DSC250: Advanced Data Mining

Topic Models

Zhiting Hu
Lecture 5, October 12, 2023

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Representations of Text and Topics
- Topic Model v1: Multinomial Mixture Model
- Topic Model v2: Probabilistic Latent Semantic Analysis (pLSA)
- Topic Model v3: Latent Dirichlet Allocation (LDA)

Slides adapted from:
- Y. Sun, CS 247: Advanced Data Mining
- M. Gormley, 10-701 Introduction to Machine Learning

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

Motivation

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

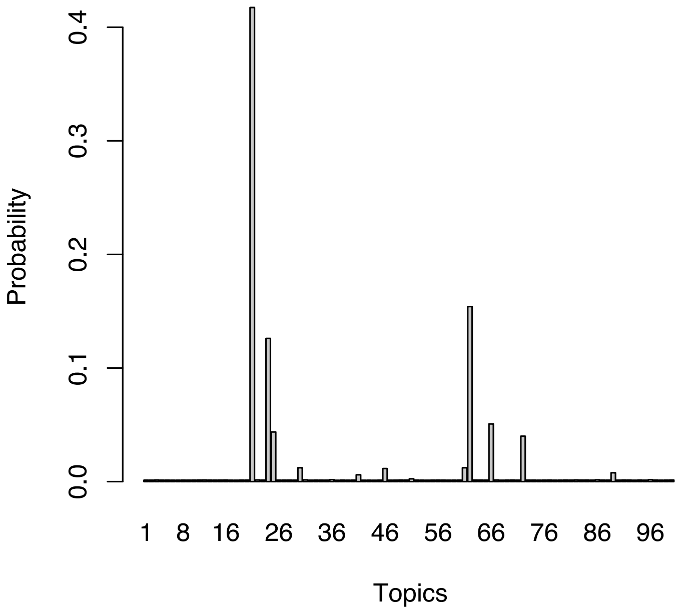
- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content

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: Examples



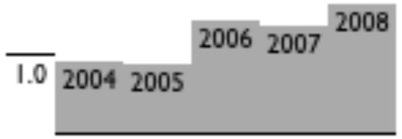
"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figure from (Blei, 2011), shows topics and top words learned automatically from reading 17,000 Science articles

Topic Modeling: Examples

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

Topic 0 [0.152]

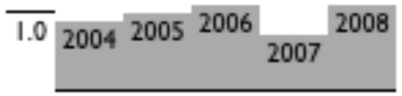


problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions

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]



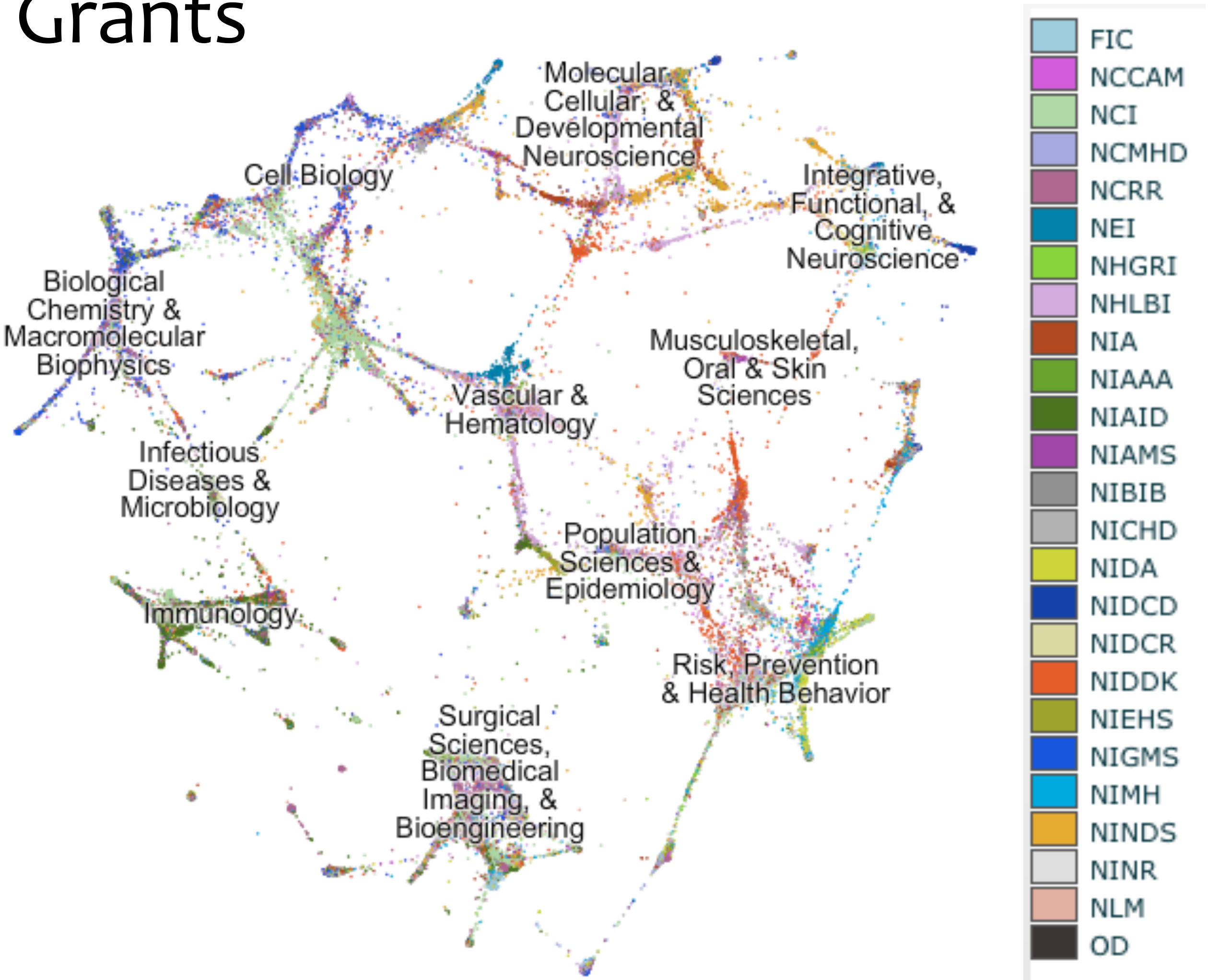
inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient

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

Topic Modeling: Examples

- Map of NIH Grants

(Talley et al., 2011)

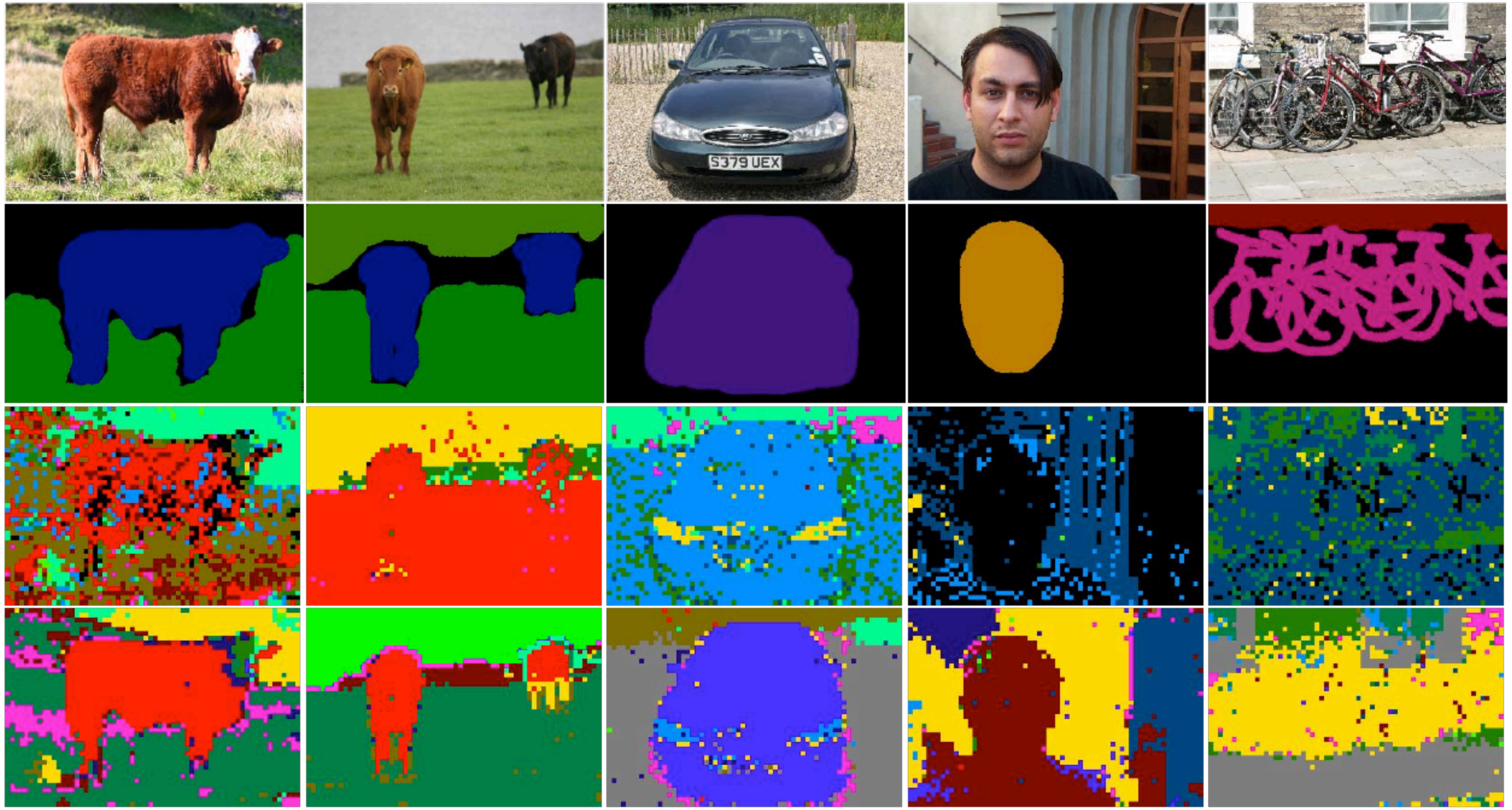


https://app.nihmaps.org/

Other Applications of Topic Models

- Spacial LDA

(Wang & Grimson, 2007)



Manual

LDA

SLDA

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

Text Data

- Word/term
- Document
  - A sequence of words
- Corpus
  - A collection of documents

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

Represent a Document

- Represent the doc as a vector where each entry corresponds to a different word and the number at that entry corresponds to how many times that word was present in the document (or some function of it)
  - Number of words is huge
  - Select and use a smaller set of words that are of interest
  - E.g. uninteresting words: ‘and’, ‘the’ ‘at’, ‘is’, etc. These are called stop-words
  - Stemming: remove endings. E.g. ‘learn’, ‘learning’, ‘learnable’, ‘learned’ could be substituted by the single stem ‘learn’
  - Other simplifications can also be invented and used
  - The set of different remaining words is called dictionary or vocabulary. Fix an ordering of the terms in the dictionary so that you can operate them by their index.
  - Can be extended to bi-gram, tri-gram, or so

Limitations of Bag-of-Words

- Dimensionality
  - High dimensionality
- Sparseness
  - Most of the entries are zero
- Shallow representation
  - The vector representation does not capture semantic relations between words

Ex: "Tom loves Kate."

Represent a Topic

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

universe	0.0439
galaxies	0.0375
clusters	0.0279
matter	0.0233
galaxy	0.0232
cluster	0.0214
cosmic	0.0137
dark	0.0131
light	0.0109
density	0.01

drug	0.0672
patients	0.0493
drugs	0.0444
clinical	0.0346
treatment	0.028
trials	0.0277
therapy	0.0213
trial	0.0164
disease	0.0157
medical	0.00997

cells	0.0675
stem	0.0478
human	0.0421
cell	0.0309
gene	0.025
tissue	0.0185
cloning	0.0169
transfer	0.0155
blood	0.0113
embryos	0.0111

sequence	0.0818
sequences	0.0493
genome	0.033
dna	0.0257
sequencing	0.0172
map	0.0123
genes	0.0122
chromosome	0.0119
regions	0.0119
human	0.0111

years	0.156
million	0.0556
ago	0.045
time	0.0317
age	0.0243
year	0.024
record	0.0238
early	0.0233
billion	0.0177
history	0.0148

bacteria	0.0983
bacterial	0.0561
resistance	0.0431
coli	0.0381
strains	0.025
microbiol	0.0214
microbial	0.0196
strain	0.0165
salmonella	0.0163
resistant	0.0145

male	0.0558
females	0.0541
female	0.0529
males	0.0477
sex	0.0339
reproductive	0.0172
offspring	0.0168
sexual	0.0166
reproduction	0.0143
eggs	0.0138

theory	0.0811
physics	0.0782
physicists	0.0146
einstein	0.0142
university	0.013
gravity	0.013
black	0.0127
theories	0.01
aps	0.00987
matter	0.00954

immune	0.0909
response	0.0375
system	0.0358
responses	0.0322
antigen	0.0263
antigens	0.0184
immunity	0.0176
immunology	0.0145
antibody	0.014
autoimmune	0.0128

stars	0.0524
star	0.0458
astrophys	0.0237
mass	0.021
disk	0.0173
black	0.0161
gas	0.0149
stellar	0.0127
astron	0.0125
hole	0.00824

Topic Models

- Topic modeling
  - Get topics automatically from a corpus
  - Assign documents to topics automatically
- Most frequently used topic models
  - pLSA
  - LDA

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

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 nth word in the vocabulary occurs in document d
- Word distribution for each topic (β_z)
  - β_{zw}: $p(w|z)$

Recap: Multinomial distribution

- Multinomial distribution
  - Discrete random variable $\boldsymbol{x}$ that takes one of M values $\{1, \dots, M\}$
  - $p(\boldsymbol{x} = i) \;=\; \pi_i, \qquad \sum_i \pi_i = 1$
  - Out of n independent trials, let k_i be the number of times $\boldsymbol{x} = i$ was observed
  - The probability of observing a vector of occurrences $\boldsymbol{k} = [k_1, \dots, k_M]$ is given by the *multinomial distribution* parametrized by $\boldsymbol{\pi}$

$$p(\mathbf{k}|\boldsymbol{\pi}, n) \;=\; p(k_1, \dots, k_m|\pi_1, \dots, \pi_m, n) \;=\; \frac{n!}{k_1!k_2!\dots k_m!}\prod_{i=1}\pi_i^{k_i}$$

  - E.g., describing a text document by the frequency of occurrence of every distinct word
  - For $n = 1$, a.k.a. categorical distribution
    - $p(\boldsymbol{x} = i \mid \boldsymbol{\pi}) \;=\; \pi_i$
    - In $\boldsymbol{k} = [k_1, \dots, k_M]$: $k_i = 1$, and $k_j = 0$ for all $j \neq i \quad \rightarrow \quad a.k.a.$, one-hot representation of i

Topic Model v1: Multinomial Mixture Model

- For documents with bag-of-words representation
  - $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \dots, x_{dN})$, x_{dn} is the number of words for nth word in the vocabulary
- Generative model

Topic Model v1: Multinomial Mixture Model

- For documents with bag-of-words representation
  - $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dN})$, x_{dn} is the number of words for nth word in the vocabulary
- Generative model

Formulating the statistical relationship between words, documents and latent topics as a generative process describing how documents are created

Topic Model v1: Multinomial Mixture Model

- For documents with bag-of-words representation
  - $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \dots, x_{dN})$, x_{dn} is the number of words for nth word in the vocabulary
- Generative model
  - For each document
    - Sample its cluster label $z \sim Categorical(\boldsymbol{\pi})$
      - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$, π_k is the proportion of jth cluster
      - $p(z = k) = \pi_k$
    - Sample its word vector $\boldsymbol{x}_d \sim multinomial(\boldsymbol{\beta}_z)$
      - $\boldsymbol{\beta}_z = (\beta_{z1}, \beta_{z2}, \dots, \beta_{zN})$, β_{zn} is the parameter associate with nth word in the vocabulary
      - $p(\boldsymbol{x}_d | z = k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}} \propto \prod_n \beta_{kn}^{x_{dn}}$

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 $z \sim Categorical(\boldsymbol{\pi})$
      - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$, π_k is the proportion of jth cluster
      - $p(z = k) = \pi_k$
    - Sample its word vector $\boldsymbol{x}_d \sim multinomial(\boldsymbol{\beta}_z)$
      - $\boldsymbol{\beta}_z = (\beta_{z1}, \beta_{z2}, \dots, \beta_{zN})$, β_{zn} is the parameter associate with nth word in the vocabulary
      - $p(\boldsymbol{x}_d | z = k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}} \propto \prod_n \beta_{kn}^{x_{dn}}$

Likelihood Function

Likelihood Function

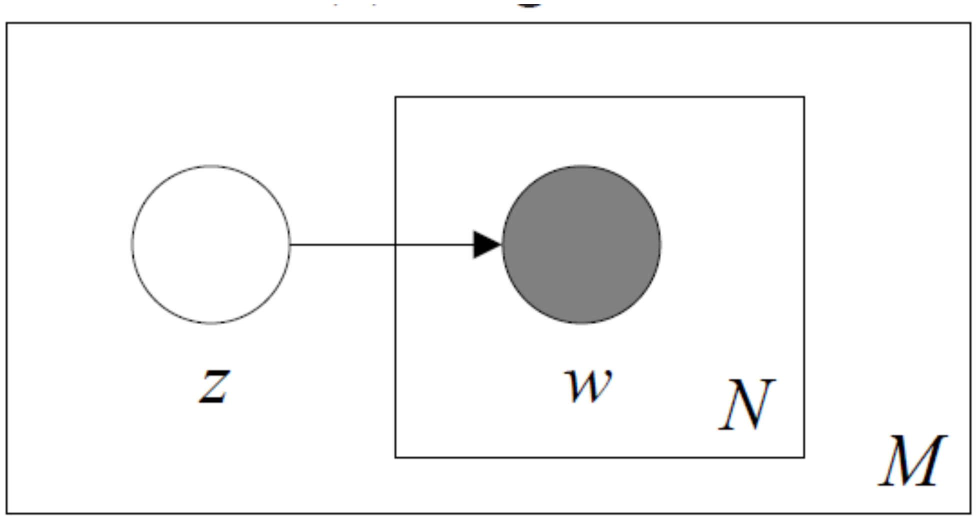
$$L = \prod_d p(\boldsymbol{x}_d) \; = \prod_d \sum_k p(\boldsymbol{x}_d, z = k)$$

$$= \prod_d \sum_k p(\boldsymbol{x}_d | z = k) p(z = k)$$

$$= \prod_d \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \sum_k p(z = k) \prod_n \beta_{kn}^{x_{dn}}$$

Limitations of Multinomial Mixture Model

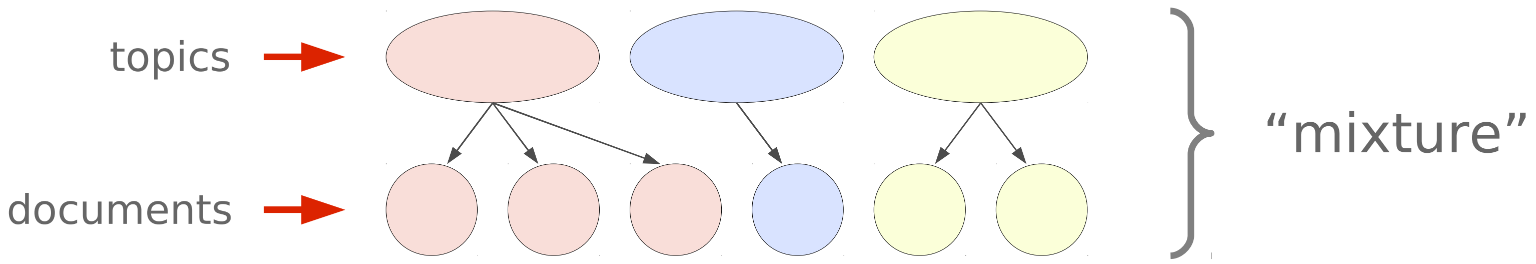
- All the words in the same documents are sampled from the same topic

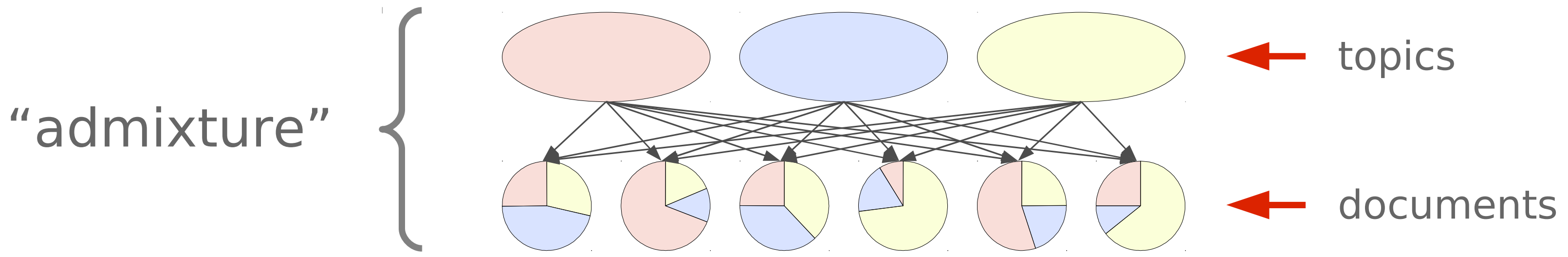


- In practice, people switch topics during their writing

Limitations of Multinomial Mixture Model

Mixture vs. Admixture





Diagrams from Wallach, JHU 2011, slides

Topic Model v2: Probabilistic Latent Semantic Analysis (pLSA)

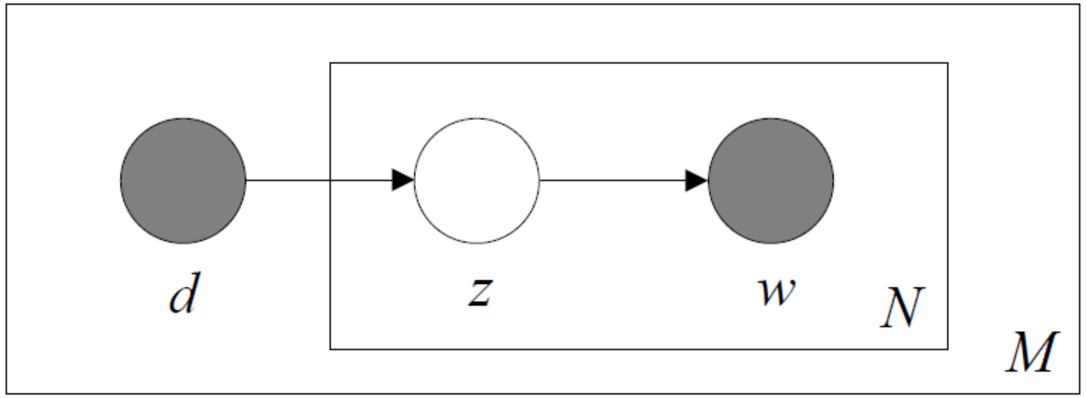
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Generative Model for pLSA

- For each position in d, $n = 1, \ldots, N_d$
  - Generate the topic for the position as
    $$z_n|d \sim Categorical(\boldsymbol{\theta}_d), i.e., p(z_n = k|d) = \theta_{dk}$$
    (Note, 1 trial multinomial)
  - Generate the word for the position as
    $$w_n|z_n \sim Categorical(\boldsymbol{\beta}_{z_n}), i.e., p(w_n = w|z_n) = \beta_{z_n w}$$

Graphical Model for pLSA



Note: Sometimes, people add parameters such as θ and β into the graphical model

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

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

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

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

$\boxed{\pi_d \textit{ is usually considered as uniform}\text{, i.e., 1/M}}$

Re-arrange the Likelihood Function

- Group the same word from different positions together

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

$$s.t. \sum_{z} \theta_{dz} = 1 \; and \; \sum_{w} \beta_{zw} = 1$$

Limitations of pLSA

- Not a proper generative model
  - $\boldsymbol{\theta}_d$ is treated as a parameter
  - Cannot model new documents

- Solution:
  - Make it a proper generative model by adding priors to θ and β

Limitations of pLSA

- Not a proper generative model
  - $\boldsymbol{\theta}_d$ is treated as a parameter
  - Cannot model new documents

- Solution:
  - Make it a proper generative model by adding priors to θ and β

⇩

Topic Model v3: Latent Dirichlet Allocation (LDA)

Review: Dirichlet Distribution

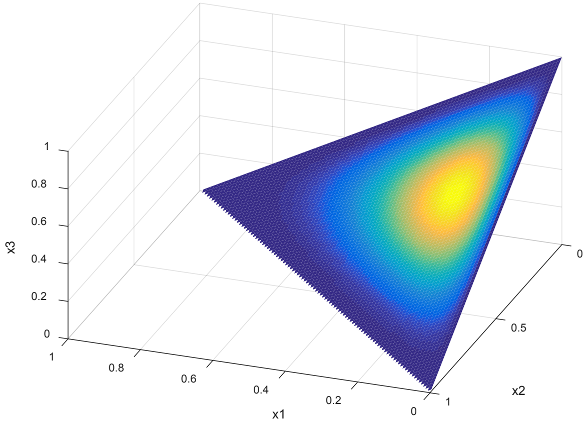
- Dirichlet distribution: $\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$
  - $i.e., p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$, where $\alpha_k > 0$
    - $\Gamma(\cdot)$ *is gamma function*: $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$
      - $\Gamma(z+1) = \mathrm{z}\Gamma(z)$

Review: Dirichlet Distribution

- Dirichlet distribution: $\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$
  - $i.e., p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$, where $\alpha_k > 0$
    - $\Gamma(\cdot)$ $is\ gamma\ function$: $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$
      - $\Gamma(z+1) = \mathrm{z}\Gamma(z)$

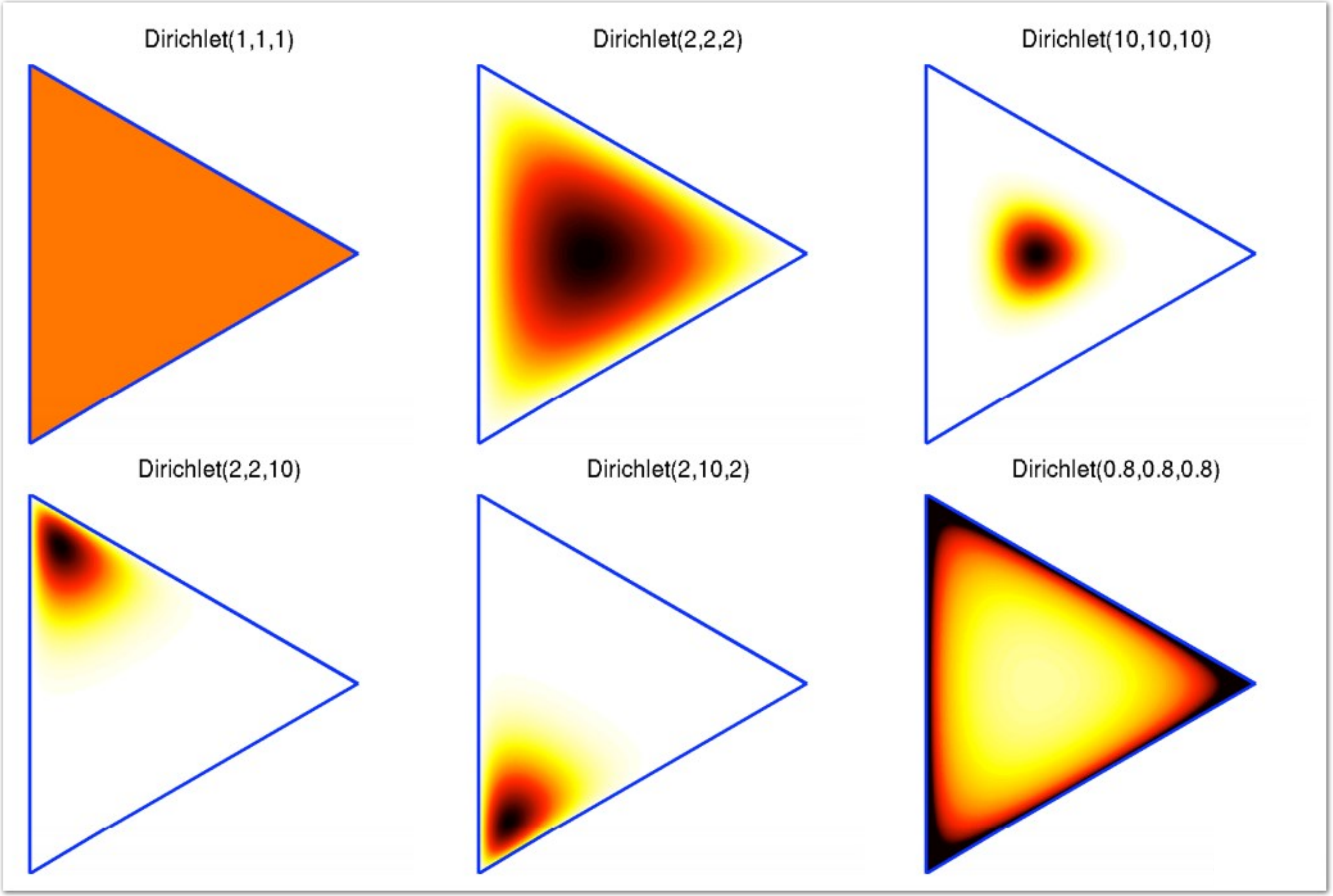
Simplex view:

- $x = x_1(1,0,0) + x_2(0,1,0) + x_3(0,0,1)$
  - Where $0 \leq x_1, x_2, x_3 \leq 1$ and $x_1 + x_2 + x_3 = 1$

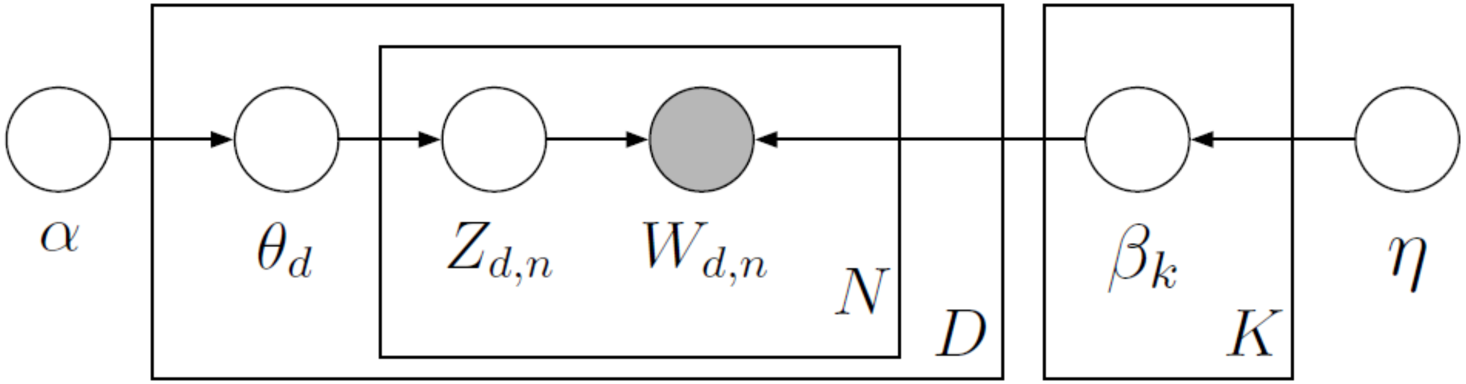


$x|\alpha \sim Dir(\alpha), \alpha = (2,3,4)$

More Examples in the Simplex View

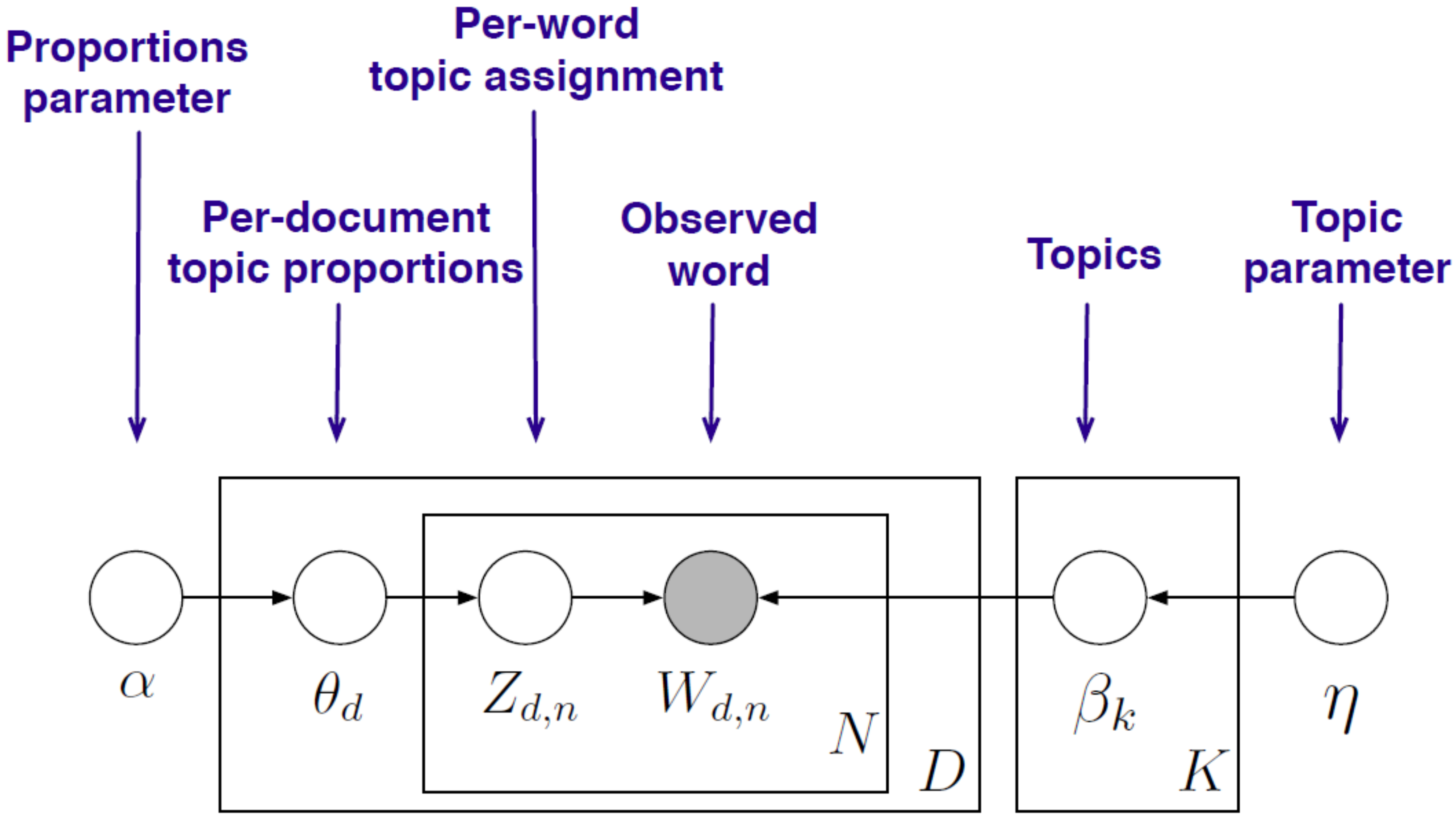


Topic Model v3: Latent Dirichlet Allocation (LDA)



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

Topic Model v3: Latent Dirichlet Allocation (LDA)

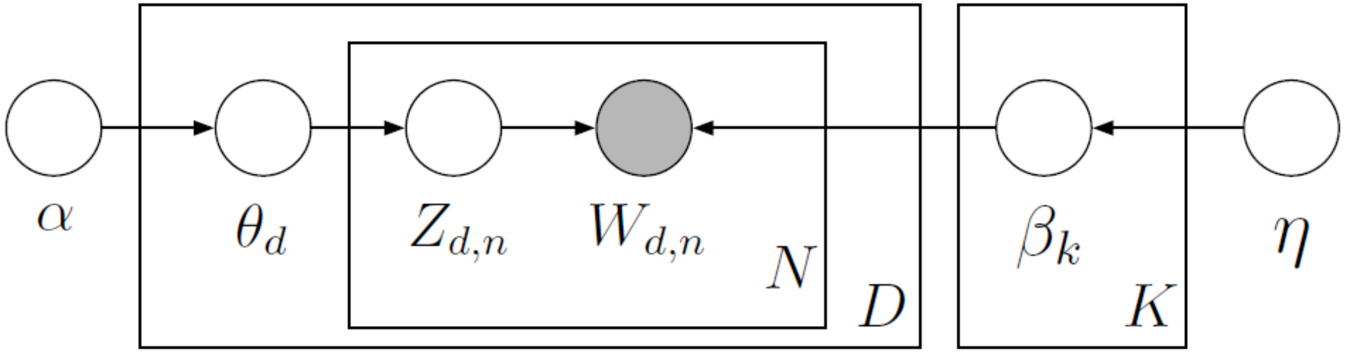


$\boldsymbol{\theta_d \sim Dirichlet(\alpha)}$: **address topic distribution for unseen documents**

$\boldsymbol{\beta_k \sim Dirichlet(\eta)}$: **smoothing over words**

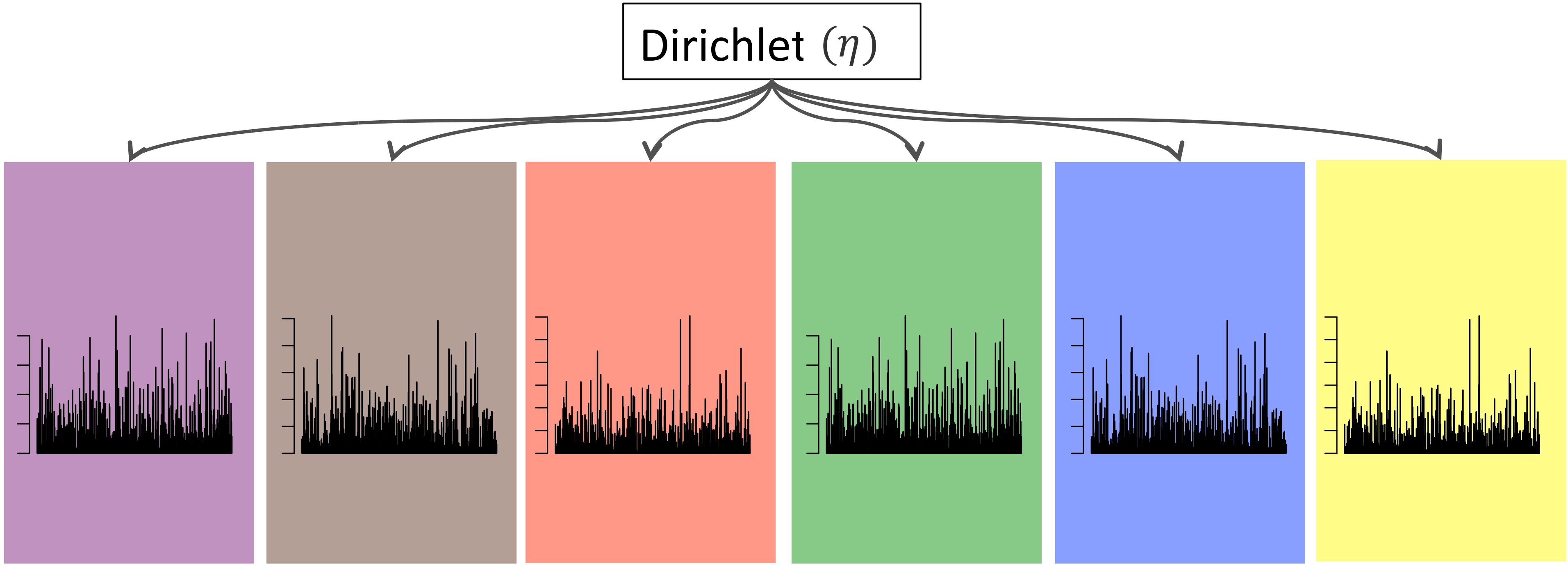
Generative Model for LDA

For each topic $k \in \{1, \ldots, K\}$:
  $\beta_k \sim \mathrm{Dir}(\eta)$ [*draw distribution over words*]
For each document $d \in \{1, \ldots, D\}$
  $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$ [*draw distribution over topics*]
  For each word $n \in \{1, \ldots, N_d\}$
    $z_{d,n} \sim \mathrm{Mult}(1, \boldsymbol{\theta}_d)$ [*draw topic assignment*]
    $w_{d,n} \sim \theta_{z_{d,n}}$ [*draw word*]



(Blei, Ng, & Jordan, 2003)

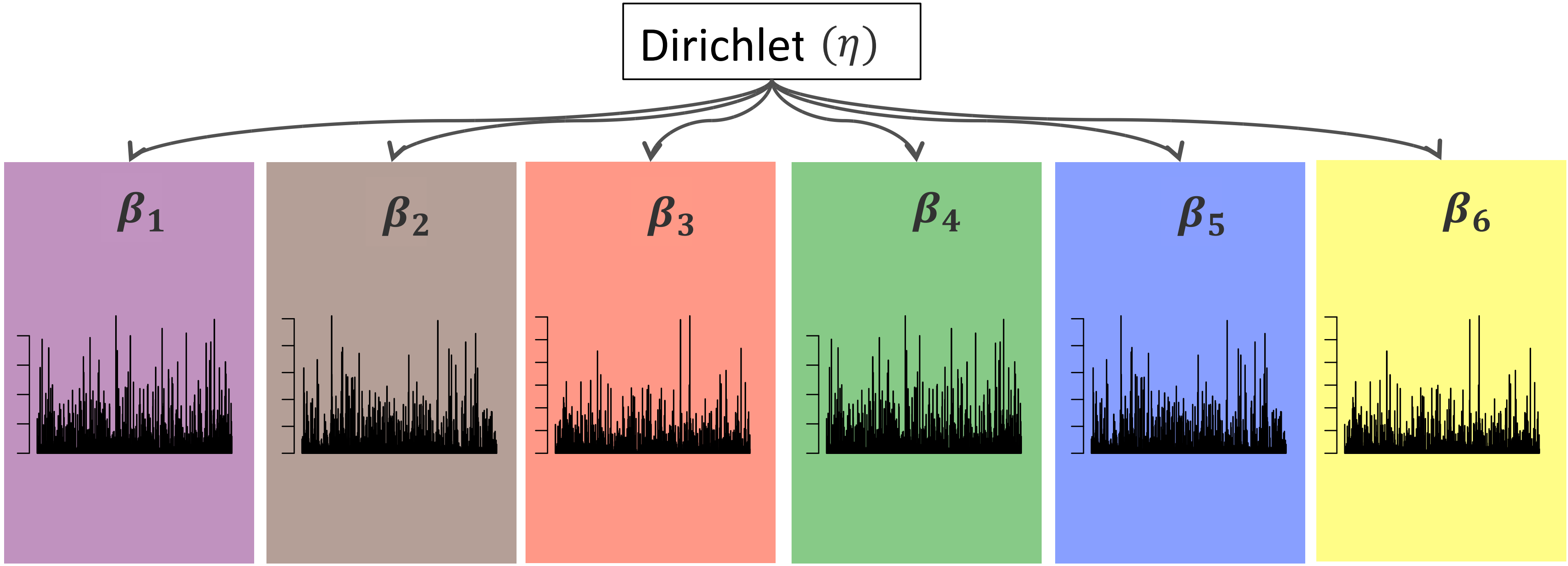
LDA for Topic Modeling



- 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 $\boldsymbol{\beta}_k$

(Blei, Ng, & Jordan, 2003)

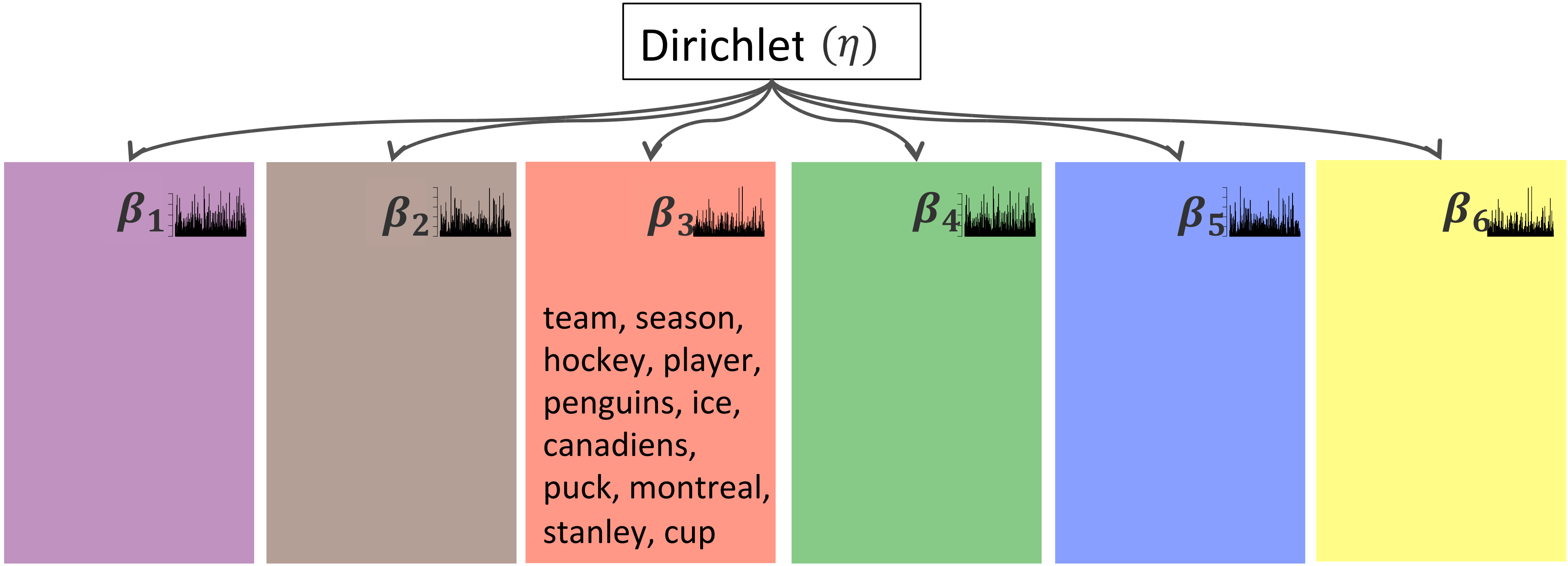
LDA for Topic Modeling



- 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 $\boldsymbol{\beta}_k$

(Blei, Ng, & Jordan, 2003)

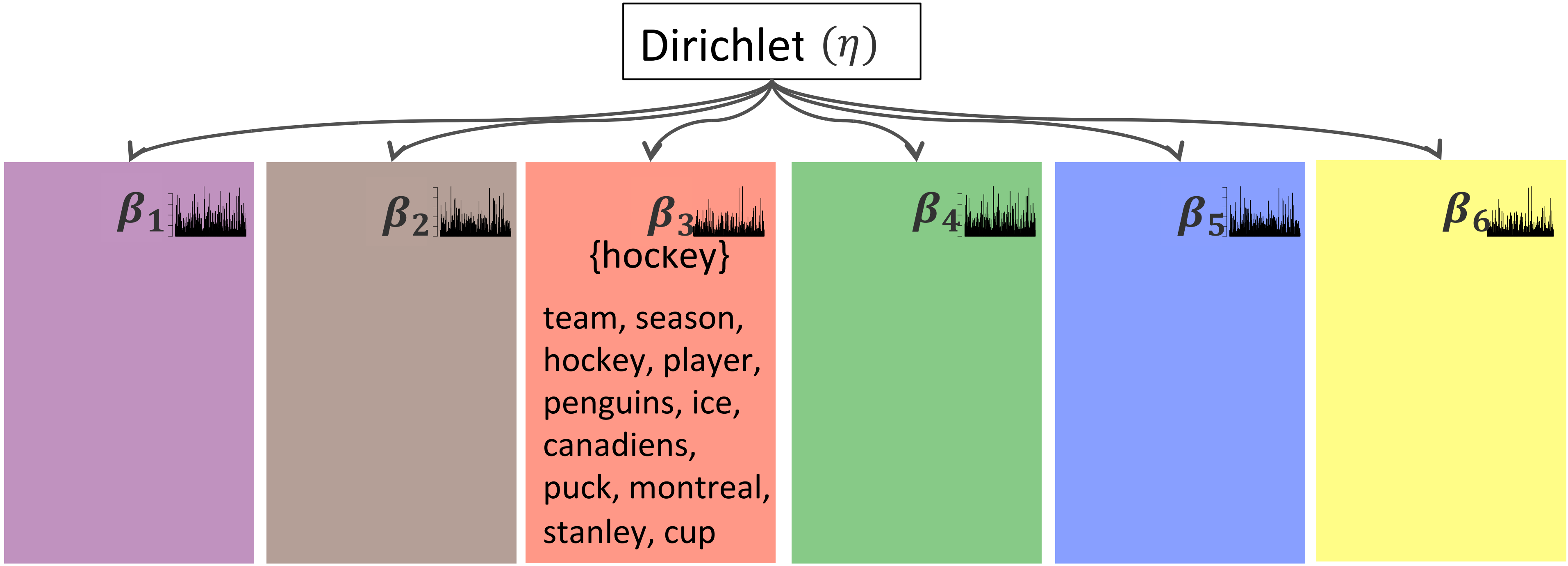
LDA for Topic Modeling



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

(Blei, Ng, & Jordan, 2003)

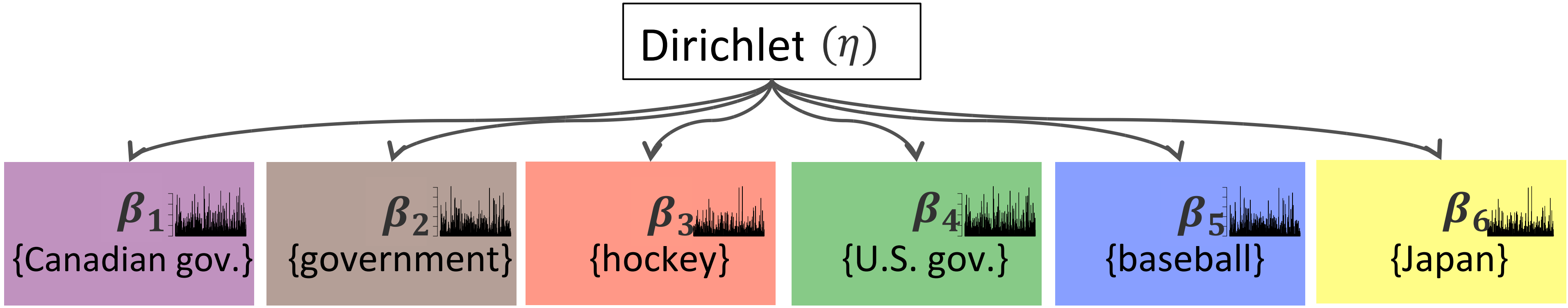
LDA for Topic Modeling



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

(Blei, Ng, & Jordan, 2003)

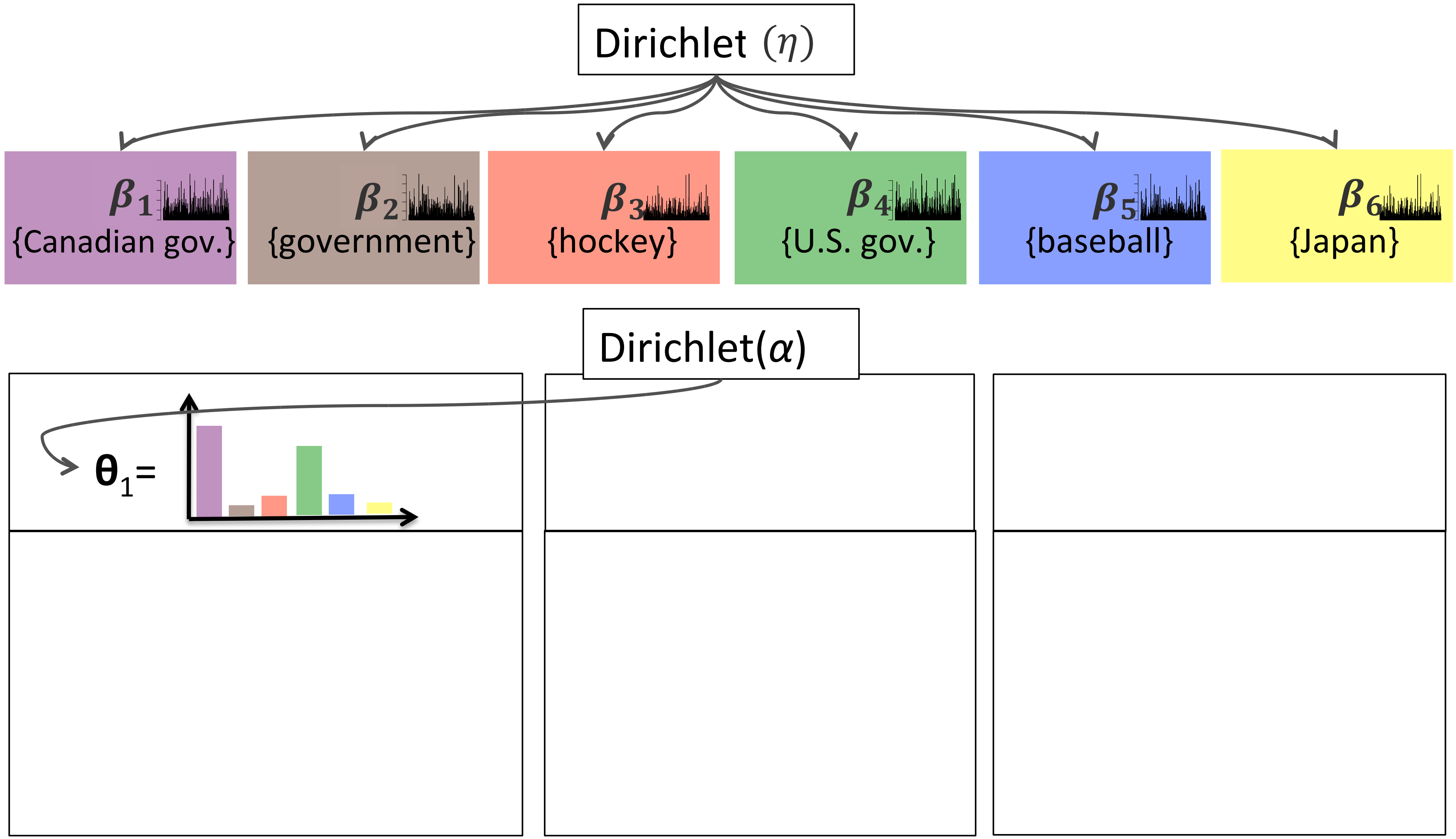
LDA for Topic Modeling



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

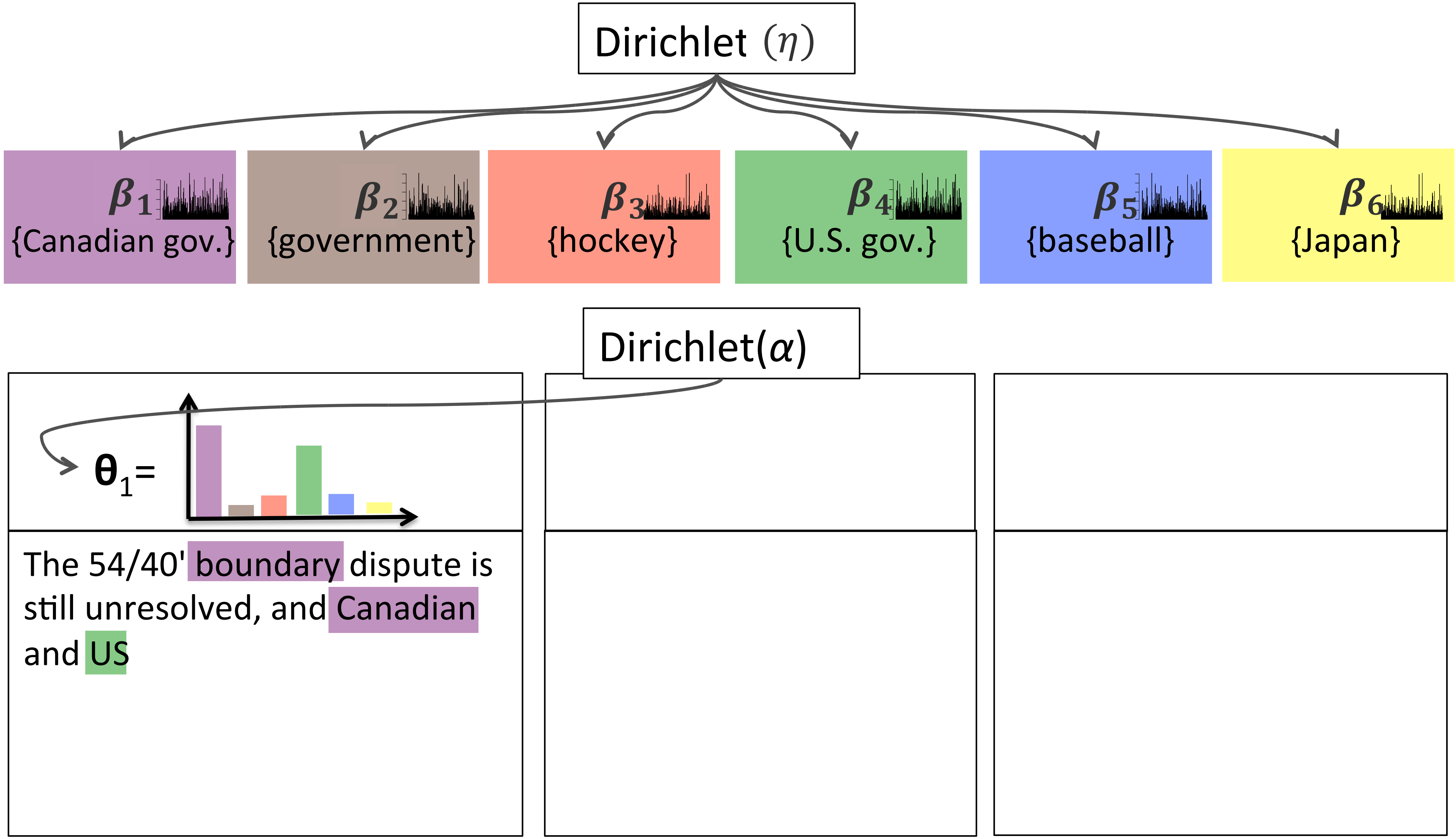
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



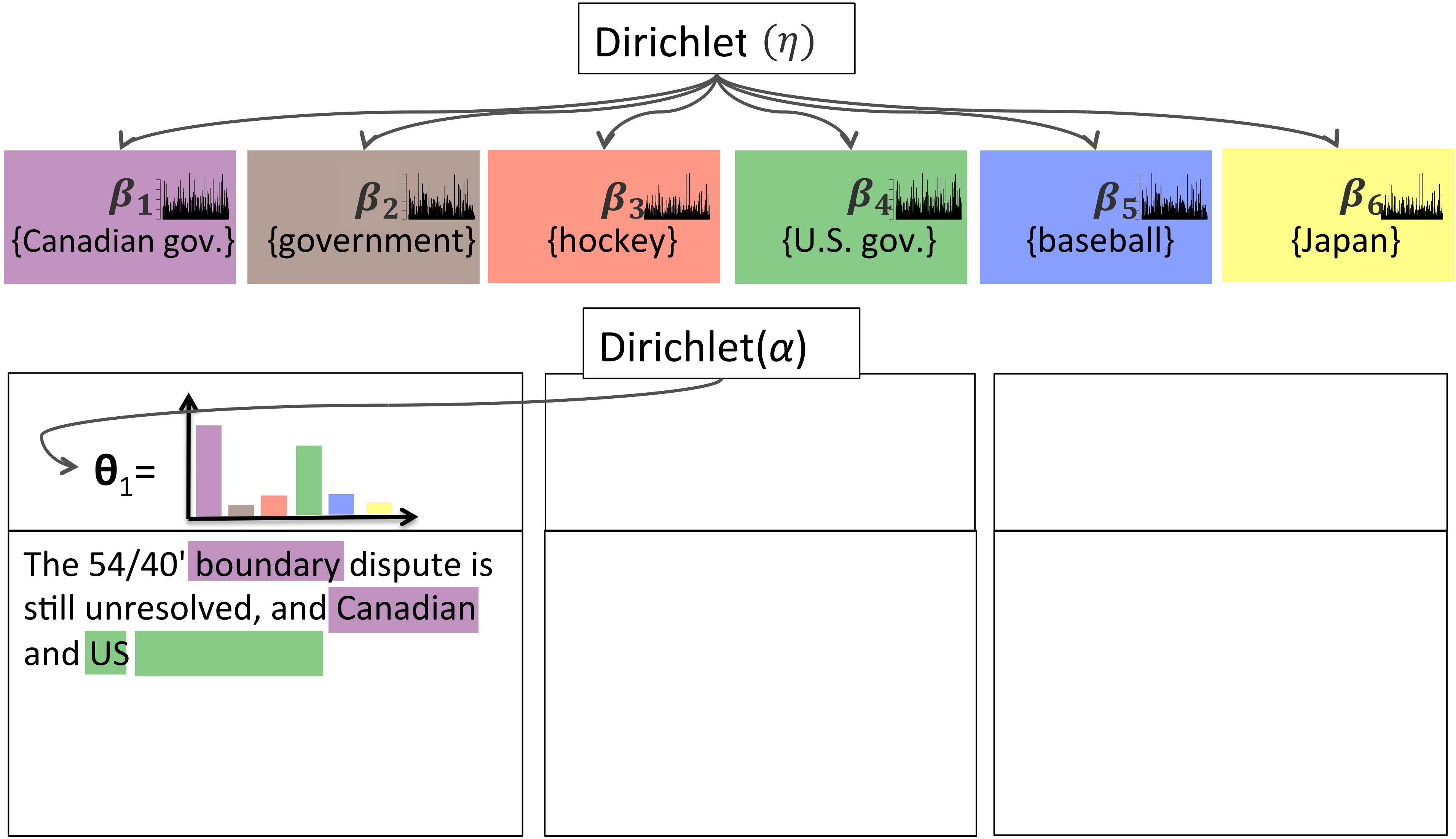
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



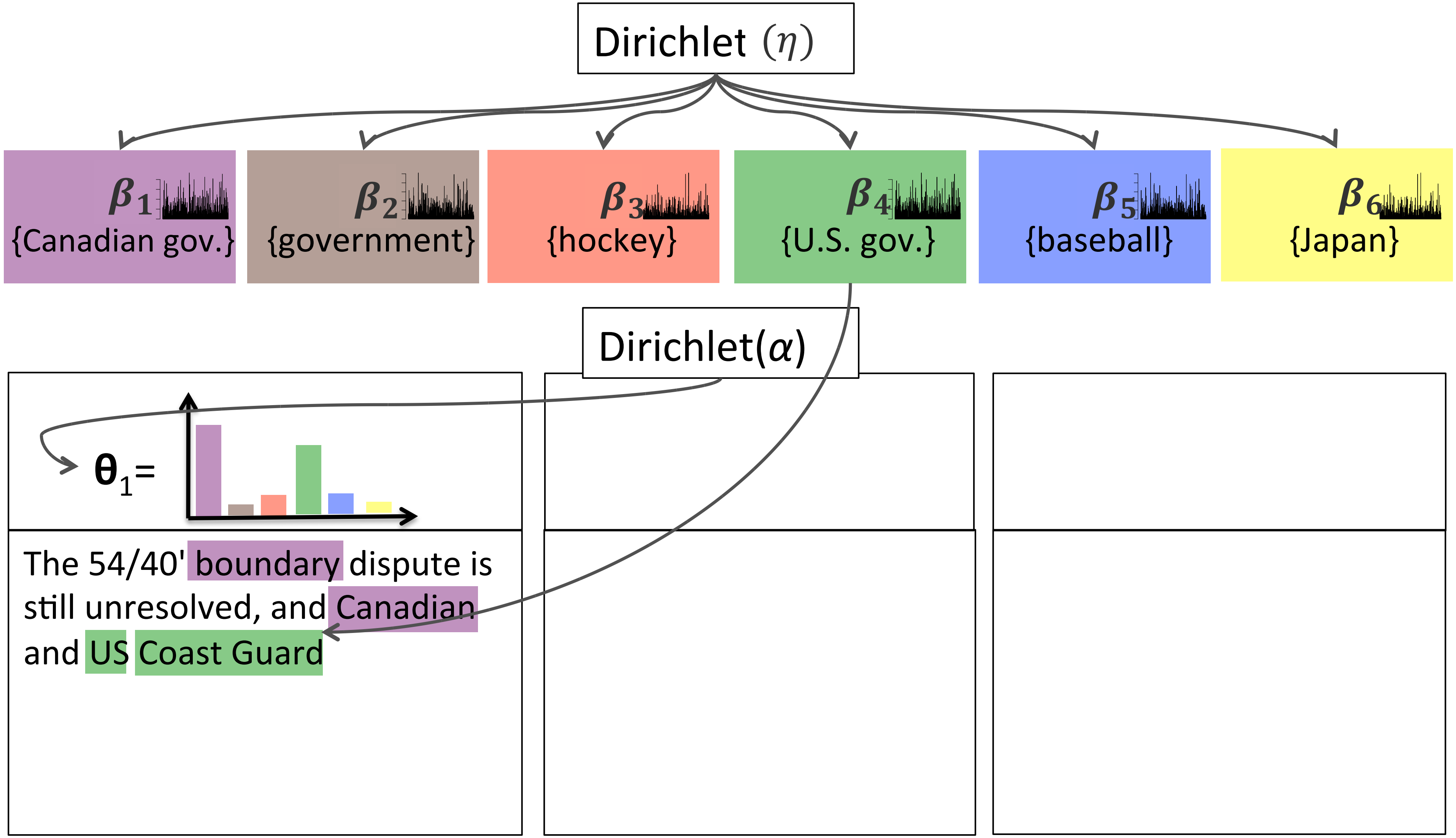
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



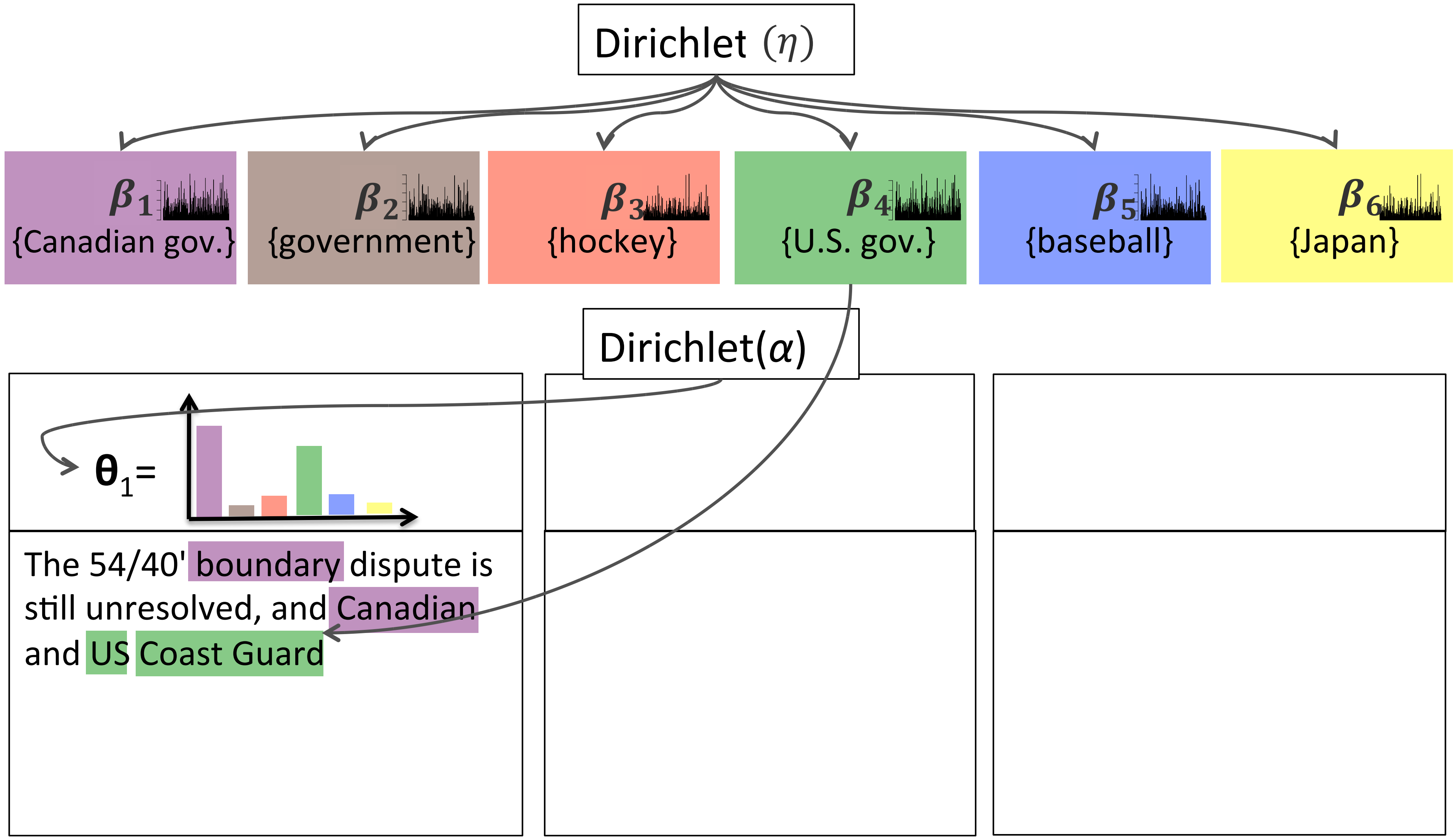
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



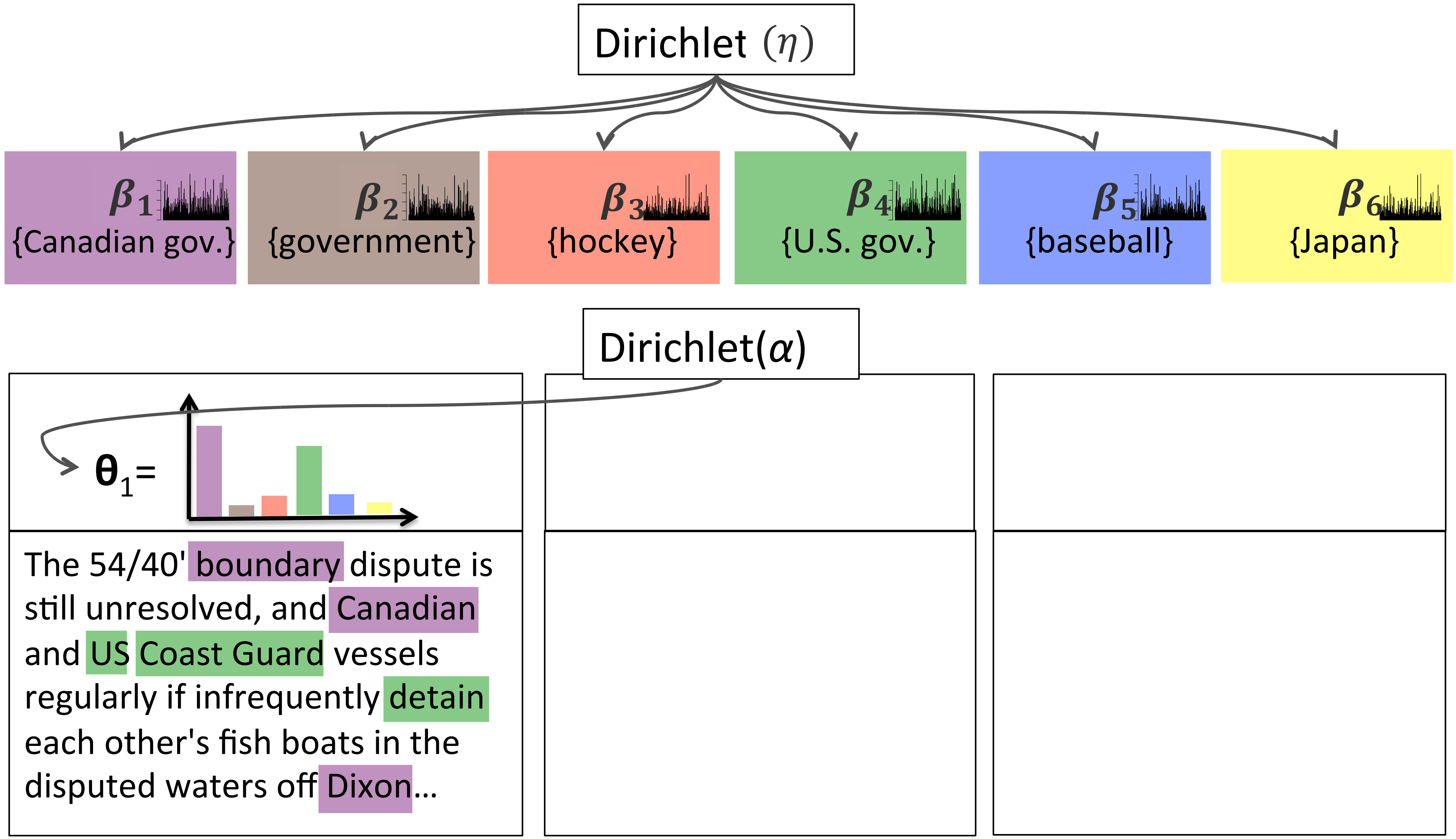
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



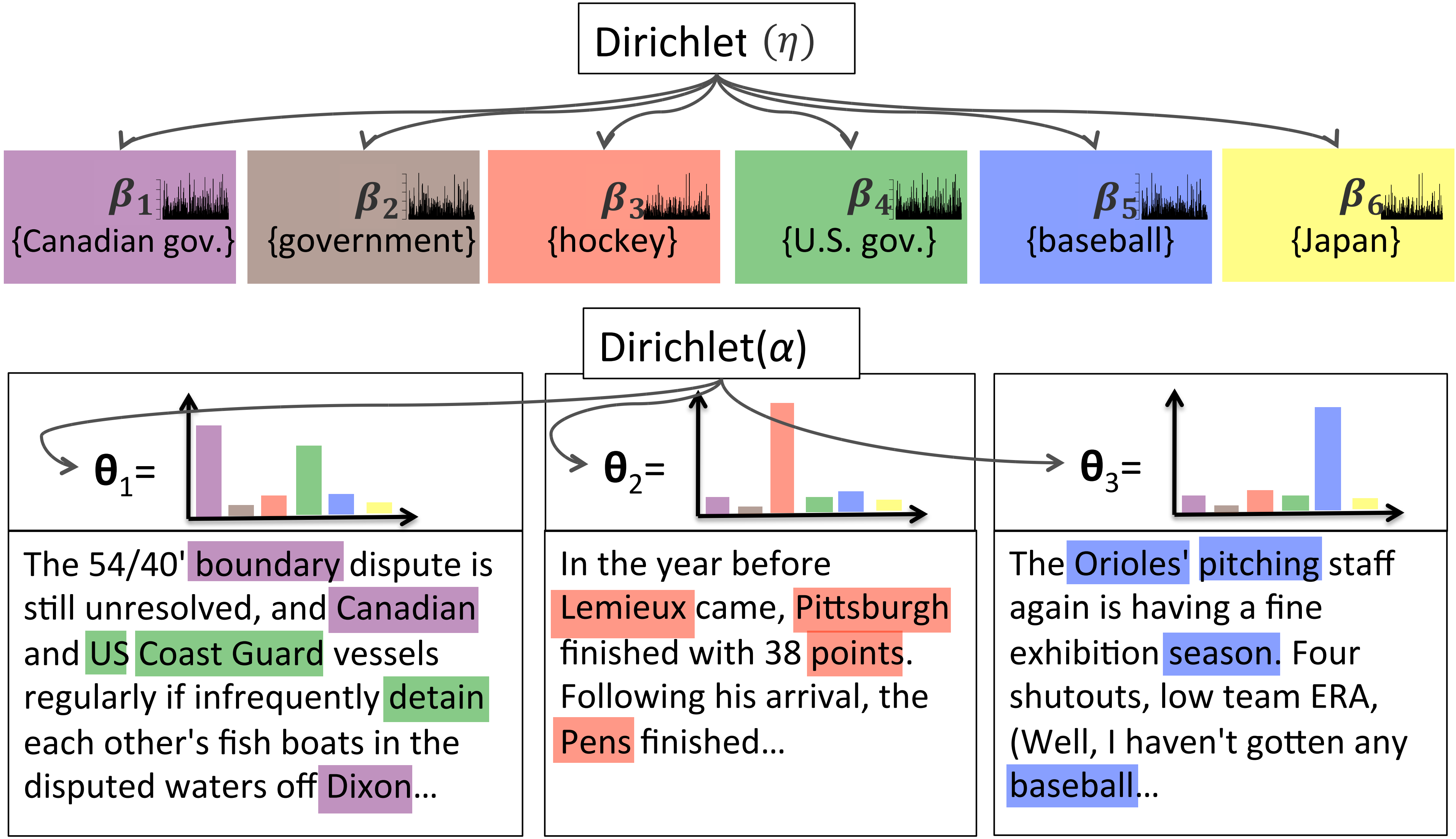
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



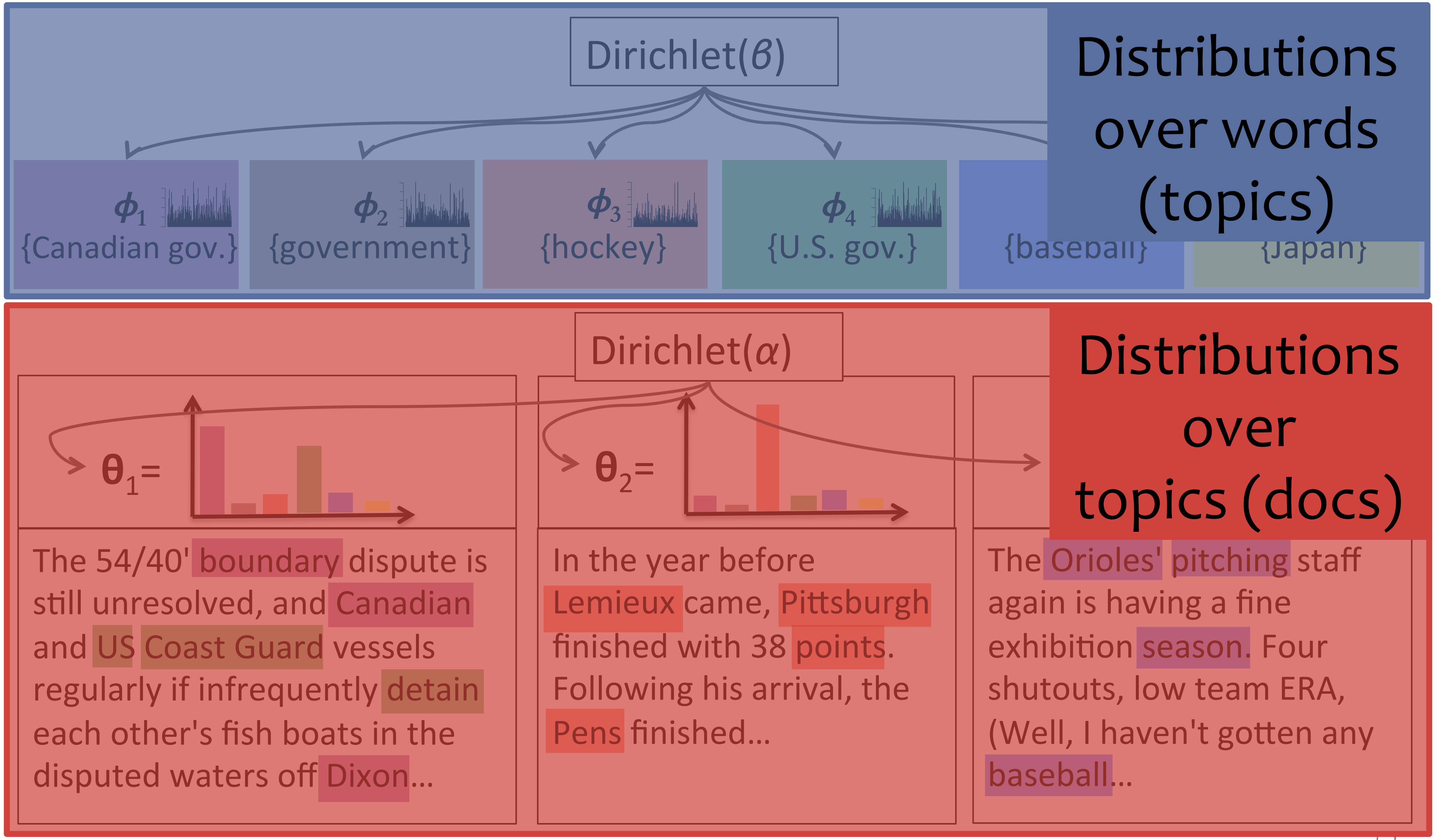
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling

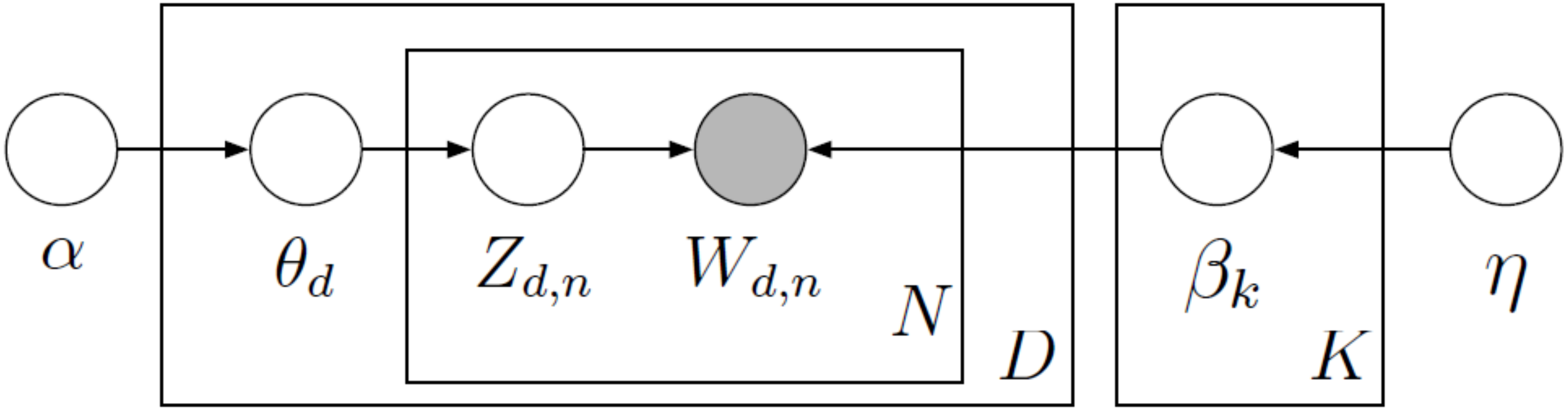


(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



Joint Distribution for LDA



- Joint distribution of latent variables and documents is:
  $p(\boldsymbol{\beta}_{1:K}, \boldsymbol{z}_{1:D}, \boldsymbol{\theta}_{1:D}, \boldsymbol{w}_{1:D} | \alpha, \eta) =$

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

Questions?