CS 129.18

K-Means Clustering

Unsupervised Learning

Class of pattern recognition used to draw inferences from data sets consisting of input data without labeled responses.

The most common unsupervised learning method is **cluster analysis,** which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are modeled using a similarity score which is defined upon metrics such as Euclidean or probabilistic distance.

K-Means Clustering

Finding *k* groups in data that occur organically

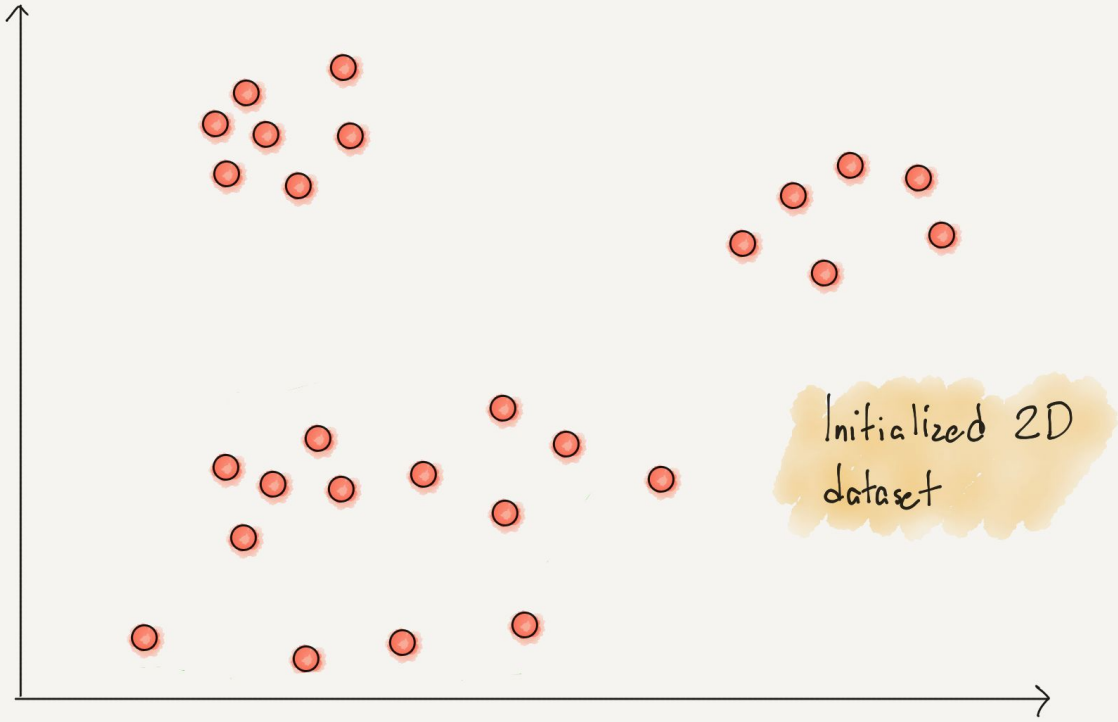
1. Data Assignment

Assign cluster membership for points

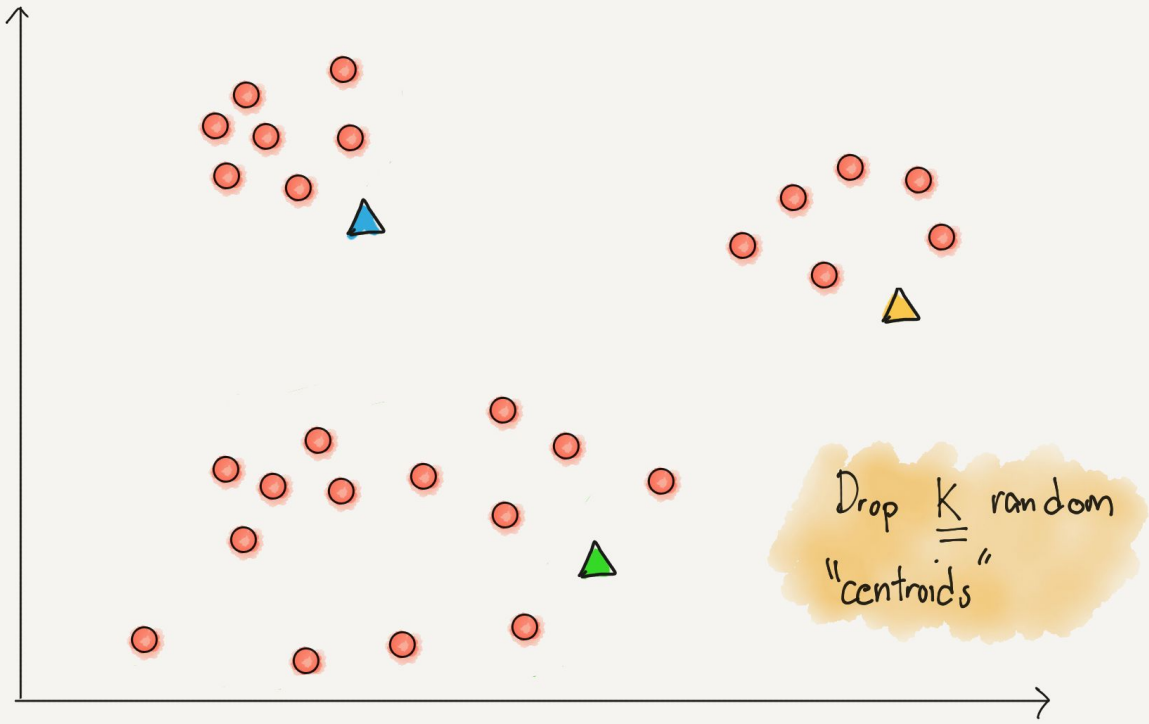
2. Centroid Update

Adjust the centroid to the arithmetic mean location

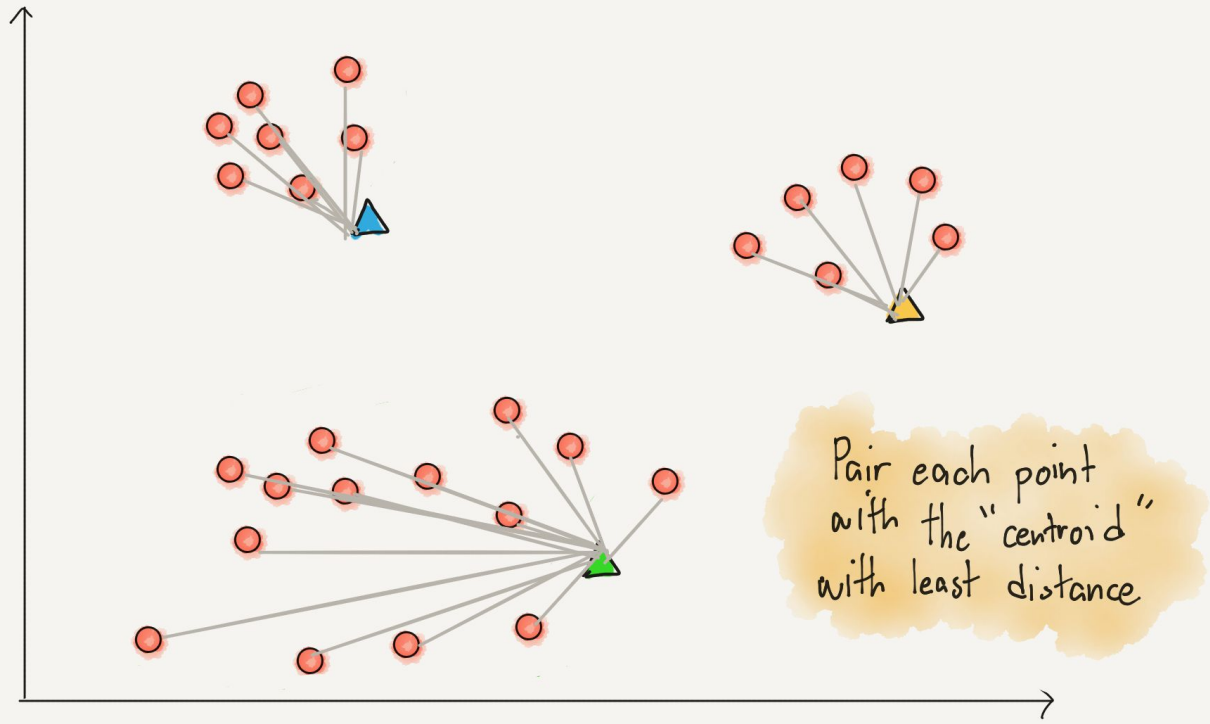
K-Means Clustering



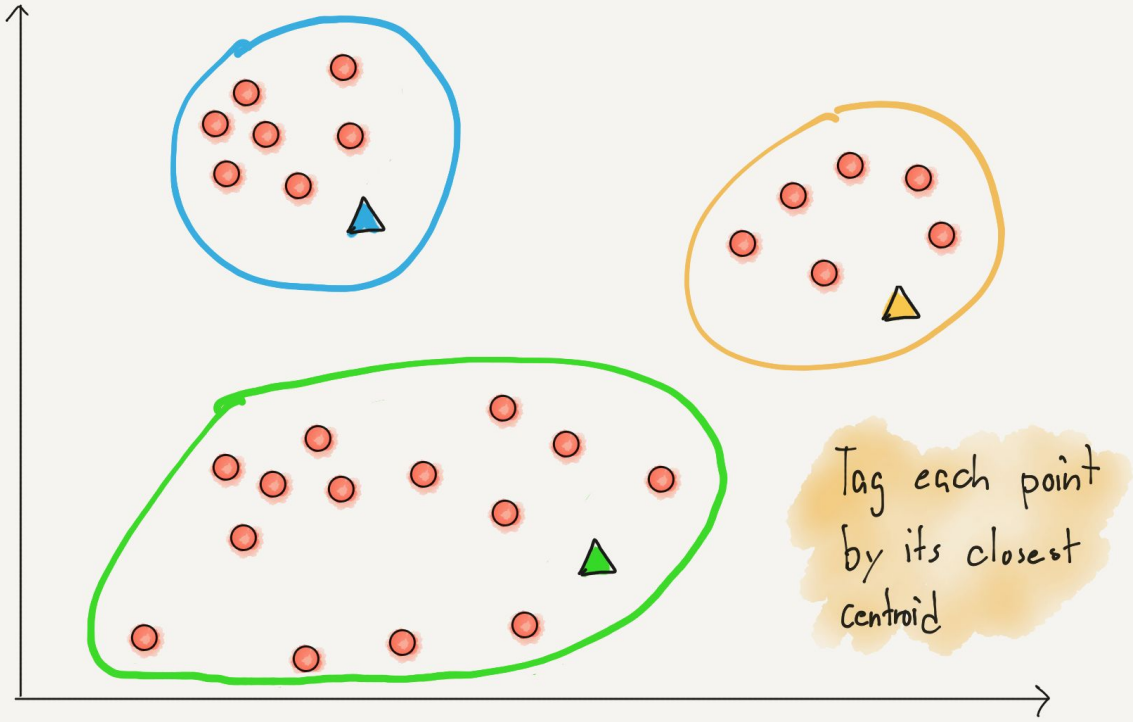
K-Means Clustering



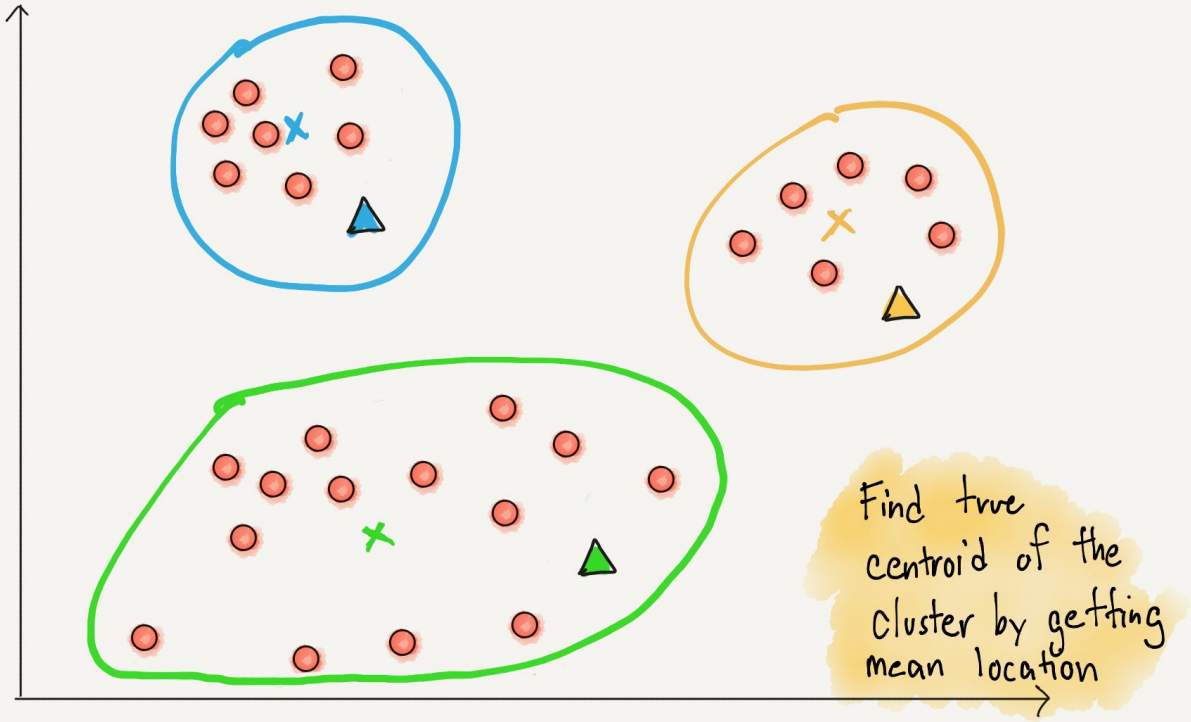
K-Means Clustering



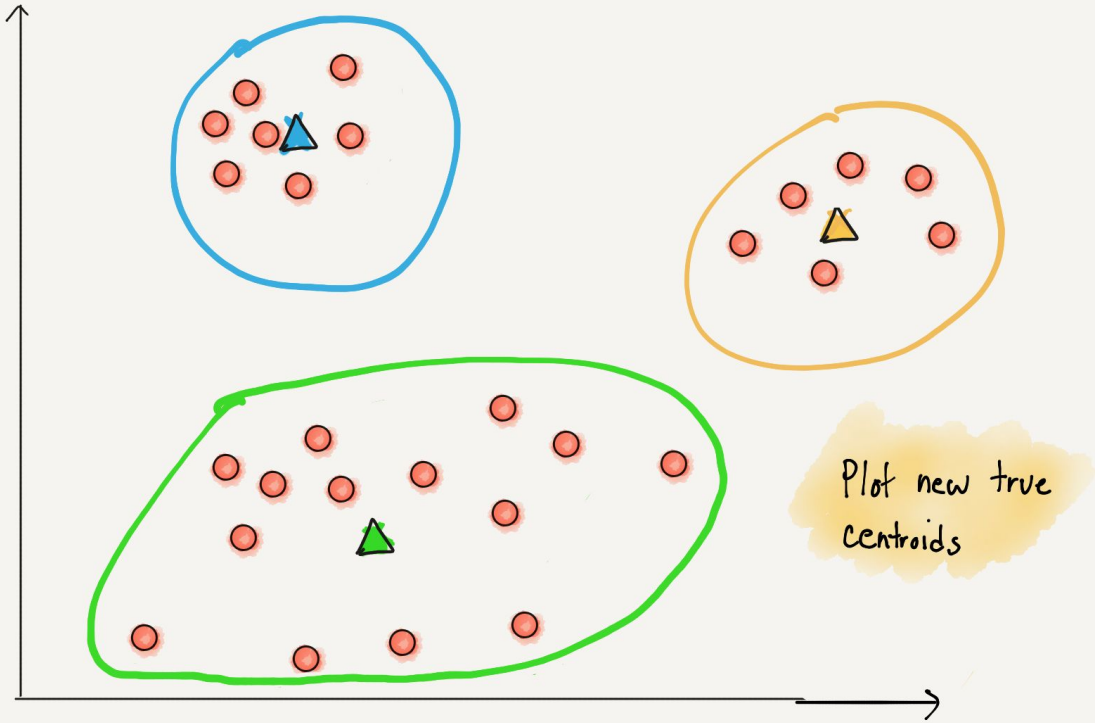
K-Means Clustering



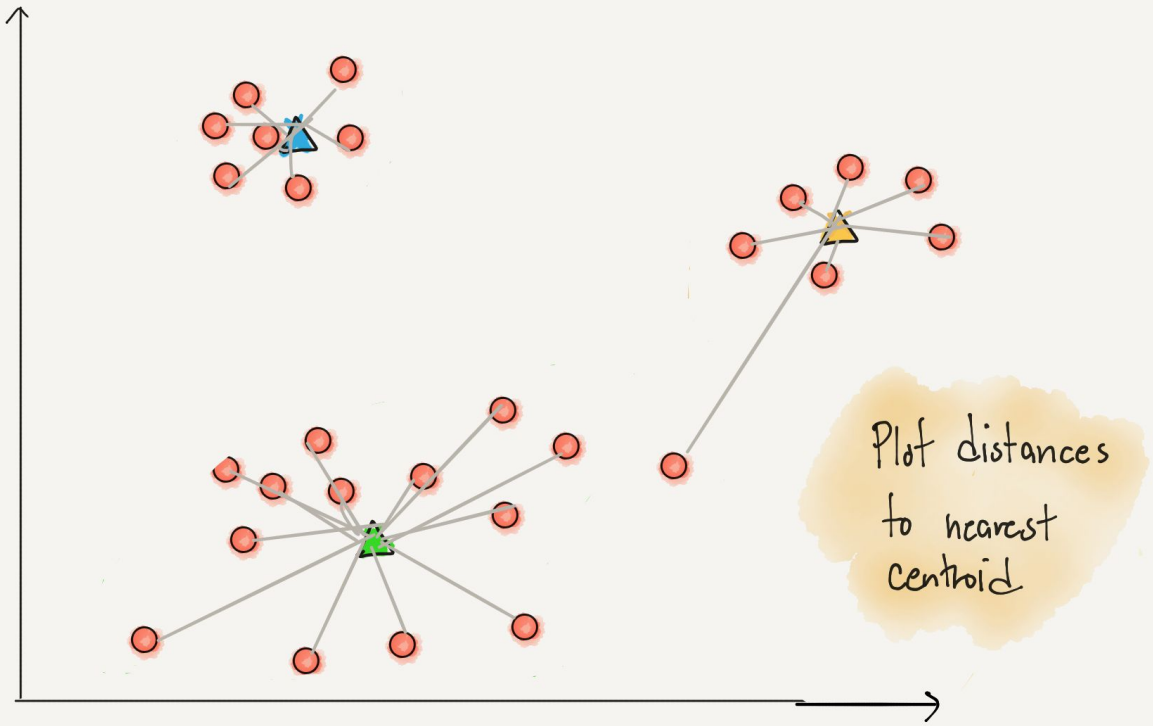
K-Means Clustering



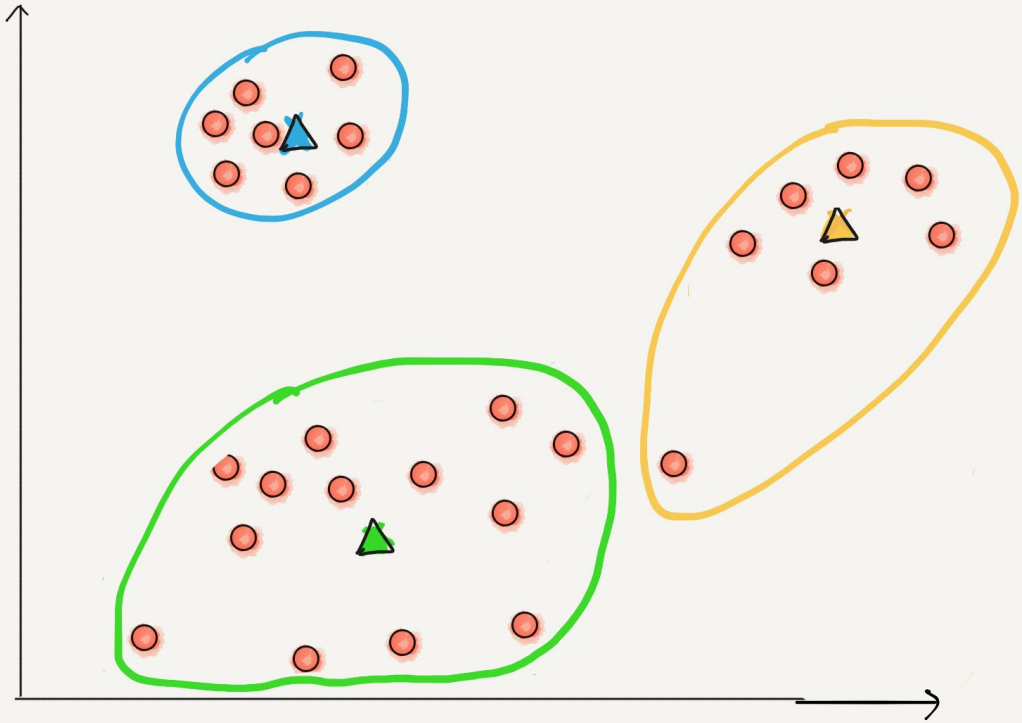
K-Means Clustering



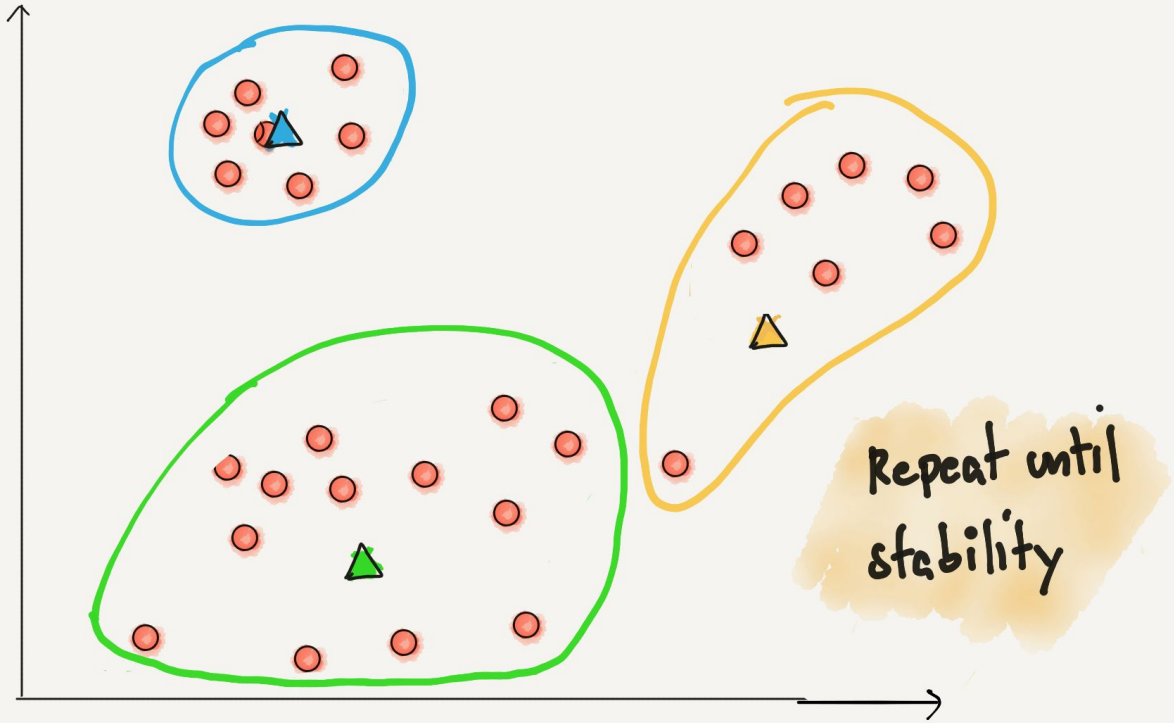
K-Means Clustering



K-Means Clustering



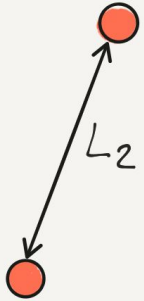
K-Means Clustering



Data Assignment Step

$$\text{argmin } \text{dist}(c_i, x)^2 ; c_i \in C$$

* C is the set of all centroids
* c_i each point in C
* x data point
* dist is Euclidean Distance or L2 distance

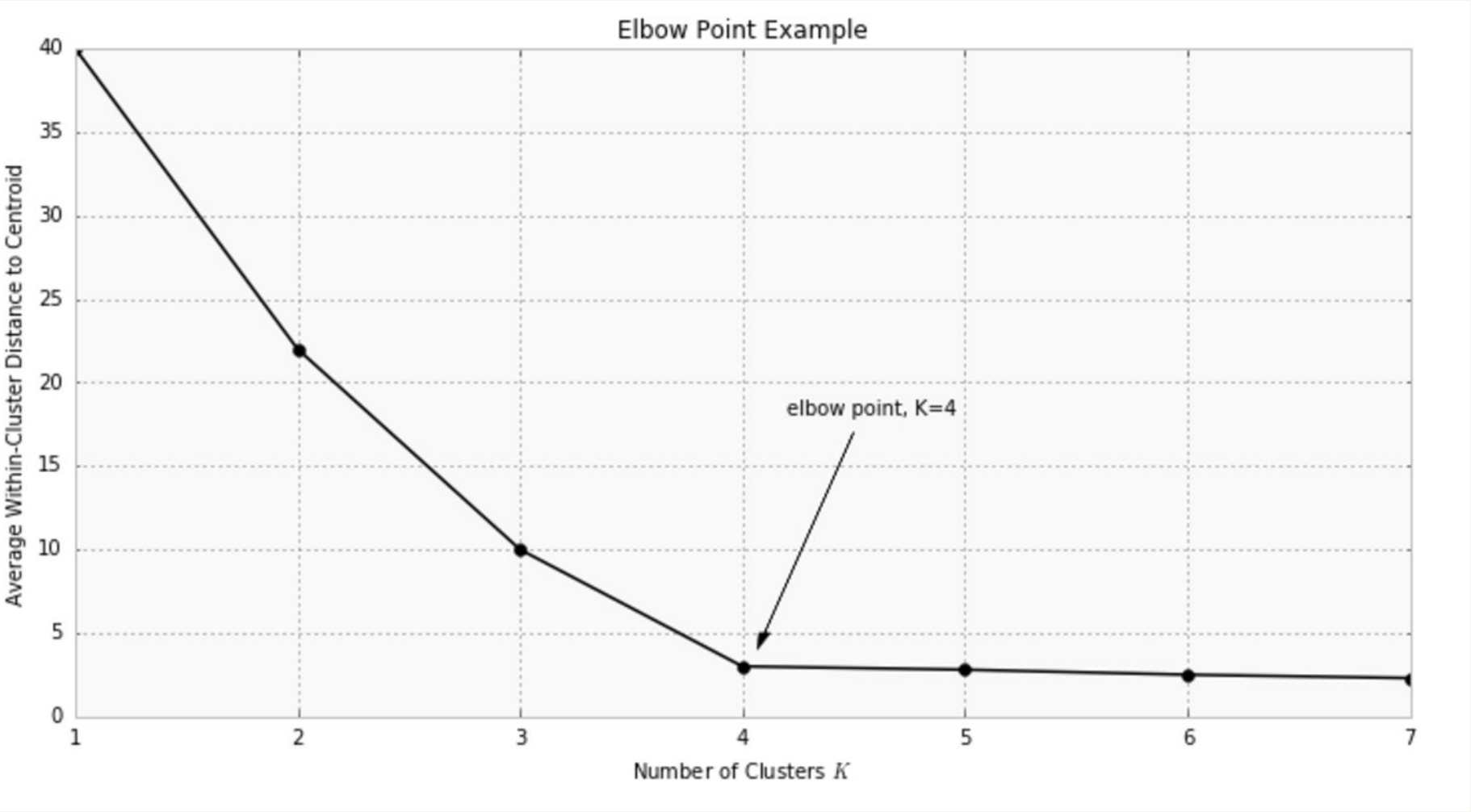


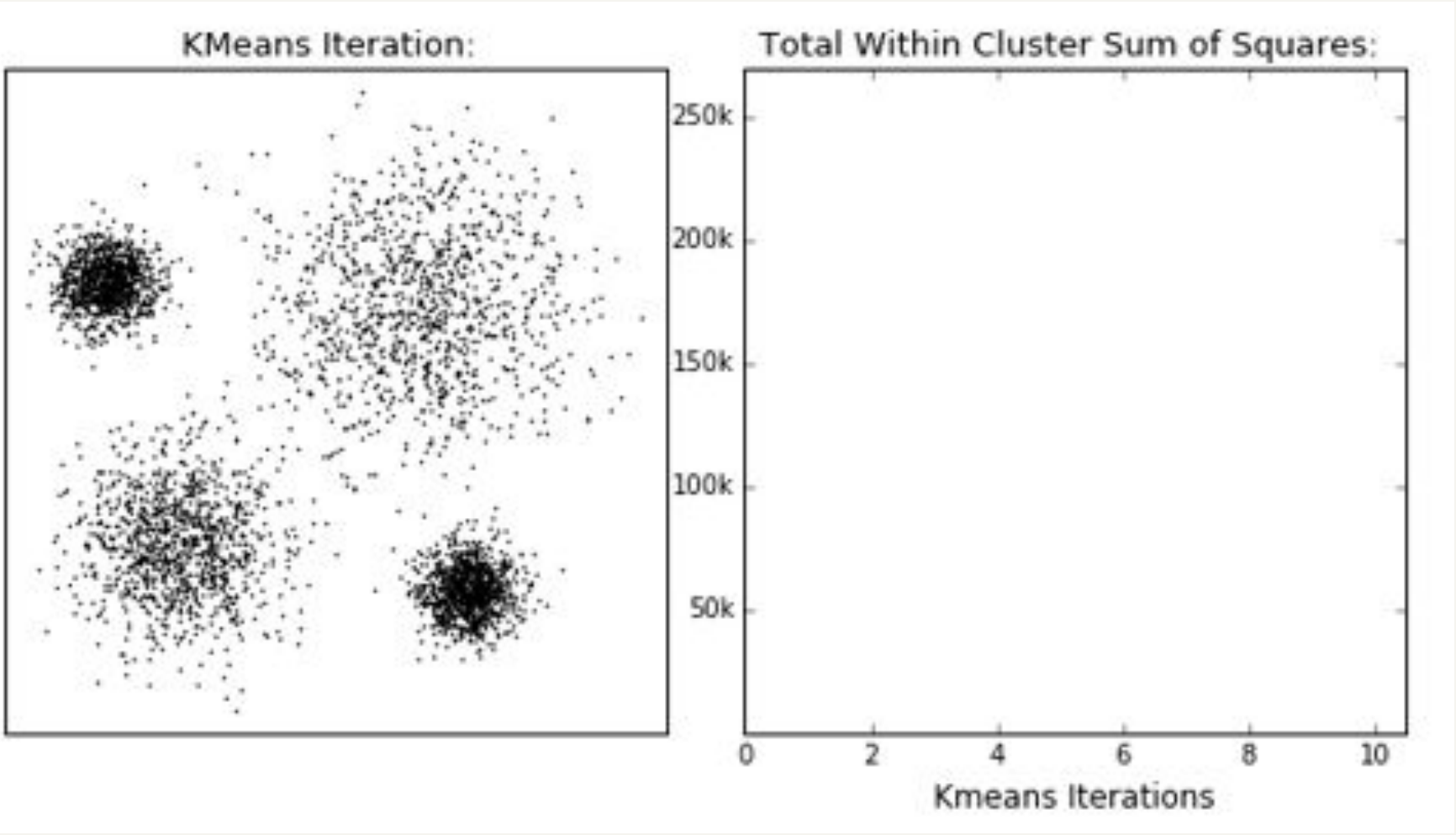
Centroid Update Step

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

S_i = set of data points per ith cluster

Take the mean of all points in the cluster and set that as new centroid.


Elbow Point Example
Average Within-Cluster Distance to Centroid
Number of Clusters K
elbow point, K=4
0
5
10
15
20
25
30
35
40
1
2
3
4
5
6
7


KMeans Iteration:
Total Within Cluster Sum of Squares:
250k
200k
150k
100k
50k
0
2
4
6
8
10
Kmeans Iterations

DBSCAN Clustering

Density-based Clustering of Applications with Noise

It doesn’t require that you input the number of clusters in order to run. But in exchange, you have to tune two other parameters.

ε - Epsilon Value

Minimum Points Value

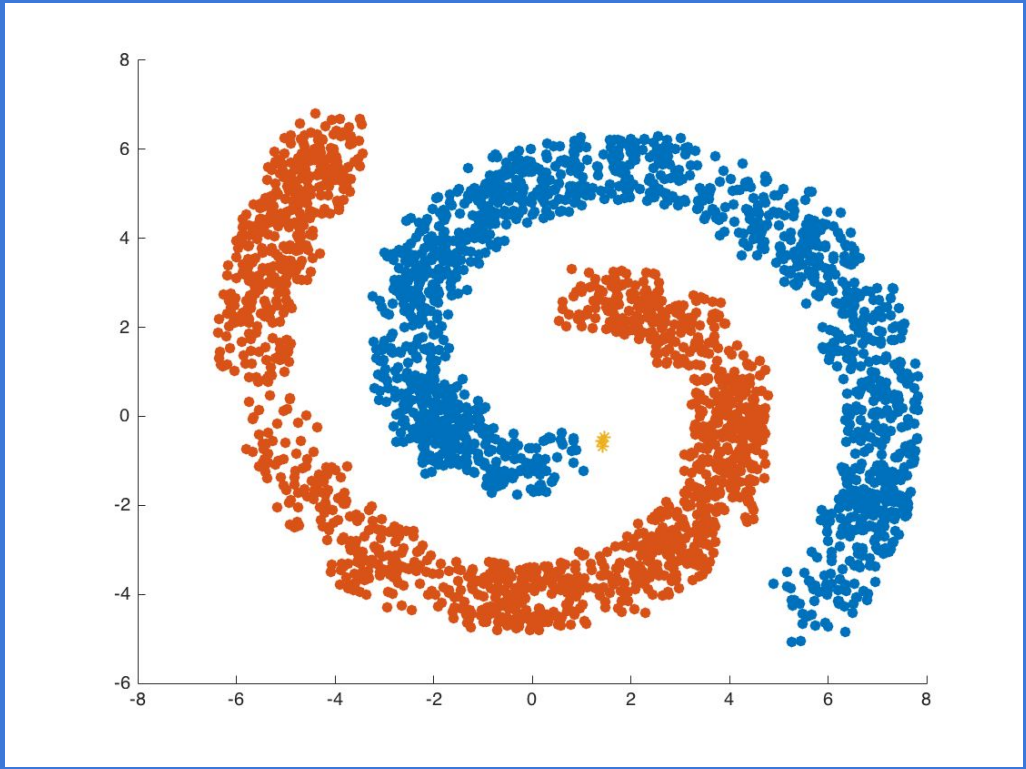
ε - Epsilon Value

Epsilon is the inverse base density the algorithm considers in grouping points

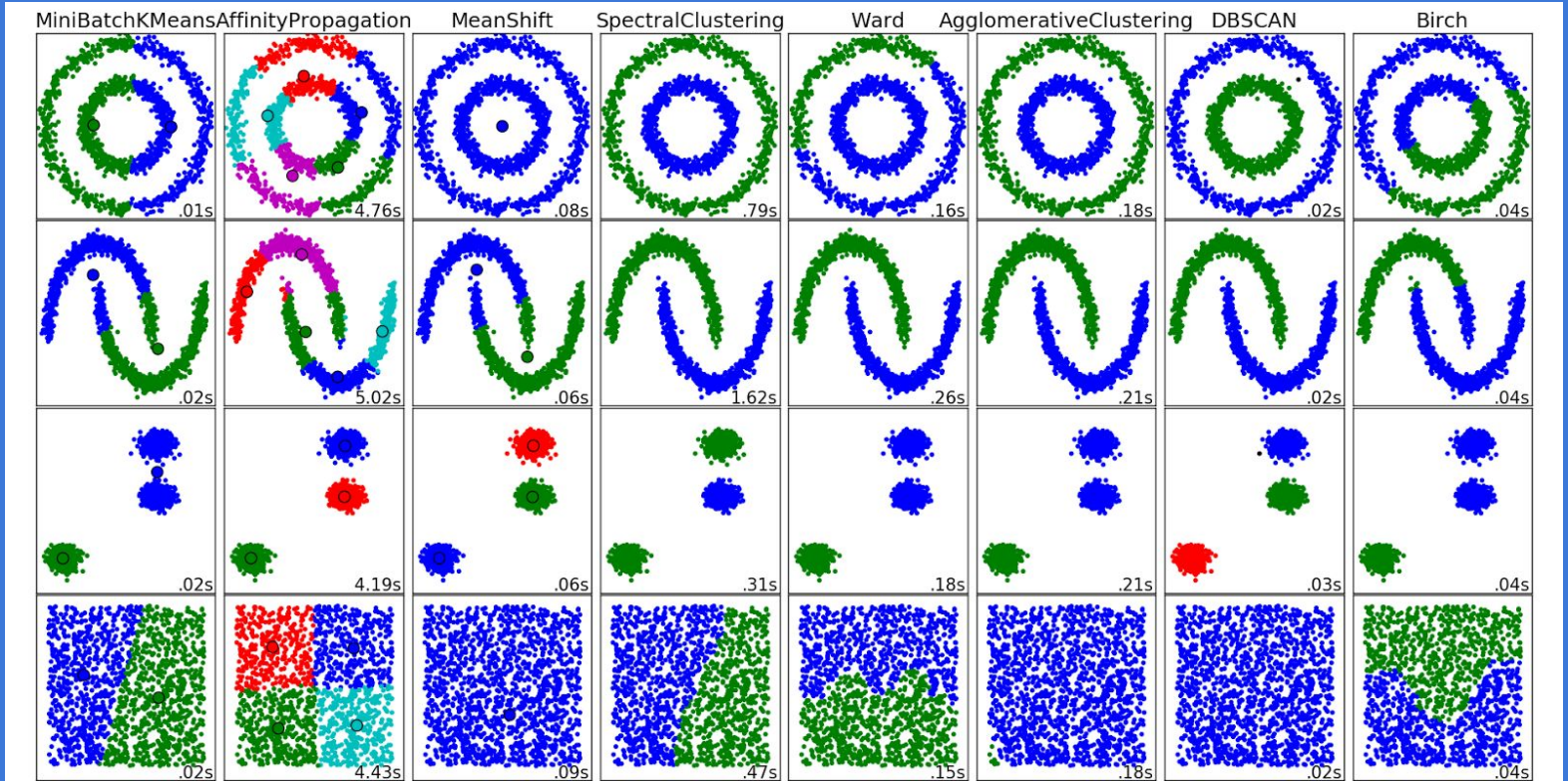
Minimum Points Value

Number of minimum points to declare as a cluster. Otherwise, classify as outlier.

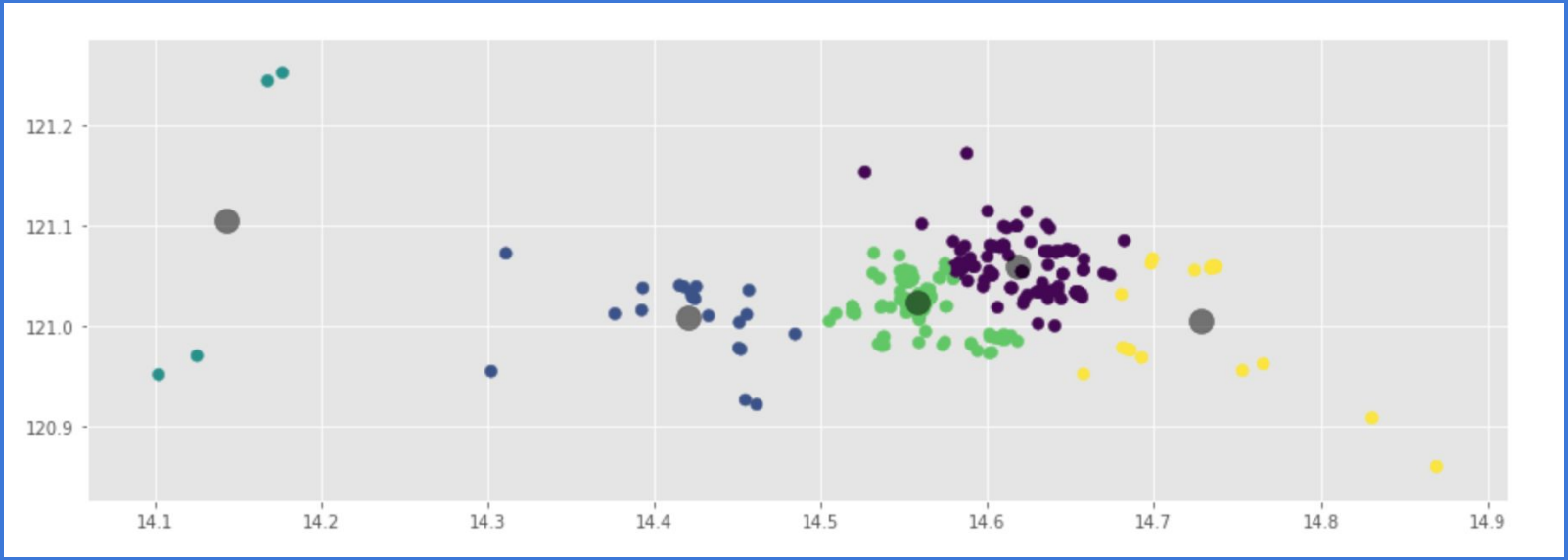
Benefits: Discover any number of clusters



Benefits: Clusters of any shape



Benefits: Can ignore outliers



Thank you