FIT3152 Data analytics Lecture 5

Cluster analysis

- Supervised vs Unsupervised learning
- k-Means Clustering
- Hierarchical Clustering
- Cluster analysis in R

- Notes:
  > Lecture recording finishes at 12:52pm.

  > Good Friday tutorials held during week. See unit info.

Consultations

Consultations have commenced.

Most are on Zoom.

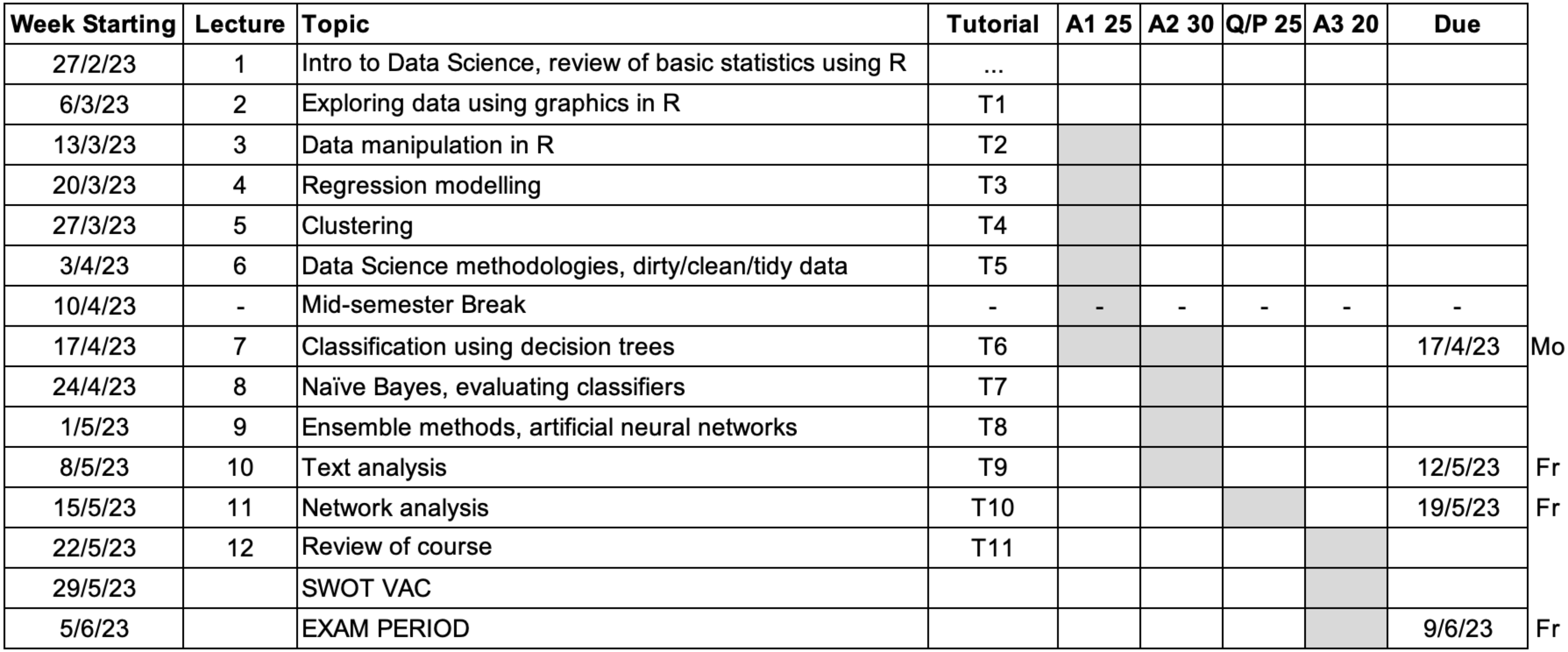
Check Moodle for days/times:

https://lms.monash.edu/course/view.php?id=153815&section=2

Unit outline (week-by-week)

Clayton lecture is Wednesday 11:00am – 1:00pm (AEDT).

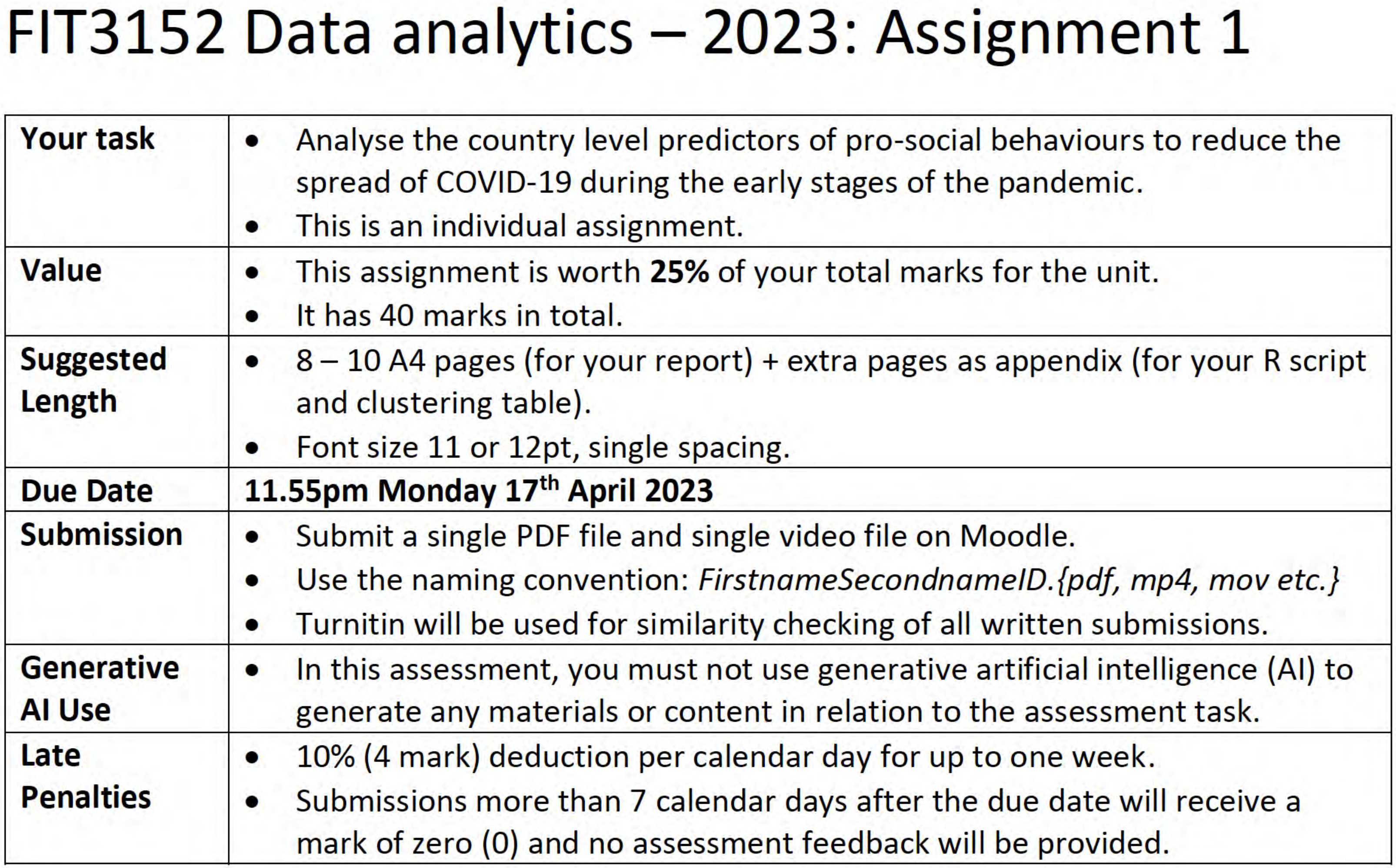
Tutorials begin Week 2 and follow lecture by a week.



Week Starting	Lecture	Topic	Tutorial	A1 25	A2 30	Q/P 25	A3 20	Due	
27/2/23	1	Intro to Data Science, review of basic statistics using R	...						
6/3/23	2	Exploring data using graphics in R	T1						
13/3/23	3	Data manipulation in R	T2	▨					
20/3/23	4	Regression modelling	T3	▨					
27/3/23	5	Clustering	T4	▨					
3/4/23	6	Data Science methodologies, dirty/clean/tidy data	T5	▨					
10/4/23	-	Mid-semester Break	-	▨ -	-	-	-	-	
17/4/23	7	Classification using decision trees	T6	▨	▨			17/4/23	Mo
24/4/23	8	Naïve Bayes, evaluating classifiers	T7		▨				
1/5/23	9	Ensemble methods, artificial neural networks	T8		▨				
8/5/23	10	Text analysis	T9		▨			12/5/23	Fr
15/5/23	11	Network analysis	T10			▨		19/5/23	Fr
22/5/23	12	Review of course	T11				▨		
29/5/23		SWOT VAC					▨		
5/6/23		EXAM PERIOD					▨	9/6/23	Fr

Assignment 1

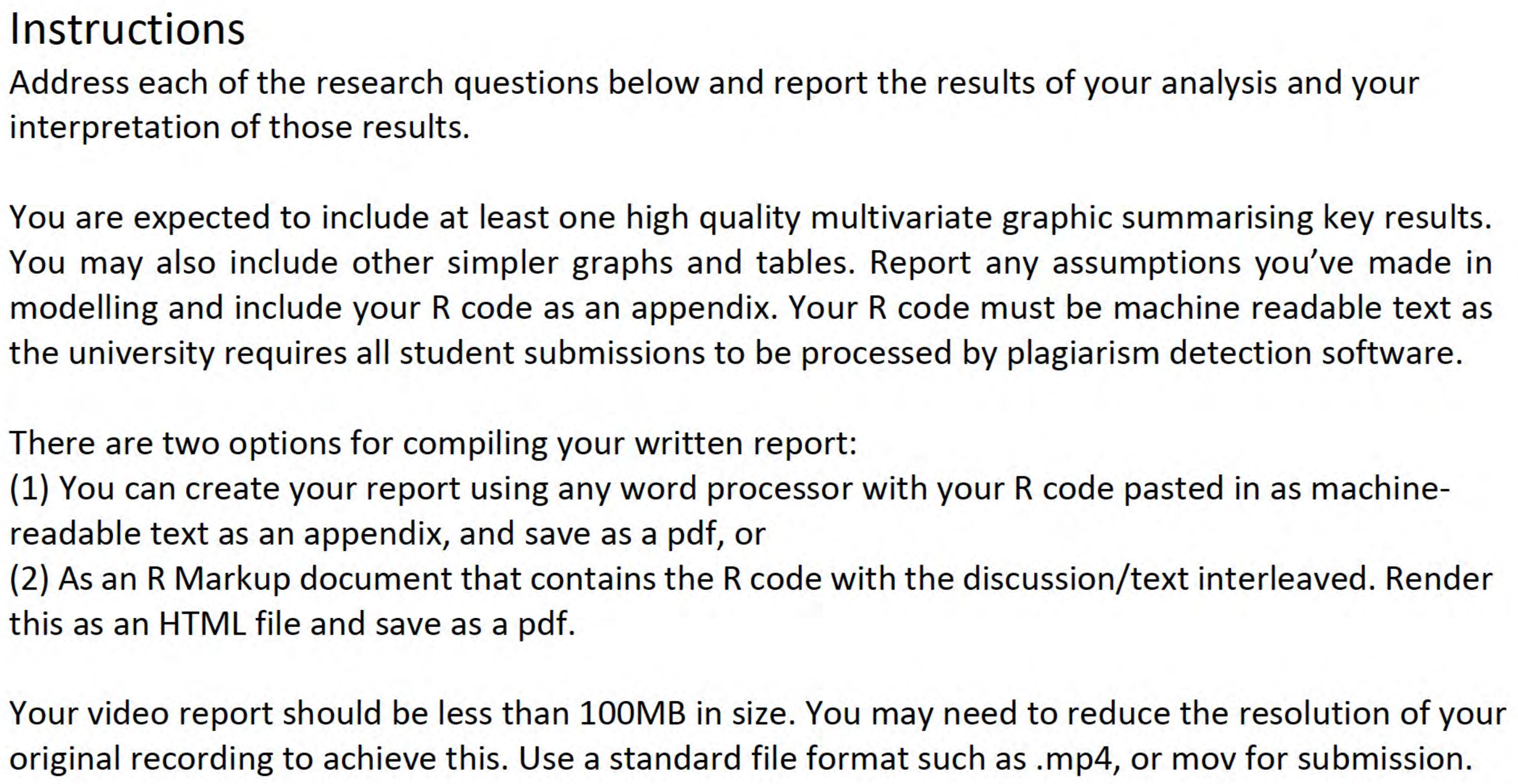
Assignment 1: Summary



FIT3152 Data analytics – 2023: Assignment 1

Your task	• Analyse the country level predictors of pro-social behaviours to reduce the spread of COVID-19 during the early stages of the pandemic. • This is an individual assignment.
Value	• This assignment is worth **25%** of your total marks for the unit. • It has 40 marks in total.
Suggested Length	• 8 – 10 A4 pages (for your report) + extra pages as appendix (for your R script and clustering table). • Font size 11 or 12pt, single spacing.
Due Date	**11.55pm Monday 17th April 2023**
Submission	• Submit a single PDF file and single video file on Moodle. • Use the naming convention: *FirstnameSecondnameID.{pdf, mp4, mov etc.}* • Turnitin will be used for similarity checking of all written submissions.
Generative AI Use	• In this assessment, you must not use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.
Late Penalties	• 10% (4 mark) deduction per calendar day for up to one week. • Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Assignment 1: Instructions



Instructions

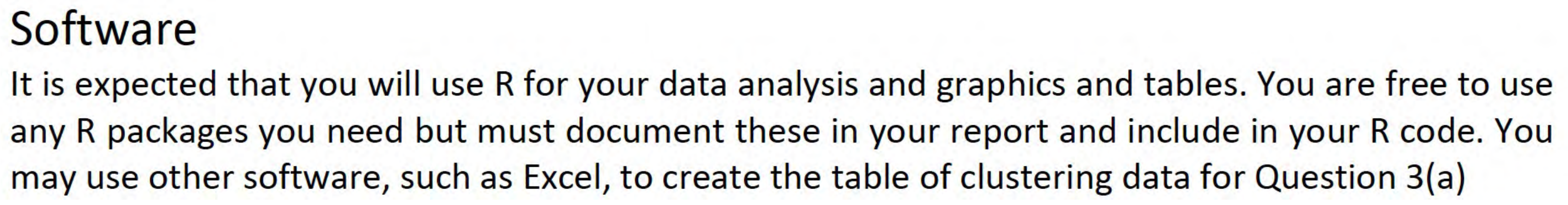
Address each of the research questions below and report the results of your analysis and your interpretation of those results.

You are expected to include at least one high quality multivariate graphic summarising key results. You may also include other simpler graphs and tables. Report any assumptions you've made in modelling and include your R code as an appendix. Your R code must be machine readable text as the university requires all student submissions to be processed by plagiarism detection software.

There are two options for compiling your written report:
(1) You can create your report using any word processor with your R code pasted in as machine-readable text as an appendix, and save as a pdf, or
(2) As an R Markup document that contains the R code with the discussion/text interleaved. Render this as an HTML file and save as a pdf.

Your video report should be less than 100MB in size. You may need to reduce the resolution of your original recording to achieve this. Use a standard file format such as .mp4, or mov for submission.

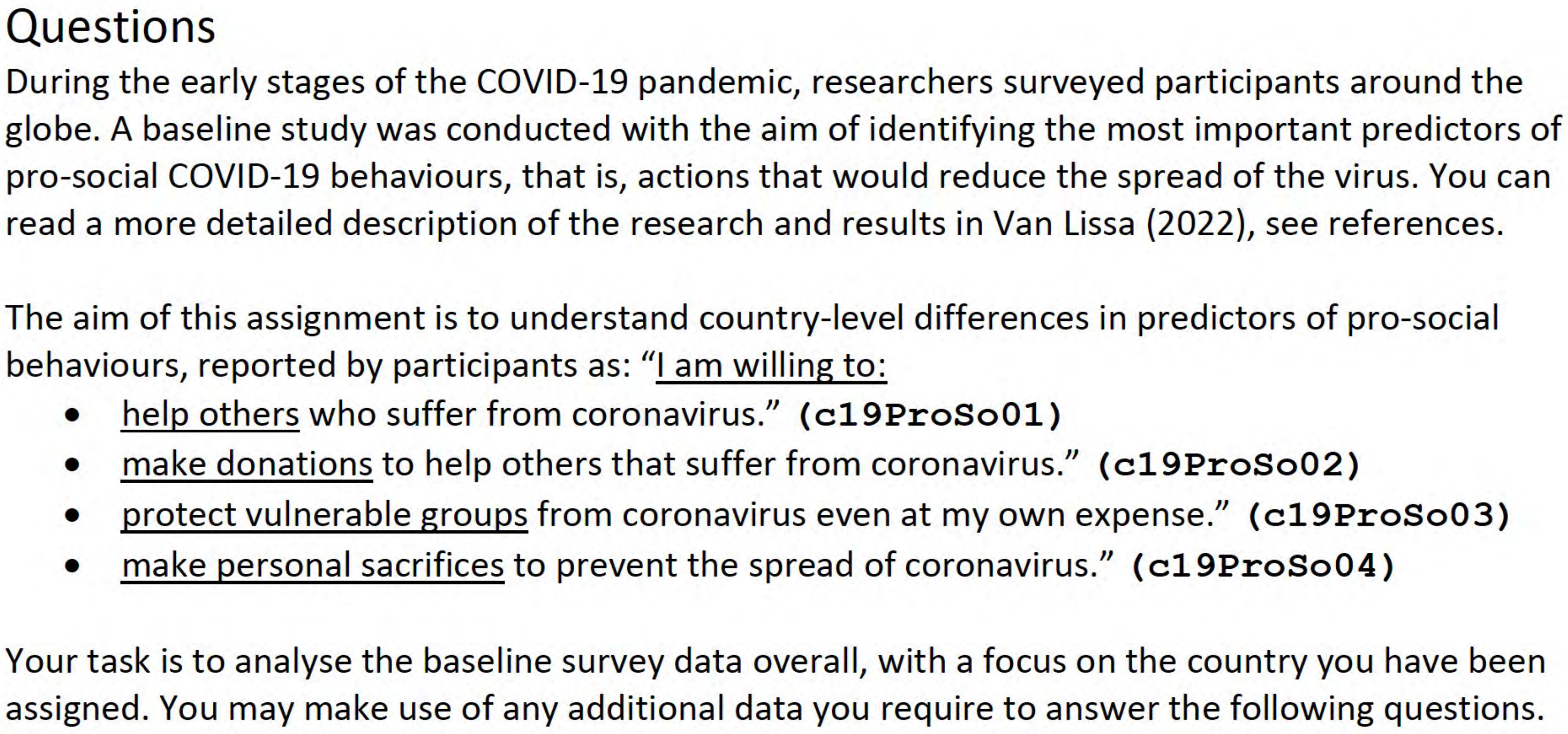
Assignment 1: Software



Software

It is expected that you will use R for your data analysis and graphics and tables. You are free to use any R packages you need but must document these in your report and include in your R code. You may use other software, such as Excel, to create the table of clustering data for Question 3(a)

Assignment 1: Questions:



Questions

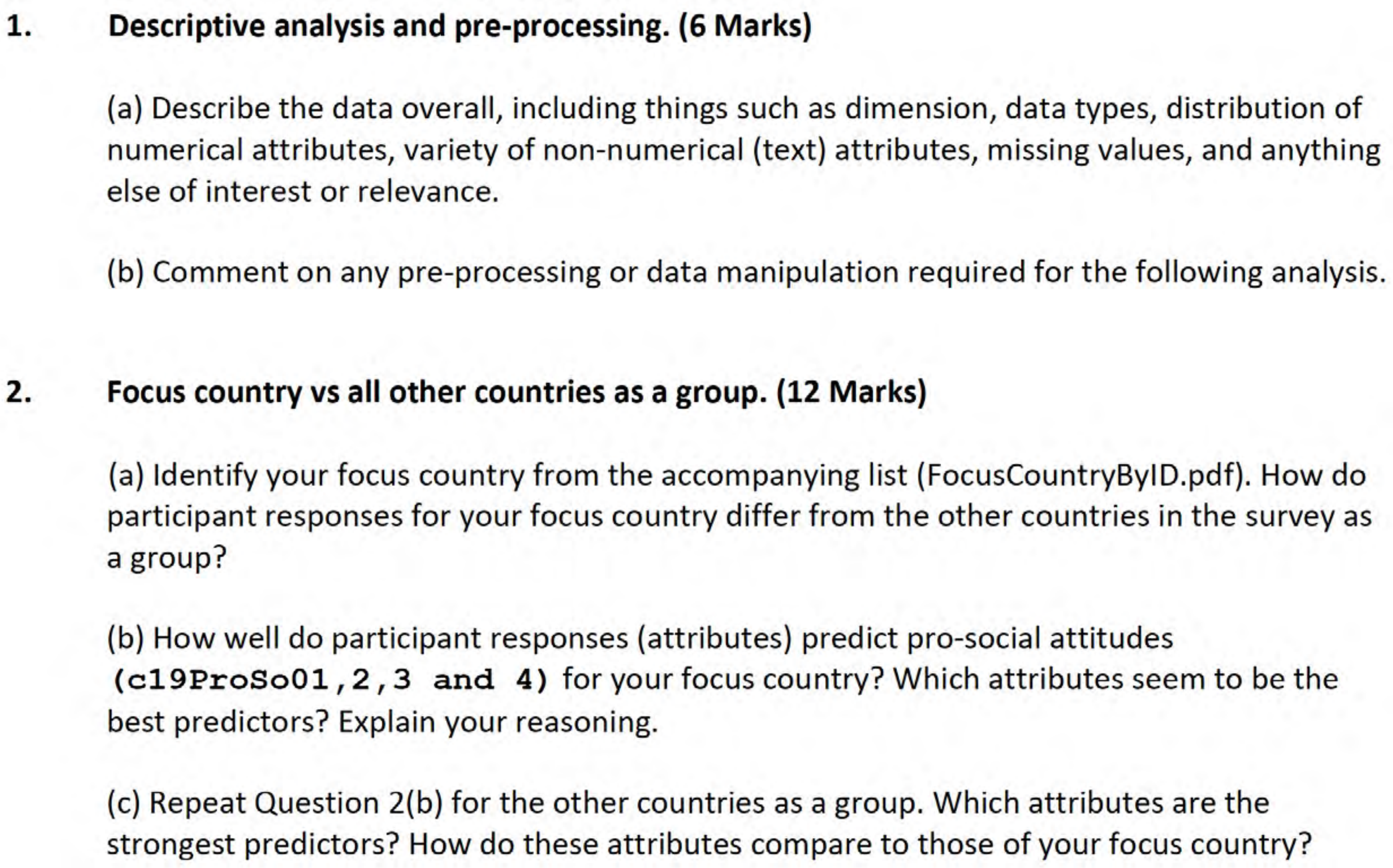
During the early stages of the COVID-19 pandemic, researchers surveyed participants around the globe. A baseline study was conducted with the aim of identifying the most important predictors of pro-social COVID-19 behaviours, that is, actions that would reduce the spread of the virus. You can read a more detailed description of the research and results in Van Lissa (2022), see references.

The aim of this assignment is to understand country-level differences in predictors of pro-social behaviours, reported by participants as: “I am willing to:

- help others who suffer from coronavirus.” **(c19ProSo01)**
- make donations to help others that suffer from coronavirus.” **(c19ProSo02)**
- protect vulnerable groups from coronavirus even at my own expense.” **(c19ProSo03)**
- make personal sacrifices to prevent the spread of coronavirus.” **(c19ProSo04)**

Your task is to analyse the baseline survey data overall, with a focus on the country you have been assigned. You may make use of any additional data you require to answer the following questions.

Assignment 1: Questions 1 & 2



1. **Descriptive analysis and pre-processing. (6 Marks)**

   (a) Describe the data overall, including things such as dimension, data types, distribution of numerical attributes, variety of non-numerical (text) attributes, missing values, and anything else of interest or relevance.

   (b) Comment on any pre-processing or data manipulation required for the following analysis.

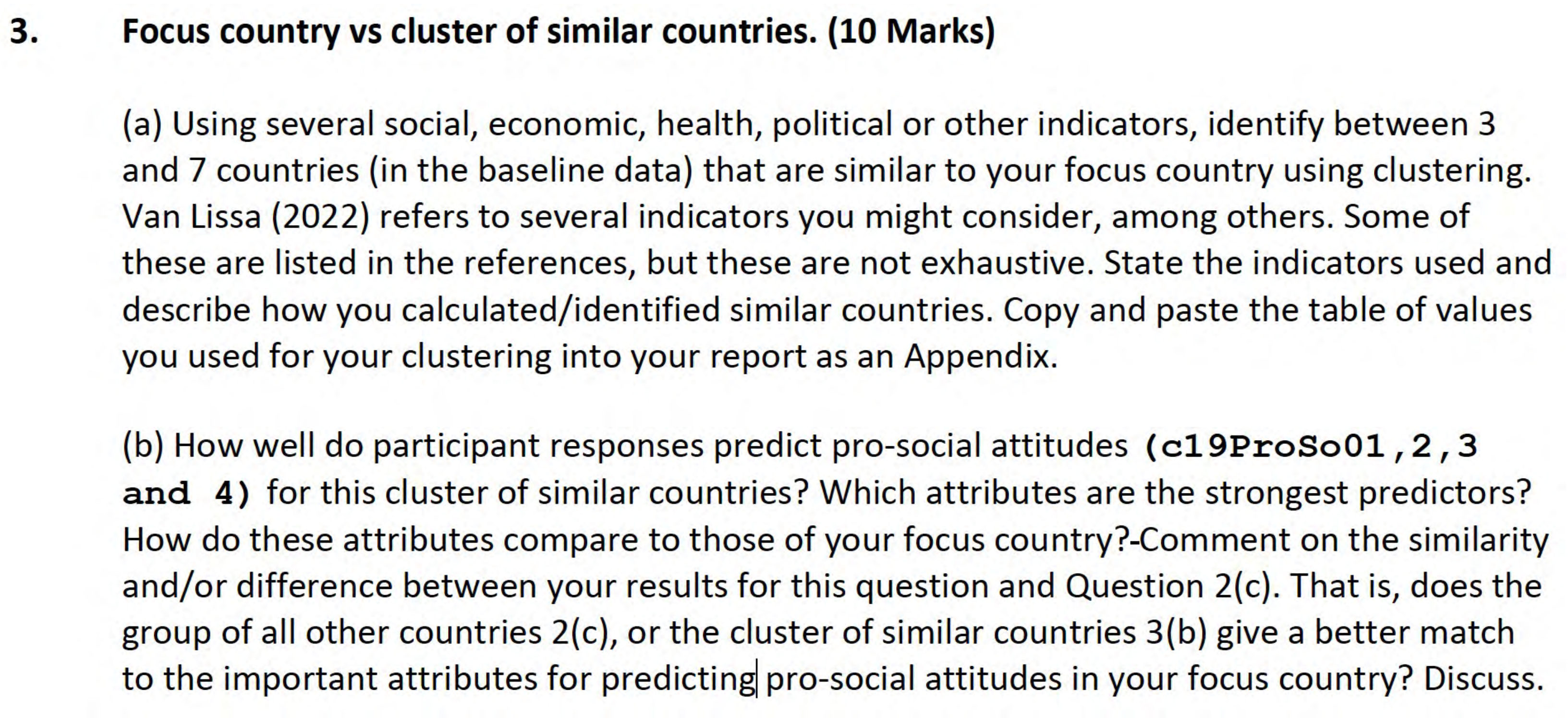
2. **Focus country vs all other countries as a group. (12 Marks)**

   (a) Identify your focus country from the accompanying list (FocusCountryByID.pdf). How do participant responses for your focus country differ from the other countries in the survey as a group?

   (b) How well do participant responses (attributes) predict pro-social attitudes **`(c19ProSo01,2,3 and 4)`** for your focus country? Which attributes seem to be the best predictors? Explain your reasoning.

   (c) Repeat Question 2(b) for the other countries as a group. Which attributes are the strongest predictors? How do these attributes compare to those of your focus country?

Assignment 1: Question 3

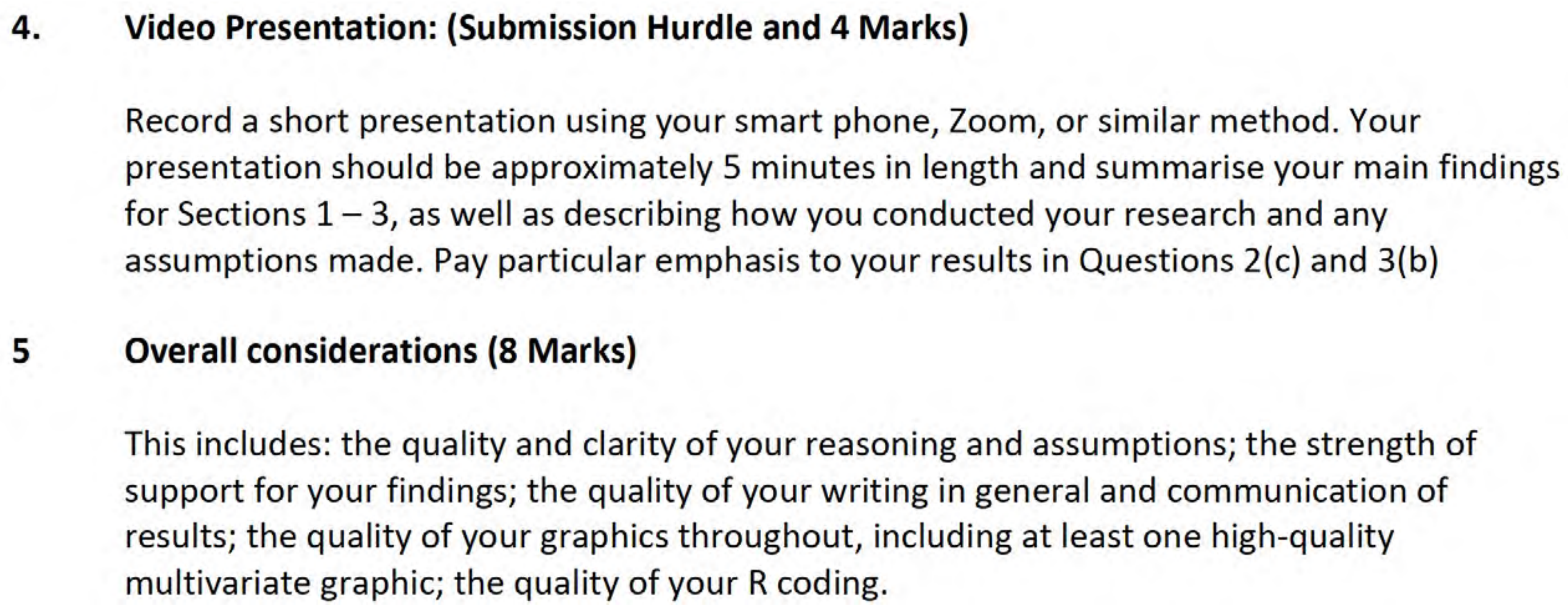


3. **Focus country vs cluster of similar countries. (10 Marks)**

(a) Using several social, economic, health, political or other indicators, identify between 3 and 7 countries (in the baseline data) that are similar to your focus country using clustering. Van Lissa (2022) refers to several indicators you might consider, among others. Some of these are listed in the references, but these are not exhaustive. State the indicators used and describe how you calculated/identified similar countries. Copy and paste the table of values you used for your clustering into your report as an Appendix.

(b) How well do participant responses predict pro-social attitudes **`(c19ProSo01,2,3 and 4)`** for this cluster of similar countries? Which attributes are the strongest predictors? How do these attributes compare to those of your focus country?-Comment on the similarity and/or difference between your results for this question and Question 2(c). That is, does the group of all other countries 2(c), or the cluster of similar countries 3(b) give a better match to the important attributes for predicting pro-social attitudes in your focus country? Discuss.

Assignment 1: Presentation/Overall



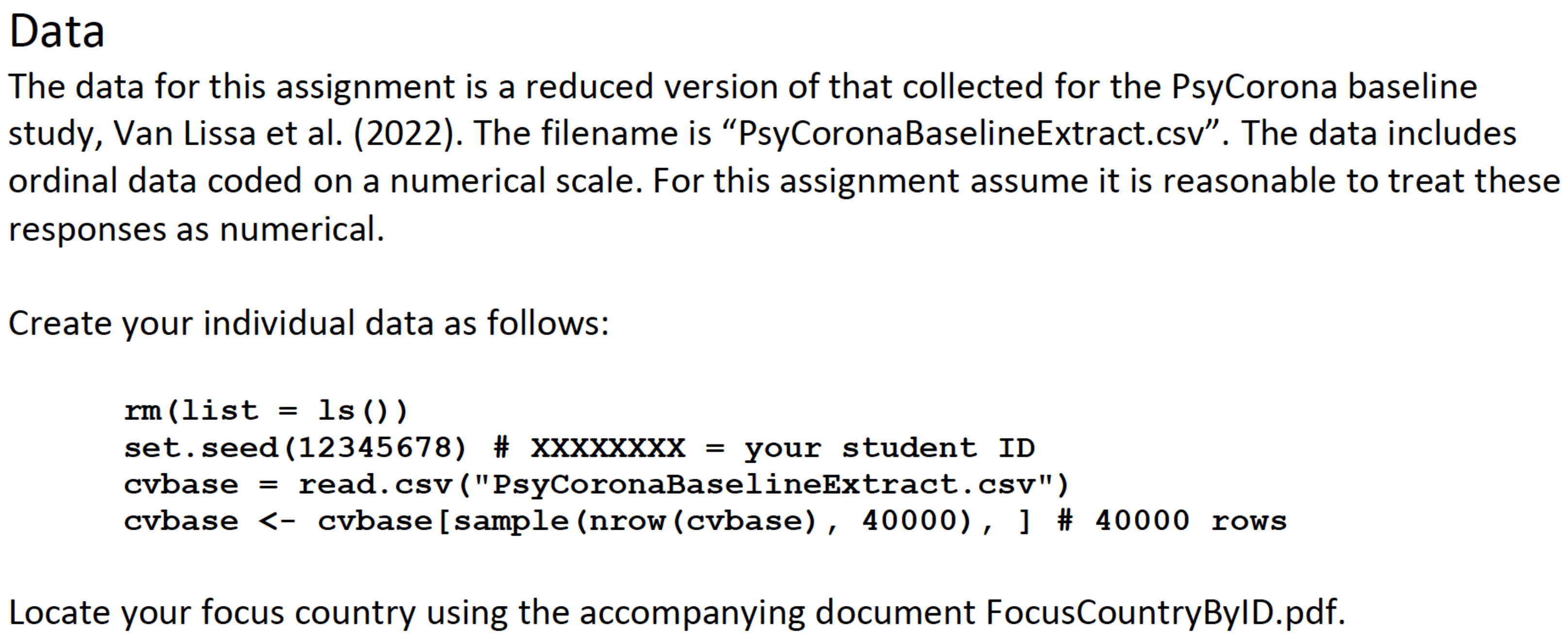
4. Video Presentation: (Submission Hurdle and 4 Marks)

Record a short presentation using your smart phone, Zoom, or similar method. Your presentation should be approximately 5 minutes in length and summarise your main findings for Sections 1 – 3, as well as describing how you conducted your research and any assumptions made. Pay particular emphasis to your results in Questions 2(c) and 3(b)

5 Overall considerations (8 Marks)

This includes: the quality and clarity of your reasoning and assumptions; the strength of support for your findings; the quality of your writing in general and communication of results; the quality of your graphics throughout, including at least one high-quality multivariate graphic; the quality of your R coding.

Assignment 1: Data generation



Data

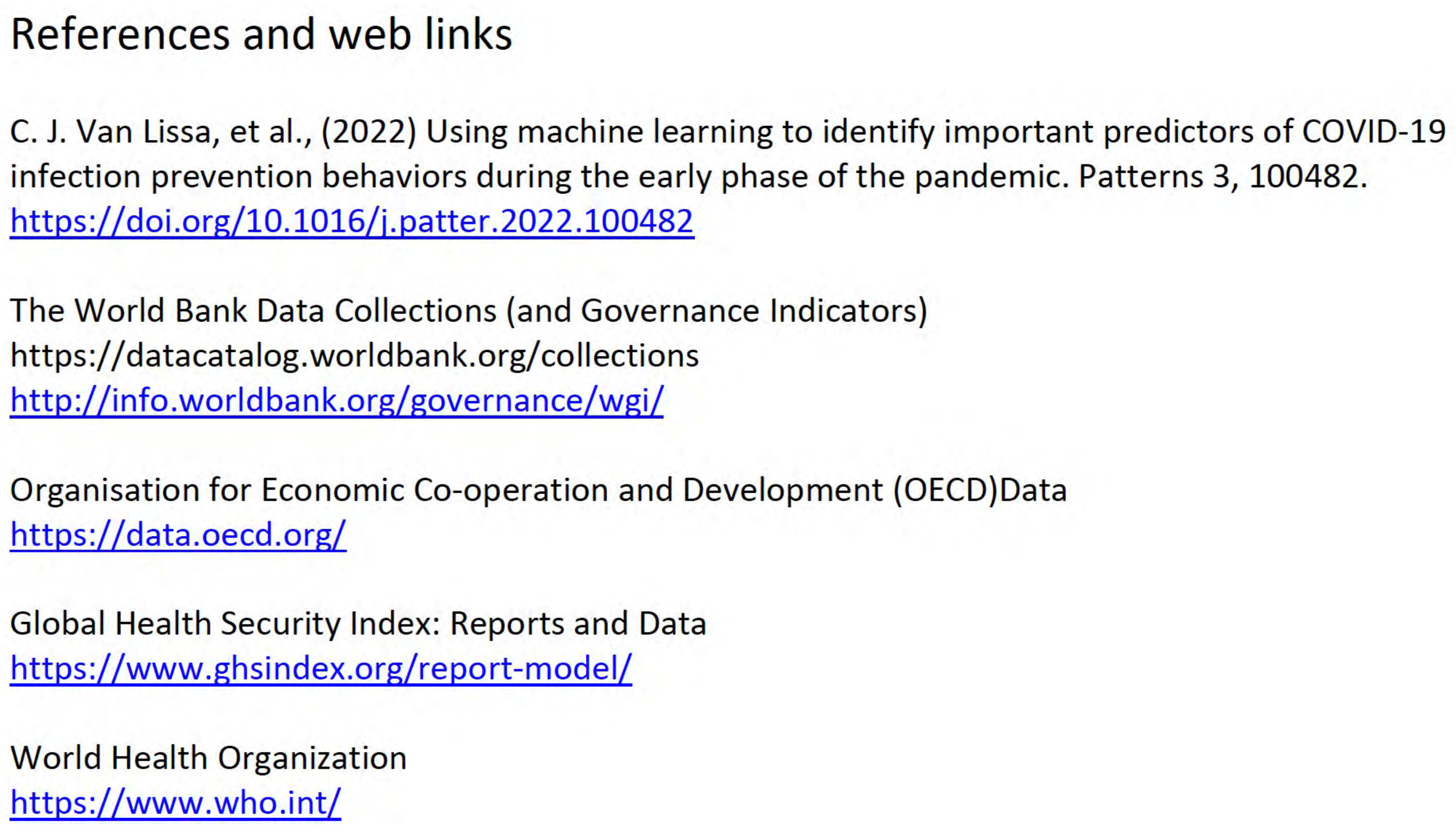
The data for this assignment is a reduced version of that collected for the PsyCorona baseline study, Van Lissa et al. (2022). The filename is “PsyCoronaBaselineExtract.csv”. The data includes ordinal data coded on a numerical scale. For this assignment assume it is reasonable to treat these responses as numerical.

Create your individual data as follows:

```
rm(list = ls())
set.seed(12345678) # XXXXXXXX = your student ID
cvbase = read.csv("PsyCoronaBaselineExtract.csv")
cvbase <- cvbase[sample(nrow(cvbase), 40000), ] # 40000 rows
```

Locate your focus country using the accompanying document FocusCountryByID.pdf.

Assignment 1: Selected references



References and web links

C. J. Van Lissa, et al., (2022) Using machine learning to identify important predictors of COVID-19 infection prevention behaviors during the early phase of the pandemic. Patterns 3, 100482.
https://doi.org/10.1016/j.patter.2022.100482

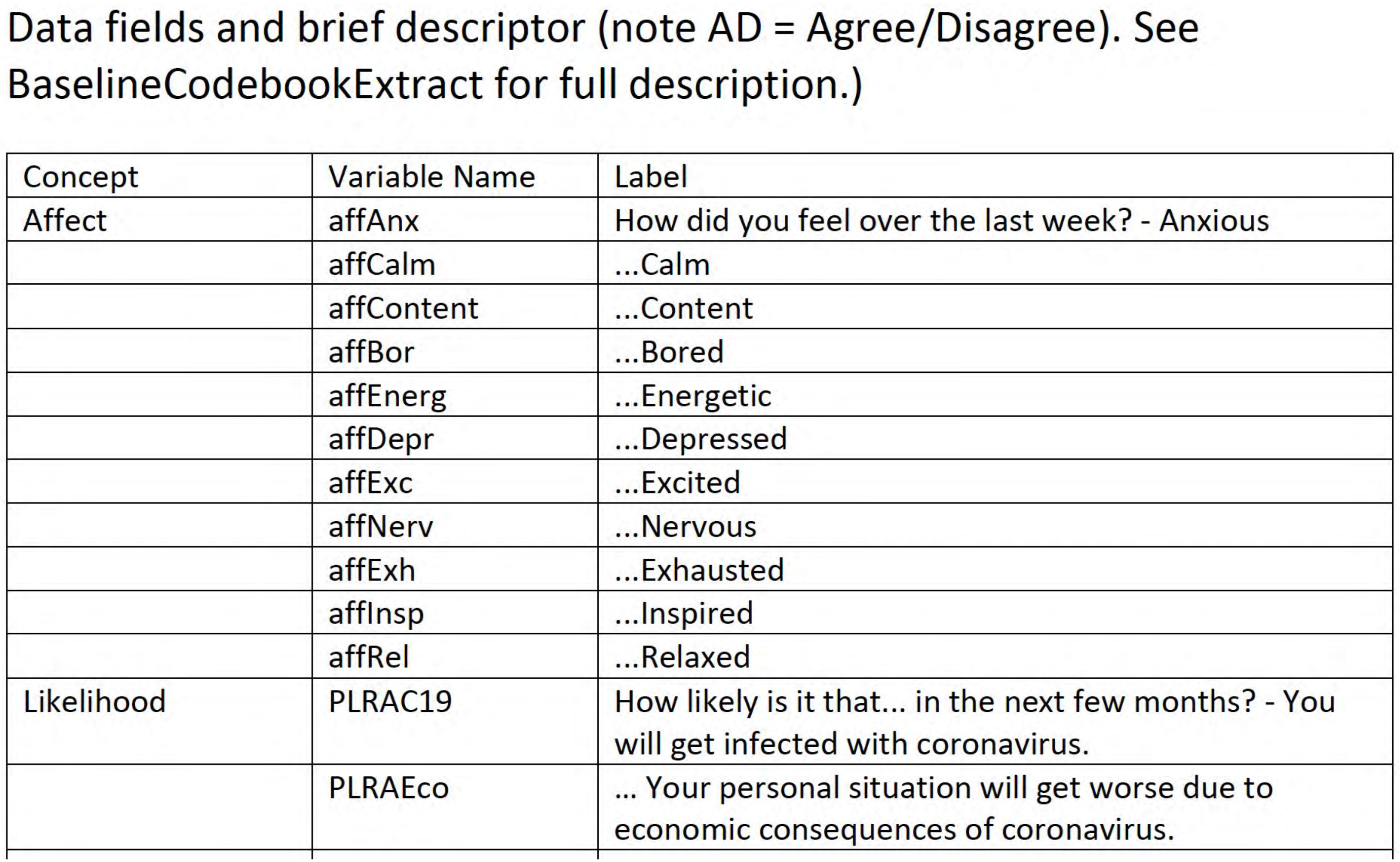
The World Bank Data Collections (and Governance Indicators)
https://datacatalog.worldbank.org/collections
http://info.worldbank.org/governance/wgi/

Organisation for Economic Co-operation and Development (OECD)Data
https://data.oecd.org/

Global Health Security Index: Reports and Data
https://www.ghsindex.org/report-model/

World Health Organization
https://www.who.int/

Assignment 1: Code Book extract



Data fields and brief descriptor (note AD = Agree/Disagree). See BaselineCodebookExtract for full description.)

Concept	Variable Name	Label
Affect	affAnx	How did you feel over the last week? - Anxious
	affCalm	...Calm
	affContent	...Content
	affBor	...Bored
	affEnerg	...Energetic
	affDepr	...Depressed
	affExc	...Excited
	affNerv	...Nervous
	affExh	...Exhausted
	affInsp	...Inspired
	affRel	...Relaxed
Likelihood	PLRAC19	How likely is it that... in the next few months? - You will get infected with coronavirus.
	PLRAEco	... Your personal situation will get worse due to economic consequences of coronavirus.

Response to student questions

- Are we expected to clean our data? For example, could we could remove rows where the participant fails to answer a question?
  > Yes you can although that is really more like pre-processing the data (to remove NAs, for example). Identifying and identifying issues like this is part of the assignment.

Response to student questions

- Should we be looking closer at the answers to ensure that responses don't contradict each other: (e.g., can a person strongly agree with both statements: I am sure I can keep my job and I feel insecure about the future of my job)?
  - > Not so important, but you could investigate and report things like this if you were interested and had time.

Response to student questions

- When I clean the data to remove NAs , this cuts my data from 40,000 rows to roughly 30,000. Is this ok?
  > This is fine.
  >
  > You might also drop some columns (attributes) if you find they have too many NAs to work with.
  >
  > You can also work with individual columns, or pairs of columns, separate from the main data file if you need to remove NA.
  >
  > It is not necessary to analyse all attributes if you give a good reason for not doing so.

Response to student questions

- There seem to be many ways to approach the assignment. Are specific methods required or is it enough to justify our approach?
  - > There are many different ways to approach the assignment. There is not one single right answer. We are not looking for everyone to do the same thing, but some tasks will use a similar approaches because they are the best method for the task.

```
DFNew = DF[(coded_country == Japan),]
      = DF (Cnd_country != Japan ),]
```

Response to student questions

- When describing the distribution of numerical attributes in Question 1 (a), would my approach of using boxplots be penalized if I don't include other methods?
  - > Using a boxplot to show the distribution of a numerical attribute is a correct approach, although not the only way to show a distribution.

Response to student questions

- How do we know what the numbers are referring to if we don’t have a key?
  > The PsyCoronaBaselineCodebookExtract.pdf should give you some guidance.
- For example, with variable "affAnx" do we assume the lower numbers correspond to lower levels of anxiousness?
  > When the results are a positive scale then assume that numbers show the degree.

Response to student questions

- For variables that have an Agree/Disagree response, what number corresponds to Strongly disagree /Disagree /Neither agree nor disagree /Agree /Strongly agree?
  > When there is an agree/disagree response you can assume the negative values show disagreement and positive values show agreement, and the number shows degree.

Response to student questions

- For Part 2 of the assignment where we have to compare the responses between our focus country and the rest, do we have to compare all the attributes or just a few that we choose to compare?
  > You are expected you to consider the majority of attributes. However if there were some that it did not make sense to include, or you were unable to include for certain reasons (for example if there were too many missing values, or not enough data etc.) then it would be reasonable to exclude them.

Review questions from last lecture

mtcars

The (inbuilt) data set Motor Trend Car Road Tests gives summary statistics including fuel usage (mpg), engine size (disp), number of cylinders (cyl), power (hp), body weight (wt) and the number of gears for a variety of cars.

How well do these variables predict fuel economy?

Summary data

> head(mtcars)

```
                   mpg cyl disp  hp drat   wt qsec vs am gear
Mazda RX4         21.0   6  160 110 3.90 2.62 16.5  0  1    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.88 17.0  0  1    4
Datsun 710        22.8   4  108  93 3.85 2.32 18.6  1  1    4
Hornet 4 Drive    21.4   6  258 110 3.08 3.21 19.4  1  0    3
Hornet Sportabout 18.7   8  360 175 3.15 3.44 17.0  0  0    3
Valiant           18.1   6  225 105 2.76 3.46 20.2  1  0    3
```

Model

```
>   attach(mtcars)
>   fitted = lm(mpg ~ cyl + disp + hp + wt + gear)
>   fitted
```

```
Call:
lm(formula = mpg ~ cyl + disp + hp + wt + gear)

Coefficients:
(Intercept)          cyl         disp           hp           wt
    37.3626      -1.1186       0.0138      -0.0279      -3.7143
       gear
     0.6788
```

Summary (a)

```
Call:
lm(formula = mpg ~ cyl + disp + hp + wt + gear)

Residuals:
   Min     1Q Median     3Q    Max
-3.223 -1.686 -0.383  1.293  5.943
```

Summary (b)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.3626     5.9725    6.26  1.3e-06 ***
cyl          -1.1186     0.7144   -1.57   0.1295
disp          0.0138     0.0123    1.12   0.2727
hp           -0.0279     0.0166   -1.68   0.1052
wt           -3.7143     1.0482   -3.54   0.0015 **
gear          0.6788     1.0345    0.66   0.5175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 2.54 on 26 degrees of freedom
Multiple R-squared:  0.851,   Adjusted R-squared:  0.822
F-statistic: 29.7 on 5 and 26 DF,  p-value: 5.72e-10
```

Question 1

Ignoring the constant term, the most reliable predictor of fuel economy (mpg) is:

```
                         Estimate Std. Error t value Pr(>|t|)
           (Intercept)  37.3626     5.9725    6.26  1.3e-06 ***
(a)        cyl          -1.1186     0.7144   -1.57   0.1295
(b)        disp          0.0138     0.0123    1.12   0.2727
(c)        hp           -0.0279     0.0166   -1.68   0.1052
(d)        wt           -3.7143     1.0482   -3.54   0.0015 **
(e)        gear          0.6788     1.0345    0.66   0.5175
```

Question 2

The least reliable predictor of fuel economy is:

```
                  Estimate Std. Error t value Pr(>|t|)
      (Intercept)  37.3626     5.9725    6.26  1.3e-06 ***
(a)   cyl          -1.1186     0.7144   -1.57   0.1295
(b)   disp          0.0138     0.0123    1.12   0.2727
(c)   hp           -0.0279     0.0166   -1.68   0.1052
(d)   wt           -3.7143     1.0482   -3.54   0.0015 **
(e)   gear          0.6788     1.0345    0.66   0.5175
```

Question 3

Cars with more gears are more economical:

A. True

B. False

C. Can’t tell

```
                       Estimate  Std. Error  t value  Pr(>|t|)
         (Intercept)   37.3626     5.9725      6.26   1.3e-06 ***
(a)      cyl           -1.1186     0.7144     -1.57    0.1295
(b)      disp           0.0138     0.0123      1.12    0.2727
(c)      hp            -0.0279     0.0166     -1.68    0.1052
(d)      wt            -3.7143     1.0482     -3.54    0.0015 **
(e)      gear           0.6788     1.0345      0.66    0.5175
```

Question 4

Heaver cars are less economical:

A. True

B. False

C. Can’t tell

```
                     Estimate Std. Error t value Pr(>|t|)
       (Intercept)  37.3626     5.9725    6.26  1.3e-06 ***
(a)    cyl          -1.1186     0.7144   -1.57   0.1295
(b)    disp          0.0138     0.0123    1.12   0.2727
(c)    hp           -0.0279     0.0166   -1.68   0.1052
(d)    wt           -3.7143     1.0482   -3.54   0.0015 **
(e)    gear          0.6788     1.0345    0.66   0.5175
```

Question 5

Overall, the predictive power of the model is high:

A. True (better than 70%)

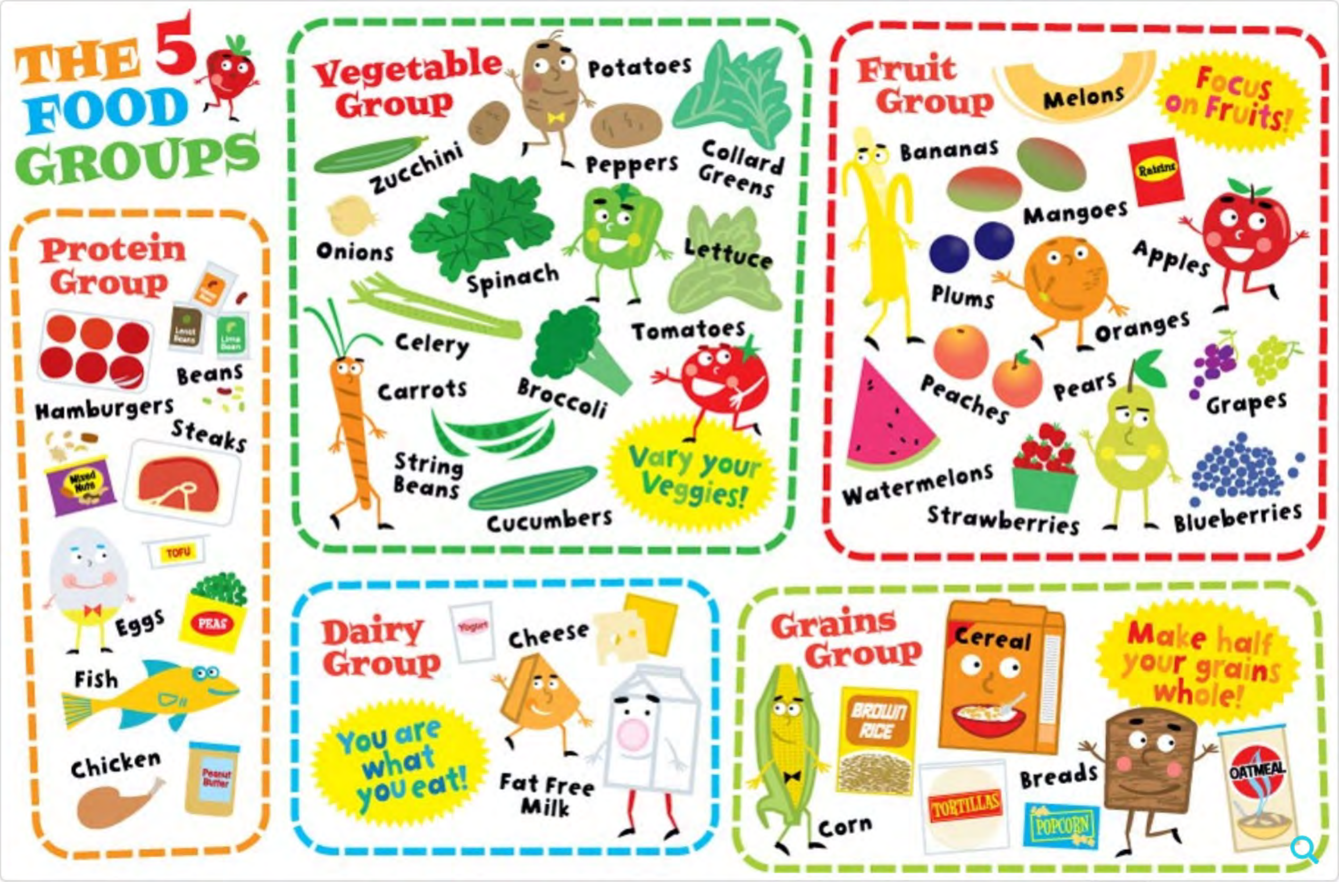
B. False (worse than 30%)

C. Can’t tell

```
Residual standard error: 2.54 on 26 degrees of freedom
Multiple R-squared:  0.851,   Adjusted R-squared:  0.822
F-statistic: 29.7 on 5 and 26 DF,  p-value: 5.72e-10
```

Cluster analysis

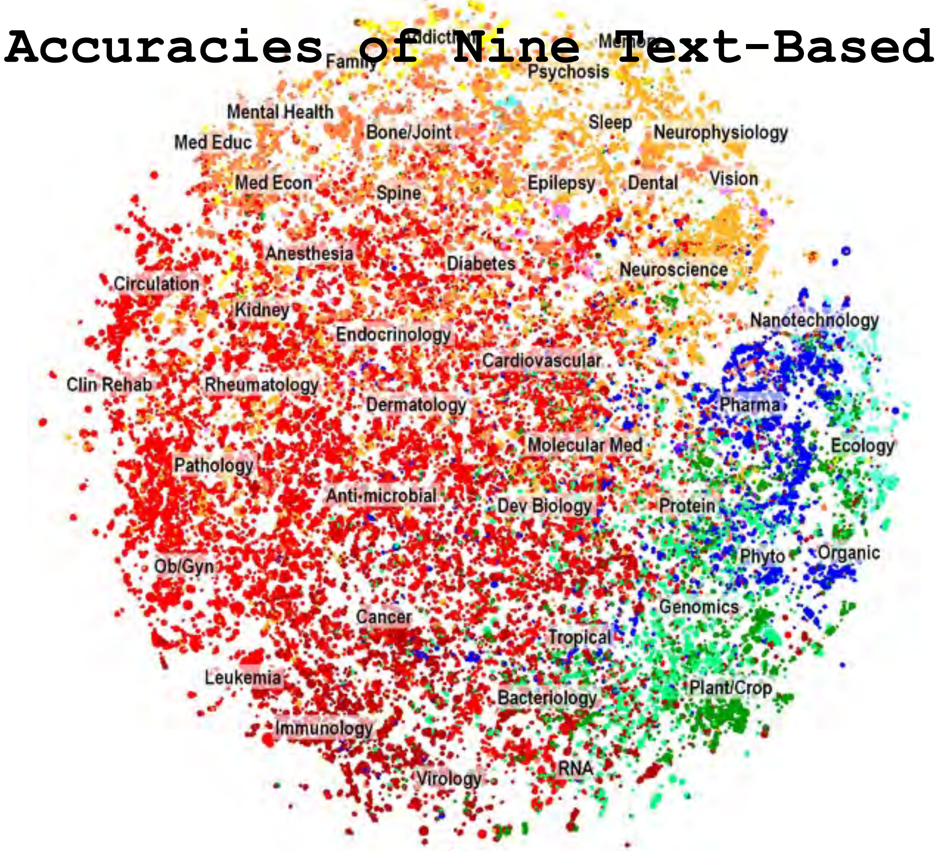
Food groups



https://www.muralsyourway.com/p/food-groups-mural/

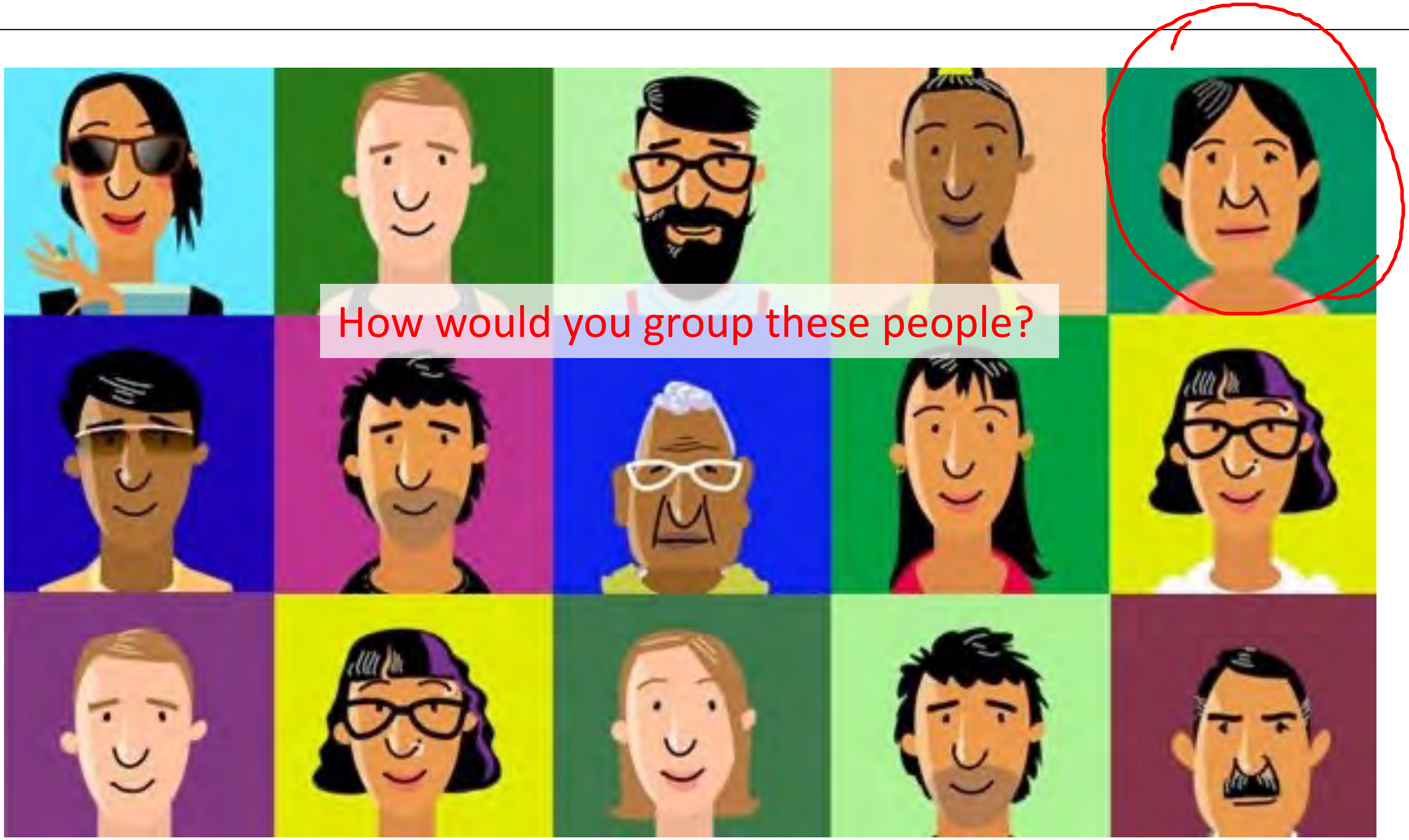
Document clustering

Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018029

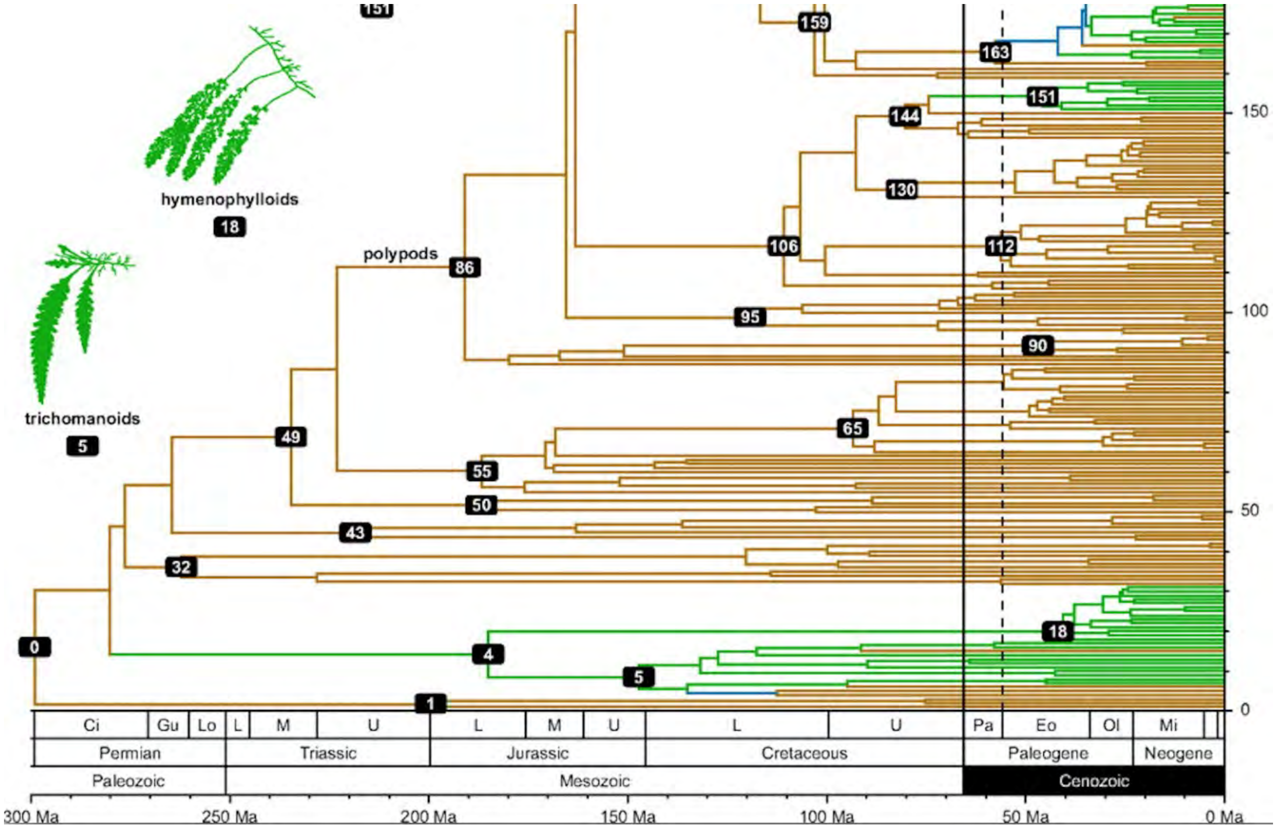
7 Australian political personas



How would you group these people?

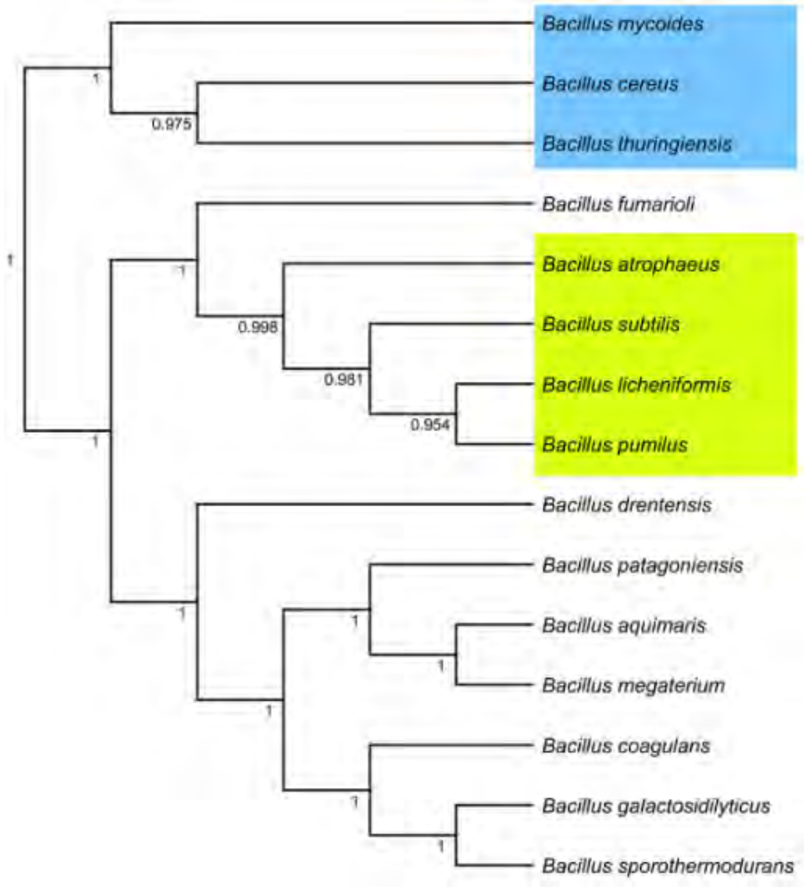
https://www.smh.com.au/

Phylogenetic tree, fern evolution



https://www.pnas.org/content/106/27/11200/F1.expansion.html

Phylogenetic tree, Bacillus species



https://openi.nlm.nih.gov/detailedresult.php?img=PMC2828439_1471-2105-11-69-1&req=4

COVID-19

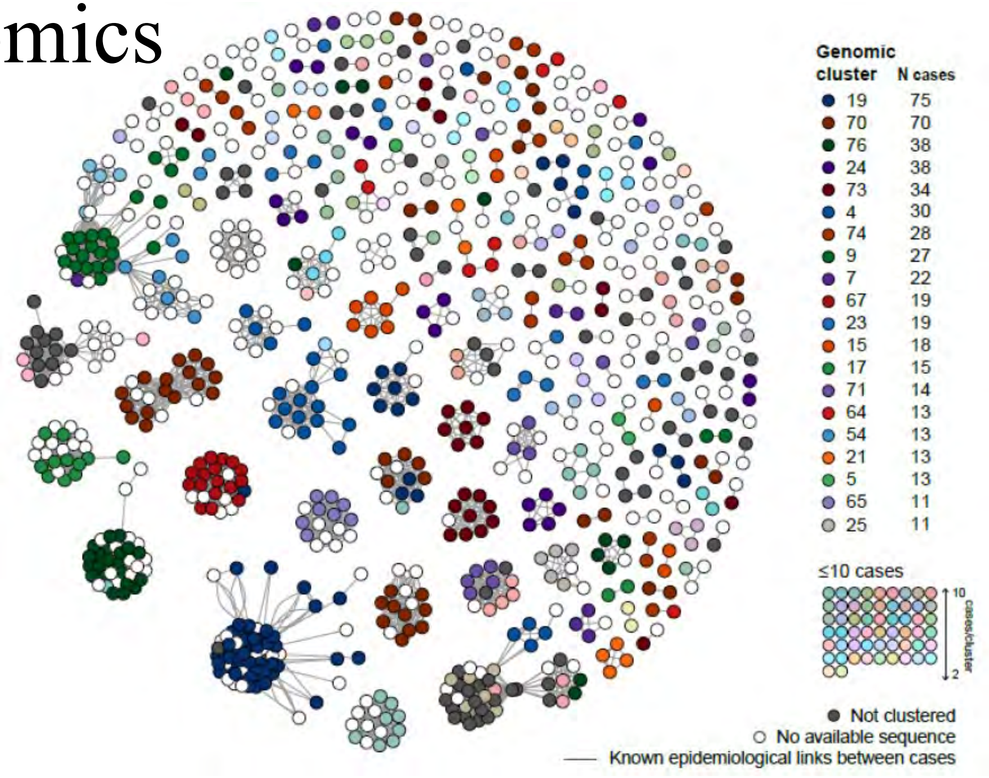
Tracking the COVID-19 pandemic in Australia using genomics

> **Sequenced samples from Australia were representative of the global diversity of SARS-CoV-2, ... In total, 76 distinct genomic clusters were identified; these included large clusters associated with social venues, healthcare facilities and cruise ships. Sequencing of sequential samples from 98 patients revealed minimal intra-patient SARS-CoV-2 genomic diversity.**

https://www.medrxiv.org/content/10.1101/2020.05.12.20099929v1

COVID-19

Tracking the COVID-19 pandemic in Australia using genomics



https://www.medrxiv.org/content/10.1101/2020.05.12.20099929v1

SARS-CoV-2 antigenic variants

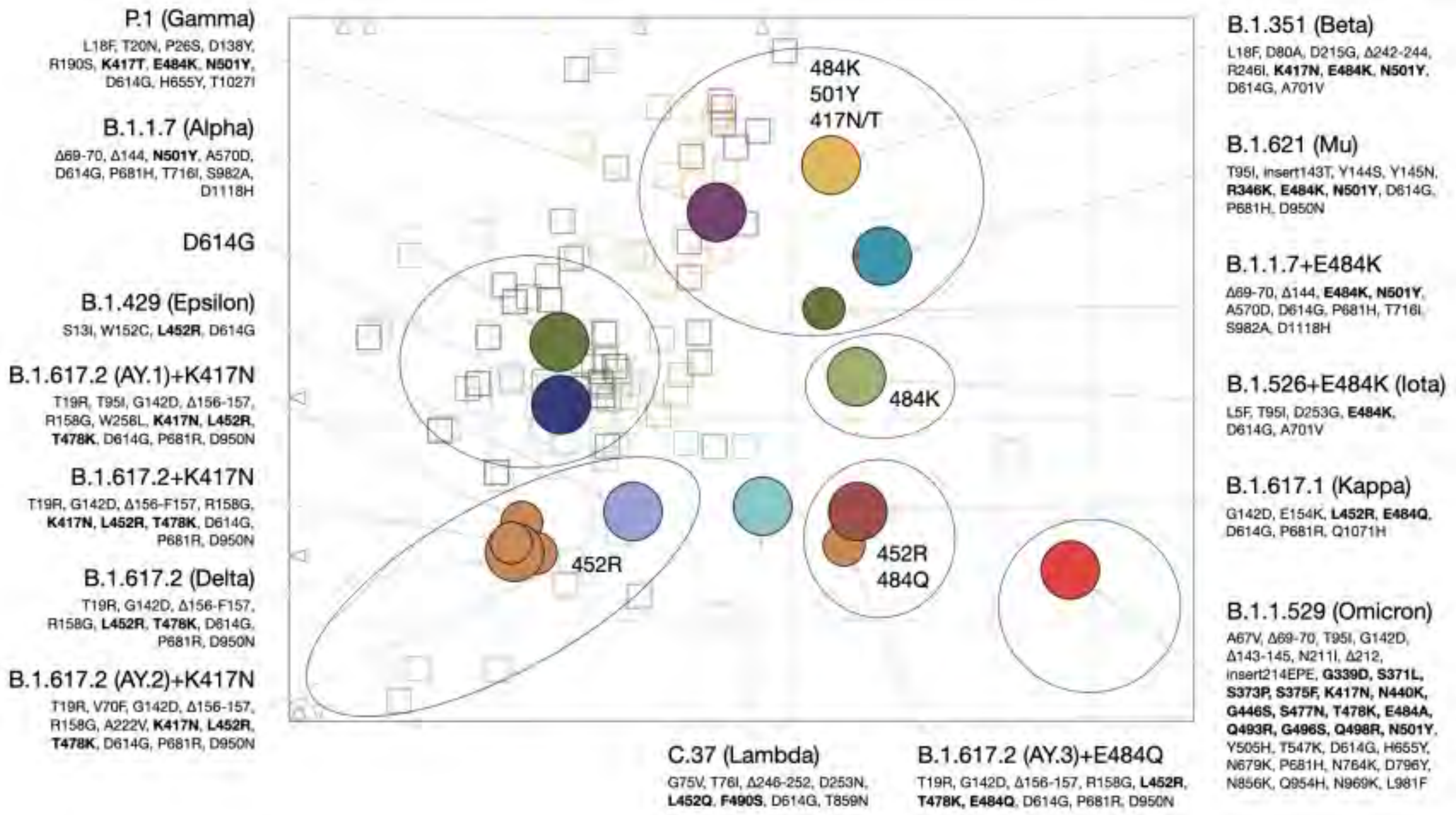


Fig. 2: Antigenic map of SARS-CoV-2 variants and selected substitutions. Variants are shown as circles, sera as squares.

https://www.biorxiv.org/content/10.1101/2022.01.28.477987v1.full.pdf

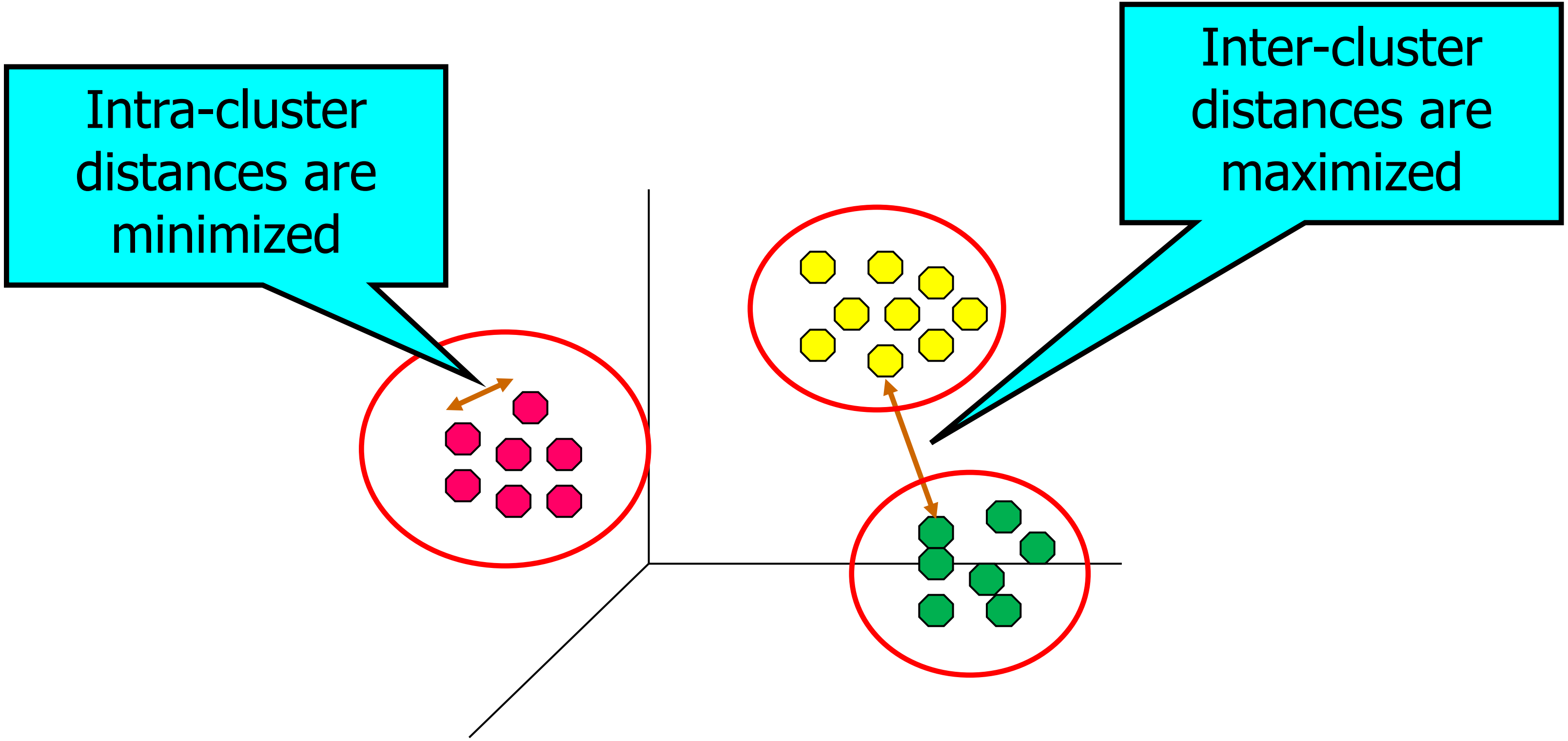
Supervised *vs* unsupervised learning

There are two main approaches to machine learning:

- Supervised learning algorithms:
  - > Algorithms are given labelled examples (target class) for the various types of data that need to be learned.
  - > For example: regression.
- Unsupervised learning algorithms:
  - > Data is unlabeled (has no predefined classes), and the learning algorithms attempt to find patterns within the data to put into groups or sets.
  - > For example, clustering algorithms.

What is Cluster Analysis?

Finding groups of points such that the points in a group will be similar (or related) to one another and different from (or unrelated to) the points in other groups



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Clustering – applications

Examples:

- Segment customer database based on similar buying patterns.
- Group houses in a town into neighborhoods based on similar features.
- Identify similar Internet usage patterns.
- Clustering emails by content.
- Gene clustering in biology.
- Group documents that have similar content.

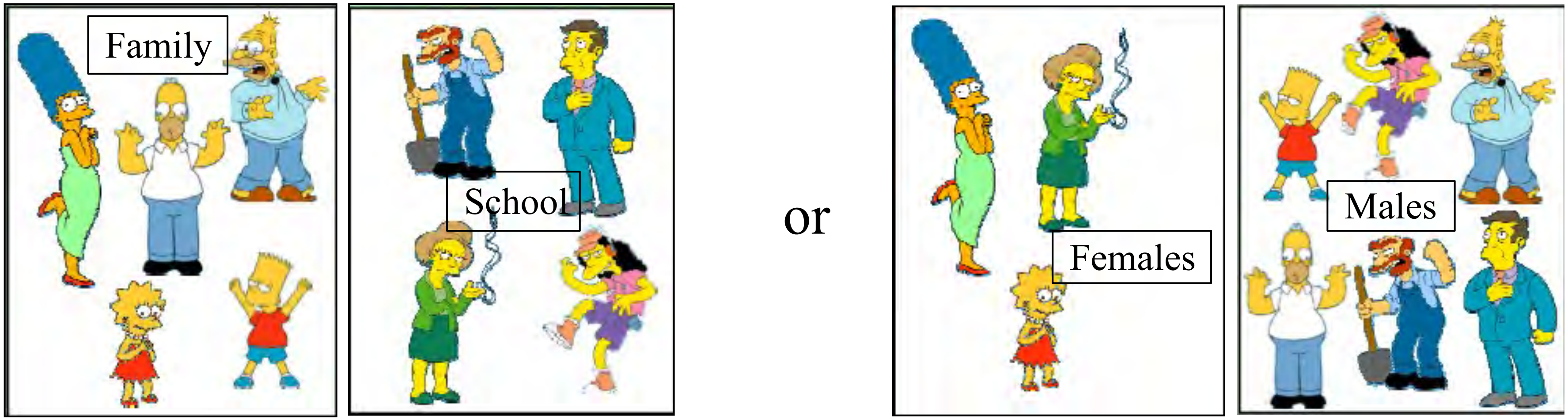
Are these clusters pre-defined?

- No, it depends how the distance between points are measured. There are no class labels.

Illustrating clustering

Are there natural groupings amongst this group?

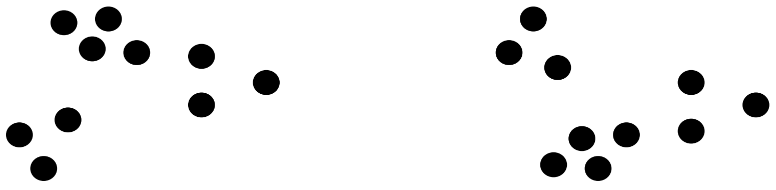
Possible clusters



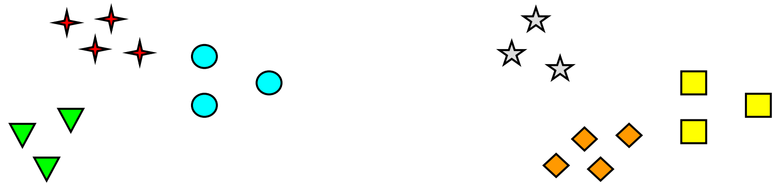
Clustering Definition

- Clustering identifies natural groups in a data set:
  - > Given a set of data points, each having a set of attributes, and a similarity measure, find clusters such that:
  - > Data points in each cluster are more similar to each other.
  - > Data points in separate clusters are less similar.
- Similarity Measures:
  - > Euclidean Distance (e.g., Pythagoras' theorem).
  - > Other distance-based measures (for example, Manhattan).
  - > Other measures if the attribute values are not continuous, e.g., cosine distance for text.

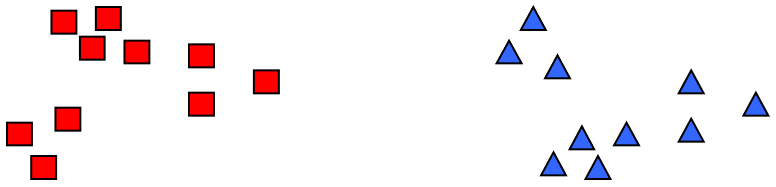
Notion of a Cluster can be Ambiguous



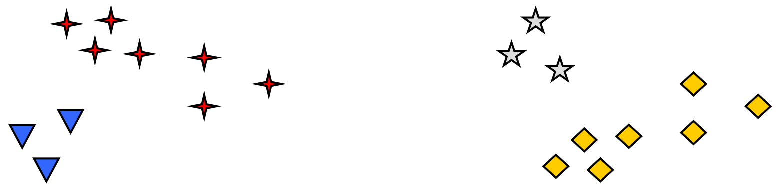
How many clusters?



Six Clusters



Two Clusters

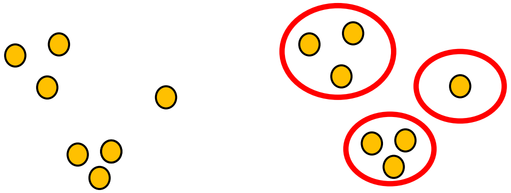


Four Clusters

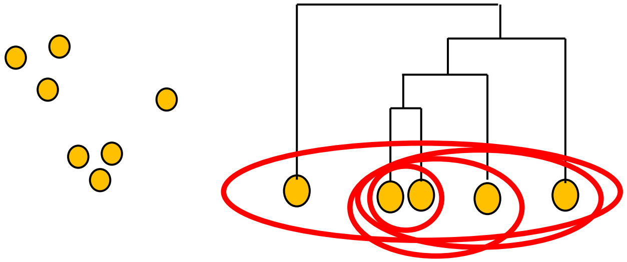
Types of clustering

Two main approaches: partitional and hierarchical.

- Partitional: the division of data points into non-overlapping subsets (clusters) such that each data point is in exactly one subset.



- Hierarchical: a set of nested clusters organized as a hierarchical tree.



k-Means clustering

k-Means Clustering

Partitional clustering approach

Each cluster is associated with a **centroid** (center point)

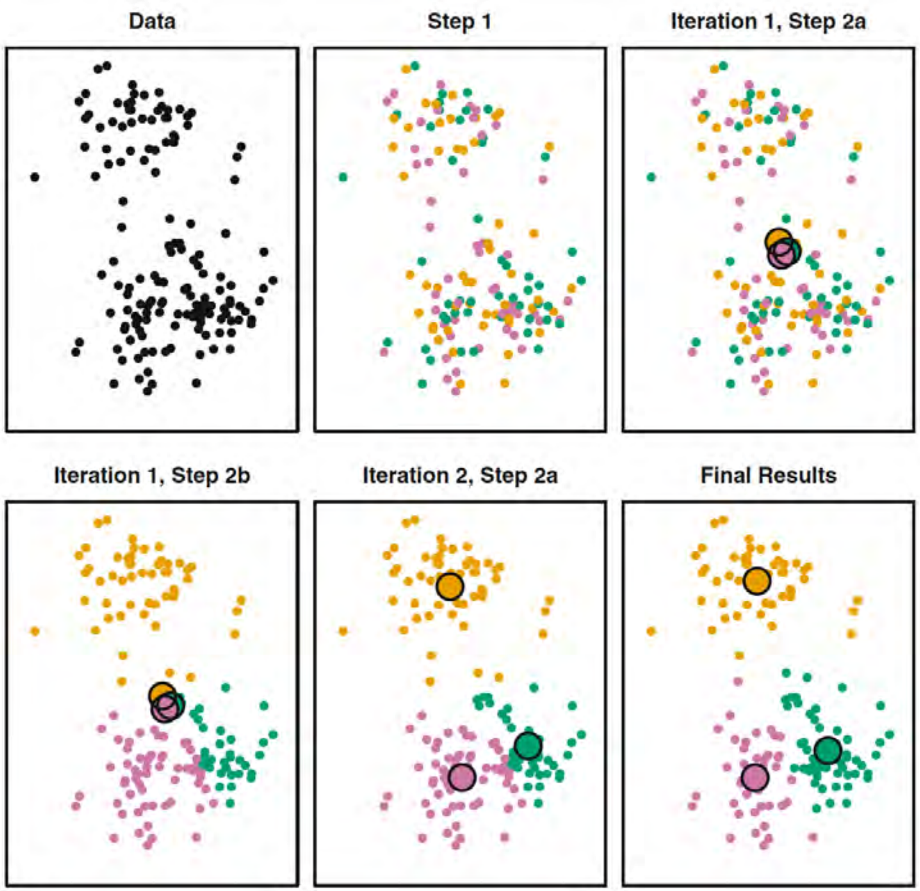
Each point is assigned to the cluster with the closest centroid

Number of clusters, *k*, must be specified

The basic algorithm is very simple:

1. Select *k* points (at random) as the initial centroids
2. **Repeat**
    3. Form *k* clusters by assigning all points to the closest centroid
    4. Re-compute the centroid of each cluster
5. **Until** the centroids don’t change

k-Means demonstration



In Step 1 each point is randomly assigned to a cluster. Top right: in Step 2(a), the cluster centroids are computed. In Step 2(b), each point is assigned to the nearest centroid. Bottom center: Step 2(a) is once again performed, leading to new cluster centroids. Bottom right: the results obtained after ten iterations.

James et al., An Introduction to Statistical Learning

Finding the centroids

How do we decide which is the **closest centroid**?

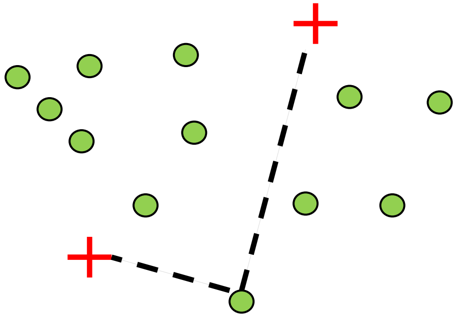
We need to find the ‘distance’ between each point and all the centroids

What does ‘distance’ mean?

There are many ways of defining ‘distance’. We need to use a distance metric.

Data points ●

Centroids +



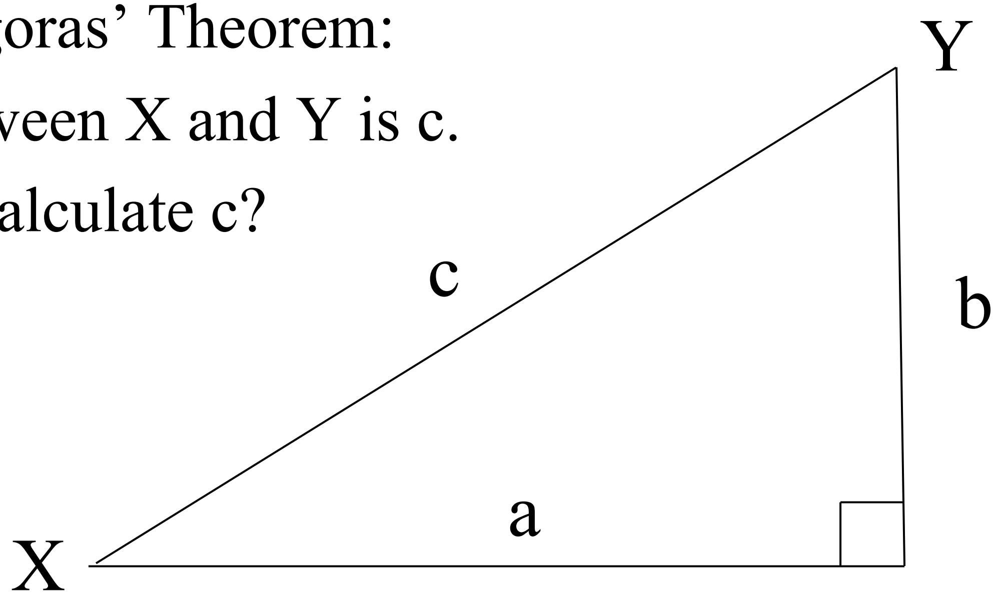
Euclidean distance

Euclidean distance is the shortest distance between two points.

Using Pythagoras' Theorem:

Distance between X and Y is c.

How do we calculate c?



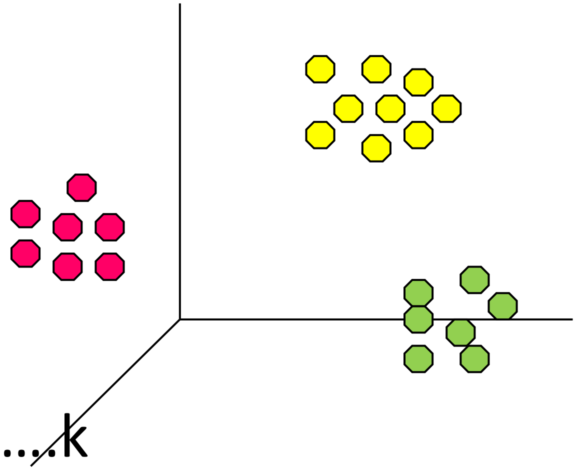
$$c^2 = a^2 + b^2, therefore\ \ c = \sqrt{(a^2 + b^2)}$$

This model can be applied to multiple dimensions!

What k-Means is aiming to do

The objective of the k-Means algorithm is to minimise the total squared distance of each point to its centroid:

$\sum_{i=1}^{k}\sum_{j=1}^{n} d\left(c_i, x_{i,j}\right)^2$ where:



- k is the number of clusters
- c_i is the centroid of each cluster for i=1,….k
- n_i is the number in cluster i
- $x_{i,j}$is the jth point of cluster i
- $d(c_i$,$x_{i,j})$ is the distance between c_i and $x_{i,j}$.

Evaluating k-Means Clusters

Most common measure: Sum of Squared Error (SSE)

- For each point, the error is the distance to the nearest cluster.
- To get SSE, we square these errors and sum them.
- x_i is a data point in cluster C_i and c_i is the centroid of cluster C_i .
- From previous slide: SSE = $\sum_{i=1}^{k} \sum_{j=1}^{n} d\left(c_i, x_{i,j}\right)^2$
- Given two sets of clusters, we can choose the one with the smallest error.
- Note: the easiest way to reduce SSE is to increase *k*, the number of clusters.
- We'll look at some alternative squared measures using R in the following examples.

Normalising attributes

It is a good idea to normalise the data before clustering, otherwise large valued attributes will exert greater influence on the clustering.

This is achieved by rescaling each attribute to fit within the same range (for example, between 0 and 1). To normalize attribute A:

MaxA and *MinA* are the maximum and minimum of A. Then, the normalized values of *A* are: $x_{new} = \frac{x - MinA}{MaxA - MinA}$

R software has a function (scale) which performs a similar – but not identical function.

Pre-processing and post-processing

Pre-processing

- Normalise the data
- Eliminate outliers

not so important

Post-processing

- Eliminate small clusters that may represent outliers
- Split 'loose' clusters, i.e., clusters with relatively high SSE.
- Merge clusters that are 'close' and that have relatively low SSE.

k-Means clustering in R

The k-Means function is built into the Stats package, which is loaded by default.

Using the iris data:

```
>   set.seed(9999) # makes “random” method repeatable
>   data("iris")
```

k-Means clustering in R

Using sepals (Cols 1 & 2), create 3 clusters, taking the best out of 20 starting configurations.

```
>   ikfit = kmeans(iris[,1:2], 3, nstart = 20) # create ikfit object
>   ikfit
>   table(actual = iris$Species, fitted = ikfit$cluster)
```

```
              fitted
actual         1  2  3
  setosa       0 50  0
  versicolor  12  0 38
  virginica   35  0 15
```

k-Means clustering in R

Looking at the ikfit object:

```
> ikfit
  K-means clustering with 3 clusters of sizes 47,
  50, 53

  Cluster means:  Sepal.Length Sepal.Width
             1     6.812766    3.074468
             2     5.006000    3.428000
             3     5.773585    2.692453

  Clustering vector:  [1] 2 2 2 2 2 2 ...
```

k-Means clustering in R

Looking at the ikfit object:

```
...
Within cluster sum of squares by cluster:
[1] 12.6217 13.1290 11.3000
(between_SS / total_SS =  71.6 %)

Available components:
[1] "cluster"      "centers"      "totss"
[4] "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

k-Means clustering in R

Looking at the sums of squares calculations:

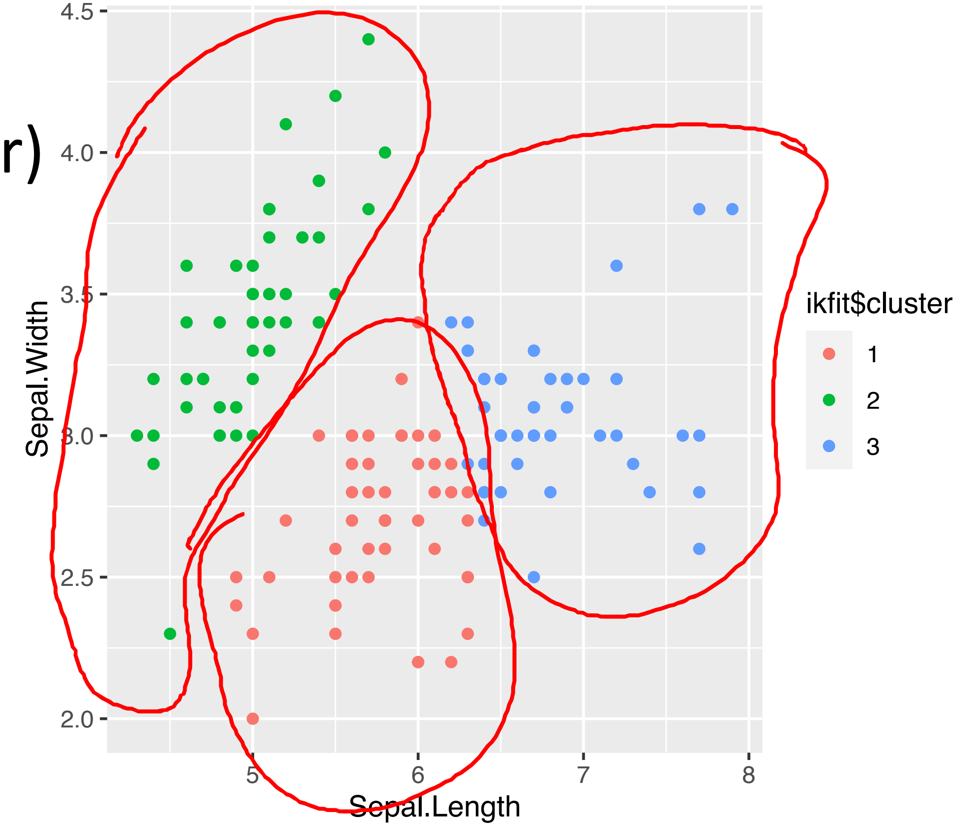
```
>   ikfit$totss   # Total SS from a single centroid (treats data as one cluster).
    [1] 130.4753
>   ikfit$withinss   # SS within each cluster.
    [1] 12.6217 13.1290 11.3000
>   ikfit$tot.withinss   # Total within clusters.  (Sum of Squared Error)
    [1] 37.0507
>   ikfit$betweenss   # Total sum of squares - total SS within clusters.
    [1] 93.42456
```

k-Means clustering in R

Plotting the clusters:

```
>   ikfit$cluster =
    as.factor(ikfit$cluster)

>   ggplot(iris,
    aes(Sepal.Length,
    Sepal.Width, color =
    ikfit$cluster)) +
    geom_point()
```



? kmeans *

- Description

```
Perform k-means clustering on a data matrix.
```

- Usage

```
kmeans(x, centers, iter.max = 10, nstart = 1,
algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
"MacQueen"), trace=FALSE)
```

`x`	`data`
`centers`	`number of clusters (k)`
`nstart`	`random starting positions to test`
`iter.max`	`maximum number of iterations`

```
...
```

k-Means clustering in R

Note that using both petals and sepals improves the accuracy of the clustering for these data:

```
>   ikfit = kmeans(iris[,1:4], 3, nstart = 20)
>   ttable(actual = iris$Species, fitted = ikfit$cluster)
                fitted
    actual        1  2  3
      setosa      0  0 50
      versicolor  2 48  0
      virginica  36 14  0
>   ikfit$tot.withinss
    [1] 78.85144
```

k-Means for classification…

From previous slide, using both petals and sepals for the clustering:

```
> ttable(actual = iris$Species, fitted = ikfit$cluster)
           fitted
actual        1  2  3
  setosa      0  0 50
  versicolor  2 48  0
  virginica  36 14  0
```

- If we classify Setosa = Group 3, Versicolor = Group 2 and Virginica = Group 1, this has an accuracy of:

> (50 + 48 + 36)/150 = 0.89.

k-Means clustering in R

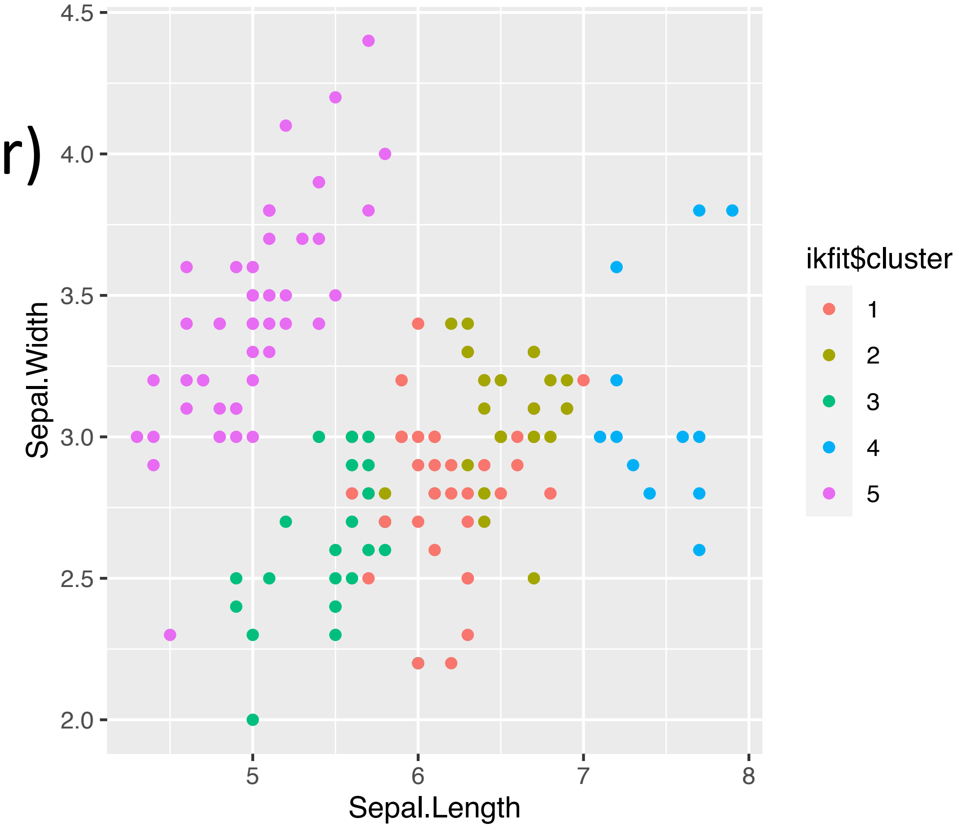
But the number of clusters is arbitrary, for example:

```
>   ikfit = kmeans(iris[,1:4], 5, nstart = 20)
>   ttable(actual = iris$Species, fitted = ikfit$cluster)
                 Fitted
actual          1  2  3  4  5
  setosa        0  0  0  0 50
  versicolor 26  0 24  0  0
  virginica  13 24  1 12  0
>   ikfit$tot.withinss
    [1] 46.44618    # compared to 78.85 for 3 clusters!
```

k-Means clustering in R

Plotting the clusters:

```
>   ikfit$cluster = as.factor(ikfit$cluster)
>   ggplot(iris, aes(Sepal.Length, Sepal.Width, color = ikfit$cluster)) + geom_point()
```



Countries data

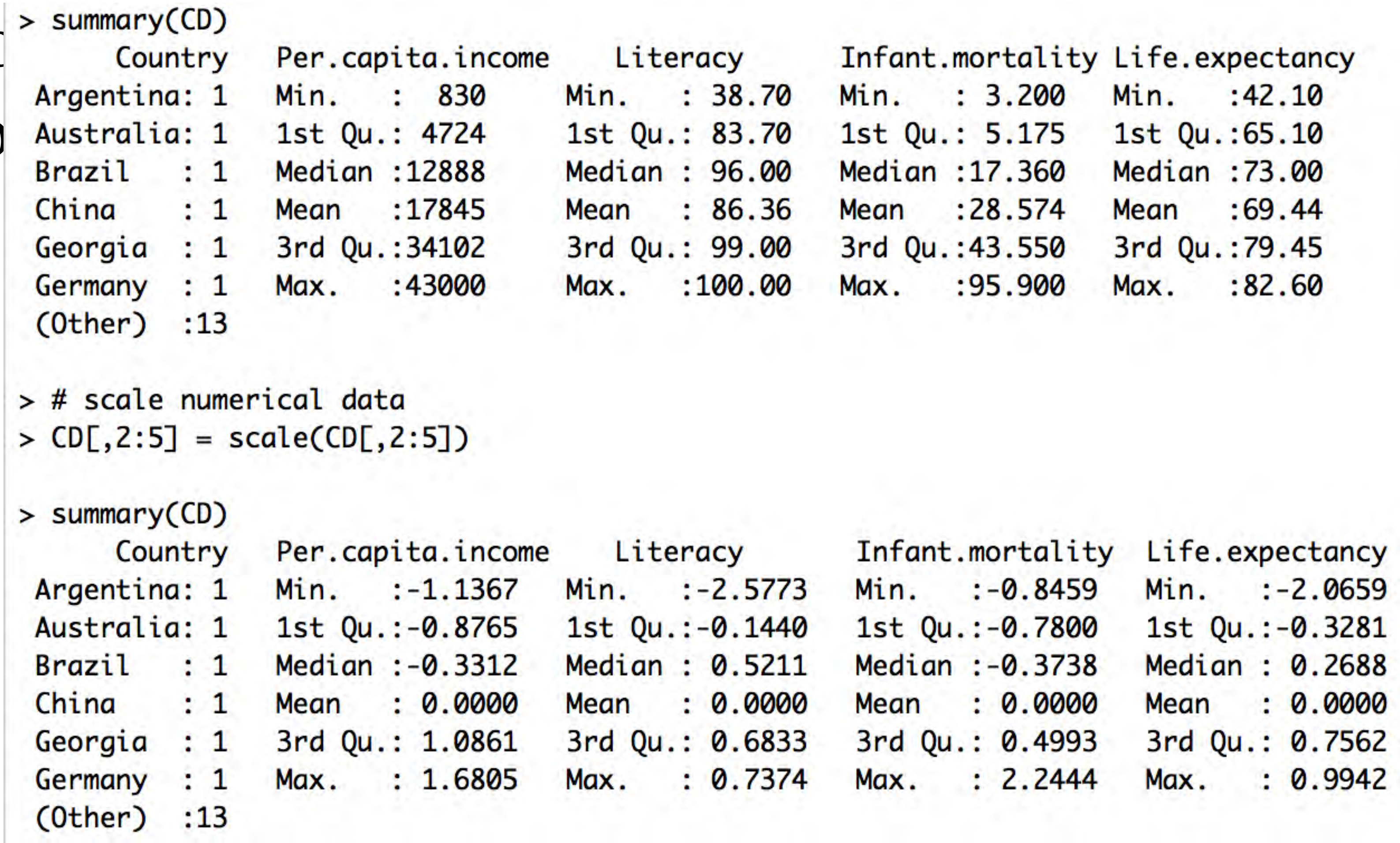
Sample socio-economic data for 19 countries.

Country	Per capita income	Literacy	Infant mortality	Life expectancy
Brazil	10326	90	23.6	75.4
Germany	39650	99	4.08	79.4
Mozambique	830	38.7	95.9	42.1
Australia	43000	99	4.57	81.2
China	5300	90.9	23	73
Argentina	13308	97.2	13.4	75.3
United Kingdom	34105	99	5.01	79.4
South Africa	10600	82.4	44.8	49.3
Zambia	1000	68	92.7	42.4
Namibia	5249	85	42.3	52.9
Georgia	4200	100	17.36	71
Pakistan	3320	49.9	67.5	65.5
India	2972	61	55	64.7
Turkey	12888	88.7	27.5	71.8
Sweden	34735	99	3.2	80.9
Lithuania	19730	99.6	8.5	73
Greece	36983	96	5.34	79.5
Italy	26760	98.5	5.94	80
Japan	34099	99	3.2	82.6

Countries data: scaling

Sc

co



```
> summary(CD)
      Country    Per.capita.income    Literacy        Infant.mortality Life.expectancy
 Argentina: 1    Min.    :  830     Min.    : 38.70   Min.    : 3.200   Min.    :42.10
 Australia: 1    1st Qu.: 4724      1st Qu.: 83.70   1st Qu.: 5.175   1st Qu.:65.10
 Brazil    : 1   Median :12888      Median : 96.00   Median :17.360   Median :73.00
 China     : 1   Mean   :17845      Mean   : 86.36   Mean   :28.574   Mean   :69.44
 Georgia   : 1   3rd Qu.:34102      3rd Qu.: 99.00   3rd Qu.:43.550   3rd Qu.:79.45
 Germany   : 1   Max.   :43000      Max.   :100.00   Max.   :95.900   Max.   :82.60
 (Other)   :13

> # scale numerical data
> CD[,2:5] = scale(CD[,2:5])

> summary(CD)
      Country    Per.capita.income    Literacy         Infant.mortality  Life.expectancy
 Argentina: 1    Min.    :-1.1367   Min.    :-2.5773   Min.    :-0.8459   Min.    :-2.0659
 Australia: 1    1st Qu.:-0.8765    1st Qu.:-0.1440   1st Qu.:-0.7800   1st Qu.:-0.3281
 Brazil    : 1   Median :-0.3312    Median : 0.5211   Median :-0.3738   Median : 0.2688
 China     : 1   Mean   : 0.0000    Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 Georgia   : 1   3rd Qu.: 1.0861    3rd Qu.: 0.6833   3rd Qu.: 0.4993   3rd Qu.: 0.7562
 Germany   : 1   Max.   : 1.6805    Max.   : 0.7374   Max.   : 2.2444   Max.   : 0.9942
 (Other)   :13
```

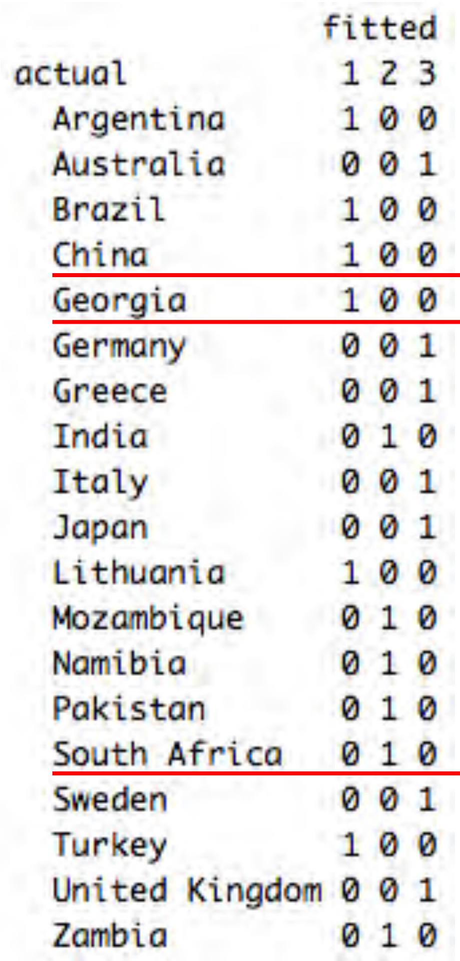
Countries data: k-Means

k-Means for the scaled data set

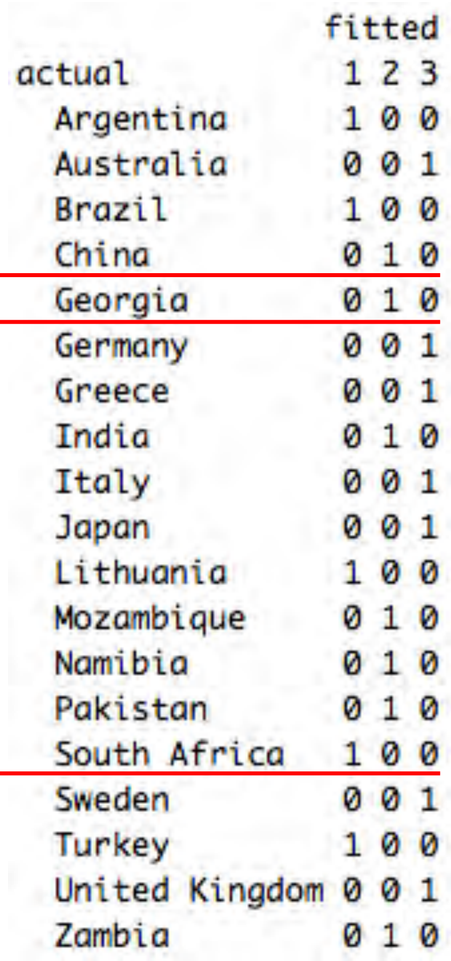
```
>   set.seed(9999)
>   CD <- read.csv("CountriesData.csv")
>   # scale numerical data
>   CD[,2:5] = scale(CD[,2:5])
>   CDkfit = kmeans(CD[,2:5], 3, nstart = 20)
>   CDkfit
>   table(actual = CD$Country, fitted = CDkfit$cluster)
```

Non–scaled *v* scaled clusters

Not-scaled



```
                 fitted
actual            1 2 3
  Argentina       1 0 0
  Australia       0 0 1
  Brazil          1 0 0
  China           1 0 0
  Georgia         1 0 0
  Germany         0 0 1
  Greece          0 0 1
  India           0 1 0
  Italy           0 0 1
  Japan           0 0 1
  Lithuania       1 0 0
  Mozambique      0 1 0
  Namibia         0 1 0
  Pakistan        0 1 0
  South Africa    0 1 0
  Sweden          0 0 1
  Turkey          1 0 0
  United Kingdom  0 0 1
  Zambia          0 1 0
```



```
                 fitted
actual            1 2 3
  Argentina       1 0 0
  Australia       0 0 1
  Brazil          1 0 0
  China           0 1 0
  Georgia         0 1 0
  Germany         0 0 1
  Greece          0 0 1
  India           0 1 0
  Italy           0 0 1
  Japan           0 0 1
  Lithuania       1 0 0
  Mozambique      0 1 0
  Namibia         0 1 0
  Pakistan        0 1 0
  South Africa    1 0 0
  Sweden          0 0 1
  Turkey          1 0 0
  United Kingdom  0 0 1
  Zambia          0 1 0
```

Scaled

Scaling changes the clusters of these countries.

k-Means: some considerations

The location of the initial centroids influences final clusters, some ways to address this:

- Multiple runs (which k-Means does), or
- Select more than *k* initial centroids and then select among these initial centroids.

k-Means: some considerations

How do we decide which k to use?

- Trial and error
- There is no single best way of doing this. One reference with some good approaches is https://uc-r.github.io/kmeans_clustering.
- The following shows one method, the average silhouette, adapted from Giordani et al., An Introduction to Clustering with R.

K-Means: Silhouette

- The average silhouette calculates how well each data point sits within its cluster. It is a proxy measure for the quality of the clustering. For each point, i ,

$$s_i = \frac{b_i \; - a_i}{max(b_i, a_i)}, i = 1, 2, 3, \ldots$$

- where a_i is the average distance between that point and all other points in the same cluster, and
- b_i is smallest average distance to any cluster it does not belong to. Ideally a_i is small and b_i is large.
- The average s can then be evaluated across all i. at different values of k.

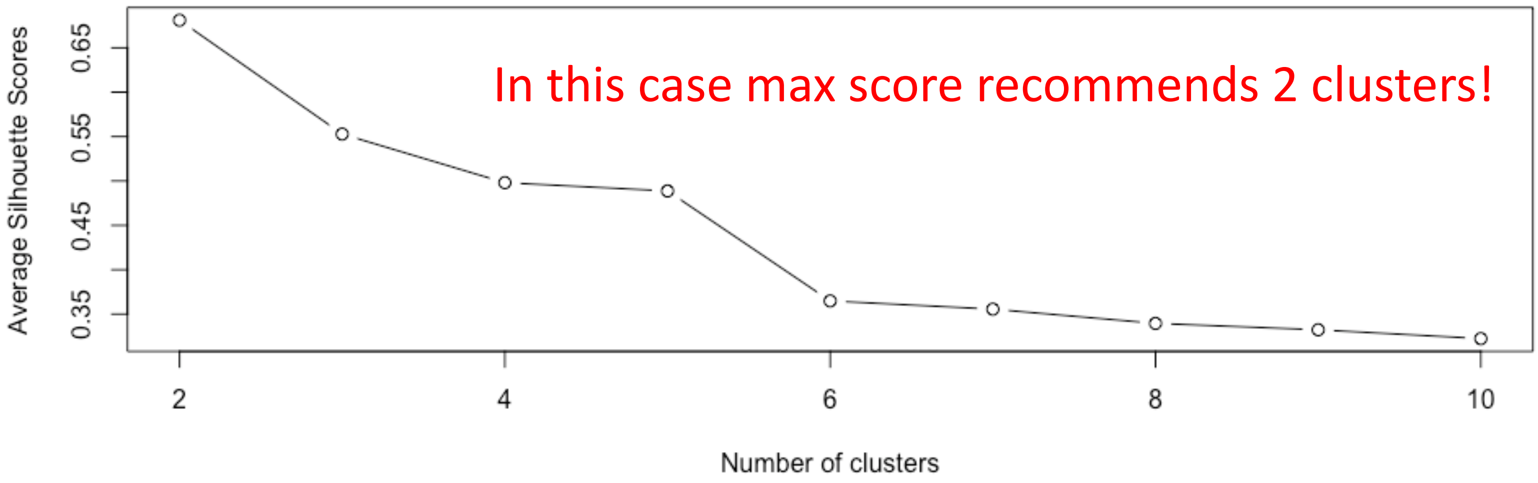
K-Means: Silhouette in R

For the iris data, using sepals and petals:

```
>   library(cluster)
>   #make function to get average silhouette score
>   i_silhouette_score <- function(k){
>     km <- kmeans(iris[,1:4], centers = k, nstart=25)
>     ss <- silhouette(km$cluster, dist(iris[,1:4]))
>     mean(ss[, 3])
>   }
```

K-Means: Silhouette in R

```
> #calc and plot average silhouette for 2-10 clusters
> k <- 2:10
> avg_sil <- sapply(k, i_silhouette_score)
> plot(k, type='b', avg_sil, xlab='Number of clusters',
  ylab='Average Silhouette Scores')
```



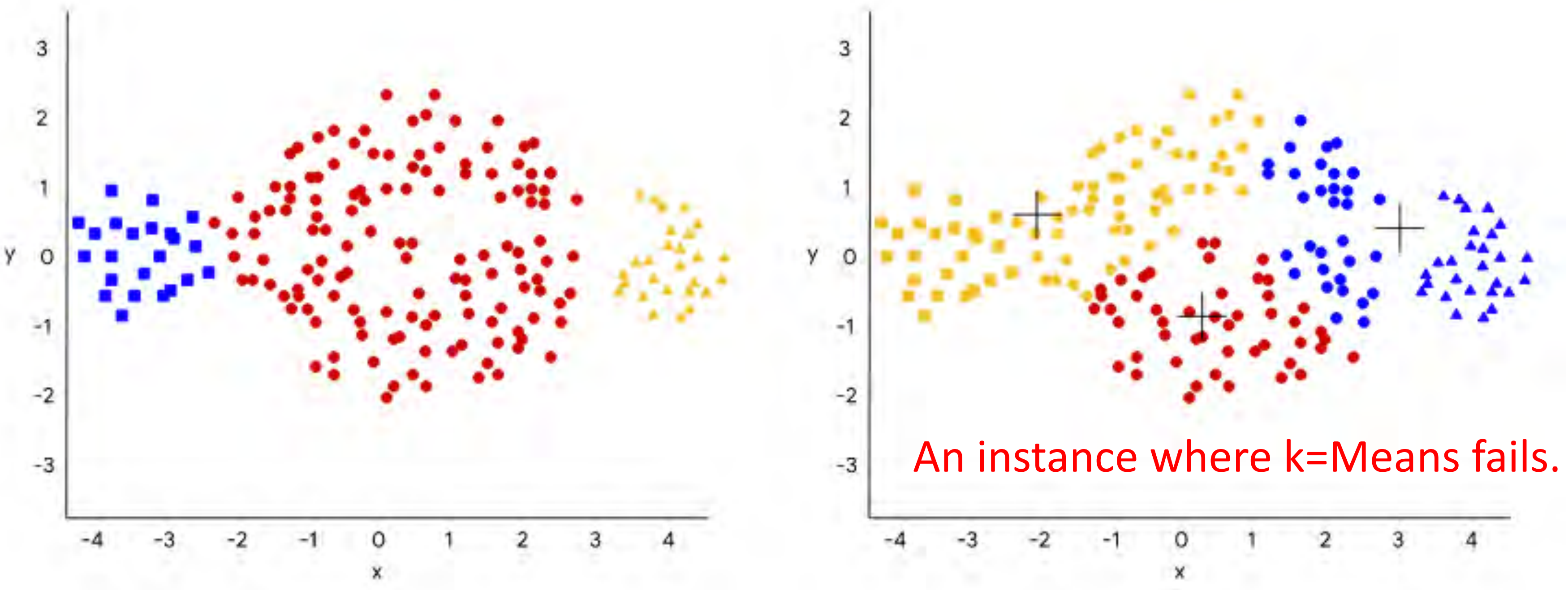
In this case max score recommends 2 clusters!

K-Means: further thoughts…

- Advantages:
  - > Relatively simple to implement. Scales to large data sets. Guarantees convergence. Can warm-start the positions of centroids. Easily adapts to new examples. Generalizes to clusters of different shapes and sizes.
- Disadvantages:
  - > Dependent on initial values. Clusters of varying sizes and density. Centroids can be dragged by outliers. Outliers might become their own cluster. Non-globular clusters hard to identify.

https://developers.google.com/

K-Means: further thoughts…



An instance where k=Means fails.

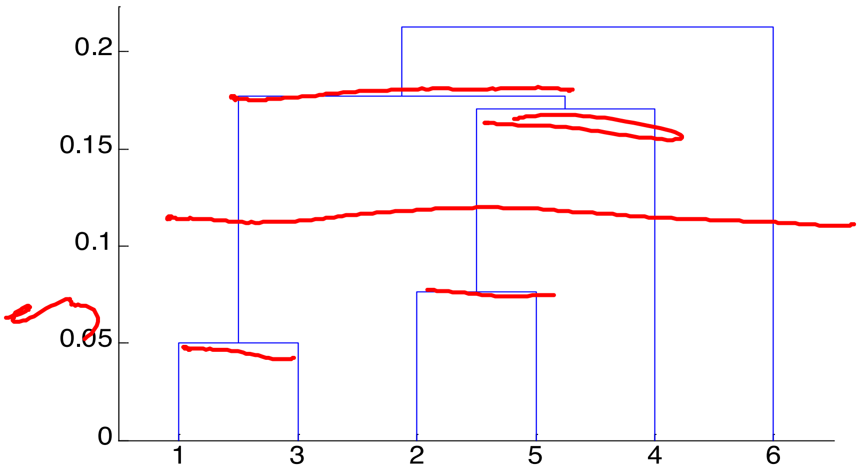
https://developers.google.com/

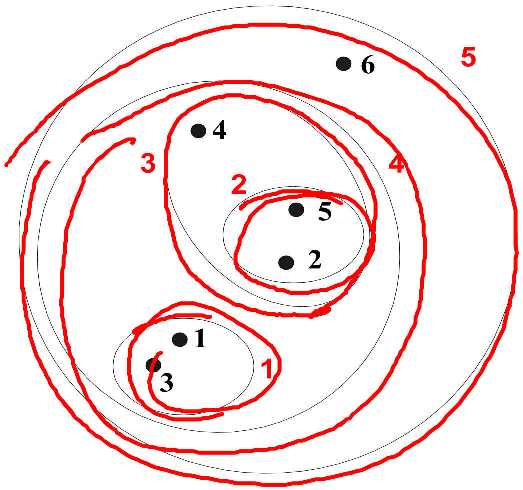
Hierarchical clustering

Hierarchical clustering

- Creates a set of nested clusters organized as a hierarchical tree that:
  - > Records the sequences of merges or splits
  - > Can be visualized as a dendrogram

4 clusters





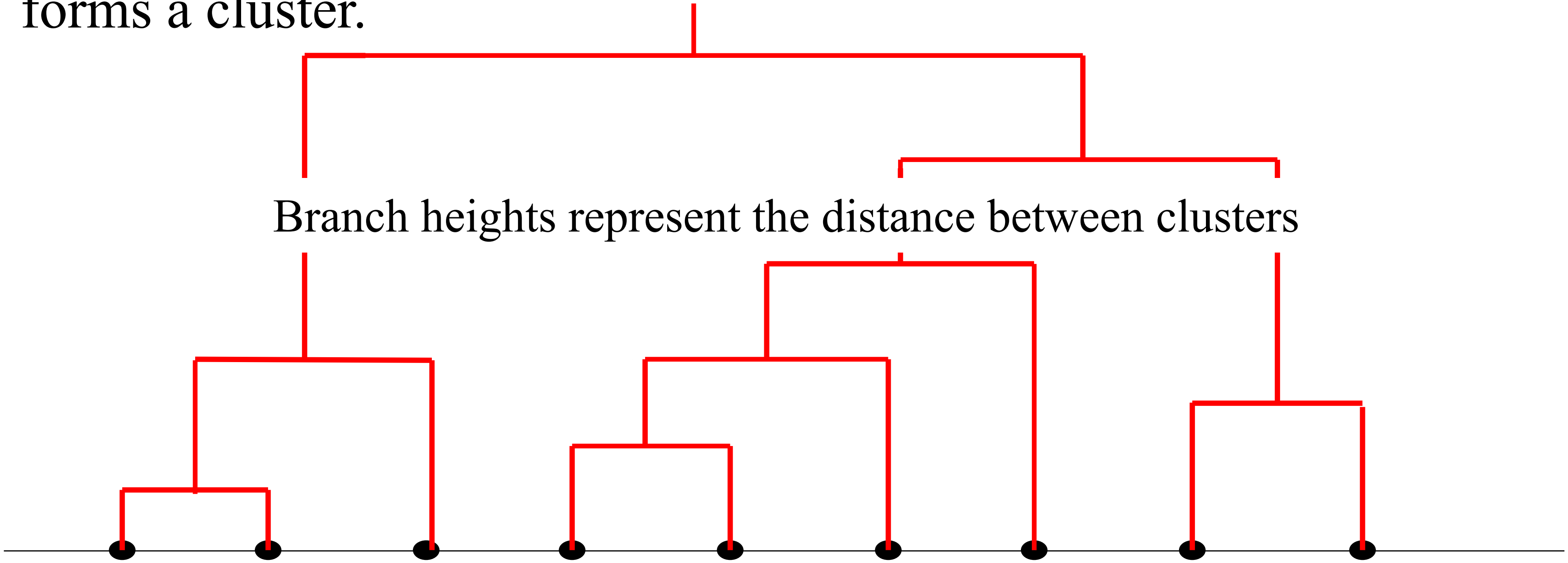
Advantages of hierarchical clustering

- Do not have to assume any particular number of clusters:
  - > Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the appropriate level.
- They may correspond to meaningful taxonomies:
  - > For example, in biological sciences (e.g., plant and animal kingdoms).

Dendrogram and hierarchies

Decompose data points into a several levels of nested partitioning (**tree** of clusters), called a **dendrogram**.

A **clustering** of the data points is obtained by **cutting** the dendrogram at the desired level, then each **connected component** forms a cluster.

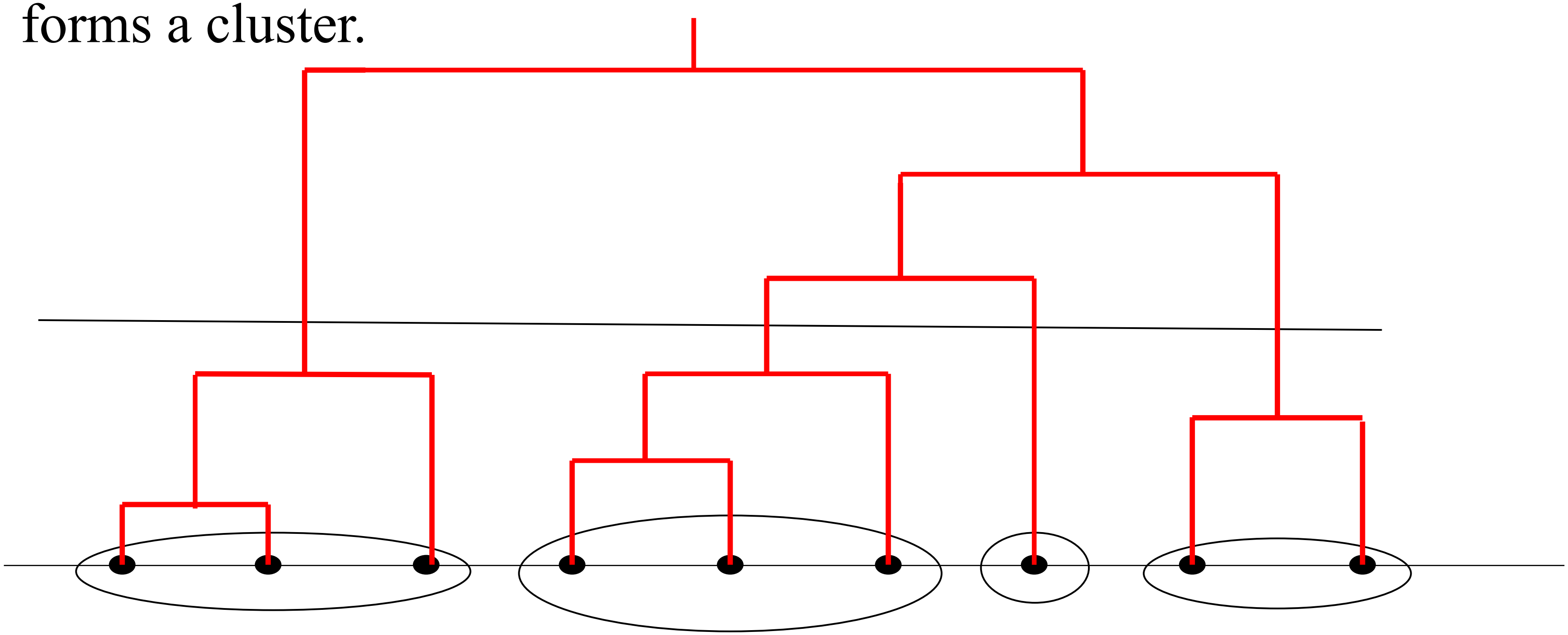


Branch heights represent the distance between clusters

Dendrogram and hierarchies

Decompose data points into a several levels of nested partitioning (**tree** of clusters), called a **dendrogram**.

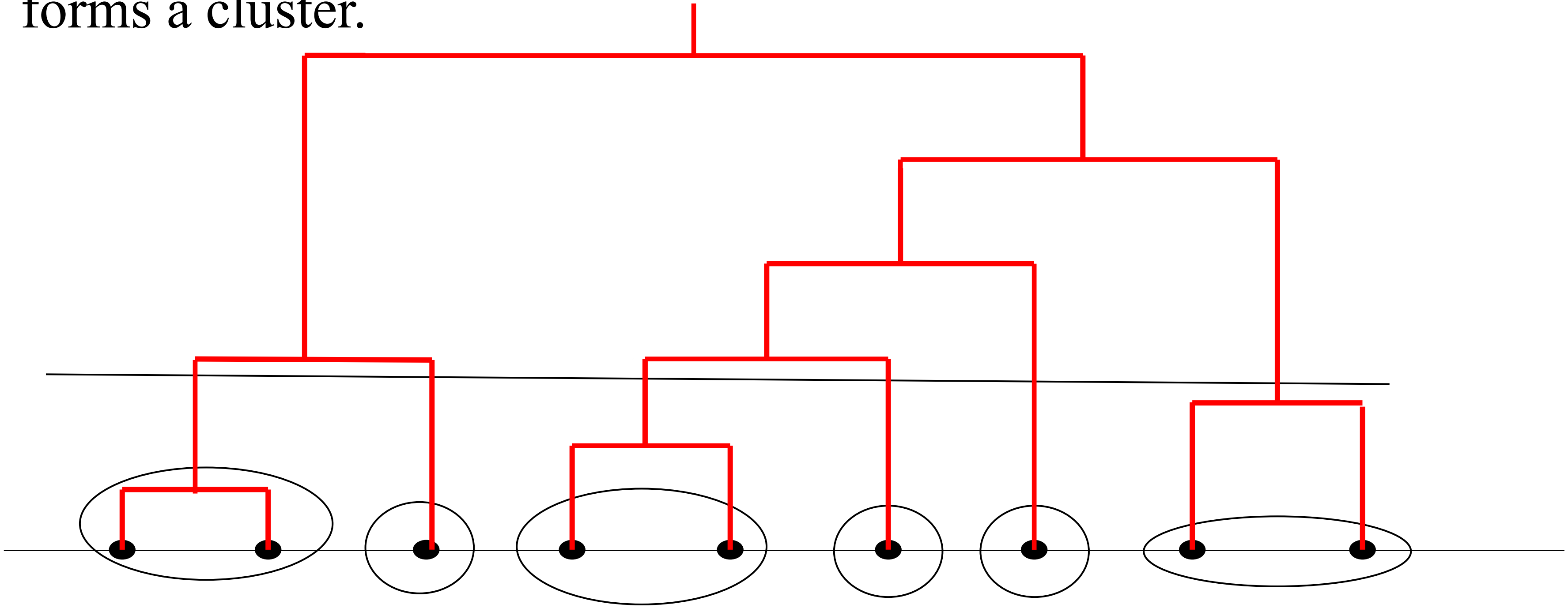
A **clustering** of the data points is obtained by **cutting** the dendrogram at the desired level, then each **connected component** forms a cluster.



Dendrogram and hierarchies

Decompose data points into a several levels of nested partitioning (**tree** of clusters), called a **dendrogram**.

A **clustering** of the data points is obtained by **cutting** the dendrogram at the desired level, then each **connected component** forms a cluster.



Hierarchical Clustering

Two main types of hierarchical clustering:

- Agglomerative (the more usual method):
  - > Start with the points as individual clusters.
  - > At each step, merge the closest pair of clusters until only one cluster (or k clusters) left.
- Divisive:
  - > Start with one, all-inclusive cluster.
  - > At each step, split a cluster until each cluster contains a point (or there are k clusters).
- Traditional hierarchical algorithms use a similarity or distance matrix and merge or split one cluster at a time

Agglomerative Clustering Algorithm

More popular hierarchical clustering technique

Distance matrix stores the distances between each cluster

Basic algorithm is straightforward

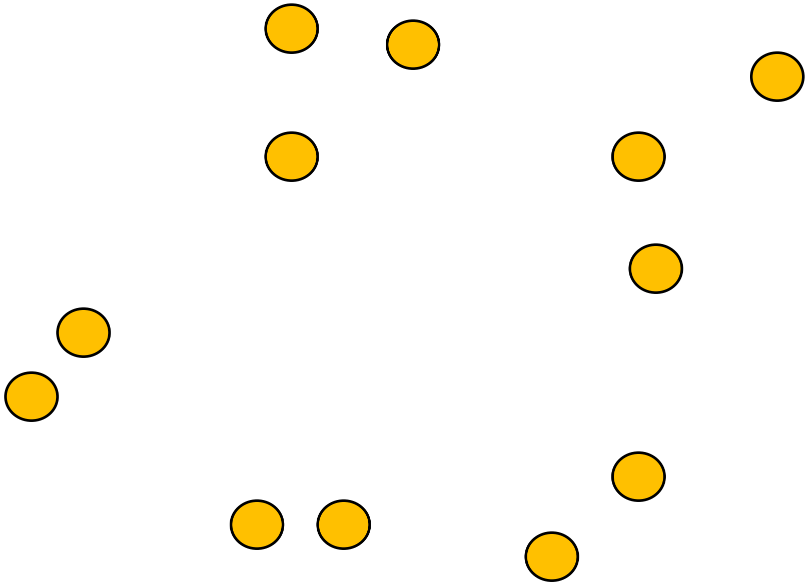
```
1. Compute the distance matrix
2. Let each data point be a cluster
3. Repeat
4.     Merge the two closest clusters
5.     Update the distance matrix
6. Until only a single cluster remains
```

Key operation is the computation of distance between two clusters

Different approaches to defining the distance distinguish the different algorithms

Starting Situation

Start with clusters of individual points and a distance/proximity matrix

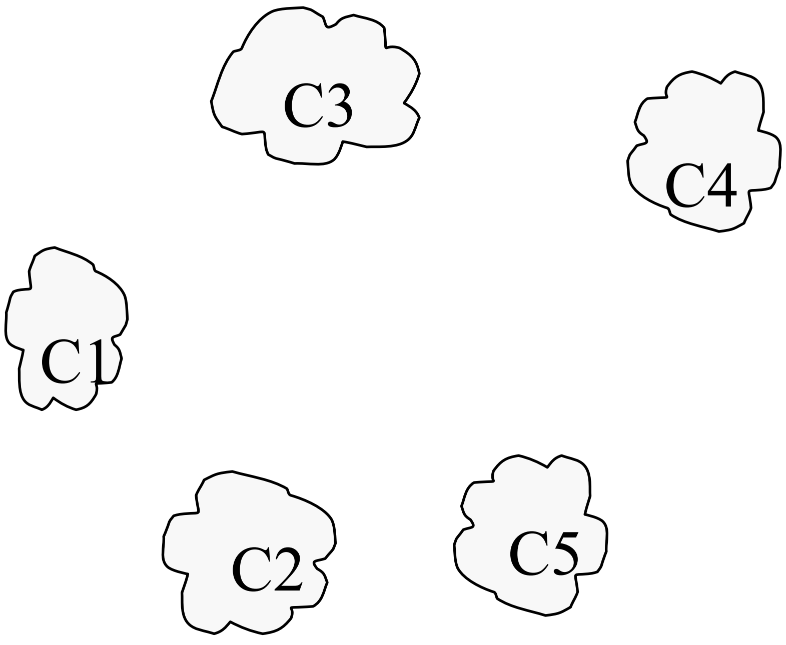


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Distance/ Proximity Matrix

Intermediate Situation

After some merging steps,
we have some clusters

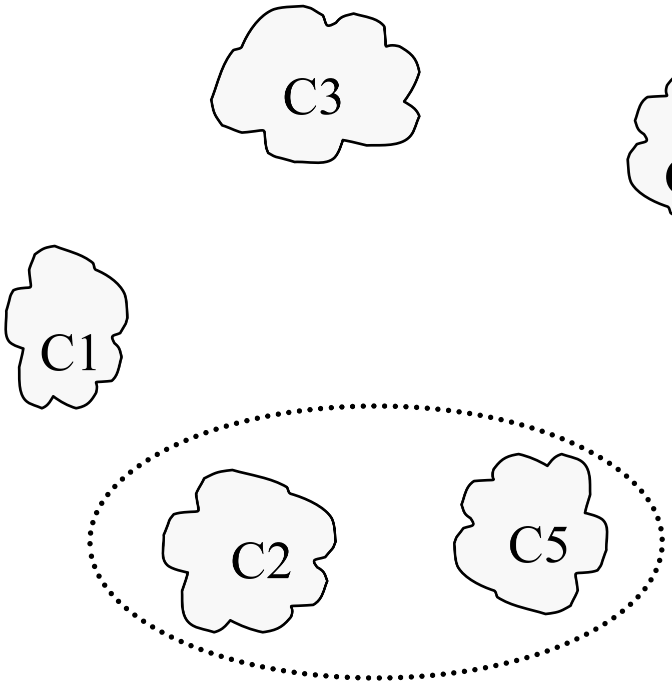


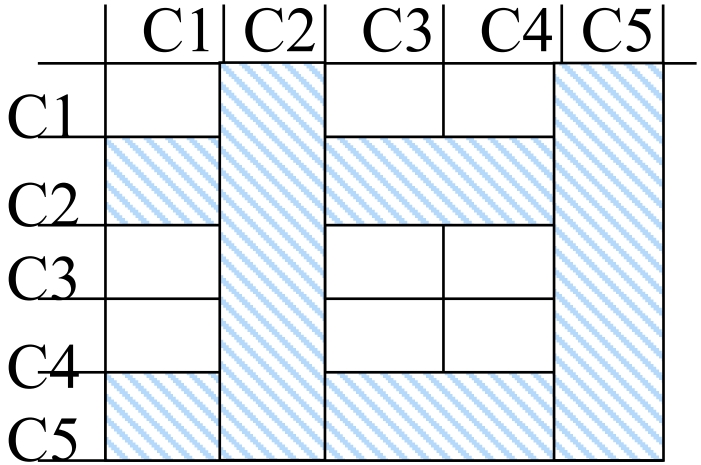
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance/ Proximity Matrix

Intermediate Situation

We want to merge the two closest clusters (C2 and C5) and update the distance matrix.





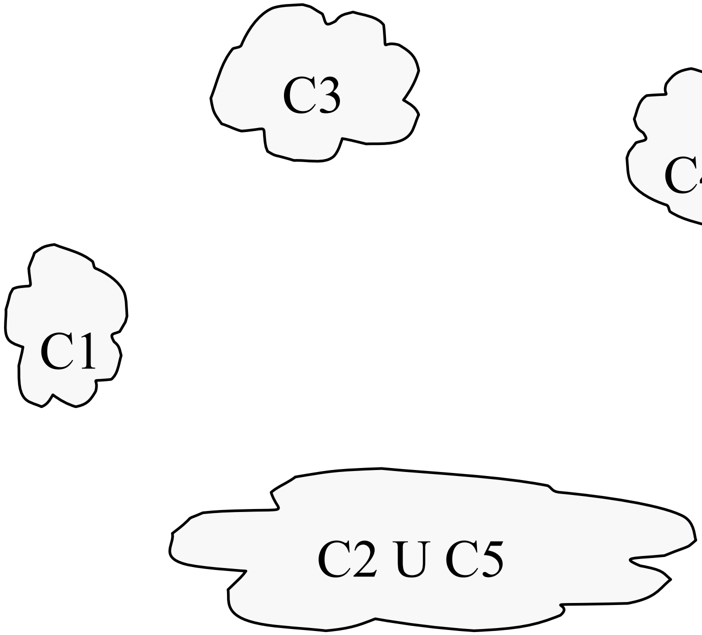
	C1	C2	C3	C4	C5
C1		▨			▨
C2	▨	▨	▨	▨	▨
C3		▨			▨
C4		▨			▨
C5	▨	▨	▨	▨	▨

Distance/ Proximity Matrix

After Merging

The question is

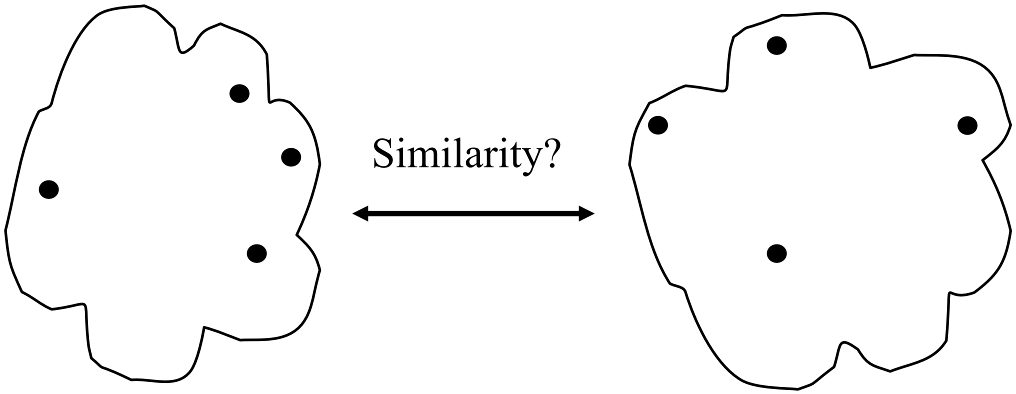
"How do we update the distance/ proximity matrix?"



	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Distance/ Proximity Matrix

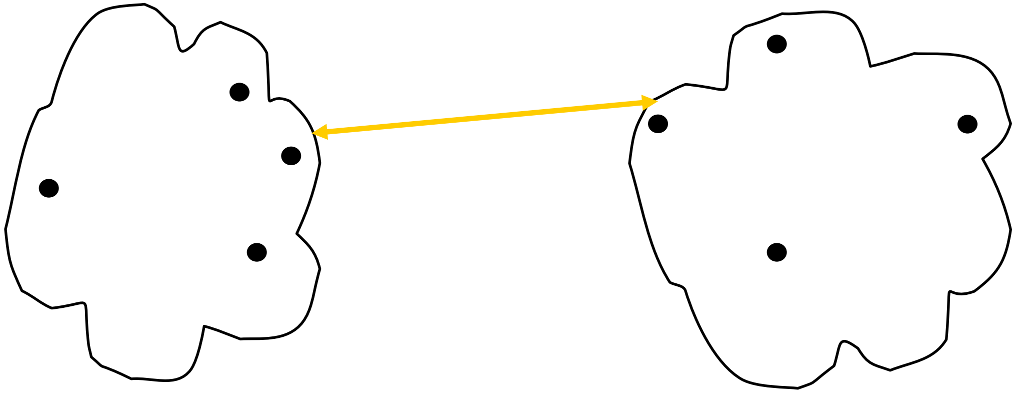
How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

- MIN
- MAX
- Group Average
- Distance Between Centroids

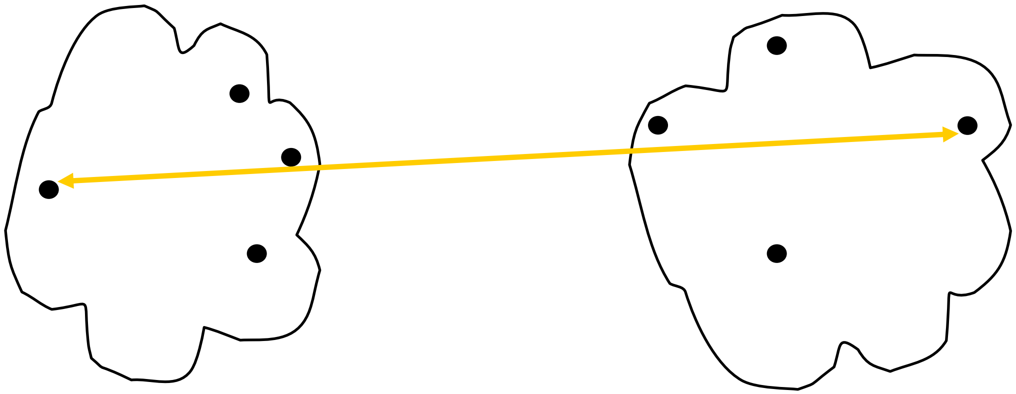
How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

- MIN
- MAX
- Group Average
- Distance Between Centroids

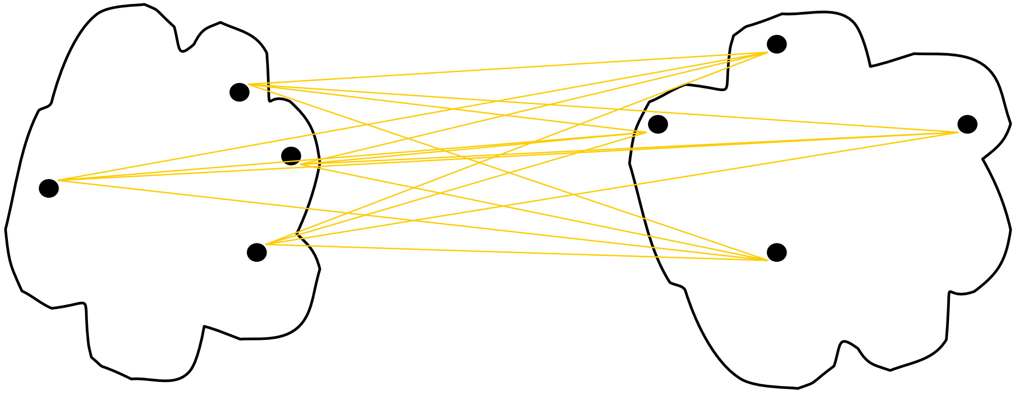
How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

- MIN
- MAX
- Group Average
- Distance Between Centroids

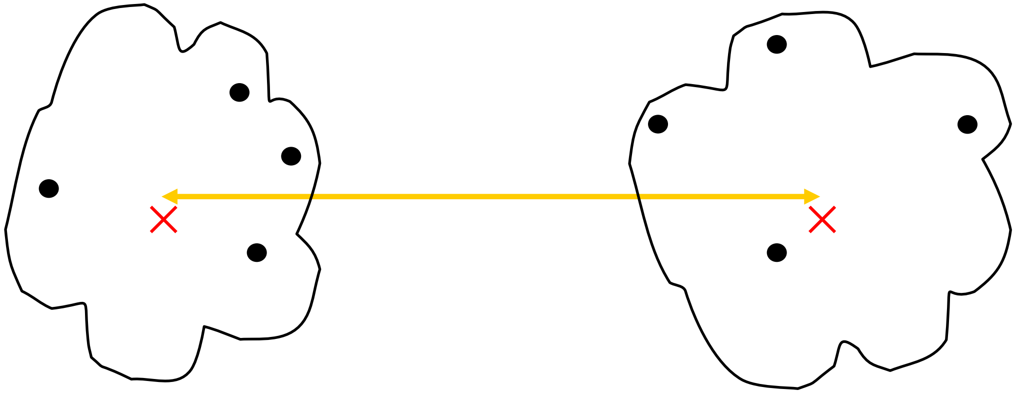
How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						

.

.

.

- MIN
- MAX
- Group Average
- Distance Between Centroids

Class Activity

Merging with MIN, let’s try first merge for a hypothetical data set...

	P1	P2	P3	P4	P5	P6
P1	0	0.24	0.22	0.37	0.34	0.23
P2	0.24	0	0.15	0.2	0.14	0.25
P3	0.22	0.15	0	0.15	0.28	0.11
P4	0.37	0.14	0.15	0	0.29	0.22
P5	0.23	0.25	0.28	0.29	0	0.39
P6	0.23	0.25	0.11	0.22	0.39	0

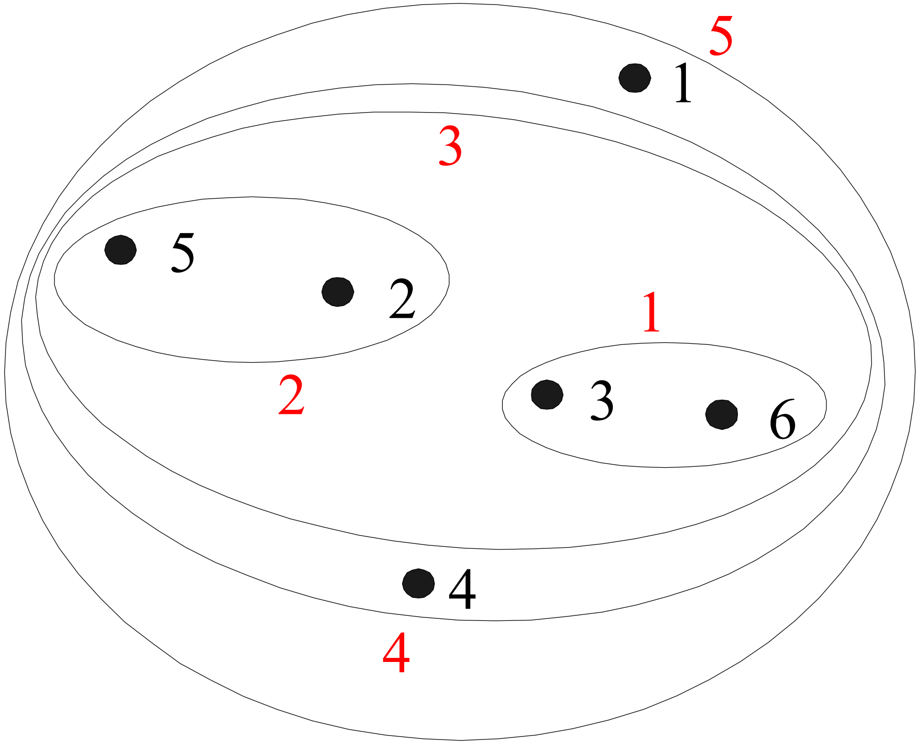
error

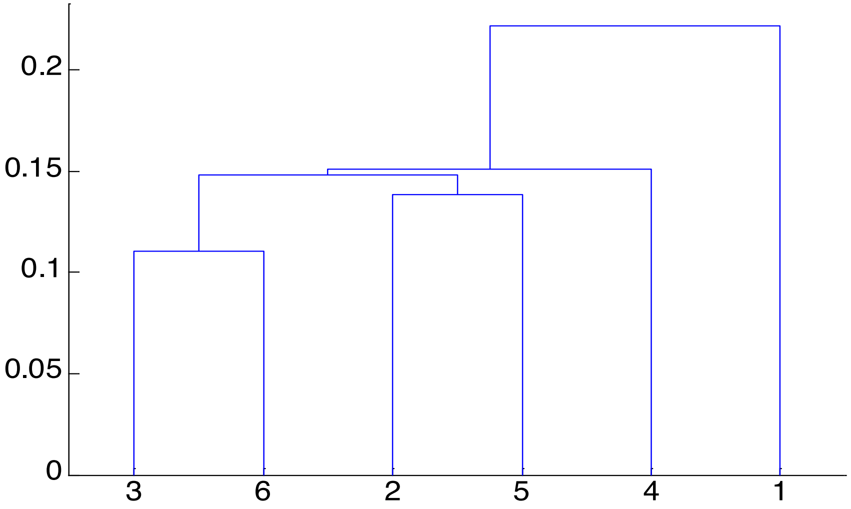
First Join					
	P1	P2	P36	P4	P5
P1		0.24	0.22	0.37	0.34
P2			0.15	…	
P36					
P4					
P5					

Effect of clustering method

The following slides show the slightly different clustering obtained by MIN, MAX and Group Average distance measures…

Hierarchical Clustering: MIN *

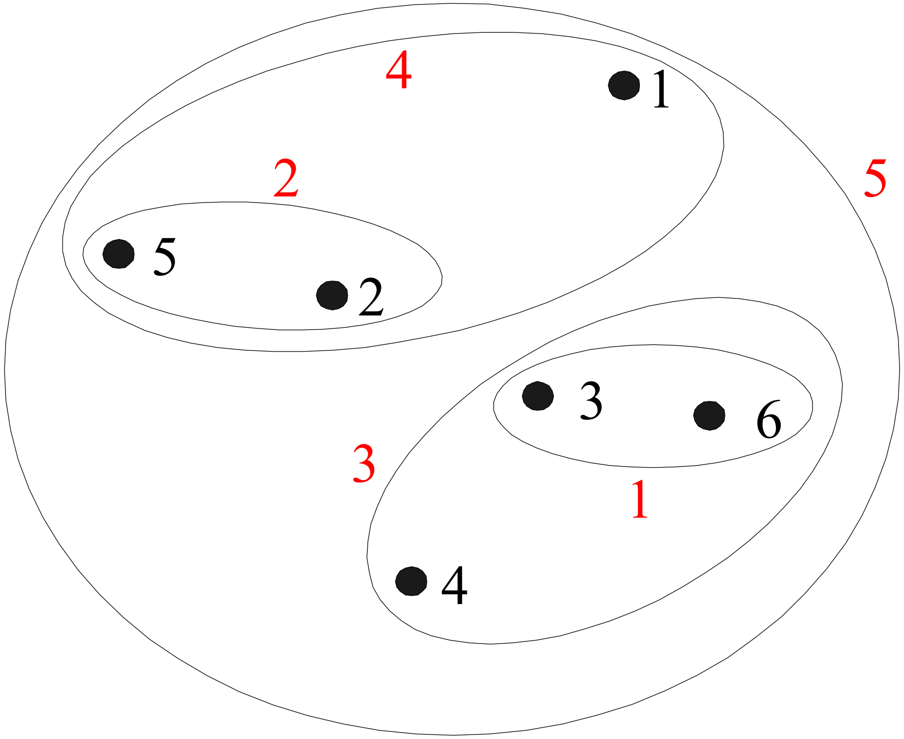


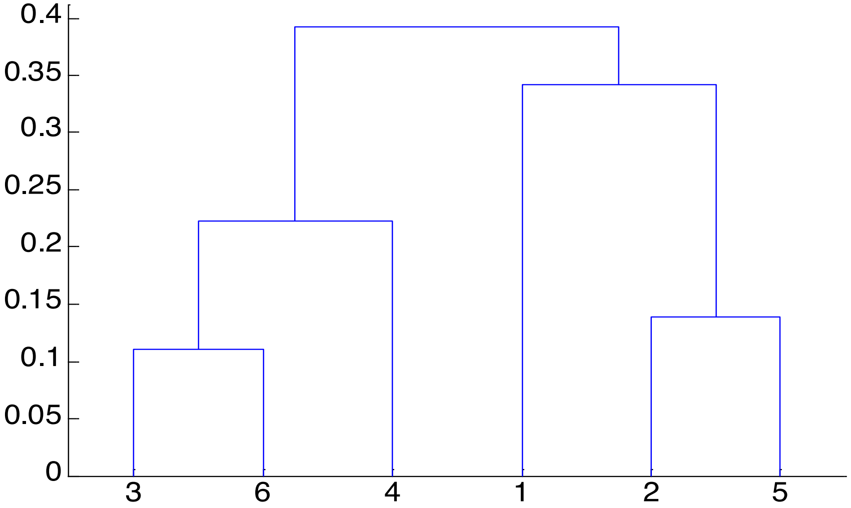


Nested Clusters

Dendrogram

Hierarchical Clustering: MAX *

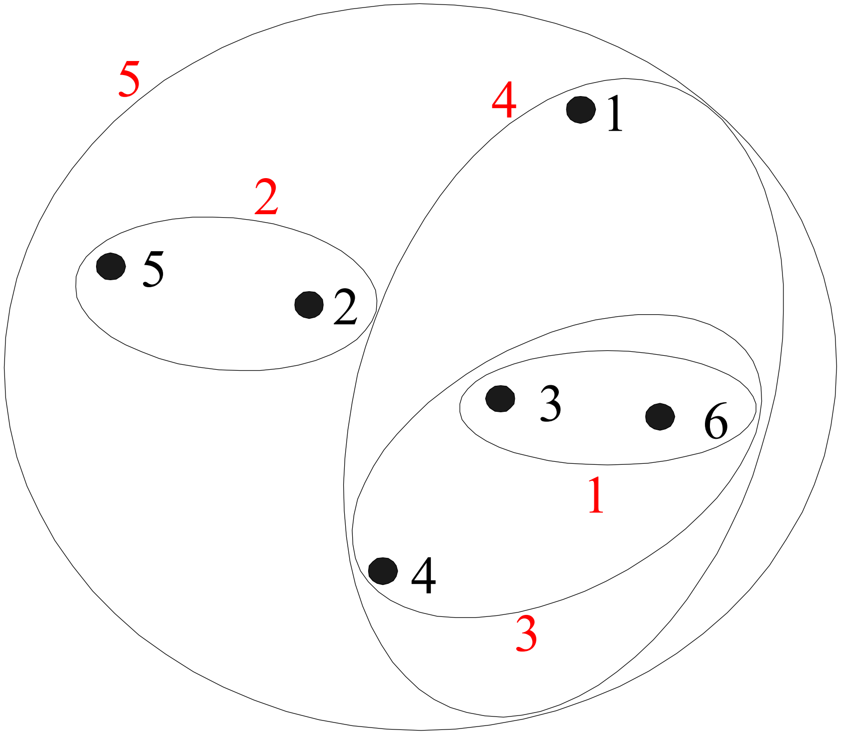


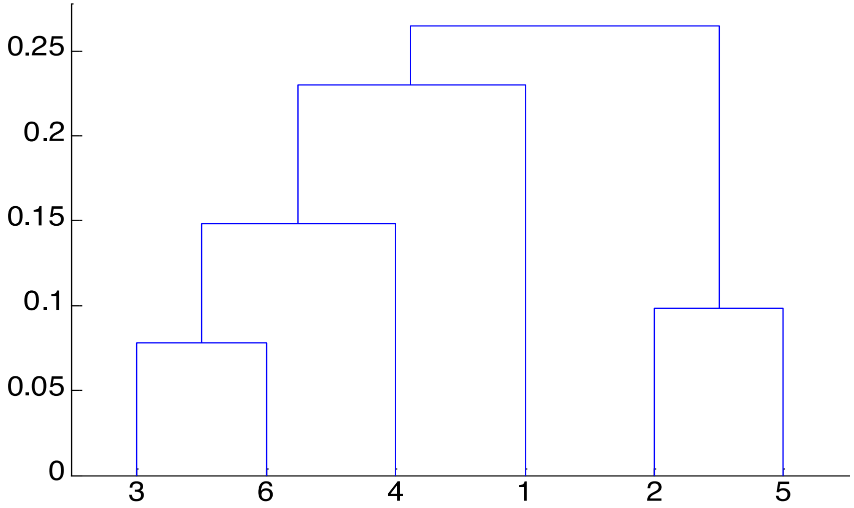


Nested Clusters

Dendrogram

Hierarchical Clustering: Group Average *

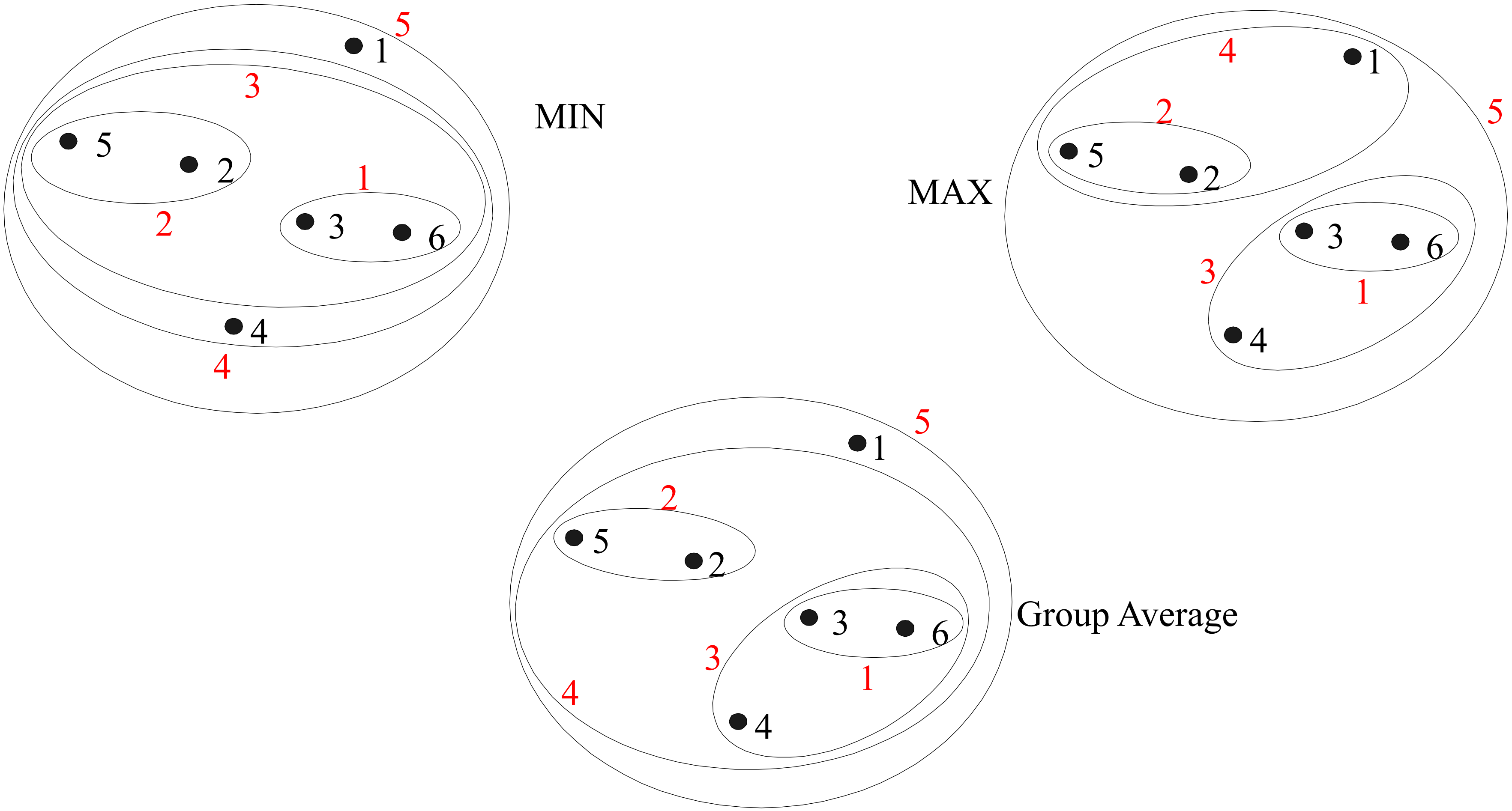




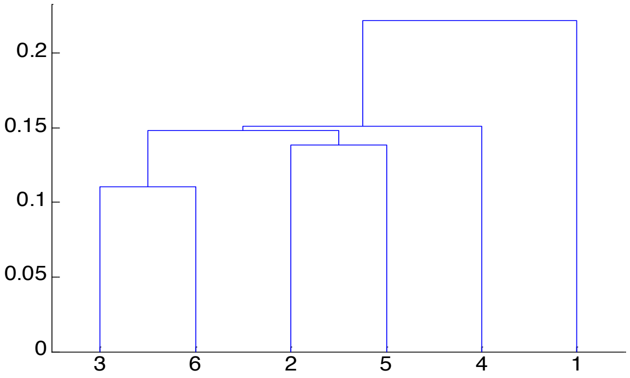
Nested Clusters

Dendrogram

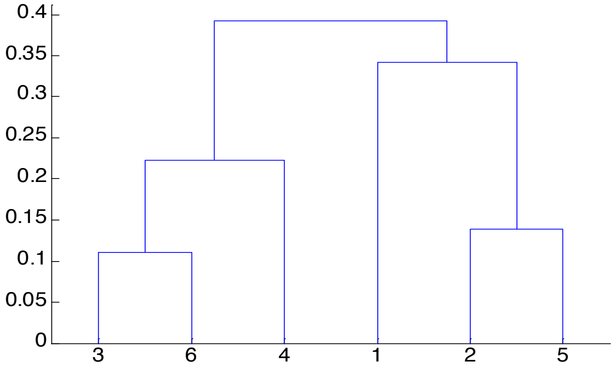
Hierarchical Clustering: Comparison



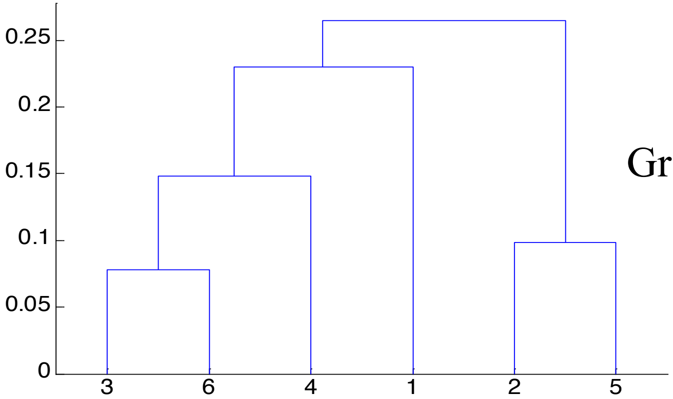
Hierarchical Clustering: Comparison



MIN



MAX



Group Average

Similarity measures: pros and cons

- MIN
  - > Can handle non-elliptical shapes
  - > Sensitive to noise and outliers
- MAX
  - > Less susceptible to noise and outliers
  - > Tends to break large clusters, biased towards elliptical shapes
- Group Average
  - > Compromise between Single and Complete Link
  - > Less susceptible to noise and outliers
  - > Biased towards globular clusters

Hierarchical clustering: considerations

- Once a decision is made to combine two clusters, it cannot be undone.
- No objective function is directly minimized, unlike k-Means.
- Different schemes have problems with one or more of the following:
  - > Sensitivity to noise and outliers.
  - > Difficulty handling different sized clusters and convex shapes.
  - > Breaking large clusters.

Hierarchical clustering in R

Hierarchical clustering of the Iris data using the function hclust (also part of the Stats package):

```
>   set.seed(9999)
>   data("iris")
>   niris = iris
>   #scale numerical data this gives poorer result for iris
>   #niris[,1:4] = scale(niris[,1:4])
>   ihfit = hclust(dist(niris[,1:4]), "ave")
>   plot(ihfit, hang = -1)
```

Hierarchical clustering in R

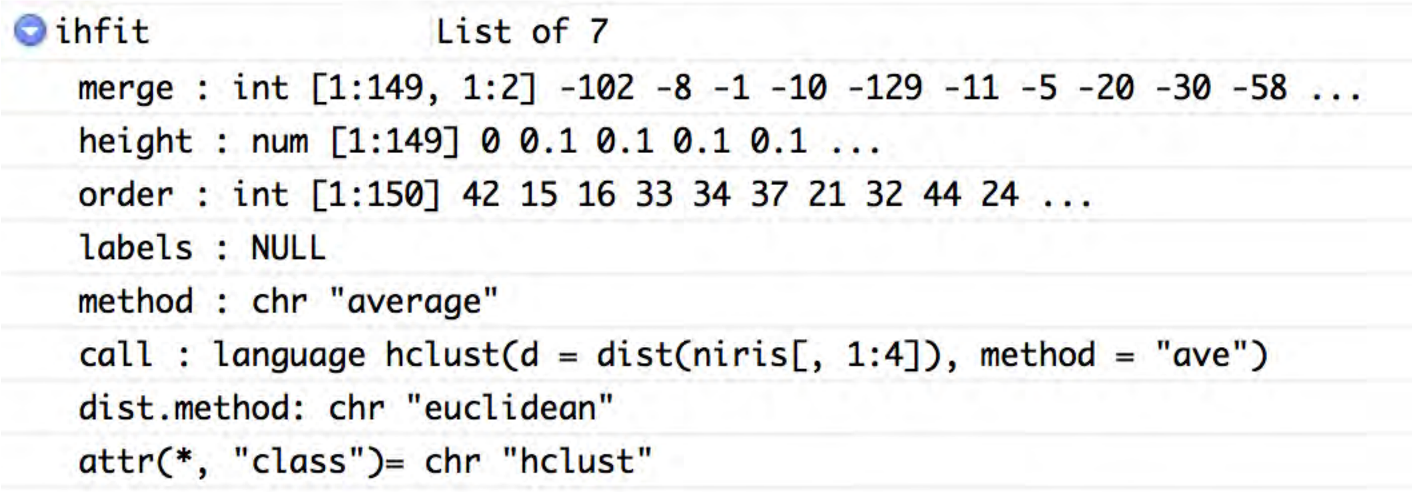
The fitted object:

```
> ihfit
Call:
hclust(d = dist(niris[, 1:4]), method = "ave")

Cluster method   : average Distance         :
euclidean
Number of objects: 150
```

Hierarchical clustering in R

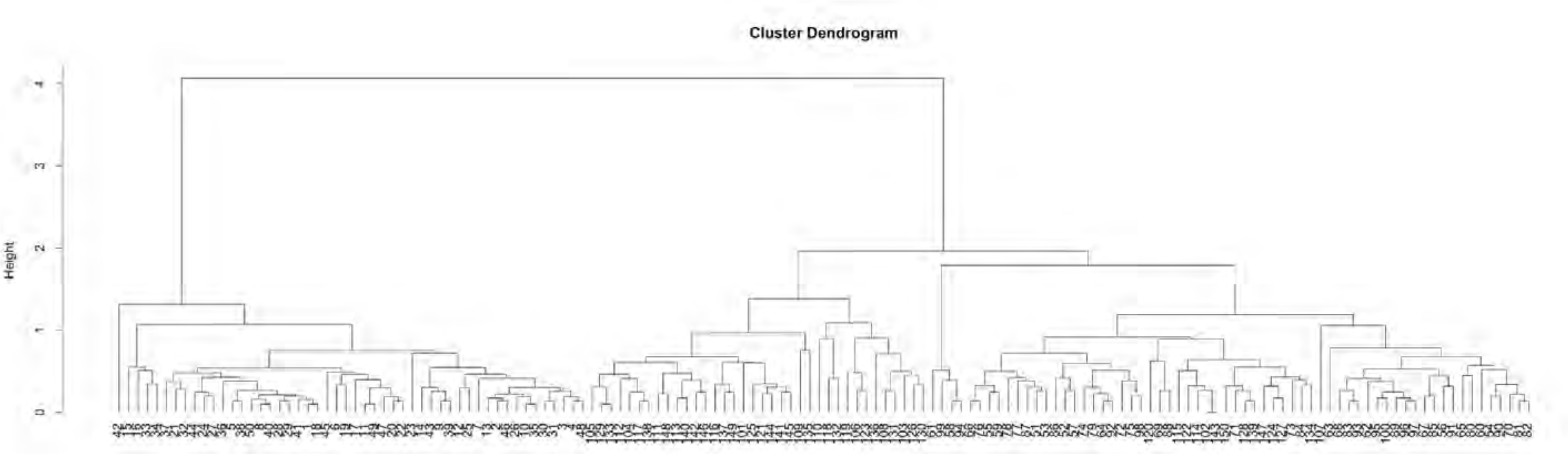
Viewing in the environment browser:



```
ihfit                    List of 7
  merge : int [1:149, 1:2] -102 -8 -1 -10 -129 -11 -5 -20 -30 -58 ...
  height : num [1:149] 0 0.1 0.1 0.1 0.1 ...
  order : int [1:150] 42 15 16 33 34 37 21 32 44 24 ...
  labels : NULL
  method : chr "average"
  call : language hclust(d = dist(niris[, 1:4]), method = "ave")
  dist.method: chr "euclidean"
  attr(*, "class")= chr "hclust"
```

Hierarchical clustering in R

Dendrogram:



Where are the clusters?

How many do you want?

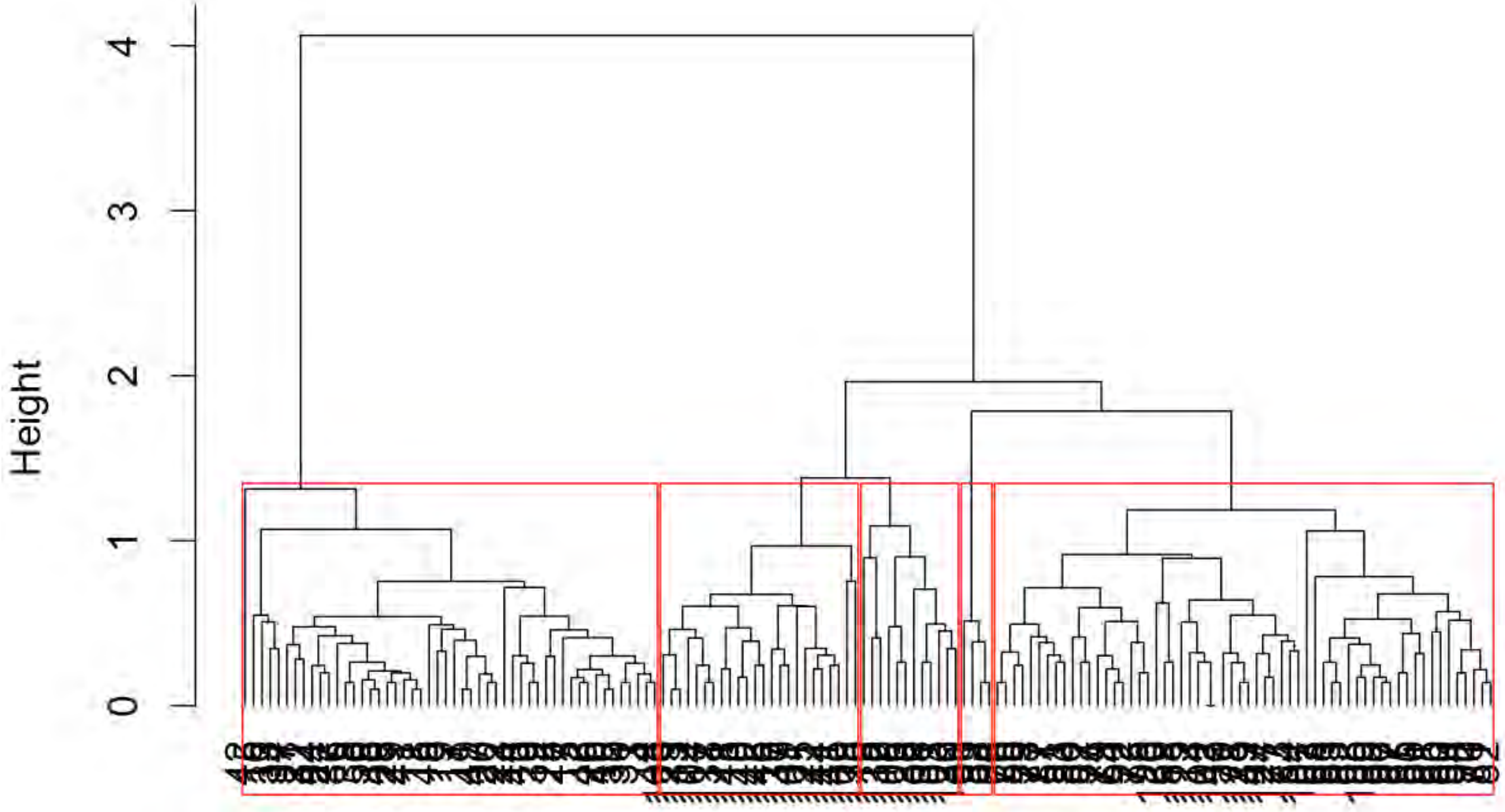
Hierarchical clustering in R

Setting a particular number of clusters:

```
> # pruning the tree into 5 clusters
> cutihfit = cutree(ihfit, k = 5)
> rect.hclust(ihfit, k = 5, border = "red")
> table(actual = niris$Species, fitted = cutihfit)
            fitted
actual        1  2  3  4  5
  setosa     50  0  0  0  0
  versicolor  0 46  4  0  0
  virginica   0 14  0 24 12
```

Hierarchical clustering in R

Dendrogram showing 5 clusters:



? hclust *

- Description

  Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.

- Usage

```
hclust(d, method = "complete", members = NULL)

d            dissimilarity structure
method       agglomeration method (many to choose)
...
```

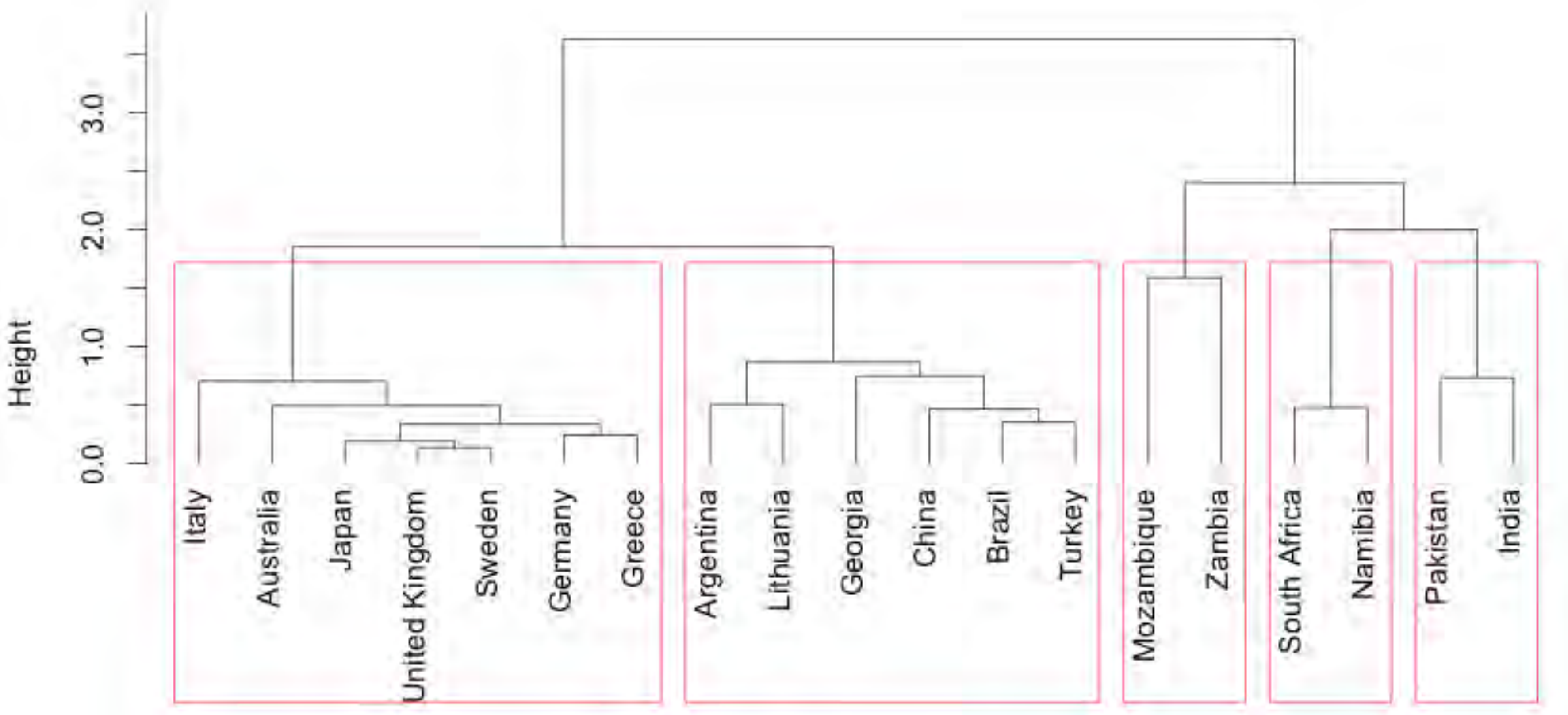
Countries data (scaled)

Reading data, scaling, setting row names to country names (to appear in dendrogram)

```
>   CD <- read.csv("CountriesData.csv")
>   CD[,2:5] = scale(CD[,2:5])
>   rownames(CD) = CD$Country
>   hfit = hclust(dist(CD[,2:5]), "average")
>   plot(hfit)
>   plot(hfit, hang = -1)
>   hfit = cutree(hfit, k = 5) #Pruning
>   rect.hclust(hfit, k = 5, border = "red")
```

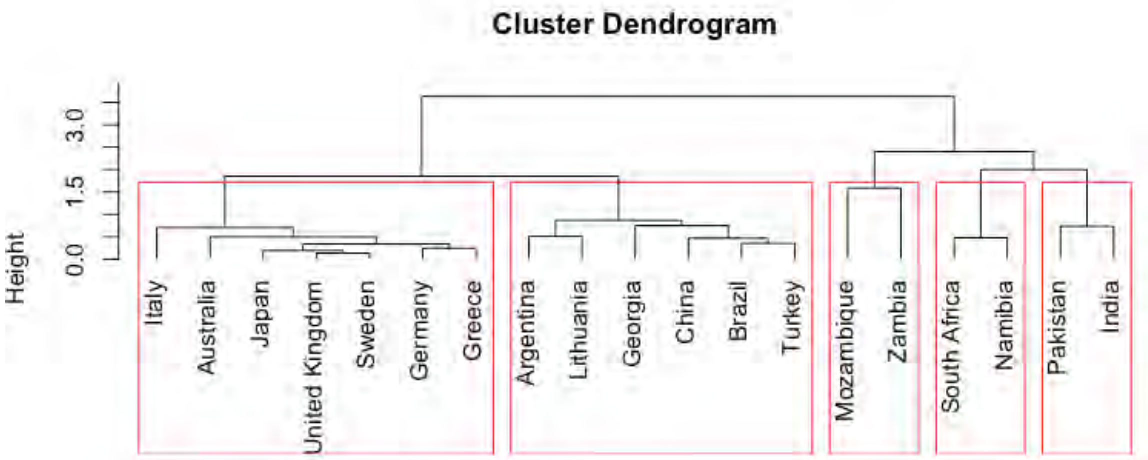
Countries data (scaled)

Dendrogram

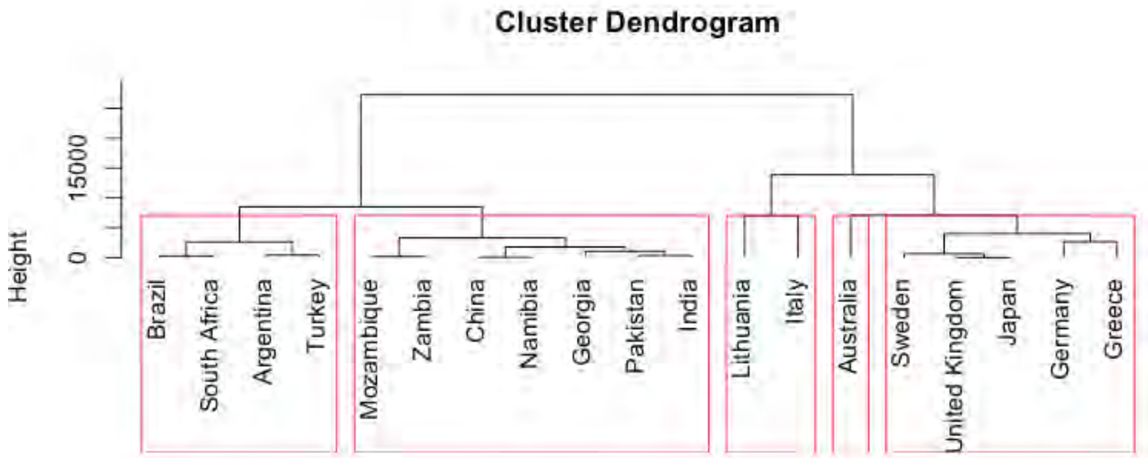


Countries data: effect of scaling

Scaled



Not-scaled



Countries data (normalised)

Normalised input gives similar tree to scaled data:

```
>   CD <- read.csv("CountriesData.csv")
>   # for loop to normalise cols 2 - 5
>   for (i in 2:5){
>   CD[,i] = (CD[,i]-min(CD[,i]))/(max(CD[,i])-min(CD[,i]))
>   }
>   rownames(CD) = CD$CountryCD
>   hfit = hclust(dist(CD[,2:5]), "average")
>   ...
```

Closing remarks

Clustering:

- An important unsupervised learning tool for grouping data.
- Enables data reduction (i.e. to identify representative subsets of the data).

Many R packages for cluster analysis:

- Cluster – is one of these which gives more control over clustering algorithm and additional analysis tools.

Solutions to review questions

1. D
2. E
3. C
4. A
5. A

References to this lecture

- James et al., An Introduction to Statistical Learning with Applications in R, 2nd Ed. Springer, 2021. Section 12.4.
- Giordani, Ferraro and Martella, An Introduction to Clustering with R. Springer, 2020.

download via Library

Notes on the presentation

This presentation contains slides created to accompany: *Introduction to Data Mining*, Tan, Steinbach, Kumar. Pearson Education Inc., 2006.

Presentation originally created by Dr. Sue Bedingfield, with additions by Rui Jie Chow & Dr. Parthan Kasarapu.