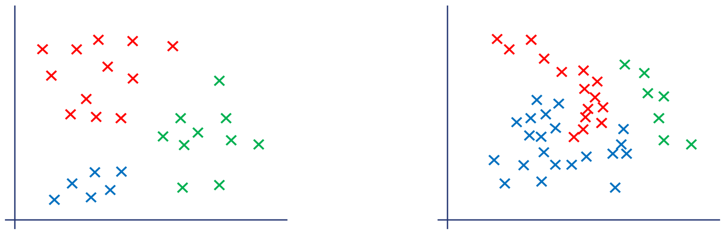
FH KREMS
UNIVERSITY OF APPLIED
SCIENCES / AUSTRIA

Machine Learning, Artificial Intelligence, and Big Data Analytics (IL, 4th Semester)

Lecture 8

Summary of previous lecture (1/4)

- Clustering: Finding natural groupings among objects in a dataset



- Maximize similarity within clusters (intra-cluster): *cohesive* within clusters
- Minimize similarity between clusters (inter-cluster): *distinctive* between cluster

Summary of previous lecture (2/4)

1. Formulate the problem
2. Select a distance/dissimilarity measure
3. Select a clustering technique
4. Decide on the number of clusters
5. Interpret the clusters
6. Assess the clustering

- Given a set of <…>, I want to find <…>, based on <…>
- Euclidean, Manhattan, Cosine, Jaccard, …
- K-means, PAM, Agnes, Diana, DBScan, …
- Using Silhouette, Elbow, Dendograms, …

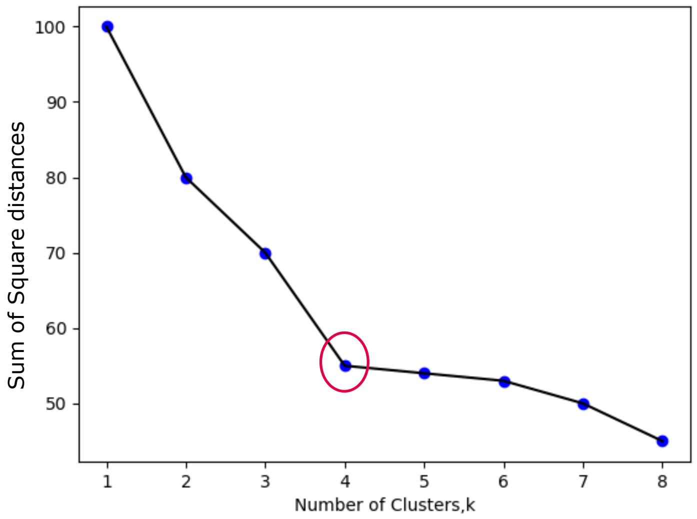
Summary of previous lecture (3/4)

- Partitioning techniques:
  - Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters,
- K-means:
  - Random initialization
  - Repeat until converge
    - Cluster assignment step
    - Centroid update step
- Variants:
  - K-median
  - K-medoid

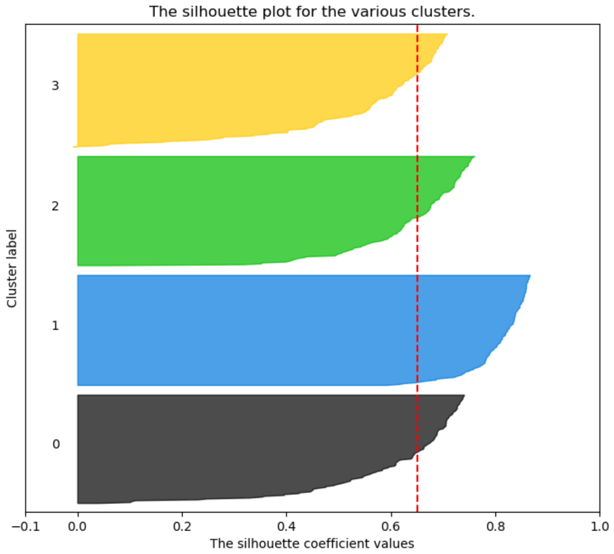
Summary of previous lecture (4/4)

Validating clustering with intrinsic measures

- The “Elbow method”: Find the elbow in the decrease of SSE



- The “Silhouette Score”: a generic measure of cluster cohesiveness and distinction

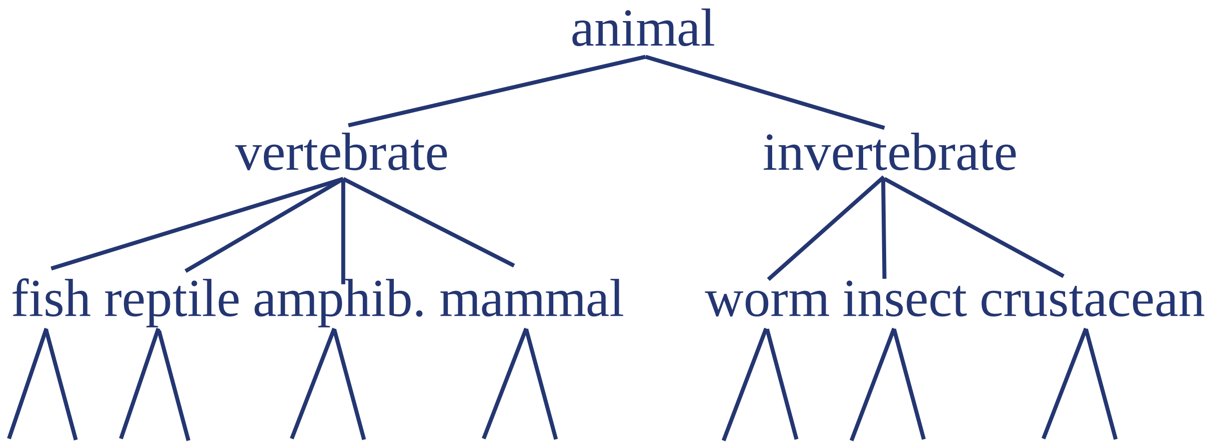


Agenda

- Hierarchical Clustering
  - Agglomerative
  - Divisive
- Linkage
- Interpreting a Dendrogram

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy from a set of observation



Two types of hierarchical clustering

- Agglomerative (bottom-up):

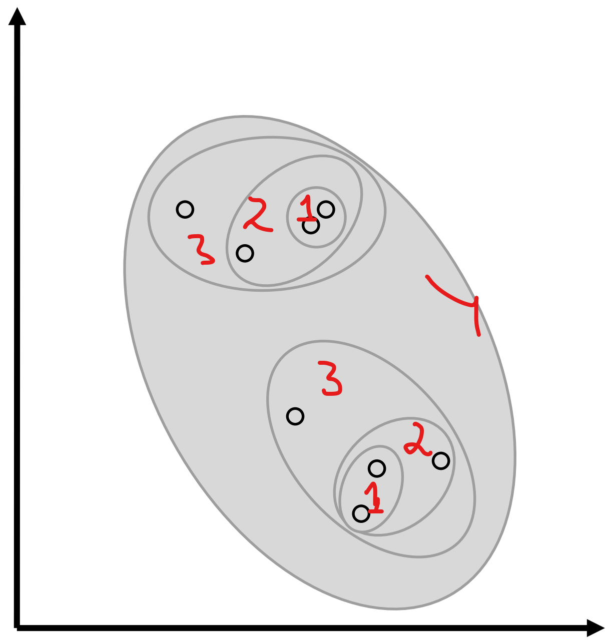
  merging single unit clusters into larger clsters

  - Starting with each item in its own cluster (singletons)
  - Find the best pair to merge into a new cluster (the points that are closest to each other)
  - Repeat until all clusters are fused together
  - It is the most common approach.
- Divisive (top-down):

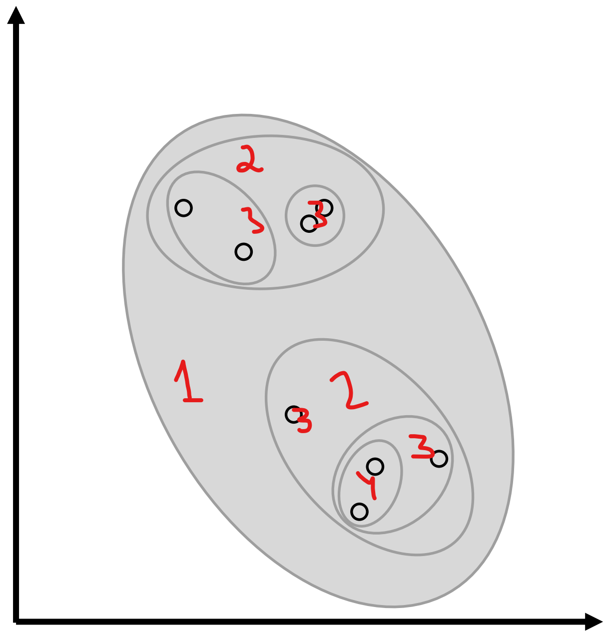
  start from the large unit and split

  - Starting with all the data in a single cluster
  - Consider every possible way to divide the cluster into two.
  - Choose the best division and recursively operate on both sides until singleton sets are reached.

Agglomerative vs. Divisive



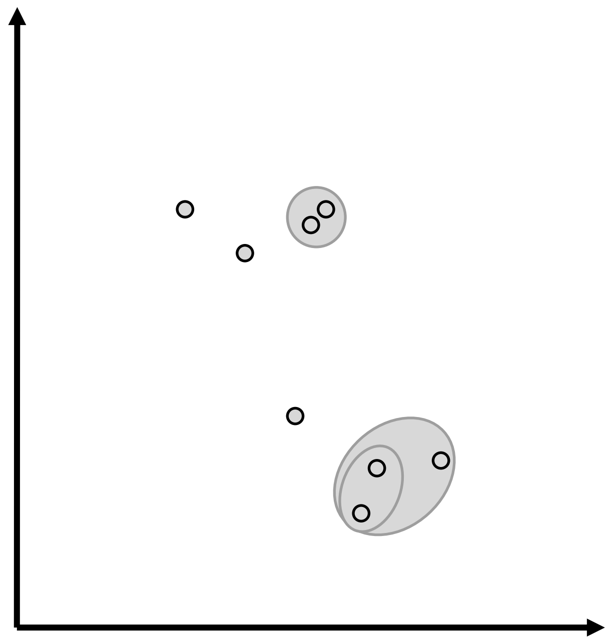
Agglomerative



Divisive

Disclaimer: Only for illustrative purpose. The points are just drawn by hand and splits/merges were done based on my perception of distance

Agglomerative vs. Divisive



Agglomerative

linkage - the concept how to measure distance

Let's go back to step 3 of agglomerative

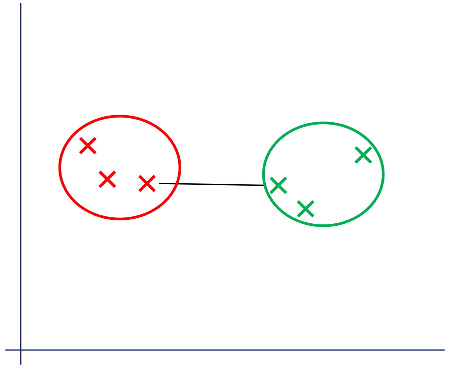
How did we decide what to merge at this step?

We know how to measure the distance between two points, but what about the distance between clusters or between points and clusters?

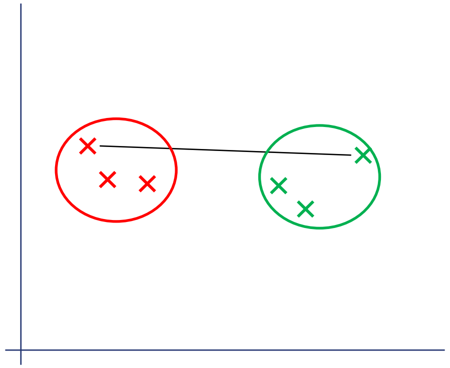
We need to introduce the concept of linkage

Linkage measures

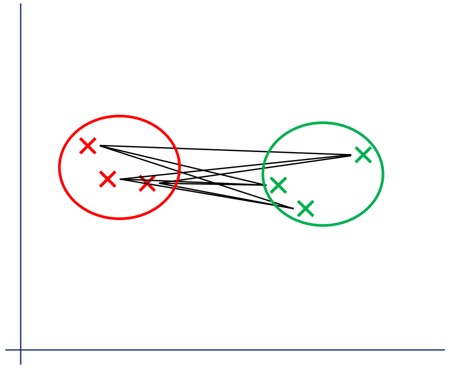
How to measure the distance between two clusters



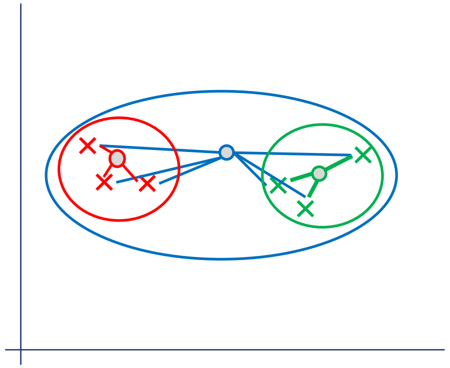
- **Single linkage**: the distance between two clusters is the distance of the two closest objects in the different clusters



- **Complete linkage:** the distance between two clusters is the distance of the two furthest objects in the different clusters



- **Group Average linkage:** the distance between two clusters is the average distance between all pairs of objects in the two different clusters

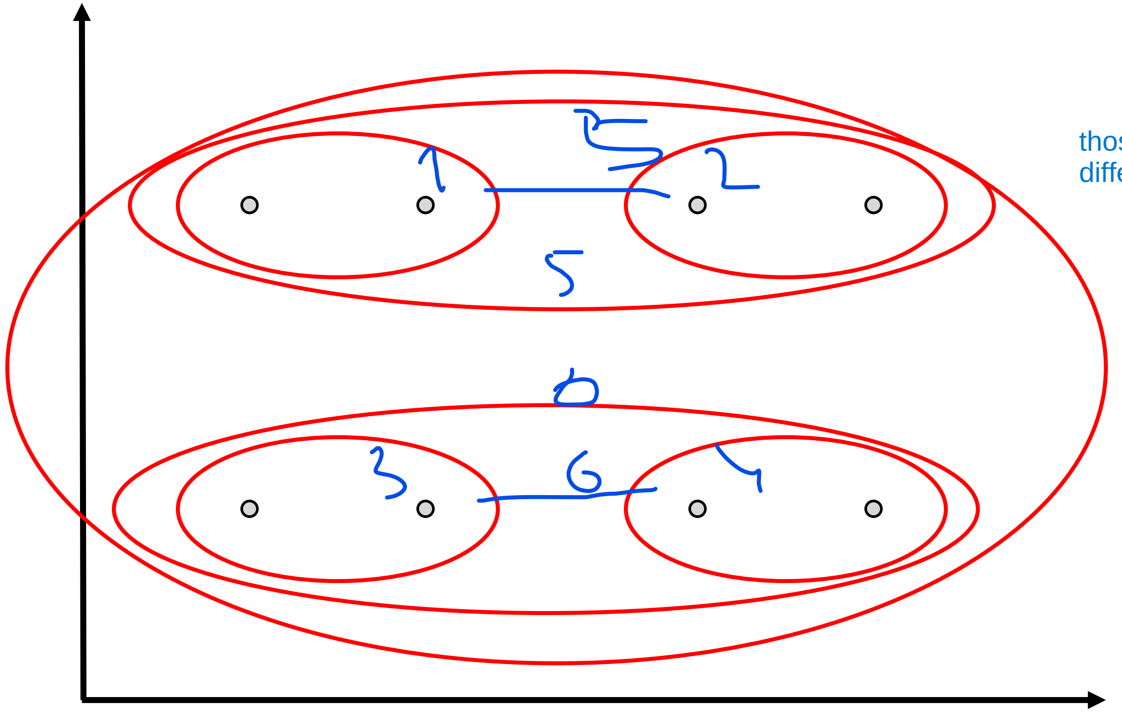


- **Ward linkage:** if you merge two clusters, how does it change the total distance from centroids.

works better than others

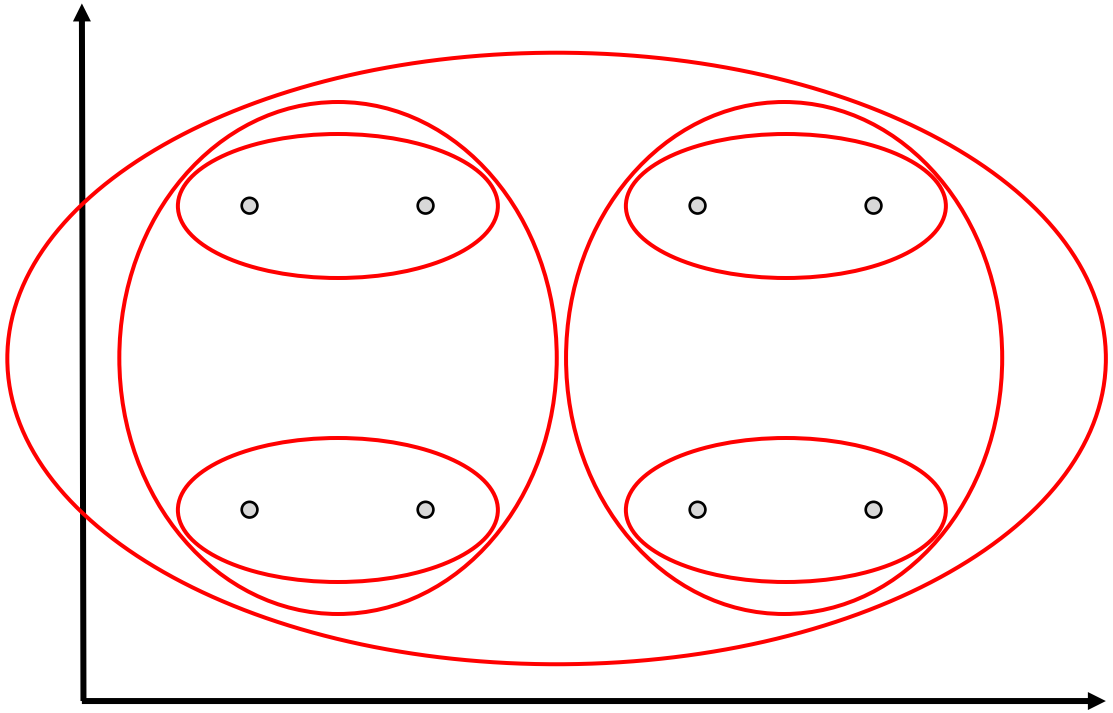
Example – Single Linkage

First group them in pairs
then



those shrtest distance between the closest points of different clusters first group them in piars

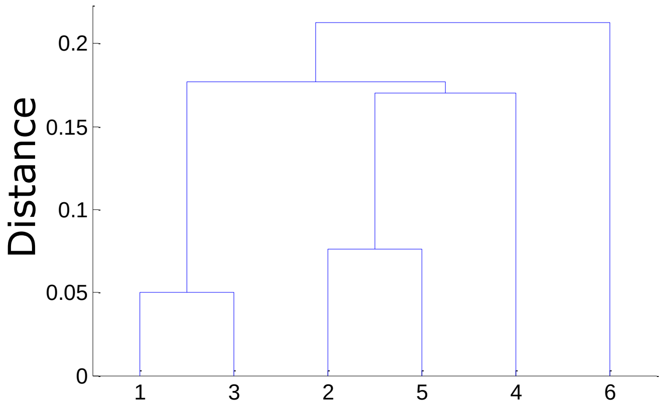
Example – Complete Linkage

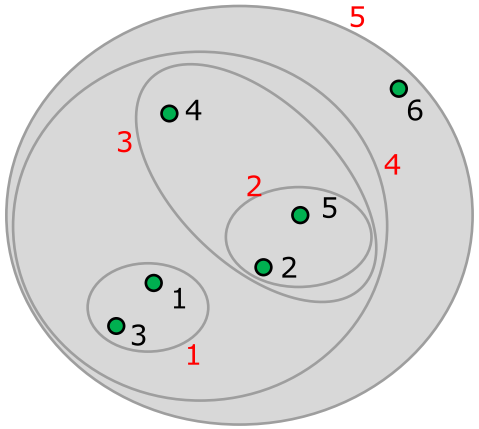


take the shortest distance in between
the fatrhest distances

The Dendrogram

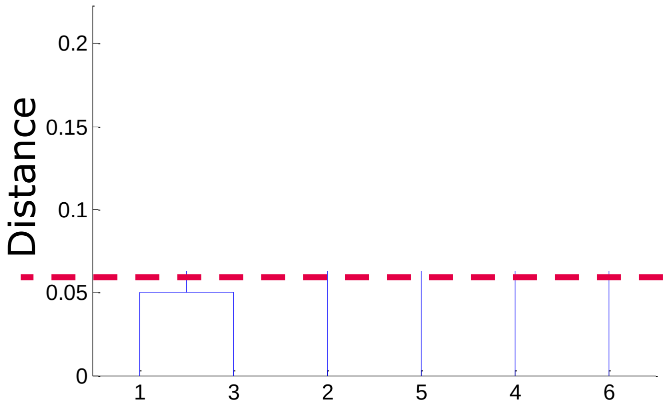
- A graphical representation displaying clustering results.
  - It records the sequences of merges and splits
  - horizontal lines represent clusters that are joined together.
  - The position of the line on the vertical axis indicates distances where clusters were joined.

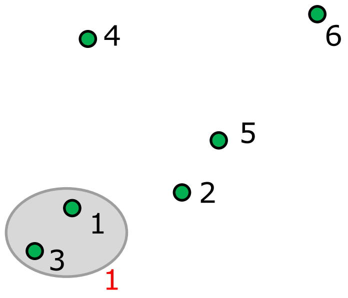




The Dendrogram

- Placing a horizontal line at a specific distance gives us the clustering level
- The connected points are clustered

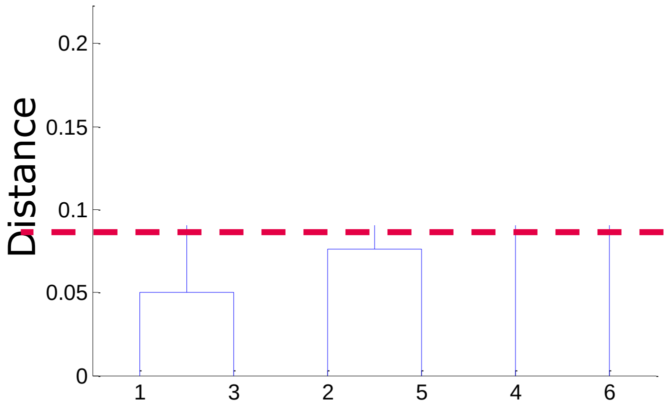


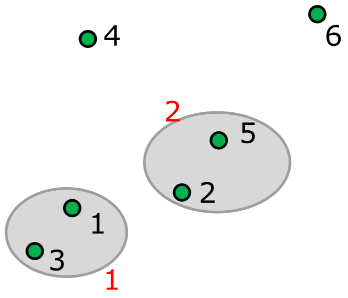


The Dendrogram

- Placing a horizontal line at a specific distance gives us the clustering level
- The connected points are clustered

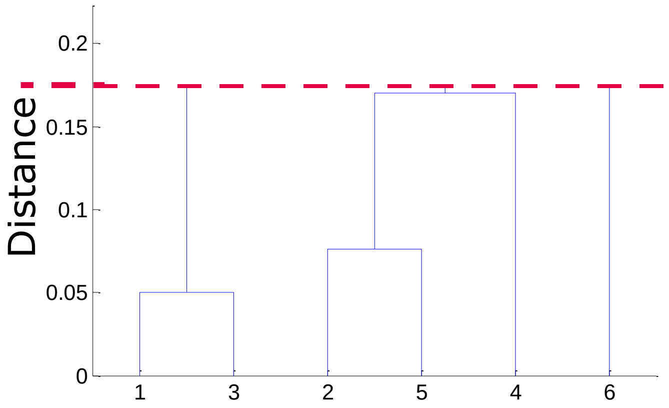
here this is the best

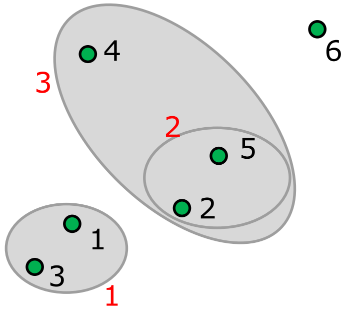




The Dendrogram

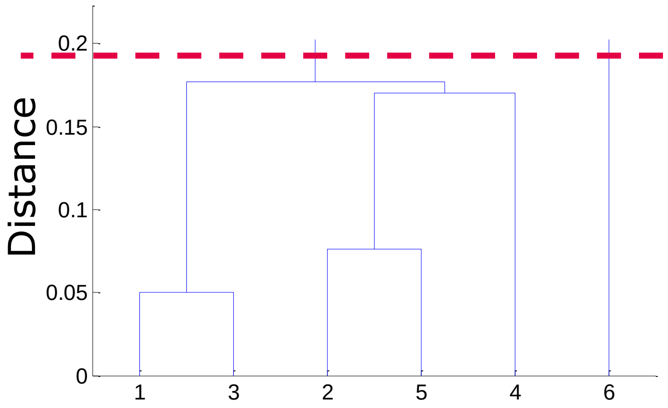
- Placing a horizontal line at a specific distance gives us the clustering level
- The connected points are clustered

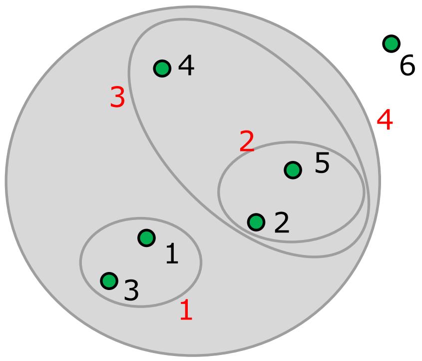




The Dendrogram

- Placing a horizontal line at a specific distance gives us the clustering level
- The connected points are clustered

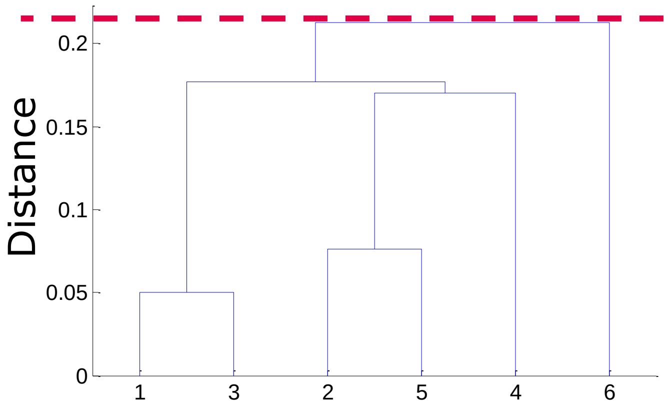


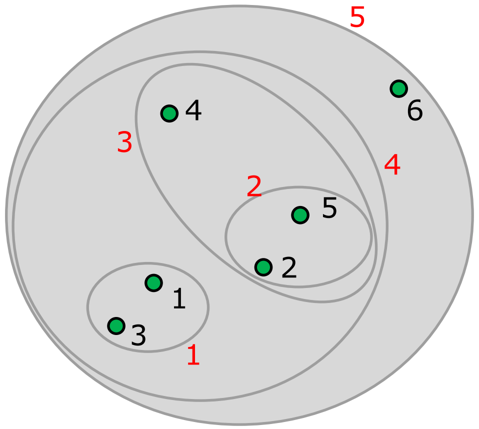


The Dendrogram

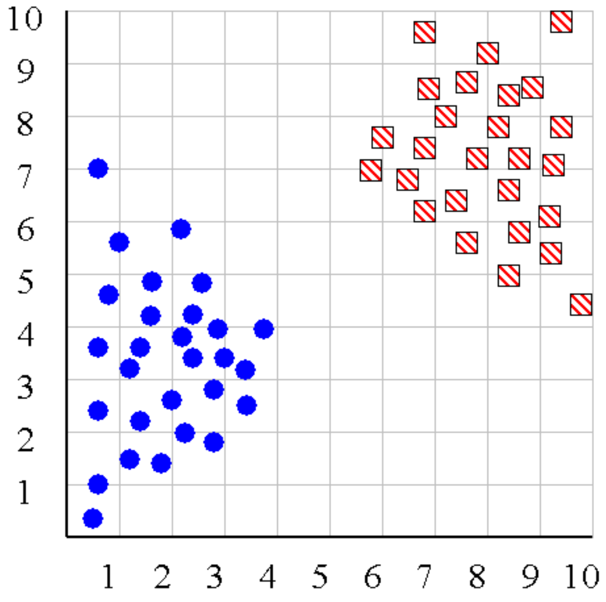
cut at first big distance jump

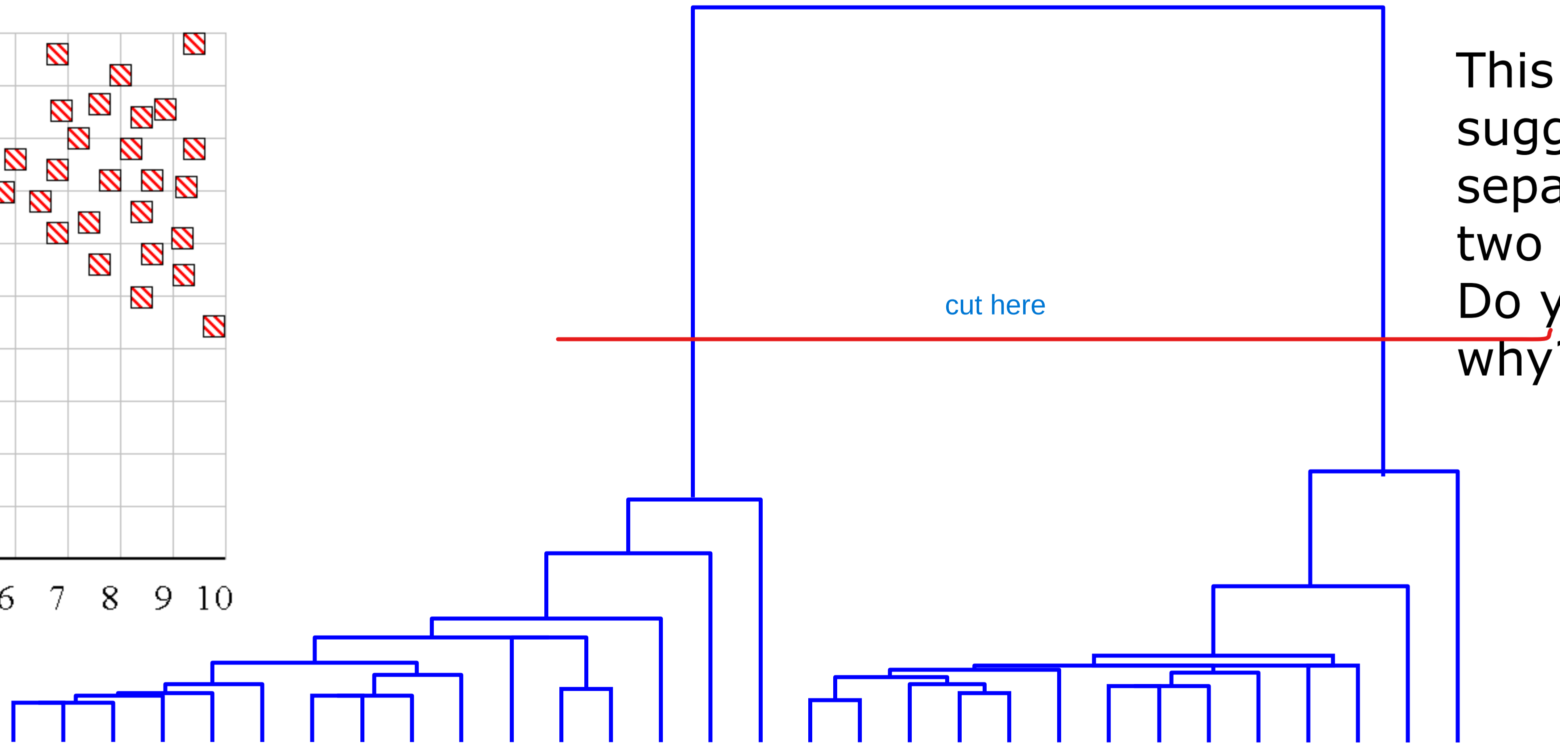
- Placing a horizontal line at a specific distance gives us the clustering level
- The connected points are clustered





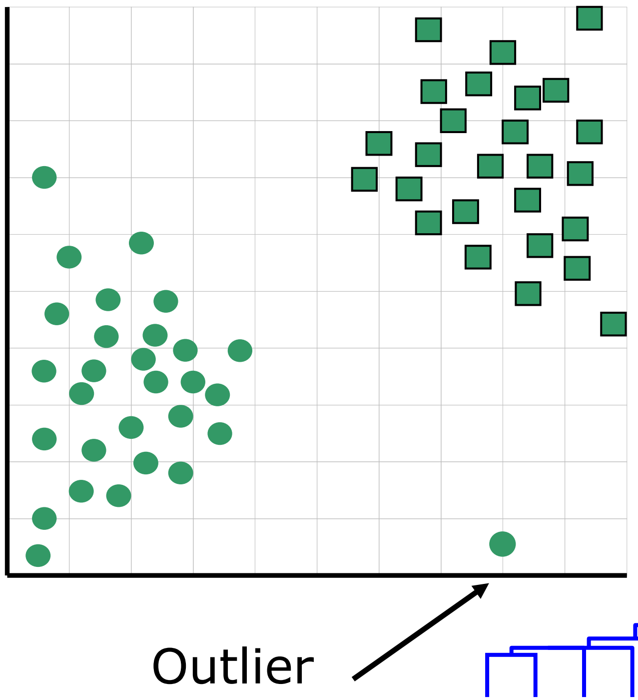
Interpreting a Dendrogram

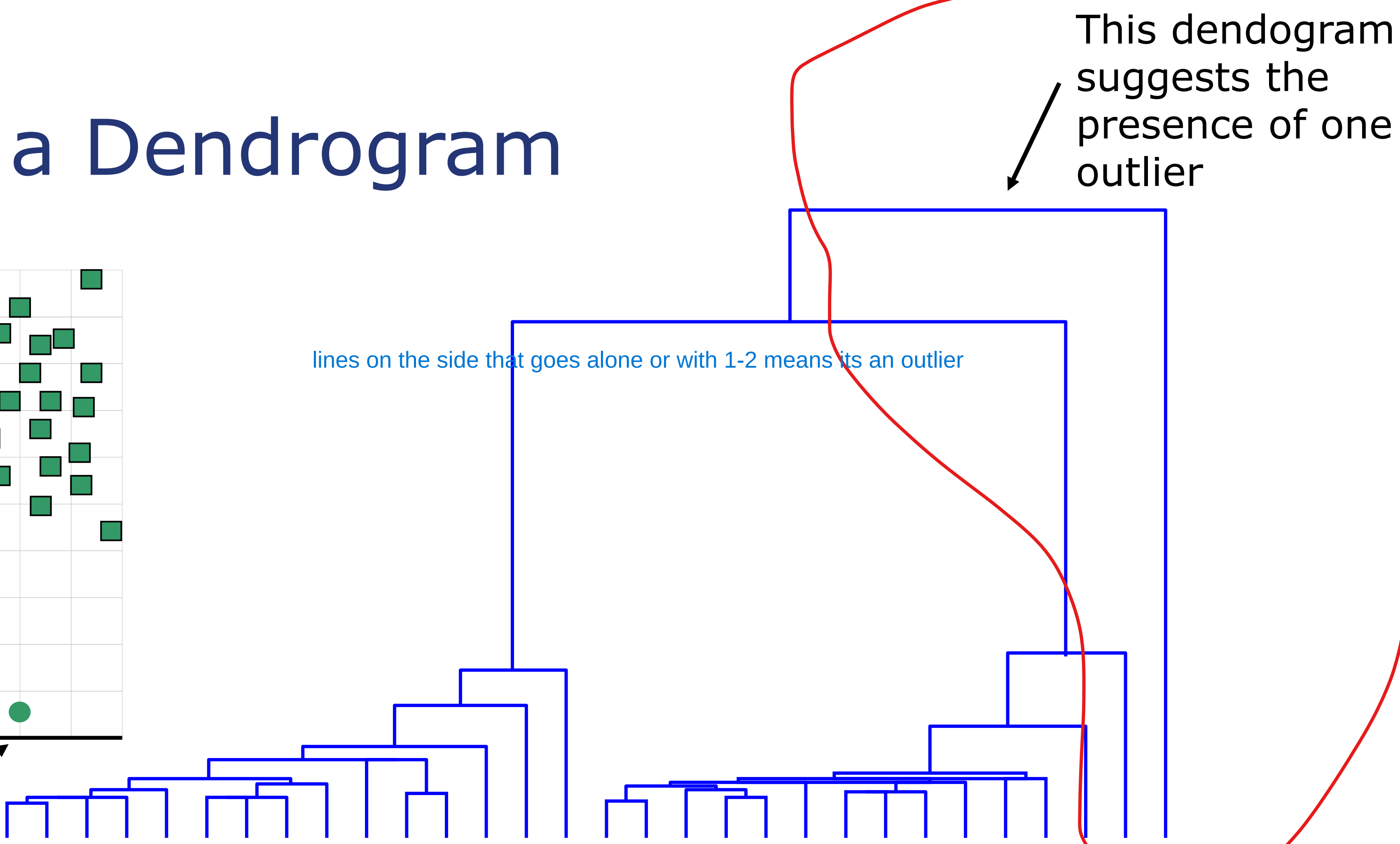




This dendogram suggests a clear separation in two clusters. Do you know why?

Interpreting a Dendrogram





lines on the side that goes alone or with 1-2 means its an outlier

Considerations on the linkage

- *Single Linkage* is sensitive to noise and outliers, but can handle clusters of different dimensions
- *Complete Linkage* is less sensitive to noise but tends to break large clusters and results in clusters of the same diameter.
- *Average-Group* and *Ward* are good trade-offs

Considerations on similarity/dissimilarity

Difference between dissimilarity and distance

- Dissimilarity and distance are often used interchangeably, but this is wrong!
- A dissimilarity metric to be a proper distance function should respect following criteria:
    - Symmetry: $d(i,j) = d(j,i)$
    - Self-similarity: $d(i,i) = 0$
    - Reflexivity: $d(i,j) = 0$ IIf $i=j$
    - Triangular Inequality: $d(i,j) \leq d(i,k) + d(k,j)$
- For example, cosine dissimilarity is not a distance function!
- Clustering techniques do often use dissimilarities instead of distances

Exercise 4 - Clustering

- **Exercise 4.1**
- The dataset “wheat.csv” contains measurements of different cereal grains. Load it and:
    - Explorative Data Analysis
    - Compute the distance matrix of the observations
    - Hierarchical clustering
        - Plot a Dendrogram for different linkage strategies.
        - Choose the best linkage and an appropriate n of clusters.
        - Create the final clustering by “cutting the tree”
        - Visualize the cluster on its principal components
    - Clustering with K-Means
        - Evaluate K with the elbow method. Is it equal to HCL?
        - Visualize the clusters on its principal components

- **Exercise 4.2**
- Download hundreds of tweets* from a hashtag of your choice (anything that interests you) and:
    - Create a corpus
    - Clean-up corpus (remove irrelevant words)
    - Create a Document-Term-Matrix (DTM)
    - Reduce sparsity if necessary.
    - Compute dissimilarity matrix (using cosine)
    - Hierarchical clustering
        - Plot a Dendrogram for different linkage strategies.
        - Choose the best linkage and an appropriate n of clusters.
        - Create the final clustering by “cutting the tree”
        - Short manual exploration of the clusters

In R you can use the package “rtweet” (it requires a twitter account).