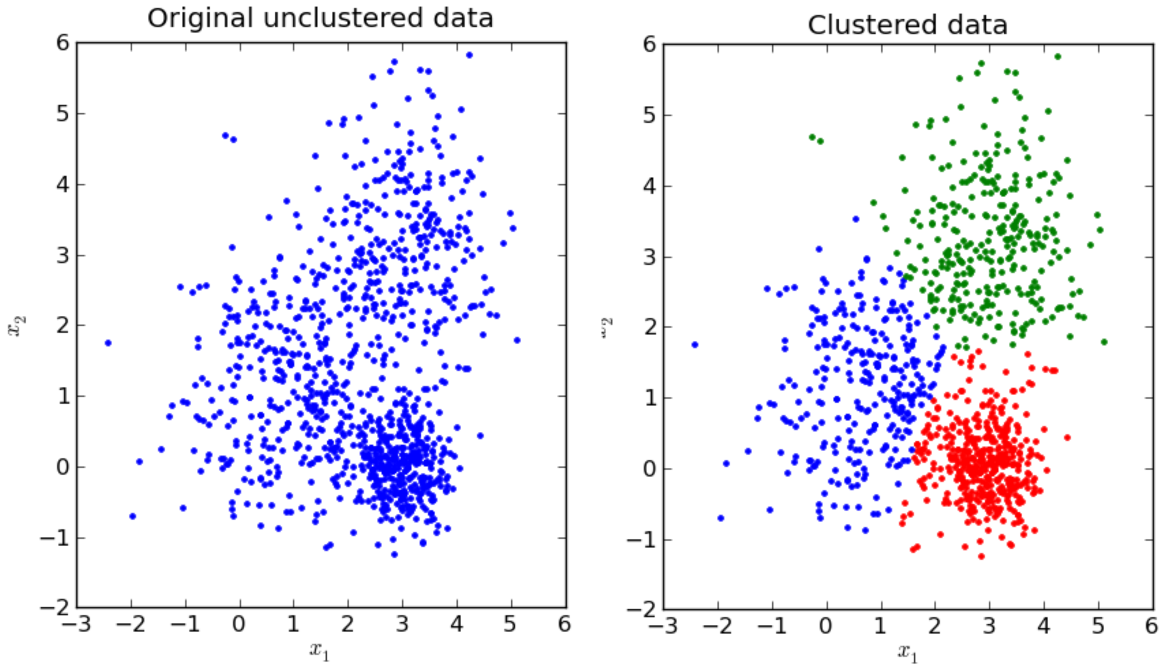
Module 12
Clustering (K-Mean & Hierarchical)

Prof. Pedram Jahangiry



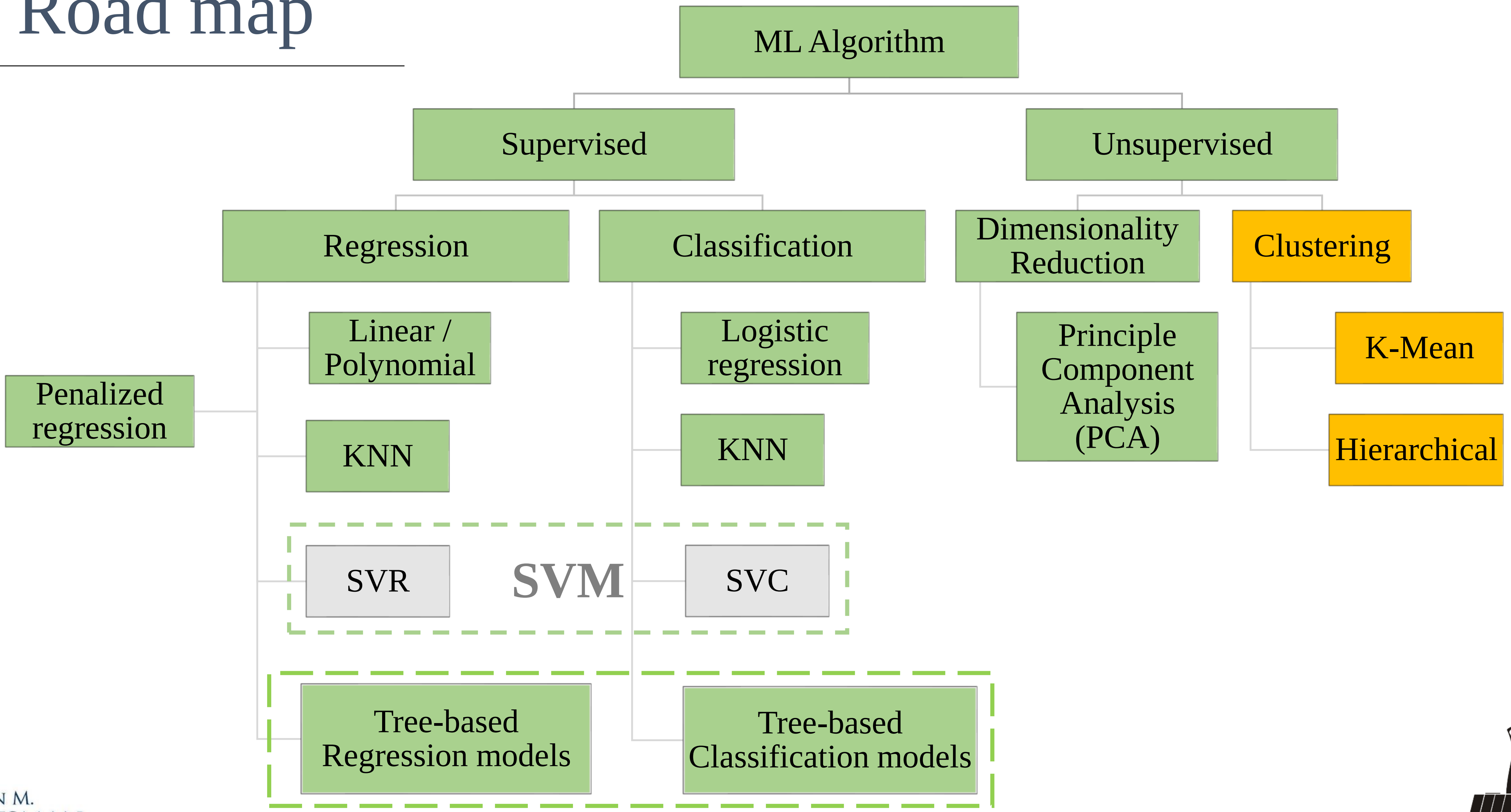
Jon M. Huntsman School of Business
Utah State University

Pedram Jahangiry

Class Modules

- Module 1- Introduction to Machine Learning
- Module 2- Setting up Machine Learning Environment
- Module 3- Linear Regression (Econometrics approach)
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)

Road map


ML Algorithm
Supervised
Unsupervised
Regression
Classification
Dimensionality Reduction
Clustering
Linear / Polynomial
Penalized regression
KNN
Logistic regression
KNN
Principle Component Analysis (PCA)
K-Mean
Hierarchical
SVR
SVM
SVC
Tree-based Regression models
Tree-based Classification models

Topics

Part I

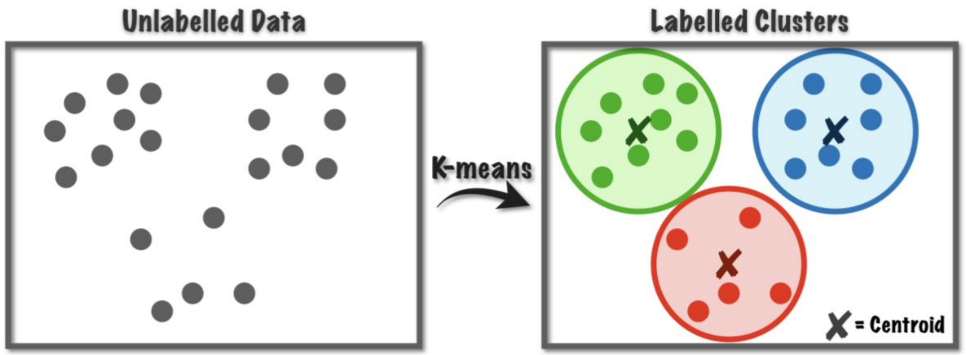
- What is clustering?
- Similarity/Dissimilarity metrics
- Applications in finance

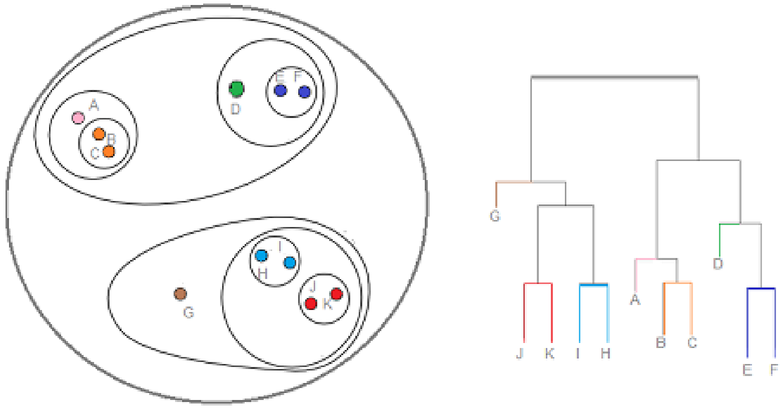
Part II

- K-Means Clustering
- K-modes and K-Prototyping

Part III

- Hierarchical Clustering





Part I

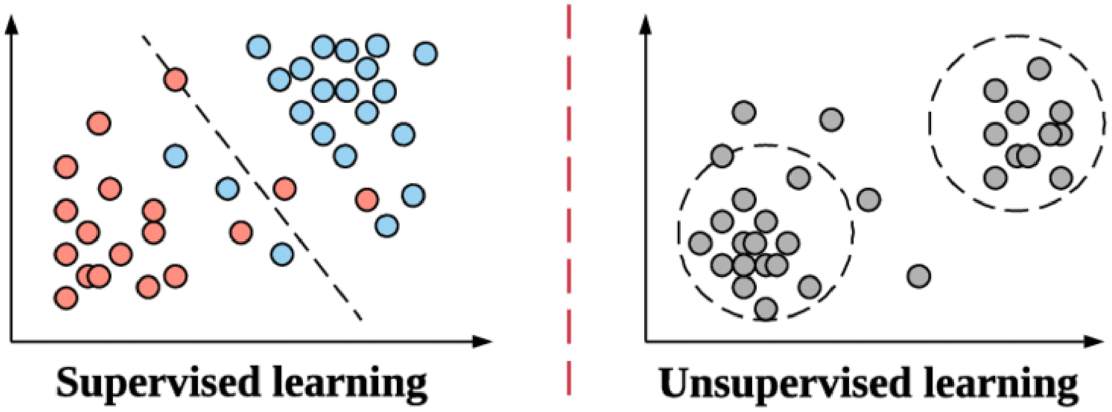
- What is clustering?
- Similarity/Dissimilarity metrics
- Applications in finance

JON M. HUNTSMAN SCHOOL OF BUSINESS UtahStateUniversity

Pedram Jahangiry

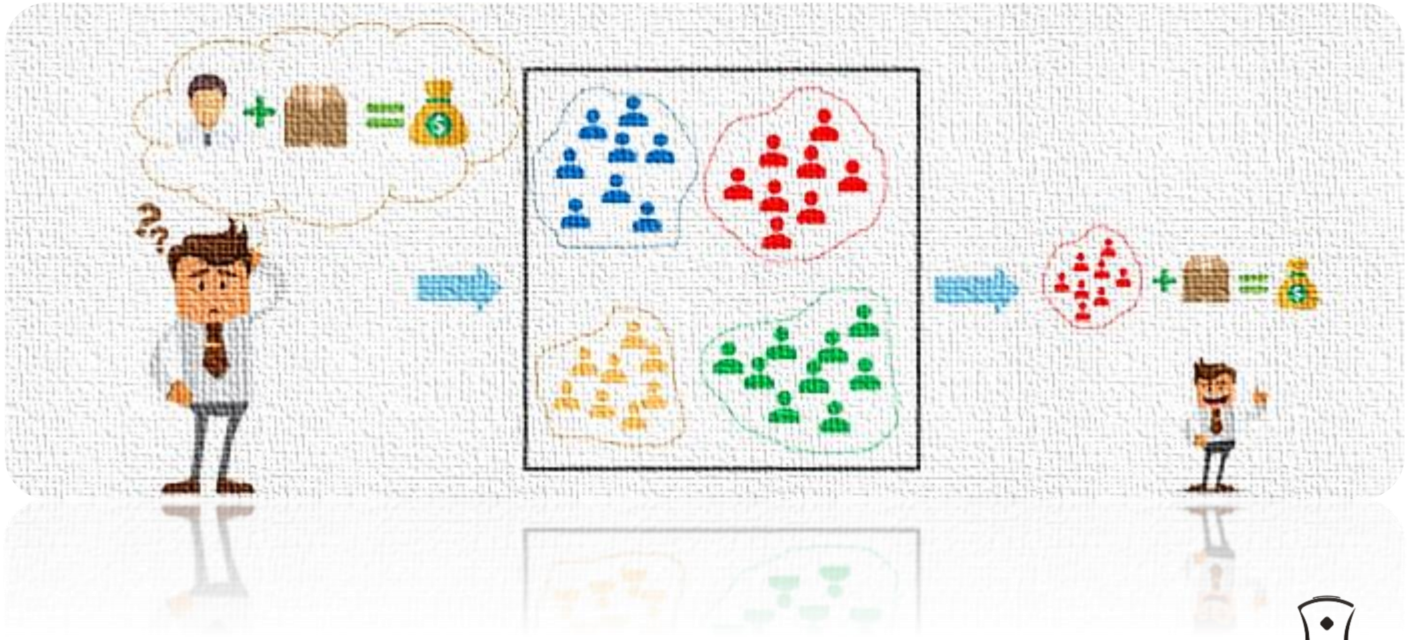
Unsupervised Learning

- Unsupervised learning is a type of machine learning where the algorithm is not given any labeled training data.
- The goal is to discover the underlying patterns and find groups of samples that behave similarly. Find something interesting!
- The two main types of unsupervised learning algorithms are:
  1) Dimension reduction algorithm
     - Principal Component Analysis
  2) Clustering: group similar data
     - K-Mean
     - Hierarchical



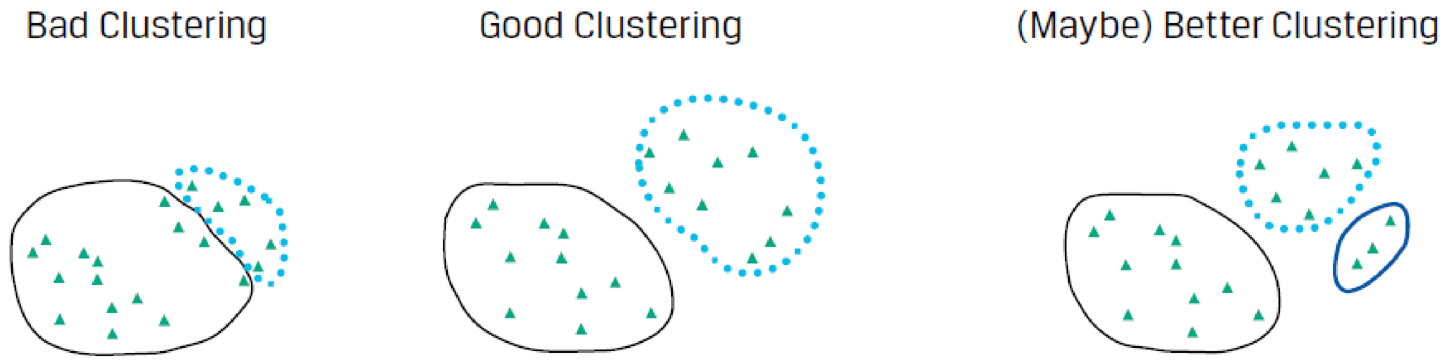
Motivation

1) Dimension reduction algorithm: Principal Component Analysis
2) Clustering techniques: K-Mean for example



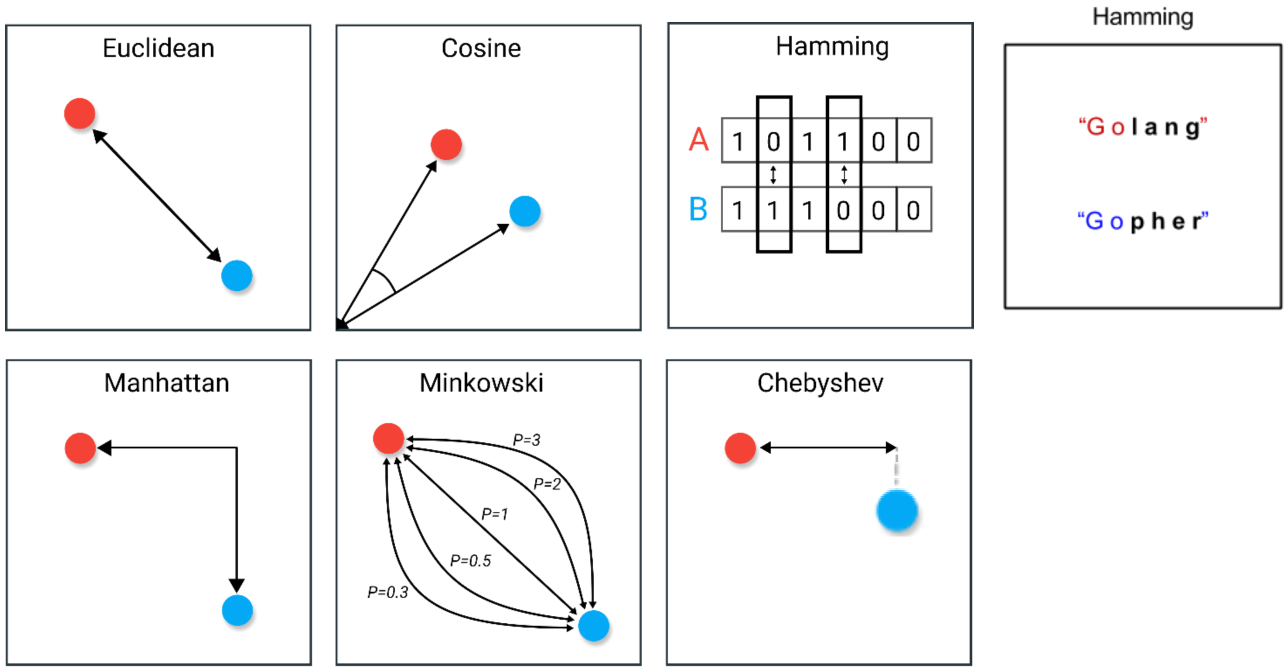
What is Clustering?

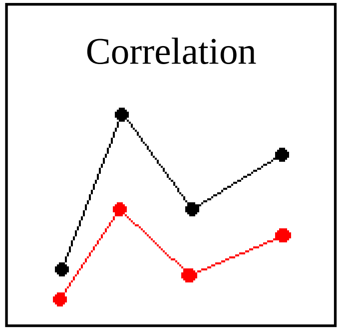
- Clustering is an unsupervised machine learning which is used to organize data points into similar groups called clusters.
- A cluster contains a subset of observations from the dataset such that all the observations within the same cluster are "similar."
- The goal is to maximize the intra-clusters (within) **similarities** or equivalently to maximize the inter-clusters (between) **dissimilarities**.



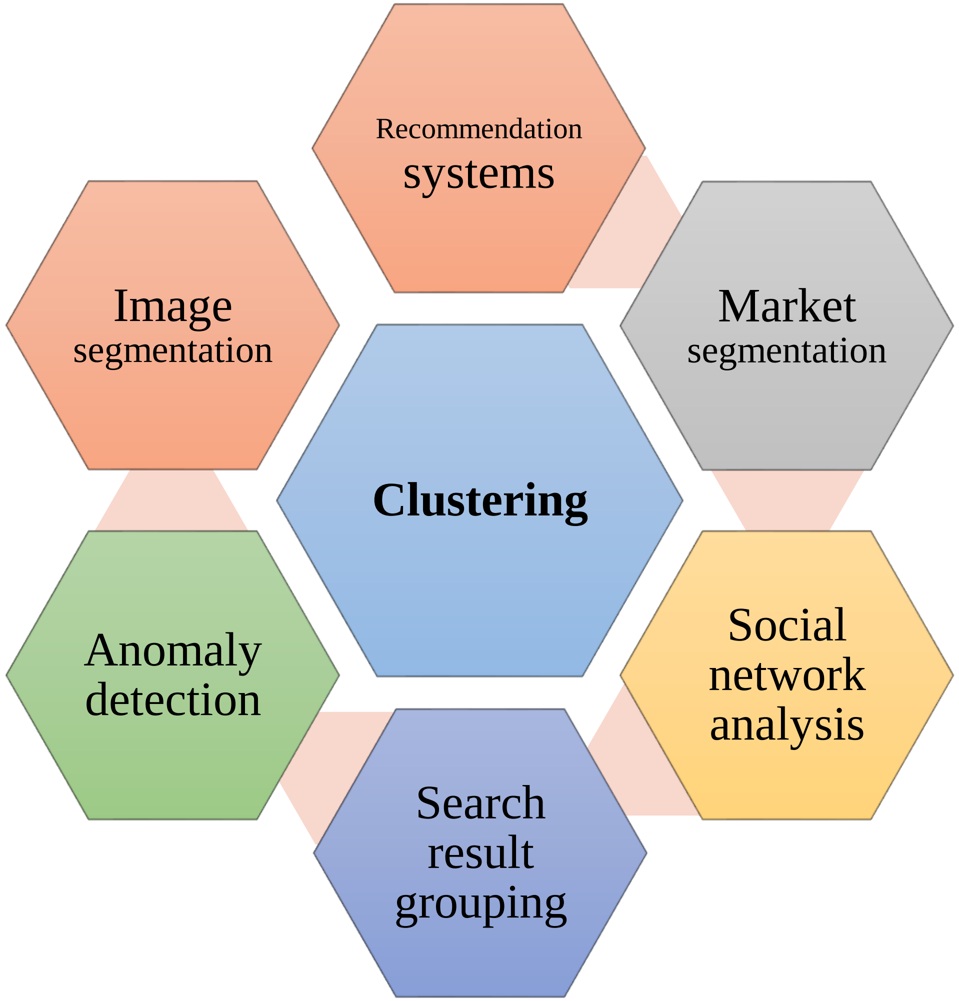
Similarity/Dissimilarity metrics

- Similarity/Dissimilarity between observations can be thought of as the distance between them.
- The <u>smaller the distance, the more similar the observations</u>; the larger the distance, the more dissimilar the observations.



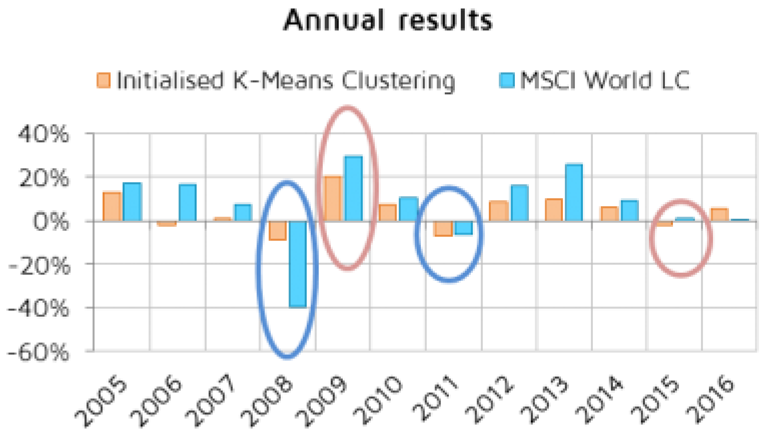


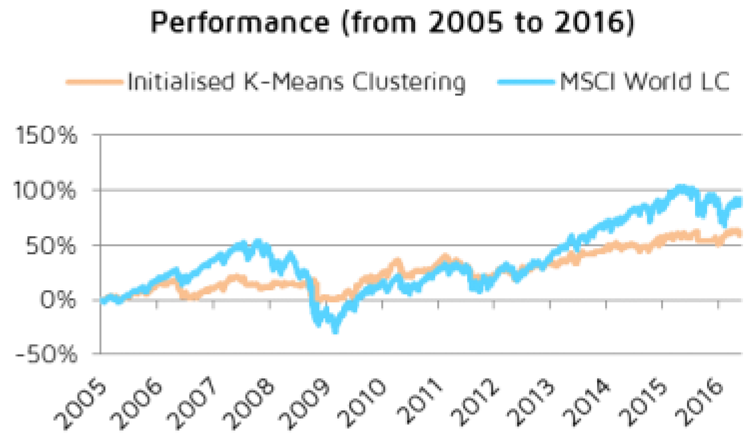
Applications of clustering



Applications in Finance

- Applied to grouping companies, for example, clustering may uncover important similarities and differences among companies that are not captured by standard classifications of companies by industry and sector.
- In portfolio management, clustering methods have been used for improving portfolio diversification by investing in assets from multiple different clusters.





Part II

- K-Mean clustering
- K-Mode and K-prototyping

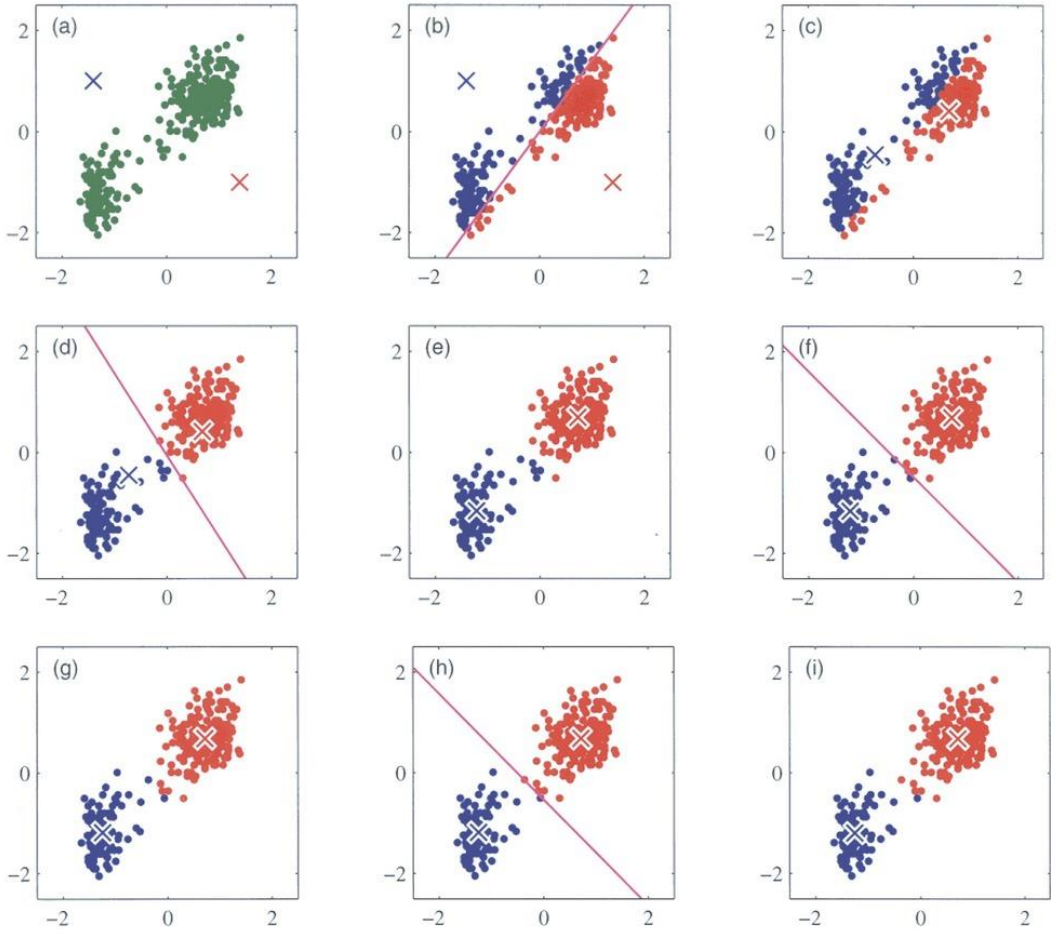
K-Means Clustering

- K-means is an algorithm that repeatedly partitions observations into a
  - fixed
  - pre-specified and
  - non-overlapping

  number of clusters, k (a hyperparameter)
- Each cluster is characterized by its centroid (arithmetic mean position).
- K-means minimizes intra-cluster (within-cluster) distance

Algorithm 1 k-means algorithm

Specify the number k of clusters to assign.
Randomly initialize k centroids.
repeat
  expectation: Assign each point to its closest centroid.
  maximization: Compute the new centroid (mean) of each cluster.
until The centroid positions do not change.



Jon M. Huntsman School of Business
Utah State University

K-Means clustering (details)

- Objective function: Minimizing the within-cluster variation (WCV)

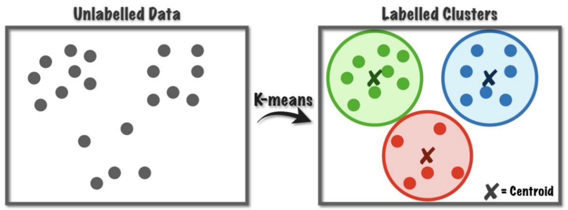
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\}$$

- This optimization says that we want to partition the observations into K clusters such that the **total** within-cluster variation is as small as possible.
- If we use Euclidian distance, then:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Initial positioning of the centroids matter!

K-Means clustering pros and cons



Pros

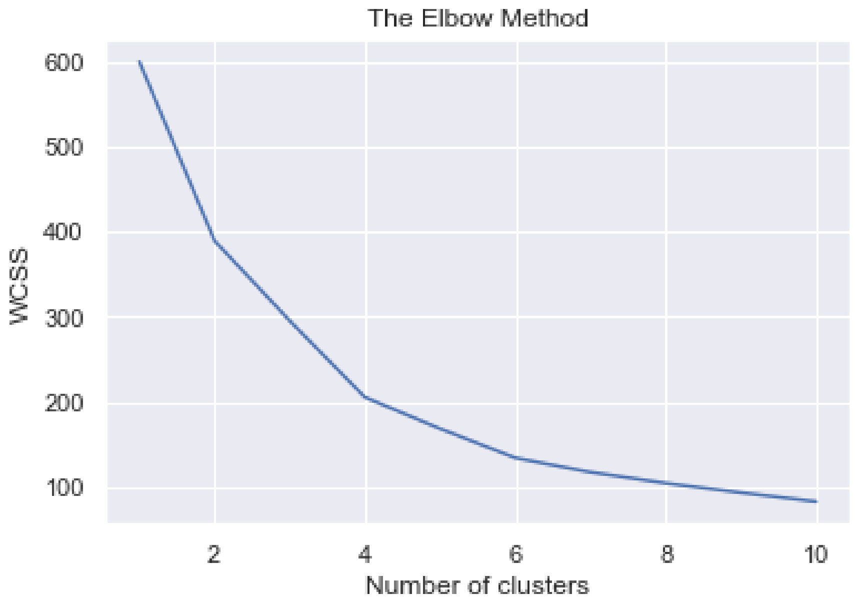
- Simple!
- The k-means algorithm is fast and works well on very large datasets.
- Can help visualize the data and facilitate detecting trends or outliers.
- The k-means algorithm is among the most used algorithms in investment practice.

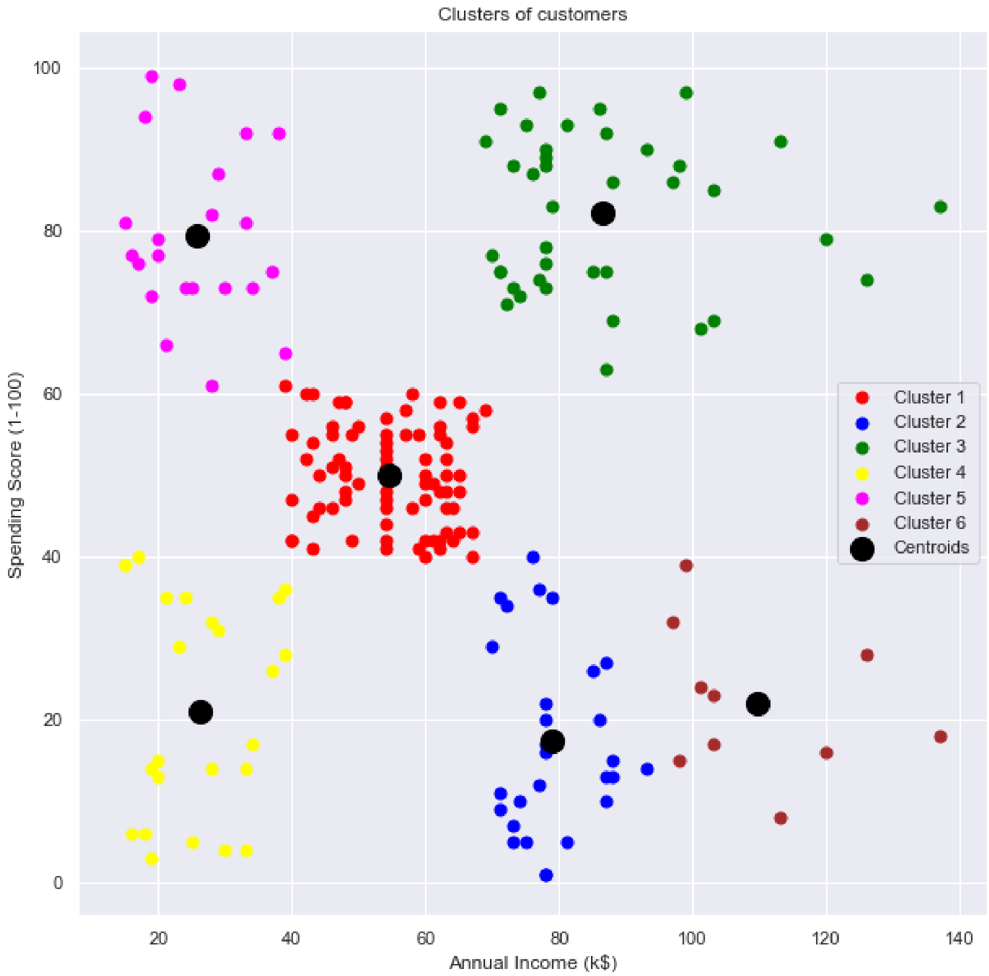
Cons

- The hyperparameter, k, must be decided before the algorithm can be run.
- The final assignment of observations to clusters can depend on the initial location of the centroids. Local optimum vs global optimum. (should be rerun with several initialization)

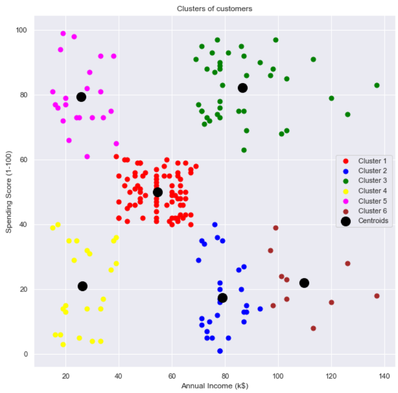
Optimal number of K (the elbow method)

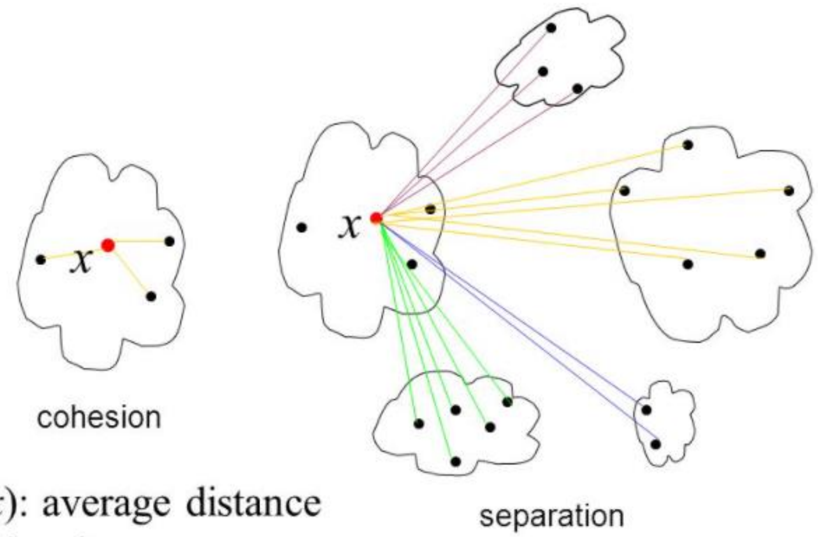
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\}$$





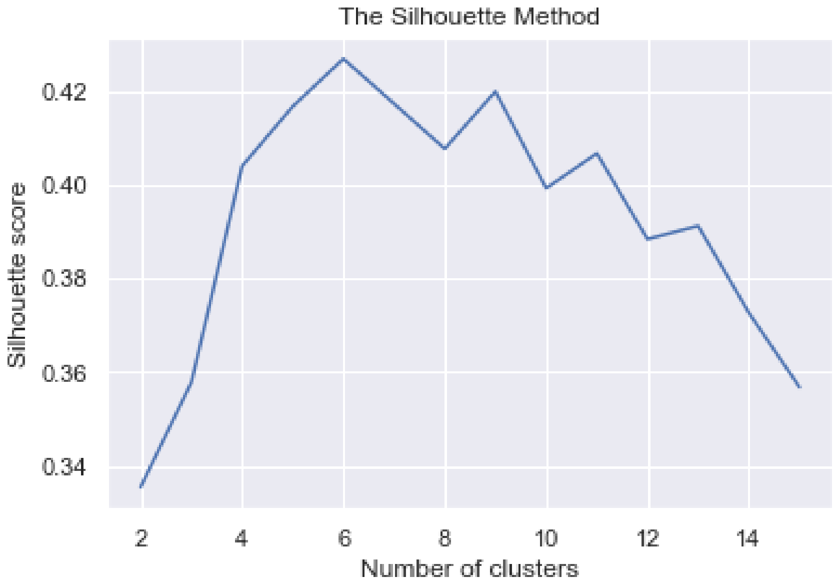
Optimal number of K (the Silhouette method)





$a(x)$: average distance in the cluster

$b(x)$: average distances to others clusters, find minimal



$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad -1 \leq s(i) \leq 1$$

K-Modes Clustering

- K-Modes technique extends K-Means for categorical data
- It uses modes instead of means for cluster centers.
- Optimization: Minimizes the dissimilarities within clusters to form homogeneous groups.

	Investment Type	Risk Tolerance	Investment Duration	cluster
0	Stocks	High	Long-term	1
1	Bonds	Low	Short-term	0
2	Mutual Funds	Medium	Medium-term	0
3	ETFs	High	Long-term	1
4	Stocks	Medium	Short-term	1
5	Bonds	Low	Medium-term	0

- K-Modes algorithm:
  - Initialization: The centroids in K-Modes are actual data points from the dataset
  - Assignments: Each data point is assigned to the nearest centroid (Hamming distance)
  - Updates: the centroids are updated to be the Mode of the clusters
  - Iterations: The assignment and update steps are repeated iteratively until the centroids stabilize

K-Prototypes Clustering

- Hybrid Approach: Combines K-Means and K-Modes to handle datasets with both numerical and categorical features.
- Centroids: Calculated using the mean for numerical attributes and modes for categorical attributes.
- Assignment: Classify data points to the closest centroid by a cost function that combines distances for numerical and categorical data.

	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

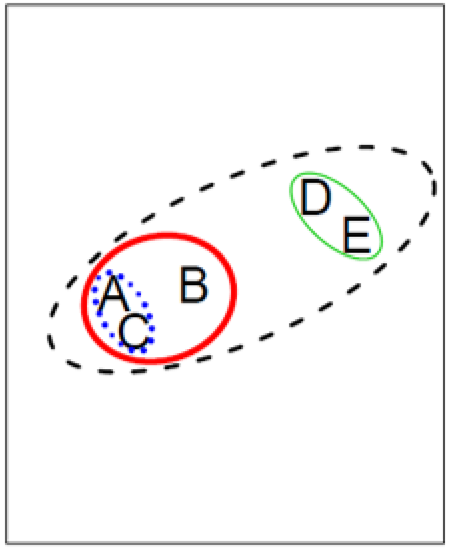
Part III

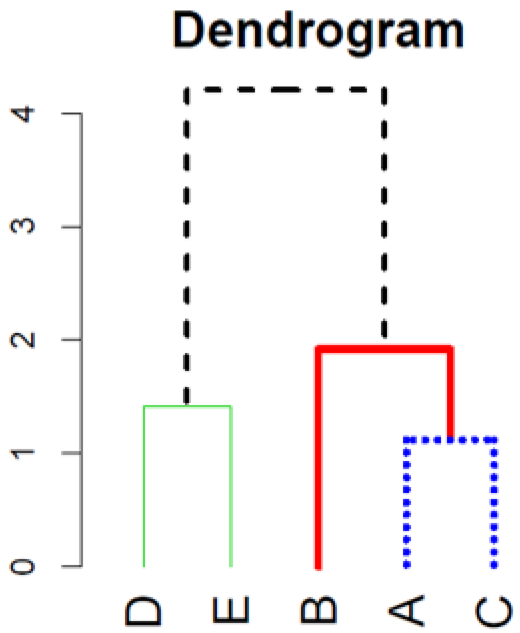
- Hierarchical Clustering
  1. Agglomerative
  2. Divisive

Hierarchical Clustering

- In k-means clustering, the algorithm seeks to partition the data into a pre-specified number of clusters k. All clusters are found simultaneously.
- In hierarchical clustering, the algorithm does not require a pre-specified choice of K. Clusters are found sequentially.

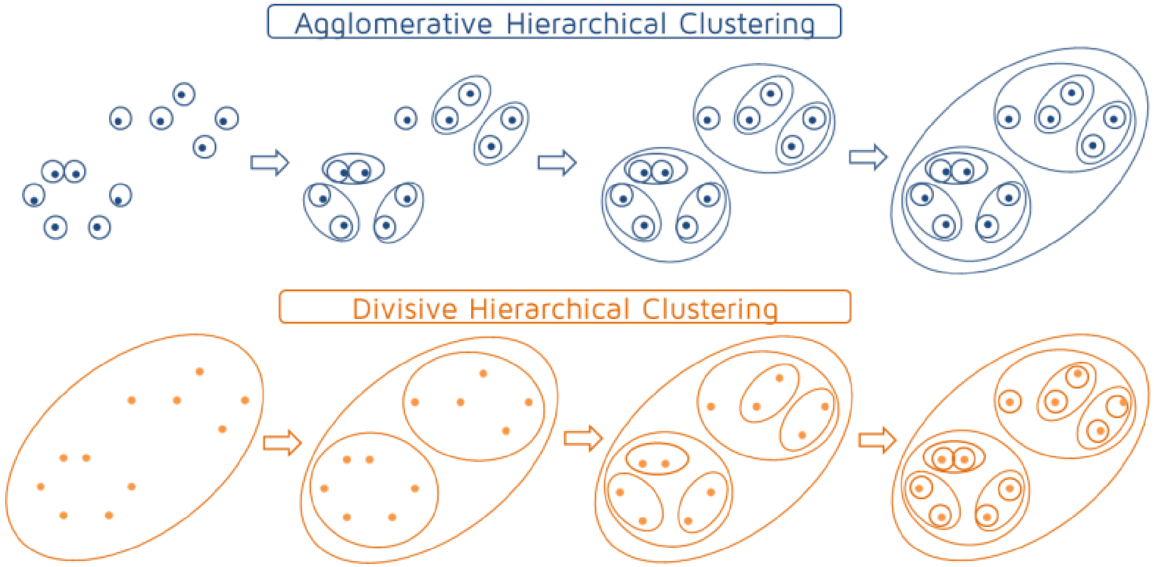
- Hierarchical clustering is an iterative procedure used to build a **hierarchy of clusters**.
- Using a dendrogram (a type of tree diagram which highlights the hierarchical relationships among the clusters), hierarchical clustering has the advantage of allowing the analyst to examine alternative partitioning of data of different granularity **before** deciding which one to use.

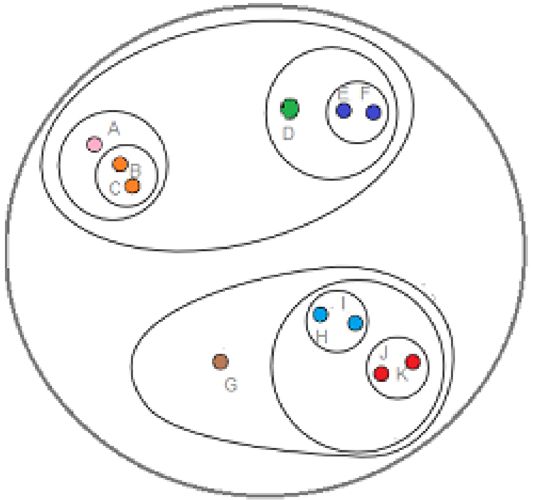


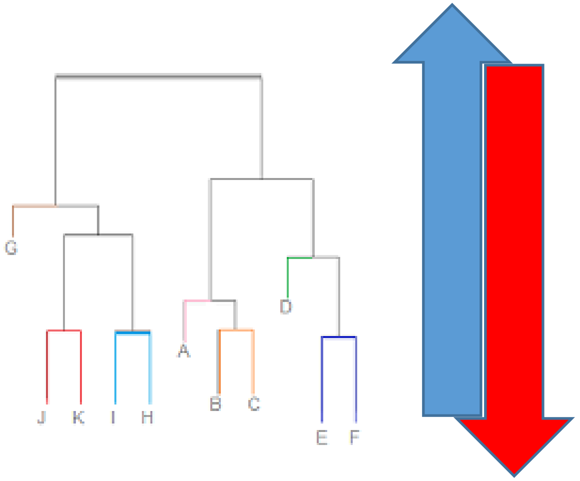


Agglomerative (bottom-up) vs Divisive (top-down) HCA

- Agglomerative: start with each observation being treated as its own cluster
- Divisive: starts with all the observations belonging to a single cluster.

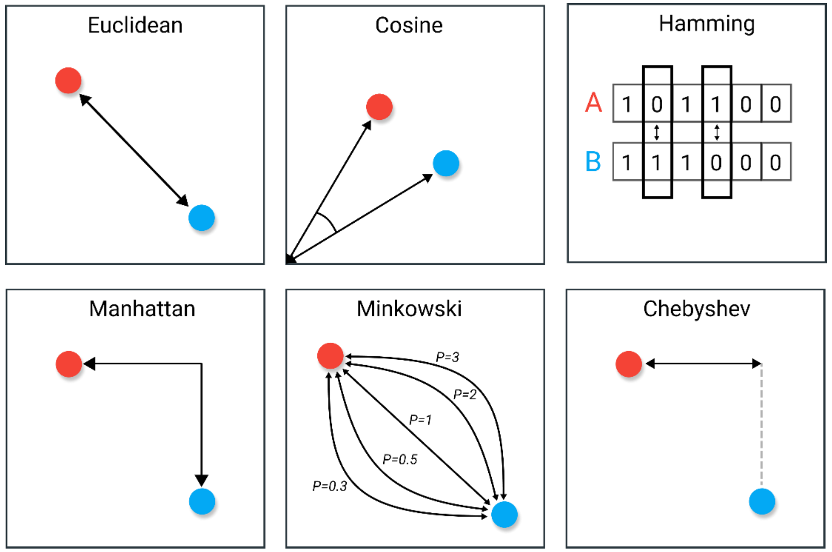


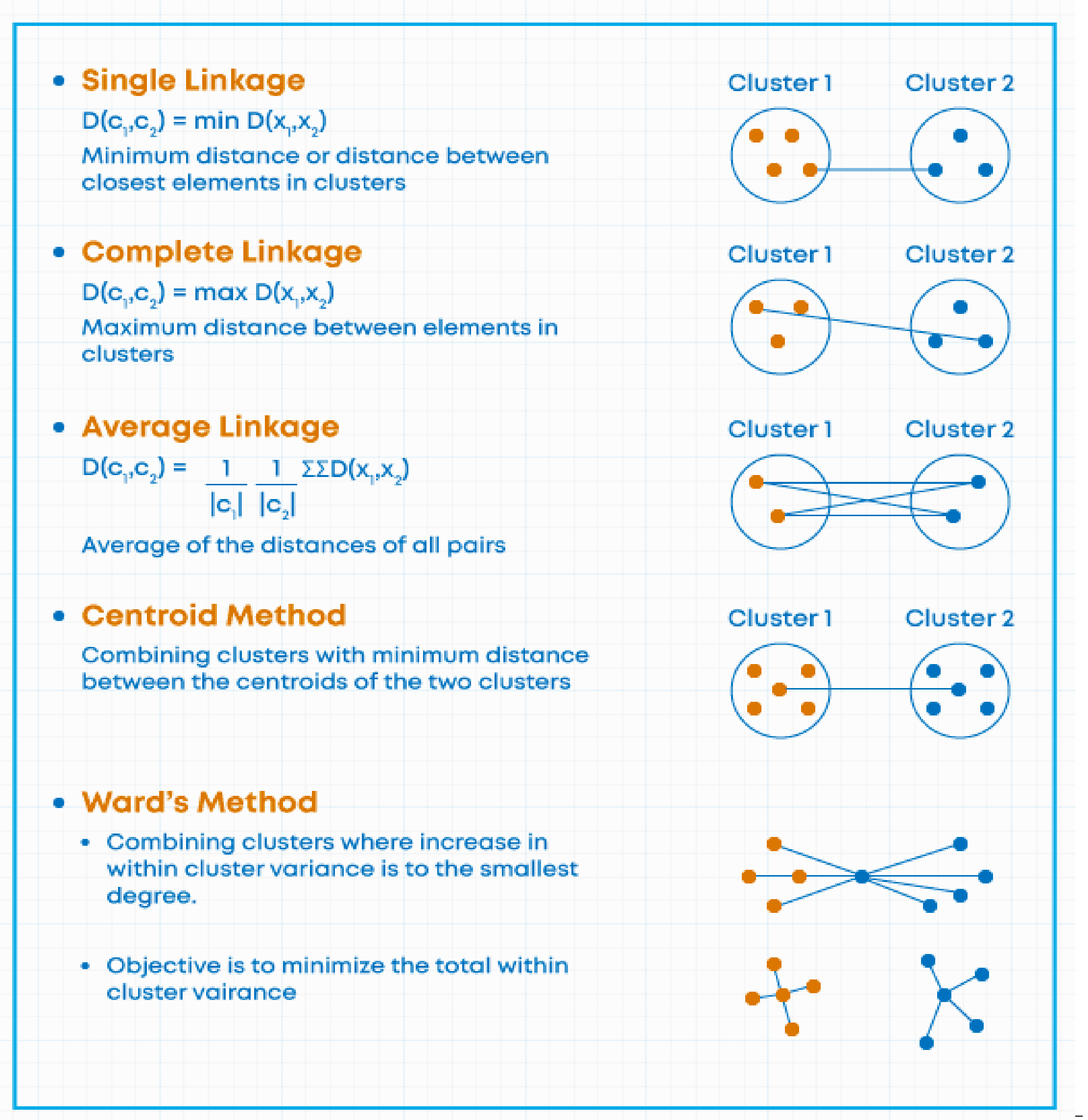




Types of Linkage (distance between two clusters)

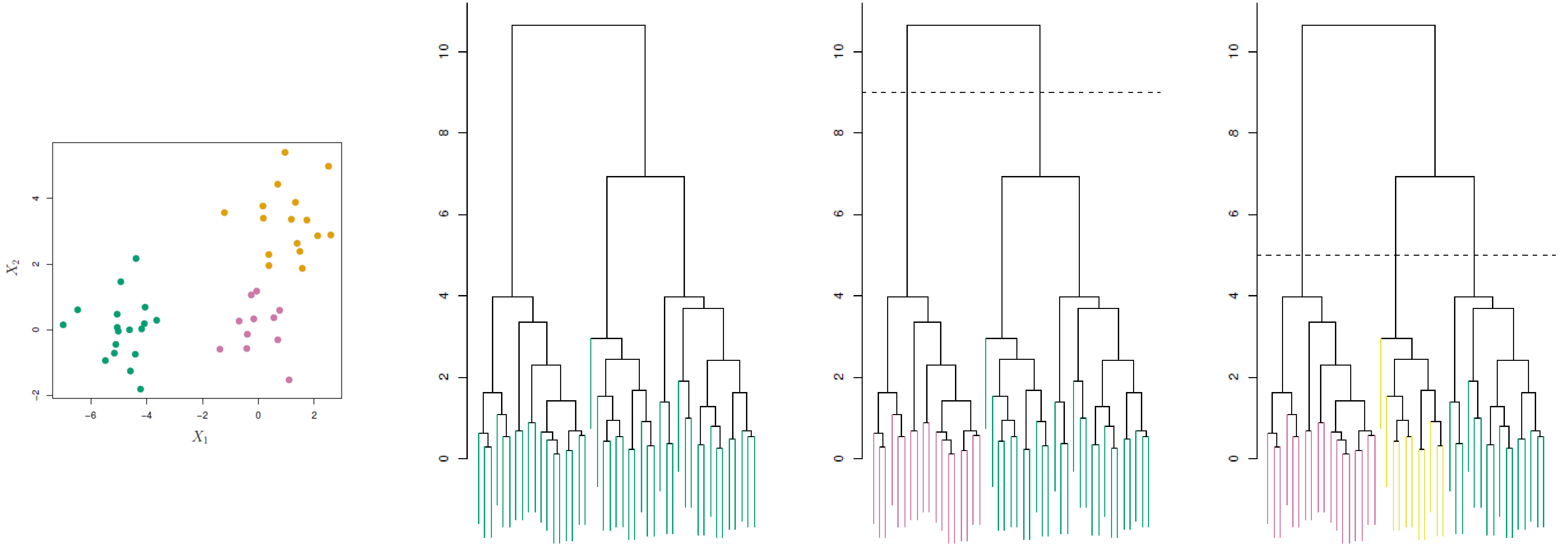
- To decide on the closest clusters, an explicit definition for the distance between two clusters is required (linkage)
- Recall: We have already defined the within-cluster distance metrics.





Source: https://dataaspirant.com/hierarchical-clustering-algorithm/

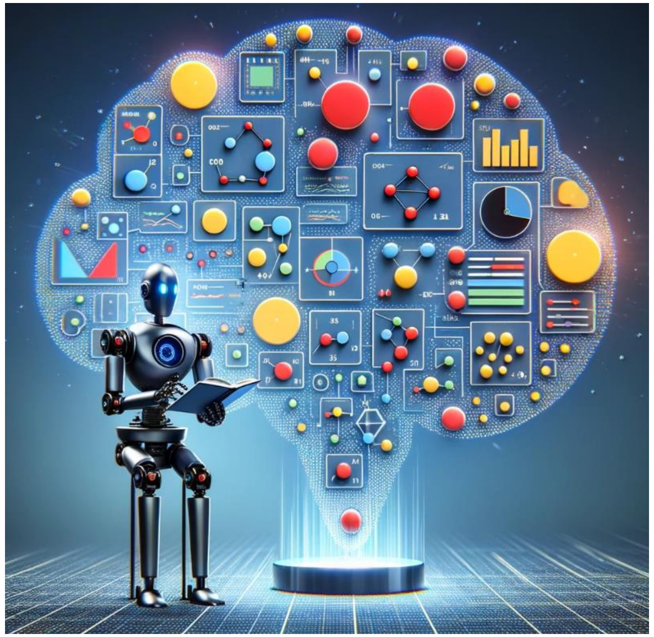
An example



Hierarchical Clustering discussion

- Agglomerative Clustering (Bottom-Up)
  - Ideal for small to medium datasets.
  - Commonly used due to simplicity and available efficient algorithms.

- Divisive Clustering (Top-Down)
  - Suitable for larger datasets.
  - Gives a global perspective of data structure.
  - Effective in early identification of outliers.



- There's no universal rule for the best distance metric or linkage type.
- There's no consensus on the best level to cut the dendrogram.

Class Modules

- ✓ Module 1- Introduction to Deep Learning
- ✓ Module 2- Setting up Machine Learning Environment
- ✓ Module 3- Linear Regression (Econometrics approach)
- ✓ Module 4- Machine Learning Fundamentals
- ✓ Module 5- Linear Regression (Machine Learning approach)
- ✓ Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- ✓ Module 7- Logistic Regression
- ✓ Module 8- K-Nearest Neighbors (KNN)
- ✓ Module 9- Classification and Regression Trees (CART)
- ✓ Module 10- Bagging and Boosting
- ✓ Module 11- Dimensionality Reduction (PCA)
- ✓ Module 12- Clustering (KMeans – Hierarchical)