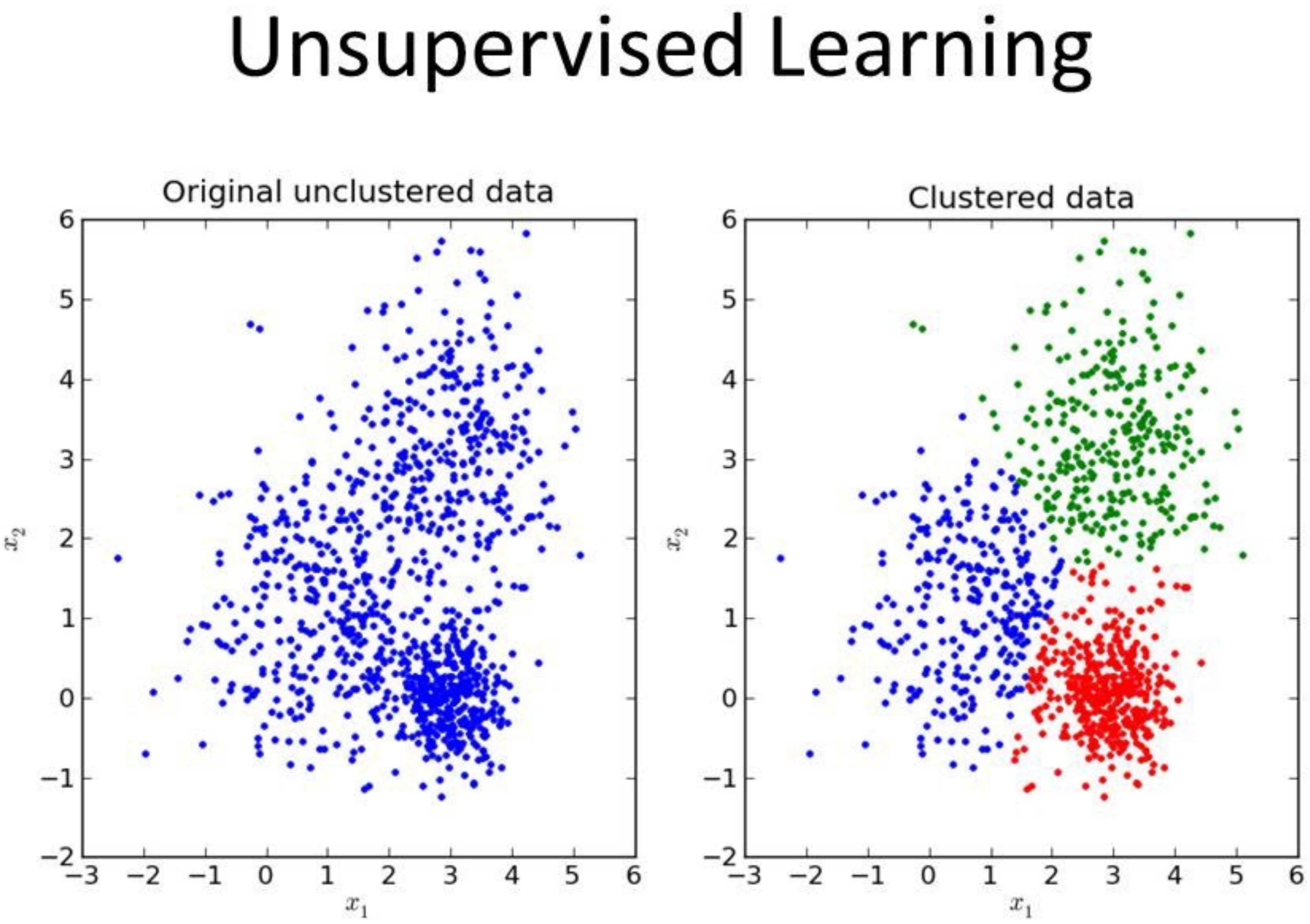
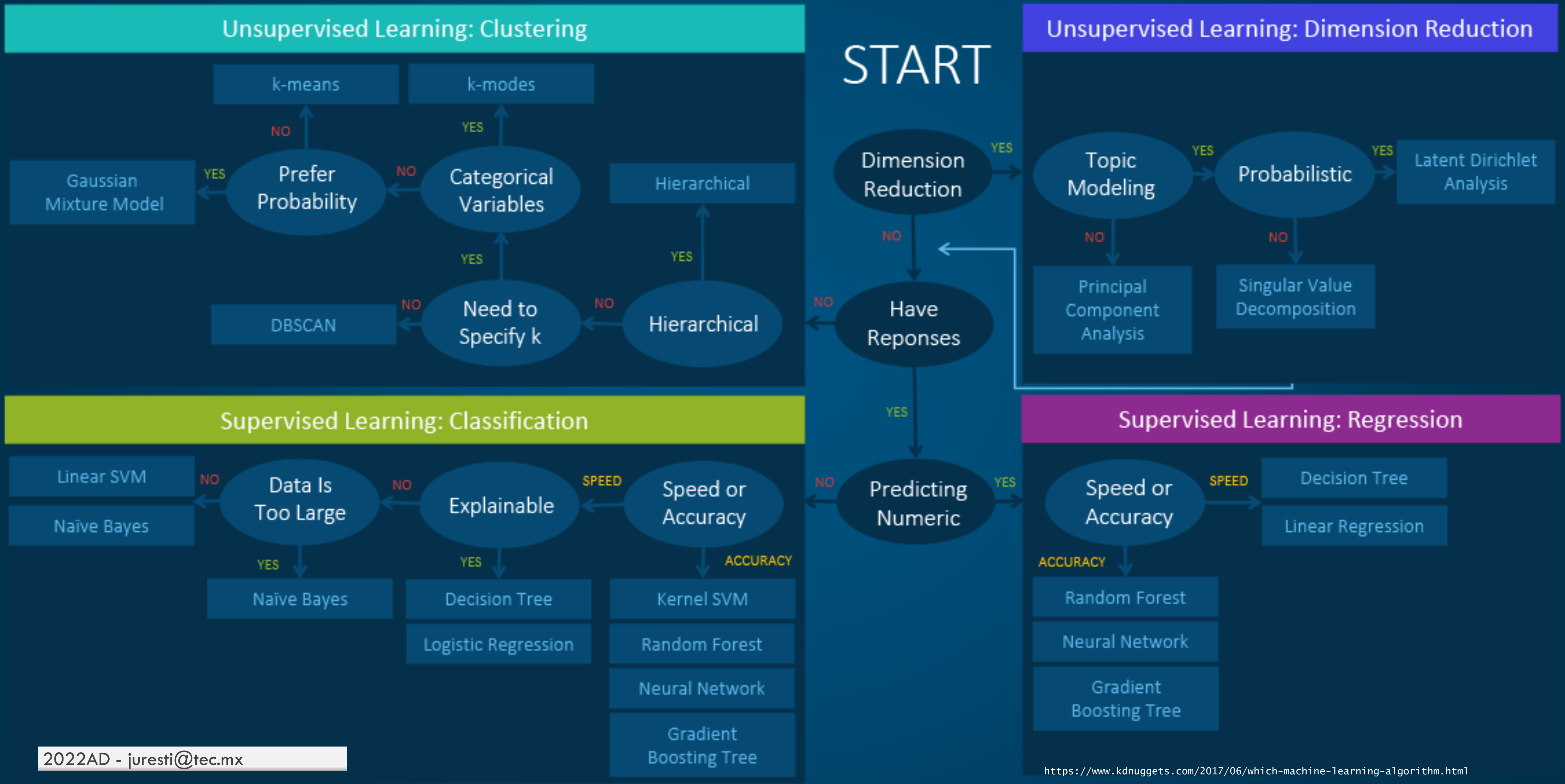


https://miro.medium.com/max/1920/1*VACikYaZIHb2OctTRohn8A.jpeg

Aprendizaje No Supervisado

TC2034 – Modelación del Aprendizaje con Inteligencia Artificial

Machine Learning Algorithms Cheat Sheet



2022AD - juresti@tec.mx

https://www.kdnuggets.com/2017/06/which-machine-learning-algorithm.html

Unsupervised Learning

Definition

- Uses information that is neither classified nor labelled
- The algorithm (learner) has to discover patterns without guidance
- Usually groups information by:
  - Similarities
  - Patterns
  - Differences
- There are two categories of algorithms:
  - Clustering
    - Discover inherent grouping in data
  - Dimension Reduction (Association)
    - Discover rules that describe large portions of data

Unsupervised Learning…

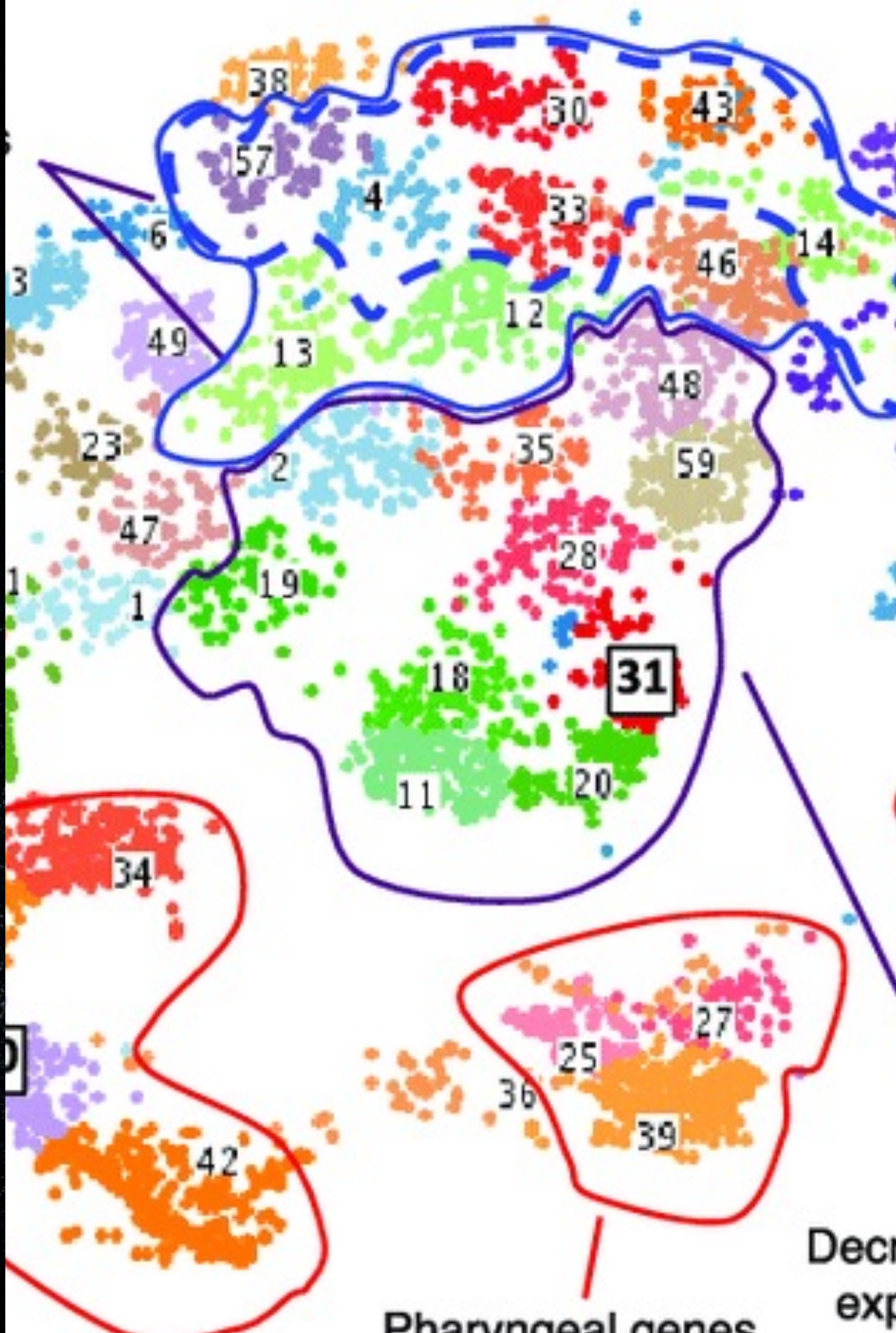
Applications

- News sections: categorize articles
- Computer vision: object recognition
- Medical imaging: image detection, classification, segmentation
- Anomaly detection: discover atypical points
- Customer persona profiles
- Recommendation engines

Challenges

- Computational complexity – high volume of data
- Longer training times
- High risk of inaccurate results
- Need of human intervention for validation
- Lack of transparency to know exactly how data was organized

Clustering



https://www.researchgate.net/profile/Adam-Warner-3/publication/333333035/figure/fig2/AS:761900535119875@1558662651725/Clustered-gene-expression-visualized-with-t-SNE-Gene-expression-values-TPM-were.png

Clustering

Definition

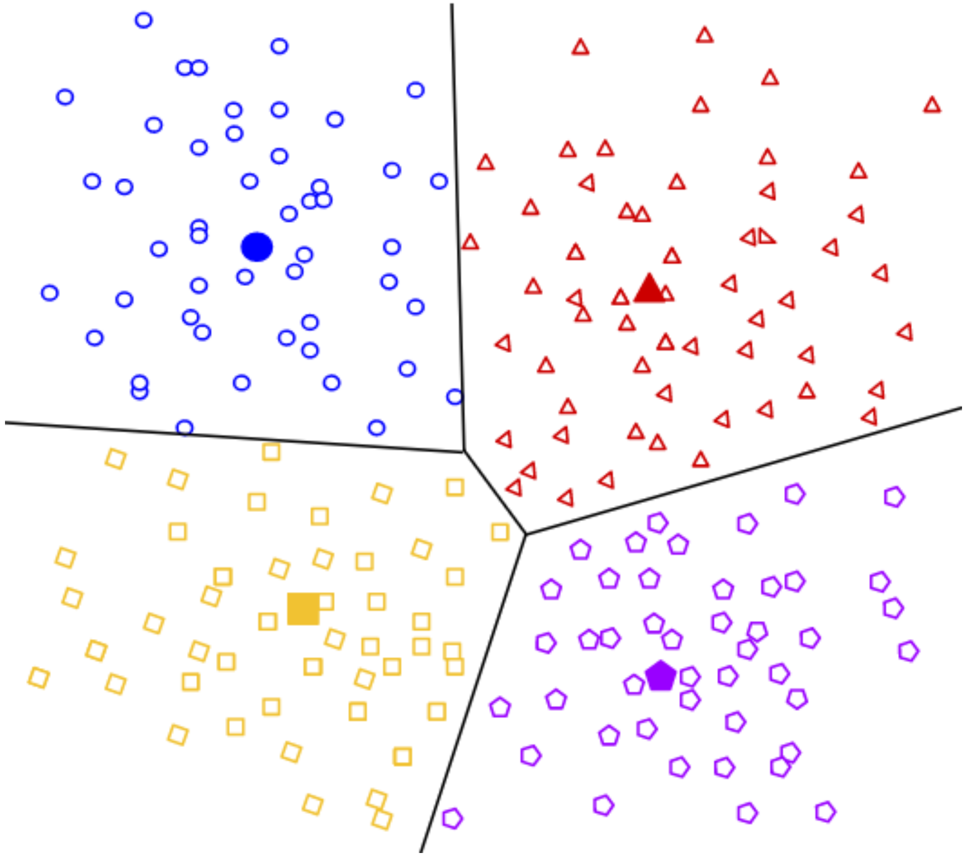
- Grouping of unlabeled instances
- Useful to understand a dataset
- Need for a ***similarity measure*** to group similar instances
- Groups are classified by features
  - The more features, the more difficult grouping is

Applications

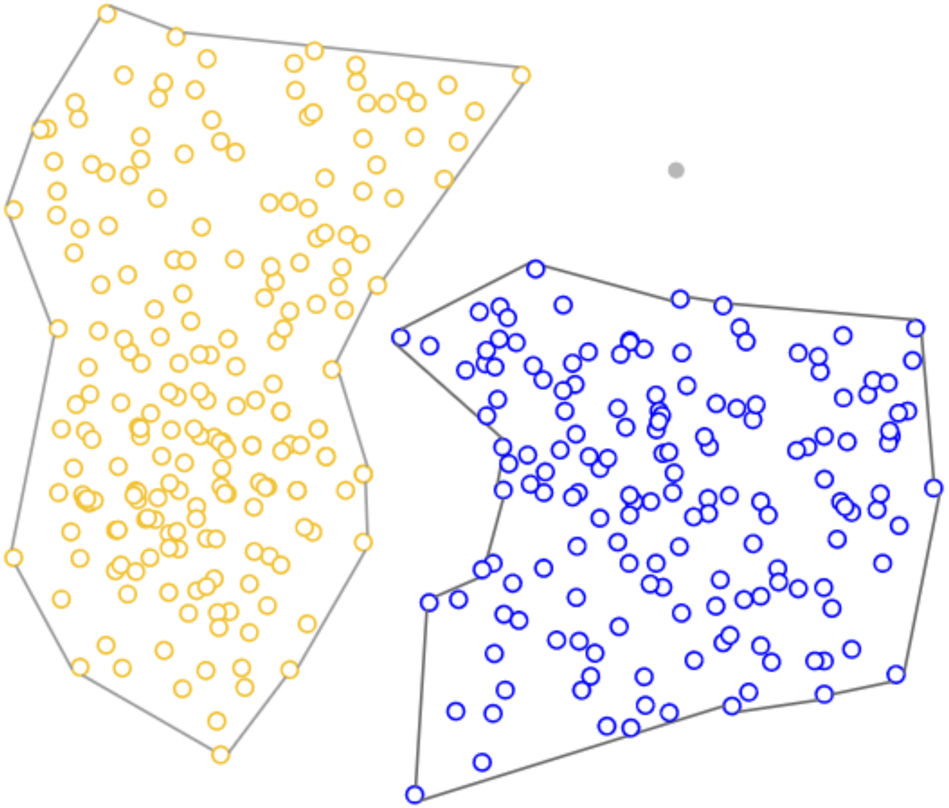
- Market segmentation
- Social network analysis
- Medical imaging
- Image segmentation
- Anomaly detection

Types of clustering

Centroid-based / Partitional



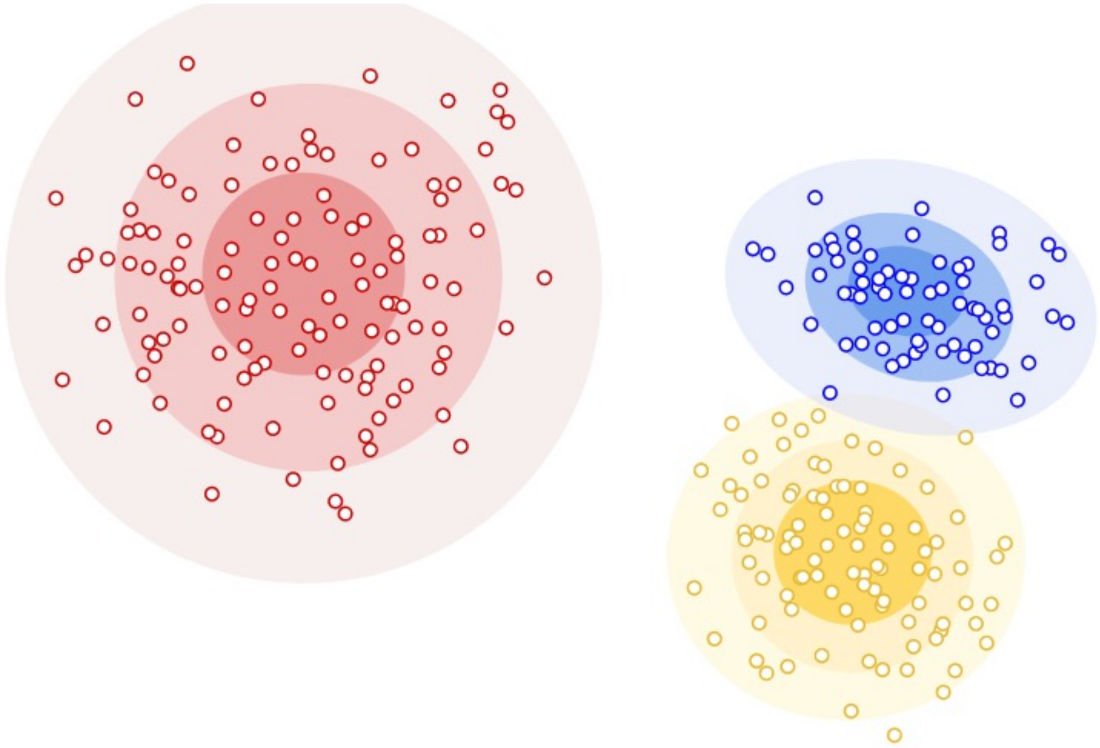
Density-based



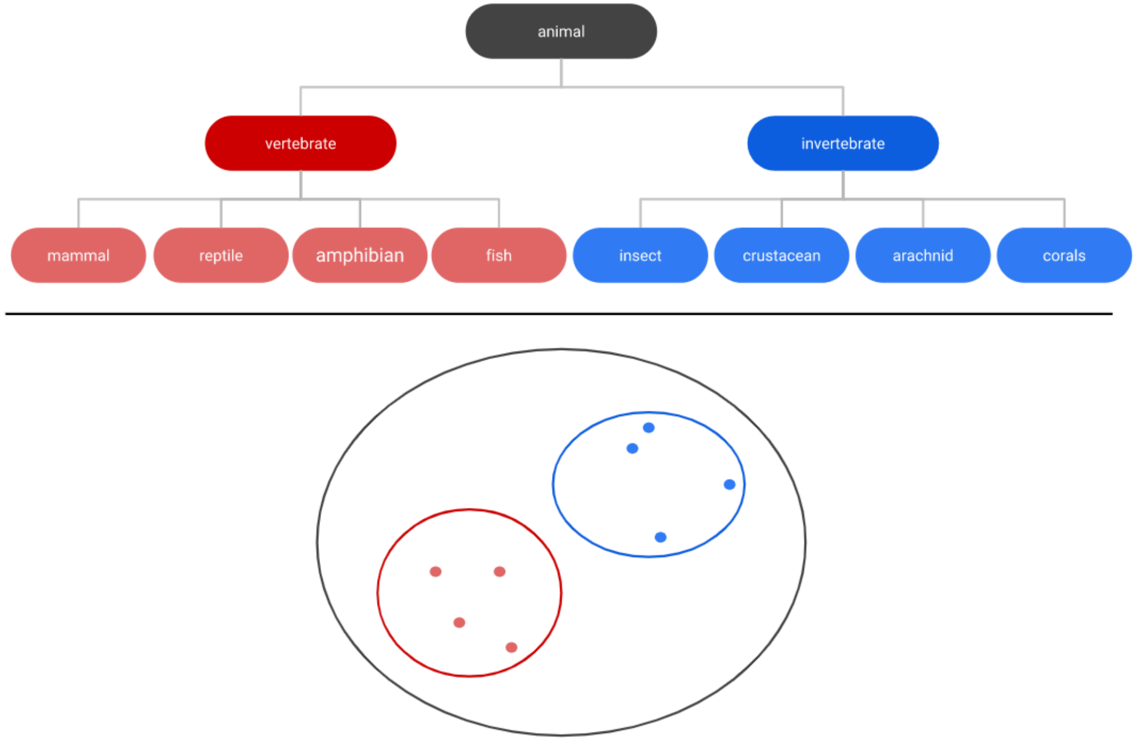
http: 2022AD - juresti@tec.mx ing/clustering/

Types of clustering…

Distribution-based



Hierarchical



Clustering

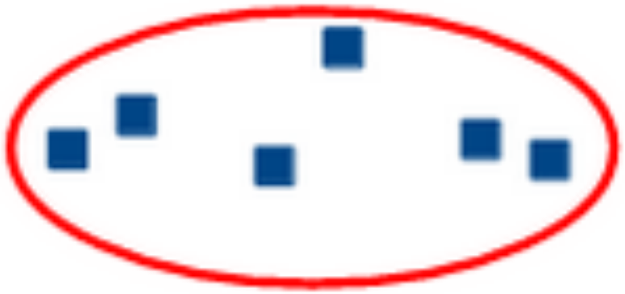
Main categories

Partitional Clustering	Hierarchical Clustering	Density-based Clustering

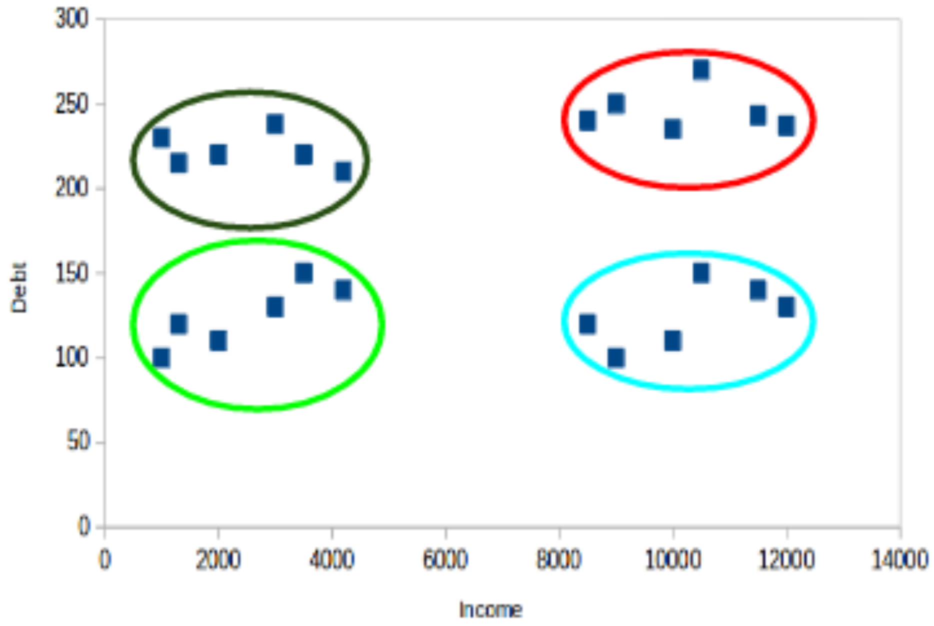
• Strengths
  • Suited for spherical shape data
  • Scalable in terms of algorithm complexity
 	
• Weaknesses
  • Not suitable for complex shape and different sizes clusters
	• Strengths
	  • Can use a dendogram
	  • Reveal finer details about relationships between data points
	
	• Weaknesses
	  • Computationally expensive
	  • Sensitive to noise and outliers
		• Strengths
		  • Suited for non-spherical shapes
		  • Resistant to outliers
		
		• Weaknesses
		  • Not suited for high-dimensional spaces
		  • Problems when density varies in clusters

Properties of clusters

All data points in a cluster must be similar to each other



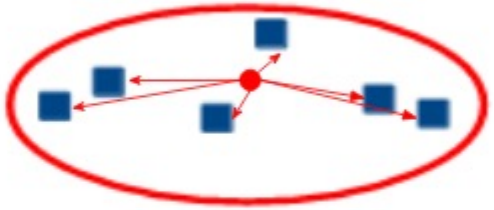
Data points from different clusters should be as different as possible



Evaluation metrics for clustering

Intracluster distance (*inertia*)

- Calculate distance from all points to the center of the cluster
- Tries to create very compact clusters
- The smaller the *inertia*, the better the cluster



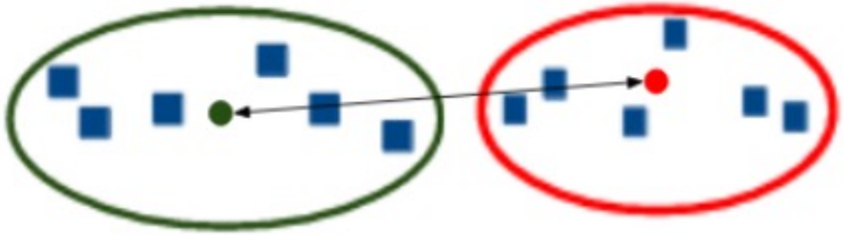
Intra cluster distance

Intercluster distance (*Dunn index*)

- Measures distance between 2 clusters

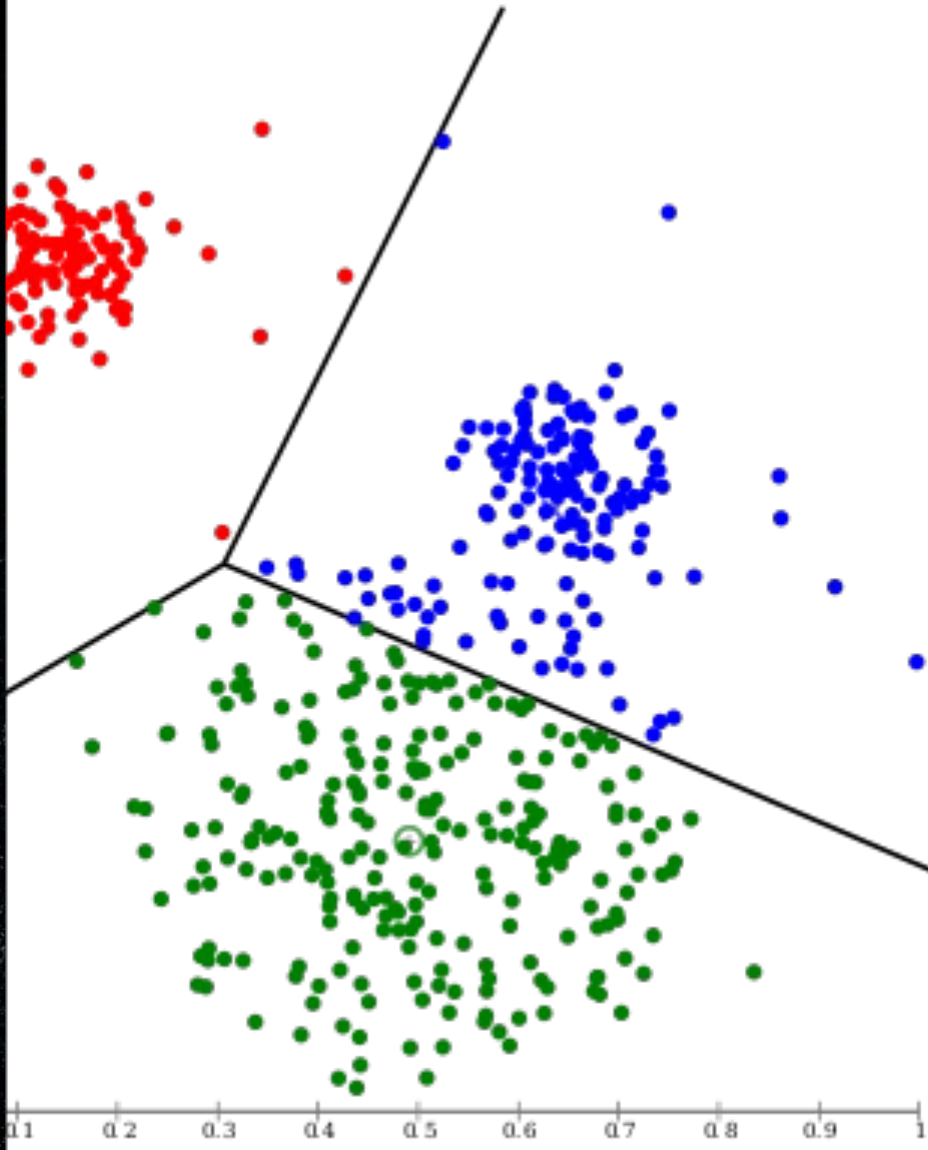
$$Dunn\ Index = \frac{\min(intercluster\ distance)}{\max\,(Inertia)}$$

- The bigger the *Dunn Index*, the more separated clusters are.



Inter cluster distance

K-means



https://upload.wikimedia.org/wikipedia/commons/thumb/e/e5/KMeans-Gaussian-data.svg/434px-KMeans-Gaussian-data.svg.png

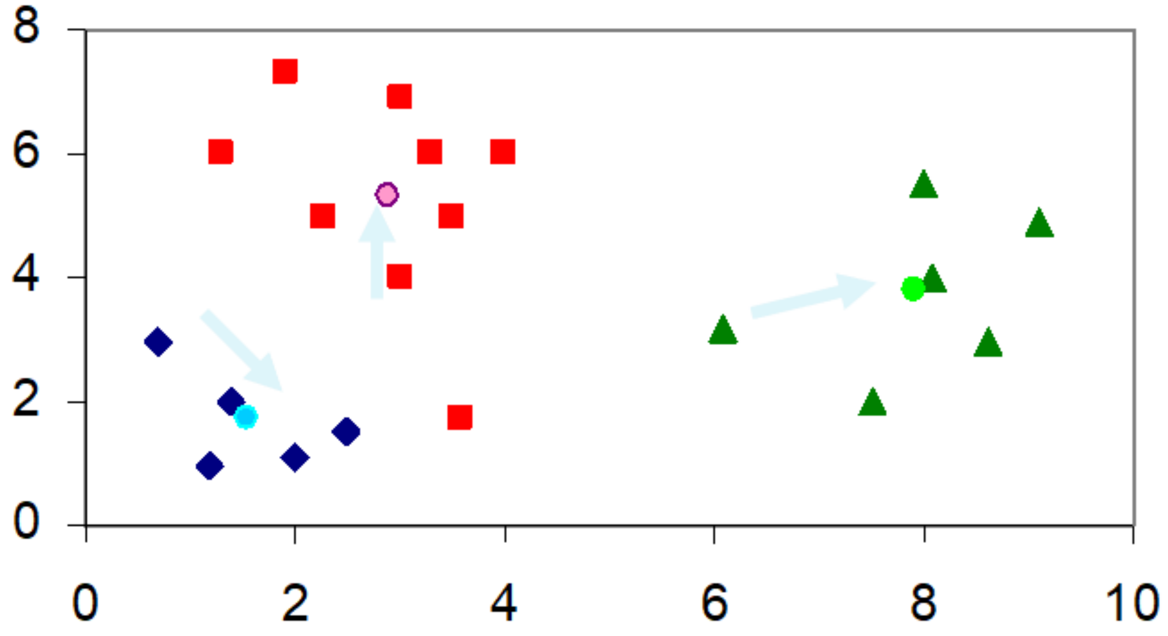
K-means

Definition

- Clustering method that tries to minimize the distance between points in a cluster (inertia)
    - Centroid-based algorithm
    - Distance-based algorithm

- Assign each point to a cluster based on its distance to the cluster's centroid

- Each cluster is associated with a centroid



K-means

Algorithm

1. Select number of clusters (K)

2. Select K random points from the dataset
    1. These will be used as centroids of our clusters

3. Assign the rest of the points to the *closest* cluster

4. Calculate *new centroids* for the current clusters

5. Repeat steps 3 and 4 until termination criterion is met

Distance from a point to another point

- Euclidian distance is used

$$P_1 = x_1, x_2, \ldots x_n \quad P_2 = y_1, y_2, \ldots y_n$$

$$D(P_1, P_2) = \sqrt{\sum_{i=1}^{i=n} (x_i - y_i)^2}$$

- In the case of 2 points (2D) use:

$$P_1 = x_1, x_2 \quad P_2 = y_1, y_2$$

$$D(P_1, P_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

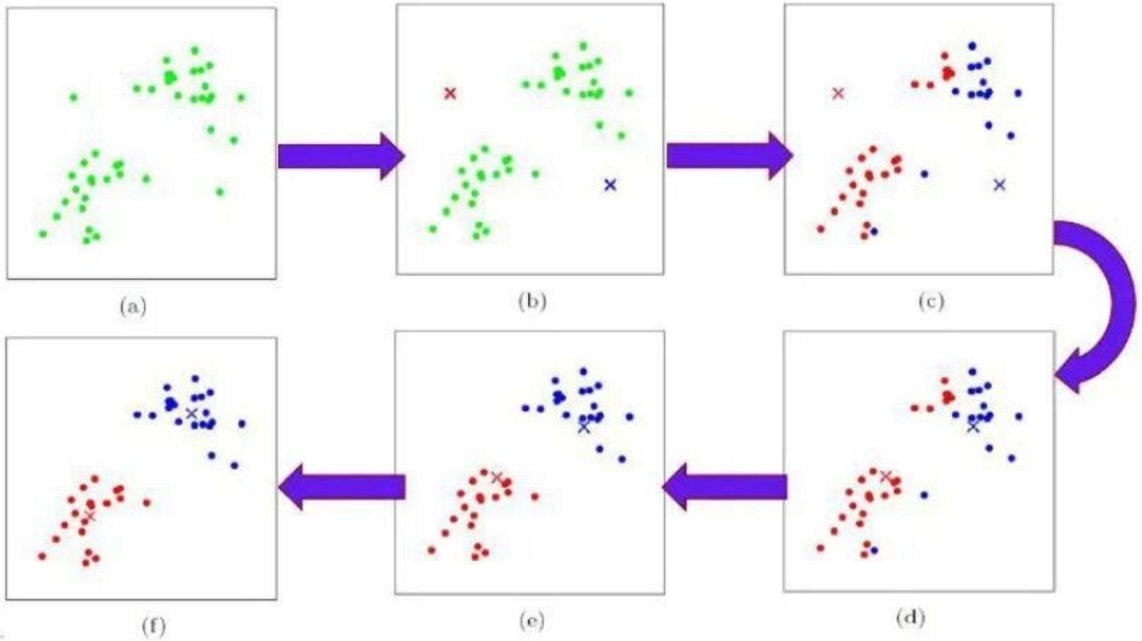
K-means

Termination Criteria

1. Centroids do not change
    1. After some iterations, centroids remain the same
    2. Algorithm is not learning new patterns

2. Points remain in the same cluster
    1. Centroids are changing
    2. No data points are changing from cluster after several iterations

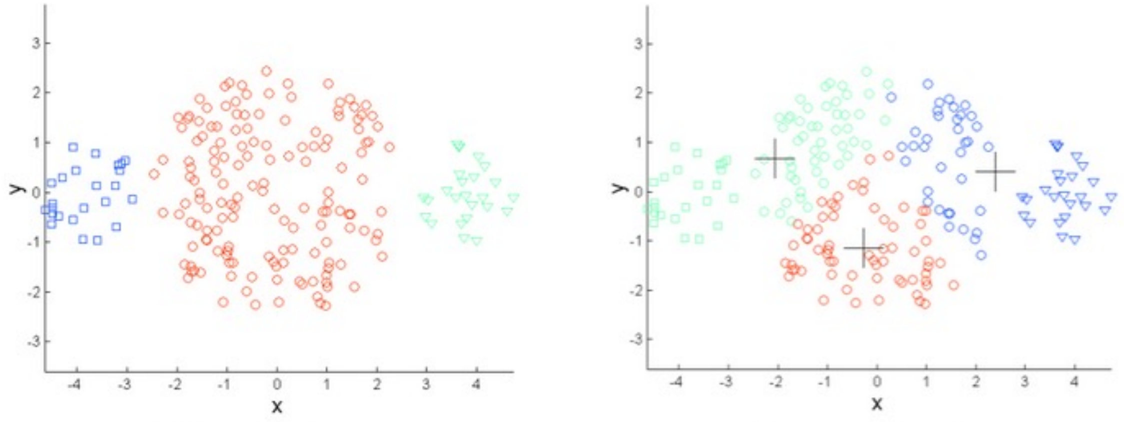
3. Maximum number of iterations reached



https://www.programmersought.com/images/437/0676d01a03e72b72253d6e62590a88c5.JPEG

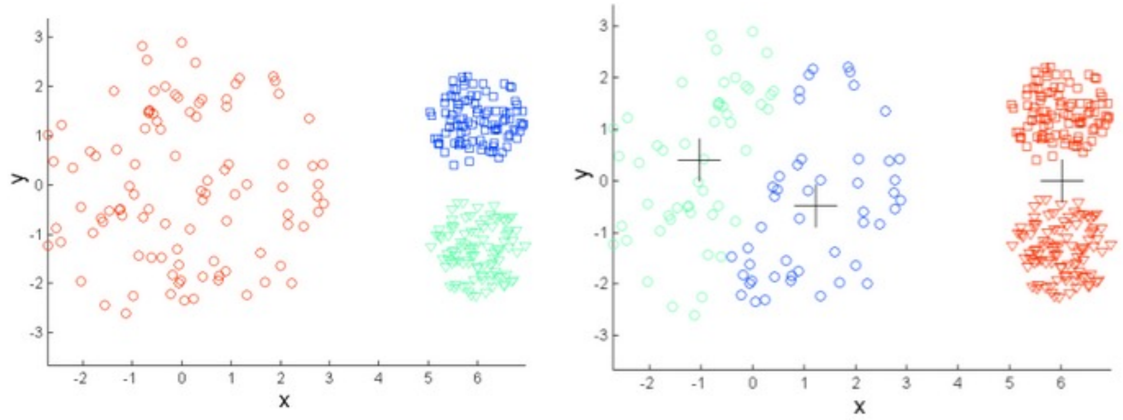
K-means: disadvantages

Clusters may have different sizes



Original Points K-means (k = 3)

Clusters may have different densities



Original Points K-means (k = 3)

Solution is usually to add more clusters

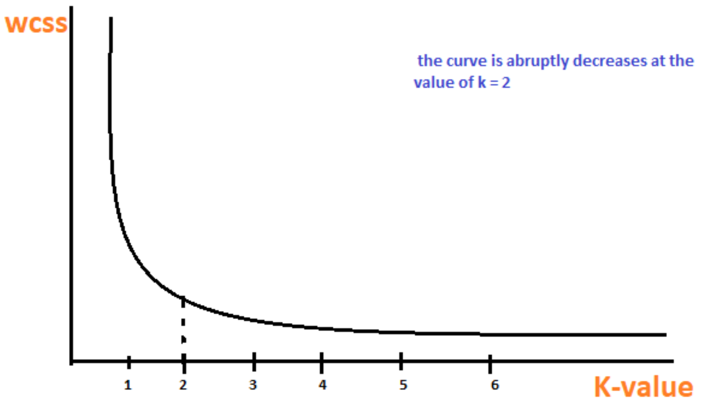
Choosing K

Elbow method

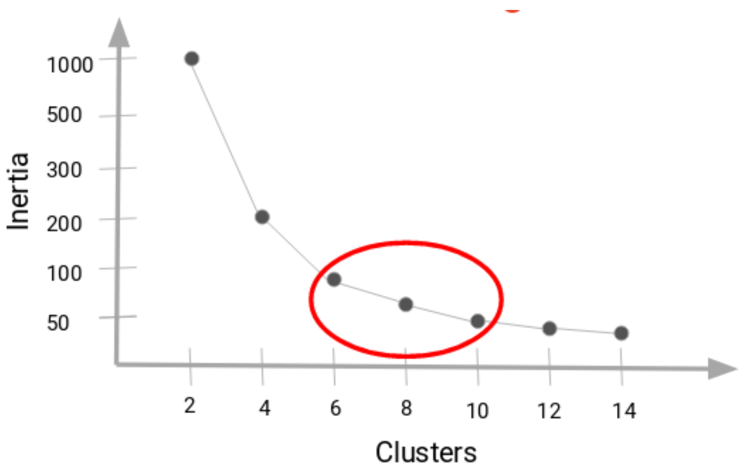
- Empirical method
- Perhaps the most famous method for choosing K

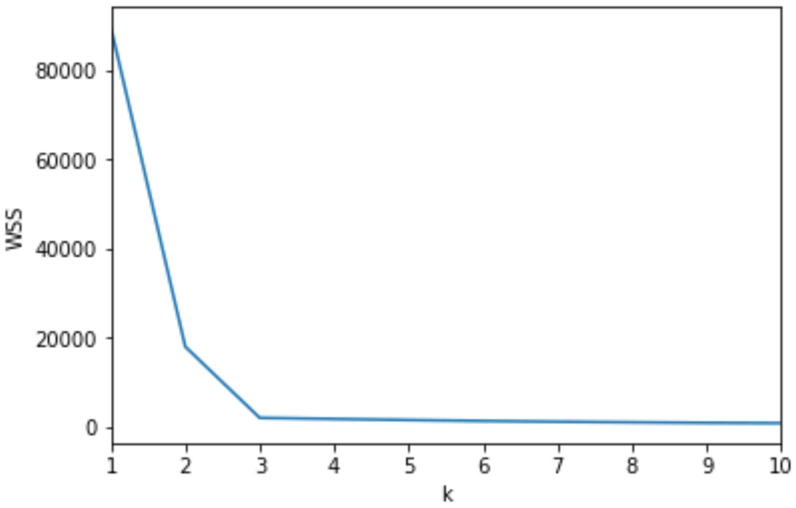
- Method
  - Calculate the Within-Cluster-Sum of Squared Errors (WCSS) for different values of K (intertia)
    - WCSS = sum(square(point distance to cluster centroid))

  - Choose the K where there is a decrease in the value of WCSS



https://editor.analyticsvidhya.com/uploads/12158wcss.png





https://miro.medium.com/max/407/1*O_JmBi6rM6PlrPFx_uNrVQ.png

p-content/uploads/2019/08/Screenshot-from-2019-08-09-16-01-34.png

Choosing K…

Silhouette Method

- Measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

- *a(i)* = similarity measure of the point ***i*** to its own cluster

For each data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

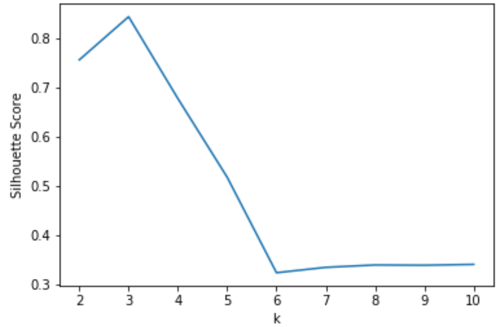
- *b(i)* = dissimilarity measure of i from points in other clusters

For each data point $i \in C_i$, we now define

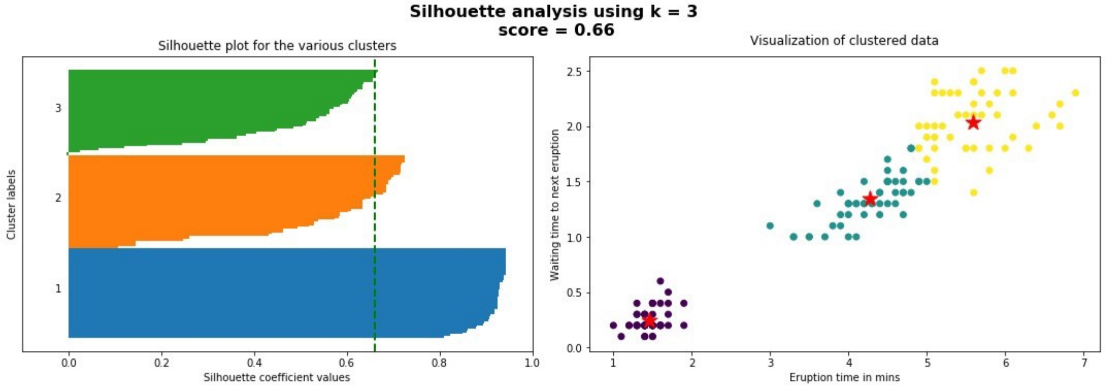
$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

- *d(i,j)* = distance from point i to point j

- *s(i)* varies from 1 to -1
  - A high value indicates a good number of clusters
  - A negative value indicates too many or too few clusters



https://miro.medium.com/max/389/1*Otg9XRLPq0jb8OPng6B9Og.png



https://medium.com/@cmukesh8688/silhouette-analysis-in-k-means-clustering-cefa9a7ad111

K-means

Advantages

- Very simple to implement
- Adapts new examples (instances) easily
- Can generalize clusters of different sizes and shapes

Disadvantages

- Very sensitive to outliers
- Choosing K is not an easy task
- Not so efficient when the number of dimensions increases

K-means…

Class Example

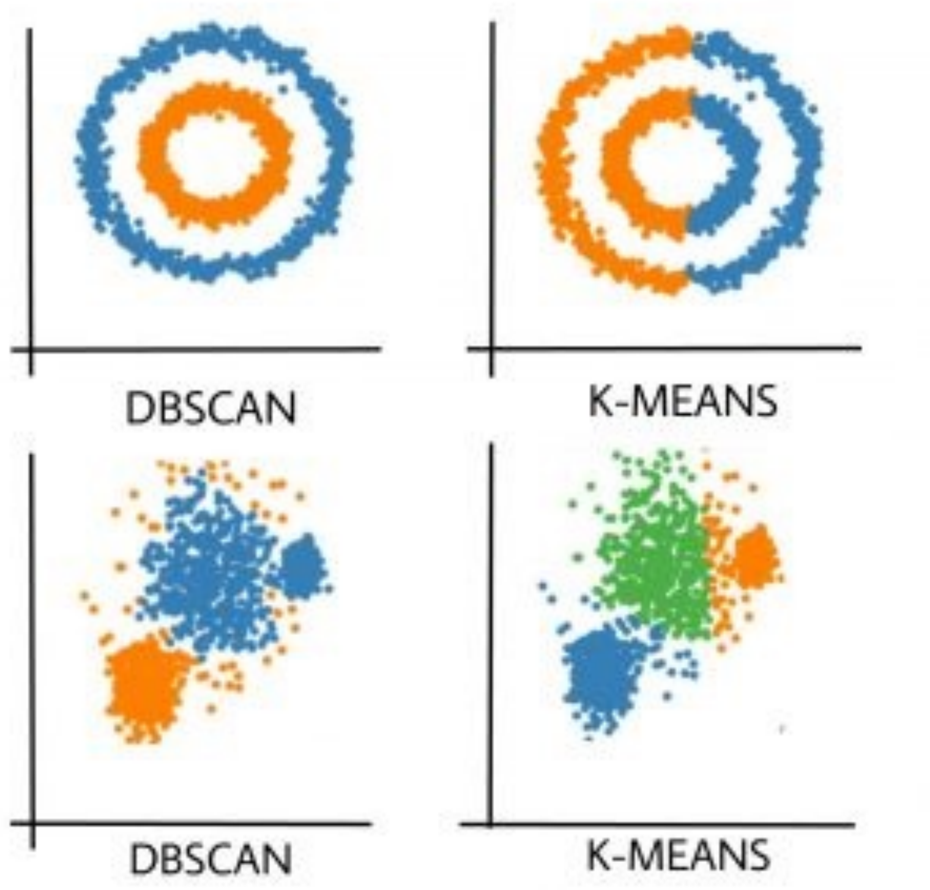
- K-means.ipynb

- Check Github

DBSCAN

DBSCAN

Density-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise



https://media.geeksforgeeks.org/wp-content/uploads/PicsArt_11-17-08.07.10-300x300.jpg

- A point belongs to a class if it is close to *many* points in that class (cluster)

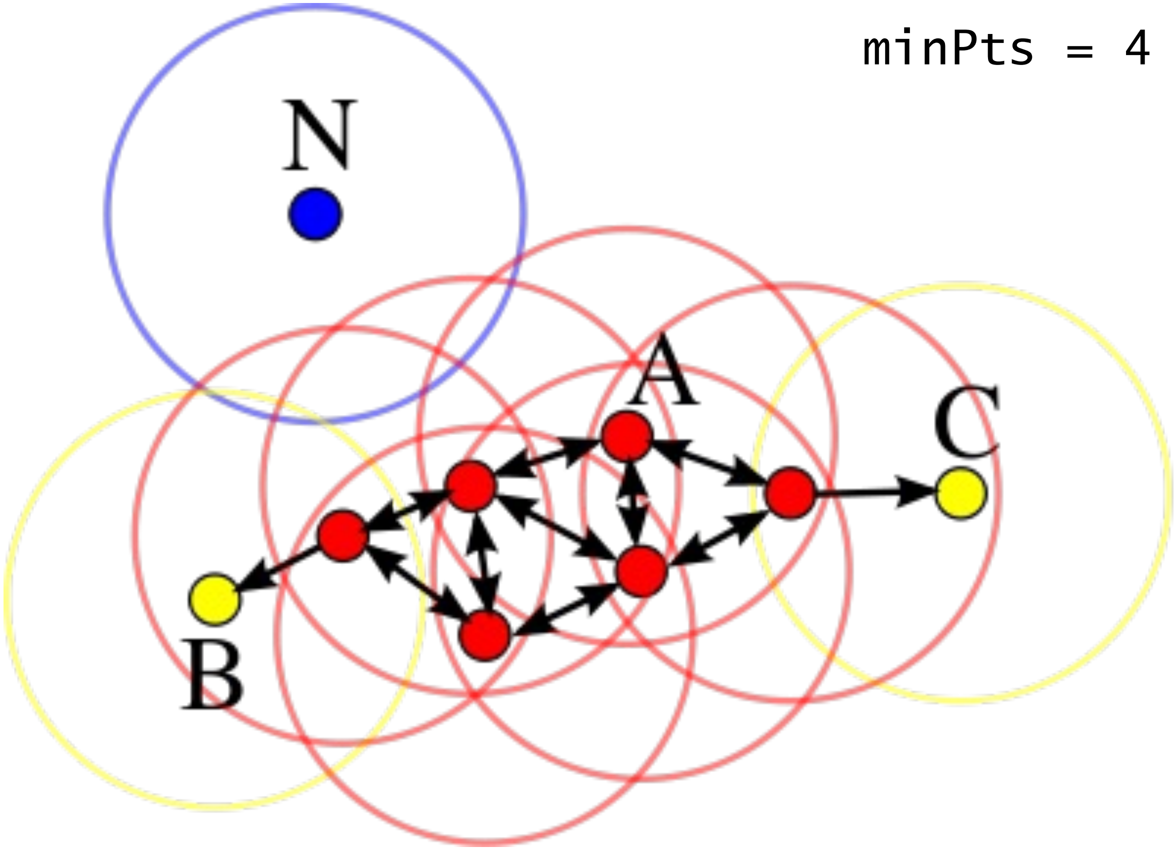
- DBSCAN hyper-parameters
    - EPS: maximum distance between two neighboring points
    - minPts: minimum number of points in a cluster

DBSCAN…

Types of points

- Core point
  - Point that has at least the minimum number of points (minPts) around it (including itself)
- Border point
  - Not a Core point
  - Is reachable by a Core point
- Outlier
  - Not a Core point
  - Not reachable by a Core point

minPts = 4



https://upload.wikimedia.org/wikipedia/commons/thumb/a/af/DBSCAN-Illustration.svg/800px-DBSCAN-Illustration.svg.png

DBSCAN…

Algorithm

1. Define all the Core points
   1. Visit each point and check if it is a Core point

2. Randomly select a Core point.
   1. Assign it to a new cluster.
   2. All Core points surrounding the initial Core point are added to the same new cluster.
   3. Continue adding **only Core points,** until all adjacent Core points have been added to the new cluster

3. Select all non-Core points that are close to the new cluster (Border points), and add them to the same new cluster
   1. When deciding if a non-Core point is close to the new cluster, the non-Core point **must be close to a Core point** (member of the new cluster), and not be part of another cluster

4. Randomly select a Core point that is not part of a cluster
   4. Repeat steps 2 and 3 to create another new cluster

5. Repeat step 4 until all Core points have been assigned a cluster

6. The remaining non-Core points are considered Outlier points

DBSCAN…

Advantages

- Number of clusters does not need to be defined
- Good for arbitrary cluster shapes
- Able to detect outliers

Disadvantages

- Function to define distance between points can be difficult to select
  - Requires domain knowledge
- Sensitive to density
  - Distance of points in a cluster varies for each expected cluster
  - Algorithm dependent on minimum distance and number of neighbors

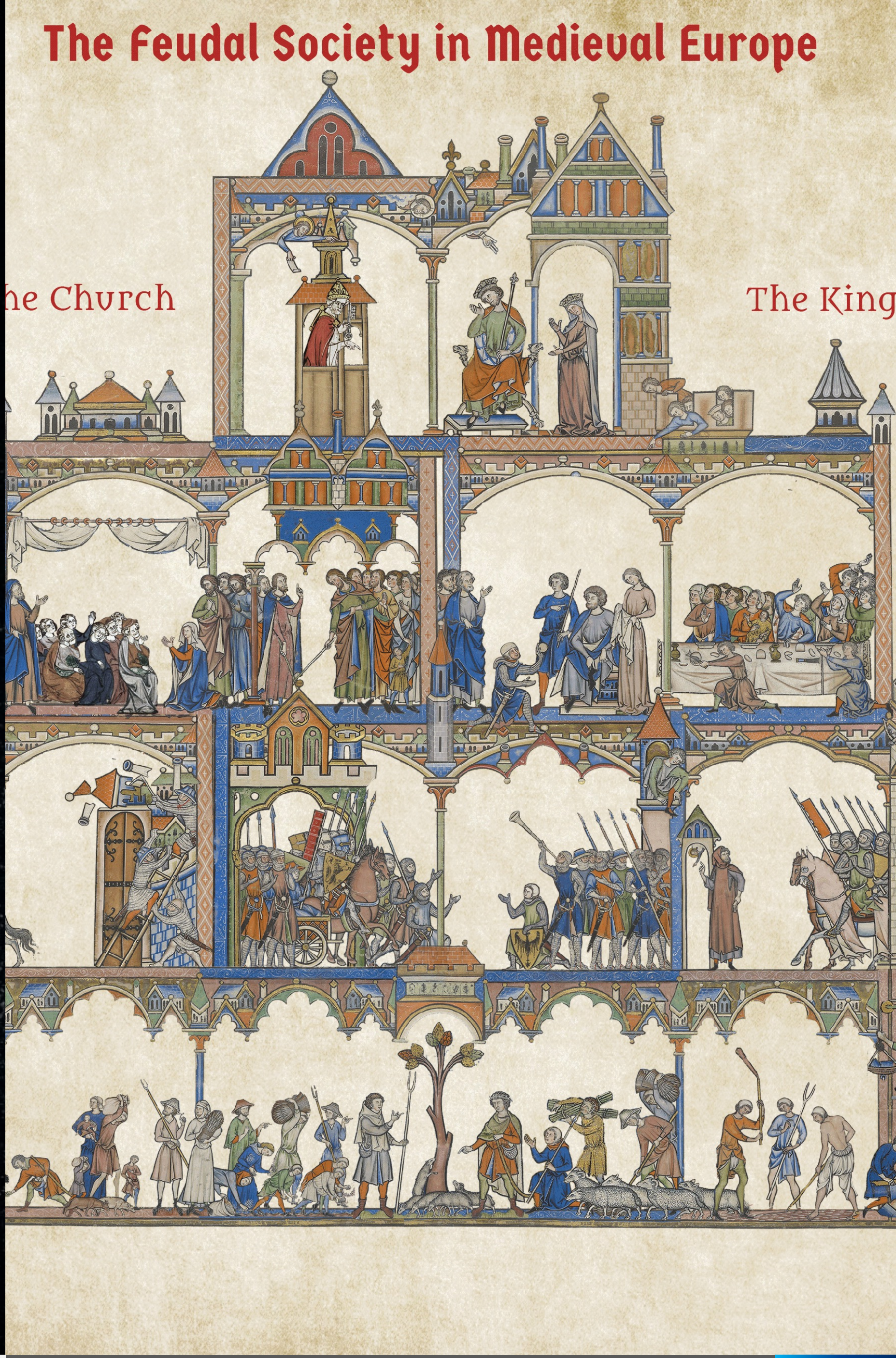
DBSCAN…

Class Example

- DBSCAN Example.ipynb

- Check Github

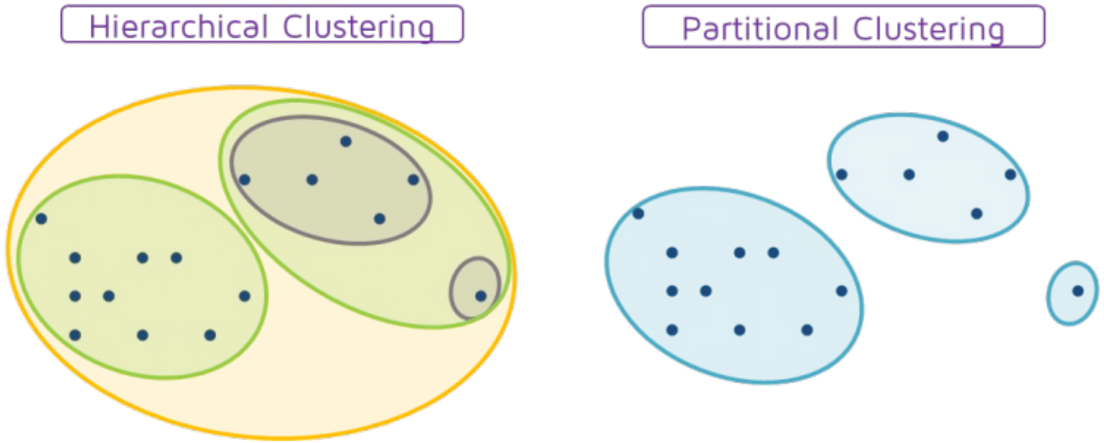
Hierarchical Clustering



Hierarchical clustering

Definition

- Method for cluster analysis
- Similar instances are cluster together
- Partitional clustering (i.e. K-means) divide data in non-overlapping clusters
  - Each instance is only in one cluster
- Hierarchical clustering allows instances to be in one or more clusters
  - Each cluster is a union of sub-clusters
  - The minimal cluster has only one instance



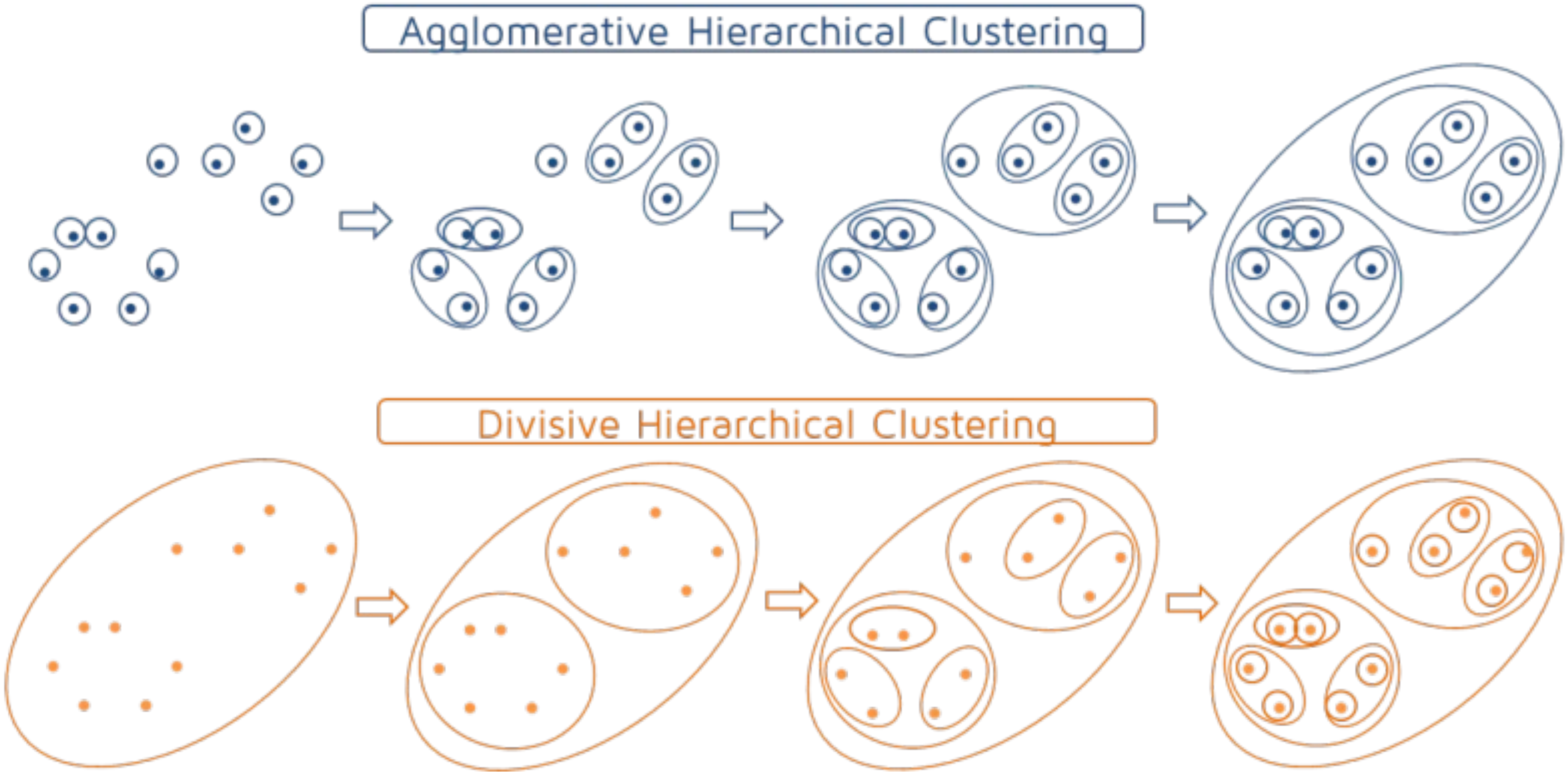
https://quantdare.com/hierarchical-clustering/#:~:text=the%20hierarchical%20clustering.-,One,-of%20the%20differences

Hierarchical clustering…

Two types

Agglomerative clustering (bottom-up)

Divisive clustering (top-down)



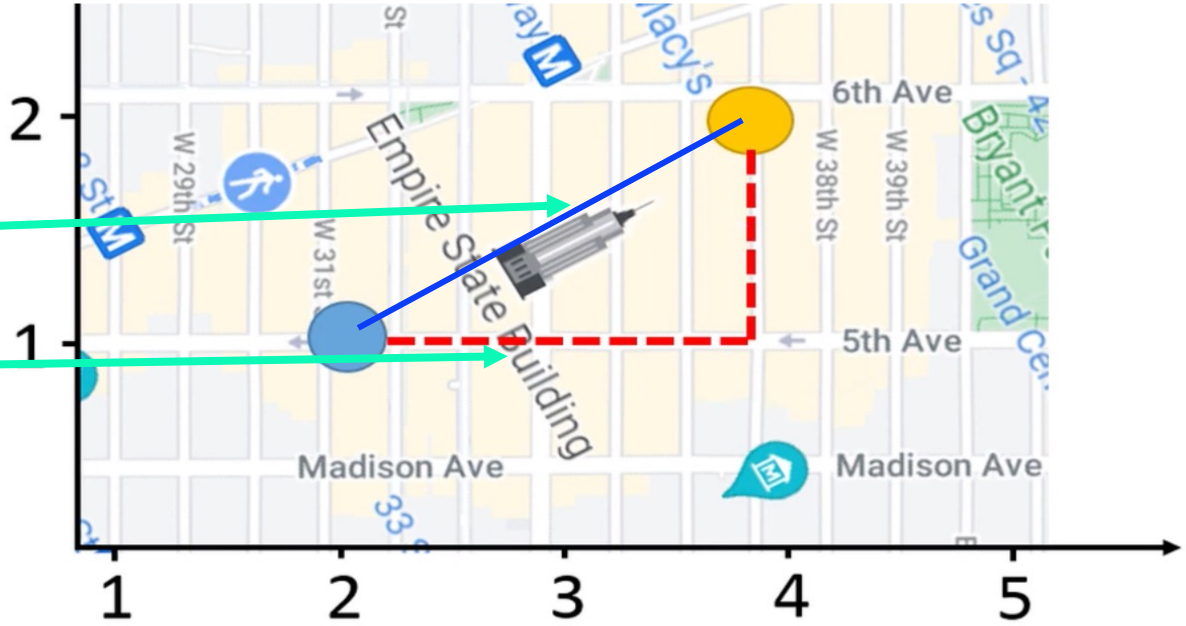
https://quantdare.com/wp-content/uploads/2016/06/AggloDivHierarClustering.png

Hierarchical clustering…

Distance between instances

- Hierarchical clustering needs to use a function to define the distance between two instances in a dataset

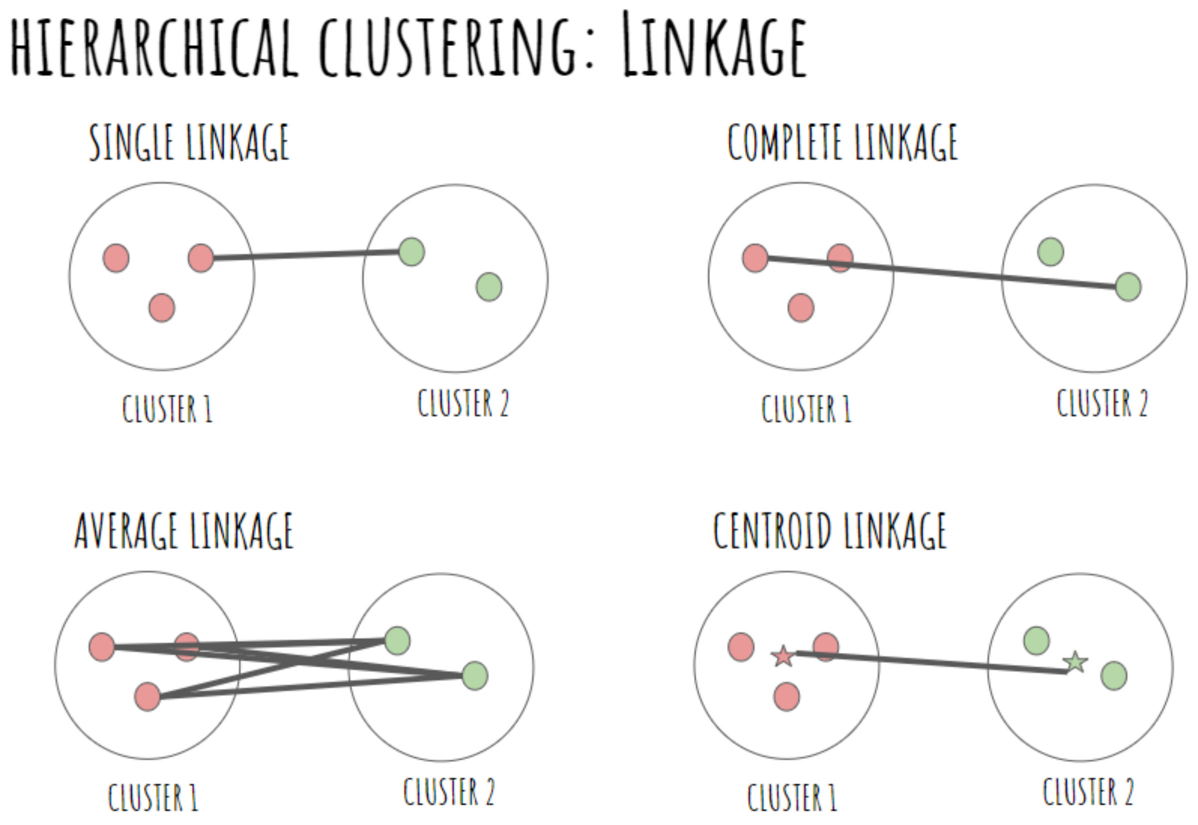
- The two most commonly used are
  - Euclidian distance
  - Manhattan distance



Hierarchical clustering…

Linkage metrics

- Define how “close” a cluster is from another cluster
- Single-linkage (nearest neighbor)
  - Shortest distance between a pair of instances in two clusters
- Complete-linkage (farthest neighbor)
  - Distance between the two farthest instances in two clusters
  - Very used in practice
- Average-linkage
  - Distance between each pair of observations in each cluster are averaged
  - Very used in practice
- Centroid-linkage
  - Distance between centroids of two clusters

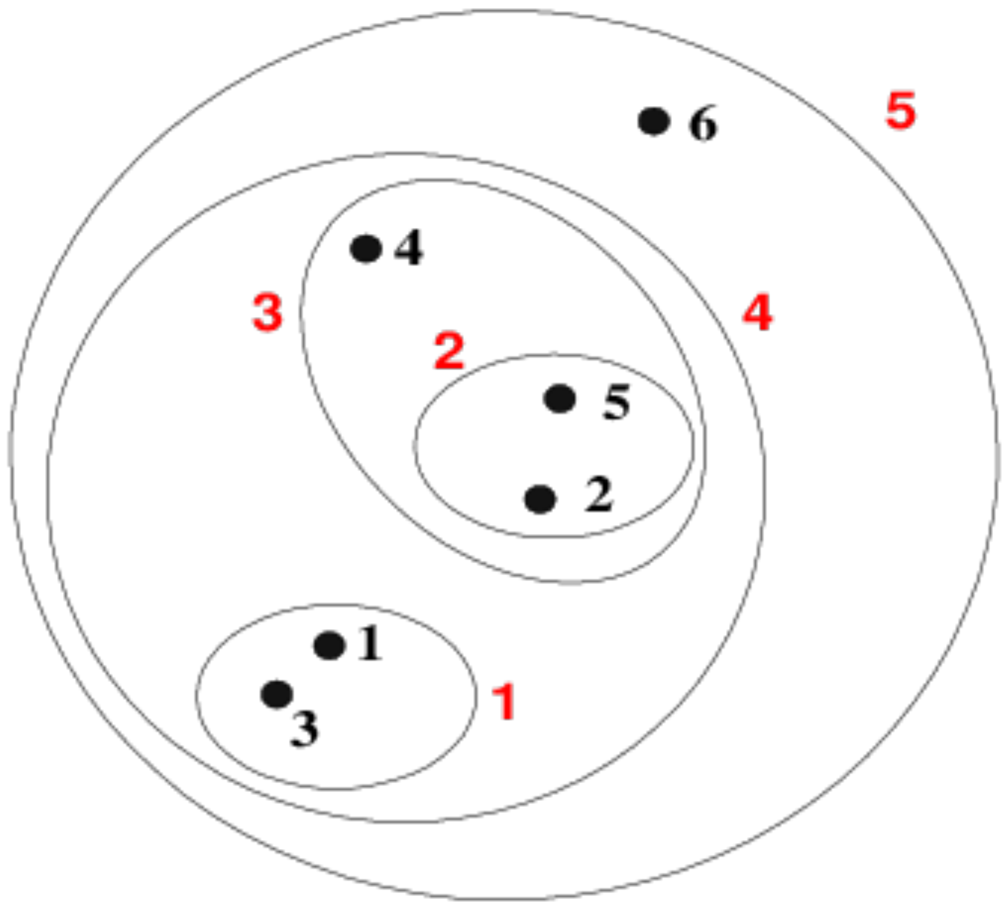


https://miro.medium.com/max/1400/1*vv5YongFHeRldYa2bujSPg.png

Hierarchical clustering…

Agglomerative

- Procedure
  1) Each instance is its own cluster
  2) Calculate the similarity (distance) between all clusters
  3) Joint together the two closest clusters into one single cluster
  4) Repeat step 2 until all clusters have been joined into one single cluster



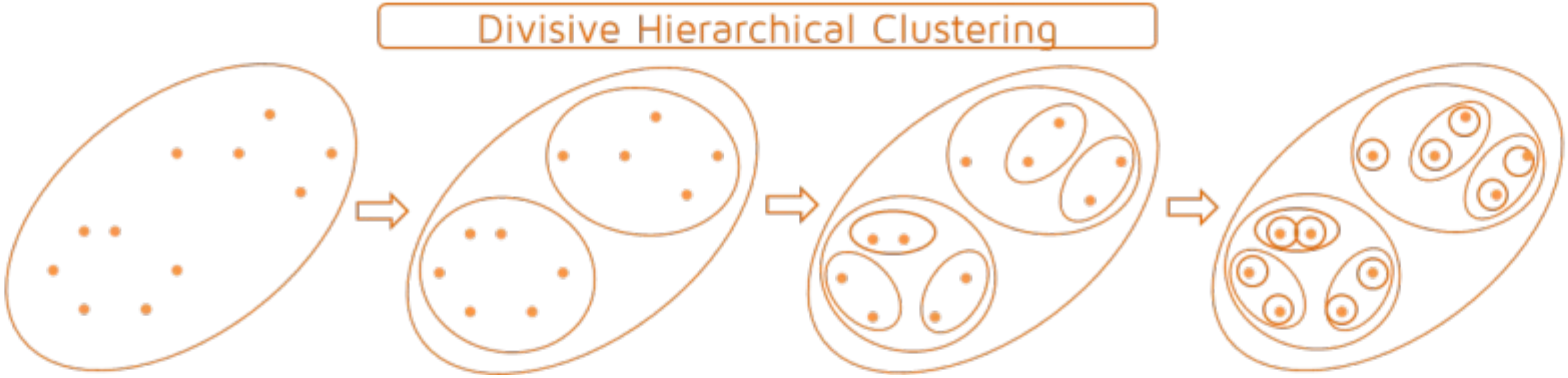
Hierarchical clustering…

Divisive

- Procedure
  1) Complete dataset starts as one single cluster
  2) Divide in two the cluster using a similarity (distance) function
  3) Repeat step 2 until all clusters contain only one instance

- This type of hierarchical clustering is very rarely used

- The most common implementation uses K-means as the divisive method



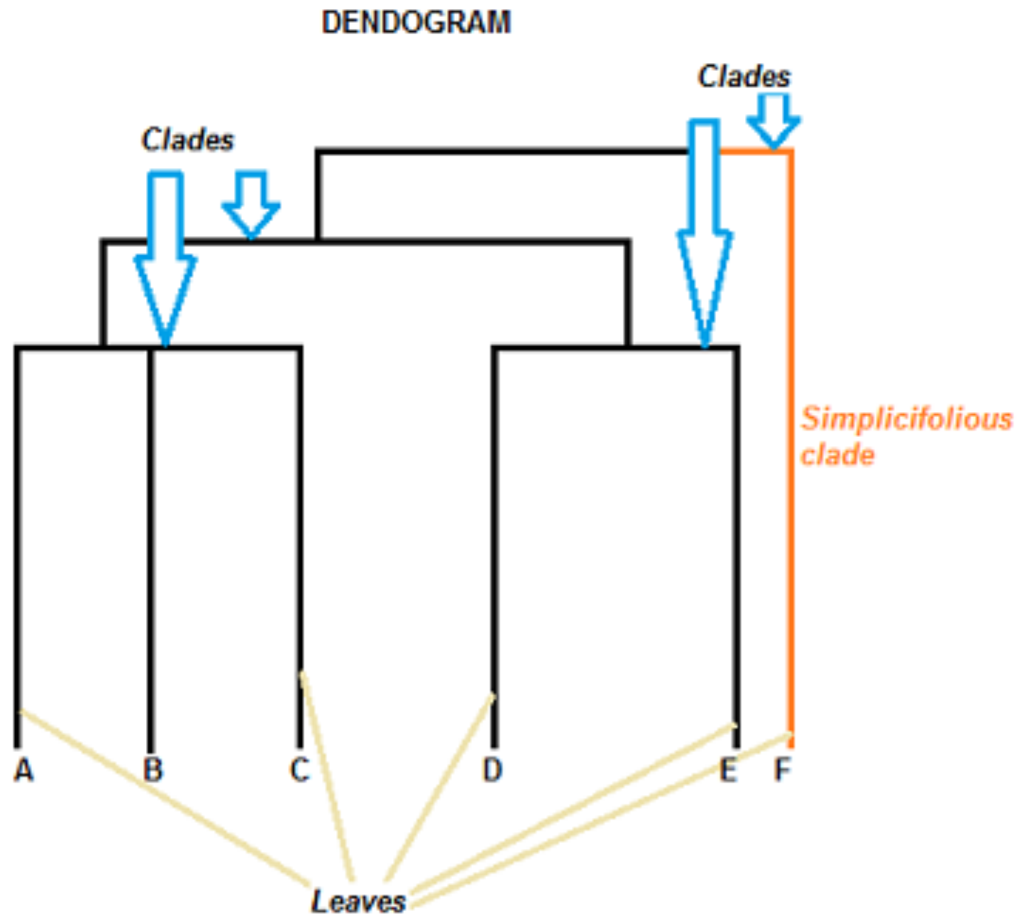
Hierarchical clustering…

Dendrograms

Hierarchical clustering…

Dendrograms

- Kind of tree diagram
- Shows hierarchical relationships between sets of data
- Frequently used in biology
- Can represent any type of data
- Usually represented as a row or column graph

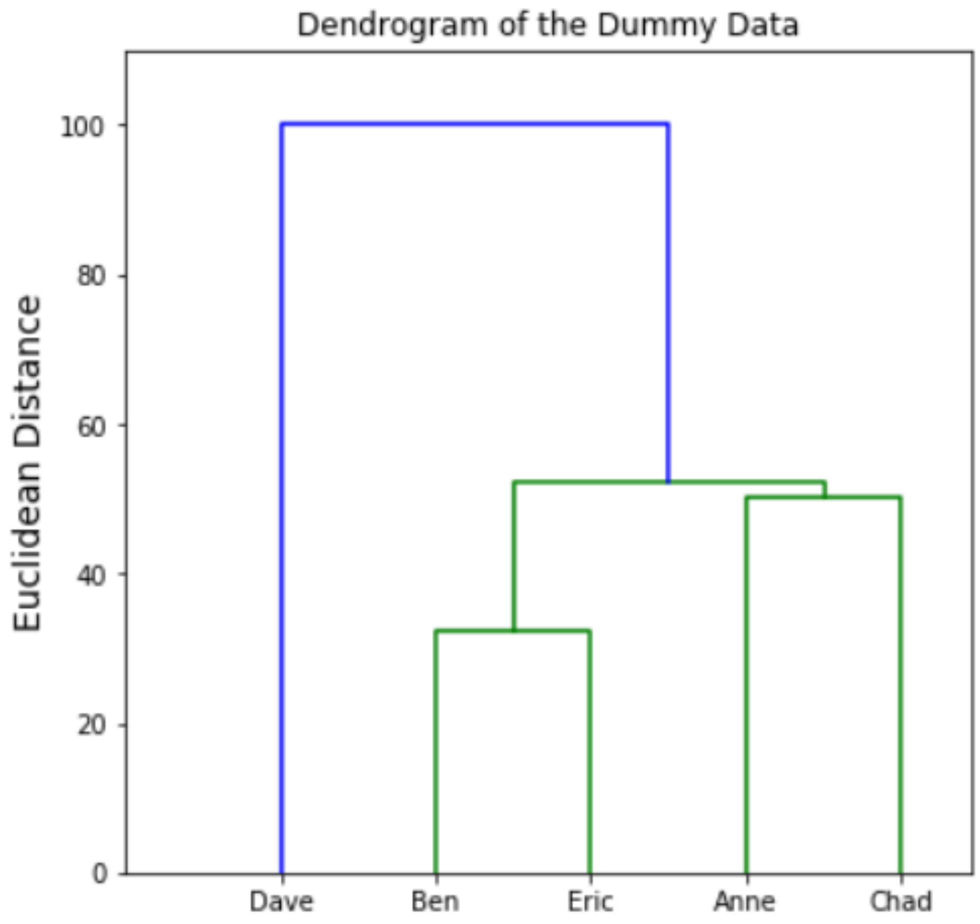


https://www.statisticshowto.com/wp-content/uploads/2016/11/dendrogram.png

Hierarchical clustering…

Dendograms…

- How to read a dendogram
  - Clades are organized by similarity
  - The closer (same height) clades are to each other, the more similar they are
  - The closer clades are to the bottom of the graph, the more similar their two elements are between each other



https://miro.medium.com/max/1048/1*pp17u6Li2D_LzIbiZumgHQ.png

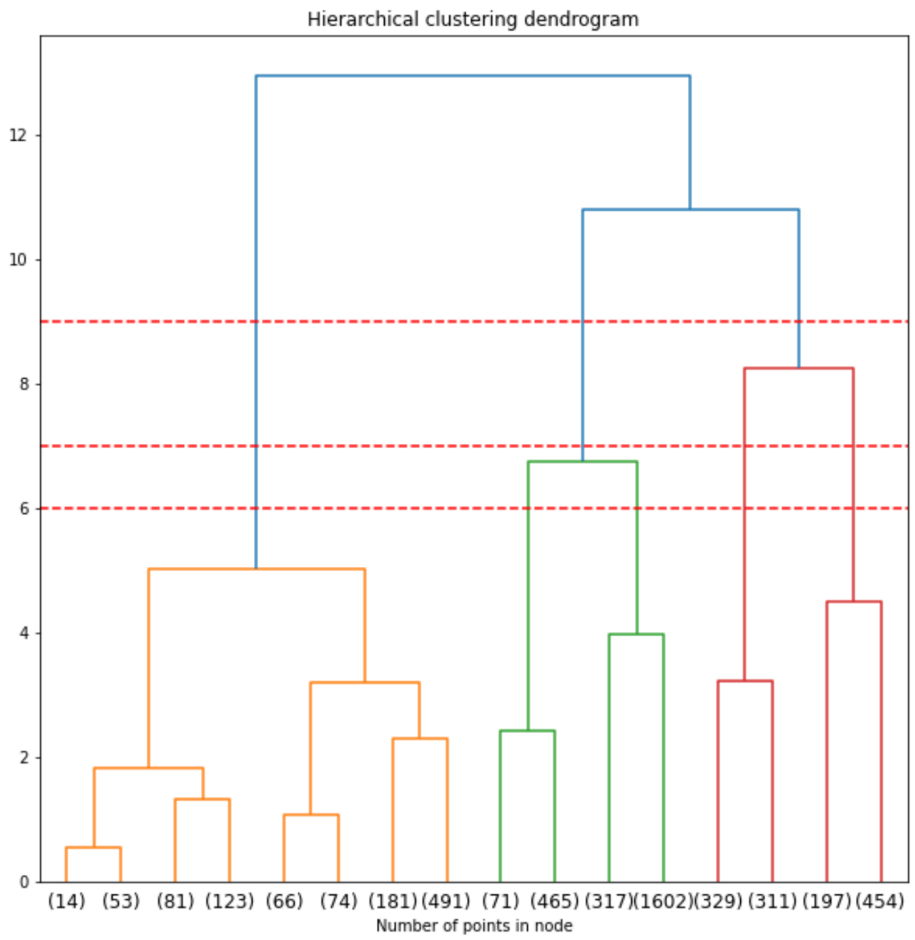
Hierarchical clustering…

Dendograms…

- Selecting Cut-points
  - A Cut-point can be selected any where in the *y* axis
    - Depends on the similarity that best fits a problem

  - The Cut-point helps to define the number of clusters in the dataset
    - The biggest clades below the cut-point are the clusters
      - Number of vertical lines crossed by the cut-point

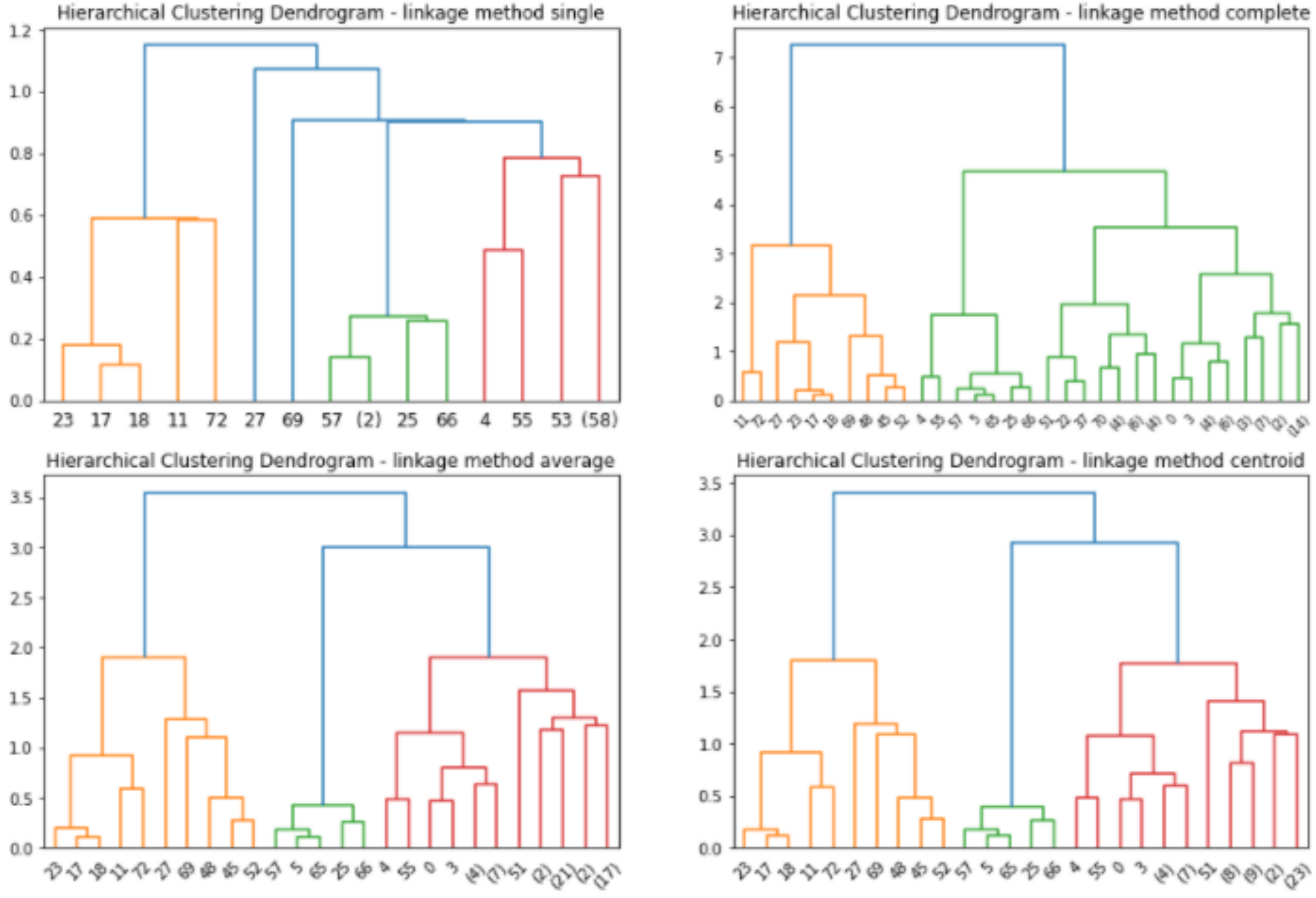
    - In the image:
      - *Cut-point* = 6 defines 5 clusters
      - *Cut-point* = 7 defines 4 clusters
      - *Cut-point* = 9 defines 3 clusters



https://towardsdatascience.com/a-practical-introduction-to-hierarchical-clustering-from-scikit-learn-ffaf8ee2670c

Hierarchical clustering…

Effect of linkage metrics in clustering



https://levelup.gitconnected.com/evaluating-socio-economical-indicators-with-exploratory-analysis-and-hierarchical-clustering-af1575782141

Hierarchical clustering…

Class example

- Program
  - Hierarchical Clustering Example.ipynb

- Check Github

References

- Russell, S. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Third Edition. Prentice Hall.

- Mitchell, T. (1997). Machine Learning. McGraw-Hill.

- Creator of the presentation design: Pedro Armijo. MS Office theme.