



# Bayesian Inference for Parameter Estimation + Topic Modeling

Matt Gormley  
Lecture 14  
Oct. 24, 2022

# Reminders

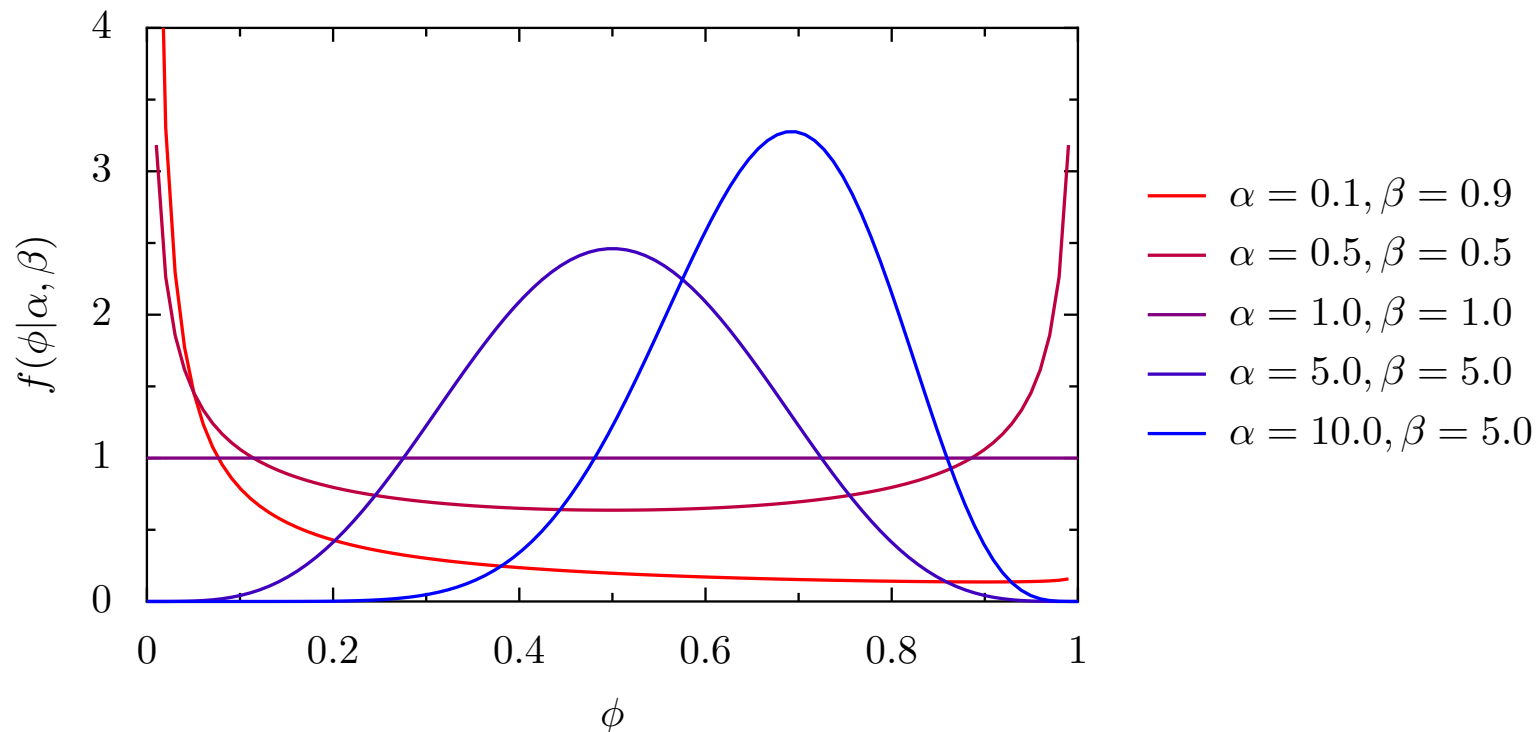
- Poll Questions 0a, 0b, 0c
- Grade Summary 1
- Homework 4: MCMC
  - Out: Mon, Oct 24
  - Due: Fri, Nov 3 at 11:59pm
- Recitation: Homework 4
  - today! 6pm, GHC 6121

# **BAYESIAN INFERENCE FOR NAÏVE BAYES**

# 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

|                                        |                                       |
|----------------------------------------|---------------------------------------|
| $\phi \sim \text{Beta}(\alpha, \beta)$ | <i>[draw distribution over words]</i> |
| For each word $n \in \{1, \dots, N\}$  |                                       |
| $x_n \sim \text{Bernoulli}(\phi)$      | <i>[draw word]</i>                    |

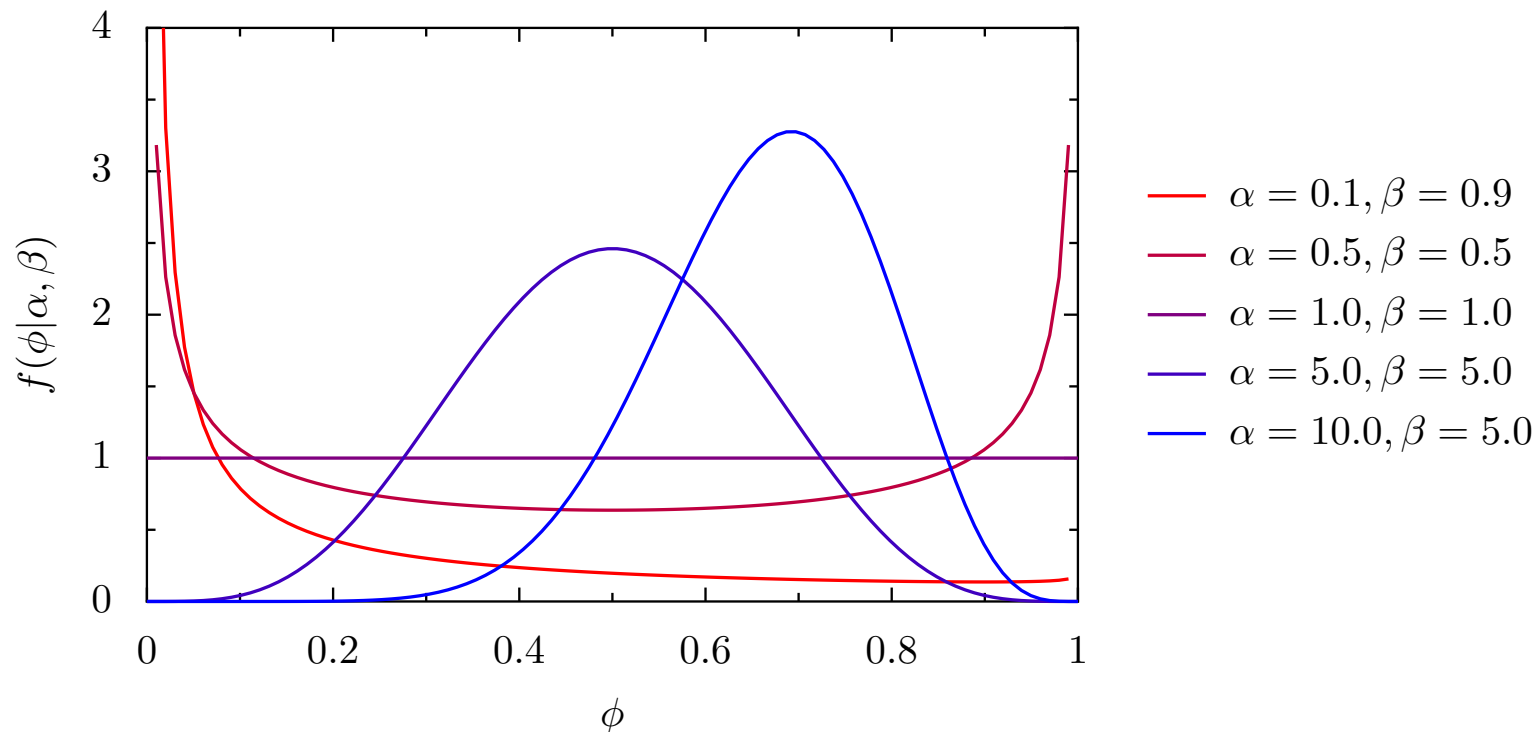
- Example corpus (heads/tails)

|       |       |       |       |       |       |       |       |       |          |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| 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

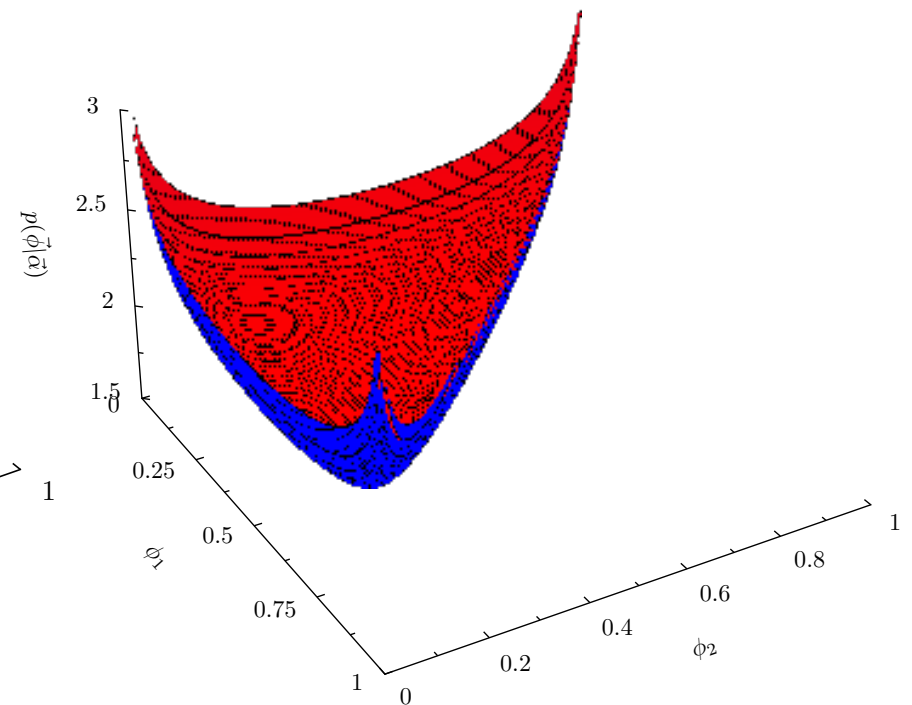
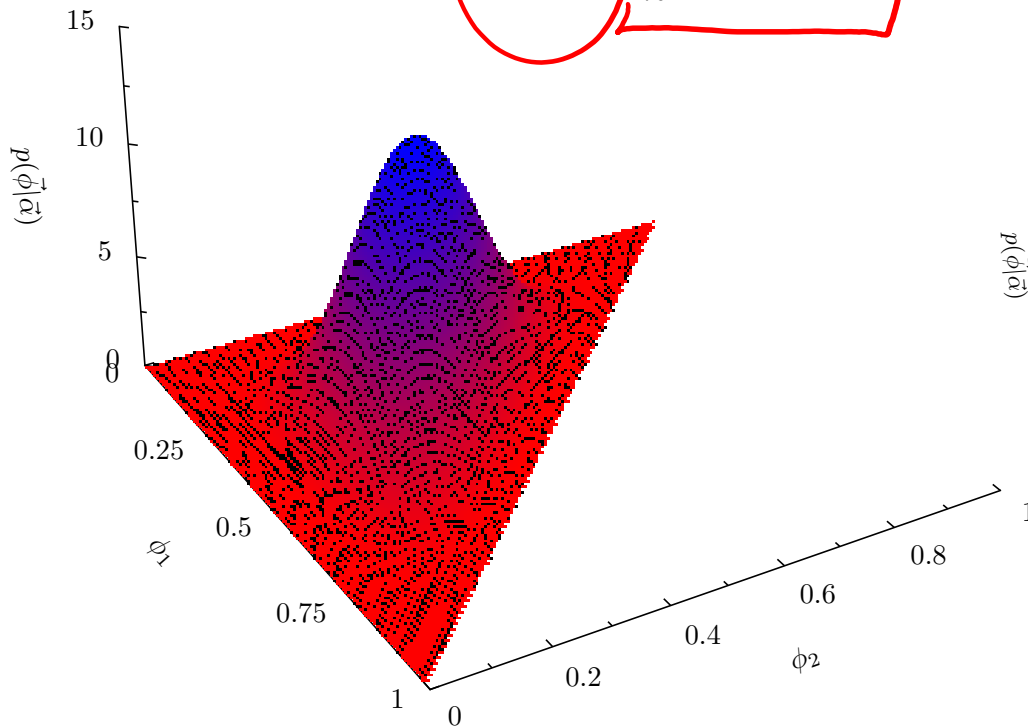
$$f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$



# Dirichlet-Multinomial Model

- Dirichlet Distribution

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



# Dirichlet-Multinomial Model

- Generative Process

$$\phi \sim \text{Dir}(\beta)$$

*[draw distribution over words]*

For each word  $n \in \{1, \dots, N\}$

$$x_n \sim \text{Mult}(1, \phi)$$

*[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 Model

The Dirichlet is **conjugate** to the Multinomial

likelihood prior

$$\phi \sim \text{Dir}(\beta)$$

[draw distribution over words]

For each word  $n \in \{1, \dots, N\}$

$$x_n \sim \text{Mult}(1, \phi)$$

[draw word]

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

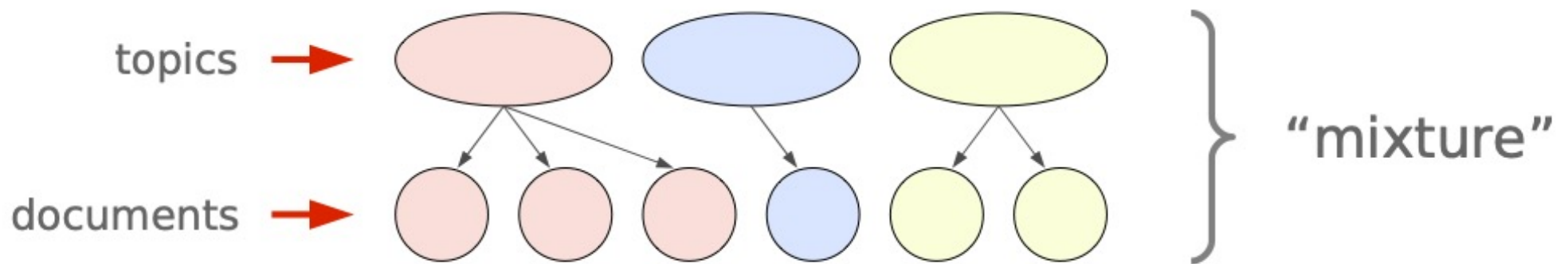
$$\begin{aligned} p(\vec{\phi} | \vec{x}, \vec{\beta}) &= p(\vec{x} | \vec{\phi}) p(\vec{\phi}) \\ &= \left[ \prod_{i=1}^N p(x^{(i)} | \vec{\phi}) \right] p(\vec{\phi}) \\ &= \left[ \prod_{k=1}^K \phi_k^{\beta_k} \right] \left[ \prod_{i=1}^N \prod_{k=1}^K \phi_k^{\mathbb{1}(x^{(i)}=k)} \right] \\ &= \prod_{k=1}^K \phi_k^{\beta_k - 1 + \sum_{i=1}^N \mathbb{1}(x^{(i)}=k)} \end{aligned}$$

$$\Rightarrow p(\vec{\phi} | \vec{x}, \vec{\beta}) \sim \text{Dirichlet}(\vec{\beta} + \vec{n})$$

where  $n_k = \# \text{ times } x^{(i)} = k$

# 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

$$\Phi_{ML} = \begin{bmatrix} 0.01 & 0.02 & 0.04 & 0.01 & 0.0 \dots \end{bmatrix}$$

Bayes      opt.      the      is      zebra...

## • Generative Process

For each topic  $k \in \{1, \dots, K\}$ :

$$\vec{\phi}_k \sim \text{Dir}(\beta)$$

→ [draw distribution over words]

$$\theta \sim \text{Dir}(\alpha)$$

[draw distribution over topics]

For each document  $m \in \{1, \dots, M\}$

$$\rightarrow z_m \sim \text{Mult}(1, \theta)$$

[draw topic assignment]

For each word  $n \in \{1, \dots, N_m\}$

$$x_{mn} \sim \text{Mult}(1, \phi_{z_m})$$

[draw word]

## • Example corpus

$$\Phi_{Bio} = \begin{bmatrix} 0.0 & 0.001 & 0.04 & 0.01 & 0.05 \dots \end{bmatrix}$$

Bayes      opt.      the      is      zebra...

$z_1 = ML$

$z_2 = Bio$

$z_3 = ML$

Bayes      opt.      is

|          |          |          |
|----------|----------|----------|
| the      | opt.     | is       |
| $x_{11}$ | $x_{12}$ | $x_{13}$ |

gene      the      RNA

|          |          |          |
|----------|----------|----------|
| gene     | the      | RNA      |
| $x_{21}$ | $x_{22}$ | $x_{23}$ |

she      is      turn      is ML

|          |          |          |          |
|----------|----------|----------|----------|
| she      | is       | turn     | is       |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ |

Document 1

Document 2

Document 3

# Bayesian Inference

- The **key idea** behind **Bayesian Inference** is to treat your parameters as though they are like any other variables in your graphical model
- Bayesian Inference Summary:
  1. **Given:** data,  $D$
  2. **Goal:** learn the posterior distribution over parameters  $\theta$  given data  $D$ , i.e.  $p(\theta | D)$
  3. **Store:** a distribution  $p(\theta | D)$  as a probability mass function (pmf) or probability density function (pdf) or via some approximation
  4. **Afterwards:** marginalize over the parameters to work directly with other latent variables, e.g.  
$$p(z | D) = \int_{\theta} p(z | \theta, D) p(\theta | D) d\theta$$

$$p(\theta, z | D)$$



# Naïve Bayes Model

## Data:

$$\mathcal{D} = \{(z^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^N$$

where  $z^{(i)} \in \{1, \dots, L\}$ ,

$$\mathbf{x}^{(i)} \in \{1, \dots, V\}^M$$

## Generative Story:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\phi_k \sim \text{Dirichlet}(\boldsymbol{\beta}), \forall k$$

$$z^{(i)} \sim \text{Categorical}(\boldsymbol{\theta}), \forall i$$

$$x_m^{(i)} \sim \text{Categorical}(\phi_{z^{(i)}}), \forall i, \forall m$$

## 1) MLE:

$$\hat{\boldsymbol{\theta}}, \hat{\phi} = \underset{\boldsymbol{\theta}, \phi}{\operatorname{argmax}} p(\mathcal{D} \mid \boldsymbol{\theta}, \phi)$$

## 2) MAP Estimation:

$$\begin{aligned} \hat{\boldsymbol{\theta}}, \hat{\phi} &= \underset{\boldsymbol{\theta}, \phi}{\operatorname{argmax}} p(\mathcal{D} \mid \boldsymbol{\theta}, \phi) p(\boldsymbol{\theta}, \phi) \\ &= \underset{\boldsymbol{\theta}, \phi}{\operatorname{argmax}} p(\boldsymbol{\theta}, \phi \mid \mathcal{D}) \end{aligned}$$

## 3) Bayesian Parameter Estimation:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \text{Dirichlet}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}')$$

$$p(\phi_k \mid \mathcal{D}) = \text{Dirichlet}(\phi_k \mid \boldsymbol{\beta}'_k)$$

# Naïve Bayes Model

The standard presentation of Naïve Bayes (i.e. MLE and MAP Estimation) is not Bayesian!



## 1) MLE:

$$\hat{\theta}, \hat{\phi} = \operatorname{argmax}_{\theta, \phi} p(\mathcal{D} \mid \theta, \phi)$$

## 2) MAP Estimation:

$$\begin{aligned} \hat{\theta}, \hat{\phi} &= \operatorname{argmax}_{\theta, \phi} p(\mathcal{D} \mid \theta, \phi) p(\theta, \phi) \\ &= \operatorname{argmax}_{\theta, \phi} p(\theta, \phi \mid \mathcal{D}) \end{aligned}$$

## 3) Bayesian Parameter Estimation:

$$\begin{aligned} p(\theta \mid \mathcal{D}) &= \text{Dirichlet}(\theta \mid \alpha') \\ p(\phi_k \mid \mathcal{D}) &= \text{Dirichlet}(\phi_k \mid \beta'_k) \end{aligned}$$

# Naïve Bayes Model

## Generative Story:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta}), \forall k$$

$$z^{(i)} \sim \text{Categorical}(\boldsymbol{\theta}), \forall i$$

$$x_m^{(i)} \sim \text{Categorical}(\boldsymbol{\phi}_{z^{(i)}}), \forall i, \forall m$$

## 1) MLE:

$$\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\theta}, \boldsymbol{\phi}}{\operatorname{argmax}} p(\mathcal{D} \mid \boldsymbol{\theta}, \boldsymbol{\phi})$$

- pro: solved by counting

$$\theta_k \propto \sum_{i=1}^N \mathbb{1}(z^{(i)} = k)$$

$$\phi_{kv} \propto \sum_{i=1}^N \sum_{m=1}^M \mathbb{1}(x_m^{(i)} = v) \mathbb{1}(z^{(i)} = k)$$

- con: single point estimate
- con: ignores the priors over parameters

## 2) MAP Estimation:

$$\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\theta}, \boldsymbol{\phi}}{\operatorname{argmax}} p(\mathcal{D} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}, \boldsymbol{\phi})$$

$$= \underset{\boldsymbol{\theta}, \boldsymbol{\phi}}{\operatorname{argmax}} p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathcal{D})$$

- pro: takes prior into account
- pro: solved by counting

$$\theta_k \propto (\alpha - 1) + \sum_{i=1}^N \mathbb{1}(z^{(i)} = k)$$

$$\phi_{kv} \propto (\beta - 1) + \sum_{i=1}^N \sum_{m=1}^M \mathbb{1}(x_m^{(i)} = v) \mathbb{1}(z^{(i)} = k)$$

- con: single point estimate

# Naïve Bayes Model

## 1) MLE:

$$\hat{\theta}, \hat{\phi} = \operatorname{argmax}_{\theta, \phi} p(\mathcal{D} \mid \theta, \phi)$$

## 2) MAP Estimation:

$$\hat{\theta}, \hat{\phi} = \operatorname{argmax}_{\theta, \phi} p(\mathcal{D} \mid \theta, \phi) p(\theta, \phi)$$

$$= \operatorname{argmax}_{\theta, \phi} p(\theta, \phi \mid \mathcal{D})$$

- pro: takes prior into account
- pro: solved by counting

$$\theta_k \propto (\alpha - 1) + \sum_{i=1}^N \mathbb{1}(z^{(i)} = k)$$

$$\phi_{kv} \propto (\beta - 1) + \sum_{i=1}^N \sum_{m=1}^M \mathbb{1}(x_m^{(i)} = v) \mathbb{1}(z^{(i)} = k)$$

- con: single point estimate

## 3) Bayesian Parameter Estimation:

$$p(\underline{\theta} \mid \mathcal{D}) = \text{Dirichlet}(\theta \mid \underline{\alpha}')$$

$$p(\phi_k \mid \mathcal{D}) = \text{Dirichlet}(\phi_k \mid \beta'_k)$$

- pro: takes uncertainty over parameters into account
- pro: compactly represented b/c of Dirichlet-Multinomial conjugacy

$$\left[ \begin{aligned} \alpha'_k &= \alpha_k + \sum_{i=1}^N \mathbb{1}(z^{(i)} = k) \\ \beta'_{kv} &= \beta_k + \sum_{i=1}^N \sum_{m=1}^M \mathbb{1}(x_m^{(i)} = v) \mathbb{1}(z^{(i)} = k) \end{aligned} \right.$$

# Naïve Bayes Model

## 2) MAP Estimation:

$$\begin{aligned}\hat{\theta}, \hat{\phi} &= \operatorname{argmax}_{\theta, \phi} p(\mathcal{D} \mid \theta, \phi) p(\theta, \phi) \\ &= \operatorname{argmax}_{\theta, \phi} p(\theta, \phi \mid \mathcal{D})\end{aligned}$$



### Question:

Given a new point  $\mathbf{x}$  and point estimates of  $\theta$  and  $\phi$  how do we do inference over  $z$ ?

### Answer:

$$\begin{aligned}p(z=k \mid \vec{x}, \hat{\theta}, \hat{\phi}) &\propto \\ p(z=k \mid \hat{\theta}) \prod_{m=1}^M p(x_m \mid z=k, \hat{\phi})\end{aligned}$$

## 3) Bayesian Parameter Estimation:

$$p(\theta \mid \mathcal{D}) = \text{Dirichlet}(\theta \mid \alpha')$$

$$p(\phi_k \mid \mathcal{D}) = \text{Dirichlet}(\phi_k \mid \beta'_k)$$

### Question:

Given a new point  $\mathbf{x}$  and distributions  $p(\theta \mid \mathcal{D})$  and  $p(\phi \mid \mathcal{D})$  how do we do inference over  $z$ ?

### Answer:

$$p(z=k \mid \vec{x}, \alpha', \beta') = \int_{\theta} \int_{\phi} p(z=k \mid \vec{x}, \theta, \phi) p(\theta \mid \alpha') p(\phi \mid \beta') d\theta d\phi$$

use MCMC

# Plate Diagrams

## ***Whiteboard:***

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

# 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]



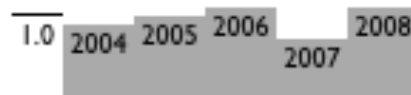
problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions

## 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]



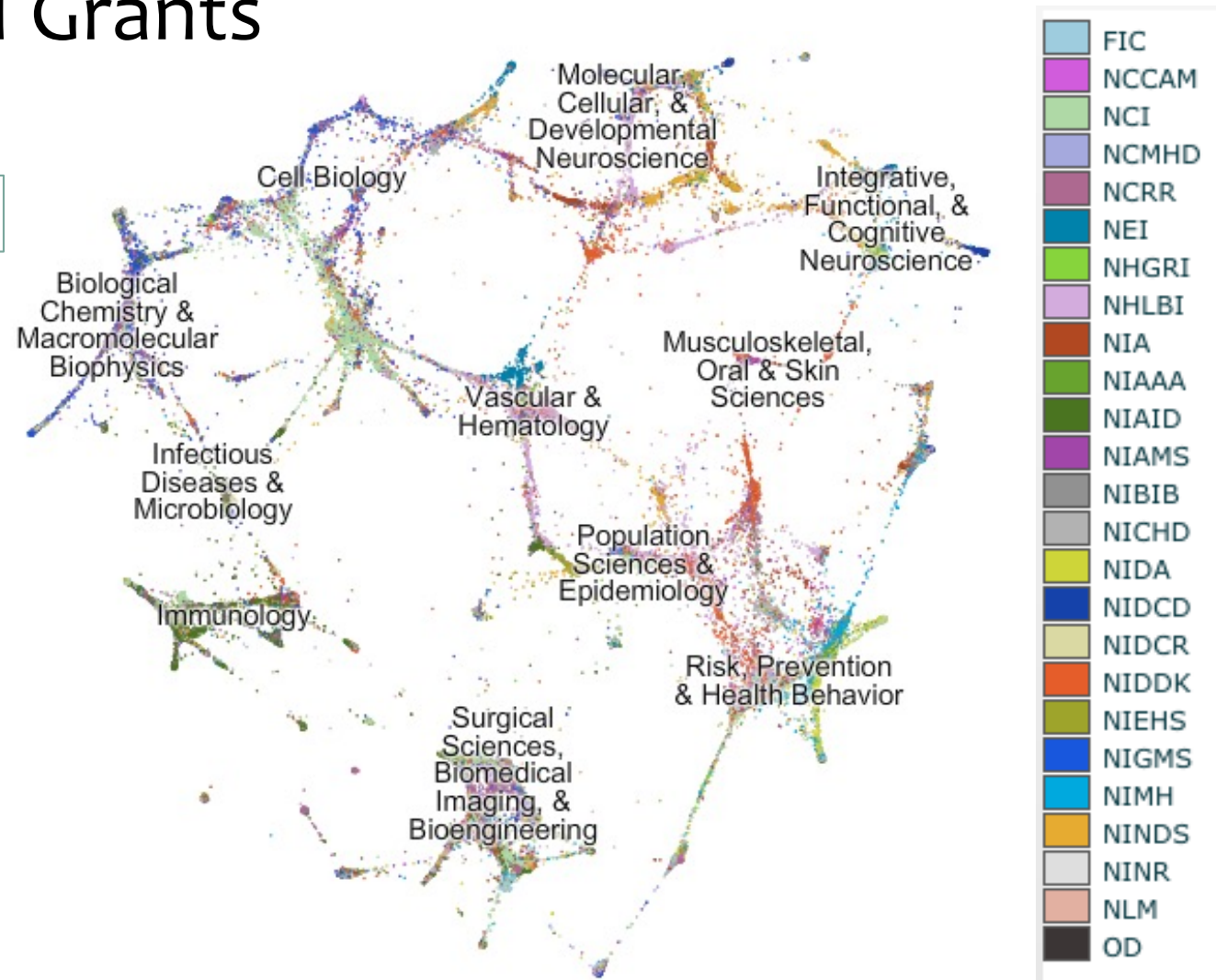
inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient

[http:// www.cs.umass.edu/~mimno/icml100.html](http://www.cs.umass.edu/~mimno/icml100.html)

# Topic Modeling

- Map of NIH Grants

(Talley et al., 2011)

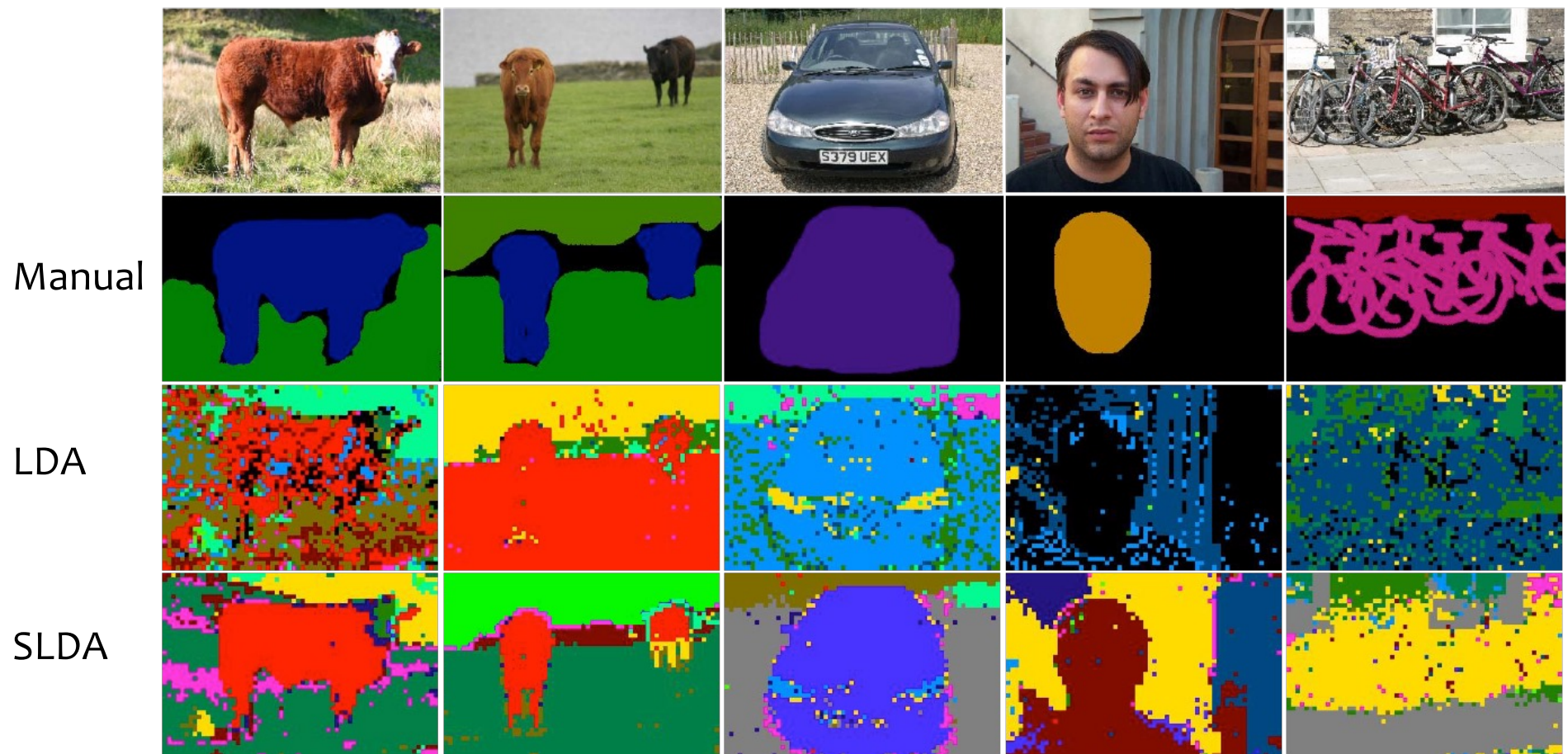


<https://app.nihmaps.org/>

# Other Applications of Topic Models

- Spatial LDA

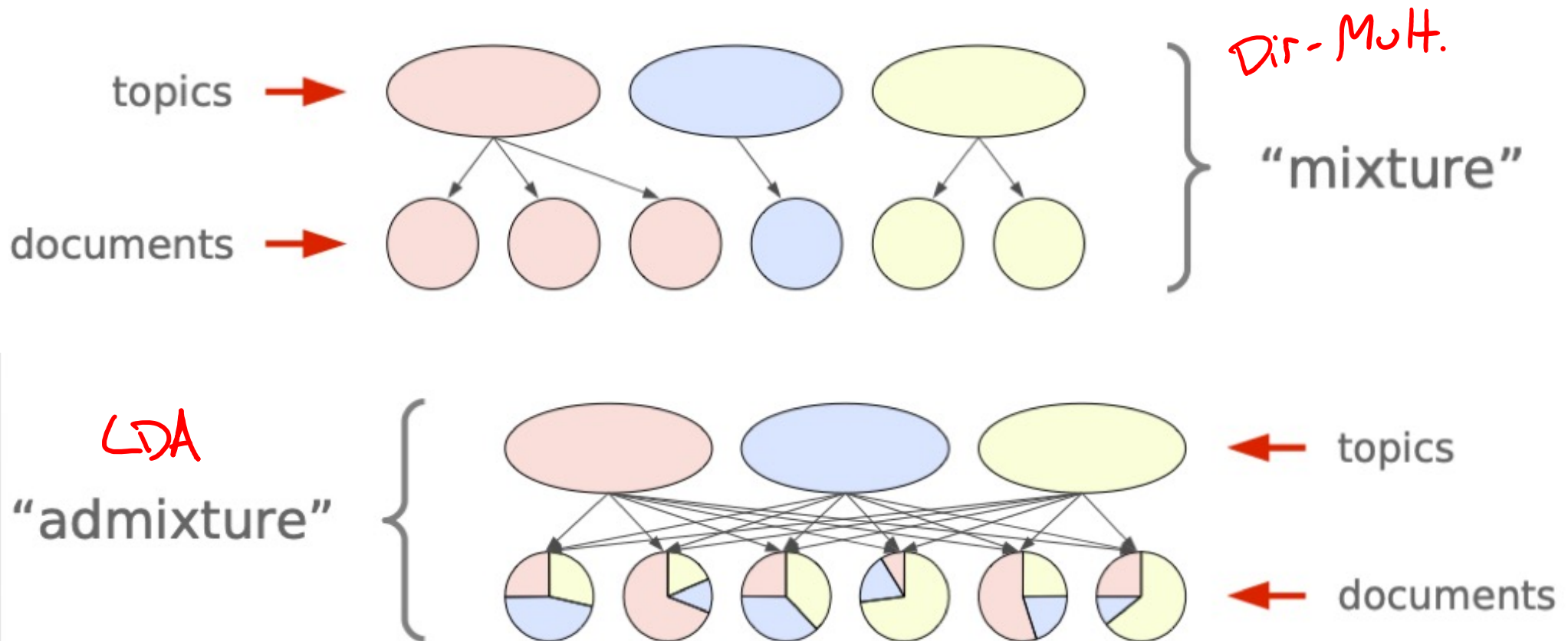
(Wang & Grimson, 2007)



# **LATENT DIRICHLET ALLOCATION (LDA)**

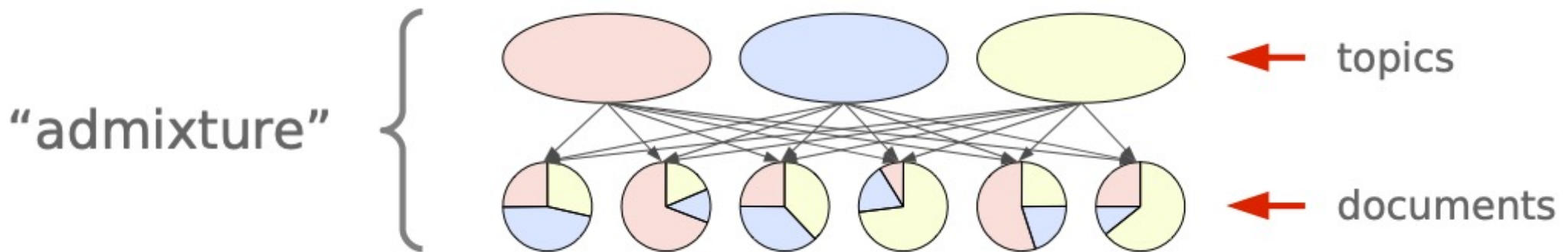


# Mixture vs. Admixture (LDA)



# 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  $k \in \{1, \dots, K\}$ :

$$\phi_k \sim \text{Dir}(\beta)$$

[draw distribution over words]

For each document  $m \in \{1, \dots, M\}$

$$\theta_m \sim \text{Dir}(\alpha)$$

[draw distribution over topics]

For each word  $n \in \{1, \dots, N_m\}$

$$z_{mn} \sim \text{Mult}(1, \theta_m)$$

[draw topic assignment]

$$x_{mn} \sim \phi_{z_{mn}}$$

[draw word]

## Example corpus

|          |          |          |
|----------|----------|----------|
| the      | he       | is       |
| $x_{11}$ | $x_{12}$ | $x_{13}$ |

Document 1

|          |          |          |
|----------|----------|----------|
| the      | and      | the      |
| $x_{21}$ | $x_{22}$ | $x_{23}$ |

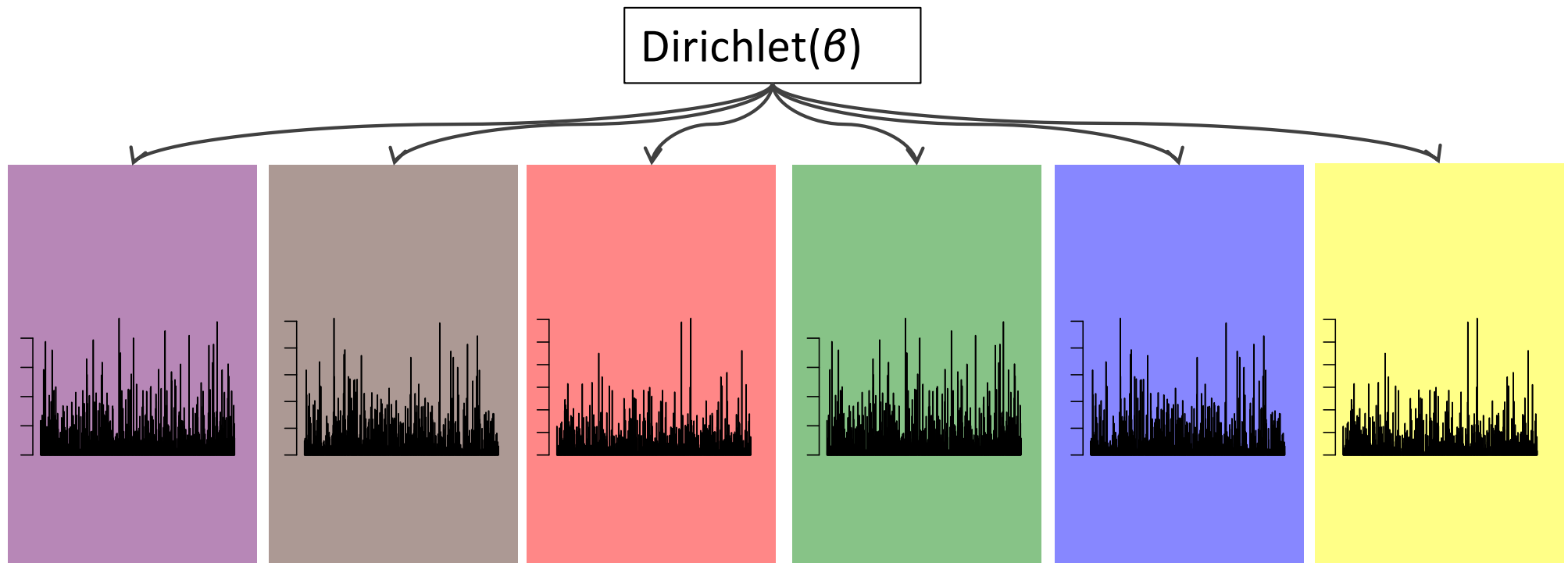
Document 2

|          |          |          |          |
|----------|----------|----------|----------|
| she      | she      | is       | is       |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ |

Document 3

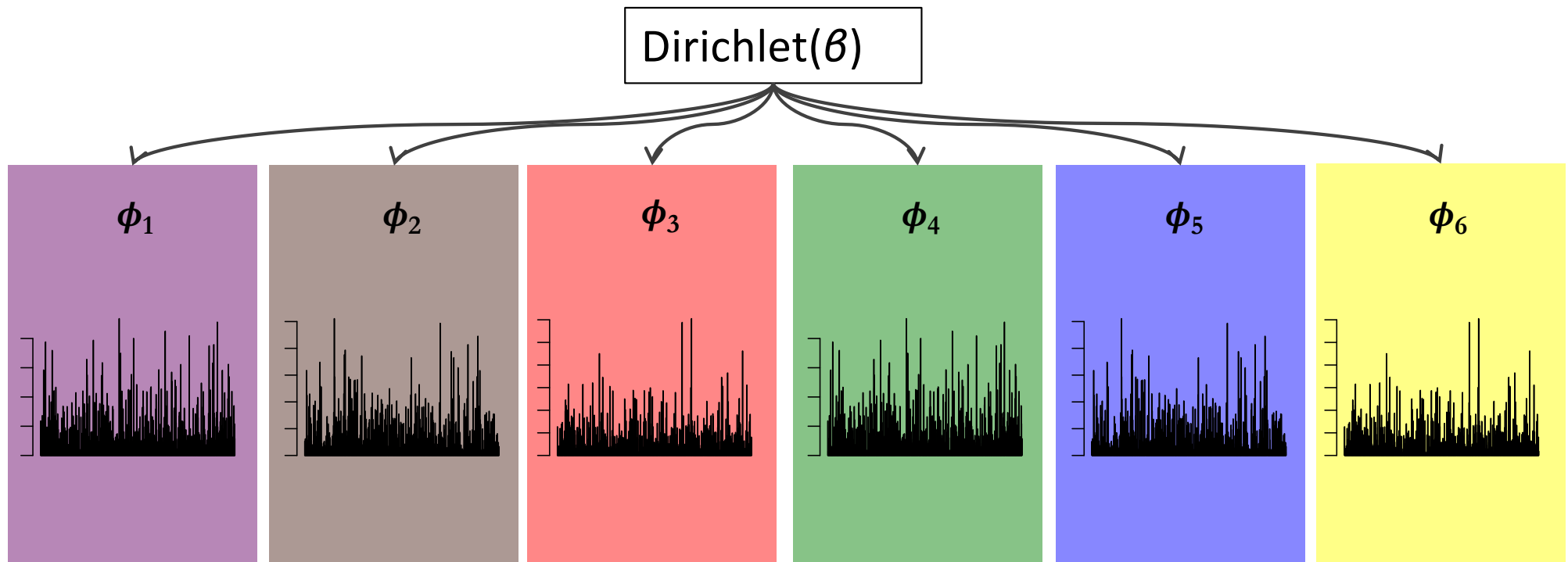


# LDA for Topic Modeling



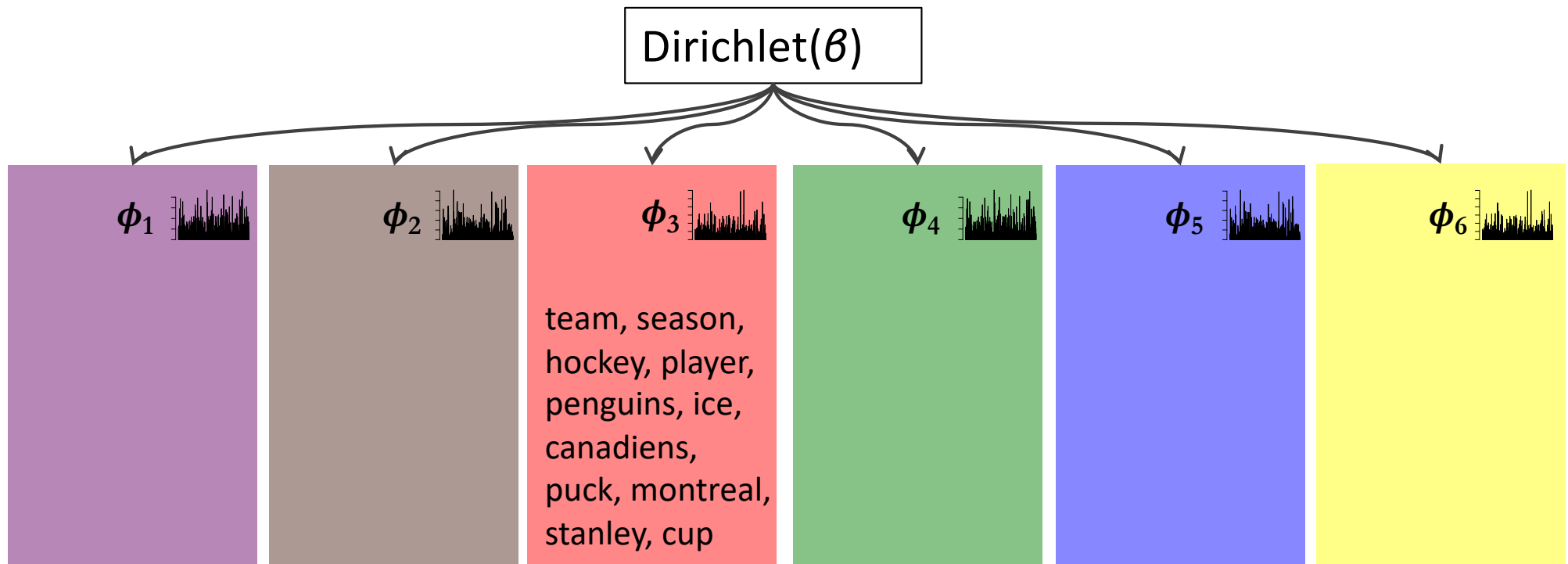
- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by  $\phi_k$

# LDA for Topic Modeling



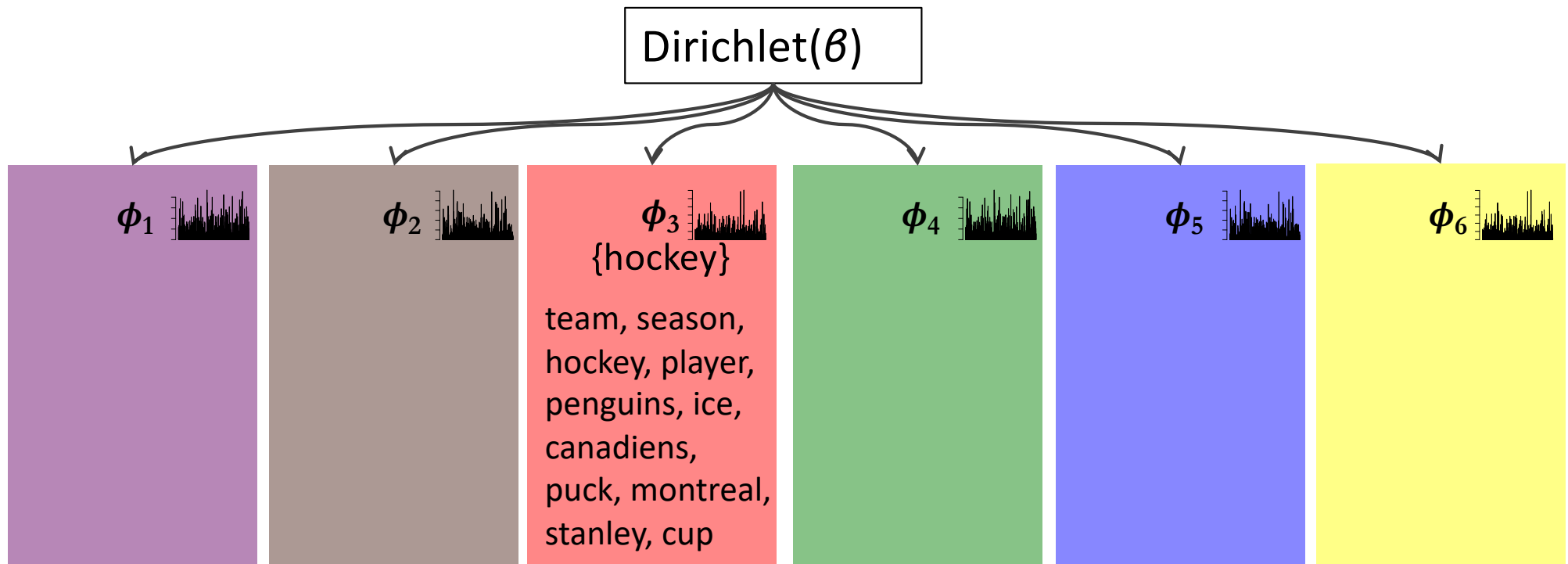
- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by  $\phi_k$

# LDA for Topic Modeling



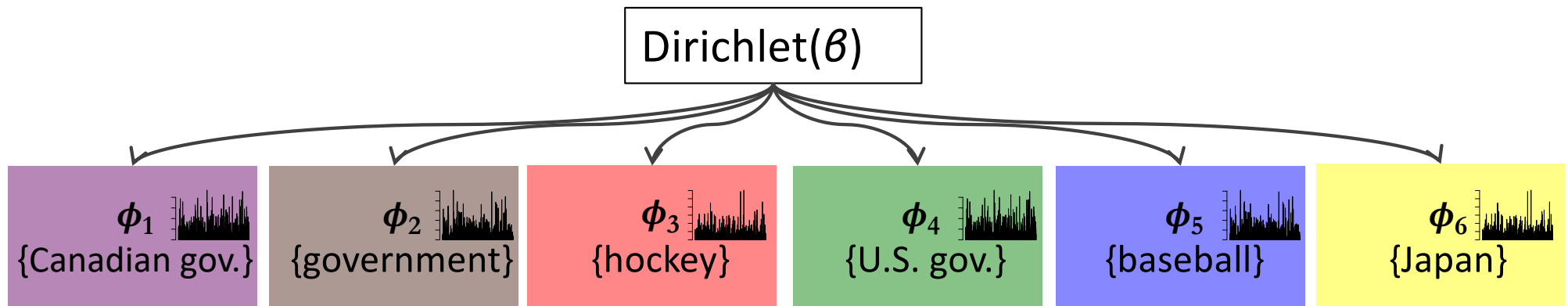
- A topic is visualized as its **high probability words**.

# LDA for Topic Modeling



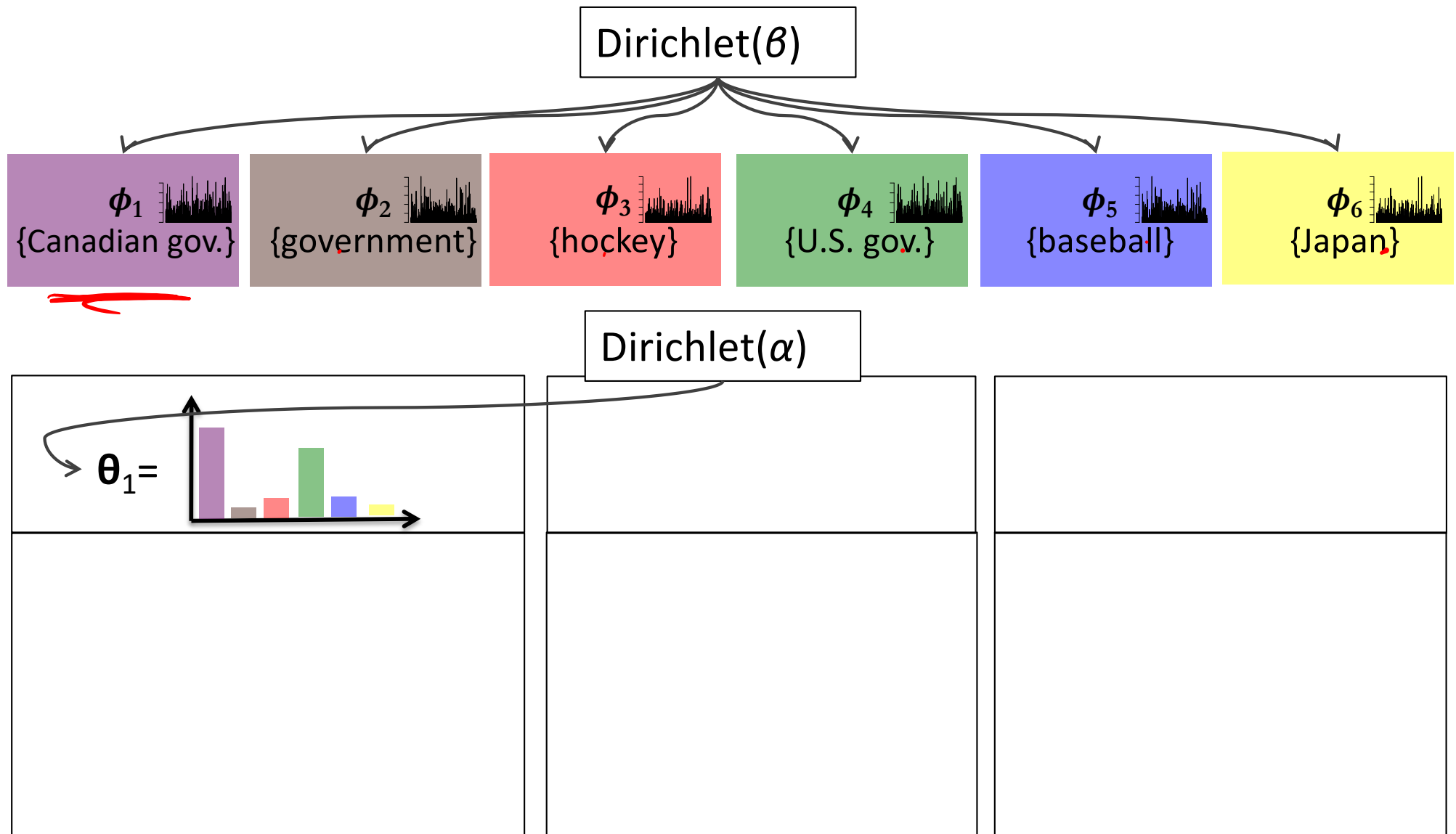
- A topic is visualized as its **high probability words**.
- A pedagogical **label** is used to identify the topic.

# LDA for Topic Modeling

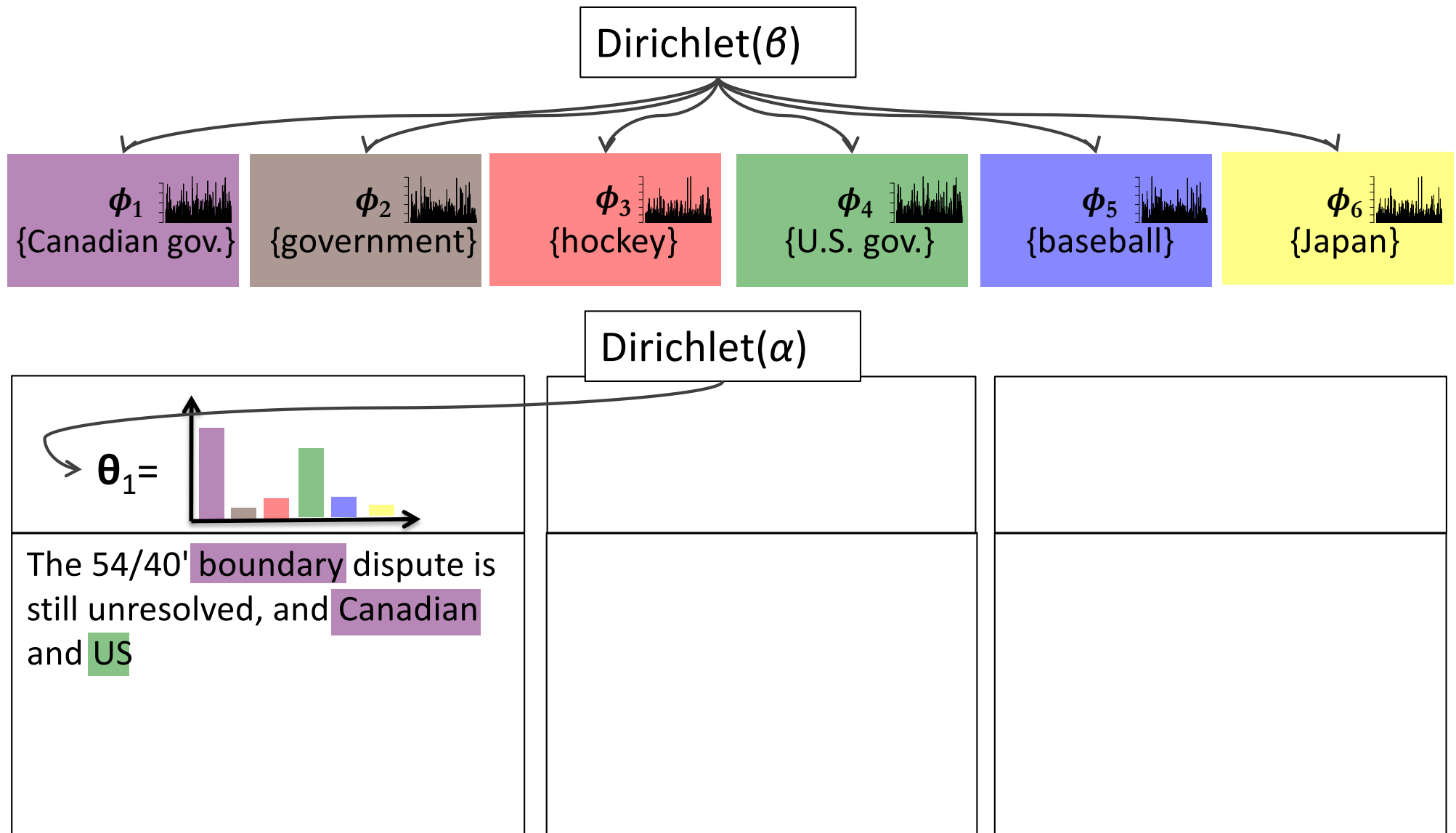


- A topic is visualized as its high probability words.
- A pedagogical **label** is used to identify the topic.

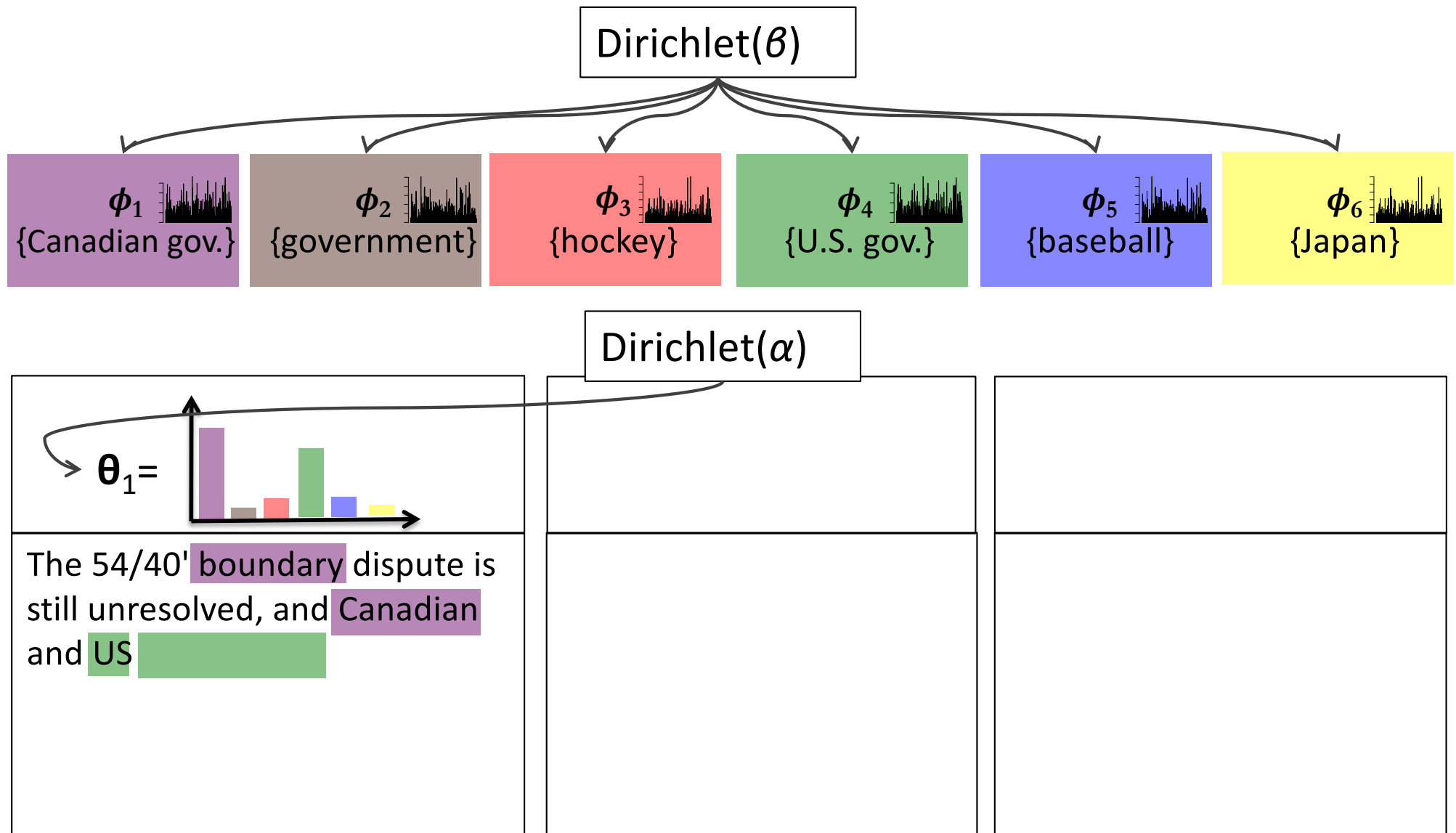
# LDA for Topic Modeling



# LDA for Topic Modeling

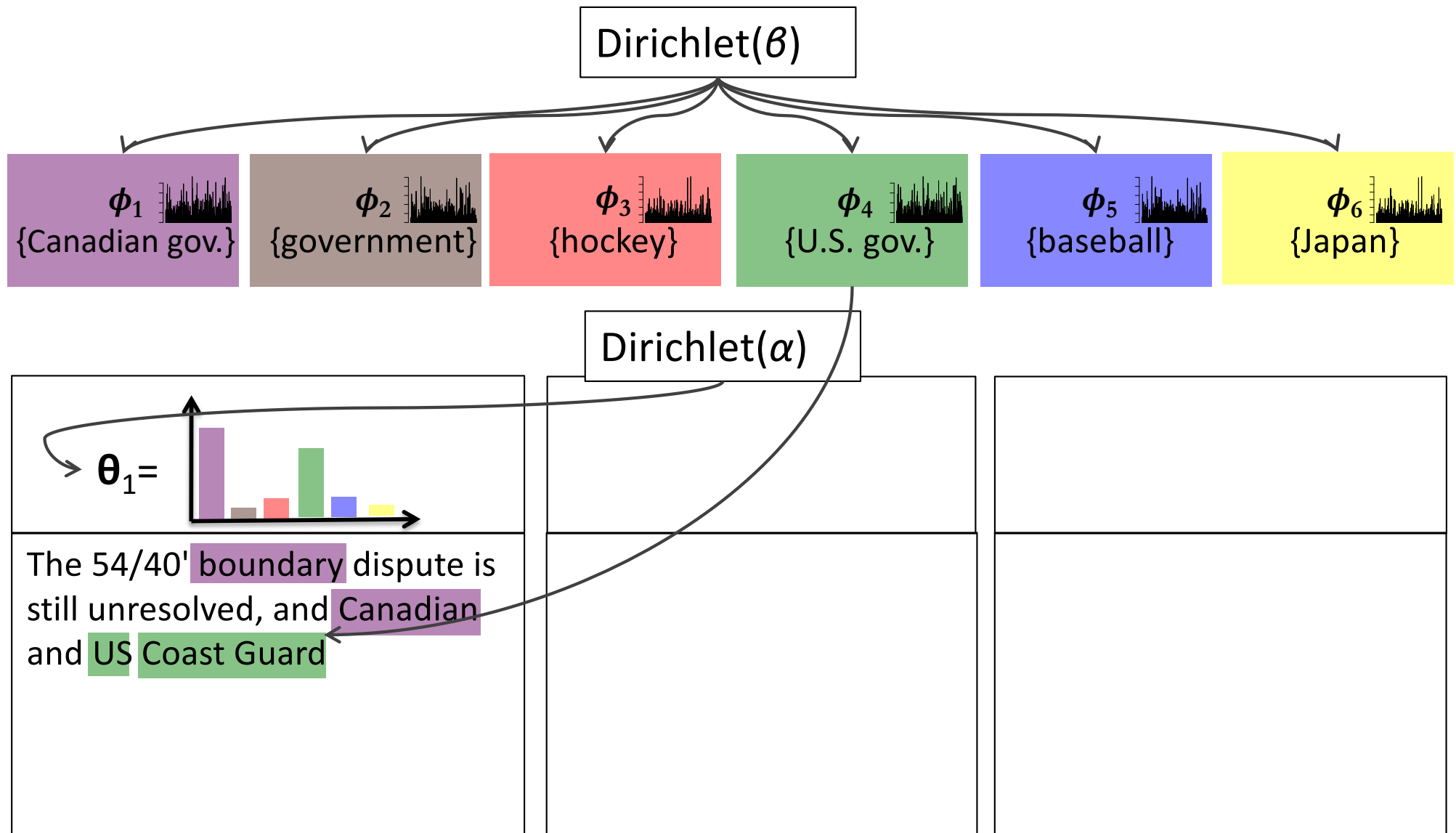


# LDA for Topic Modeling

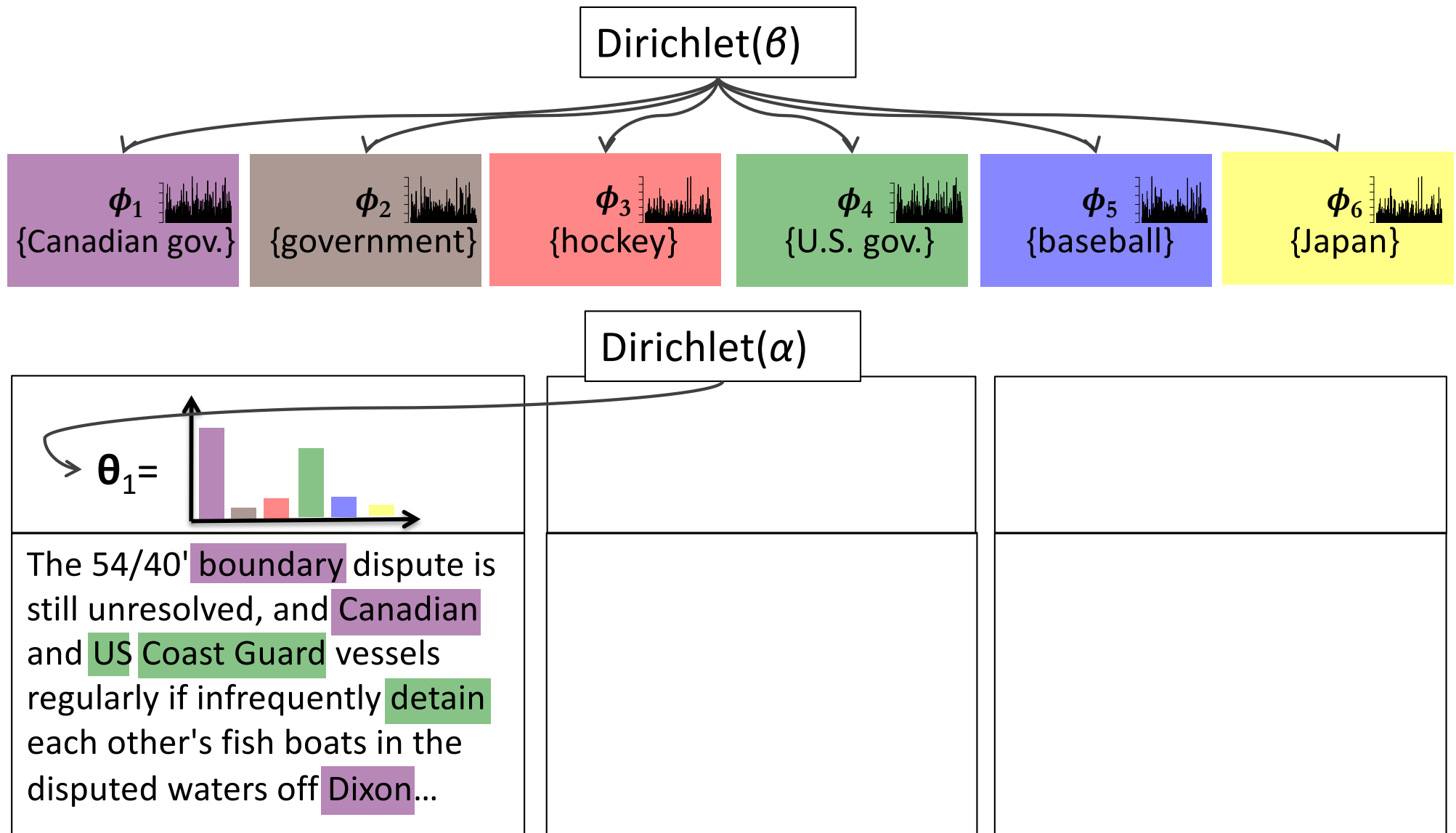




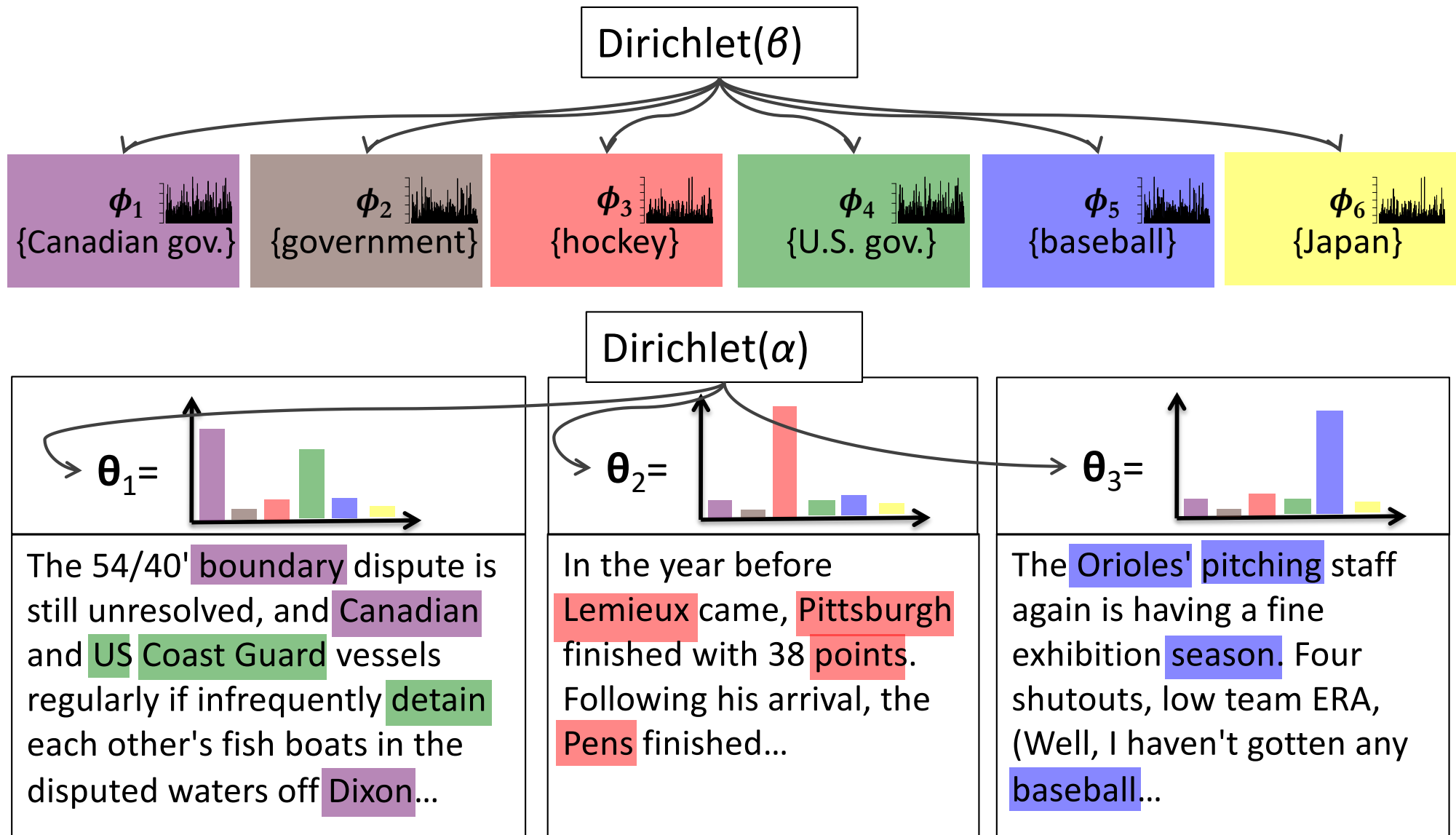
# LDA for Topic Modeling



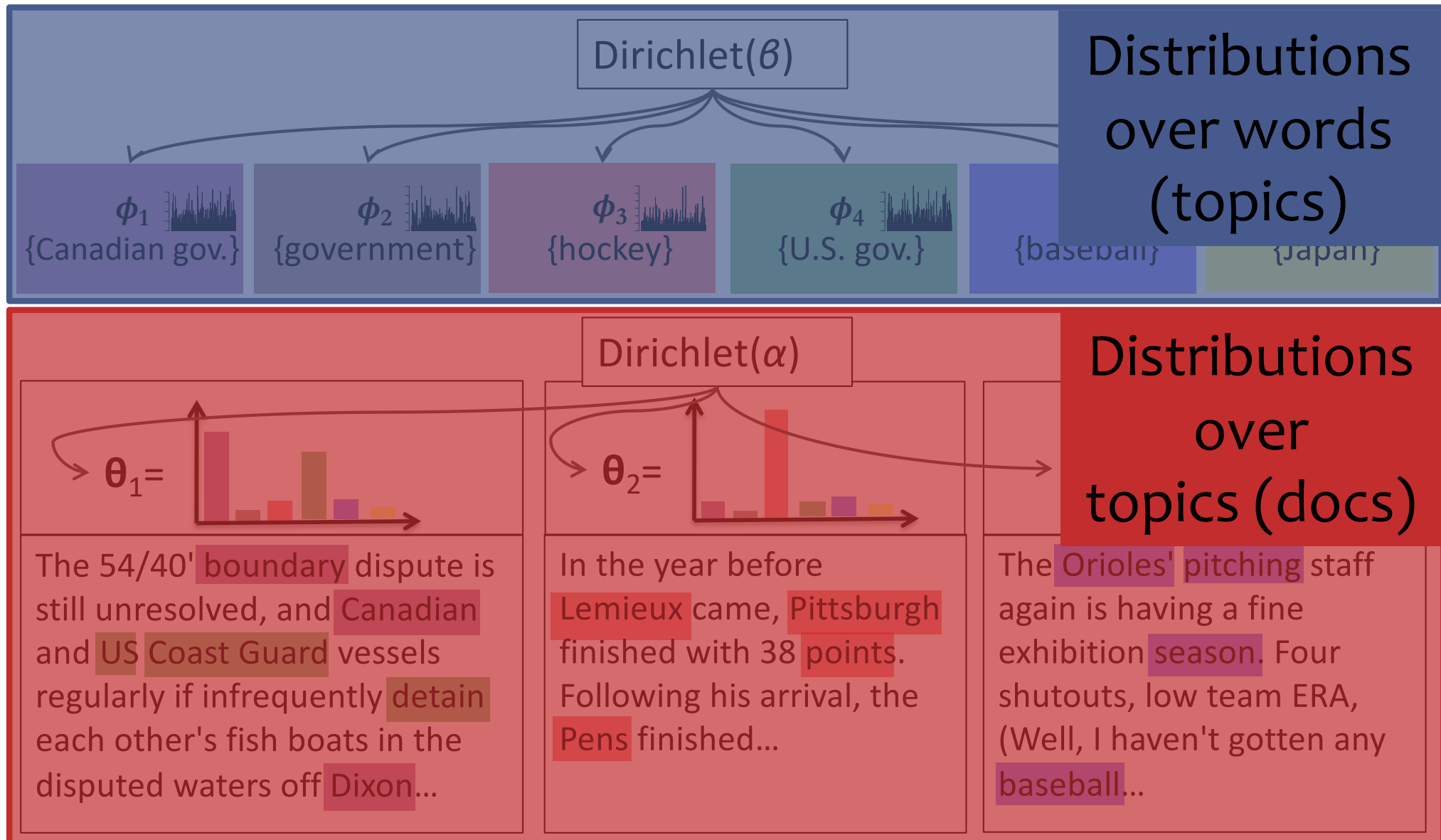
# LDA for Topic Modeling



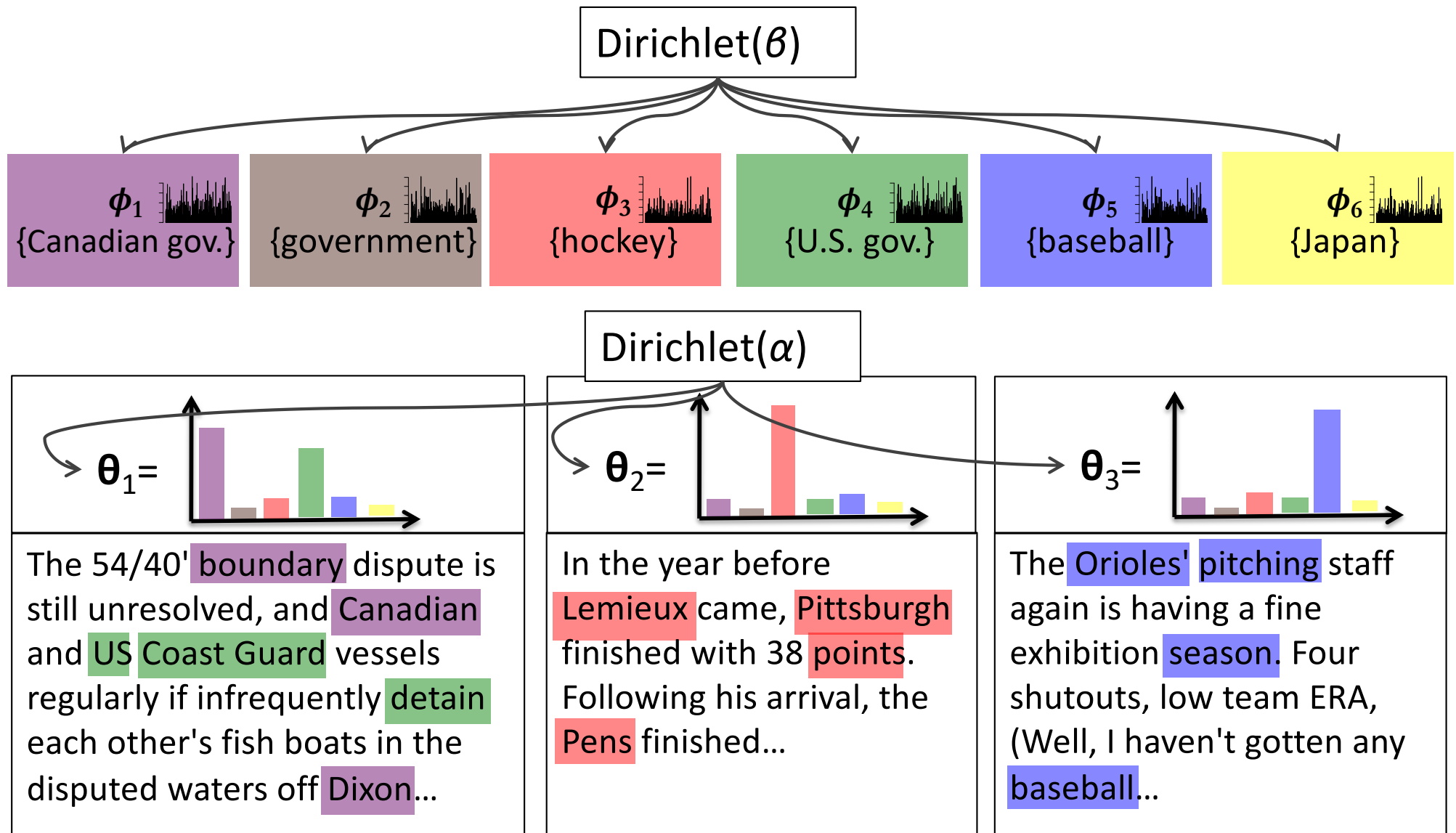
# LDA for Topic Modeling



# LDA for Topic Modeling

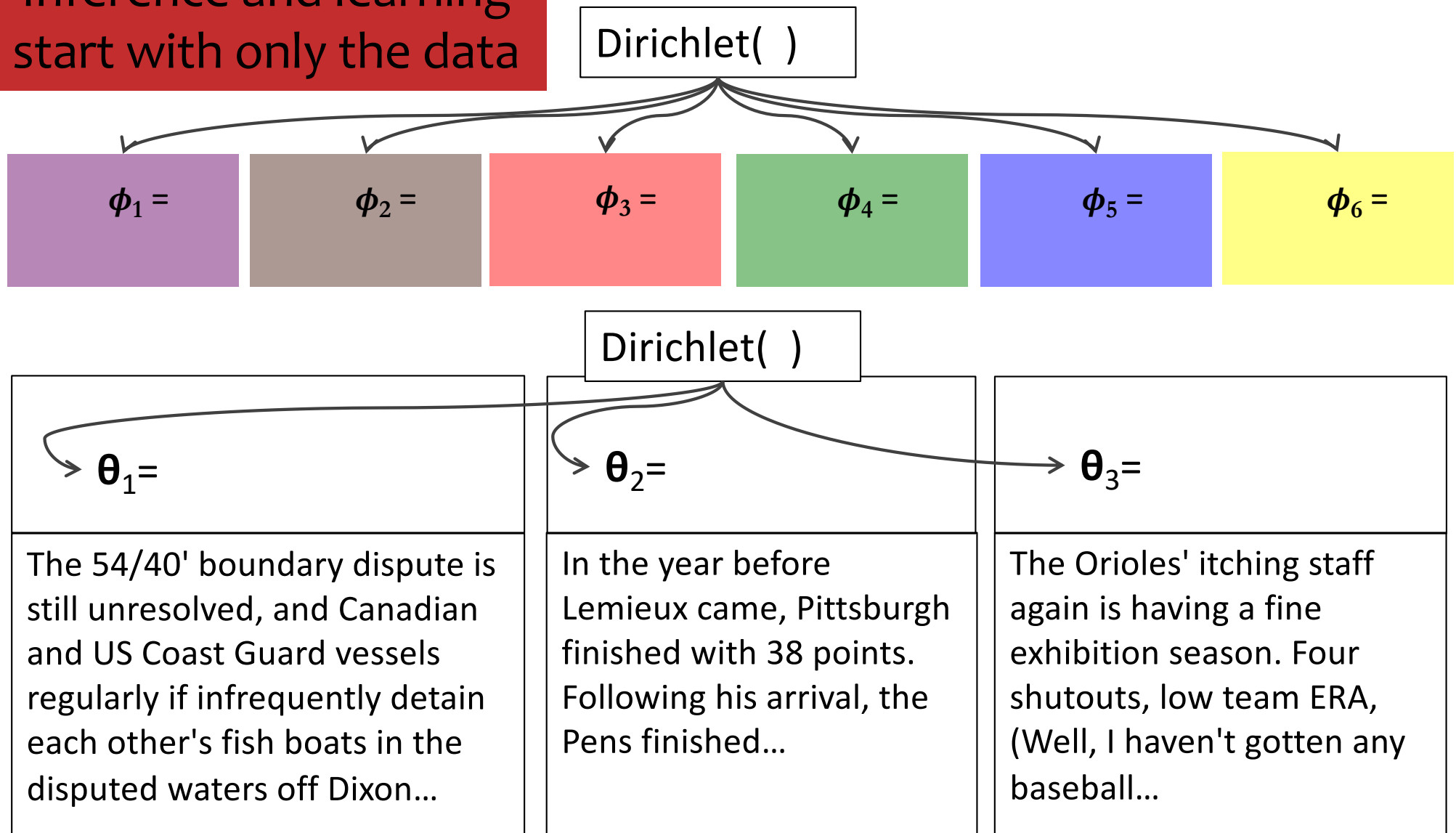


# LDA for Topic Modeling



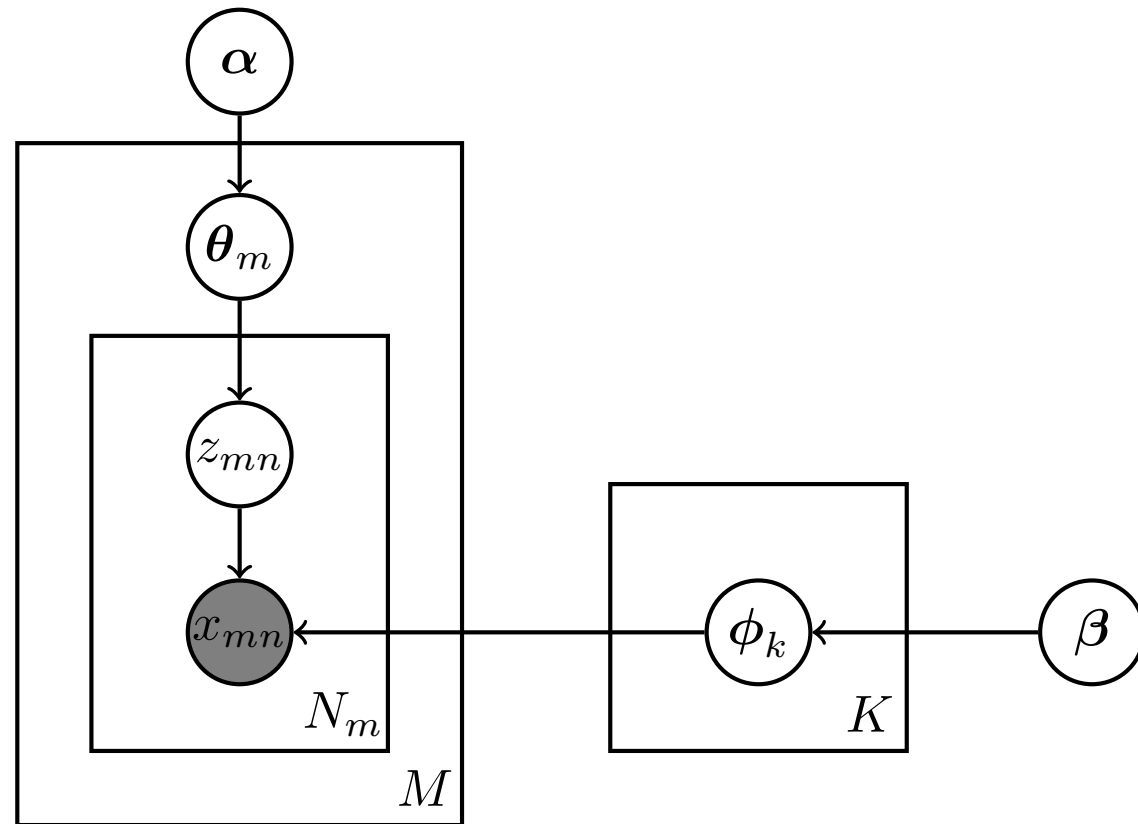
# LDA for Topic Modeling

Inference and learning start with only the data



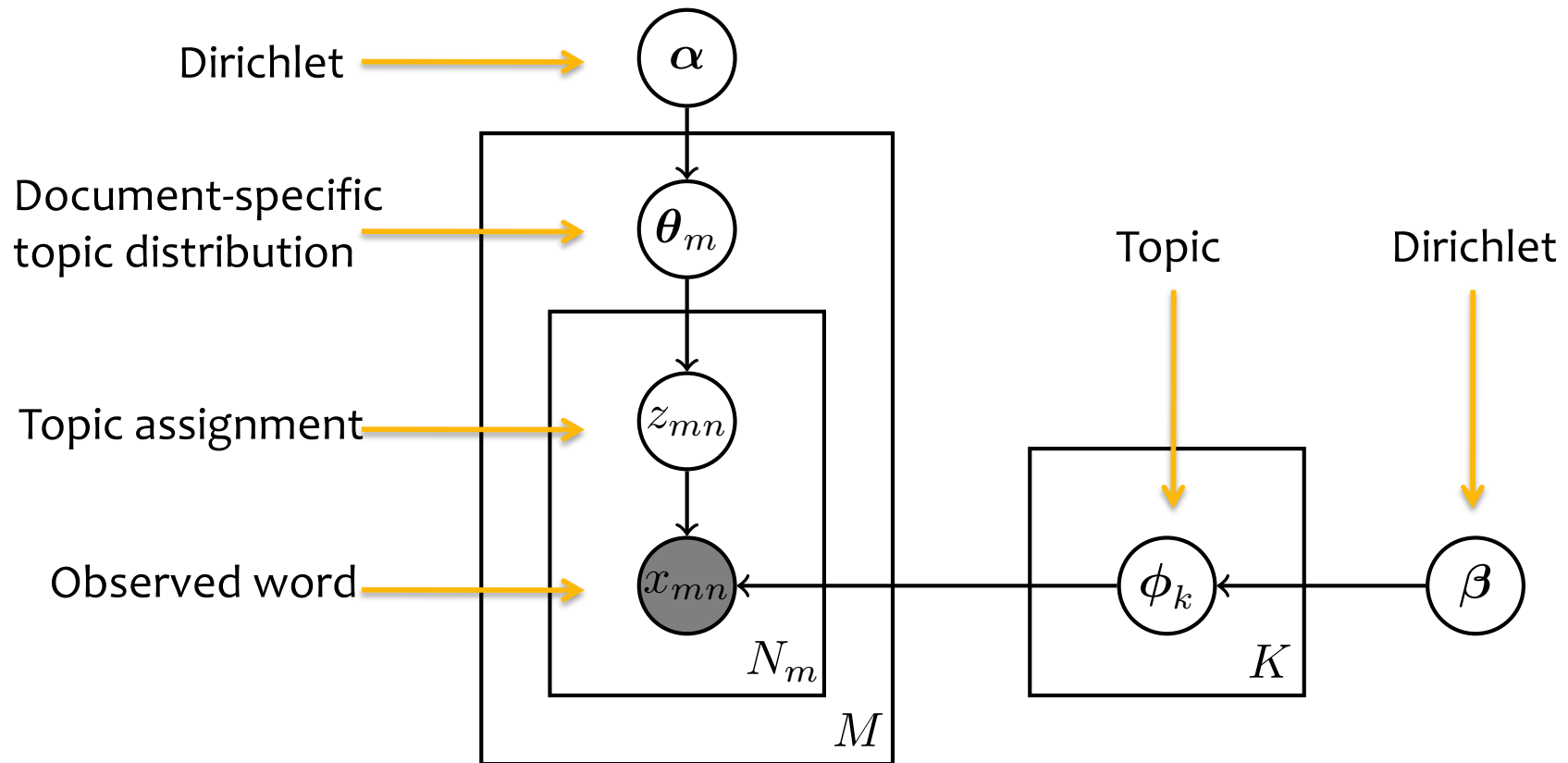
# Latent Dirichlet Allocation

- Plate Diagram



# Latent Dirichlet Allocation

- Plate Diagram





# Latent Dirichlet Allocation

**Question:** Q1

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

**Answer:**

No

**Question:** Q2

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)

# Outline

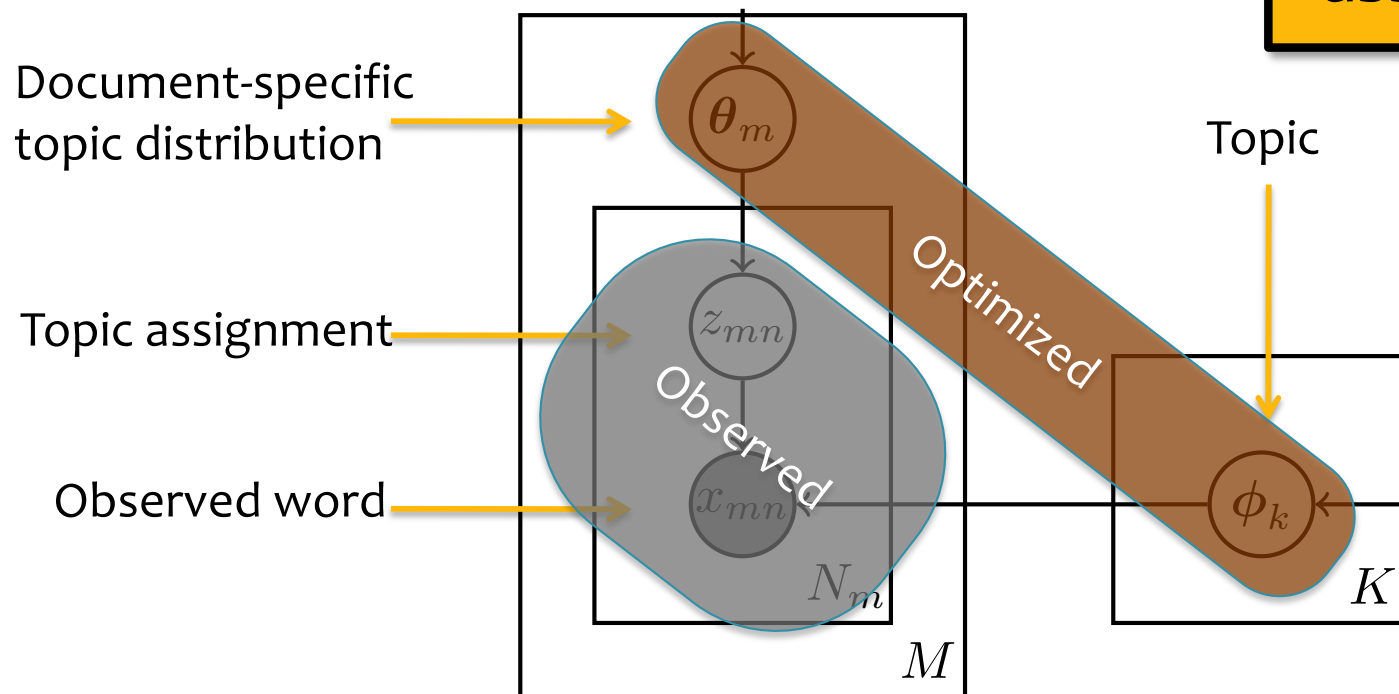
- Applications of Topic Modeling
- Latent Dirichlet Allocation (LDA)
  1. Beta-Bernoulli
  2. Dirichlet-Multinomial
  3. Dirichlet-Multinomial Mixture Model
  4. LDA
- **Bayesian Inference for Parameter Estimation**
  - Exact inference
  - EM
  - Monte Carlo EM
  - Gibbs sampler
  - Collapsed Gibbs sampler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

# **BAYESIAN INFERENCE FOR PARAMETER ESTIMATION**

# LDA Inference

- Fully Observed MLE

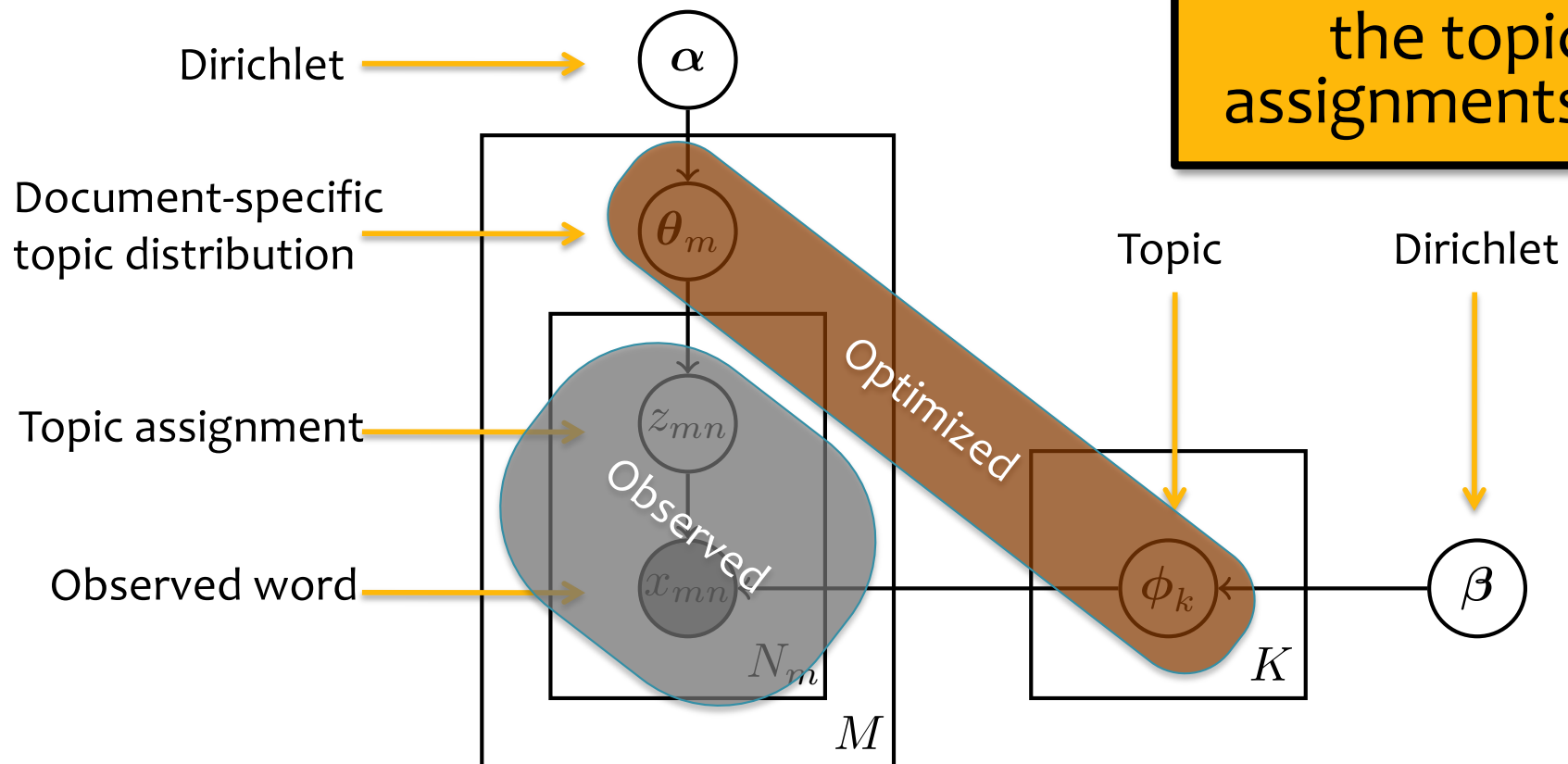
Learning like this  
would be easy,  
but in practice we  
do not observe  
the topic  
assignments  $z_{mn}$



# LDA Inference

- Full Observed MAP Estimation

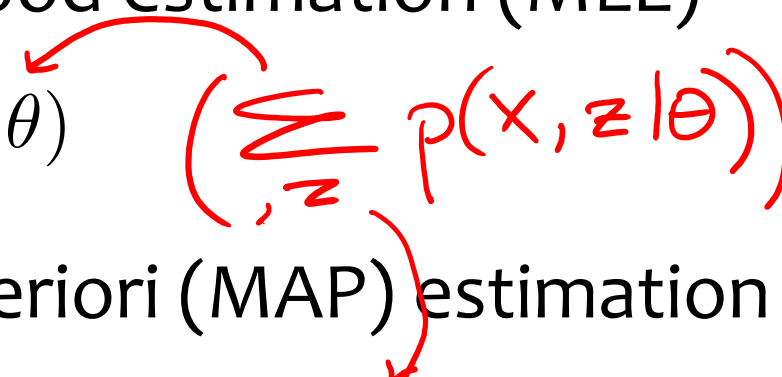
Learning like this would be easy, but in practice we do not observe the topic assignments  $z_{mn}$



# Unsupervised Learning

Three learning paradigms:

1. Maximum likelihood estimation (MLE)

$$\arg \max_{\theta} p(X|\theta) \quad \left( \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta) \right)$$
Two red arrows originate from the handwritten red expression. One arrow points from the summation symbol to the likelihood term  $p(X|\theta)$  in the MLE equation above. The other arrow points from the same summation symbol to the likelihood term  $p(X|\theta)$  in the MAP equation below.

2. Maximum a posteriori (MAP) estimation

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

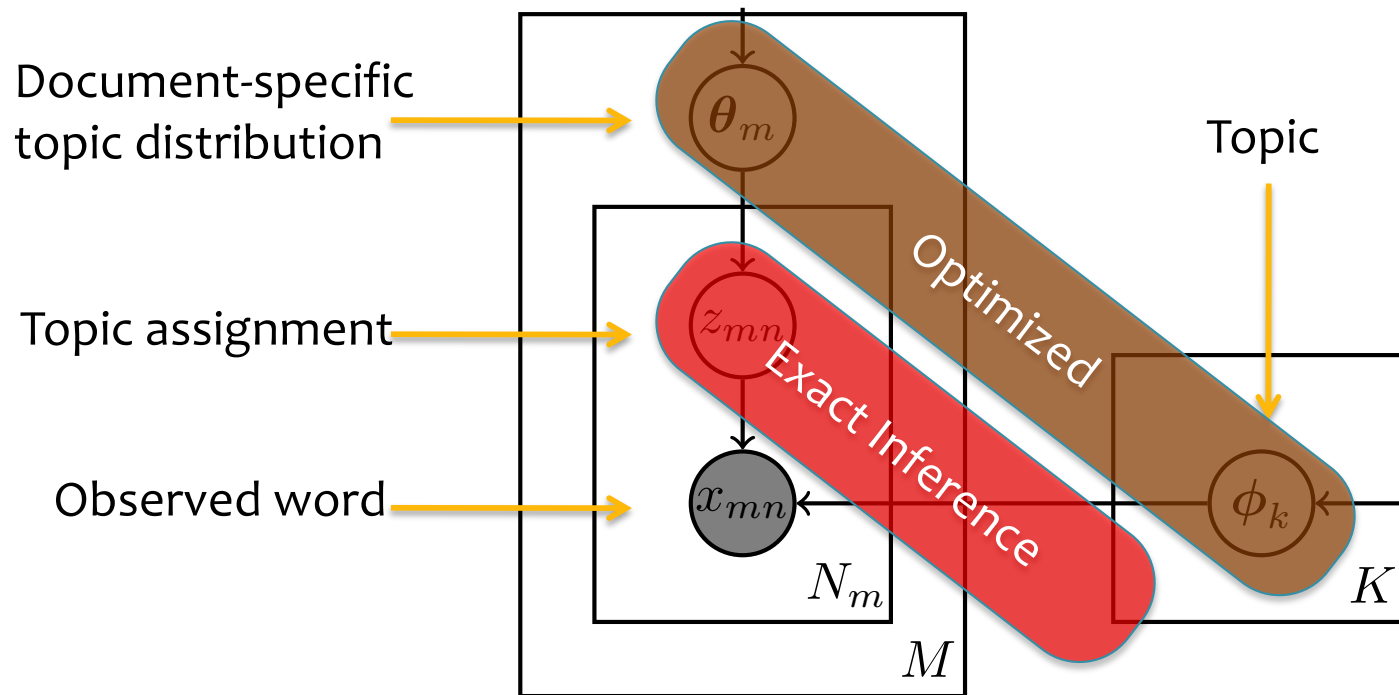
3. Bayesian approach

Estimate the posterior:

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

# LDA Inference

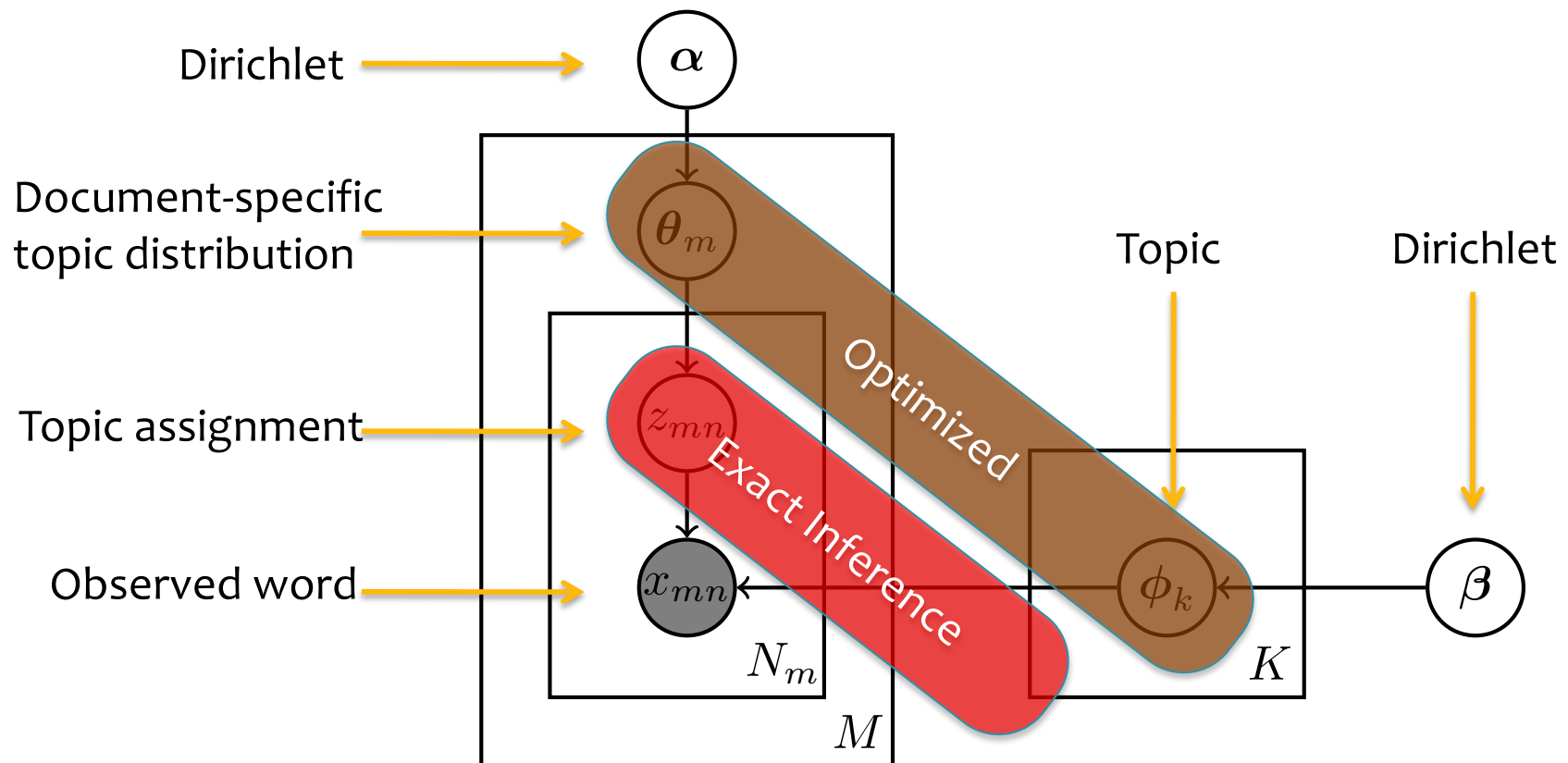
- Standard EM (MLE)





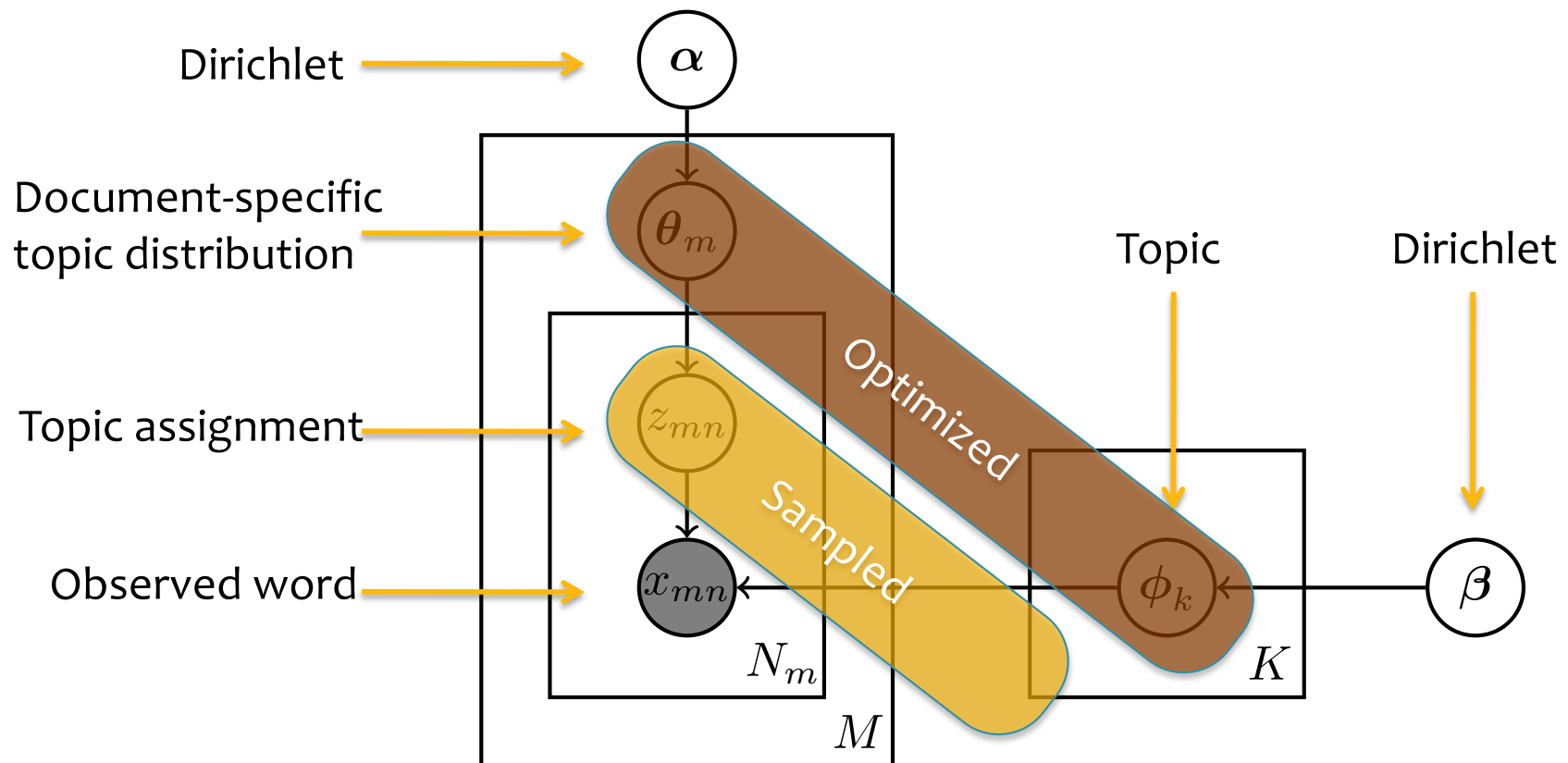
# LDA Inference

- Standard EM (MAP Estimation)



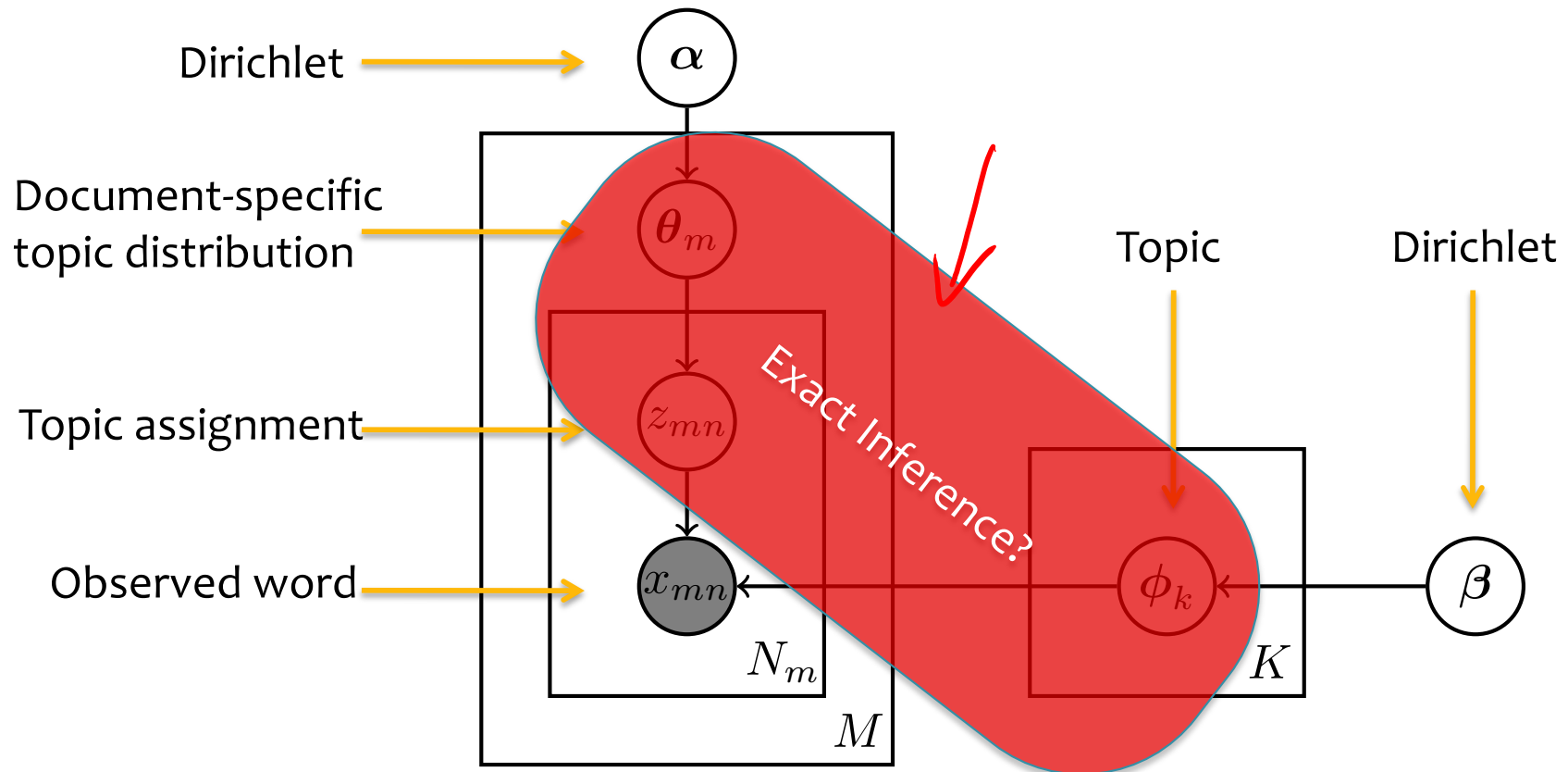
# LDA Inference

- Monte Carlo EM (MAP Estimation)



# LDA Inference

- Bayesian Approach



# LDA Inference

- Bayesian Approach

