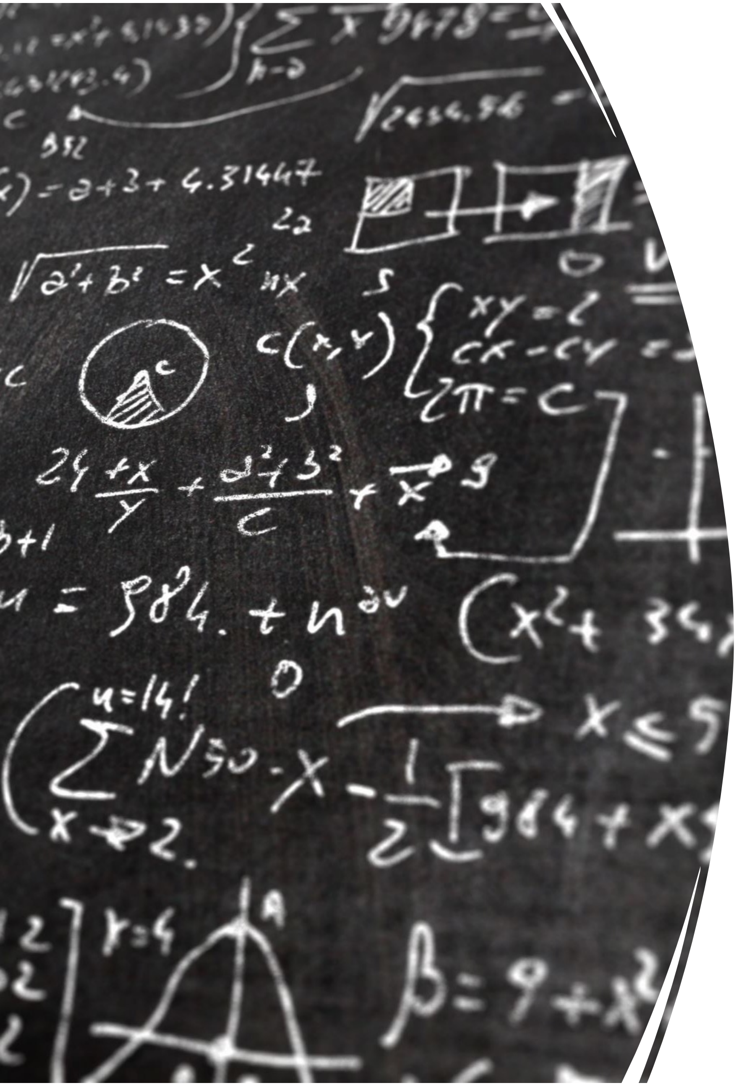


CS 383 – Computational Text Analysis

Lecture 6
Clustering, LDA

Adam Poliak

02/06/2023

Slides adapted David Mimno, Jordan Boyd-Graber

Announcements

- Office Hours:
  - This week: Thursday 3:30-4:30pm
- HW02 due Wednesday 02/08
- Reading 03 released today
  - Due Monday 02/13
- HW03 due Wednesday 02/15
  - Released today

Outline

- Clustering
- Topic Modeling - LDA

Clustering

Different Types of Machine Learning

- Supervised Learning
  - Given labeled examples, learn rules

- Unsupervised Learning
  - Given unlabeled example, learn patterns

[Slide from Tony Liu]

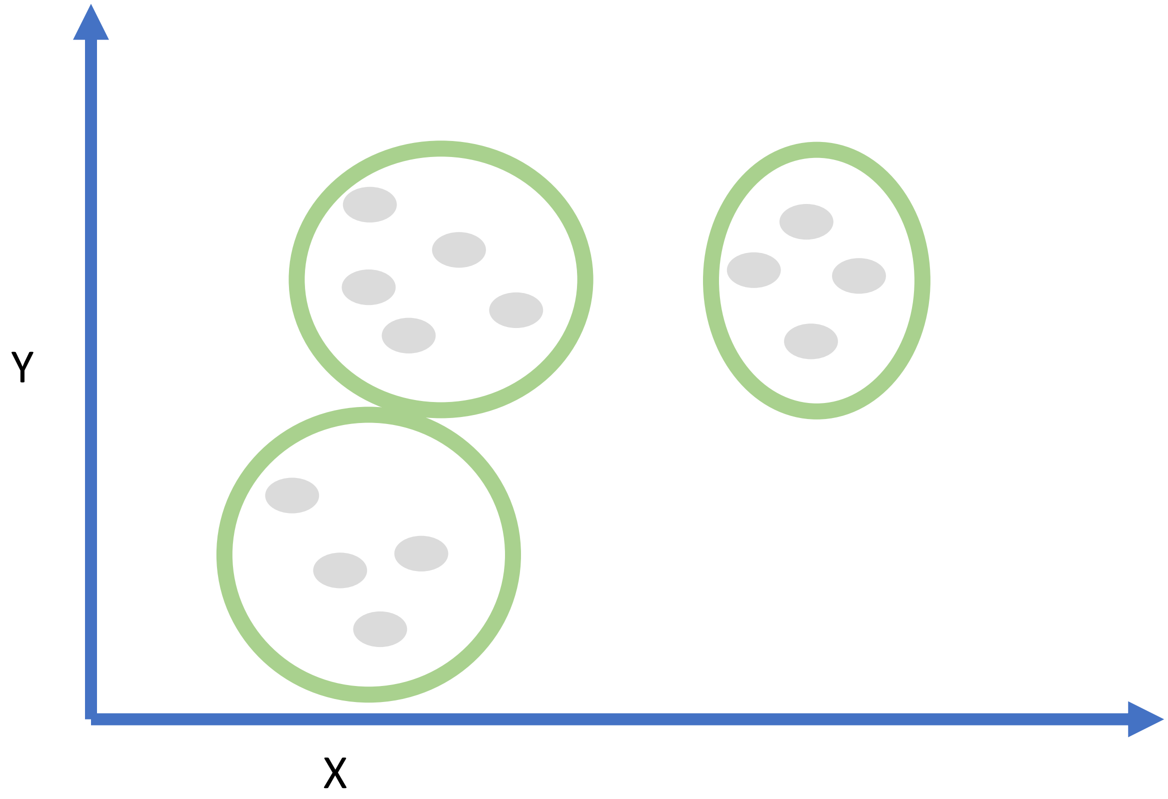
Clustering

- Unsupervised learning
  - Requires data, but no labels
- Detect patterns e.g. in
  - Group emails
  - Group obituaries
  - Group any documents
- Useful when don't know what you're looking for
- Good way to explore your data

Slide from David Sontag

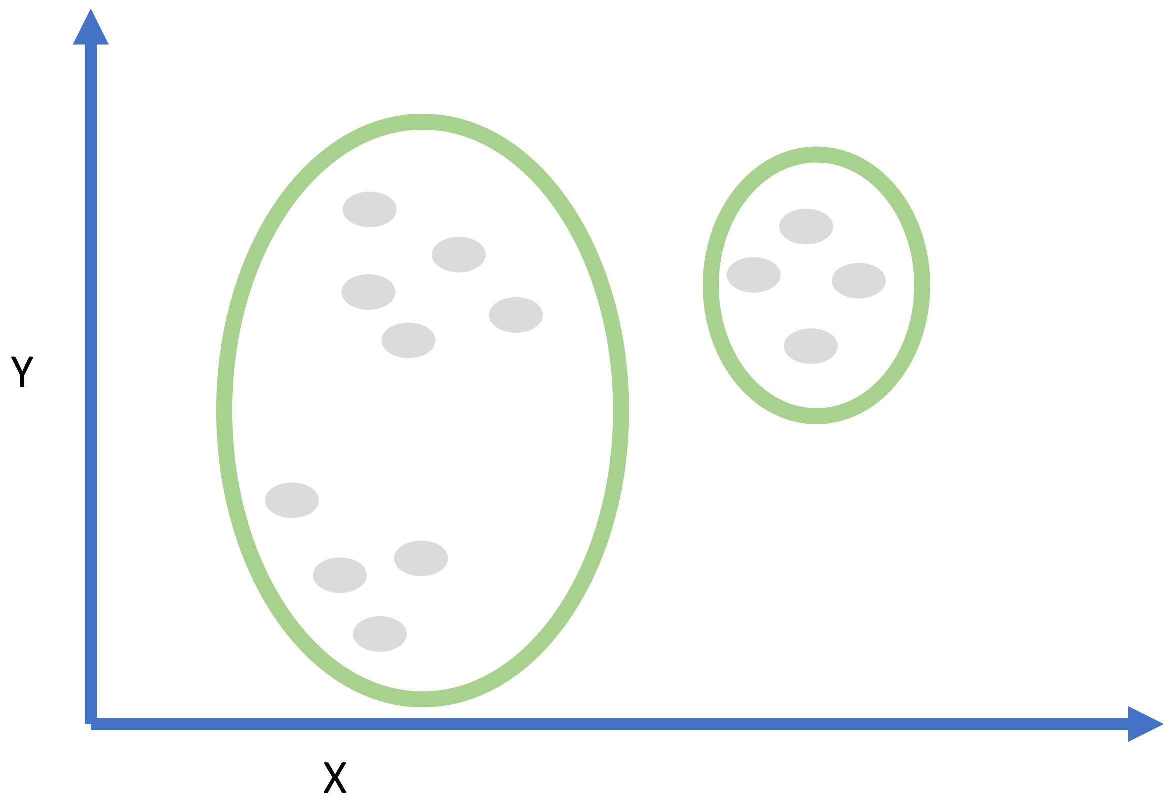
Idea: group together similar instances

- Example: 2D point patterns



Idea: group together similar instances

- Example: 2D point patterns



Clustering HW02

- HW02 analyzed obits

- Why might we want to cluster obits?

  - Find groups of similar obituaries
  - Find topics of obituaries
  - …

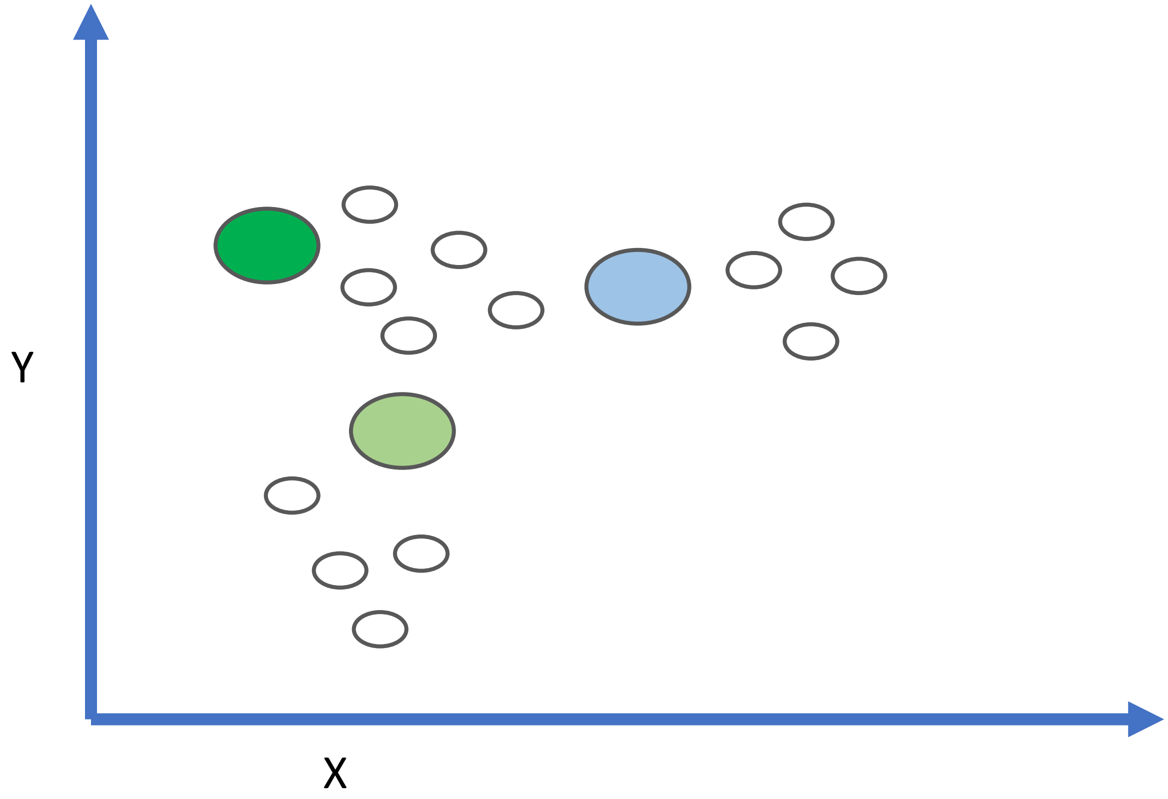
K-Means Algorithm

K-means Algorithms

1. Initialize: Randomly pick K points as cluster centers

Randomly pick K points as centers

- Example: 2D point patterns



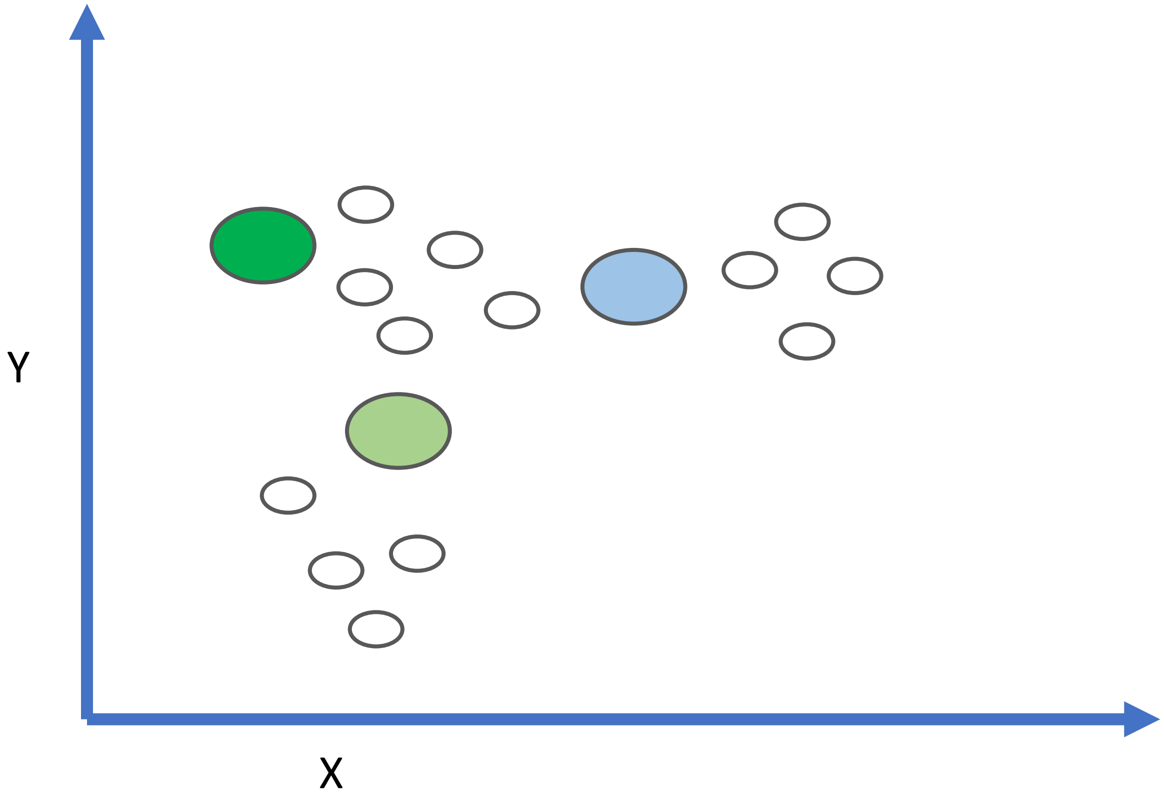
K-means Algorithms

1. Initialize: Randomly pick K points as cluster centers

2. Assign data points to each cluster
    1. Based on distance between point and cluster's center

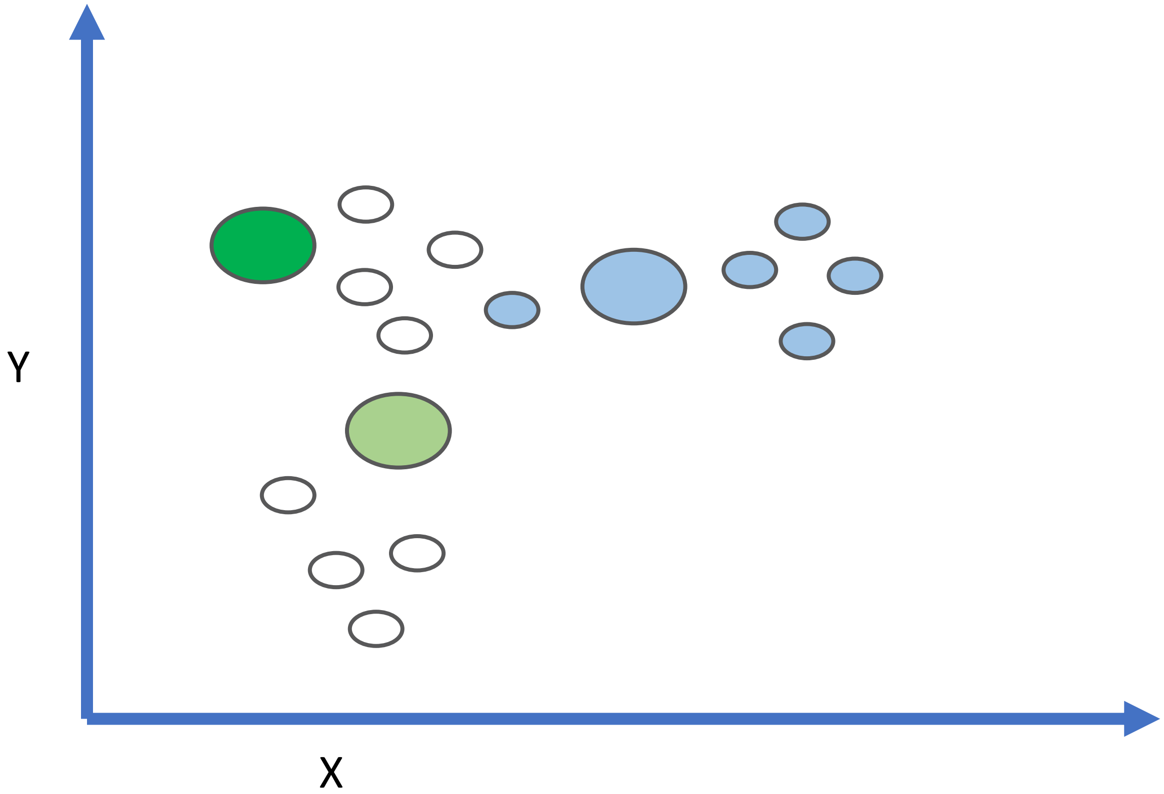
Assign data points to each cluster

- Example: 2D point patterns



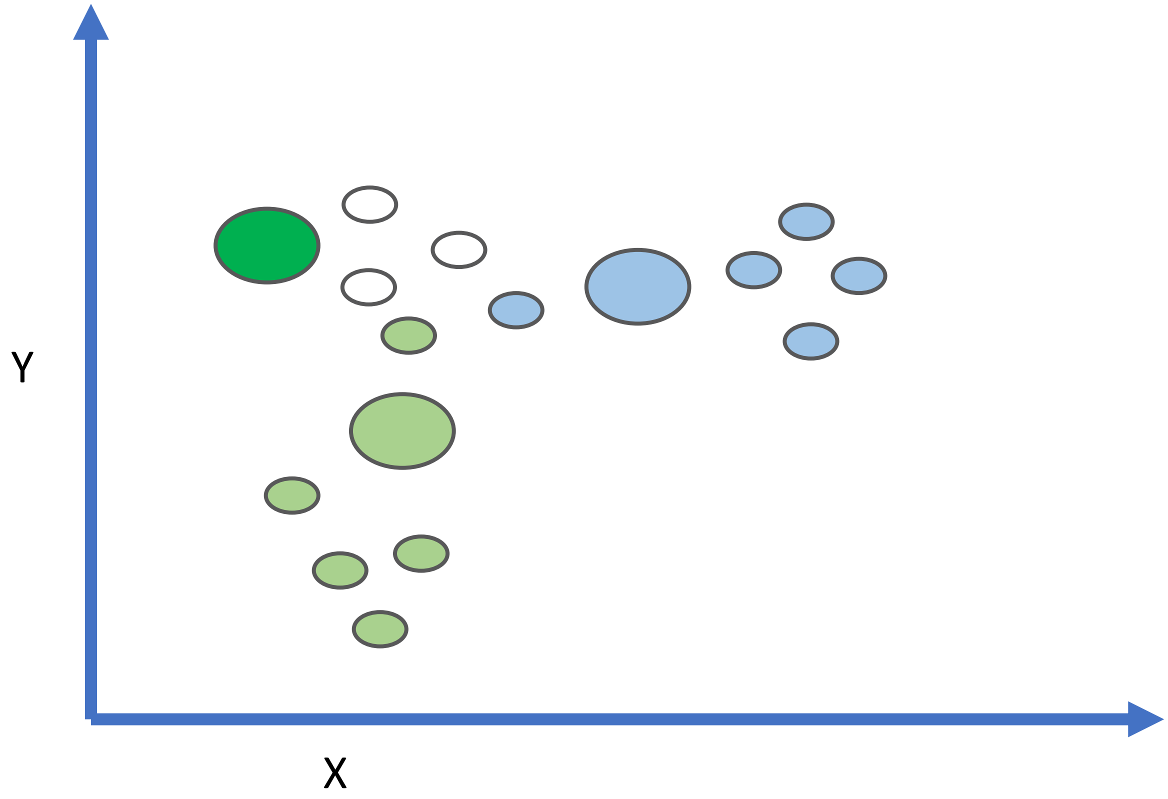
Assign data points to each cluster

- Example: 2D point patterns



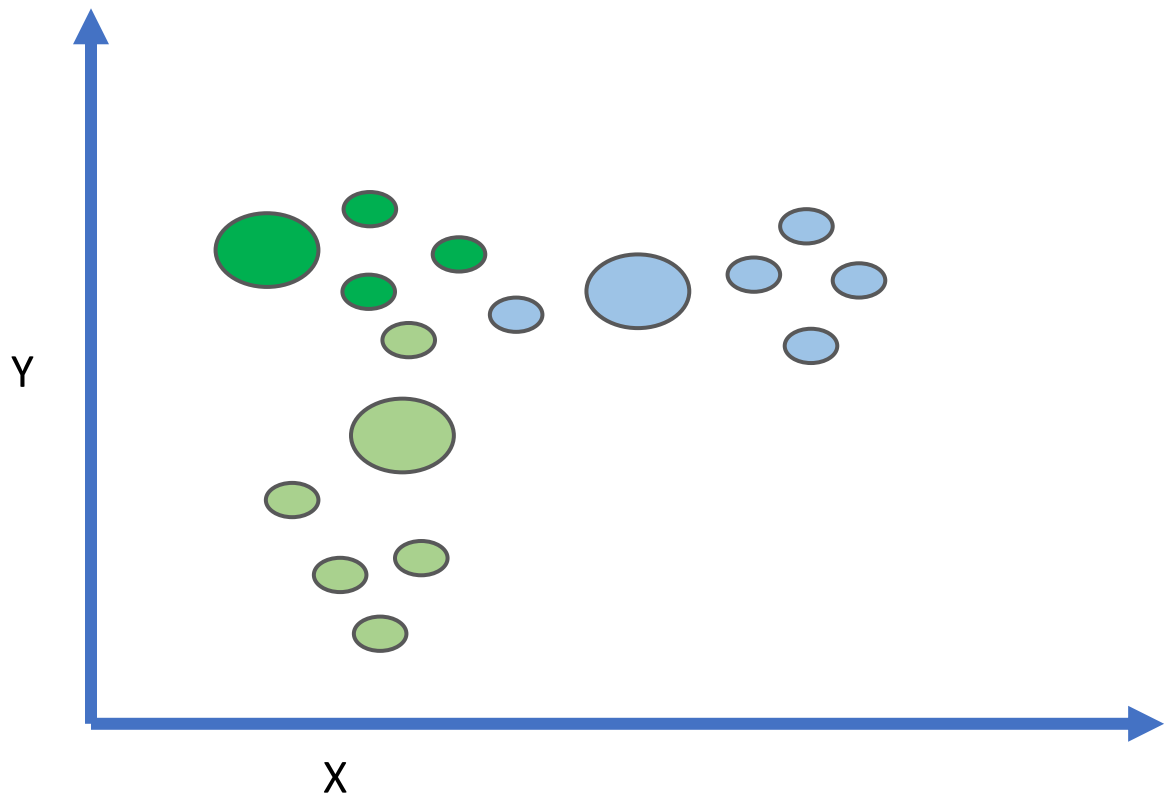
Assign data points to each cluster

- Example: 2D point patterns



Assign data points to each cluster

- Example: 2D point patterns



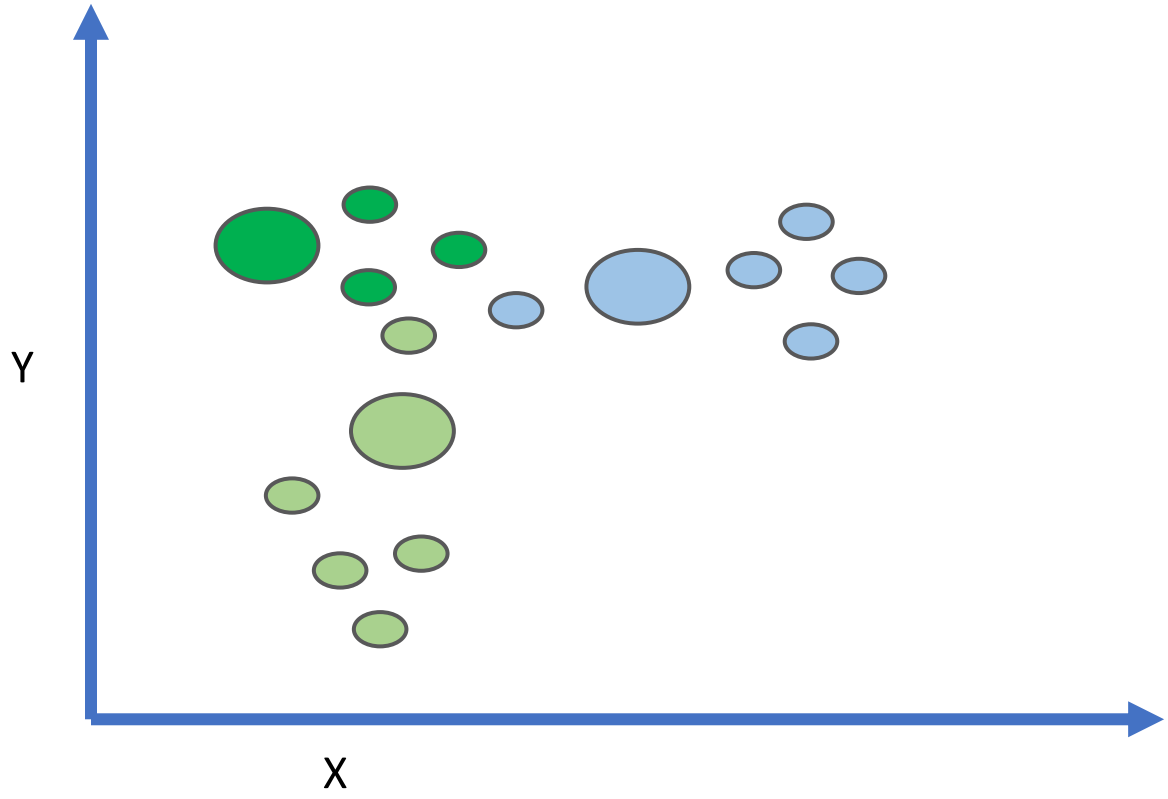
K-means Algorithms

1. Initialize: Randomly pick K points as cluster centers

2. Assign data points to each cluster
    1. Based on distance between point and cluster's center
3. Update the center of each cluster
    1. The average of its assigned points

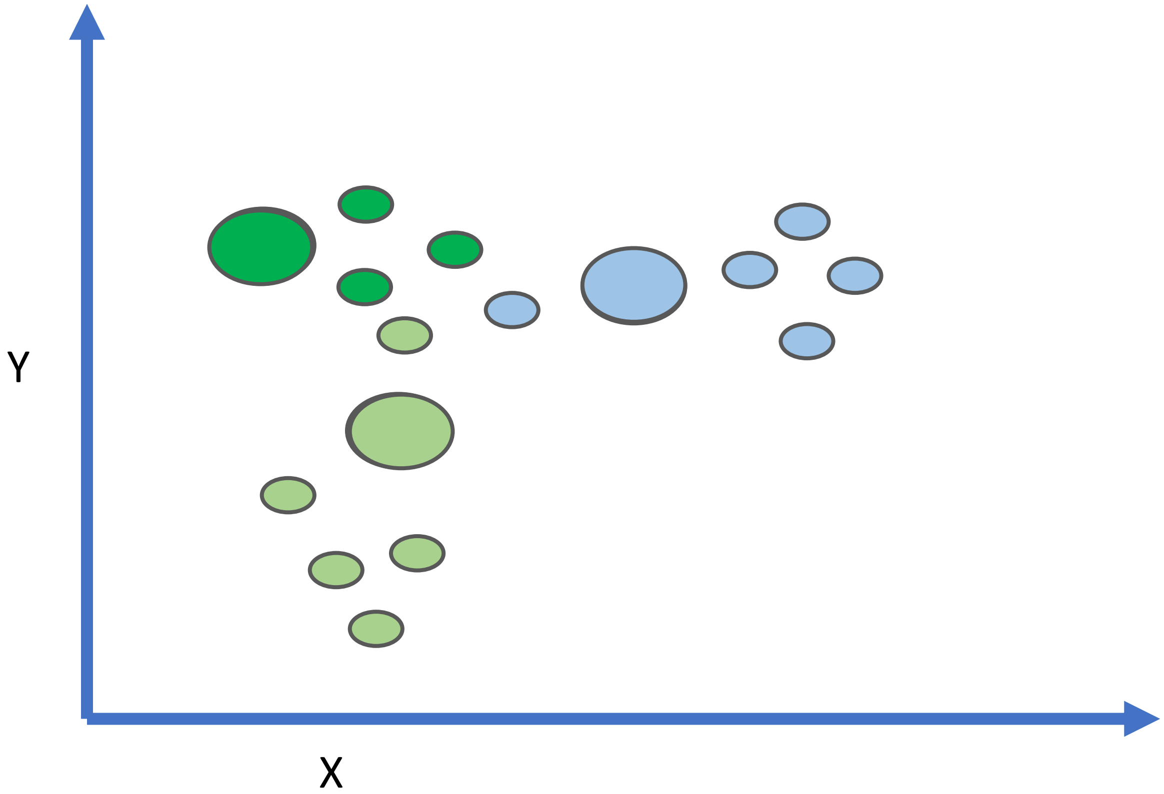
Update Centers

- Example: 2D point patterns



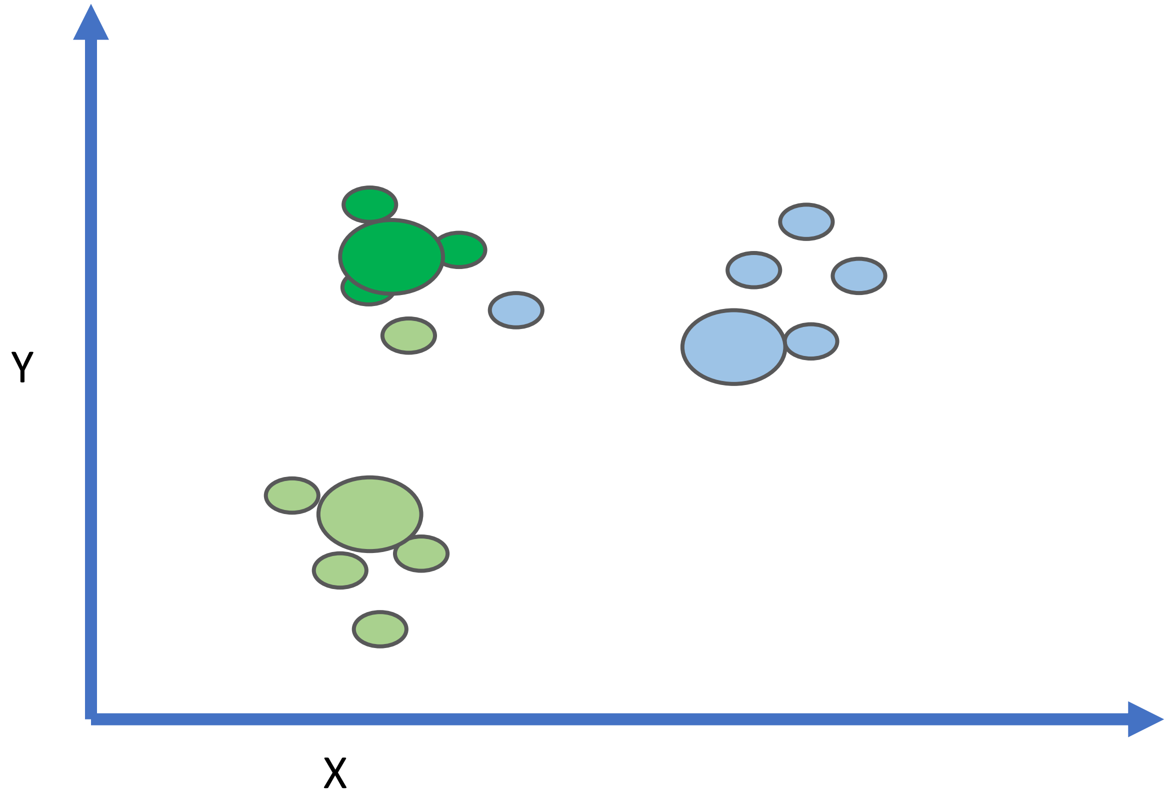
Update Centers

- Example: 2D point patterns



Updated Centers

- Example: 2D point patterns



K-means Algorithms

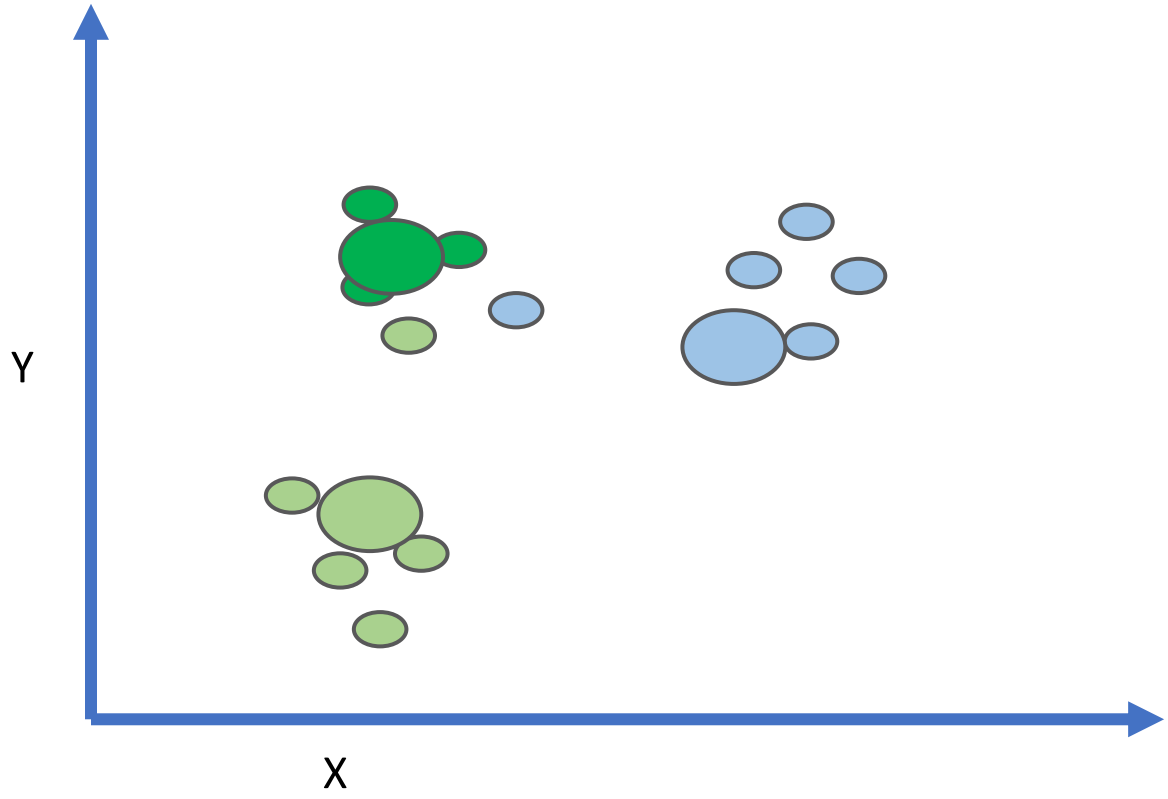
1. Initialize: Randomly pick K points as cluster centers

2. Assign data points to each cluster
    1. Based on distance between point and cluster's center
3. Update the center of each cluster
    1. The average of its assigned points

4. Repeat 2 & 3 until the assignments stop changing

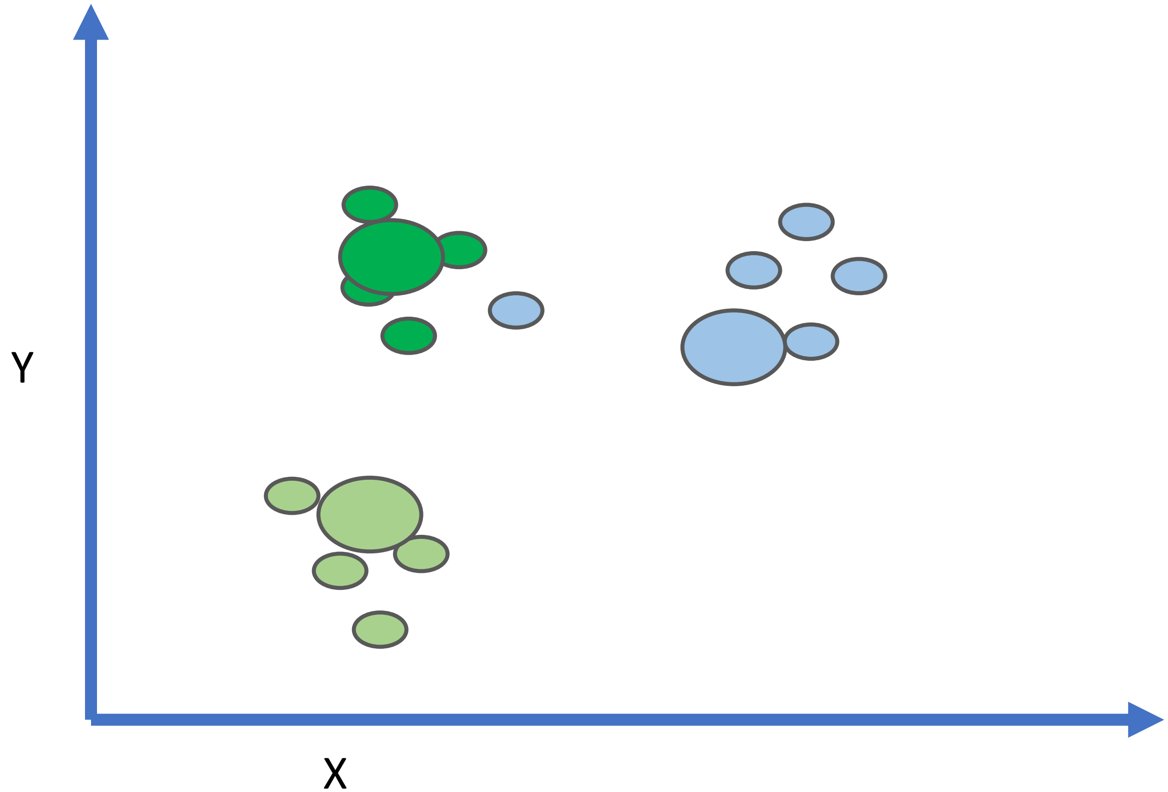
Reassign data points to each cluster

- Example: 2D point patterns



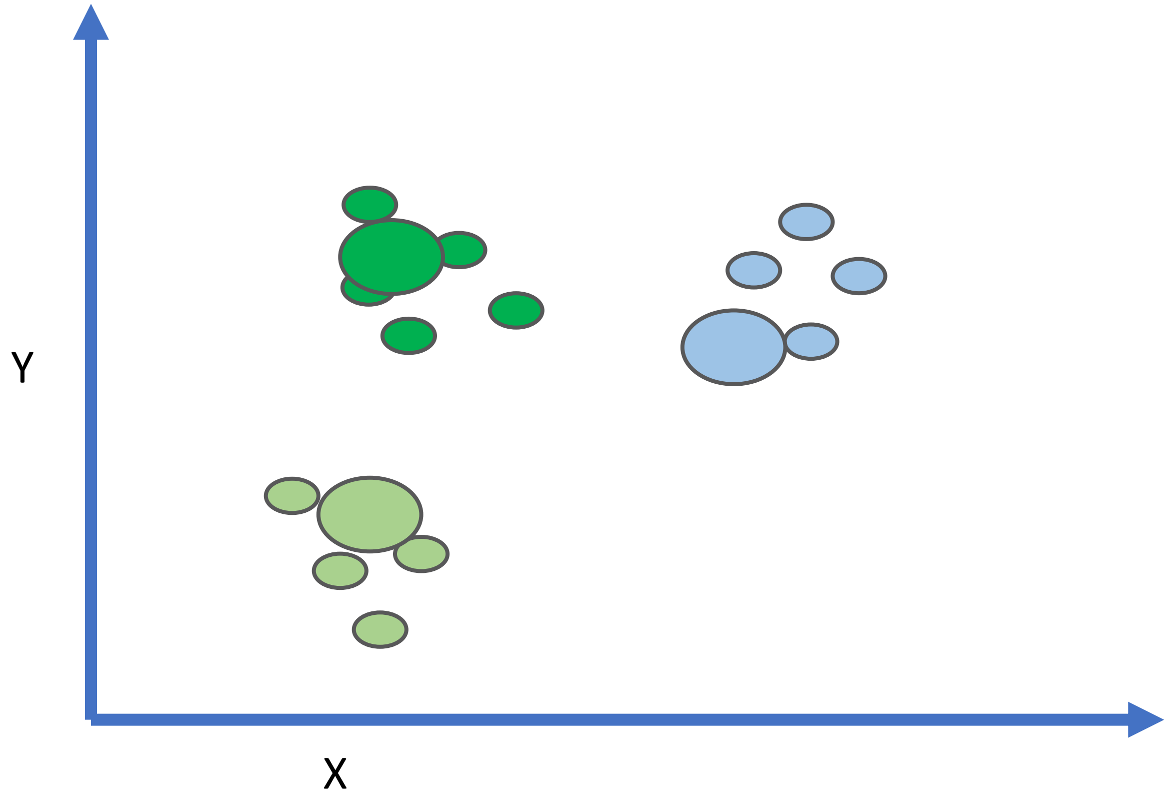
Reassign data points to each cluster

- Example: 2D point patterns



Reassign data points to each cluster

- Example: 2D point patterns



K-means Algorithms

1. Initialize: Randomly pick K points as cluster centers

2. Assign data points to each cluster
    1. Based on distance between point and cluster's center
3. Update the center of each cluster
    1. The average of its assigned points

4. Repeat 2 & 3 until the assignments stop changing

K-means Algorithms

1. Initialize: Randomly pick K points as cluster centers

2. Assign data points to each cluster
   1. Based on **distance between** point and cluster's center
3. Update the center of each cluster
   1. The average of its assigned points

4. Repeat 2 & 3 until the assignments stop changing

How do we quantify similarity/distance?

We need to define similarity/distance

Similarity metrics we've seen so far:

cos similarity

Euclidian distance between two documents x_1 and x_2

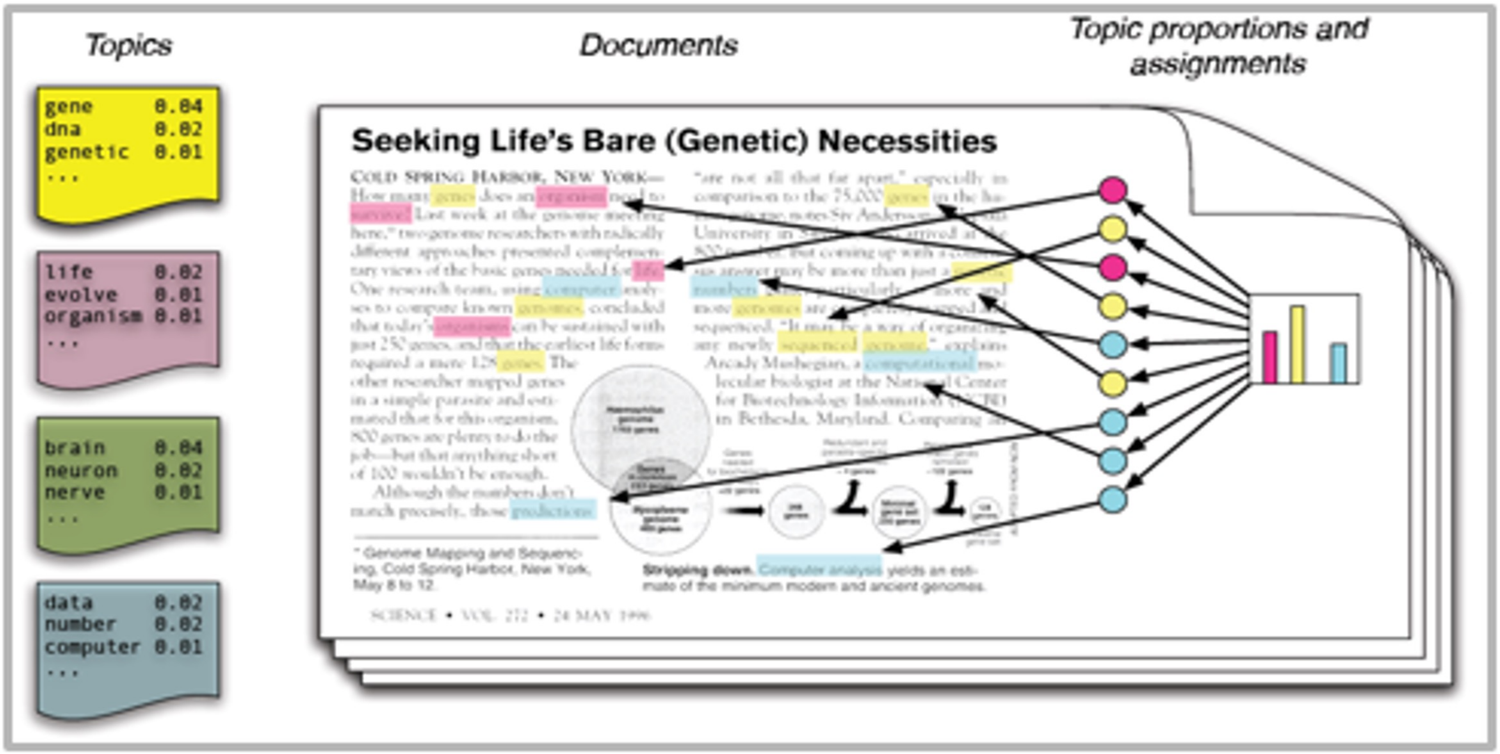
$$D = \sqrt{\sum_i (x_{1_i} - x_{2_i})^2}$$

Outline

- Clustering
- Topic Modeling – LDA
    - Background: Multinomial, Dirichlet Distributions

Topic Modeling

- Goal: Identify underlying topics across documents



 Slide from Federico Nanni

What are topics?

Observation

Tokens that are likely to appear in the same context

Hidden structure that determines how **tokens** appear in a corpus

Want to uncover

 Slide from Federico Nanni

Topic Modeling: Corpora -> Topics

Input:

Millions of Books

Output: topics

(distributions over words)

killed wounded sword slain arms military rifle wounds loss
human Plato Socrates universe philosophical minds ethics
inflammation affected abdomen ulcer circulation heart
ships fleet sea shore Admiral vessels land boats admiral
sister child tears pleasure daughters loves wont sigh warm
sentence clause syllable singular examples clauses syllables
provinces princes nations imperial possessions invasion
women Quebec Women Iroquois husbands thirty whom
steam engines power piston boilers plant supplied chimney
lines points direction planes Lines scale sections extending

Each row is a topic

 Slide from David Mimno

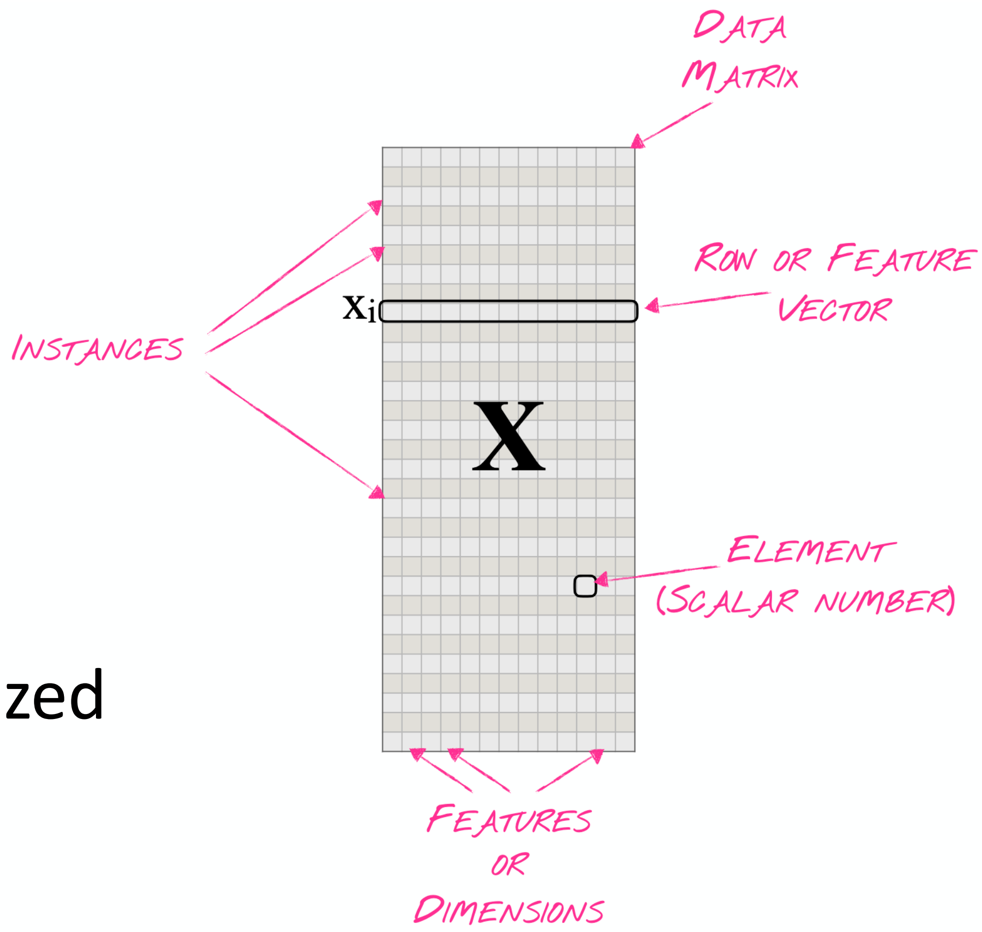
Background

Each row represents a Document vector

Number of times each word appeared

A distribution of discrete outcomes, when normalized sums to 1

Multinomial Distribution!

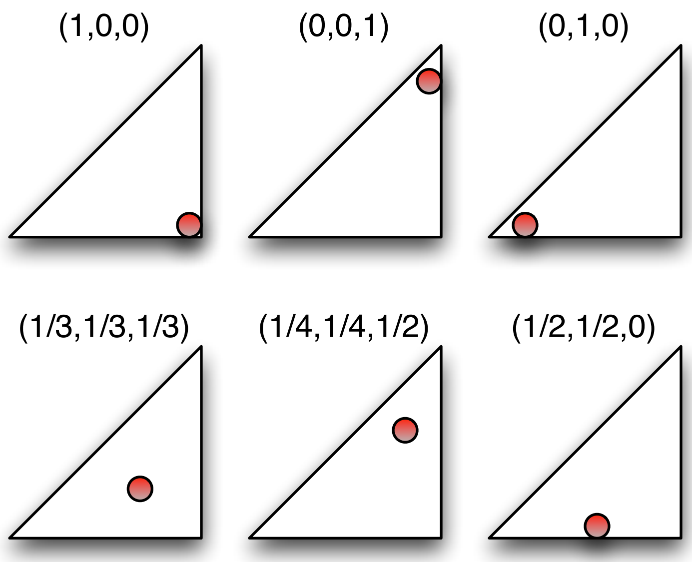


Background

Each row represents a Document vector

Number of times each word appeared

A distribution of discrete outcomes, when normalized sums to 1



Multinomial Distribution!

[Slide from Jordan Boyd-Graber]

Background

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

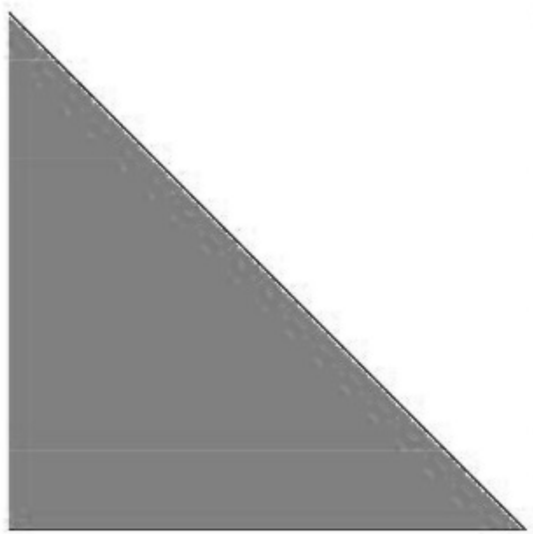
Dirichlet Distribution: Distribution over the multinomial distributions

Slide from Jordan Boyd-Graber

Background

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

Dirichlet Distribution:
Distribution over the multinomial distributions



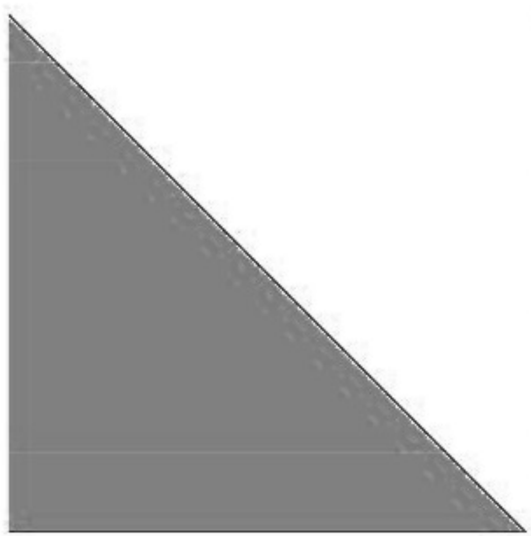
$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Slide from Jordan Boyd-Graber

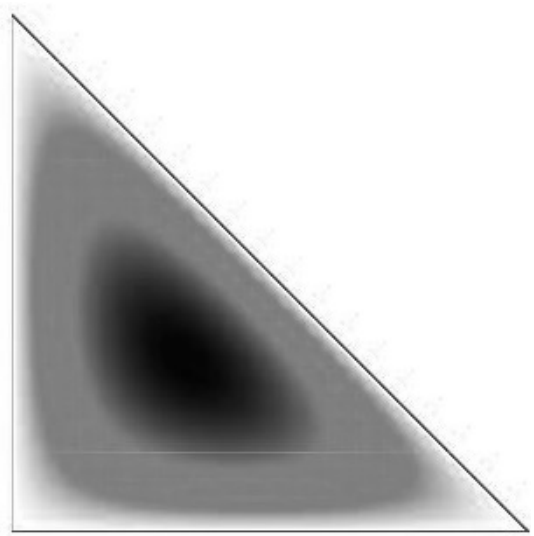
Background

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

Dirichlet Distribution:
Distribution over the multinomial distributions



$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$



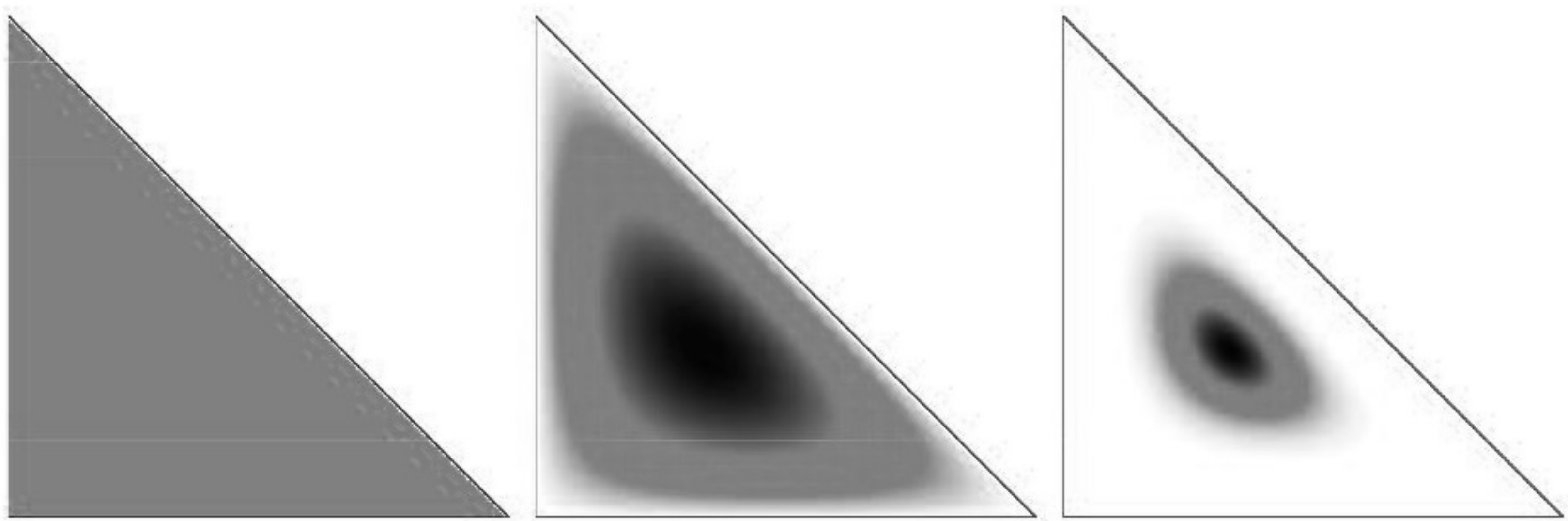
$\alpha = 6, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Slide from Jordan Boyd-Graber

Background

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

Dirichlet Distribution:
Distribution over the multinomial distributions



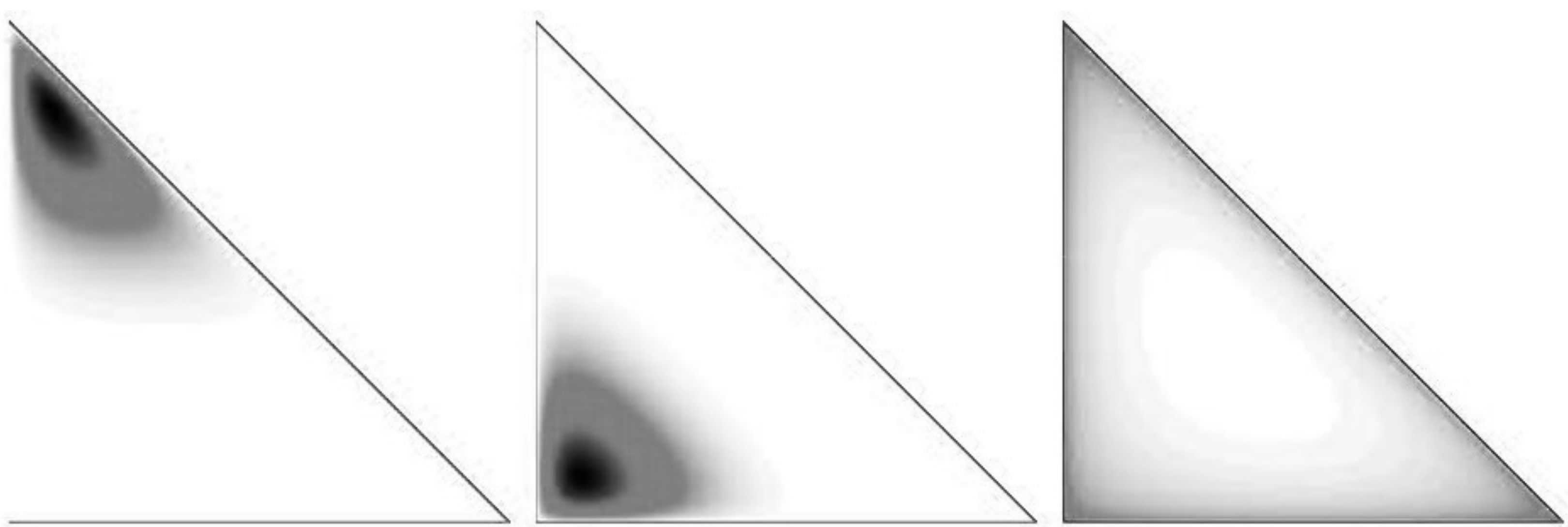
$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 6, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 30, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Slide from Jordan Boyd-Graber≈

Background

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

Dirichlet Distribution:
Distribution over the
multinomial distributions



$\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7})$ $\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7})$ $\alpha = 2.7, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Slide from Jordan Boyd-Graber

Discovering Topics

How do we discover topics?

- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Latent Dirichlet Allocation

How do we discover topics?

- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- **Latent Dirichlet Allocation**

LDA

- Probabilistic model
- Generative model

LDA Generative Story

- Each word appears independent of each other
- Each word depends on the topic
  - Topics have a distribution of words

Distribution of topics over words

TOPIC 1

computer, technology, system, service, site, phone, internet, machine

TOPIC 2

sell, sale, store, product, business, advertising, market, consumer

TOPIC 3

play, film, movie, theater, production, star, director, stage

- Each topic is a multinomial distribution over words

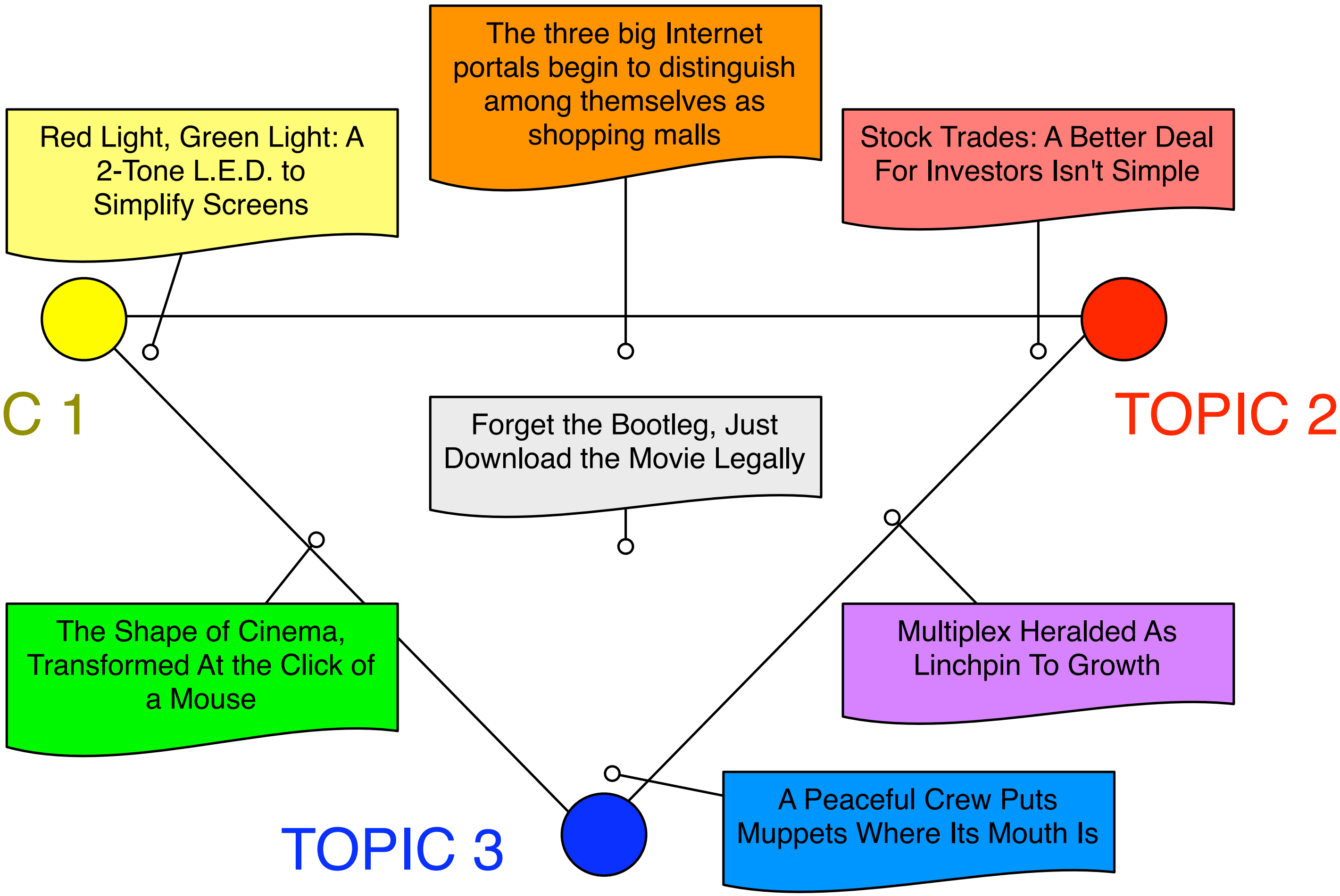
Slide from Jordan Boyd-Graber≈

LDA Generative Story

- Each word appears independent of each other

- Each word depends on the topic
  - Topics have a distribution of words
  - Topics have a distribution of documents

Distribution of topics over documents



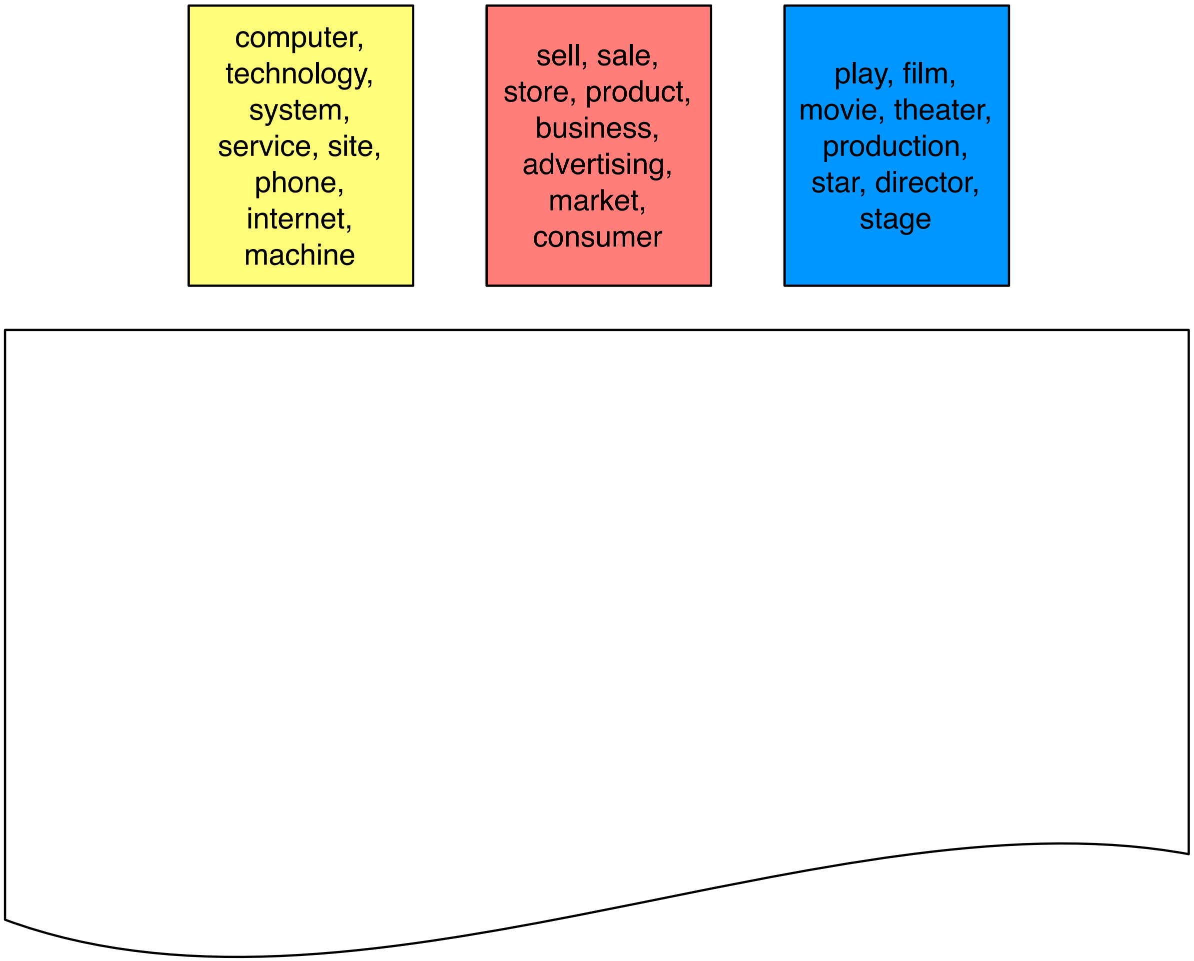
Slide from Jordan Boyd-Graber≈

LDA Generative Story

- Each word appears independent of each other

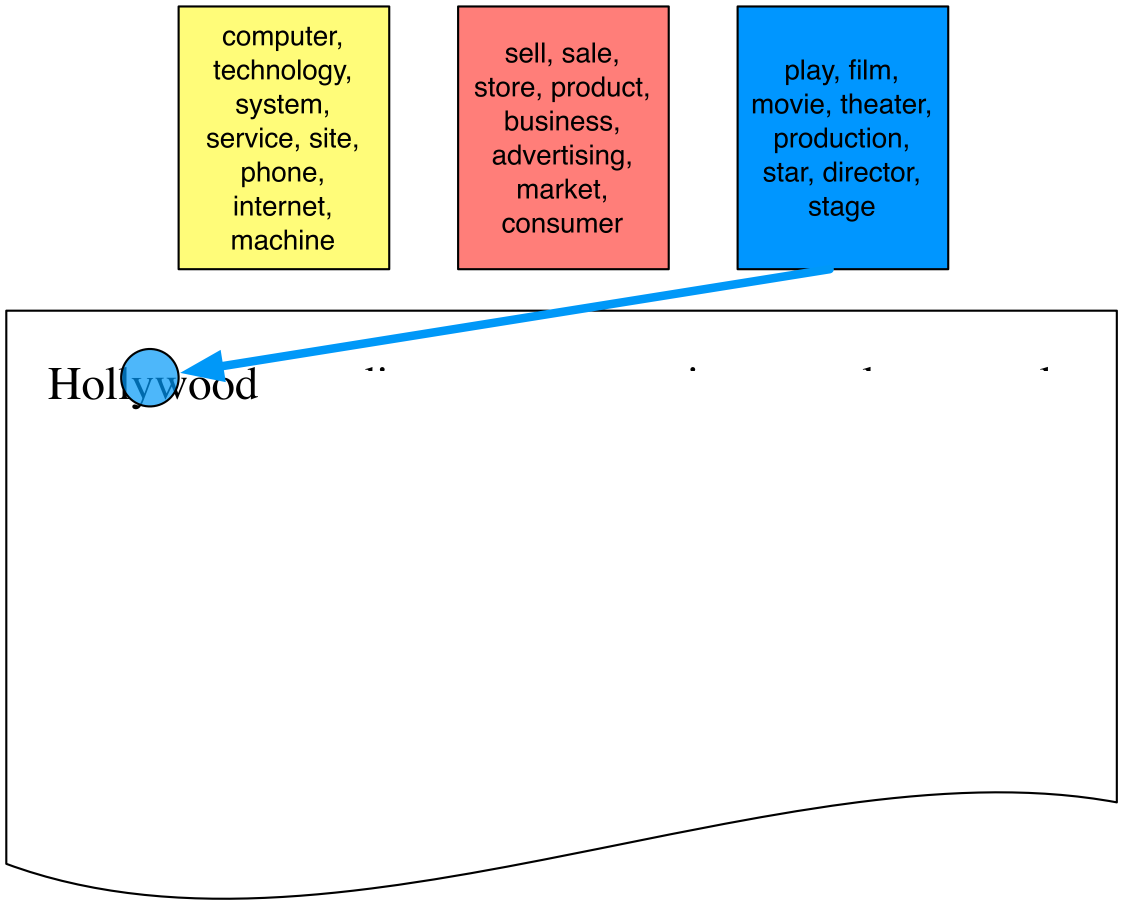
- Each word depends on the topic
  - Topics have a distribution of words
  - Topics have a distribution of documents
  - Both are multinomial distributions!

Generating a document



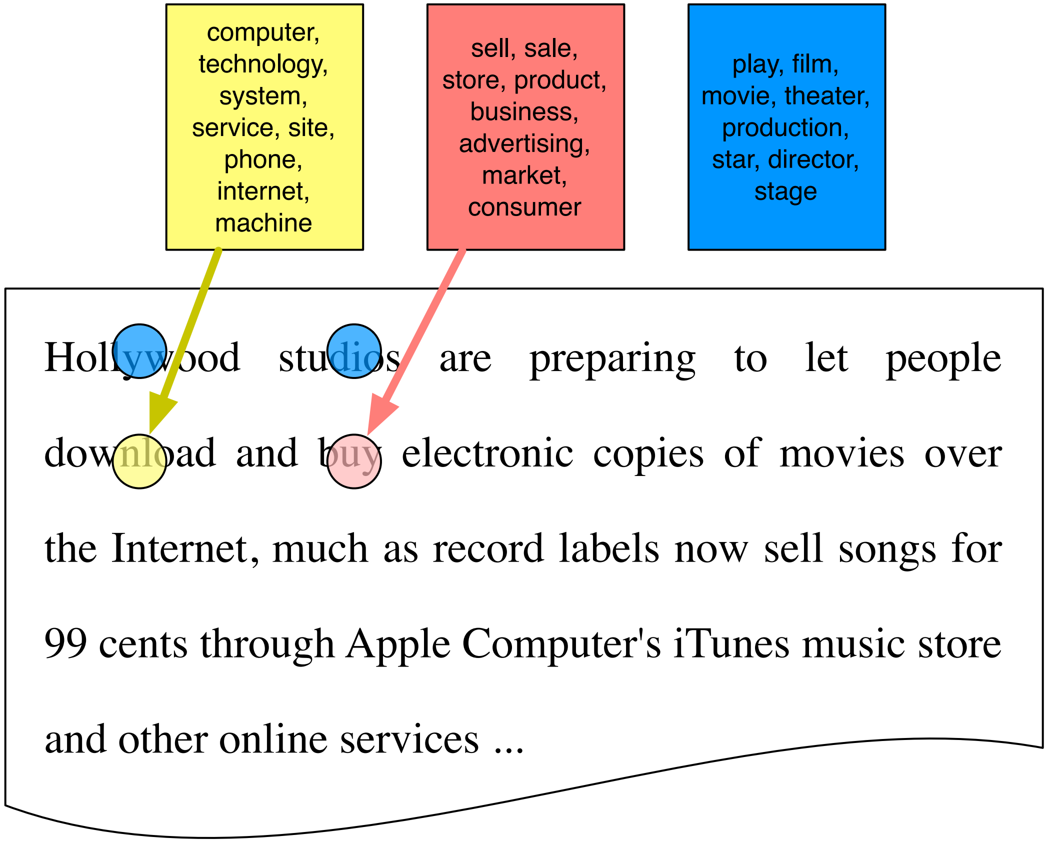
Slide from Jordan Boyd-Graber≈

Generating a document



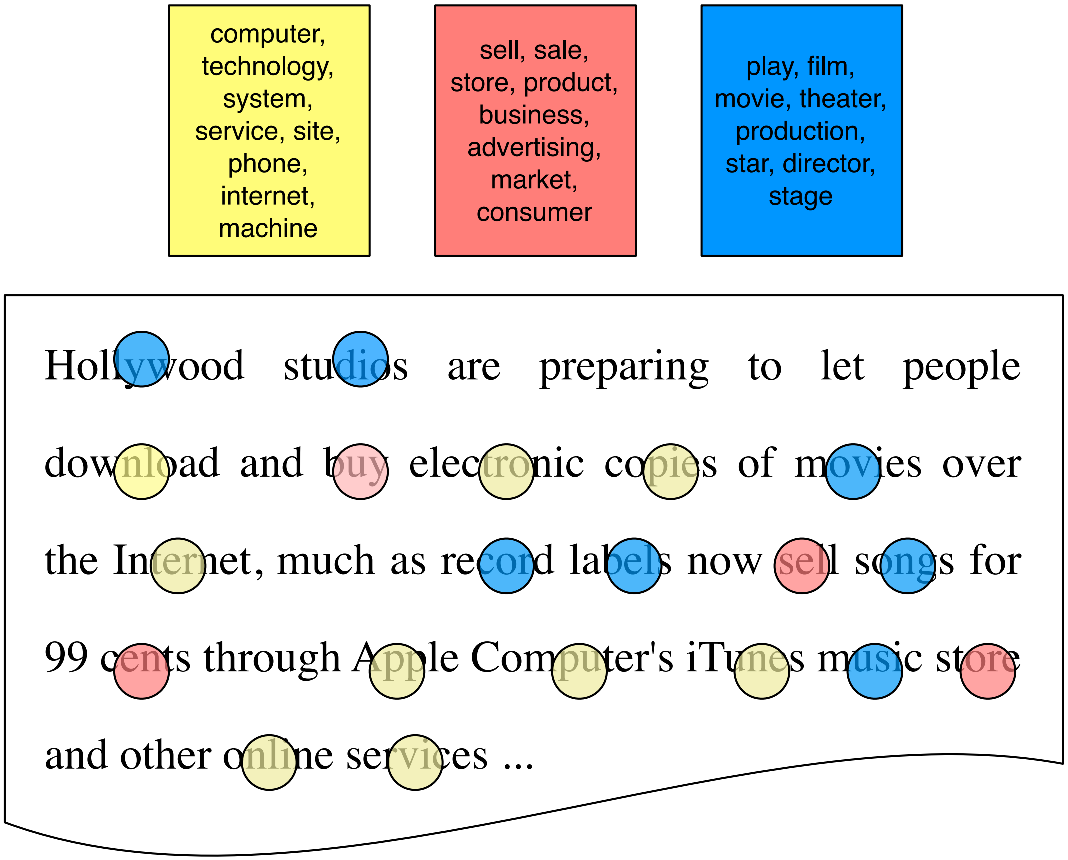
[Slide from Jordan Boyd-Graber]

Generating a document



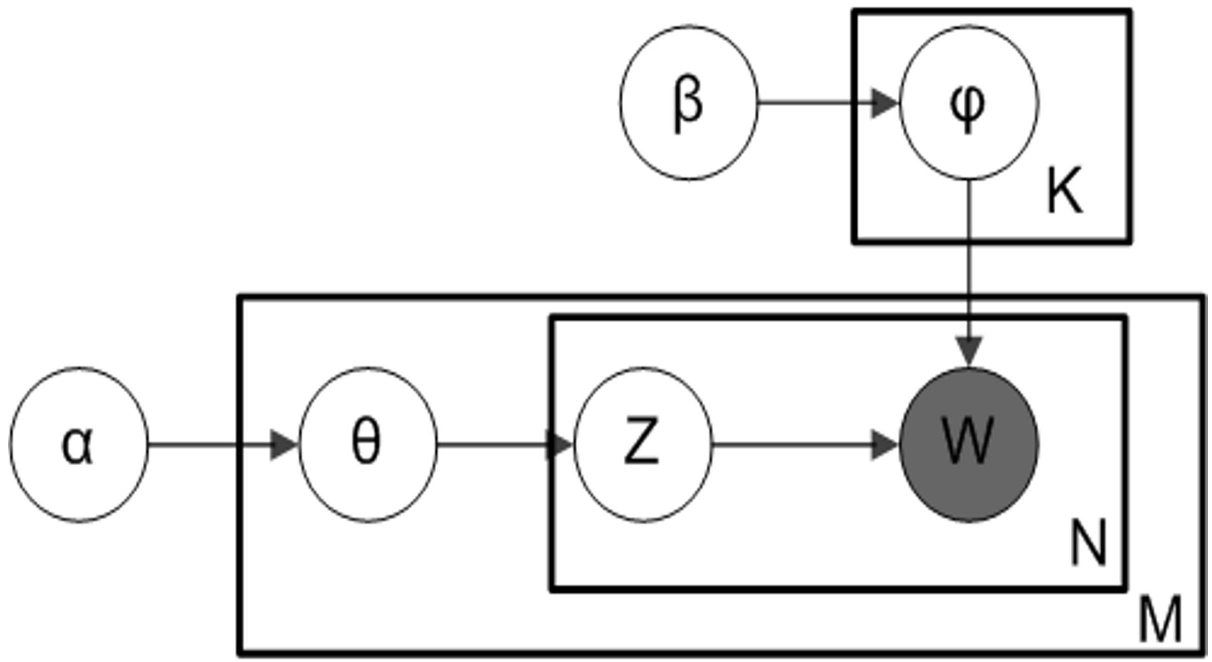
Slide from Jordan Boyd-Graber

Generating a document



Slide from Jordan Boyd-Graber

LDA Plate Notation

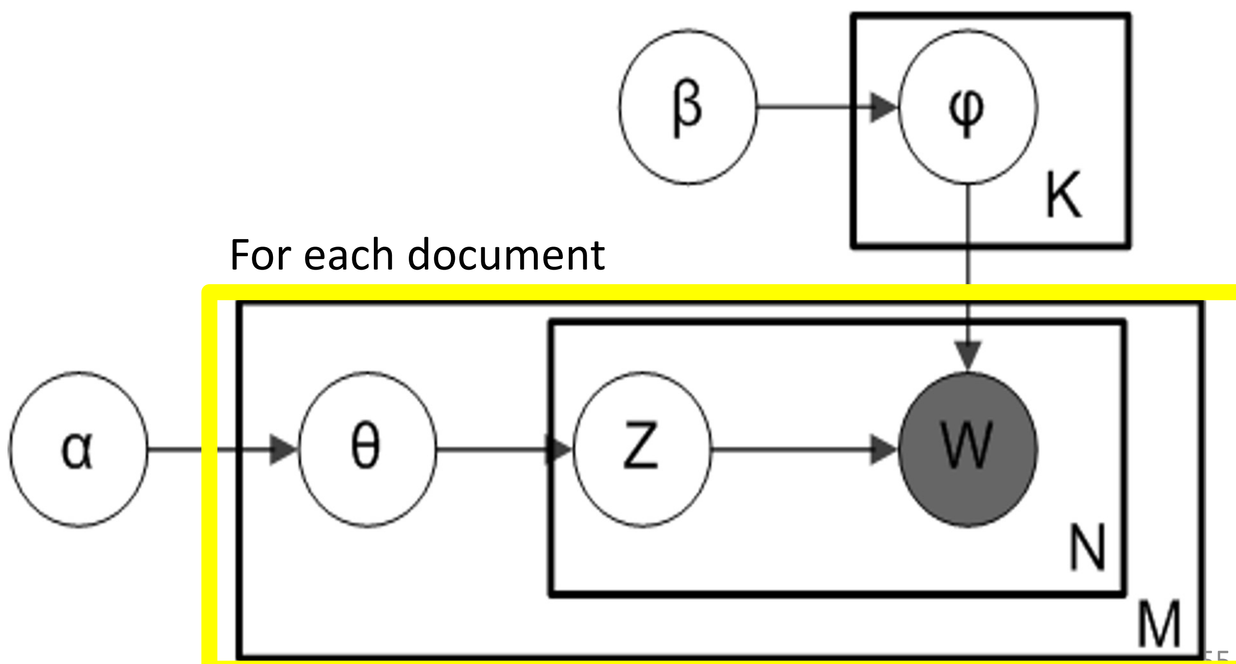


LDA Plate Notation

M = number of documents

N = number of words in a document

K = number of topics (we choose this)

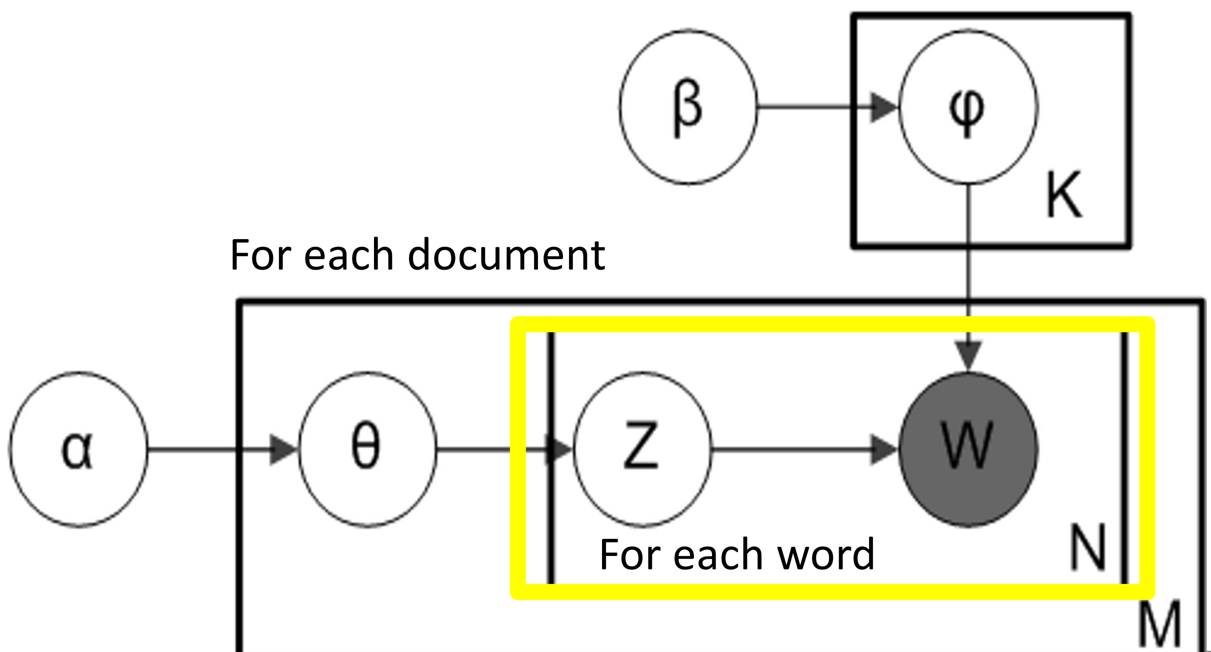


LDA Plate Notation

M = number of documents

N = number of words in a document

K = number of topics (we choose this)

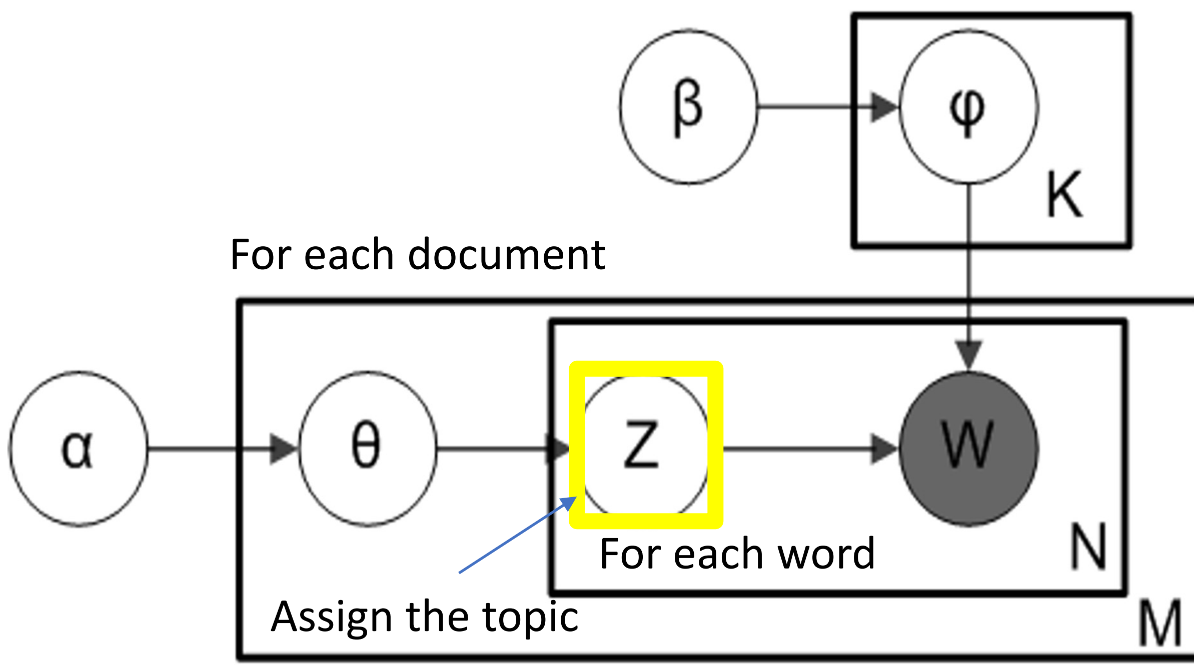


LDA Plate Notation

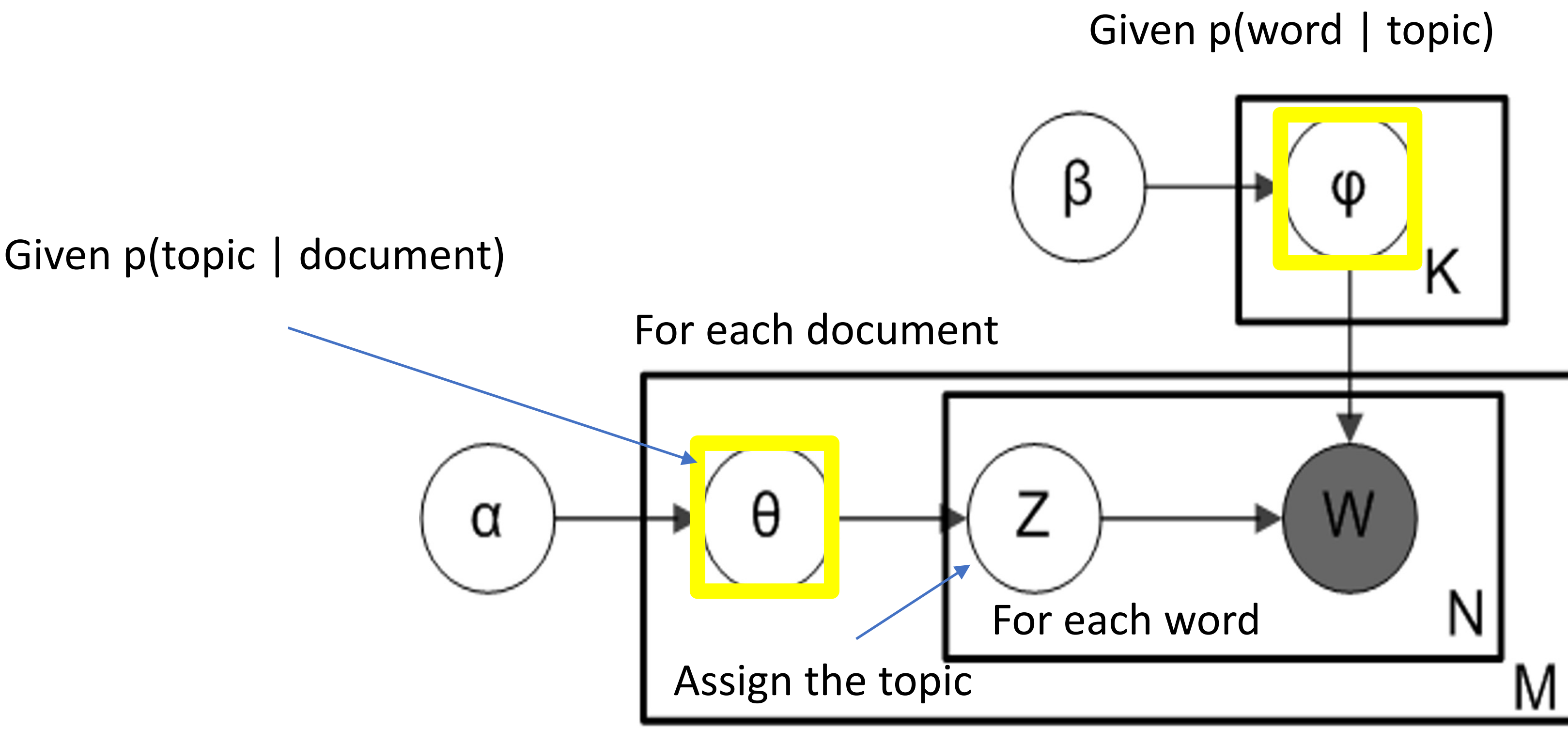
M = number of documents

N = number of words in a document

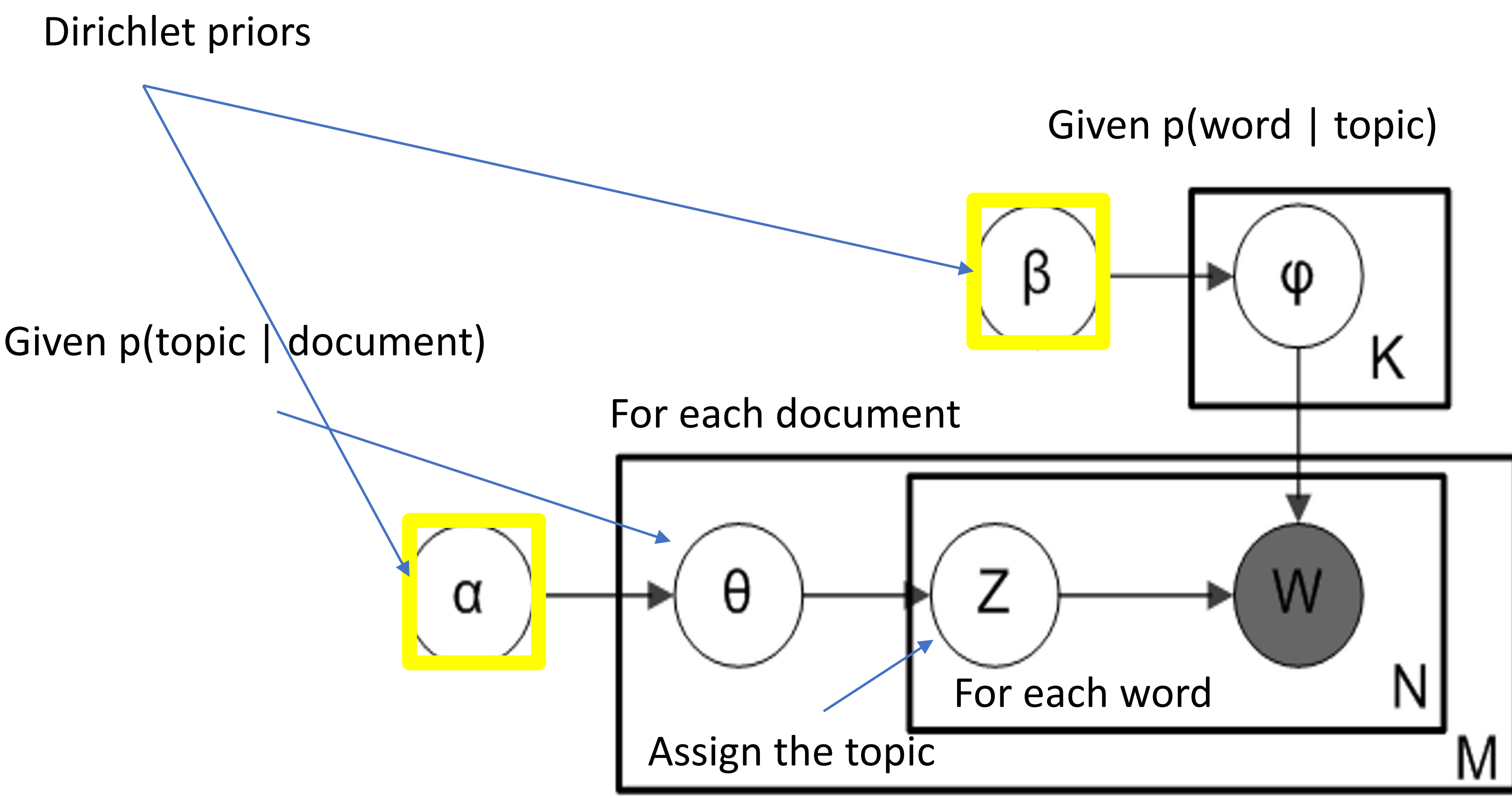
K = number of topics (we choose this)



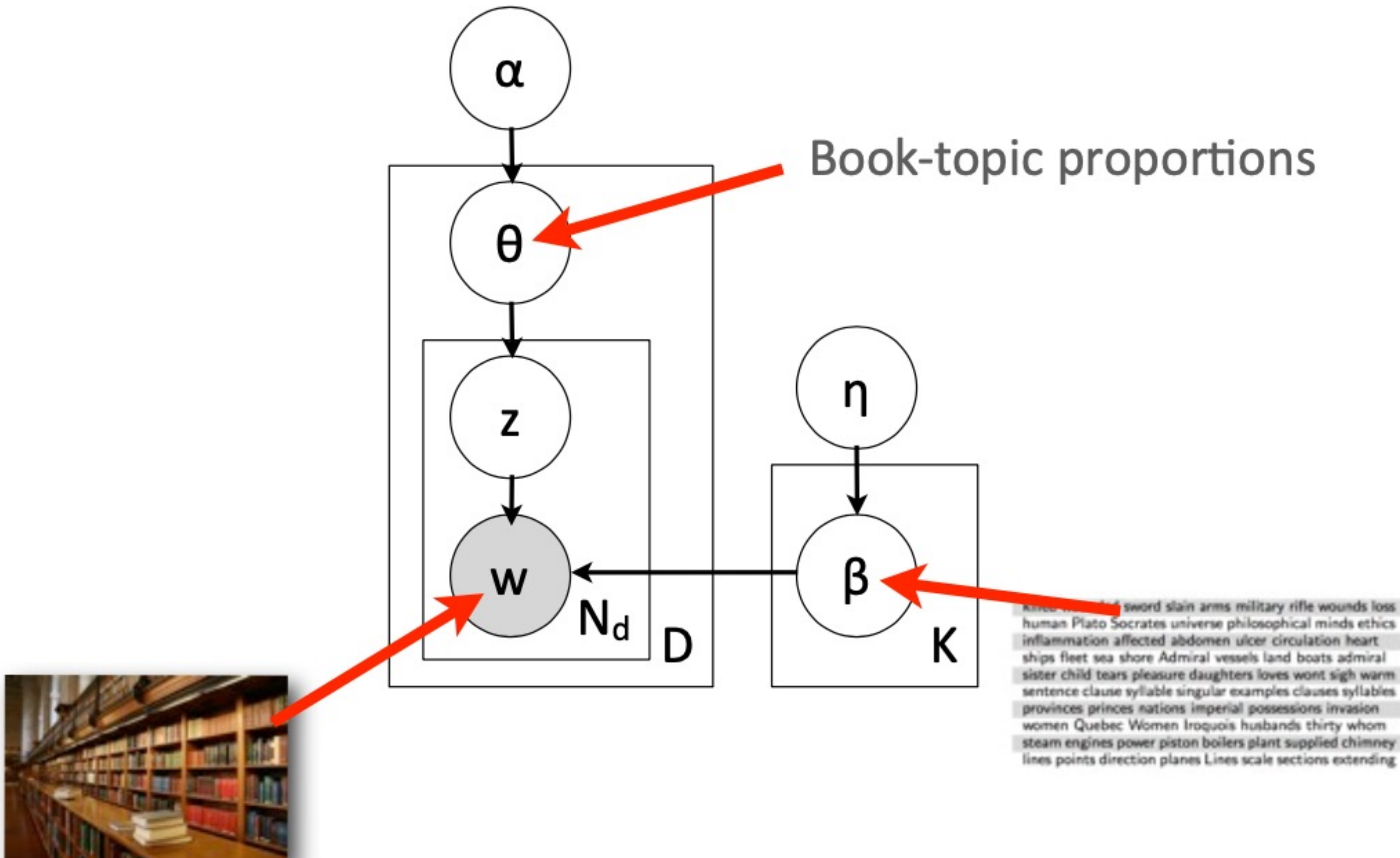
LDA Plate Notation



LDA Plate Notation



LDA Plate Notation



Slide from David Mimno

LDA Plate notation

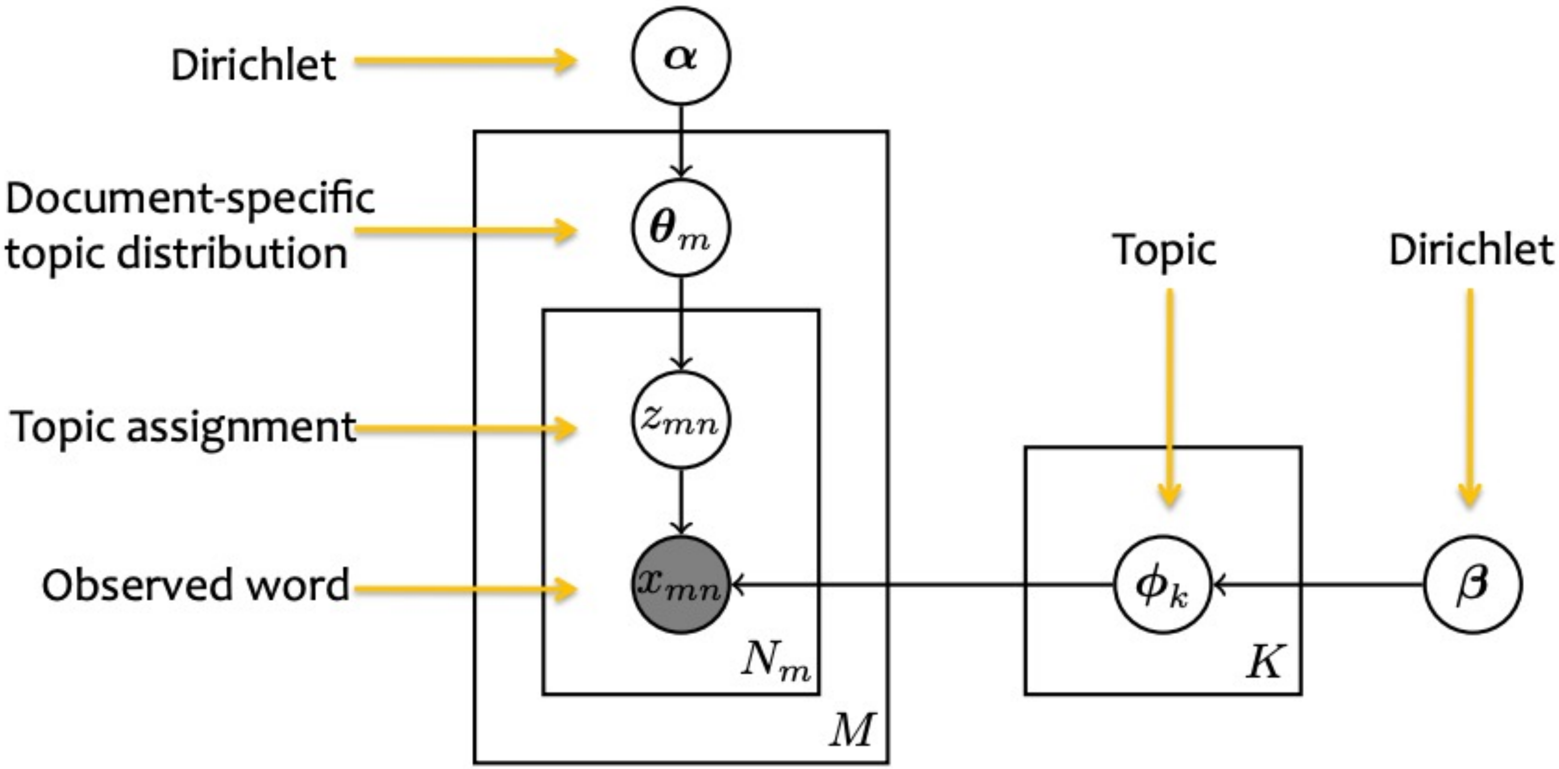
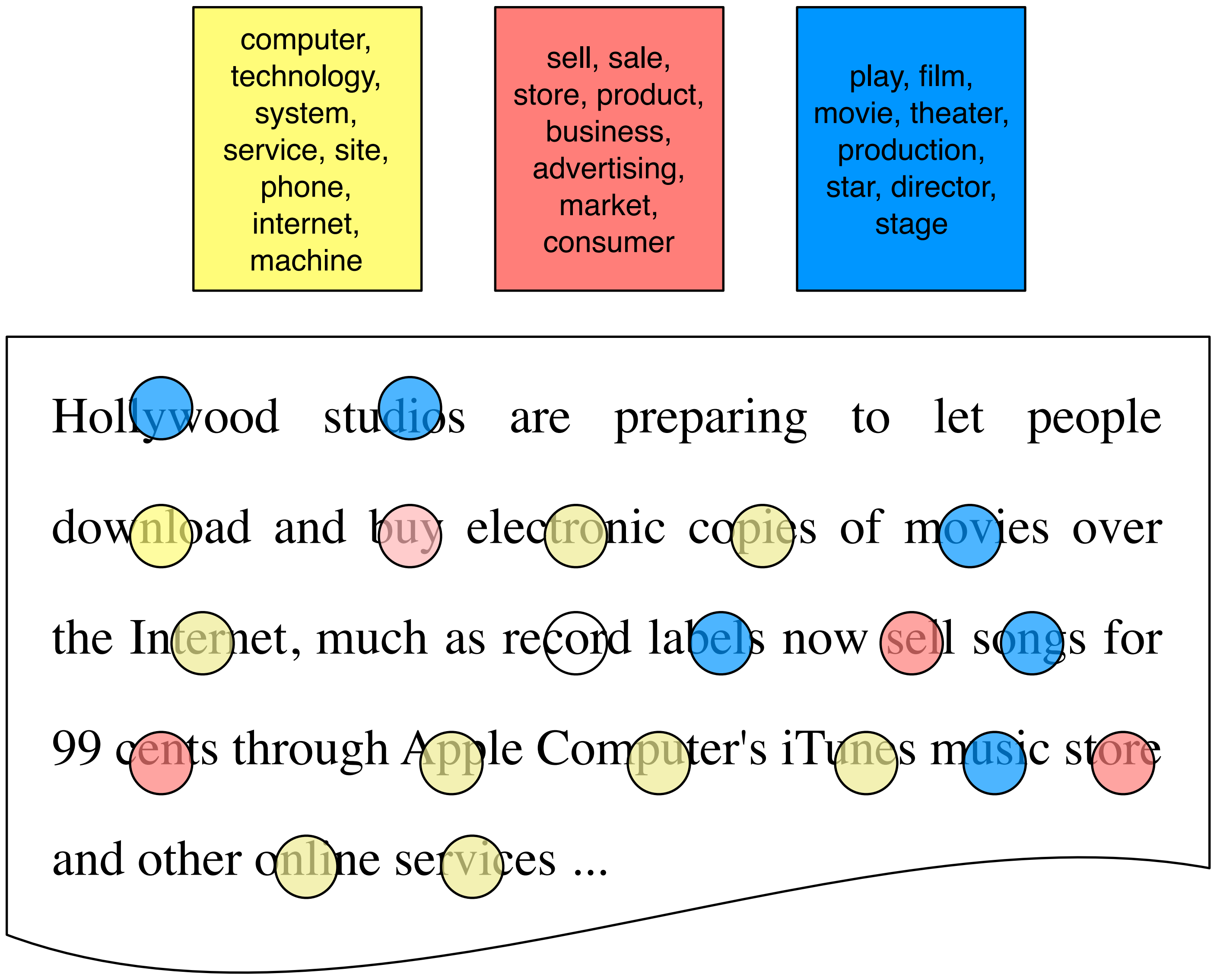
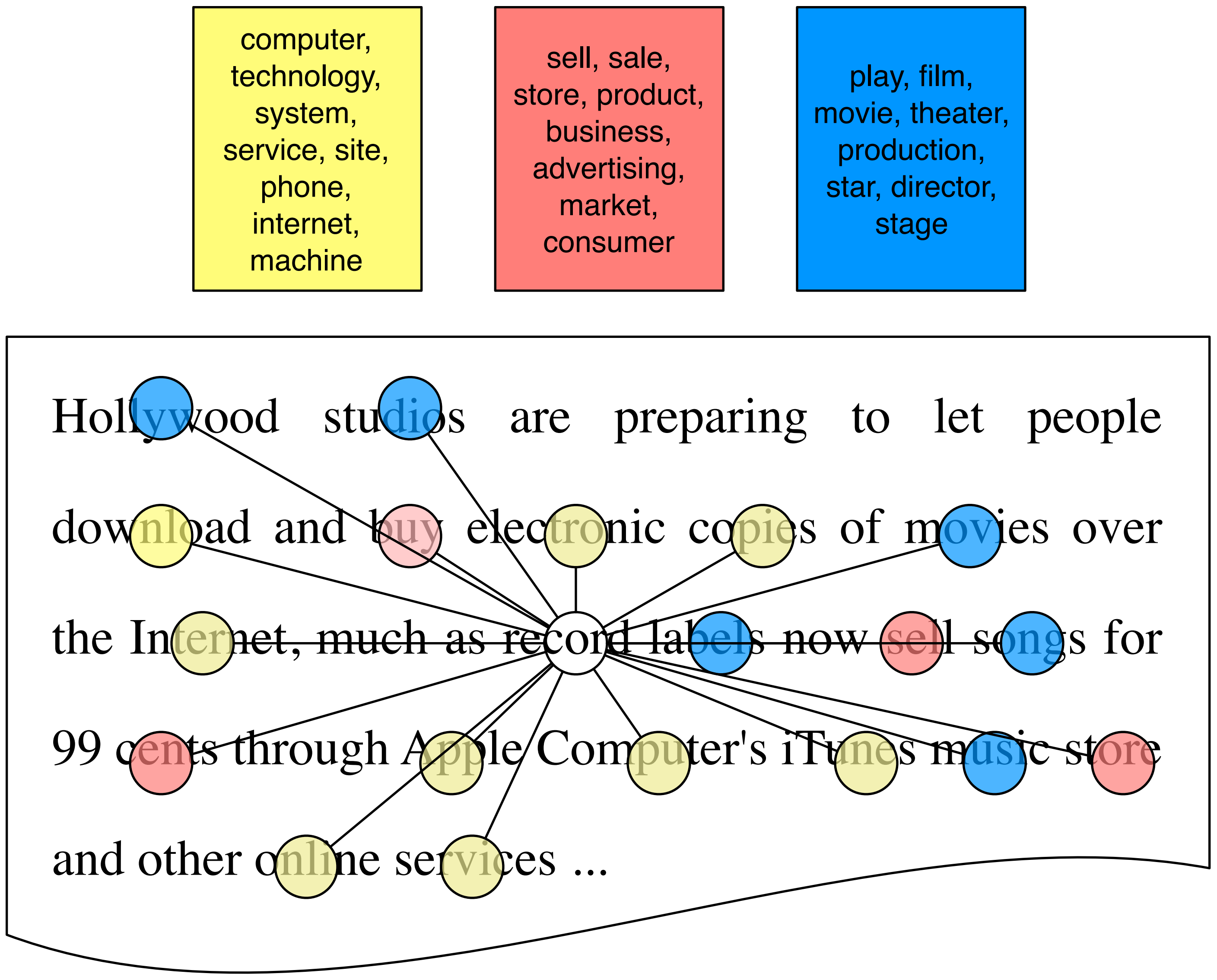


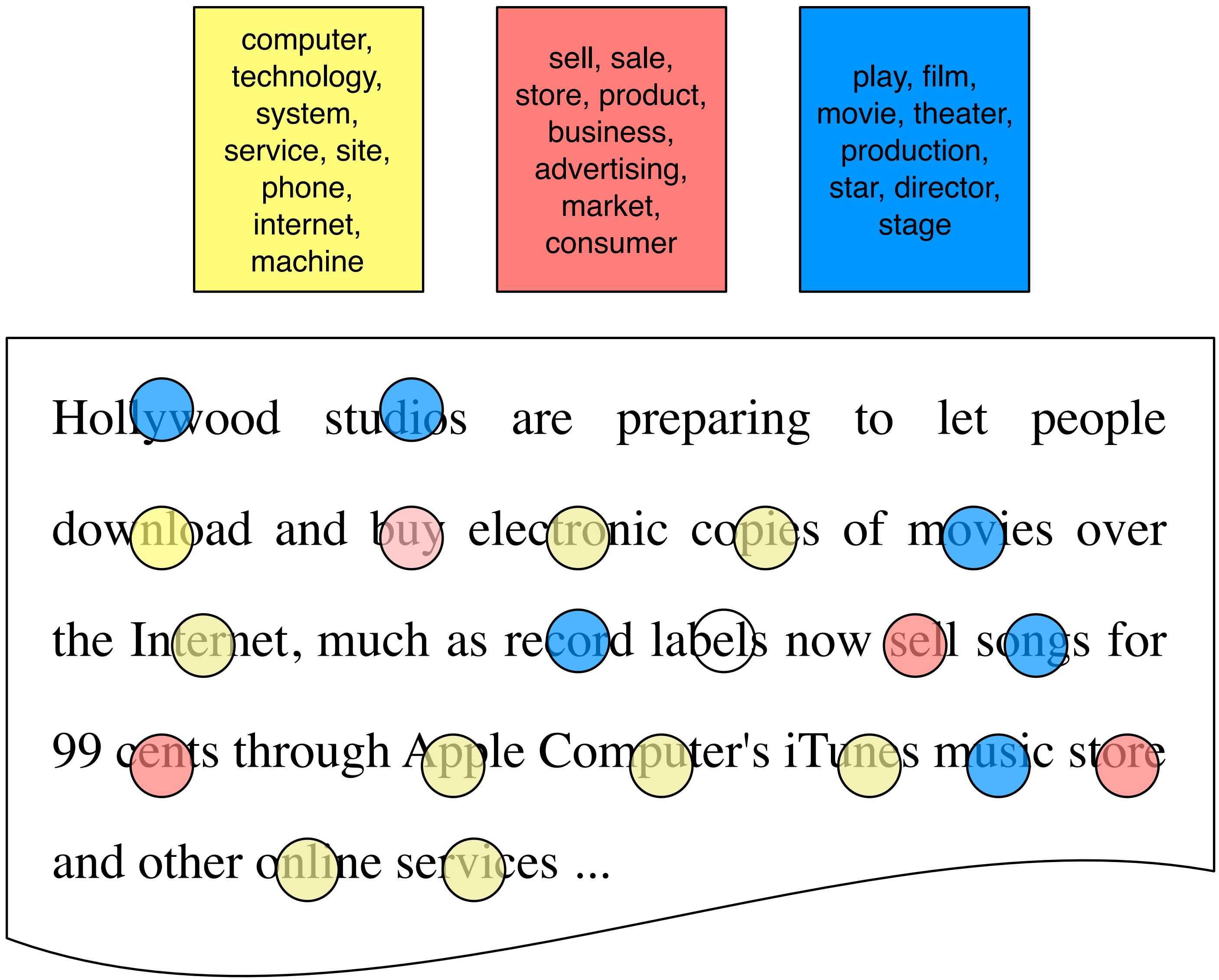
 Figure from Matt Gormley

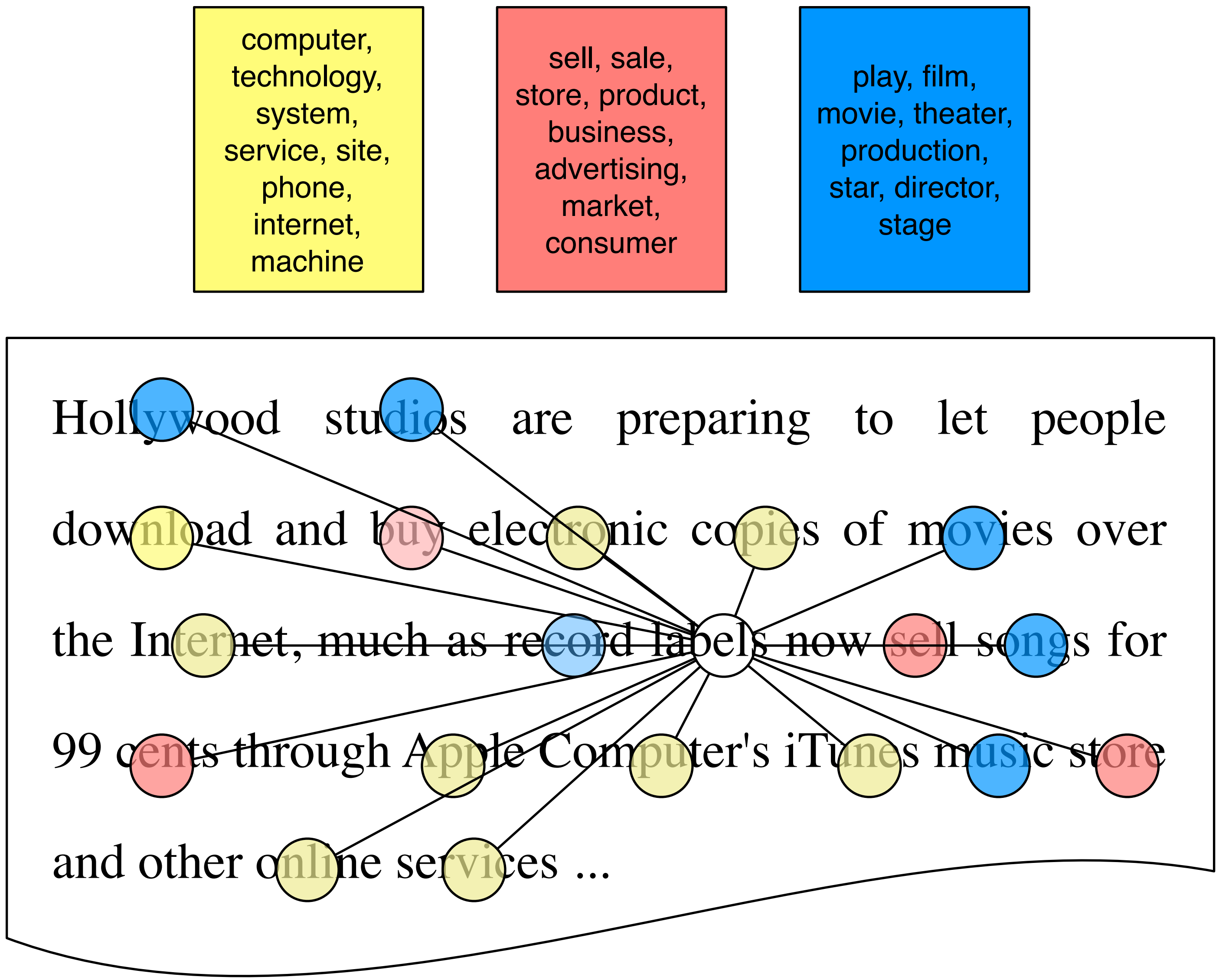
LDA Algorithm


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...



computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

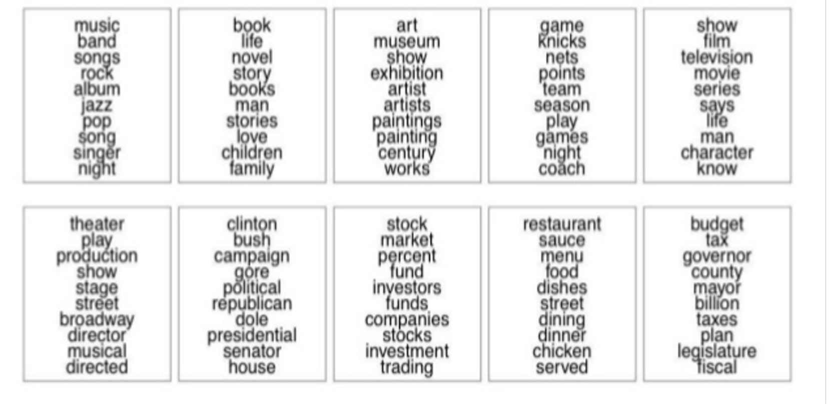

computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...



computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Training LDA Model

1. Randomly assign words to topics
2. Repeat many times:
    1. For each document:
        1. For each token, re-assign the topic based on:
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens



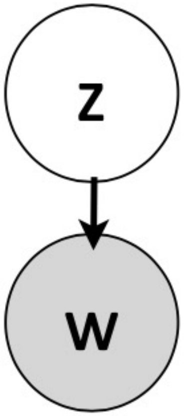
Training LDA Model

1. Randomly assign words to topics
2. Repeat many times:
    1. For each document:
        1. For each token, re-assign the topic based on:
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

Randomly assign words to topics

Etruscan	trade	price	temple	market

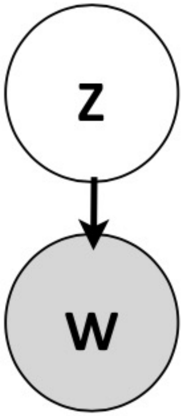
Randomly assign words to topics



Etruscan	trade	price	temple	market

 Example David Mimno

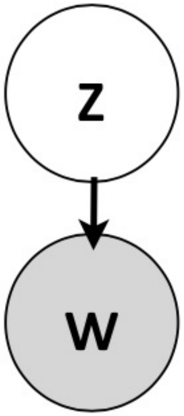
Randomly assign words to topics



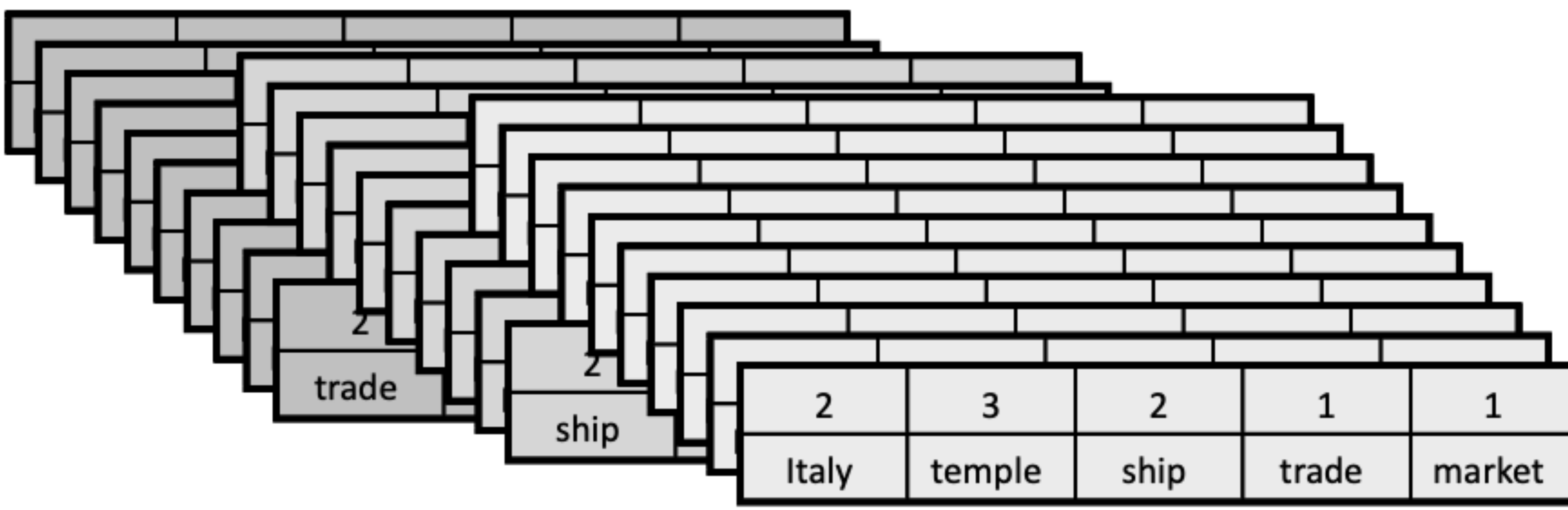
3	2	1	3	1
Etruscan	trade	price	temple	market

 Example David Mimno

Randomly assign words to topics



3	2	1	3	1
Etruscan	trade	price	temple	market



Example David Mimno

Global Statistics from Random Topic Assignments

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts across corpus

	1	2	3
Etruscan	1	0	35
trade	10	8	1
price	42	1	0
market	50	0	1
temple	0	0	20
…			

Example David Mimno

Training LDA Model

1. ~~Randomly assign words to topics~~
2. Repeat many times:
   1. For each document:
      1. For each token, re-assign the topic based on:
         1. Topic assignment for every other token in the document
         2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

Training LDA Model

1. ~~Randomly assign words to topics~~
2. Repeat many times:
   1. For each document:
      1. For each token, re-assign the topic based on:
         1. Topic assignment for every other token in the document
         2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

Reassign topic for “Trade”

3	2	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
trade	10	8	1
price	42	1	0
market	50	0	1
temple	0	0	20
…			

Reassign topic for "Trade"

3	2	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
trade	10	8	1
price	42	1	0
market	50	0	1
temple	0	0	20
…			

Reassign topic for "Trade"

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
trade	10	8	1
price	42	1	0
market	50	0	1
temple	0	0	20
…			

Reassign topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
trade	10	7	1
price	42	1	0
market	50	0	1
temple	0	0	20
…			

Pick a topic for "Trade"

3	?	1	3	1
Etruscan	trade	price	temple	market

Which topics occur in this document?

Pick a topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

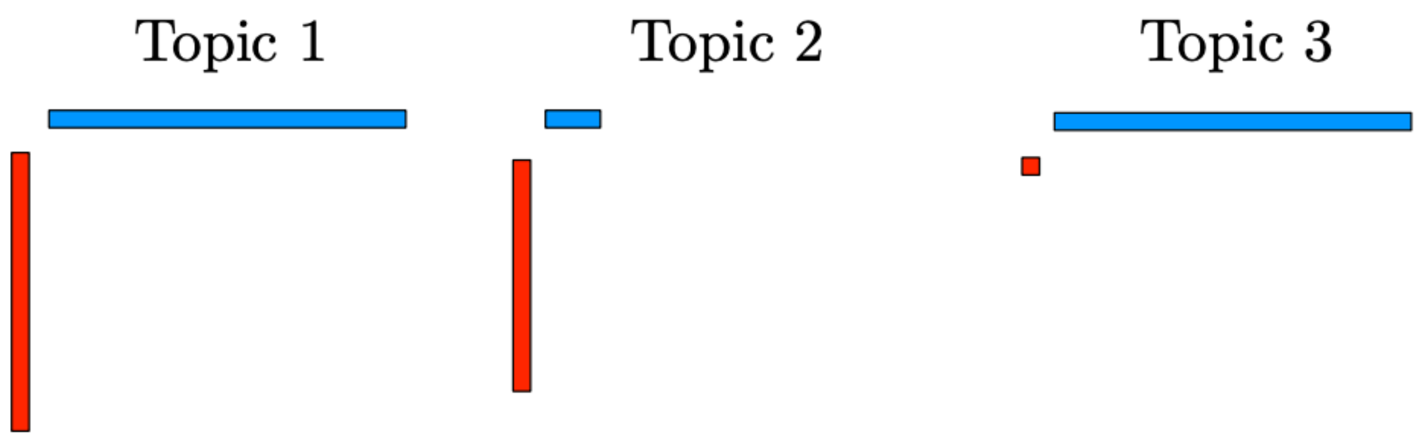
Which topics like the word-type “trade”?

	1	2	3
trade	10	7	1

Pick a topic for "Trade"

3	?	1	3	1
Etruscan	trade	price	temple	market

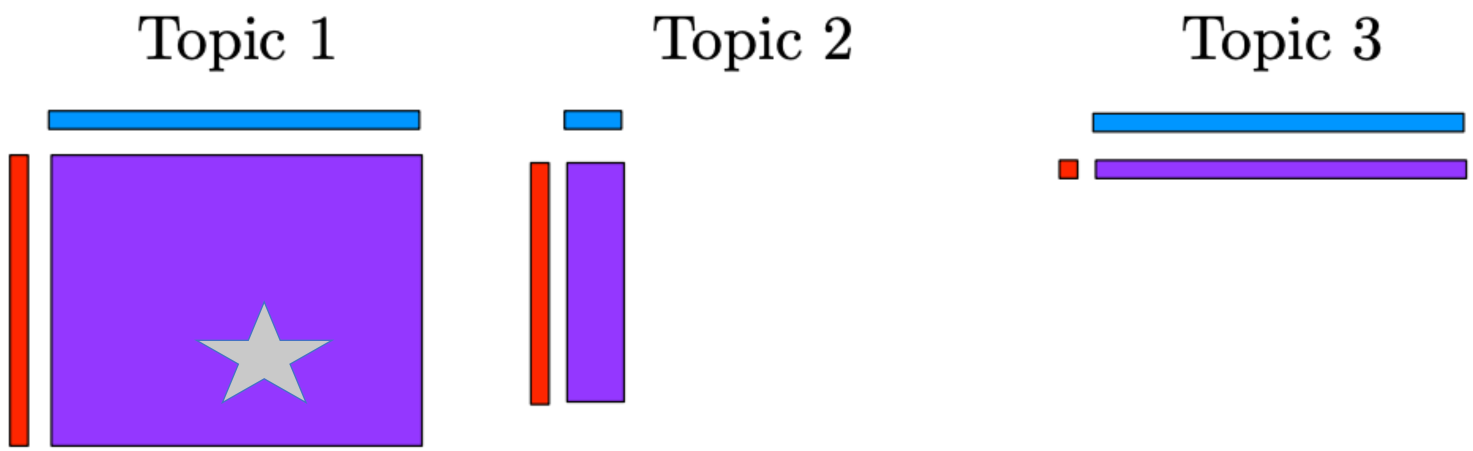
Which topics like the word "trade"?



Pick a topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

Pick a topic for “trade”?



Update topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
trade	10	7	1
price	42	1	0
market	50	0	1
temple	0	0	20
…			

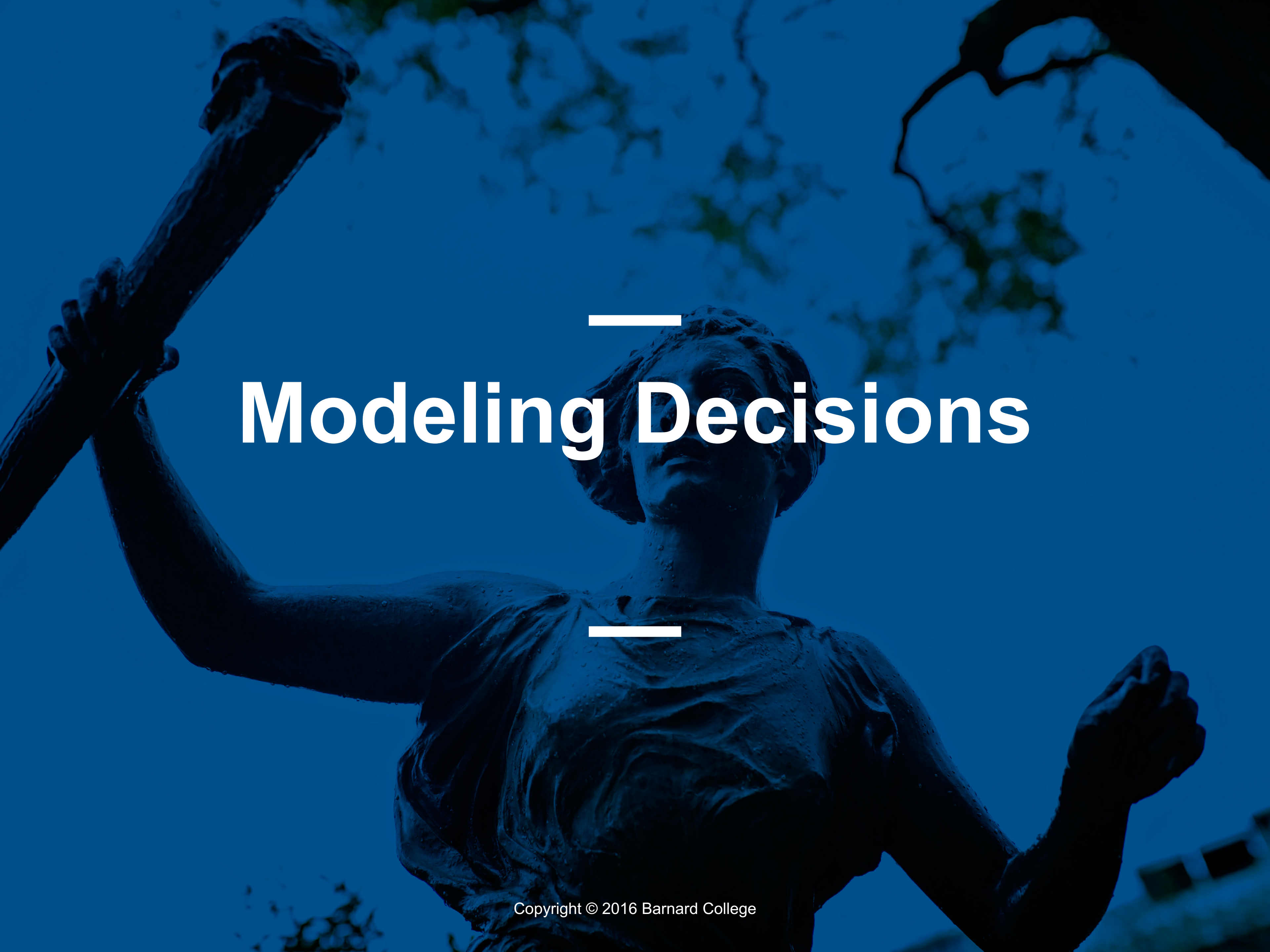
Update topic for “Trade”

3	1	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
trade	11	7	1
price	42	1	0
market	50	0	1
temple	0	0	20
…			

Training LDA Model – Gibbs Sampling

1. **Randomly assign words to topics**
2. Repeat many times:
    1. For each document:
        1. **For each token, re-assign the topic based on:**
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens



Modeling Decisions

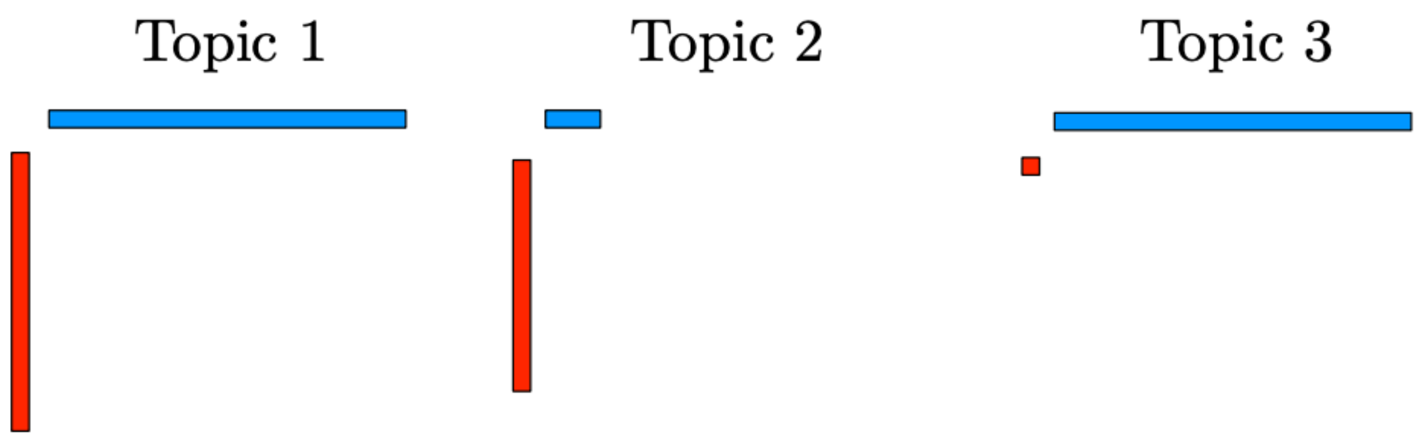
Modeling decisions – hard choices

- Document definition
- Interesting words
- Knobs:
  - K - Number of topics
  - Hyper-parameters

Hyperparameters

3	?	1	3	1
Etruscan	trade	price	temple	market

Which topics like the word “trade”?

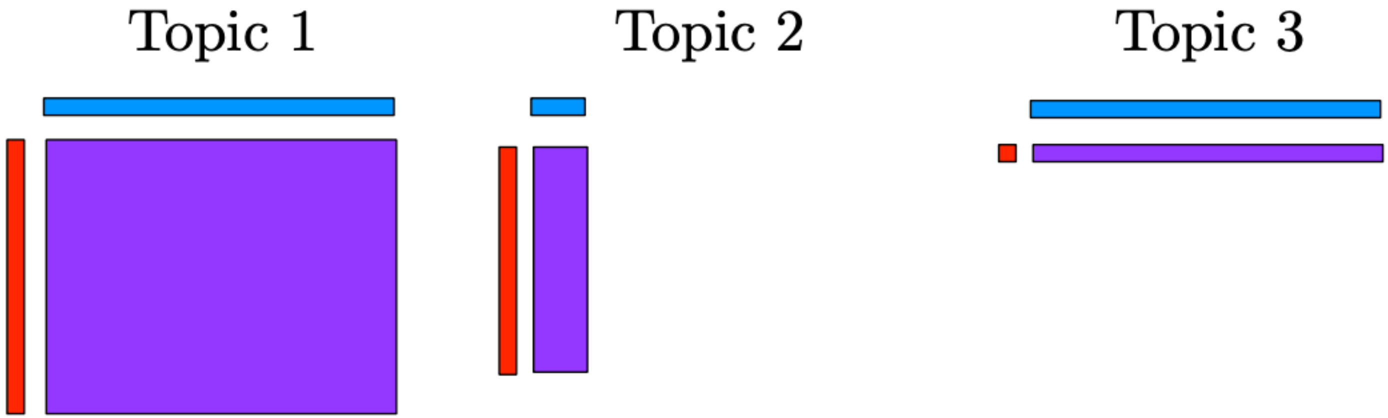


Hyperparameters - alpha

Hyperparameters

3	?	1	3	1
Etruscan	trade	price	temple	market

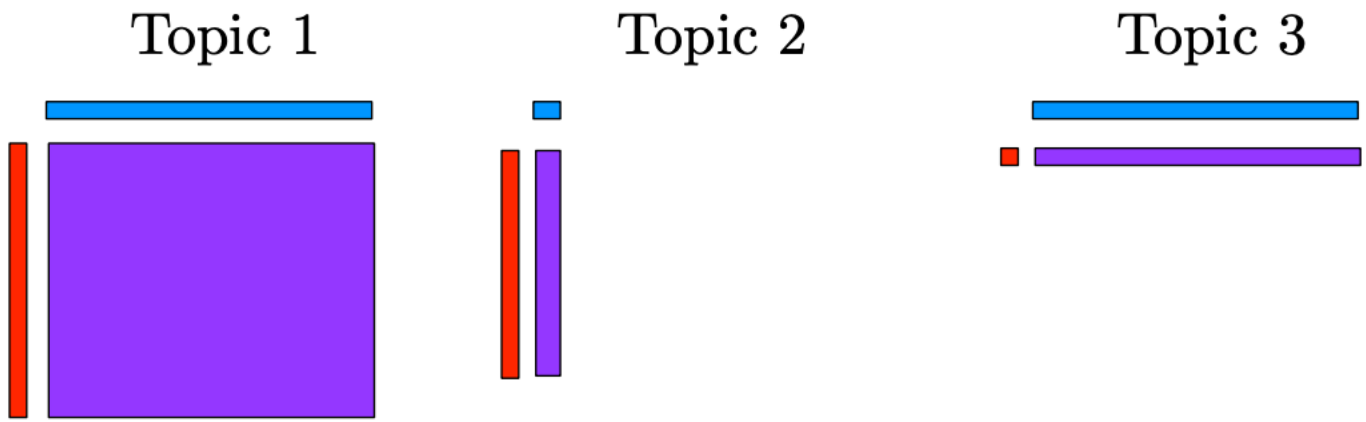
Which topics like the word "trade"?

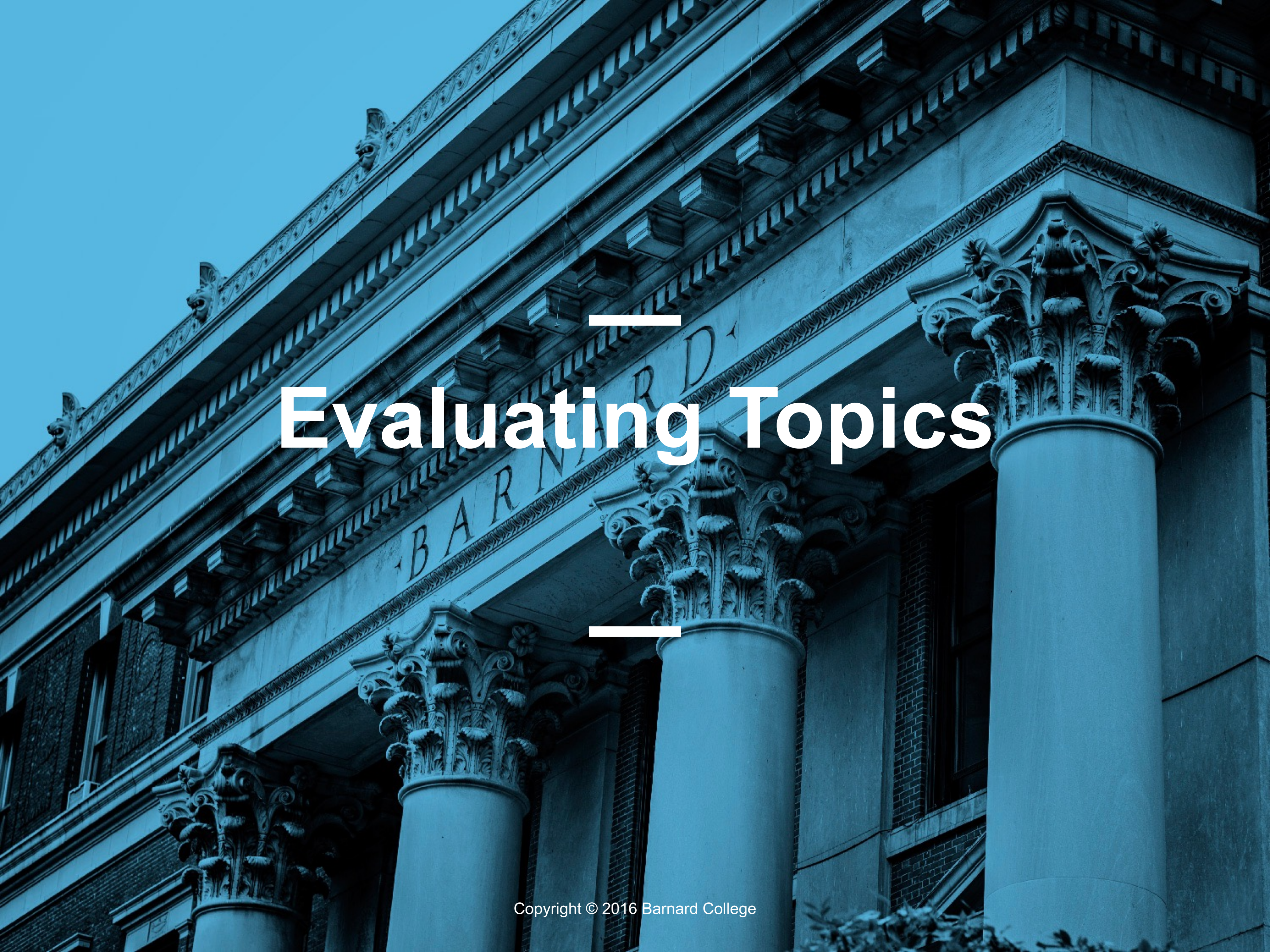


Hyperparameters

3	?	1	3	1
Etruscan	trade	price	temple	market

Which topics like the word "trade"?





Evaluating Topics

Output of topic models

Top 10 topic terms
face, problem, depress, econom, suffer, economi, caus, great depress, crisi, prosper
bank, money, tax, pay, debt, loan, rais, fund, paid, govern
worker, labor, work, union, job, employ, strike, factori, industri, wage
govern, power, feder, nation, peopl, author, constitut, state, system, unit
roosevelt, wilson, peac, presid, treati, negoti, theodor roosevelt, taft, leagu, agreement
men, women, famili, children, young, work, woman, home, mother, husband
citi, york, urban, hous, live, town, center, communiti, move, chicago
railroad, build, line, technolog, transport, road, develop, travel, invent, canal
good, trade, product, manufactur, market, import, produc, economi, consum, tariff
farmer, farm, planter, small, land, cotton, plantat, crop, famili, larg

What makes topics bad?

- **Random**, unrelated words
- ***Intruder*** words
- Boring, **overly general** words
- **Chimaeras:**
  - Multiple topics combined

Evaluation – Word Intrusion Task

- Take top k words in a topic
  - Usually 5 or 10
- Substitute 1 word with a top word from another topic
- Shuffle the works
- Ask someone to pick the intruder
  - If they can pick the intruder – it's a good topic