

# Advanced Data & Network Mining

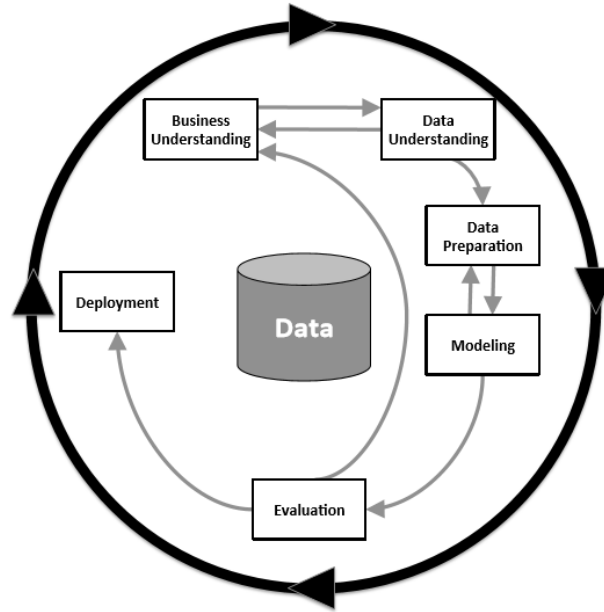
Modelling - Unsupervised Learning  
Clustering

2023-24

*terri.hoare@dbs.ie*

# Recap on Challenges and Methodology

## CRISP-DM (Cross Industry Standard Process for Data Mining)



**The CRISP-DM process, including the six key phases and the important relationships between them (adapted from Wirth and Hipp, 2000, repr. in Kelleher *et al.*, 2020, p.14)**

# Data Mining Taxonomy

## Matching Problems to Data Mining Algorithms

### Classification

Predicting a Categorical Target Variable (**supervised**)

### Regression

Predicting a Numeric Target Variable (**supervised**)

### Association

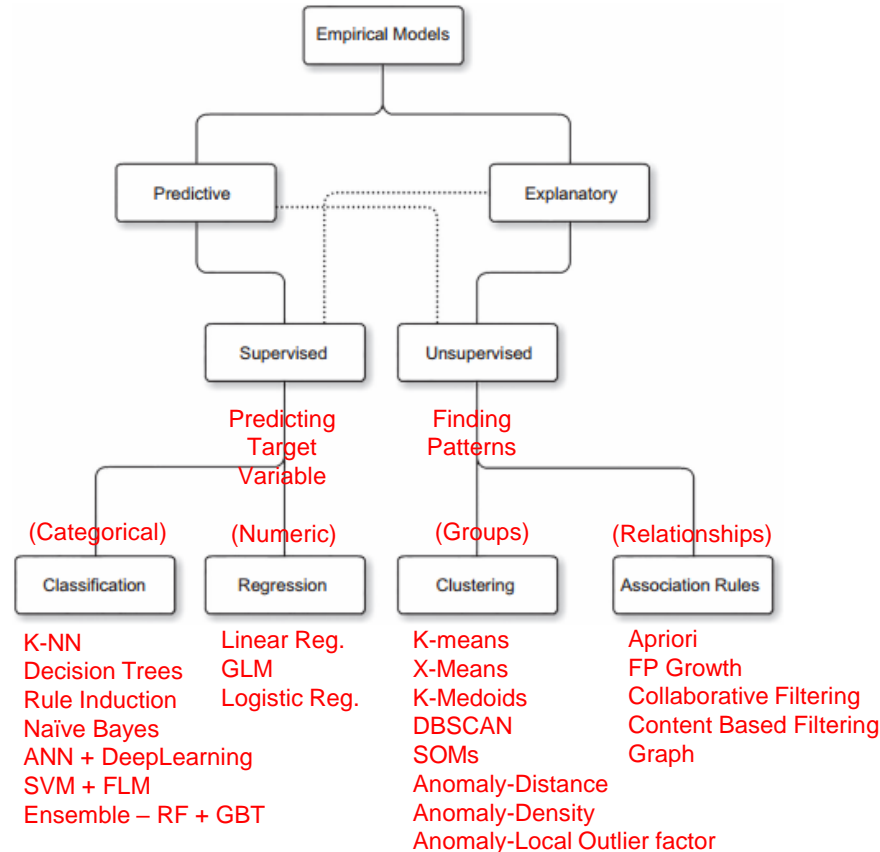
**Unsupervised** process for finding relationships between Items

### Clustering

**Unsupervised** process for finding meaningful groups in Data

# Data Mining

## Taxonomy of Algorithms



## Data Mining Unsupervised Clustering

Unsupervised learning task of **finding meaningful groups in data**.  
Important to distinguish between :-

- **Classification**, the process of identifying whether a data point belongs to a particular **known group**

(For example, categorizing a given voter as an known group “soccer mom” or not)

and

- **Clustering**, the process of **finding** and dividing data into **meaningful groups**

(For example, segregating a population of electorates into different groups based on similar demographics)

## Data Mining Unsupervised Clustering : classes of applications

Two different classes of applications :-

- **Describing a given data set**

and

- **A pre-processing step for other predictive algorithms**

## Data Mining Unsupervised

### Clustering applications : To describe the data

- **Marketing** Finding common groups of customers based on all past customer behaviours, potential customers' attributes and/or purchase patterns. Helpful to segment the customers, identify prototype customers (description of a typical customer of a group) and tailor a marketing message to the customers in a group
- **Document clustering** in text mining to group documents into groups of similar topics. (Used in routing customer support incidents, online content sites, forensic investigations etc.)
- **Session grouping** in web analytics. Understanding clusters of clickstream patterns and discovering different kinds of clickstream profile (for example one profile proceeding straight to checkout, another reading customer reviews to make purchase at a later session)

## Data Mining Unsupervised Clustering applications : For Pre-processing

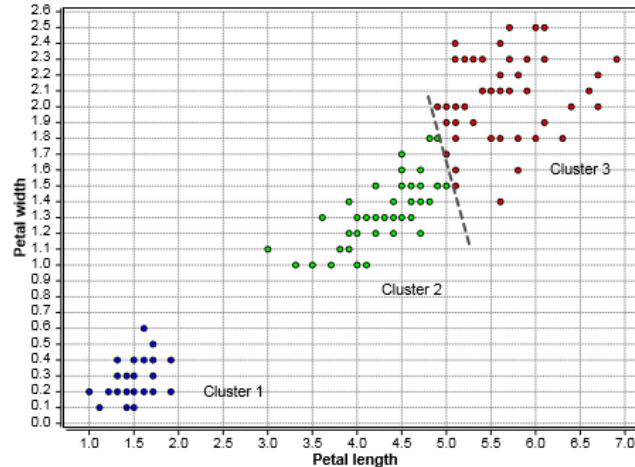
Since clustering processes consider all the attributes of the data set and “reduce” the information to a cluster which is in reality just another attribute, clustering can be used as a data compression technique with the output the cluster name for each record which can be used as input to other predictive data mining tasks. Two types of pre-processing :-

- **Clustering to reduce dimensionality** Computational complexity is proportional to number of dimensions
- **Clustering for object reduction** The prototype of a cluster is the most common representation of all the data objects. Reducing millions of customer records to say 100 prototype records allows algorithms such as k-NN (for which computational complexity is dependent on the number of records) to process the prototypes more efficiently for further classification or regression tasks



## Data Mining Unsupervised Clustering : Types of Techniques

All clustering techniques aim to find the groupings in the data in such a way that **data points within a cluster are more “similar” to each other than to data points in other clusters**. Below illustration on the Iris data set using Euclidean distance. All data points in cluster 2 are closer to other data points in cluster 2 than other data points in cluster 1



Example clustering of Iris data set without class labels.

## Data Mining Unsupervised

### Clustering : Techniques : Types of Cluster

- **Exclusive or strict partitioning clusters**

Each data object belongs to one exclusive cluster

Eg. Iris data set clusters

- **Overlapping clusters**

Multiview. Cluster groups are not exclusive and each data object may belong to more than one cluster

Eg. Customers can be grouped in a high-profit and a high-volume customer cluster at the same time

- **Hierarchical clusters**

Each child cluster can be merged to form a parent cluster

Eg. The most profitable cluster can be further divided into a long term customer cluster and a cluster with new customers with high value purchases

- **Fuzzy or probabilistic clusters**

Each data point belongs to all cluster groups with varying degrees of membership from 0 to 1

## Data Mining Unsupervised

### Clustering : Techniques : Algorithmic Approach

- **Prototype based clustering**

Eg. In clustering customer segments each customer cluster will have a central prototype customer and customers with similar properties are associated with the prototype customer of a cluster

- **Density clustering**

A cluster is defined as a dense region where data objects are concentrated, surrounded by a low density area where data objects are sparse. The data objects in low density areas are discarded as noise

## Data Mining Unsupervised Clustering : Techniques : Algorithmic Approach

- **Hierarchical clusters**

Used when data size is limited with interactive feedback. A tree diagram or **dendogram** is created either bottom up (agglomerative) where each data point is a cluster and clusters are merged or top down (divisive) where the data set is recursively divided into sub clusters until each data point forms a separate cluster

- **Model based clustering**

Distribution based clustering. (Eg. Gaussian or Poisson) where the parameter of the distribution is iteratively optimised. The entire data set can be represented by a mix of models

## Data Mining Unsupervised

### Clustering : Telecoms Customer Segmentation

Data Set for Customer Segmentation					
Customer ID	Location	Demographics	Call Usage	Data Usage	Monthly Bill
01	San Jose, CA	Male	1400	200 MB	\$75.23
02	Miami, FL	Female	2103	5,000 MB	\$125.78
03	Los Angeles, CA	Male	292	2,000 MB	\$89.90
04	San Jose, CA	Female	50	40 MB	\$59.34

- De-normalised telecommunications customer data including demographics, address, products used, revenue details, usage details of product, call volume, type of calls, call duration, time of call etc.
- Clustering allows going beyond traditional ways of grouping the data. A clustering algorithm consumes the data and groups the data with similar patterns into clusters based on **all** the attributes. The resulting clusters could be a group of customers who have low data usage but with high bills at a location where there is weak cellular coverage, which may indicate dissatisfied customers.

## Data Mining Unsupervised Clustering : Notes and Selected Algorithms

- The clustering algorithm doesn't explicitly provide the reason for clustering and doesn't intuitively label the cluster groups
- While clustering can be performed using a large number of attributes, it is up to the data mining practitioner to carefully select the attributes that will be relevant for clustering
- Common implementations of clustering :-

- **K-means**

Based on the prototype clustering technique

- **DBSCAN**

Density Based Spatial Clustering of Applications with Noise

- **SOM**

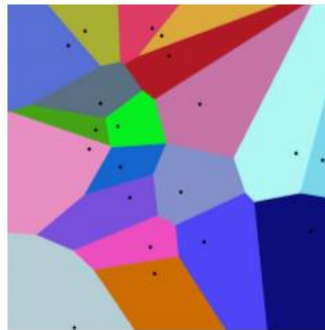
Novel approach called self-organising maps

## Data Mining Unsupervised Clustering : K-means

**Description** Data set is divided into k clusters by finding k centroids

The objective of K-means clustering is to find a prototype data point (centroid) for each cluster, all the data points are then assigned to the nearest prototype which then forms a cluster.

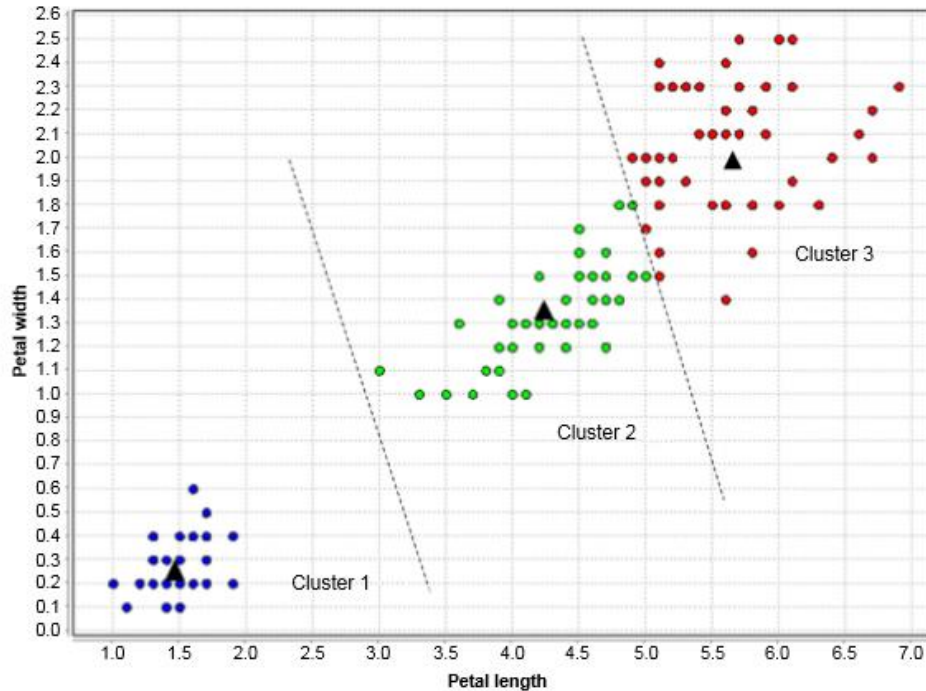
Based on the Lloyd-Forgy algorithm (1982). Visually the K-means algorithm divides the data space into k **Voronoi partitions** with each prototype a **seed** in a **Voronoi partition** and all other points associated to the nearest seed.



Voronoi partition.

# Data Mining Unsupervised

## Clustering : K-means : Iris Data Set (n=2, k=3)



Prototype-based clustering and boundaries.

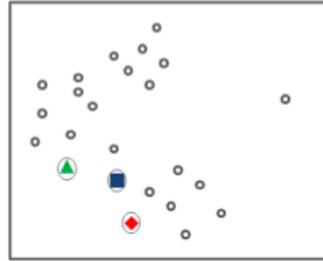


# Data Mining Unsupervised

## Clustering : K-means : How it Works

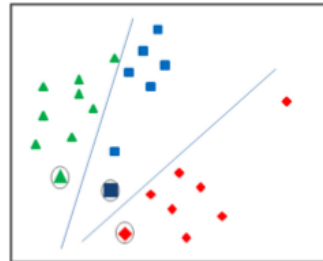
- **Step 1 – Initiate Centroids**

Initiate k random centroids



- **Step 2 – Assign Data Points**

All data points are assigned to “nearest” (in this case Euclidean distance) centroid to form a cluster.

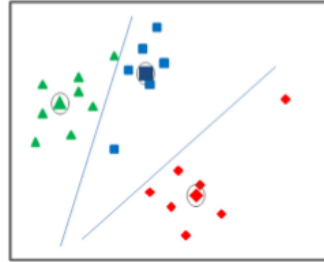


## Data Mining Unsupervised

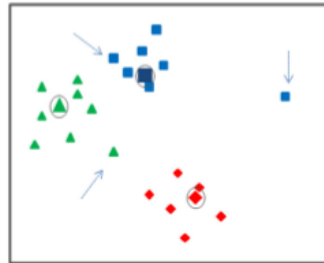
### Clustering : K-means : How it Works cont.

- **Step 3 – Calculate New Centroids**

For each cluster, calculate a new centroid by minimising the sum of squared errors (SSE) of all data points in a cluster to the centroid of the cluster. The centroid with minimal SSE is the new mean of the cluster



- **Step 4 – Repeat Assignment and Calculate New Centroids** (Note new – arrows)



- **Step 5 – Termination once iteration shows no significant change in centroids**

## Data Mining Unsupervised

### Clustering : K-means : Cluster Evaluation

Unsupervised or internal evaluation is required. (no known external labels).

- **SSE**

SSE is the average within cluster distance and can be calculated for each cluster and averaged across clusters.

**Good models will have low SSE both within and across clusters**

- **Davies-Bouldin Index**

Measure of uniqueness of the clusters and takes into account both the cohesiveness of the cluster (distance between the data points and the centre of the cluster) and separation between clusters. It is the function of the ratio of within cluster separation to the separation between the clusters.

**The lower the index, the better the cluster**

Both SSE and Davies Bouldin Index have the limitation of not guaranteeing better clustering when they have lower scores

# Data Mining Unsupervised

## Clustering : K-means : How to Implement cont.

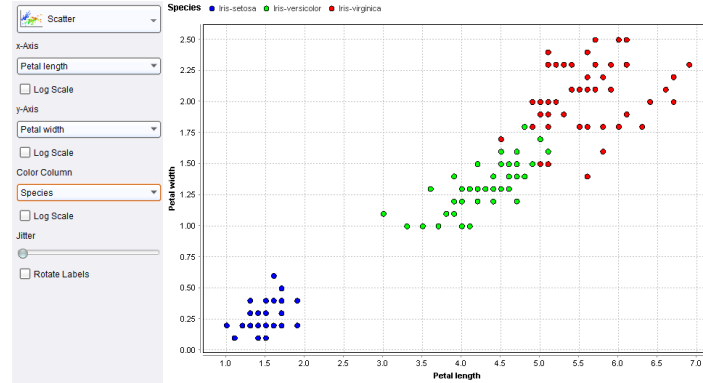
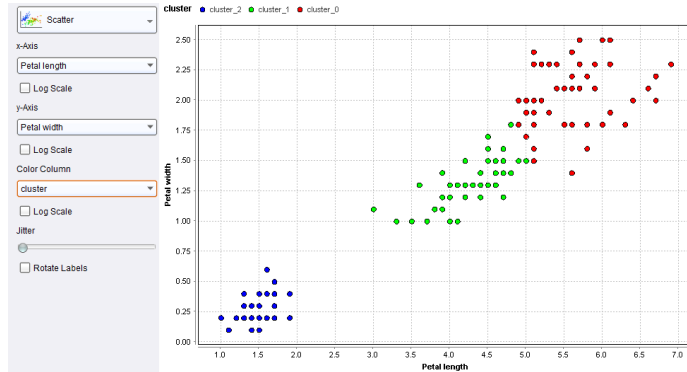
A labelled data set with inclusion of a clustering attribute

 Data	ExampleSet (150 examples, 3 special attributes, 2 regular attributes)					
 Statistics	Row No.	ID	Species	cluster	Petal length	Petal width
 Charts	1	id_1	Iris-setosa	cluster_2	1.400	0.200
 Advanced Charts	2	id_2	Iris-setosa	cluster_2	1.400	0.200
 Annotation	3	id_3	Iris-setosa	cluster_2	1.300	0.200
	4	id_4	Iris-setosa	cluster_2	1.500	0.200
	5	id_5	Iris-setosa	cluster_2	1.400	0.200
	6	id_6	Iris-setosa	cluster_2	1.700	0.400
	7	id_7	Iris-setosa	cluster_2	1.400	0.300
	8	id_8	Iris-setosa	cluster_2	1.500	0.200
	9	id_9	Iris-setosa	cluster_2	1.400	0.200
	10	id_10	Iris-setosa	cluster_2	1.500	0.100
	11	id_11	Iris-setosa	cluster_2	1.500	0.200
	12	id_12	Iris-setosa	cluster_2	1.600	0.200

# Data Mining Unsupervised

## Clustering : K-means : How to Implement cont.

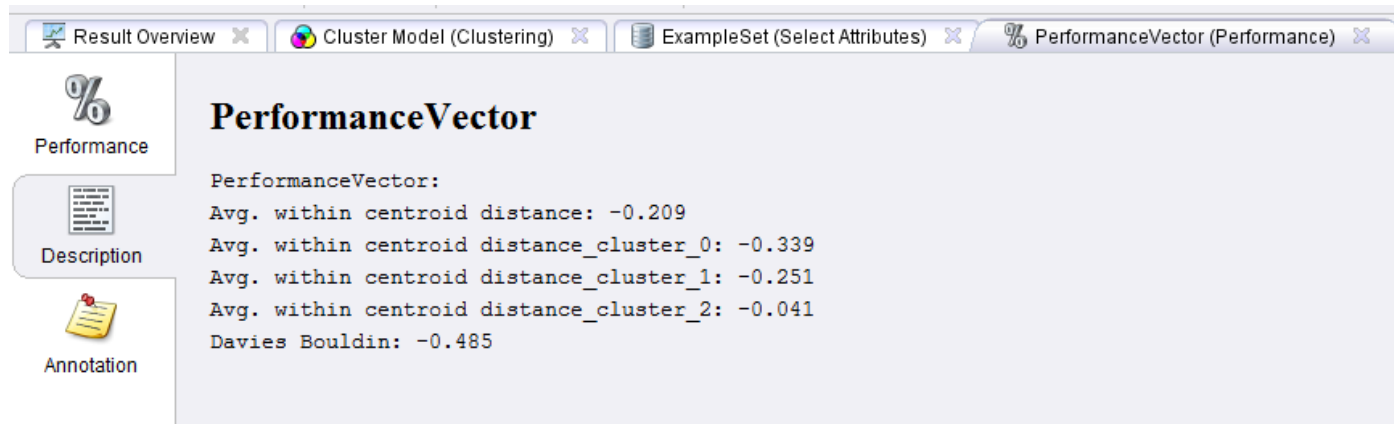
Charting cluster and species, observe only five data points in the border of versicolor and virginica are miss-clustered! The K-means process identified the different species in the data set almost exactly



# Data Mining Unsupervised

## Clustering : K-means : How to Implement cont.

The Performance vector includes the average distance measured and the Davies-Bouldin index. Amongst multiple clustering runs, a low average within-centroid distance and low Davies-Bouldin index yield better clusters, because they indicate cohesiveness of the cluster



The screenshot shows a software interface with four tabs: 'Result Overview', 'Cluster Model (Clustering)', 'ExampleSet (Select Attributes)', and 'PerformanceVector (Performance)'. The 'PerformanceVector' tab is active, displaying a sidebar with icons for 'Performance' (a percentage sign), 'Description' (a document), and 'Annotation' (a notepad). The main area of the 'PerformanceVector' window is titled 'PerformanceVector' and contains the following text:

```
PerformanceVector:  
Avg. within centroid distance: -0.209  
Avg. within centroid distance_cluster_0: -0.339  
Avg. within centroid distance_cluster_1: -0.251  
Avg. within centroid distance_cluster_2: -0.041  
Davies Bouldin: -0.485
```

## Data Mining Unsupervised

### Clustering : K-means : Discussion Points

- Simple and always converges to a local optimal solution
- Finding the global optimum dependent on choice of initiating centroids. This limitation can be addressed by having multiple random initiators or “runs” to find the clusters with minimal total SSE
- Susceptible to outliers which cannot always be removed by pre-processing (for example when identifying fraudulent transactions)
- Although the algorithm can effectively handle an n-dimensional data set, the operation will be expensive with a higher number of iterations and runs
- It relies on the user to assign the value of k, an arbitrary number can limit the ability to find the right number of natural clusters in the data set

## Data Mining Unsupervised

### Clustering : K-means : Discussion Points

- There are a number of methods to estimate the right number for  $k$  ranging from the Bayesian Information Criterion (BIC) to hierarchical methods that increase the value of  $k$  until the data points assigned to the cluster are Gaussian (bell-shaped, describing normal distribution)
- Recommend starting with a low single digit  $k$  and increasing until it fits
- K-means tends to find globular clusters whereas natural clusters are all shapes and sizes
- Good algorithm for quick evaluation of globular clusters and a pre-processing technique for predictive modelling and dimension reduction



## Clustering Models

### K-means Summary

#### **Model**

Algorithm finds  $k$  centroids and all the data points are assigned to the nearest centroids, which form a cluster

#### **Input**

No restrictions. However, the distance calculations work better with numeric data. Data should be normalized

#### **Output**

Data set is appended by one of  $k$  cluster labels

#### **Pros**

Simple to implement. Can be used for dimension reduction

#### **Cons**

Specification of  $k$  is arbitrary and may not find natural clusters. Sensitive to outliers

#### **Use Cases**

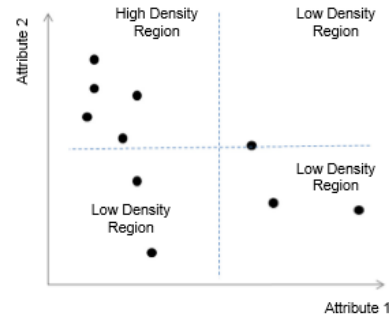
Customer segmentation, anomaly detection, applications where globular clustering is natural

## Data Mining Unsupervised Clustering : DBSCAN

**Description** Identifies clusters as a high-density area surrounded by low-density areas

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is one of the most commonly used density clustering algorithms (1996)

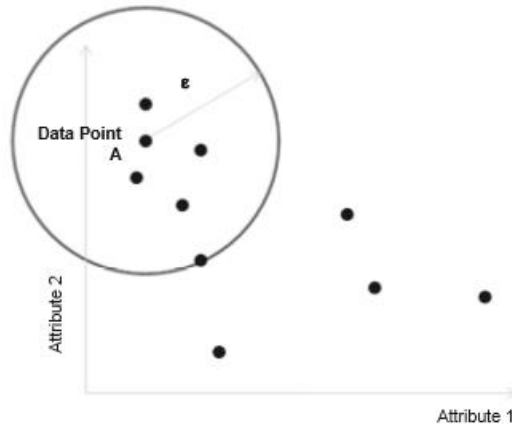
Density can be defined as the number of data points in a unit n-dimensional space.



Data set with two attributes.

## Data Mining Unsupervised Clustering : DBSCAN cont.

Density can also be measured within circular space around a point. The number of points within a circular space with radius  $\epsilon$  (epsilon) around a data point A as below is six. This measure is called centre-based density since the space considered is globular with the centre being the point considered



Density of a data point within radius  $\epsilon$ .

## Data Mining Unsupervised Clustering : DBSCAN : How it Works

- **Step 1 – Defining Epsilon and MinPoints**

We have to define a threshold of data points (MinPoints) above which the neighbourhood is considered high density. The number of data points inside the space is defined by radius  $\epsilon$

## Data Mining Unsupervised Clustering : DBSCAN : How it Works

- **Step 2 – Classification of Data Points**

In a data set with a given  $\epsilon$  and MinPoints, data points can be classified :-

- **Core Points**

All data points inside a high-density region of at least one data point are considered a core point. A high-density region is a space where there are at least MinPoints data points within a radius of  $\epsilon$  for any data points

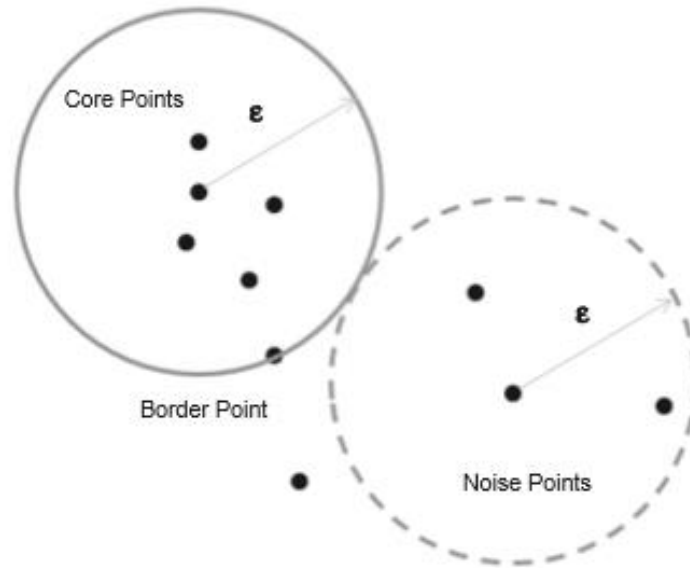
- **Border Points**

Sit on the circumference defines by  $\epsilon$ . A border point is the boundary between high density and low density space. Border points are counted within the high density space calculation

- **Noise Point**

Any point neither a core nor border point. They form a low density region around the high-density region

## Data Mining Unsupervised Clustering : DBSCAN : How it Works cont.



Core, border, and density points.

## Data Mining Unsupervised

### Clustering : DBSCAN : How it Works cont.

- **Step 3 – Clustering**

Once all data points in the data set are classified into density points, clustering is a straightforward task. Group of core points form distinct clusters. If two core points are within  $\varepsilon$  of each other, then both core points are within the same cluster. All these clustered core points form a cluster, which is surrounded by low density noise points. All noise points form low density regions around the high density cluster, and noise points are not classified in any cluster.

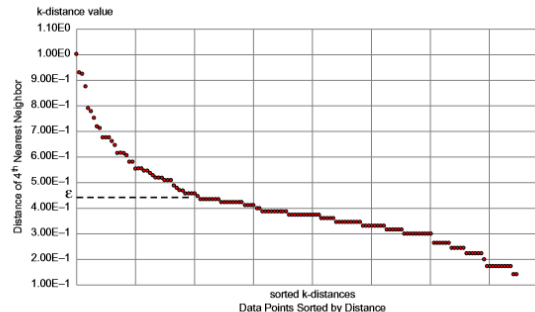
Note that as DBSCAN is a ‘partial’ clustering algorithm a few data points are left unlabelled or associated to a default noise cluster

# Data Mining Unsupervised

## Clustering : DBSCAN : How it Works cont.

- **Optimising MinPoints and  $\epsilon$**

We can estimate initial values by building a k-distribution graph. For a user specified value of k (say four data points) we can calculate the distance to the kth nearest neighbour for all data points in a data set. A k-distribution graph can be built by arranging all the k-distance values of individual data points in descending order. Points on the right of the chart will belong to data points inside a cluster because the distance is smaller. In most data sets the value of k-distance rises sharply after a particular value. The distance at which the chart rises is  $\epsilon$  and the value of k can be used for MinPoints



k-distribution chart for Iris data set with k = 4.



## Data Mining Unsupervised

### Clustering : DBSCAN : Discussion Points

- One of the key advantages is that there is no need to specify the number of clusters  $k$ . In many practical applications like finding unique customers or electoral segments, the number of clusters to be discovered will be unknown
- A K-distribution graph may be used to estimate  $\varepsilon$  and MinPoints
- K-means clustering is better at partitioning data sets with varying densities
- DBSCAN partitions data based on a certain threshold density. DBSCAN can find clusters of any shape and is not limited to finding the globular clusters typical of k-Means clustering. Given the complementary pros and cons of the two methods, it is advisable to cluster the data set by both methods and understand the patterns of both result sets

# Clustering Models

## DBSCAN: Summary

### **Model**

List of clusters and assigned data points, Default Cluster 0 contains noise points

### **Input**

No restrictions. However, the distance calculations work better with numeric data. Data should be normalized

### **Output**

Cluster labels based on identified clusters

### **Pros**

Finds the natural clusters of any shape. No number of clusters needed

### **Cons**

Specification of density parameters. A bridge between two clusters can merge the cluster. Can not cluster varying density data set

### **Use Cases**

Applications where clusters are non-globular shapes and when the prior number of natural groupings is not known

## Data Mining Unsupervised

### Clustering : SOMs (Self Organising Maps)

**Description** A visual clustering and data exploration technique with roots from neural networks and prototype clustering

First proposed by Teuvo Kohonen (1982), also known as Kohonen networks. They effectively arrange the data points in a lower dimensional space. A SOM is a form of neural network where the output is an organised visual matrix usually a two-dimensional grid with rows and columns. The objective of this neural network is to transfer all input data objects with  $n$  attributes ( $n$  dimensions) to the output lattice in such a way that objects next to each other are closely related to each other. In the hexagonal grid below, countries with similar GDP profiles are placed either in the same cells or next to each other.



## Data Mining Unsupervised Clustering : SOMs : How it Works

The Algorithm for building a SOM is similar to centroid-based clustering but with a neural network foundation. Since a SOM is essentially a neural network, the model only accepts numerical attributes. As it is an unsupervised learning model, there is no target variable. The objective of the algorithm is to find a set of centroids (neurons) to represent the cluster on an output grid. All the data objects from the data set are assigned to each centroid. Centroids closer to each other in the grid are more closely “related” to each other than to centroids further away in the grid.

As with many data mining algorithms a SOM tends to converge to a solution in most cases but doesn't guarantee an optimal solution. To tackle this problem, it is necessary to have multiple runs with various initiation measures and compare the results.

After the grids with the desired number of centroids have been built, any new data object can be quickly given a location on the grid space, based on its proximity to the centroids.

## Data Mining Unsupervised

### Clustering : SOMs : Implement Example

07\_Cluster\_7.3\_SOM\_extension\_country\_imf.rmp (07\_Cluster\_7.3\_IMFdata.csv)

#### **World Economic Outlook Database October 2012 by the IMF**

The data set has 186 records one for each country and four attributes in percentage of GDP, relative GDP invested, relative GDP saved, government revenue, and current account balance.

The objective of the clustering is to compare and contrast countries based on their percentage of GDP invested and saved, government revenue, and current account balance. Note that we are not comparing the size of the economy through absolute GDP but size of investment, national savings, current account, and size of government relative to the country's GDP.

The goal of this modelling exercise is to arrange countries in a grid so that countries with similar characteristics of investing, savings, size of government, and current account are placed next to each other. We are compressing four-dimensional information to a two-dimensional map or grid.

# Data Mining Unsupervised

## Clustering : SOMs : Implement Example

07\_Cluster\_7.3\_SOM\_extension\_country\_imf.rmp (07\_Cluster\_7.3\_IMFdata.csv)

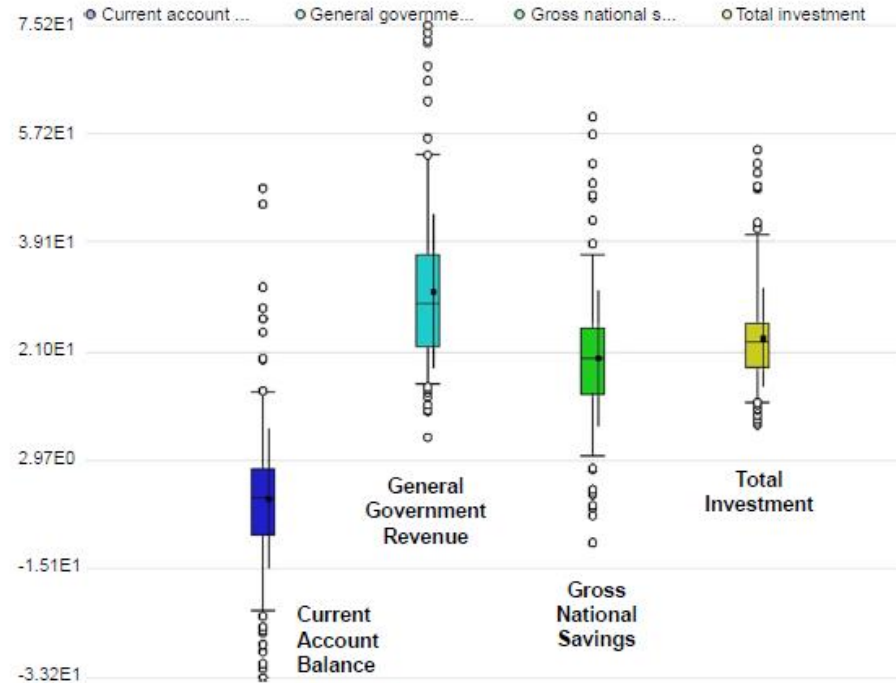
Row No.	Country	Current account balance	General government revenue	Gross national savings	Total investment
1	Afghanistan	3.877	21.977	30.398	26.521
2	Albania	-11.372	25.835	14.509	25.886
3	Algeria	7.489	36.458	48.947	41.428
4	Angola	9.024	43.479	21.692	12.668
5	Antigua and	-13.109	22.430	16.194	29.303
6	Argentina	0.658	37.199	22.595	24.451
7	Armenia	-14.653	20.970	16.660	31.313
8	Australia	-2.870	31.846	23.925	26.794
9	Austria	3.009	48.105	24.611	21.602
10	Azerbaijan	28.423	45.652	46.955	18.532
11	Bahrain	3.578	27.174	34.544	30.965

GDP by country data set.

# Data Mining Unsupervised

## Clustering : SOMs : Implement Example cont.

07\_Cluster\_7.3\_SOM\_extension\_country\_imf.rmp (07\_Cluster\_7.3\_IMFdata.csv)



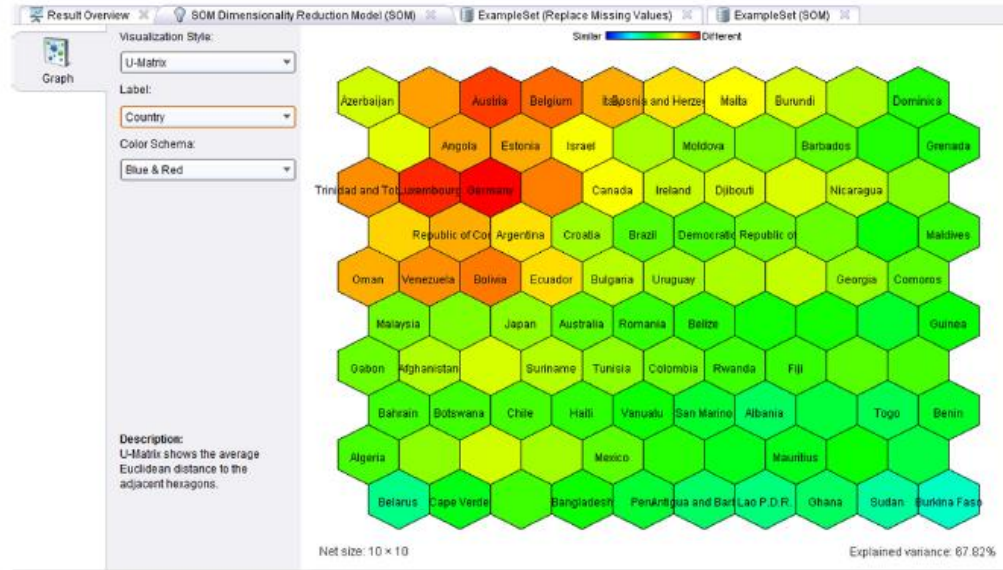
GDP by country: box-whisker (quartile) plot for all four attributes.

# Data Mining Unsupervised

## Clustering : SOMs : How to Implement cont.

### Execution and Interpretation

The output of the SOM modelling operator is a visual model (a lattice with centroids and mapped data points) and a grid data set (the example set labelled with location coordinates for each record in the grid lattice).



SOM output in the hexagonal grid.

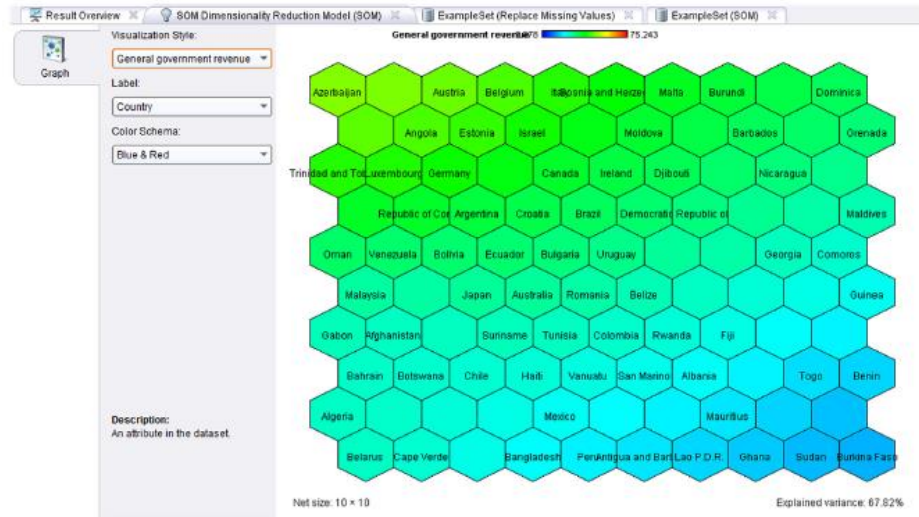


# Data Mining Unsupervised

## Clustering : SOMs : How to Implement cont.

### Execution and Interpretation cont.

Selecting “government revenue as a percentage of GDP” and countries with high government revenue as percentage of GDP are displayed top left of grid (Belgium 48%) with countries with low government revenue at bottom of grid (Bangladesh 11%)



SOM output with color overlay related to government revenue.

## Execution and Interpretation cont.

Visualization Style: Gross national savings

Label: Country

Color Schema: Blue & Red

Gross national savings: 0.000 to 60.015

Net size: 10 x 10

Explained variance: 67.82%

Country names visible in the grid include: Azerbaijan, Austria, Belgium, Bosnia and Herzegovina, Malta, Burundi, Dominica, Angola, Estonia, Israel, Moldova, Barbados, Grenada, Trinidad and Tobago, Luxembourg, Germany, Canada, Ireland, Djibouti, Nicaragua, Republic of Congo, Argentina, Croatia, Brazil, Democratic Republic of the Congo, Maldives, Oman, Venezuela, Bolivia, Ecuador, Bulgaria, Uruguay, Georgia, Comoros, Malaysia, Japan, Australia, Romania, Belize, Guinea, Gabon, Afghanistan, Suriname, Tunisia, Colombia, Rwanda, Fiji, Bahrain, Botswana, Chile, Haiti, Vanuatu, San Marino, Albania, Togo, Benin, Algeria, Belarus, Cape Verde, Bangladesh, Pakistan, Laos P.D.R., Ghana, Sudan, Burkina Faso, Mauritius, Mexico, and Mauritania.

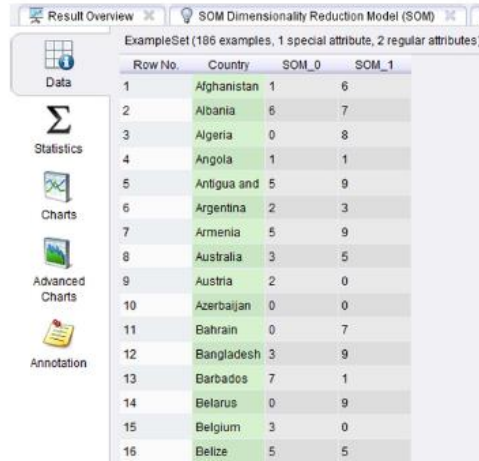
SOM output with color overlay related to national savings rate.

# Data Mining Unsupervised

## Clustering : SOMs : How to Implement cont.

### **Execution and Interpretation cont.**

Grid location co-ordinates for each country are produced. Row\_0 (X co-ordinate) and Row\_1 (Y co-ordinate) to allow post processing as required



Row No.	Country	SOM_0	SOM_1
1	Afghanistan	1	6
2	Albania	6	7
3	Algeria	0	8
4	Angola	1	1
5	Antigua and	5	9
6	Argentina	2	3
7	Armenia	5	9
8	Australia	3	5
9	Austria	2	0
10	Azerbaijan	0	0
11	Bahrain	0	7
12	Bangladesh	3	9
13	Barbados	7	1
14	Belarus	0	9
15	Belgium	3	0
16	Belize	5	5

SOM output with location coordinates.

## Data Mining Unsupervised

### Clustering : Self Organising Maps : Discussion Points

- Derived from both neural network and prototype clustering approaches
- An effective visual clustering tool to understand numeric high dimensional data. They reduce the features to two or three features used to specify the topology of the layout
- Predominately used as a visual discovery and data exploration technique. SOMs are used in combination with graph mining, text mining, and speech recognition

# Clustering Models

## SOMs (Self Organising Maps) Summary

### **Model**

A two-dimensional lattice where similar data points are arranged next to each other

### **Input**

No restrictions. However, the distance calculations work better with numeric data. Data should be normalized

### **Output**

No explicit clusters identified. Similar data points occupy either the same cell or are placed next to each other in the neighbourhood

### **Pros**

A visual way to explain the clusters. Reduces multidimensional data to two dimensions

### **Cons**

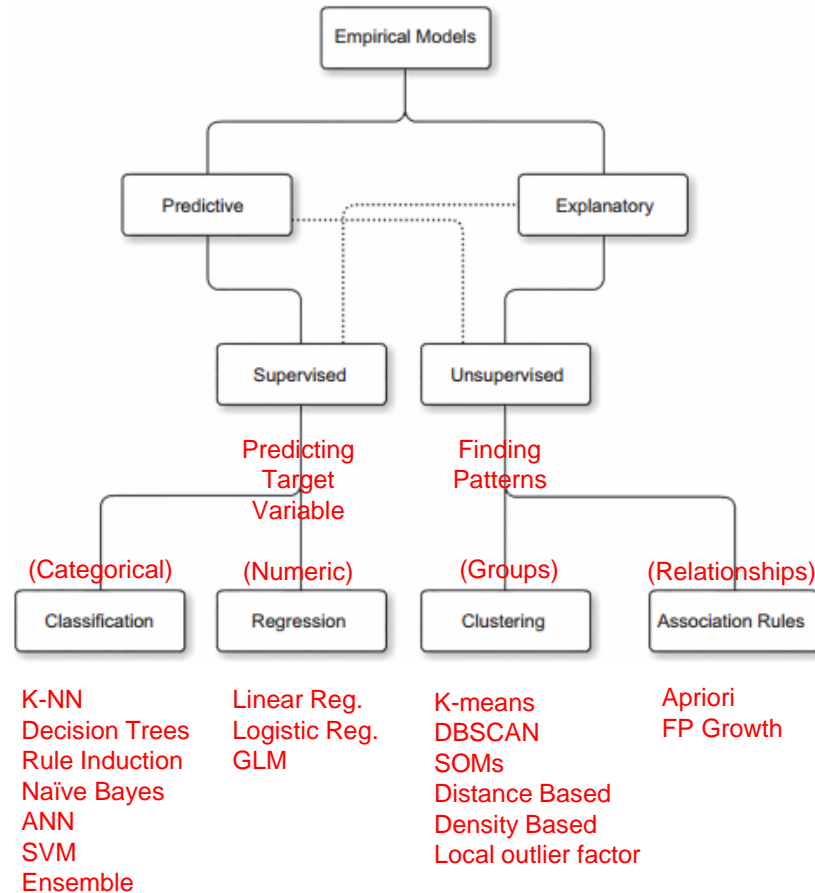
Number of centroids (topology) is specified by the user. Does not find natural clusters in the data

### **Use Cases**

Diverse applications including visual data exploration, content suggestions and dimension reduction

# Data Mining

## Recap on Taxonomy



## References

- Bache, K., & Lichman, M. (2013). UCI machine learning repository. University of California, School of Information and Computer Science. Retrieved from ,<http://archive.ics.uci.edu/ml..>
- Berry, M. J., & Linoff, G. (2000a). Converging on the customer: Understanding the customer behavior in the telecommunications industry. In M. J. Berry, & G. Linoff (Eds.), *Mastering data science: The art and science of customer relationship management* (pp. 357394). John Wiley & Sons, Inc.
- Berry, M. J., & Linoff, G. (2000b). Data science techniques and algorithms. In M. J. Berry, & G. Linoff (Eds.), *Mastering data science: The art and science of customer relationship management* (pp. 103107). John Wiley & Sons, Inc.
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224227.

## References

- Ester, M., Kriegel, H. -P., Sander, J., & Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In AAAI Press proceedings of 2nd international conference on knowledge discovery and data science KDD-96 (Vol. 96, pp. 226231). AAAI Press.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7, 179188. Retrieved from ,<https://doi.org/10.1111/j.1469-1809.1936.tb02137.x..>
- Germano, T. (March 23, 1999) Self-organizing maps. Retrieved from ,<http://davis.wpi.edu/Bmatt/courses/soms/>. Accessed 10.12.13



## References

- Hamerly, G., & Elkan, C. (2003). Learning the k in k-means. Advances in Neural Information Processing Systems, 17, 18. Available from <http://dx.doi.org/10.1.1.9.3574>. IMF, (2012, October). World economic outlook database. International Monetary Fund. Retrieved from <http://www.imf.org/external/pubs/ft/weo/2012/02/weodata/index.aspx>. Accessed 15.03.13.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 5969. Kohonen, T. (1988). The “neural” phonetic typewriter. Computer, IEEE, 21(3), 1122. Available from <https://doi.org/10.1109/2.28>.
- Liu, Y., Liu, M., & Wang, X. (2012). Application of self-organizing maps in text clustering: A review. In M. Johnsson (Ed.), Applications of self-organizing maps (pp. 205220). InTech. Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28, 129137.

## References

- Motl, J. (2012). SOM extension for rapid miner. Prague: Czech Technical University.
- Pearson, P., & Cooper, C. (2012). Using self organizing maps to analyze demographics and swing state voting in the 2008 U.S. presidential election. In N. Mana, F. Schwenker, & E. Trentin (Eds.), Artificial neural networks in pattern recognition ANNPR'12 Proceedings of the 5th INNS IAPR TC 3 GIRPR conference (pp. 201212). Heidelberg: Springer Berlin Heidelberg, Berlin10.1007/978-3-642-33212-8.

## References

- Resta, M. (2012). Graph mining based SOM: A tool to analyze economic stability. In M. Johnsson (Ed.), Applications of self-organizing maps (pp. 126). InTech. Retrieved from <http://www.intechopen.com/books/applications-of-self-organizing-maps..>
- Tan, P.-N., Michael, S., & Kumar, V. (2005). Clustering analysis: Basic concepts and algorithms. In P.-N. Tan, S. Michael, & V. Kumar (Eds.), Introduction to data science (pp. 487555). Boston, MA: Addison-Wesley.
- Witten, I. H., & Frank, E. (2005). Algorithms: The basic methods. Data science: Practical machine learning tools and techniques (pp. 136139). San Francisco, CA: Morgan Kaufmann