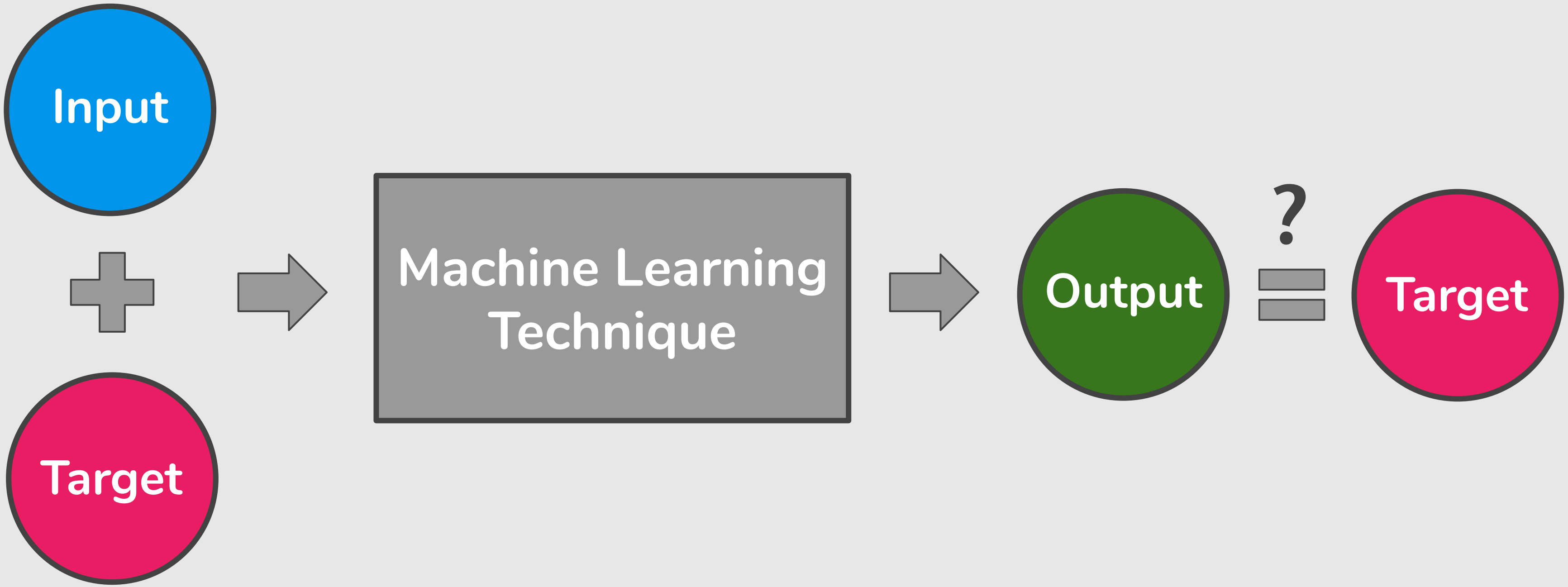
Clustering Algorithms

Machine Learning

Prof. Sandra Avila
Institute of Computing (IC/Unicamp)

MC886, September 23, 2019

Supervised Learning

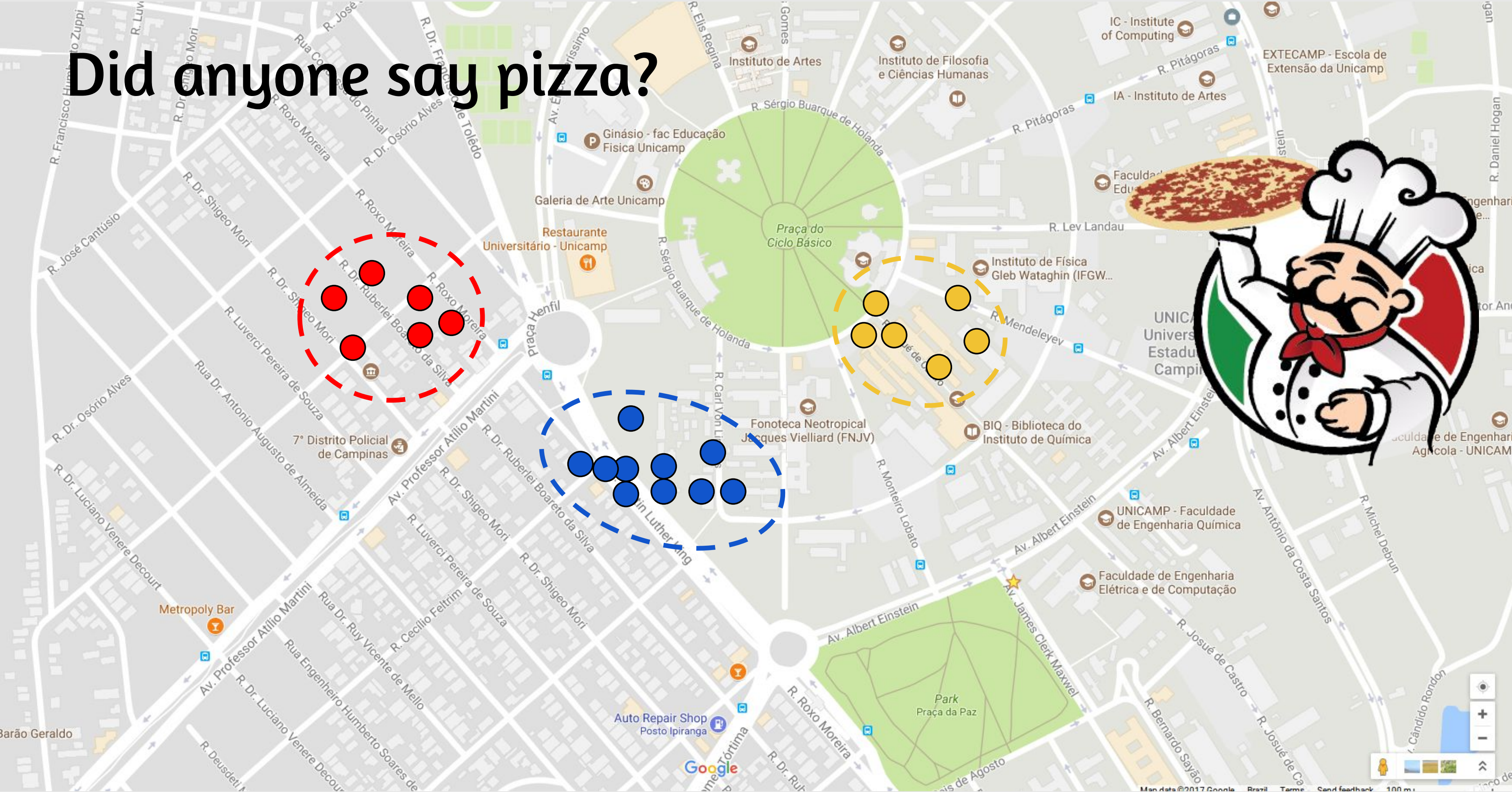


Unsupervised Learning

The goal of unsupervised learning is **to find patterns** in the data, and build new and useful representations of it.

Clustering
k-Means Algorithm

Did anyone say pizza?



k-Means: Image Segmentation

Original | K = 10 | K = 3 | K = 2

Credit: Christopher Bishop

k-Means Algorithm

1. Define the *k* centroids.
2. Find the closest centroid & update cluster assignments.
3. Move the centroids to the center of their clusters.
4. Repeat steps 2 and 3 until the centroid stop moving a lot at each iteration (i.e., until the algorithm converges).

k-Means Algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

$$\min_k ||x^{(i)} - \mu_k||$$

repeat {

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid **closest** to $x^{(i)}$

for k = 1 to K

μ_k := mean of points assigned to cluster k

}

Clustering

Optimization Objective

k-Means Optimization Objective

$c^{(i)}$ = index of cluster (from 1 to K) to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid k

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

k-Means Optimization Objective

$c^{(i)}$ = index of cluster (from 1 to K) to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid k

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

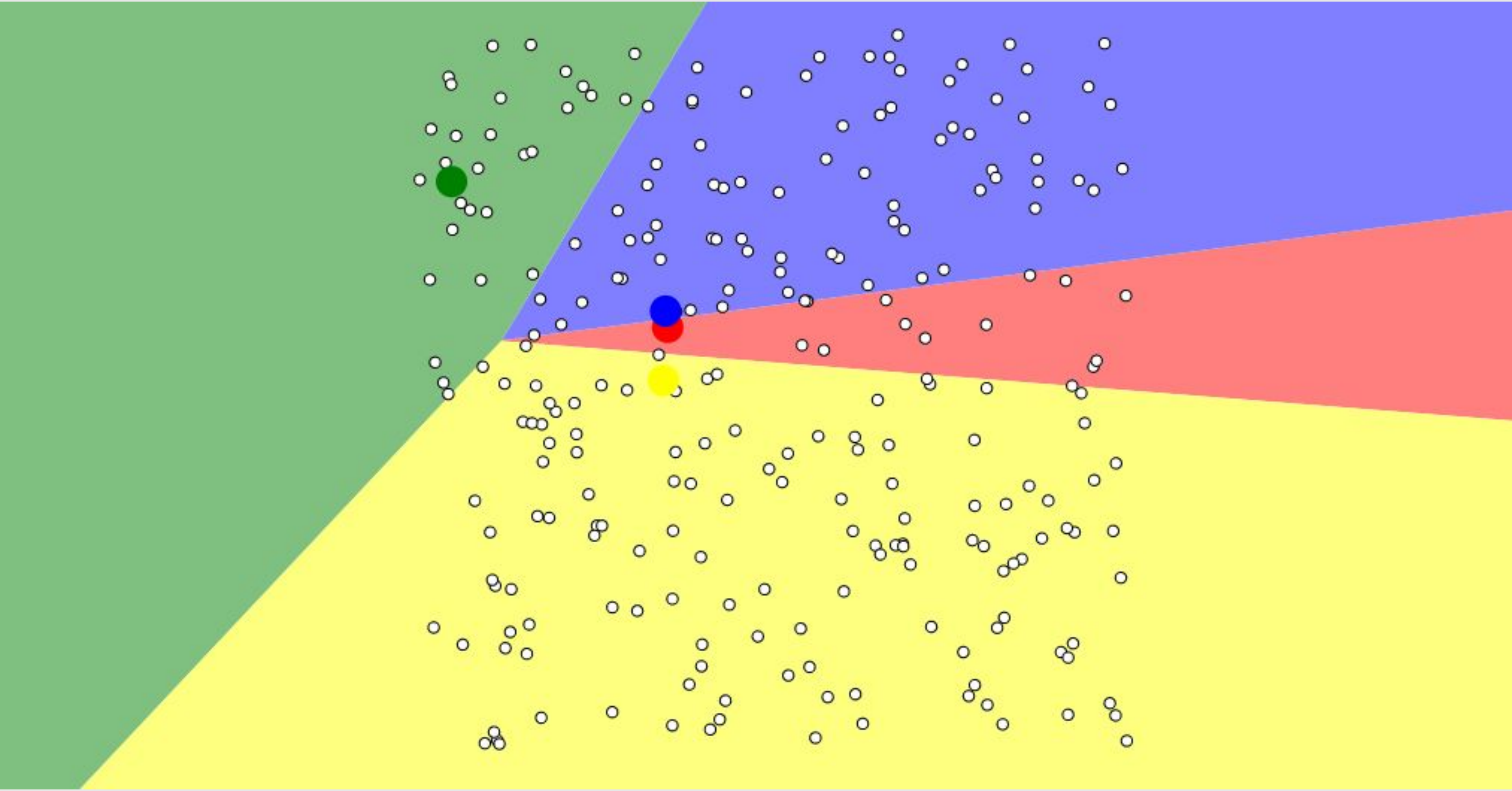
Optimization objective:

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|$$

$$\min_{\substack{c^{(1)}, \ldots, c^{(m)} \\ \mu_1, \ldots, \mu_K}} J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

Clustering

Random Initialization



Random Initialization

for i = 1 to 100 {

Randomly initialize k-Means.

Run k-Means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.

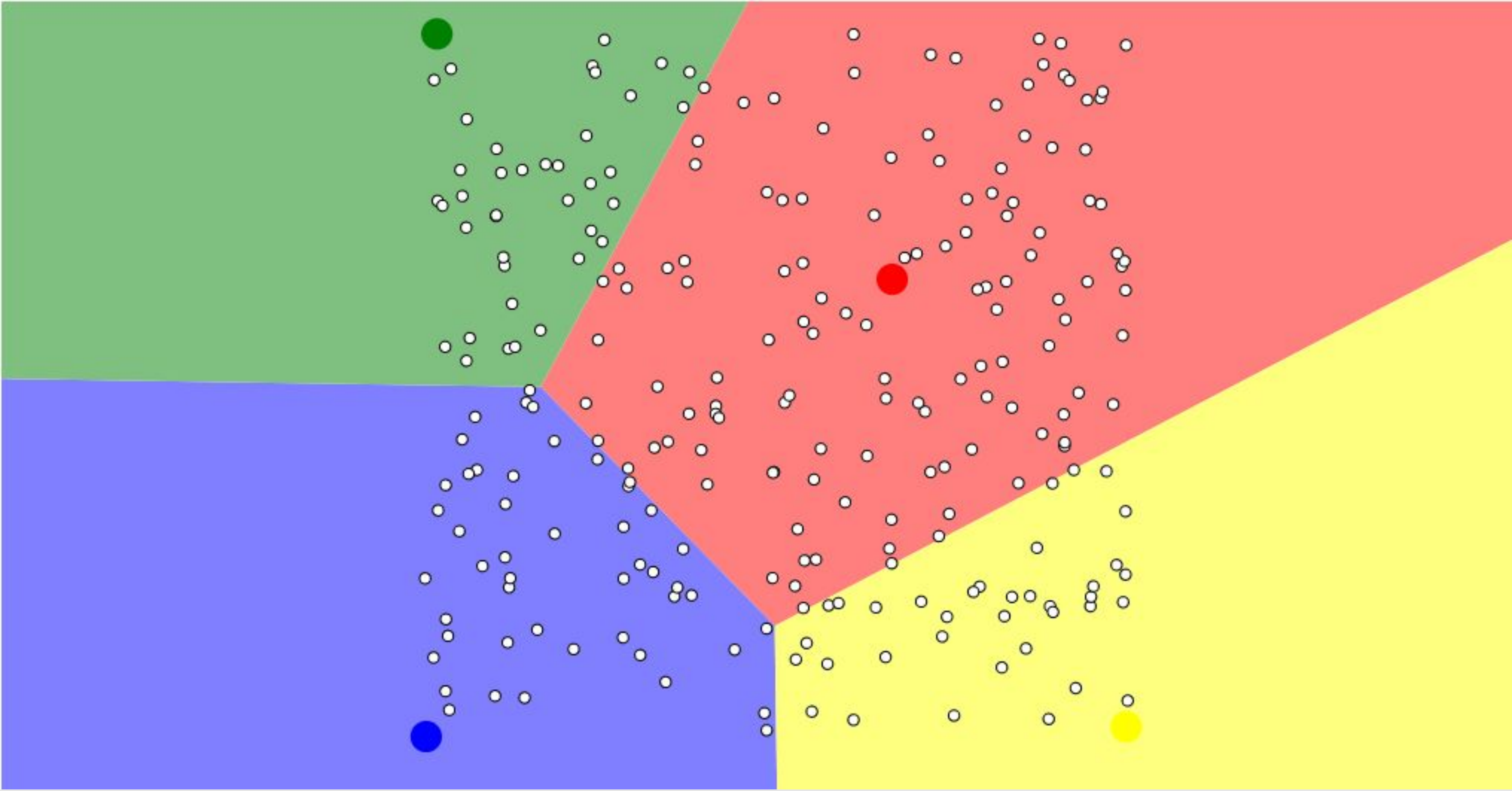
Compute cost function J.

}

Pick clustering that gave lowest cost $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$.

Can we do better?

- One idea for initializing k-Means is to use a farthest-first traversal on the data set, **to pick K points that are far away from each other**.



Can we do better?

- One idea for initializing k-Means is to use a farthest-first traversal on the data set, to pick K points that are far away from each other.

- However, this is **too sensitive to outliers**.

k-Means++ (Arthur & Vassilvitski, 2007)

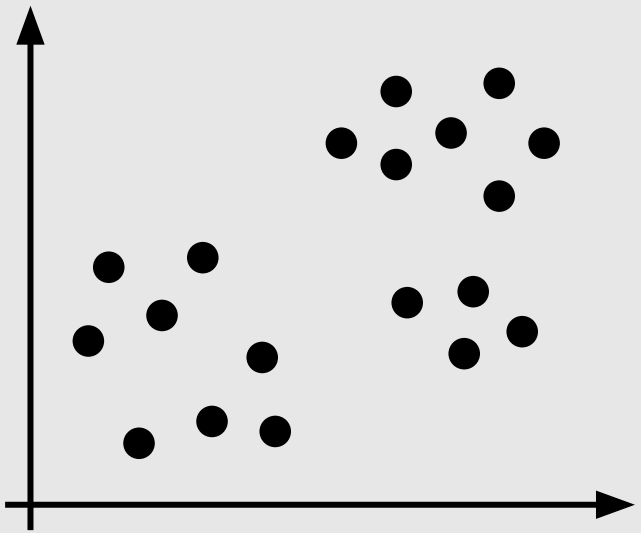
- It works similarly to the “farthest” heuristic.
- Choose each point at random, with probability proportional to its squared distance from the centers chosen already.

scikit-learn (default)

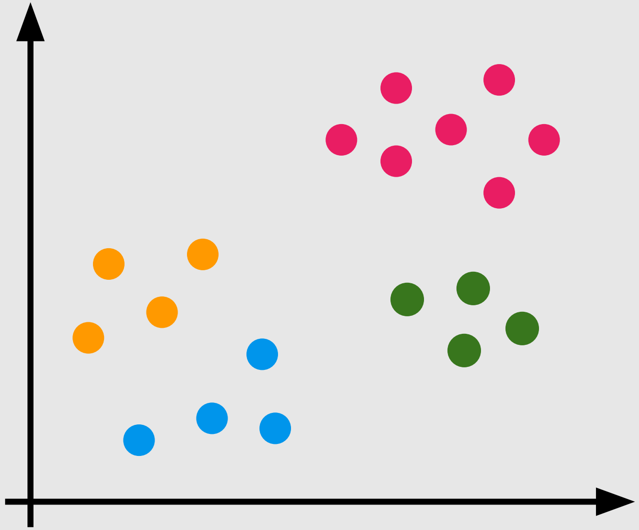
Clustering

Choosing the number of clusters

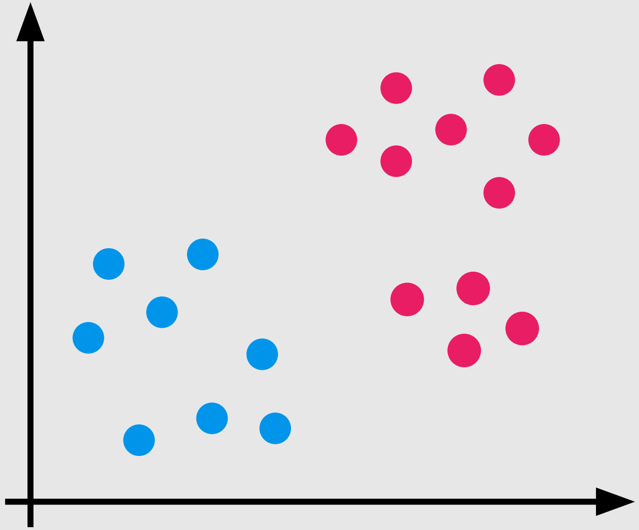
What is the right value of K?



What is the right value of K?

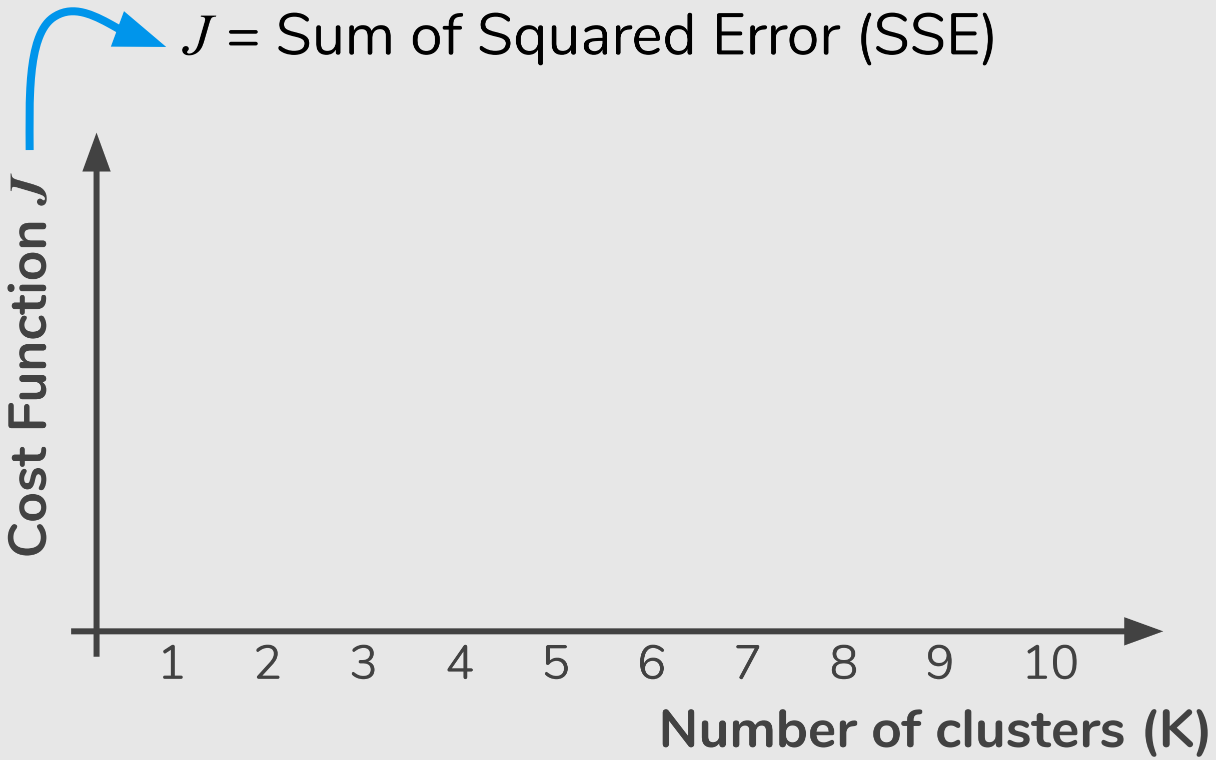


What is the right value of K?

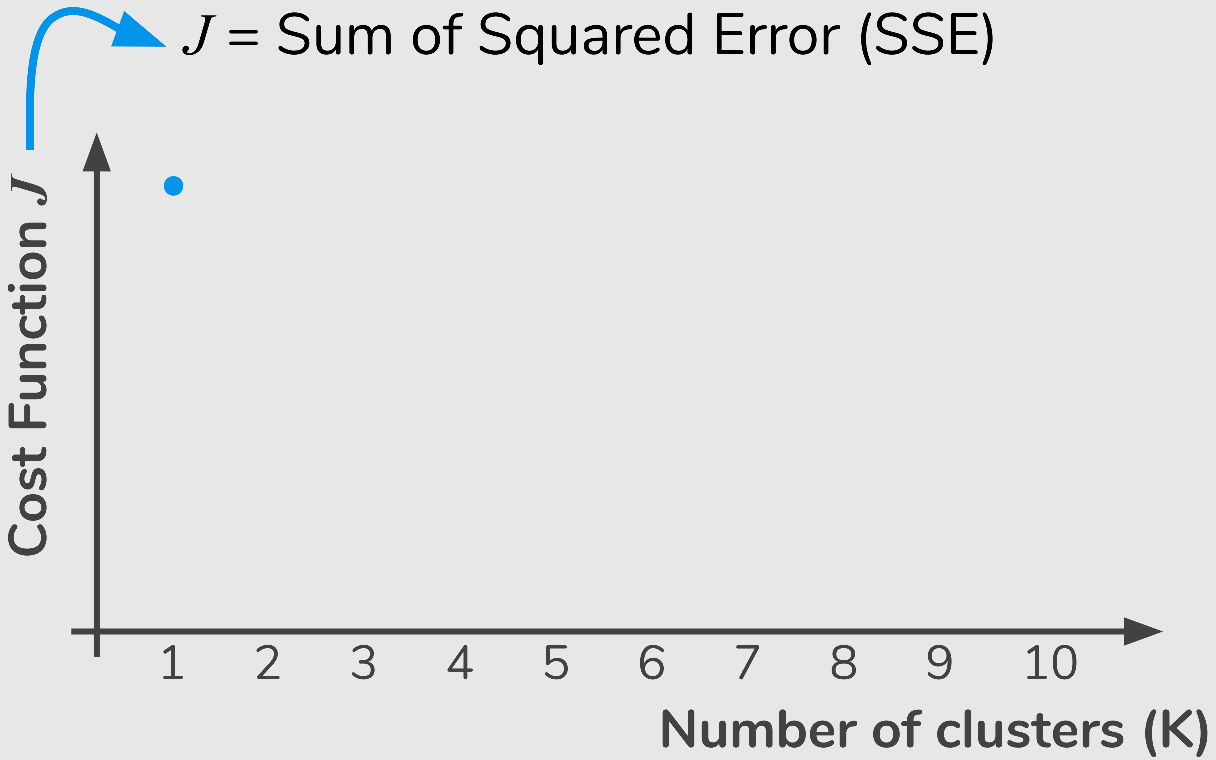


Elbow Method

Elbow Method

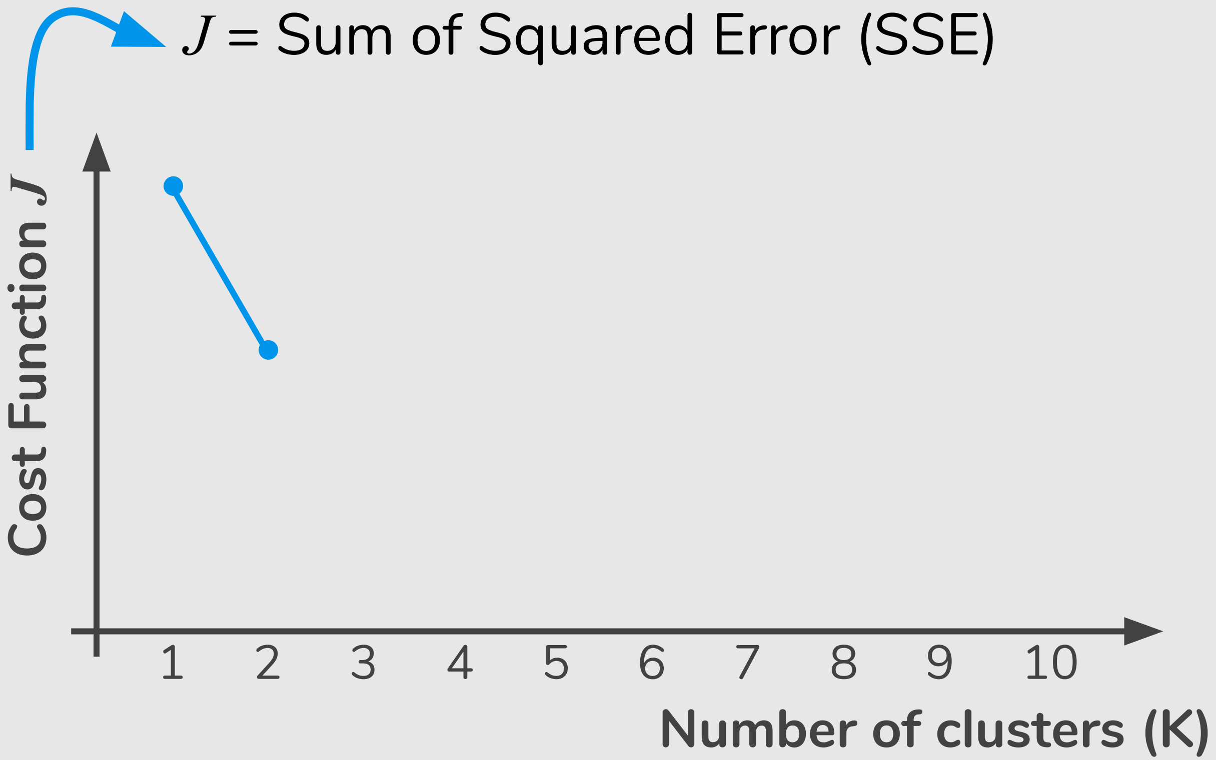


Elbow Method

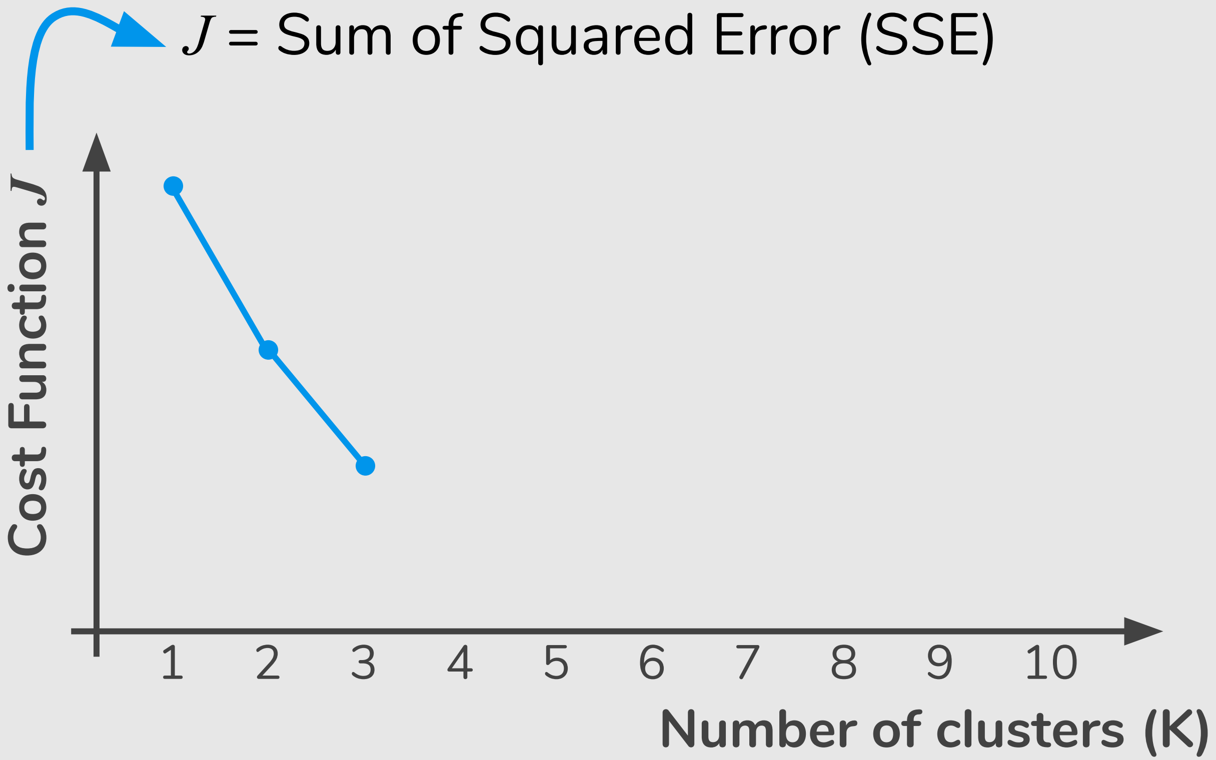


Elbow Method

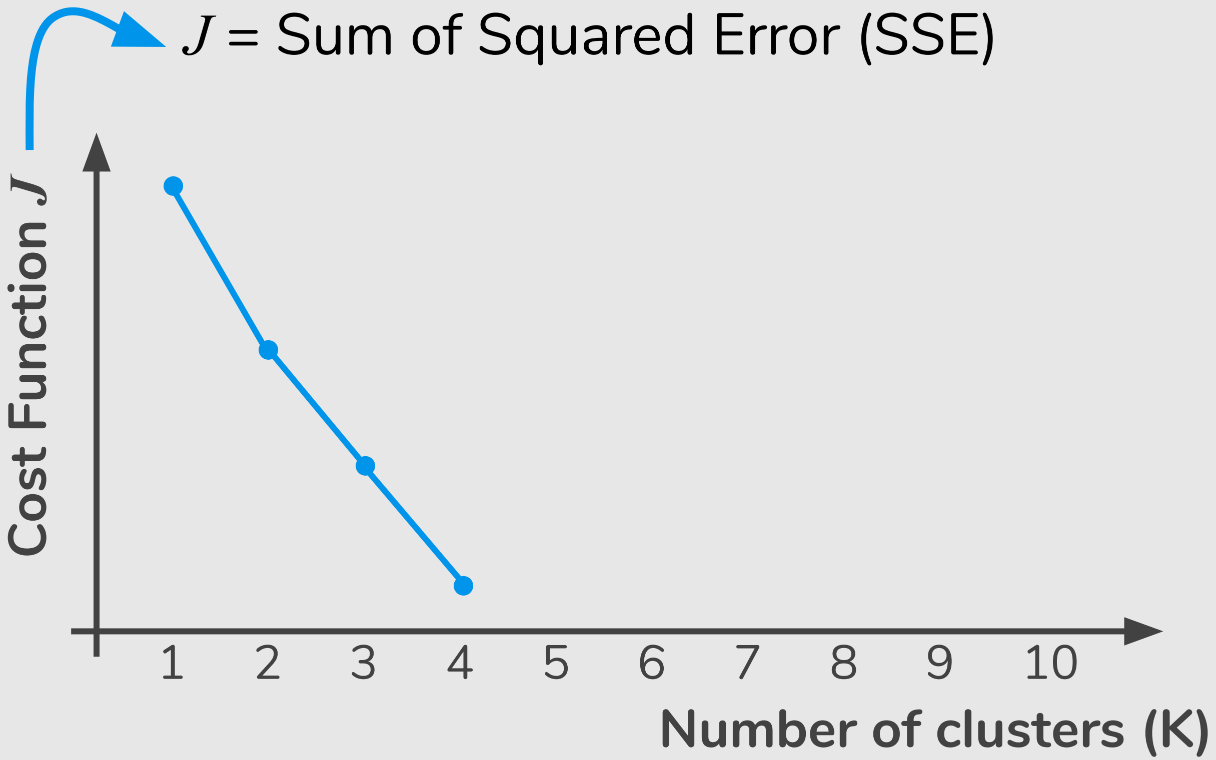
J = Sum of Squared Error (SSE)



Elbow Method

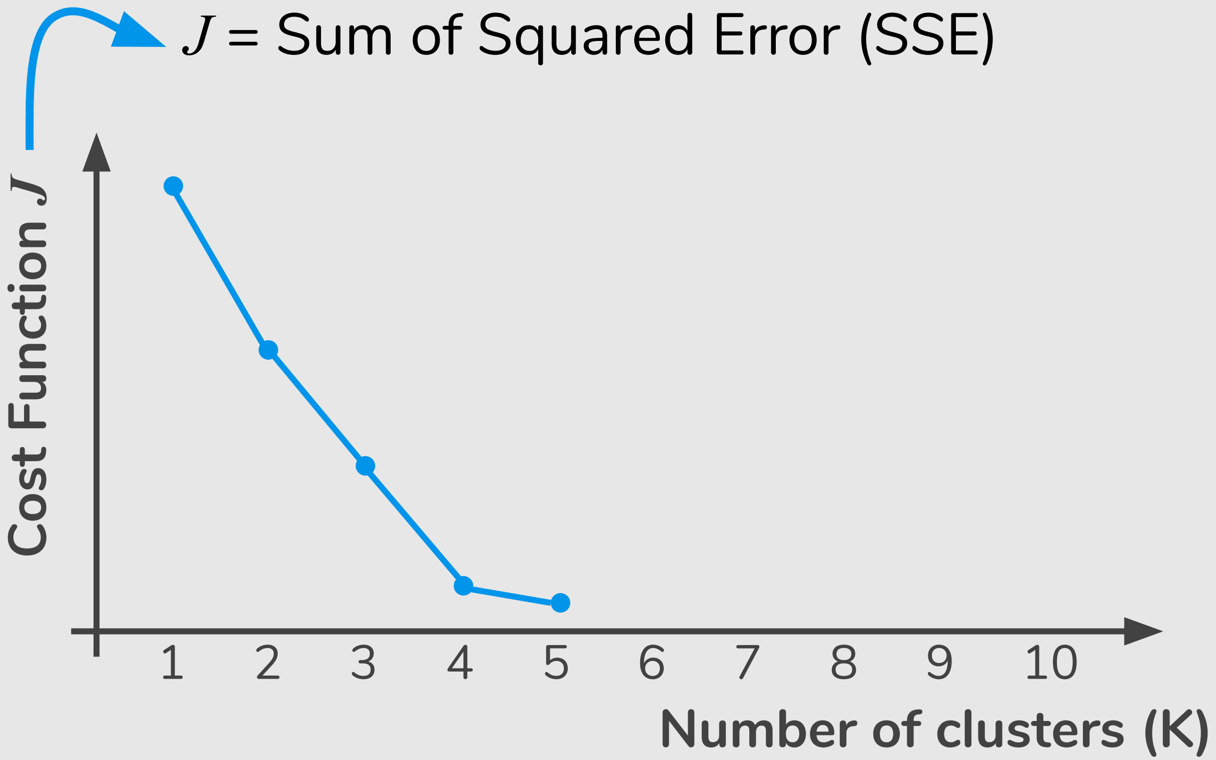


Elbow Method

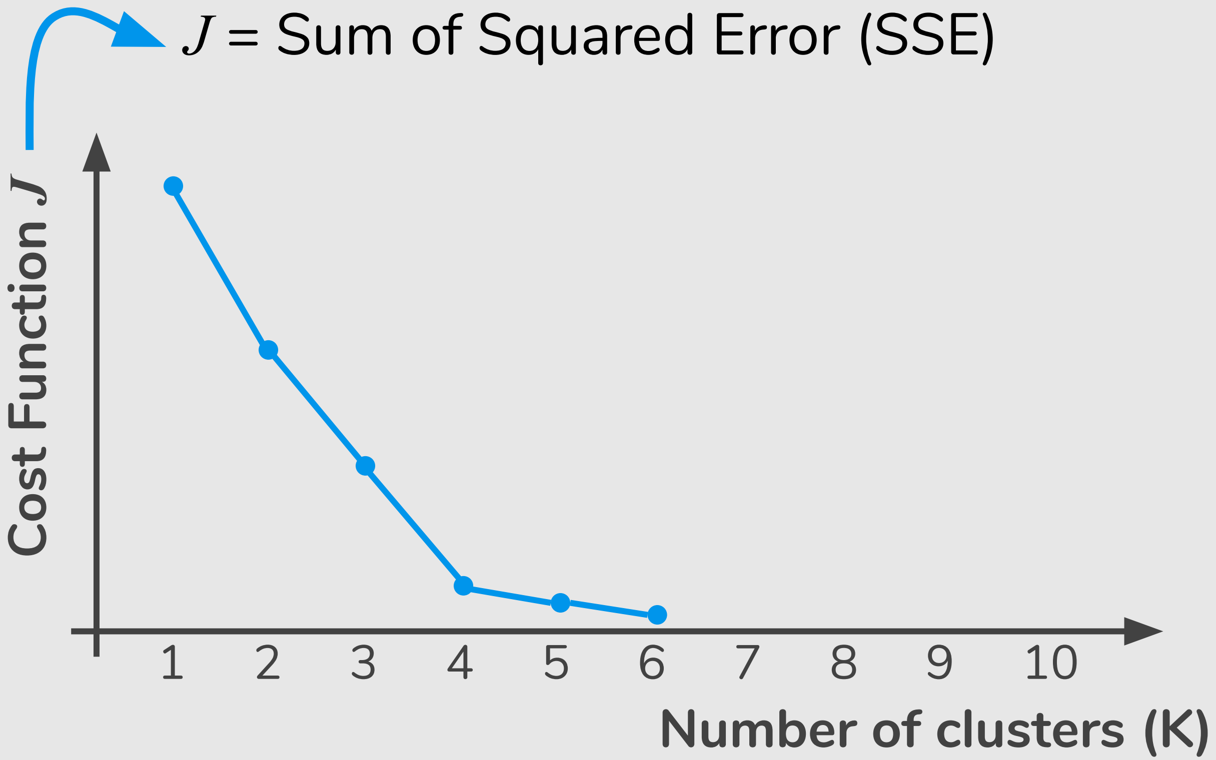


Elbow Method

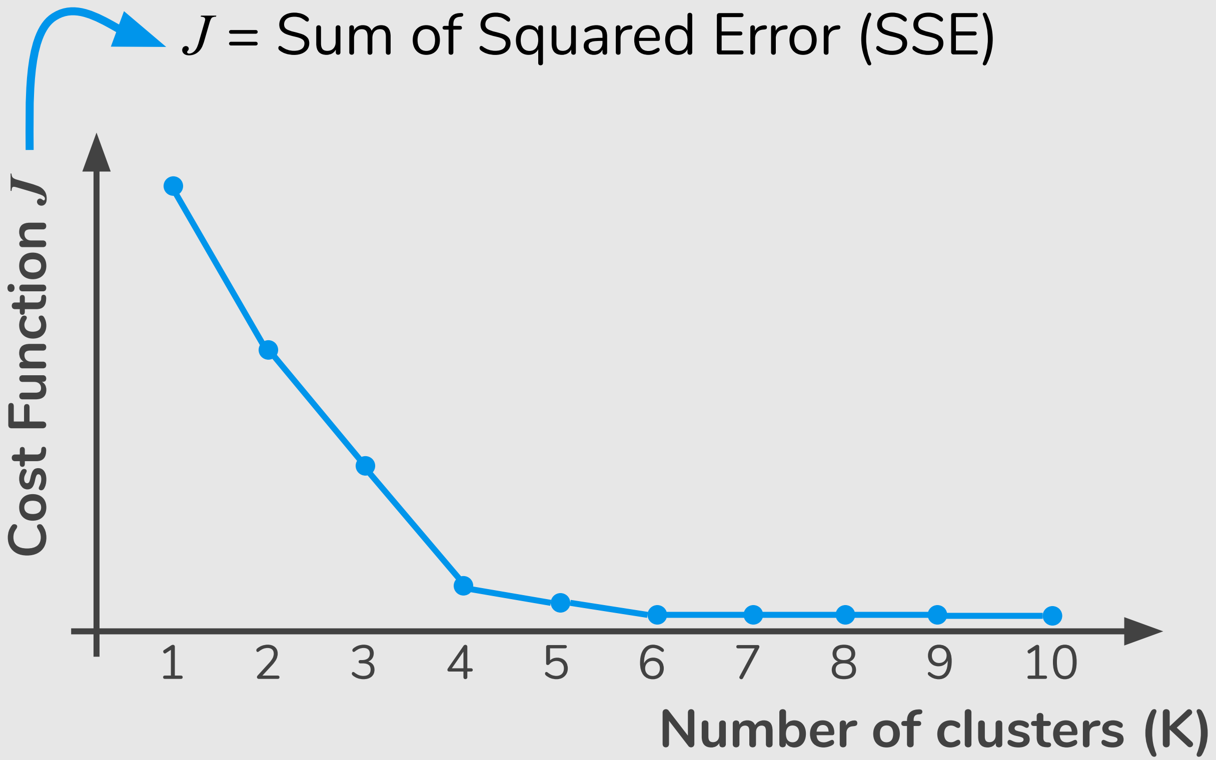
J = Sum of Squared Error (SSE)



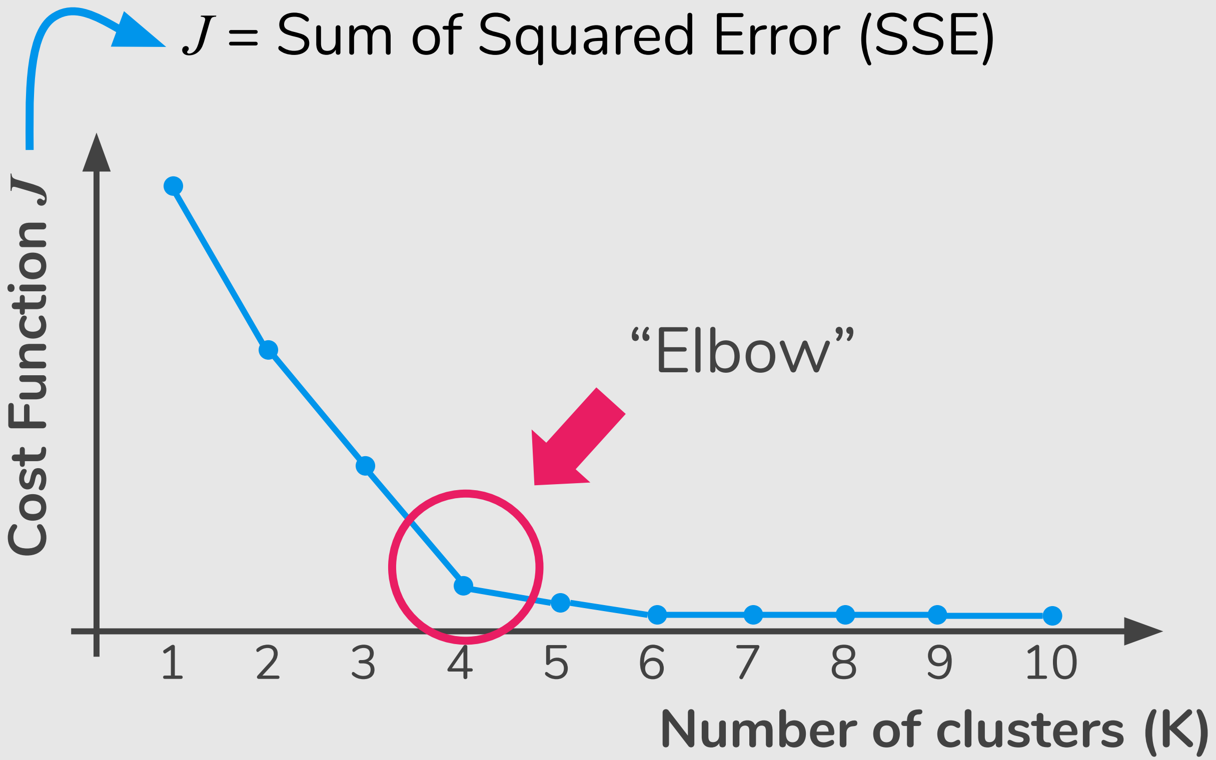
Elbow Method



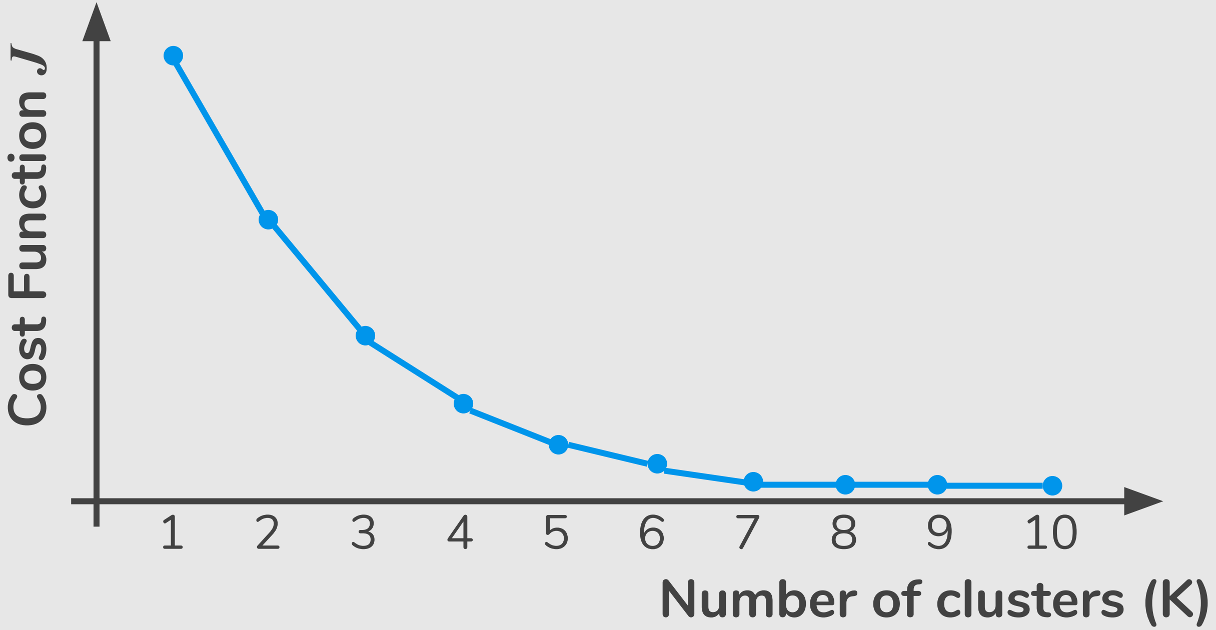
Elbow Method



Elbow Method

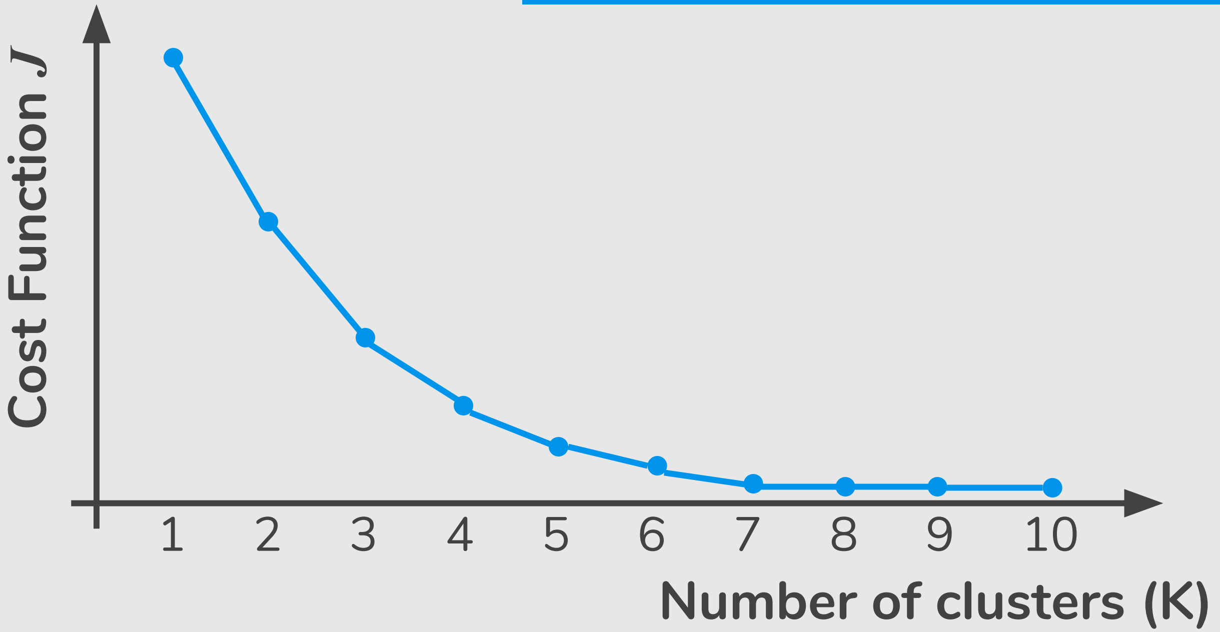


~~Elbow~~ Method

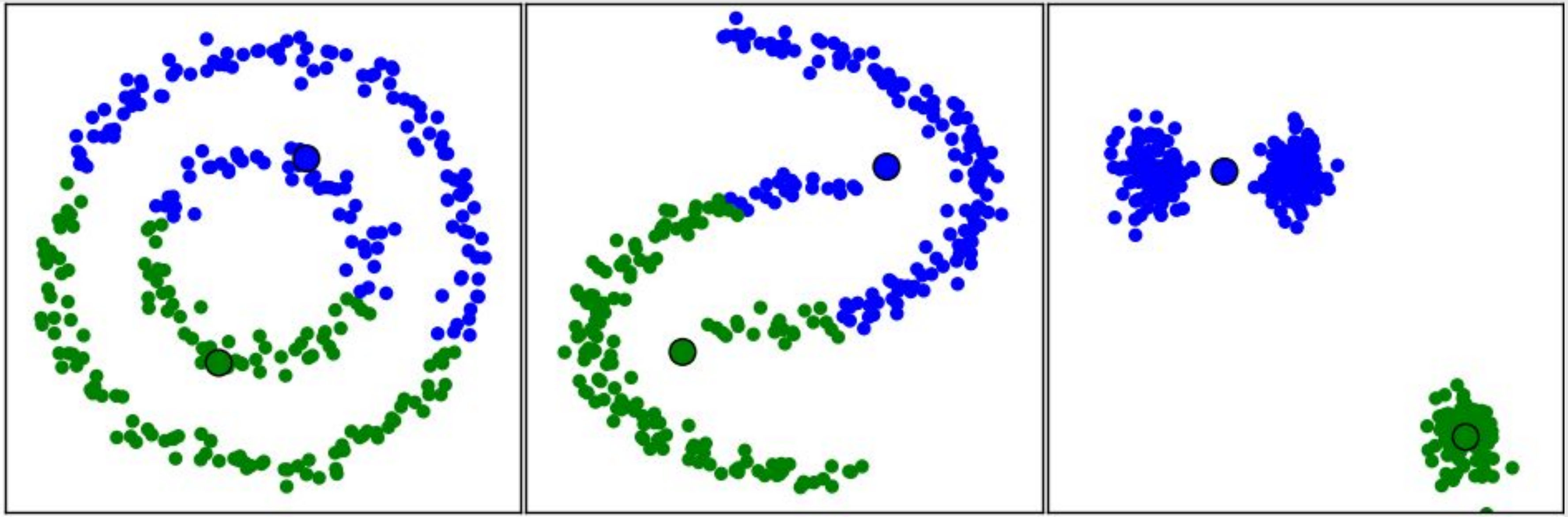


~~Elbow~~ Method

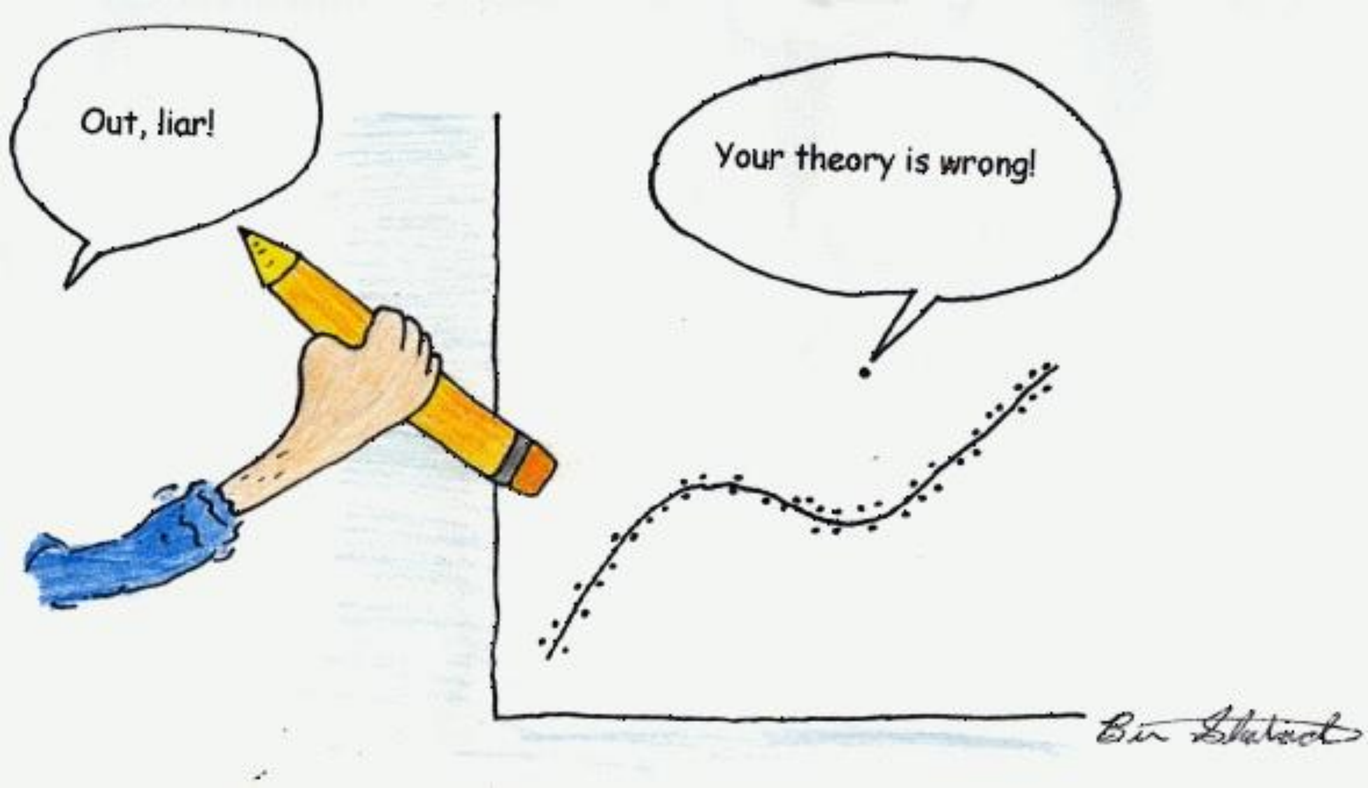
Q: You find that cost function J is much higher for k = 5 than for k =3. What can you conclude?



k-Means: Additional Issues



Outliers

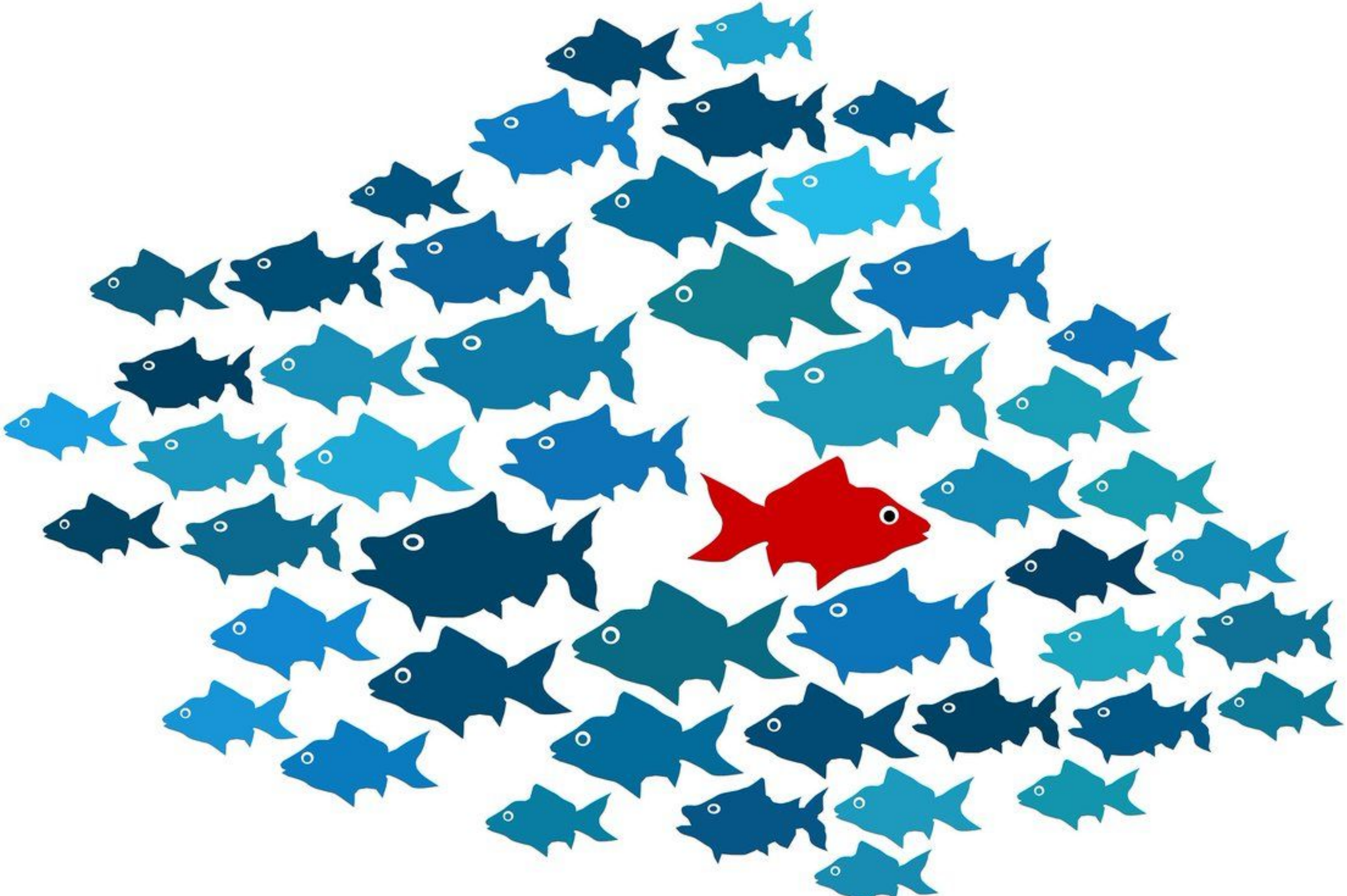


Outliers

- It is often useful to discover outliers and eliminate them before clustering.

Outliers

- It is often useful to discover outliers and eliminate them before clustering.
- Techniques for identifying outlier: “*Anomaly Detection*” *[chap. 9]*, *Introduction to Data Mining*, 2018.



Outliers

- It is often useful to discover outliers and eliminate them before clustering.
- Techniques for identifying outlier: “*Anomaly Detection*” *[chap. 9]*, *Introduction to Data Mining*, 2018.
- Also, we often want to eliminate small clusters because they frequently represent groups of outliers.

Reducing the SSE with Postprocessing

- **Split a cluster**: the cluster with the largest SSE is usually chosen.

Reducing the SSE with Postprocessing

- **Split a cluster**: the cluster with the largest SSE is usually chosen.
- **Introduce a new cluster centroid**: often the point that is farthest from any cluster center is chosen.

Reducing the SSE with Postprocessing

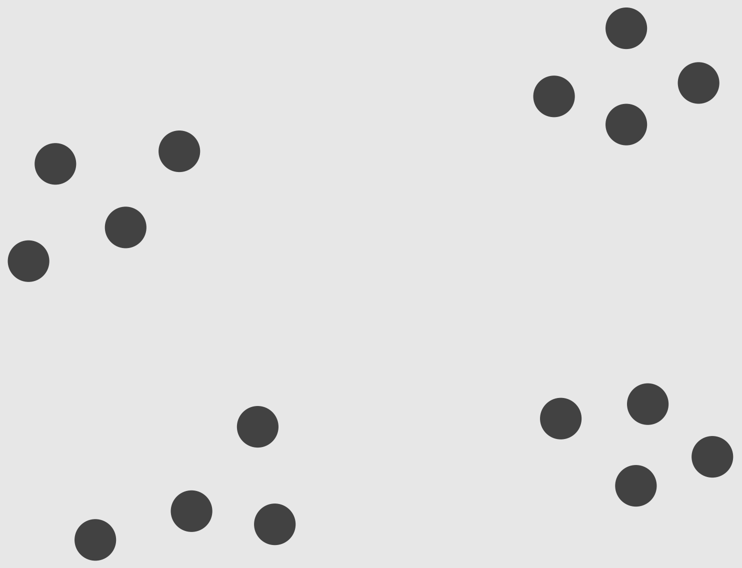
- **Split a cluster**: the cluster with the largest SSE is usually chosen.
- **Introduce a new cluster centroid**: often the point that is farthest from any cluster center is chosen.
- **Merge two clusters**: The clusters with the closest centroids are typically chosen.

k-Means Variations

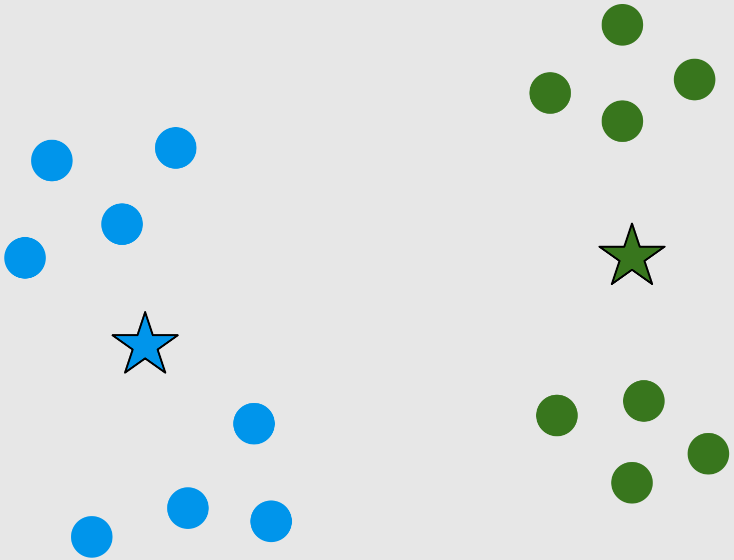
Bisecting k-Means

- A straightforward extension of the basic k-means.
- To obtain k clusters:
  - Split the set of all points into two clusters,
  - Select one of these clusters to split,
  - Repeat until k clusters have been produced.

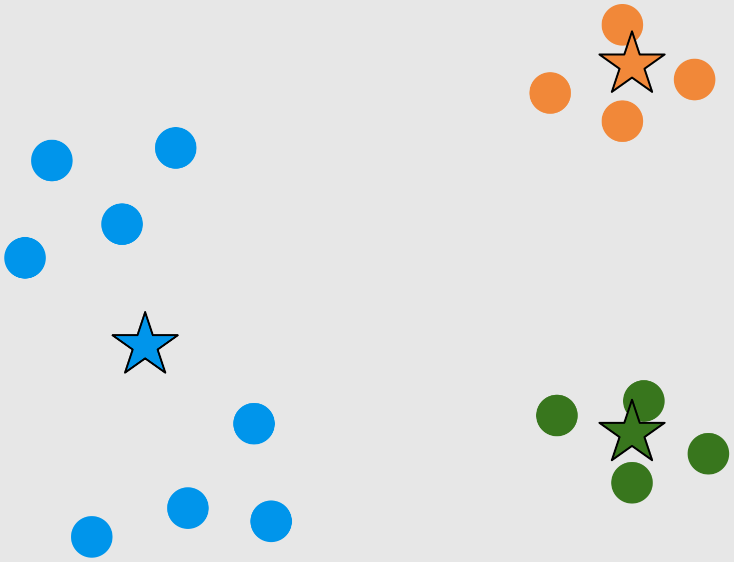
Bisecting k-Means



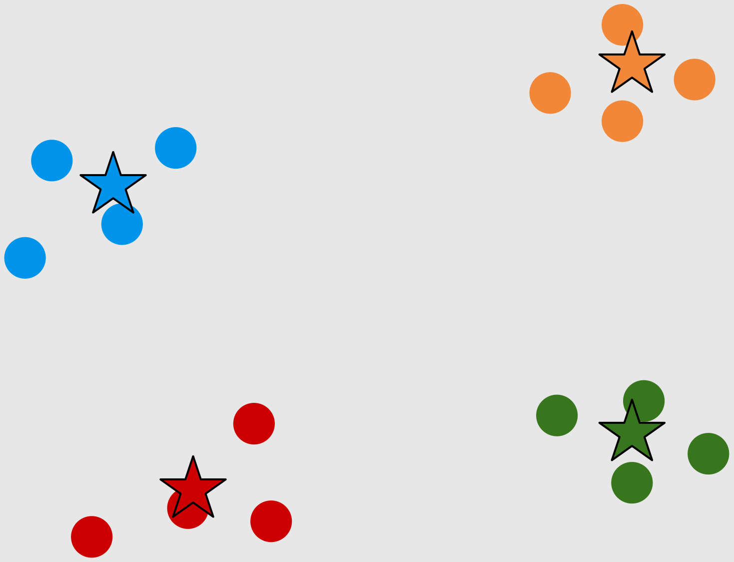
Bisecting k-Means



Bisecting k-Means



Bisecting k-Means



Mini-batch k-Means

- Uses mini-batches to reduce the computation time, while still attempting to optimize the same objective function.

- Converges faster than k-Means, but the quality of the results is reduced.

k-Medians Clustering

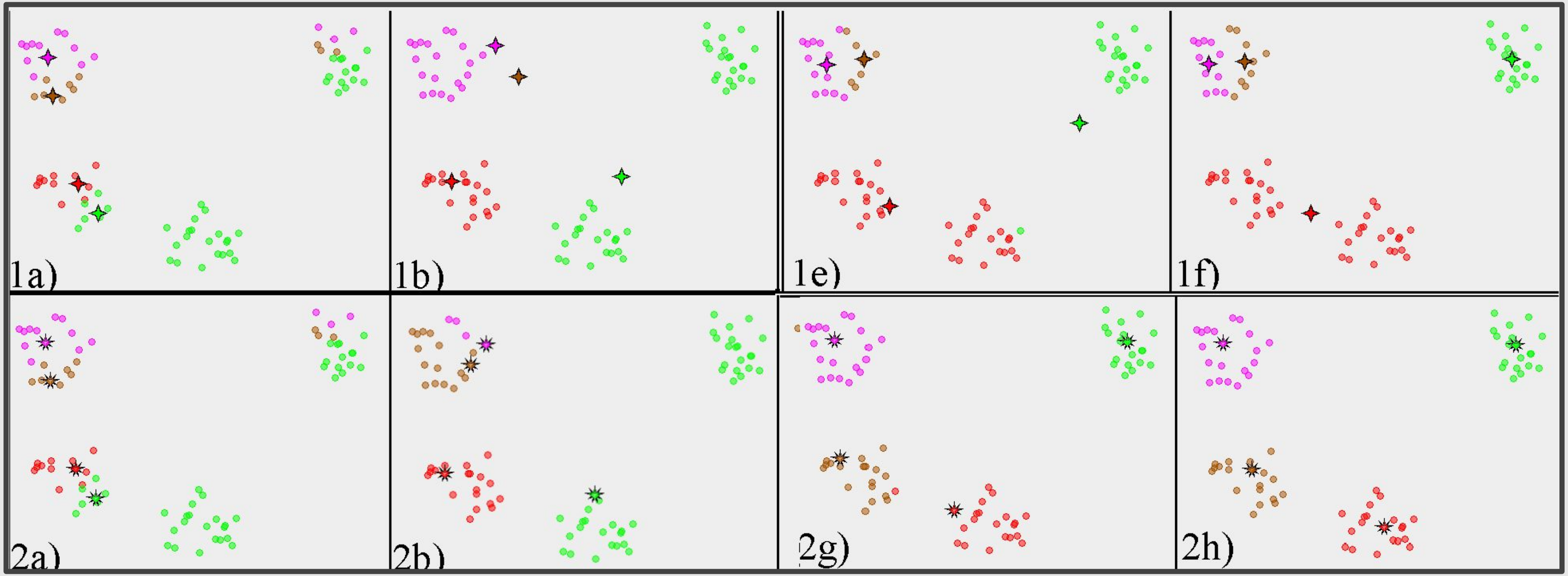
- Instead of calculating the mean for each cluster to determine its centroid, one instead **calculates the median**.

- Minimizing error over all clusters with respect to the **1-norm distance metric**, as opposed to the square of the 2-norm distance metric (which k-Means does).

k-Medoids Clustering

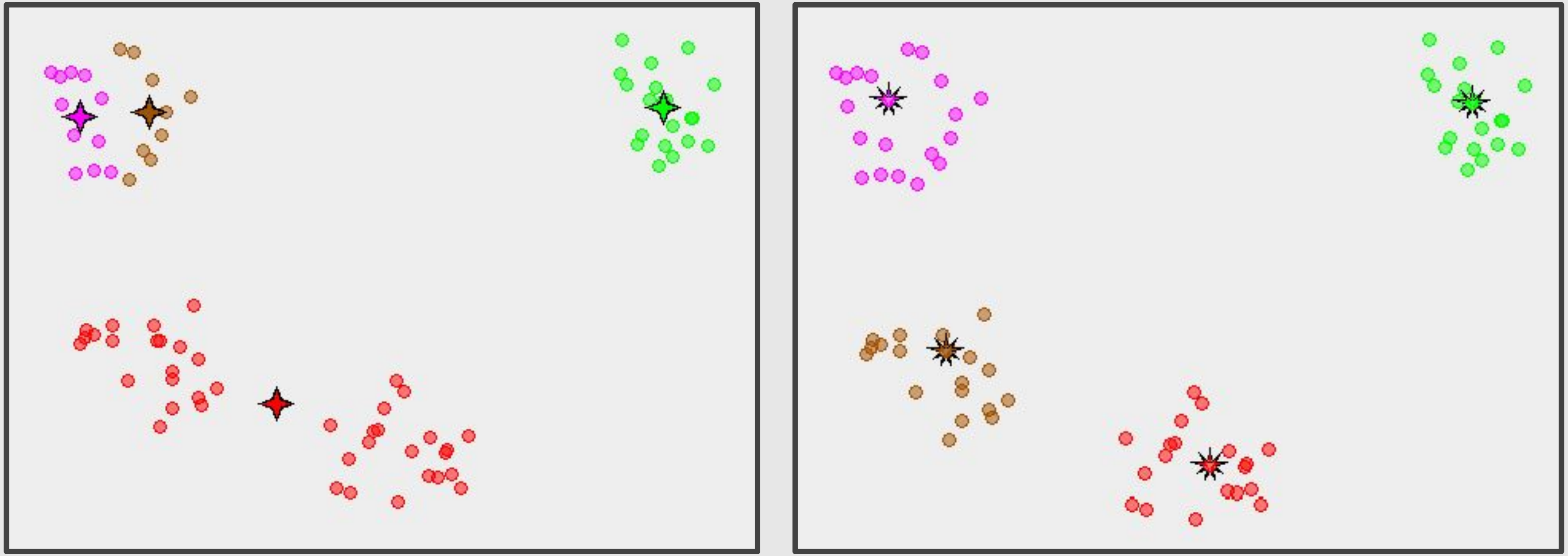
- Instead of calculating the mean for each cluster to determine its centroid, one instead **calculates the medoid**.
- Minimizing error over all clusters with respect to the **1-norm distance metric**.
- In contrast to the k-Means, k-Medoids **chooses data points as centroids**.

k-Means (top) vs k-Medoids (bottom)



Credit: https://commons.wikimedia.org/wiki/File:K-means_versus_k-medoids.png

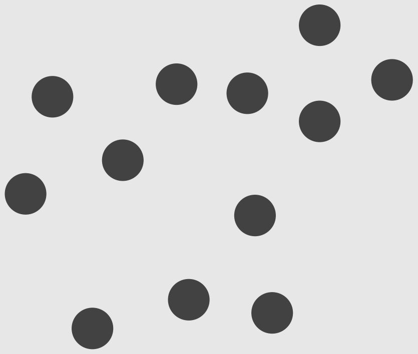
k-Means (left) vs k-Medoids (right)



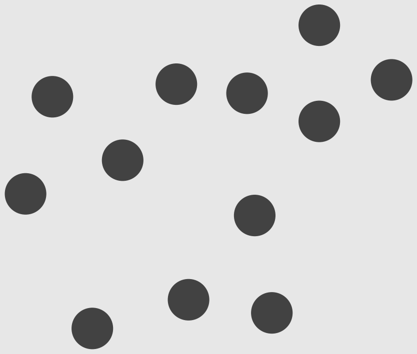
Credit: https://commons.wikimedia.org/wiki/File:K-means_versus_k-medoids.png

Fuzzy Clustering (Soft Clustering)

- Each data point can belong to more than one cluster.



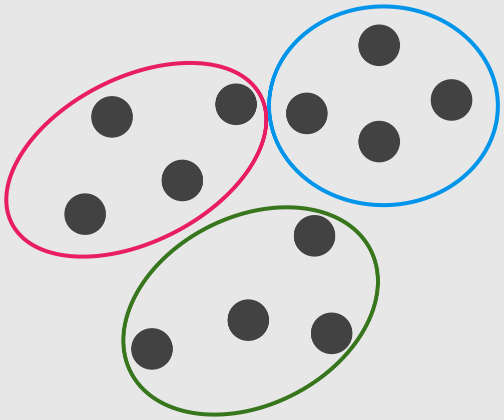
Hard clustering



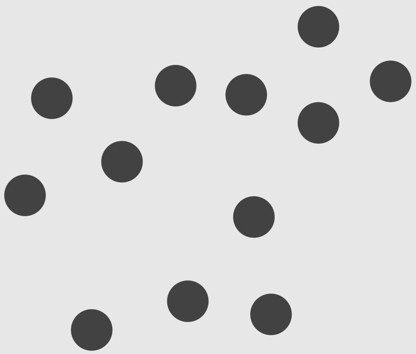
Soft clustering

Fuzzy Clustering (Soft Clustering)

- Each data point can belong to more than one cluster.



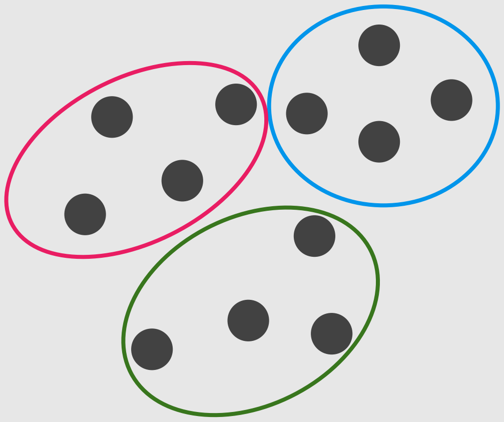
Hard clustering



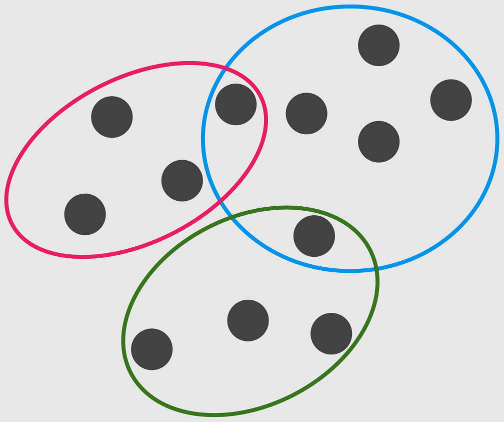
Soft clustering

Fuzzy Clustering (Soft Clustering)

- Each data point can belong to more than one cluster.

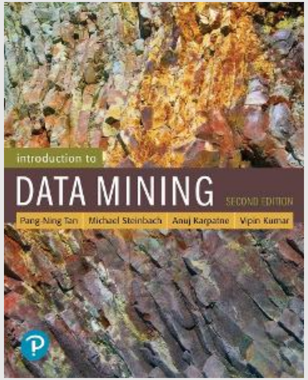


Hard clustering



Soft clustering

References



Machine Learning Books

- Pattern Recognition and Machine Learning, Chap. 9 “Mixture Models and EM”
- Pattern Classification, Chap. 10 “Unsupervised Learning and Clustering”
- “Introduction to Data Mining”, https://www-users.cs.umn.edu/~kumar001/dmbook/ch7_clustering.pdf

Machine Learning Courses

- https://www.coursera.org/learn/machine-learning, Week 8