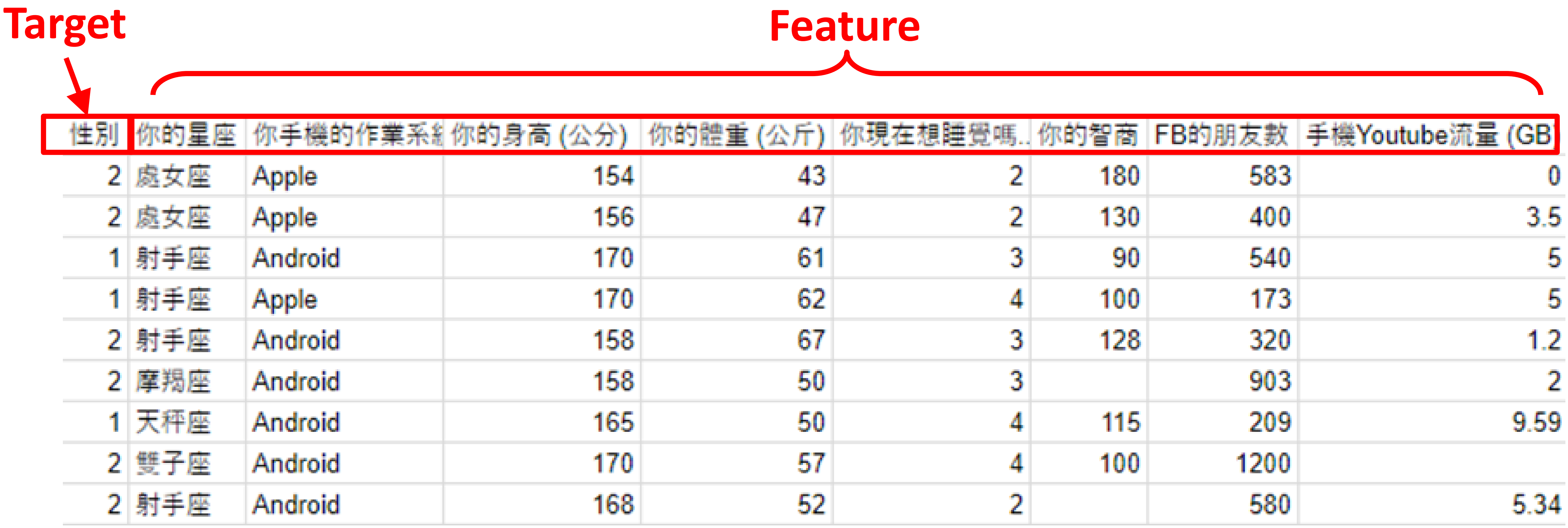
Feature Engineering & Clustering

DATA SCIENCE & MACHINE LEARNING

Feature Engineering

Given we are dealing with a classification problem and the target is the gender.



Target

Feature

性別	你的星座	你手機的作業系	你的身高 (公分)	你的體重 (公斤)	你現在想睡覺嗎..	你的智商	FB的朋友數	手機Youtube流量 (GB)
2	處女座	Apple	154	43	2	180	583	0
2	處女座	Apple	156	47	2	130	400	3.5
1	射手座	Android	170	61	3	90	540	5
1	射手座	Apple	170	62	4	100	173	5
2	射手座	Android	158	67	3	128	320	1.2
2	摩羯座	Android	158	50	3		903	2
1	天秤座	Android	165	50	4	115	209	9.59
2	雙子座	Android	170	57	4	100	1200	
2	射手座	Android	168	52	2		580	5.34

Feature Engineering

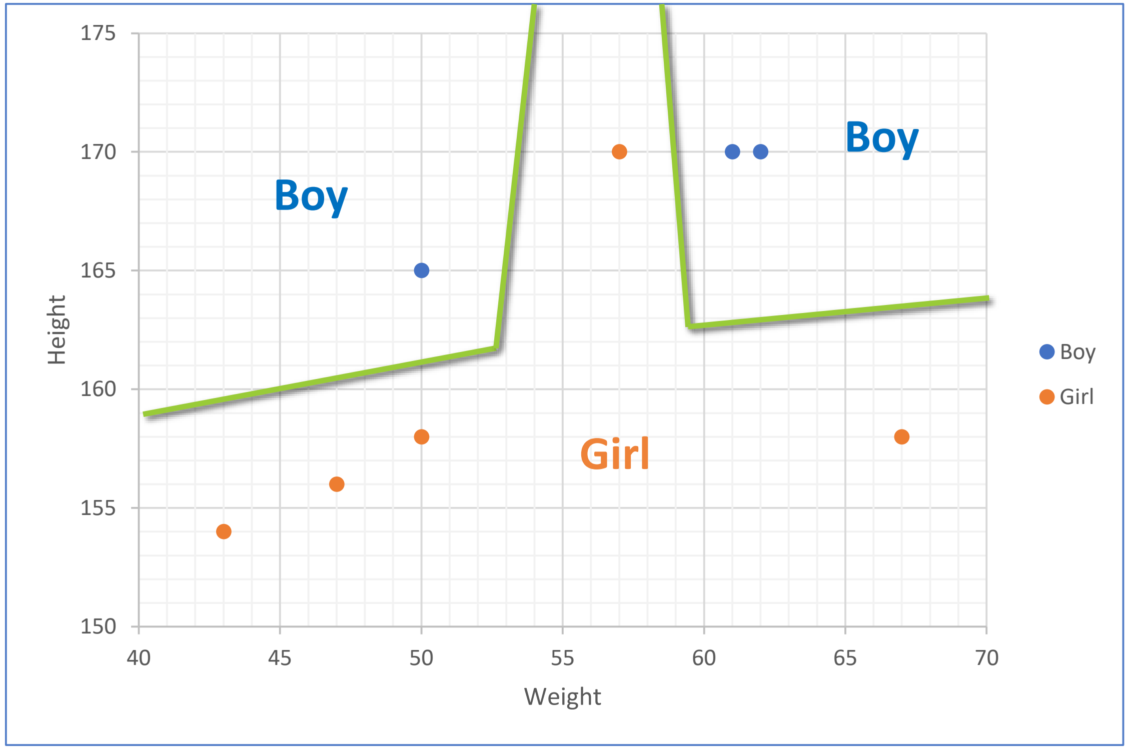
Regardless of supervised or unsupervised learning models, they also perform their “jobs” on features.

For supervised learning models, e.g. classification or regression models, they learn to associate features with targets/labels.

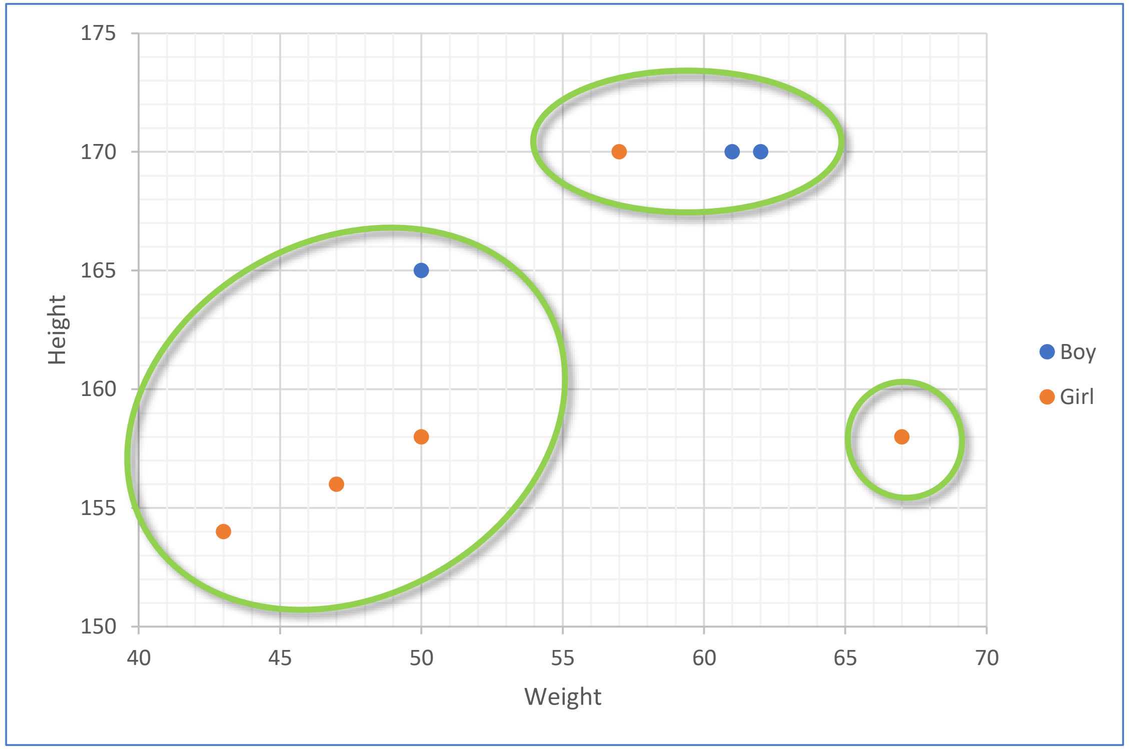
For unsupervised learning models, e.g. clustering, they try to group similar instances together based on features.

Feature Engineering

Classification



Clustering



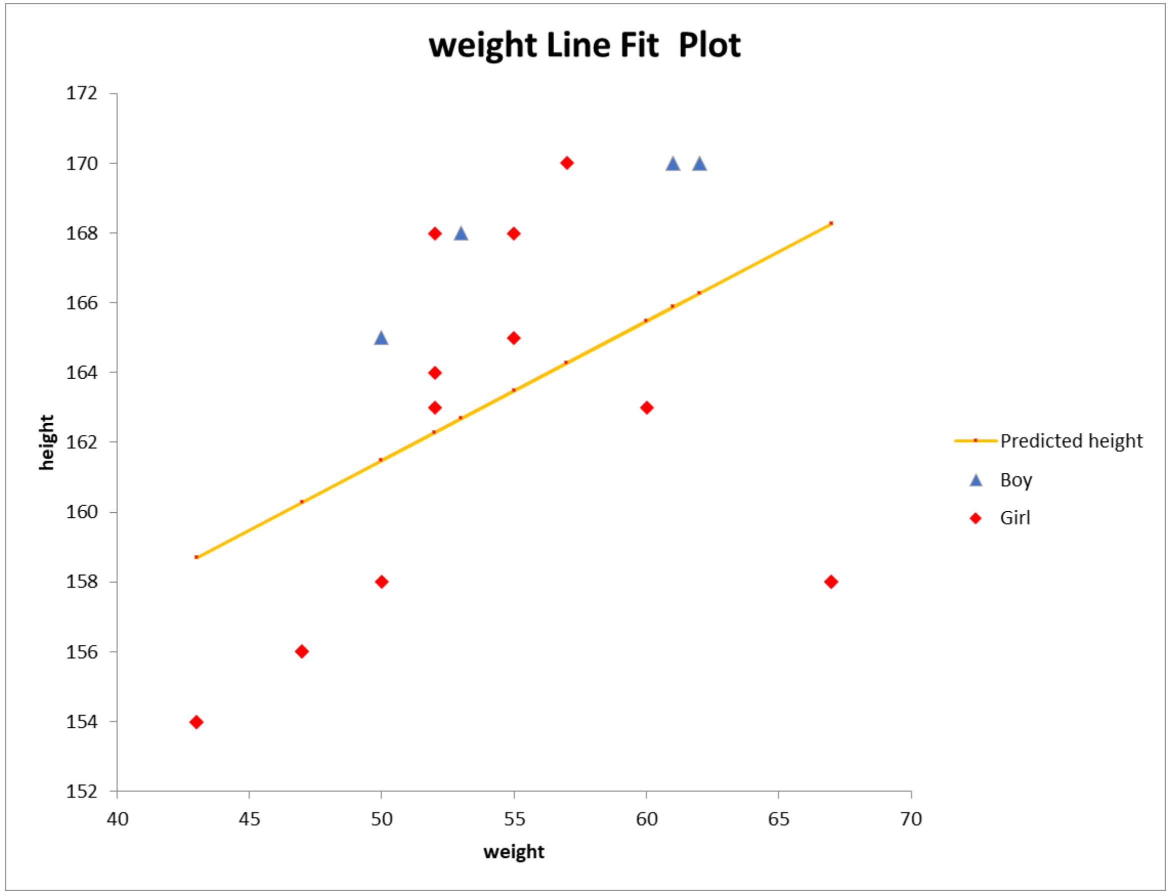
Feature Engineering

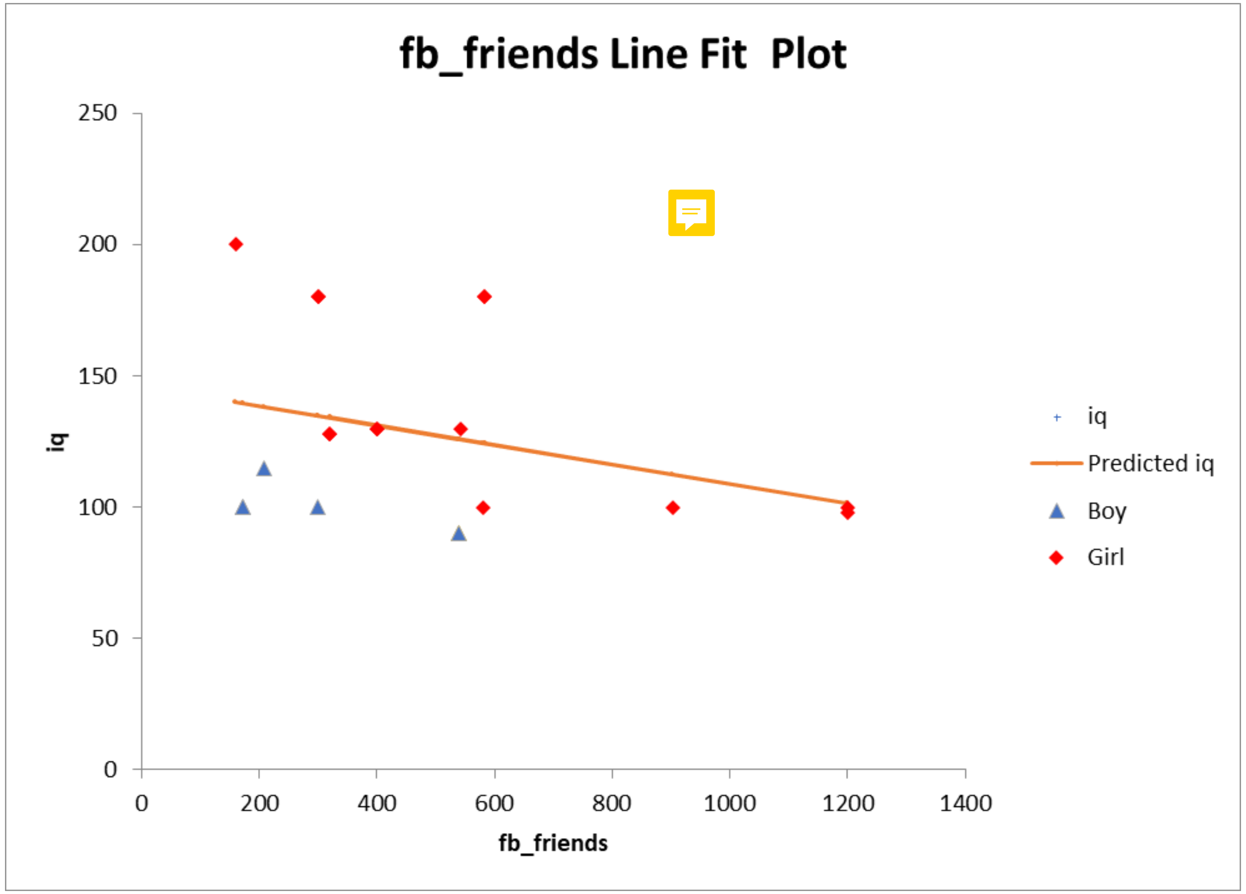
The goal of feature engineering is simply to make your data better suited to the problem at hand.

For a feature to be useful, it must have a ***relationship*** to the target that your model is able to learn.

Ref: https://www.kaggle.com/ryanholbrook/what-is-feature-engineering

Feature Engineering





Feature Engineering

Reasons to apply feature engineering:

- improve a model's predictive performance
- reduce computational or data needs
- improve interpretability of the results

Ref: https://www.kaggle.com/ryanholbrook/what-is-feature-engineering

Feature Engineering

- Feature selection
  - Filter
  - Wrapper
  - Embedded
- Feature extraction
  - PCA: Principle Component Analysis
  - t-SNE: t-distributed Stochastic Neighboring Embedding
  - Autoencoder
  - LDA: Linear Discriminant Analysis
  - … and many more

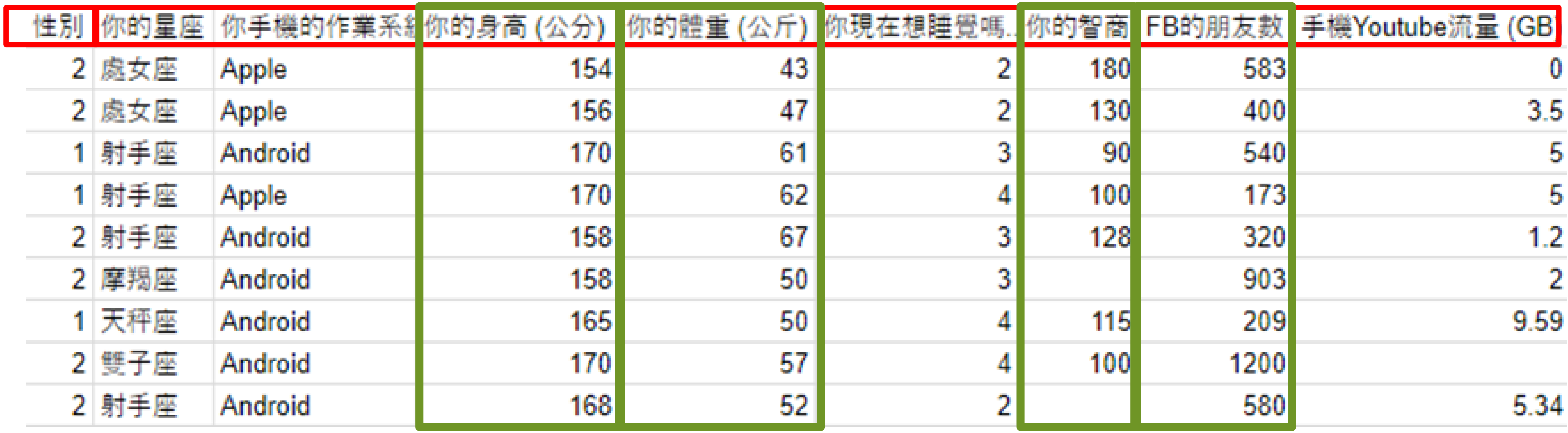
- Feature Engineering techniques*
  - Imputation
  - Handling Outliers
  - Binning
  - Log Transform
  - One-Hot Encoding
  - Grouping Operations
  - Scaling

*https://www.analyticsvidhya.com/blog/2020/10/7-feature-engineering-techniques-machine-learning/

Feature Selection

Use a subset of features *t* from *T*, where |*t*| < |*T*| and |*T*| is the dimension of the data, according to some form of measures or sometimes “hunch”.

Mutual Information



性別	你的星座	你手機的作業系	你的身高 (公分)	你的體重 (公斤)	你現在想睡覺嗎..	你的智商	FB的朋友數	手機Youtube流量 (GB
2	處女座	Apple	154	43	2	180	583	0
2	處女座	Apple	156	47	2	130	400	3.5
1	射手座	Android	170	61	3	90	540	5
1	射手座	Apple	170	62	4	100	173	5
2	射手座	Android	158	67	3	128	320	1.2
2	摩羯座	Android	158	50	3		903	2
1	天秤座	Android	165	50	4	115	209	9.59
2	雙子座	Android	170	57	4	100	1200	
2	射手座	Android	168	52	2		580	5.34

Feature Selection

FILTER METHODS

“Filter methods select features based on a performance measure regardless of the employed data modeling algorithm. Only after the best features are found, the modeling algorithms can use them. Filter methods can rank individual features or evaluate entire feature subsets.”

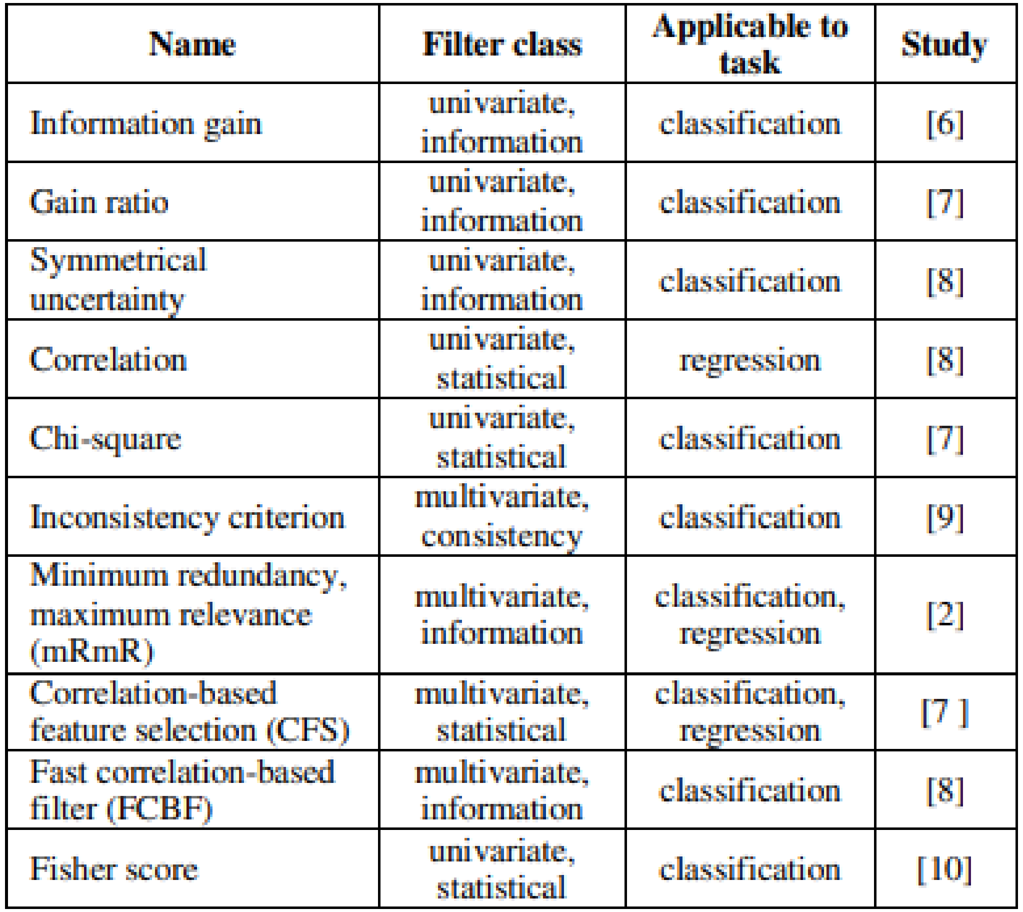
WRAPPER METHODS

“Wrappers consider feature subsets by the quality of the performance on a modelling algorithm, which is taken as a black box evaluator. Thus, for classification tasks, a wrapper will evaluate subsets based on the classifier performance (e.g. Naïve Bayes or SVM) [18,19], while for clustering, a wrapper will evaluate subsets based on the performance of a clustering algorithm (e.g. K-means) [20].”

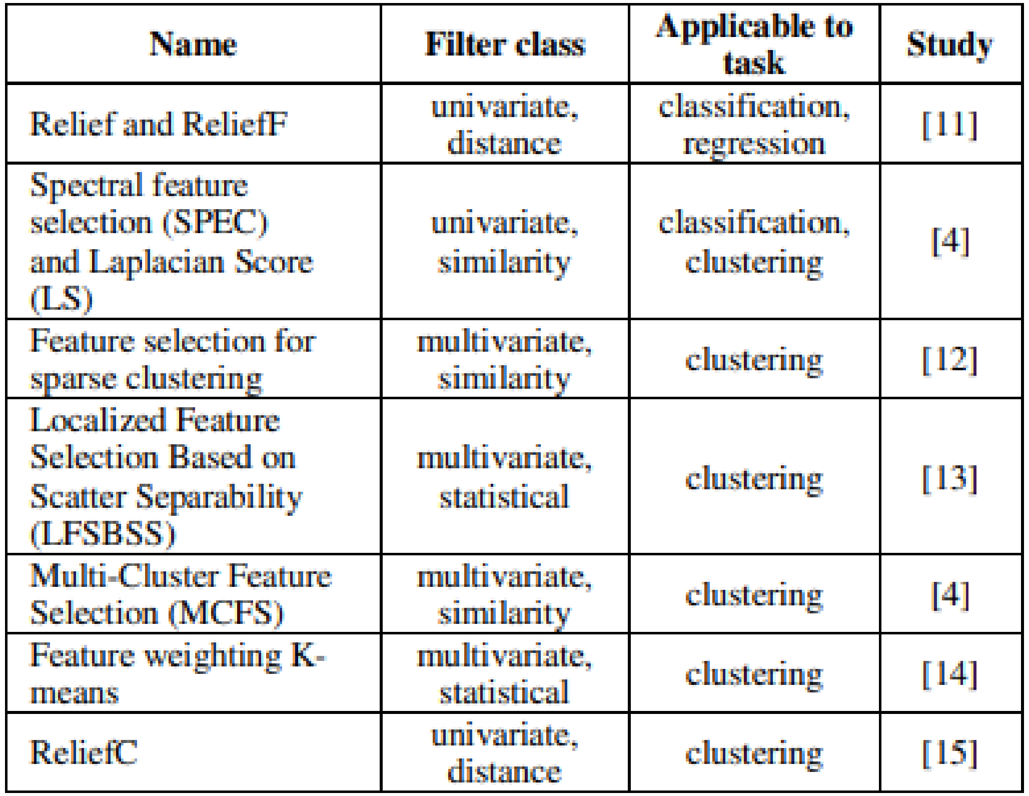
A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

Feature Selection

Refer to the article at the bottom for numbered references.



Name	Filter class	Applicable to task	Study
Information gain	univariate, information	classification	[6]
Gain ratio	univariate, information	classification	[7]
Symmetrical uncertainty	univariate, information	classification	[8]
Correlation	univariate, statistical	regression	[8]
Chi-square	univariate, statistical	classification	[7]
Inconsistency criterion	multivariate, consistency	classification	[9]
Minimum redundancy, maximum relevance (mRmR)	multivariate, information	classification, regression	[2]
Correlation-based feature selection (CFS)	multivariate, statistical	classification, regression	[7]
Fast correlation-based filter (FCBF)	multivariate, information	classification	[8]
Fisher score	univariate, statistical	classification	[10]



Name	Filter class	Applicable to task	Study
Relief and ReliefF	univariate, distance	classification, regression	[11]
Spectral feature selection (SPEC) and Laplacian Score (LS)	univariate, similarity	classification, clustering	[4]
Feature selection for sparse clustering	multivariate, similarity	clustering	[12]
Localized Feature Selection Based on Scatter Separability (LFSBSS)	multivariate, statistical	clustering	[13]
Multi-Cluster Feature Selection (MCFS)	multivariate, similarity	clustering	[4]
Feature weighting K-means	multivariate, statistical	clustering	[14]
ReliefC	univariate, distance	clustering	[15]

A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

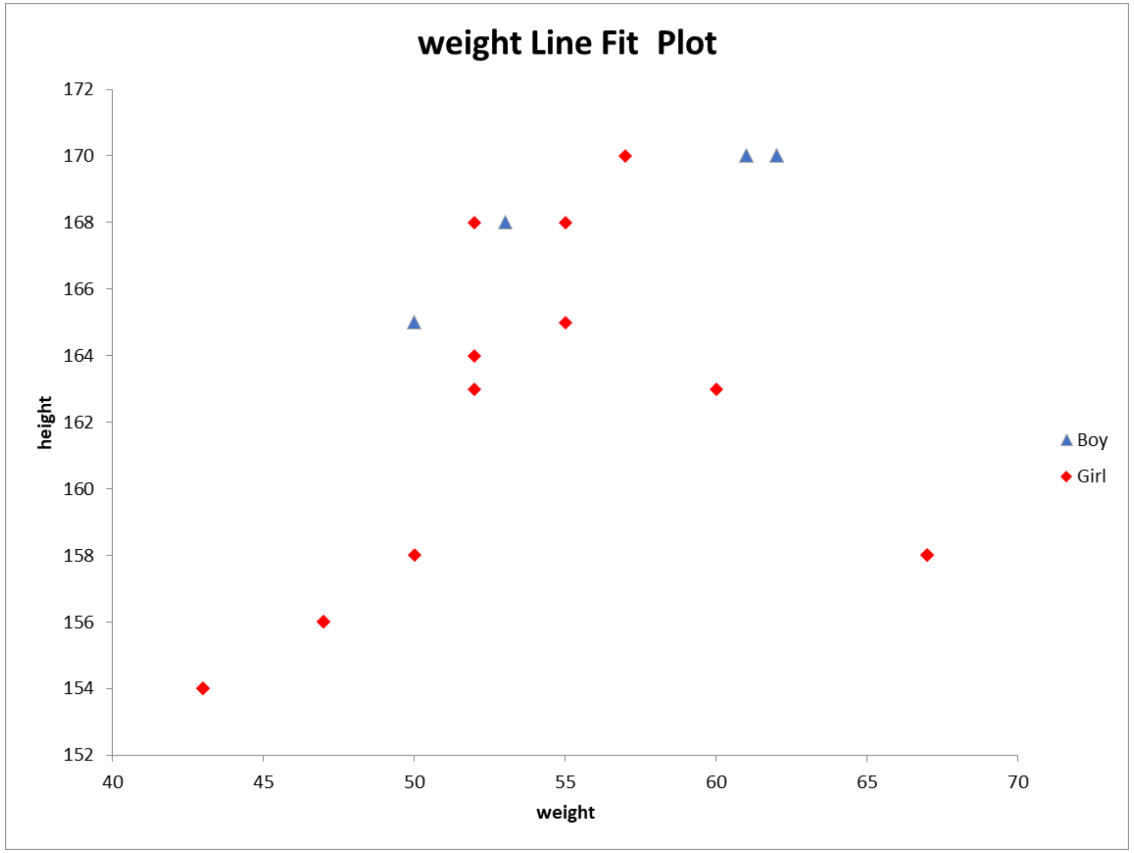
Feature Extraction: PCA

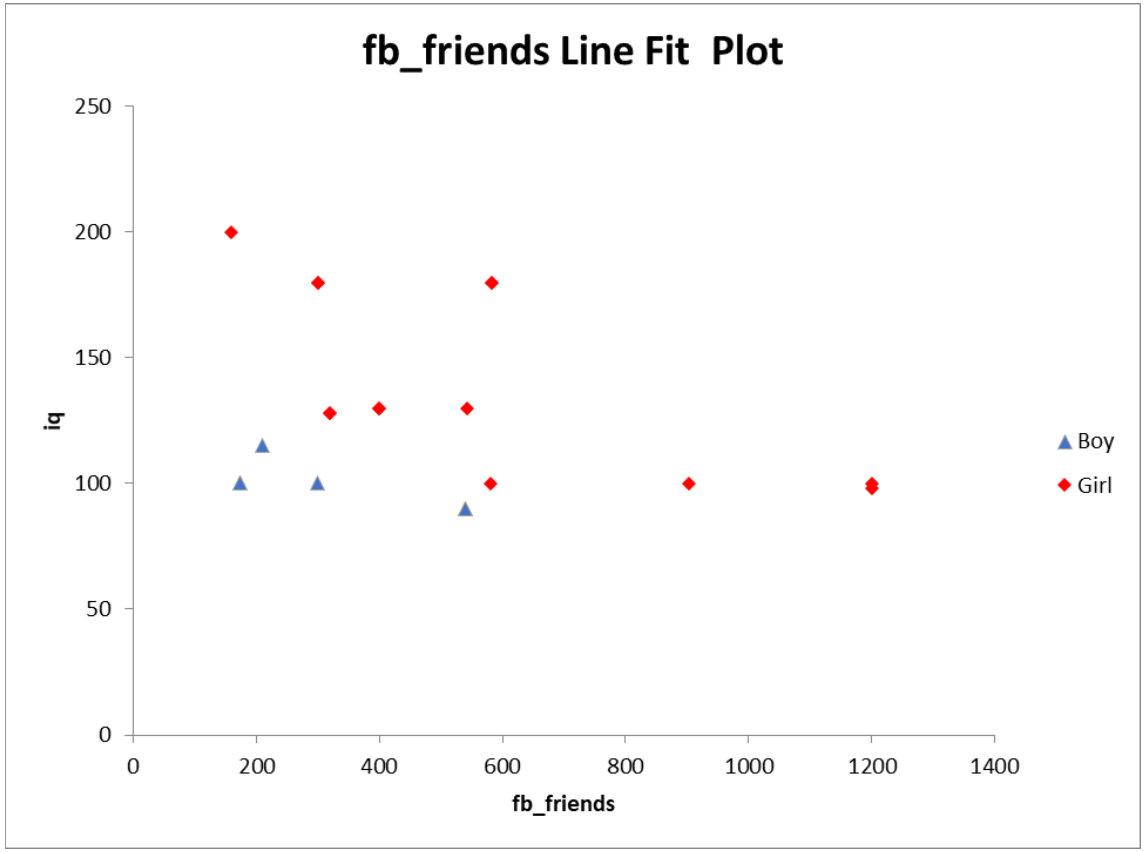
- PCA: Principle Component Analysis is a linear transformation algorithm that seeks to project the original features of our data onto a smaller set of features (or **subspace**) while still retaining most of the information. To do this the algorithm tries to find the most appropriate directions/angles (which are the principal components) that maximise the variance in the new subspace.

- https://setosa.io/ev/principal-component-analysis/
- http://www2.imm.dtu.dk/courses/02450/DemoPCA.html
- https://plotly.com/python/pca-visualization/
- https://youtu.be/iwh5o_M4BNU?t=950
- https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.09-Principal-Component-Analysis.ipynb
- https://www.kaggle.com/arthurtok/interactive-intro-to-dimensionality-reduction
- https://www.kaggle.com/ryanholbrook/principal-component-analysis

Ref: https://www.kaggle.com/arthurtok/interactive-intro-to-dimensionality-reduction

Feature Extraction: PCA



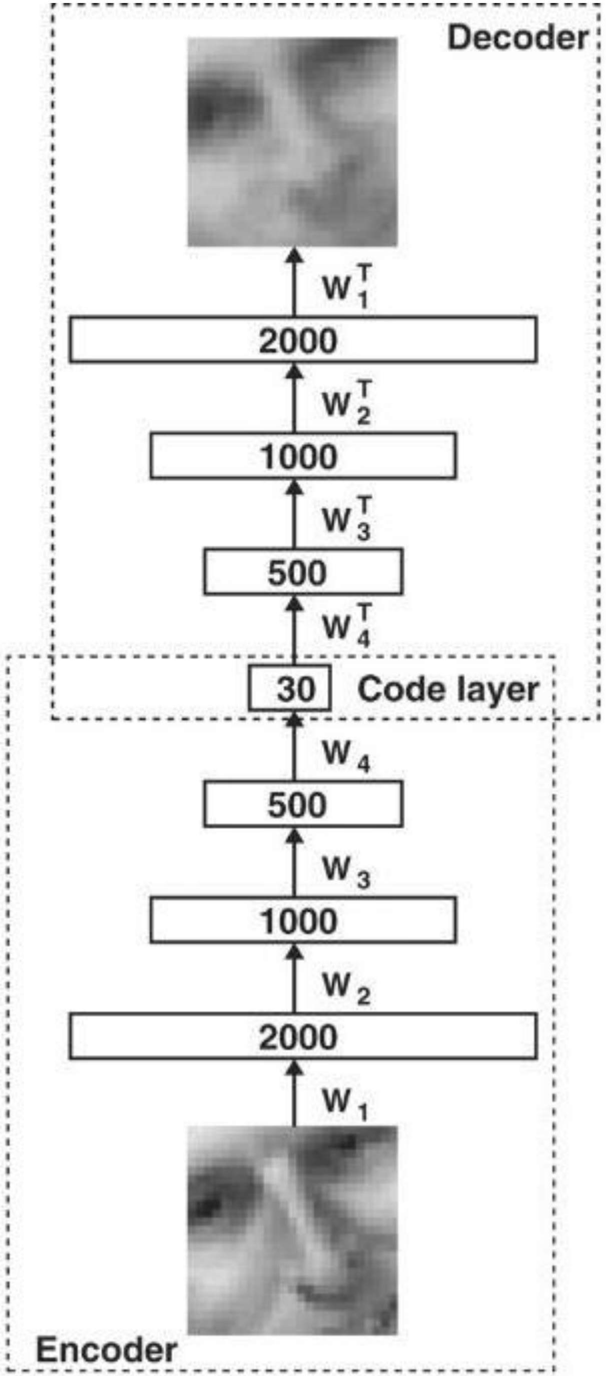


Autoencoder

Reducing the Dimensionality of Data with Neural Networks

DOI: 10.1126/science.1127647

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network to transform the high-dimensional data into a low-dimensional code and a similar "decoder" network to recover the data from the code.



Surveys / Literature Reviews

- A review of feature selection methods with applications
  - DOI: 10.1109/MIPRO.2015.7160458
- A survey of feature selection and feature extraction techniques in machine learning
  - DOI: 10.1109/SAI.2014.6918213
- Feature Extraction: A Survey of the Types, Techniques, Applications
  - DOI: 10.1109/ICSC45622.2019.8938371
- A Survey on Feature Selection
  - https://doi.org/10.1016/j.procs.2016.07.111
- A survey on deep learning feature extraction techniques
  - https://doi.org/10.1063/5.0028564
- **Google Scholar**

Clustering

Clustering tries to divide a collection of data into sub-collections, namely clusters, in a way that the data in the same cluster is similar in certain sense or share some characteristics.

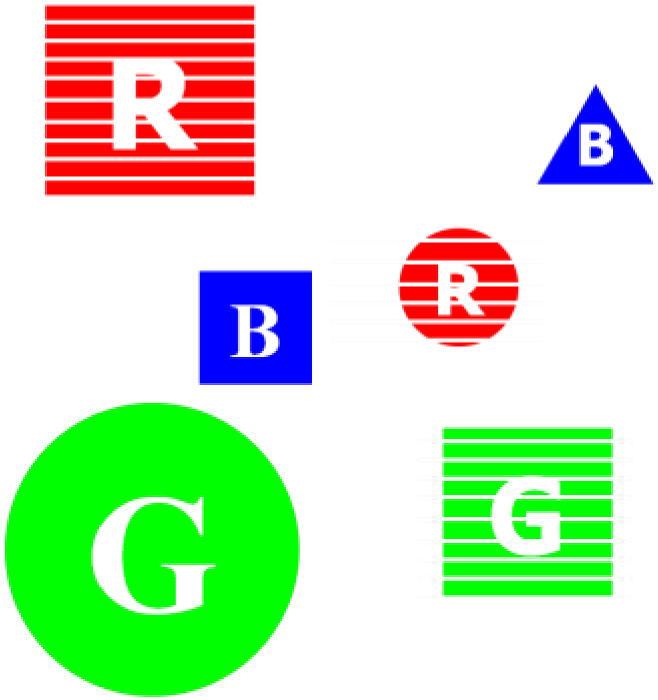
Clustering v.s. Classification

In contrast to supervised learning where data is usually labelled by human, data clustering process takes place “unsupervised” meaning that the clusters are “learned” without human experts’ involvement.

Clustering is an exploratory technique that can lead to different interpretations of the clustering results.

Clustering Task

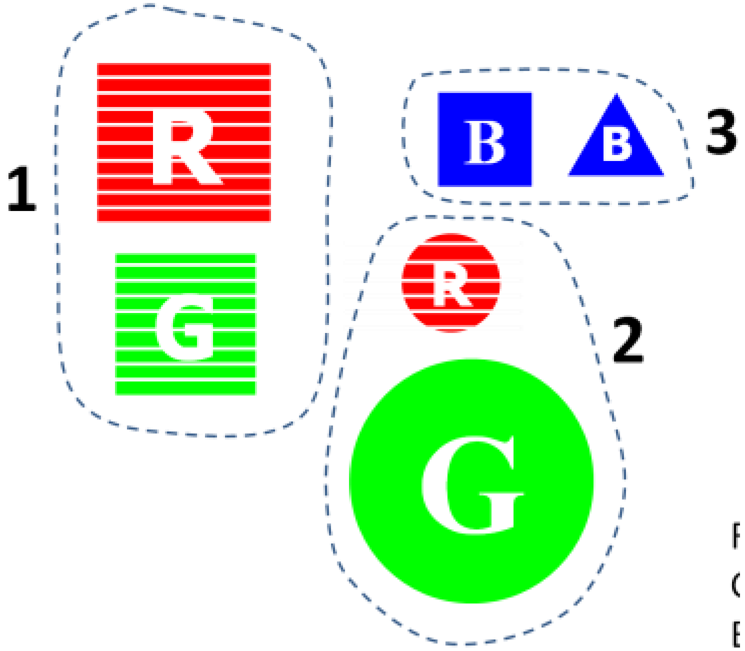
You have been asked to group the objects into meaningful groups and yet no one has told you what meaning is sought after (unsupervised).



Clustering Task

You can group objects by:

- Colour
- Shape
- Solid or striped
- Size?
- Font?



Steps of Clustering Task

Data representation

Measuring similarity

Clustering data

Abstraction of the clusters

Evaluation of the clusters

Data Representation

Data needs to be prepared and manipulated so that it is represented in a way that can be read and understood by the clustering algorithm; so the above objects can be represented by their shape, colour, background, font, and so on in a way that object 1 = {square, red, striped, Tahoma}

Data Representation

Bag-of-Words

n-Grams

Multiple word representation

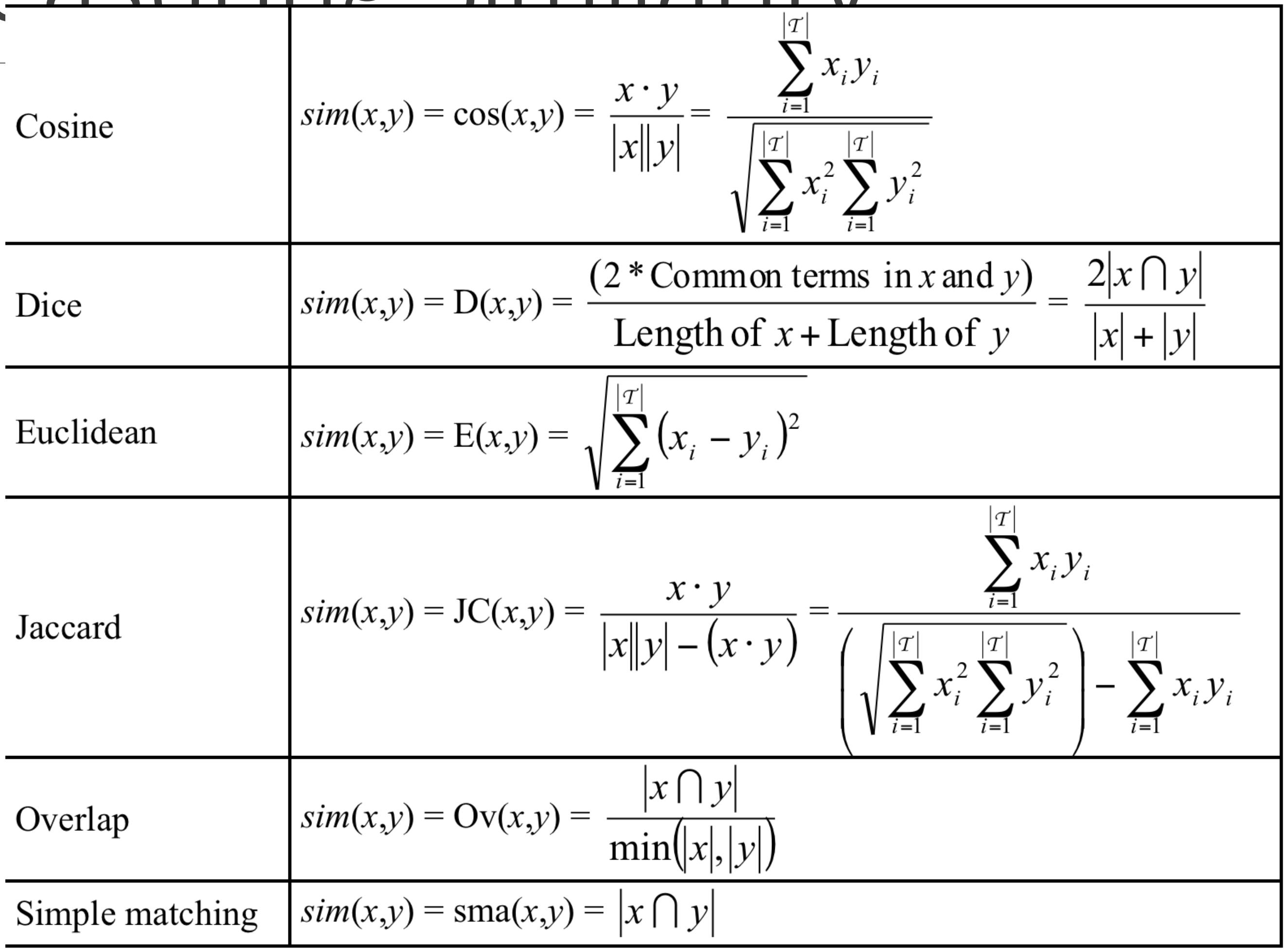
Term weight?

Measuring Similarity

The similarity between two data can be measured by a variety of functions and choosing an appropriate function is no trivial task as they can produce different results.

So the similarity measure in my head determined that the red square is more similar to the green square than to the red circle.

Measuring Similarity



Cosine	$sim(x,y) = \cos(x,y) = \frac{x \cdot y}{\lvert x\rvert\lvert y\rvert} = \frac{\sum_{i=1}^{\lvert T\rvert} x_i y_i}{\sqrt{\sum_{i=1}^{\lvert T\rvert} x_i^2 \sum_{i=1}^{\lvert T\rvert} y_i^2}}$
Dice	$sim(x,y) = \mathrm{D}(x,y) = \frac{(2 * \text{Common terms in } x \text{ and } y)}{\text{Length of } x + \text{Length of } y} = \frac{2\lvert x \cap y\rvert}{\lvert x\rvert + \lvert y\rvert}$
Euclidean	$sim(x,y) = \mathrm{E}(x,y) = \sqrt{\sum_{i=1}^{\lvert T\rvert} (x_i - y_i)^2}$
Jaccard	$sim(x,y) = \mathrm{JC}(x,y) = \frac{x \cdot y}{\lvert x\rvert\lvert y\rvert - (x \cdot y)} = \frac{\sum_{i=1}^{\lvert T\rvert} x_i y_i}{\left(\sqrt{\sum_{i=1}^{\lvert T\rvert} x_i^2 \sum_{i=1}^{\lvert T\rvert} y_i^2}\right) - \sum_{i=1}^{\lvert T\rvert} x_i y_i}$
Overlap	$sim(x,y) = \mathrm{Ov}(x,y) = \frac{\lvert x \cap y\rvert}{\min(\lvert x\rvert, \lvert y\rvert)}$
Simple matching	$sim(x,y) = \mathrm{sma}(x,y) = \lvert x \cap y\rvert$

x and *y* denote the document vectors of two separate documents, *sim*(*x*,*y*) denotes the similarity between *x* and *y* and |T| denotes the dimensionality of the *x* and *y*.

Measuring Similarity

The simplicity and intuitiveness of Euclidean distance measure made it the most popular similarity metric in text clustering (Jain et al., 1999) while Cosine similarity measure is commonly used with vector space model (Zhao and Karypis, 2002) and found to be the most effective one in text classification (Tresch et al., 1995; Joachims, 1998).

Clustering Data

The data collection is now divided or clustered into sub-collections according to the similarity of the data, and this clustering process can be performed in a number of different ways.

In the above task, I decided that each object exclusively belong to only one group.

Abstraction of the Clusters

This step is optional.

It provides an abstract or compact description of each cluster. For example, the most frequent words in a cluster can provide clues on what this cluster is about.

So the abstraction of group 1 can be “square with striped background” in this case although I did not have to derive such information from the cluster’s content.

Evaluation of the Clusters

This step is carried out to assess the quality of the resulted clusters in terms of their meaningfulness and the distinction between the clusters.

I could have evaluated my clustering by asking if someone agrees with it

Hard v.s. Soft Clustering

A hard clustering algorithm assigns each observation to only one cluster in its outcome. That is, each observation can belong to only and exactly one cluster.

A soft clustering (a.k.a. fuzzy clustering) algorithm computes the likelihoods, called degrees of membership, of each observation belonging to all clusters.

- One can assign an observation to multiple clusters if the degrees of membership of this observation to these clusters are greater than a pre-determined threshold.

Hard v.s. Soft Clustering

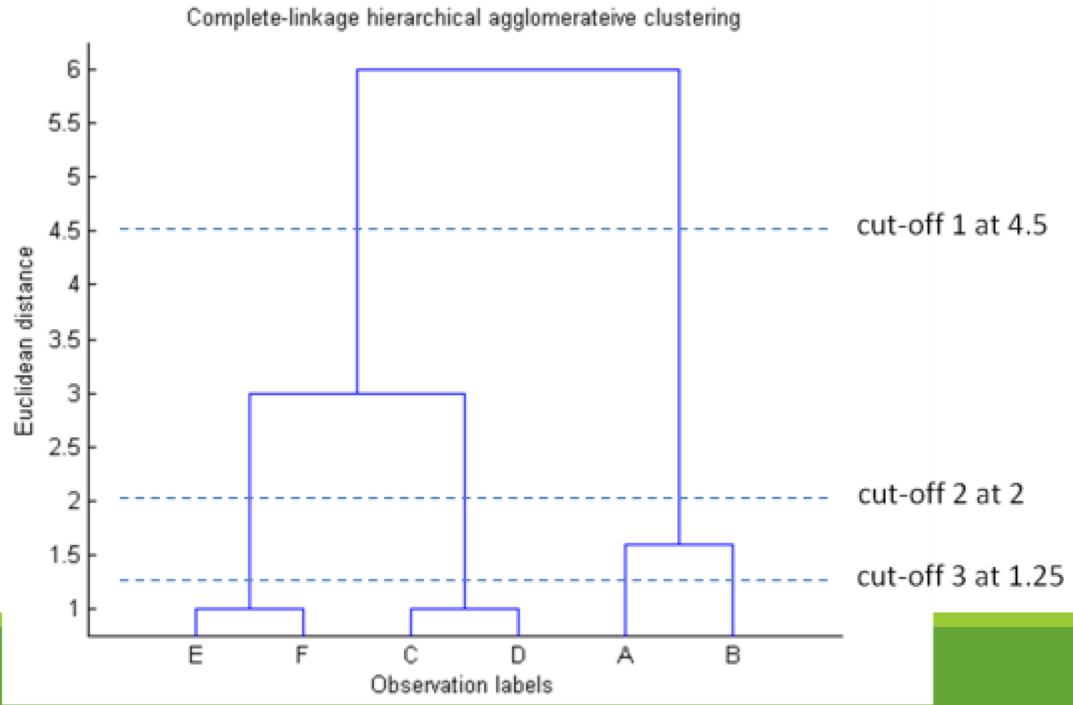
Choosing between hard and soft clustering is entirely up to the experimenters.

One application that can benefit from soft clustering is document searching and browsing where the system can return more documents if the query is ambiguous (Lin and Kondadadi, 2001).

For example, when a news article belongs to both financial and political clusters the system can suggest “more like this” news from either or both clusters to the user.

Hierarchical Clustering

A hierarchical clustering produces hierarchically structured clusters and in this structure relationships between the clusters are often visualised as a dendrogram.


Complete-linkage hierarchical agglomerateive clustering

Hierarchical Clustering

The algorithm can either merge individual observations into clusters from the bottom of the dendrogram (known as **agglomerative**) or decide where to split all the observations into smaller clusters (known as **divisive**).

Hierarchical: Agglomerative

An agglomerative clustering first treats each observation as an individual cluster and then in each iteration it merges the pair of clusters with the highest similarity into one cluster.

This iteration repeats until all observations are merges in a single cluster. An agglomerative clustering is also referred to as ***bottom-up*** approach.

Hierarchical: Divisive

A divisive clustering sees the entire collection of observations as a single large cluster and then iteratively splits this cluster into smaller clusters that maximise the distance (dissimilarity) between them.

This process stops when each cluster contains one observation. A divisive clustering is also referred to as ***top-down*** approach.

Agglomerative v.s. Divisive

Agglomerative approach is commonly used in hierarchical clustering because of its performance observed in the literature (Jain et al., 1999; Manning et al., 2008).

Agglomerative has quadratic time complexity makes it less efficient with large dataset.

Agglomerative v.s. Divisive

Some have shown that divisive approach can produce more accurate hierarchical clustering (Steinback et al., 2000) because clusters are formed based on global distribution of the observations – whereas an agglomerative approach forms clusters without a complete picture of the observations.

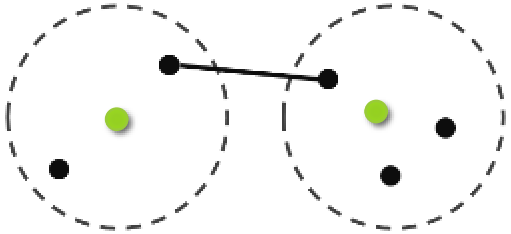
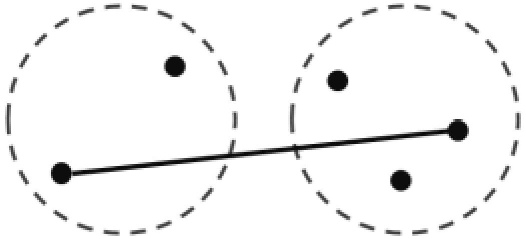
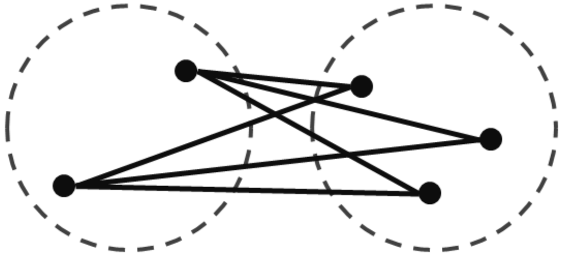
The time complexity of a divisive approach is linear to the size of the dataset and hence less computationally expensive than agglomerative ones.

Linkage

A linkage criterion is used in hierarchical agglomerative clustering (HAC) to determine the similarity between two clusters.

The two clusters with the shortest distance (or highest similarity) are then merged together.

Linkage

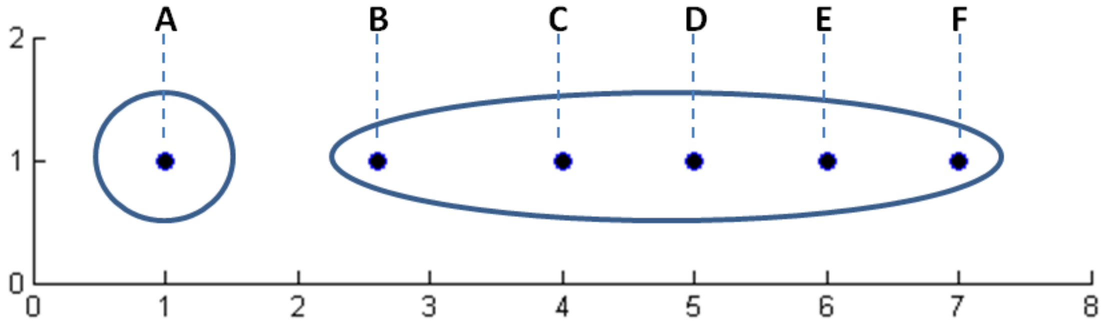
		
single-linkage	complete-linkage	average-linkage

Single-linkage

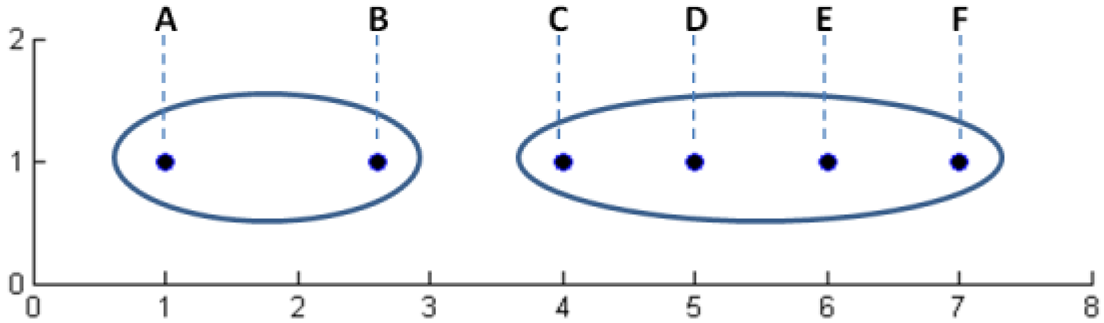
With single-linkage criterion, the similarity of two clusters is determined by the closest (most similar) members between the two clusters. Single-linkage clustering therefore disregards the possibility that other members of the two clusters may be very distant to each other.

This characteristic makes single-linkage clustering sensitive to noisy data and inherently prone to chaining effect (Guha et al., 1998).

Single v.s. Complete



Single-linkage: Chaining effect



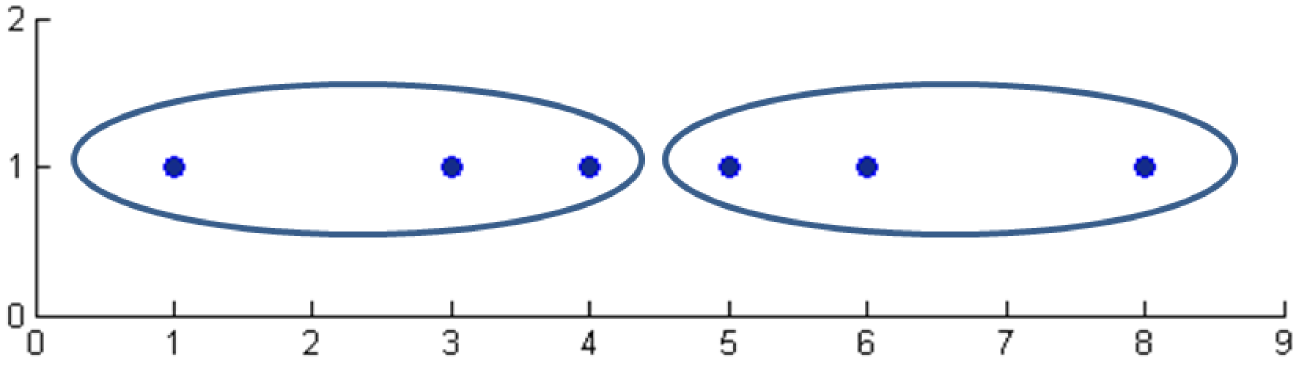
Complete-linkage

Complete-linkage

Complete-linkage criterion, on the other hand, determines the similarity of two clusters based on the two most distant (most dissimilar) members between the two clusters. This means that the two most distant members in the resulting cluster have the smallest pair-wise distance amongst all clusters.

Complete-linkage

However, a complete-linkage clustering also has its own problem. It focuses on the two most distant members from the pair of clusters, and these outliers sometimes can produce cluster boundary right in the middle of closely distant data.



Average-linkage

Average-linkage criterion tries to address the shortcomings of single- and complete-linkages by including all similarities between each pair of members from the two clusters.

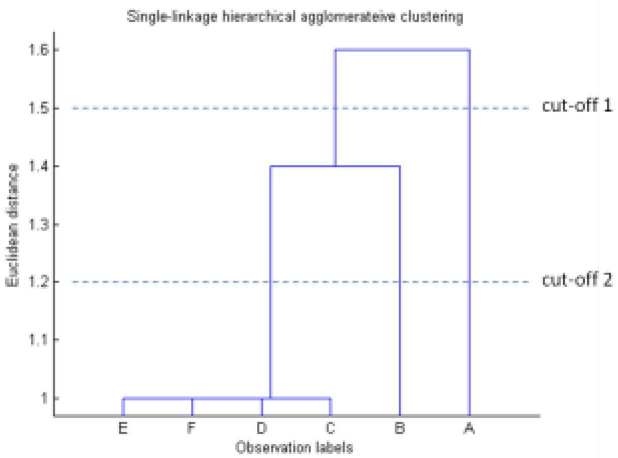
But it suffers worse time complexity.

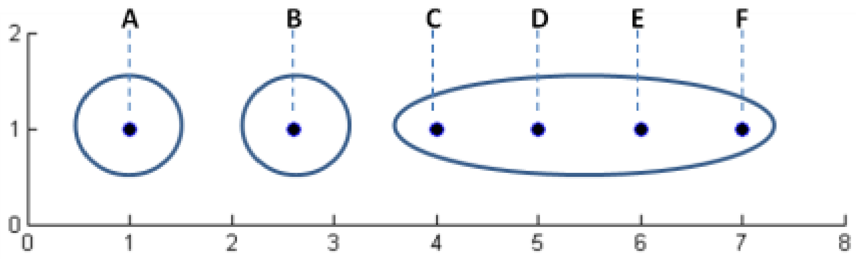
Determine cluster number: cutting the dendrogram

The crudest way is to specify the number *k* of clusters beforehand and then cut the dendrogram where it yields *k* clusters. One rule of thumb by Mardia et al. (1979) is to approximate *k* to , where *n* is the number of observation.

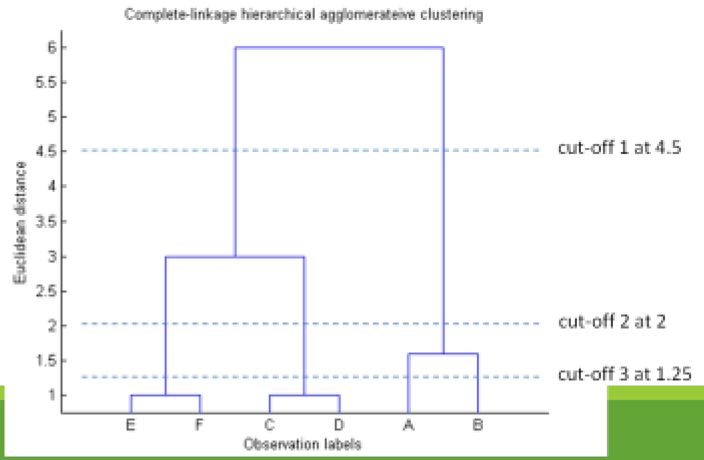
Determine cluster number: cutting the dendrogram

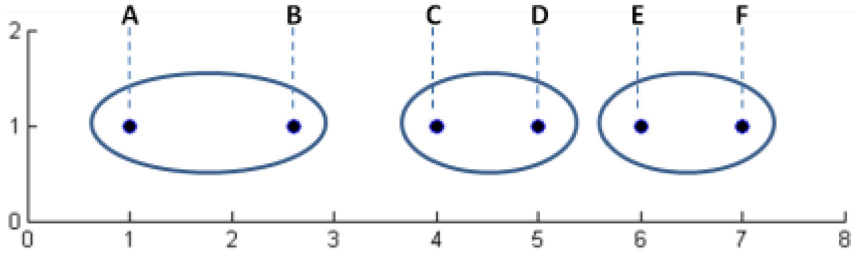
$$k = \sqrt{6/2} \approx 2$$





The clusters produced by the same dendrogram when cut-off 2 is applied, i.e. k = 3





Clustering results produced by the complete-linkage dendrogram when cut-off 2 (k =3) is applied

Partitioning Clustering

k-means is arguably the most used and important partitioning clustering method.

The steps of the *k*-means algorithm can be described as follows:

- Initial centroids are found for the whole document collection by randomly choosing *k* document vectors.
- Assign each document vectors to their closest centroid.
- Re-calculate the centroid vector for each cluster using current cluster memberships.
- Compute the quality function. If a better quality partition is produced, the process is repeated from step 2 until the optimal partitioning is reached.

k-Means

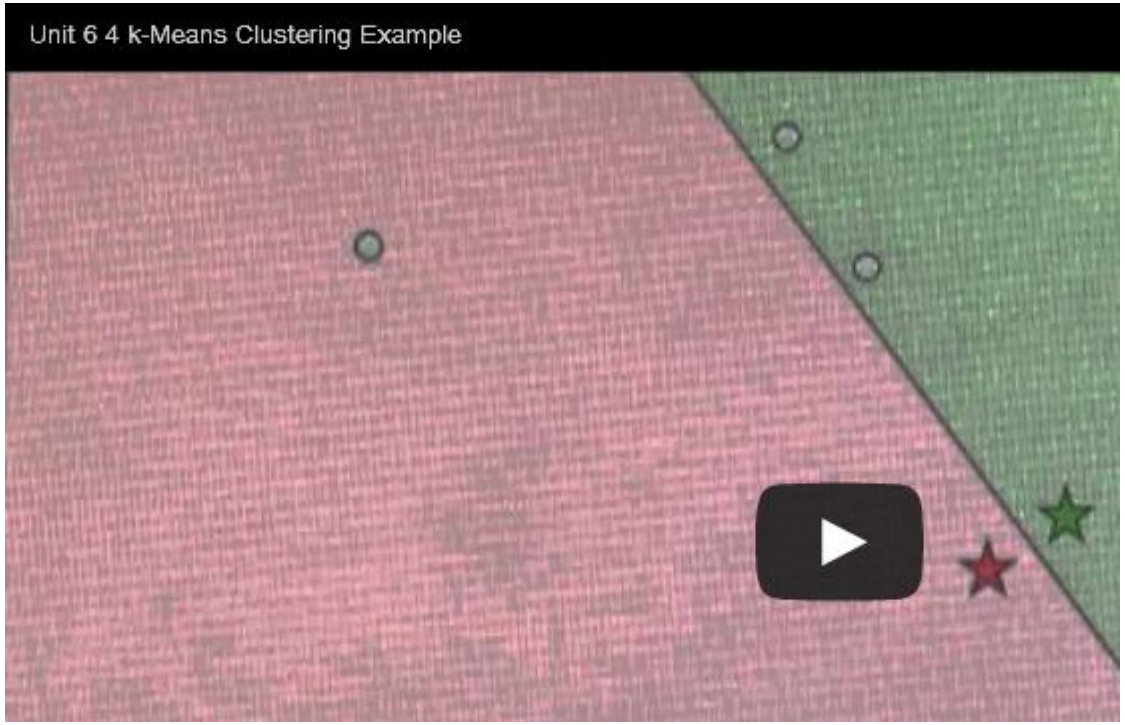
k-means aims to minimise the average squared Euclidean distance of the cluster members to the cluster centroid.

Centroid of cluster $\vec{\mu}$ is defined ω as:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{d} \in \omega} \vec{d}$$

where is a docum $\vec{d}$ ent vector in the cluster , and ω
| |is the size ω of the cluster .

k-Means



http://youtu.be/zaKjh2N8jN4

k-Means: drawbacks

The quality of clustering result depends on the randomly selected initial centroids.

Selecting the optimal *k* is not always an easy task, where different *k* values will have to be evaluated on the same data collection

Clustering on Translational Corpus

The corpus contains 4 Norwegian documents in 4 different topics, labelled A, J, O and T respectively.

Each Norwegian document has 18 corresponding translations in English, except for one that has only 17, making a total of 71 English documents.

Some of these corresponding translations in English are approved by a panel of examiners.

Clustering on Translational Corpus

Each English translation is coded, for example A03B-A:

- first letter is the topic label for the original Norwegian document
- the two digits here, 03, indicates this is the third English translation for document A
- the forth letter indicates whether this translation is in British English (B) or American English (A)
- the last letter indicates where this translation was accepted (A) or rejected (R) by the examiners

Pre-processing

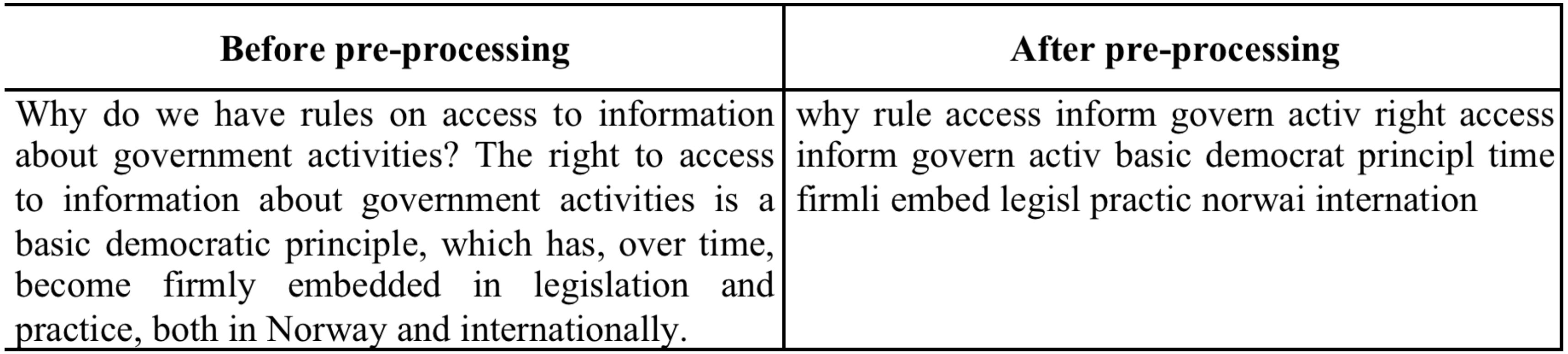
Standard pre-processing on English translations:

- Stopword removal
- Stemming

Bag-of-word representation

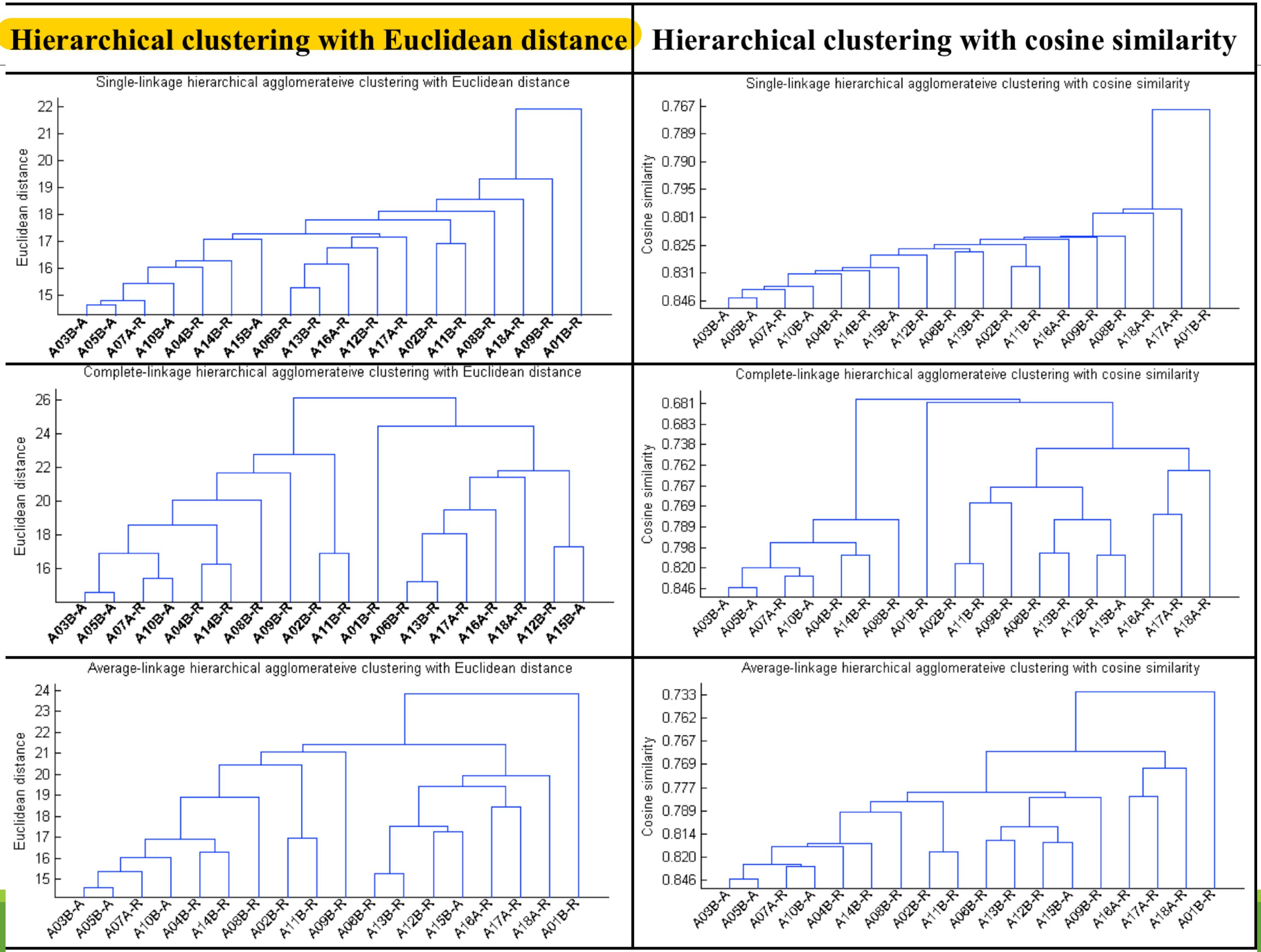
Term frequency for term weighting

Pre-processing



Before pre-processing	After pre-processing
Why do we have rules on access to information about government activities? The right to access to information about government activities is a basic democratic principle, which has, over time, become firmly embedded in legislation and practice, both in Norway and internationally.	why rule access inform govern activ right access inform govern activ basic democrat principl time firmli embed legisl practic norwai internation

Results: Hierarchical Clustering



Hierarchical clustering with Euclidean distance

Hierarchical clustering with cosine similarity

Single-linkage hierarchical agglomerateive clustering with Euclidean distance

Single-linkage hierarchical agglomerateive clustering with cosine similarity

Complete-linkage hierarchical agglomerateive clustering with Euclidean distance

Complete-linkage hierarchical agglomerateive clustering with cosine similarity

Average-linkage hierarchical agglomerateive clustering with Euclidean distance

Average-linkage hierarchical agglomerateive clustering with cosine similarity

Results: Hierarchical Clustering

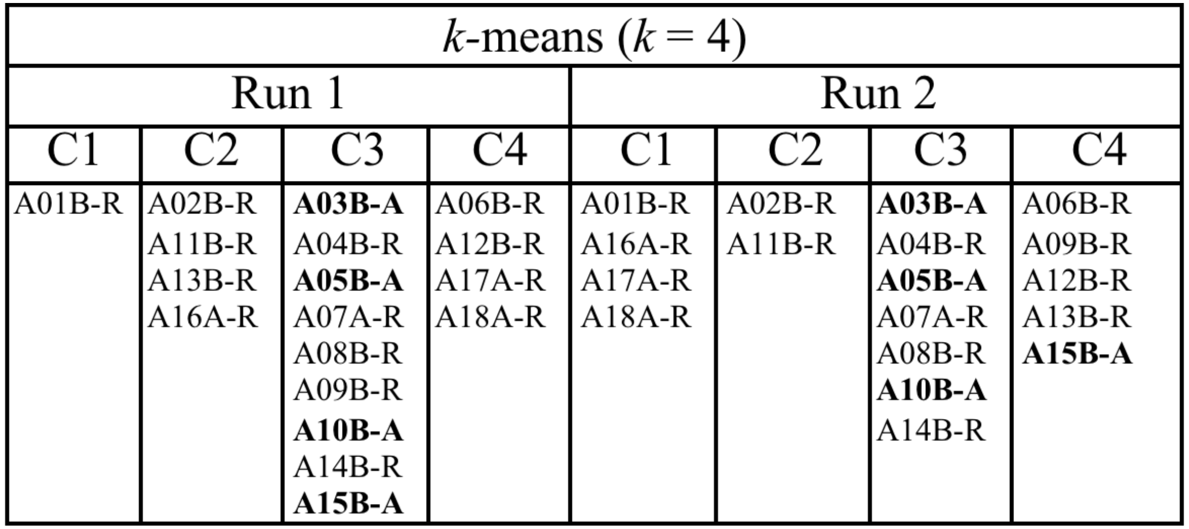
The clustering results differ more between linkages than between similarity metrics.

The chaining effect can be clearly observed in the single linkage clusterings.

An outlier problem in complete linkage was observed with document A01B-R.

With the single and average linkages, document A01B-R was in its own cluster if we cut the dendrogram at *k* = 2. However, with the complete linkages, document A01B-R essentially created cluster boundary right in the middle the data when *k* = 2, splitting up closely distant or similar documents.

Results: *k*-means



k-means (*k* = 4)							
Run 1				Run 2			
C1	C2	C3	C4	C1	C2	C3	C4
A01B-R	A02B-R	**A03B-A**	A06B-R	A01B-R	A02B-R	**A03B-A**	A06B-R
	A11B-R	A04B-R	A12B-R	A16A-R	A11B-R	A04B-R	A09B-R
	A13B-R	**A05B-A**	A17A-R	A17A-R		**A05B-A**	A12B-R
	A16A-R	A07A-R	A18A-R	A18A-R		A07A-R	A13B-R
		A08B-R				A08B-R	**A15B-A**
		A09B-R				**A10B-A**	
		A10B-A				A14B-R	
		A14B-R					
		A15B-A					

Results: *k*-means

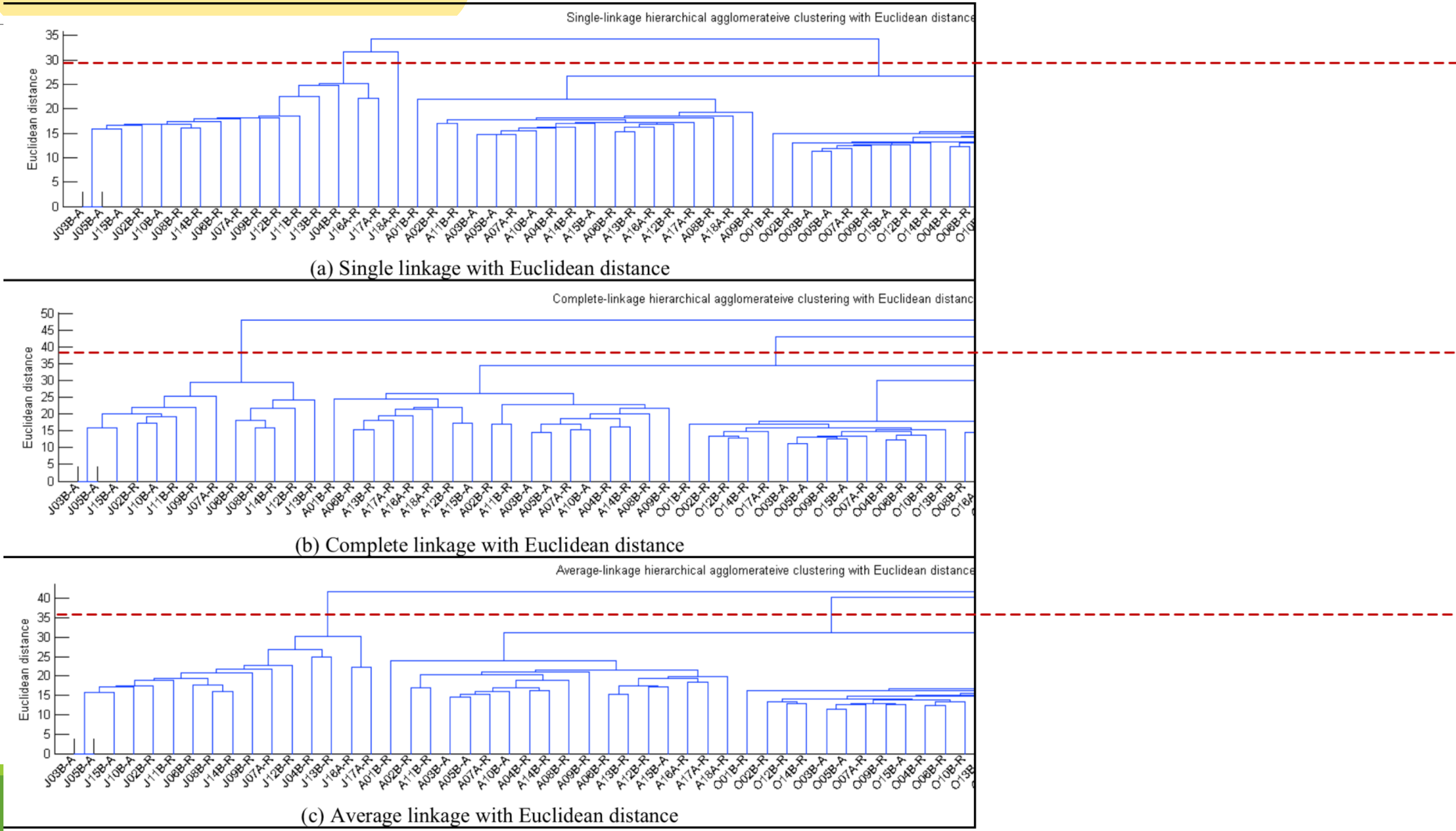
The discrepancy in the clustering results from the two runs confirms that both k-means is sensitive to the initial selection of centroids for k-means.

While k-means is successful in producing the clusters, it is less informative than the hierarchical methods above in telling the relationships between the clusters.

Clustering by Topics

We are going to apply the same methods again but to all of 71 English translations across four topics.

Results: Hierarchical



(a) Single linkage with Euclidean distance

(b) Complete linkage with Euclidean distance

(c) Average linkage with Euclidean distance

Clustering Un-preprocessed Docs

Stopwords in the English translation are not removed and the texts are not stemmed.

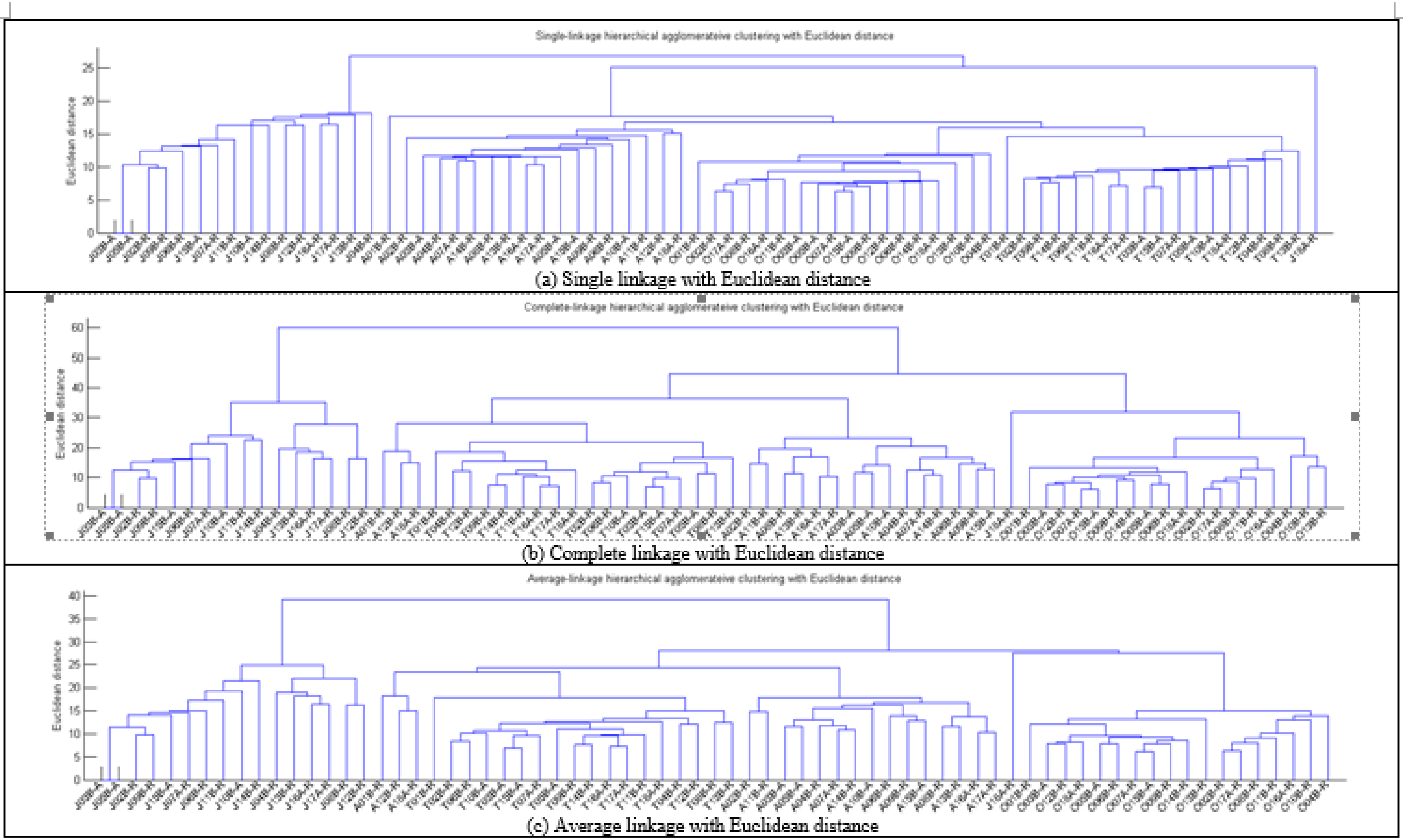
Only top 50 frequently occurring words are selected to make the document vectors.

Using unprocessed text and high frequency words is that it could reflect differences in writing style.

We would be able to see whether we could produce good clusters with far less information and whether we could detect any difference in writing style in our dataset.

Burrows, J., 1987. Word-patterns and story-shapes: the statistical analysis of narrative style. *Library and Linguistic Computing*, vol. 2, pp. 61-70
Burrows, J. 2002. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Library and Linguistic Computing*, vol. 17(3), pp. 267-287

Results: Hierarchical (1)



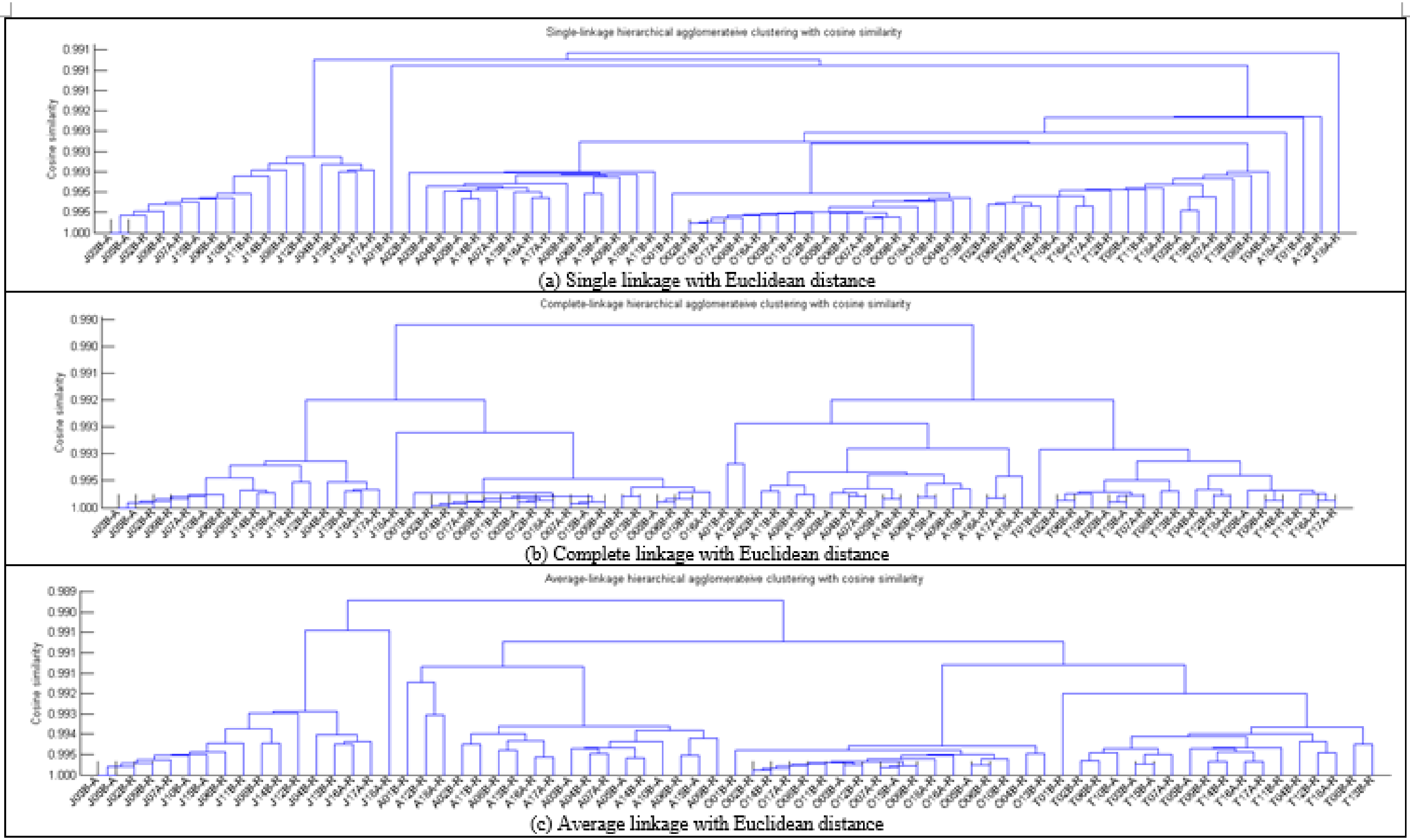
(a) Single linkage with Euclidean distance

(b) Complete linkage with Euclidean distance

(c) Average linkage with Euclidean distance

Figure 6.3 HAC with Euclidean distance applied to all translations across four topics, text un-processed

Results: Hierarchical (2)



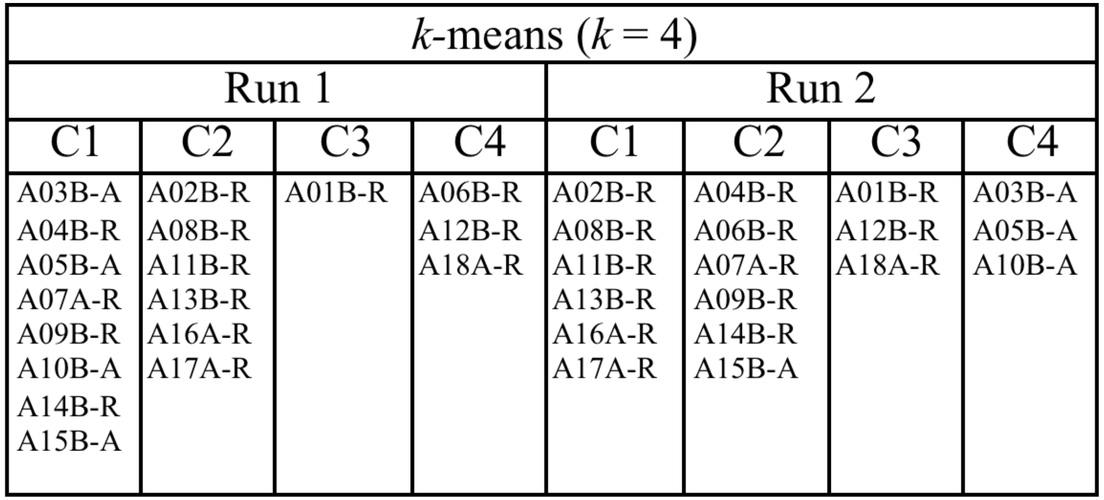
(a) Single linkage with Euclidean distance

(b) Complete linkage with Euclidean distance

(c) Average linkage with Euclidean distance

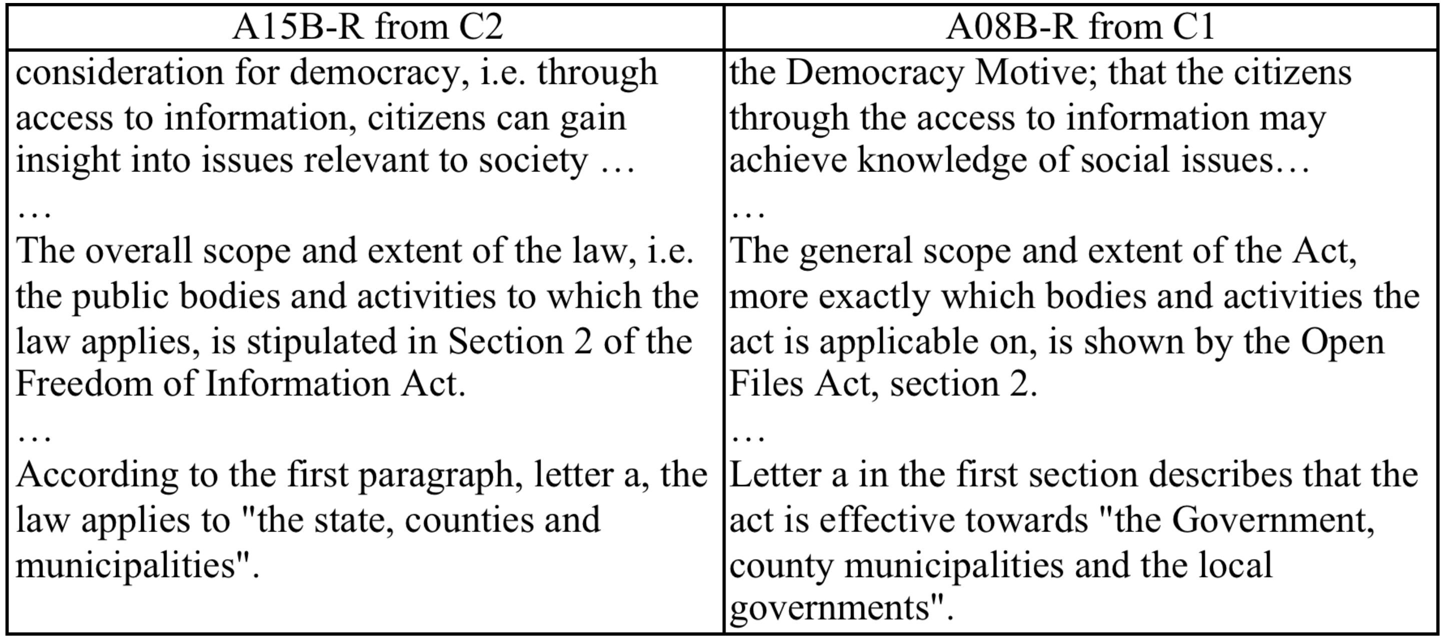
Figure 6.4 HAC with cosine similarity applied to all translations across four topics, text un-processed

Results: k-means



k-means (*k* = 4)							
Run 1				Run 2			
C1	C2	C3	C4	C1	C2	C3	C4
A03B-A	A02B-R	A01B-R	A06B-R	A02B-R	A04B-R	A01B-R	A03B-A
A04B-R	A08B-R		A12B-R	A08B-R	A06B-R	A12B-R	A05B-A
A05B-A	A11B-R		A18A-R	A11B-R	A07A-R	A18A-R	A10B-A
A07A-R	A13B-R			A13B-R	A09B-R		
A09B-R	A16A-R			A16A-R	A14B-R		
A10B-A	A17A-R			A17A-R	A15B-A		
A14B-R							
A15B-A							

Results: k-Means



A15B-R from C2	A08B-R from C1
consideration for democracy, i.e. through access to information, citizens can gain insight into issues relevant to society … … The overall scope and extent of the law, i.e. the public bodies and activities to which the law applies, is stipulated in Section 2 of the Freedom of Information Act. … According to the first paragraph, letter a, the law applies to "the state, counties and municipalities".	the Democracy Motive; that the citizens through the access to information may achieve knowledge of social issues… … The general scope and extent of the Act, more exactly which bodies and activities the act is applicable on, is shown by the Open Files Act, section 2. … Letter a in the first section describes that the act is effective towards "the Government, county municipalities and the local governments".

Discussions

One must remember that clustering is an exploratory technique, which does not necessarily provide an absolute analysis on the given data.

So it usually takes numerous empirical observations on the clustering results before one can confidently recommend a particular clustering technique for any given task.

The clustering results can be interpreted differently depending on what is sought after in the current analysis.

References & Reading List

- https://www.kaggle.com/learn/feature-engineering
- https://towardsdatascience.com/clustering-for-data-nerds-ebbfb7ed4090
- https://towardsdatascience.com/why-data-is-represented-as-a-vector-in-data-science-problems-a195e0b17e99
- https://www.kaggle.com/arthurtok/interactive-intro-to-dimensionality-reduction