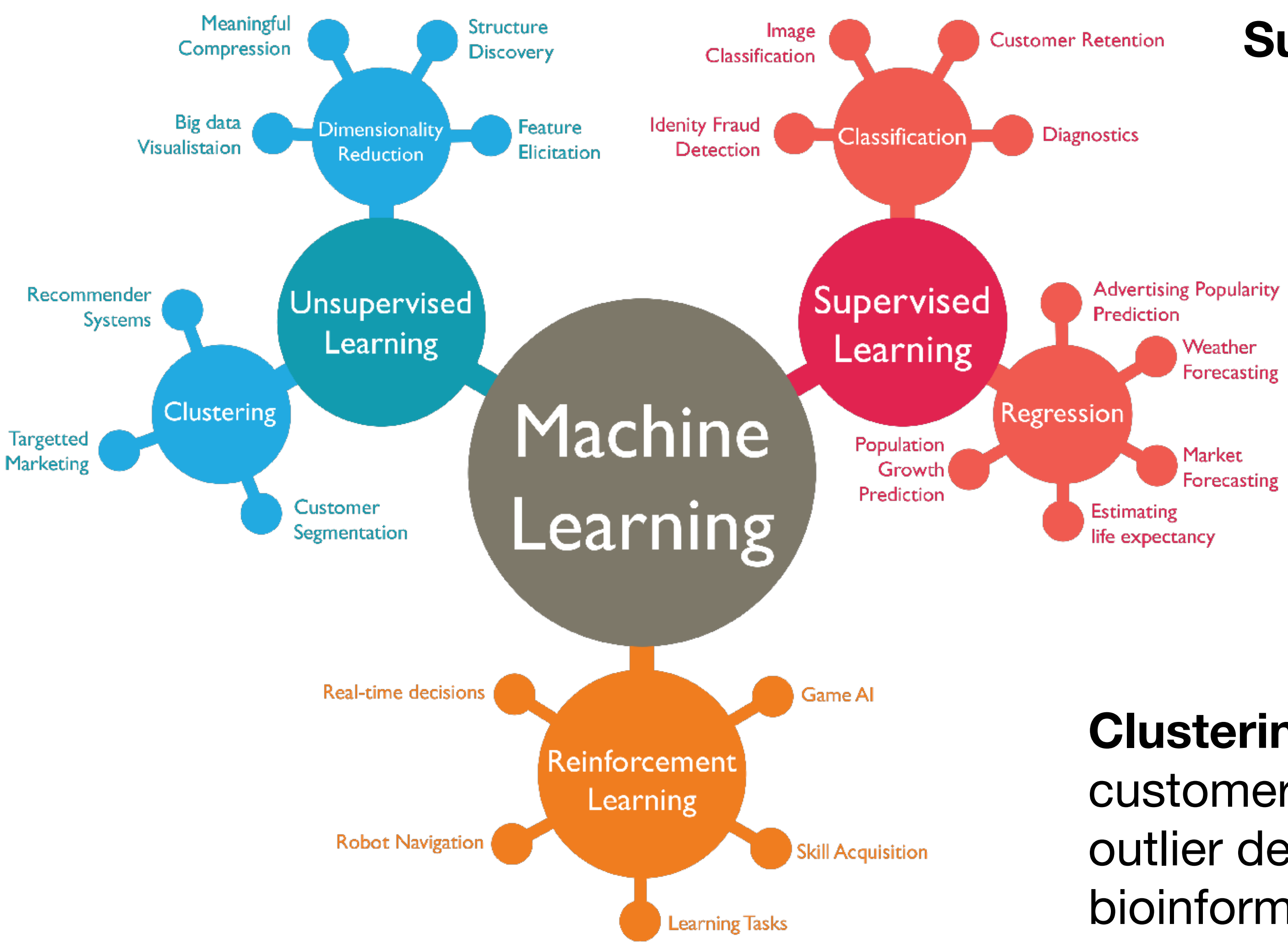
Content

Content

AI & Machine Learning Basics



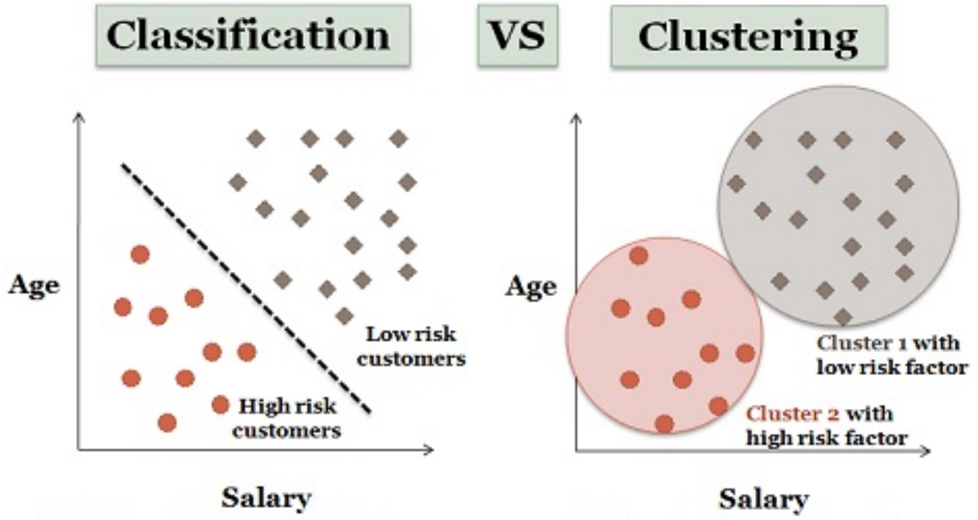
Supervised Learning: Classification & Regression

Labeled dataset
Input —> machine/model —> Output
Correct outputs are provided by the supervisor

Unsupervised Learning: only have input data

Unlabeled dataset
Find regularities from the input

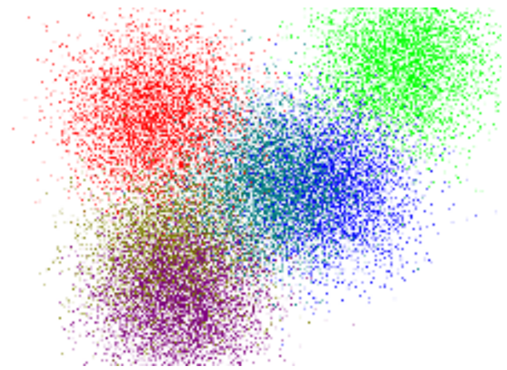
Clustering:
customer segmentation, customer relationship management, outlier detection; Image compression
bioinformatics: DNA, RNA, amino acids, Motif, Proteins, sequence alignments



Risk classification for the loan payees on the basis of customer salary

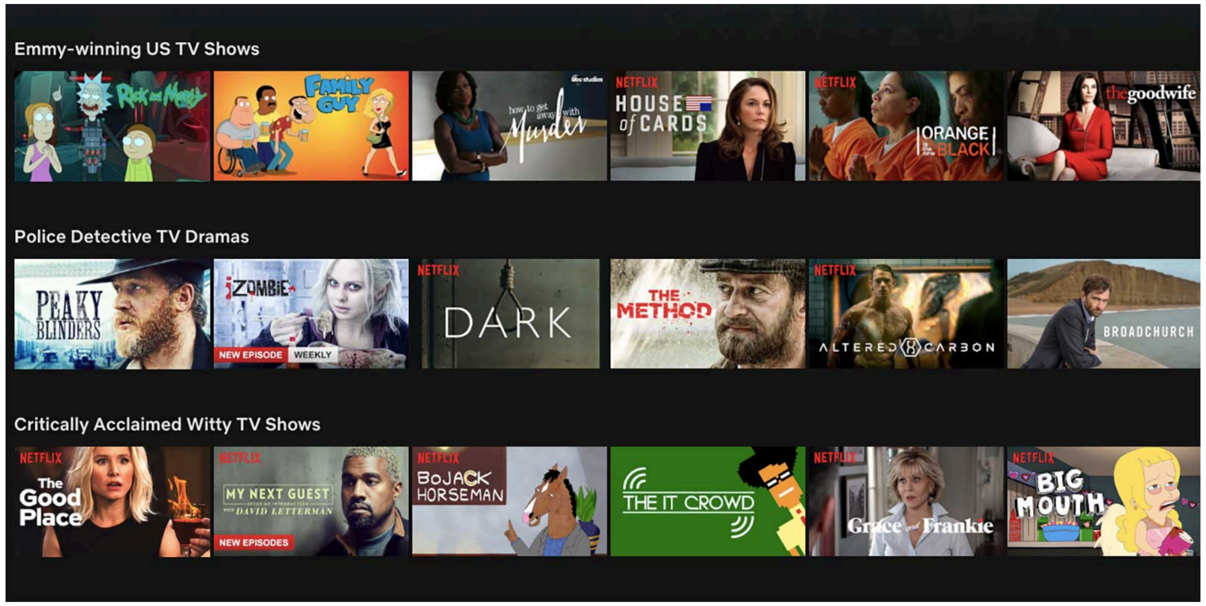
Clustering

- Grouping of data points



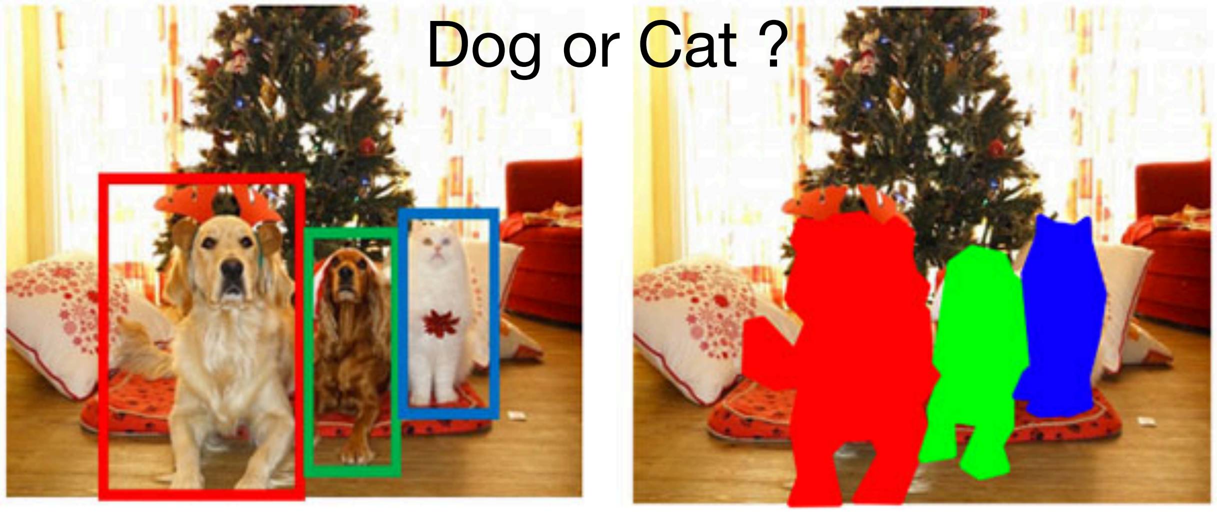
"Clustering" literally means
grouping similar things together

- Recommendation Engines



- Image Segmentation

Dog or Cat ?



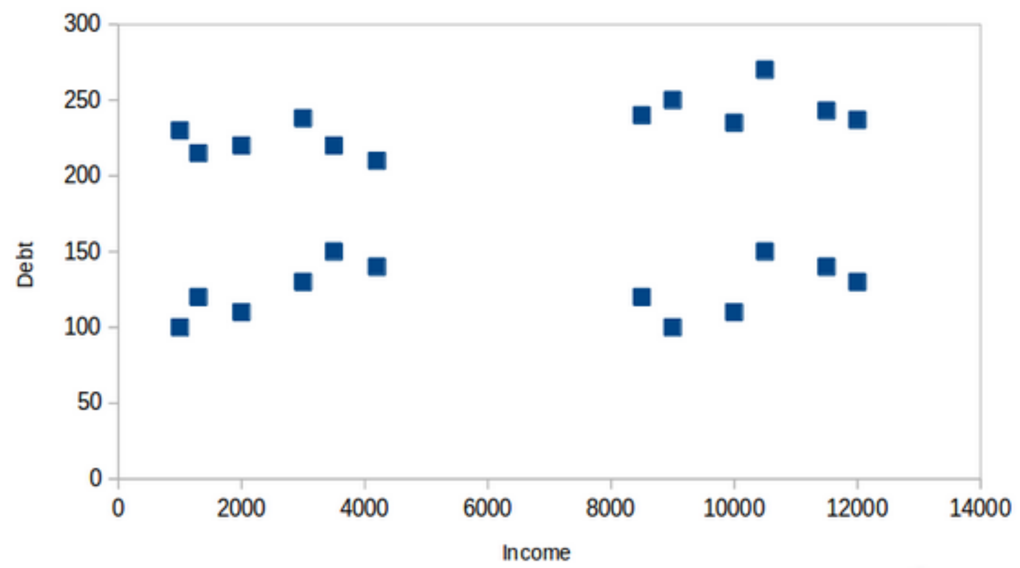
Good references:

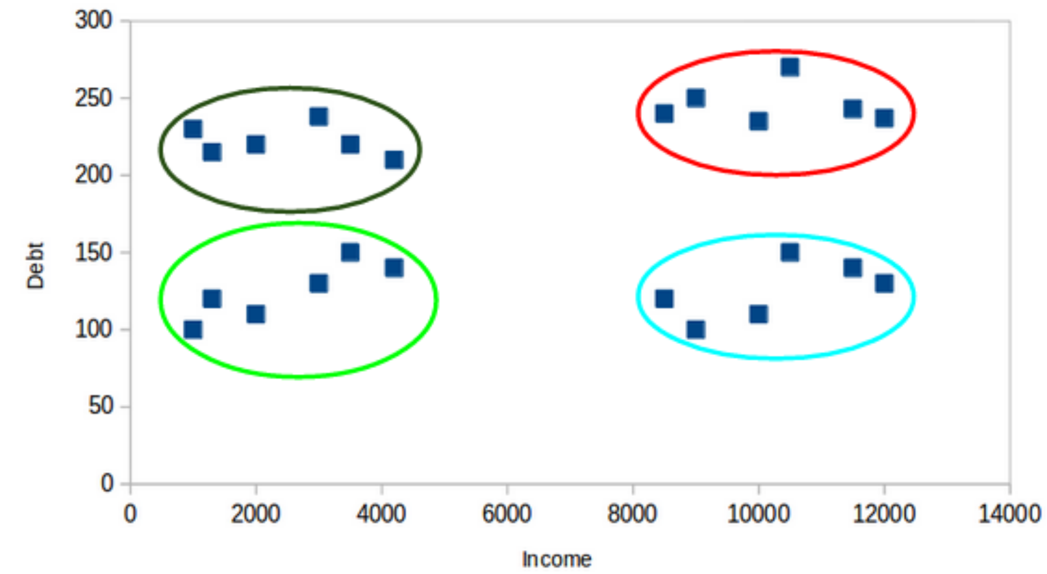
https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a

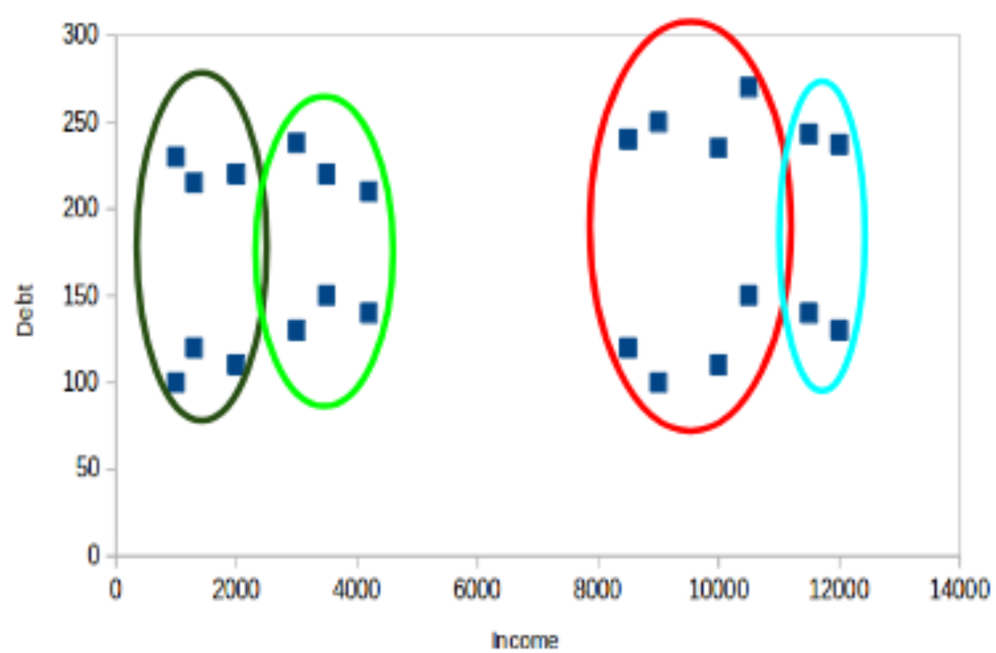
Clustering

- All the data points in a cluster should be similar to one another

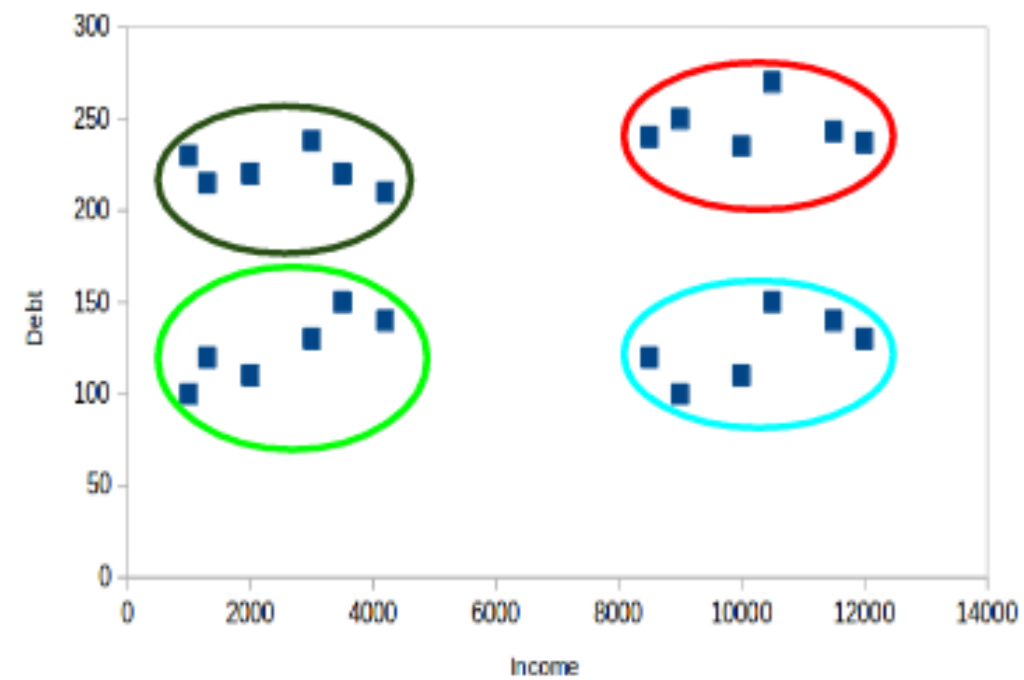




- The data points from different clusters should be as different as possible



Case - I



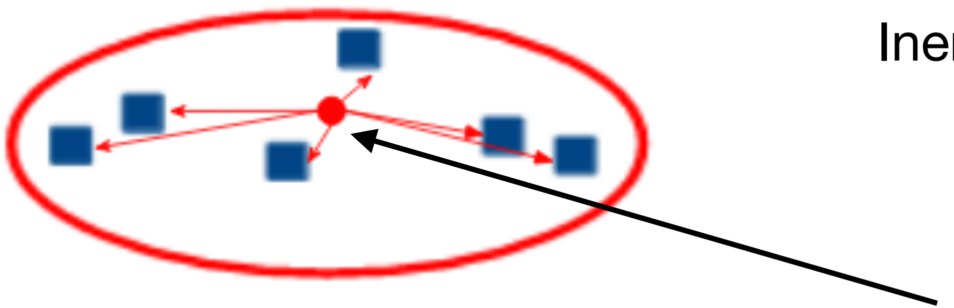
Case - II

Evaluation Metrics for Clustering

- Inertia: Sum of intracluster distances

The lesser the inertia value, the better the cluster is

Inertia: Sum of intracluster distances $\sqrt{\sum_{i=1}^{m} |\overrightarrow{x}_i - \overrightarrow{c}_i|^2}$



Centroid

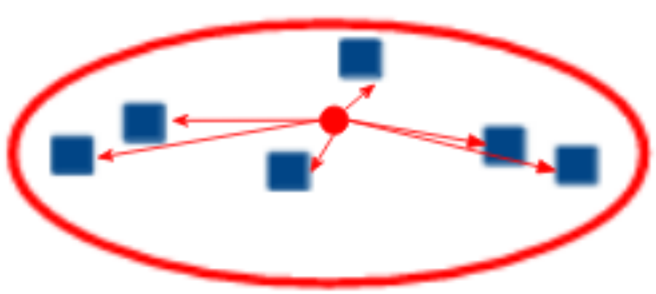
Intra cluster distance

- Dunn Index:

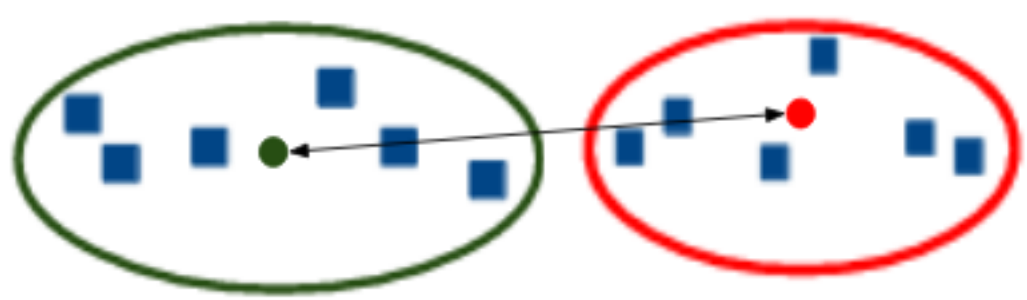
$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

Clusters are far apart

Clusters are compact



Intra cluster distance



Inter cluster distance

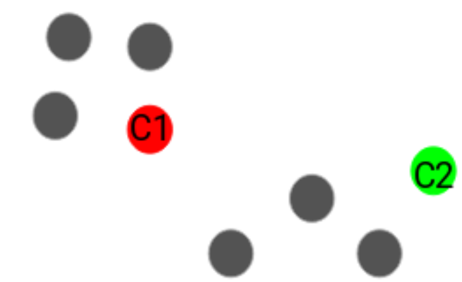
K-Means Clustering

Centroid-based or distance-based algorithm, minimise the sum of distances

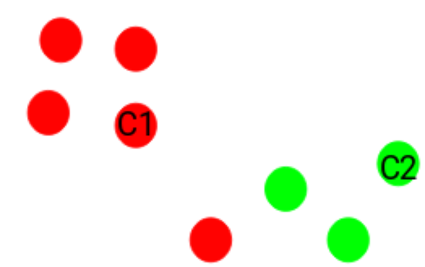
- Step 1: Choose the number of clusters k

take k=2

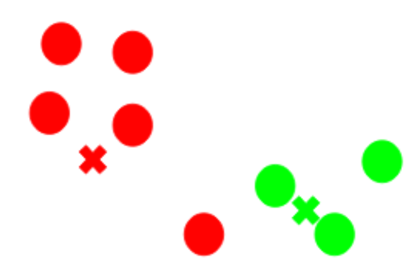
- Step 2: Select k random points from the data as centroids



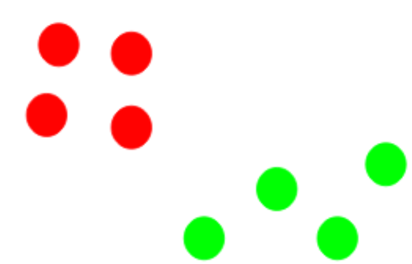
- Step 3: Assign all the points to the closest cluster centroid



- Step 4: Recompute the centroids of newly formed clusters



- Step 5: Repeat steps 3 and 4

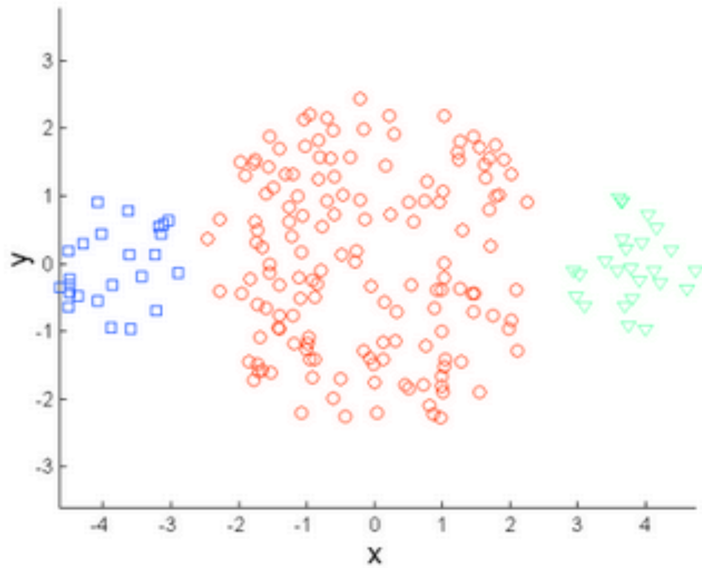


Stopping Criteria

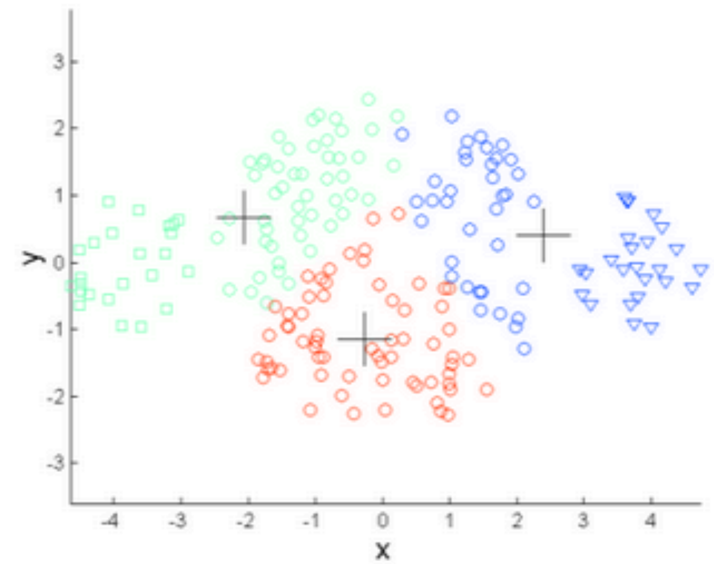
1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations are reached

Challenges with the K-Means Clustering

The size of clusters is different

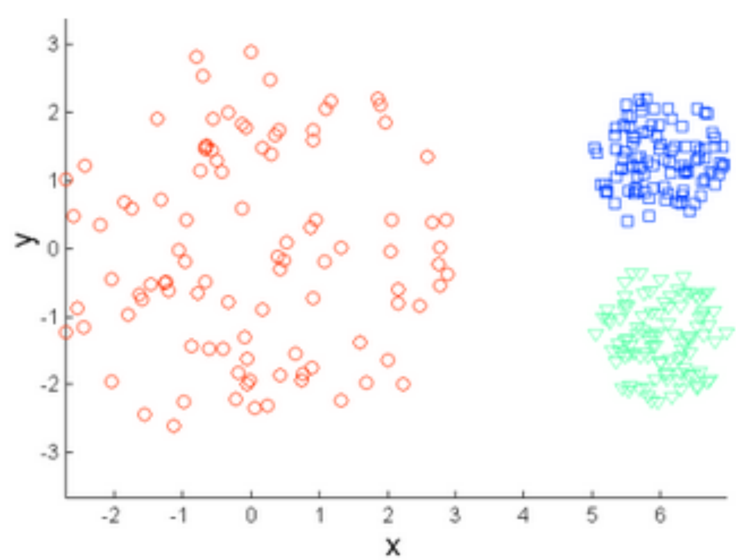


Original Points

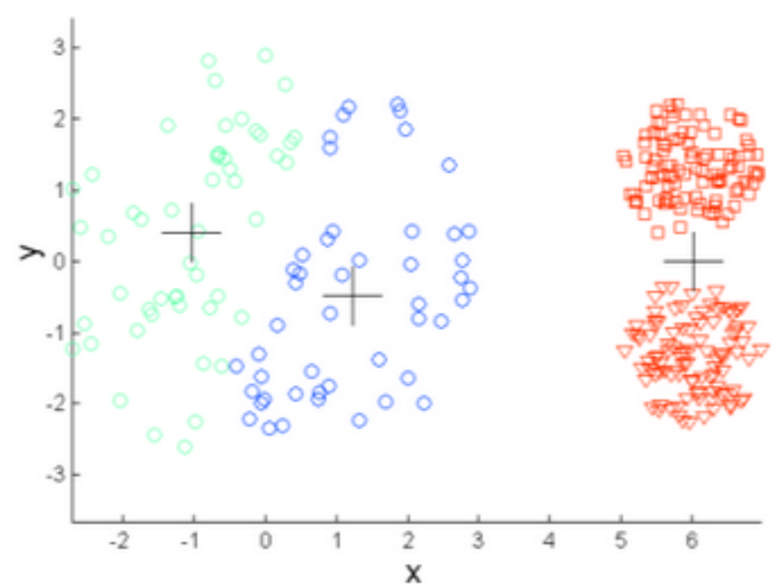


K-means (k = 3)

The densities of the original points are different



Original Points



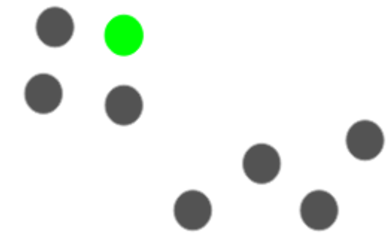
K-means (k = 3)

K-Means++ Clustering

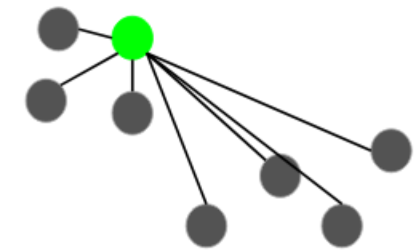
Specifies a procedure to initialise the cluster centres before moving forward with k-means, take k=3

- Step 1: randomly pick **a** data point as **a** cluster centroid

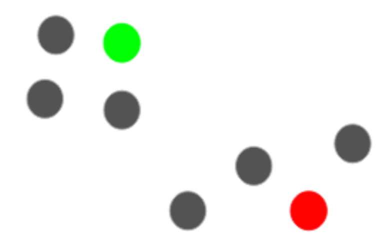
  (not all the centroids but one)



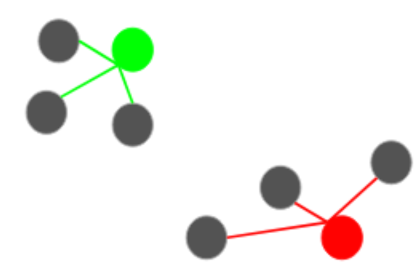
- Step 2: calculate the distance of each data point with this centroid



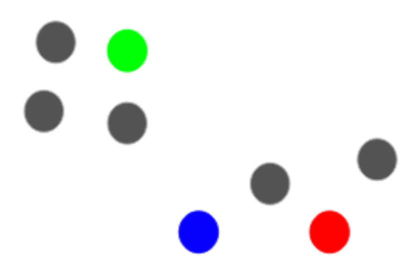
- Step 3: the next centroid is the one whose distance is the farthest from the current centroid



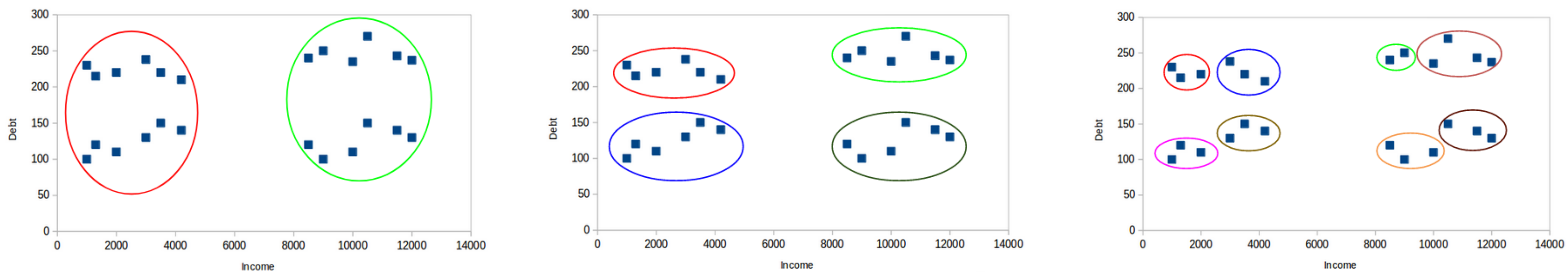
- Step 4: take the distance of each point from its closest centroid and the point having the largest distance will be selected as the next centroid



- Step 5: continue with the K-means after initialising the centroids



How to choose the right number of clusters



Elbow curve, x-axis represent the number of clusters and y-axis the evaluation metric