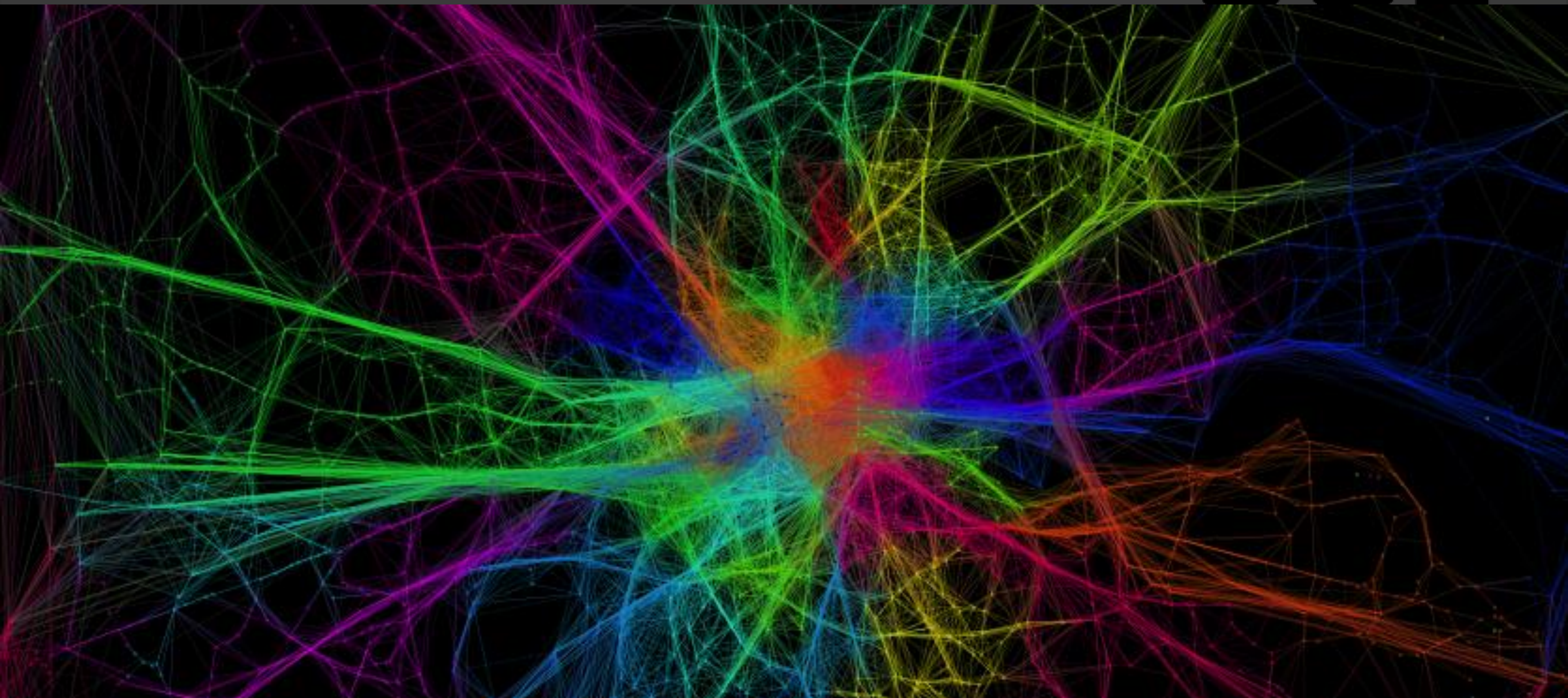


Advanced Clustering

CASA0006: Spatial Data Capture and Analysis
CASA0009: Data Science for Spatial Systems

Huanfa Chen

Thanks to Ed and Thomas for some slides

CASA0006

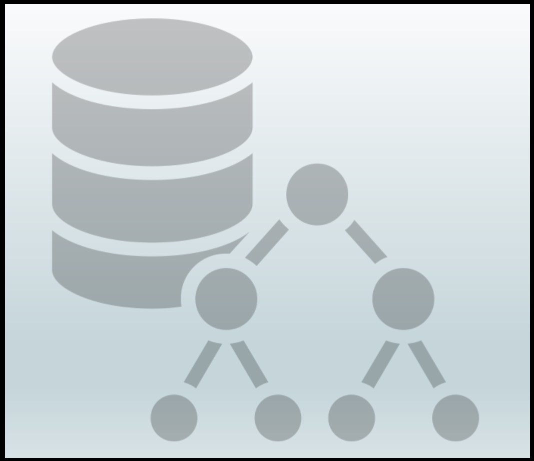
1. Introduction to Databases
2. Introduction to SQL
3. Advanced SQL
4. Data Munging
5. Advanced Clustering
6. Advanced Regression
7. Classification
8. Dimension Reduction
9. Unstructured Data
10. Analysis Workflow

CASA0009

1. Introduction to Databases
2. Introduction to SQL
3. Advanced SQL
4. Data Munging
5. Advanced Clustering
6. Advanced Regression
7. Interactive Viz 1: HTML + CSS
8. Interactive Viz 2: Javascript
9. Server Side Coding: Node.JS
10. Real-time data visualisation

Recap

What we already know



Database and SQL

Different data formats

Data cleaning using Python and Pandas

Data Analysis?

Connecting with *Quantitative Methods*

Clustering : Plan of Attack

Please find the lecture note and video of QM lecture 'Cluster Analysis' on Moodle

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

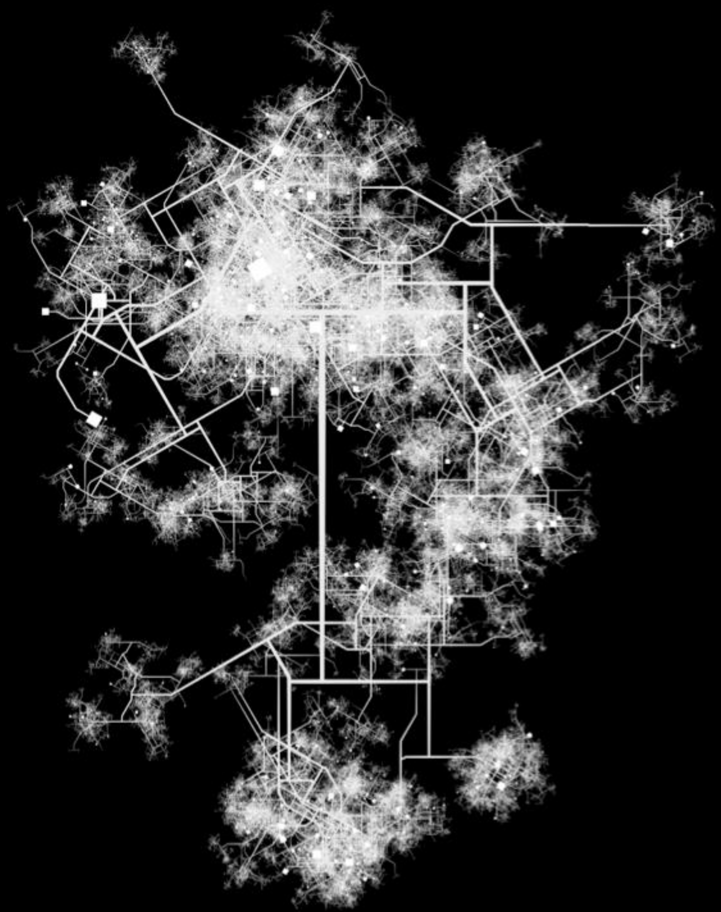
Clustering Methods
K-Means
Hierarchical

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

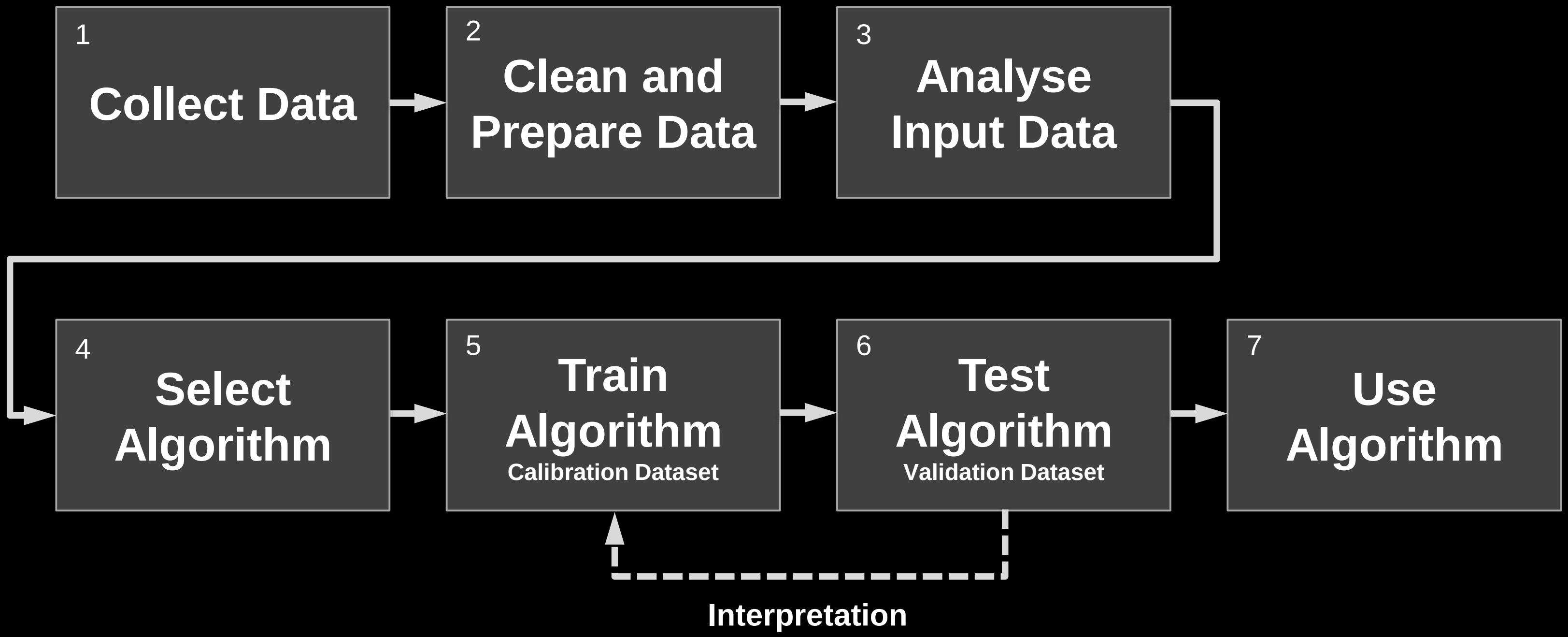
Outline



1. Overview
   a. Data Analysis Approaches
   b. Definition of Clustering
   c. Standardisation
2. Clustering Methods
   a. K-Means
   b. Hierarchical
   c. DBSCAN
3. Measuring Clustering Quality
   a. SSE/Elbow Method
   b. Silhouette Analysis
4. Next steps
   a. Visualisation
   b. Observe and interpret
   c. Consider analysing separately

Data Mining
Analysis Approach

All data analyses follow a similar methodology, regardless of the dataset and data mining approach being used



Data Analysis

Picking an Approach

The approach to take towards analyzing your data depends on what you want to understand from it

Input Dataset	Method		Output
	Clustering	→	Creation of Groupings
	Regression	→	Identify Data Relationships
	Classification	→	Identify Discrete Class
	Dimensionality Reduction	→	Understand Influential Factors
	Association Rule Mining	→	Identify Dependencies
	Anomaly Detection	→	Identify Outliers

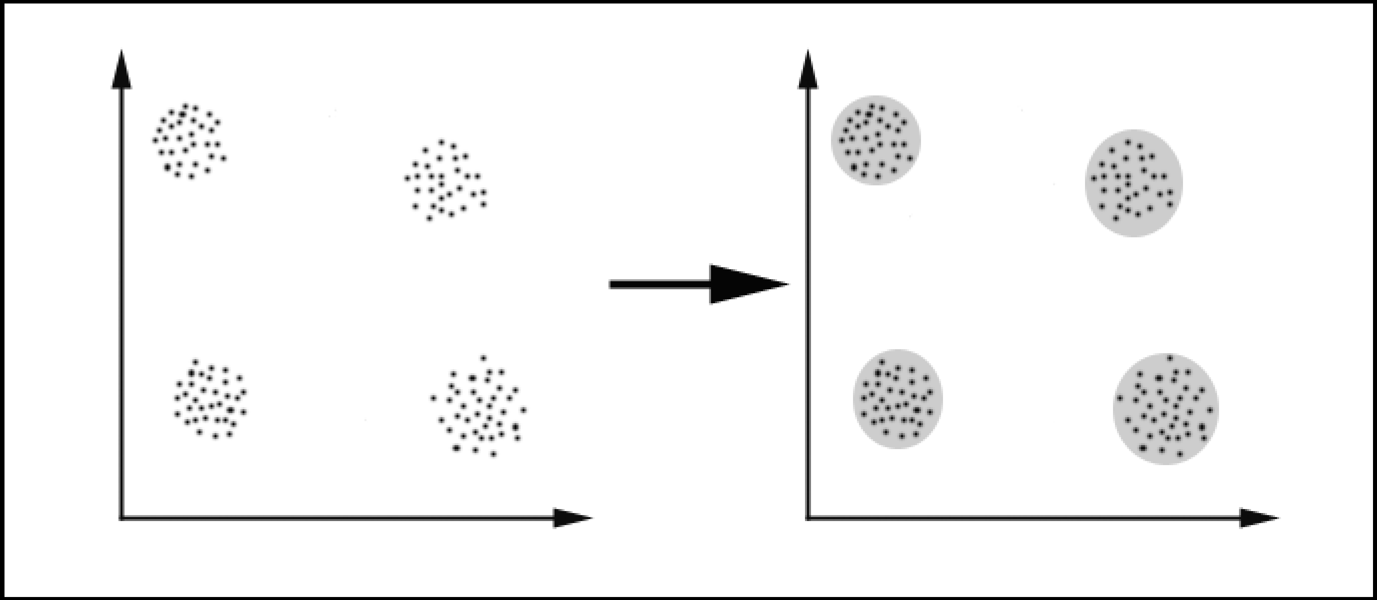
Unsupervised: no ground truth

Supervised: with ground truth

Clustering

Definition

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



Clustering

- Goals of clustering
  - Discover groups of similar observations
  - Reduce data size

- Issues with clustering
  - Unsupervised learning: no ground truth or accuracy measure available to check the result
  - Good news is that there are some ways to measure clustering quality
  - Clustering often involves many dimensions, so standardisation is necessary to make these dimensions comparable

Standardisation

Z score

(for not highly skewed data)

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

Min-Max Rescaling

(for highly skewed data)

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

IDR Standardisation

(Non-normal data with significant outliers)

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

Criteria

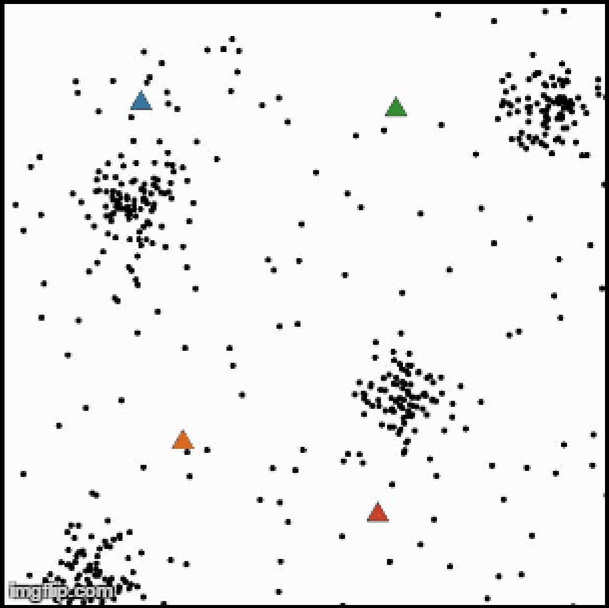
1. Highly skewed distribution?
2. Significant outliers?

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

Clustering

Divisive – K-Means Clustering

K-Means clustering **breaks down** a dataset into groups, based on proximity of points within a multidimensional space.



Iterative Algorithm

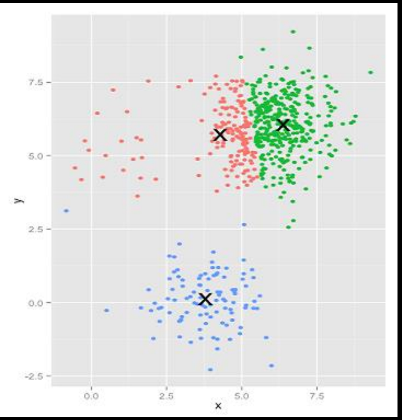
1. Place *k* centroids randomly within space
2. Assign points to nearest centroid
3. Recalculate centroids as the new mean of the cluster
4. Continue until centroid assignments no longer change

Clustering

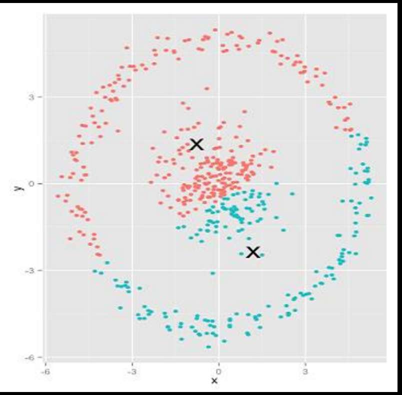
Problems with K-Means Clustering

- Requires knowledge of the number of clusters, which you may not know in advance (solution: Elbow method);
- Sensitive to initialisation, which can lead to poor solutions (solution: try different random initialisation and get a best one);
- Sensitive to outliers, which can results in inaccurate clusters (solution: use another clustering method or remove outliers);
- Incapable of handling clusters of a non-convex shape;
- Inapplicable to categorical data (solution: k-modes).

Choose k wisely

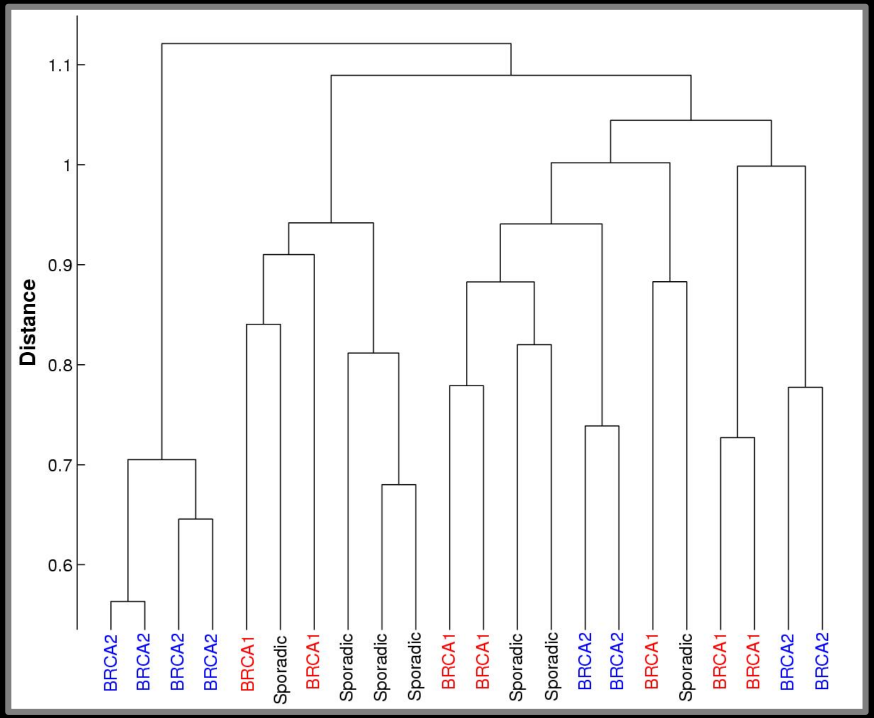


Non-convex shape



Hierarchical Clustering
Agglomerative

Hierarchical clustering **builds up** clusters based on proximity of instances, ending on reaching predefined number of points



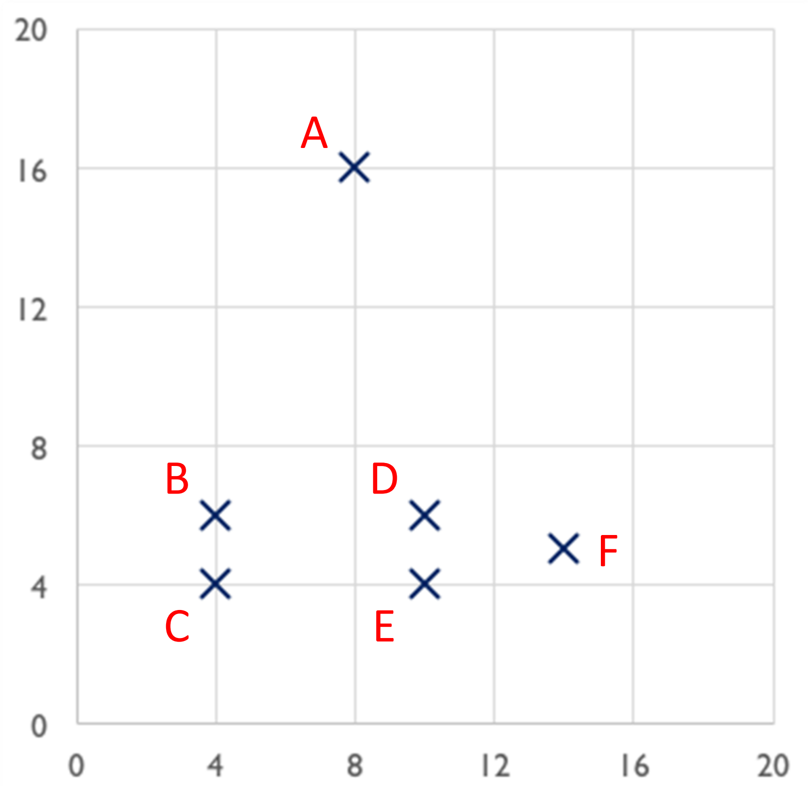
Iterative Algorithm

1. Start with every point in its own cluster
2. Merge points according to a *linkage criterion (or distance)*
3. Compute centroid of new clusters
4. Expand linkage threshold and continue until all points in one cluster

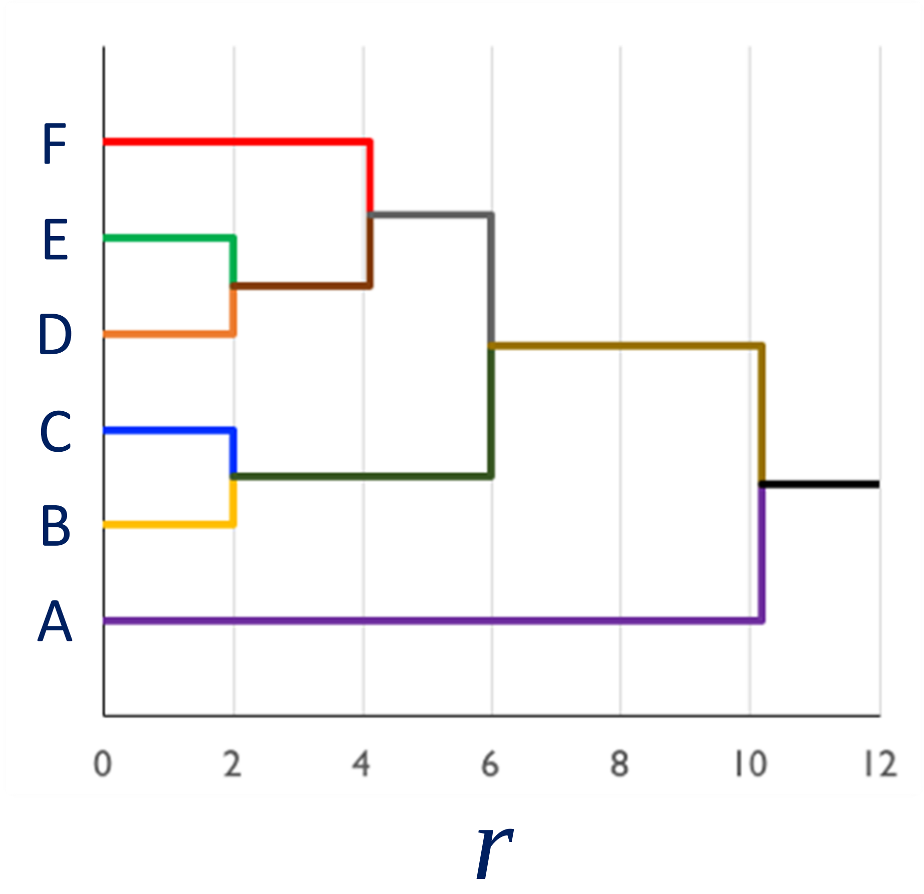
\+ Hierarchical structure
No a priori knowledge of data required

\- Can not un-agglomerate after cluster formed

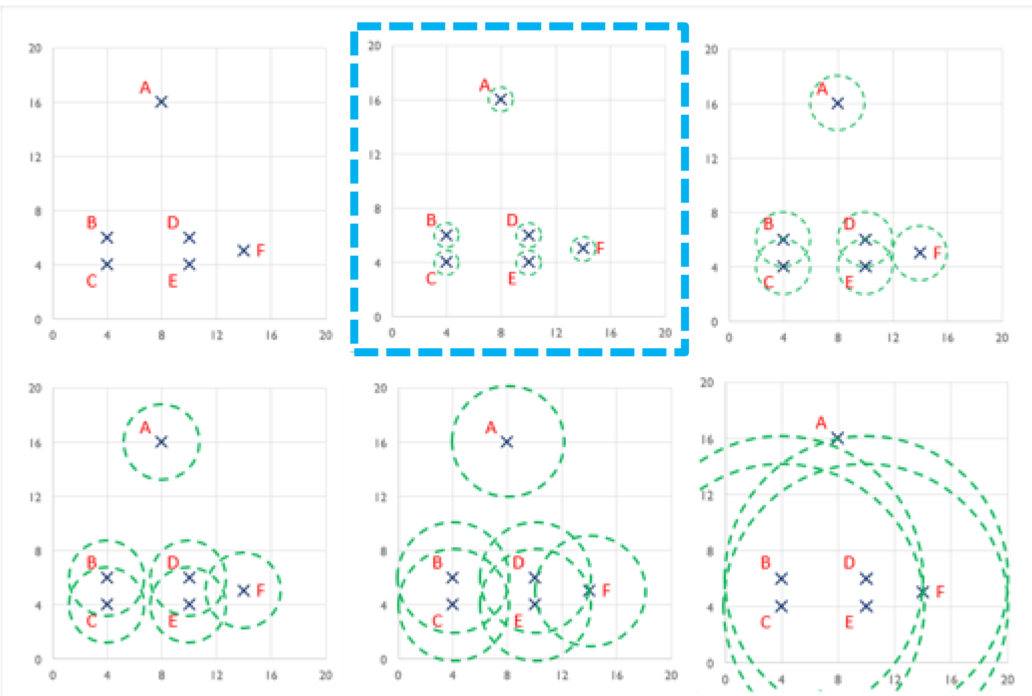
Hierarchical Clustering

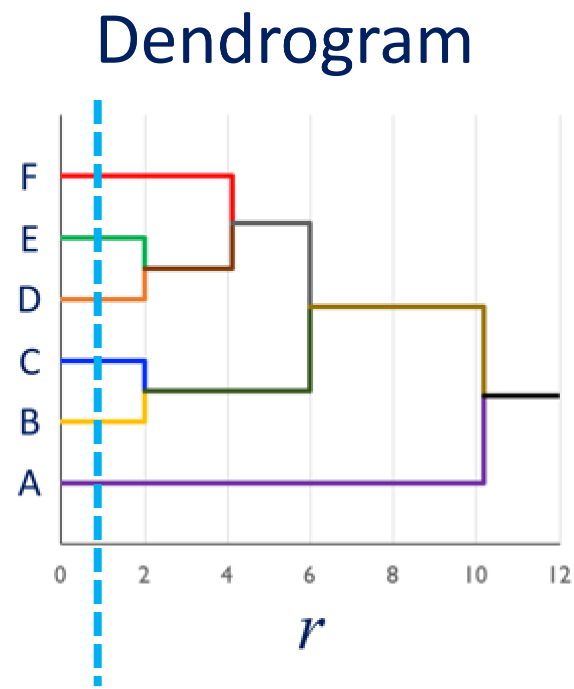


Dendrogram

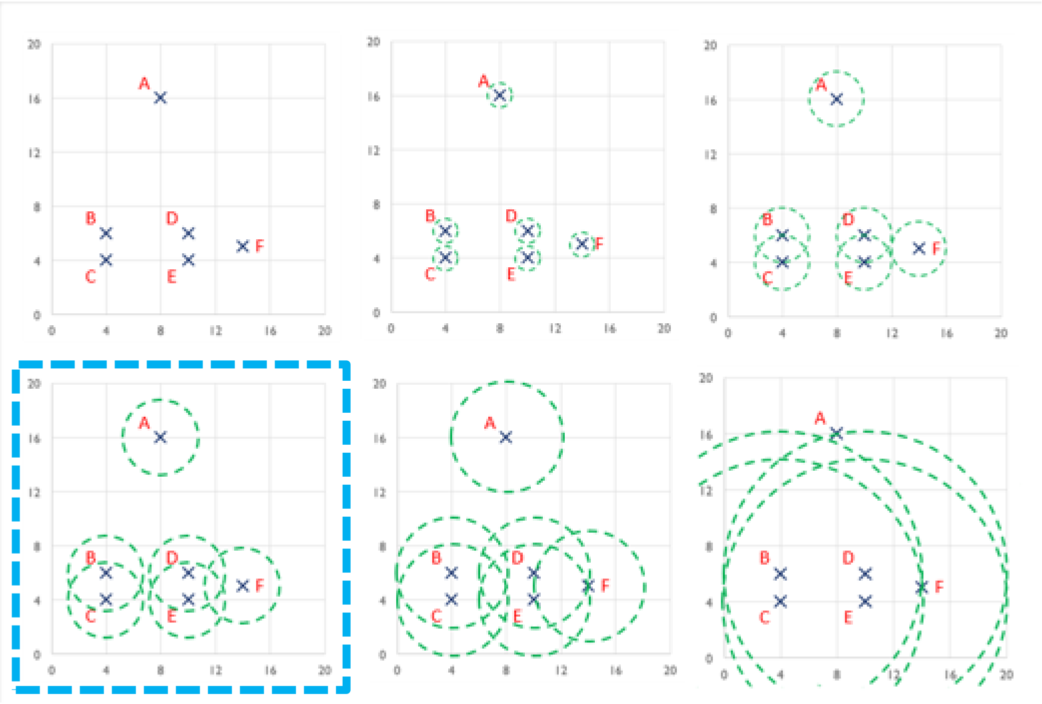


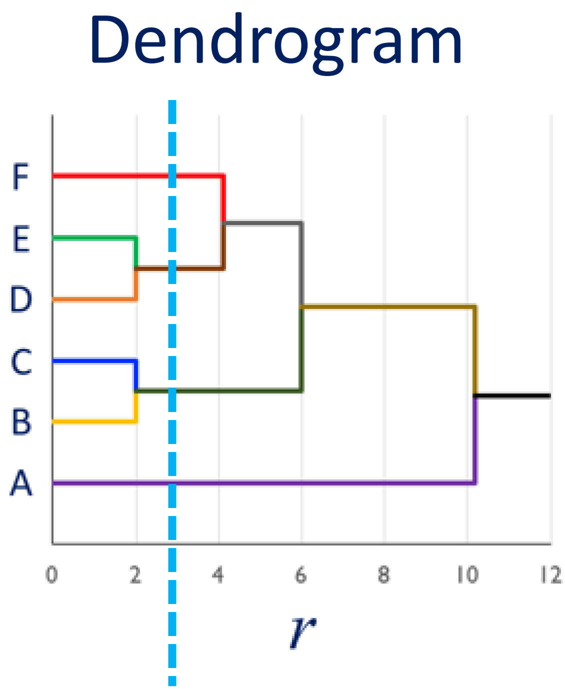
Hierarchical Clustering





Hierarchical Clustering





Hierarchical Clustering

Agglomerative

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

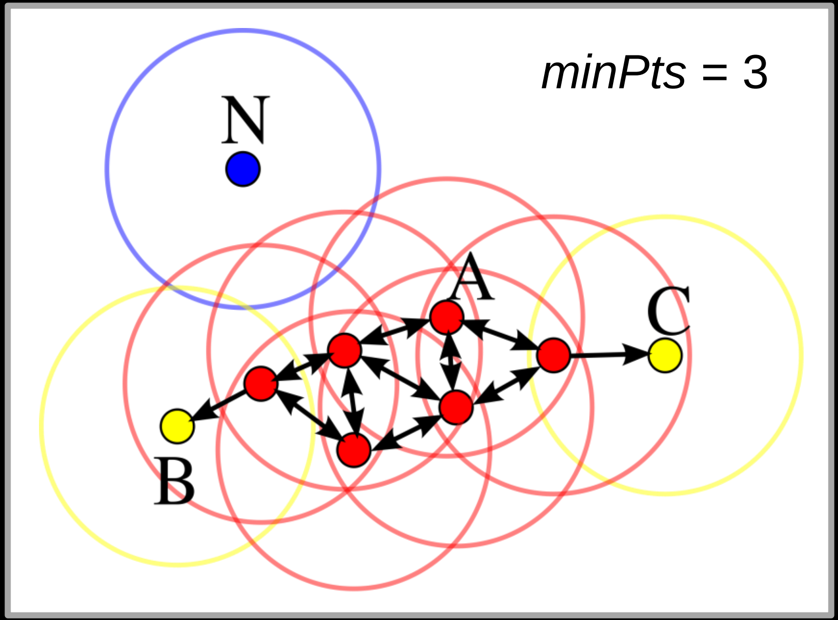
Divisive

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

Clustering

Density-based – DBSCAN Clustering

DBSCAN clustering joins builds clusters of points based on local proximity, only where falling within a maximum distance threshold



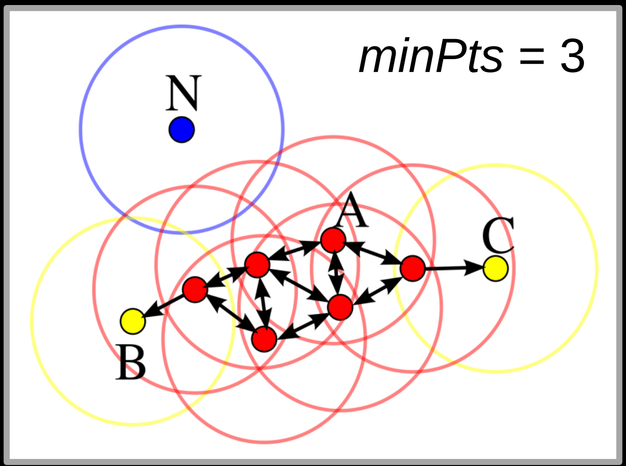
Given ε (search radius), points are classified into three classes:

1. Point p is **core point**: if at least minPts points are within distance ε of it (including p)
2. Point p is **edge point**: if p is not a core point but it is reachable from a core point
3. Point p is **outlier**: all points not reachable from any core points

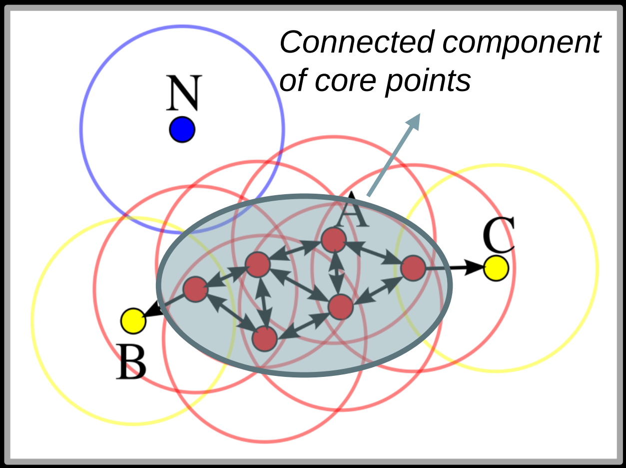
Clustering

Density-based – DBSCAN Clustering

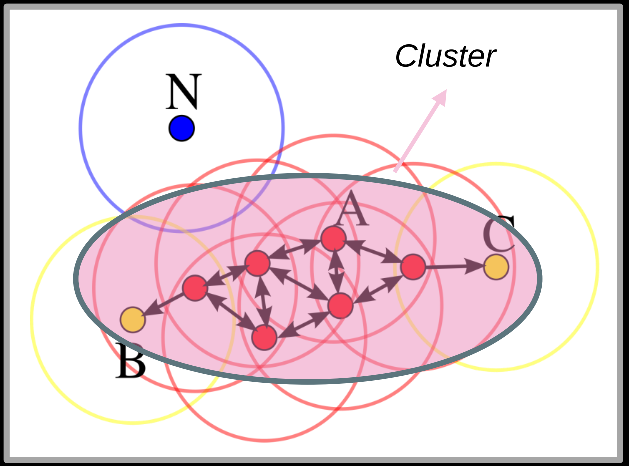
Step 1



Step 2



Step 3



Process

1 Find the points in the ε neighborhood of every point, and identify the core points

2 Find the connected components of *core* points, ignoring all non-core points

3 Assign each non-core point to a nearby cluster if the cluster is an ε neighbour, otherwise assign it to noise

Summary

Three clustering methods

method	required parameters
kmeans	number of clusters
hierarchical	number of clusters (but you can get a sense of it from the dendrogram)
DBSCAN	eps and minPts

Measuring Clustering Quality

How do you know our groups make sense?

Necessary when...

- **Comparing different implementations of a clustering method (e.g. k-means)**
- **Comparing clusterings with different numbers of clusters**
- **Comparing different clustering techniques**

Method 1: SEE / Elbow Method

- SSE: Sum of Squared Errors

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w^{(i,j)} \mathrm{dist}(x^{(i)}, \mu^{(j)})$$

Where: *i* is a observation*,* *j* is a cluster, *and w*$^{(i,j)}$*=1* when *i* is in cluster *j.*

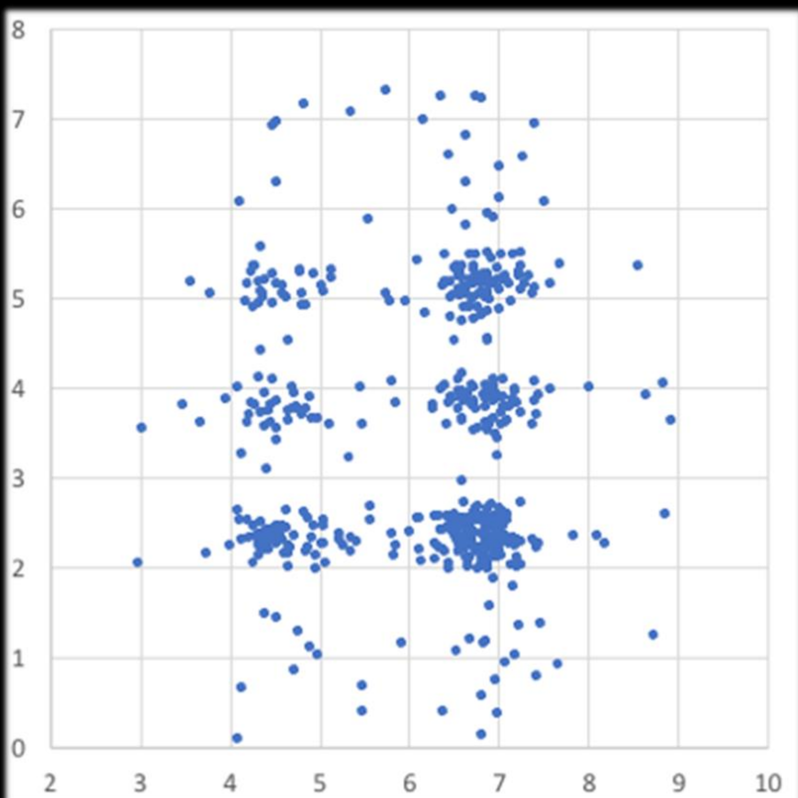
What is the range of SSE? [0, infinity)

- When the points in each cluster are identical, SSE = 0
- When #observation = #cluster, SSE = 0

Method 1: SEE / Elbow Method

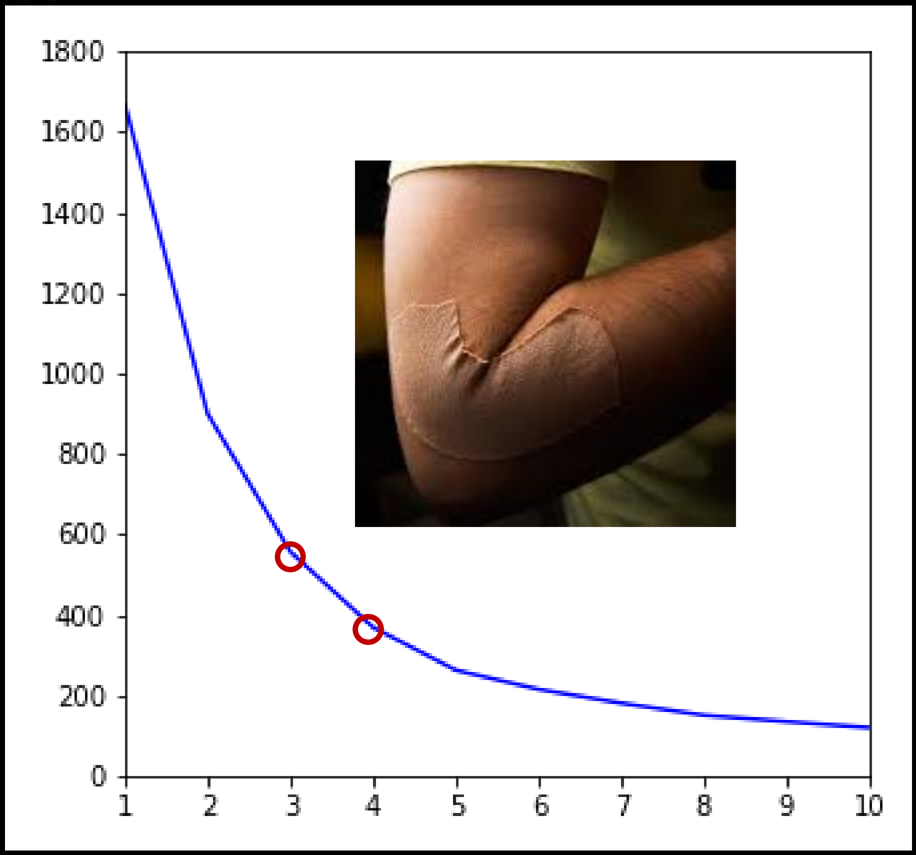
Elbow diagram: help choose k for k-means

Y



X

SSE

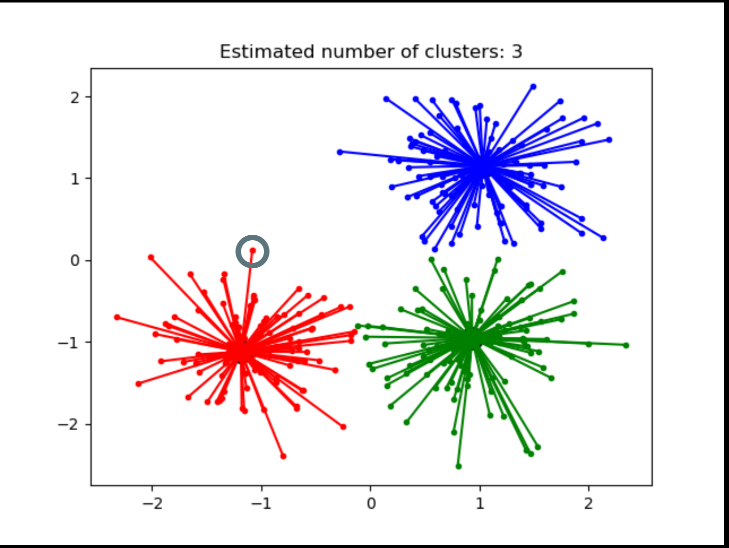


k (Number of Clusters)

Method 2: Silhouette Analysis

Silhouette of a point

"Is this point closer to points of the same cluster, or any other cluster? "



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

$a(i)$: mean distance to points of the same cluster

$b(i)$: minimum mean distance to points of another cluster

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

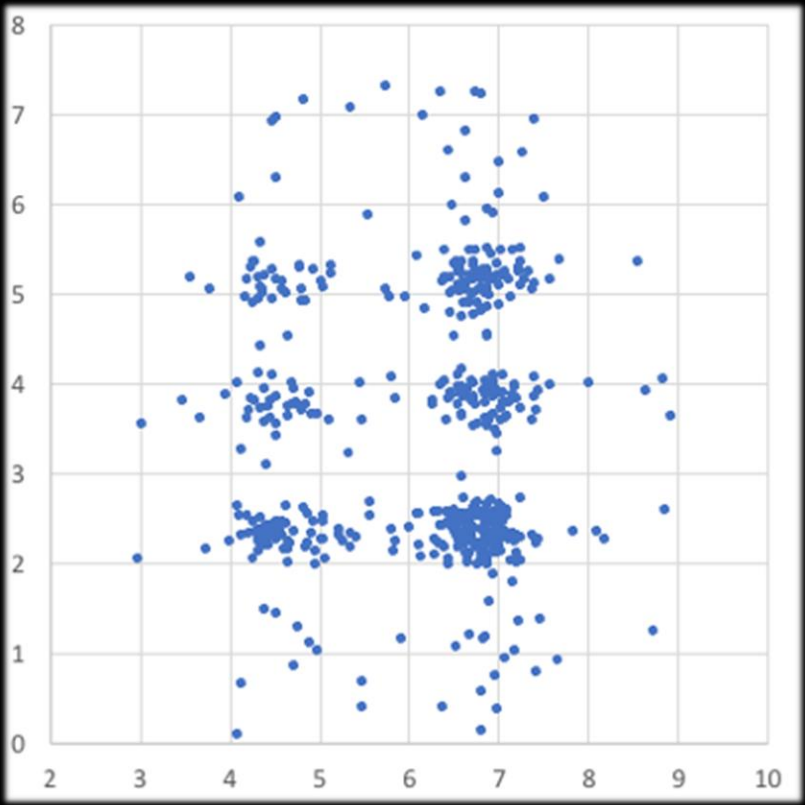
poorly clustered   well clustered

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

Method 2: Silhouette Analysis

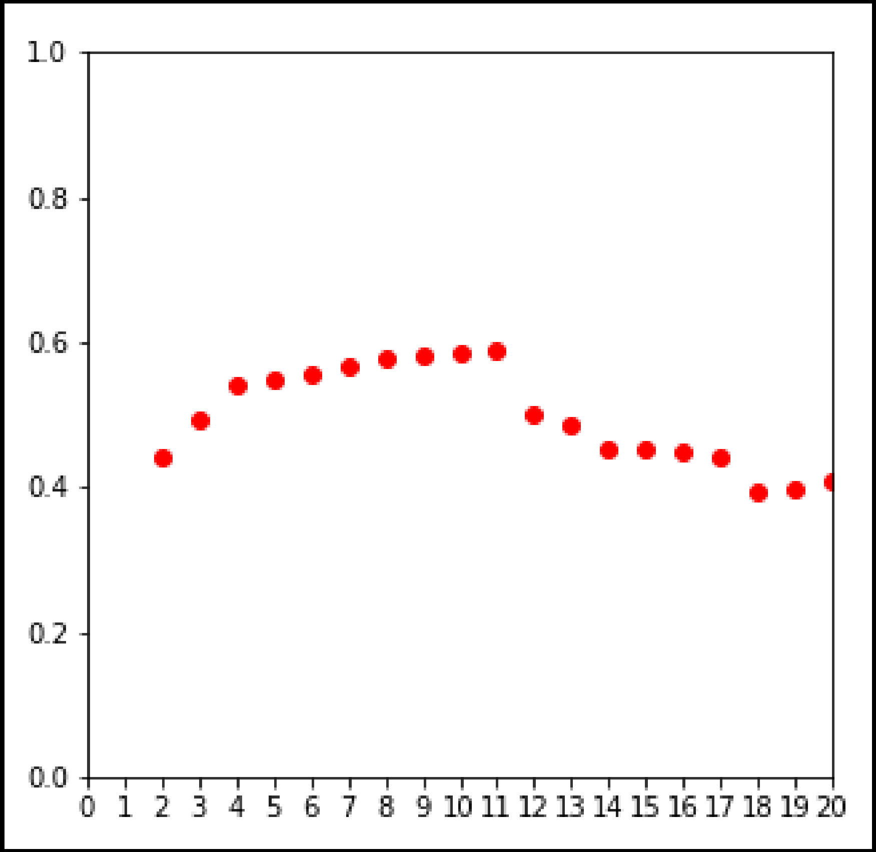
Choose k for k-means

Y



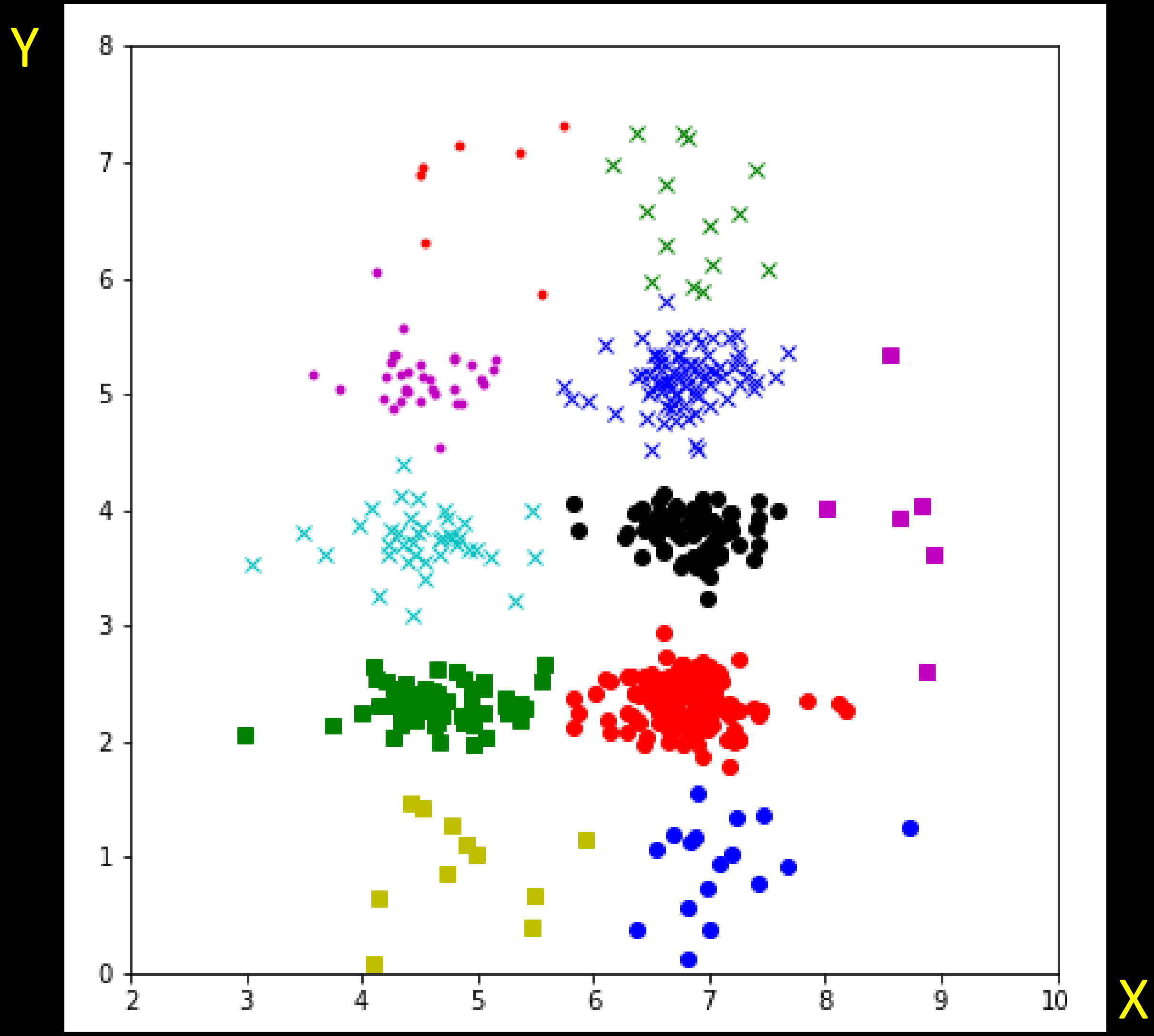
X

Silhouette Score



k (Number of Clusters)

Method 2: Silhouette Analysis



'Optimal' k-Means

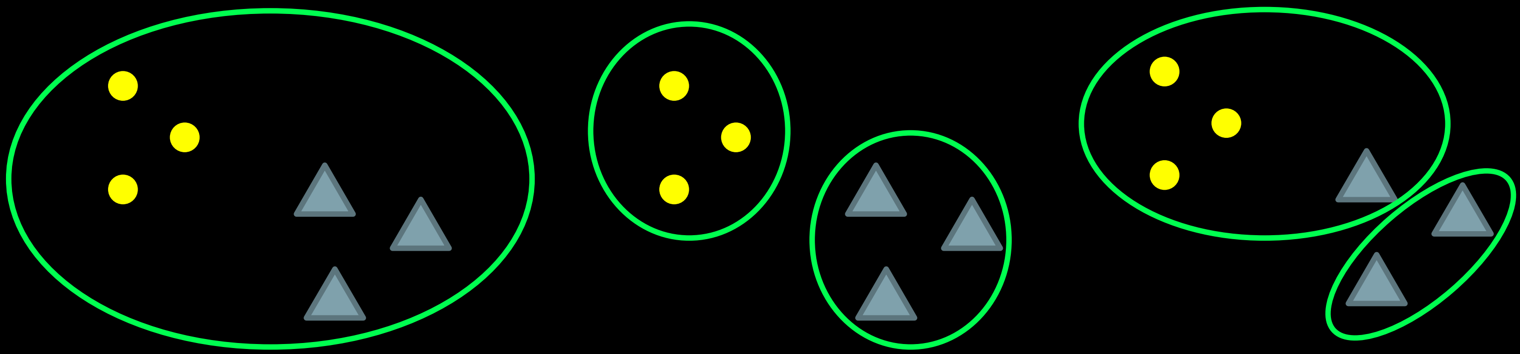
k = 11

S. Score = 0.59

Method 3: Comparing against ‘ground truth’

Homogeneity All clusters contain only points from a single observed class – expressed as a proportion of clusters for which this is true

Completeness All members of given class are within the same cluster – expressed as a proportion of classes for which this is true



Homogeneity	**0**	**1**	**0.5**
Completeness	**1**	**1**	**0**

https://scikit-learn.org/stable/modules/clustering.html#homogeneity-completeness

Method 3: Comparing against ‘ground truth’

- But … where is the ‘*ground truth*’ from?
  - Possibly you have some ground truth available, and you want to create a clustering for later use
  - The ‘ground truth’ can come from a different but relevant task. Be sure to prove they are really relevant.
  - You can ask some experts for ground truth. This is very common and useful.

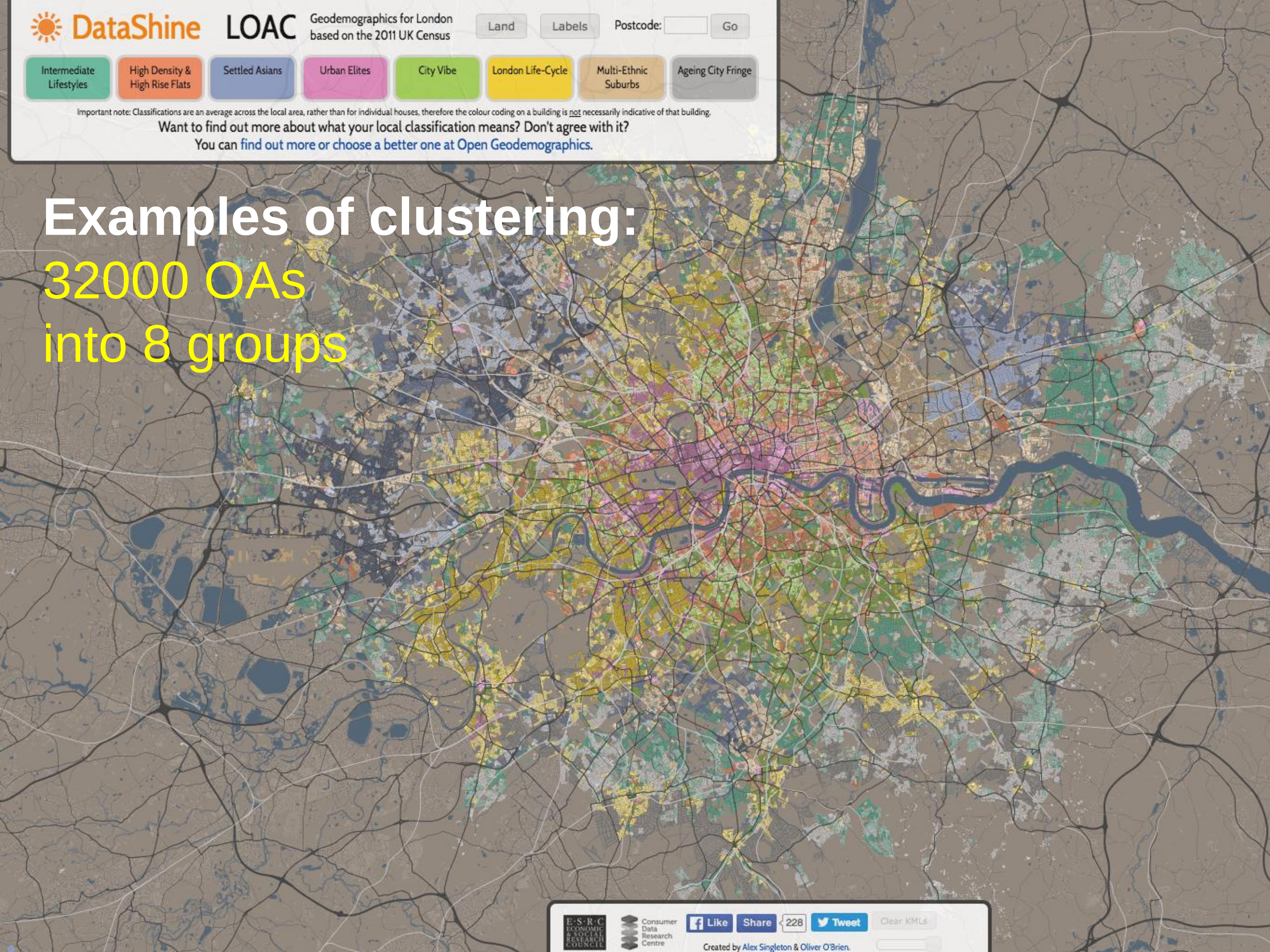
https://scikit-learn.org/stable/modules/clustering.html#homogeneity-completeness

Next steps of clustering
Very important !

- Visualisation (often combined with dimension reduction, e.g. PCA)
- Describe cluster characteristics
- Compare against unconsidered variables or geography (do these clusters cluster in space?)
- Compare against expert knowledge
- Consider analysing clusters separately

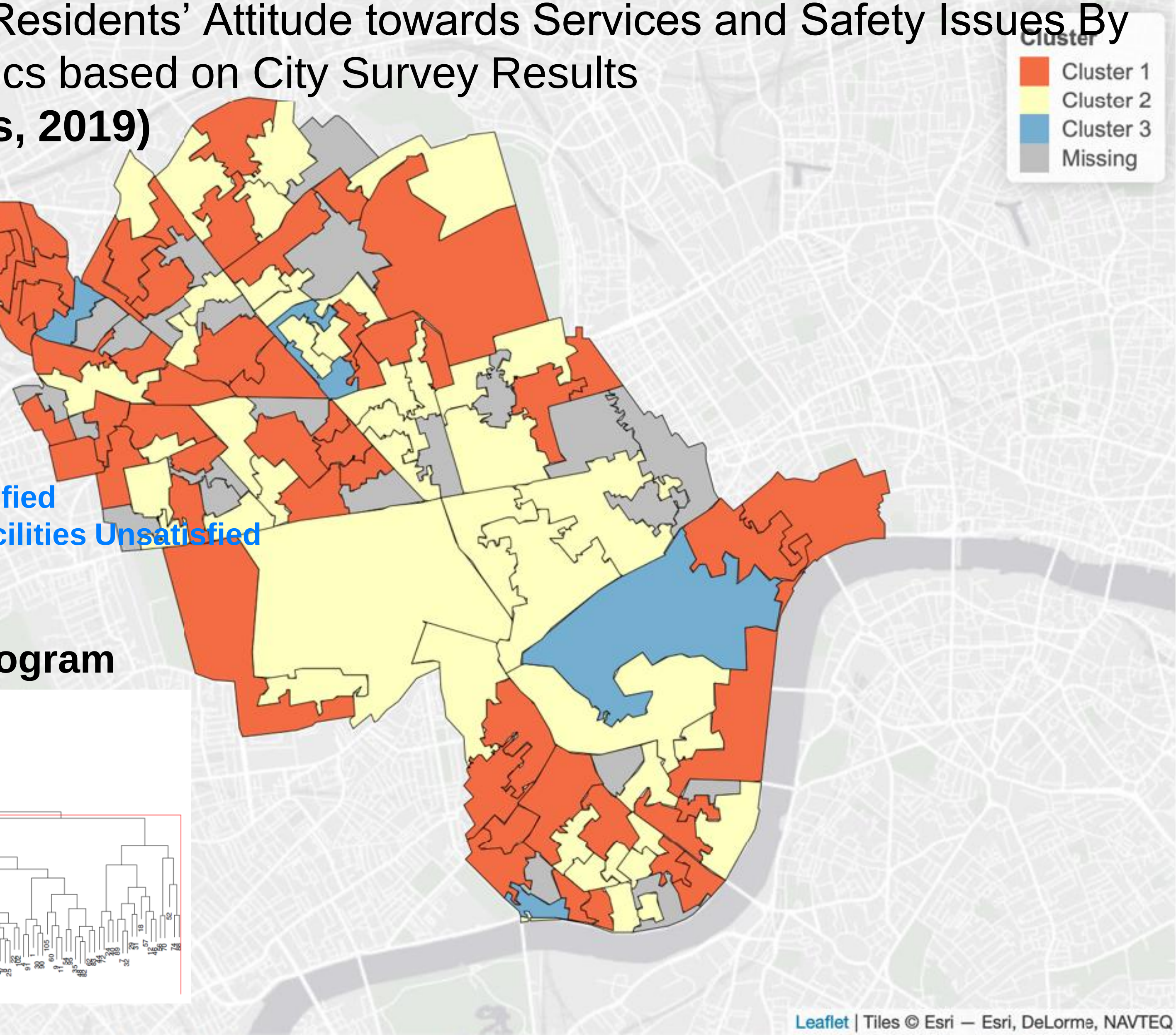
Examples of clustering:

32000 OAs
into 8 groups



Understanding Residents' Attitude towards Services and Safety Issues By Geodemographics based on City Survey Results

(Yafei Ye, MRes, 2019)



Input data:
20 survey questions

Result interpretation

- **Cluster 1: Env Unsatisfied**
- **Cluster 2: Outdoor Facilities Unsatisfied**
- **Cluster 3: Newcomers**

Cluster Dendrogram

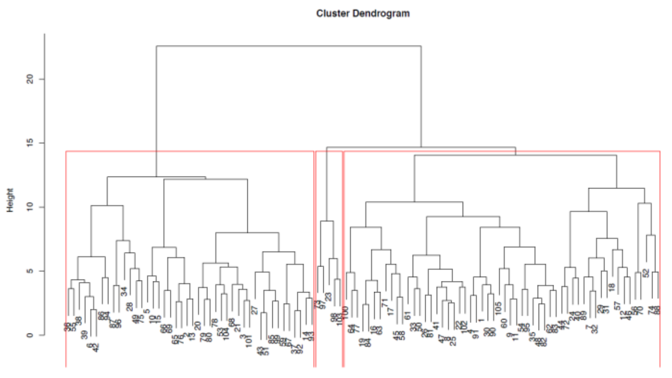
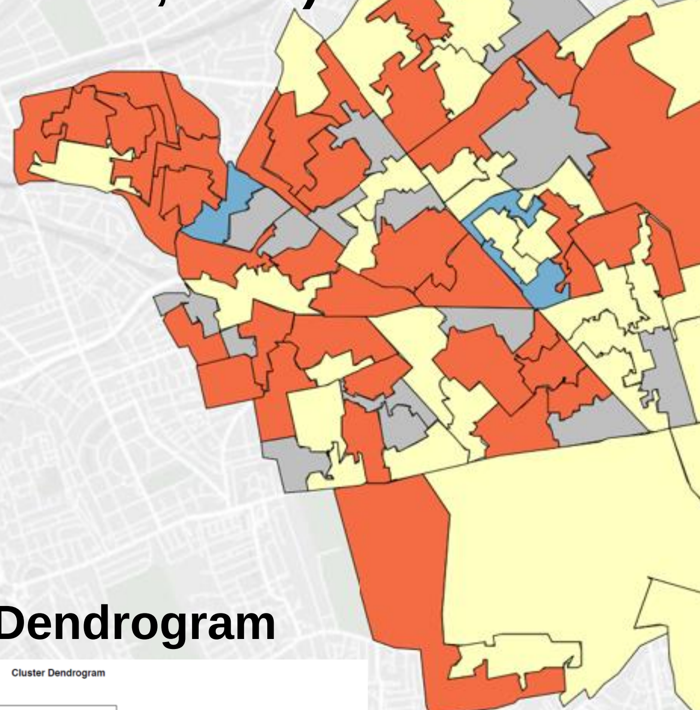


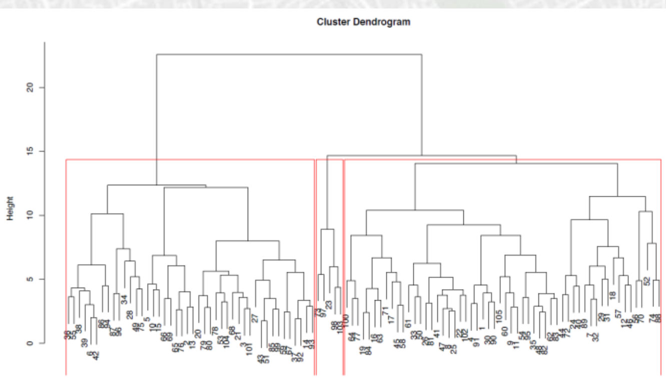
Figure 33. Cluster Map of HAC for Index of Service Usage Rate and Satisfaction

Understanding Residents' Attitude towards Services and Safety Issues By Geodemographics based on City Survey Results

(Yafei Ye, MRes, 2019)



Cluster Dendrogram



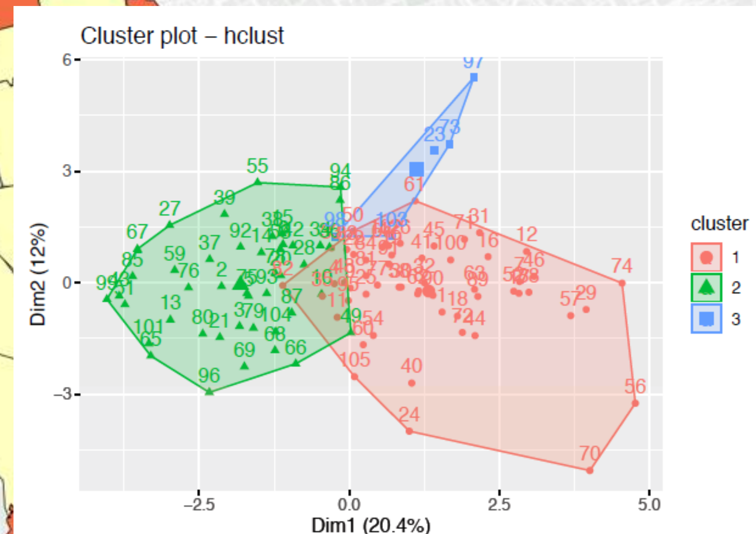


Figure 30. Cluster Plot of HAC for Index of Service Usage Rate and Satisfaction

Figure 33. Cluster Map of HAC for Index of Service Usage Rate and Satisfaction

Thank You

Questions?

Huanfa Chen

huanfa.chen@ucl.ac.uk

Workshop

Data Mining

- This workshop will focus on using clustering methods to analyse a multivariate dataset
- You'll learn how to use the scikit-learn Python library, which offers a number of useful tools for running data analysis methods
- Don't worry about the maths for clustering. You're not expected to understand all of the maths and algorithms. The key skill is the application, validation, and interpretation of the methods and results.
- **Download this week's iPython Notebook from Moodle, open it in Anaconda and work through**