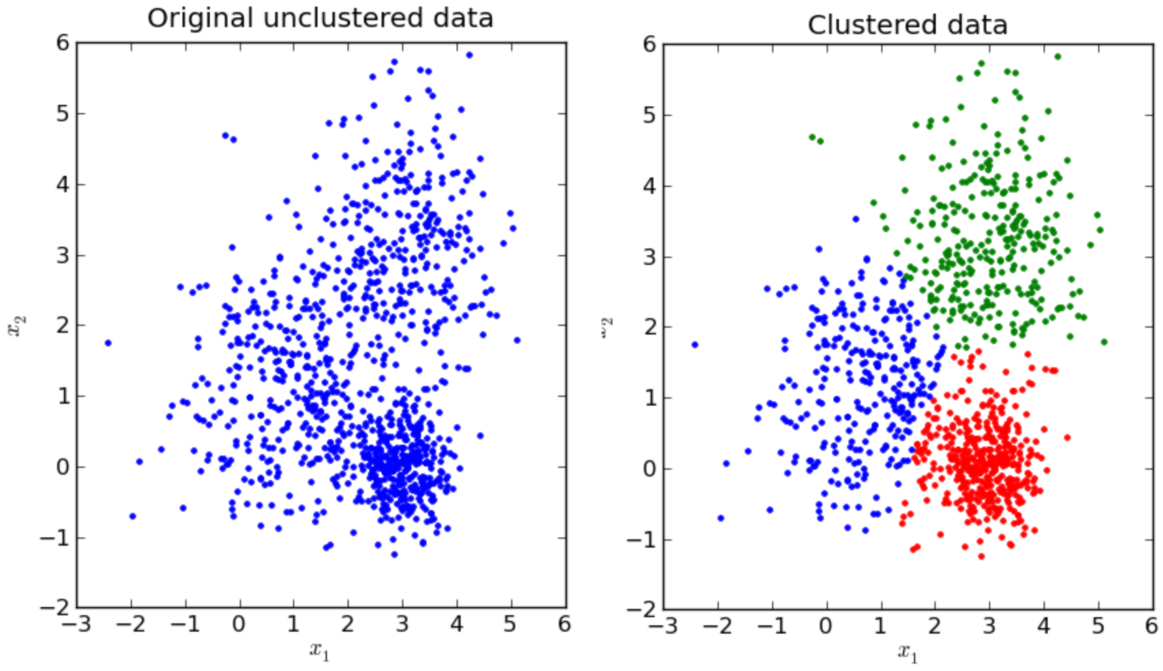
Class -23
Clustering (K-Mean & Hierarchical)

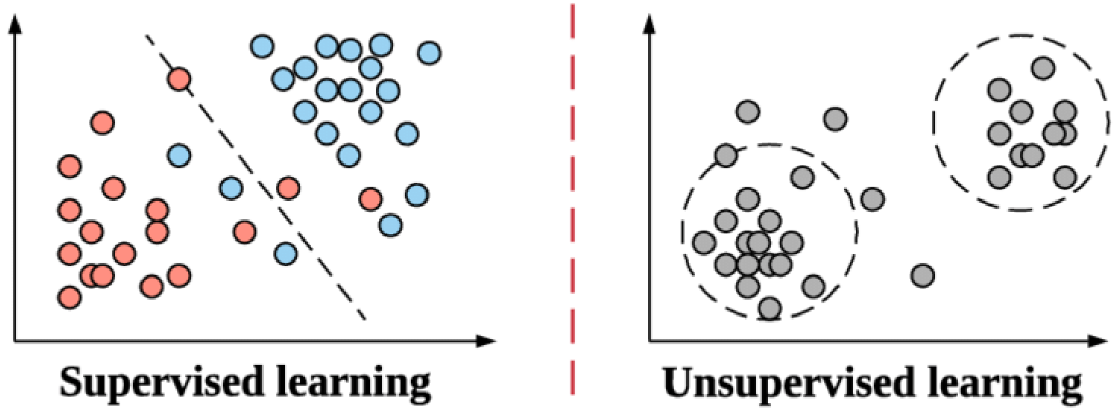
Prof. Pedram Jahangiry



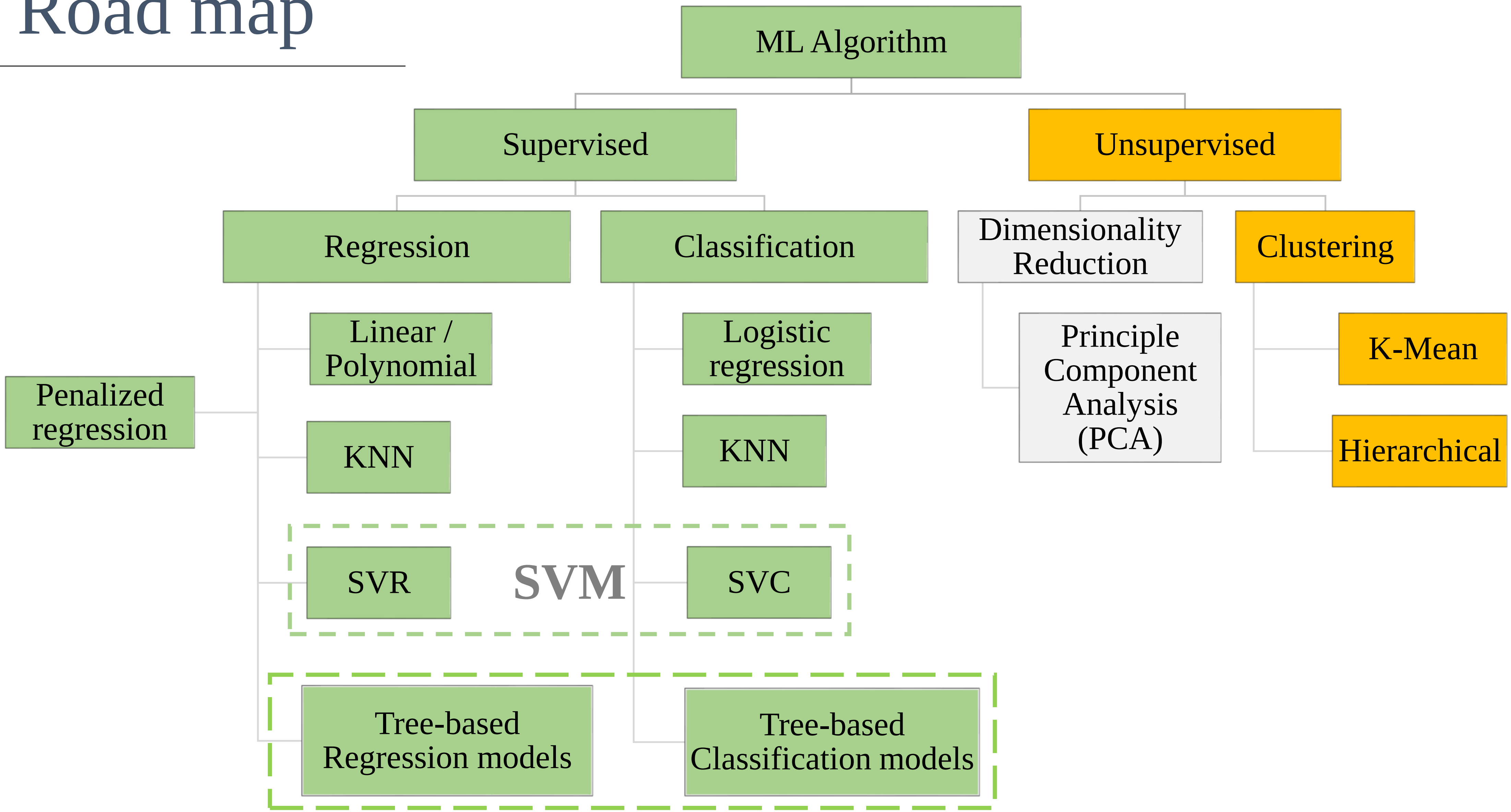
JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Unsupervised Learning

- Unsupervised learning is a machine learning technique that does not use labeled data (no target variable)
- The goal is to discover the underlying patterns and find groups of samples that behave similarly.
- The two main types of unsupervised learning algorithms are:
  1) Dimension reduction algorithm
     - Principal Component Analysis
  2) Clustering techniques
     - K-Mean
     - Hierarchical



Road map


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

Topics

Part I

1. What is clustering?
2. Similarity/Dissimilarity metrics

Part II

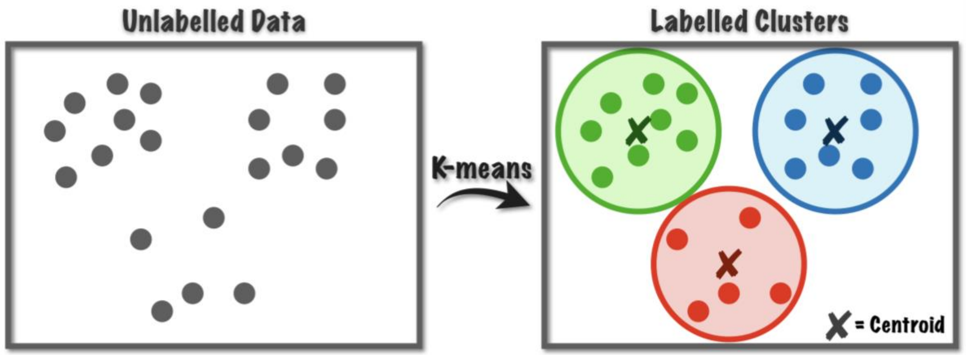
✓ K-Mean clustering

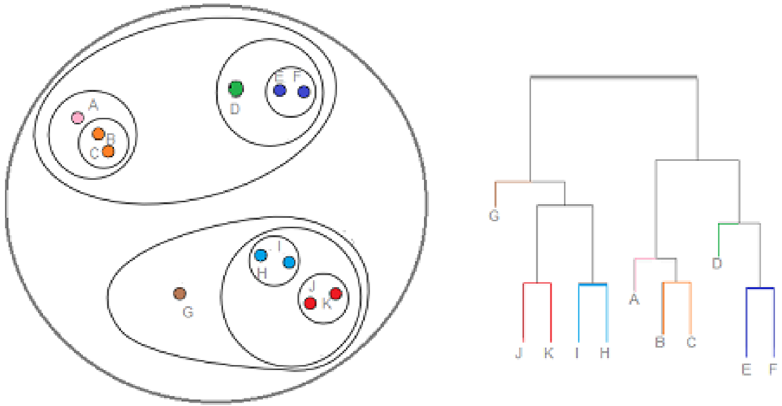
Part III

✓ Hierarchical Clustering

Part IV

✓ Applications in finance



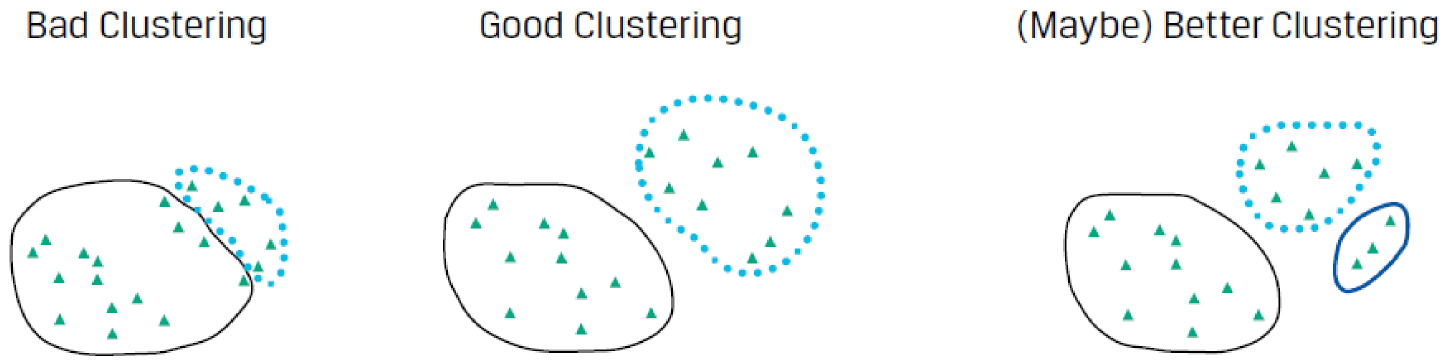


Part I

1. What is Clustering?
2. Similarity/Dissimilarity metrics

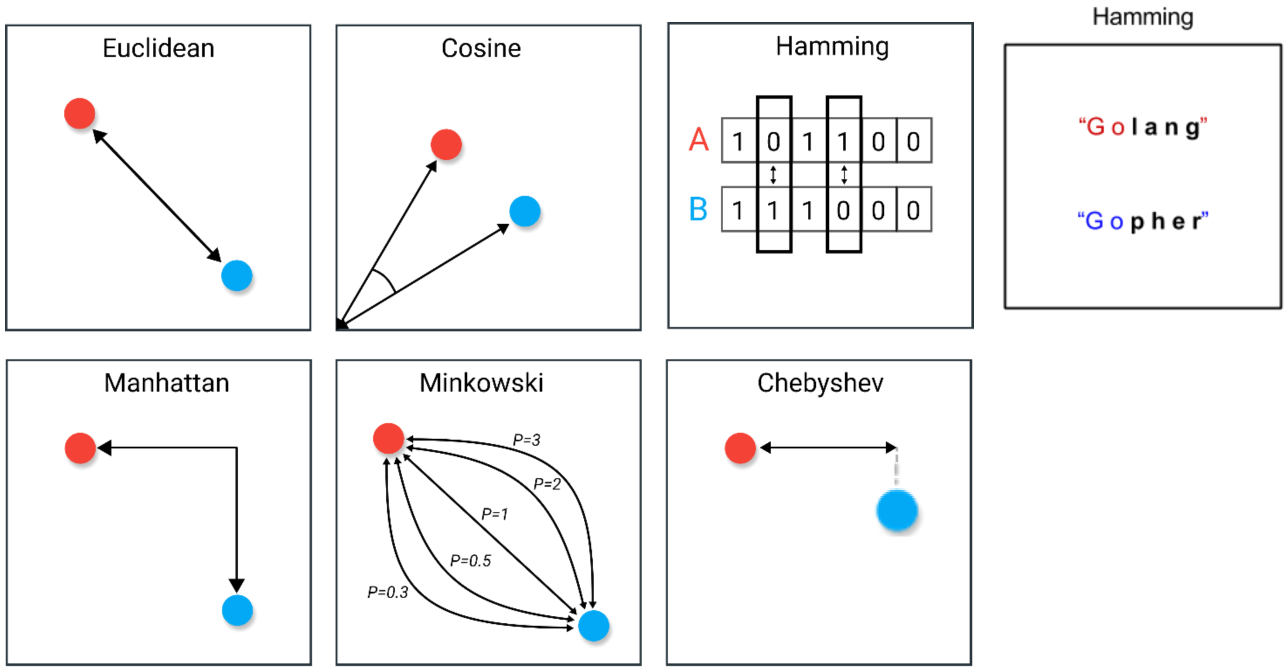
What is Clustering?

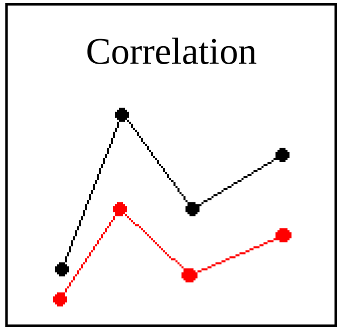
- Clustering is an unsupervised machine learning which is used to organize data points into similar groups called clusters.
- A cluster contains a subset of observations from the dataset such that all the observations within the same cluster are "similar."
- The goal is to maximize the intra-clusters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





Part II

- K-Mean clustering

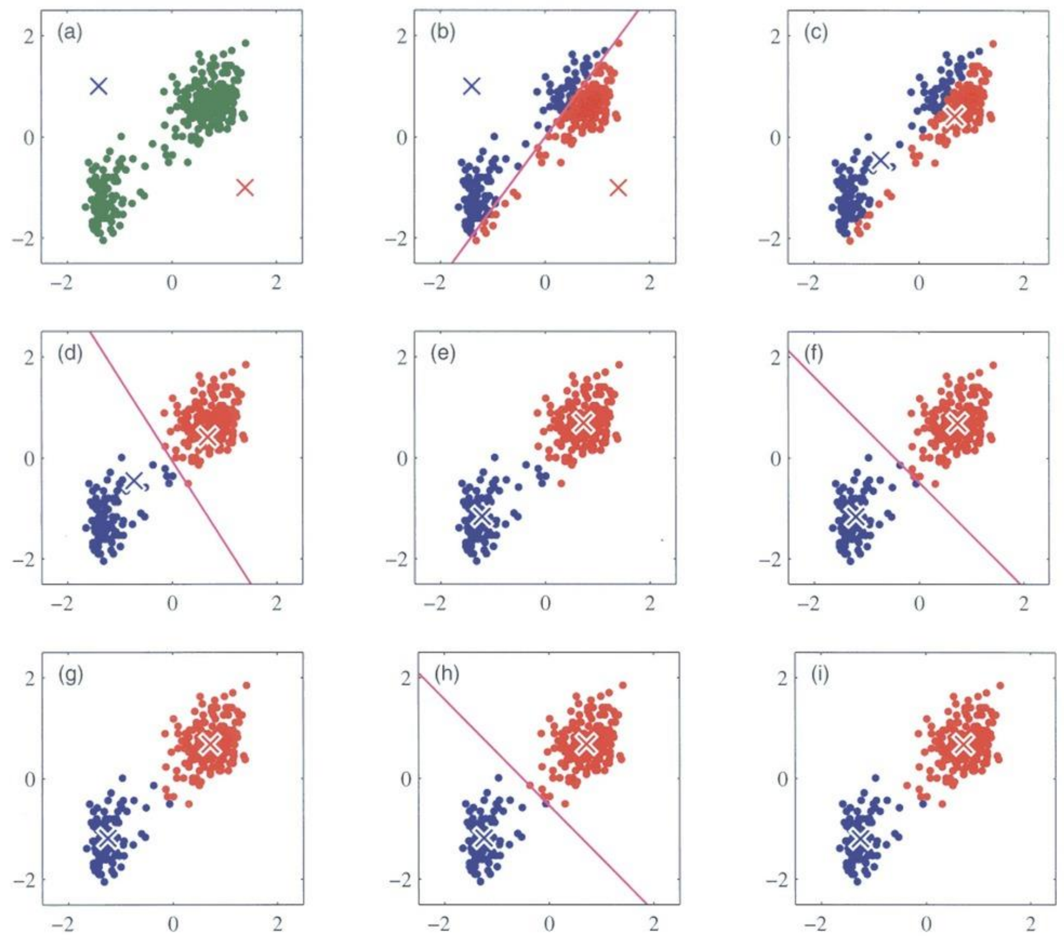
K-Means Clustering

- 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.
```



K-Means clustering (details)

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

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

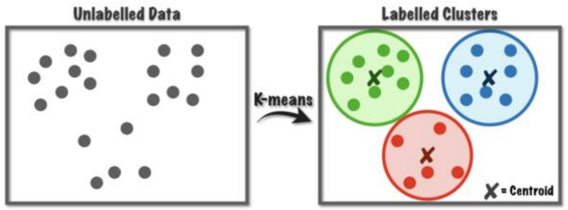
- This optimization says that we want to partition the observations into K clusters such that the **total** within-cluster variation is as small as possible.
- If we use Euclidian distance, then:

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

Initial positioning of the centroids matter!

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

K-Means clustering pros and cons



Pros

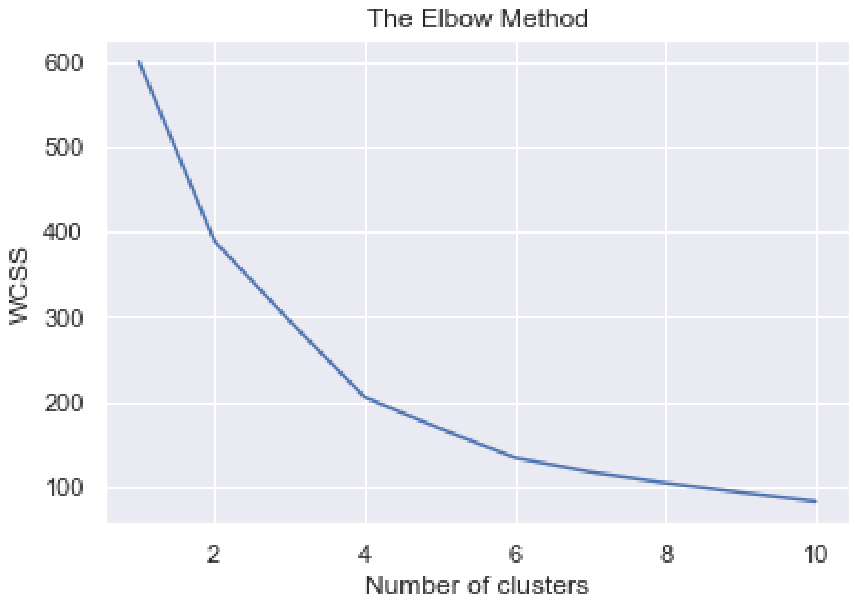
- Simple!
- The k-means algorithm is fast and works well on very large datasets.
- Can help visualize the data and facilitate detecting trends or outliers.
- The k-means algorithm is among the most used algorithms in investment practice.

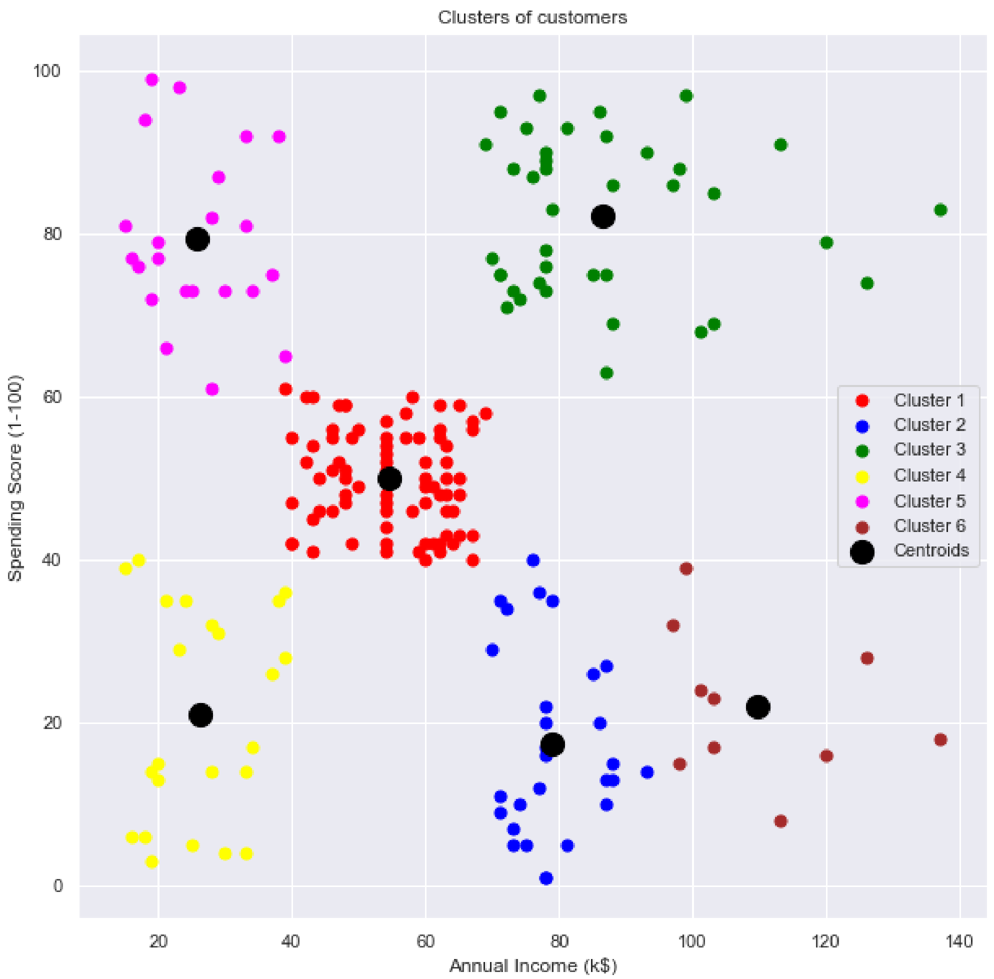
Cons

- 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)

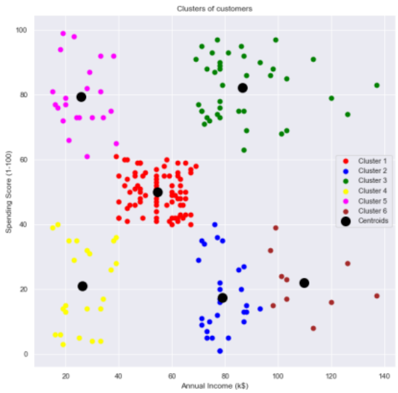
Optimal number of K (the elbow method)

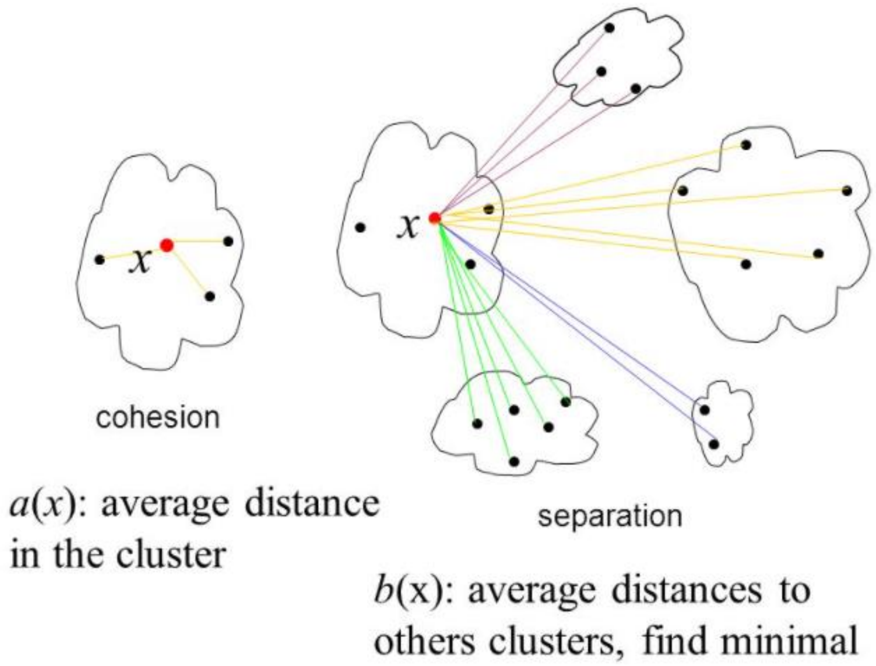
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \mathrm{WCV}(C_k) \right\}$$

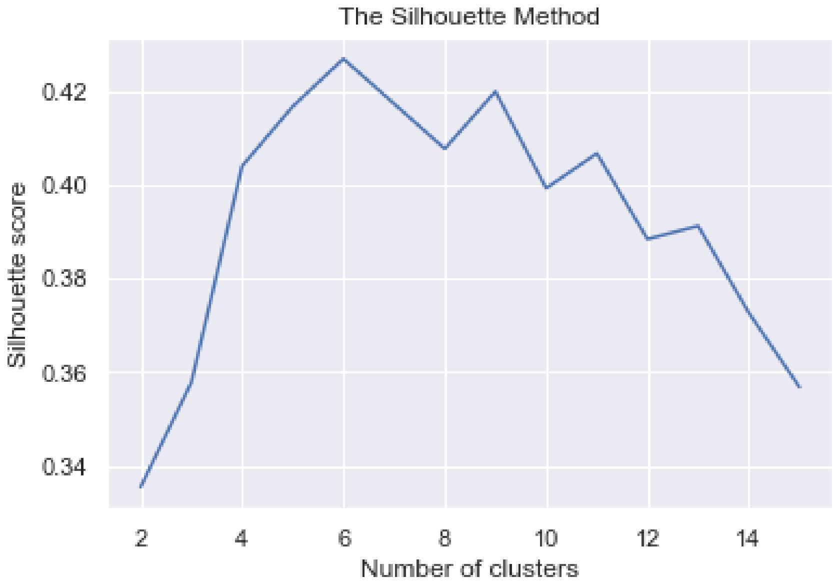




Optimal number of K (the Silhouette method)







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

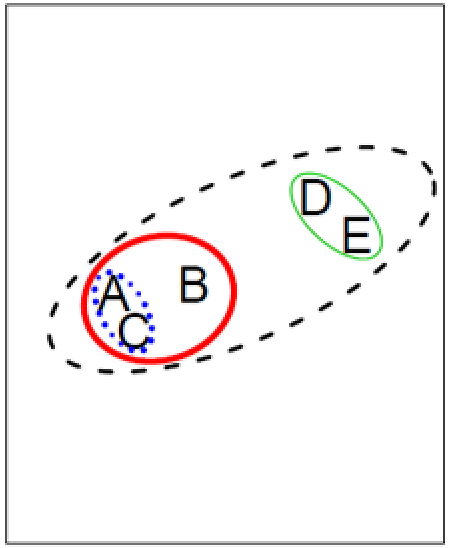
Part II

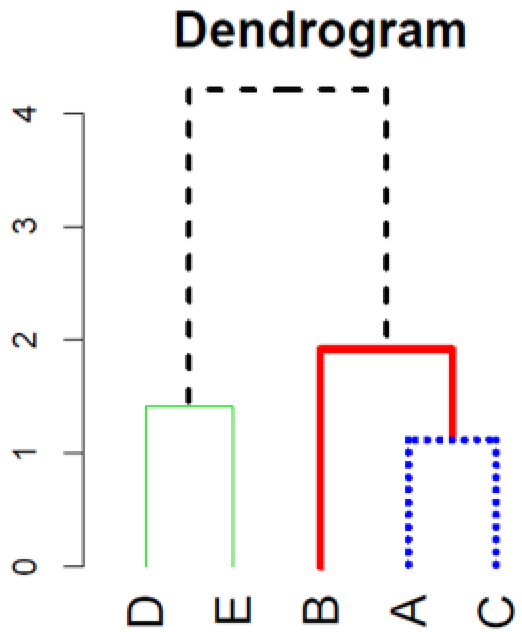
- Hierarchical Clustering
  1. Agglomerative
  2. Divisive

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.

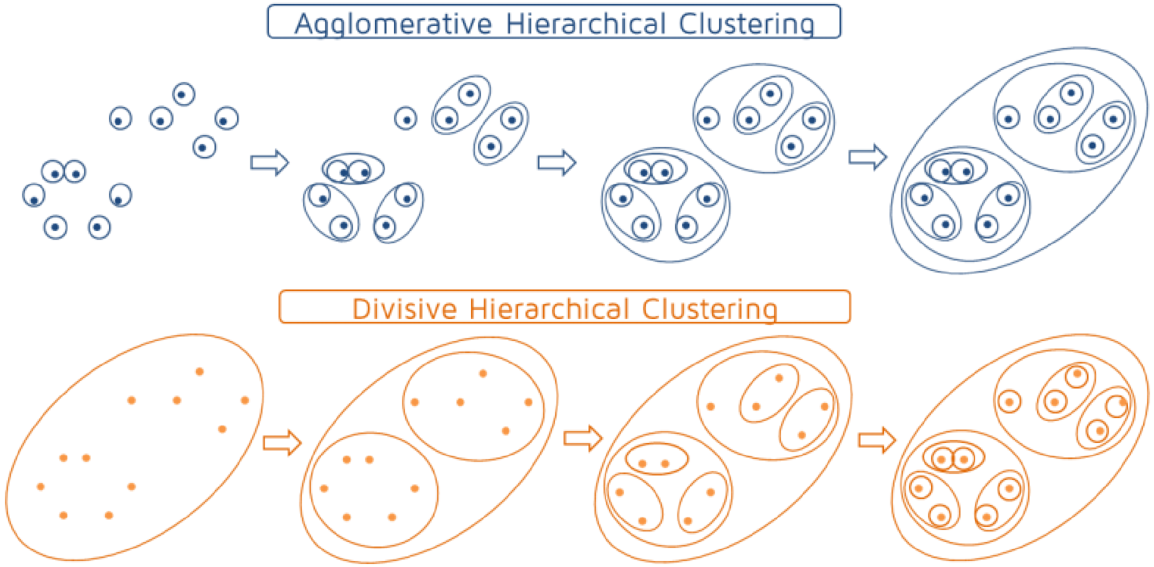
- Hierarchical clustering is an iterative procedure used to build a **hierarchy of clusters**.
- Using a dendrogram (a type of tree diagram which highlights the hierarchical relationships among the clusters), hierarchical clustering has the advantage of allowing the analyst to examine alternative partitioning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

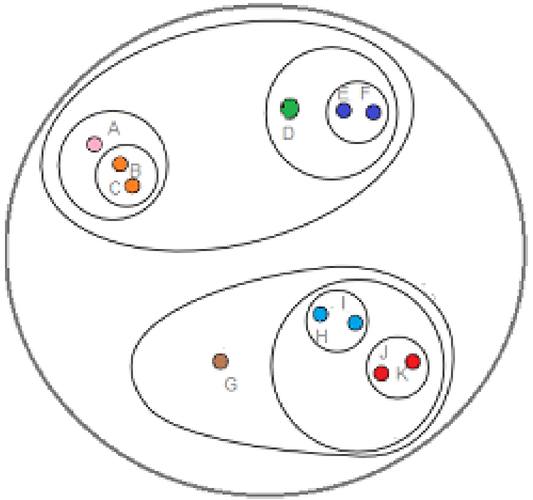


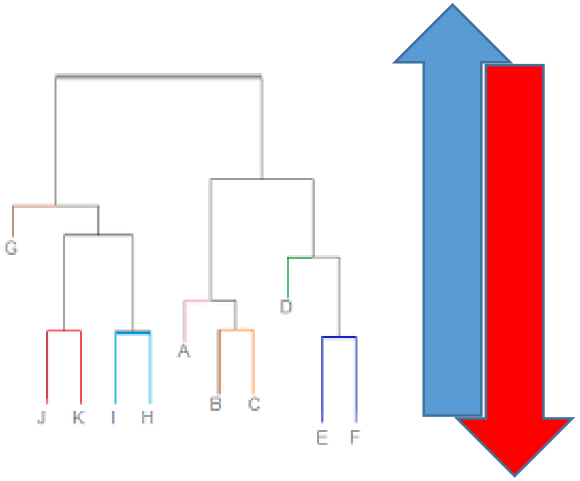


Agglomerative (bottom-up) vs Divisive (top-down) HCA

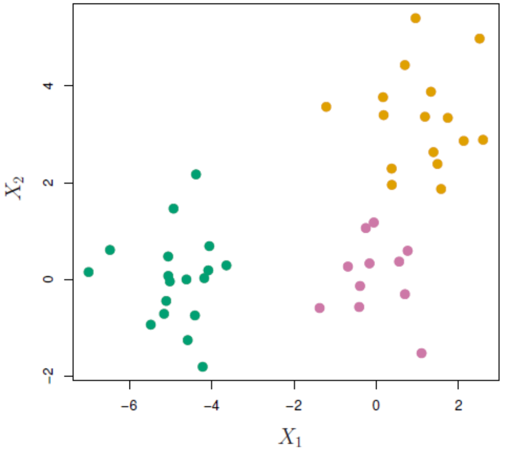
- Agglomerative: start with each observation being treated as its own cluster
- Divisive: starts with all the observations belonging to a single cluster.

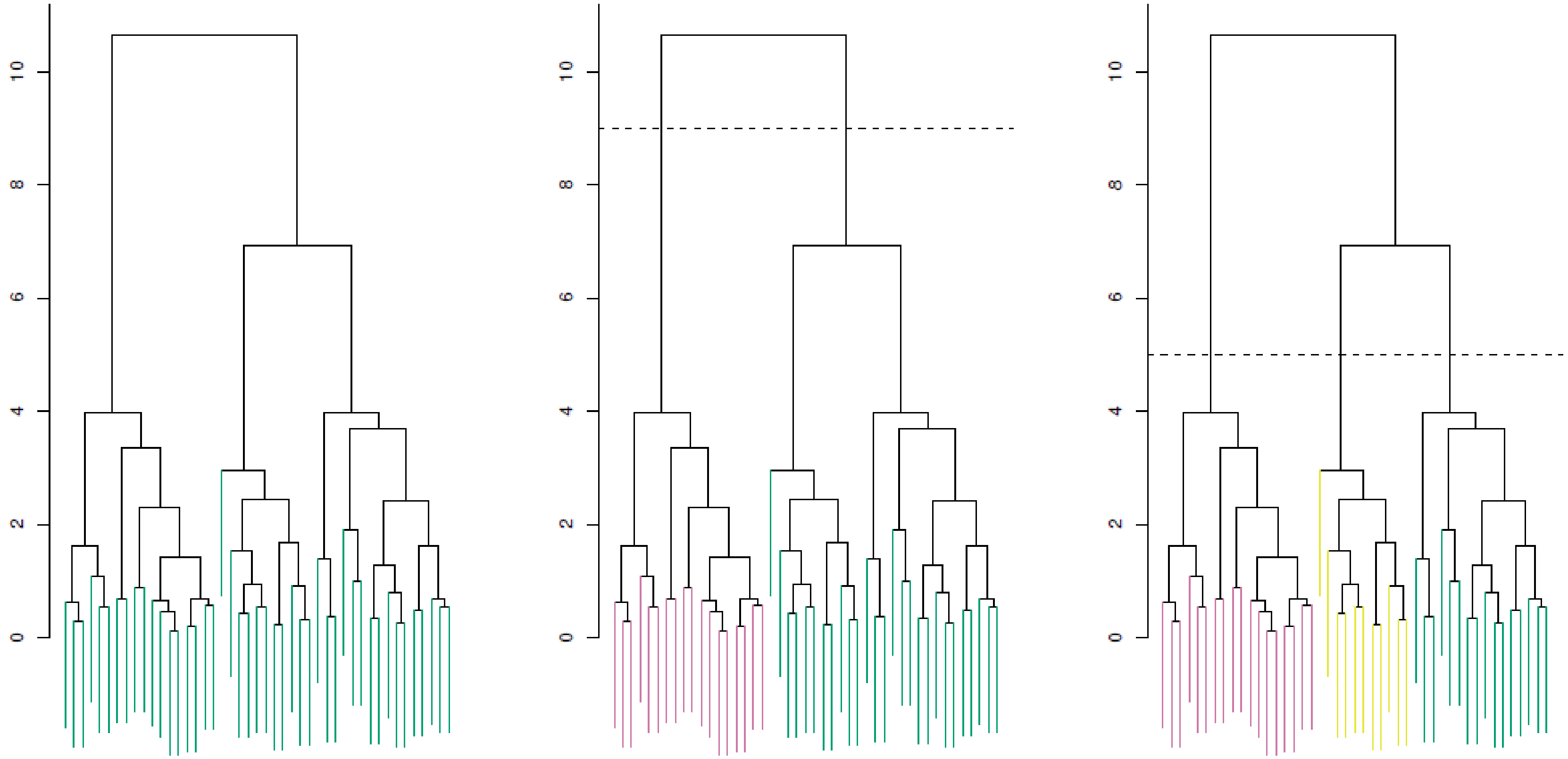






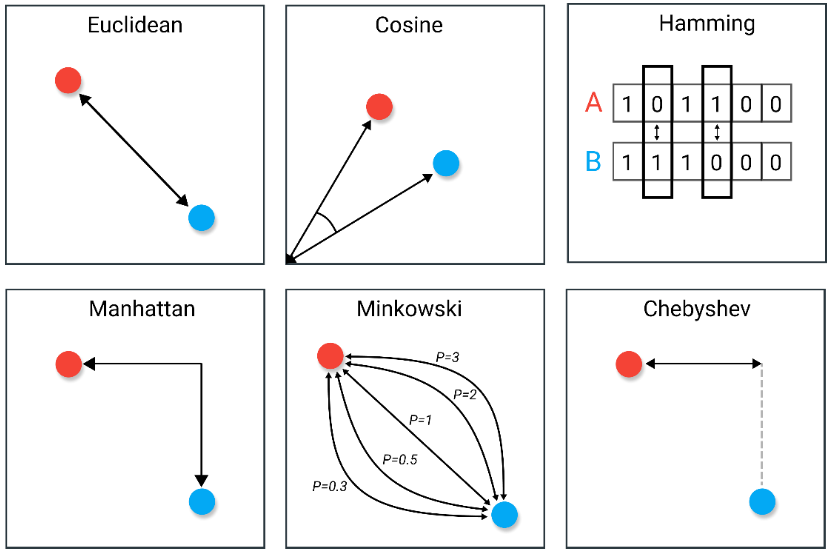
An example

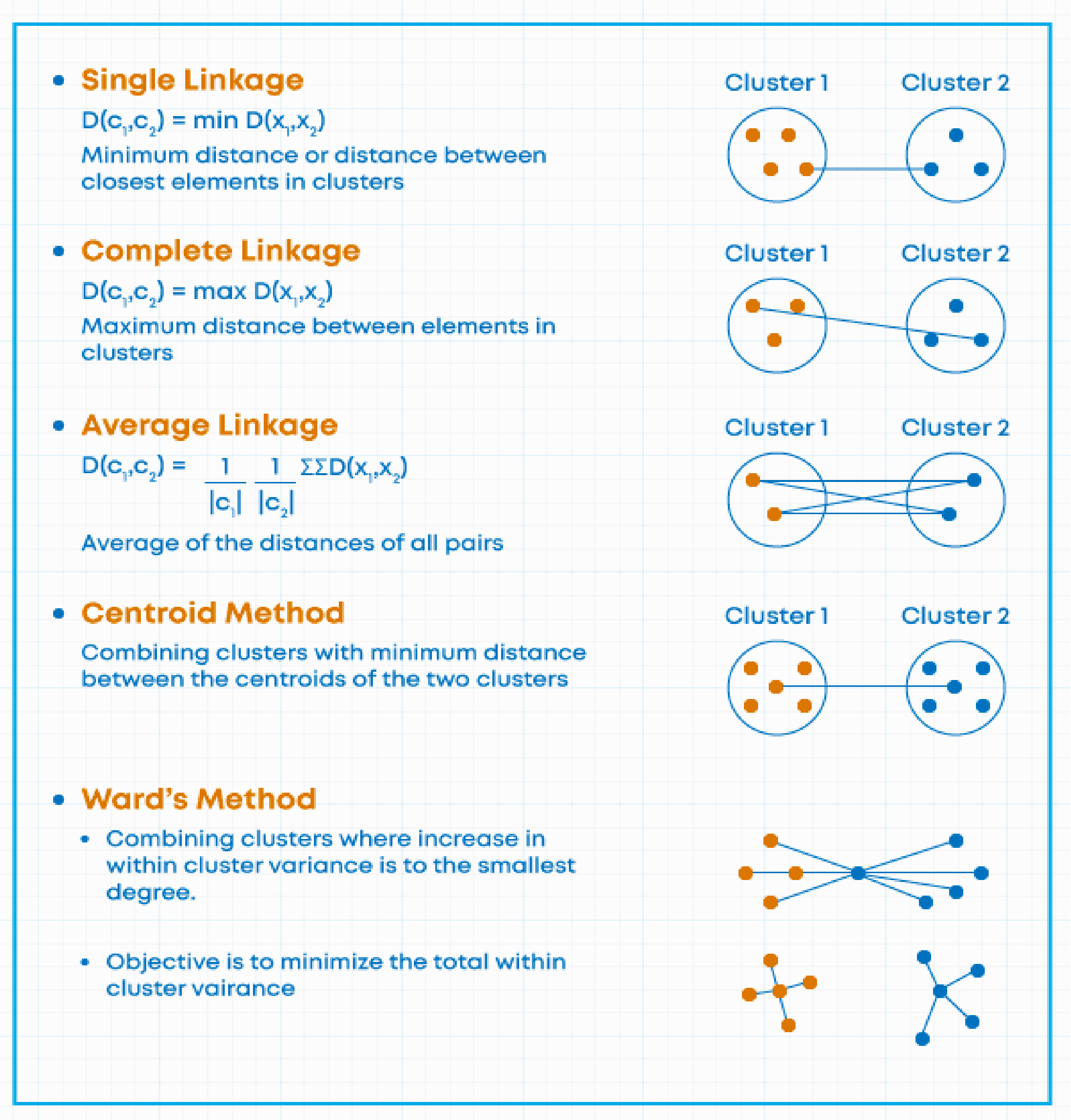




Types of Linkage (distance between two clusters)

- 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.





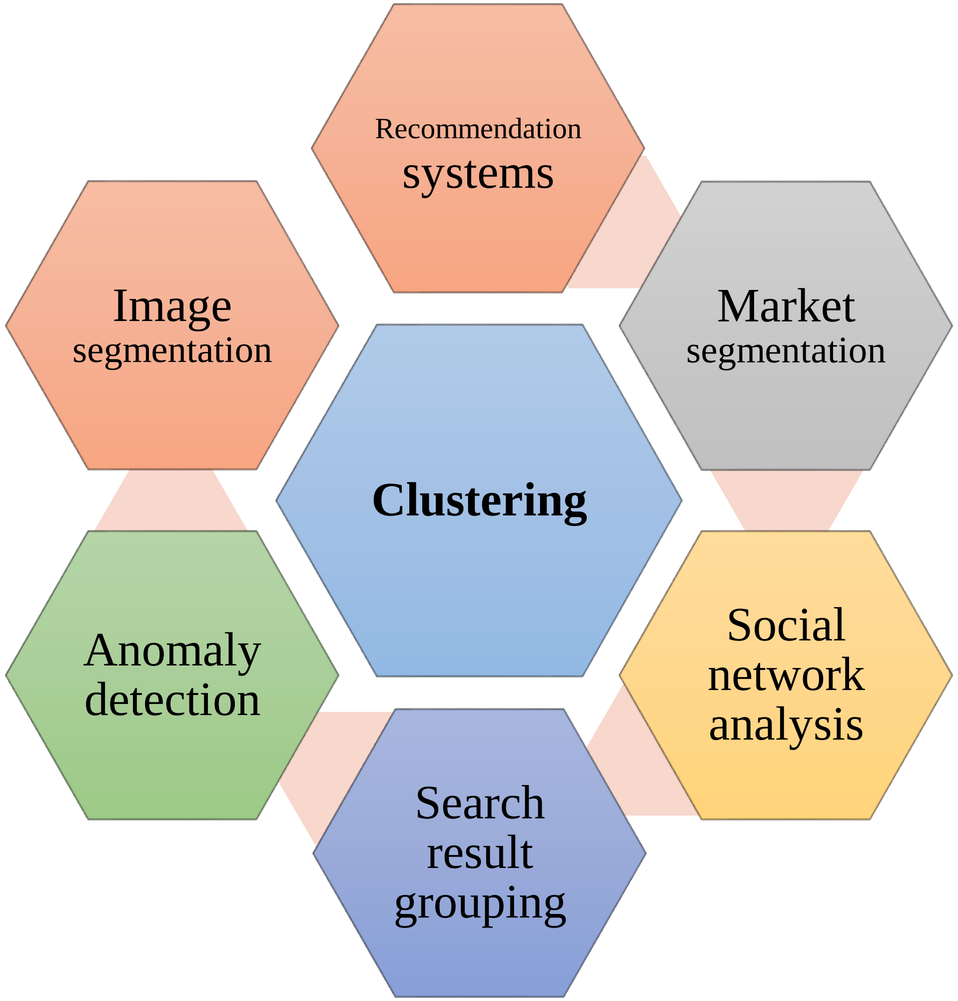
Hierarchical Clustering discussion

- The agglomerative method is the approach typically used with large datasets because of the algorithm's fast computing speed.
- The agglomerative clustering algorithm makes clustering decisions based on local patterns without initially accounting for the global structure of the data. As such, the agglomerative method is well suited for identifying small clusters.
- The divisive method starts with a holistic representation of the data, so it is designed to account for the global structure of the data and thus is better suited for identifying large clusters.
- What dissimilarity measure should be used?
- What type of linkage should be used?
- There is no commonly agreed-upon way to decide where to cut the tree.

Part III

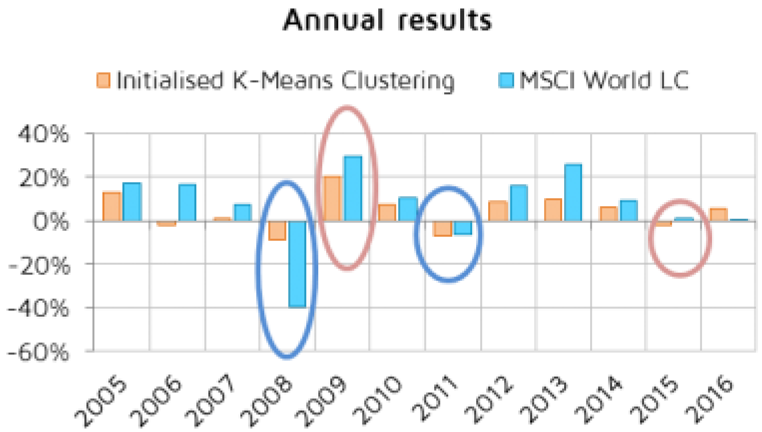
Applications in finance

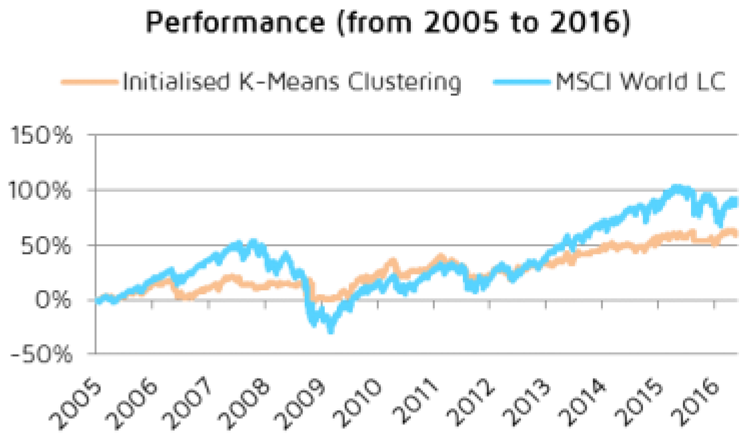
Applications of clustering



Applications in Finance

- Clustering algorithms are particularly useful in the many investment problems and applications in which the concept of similarity is important.
- 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.





Clustering stocks based on co-movement similarity

Exhibit 23 Dataset of Eight Stocks from the S&P 500 Index

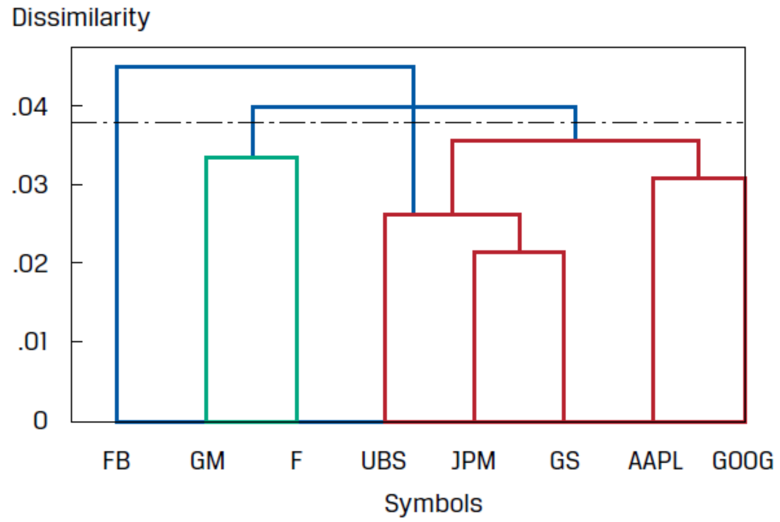
Description: Daily adjusted closing prices of eight S&P 500 member stocks

Trading Dates: 30 May 2017 to 24 May 2019

Number of Observations: 501

Stocks (Ticker Symbols): AAPL, F, FB, GM, GS, GOOG, JPM, and UBS

Exhibit 26 Dendrogram for Hierarchical Agglomerative Clustering

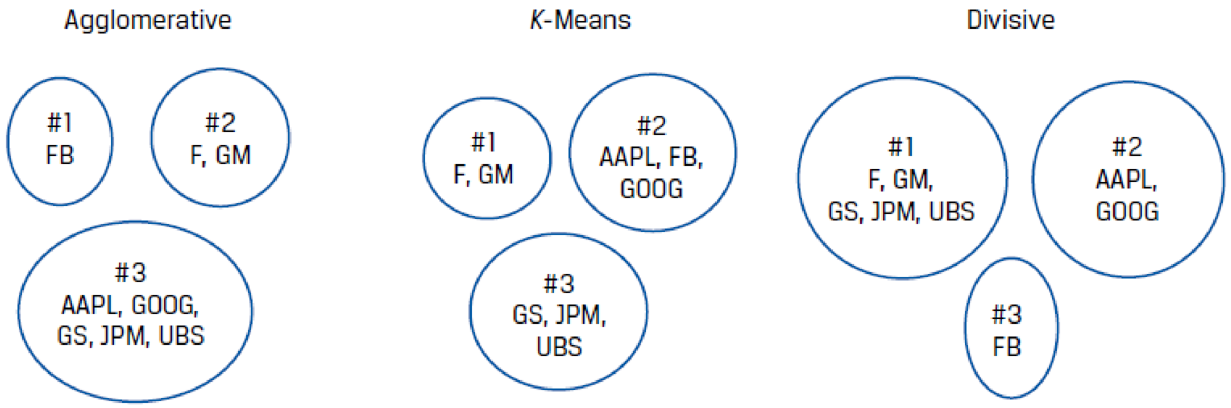


Developed and written by Matthew Dixon, PhD, FRM.

Comparing clustering methods

Exhibit 27 Comparison of Results of Different Clustering Algorithms

	Agglomerative	K-means	Divisive
AAPL	3	2	2
F	2	1	1
FB	1	2	3
GM	2	1	1
GOOG	3	2	2
GS	3	3	1
JPM	3	3	1
UBS	3	3	1



Developed and written by Matthew Dixon, PhD, FRM.