

Introduction to Data Mining

Paweł Lula, Department of Computational Systems,
Cracow University of Economics
pawel.lula@uek.krakow.pl

Agenda

- Data mining definition
- Determinants of data mining development
- Scope of data mining applications
- Neural network as an example of typical data mining models
- Text mining and its application
- R as a platform for data mining analysis

DATA MINING DEFINITION

Data Mining

- The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. (Usama Fayyad, 1996).

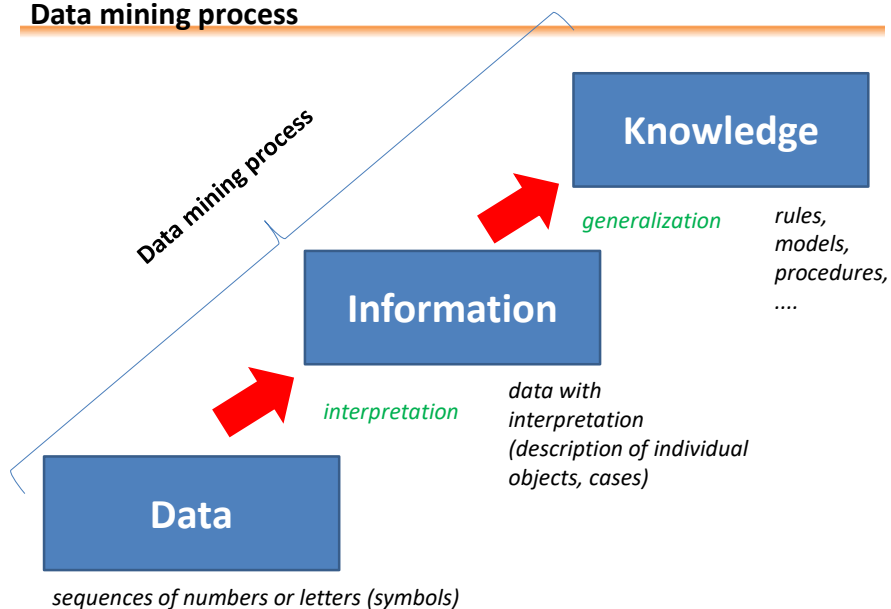


*Usama Fayyad (born July, 1965)
American data scientist*

Data Mining

- Goal:
 - discovering and modelling hidden, previously unknown patterns, relationships, correlations, structures, rules, ...
- Scope:
 - large and often heterogeneous data sets
- Results:
 - useful,
 - easy to understand and interpret.

Data mining process



DETERMINANTS OF DATA MINING DEVELOPMENT

Determinants of data mining development

- Lack of theory describing observed phenomena
- Information overload problem
- Progress in capabilities of computer systems

LACK OF THEORY DESCRIBING OBSERVED PHENOMENA

Paweł Lula, Cracow University of Economics

9

Theoretical background and data mining approach

- Lack of theory describing observed phenomena



we have only observations!!!

data-based character of data mining analysis

knowledge discovery in databases

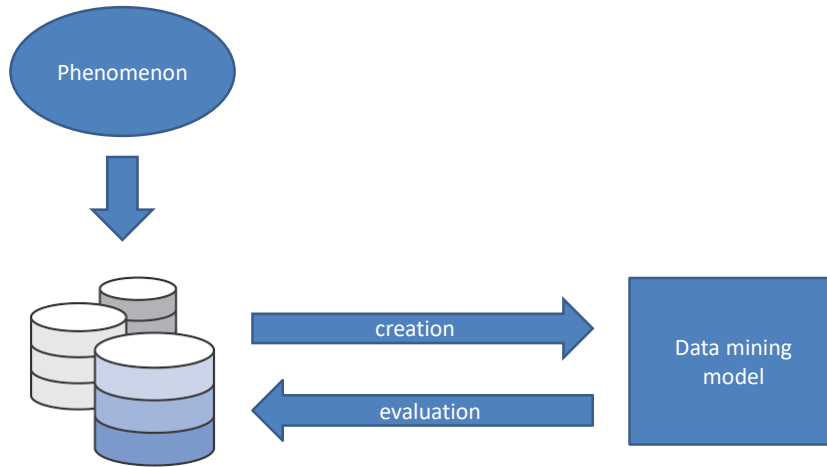
role of DM methods:

- ***phenomena description (prediction, decision making, ...),***
- ***initial step in theory creation***

Paweł Lula, Cracow University of Economics

10

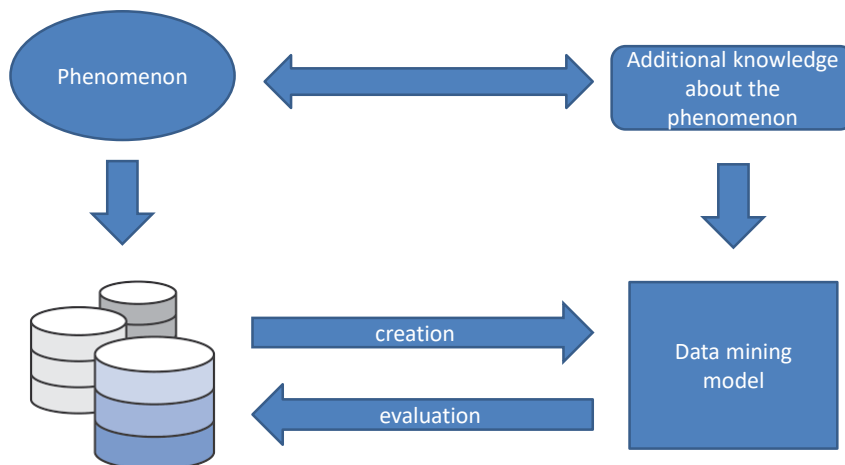
Data-based character of data mining



Paweł Lula, Cracow University of Economics

11

Hybrid models in data mining approach



Paweł Lula, Cracow University of Economics

12

INFORMATION OVERLOAD PROBLEM

Paweł Lula, Cracow University of Economics

13

How much information is there in the world?



Paweł Lula, Cracow University of Economics

14

The Royal Library of Alexandria



The largest and most significant library of the ancient world. The library was conceived and opened either during the reign of Ptolemy I Soter (323–283 BC) or during the reign of his son Ptolemy II (283–246 BC).

At its height, the library held nearly 750,000 scrolls.



Paweł Lula, Cracow University of Economics

15

How much information is there in the world?

Science 1 April 2011:
Vol. 332 no. 6025 pp. 60-65
DOI: 10.1126/science.1200970

RESEARCH ARTICLE

The World's Technological Capacity to Store, Communicate, and Compute Information

Martin Hilbert and Priscila López

Paweł Lula, Cracow University of Economics

16

The total amount of information in the world

	1986	1993	2000	2007
The total amount of information in exabytes (1 EB = 10^{18} B)	2,6	15,8	54,5	295
The total amount of information per person in MB	539	2866	8988	44716
The total amount of information per person (CD equivalent)	1	4	12	61

295 EB = $410 * 10^9$ CD disks = a stack from the Earth to the Moon and a quarter of this distance beyond (with 1.2 mm thickness per CD).

How many books are there in the world?

- Google: 129,864,880 books have ever been published in the entire World (unique different titles)
- The Library of Congress: in the catalogue there are 34,5 million books



How many web sites are there in the Internet?

internet live stats

live

1 second

watch

trends i

Home > Trends and More > Total Number of Websites

Total number of Websites

1,266,201,077

Websites online right now



Information overload



Information overload

Information overload: a situation in which you get more information than you can deal with at one time and become tired and confused.



Paweł Lula, Cracow University of Economics

21

Needle in a haystack

As long as the centuries continue to unfold, the number of books will grow continually, and one can predict that a time will come when it will be almost as difficult to learn anything from books as from the direct study of the whole universe.

It will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes.

–Denis Diderot, "Encyclopédie" (1755)



Paweł Lula, Cracow University of Economics

22

Information overload

- 1964, the „information overload” term was coined by Bertram Gross ...



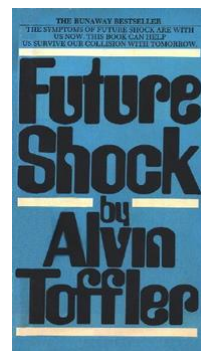
Pawel Lula, Cracow University of Economics

23

Information overload

- ...
- and popularized by Alvin Toffler in his book „Future Shock” (1970)

„too much change in too short a period of time”



Pawel Lula, Cracow University of Economics

24

Can computers solve the information overload problem?



*Computers have promised us a fountain of wisdom
but delivered a flood of data*

A.Piatetsky-Shapiro, 1992

Dealing with the information overload problem

1. Specify your information needs
2. Evaluate information quality
3. Use adequate methods of information storing and processing

Specify your information needs!

- What are your goals?
- What information do you need to achieve them?
 - about your work (study),
 - about tasks which you ought to perform,
 - about people you have around you (clients, competitors),
 - about legal, finance, political, system,
 - about society in which you live.



Paweł Lula, Cracow University of Economics

27

Evaluate the information quality!

Information quality means:

- The right information = which we need to achieve our goals

'MR & MRS' Jamie O'Hara's fiancée flashes engagement ring as they confirm marriage plans

The footballer, 31, and lingerie model, 25, are currently in Ibiza where Jamie popped the question earlier this week. They appear to finally be back on track after breaking up earlier this year.

*Do not read all what is nice to know!
Read what you need to know!*



Paweł Lula, Cracow University of Economics

28

Evaluate the information quality!



Information quality means:

- Information with the right completeness = all data we need!



Pawel Lula, Cracow University of Economics

29

Evaluate the information quality!



Information quality means:

- Information with the right accuracy = must reflect reality!



Pawel Lula, Cracow University of Economics

30

Evaluate the information quality!



Information quality means:

- Information at the right level of generality = not too detailed, not too general!



Pawel Lula, Cracow University of Economics

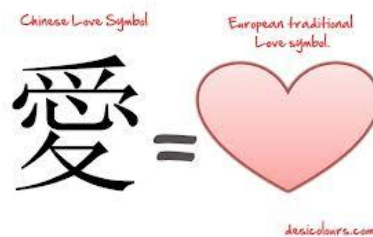
31

Evaluate the information quality!



Information quality means:

- Information in the right format = easy to understand and use!



Pawel Lula, Cracow University of Economics

32

Evaluate the information quality!



Information quality means:

- Information at the right time = not too late, not too earlier, just in time!



Pawel Lula, Cracow University of Economics

33

Evaluate the information quality!



Information quality means:

- Information at the right place = where somebody needs it!



www.shutterstock.com · 54705181

Pawel Lula, Cracow University of Economics

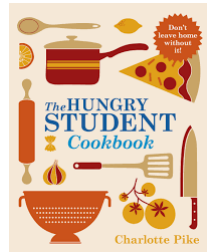
34

Evaluate the information quality!



Information quality means:

- Information for the right purpose = to solve the problem!

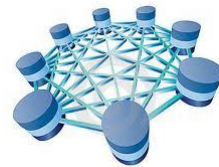


Paweł Lula, Cracow University of Economics

35

Use adequate methods of information storing

- large capacity,
- support for various types of information – heterogeneous information (numbers, texts, multimedia, ...),
- tools for data cleaning,
- possibility of data integration,
- tools for information filtering and searching,
- flexible structure,
- security assurance (availability, integrity and confidentiality).



Paweł Lula, Cracow University of Economics

36

Use adequate methods of information processing

- Goal: knowledge discovery
- Main requirements:
 - support for heterogeneous types of information (numbers, nominal values, texts, audio and video stream, ...)
 - support for complex structures (tables /sometimes dimension and size is very large/, time series, sequences, associate list /maps/, trees, graphs, ...)

	Product A	Product B	Product C	Product D	Product E
Chicago	20M	2M	12M	2M	21M
Cincinnati	30M	4M	10M	8M	20M
Dallas	14M	3M	14M	9M	24M
Louisville	10M	5M	11M	4M	23M

2002
2004



Pawel Lula, Cracow University of Economics

37

Use adequate methods of information processing

- Goal: knowledge discovery
- Main requirements:
 - results are easy to interpret, understand and use,
 - need for aggregation, summarisation,
 - visualization.



Pawel Lula, Cracow University of Economics

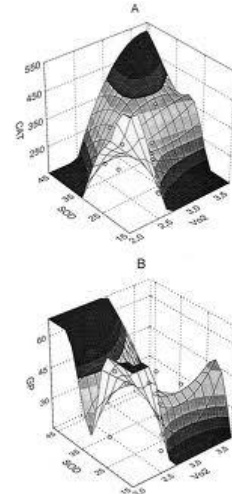
38

Use adequate methods of information processing

- Requirements for methods:
 - nonlinearity,
 - lack of theory /data mining methods/
 - variability in time,
 - short time for processing.



Pawel Lula, Cracow University of Economics



39

Big Data concept

Application-Controlled Demand Paging for Out-of-Core Visualization

Michael Cox
MRJ/NASA Ames Research Center
Microcomputer Research Labs, Intel Corporation
<mbc@nas.nasa.gov>

David Ellsworth
MRJ/NASA Ames Research Center
<ellswor@nas.nasa.gov>

Abstract

In the area of scientific visualization, input data sets are often very large. In visualization of Computational Fluid Dynamics (CFD) in particular, input data sets today can surpass 100 Gbytes, and are expected to scale with the ability of supercomputers to generate them. Some visualization tools already partition large data sets into segments, and load appropriate segments as they are needed. However, this does

1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the *problem of big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources. This *write-a-check* algorithm has two drawbacks. First, if visualization algorithms and tools are

*“Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the **problem of big data**. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources.”*

October 1997, first use of 'big data' term

Pawel Lula, Cracow University of Economics

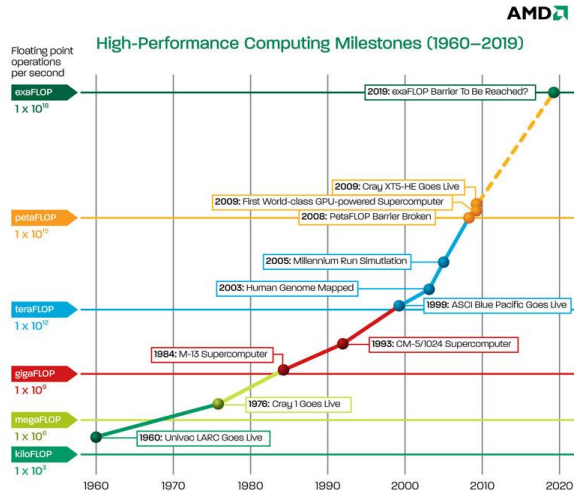
40

Big Data concept

- enormous volume of data
- heterogenous forms, unstructured format
- high pace of data flows

PROGRESS IN CAPABILITIES OF COMPUTER SYSTEMS

Computer performance over time

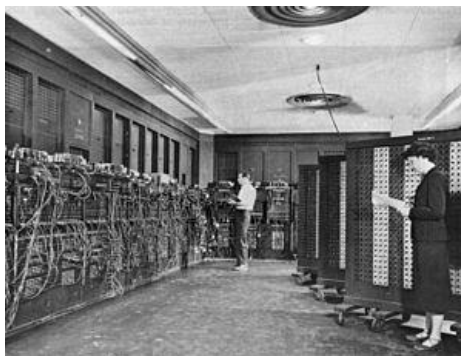


Source: <https://www.cnet.com/news>

Pawel Lula, Cracow University of Economics

43

The best in 1945



ENIAC (1945)

5000 operations per second
= 5000 FLOPS

Pawel Lula, Cracow University of Economics

44

The best in July 2017

Sunway TaihuLight - Sunway MPP, Sunway
SW26010 260C 1.45GHz, Sunway



Site:	National Supercomputing Center in Wuxi
Manufacturer:	NRCCPC
Cores:	10,649,600
Memory:	1,310,720 GB
Processor:	Sunway SW26010 260C 1.45GHz
Interconnect:	Sunway
Performance	
Linpack Performance (Rmax)	93,014.6 TFlop/s
Theoretical Peak (Rpeak)	125,436 TFlop/s
Nmax	12,288,000
HPCG [TFlop/s]	480.8
Power Consumption	
Power:	15,371.00 kW (Submitted)
Power Measurement Level:	2
Software	
Operating System:	Sunway RaiseOS 2.0.5



Source: <https://www.top500.org/>

93 petaFLOPS = 93 * 10¹⁵ FLOPS = 9300000000000000 FLOPS

Paweł Lula, Cracow University of Economics

45

SCOPE OF DATA MINING APPLICATIONS

Paweł Lula, Cracow University of Economics

46

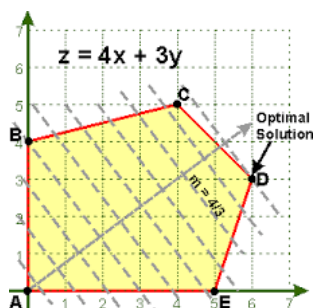
Types of problems

1. optimization problems,
2. regression
3. classification
4. cluster analysis
5. trends
6. network analysis
7. rules identification
8. sequence analysis
9. associate rules
10. text analysis

Paweł Lula, Cracow University of Economics

47

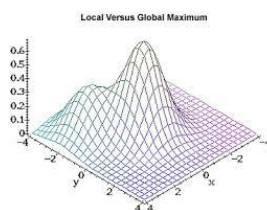
Types of problems: optimization problems



Main goal: finding optimal solution



the shortest path



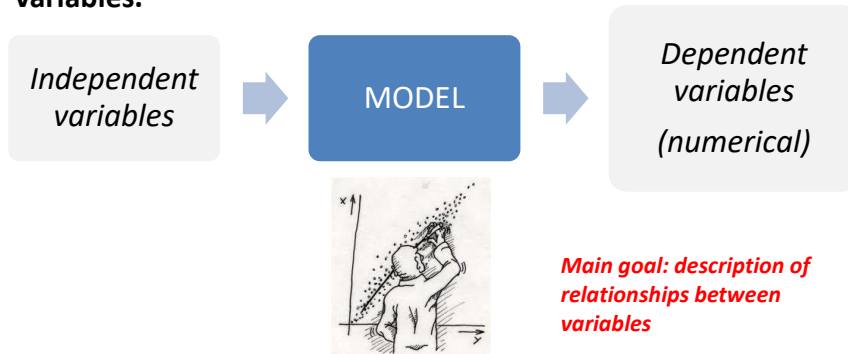
choice problem

Paweł Lula, Cracow University of Economics

48

Types of problems: regression

Regression analysis: techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

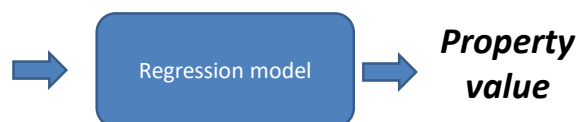


Paweł Lula, Cracow University of Economics

49

Types of problems: regression

- Location
- Size
- Number of bedrooms and bathrooms
- Age
- Garage size (number of cars)
- Central heating system
- Air conditioning system
- Swimming pool



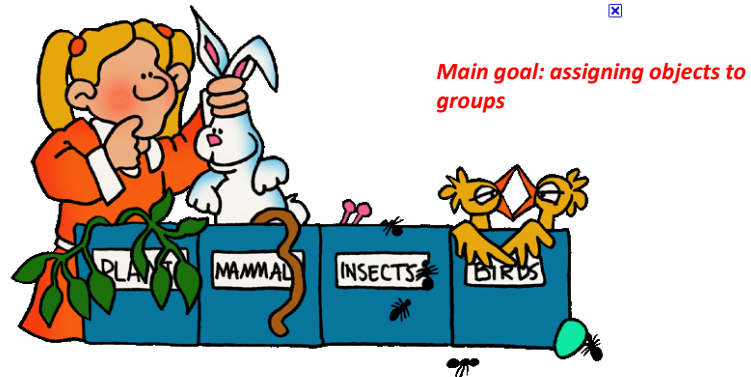
Main goal: description of relationships between variables

Paweł Lula, Cracow University of Economics

50

Types of problems: classification

Classification is the problem of identifying the group to which objects belong.



Pawel Lula, Cracow University of Economics

51

Types of problems: classification

Customer segmentation is the practice of dividing customers into groups of individuals that are similar in specific ways relevant to marketing, such as age, level of income, gender, interests, spending habits, and so on.



Main goal: assigning objects to groups

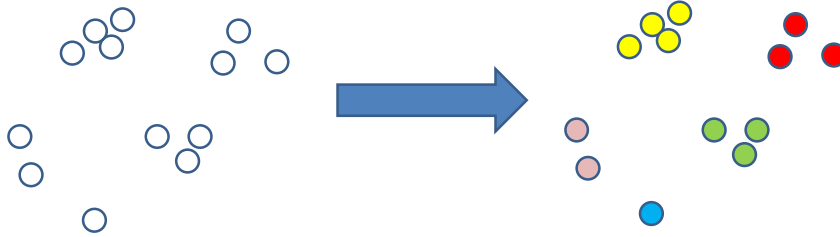
Pawel Lula, Cracow University of Economics

52

Types of problems: cluster analysis

Main goal: analysis the structure of a given set of objects

- Cluster analysis or clustering is a process of grouping of a set of observations into subsets (called clusters).
- A cluster is a group of relatively homogeneous cases or observations. Objects in a cluster are similar to each other. They are also dissimilar to objects outside the cluster, particularly objects in other clusters.



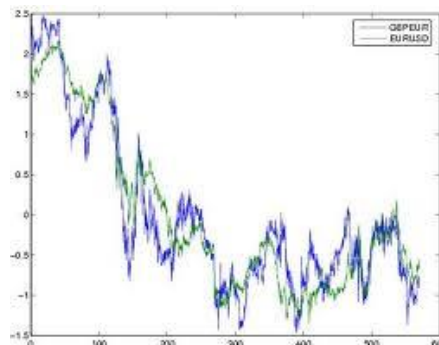
Pawel Lula, Cracow University of Economics

53

Types of problems: trends



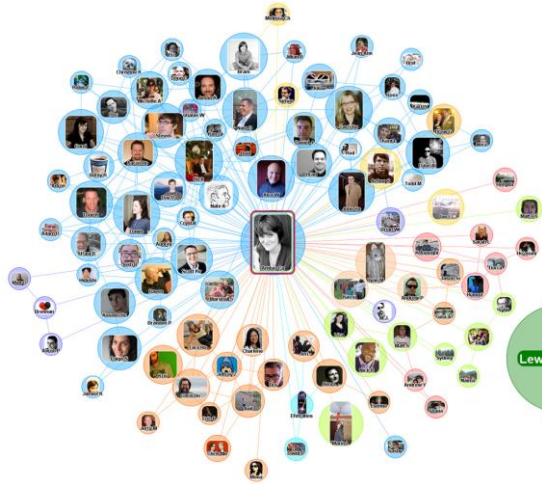
Main goal: description of relationships between consecutive observations



Pawel Lula, Cracow University of Economics

54

Types of problems: network analysis

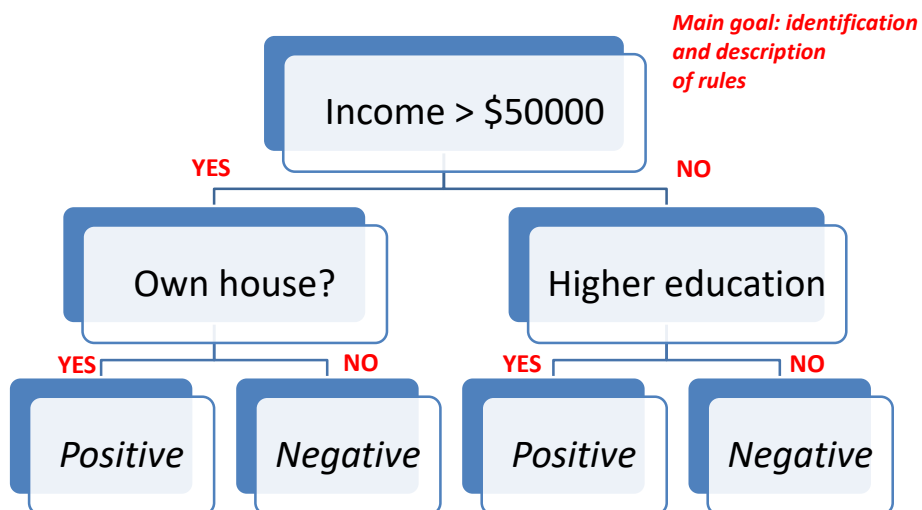


Main goal:
 a) description of relationships between objects,
 b) importance evaluation of objects and links,
 c) analysis the structure of relationships.

Paweł Lula, Cracow University of Economics

55

Types of problems: rules identification

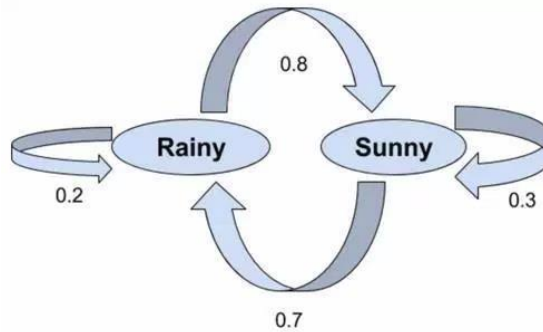


Main goal: identification and description of rules

Paweł Lula, Cracow University of Economics

56

Types of problems: sequence analysis



Main goal: analysis, description and prediction of sequences of events

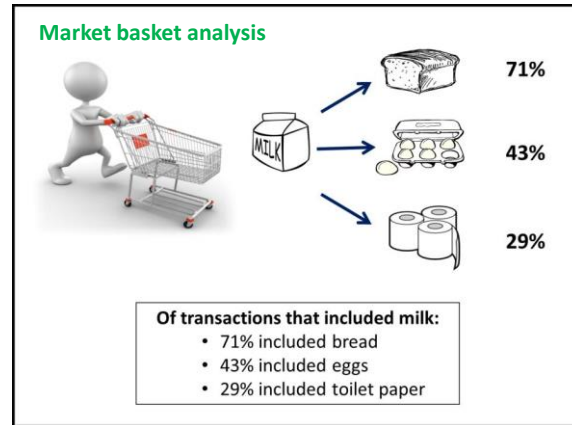
Types of problems: associate rules

Associate analysis: technique that allows to identify how the data items are associated each other



Main goal: analysis of co-occurrence of events

Types of problems: associate rules



Main goal: analysis of co-occurrence of events

Types of problems: text analysis

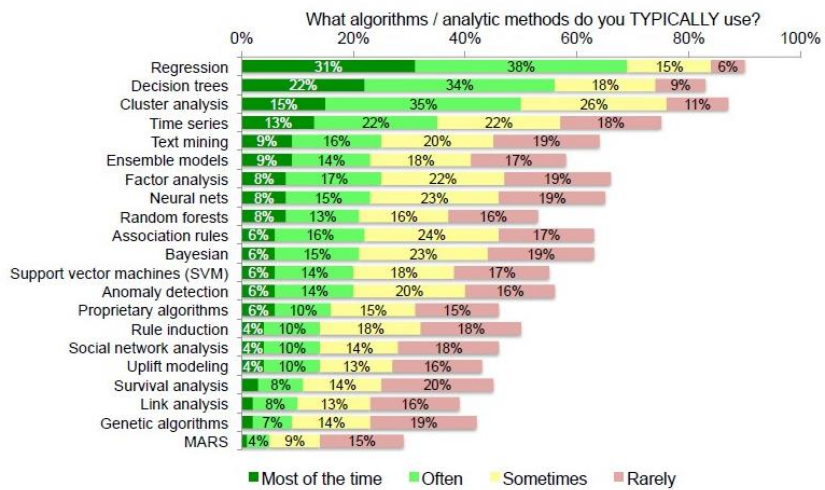


Main goal:

- information retrieval,
- document classification,
- document clustering,
- identification of key-words,
- similarity evaluation.

DATA MINING METHODS

Data mining methods



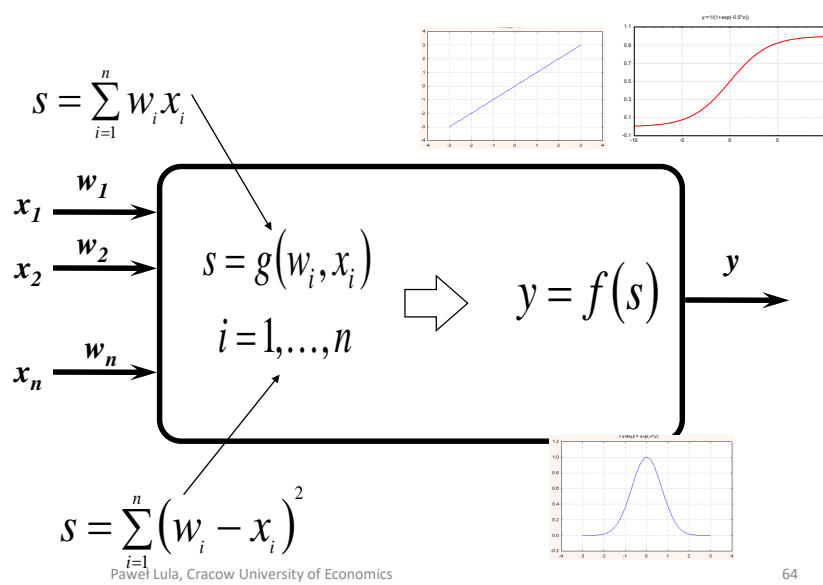
Source: https://gerardnico.com/wiki/data_mining/algorithm

NEURAL NETWORKS MODELS AS AN EXAMPLE OF DATA MINING APPROACH

Paweł Lula, Cracow University of Economics

63

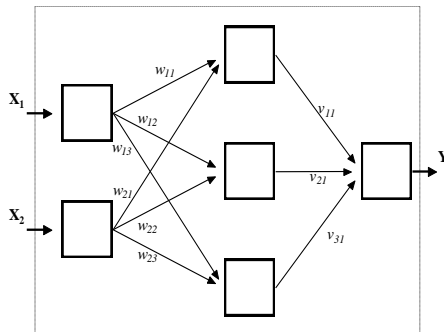
Neuron's model



Paweł Lula, Cracow University of Economics

64

Artificial neural network



- neurons
- layers
 - input
 - output
 - hidden
- connections
- weights
- input values
- output values

Pawel Lula, Cracow University of Economics

65

Applications of NN models: regression

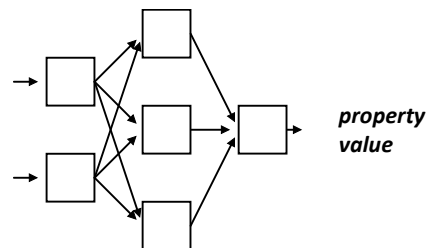


$$Y = NN(X_1, X_2, \dots, X_N)$$

Y – numerical variable

X_i – numerical or nominal variables

- Location
- Size
- Number of bedrooms and bathrooms
- Age
- Garage size (number of cars)
- Central heating system
- Air conditioning system
- Swimming pool



Pawel Lula, Cracow University of Economics

66

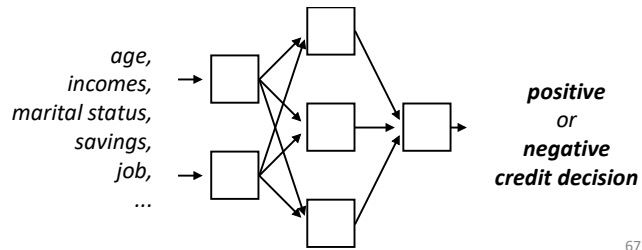
Applications of NN models: classification



$$Y = NN(X_1, X_2, \dots, X_N)$$

Y – nominal variable

X_i – numerical or nominal variables

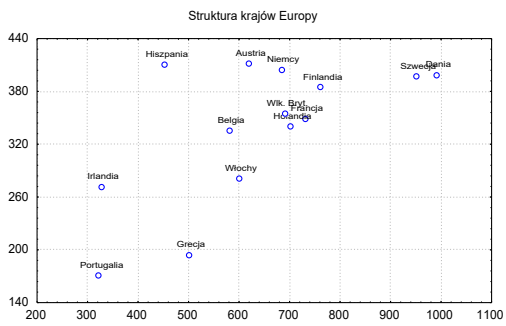


67

Pawel Lula, Cracow University of Economics

Applications of NN models: cluster analysis

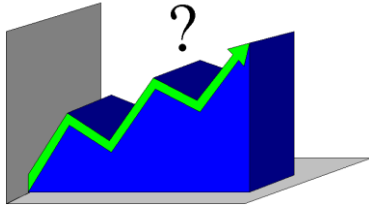
description of European countries



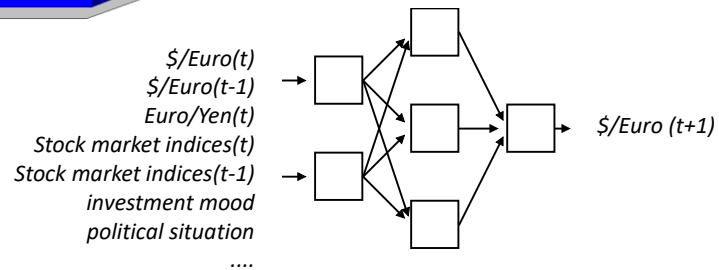
Pawel Lula, Cracow University of Economics

68

Applications of NN models: time series analysis

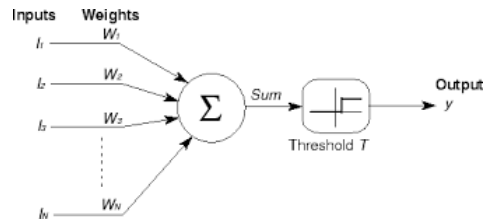
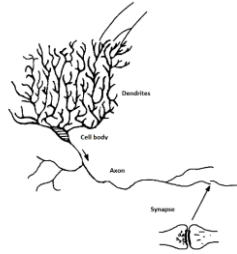


$$Y_{t+1} = NN(Y_t, Y_{t-1}, \dots, Y_{t-k}, X_t, \dots, X_{t-1})$$



THE HISTORY OF NEURAL NETWORK MODELS

1943 – Model of artificial neuron



A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

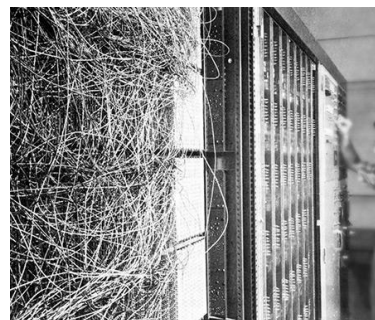
- WARREN S. MCCULLOCH AND WALTER PITTS
University of Illinois, College of Medicine,
Department of Psychiatry at the Illinois Neuropsychiatric Institute,
University of Chicago, Chicago, U.S.A.

Paweł Lula, Cracow University of Economics

71

1958 – Perceptron

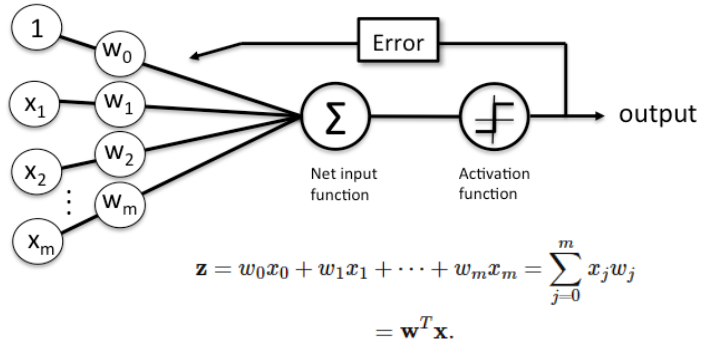
- 1958 – Perceptron is a model for classification tasks proposed by Frank Rosenblatt



Paweł Lula, Cracow University of Economics

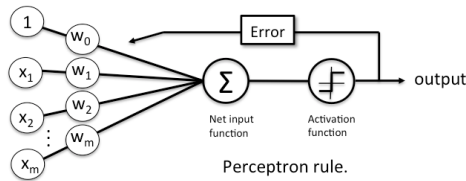
72

1958 – Perceptron



$$g(\mathbf{z}) = \begin{cases} 1 & \text{if } \mathbf{z} \geq 0 \\ -1 & \text{otherwise.} \end{cases} \quad \text{Two classes}$$

Learning process for Perceptron



The structure of data:

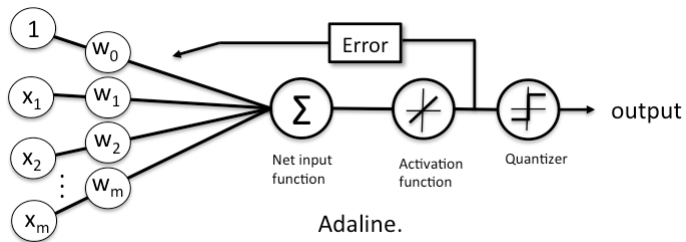
Inputs	Output value	Target value
....	1	1
....	1	-1
....	-1	1
....	-1	-1
....	1	1

Error = target - output

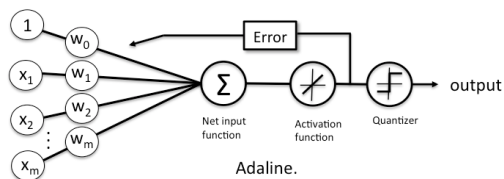
**Goal of learning process:
output class should be
equal to target class!!!**

1960 - ADALINE (Adaptive Linear Neuron)

Proposed by:
Bernard Widrow and Tedd Hoff
at Stanford University, 1960



Learning process for ADALINE neuron



The structure of data:

Inputs	Output value	Target value
....	real value	real value
....	real value	real value
....	real value	real value
....	real value	real value
....	real value	real value

Error = target - output

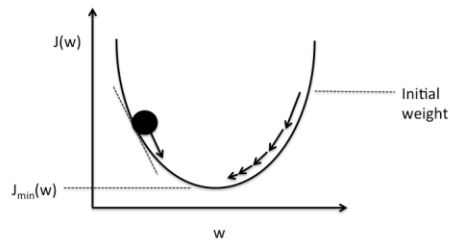
**Goal of learning process:
error minimization!!!**

Learning process for ADALINE neuron

$$J(\mathbf{w}) = \frac{1}{2} \sum_i (\text{target}^{(i)} - \text{output}^{(i)})^2 \quad \text{output}^{(i)} \in \mathbb{R}$$

weights
(parameters)

inputs



$$\Delta \mathbf{w} = -\eta \nabla J(\mathbf{w}),$$

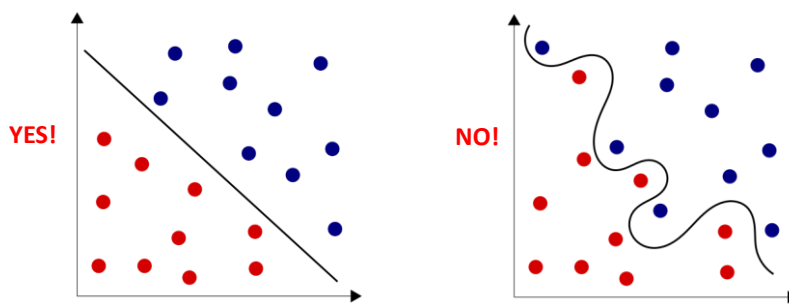
**Gradient descent
method**
(movement into the direction
opposite to the gradient)

$$\Delta w_j = -\eta \frac{\partial J}{\partial w_j}$$

Paweł Lula, Cracow University of Economics

77

Main limitation of ADALINE and Perceptron network



For regression problems:

- equivalent to linear models.

For classification problems:

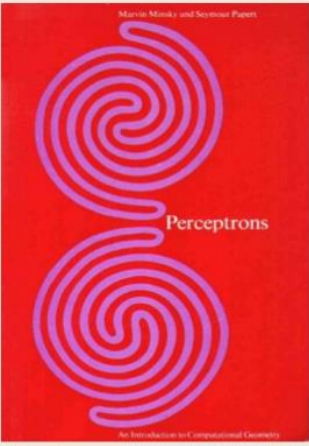
- appropriate only for linear separable problems (linear separability)

Paweł Lula, Cracow University of Economics

78

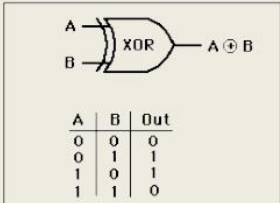
1969 – Minsky and Papert publish the book "Perceptrons"

1969: Perceptrons can't do XOR!



Marvin Minsky and Seymour Papert
Perceptrons
An Introduction to Computational Geometry


<http://www.i-programmer.info/images/stories/BabBag/AI/book.jpg>



A B XOR A ⊕ B

A	B	Out
0	0	0
0	1	1
1	0	1
1	1	0

<http://hyperphysics.phy-astr.gsu.edu/hbase/electronic/ietron/xor.gif>



Minsky & Papert

<https://constructingkids.files.wordpress.com/2013/05/minsky-papert-71-csolomon-x640.jpg>

Paweł Lula, Cracow University of Economics

79

1957 – Kolmogorov theorem on approximation

THEOREM 2.3.1 (Kolmogorov, 1957) Any continuous real-valued functions $f(x_1, x_2, \dots, x_n)$ defined on $[0, 1]^n$, $n \geq 2$, can be represented in the form

$$f(x_1, x_2, \dots, x_n) = \sum_{j=1}^{2n+1} g_j \left[\sum_{i=1}^n \phi_{ij}(x_i) \right] \quad (2.3.1)$$

where the g_j terms are properly chosen continuous functions of one variable, and the ϕ_{ij} functions are continuous monotonically increasing functions independent of f .

Paweł Lula, Cracow University of Economics

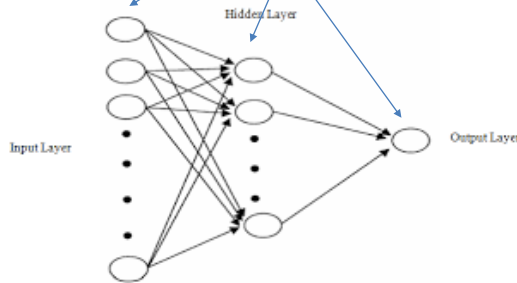
80

Neural network as a realization of Kolmogorov's idea

THEOREM 2.3.1 (Kolmogorov, 1957) Any continuous real-valued functions $f(x_1, x_2, \dots, x_n)$ defined on $[0, 1]^n$, $n \geq 2$, can be represented in the form

$$f(x_1, x_2, \dots, x_n) = \sum_{j=1}^{2n+1} g_j \left[\sum_{i=1}^n \phi_{ij}(x_i) \right] \quad (2.3.1)$$

where the g_j terms are properly chosen continuous functions of one variable, and the ϕ_{ij} functions are continuous monotonically increasing functions independent of f .



1989 – Cybenko theorem on approximation

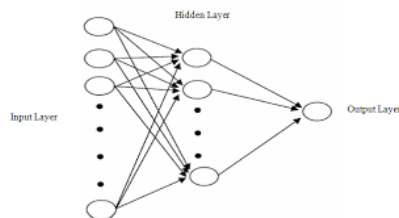
THEOREM 2.3.2 (Cybenko, 1989) Let φ be any continuous sigmoid-type function [e.g., $\varphi(\xi) = 1/(1 + e^{-\xi})$]. Then, given any continuous real-valued function f on $[0, 1]^n$ (or any other compact subset of R^n) and $\varepsilon > 0$, there exists vectors w_1, w_2, \dots, w_N , α , and θ and a parameterized function $G(\cdot, w, \alpha, \theta): [0, 1]^n \rightarrow R$ such that

$$|G(x, w, \alpha, \theta) - f(x)| < \varepsilon \quad \text{for all } x \in [0, 1]^n$$

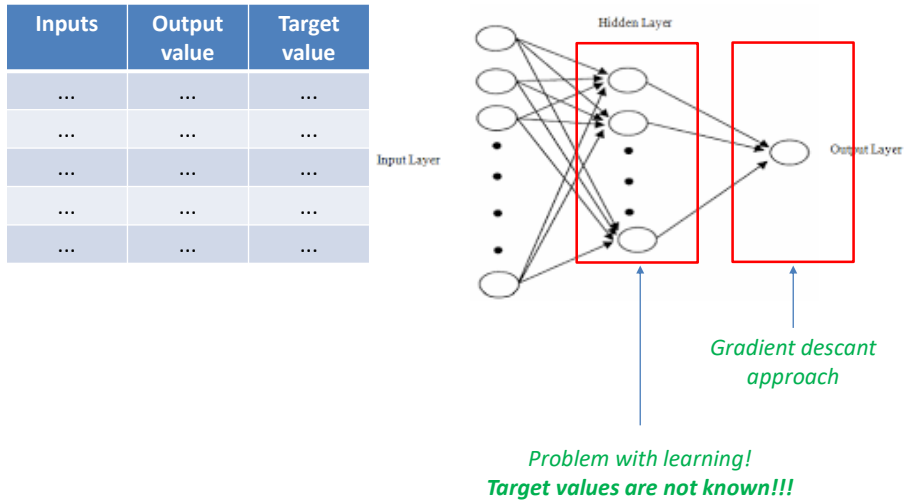
where

$$G(x, w, \alpha, \theta) = \sum_{i=1}^N \alpha_i \varphi(w_i^T x + \theta_i) \quad (2.3.2)$$

and $w_j \in R^n$, $\alpha_j, \theta_j \in R$, $w = (w_1, w_2, \dots, w_N)$, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)$, and $\theta = (\theta_1, \theta_2, \dots, \theta_N)$.



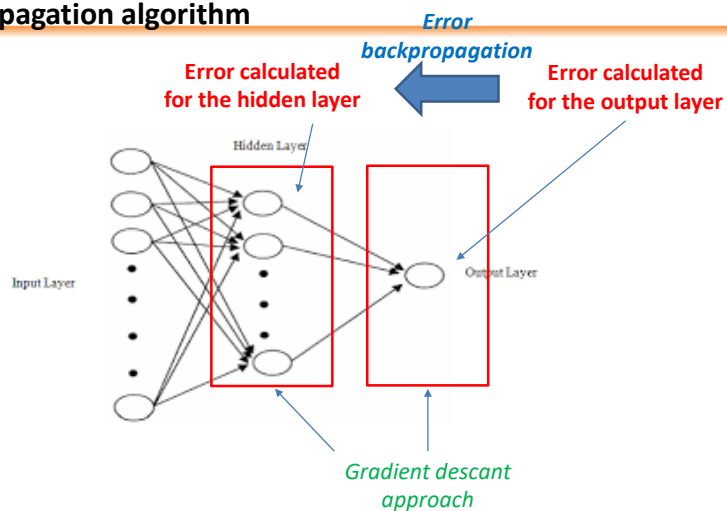
Problem with learning of three-layer network



Paweł Lula, Cracow University of Economics

83

Backpropagation algorithm



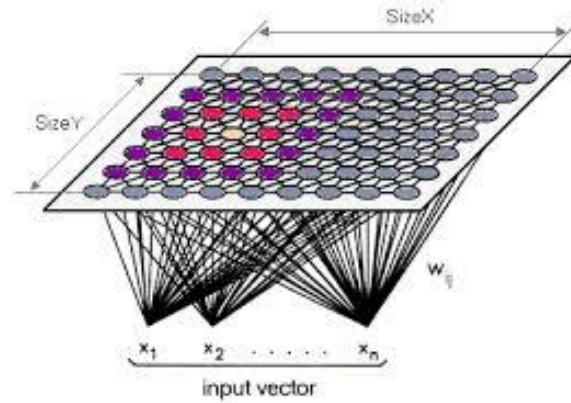
Backpropagation algorithms was proposed two times:

- 1974 - Paul Werbos,
- 1986 - David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams.

Paweł Lula, Cracow University of Economics

84

1982 – Kohonen network



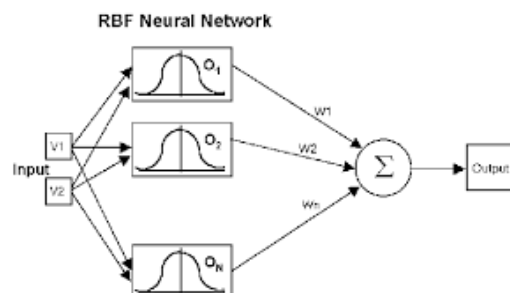
- Learned in unsupervised mode (target values are not known)
- Kohonen network is appropriate for cluster analysis.

Pawel Lula, Cracow University of Economics

85

1988 – Radial basis function network (RBF network)

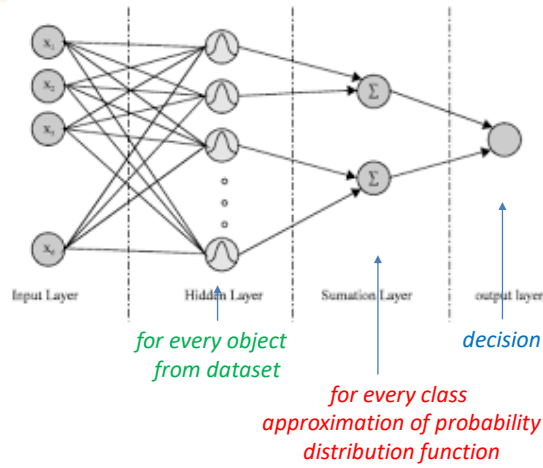
- D.S. Broomhead and D. Lowe. Multivariate functional interpolation and adaptive networks. Complex Systems, 2:321-355, 1988.



Pawel Lula, Cracow University of Economics

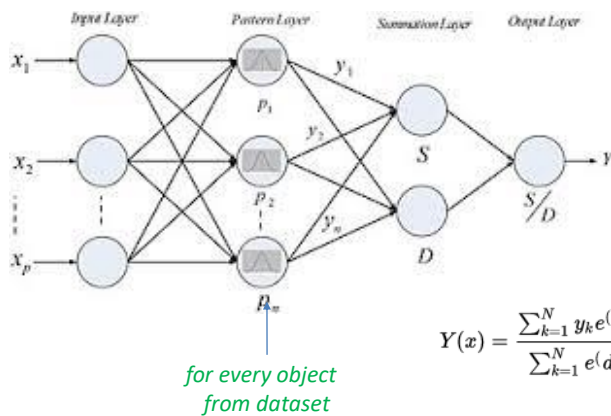
86

1990 – Probabilistic neural networks (for classification problems)



1990 - Speckt, D.F. (1990). Probabilistic Neural Networks. Neural Networks 3 (1), 109-118

1991 – Generalized regression neural network



$$Y(x) = \frac{\sum_{k=1}^N y_k e^{-(d_k/2\sigma)} }{\sum_{k=1}^N e^{-(d_k/2\sigma)}}$$

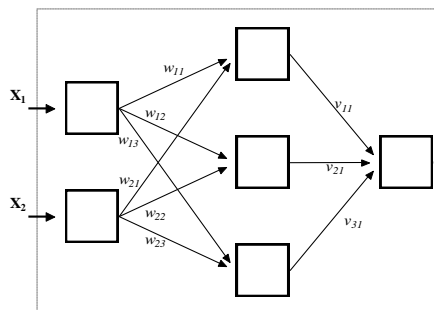
1991 - Speckt, D.F. (1991). A Generalized Regression Neural Network. IEEE Transactions on Neural Networks 2 (6), 568-576

REQUIREMENTS FOR GOOD NEURAL NETWORK MODEL

Paweł Lula, Cracow University of Economics

89

Requirements for good neural model



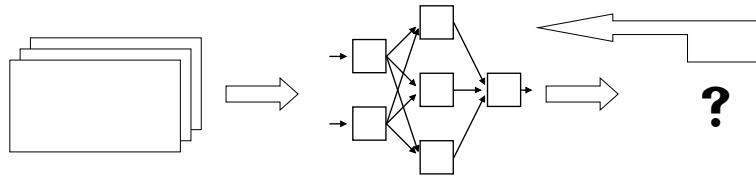
- proper neuron model
- proper values of neuron weights
- proper structure of neural model

Neural network model can store a knowledge on studied phenomenon
The network knowledge is stored in its weights.

Paweł Lula, Cracow University of Economics

90

The learning process



- Finding proper values of network's weights is the main goal of learning process
- The learning process is based on data.
- Learning algorithm – a method which is used for weights' adjustment.

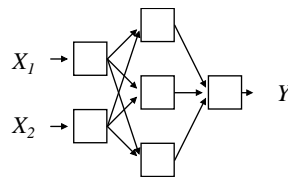
Pawel Lula, Cracow University of Economics

91

Two types of learning: *Learning with a Teacher* (supervised learning)

Training set:

X_1	X_2	D
x_{11}	x_{12}	d_1
x_{21}	x_{22}	d_2
...
x_{n1}	x_{n2}	d_n



*Does y_i is equal to d_i ?
if not than adjust
to decrease the
difference between
 y_i and d_i .*

Epoch – one complete presentation of the entire training set.

The goal of learning process:
output values (y_i) should be equal to (or very close to) desired (target) values (d_i) from training set.

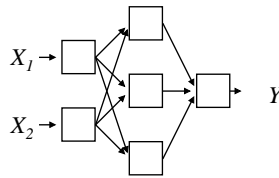
Pawel Lula, Cracow University of Economics

92

Two types of learning: *Learning without a Teacher* (*unsupervised learning*)

Training set:

X_1	X_2
x_{11}	x_{12}
x_{21}	x_{22}
...	...
x_{n1}	x_{n2}



Does a network achieve the equilibrium state?

Epoch – one complete presentation of the entire training set.

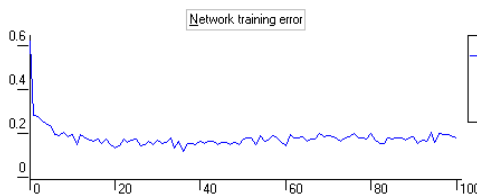
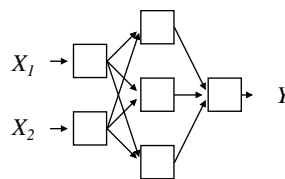
The goal of learning process: Achieving the equilibrium state.

Learning with a Teacher

Training set:

X_1	X_2	D
x_{11}	x_{12}	d_1
x_{21}	x_{22}	d_2
...
x_{n1}	x_{n2}	d_n

$$SSE = \sum_{i=1}^N (d_i - y_i)^2$$



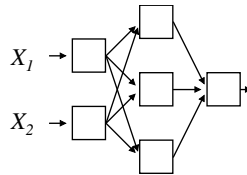
Learning process = training error minimizing.

Learning process is based **on training set**.

Overtraining – lack of ability to generalization

Training set:

X_1	X_2	D
x_{11}	x_{12}	d_1
x_{21}	x_{22}	d_2
...
x_{k1}	x_{k2}	d_k



Ability to approximation:

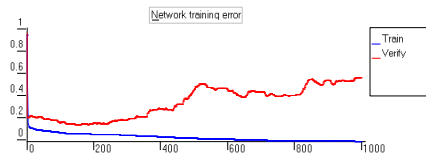
$$SSE_U = \sum_{i=1}^k (d_i - y_i)^2$$

Ability to generalization:

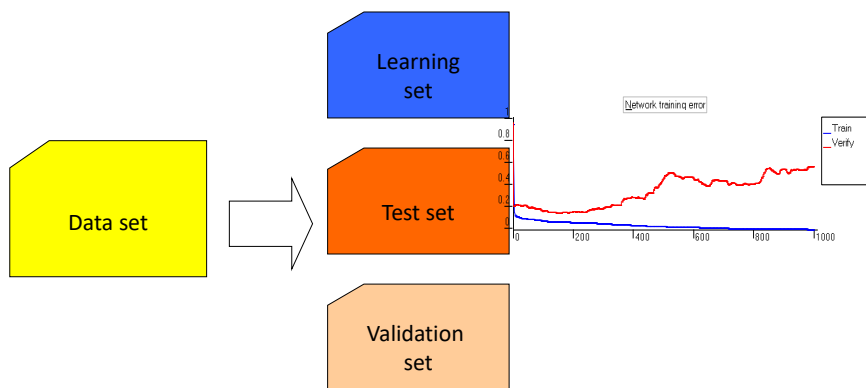
$$SSE_W = \sum_{i=k+1}^n (d_i - y_i)^2$$

Test set:

X_1	X_2	D
$x_{(k+1)1}$	$x_{(k+1)2}$	d_{k+1}
...
x_{n1}	x_{n2}	d_n

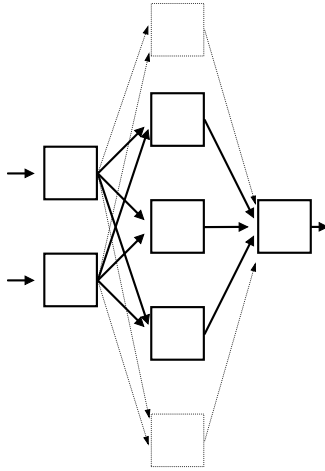


Data sets used during network building process



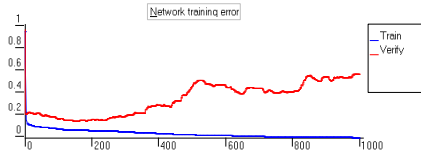
- **Learning set** – a basis for weights' modification,
- **Test set** – a basis for overtraining identification (during learning process),
- **Validation set** – a basis for final evaluation of network model (after learning process).

The effect of network's structure on its quality



Simple network's structure:

- lack of ability to describe complex relationships,
- no problems with learning process



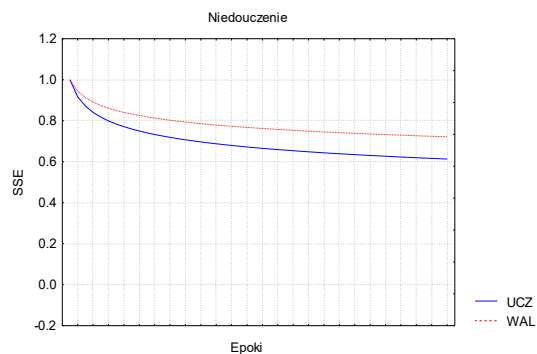
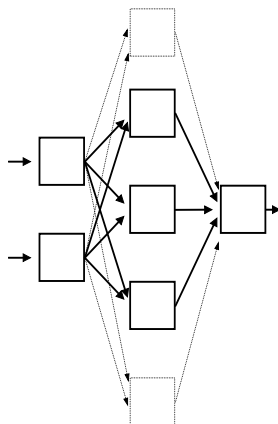
Complex network's structure:

- very good quality for learning set,
- (sometimes) poor quality for test set - overtraining effect,
- problems with learning process.

Paweł Lula, Cracow University of Economics

97

The effect of network's structure on its quality

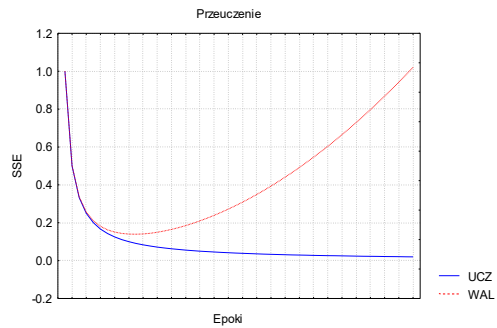
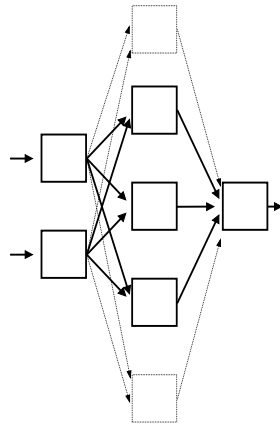


Network structure is too simple!

Paweł Lula, Cracow University of Economics

98

The effect of network's structure on its quality

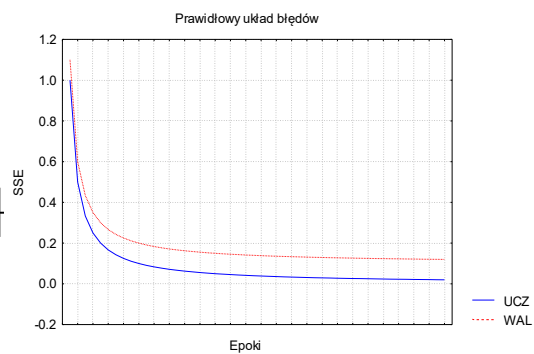
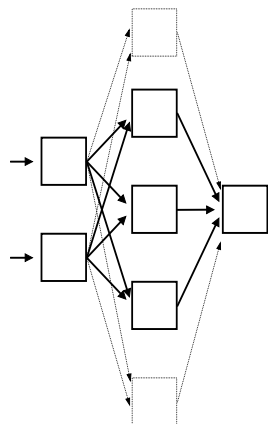


Network structure is too complex!

Pawel Lula, Cracow University of Economics

99

The effect of network's structure on its quality



Network structure is OK!

Compromise between ability to approximation and ability to generalization.

Pawel Lula, Cracow University of Economics

100

TEXT MINING

Text Mining

- **Text Mining** is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.

Marti A. Hearst, 1999



Frequency matrix as a starting point for further analysis

- Pieces of information are represented by words,
- Stages:
 - cutting text into words,
 - removing irrelevant words from documents (*stop-list* = a collection of irrelevant words)
 - transformation to the base form
 - calculation of word occurrence frequencies,
 - forming frequency matrix

$$\begin{array}{c} \text{words} \end{array} \begin{array}{c} \text{documents} \\ \left[\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{array} \right] \end{array}$$

Transformation of frequency matrix

- Binary representation

$$\mathbf{X} = \begin{bmatrix} 2 & 0 & 4 & \dots & 4 \\ 1 & 0 & 3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 2 & \dots & 1 \end{bmatrix} \xrightarrow{\text{red arrow}} \mathbf{X}^{\text{bin}} = \begin{bmatrix} 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 1 & \dots & 1 \end{bmatrix}$$

Transformation of frequency matrix

- Logarithmic representation

$$\mathbf{X} = \begin{bmatrix} 2 & 0 & \dots & 4 \\ 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 1 & \dots & 2 \end{bmatrix} \rightarrow \mathbf{X}^{\log} = \begin{bmatrix} 1,301 & 0,000 & \dots & 1,602 \\ 1,000 & 0,000 & \dots & 0,000 \\ \dots & \dots & \dots & \dots \\ 0,000 & 1,000 & \dots & 1,301 \end{bmatrix}$$

$$x_{ij} \rightarrow 1 + \log(x_{ij})$$

- Weighted logarithmic representation (*TFIDF* model: TF – term frequency, IDF – inverse document frequency)

$$x_{ij} \rightarrow (1 + \log(x_{ij})) * \log(N/df_i)$$

N – number of documents

df_i – number of documents containing word f_i

(Transformed) frequency matrix

$$\begin{array}{c} \text{documents} \\ \left[\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{array} \right] \end{array} \xrightarrow{\text{Data mining methods}}$$

words

Distance between documents and between words

$$\begin{array}{c}
 \text{documents} \\
 \left[\begin{array}{cccc}
 x_{11} & x_{12} & \dots & x_{1m} \\
 x_{21} & x_{22} & \dots & x_{2m} \\
 \dots & \dots & \dots & \dots \\
 x_{n1} & x_{n2} & \dots & x_{nm}
 \end{array} \right]
 \end{array}$$

words

Distance between documents =
distance between columns

Distance between words =
distance between rows

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad \text{Euclidean distance}$$

$$d(x, y) = 1 - \frac{\sum_{k=1}^n x_k y_k}{\sqrt{\sum_{k=1}^n x_k^2 \sum_{k=1}^n y_k^2}}$$


Cosine distance

$$d(x, y) = \sum_{k=1}^n |x_k - y_k| \quad \text{Manhattan distance}$$


EXAMPLE 1: ANALYSIS OF RESEARCH PROJECTS CONDUCTED AT THE CRACOW UNIVERSITY OF ECONOMICS

Latent Dirichlet Allocation (LDA) (Blei et al. 2003)

Documents



↓




Topics

Latent Dirichlet Allocation
– completely **unsupervised**
method of topics
identification.

Pawel Lula, Cracow University of Economics 109

Latent Dirichlet Allocation (LDA) (Blei et al. 2003)

Documents



↓

Topic 1

$word_i$

.....

$word_j$

$word_k$

.....

$word_l$

$word_m$

.....

$word_n$

Topic 2

$word_i$

.....

$word_j$

$word_k$

.....

$word_l$

$word_m$

.....

$word_n$

Topic 3

$word_i$

.....

$word_j$

$word_k$

.....

$word_l$

$word_m$

.....

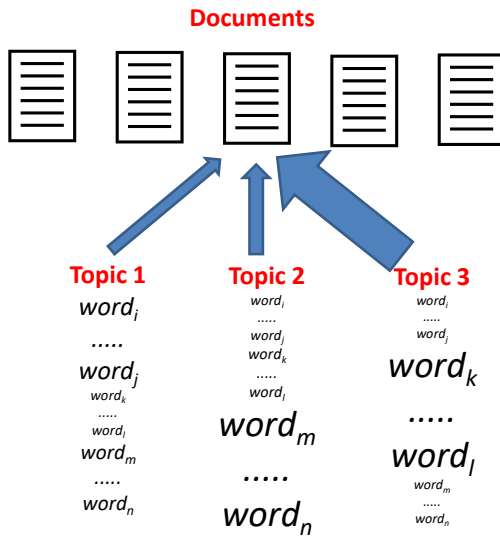
$word_n$

Latent Dirichlet Allocation
– completely **unsupervised**
method of topics
identification.

*Topics are described in
terms of discrete
probabilities over words.*

Pawel Lula, Cracow University of Economics 110

Latent Dirichlet Allocation (LDA) (Blei et al. 2003)



Latent Dirichlet Allocation
 – completely **unsupervised**
 method of topics
 identification.

Topics are described in
 terms of discrete
 probabilities over words.

Each document can be
 modeled as a mixture of
 topics. Documents are
 describes in terms of
 discrete probabilities over
 topics.

Research projects at CUE in 2014

- 80 projects:
 - Faculty of Economics and International Relations (24 projects),
 - Faculty of Finance (17 projects),
 - Faculty of Commodity Science (11 projects),
 - Faculty of Management (28 projects).

Topics identified during analysis

Words with highest probability in topics		
Topic 1	Topic 2	Topic 3
rozwój (development), gospodarczy (economic), analiza (analysis), proces (process), wpływ (impact), gospodarka (economy), kraj (country), ekonomiczny (economic), rynek (market), polski (Polish).	badanie (study, proving), finansowy (financial), zakres (scope), cel (goal), rachunkowość (accounting), wynik (result), publiczny (public), zmiana (change), sprawozdanie (report), analiza (analysis).	zarządzanie (management), analiza (analysis), system (system), badanie (study, proving), przedsiębiorstwo (enterprise), efektywność (efficiency), organizacja (organization), rozwój (development), ocena (assessment), metoda (method).

Macroeconomics

**Finance and
accounting**

Management

Paweł Lula, Cracow University of Economics

113

Topic 1: Macroeconomics



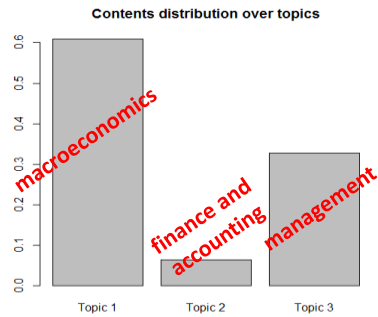
**Topic 1:
Macroeconomics**

Paweł Lula, Cracow University of Economics

114

The analysis of an exemplary project description

*Przegląd i analiza współczesnych procesów globalizacji w wymiarze społecznym i gospodarczym, a w szczególności zagadnień: wzrostu gospodarczego; rynku pracy (zwłaszcza wśród ludzi młodych oraz starszych); ubóstwa; bezpieczeństwa energetycznego na świecie.
Opracowanie naukowe poświęcone zagadnieniom globalizacji we wskazanym zakresie.*

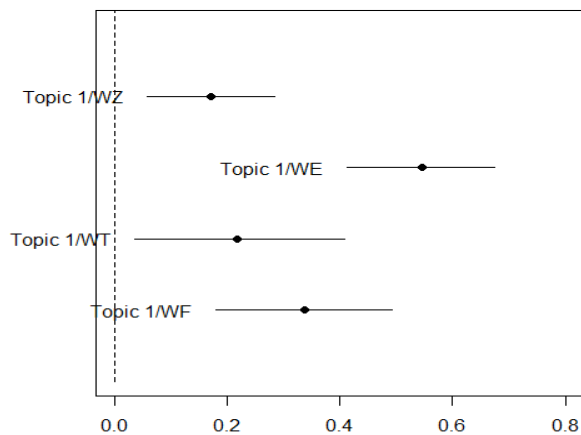


Pawel Lula, Cracow University of Economics

117

Distribution of research topics over CUE faculties.

Topic 1 distribution over CUE faculties

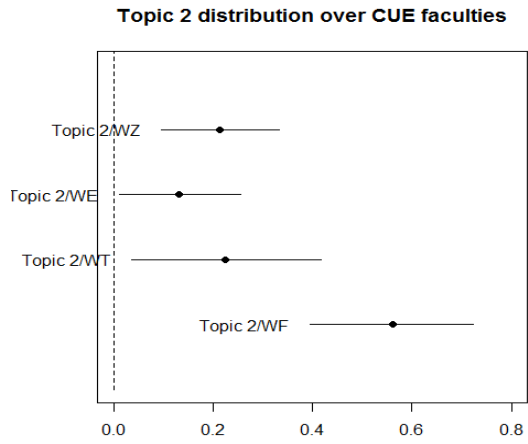


**Topic 1:
Macroeconomics**

Pawel Lula, Cracow University of Economics

118

Distribution of research topics over CUE faculties.

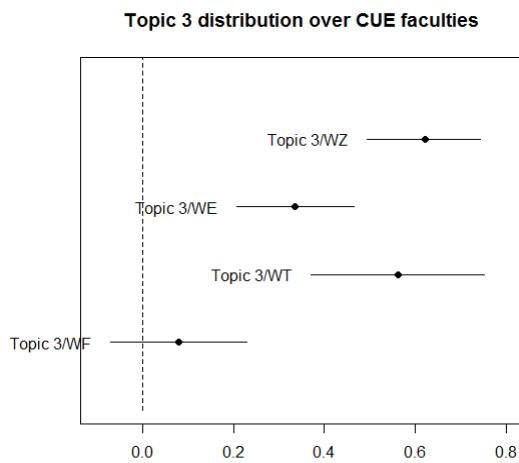


**Topic 2:
Finance and
accounting**

Pawel Lula, Cracow University of Economics

119

Distribution of research topics over CUE faculties.



**Topic 3:
Management**

Pawel Lula, Cracow University of Economics

120

EXAMPLE 2: CONSUMER OPINION MINING

Consumer opinion mining

- Opinions about hotels in London
- Source: <http://kavita-ganesan.com/opinosis-opinion-dataset>

Labels

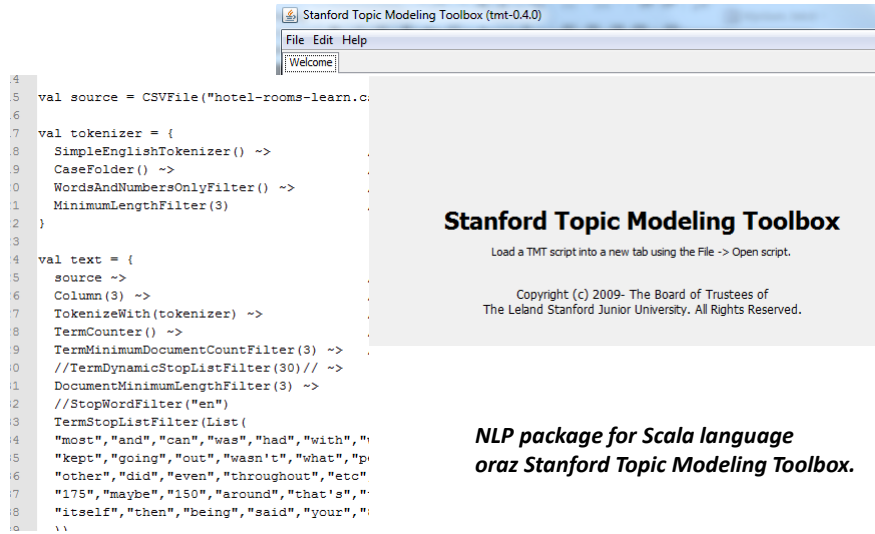
Labels – positive or negative opinion about features listed below:

- general
- size
- staff
- bed
- clean
- equip
- readiness
- location
- quiet
- comfort
- light
- internet
- price
- temperature
- bathroom
- food
- view
- secure
- decor
- reservation

Data set

	A	B
1	LABELS	DESCRIPTIONS
2	staff-neg	We arrived at 23,30 hours and they could not recommend a restaurant so we decided to go to Tesco, with very limited choices but when you are hungry you do not careNext day they rang the bell at 8,00 hours to clean the room, not being very nice being waken up so earlyEvery day
3	size-pos	We had a room with two double beds which was surprisingly roomy, considering the small hotel rooms I have in previous trips to London .
4	staff-pos clean-pos bed-pos	The room was quiet, clean, the bed and pillows were comfortable, and the service was
5	readiness-pos	We arrived about 11 am, room was ready .
6	size-pos clean-pos	Room was good size for Europe , clean throughout .
7	staff-pos	The Concierge desk called our room to ask if we needed any information or assistance .
8	size-pos clean-pos bed-pos	Room was plenty big enough and clean and tidy, bed was comfortable .
9	equip-neg	First, we walked in and the restroom door was broken .
10	clean-pos	Our room was typical holiday inn the bathroom could have done with updating but was
11	readiness-neg	Our rooms were not ready, we were promised rooms at a later time, etc .
12	size-pos	My room was positively huge by European standards .

Topic model building



```

4 val source = CSVFile("hotel-rooms-learn.c
5
6
7 val tokenizer = {
8   SimpleEnglishTokenizer() ~>
9   CaseFolder() ~>
10  WordsAndNumbersOnlyFilter() ~>
11  MinimumLengthFilter(3)
12 }
13
14 val text = {
15   source ~>
16   Column(3) ~>
17   TokenizeWith(tokenizer) ~>
18   TermCounter() ~>
19   TermMinimumDocumentCountFilter(3) ~>
20   //TermDynamicStopListFilter(30) // ~>
21   DocumentMinimumLengthFilter(3) ~>
22   //StopWordFilter("en")
23   TermStopListFilter(List(
24     "most", "and", "can", "was", "had", "with", "i
25     "kept", "going", "out", "wasn't", "what", "p
26     "other", "did", "even", "throughout", "etc"
27     "175", "maybe", "150", "around", "that's", "
28     "itself", "then", "being", "said", "your", "i
29     \
30   )

```

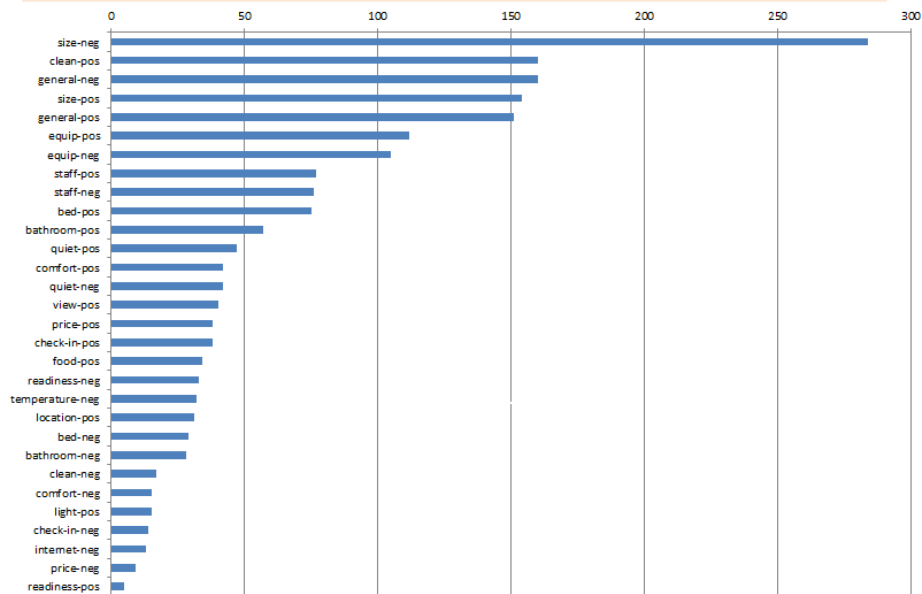
Stanford Topic Modeling Toolbox

Load a TMT script into a new tab using the File -> Open script.

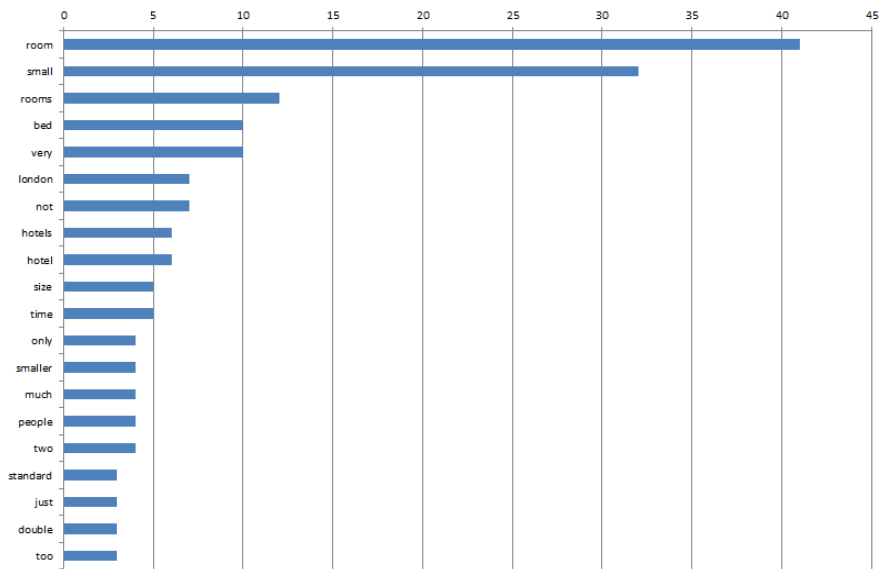
Copyright (c) 2009- The Board of Trustees of The Leland Stanford Junior University. All Rights Reserved.

NLP package for Scala language oraz Stanford Topic Modeling Toolbox.

Topics and their importance



Frequencies of words for topic 'size-neg'



Pawel Lula, Cracow University of Economics

127

Model execution

Opinion:

The bathroom is a good size .

Label assigned by customer:

bathroom-pos

Label predicted by the model:

bathroom-pos (1,0)

Pawel Lula, Cracow University of Economics

128

Model execution

Opinion:

When we tried to use a phone card from our room it would not work so I asked the front desk to help me and was told they couldn't really !

Label assigned by customer:

staff-neg

Label predicted by the model:

staff-neg (1,00)

Model execution

Opinion:

The hotel room was very clean and the cleaning staff and breakfast staff were very attentive .

Labels assigned by customer:

clean-pos

staff-pos

Labels predicted by the model:

staff-pos (0,7)

clean-pos (0,3)

SOFTWARE FOR DATA MINING ANALYSIS

R - <http://www.r-project.org/>

Bezpieczna | <https://www.r-project.org>



[\[Home\]](#)

Download

[CRAN](#)

R Project

[About R](#)

[Logo](#)

[Contributors](#)

The R Project for Statistical Computing

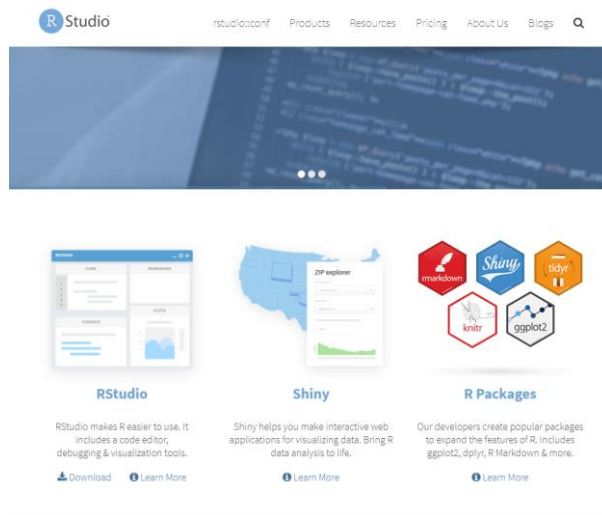
Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

<https://www.rstudio.com/>



RStudio

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

[Download](#) [Learn More](#)

Shiny

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

[Learn More](#)

R Packages

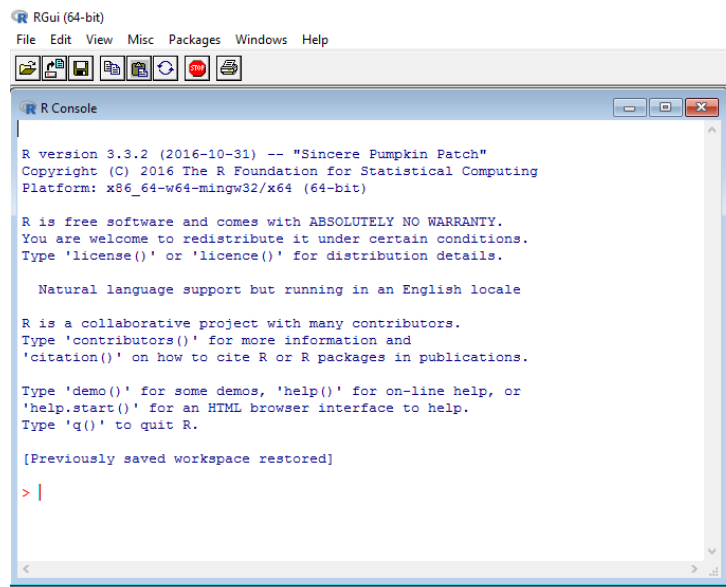
Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

[Learn More](#)

Paweł Lula, Cracow University of Economics

133

R Console



```
RGui (64-bit)
File Edit View Misc Packages Windows Help
[Icons]

R Console
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]
> |
```

Paweł Lula, Cracow University of Economics

134

Thank you for your attention!