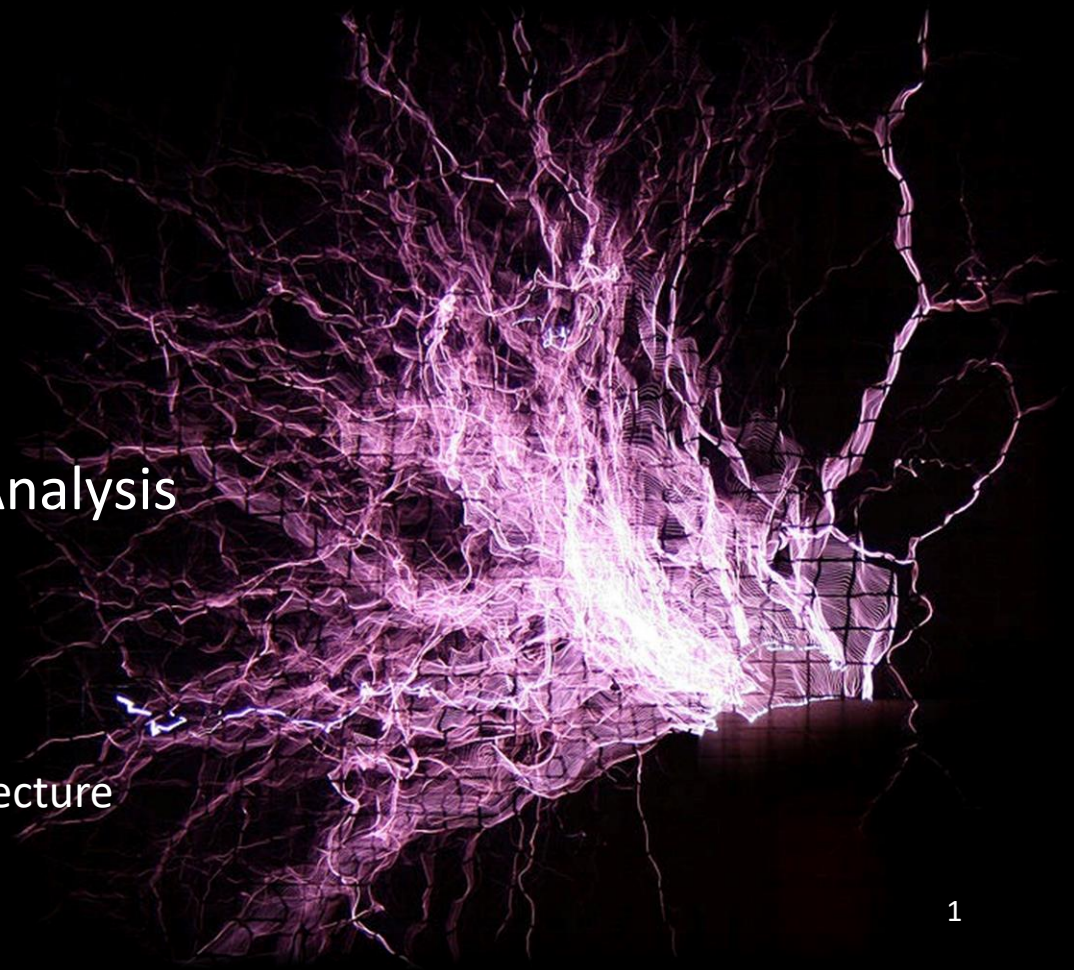
CASA0007: Quantitative Methods

Dr Huanfa CHEN
huanfa.chen@ucl.ac.uk

Dr Hannah Fry
hannah.fry@ucl.ac.uk

Moodle password: QM2020

Centre for Advanced Spatial Analysis



You are all ‘attendant’ for this week’s lecture

https://commons.wikimedia.org/wiki/File:Tesla_coil_Vienna_Technical_Museum_2007_a.jpg

LECTURE 6
Cluster Analysis

https://commons.wikimedia.org/wiki/Category:Tesla_coils#/media/File:Tesla_coil_Vienna_Technical_Museum_2007_b.jpg

Week 1: Introduction to Quantitative Problems

Week 2: Approaching & Communicating Data

Week 3: Measuring Relationships

Week 4: Advanced Regression

Week 5: Hypothesis Testing

READING WEEK

Week 6: Cluster Analysis

Week 7: Optimising Limited Resources

Week 8: Modelling the World

Week 9: Statistical Traps & Advanced Topics

PRESENTATION WEEK

OBJECTIVES

1. Understand the purpose of cluster analysis and know the algorithms for k-means and hierarchical clustering.

2. Understand the concept of data standardisation and recognise different approaches.

3. Understand how SSE and silhouette scores can be used to assess the quality of a clustering.

4. Consider how to structure a piece of quantitative writing.

https://commons.wikimedia.org/wiki/File:Tesla_coil_Vienna_Technical_Museum_2007_c.jpg

- Part 1: Hypothesis testing recap
- Part 2: Cluster analysis – why should I care?
- Part 3: Before you start
- Part 4: K means
- Part 5: Hierarchical clustering
- Part 6: How good are your clusters?
- Part 7: Some tips and tricks for your written work

Last time: Hypothesis Testing

Five Steps

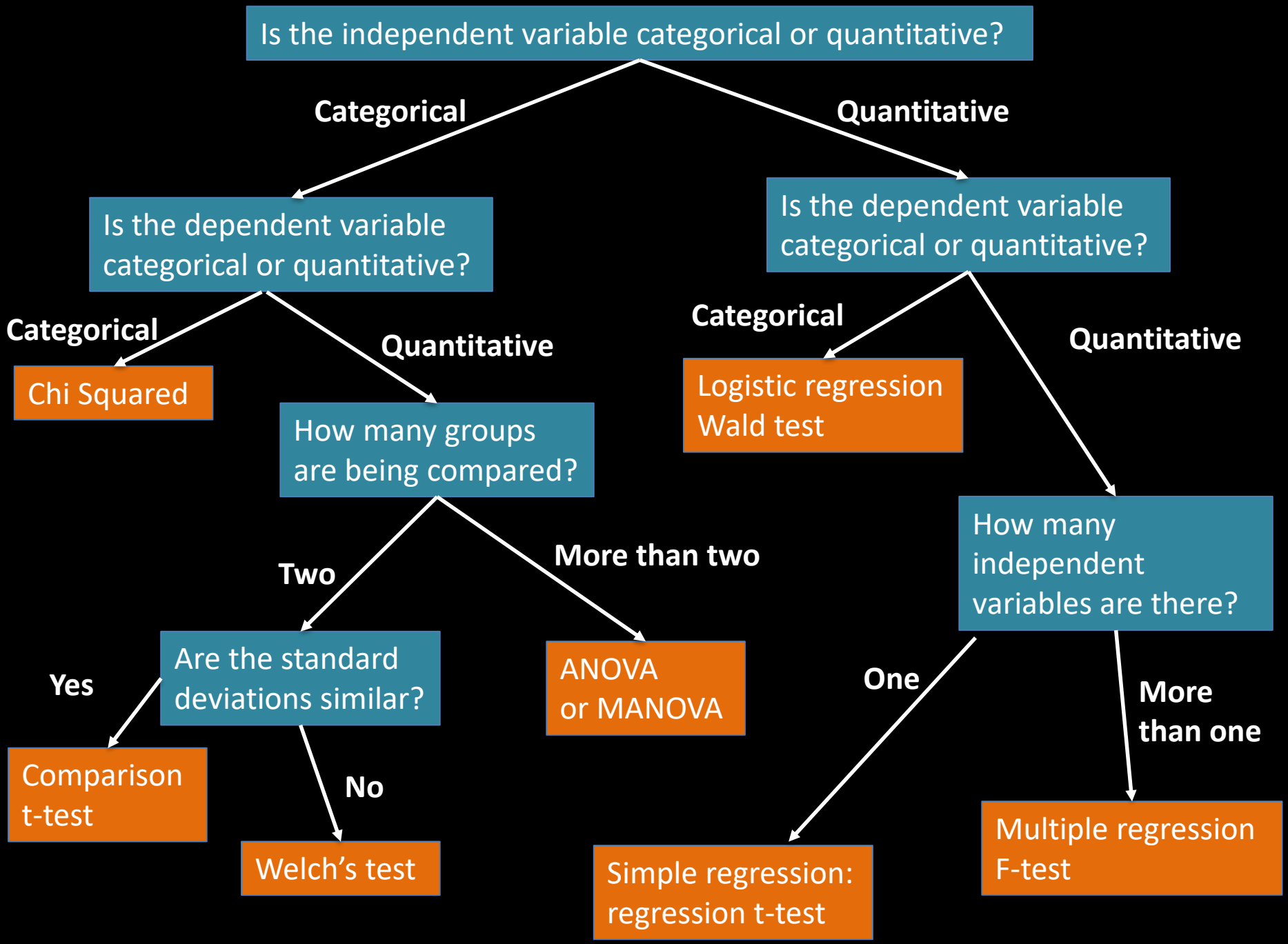
1. State your hypotheses.

2. State your significance level: α

3. What is the evidence (E)? (The Test Statistic)

4. Calculate the probability of seeing evidence at least as extreme as E, *if* H_0 *is true*. (The "p-value")

5. If the p-value is smaller than the significance, reject H_0 and accept H_1. Otherwise there is not enough evidence to reject H_0.


Is the independent variable categorical or quantitative?
Categorical
Quantitative
Is the dependent variable categorical or quantitative?
Is the dependent variable categorical or quantitative?
Categorical
Quantitative
Chi Squared
Categorical
Quantitative
Logistic regression
Wald test
How many groups are being compared?
How many independent variables are there?
Two
More than two
Yes
One
More than one
Are the standard deviations similar?
ANOVA or MANOVA
Comparison t-test
No
Welch's test
Simple regression: regression t-test
Multiple regression F-test


Pfizer
COVID - 19
Coronavirus
Vaccine
Injection only
COVID - 19
Coronavirus
Vaccine
Injection only
COVID - 19
Coronavirus
Vaccine
Injection only
COVID - 19
Coronavirus
Vaccine
Injection only

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

1. H0:?

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

1. H0: The vaccine has no effect

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

1. H0: The vaccine has no effect
   H1: The vaccine has some effect

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

1. H0: The vaccine has no effect
   H1: The vaccine has some effect

2. α = 0.000001 (That is, willing to tolerate a one in a million chance that we'd see these results if the null hypothesis were true)

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

1. H0: The vaccine has no effect
   H1: The vaccine has some effect

2. α = 0.000001 (That is, willing to tolerate a one in a million chance that we'd see these results if the null hypothesis were true)

3. The evidence is as above. Calculate the test statistic.

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

1. H0: The vaccine has no effect
   H1: The vaccine has some effect

2. α = 0.000001 (That is, willing to tolerate a one in a million chance that we'd see these results if the null hypothesis were true)

3. The evidence is as above. Calculate the test statistic.

4. Calculate the p value

5. If p< α then reject the null hypothesis.

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

1. H0: The vaccine has no effect
   H1: The vaccine has some effect

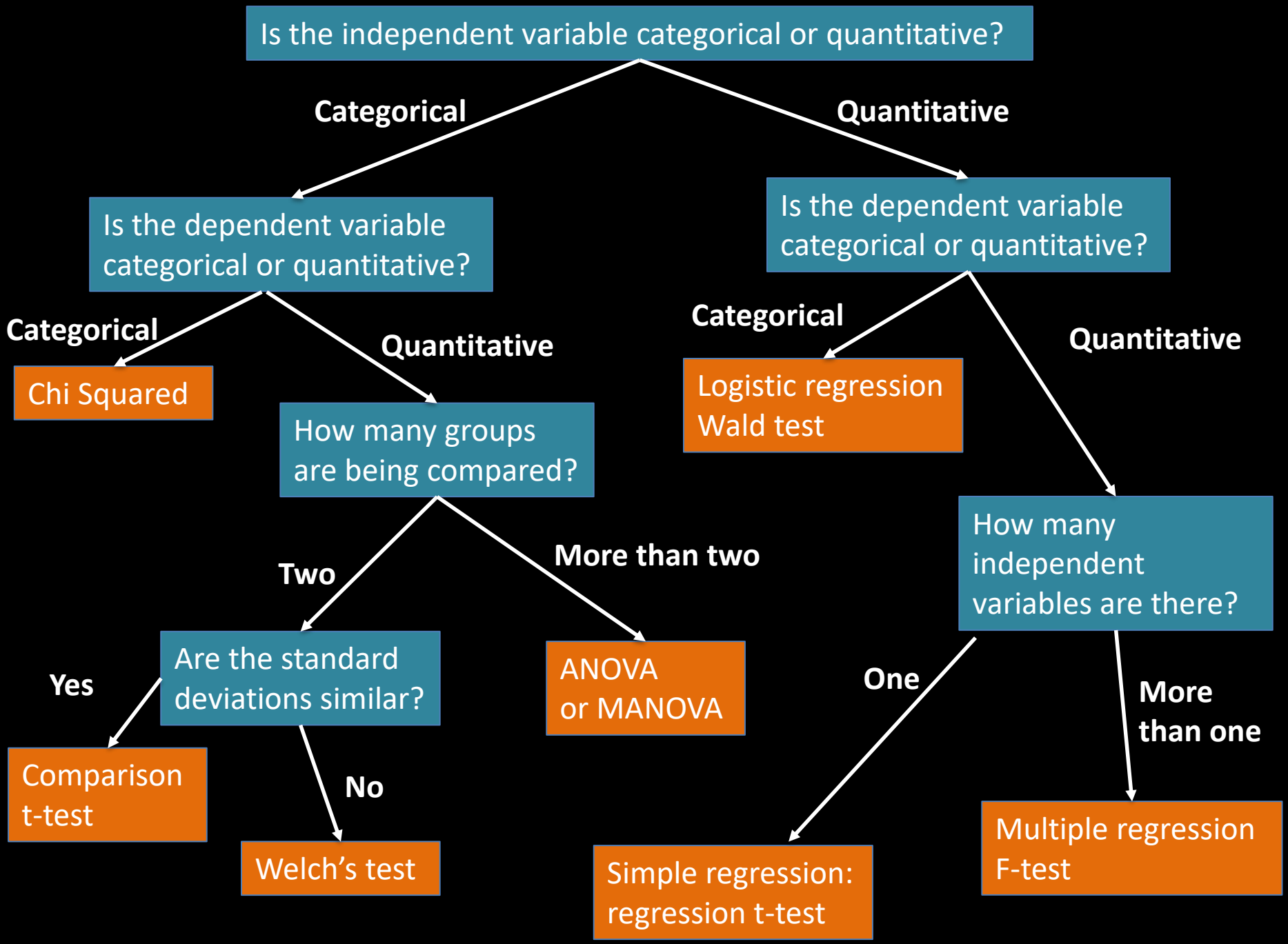
2. α = 0.000001 (That is, willing to tolerate a one in a million chance that we'd see these results if the null hypothesis were true)

3. The evidence is as above. Calculate the test statistic.

4. Calculate the p value

5. If p< α then reject the null hypothesis.

Only the bits in yellow change. Everything else is the same every time


Is the independent variable categorical or quantitative?
Categorical
Quantitative
Is the dependent variable categorical or quantitative?
Is the dependent variable categorical or quantitative?
Categorical
Quantitative
Chi Squared
Categorical
Quantitative
Logistic regression
Wald test
How many groups are being compared?
How many independent variables are there?
Two
More than two
Yes
One
More than one
Are the standard deviations similar?
ANOVA or MANOVA
Comparison t-test
No
Welch's test
Simple regression: regression t-test
Multiple regression F-test

The Test Statistic

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

	Vaccine	Placebo
Symptoms		
No symptoms		

Independent variable

Dependent variable

The Test Statistic – Chi squared

40,000 participants
Split equally between vaccine and placebo

Of those who had symptoms:

8 people in the vaccine group
86 people in the placebo group

	Vaccine	Placebo
Symptoms	8	86
No symptoms	19992	19914

The Test Statistic – Chi squared

OBSERVED	Vaccine	Placebo
Symptoms	8	86
No symptoms	19992	19914

94 sick

The Test Statistic – Chi squared

OBSERVED	Vaccine	Placebo
Symptoms	8	86
No symptoms	19992	19914

94 sick

If the Null Hypothesis were true..

EXPECTED	Vaccine	Placebo
Symptoms		
No symptoms		

The Test Statistic – Chi squared

OBSERVED	Vaccine	Placebo
Symptoms	8	86
No symptoms	19992	19914

94 sick

If the Null Hypothesis were true..

EXPECTED	Vaccine	Placebo
Symptoms	47	47
No symptoms		

The Test Statistic – Chi squared

OBSERVED	Vaccine	Placebo
Symptoms	8	86
No symptoms	19992	19914

94 sick

If the Null Hypothesis were true..

EXPECTED	Vaccine	Placebo
Symptoms	47	47
No symptoms	19953	19953

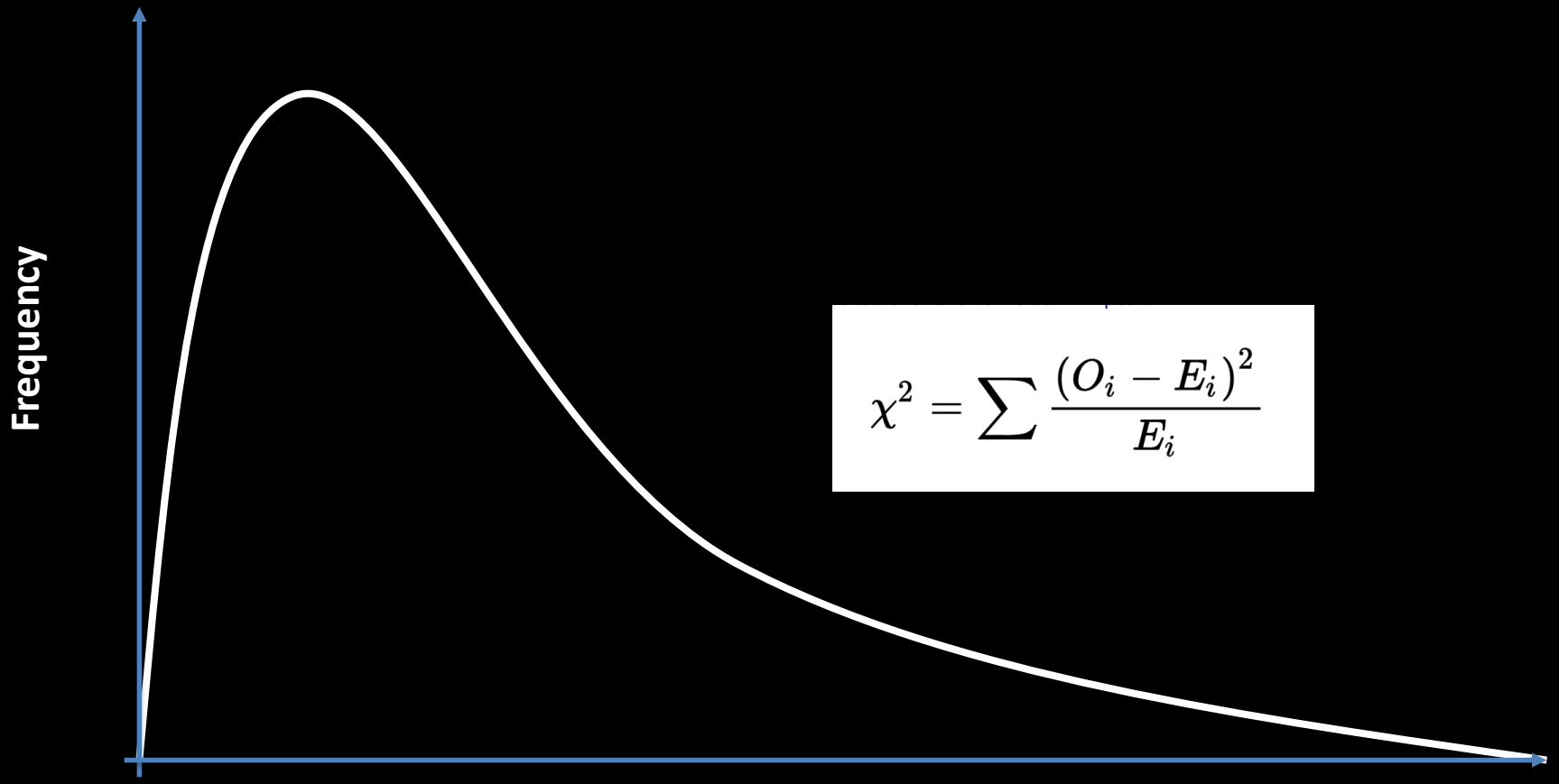
The Test Statistic – Chi squared

OBSERVED	Vaccine	Placebo
Symptoms	8	86
Symptoms	47	47
No symptoms	19953	19953

The Chi Squared statistic measures how much the observed values deviate from what would be expected if the null hypothesis were true

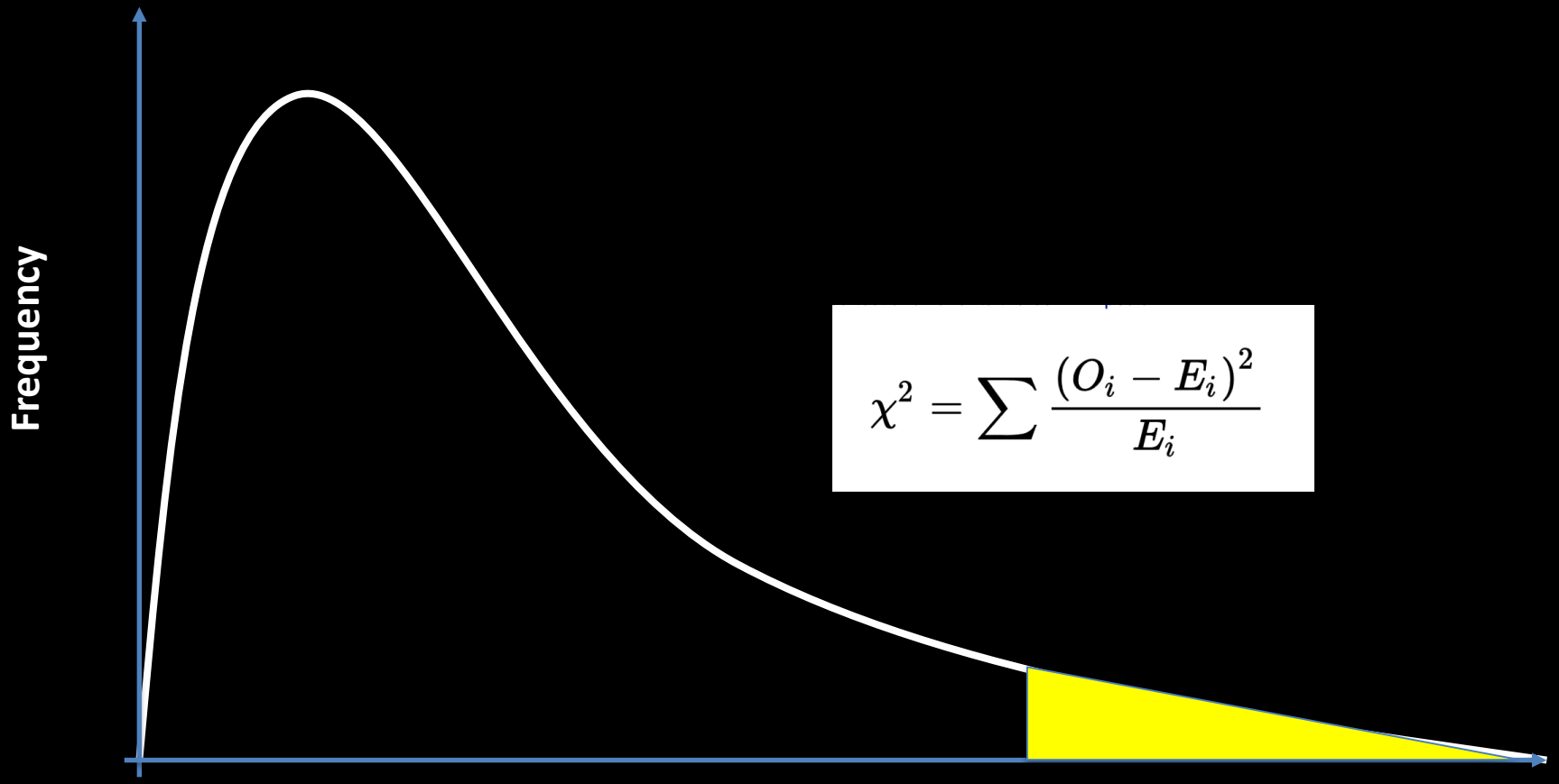
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi squared – the distribution



$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

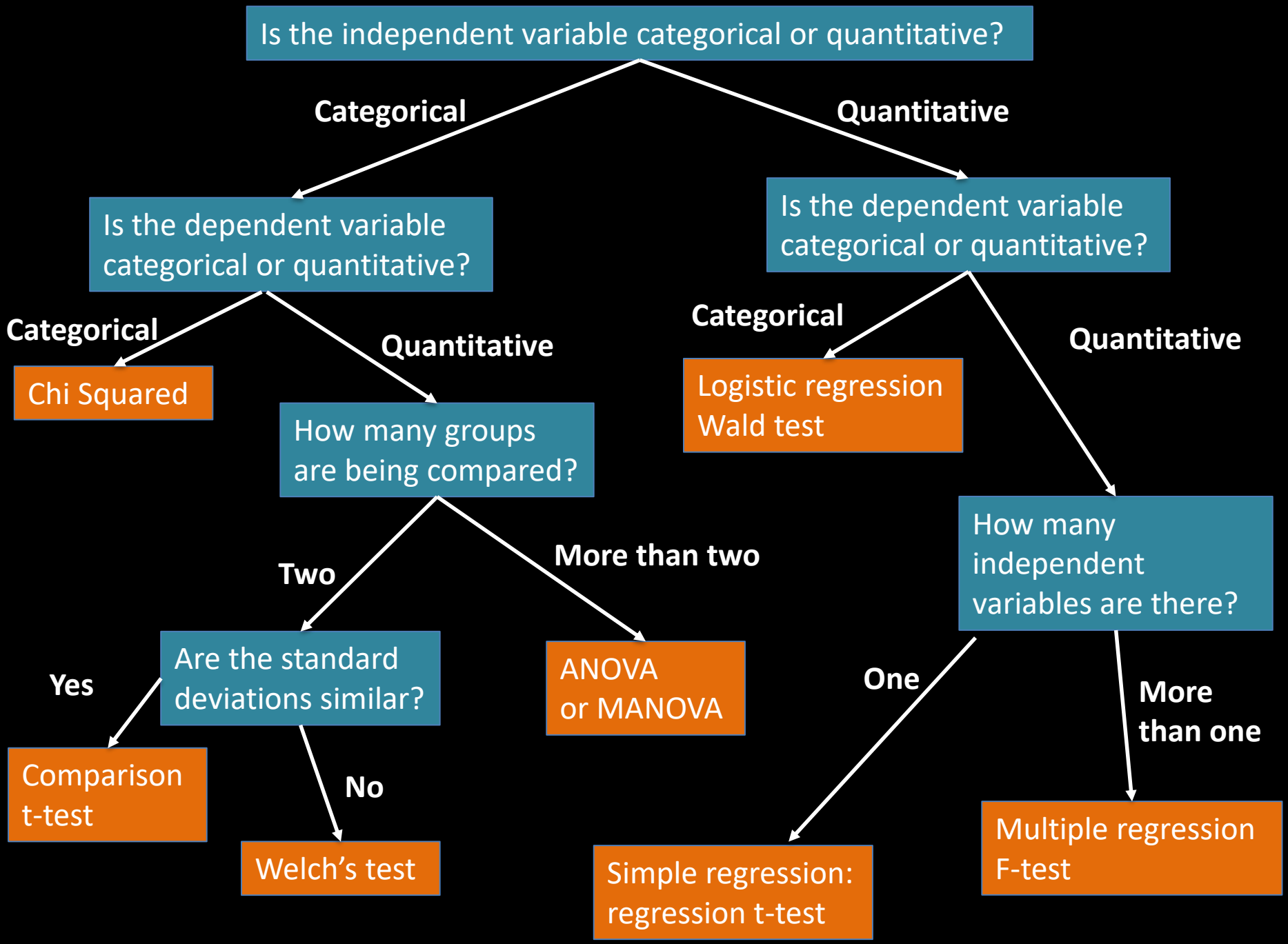
Chi squared – the distribution



$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

TO PYTHON!

QUIZ!


Is the independent variable categorical or quantitative?
Categorical
Quantitative
Is the dependent variable categorical or quantitative?
Is the dependent variable categorical or quantitative?
Categorical
Quantitative
Chi Squared
Categorical
Quantitative
Logistic regression
Wald test
How many groups are being compared?
How many independent variables are there?
Two
More than two
Yes
One
More than one
Are the standard deviations similar?
ANOVA or MANOVA
Comparison t-test
No
Welch's test
Simple regression: regression t-test
Multiple regression F-test

Lecture 5 – Assignment

Read the following four articles (available on Moodle):

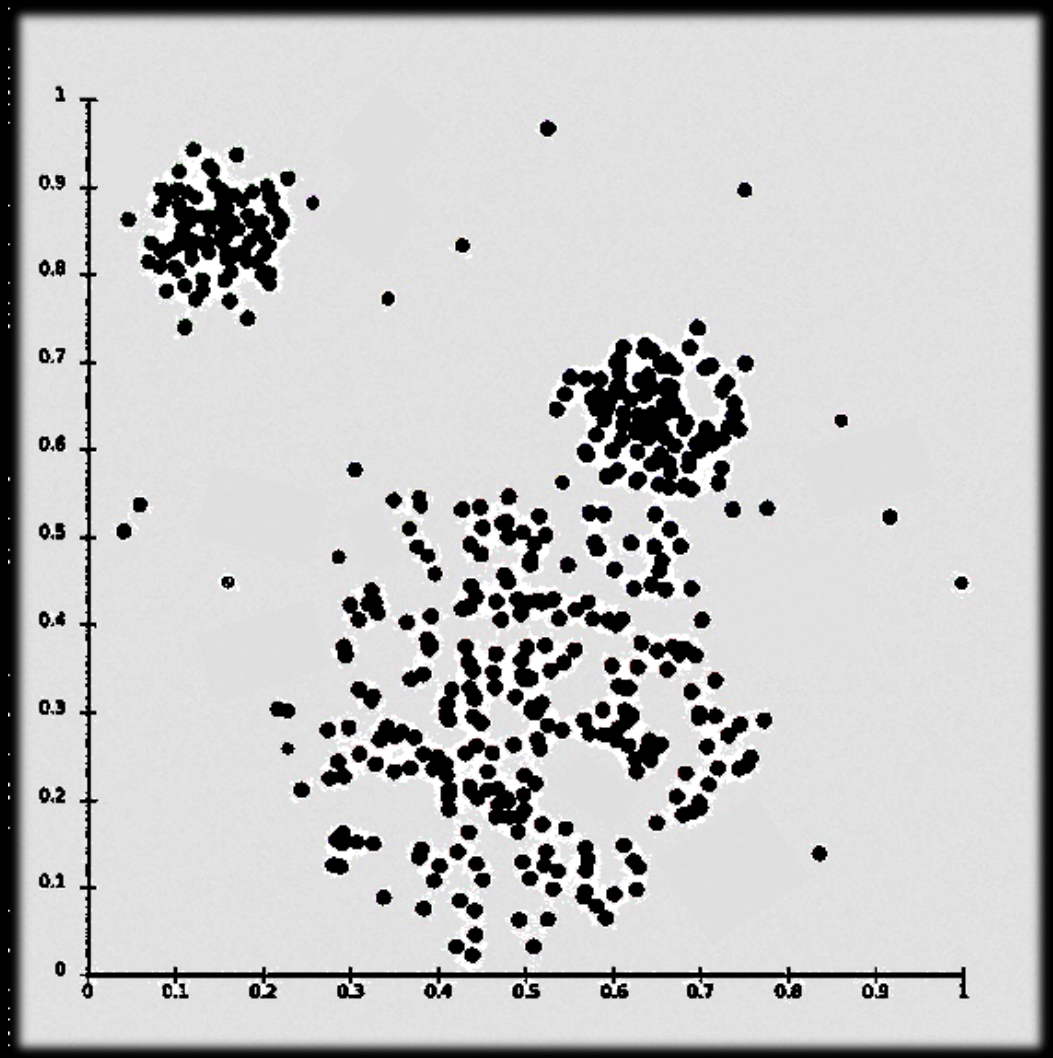
London murders: a predictable pattern? (Spiegelhalter & Barnet, 2009)
Have London's roads become more dangerous for cyclists? (Aberdeing & Spiegelhalter, 2009)
https://www.newyorker.com/magazine/2019/09/09/what-statistics-can-and-cant-tell-us-about-ourselves (Fry)
Science isn't broken (fivethirtyeight.com)

Consider what can be learnt from these articles on understanding and communicating data and on hypothesis testing more specifically.

It’s Tea Time

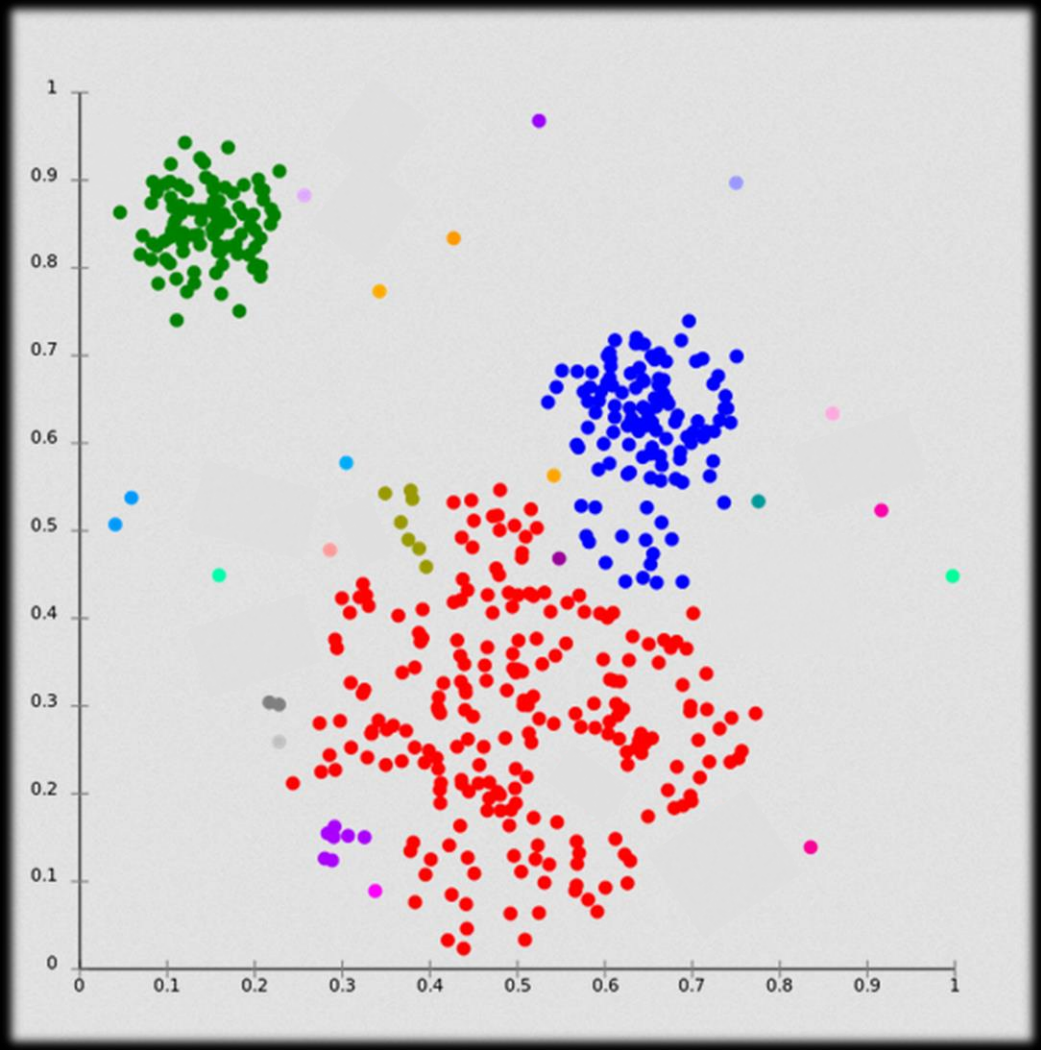
- **Part 1: Hypothesis testing recap**
- **Part 2: Cluster analysis – why should I care?**
- **Part 3: Before you start**
- **Part 4: K means**
- **Part 5: Hierarchical clustering**
- **Part 6: How good are your clusters?**
- **Part 7: Some tips and tricks for your written work**

Clustering – Concept



https://commons.wikimedia.org/wiki/File:SLINK-Gaussian-data.svg

Clustering – Concept



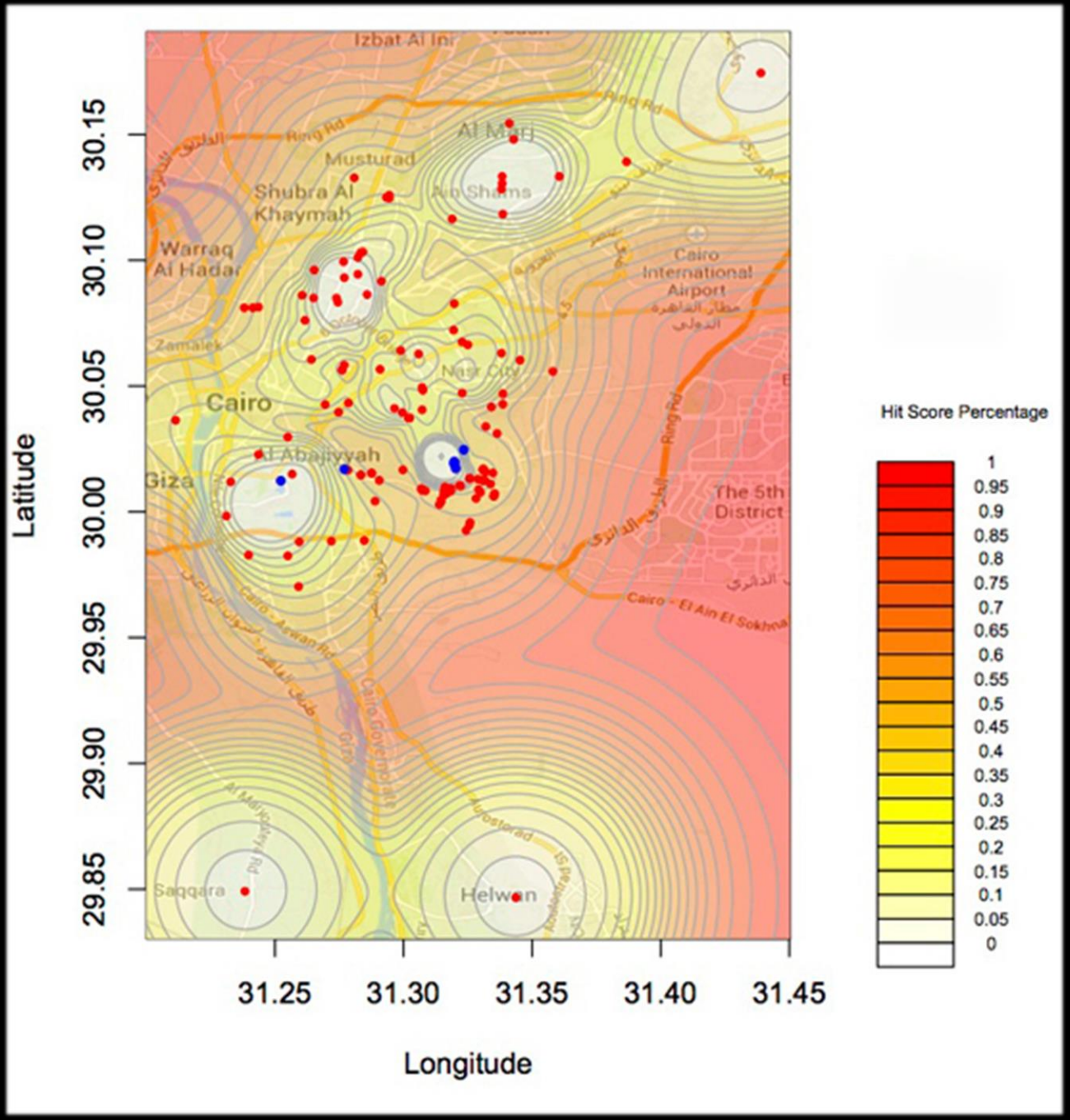
https://commons.wikimedia.org/wiki/File:SLINK-Gaussian-data.svg

Clustering

Definition:

Type of analysis that divides observations into on groups based on some similarity criteria (distance).

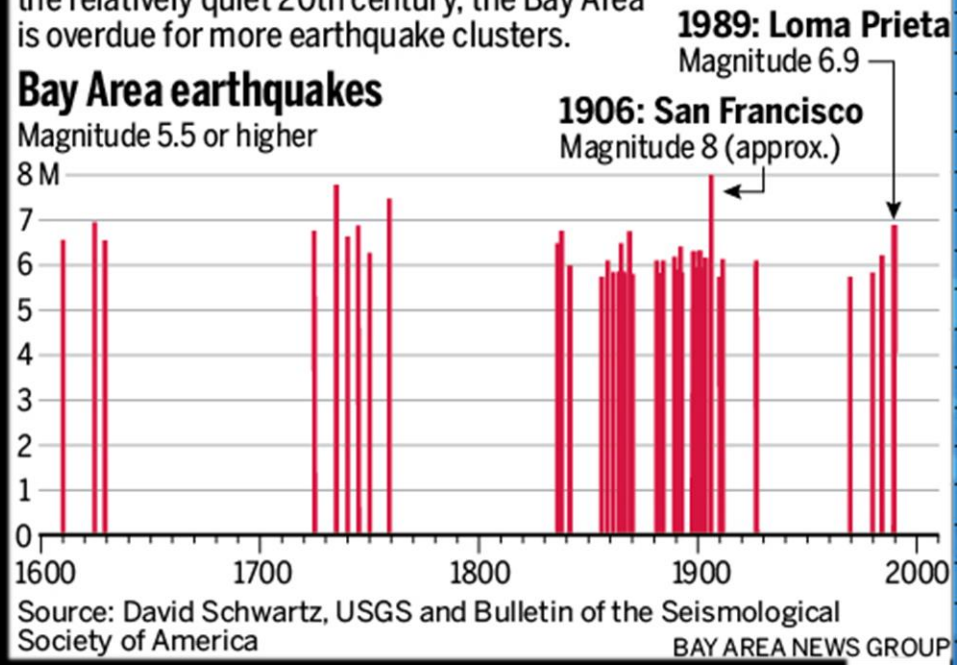
Examples: K-Means, Hierarchical clustering



http://www.qmul.ac.uk/media/news/items/se/130968.html

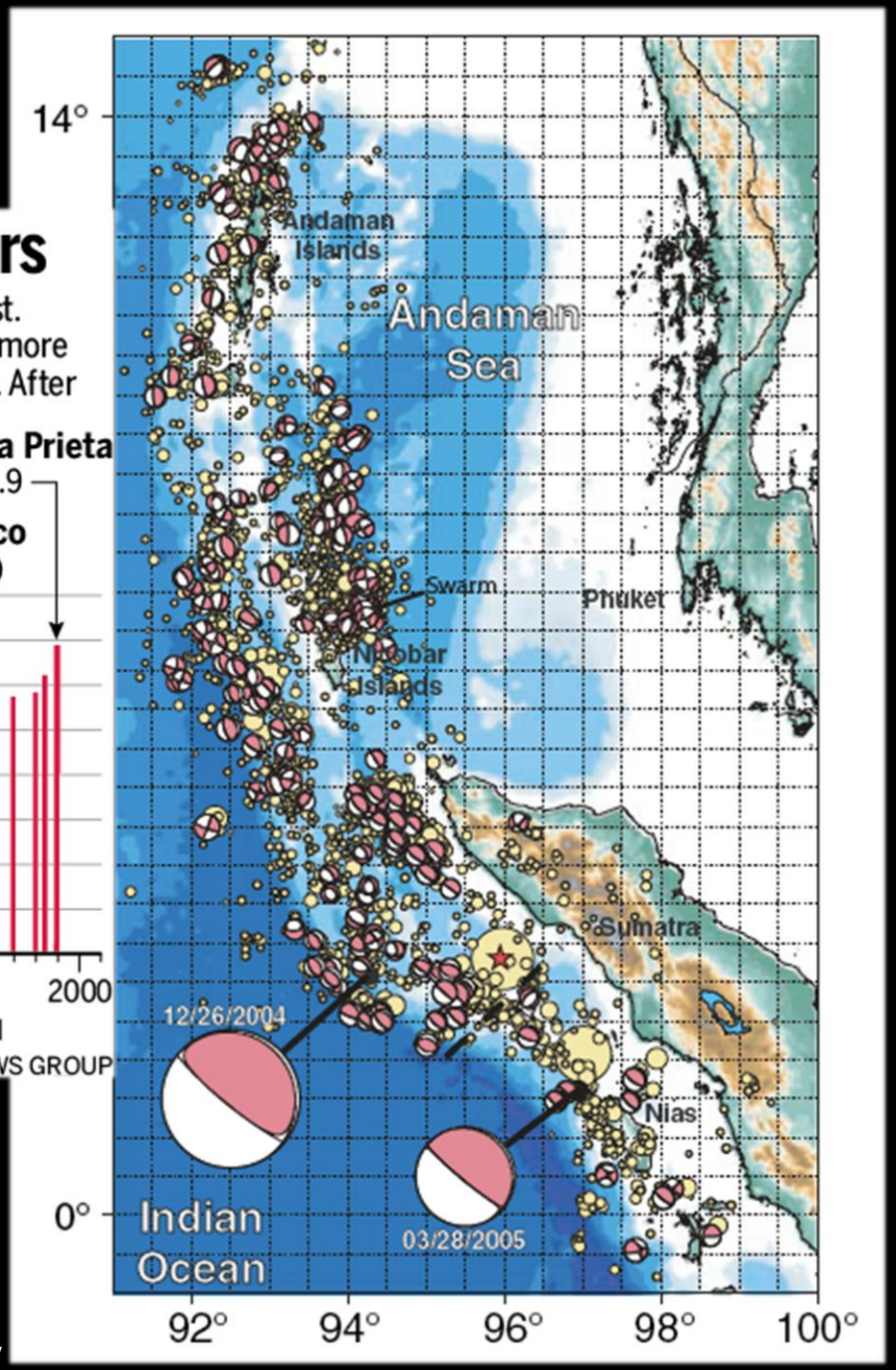
Earthquakes may come in clusters

Clusters of large earthquakes rattled the Bay Area in the past. Scientists suspect that these groupings of earthquakes are more common than single, giant earthquakes like the one in 1906. After the relatively quiet 20th century, the Bay Area is overdue for more earthquake clusters.



Source: David Schwartz, USGS and Bulletin of the Seismological Society of America

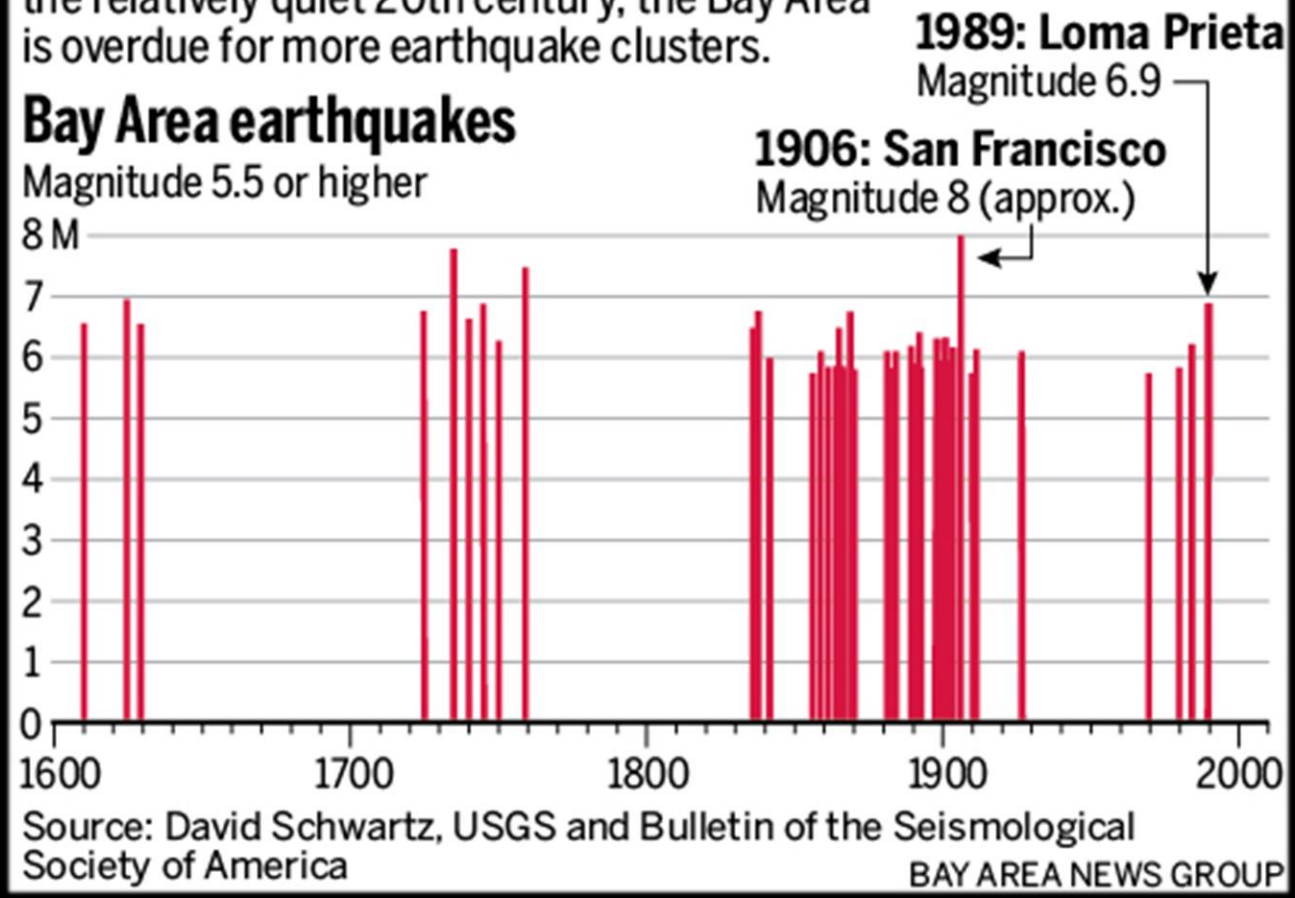
BAY AREA NEWS GROUP



http://www.mercurynews.com/2014/05/19/earthquake-cluster-likely-to-strike-bay-area-scientists-say/
http://science.sciencemag.org/content/308/5725/1127/tab-pdf

Earthquakes may come in clusters

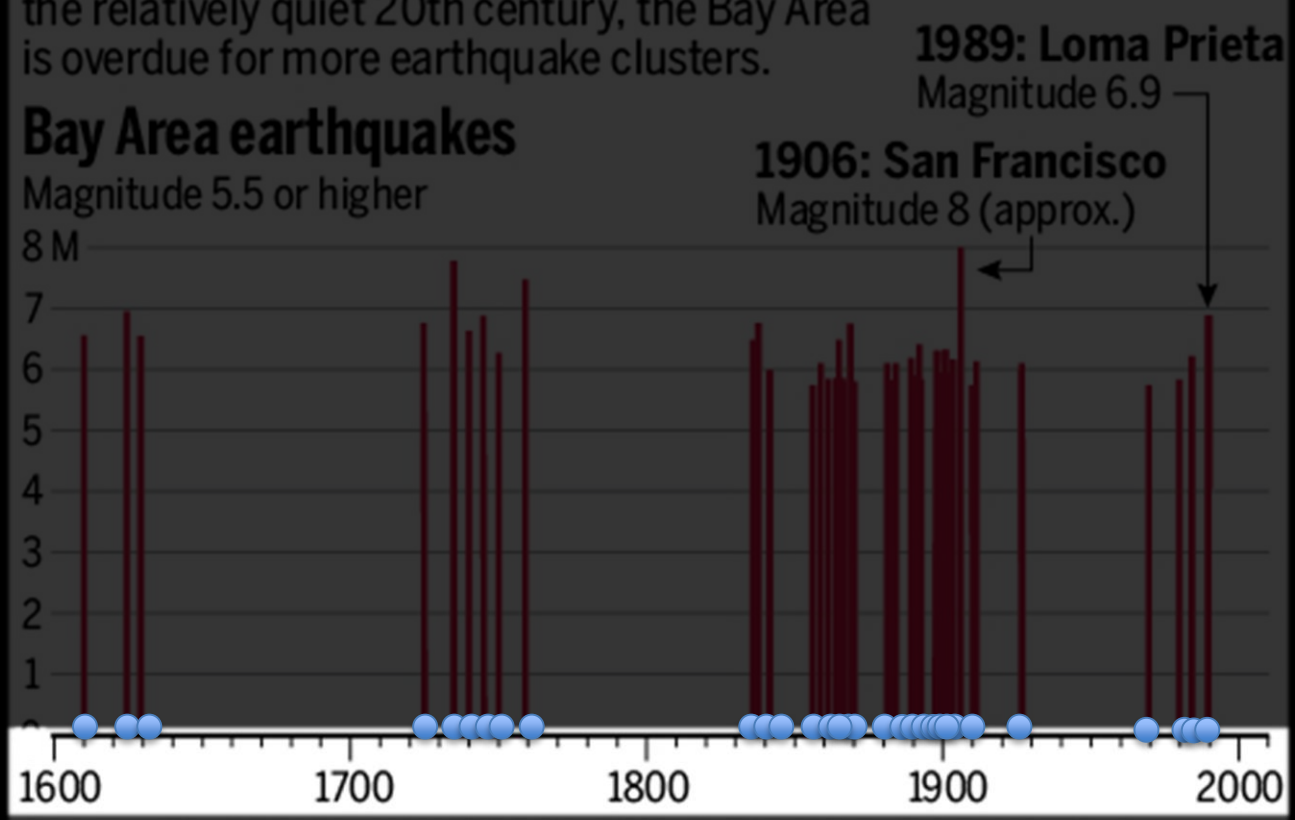
Clusters of large earthquakes rattled the Bay Area in the past. Scientists suspect that these groupings of earthquakes are more common than single, giant earthquakes like the one in 1906. After the relatively quiet 20th century, the Bay Area is overdue for more earthquake clusters.



http://www.mercurynews.com/2014/05/19/earthquake-cluster-likely-to-strike-bay-area-scientists-say/
http://science.sciencemag.org/content/308/5725/1127/tab-pdf

Earthquakes may come in clusters

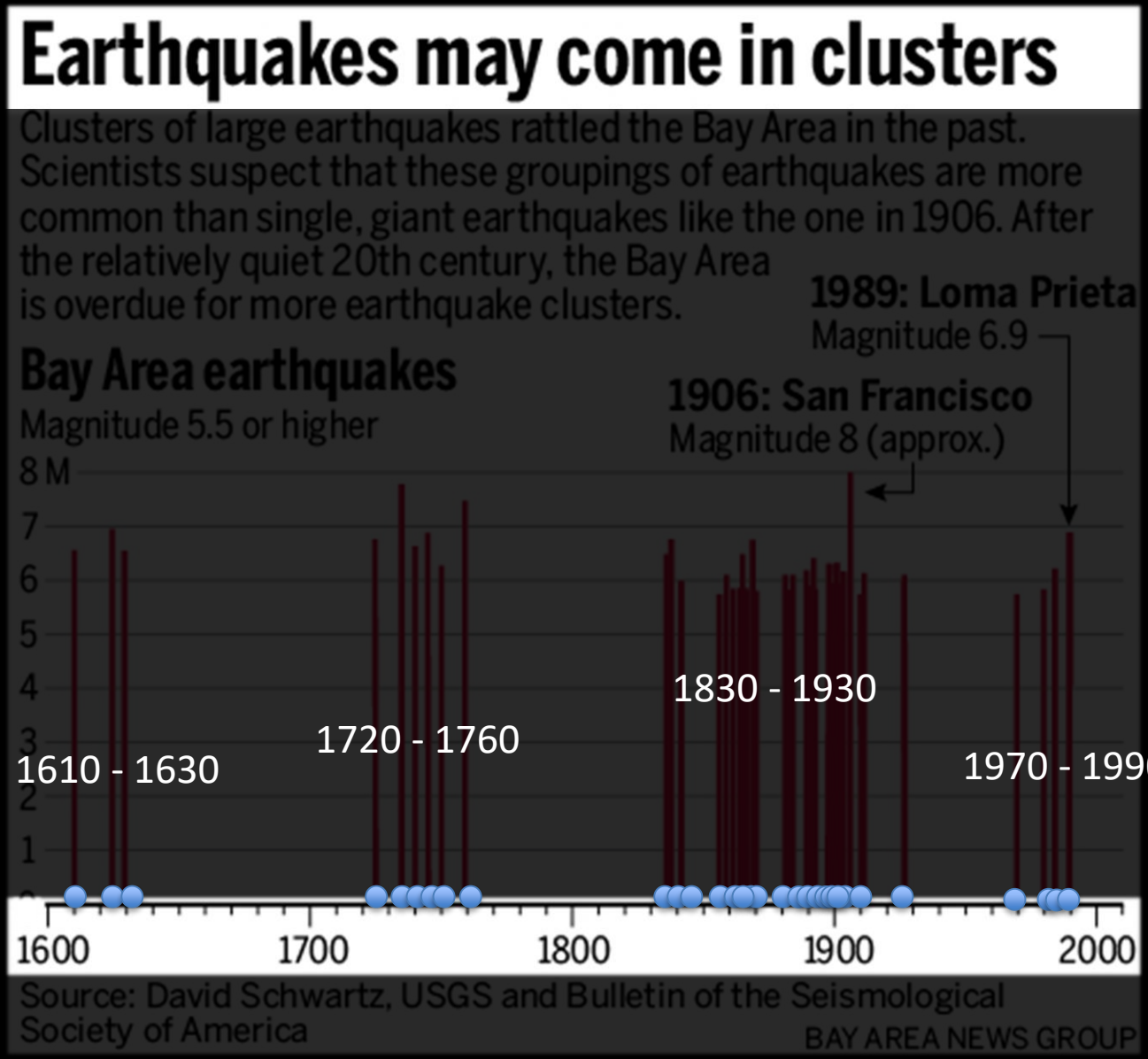
Clusters of large earthquakes rattled the Bay Area in the past. Scientists suspect that these groupings of earthquakes are more common than single, giant earthquakes like the one in 1906. After the relatively quiet 20th century, the Bay Area is overdue for more earthquake clusters.



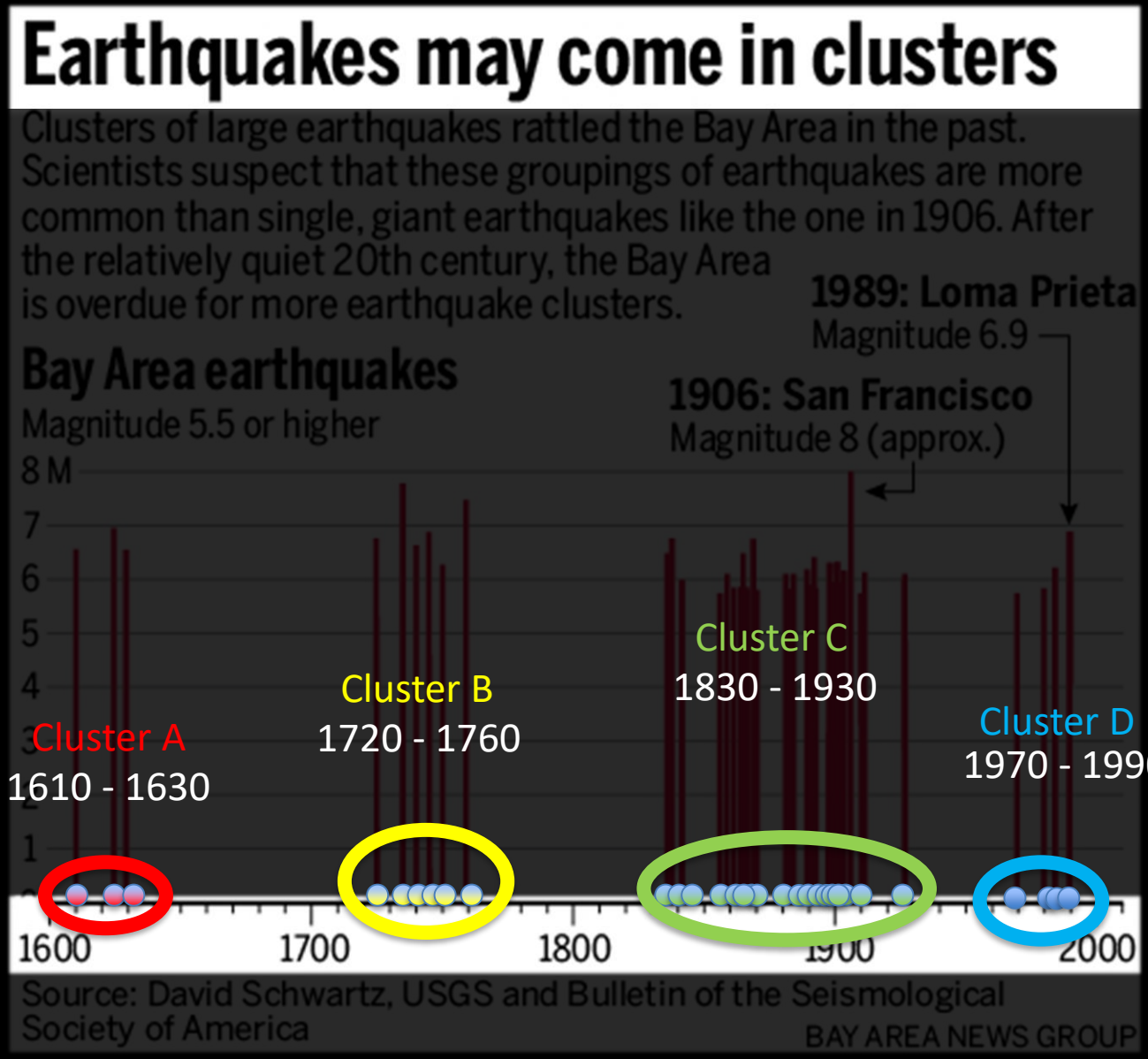
Source: David Schwartz, USGS and Bulletin of the Seismological Society of America

BAY AREA NEWS GROUP

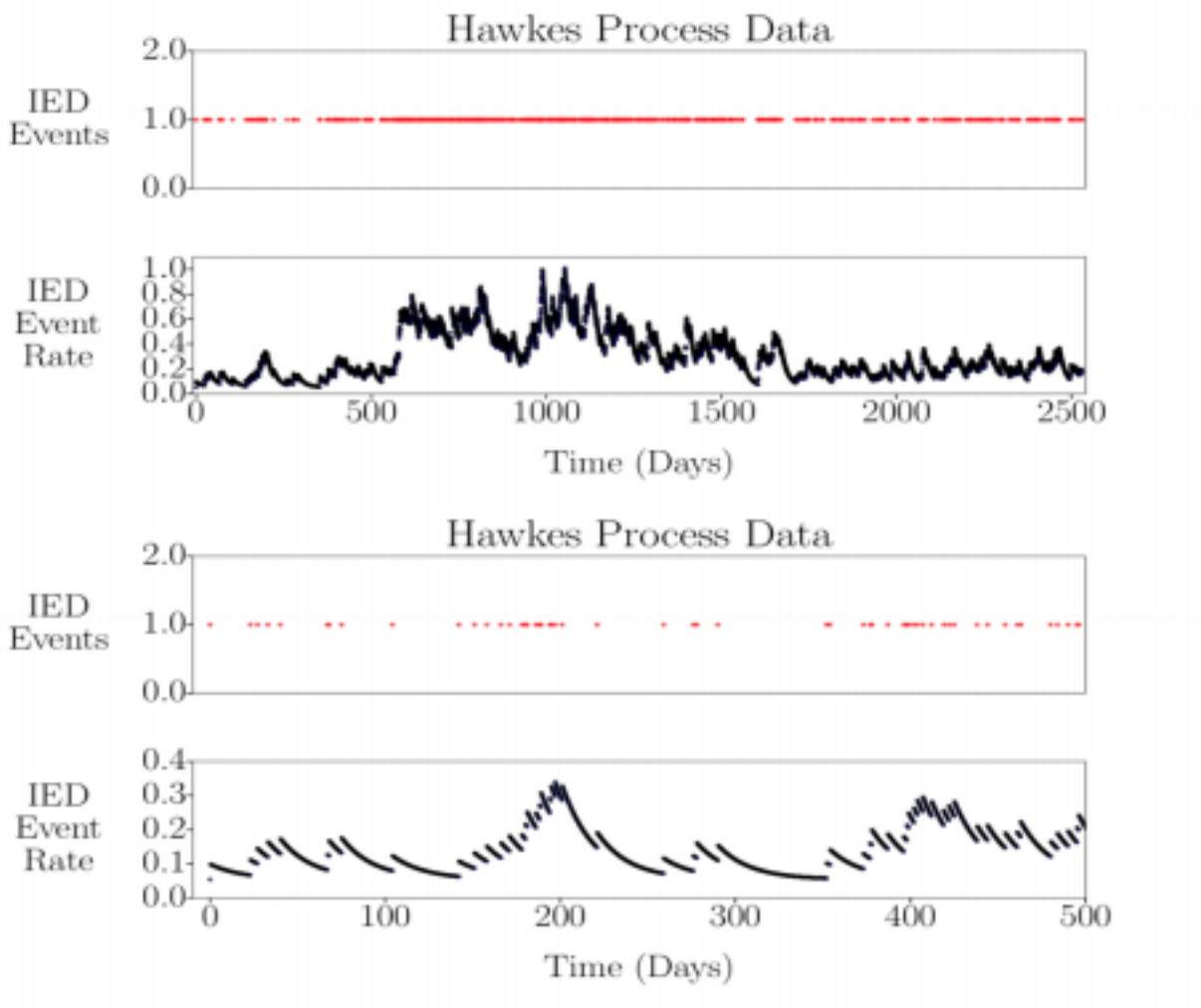
http://www.mercurynews.com/2014/05/19/earthquake-cluster-likely-to-strike-bay-area-scientists-say/
http://science.sciencemag.org/content/308/5725/1127/tab-pdf



http://www.mercurynews.com/2014/05/19/earthquake-cluster-likely-to-strike-bay-area-scientists-say/
http://science.sciencemag.org/content/308/5725/1127/tab-pdf



http://www.mercurynews.com/2014/05/19/earthquake-cluster-likely-to-strike-bay-area-scientists-say/
http://science.sciencemag.org/content/308/5725/1127/tab-pdf



Tench, S., Fry, H. & Gill, P. *Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army* (2015) European Journal of Applied Mathematics [PDF]

Cooper, M., Foote, J., Girgensohn, A., & Wilcox, L. (2005). *Temporal event clustering for digital photo collections*. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 1(3), 269-288.

TikTok

For You

Following

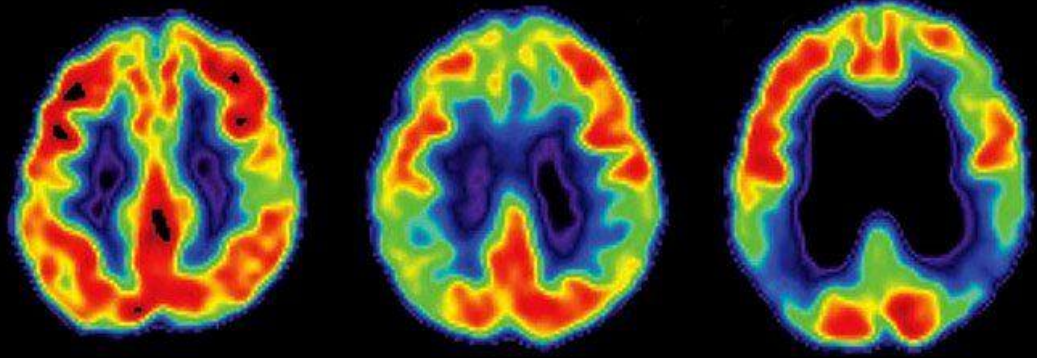
Who to follow

Cerys Bradley
@hashtagcerys
Follow

Frank Farris
@Farris_Frank
Follow

Lewisham Music
@LewishamMusic
Follow

Show more



Normal

Mild cognitive impairment

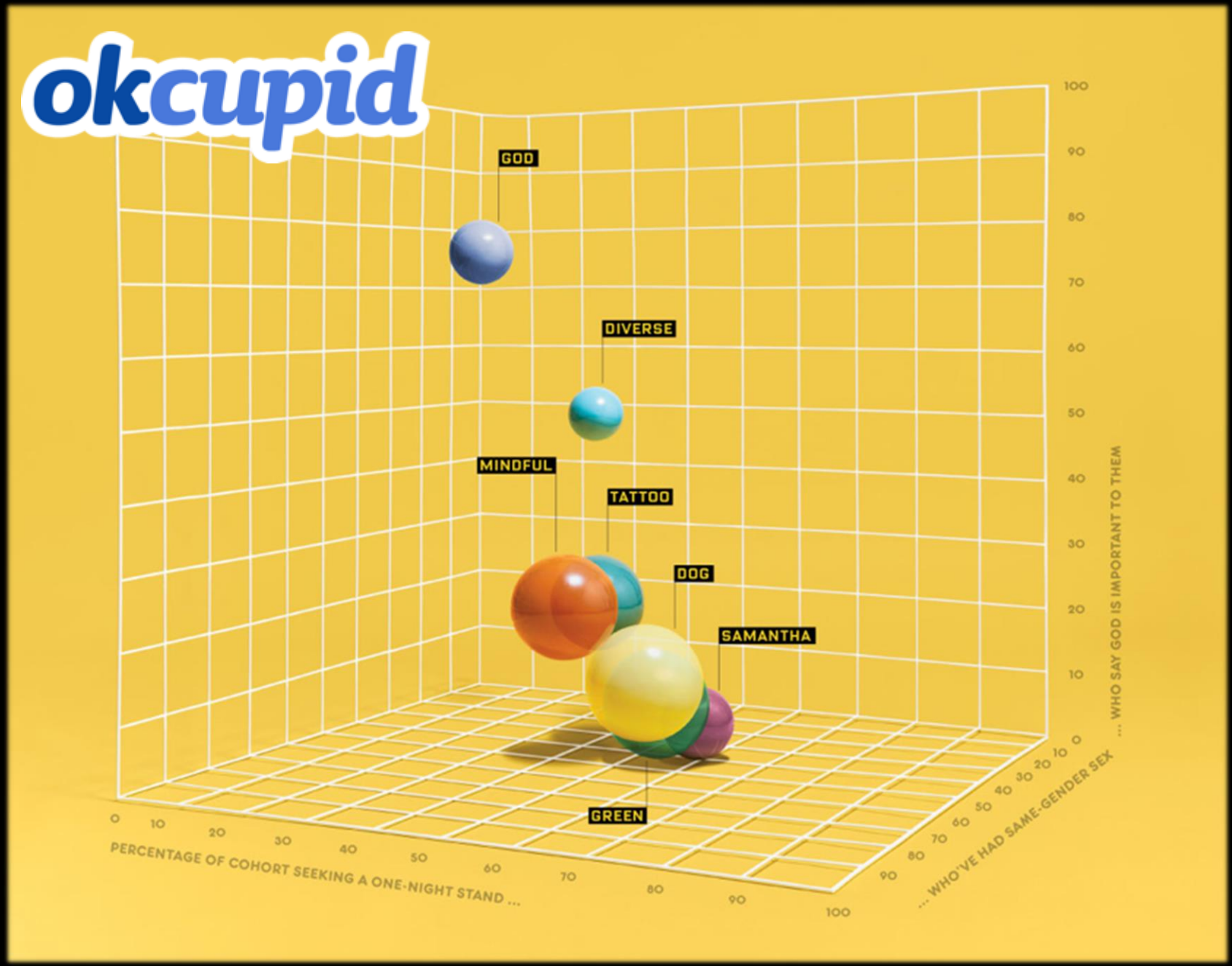
Alzheimer's disease

Clustering

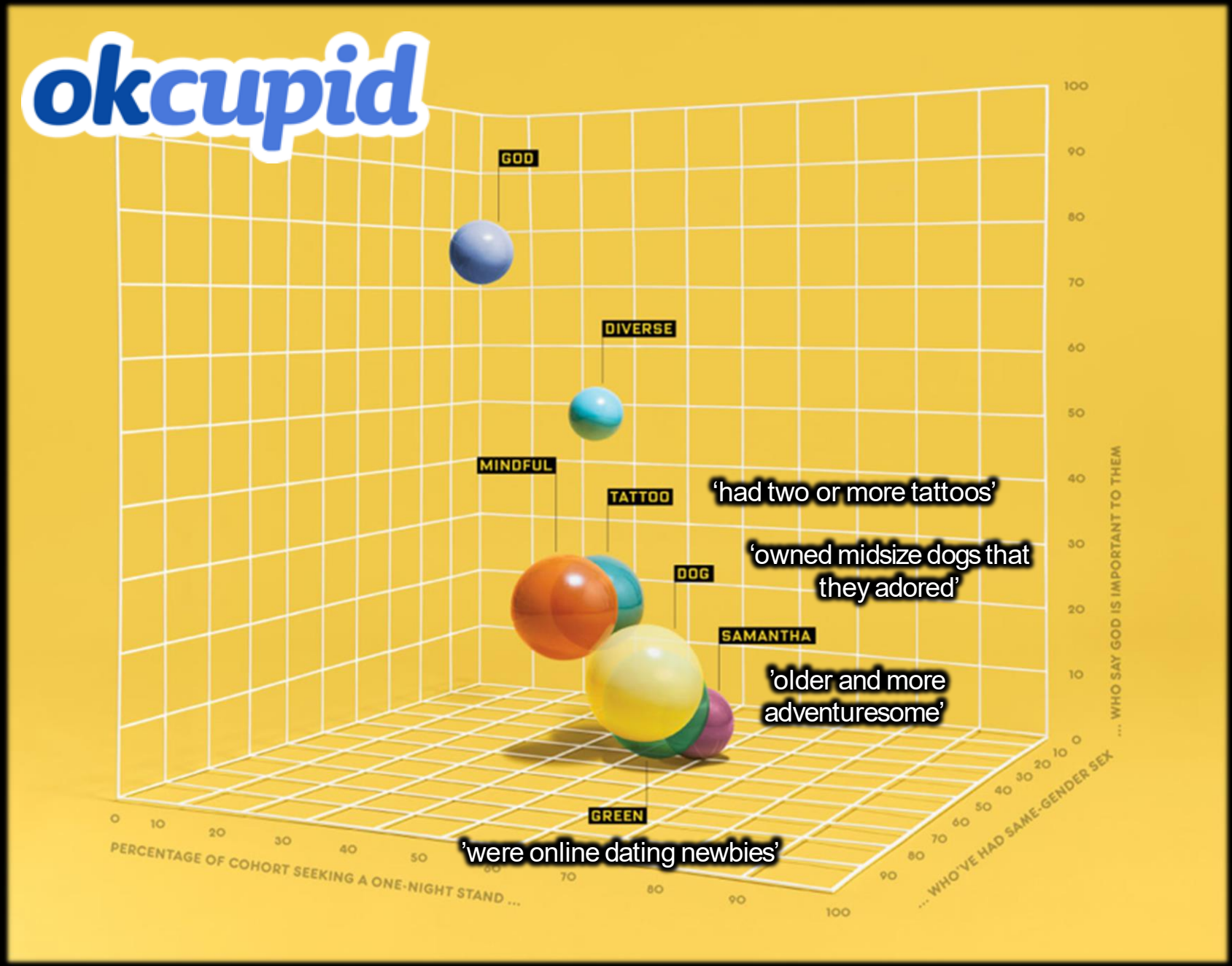
Look at your data.

Can it reasonably be described as representing one homogenous group, described with a single pattern…

… or are there hidden subgroups, each with different patterns?



https://www.wired.com/2014/01/how-to-hack-okcupid/



https://www.wired.com/2014/01/how-to-hack-okcupid/

And after only …

And after only 87 dates…

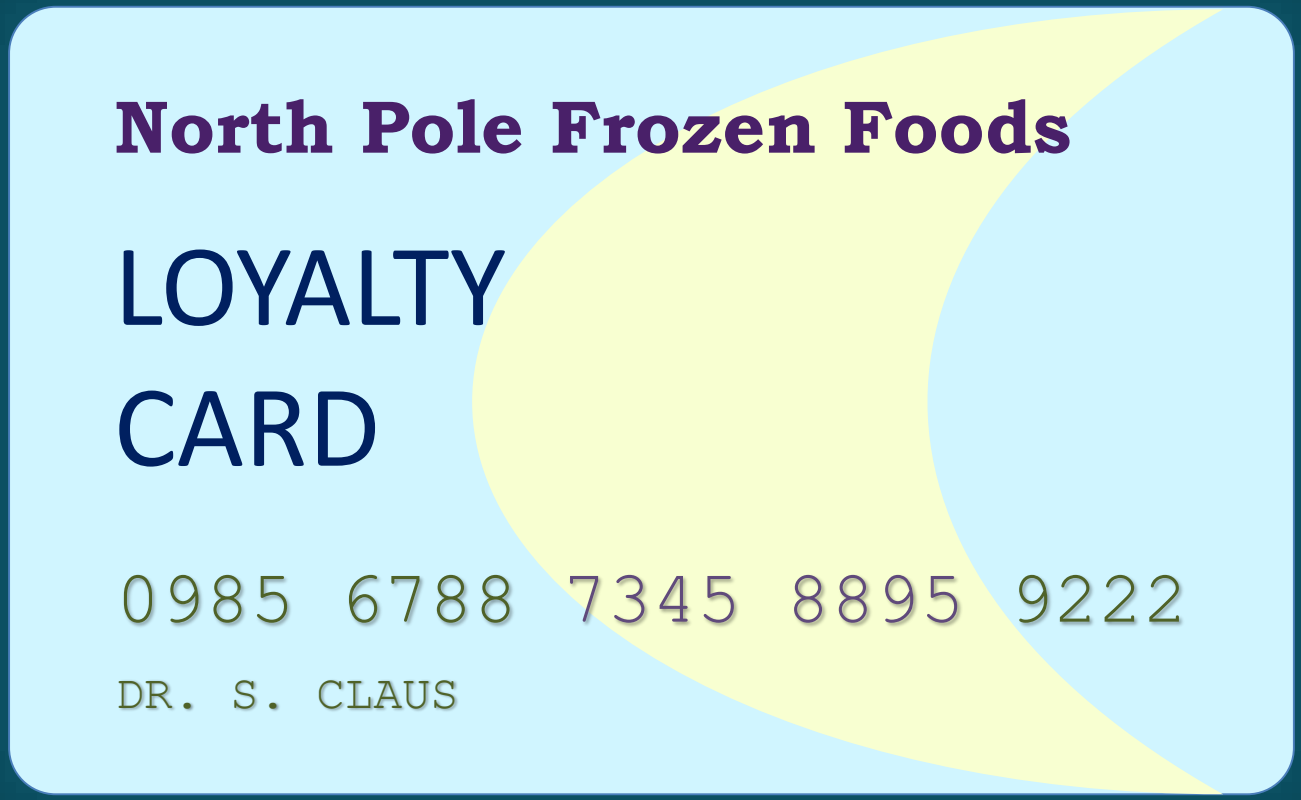
And after only 87 dates . .

http://a.abcnews.com/images/Lifestyle/HT_chris_mckinlay_jtm_140122_16x9_608.jp

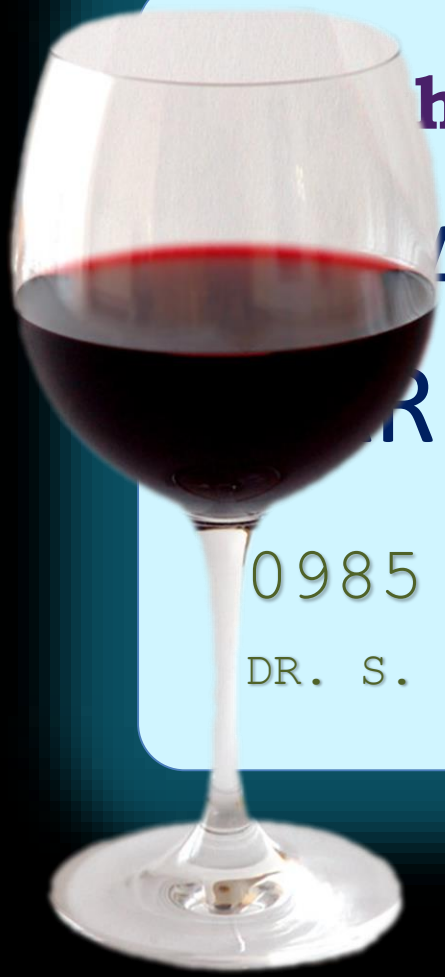
It’s Tea Time

- Part 1: Hypothesis testing recap
- Part 2: Cluster analysis – why should I care?
- Part 3: K means
- Part 4:Before you start
- Part 5: Hierarchical clustering
- Part 6: How good are your clusters?
- Part 7: Some tips and tricks for your written work

Shopping Example



Shopping Example



h Pole Frozen Foods

ALTY

RD

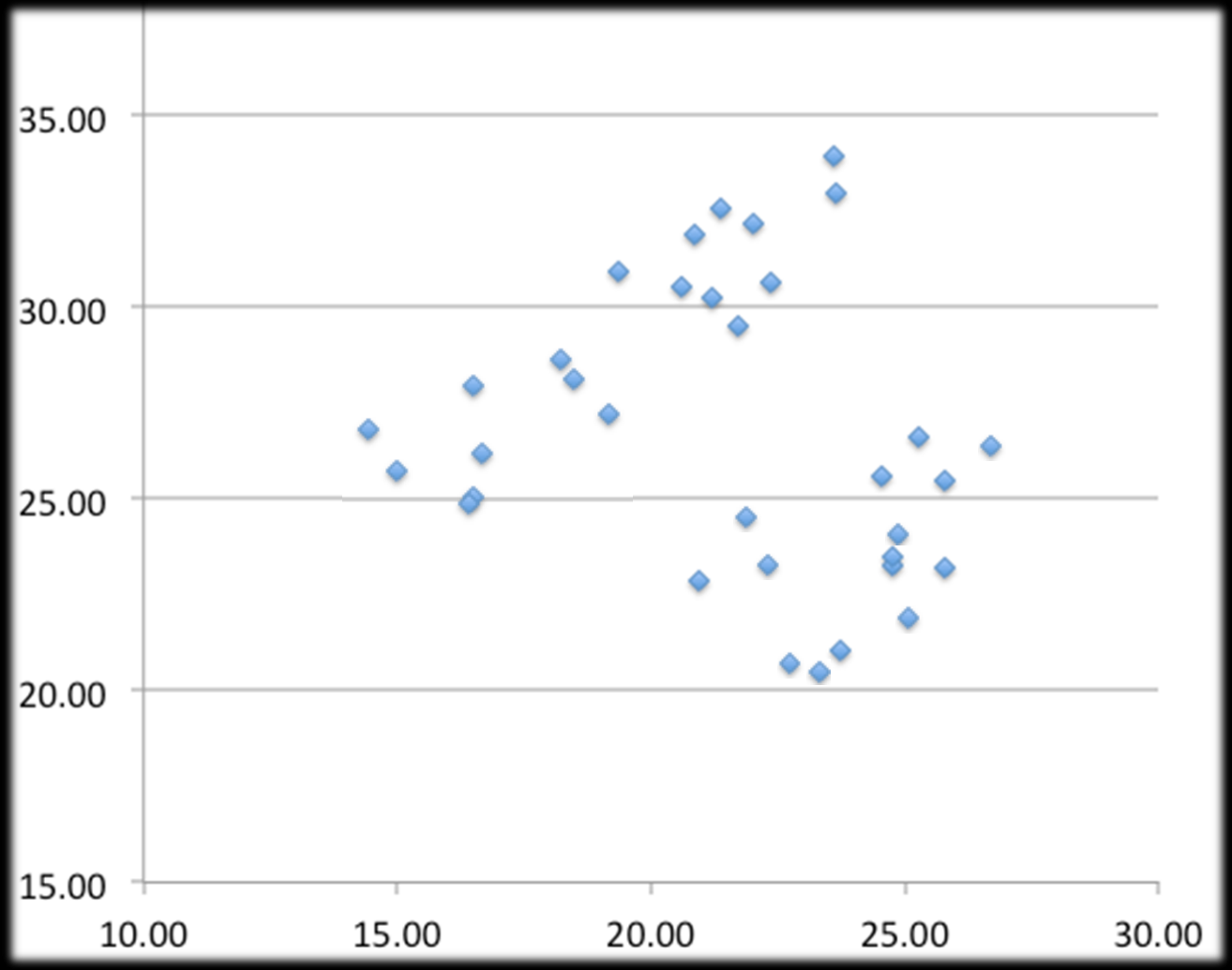
0985 6788 73

DR. S. CLAUS

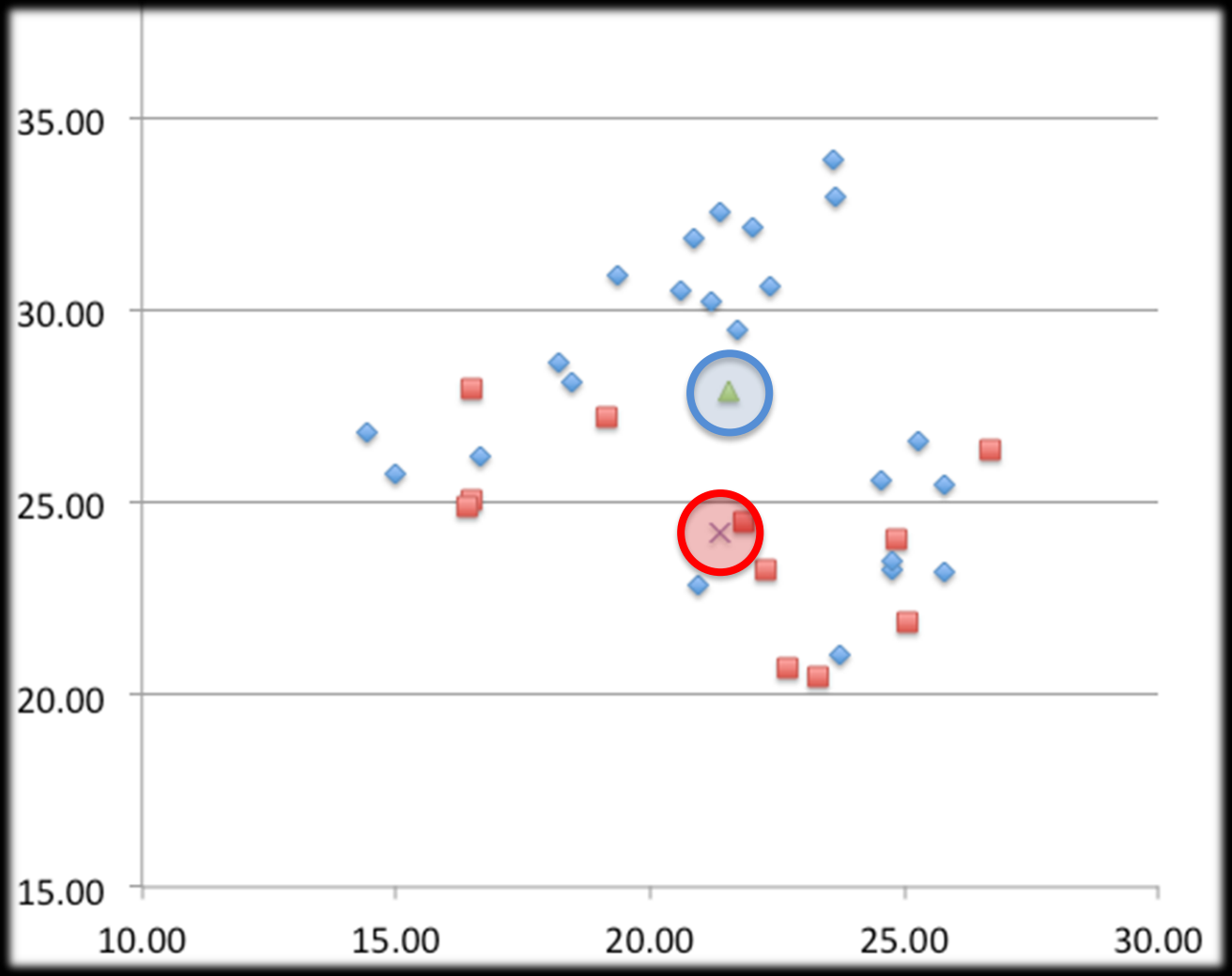
https://commons.wikimedia.org/wiki/File:Red_Wine_Glass.jpg

https://commons.wikimedia.org/wiki/File:Meal_prep_food_container_(43991880552).jpg

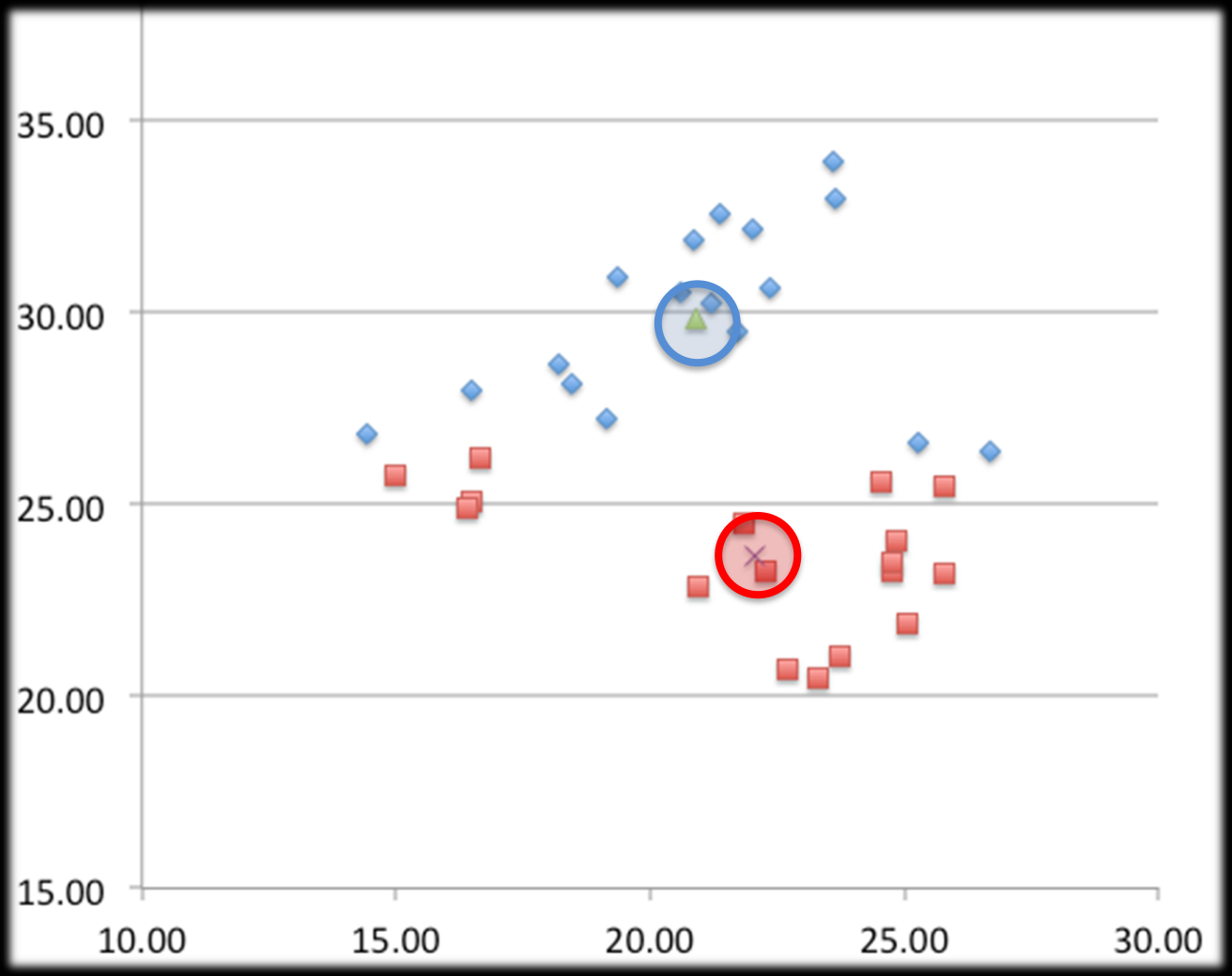
K-means Clustering



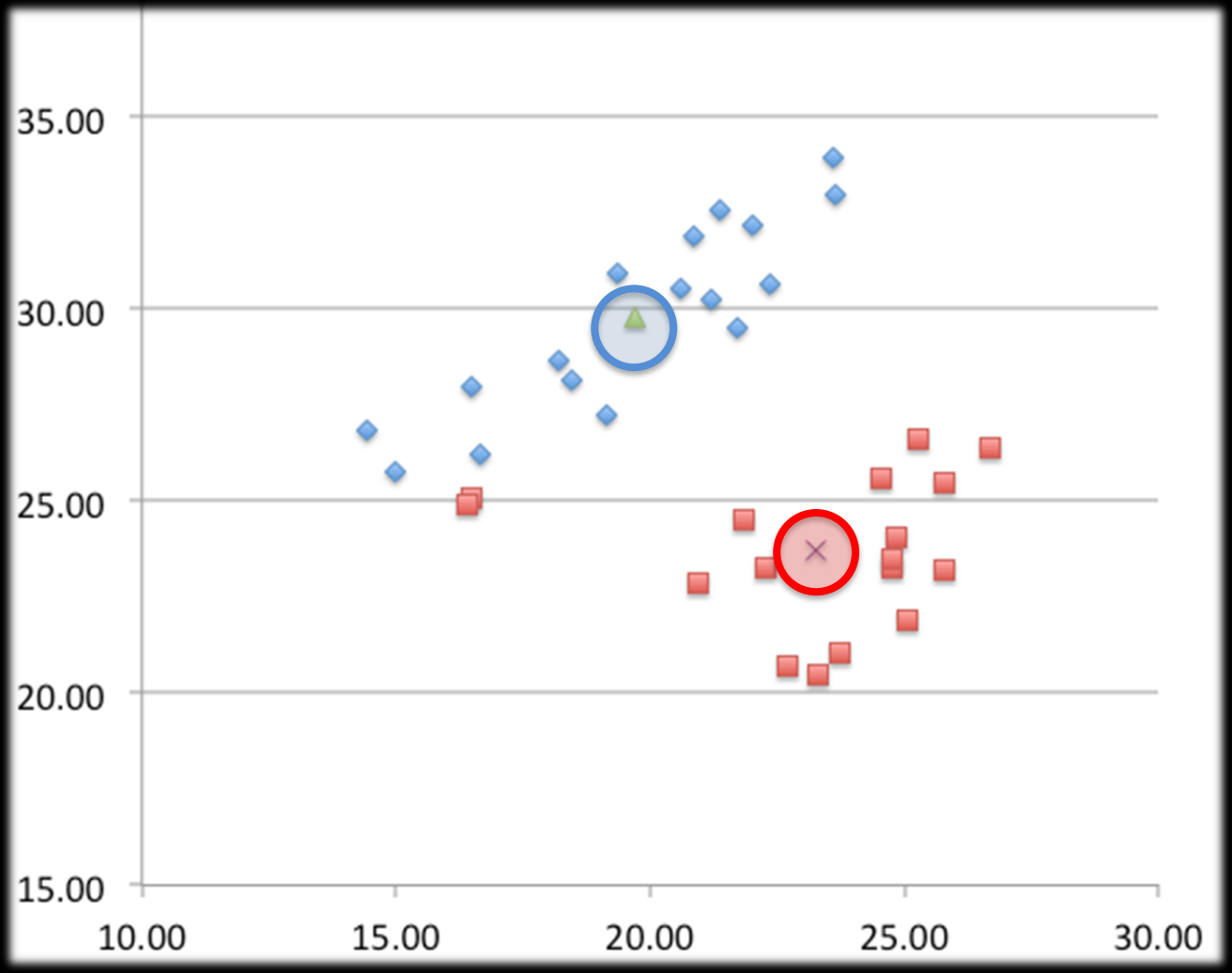
K-means Clustering



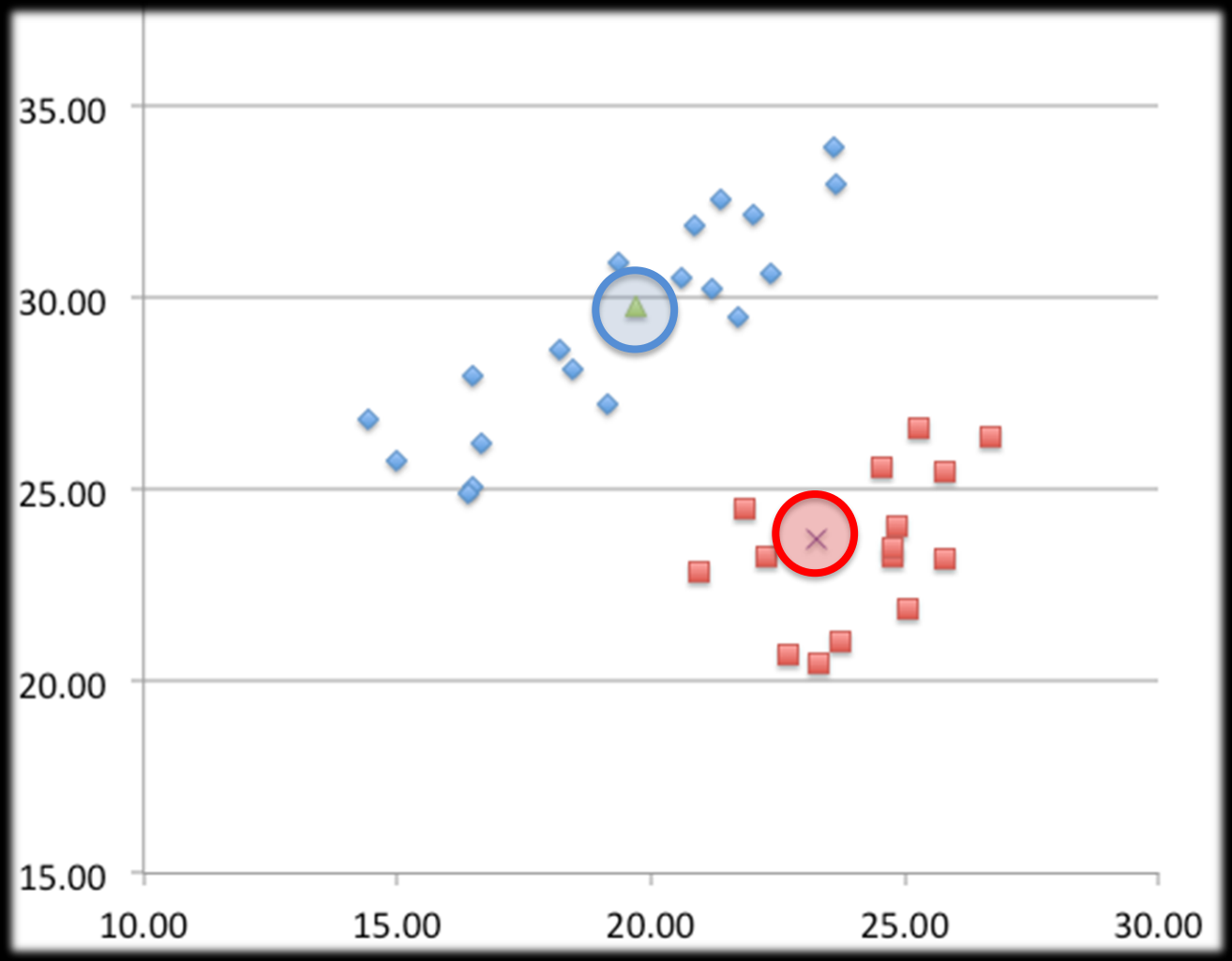
K-means Clustering



K-means Clustering



K-means Clustering



k-Means Clustering

1. Randomly assign your points to k clusters.
2. Calculate the centroid (average point) of each cluster.
3. Reassign all points to the cluster of the nearest centroid.
4. Repeat steps 2-3 until there are no changes.

k-Means Clustering

1. Randomly assign your points to k clusters.
2. Calculate the centroid (average point) of each cluster.
3. Reassign all points to the cluster of the nearest centroid.
4. Repeat steps 2-3 until there are no changes.

Note that “distance” here doesn’t need to be physical. In this example, it’s money, but it can be anything quantitative – think OKCupid example

k-Means Disadvantages

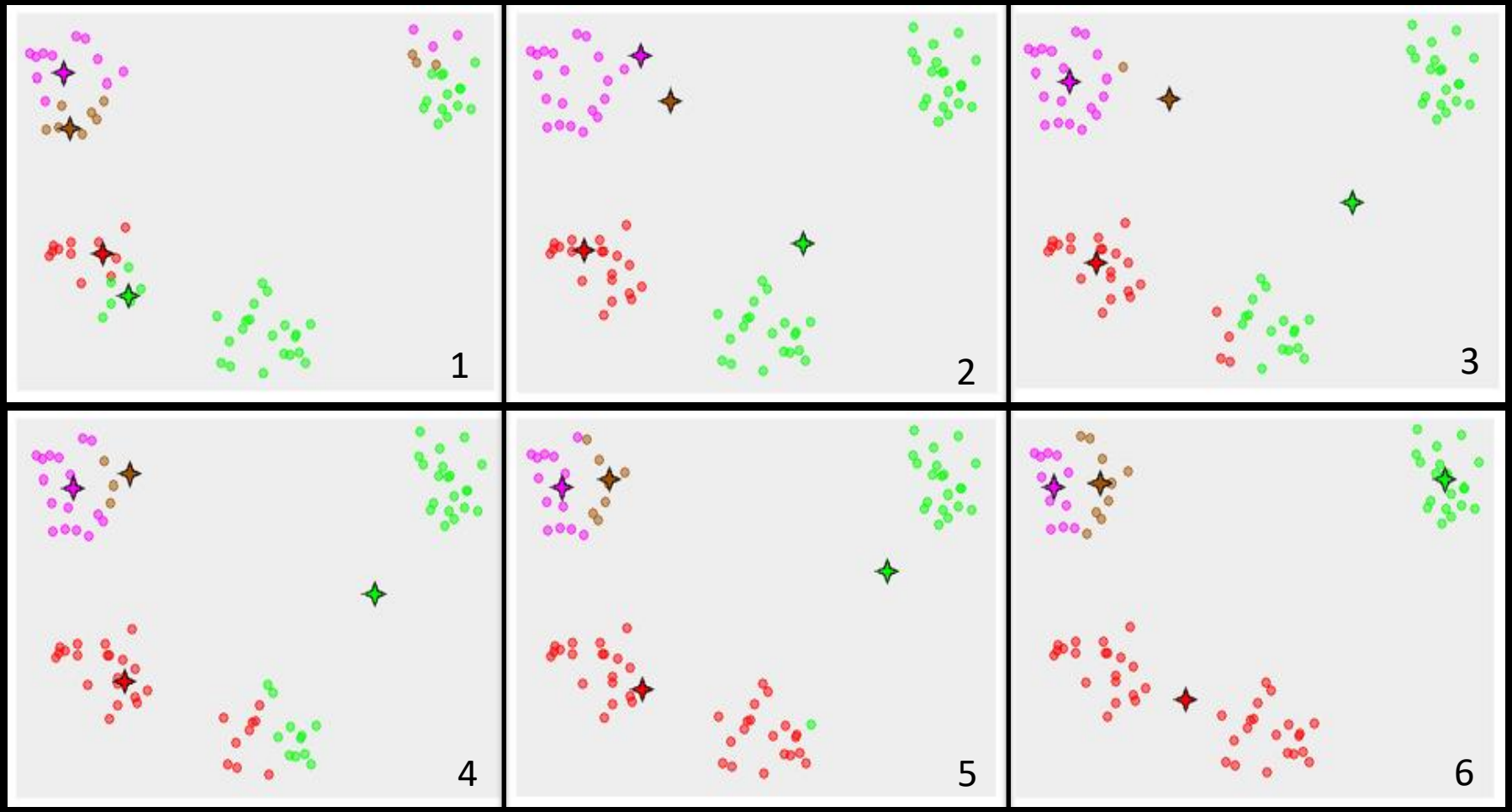
1. Requires knowledge of the number of clusters;

2. Sensitive to initialisation, which can lead to poor solutions;

3. Sensitive to outliers, which can results in inaccurate clusters;

k-Means Disadvantages

2. Sensitive to initialisation, which can lead to poor solutions;



k-Means Disadvantages

1. Requires knowledge of the number of clusters;

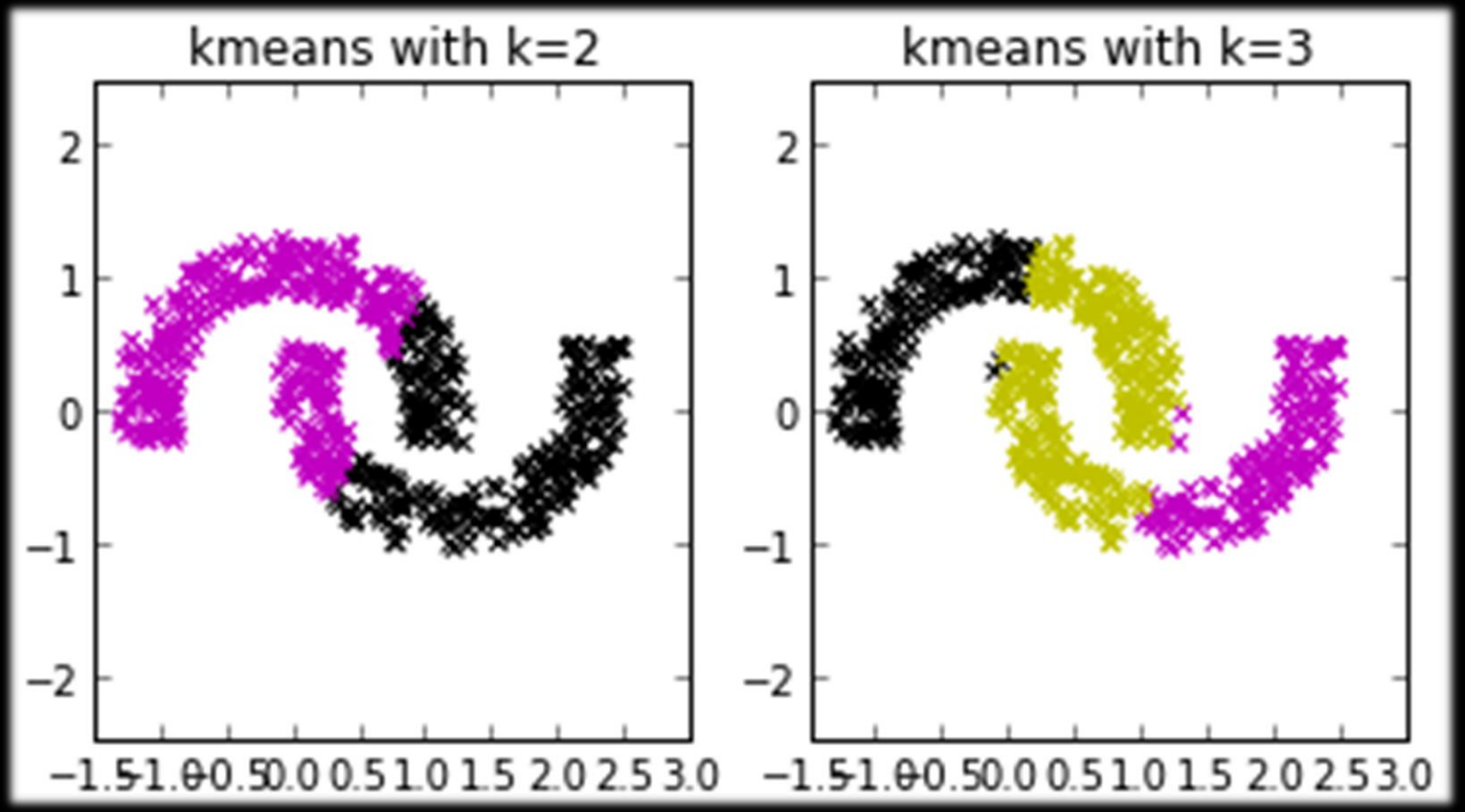
2. Sensitive to initialisation, which can lead to poor solutions;

3. Sensitive to outliers, which can results in inaccurate clusters;

4. Incapable of handling clusters of a non-convex shape;

k-Means Disadvantages

4. Incapable of handling clusters of a non-convex shape;



k-Means Disadvantages

1. Requires knowledge of the number of clusters;
2. Sensitive to initialisation, which can lead to poor solutions;
3. Sensitive to outliers, which can results in inaccurate clusters;
4. Incapable of handling clusters of a non-convex shape;
5. Inapplicable to categorical data.

It’s Tea Time

- Part 1: Hypothesis testing recap
- Part 2: Cluster analysis – why should I care?
- Part 3: K means
- Part 4:Before you start
- Part 5: Hierarchical clustering
- Part 6: How good are your clusters?
- Part 7: Some tips and tricks for your written work

IMPORTANT FIRST STEP: Standardisation

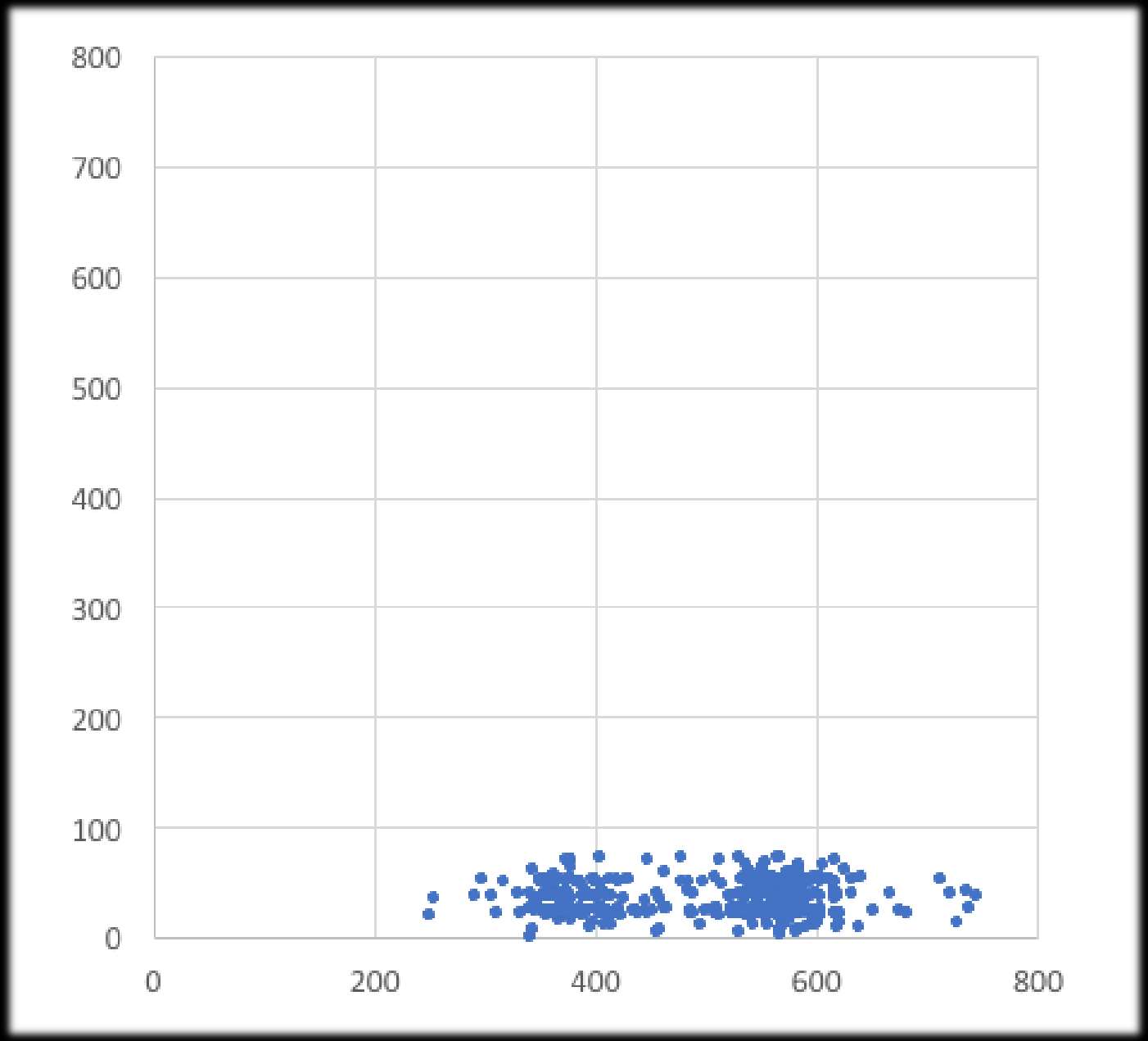
Comparing Variables with Different Units

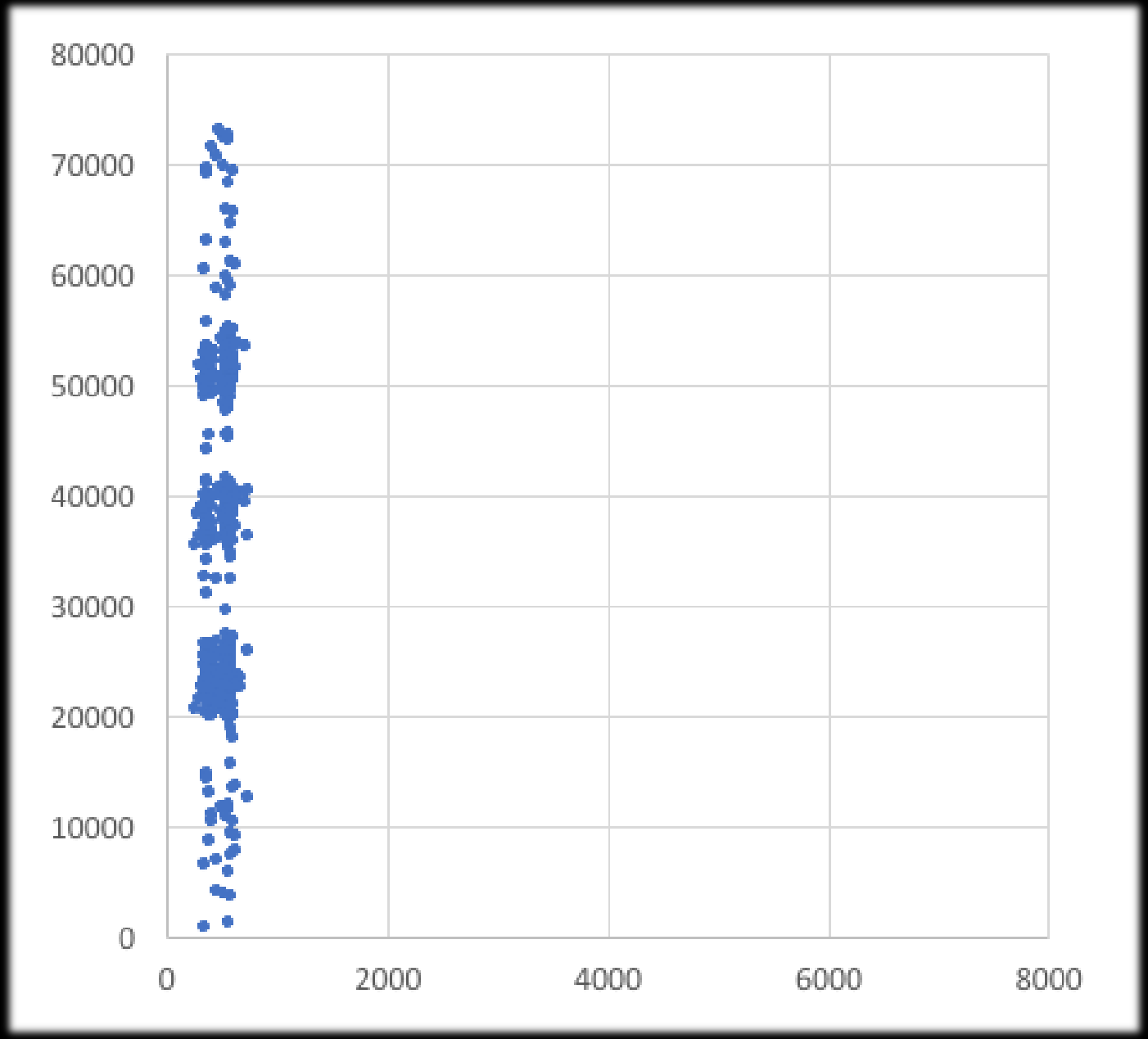
Putting everything on the same scale

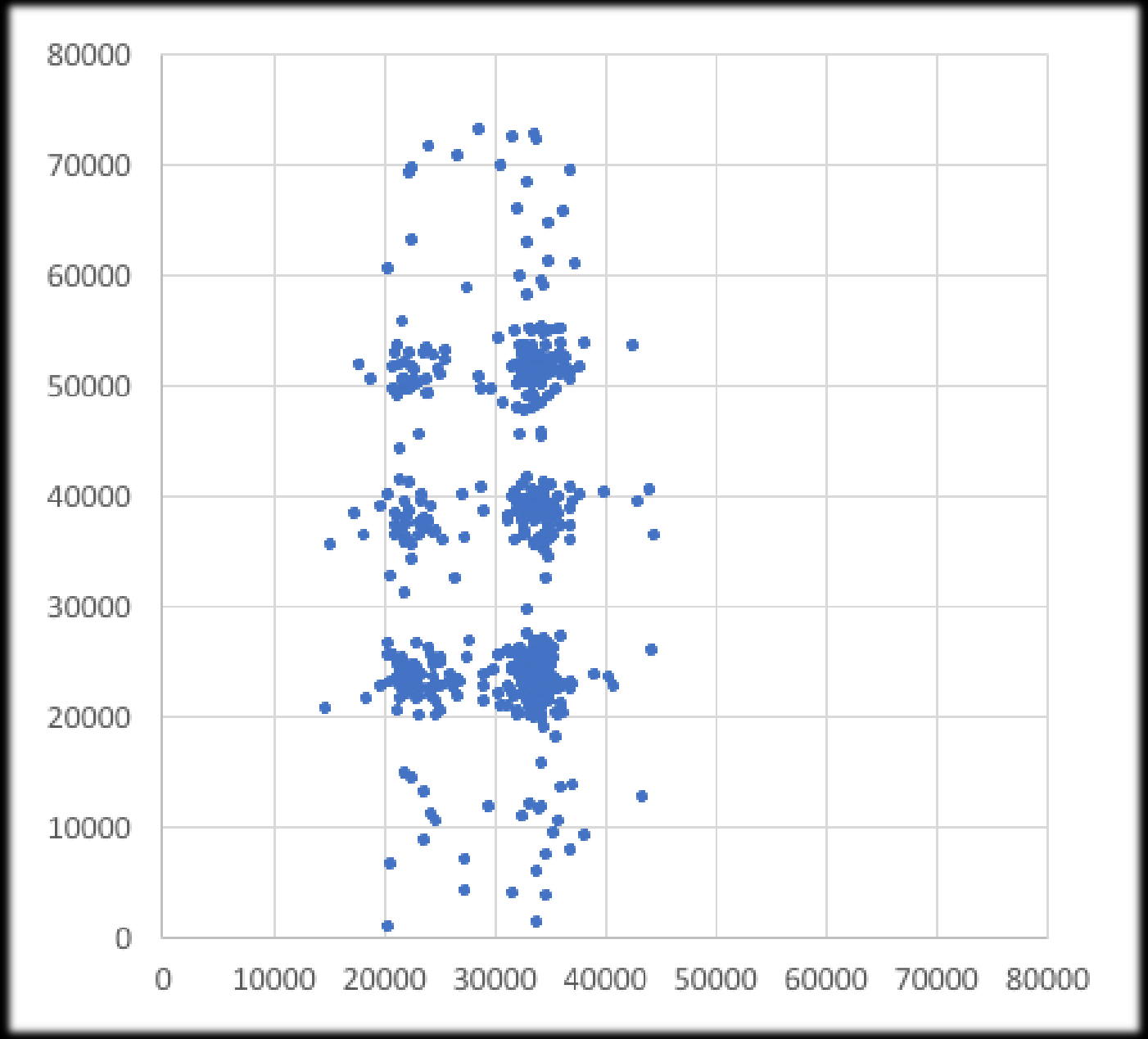
http://www.imdb.com/title/tt2126355/

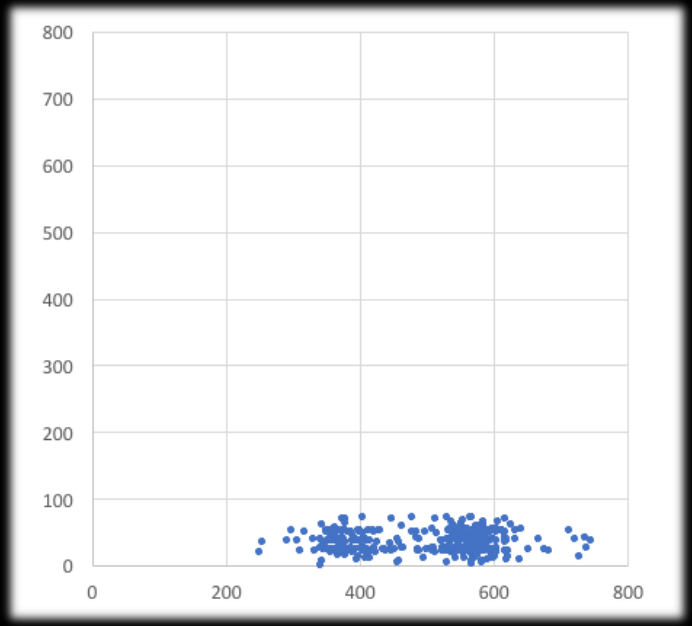
tremor_data.csv

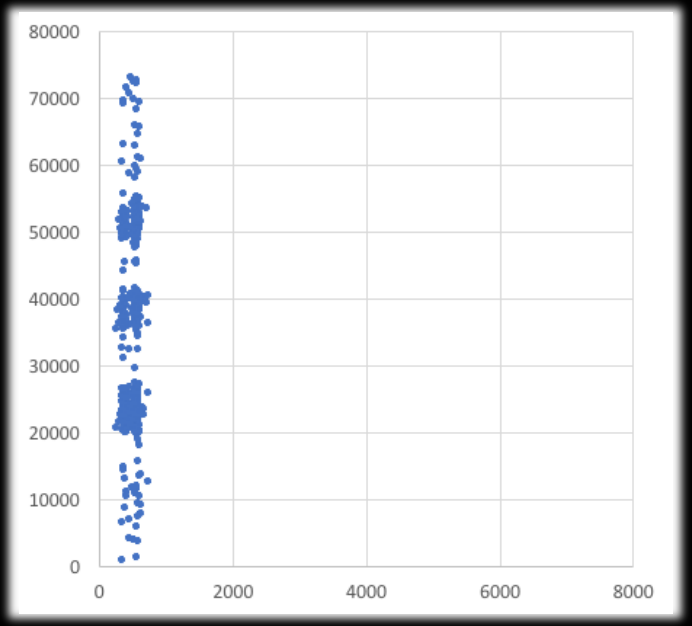
	A	B	C	D
1	Time (mins)	Distance (m)	Size (Moment Magnitude)	
2	14963	20518	3	
3	15229	35321	1.8	
4	17442	38120	1.9	
5	17902	51661	2	
6	18393	36197	1.7	
7	18688	21457	2.4	
8	18985	50482	1.5	
9	19827	38701	2.5	
10	20000	22431	2	
11	20403	40080	2.4	
12	20465	873	2.5	
13	20507	26386	2.3	
14	20595	60527	2	
15	20619	25368	2.7	
16	20679	6520	2.9	
17	20698	23061	2.5	

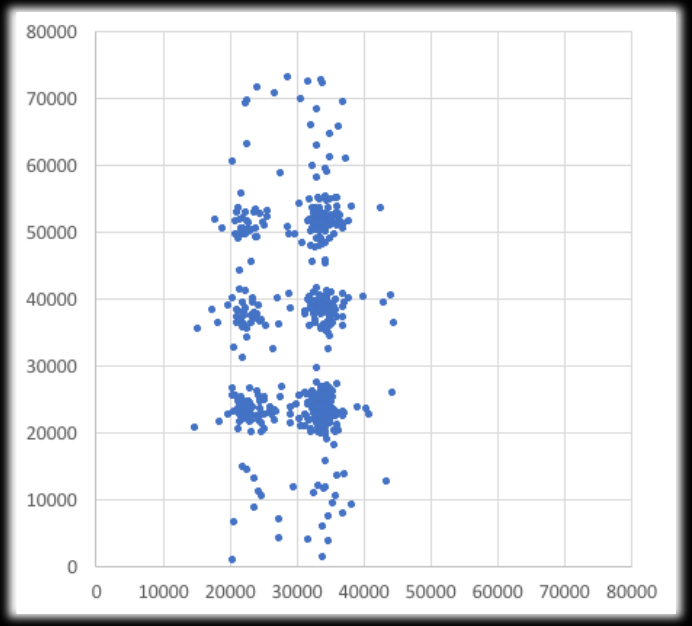

Distance along fault line (km)
800
700
600
500
400
300
200
100
0
0
200
400
600
800
Time (Hours after start of study)


80000
70000
60000
50000
40000
30000
20000
10000
0
0
2000
4000
6000
8000
Distance along fault line (m)
Time (Hours after start of study)


Distance along fault line (m)
80000
70000
60000
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
60000
70000
80000
Time (Minutes after start of study)







http://www.imdb.com/title/tt2126355/

IMPORTANT FIRST STEP: Standardisation

Comparing Variables with Different Units

Putting everything on the same scale

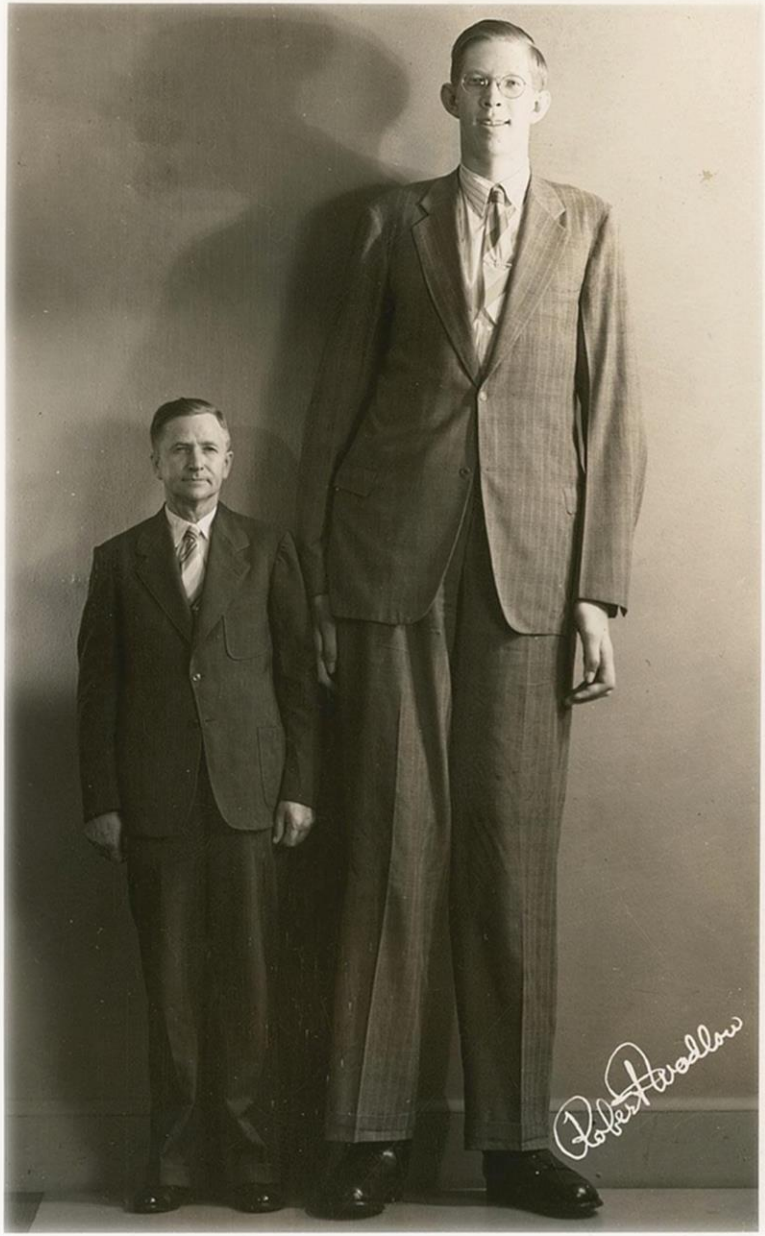
Standardisation

Clustering generally requires that variables are standardised, because:

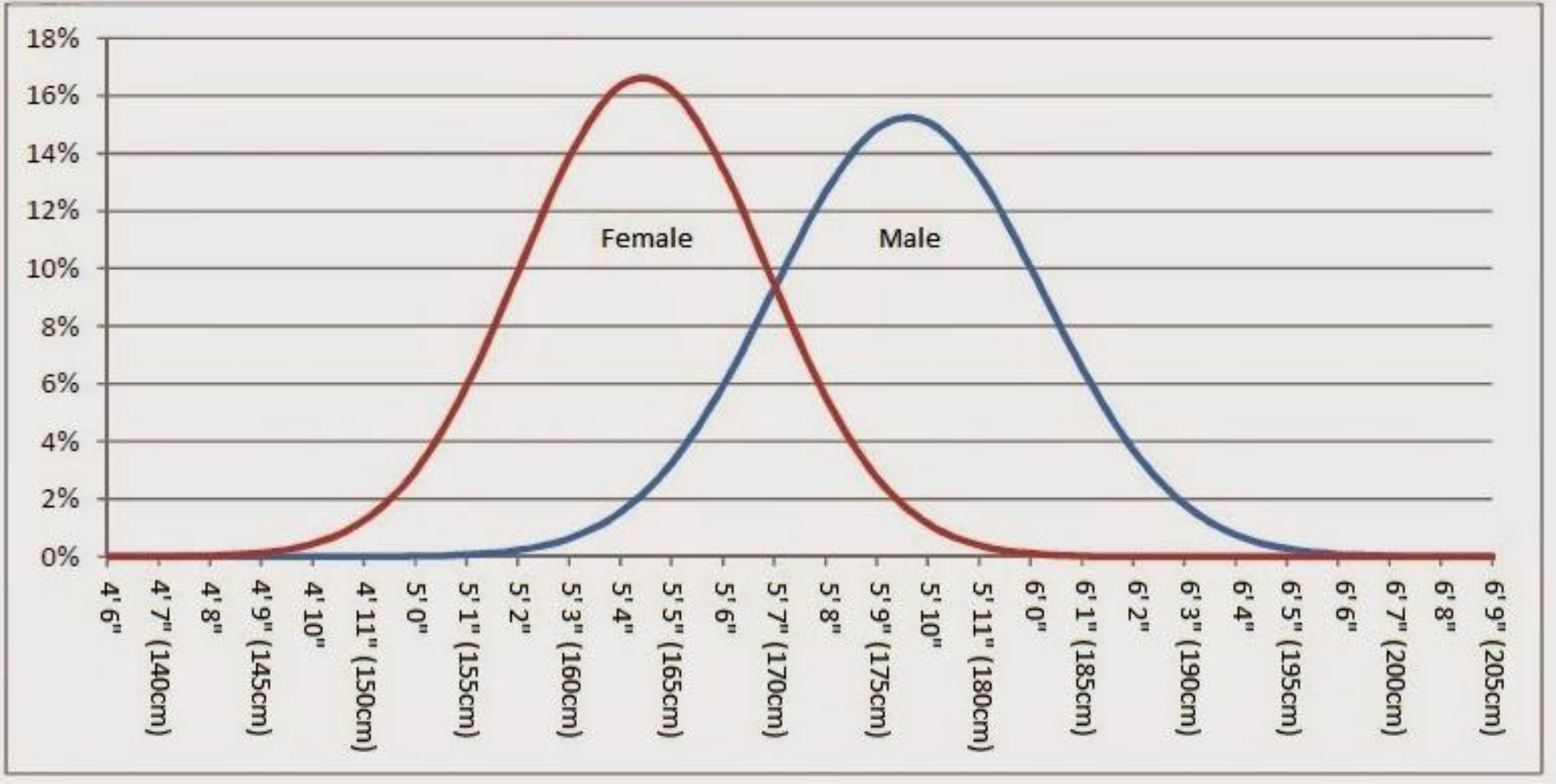
- Data in different units cannot be meaningfully compared without scaling.
- How you scale your data can be very important.
- Without standardisation, clustering may be dominated by the variable with the greatest range.

EXAMPLE

How to rescale if your data is normally distributed


Robert Wadlow


Jyoti Amge



Female Heights

Mean: 163 cm

Standard deviation: 6 cm.

Jyoti Amge: 63 cm

Male Heights

Mean: 178 cm

Standard Deviation: 7 cm.

Robert Wadlow: 272 cm

Female Heights

Mean: 163 cm

Standard deviation: 6 cm.

Jyoti Amge: 63 cm

Male Heights

Mean: 178 cm

Standard Deviation: 7 cm.

Robert Wadlow: 272 cm

Whose height is the greater outlier?

Female Heights

Mean: 163 cm

Standard deviation: 6 cm.

Jyoti Amge: 63 cm

Male Heights

Mean: 178 cm

Standard Deviation: 7 cm.

Robert Wadlow: 272 cm

Whose height is the greater outlier?

(63 – 163) / 6 = -16.7

16.7 standard deviations

Below the mean

(272 – 178) / 7 = 13.4

13.4 standard deviations

above the mean

Standardisation

For roughly normally distributed data
(i.e. not highly skewed)

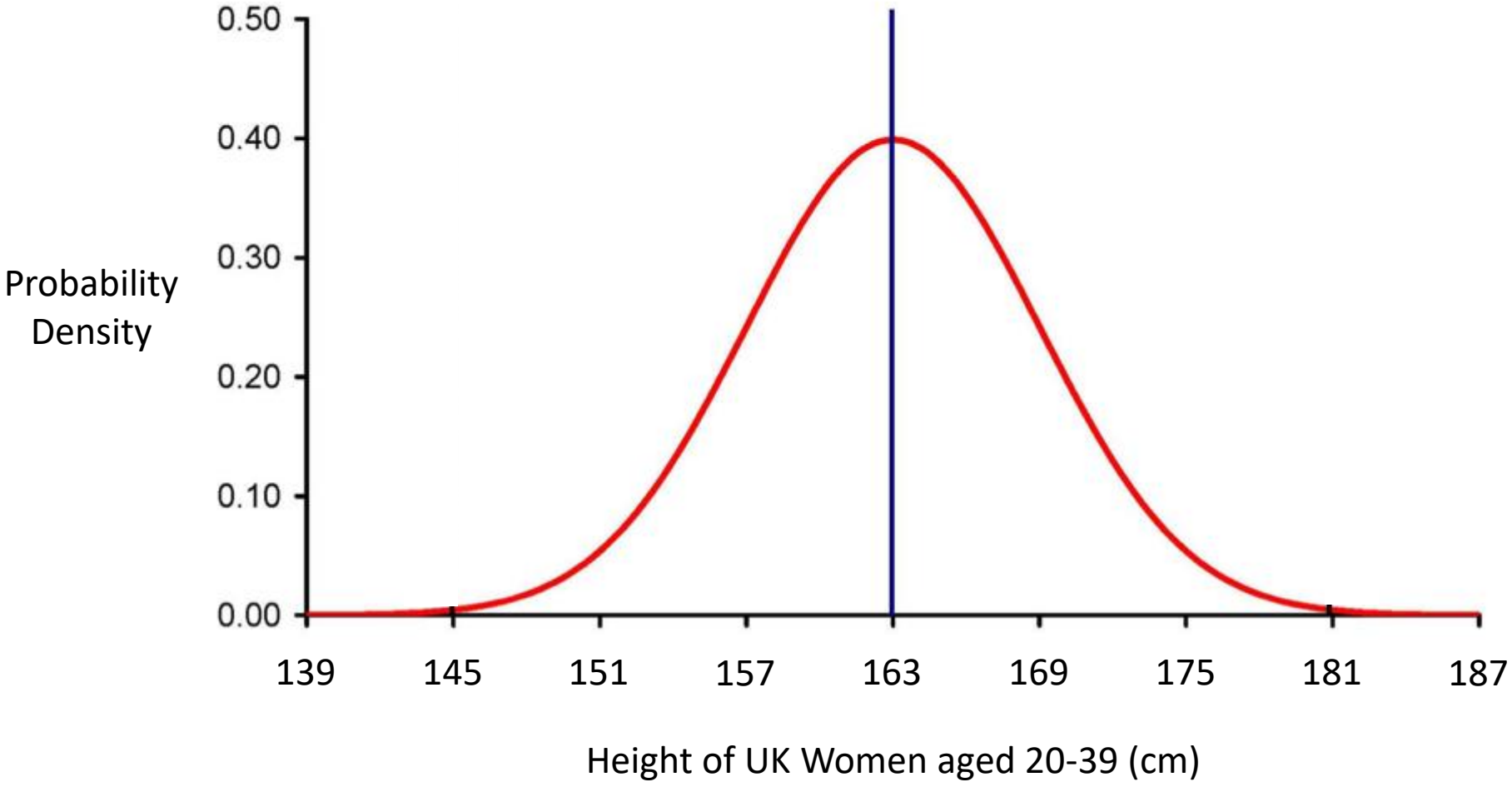
Data Mean

Standardised Data

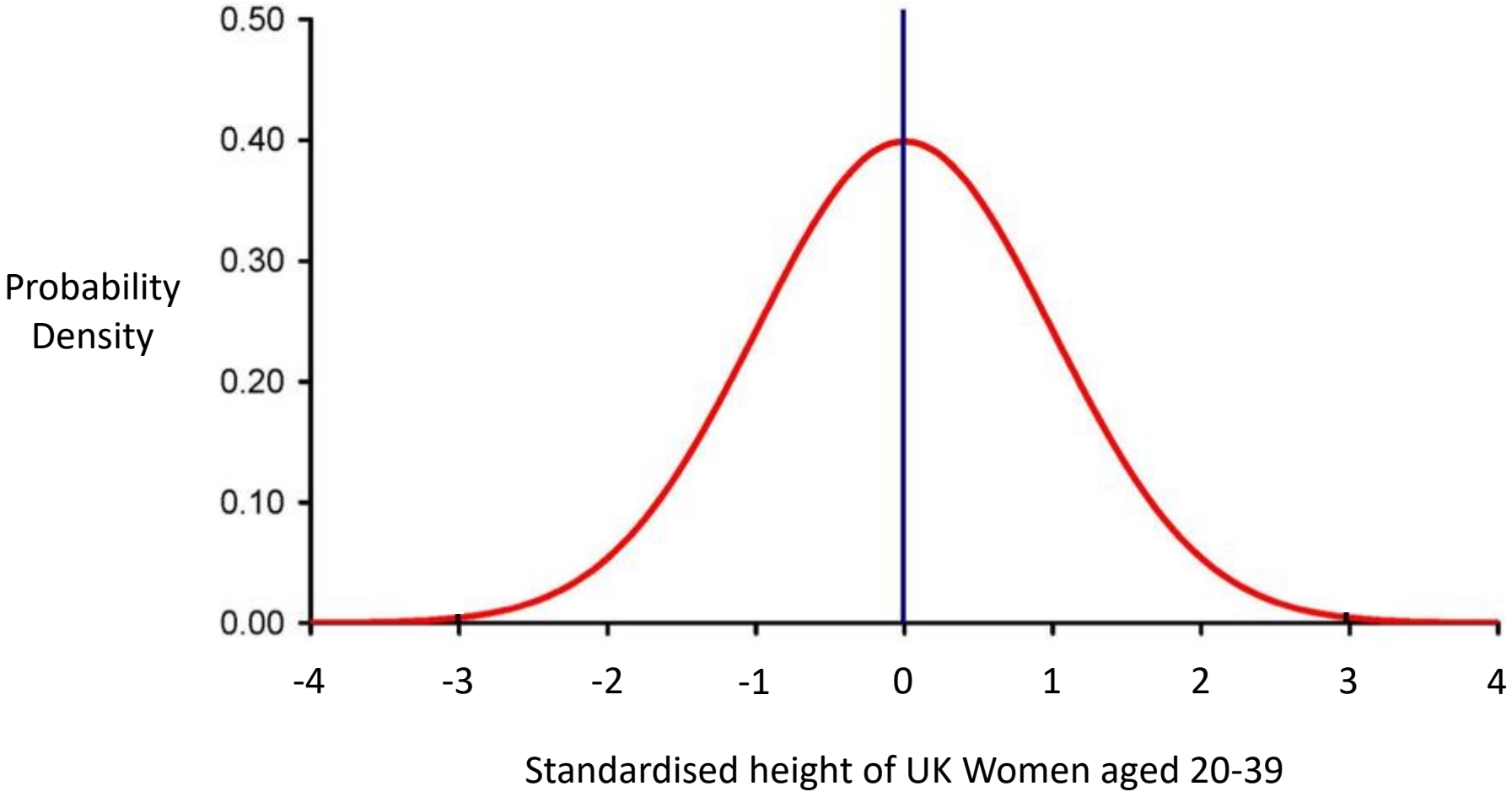
$$z = \frac{x - \mu}{\sigma}$$

Standard Deviation

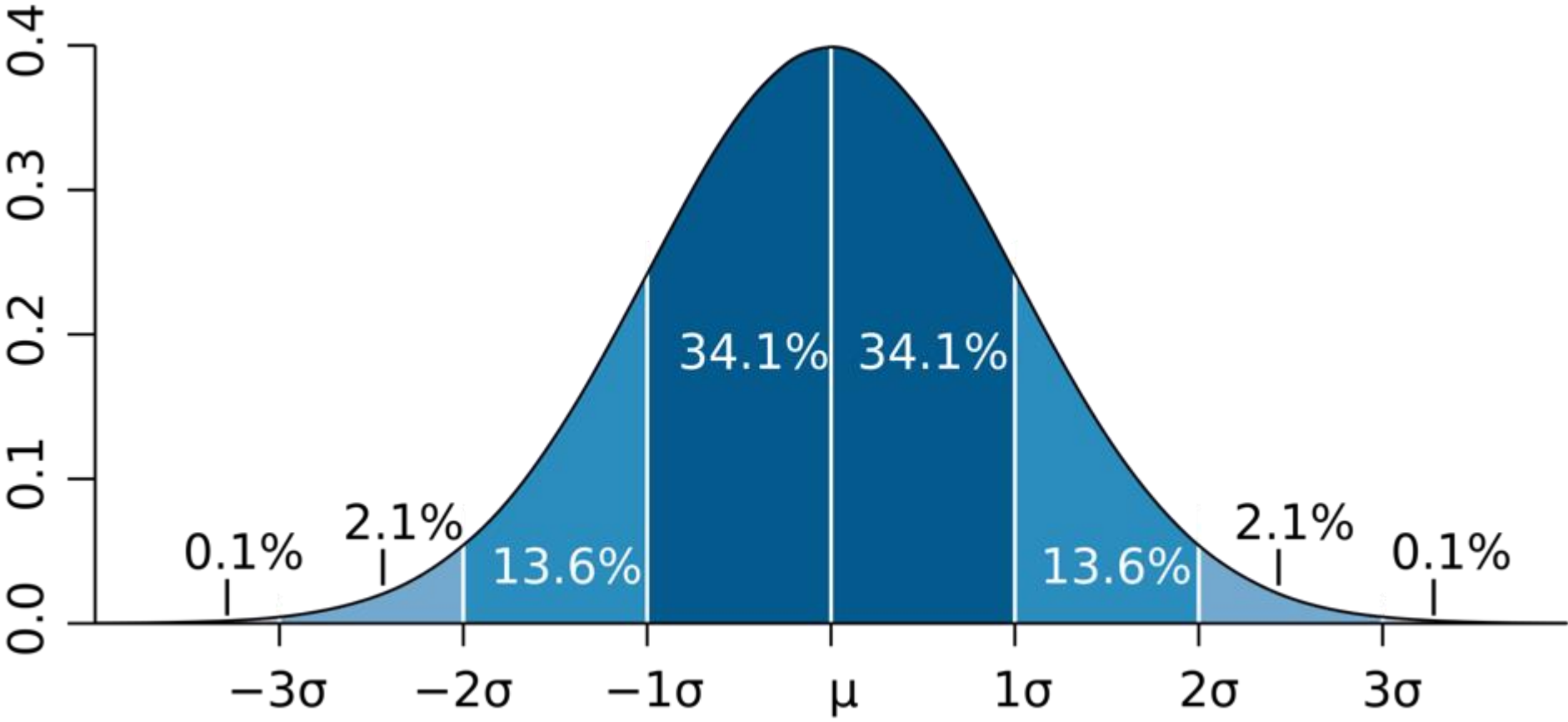
Standardisation



Standardisation



Normal Distribution



What if your data isn’t normally distributed

Min-Max Rescaling

For highly non-normal data (i.e. highly skewed)

Original Data

Rescaled Data

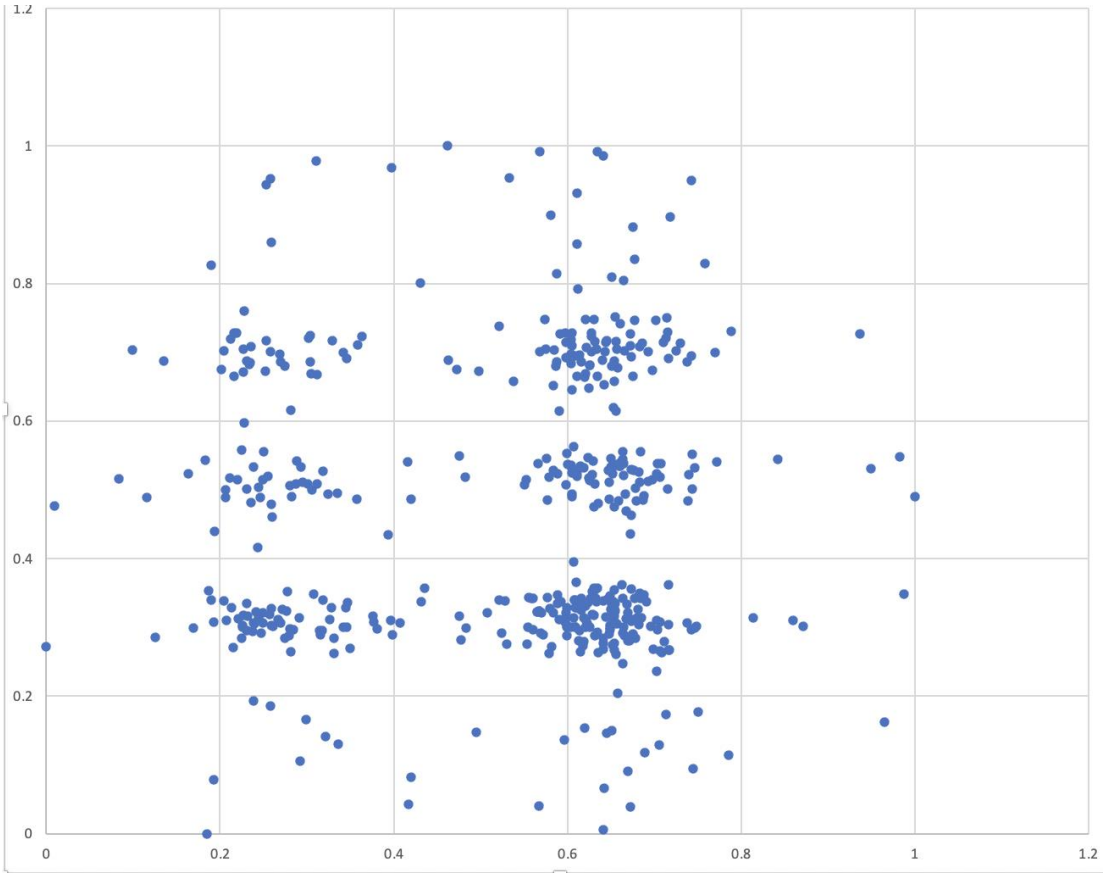
$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Original Data

Rescaled Data

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Distance rescaled between 0 and 1



Time, rescaled between 0 and 1

What if your data has mega outliers?

IDR Standardisation

Non-normal data with significant outliers

Original Data

$$x^{\mathrm{IDR}} = \begin{cases} \dfrac{x - P_{50}}{P_{90} - P_{50}}, & x \geq P_{50} \\ \dfrac{x - P_{50}}{P_{50} - P_{10}}, & x < P_{50} \end{cases}$$

Rescaled Data

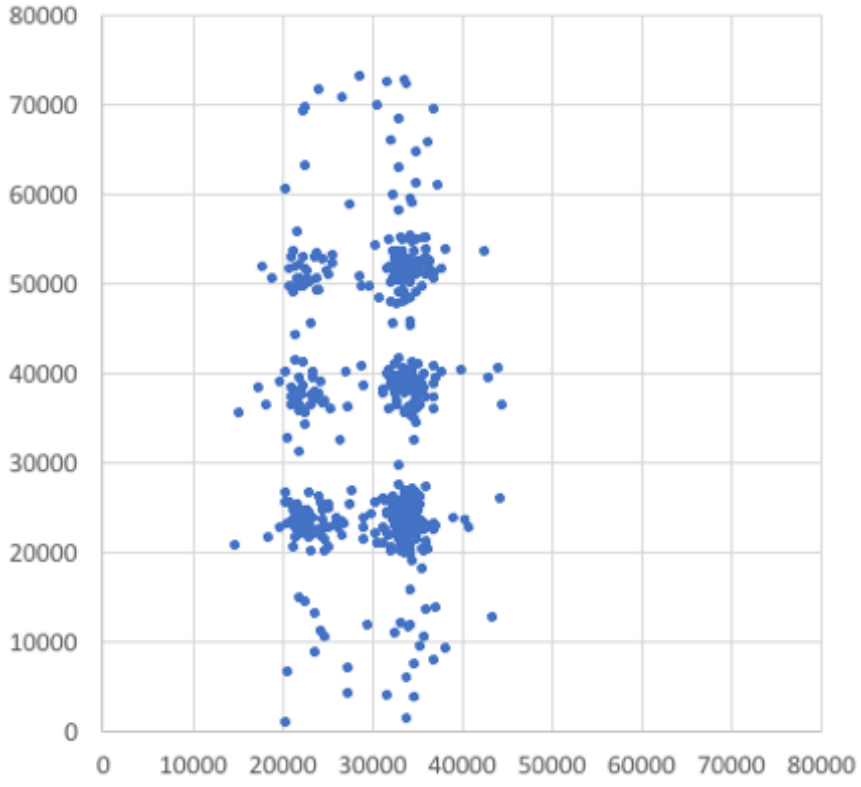
(See e.g. https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/methodologyandvariables)

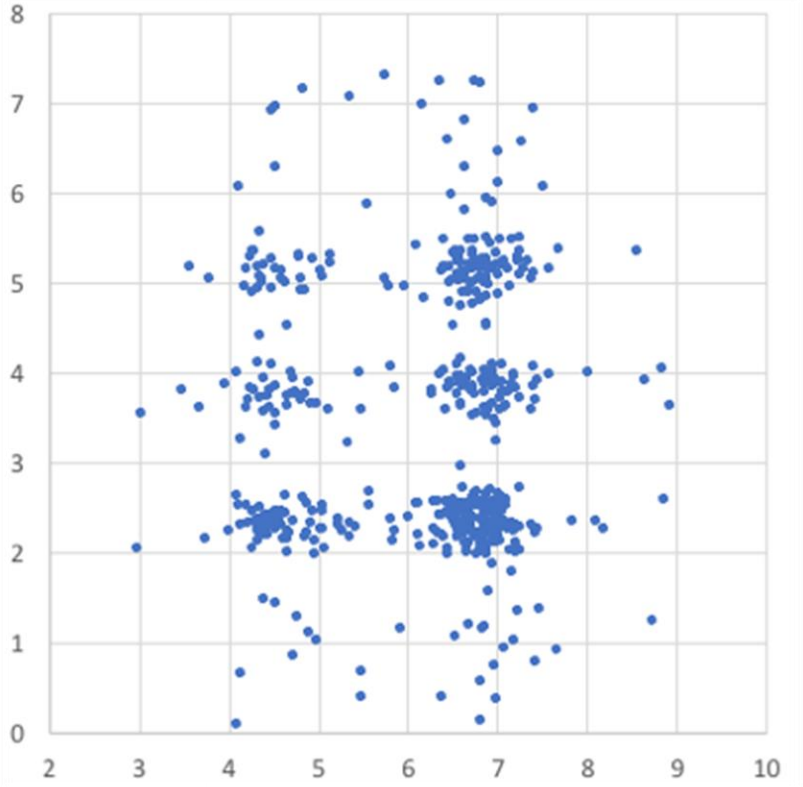
One last option..

Explicit Rescaling

y : Distance along fault line (m)

Y = *y* / 10000





x : Time (Minutes after start of study)

X = 2 *x* / 10000

IMPORTANT FIRST STEP: Standardisation

Get your data on the same scale before you begin clustering

It’s Tea Time

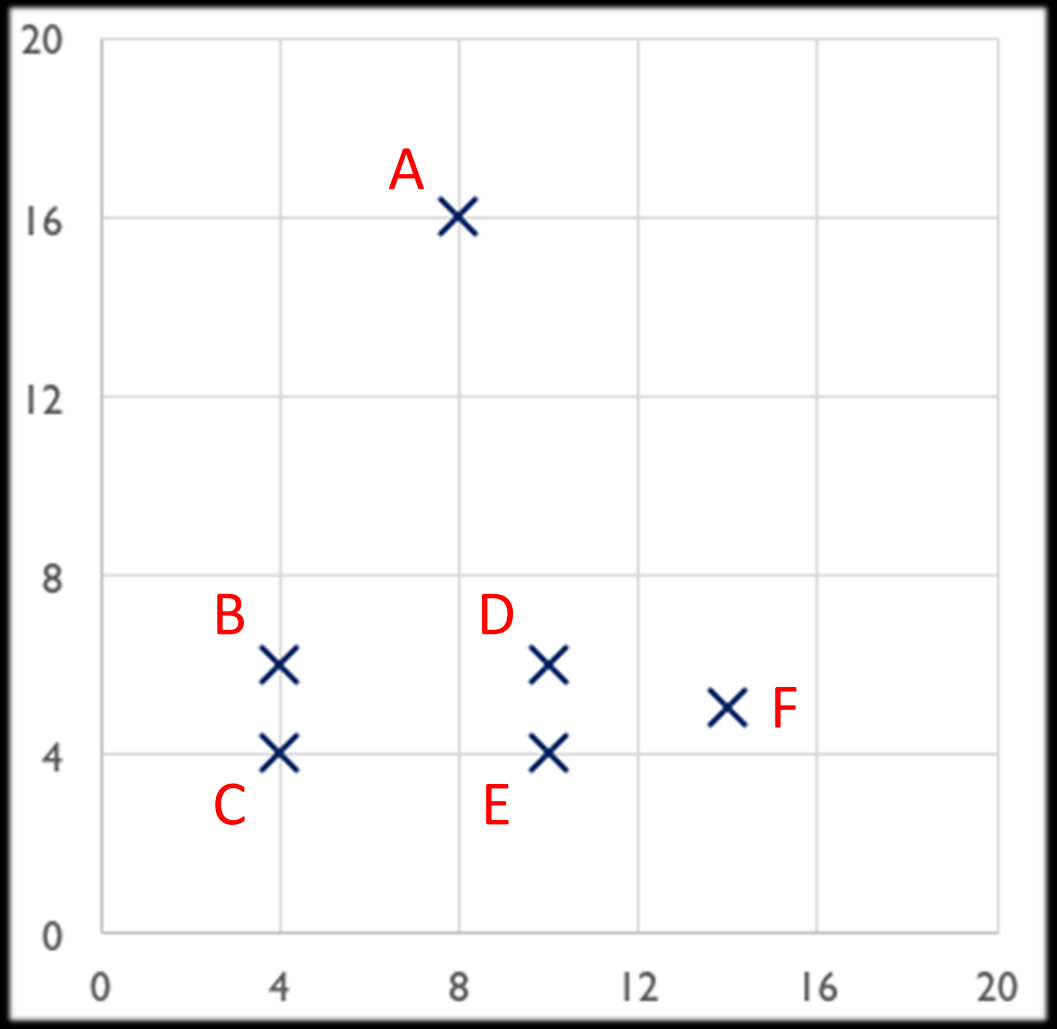
- Part 1: Hypothesis testing recap
- Part 2: Cluster analysis – why should I care?
- Part 3: K means
- Part 4:Before you start
- Part 5: Hierarchical clustering
- Part 6: How good are your clusters?
- Part 7: Some tips and tricks for your written work

k-Means Clustering

1. Randomly assign your points to k clusters.
2. Calculate the centroid (average point) of each cluster.
3. Reassign all points to the cluster of the nearest centroid.
4. Repeat steps 2-3 until there are no changes.

But there are other methods

Hierarchical Clustering



$r = 0$

Clusters (6)

A

B

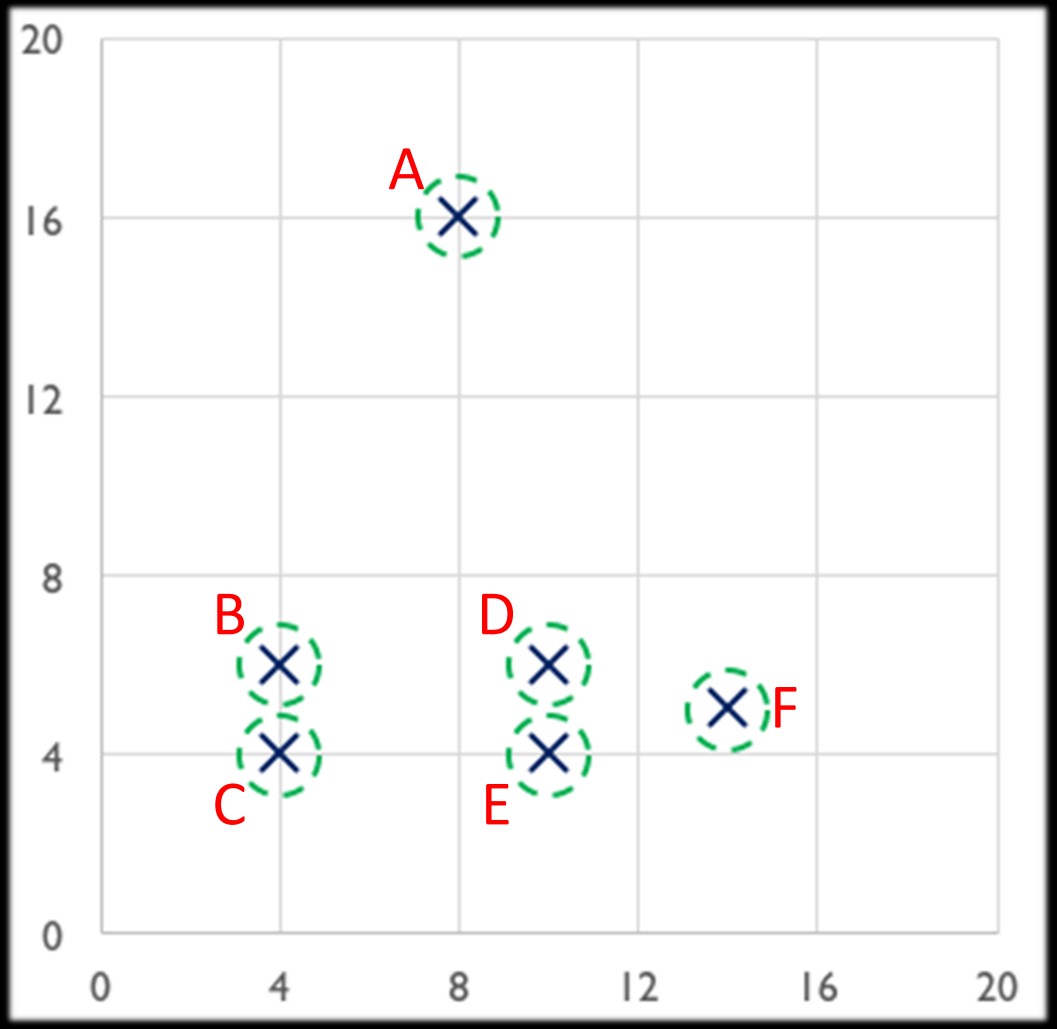
C

D

E

F

Hierarchical Clustering



$r = 0.9$

Clusters (6)

A

B

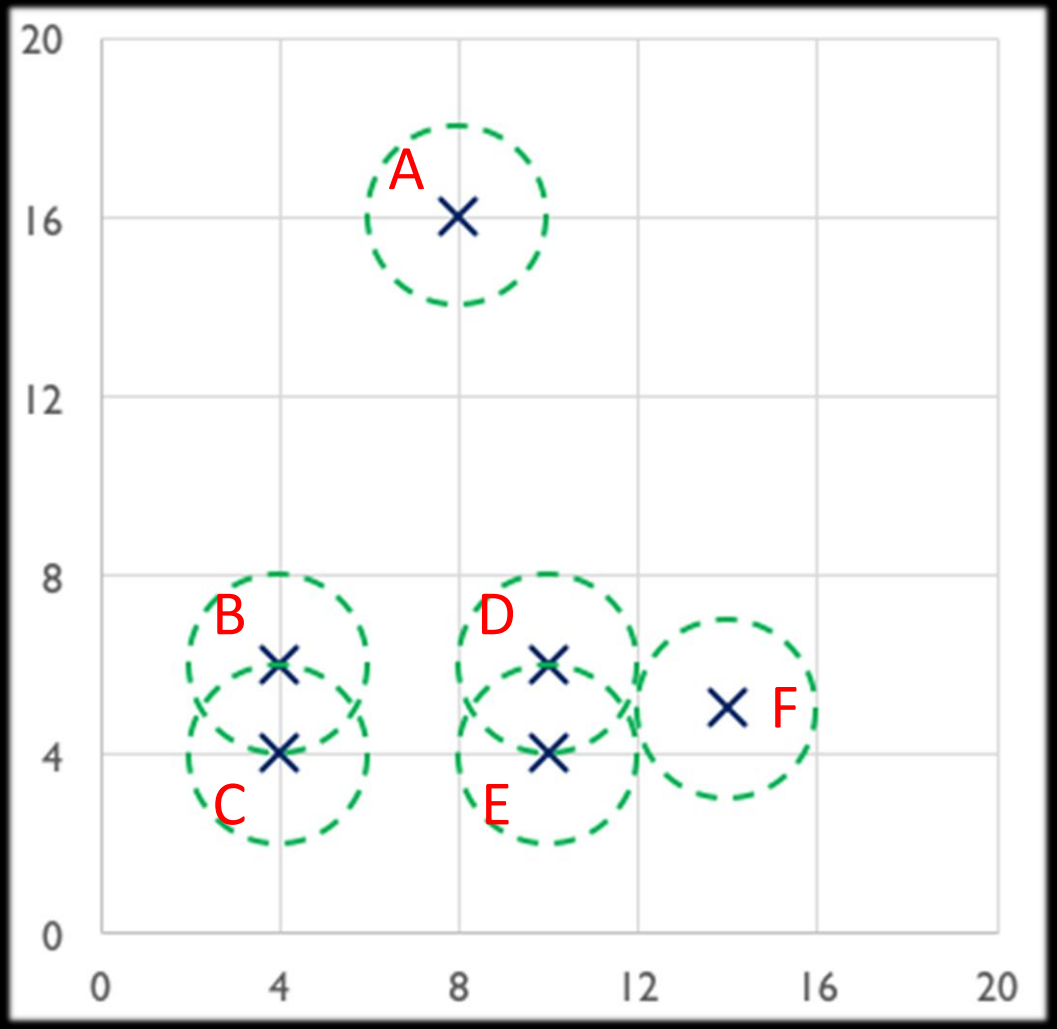
C

D

E

F

Hierarchical Clustering



$r = 2$

Clusters (4)

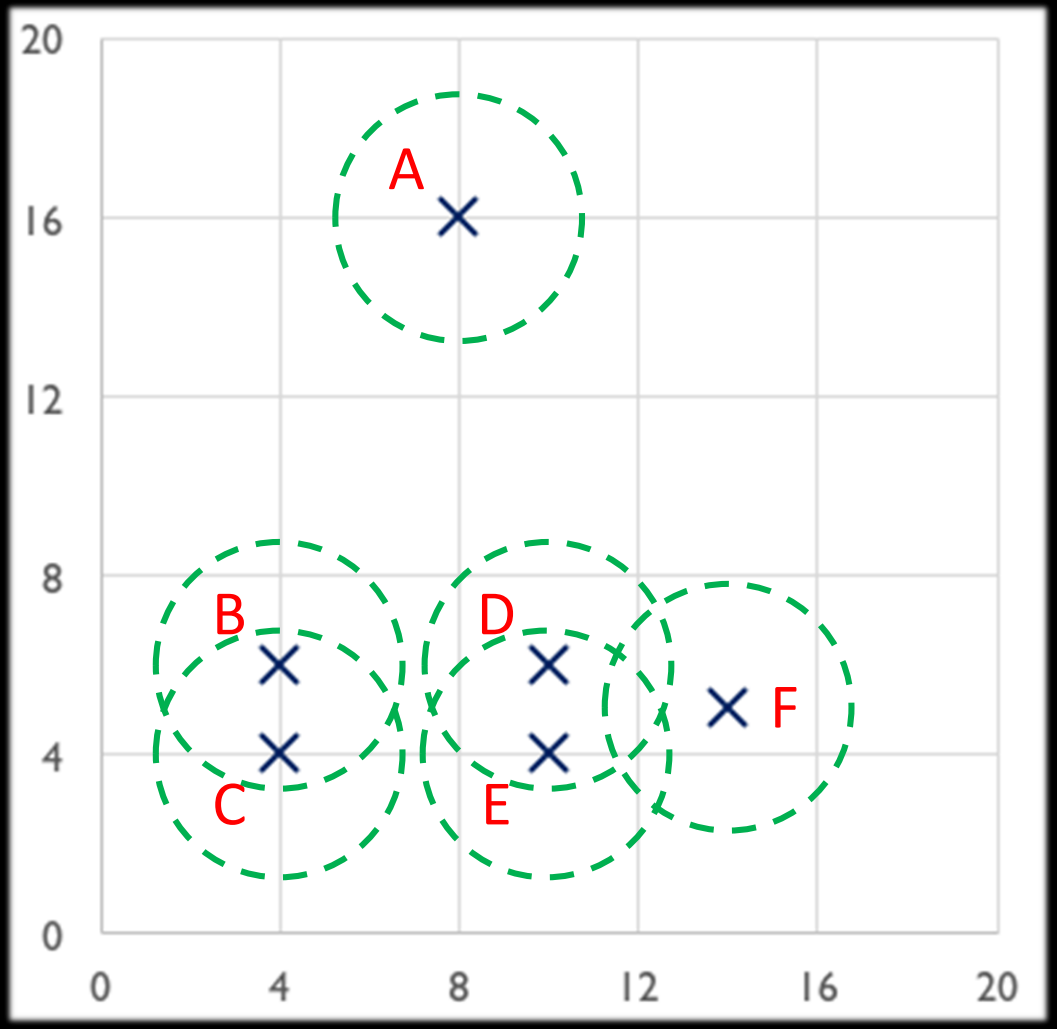
A

B C

D E

F

Hierarchical Clustering



$r = 2.7$

Clusters (4)

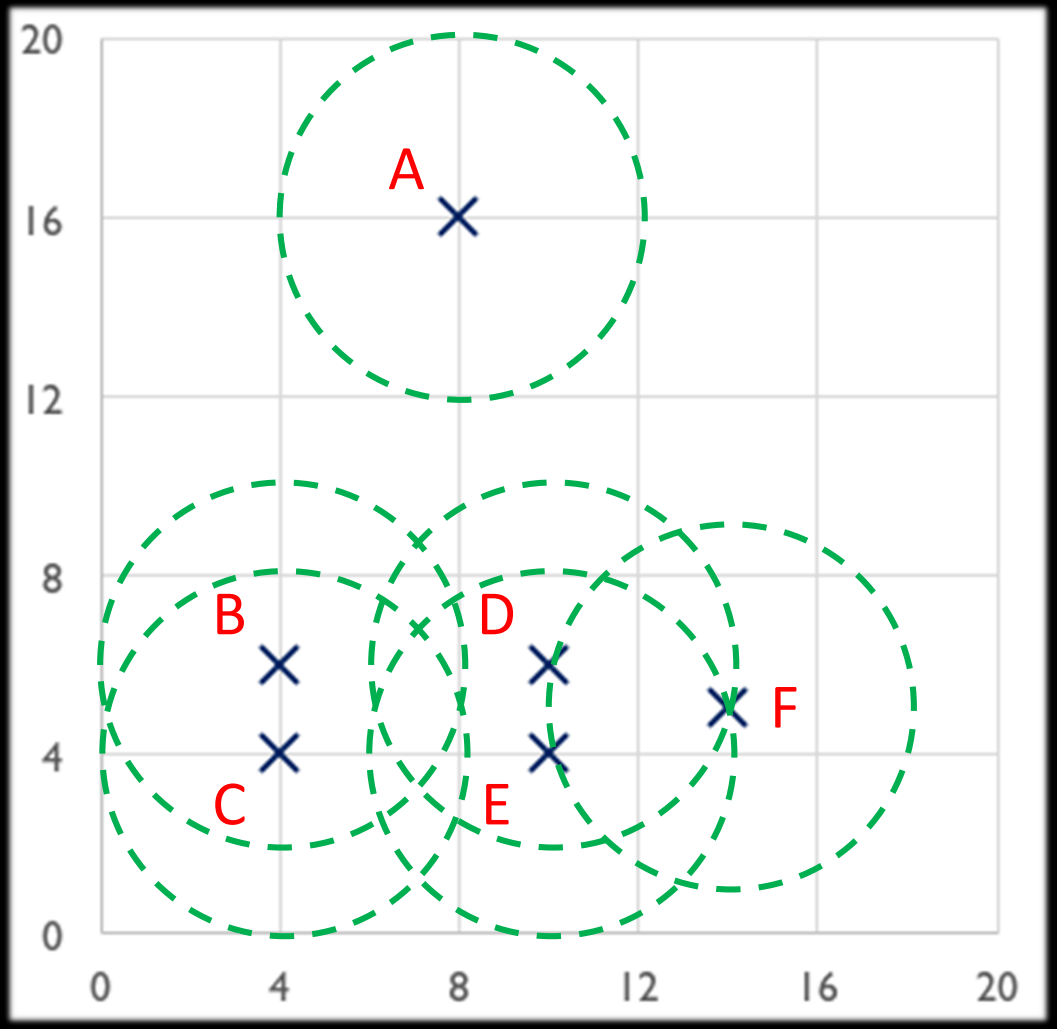
A

B C

D E

F

Hierarchical Clustering



$r = 4.2$

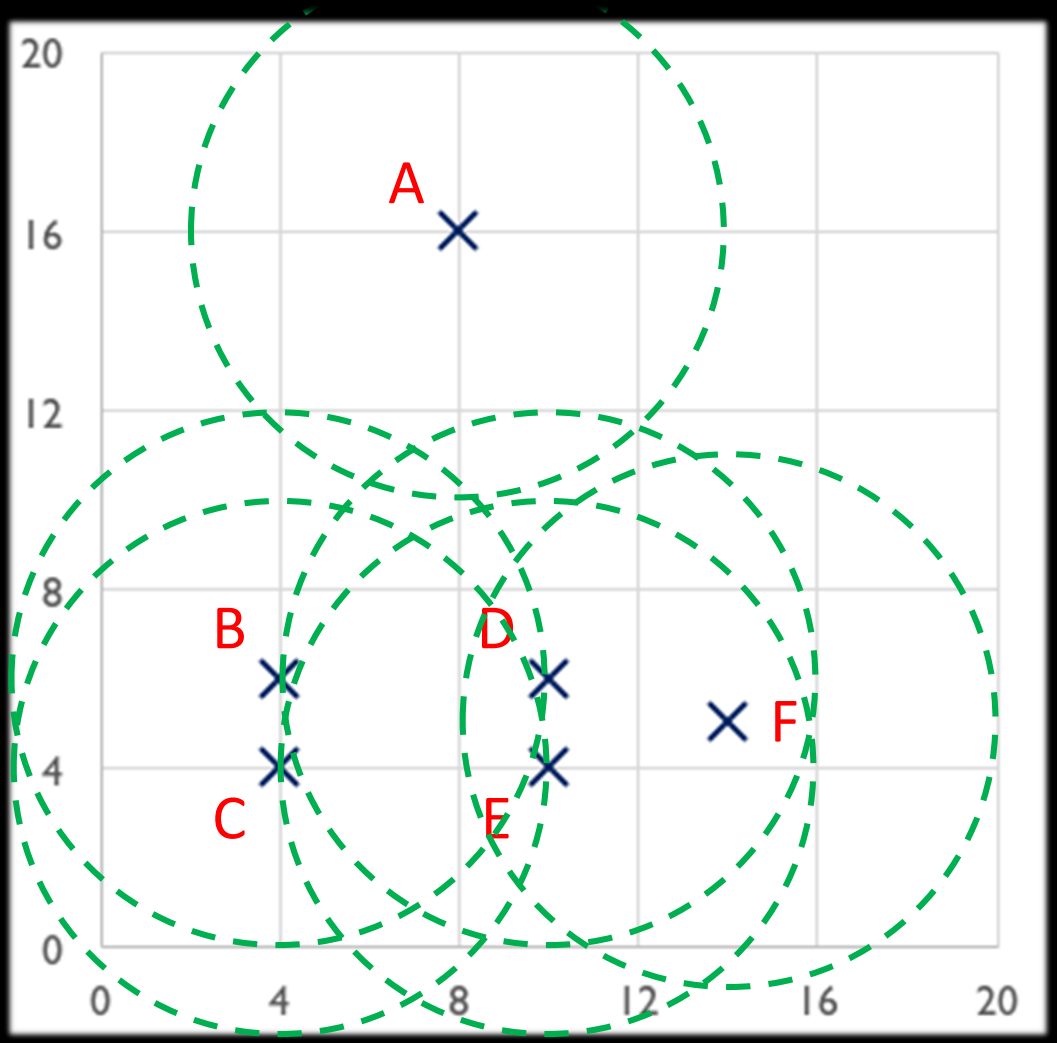
Clusters (3)

A

B C

D E F

Hierarchical Clustering



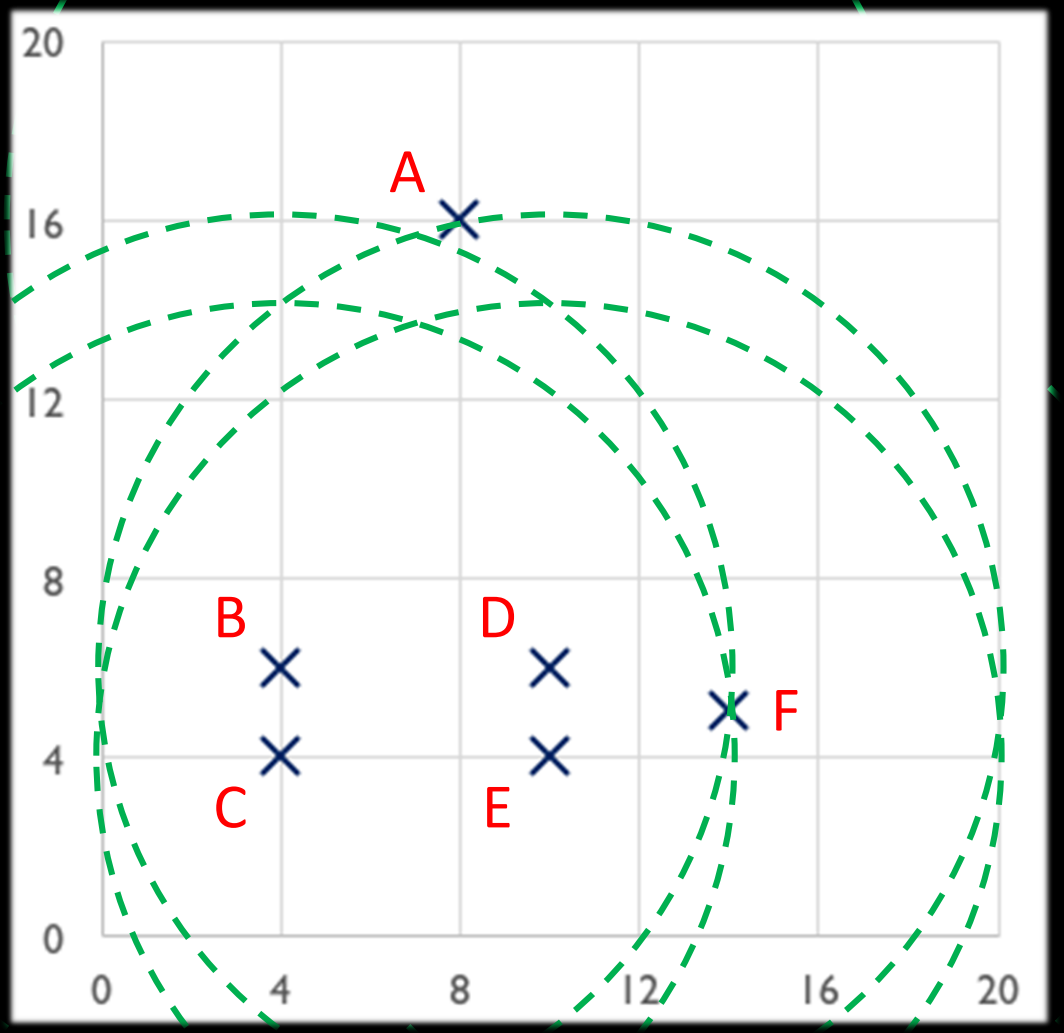
$r = 6$

Clusters (2)

A

B C D E F

Hierarchical Clustering



$r = 10.2$

Clusters (1)

A B C D E F

Hierarchical clustering

Agglomerative

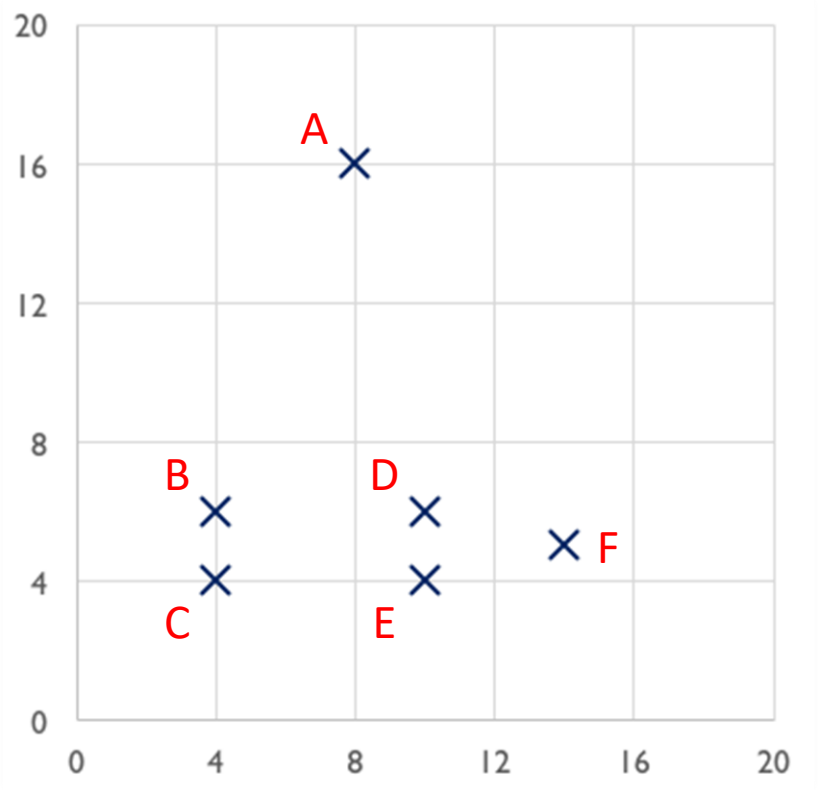
Bottom Up: Begins with one cluster per data point;
Gradually merges into larger clusters.

Divisive

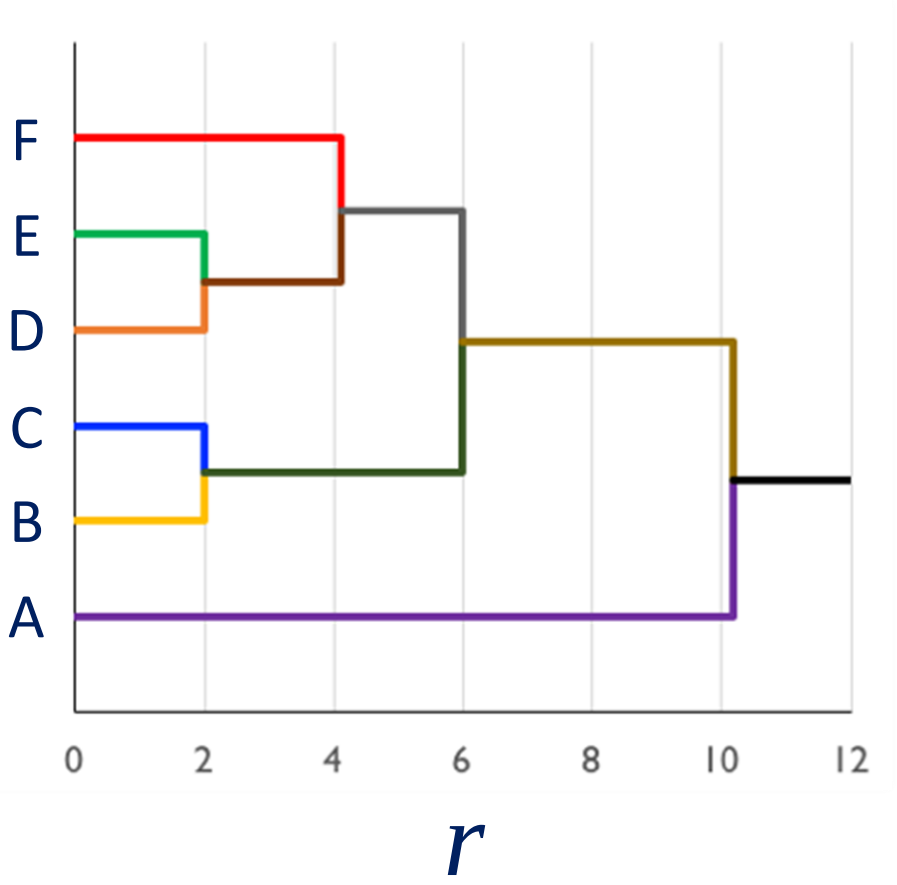
Top Down: Begins with one big cluster;
Gradually splits into smaller clusters.

Requires… A dissimilarity measure (a distance)

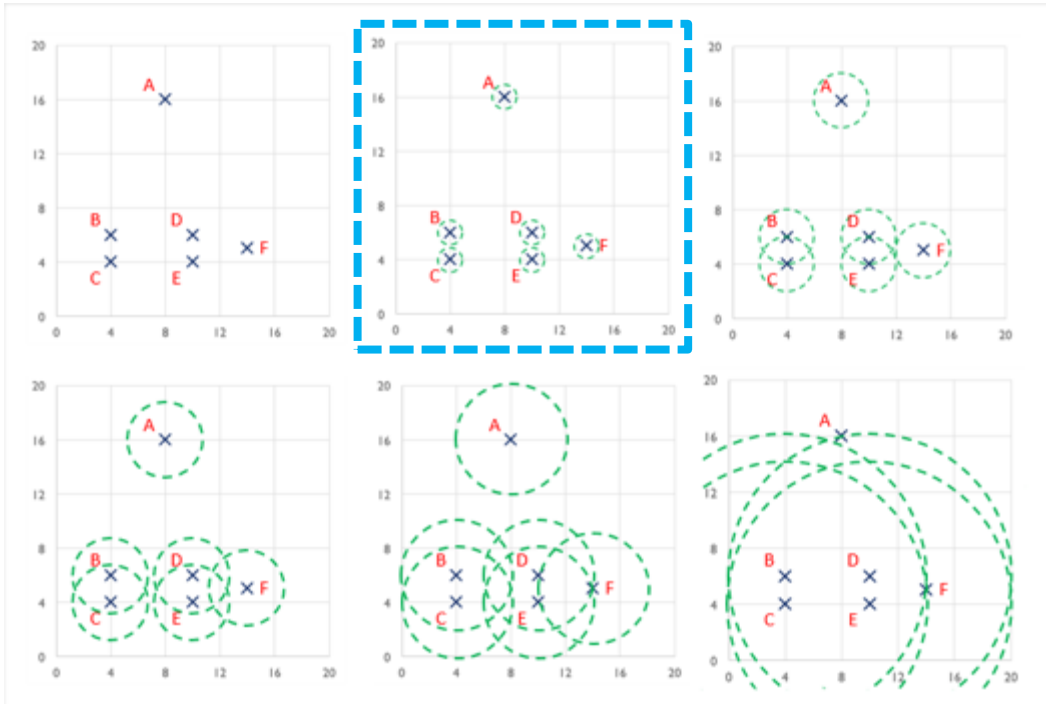
Hierarchical Clustering



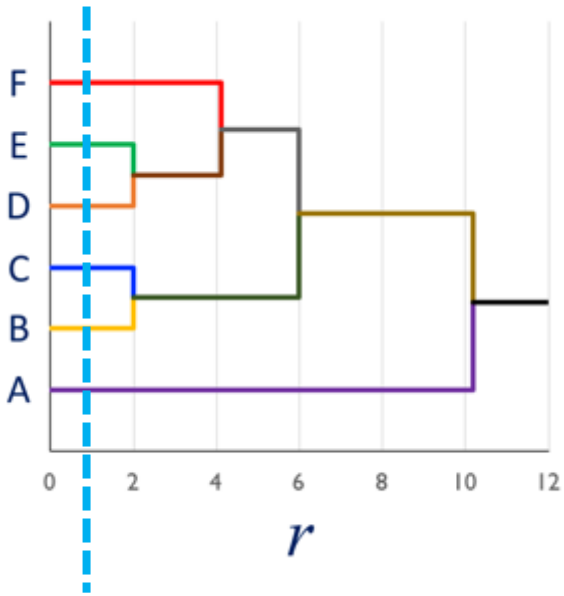
Dendrogram



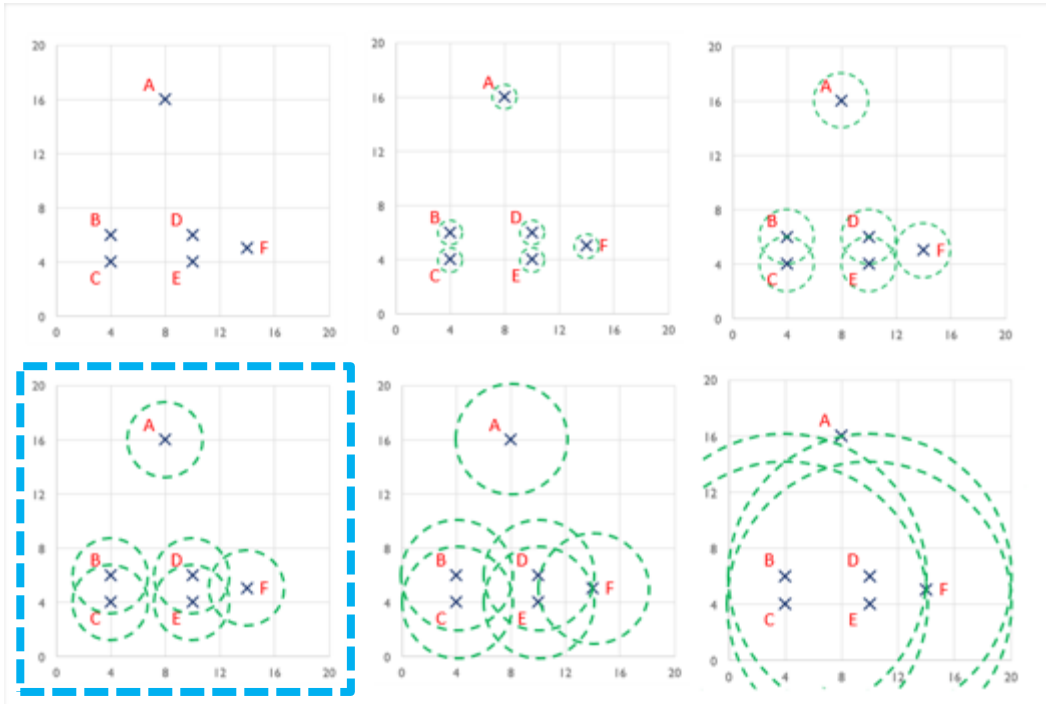
Hierarchical Clustering



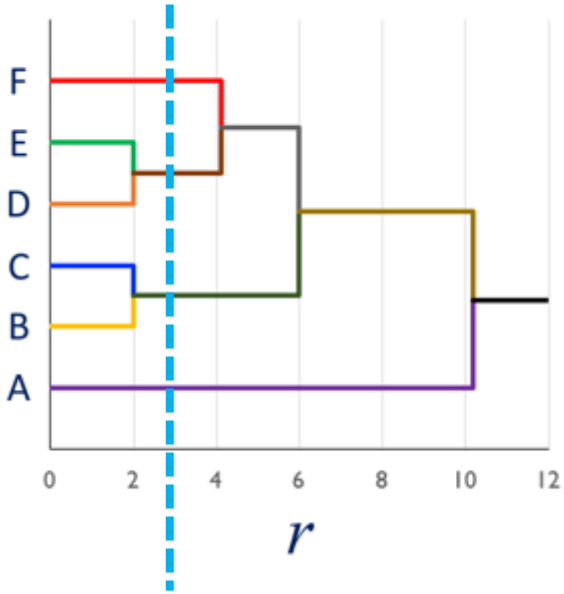
Dendrogram



Hierarchical Clustering

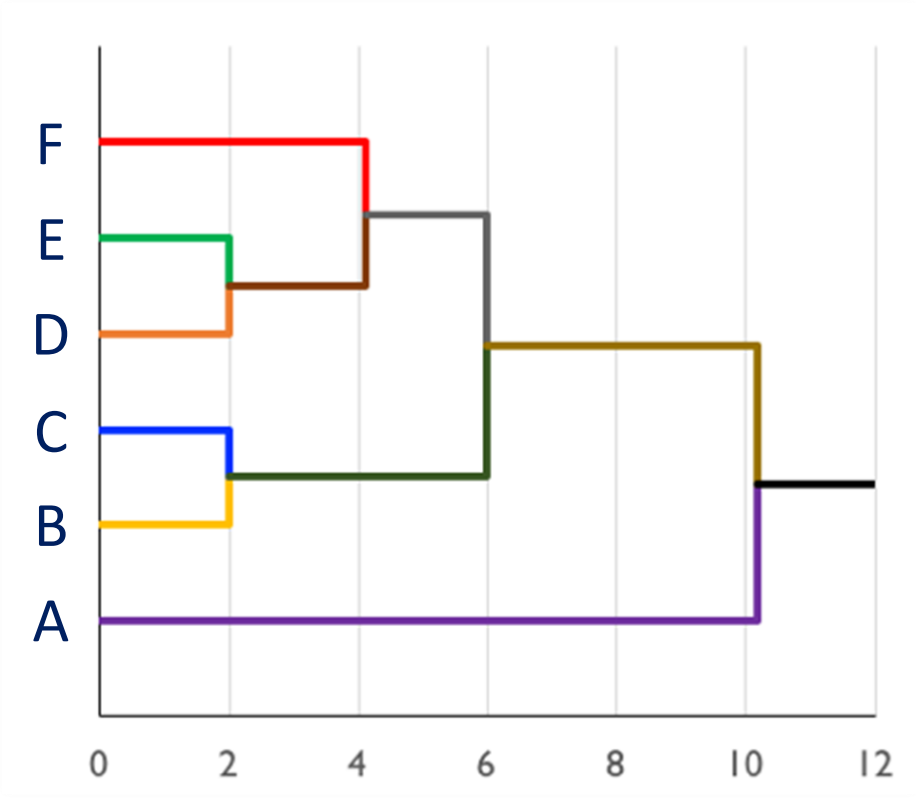


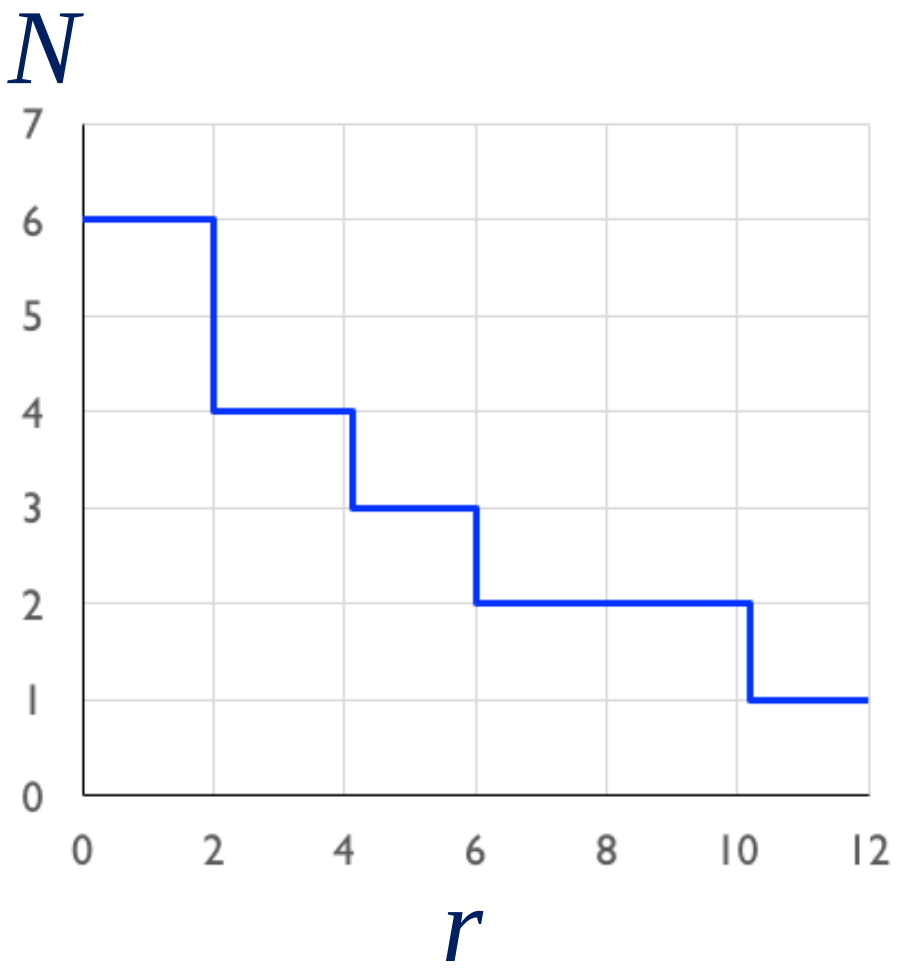
Dendrogram



Hierarchical Clustering

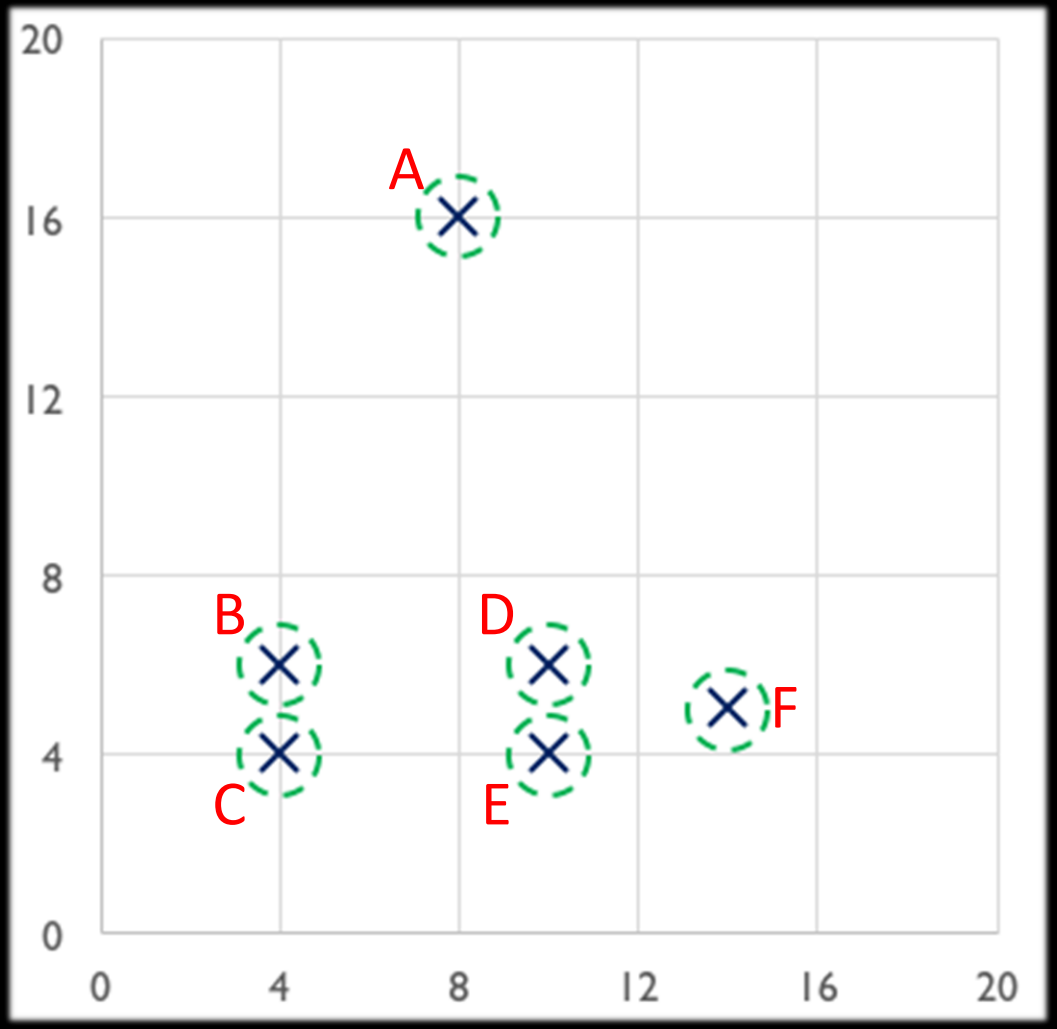
Dendrogram





Hierarchical Clustering

Different Algorithms



$r = 0.9$

Clusters (6)

A

B

C

D

E

F

It’s Tea Time

- Part 1: Hypothesis testing recap
- Part 2: Cluster analysis – why should I care?
- Part 3: K means
- Part 4:Before you start
- Part 5: Hierarchical clustering
- Part 6: How good are your clusters?
- Part 7: Some tips and tricks for your written work

Clustering

Definition:

Type of analysis that divides observations into on groups based on some similarity criteria (distance).

Examples: K-Means, Hierarchical clustering

How do we know our groups make sense?

Measuring Clustering Quality

Necessary when…

- Comparing different random implementations of k-means
- Comparing clusterings with different numbers of clusters
- Comparing different clustering techniques
  (e.g. k-means vs hierarchical)

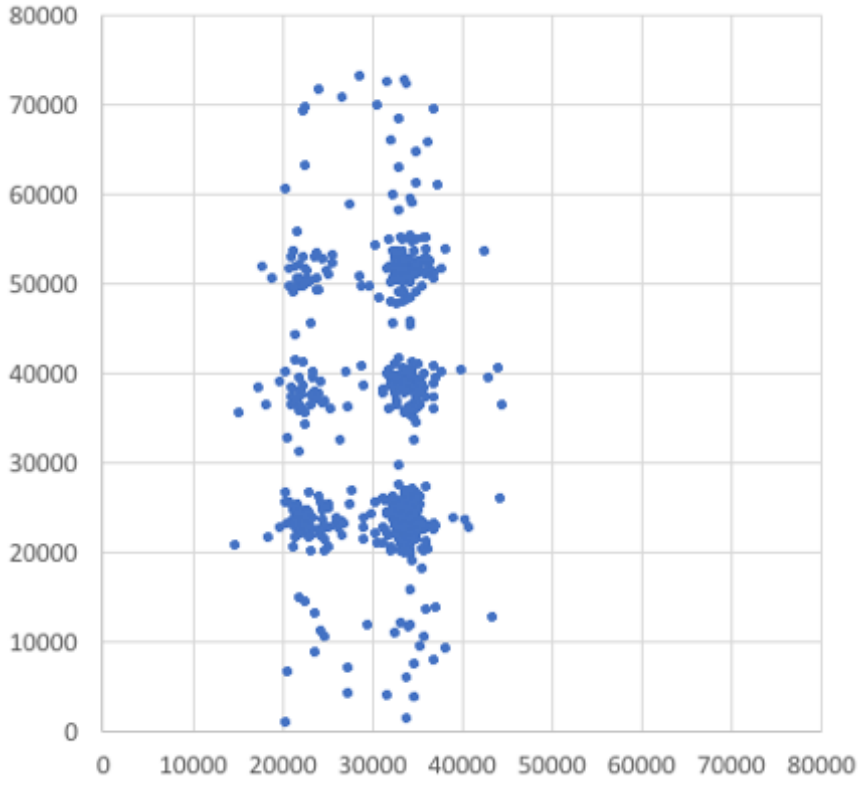
Method I : SSE / Elbow Method

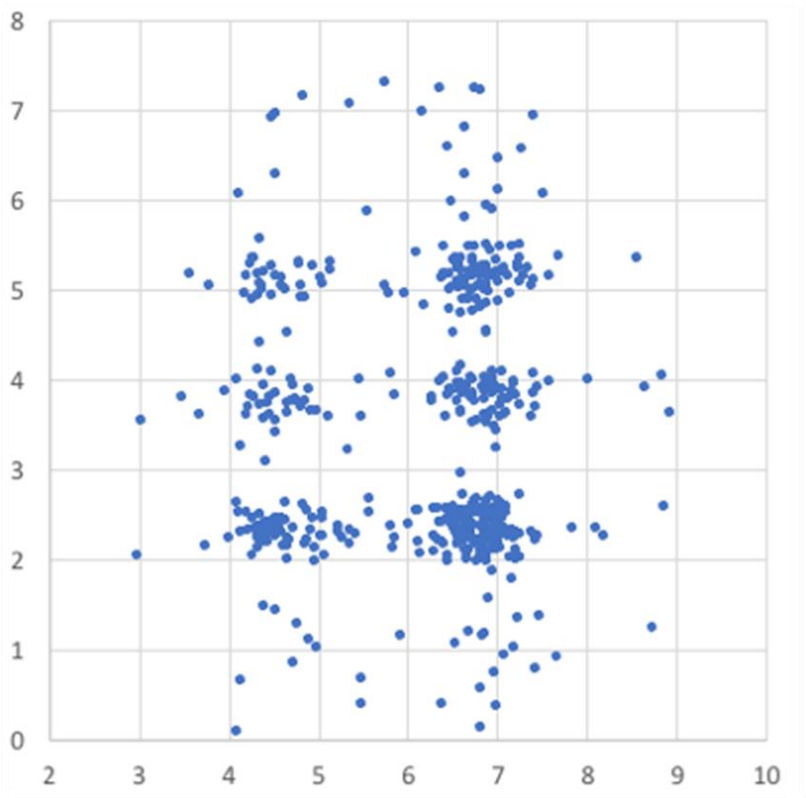
Measuring Clustering Quality

Method 1 – SSE / Elbow Method

y : Distance along fault line (m)

Y = *y* / 10000





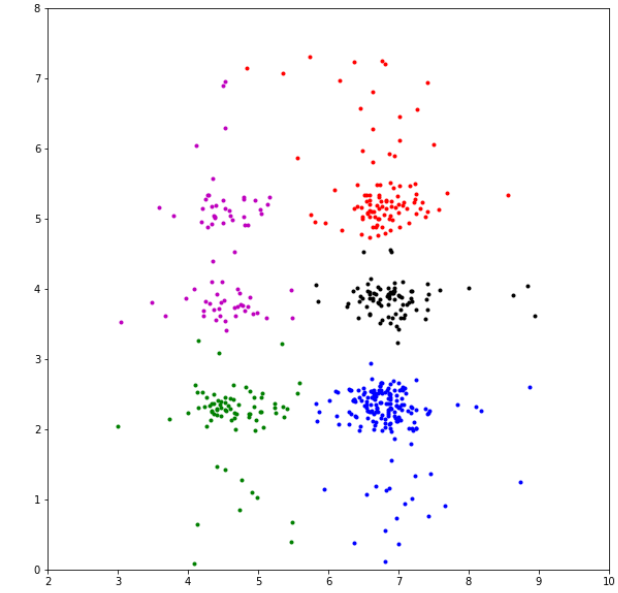
x : Time (Minutes after start of study)

X = 2 *x* / 10000

Measuring Clustering Quality

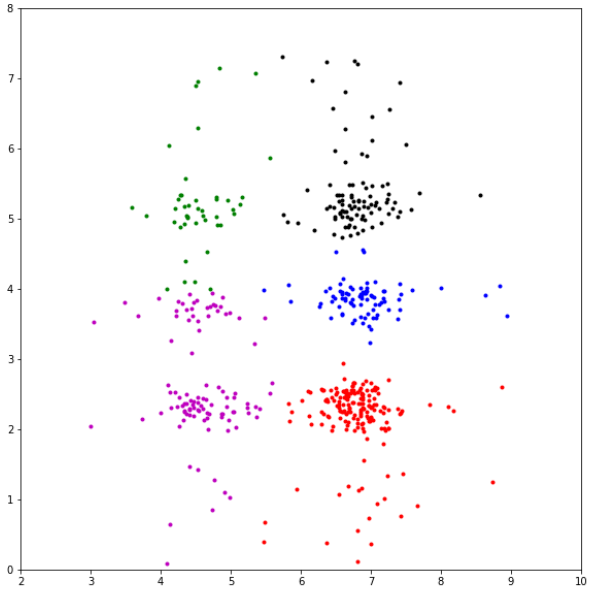
Method 1 – Sum Squared Error (SSE)

Comparing Different Random k-Means Clusterings



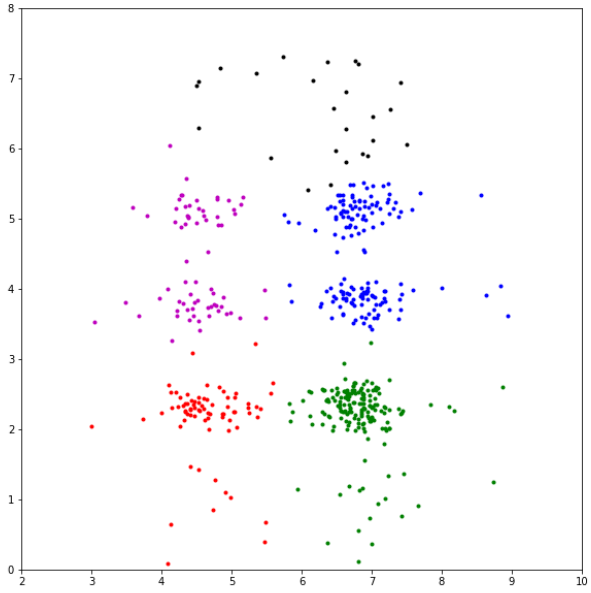
k = 5

SSE = 262.4



k = 5

SSE = 275.6



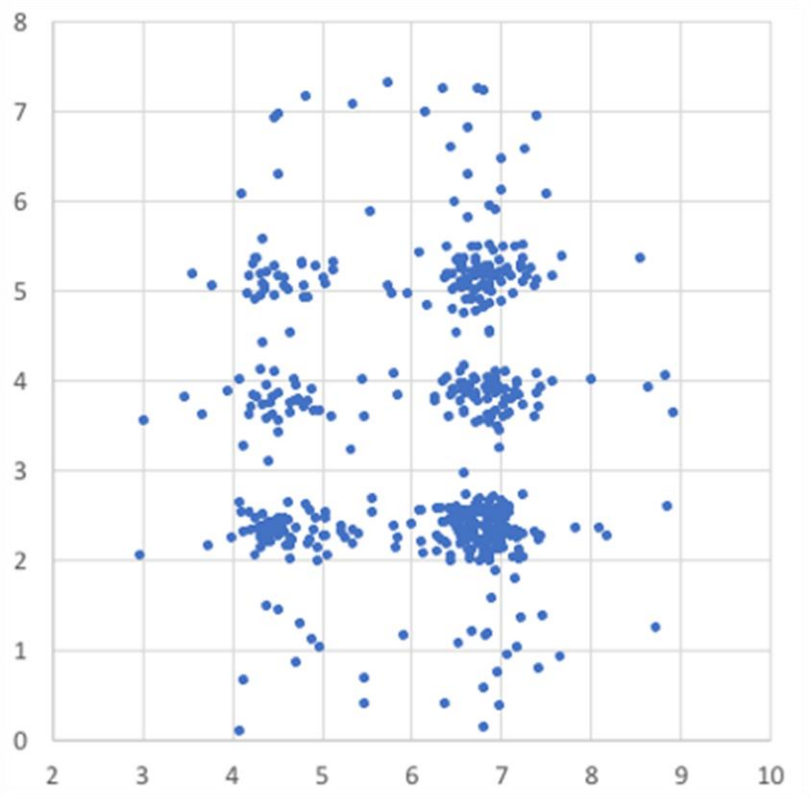
k = 5

SSE = 289.9

Measuring Clustering Quality

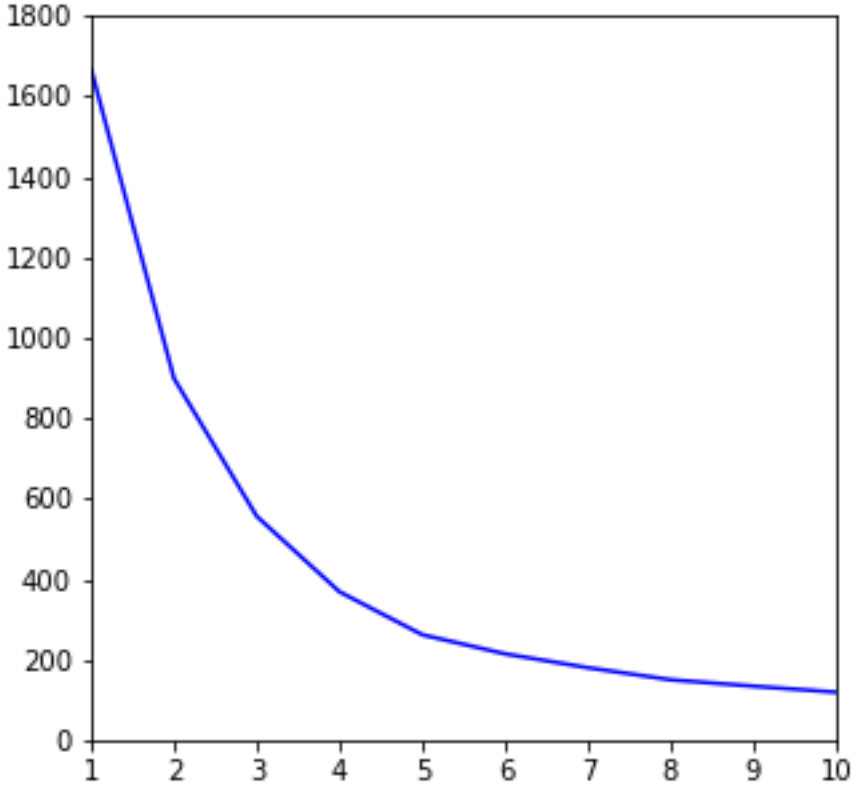
Method 1 – Elbow Diagram
Choosing the number of clusters

Y



X

SSE



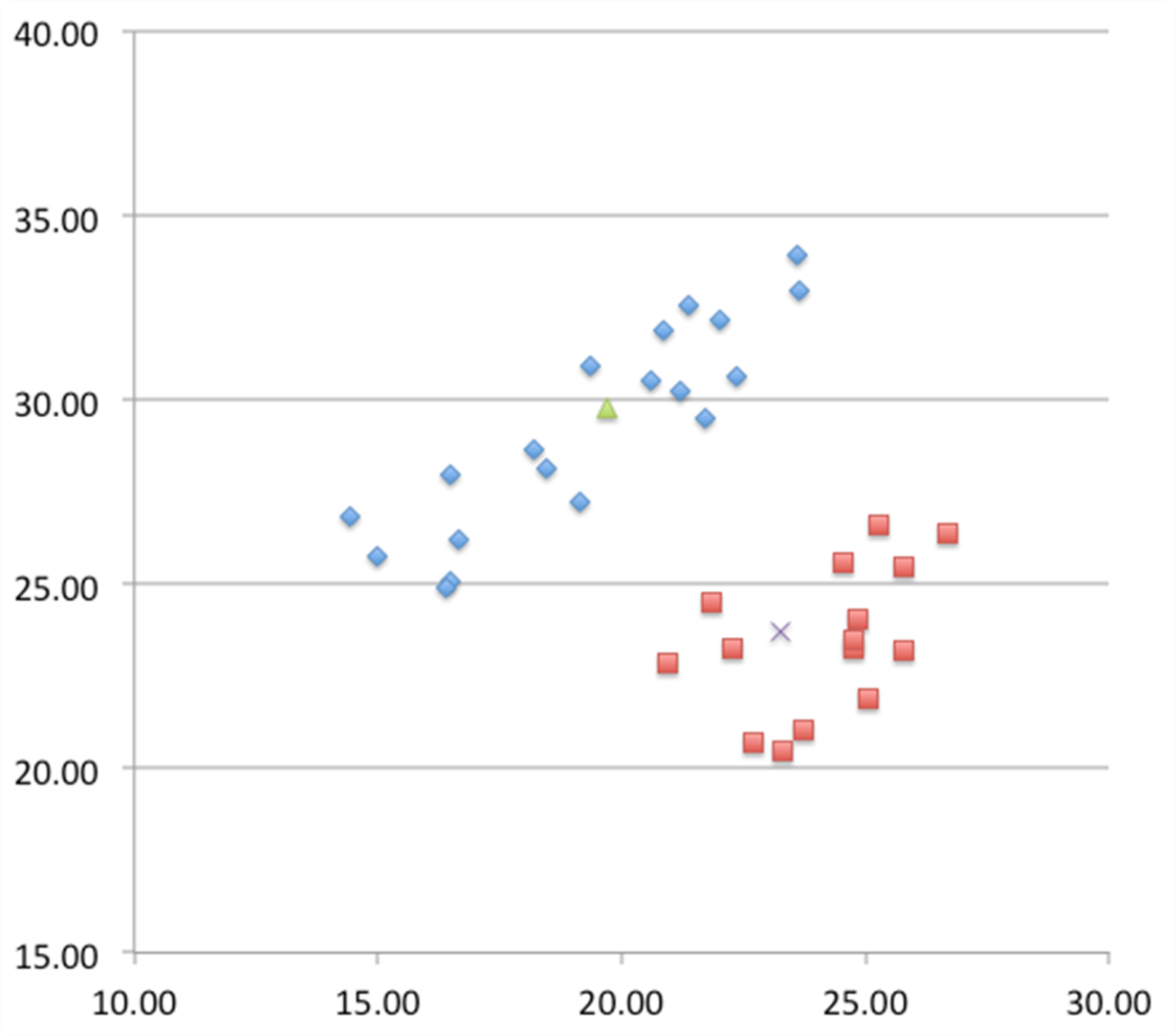
k (Number of Clusters)

Method II : Silhouette Analysis

Choosing the right number of clusters

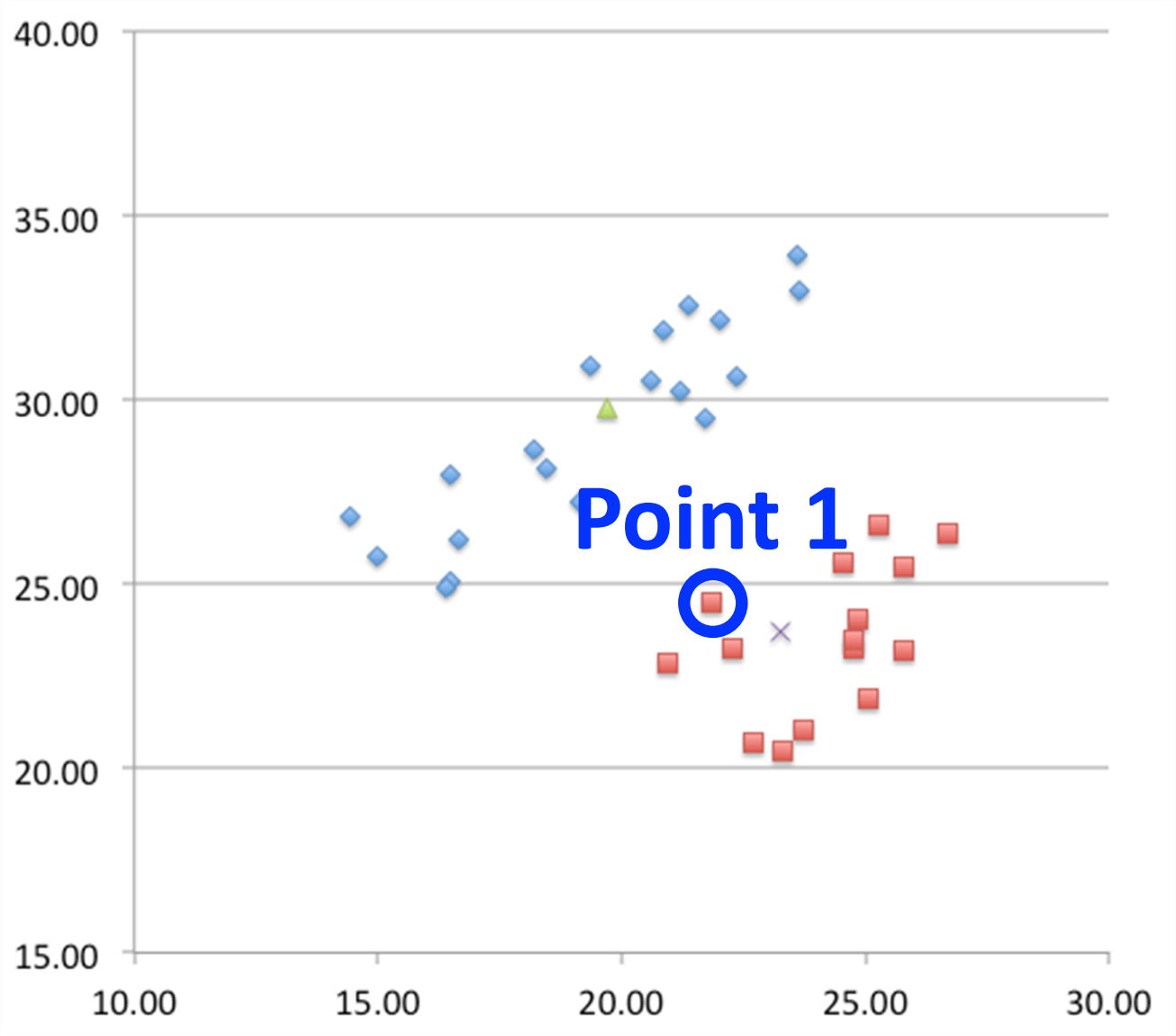
How well clustered is this data?

Mean spend on alcohol (£ / wk)



Mean spend on ready meals (£ / wk)

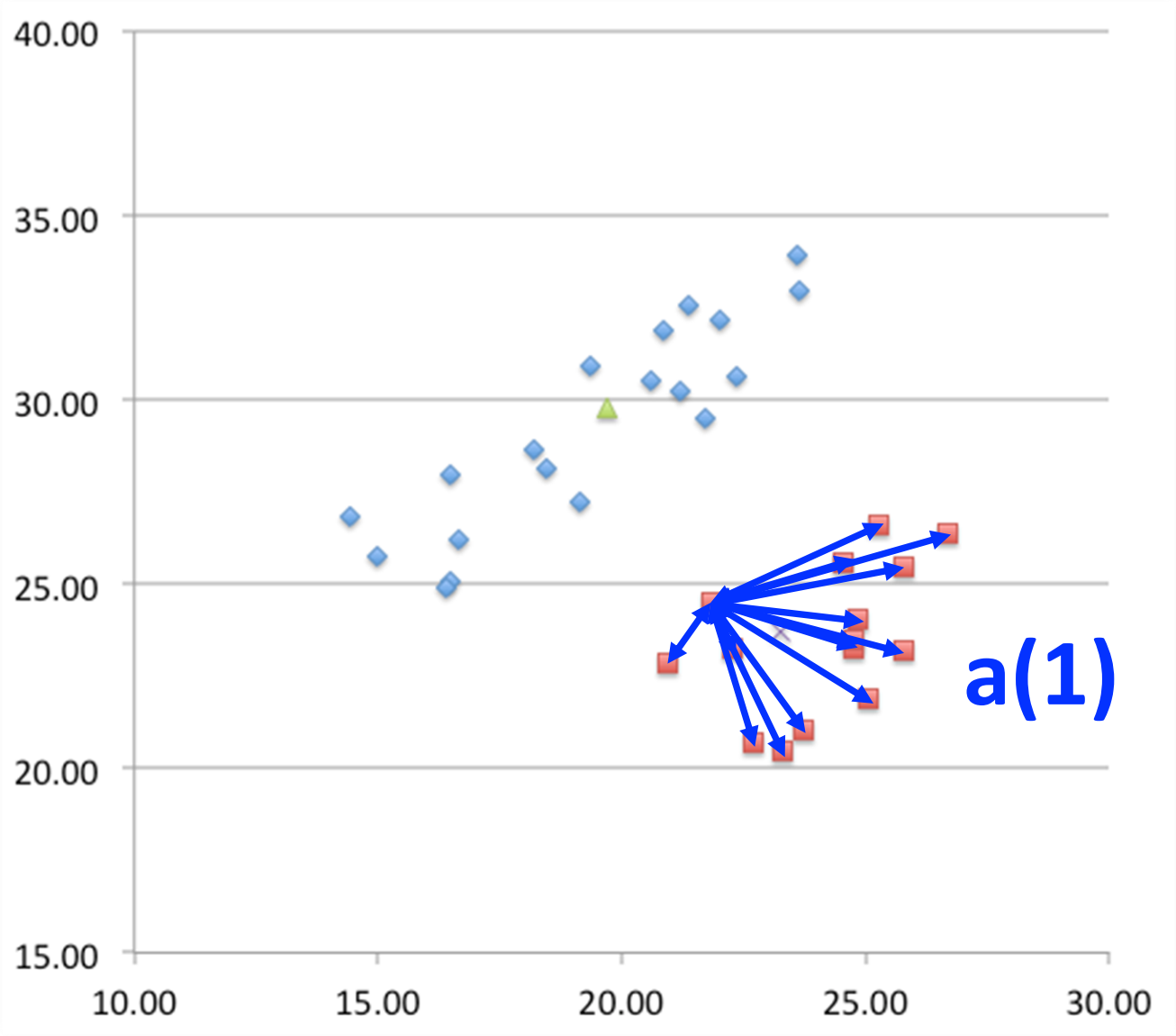
How well clustered is this data?



Mean spend on alcohol (£ / wk)

Mean spend on ready meals (£ / wk)

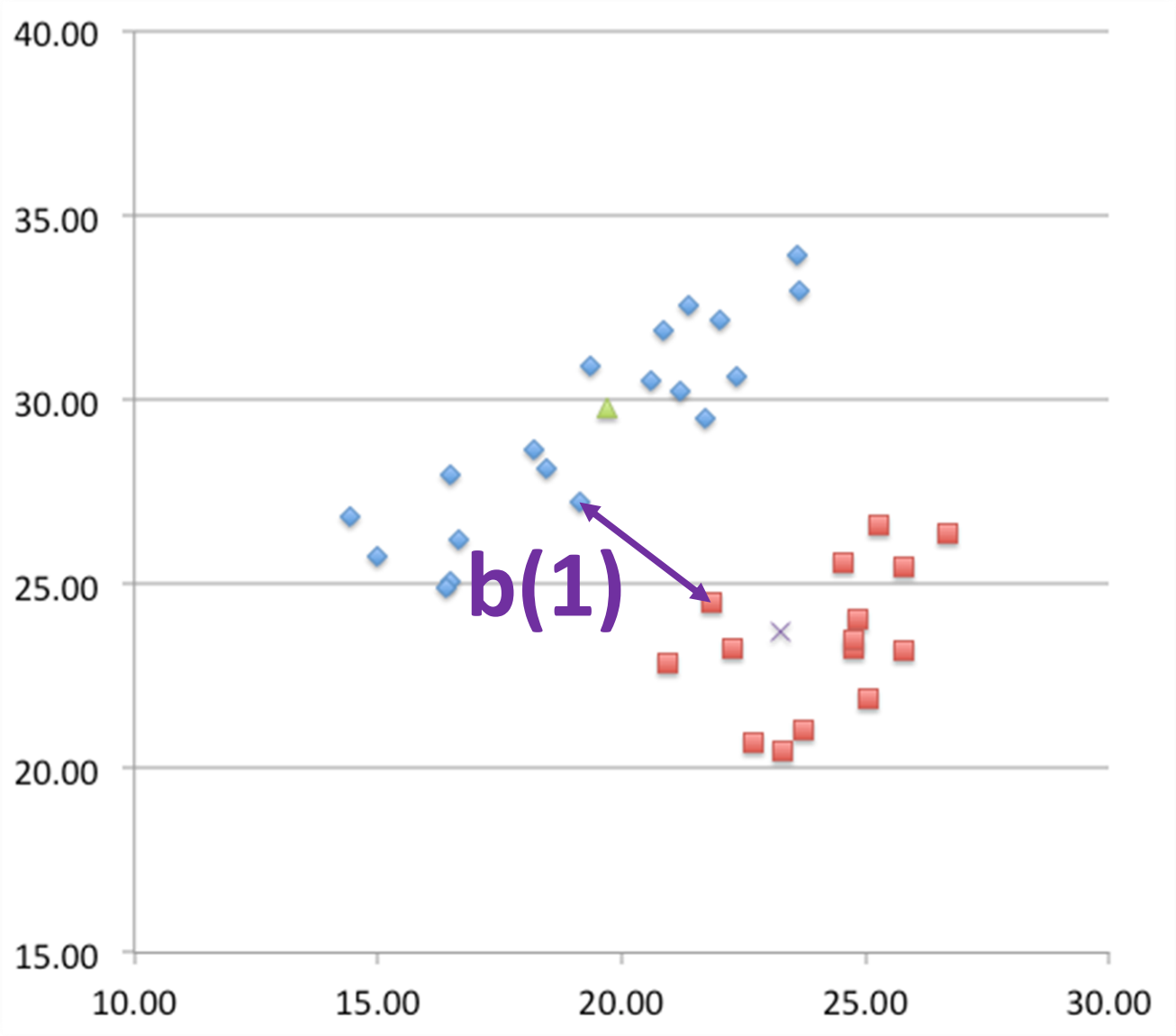
How well clustered is this data?



Mean spend on alcohol (£ / wk)

Mean spend on ready meals (£ / wk)

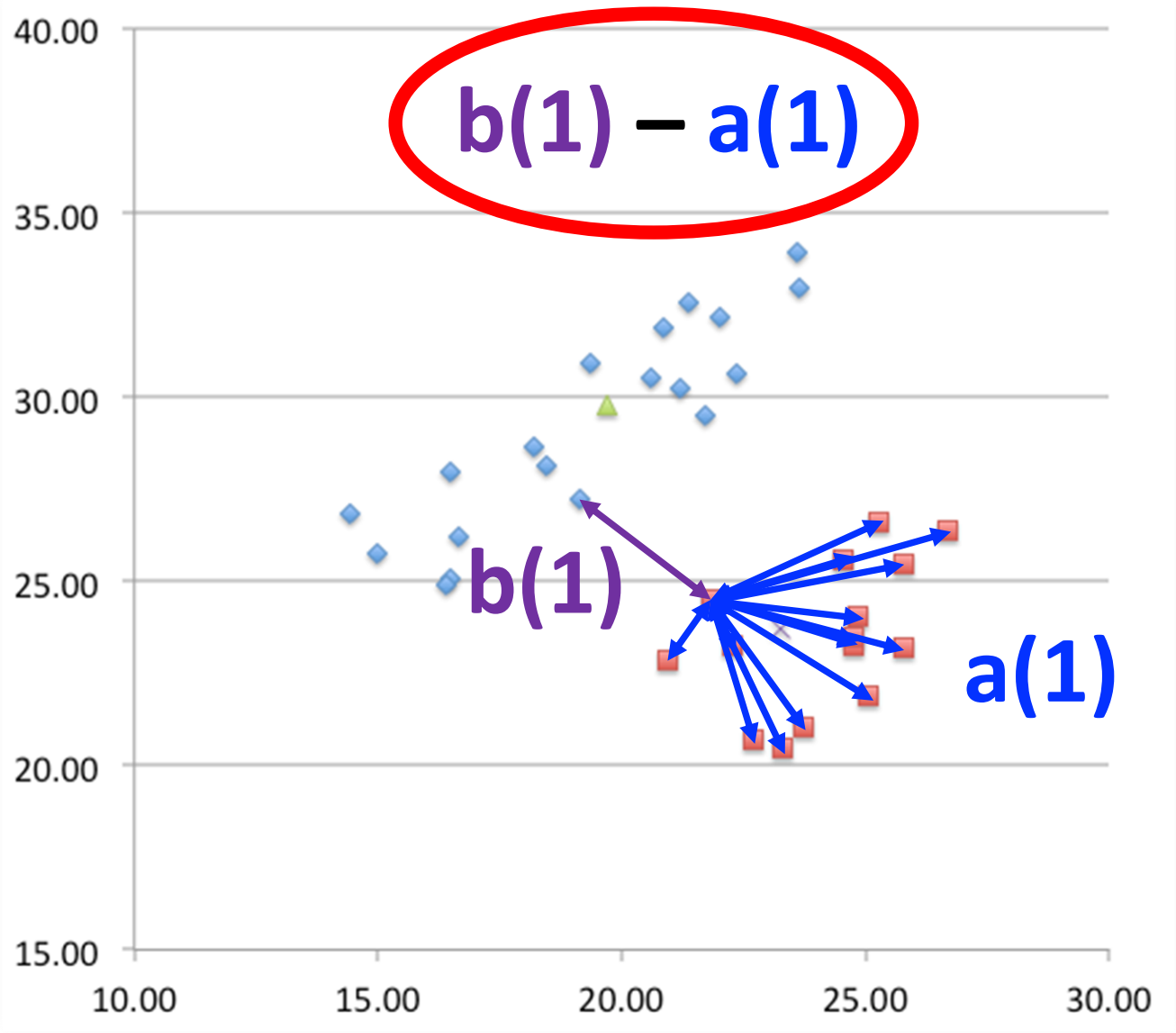
How well clustered is this data?



Mean spend on alcohol (£ / wk)

Mean spend on ready meals (£ / wk)

How well clustered is this data?



Mean spend on alcohol (£ / wk)

Mean spend on ready meals (£ / wk)

The Silhouette of a Point

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

The Silhouette of a Point

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

$$-1 \leq s(i) \leq 1$$

The Silhouette of a Point

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

$$-1 \leq s(i) \leq 1$$

poorly clustered

well clustered

The Silhouette of a Point

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

$$-1 \leq s(i) \leq 1$$

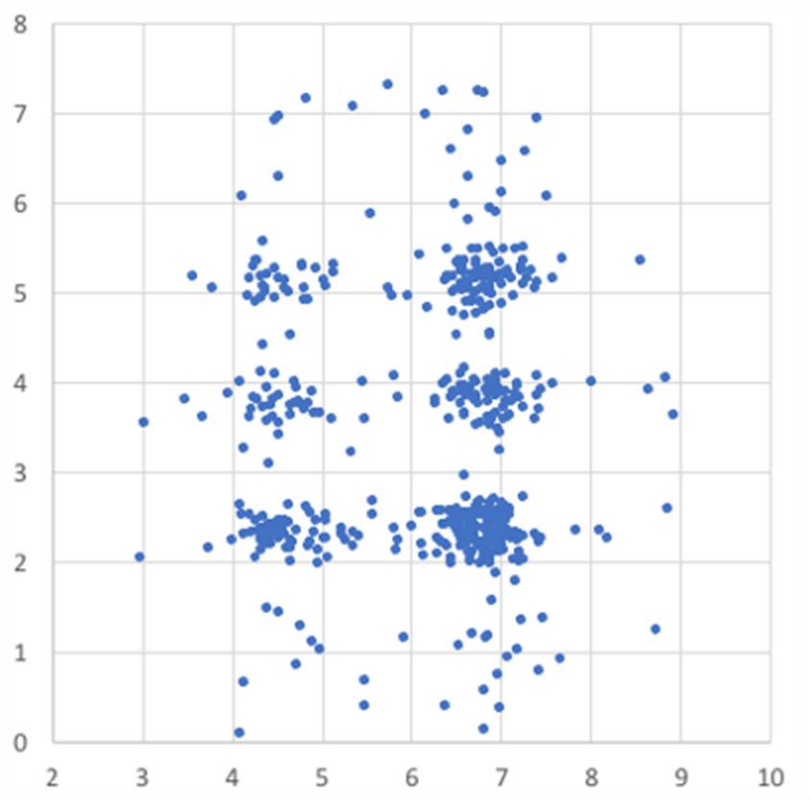
poorly clustered

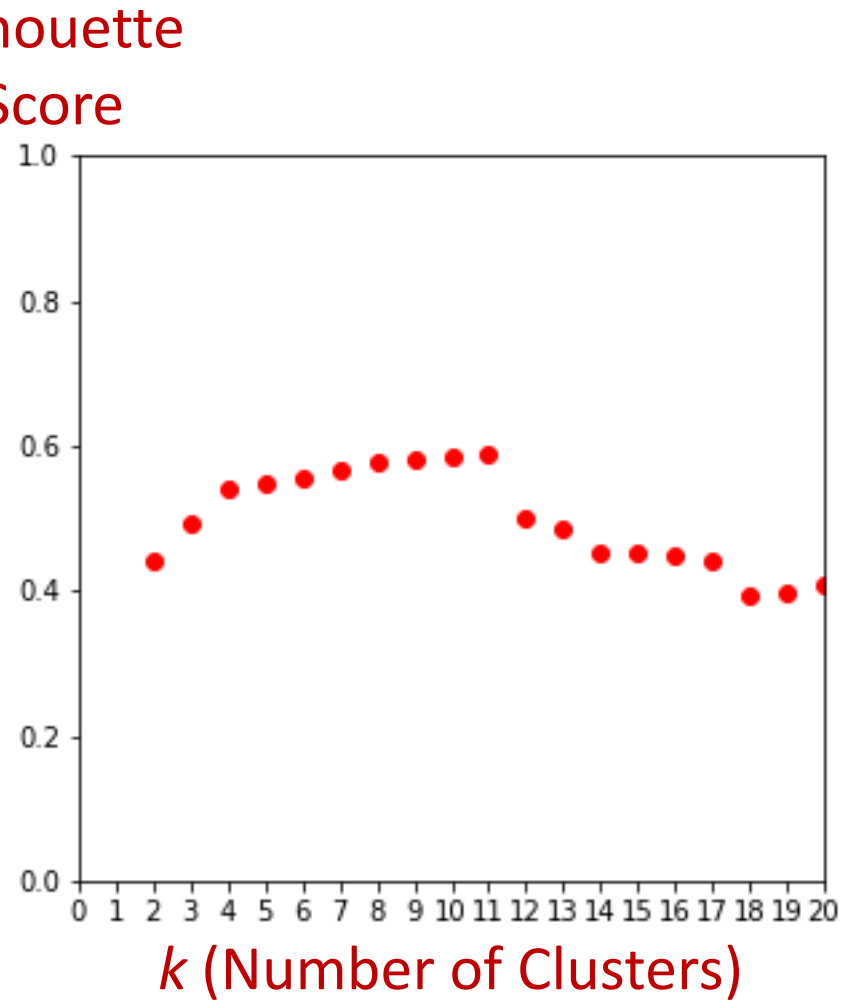
well clustered

Silhouette Score for a Clustering = *Average of s(i) for all points i*

Measuring Clustering Quality

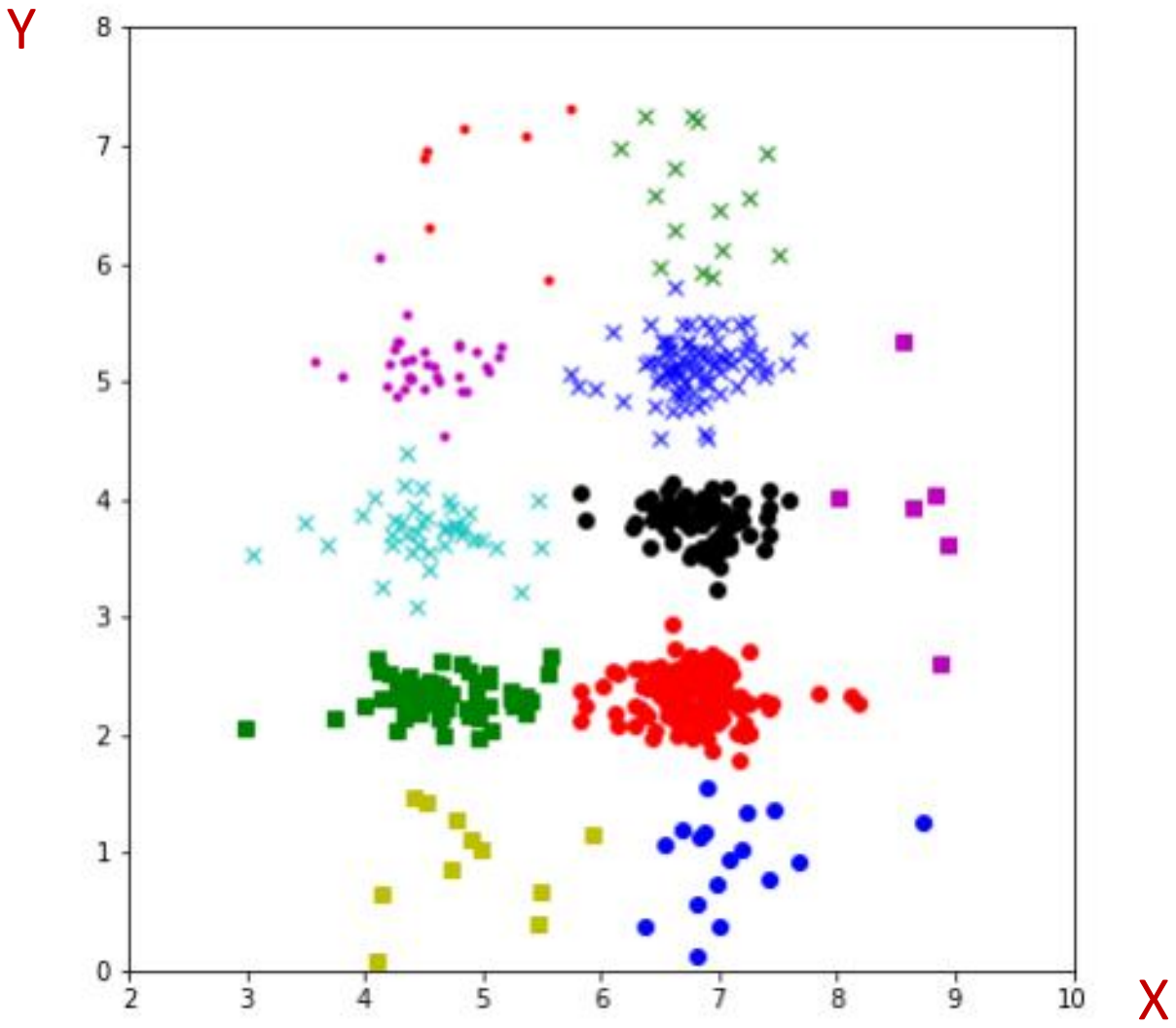
Method 1I – Compare Silhouette Scores (k-means)





Measuring Clustering Quality

Method 1I – Compare Silhouette Scores (k-means)



'Optimal' k-Means

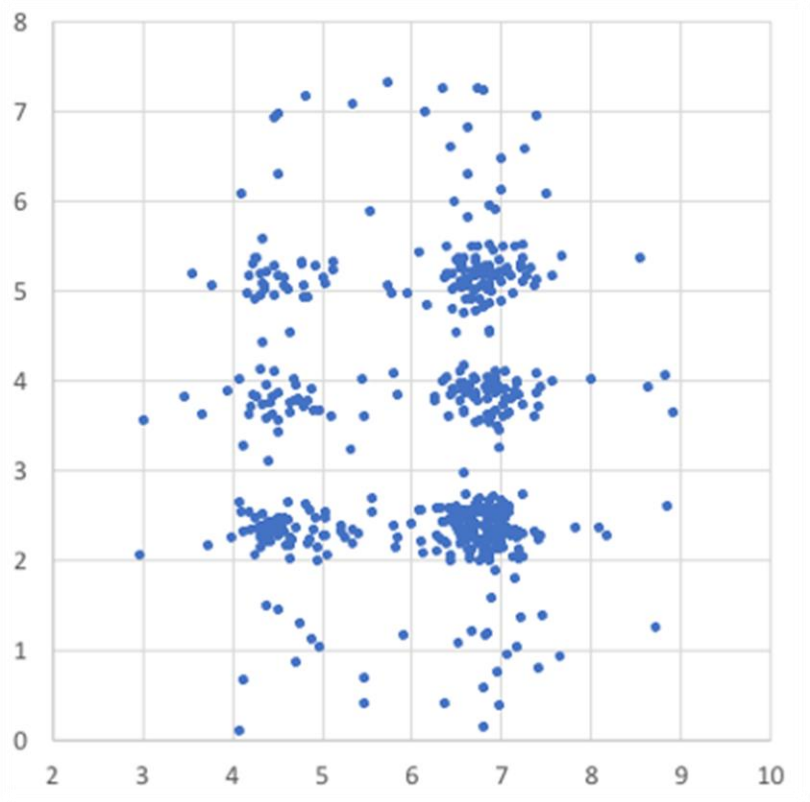
k = 11

S. Score = 0.59

Measuring Clustering Quality

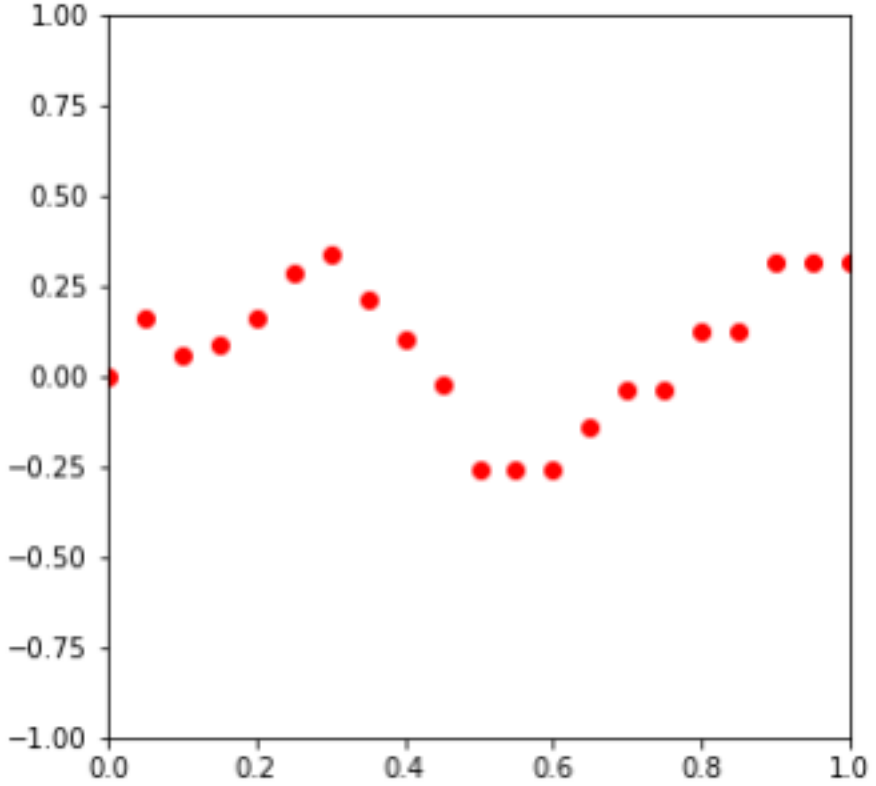
Method 1I – Compare Silhouette Scores (Hierarchical)

Y



X

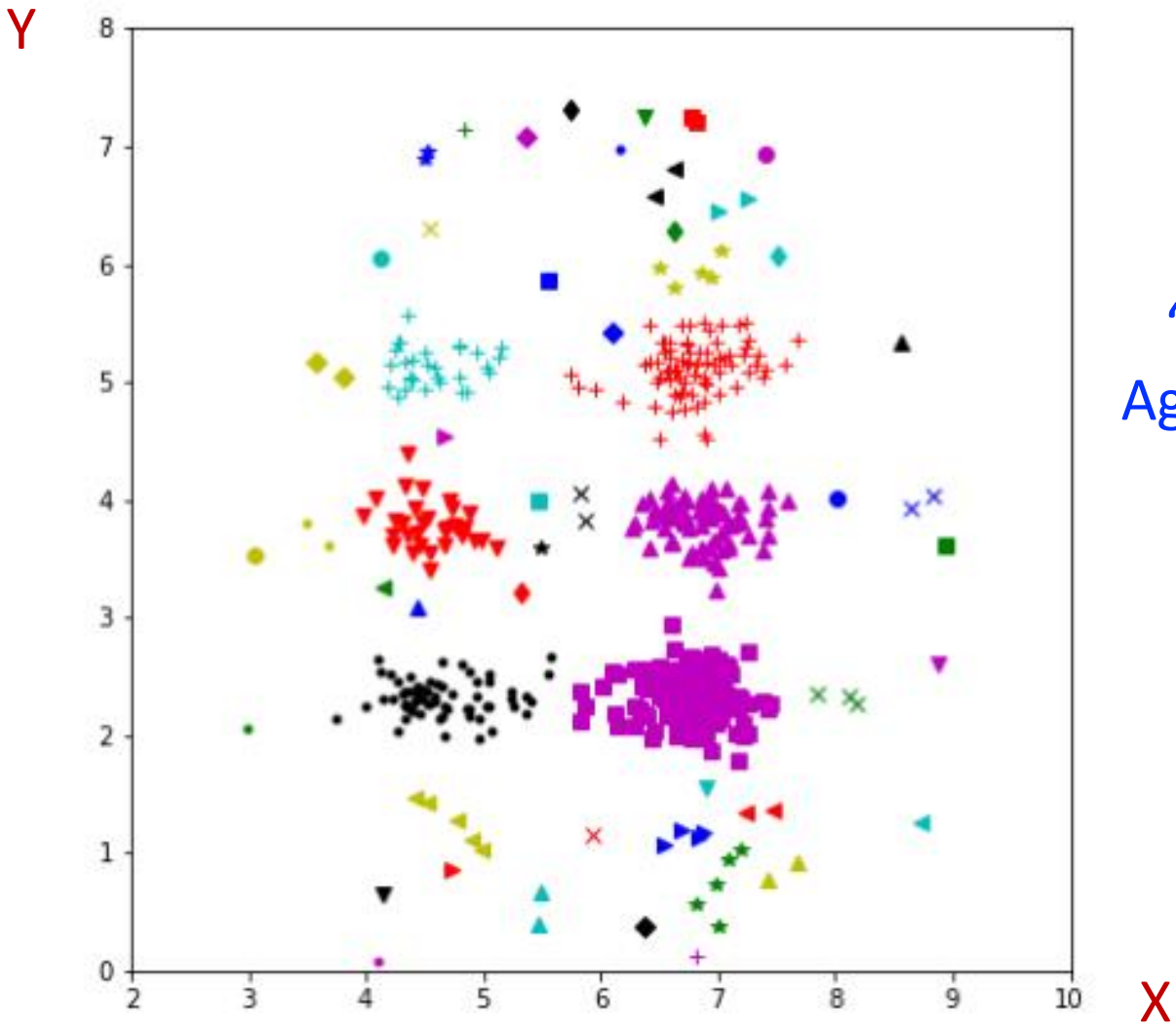
Silhouette Score



r (agglomeration threshold)

Measuring Clustering Quality

Method 1I – Compare Silhouette Scores (Hierarchical)



'Optimal' Hierarchical Agglomerative Clustering

$r = 0.3$

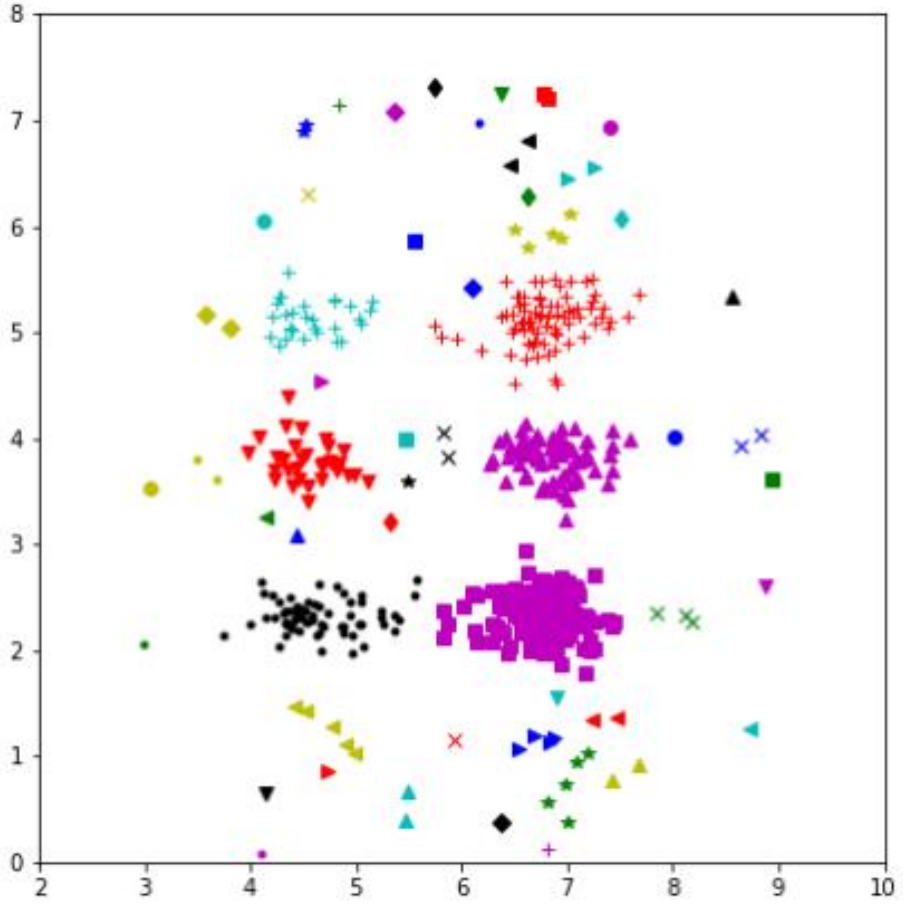
$k = 55$

S. Score = 0.34

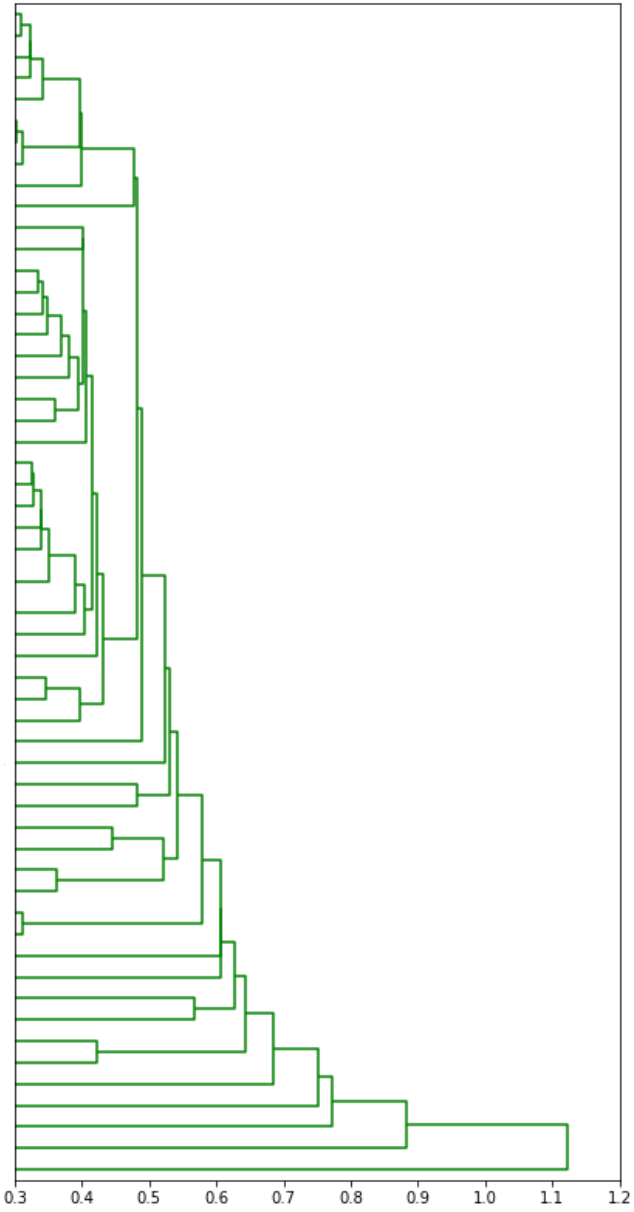
Measuring Clustering Quality

Method 1I – Compare Silhouette Scores (Hierarchical)

Y



X



r (agglomeration threshold)

Clustering : What Next?

- Identify clusters
- Observe and interpret
- Consider analysing separately

Clustering : Plan of Attack

Standardisation Methods
Z-Score (roughly symmetrical data)
Min-Max rescaling (asymmetric data)
IDR rescaling (data with significant outliers)
Explicit rescaling

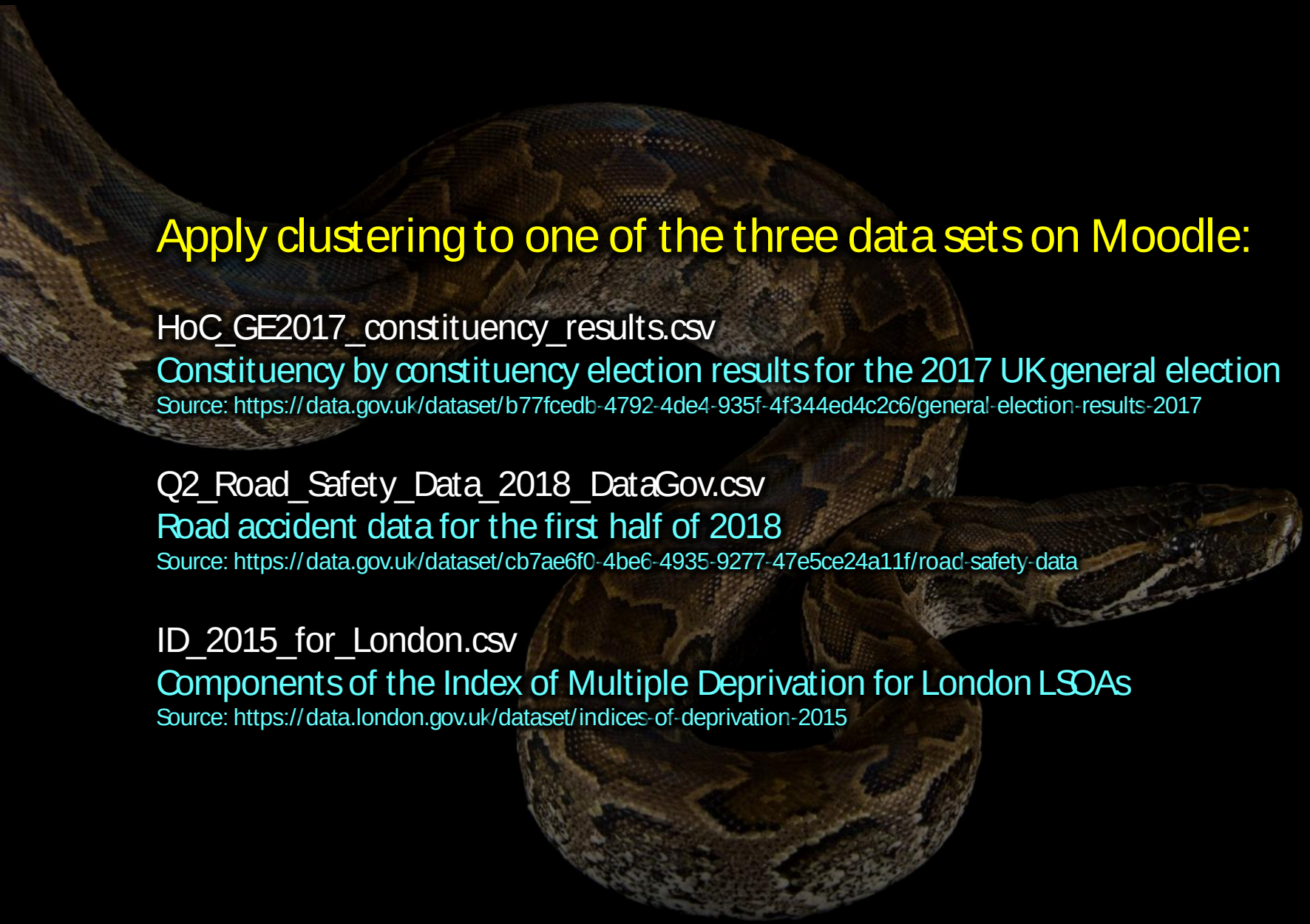
Clustering Methods
K-Means
Hierarchical

Clustering Quality
SSE
Silhouette Analysis

Visualisation
Elbow Diagram
Silhouette Plot
Dendrogram
Scatter Plots

Follow Up
Examine cluster centroids
Describe cluster characteristics
Compare against unconsidered variables / categories / geography
Consider analysing clusters separately

It’s Tea Time



Apply clustering to one of the three data sets on Moodle:

HoC_GE2017_constituency_results.csv

Constituency by constituency election results for the 2017 UK general election

Source: https://data.gov.uk/dataset/b77fcedb-4792-4de4-935f-4f344ed4c2c6/general-election-results-2017

Q2_Road_Safety_Data_2018_DataGov.csv

Road accident data for the first half of 2018

Source: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

ID_2015_for_London.csv

Components of the Index of Multiple Deprivation for London LSOAs

Source: https://data.london.gov.uk/dataset/indices-of-deprivation-2015

- **Part 1: Hypothesis testing recap**
- **Part 2: Cluster analysis – why should I care?**
- **Part 3: K means**
- **Part 4: Before you start**
- **Part 5: Hierarchical clustering**
- **Part 6: How good are your clusters?**
- **Part 7: Some tips and tricks for your written work**

What sections should be included in a quantitative essay?

What sections should be included in a quantitative essay?

Introduction

Literature Review

Research Question

(Hypothesis)

Presentation of Data

Methodology

Results

Discussion

Conclusion

What sections should be included in a quantitative essay?

Introduction

Literature Review

Research Question

(Hypothesis)

Presentation of Data

Methodology

Results

Discussion

Conclusion

What sections should be included in a quantitative essay?

Introduction

Literature Review

Research Question

(Hypothesis)

Presentation of Data

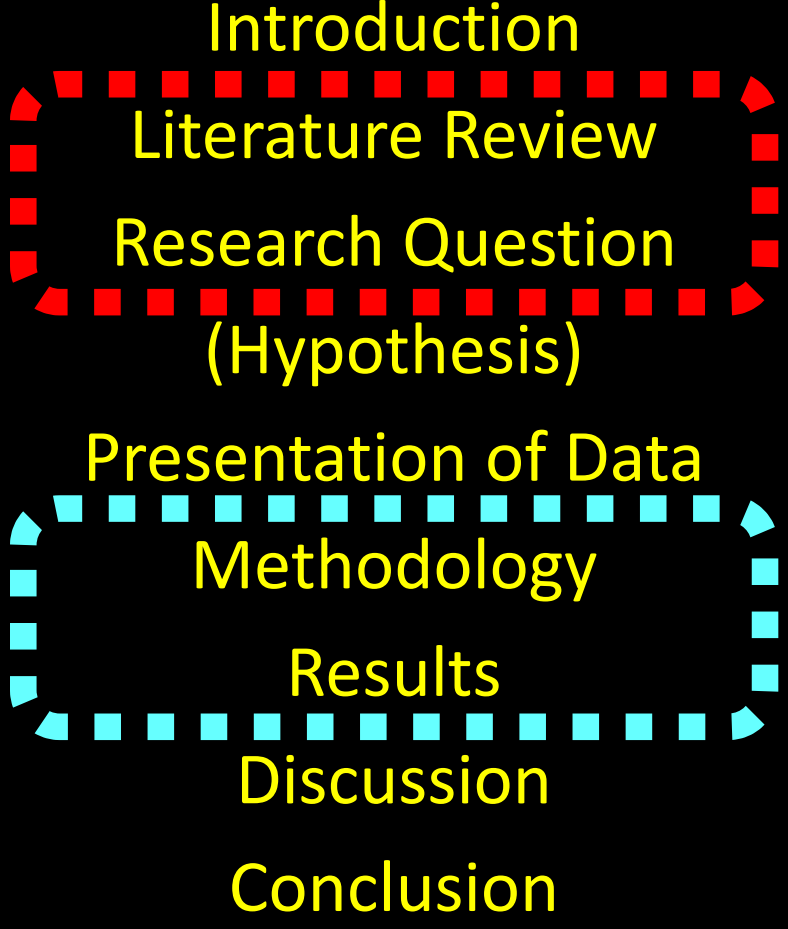
Methodology

Results

Discussion

Conclusion

What sections should be included in a quantitative essay?

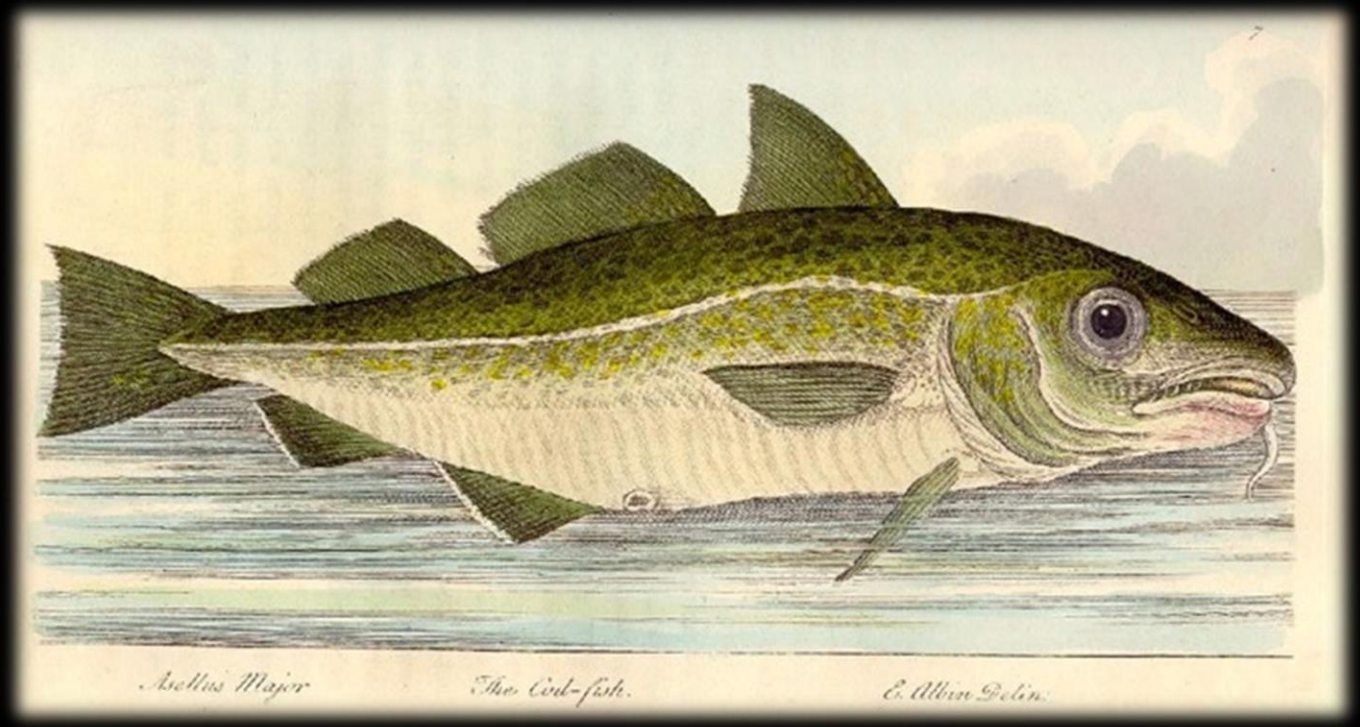


Good Practice

Avoiding common pitfalls

Good Practice

Avoiding common pitfalls



Good Practice

Introduction

Introduction

Humans have been fishing since the dawn of time. Issues surrounding fishing are gaining in importance.

I decided to do an investigation into cod fishing because I have always been amazed by the beauty of the cod. My mother worked in the cod fishing industry and my father was a cod liver oil salesman, so these personal links have also inspired this research.

Good Practice

Introduction

Avoid vague context.
Get to the point!

Introduction

Humans have been fishing since the dawn of time. Issues surrounding fishing are gaining in importance.

I decided to do an investigation into cod fishing because I have always been amazed by the beauty of the cod. My mother worked in the cod fishing industry and my father was a cod liver oil salesman, so these personal links have also inspired this research.

Good Practice

Introduction

Avoid vague context.
Get to the point!

Introduction

Humans have been fishing since the dawn of time. Issues surrounding fishing are gaining in importance.

I decided to do an investigation into cod fishing because I have always been amazed by the beauty of the cod. My mother worked in the cod fishing industry and my father was a cod liver oil salesman, so these personal links have also inspired this research.

Academic writing should be impersonal & objective.

Good Practice

Introduction

- Gives the reader a feel for your investigation.
- Begins to justify your investigation.

Good Practice

Literature Review

Literature Review

The cod is the most beautiful of all fish (welovecod.com), so if the world runs out of cod, this would be a great aesthetic loss. "The two most common species of cod are the Atlantic cod (*Gadus morhua*), which lives in the colder waters and deeper sea regions throughout the North Atlantic, and the Pacific cod (*Gadus macrocephalus*), found in both eastern and western regions of the northern Pacific. *Gadus morhua* was named by Linnaeus in 1758." (Wikipedia, 2019)

Good Practice

Literature Review

Take a critical approach!

Literature Review

The cod is the most beautiful of all fish (welovecod.com), so if the world runs out of cod, this would be a great aesthetic loss. "The two most common species of cod are the Atlantic cod (*Gadus morhua*), which lives in the colder waters and deeper sea regions throughout the North Atlantic, and the Pacific cod (*Gadus macrocephalus*), found in both eastern and western regions of the northern Pacific. *Gadus morhua* was named by Linnaeus in 1758." (Wikipedia, 2019)

Good Practice

Literature Review

Take a critical approach!

Literature Review

The cod is the most beautiful of all fish (welovecod.com), so if the world runs out of cod, this would be a great aesthetic loss. "The two most common species of cod are the Atlantic cod (*Gadus morhua*), which lives in the colder waters and deeper sea regions throughout the North Atlantic, and the Pacific cod (*Gadus macrocephalus*), found in both eastern and western regions of the northern Pacific. *Gadus morhua* was named by Linnaeus in 1758." (Wikipedia, 2019)

Keep quotes concise!

Good Practice

Literature Review

Take a critical approach!

Literature Review

The cod is the most beautiful of all fish (welovecod.com), so if the world runs out of cod, this would be a great aesthetic loss. "The two most common species of cod are the Atlantic cod (*Gadus morhua*), which lives in the colder waters and deeper sea regions throughout the North Atlantic, and the Pacific cod (*Gadus macrocephalus*), found in both eastern and western regions of the northern Pacific. *Gadus morhua* was named by Linnaeus in 1758." (Wikipedia, 2019)

Keep quotes concise!

How has this informed your approach?

Good Practice

Literature Review

Take a critical approach!

Literature Review

The cod is the most beautiful of all fish (welovecod.com), so if the world runs out of cod, this would be a great aesthetic loss. "The two most common species of cod are the Atlantic cod (*Gadus morhua*), which lives in the colder waters and deeper sea regions throughout the North Atlantic, and the Pacific cod (*Gadus macrocephalus*), found in both eastern and western regions of the northern Pacific. *Gadus morhua* was named by Linnaeus in 1758." (Wikipedia, 2019)

Keep quotes concise!

How has this informed your approach?

Don't cite Wikipedia!

Good Practice

Literature Review

Take a critical approach!

Literature Review

The cod is the most beautiful of all fish (welovecod.com), so if the world runs out of cod, this would be a great aesthetic loss. "The two most common species of cod are the Atlantic cod (*Gadus morhua*), which lives in the colder waters and deeper sea regions throughout the North Atlantic, and the Pacific cod (*Gadus macrocephalus*), found in both eastern and western regions of the northern Pacific. *Gadus morhua* was named by Linnaeus in 1758." (Wikipedia, 2019)

Keep quotes concise!

How has this informed your approach?

Are these references in Harvard format?

Don't cite Wikipedia!

Good Practice

Literature Review

- Sources for context vs. sources for methods
- Is all the literature relevant to your investigation?
- How has your reading informed your methodology?
- Demonstrates the importance of your investigation
- Takes a critical perspective
- (Engaging narrative)

Good Practice

Research Question

- Explicit
- Precise
- Real world relevance
- Informed by literature review
- Possibly split into sub-questions

Good Practice

Presentation of Data

- Explains and visualises your data for the reader

Good Practice

Methodology

Methodology & Results

To discover how many cod there are in the North Sea, I will use Fermi estimation, making appropriate assumptions. I will use the geometric mean and the formula for the volume of a cuboid.

I did this and the answer was 74.

Good Practice

Methodology

Could your work be reproduced by another researcher?

Methodology & Results

To discover how many cod there are in the North Sea, I will use Fermi estimation, making appropriate assumptions. I will use the geometric mean and the formula for the volume of a cuboid.

I did this and the answer was 74.

How can we have confidence that this is not just a guess?

Good Practice

Methodology

- Precise
- Reproducible
- Explicitly informed by literature review
- Explicitly designed to answer **research question**

Good Practice

Results

- Clear
- Complete
- Minimal Text
- Presented in most informative way possible

Good Practice

Discussion

- Highlights key points from results
  - Interprets and synthesises
  - Acknowledges limitations
- Relates back to **research question**

Good Practice

Conclusion

Conclusion

The question of how many cod there are in the North Sea is clearly an important one. There is much more to be done than could have been included in this piece of work. Future research could consider how many cod there are in other seas or how many salmon there are in the North Sea.

Good Practice

Conclusion

Conclusion

The question of how many cod there are in the North Sea is clearly an important one. There is much more to be done than could have been included in this piece of work. Future research could consider how many cod there are in other seas or how many salmon there are in the North Sea.

Where is the answer to
the research question ???

Good Practice

Conclusion

- Explicitly answers **research question** (as far as possible)
- Proposes further research

Good Practice

Overall Style

- Impersonal
  Use “we” or the passive (“A simulation was run.”)
  Never use “I”

- Not narrative
  Your Goal – Communicate your results
  Not your goal – Describe your every thought and action in chronological order

What sections should be included in a quantitative essay?

Introduction

Literature Review

Research Question

(Hypothesis)

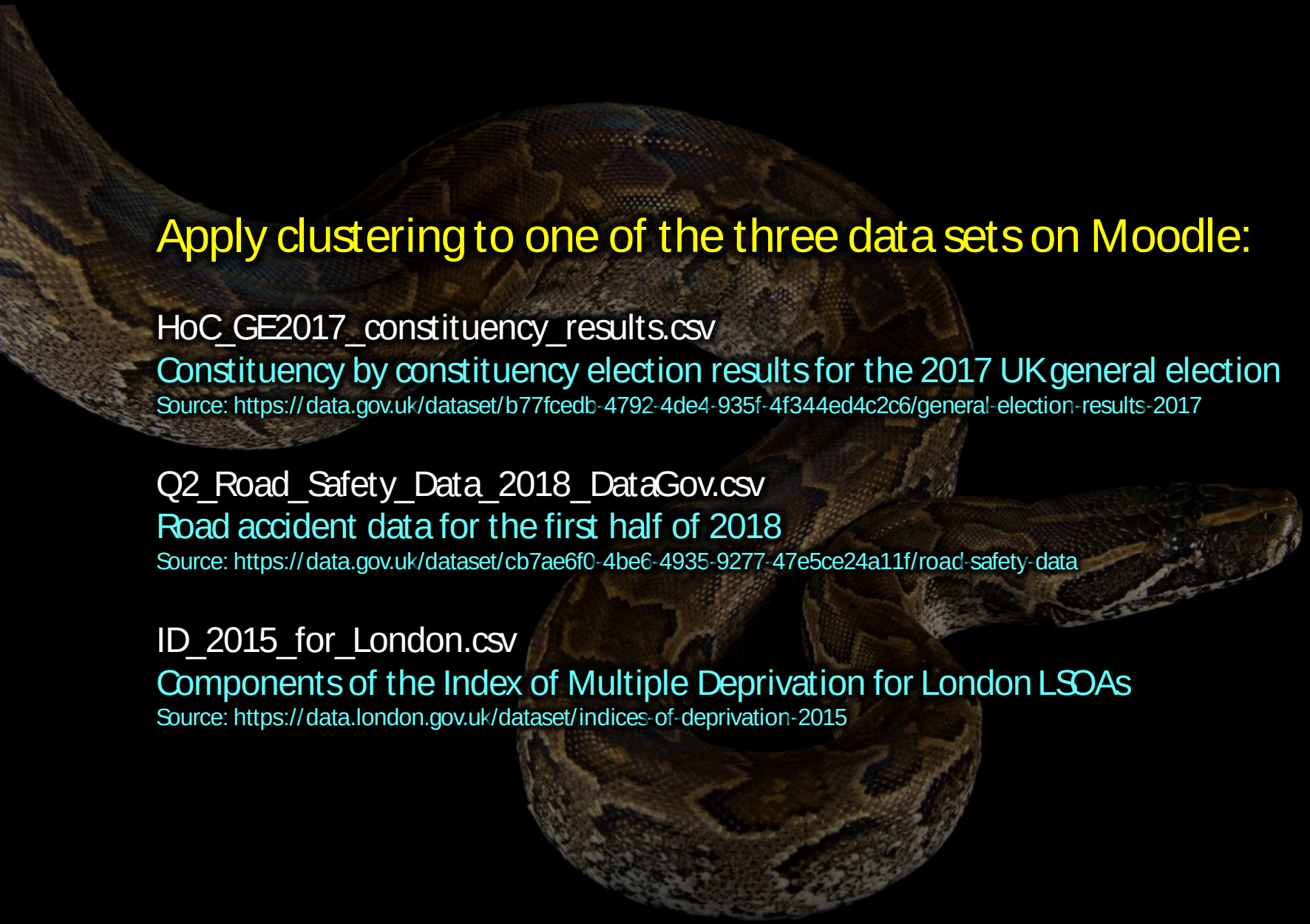
Presentation of Data

Methodology

Results

Discussion

Conclusion



Apply clustering to one of the three data sets on Moodle:

HoC_GE2017_constituency_results.csv

Constituency by constituency election results for the 2017 UK general election

Source: https://data.gov.uk/dataset/b77fcedb-4792-4de4-935f-4f344ed4c2c6/general-election-results-2017

Q2_Road_Safety_Data_2018_DataGov.csv

Road accident data for the first half of 2018

Source: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

ID_2015_for_London.csv

Components of the Index of Multiple Deprivation for London LSOAs

Source: https://data.london.gov.uk/dataset/indices-of-deprivation-2015