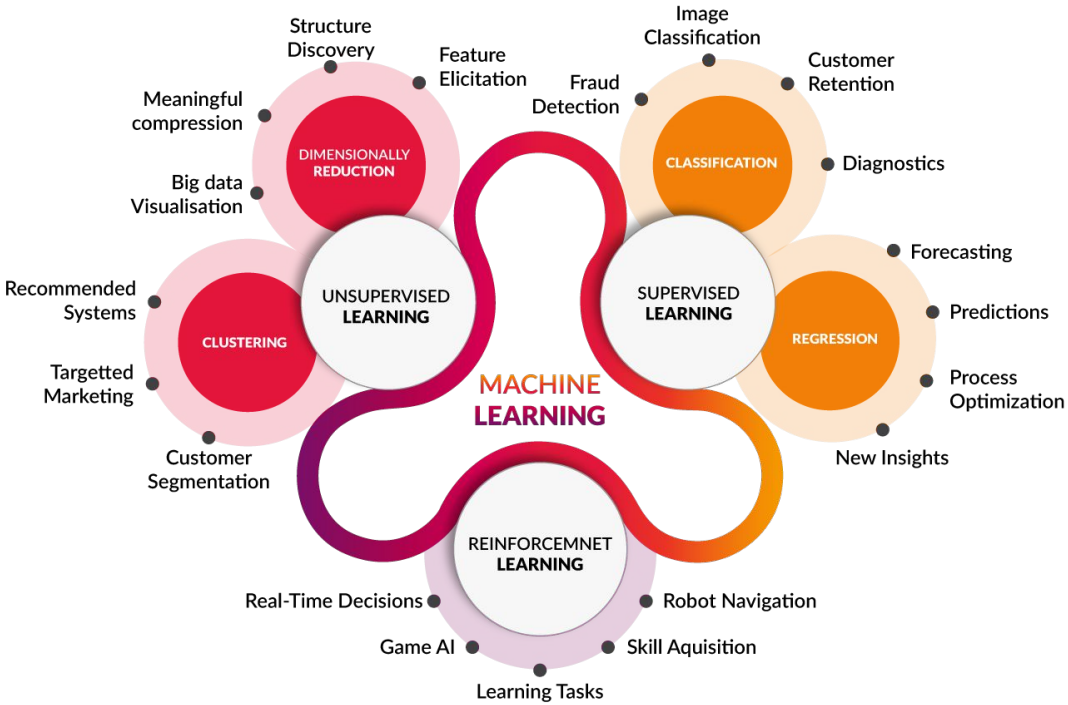
K-Means

Clustering

Machine Learning types



Unsupervised Learning usecases

- **Dimensionality Reduction** ⇒ Compress datasets
- **Clustering** ⇒ Divide data into meaningful groups

Supervised vs Unsupervised

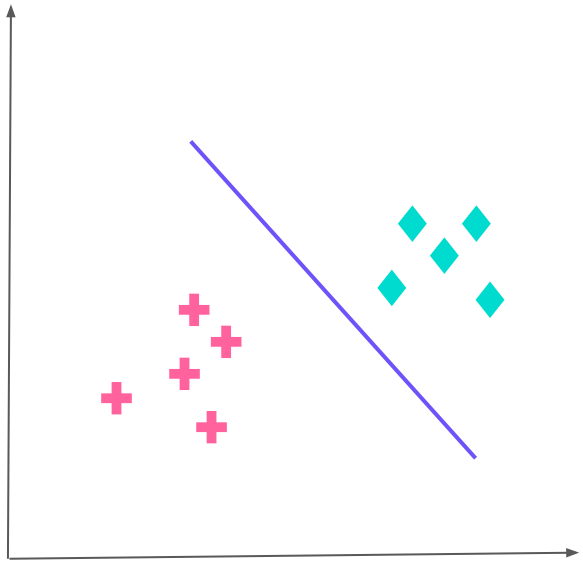
Predict

X1	X2	Y
x1	*x2*	*y1*
...	...	...
xn	*xn*	*yn*

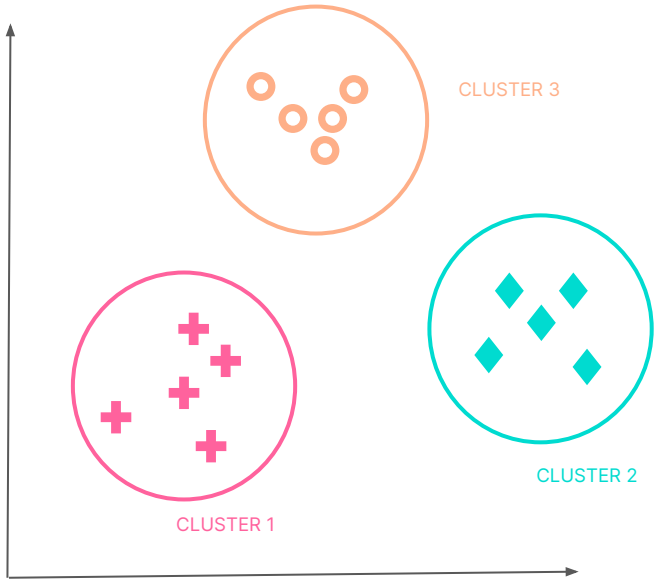
VS

X1	X2	~~Y~~
x1	*x2*	~~*y1*~~
...	...	~~...~~
xn	*xn*	~~*yn*~~

Supervised vs Unsupervised



Supervised classification



Unsupervised Clustering

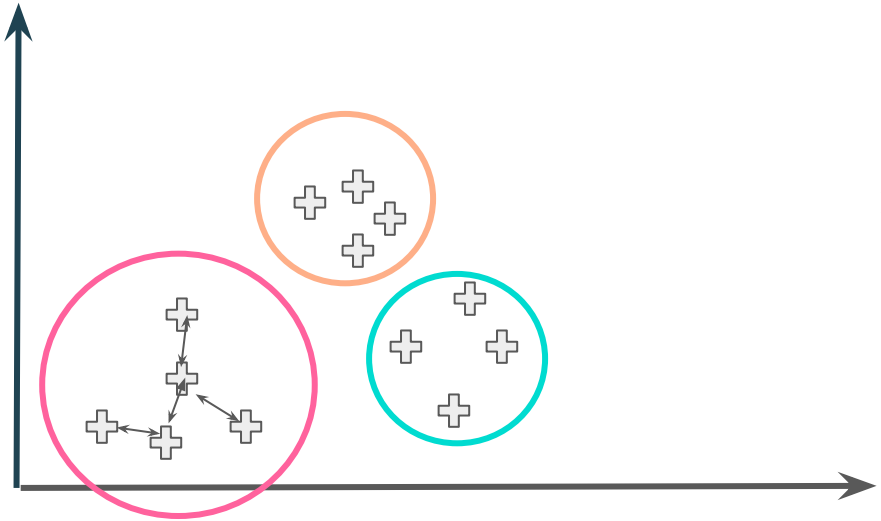
K-Means

K-Means Algorithm

*The **KMeans** algorithm clusters data by trying to **separate samples in n groups** of equal variance, minimizing a criterion known as the **inertia or within-cluster sum-of-squares***

Source: Sklearn

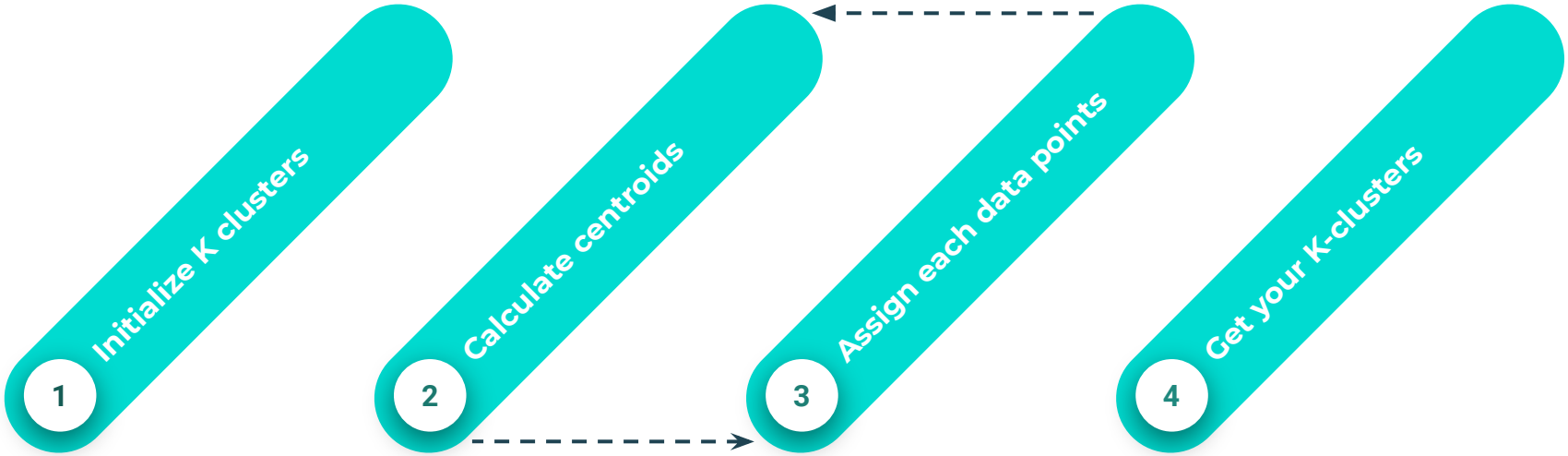
K-Means Algorithm



WCSS = $$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

C = clusters
μ = mean within a given cluster
x = data point

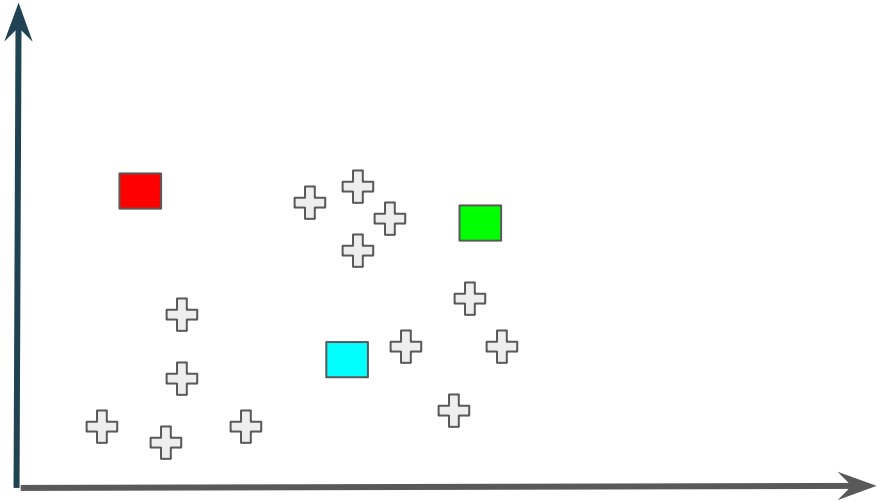
Process


1
Initialize K clusters
2
Calculate centroids
3
Assign each data points
4
Get your K-clusters

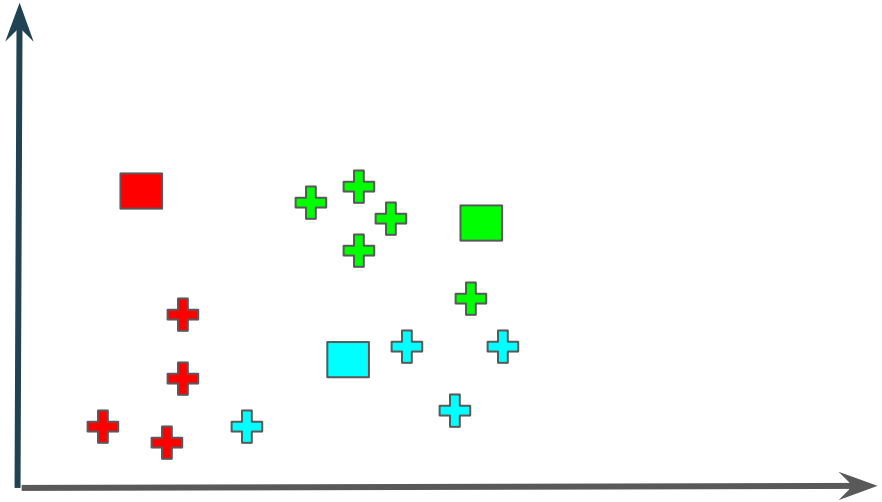
Step 1 - Initialize K clusters

$$K = 3$$

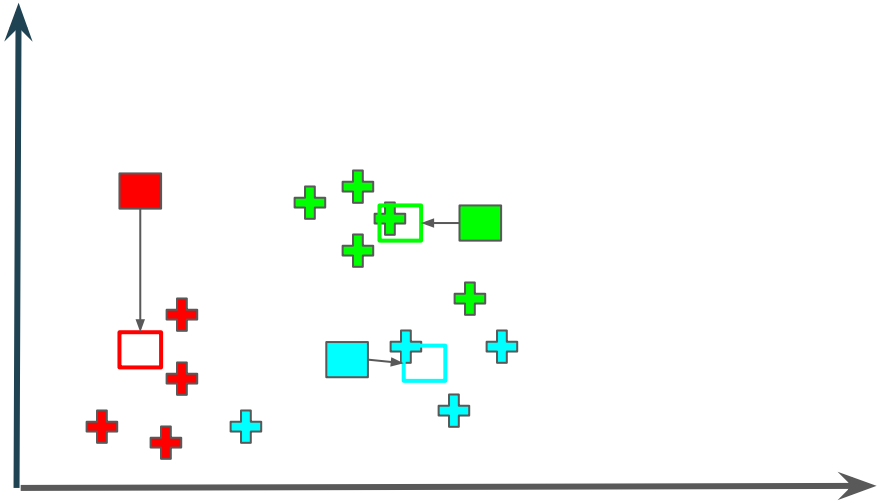
Step 2 - Calculate centroïds



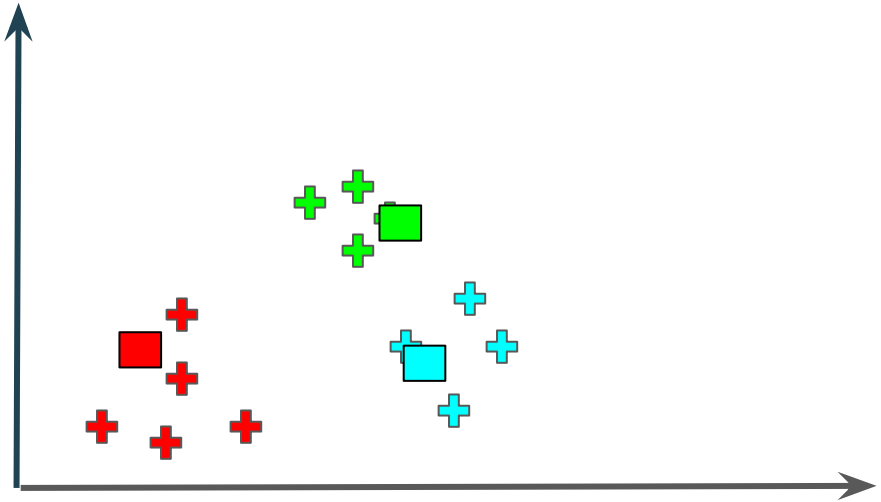
Step 3 - Assign data points to the closest centroid



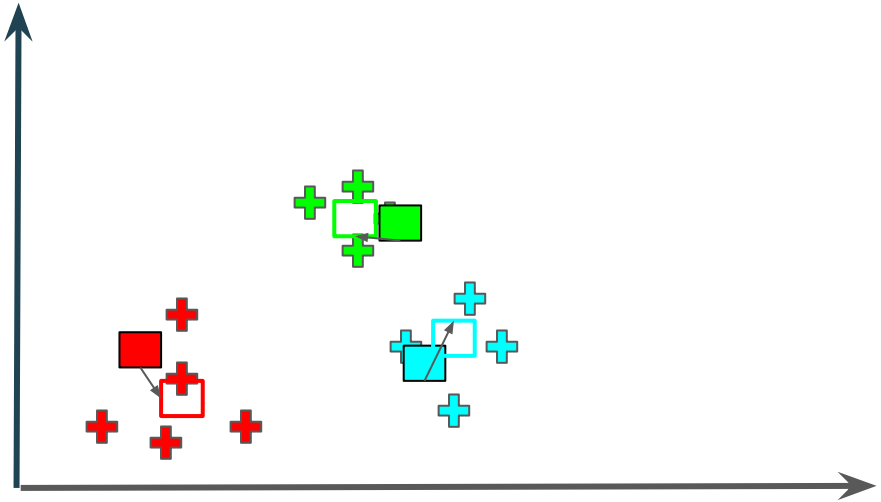
Step 2 - Calculate centroids



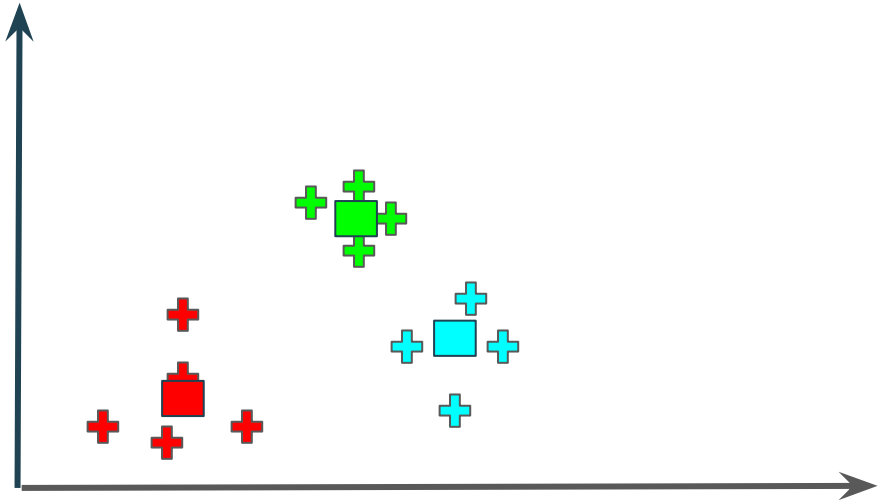
Step 3 - Assign data points to the closest centroid



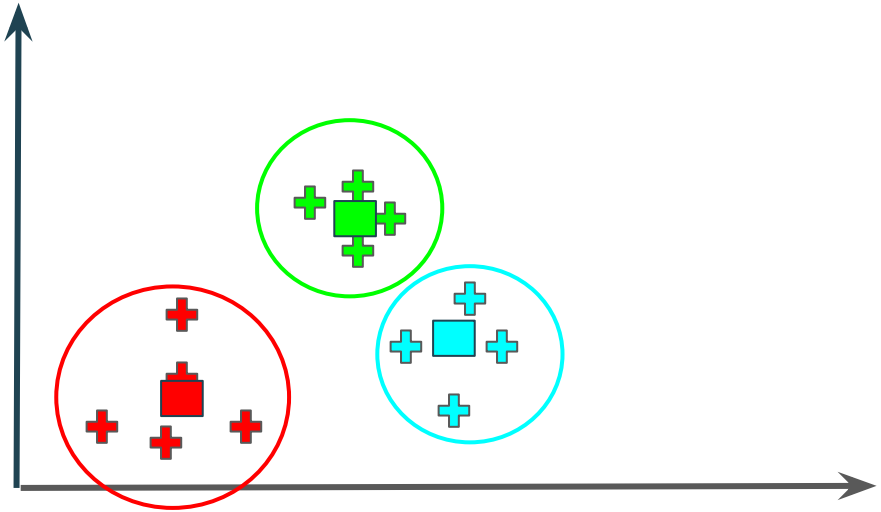
Step 2 - Calculate centroids



Step 3 - Assign data points to the closest centroid



Step 4 - Get our K clusters

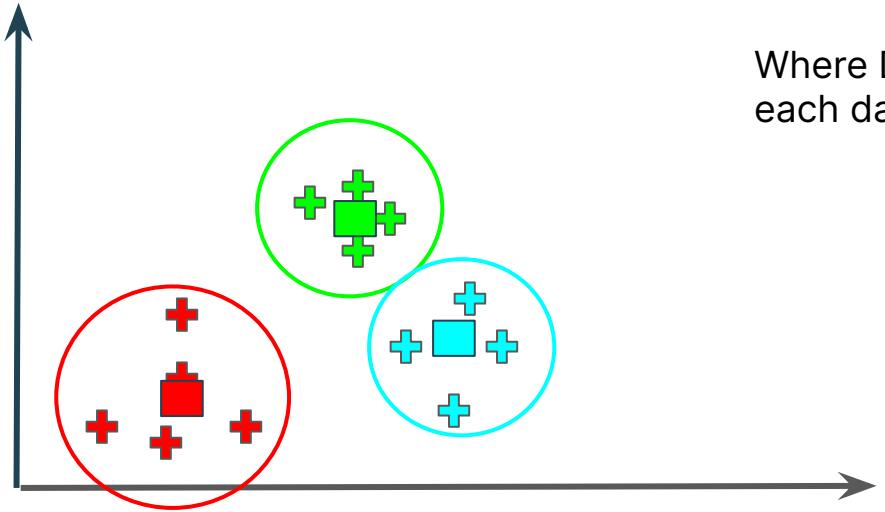


How to get the optimal number K?

Two methods

- **Elbow** ⇒ Check if data points within a cluster are close from their centroid
- **Silhouette** ⇒ Check if clusters are far from each other

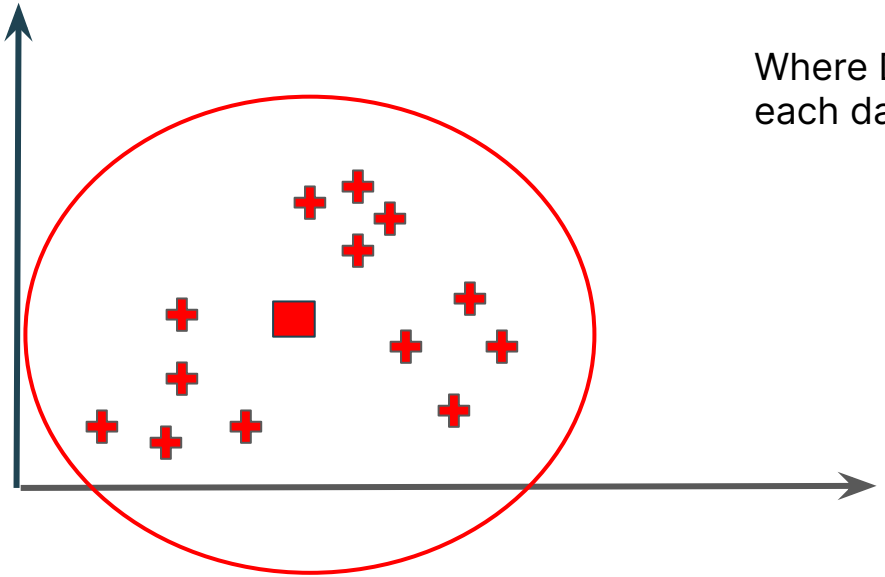
Elbow



WCSS = D(,) + D(,) + D(,)

Where D() is the square distance from each data point to its centroid.

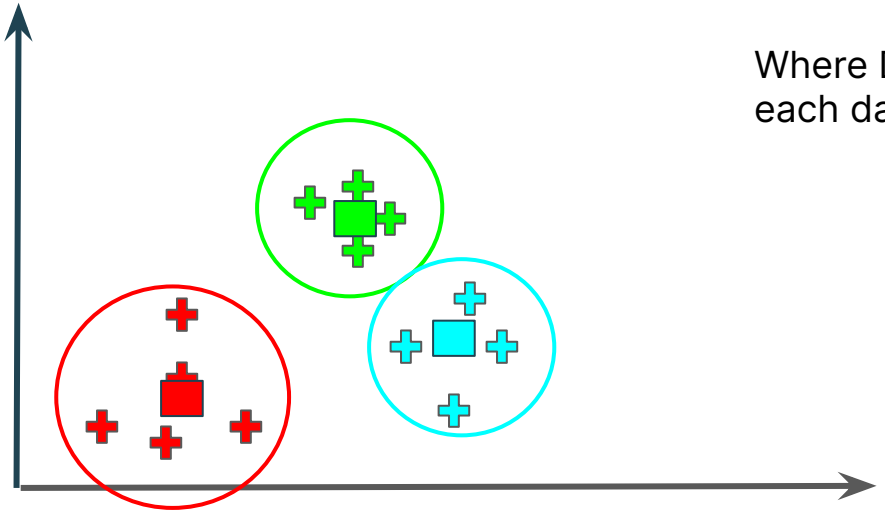
Elbow - WCSS with 1 cluster



WCSS = D(✚, ■)

Where D() is the square distance from each data point to its centroid.

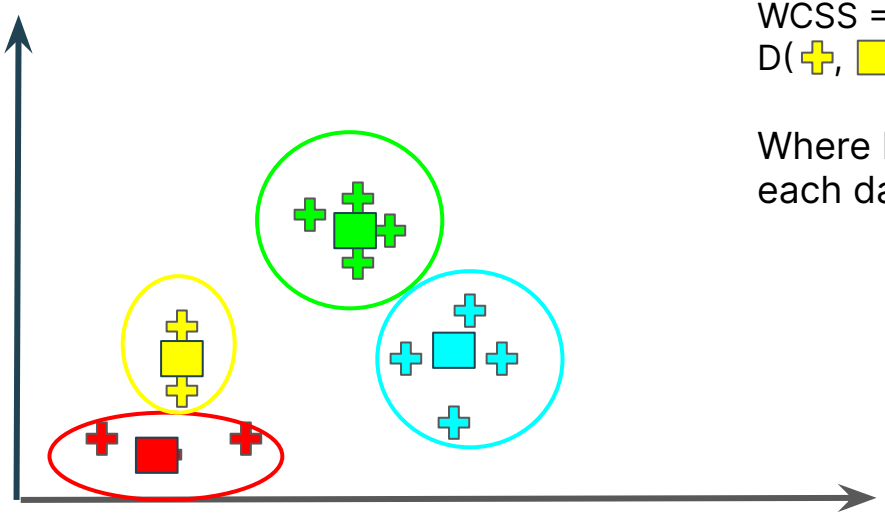
Elbow - WCSS with 3 clusters



WCSS = D(,) + D(,) + D(,)

Where D() is the square distance from each data point to its centroid.

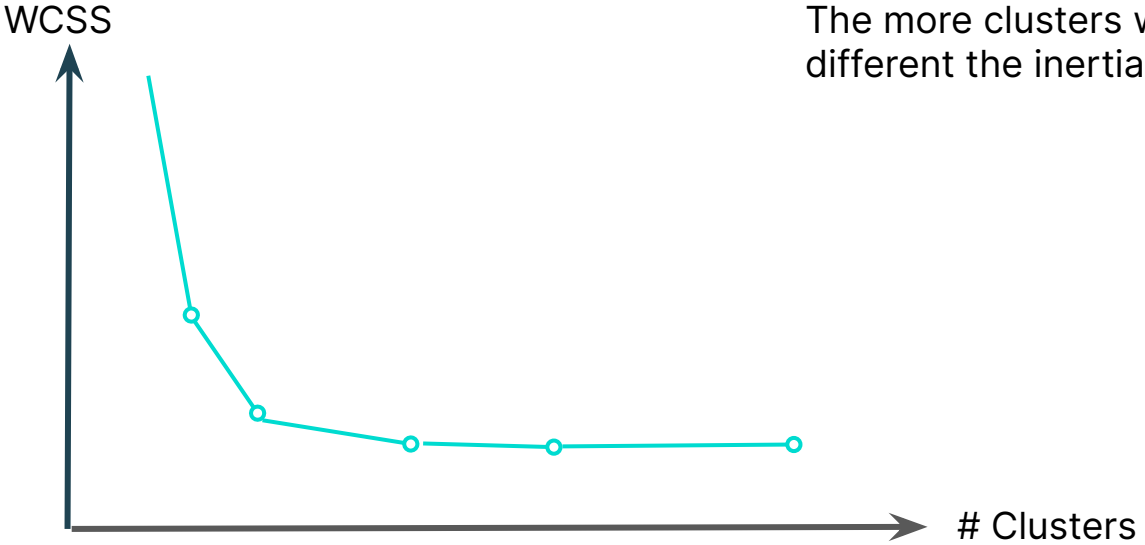
Elbow - WCSS with 4 clusters



WCSS = D(,) + D(,) + D(,) + D(,)

Where D() is the square distance from each data point to its centroid.

Elbow - WCSS with 4 clusters



The more clusters we get the less different the inertia (WCSS) is going to be

Silhouette

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

Where

a → average distance from all data points in the same cluster
b → average distance from all data points in the closest cluster

Silhouette

- **Close to 0** ⇒ Clusters are close from each other
- **Close to 1** ⇒ Clusters are far from each other

Thanks!

See you in the next course