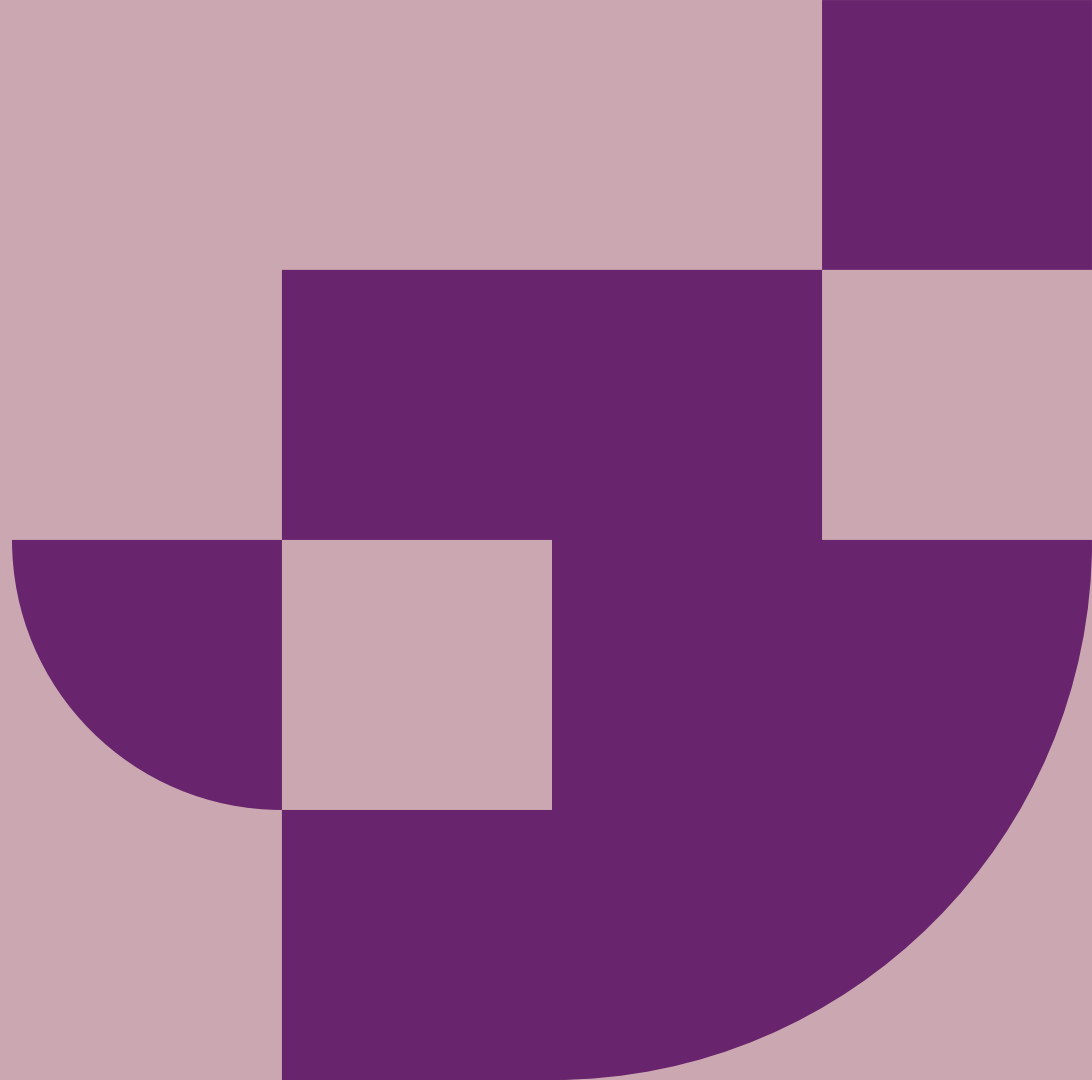




Durham  
University

# Case Based Complexity

Computational Social Science  
Lecture 2



# Overview

Big data is not just about amount of data, there is difference in kind

Data mining methods are big data algorithms to analyse cases

- Look for patterns of similarity and difference

Algorithms conceptualise cases as points in multidimensional space, so similar cases are close to each other

Data mining methods create clusters of cases that are similar to each other and different from other cases

- Introduce k-means clustering, hierarchical clustering, self organised maps
- Introduce tools that can interpret clusters

# Big Data



# What is Big Data?

Original definitions emphasised 3Vs

- Volume
- Velocity
- Variety

Data with these characteristics could not be managed with existing technology (storage, relational datasets)

Big data refers to large amounts of data produced very quickly by a high number of diverse sources. Data can either be created by people or generated by machines, such as sensors gathering climate information, satellite imagery, digital pictures and videos, purchase transaction records, GPS signals, and more. It covers many sectors, from healthcare to transport to energy.

# Volume is large

Volume is simply the size of the data

For big data, the data exceeds the storage of individual computers

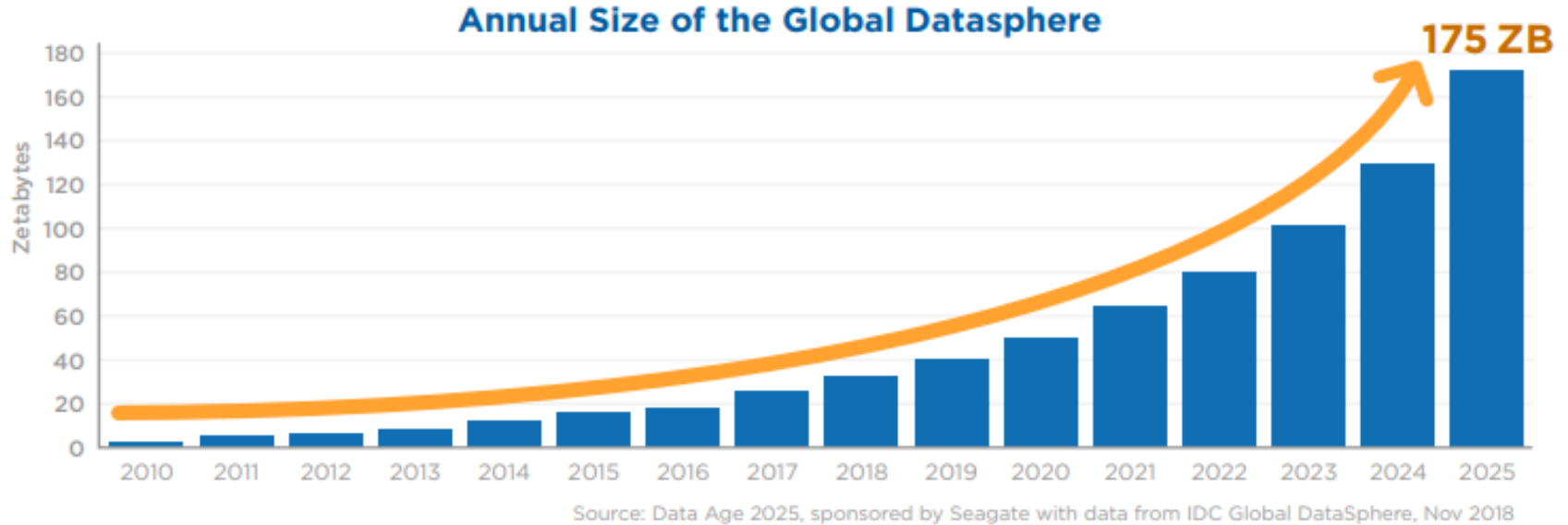
1.2 ZB internet traffic in 2016 (Cisco)

Barnett, The Zettabyte Era Officially Begins  
<https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-that>

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies

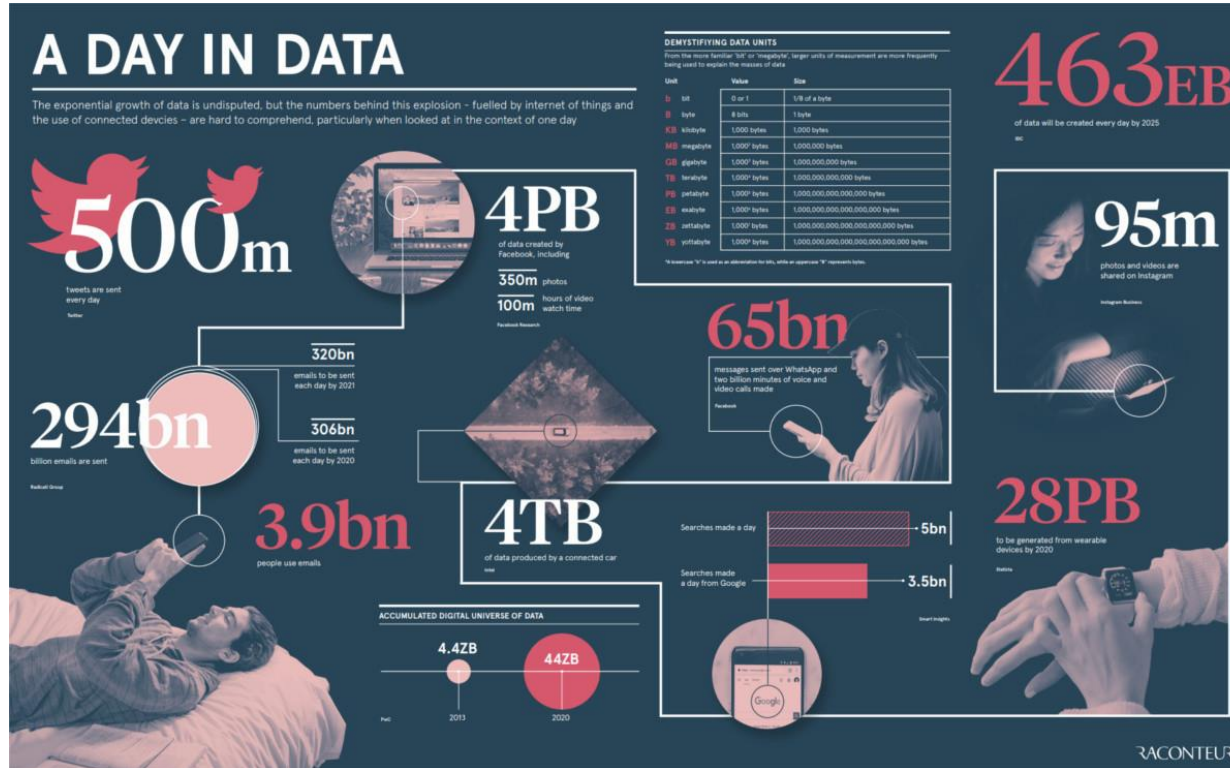
NASA, Data volume units  
<https://mynasadata.larc.nasa.gov/basic-page/data-volume-units>  
From Roy Williams, Powers of 10

# Volume is growing



Reinsel, Gantz, Rydning (2018), *Data Age 2025: The Digitization of the World*, IDC White Paper – #US44413318  
<https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>

# Volume driven by social media, video and digital devices



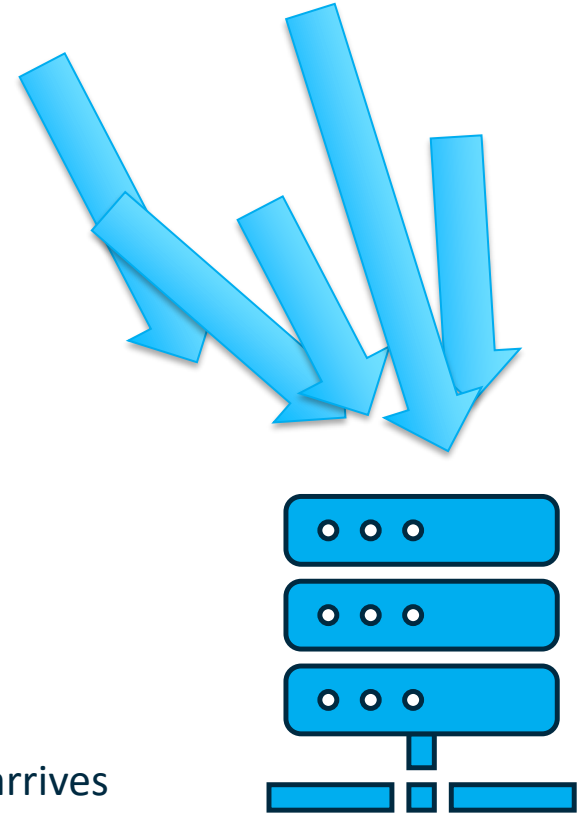
# Velocity

Data streams generated and arriving at high speed

- Real time sensor data
- Transactional data such as financial data

Data must be processed at speed

- Decisions
  - Importance of being first for financial decisions
  - Data protection and security
  - Self driving cars
- On average, must be processed at least as quickly as arrives

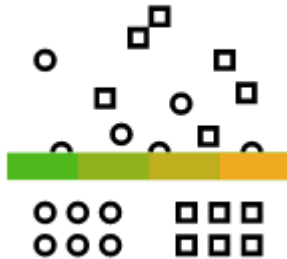




# Variety

## Different types of data

- Structured datasets, easy to store and process
- Unstructured data such as text and video
- Geospatial
- Sensor traces



SAP Insights, *What is Big Data?*

<https://www.sap.com/uk/insights/what-is-big-data.html>

# Other challenges of big data

## More Vs and other letters

Veracity: unstructured data has different accuracy issues

- Fake news

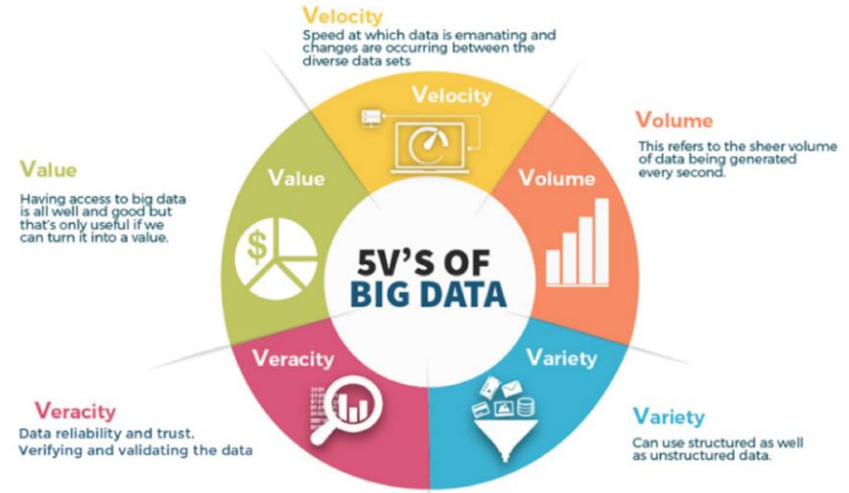
Variability: inconsistencies in the flow (volume, velocity) can crash infrastructure

- Searches when news item

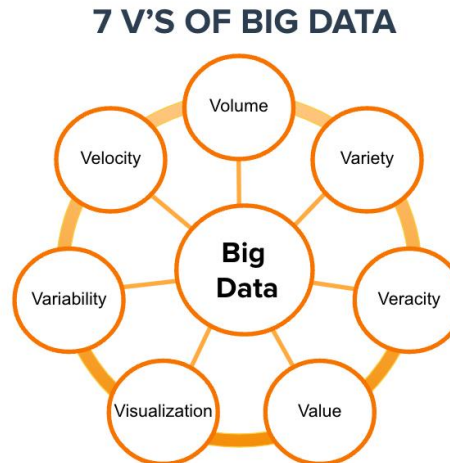
Value: is it useful?

Vulnerability: how to keep data secure?

Visualisation: how to understand big data?



<https://sis.binus.ac.id/2020/09/28/karakteristik-big-data/>



<https://impact.com/marketing-intelligence/7-vs-big-data/>

# Other aspects

No agreed definition or conceptual clarity

Description of society

Buzzword for funding and marketing

Defined by methods required to analyse and interpret the data

## PLOS ONE


OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

### What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade

Maddalena Favaretto, Eva De Clercq, Christophe Olivier Schneble, Bernice Simone Elger

Published: February 25, 2020 • <https://doi.org/10.1371/journal.pone.0228987>

Article	Authors	Metrics	Comments	Media Coverage
				
Abstract	Abstract			
Introduction	The term Big Data is commonly used to describe a range of different concepts: from the collection and aggregation of vast amounts of data, to a plethora of advanced digital techniques designed to reveal patterns related to human behavior. In spite of its widespread use, the term is still loaded with conceptual vagueness. The aim of this study is to examine the understanding of the meaning of Big Data from the perspectives of researchers in the fields of psychology and sociology in order to examine whether researchers consider currently existing definitions to be adequate and investigate if a standard discipline centric definition is possible.			
Methods				
Results				
Discussion				
Limitations				
Conclusions				

<https://doi.org/10.1371/journal.pone.0228987>

# Digital sociology

Broader than new methods to research old questions

Interplay between people, ideas and technology

- Samaritan app intended to identify those at suicide risk from tweets -> concern about stigma and self-censoring

New questions about social phenomena

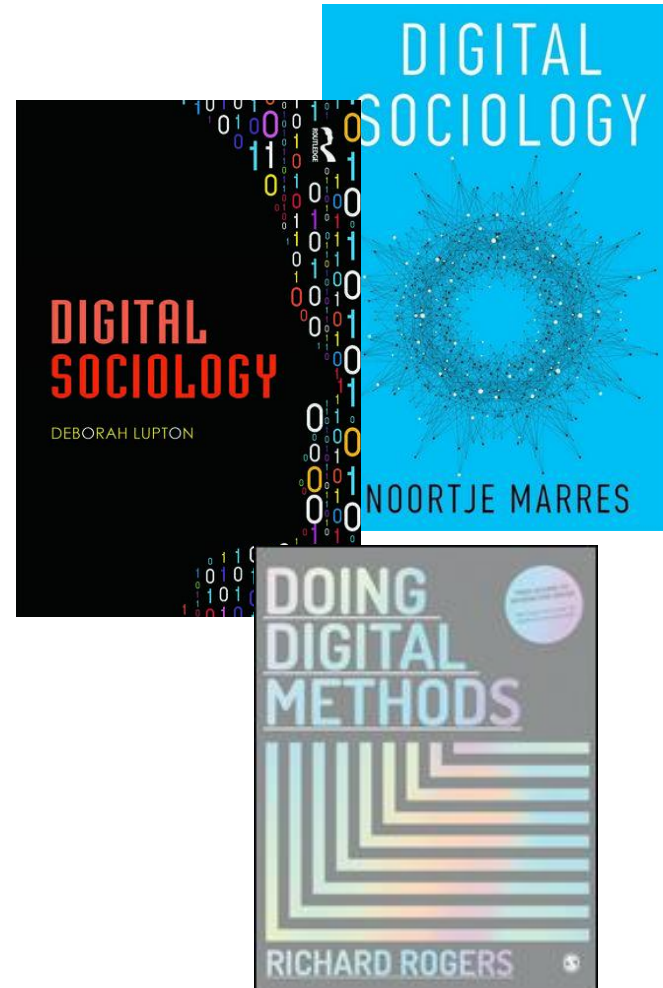
- Digital activism contributes to civil resistance but also exposes activists to surveillance

New questions about methods

- Social media and by-product data has no real consent

If social scientists are to make sense of digital saturated world

- Embrace new and different methods and methodologies
- Critically advance digital methods



# Machine Learning



# What is artificial intelligence (AI)?

Developing computers (including robots) that can:

- Emulate human thought
- Deal with real world

Turing test: can a computer fool a human (written) interrogator into thinking it is human?

Outcome is measured, not whether it thinks like a human

- Aeroplanes and birds both use wings to fly, but in different ways



# What is machine learning (ML)?

Machines don't 'learn': An analyst gives the computer a rule to follow about how to solve a problem

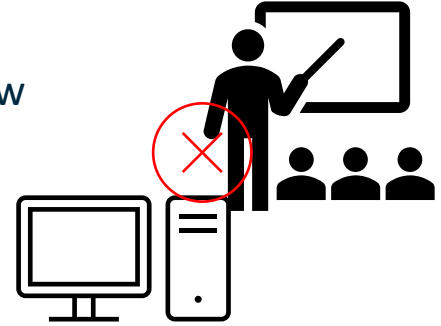
Problems in ML are about finding patterns in data

- ML is a path to artificial intelligence

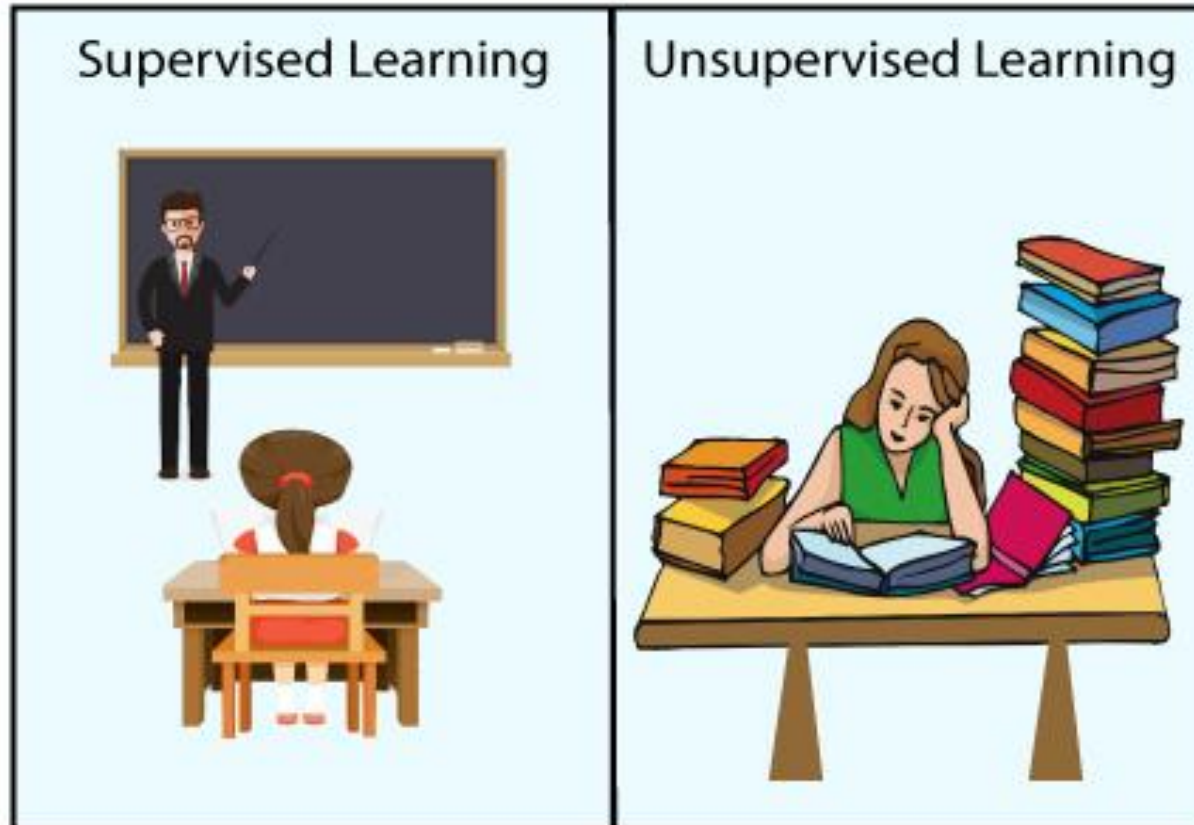
The discipline of ML is about

- Developing algorithms to detect new types of patterns
- Developing new algorithms to better detect patterns
- Making existing algorithms faster or otherwise better

Social science uses those algorithms to understand social phenomena

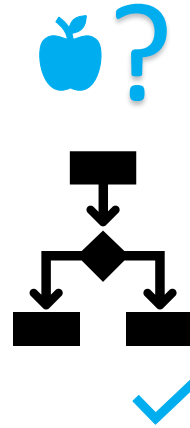
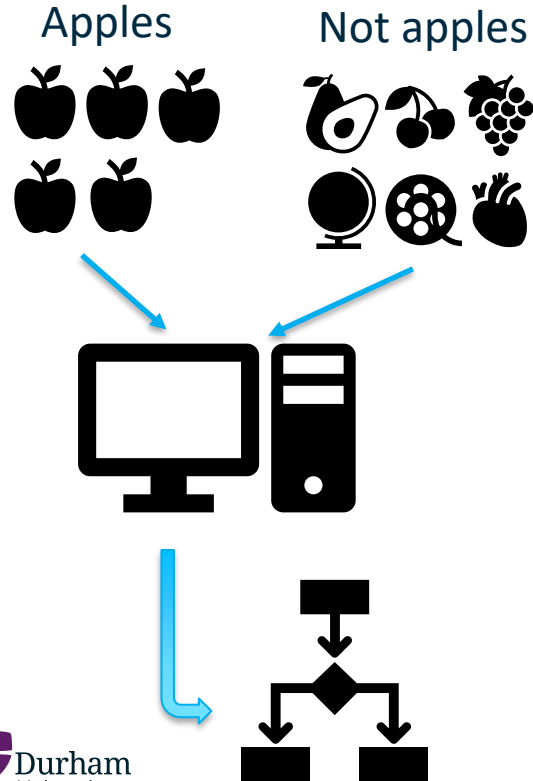


# Two broad types of machine learning algorithms





# Supervised learning



Training data to the algorithm

- Each input is labelled (ground truth)
  - Is or not an apple

Algorithm develops a rule for how to classify

Rule is used on new input to predict class or value

# Supervised learning

Two types of output

- Classification: predict label
- Regression: predict value

Examples

- Spam detection
- Image recognition
- Sentiment analysis
- Credit risk

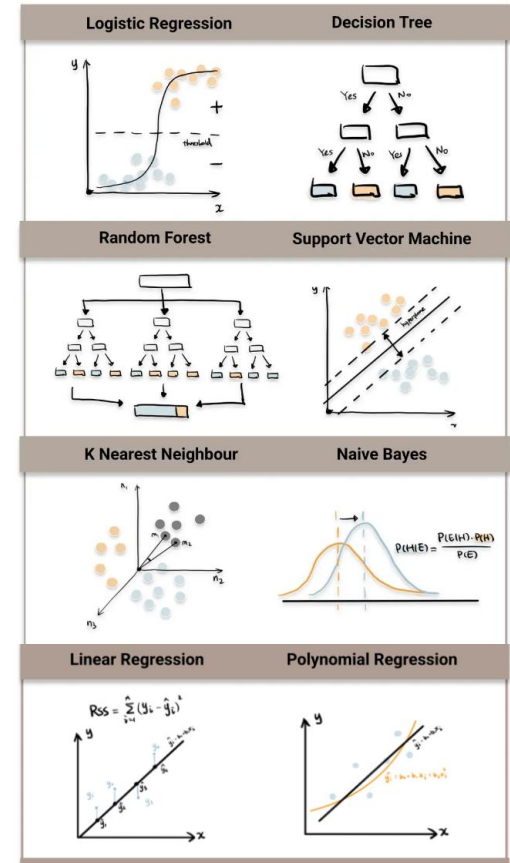
Supervised learning algorithms maximise predictive accuracy

Developed rules

- Ignore input similarity
- Ignore input meaning
- May not be interpretable, human cannot understand how the prediction is made

Not suitable for our purposes

- Complexity stripped away



# Unsupervised learning

When applied, two older terms are used

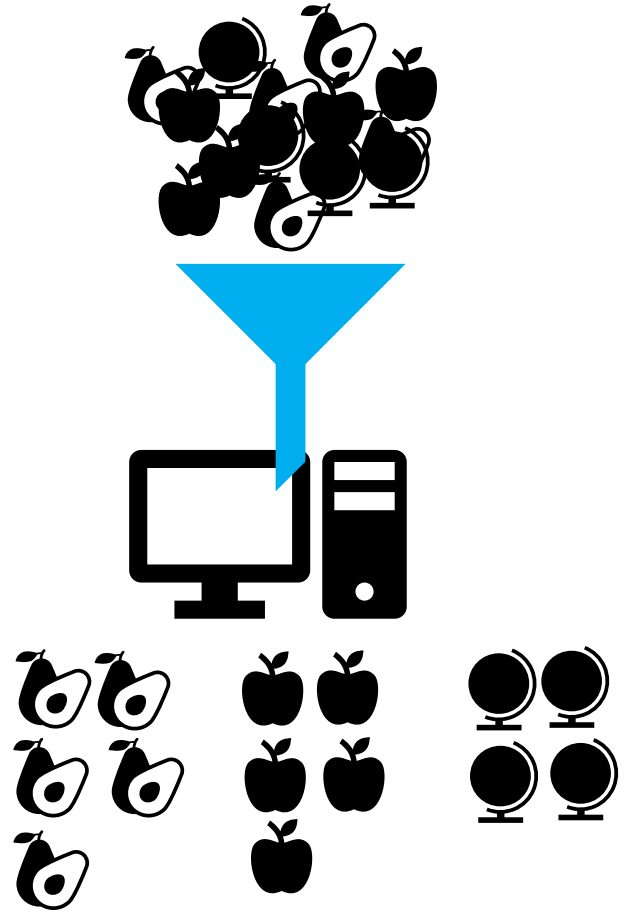
- Data Mining
- Knowledge Discovery [in Databases, hence KDD]

Unsupervised learning is exploratory

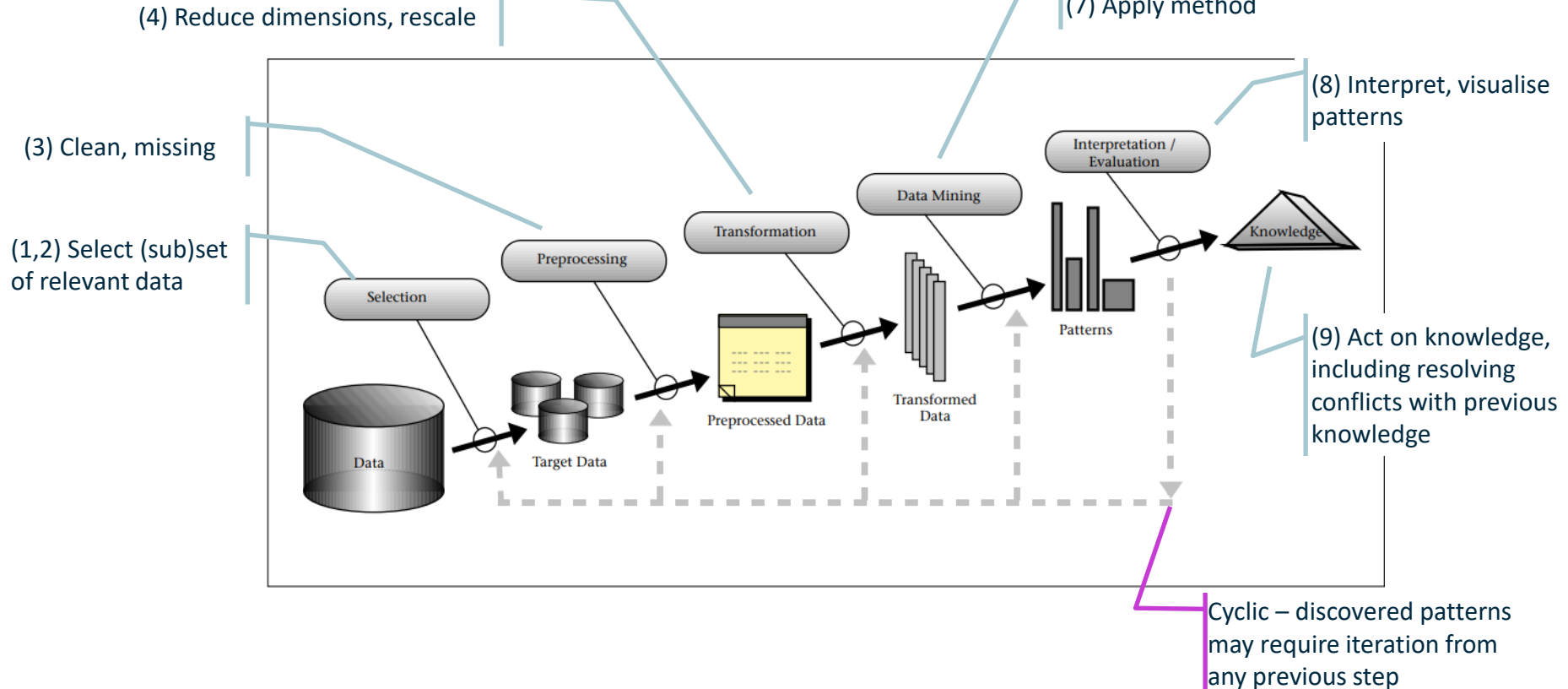
- Finds patterns
- Many algorithms do not require training

Patterns illuminate how a complex social phenomenon is organised

- Case based complexity
- Gain insight into the phenomenon



# General approach to KDD



# KDD is explicitly exploratory and iterative

Interplay between human interpreter and computer algorithms

Data mining is the pattern extraction step made by the computer in KDD process

Human role is twofold

- Prepare the data so of sufficient quality
- Decide if pattern is knowledge

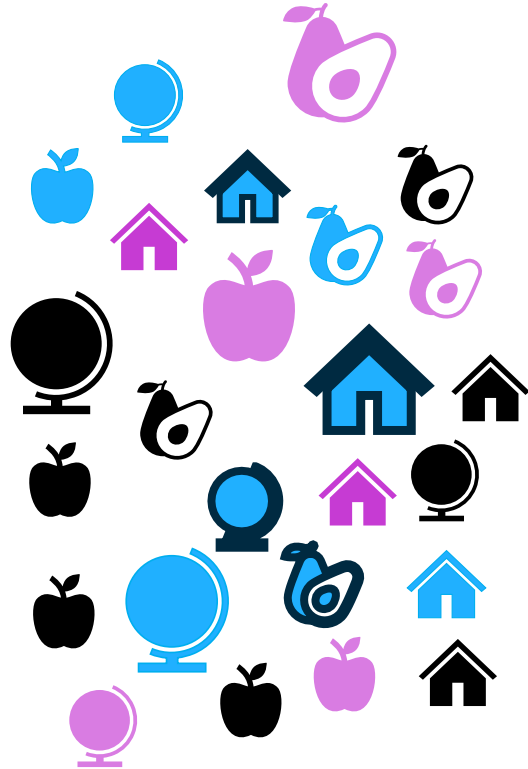
Pattern is 'knowledge' if it is sufficiently interesting

- Valid, novel, simple, useful

*Blind application of data-mining methods can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.*

Fayyad, Piatetsky-Shapiro, Smyth (1996)

# How would you group these items?



How many groups?

What attributes?

What if you were forced to  
organise the items into 3 groups?

How are you deciding similarity?

- Number of identical attributes
- Ranked importance

Colour

- Number of colours
- Main colour

Type

- Fruit or inanimate
- Specific item (four different)

Size

# Case Based Complexity

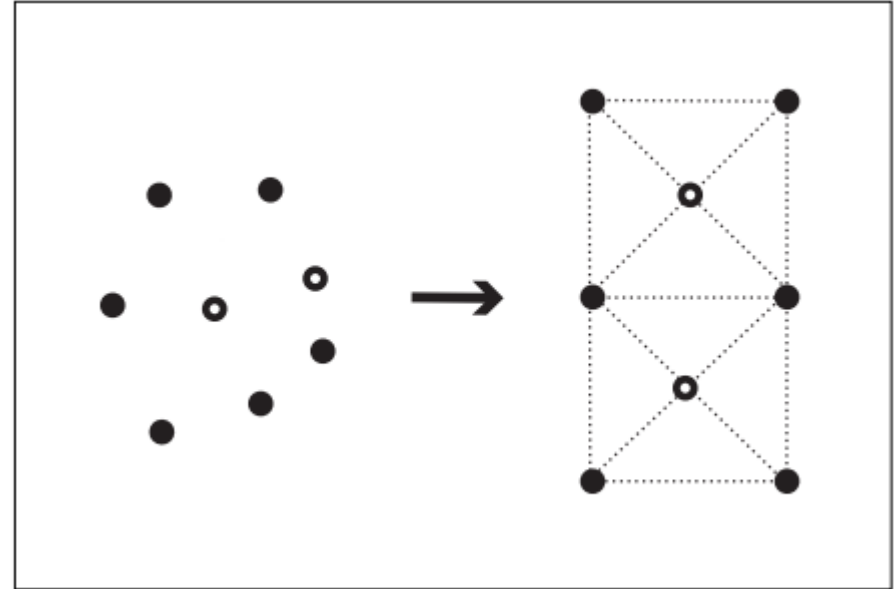
Each case is being considered as a whole

- All the attributes are involved in the consideration of similarity

Each case is treated as a complex system

- The parts (attributes) are not separated from the whole (case) as variables that operate separately

Potential for similarity with other cases is an emergent property of each case



# Qualitative and quantitative attributes

Some qualitative attributes do support similarity assessment

- Human and Chimpanzee are more similar than Human and Tree, and
- Human and Tree are more similar than Human and Sun

But that relies on an implicit distance measure from a taxonomy of life

Nominal variables are defined as being non-comparable and unordered

- Eye colour
- Nationality
- Mode of transport

Will take that up with QCA later, and instead cluster over only quantitative attributes



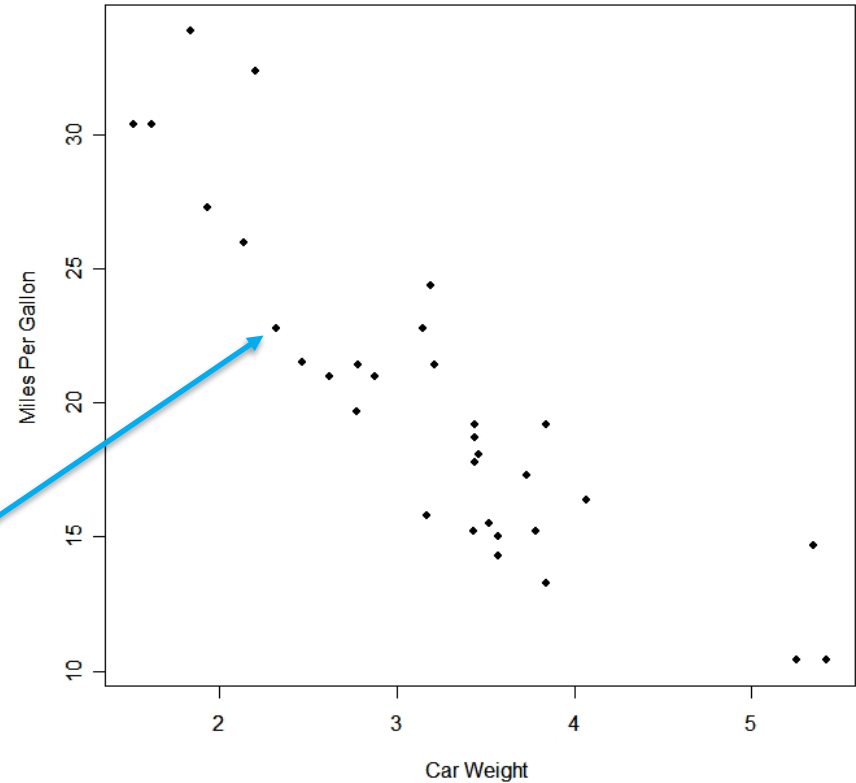
# Conceptualising a case in multidimensional space

If there are only two attributes

- Position on 2D coordinate system describes the whole case

Example - R built-in dataset: cars

- Each car has two attributes, car weight (in 1000 lbs) and fuel economy (in miles per US gallon)
- Datsun 710 attribute values are 2.320 weight and 22.8 fuel economy



# Mathematical formulation

Assume there are  $n$  cases, each with  $k$  attributes (features in ML literature)

- Attribute values are notated  $x_{ij}$
- With  $i \in \{1, 2 \dots n\}$  identifies case
- With  $j \in \{1, 2 \dots k\}$  identifies attribute

So the case profile of the  $i$ th case  $c_i$  is given by the row vector

$$c_i = (x_{i1}, x_{i2}, \dots, x_{ik})$$

The database  $D$  of all cases is given by the matrix (or array)

$$D = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

Actually, these are all dependent on time, so really  $D(t)$  and  $c_i(t)$

# Case conceptualisation includes values of all attributes

Different from variable orientation

Considering *all* attributes simultaneously

Case →

	Income	Employment	Health						
Area Code	People in income deprivation (%)	Working-age people in employment deprivation (%)	GP-recorded chronic condition (rate per 100)	Limiting long-term illness (rate per 100)	Premature death (rate per 100,000)	GP-recorded mental health condition (rate per 100)	Cancer incidence (rate per 100,000)	Low birth weight (live single births less than 2.5kg) (%)	Children aged 4-5 who are obese (%)
Isle of Anglesey	15	10	13.4	20.8	359.6	23.0	601.1	5.5	12.7
Gwynedd	15	8	12.9	19.5	348.9	20.3	604.0	5.0	13.0
Conwy	15	10	12.9	20.6	375.6	23.7	593.6	5.1	11.4
Denbighshire	17	11	14.7	21.8	397.4	28.4	639.3	6.1	12.2
Flintshire	12	8	14.1	19.7	358.1	23.1	647.8	5.4	11.2
Wrexham	15	9	14.3	21.5	393.7	24.3	637.3	6.4	12.4
Powys	11	7	12.8	18.8	309.1	19.0	579.8	4.7	10.5
Ceredigion	12	8	12.7	20.0	322.4	19.9	545.5	4.8	10.5
Pembrokeshire	15	10	13.1	20.5	345.8	22.1	606.1	5.2	12.5
Carmarthenshire	15	11	13.9	23.7	365.5	20.0	602.6	5.4	12.8

← Case Profile

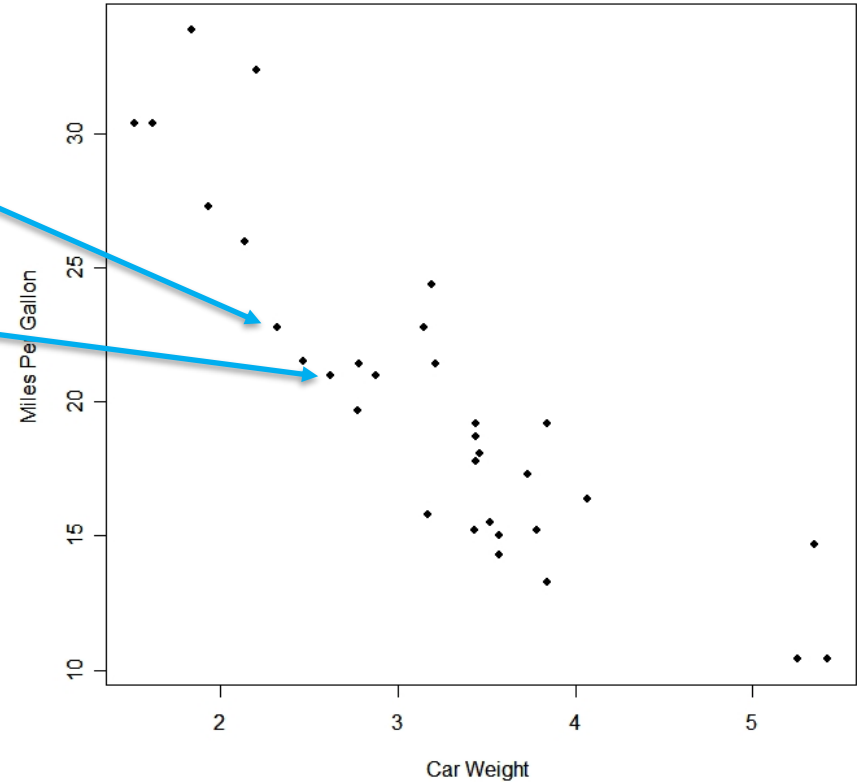
# Similarity in 2-dimensional space

Consider two cases as if they only had two attributes

- Datsun 710: attribute values are 2.320 weight and 22.8 fuel economy
- Mazda RX4: attribute values are 2.620 weight and 21.0 fuel economy

What does similarity mean?

- These two cars are near each other
- Indicates similar attribute values
- Distance is a measure of similarity



# Distance measures in 2-dimensional space

Manhattan: add the differences

- $(2.62 - 2.32) + (22.8 - 21.0)$

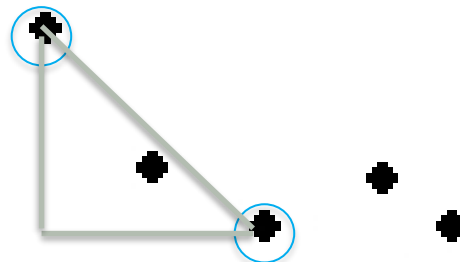
Euclidean: straight line

- $\sqrt{(2.62 - 2.32)^2 + (22.8 - 21.0)^2}$

Or create your own if that's meaningful

- Largest difference: 1.8
- Smallest difference: 0.3
- Number of attributes within 10%
- Weighted with importance

Datsun 710: 2.320 and 22.8



Mazda RX4: 2.620 and 21.0

# Additional attributes require extension to distance metric

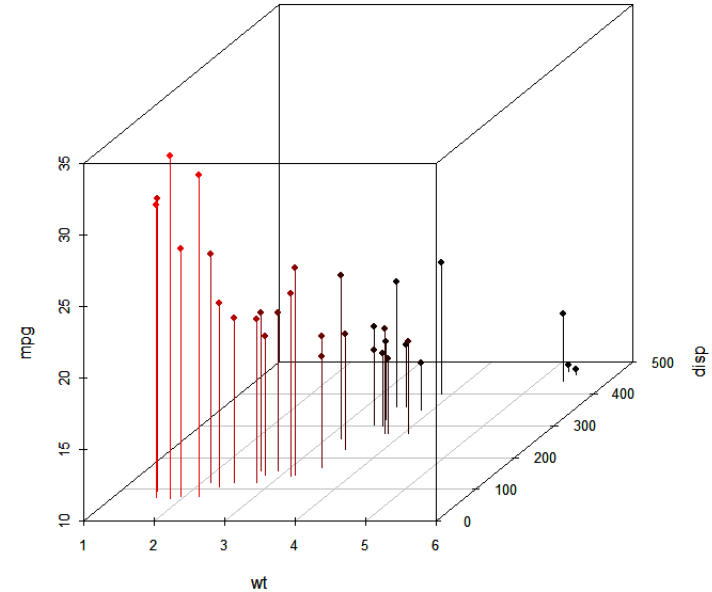
Cases located in n-dimensional space, where n is the number of attributes

- Distance measured in n dimensions

Euclidean distance:  $\sqrt{\sum_{j=1}^n (A_j - B_j)^2}$

Where

- A and B are cases
- index  $j$  refers to a specific attribute



Same set of cars, but now have three attributes, also engine size (displacement)

# Conceptualising a case in multidimensional space: key ideas

Case is conceptually located in multidimensional space

- One dimension for each (quantitative) attribute
- Coordinate is determined by the value of the relevant attribute

Similarity of cases is equivalent to “close to” in the multidimensional space

Similarity of cases is with respect to a specific way of measuring distance

- Euclidean distance is common distance metric for quantitative attributes

# Cases are dynamic (attribute values change over time)

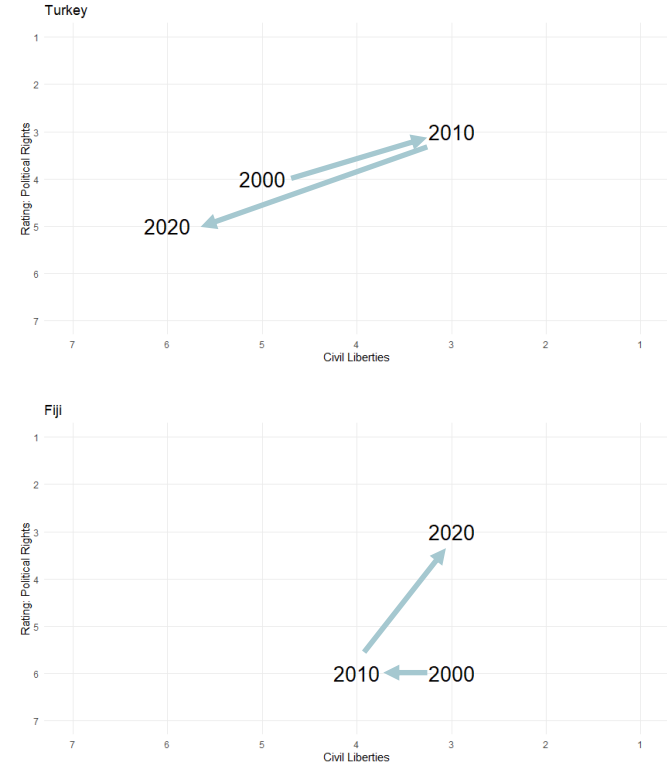
As attribute values change, the case 'moves' in multidimensional space

Trajectories can be similar

- Start points near each other and end points near each other
- Similar directions and distance moved

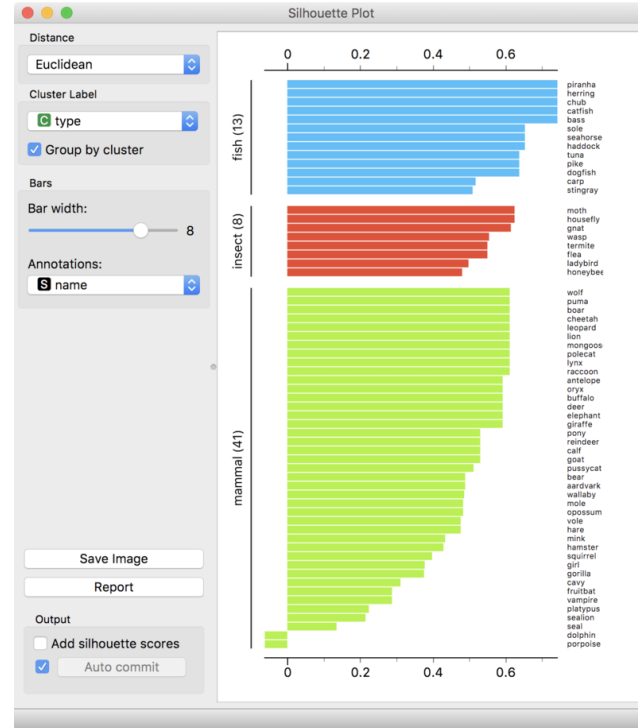
Cases that are similar at one time point may be dissimilar at another time point

- Trajectories are different
- Complex systems are path dependent





# Clustering Algorithms and Examples



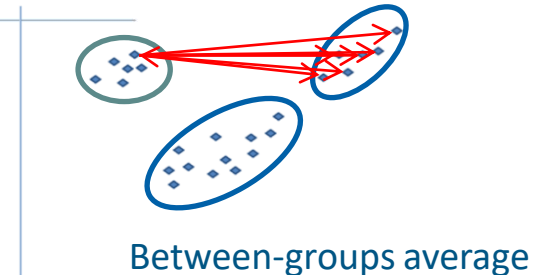
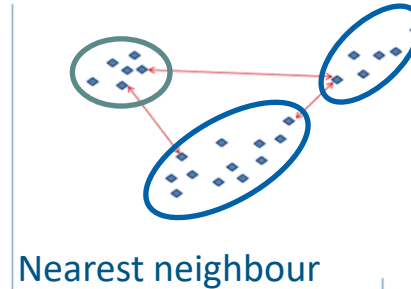
# Objective of clustering

Goal is to allocate cases to clusters

- Cases within a cluster are close to each other (similar)
- Clusters are far from each other (different)

Both use same distance metric

Many cases in each cluster so can measure difference between clusters in different ways



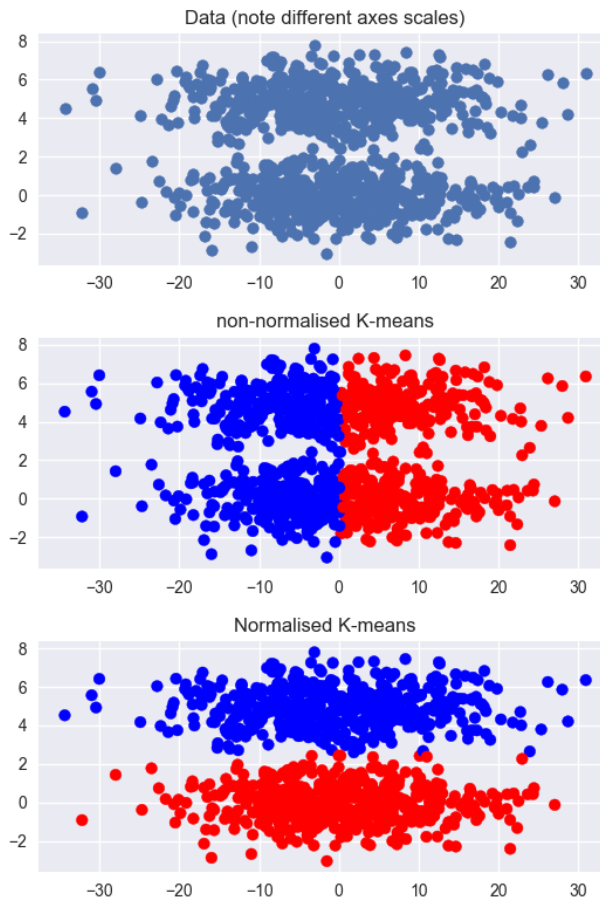
# Preparing data for clustering

Remove cases with missing values

- Case is the meaningful unit
- Case is all of the attribute values

Consider whether attribute values should be scaled to a common range

- Distance metrics are affected by scales
- If all attribute values are in same range (such as  $[0:1]$ , or to mean 0 and SD 1) then they have similar weight in the distance calculation



# Clustering with *k-means*

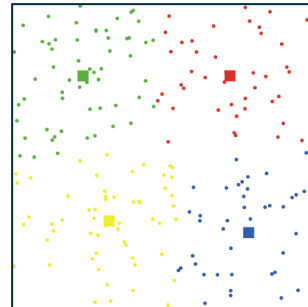
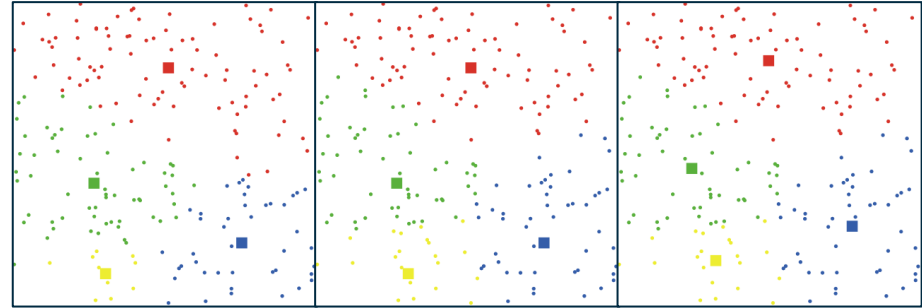
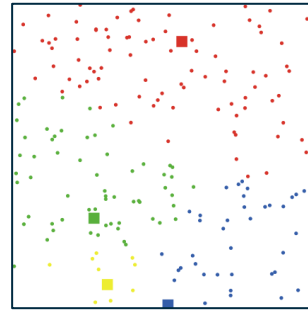
Goal is to group cases for closeness within group and farness between groups

- Choose number of clusters ( $k$ )
- Randomly select  $k$  cases and set centroids in those locations
- Assign each case to closest centroid
- Calculate new centroid position as mean of positions of cases in the cluster

Iterate until centroids stable

Note: random start so try multiple choices

- Can use ensemble



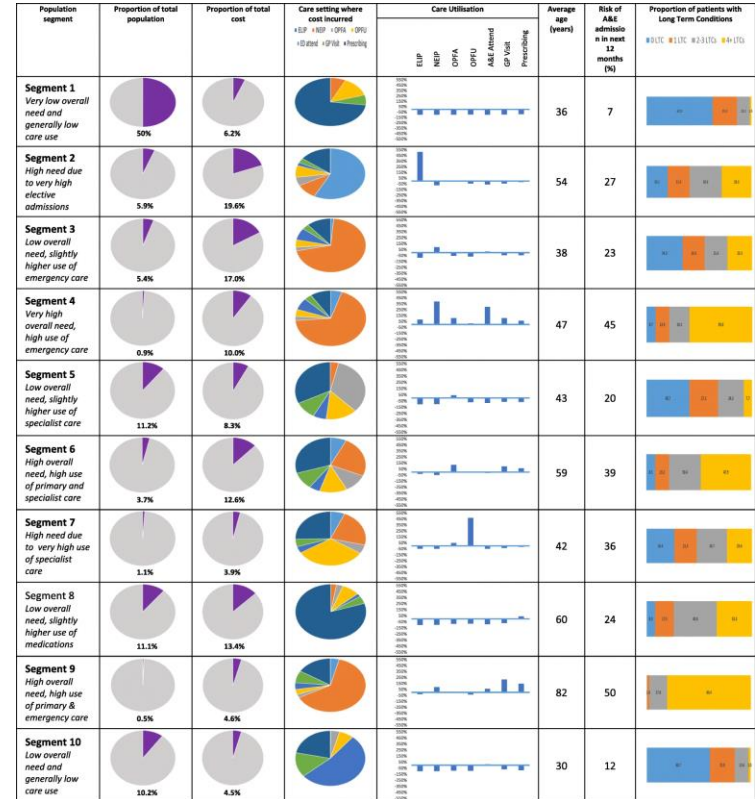
# k-means, population segmentation on health service use

Linked database of primary and secondary care

79,607 patients with 7 variables about healthcare utilisation

Found 10 clusters with homogenous needs

- Potential for tailored programmes
- Prevention (low need) or active monitoring (high need)



# Hierarchical clustering: Agglomerative

Bottom up (agglomerative) approach

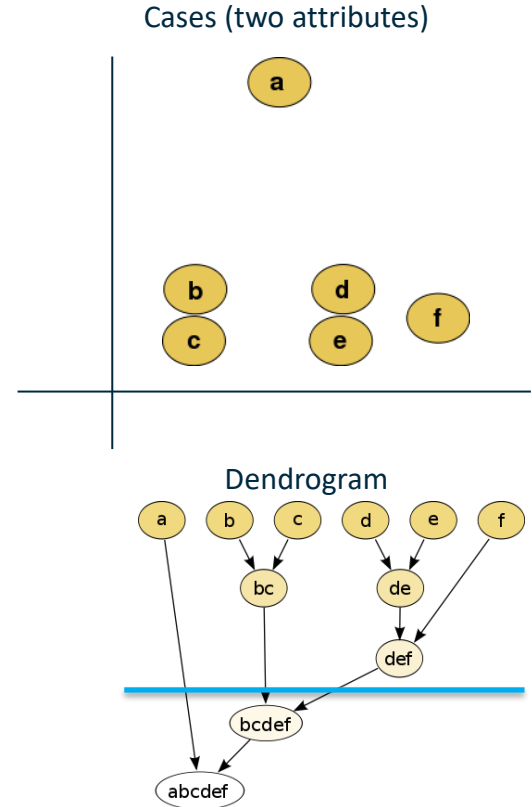
Initialise with each case assigned to its own cluster

Iteratively, until all in a single cluster

- Find the pair of clusters that are closest together
- Combine that pair

Cut the dendrogram at height that gives good outcome

- Cutting at blue line would give three clusters: {a} and {b,c} and {d,e,f}



# Hierarchical clustering, UK universities



## Questions

- Are there Distinctive Clusters of Higher and Lower Status Universities in the UK?
- If there are clusters, what differentiates them?



## Research activity

- Adjusted research income
- % postgraduates
- REF 2008

## Teaching environment

- % Satisfaction with teaching
- % satisfied with feedback
- 'Value-added' score

## Economic resources

- Investment income
- % spend on academic services
- Student-staff ratio

## Academic selectivity

- Average UCAS points on entry
- % degree completion
- % 'good degree'

## Social mix

- % Undergraduates not from low participation neighbourhoods
- % from NS-SEC 1-3
- % from private schools

# Hierarchical clustering, UK universities

Found 4 clusters with different attribute profiles

- Oxford and Cambridge separate cluster
- Other Russell Group not distinct from other old universities

	Oxbridge	Most other Old	Mainly New	Struggling New
	Cluster 1(N=2)	Cluster 2(N=39)	Cluster 3(N=67)	Cluster 4(N=19)
<b>Research activity</b>				
Research income adjusted	37,100 (6,131)	21,356 (6,128)	5,740 (4,727)	1,620 (1,617)
% postgraduates	37.0 (3.9)	31.2 (8.0)	20.0 (7.7)	20.9 (11.7)
RAE score in 2008	3.0 (0.0)	2.9 (0.3)	2.0 (0.2)	1.7 (0.4)
<b>Teaching quality</b>				
% students satisfied with teaching	92.5 (0.7)	88.7 (1.9)	84.5 (3.4)	84.1 (2.9)
% students satisfied with feedback	73.0 (2.8)	67.9 (4.4)	69.4 (4.2)	69.5 (3.6)
<i>Guardian</i> value-added score out of 10	6.5 (0.7)	6.0 (0.7)	5.5 (1.1)	4.1 (1.4)
<b>Economic resources</b>				
Endowment/investment income (£000s)	23,871 (5,481)	4,266 (4,345)	687 (555)	392 (340)
Academic services spending per capita	2,812 (384)	1,514 (331)	1,055 (289)	724 (379)
Student-staff ratio	11.3 (0.4)	14.4 (1.8)	18.8 (2.2)	22.6 (3.9)
<b>Academic selectivity</b>				
Average UCAS points on entry	595 (18)	442 (47)	308 (33)	251 (28)
% students completing their degree	98.7 (0.4)	92.2 (4.5)	82.9 (5.0)	78.5 (7.0)
% students achieving a "good degree"	89.3 (3.5)	78.2 (5.0)	55.0 (5.5)	63.4 (5.6)
<b>Socioeconomic student mix</b>				
% students not from low participation neighbourhoods	96.6 (0.3)	93.2 (2.8)	87.4 (5.1)	82.3 (6.2)
% students from more advantaged social class backgrounds	89.4 (1.5)	77.4 (5.0)	61.6 (6.7)	56.4 (5.5)
% students from private schools	34.9 (2.9)	16.1 (8.2)	3.6 (3.0)	1.4 (0.9)



# Hierarchical clustering paths, mental health and offending

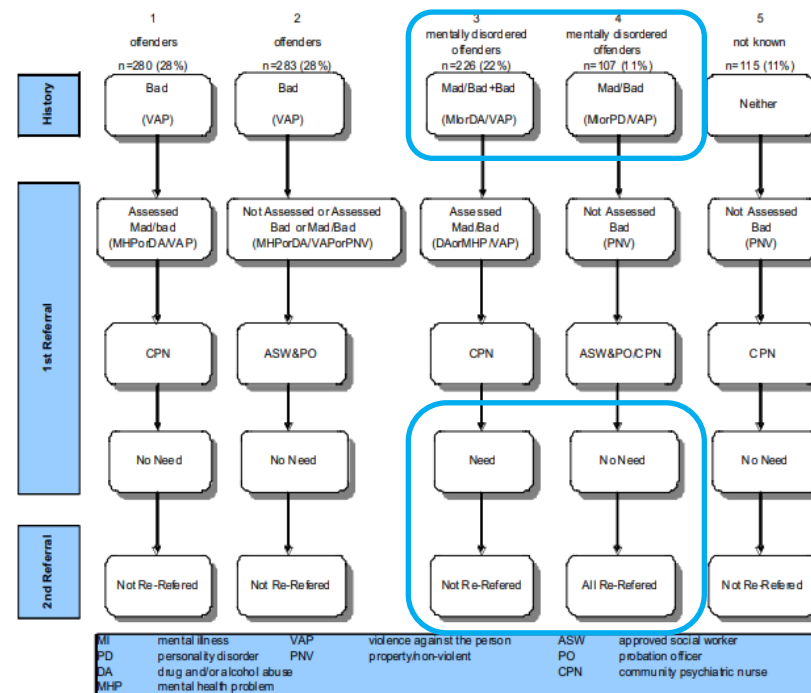
Three time points about same 1011 offenders referred to NE custody diversion team over 2.5 years

- Criminal and psychiatric history
- First referral to CDT
- Second referral (if exists) to CDT

Hierarchical clustering used in two ways

- Clusters for each time point
- Clusters for trajectories through the 3 time points

Found 5 pathways, key is whether assessed



Dyer (2006). "The Psychiatric and Criminal Careers of Mentally Disordered Offenders Referred to a Custody Diversion Team in the United Kingdom, *International Journal of Forensic Mental Health*, 5:1, 15-27, doi: 10.1080/14999013.2006.10471227

# Organising clusters

Clusters exist in multi-dimensional space

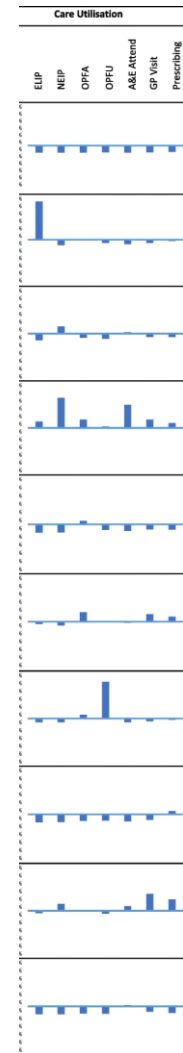
- Many attributes so many dimensions

Difficult to visualise and hence interpret

Visualisations focus on attribute values in each cluster

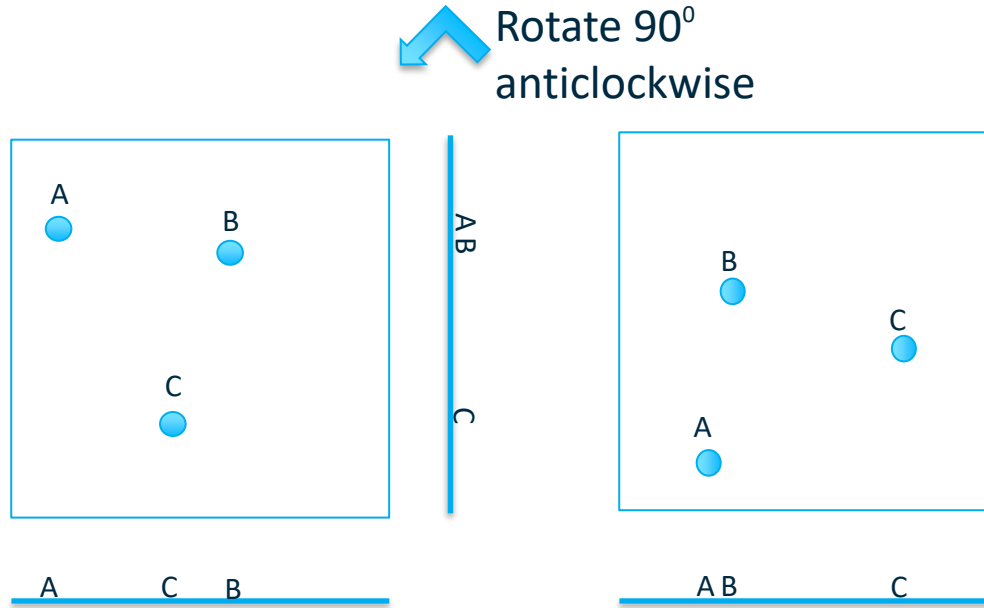
- More difficult as number of attributes increases

Dimension reduction orders the clusters so that comparisons can focus on 'nearby' clusters

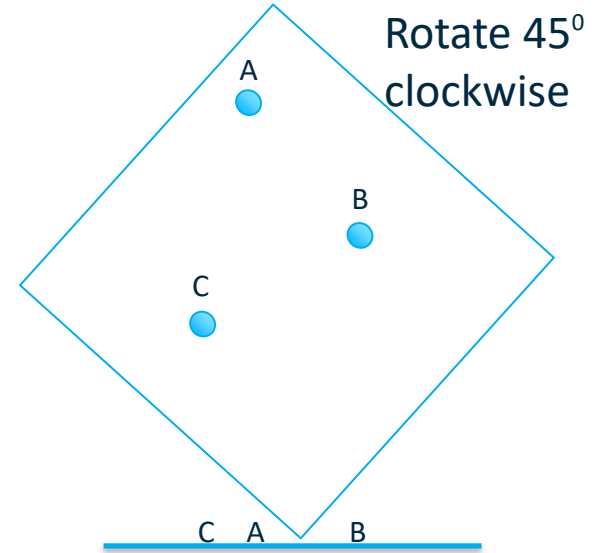


# Dimension reduction distorts distances and position

## For example, moving from 2D to 1D



Same as viewing from  
left of slide in first image



Aside: PCA angle to maximise  
variation, hence separation

# Self organising maps

## Self organising maps (SOM)

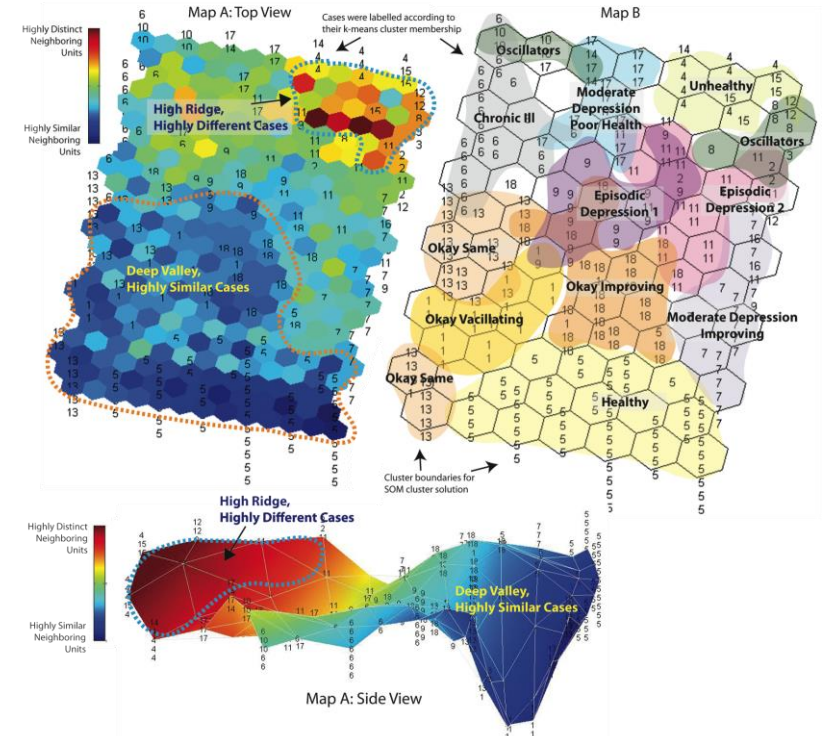
- Is a clustering method
  - Assigns cases to clusters
- Is a dimension reduction method
  - Clusters that are close in higher dimensional attribute-space are close in the lower dimensional space

Used for visualisation as well as clustering

- Reduce to 2D for ease of visualisation

Also referred to as Kohonen maps

Figure 2: Self-Organizing Topographical Map of Eleven Major and Minor Trends



# SOM algorithm

## Initialise

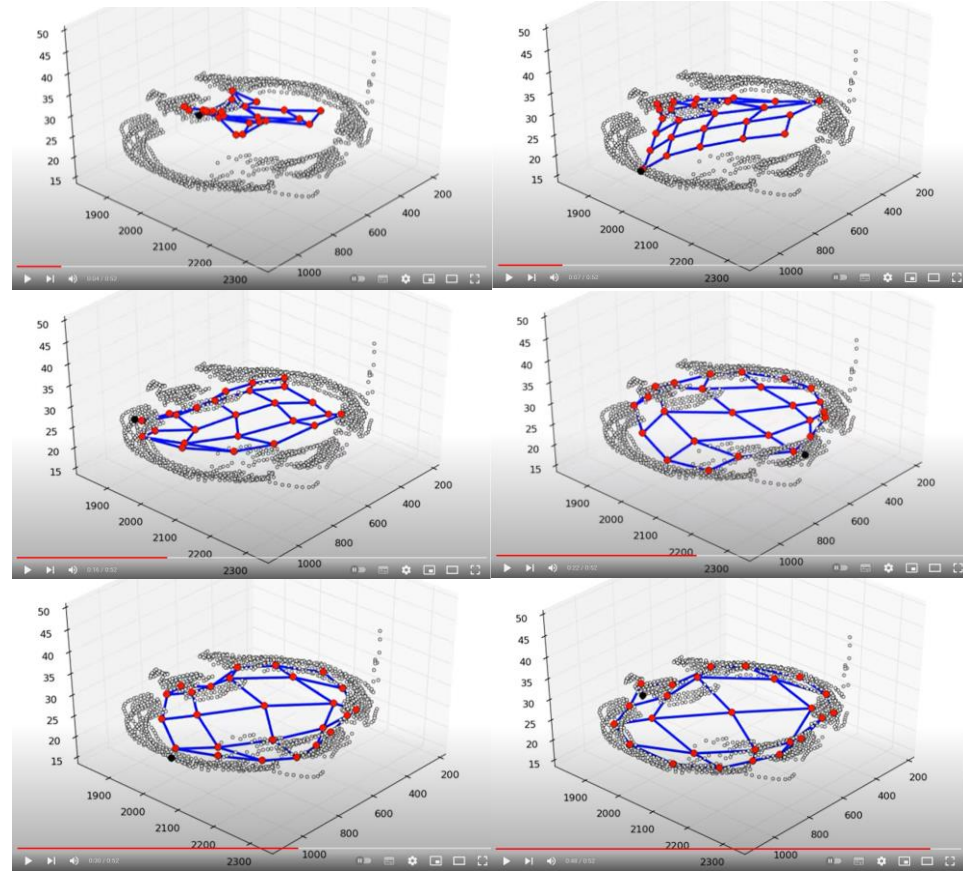
- Mesh of prototype clusters (neurons, synapses)
- Each neuron assigned random values for all attributes (weights)

## Iterate through all cases, random order

- Find closest neuron
- Update attribute values to move it toward case
- Move neighbouring neurons toward case

Neurons eventually located in areas with high case density but also connected to each other

- Cases in cluster of close neuron



# SOM hyperparameters

## Mesh of neurons

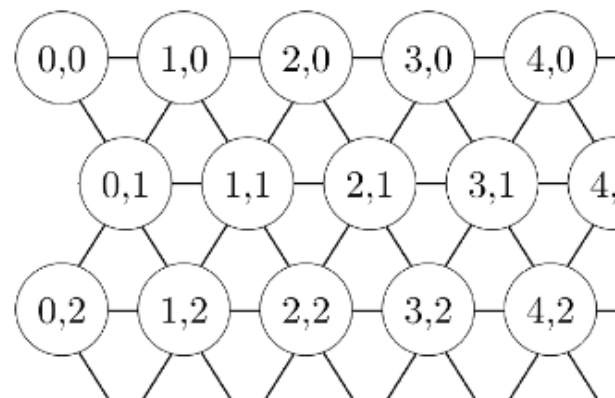
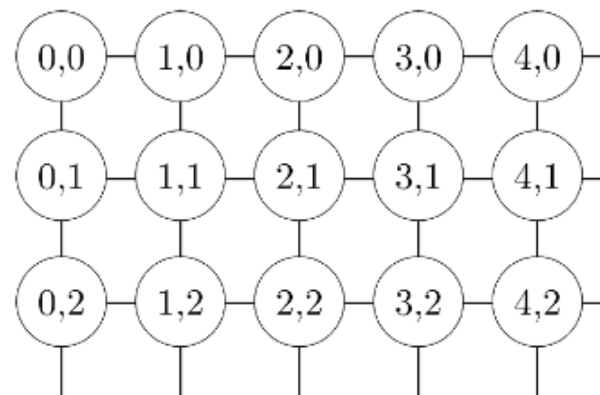
- Topology: rectangular or hexagonal
- Size: need more neurons than expected clusters

## Learning rate

- How much the best matching unit (closest neuron) moves
- Decreases over time so map converges

## Neighbourhood

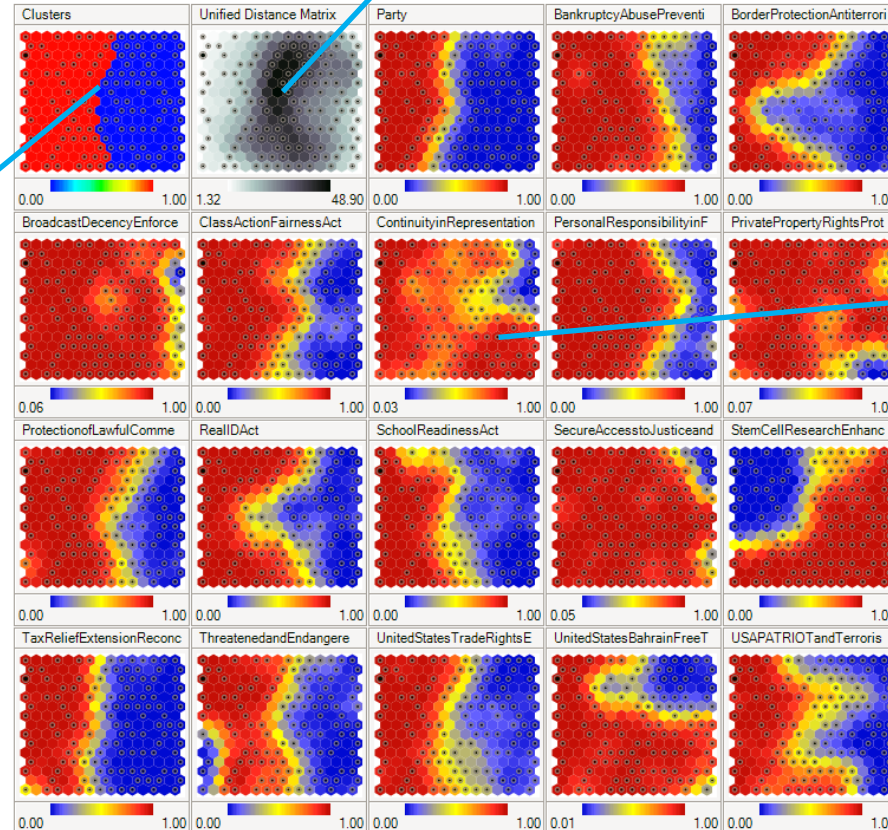
- Radius: what neurons are also moved
- Rate function: how much they move



# SOM visualisation

U-matrix (unified distance): distance between neurons (clusters), dark=far

Clusters formed from voting patterns



Single attribute values (vote)



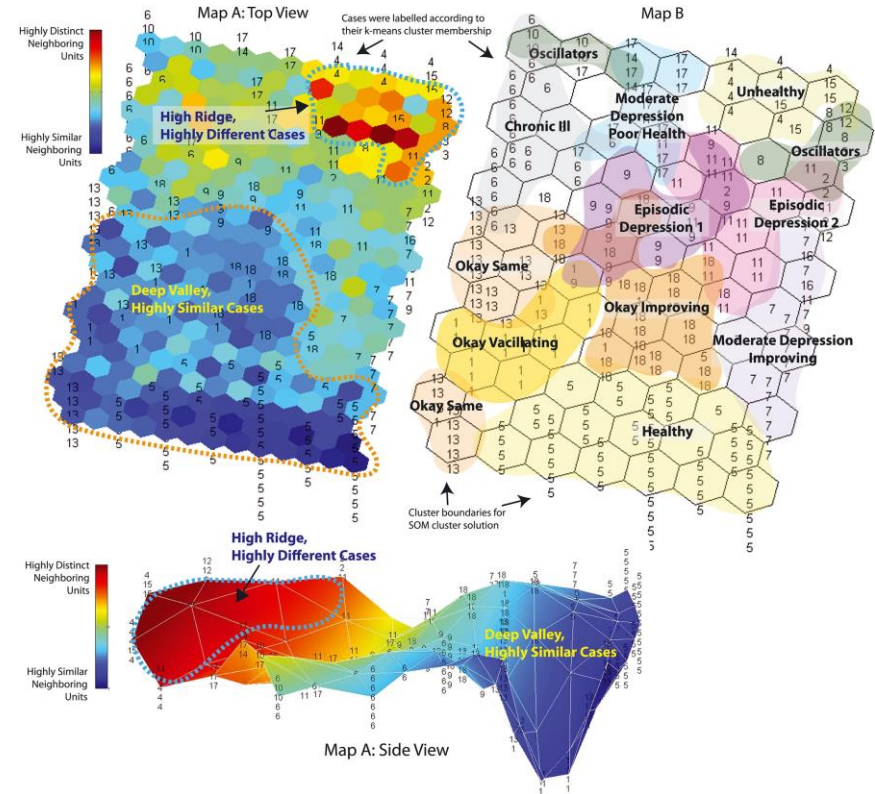
# SOM depression and physical health trajectories

259 longitudinal cases

- 20 biomarkers for stress
- 23 health outcomes
- 11 time points

Found 11 trajectories

- Used *k-means* to identify 18 clusters
- Constructed 7x12 SOM
- SOM to verify the *k-means* solution



Castellani et al (2018). Exploring comorbid depression and physical health trajectories: A case-based computational modelling approach. doi:10.1111/jep.13042



# How many clusters?

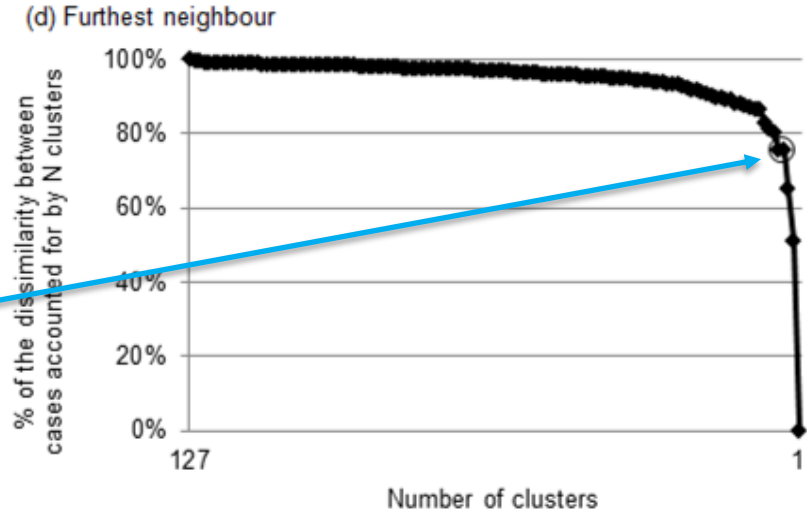
Each additional cluster increase the amount of variation between attribute values 'explained' by the clustering

- Fewer cases in each cluster, so more similar

Plot number of clusters against explained variation

- Judgement as to the point at which additional clusters offer little improvement
- Referred to as elbow
- Justifies choice of number of clusters

Heuristic only: no correct method



# Silhouette plots

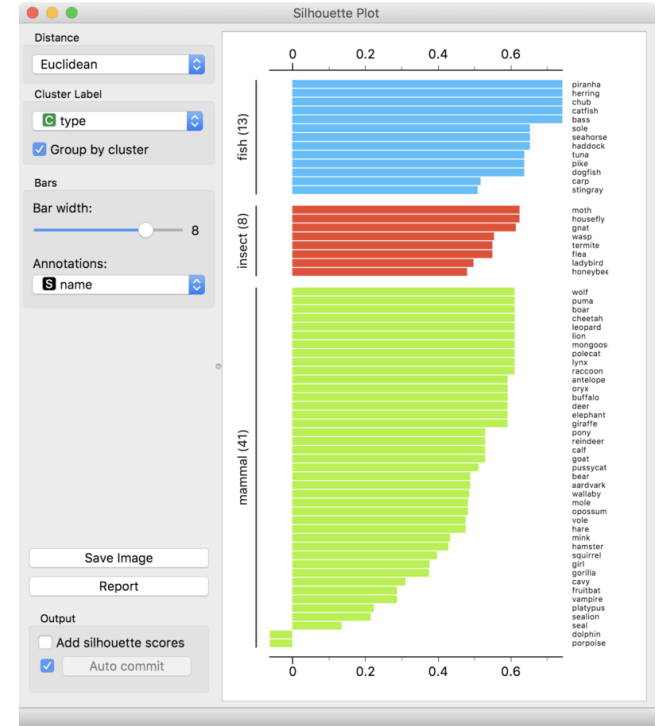
Silhouette plots display the similarity of the cases within each cluster to help verify the clustering solution, whether clusters are meaningful

Silhouette value is a function of the case attributes and the assigned cluster

- Let  $S$  = average distance to cases in same cluster
- Let  $N$  = average distance to closest neighbour cluster
- Silhouette value given by
  - $1 - S / N$  for  $S < N$  (hence +1 for perfect fit with own cluster)
  - $1 - N / S$  for  $N > S$  (hence -1 for perfect fit with neighbour)

Plot shows

- Cluster assigned for each case
- Silhouette value for each case



# Clusters found depend on the attributes and method used

All of these are important

- Specific attributes and how are they measured
- Composition of the sample (what cases selected)
- Clustering algorithm (hierarchical agglomerative, etc.)
- Choices in the algorithm (hyperparameters such as number of clusters)
- Distance metric
- How distance between clusters is measured (eg nearest neighbour, average)

So how do we know if the groups we have found are real or meaningful and not just statistical artefacts?

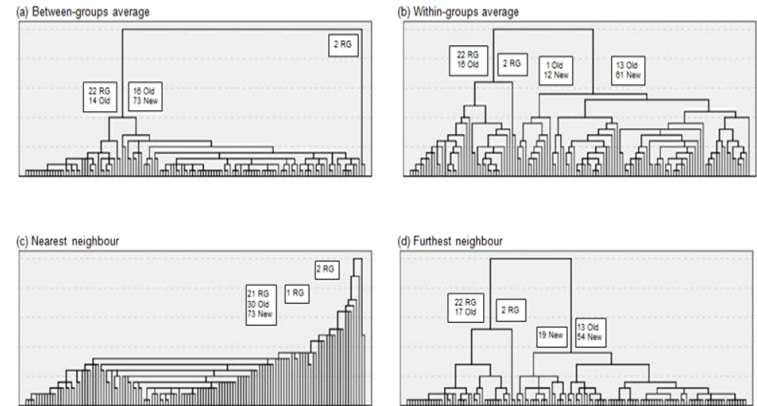
# Is the solution robust?

## Try different clustering methods and implementation details

Try with different hyperparameters such as number of clusters, initialisation, distance metric, and how to measure cluster distance

Remove some of the attributes: same results suggests robust

Introduce perturbations: slightly vary some attribute values or add fake cases with random attribute values, different results suggests not robust



Boliver (2015): Are there Distinctive Clusters of Higher and Lower Status Universities in the UK? doi: 10.1080/03054985.2015.1082905

# Are the clusters meaningful and distinct?

## Are average attribute values different?

	Oxbridge Cluster 1(N=2)	Most other Old Cluster 2(N=39)	Mainly New Cluster 3(N=67)	Struggling New Cluster 4(N=19)
<b>Research activity</b>				
Research income adjusted	37,100 (6,131)	21,356 (6,128)	5,740 (4,727)	1,620 (1,617)
% postgraduates	37.0 (3.9)	31.2 (8.0)	20.0 (7.7)	20.9 (11.7)
RAE score in 2008	3.0 (0.0)	2.9 (0.3)	2.0 (0.2)	1.7 (0.4)
<b>Teaching quality</b>				
% students satisfied with teaching	92.5 (0.7)	88.7 (1.9)	84.5 (3.4)	84.1 (2.9)
% students satisfied with feedback	73.0 (2.8)	67.9 (4.4)	69.4 (4.2)	69.5 (3.6)
Guardian value-added score out of 10	6.5 (0.7)	6.0 (0.7)	5.5 (1.1)	4.1 (1.4)
<b>Economic resources</b>				
Endowment/investment income (£000s)	23,871 (5,481)	4,266 (4,345)	687 (555)	392 (340)
Academic services spending per capita	2,812 (384)	1,514 (331)	1,055 (289)	724 (379)
Student-staff ratio	11.3 (0.4)	14.4 (1.8)	18.8 (2.2)	22.6 (3.9)
<b>Academic selectivity</b>				
Average UCAS points on entry	595 (18)	442 (47)	308 (33)	251 (28)
% students completing their degree	98.7 (0.4)	92.2 (4.5)	82.9 (5.0)	78.5 (7.0)
% students achieving a "good degree"	89.3 (3.5)	78.2 (5.0)	55.0 (5.5)	63.4 (5.6)
<b>Socioeconomic student mix</b>				
% students not from low participation neighbourhoods	96.6 (0.3)	93.2 (2.8)	87.4 (5.1)	82.3 (6.2)
% students from more advantaged social class backgrounds	89.4 (1.5)	77.4 (5.0)	61.6 (6.7)	56.4 (5.5)
% students from private schools	34.9 (2.9)	16.1 (8.2)	3.6 (3.0)	1.4 (0.9)

# Are the clusters consistent with other evidence?

Validity might also be ascertained through triangulation and/or mixed-methods research

Do the results align with those of other research?

- If so, this lends credence to their validity

Can you verify the results using other methods?



# Case-Based Complexity

(wrap up)



# In what ways is clustering a case based complexity method?

Case is a complex system

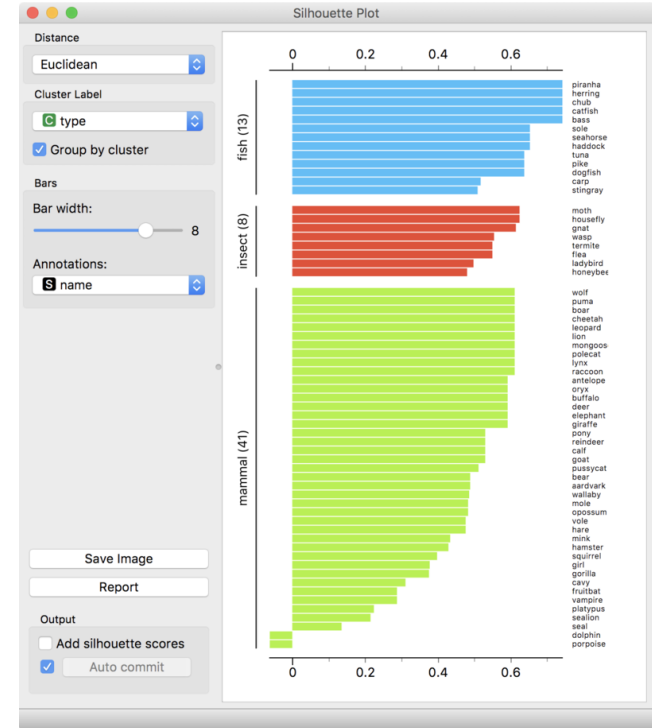
- Cases are treated as a composite of their attributes
- Attributes are interdependent, reflecting the history of the case
- Cases are dynamic, attribute values change over time

Case is the unit of analysis in clustering methods

- All attributes are used to assess similarity and difference

Outcome of the analysis is a catalogue of case profiles (at a particular point in time)

- Method is about identifying patterns in the cases
- Similar profiles clustered together
- Clusters help to understand diversity of cases





# Diversity of cases

Statistical methods and case based complexity differ in their attitude to diversity

- In statistics, average is meaningful for a variable
  - Good at interchangeable (fungible) cases
- In case based methods, all clusters are important
  - There is no average cluster
  - Small clusters show meaningful differences


A set of cases is more diverse where



- More clusters are required to cover the cases
- Cases are distributed evenly between clusters



## An entropy based measure for comparing distributions of complexity

R. Rajaram <sup>a</sup> , B. Castellani <sup>b</sup>

[Show more](#) 

[+ Add to Mendeley](#)  [Share](#)  [Cite](#)

<https://doi.org/10.1016/j.physa.2016.02.007>

[Get rights and content](#)

### Highlights

- An entropy based measure for comparison of diversity of complexity is proposed.
- The measure allows for comparison of diversity both within and across distributions.
- The measure is multiplicative i.e., a doubling of value implies a doubling of diversity.

# Review: How does clustering work?

Clustering is an exploratory method in the big data toolbox

Cases are conceptually located in multidimensional space

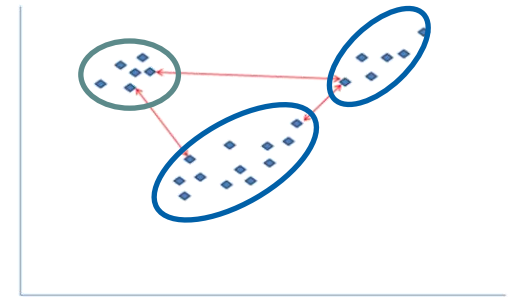
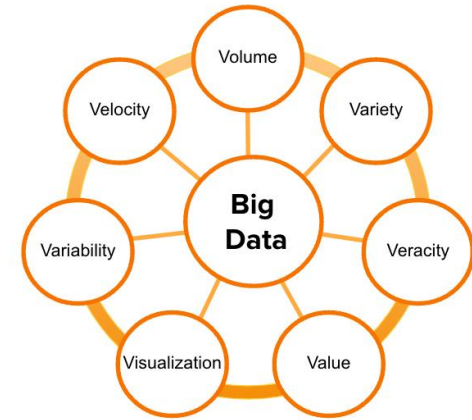
- Each dimension refers to one attribute

Goal is to allocated cases to clusters so that

- Cases within a cluster are close to each other
- Clusters are distant from each other

Euclidean (straight line) is common distance metric

## 7 V'S OF BIG DATA



# Challenges

Essential to treat clustering as an exploratory method

Finding clusters does not mean the cases are structured

- Any dataset can be clustered, even random noise

Always investigate the meaning of the clusters

- How robust are the clusters?
- Are the differences between attribute values meaningful?
- Are the clusters far apart?



# Next session: Case Based Complexity Workshop

## Dataset

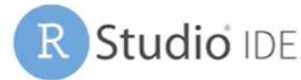
- USArrests – built in R dataset
- Cases are US States
- Attributes are criminal offence rates

## Cluster and visualise

- k-means
- Hierarchical clustering
- Self organising map

## R packages (+ tidyverse)

- cluster
- factoextra, dendextend
- kohonen, SOMbrero



- Data analysis scripts
- Interactive web applications
- Documents
- Reports
- Graphs



# Alternative software: COMPLEX-IT



Exploring complex data from a case-based perspective

**Build the Model**

1. Build Database and Import Cases
2. Cluster Cases

**Test the Model**

3. The Computer's turn
4. Compare and Visualise Results

**Extend the Model**

5. Simulate Interventions
6. Predict New Cases

**Export Results**

7. Generate Report

beta version  
release 2019

COMPLEX-IT is a web-based and downloadable software tool designed to increase your access to the tools of computational social science (i.e., artificial intelligence, micro-simulation, predictive analytics). It does this through a user friendly interface, with quick access to introductions on concepts and methods; and with directions to richer detail and information for those who want it.

The result is a seamless and visually intuitive learning environment for exploring your complex data -- from data classification and visualisation to exploring simulated interventions and policy changes to data forecasting.

**You don't need any technical expertise to start using COMPLEX-IT, all that is required is a data set you want to explore, and a curious mind!**



**DOWNLOAD  
VERSION**

**USER  
RESOURCES**

Video Tutorials  
Step-by-step User Guide  
Additional Readings




**WEB  
VERSION**


Meet the team















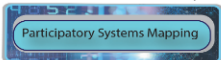
## COMPLEX-IT

<https://www.art-sciencefactory.com/complexit.html>

**For each of the major TABS and STEPS in COMPLEX-IT, we have provided a quick video tutorial and supporting documentation.**

- |              |            |   |
|--------------|------------|---|
| <b>VIDEO</b> | <b>PDF</b> | Basic Introduction - What is Case-based modeling?   |
| <b>VIDEO</b> | <b>PDF</b> | OVERVIEW OF COMPLEX-IT.  <b>This is the key article!</b> |
| <b>VIDEO</b> | <b>PDF</b> | Thinking about data as cases and preparing the database.  |
- CSV File  
Visual Multiple  
Documentation Index

EXCEL File  
Visual Multiple  
Documentation Index

 **Tutorial Datasets**
- |              |            |  |
|--------------|------------|--|
| <b>VIDEO</b> | <b>PDF</b> | How to run & interpret the cluster analysis tab in COMPLEX-IT. |
| <b>VIDEO</b> | <b>PDF</b> | How to run the artificial neural net in COMPLEX-IT.            |
| <b>VIDEO</b> | <b>PDF</b> | Making sense of the visualisation map.                         |
| <b>VIDEO</b> | <b>PDF</b> | Comparing the artificial neural net to your cluster analysis.  |
| <b>VIDEO</b> | <b>PDF</b> | Rerunning your cluster analysis & artificial neural net.       |
- |              |            |   |
|--------------|------------|---|
| <b>VIDEO</b> | <b>PDF</b> | Participatory Systems Mapping -- <b>here is the software package.</b>  |
| <b>VIDEO</b> | <b>PDF</b> | Thinking about modelling and causality  |
- 
- |              |            |  |
|--------------|------------|--|
| <b>VIDEO</b> | <b>PDF</b> | How to run a case-based scenario simulation in COMPLEX-IT. |
| <b>VIDEO</b> | <b>PDF</b> | How to run the sensitivity analysis for your simulation.   |
- |              |            |  |
|--------------|------------|--|
| <b>VIDEO</b> | <b>PDF</b> | How to run the prediction/forecasting tab. |
| <b>VIDEO</b> | <b>PDF</b> | RUNNING THE SOM TO DO DATA FORECASTING!    |
- |              |            |   |
|--------------|------------|---|
| <b>VIDEO</b> | <b>PDF</b> | Making sense of the REPORT provided at the end of your analyses |
|--------------|------------|---|

## Unsupervised

