Oleksandr Romanko, Ph.D.
Senior Research Analyst, Risk Analytics, Watson Financial Services, IBM Canada
Adjunct Professor, University of Toronto

MIE1624H – Introduction to Data Science and Analytics

Lecture 7 – Data Mining and Machine Learning

University of Toronto
November 6, 2018

Data mining

- Data mining application classes of problems
  - Classification
  - Clustering
  - Regression
  - Forecasting
  - Others
- Hypothesis or discovery driven
- Iterative
- Scalable

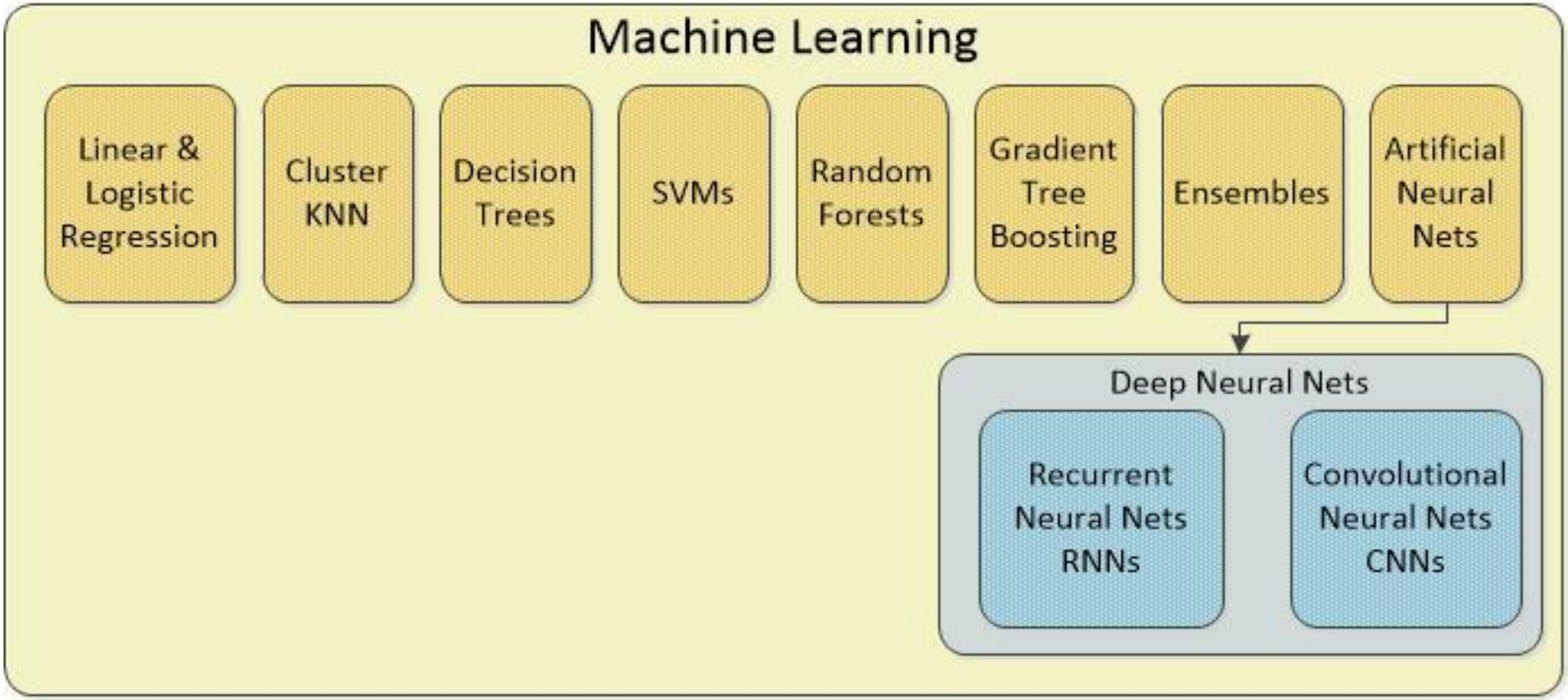
Machine learning

Machine learning gives computers the ability to learn without being explicitly programmed

- **Supervised learning**: decision trees, ensembles (bagging, boosting, random forests), k-NN, linear regression, Naive Bayes, neural networks, logistic regression, SVM
  - ❑ Classification
  - ❑ Regression (prediction)
- **Unsupervised learning**: k-means, c-means, hierarchical clustering, DBSCAN
  - ❑ Clustering
- **Dimensionality reduction**: PCA, LDA, factor analysis, t-SNE
- **Reinforcement learning**
  - ❑ Dynamic programming
- **Association rules**
  - ❑ Market basked analysis
- **Neural nets**: deep learning, multilayer perceptron, recurrent neural network (RNN), convolutional neural network (CNN)

Machine learning

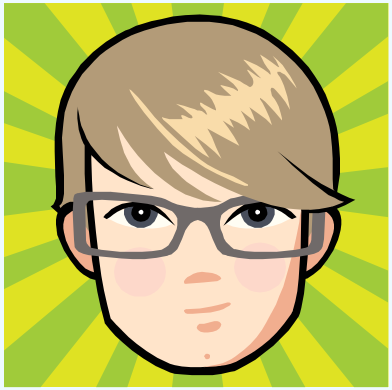
Machine learning gives computers the ability to learn without being explicitly programmed



What is the difference between descriptive (BI) and predictive analytics?

Descriptive

John
Lives in Seattle, zip: 98109
21 years old
iPhone 5
Plan: $98 a month
Talk: 400 minutes
Data: 1.9Gb
SMS: 370
Complaints: 0
Customer care calls: 1
Dropped calls: low



Mike
Lives in Atlanta, zip: 30308
38 years old
Samsung Galaxy S3
Plan: $78 a month
Talk: 1200 minutes
Data: 0.2 Gb of data
SMS: 8
Customer care calls: 6
Dropped calls: high

Predictive

Low churn risk

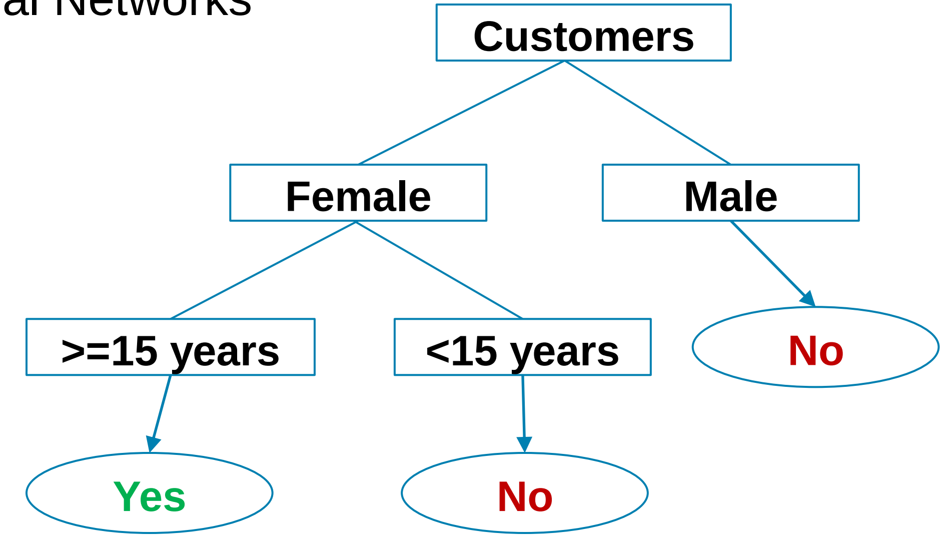
High churn risk

📖 Classification – Decision Trees

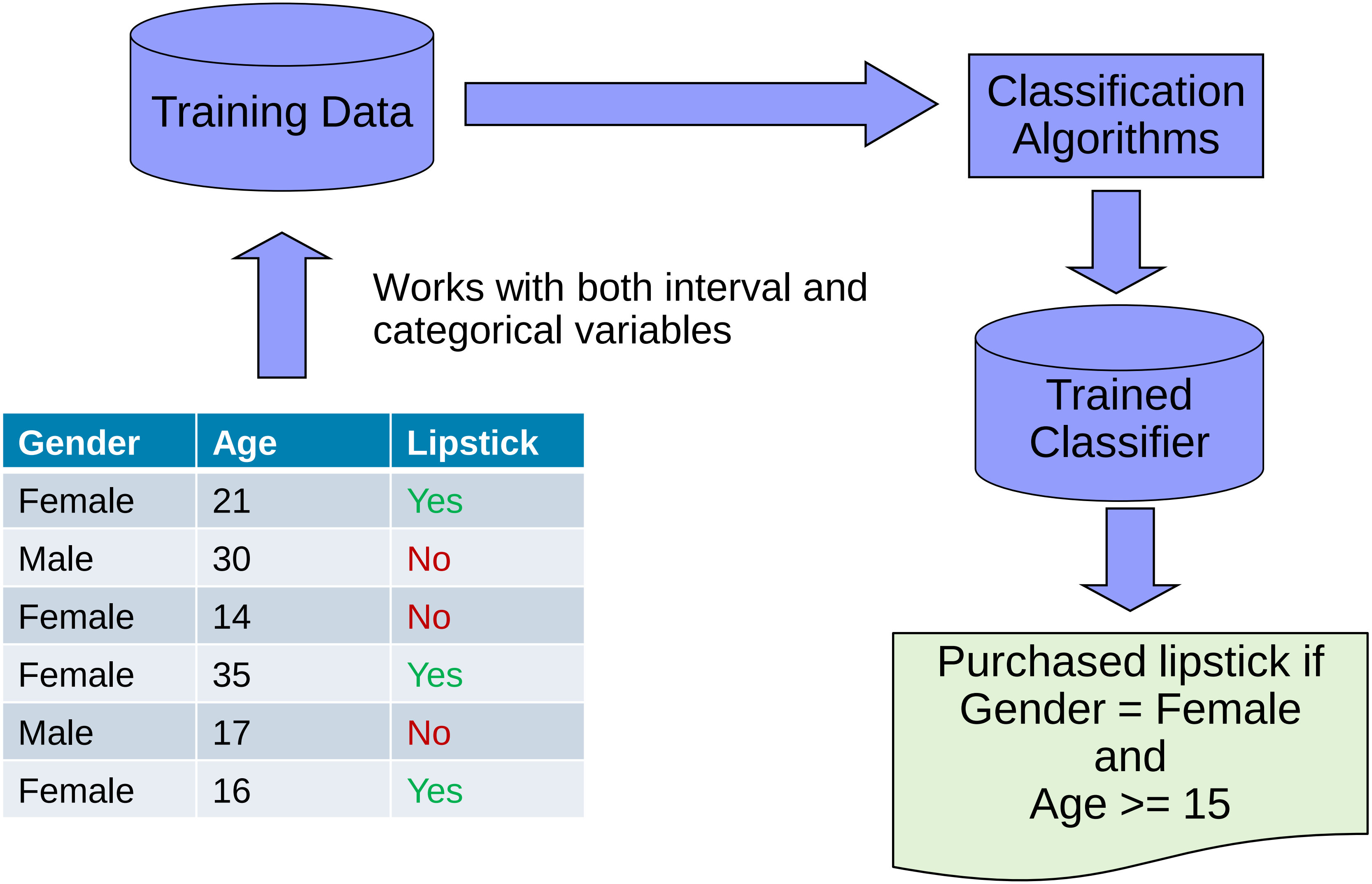
Classification

- Classification is a supervised learning technique, which maps data into predefined classes or groups
- Training set contains a set of records, where one of the records indicates class
- Modeling objective is to assign a class variable to all of the records, using attributes of other variables to predict a class
- Data is divided into test / train, where “train” is used to build the model and “test” is used to validate the accuracy of classification
- Typical techniques: Decision Trees, Neural Networks

Gender	Age	Lipstick
Female	21	Yes
Male	30	No
Female	14	No
Female	35	Yes
Male	17	No
Female	16	Yes



Classification: creating model



Works with both interval and categorical variables

Gender	Age	Lipstick
Female	21	Yes
Male	30	No
Female	14	No
Female	35	Yes
Male	17	No
Female	16	Yes

Purchased lipstick if Gender = Female and Age >= 15

Classification: applying rules

Gender	Age	Lipstick
Female	27	?
Male	55	?
Female	47	?
Male	39	?
Female	27	?
Male	19	?

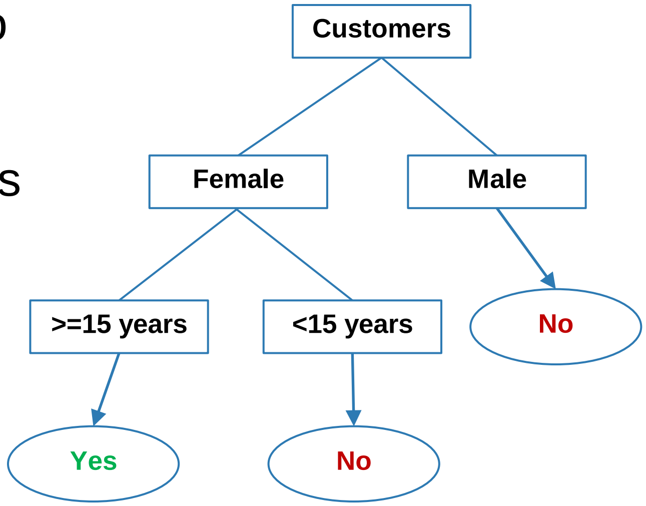
Apply Scoring

Gender	Age	Lipstick
Female	27	P Yes
Male	55	P No
Female	47	P Yes
Male	39	P No
Female	27	P Yes
Male	19	P No

If
Gender = Female
and
Age >= 15 then
Purchase lipstick = YES

Decision (classification) trees

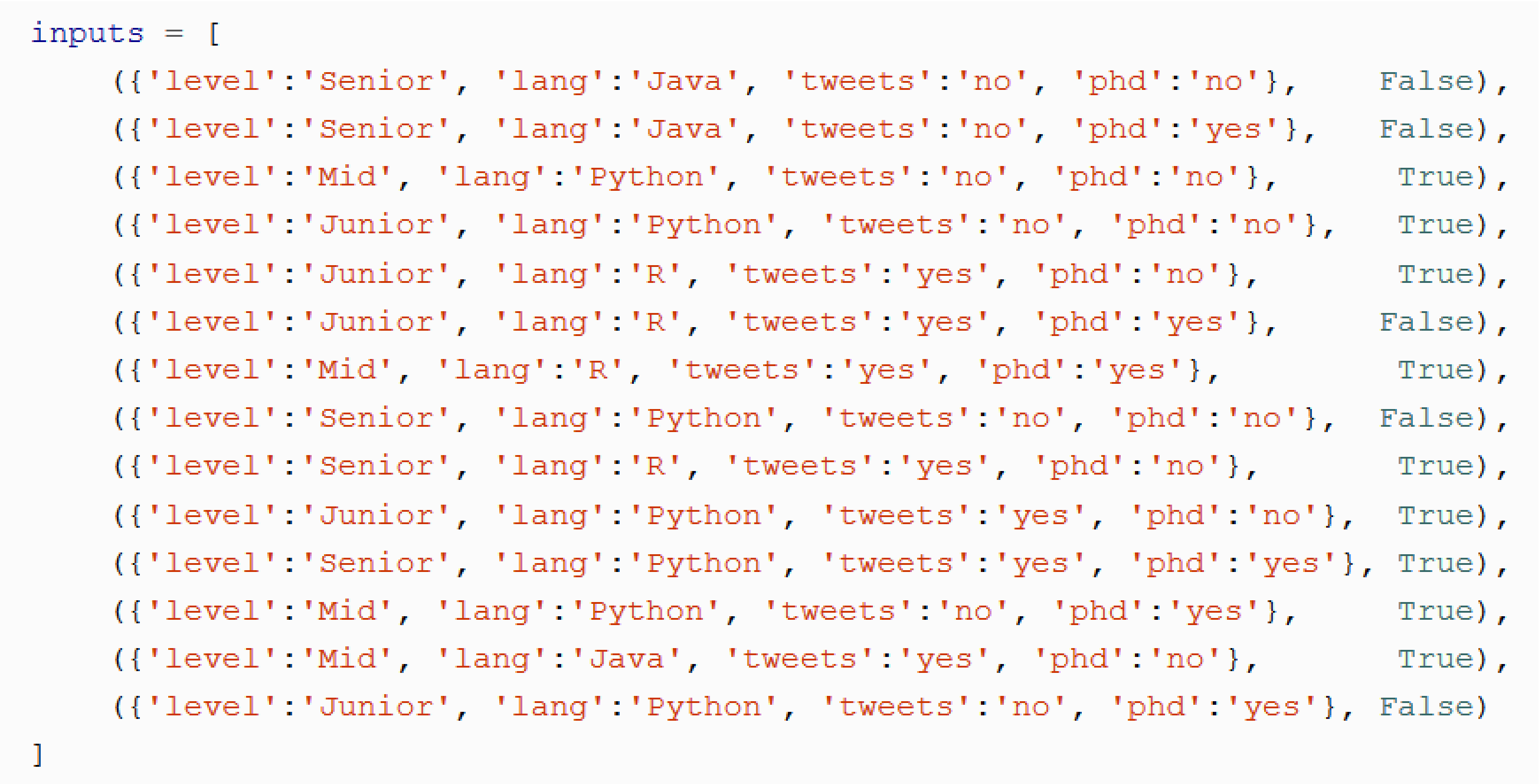
- A tree can be "learned" by splitting the source set into subsets based on an attribute value test
- Tree partitions samples into mutually exclusive groups by selecting the best splitting attribute, one group for each terminal node
- The process is repeated recursively for each derived subset, until the stopping criteria is reached



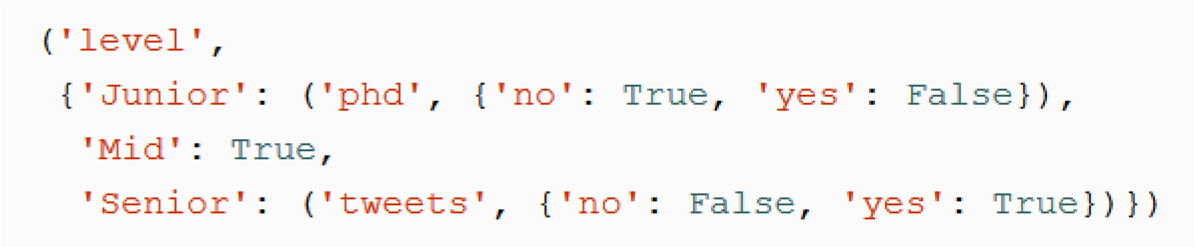
- ➢ Works with both interval and categorical variables
- ➢ No need to normalize the data
- ➢ Intuitive if-then rules are easy to extract and apply
- ➢ Best applied to binary outcomes

- Decision trees can be used to support multiple modeling objectives
  - ○ Customer segmentation
  - ○ Investment / portfolio decisions
  - ○ Issuing a credit card or loan
  - ○ Medical patient / disease classification

Decision (classification) trees

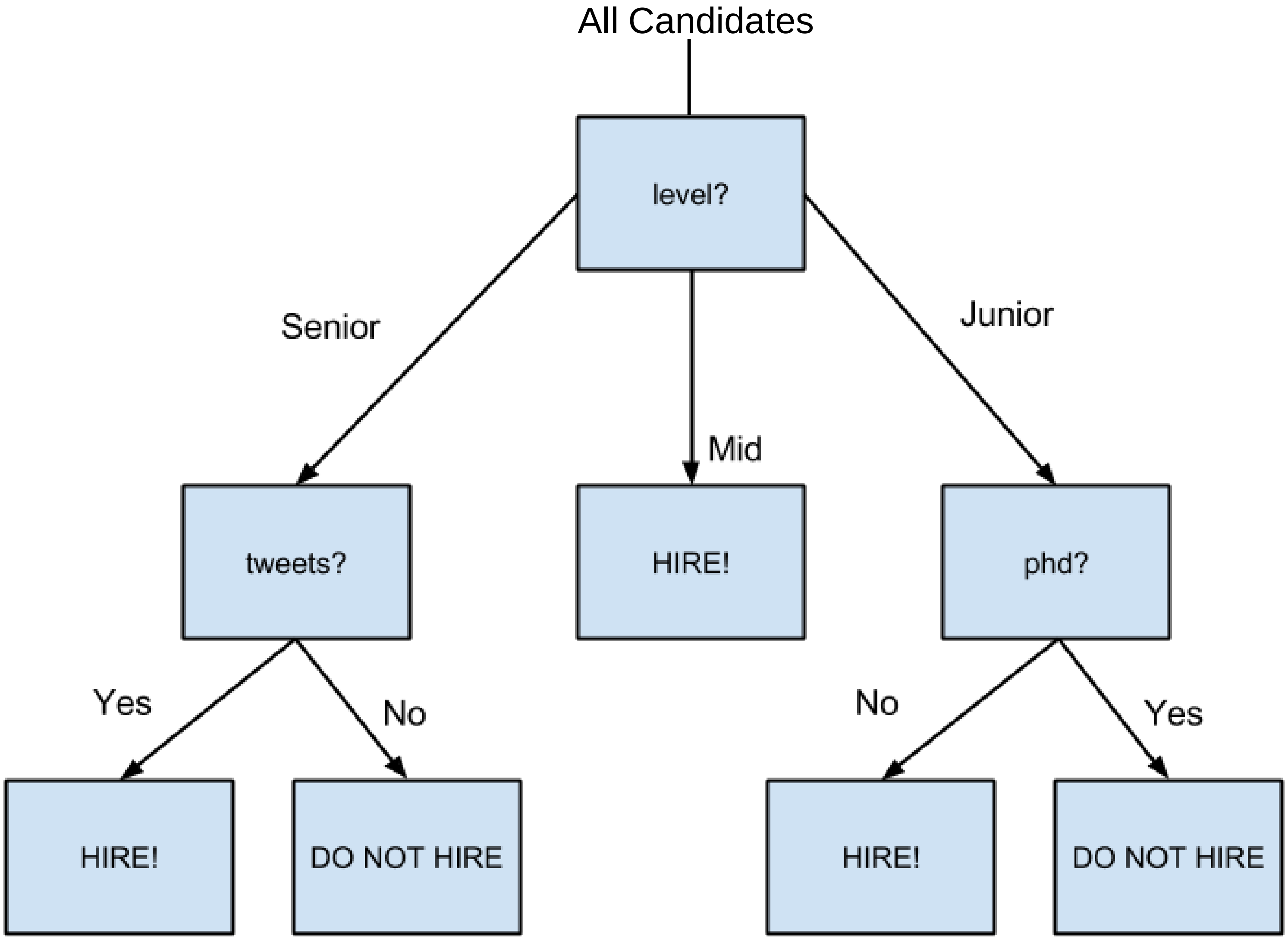


```
inputs = [
    ({'level':'Senior', 'lang':'Java', 'tweets':'no', 'phd':'no'},    False),
    ({'level':'Senior', 'lang':'Java', 'tweets':'no', 'phd':'yes'},   False),
    ({'level':'Mid', 'lang':'Python', 'tweets':'no', 'phd':'no'},      True),
    ({'level':'Junior', 'lang':'Python', 'tweets':'no', 'phd':'no'},   True),
    ({'level':'Junior', 'lang':'R', 'tweets':'yes', 'phd':'no'},       True),
    ({'level':'Junior', 'lang':'R', 'tweets':'yes', 'phd':'yes'},     False),
    ({'level':'Mid', 'lang':'R', 'tweets':'yes', 'phd':'yes'},         True),
    ({'level':'Senior', 'lang':'Python', 'tweets':'no', 'phd':'no'},  False),
    ({'level':'Senior', 'lang':'R', 'tweets':'yes', 'phd':'no'},       True),
    ({'level':'Junior', 'lang':'Python', 'tweets':'yes', 'phd':'no'},  True),
    ({'level':'Senior', 'lang':'Python', 'tweets':'yes', 'phd':'yes'}, True),
    ({'level':'Mid', 'lang':'Python', 'tweets':'no', 'phd':'yes'},     True),
    ({'level':'Mid', 'lang':'Java', 'tweets':'yes', 'phd':'no'},       True),
    ({'level':'Junior', 'lang':'Python', 'tweets':'no', 'phd':'yes'}, False)
]
```



```
('level',
 {'Junior': ('phd', {'no': True, 'yes': False}),
  'Mid': True,
  'Senior': ('tweets', {'no': False, 'yes': True})})
```

Decision (classification) trees



Root node

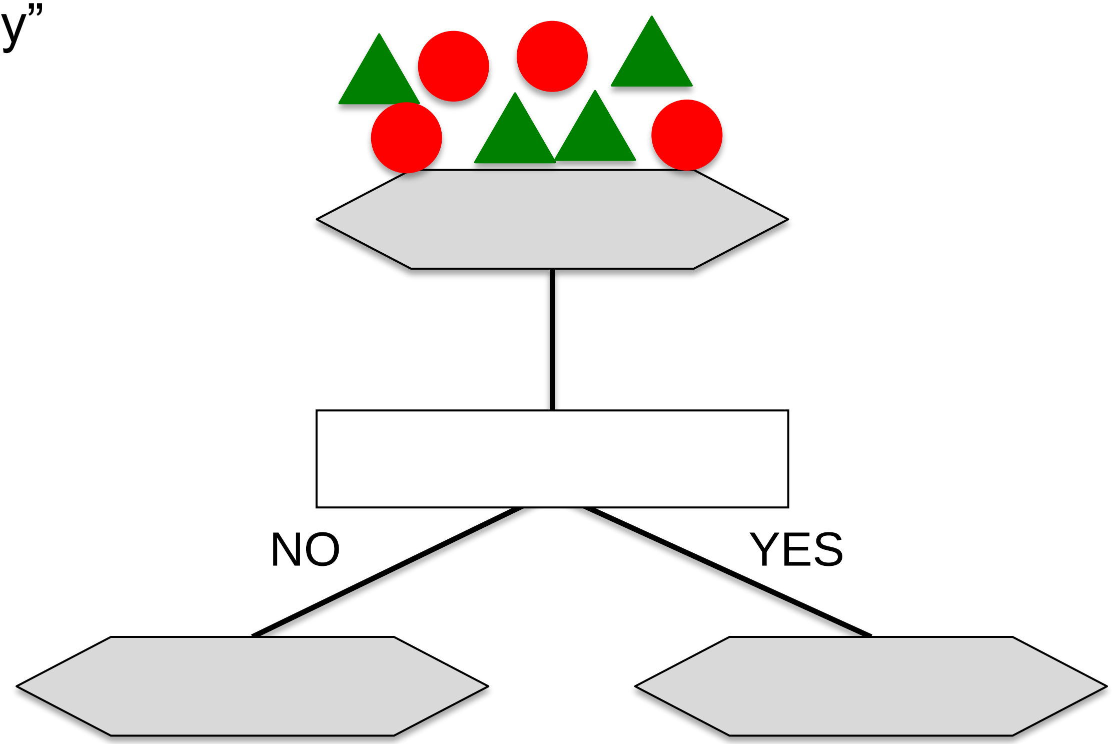
Decision node

Branches

Leaf nodes

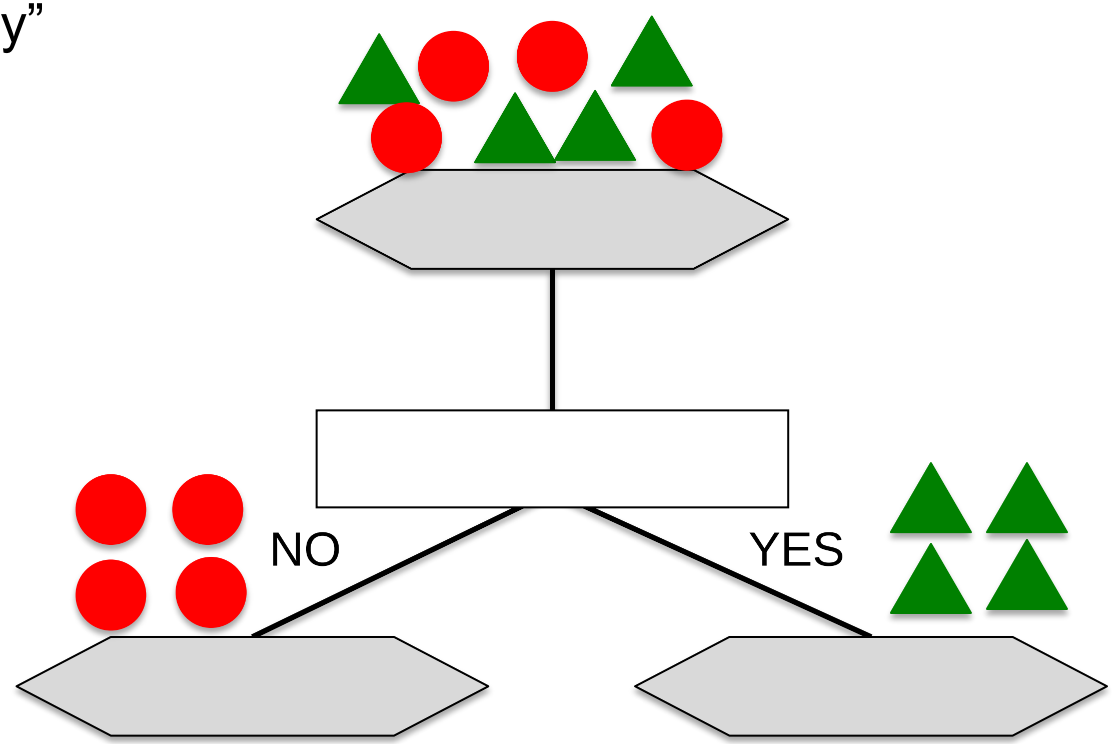
Understanding decision trees

- Decision trees are built using recursive partitioning to classify the data
- The algorithm chooses the most predictive feature to split the data on
- “Predictiveness” is based on decrease in entropy (gain in information) or “impurity”



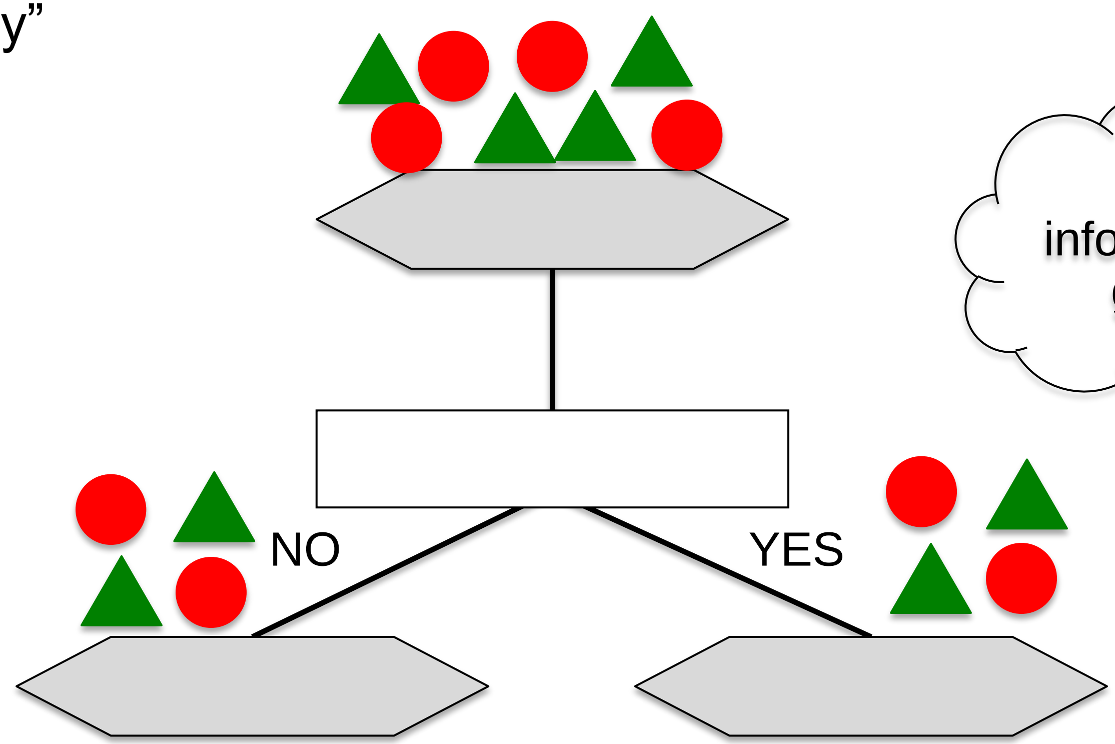
Understanding decision trees

- Decision trees are built using recursive partitioning to classify the data
- The algorithm chooses the most predictive feature to split the data on
- “Predictiveness” is based on decrease in entropy (gain in information) or “impurity”



Understanding decision trees

- Decision trees are built using recursive partitioning to classify the data
- The algorithm chooses the most predictive feature to split the data on
- “Predictiveness” is based on decrease in entropy (gain in information) or “impurity”



No information gain!

Understanding decision trees

- Root node partitions the data using the feature that provides the most information gain

- Information gain tells us how important a given attribute of the feature vectors is:

$$\text{Information Gain} = \text{entropy(parent) - average entropy(children)}$$

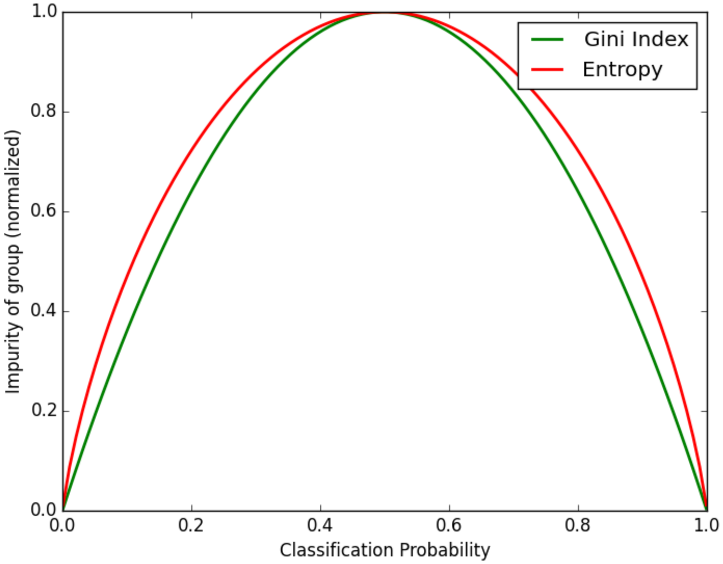
- Entropy is a common measure of target class impurity (*i* is each of the target classes, p_i is proportion of the number of elements in class 0 or 1):

$$\text{Entropy} = \sum_i -p_i \log_2 p_i$$

- Gini Index is another measure of impurity:

$$\text{Gini} = 1 - \sum_i p_i^2$$

Gini impurity is computationally faster as it doesn't require calculating logarithmic functions, though in reality which of the two methods is used rarely makes too much of a difference.



Characteristics of decision trees

Pros	Cons
Easy to interpret	Easy to overfit or underfit the model
Can handle numeric or categorical features	Cannot model interactions between features
Can handle missing data	Large trees can be difficult to interpret
Uses only the most important features	
Can be used on very large or small data	

A tree stops growing at a node when…

- pure or nearly pure
- no remaining variables on which to further subset the data
- the tree has grown to a preselected size limit

Bias-variance tradeoff

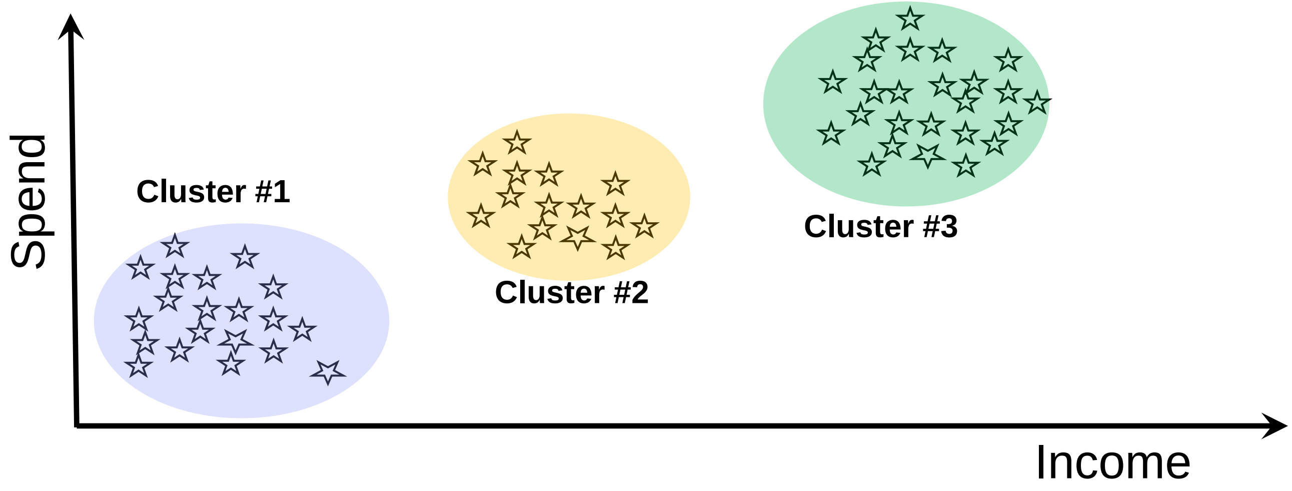
error = bias + variance

- **Bias-variance tradeoff** is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:
  - ❑ **Bias** is error from erroneous assumptions in the learning algorithm, high bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting)
  - ❑ **Variance** is error from sensitivity to small fluctuations in the training set, high variance can cause overfitting, i.e., modeling the random noise in the training data, rather than the intended outputs
- **Ensemble tree methods**:
  - ❑ Gradient Boosting (GBoost) is based on weak learners (high bias, low variance). In terms of decision trees, weak learners are shallow trees, sometimes even as small as decision stumps (trees with two leaves). Boosting reduces error mainly by reducing bias.
  - ❑ Random Forest uses fully grown decision trees (low bias, high variance). It tackles the error reduction task by reducing variance. The trees are made uncorrelated to maximize the decrease in variance, but the algorithm cannot reduce bias (which is slightly higher than the bias of an individual tree in the forest).

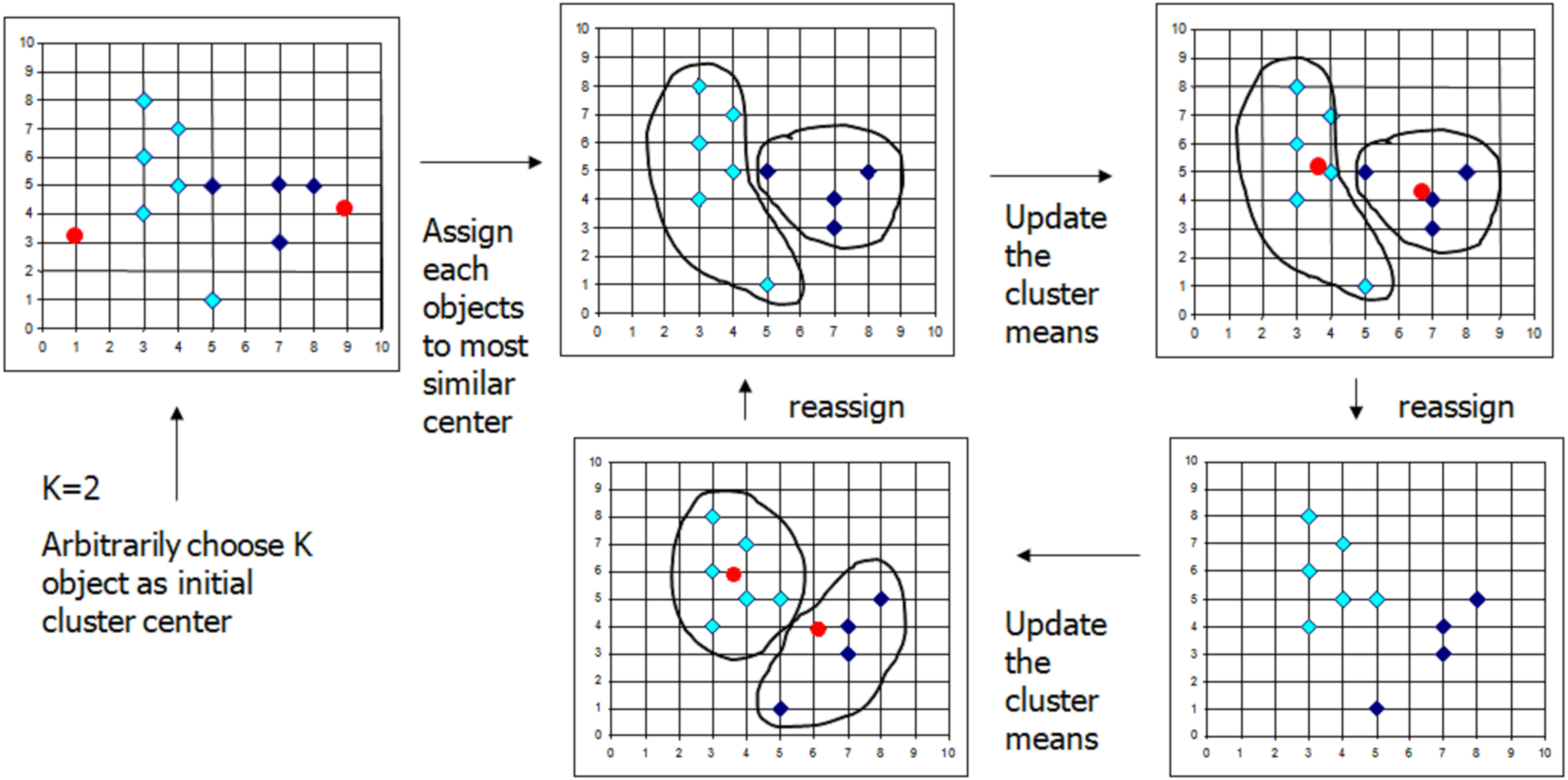
📖 Clustering

Cluster analysis (segmentation)

- Unsupervised learning algorithm
  - Unlabeled data and no “target” variable
- Frequently used for segmentation (to identify natural groupings of customers)
  - Market segmentation, customer segmentation
- Most cluster analysis methods involve the use of a distance measure to calculate the closeness between pairs of items
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another



K-means clustering



K=2

Arbitrarily choose K object as initial cluster center

Assign each objects to most similar center

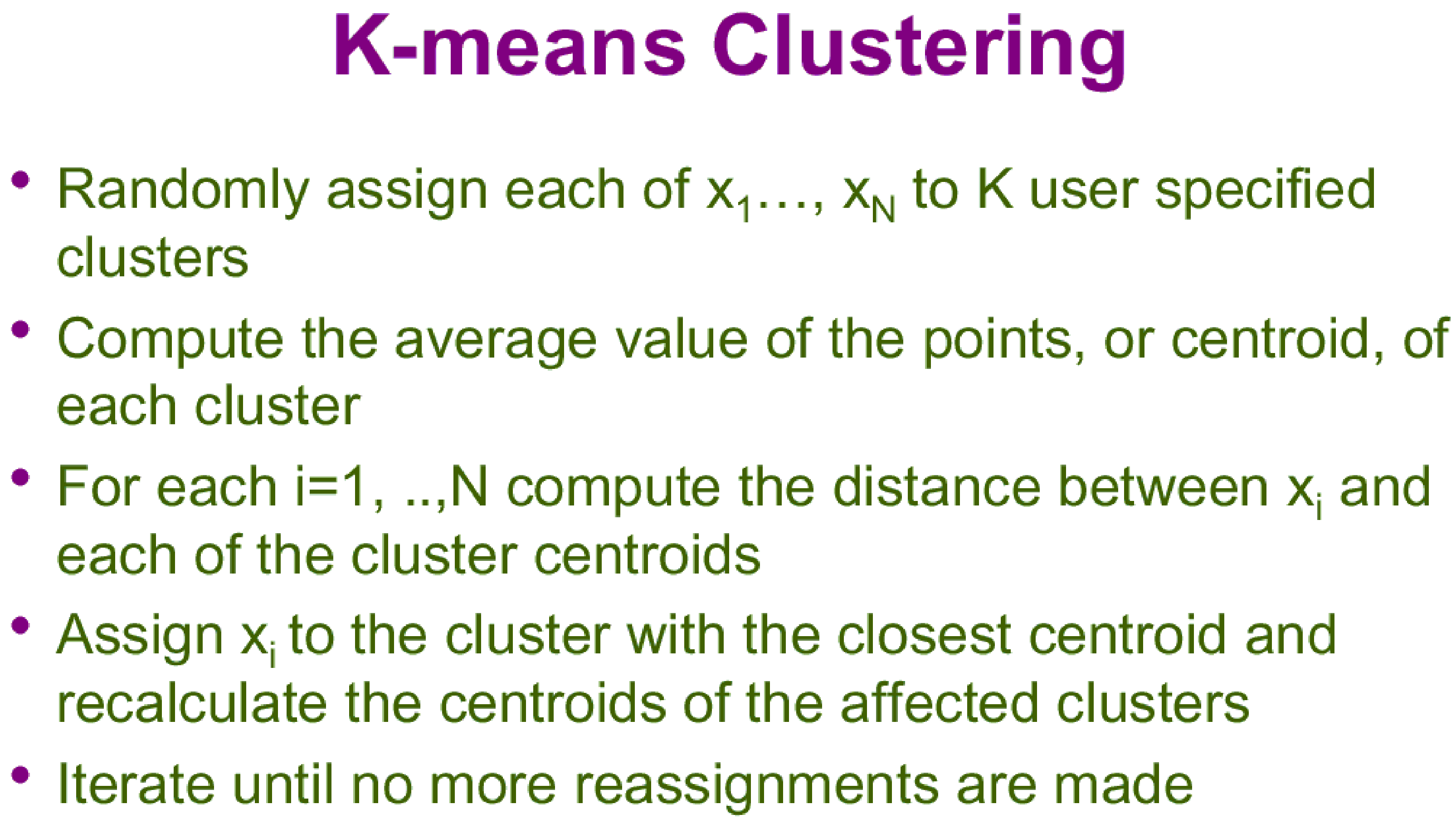
Update the cluster means

reassign

Update the cluster means

reassign

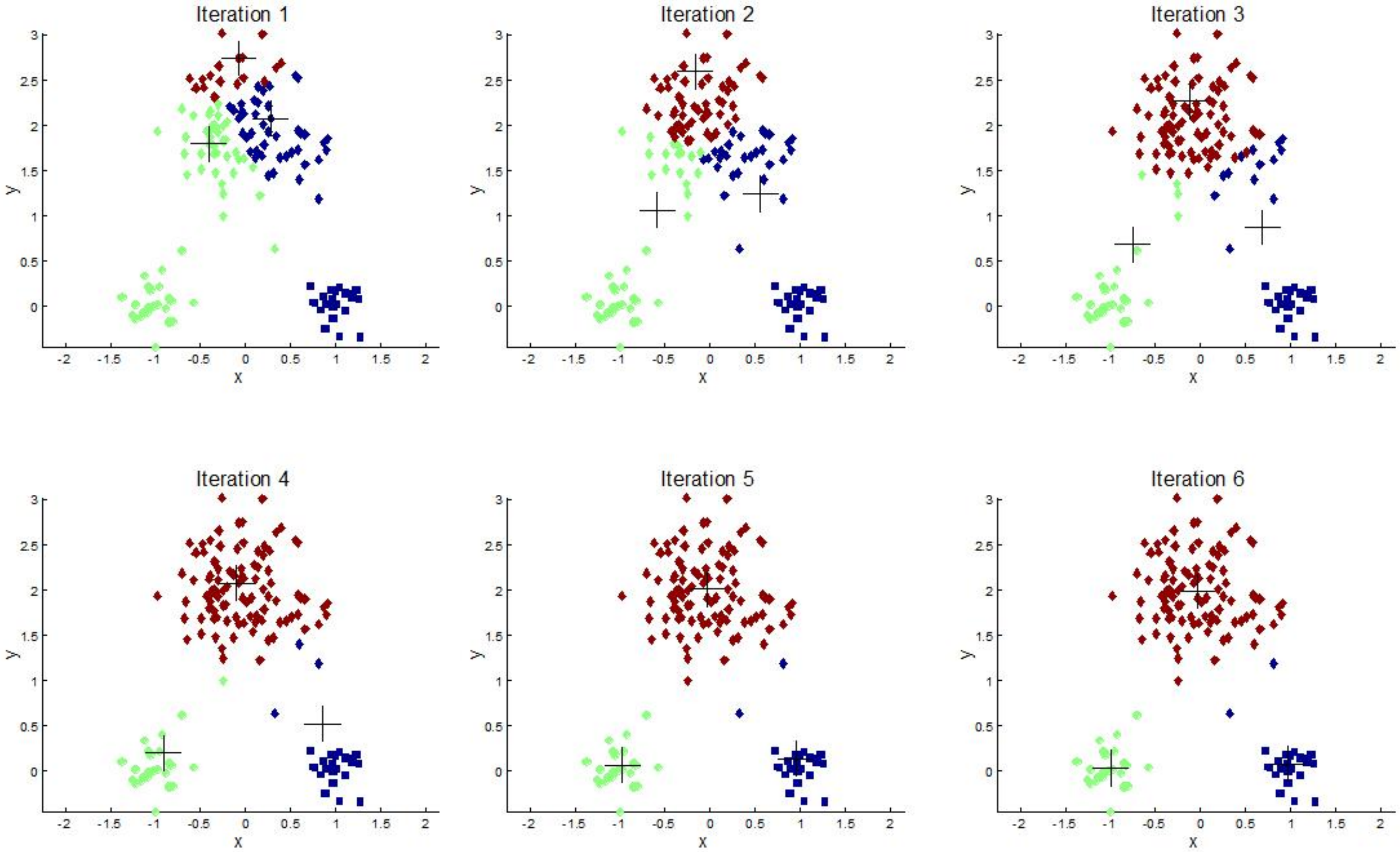
K-means clustering



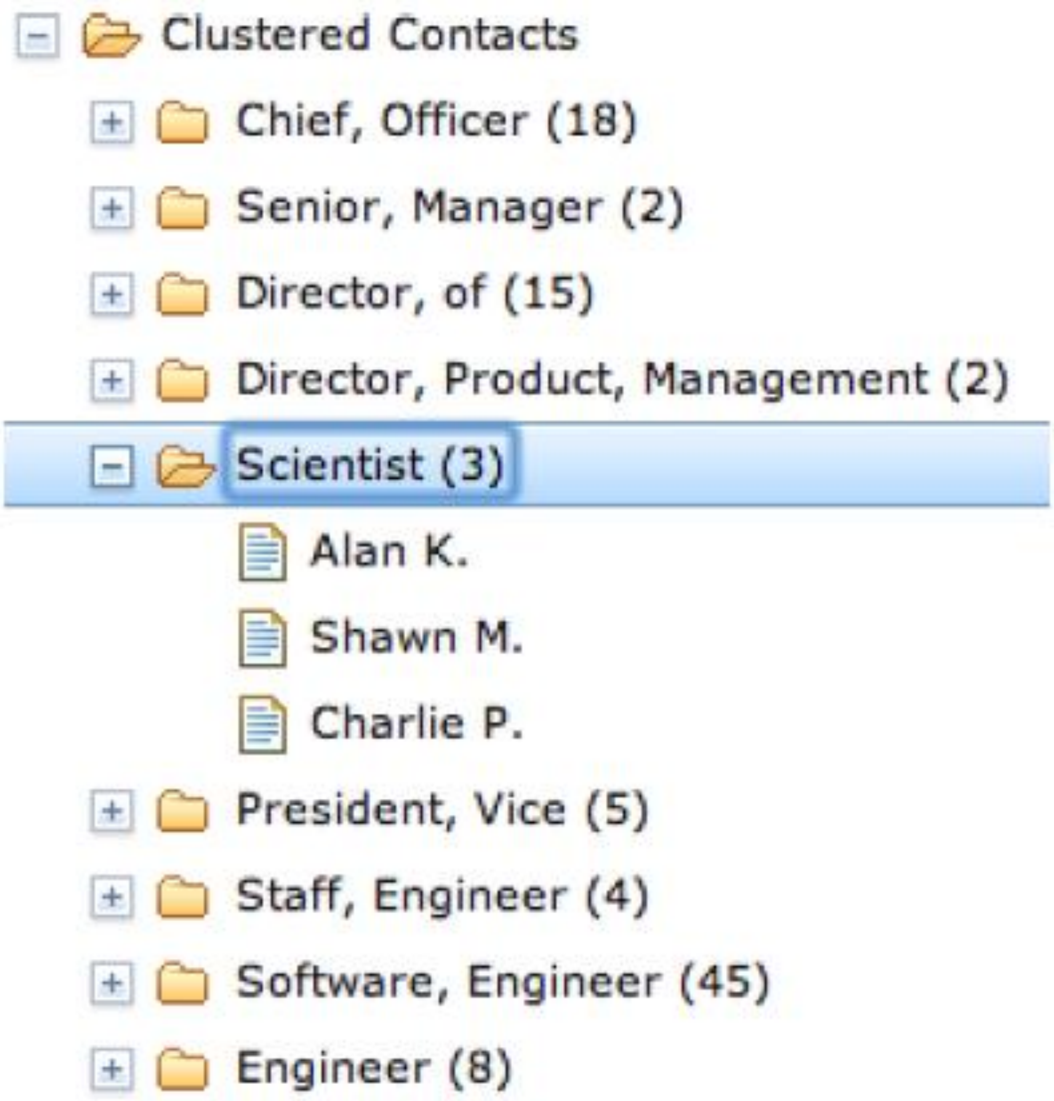
K-means Clustering

- Randomly assign each of $x_1, \ldots, x_N$ to K user specified clusters
- Compute the average value of the points, or centroid, of each cluster
- For each i=1, ..,N compute the distance between x_i and each of the cluster centroids
- Assign x_i to the cluster with the closest centroid and recalculate the centroids of the affected clusters
- Iterate until no more reassignments are made

K-means clustering

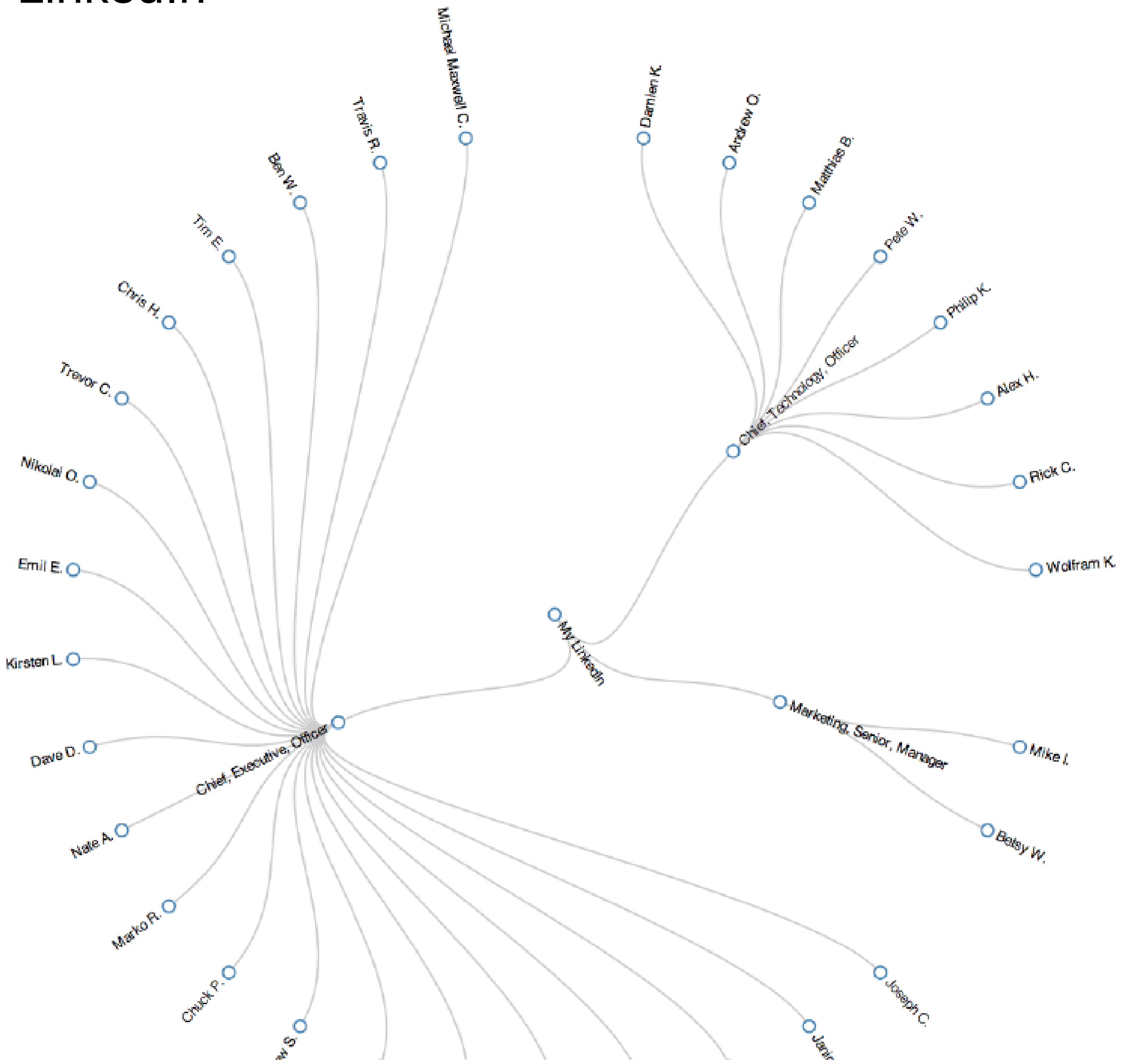


Clustering: LinkedIn



- Clustered Contacts
  - Chief, Officer (18)
  - Senior, Manager (2)
  - Director, of (15)
  - Director, Product, Management (2)
  - Scientist (3)
    - Alan K.
    - Shawn M.
    - Charlie P.
  - President, Vice (5)
  - Staff, Engineer (4)
  - Software, Engineer (45)
  - Engineer (8)

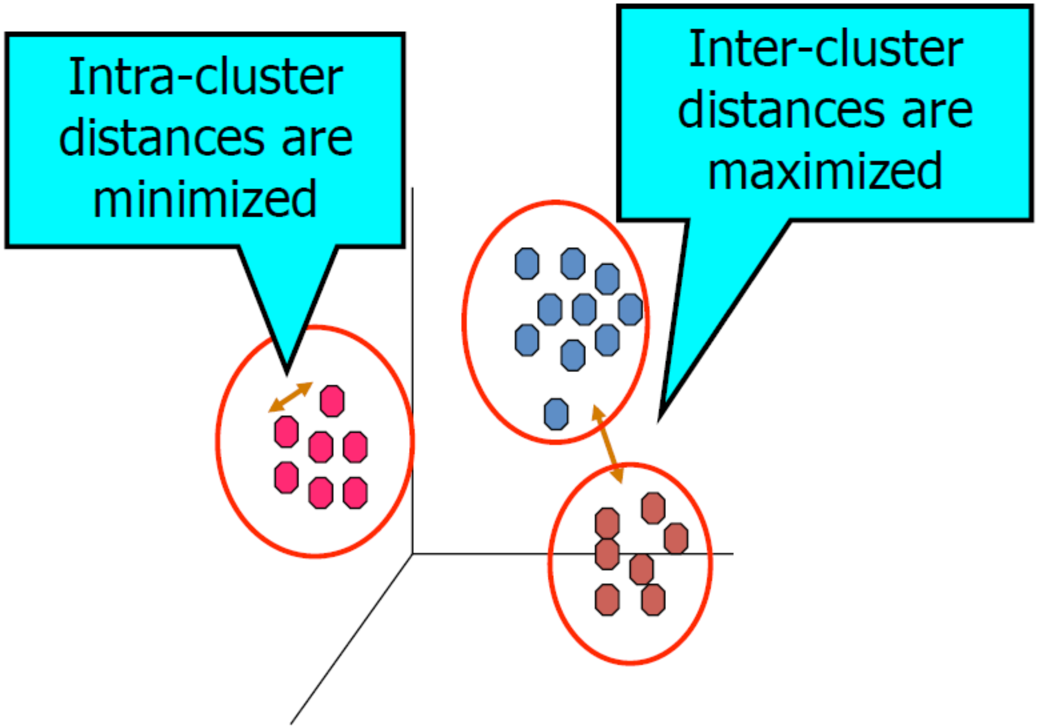
Clustering: LinkedIn



Cluster analysis - K-means clustering

- K-Means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure

- Examples within a cluster are very similar
- Examples across different clusters are very different

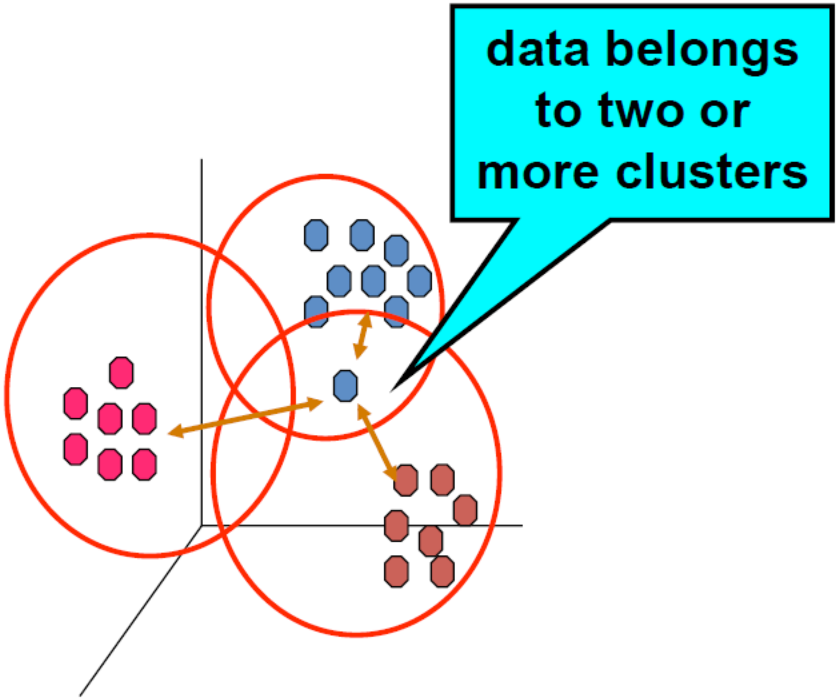

Intra-cluster distances are minimized
Inter-cluster distances are maximized

Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis - Fuzzy C-means clustering (FCM)

- Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters.

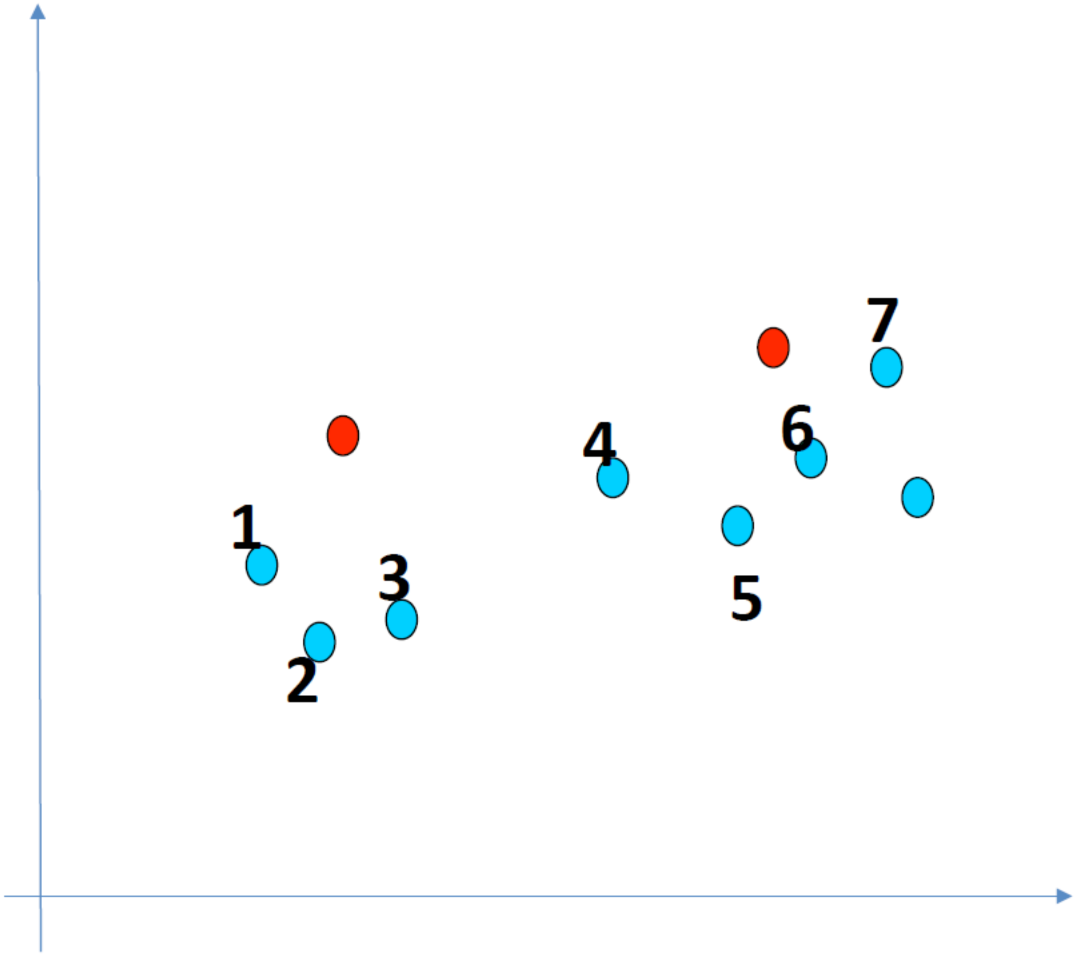
- Always converges
- Clustering noisy data samples



data belongs to two or more clusters

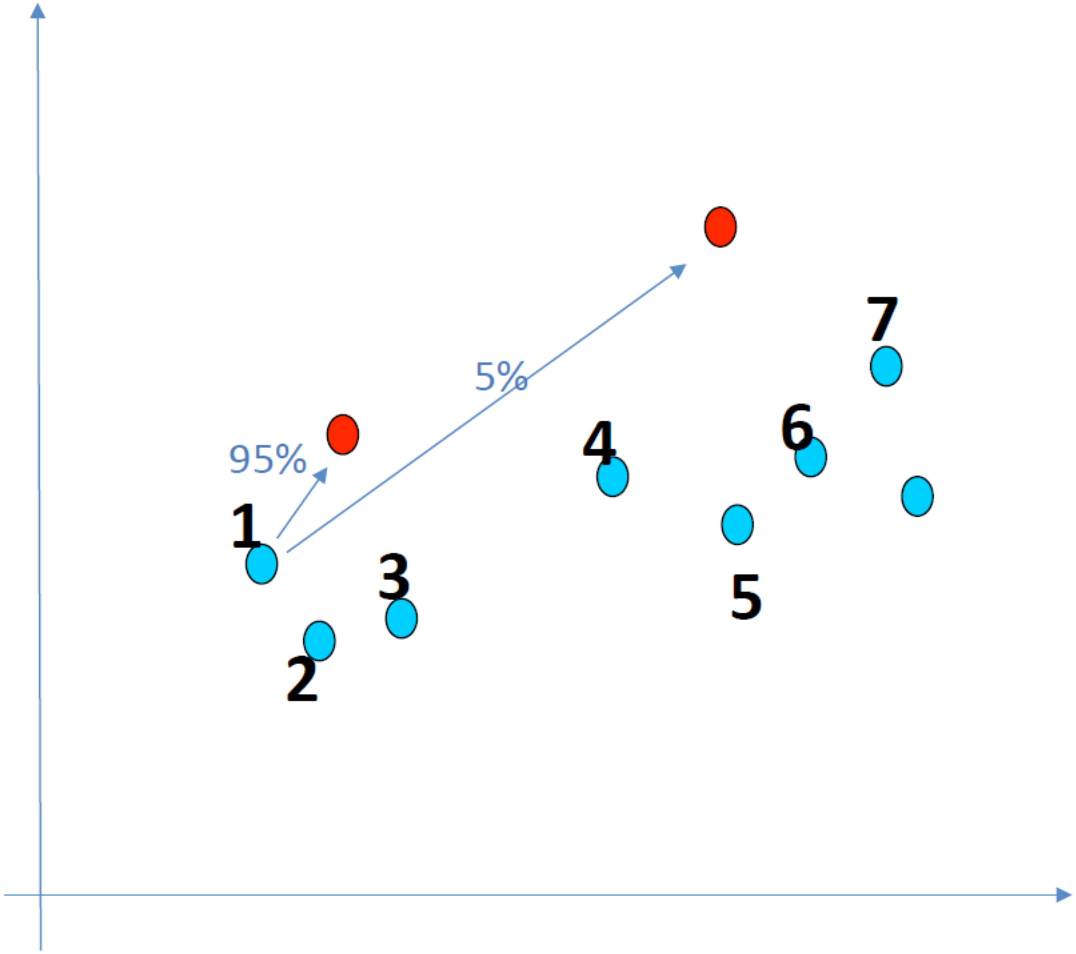
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis - Fuzzy C-means clustering (FCM)



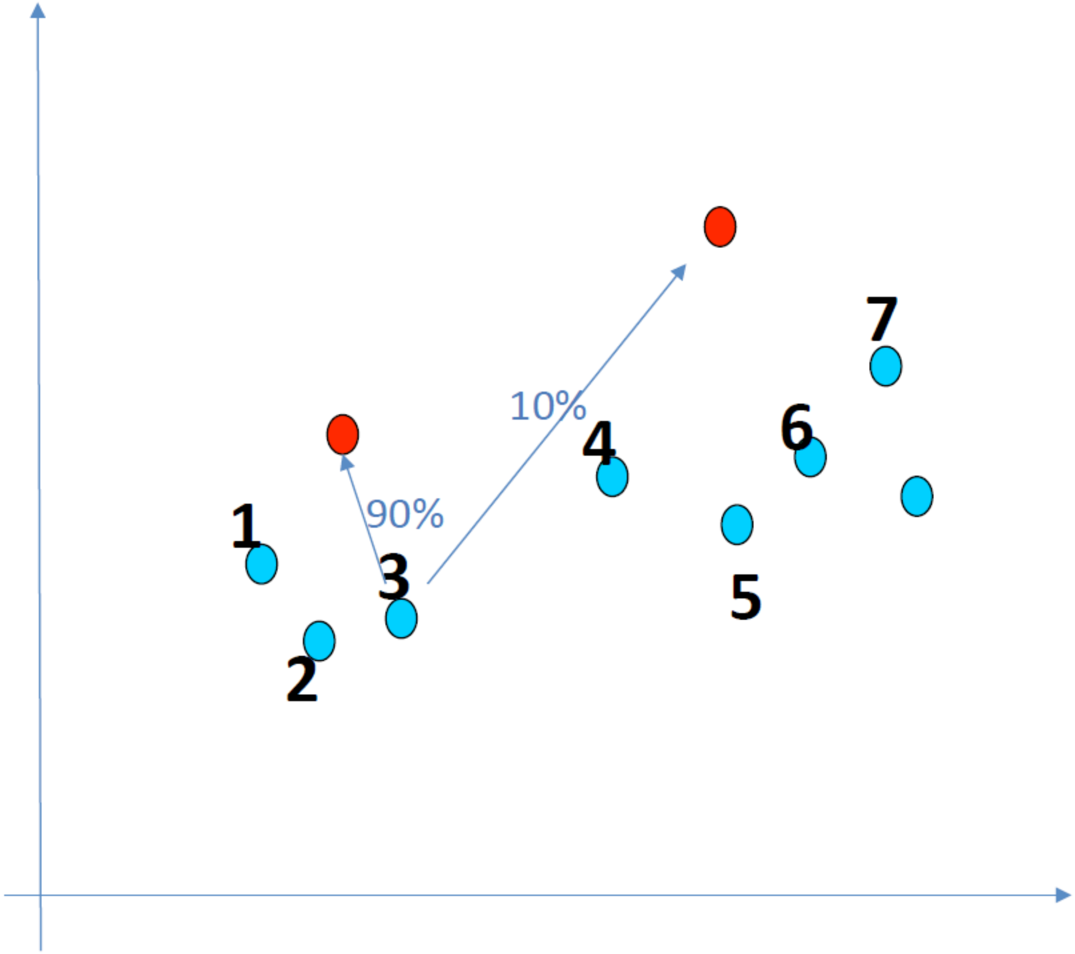
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis - Fuzzy C-means clustering (FCM)



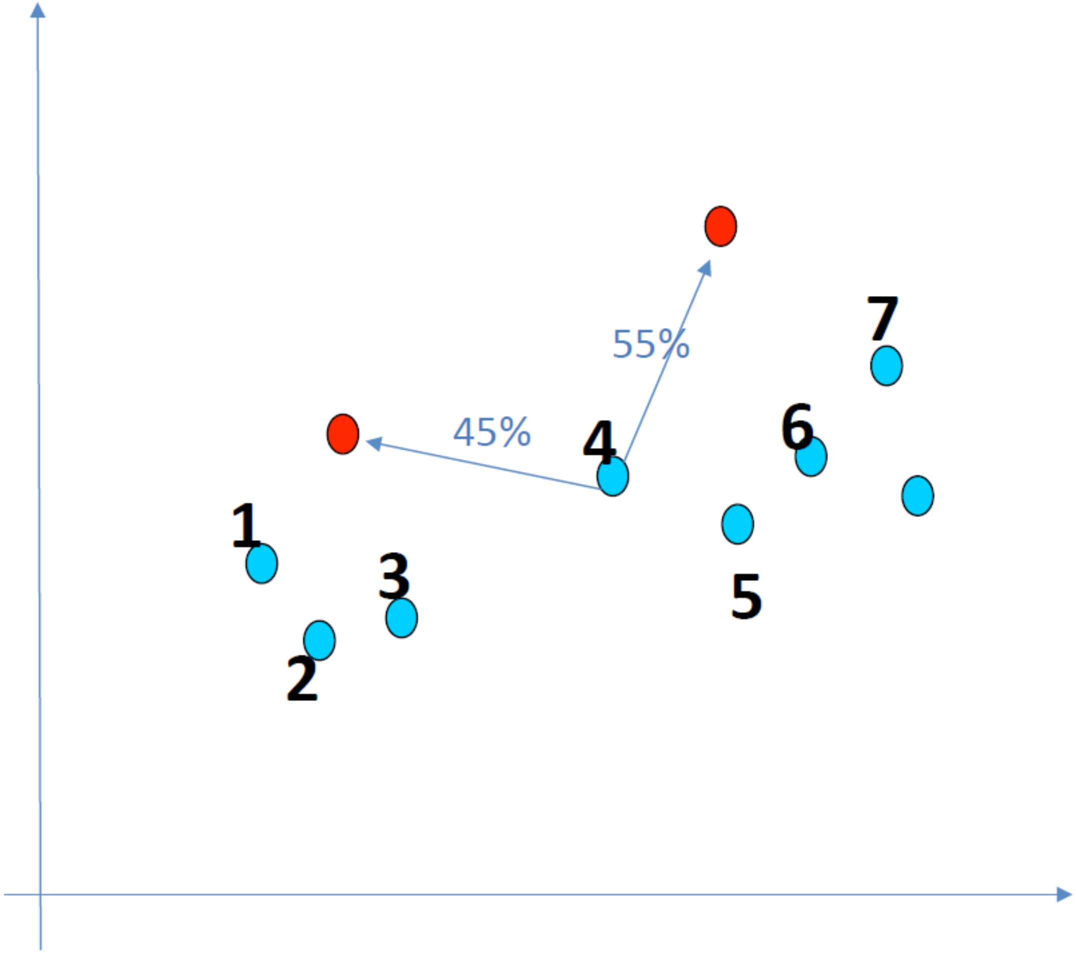
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis - Fuzzy C-means clustering (FCM)



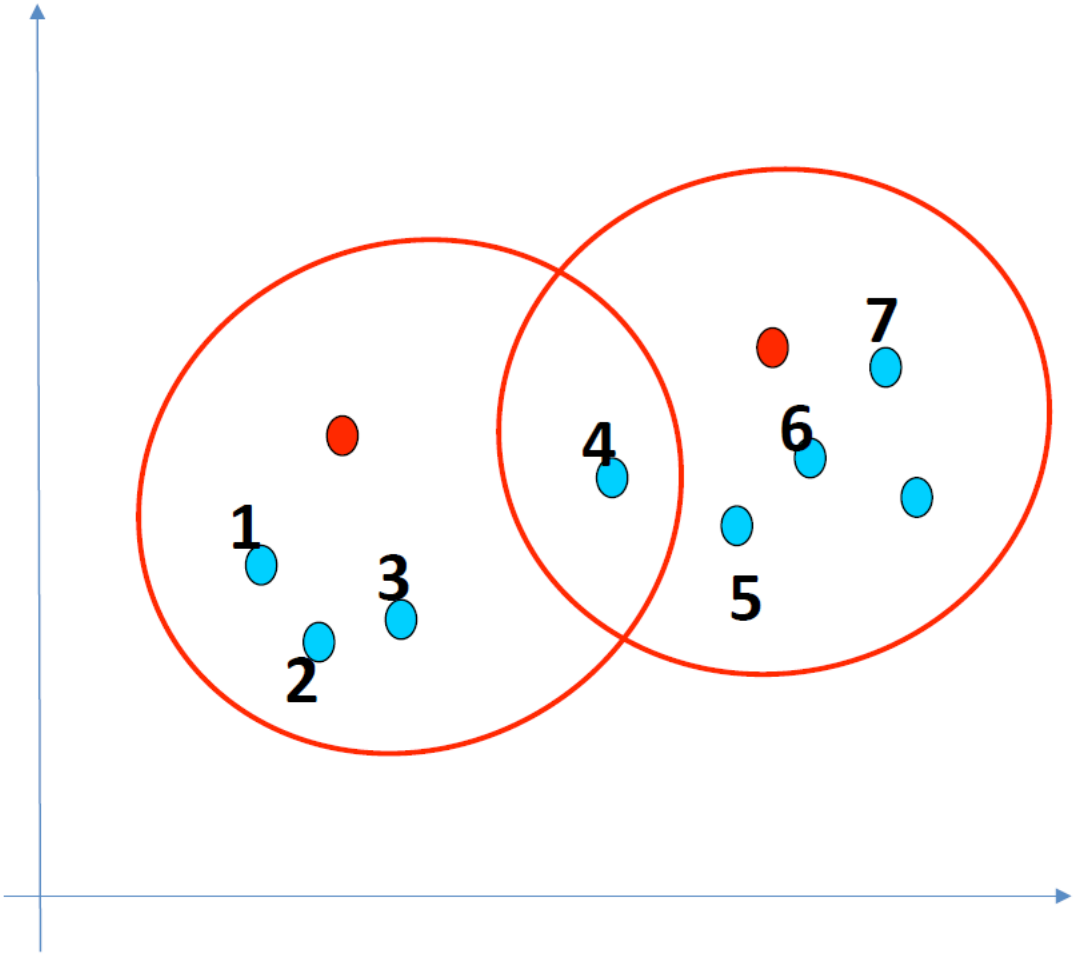
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis - Fuzzy C-means clustering (FCM)



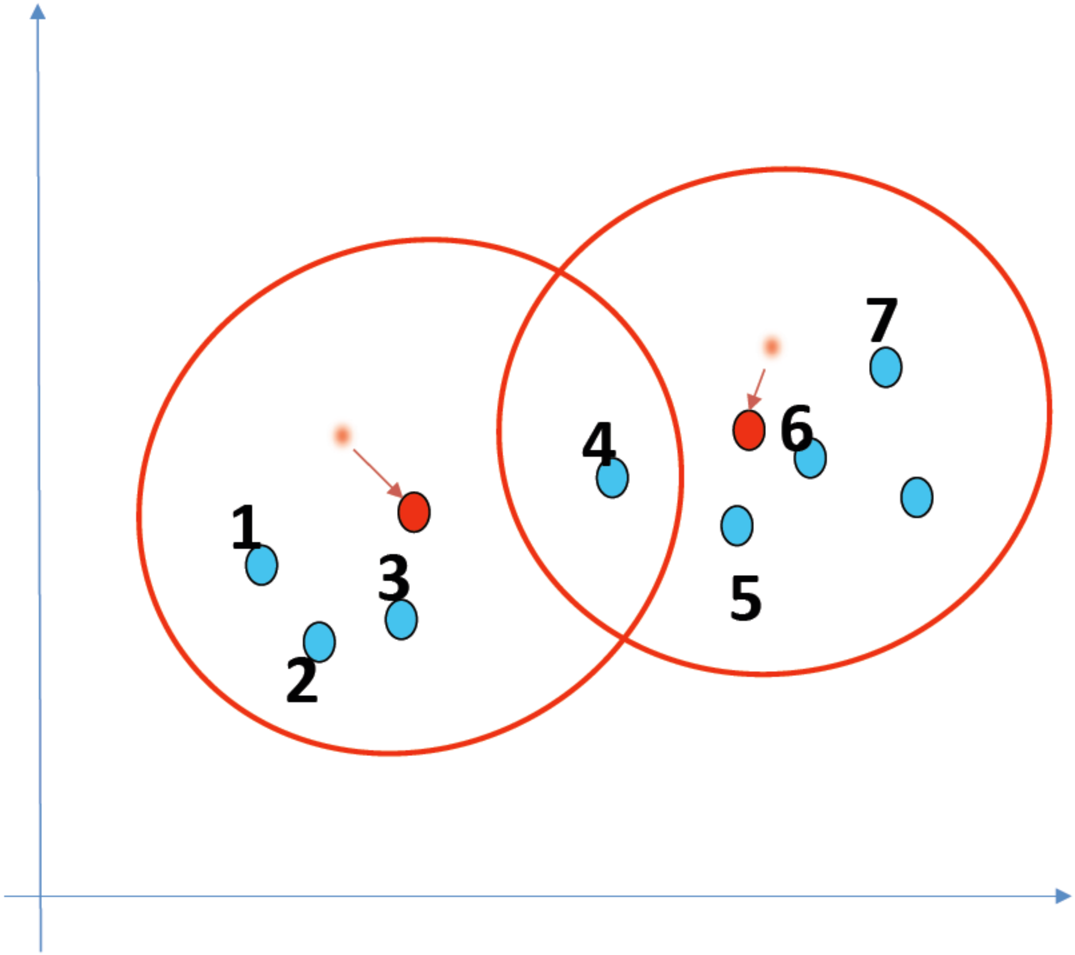
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis - Fuzzy C-means clustering (FCM)



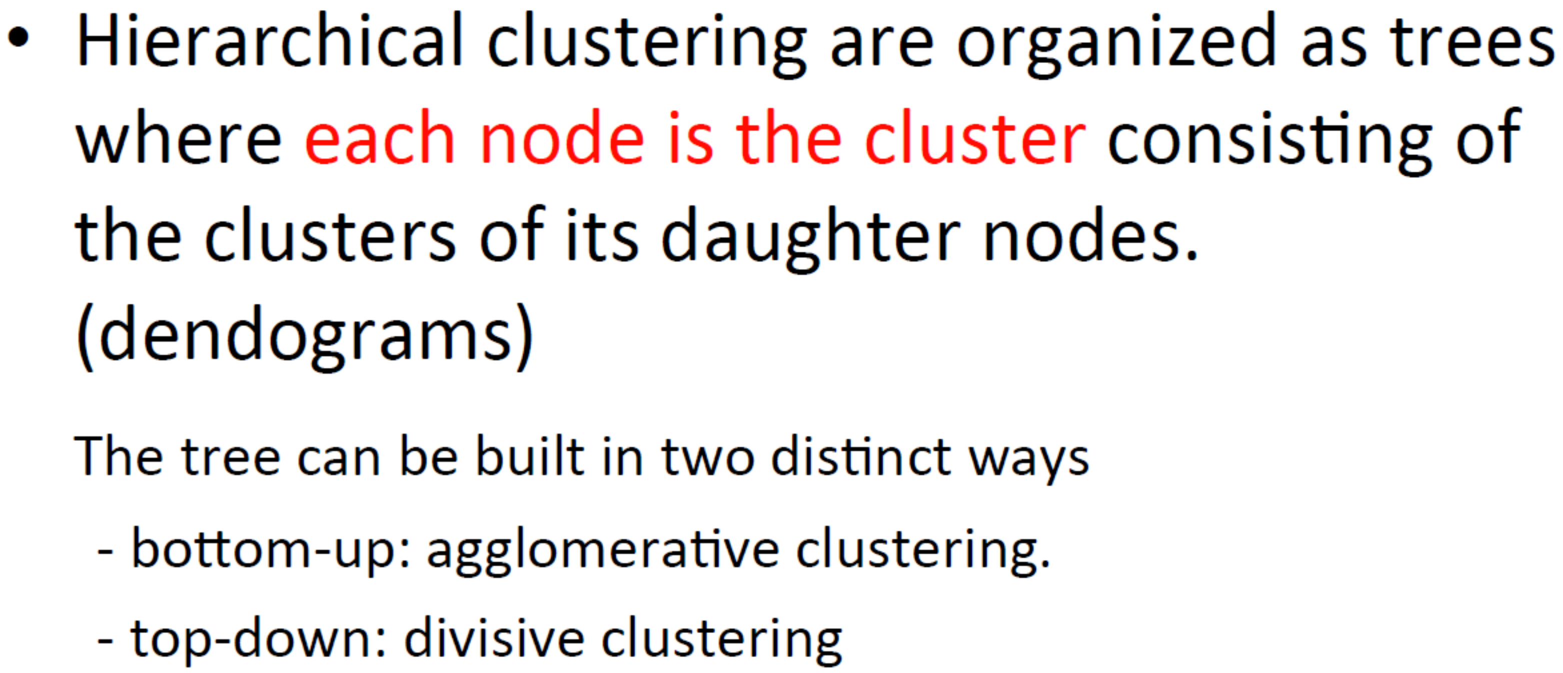
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis - Fuzzy C-means clustering (FCM)



Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – Hierarchical clustering



- Hierarchical clustering are organized as trees where each node is the cluster consisting of the clusters of its daughter nodes. (dendograms)

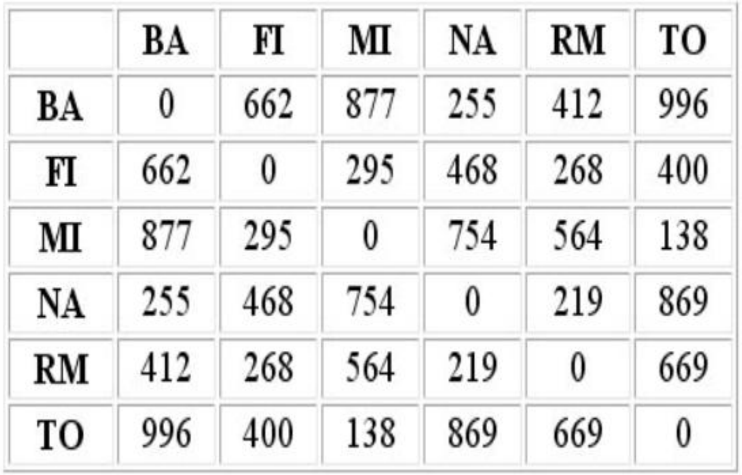
The tree can be built in two distinct ways

- bottom-up: agglomerative clustering.

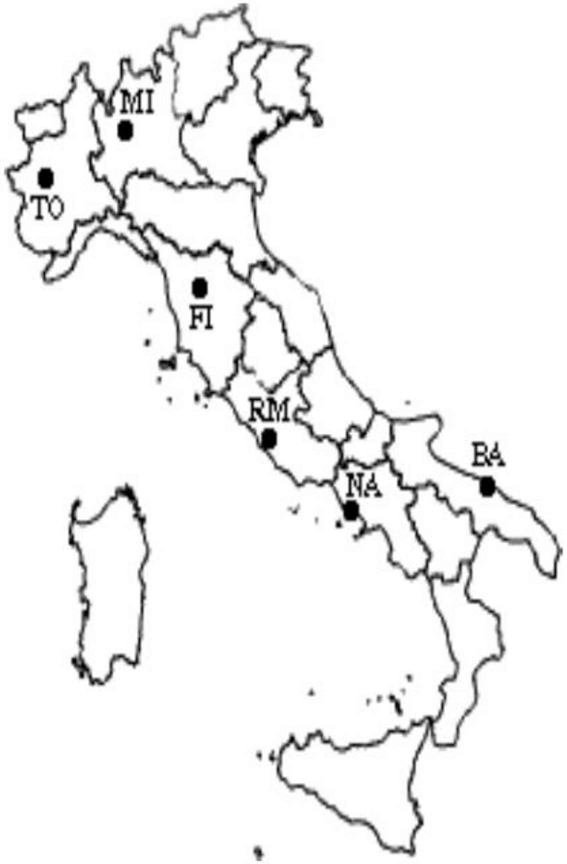
- top-down: divisive clustering

Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – Hierarchical clustering



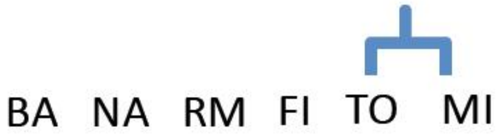
	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

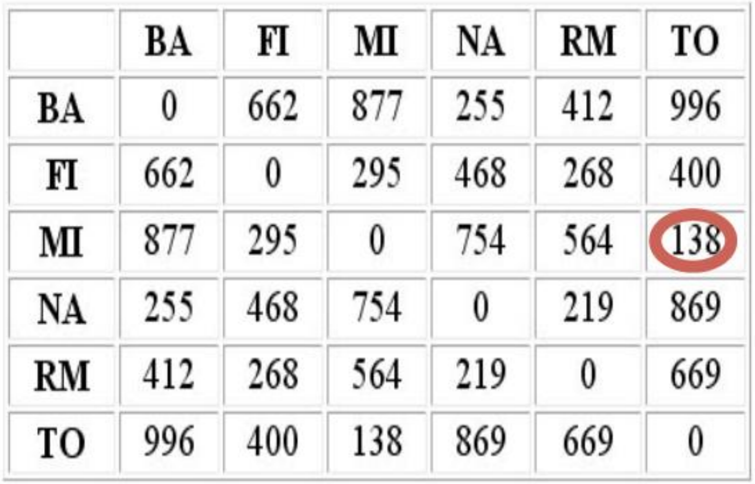


DATA MINING WITH CLUSTERING AND CLASSIFICATION, Spring 207, SJSU, Benjamin Lam

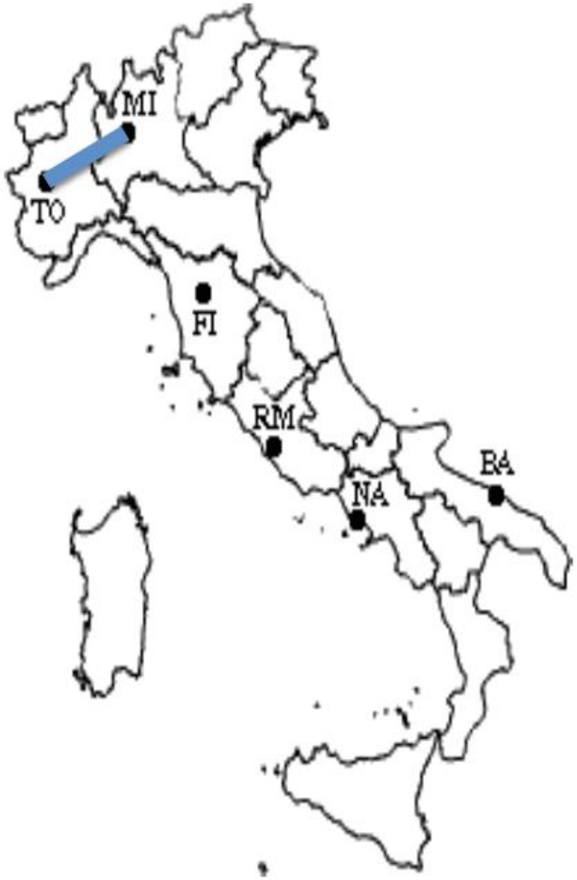
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – Hierarchical clustering



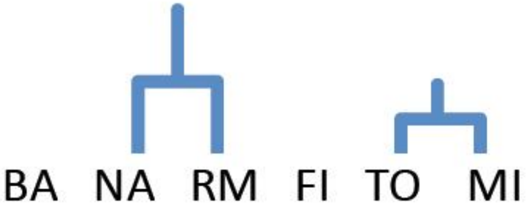


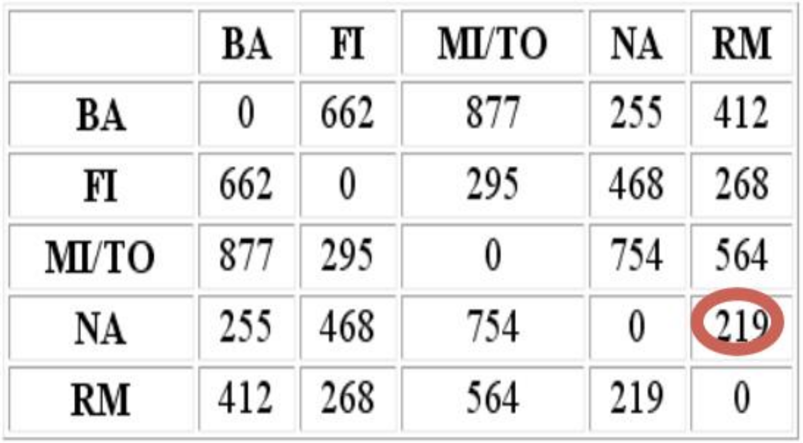
	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0



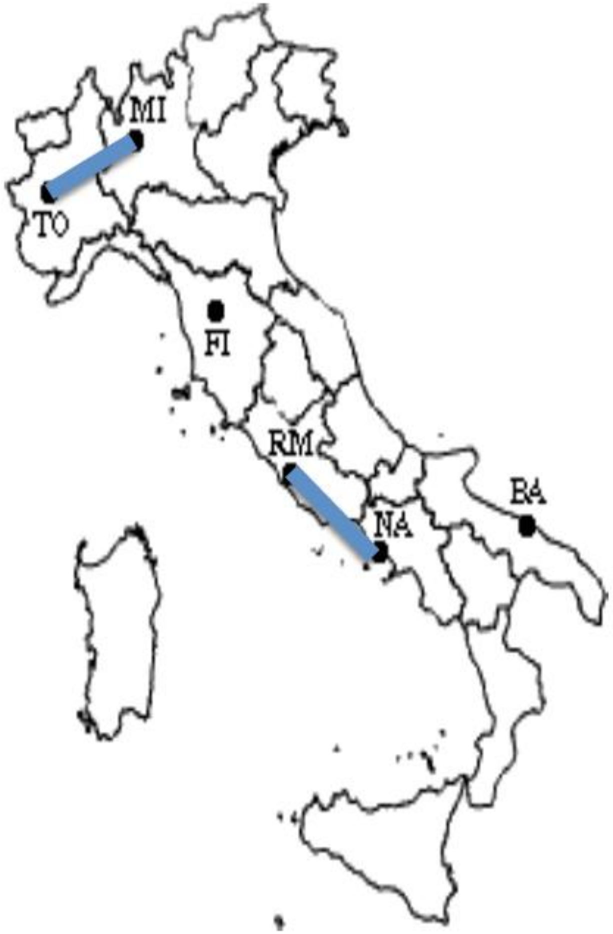
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – Hierarchical clustering



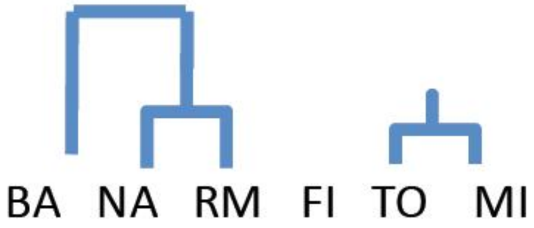


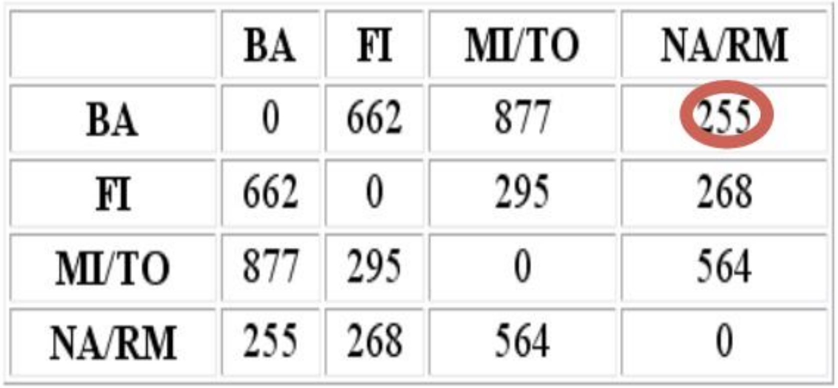
	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0



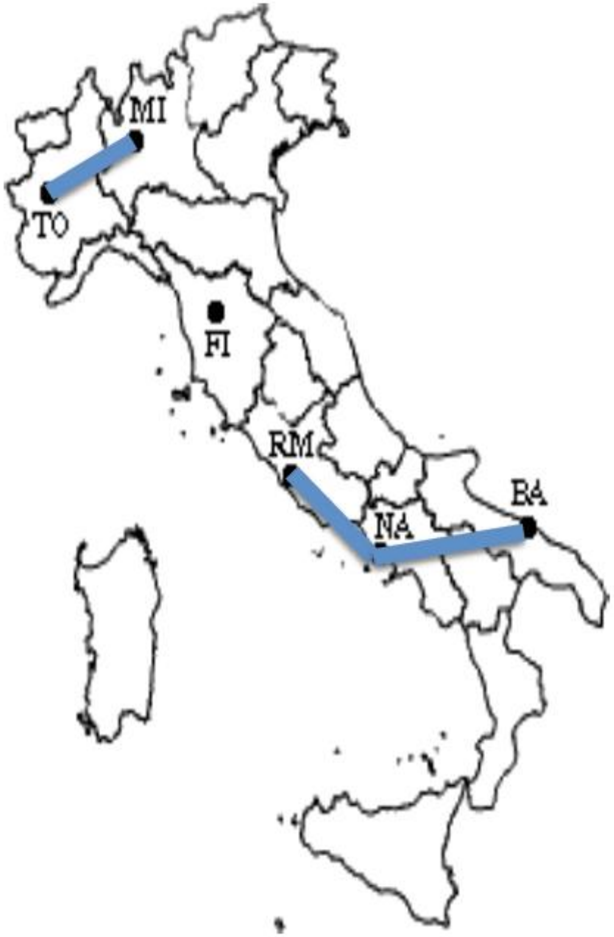
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – Hierarchical clustering



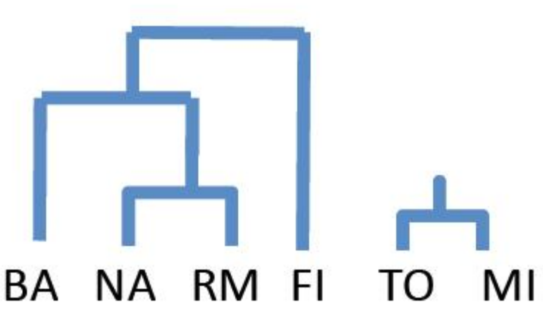


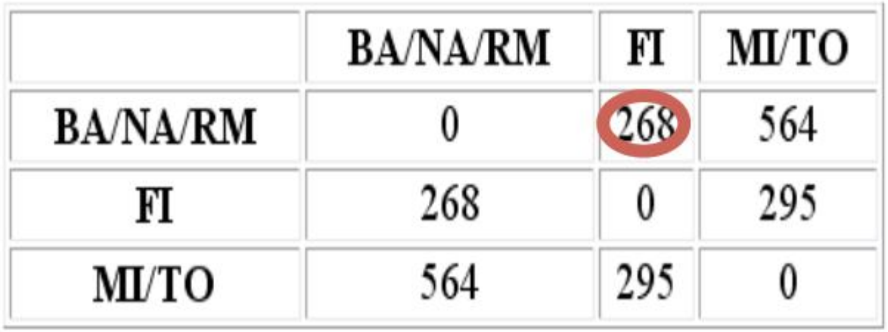
	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0



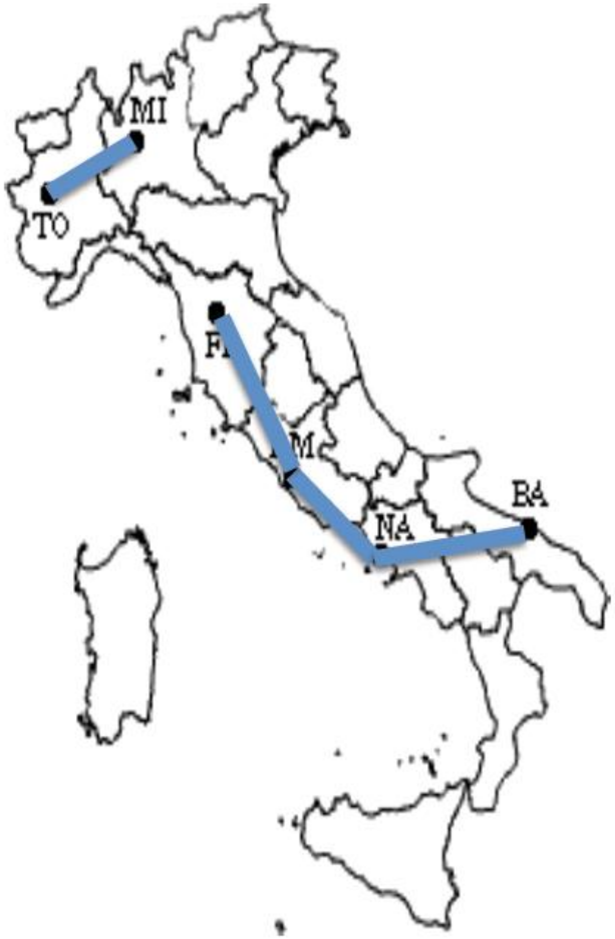
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – Hierarchical clustering



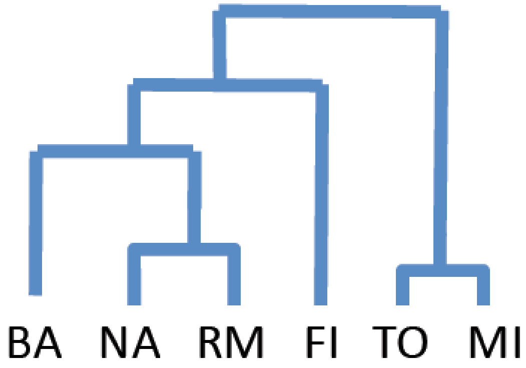


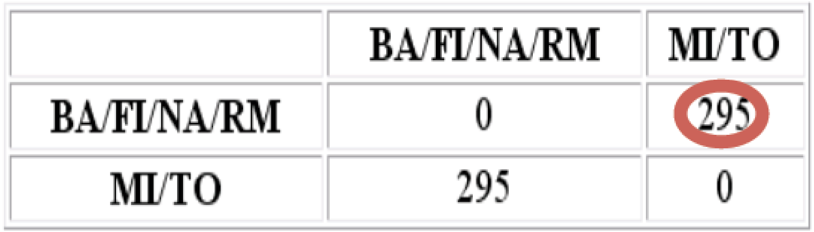
	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0



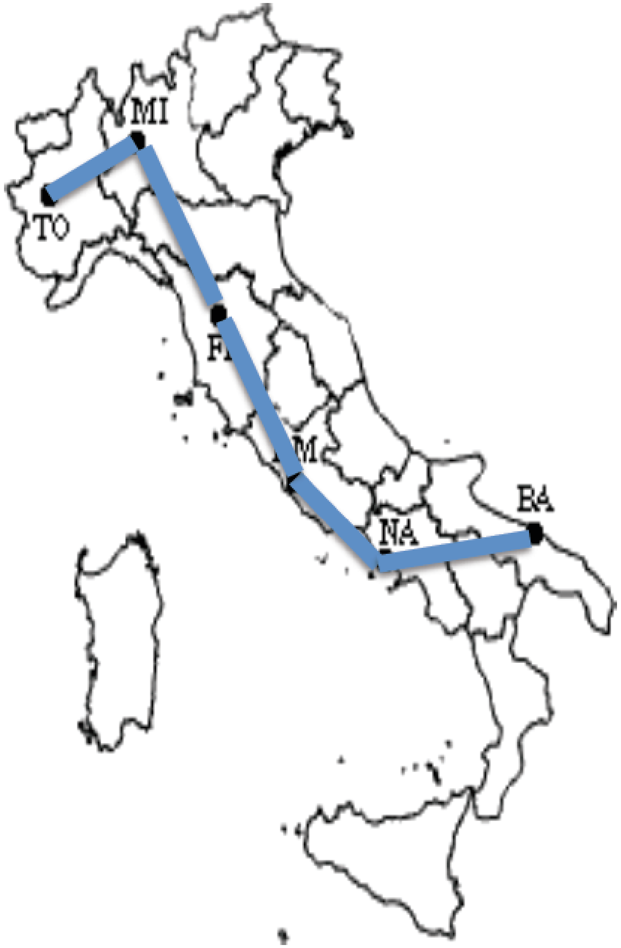
Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – Hierarchical clustering



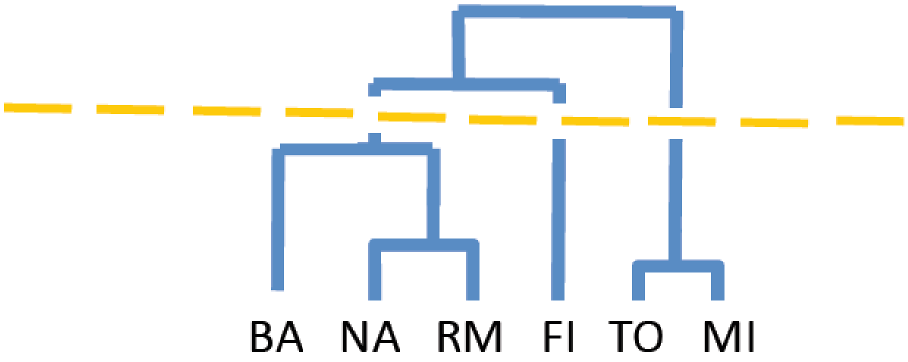


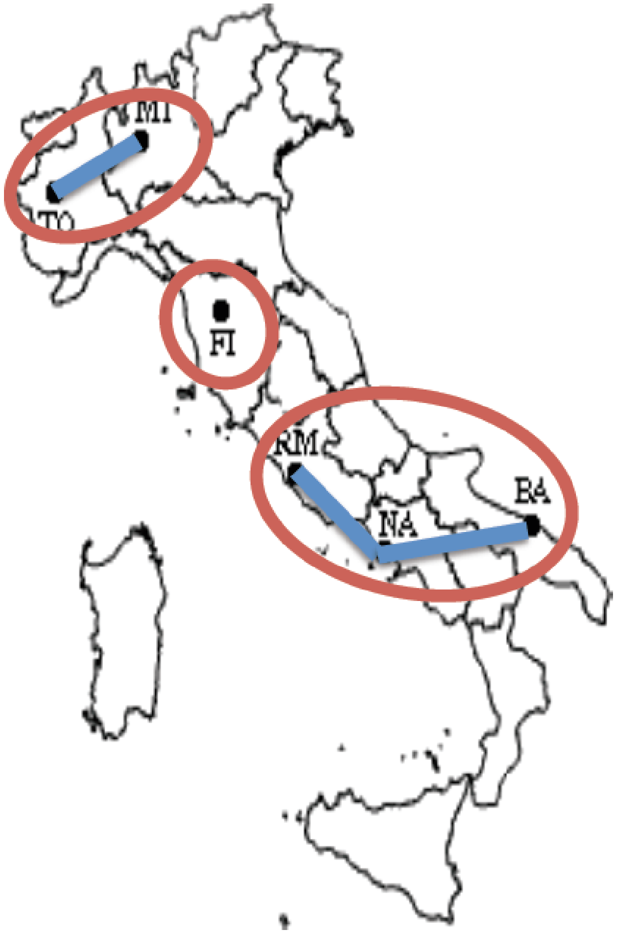
	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0



Source: Saeed Aghabozorgi, Cluster Analysis

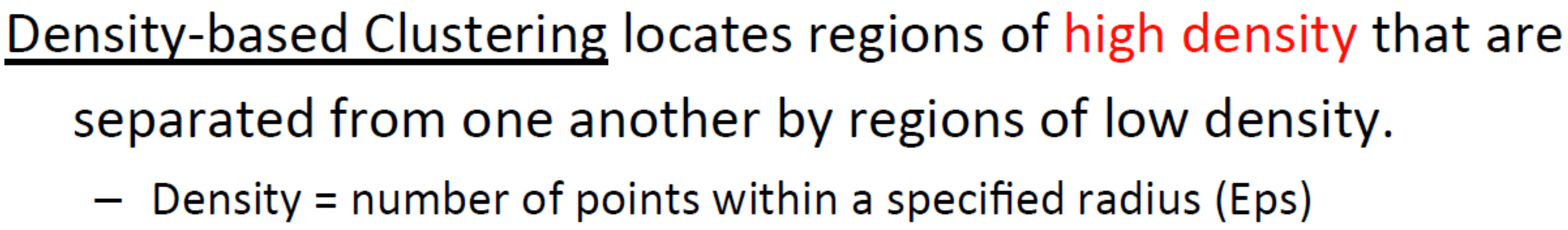
Cluster analysis – Hierarchical clustering





Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – DBSCAN

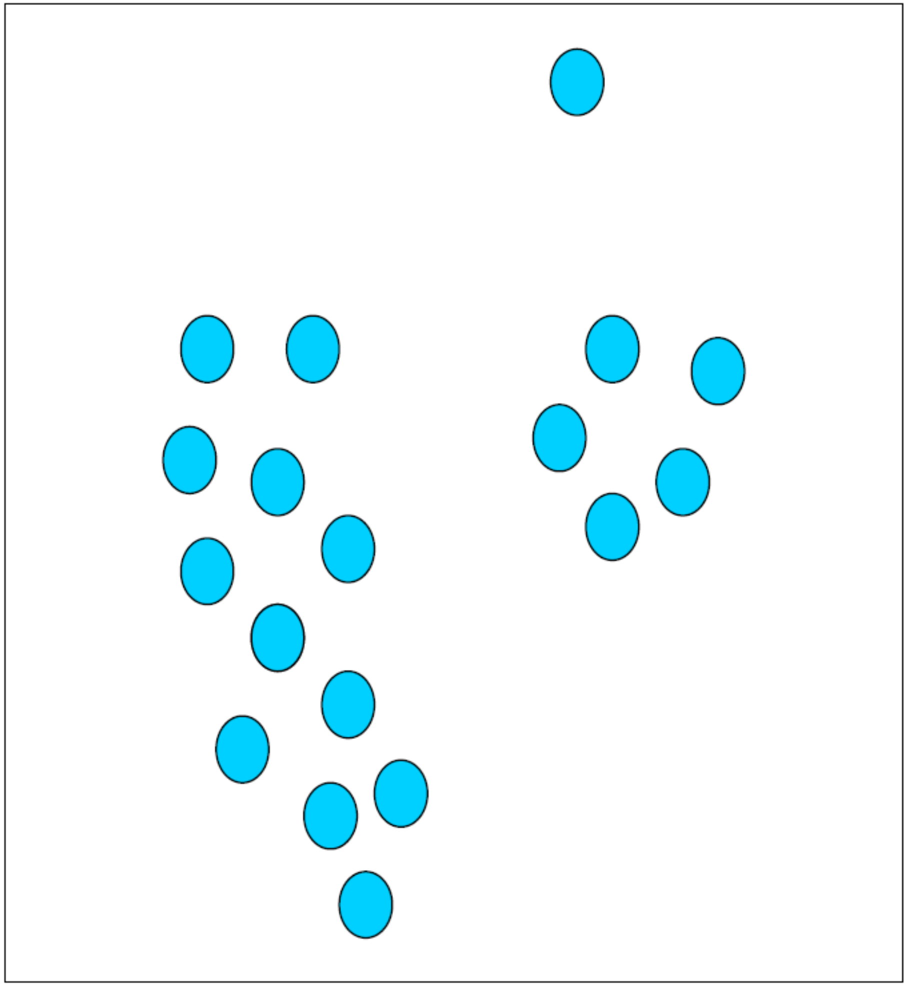


Density-based Clustering locates regions of high density that are separated from one another by regions of low density.

- Density = number of points within a specified radius (Eps)

Source: Saeed Aghabozorgi, Cluster Analysis

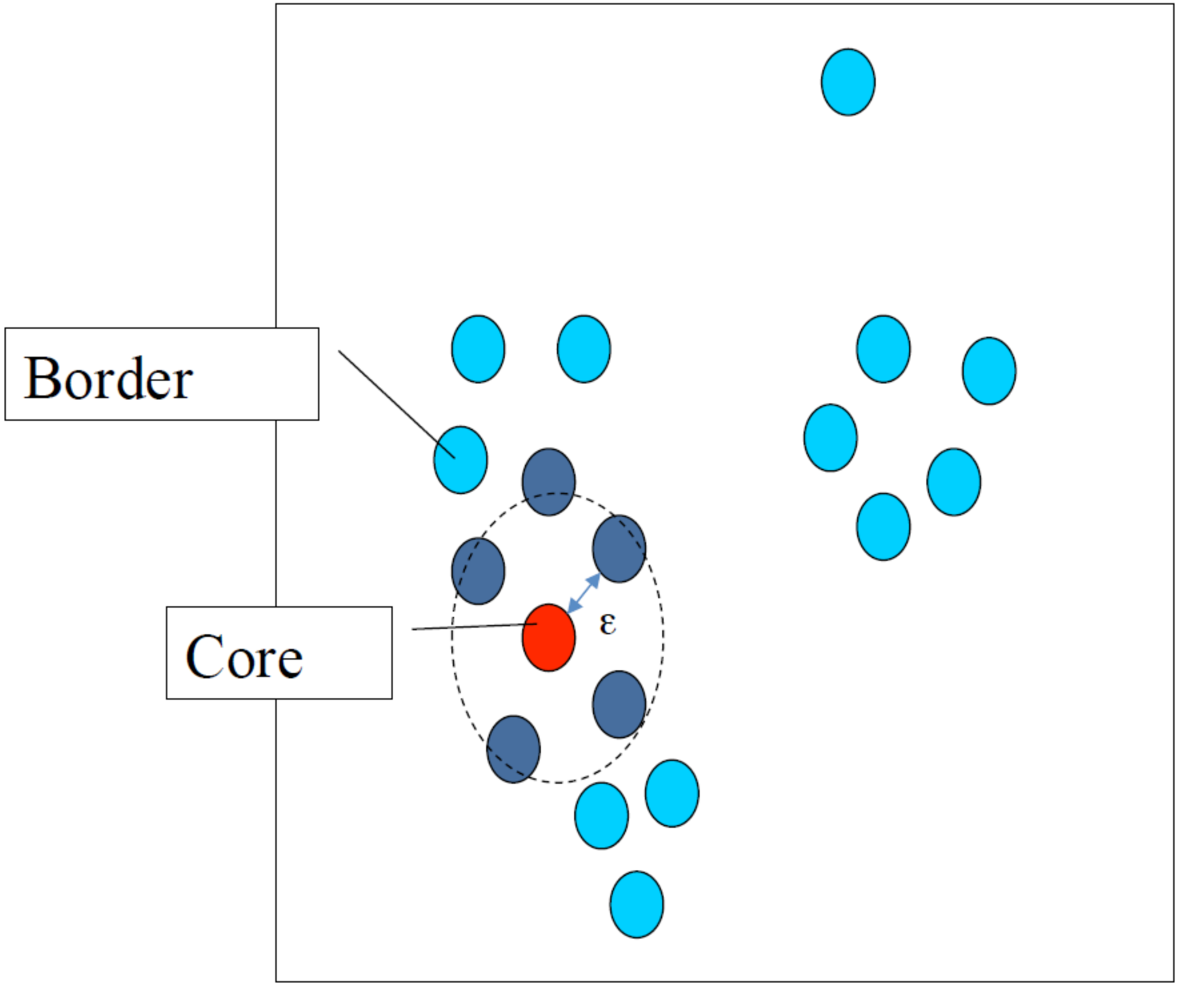
Cluster analysis – DBSCAN



ε = 1unit, MinPts = 5

Source: Saeed Aghabozorgi, Cluster Analysis

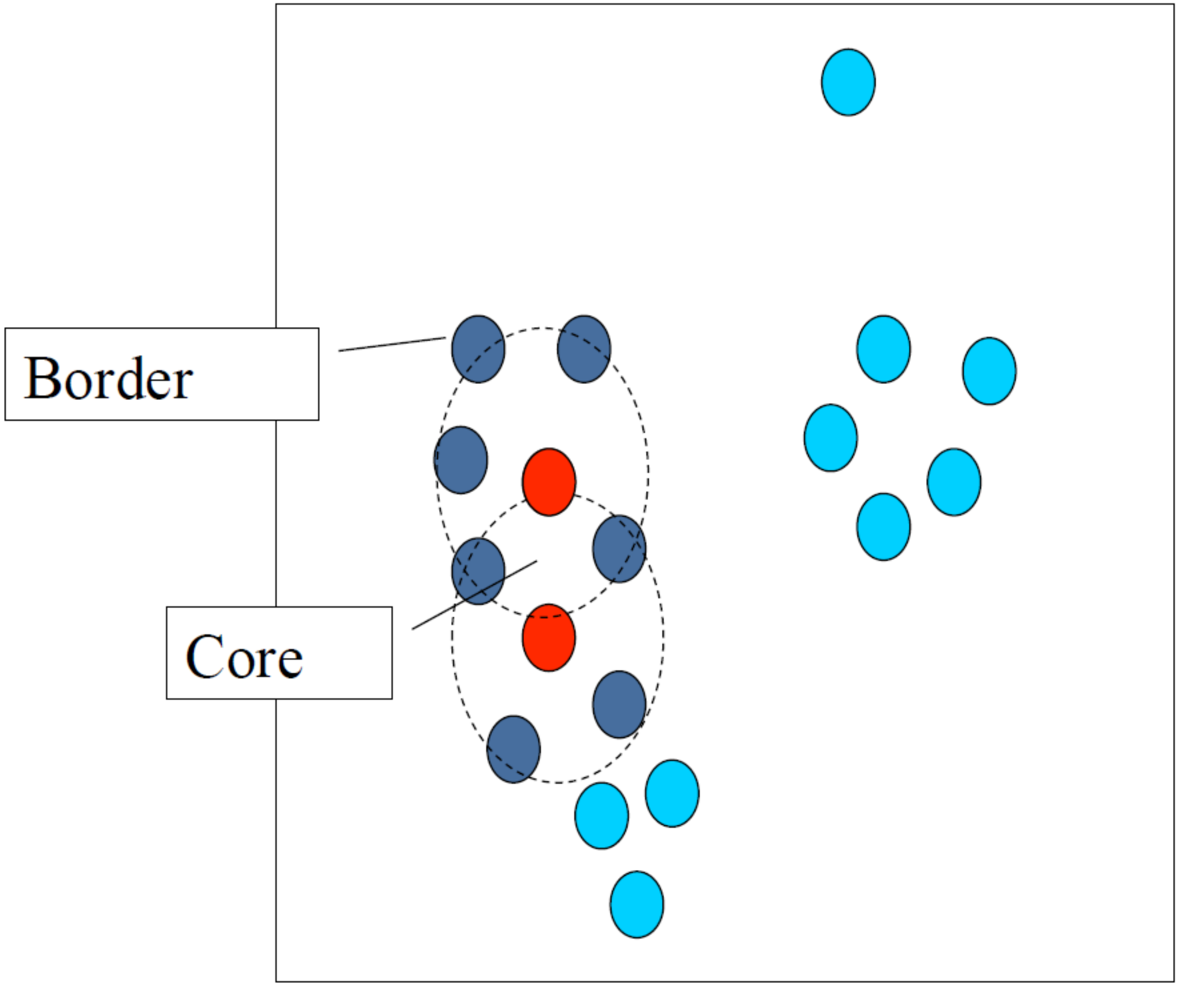
Cluster analysis – DBSCAN



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

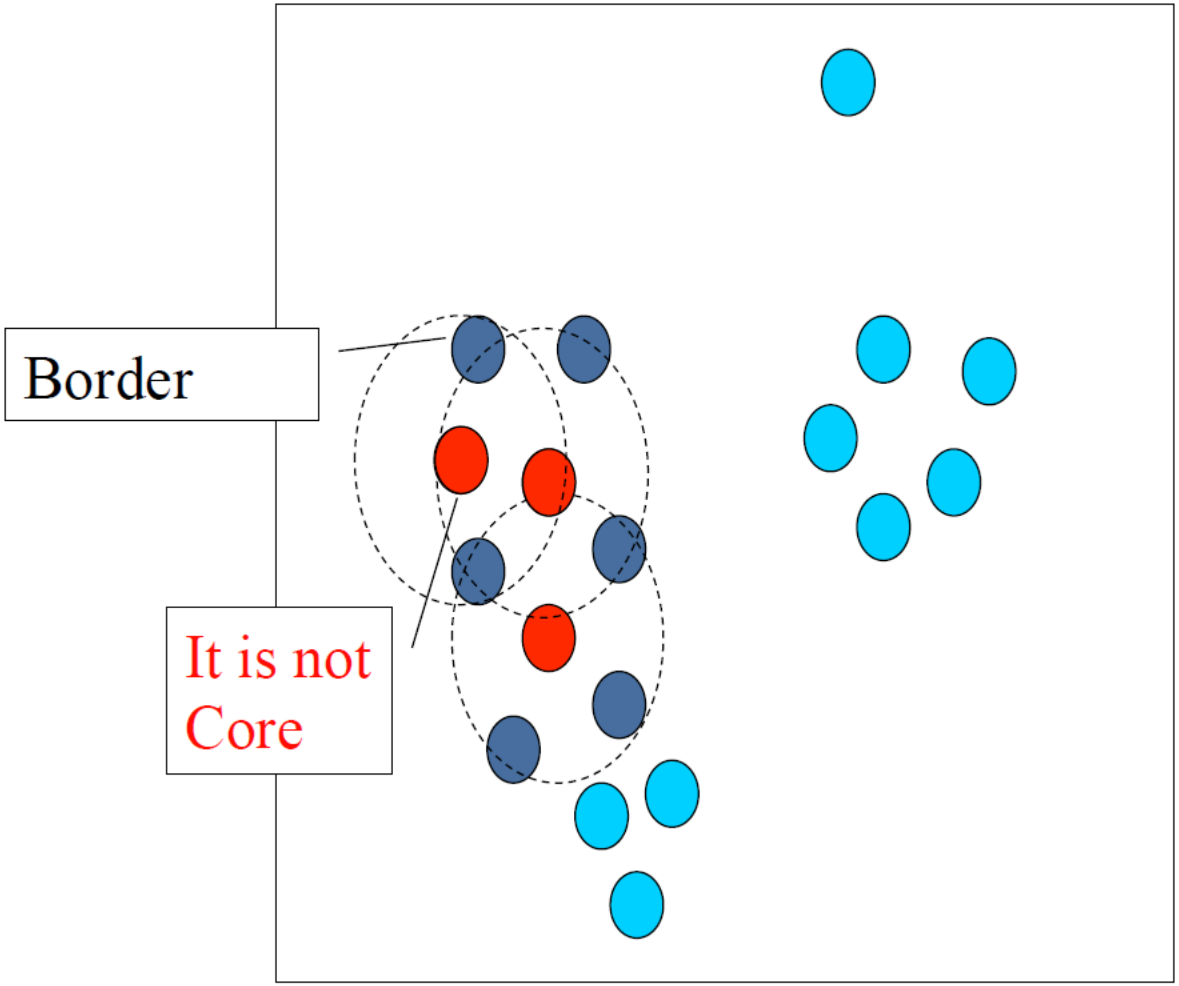
Cluster analysis – DBSCAN



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

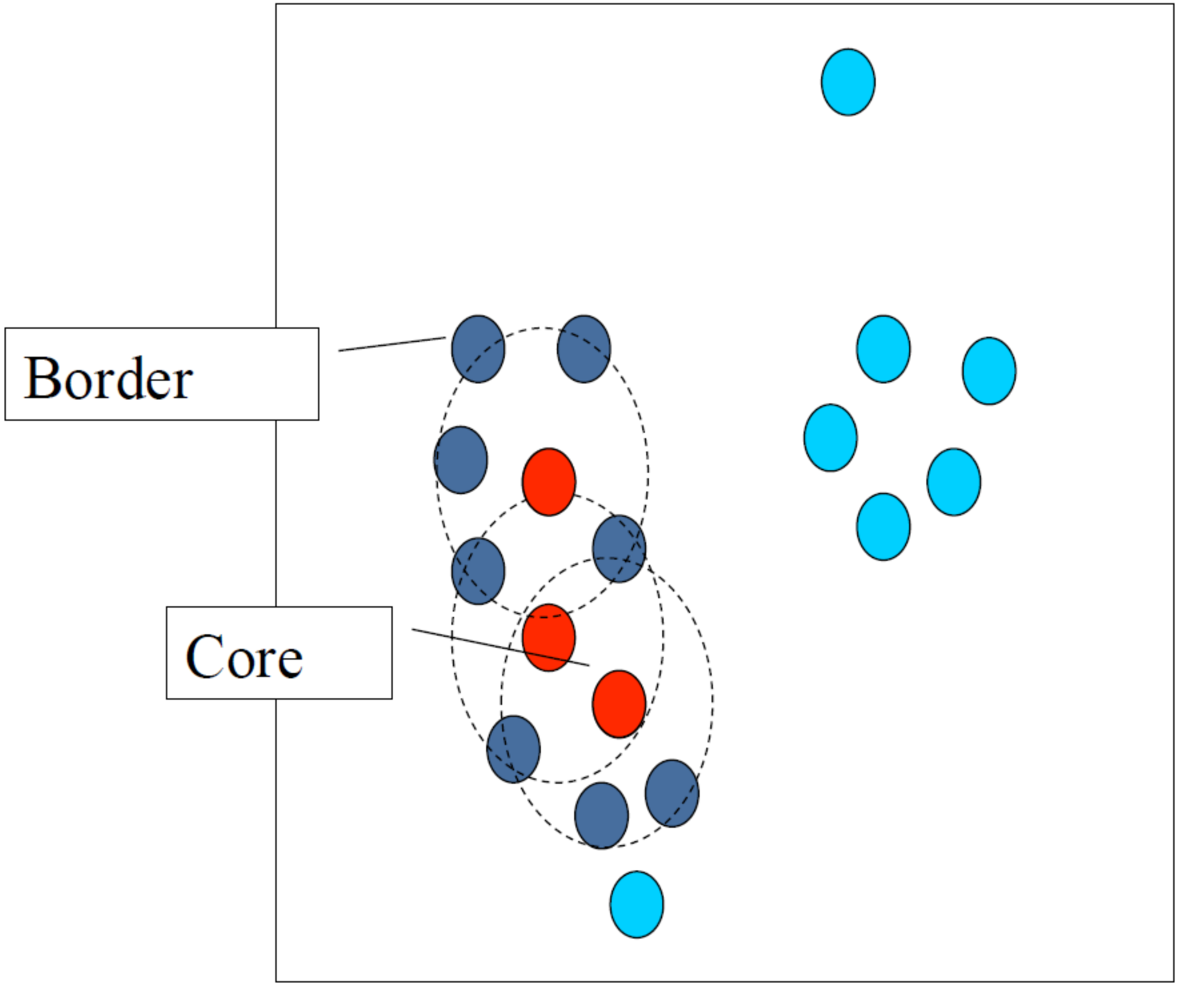
Cluster analysis – DBSCAN



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

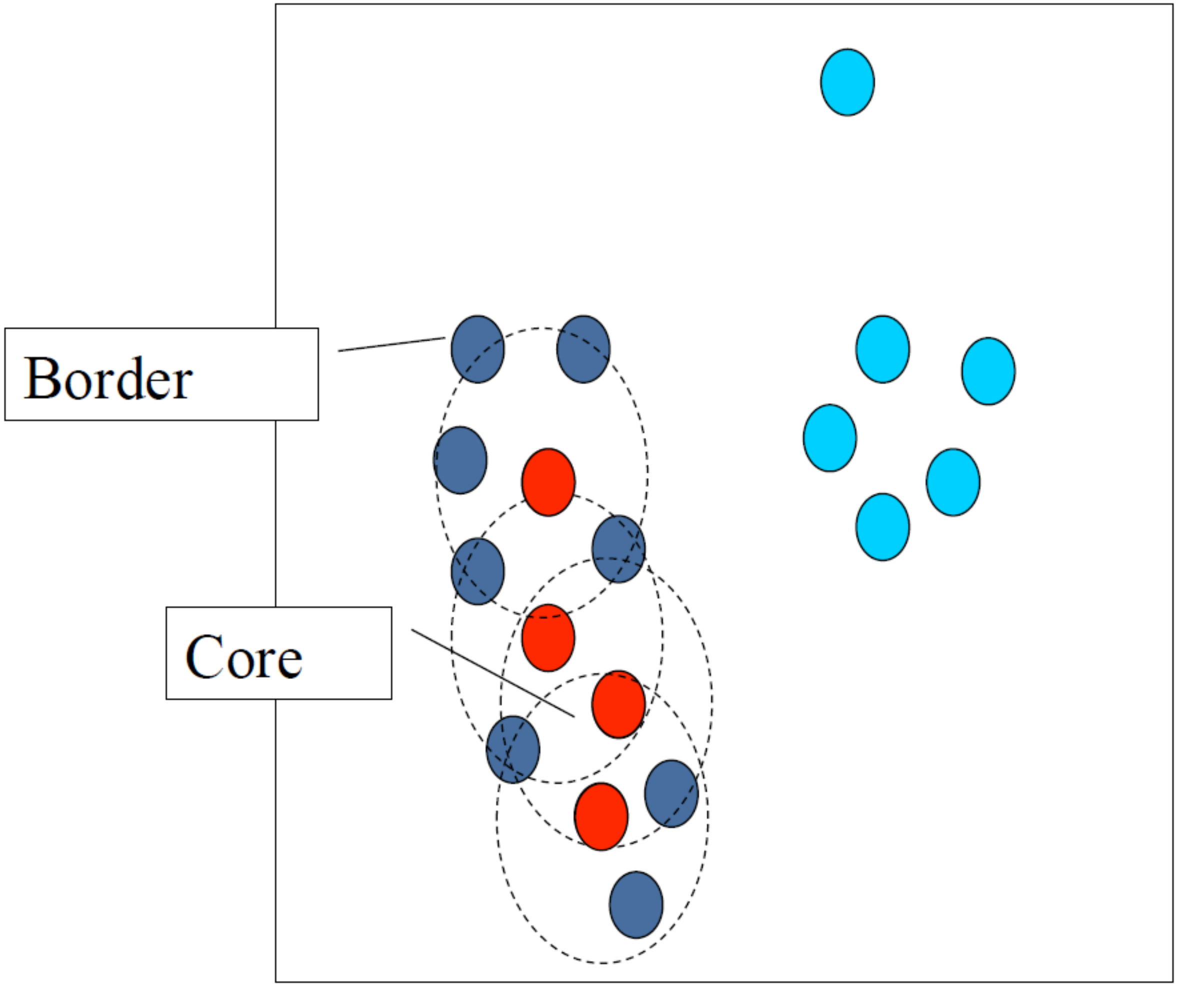
Cluster analysis – DBSCAN



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

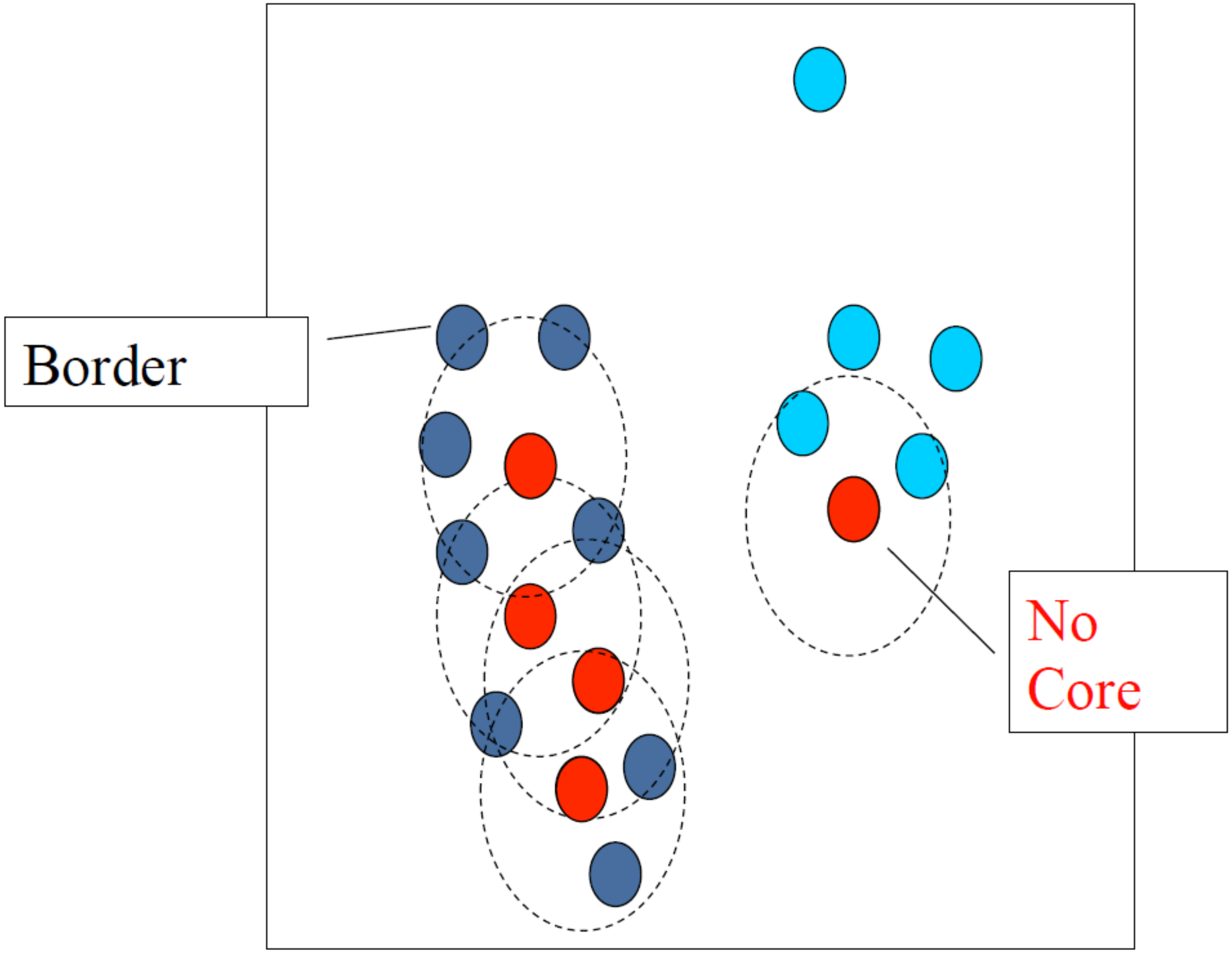
Cluster analysis – DBSCAN



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – DBSCAN



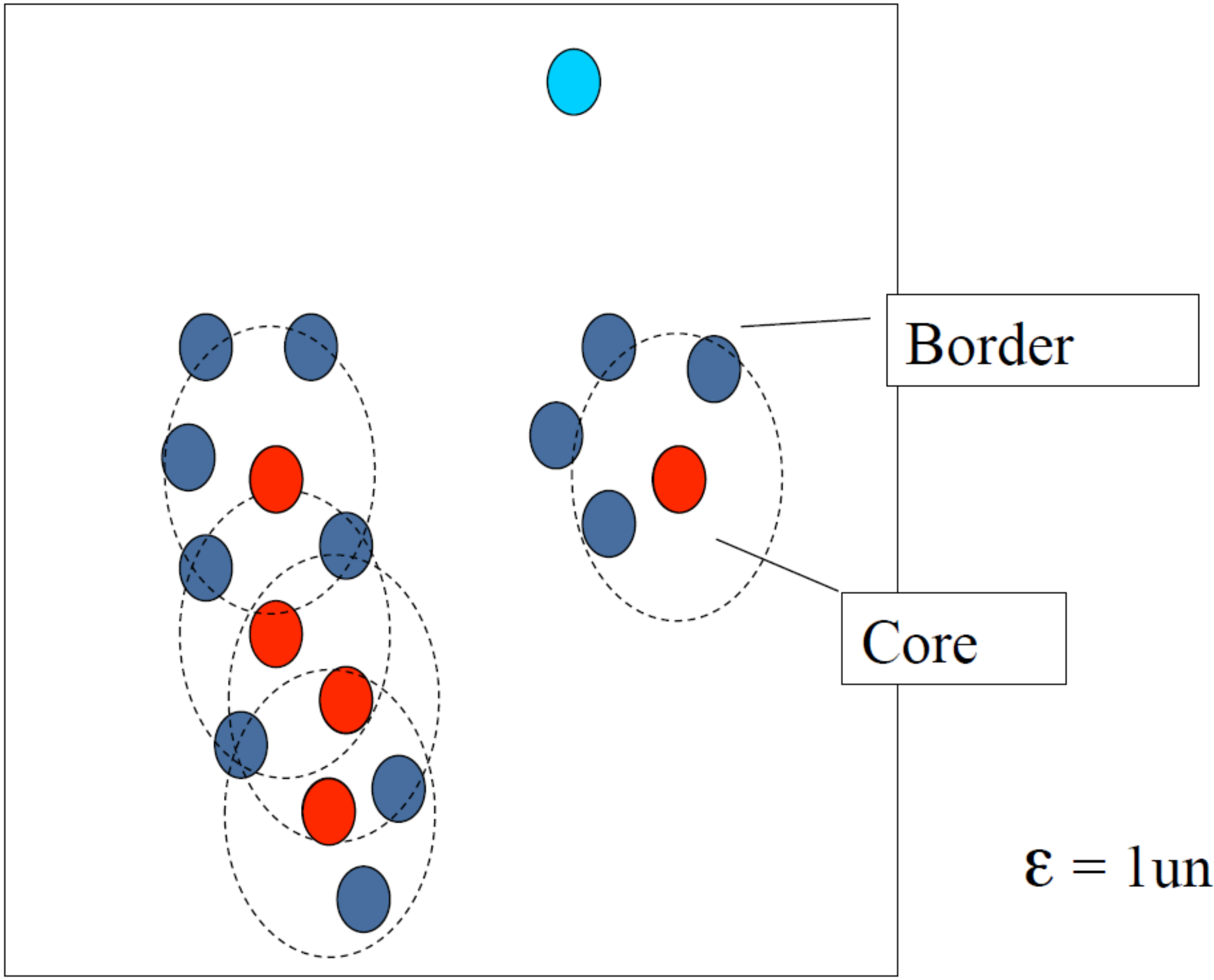
Border

No Core

$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – DBSCAN



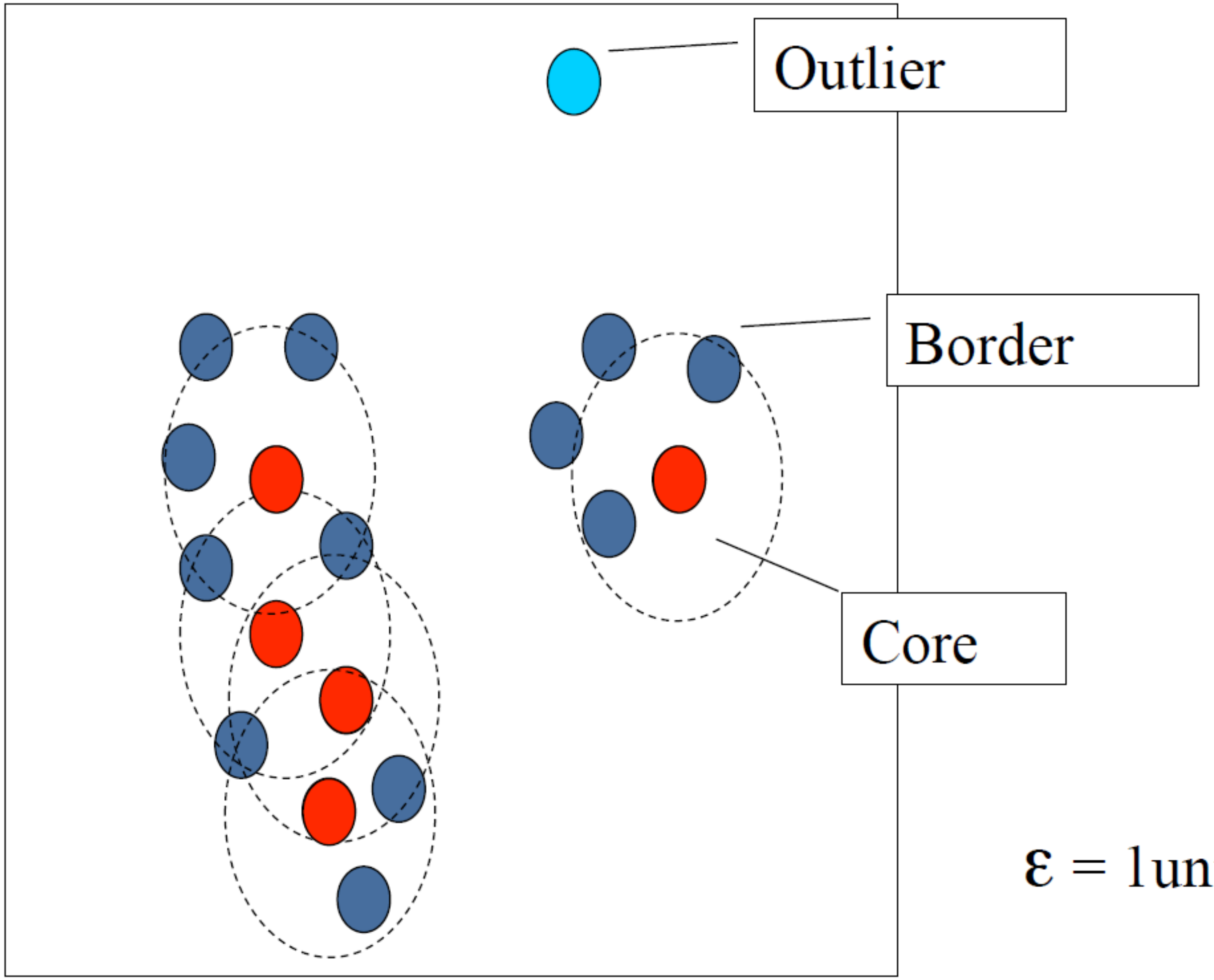
Border

Core

$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

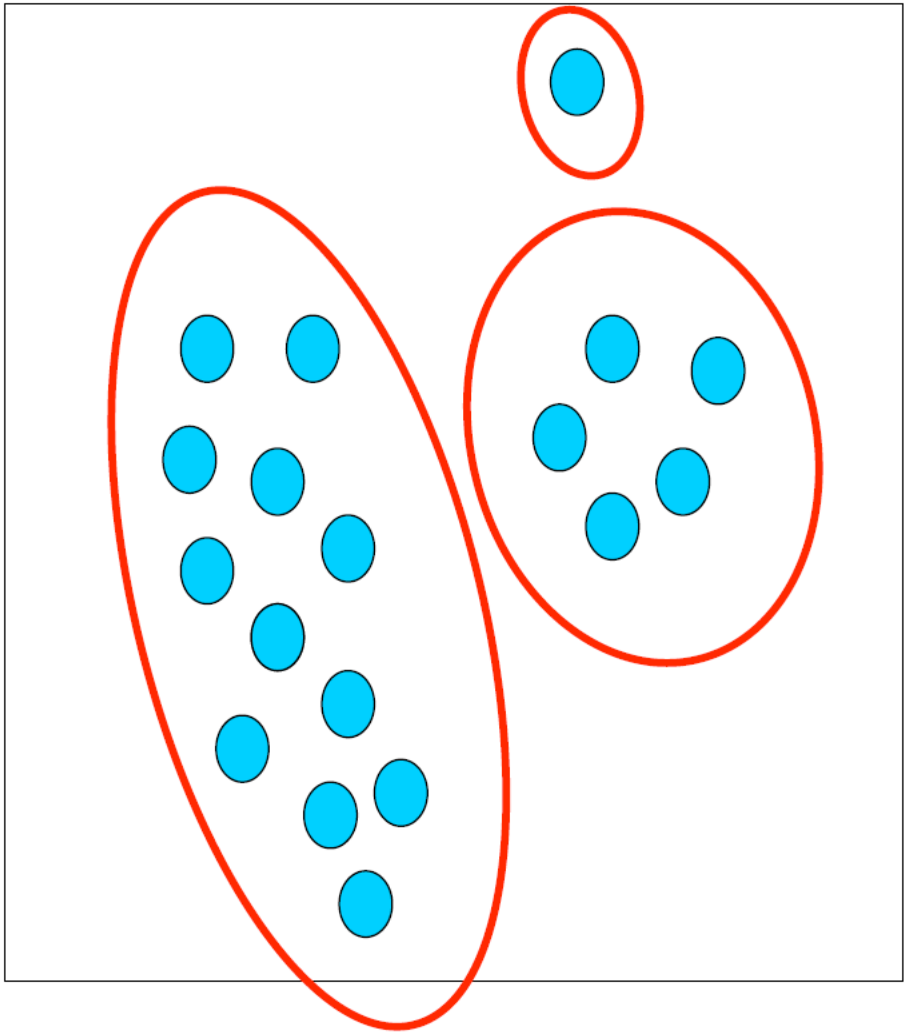
Cluster analysis – DBSCAN



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

Cluster analysis – DBSCAN



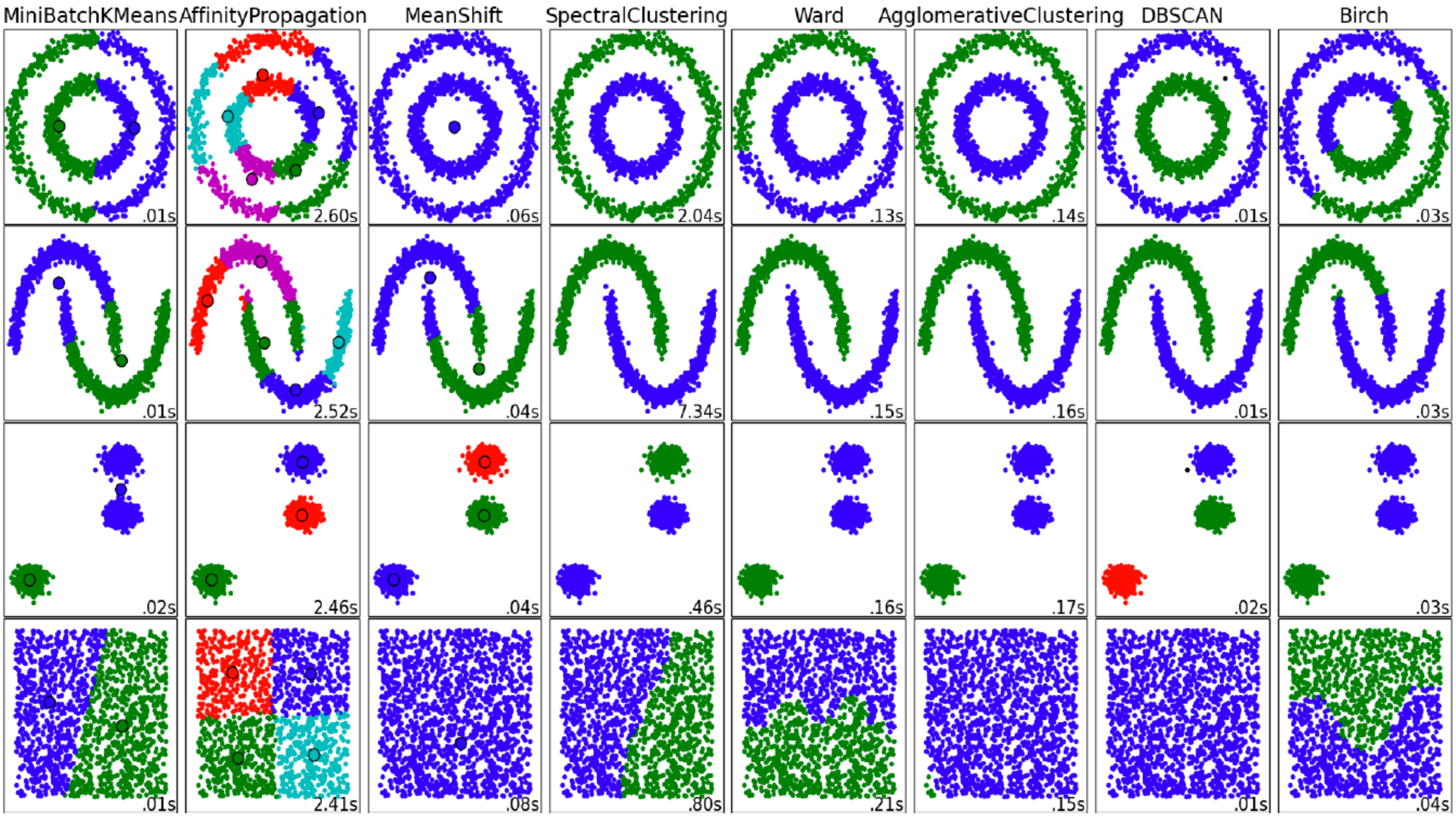
$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Source: Saeed Aghabozorgi, Cluster Analysis

Main clustering algorithms

- **Partition based (K-means)**:
  - Medium and large sized databases (relatively efficient)
  - Produces sphere-like clusters
  - Needs number of clusters (K)
- **Partition based (FCM)**:
  - Produces fuzzy clusters
  - Long computational time
- **Hierarchical based (agglomerative)**:
  - Produces trees of clusters
- **Density based (DBScan)**:
  - Produces arbitrary shaped clusters
  - Good when dealing with spatial clusters (maps)

Cluster analysis – comparison



Source: Saeed Aghabozorgi, Cluster Analysis

Applications of clustering

- **Retail / Marketing**:
  - ❑ Identifying buying patterns of customers
  - ❑ Finding associations among customers demographic characteristics
  - ❑ To recommend a new book, or to new customer by identifying clusters of books or clusters of customer preferences
- **Education**:
  - ❑ Education professionals may want to know the likes and dislikes of their students, they can create and understand the different groups and then package and market the various courses
- **Banking**:
  - ❑ Clustering normal transactions to find patterns of fraudulent credit card use
  - ❑ Identifying clusters of customers, e.g., loyal
  - ❑ Determining credit card spending by customer groups

Applications of clustering

- **Insurance**:
  - ❑ Fraud detection in claims analysis
  - ❑ Insurance risk of customers
- **Publishing / Media**:
  - ❑ Automatically categorizing news based on their content
  - ❑ Recommending similar news articles
  - ❑ Tagging news
  - ❑ Automatic fact checking
- **Medicine**:
  - ❑ Characterizing patient behaviour based on similar characteristics
  - ❑ Identifying successful medical therapies for different illnesses
- **Biology**:
  - ❑ Clustering genetic markers to identify family tries

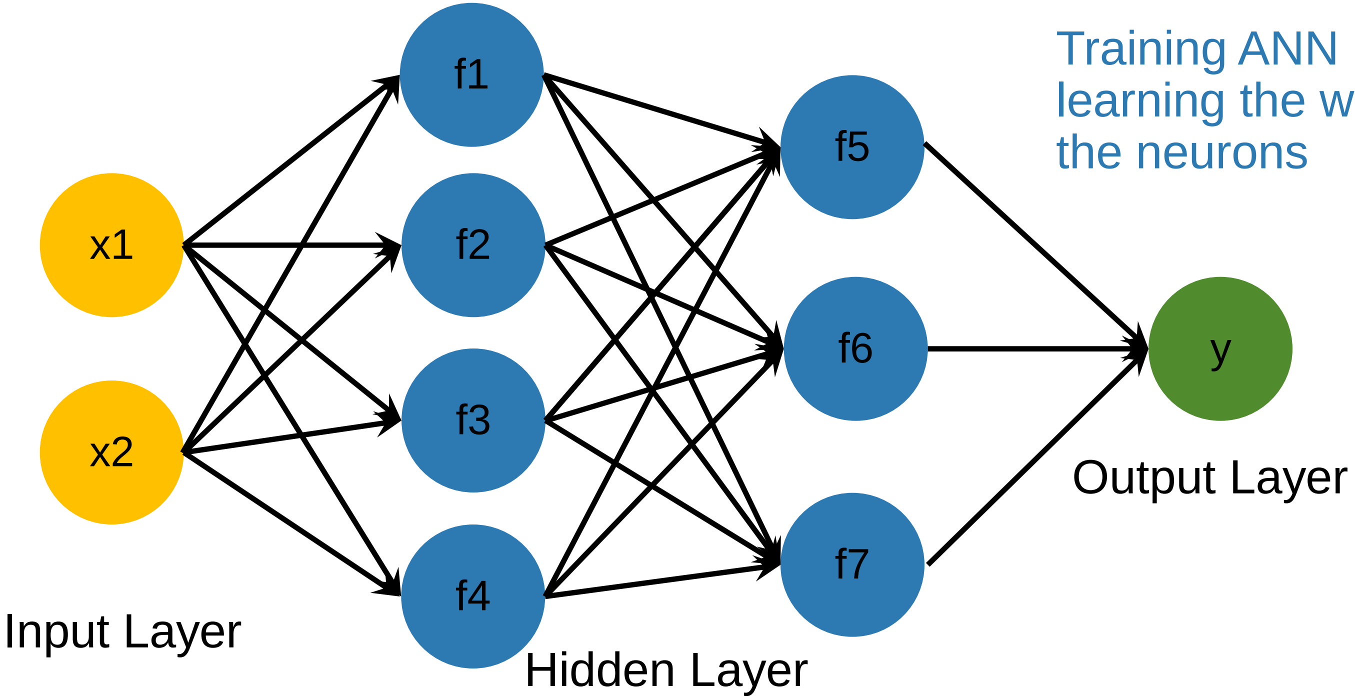
📖 Other Machine Learning Algorithms

Association rules

- Frequently called Market Basket Analysis is an unsupervised learning algorithm (no target variable)
- Detects associations (affinities) between variables (items or events)
- If customer purchased bread and bananas, s/he has an 80% probability to purchase milk during the same trip
- Multiple applications:
  - Cross-sell and up-sell
  - Targeted Promotions
  - Product bundling
  - Store planograms
  - Assortment optimization

Neural networks

- Based loosely on computer models of how brains work
- Model is an assembly of inter-connected neurons (nodes) and weighted links
- Each neuron applies a nonlinear function to its inputs to produce an output
- Output node sums up each of its input value according to the weights of its links
- Used for classification, pattern recognition, speech recognition
- “Black Box” model – no explanatory power, very hard to interpret the results

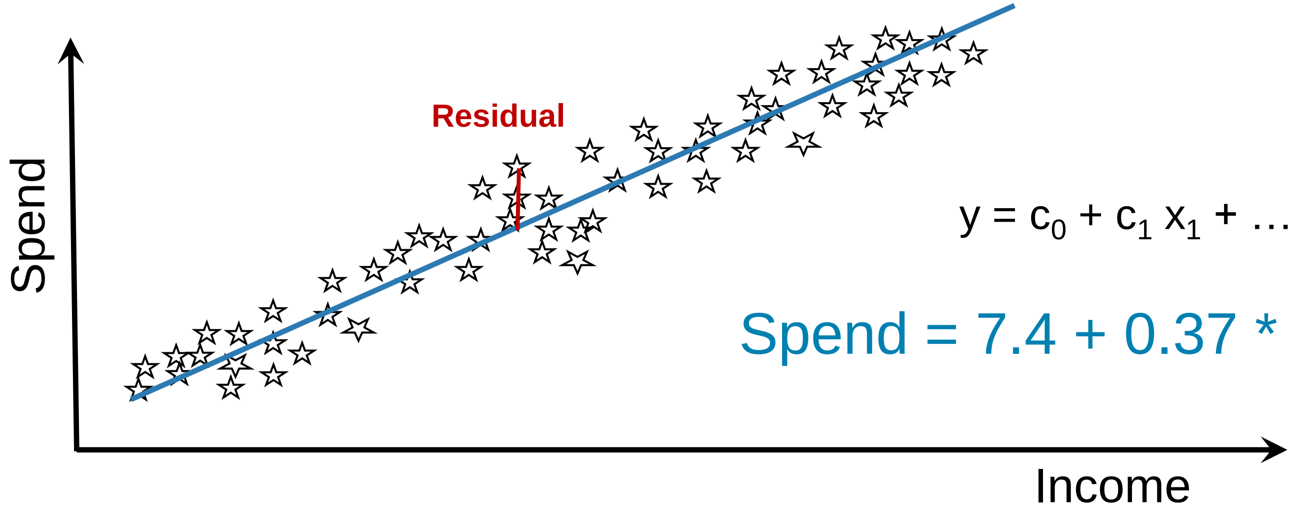


Training ANN means learning the weights of the neurons

Linear regression

- Predict a value of a given continuous variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Virtually endless applications:
  - Election outcomes
  - Future product revenues or commodity prices
  - Wind velocity

✓Both predictive and explanatory power



$y = c_0 + c_1 x_1 + \ldots + c_n x_n$

Spend = 7.4 + 0.37 * Income

Other types of regression analysis

Quantile regression

- Ordinary least squares regression approximates the conditional mean of the response variable, while quantile regression is estimating either the conditional median or other quantiles of the response variable
- This is very helpful in case of skewed data (i.e. income distribution in the US) or to deal with data without suppressing outliers

Logistic regression

- Logistic regression is used to predict categorical target variable
- Most often a variable with a binary outcome
  - ➢ Logit and Probit regressions can also be used to predict binary outcome. While , the underlying distributions are different, all three models will produce rather similar outcomes
- It is frequently used to estimate the probability of an event
  - o Bank customer defaulting on the loan
  - o Customer responding to a marketing promotion