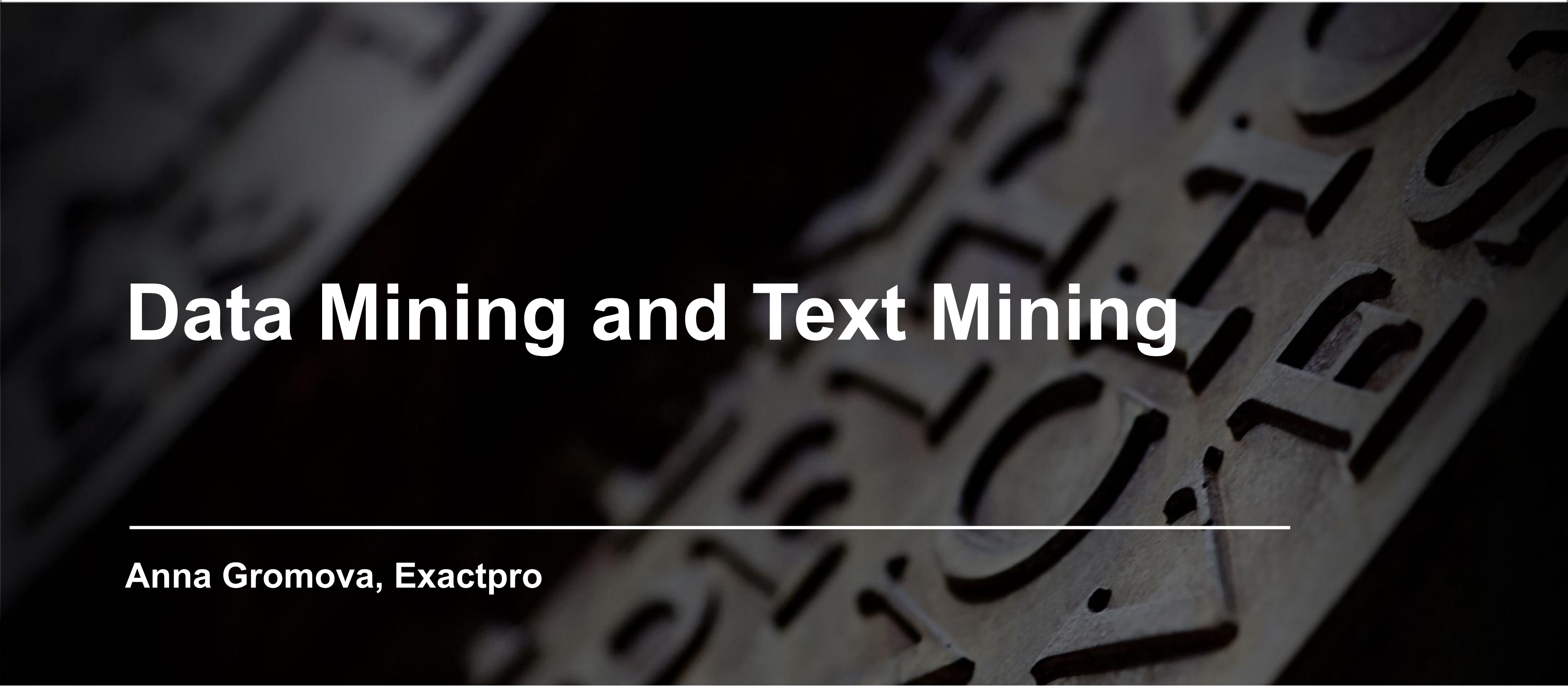
exactpro
EXITUS ACTA PROBAT

London
Stock Exchange Group



Data Mining and Text Mining

Anna Gromova, Exactpro

Open Access Quality Assurance & Related Software Development for Financial Markets
Tel: +7 495 640 2460, +1 415 830 38 49
www.exactpro.com

Key definitions

Artificial intelligent

An area of study in the field of computer science. Artificial intelligence is concerned with the development of computers able to engage in human-like thought processes such as learning, reasoning and self-correction.

The concept that machines can be improved to assume some capabilities normally thought to be like human intelligence such as learning, adapting, self-correction, etc.

The extension of human intelligence though the use of computers, as in times past physical power was extended through the use of mechanical tools.

In restricted sense, the study of techniques to use computers more effectively by improved programming techniques.

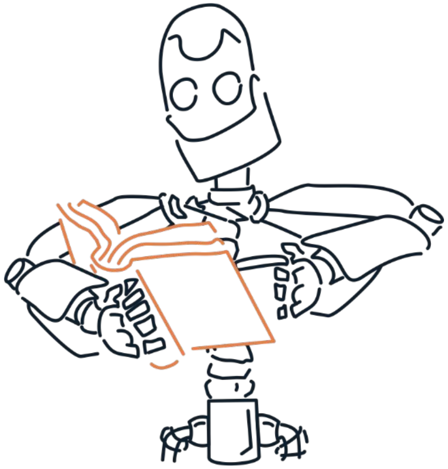
The New International Webster's Comprehensive Dictionary of the English Language

Key definitions

Machine learning

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.
T. Mitchell "Machine learning"

Vast amounts of data are being generated in many fields, and the statisticians's job is to make sense of it all: to extract important patterns and trends, and to understand "what the data says". We call this learning from data.
T. Hastie, R. Tibshirani, J. Friedman "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition"



One of the most interesting features of machine learning is that it lies on the boundary of several different academic disciplines, principally computer science, statistics, mathematics, and engineering. …machine learning is usually studied as part of artificial intelligence, which puts it firmly into computer science …understanding why these algorithms work requires a certain amount of statistical and mathematical sophistication.
S. Marsland "Machine Learning: An Algorithmic Perspective"

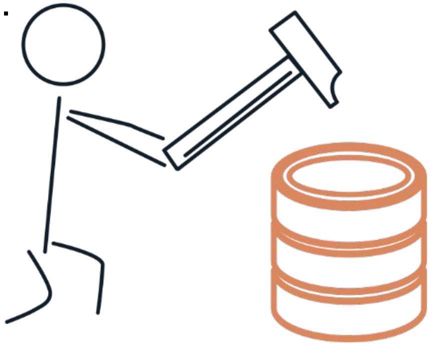
Key definitions

Data mining

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data. … Machine learning provides the technical basis for data mining. It is used to extract information from the raw data in databases…
I. Witten, E. Frank “Data Mining: Practical Machine Learning Tools and Techniques“

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.”
J.i Han, M. Kamber «Data Mining: Concepts and Techniques

KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data.
U. Fayyad, G. Piatetsky-Shapiro, P. Smyth “From Data Mining to Knowledge Discovery in Databases”



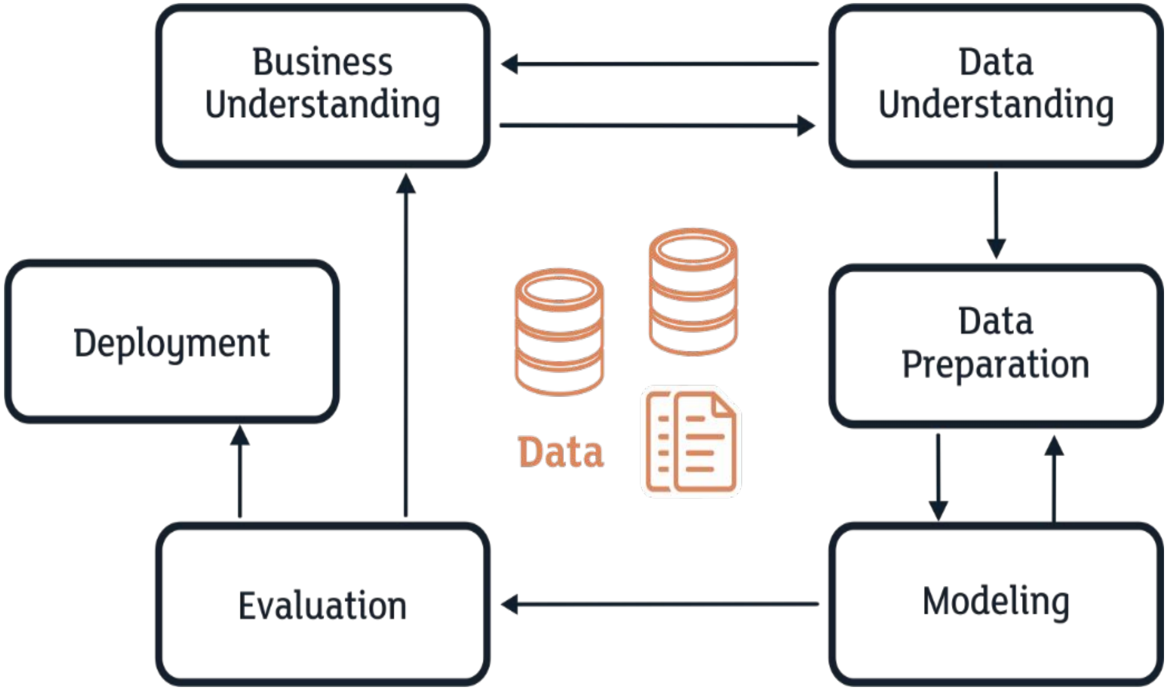
Key definitions

Text mining

Text mining is a variation on a field called data mining,that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text.

V. Gupta and G. S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Web Technologies in Web Technologies, Vol. 1, No 1, 2009

Process model for Data/Text mining



Cross Industry Standard Process for Data Mining

Data mining

Application:

- Financial data analysis (loan payment prediction, consumer credit policy analisys, price movement, detection of money laundering and etc.)
- Biomedical data analysis (diagnostic tasks, prediction of disease)
- Retail industry (identify customer buying behaviours, discover customer shopping paterns, design more effective goods transportation and etc.)

Data mining

Type of attributes:

- Nominal (categorical)
- Binary
- Ordinal
- Numeric

Data mining

Data preparation:

- Representative samples
- Categorial value
- Normalization
- Missing and empty value
- Anomaly detection
- Smooth noisy data

Data mining

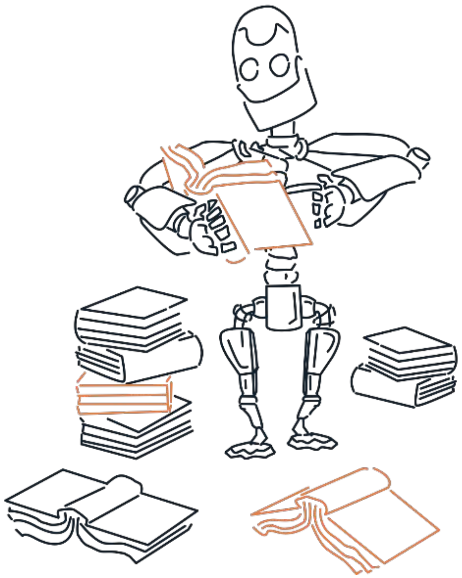
Tasks:

- Classification
- Regression
- Clustering
- Associating rule learning

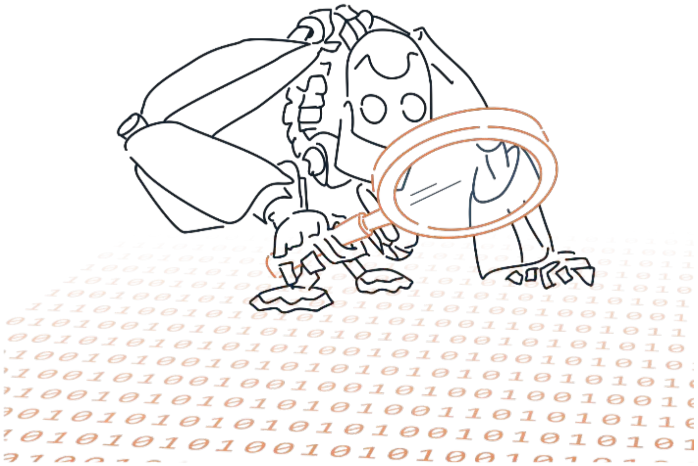
Data mining

Type of learning:

Supervised Learning



Unsupervised Learning



- Hold-out=Training set (70%) + Validation set (30%)
- Cross-validation

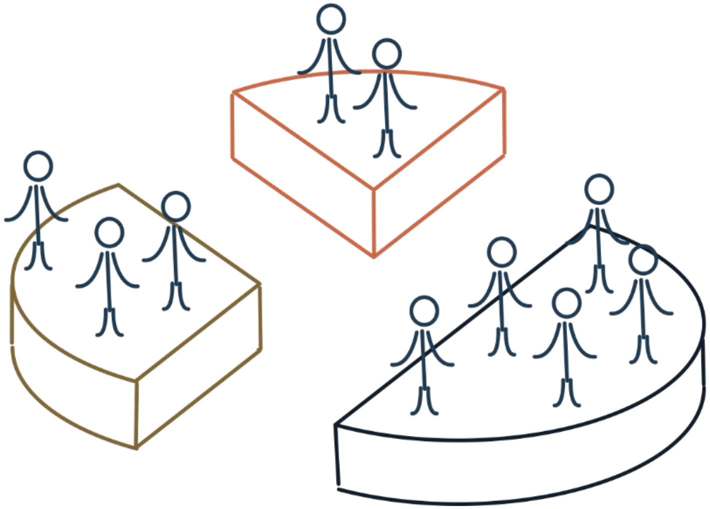
Data mining

Classification:

- $I \ = \ \{i_1, i_2, ..i_j, ..i_n\} i_j - \ object$
- $i_j \ = \ \{x_1, x_2, ..x_h, ..x_m, Y\},$

X – independent variables, Y – depended variable

- $v_h \ = \ \{v_{h1}, v_{h2} \ .... \}$
- $v_y \ = \ \{v_{y1}, v_{y2} .. v_{yk}\}$



Data mining

Example: “Heart desease prediction”

I = {id1, id2....} //patient

Ij = {gender, age, smoking, overweight, alcohol_intake, high_salt_diet, high_saturated_fat_diet, exercise, hereditary, bad_cholesterol, blood_ pressure, blood_shugar, heart_rate, **heart_desease** }

Gender = {0,1}, alcohol ={never, past, current}, blood_shugar= {<90, >90&<120, >120}

Heart_desease = {0,1}

Jyoti Soni, Ujma Ansari, Dipesh Sharma, “Predictive data mining for Medical Diagnosis: an overview of heart disease prediction ”

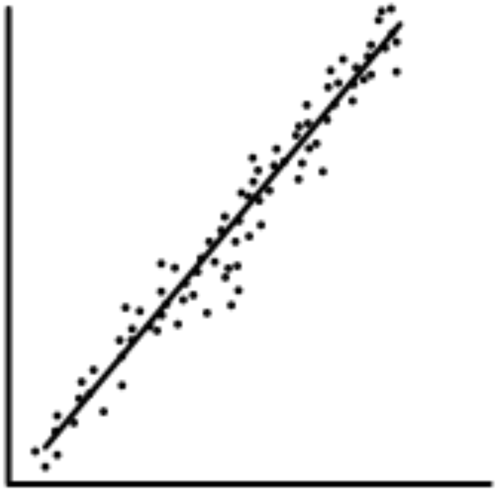
Data mining

Regression:

- $I \ = \ \{i_1, i_2, ..i_j, ..i_n\} i_j - \ object$
- $i_j \ = \ \{x_1, x_2, ..x_h, ..x_m, Y\},$

X – independent variables, Y – depended variable

- $v_h \ = \ \{v_{h1}, v_{h2} \ ....\}$
- $v_y \in R$



Data mining

Example: Electricity market price forecast

I = {id1, id2....} //time

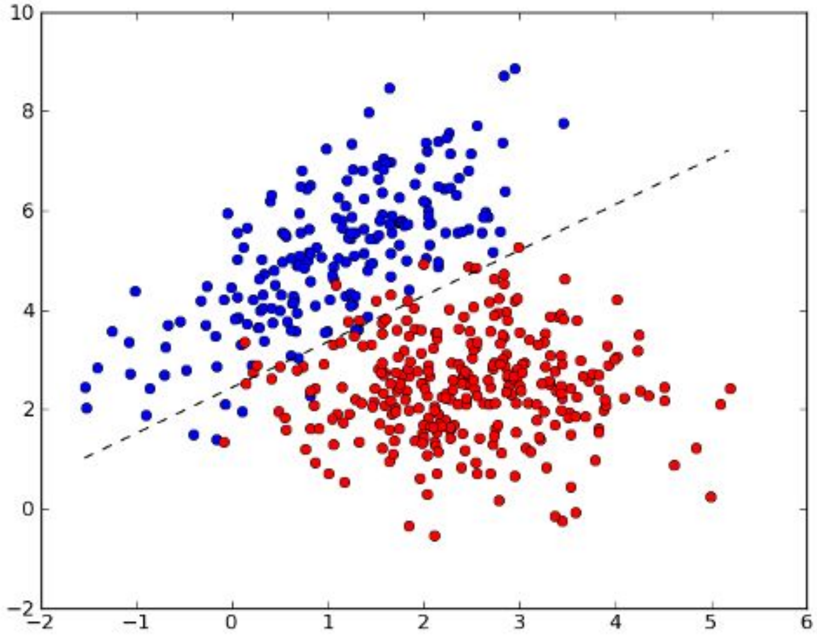
Ij = {Date, time, demand_el, supply_el, reserve_el, ∆demand_el, ∆ supply_el, ∆ reserve_el **regional_ref_price** }

Xin Lua, Zhao Yang Dongb, Xue Li "Electricity market price spike forecast with data mining techniques"

Data mining

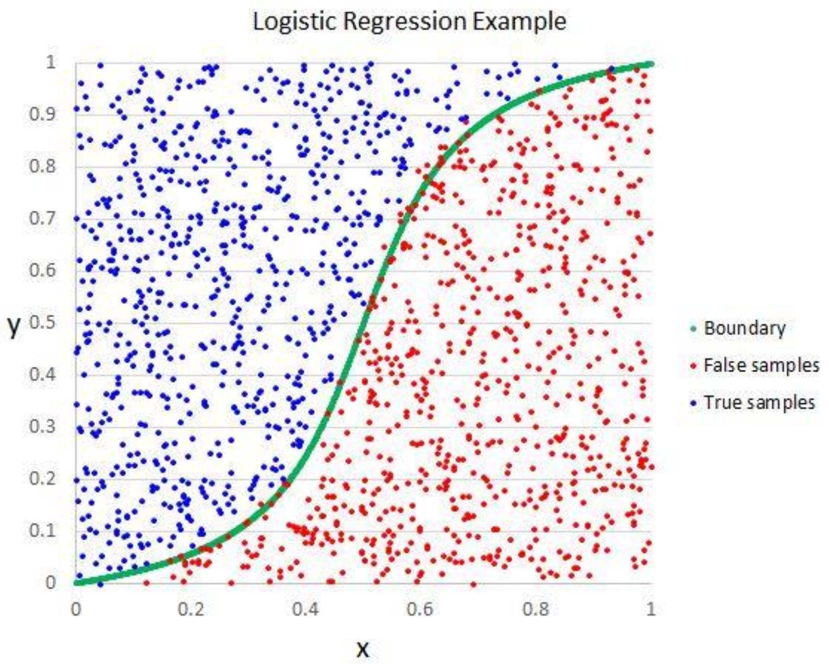
Linear regression

- $I_j = \{x_1, x_2, .. x_h .. x_m, Y\}$
- $Yi \; = \vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + .. + \vartheta_m x_m$



Logistic regression

- $I_j = \{x_1, x_2, ..x_h..x_m, Y\}$
- $Yi \ = \vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + .. + \vartheta_m x_m$
- $f(y) = \frac{1}{1+e^{-y}}$

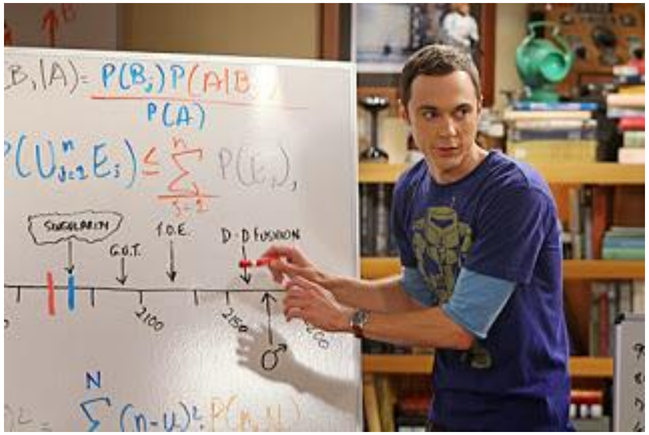


Data mining

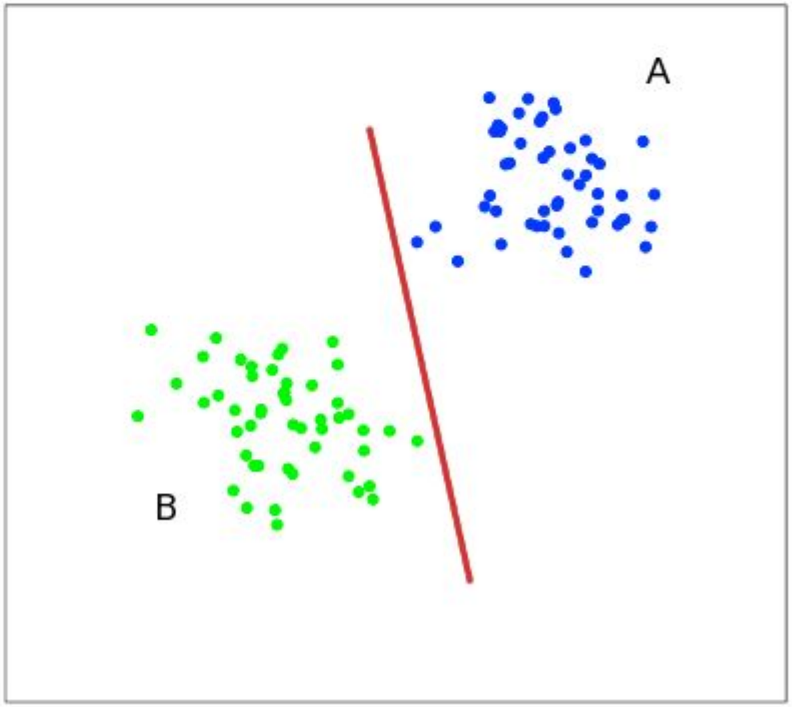
Naive Bayes

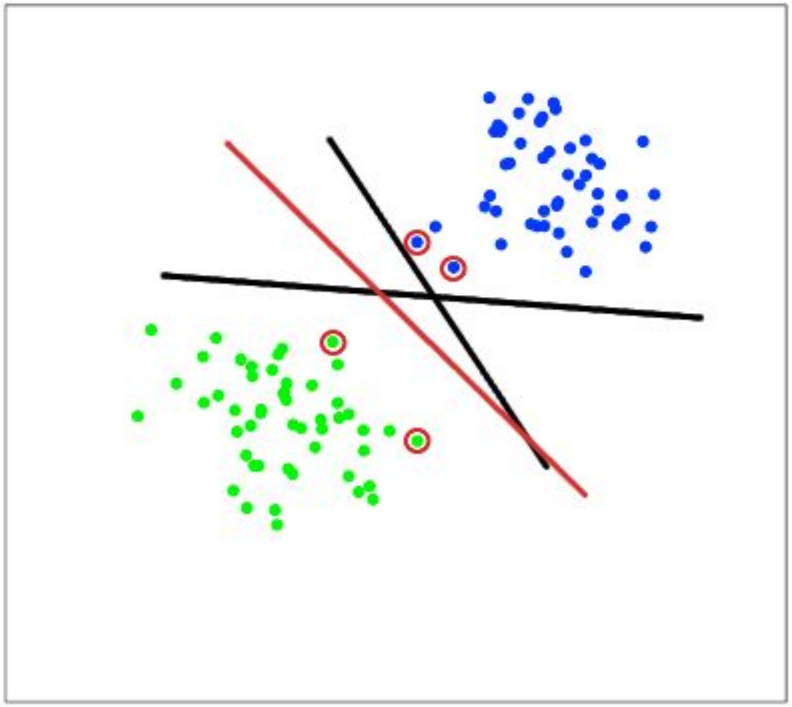
$$P(H_k \mid A) = \frac{P(H_k) * P(A \mid H_k)}{P(A)},$$

$P(H_k)$ - априорная вероятность события H_k
$P(H_k \mid A)$ - вероятность события H_k при наступлении A
$P(A \mid H_k)$ - вероятность наступления A при истинности H_k
$P(A)$ - полная вероятность события A



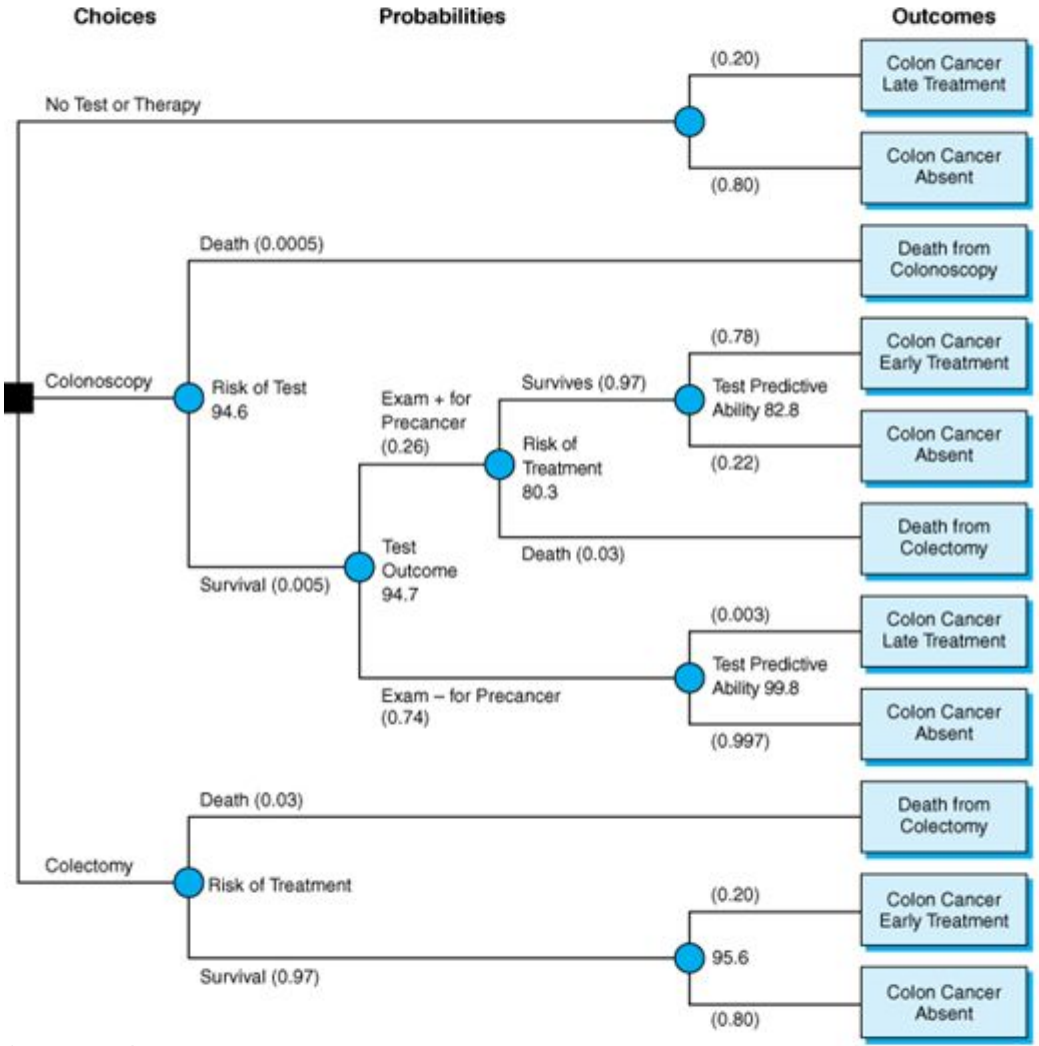
Support Vector Machine (SVM)





Data mining

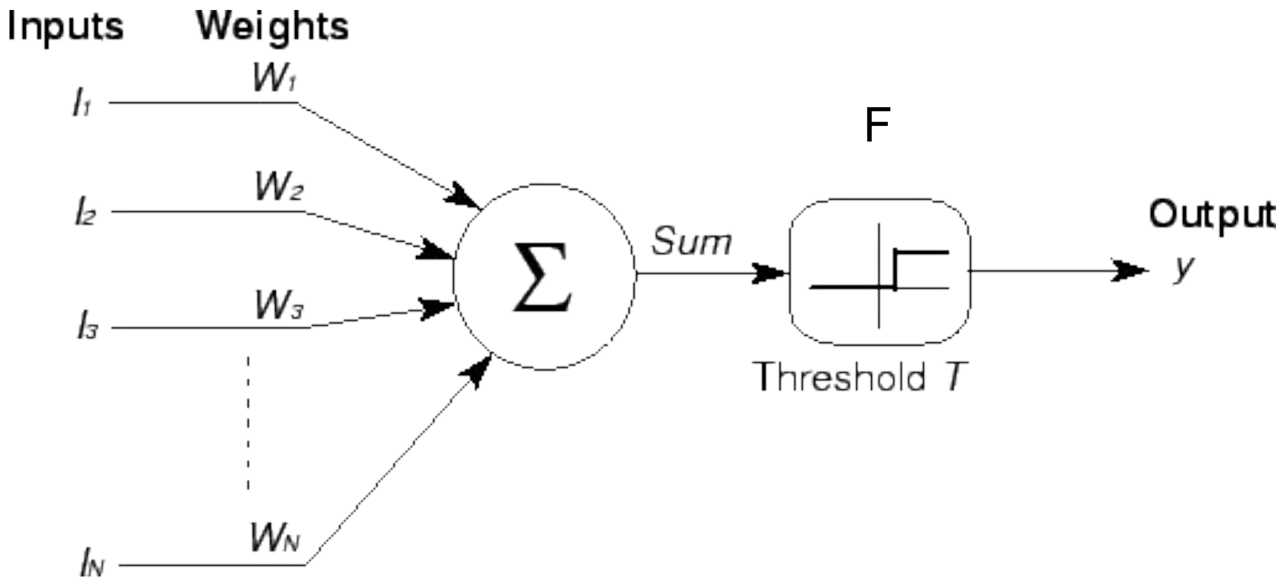
Decision tree



B. Dawson, R.G. Trapp “Basic &amp; Clinical Biostatistics, 4e”

Data mining

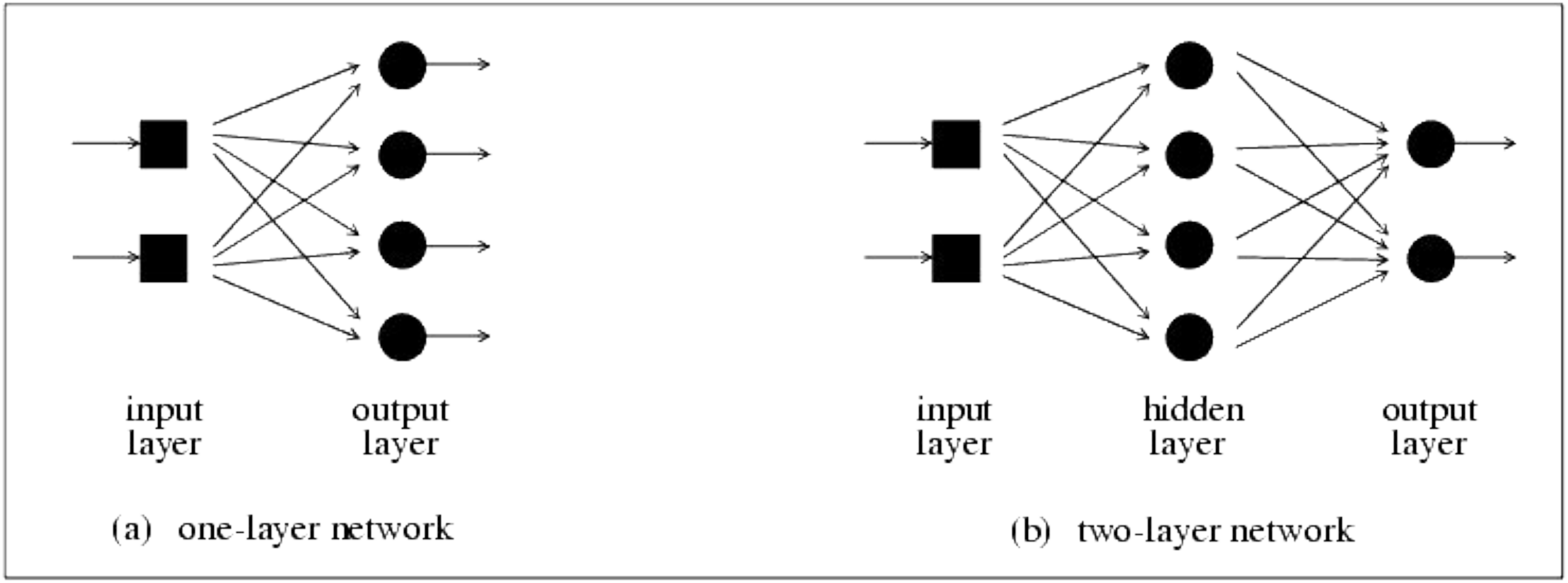
Neural network: formal neuron



$$Output = F(\sum_{j} (I_j * W_j) - \mathrm{T})$$

Data mining

Neural network



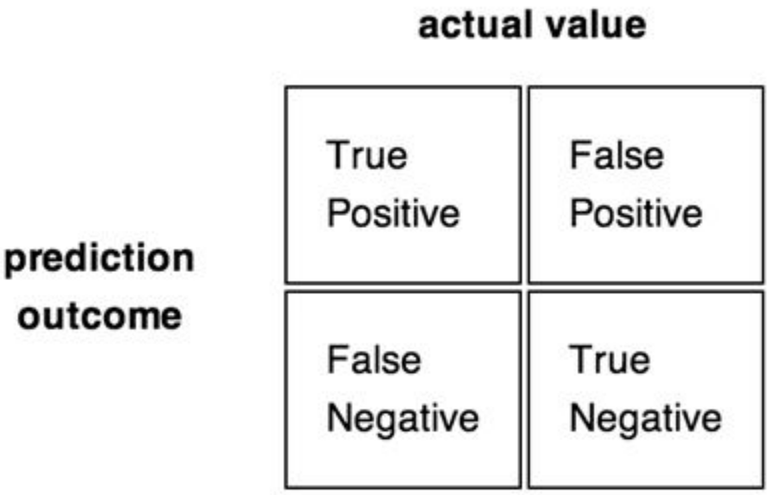
Data mining

Metrics:

Precision (positive predictive value): TP/(TP+FP)

Recall (true positive rate): TP/(TP+FN)

F-measure: $2 * \frac{precision*recall}{precision+recall}$



	actual value	
prediction outcome	True Positive	False Positive
	False Negative	True Negative

Data mining

Clustering:

- $I \; = \; \{i_1, i_2, ..i_j, ..i_n\}, i_j - \; object$
- $i_j \; = \; \{x_1, x_2, ..x_h, ..x_m\},$
- $v_h \; = \; \{v_{h1}, v_{h2} \; ....\}$
- $C = \{c_1, c_2..c_k..c_g\}$
- $c_k \; = \{i_j, i_p | i_j, i_p \in \mathrm{I} \; \& \, d(i_j, i_p) < \delta\}$

Data mining

Example: Clustering e-Banking Customer

I = {id1, id2....} //transaction

Ij ={date, time, status_of_transaction, type_of_transaction, RFM_score)

Date={d1, d2}, time={tI1, tI2, tI3, tI4},

Status_of_transaction={Real-time, schedule}

Type_of_transaction={balance, report, money_transfer, payment}

Waminee Niyagas, Anongnart Srivihok, Sukumal Kitisin “” Clustering e-Banking Customer using Data Mining and Marketing Segmentation

Data mining

K-means

number of clusters

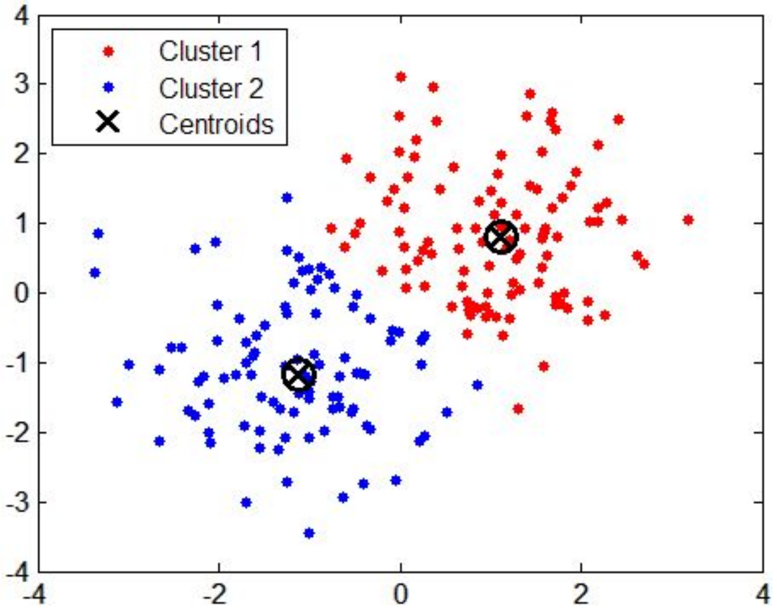
number of cases

case *i*

centroid for cluster *j*

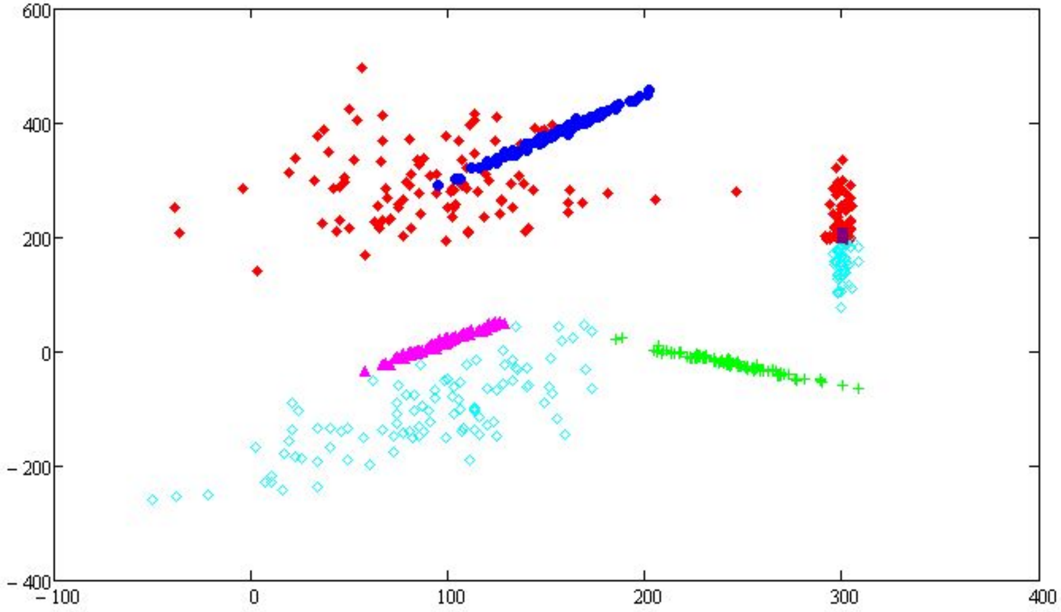
objective function ←

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$



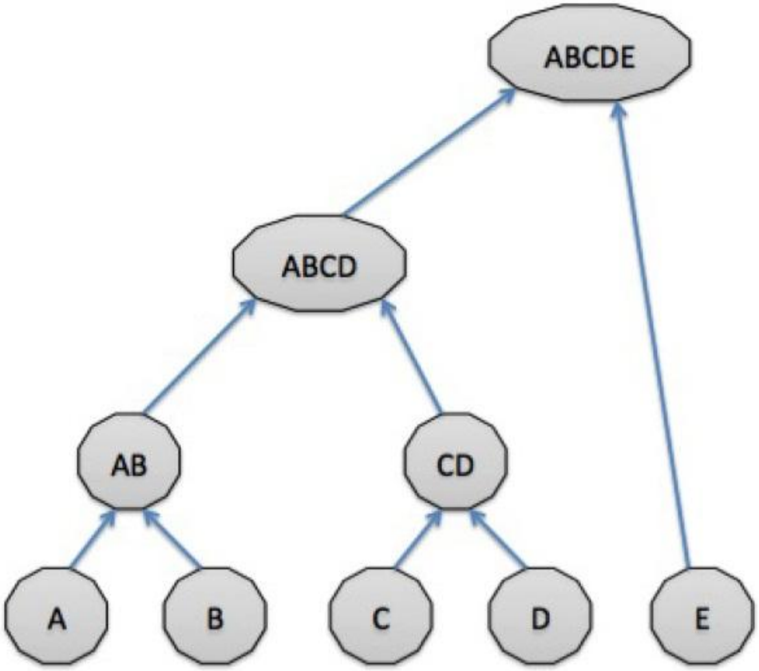
Data mining

EM-algorithm

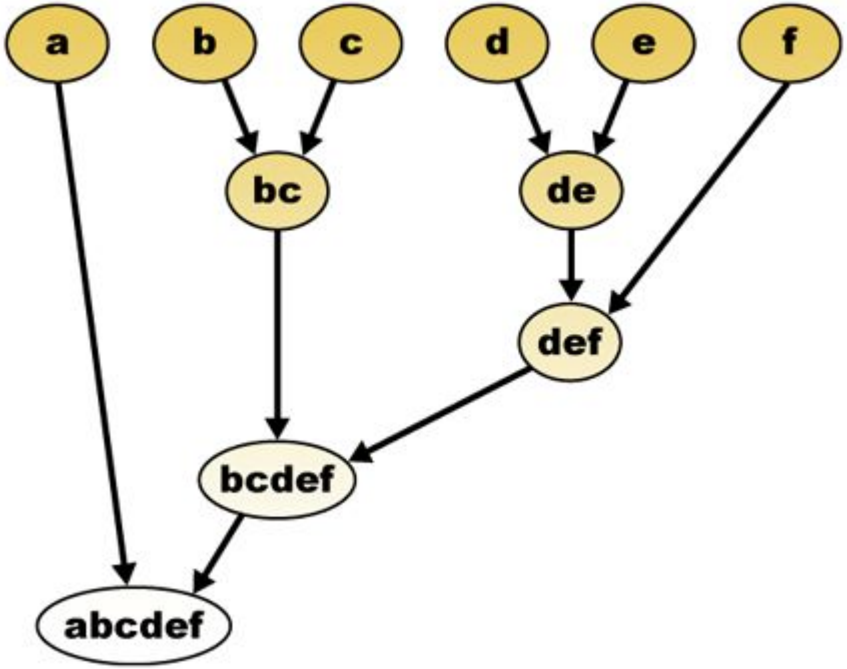


Data mining

Divisive algorithm



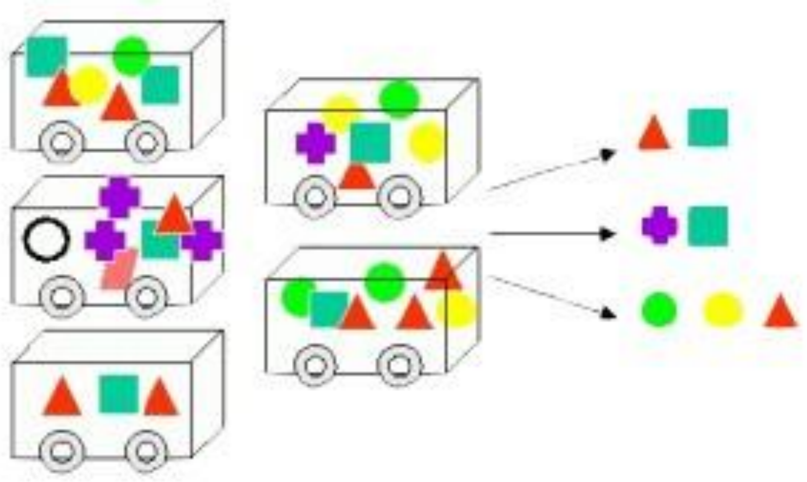
Agglomerative algorithm



Data mining: tasks

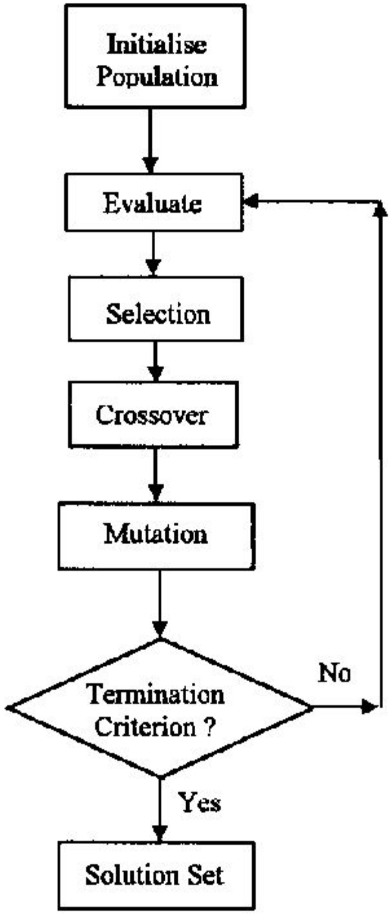
Associating rule :

- $I \ = \ \{i_1, i_2, ..i_j, ..i_n\}, i_j\text{–}\ object$
- $T = \{i_j\ |i_j \in I\}, T\ - transaction$
- $D = \{T_{1,} T_2\ ..T_r, ..T_m\}$
- $D_{i_j} = \{T_r|i_j \in T_r; j = 1..n; r = 1..m\}$
- $F\ = \{i_j\ |i_j \in I; j = 1..n\}$
- $D_f\ = \{T_r|F \subseteq T_r; \mathrm{r} = 1..\mathrm{m}\}$
- $Supp(F) = \frac{|D_F|}{|D|}$
- $L = \{F|Supp(F) > Supp_{min}\}$



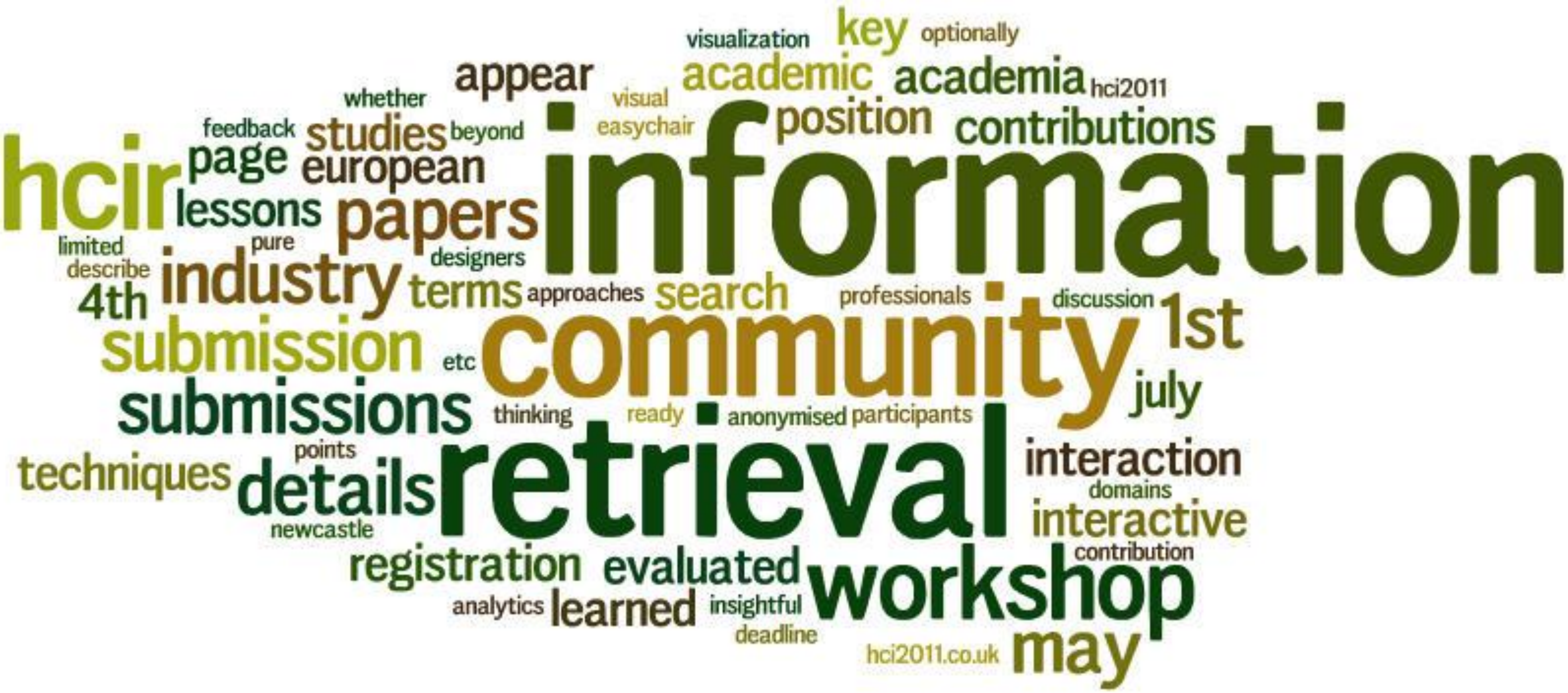
Data mining: tasks

Genetic algorithms:



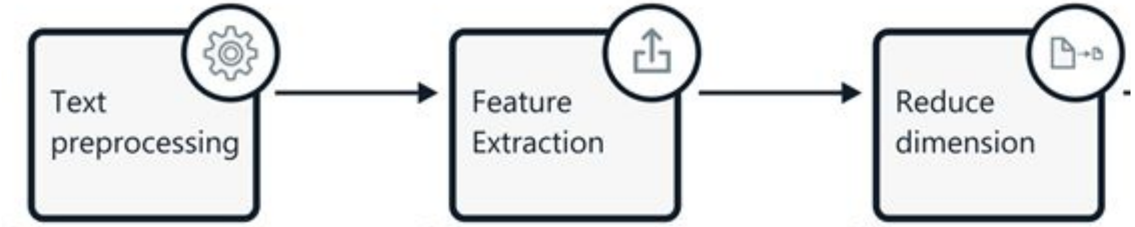
Text Mining

Information retrieval (IR) + natural language processing (NLP)



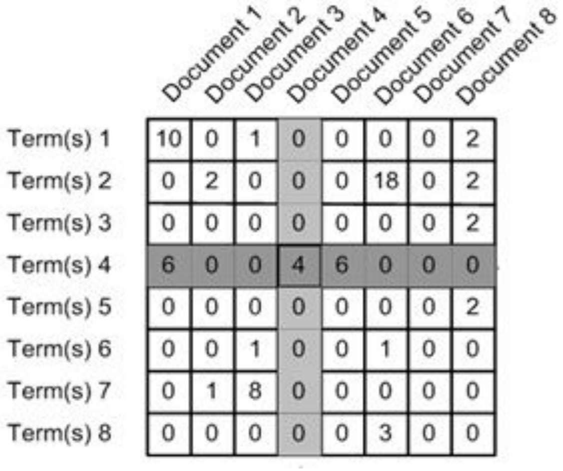
Text mining

Text preparation:



- Tokenization
- Removal stop-words
- Stemming
- Lemmatization

Bag-of-Words
(TF-IDF)



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

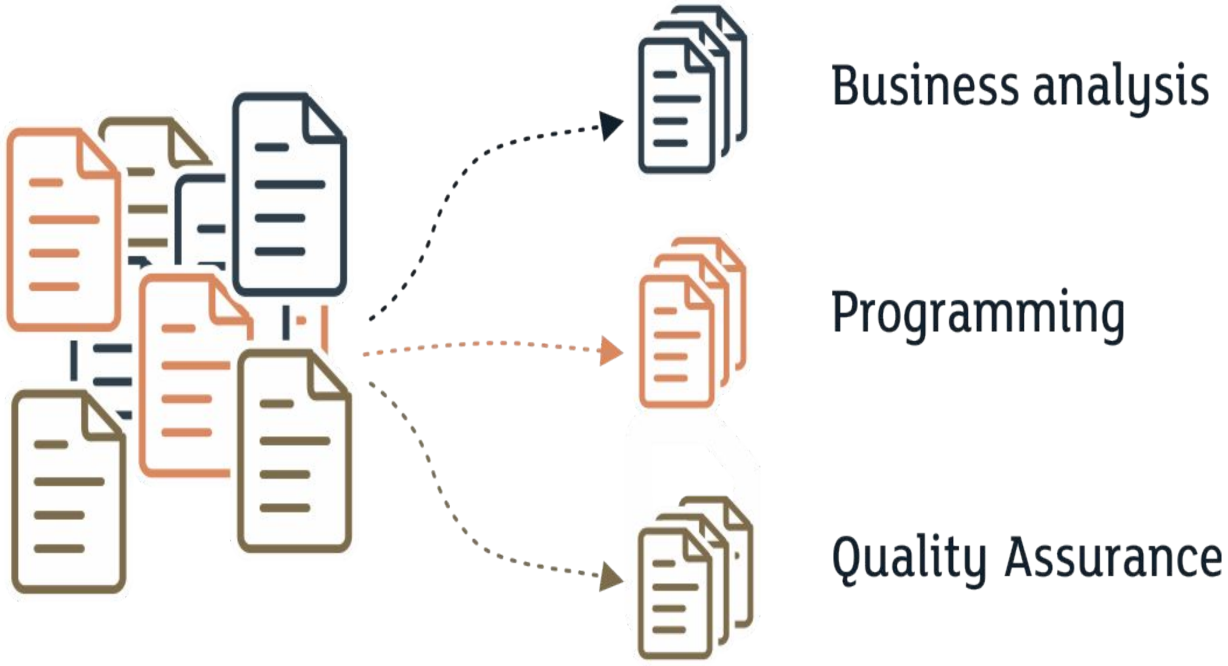
Text mining

Tasks:

- Classification
- Clustering
- Building ontology
- Information extraction
- Sentiment analysis
- Document summarisation

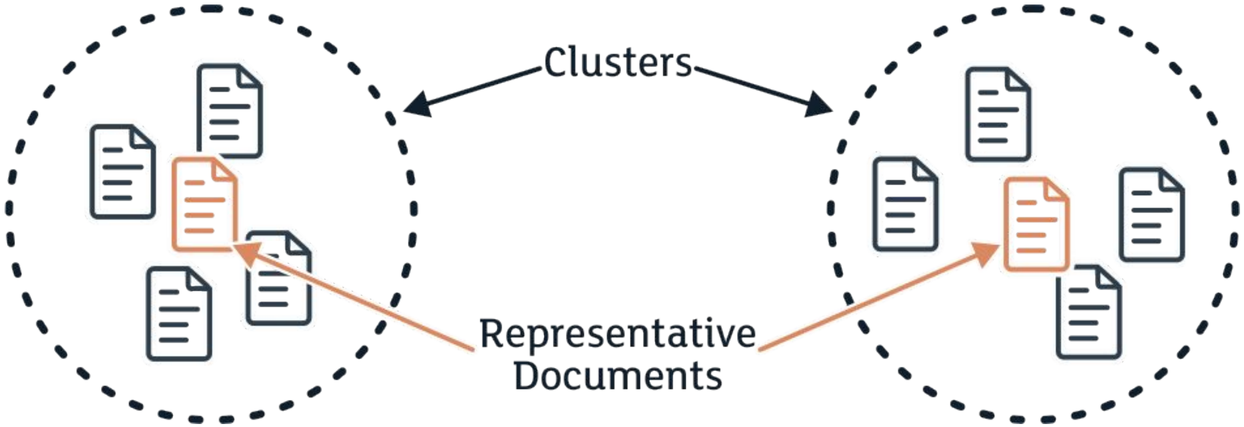
Text Mining

Text classification:



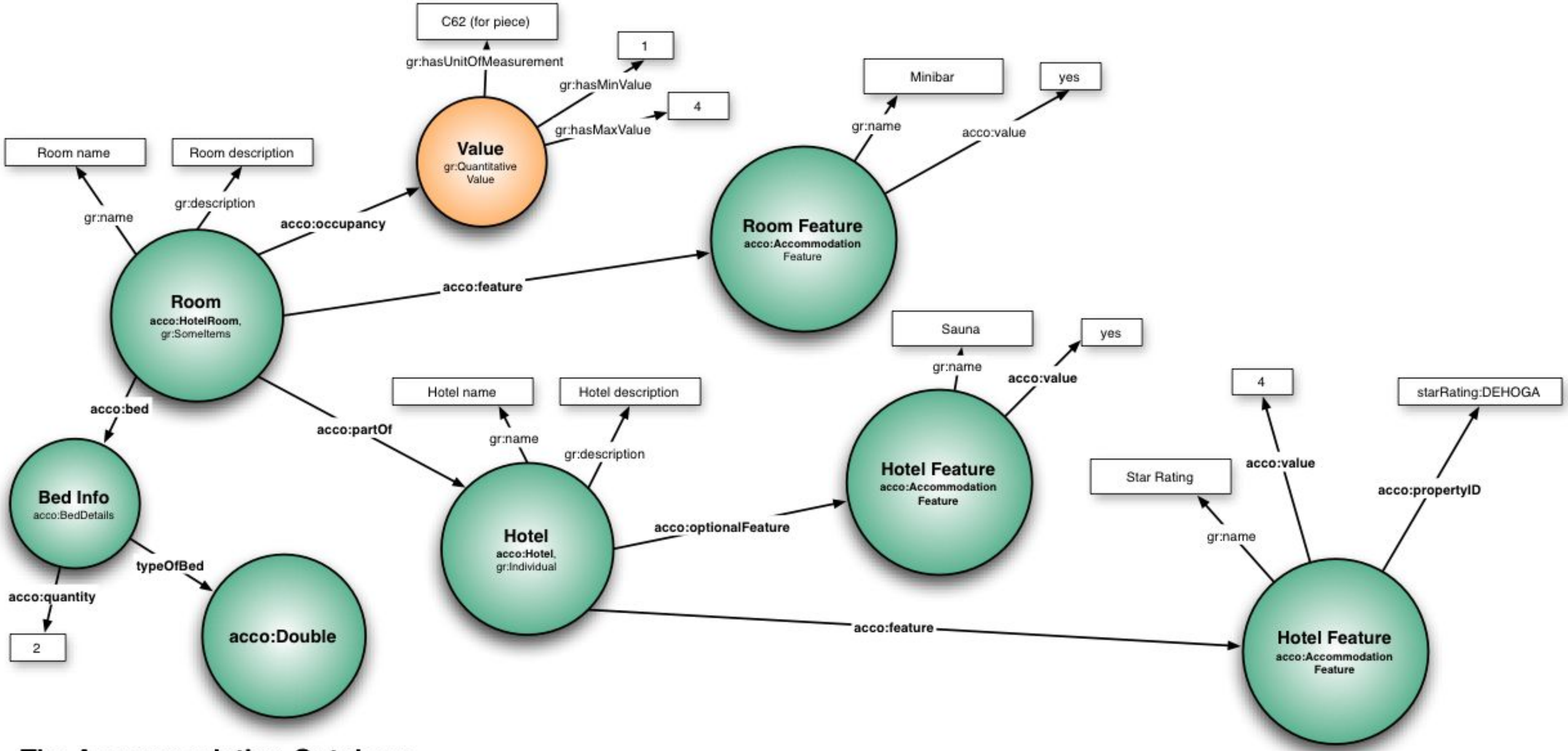
Text Mining

Clustering:



Text Mining

Ontology:

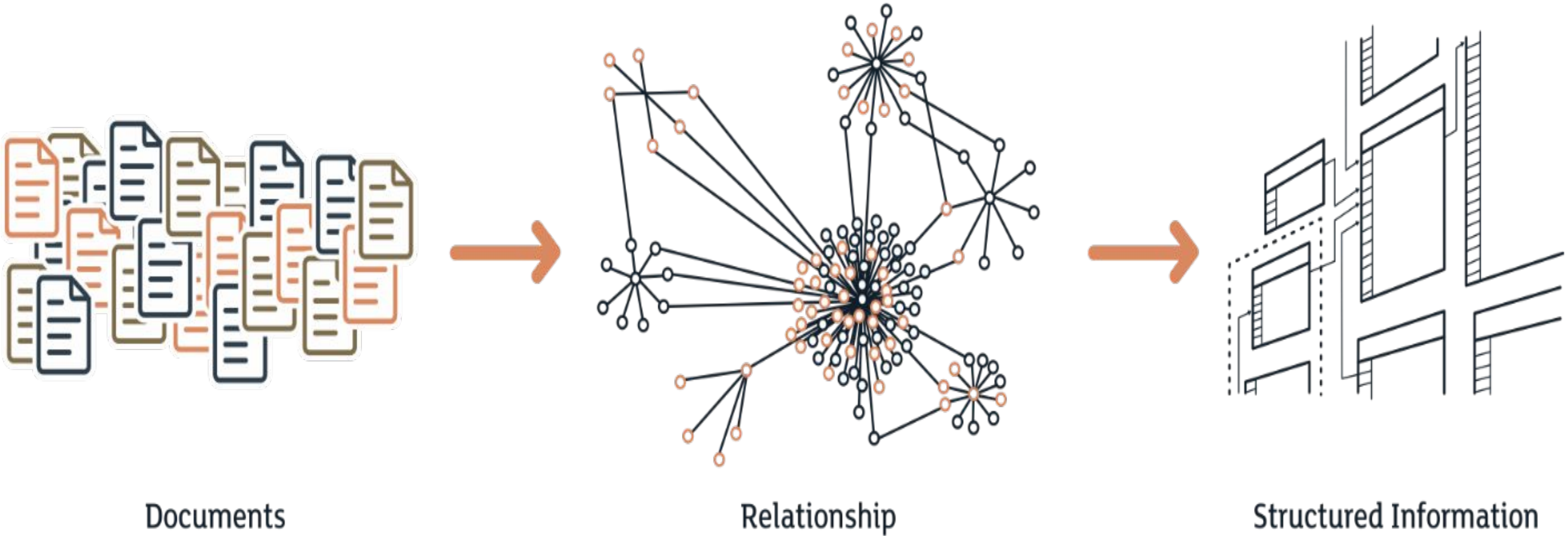


The Accommodation Ontology
http://purl.org/acco/ns

http://ontologies.sti-innsbruck.at/acco/ns.html

Text Mining

Information extraction:

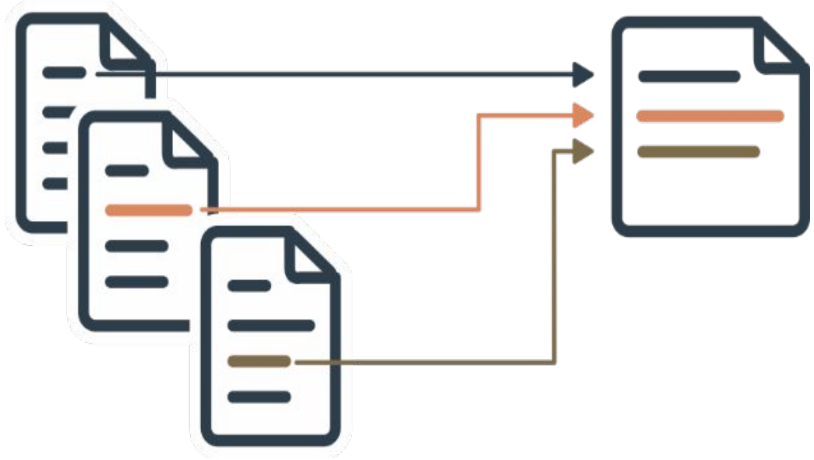


Text Mining

Sentiment analysis:

Text Mining

Document summarization:



Text Mining

Not covered in this lecture:

- Mathematical apparatus
- Time series
- Feature selection
- Fuzzy logic
- Genetic algorithms
- PCA
- Cobweb (clustering)
- LSA

References

Books:

- Чубукова И. А. Data Mining: учебное пособие.
- Барсегян А. А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-е издание.
- T. Mitchell “Machine learning”
- T. Hastie, R. Tibshirani, J. Friedman “The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition”
- S. Marsland “Machine Learning: An Algorithmic Perspective”
- I. Witten, E. Frank “Data Mining: Practical Machine Learning Tools and Techniques“
- J.i Han, M. Kamber “Data Mining: Concepts and Techniques”
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth “From Data Mining to Knowledge
- Discovery in Databases”
- C. D. Manning, P. Raghavan, H. Schutze “Introduction to Information retrieval”
- B. Dawson, R.G. Trapp “Basic &amp; Clinical Biostatistics, 4e” (example for decision tree)

Papers:

- V. Gupta and G. S. Lehal, “A Survey of Text Mining Techniques and Applications”
- Jyoti Soni, Ujma Ansari, Dipesh Sharma, “Predictive data mining for Medical Diagnosis: an overview of heart disease prediction ”
- Xin Lua, Zhao Yang Dongb, Xue Li “Electricity market price spike forecast with data mining techniques”
- Waminee Niyagas, Anongnart Srivihok, Sukumal Kitisin “Clustering e-Banking Customer using Data Mining and Marketing Segmentation”

Thank you!