw word*]

- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

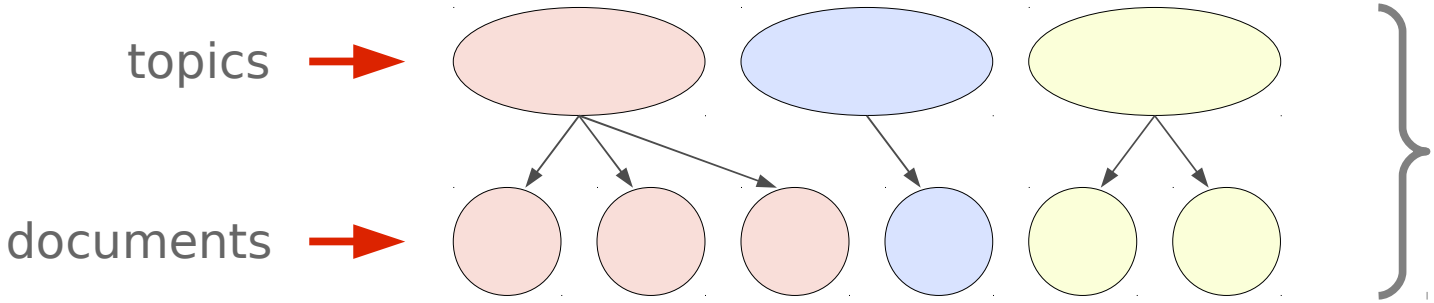
the	and	the
x_{21}	x_{22}	x_{23}

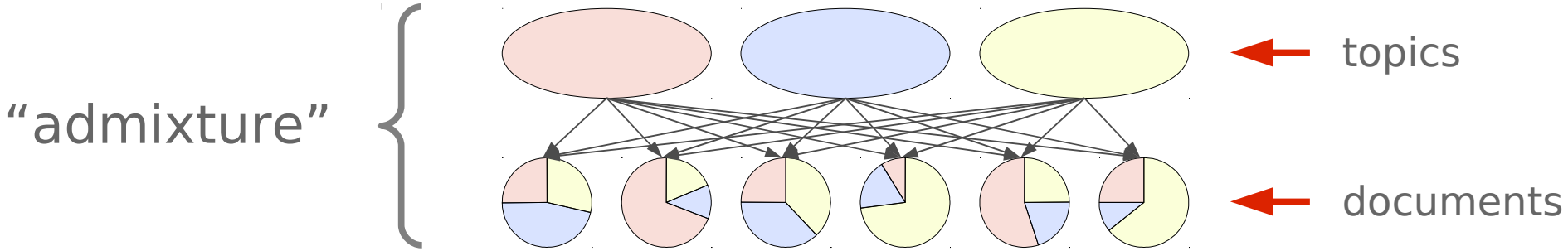
Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

Mixture vs. Admixture (LDA)

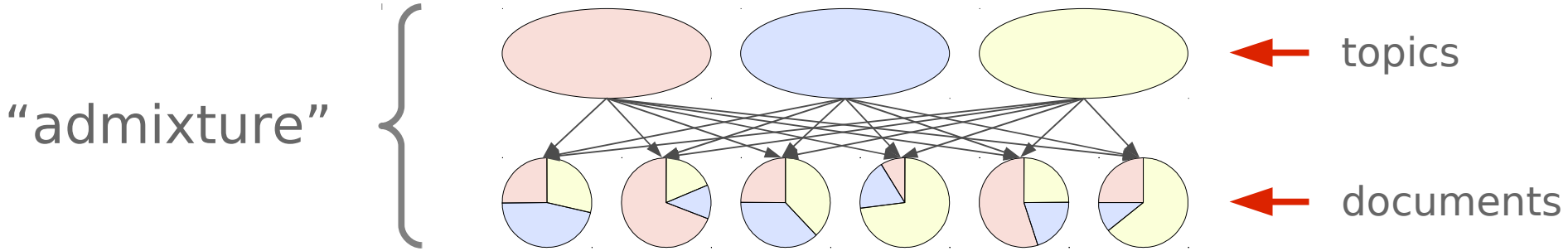




Diagrams from Wallach, JHU 2011, slides

Latent Dirichlet Allocation

- Generative Process



- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

Latent Dirichlet Allocation

- Generative Process

For each topic $k \in \{1, \ldots, K\}$:
  $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\beta})$ [*draw distribution over words*]
For each document $m \in \{1, \ldots, M\}$
  $\boldsymbol{\theta}_m \sim \text{Dir}(\boldsymbol{\alpha})$ [*draw distribution over topics*]
  For each word $n \in \{1, \ldots, N_m\}$
    $z_{mn} \sim \text{Mult}(1, \boldsymbol{\theta}_m)$ [*draw topic assignment*]
    $x_{mn} \sim \boldsymbol{\phi}_{z_{mi}}$ [*draw word*]

- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

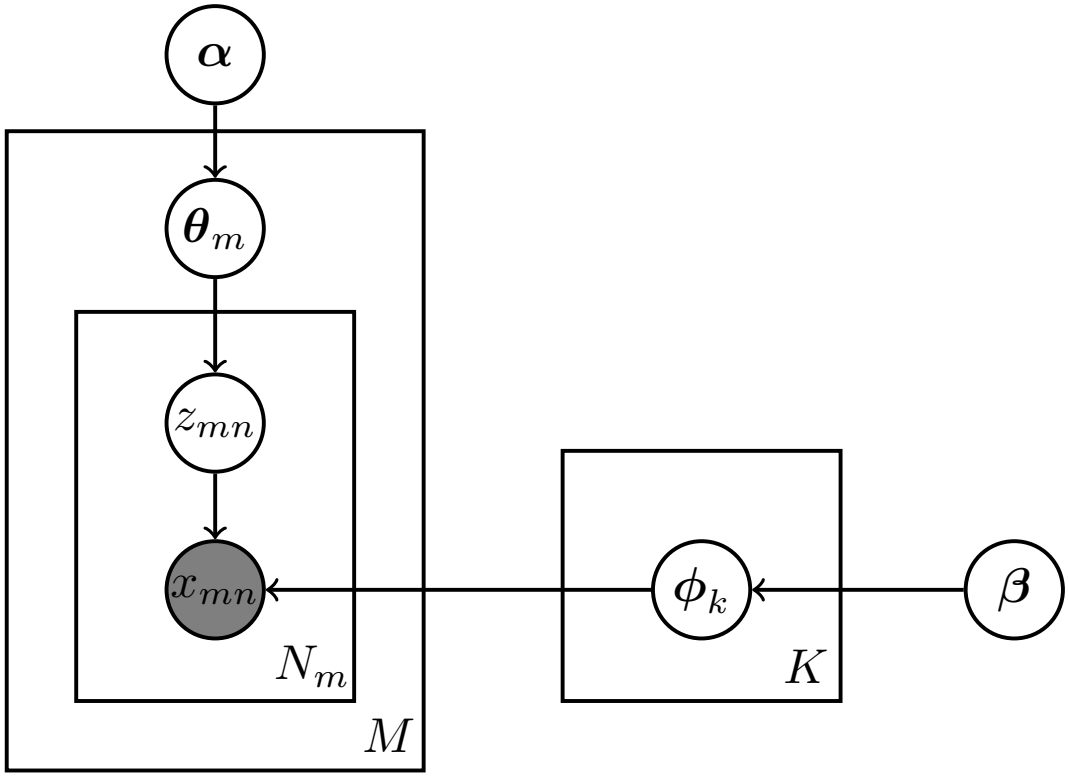
Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

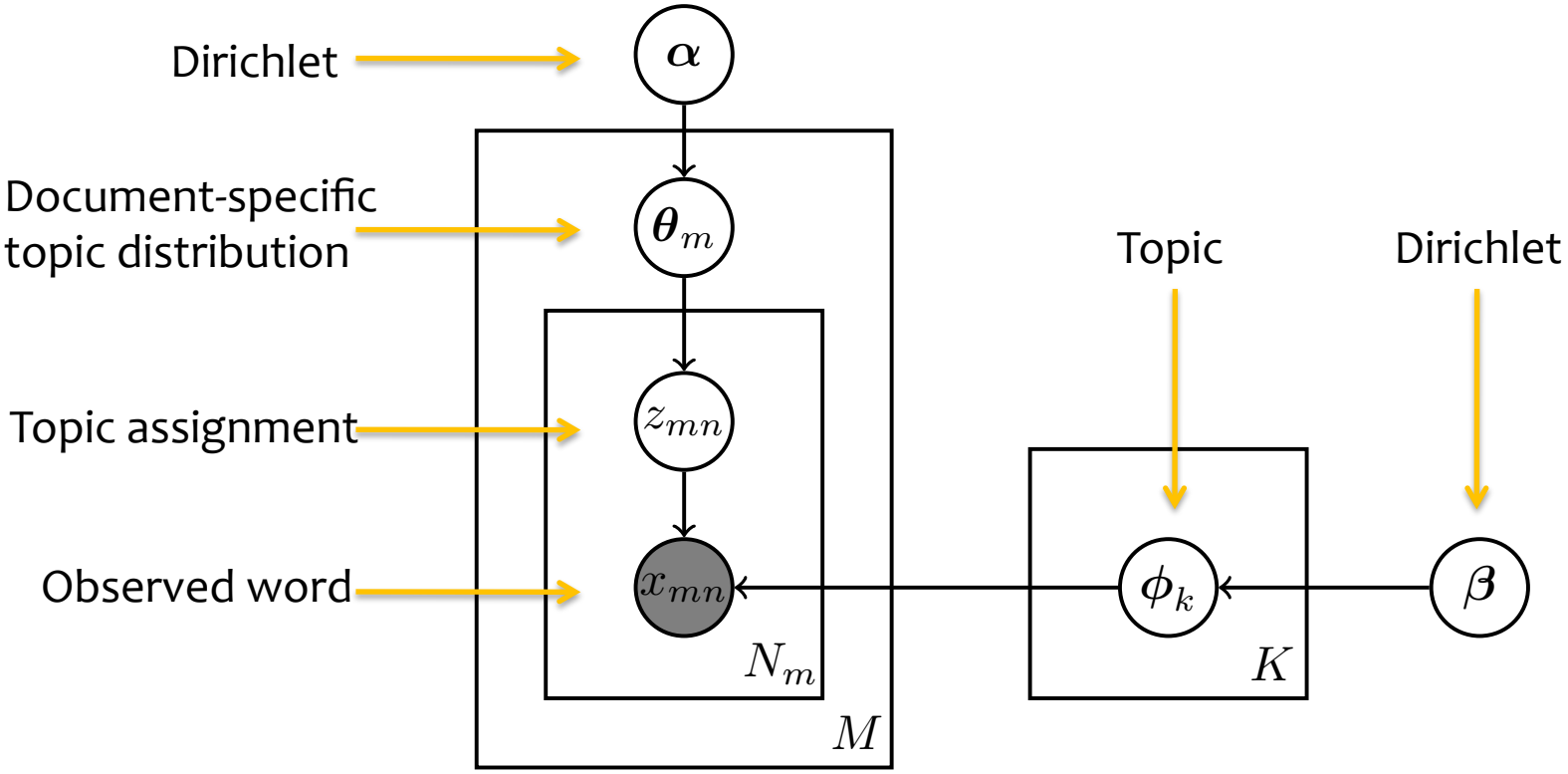
Latent Dirichlet Allocation

- Plate Diagram



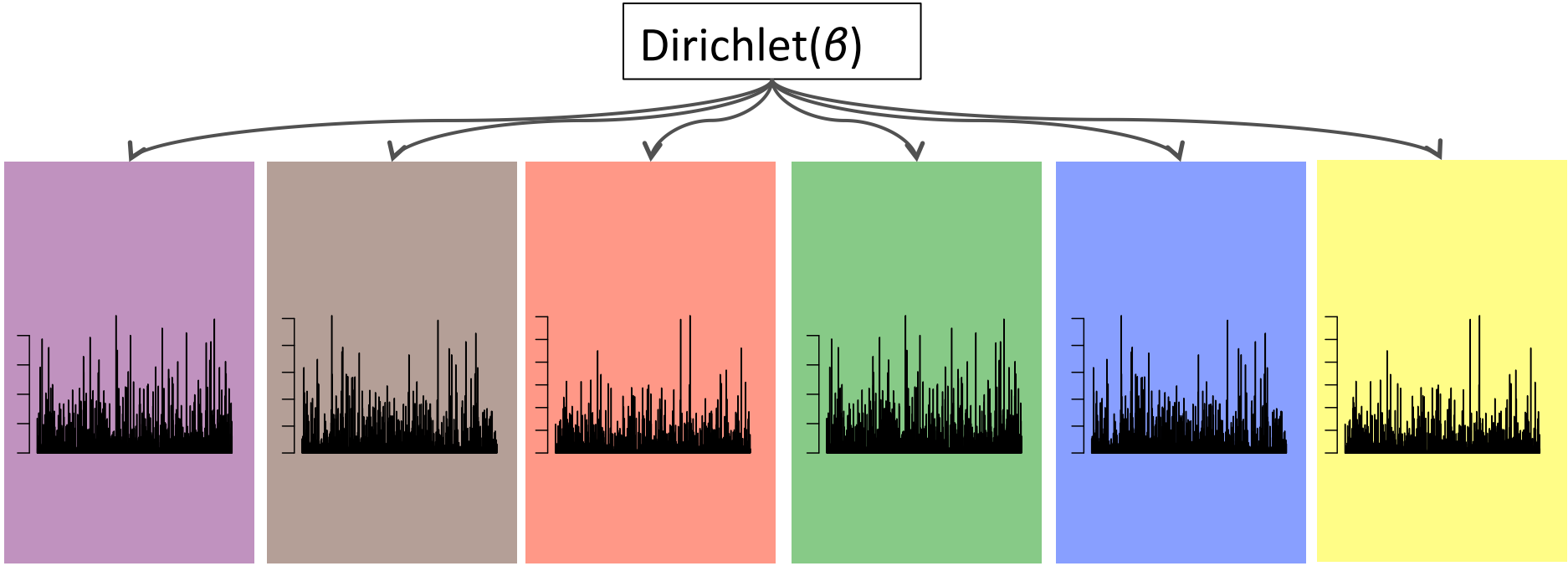
Latent Dirichlet Allocation

- Plate Diagram



(Blei, Ng, & Jordan, 2003)

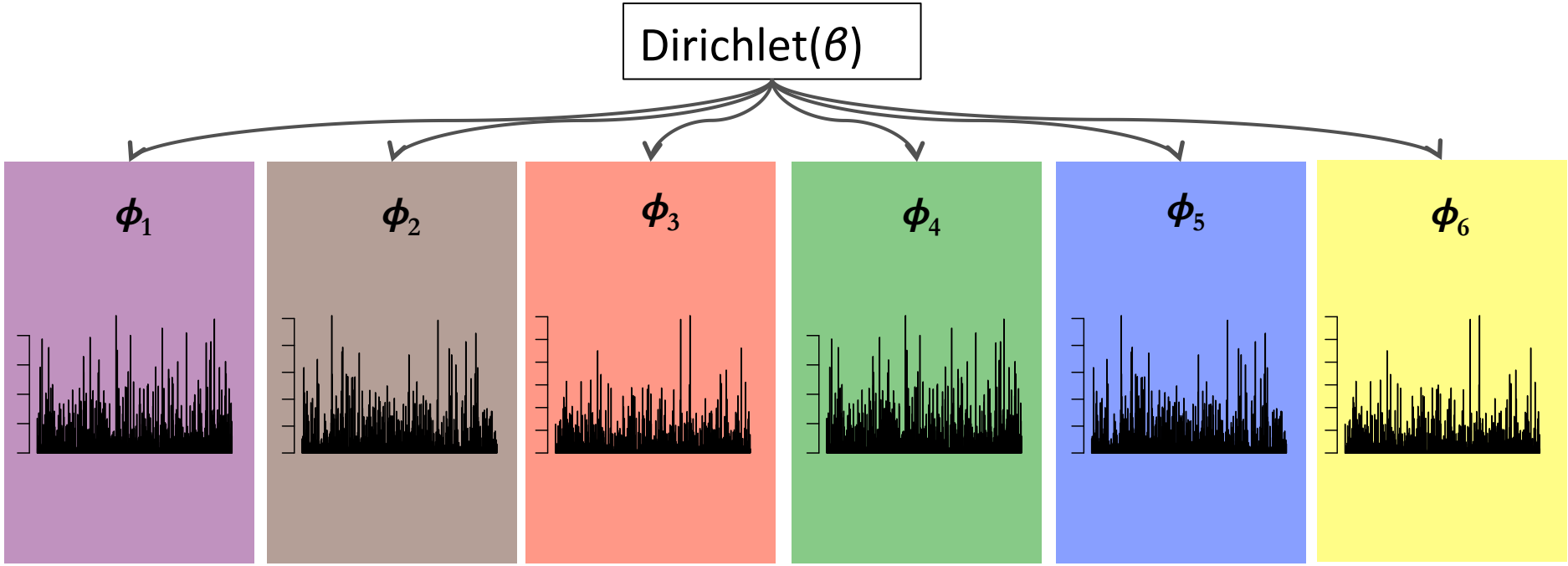
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\phi}_{\mathbf{k}}$

(Blei, Ng, & Jordan, 2003)

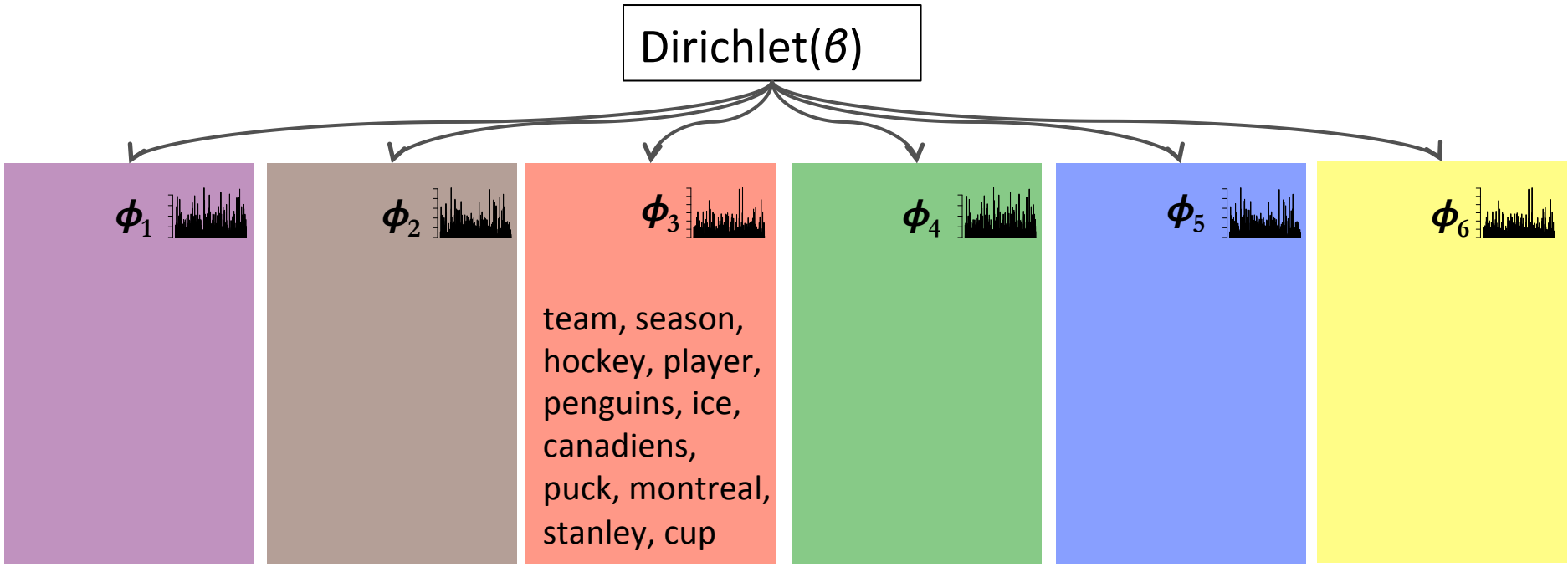
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\phi}_k$

(Blei, Ng, & Jordan, 2003)

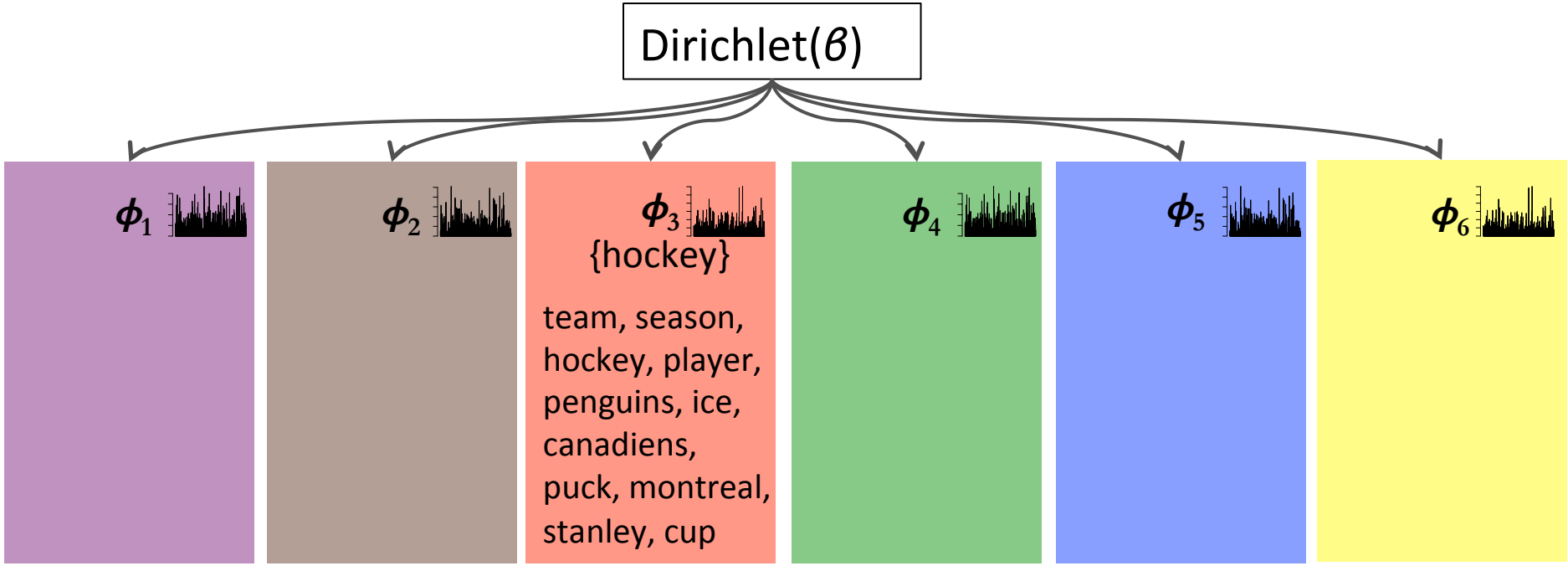
LDA for Topic Modeling



- A topic is visualized as its **high probability words**.

(Blei, Ng, & Jordan, 2003)

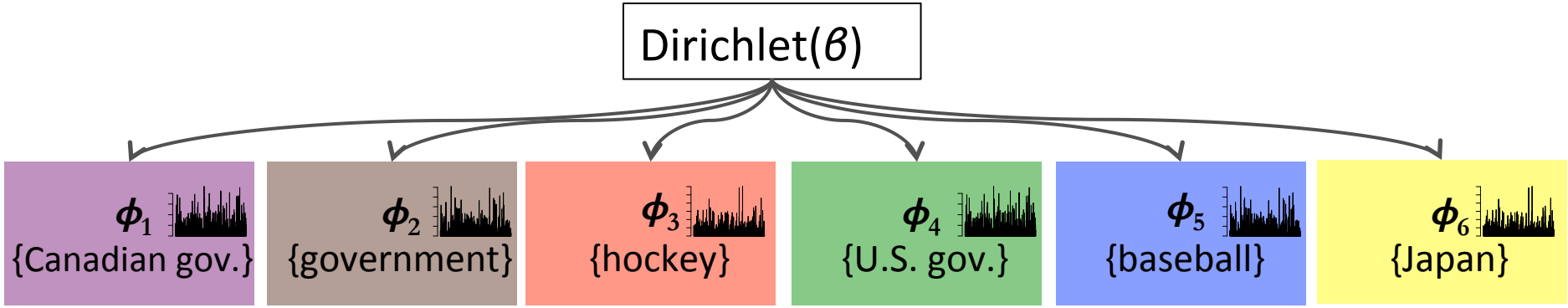
LDA for Topic Modeling



- A topic is visualized as its **high probability words**.
- A pedagogical **label** is used to identify the topic.

(Blei, Ng, & Jordan, 2003)

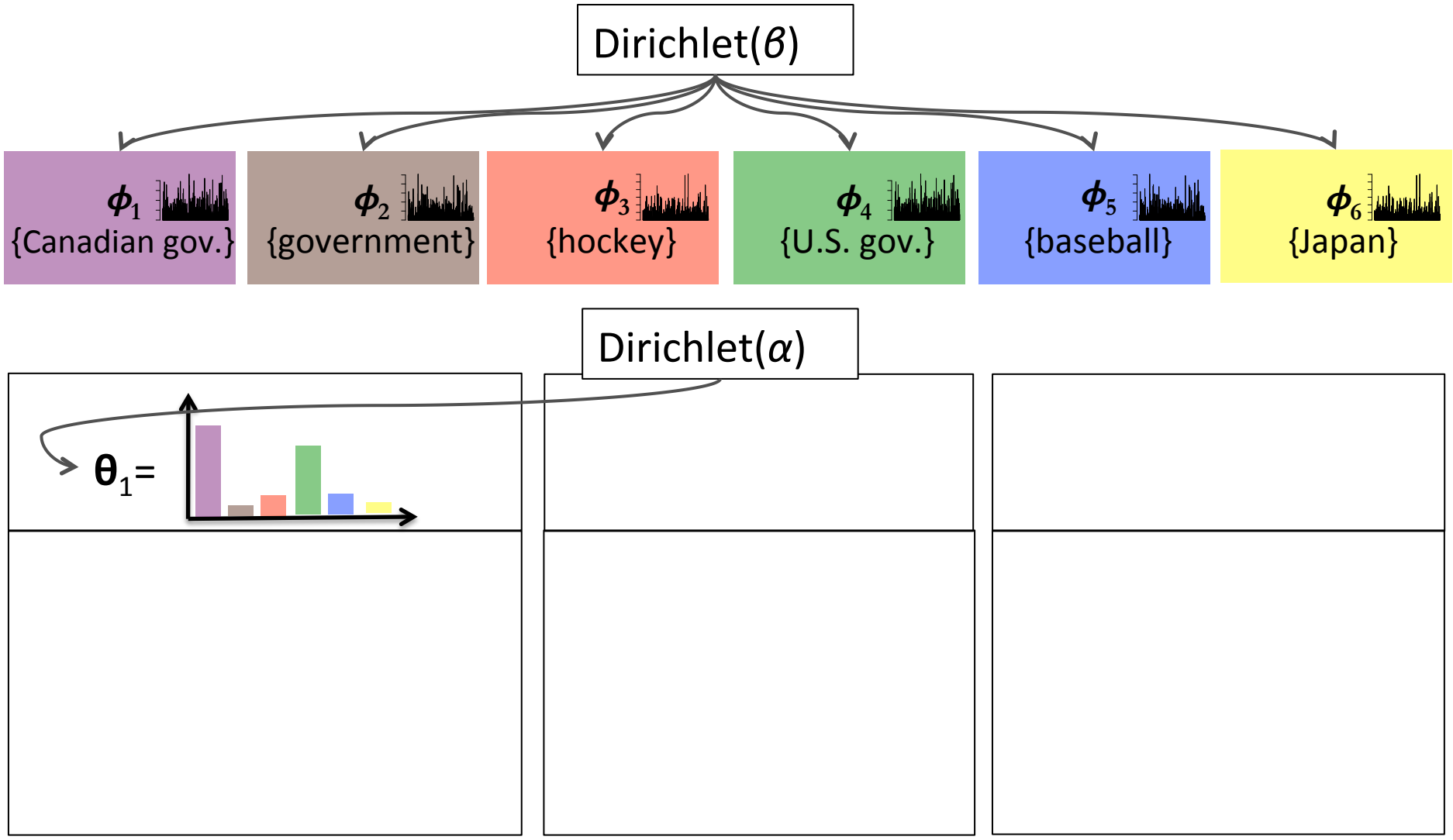
LDA for Topic Modeling



- A topic is visualized as its high probability words.
- A pedagogical **label** is used to identify the topic.

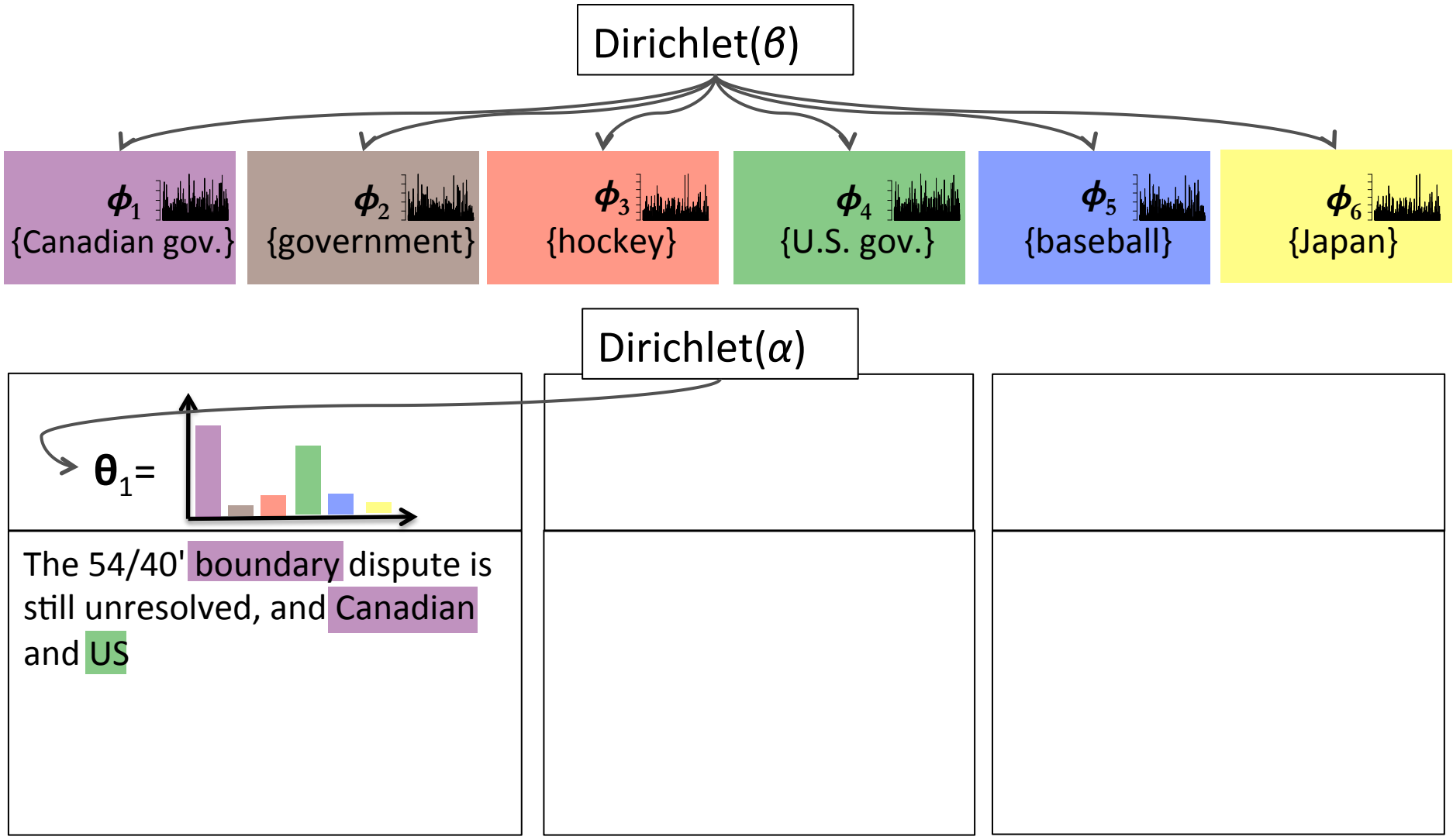
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



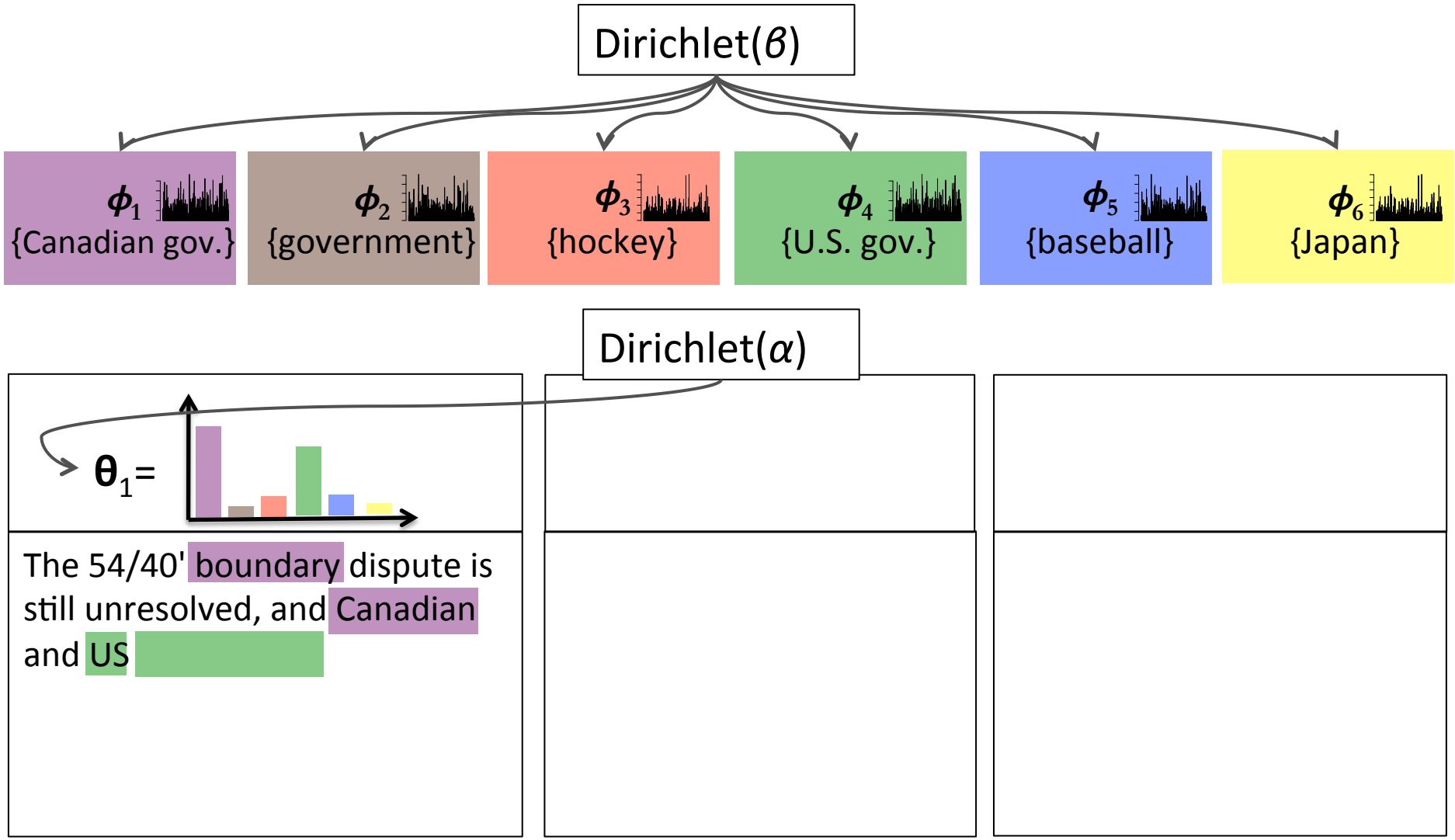
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



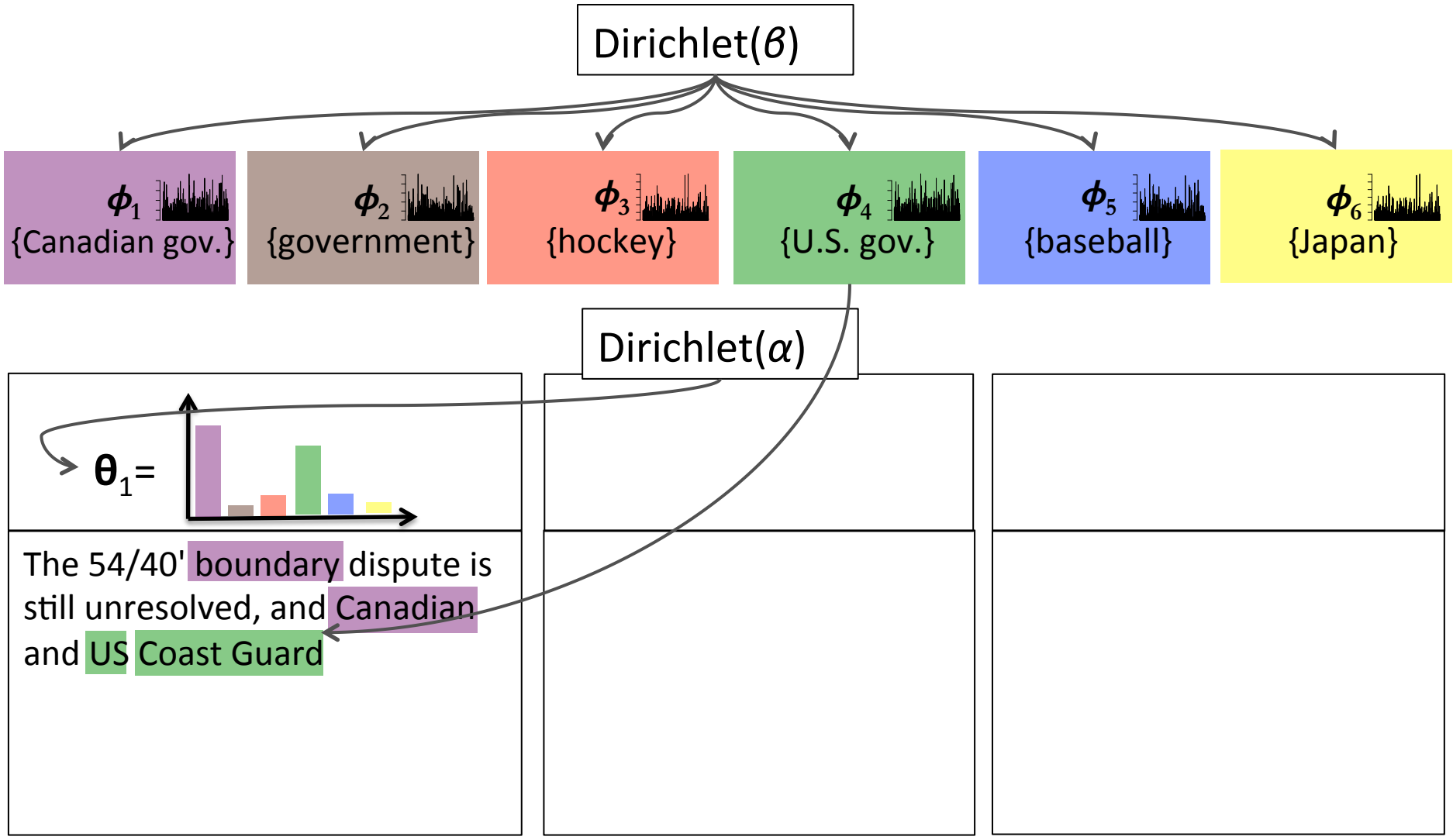
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



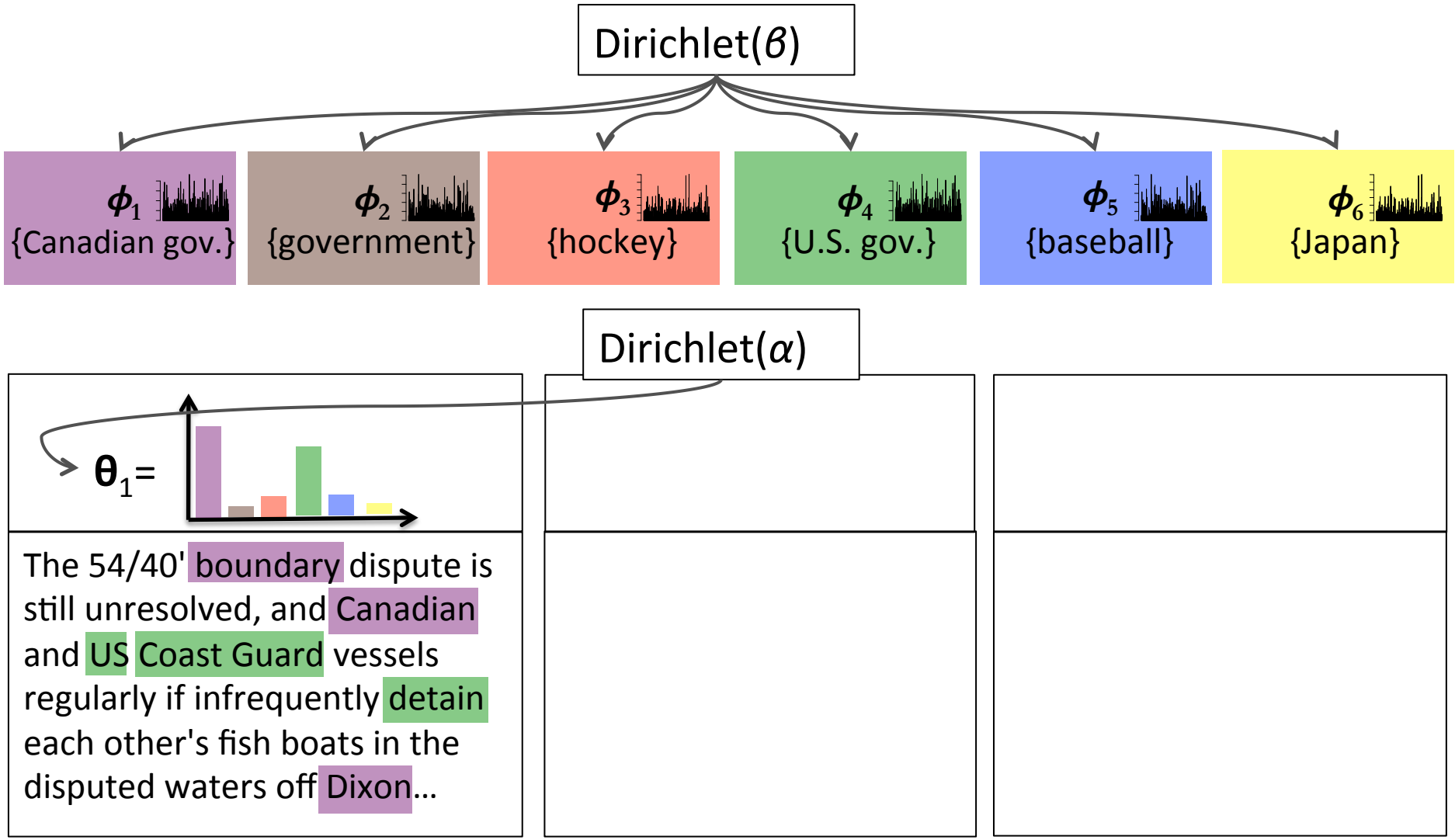
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



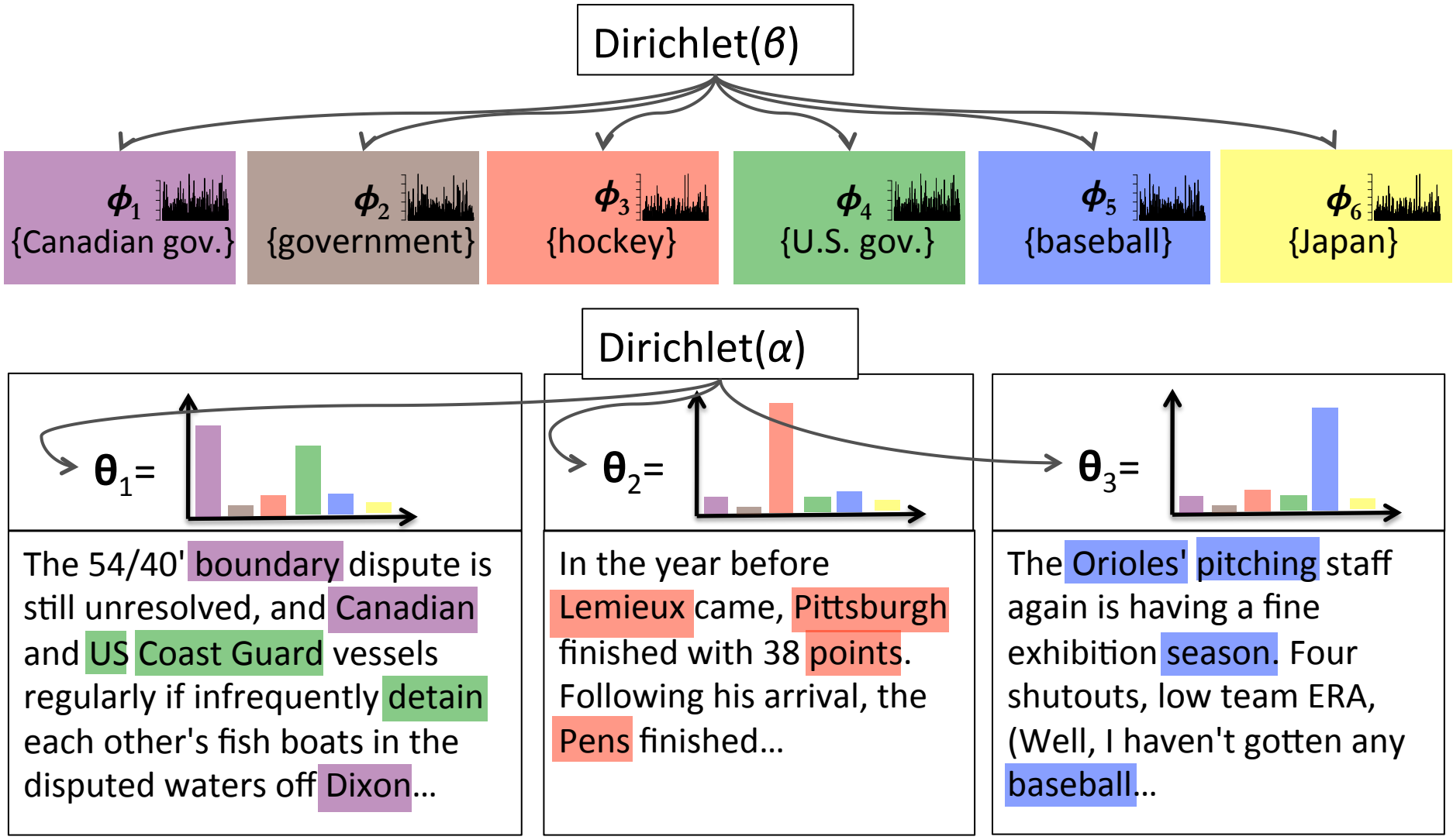
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



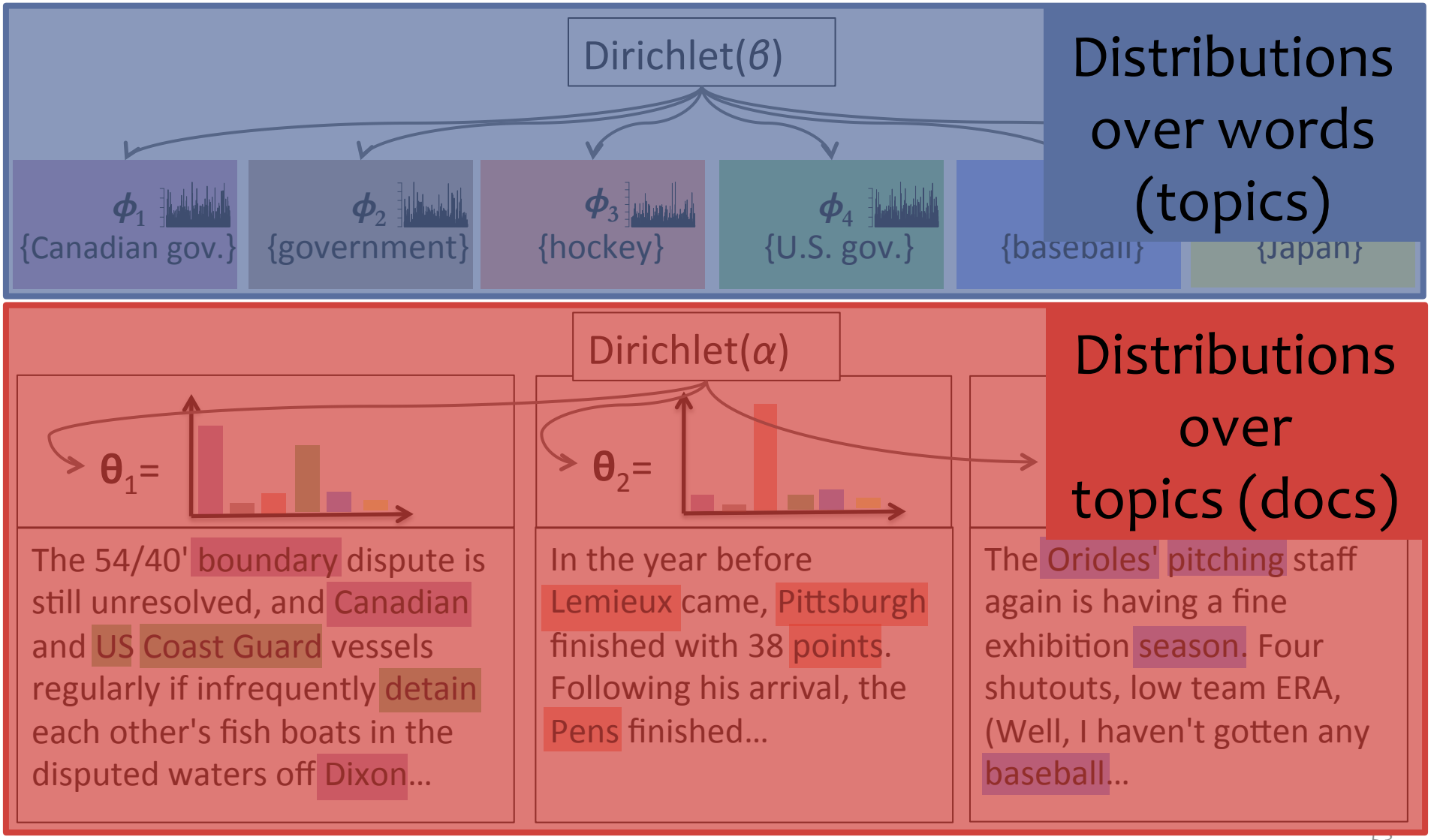
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



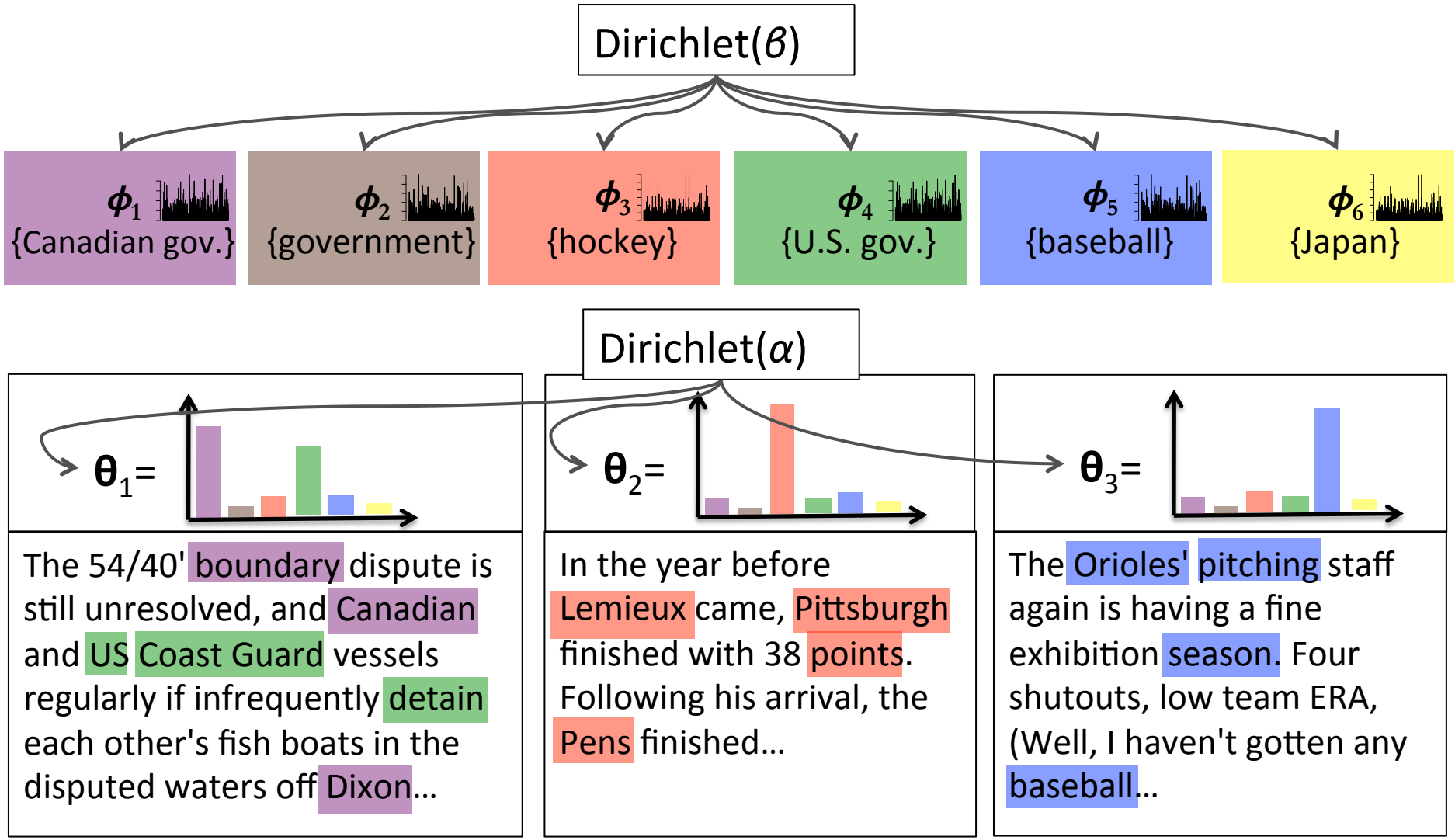
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



(Blei, Ng, & Jordan, 2003)

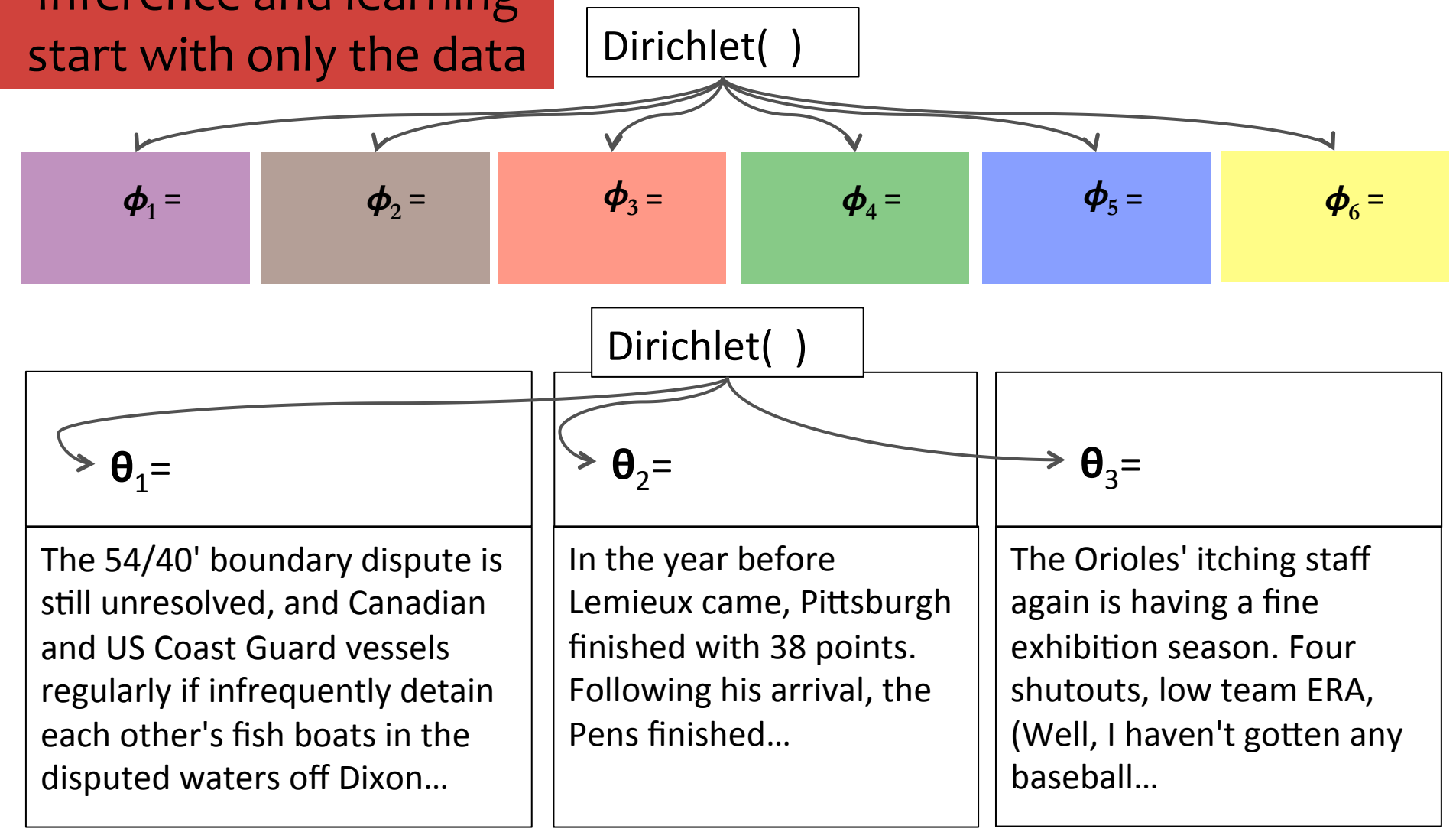
LDA for Topic Modeling



(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling

Inference and learning start with only the data



Latent Dirichlet Allocation

Questions:

- Is this a believable story for the generation of a corpus of documents?

- Why might it work well anyway?

Latent Dirichlet Allocation

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.

Slide from David Blei, MLSS 2012

Latent Dirichlet Allocation

How does this relate to my other favorite model for capturing low-dimensional representations of a corpus?

- 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)

Slide from David Blei, MLSS 2012

LDA INFERENCE / LEARNING

Unsupervised Learning

Three learning paradigms:

1. Maximum likelihood

$$\arg\max_{\theta} p(X|\theta)$$

2. Maximum a posteriori (MAP)

$$\arg\max_{\theta} p(\theta|X) \propto p(X|\theta)p(\theta)$$

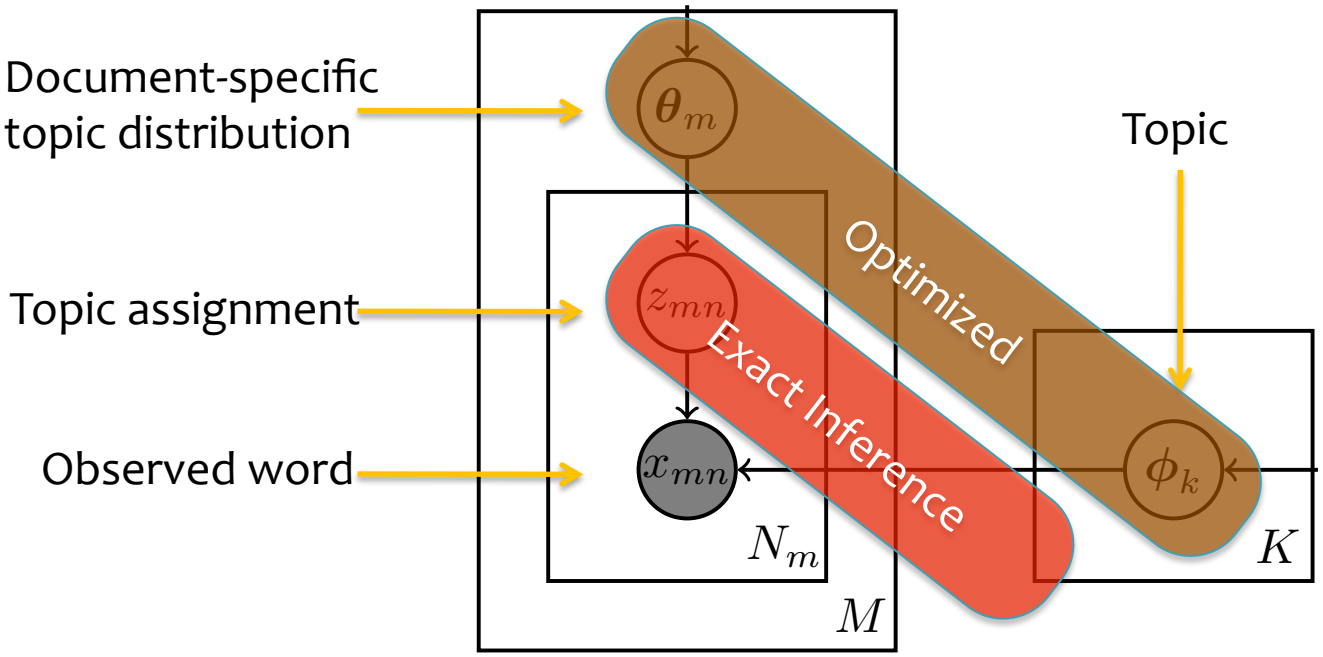
3. Bayesian approach

Estimate the posterior:

$$p(\theta|X) = \ldots$$

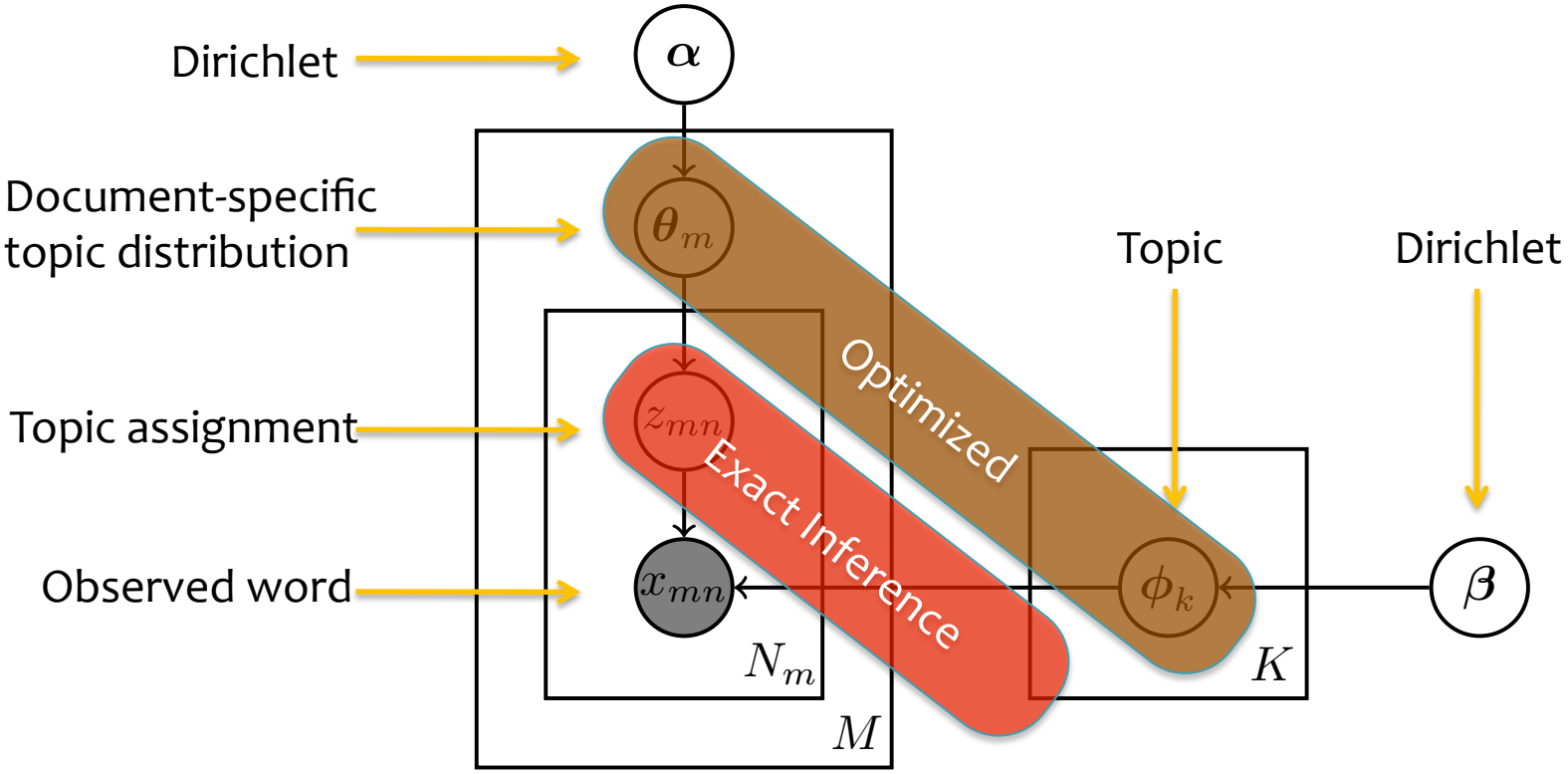
LDA Inference

- Standard EM (Maximum Likelihood)



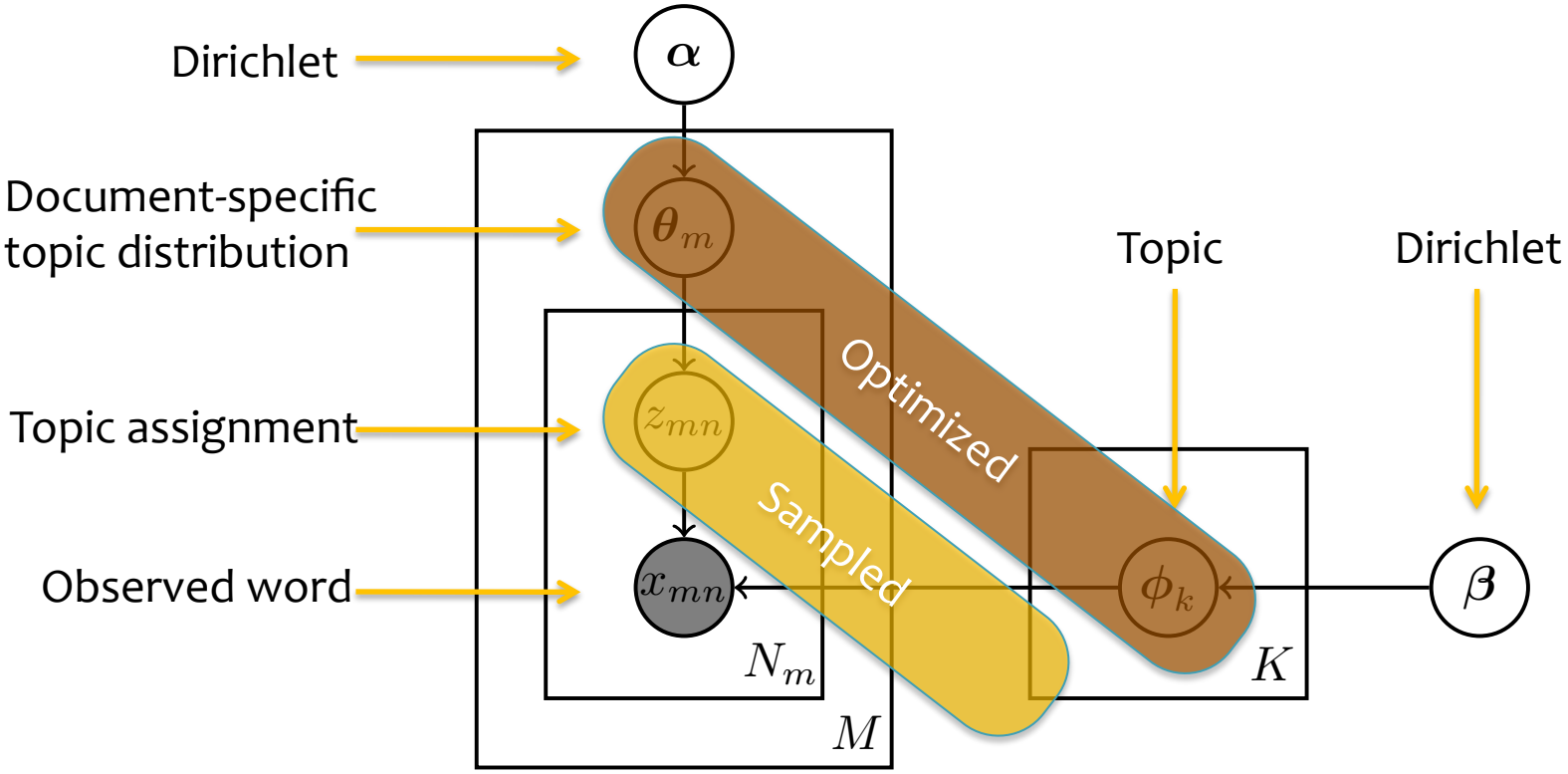
LDA Inference

- Standard EM (MAP)



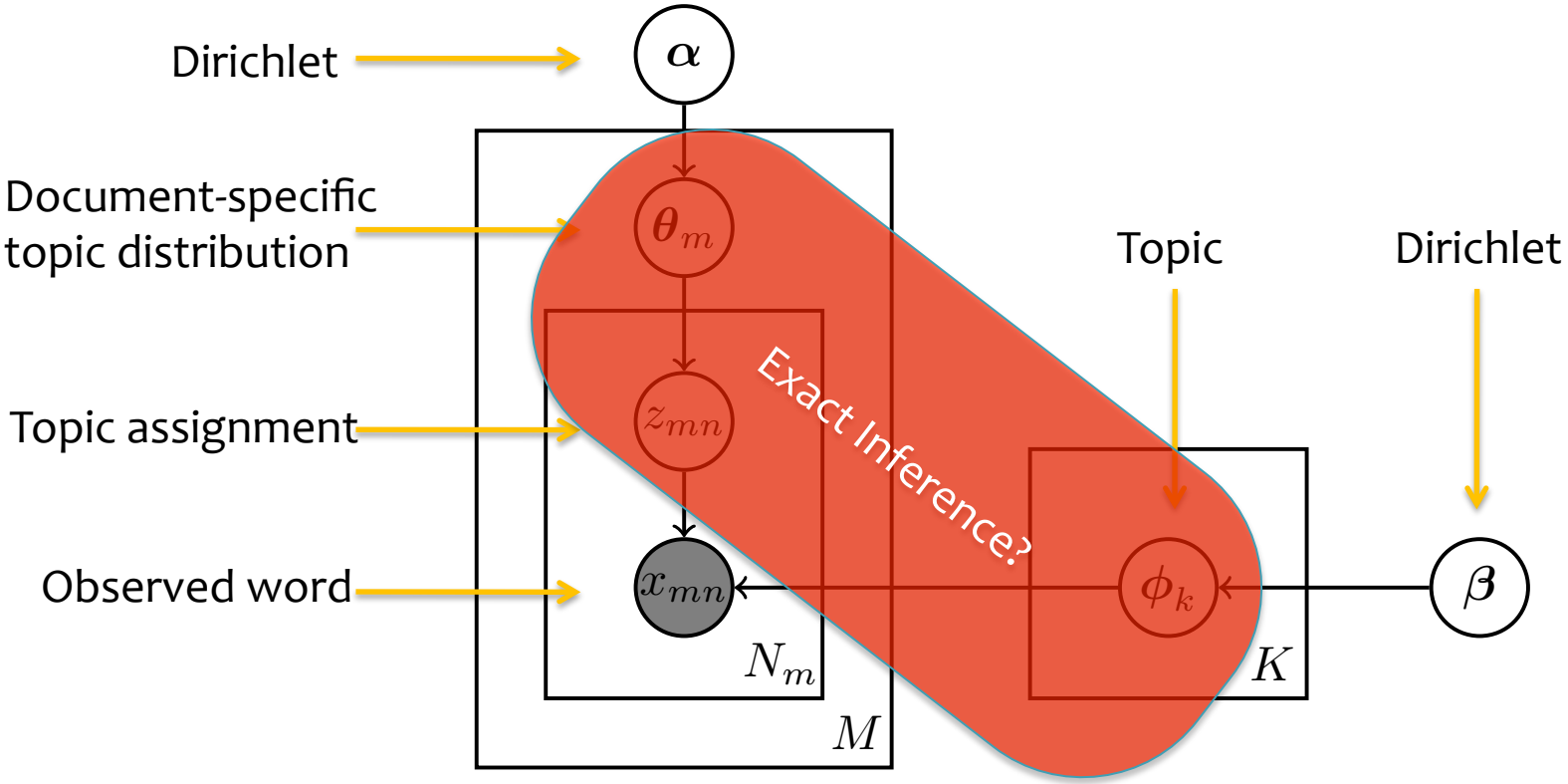
LDA Inference

- Monte Carlo EM



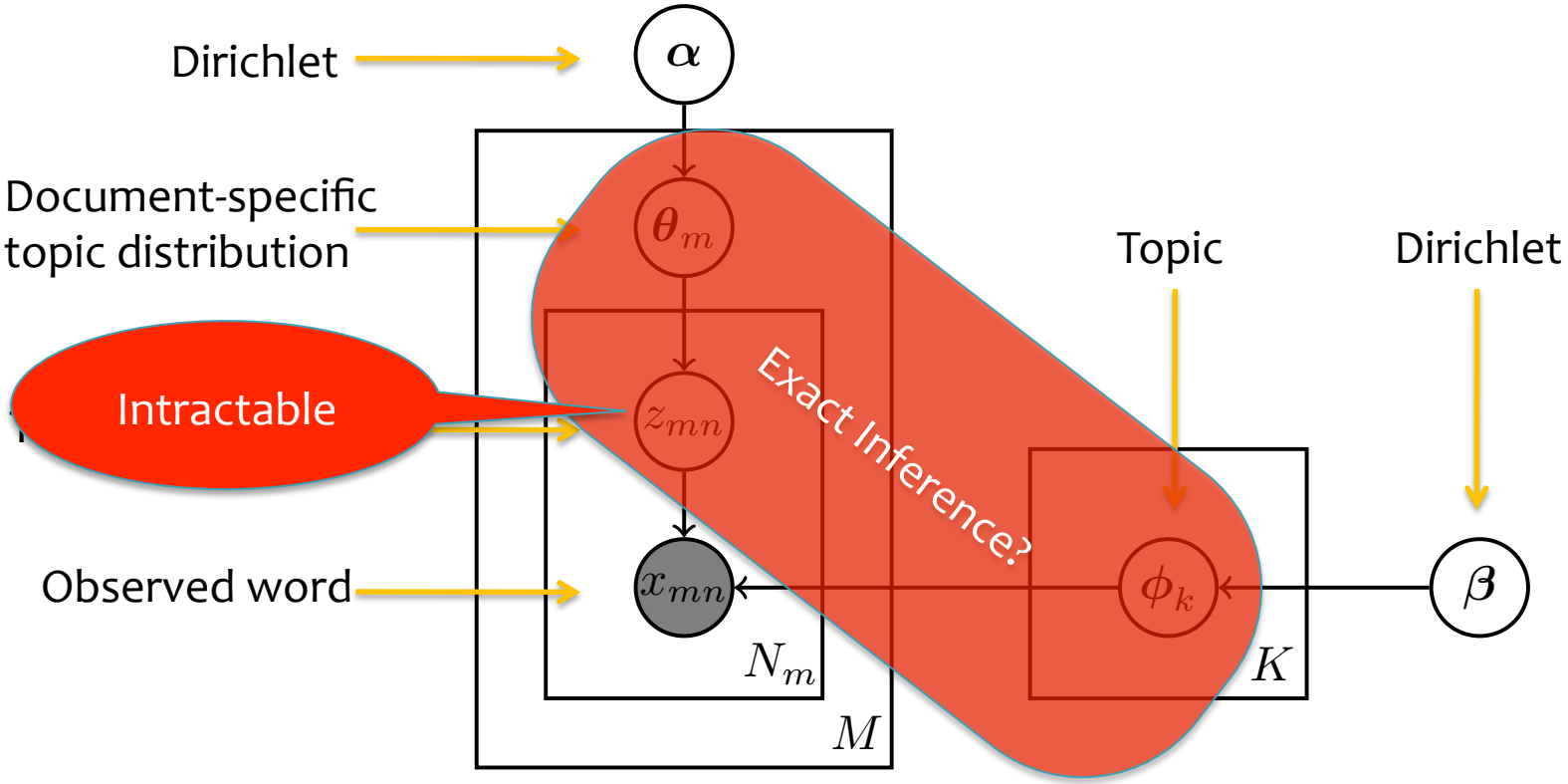
LDA Inference

- Bayesian Approach



LDA Inference

- Bayesian Approach

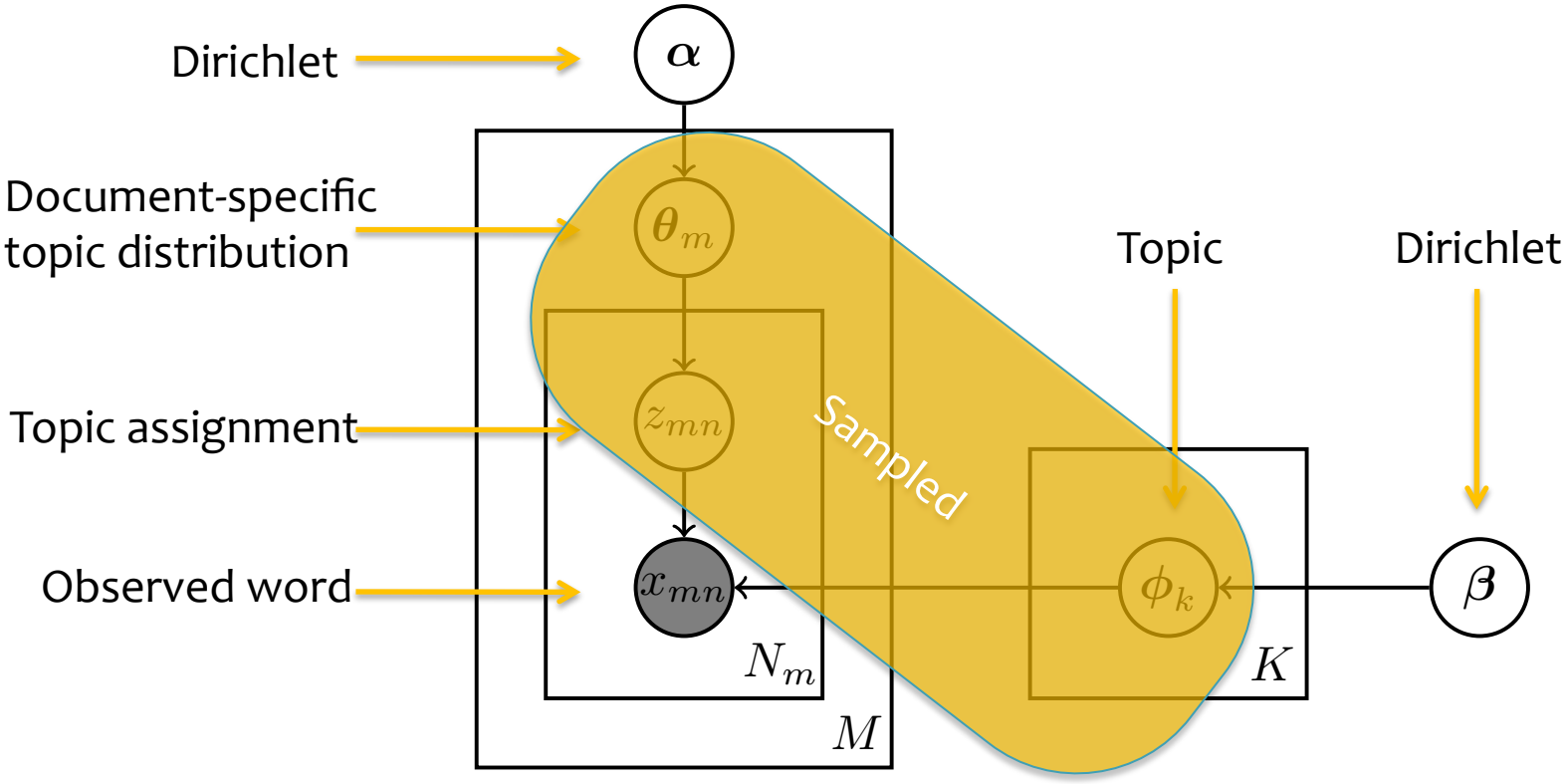


Exact Inference in LDA

- Exactly computing the posterior is intractable in LDA
  - Junction tree algorithm: exact inference in general graphical models
    1. "moralization" converts directed to undirected
    2. "triangulation" breaks 4-cycles by adding edges
    3. Cliques arranged into a junction tree
  - Time complexity is exponential in size of cliques
  - LDA cliques will be large (at least O(# topics)), so complexity is $O(2^{\#\text{ topics}})$
- Exact MAP inference in LDA is NP-hard for a large number of topics (Sontag & Roy, 2011)

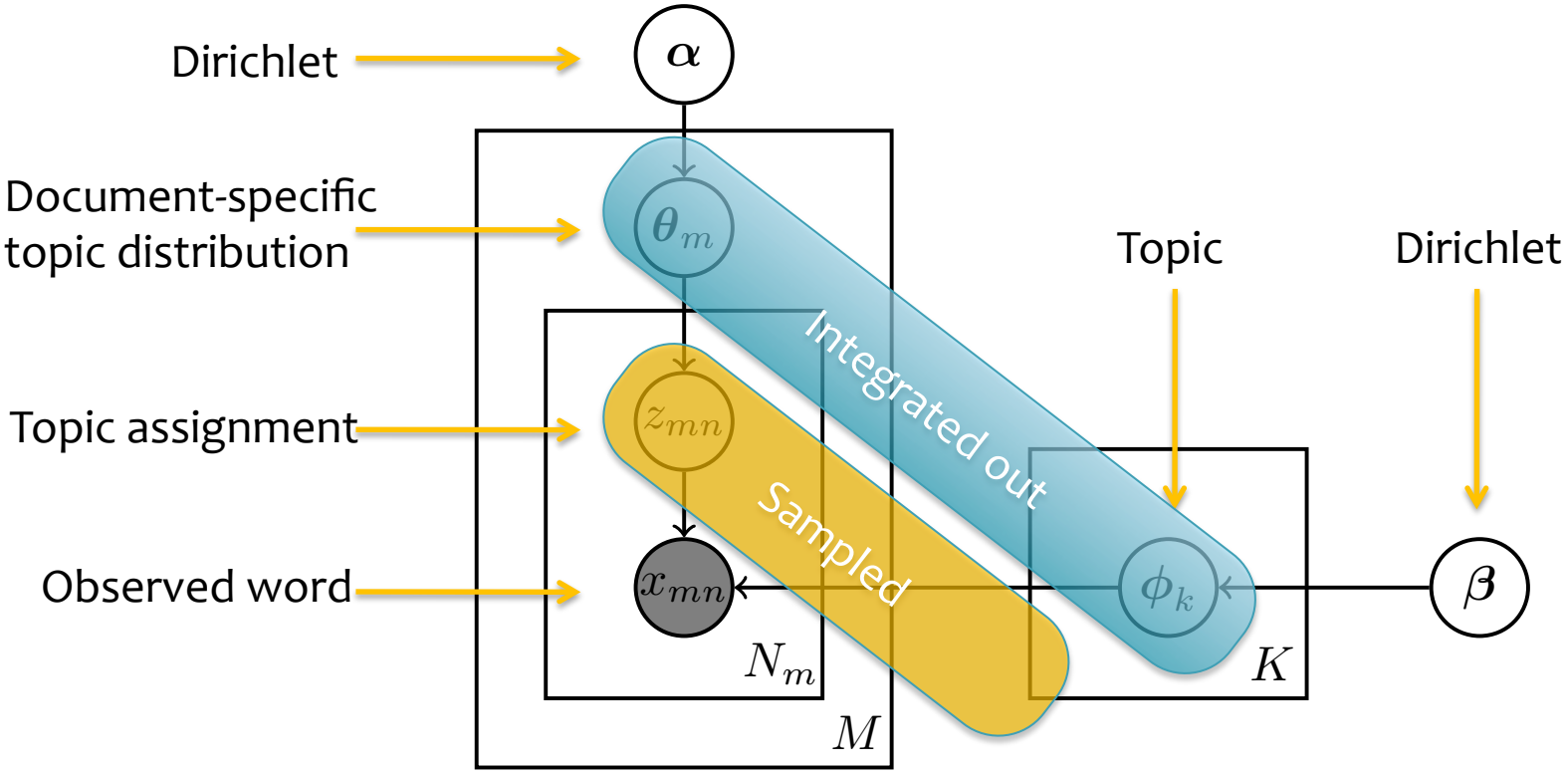
LDA Inference

- Explicit Gibbs Sampler



LDA Inference

- Collapsed Gibbs Sampler



Sampling

Goal:

- Draw samples from the posterior $p(Z|X, \alpha, \beta)$
- Integrate out topics ϕ and document-specific distribution over topics θ

Algorithm:

- While not done…
  - For each document, m:
    - For each word, n:
      - » Resample a single topic assignment using the full conditionals for z_{mn}

Sampling

- What can we do with samples of z_{mn}?
  - Mean of z_{mn}
  - Mode of z_{mn}
  - Estimate posterior over z_{mn}
  - Estimate of topics ϕ and document-specific distribution over topics θ

Gibbs Sampling for LDA

- Full conditionals

$$p(z_i = k | Z^{-i}, X, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{n_{kt}^{-i} + \beta_t}{\sum_{v=1}^{T} n_{kv}^{-i} + \beta_v} \cdot \frac{n_{mk}^{-i} + \alpha_k}{\sum_{j=1}^{K} n_{mj}^{-i} + \alpha_j}$$

where t, m are given by i

n_{kt} = # times topic *k* appears with type *t*

n_{mk} = # times topic *k* appears in document m

Gibbs Sampling for LDA

- Sketch of the derivation of the full conditionals

$$
\begin{aligned}
p(z_i = k | Z^{-i}, X, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{p(X, Z | \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(X, Z^{-i} | \boldsymbol{\alpha}, \boldsymbol{\beta})} \\
&\propto p(X, Z | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= p(X | Z, \boldsymbol{\beta}) p(Z | \boldsymbol{\alpha}) \\
&= \int_{\Phi} p(X | Z, \Phi) p(\Phi | \boldsymbol{\beta}) \, d\Phi \int_{\Theta} p(Z | \Theta) p(\Theta | \boldsymbol{\alpha}) \, d\Theta \\
&= \left(\prod_{k=1}^{K} \frac{B(\vec{n}_k + \boldsymbol{\beta})}{B(\boldsymbol{\beta})} \right) \left(\prod_{m=1}^{M} \frac{B(\vec{n}_m + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \right) \\
&= \frac{n_{kt}^{-i} + \beta_t}{\sum_{v=1}^{T} n_{kv}^{-i} + \beta_v} \cdot \frac{n_{mk}^{-i} + \alpha_k}{\sum_{j=1}^{K} n_{mj}^{-i} + \alpha_j}
\end{aligned}
$$

where t, m are given by i

Dirichlet-Multinomial Model

- The Dirichlet is conjugate to the Multinomial

$\boldsymbol{\phi} \sim \text{Dir}(\boldsymbol{\beta})$	[*draw distribution over words*]
For each word $n \in \{1, \ldots, N\}$	
$x_n \sim \text{Mult}(1, \boldsymbol{\phi})$	[*draw word*]

- The posterior of ϕ is $p(\phi|X) = \frac{p(X|\phi)p(\phi)}{P(X)}$
- Define the count vector $\boldsymbol{n}$ such that n_t denotes the number of times word t appeared
- Then the posterior is also a Dirichlet distribution: $p(\phi|X) \sim \text{Dir}(\boldsymbol{\beta} + \boldsymbol{n})$

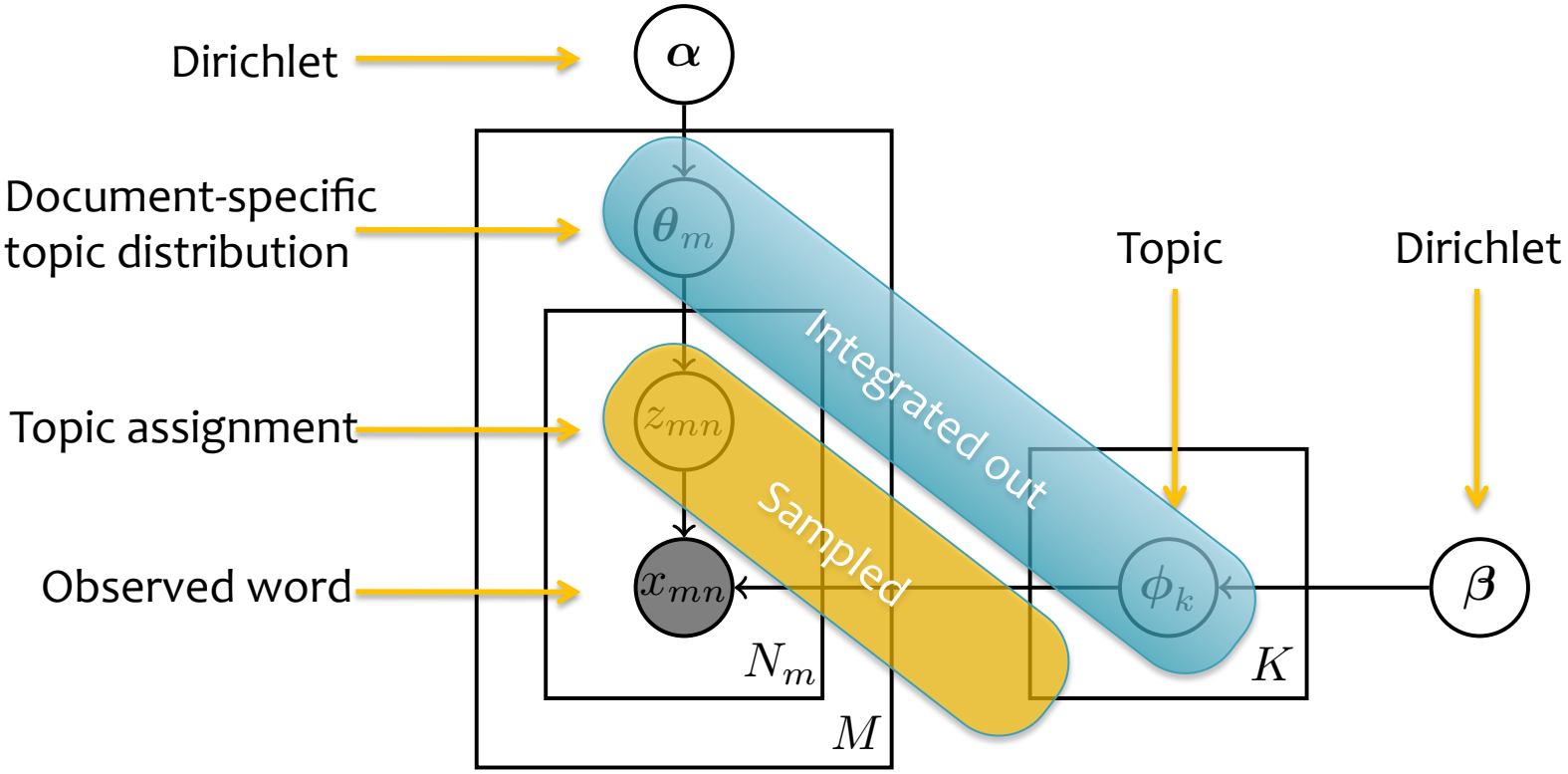
Dirichlet-Multinomial Model

- Why conjugacy is so useful

$$
\begin{aligned}
p(X|\boldsymbol{\alpha}) &= \int_{\phi} p(X|\vec{\phi})p(\vec{\phi}|\boldsymbol{\alpha})\; d\phi \\
&= \int_{\phi} \left(\prod_{v=1}^{V} \phi_v^{n_v} \right) \left(\frac{1}{B(\boldsymbol{\alpha})} \prod_{v=1}^{V} \phi_v^{\alpha_v - 1} \right) d\phi \\
&= \frac{1}{B(\boldsymbol{\alpha})} \int_{\phi} \prod_{v=1}^{V} \phi_v^{n_v + \alpha_v - 1}\; d\phi \\
&= \frac{1}{B(\boldsymbol{\alpha})} \int_{\phi} \frac{B(\vec{n} + \boldsymbol{\alpha})}{B(\vec{n} + \boldsymbol{\alpha})} \prod_{v=1}^{V} \phi_v^{n_v + \alpha_v - 1}\; d\phi \\
&= \frac{B(\vec{n} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \int_{\phi} \underbrace{\frac{1}{B(\vec{n} + \boldsymbol{\alpha})} \prod_{v=1}^{V} \phi_v^{n_v + \alpha_v - 1}}_{Dir(\vec{n} + \boldsymbol{\alpha})}\; d\phi \\
&= \frac{B(\vec{n} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}
\end{aligned}
$$

LDA Inference

- Collapsed Gibbs Sampler



Gibbs Sampling for LDA

Algorithm

```
// initialisation
zero all count variables, n_m^(k), n_m, n_k^(t), n_k
for all documents m ∈ [1, M] do
    for all words n ∈ [1, N_m] in document m do
        sample topic index z_{m,n}=k ~ Mult(1/K)
        increment document–topic count: n_m^(k) += 1
        increment document–topic sum: n_m += 1
        increment topic–term count: n_k^(t) += 1
        increment topic–term sum: n_k += 1
```

Figure from Heinrich (2009)

Gibbs Sampling for LDA

Algorithm

`// Gibbs sampling over burn-in period and sampling period`

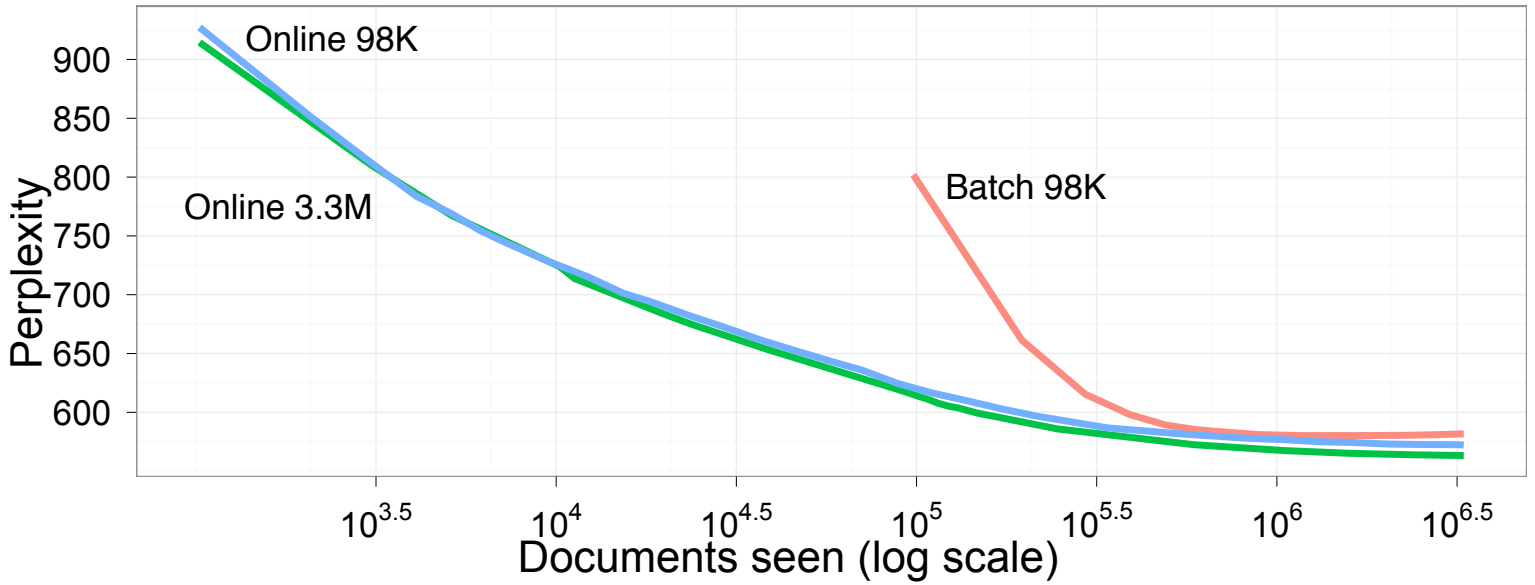
while not finished **do**

- **for** all documents $m \in [1, M]$ **do**
  - **for** all words $n \in [1, N_m]$ in document m **do**
    - `// for the current assignment of` k `to a term` t `for word` $w_{m,n}$`:`
    - decrement counts and sums: $n_m^{(k)} \mathrel{-}= 1; n_m \mathrel{-}= 1; n_k^{(t)} \mathrel{-}= 1; n_k \mathrel{-}= 1$
    - `// multinomial sampling acc. to Eq. 78 (decrements from previous step):`
    - sample topic index $\tilde{k} \sim p(z_i | \vec{z}_{\neg i}, \vec{w})$
    - `// for the new assignment of` $z_{m,n}$ `to the term` t `for word` $w_{m,n}$`:`
    - increment counts and sums: $n_m^{(\tilde{k})} \mathrel{+}= 1; n_m \mathrel{+}= 1; n_{\tilde{k}}^{(t)} \mathrel{+}= 1; n_{\tilde{k}} \mathrel{+}= 1$

Figure from Heinrich (2009)

(Hoffman et al., 2010)

Online Variational Inference for LDA



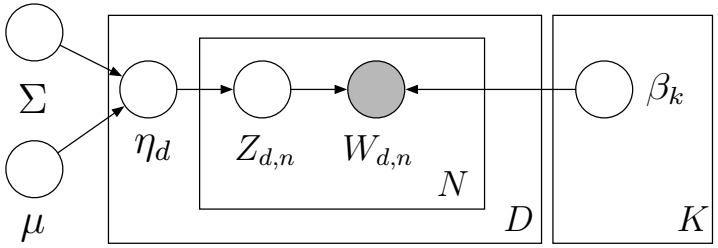
Documents analyzed	2048	4096	8192	12288	16384	32768	49152	65536
Top eight words	systems road made service announced national west language	systems health communication service billion language care road	service systems health companies market communication company billion	service systems companies business company billion health industry	service companies systems business company industry market billion	business service companies industry company management systems services	business service companies industry services company management public	business industry service companies services company management public

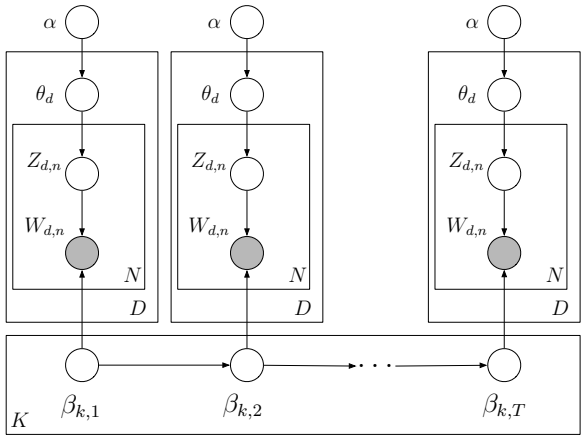
EXTENSIONS OF LDA

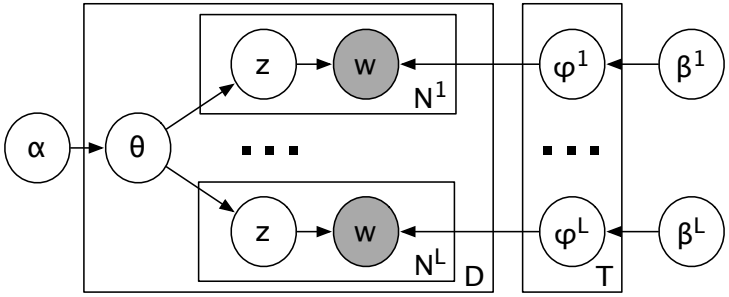
Extensions to the LDA Model

- Correlated topic models
  - Logistic normal prior over topic assignments
- Dynamic topic models
  - Learns topic changes over time
- Polylingual topic models
  - Learns topics aligned across multiple languages

…

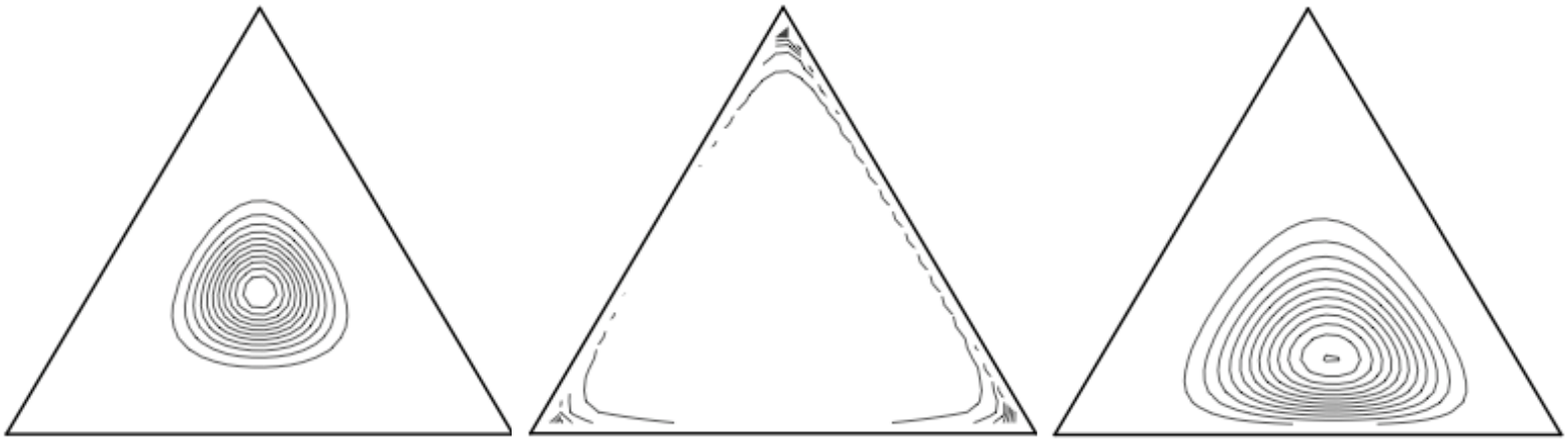






(Blei & Lafferty, 2004)

Correlated Topic Models

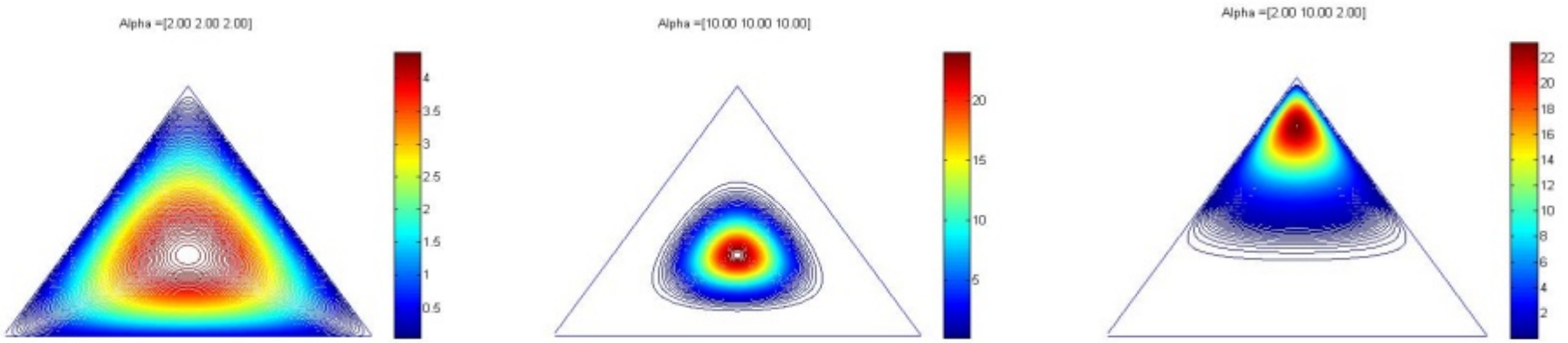


- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models

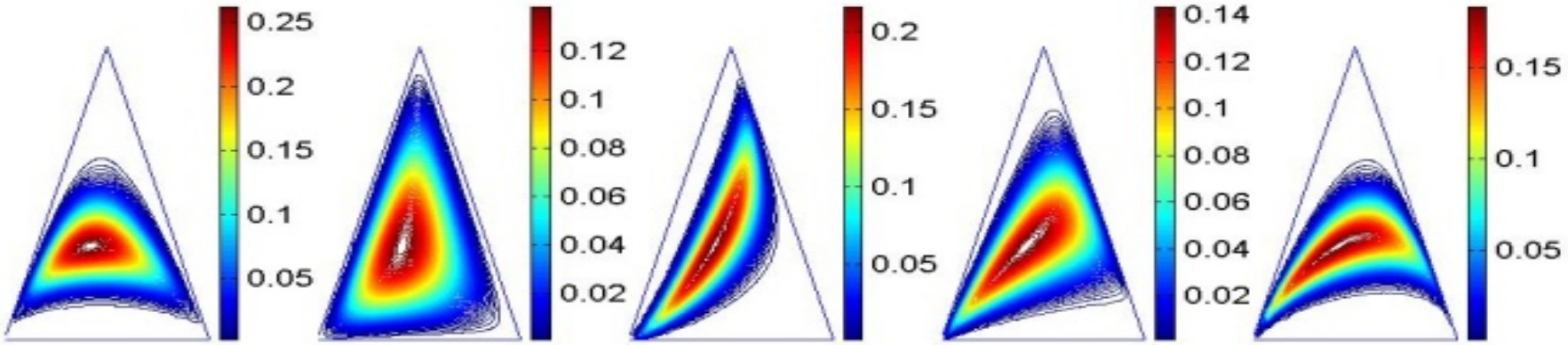


- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models



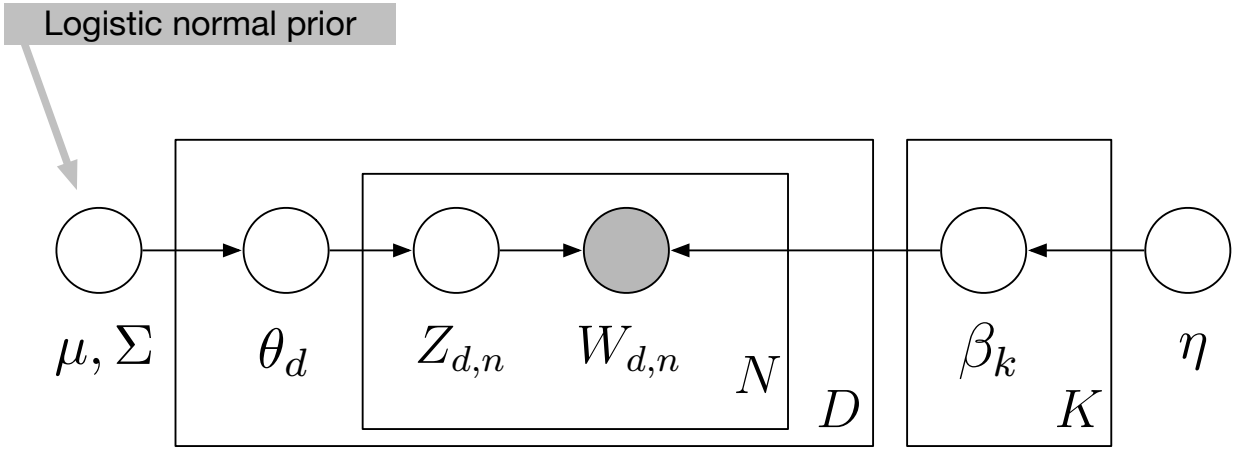
- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$X \sim \mathcal{N}_K(\mu, \Sigma)$$
$$\theta_i \propto \exp\{x_i\}.$$

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models

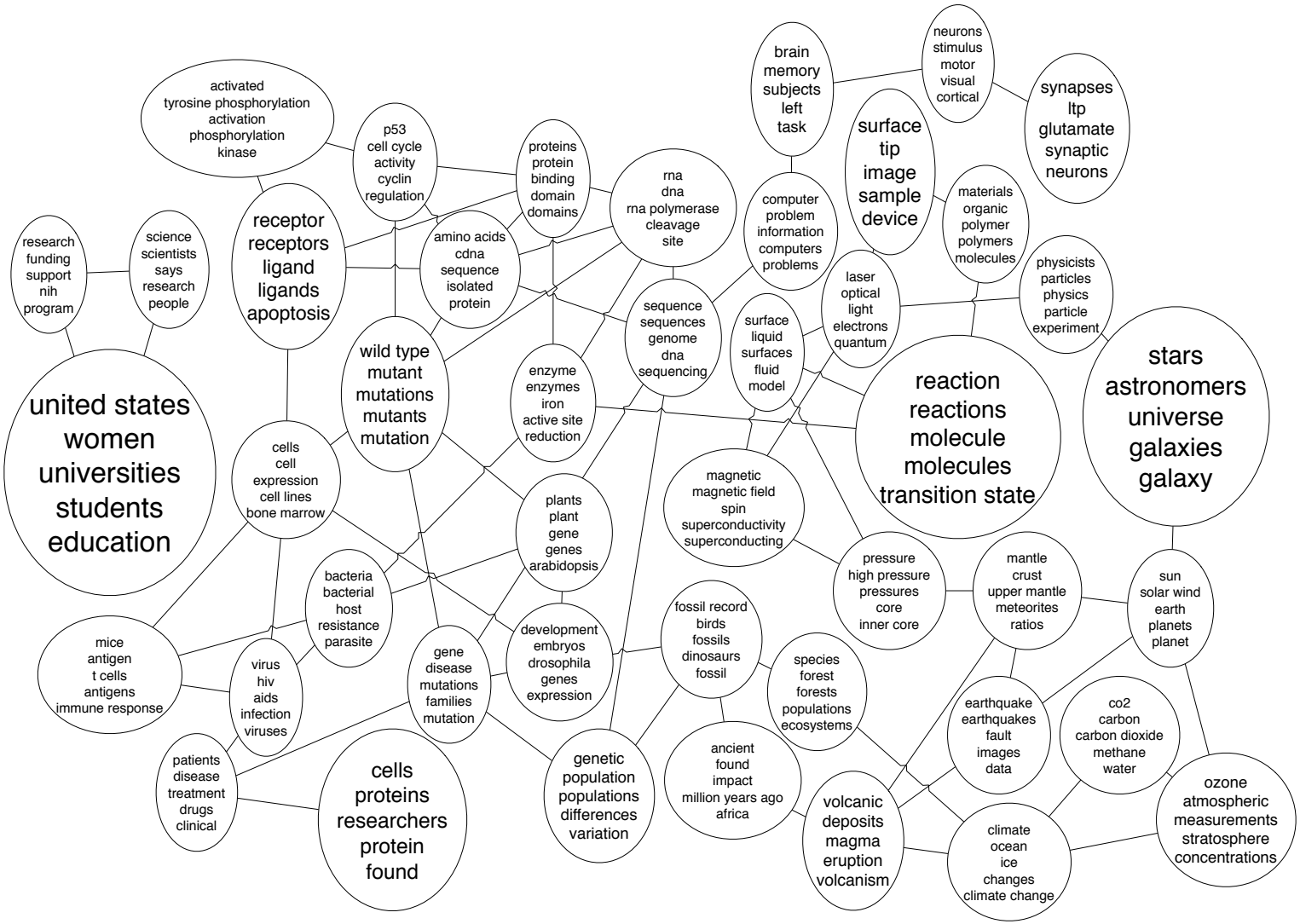


- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a "map" of topics and how they are related
- Provides a better fit to text data, but computation is more complex

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models



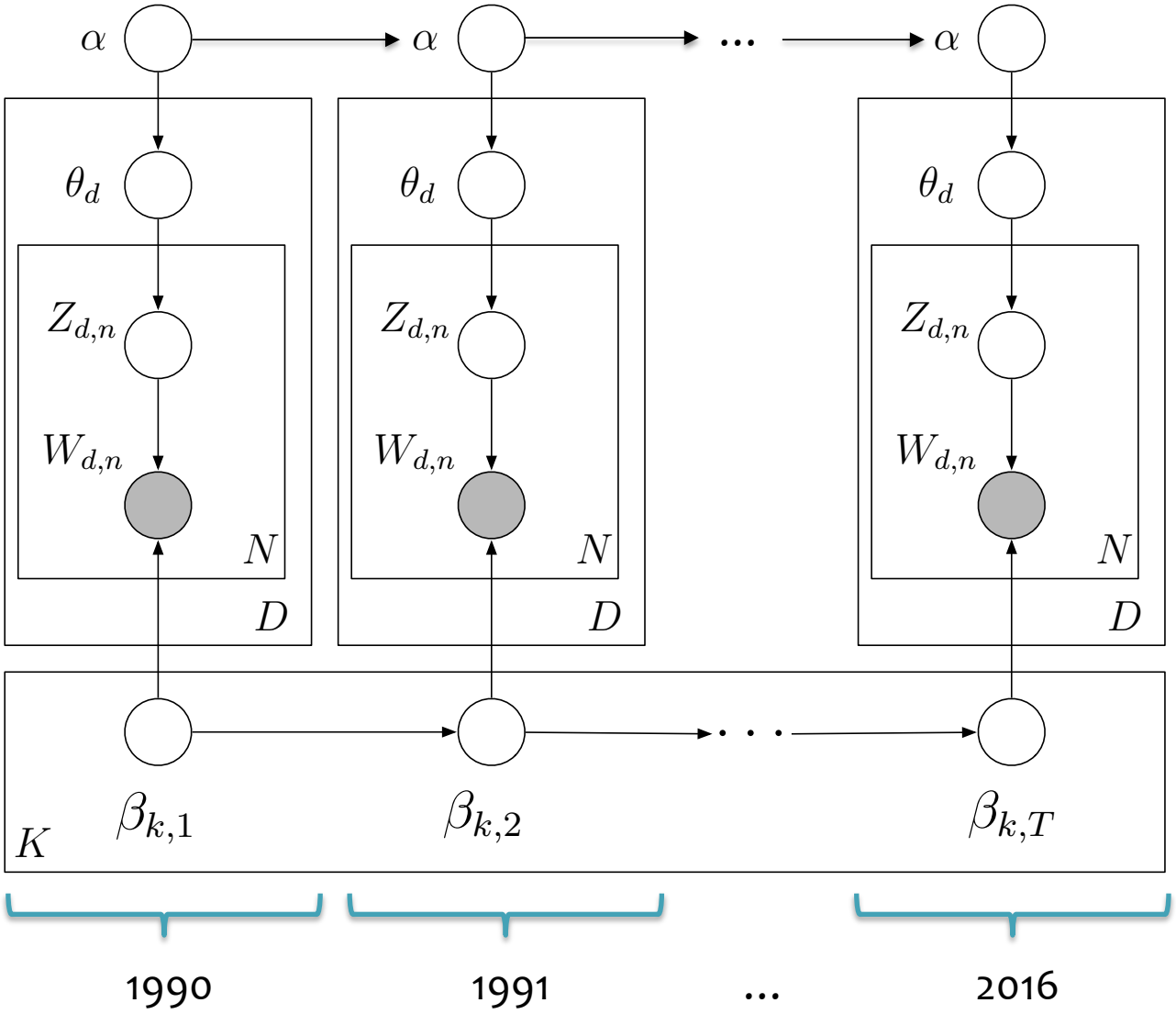
Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

High-level idea:

- Divide the documents up by year
- Start with a separate topic model for each year
- Then add a dependence of each year on the previous one



(Blei & Lafferty, 2006)

Dynamic Topic Models

1789

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

Inaugural addresses

2009

AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time.
- Dynamic topic models let the topics *drift* in a sequence.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

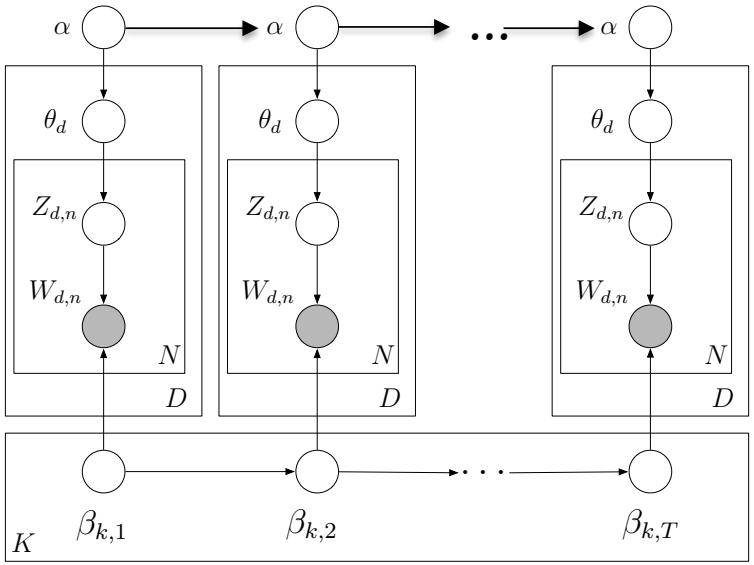
Generative Story

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim \mathit{Mult}(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim \mathit{Mult}(\pi(\beta_{t,z}))$.

Logistic-normal priors

The pi function maps from the natural parameters to the mean parameters:

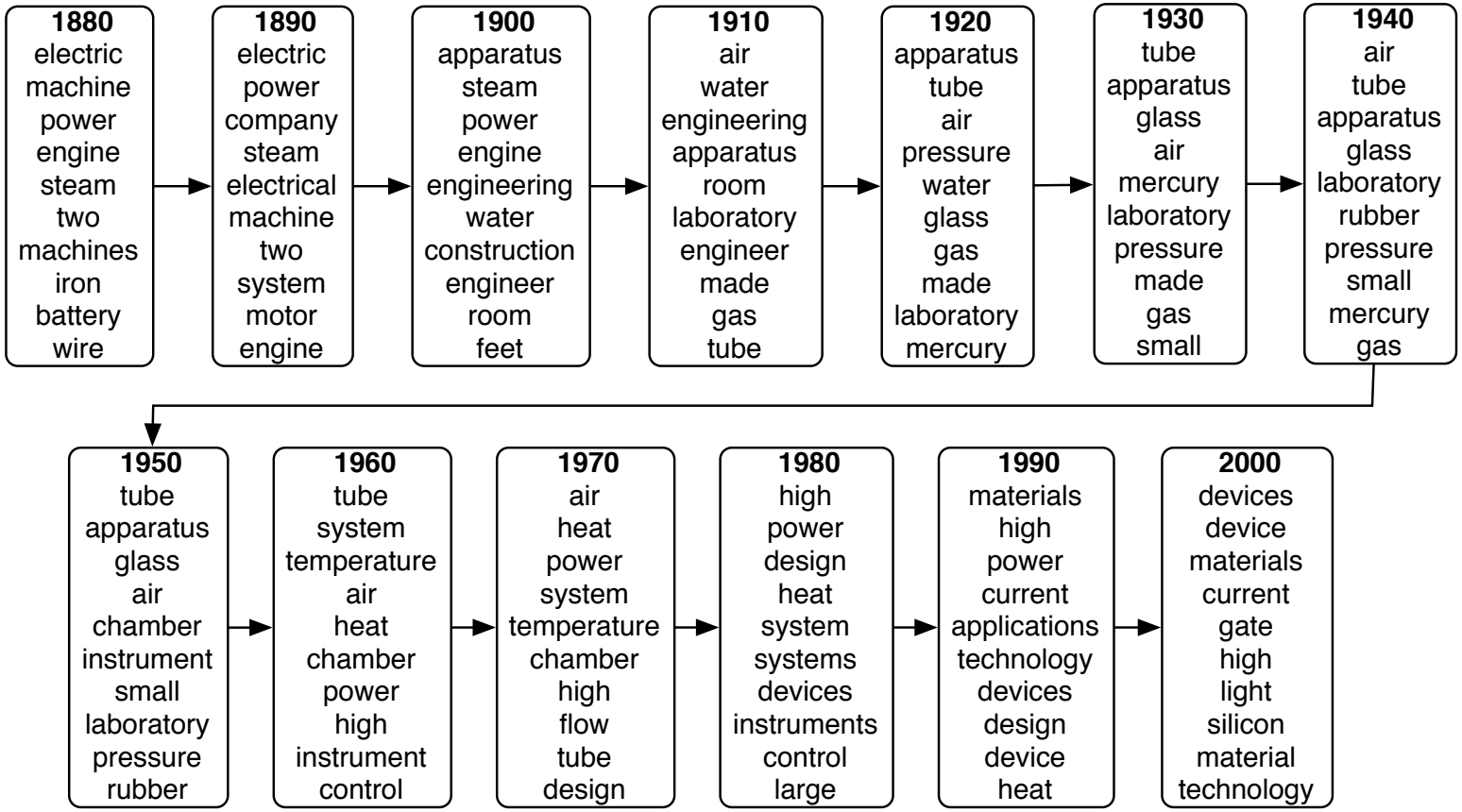
$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}.$$



(Blei & Lafferty, 2006)

Dynamic Topic Models

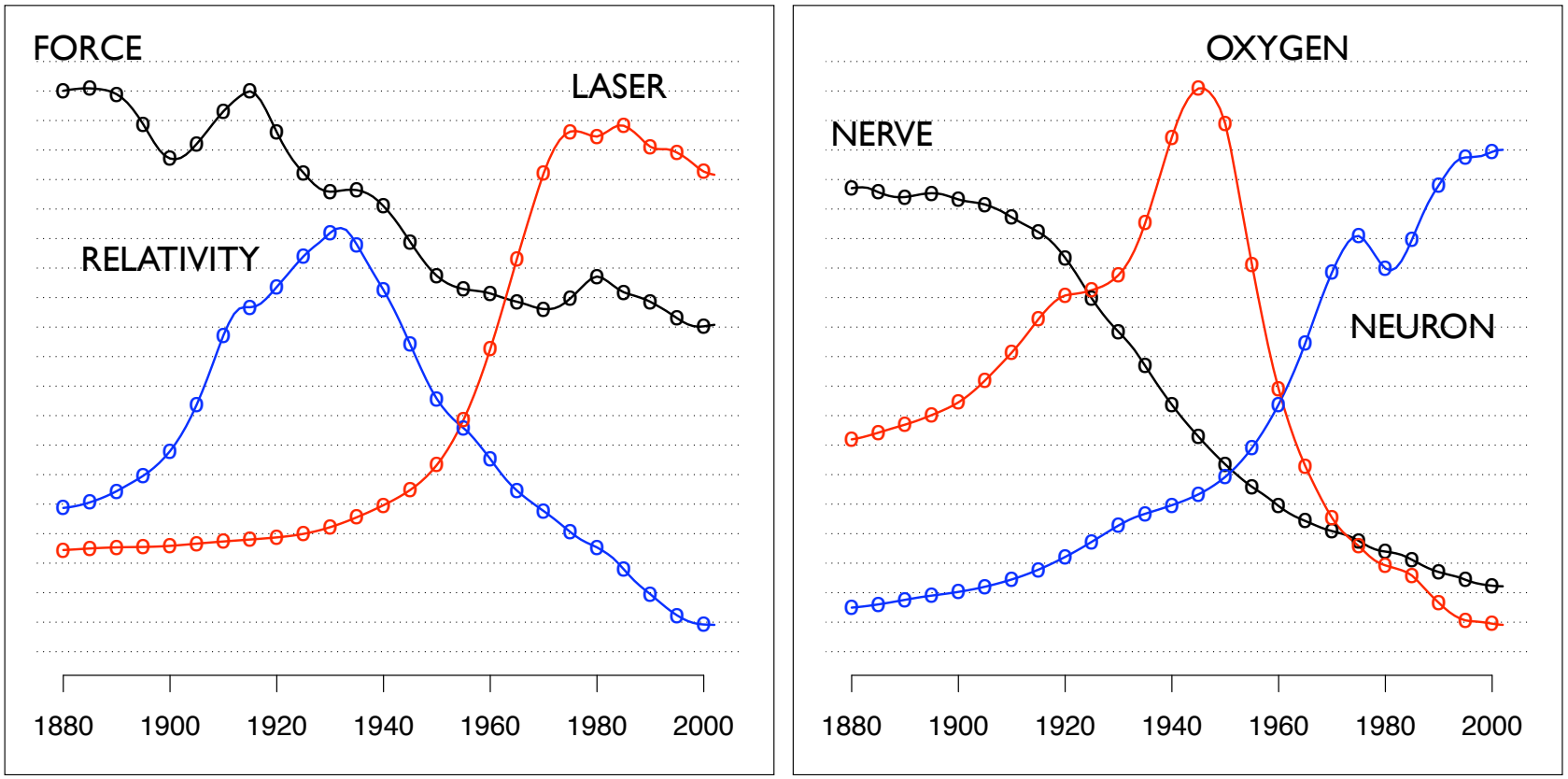
Top ten most likely words in a "drifting" topic shown at 10-year increments



(Blei & Lafferty, 2006)

Dynamic Topic Models

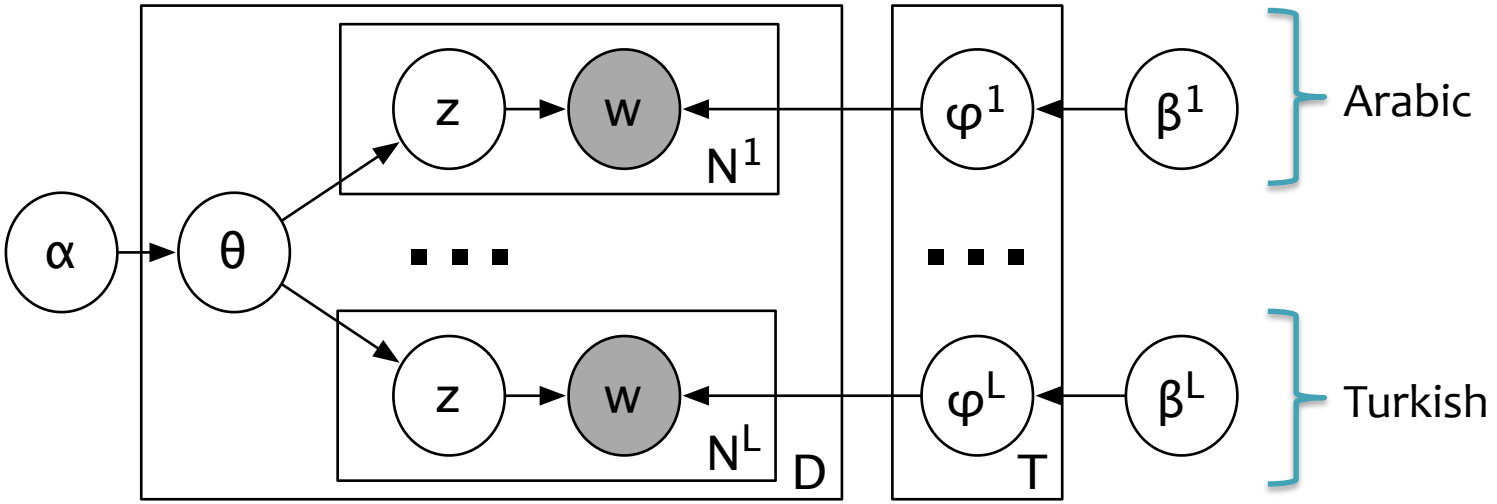
Posterior estimate of **word frequency as a function of year** for three words each in two separate topics:



(Mimno et al., 2009)

Polylingual Topic Models

- **Data Setting:** Comparable versions of each document exist in multiple languages (e.g. the Wikipedia article for "Barak Obama" in twelve languages)
- **Model:** Very similar to LDA, except that the topic assignments, z, and words, w, are sampled separately for each language.



(Mimno et al., 2009)

Polylingual Topic Models

Topic 1 (twelve languages)

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

(Mimno et al., 2009)

Polylingual Topic Models

Topic 2 (twelve languages)

CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	اسپانیا اسپانیایی کوبا مادرید de ترین
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

(Mimno et al., 2009)

Polylingual Topic Models

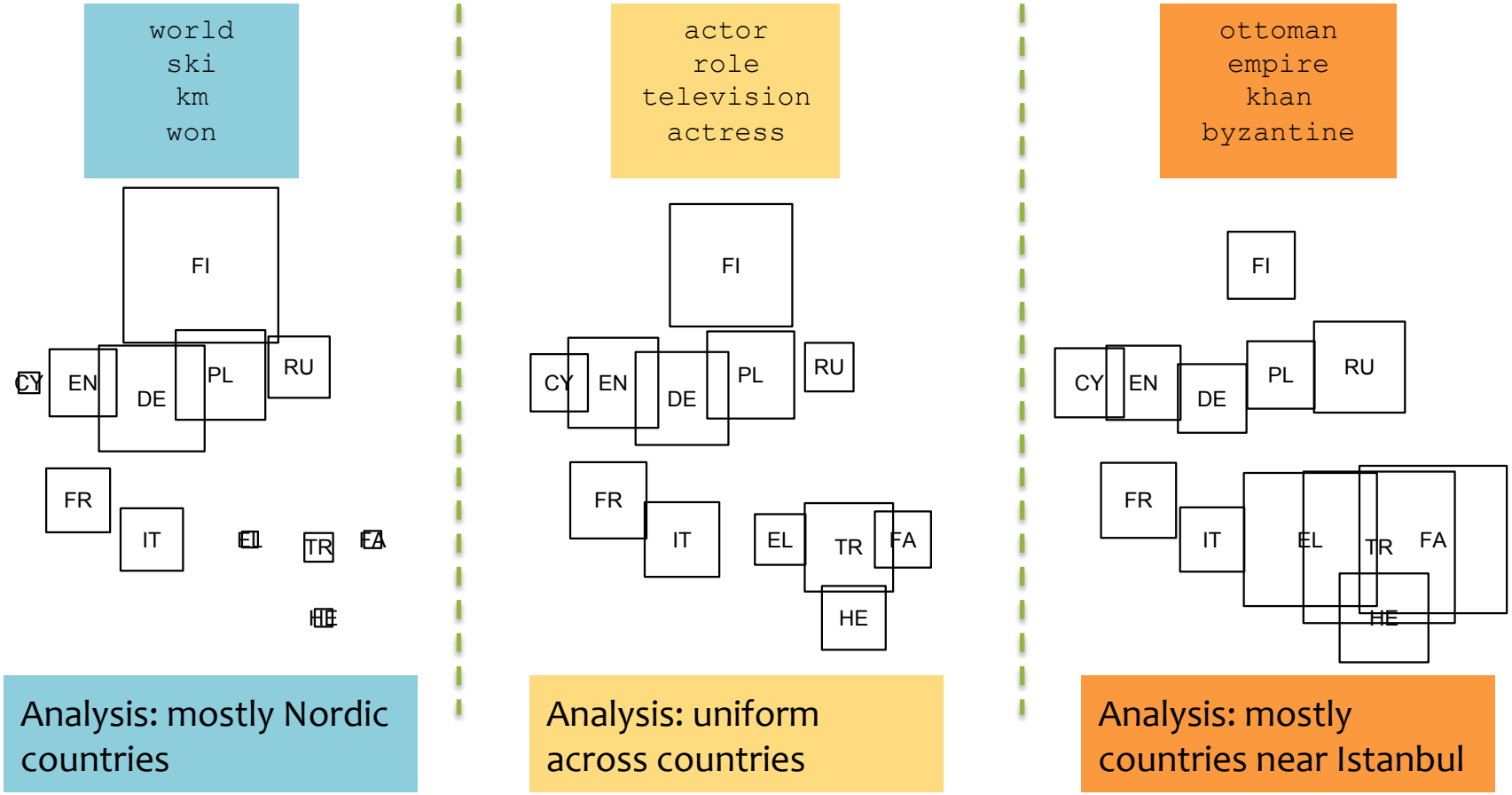
Topic 3 (twelve languages)

CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

(Mimno et al., 2009)

Polylingual Topic Models

Size of each square represents proportion of tokens assigned to the specified topic.



Analysis: mostly Nordic countries

Analysis: uniform across countries

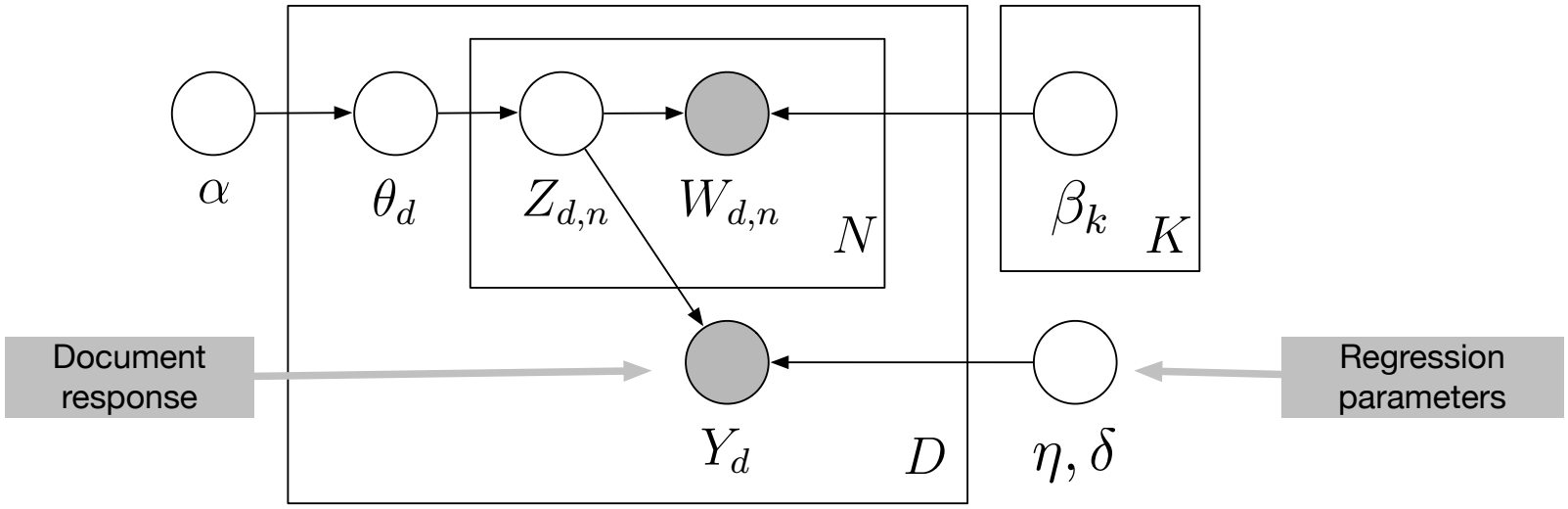
Analysis: mostly countries near Istanbul

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.

Slide from David Blei, MLSS 2012

Supervised LDA

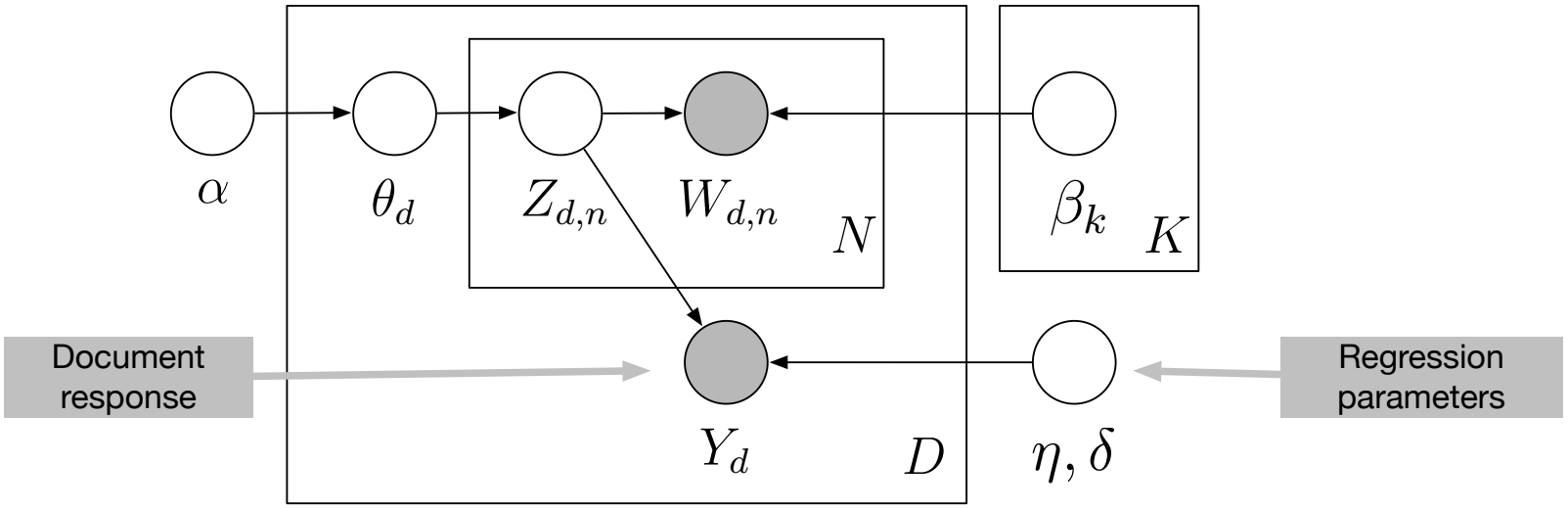


1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim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

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

Slide from David Blei, MLSS 2012

Supervised LDA



- Fit sLDA parameters to documents and responses.
  This gives: topics $\beta_{1:K}$ and coefficients $\eta_{1:K}$.
- Given a new document, predict its response using the expected value:

$$\mathrm{E}\left[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2\right] = \eta^\top \mathrm{E}\left[\bar{Z} \mid w_{1:N}\right]$$

- This blends generative and discriminative modeling.

Slide from David Blei, MLSS 2012

What if we don't know the number of topics, K, ahead of time?

Take 10-708 to learn **Bayesian Nonparametrics**

- New modeling constructs:
  - Chinese Restaurant Process (Dirichlet Process)
  - Indian Buffet Process
- e.g. an **infinite number of topics** in a finite amount of space

Summary: Topic Modeling

- **The Task of Topic Modeling**
  - Topic modeling enables the **analysis of large** (possibly unannotated) **corpora**
  - Applicable to more than just bags of words
  - Extrinsic evaluations are often appropriate for these unsupervised methods
- **Constructing Models**
  - LDA is comprised of **simple building blocks** (Dirichlet, Multinomial)
  - LDA itself can act as a building block **for other models**
- **Approximate Inference**
  - Many different approaches to inference (and learning) can be applied to the same model