MIS772
Predictive Analytics

Clustering

Refer to your textbook by Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed, Elsevier, 2018.

Data Clustering

- **k-Means clustering**
  Centroids
  k-Means algorithm
- **Other types of clustering**
  k-Medoids
  Density Based (DBSCAN)
  Agglomerative
- **Cluster visualisation**
  with help from Principal Component Analysis (PCA)
- **Optimisation of clustering**

DEAKIN BUSINESS SCHOOL

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

Data Clustering

- The task of grouping data so that data points in the same group (called cluster) are more similar to each other than to those in other groups (also clusters)
- Data clustering can be viewed as multi-objective optimisation
- Cluster analysis is an iterative *knowledge discovery* process by trial and error until results achieve some desired properties

Purposes

- Understanding of data
- Dimension reduction

Applications

- Business, e.g. marketing
- Science/Eng., e.g. materials
- Medicine, e.g. diagnostics
- IT, e.g. pattern recognition

Sample Clustering Methods

- *k-Means/k-Medoids*: searches for centers of groups
- *DBSCAN*: searches for high density data groups
- *Agglomerative*: searches for group hierarchies (trees)

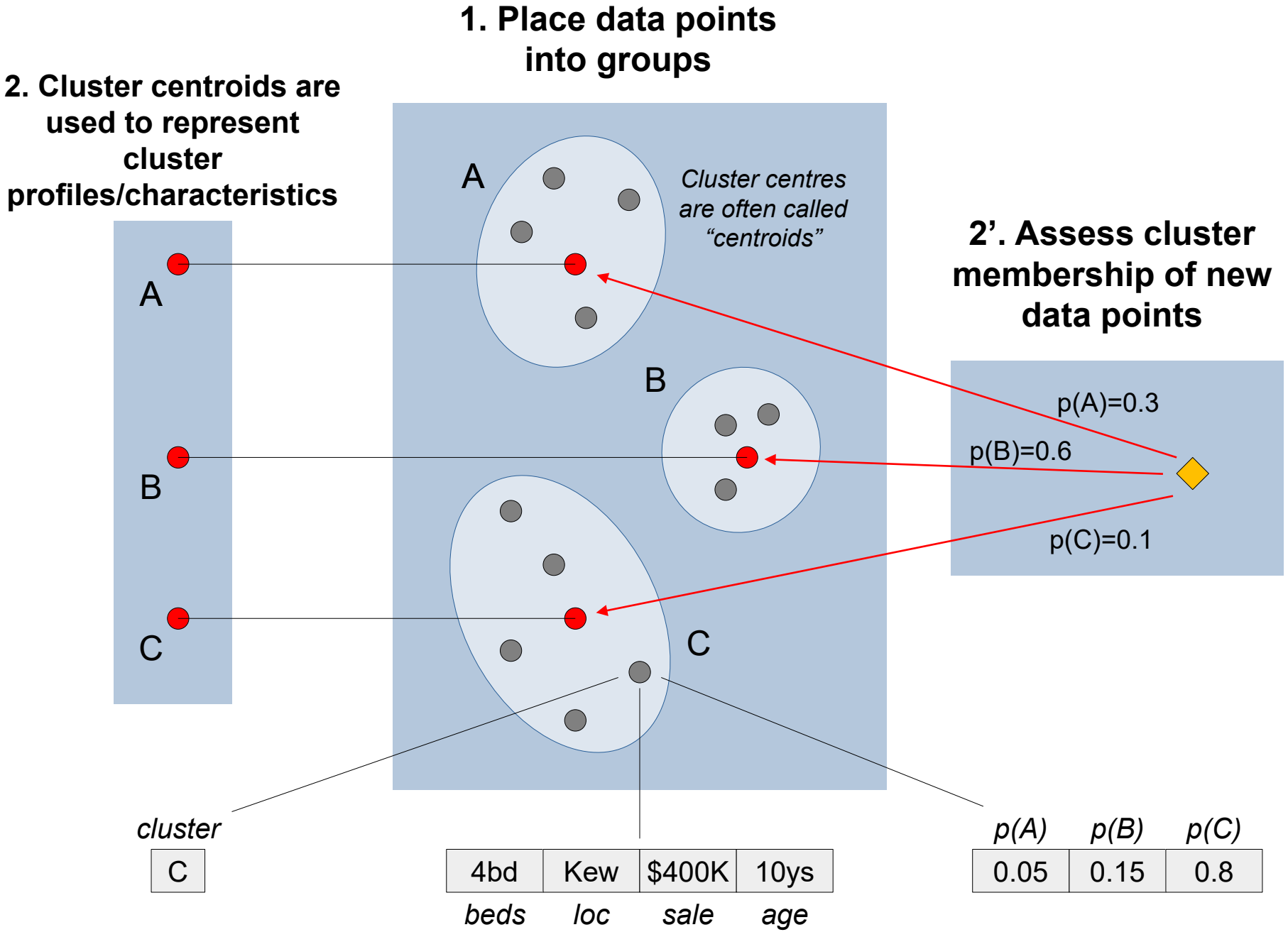
Also see KD 7 on Data clustering

Some definitions

- ***Data clustering*** **(in some disciplines *data segmentation*) is the unsupervised method of grouping data so that data points in the same group (called cluster) are more similar to each other than to those in other groups.**

- ***Cluster diagnostics*** **is the process of assessing the quality of the clustering system. It usually focuses on ensuring *cluster cohesion* (i.e. similar data points belong to the same cluster), *cluster separation* (i.e. different data points belong to different clusters and cluster centres are far apart), *minimisation of fragmentation* (i.e. few small clusters).**

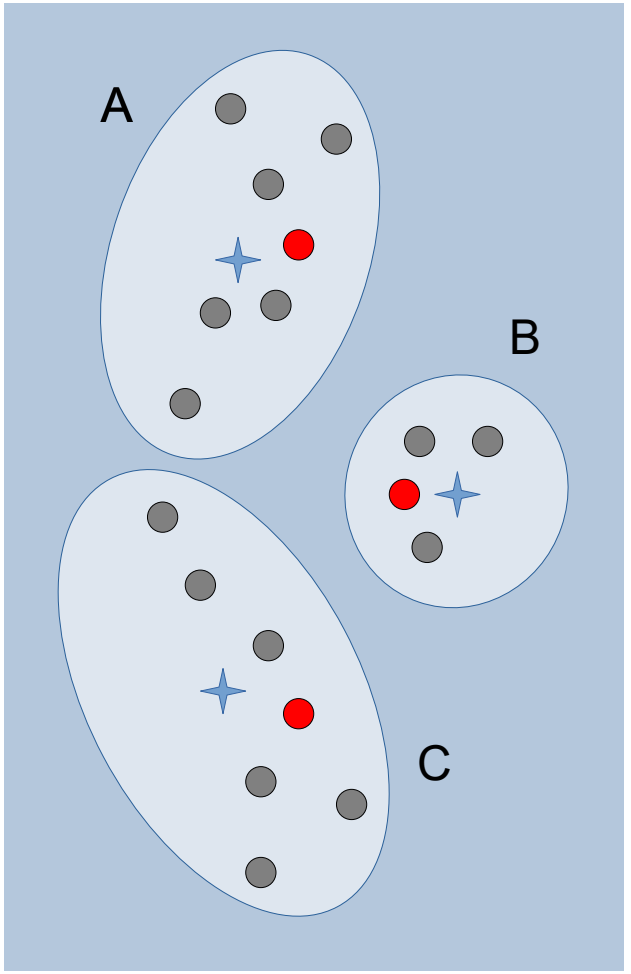
Aims of Data Clustering



DEAKIN BUSINESS SCHOOL

Centroid-based clustering:

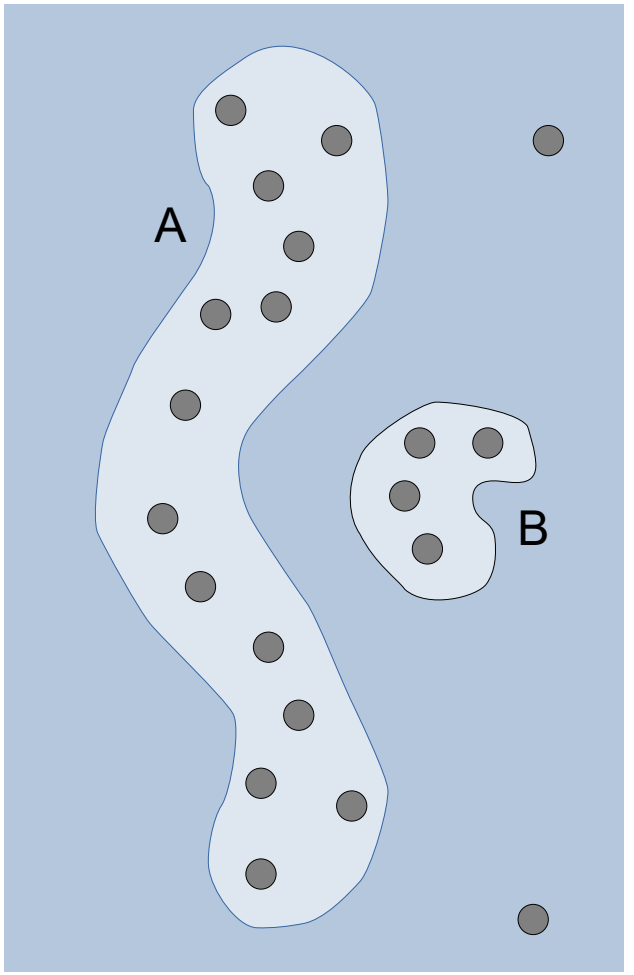
- Focuses on identifying the centers of clusters and allocating cluster members to optimise some quality measure.
- Usually require users to specify the number of clusters (*k*) apriori.
- Cluster centroids are defined differently in different clustering techniques:
  - K-Means: cluster centroids are defined by new data points, whose attributes are the means/averages of the attributes belonging to data points assigned to each cluster. (most popular and quite fast)
  - K-Medoids: similar to k-means, but the cluster centroids are defined by an actual data points, which minimises cluster dissimilarity measures
  - K-medians, and other kernel-based varieties



k-means (blue),
k-medoids (red)

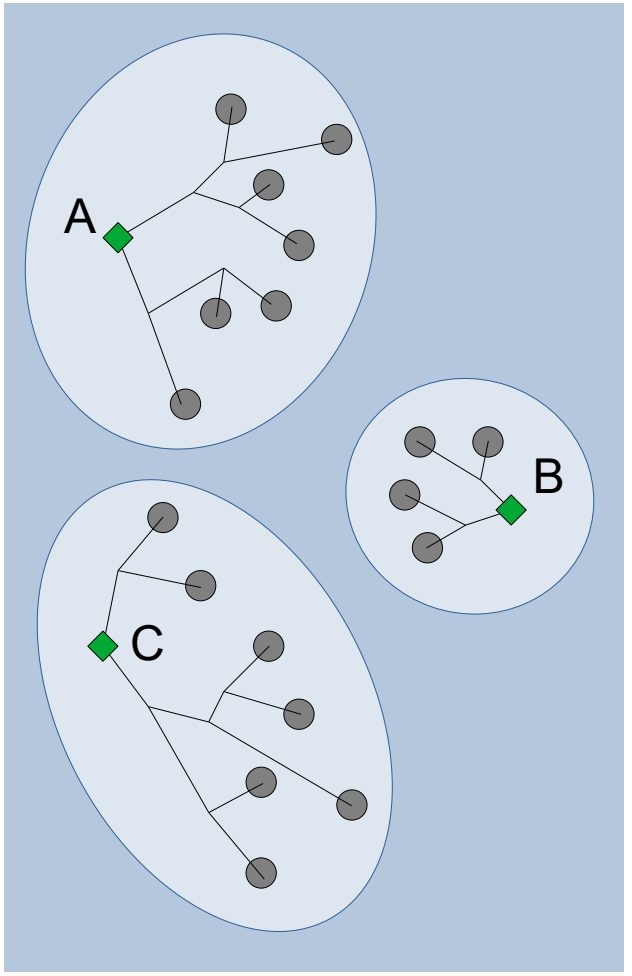
Density-based clustering

- It aims to separate data points into high-density and low-density groups (considered as noise).
- A cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. The groups are formed by linking data within their neighbourhood.
- The method is resistant to outliers and has irregularly shaped clusters.
- It struggles with high-dimensional data and data of varying densities
- It requires numerical and scaled distance measures.
- Usually require a density threshold specified by user.
- Examples include: DBSCAN (popular) and OPTICS



Hierarchical clustering

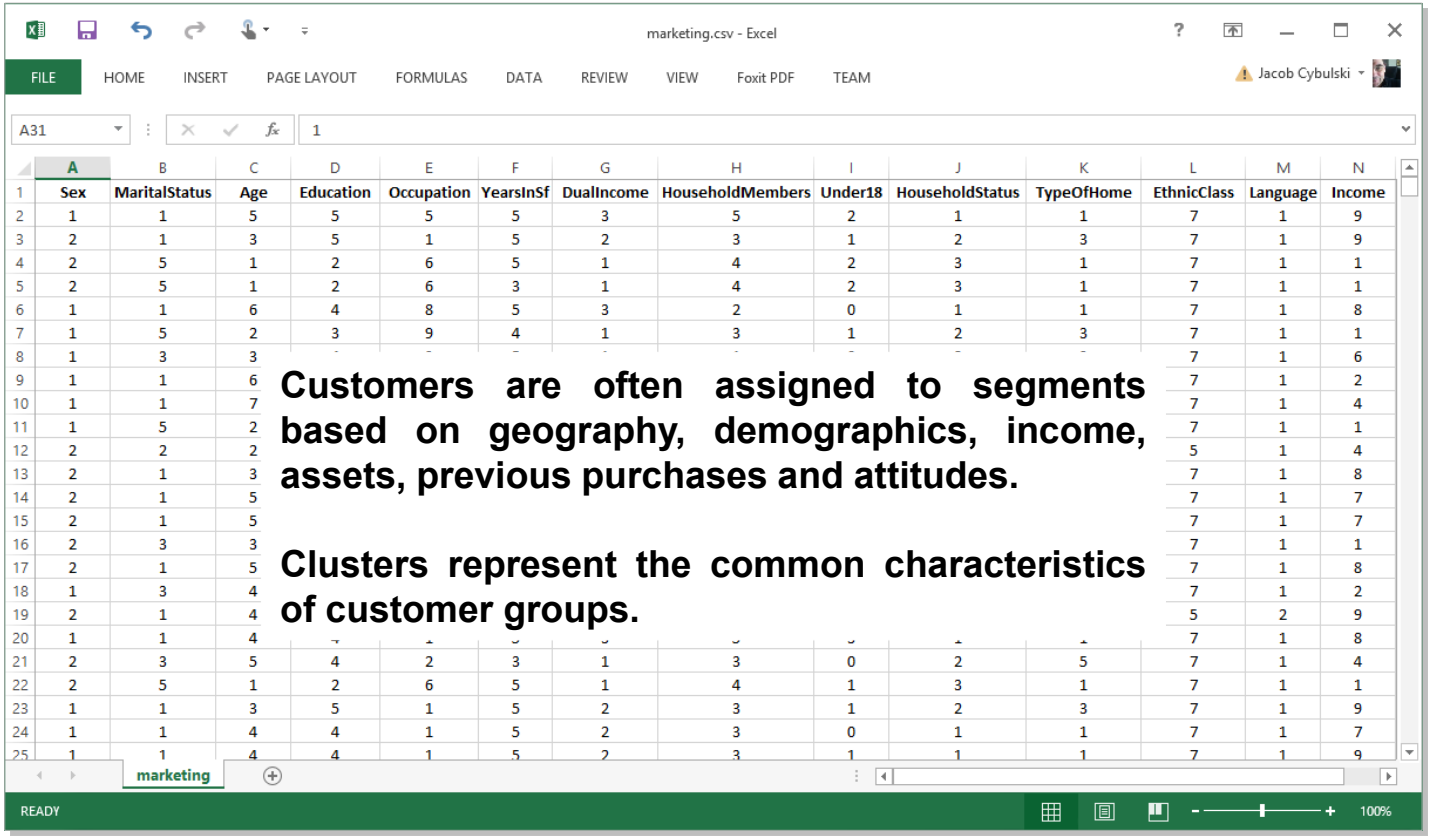
- It derives clusters from hierarchical relationships in data
- Types of cluster formation include:
  - *agglomerative* (bottom-up): each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
  - *divisive* (top-down): all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
- Both suffer from the presence of outliers and divisive is slow.



Application and Example

- Segmenting a customer base into similar groups is important for marketing purposes.
- The segments can be used to attract new customers or determine their needs to be matched with existing products and services.
- This is done by performing clustering of enterprise data, which is sometimes combined with other sources of open data.

- This following dataset contains data from a survey of customers in a shopping mall in the San Francisco Bay area.
- The goal of the clustering was to identify segments of customers which could be estimated by 13 demographic attributes.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Sex	MaritalStatus	Age	Education	Occupation	YearsInSf	DualIncome	HouseholdMembers	Under18	HouseholdStatus	TypeOfHome	EthnicClass	Language	Income
2	1	1	5	5	5	5	3	5	2	1	1	7	1	9
3	2	1	3	5	1	5	2	3	1	2	3	7	1	9
4	2	5	1	2	6	5	1	4	2	3	1	7	1	1
5	2	5	1	2	6	3	1	4	2	3	1	7	1	1
6	1	1	6	4	8	5	3	2	0	1	1	7	1	8
7	1	5	2	3	9	4	1	3	1	2	3	7	1	1
8	1	3	3									7	1	6
9	1	1	6									7	1	2
10	1	1	7									7	1	4
11	1	5	2									7	1	1
12	2	2	2									5	1	4
13	2	1	3									7	1	8
14	2	1	5									7	1	7
15	2	1	5									7	1	7
16	2	3	3									7	1	1
17	2	1	5									7	1	8
18	1	3	4									7	1	2
19	2	1	4									5	2	9
20	1	1	4									7	1	8
21	2	3	5	4	2	3	1	3	0	2	5	7	1	4
22	2	5	1	2	6	5	1	4	1	3	1	7	1	1
23	1	1	3	5	1	5	2	3	1	2	3	7	1	9
24	1	1	4	4	1	5	2	3	0	1	1	7	1	7
25	1	1	4	4	1	5	2	3	1	1	1	7	1	9

Customers are often assigned to segments based on geography, demographics, income, assets, previous purchases and attitudes.

Clusters represent the common characteristics of customer groups.

DEAKIN BUSINESS SCHOOL

Example: Attributes And Their Coding

1. HOUSEHOLD INCOME PA

1. Less than $10,000
2. $10,000 to $14,999
3. $15,000 to $19,999
4. $20,000 to $24,999
5. $25,000 to $29,999
6. $30,000 to $39,999
7. $40,000 to $49,999
8. $50,000 to $74,999
9. $75,000 or more

2. SEX

1. Male
2. Female

3. MARITAL STATUS

1. Married
2. Living together, not married
3. Divorced or separated
4. Widowed
5. Single, never married

4. AGE

1. 14 thru 17
2. 18 thru 24
3. 25 thru 34
4. 35 thru 44
5. 45 thru 54
6. 55 thru 64
7. 65 and Over

5. EDUCATION

1. Grade 8 or less
2. Grades 9 to 11
3. Graduated high school
4. 1 to 3 years of college
5. College graduate
6. Grad Study

6. OCCUPATION

1. Professional/Managerial
2. Sales Worker
3. Laborer/Driver
4. Clerical/Service Worker
5. Homemaker
6. Student, HS or College
7. Military
8. Retired
9. Unemployed

7. HOW LONG LIVED IN SF AREA?

1. Less than one year
2. One to three years
3. Four to six years
4. Seven to ten years
5. More than ten years

8. DUAL INCOMES (IF MARRIED)

1. Not Married
2. Yes
3. No

9. PERSONS IN YOUR HOUSEHOLD

1. One… 9. Nine or more

10. PERSONS IN HOUSEHOLD UNDER 18

0. None… 9. Nine or more

11. HOUSEHOLDER STATUS

1. Own
2. Rent
3. Live with Parents/Family

12. TYPE OF HOME

1. House
2. Condominium
3. Apartment
4. Mobile Home
5. Other

13. ETHNIC CLASSIFICATION

1. American Indian
2. Asian
3. Black
4. East Indian
5. Hispanic
6. Pacific Islander
7. White
8. Other

14. LANGUAGE SPOKEN AT HOME?

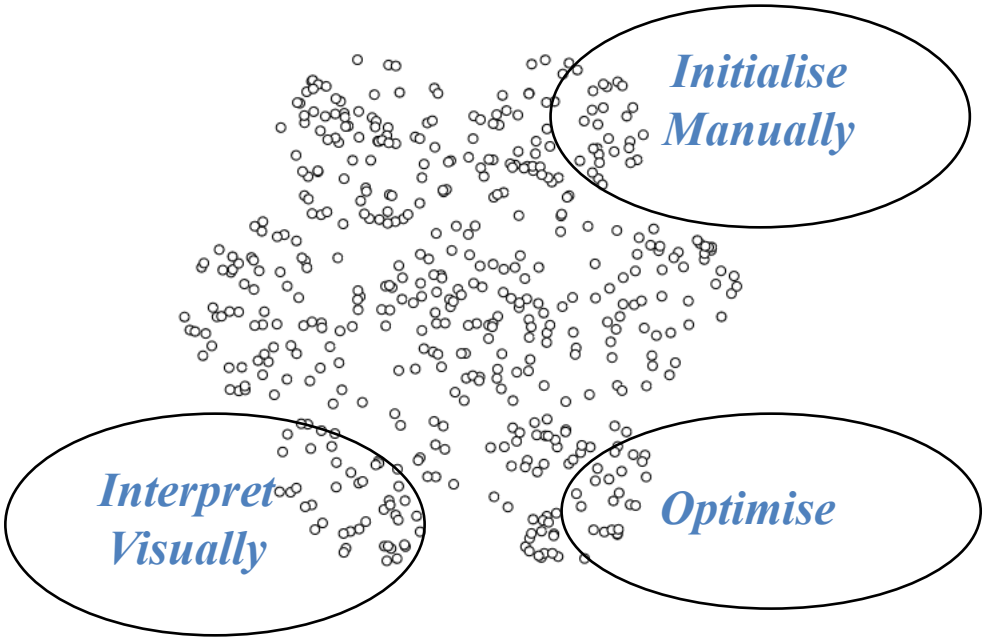
1. English
2. Spanish
3. Other

DEAKIN BUSINESS SCHOOL

K-Means Clustering

- ***k-Means clustering*** **is the most commonly used technique for cluster detection in data.**
- **It relies on geometric interpretation of data as points in space separated by some distance, e.g. measured as *Euclidean distance*.**
- ***k*** **stands for the number of clusters to be detected and is selected by the user.**
- **The objective is to find k points that could be used as cluster centres – *centroids*.**
- **The best centres minimise *mean square distances* of every data point to its nearest centroid.**

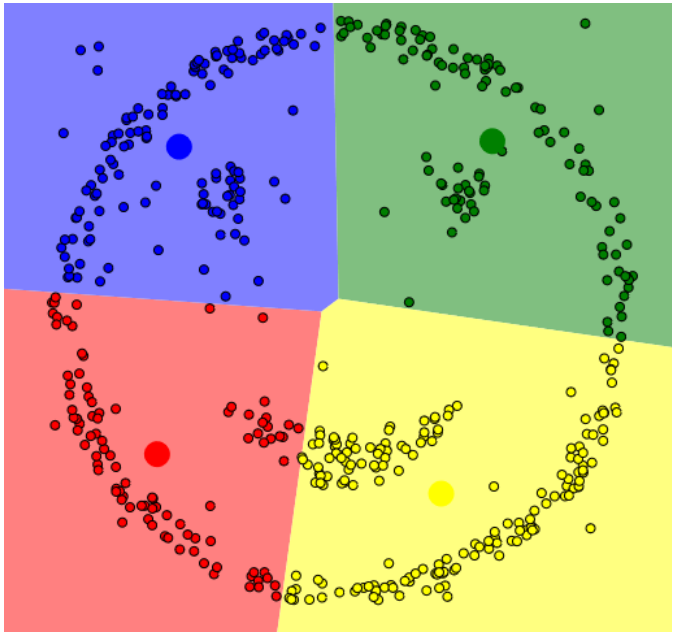
- **k-Means algorithm seeks optimal solution by *iterative improvement* of its initial "guess" of cluster centroids.**
- **Typical problems:**
  - **How many clusters?**
  - **How to select the initial centres?**
  - **How to interpret the result?**



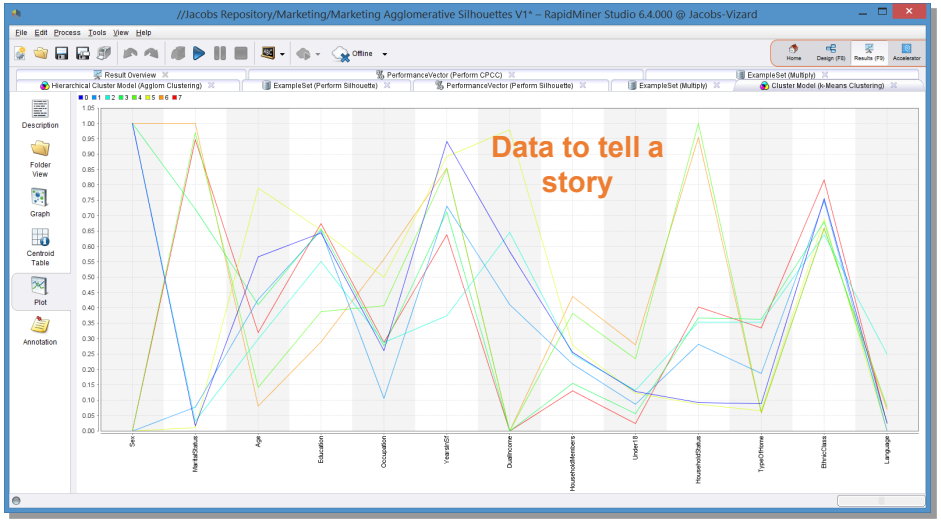
DEAKIN BUSINESS SCHOOL

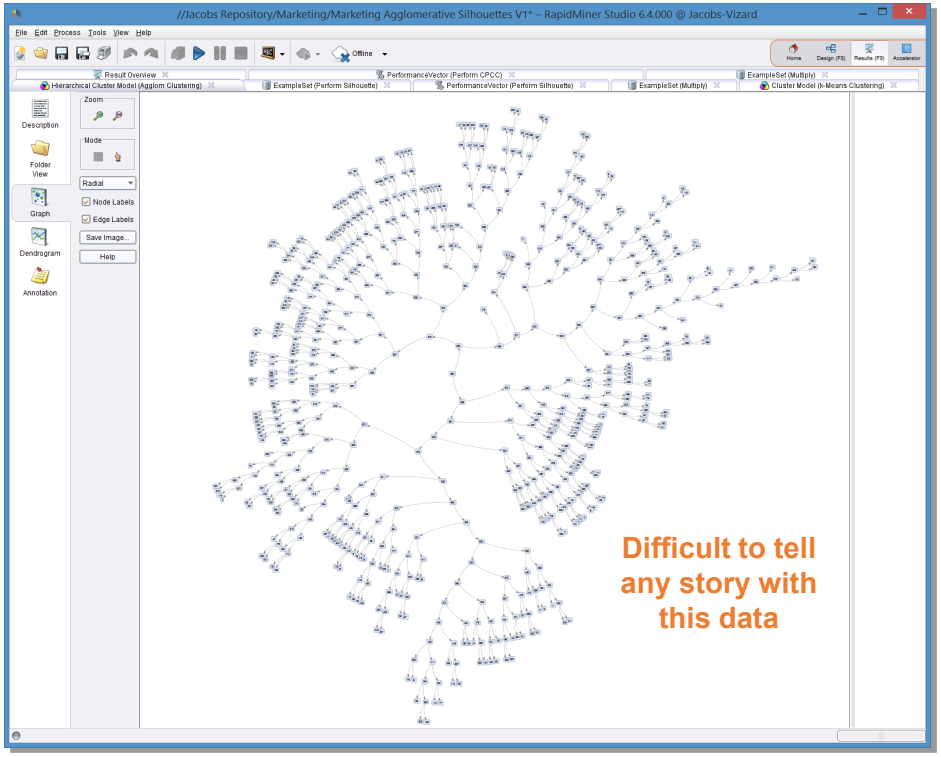
- In *k-means clustering* each cluster point should be closest to the cluster centre – the *centroid*.
- First we select *k* - the number of clusters we want to find.
- The centres of those k clusters are initialised, e.g. using:
  - Random allocation
  - Farthest points
  - Probabilistic measures
- Then iterating these steps:
  - *Reassign Points*: we assign every data point to the cluster with the nearest centroid.
  - *Update Centroids*: we recalculate each centroid's location as the mean (centre) of all cluster points.
- This is done until centroids and the points stop moving.

- In *k-medoid clustering*, cluster centres are the actual data points, which are iteratively reassigned to other data points by minimising the sum of pairwise dissimilarities between points.



How to analyse clusters





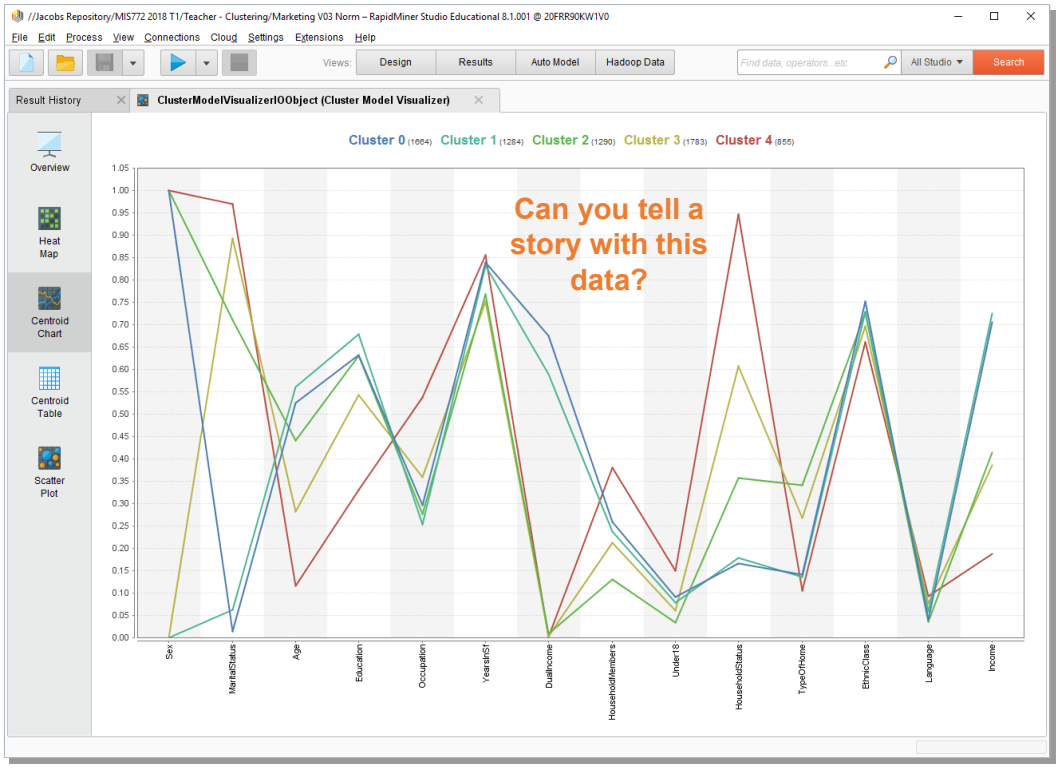
How to analyse clusters?

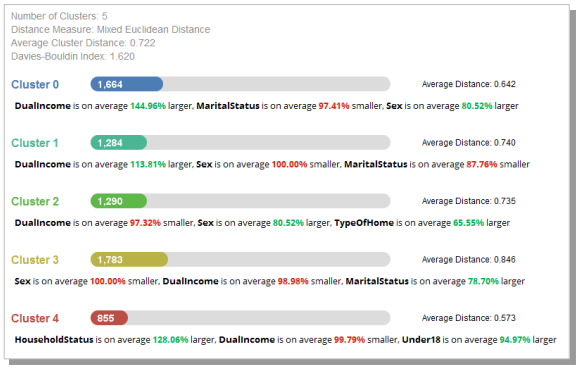
- Analyse clusters by features and relationships between their centroids
- For *flat clusters*, use web or bar charts to identify which attributes differentiate what clusters and by how much
- For *hierarchical clusters* investigate the tree structure, also decision trees are a good start for this analysis as well
- Name or describe your clusters – if you are unable to do this then perhaps they make no sense
- Beware of numeric vs. nominal data
- Clustering using inter-related attributes may lead to bias
- Treat all your attributes equally by normalizing them

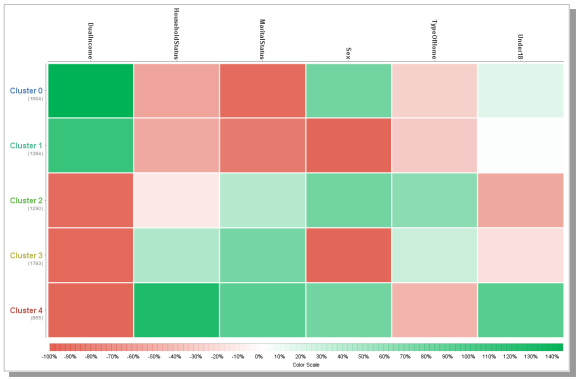
DEAKIN BUSINESS SCHOOL

Visualising Clusters for Understanding

Further analysis is undertaken to understand cluster properties. Many visualisation techniques can assist this process, e.g. in RapidMiner *"Cluster Model Visualizer" operator* helps analysing centroids generated by k-means clusters.







Sample analysis of the cluster plot (above)

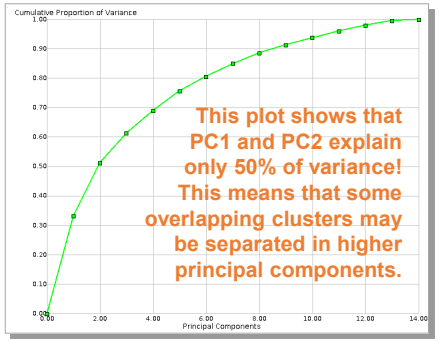
- **Cluster 4 (red line): young single men, uneducated and unemployed, living with parents and many siblings in a house, mainly students, no income.**
- **Cluster 1 (turquoise): older, well-educated, professional women, living together but not married, on very good dual-income, mostly owning their own house.**

DEAKIN BUSINESS SCHOOL

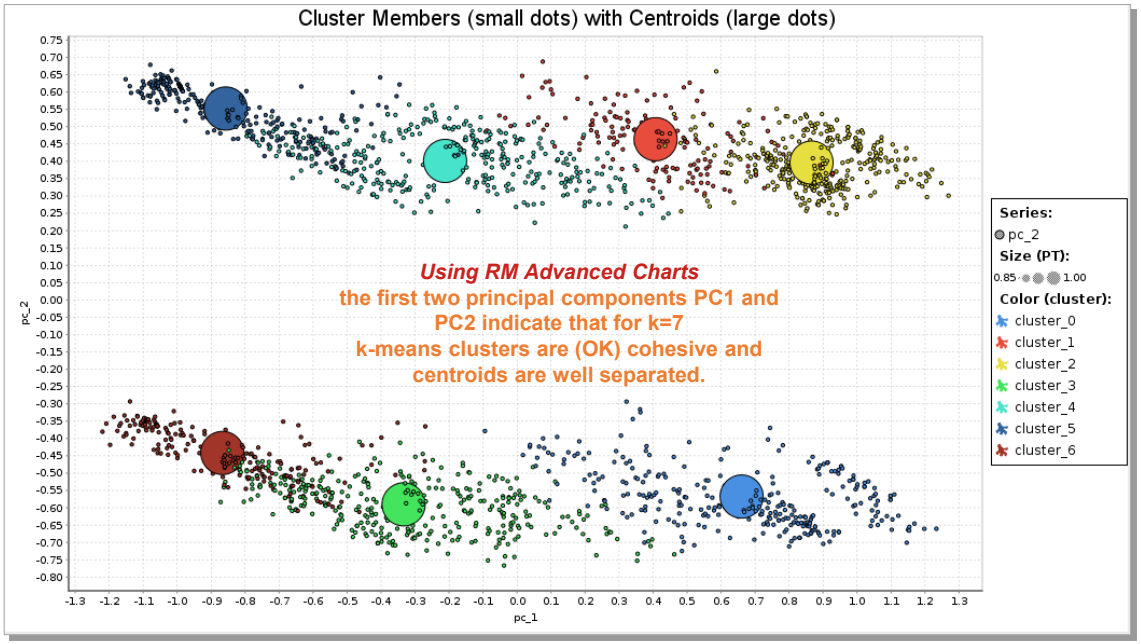
Cluster Evaluation by Visualisation

- Cluster visualisation can determine *cluster cohesion* or *overlap* and *centroids separation*.
- Unfortunately this is not easy in multi-dimensional data.
- To overcome this problem, various methods are used to "cast" the data into 2D.
- The commonly used techniques is PCA, which transforms all data into new axes which can be plotted in 2D.
- The PCA transformation, which was used in plotting all clustered data, can also be applied to cluster centroids, so that their position could then be determined against each other and in relation to their cluster members.

- *PCA (Principal Components Analysis)* is commonly used in plotting clusters.
- PCA transforms (rotates) existing attributes into new attributes called *principal components*, while preserving their geometry, i.e. relative distances between all data points stay the same.
- PCA also ensures that principal components are independent and sequenced in the order of variance, i.e. PC1 explains the most variance in data, then PC2, PC3, etc.
- We can plot data in PC1 and PC2, thus depicting the majority (if we are lucky) of variance in data.
- Cumulative variance plot needs to be consulted to determine how much variance is explained by PC1 and PC2.



Size denotes cluster members (small) and centroids (large)
Colour denotes clusters (sample reduced to 2000)



DEAKIN BUSINESS SCHOOL

Principal Components Analysis (PCA)

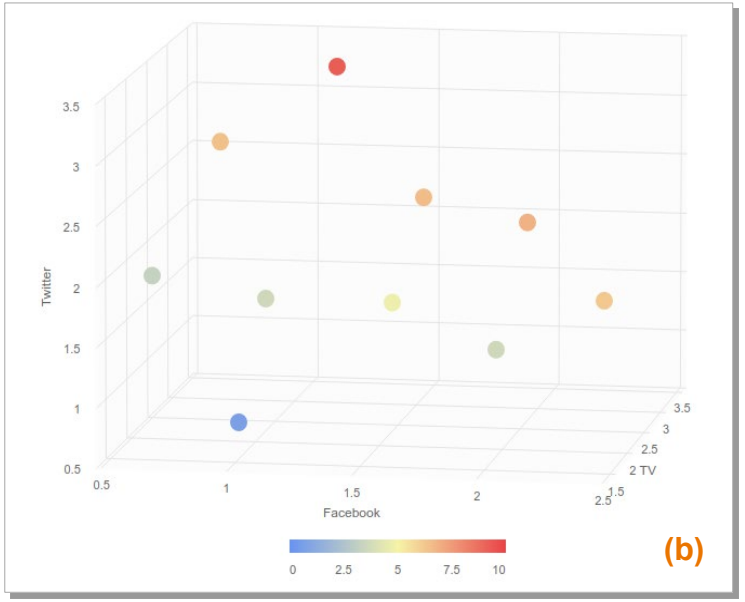
- PCA is a mathematical transformation of data which aims to reduce its dimensionality while preserving geometric properties of data points, such as distance between them.
- PCA achieves this by "rotating" existing the axes (attributes which may be dependent) of the multi-dimensional data into new axes (which are independent), which also capture data variance from the highest (PC1 and PC2) to lowest (PC n).
- PCA creates a matrix of *loadings* which represent linear combinations of attributes in the new axes, and defining a "rotation" matrix of *eigenvectors* (defining directions of new axes) and a vector of *eigenvalues* (defining variance of each component, and used for data scaling in each direction).
- This is useful in *visualising complex multi-dimensional data* on a 2D screen, e.g. plotting data clusters.
- It is also useful to eliminate multi-colinearities in data.

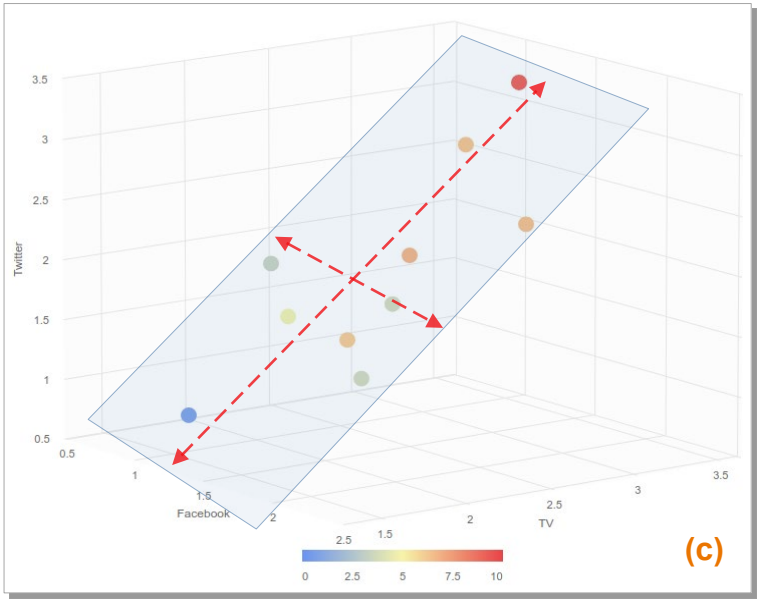
PCA – The works, Part 1 cont'd

Company	Facebook	TV	Twitter	Revenue
No Frills Leader	2	1.800	1.400	3.600
Abracadabra	2.500	1.300	2	6.200
Go With Flow	1.600	1.700	1.800	4.700
Solid Bucks	1	1.600	0.800	0.700
Shy Winner	0.500	2.500	1.700	3.300
Cowboy Patch	0.900	2.900	1.400	3.600
Down Under Star	1.500	3.200	2.200	6.500
Pluck Your Luck	2.100	2	2.400	6.800
Golden Racer	0.600	3.600	2.500	6.400 (a)
Timely Reminder	1.100	3.500	3.200	9.200

We cannot see any relations in **the tabular format (a)**.

When plotting data in **3D**, it all **seems random (b)**.





However, it is possible to rotate the 3D chart in such a way that all companies seem to line up and it seem that the majority of variance occurs in a **2D plane (c)**.

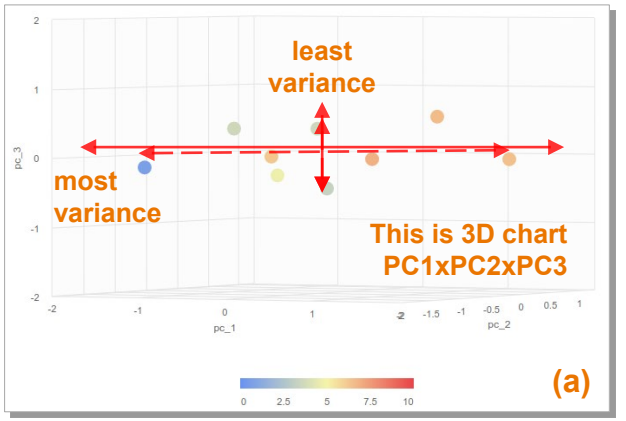
This opens the possibility of dealing with data complexity by reducing its dimension from 3D to 2D.

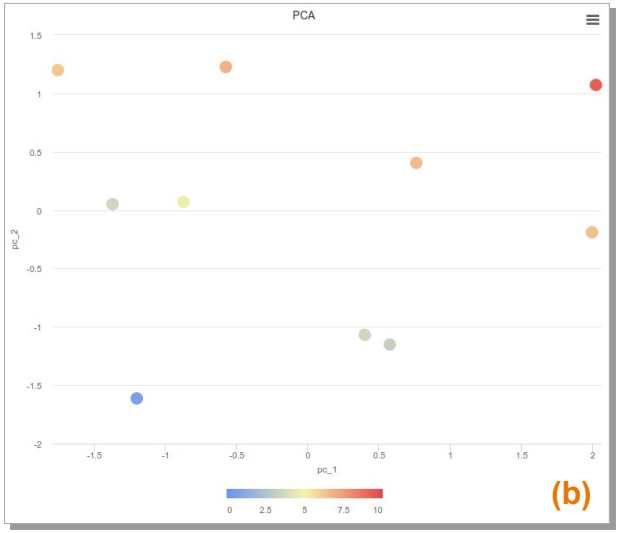
DEAKIN BUSINESS SCHOOL

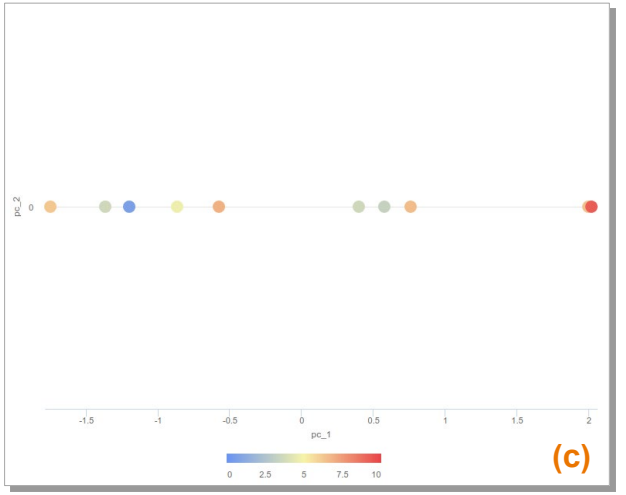
PCA – The works, Part 2

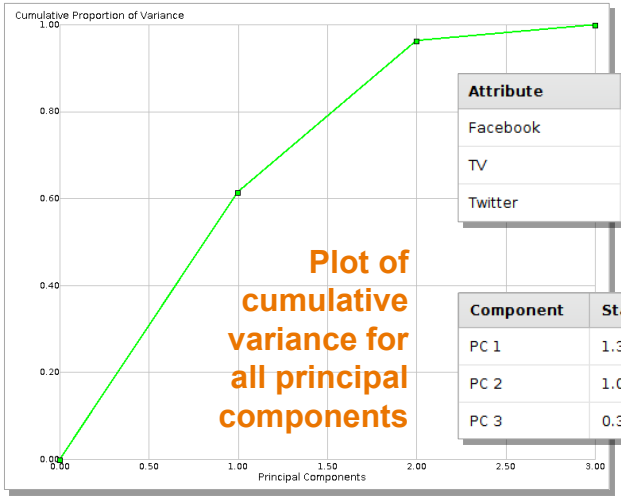
Example – Applying the PCA

- For PCA to work, data must be centered (around the mean) and standardised (with standard deviation) – in RM this can be achieved using the "Normalize" operator.
- When PCA is applied to normalised data, it generates a model consisting of eigenvectors and eigenvalues, which define the "rotation" needed to be applied to all new data, so that it ends up in the new coordinate system of PCs.
- It is also important to plot cumulative variance for all PCs, which tells how much variance each PC explains, also to determine if PCA visualisation is going to be effective.
- In this example, when a PCA model is applied, its first two PCs explain the most variance in data (here 90%), the remaining PCs explain much less variance (see 3D plot – a).
- When plotting transformed data in PC1 and PC2 axes and ignoring PC3, we still capture over 90% of data variance (b).
- We can further reduce data to a single dimension PC1, however, in this case the plot retains only 60% of data variance and as such a lot of information is lost (c).









Eigenvectors

Attribute	PC 1	PC 2	PC 3
Facebook	-0.510	0.685	0.520
TV	0.716	0.003	0.698
Twitter	0.476	0.729	-0.492

Eigenvalues

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.358	0.615	0.615
PC 2	1.021	0.348	0.963
PC 3	0.335	0.037	1.000

DEAKIN BUSINESS SCHOOL

How to evaluate clusters

- Clusters should consist of data points that have high degree of similarity (small average distance between cluster members and *centroid*).
- Clusters themselves (or their centroids) should be relatively dissimilar (large average distance between centroids).
- For many applications clusters should have a similar number of members (but not always).
- There should be a minimum number of unclustered data points (or small clusters).
- There are several approaches to measure the "goodness" of data clustering. RapidMiner provides several performance metrics for flat clusters, e.g.
  - *Distance measures*
  - *Density measures*
  - *Distribution measures*
- Such measures can be taken iteratively while varying model parameters, e.g.
  - k (the number of clusters).
- By plotting performance against clustering parameters, e.g.
  - We can select the best value of k by finding the smallest value of clustering performance metric, such as *Davies-Bouldin* or *WSS (within-cluster sum of squared errors)*

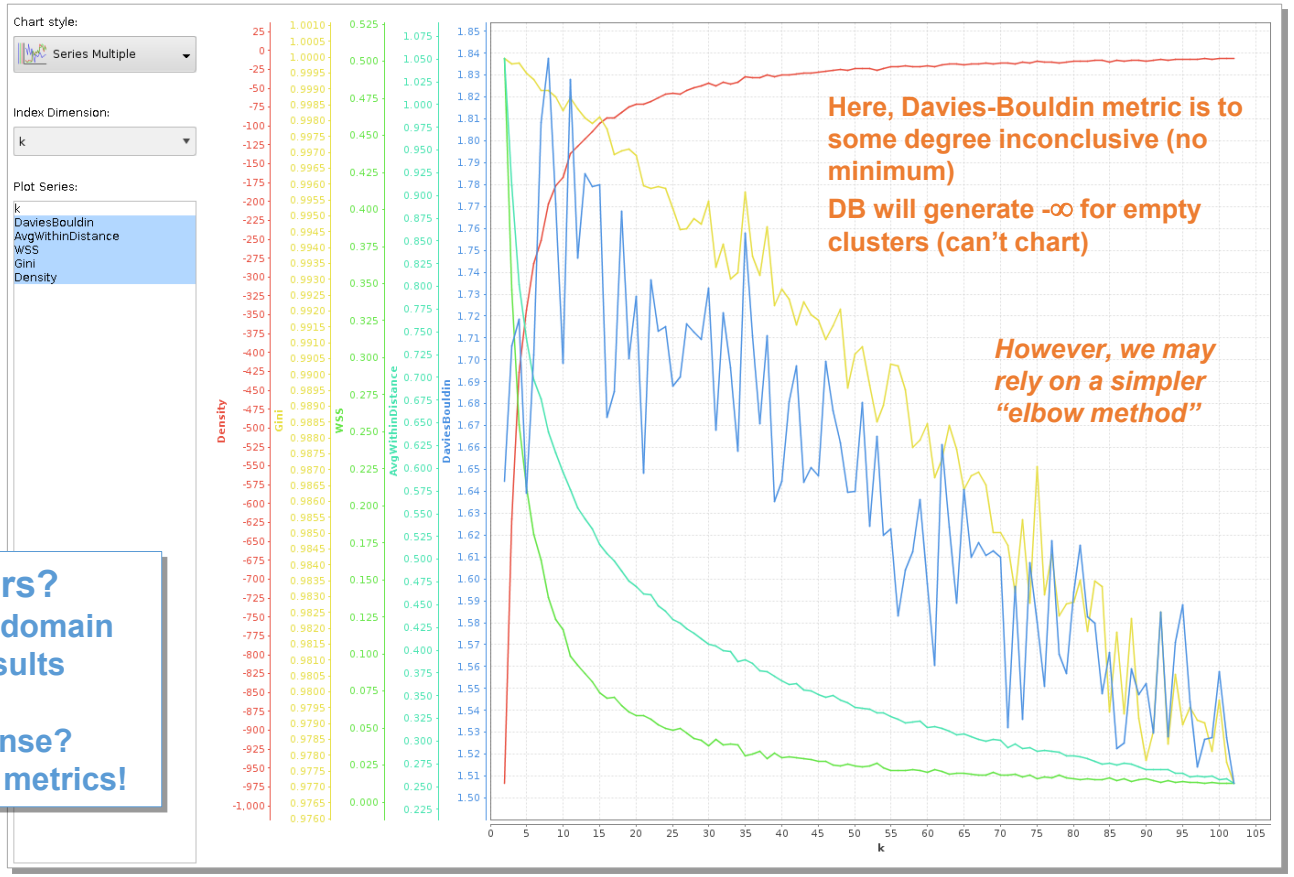
There are many other ways of measuring cluster quality, e.g. using Gini coefficient or Partition Coefficient, Entropy, Dunn index, Separation Index, Xie and Beni's Index, etc.

DEAKIN BUSINESS SCHOOL

How many clusters?

Example performance metrics that could be used in RapidMiner

- **Davies-Bouldin metric (dark blue)** is lowest (near zero) when clustering produces low intra-cluster distances (high similarity) and high inter-cluster distances (low similarity) – considered best
- **Sum of squares (green)** is a distribution measure, higher values indicating higher distribution



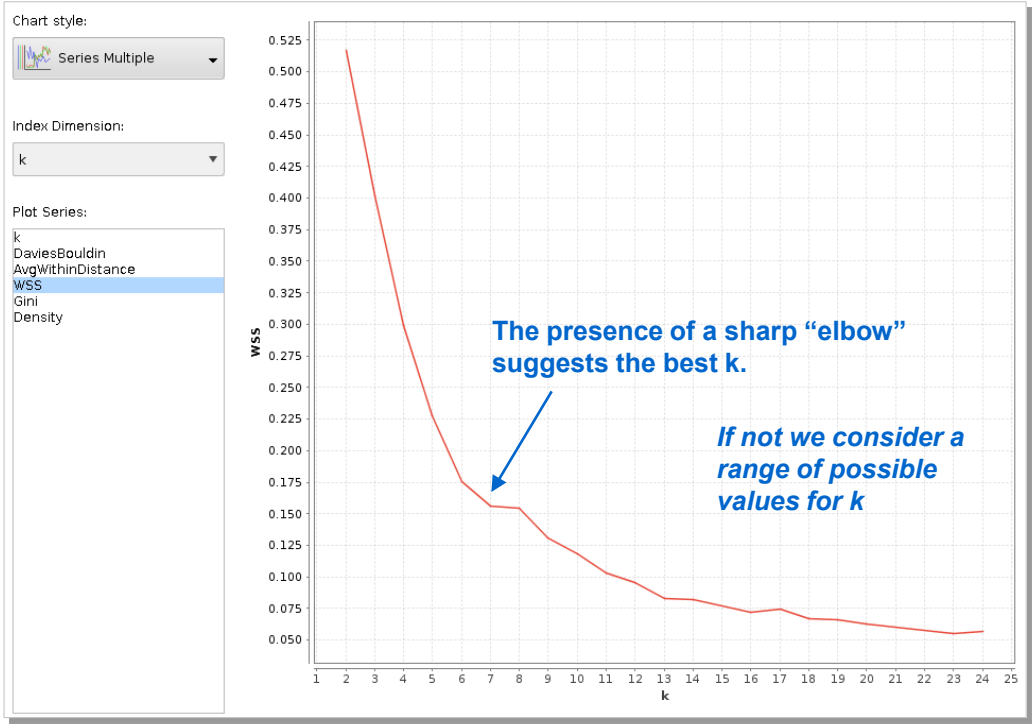
How many clusters?

- Understand the domain
- Visualise the results
- Experiment
- Does it make sense?
- Consult several metrics!

DEAKIN BUSINESS SCHOOL

The "Elbow Method"

- An "informal" method of finding the optimum number of clusters is using the ***elbow method***.
- The idea is to run k-means on a range of k values and for each value of k, for all data points within each cluster calculate the within-cluster sum of squared errors (***WSS***, also known as ***SSE***).
- As the number of clusters k equals the size of the data set, the WSS is zero (no error).
- So we will need to find the smallest value of k that WSS is already small and increasing the number of clusters will not significantly reduce the SSE.



- The plot of the k-WSS pairs looks like an arm. When we can identify a sharp line angle, or the "elbow" on the arm, this usually implies the best value of k (e.g. here k=7 but could also be k=13).
- If we cannot easily identify such a pronounced elbow then a range of possible values should be investigated.
- When the data is not well clustered, the elbow method will not work very well.

DEAKIN BUSINESS SCHOOL

Summary / Review

- What is data clustering?
- For what purpose is clustering used in analytics?
- What are the types of clusters used in data clustering?
- What are the main methods of data clustering?
- Describe the principle of k-means clustering.
- What are centroids?
- What are the main issues in k-means clustering?
- Are there any precautions about data attributes before applying k-means clustering?
- Describe three visualisations useful in cluster analysis.
- What is PCA? How can PCA help clusters visualisation and diagnostics?
- How can Davies-Bouldin index be used with data clustering? How should you interpret this metric?
- How can we optimise k-means clustering?
- Can data clustering improve the performance of predictive models? How?

DEAKIN BUSINESS SCHOOL