

# DSC250: Advanced Data Mining

## Topic Models

Zhitong Hu

Lecture 7, Jan 28, 2025

# Outline

- Topic models: v1, v2, v3
- Paper Presentations:
  - (1) Liyuan Jin, Riqian Hu: **Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism**
  - (2) Victoria Jin, Wenqi Li: **Large Language Models Are Human-Level Prompt Engineers**

# Recap: Represent a Document

- Most common way: Bag-of-Words
  - Ignore the order of words
  - keep the count

c1: Human machine interface for Lab ABC computer applications  
c2: A survey of user opinion of computer system response time  
c3: The EPS user interface management system  
c4: System and human system engineering testing of EPS  
c5: Relation of user-perceived response time to error measurement

m1: The generation of random, binary, unordered trees  
m2: The intersection graph of paths in trees  
m3: Graph minors IV: Widths of trees and well-quasi-ordering  
m4: Graph minors: A survey



	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Vector space model

# Recap: Represent a Topic

- A topic is represented by a word distribution
- Relate to an issue

universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148
bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
strain	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824

**TOPIC 42**  
 archaeological, landscape, historical, designations, eligible, architectural, significant, home, ordinance, properties, character, registers, survey, tourism, promote, history, structures, district, white, landmark, style, listed, past, significant, **historic**, designated, owners, process, local, rehabilitation, architectural, buildings, preserve, creative, distribution, historic, demolition, buildings, heritage, property, building, tax, category, designation, local, area, action, historically, design, plan, project, land, copy, planning, places, guidelines, effective, neighborhoods, adaptive, original, permanent, professional, integrity, general, documentation

**TOPIC 43**  
 archaeological, landscape, historical, designations, eligible, architectural, significant, home, ordinance, properties, character, registers, survey, tourism, promote, history, structures, district, white, landmark, style, listed, past, significant, **historic**, designated, owners, process, local, rehabilitation, architectural, buildings, preserve, creative, distribution, historic, demolition, buildings, heritage, property, building, tax, category, designation, local, area, action, historically, design, plan, project, land, copy, planning, places, guidelines, effective, neighborhoods, adaptive, original, permanent, professional, integrity, general, documentation

**TOPIC 45**  
 community, identified, installed, technology, recently, implemented, created, underway, assembly, inflation, urban, economic, initiated, provided, including, recently, designed, building, program, line, board, offers, miles, fall, million, source, miles, fall, potential, construction, completed, June, projects, sector, partnership, year, study, project, called, day, major, grant, based, phase, begin, million, impact, portion, strategy, lead, received, funded, addition, approximately, effort, estimated, set, facility, case, built, years, total, total, days, years, operated, fund, partners, working, plan, opened, updated, numerous, agreement, conducted, capability, beginning, significantly, high, been, visited, addressing, expected, included, December, promising

**TOPIC 46**  
 community, identify, involving, opportunities, downtown, gathering, decision, include, neighborhoods, guidance, prepared, future, vision, urban, suburban, surrounding, facilities, opportunity, region, for, development, designation, assist, employment, existing, specialty, part, broadway, community, key, study, villages, located, adopted, policies, bush, march, high, general, include, quasi, state, goals, areas, residents, airport, plan, figure, special, highway, require, down, light, provide, pha, drainage, project, residential, allowed, rubber, subject, houses, urban, diagram, generally, university, building, underground

**TOPIC 48**  
 evaluated, mechanisms, funded, animals, proposed, revised, increased, amount, analysis, increases, fee, include, fiscal, benefit, private, project, pace, grants, expenditures, need, capital, cost, investments, program, term, costs, pay, managed, improvements, support, revenue, projects, program, performance, tax, operating, condition, funding, financial, total, priority, facilitated, services, serial, listed

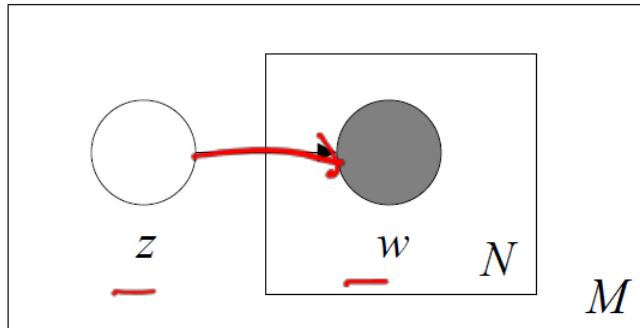
appro, government, require, individual, priorities, bring, ensuring, en, recent, ad, making, 4

# Notations

- Word, document, topic
  - $w, d, z$
- Word count in document:
  - $c(w, d)$  : number of times word  $w$  occurs in document  $d$
  - or  $x_{dn}$ : number of times the  $n$ th word in the vocabulary occurs in document  $d$
- Word distribution for each topic (  $\beta_z$  )
  - $\beta_{zw}$ :  $p(w|z)$

# Recap: Topic Model v1: Multinomial Mixture Model

Graphical Model ✓



- Plates indicate replicated variables.
- Shaded nodes are observed; unshaded nodes are hidden.

## Generative model

- For each document

- Sample its cluster label  $z \sim \text{Categorical}(\boldsymbol{\pi})$

- $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$ ,  $\pi_k$  is the proportion of jth cluster

- $p(z = k) = \pi_k$

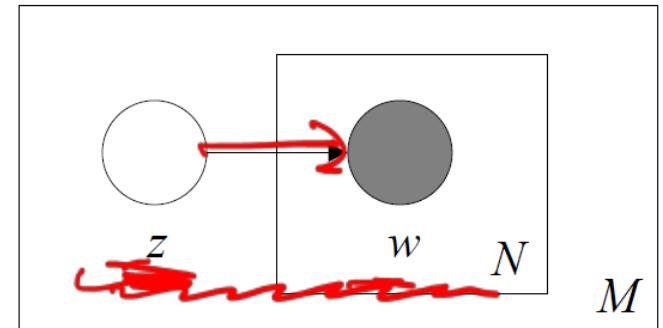
- Sample its word vector  $\mathbf{x}_d \sim \text{multinomial}(\boldsymbol{\beta}_z)$

- $\boldsymbol{\beta}_z = (\beta_{z1}, \beta_{z2}, \dots, \beta_{zN})$ ,  $\beta_{zn}$  is the parameter associate with nth word in the vocabulary

- $p(\mathbf{x}_d | z = k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}} \propto \prod_n \beta_{kn}^{x_{dn}}$

parameter

## Recap: Likelihood Function



$$\begin{aligned}
 L &= \prod_d p(\underline{x}_d) = \prod_d \sum_k p(x_d, z = k) \\
 &= \prod_d \sum_k p(x_d | z = k) p(z = k) \\
 &= \prod_d \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \sum_k p(z = k) \prod_n \beta_{kn}^{x_{dn}}
 \end{aligned}$$

## Recap: Topic Model v2: pLSA

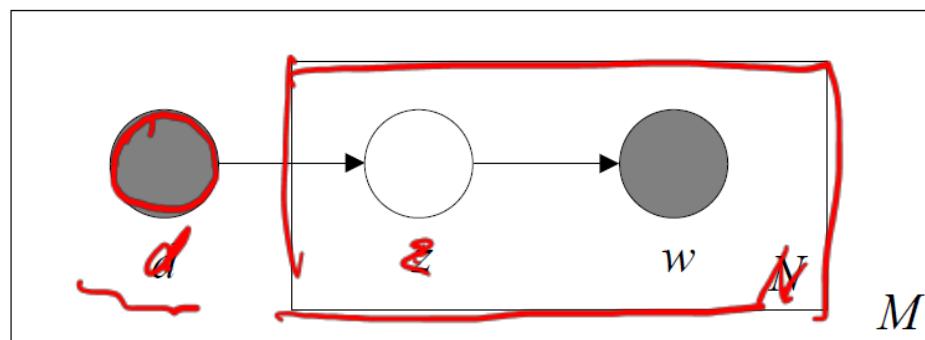
- For each position in d,  $n = 1, \dots, N_d$

- Generate the topic for the position as

$z_n | d \sim \text{Categorical}(\theta_d)$ , i.e.,  $p(z_n = k | d) = \theta_{dk}$   
(Note, 1 trial multinomial)

- Generate the word for the position as

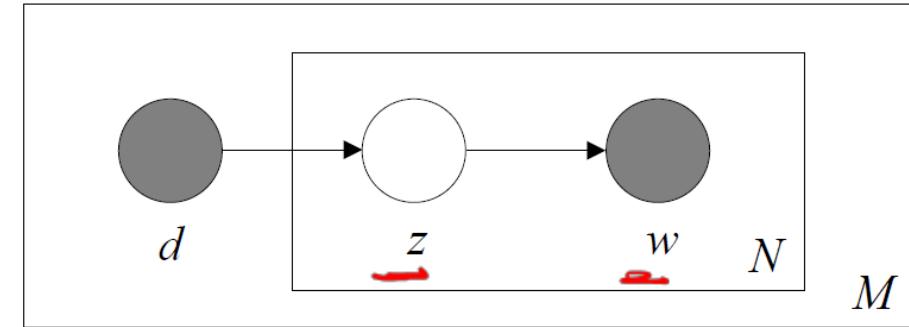
$w_n | z_n \sim \text{Categorical}(\beta_{z_n})$ , i.e.,  $p(w_n = w | z_n) = \beta_{z_n w}$



Graphical  
Model

## Likelihood Function

- Probability of a word  $w$



$$p(w|d, \theta, \beta) = \sum_k P(w, z=k | d, \theta, \beta)$$

$$\beta = \begin{bmatrix} \vdots & \swarrow \\ k & \text{---} \\ \vdots & \nearrow \end{bmatrix}$$

$\beta_{kw}$

$\beta_{k\theta}$

$$= \sum_k P(w|z=k, \beta) p(z=k|d, \theta)$$

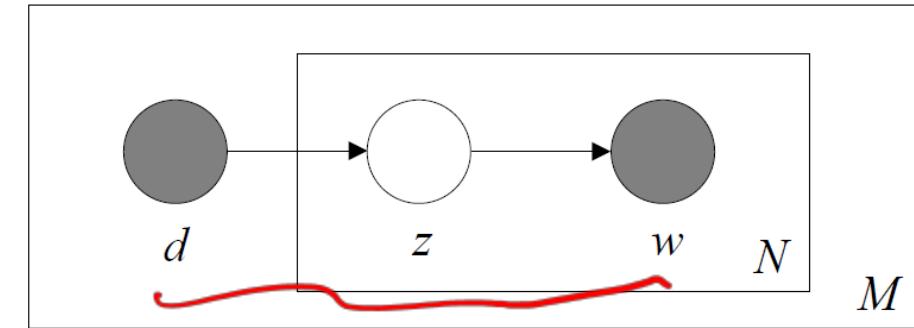
$$= \sum_k \beta_{kw} \cdot o_{dk}$$

# Likelihood Function

- Probability of a word  $w$

$$p(w|d, \theta, \beta) = \sum_k p(w, z = k | d, \theta, \beta)$$

$$= \sum_k p(w|z = k, d, \theta, \beta) p(z = k | d, \theta, \beta) = \sum_k \beta_{kw} \theta_{dk}$$



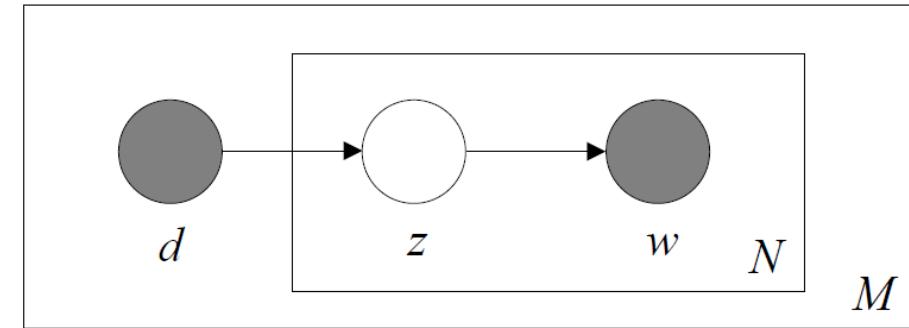
# Likelihood Function

- Probability of a word  $w$

$$p(w|d, \theta, \beta) = \sum_k p(w, z = k | d, \theta, \beta)$$

$$= \sum_k p(w|z = k, d, \theta, \beta)p(z = k | d, \theta, \beta) = \sum_k \beta_{kw} \theta_{dk}$$

- Likelihood of a corpus



# Likelihood Function

- Probability of a word  $w$

$$p(w|d, \theta, \beta) = \sum_k p(w, z = k | d, \theta, \beta)$$

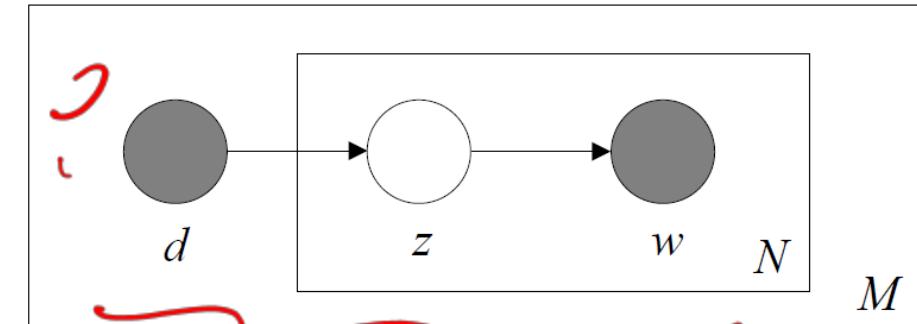
$$= \sum_k p(w|z = k, d, \theta, \beta) p(z = k | d, \theta, \beta) = \sum_k \beta_{kw} \theta_{dk}$$

- Likelihood of a corpus

$$\prod_{d=1} P(w_1, \dots, w_{N_d}, d | \theta, \beta, \pi)$$

$$= \prod_{d=1} P(d) \left\{ \prod_{n=1}^{N_d} \left( \sum_k P(z_n = k | d, \theta_d) P(w_n | \beta_k) \right) \right\}$$

$$= \prod_{d=1} \pi_d \left\{ \prod_{n=1}^{N_d} \left( \sum_k \theta_{dk} \beta_{kw_n} \right) \right\}$$



$\pi_d$  is usually considered as uniform, i.e.,  $1/M$

## Re-arrange the Likelihood Function

- Group the same word from different positions together

$$\max \log L = \sum_{dw} c(w, d) \log \sum_z \theta_{dz} \beta_{zw}$$

$$s.t. \sum_z \theta_{dz} = 1 \text{ and } \sum_w \beta_{zw} = 1$$

# Limitations of pLSA

- Not a proper generative model
  - $\theta_d$  is treated as a parameter
  - Cannot model new documents
- Solution:
  - Make it a proper generative model by adding priors to  $\theta$  and  $\beta$

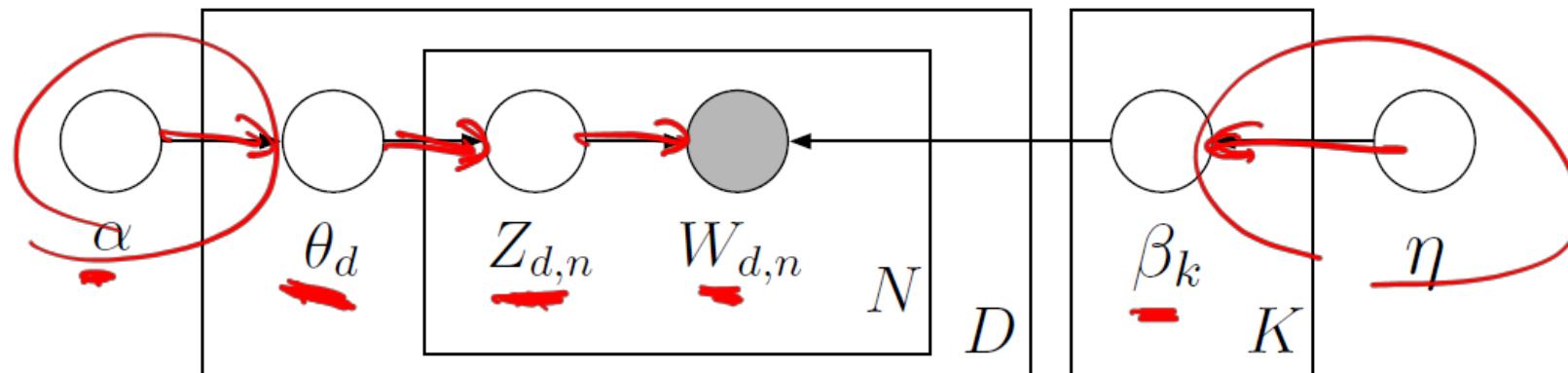
# Limitations of pLSA

- Not a proper generative model
  - $\theta_d$  is treated as a parameter
  - Cannot model new documents
- Solution:
  - Make it a proper generative model by adding priors to  $\theta$  and  $\beta$



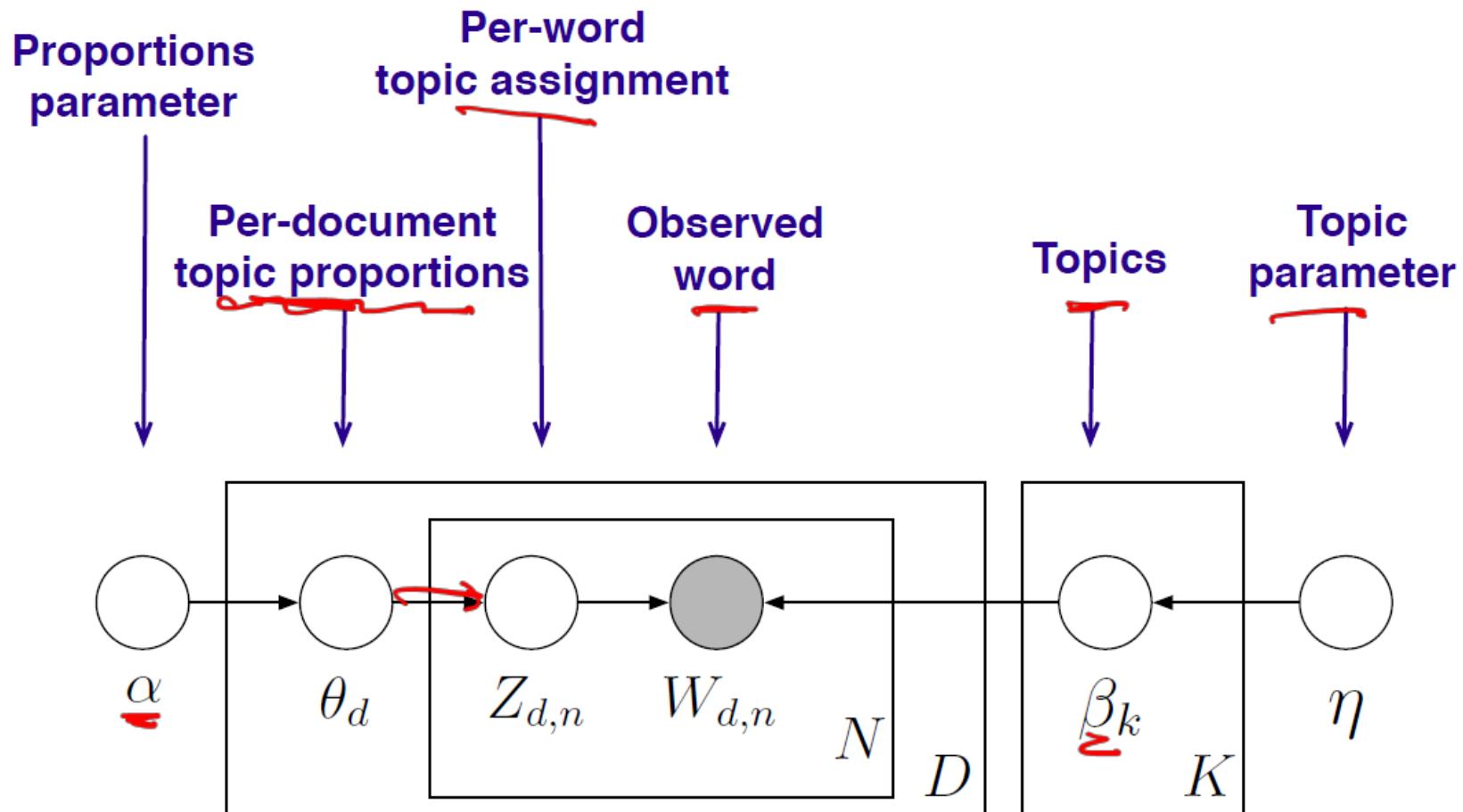
Topic Model v3: Latent Dirichlet Allocation (LDA)

# Topic Model v3: Latent Dirichlet Allocation (LDA)



$\theta_d \sim \text{Dirichlet}(\alpha)$ : address topic distribution for unseen documents  
 $\beta_k \sim \text{Dirichlet}(\eta)$ : smoothing over words

# Topic Model v3: Latent Dirichlet Allocation (LDA)



$\theta_d \sim \text{Dirichlet}(\alpha)$ : address topic distribution for unseen documents  
 $\beta_k \sim \text{Dirichlet}(\eta)$ : smoothing over words

# Generative Model for LDA

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

$\beta_k \sim \text{Dir}(\eta)$  [draw distribution over words]

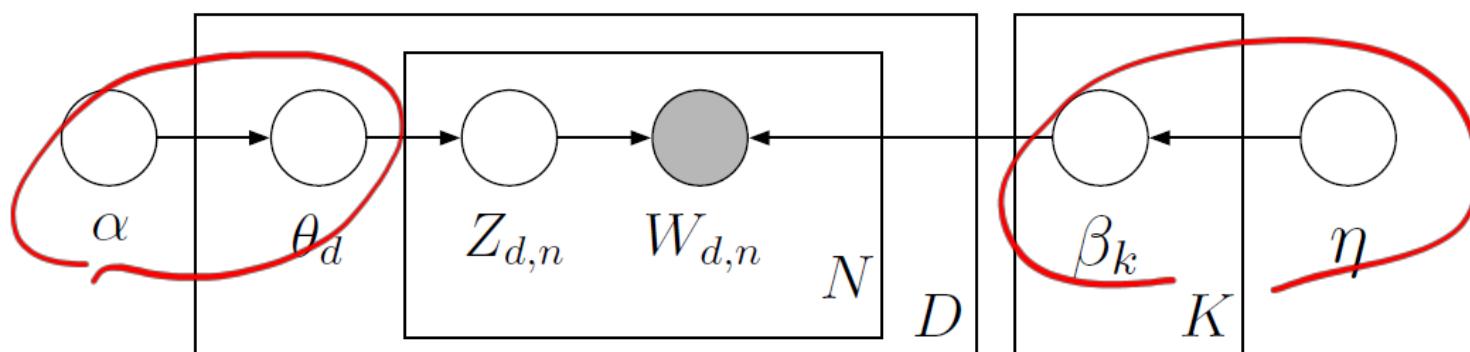
For each document  $d \in \{1, \dots, D\}$

$\theta_d \sim \text{Dir}(\alpha)$  [draw distribution over topics]

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

$z_{d,n} \sim \text{Mult}(1, \theta_d)$  [draw topic assignment]

$w_{d,n} \sim \theta_{z_{d,n}}$  [draw word]



## Review: Dirichlet Distribution

- Dirichlet distribution:  $\theta \sim Dirichlet(\alpha)$

- i.e.,  $p(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k-1}$ , where  $\alpha_k > 0$

- $\Gamma(\cdot)$  is gamma function:  $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$

- $\Gamma(z+1) = z\Gamma(z)$

## Review: Dirichlet Distribution

- Dirichlet distribution:  $\theta \sim \text{Dirichlet}(\alpha)$

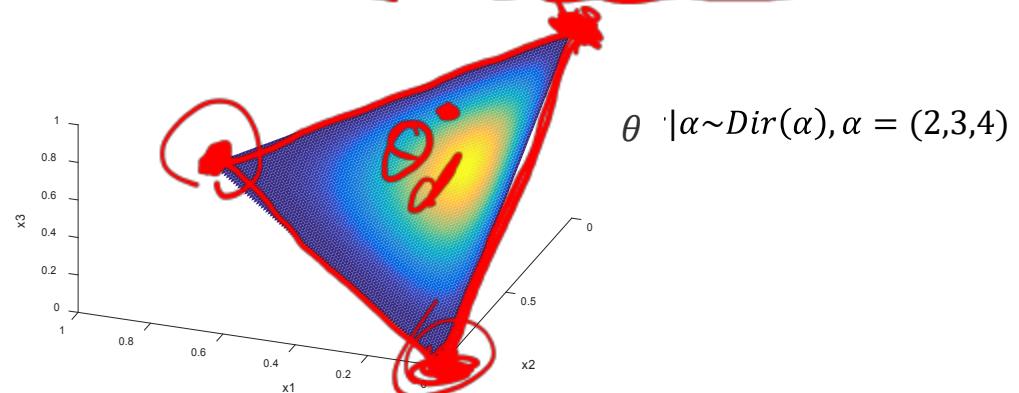
- i.e.,  $p(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k-1}$ , where  $\alpha_k > 0$

- $\Gamma(\cdot)$  is gamma function:  $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$ 
  - $\Gamma(z+1) = z\Gamma(z)$

Simplex view:

$$\theta = \theta_1(1,0,0) + \theta_2(0,1,0) + \theta_3(0,0,1) = (\theta_1, \theta_2, \theta_3)$$

Where  $0 \leq \theta_1, \theta_2, \theta_3 \leq 1$  and  $\theta_1 + \theta_2 + \theta_3 = 1$

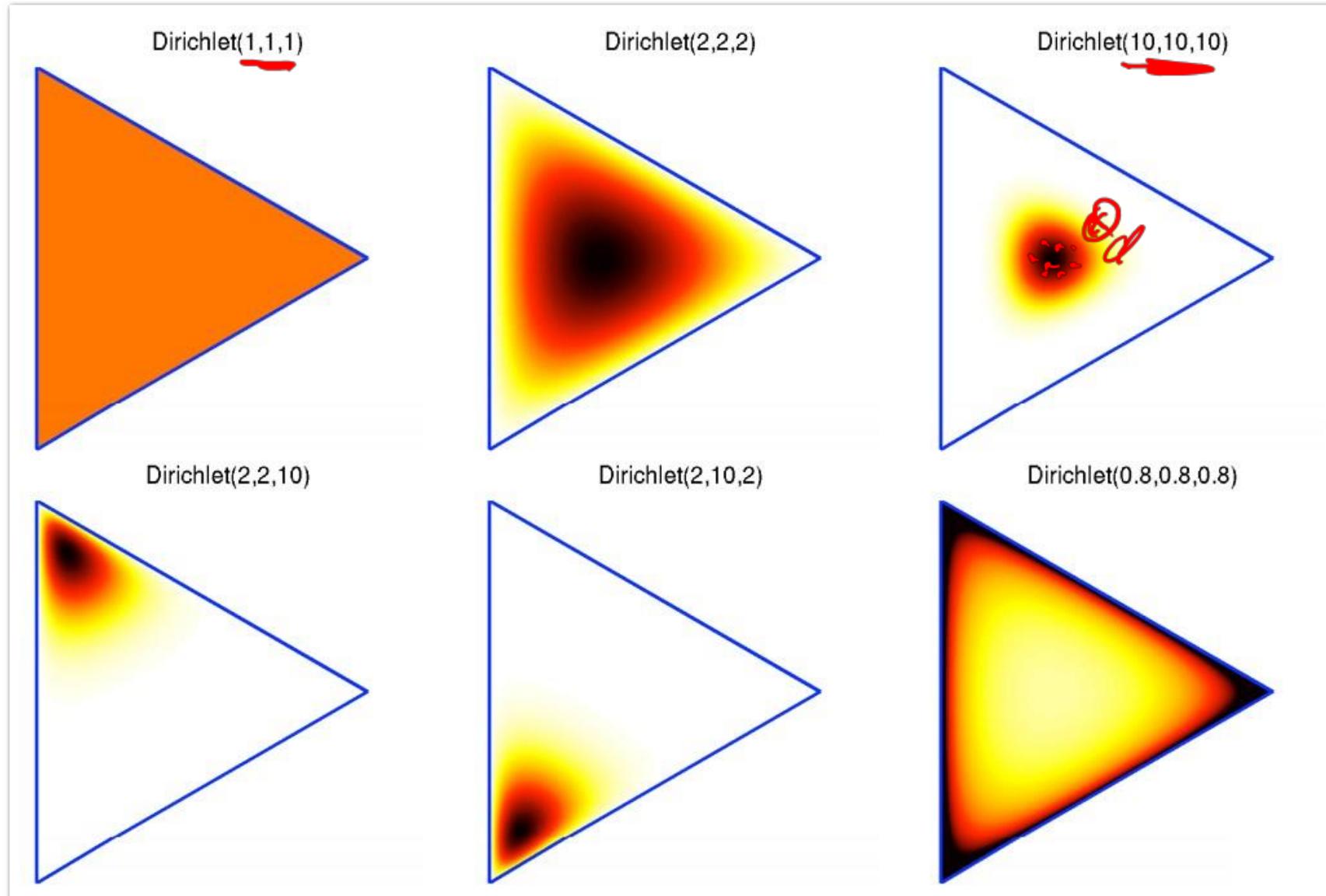


$\theta \sim \text{Dir}(\alpha), \alpha = (2,3,4)$

$\theta \sim \text{Dir}(0.5, 0.3, 0.2)$

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

# More Examples in the Simplex View



# Generative Model for LDA

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

$$\beta_k \sim \text{Dir}(\eta) \quad [\text{draw distribution over words}]$$

For each document  $d \in \{1, \dots, D\}$

$$\theta_d \sim \text{Dir}(\alpha) \quad [\text{draw distribution over topics}]$$

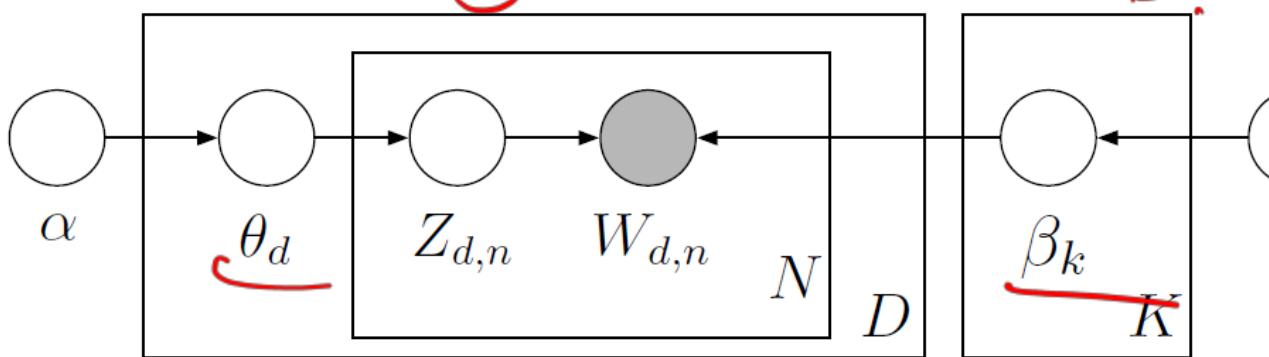
For each word  $n \in \{1, \dots, N_d\}$

$$z_{d,n} \sim \text{Mult}(1, \theta_d) \quad [\text{draw topic assignment}]$$

$$w_{d,n} \sim \theta_{z_{d,n}} \quad [\text{draw word}]$$

*training:  $\{\beta_k\}, \{\theta_d\}$*

$\alpha, \eta$

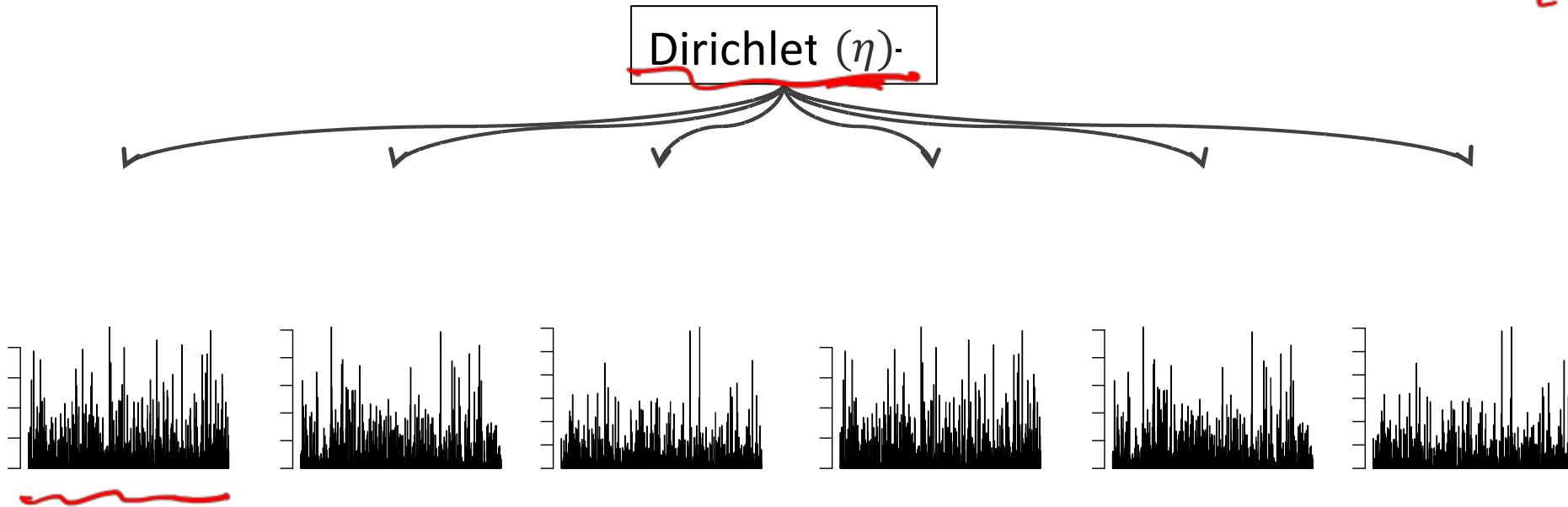


*L-M  
4.1*

*test: d', θ\_d'*

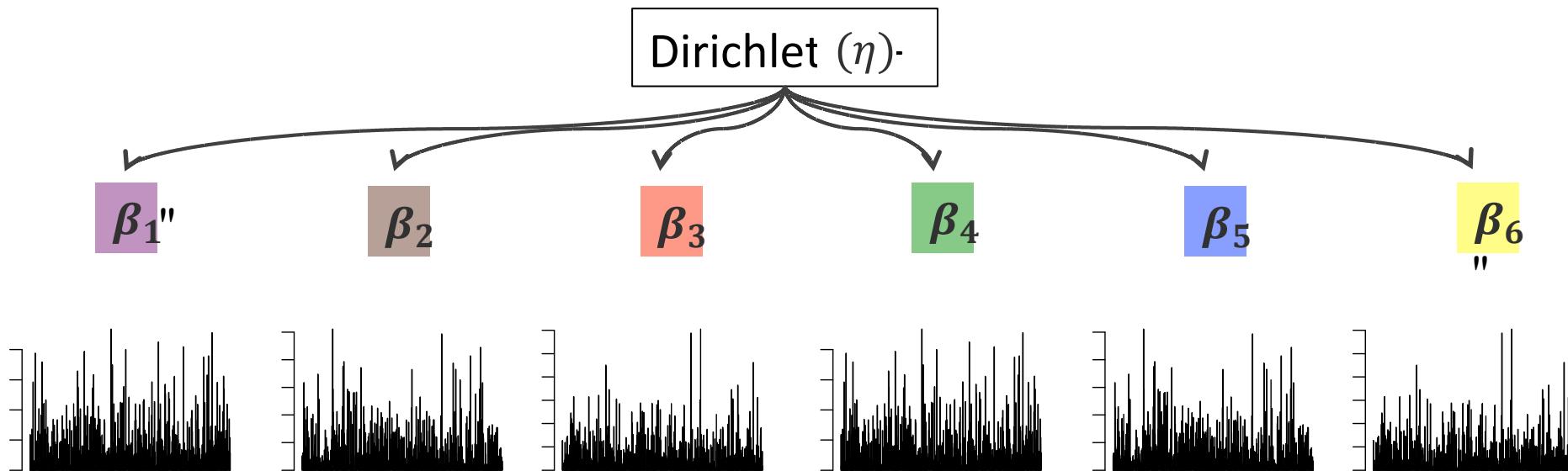
# LDA 'for' Topic 'Modeling'

$K \approx 6$



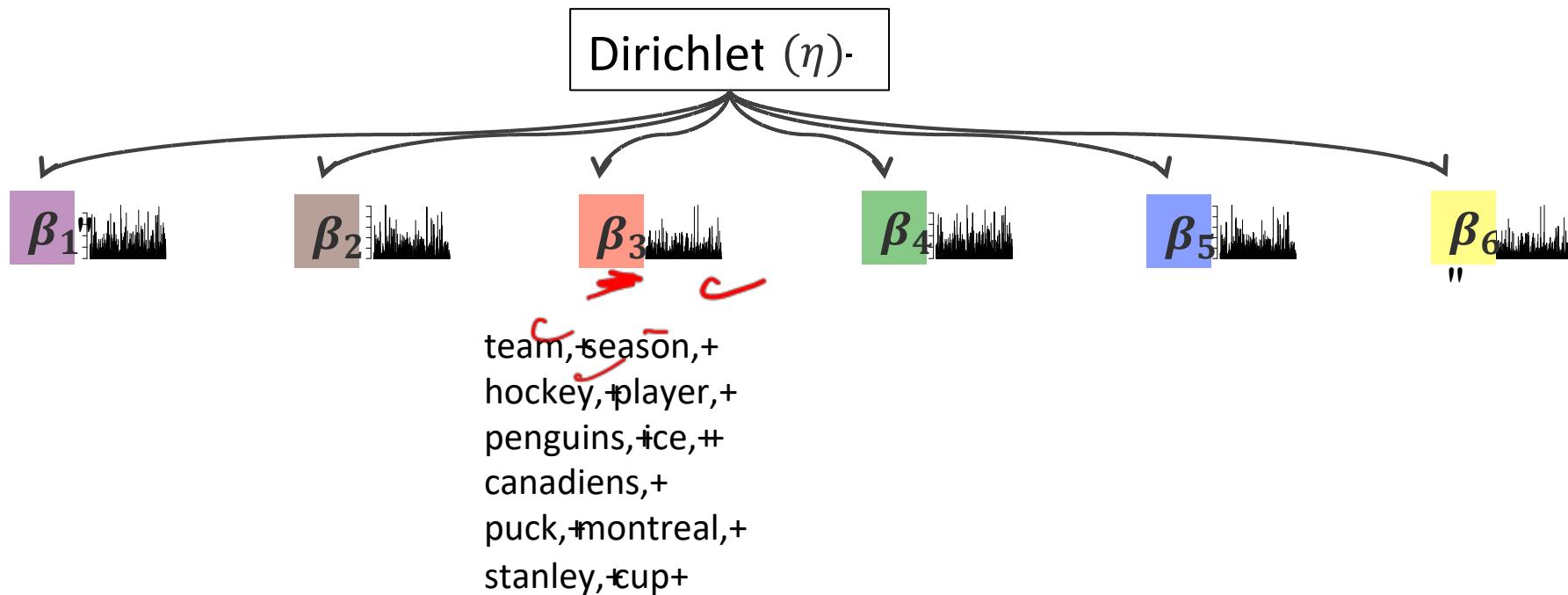
- The generative story begins with only a Dirichlet prior over the topics.
- Each topic defines as a Multinomial distribution over the vocabulary, parameterized by  $\beta_k$ .

# LDA "for" Topic Modeling"



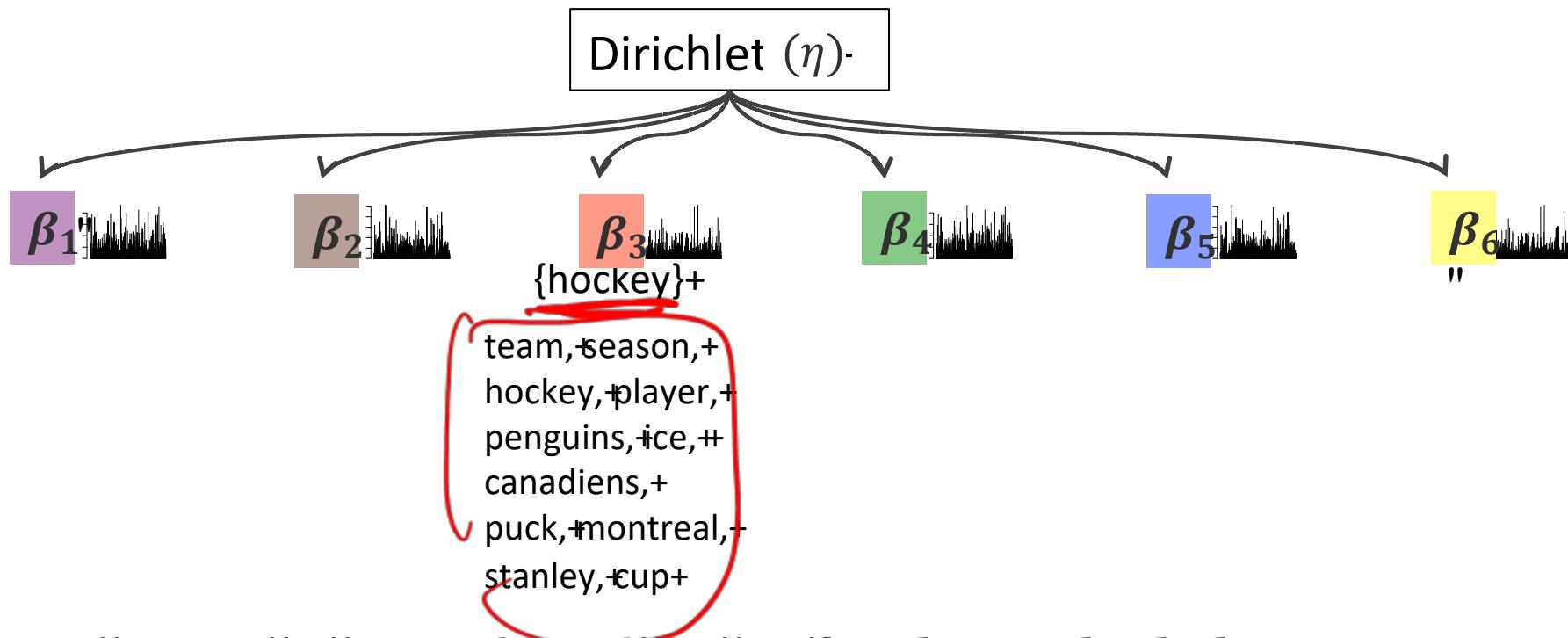
- The generative story begins with only a Dirichlet prior over the topics.
- Each topic defines as a Multinomial distribution over the vocabulary, parameterized by  $\beta_k$

# LDA "for" Topic Modeling"



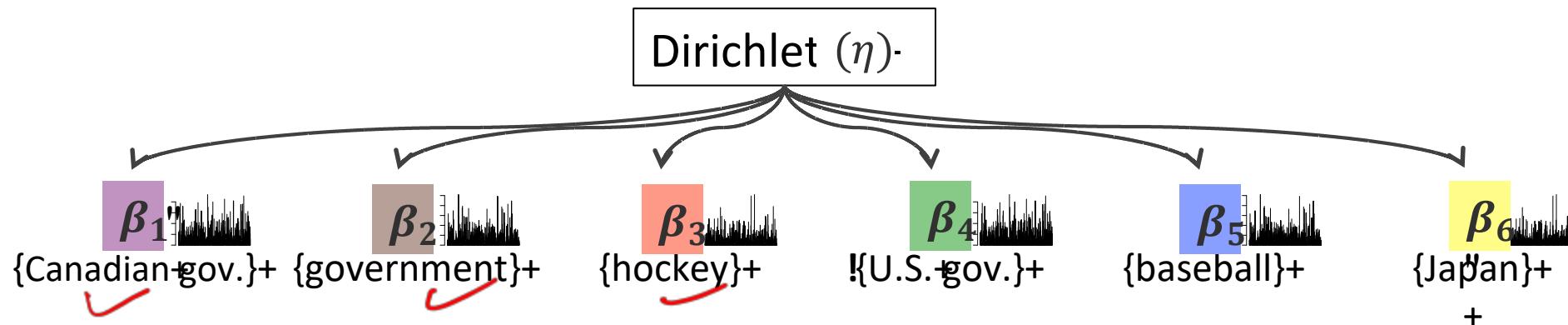
- A "topic" is visualized as its high probability words."

# LDA "for" Topic Modeling"



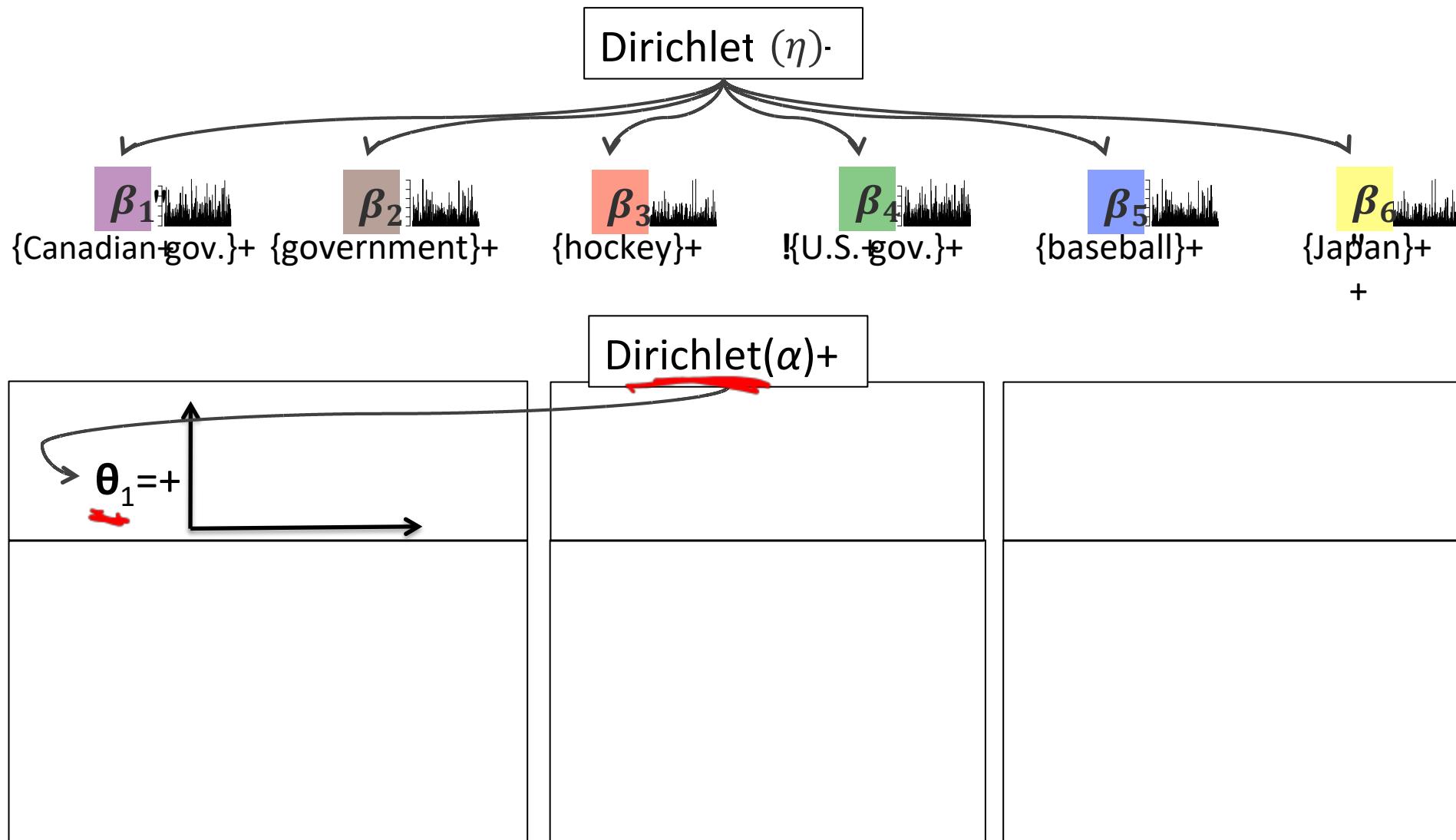
- A "topic" is visualized as its high-probability words."
- A pedagogical label is used to identify the topic."

# LDA "for" Topic Modeling"

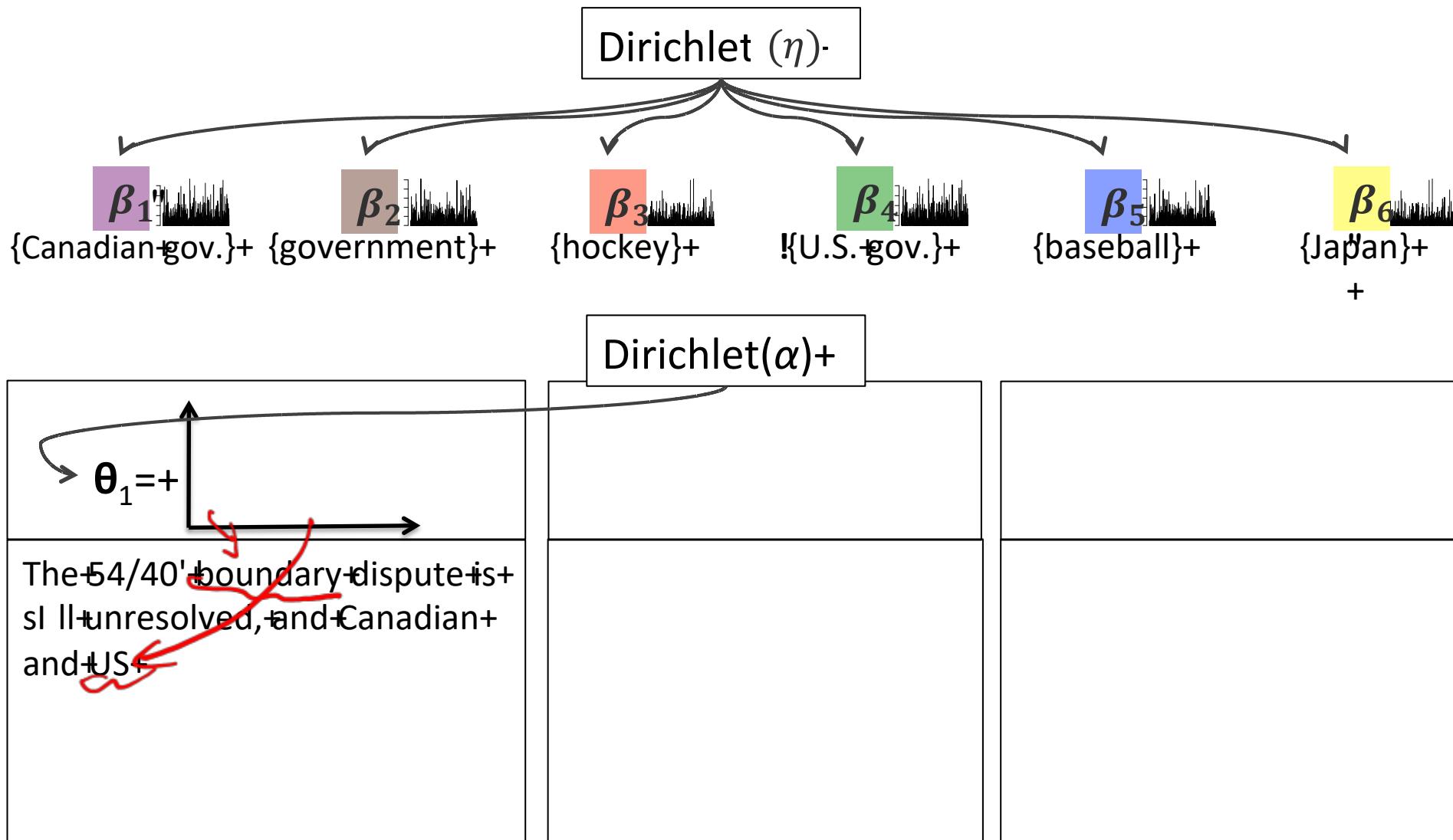


- A "topic" is visualized as its high probability words."
- A pedagogical label & used to identify the topic."

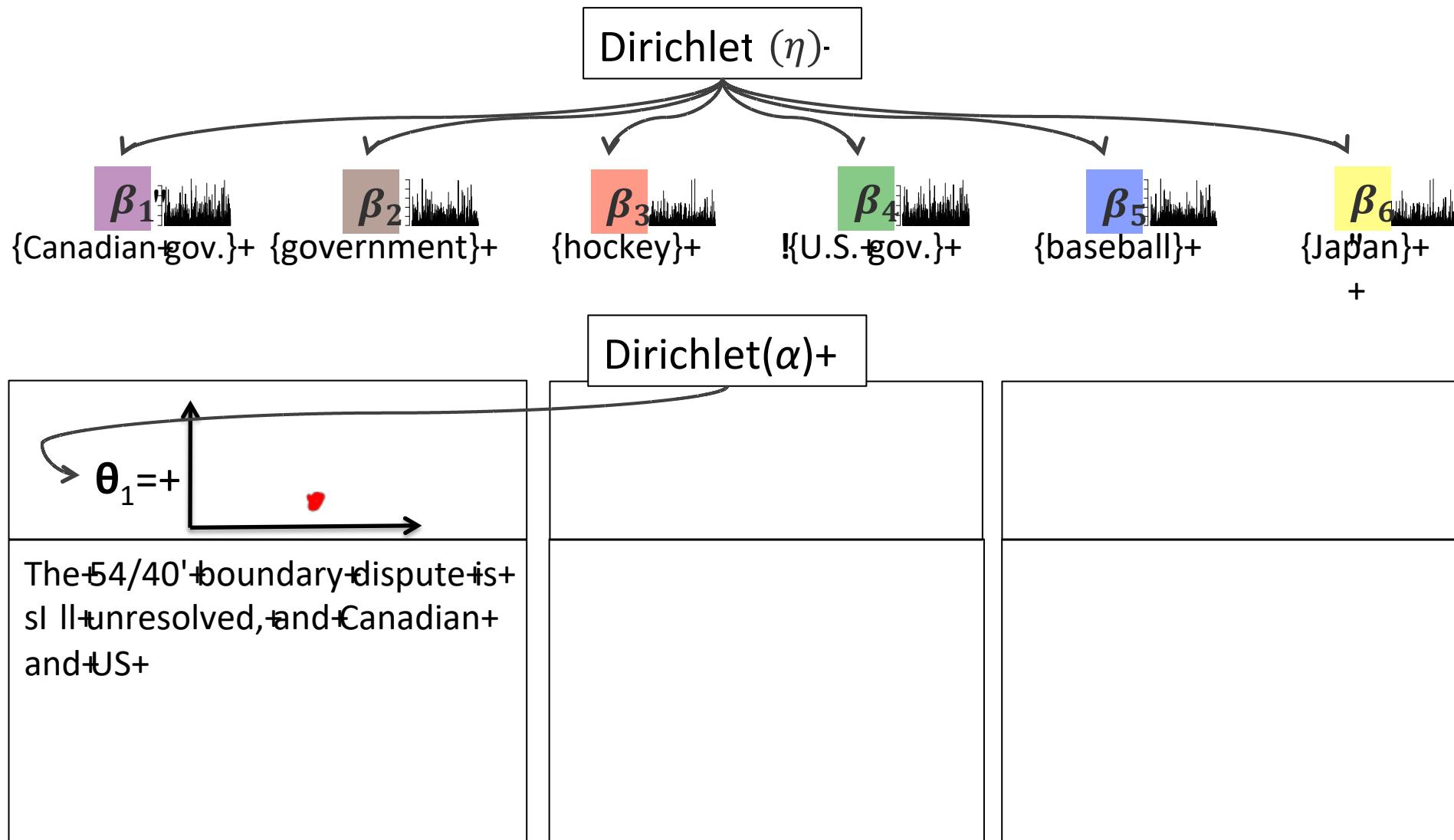
# LDA 'for' Topic Modeling"



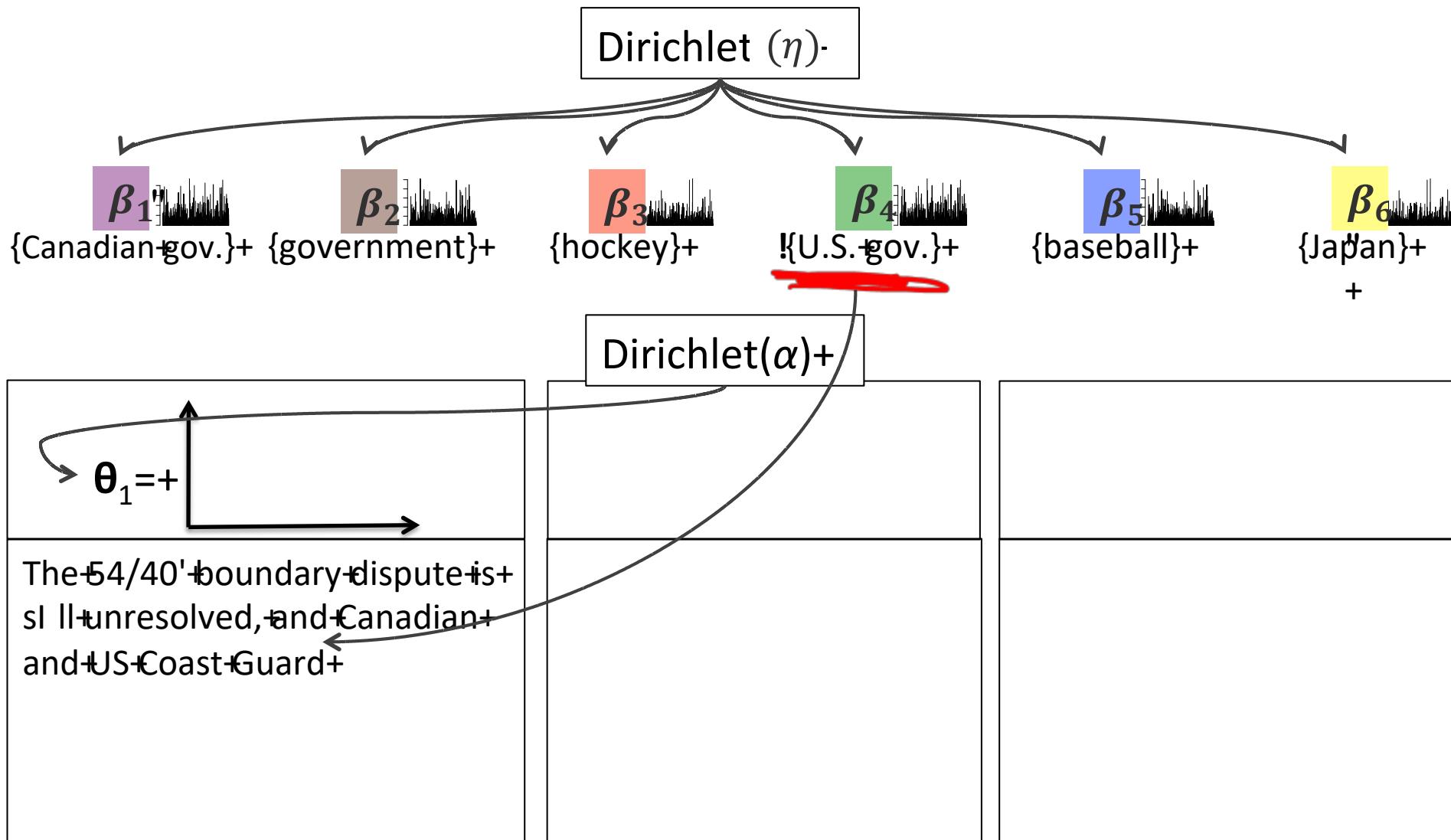
# LDA 'for' Topic Modeling'



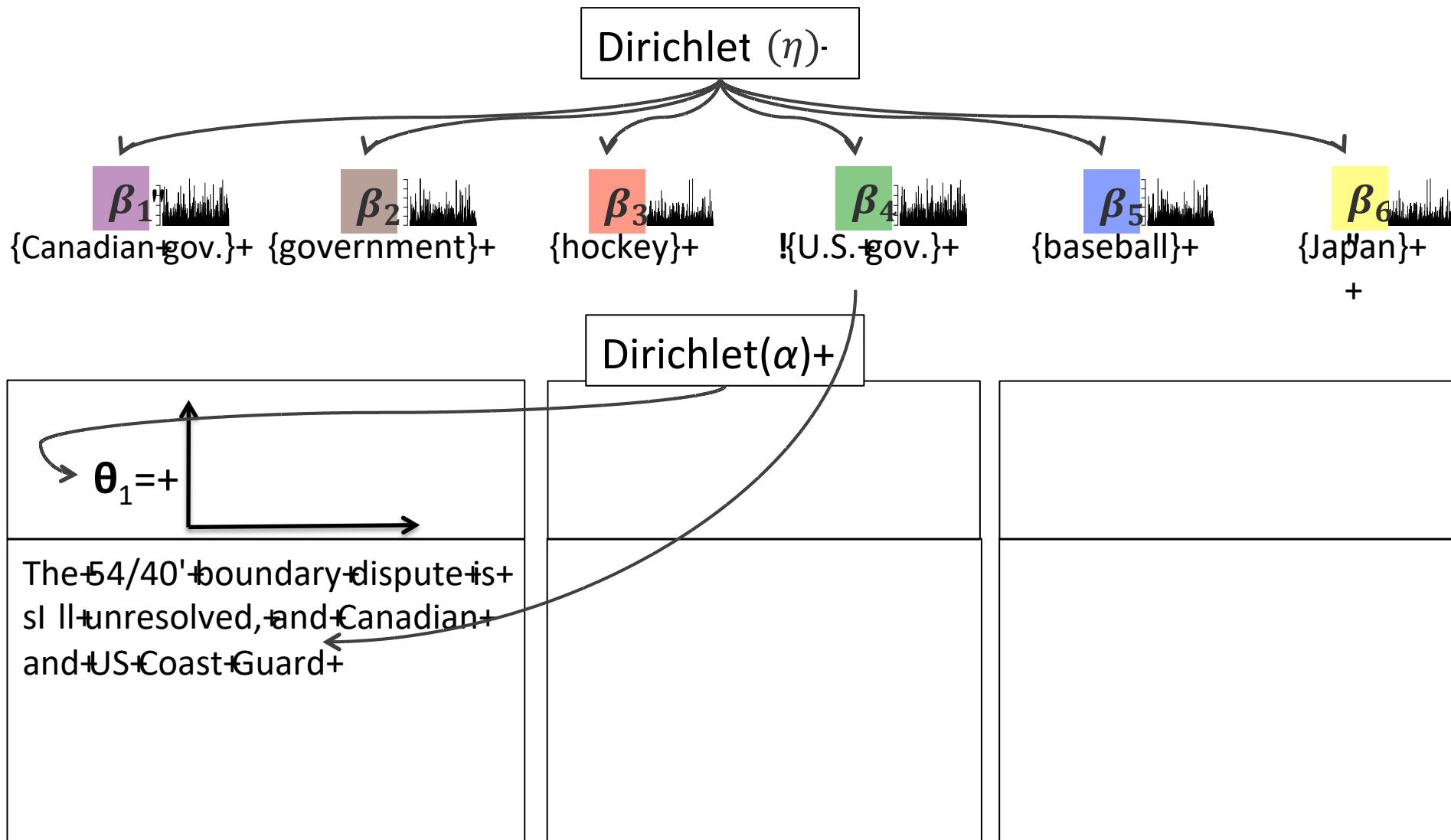
# LDA 'for' Topic Modeling"



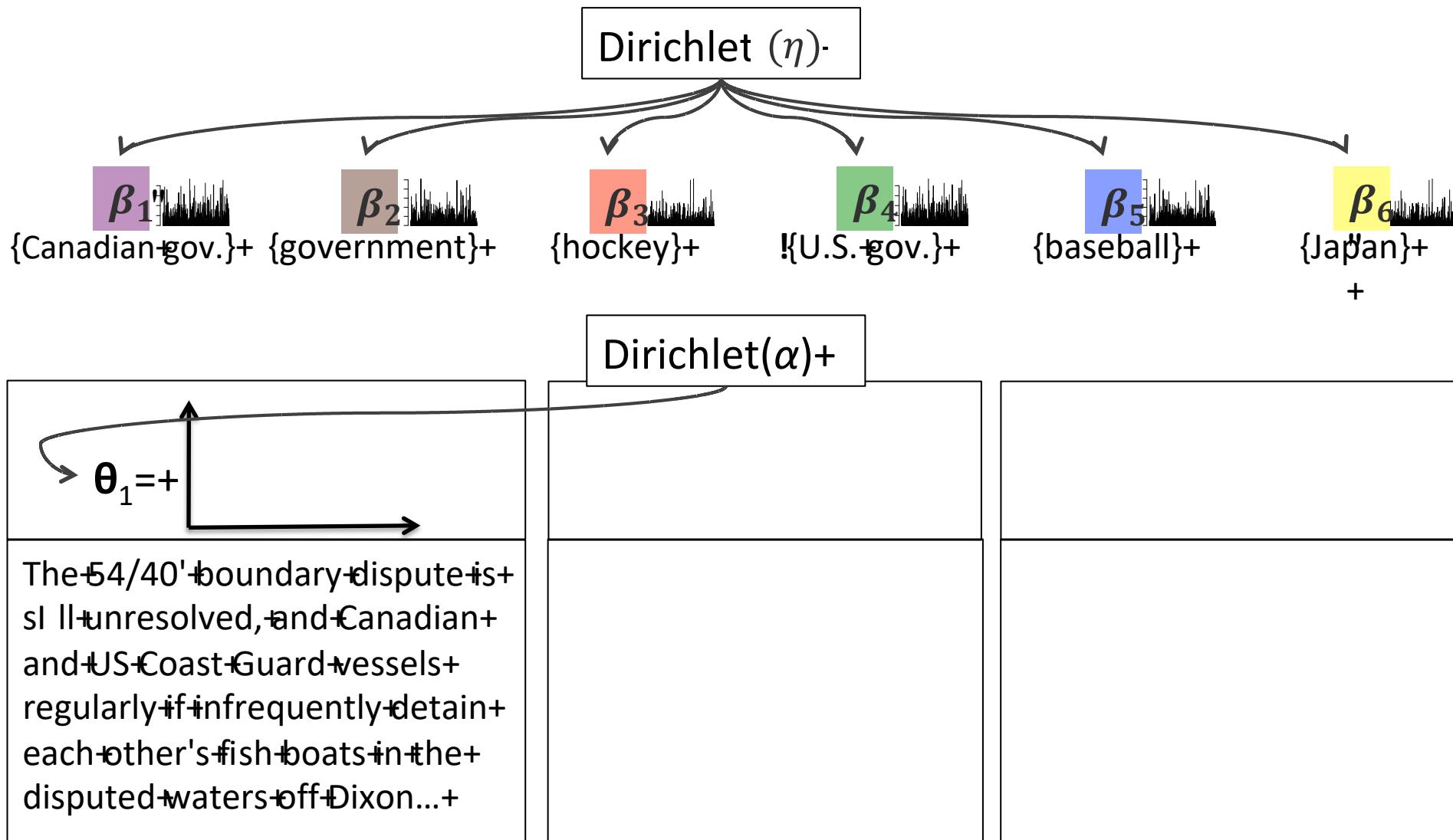
# LDA 'for' Topic Modeling"



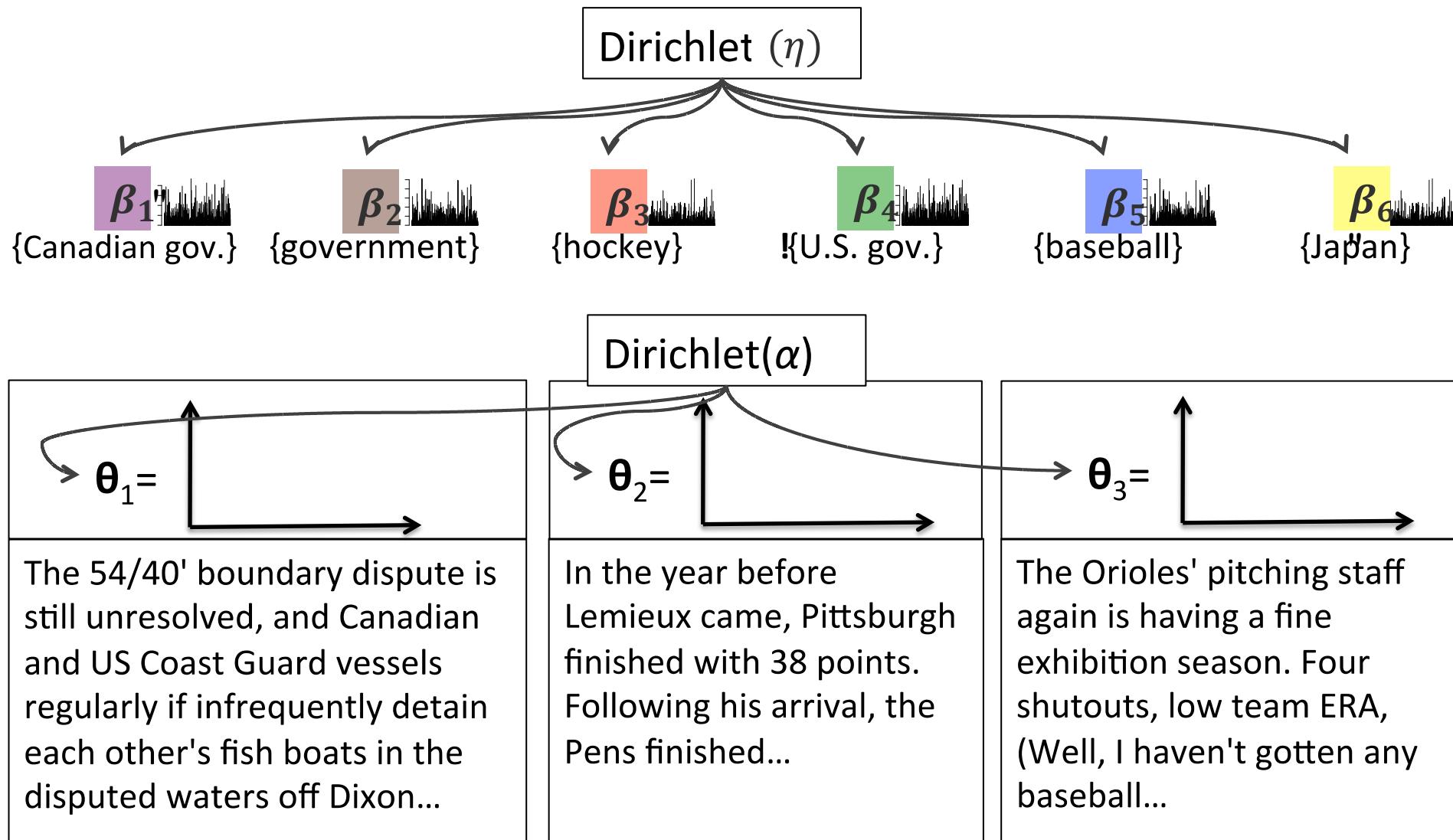
# LDA 'for' Topic Modeling'



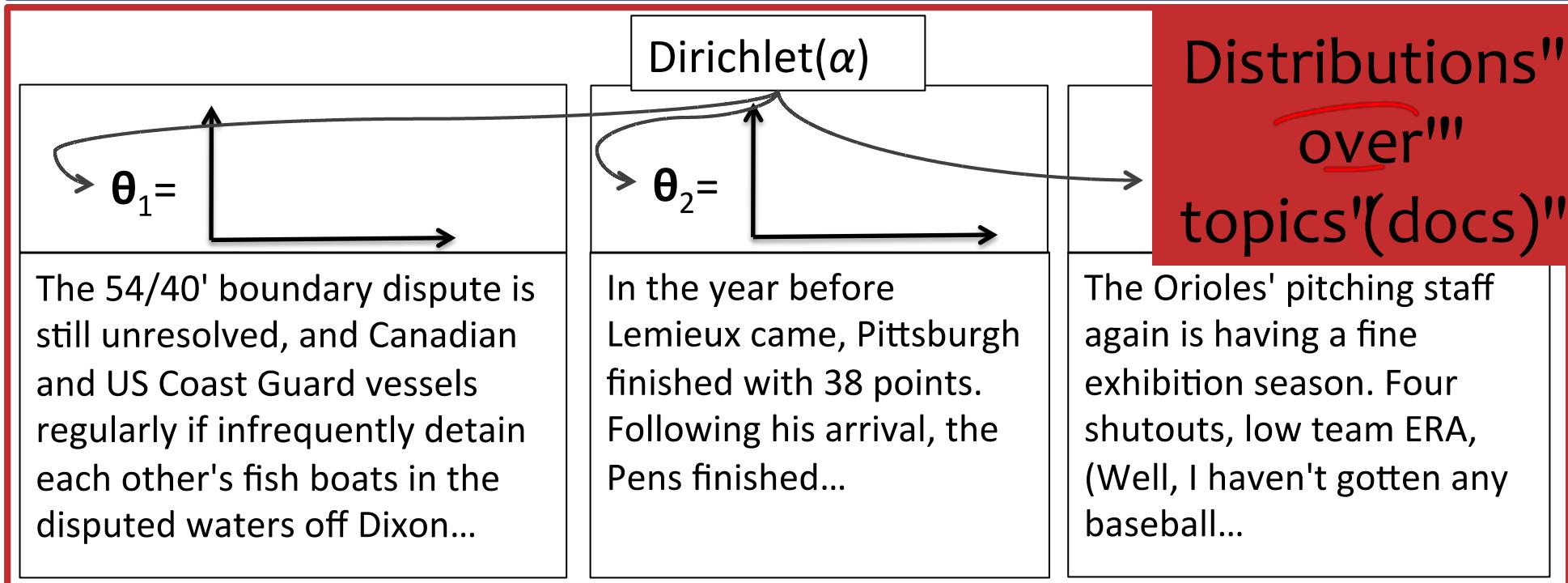
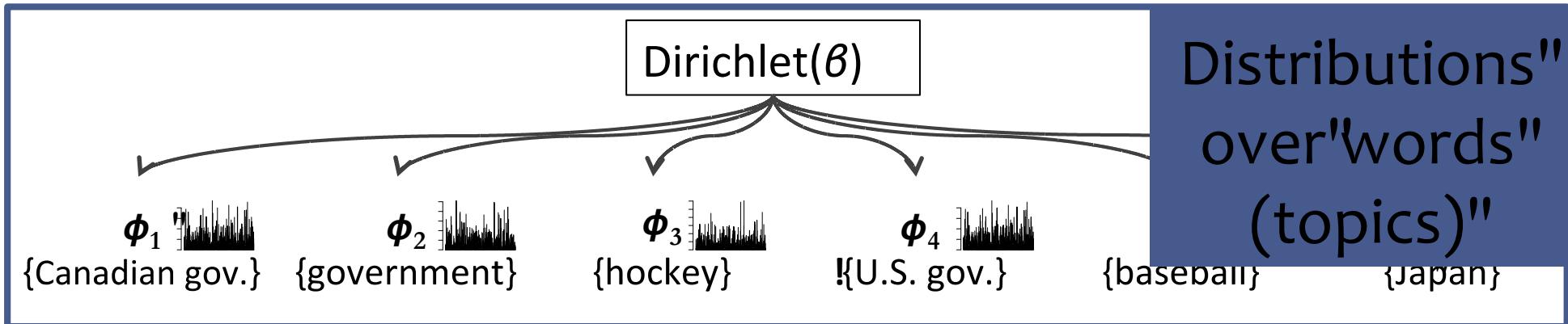
# LDA "for" Topic Modeling"



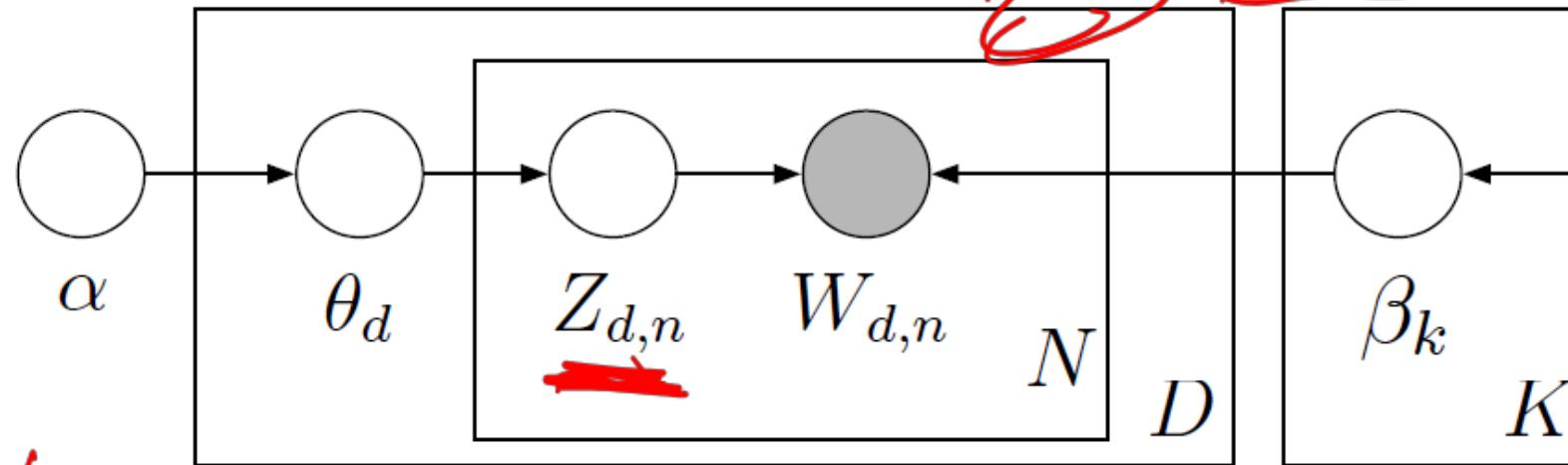
# LDA 'for' Topic Modeling"



# LDA 'for' Topic 'Modeling'



## Joint Distribution for LDA



$p(w|\alpha, \beta)$

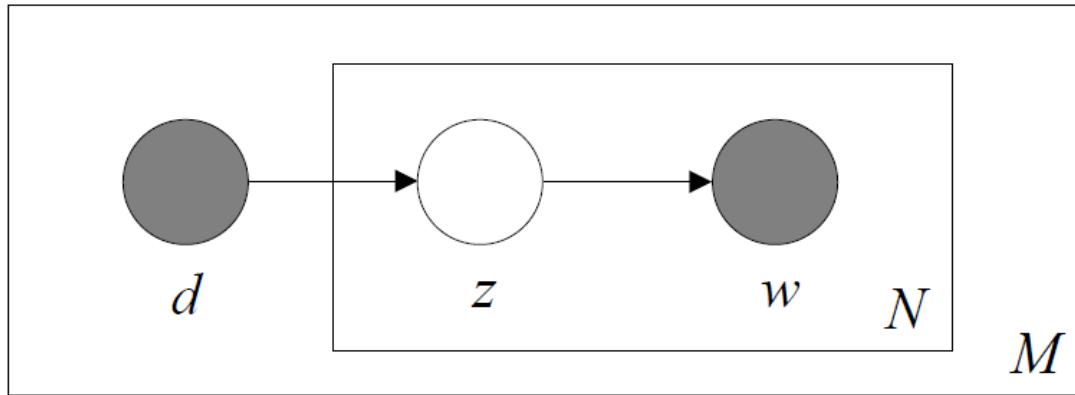
Joint distribution of latent variables and documents is:

$$p(\beta_{1:K}, z_{1:D}, \theta_{1:D}, w_{1:D} | \alpha, \eta) =$$

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

# Learning of Topic Models

# Recap: pLSA Topic Model



- Observed variables:  $d, w$  *(Unsupervised)*
  - Latent variables:  $z$  *(Supervised)*
  - Parameters:  $\theta, \beta$
- Obs. param:  $\theta, w, z$*
- param:  $\theta, \beta$*

# The General Unsupervised Learning Problem

- Each data instance is partitioned into two parts:
  - observed variables  $x$
  - latent (unobserved) variables  $z$
- Want to learn a model  $p_\theta(x, z)$

$$\max_{\theta} \log \sum_z p_\theta(x, z)$$

~~( $x, z$ )~~

~~Separable~~

~~max log P<sub>0</sub>( $x, z$ )~~

# Latent (unobserved) variables

- A variable can be unobserved (latent) because:
  - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
    - e.g., topic model, speech recognition models, ...

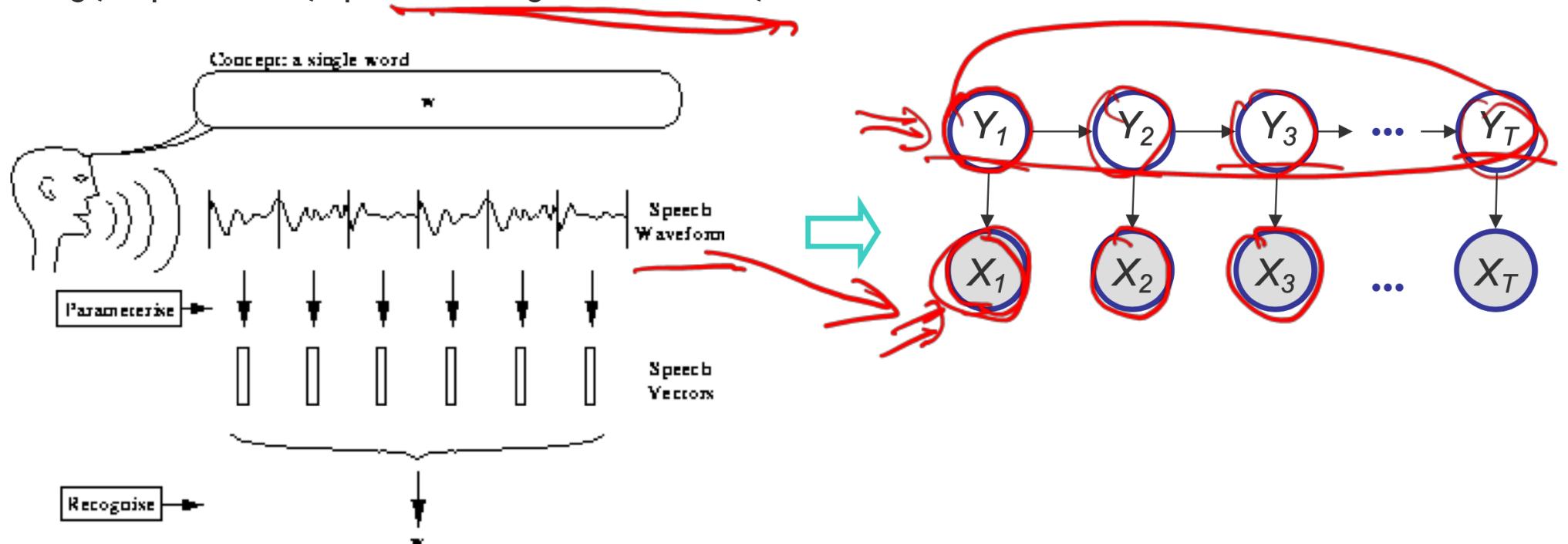


Fig. 1.2 Isolated Word Problem

# Latent (unobserved) variables

- A variable can be unobserved (latent) because:
  - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
    - e.g., topic model, speech recognition models, ...



# Latent (unobserved) variables

- A variable can be unobserved (latent) because:
  - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
    - e.g., topic model, speech recognition models, ...
  - a real-world object (and/or phenomena), but difficult or impossible to measure
    - e.g., the temperature of a star, causes of a disease, evolutionary ancestors ...
  - a real-world object (and/or phenomena), but sometimes wasn't measured, because of faulty sensors, etc.
- Discrete latent variables can be used to partition/cluster data into sub-groups
- Continuous latent variables (factors) can be used for dimensionality reduction (e.g., factor analysis, etc.)



# Questions?