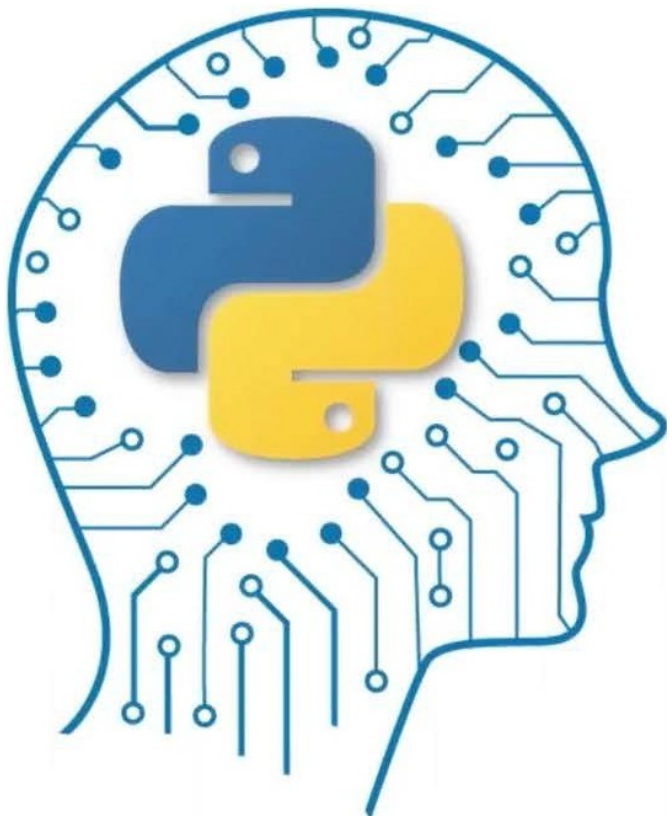
Advanced Machine Learning (AIE322)

Lecture 2

Prof. Aya M. Al-Zoghby
2024-2025

Agenda of Today's Session

- What is Clustering?
- Types of Clustering
- What is K- Means Clustering?
- How does a K-Means Algorithm works?
- K-Means with Python

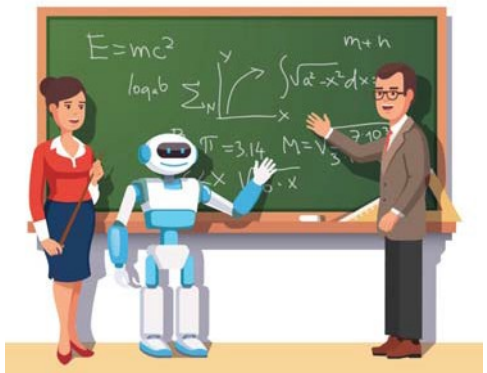


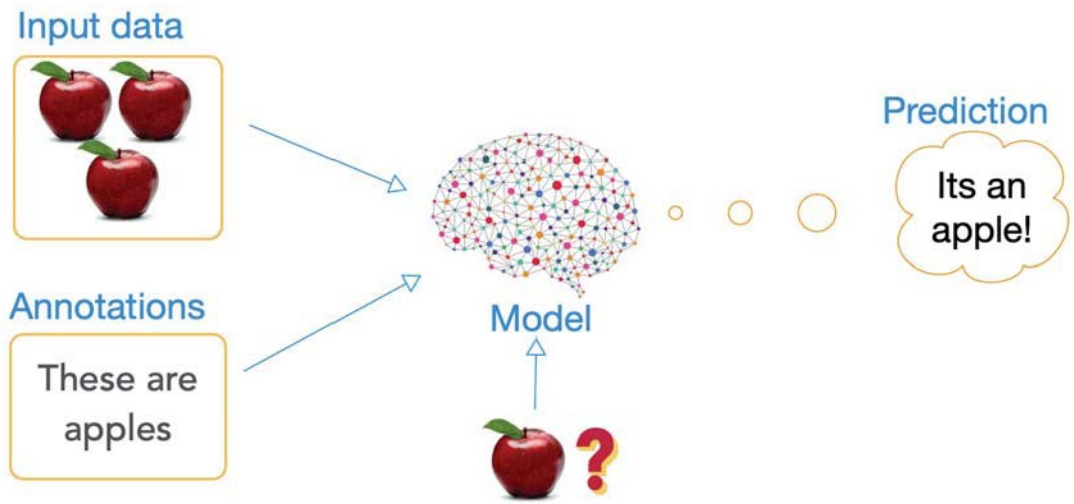
What is Clustering?

Clustering – grouping
<u>unlabelled</u> data

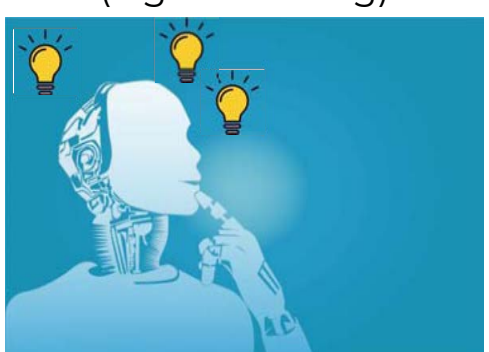
What is Clustering?

Supervised Learning
(e.g. Regression, Classification)





Unsupervised Learning
(e.g. Clustering)



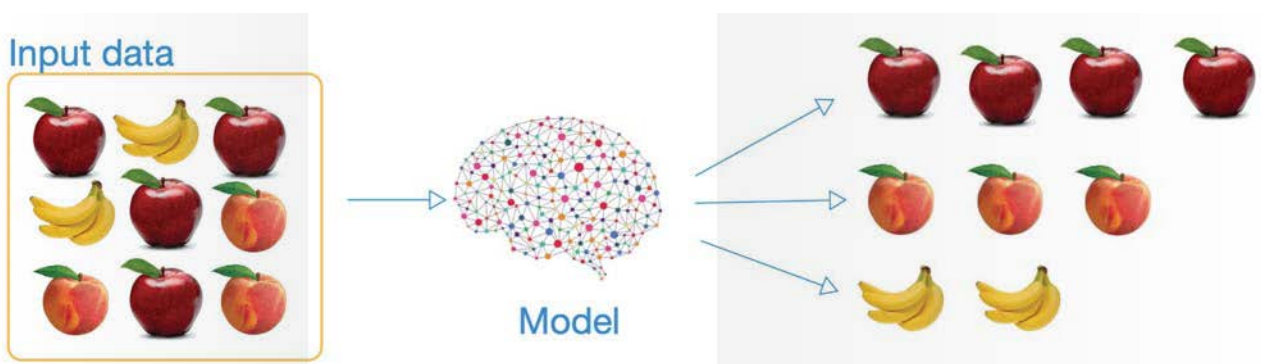
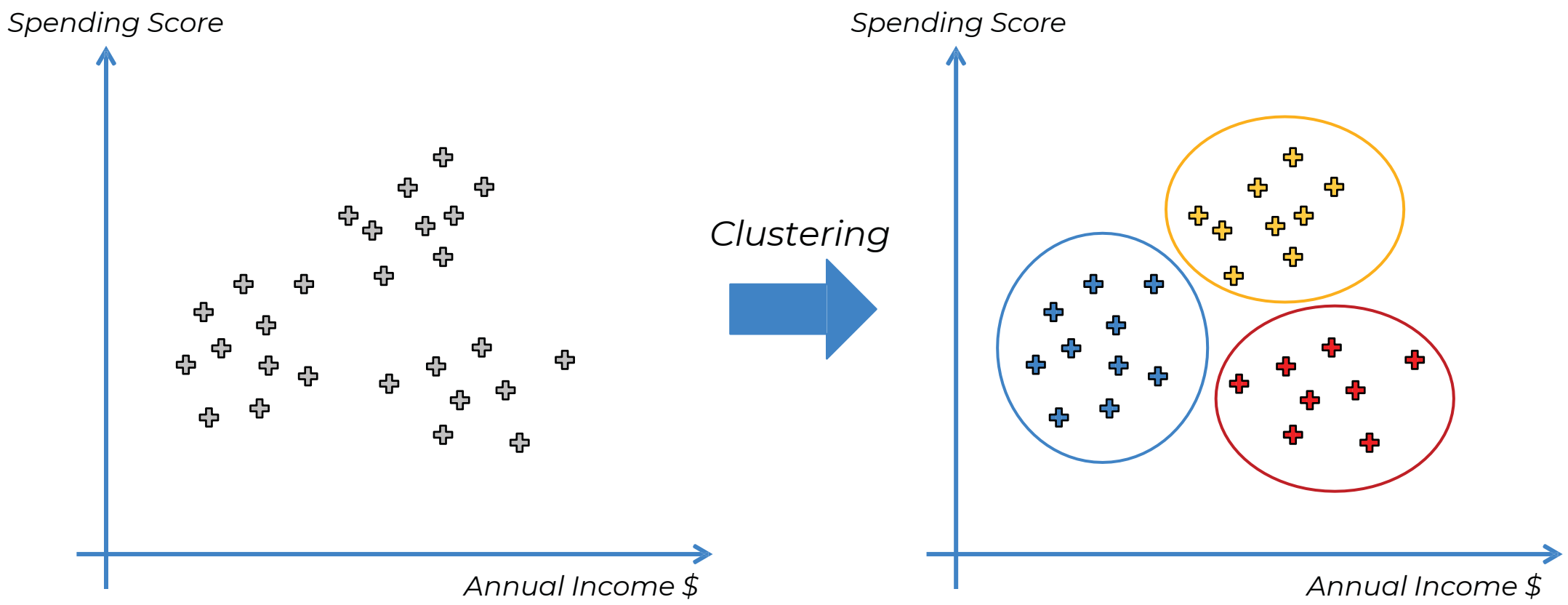


Image source: mdpi.com/2073-8994/10/12/73

What is Clustering?



edureka!

What is Clustering?

"Clustering is the process of dividing the datasets into groups, consisting of similar data-points"

- Points in the same group are as similar as possible
- Points in different group are as dissimilar as possible

edureka!

What is Clustering?

Group of diners in a restaurant

Items arranged in a mall

edureka!

Where is it Used?

amazon

Recommendation System

NETFLIX

Recommended Movies

flickr

Fickr's Photos

edureka!

How business use Clustering?

Retail Store

Banking

Insurance Companies

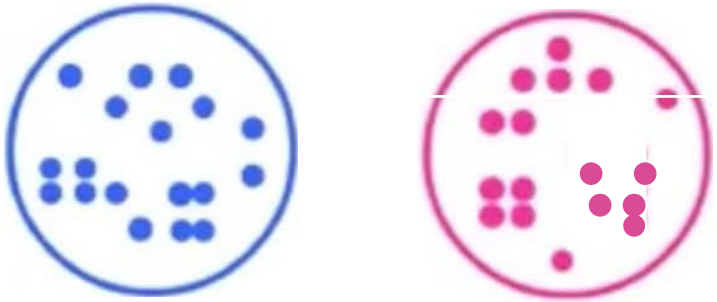
edureka!

Types of Clustering

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering

Exclusive Clustering

- Hard Clustering
- Data Point / Item belongs exclusively to one cluster
- For Example: K-Means Clustering



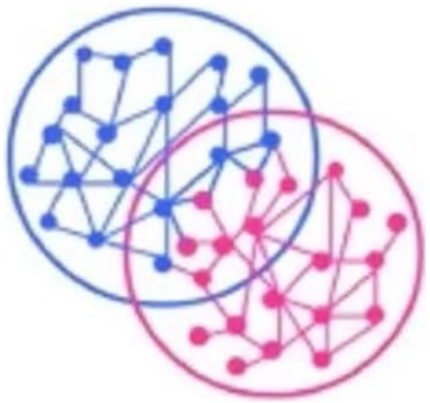
edureka!

Types of Clustering

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering

Overlapping Clustering

- Soft Cluster
- Data Point/ Item belongs to multiple cluster
- For Example: Fuzzy/ C-Means Clustering

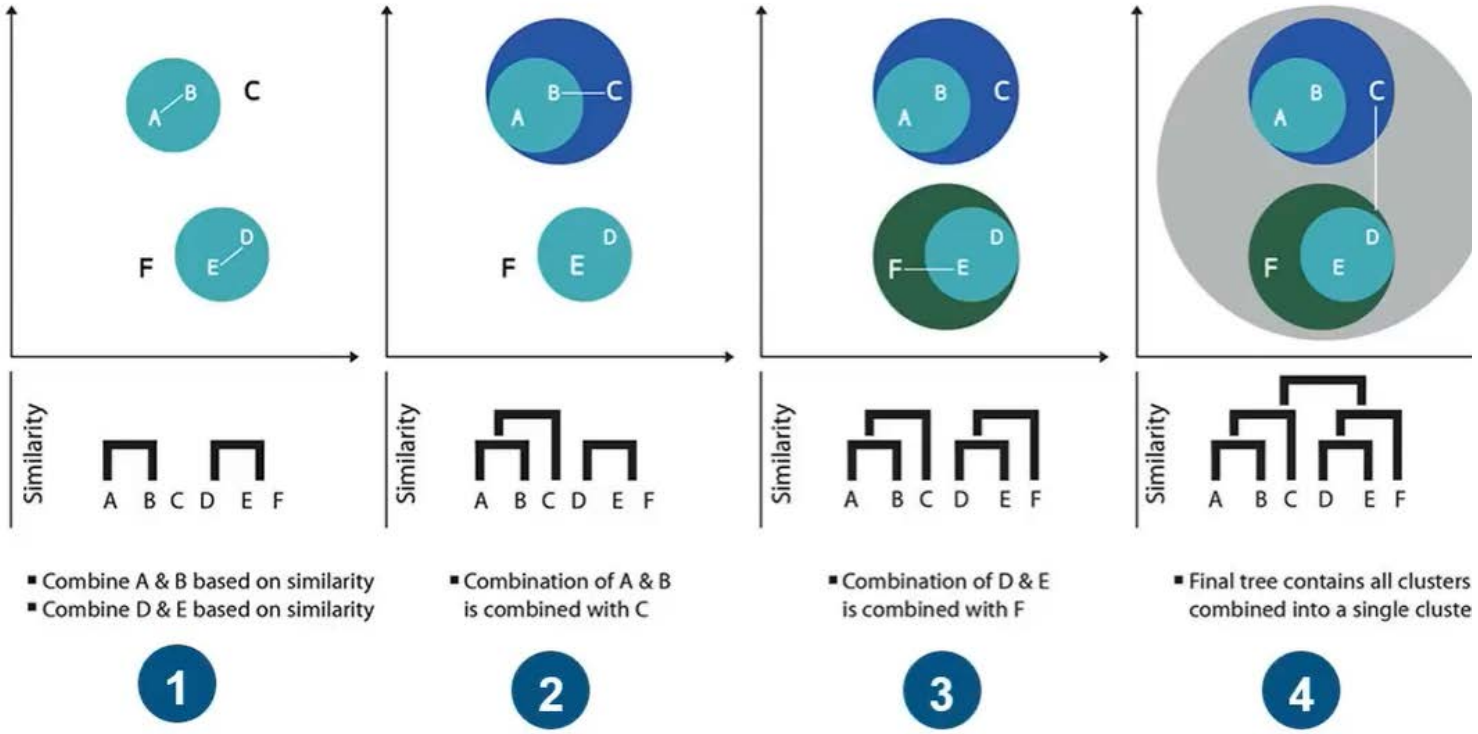


edureka!

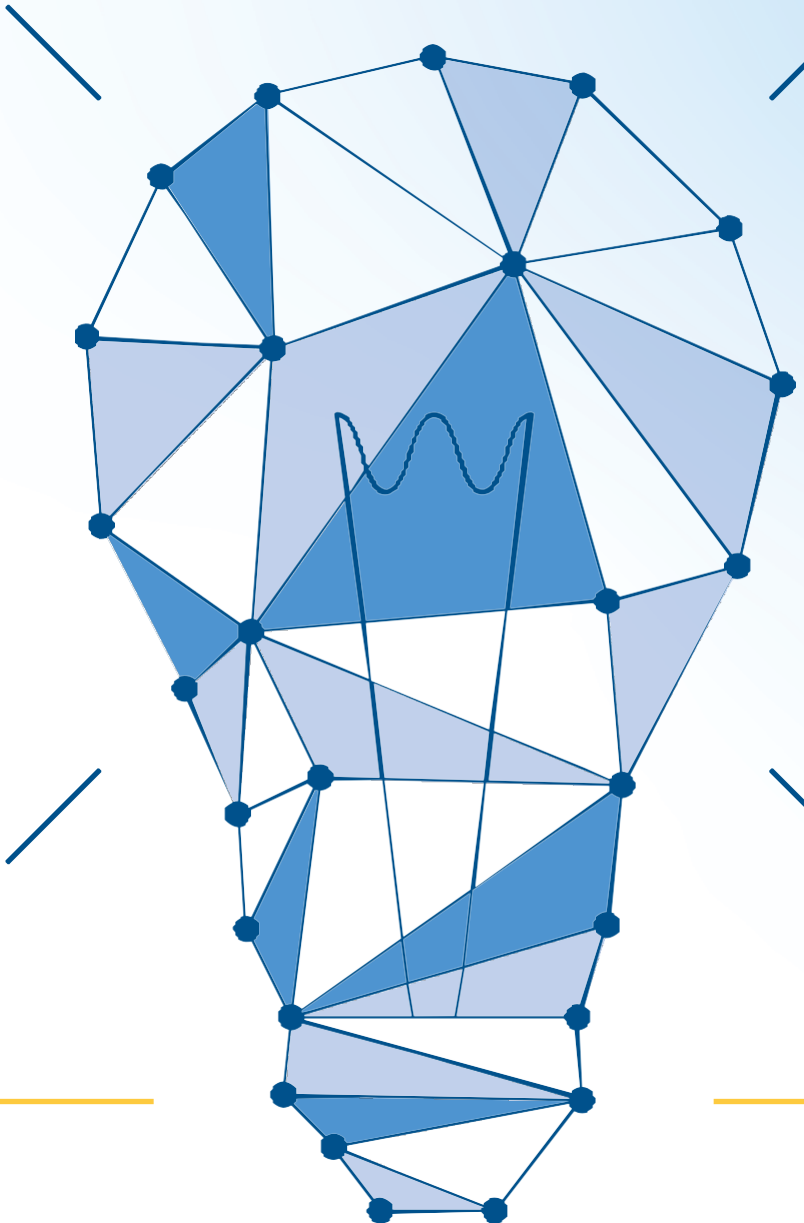
Types of Clustering

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering

Hierarchical Clustering



K-Means Clustering



edureka!

What is K-Means Clustering?

"K-Means is a clustering algorithm whose mail goal is to group similar elements or data points into a cluster."

NOTE: 'K' in K-Means represent the number of clusters

edureka!

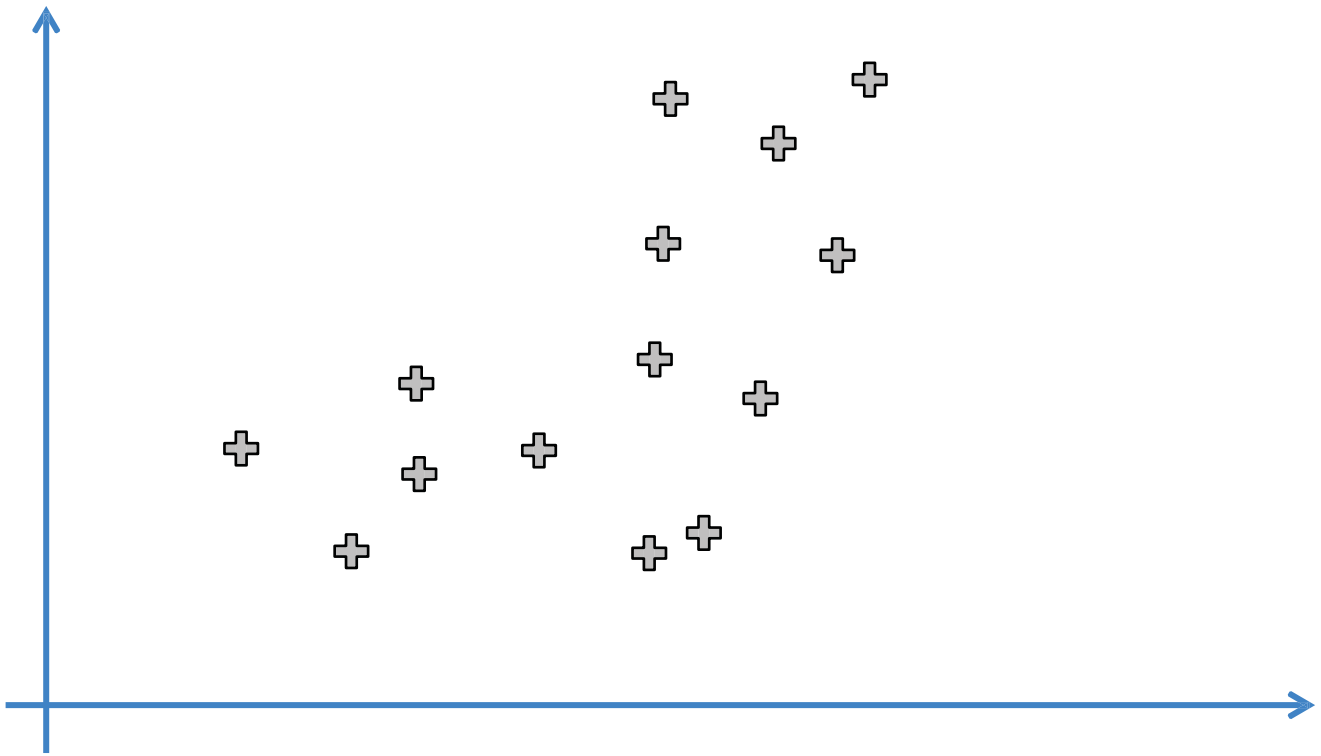
What is K-Means Clustering?

Pile of dirty clothes

K-Means Clustering

Step 1: Choosing the Number of Clusters (K):

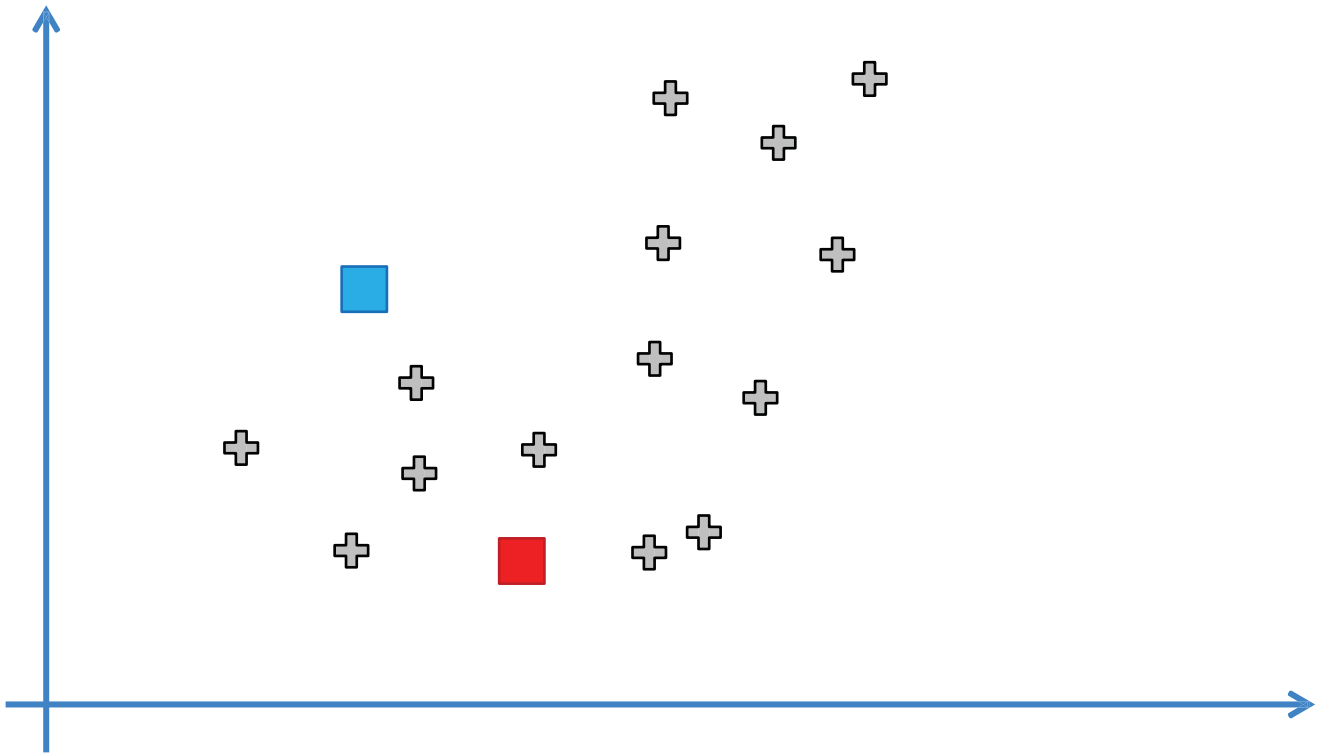
•Decide how many clusters to create based on the data or problem context. For this example, we'll assume two clusters.



K-Means Clustering

Step 2: Placing Centroids:

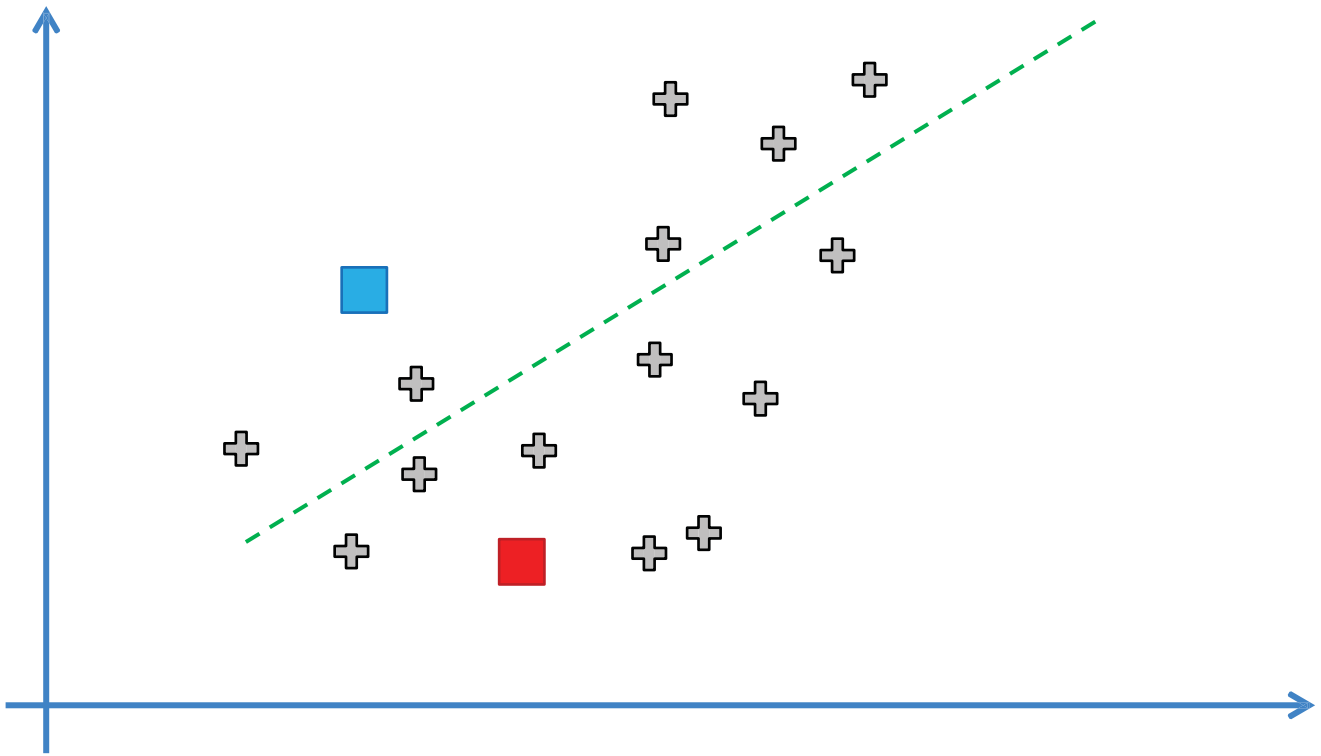
•Randomly initialize K centroids on the scatter plot. These centroids represent the centers of the clusters and do not need to align with the actual data points.



K-Means Clustering

Step 3: Assigning Data Points to Centroids:

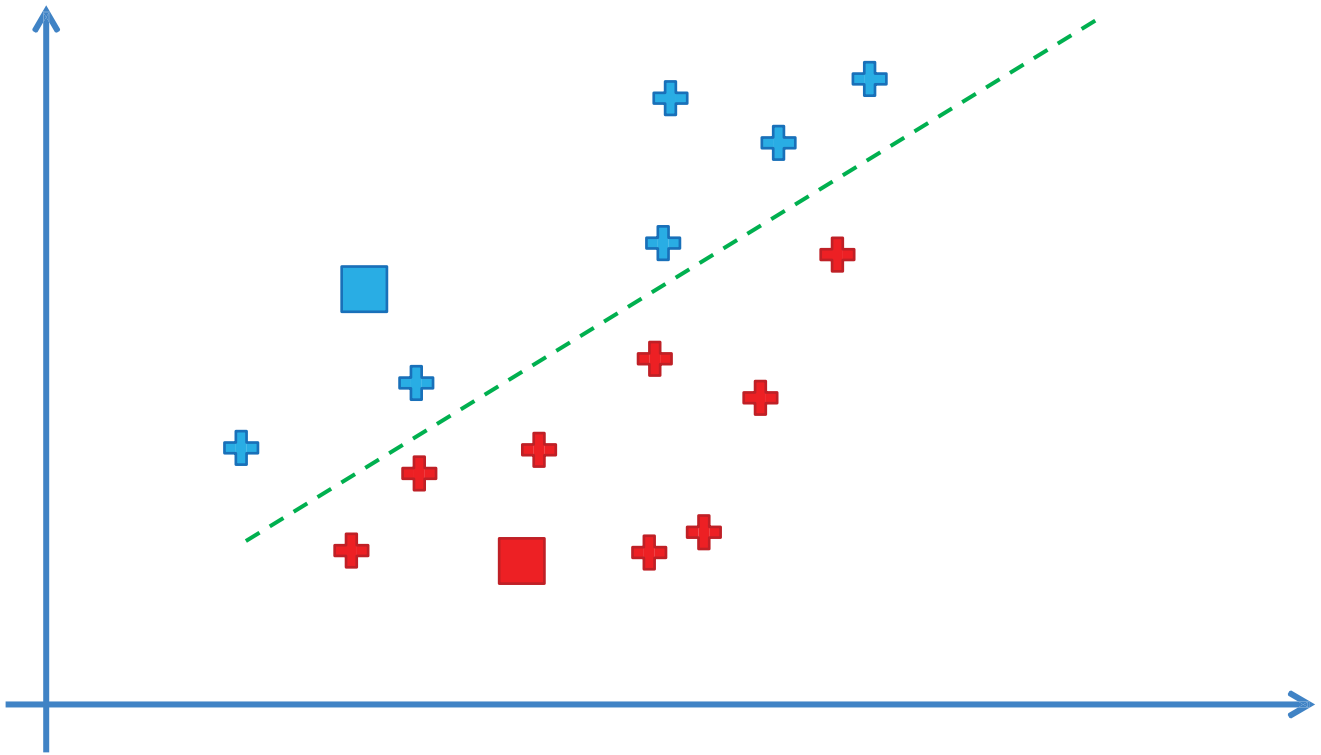
•Measure the distance between each data point and the centroids. Assign each data point to the nearest centroid, forming preliminary clusters. Visualization can include drawing equidistant lines to aid in assignment.



K-Means Clustering

Step 3: Assigning Data Points to Centroids:

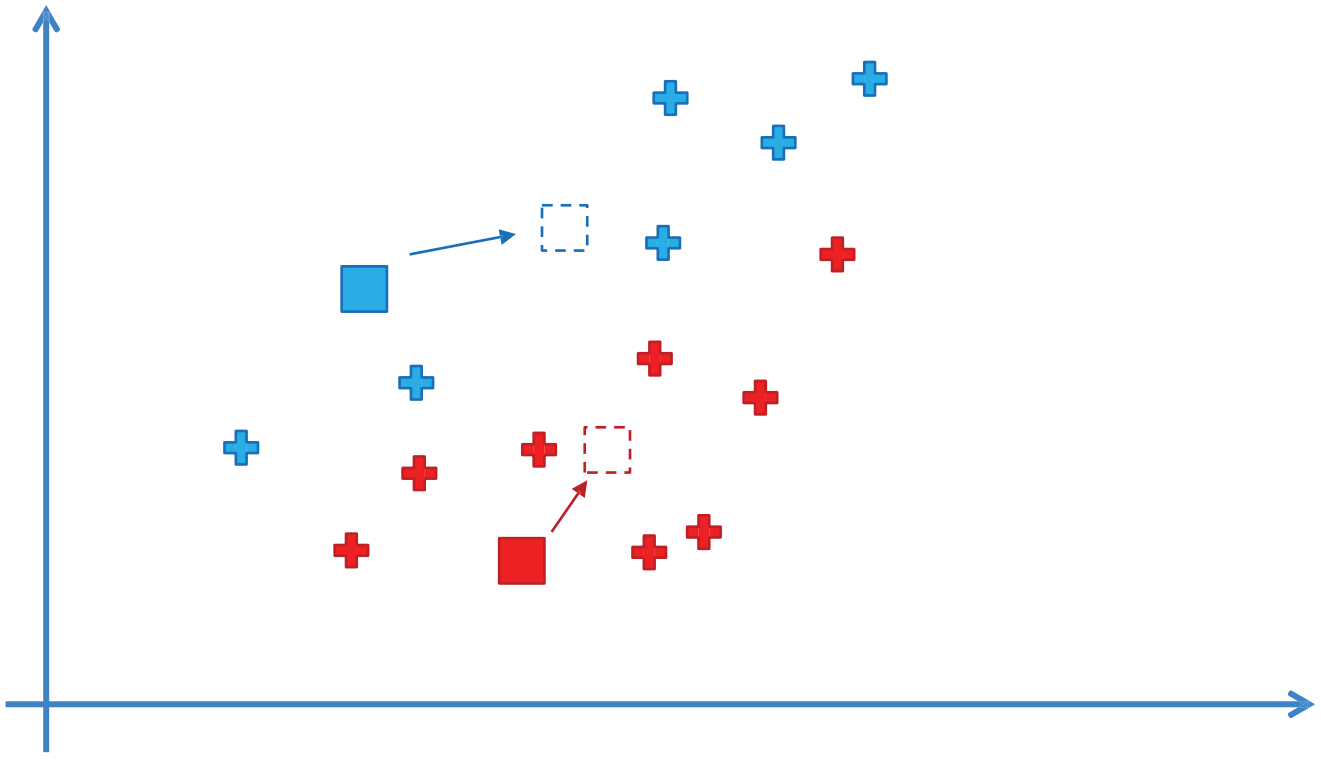
•Measure the distance between each data point and the centroids. Assign each data point to the nearest centroid, forming preliminary clusters. Visualization can include drawing equidistant lines to aid in assignment.



K-Means Clustering

Step 4: Recalculating the Centroids:

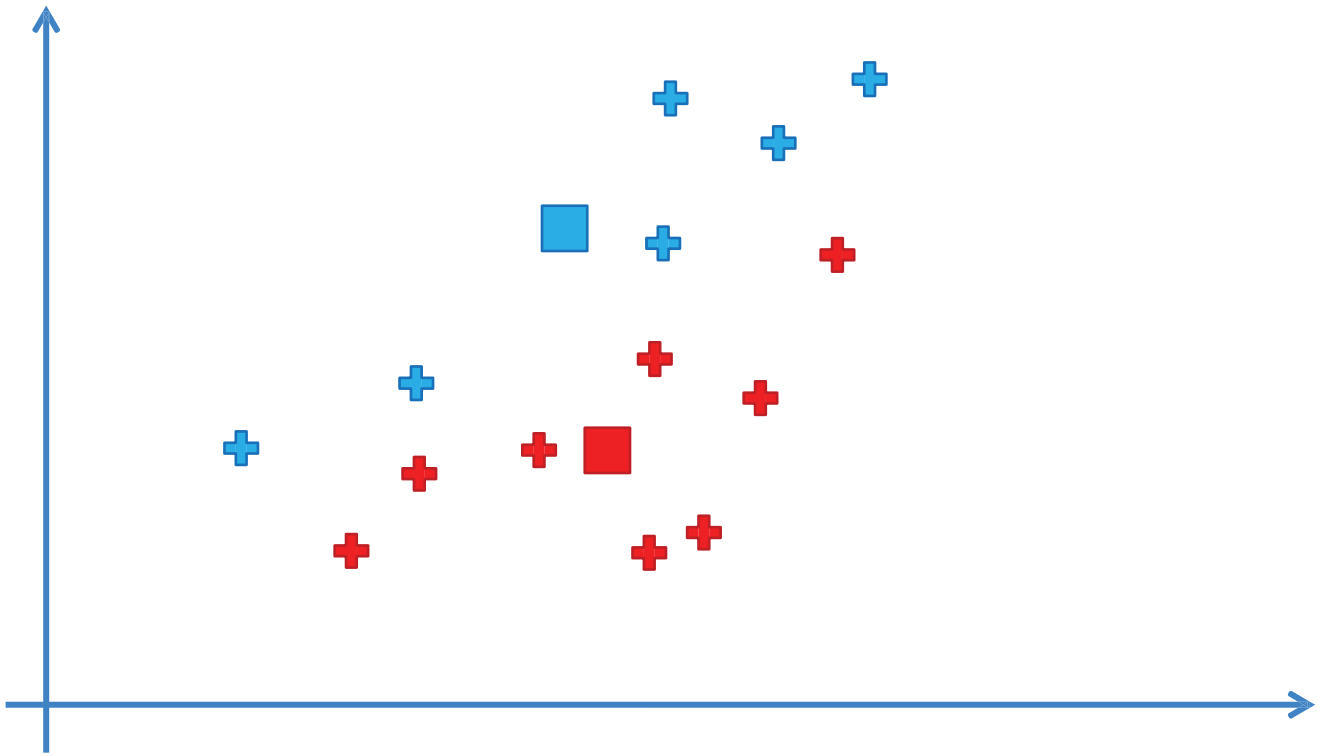
•Calculate the new positions of the centroids by averaging the coordinates of the data points assigned to each cluster, excluding the original centroids from the calculation.



K-Means Clustering

Step 4: Recalculating the Centroids:

•Calculate the new positions of the centroids by averaging the coordinates of the data points assigned to each cluster, excluding the original centroids from the calculation.

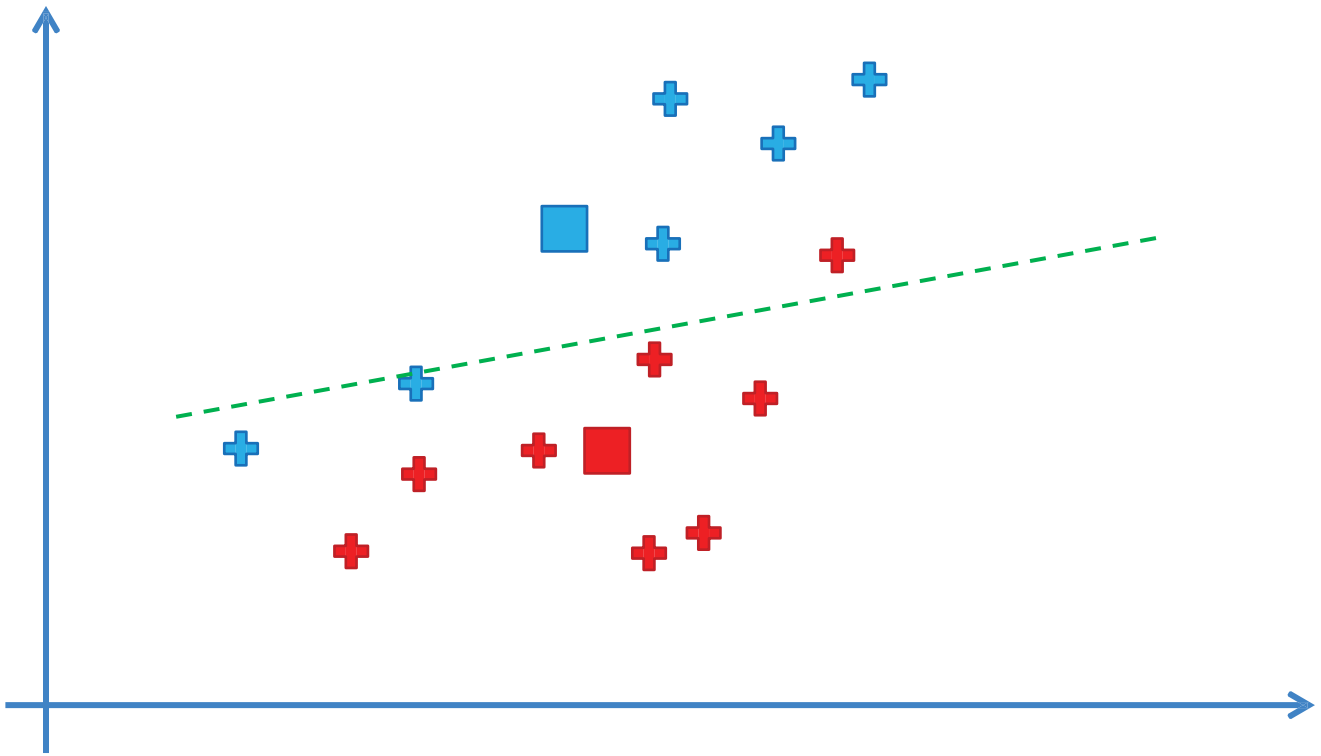


K-Means Clustering

Step 5: Reassigning Data Points:

•Repeat the assignment process:

- Reassign each data point to the closest centroid.
- Redraw equidistant lines.
- Calculate the new center of mass for each cluster.
- Update the centroid positions.

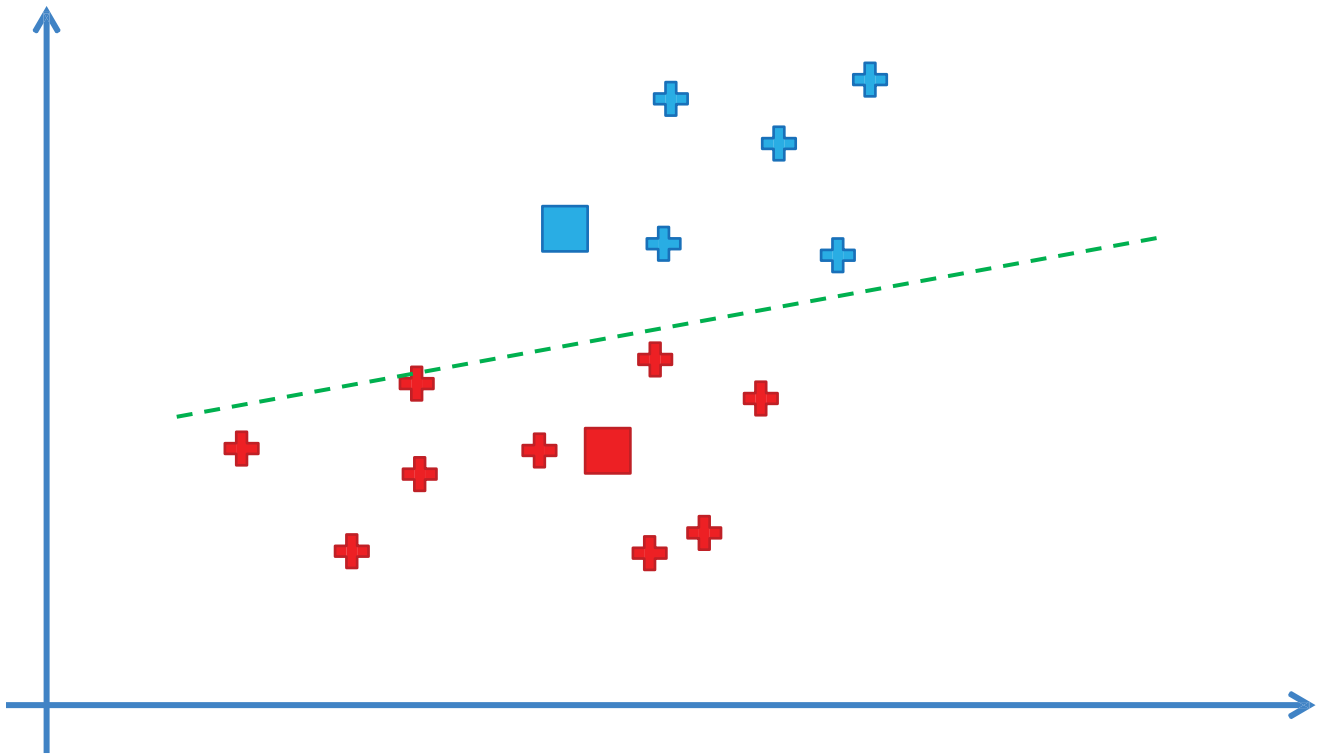


K-Means Clustering

Step 5: Reassigning Data Points:

•Repeat the assignment process:

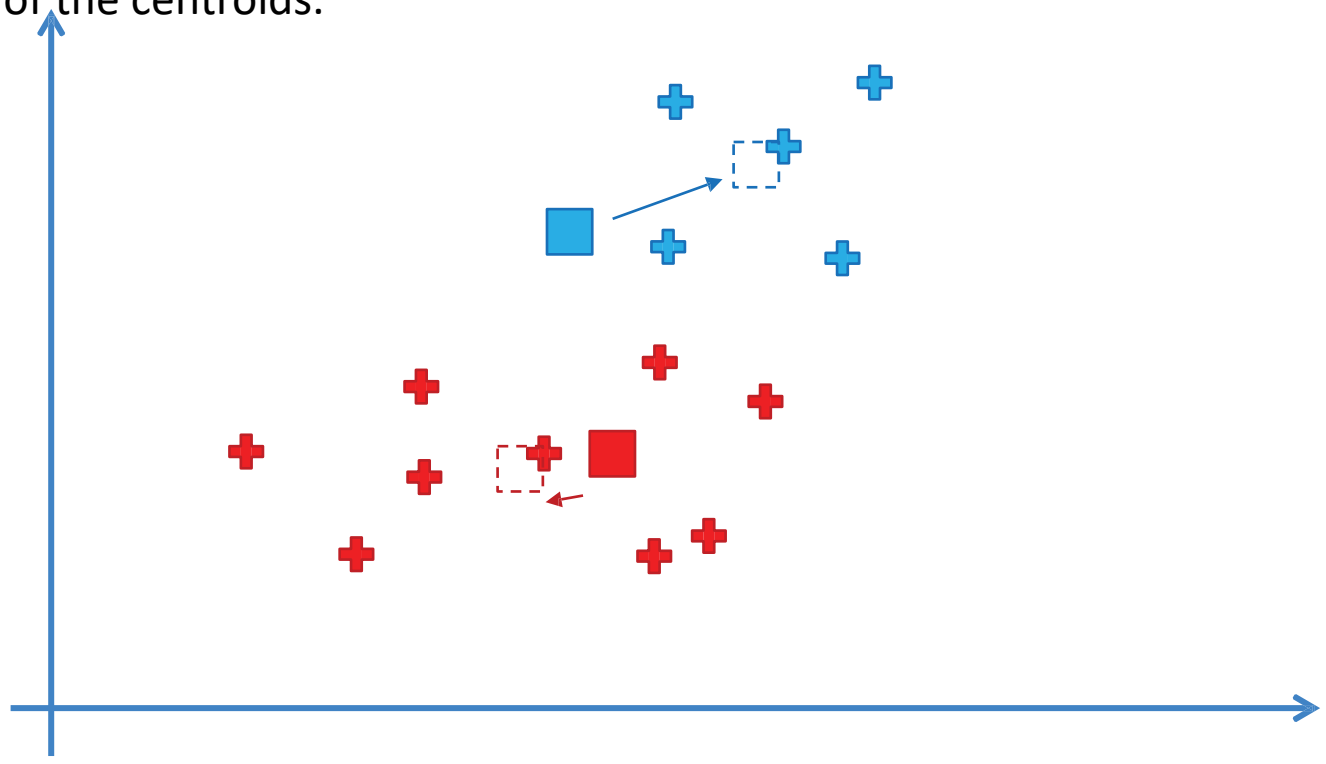
- Reassign each data point to the closest centroid.
- Redraw equidistant lines.
- Calculate the new center of mass for each cluster.
- Update the centroid positions.



K-Means Clustering

Step 6: Iterating Until Convergence:

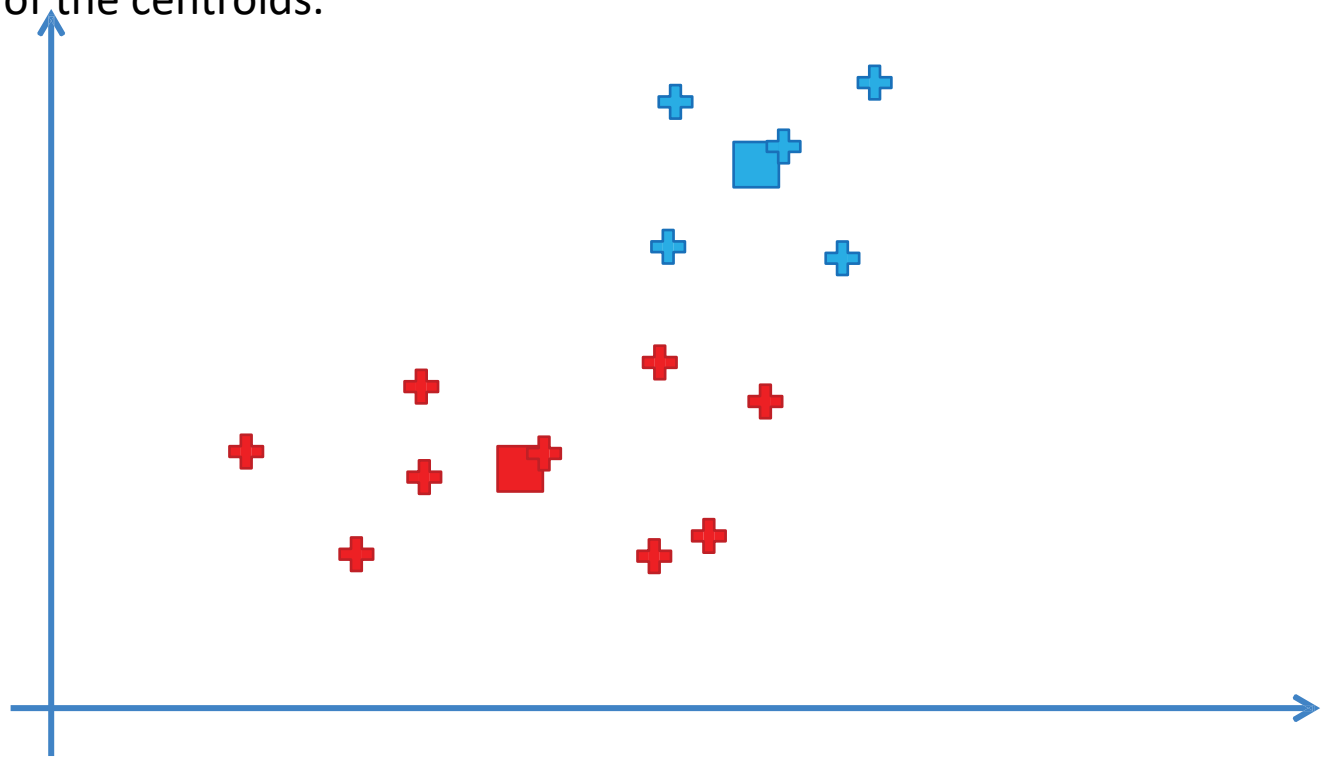
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

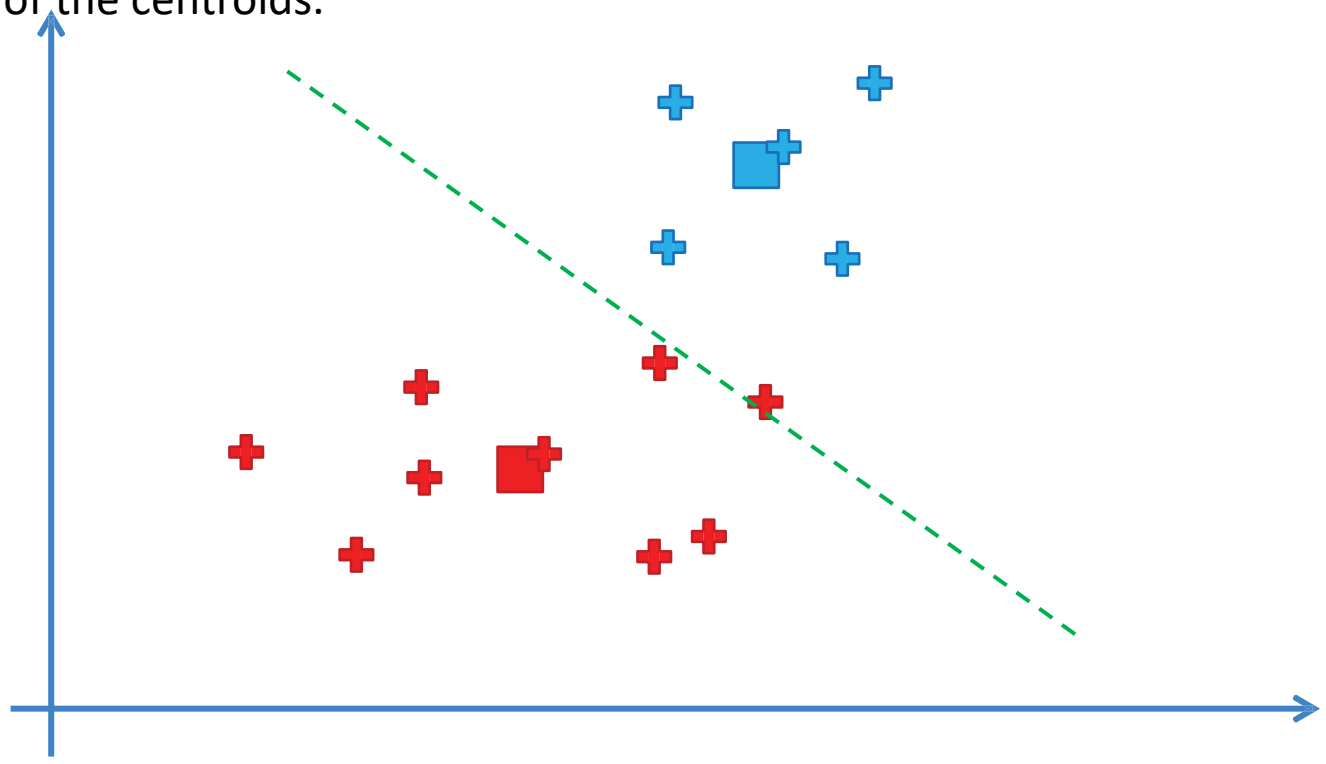
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

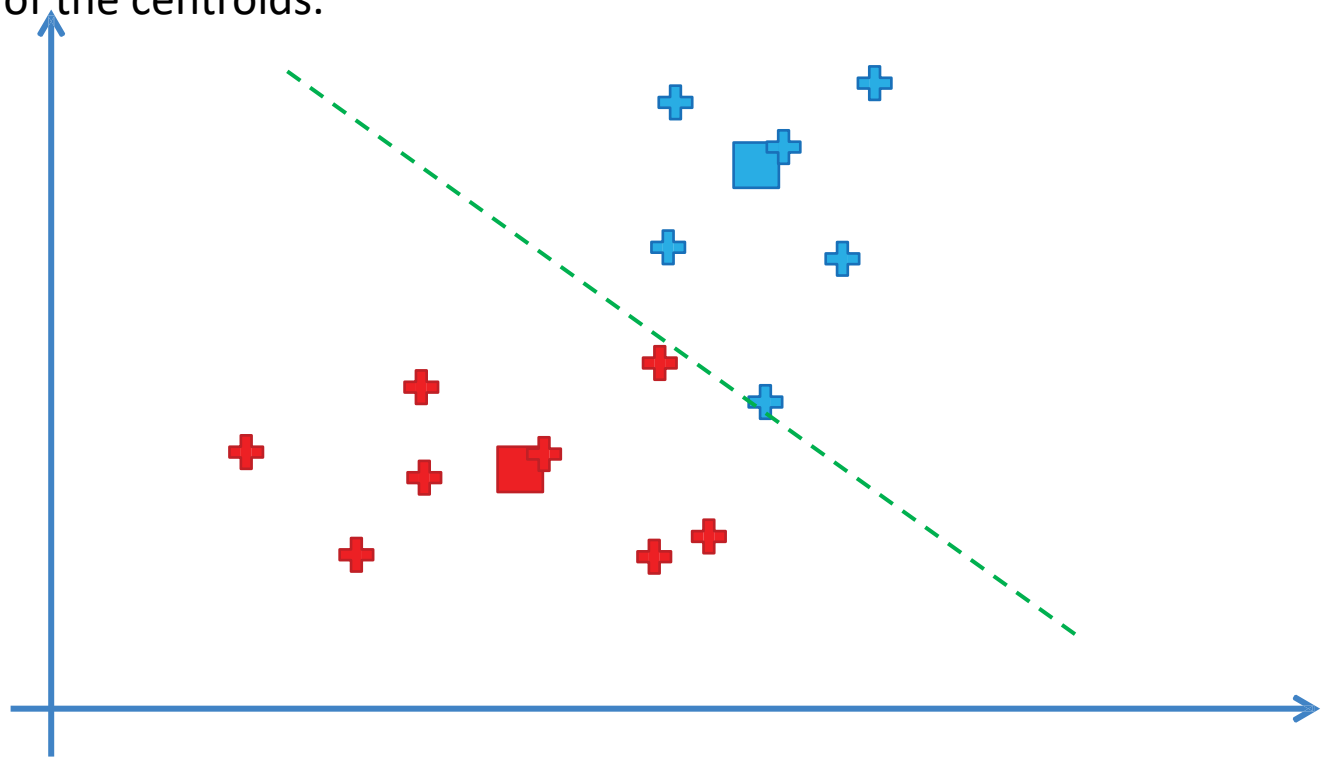
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

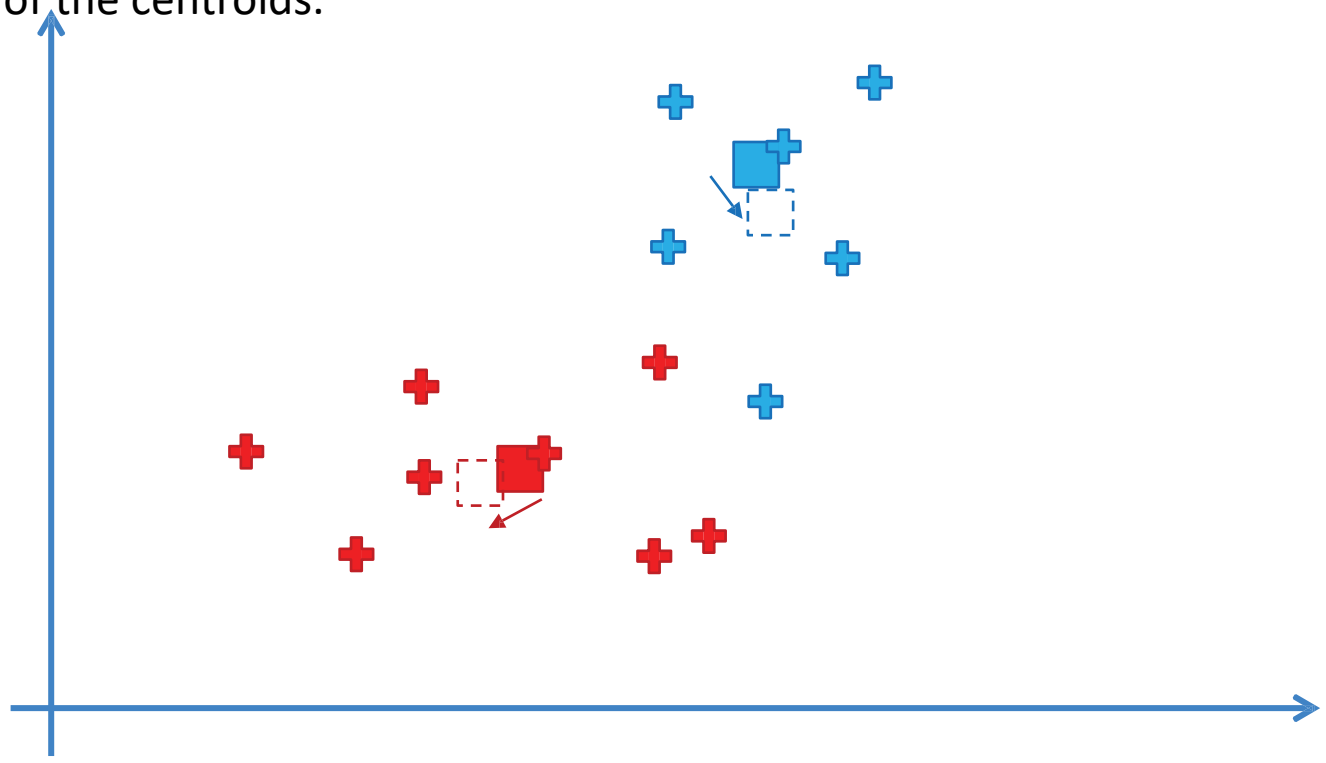
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

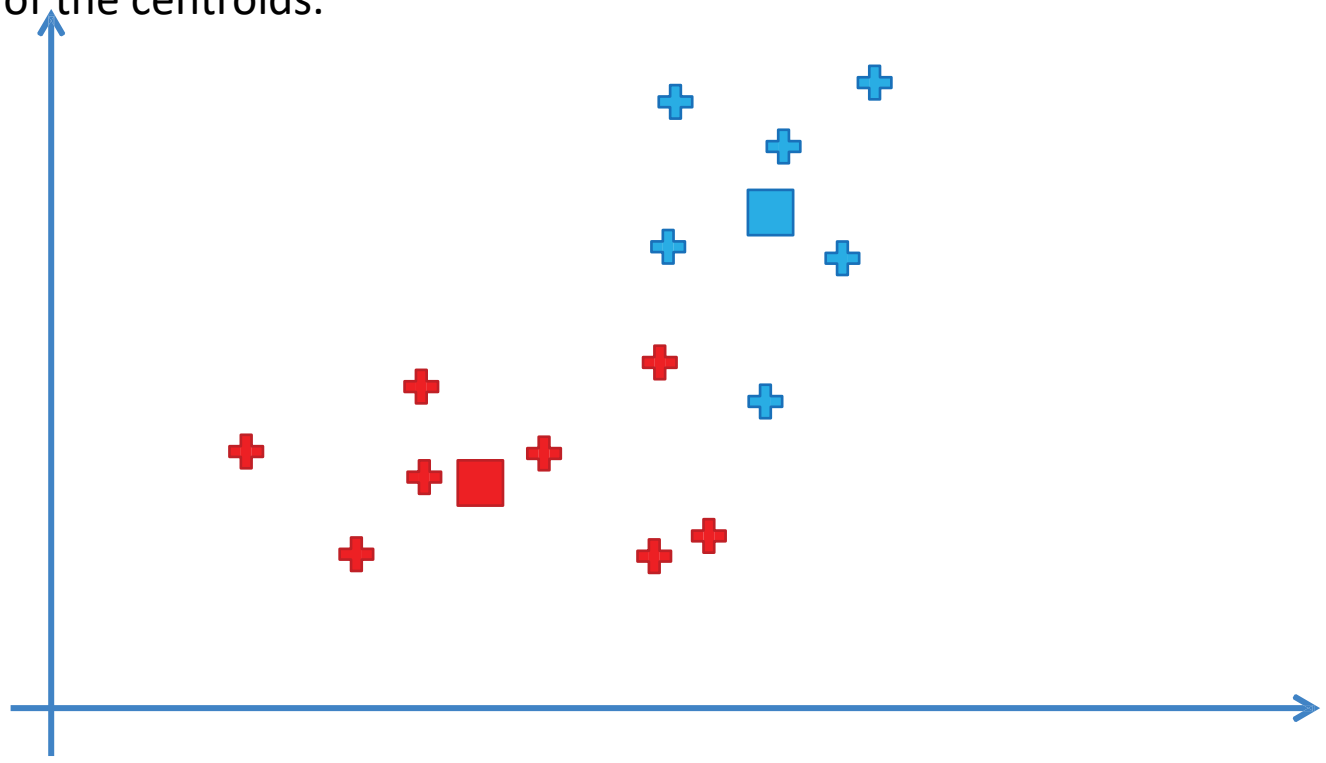
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

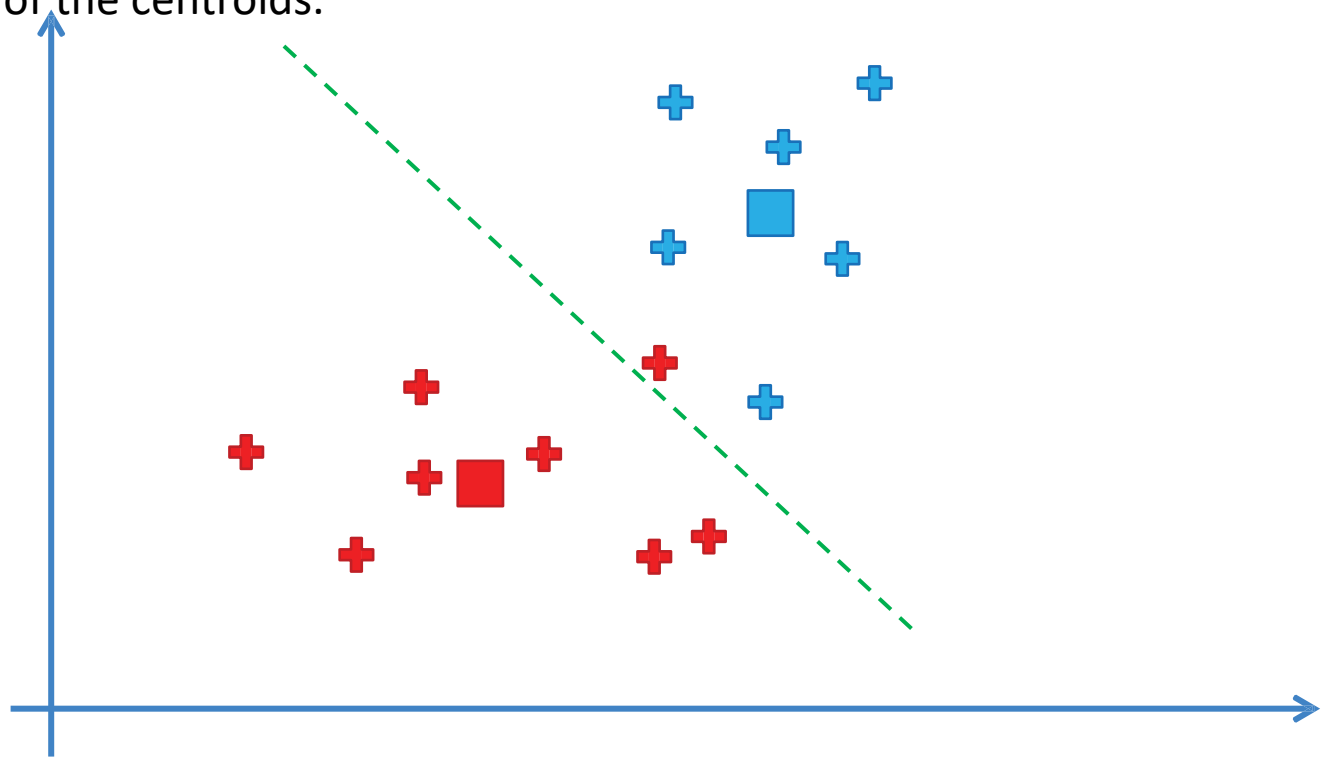
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

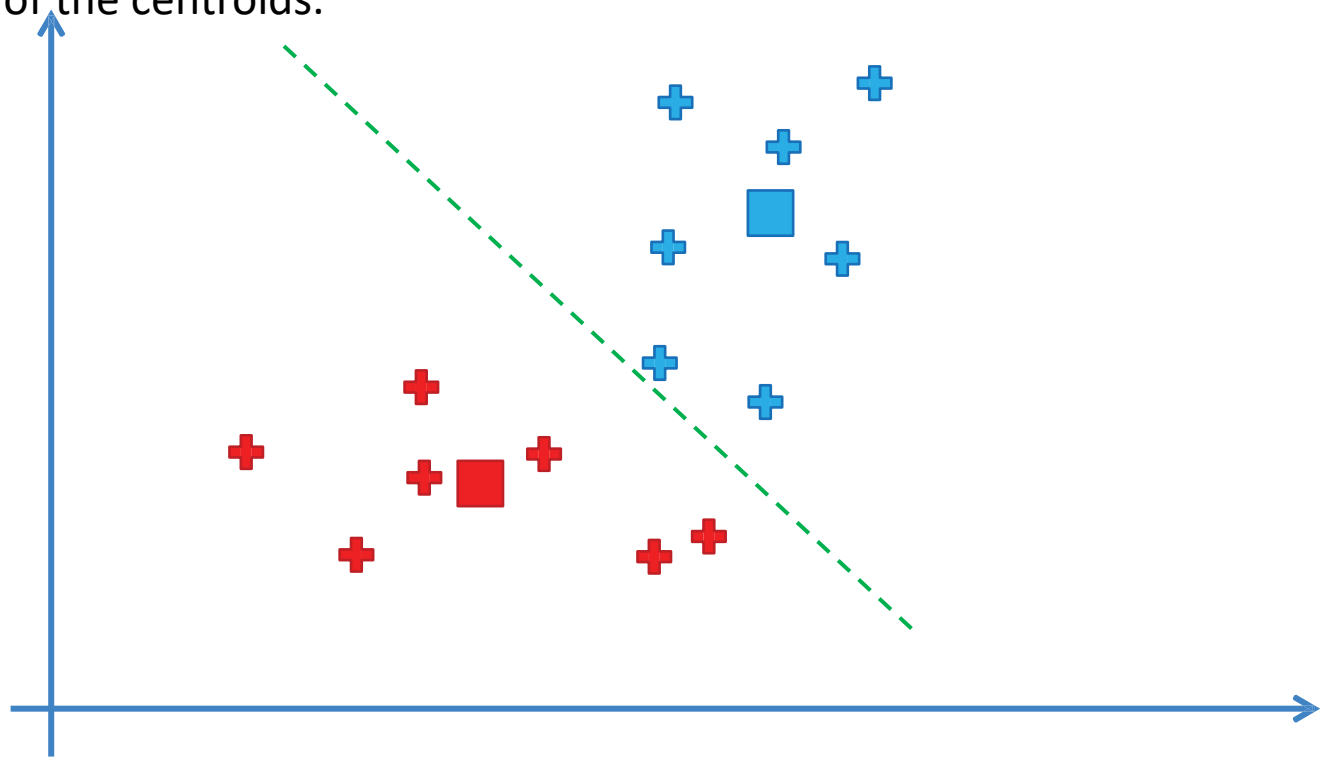
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

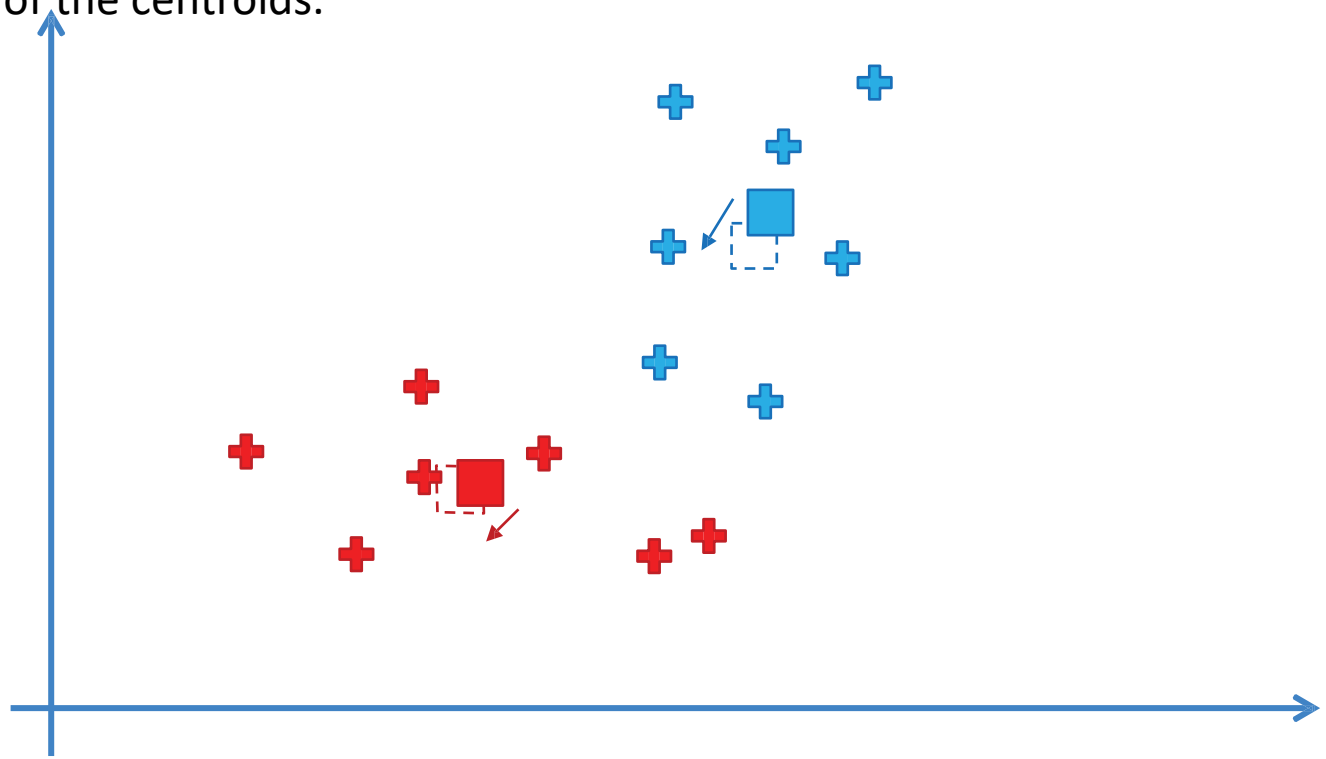
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

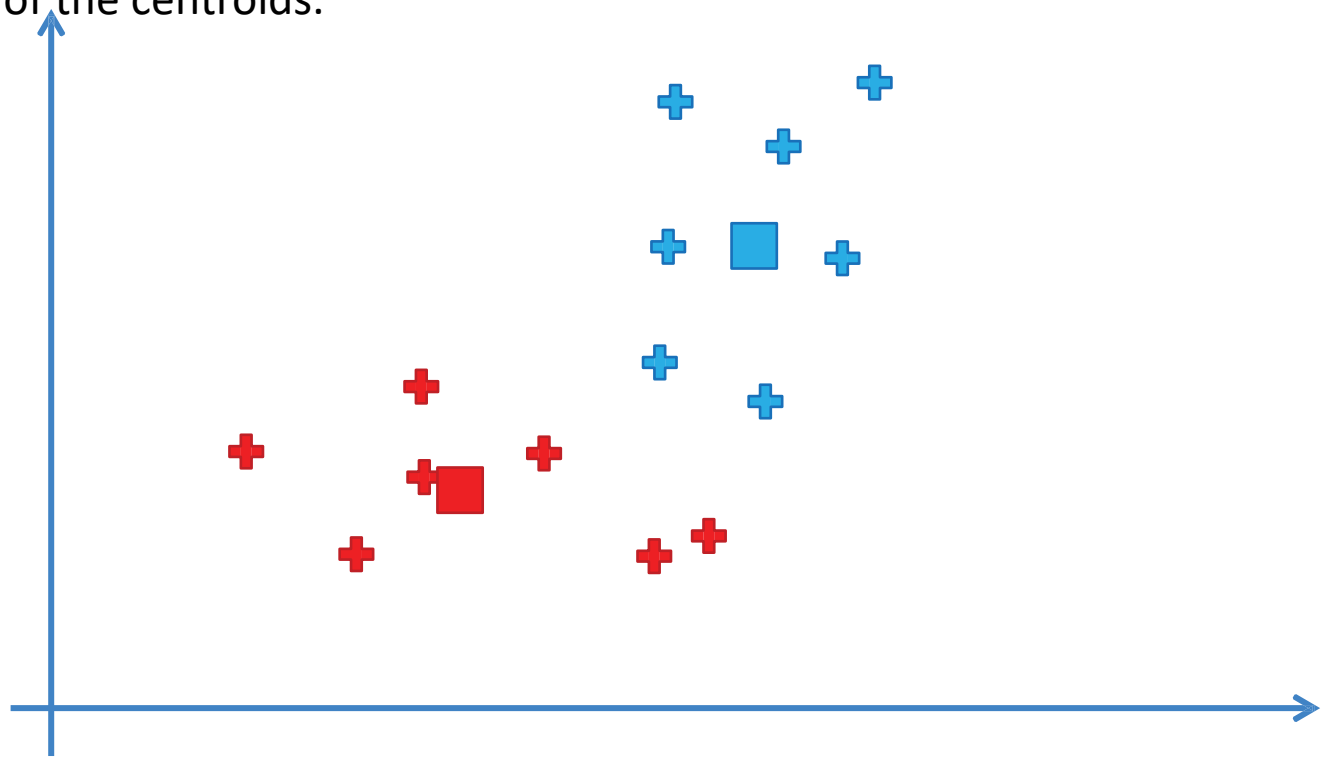
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering

Step 6: Iterating Until Convergence:

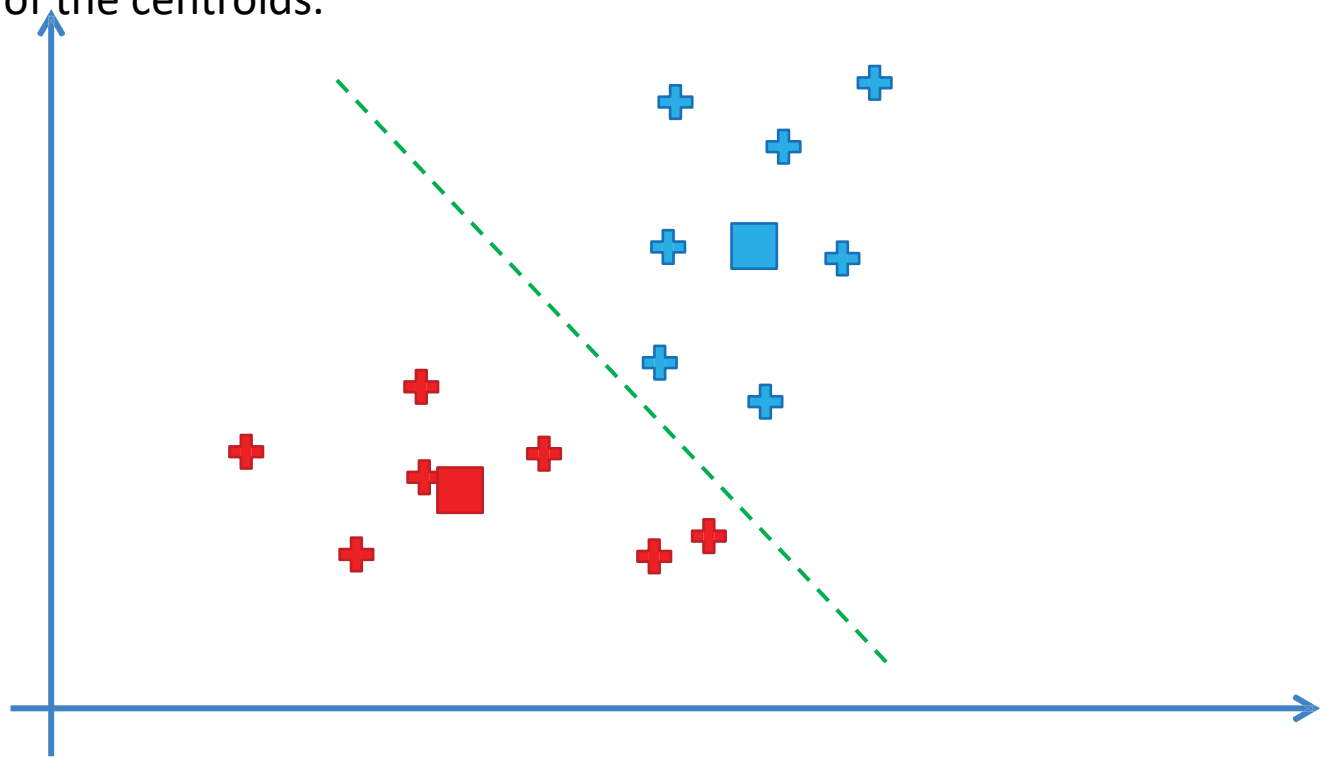
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



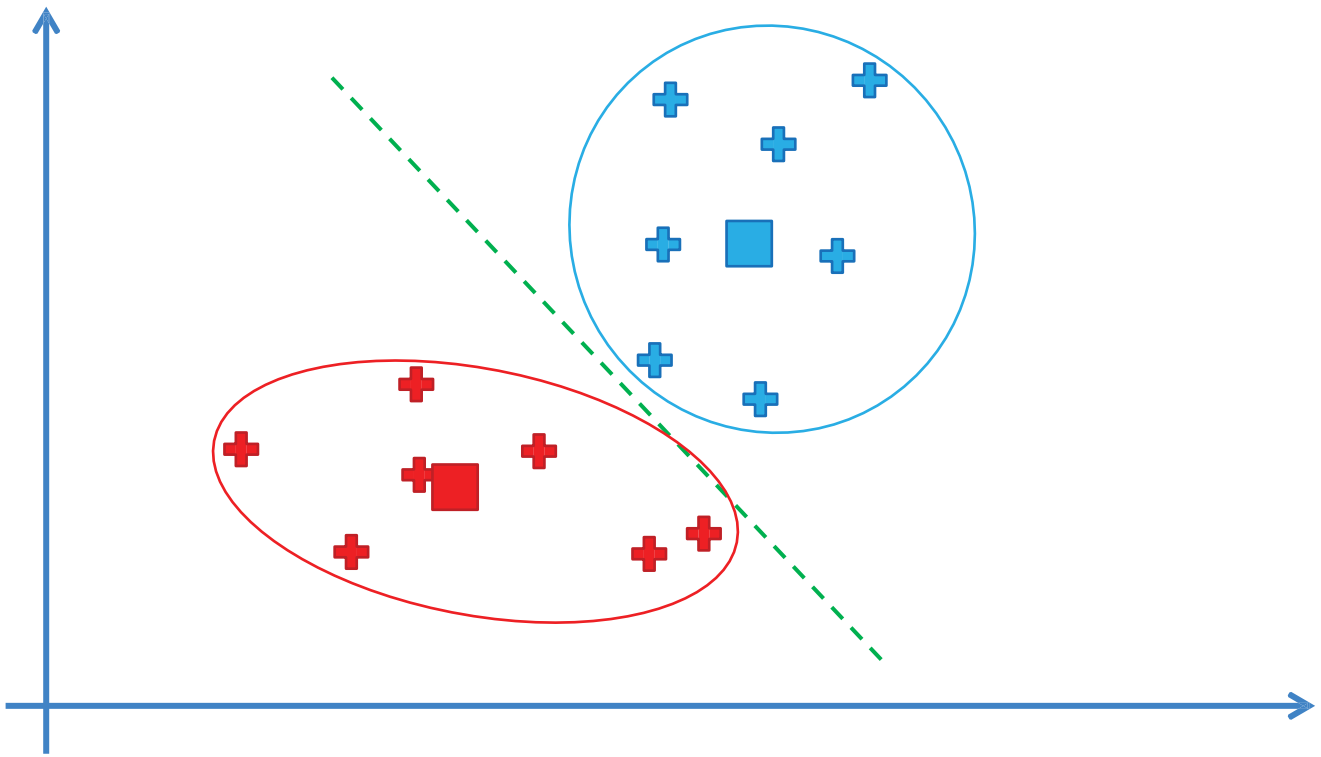
K-Means Clustering

Step 6: Iterating Until Convergence:

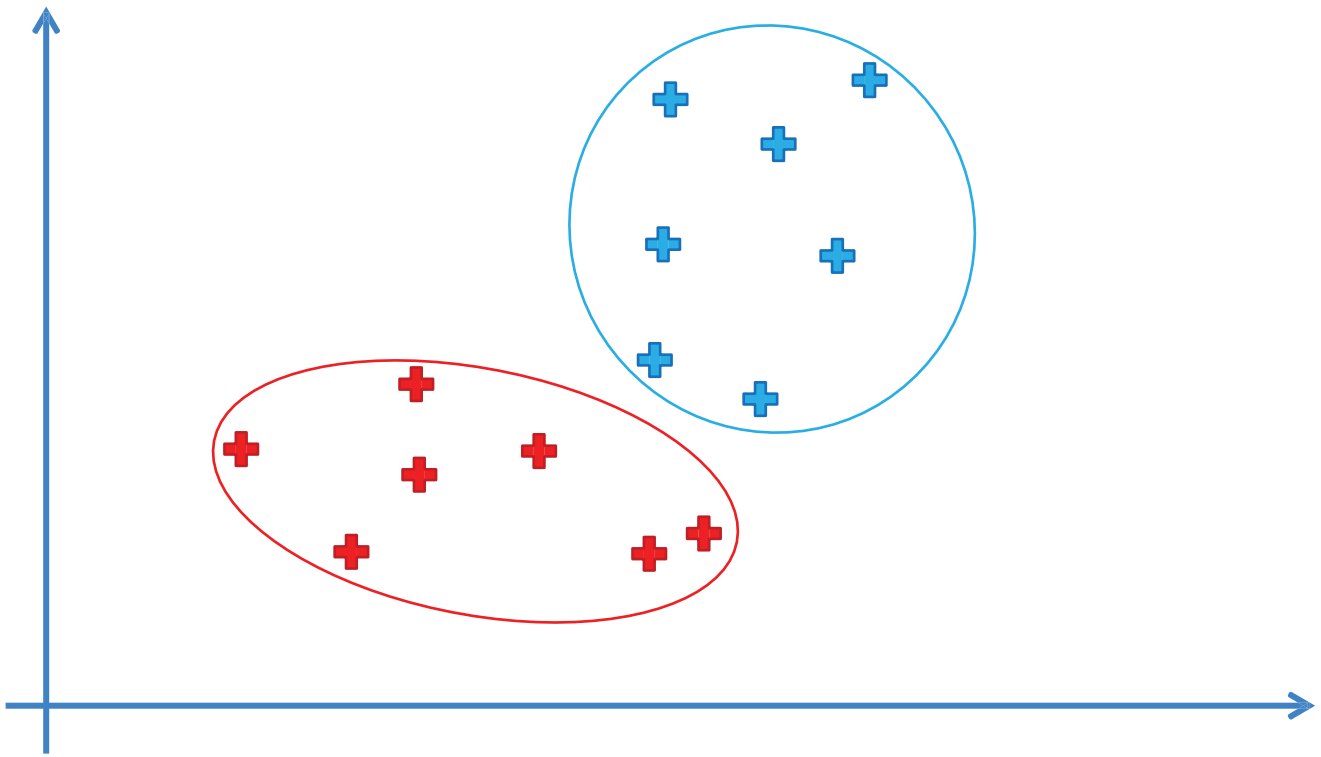
•Continue the process of reassignment, center of mass calculation, and centroid movement iteratively. The algorithm stops when centroids stabilize, meaning further reassignments do not change cluster memberships, indicating convergence and finalization of the centroids.



K-Means Clustering



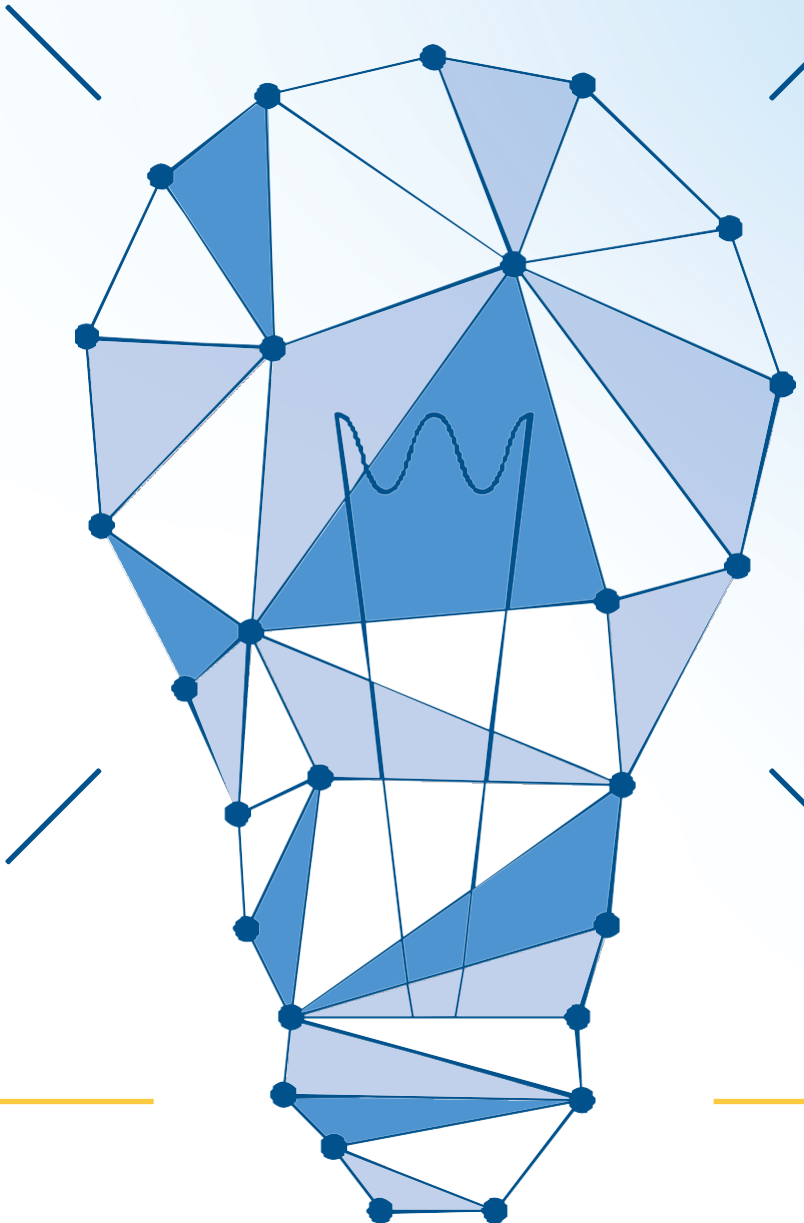
K-Means Clustering



Solved Example

Solved Example Video

The Elbow Method



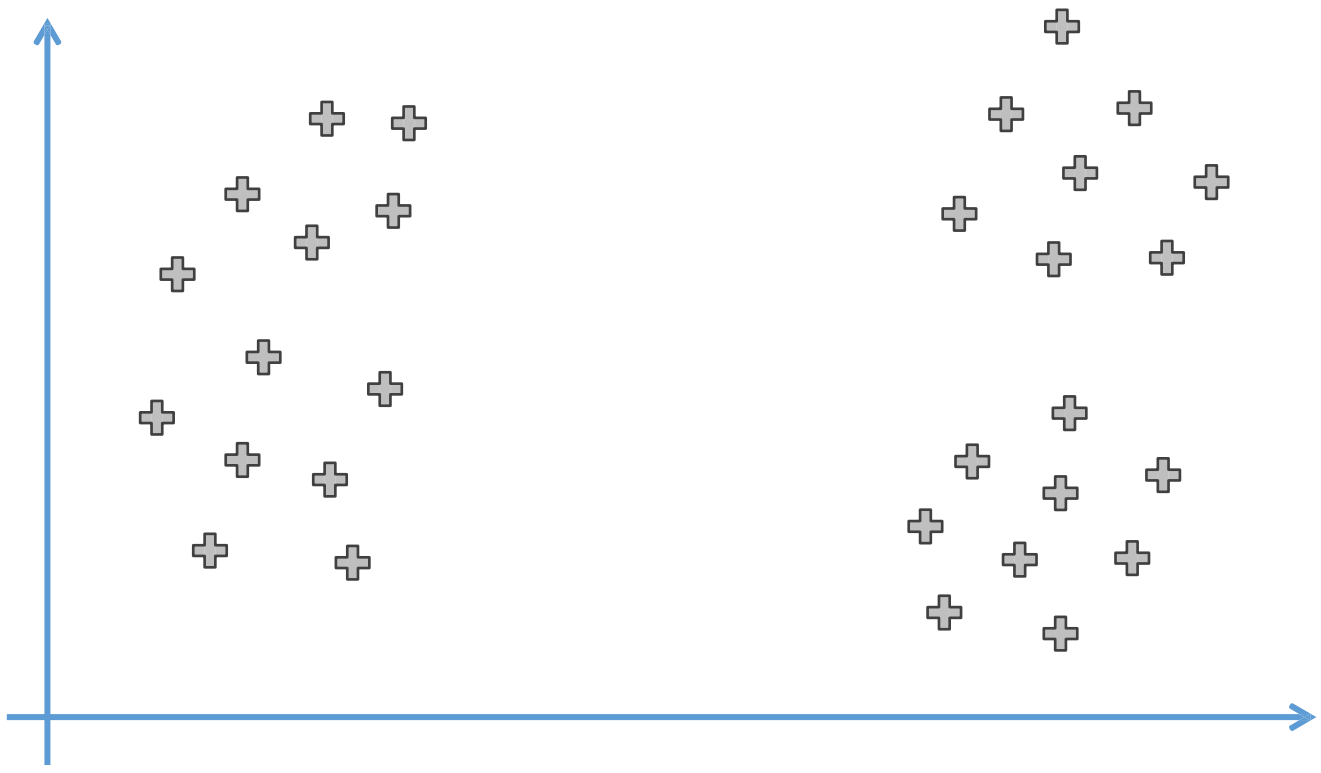
The Elbow Method

The Elbow Method

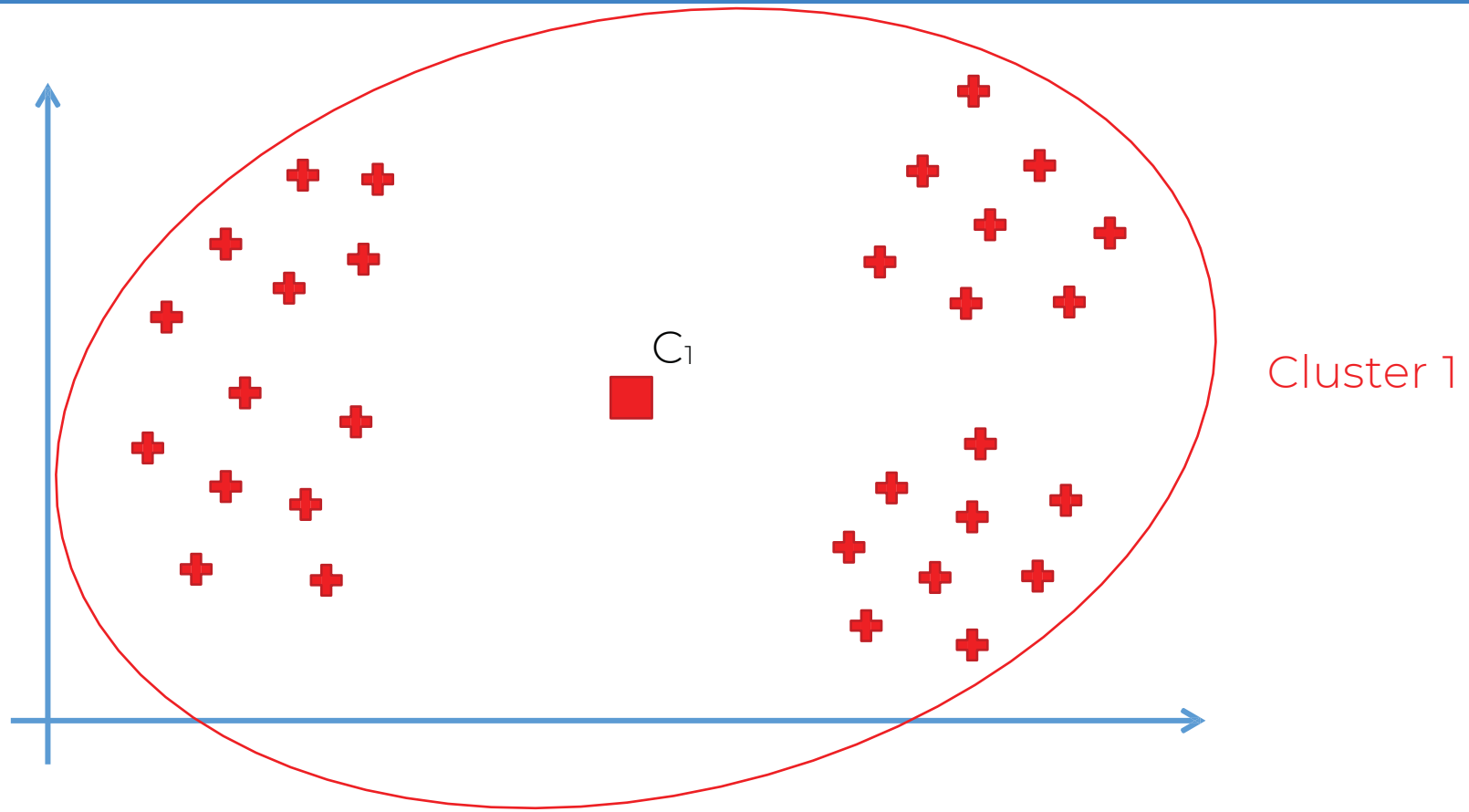
Within Cluster Sum of Squares:

$$\text{WCSS} = \sum_{\text{P}_\text{i} \text{ in Cluster 1}} \text{distance}(\text{P}_\text{i}, \text{C}_1)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 2}} \text{distance}(\text{P}_\text{i}, \text{C}_2)^2 + \ldots$$

The Elbow Method

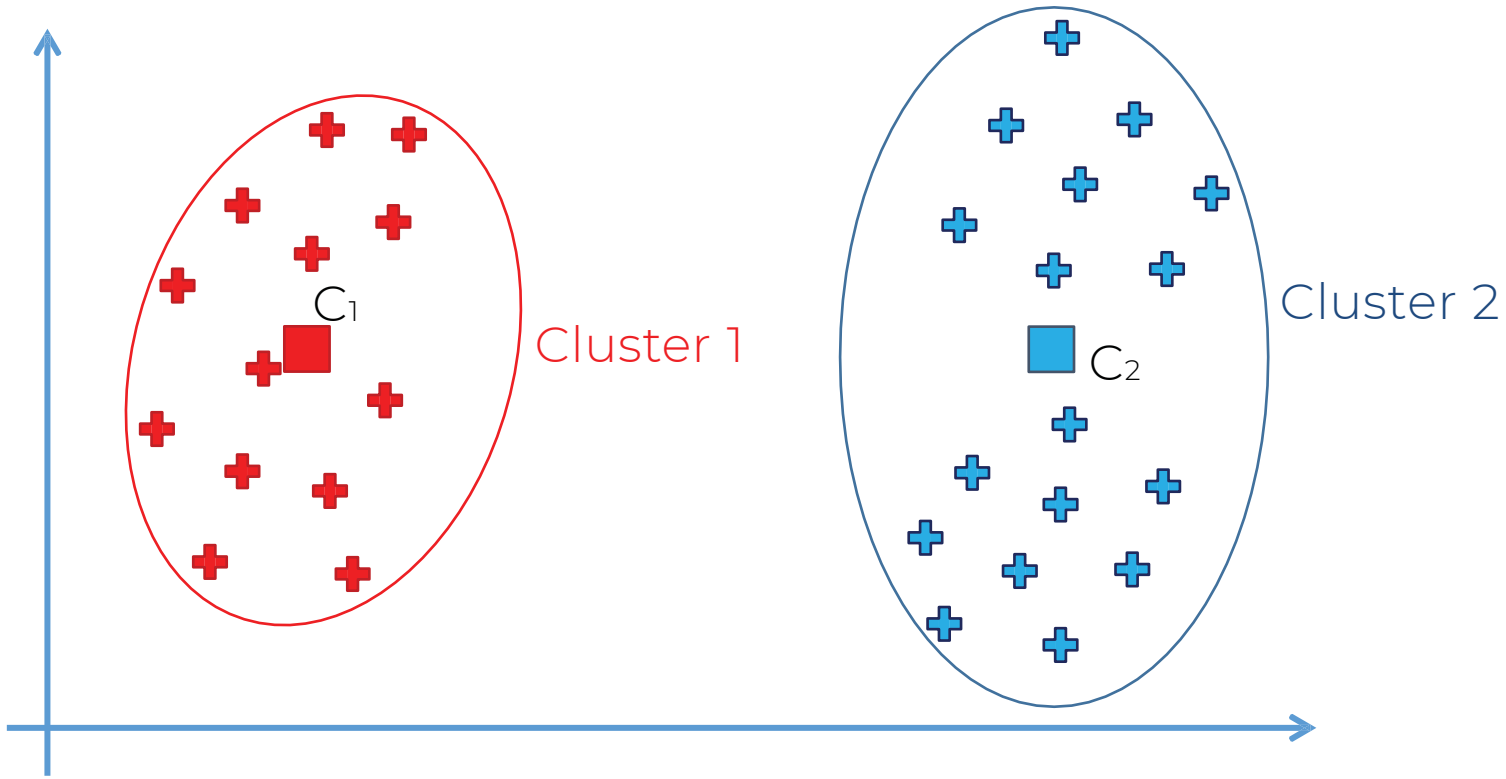


The Elbow Method



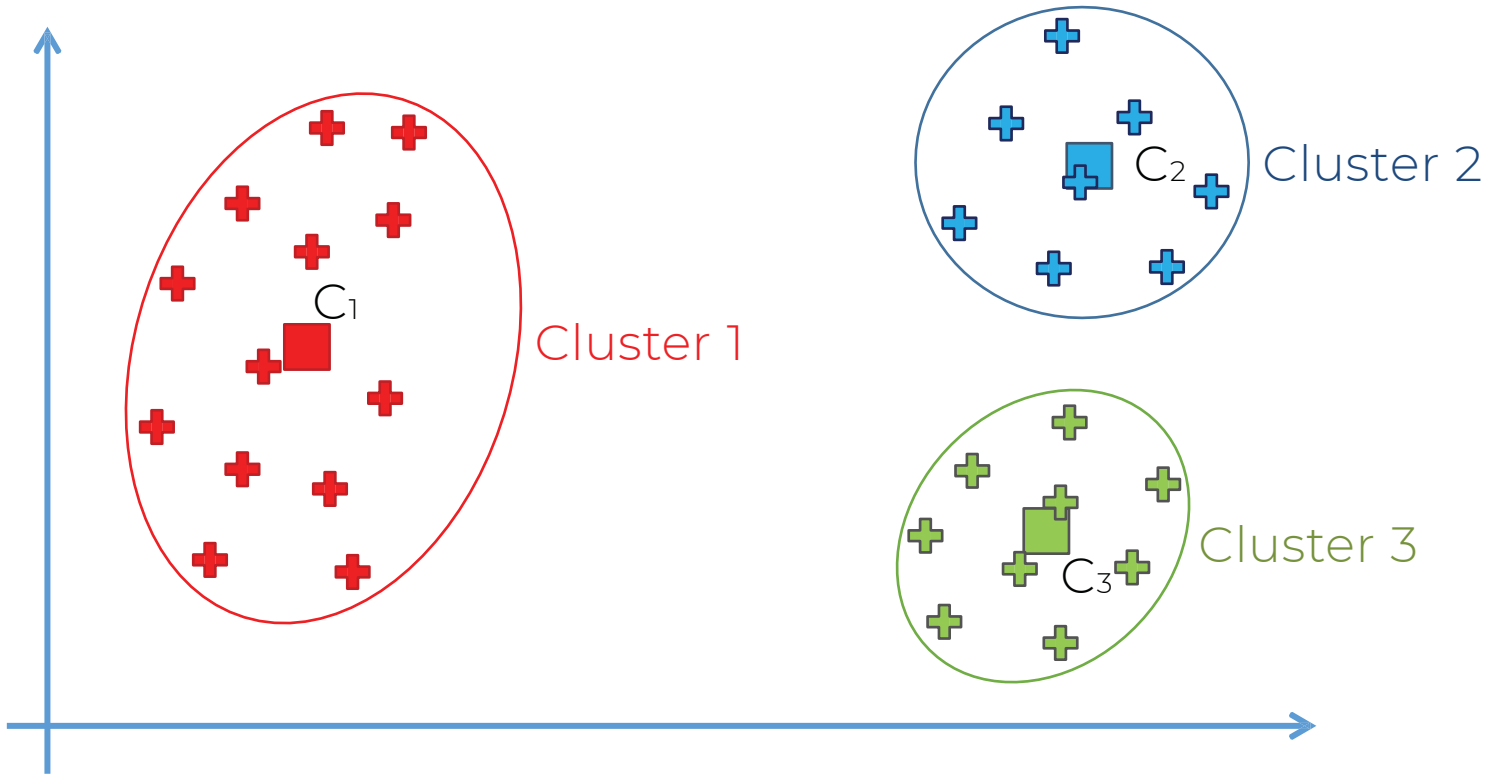
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2$$

The Elbow Method



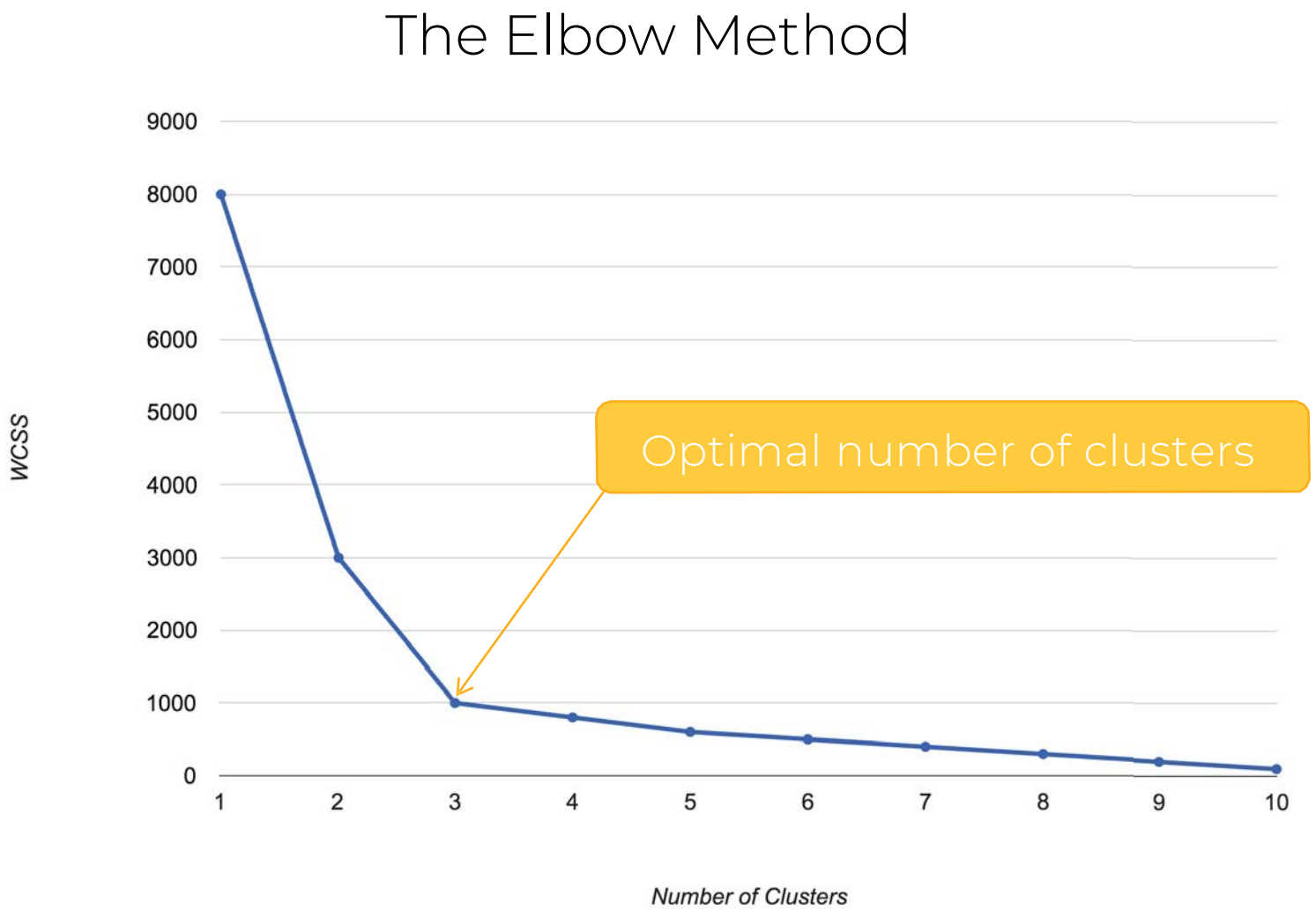
$$\text{WCSS} = \sum_{\text{P}_\text{i} \text{ in Cluster 1}} \text{distance}(\text{P}_\text{i}, \text{C}_1)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 2}} \text{distance}(\text{P}_\text{i}, \text{C}_2)^2$$

The Elbow Method



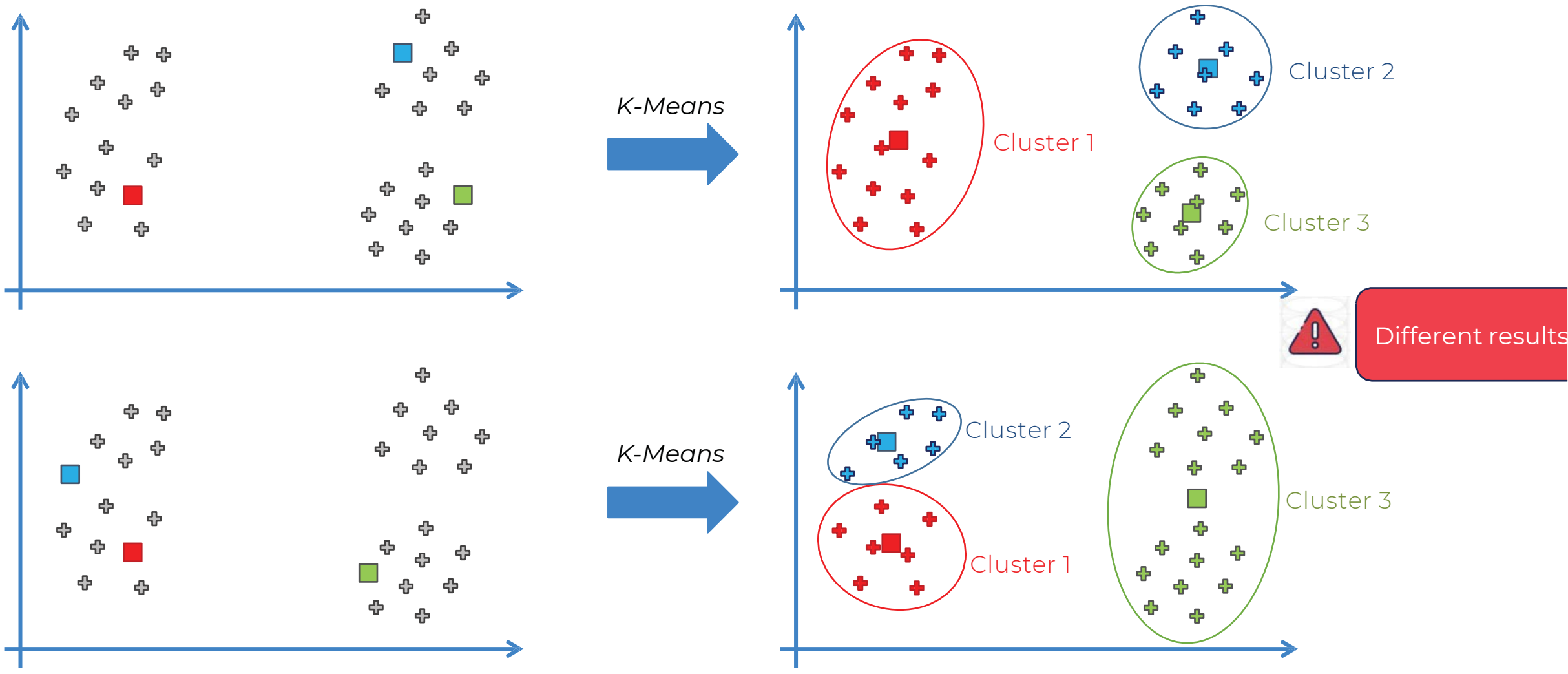
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

The Elbow Method



K-Means++

K-Means++


K-Means
Cluster 1
Cluster 2
Cluster 3
Different results
K-Means
Cluster 2
Cluster 3
Cluster 1

K-Means++

K-Means++ Initialization Algorithm:

Step 1: Choose first centroid at random among data points

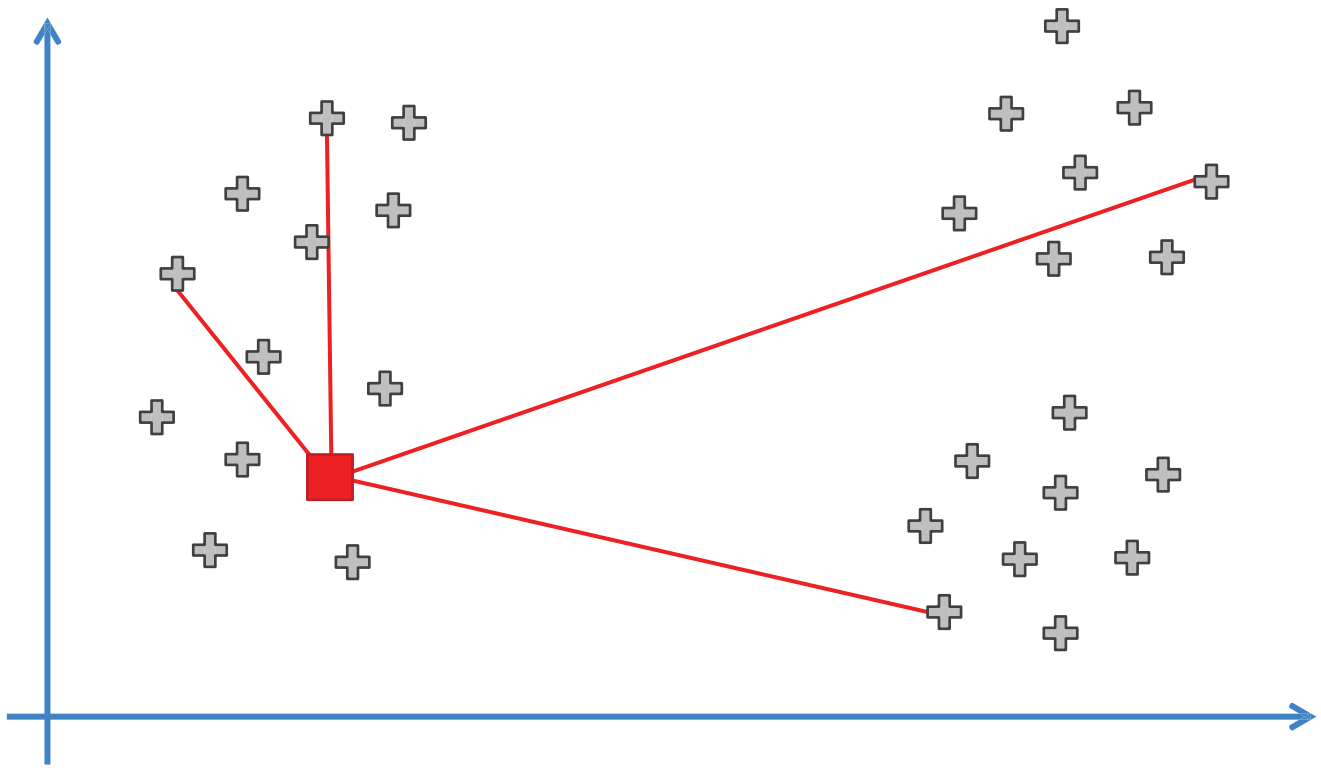
Step 2: For each of the remaining data points compute the distance (D) to the <u>nearest</u> out of already selected centroids

Step 3: Choose next centroid among remaining data points using <u>weighted</u> random selection – weighted by D^2

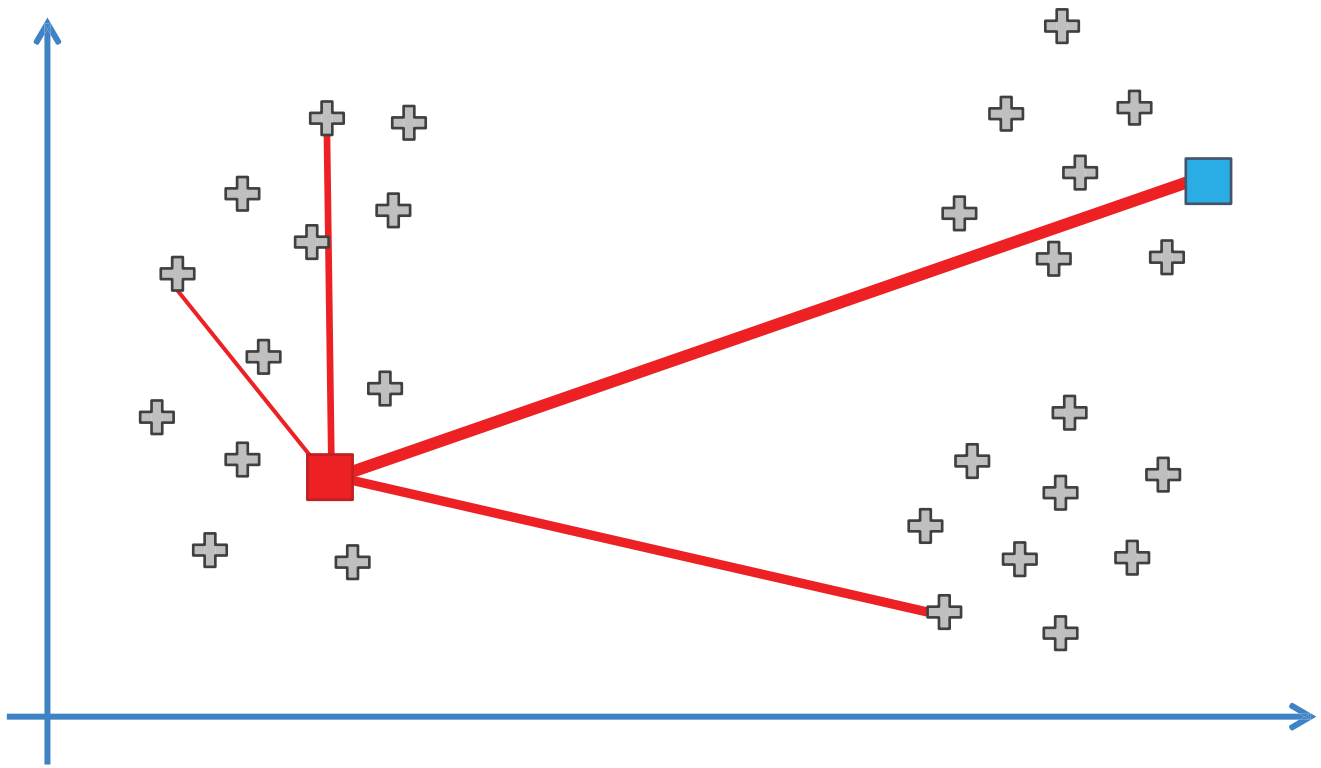
Step 4: Repeat Steps 2 and 3 until all k centroids have been selected

Step 5: Proceed with standard k-means clustering

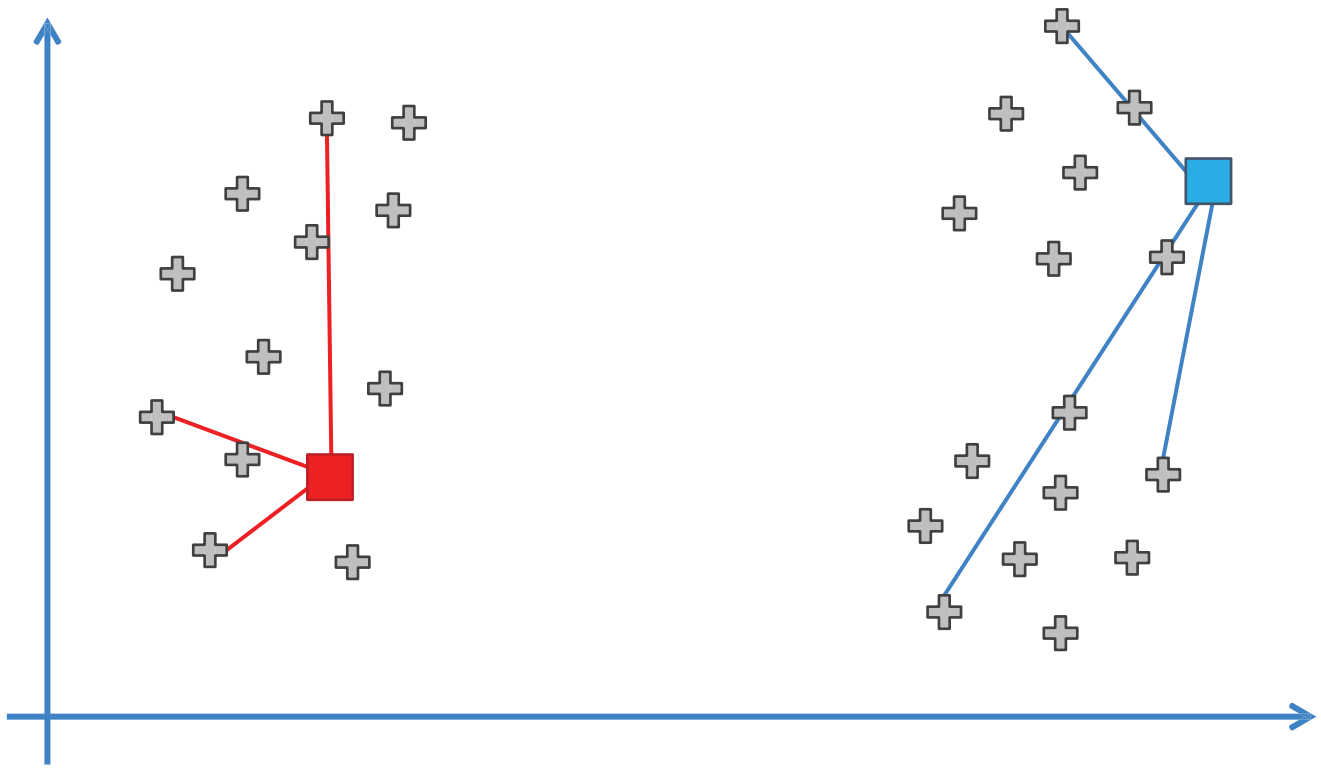
K-Means++



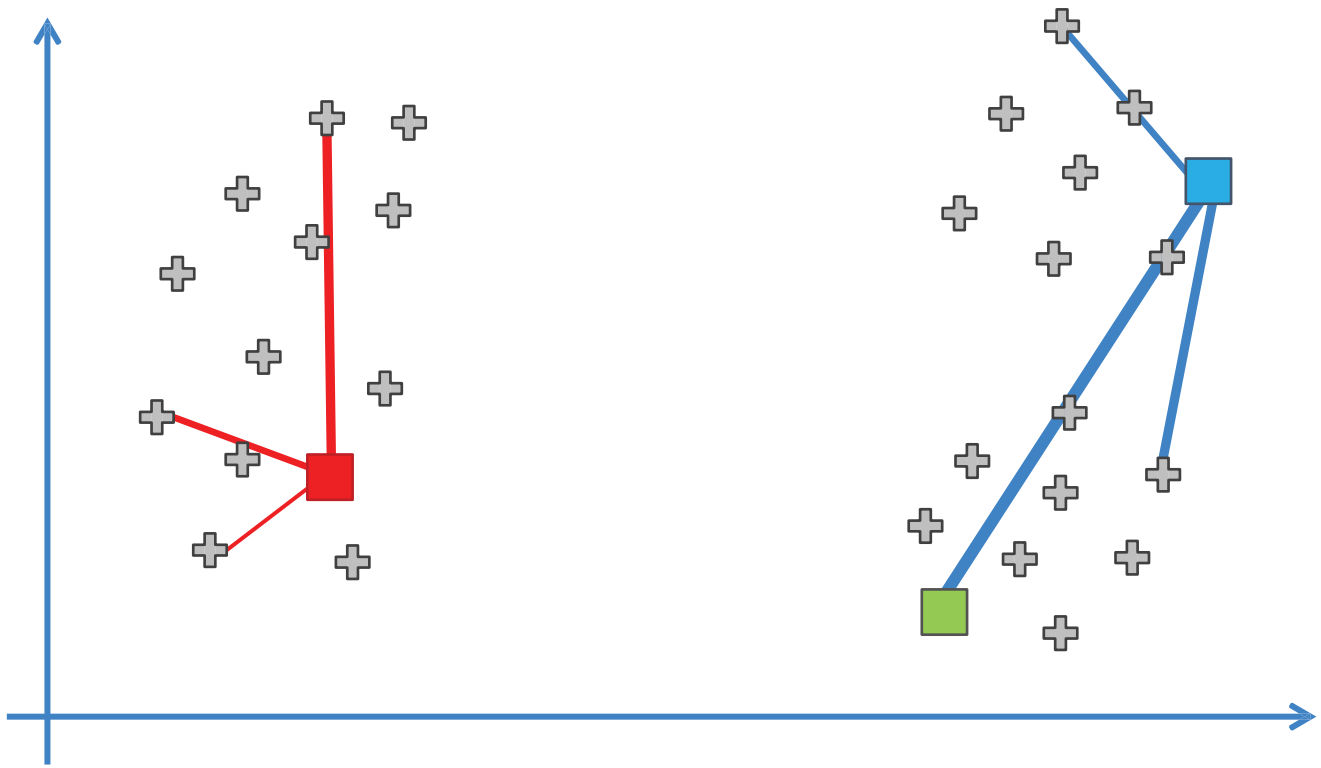
K-Means++



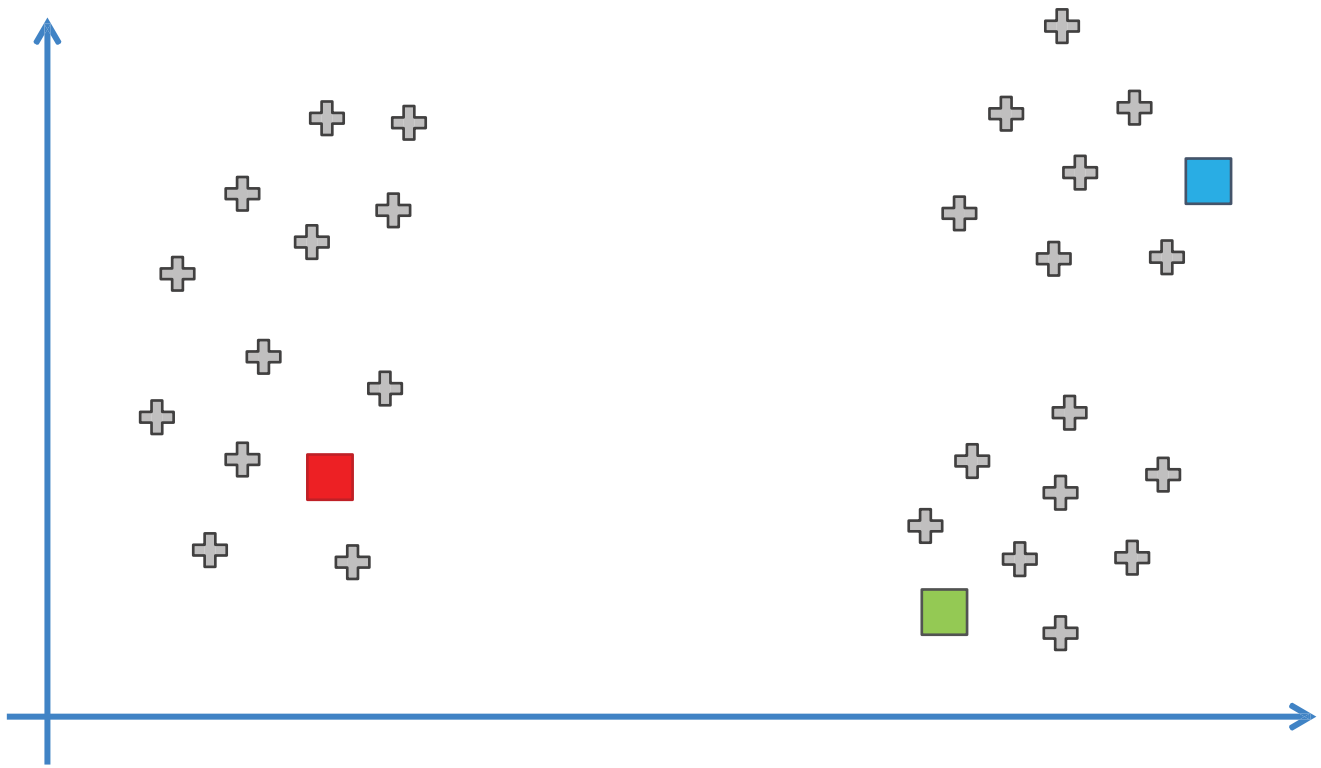
K-Means++



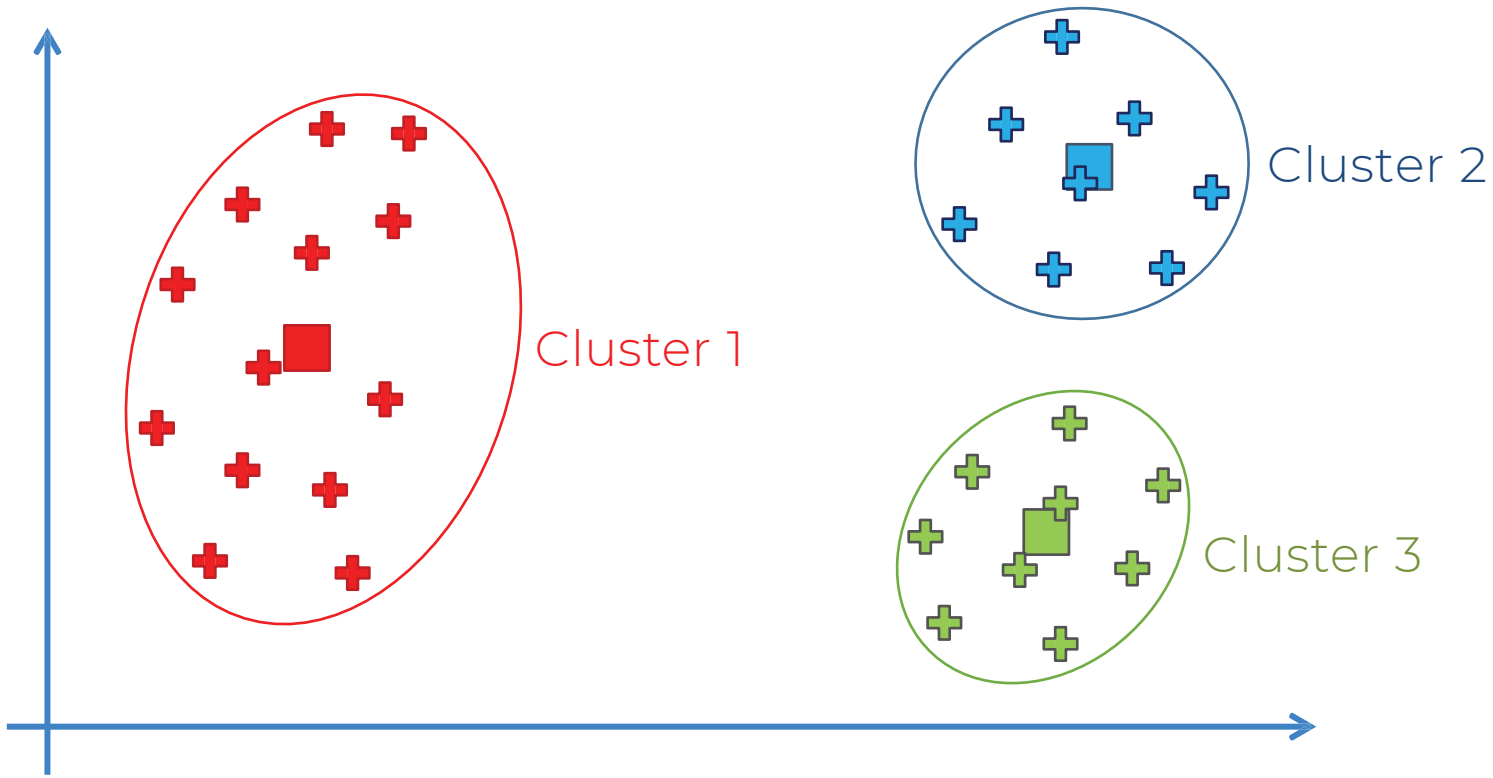
K-Means++



K-Means++

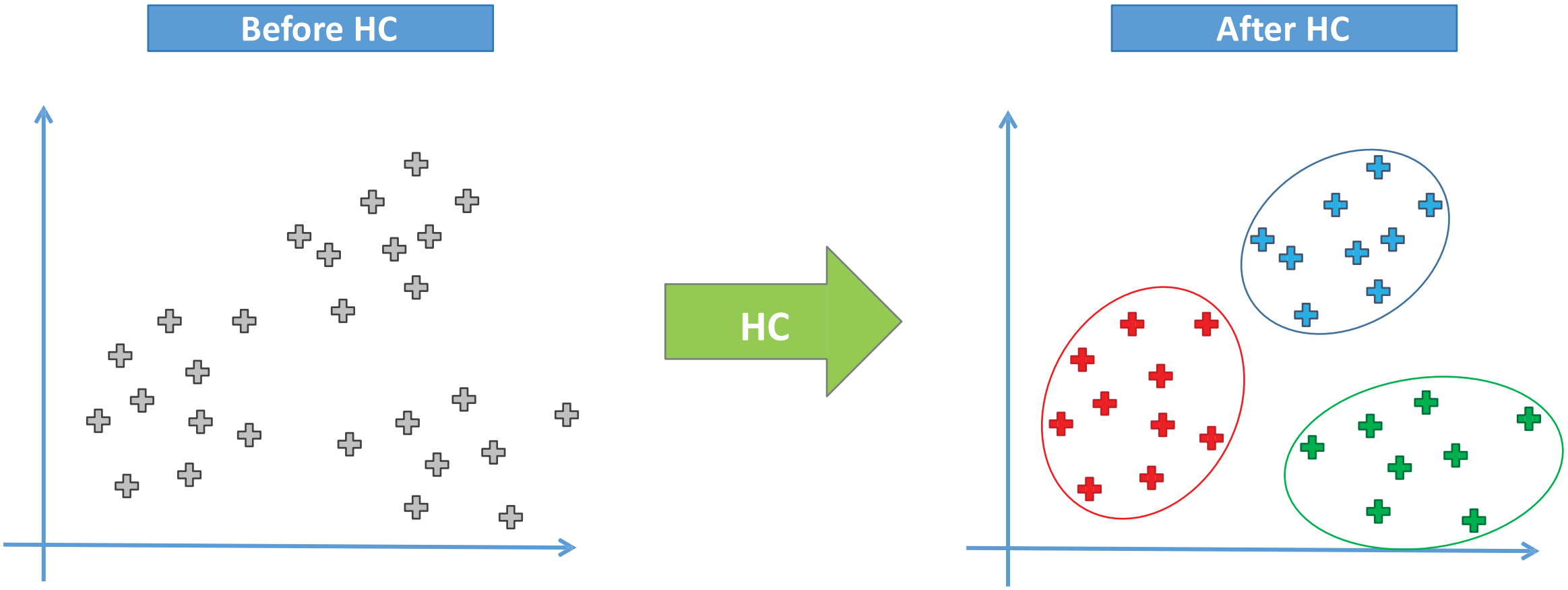


K-Means++



HC Intuition: Understanding HC

What HC does for you



Same as K-Means but different process

NOTE:
Agglomerative
&
Divisive

Agglomerative HC

STEP 1: Make each data point a single-point cluster ➡ That forms N clusters

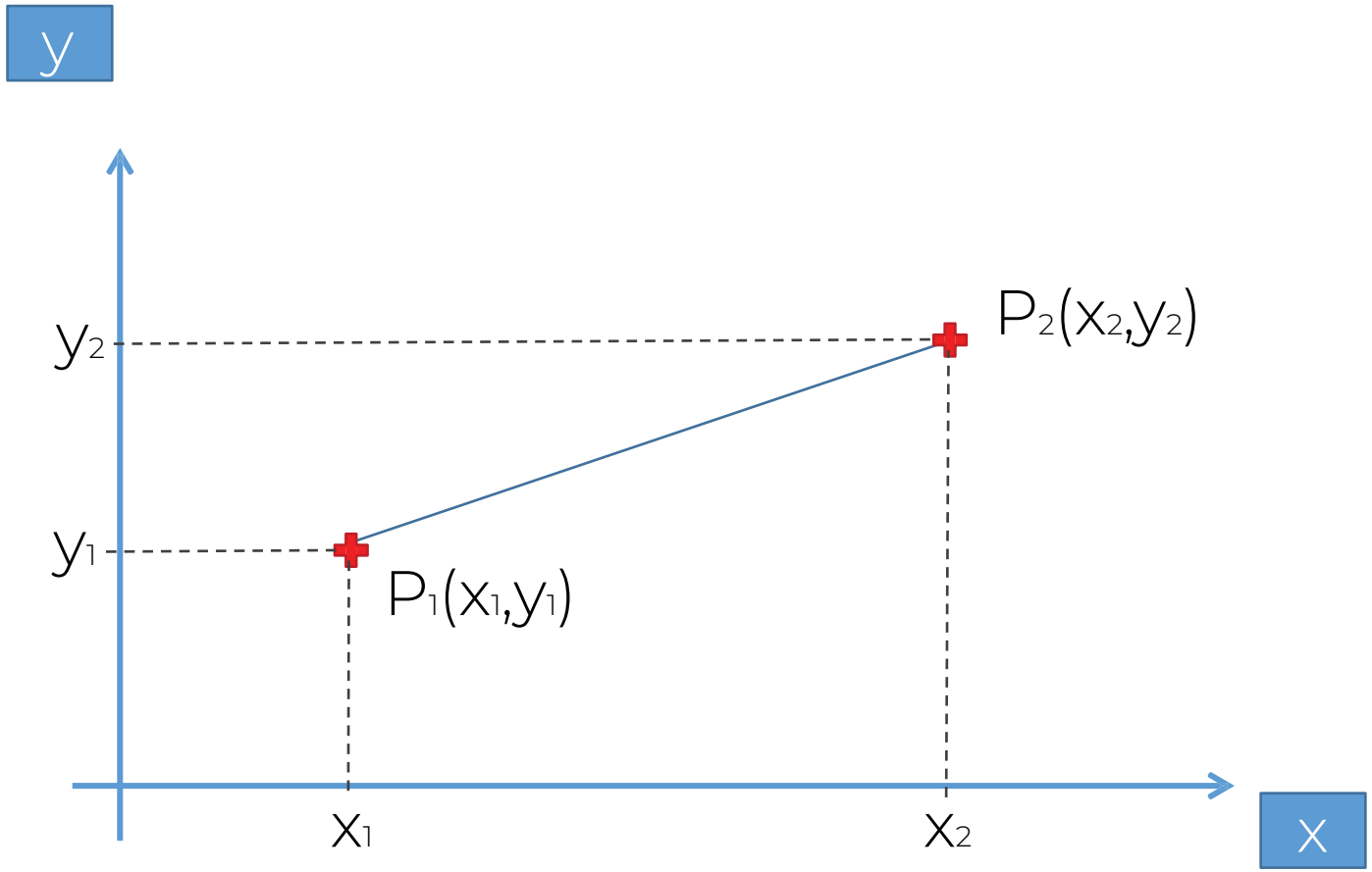
STEP 2: Take the two closest data points and make them one cluster ➡ That forms N-1 clusters

STEP 3: Take the two closest clusters and make them one cluster ➡ That forms N - 2 clusters

STEP 4: Repeat STEP 3 until there is only one cluster

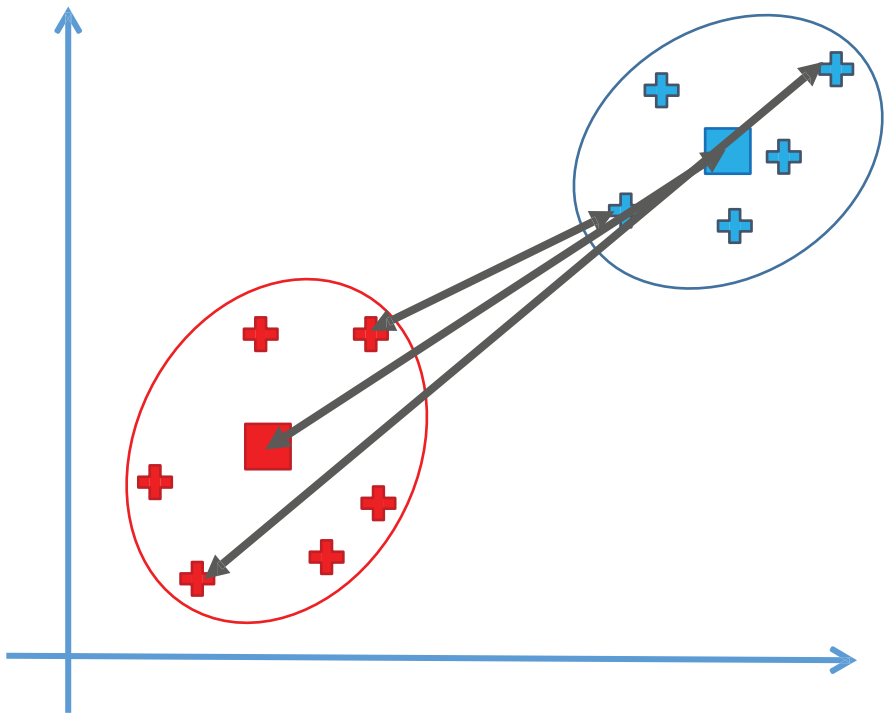
FIN

Euclidean Distance



$$\text{Euclidean Distance between } P_1 \text{ and } P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distance Between Clusters

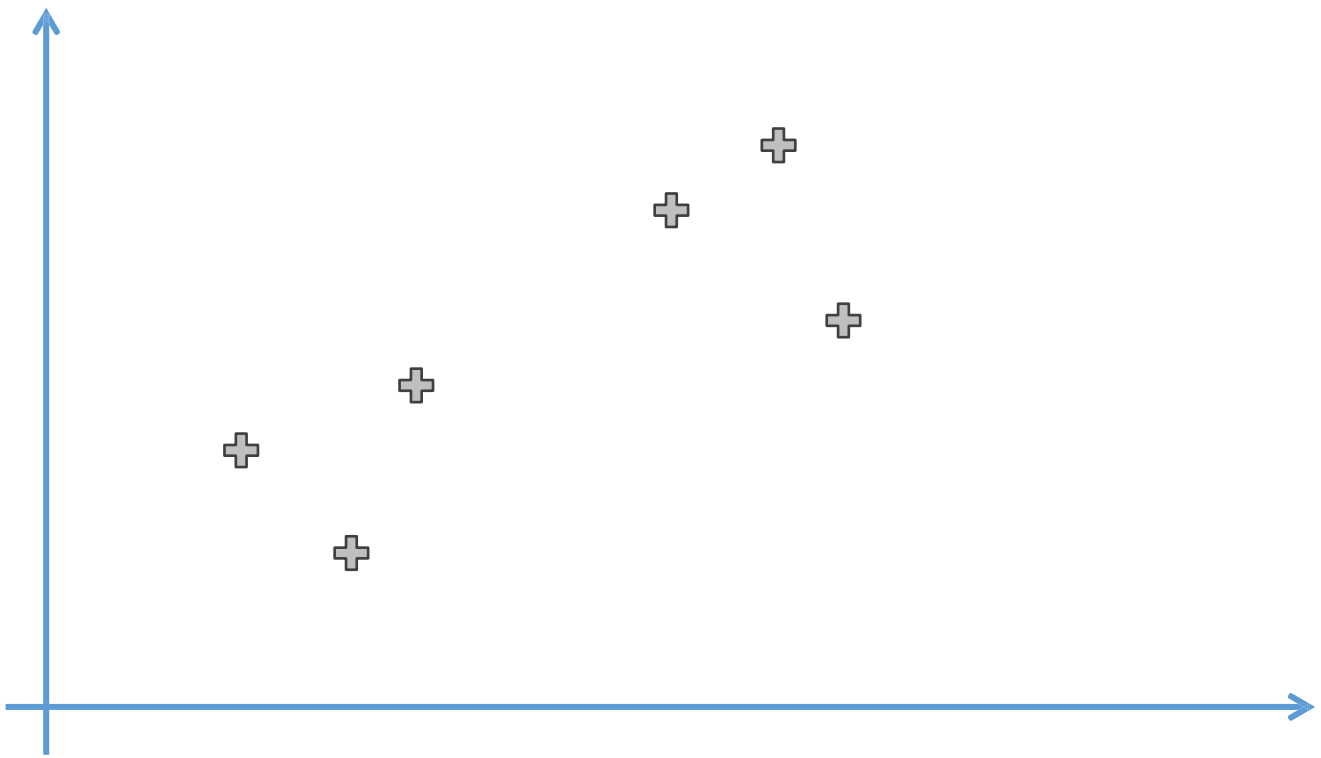


Distance Between Two Clusters:

- Option 1: Closest Points
- Option 2: Furthest Points
- Option 3: Average Distance
- Option 4: Distance Between Centroids

Agglomerative HC

Consider the following dataset of N = 6 data points

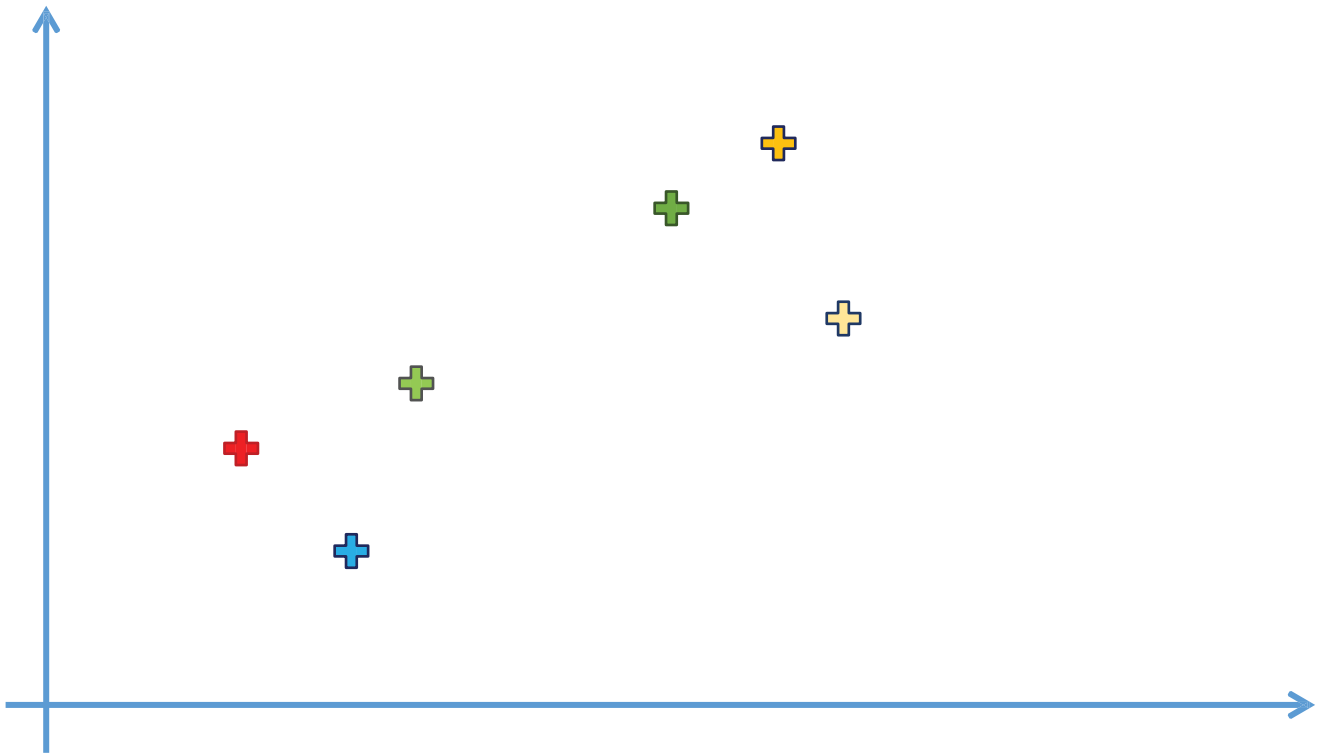


Agglomerative HC

STEP 1: Make each data point a single-point cluster ➡ That forms 6 clusters

Agglomerative HC

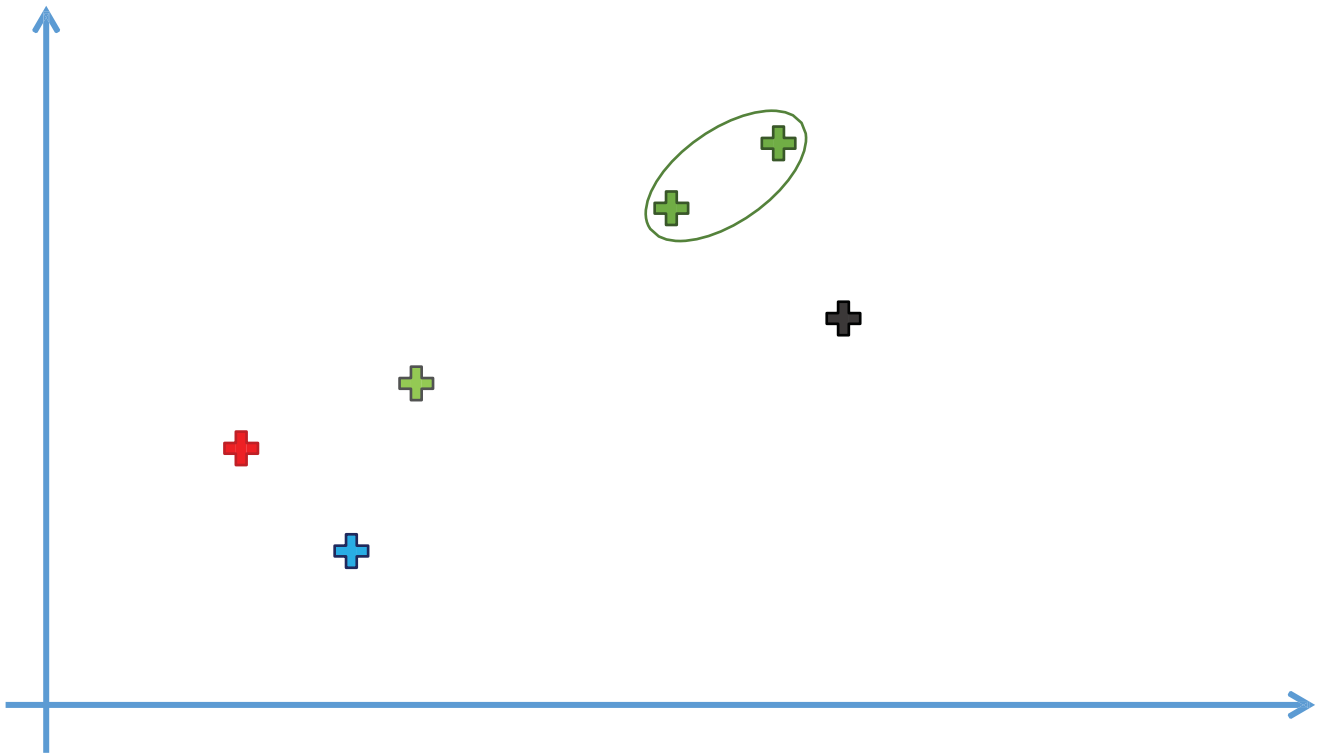
STEP 1: Make each data point a single-point cluster ➡ That forms 6 clusters



Agglomerative HC

STEP 2: Take the two closest data points and make them one cluster

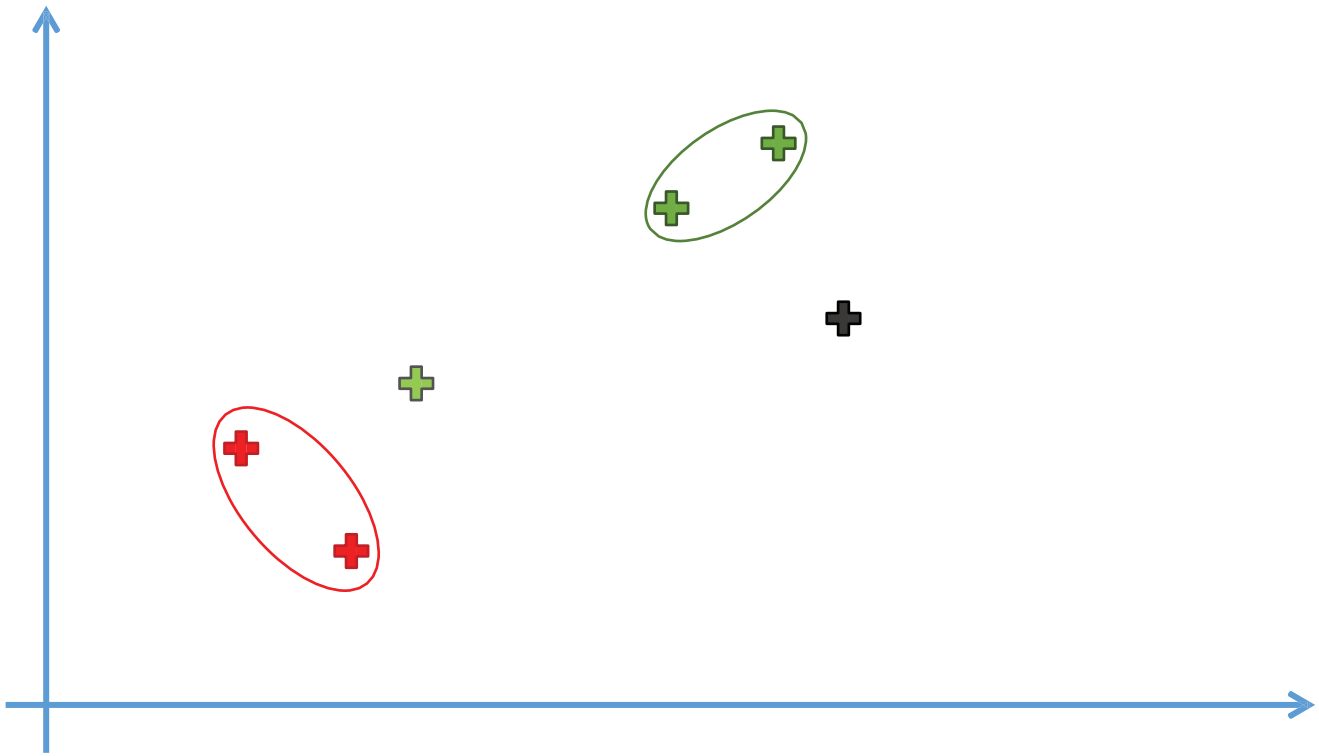
➡ That forms 5 clusters



Agglomerative HC

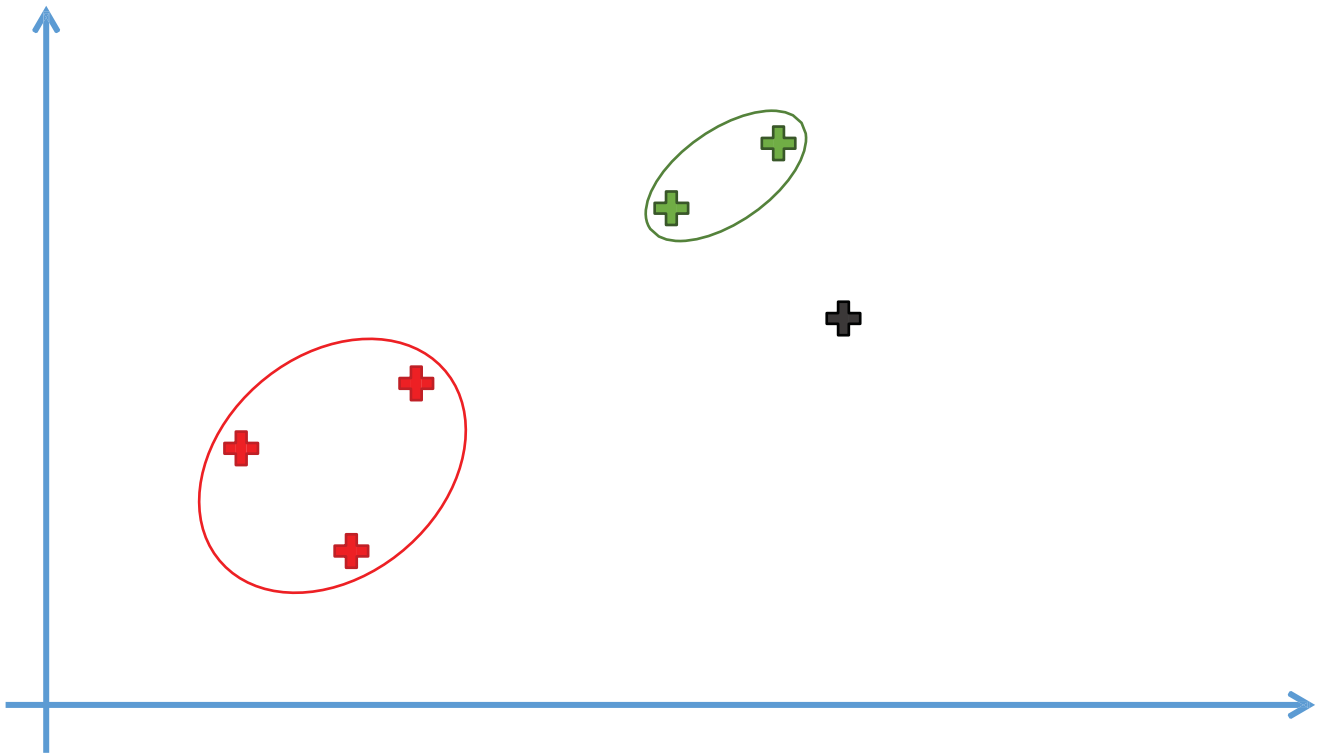
STEP 3: Take the two closest clusters and make them one cluster

➡That forms 4 clusters



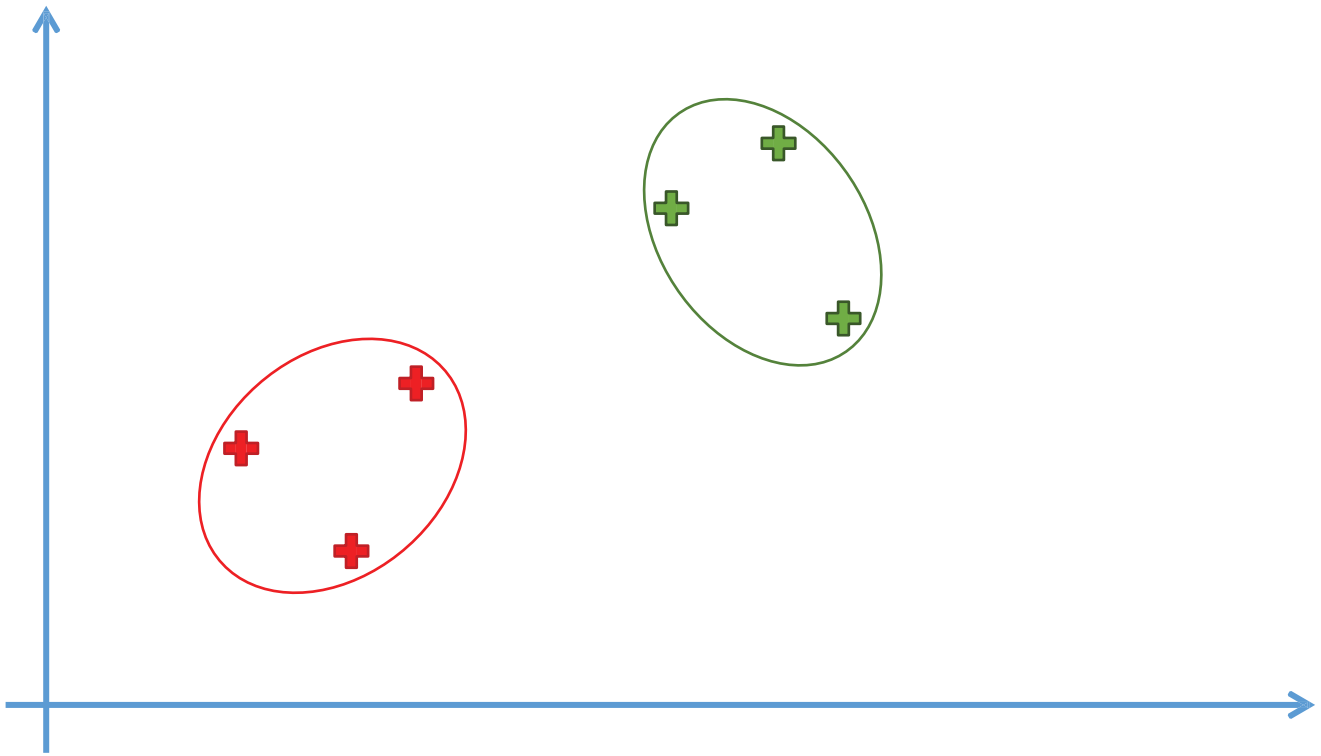
Agglomerative HC

STEP 4: Repeat STEP 3 until there is only one cluster



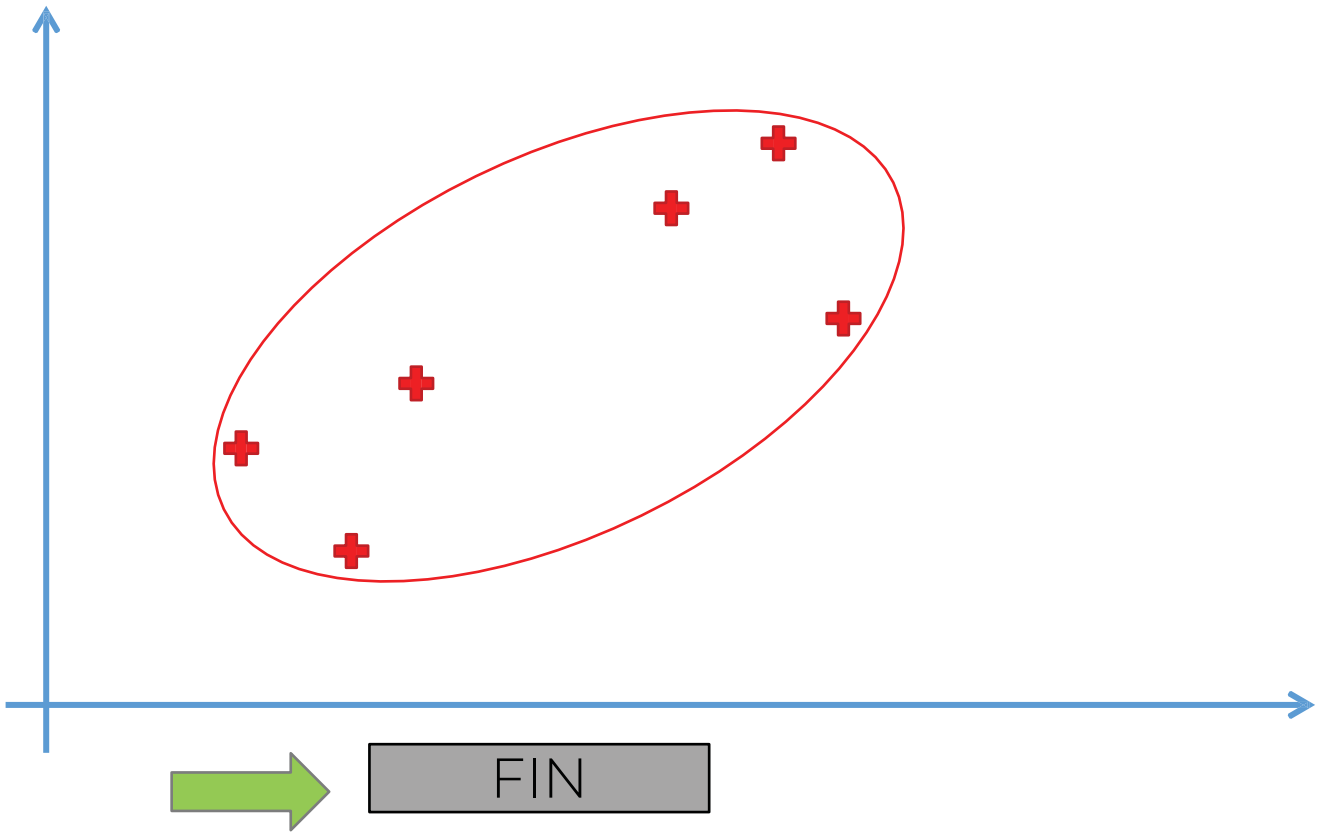
Agglomerative HC

STEP 4: Repeat STEP 3 until there is only one cluster



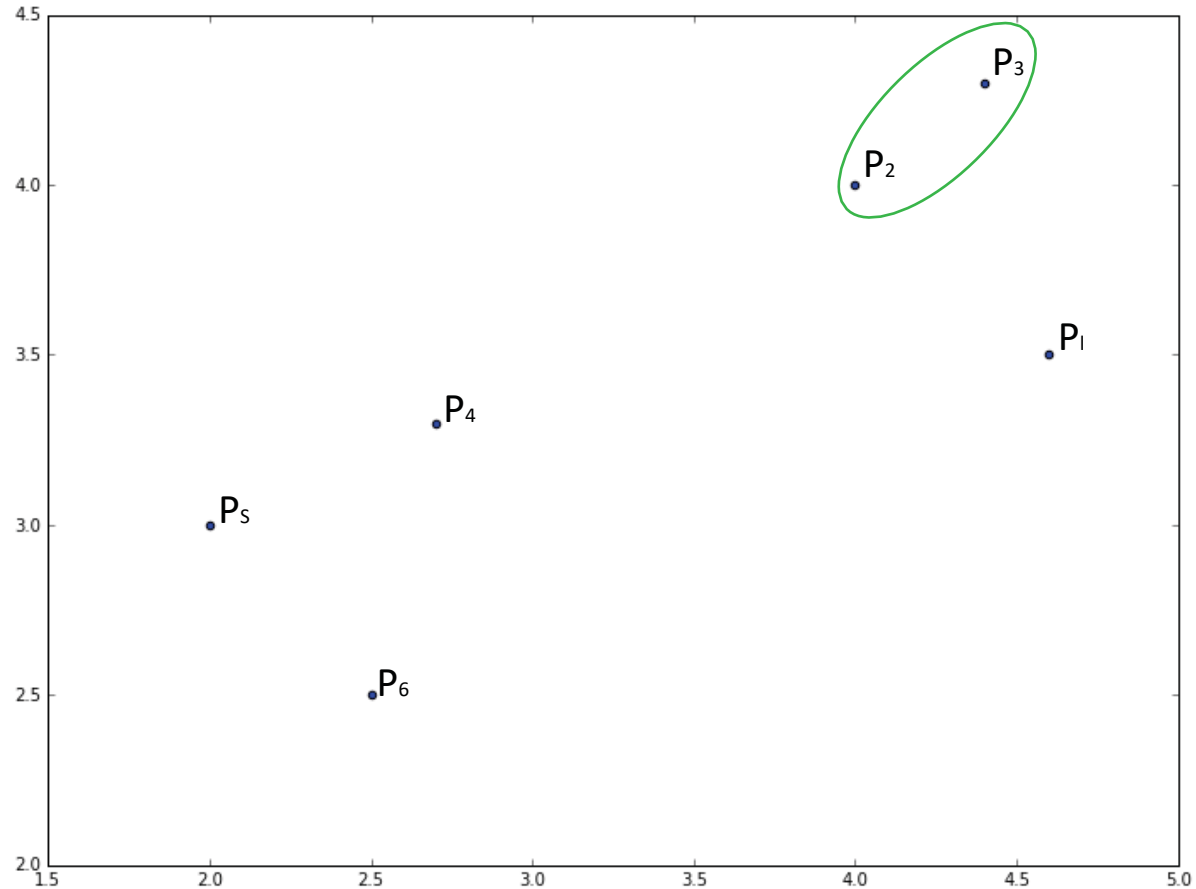
Agglomerative HC

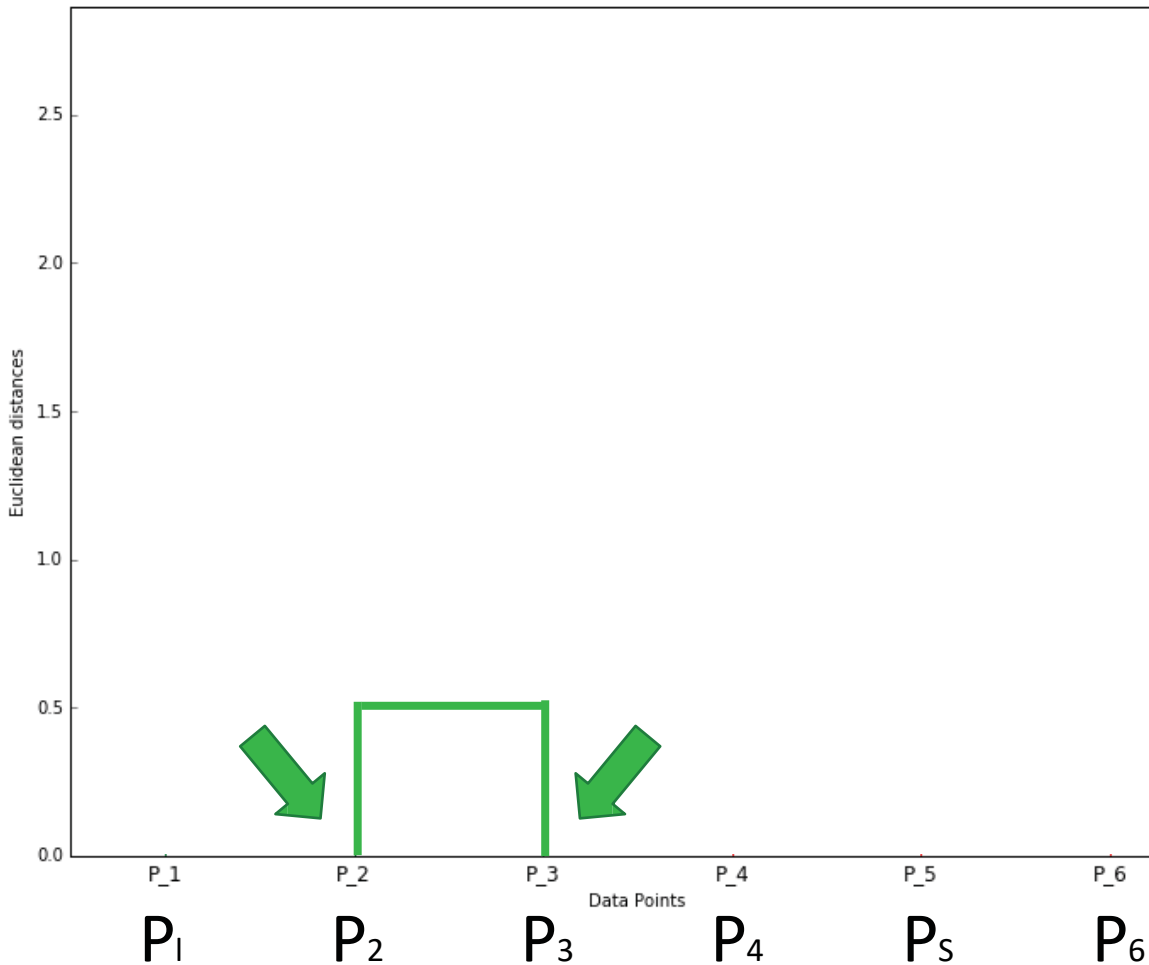
STEP 4: Repeat STEP 3 until there is only one cluster



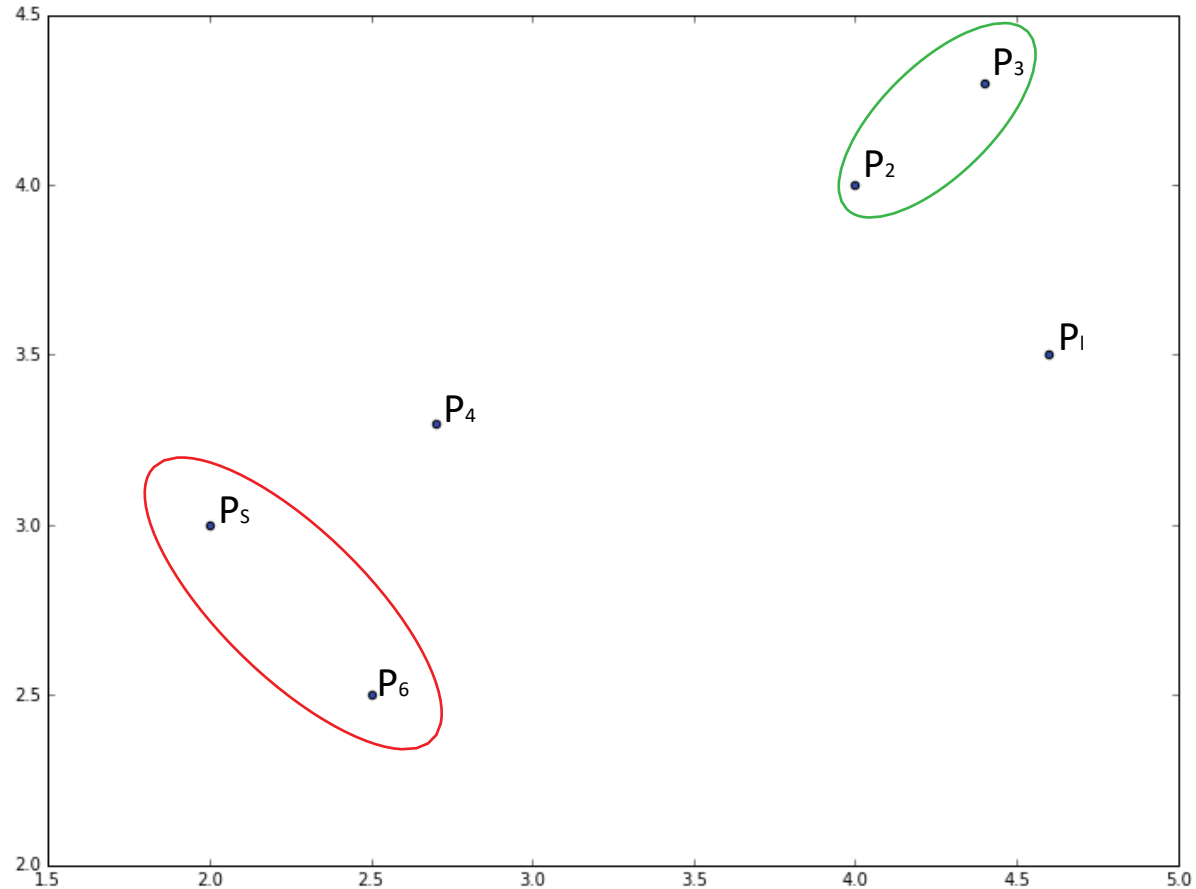
HC Intuition: How Do Dendograms Work?

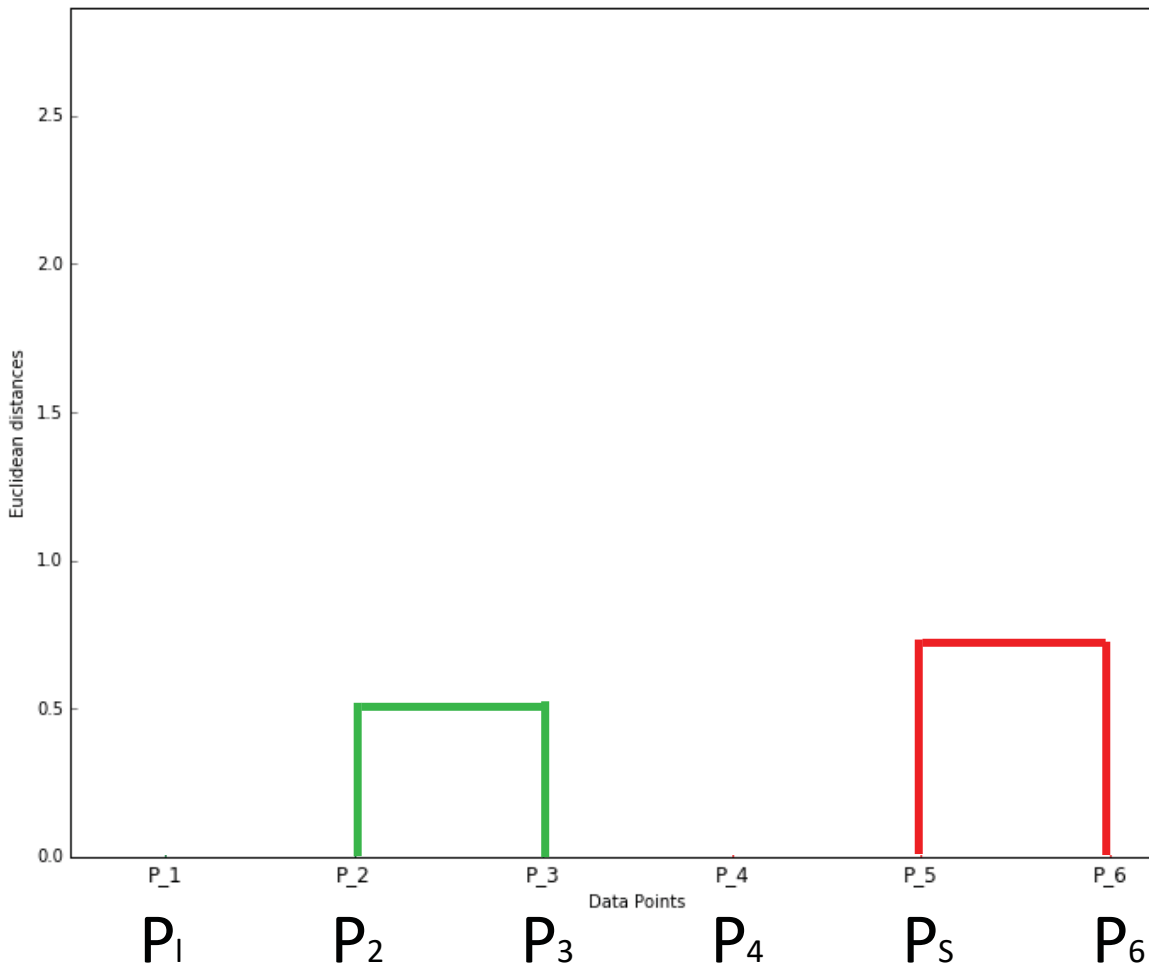
How Do Dendograms Work?



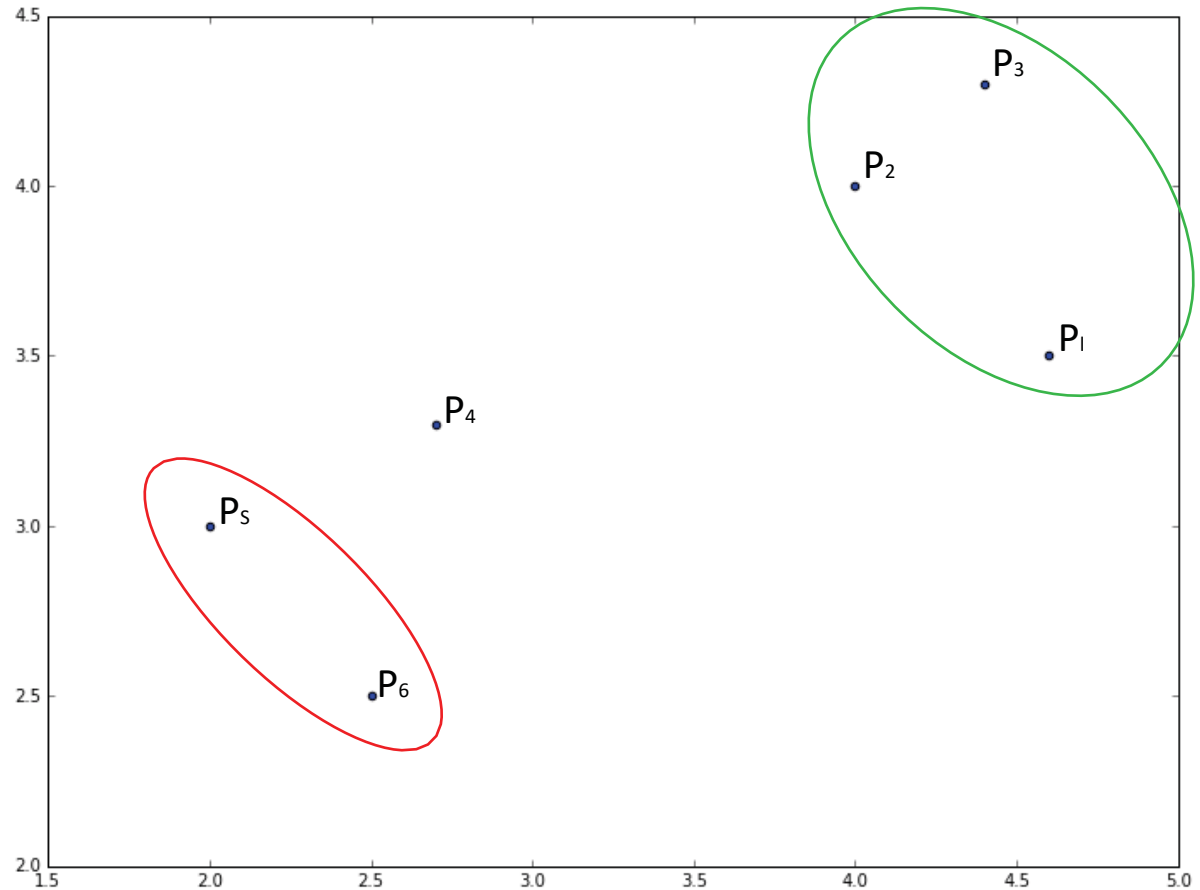


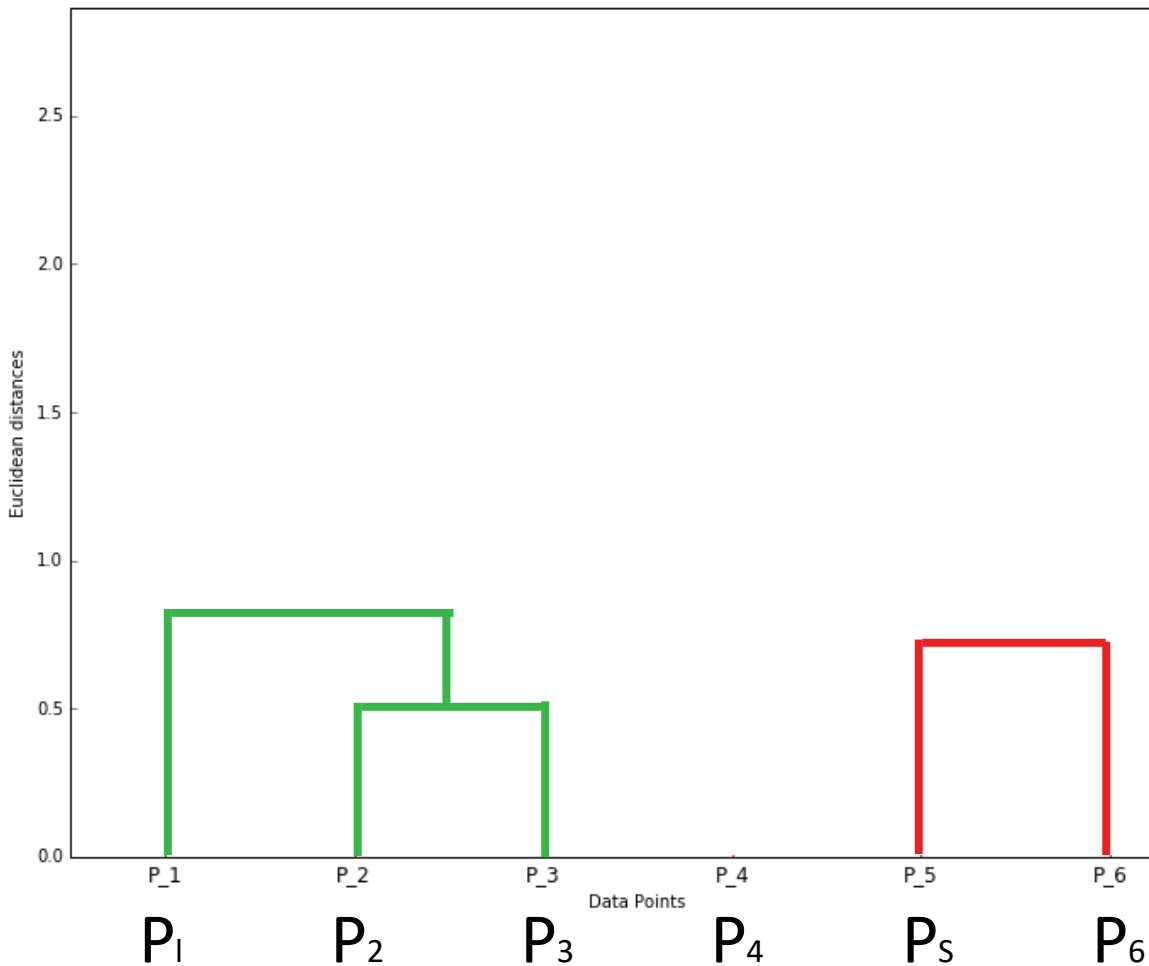
How Do Dendograms Work?



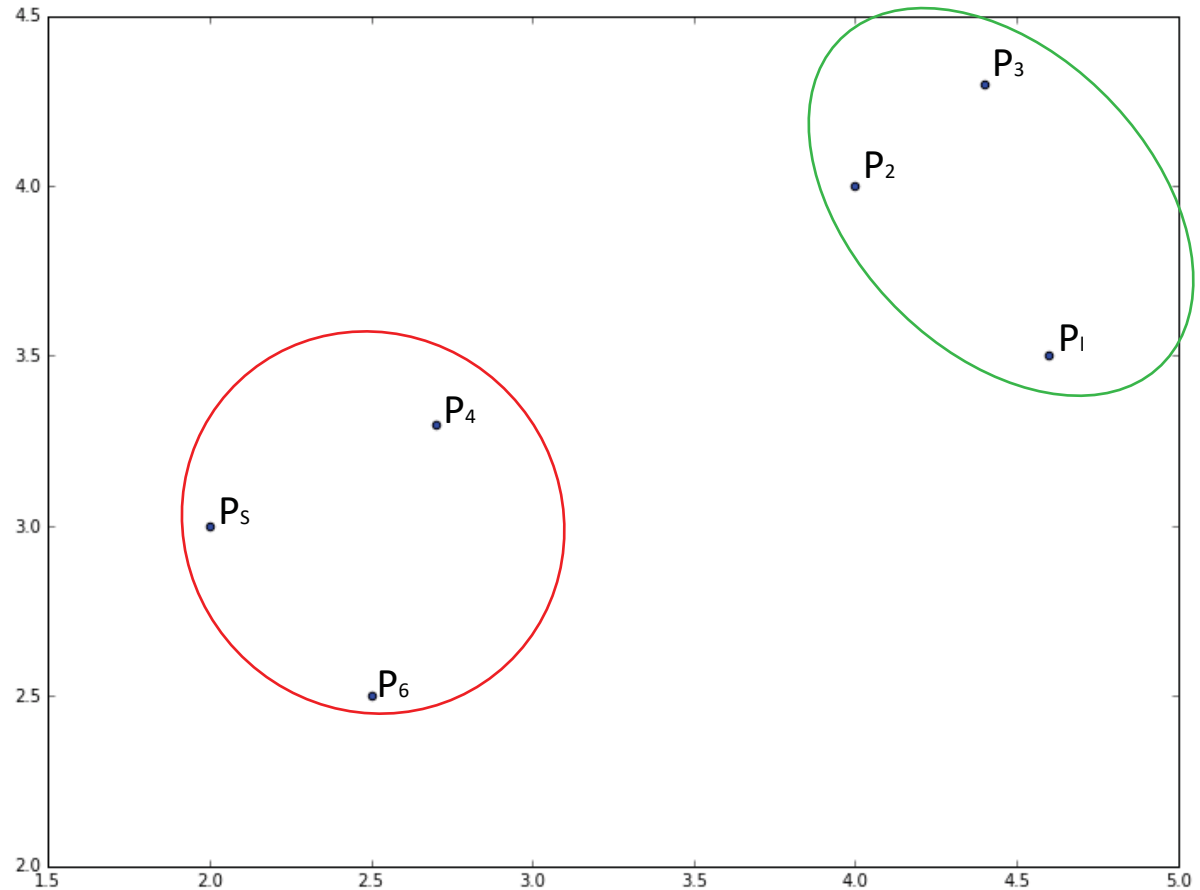


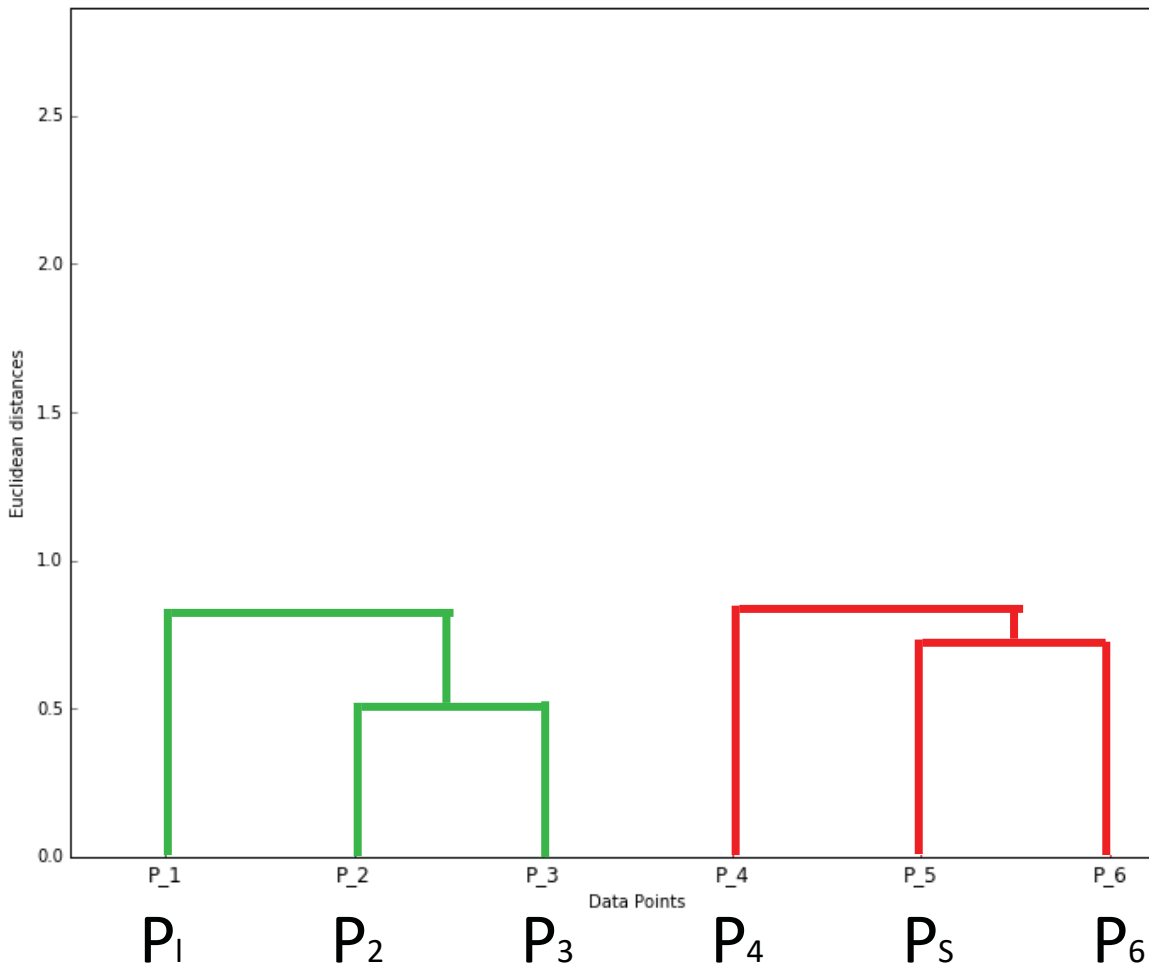
How Do Dendograms Work?



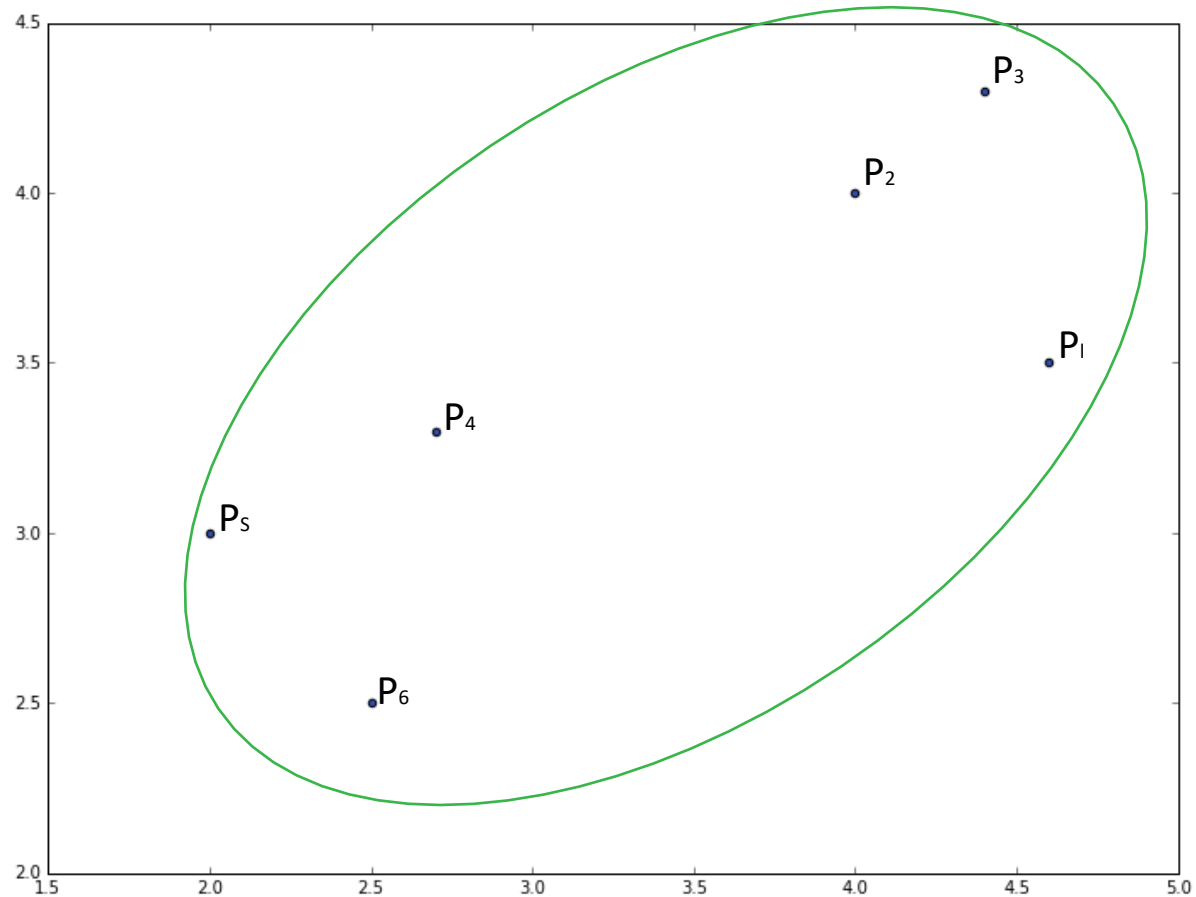


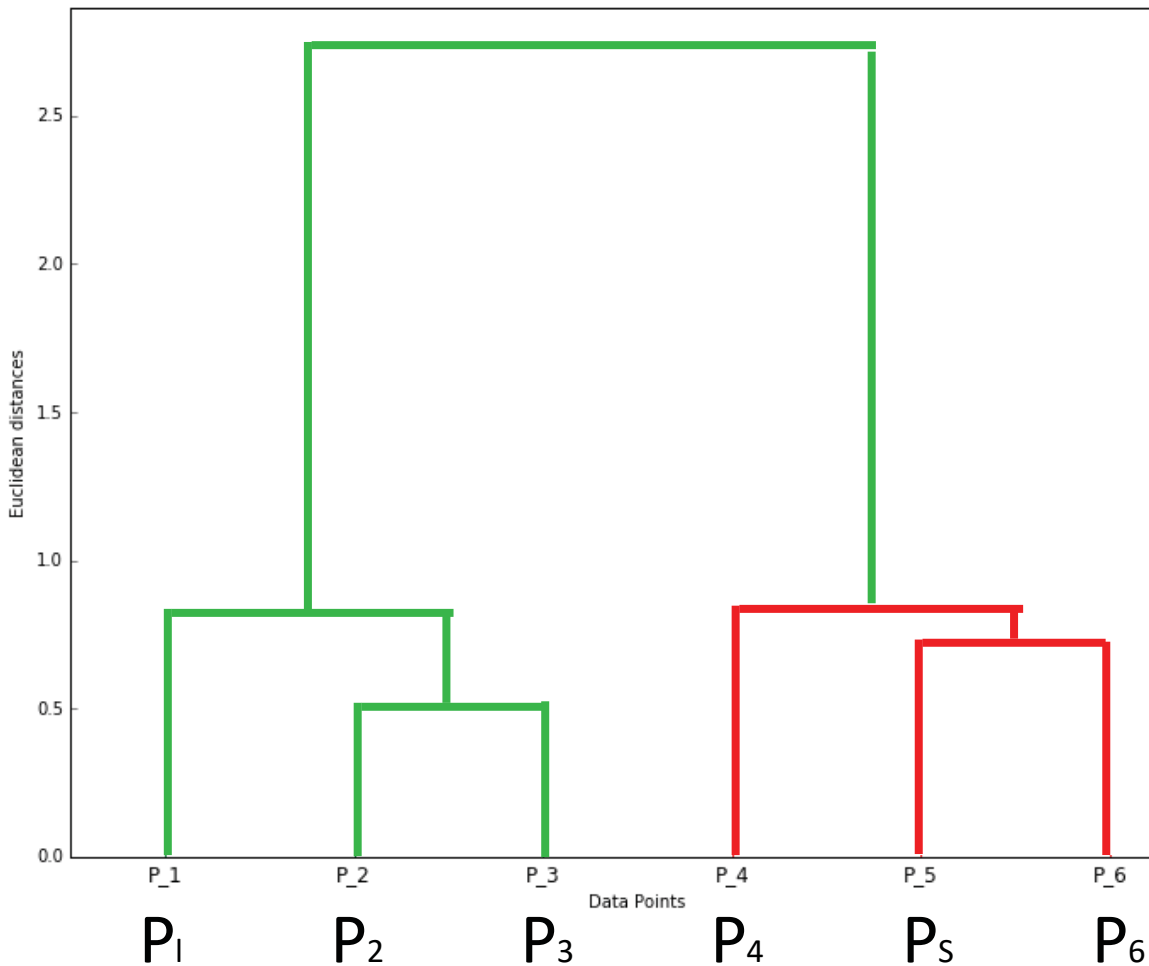
How Do Dendograms Work?



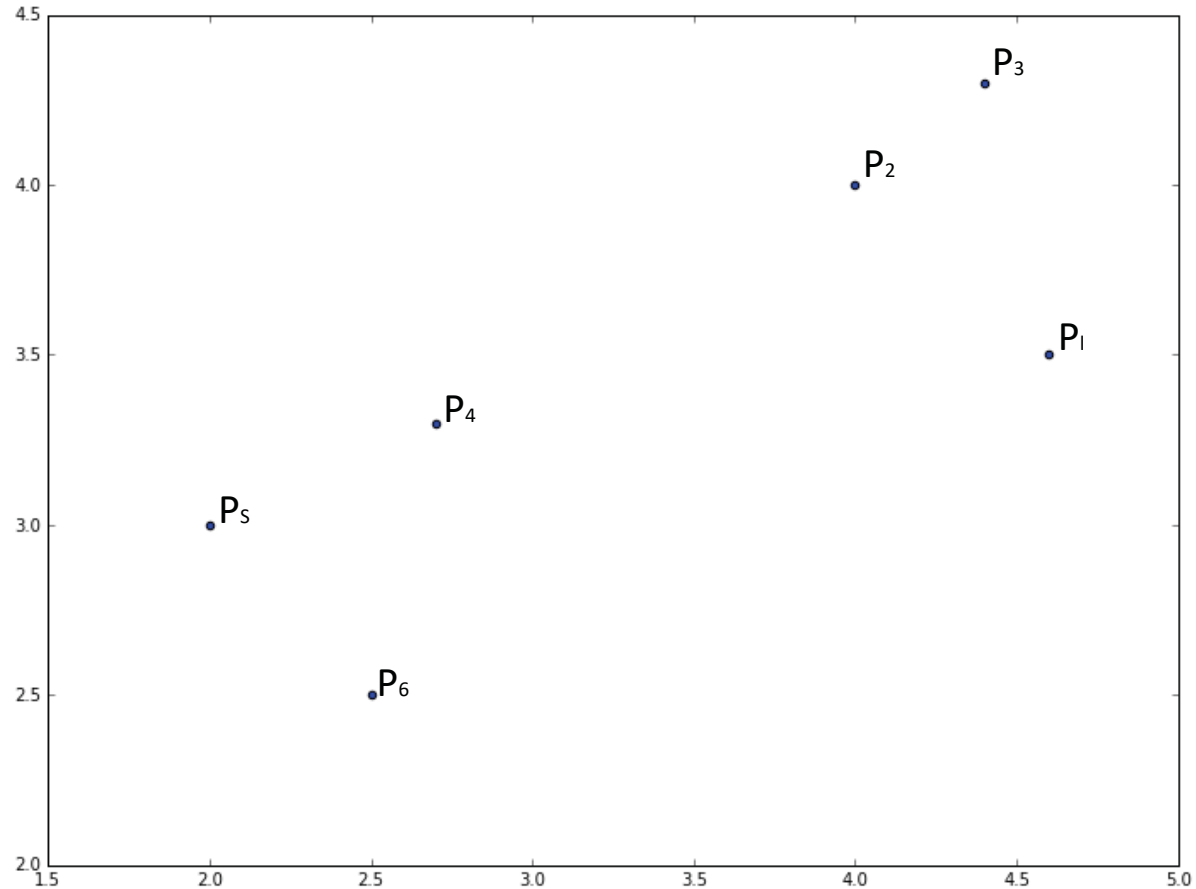


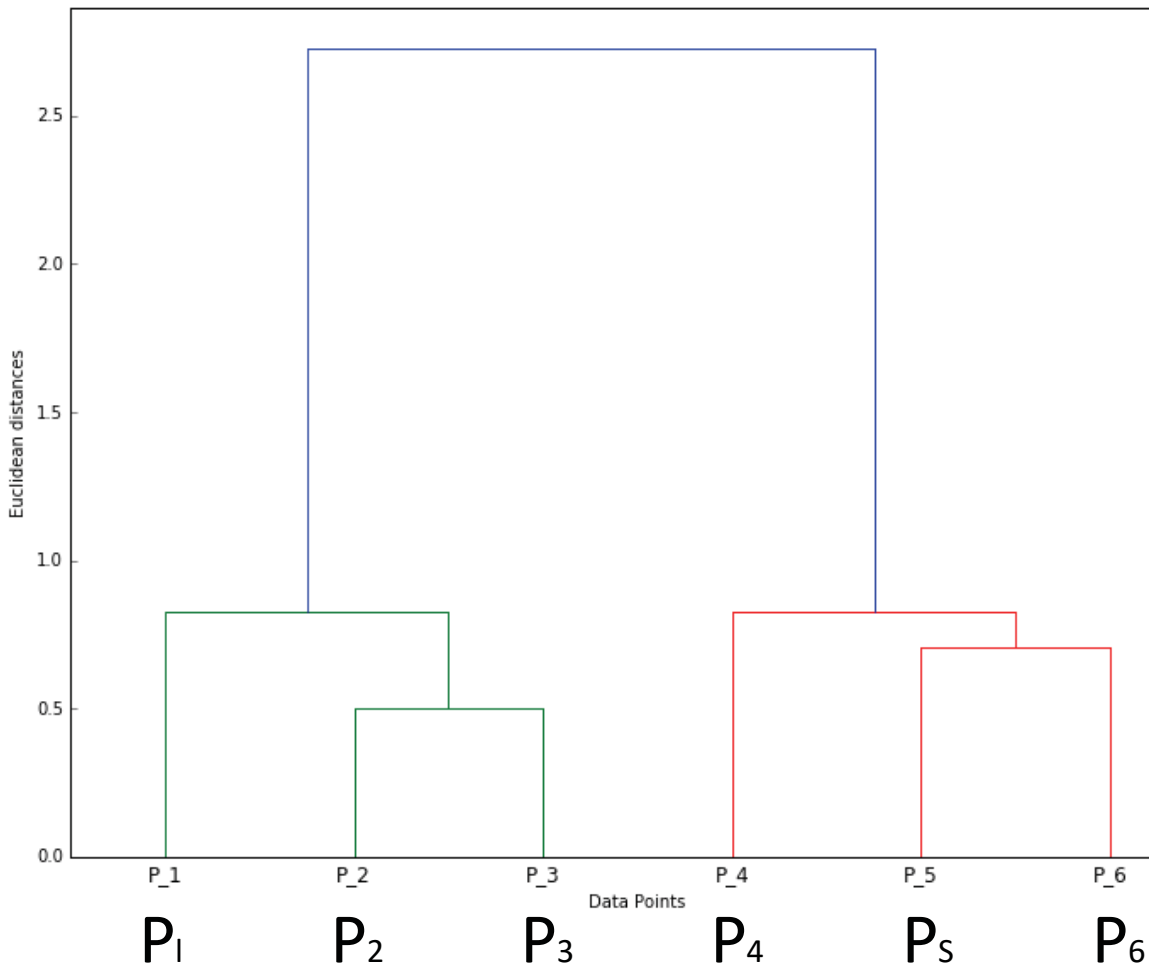
How Do Dendograms Work?





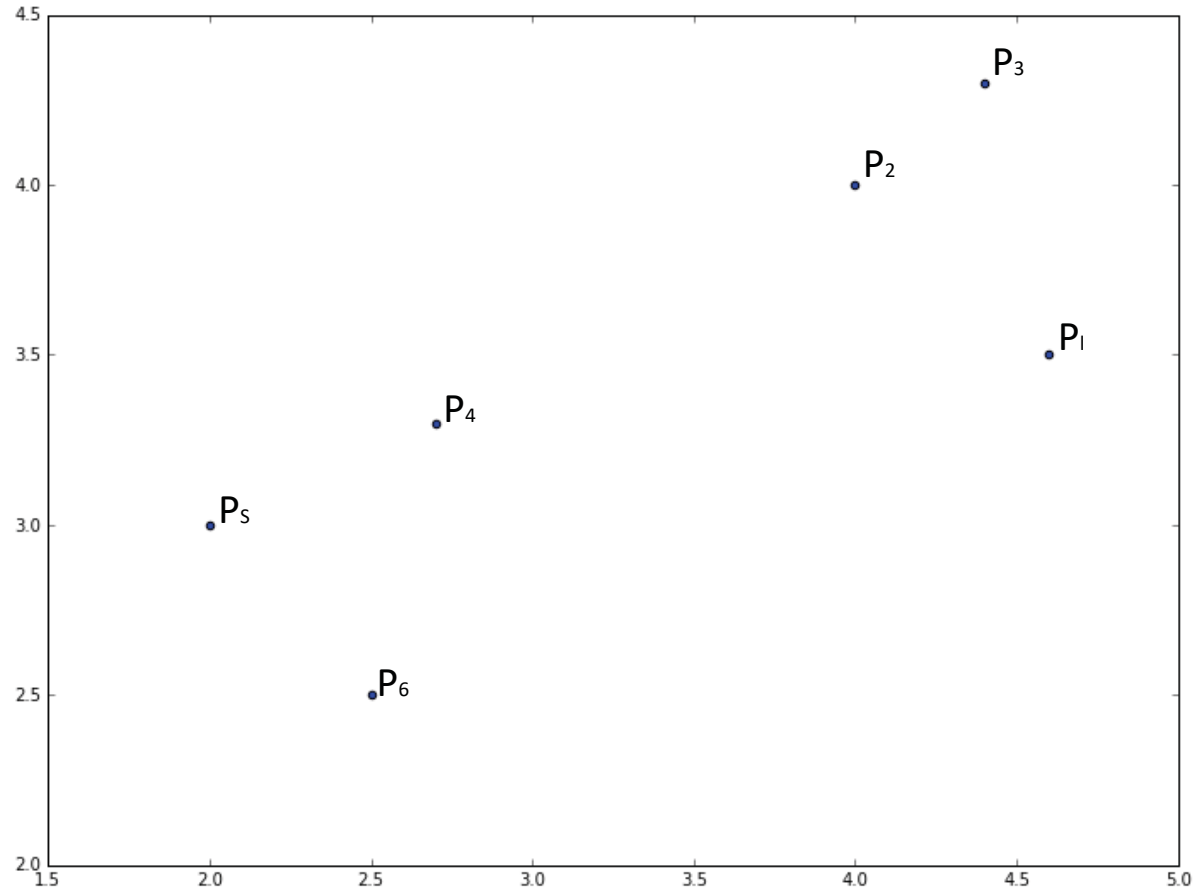
How Do Dendograms Work?

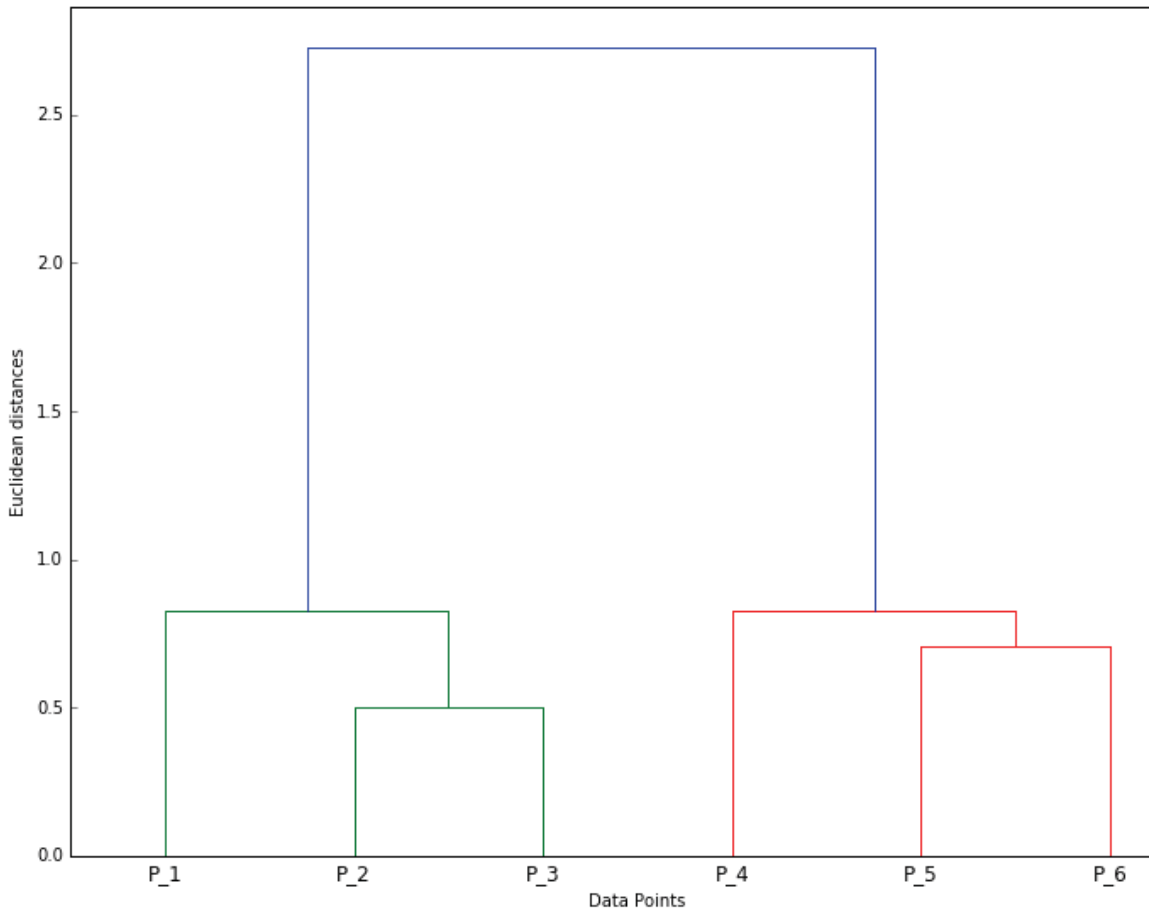




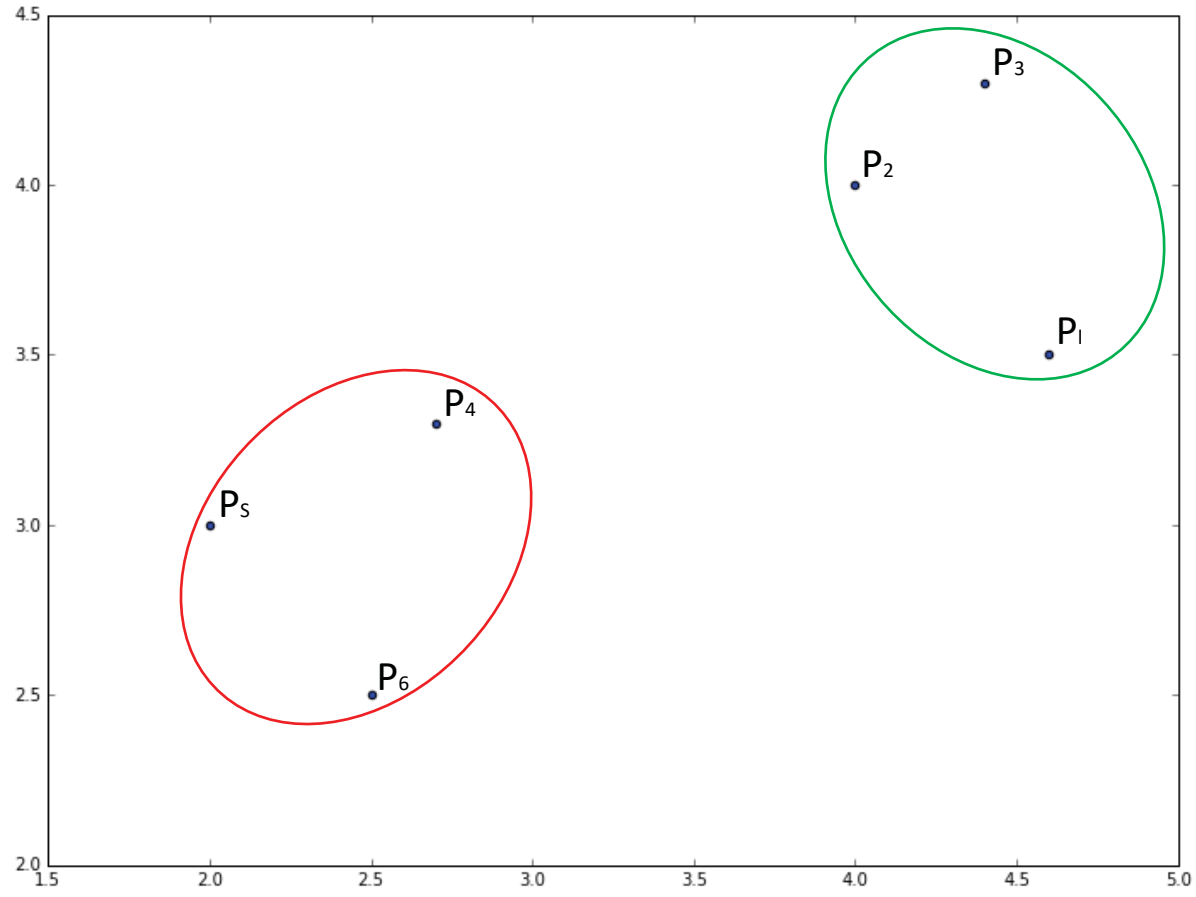
HC Intuition: Using Dendrograms

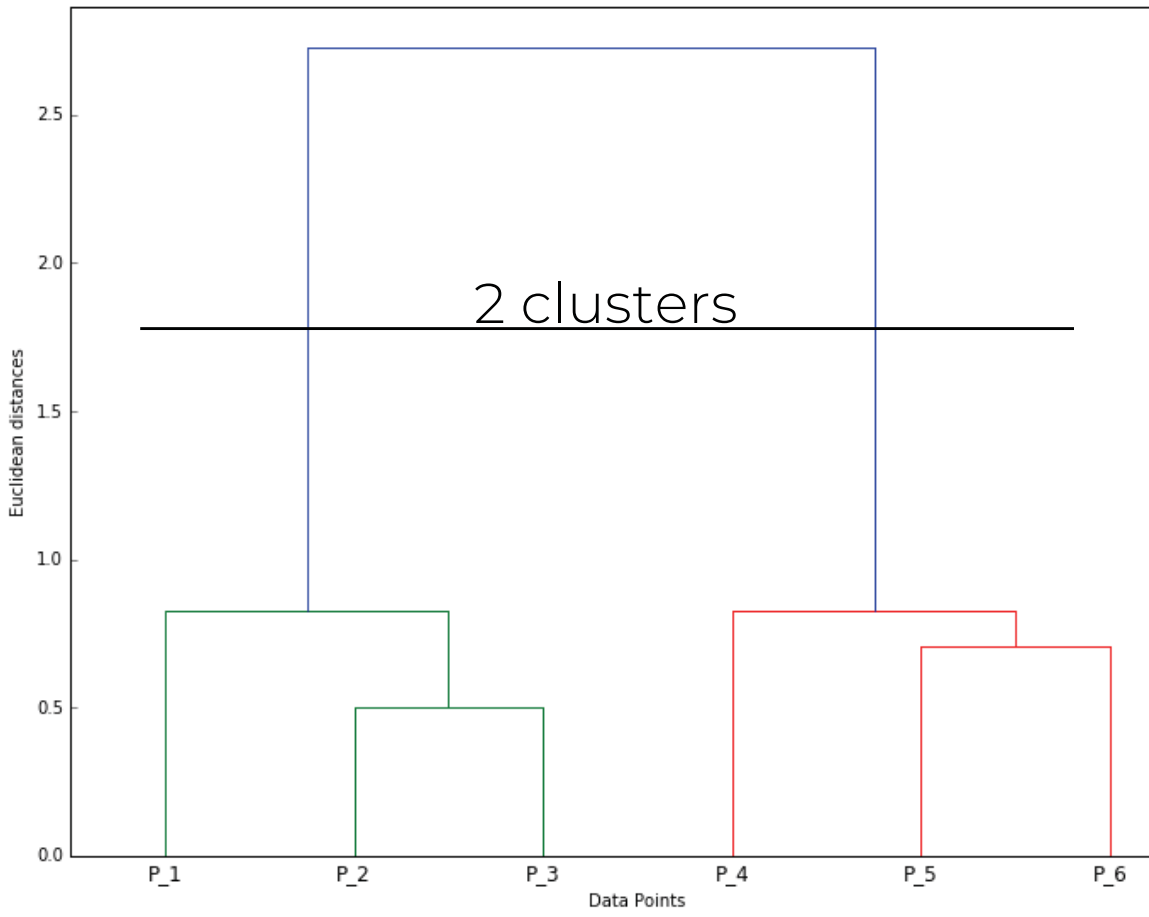
Dendrograms – Two Clusters



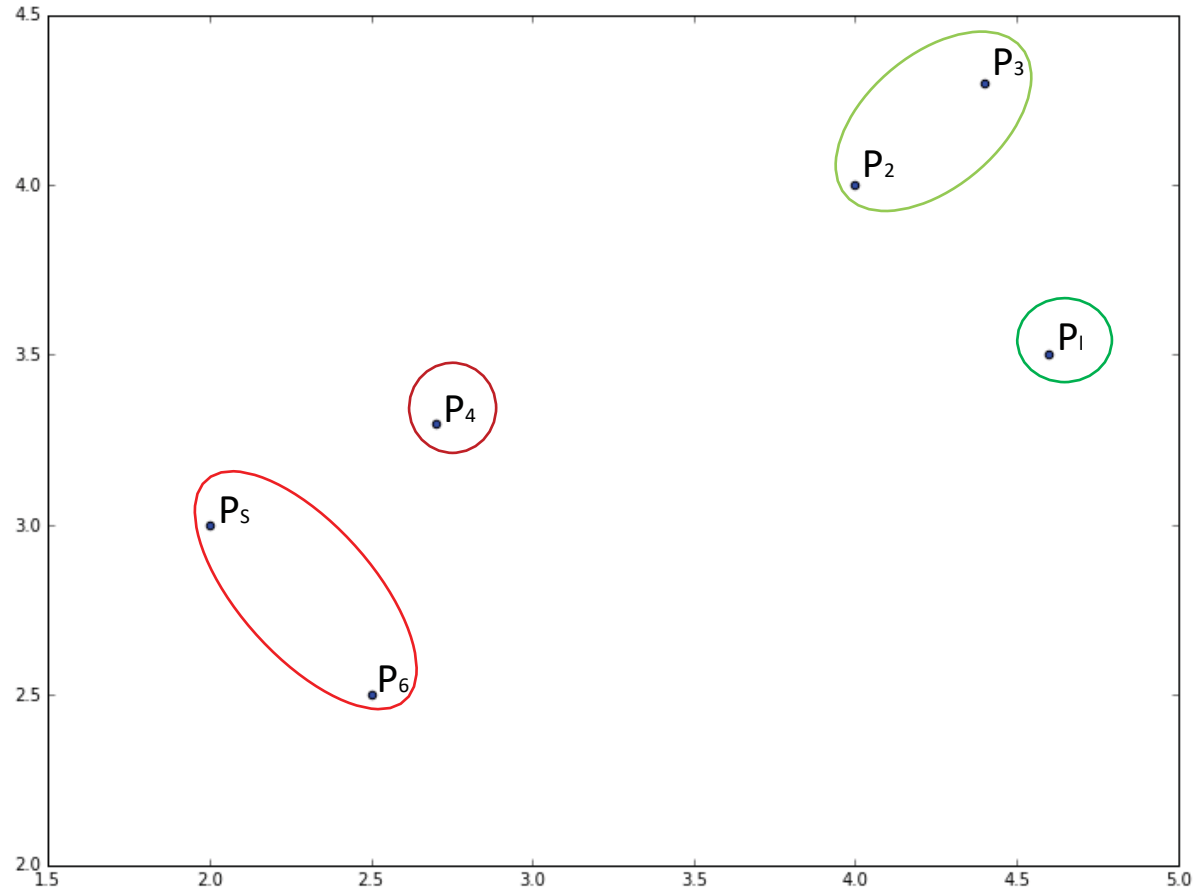


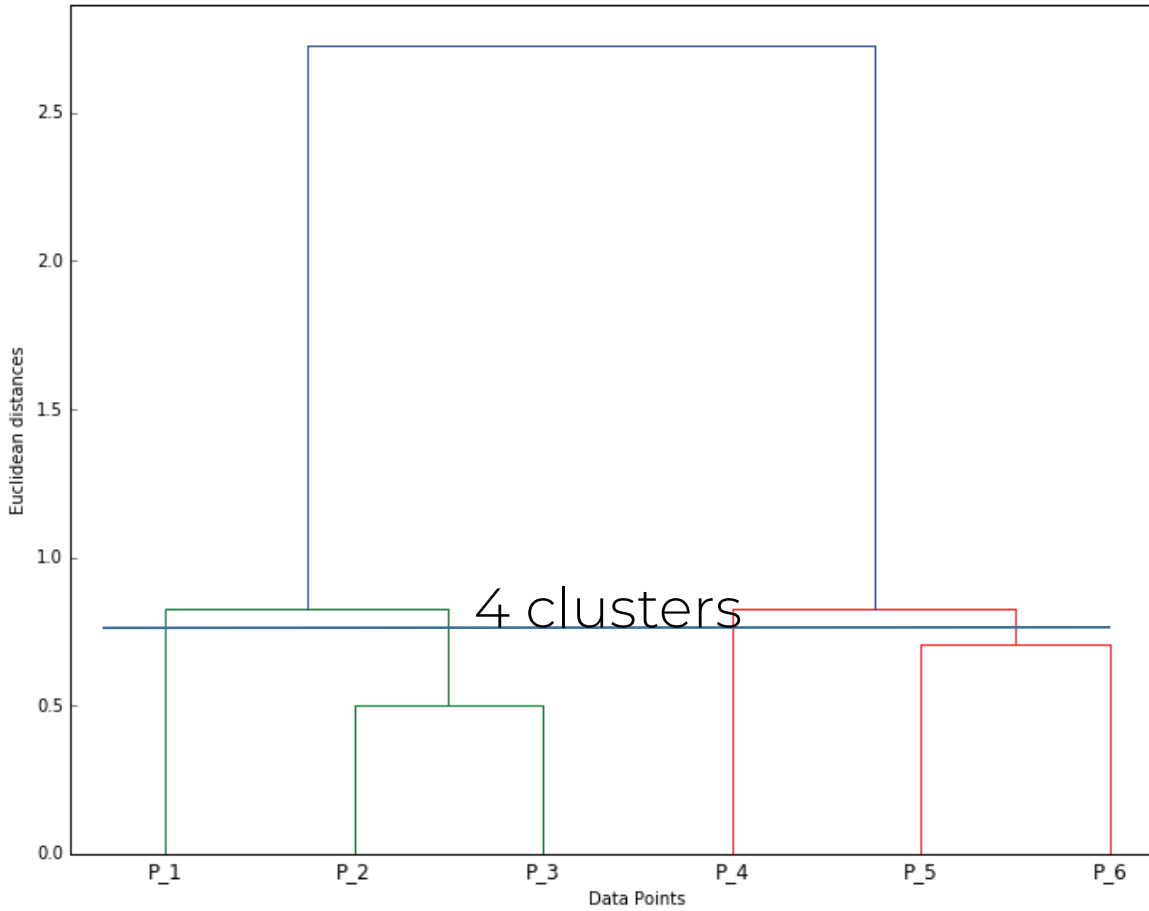
Dendrograms – Two Clusters



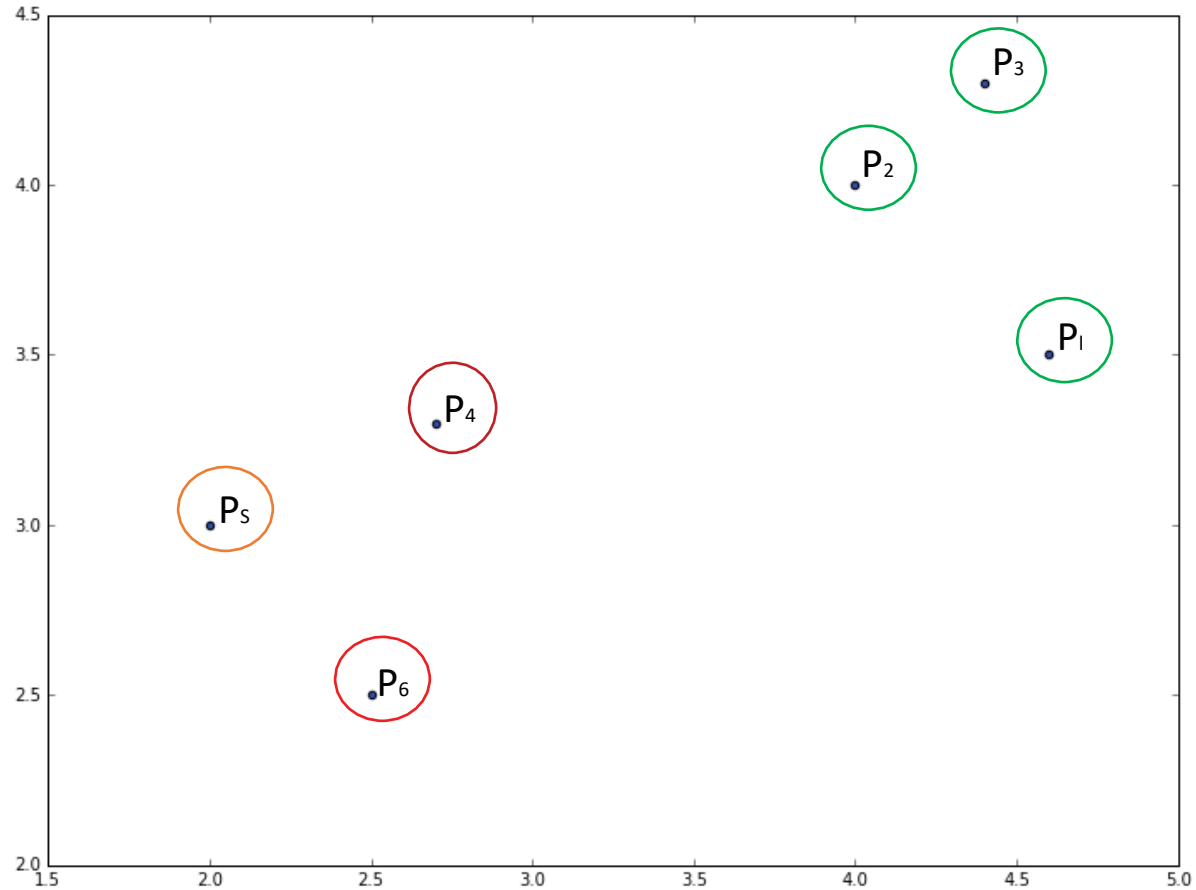


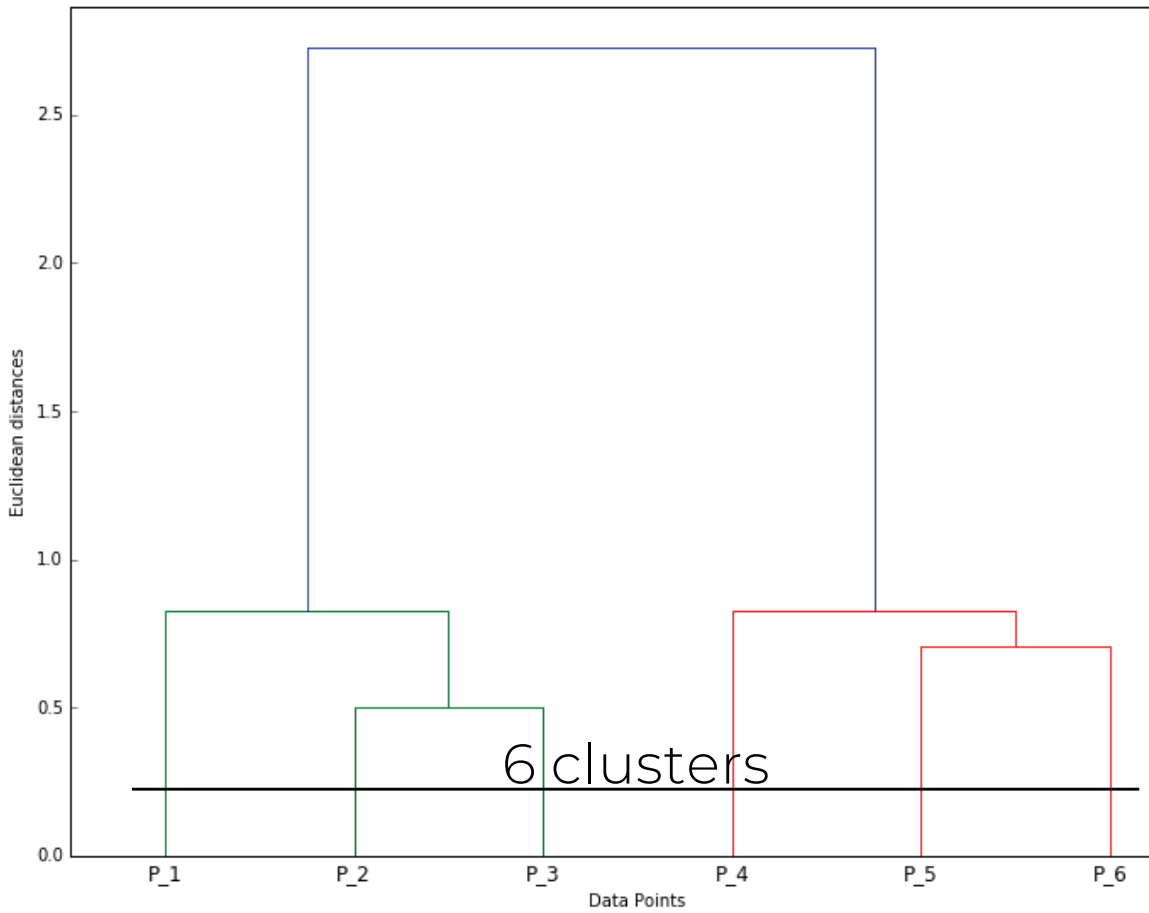
Dendrograms – Four Clusters



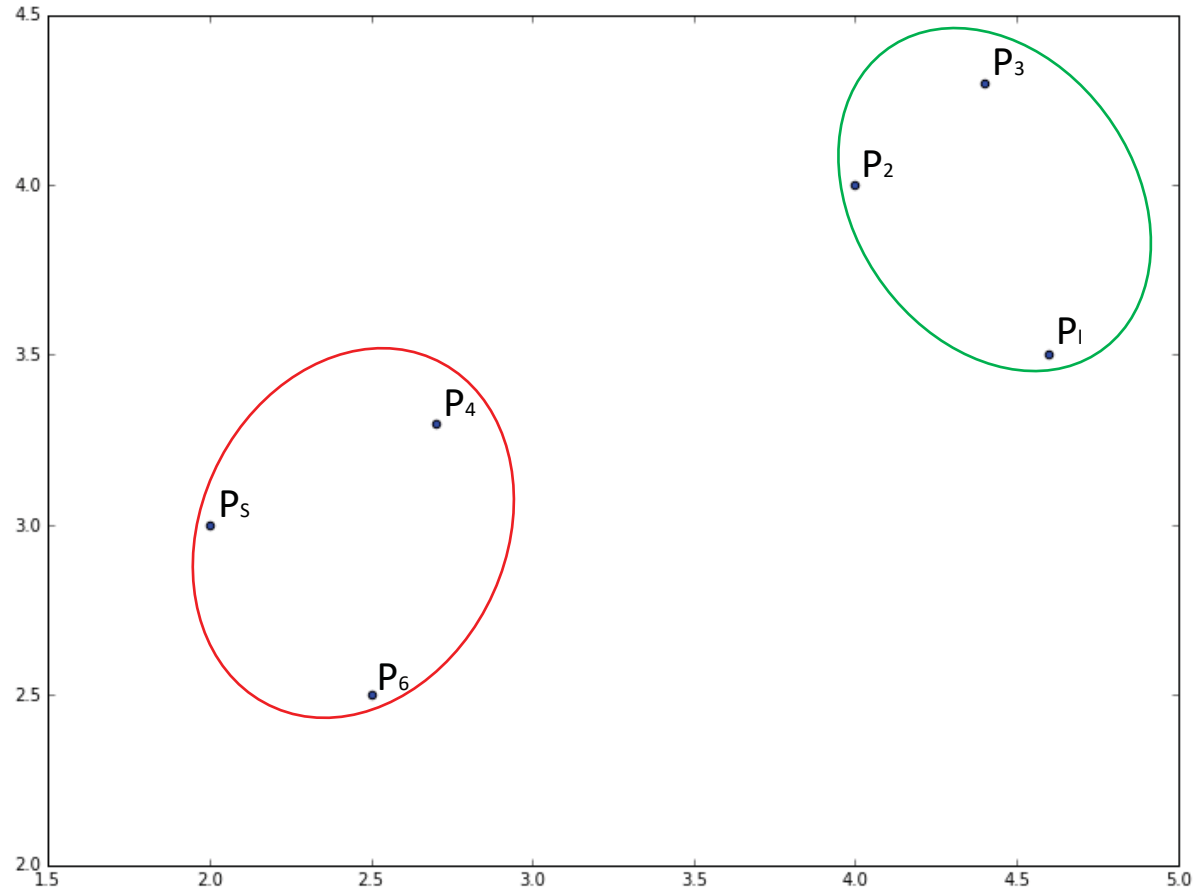


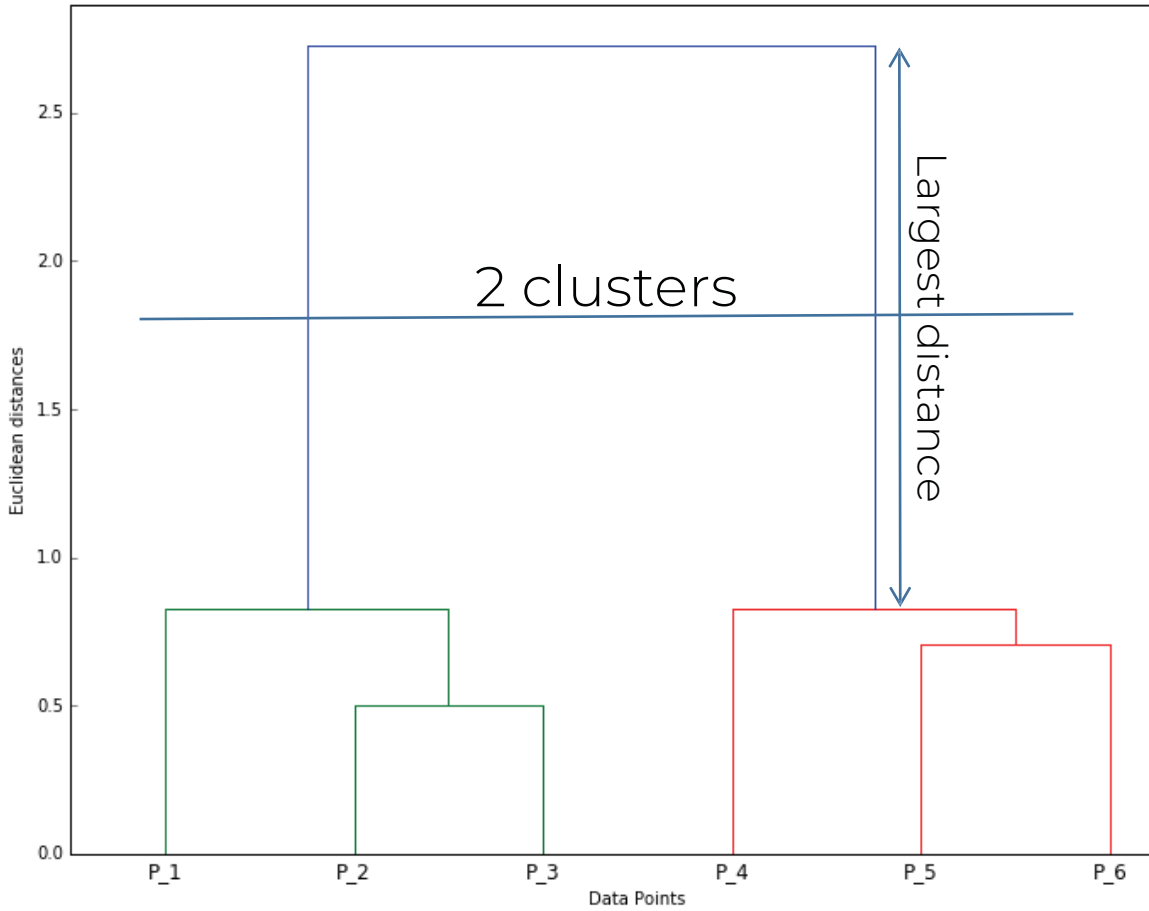
Dendrograms – Six Clusters





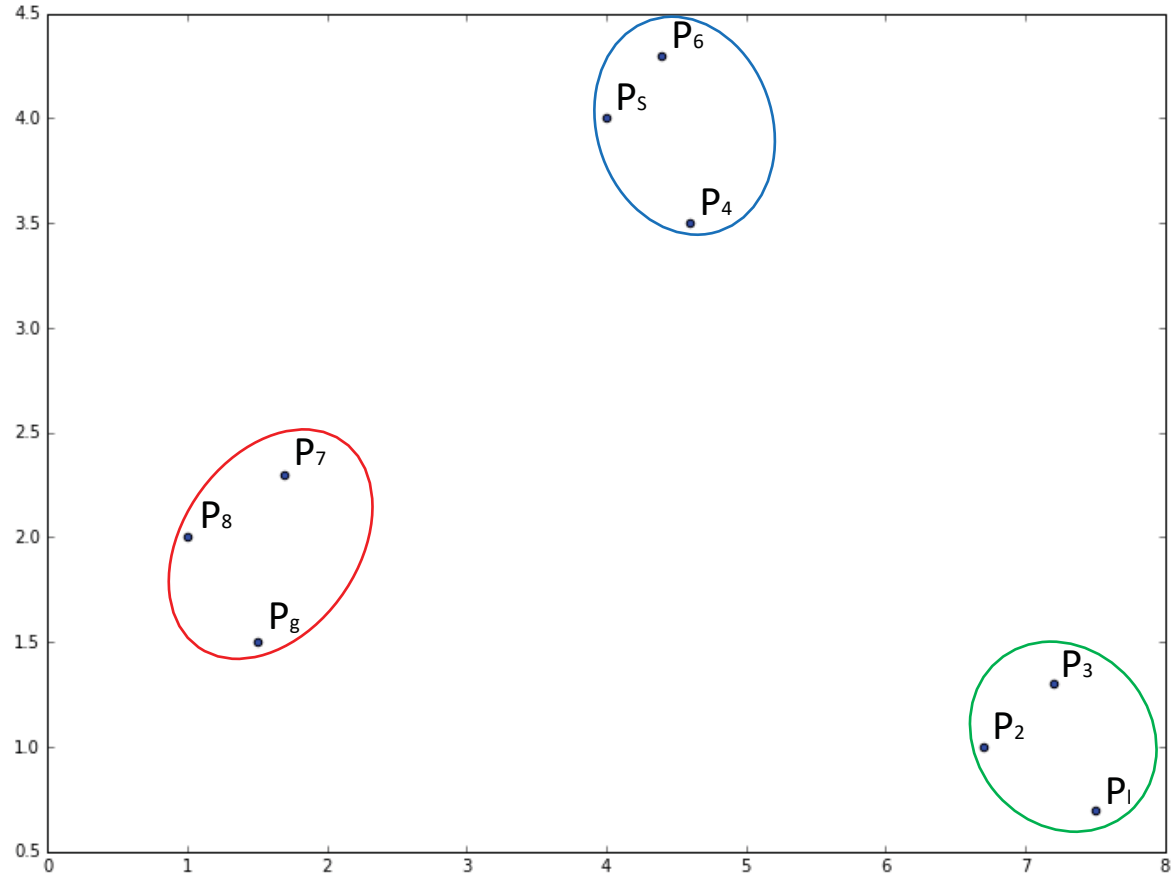
Dendrograms – Optimal # of Clusters

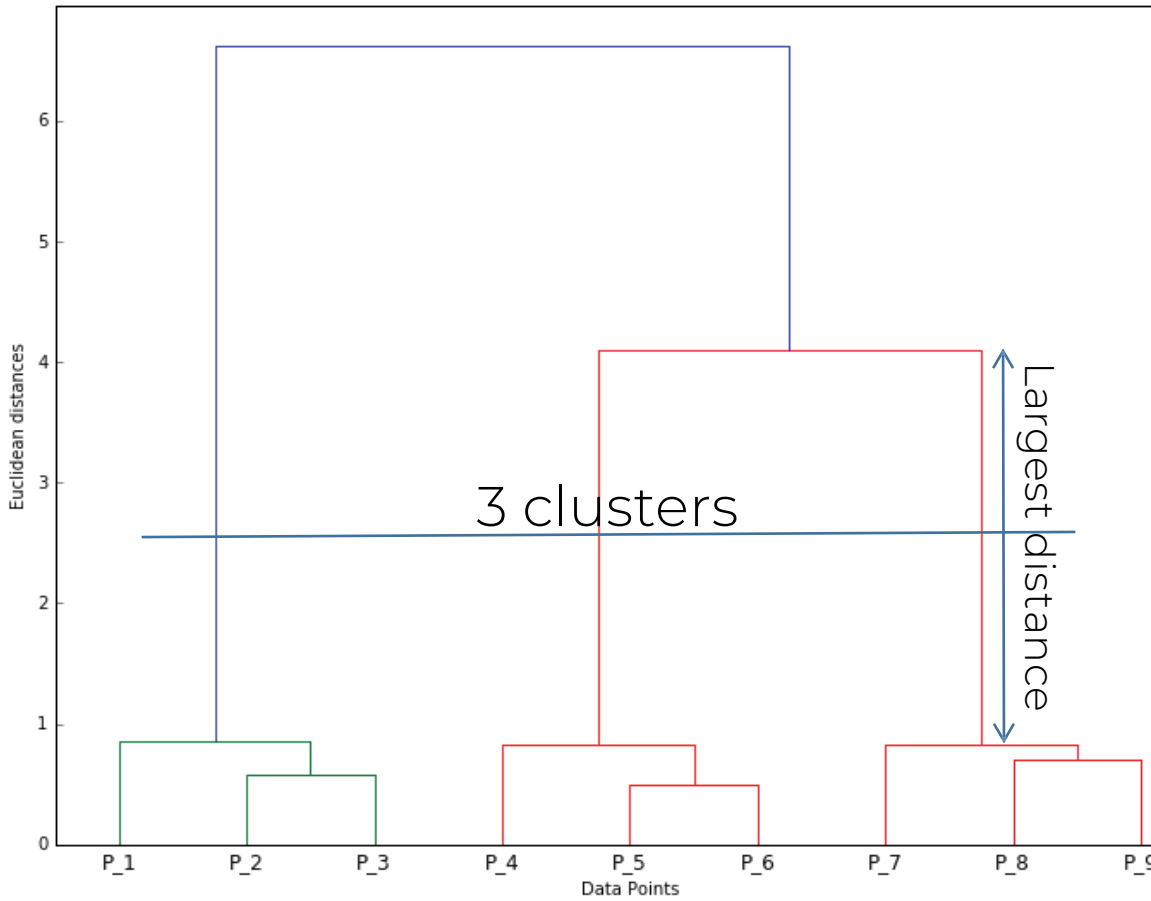




Knowledge Test

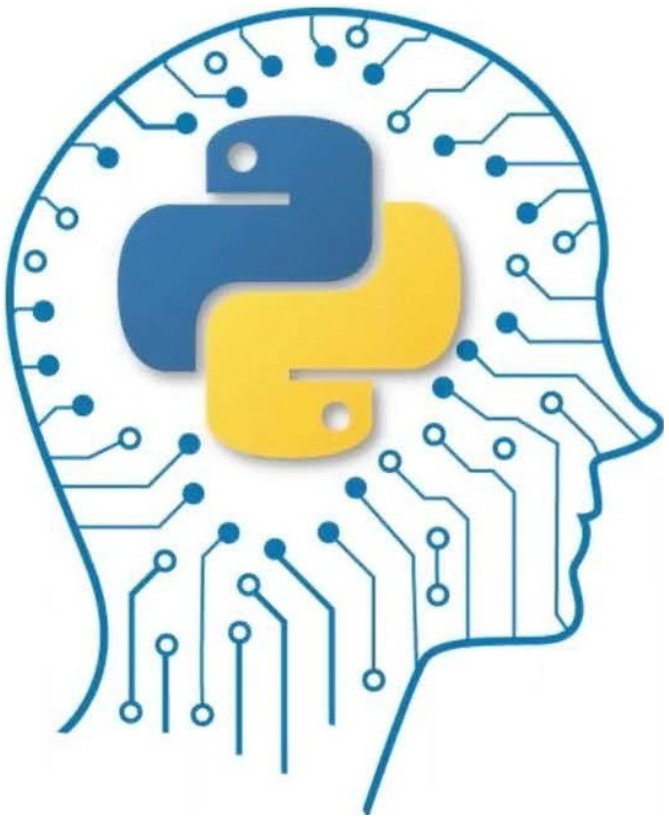
Dendrograms – Knowledge Test





K-Means vs. HC

Clustering Model	Pros	Cons
K-Means	Simple to understand, easily adaptable, works well on small or large datasets, fast, efficient and performant	Need to choose the number of clusters
Hierarchical Clustering	The optimal number of clusters can be obtained by the model itself, practical visualisation with the dendrogram	Not appropriate for large datasets

© SuperDataScience

See you next lecture!