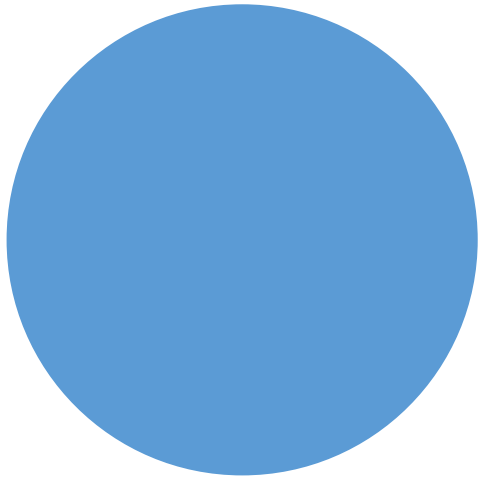
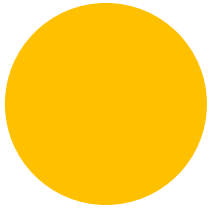
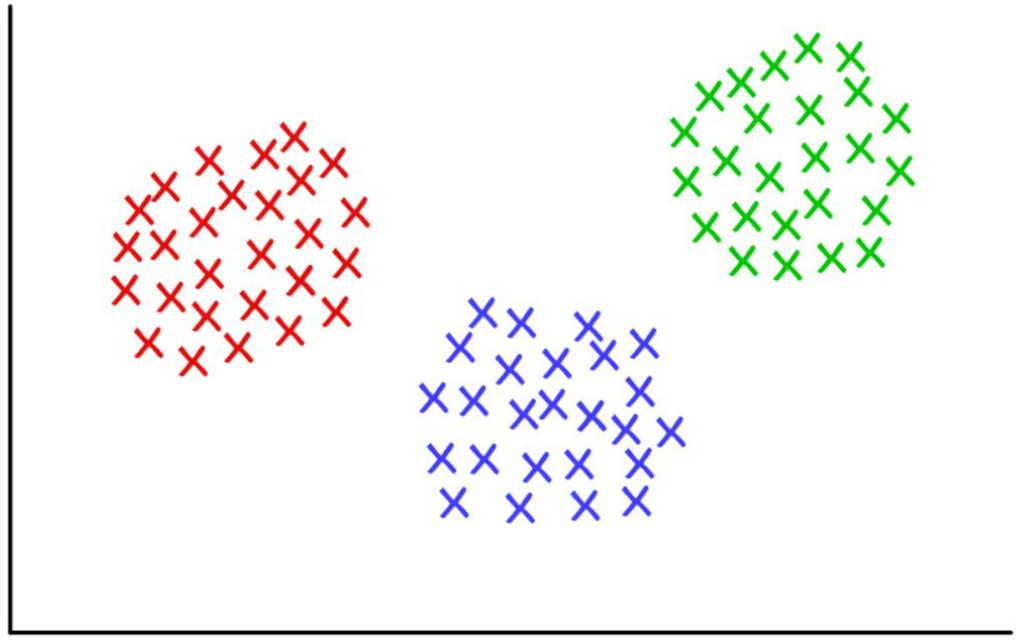



Clustering

What is it?

- *Clustering* is a method of dividing data into groups = *clusters*; with respect to their similarity.
- Number of clusters defines the *cardinality* of clustering.
- Objects within the same cluster should be similar
- Objects across the clusters should be different



Clustering vs. Classification

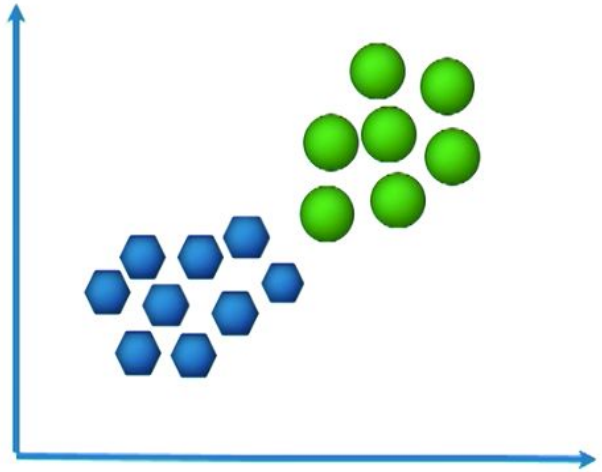
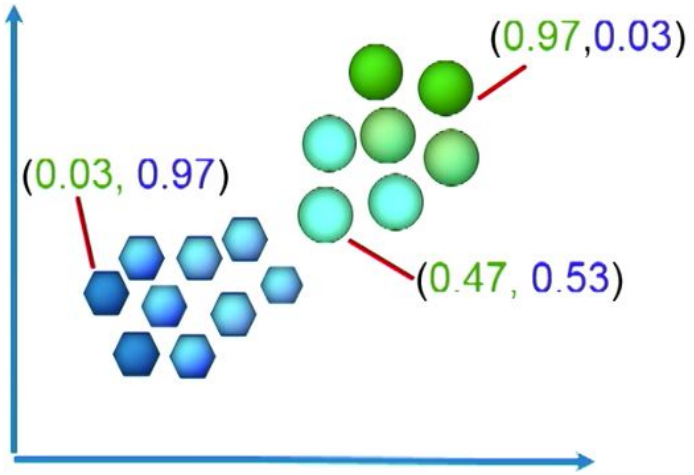
Clustering	Classification
Organises the data according to the similarity among the objects	Classifies data into one of the predefined classes
Unsupervised learning	Supervised learning
No training set	Training set

Where to use it?

- Determining the internal structure of the data
  - e.g. genes structure
- Partitioning
  - e.g. market segmentation
- Data preprocessing
  - e.g. determining patterns, extracting topics
- Profiling
  - e.g. determine correlations among individuals or groups
- . . .

Hard vs. Soft (Fuzzy) Clustering?

Hard Clustering	Soft Clustering
Each object is assigned to one and only one cluster	Each object may be assigned to more clusters
Binary approach	Probabilistic approach

Hard Clustering	Soft Clustering
	

What is used? . . . K-means

- Determine a fixed number of expected clusters
- Initialise the position of cluster centres = *centroids*
- Iteratively minimise the RSS (Residual Sum of Squares)
  - minimise distances of objects from their respective centroids
- Terminate when the condition is met
  - usually a combination of maximum number of iterations and RSS threshold
- Demo:

  http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

  https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

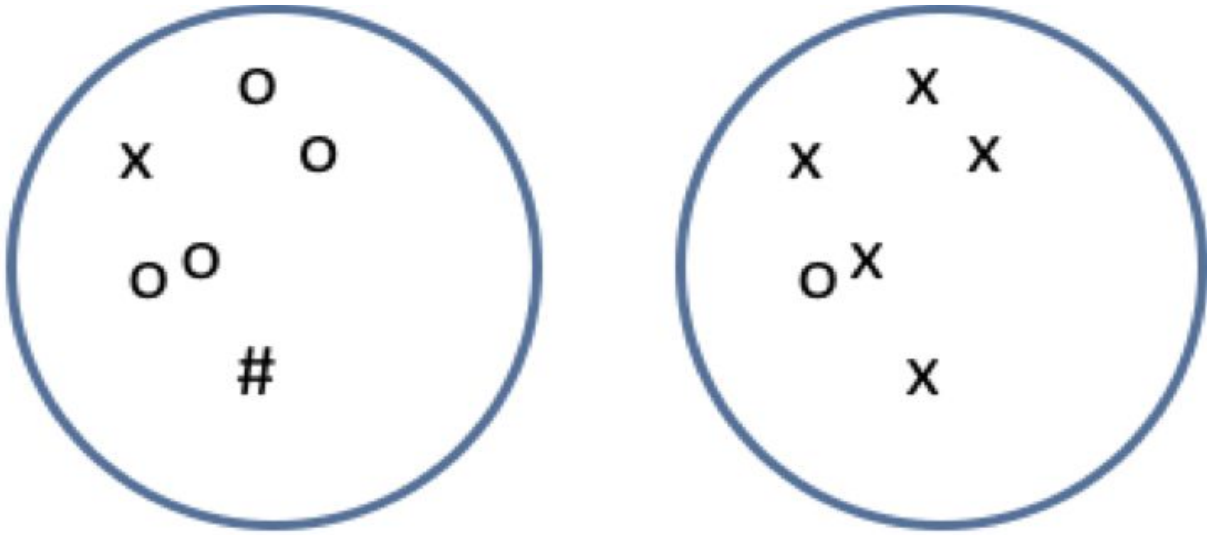
How to evaluate it?

- Optimise the objective function *= internal criterion*
  - achieve high intra-cluster similarity
  - achieve low inter-cluster similarity
- Evaluate the clustering with respect to the application *= external criterion*
  - *confusion matrix* (gold standard/benchmark)
  - *rand measure*
  - *purity*
  - *F-measure*
  - . . .

. . . purity

- Clusters are assigned classes based on the most dominant object
- Purity represents the ratio of correctly assigned objects to the total number of objects

- *If each element represents one cluster the purity is 1*



There is more!

- Different cluster models
  - centorid, connectivity, density, subspace . . .
- Level of clustering distinction
  - hierarchical, subspace, overlapping . . .

- Where to look
  - https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/
  - https://github.com/benjaminwilson/python-clustering-exercises

Q&A?