Dall-E

Topic Models

Applied Machine Learning
Derek Hoiem

Last Class: Dimensionality Reduction

- PCA finds the linear basis projection that maximizes variance
- UMAP solves for an embedding that preserves local and, to some extent, global structure of samples

1. When might PCA be a better choice than UMAP?

2. When might UMAP be a better choice than PCA?

Today's Class: Topic Modeling

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

Slide: Gormley, CMU 10-701, Lec 20, 2016

Today's Class: Topic Modeling

Represent documents as a collection of words, where only the count of words matters ("bag of words" model)

- Latent Semantic Analysis: Based on word-document co-occurrence, solve for a continuous vector to represent each word and document
- Latent Dirichlet Allocation: Model topics as a distribution of words, and documents as a distribution of topics
- BertTopic: Use deep learning to encode words or sentences and then apply topic modeling

Latent Semantic Analysis

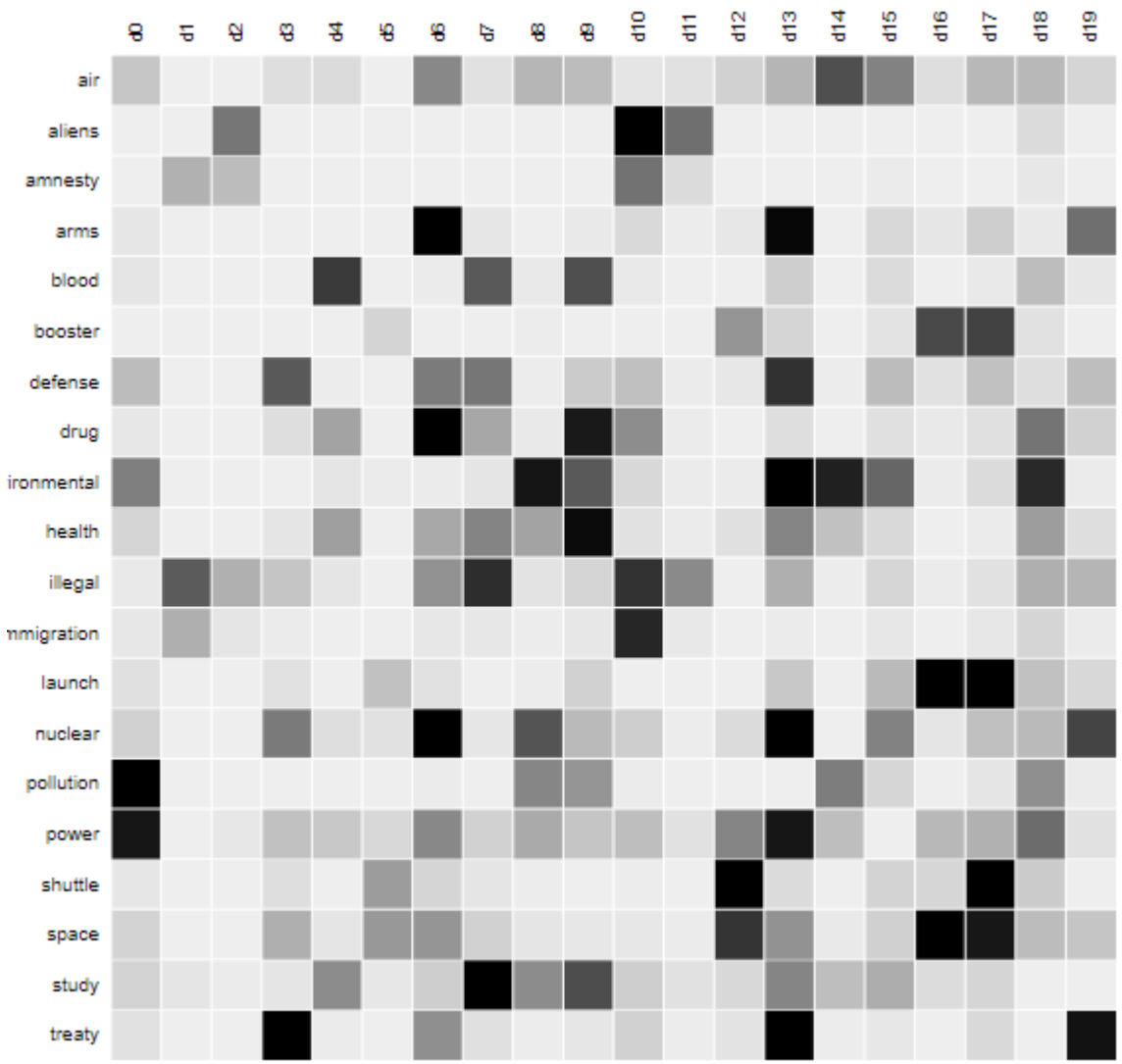
Goal: Given a collection of documents that each contains a set of terms, learn representations for documents and terms that can be used to classify, retrieve, or find relationships

Ordering of terms is not considered

Latent Semantic Analysis

1. Form a term-document matrix M of tf-idf score of each term in each document

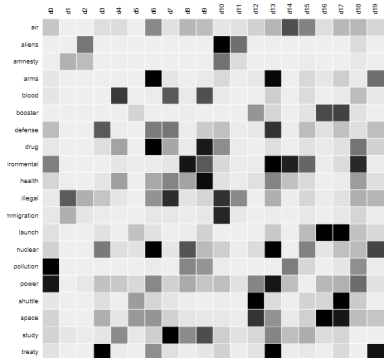
$$W_{t,d} = tf_{t,d} \cdot \log(\frac{N}{df_t})$$

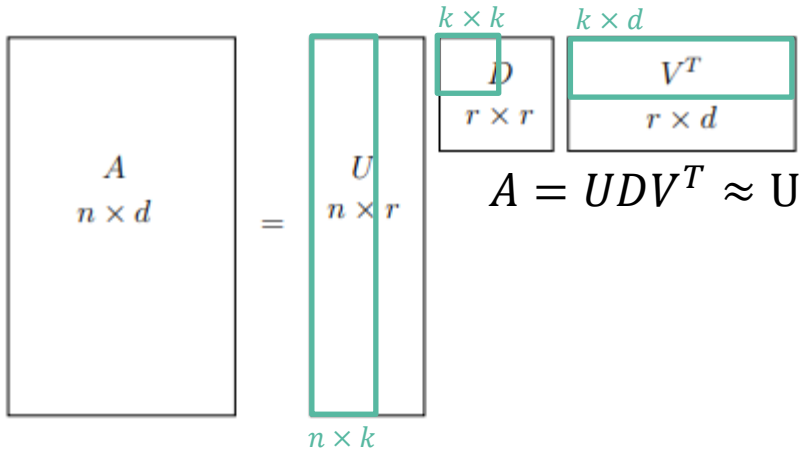


https://topicmodels.west.uni-koblenz.de/ckling/tmt/svd_ap.html

Latent Semantic Analysis

1. Form a term-document matrix M of tf-idf score of each term in each document
2. Decompose and compress the matrix using SVD and keeping only k singular values



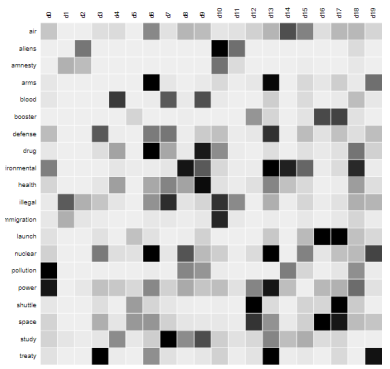


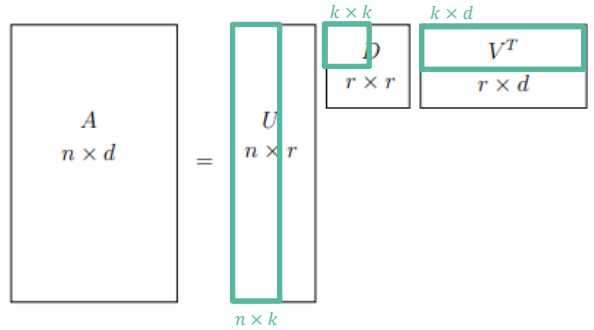
$$A = UDV^T \approx \mathrm{U}[:, :\mathrm{k}]\mathrm{D}[:\mathrm{k}, :\mathrm{k}]V^T[:\mathrm{k},:]$$

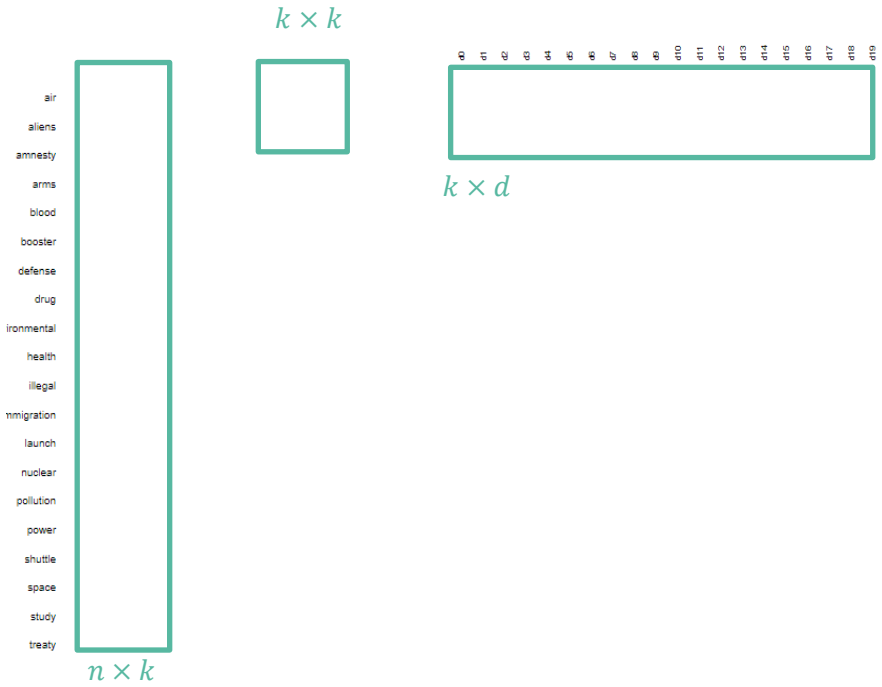
This gives the same approximation to A that you would get from PCA. If X is centered, the right columns of V are the principal components.

Latent Semantic Analysis

1. Form a term-document matrix M of tf-idf score of each term in each document
2. Compress the matrix using SVD by keeping only k singular values
3. Rows of $\Sigma_k U_k$ represent terms, and cols of $\Sigma_k V_k^T$ represent documents

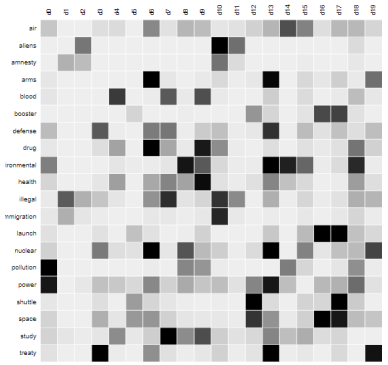


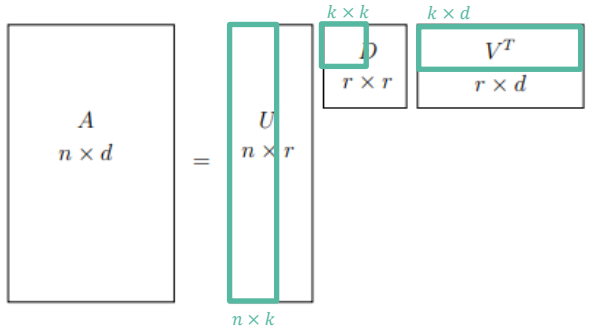




Latent Semantic Analysis

1. Form a term-document matrix M of tf-idf score of each term in each document
2. Compress the matrix using SVD by keeping only k singular values
3. Rows of $\Sigma_k U_k$ represent terms, and cols of $\Sigma_k V_k^T$ represent documents
4. Now we have a continuous "semantic space" (per-word and per-document vectors) for
    * Computing similarity
    * Retrieval
    * Classification
    * 2D embedding map
    * Discovery of relations between terms or documents



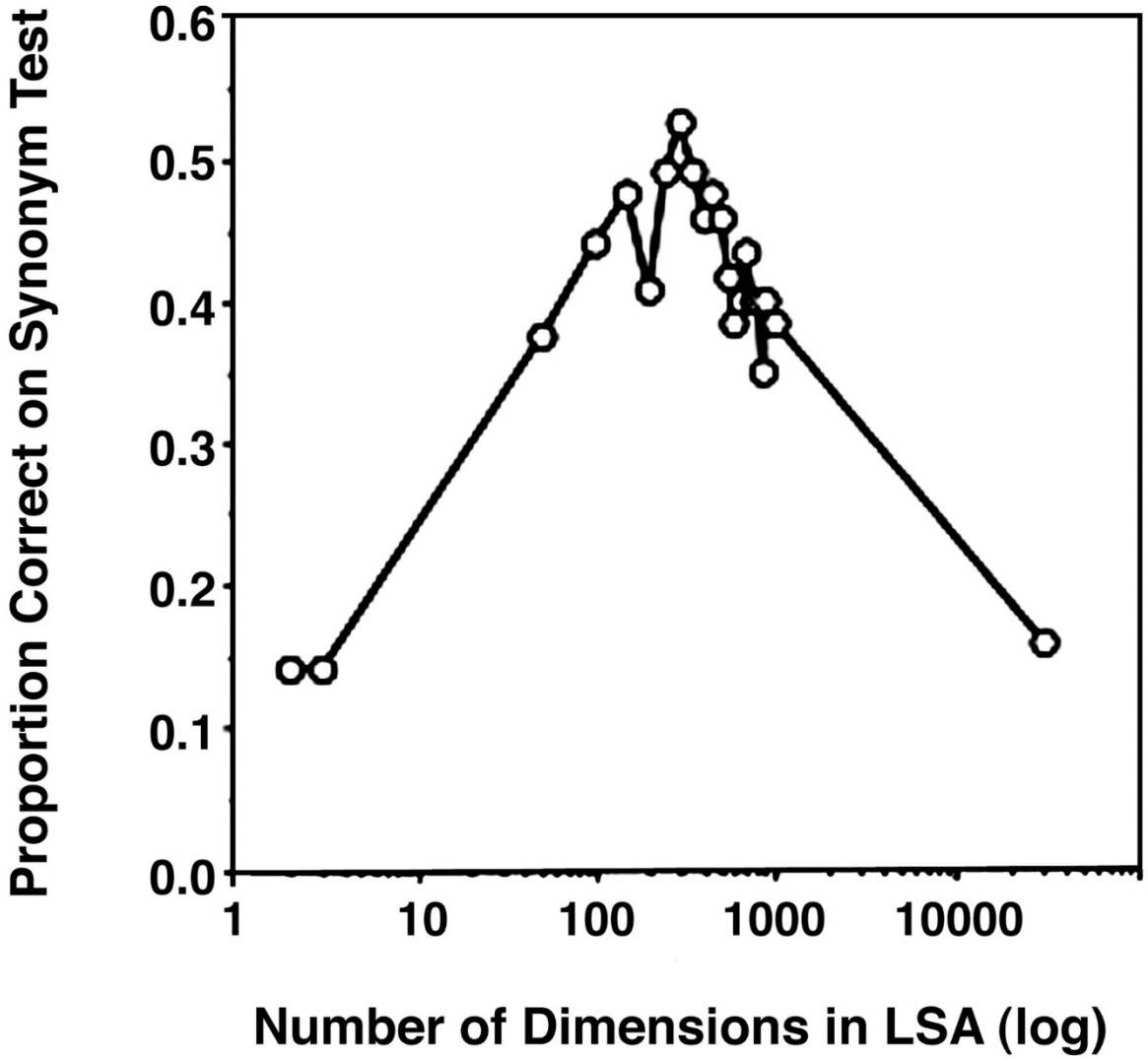


Let $\boldsymbol{t}_i = \mathrm{D}_k U_k[i]$ and $\boldsymbol{d}_j = \mathrm{D}_k V_k[j]$ be vectors representing the ith term and jth document

Similarity between documents: $\boldsymbol{d}_j \cdot \boldsymbol{d}_k$

A new document $\boldsymbol{d}_q$ can be mapped into the latent space by $\mathrm{D}_k^{-1} U_k^T \boldsymbol{d}_q$, and then it can be compared to other documents in this space

Use of LSA to identify synonyms (similar terms)



https://www.pnas.org/doi/10.1073/pnas.0400341101

Use of LSA to organize science articles

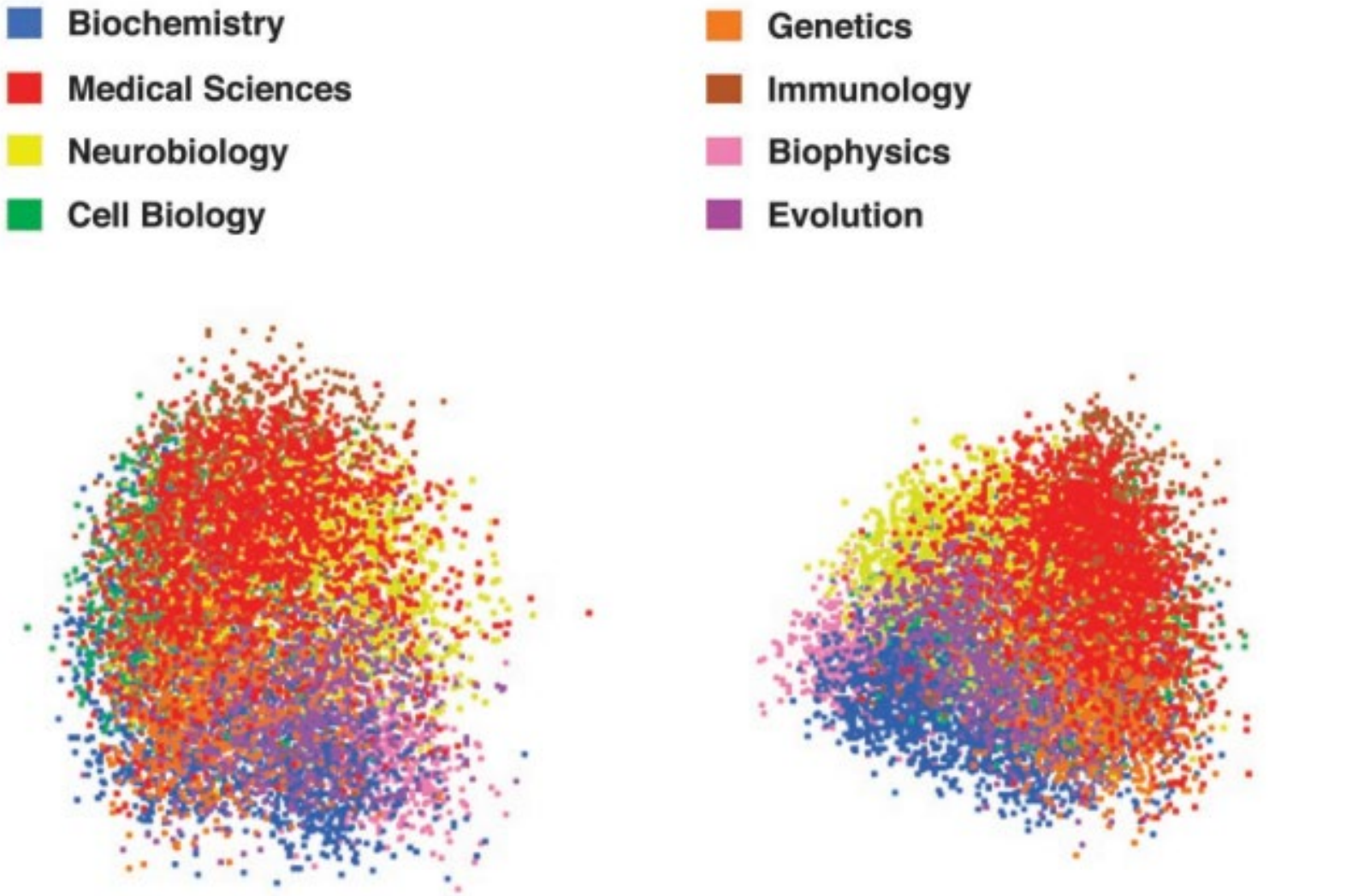


Fig. 2. PNAS articles colorized by biology subfield categories. The two-dimensional view on the three-dimensional space was selected algorithmically (*Left*) and by aided human selection (*Right*).

Discovering article relationships with LSA

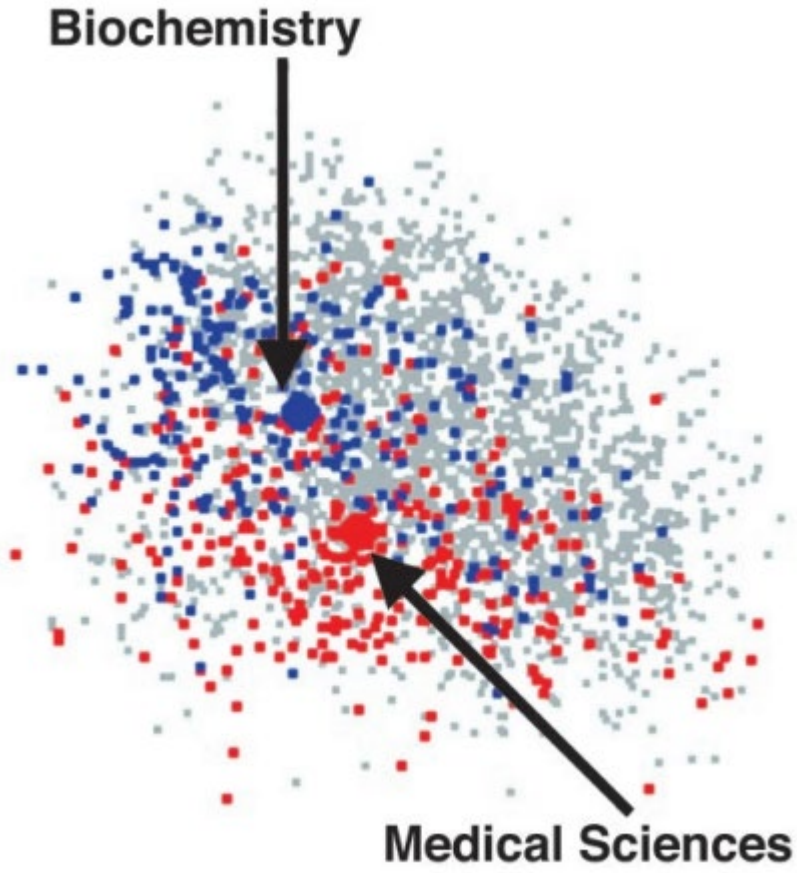


Fig. 3. Overlap of articles in categories Biochemistry (blue) and Medicine (red). Centroids of all articles in categories shown as the larger labeled dots.

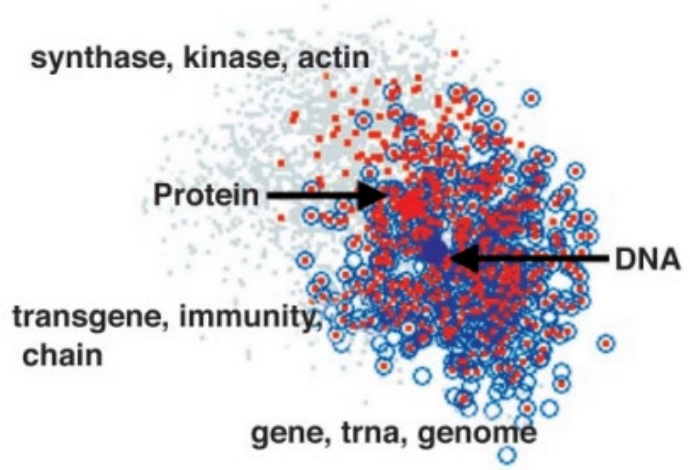


Fig. 4. Overlap between articles similar at cos ≥ 0.7 to centroids of ones with MeSH terms DNA or protein. Note the groups of bull's-eyes, articles related to both topics according in the current view, and autogenerated key words.

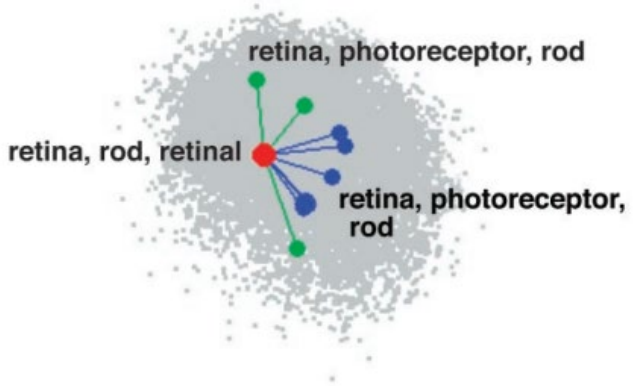


Fig. 5. Connecting similar articles across years. Red, a single article from 1998; green, similar ones from 1997; and blue, similar ones from 1999, labeled by autogenerated key words.

LDA: Latent Dirichlet Allocation

- Models document as composed of a set of topics, each of which has a distribution of words
- Word order is not considered

Recall previously the “Bad Annotator” problem

- Good annotators provide consistent scores for a given image; bad annotators assign random scores

- We can view this as a generative process:
  1. Generate a good or bad annotator with some probability
  2. That annotator then generates a score for each image, where good annotator’s score is according to that image’s good score distribution, and bad annotator’s score is uniform

- We solved for the parameters (P(good), P(score|good, image), P(score|bad)) using the EM Algorithm

Dirichlet-Multinomial Model

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

E.g.
$P_\phi(x_n = \text{"hello"}) = 0.001$
$P_\phi(x_n = \text{"is"}) = 0.06$
…

- A multinomial has a probability for each possible discrete value
  - E.g. “is” is more likely that “hello”

- The Dirichlet is a distribution of multinomials
  - E.g., for one distribution, “computer” and “hardware” may both be somewhat likely, while for another “elbow” and “forearm” are both likely

Fig: Gormley, CMU 10-701, Lec 20, 2016

Dirichlet-Multinomial Mixture Model (aka pLSA)

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

- Each document has a topic, according to a topic distribution
- Then, words are generated, according to a topic-specific word distribution
  - These word distributions can have a Dirichlet prior, which indicates how diverse the distributions are likely to be

Fig: Gormley, CMU 10-701, Lec 20, 2016

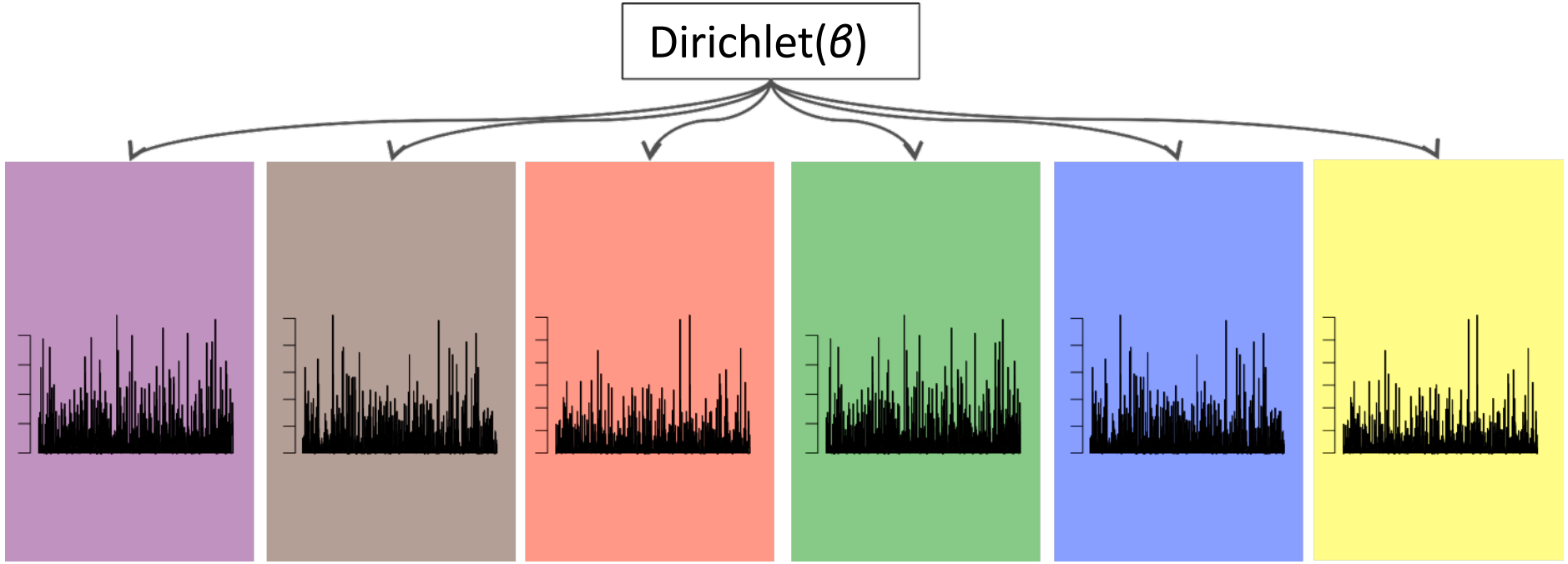
Latent Dirichlet Allocation (LDA) – an "admixture"

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

- Each document has a distribution of topics
- Each word is sampled by generating a topic, and then generating a word from the topic-specific word distribution

(Blei, Ng, & Jordan, 2003)

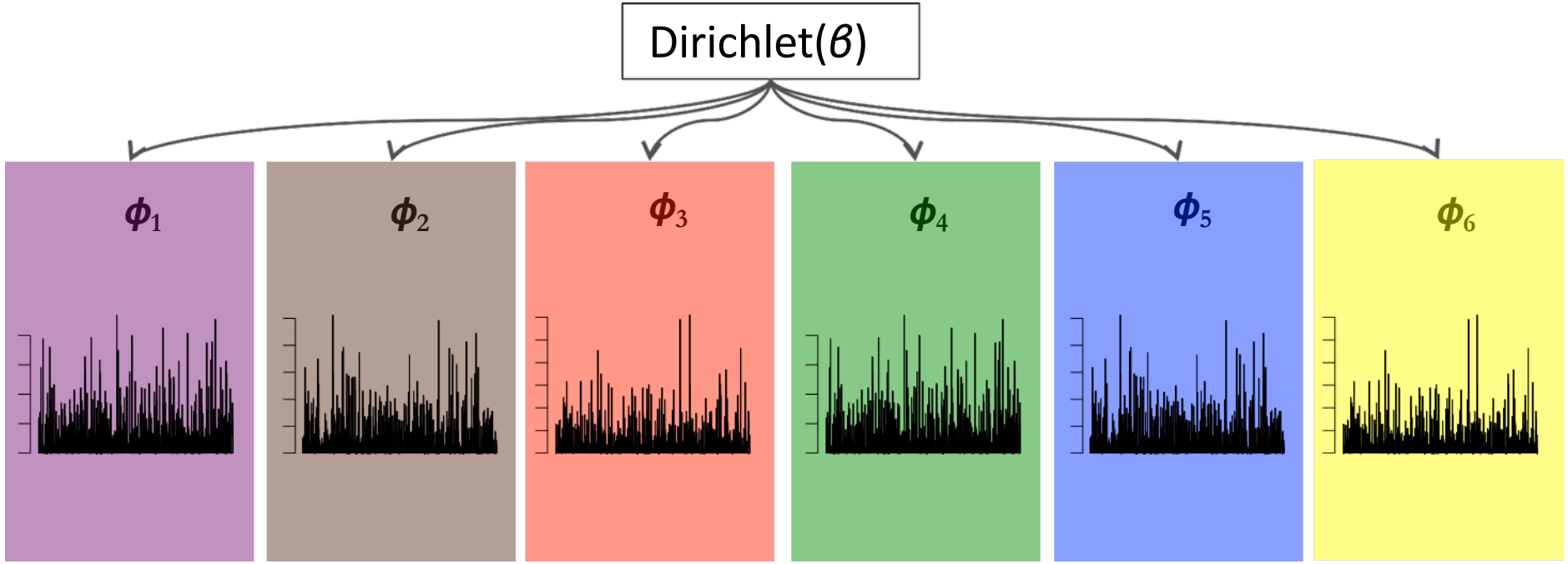
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics, which can establish how concentrated the topics are likely to be.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\phi}_{\mathrm{k}}$

(Blei, Ng, & Jordan, 2003)

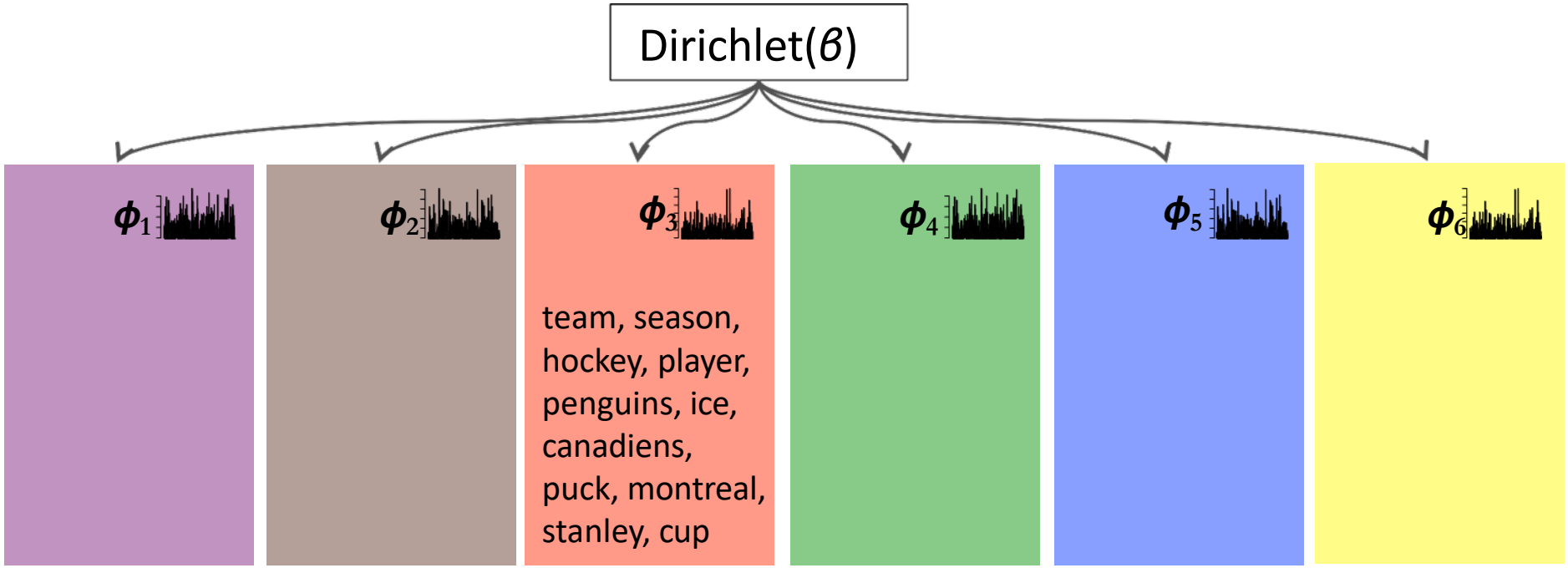
LDA for Topic Modeling



- 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

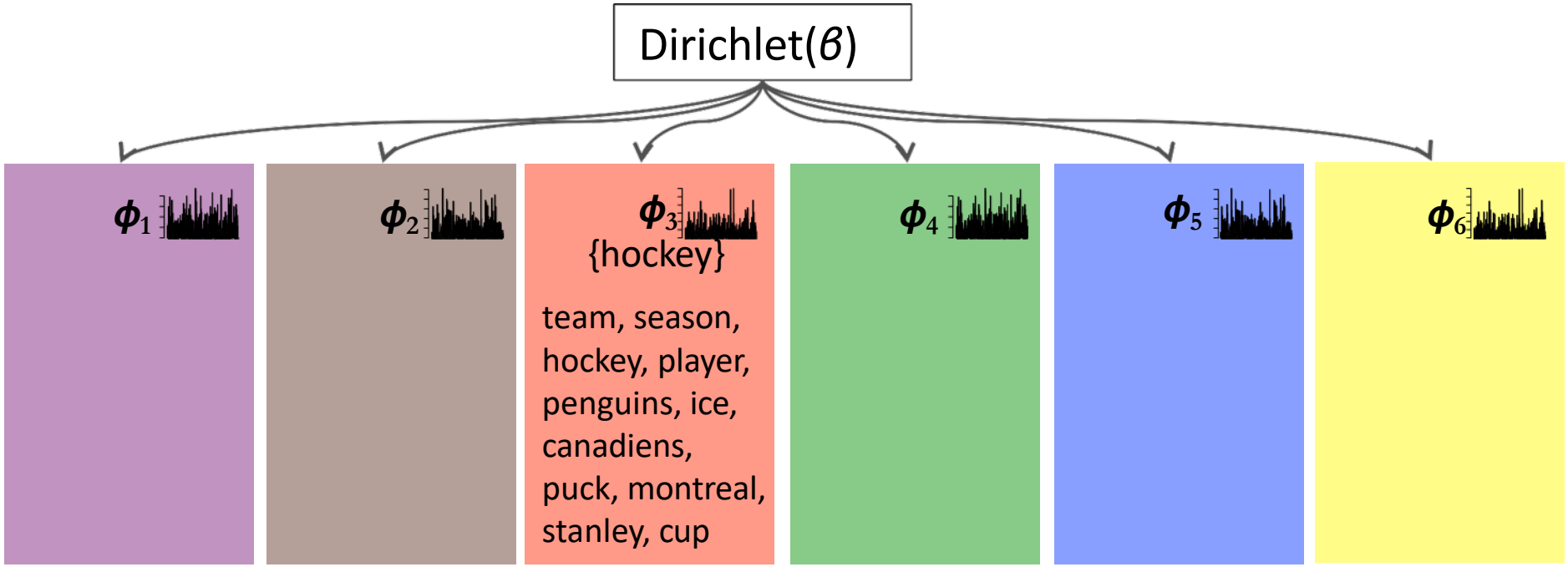


- A topic is visualized as its **high probability words.**

Slide: Gormley, CMU 10-701, Lec 20, 2016

(Blei, Ng, & Jordan, 2003)

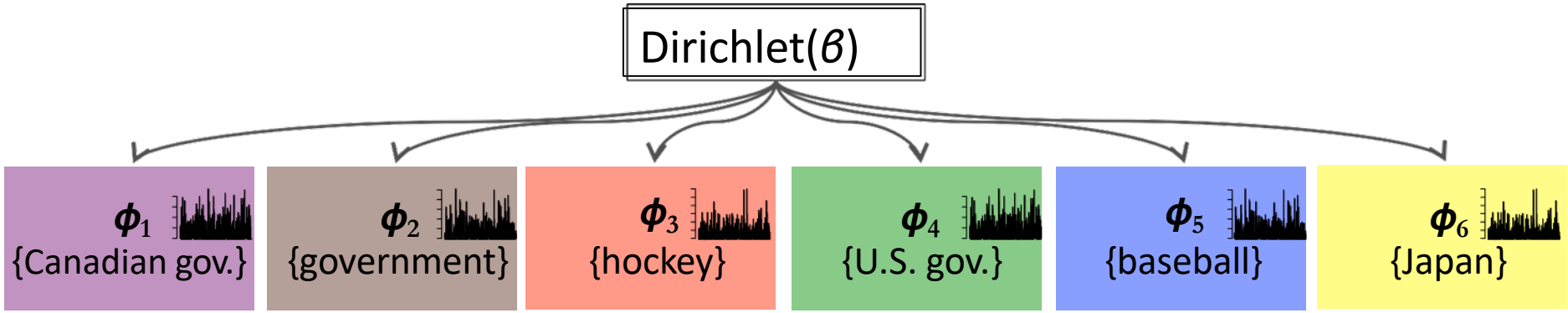
LDA for Topic Modeling



- 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

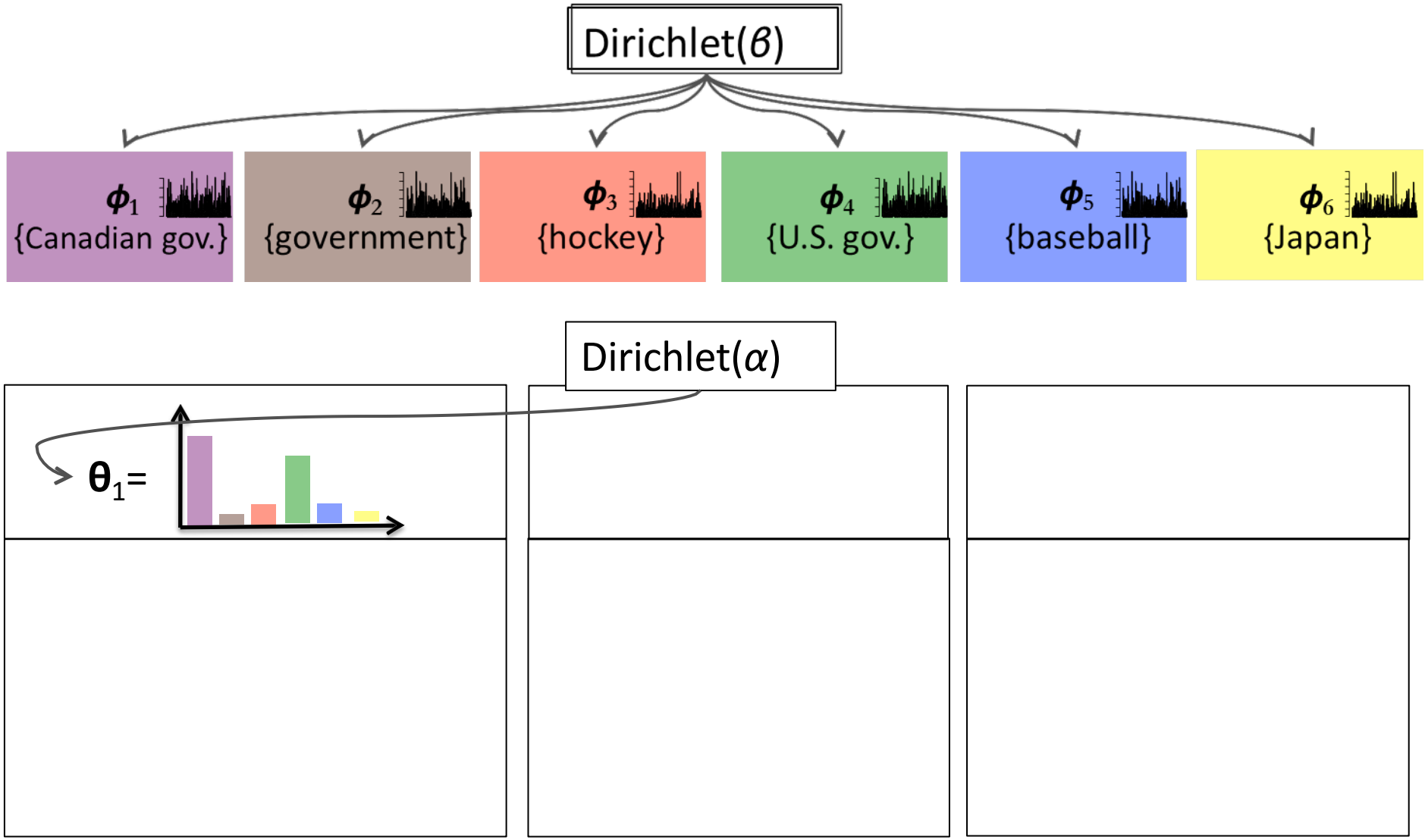


- A topic is visualized as its high probability words.
- A pedagogical **label** is used to identify the topic.

Slide: Gormley, CMU 10-701, Lec 20, 2016

(Blei, Ng, & Jordan, 2003)

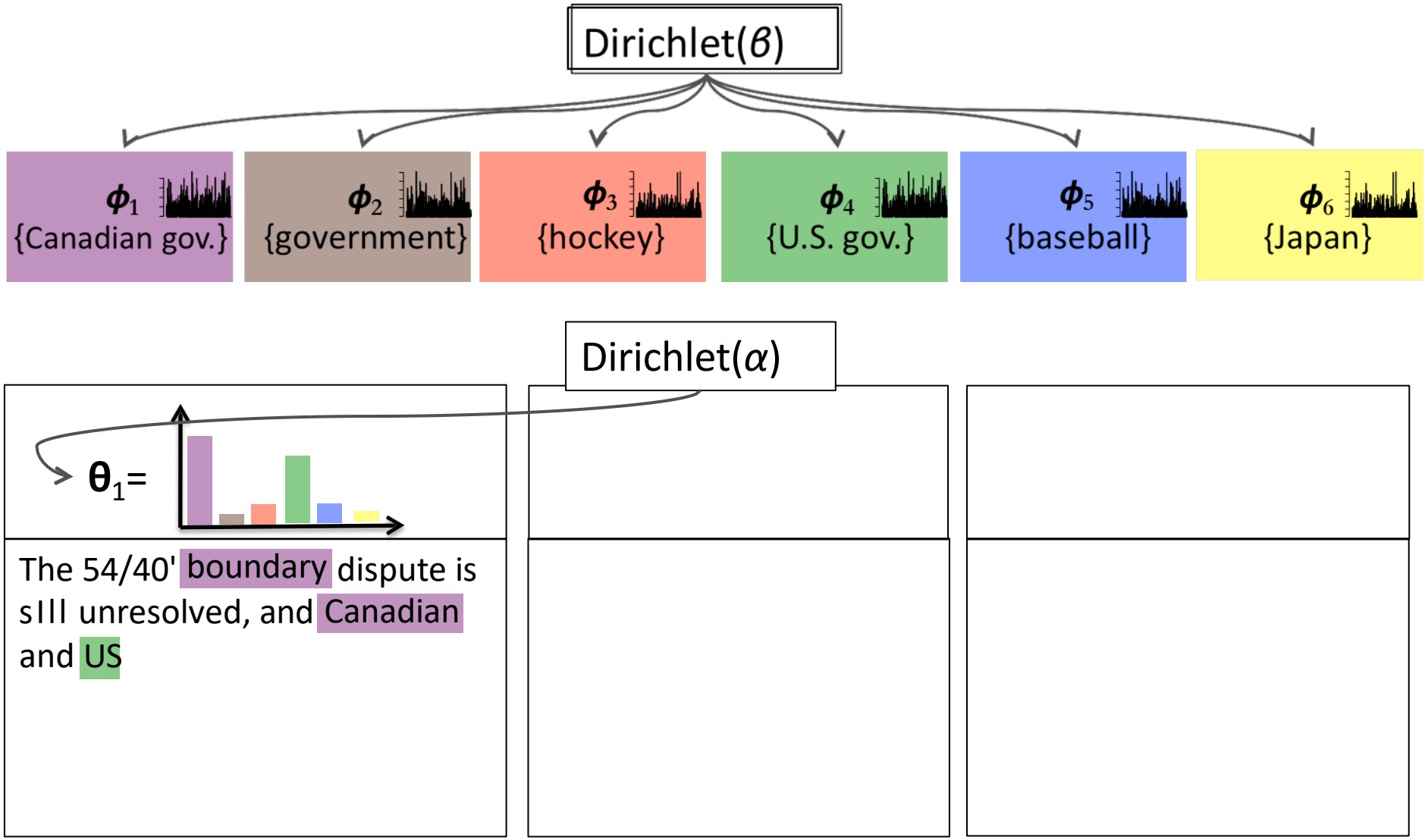
LDA for Topic Modeling



Slide: Gormley, CMU 10-701, Lec 20, 2016

(Blei, Ng, & Jordan, 2003)

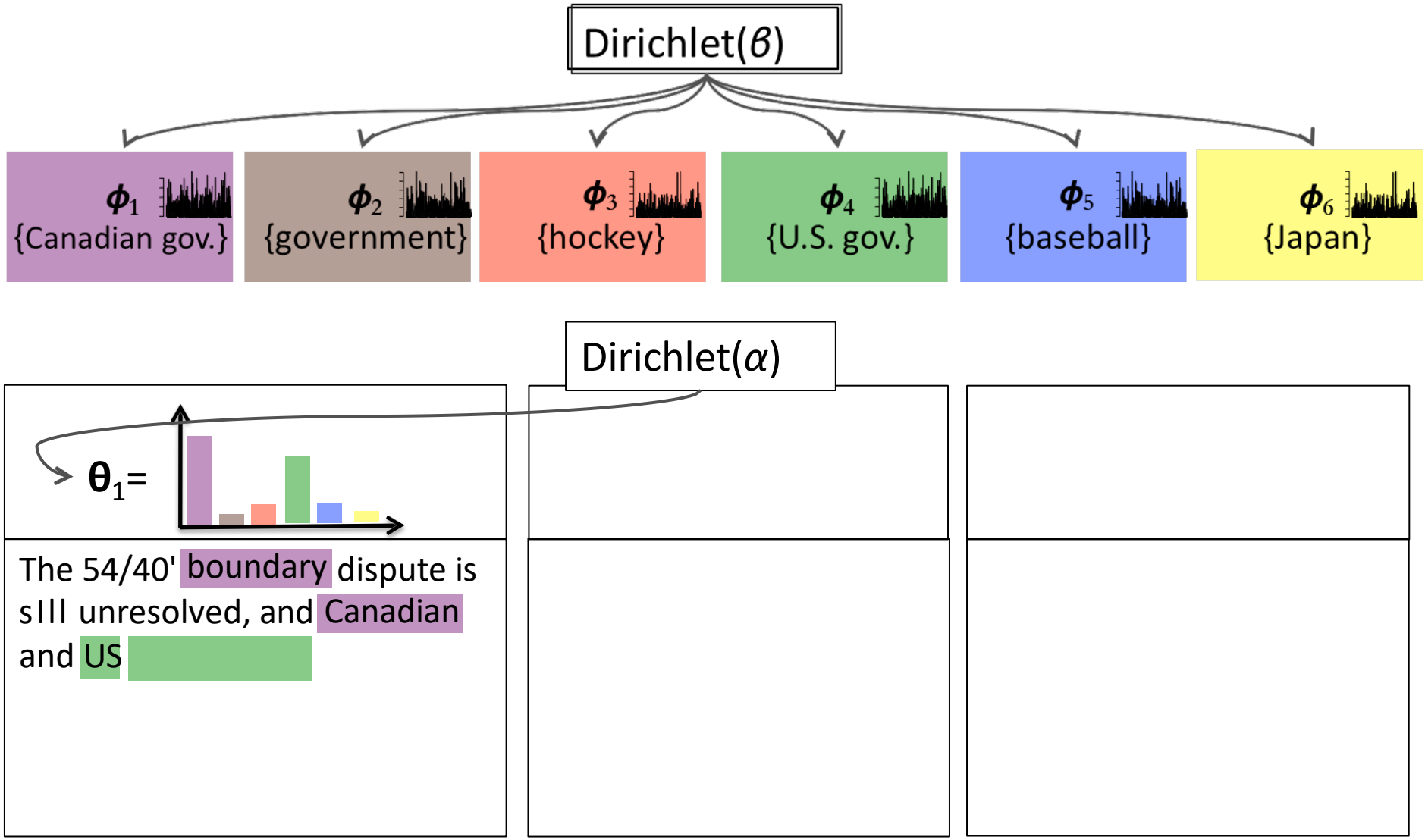
LDA for Topic Modeling



Slide: Gormley, CMU 10-701, Lec 20, 2016

(Blei, Ng, & Jordan, 2003)

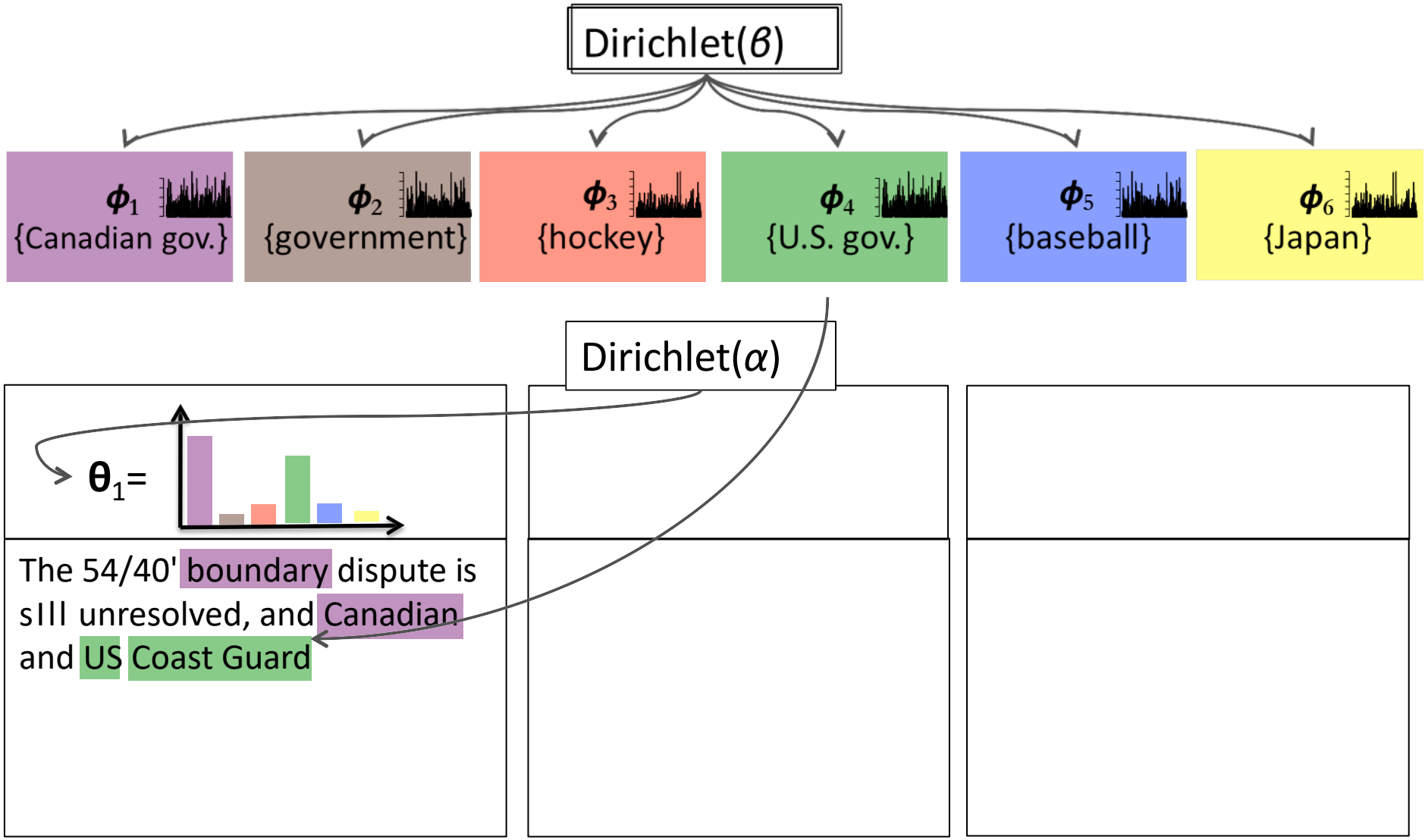
LDA for Topic Modeling



Slide: Gormley, CMU 10-701, Lec 20, 2016

(Blei, Ng, & Jordan, 2003)

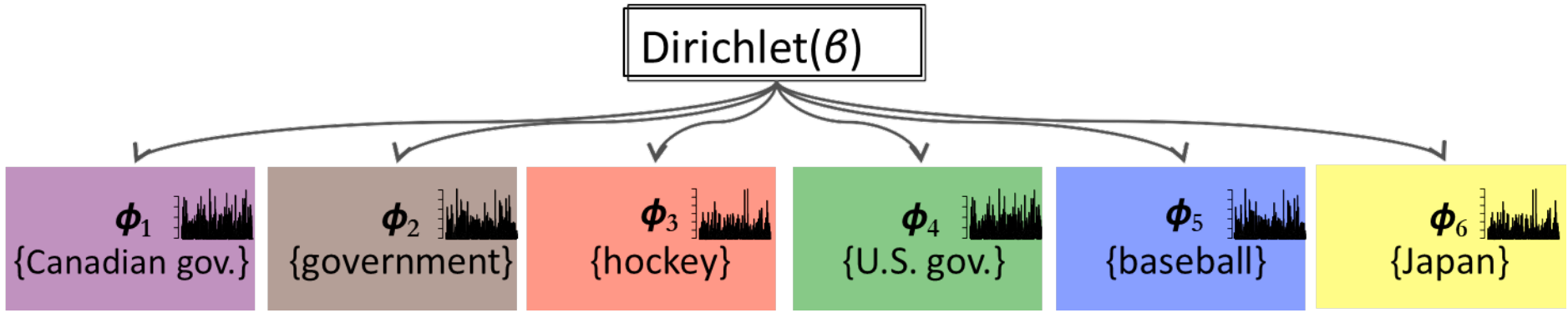
LDA for Topic Modeling

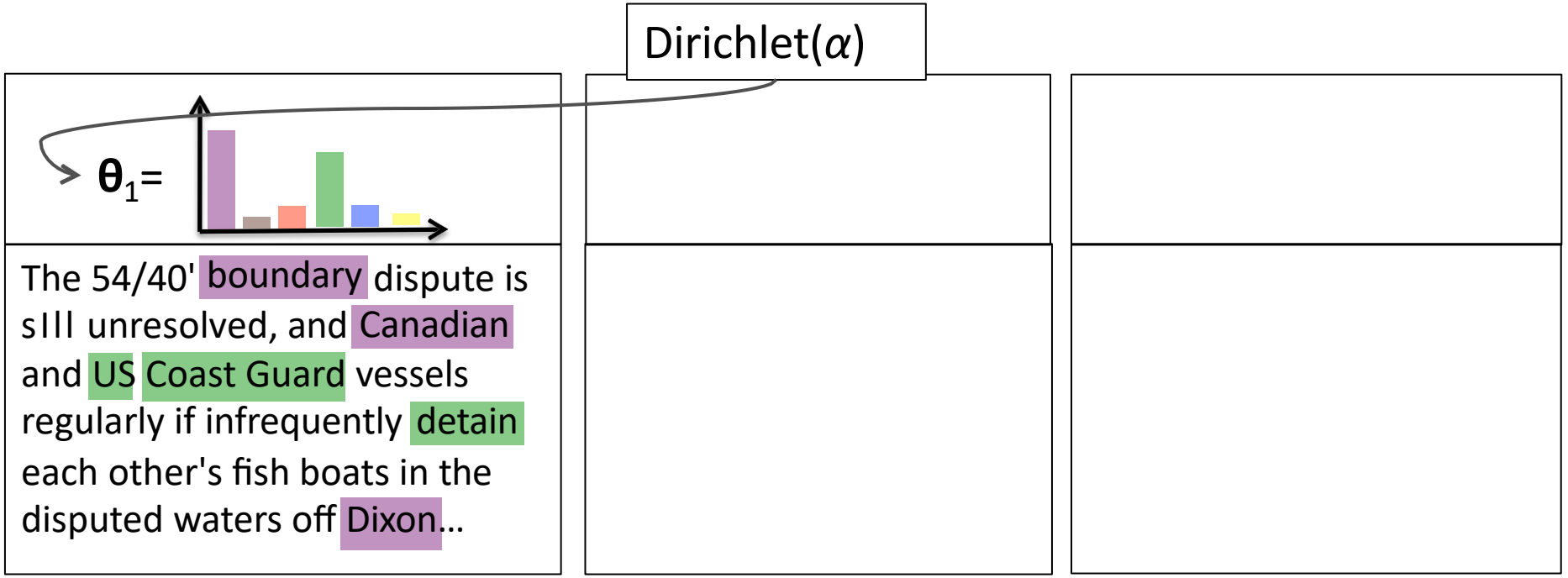


Slide: Gormley, CMU 10-701, Lec 20, 2016

(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling

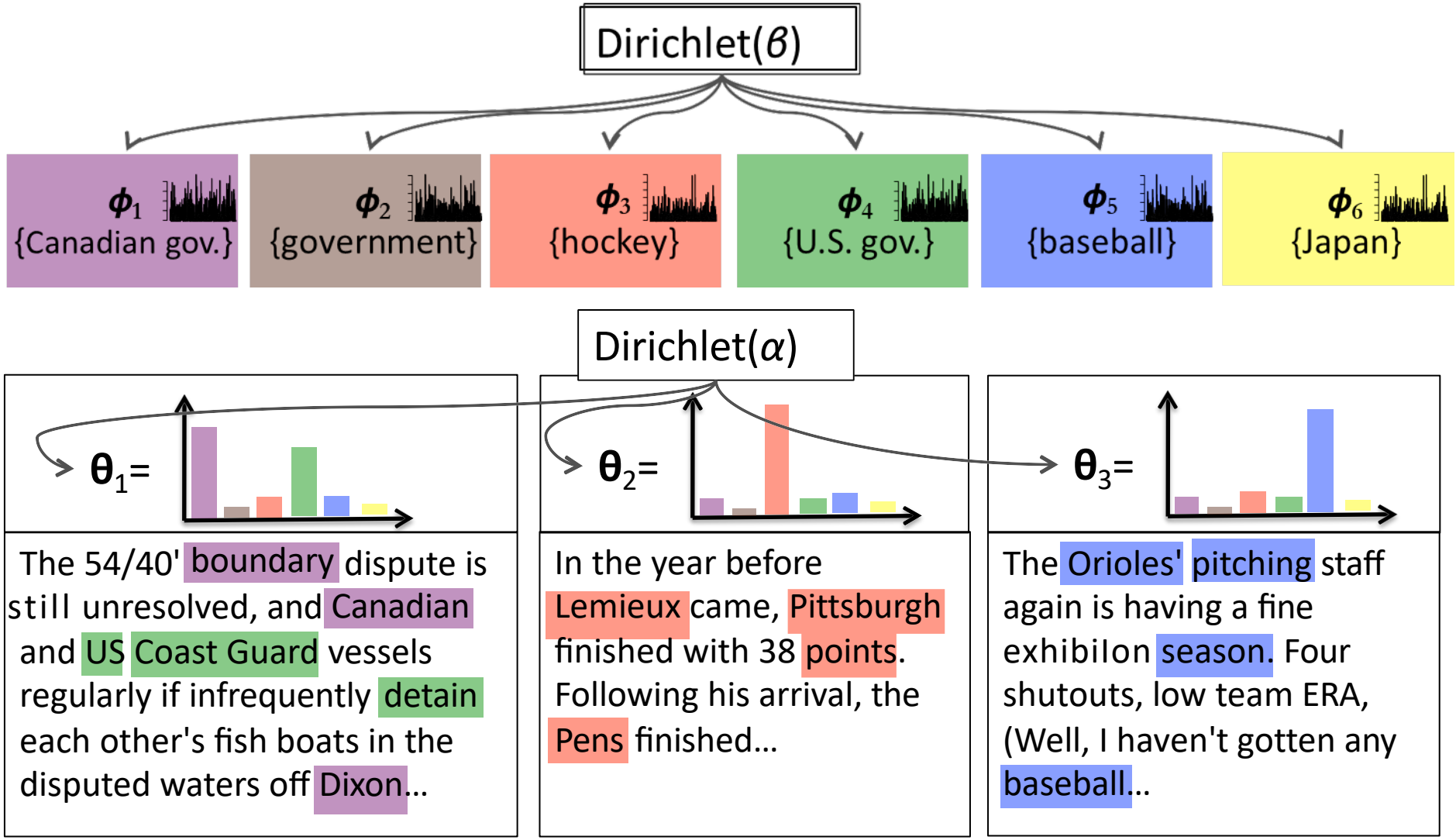




Slide: Gormley, CMU 10-701, Lec 20, 2016

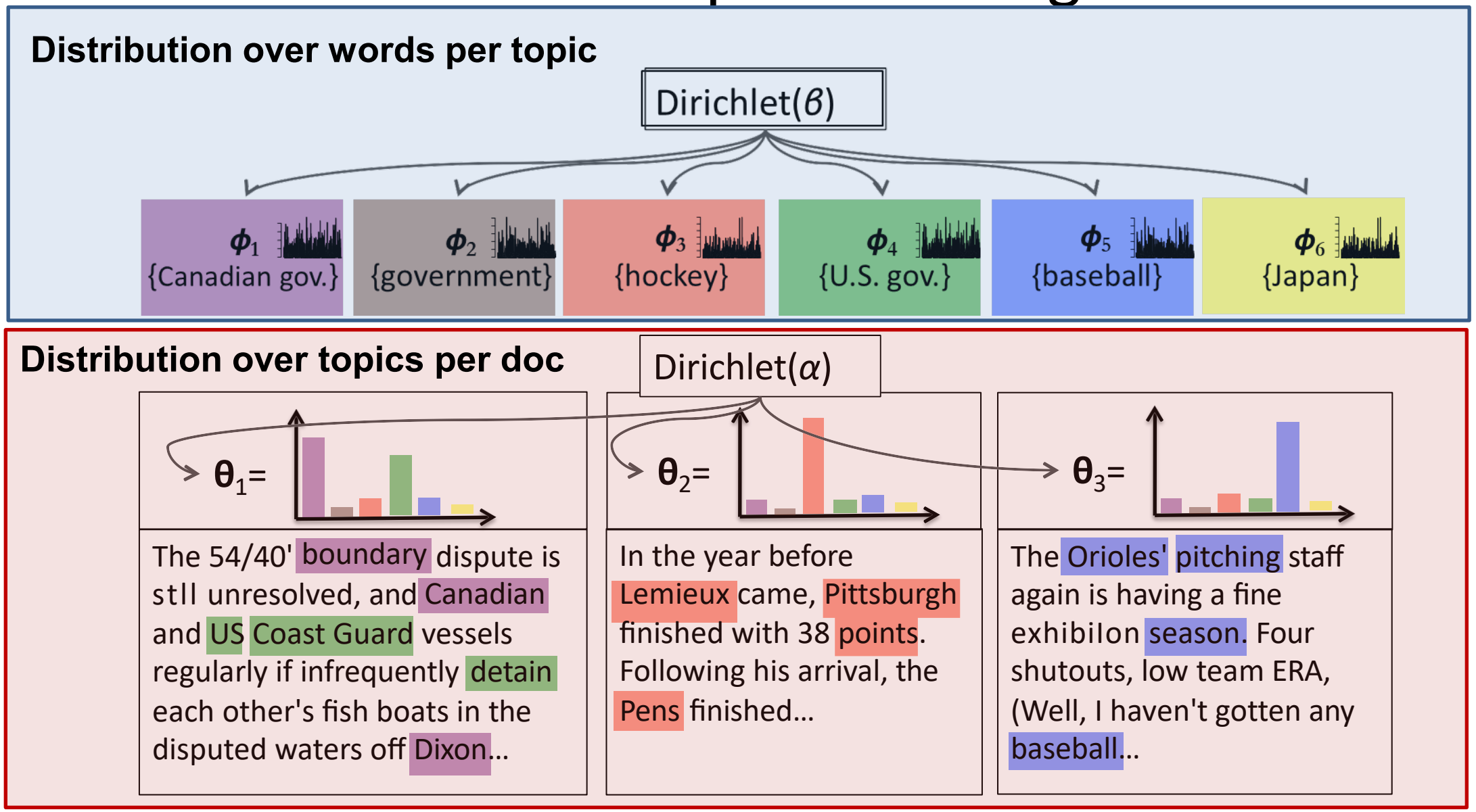
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



(Blei, Ng, & Jordan, 2003)

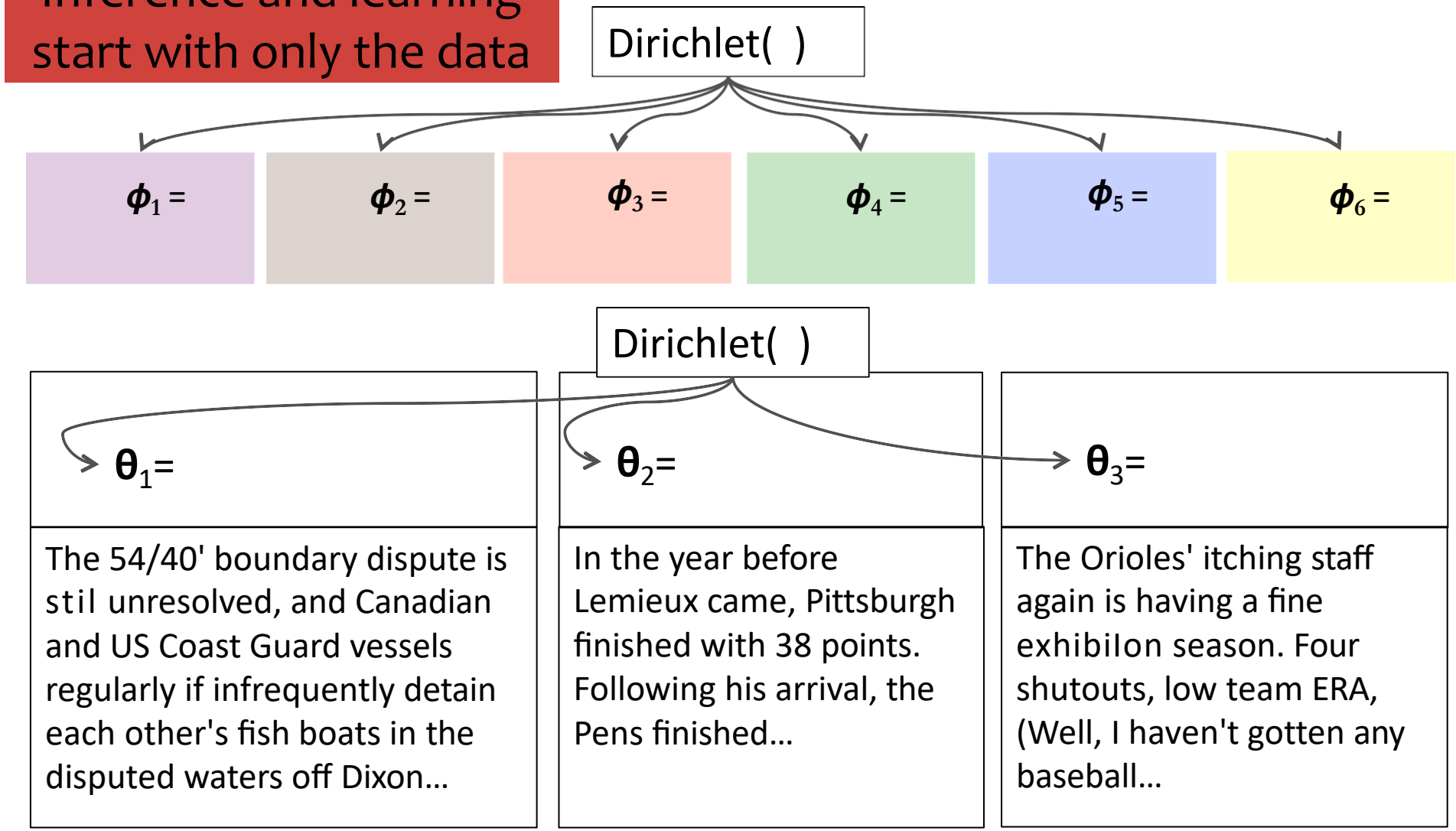
LDA for Topic Modeling



(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling

Inference and learning start with only the data



Slide: Gormley, CMU 10-701, Lec 20, 2016

Latent Dirichlet Allocation

Questions:

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

- Why might it work well anyway?

Latent Dirichlet Allocation

Why does LDA "work"?

- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible.
  2. For each topic, assign high probability to as few terms as possible.
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

Slide from David Blei, MLSS 2012

Solving for the parameters of LDA Model

- Need to solve for
  - Parameters of multinomials (topic/word distributions)
  - Latent variables of which topics are generated by each document

- This can be solved by a variational EM
  - E-step: For each document, infer the topic distribution using variational inference (involving sampling from distributions)
  - M-step: Optimize the model parameters

"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.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Fig: Blei et al. 2003

Comparison of how well the models explain the word distributions

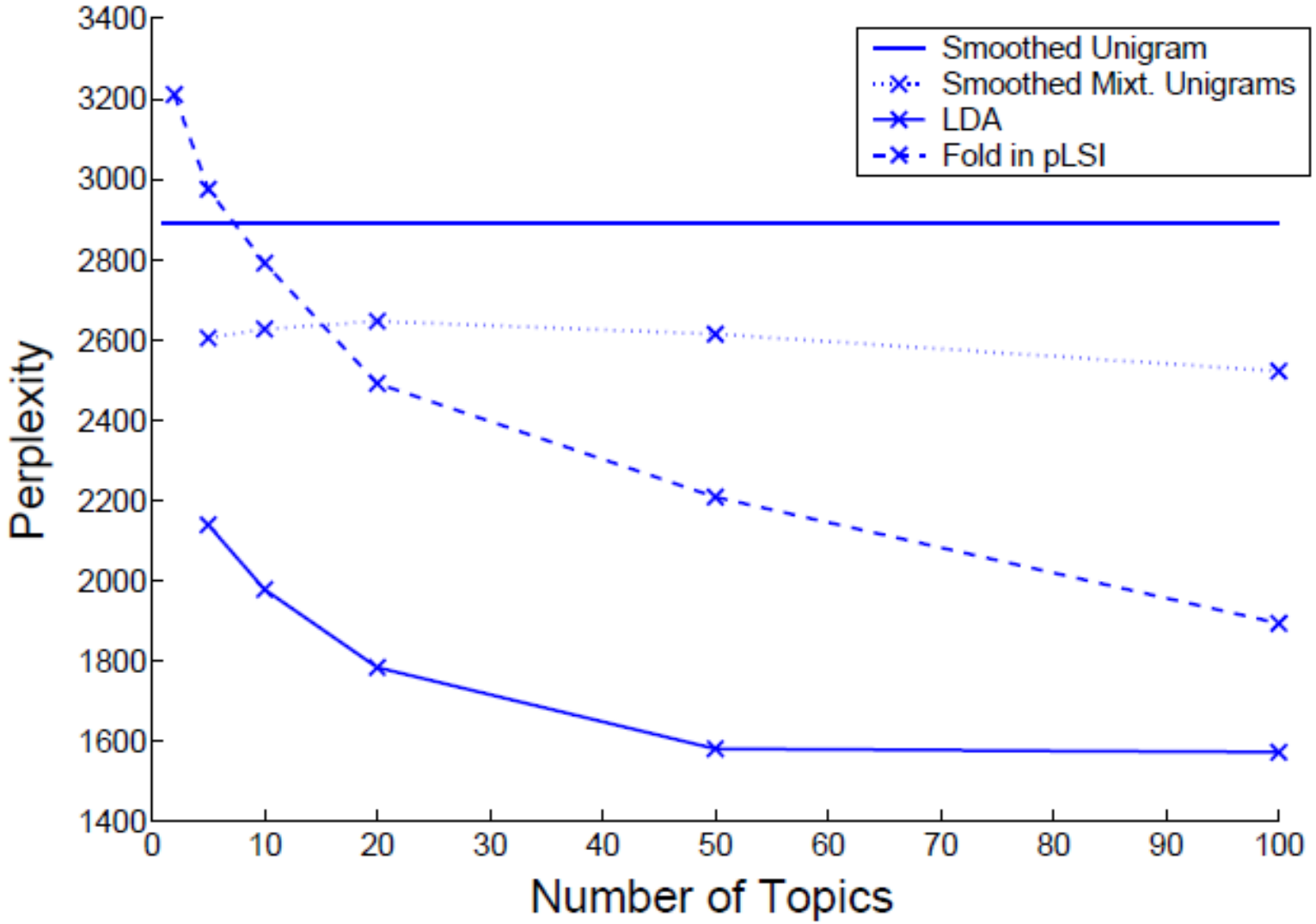


Fig: Blei et al. 2003

Classification using topic distributions as features for documents

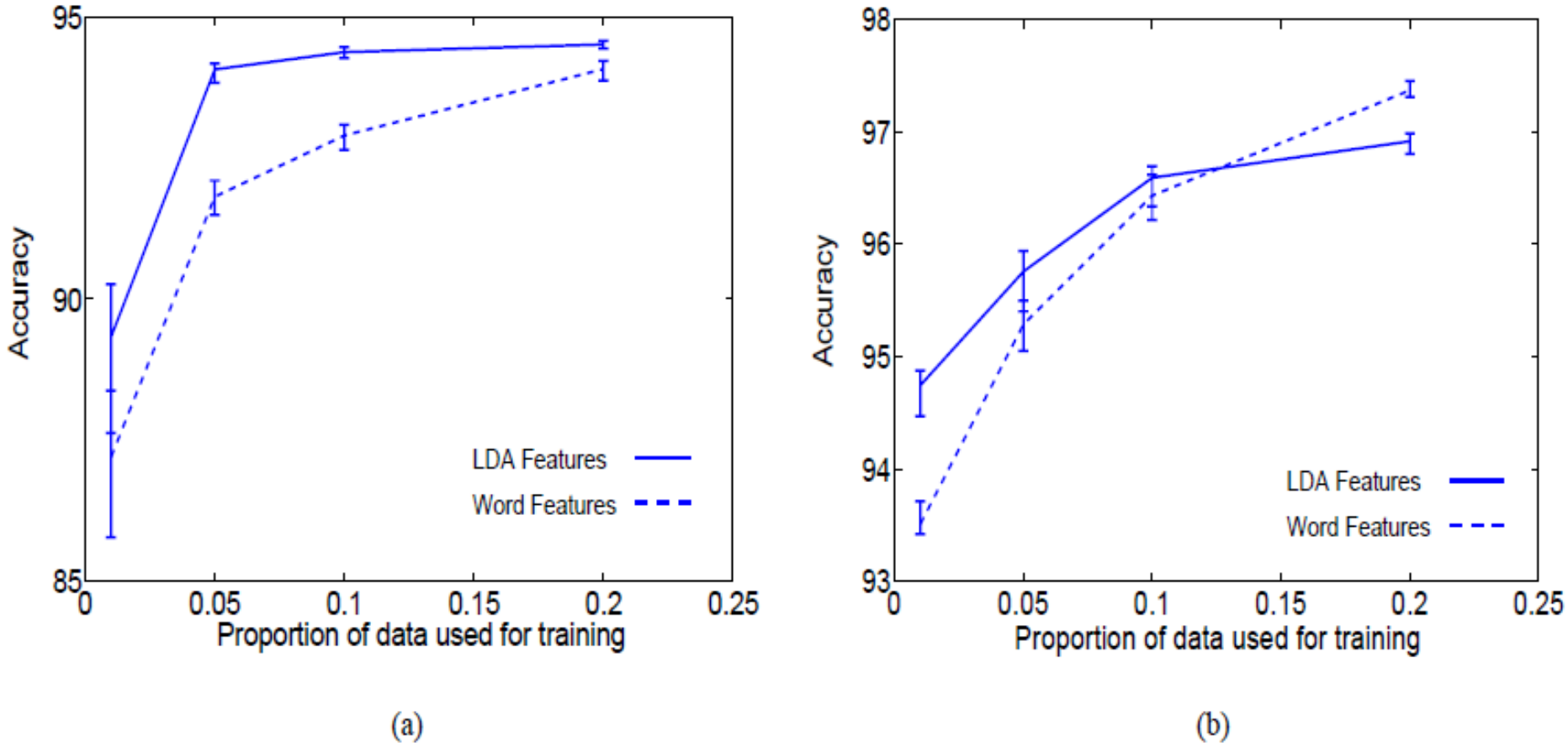


Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN VS. NOT EARN. Graph (b) is GRAIN VS. NOT GRAIN.

Fig: Blei et al. 2003

Applications of LDA

- Detect the presence of structured genetic variation in a group of individuals
  - alternative gene forms “allele” are drawn from a source population, and the specific gene form is drawn from the allele
- Identify common themes of self-images experienced by young people in social situations (Kin et al. 2022)
- Topic extraction in construction safety and health posts in Instagram (Zeng et al. 2023)
- Discover tonal structures in music corpora (Lieck et al. 2020)

Play: Guess the Topics

https://colab.research.google.com/drive/1RbYMJCXY07e3TaHfp7BaHUN5z99WUdXK?usp=sharing

BERTopic (Grootendorst 2022)

- A library for using deep learning modules as part of topic modeling
  1. Convert documents to embedding; they use Sentence-BERT so that each sentence is one vector
  2. Cluster embeddings
     1. Reduce dimensionality; they use UMAP
     2. Cluster in the lower dimension to create “terms”; they use HDBSCAN (a hierarchical density-based clustering algorithm)
  3. Compute c-tf-idf, which identifies the most salient terms for each cluster
  4. Apply topic modeling methods, like LDA or LSA

Explanation/usage: https://www.pinecone.io/learn/bertopic/

Paper: https://arxiv.org/pdf/2203.05794.pdf Code/docs: https://maartengr.github.io/BERTopic/index.html

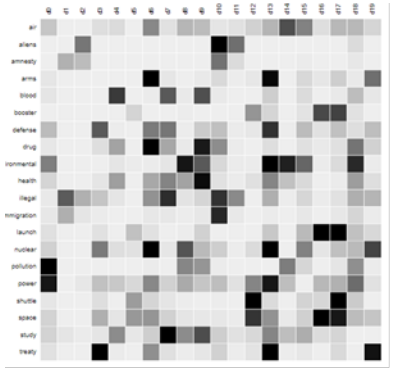
Things to remember

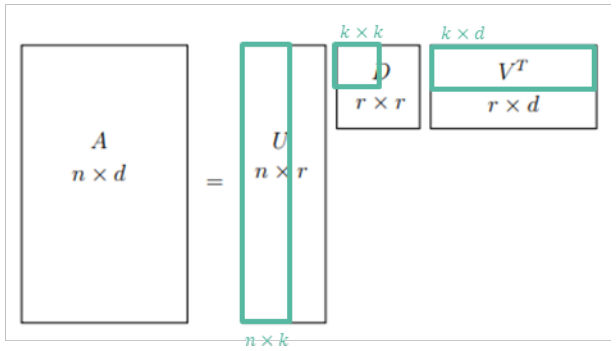
Topic modeling considers a document to be a collection of topics that compose words

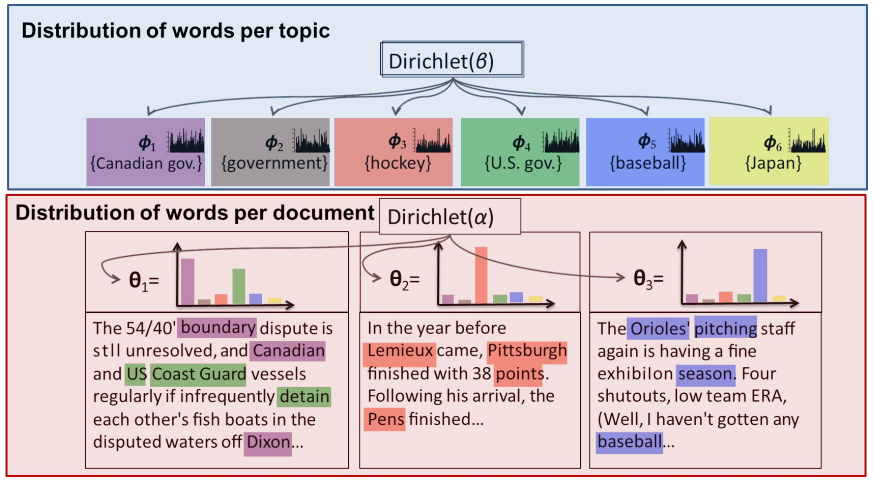
LSA: documents are a linear combination of vectors that are linear combinations of words

LDA: documents generate topics that generate words

BERTopic uses deep learning models to represent words/sentences and then applies classic methods







Next week

- Robust estimates and outliers
- Reinforcement learning – by Josh Levine (your TA)