CLASSIFICATION

Alejandro Ordonez

Assistant Professor - Department of Bioscience

Section for Ecoinformatics & Biodiversity

Center for Biodiversity Dynamics in a Changing World (BIOCHANGE)

CLASSIFICATION
WHAT WE WILL TALK ABOUT TODAY

- What is classification – Patterns in a collection of objects.
- Distance Matrices – Measuring (d)similarity.
- Classification as a multivariate approach
- Hierarchical clustering – The tree based approach.
- Non-Hierarchical clustering – A single partition.

Any questions?

Ready to start?

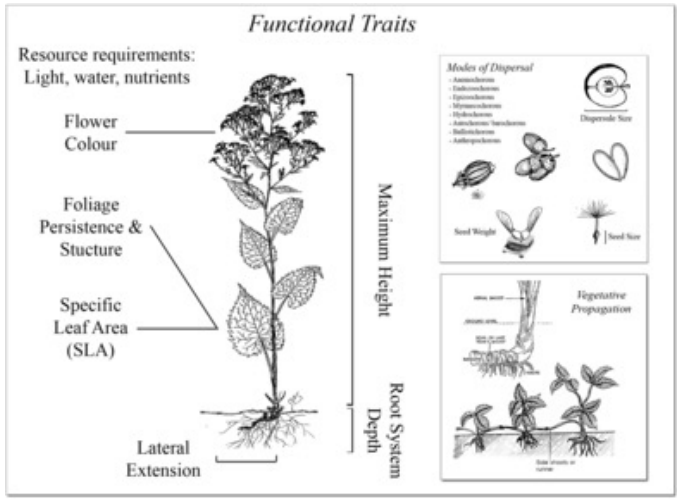
THE PROBLEM AT HAND

Biological/ecological datasets are multidimensional

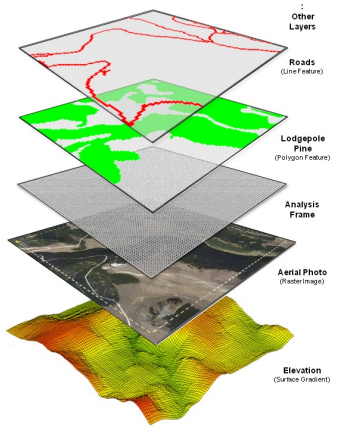
What is multidimensional? → multiple descriptors are recorded from a number of objects (i.e., replicate sampling or experimental units)

- **Organisms**: morphological or physiological measurements
- **Ecological sampling units**: the variables might be physicochemical measurements or species abundances

Organism attributes



Ecological sampling units



REGRESSION VS. MULTIVARIATE

The goal of a regression:

- **Test** for a **relation** between descriptors and response variables among objects
    - Test a H_o: is $\beta_i=0$

The goal of classification:

- **Visualise** the **association** between objects
    - **Clustering**: Group the observations.
    - **Ordination**: Reposition the observations in less dimensions

In a strict sense, classification analyses are not statistical test or models

Because no Hypothesis is tested.

CLASSIFICATION – THE GOAL

The objective

Recognize discontinuous subsets in an environment that is sometimes discrete (as in taxonomy), but most often perceived as continuous in ecology.

How?

Using the association between elements in a collection of objects, these are partition based on a set of criteria/rules

Hard partition: each observation belongs to one and only one subset.

Fuzzy partitions: the membership of each observation is continuous.

CLASSIFICATION
WHY GO TO ALL THE TROUBLE?

Classification is an way to simplify the description of the data by means of groups

What classification does?

- Group together a number of objects based on their attributes or variables
- Produces groups of objects where each object within a group is more similar to other objects in that group than to objects in other groups.

Classification is a traditional tool of GIS analysis (land cover classes), and several informal subjective systems

CLASSIFICATION - CLASSES

Hierarchic methods have several nested levels of classification.

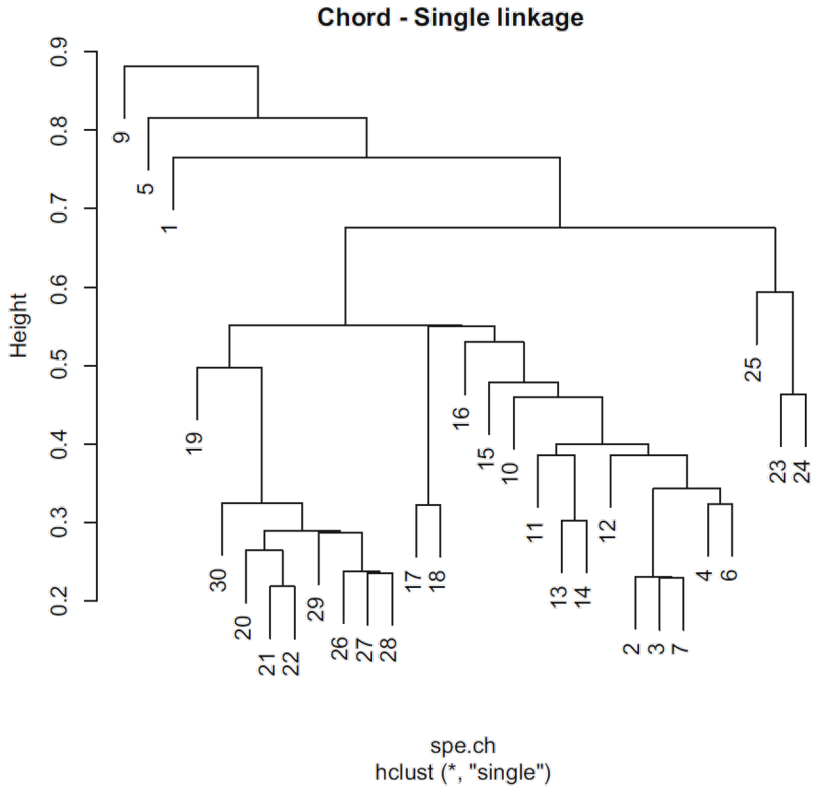
- **Agglomerative** methods combine sampling units to classes and classes to classes.
- **Divisive** methods split data and then split the splits.

Non-hierarchic methods perform classification at one level.

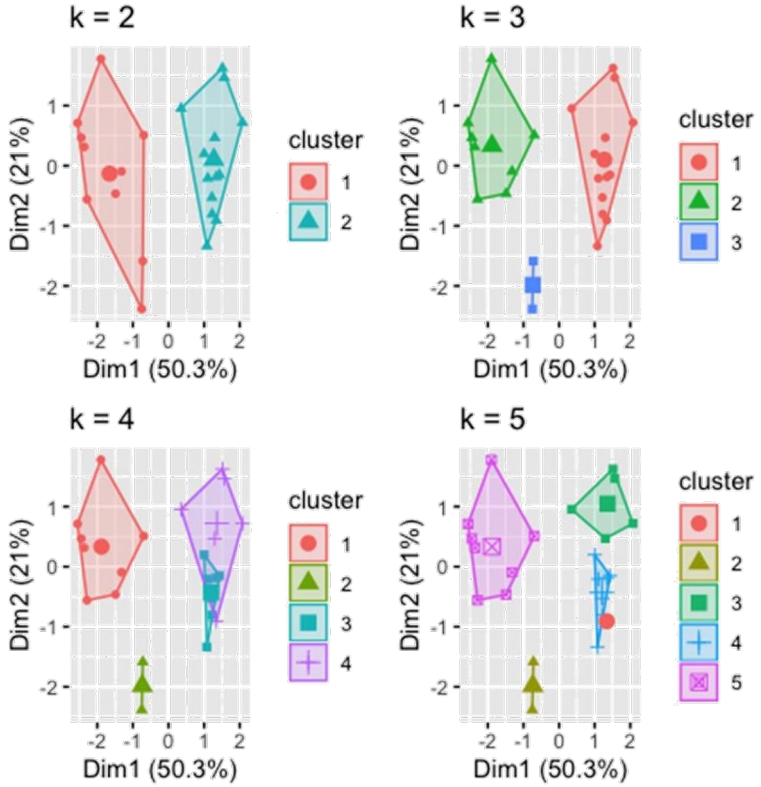
- **K-means**: Often optimised at one level iteratively.
- **Probabilistic** and **fuzzy** models: not a sharp classification, but a probability of belong to a class.

CLASSIFICATION - CLASSES

Hierarchic



Non-hierarchic



CLASSIFICATION – ALGORITHMS

Sequential or simultaneous.

- Grouping by repetition a procedure **OR** a solution in reached in a single step.

Agglomerative or divisive.

- Successively grouping into larger groups **OR** Successively dividing it into subgroups.

Monothetic versus polythetic **.**

- Single descriptor at each grouping step OR all descriptors are used (via association matrix).

Hierarchical versus non-hierarchical .

- Subgroups are nested in higher order groups OR single partition without any hierarchy.

Probabilistic versus non-probabilistic.

- Within-group association matrices have OR not given probability of being homogeneous.

Unconstrained or constrained methods**.**

- Based on a single data set OR two data set (clustered elements + explanatory variables).

CLASSIFICATION – ALGORITHMS

Sequential or simultaneous.

- Grouping by repetition a procedure **OR** a solution in reached in a single step.

Agglom

- Su

Monot

- Sin

Hierarc

- Su

Probabi

- Within-group association matrices have OR not given probability of being homogeneous.

Unconstrained or constrained methods.

- Based on a single data set OR two data set (clustered elements + explanatory variables).

Most methods presented below are sequential, agglomerative and hierarchical

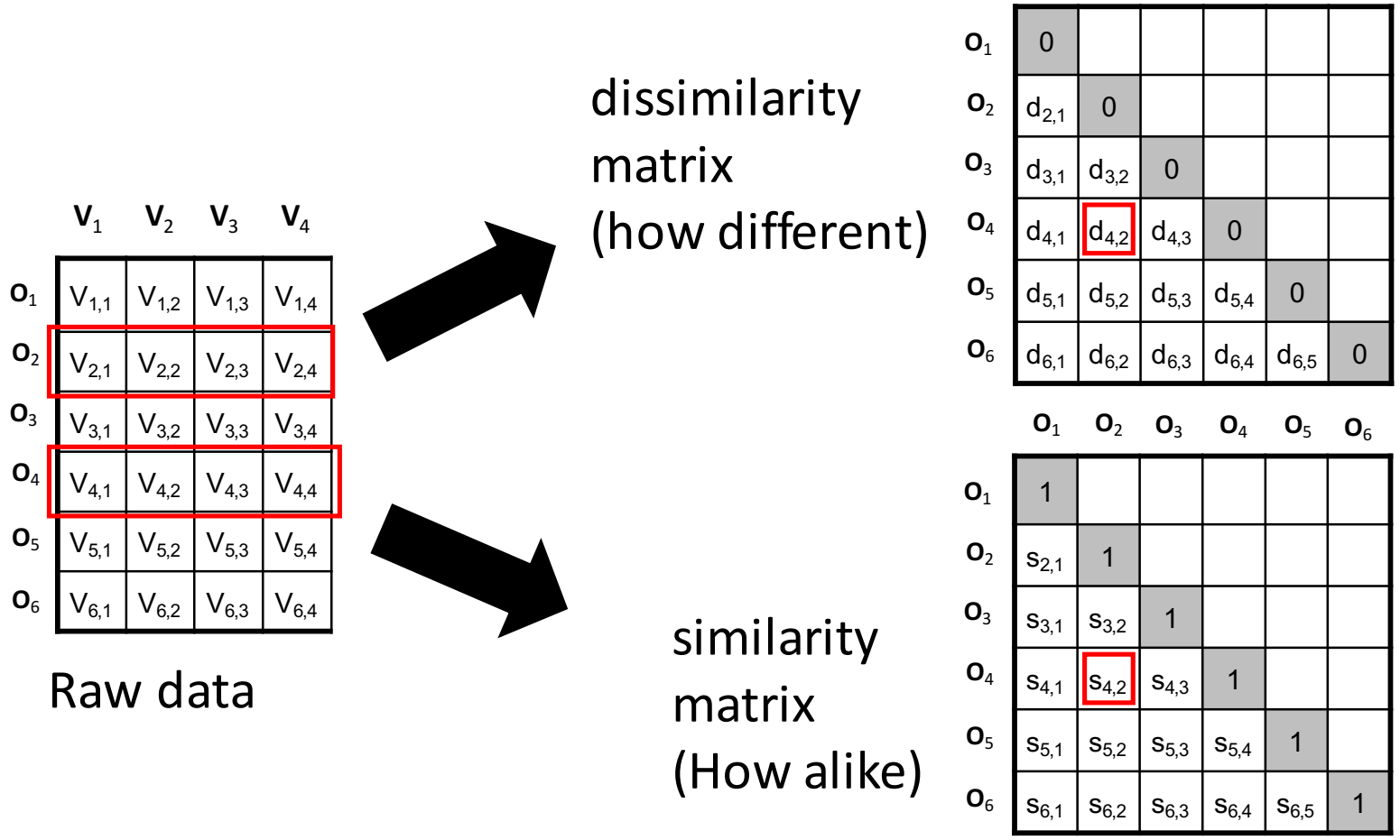
So far so good?

Any questions?

Ready to move on?

CLASSIFICATION

IT IS ALL ABOUT DISTANCES



CLASSIFICATION
ATTRIBUTES OF DISTANCE MEASUREMENTS

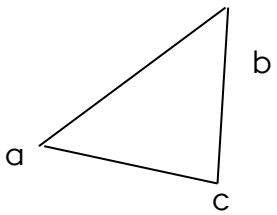
Axioms

	Metric	Semimetric	Non-metric
Coincidence: The minimum value is zero when two items are identical	+	+	+/−
Non-negativity: When two items differ, the distance is positive (negative distances *not* allowed)	+	+	+/−
Symmetry: the distance from object A to object B equals the distance from B to A	+	+	+/−
Triangle inequality: with three objects, the distance between any two of these cannot be larger than the sum of the two other distances	+	−	−

D(a, a) = 0

D(a, b) > 0

D(a, b) = D(b,a)



D(a,c) <= D(a,b)+D(b,c)

CLASSIFICATION

KINDS OF DISSIMILARITY MEASUREMENTS

Name	Metric	Data type	Range	Symmetric	Formula
Euclidean distance	Yes	Quan	[0-∞)	Yes	$\sqrt{\sum_{i=1}^{p}(x_{i,1}-x_{i,2})^2}$
Manhattan metric	Yes	Quan	[0-∞)	Yes	$\sum_{i=1}^{p}\lvert y_{1i}-y_{2i}\rvert$
Mean character difference (Czekanowski 1909)	Yes	Quan	[0-∞)	Yes	$\frac{1}{p}\sum_{i=1}^{p}\lvert y_{1i}-y_{2i}\rvert$
Bray-Curtis Distance (Bray and Curtis 1957)	Semi	Quan	[0,1]	No	$\frac{2a}{2a+b+c}$
Jaccard Distance	Yes	Binary	[0,1]	No	$1-\frac{a}{a+b+c}$

Environmental data

Species data

CLASSIFICATION

KINDS OF SIMILARITY MEASUREMENTS

a=Pres-Pres b=Pres-Abs
c=Abs-Pres d=Abs-Abs
A&B=Total Abundance

$$W = \sum_{i=1}^{S} \min(Abundance)$$

Name	Symmetry	Data type	Metric	Formula
Simple matching coefficient	Yes	Binary	Yes	$\frac{a+d}{a+b+c+d}$
Jaccard's coefficient	No	Binary	Yes	$\frac{a}{a+b+c}$
Sørensen's coefficient	No	Binary	Semi	$\frac{2a}{2a+b+c}$
Gower's coefficient	Yes	Mixed	Yes	$\frac{1}{p}\sum_{j=1}^{p} 1 - \frac{\lvert y_{1j} - y_{2j} \rvert}{R_j}$
Steinhaus' coefficient	No	Quan	Semi	$\frac{2W}{A+B}$
Kulczynski's coefficient	No	Quan	Semi	$\frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right)$

Species data

Environmental data

MEASURING DISTANCES IN R

Most of the ecological relevant distances are incorporated in the `vegdist` method of the `vegan` package

Jaccard and Sørensen coefficients are also available through `betadiver` methods also in the `vegan` package

The code for R

```
require(vegan)
data(varespec) # Vegetation and environment in lichen pastures data
vegdist(varespec)
vegdist(varespec, method="bray") # However the default distance is Bray-Curtis
vegdist(decostand(varespec, "norm"), method="euclidean")) # Euclidean Distance. The decostand
function is used for transforming community matrices
vegdist(decostand(varespec, "range"), method="gower") # Gower's Distance
```

So far so good?

Any questions?

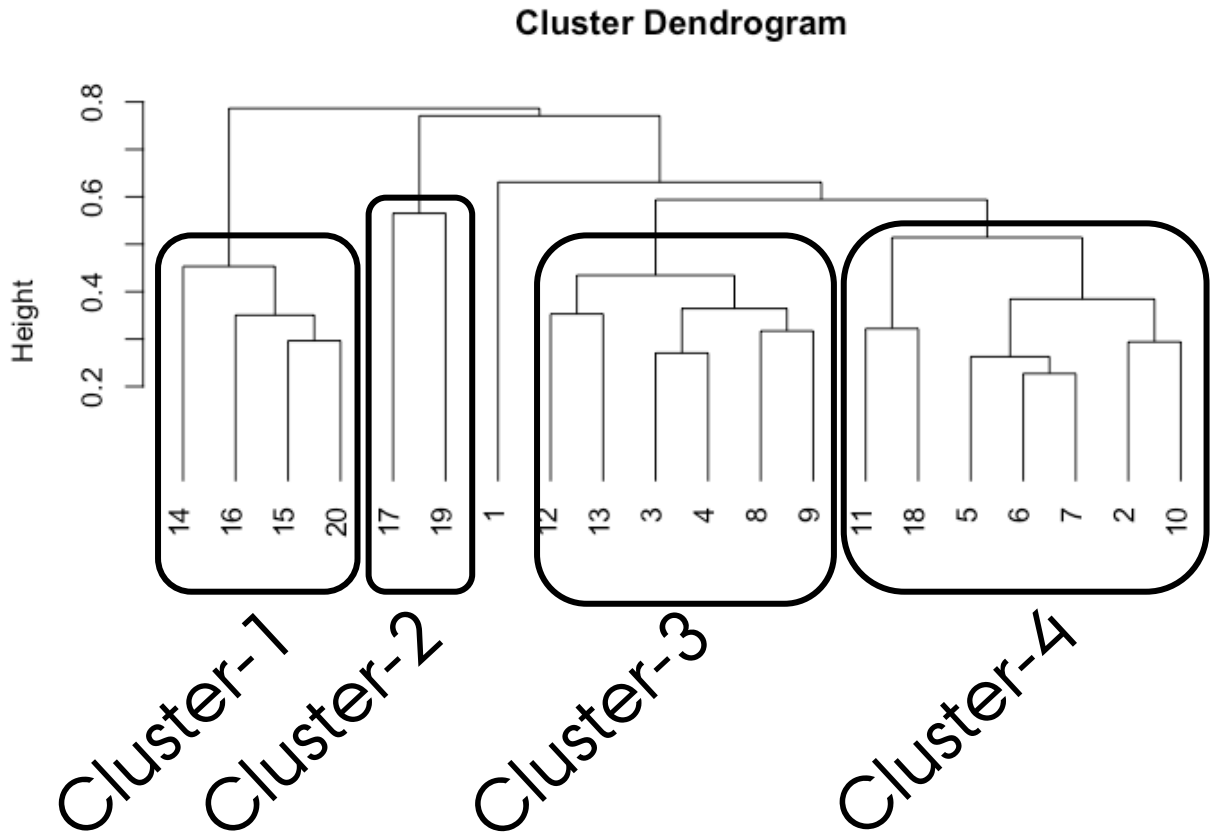
Ready to move on?

HIERARCHICAL CLUSTERING

Also called **Hierarchical cluster analysis** or **HCA** is an unsupervised clustering algorithm

The endpoint is a set of nested clusters,

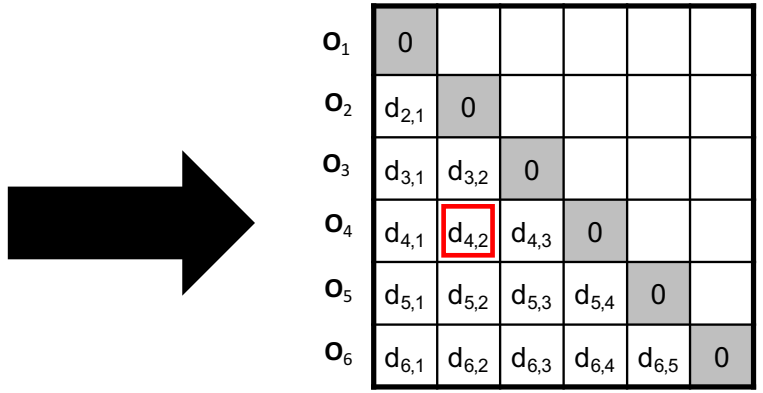
- Each cluster is distinct from each other cluster.
- Objects within each cluster are broadly similar to each other.



HIERARCHICAL CLUSTERING

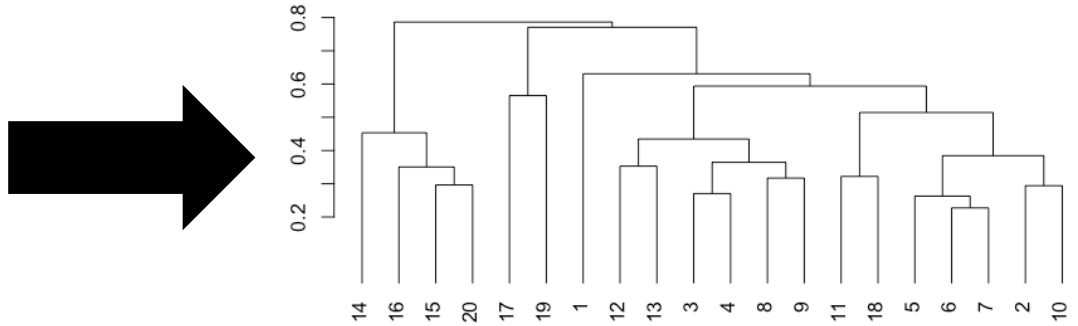
The **input** is a distance matrix that defines the "similarity" between objects.

- So the first decision to make is which distance to use.



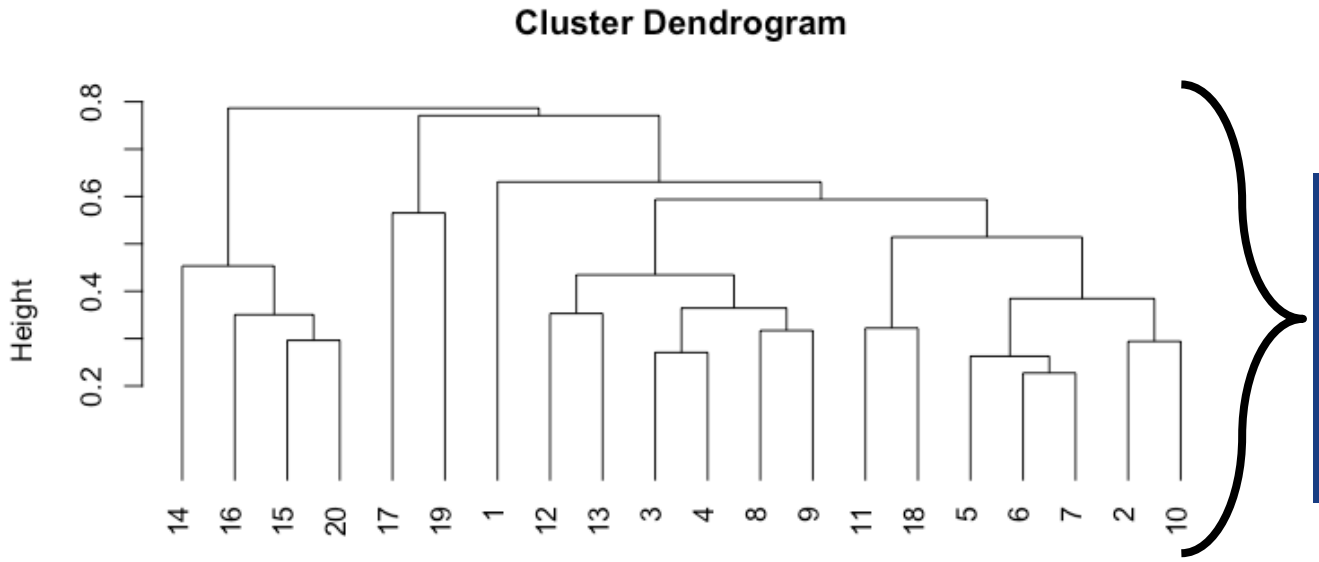
The main **output** of Hierarchical Clustering is a dendrogram.

- The appearance of it dependence on the Linkage Criteria → how are clusters put together



HIERARCHICAL CLUSTERING – THE STEPS

1. Start with combining/splitting two most similar sampling units
2. Proceed with combining two next similar or cluster to cluster
3. Continue until all objects are combined within the **tree.**

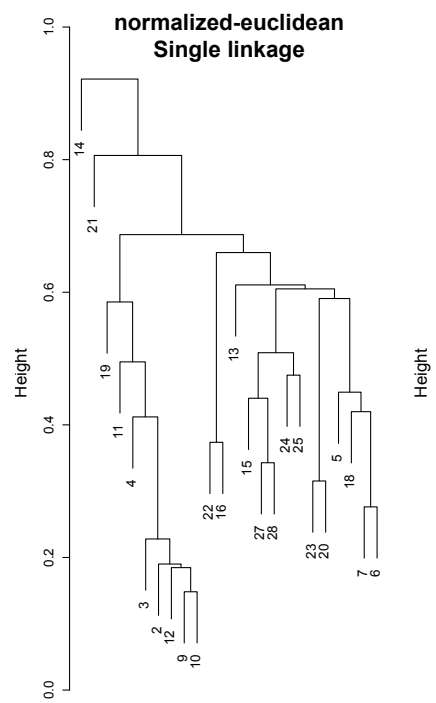


These trees are build using the `hclust()` f function on a distance matric crated using `dist()` or `vegdist()`.

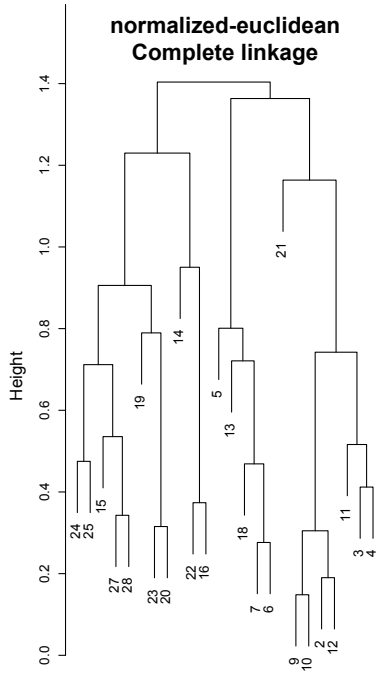
HIERARCHICAL CLUSTERING

HOW TO CLUSTER?

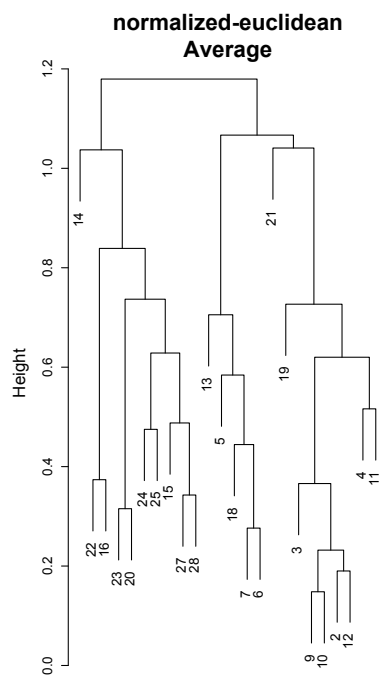
Single linkage
the most different points.
[**Furthest neighbours**]



Complete linkage
the closest points.
[**Nearest neighbour**]



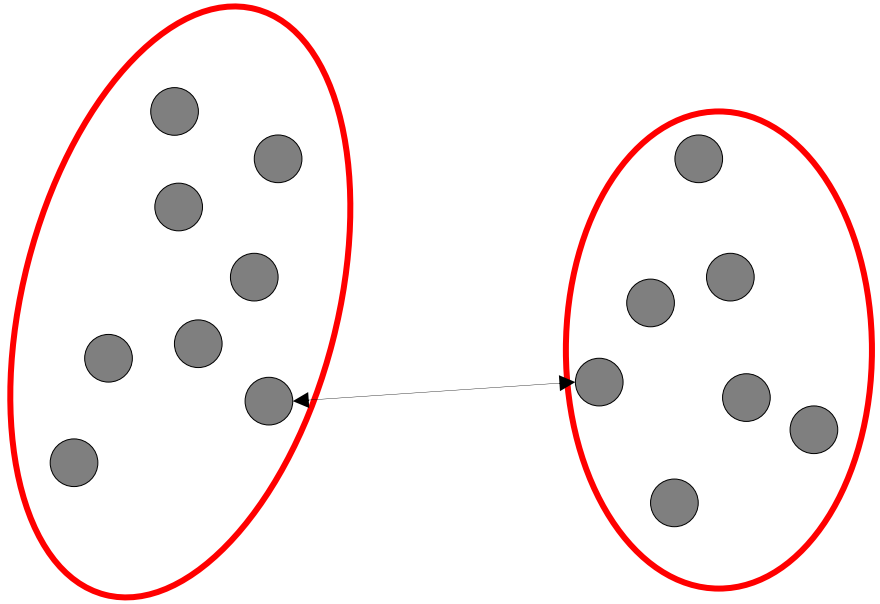
Average linkage
mean distances of group centroids.



AGGLOMERATIVE CLUSTERING
SINGLE LINKAGE

Single Linkage: individual observations are joined based on the shortest pairwise dissimilarities (=grater similarities).

- **Good** to find gradients.
- **Bad** to see partitions.
- Related to Minimum Spanning Tree: the shortest route that joins all points together.



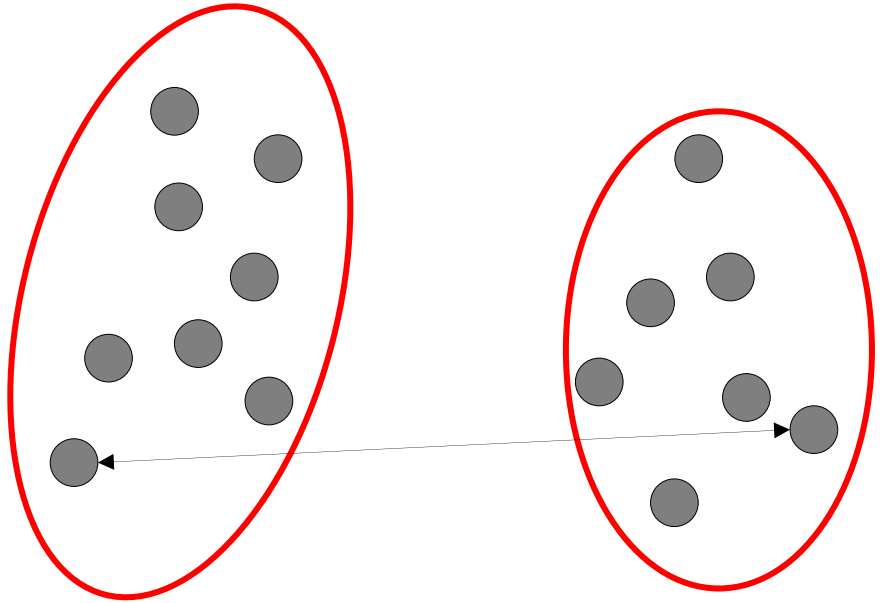
$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Distance between two clusters is defined as the shortest distance between two points in each cluster.

AGGLOMERATIVE CLUSTERING
COMPLETE LINKAGE

Complete Linkage: individual observations are joined based on the dissimilarity between the furthest object in a group

- Makes compact clusters by minimising the **diameter** of groups, or the longest possible distance within a cluster.
- Groups do not grow heterogeneous, but remain small and compact.



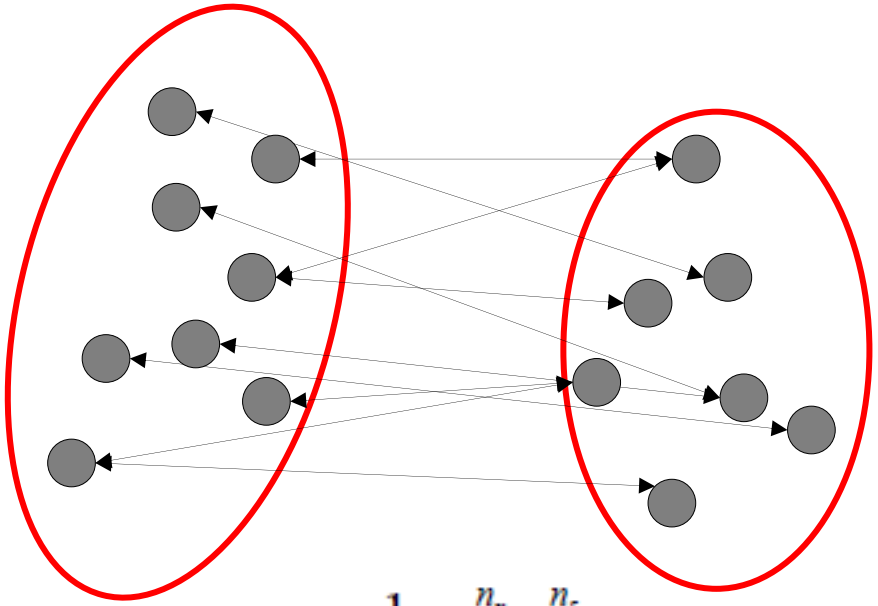
$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Distance between two clusters is defined as longest distance between two points in each cluster.

AGGLOMERATIVE CLUSTERING
AVERAGE LINKAGE

Average Linkage is a compromise between the simple and complete linkage approaches.

- There are four methods differing on how the position between groups is defined.
- UPGMA [best known] → joining is at the mean of the dissimilarities to all members of the group.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster Or the position of the centroids

AGGLOMERATIVE CLUSTERING
AVERAGE LINKAGE

Table 4.1 The four methods of average clustering. The names in quotes are the corresponding arguments of function **hclust()**

	Arithmetic average	Centroid clustering
Equal weights	Unweighted pair-group method using arithmetic averages (UPGMA) "`average`"	Unweighted pair-group method using centroids (UPGMC) "`centroid`"
Unequal weights	Weighted pair-group method using arithmetic averages (WPGMA) "`mcquitty`"	Weighted pair-group method using centroids (WPGMC) "`median`"

Here, "**Weights**" mean if the number of object in the group are consider on the "Averaging"

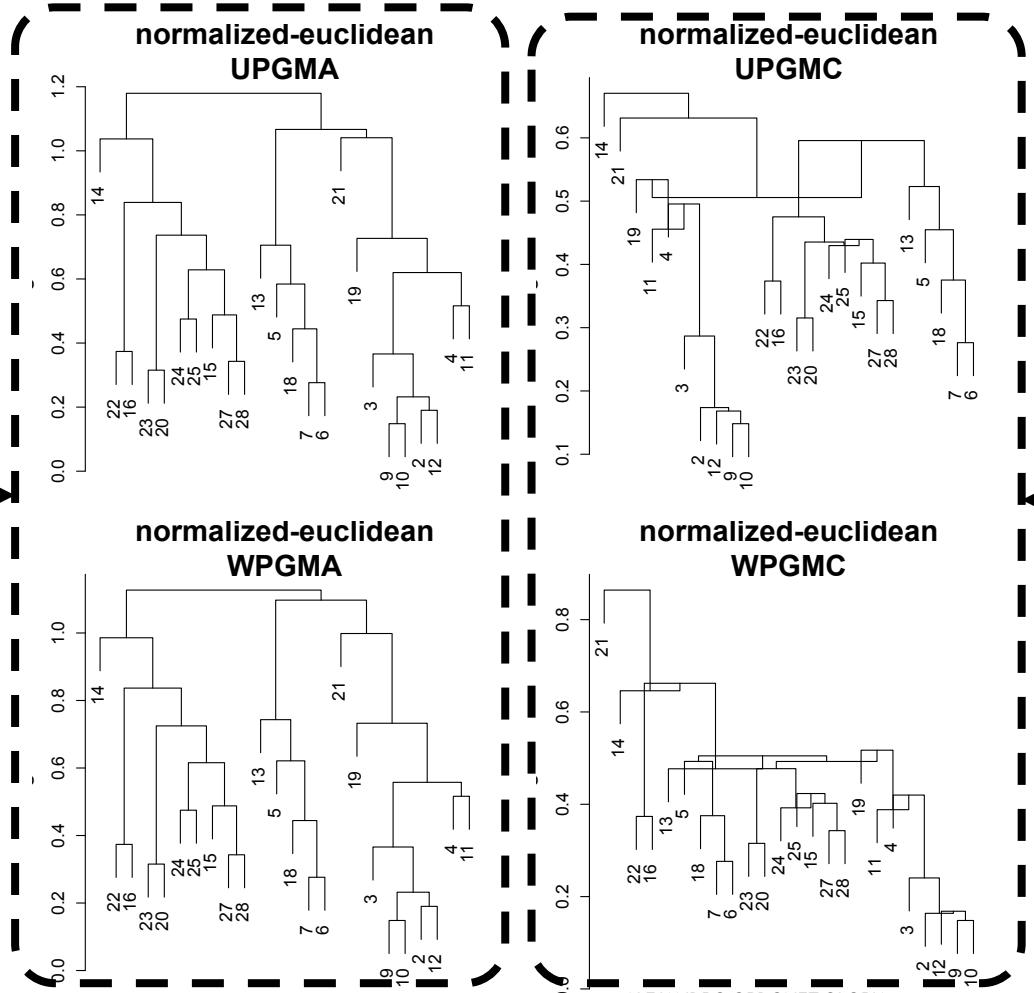
AGGLOMERATIVE CLUSTERING
AVERAGE LINKAGE

Avg distance between observations

Avg distance between centroids

Are these two different?

Are these two different?



So far so good?

Any questions?

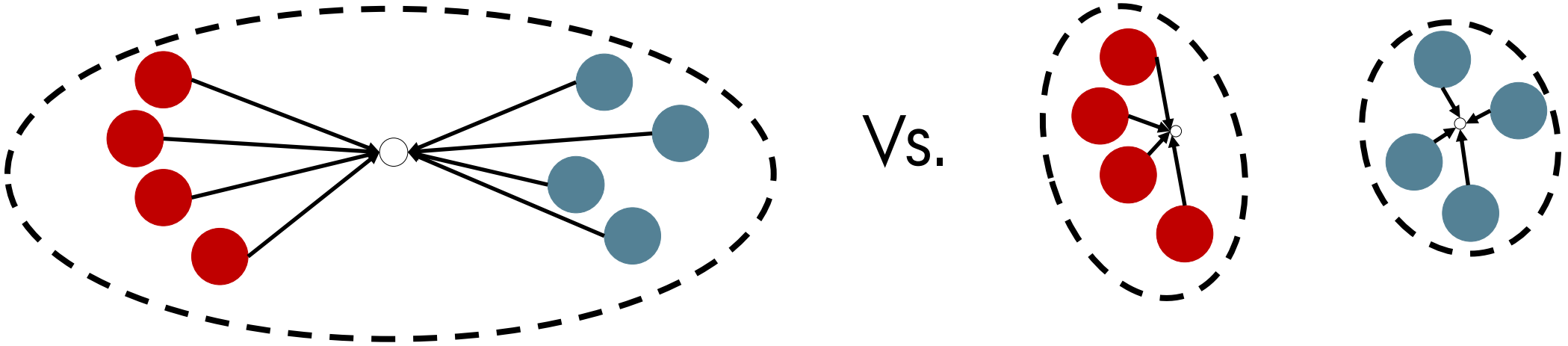
Ready to move on?

WARD'S HIERARCHICAL CLUSTERING

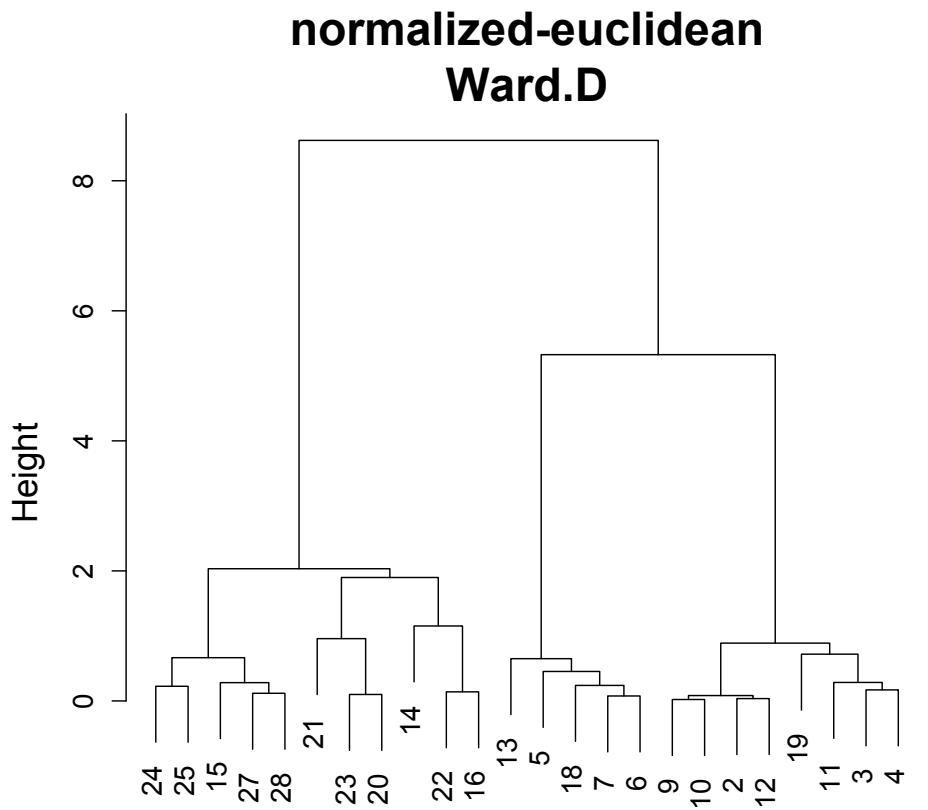
The guiding principle

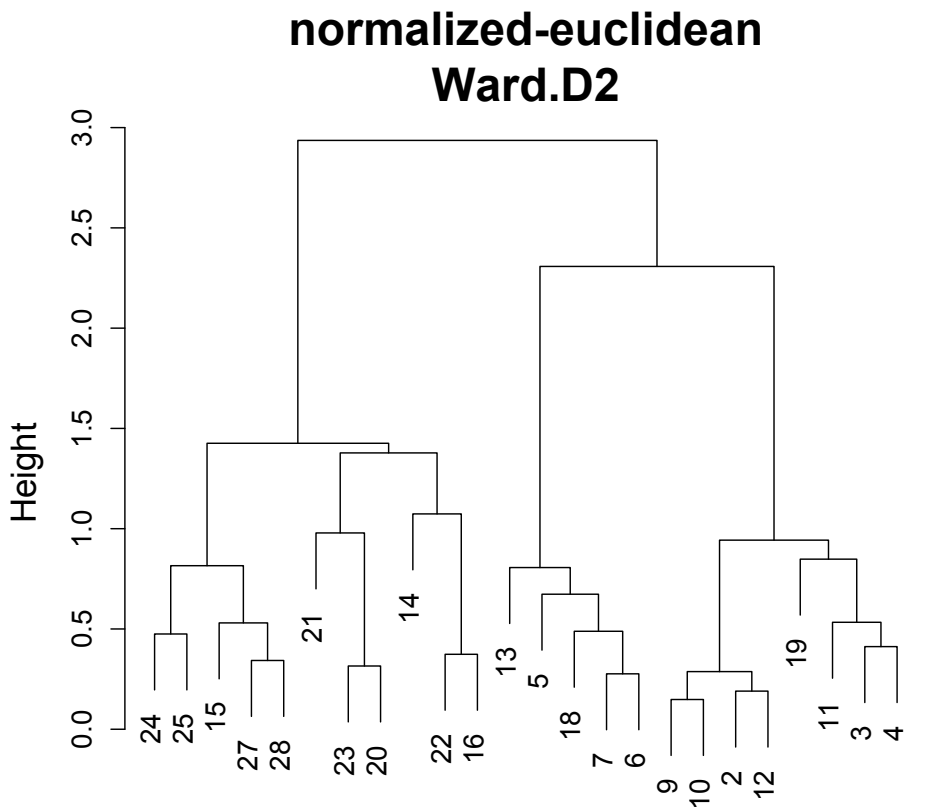
This method is based on the linear model criterion of **least squares**. [relates them to the linear model]

- The guiding principle is to minimize the within-cluster sum of squared errors. → minimising the within group variability.
- It leads to compact "spherical" clusters



WARD'S HIERARCHICAL CLUSTERING






```
hclust(dist(x)^2,method="ward.D")
```

Dissimilarities are transformed to a scale of squared distances.

```
hclust(dist(x),method="ward.D2")
```

Dissimilarities are on a scale of distances (distances are NOT squared)

PROBABILISTIC CLUSTERING

Goal:

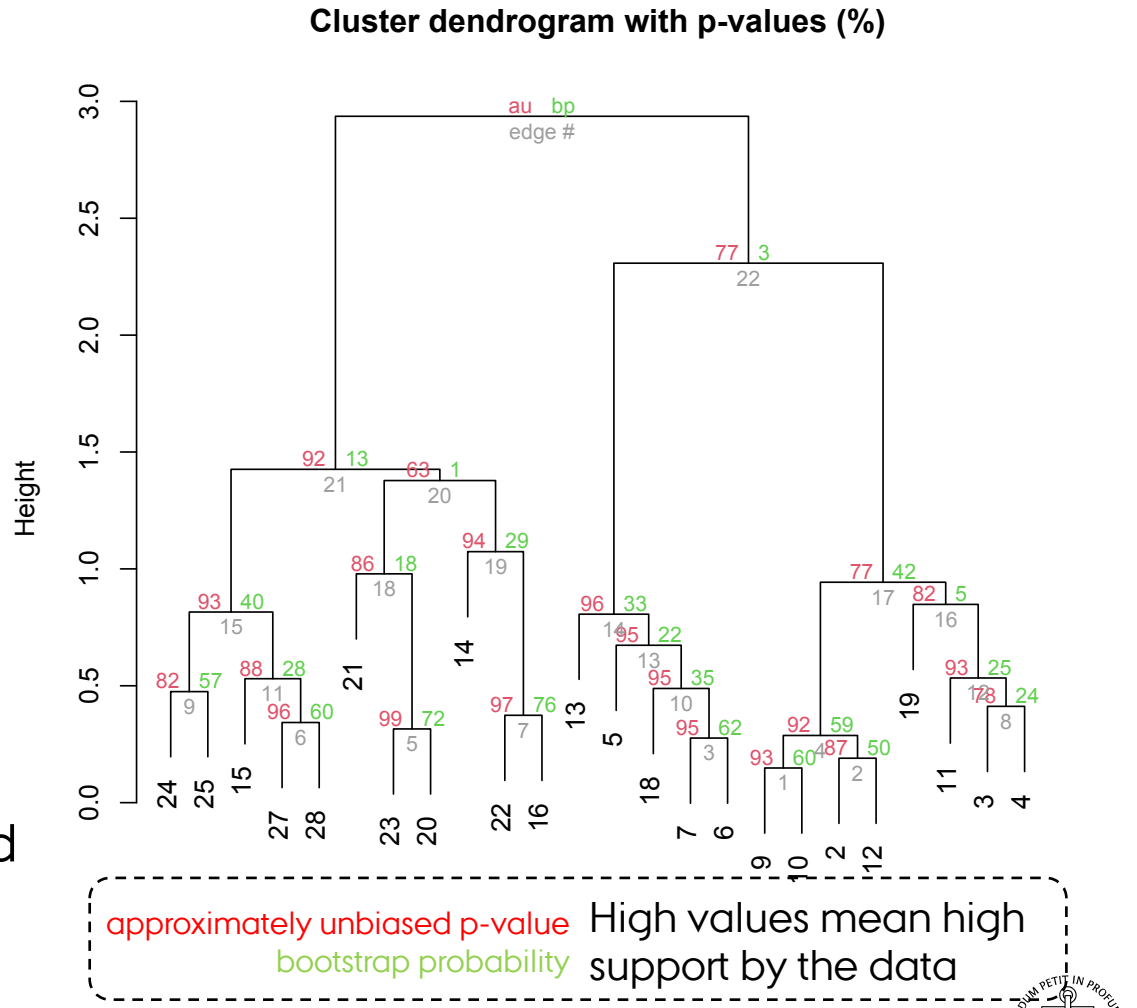
Consider (natural) variation and assess the uncertainty of a classification.

How to?

Bootstrap resampling

- randomly sampling subsets of the data and computing the clustering on these subsets.

The `pvclust` package provides functions to plot a dendrogram with bootstrap p-values associated to each cluster.



So far so good?

Any questions?

Ready to move on?

HIERARCHIC CLUSTERING
TREES AND DISTANCES

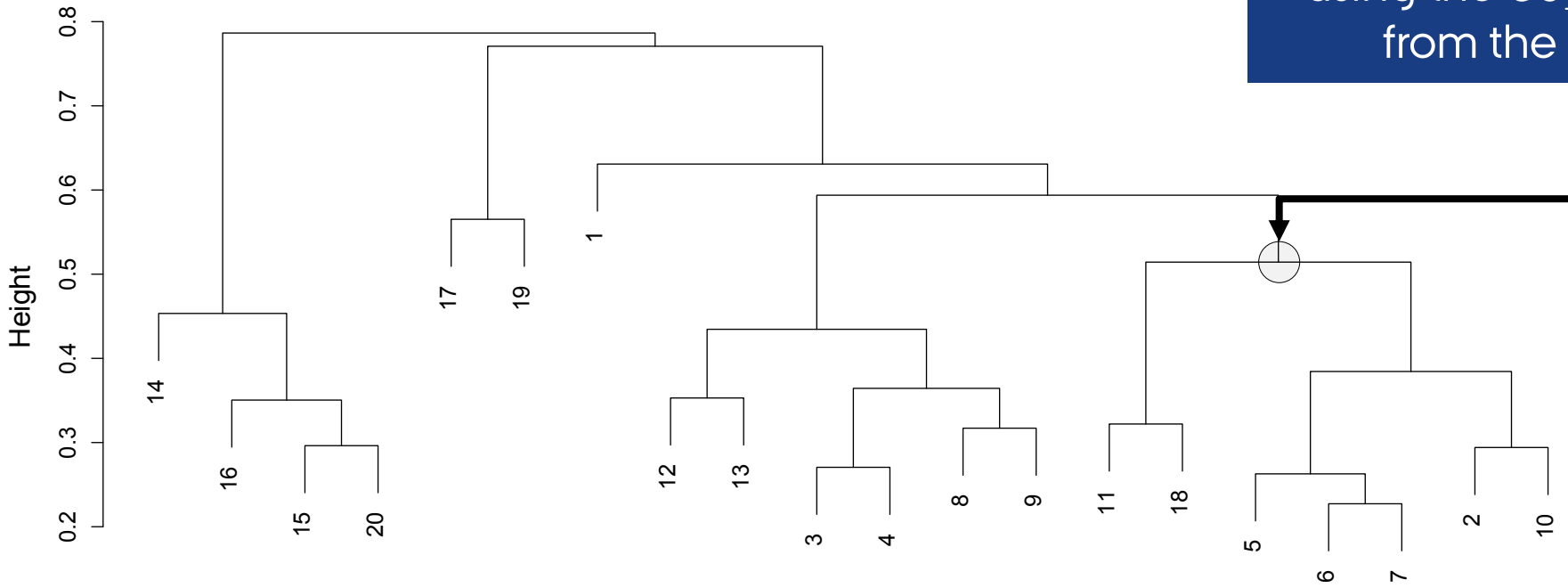
The goal of any tree is that it represents the original dissimilarity matrix.

How to test this?

- Assess the relation between original dissimilarities and cophenetic distances.
  - **Pearson's r correlation** → cophenetic correlation [should be large].
  - **Gower distance** → cophenetic distance [should be small].

The cophenetic correlation and Gower distance criteria do not always designate the same clustering result as the best.

HIERARCHIC CLUSTERING
TREES AND DISTANCES

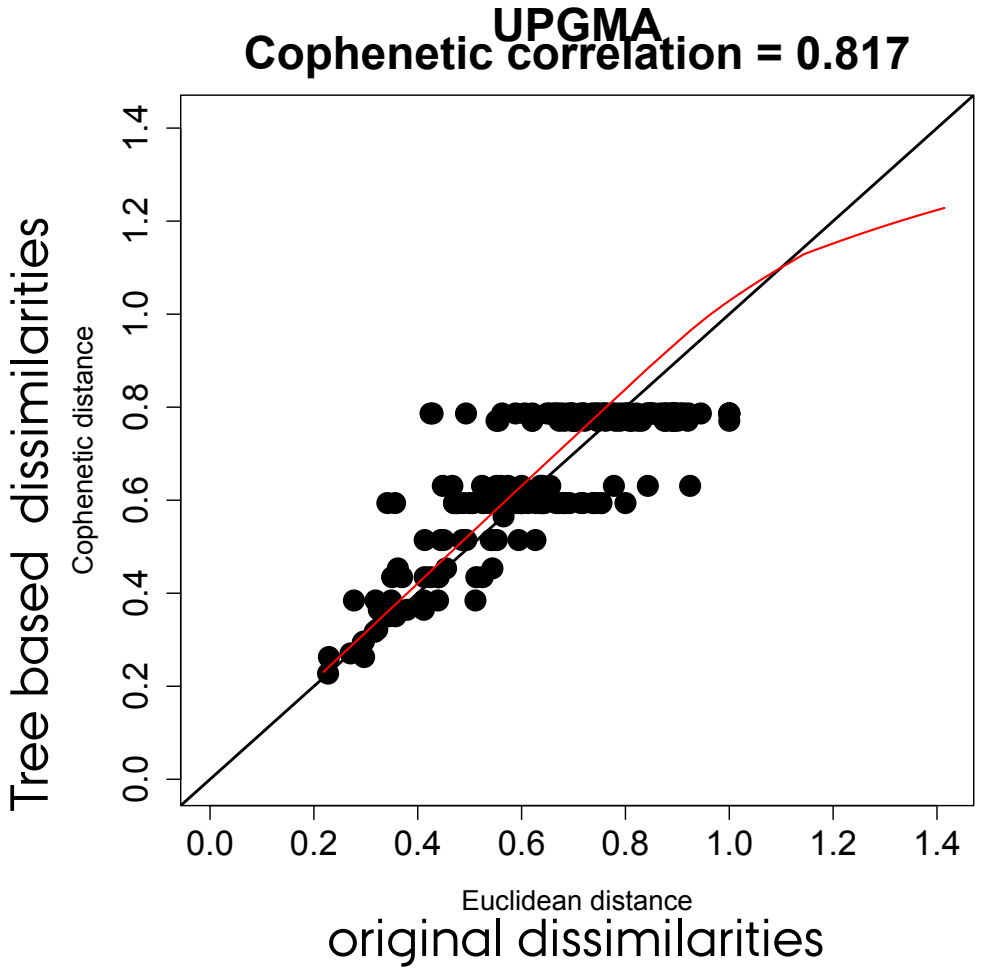


Trees dissimilarities measured in R using the `cophenetic` methods from the `stats` package

The height of the common root for two observations is their distance in a tree

HIERARCHIC CLUSTERING
TREES AND DISTANCES

Assess the relation between original dissimilarities and Tree based dissimilarities



So far so good?

Any questions?

Ready to move on?

HIERARCHIC CLUSTERING
FROM A TREE TO *N*-CLASSES

The goal:

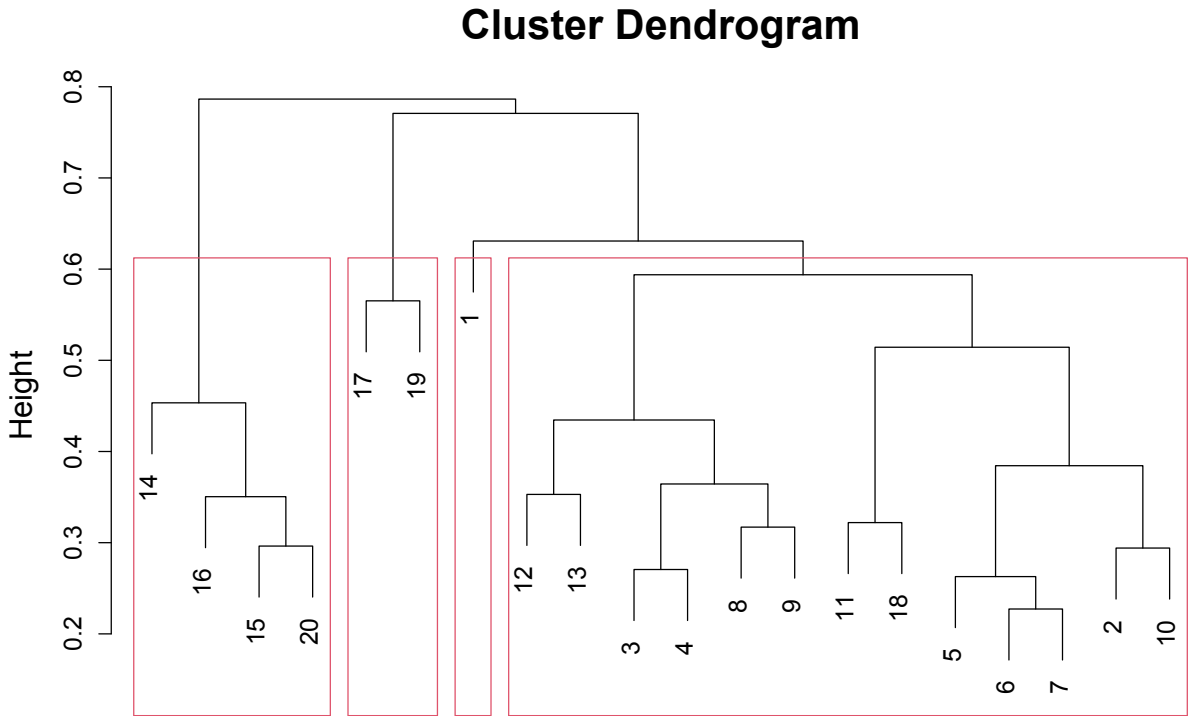
Look for interpretable clusters.

How to?

This means that a decision must be made: at what level should the dendrogram be cut?

Two criteria are possible

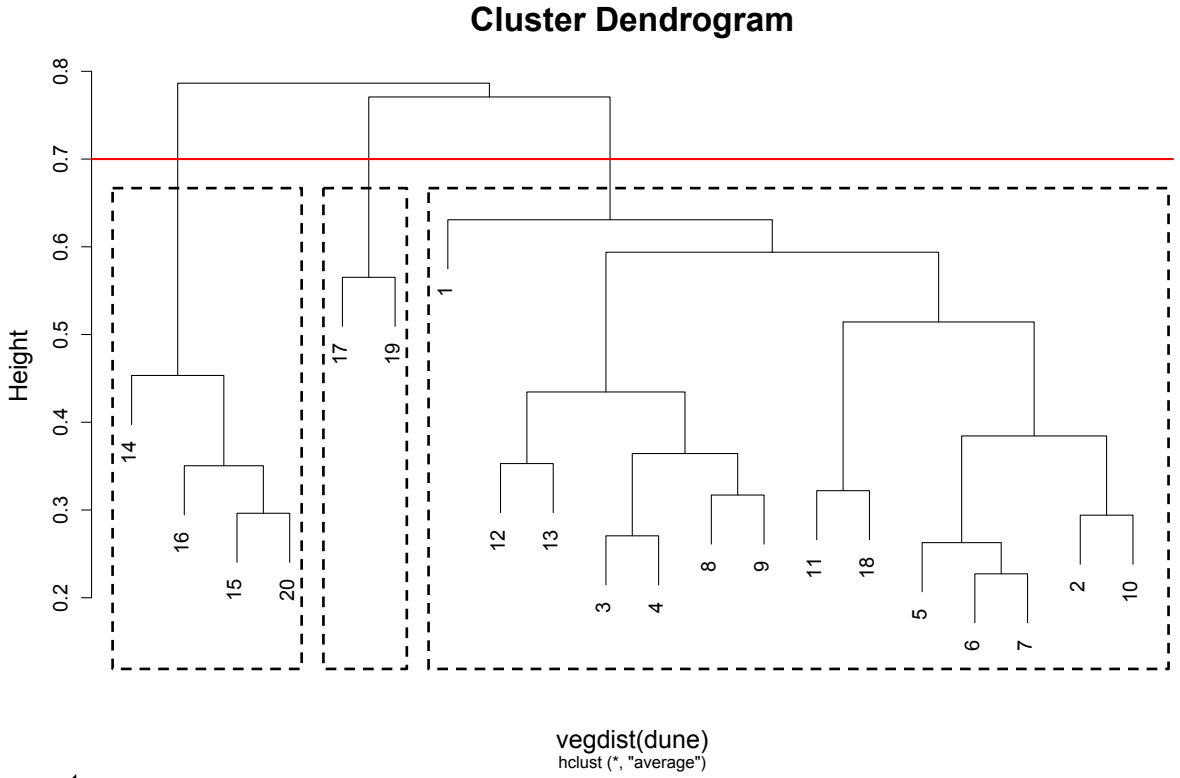
- Single cutting level for a whole dendrogram → 70%.
- Predefined a common number of groups → k-values!



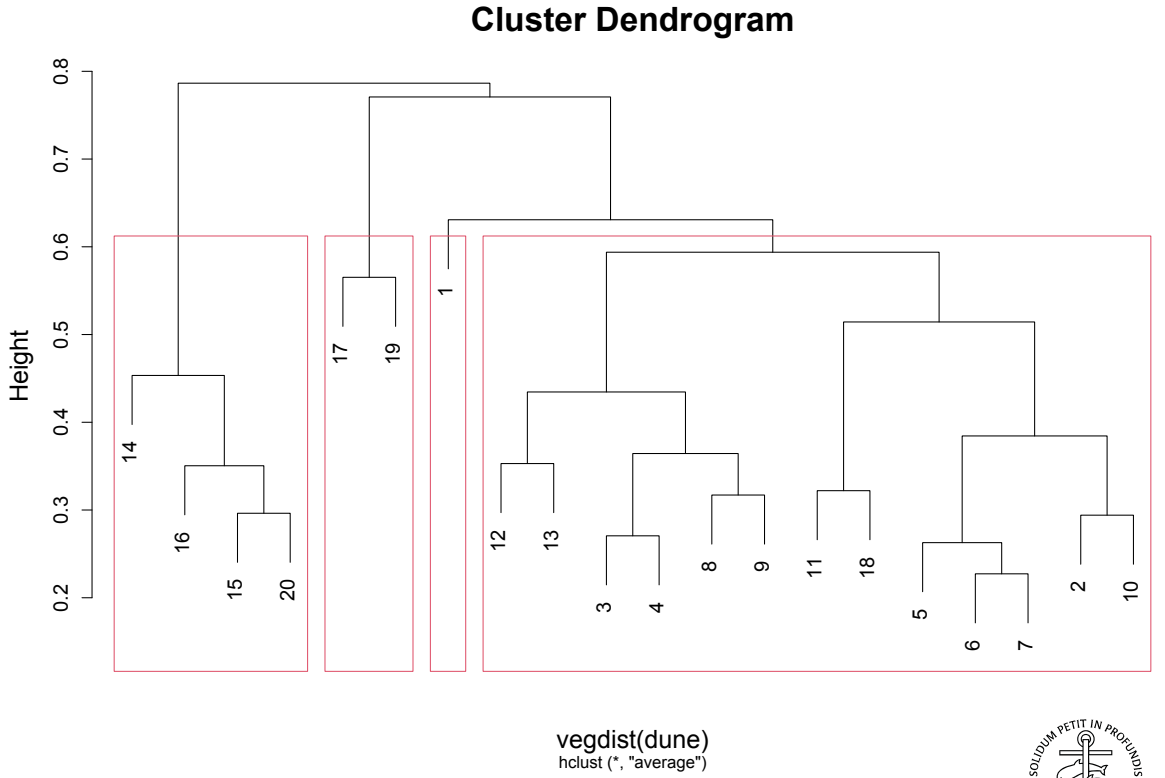
All observations are combined in a tree, but you can extract 1 … *n* classes

HIERARCHIC CLUSTERING
FROM A TREE TO N-CLASSES

A singe dissimilarity level

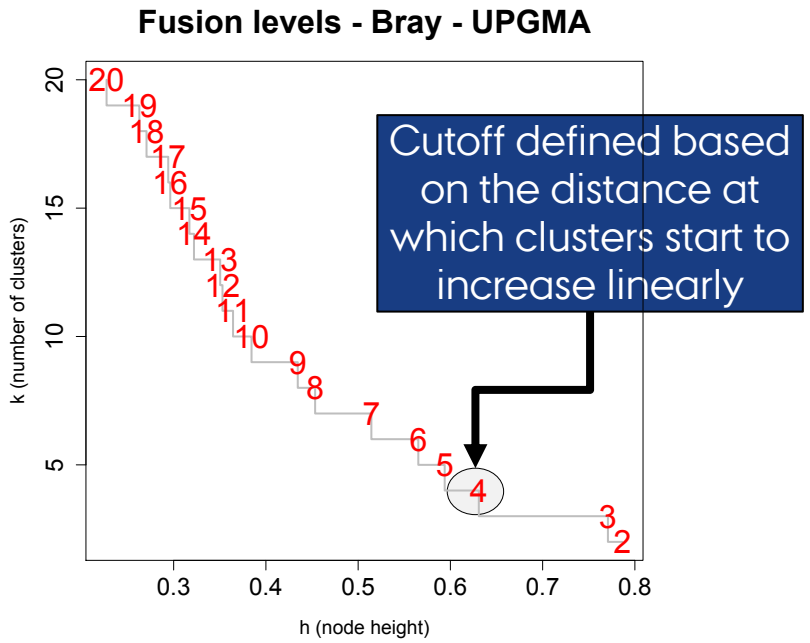


A defined number of clusters

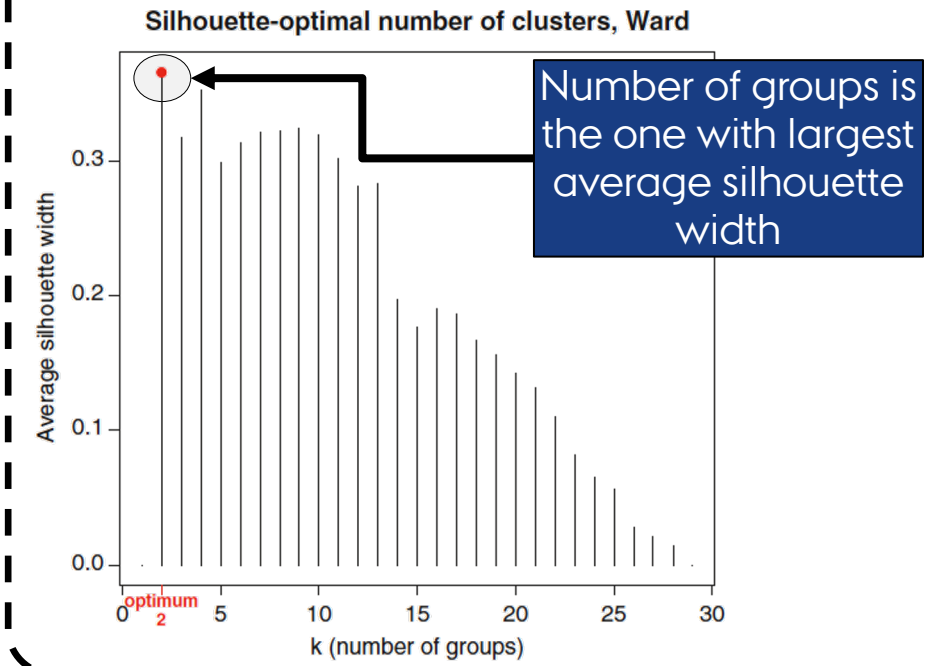


HOW MANLY CLUSTERS?

Approach 1:
Fusion Level Values.



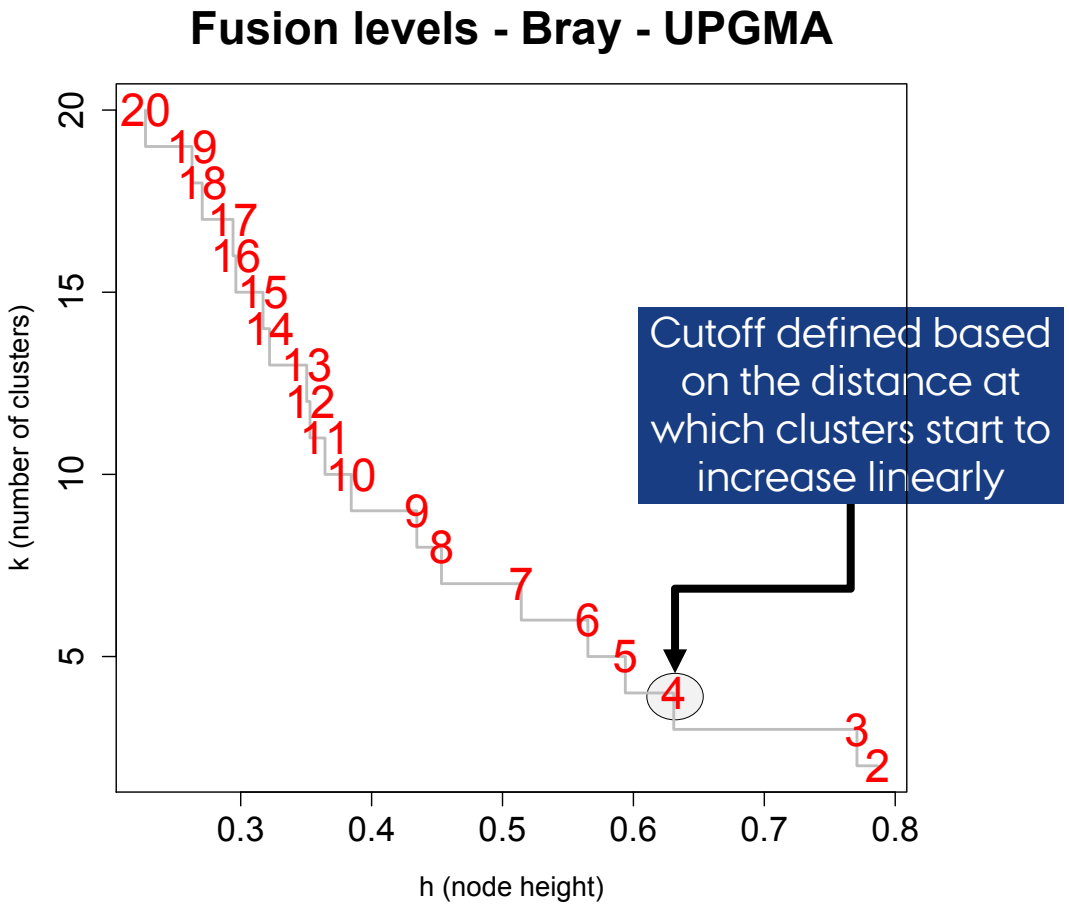
Approach 2:
Silhouette Widths.



HOW MANLY CLUSTERS?
FUSION LEVEL VALUES.

Graph of the Fusion Level Values.

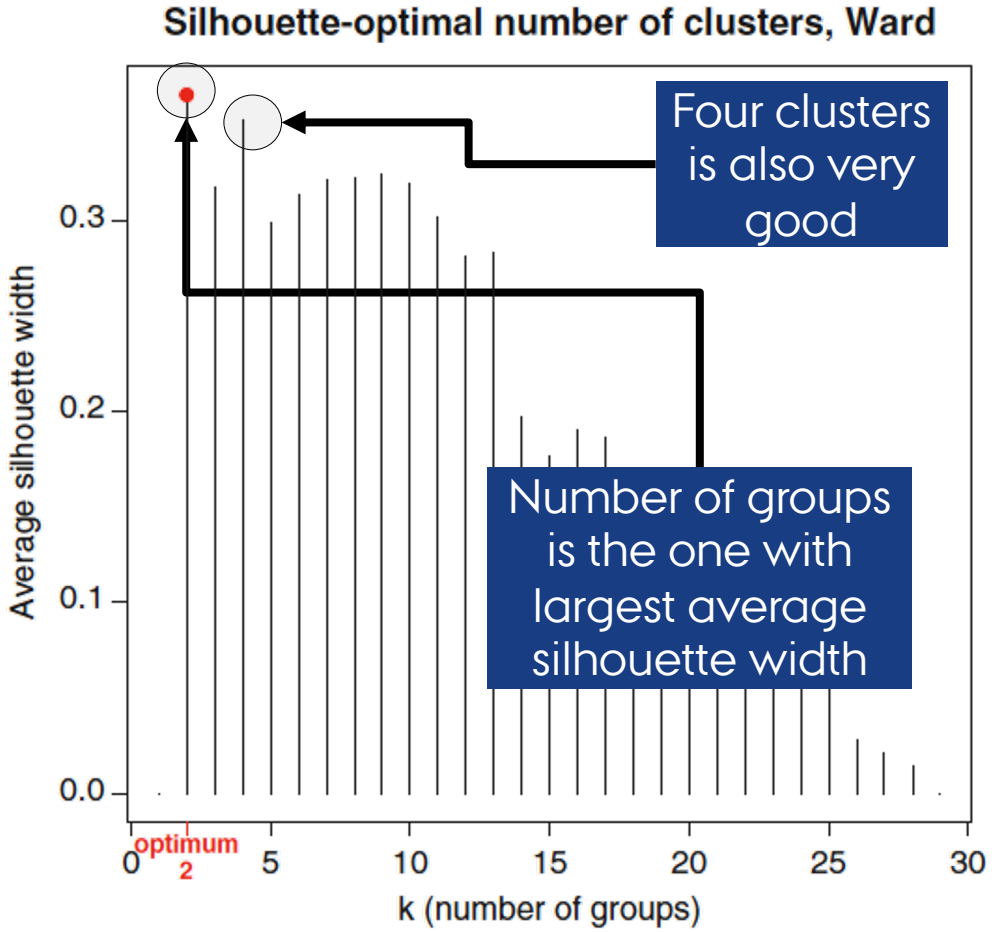
- The fusion level values = the dissimilarity where a fusion between two branches occurs.
- Plotting the fusion level values may help define cutting levels.



HOW MANLY CLUSTERS?
SILHOUETTES WIDTHS

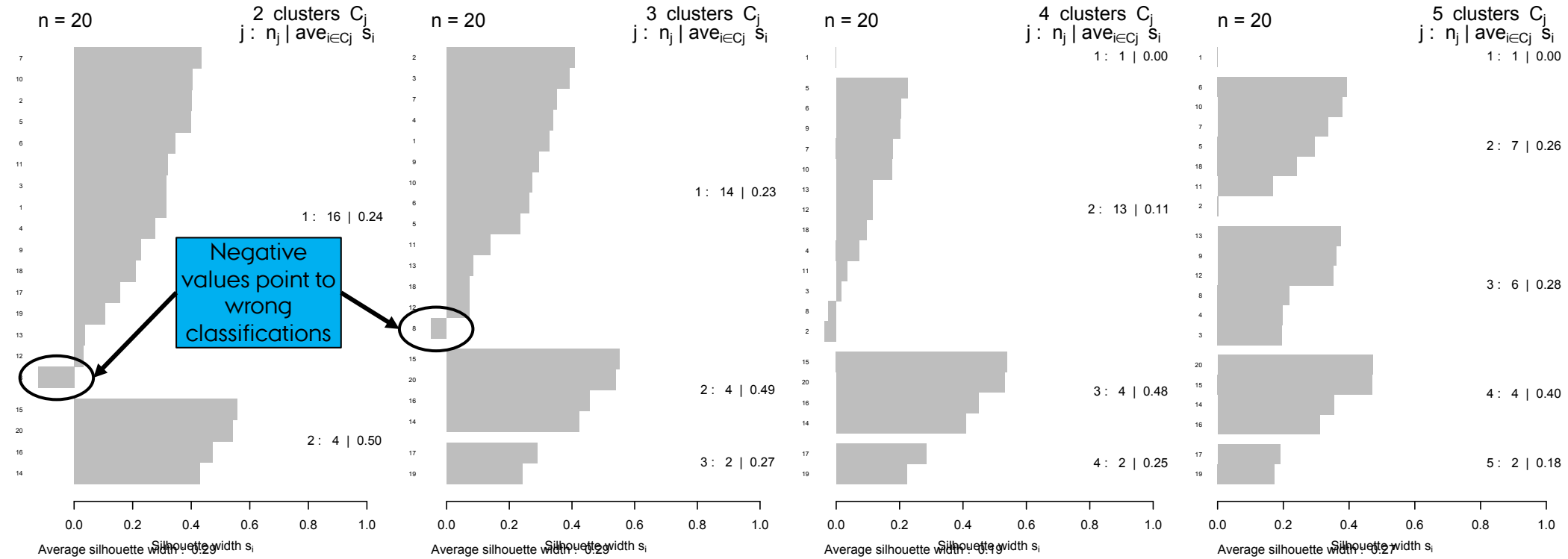
Graphs of Silhouette Widths.

- Silhouette width is a measure of the degree of membership of an object.
  - How similar an object is to its own cluster (**cohesion**) compared to other clusters (**separation**).
- High Silhouette Widths == good match



HOW MANLY CLUSTERS?

SILHOUETTES WIDTHS



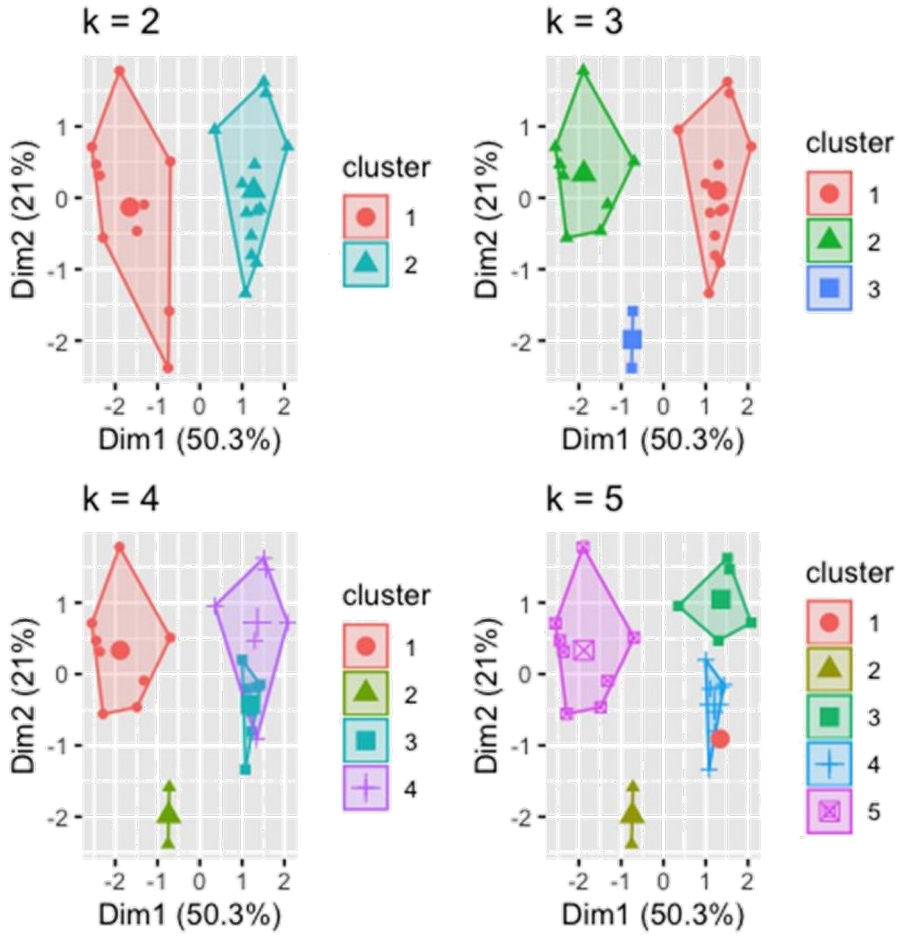
So far so good?

Any questions?

Ready to move on?

NON-HIERARCHICAL CLASSIFICATION
PARTITIONING OF DATA

Non-hierarchical partitioning classifies data into **PREDEFINED** number of groups, by identifying high-density regions in the data



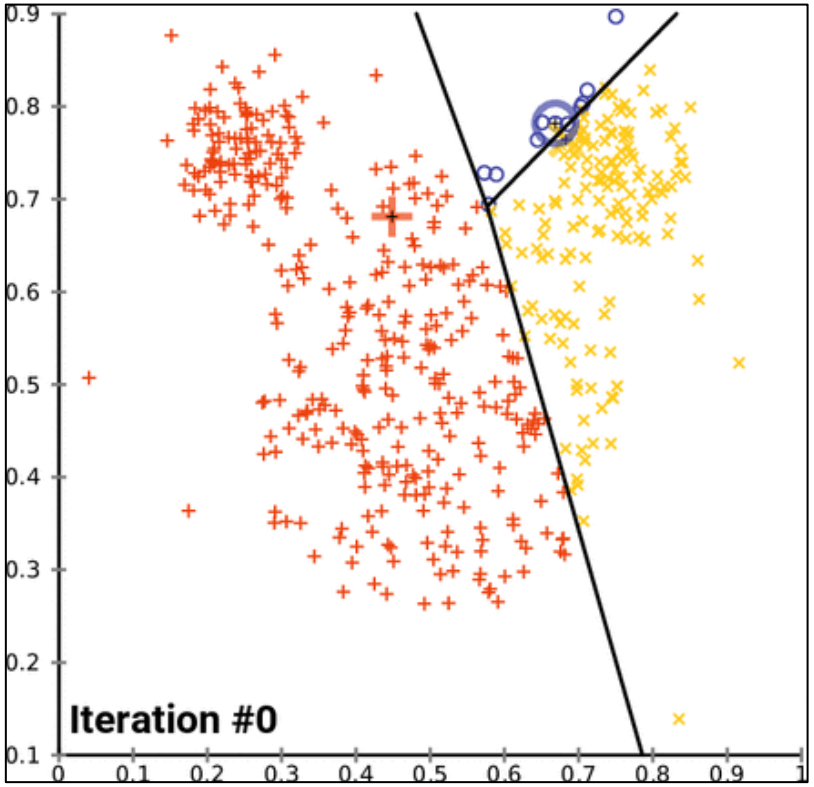
NON-HIERARCHICAL CLASSIFICATION
PARTITIONING OF DATA

The Principles of Non-hierarchical classification

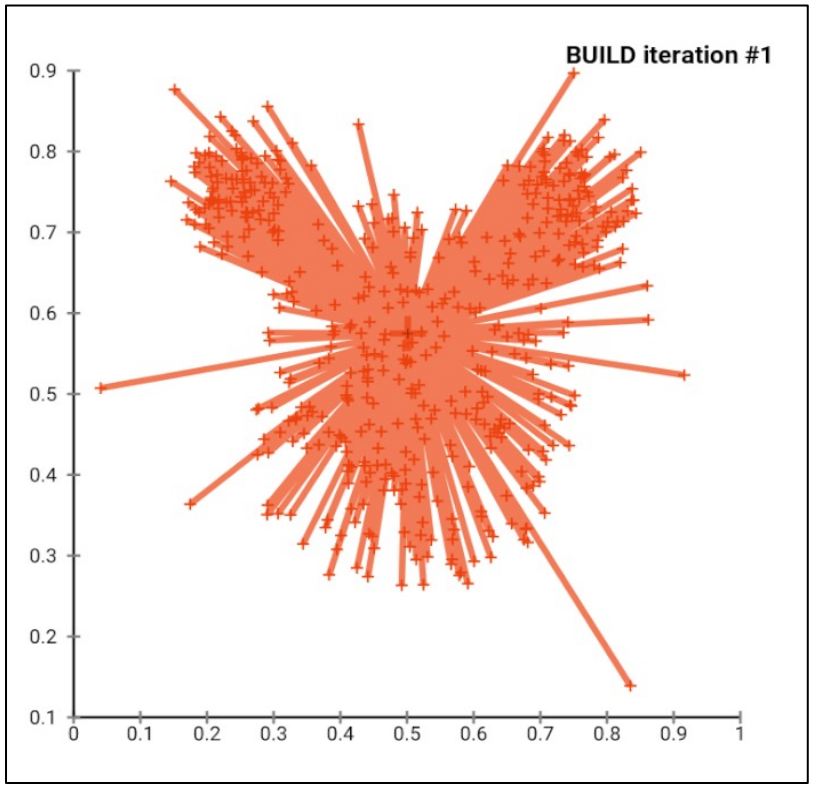
1. Operate in Euclidean space:
    - This means variables need to be in the same units so that the clusters are meaningful.
2. The process is partitional in nature:
    - Breaking the dataset up into a predefined number of groups.
3. Goal is minimize the distance on object and the cluster centre.
    - How you define this carter determine the two possible approaches

NON-HIERARCHICAL CLASSIFICATION
PARTITIONING OF DATA

k-means Clustering



k-medoids Clustering

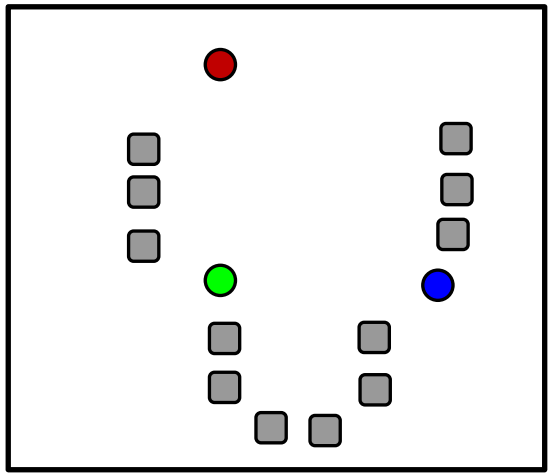


NON-HIERARCHICAL CLASSIFICATION
K-MEANS CLUSTERING

Goal:

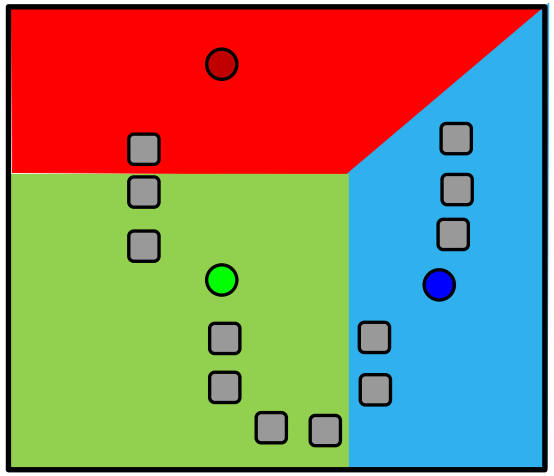
minimize the distance on object and the cluster centre/mean.

Step 1



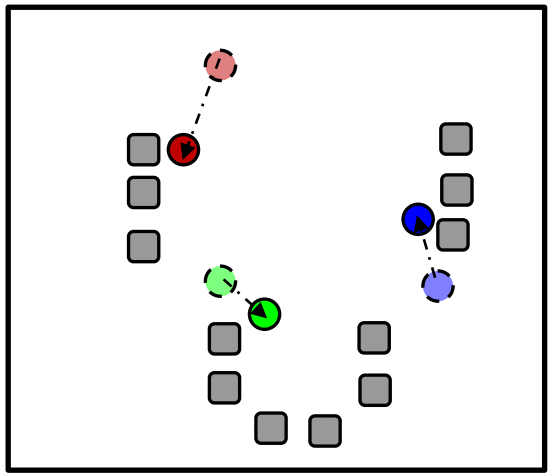
K initial "means" are randomly generated within the data domain

Step 2



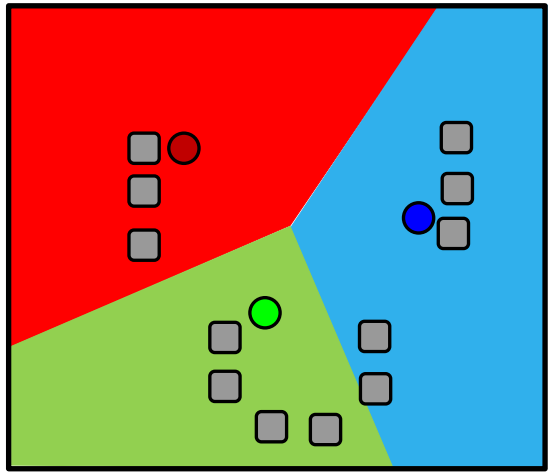
K initial "means" are randomly generated within the data domain

Step 3



The centroid of each of the k clusters becomes the new mean.

Step 4...



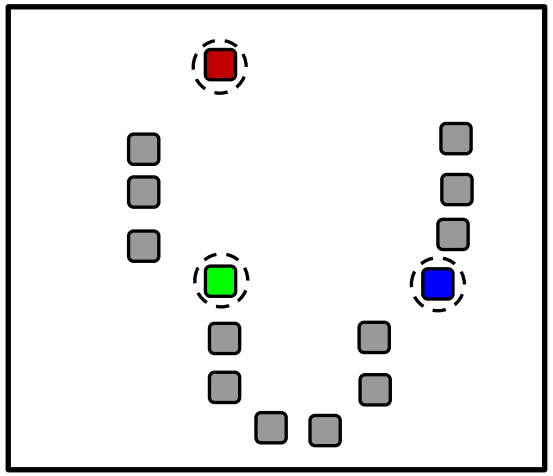
Steps 2 and 3 are repeated until convergence has been reached.

NON-HIERARCHICAL CLASSIFICATION
K-MEDOIDS CLUSTERING

Goal:

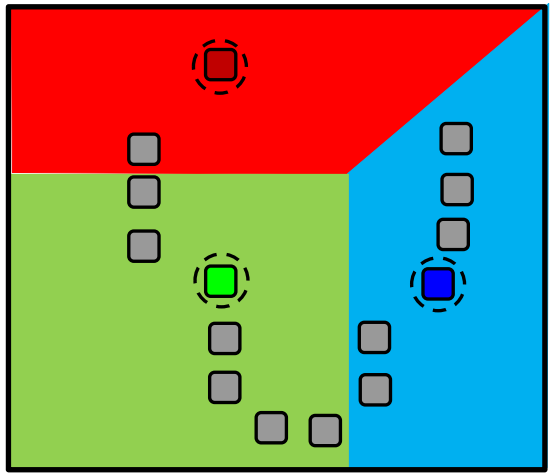
minimize the distance on object and a cluster *k-medoid*.

Step 1



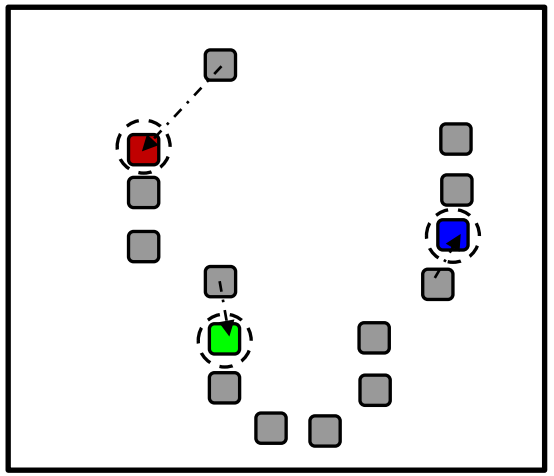
K initial "medoid" are randomly generated within the data domain

Step 2



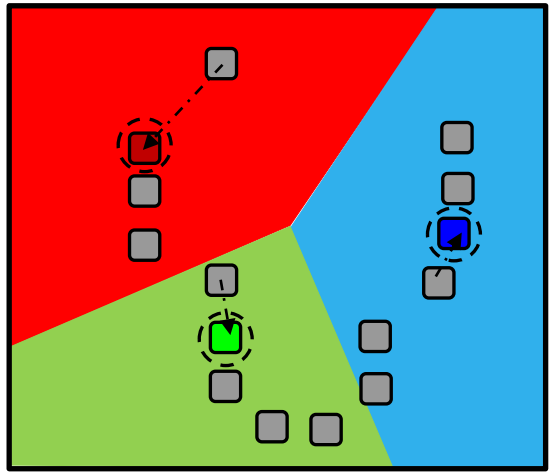
K initial "groups" are randomly generated within the data domain

Step 3



The center of each of the k clusters is moved to another medoid.

Step 4...



Steps 2 and 3 are repeated until convergence has been reached.

medoid = objects of a data set

So far so good?

Any questions?

Ready to move on?

NON-HIERARCHICAL CLASSIFICATION
FUZZY CLUSTERING

The Basis.

- Sometimes cluster limits may not be so clear-cut as one would like them to be.

The goal.

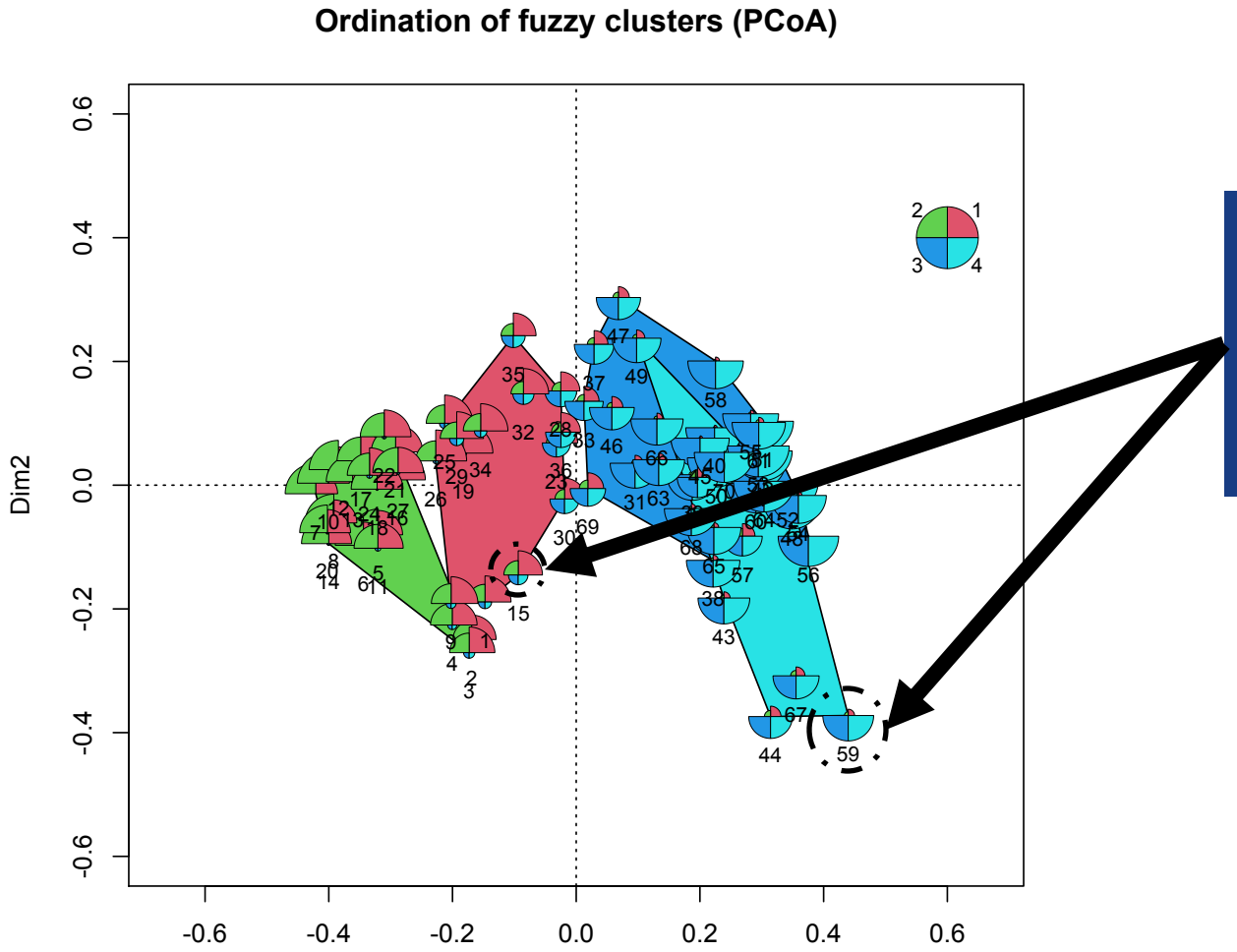
- Measure an object degree of memberships to the various clusters.

The how to.

- One simple approach is the the estimation of c-means.
- There are many functions and packages here we focus on the function `fanny()` of the `cluster()` package.
- But as k-means and k-medoids → **you specify how may groups before hand.**

NON-HIERARCHICAL CLASSIFICATION

FUZZY CLUSTERING



The size of the section represent the “degree of memberships “ to on of the preestablished 4 groups

So far so good?

Any questions?

Ready to finish?

IN SUMMARY…

- Classification provides an approach to simplify multivariate data description.
- Classification is about **defining discrete groups of objects**.
  - Grouping is based on the (di)similarity between objects
  - Once you have a (di)similarity matrix you can build a **Hierarchic** or **Non-hierarchic** groping that shows the (di)similarity of the evaluated objects.
  - Trees have no order and similarity between objects points can only interpreted through their common root
  - **Non-hierarchic** approaches are Euclidean in nature!!!

AARHUS
UNIVERSITY