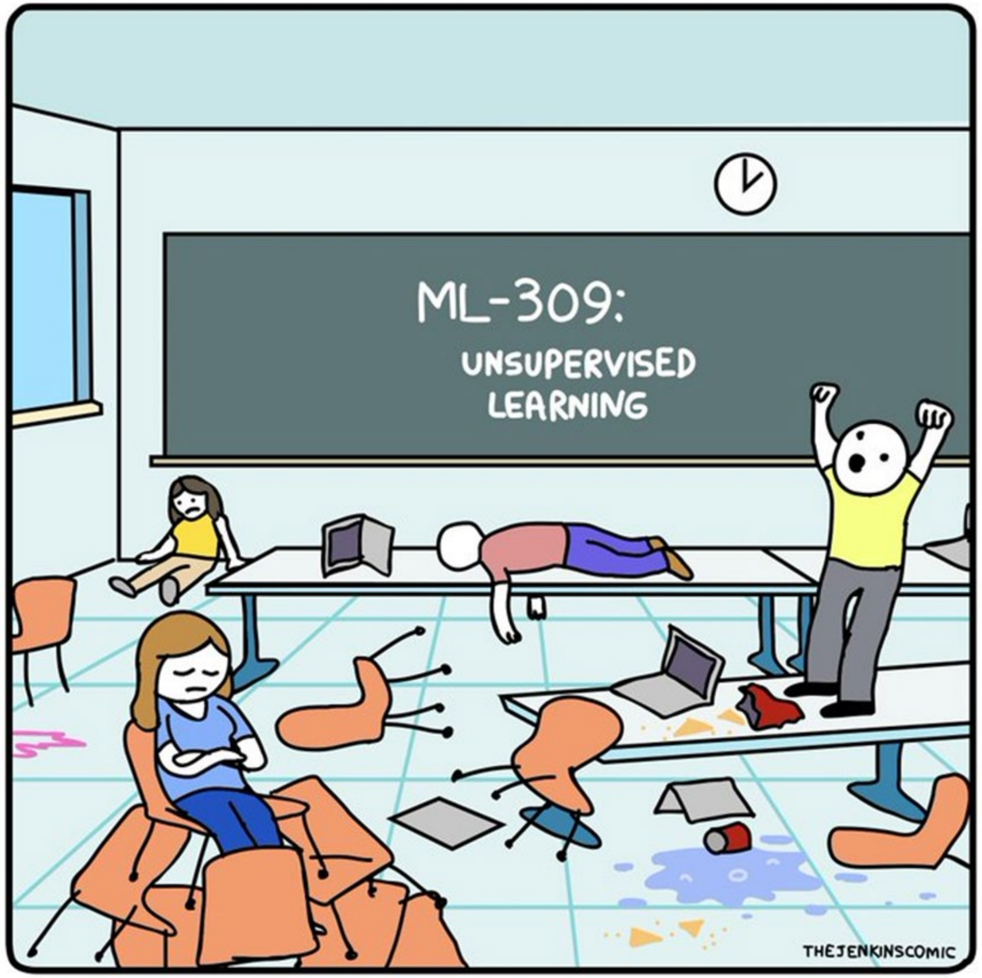
Clustering

Introduction to data analysis: Lecture 12

Ori Plonsky

Spring 2023



Supervised vs. Unsupervised Learning

- Supervised Learning
  - Data is labeled. We have the ground truth.
  - We want to predict how to label a new data point based on the input data.
  - Used in the context of classification or regression.

Supervised vs. Unsupervised Learning

- Supervised Learning
    - Data is labeled. We have the ground truth.
    - We want to predict how to label a new data point based on the input data.
    - Used in the context of classification or regression.
- Unsupervised Learning
    - Data is unlabeled. There are only “predictors”
    - The algorithm’s goal is to model the structure of the data.
    - Used is the context of clustering.

Clustering

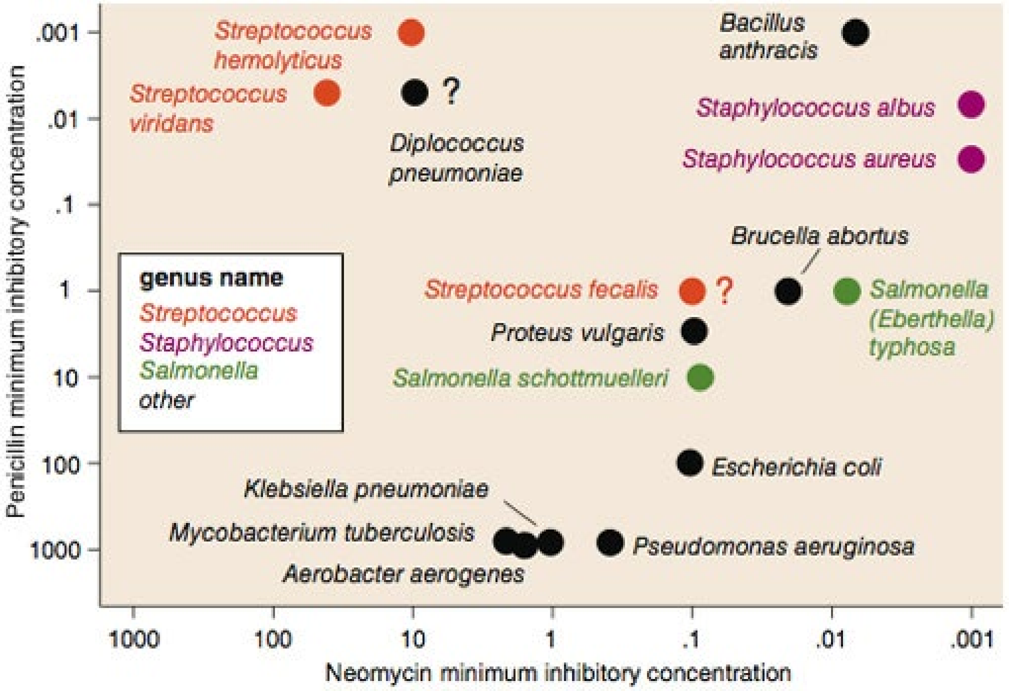
- Dividing data into groups of similar data points, when we do not have a pre-specified set of groups
- Examples:

Clustering

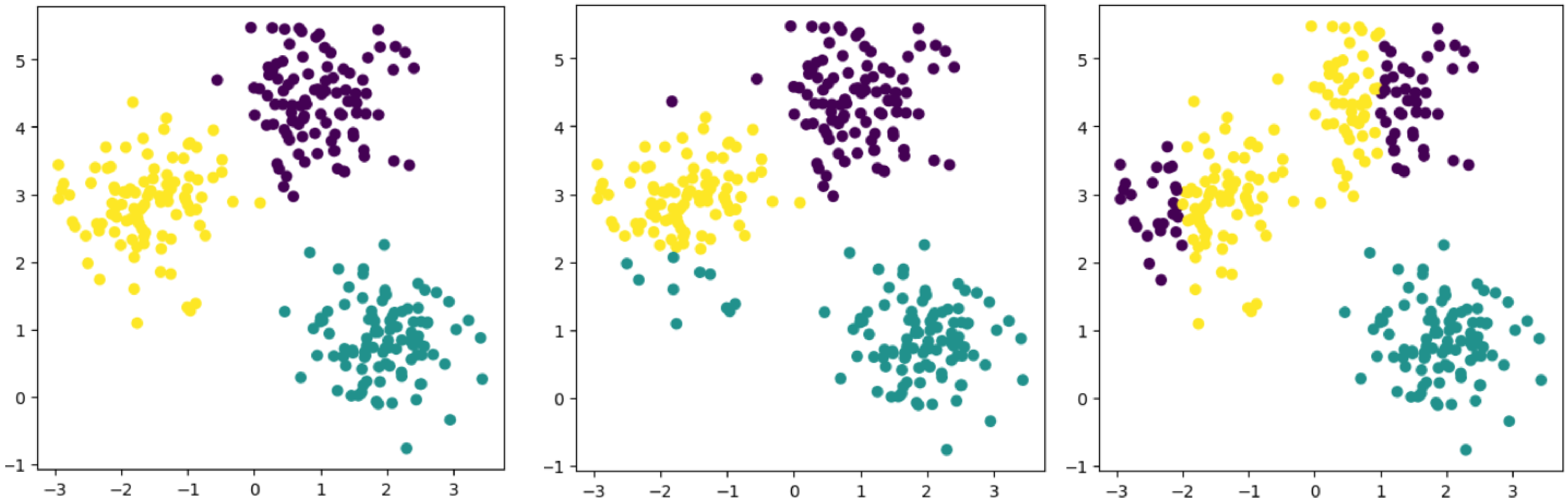
- Dividing data into groups of similar data points, when we do not have a pre-specified set of groups
- Examples:
  - Customer segmentation: Group customers to “types”
  - Group similar photos (e.g. faces)
  - Group genes/species by their attributes
  - Detect anomalies (e.g. fraud detection)

Clustering

- Dividing data into groups of similar data points, when we do not have a pre-specified set of groups
- Examples:
  - Customer segmentation: Group customers to “types”
  - Group similar photos (e.g. faces)
  - Group genes/species by their attributes
  - Detect anomalies (e.g. fraud detection)



What is good clustering?



How should we cluster our data?

How should we cluster our data?

In general, we want:

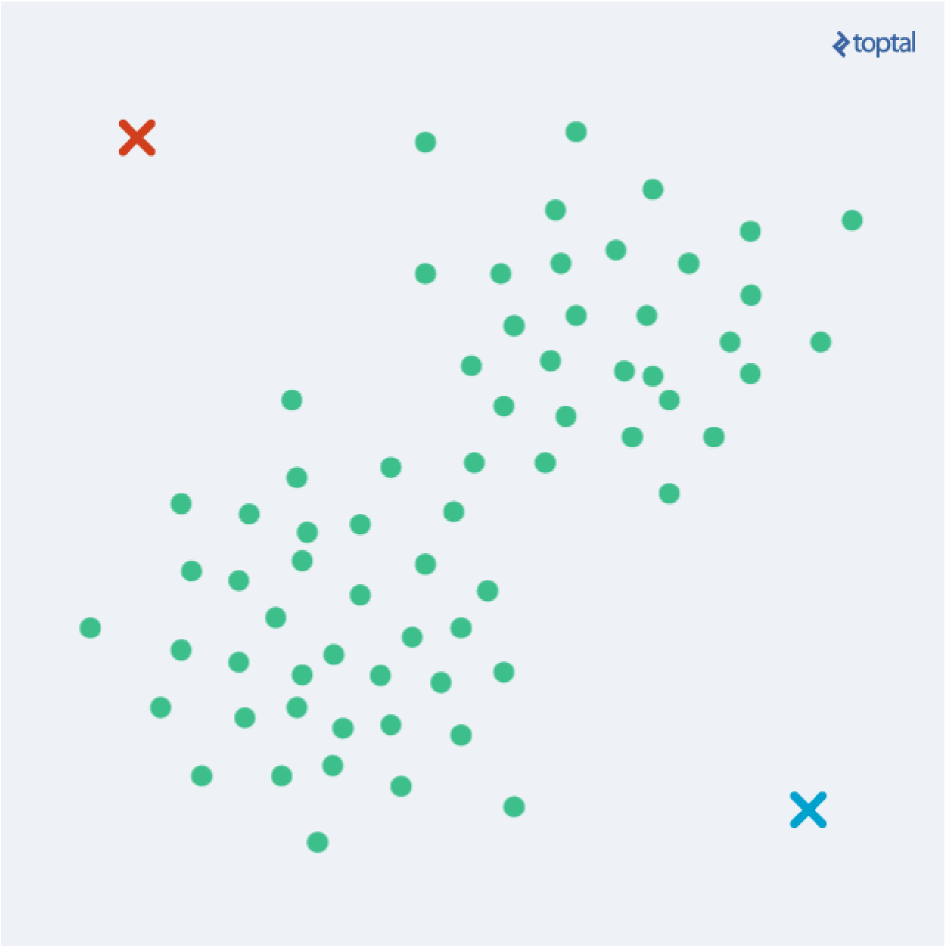
- Data points in the same cluster should be close to each other
- Data points in different clusters should be far from each other

K-Means clustering

Main steps in the algorithm:

1. Pick K
2. Initialize K centroids (centers of clusters)
3. Assign each data point to its closest centroid
4. Update centroids to be at the center of the assigned points
5. Repeat 3, 4 until no more updates in assignment of data points to clusters

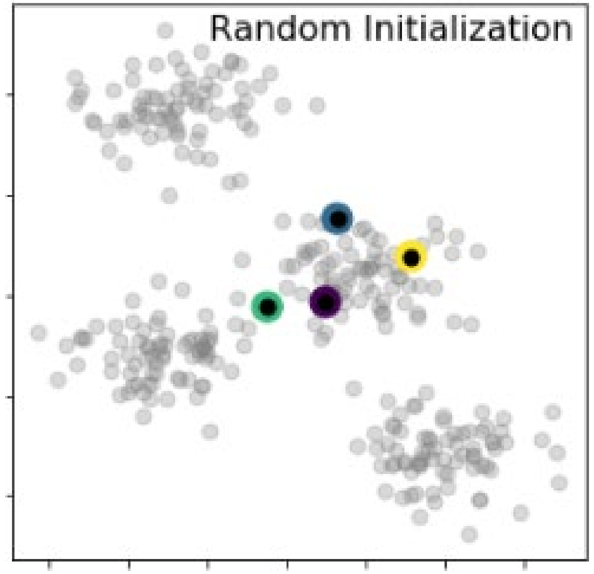
K-Means clustering



https://www.toptal.com/machine-learning/clustering-algorithms

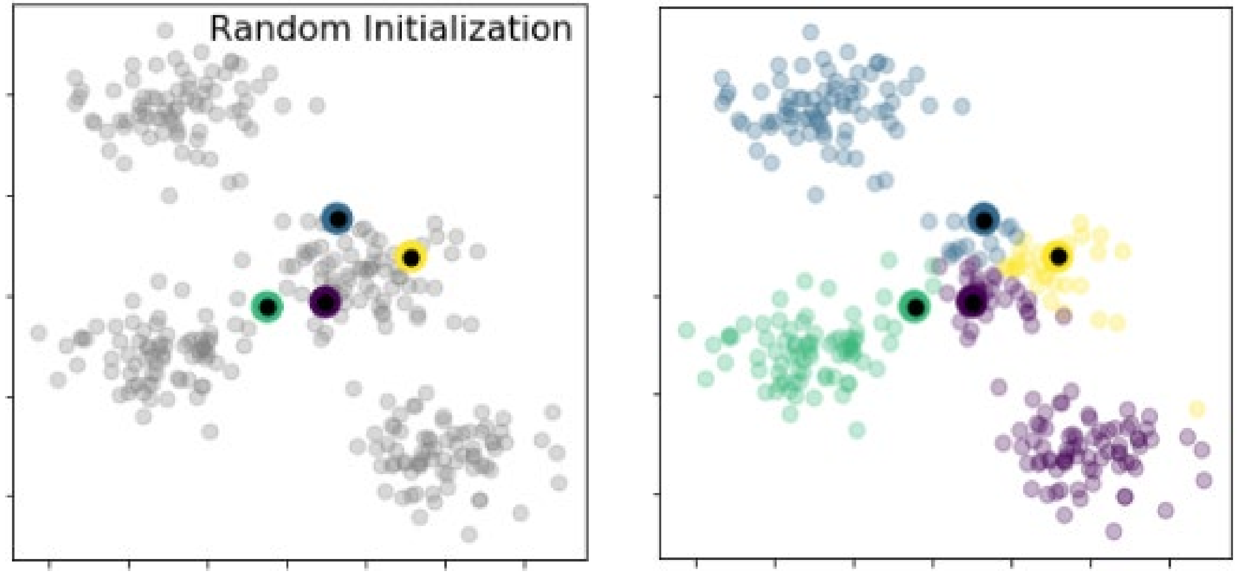
K-Means: 1st iteration

- Randomly choose K centroids μ^j that will serve as initial cluster centers (not necessarily from your data points)



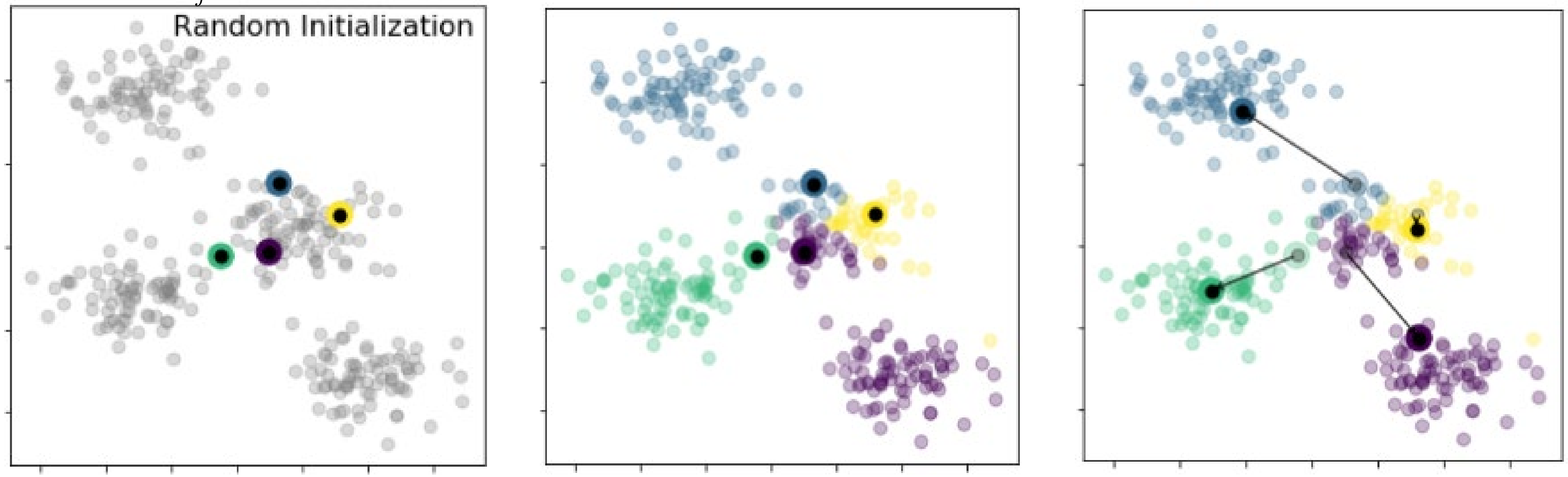
K-Means: 1st iteration

- Randomly choose K centroids μ^j that will serve as initial cluster centers (not necessarily from your data points)
- Compute distances between data points and cluster centroids, $\left\| x_i - \mu^j \right\|$ and assign each point to its closest centroid

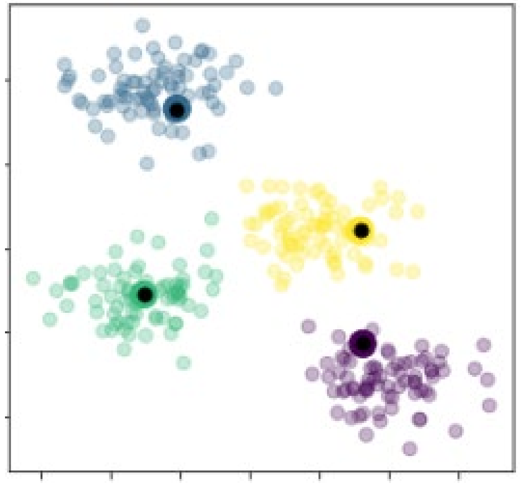


K-Means: 1st iteration

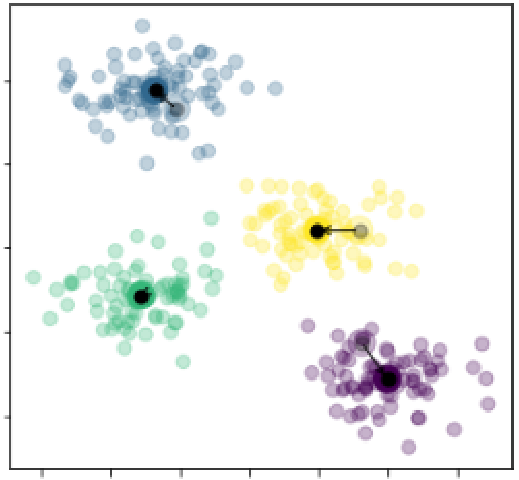
- Randomly choose K centroids μ^j that will serve as initial cluster centers (not necessarily from your data points)
- Compute distances between data points and cluster centroids, $\left\| x_i - \mu^j \right\|$ and assign each point to its closest centroid
- Update the centroids to be at the center of the data points assigned to the cluster $\mu^j = \frac{1}{N_j} \sum_{x_i \in C_j} x_i$



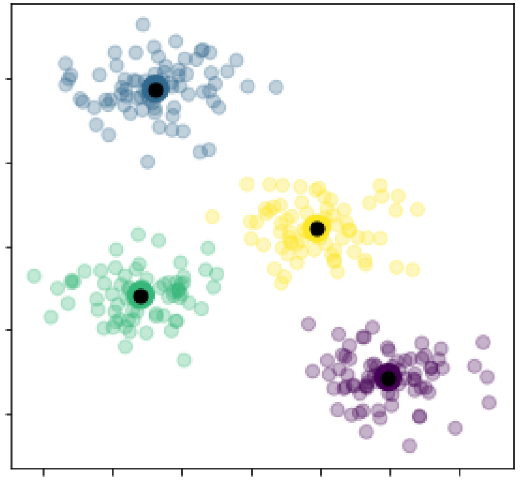
K-Means: more iterations



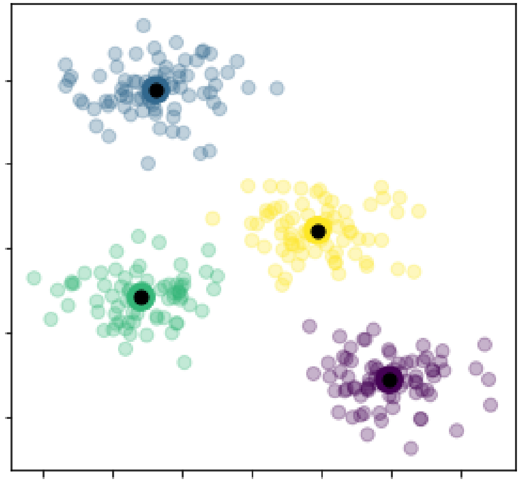
Assign points to clusters



Update centroids

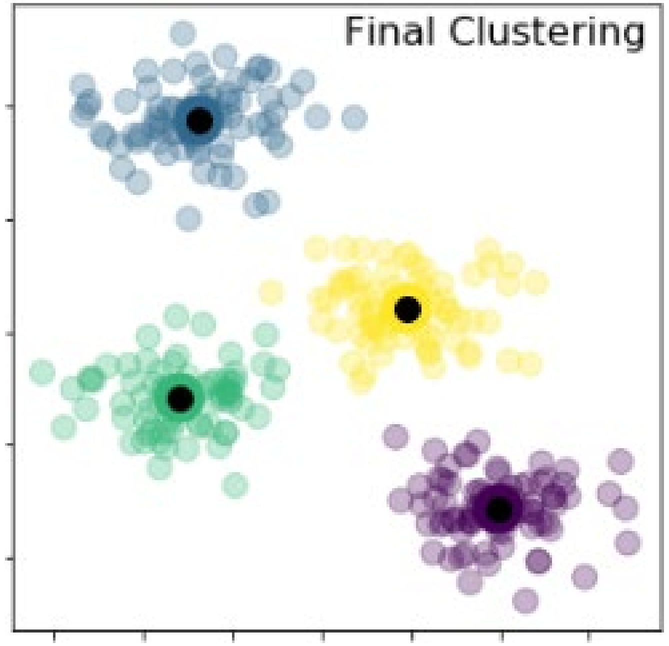


Assign points to clusters



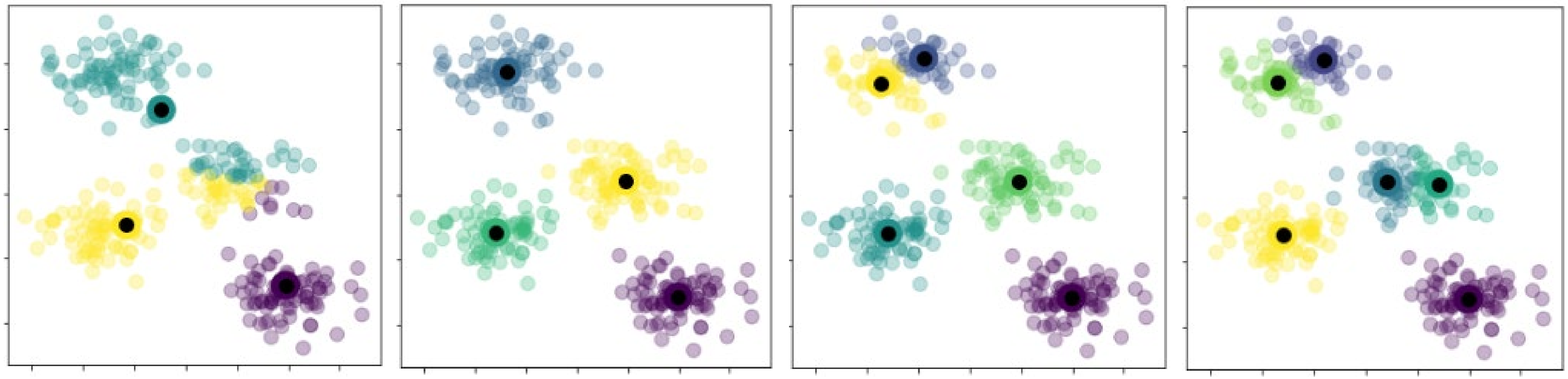
Update centroids

...



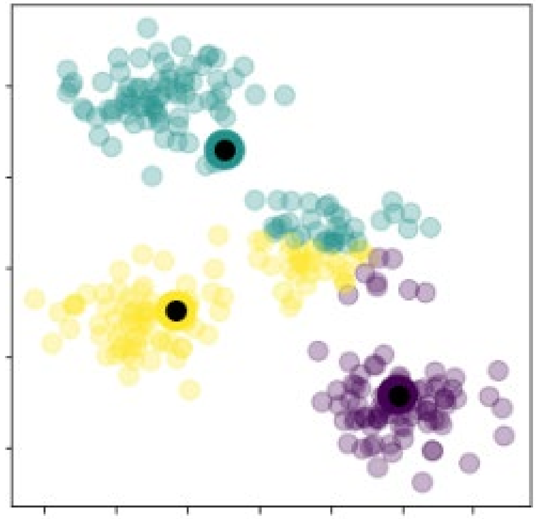
Final clusters

The effect of the choice of K

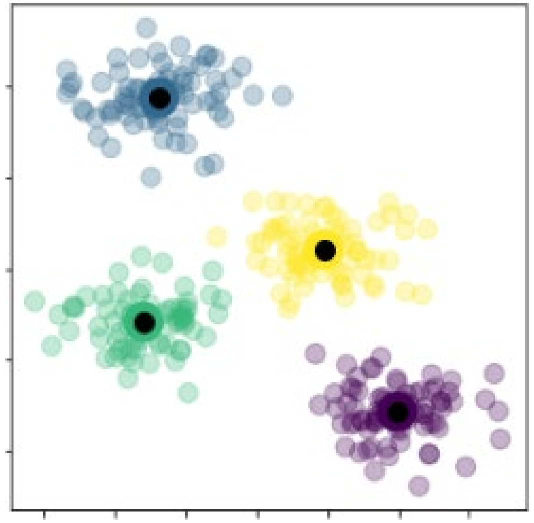


K = 3 K = 4 K = 5 K = 6

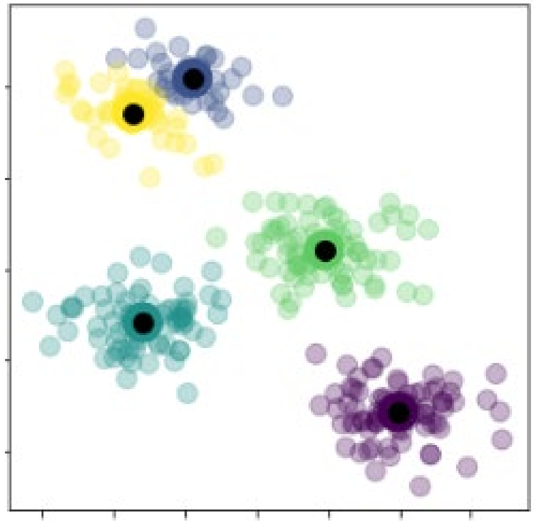
The effect of the choice of K



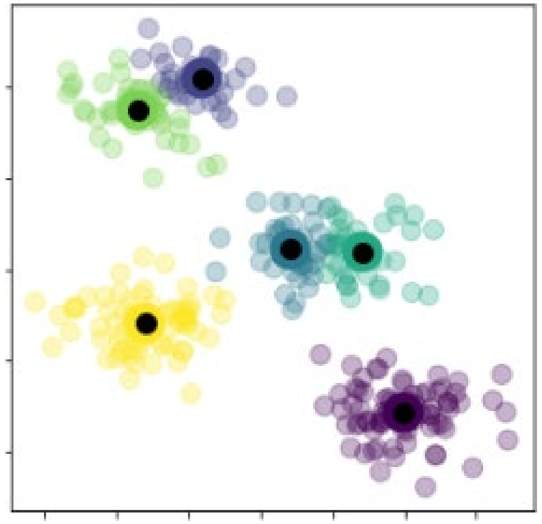
K = 3



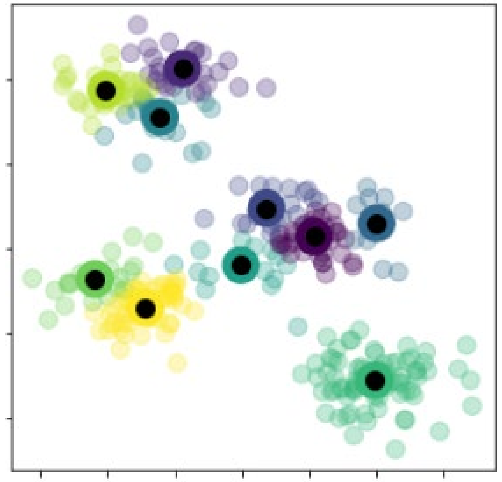
K = 4



K = 5

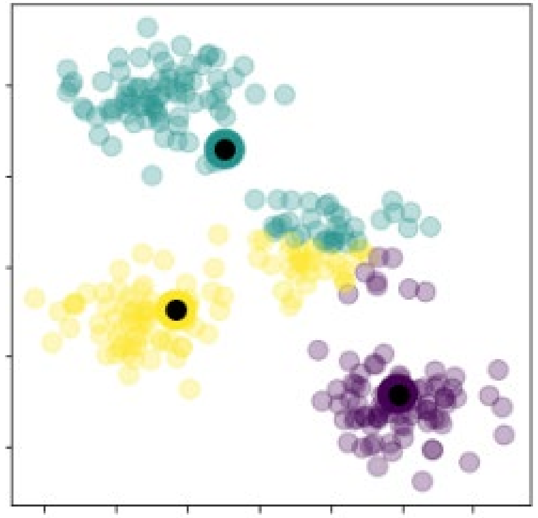


K = 6

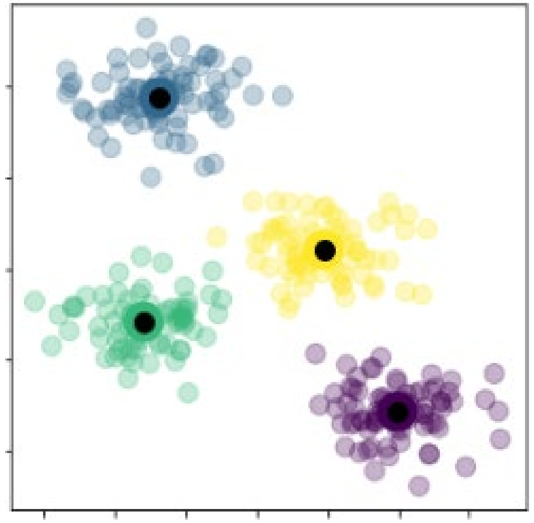


K = 10

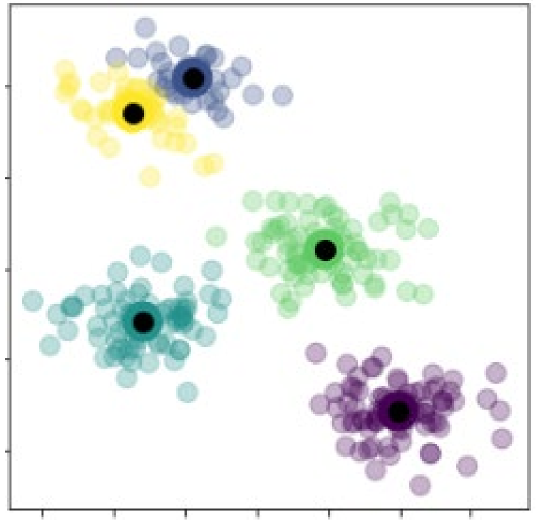
The effect of the choice of K



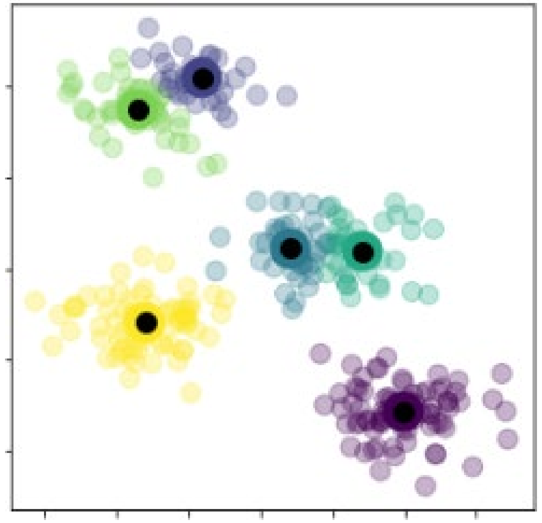
K = 3



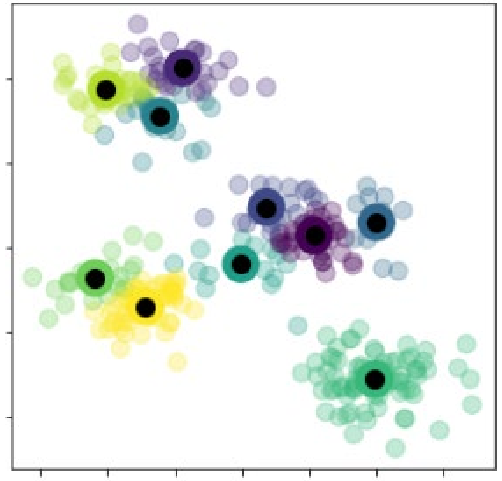
K = 4



K = 5



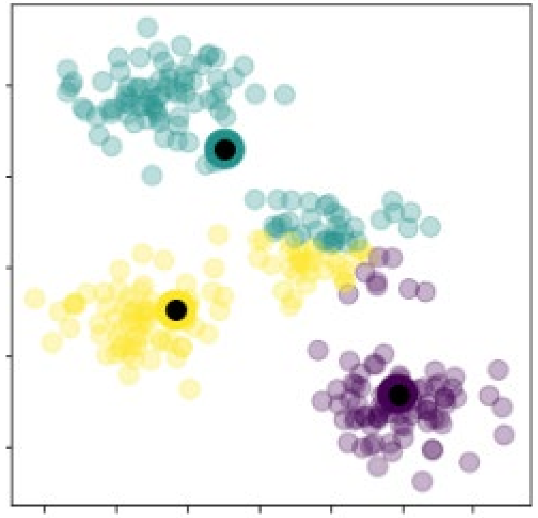
K = 6



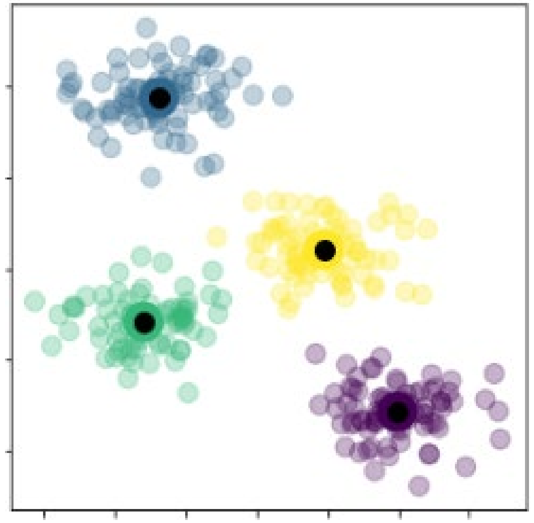
K = 10

- What will happen when K = N?

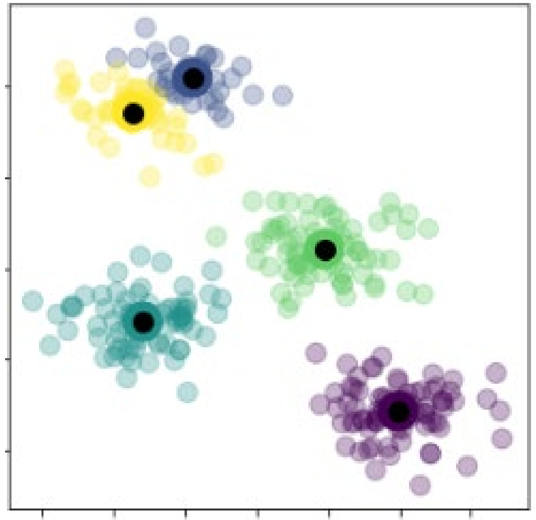
The effect of the choice of K



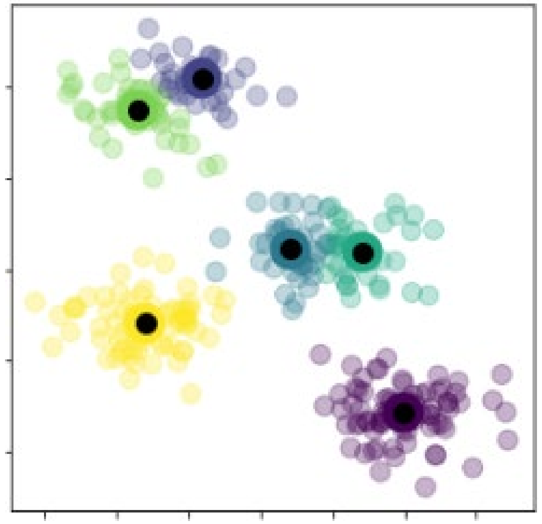
K = 3



K = 4

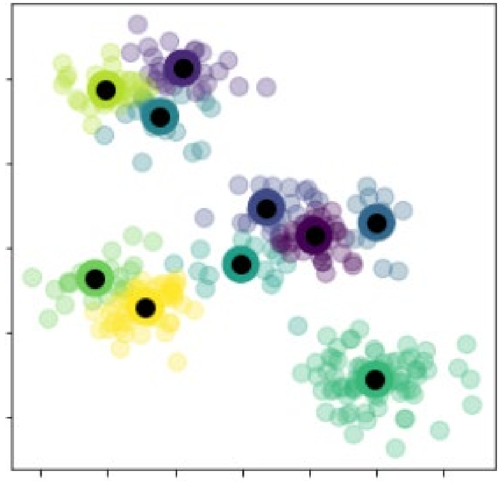


K = 5



K = 6

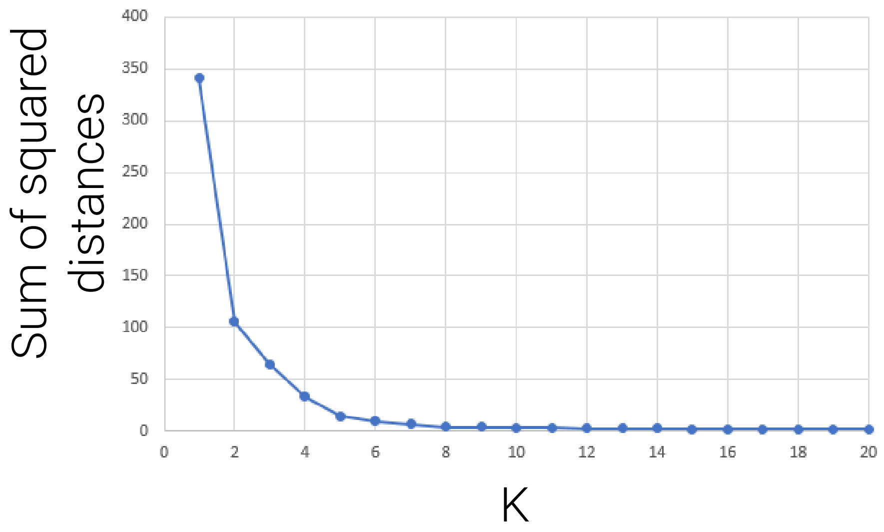
- What will happen when K = N?
  - (but no guarantee that any points are assigned to a cluster)



K = 10

The elbow method - choosing K

- Run K Means with different values of *K*
- For each *K*, compute the sum of squared distances between each point and the centroid of its cluster
  - Distances for cluster *j* is: $I_j = \sum_{i=1}^{N_j} d\left(x_{ij}, \mu_j\right)^2$
  - Sum of distances: $S_K = \sum_{j=1}^{K} I_j$
- Plot the sum of distances as a function of *K*
- Pick *K* where there is an “elbow” in the plot: adding more clusters doesn’t reduce distances by much

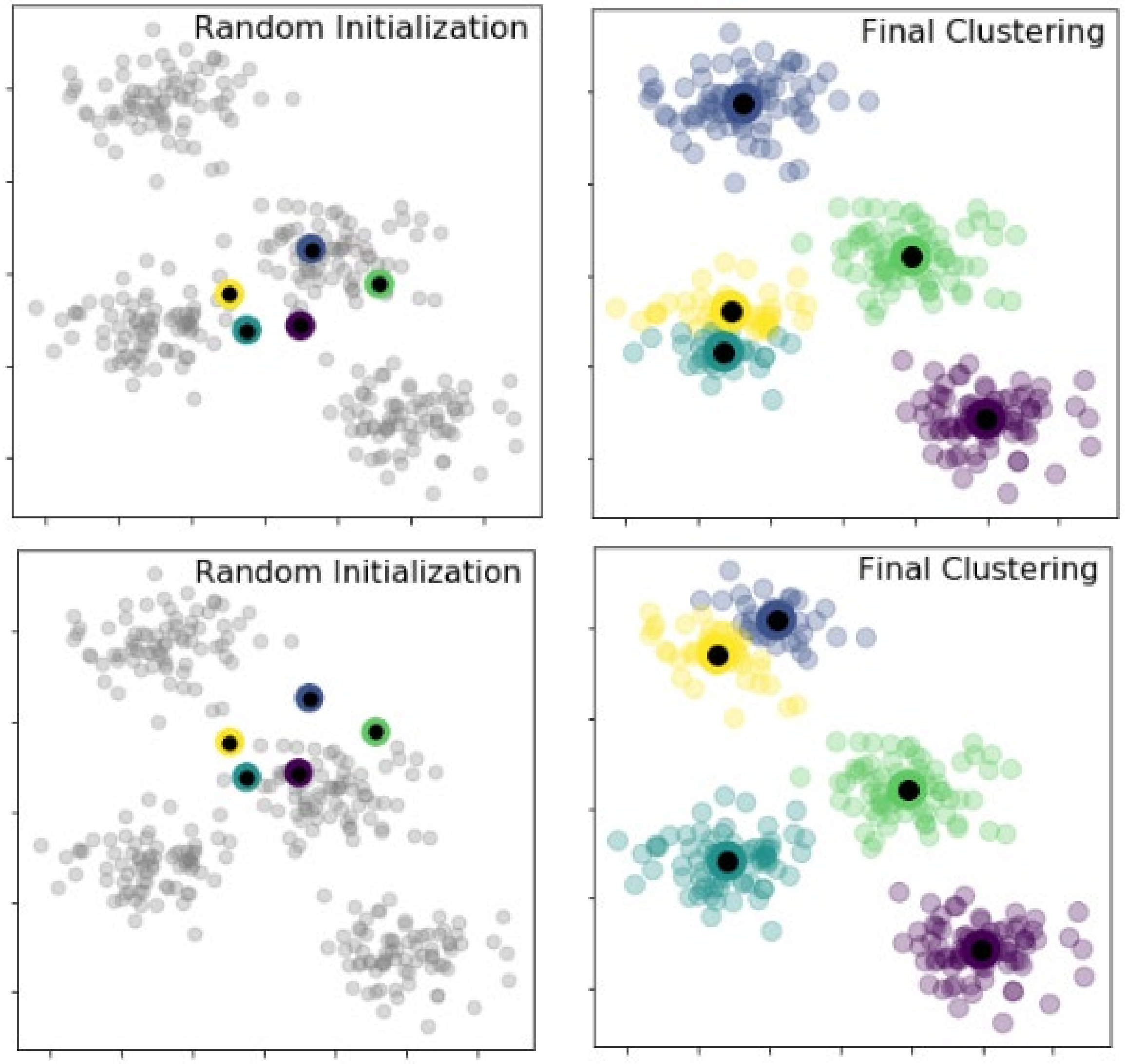


Effect of random initialization

K Means finds a *local* minimum

Different initializations →

Different final clusters



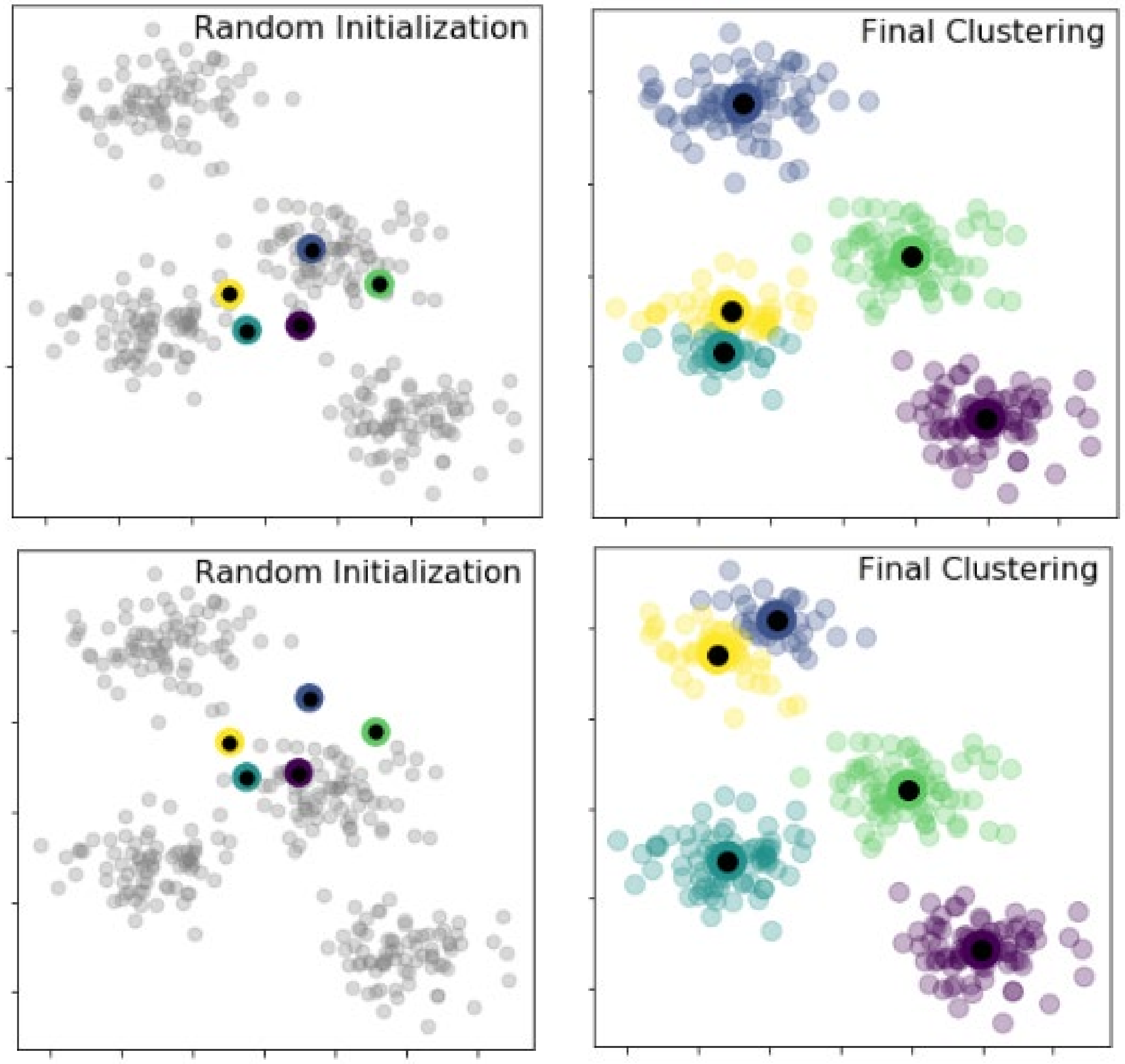
Effect of random initialization

K Means finds a *local* minimum

Different initializations →

Different final clusters

Should run the algorithm multiple times (with different initializations) and pick the best clustering



K-Means: things to consider

- Algorithm converges to local solutions

K-Means: things to consider

- Algorithm converges to local solutions
- Everything that applies to kNN distance issues
    - Scaling
    - Categorical variables
    - High dimensional data

K-Means: things to consider

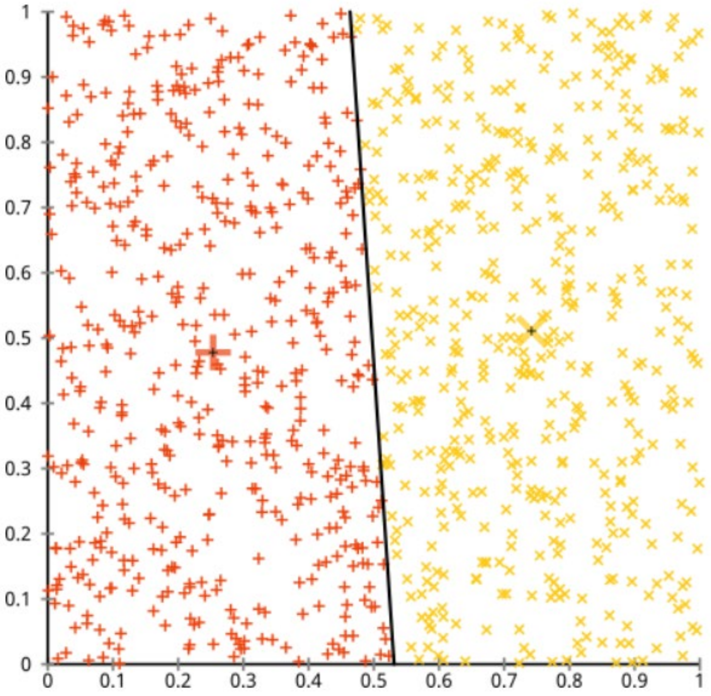
- Algorithm converges to local solutions
- Everything that applies to kNN distance issues
  - Scaling
  - Categorical variables
  - High dimensional data
- If the data has no clusters, K-Means still “finds” K clusters
  - (special case of wrong choice for K)

K-Means: things to consider

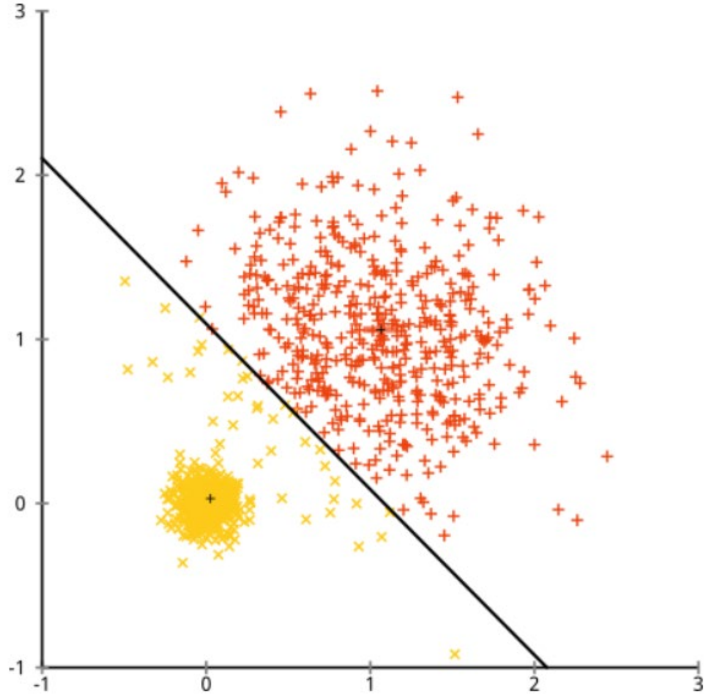
- Algorithm converges to local solutions
- Everything that applies to kNN distance issues
  - Scaling
  - Categorical variables
  - High dimensional data
- If the data has no clusters, K-Means still “finds” K clusters
  - (special case of wrong choice for K)
- K-Means learns linear separation between clusters, will not handle more complex geometry

K-Means: things to consider

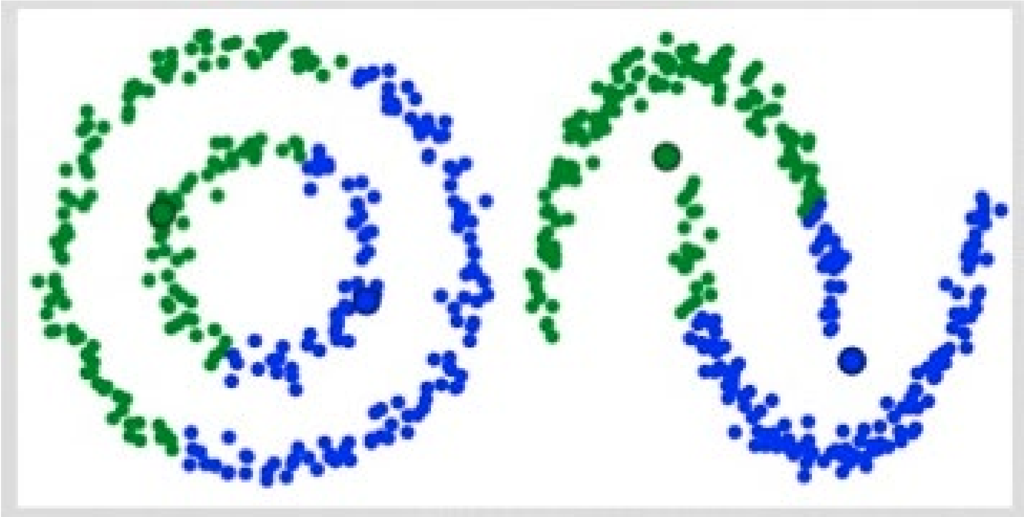
- Algorithm converges to local solutions
- Everything that applies to kNN distance issues
  - Scaling
  - Categorical variables
  - High dimensional data
- If the data has no clusters, K-Means still “finds” K clusters
  - (special case of wrong choice for K)
- K-Means learns linear separation between clusters, will not handle more complex geometry
- K-Means assumes variance of the clusters is the same



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1



3
2
1
0
-1
-1
0
1
2
3



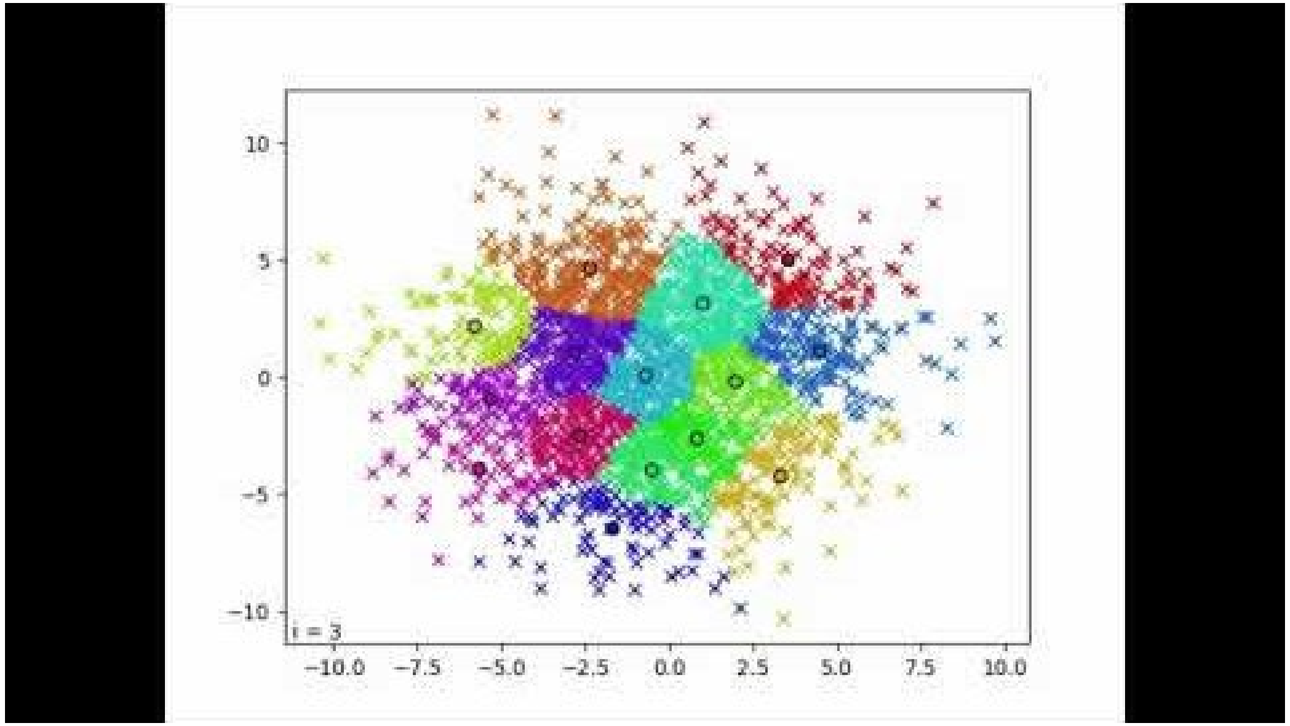
K-Means visualizations

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

https://www.youtube.com/watch?time_continue=98&v=Rrxuq2vNwCM&feature=emb_logo

K-Means visualizations

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



https://www.youtube.com/watch?time_continue=98&v=Rrxuq2vNwCM&feature=emb_logo