



THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL

# Exploratory Analysis: Clustering

(some material taken or  
adapted from slides by Hinrich  
Schutze)

Heejun Kim

June 26, 2018

# Clustering

## objective

- Grouping documents or instances into subsets or clusters
- Documents in the same cluster should be similar
- Documents in different clusters should be dissimilar
- A common form of **unsupervised learning**
- Unsupervised = no human-produced labels
- The goal is to discover structure from the data

# Clustering vs. Classification

- Classification:

- ▶ the input to the system is a set of labeled data
- ▶ the algorithm learns a model for predicting the label on new examples

- Clustering:

- ▶ the input to the system is a set of unlabeled data
- ▶ the algorithm infers the labels from the data and assigns a label to each input instance



# Clustering

## applications

- **Search engine results clustering:** grouping search engine results by topic
  - ▶ the user can identify the relevant clusters and ignore the non-relevant ones
- **Collection clustering:** grouping documents by topic to support navigation and exploration
- **Data analytics:** grouping instances to identify popular trends (big clusters) and outliers (small clusters)

# Clustering Applications

## search engine results clustering

Results 1-20 of about 15,703,845 | [Details](#)

[Sources](#) [Sites](#) [Time](#) [Topics](#)

**Top 576 Results** [remix](#)

- + Time, Festival (9)
- + Land, Rover (117)
- + Parts (88)
- + Photos (57)
- + Club (55)
- + Jacksonville Jaguars (39)
- + Classic, Cars (38)
- + Reviews (30)
- + Sports Cars (15)
- + Game (26)
- + Team (22)
- + New And Used Jaguar (17)
- + Atari (16)
- + Jaguar X Type (15)
- + Defend, Largest (6)
- + Jaguar Enthusiasts (12)
- + Jaguar For Sale (8)
- + Kits (11)
- New Jaguar dealership (6)
- + Big Cat (5)
- + University (12)
- + Tiger (9)
- Virginia, Washington (4)
- + Experiences (6)
- + Vintage, Car (8)
- + New And Used Cars (6)
- Autotrader, Jaguar Cars, Find (3)
- ...

[Market Selector | Jaguar | View the site in your preferred ...](#) [new window](#) [preview](#)

You are about to leave **Jaguar.com**. Please note that **Jaguar** cannot be responsible for any content or validity outside of this domain. Please click on Accept to go ...

<https://www.jaguar.com> - - Yippy Index V

[The week in wildlife – in pictures](#) [new window](#) [preview](#)

**Date:** 2017-03-10T14:02:09.000Z, 2017-03-10T14:02:09.000Z, 2017-03-10T14:02:09.000Z

A rare **jaguar** sighting in the US, a green toad and spring flowers are among this week's pick of images from the natural world ... A rare **jaguar** sighting in the US, a green toad and spring flowers are among this week's pick of images from the natural world ... A rare **jaguar** sighting in the US, a green toad and spring flowers are among this week's pick of images from the natural world

<https://www.theguardian.com/.../10/the-week-in-wildlife-in-pictures> - [cache](#) - Yippy News

[Jaguar Sedans, SUVs & Sports Cars - Official Site | Jaguar USA](#) [new window](#) [preview](#)

The official home of **Jaguar** USA. Explore our luxury sedans, SUVs and sports cars. Build Yours, Schedule a Test Drive or Find a Dealer Near You.

[www.jaguarusa.com/index.html](http://www.jaguarusa.com/index.html) - - Yippy Index V

[Jaguar - Wikipedia](#) [new window](#) [preview](#)

The **jaguar** (*Panthera onca*) is a big cat, a feline in the Panthera genus, and is the only extant Panthera species native to the Americas. The **jaguar** is the ...

<https://en.wikipedia.org/wiki/Jaguar> - - Yippy Index V

[Jacksonville Jaguars, Official Site of the Jacksonville ...](#) [new window](#) [preview](#)

The official team site with scores, news items, game schedule, and roster.

[www.jaguars.com](http://www.jaguars.com) - - Yippy Index V

[Jaguar Reviews - Jaguar Cars | Edmunds](#) [new window](#) [preview](#)

**Jaguar** cars: research **Jaguar** cars, read **Jaguar** reviews, find **Jaguar** car listings and get **Jaguar** pricing & dealer quotes.

<https://www.edmunds.com/jaguar> - - Yippy Index V

[Jaguar Cars, Convertible, Coupe, Sedan, SUV/Crossover ...](#) [new window](#) [preview](#)

View Motor Trend's **Jaguar** car lineup and research **Jaguar** prices, specs, fuel economy and photos. Select a **Jaguar** model and conveniently compare local dealer pricing.

[www.motortrend.com/cars/jaguar](http://www.motortrend.com/cars/jaguar) - - Yippy Index V

[2176423 Ontario Ltd. Announces Investment in Jaguar Mining Inc.](#) [new window](#) [preview](#)

**Date:** June 15, 2017 18:54 ET

# ENABLE



THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL



# Clustering Applications

## collection clustering

Google



News

U.S. edition

Modern

### Top Stories

Donald Trump  
U.S. Open  
Cristiano Ronaldo  
Robert Mueller  
Brexit  
Emmanuel Macron  
Finsbury Park, London  
James Comey  
Cladding  
Grenfell Tower  
Chapel Hill, North C...

Suggested for you

World

U.S.

Business

Technology

Entertainment

Sports

Science

Health

### Top Stories



CNN

See realtime coverage

## London mosque attack suspect named, according to media outlets

CNN - 1 hour ago

London (CNN) The driver of the van that plowed into pedestrians near a mosque in north London has been identified as Darren Osborne, 47, a resident of Cardiff in Wales, according to multiple UK media outlets.

[London Attack Near Mosque Investigated as Terrorism](#) New York Times

[Van rams Muslim worshippers in London, PM May condemns 'sickening' attack](#) Reuters

Most Referenced: [MCB on Twitter: "BREAKING: We have been informed that a van has run over worshippers as they left #FinsburyPark ..."](#) Twitter

Featured: [London's Summer of Discontent](#) CityLab

Opinion: [Why won't Donald Trump rush to tweet criticism of attacks against Muslims?](#) Chicago Tribune

Wikipedia: [Finsbury Park Mosque](#)

Related

[Finsbury Park, London »](#)  
[United Kingdom »](#)  
[London »](#)



Today.com



Aljazeera.com



Fox News



Chicago Tri...



Los Angele...



Washingto...



Vox



Fox News



Newsweek



Washington...

## Senate leaders plan to rush a health-care bill to a vote, and there's nothing Democrats can do about it

Washington Post - 3 hours ago

When the Republican-led Senate Rules Committee briefly flirted with the idea of restricting television interviews in the hallways of the Capitol last week, it became only the most obvious manifestation of how the party's leaders were handling the ...



Washington...

## Democrats just got some very good news from the Supreme Court on gerrymandering

Washington Post - 4 hours ago

The Supreme Court just made a major decision without actually issuing a decision. On Monday morning, the justices announced that they would take up a case out of Wisconsin that could result in a ruling on the constitutionality of partisan gerrymandering.



# Clustering

objective

- Grouping documents or instances into subsets or clusters
- Documents within a the same cluster should be similar
- Documents from different clusters should be dissimilar



# Clustering

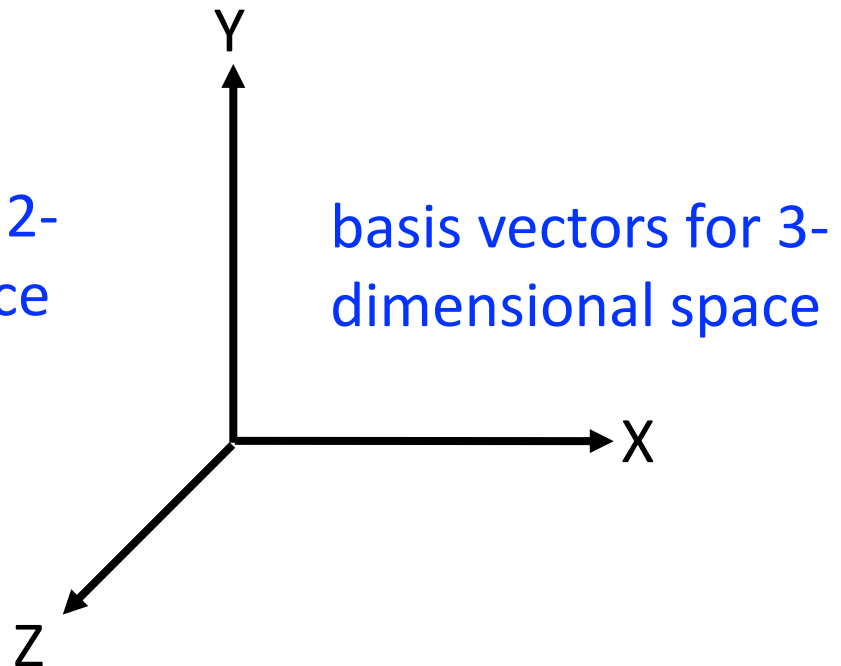
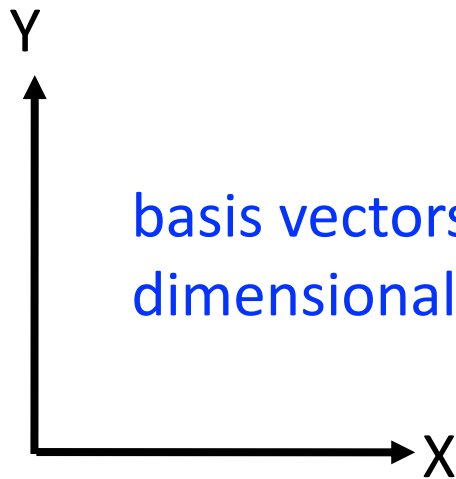
## basics

- What does it mean for documents to be similar or dissimilar?
- We need a computational way of modeling similarity
- **One solution:** model similarity using distance in a vector space representation of the collection or dataset
  - small distance = high similarity
  - long distance = low similarity

# Vector Space Representation

## review

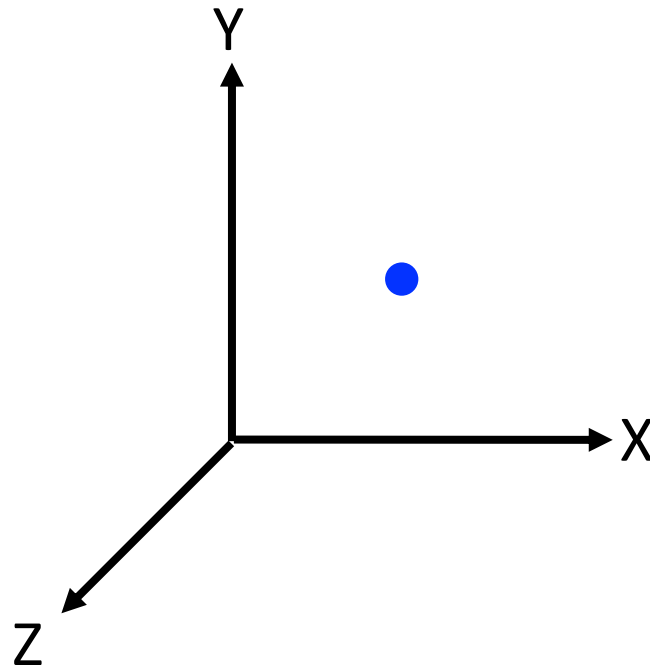
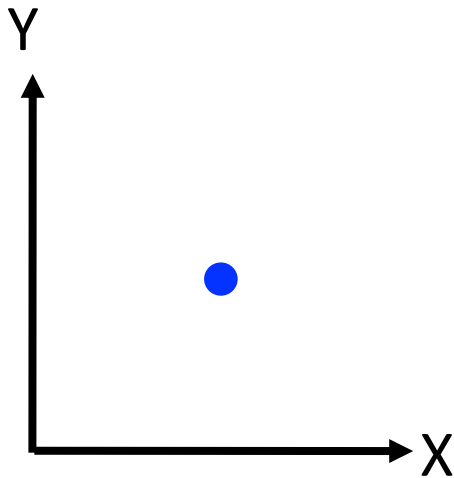
- A **vector space** is defined by a set of linearly independent basis vectors
- The **basis vectors** correspond to the dimensions or directions of the vector space



# Vector Space Representation

review

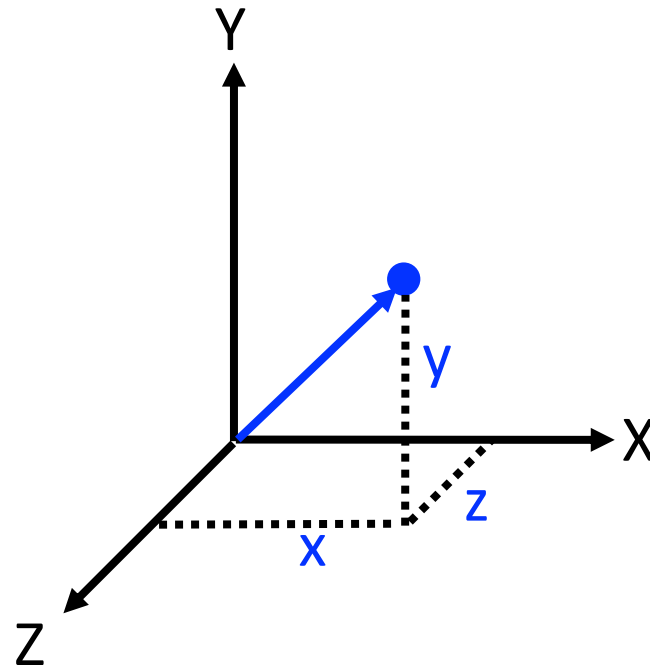
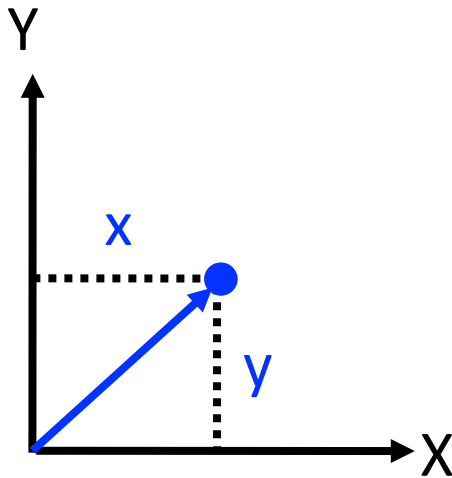
- A **vector** is a point in a vector space



# Vector Space Representation

## review

- A 2-dimensional vector can be written as  $[x,y]$
- A 3-dimensional vector can be written as  $[x,y,z]$





# Vector Space Representation

review

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1

# Vector Space Representation

review

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1

- We can represent this document as a vector in a 10-dimensional vector space

# Vector Space Representation

review

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1

- This representation assumes binary term-weights.
- Are there other term-weighting schemes?

# Vector Space Representation

review

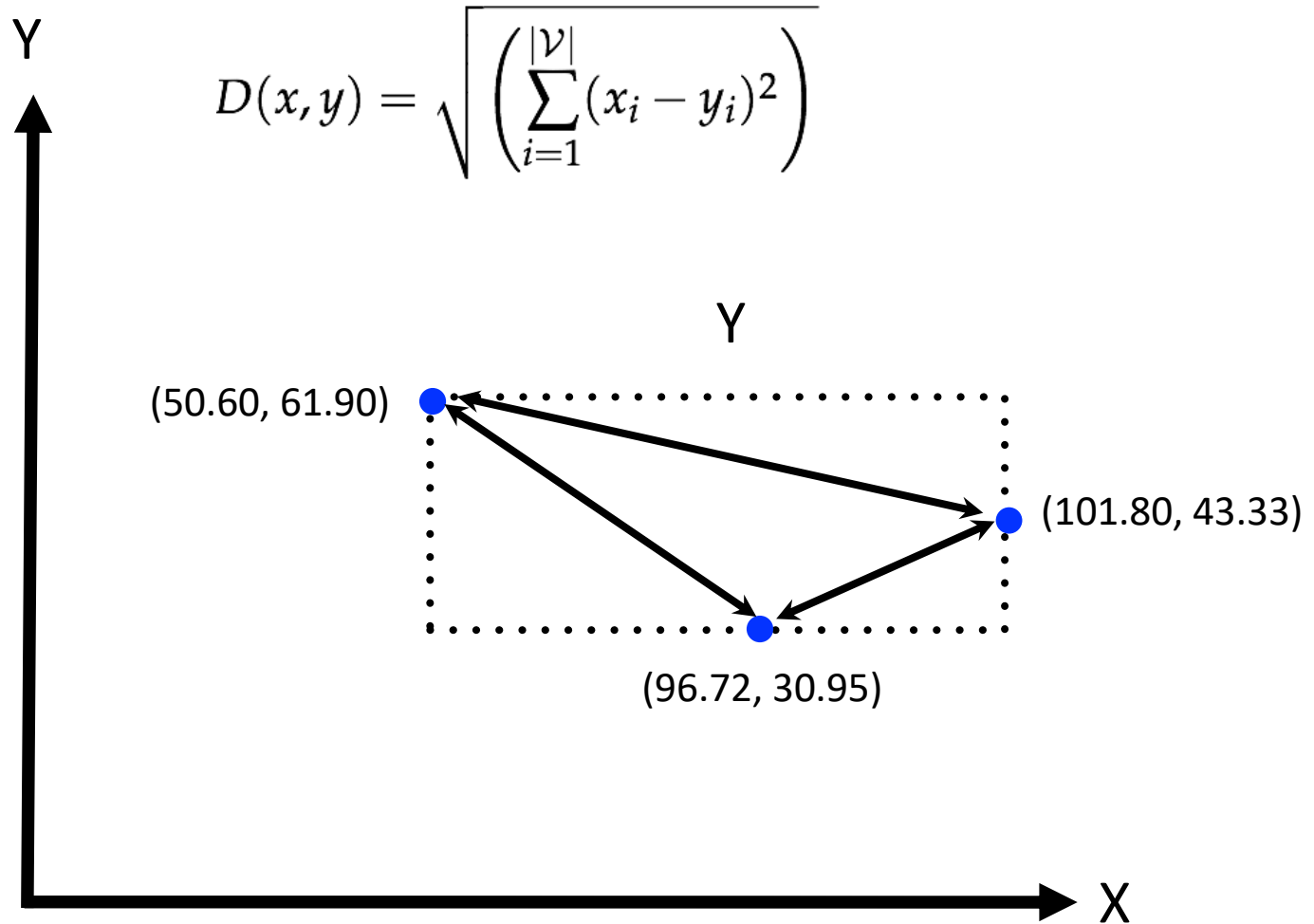
- Similarity = Euclidean Distance:

$$D(x, y) = \sqrt{\left( \sum_{i=1}^{|\mathcal{V}|} (x_i - y_i)^2 \right)}$$



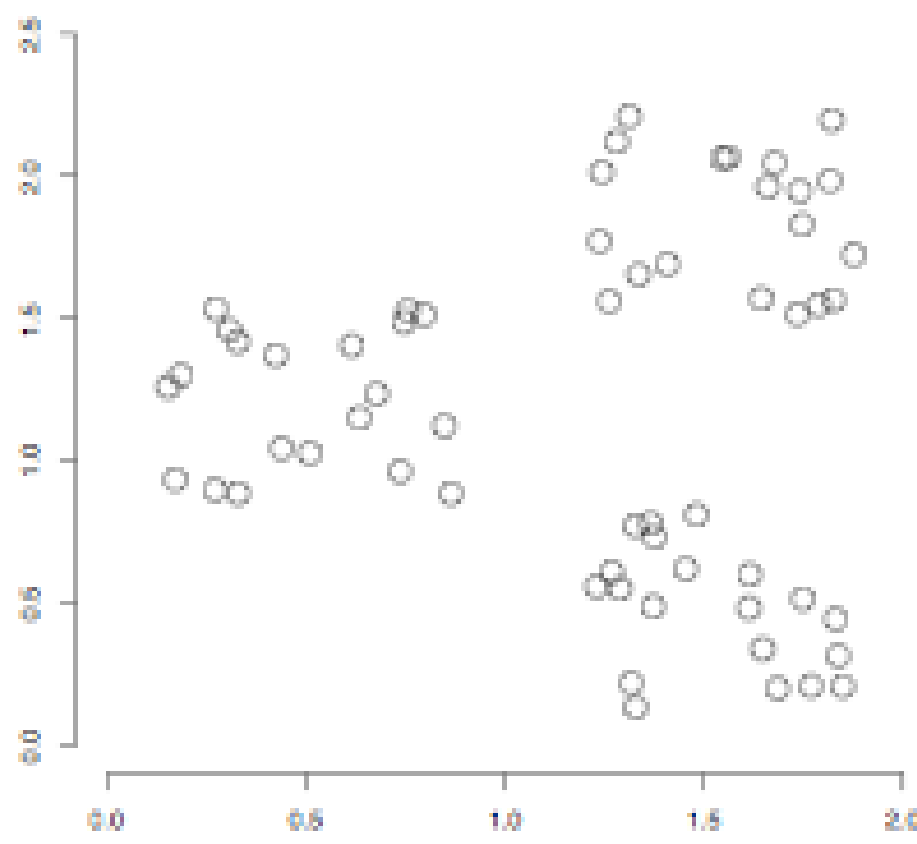
# Vector Space Representation

review



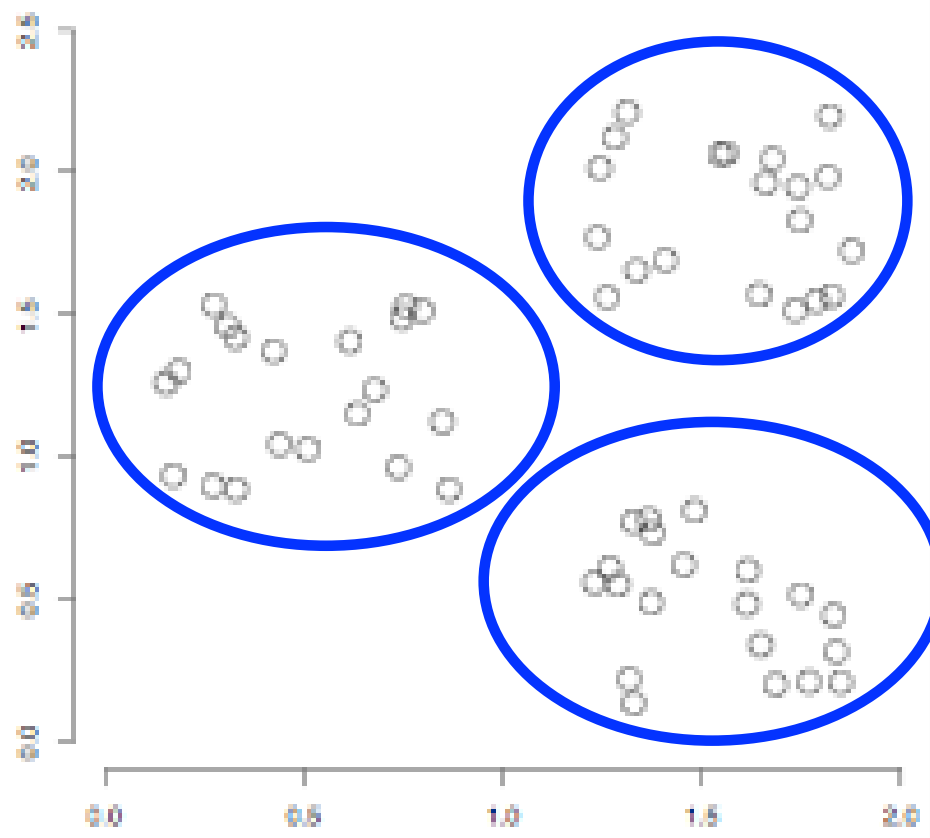
# Clustering

- What would we expect a clustering algorithm to do with this dataset?



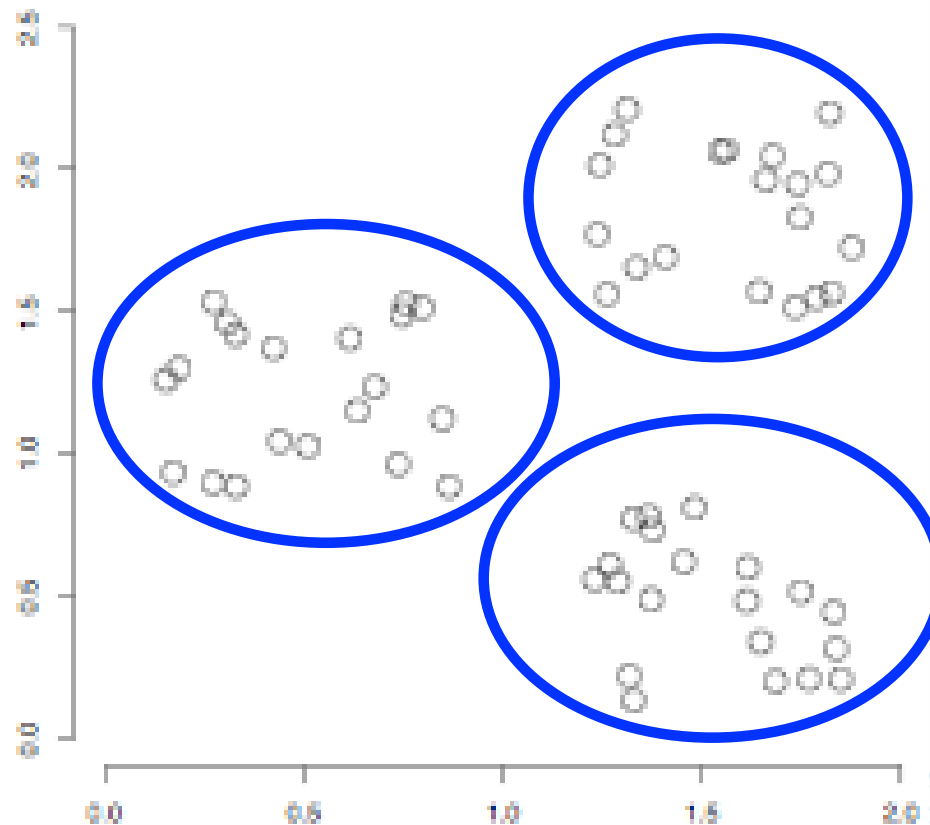
# Clustering

- What would we expect a clustering algorithm to do with this dataset?



# Clustering

- Propose an algorithm that might be able to do this!



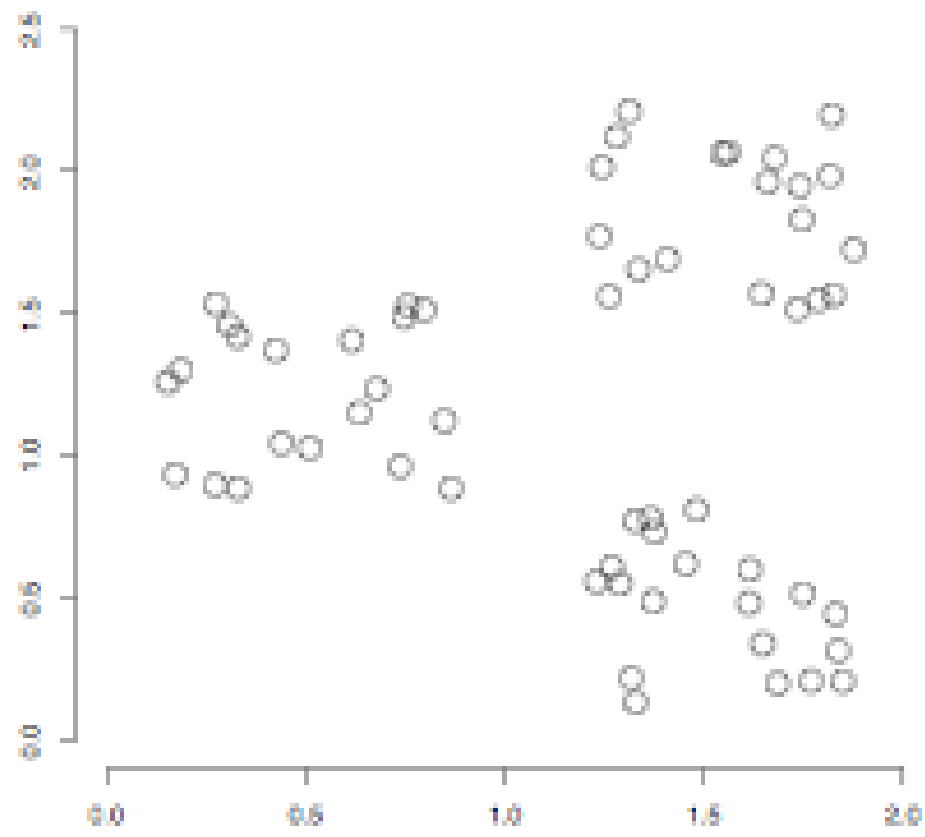


# Clustering

- **Input:** number of desired clusters  $K$
- **Output:** assignment of documents to  $K$  clusters
- **Algorithm:**
  - ▶ randomly select  $K$  documents (seeds)
  - ▶ assign each remaining document to its nearest seed
  - ▶ and so on.

# Clustering

- Could this work?



# K-Means Clustering

ENABLE



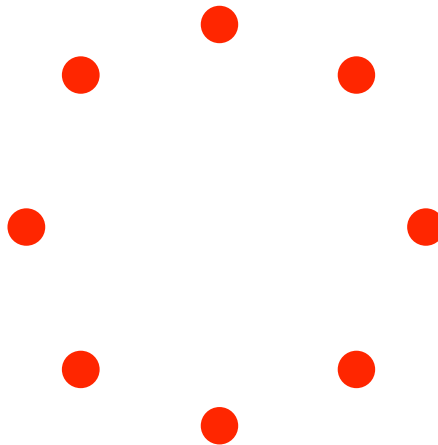
THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL



# K-means Clustering

## cluster centroid

- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”

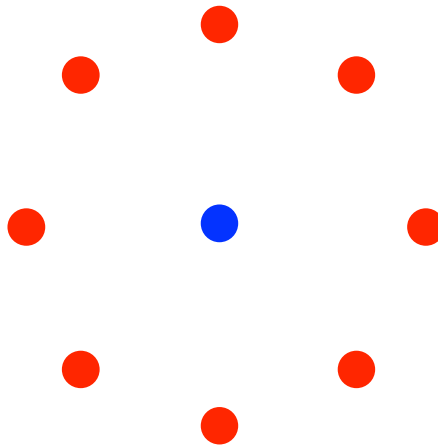




# K-means Clustering

## cluster centroid

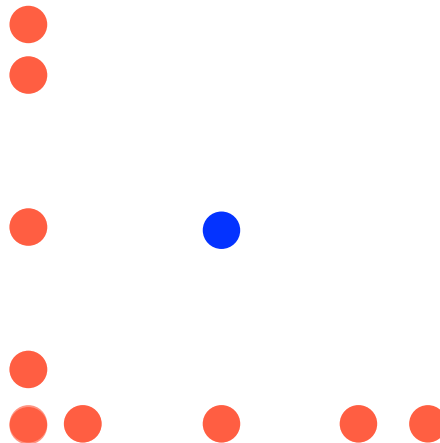
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

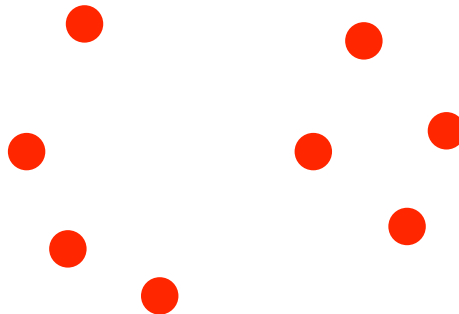
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

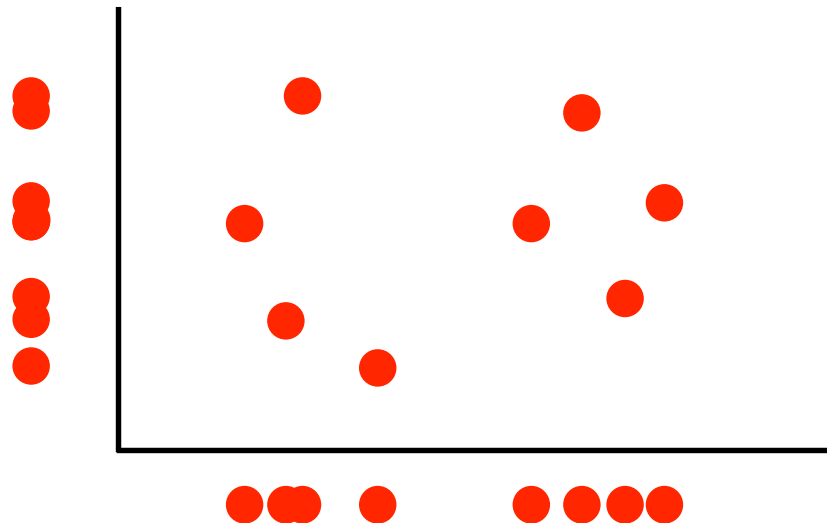
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

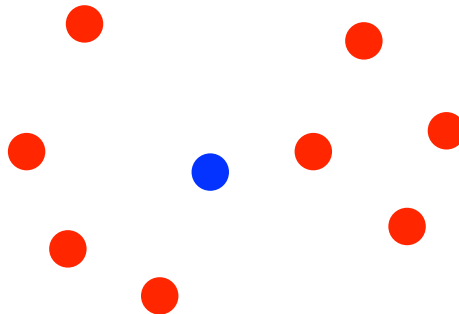
- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



# K-means Clustering

## cluster centroid

- The key to understanding K-means clustering is to understand the idea of a **cluster centroid**
- Given a cluster, you can think of its centroid as a point (or vector) that corresponds to its “center of mass”



## cluster centroid

cluster 1  
centroid

[illegible]

# K-means Clustering

cluster centroid

docs  
assigned to  
cluster 1

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0	0
0	1	0	1	1	0	1	0	0	0
0	0	1	0	1	1	0	1	1	1
0	0	1	0	1	1	0	1	1	1
1	1	0	1	1	0	0	1	0	1

cluster 1  
centroid  
(average!)

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10
0.33	0.5	0.5	0.5	1	0.33	0.33	0.83	0.5	0.5

# K-means Clustering

## cluster centroid

- For each dimension  $i$ , set:

$$c_i = \frac{1}{|C|} \sum_{d \in C} d_i$$

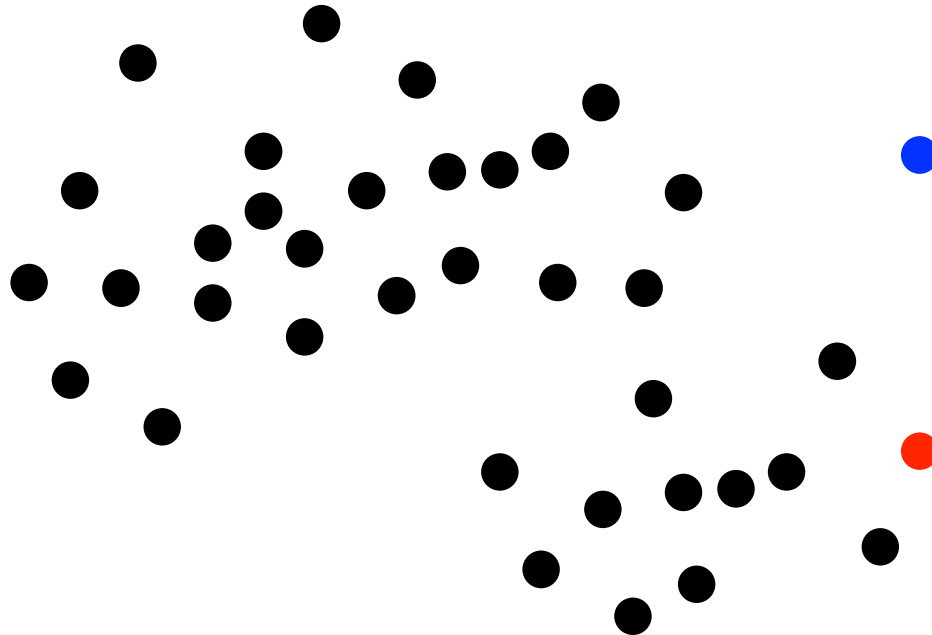


# K-means Clustering

- **Input:** number of desired clusters  $K$
- **Output:** assignment of documents to  $K$  clusters
- **Algorithm:**
  - ▶ **Step 1:** randomly select  $K$  documents (seeds)
  - ▶ **Step 2:** assign each document to its nearest seed
  - ▶ **Step 3:** compute all  $K$  cluster centroids
  - ▶ **Step 4:** re-assign each document to its nearest centroid
  - ▶ **Step 5:** re-compute all  $K$  cluster centroids
  - ▶ **Step 6:** repeat steps 4 and 5 until terminating condition

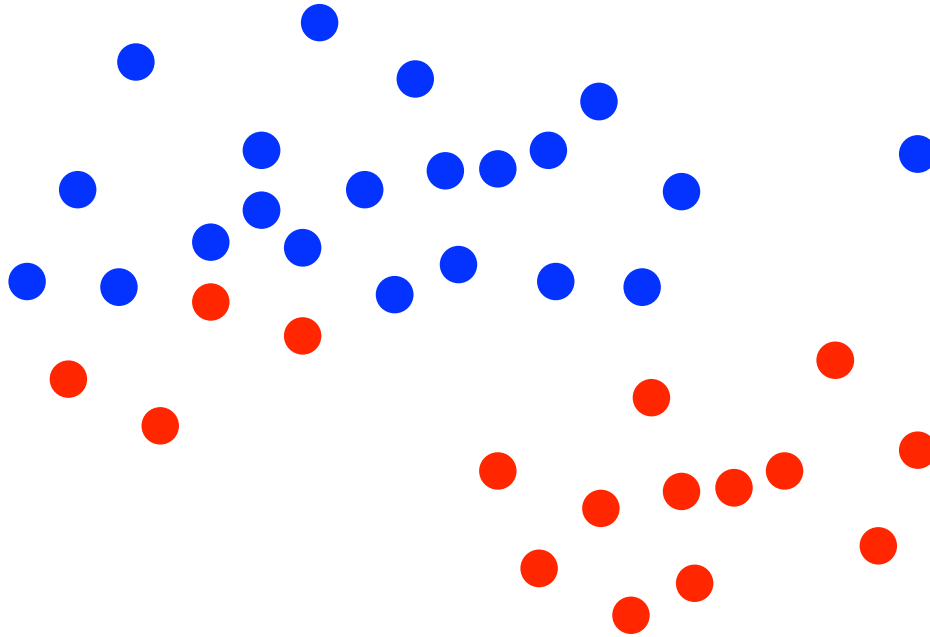
# K-means Clustering

- **Step 1:** randomly select K documents (seeds)



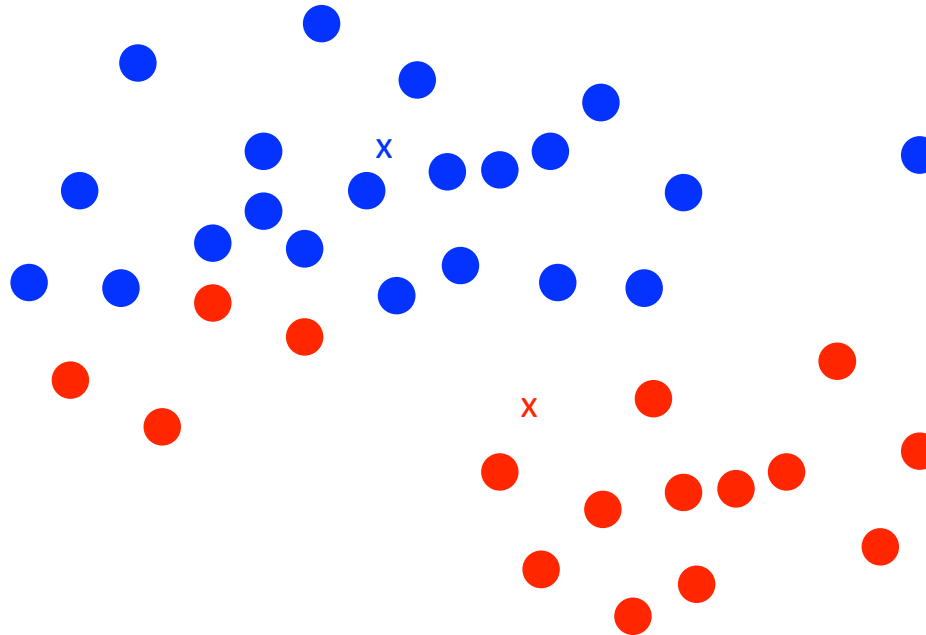
# K-means Clustering

- Step 2: assign each document to its nearest seed



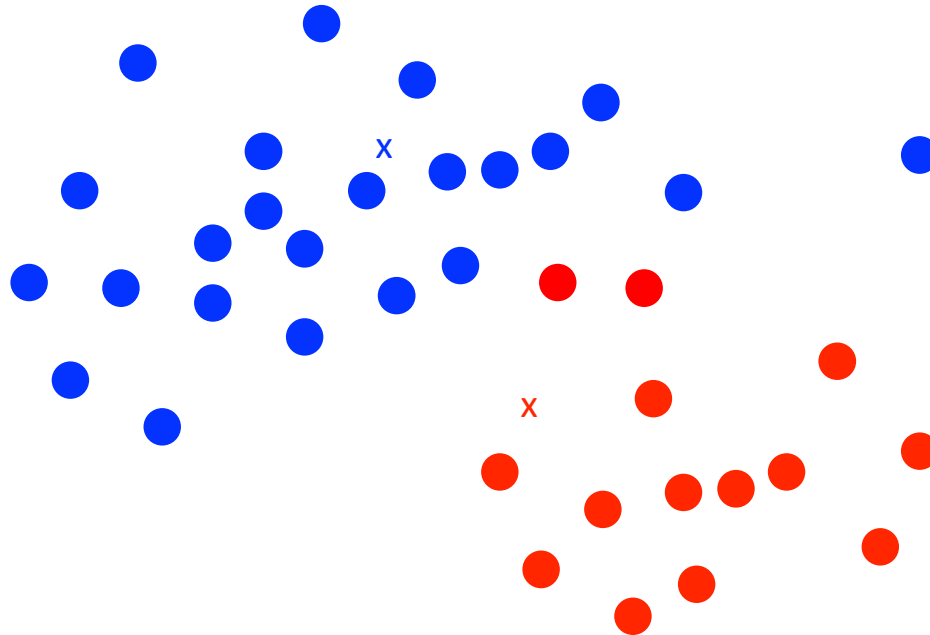
# K-means Clustering

- Step 3: compute all K cluster centroids



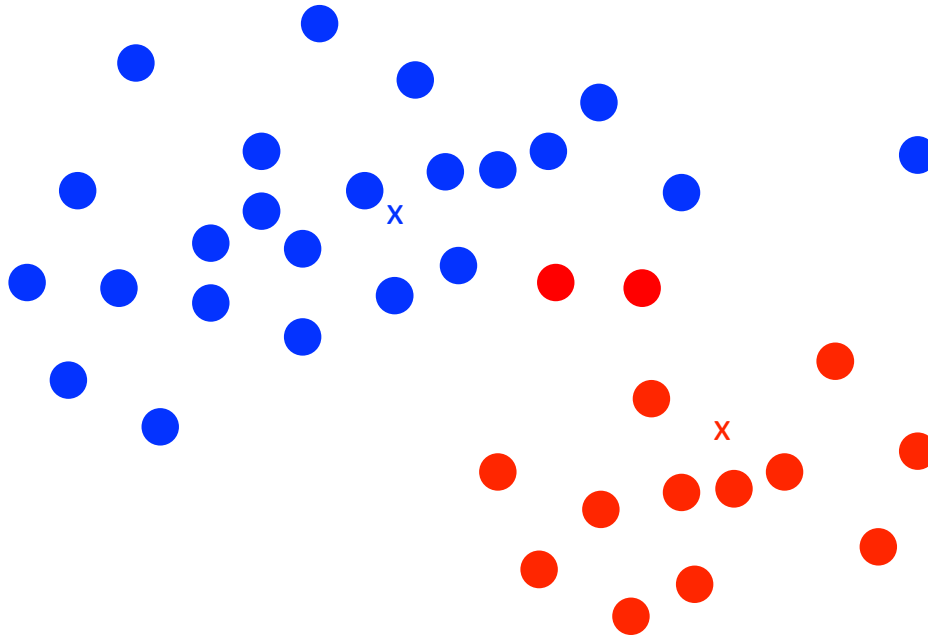
# K-means Clustering

- Step 4: re-assign each document to its nearest centroid



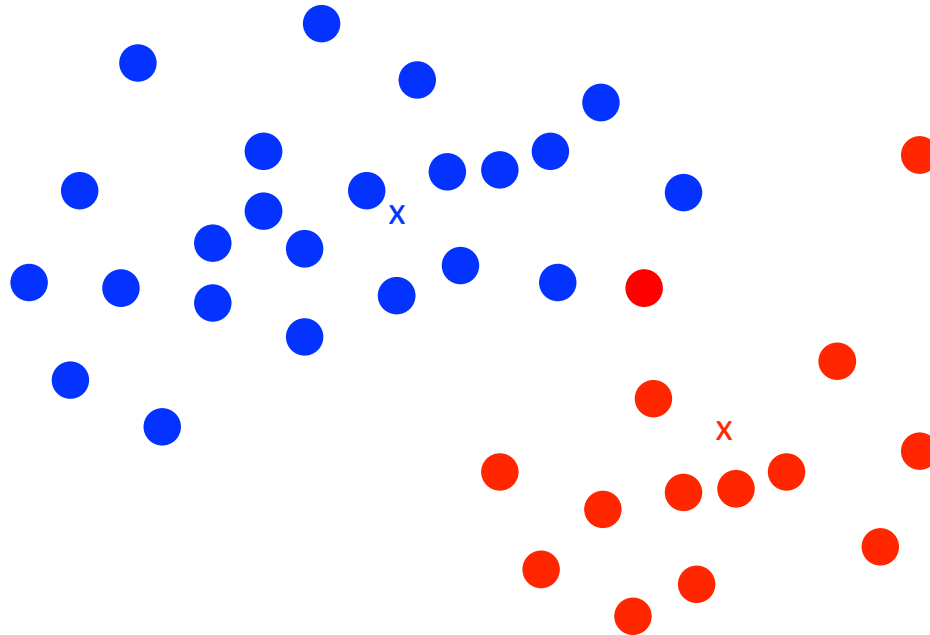
# K-means Clustering

- Step 4: re-compute all K cluster centroids



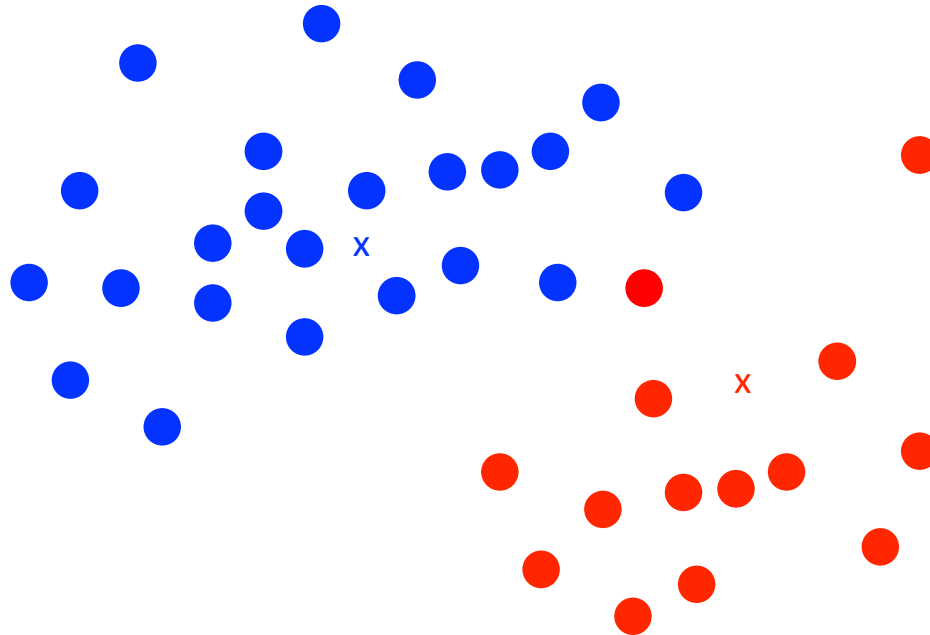
# K-means Clustering

- **Step 5:** re-assign each document to its nearest centroid



# K-means Clustering

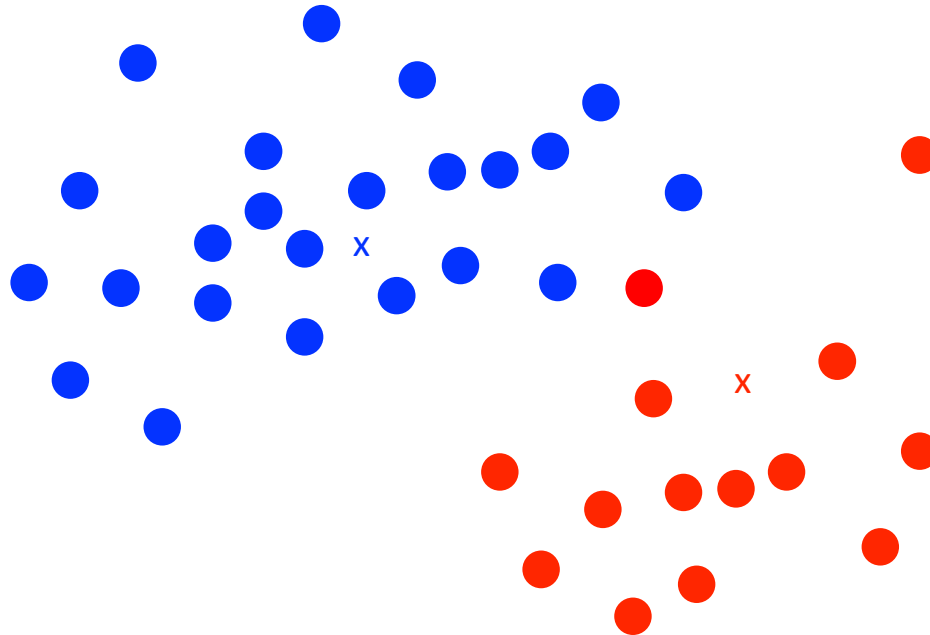
- Step 4: re-compute all K cluster centroids





# K-means Clustering

- **Step 5:** re-assign each document to its nearest centroid



# K-means Clustering

- **Input:** number of desired clusters  $K$
- **Output:** assignment of documents to  $K$  clusters
- **Algorithm:**
  - ▶ **Step 1:** randomly select  $K$  documents (seeds)
  - ▶ **Step 2:** assign each document to its nearest seed
  - ▶ **Step 3:** compute all  $K$  cluster centroids
  - ▶ **Step 4:** re-assign each document to its nearest centroid
  - ▶ **Step 5:** re-compute all  $K$  cluster centroids
  - ▶ **Step 6:** repeat steps 4 and 5 until terminating condition

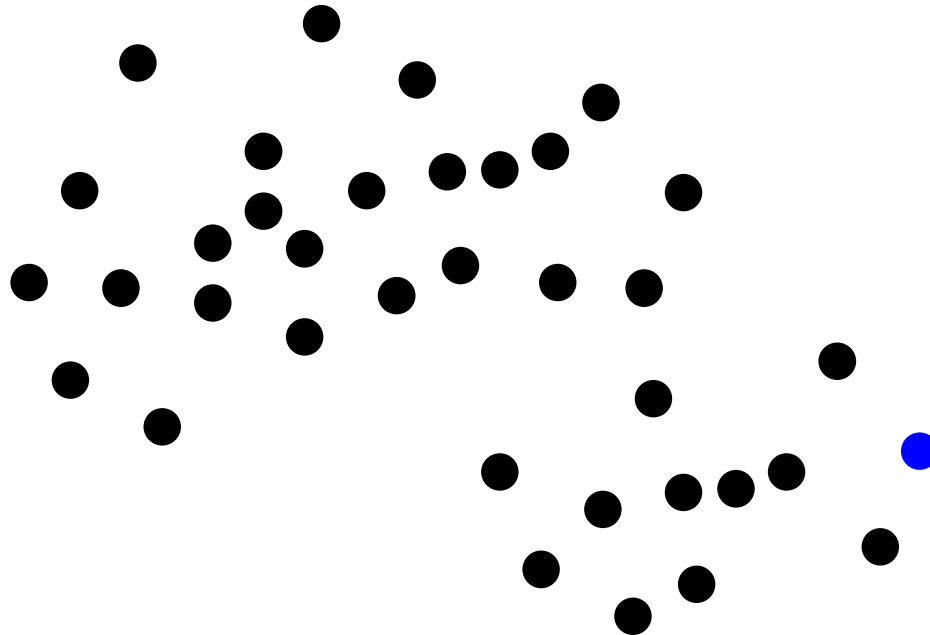
# K-means Clustering

## potential drawback

- The quality of the output clustering depends on the choice of  $K$  and on the **initial seeds**
- In many cases, the choice of  $K$  is pre-determined by the application
  - ▶ **Search engine results clustering:** grouping search engine results by topic
  - ▶ **Collection clustering:** grouping documents by topic to support navigation and exploration
- Later we'll see ways of setting  $K$  dynamically

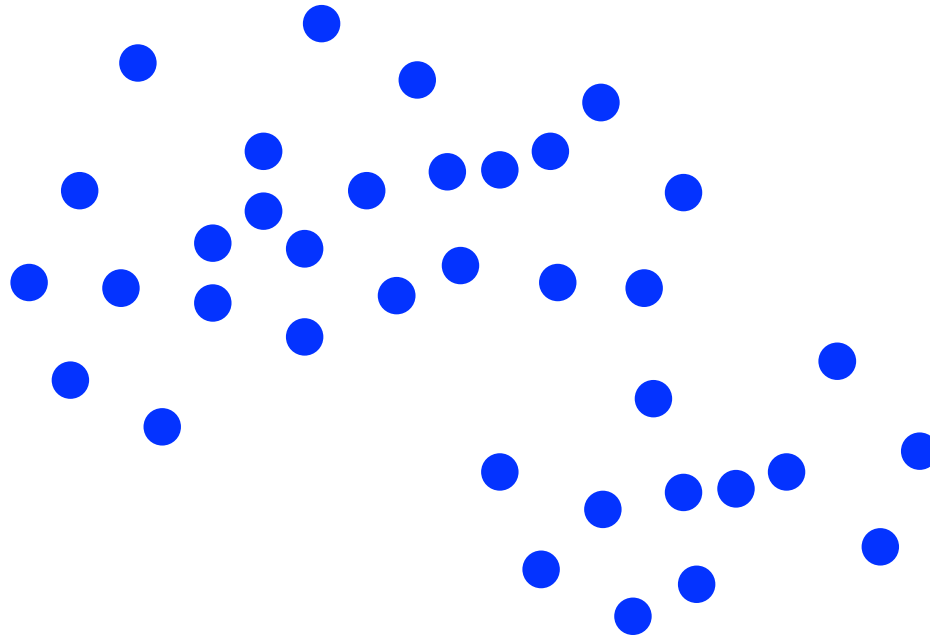
# K-means Clustering

bad seeds?



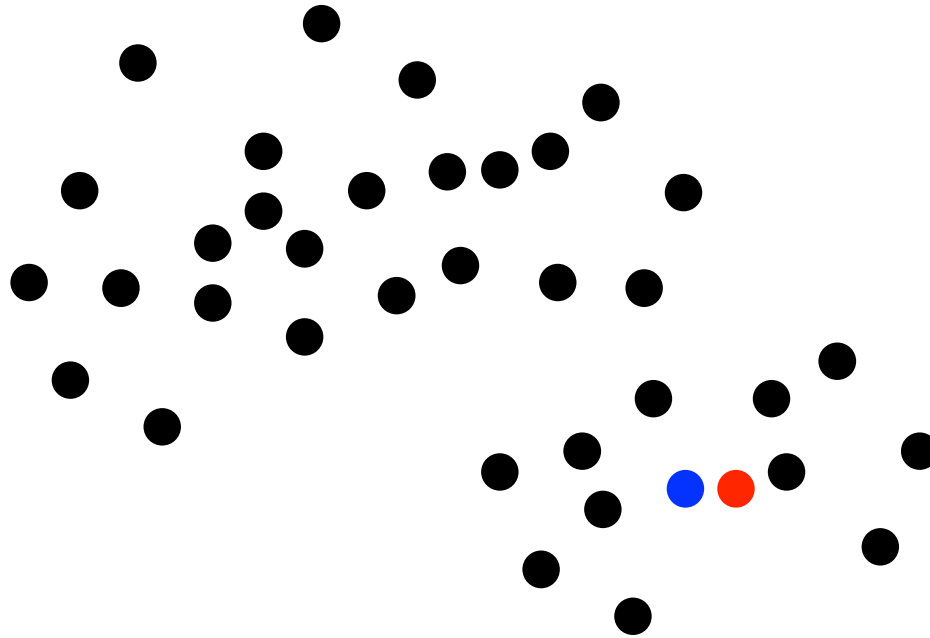
# K-means Clustering

bad seeds?



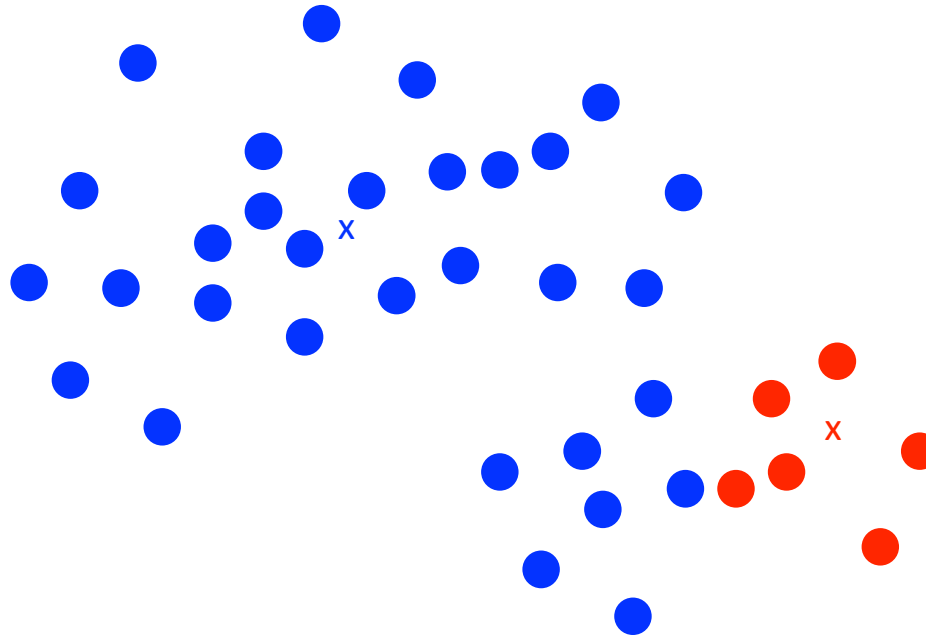
# K-means Clustering

bad seeds?



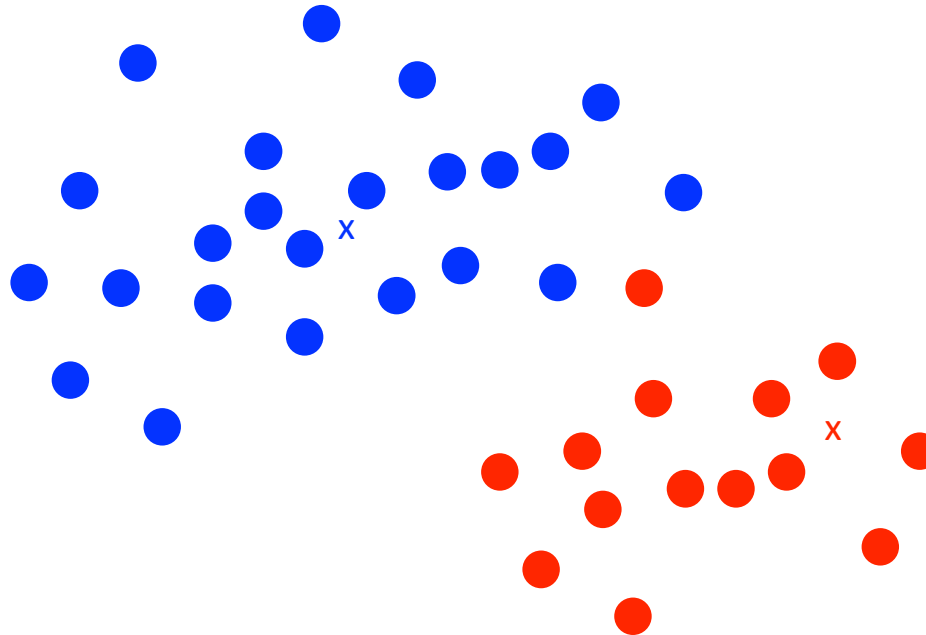
# K-means Clustering

bad seeds?



# K-means Clustering

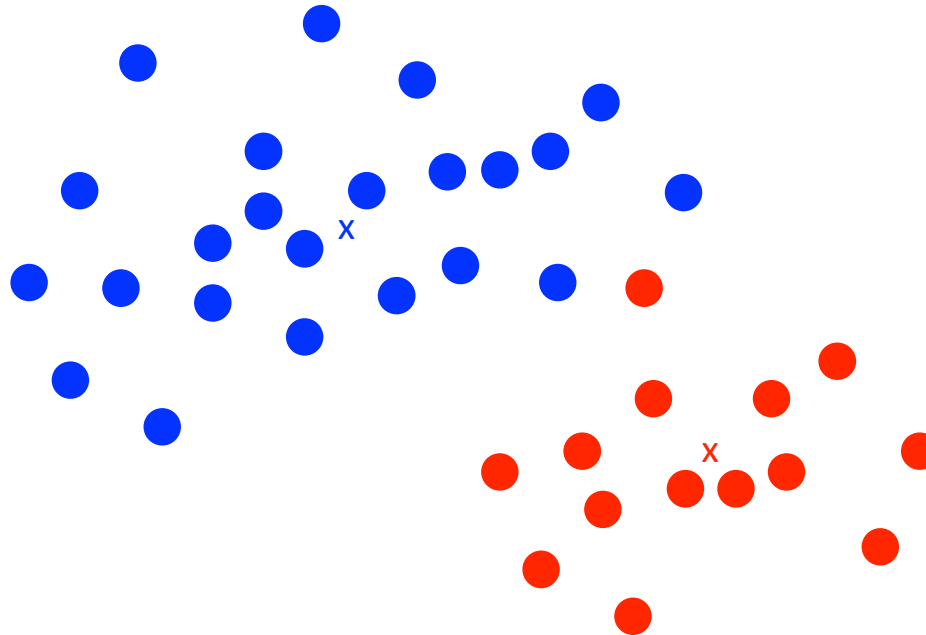
bad seeds?





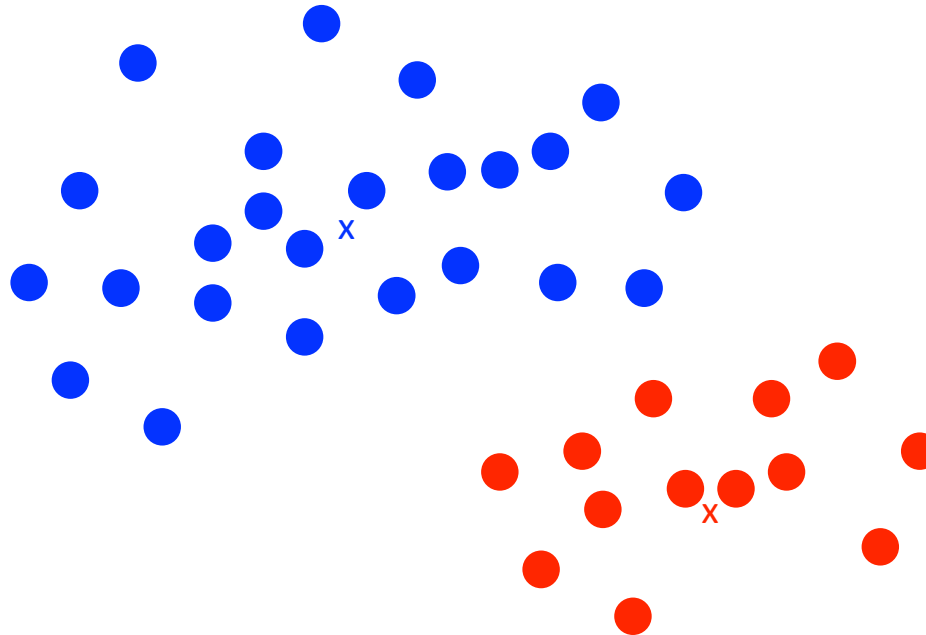
# K-means Clustering

bad seeds?



# K-means Clustering

bad seeds?



# K-means Clustering

bad seeds

- It's difficult to know which seeds will yield a high-quality clustering
- However, it's usually a good idea to avoid seeds that are outliers
- How would you detect outliers?

# K-means Clustering

## clustering evaluation

- What does it mean for a clustering to be high quality anyway?
- What is the goal of clustering again?

# K-means Clustering

## internal evaluation

- In theory, a good clustering should have:
  - ▶ Similar documents in the same clusters
  - ▶ Different documents in different clusters

# K-means Clustering

internal evaluation

$$\text{Clustering Quality} = \left( \text{Average distance between all pairs of documents in different clusters} \right) - \left( \text{Average distance between all pairs of documents in the same cluster} \right)$$

Inter-cluster distance      Intra-cluster distance

# K-means Clustering

## improved k-means

- Given a set of documents and a value  $K$ , run K-means clustering  $N$  times and keep the clustering that produces the greatest difference between the inter-cluster distance and the intra-cluster distance

# Bottom-up Agglomerative Clustering

ENABLE



THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL





# Bottom-up Clustering

- While K-means requires setting **K**, bottom-up clustering groups the data in a hierarchical fashion
- We can then set **K** after the clustering is done or use a distance threshold to set **K** dynamically (more on this later)

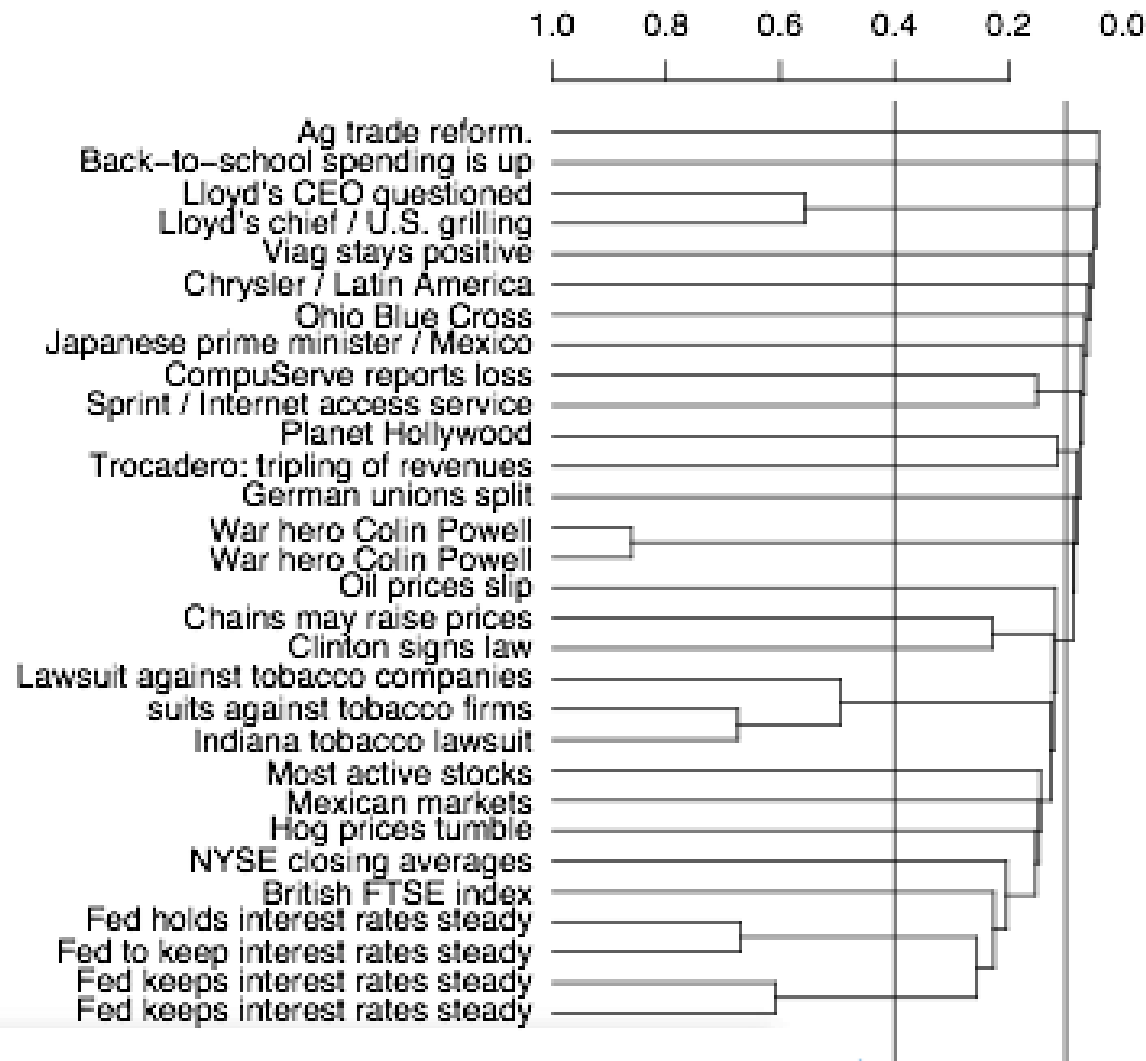
# Bottom-up Clustering

- Input: data
- Output: cluster hierarchy
- Algorithm:
  - ▶ Step 1: consider every document its own cluster
  - ▶ Step 2: compute the distance between all cluster pairs
  - ▶ Step 3: merge/combine the nearest two clusters into one
  - ▶ Step 4: repeat steps 2 and 3 until every document is in one cluster

# Bottom-up Clustering

- **Input:** data
- **Output:** cluster hierarchy
- **Algorithm:**
  - ▶ **Step 1:** consider every document its own cluster
  - ▶ **Step 2:** compute the distance between all cluster pairs
  - ▶ **Step 3:** merge/combine the nearest two clusters into one
  - ▶ **Step 4:** repeat steps 2 and 3 until every document is in one cluster

# Bottom-up Clustering



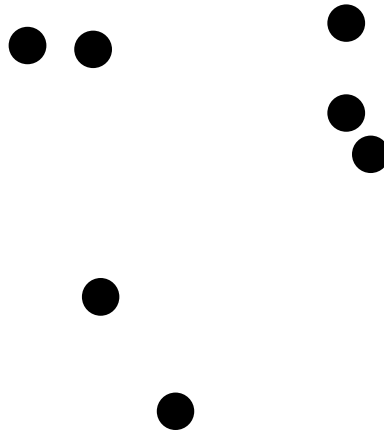
# Bottom-up Clustering

- Computing the distance between two clusters
- **Single-Link**: the distance between the two nearest documents
- **Complete-Link**: the distance between the two documents that are farthest apart
- **Average-Link**: the average distance between all document pairs in the two different clusters
  - ▶ this is equivalent to using the distance between the two cluster centroids

# Bottom-up Clustering

## single-link

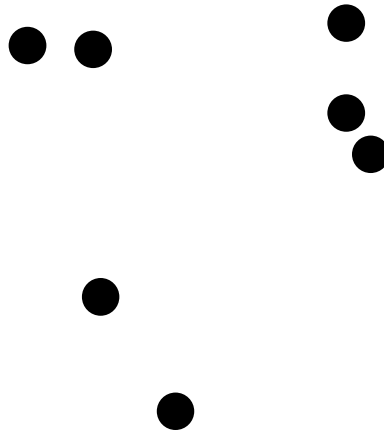
- **Step 1:** consider each document its own cluster



# Bottom-up Clustering

single-link

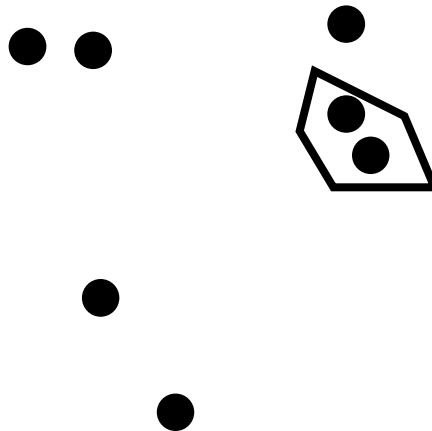
- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one



# Bottom-up Clustering

## single-link

- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one

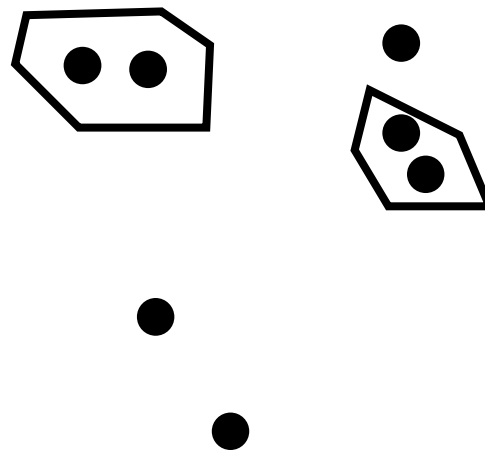




# Bottom-up Clustering

single-link

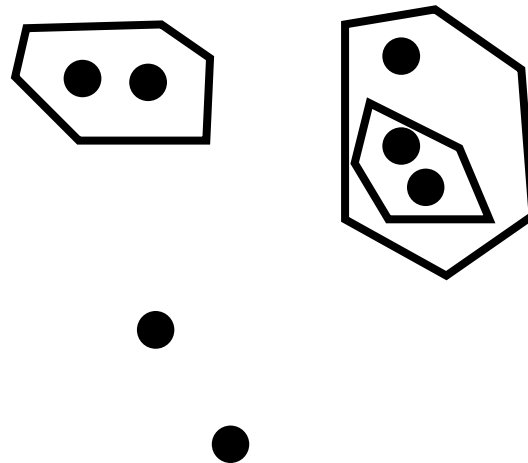
- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one



# Bottom-up Clustering

single-link

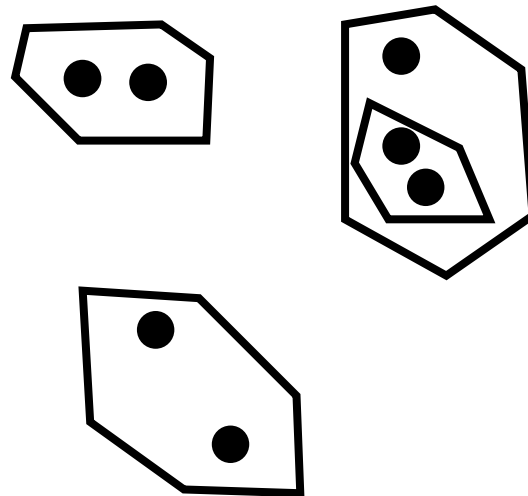
- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one



# Bottom-up Clustering

## single-link

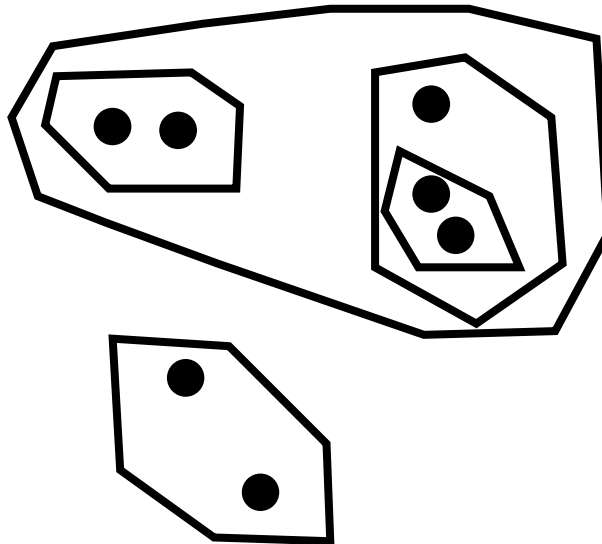
- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one



# Bottom-up Clustering

single-link

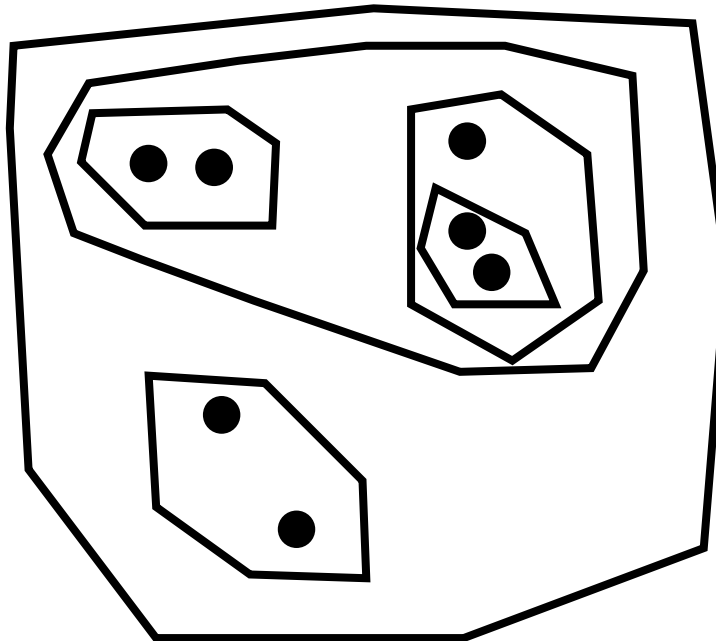
- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one



# Bottom-up Clustering

single-link

- **Step 2:** compute the distance between all cluster pairs
- **Step 3:** merge/combine the nearest two clusters into one



# Bottom-up Clustering

- Setting  $K$  dynamically
- Instead of setting  $K$ , we could set a distance threshold  $T$
- Stop merging/combining clusters when the distance between the two nearest clusters  $> T$
- Using a distance threshold can help prevent “concept drift” (especially with single-link clustering)
  - ▶ text mining --> HiDAV --> unc --> basketball

# Labeling Clusters

ENABLE

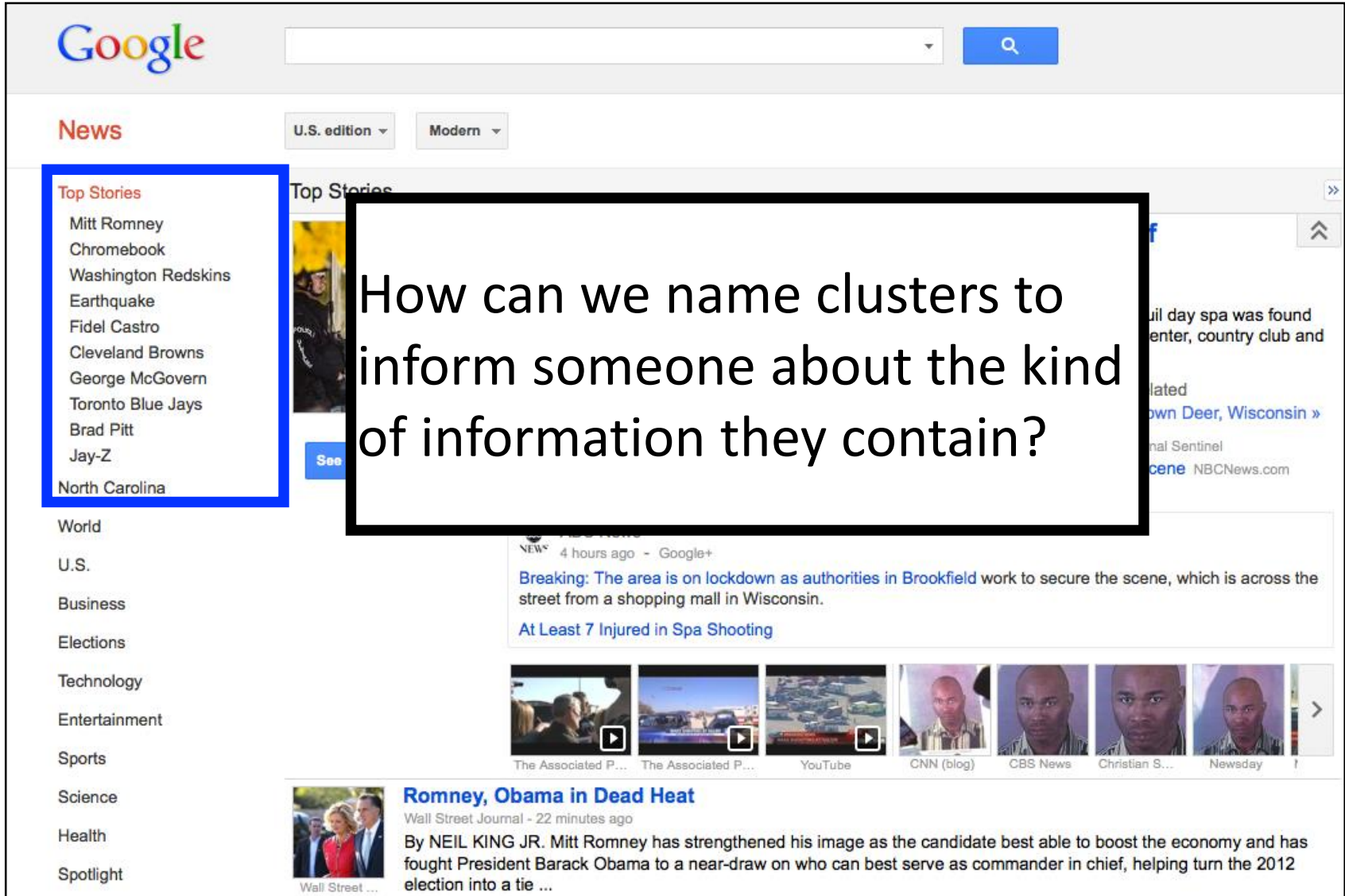


THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL



# Clustering Applications

## collection clustering



The image is a screenshot of the Google News homepage. At the top is the Google logo and a search bar. Below the logo, there are tabs for "News", "U.S. edition", and "Modern". A blue box highlights the "Top Stories" section on the left, which lists various news items: Mitt Romney, Chromebook, Washington Redskins, Earthquake, Fidel Castro, Cleveland Browns, George McGovern, Toronto Blue Jays, Brad Pitt, Jay-Z, and North Carolina. A large black text box is overlaid on the right side of the page, containing the text: "How can we name clusters to inform someone about the kind of information they contain?". Below the text box, there is a news article titled "Breaking: The area is on lockdown as authorities in Brookfield work to secure the scene, which is across the street from a shopping mall in Wisconsin." with a sub-headline "At Least 7 Injured in Spa Shooting". Below the article, there is a row of video thumbnails from various sources: The Associated P..., The Associated P..., YouTube, CNN (blog), CBS News, Christian S..., and Newsday. At the bottom, there is a section titled "Romney, Obama in Dead Heat" from the Wall Street Journal, dated 22 minutes ago, with a byline "By NEIL KING JR." and a brief summary of the article.

Google

News

U.S. edition Modern

Top Stories

Mitt Romney  
Chromebook  
Washington Redskins  
Earthquake  
Fidel Castro  
Cleveland Browns  
George McGovern  
Toronto Blue Jays  
Brad Pitt  
Jay-Z  
North Carolina

World  
U.S.  
Business  
Elections  
Technology  
Entertainment  
Sports  
Science  
Health  
Spotlight

How can we name clusters to inform someone about the kind of information they contain?

Breaking: The area is on lockdown as authorities in Brookfield work to secure the scene, which is across the street from a shopping mall in Wisconsin.  
At Least 7 Injured in Spa Shooting

Romney, Obama in Dead Heat  
Wall Street Journal - 22 minutes ago  
By NEIL KING JR. Mitt Romney has strengthened his image as the candidate best able to boost the economy and has fought President Barack Obama to a near-draw on who can best serve as commander in chief, helping turn the 2012 election into a tie ...



# Labeling Clusters

## A simple solution

- Construct a vocabulary of terms and/or phrases (n-grams) that are frequent in the data
- Assign each cluster the term(s) or phrase(s) with the highest mutual information

# Mutual Information

$$\text{MI}(w, c) = \log \left( \frac{P(w, c)}{P(w)P(c)} \right)$$

- **P(w,c)**: the probability that a document contains word **w** and belongs to cluster **c**
- **P(w)**: the probability that word **w** occurs in a document from any cluster
- **P(c)**: the probability that a document belongs to cluster **c**

# Mutual Information

$$\text{MI}(w, c) = \log \left( \frac{P(w, c)}{P(w)P(c)} \right)$$

- If  $P(w, c) = P(w) P(c)$ , it means that the word **w** is independent of cluster **c**
- If  $P(w, c) > P(w) P(c)$ , it means that the word **w** is not independent of of cluster **c**

# Mutual Information

- Every document falls under one of these quadrants

	belongs to cluster <b>c</b>	does not belong to cluster <b>c</b>
contains word <b>w</b>	a	b
does not contains word <b>w</b>	c	d

total # of instances  $N =$   
 $a + b + c + d$

$$P(w, c) = a / N$$

$$P(c) = (a + c) / N$$

$$P(w) = (a + b) / N$$

$$MI(w, c) = \log \left( \frac{P(w, c)}{P(w)P(c)} \right)$$

# Summary

- Clustering: grouping similar documents (or instances) into subsets
- Exploratory analysis: the goal is to discover common and uncommon properties of the data
- K-means and Agglomerative Bottom-up Clustering (there are many, many others)
- Labeling clusters

# The Future of Text Mining

[Link](#)



# Too Many Barriers? It really Matter Where to Apply and How to Start.

[Link](#)





# Challenge with Something You Really Like



ENABLE



THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL





Any Questions?

ENABLE



THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL



No More Next Class

ENABLE



THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL

