10-708 Probabilistic Graphical Models

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Topic Modeling

Matt Gormley
Lecture 15
Mar. 24, 2021

Reminders

- Homework 3: Structured SVM
  - Out: Wed, Mar. 10
  - Due: Wed, Mar. 24 at 11:59pm
- Project Proposal
  - Due: Wed, Mar. 31 at 11:59pm
- Homework 4: MCMC
  - Out: Wed, Mar. 24
  - Due: Wed, Apr. 7 at 11:59pm


SOLID
STOP

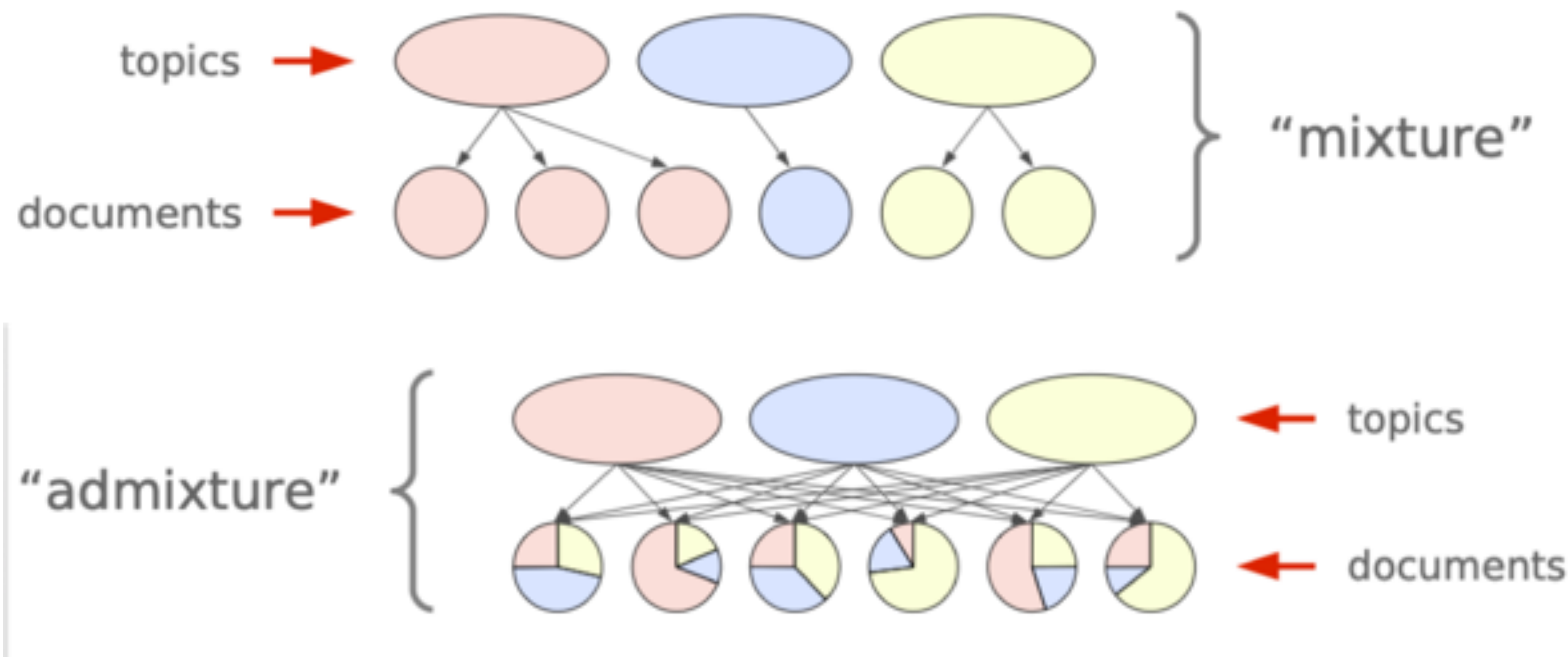
Plate Diagrams

***Whiteboard*:**

- Example: Dirichet-Multinomial as a directed graphical model
- Example: Plate diagram for Dirichlet-Multinomial model

LATENT DIRICHLET ALLOCATION (LDA)

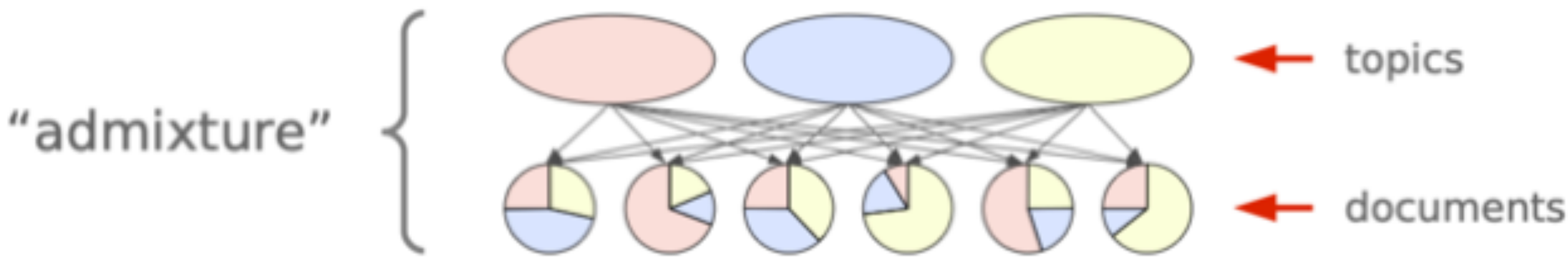
Mixture vs. Admixture (LDA)



Figures 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

Figure from Wallach, JHU 2011, slides

Latent Dirichlet Allocation

- Generative Process

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

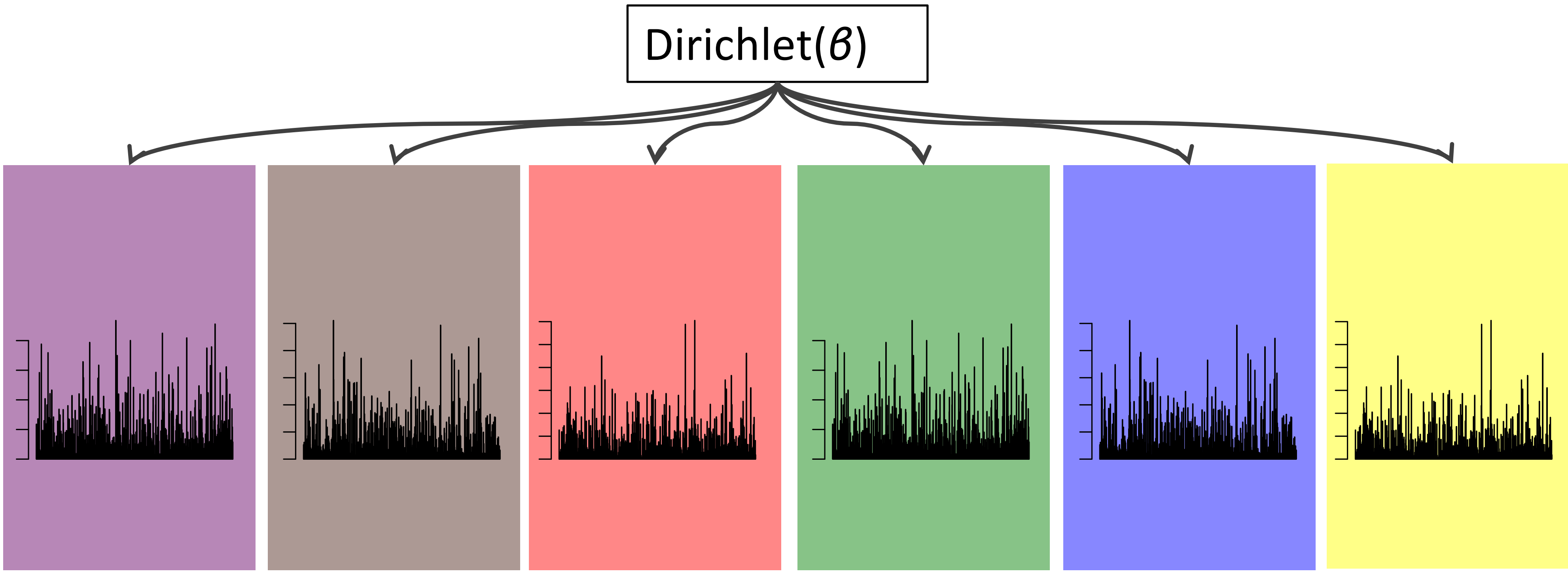
Document 2

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

Document 3

(Blei, Ng, & Jordan, 2003)

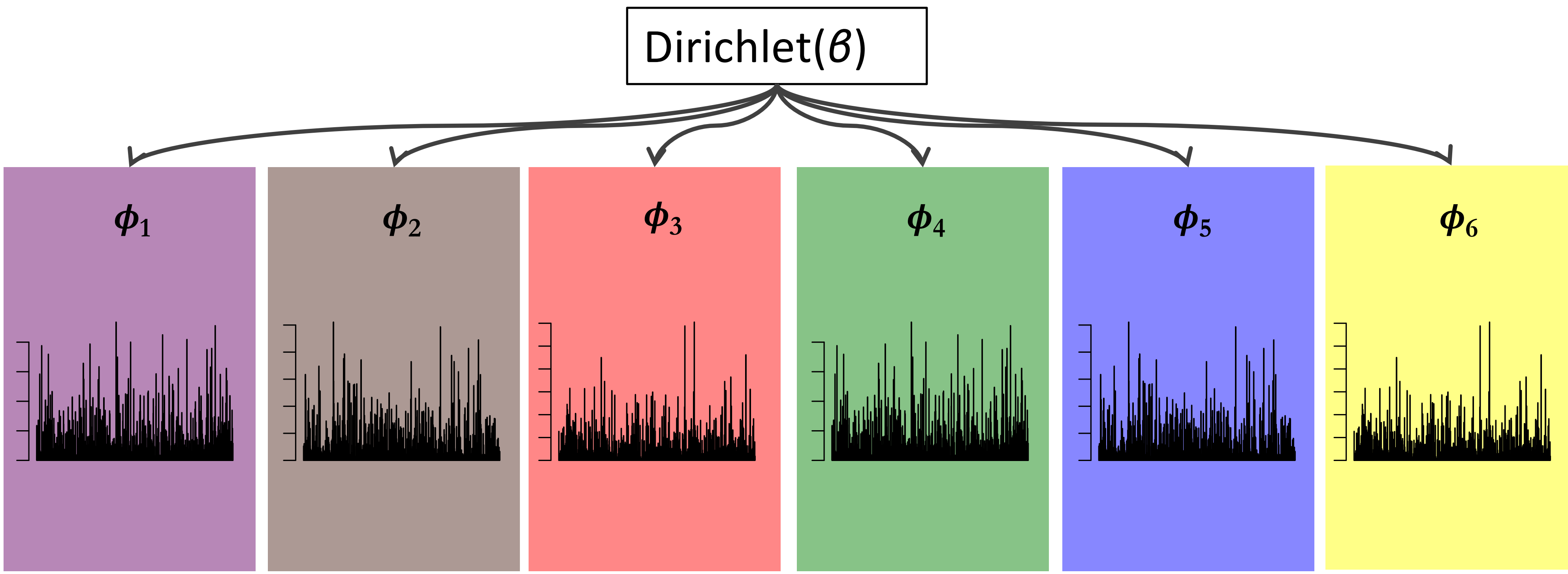
LDA for Topic Modeling



- 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)

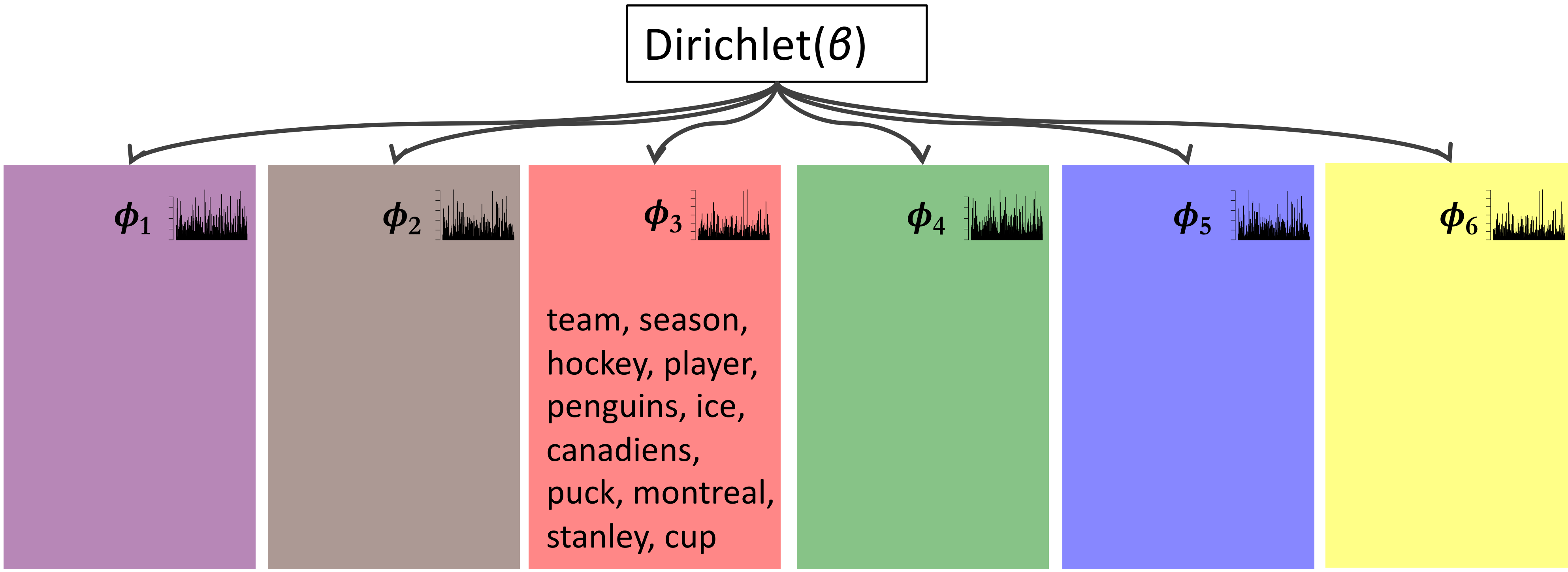
LDA for Topic Modeling



- 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)

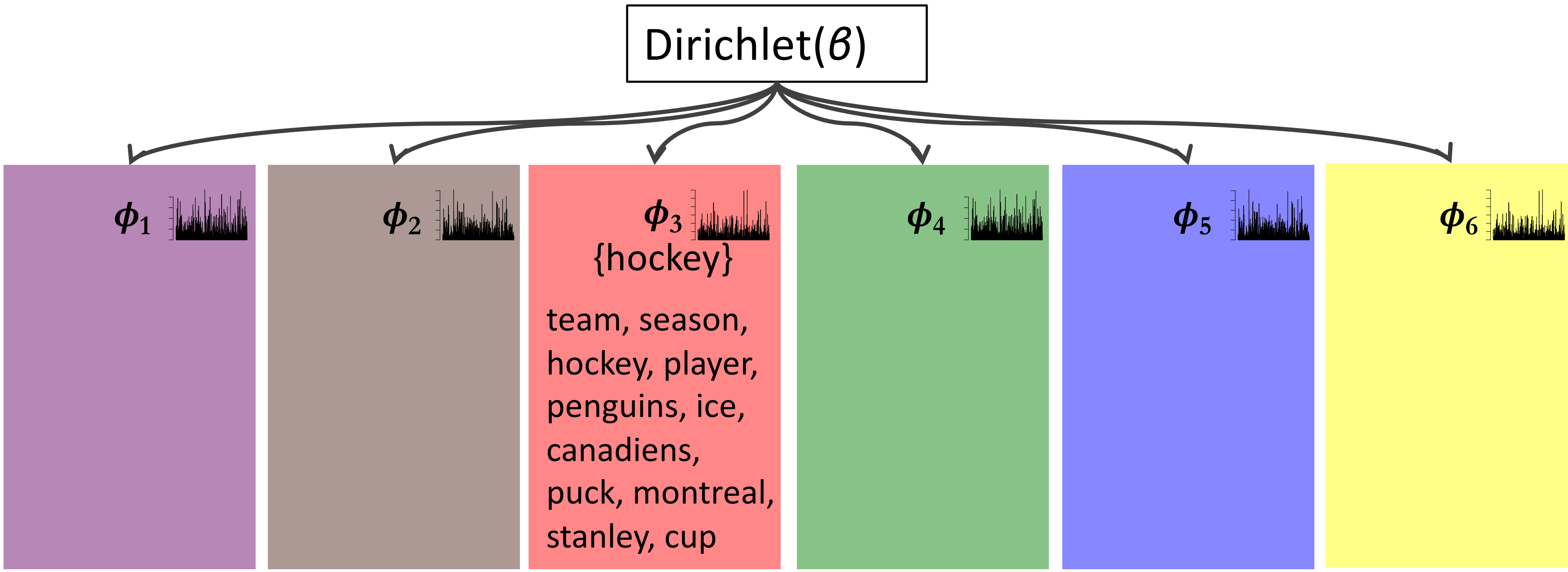
LDA for Topic Modeling



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

(Blei, Ng, & Jordan, 2003)

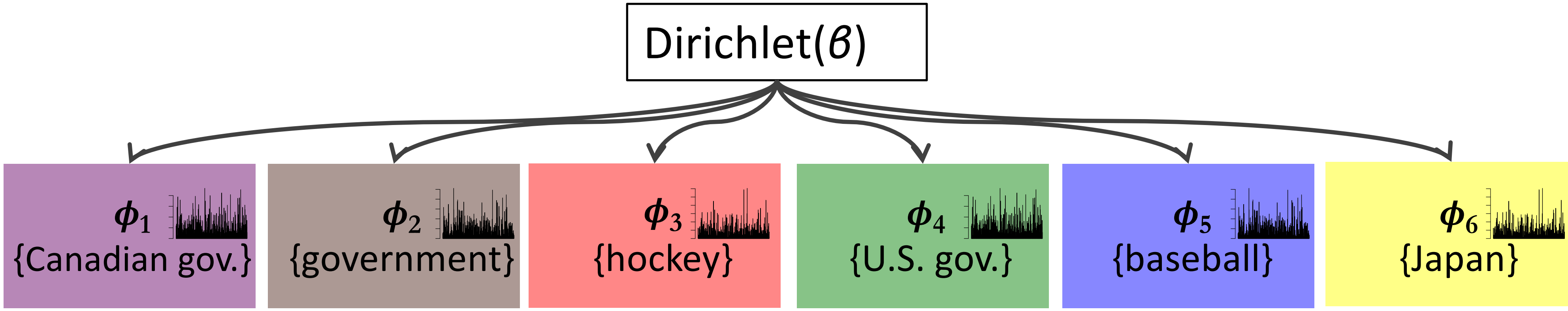
LDA for Topic Modeling



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

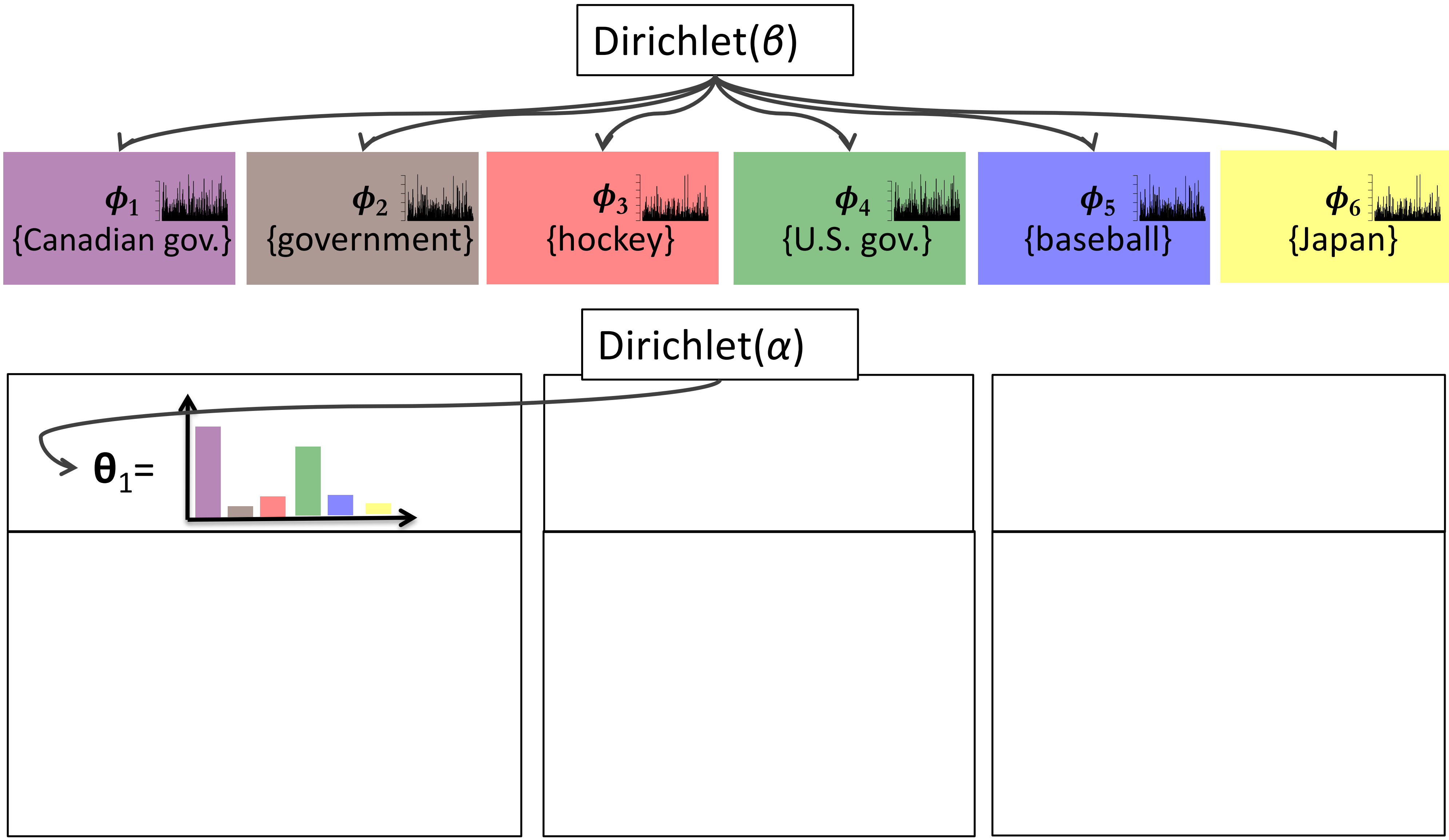
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



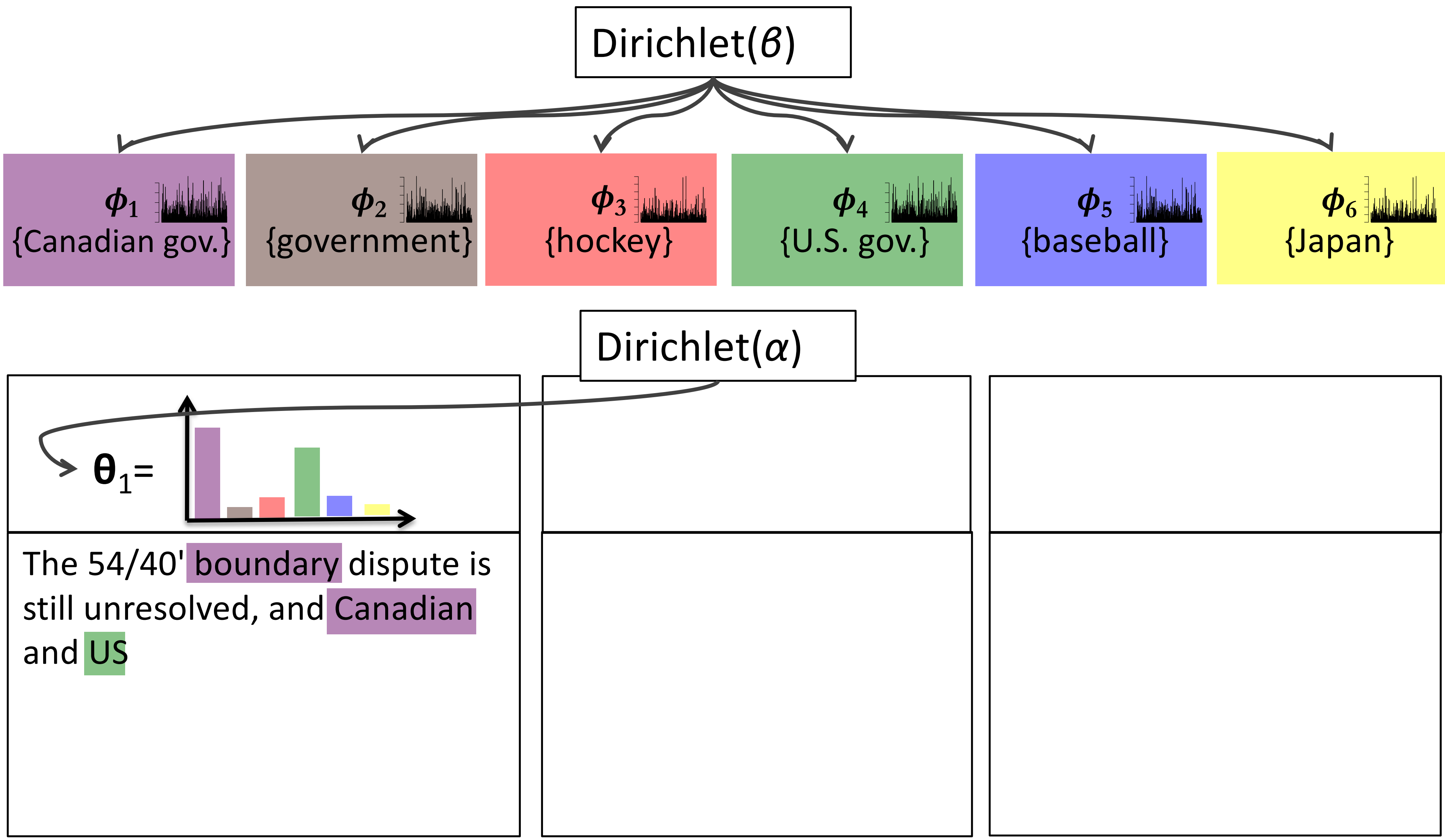
- A topic is visualized as its high probability words.
- A pedagogical **label** is used to identify the topic.

LDA for Topic Modeling



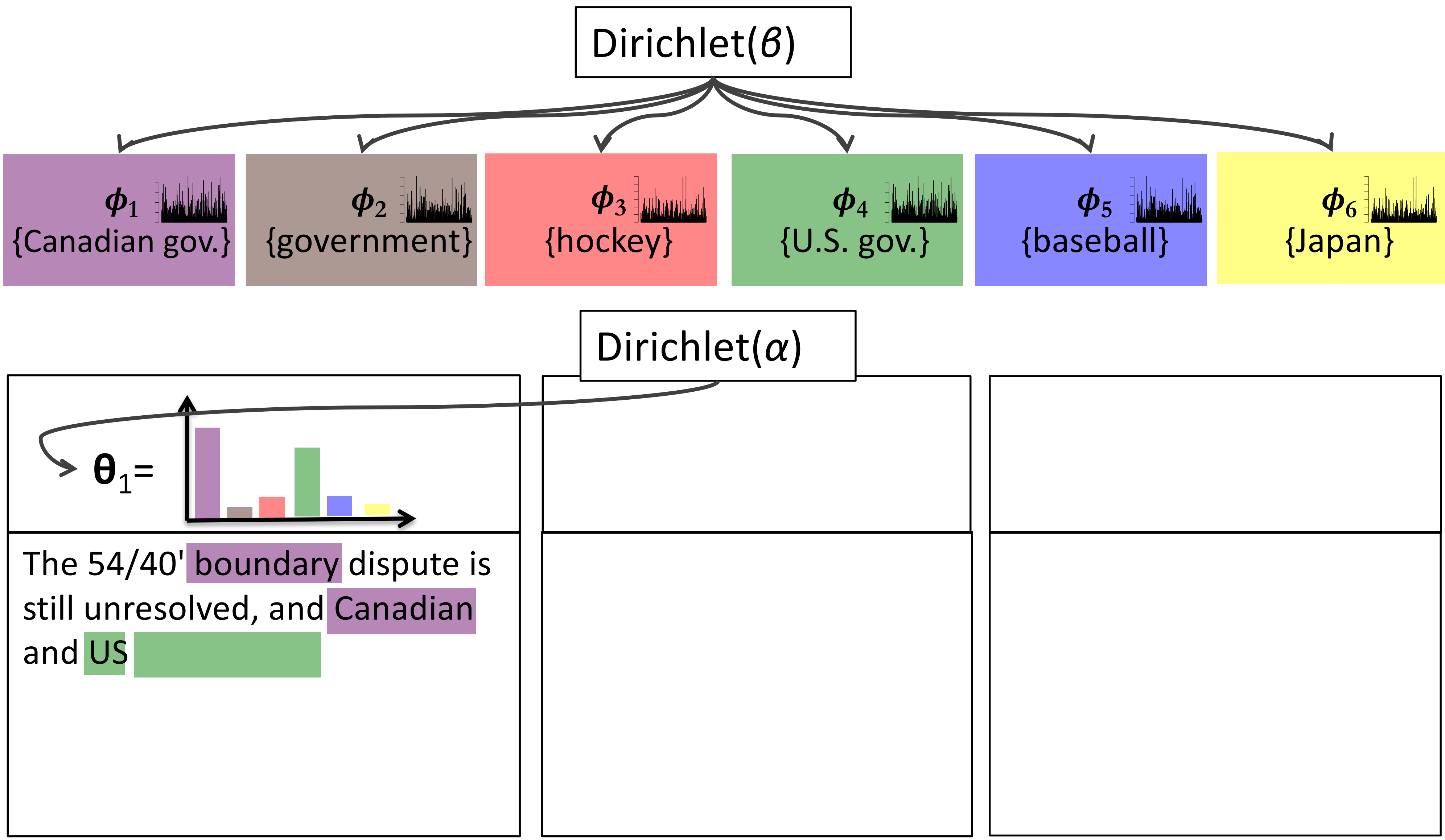
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



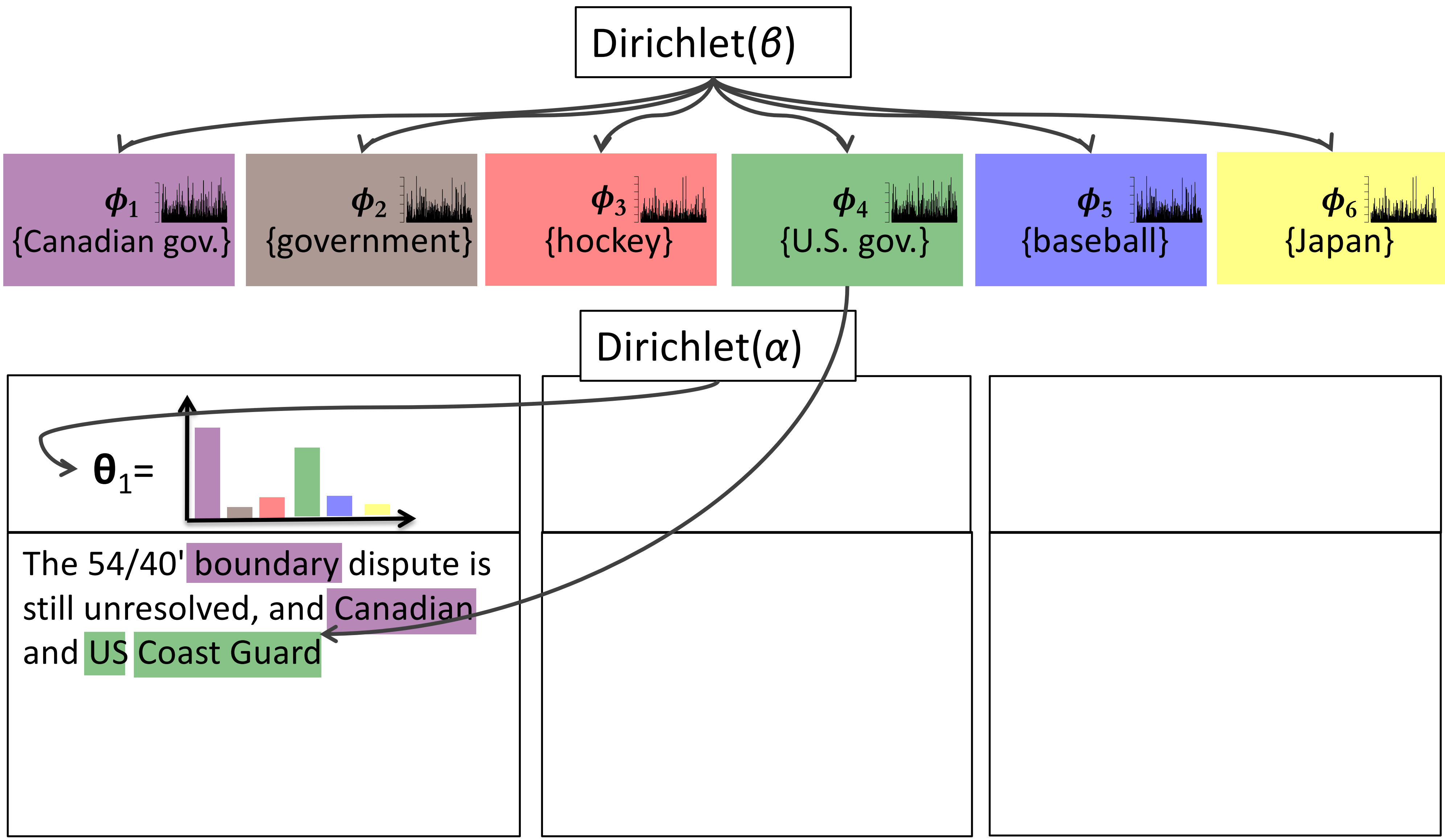
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



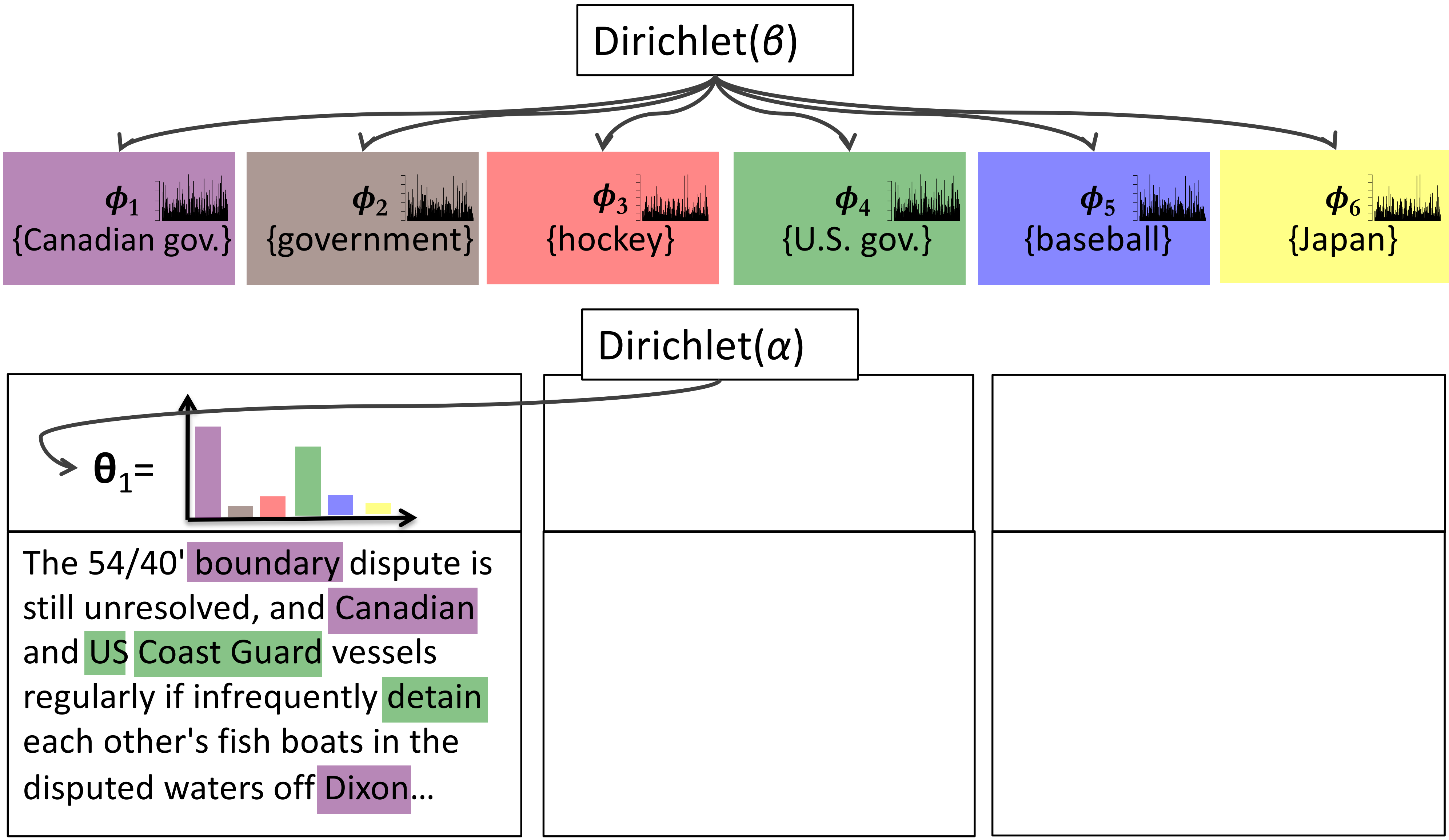
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



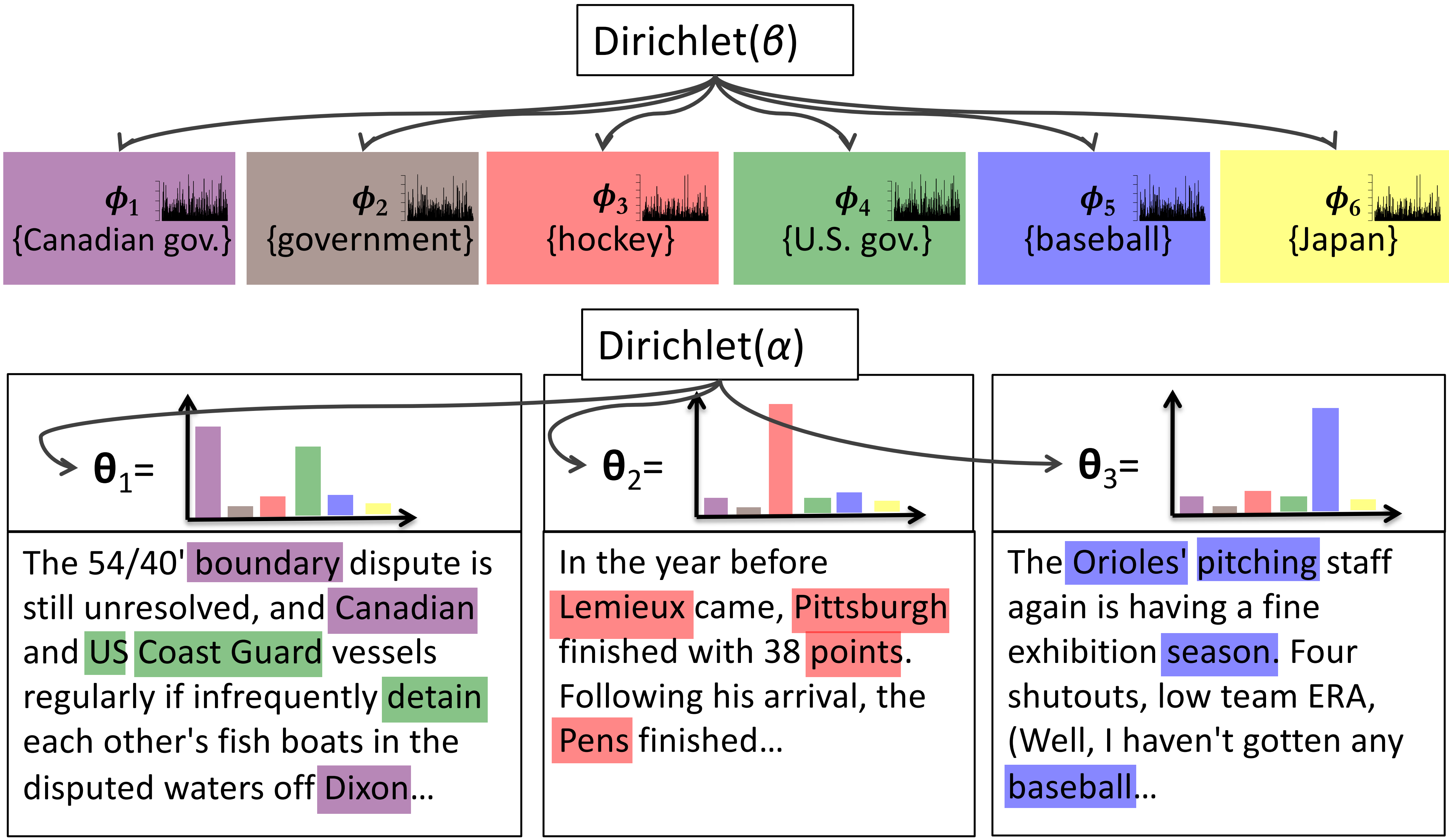
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



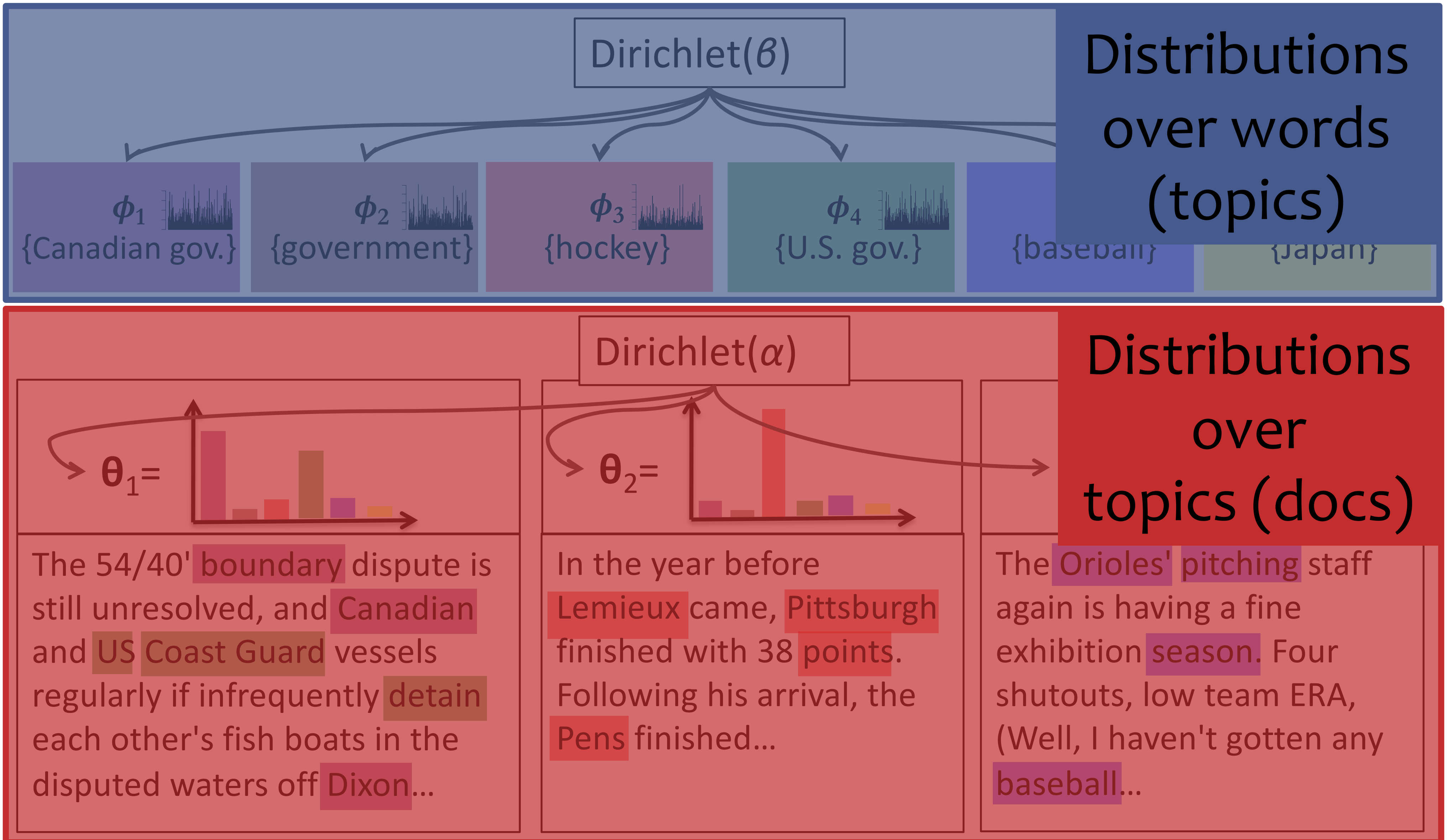
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



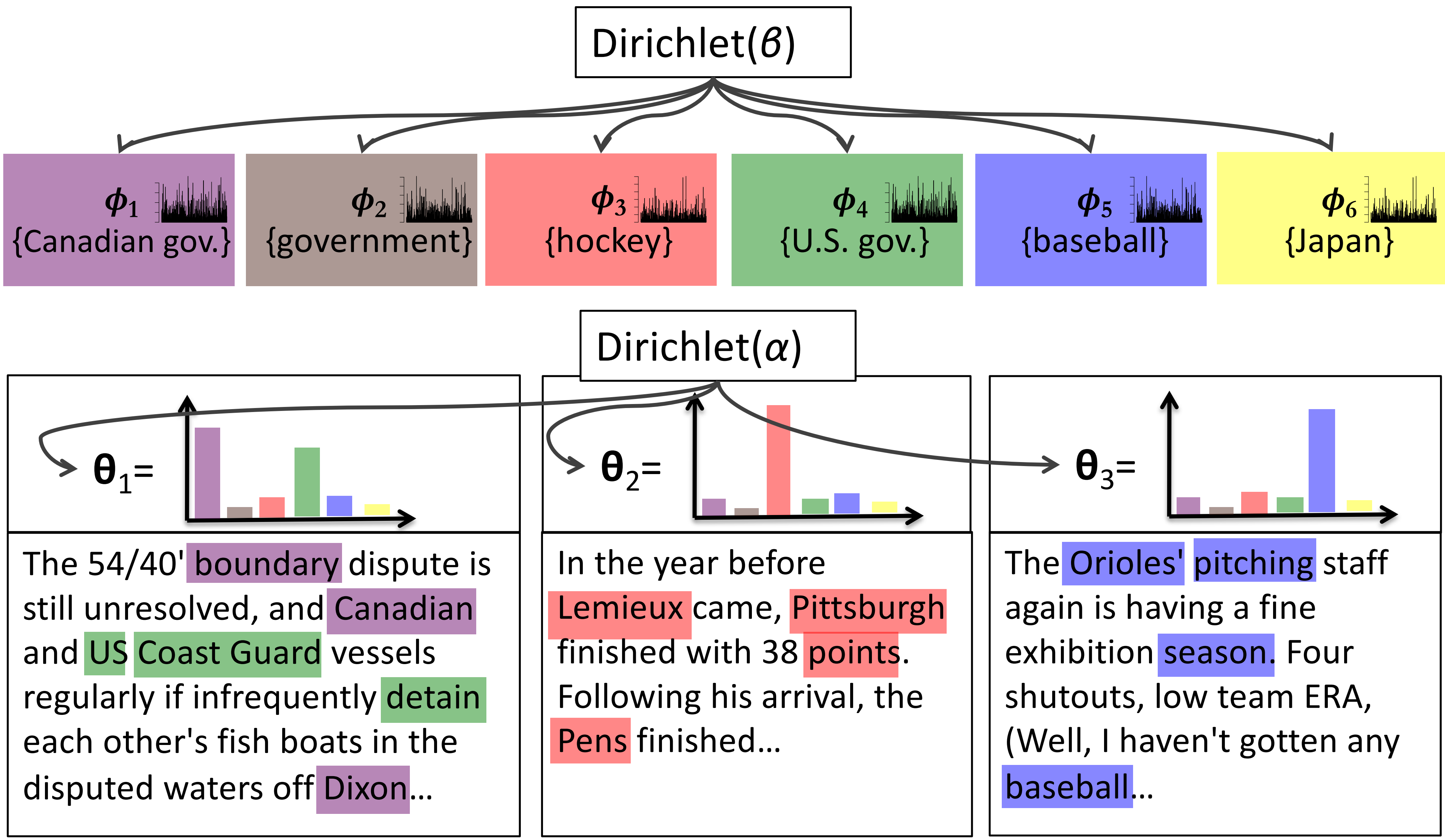
(Blei, Ng, & Jordan, 2003)

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

Dirichlet( )

ϕ_1 = | ϕ_2 = | ϕ_3 = | ϕ_4 = | ϕ_5 = | ϕ_6 =

Dirichlet( )

θ_1=

The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon…

θ_2=

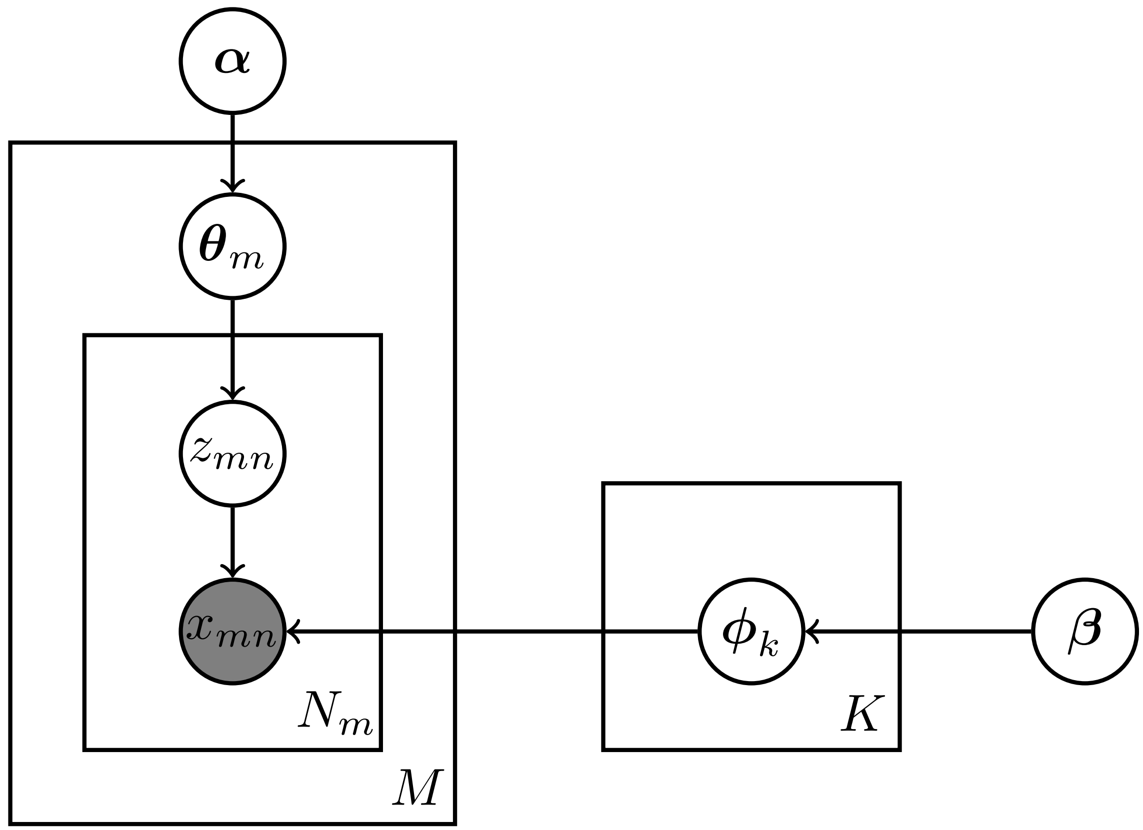
In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished…

θ_3=

The Orioles' itching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball…

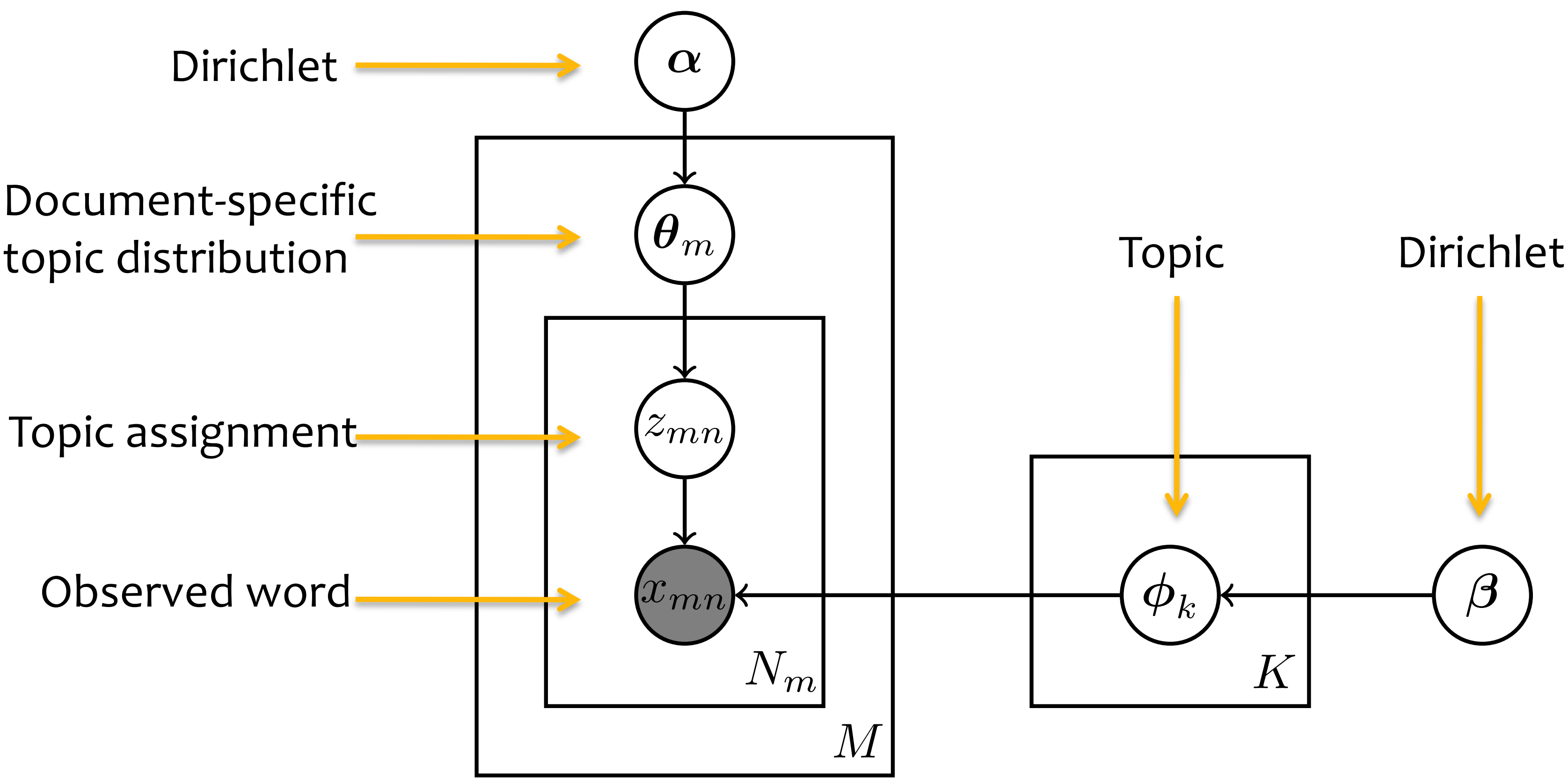
Latent Dirichlet Allocation

- Plate Diagram



Latent Dirichlet Allocation

- Plate Diagram



Latent Dirichlet Allocation

Question: Is this a believable story for the generation of a corpus of documents?	**Answer:**
Question: Why might it work well anyway?	**Answer:**

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 non-negative matrix factorization (Jakulin and Buntine, 2002)
- Was independently invented for genetics (Pritchard et al., 2000)

Slide from David Blei, MLSS 2012

Outline

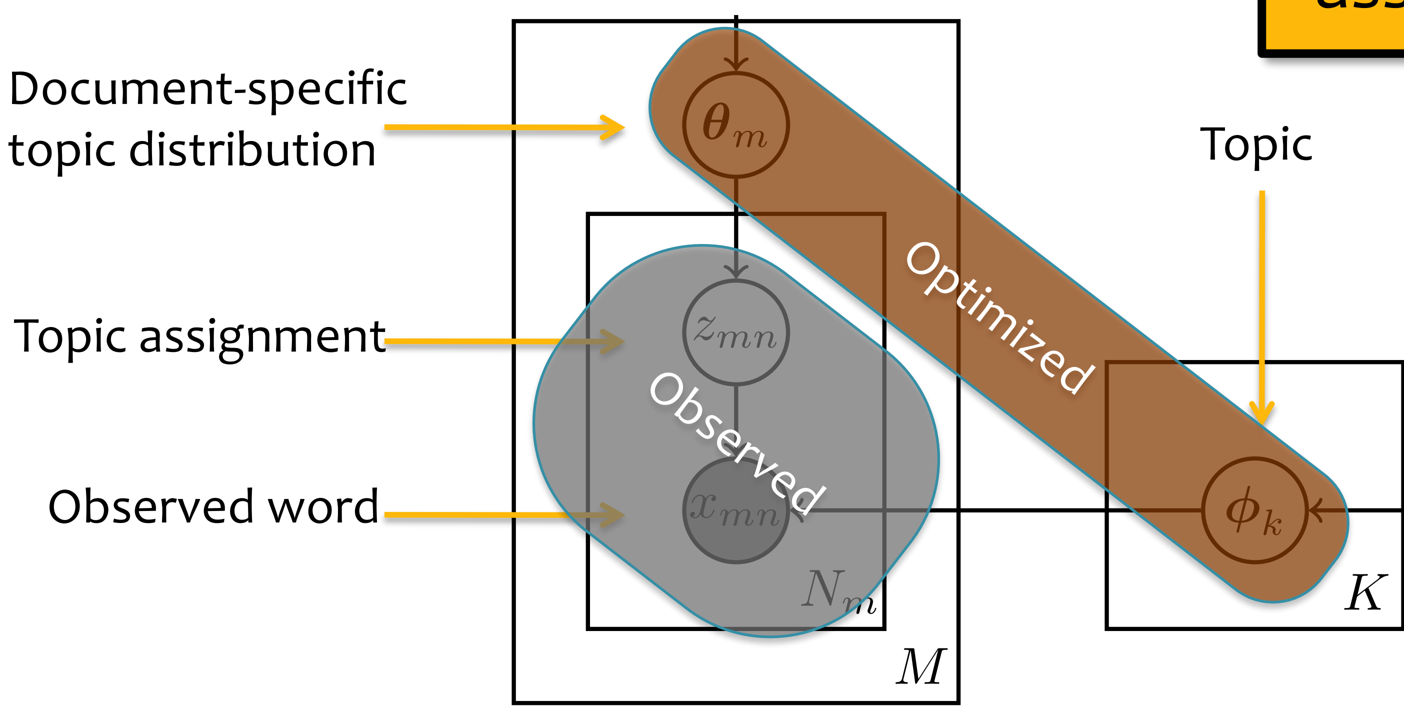
- Applications of Topic Modeling
- Latent Dirichlet Allocation (LDA)
  1. Beta-Bernoulli
  2. Dirichlet-Multinomial
  3. Dirichlet-Multinomial Mixture Model
  4. LDA
- **Bayesian Inference for Parameter Estimation**
  - Exact inference
  - EM
  - Monte Carlo EM
  - Gibbs sampler
  - Collapsed Gibbs sampler
- Extensions of LDA
  - Correlated topic models
  - Dynamic topic models
  - Polylingual topic models
  - Supervised LDA

BAYESIAN INFERENCE FOR PARAMETER ESTIMATION

LDA Inference

- Fully Observed MLE

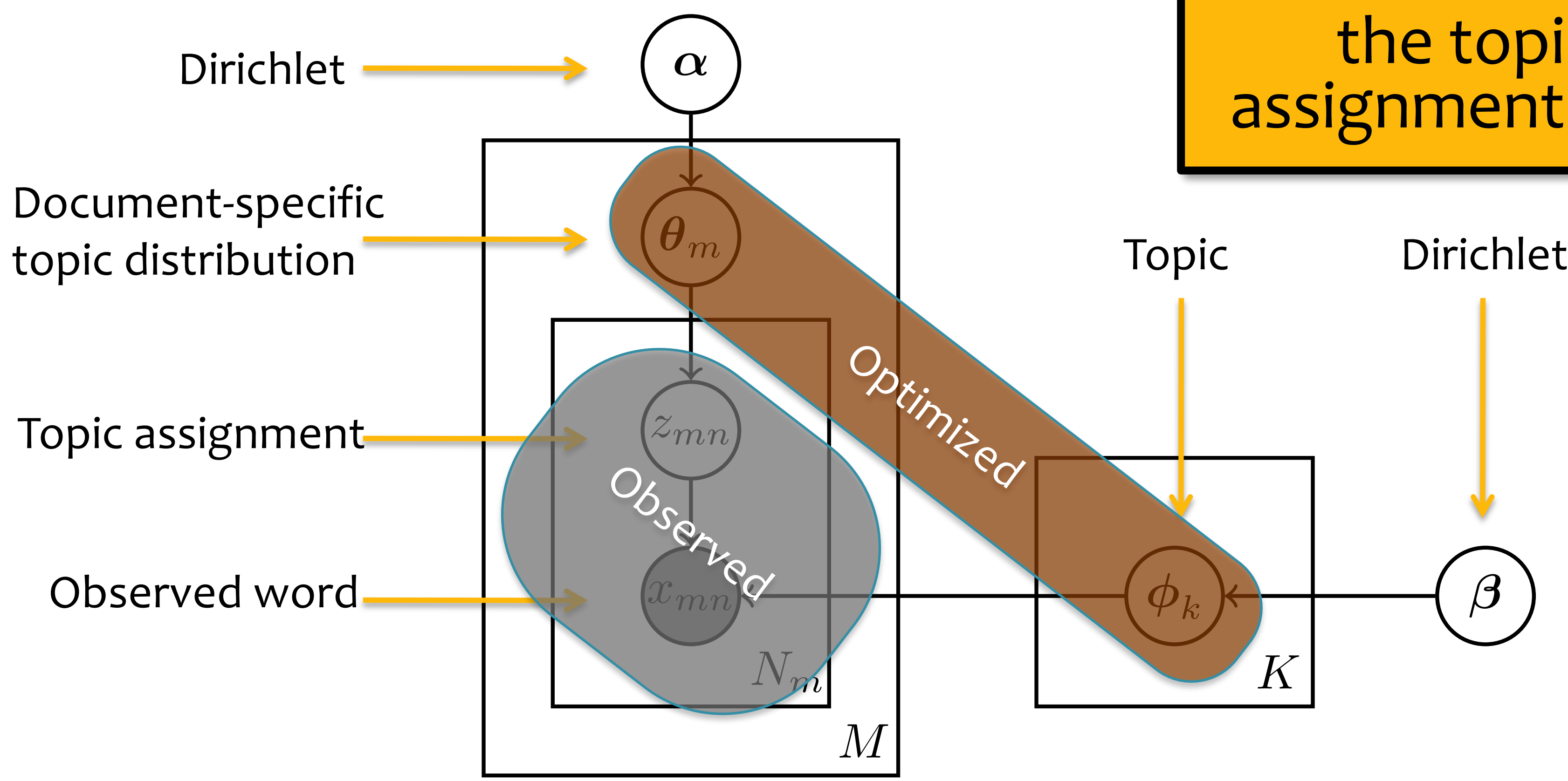
Learning like this would be easy, but in practice we do not observe the topic assignments z_{mn}



LDA Inference

- Full Observed MAP Estimation

Learning like this would be easy, but in practice we do not observe the topic assignments z_{mn}



Unsupervised Learning

Three learning paradigms:

1. Maximum likelihood estimation (MLE)

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

2. Maximum a posteriori (MAP) estimation

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

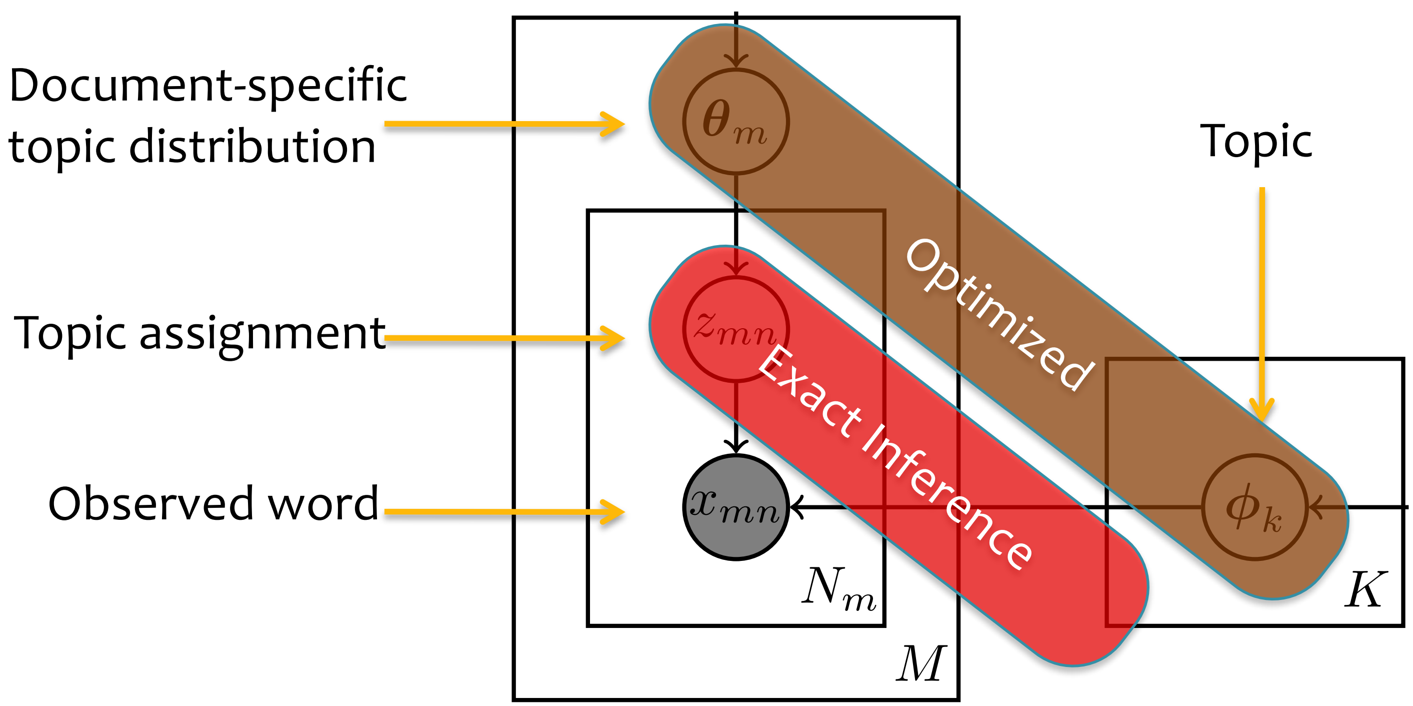
3. Bayesian approach

Estimate the posterior:

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

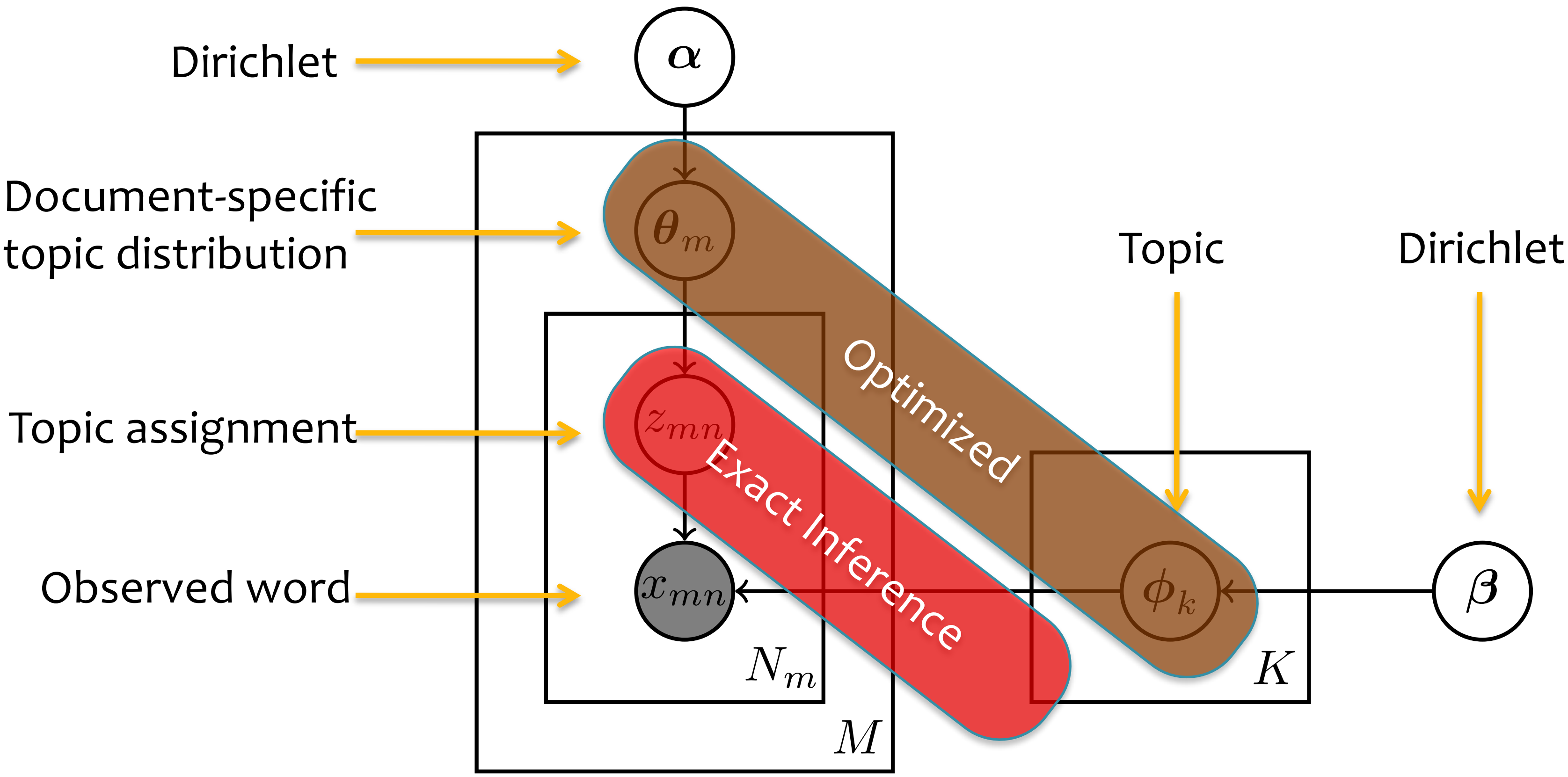
LDA Inference

- Standard EM (MLE)



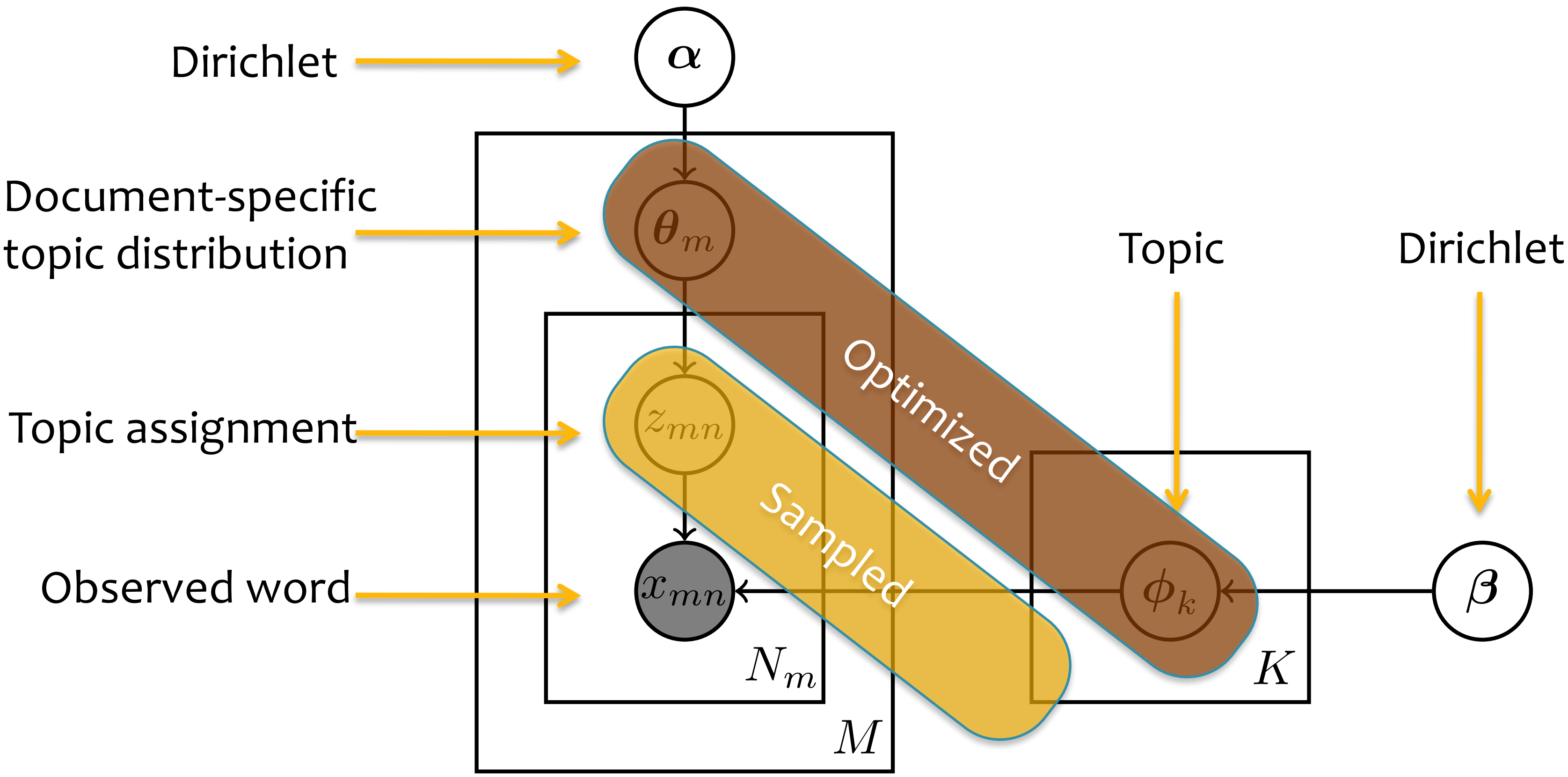
LDA Inference

- Standard EM (MAP Estimation)



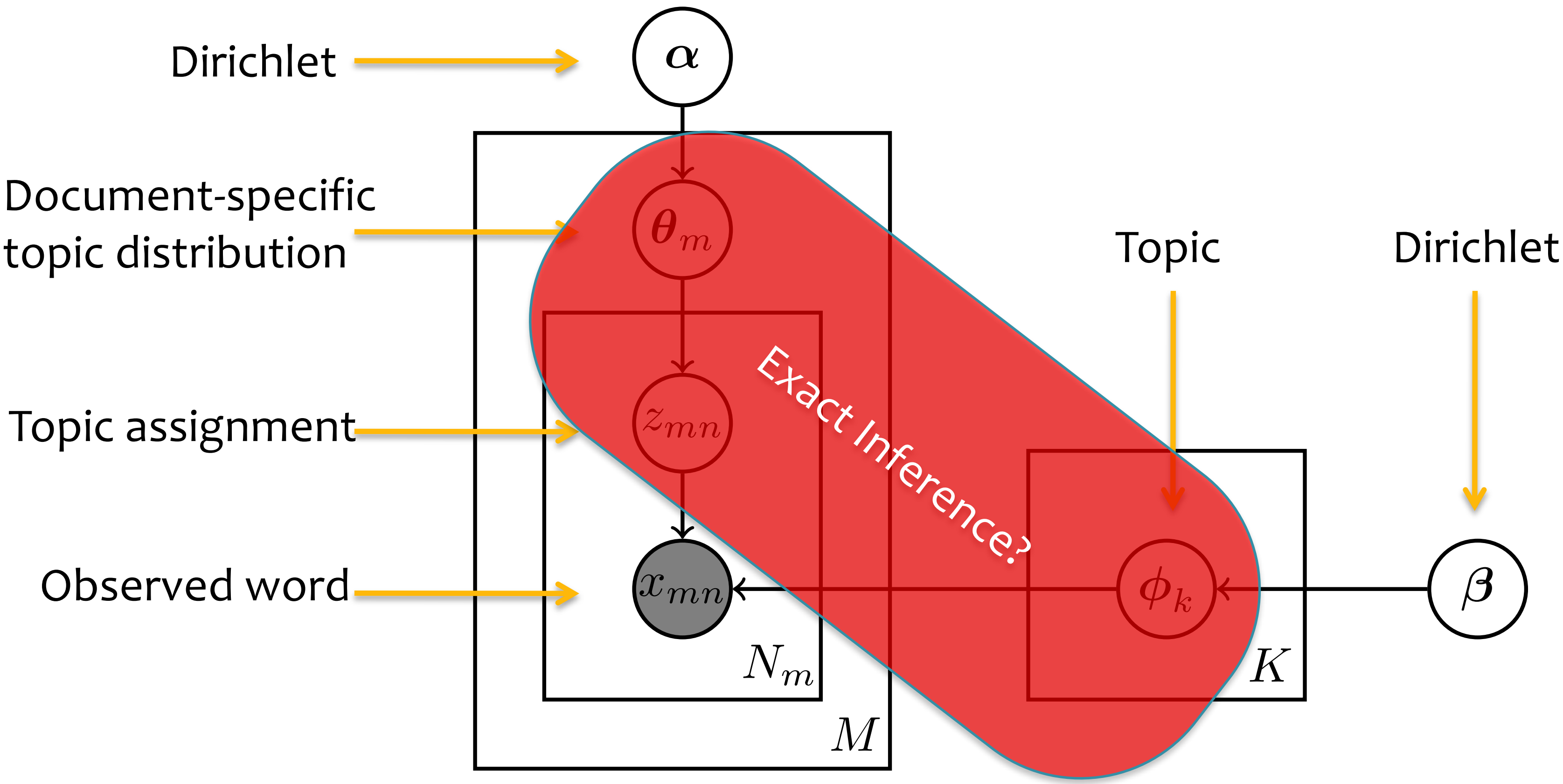
LDA Inference

- Monte Carlo EM (MAP Estimation)



LDA Inference

- Bayesian Approach



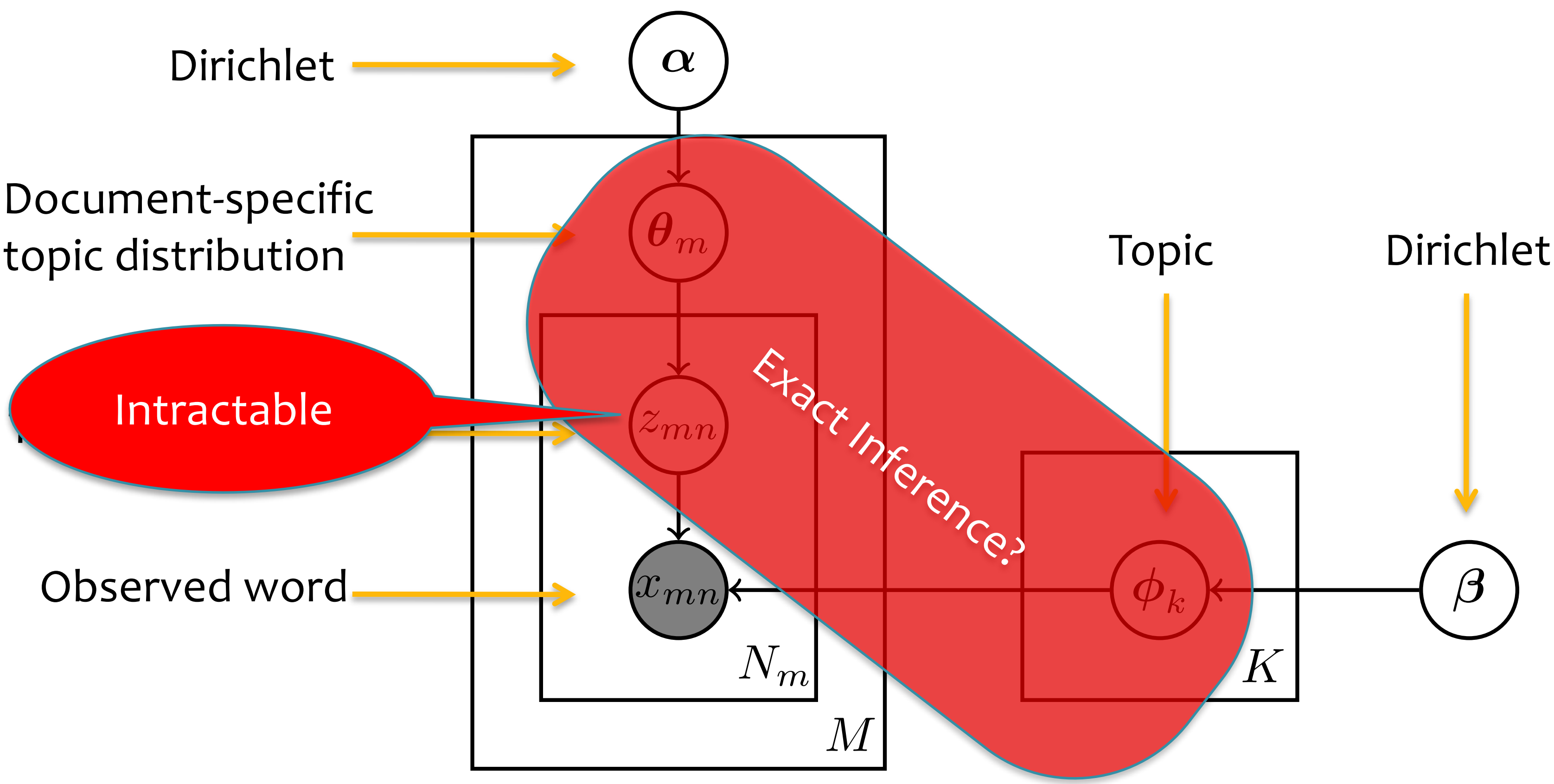
Bayesian Inference

***Whiteboard*:**

- Posteriors over parameters
- Bayesian inference for parameter estimation

LDA Inference

- Bayesian Approach

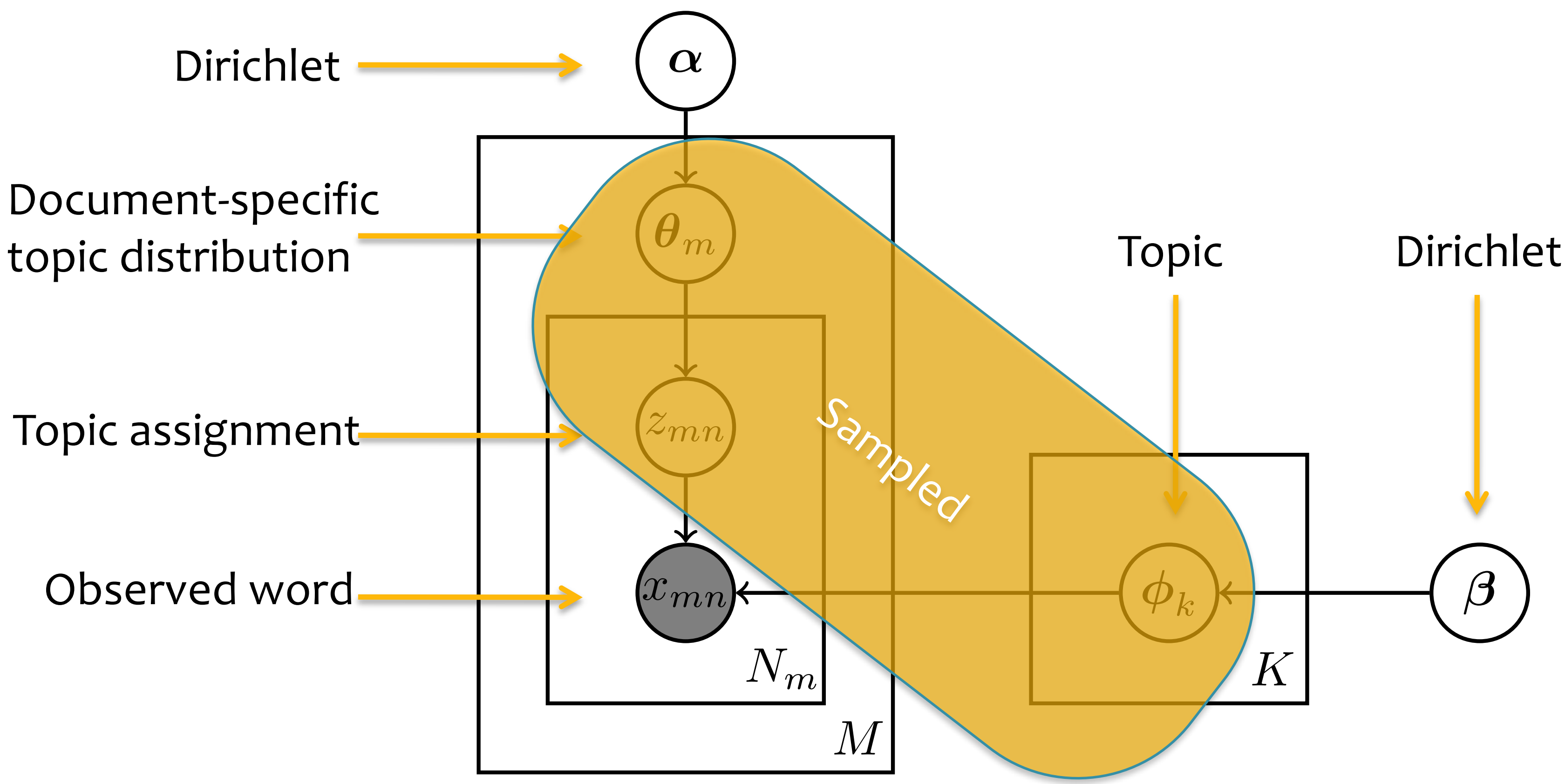


Exact Inference in LDA

- 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



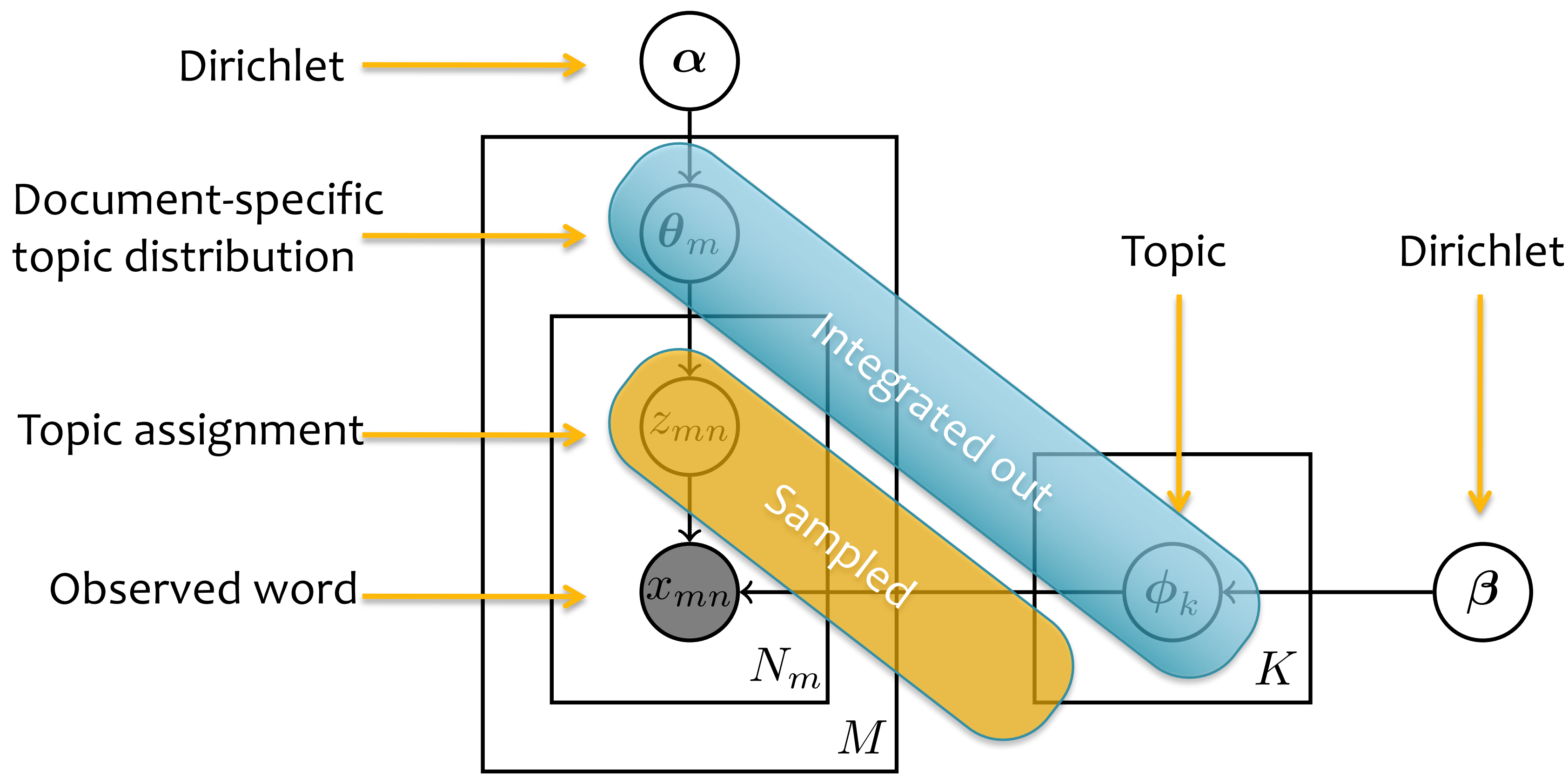
LDA Inference

Whiteboard:

– Explicit Gibbs Sampler for LDA

LDA Inference

- Collapsed Gibbs Sampler



LDA Inference

Whiteboard:

- Collapsed Gibbs Sampler for LDA

COLLAPSED GIBBS SAMPLER FOR LDA

Collapsed Gibbs Sampler for LDA

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}

Collapsed Gibbs Sampler for LDA

- 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

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t},$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}.$$

Collapsed Gibbs Sampler 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

Collapsed Gibbs Sampler for LDA

Whiteboard:

- Efficient computation of count variables

Collapsed Gibbs Sampler 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

$$\begin{array}{|ll|}\hline \boldsymbol{\phi} \sim \mathrm{Dir}(\boldsymbol{\beta}) & [\textit{draw distribution over words}] \\ \text{For each word } n \in \{1,\ldots,N\} & \\ \quad x_n \sim \mathrm{Mult}(1, \boldsymbol{\phi}) & [\textit{draw word}] \\ \hline \end{array}$$

- 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: $p(\phi|X) \sim \mathrm{Dir}(\boldsymbol{\beta} + \boldsymbol{n})$

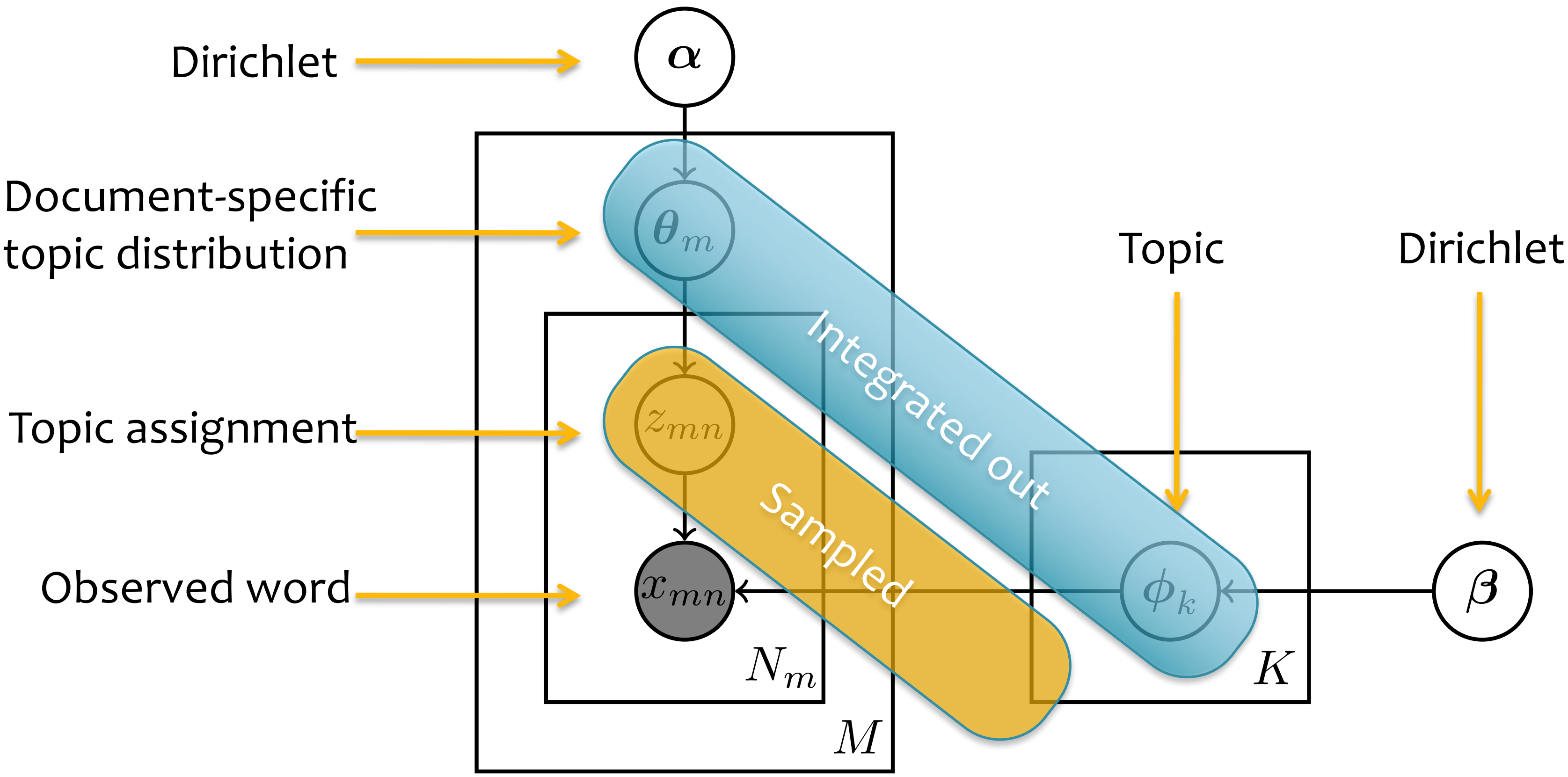
Dirichlet-Multinomial Model

- 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



Collapsed Gibbs Sampler for LDA

Algorithm

```
// initialisation
```

zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$

for all documents $m \in [1, M]$ **do**

  for all words $n \in [1, N_m]$ in document m **do**

    sample topic index $z_{m,n}=k \sim \mathrm{Mult}(1/K)$

    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

Collapsed Gibbs Sampler 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: $n_m^{(\tilde{k})} \mathrel{+}= 1; n_m \mathrel{+}= 1; n_{\tilde{k}}^{(t)} \mathrel{+}= 1; n_{\tilde{k}} \mathrel{+}= 1$

Collapsed Gibbs Sampler for LDA

Whiteboard:

– Q: How to recover parameter estimates from the collapsed Gibbs sampler?

– Dirichlet distribution over parameters

– Expected values of the parameters

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