



EXPLORATORY DATA ANALYSIS: CLUSTERING

Sandeep Soni

09/18/2023

CLASS LOGISTICS

- Hw4 due; Hw5 released

FROM LAST CLASS

- PMI demo

HW4 REVIEW

- Chloe wants to find company names strongly associated with NLP and CV. She has a background corpus (comments from NLP and CV fora). How can she find company names?
- Find words from the comments that are strongly associated with individual categories and then take intersection with the company names



Search for topics, locations & sources



Home

For you

Following

News Showcase

U.S.

World

Local

Business

Technology

Entertainment

Sports

Science

Health

Technology >

MacRumors

Apple Significantly Lowers Repair Fees for iPhone 15 Pro Models With Cracked Back...



4 hours ago



USA TODAY

iOS 17 release: Apple's new operating system has a lot to offer



13 hours ago



Newsmax

FDA to Look Into iPhone 12 Radiation Concerns



2 days ago



Entertainment >

THE WALL STREET JOURNAL

Drew Barrymore, Other Talk Shows Halt Season Premieres Until Industry Strikes End



27 minutes ago



Variety

Billy Miller, 'The Young and the Restless' and 'General Hospital' Actor, Dies at 43



6 hours ago



Daily Mail

Teyana Taylor reveals split from Iman Shumpert after seven years of marriage and insists...



1 hour ago



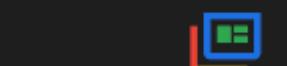
Sports >

FOX SPORTS

NFL Week 2 top plays: Falcons, Bucs win, Giants-Cardinals, 49ers-Rams live



1 hour ago



newyorkjets.com

New York Jets vs. Dallas Cowboys Game Inactives - Week 2 2023

INACTIVES	
KO	ISRAEL ADANAKANDA
WR	JASON BROWN JR.
OL	MAX MITCHELL
DL	WILL MCDONALD
LB	ZAPPI BARNES
CB	BRYCE HALL
X	GREG ZIERLERSEN

4 hours ago



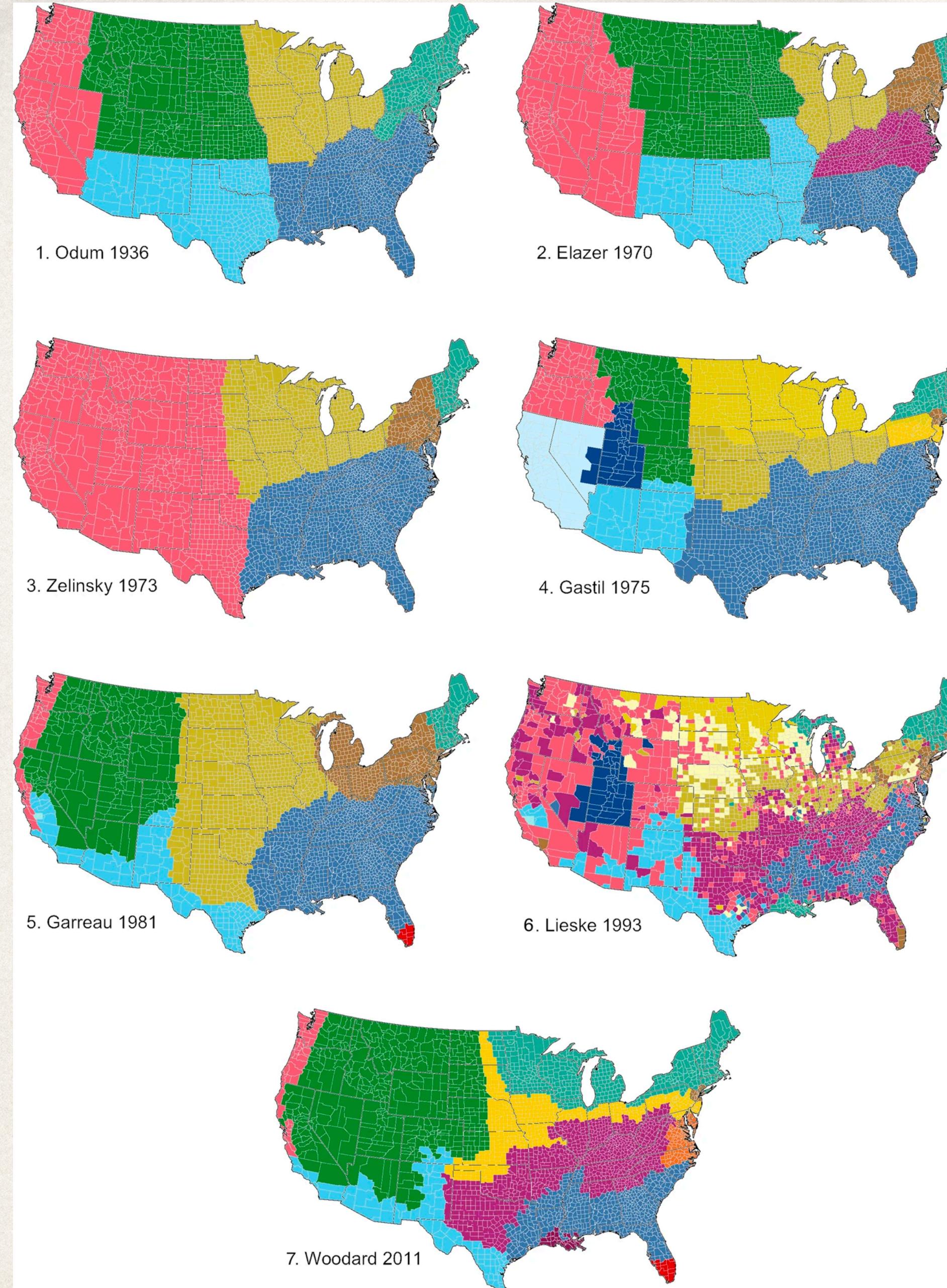
ESPN

49ers gamble with one second left, score on Brock Purdy sneak - ESPN

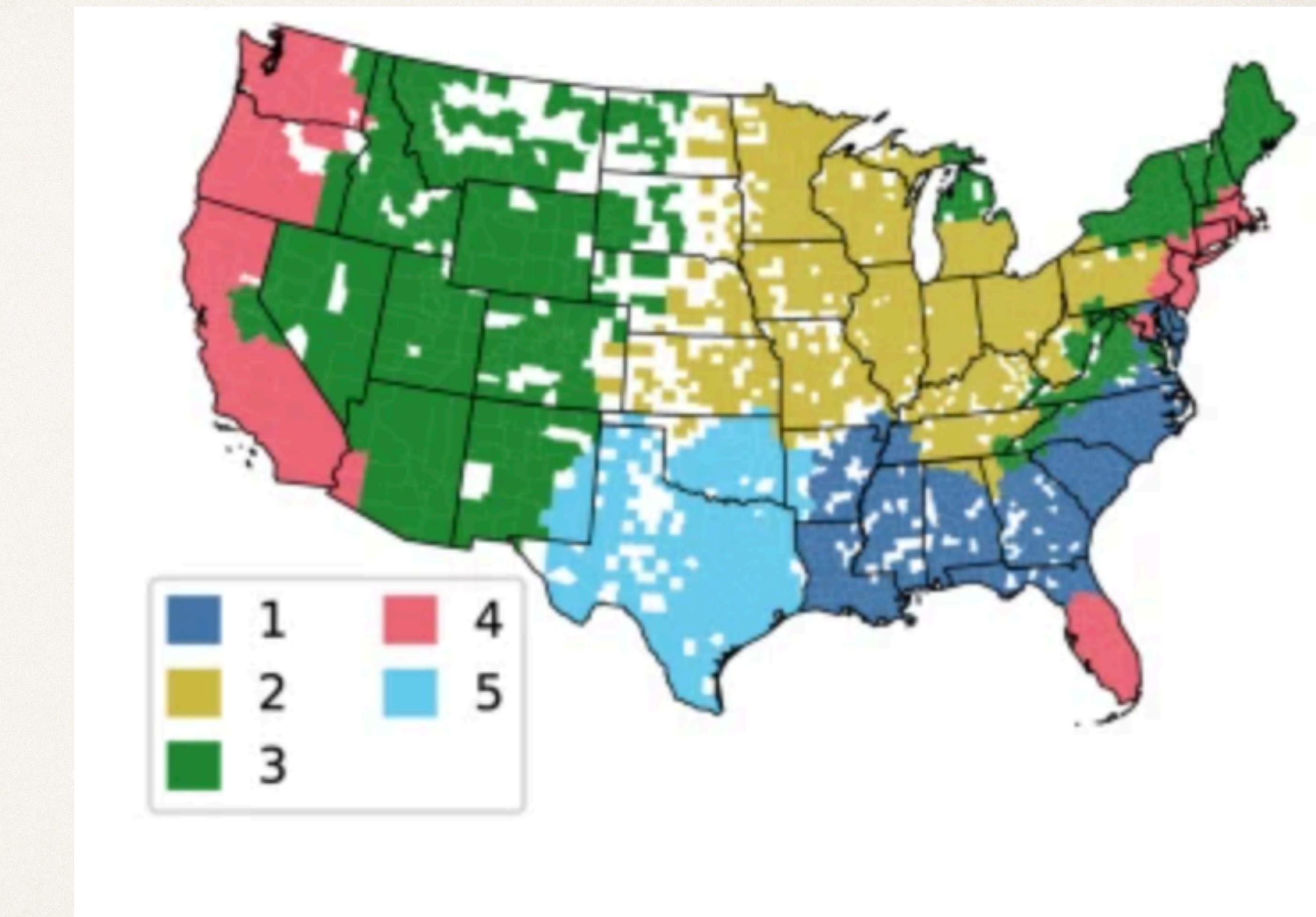


1 hour ago





American cultural regions from surveys

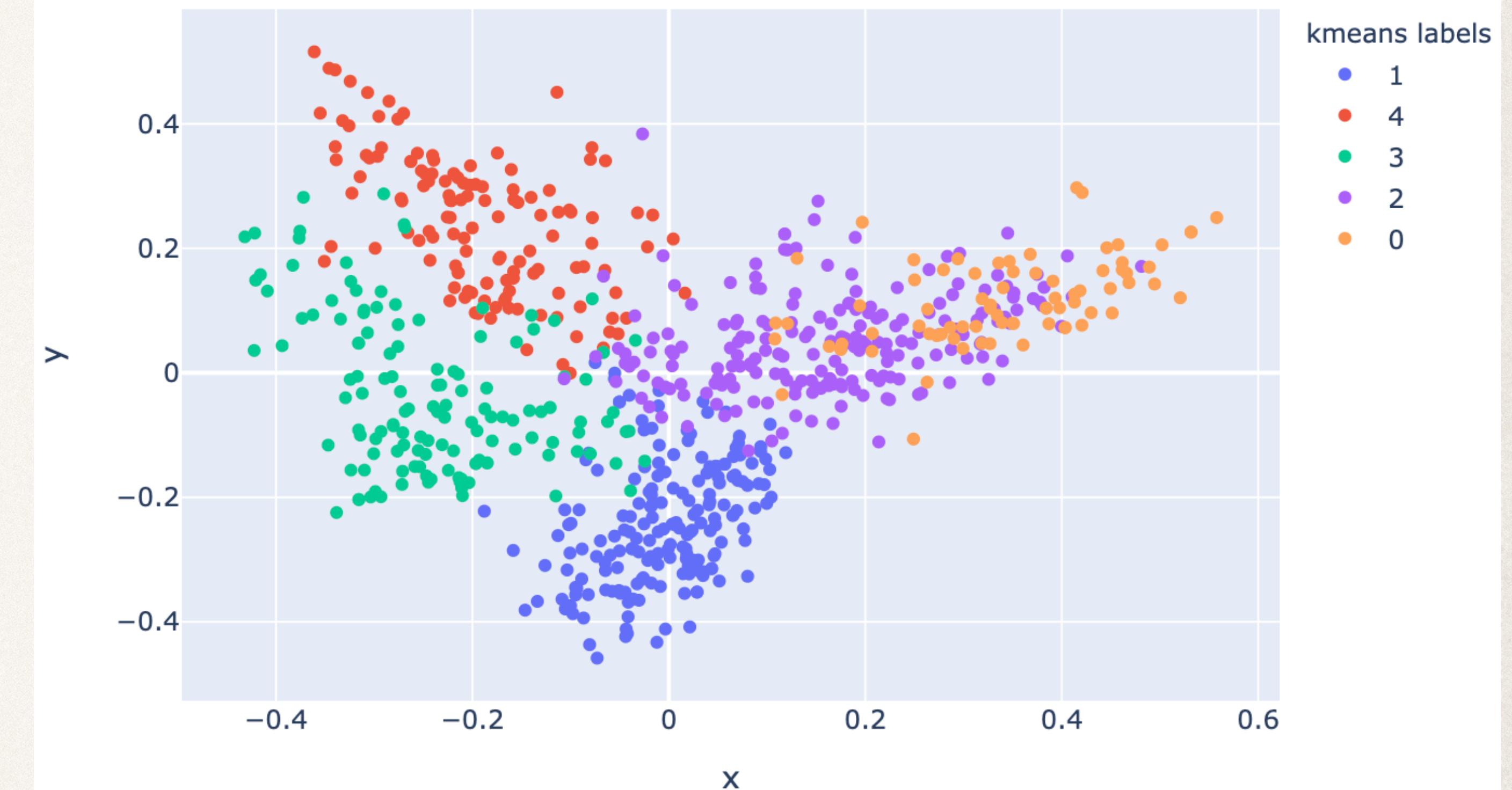


American cultural regions mapped from social media text

Input: documents

Output: partitions of
documents

K-means BBC Sport news



<https://github.com/jbesomi/texthero>

QUESTION FOR THE DAY

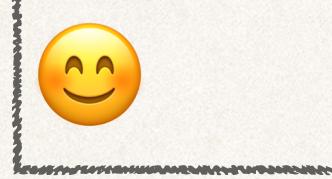
“How to group documents (e.g., speeches, lyrics, papers, tweets, etc) using their text?”

AGENDA

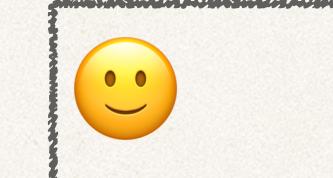
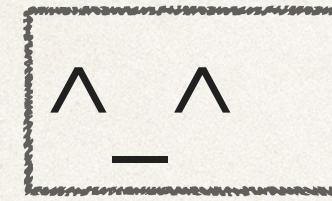
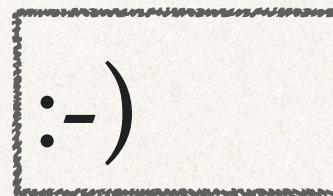
- What are the techniques of grouping documents?
- What are the challenges in grouping documents?
- How to evaluate grouping?

EXAMPLE I

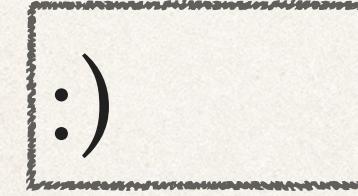
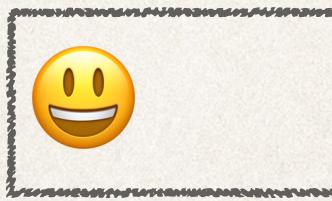
haha



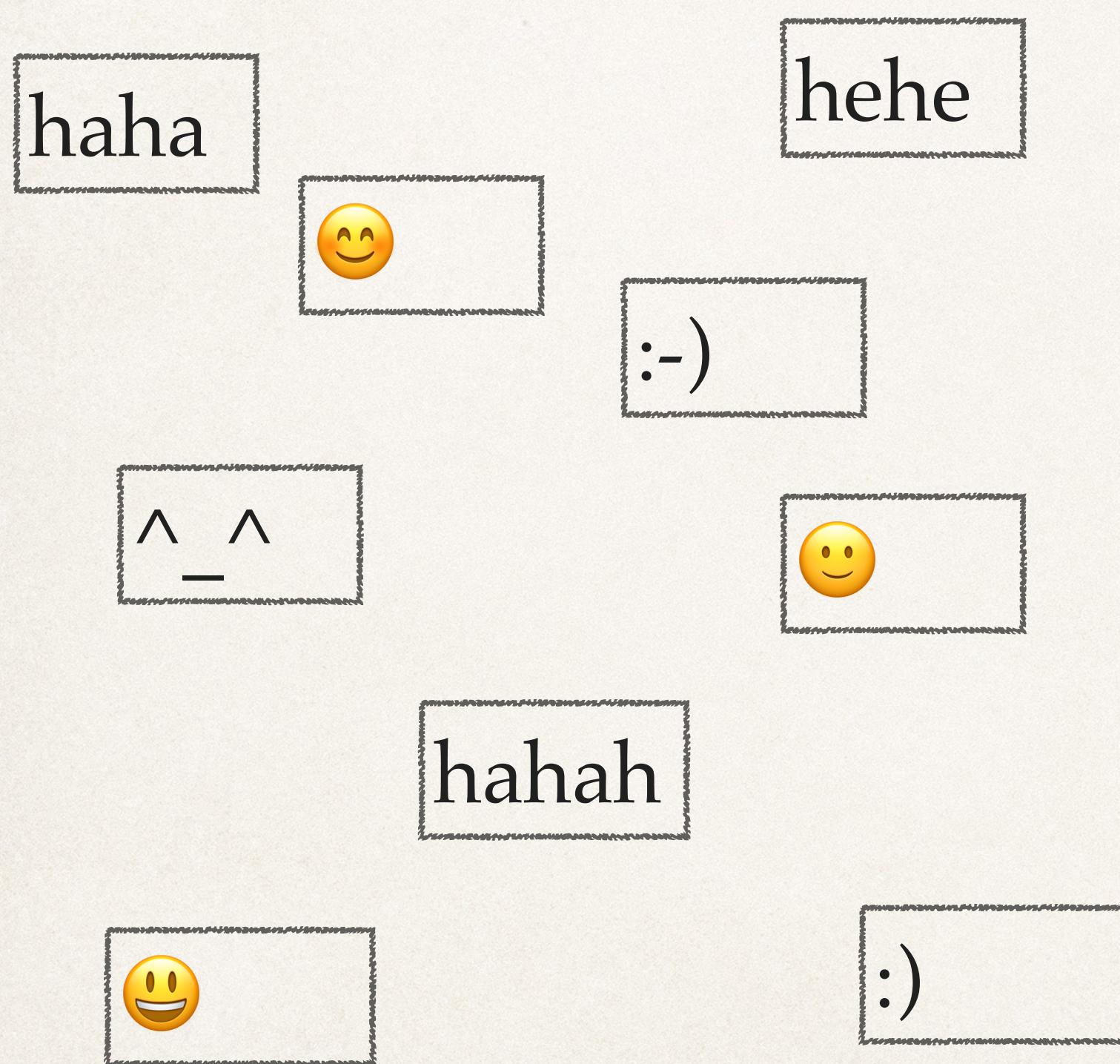
hehe



hahah



EXAMPLE I



Text-based

haha

hehe

hahah

Emoticons

:-)

:)

^-^

Emojis

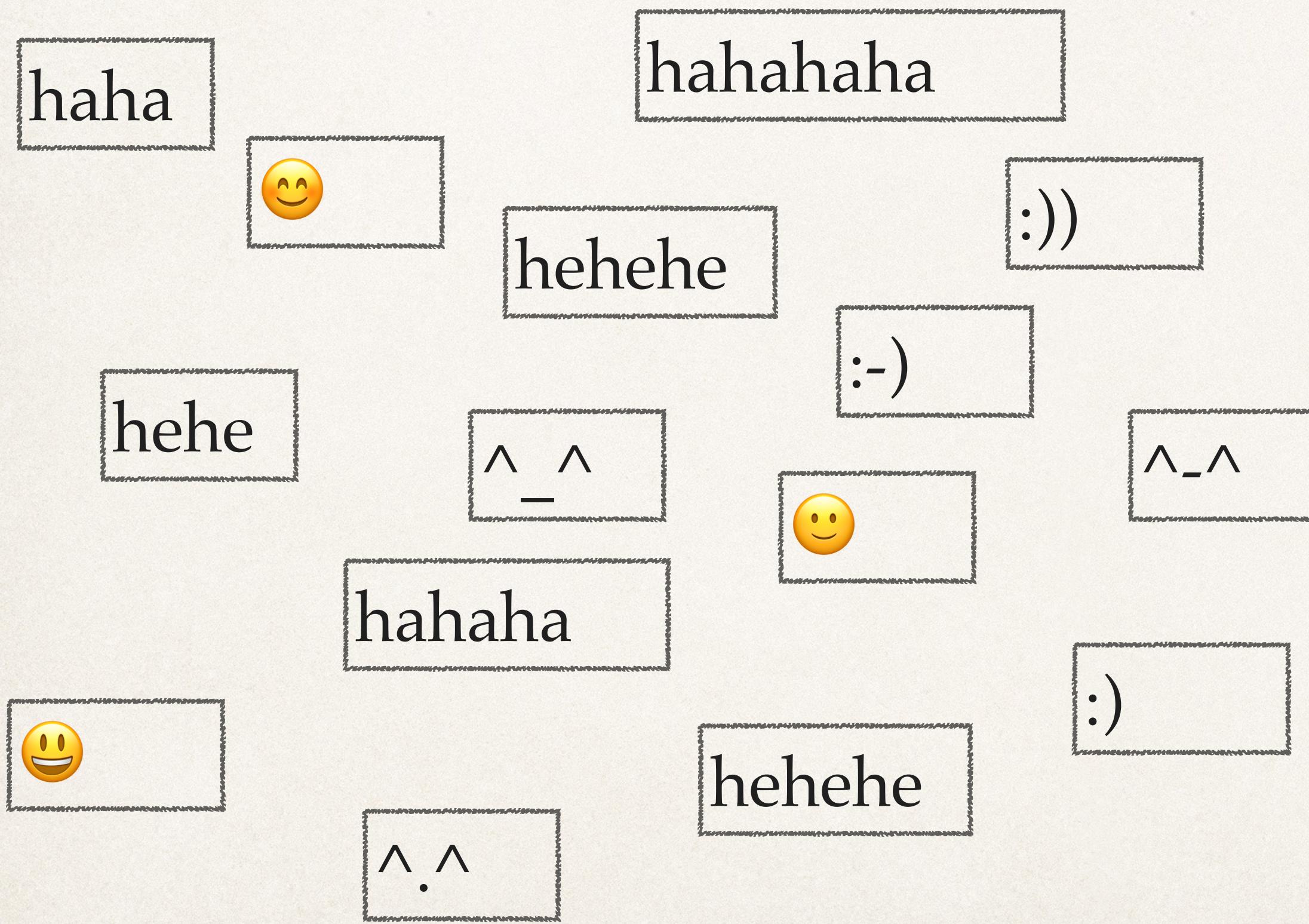
😊

😃

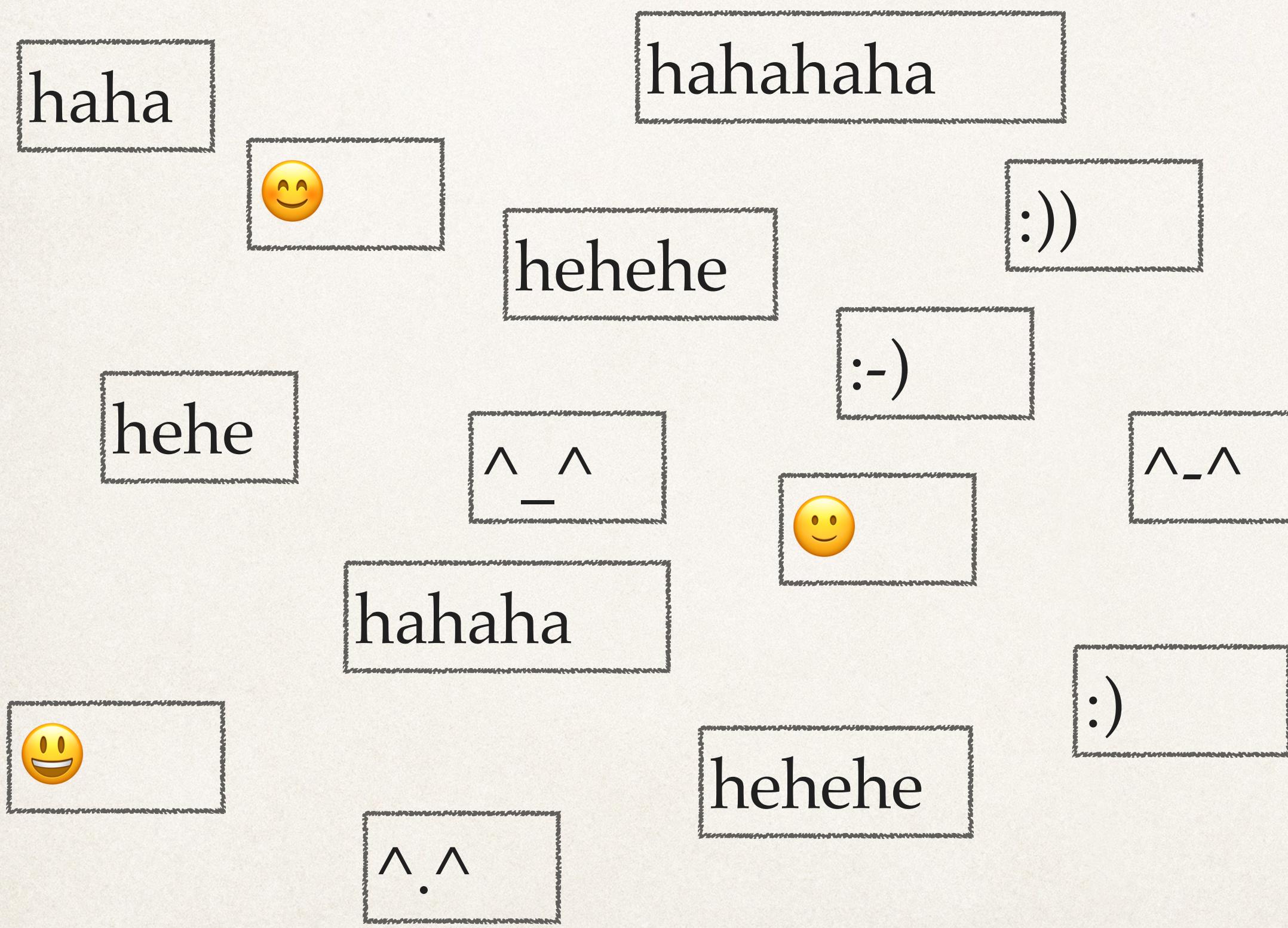
😄

😆

EXAMPLE II



EXAMPLE II



Text-based

ha-variation
ha-ha
ha-ha-ha
ha-ha-ha-ha-ha

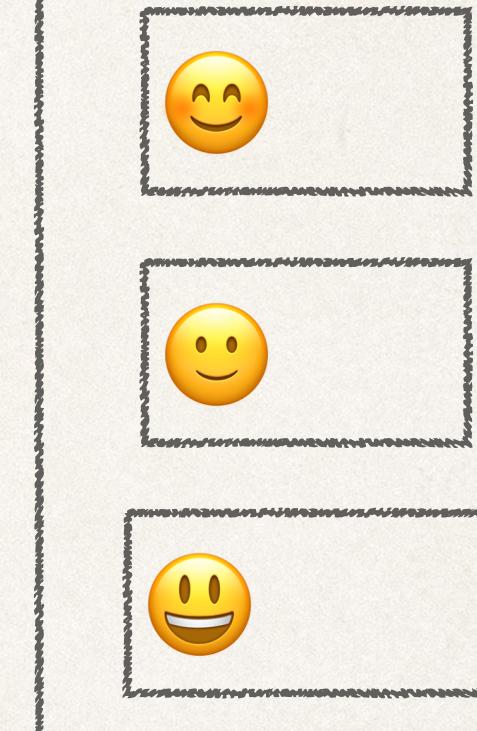
he-variation
he-he
he-he-he
he-he-he-he

Emoticons

Sideface emoticons
:-)
:-))
:-)

Frontface emoticons
^-^
^-.-^
^-_-^

Emojis



CLUSTERING TECHNIQUES

- Many ways but the most popular is kmeans clustering and its variations

KMEANS

KMEANS

- Let d be a document; $d \in \mathbb{R}^m$
- D is a set of documents; $D = [d^{(1)}; d^{(2)}; \dots; d^{(|D|)}] \in \mathbb{R}^{|D| \times m}$

KMEANS

- Let d be a document; $d \in \mathbb{R}^m$
- D is a set of documents; $D = [d^{(1)}; d^{(2)}; \dots; d^{(|D|)}] \in \mathbb{R}^{|D| \times m}$
- Objective: Find a partition set $\Pi = \{\pi_1, \pi_2, \dots, \pi_k\}$ such that:
 - $\pi_i \cap \pi_j = \emptyset$ and $\pi_i \cup \pi_j = D$
 - $\text{dist}(d_i^{(q)}, d_j^{(r)})$ should be small if $i = j$
 - $\text{dist}(d_i^{(q)}, d_j^{(r)})$ should be big if $i \neq j$

KMEANS

KMEANS

- Choose k

KMEANS

- Choose k
- Initialize k vectors of m dimensions as cluster centers

KMEANS

- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:

KMEANS

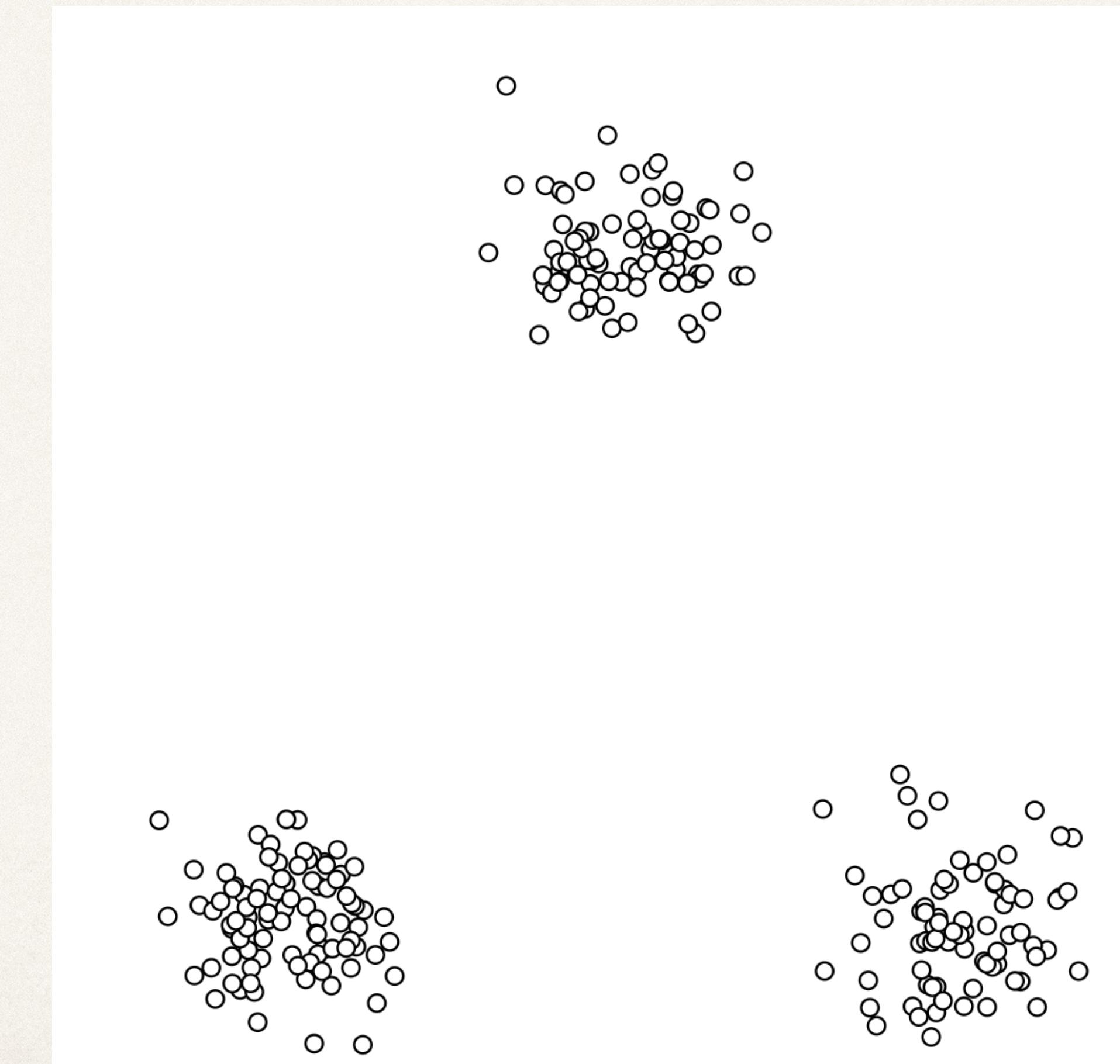
- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:
 - For every document, assign closest cluster center

KMEANS

- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:
 - For every document, assign closest cluster center
 - Recalculate cluster centers based on recent assignments

KMEANS

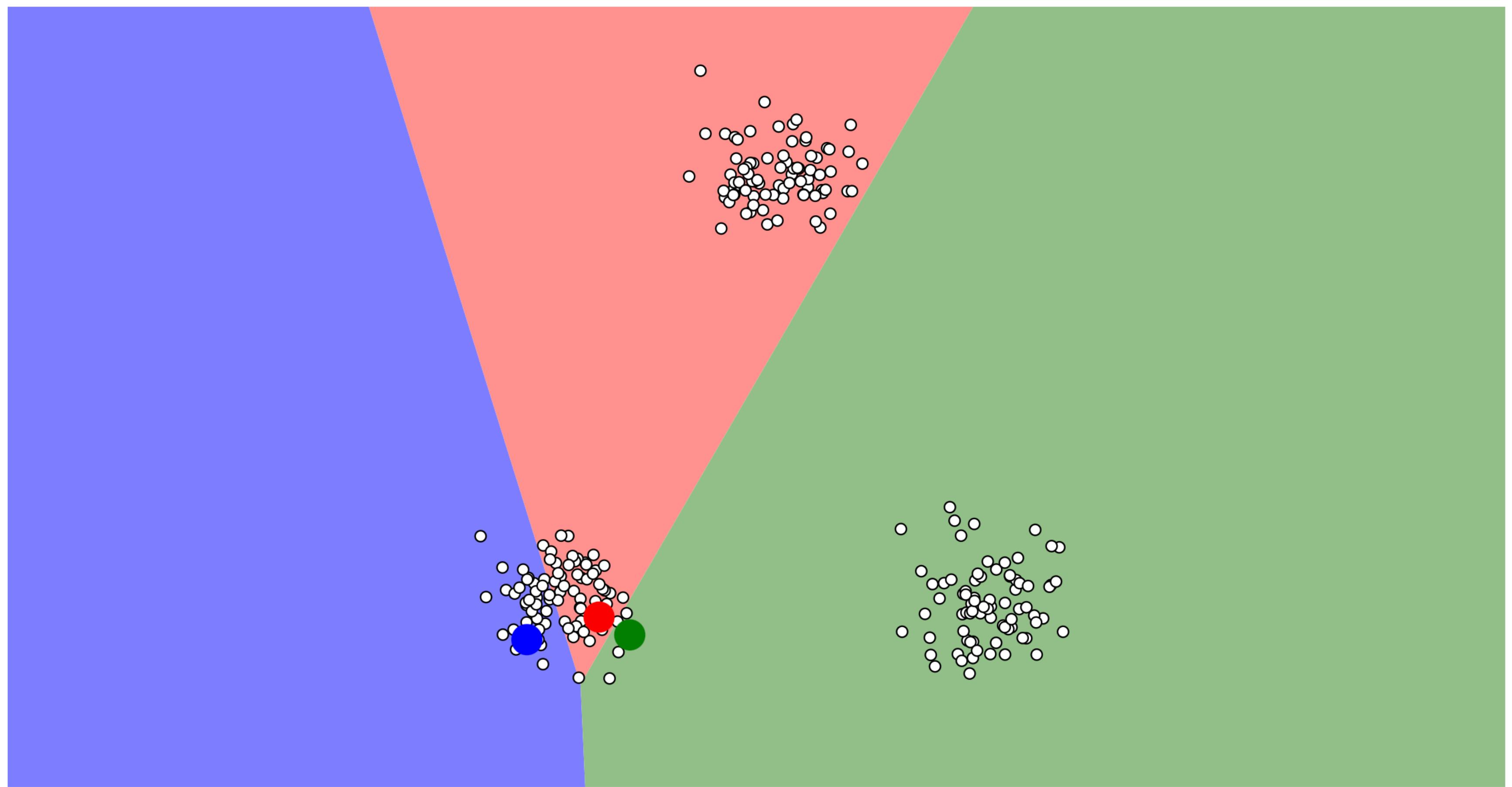
We may have a gaussian data distribution



Source: <https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>

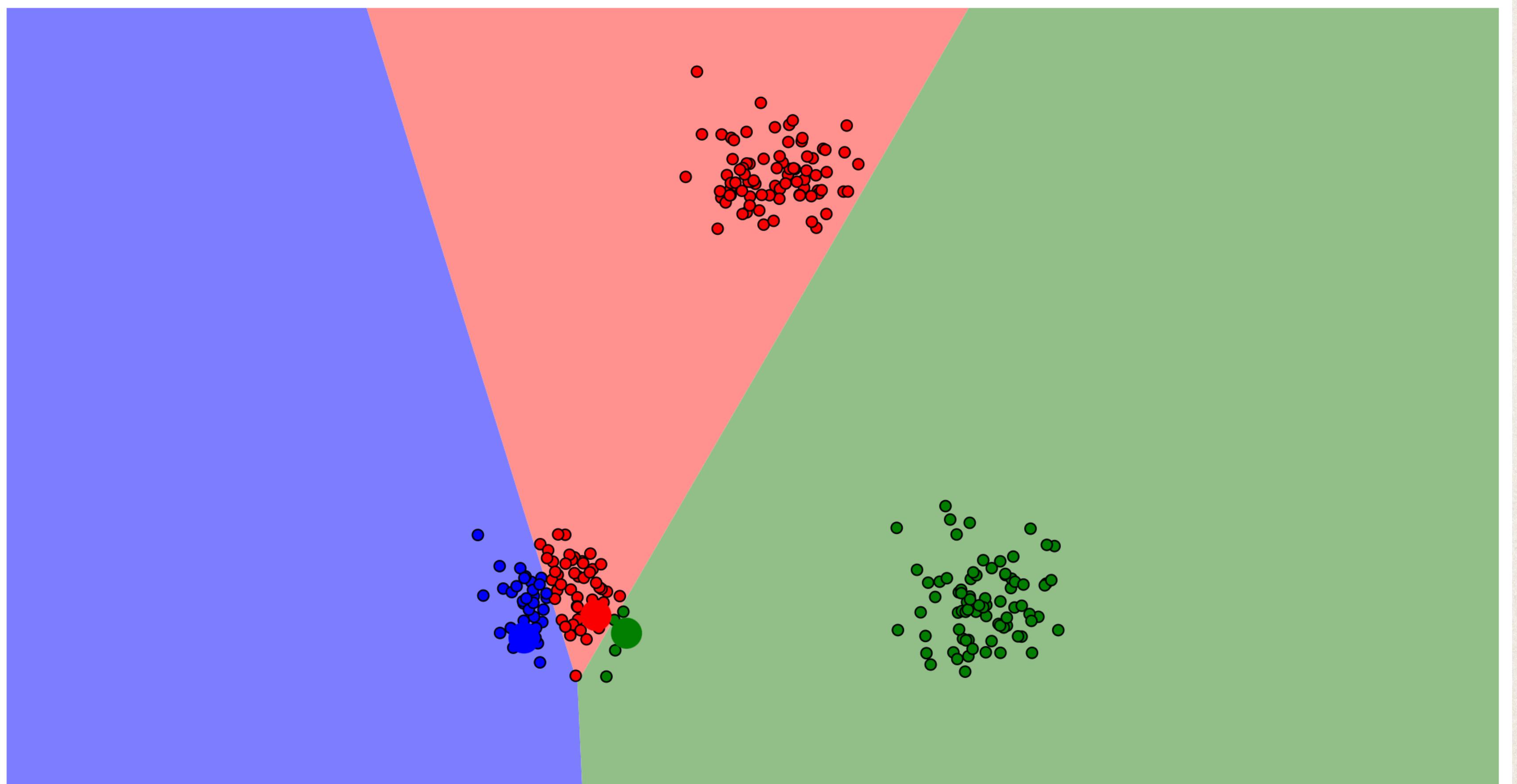
KMEANS

Initialize
the cluster
centers



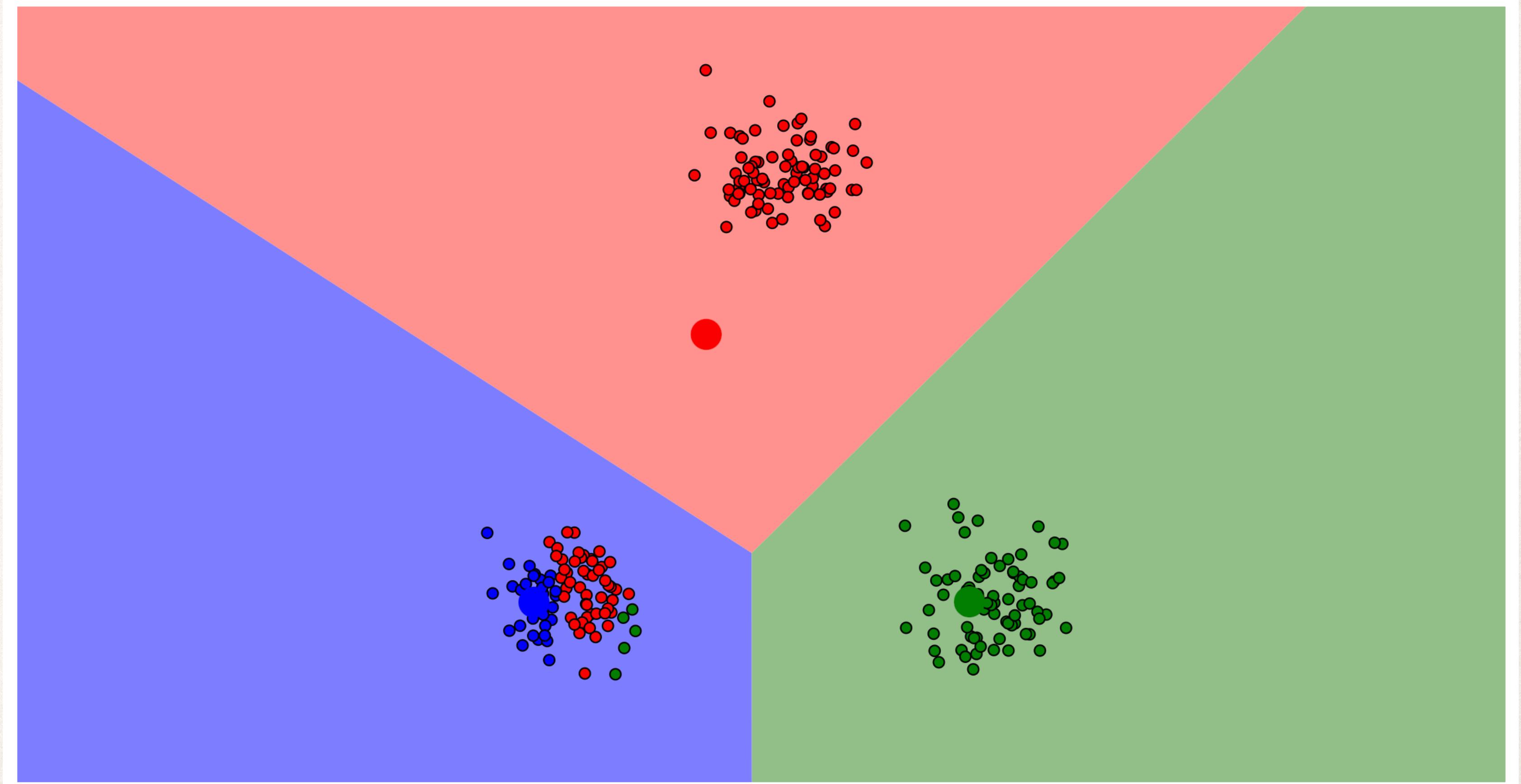
KMEANS

Partitions
formed
based on
cluster
centers



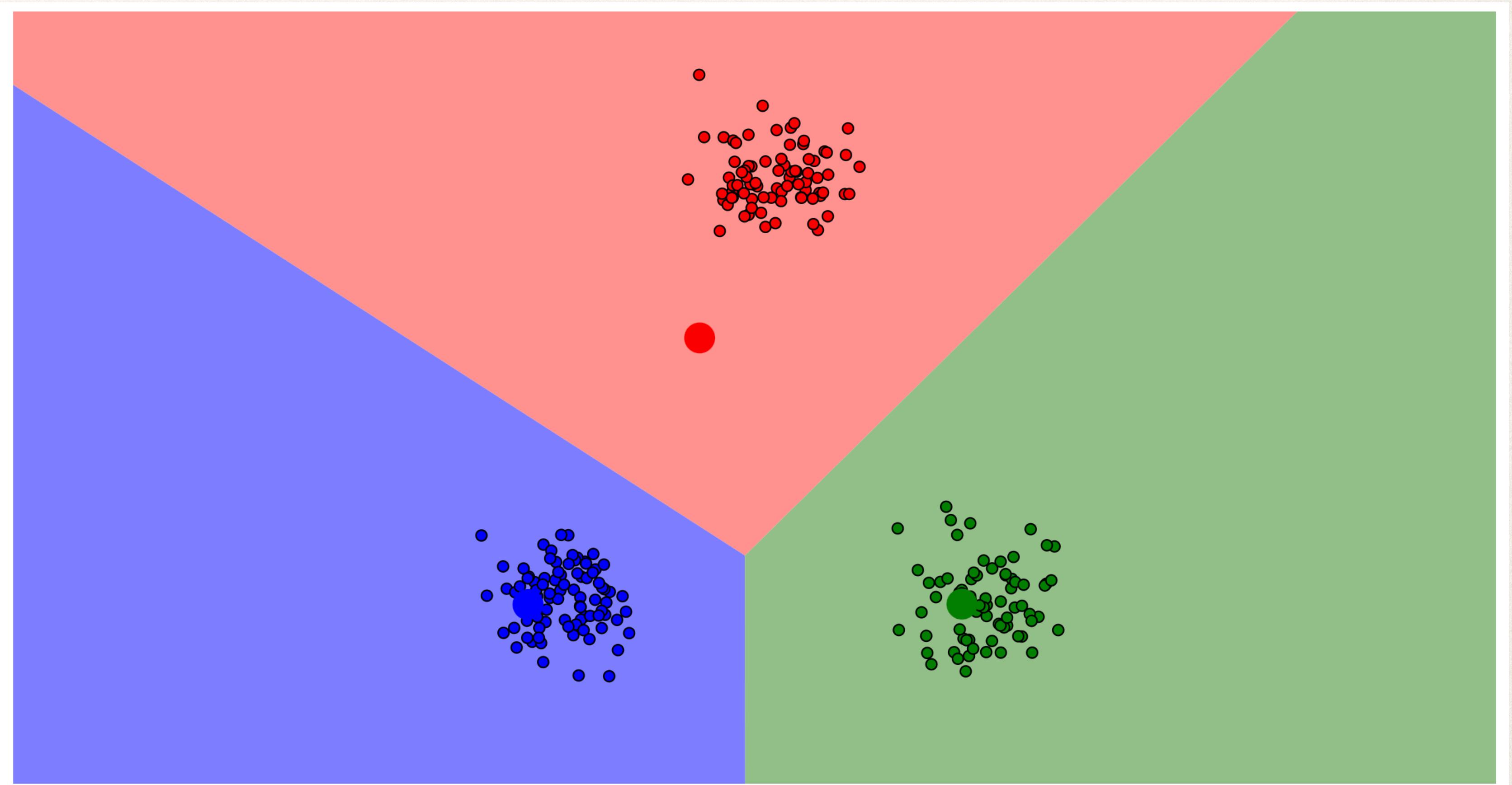
KMEANS

Update the
cluster
centers



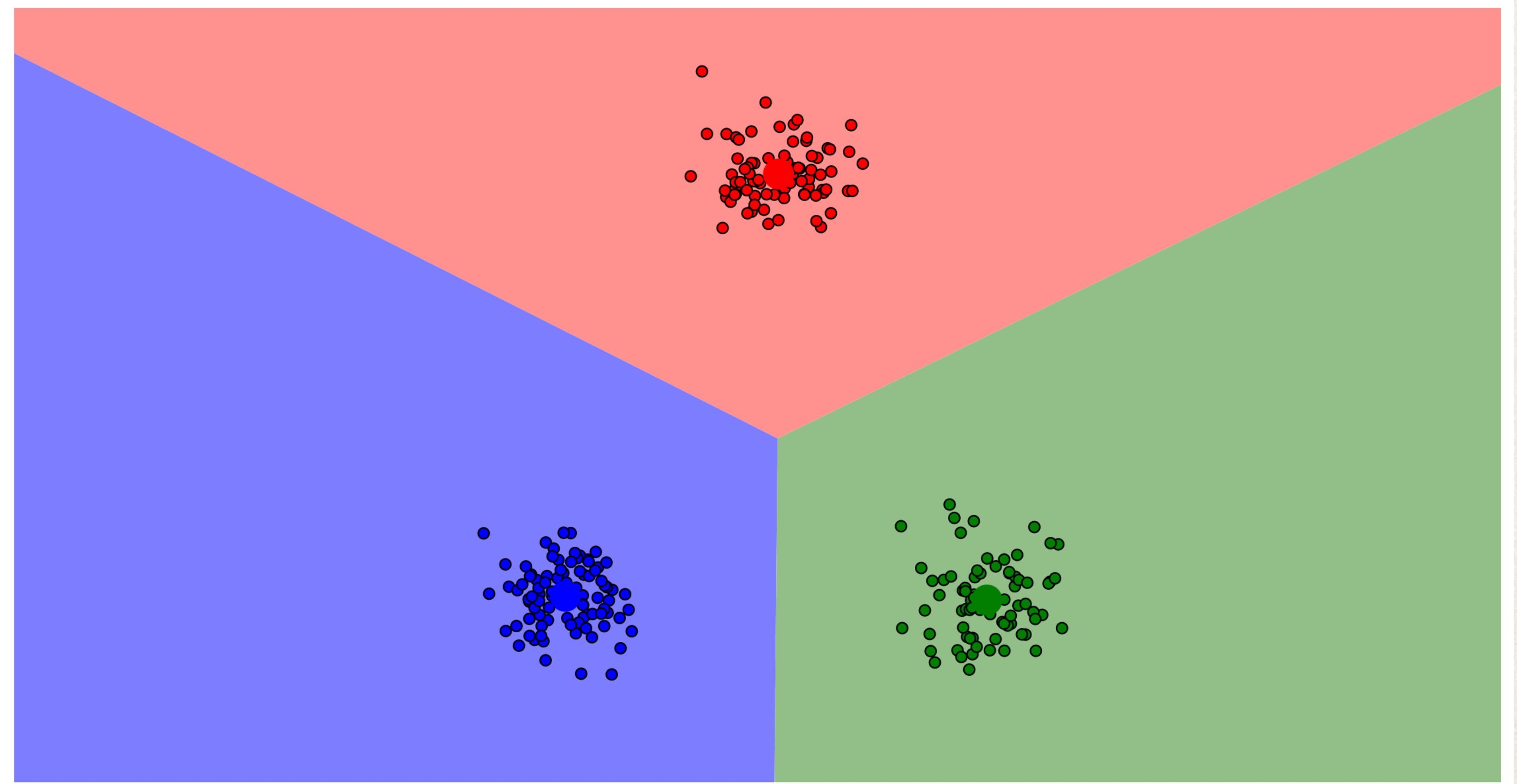
KMEANS

Reassign
the data
points



KMEANS

Repeat
until
convergence

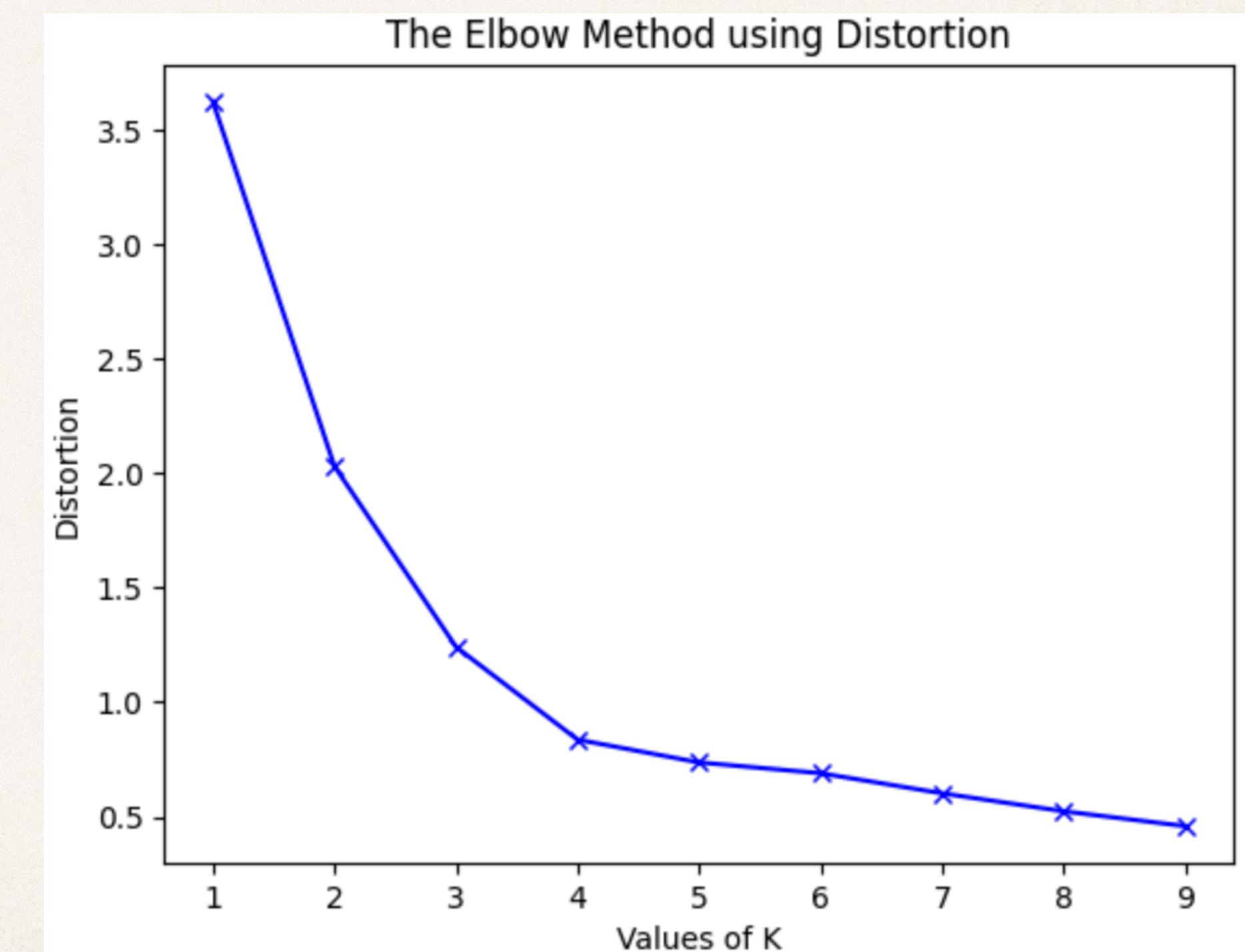


KMEANS

- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:
 - For every document, assign closest cluster center
 - Recalculate cluster centers based on recent assignments

CHOOSING K

- Sometimes we know the value k (e.g., political parties)
- In general, use your judgment or some heuristics to find k
- Bayesian non-parametric methods find the optimal k



INITIALIZATION

- Initialization may affect convergence time and optimal clustering
- There are some smart initializations but this is a non-trivial problem [see Caleb et. al]

REPEAT

- Each cluster assignment and centroid recalculation step is repeated several times
- Since the algorithm is sensitive to initialization, the entire procedure may be repeated to get stability.

MEASURING CLOSENESS

- Typically euclidean distances are used to measure closeness
- Non-euclidean metrics or kernel transformations of vectors useful when data is not linearly separable

TEXT PREPROCESSING

TEXT PREPROCESSING

- Preprocessing steps such as stop words removal, lemmatization, normalization, etc are crucial

TEXT PREPROCESSING

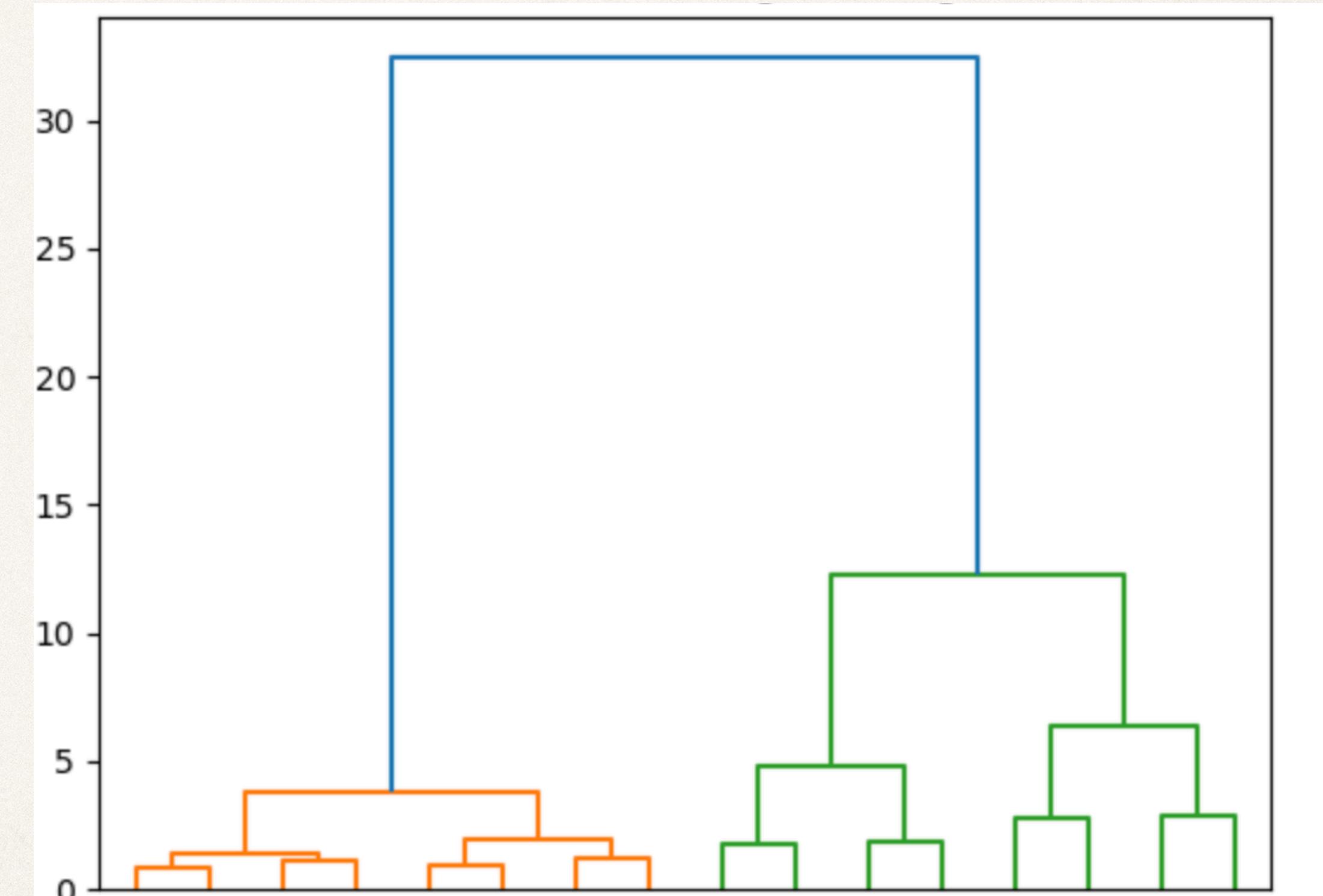
- Preprocessing steps such as stop words removal, lemmatization, normalization, etc are crucial
- In general, clustering works best with dense features than sparse, so some dimensionality reduction technique can also be useful

SHORTCOMINGS

- Knowing K is not easy (e.g., how many genres of books?)
- Outliers can have a strong effect on cluster formation
- Kmeans does not extract a hierarchical structure from the data

OTHER METHODS

- Agglomerative clustering can extract hierarchical clustering
- Affinity propagation algorithm on a graph



IN CLASS

- Clustering demo