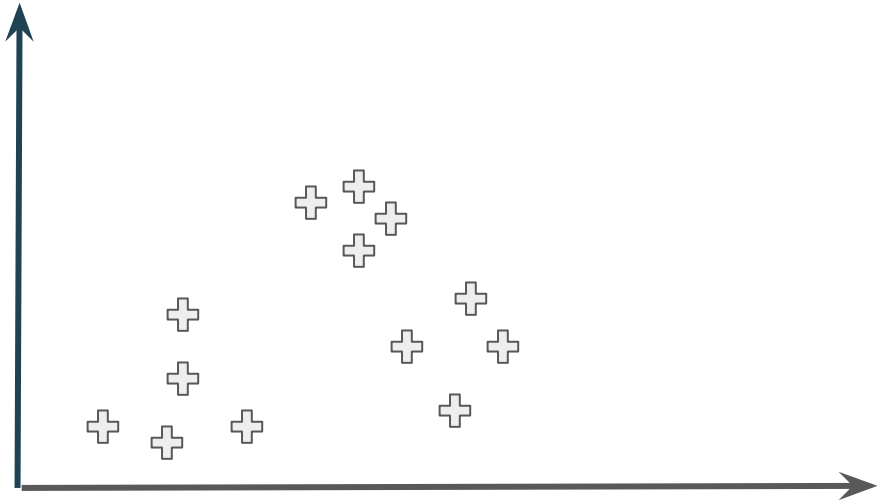
DBScan

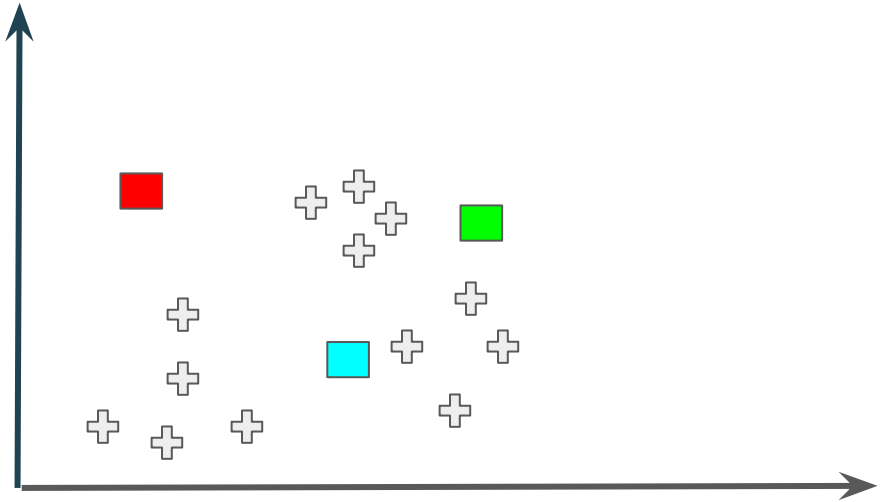
Density Clustering

When do K-Means fail?

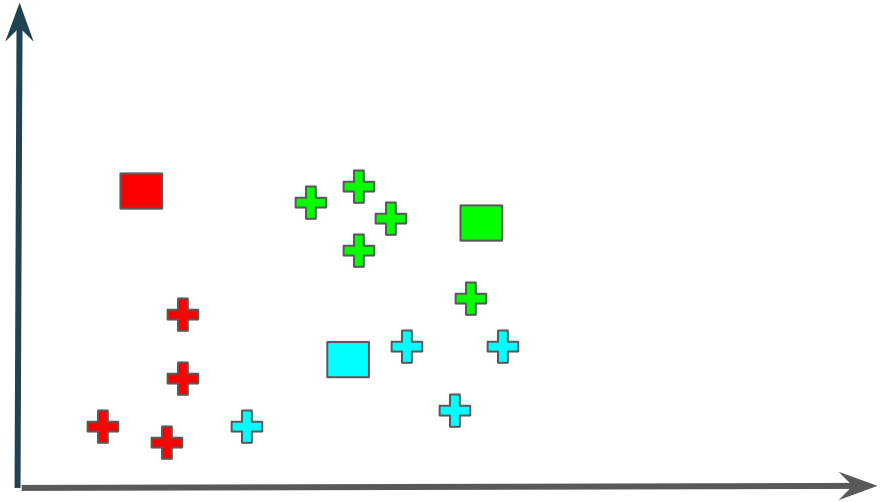
Example 1 - Data



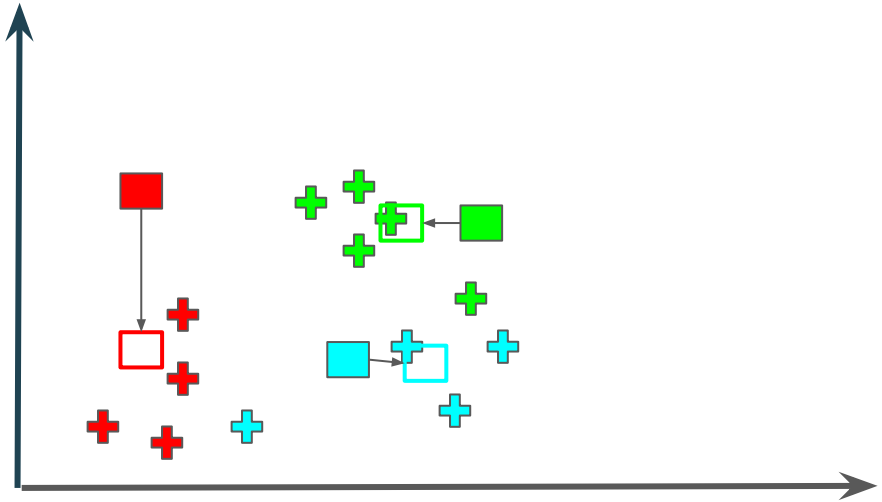
Step 2 - Calculate centroïds



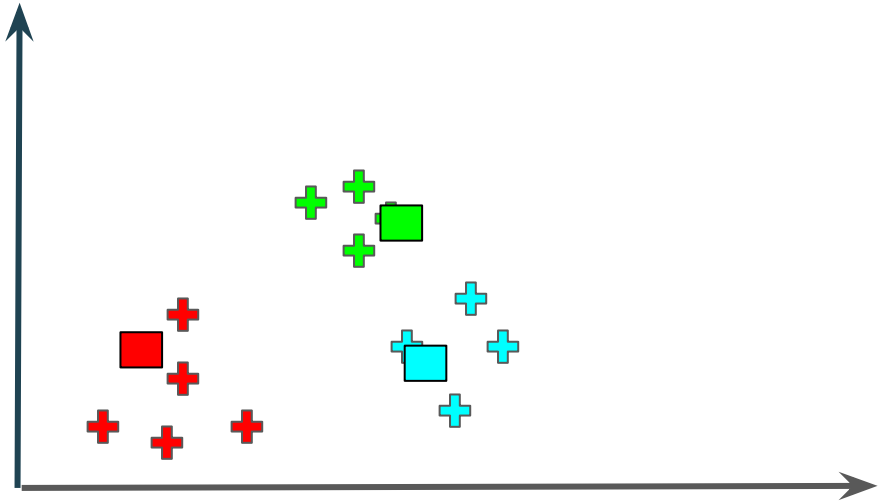
Step 3 - Assign data points to the closest centroid



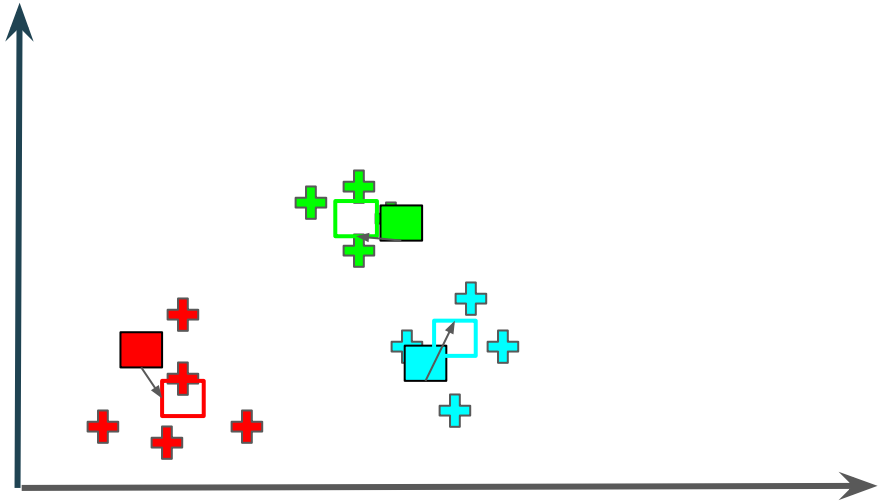
Step 2 - Calculate centroids



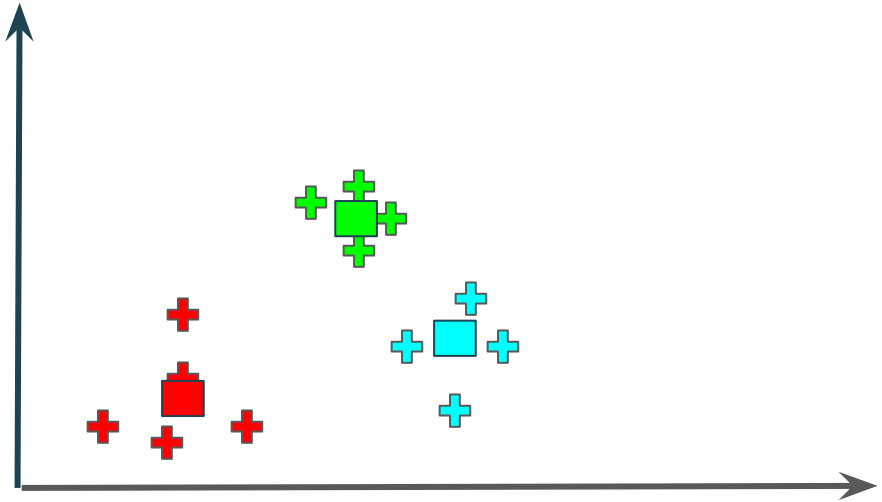
Step 3 - Assign data points to the closest centroid



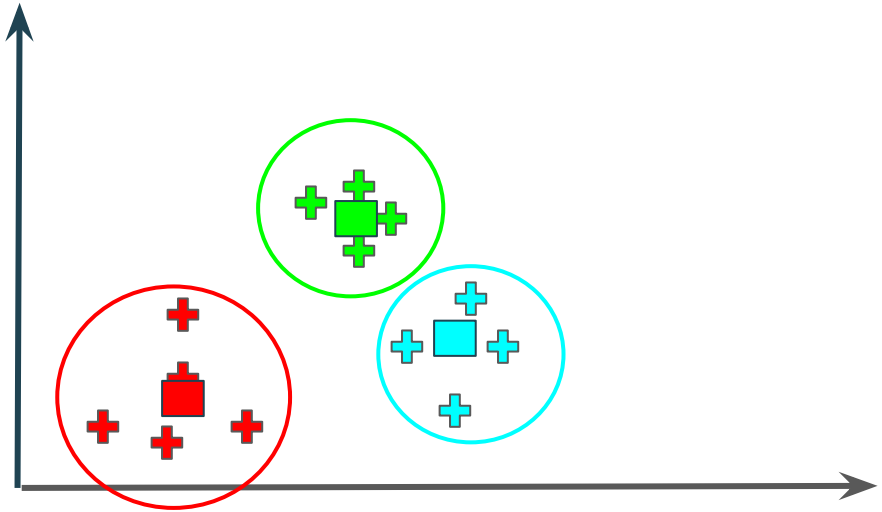
Step 2 - Calculate centroids



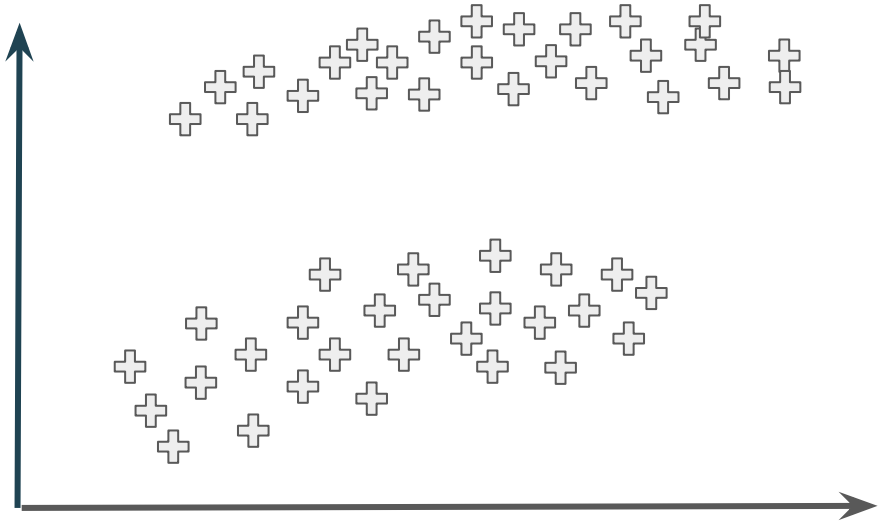
Step 3 - Assign data points to the closest centroid



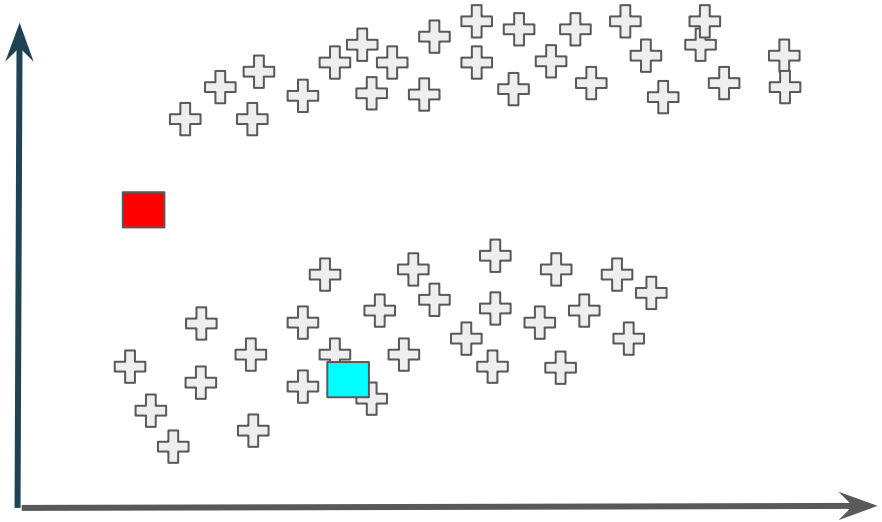
Step 4 - Get our K clusters



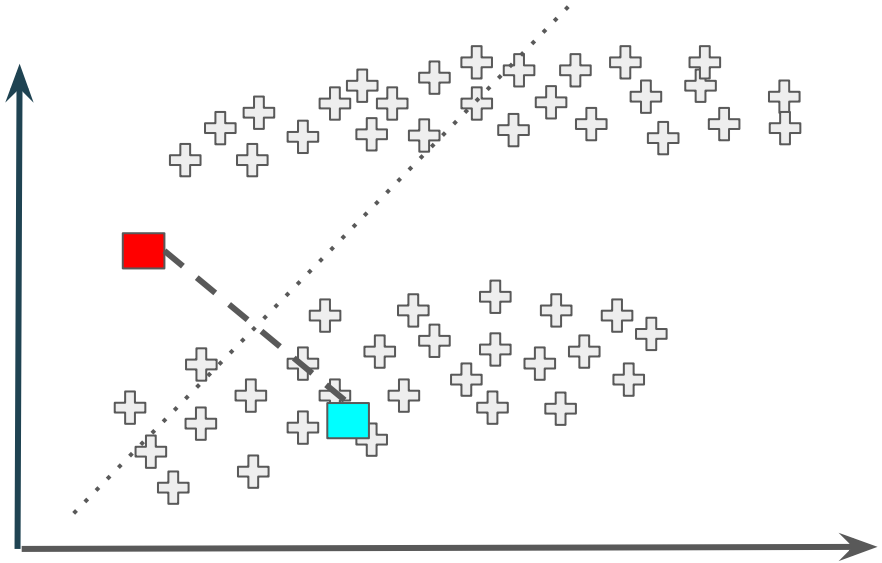
Example 2 - Data



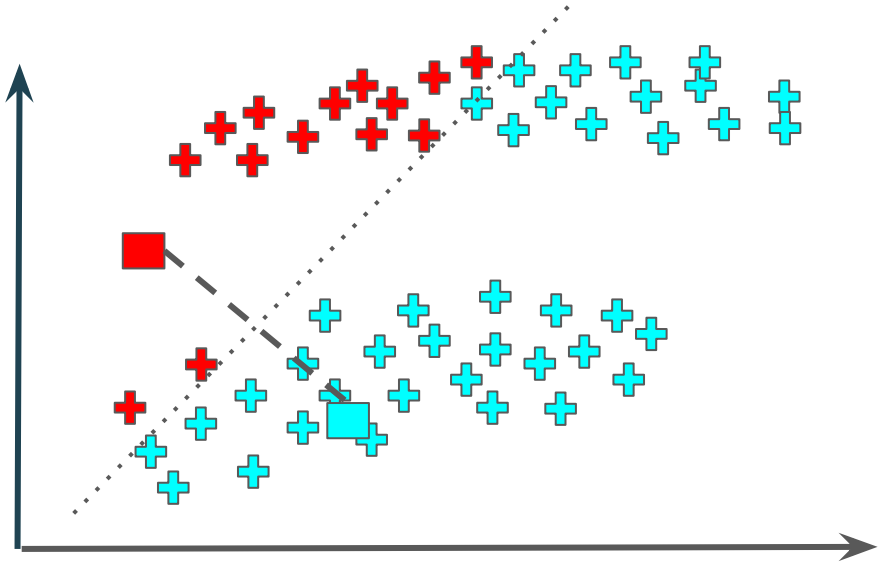
Example 2 - K=2



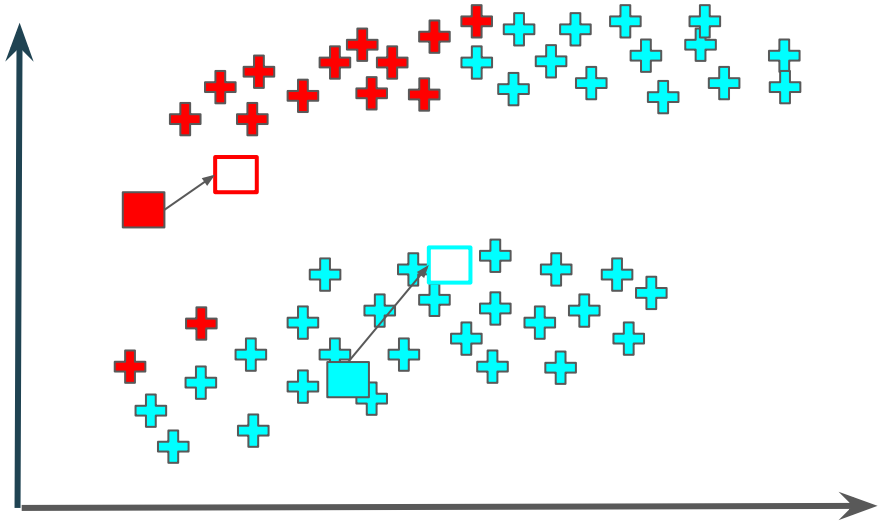
Example 2 - Assign Data Points



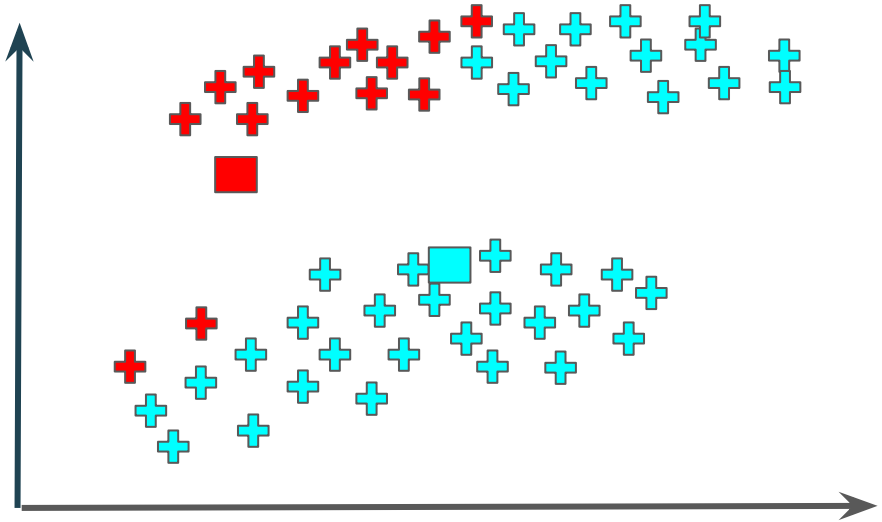
Example 2 - Assign Data Points



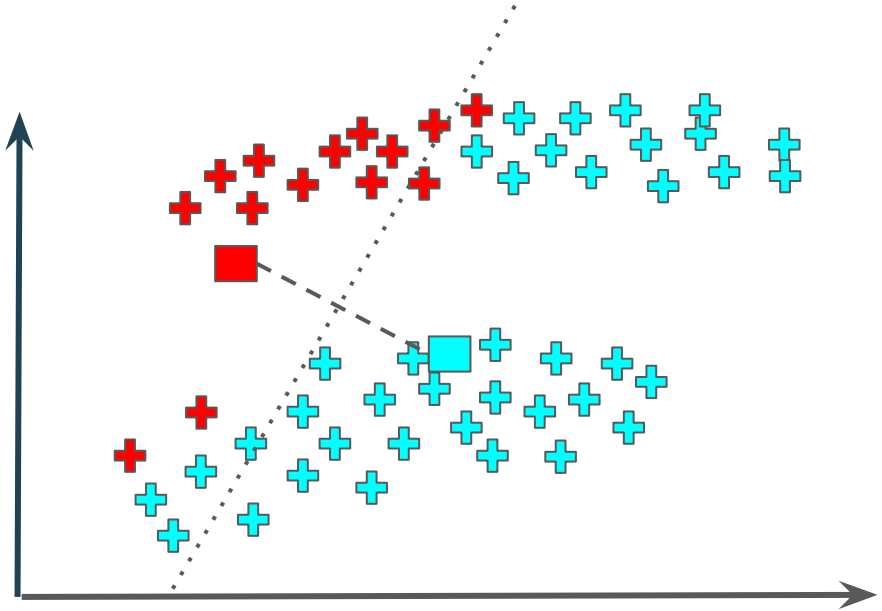
Example 2 - Calculate Centroids



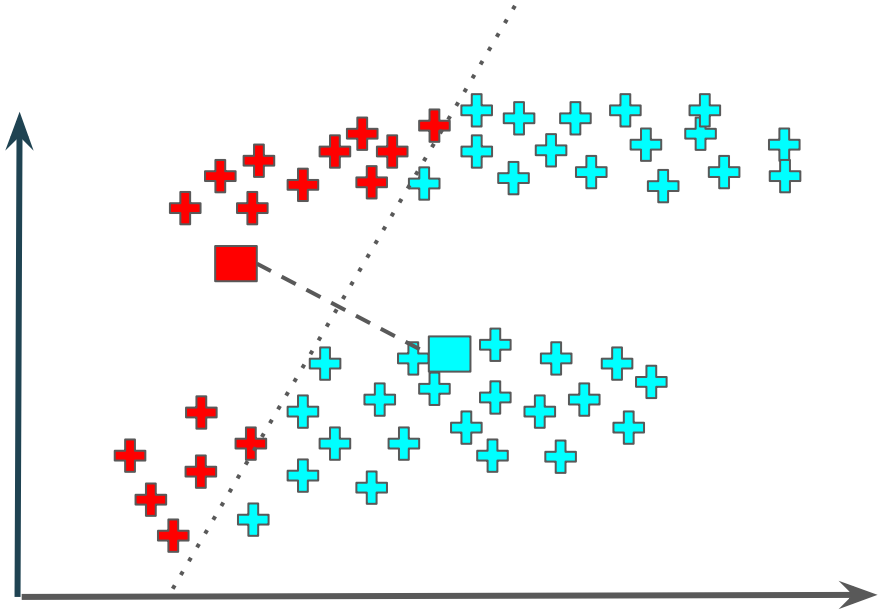
Example 2 - Calculate Centroids



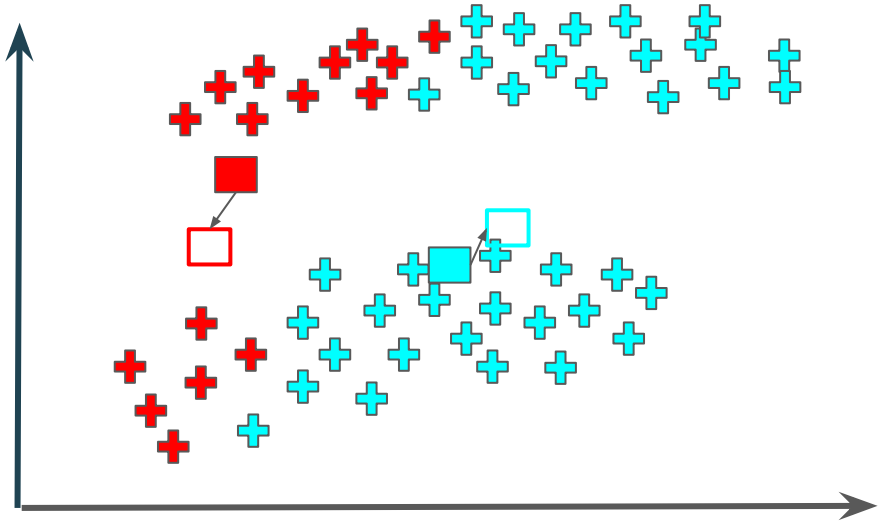
Example 2 - Assign Data Points



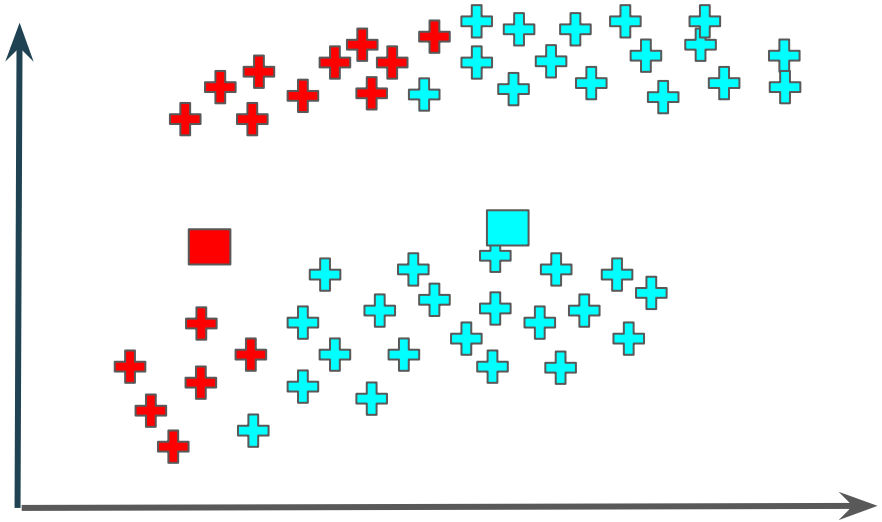
Example 2 - Assign Data Points



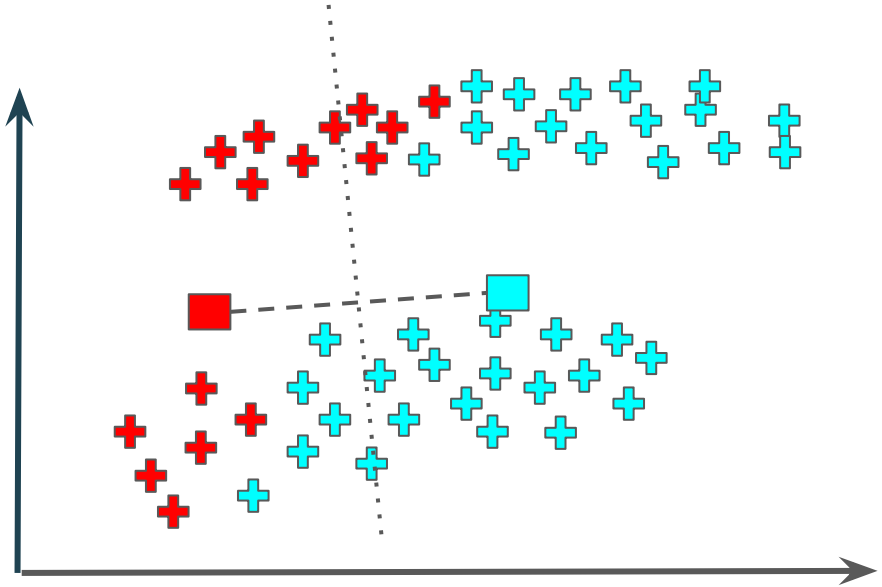
Example 2 - Calculate Centroids



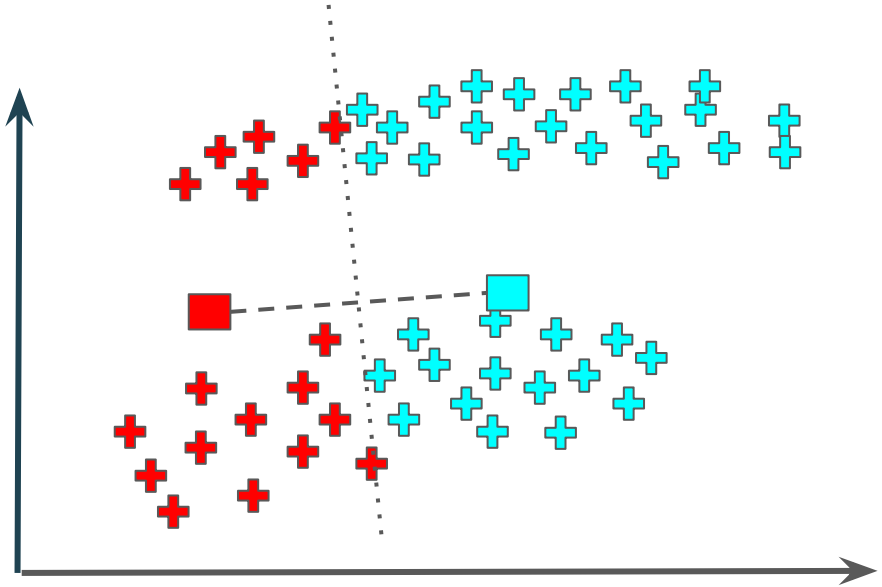
Example 2 - Calculate Centroids



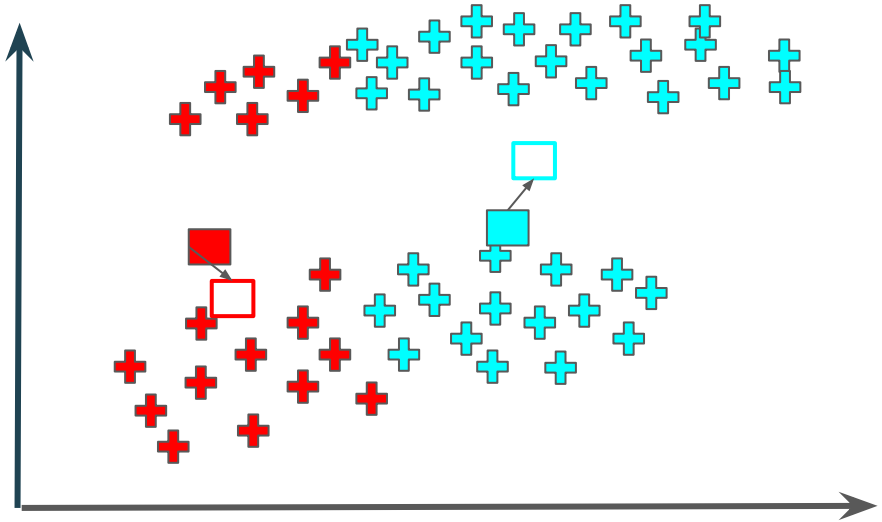
Example 2 - Assign Data Points



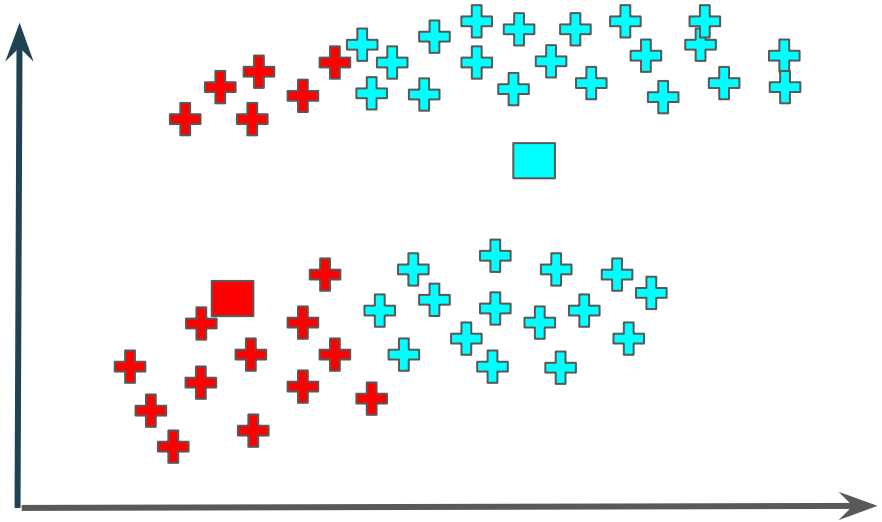
Example 2 - Assign Data Points



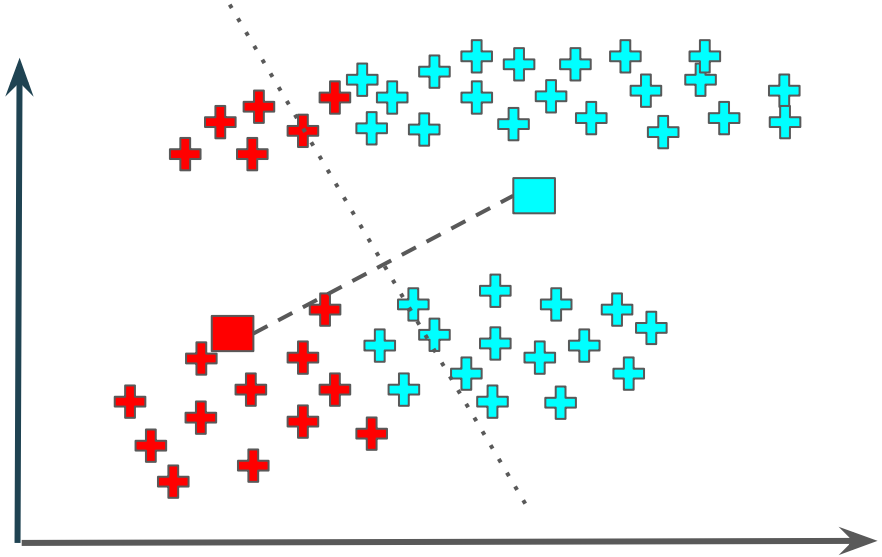
Example 2 - Calculate Centroids



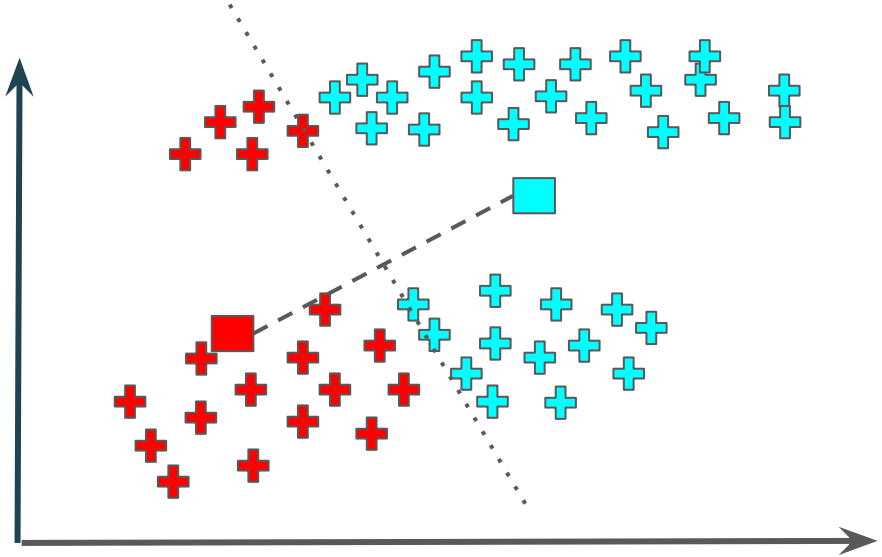
Example 2 - Calculate Centroids



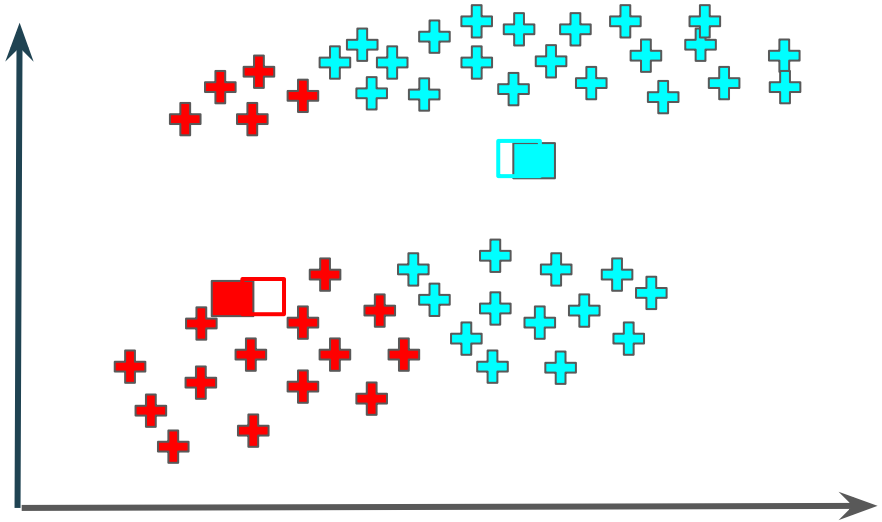
Example 2 - Assign data points



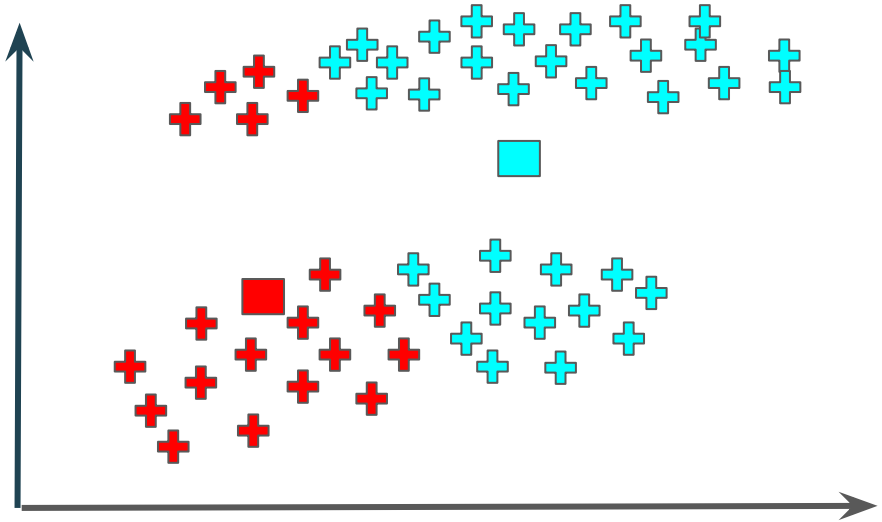
Example 2 - Assign data points



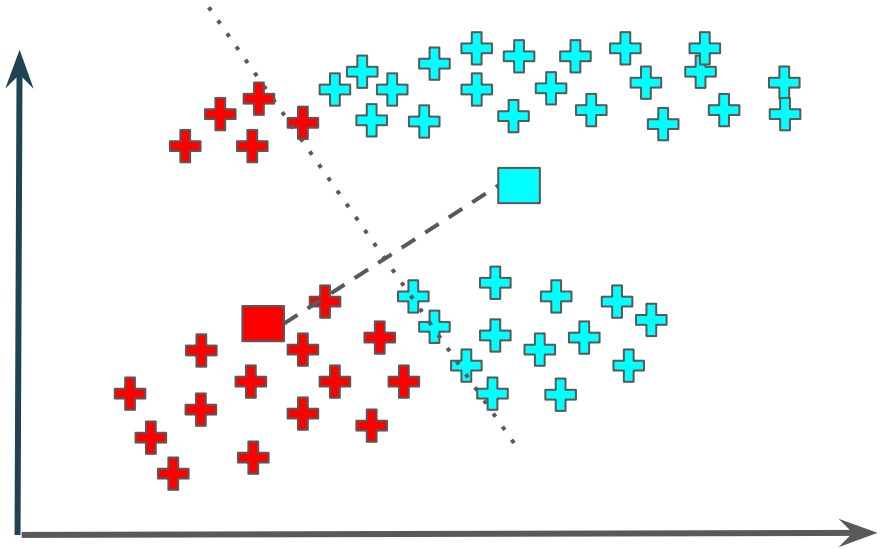
Example 2 - Calculate Centroids



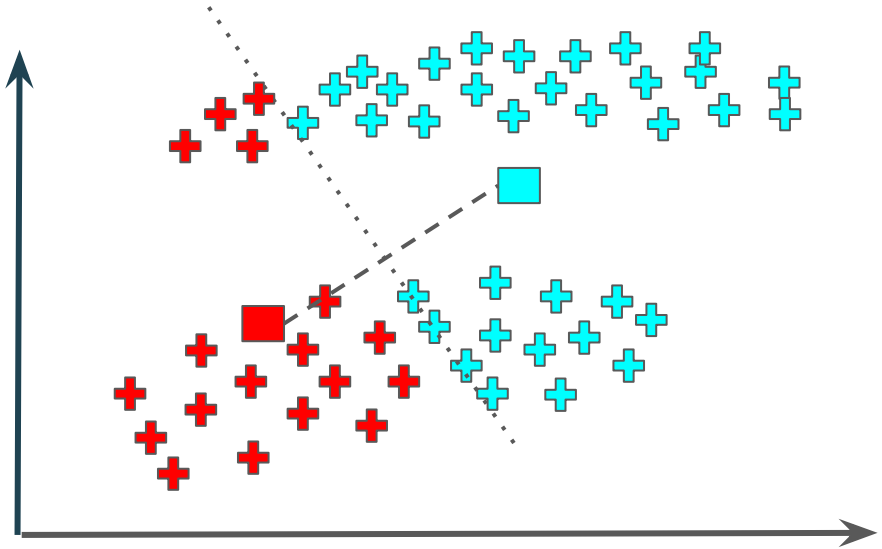
Example 2 - Calculate Centroids



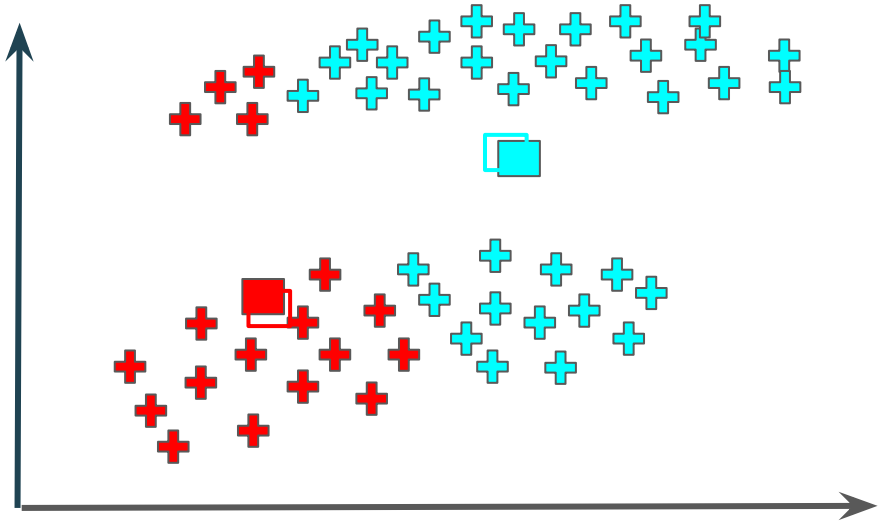
Example 2 - Assign Data Points



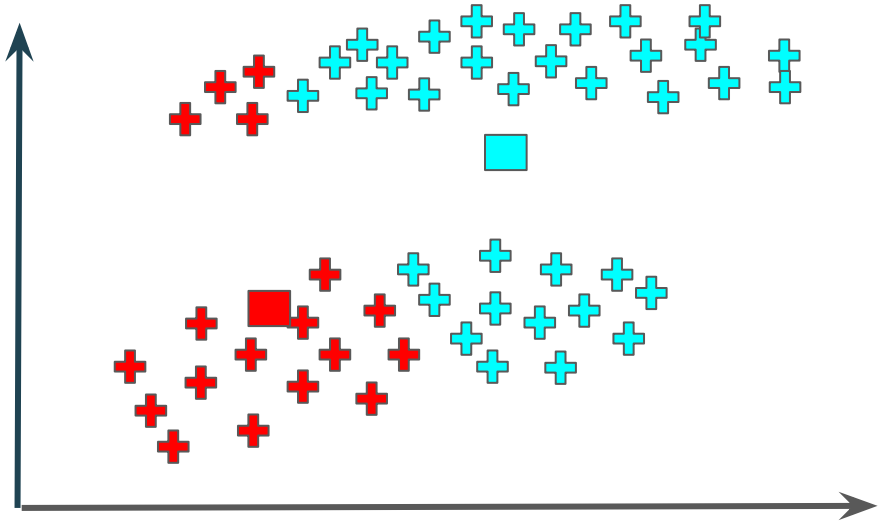
Example 2 - Assign Data Points



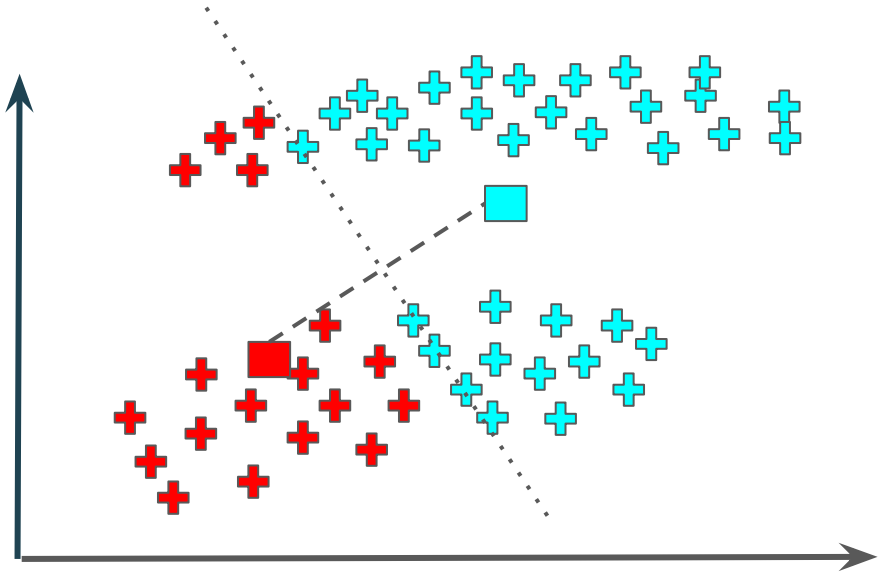
Example 2 - Calculate Centroids



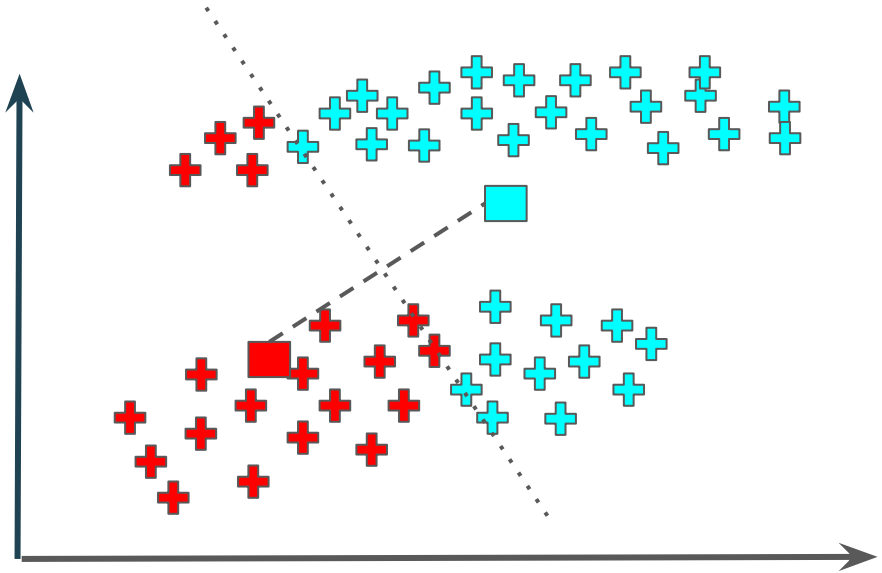
Example 2 - Calculate Centroids



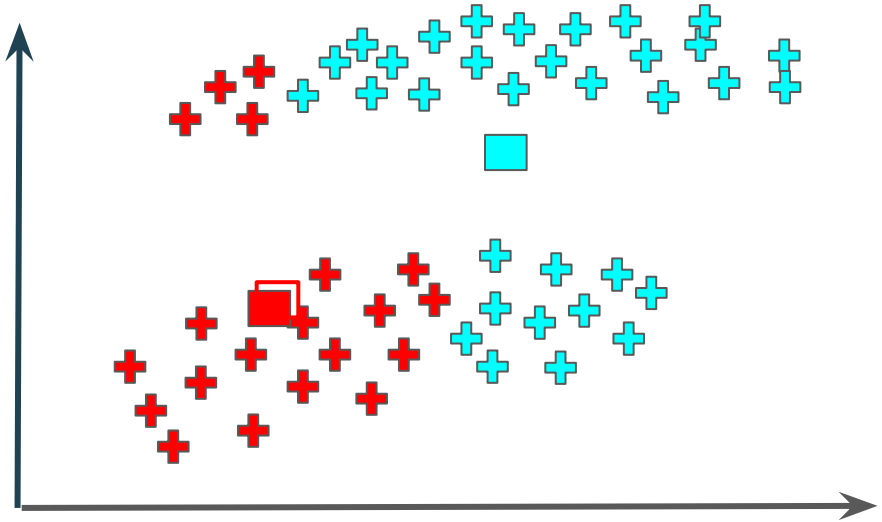
Example 2 - Assign data points



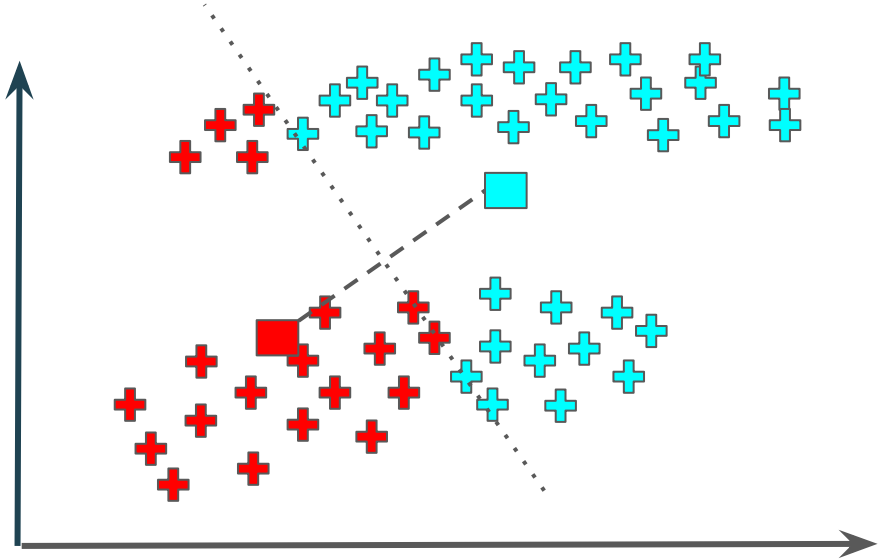
Example 2 - Assign data points



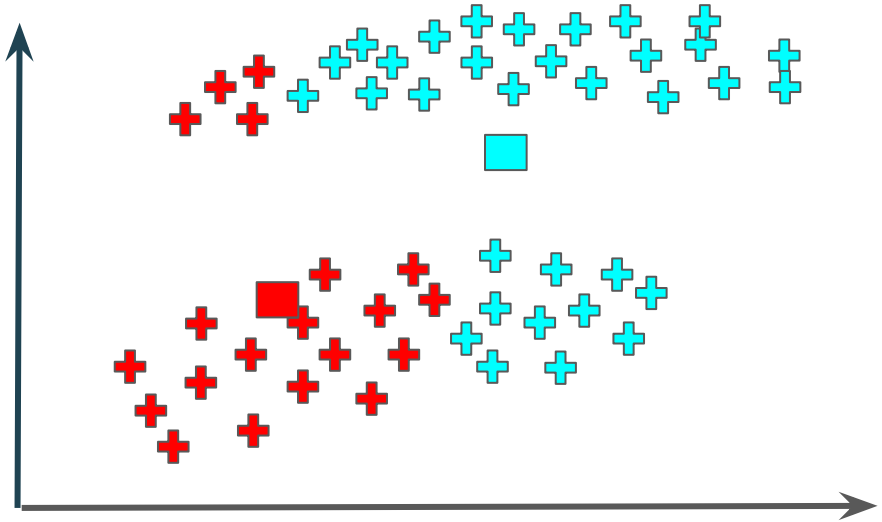
Example 2 - Calculate Centroids



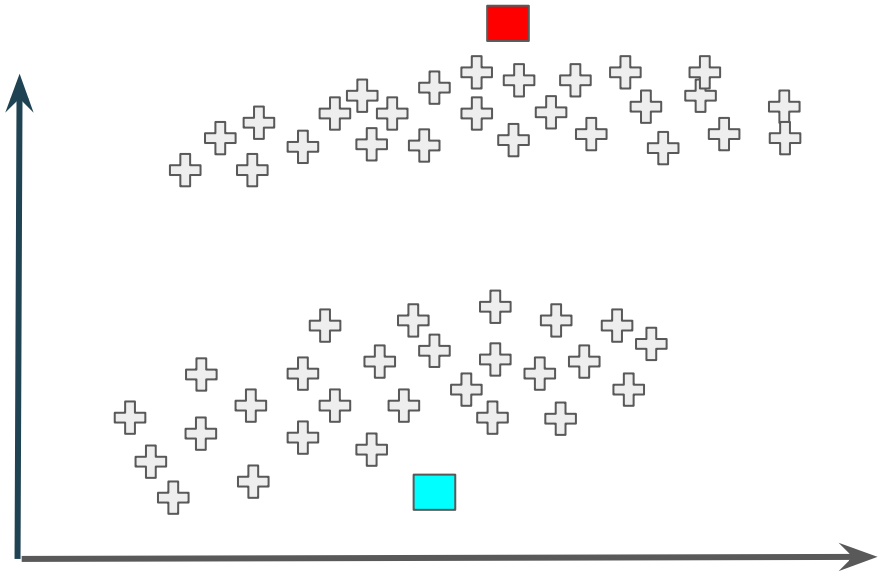
Example 2 - Calculate Centroids



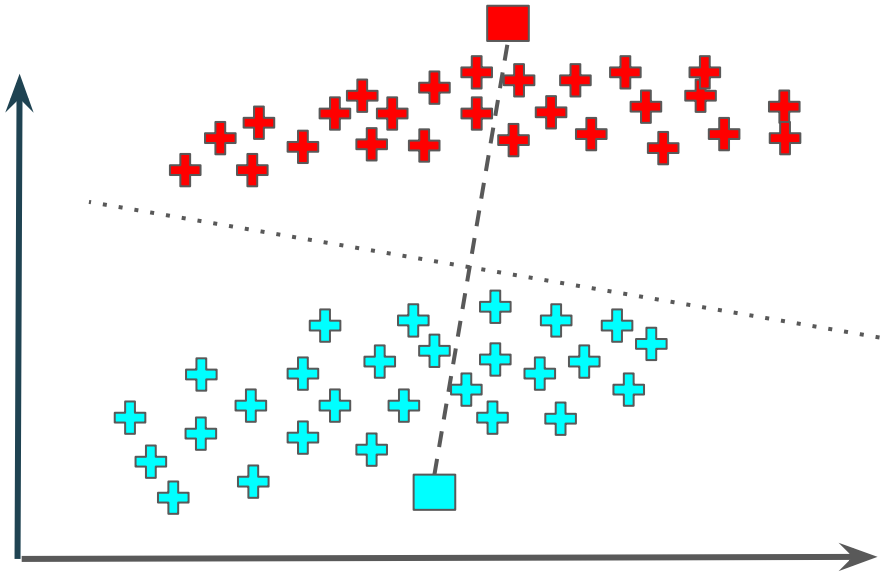
Example 2 - Calculate Centroids



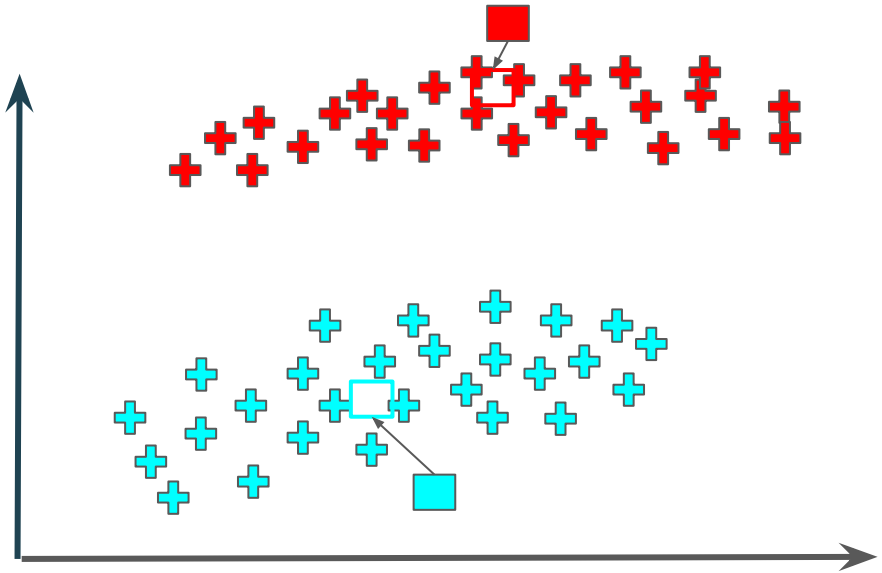
Example 3 - K=2



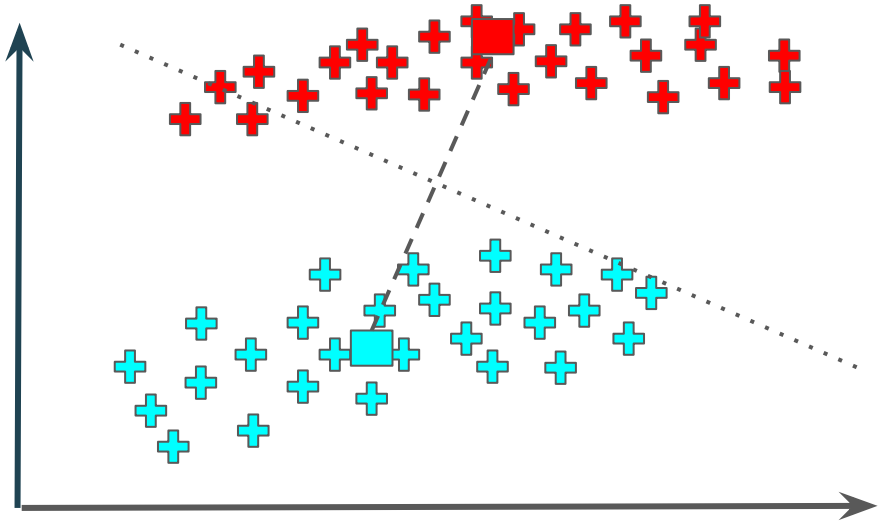
Example 3 - Assign data points



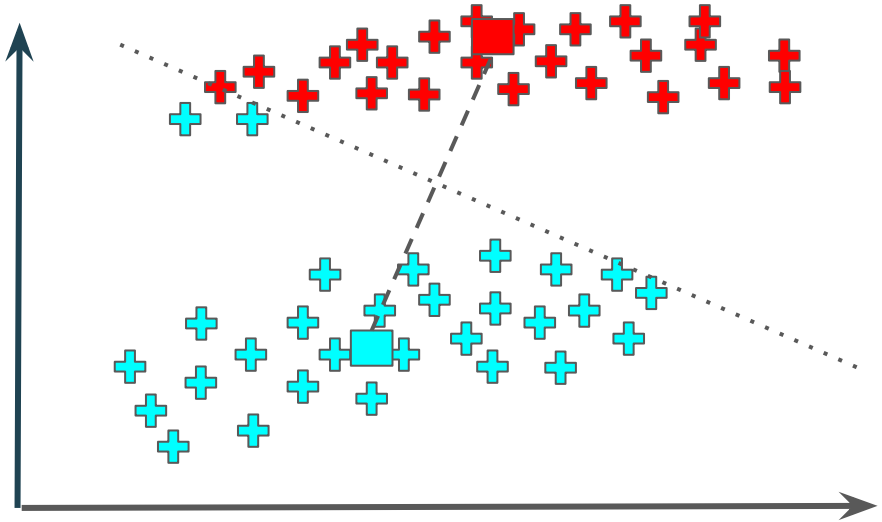
Example 3 - Calculate centroids



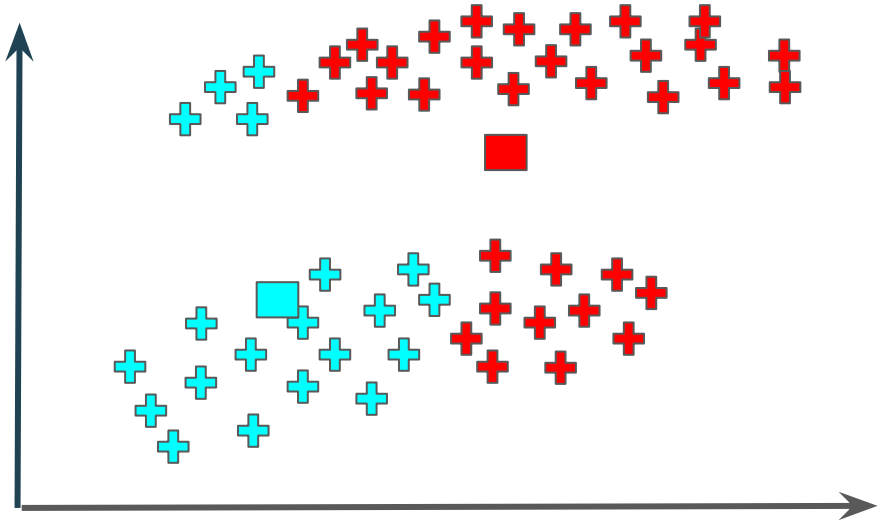
Example 3 - Assign data points



Example 3 - Assign data points



Example 3 - Fast forward

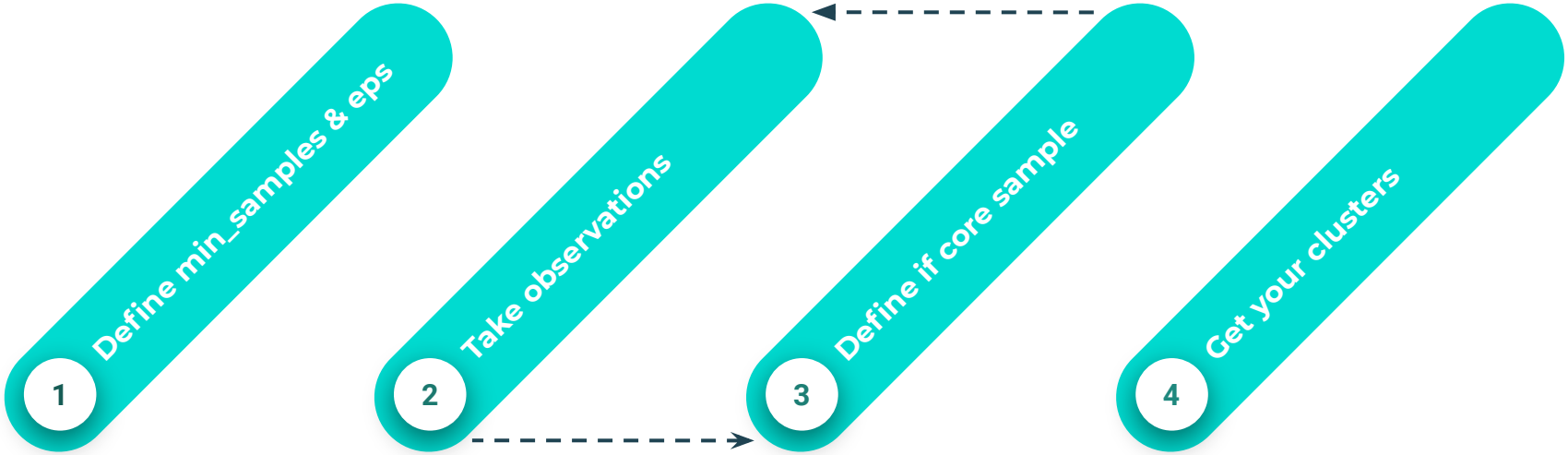


DBScan

New approach

- **Density** ⇒ DBScan will create clusters based on how close each samples are from each other

Process

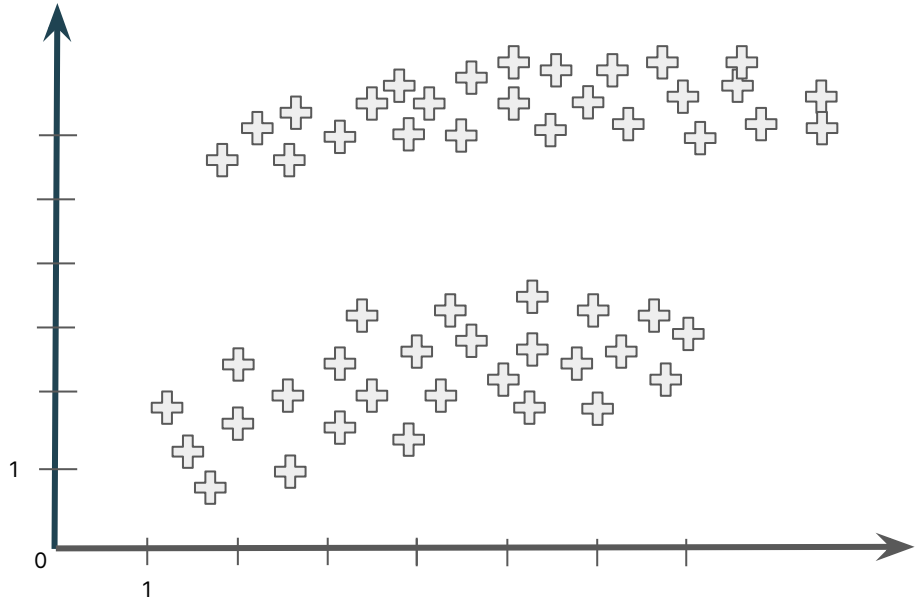

1
Define min_samples & eps
2
Take observations
3
Define if core sample
4
Get your clusters

Core metrics

- **Minimum Sample** ⇒ How many observations to create a core sample
- **Epsilon** ⇒ Maximum distance to define an observation as part of a sample

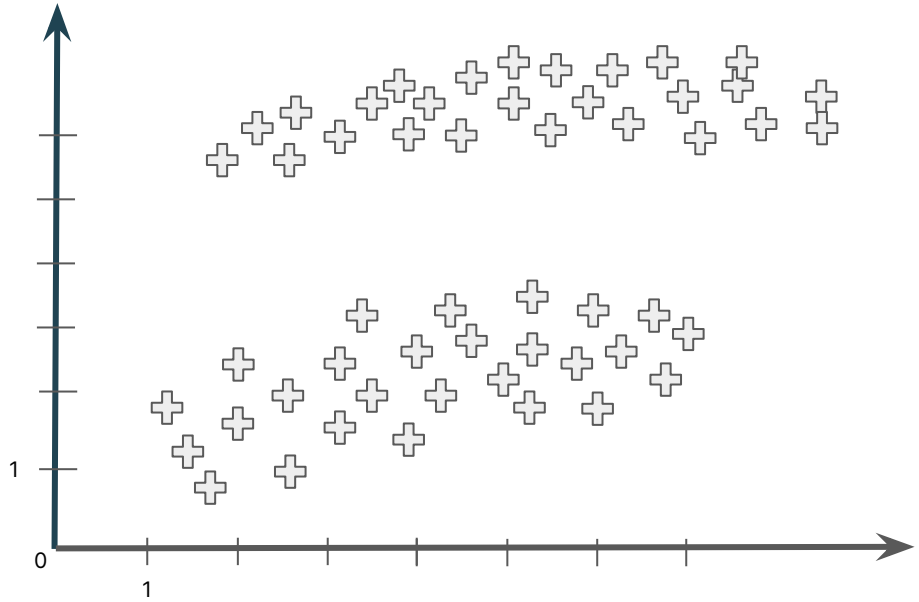
Example - Define min_sample & eps

min_sample = 4
eps = 0.2



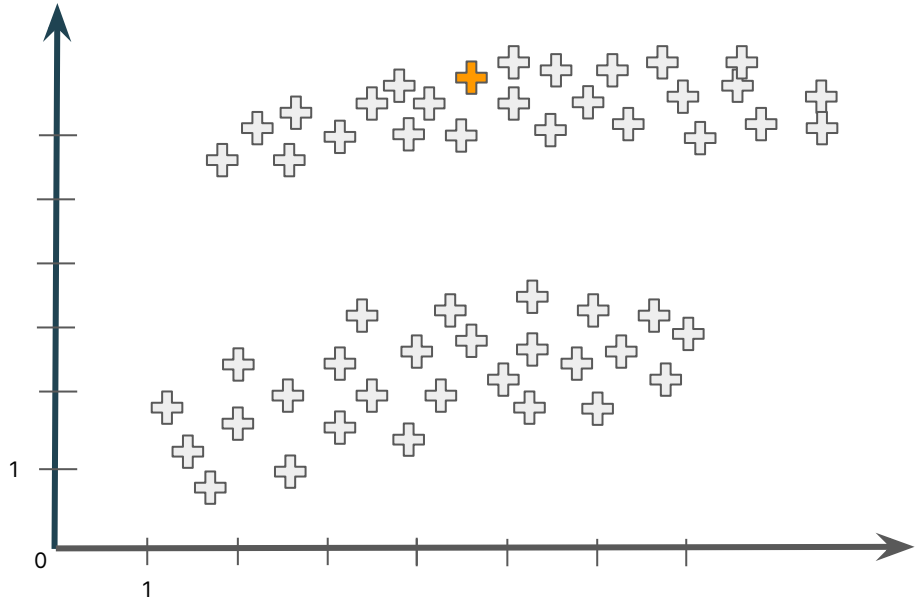
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



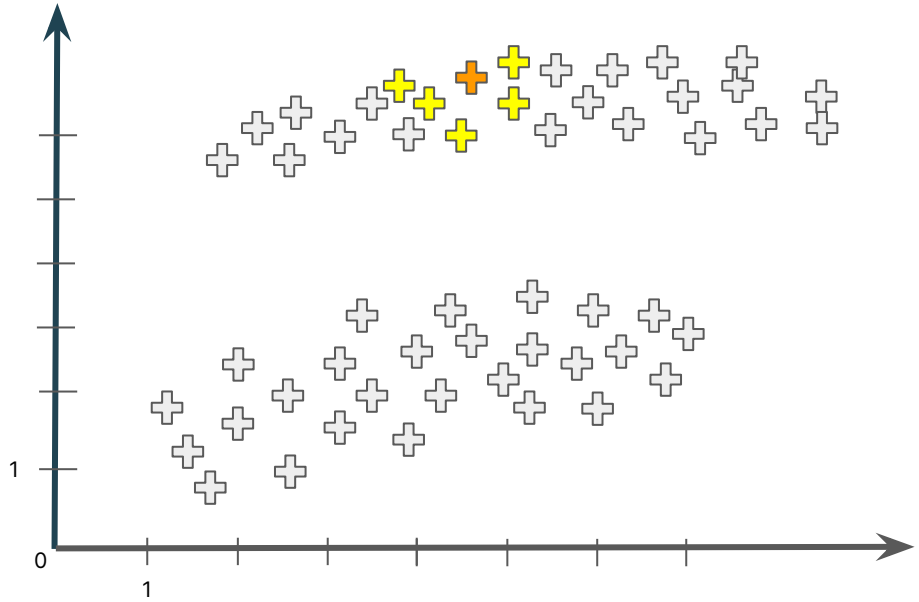
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



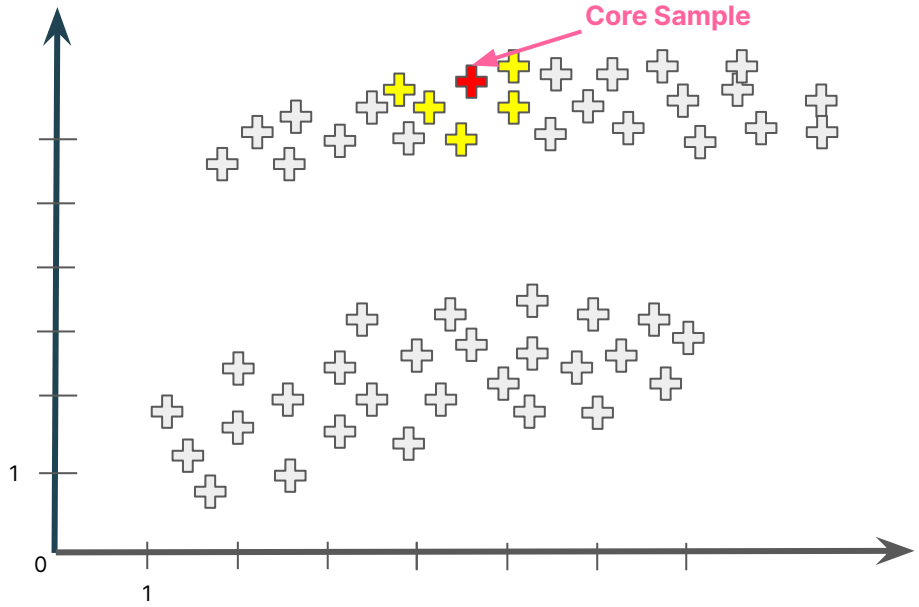
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



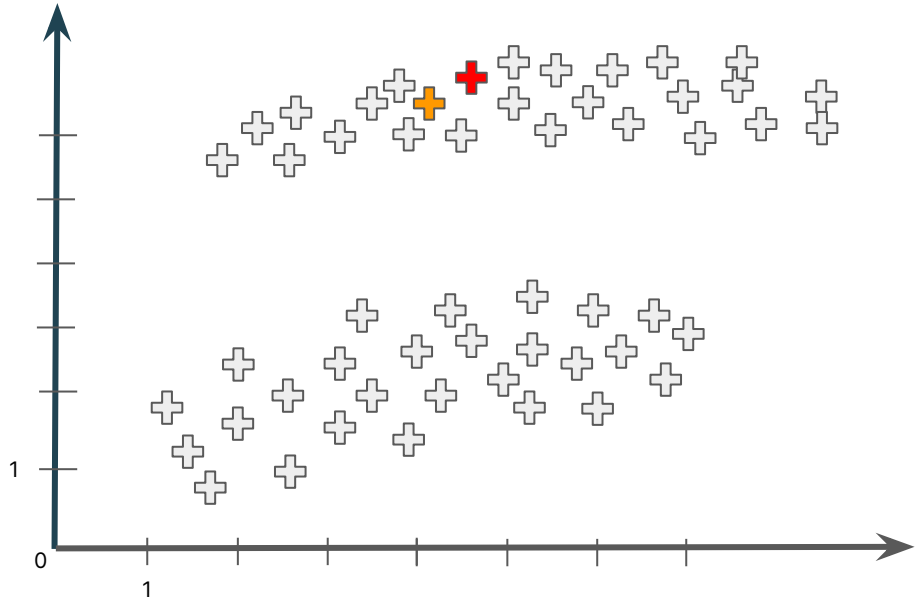
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



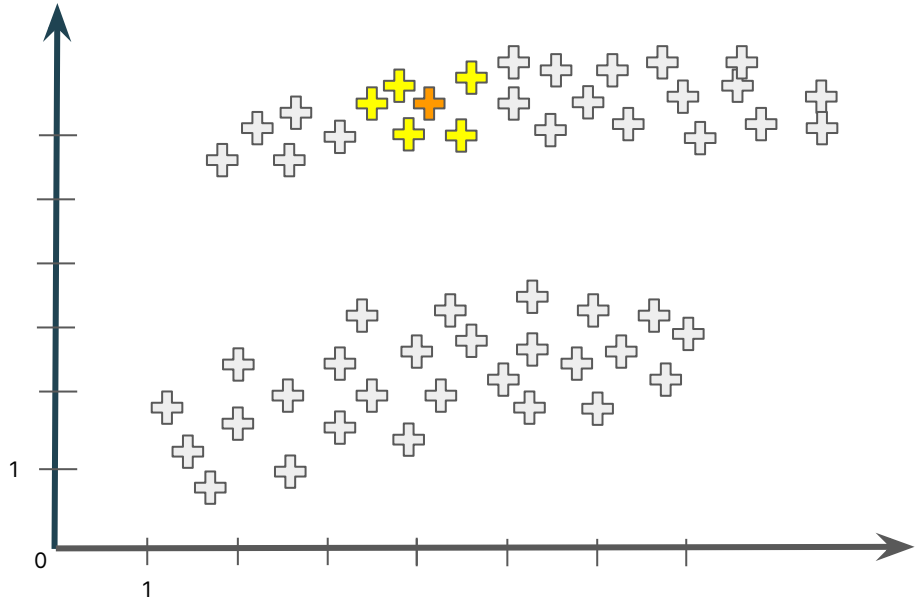
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



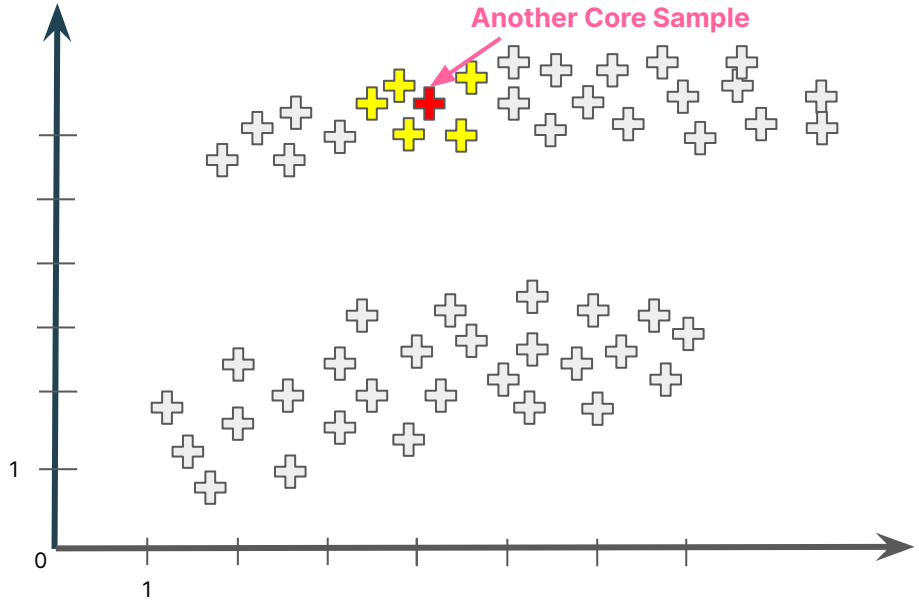
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



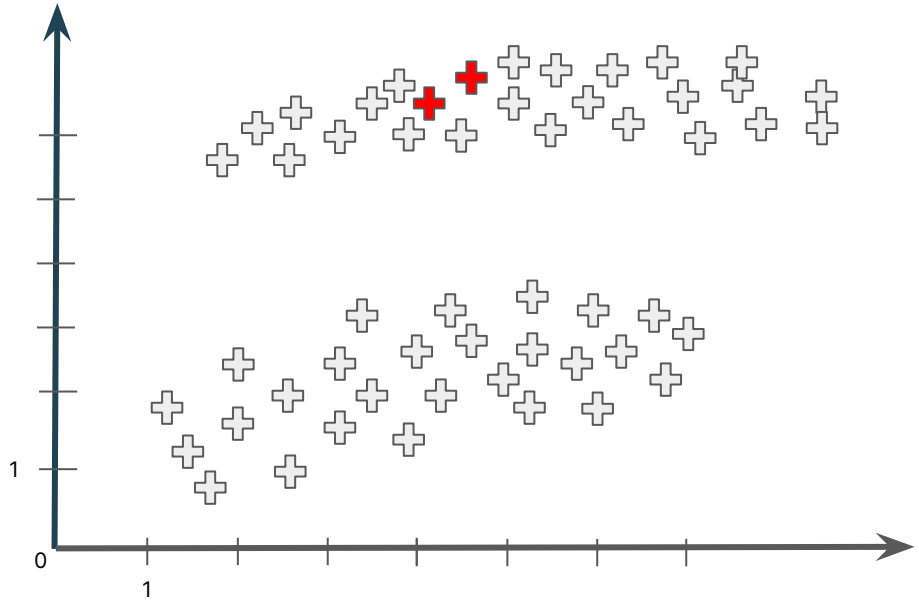
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



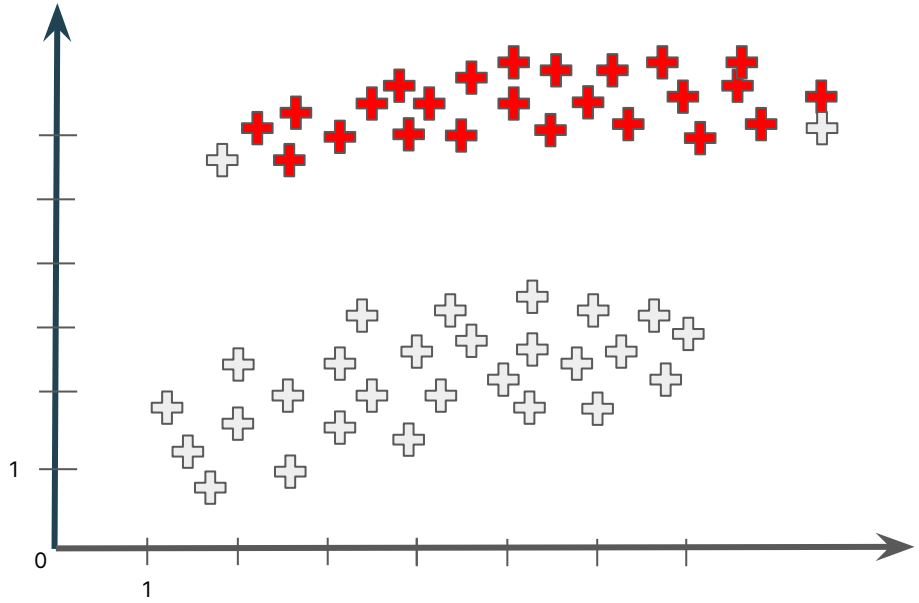
Example - Take observation & define core samples

min_sample = 4
eps = 0.2



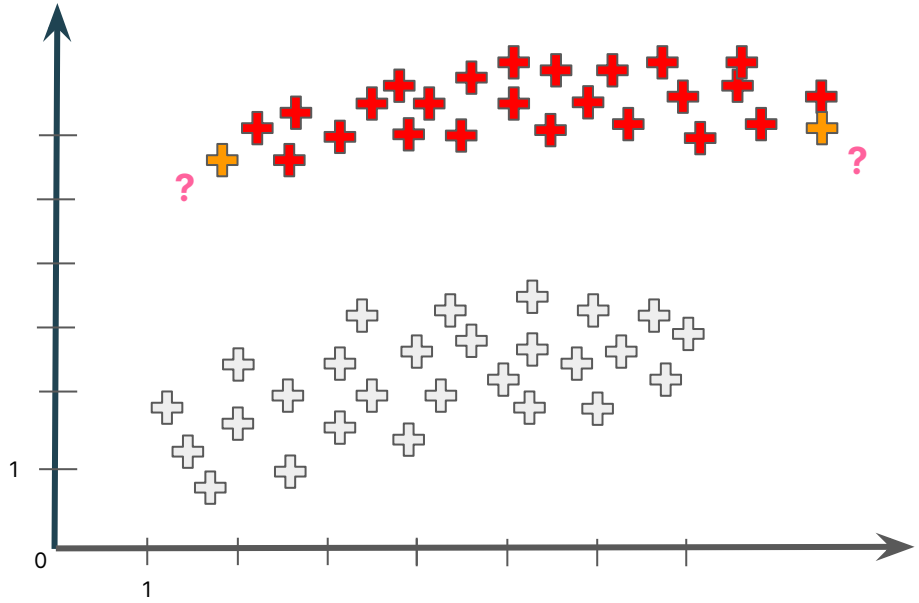
Example - Fast Forward

min_sample = 4
eps = 0.2



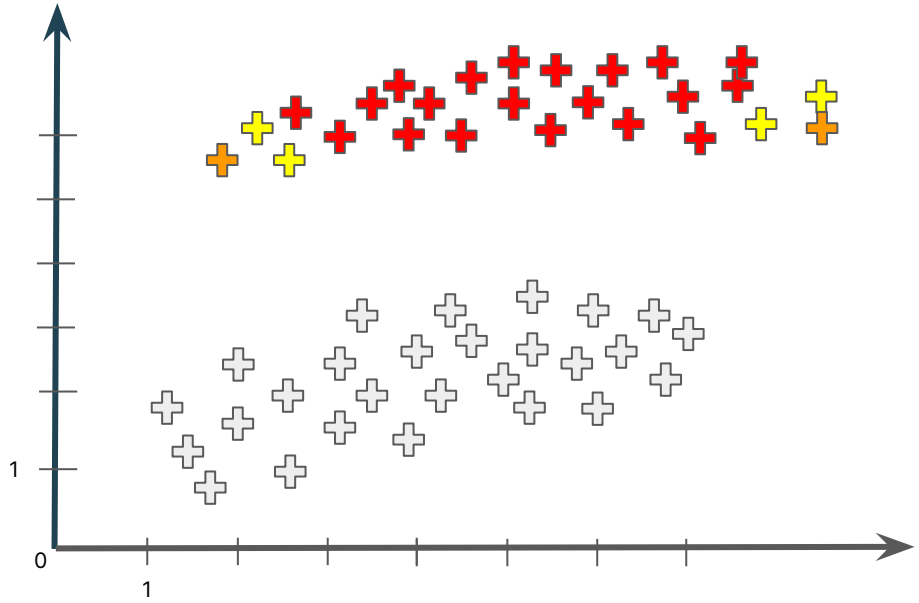
Example - Non-core samples

min_sample = 4
eps = 0.2



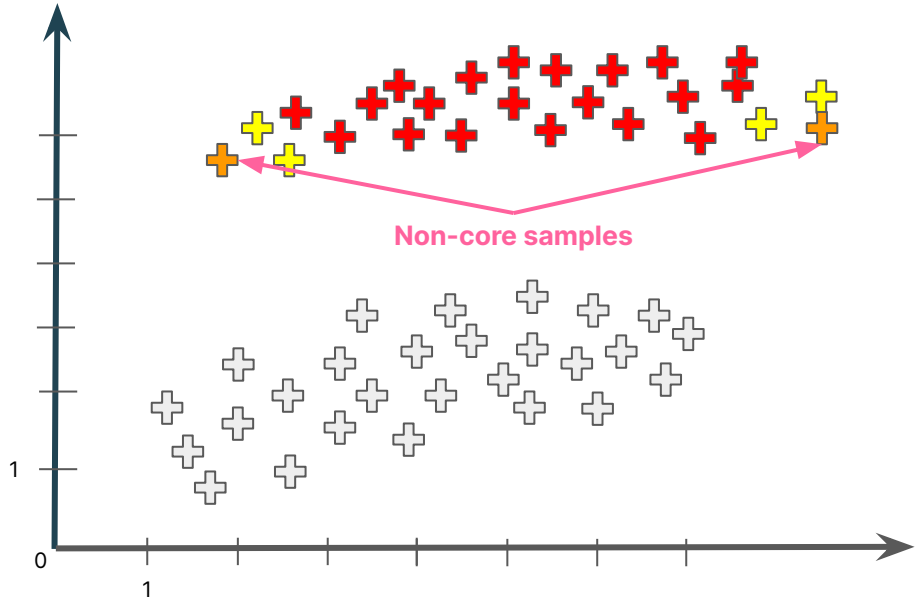
Example - Non-core samples

min_sample = 4
eps = 0.2



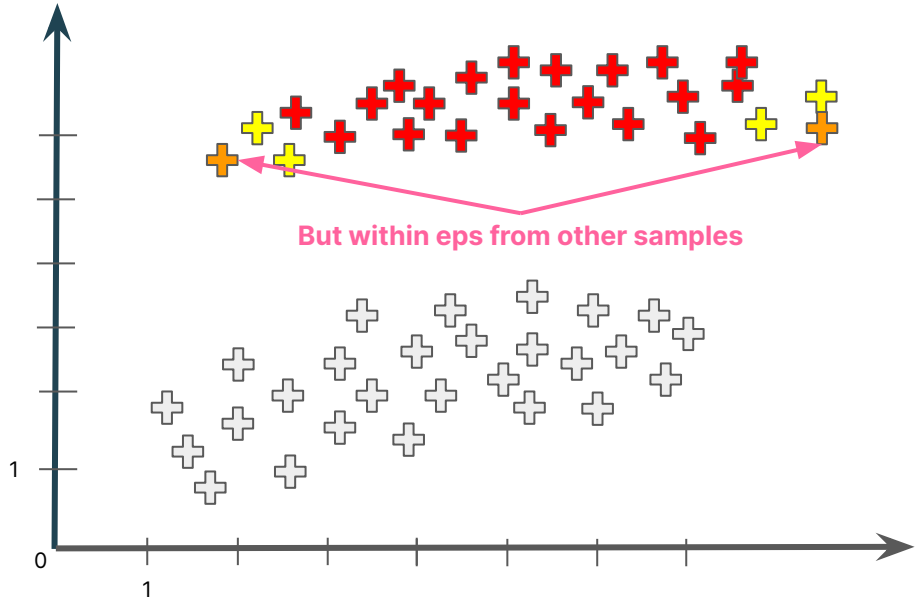
Example - Non-core samples

min_sample = 4
eps = 0.2



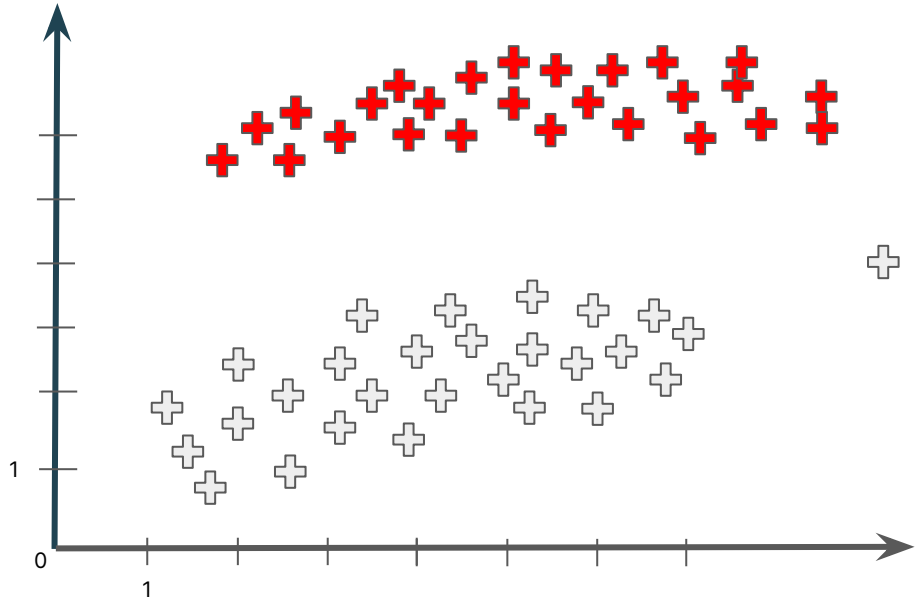
Example - Non-core samples

min_sample = 4
eps = 0.2



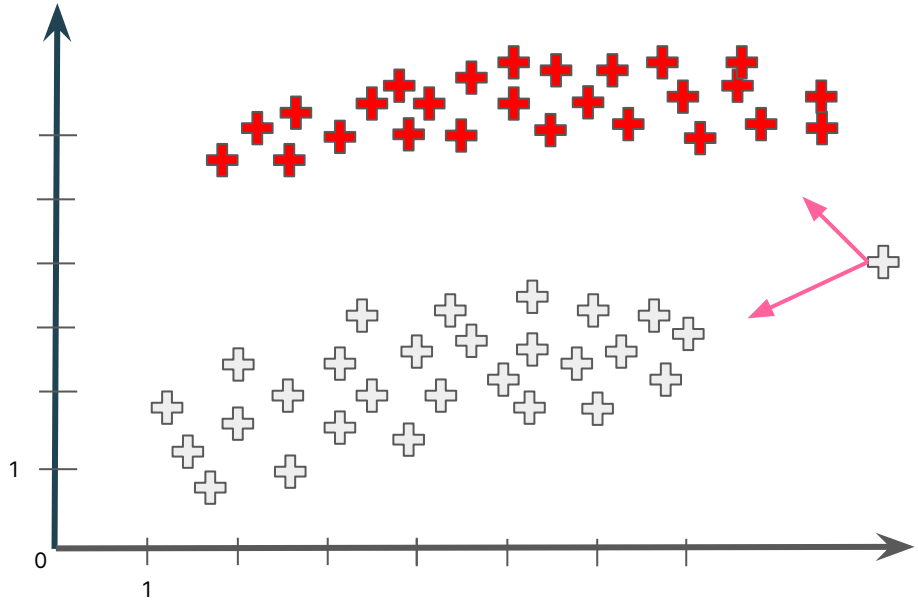
Example - Define outliers

min_sample = 4
eps = 0.2



Example - Define outliers

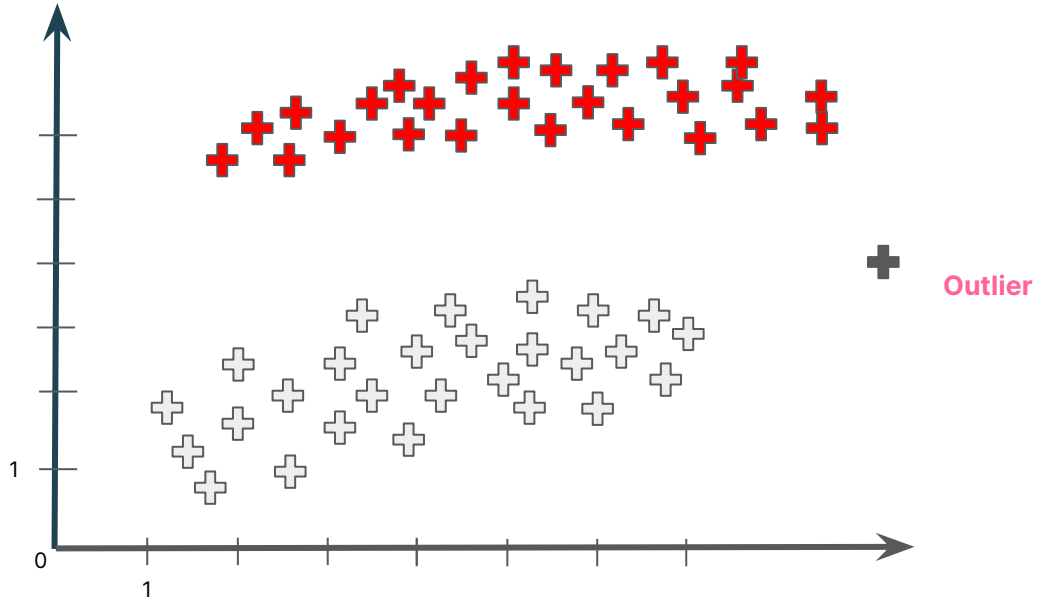
min_sample = 4
eps = 0.2



No min_samples around
Further than eps away

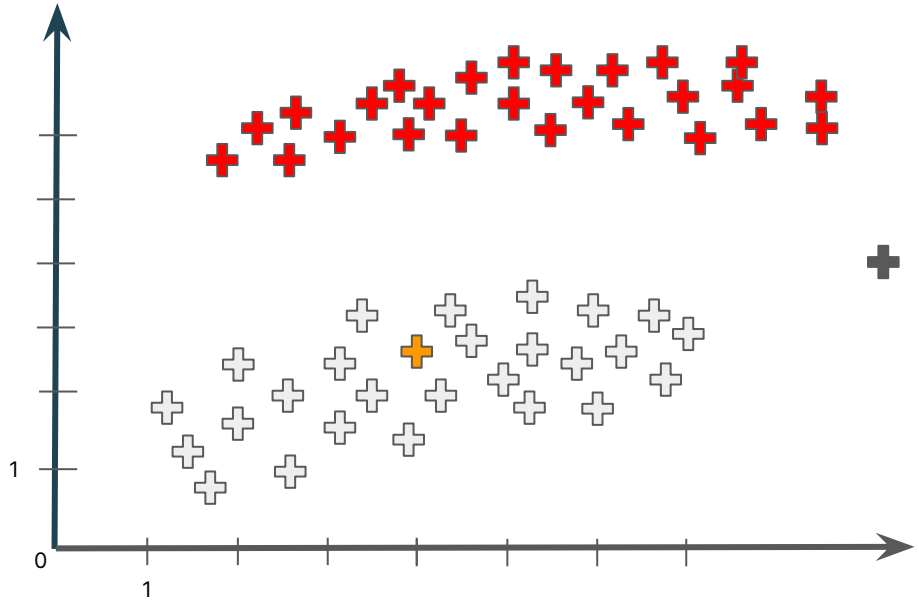
Example - Define outliers

min_sample = 4
eps = 0.2



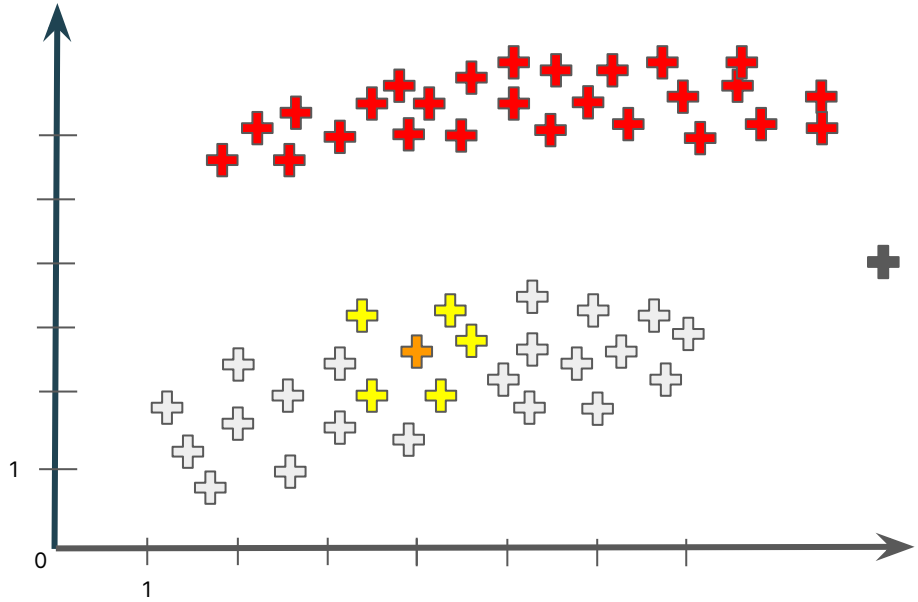
Example - Same process with second cluster

min_sample = 4
eps = 0.2



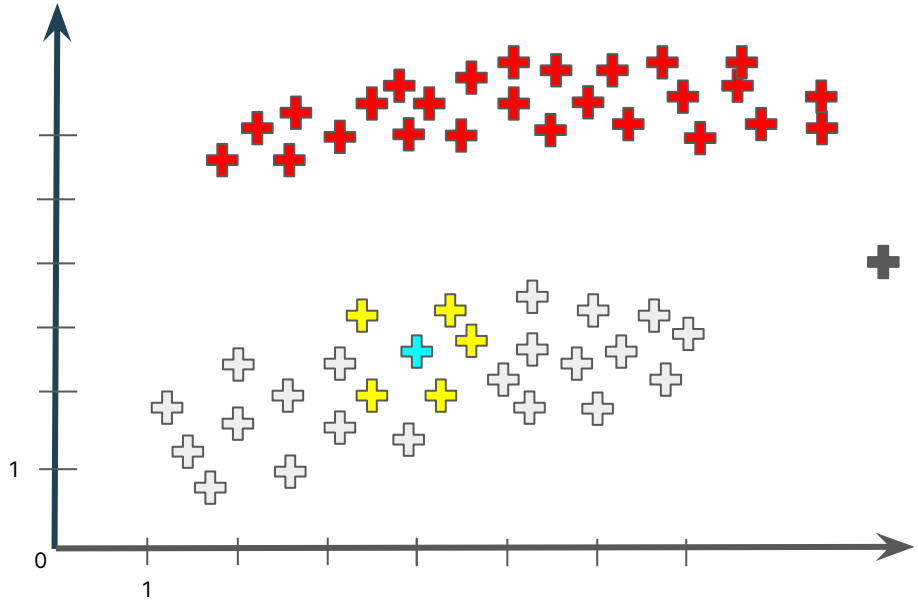
Example - Same process with second cluster

min_sample = 4
eps = 0.2



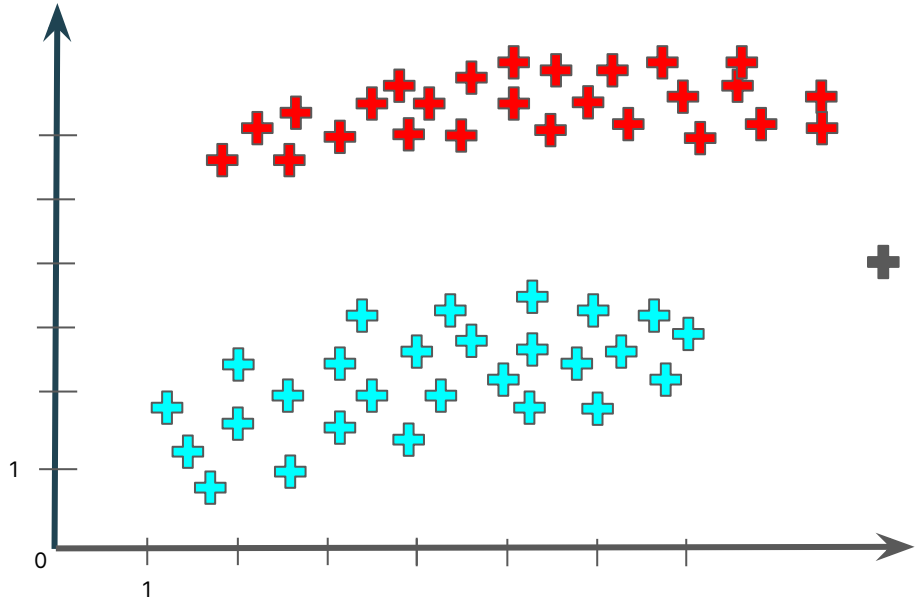
Example - Same process with second cluster

min_sample = 4
eps = 0.2



Example - Same process with second cluster

min_sample = 4
eps = 0.2



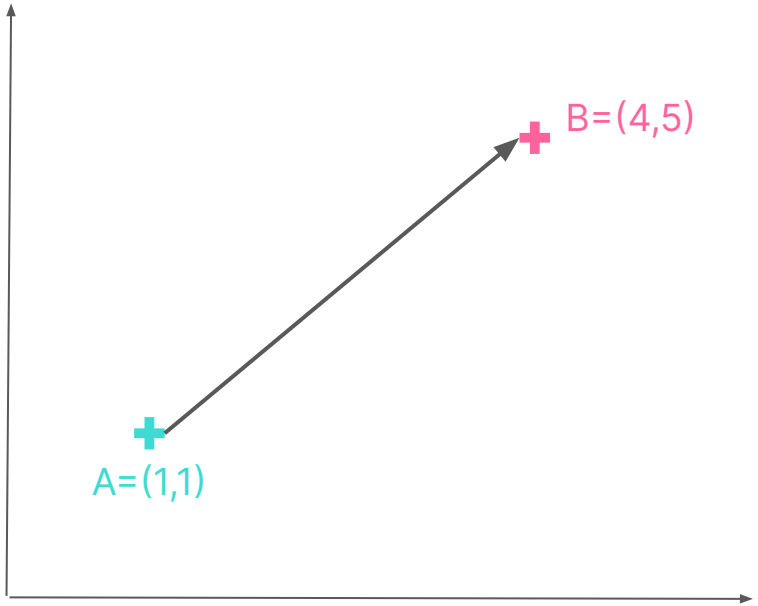
How to choose min_sample & eps?

How to set min_sample & eps?

- **Low eps & High min_sample** → High density clusters
- **High eps & Low min_sample** → Low density clusters
- **Low eps & Low min_sample** → High outliers sensitivity
- **High eps & High min_sample** → Low outliers sensitivity

Choose types of distance for eps

Euclidean Distance



$$d_2(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

$$d = \sqrt{(4-1)^2 + (5-1)^2}$$

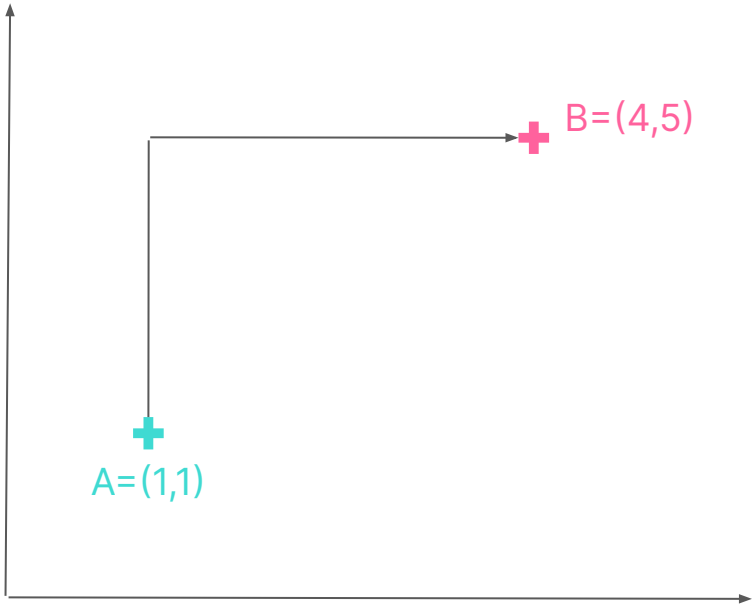
$$d = \sqrt{(3)^2 + (4)^2}$$

$$d = \sqrt{9+16}$$

$$d = \sqrt{25}$$

$$d = 5$$

Manhattan Distance



$$d_1(x, y) = \sum_{i=1}^{p} |x_i - y_i|$$

$$d = |4 - 1| + |5 - 1|$$

$$d = 3 + 4$$

$$d = 7$$

Thanks!

See you in the next course