CS989 Big Data Fundamentals
CS982 Big Data Technologies

Joseph El Gemayel, Ph.D.
Teaching Fellow, Course Director MSc Advanced Computer Science

Computer & Information Sciences
Email: joseph.el-gemayel@strath.ac.uk
Office: LT1207 - Livingstone Tower

Lecture overview

- Some questions about the coursework
- What is clustering?
- Distance measures
- Hierarchical clustering
- K-Means clustering
- Evaluation of clustering
- Discussion

Record my Attendance – CS982

Student password:

- nbgsb8

Record my Attendance – CS989

Student password:

- mx2hq5

Some questions about the coursework

- Which dataset should I choose?
  - Identify the topic that you are interested in before choosing the dataset
  - Take your time in choosing the topic and the dataset
  - Avoid time series datasets
  - Avoid dataset with very low number of relevant features
  - Try some exploratory analysis before deciding

Some questions about the coursework

- ➢ Which dataset should I choose?
- ➢ The dataset is too small or too large, what should I do?
  - ❖ If too small, look for another dataset on a similar topic to merge with
  - ❖ If too large, work on part of the dataset
  - ❖ Whatever you do, you should mention it in the report

Some questions about the coursework

- Which dataset should I choose?
- The dataset is too small or too large, what should I do?
- Fake or Real dataset?
  - Working with a real dataset is definitely better
  - But, for the coursework, fake dataset from Kaggle is fine

Some questions about the coursework

- Which dataset should I choose?
- The dataset is too small or too large, what should I do?
- Fake or Real dataset?
- When should I start working on the coursework?
    - Yesterday ☺

Some questions about the coursework

- Which dataset should I choose?
- The dataset is too small or too large, what should I do?
- Fake or Real dataset?
- When should I start working on the coursework?
- What should I add in the introduction?
  - Identification and description of key challenge(s) or problem(s) to be addressed
  - Not the problem(s) that you may face / have faced when dealing with a dataset

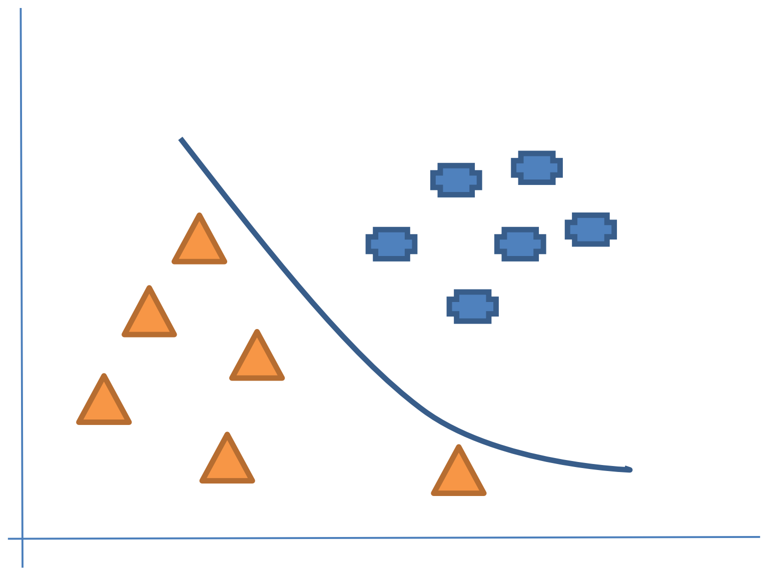
Some questions about the coursework

University of Strathclyde Glasgow

- Which dataset should I choose?
- The dataset is too small or too large, what should I do?
- Fake or Real dataset?
- When should I start working on the coursework?
- What should I add in the introduction?
- Any other question?

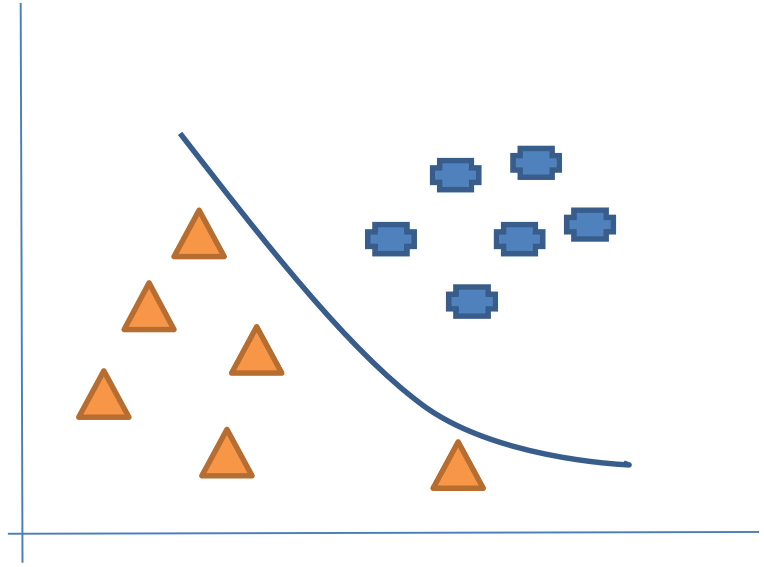
Machine Learning

- Supervised methods

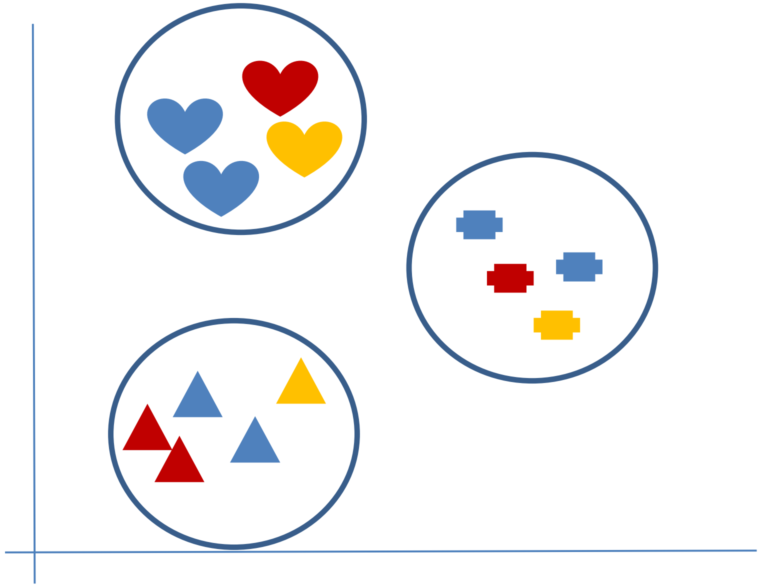


Machine Learning

- Supervised methods

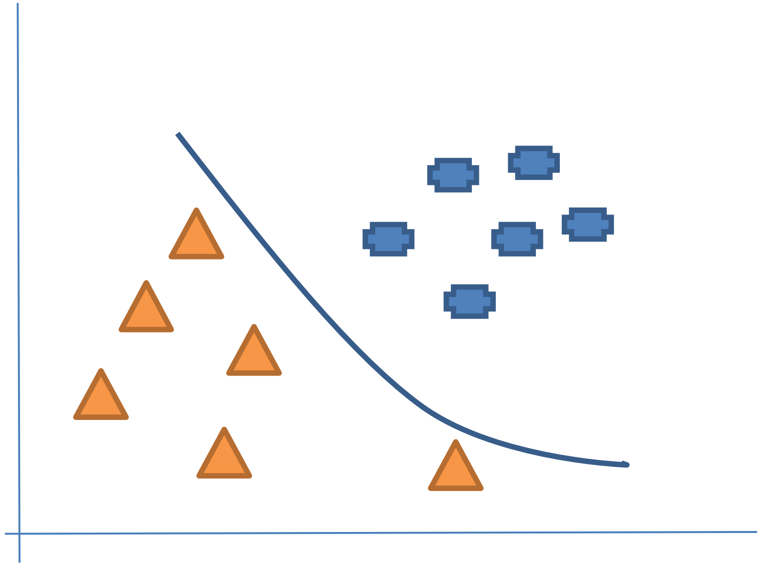


- Unsupervised methods

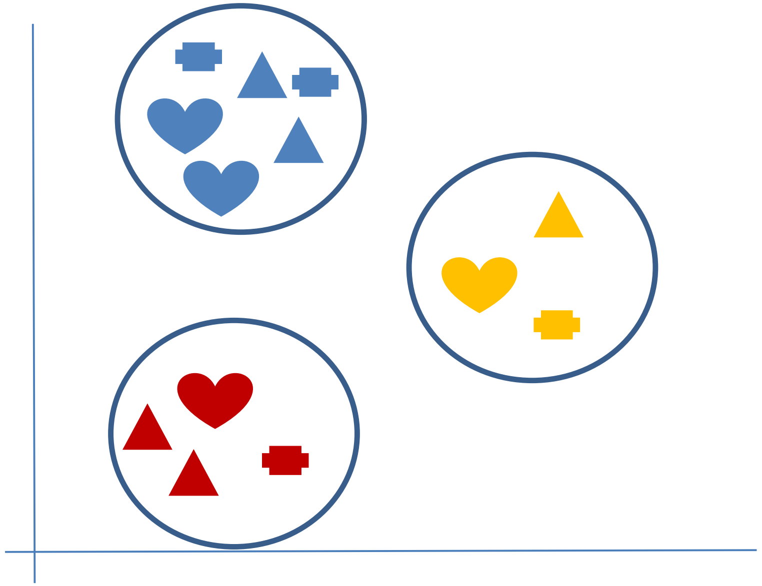


Machine Learning

- Supervised methods



- Unsupervised methods



Unsupervised Methods

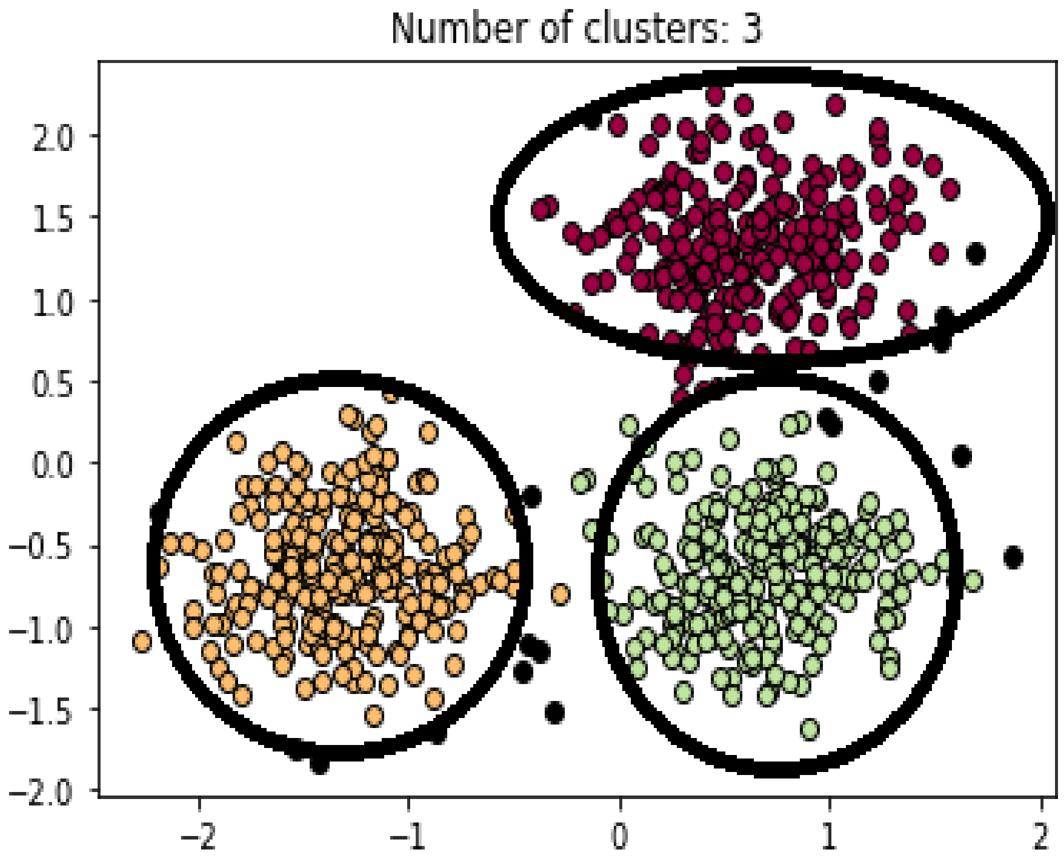
- Unsupervised methods involve no training
  - Used to discover patterns / relationships in data and in understanding the data
  - The data given to unsupervised algorithm are not labelled, which means only the input variables are given
  - Algorithms are left to themselves to discover interesting structures in the data
  - For example groups of customers with similar purchase patterns or correlations between population movement and socioeconomic factors
- The main approach that we will look at is clustering

Cluster Analysis

- Goal is to group observations in data into cluster so that the items in the same cluster are more similar to each other than items in other clusters
- For example company that offers holidays might want to cluster clients behaviour and tastes by:
  - Which sites / countries they like to visit or kind of activities they participate in
  - Whether they prefer adventure, luxury, beach or educational holidays
- This might help the company design attractive packages and target appropriate segments of their client base

Cluster Analysis

- We want to find the regions of the space where the data is densest
- If those regions are distinct or nearly distinct then we have clusters



Distance

- In order to cluster we need notions of similarity and dissimilarity
- In terms of distance, points in the same cluster are / should be closer to each other than to points in other clusters
- Common distance measures:
  - Euclidean
  - Manhattan
  - Cosine
  - Hamming

Distance

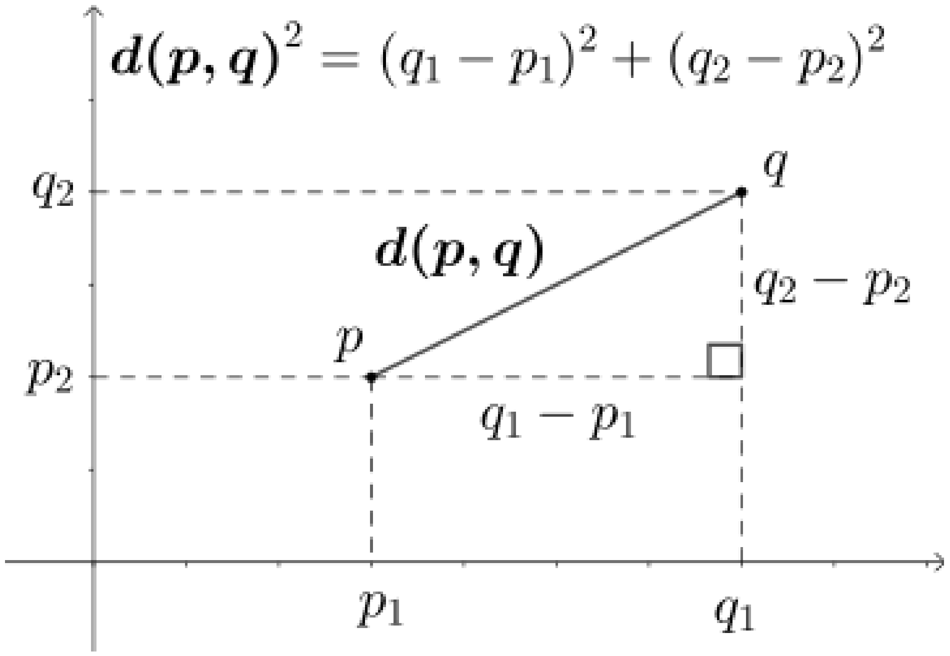
- Different distance metrics will give different clusters, as will different clustering algorithms
  - Application domain may determine what is chosen
  - Or trial and error

Euclidean Distance

- Common and simple measure
- The Euclidean distance between two vectors x and y is:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

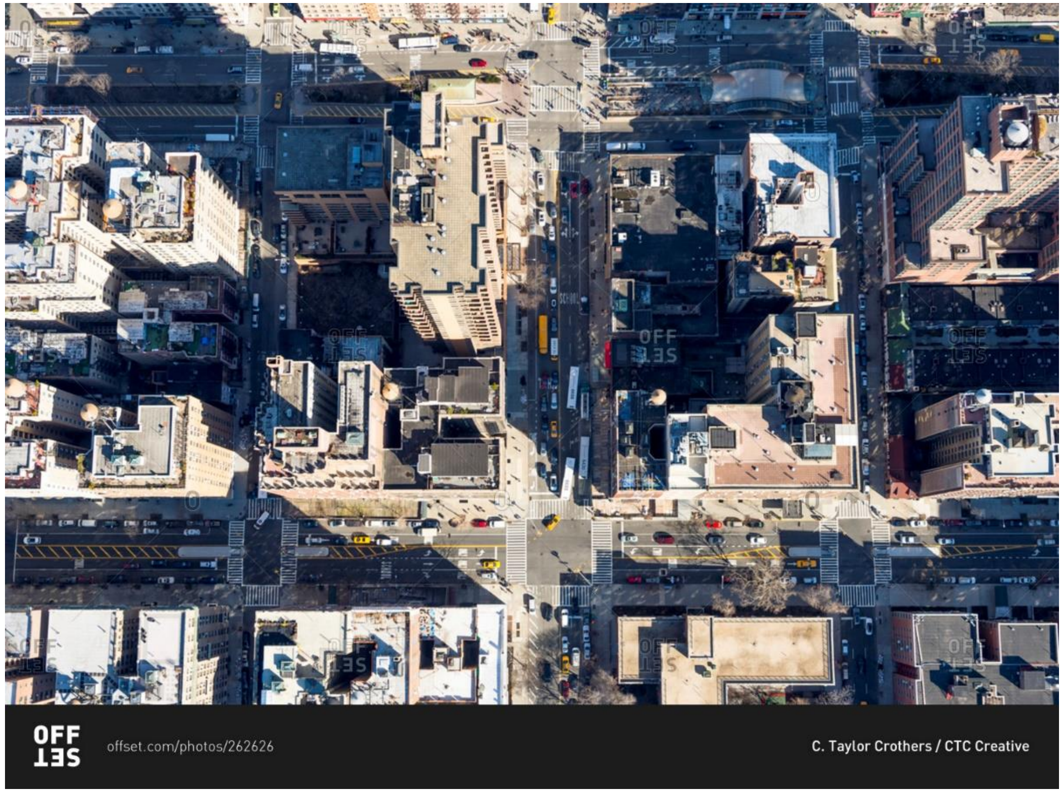
- Makes sense when all data is real-valued (quantitative)



Manhattan Distance

- Manhattan Distance measure is the number of horizontal and vertical moves it takes to get from one point to another
- No diagonal moves
- The Manhattan Distance between two vectors x and y is:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$



Source: https://www.offset.com/photos/top-down-view-of-manhattan-city-blocks-262626

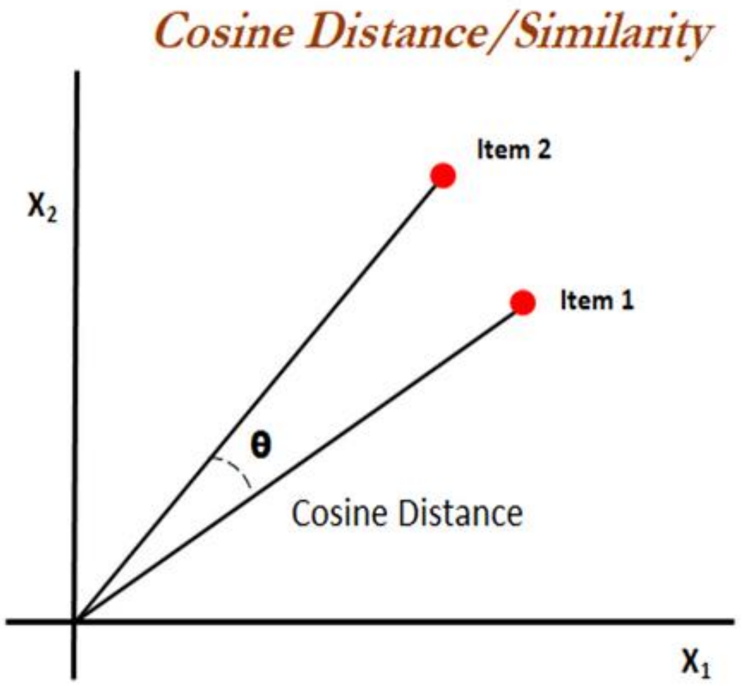
Manhattan Distance

- Manhattan Distance measure is the number of horizontal and vertical moves it takes to get from one point to another
- No diagonal moves
- The Manhattan Distance between two vectors x and y is:

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Cosine Matching

- Cosine matching is a method commonly used in text analysis e.g. search systems
- Measures the angle between two vectors
- Two perpendicular vectors are furthest apart (Cosine(90) = 0)
- Two parallel are the most similar (Cosine(0) = 1)



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Source: https://www.tyrrell4innovation.ca/miword-of-the-day-iscosine-distance/

Hamming Distance

- For categorical variables (small/medium/large), you can define the distance as 0 if two points are in the same category and 1 otherwise
- In Information Theory, the Hamming Distance between 2 strings of equal length is the number of positions at which corresponding symbols are different
- Can extend to non-binary categories
- In this example, Hamming distance is 3

Fruit	Sphere	Sweet	Sour	Crunchy
Apple	Y	Y	Y	Y
Banana	N	Y	N	N
Calculation				
Is different	Y	N	Y	Y
Value	1	0	1	1

This example and more details from:
http://people.revoledu.com/kardi/tutorial/Similarity/HammingDistance.html

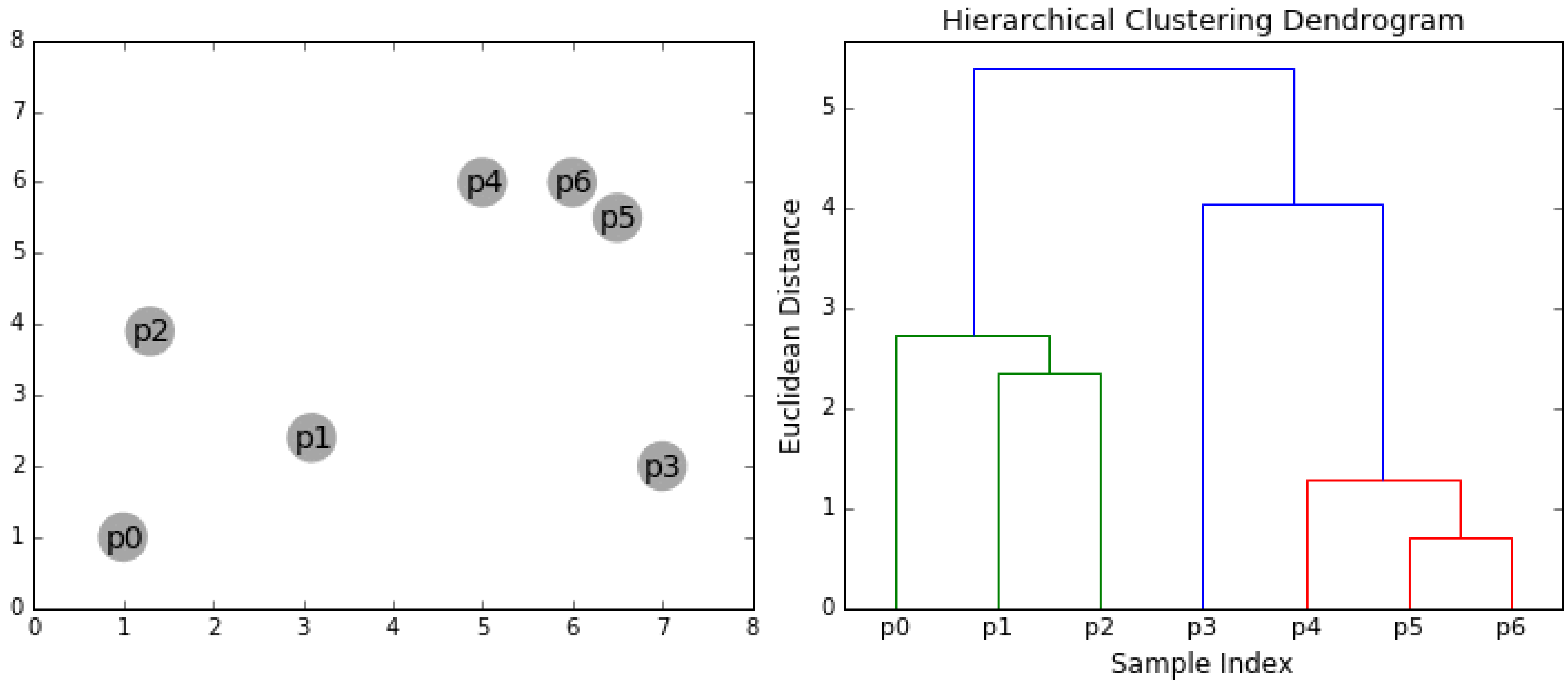
Preparing the Data - Scaling

- Units (or disparity in units) impact what clusters an algorithm will discover
- Ideally you want a unit of change in each coordinate to represent the same degree of difference
- One approach is to transform all columns to have a mean value of 0 and a standard deviation of 1
- Make the standard deviation the unit of measurement in each coordinate

Hierarchical Clustering

- Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters
- There are 2 types
  - Agglomerative, "bottom up" approach
  - Divisive, "top down" approach
- In general, the merges and splits are determined in a greedy manner
- The results of hierarchical clustering are usually presented in a dendrogram

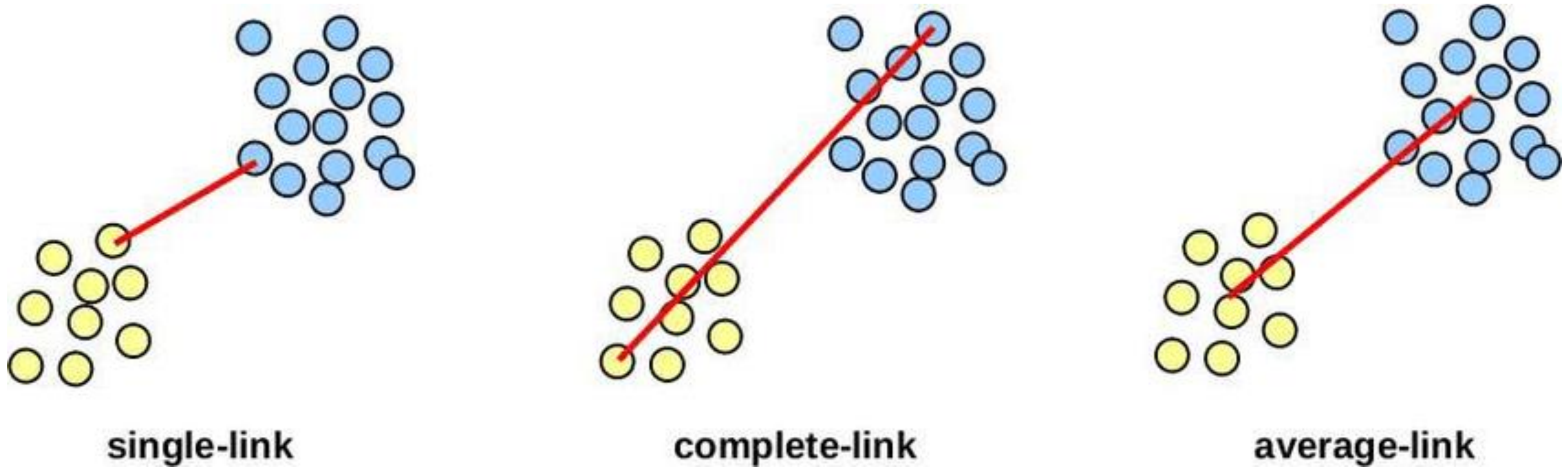
Hierarchical Clustering



Source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Linkage

- Determines how objects should be joined or divided
  - Single uses the minimum distances between all observations of the two sets
  - Average uses the average of the distances of each observation of the two sets
  - Complete uses the maximum distances between all observations of the two sets



Source: https://www.drive5.com/usearch/manual/linkage.html

Number of Clusters

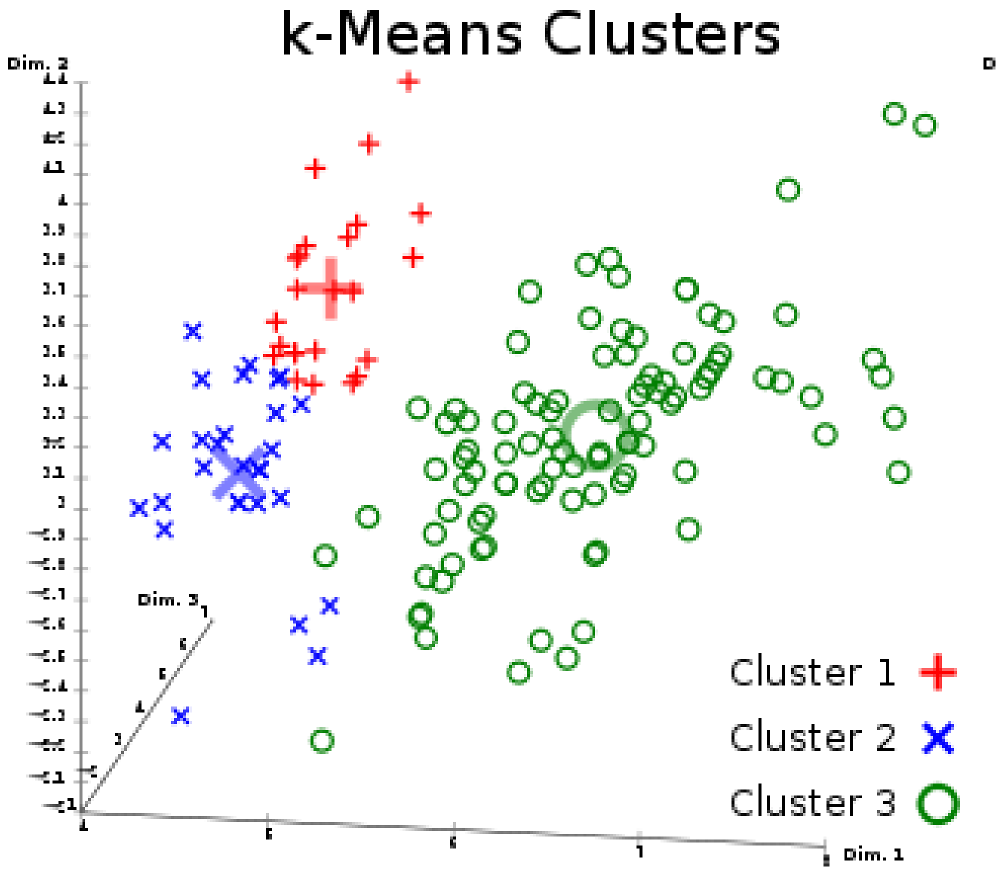
- Sometimes there is no problem specifying the number of clusters in advance e.g. segmenting a client database into X clusters for X salesman
- Sometimes the cut off is implicit in stopping at a certain point e.g. placing cell phone towers
- However in most exploratory applications, the number of clusters is not known in advance
- So how do we decide how many clusters there should be?

Number of Clusters

- Silhouette Score
  - Measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)
  - Value ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters
  - High values indicate appropriate clustering
- Calinski-Harabasz Index
  - Aims to trade off between within cluster variation and between cluster variation

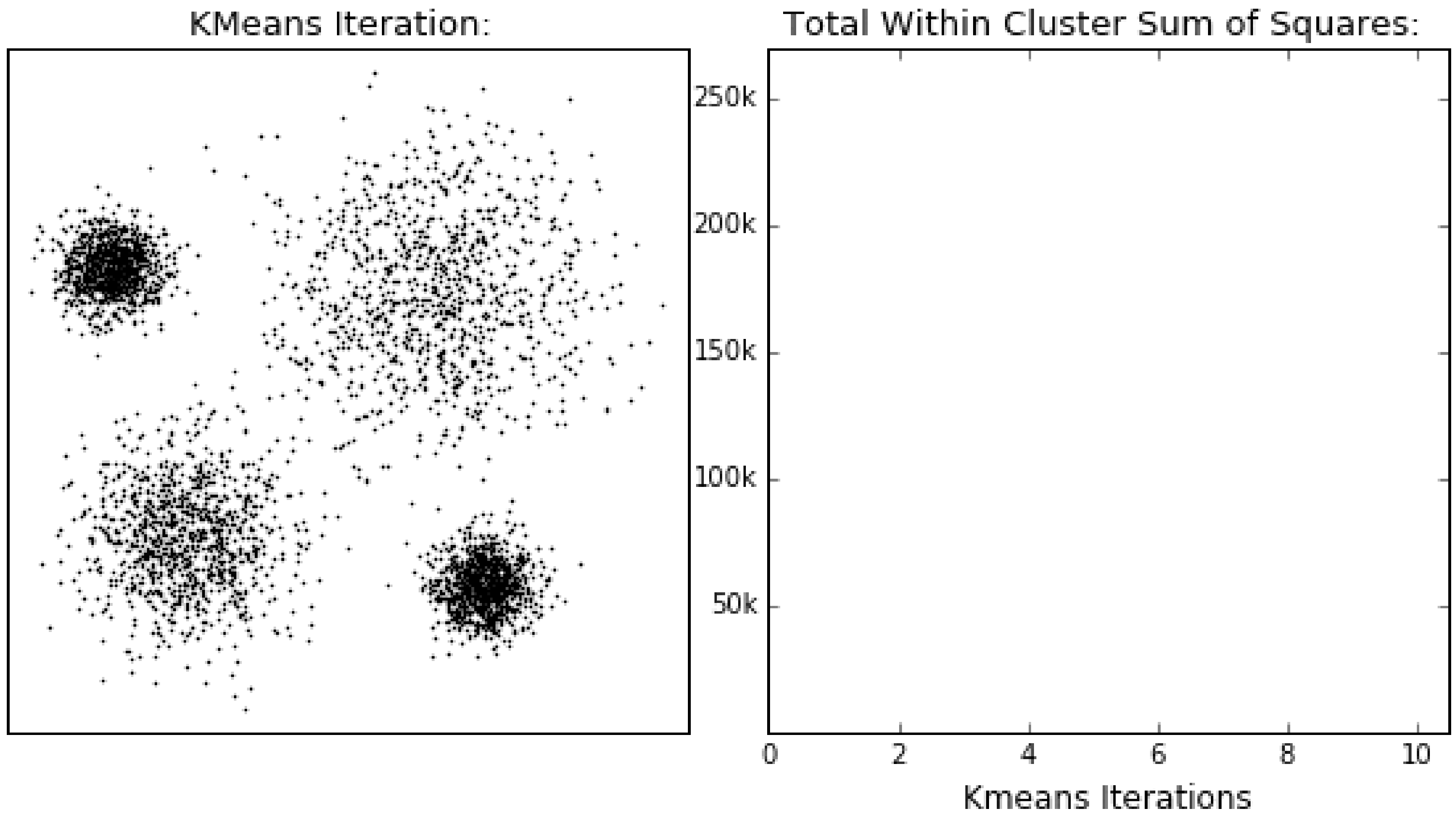
K-Means Clustering

- Alternative to hierarchical clustering
- Need to specify number of clusters



This image is released to the public domain
https://commons.wikimedia.org/wiki/File:Iris_Flowers_Clustering_kMeans.svg

K-Means Clustering



Source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Clustering Using K-Means

- Algorithm is not guaranteed to have a unique stopping point
- Final clusters depend on initial cluster centres
- Can run K-means several times with different random starts and then select the best results

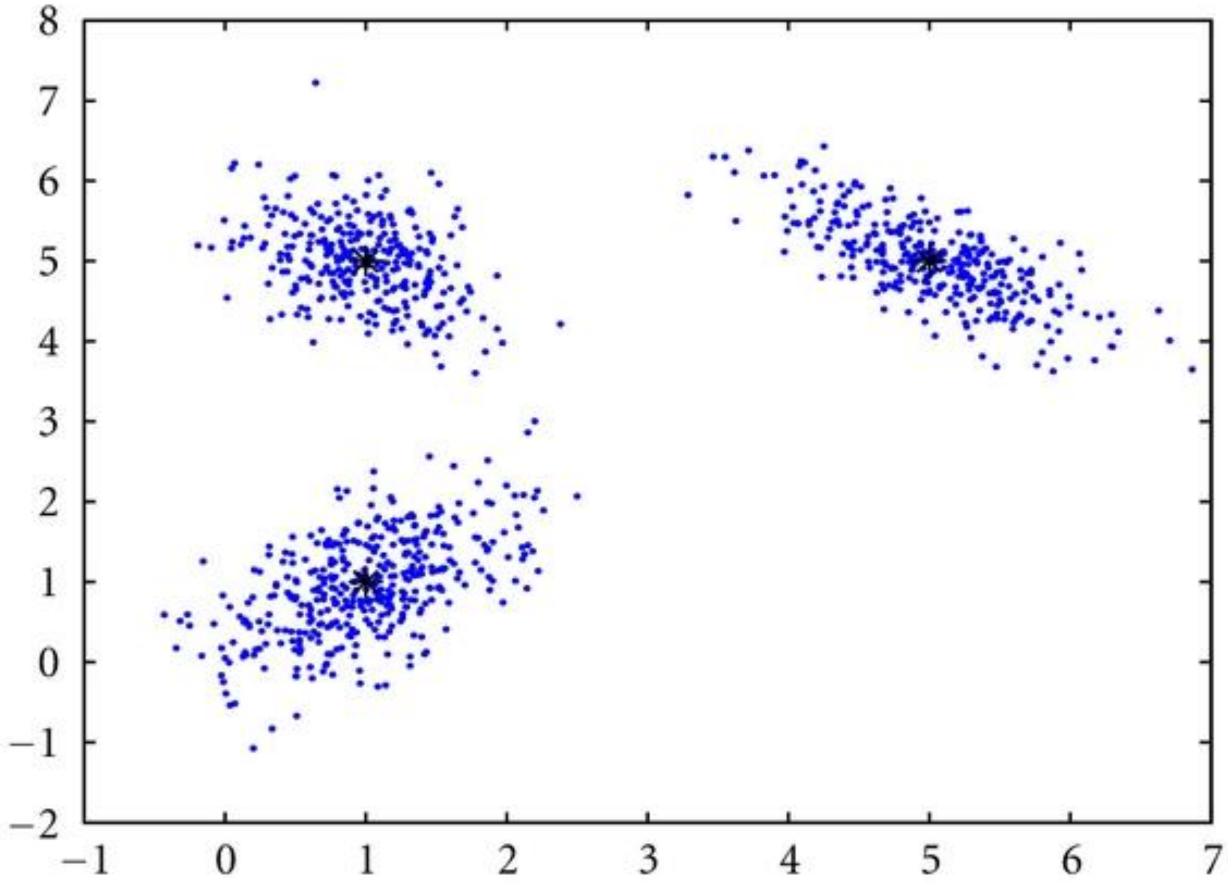
Evaluation of clustering

- Calinski-Harabasz Index
- Silhouette Score
- Homogeneity
  - A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class
- Completeness
  - A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster

Summary

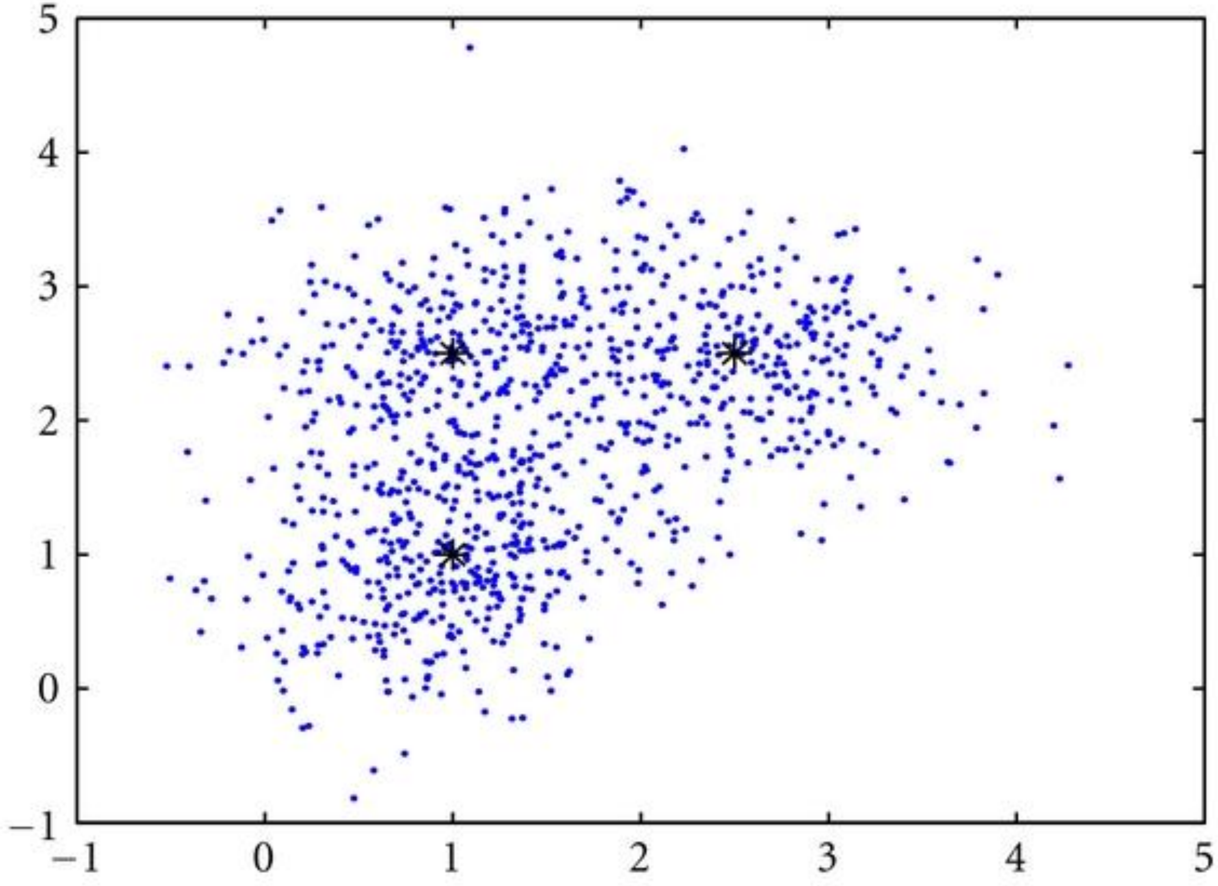
- Goal of clustering is to discover or draw out similarities among subsets of your data
- Points in the same cluster should be more similar to each other than they are to points in other clusters
- Scaling data may be necessary
- Clustering is an iterative process and often used for data exploration or as a precursor to supervised learning methods

Discussion



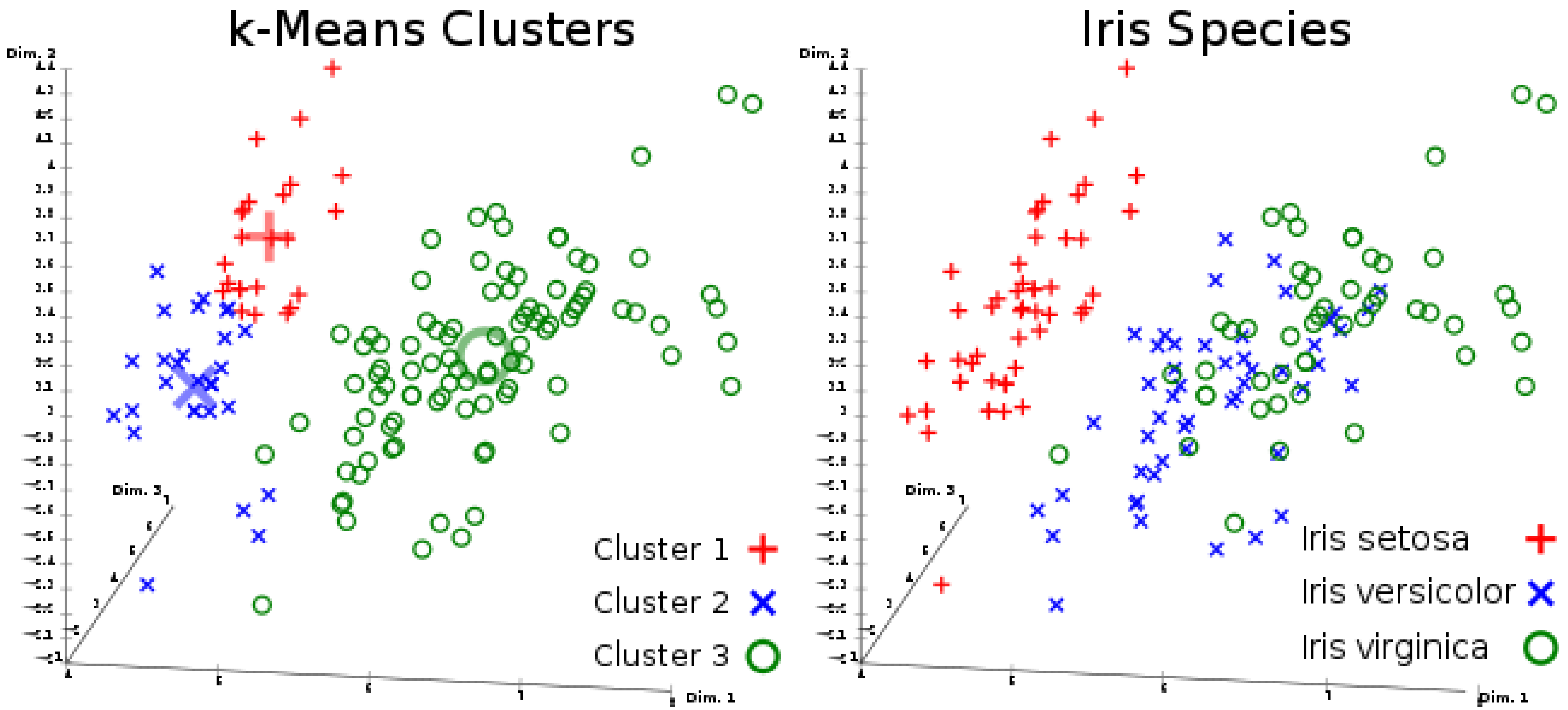
Source: https://developers.google.com/machine-learning/clustering/interpret

Discussion



Source: https://developers.google.com/machine-learning/clustering/interpret

Discussion



This image is released to the public domain
https://commons.wikimedia.org/wiki/File:Iris_Flowers_Clustering_kMeans.svg

Discussion

- Topic
  - Interpret Results of Unsupervised Learning - Clustering

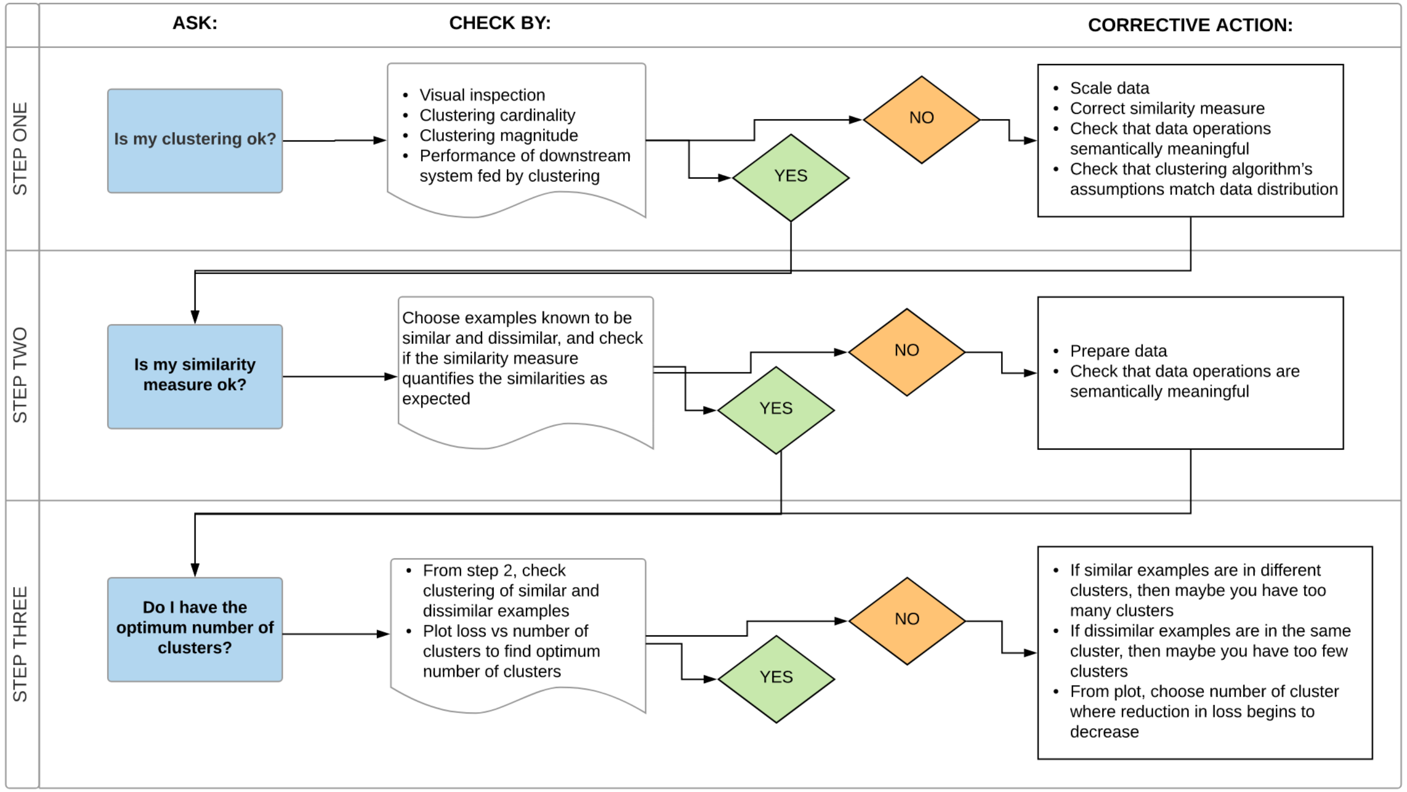
- Things to consider:
  - No right or wrong answer
  - Discuss and agree (or disagree)

Discussion

- What Is Good Clustering?
  - A good clustering method will produce high quality clusters in which the intra-class (that is, intra-cluster) similarity is high and the inter-class similarity is low
  - The quality of a clustering result also depends on both the similarity measure used by the method and its implementation
  - The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns
  - However, objective evaluation is problematic: usually done by human / expert inspection

Discussion

➢ Interpret Results and Adjust Clustering



Source: https://developers.google.com/machine-learning/clustering/interpret

University of
Strathclyde
Glasgow