Marketing Information Systems: part 4

Course code: PV250

Dalia Kriksciuniene, PhD

Faculty of Informatics, Lasaris lab., ERCIM research program

Autumn, 2012

Timetable

Part 1: Oct.22 Mon 14:00–17:50 C525

Part 2: Oct.23 Tue 8:00–11:50 G101

Part 3: Nov. 05 Mon 14:00–17:50 C525

Part 4: Nov. 05 Tue 8:00–11:50 G101

Part 5: Dec.10 Mon 14:00–17:50 C525

Part 6: Dec.11 Tue 8:00–11:50 G101

Assessment session: 1-2nd week of January

Syllabus 3

Management processes of the marketing manager. Information supply for their performance:

- analytical and control applications:
- pivot tools,
- dashboards
- computational intelligence methods for marketing

Tools &software: *MS Excel* pivot module, *Statistica* advanced models, *Viscovery SoMine* trial

Computational methods for marketing

- Business intelligence: analytical reporting (pivoting)
- Statistical methods: probabilistic
- Artificial intelligence: directed learning:
  - Neural networks NN
  - Memory-Based Reasoning MBR
  - Survival analysis
- Artificial intelligence: undirected learning:
  - Segmentation
  - Clustering
  - Association rules
- Fuzzy inference (possibilities, natural language reasoning)
- Web data mining

Data Mining Techniques Applications

- **Marketing** – Predictive DM techniques, like artificial neural networks (ANN), have been used for ***target marketing*** including market segmentation.
- **Direct marketing** – customers are likely to respond to new products based on their previous consumer behavior.
- **Retail** – DM methods have likewise been used for ***sales forecasting***.
- **Market basket analysis** – uncover which products are likely to be purchased together.

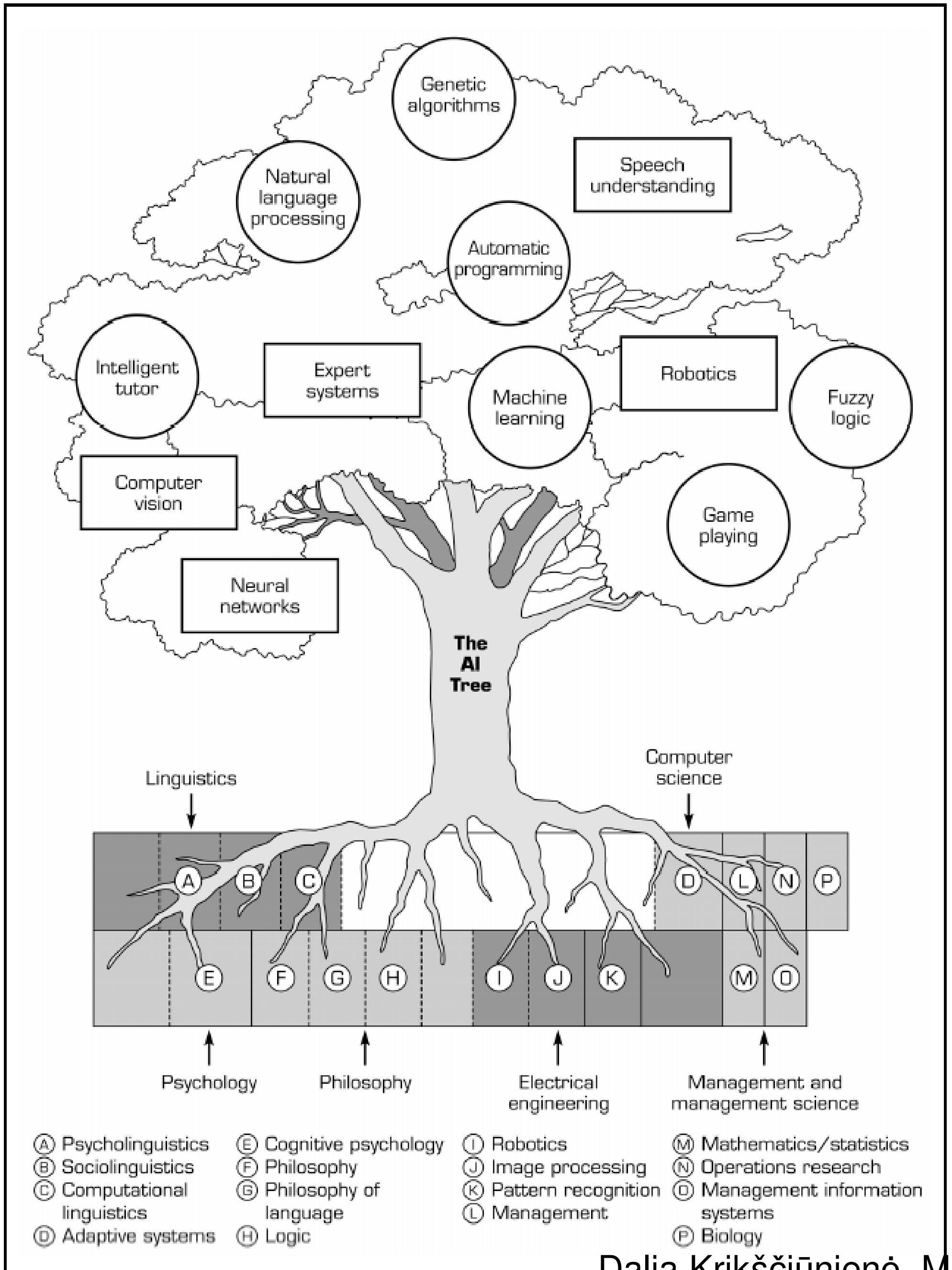


FIGURE 12.2 The Disciplines and Applications of AI

Artificial intelligence (AI): The subfield of computer science concerned with symbolic reasoning and problem solving

Characteristics of artificial intelligence

Symbolic processing (versus Numeric)

Heuristic (versus algorithmic)

Inferencing

Machine learning

- **Heuristics**

  Informal, judgmental knowledge of an application area that constitutes the “rules of good judgment” in the field. Heuristics also encompasses the knowledge of how to solve problems efficiently and effectively, how to plan steps in solving a complex problem, how to improve performance, and so forth.

It can be transferred as tacit knowledge

Marketing activities are heuristic to high extent

Characteristics of artificial intelligence

Inferencing

Reasoning capabilities that can build higher-level knowledge from existing heuristics

Expert knowledge and experience capturing

Machine learning

Learning capabilities that allow systems to adjust their behavior and react to changes in the outside environment

Designing the Knowledge Discovery System

1. Business Understanding – To obtain the highest benefit from data mining, there must be a clear statement of the business objectives.
2. Data Understanding – Knowing the data well can permit the designer to tailor the algorithm or tools used for data mining to his/her specific problem.
3. Data Preparation – Data selection, variable construction and transformation, integration, and formatting
4. Model building and validation – Building an accurate model is a trial and error process. The process often requires the data mining specialist to iteratively try several options, until the best model emerges.
5. Evaluation and interpretation – Once the model is determined, the validation dataset is fed through the model.
6. Deployment – Involves implementing the 'live' model within an organization to aid the decision making process.

CRISP-DM Data Mining Process Methodology

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements
Assumptions
Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan

Data Understanding

Initial Data Collection
Initial Data Collection Report

Data Description
Data Description Report

Data Quality Verification
Data Quality Report

Exploratory Analysis
Exploratory Analysis Report

Data Preparation

Data Set
Data Set Description

Selection
Rationale for Inclusion / Exclusion

Cleaning
Data Cleaning Report

Construction
Derived Variables
Generated Records
Transformation

Integration
Merging
Aggregation

Formatting
Rearranging Attributes
Reordering Records
Within-Value Reformatting

Modelling

Generate Test Design
Test Design

Build Model
Parameter Settings
Models

Model Evaluation
Model Description
Assessment

Evaluation

Evaluate Results
Approved Models
Assessment of Data Mining Results w.r.t. Business Success Criteria

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

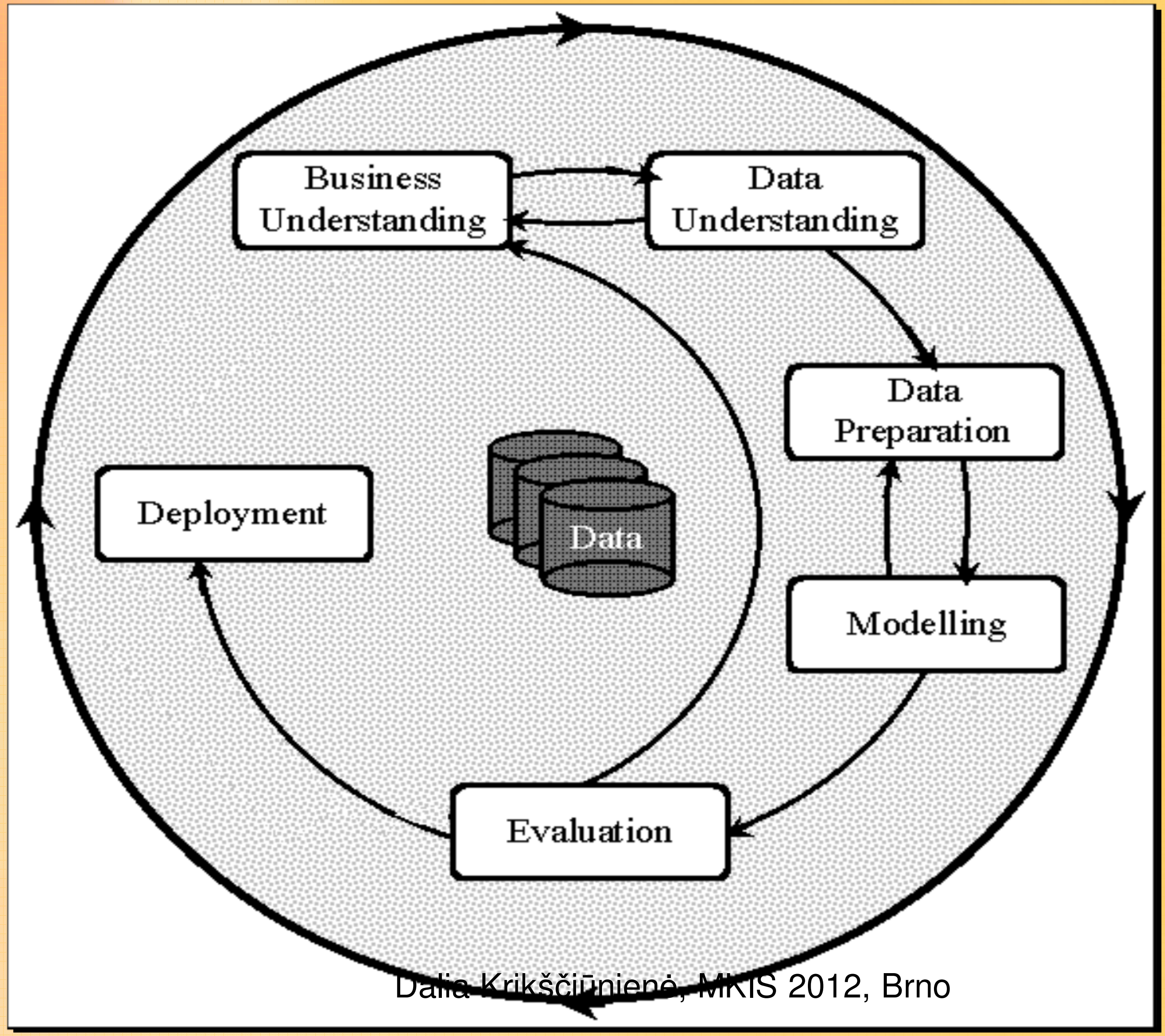
Plan Deployment
Deployment Plan

Produce Final Report
Final Report
Final Presentation

Plan Monitoring and Maintenance
Maintenance Plan

Review Project
Experience Documentation

The Iterative Nature of the Knowledge Discovery process



Data Mining Technique categories

1. **Predictive Techniques**
   - **Classification:** serve to classify the discrete outcome variable.
   - **Prediction or Estimation:** predict a continuous outcome (as opposed to classification techniques that predict discrete outcomes).
2. **Descriptive Techniques**
   - **Affinity or association:** serve to find items closely associated in the data set.
   - **Clustering:** create clusters according to similarity defined by complex of variables of input objects, rather than an outcome variable.

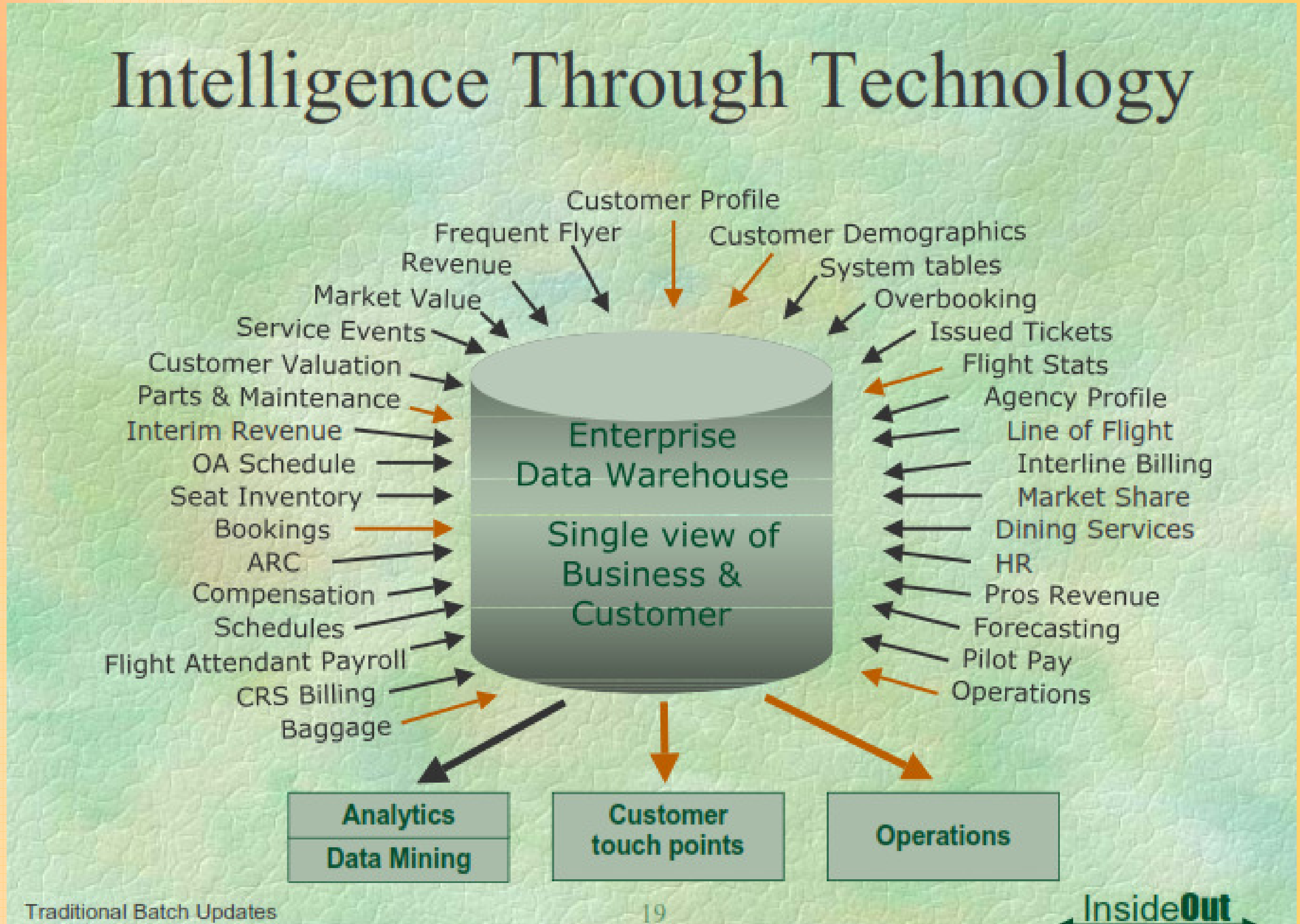
Web Data Mining - Types

1. **Web structure mining –** Examines how the Web documents are structured, and attempts to discover the model underlying the link structures of the Web.
    - ***Intra-page structure mining*** evaluates the arrangement of the various HTML or XML tags within a page
    - ***Inter-page structure*** refers to hyper-links connecting one page to another.
2. **Web usage mining *(Clickstream Analysis)* –** Involves the identification of patterns in user navigation through Web pages in a domain.
    - Processing, Pattern analysis, and Pattern discovery
3. **Web content mining –** Used to discover what a Web page is about and how to uncover new knowledge from it.

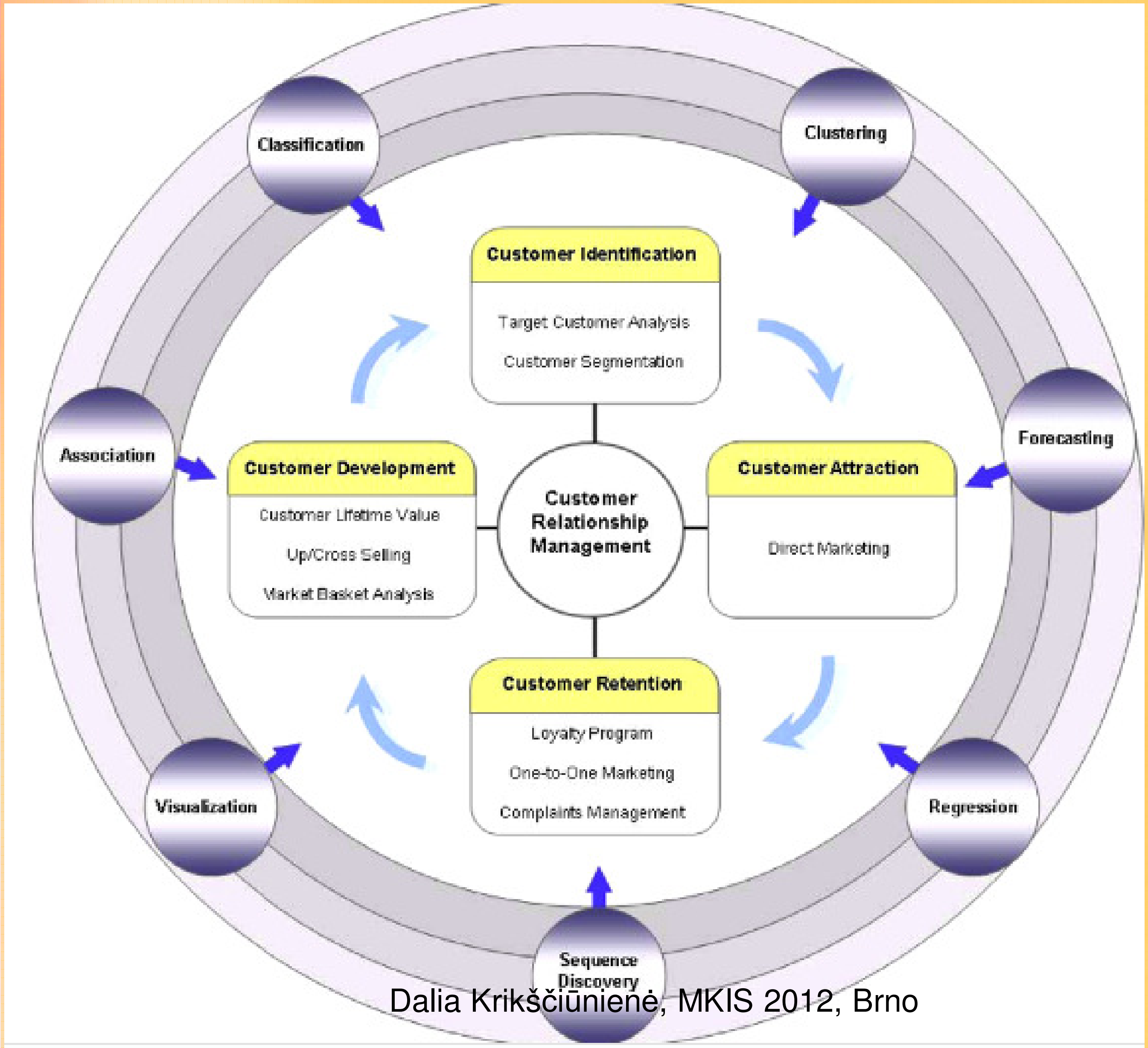
Barriers to the use of DM

- Two of the most significant barriers that prevented the earlier deployment of knowledge discovery in the business relate to:
  - •Lack of data to support the analysis
  - •Limited computing power to perform the mathematical calculations required by the data mining algorithms.

Variables for consideration in airline planning



Classification of data mining methods for CRM



Neural networks

- They are used for classification, regression, time series forecasting tasks
- Supervised and unsupervised learning
- Supervised means, that you have data samples with the known outcome (e.g. credit success and failure cases). Theses samples are used for creating NN model by learning. The outcome for new unknown samples is computed according to NN model
- Unsupervised means, that we do not know the outcome for samples, but we can cluster them according to their similarity by taking into account all known information, put into data records consinsting of many variables.

Good NN problem has following characteristics

- Inputs are well understood. You know which features (indicators) are important, but not necessarily know how to combine them
- Outputs are well understood. You know wht you try to model
- Experience is available- you have enough examples where both input and output are known. These cases will be used to train network
- A black box model is acceptable. Explaining and interpreting model is not necessary

Neural network analysis

- Neural network performance is based on node's activation function
- Inputs are combined into single value, then passed to transfer function to produce output
- Each input has its own weight
- Usually combination function is a weighted sum
- Other possibilities-max function (e.g. radial basis network has other combination)
- Transfer function is made by 0-1 or sigmoid (continuous)
- If linear- neural network is the same as linear regression
- Sigmoid is sensitive in middle range: small change makes big difference

Neural network analysis

- NN has linear behavior similarity in large ranges and non-linear in small
- Power of NN is in non-linear behavior due to activation of constituent unite
- It leads to requirement to have similar ranges of inputs (standardized or near to 0)
- In this case weight adjustment will have bigger impact

Neural network models

The generally applied network types for designing neural network models are Multilayer Perceptron, Radial Basis Function and Probabilistic Neural Network.

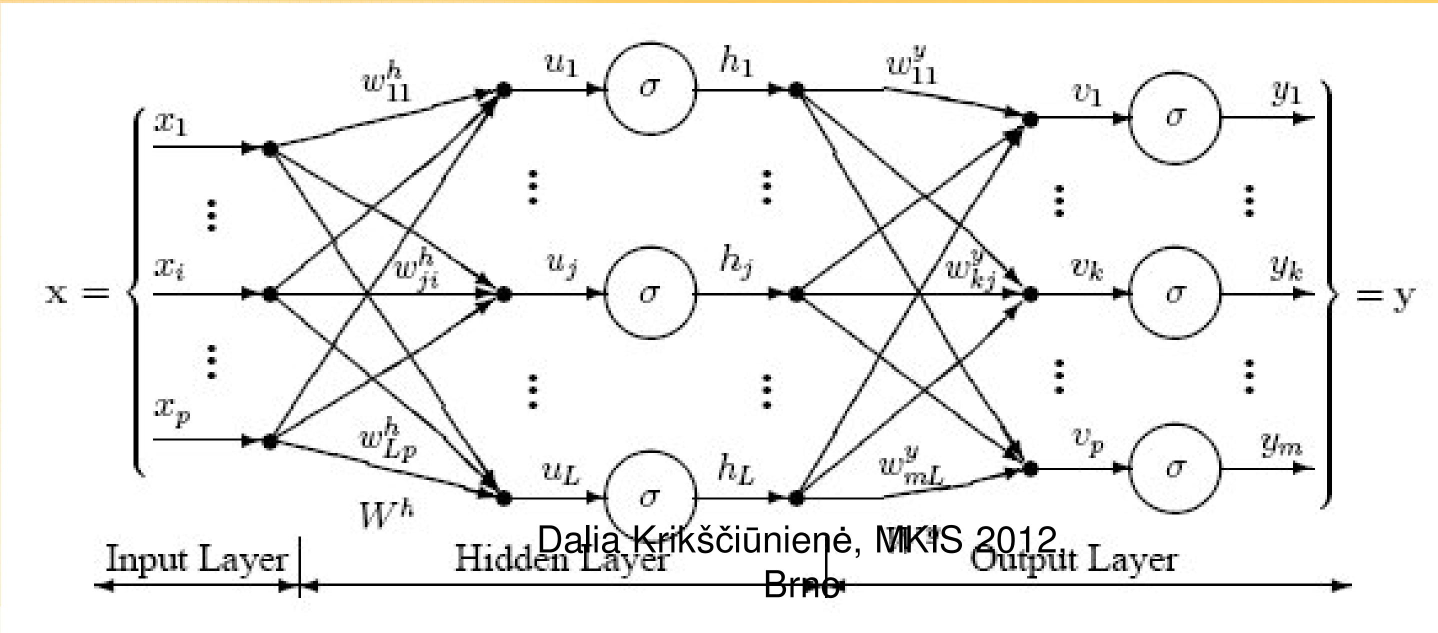
The main difference is in their algorithms, used for analysis and grouping of the input cases for further classification.

The Multilayer Perceptron NN model

The following diagram illustrates a perceptron network with three layers:

This network has an **input layer** (on the left) with three neurons, one **hidden layer** (in the middle) with three neurons and an **output layer** (on the right) with three neurons.

There is one neuron in the input layer for each predictor variable. In the case of categorical variables, *N*-1 neurons are used to represent the *N* categories of the variable.



Multilayer perceptron

- Hidden layer gets inputs from all nodes in input layer
- Standardization is important
- In hidden layer – hyperbolic tangent is preferred, as it gives positive and negative values
- Transfer function depends on target
  - For continuous- linear is preferred
  - For binary- logistic, which behaves as probability
- One hidden layer is usually sufficient
- The wider it is, the bigger capacity NN gains
- The drawback of increasing hidden layer is memorizing instead of generalizing (overfit)

Multilayer perceptron

- A small number of hidden layer nodes with non-linear transfer functions are sufficient for very flexible models
- Output is weighted linear combination
- Usually output is one value and is calculated from all nodes of hidden layer
- One additional input- constant which is weighted as well
- Topologies can vary- NN can have more outputs (e.g. calculating probability that customer will by in each of the departments NN has output for each department)
- The results can be used in different ways, usually selected by experimenting: take max, take top 3, take those above threshold, take meeting percentage from maxs

Multilayer perceptron

- Training is performed for one set in order to test performance with the other
- It is similar to finding one best fit line for regresssion
- In NN there is no single case of best fit, it uses optimization
- Goal is to find set of weights which minimize the overall error function, e.g. average square error

Multilayer perceptron

First successful training method- back propogation, 3 steps:

- Get data, compute outputs with existing weights of the system (e.g. random)
- Calculate overall error by taking difference of actual values
- Error is sent back to network, weights are adjusted

Then blame is adjusted to nodes, and weights adjusted for these nodes

(complex math procedure of partial derivatives is used)

- After sufficient generations and showing sufficient training samples the error no longer decreases- stop

Multilayer perceptron

- The weights are adjusted: if their change decrease overall error (not eliminate)
- After sufficient generations and showing sufficient training samples the error no longer decreases- stop
- Training set has to be balances to have enough various cases as goal is to generalize
- This technique is called generalization delta rule-2 param:
  - Momentum- weight remembers which direction is was changing, it tries to go same direction. If momentum is high the NN responds slowly to samples which try to change direction. Low momentum allows flexibility
  - Learning rate controls how quickly weights change. Best approach is to start big and decrease slowly as NN is being trained.

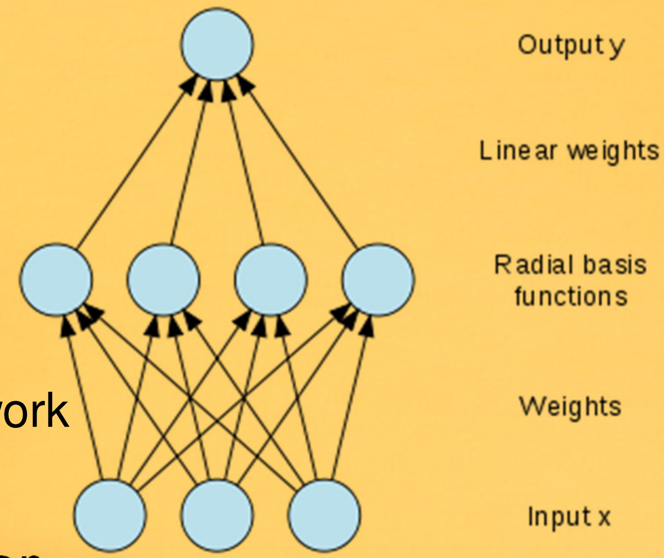
Multilayer perceptron

- Initially weights are random
- Large oscillations are useful
- Getting closer to optimal, learning rate should decrease
- There are more methods, the goal for all of them – to arrive quickly to optimal

Radial basis function network

- Fitting a curve exactly thr ough a set of points
  - Weighted distances are computed between the input x and a set of prototypes
  - These scale distances are then transformed through a set of nonlinear basis functions h, and these outputs are summed up in a linear combination with the original inputs and a constant.

Radial basis function network



$$\varphi(\mathbf{x}) = \sum_{i=1}^{N} a_i \rho(||\mathbf{x} - \mathbf{c}_i||)$$

$$\rho\big(\,\|\mathbf{x} - \mathbf{c}_i\|\,\big) = \exp\left[-\beta\,\|\mathbf{x} - \mathbf{c}_i\|^2\right].$$

Radial basis function network RBF

- They differ from MLPin 2 ways:
  - Interpretation relies on geometry rather than biology
  - Training method is different as in addition to optimizing weights used to combine outputs of RBF nodes , the nodes themselves have parameters that can be optimized
- As with other types of NN the data processed is always numeric, so it is possibles to interpret any input record as point in space

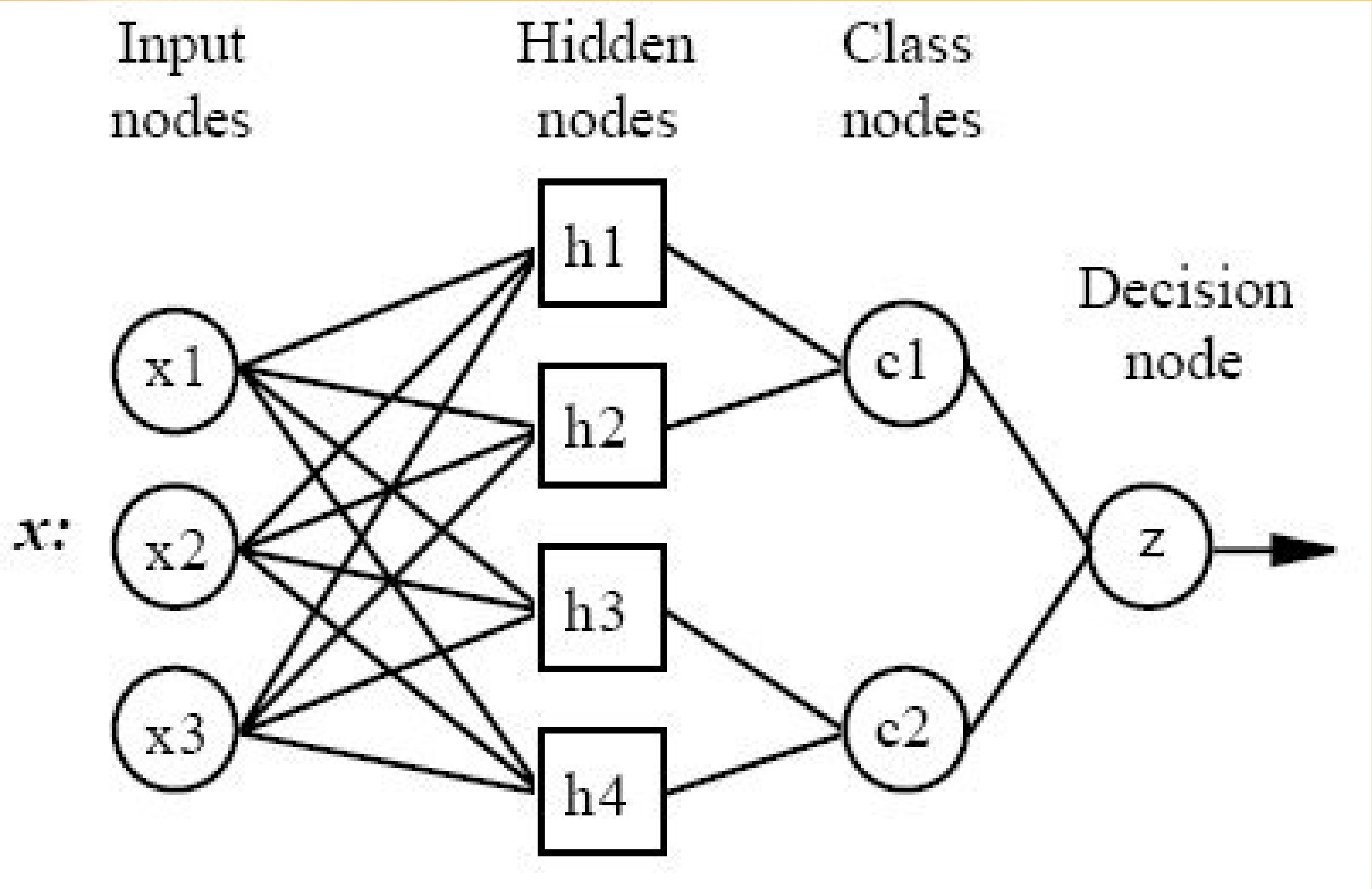
Radial basis function network

- In RBF network hidden layer nodes are also points in same space, Each has address specified by vector of elements which number equals to no. of variables
- Instead of combination and transfer functions the RBF have distance and transfer functions
- Distant function os standard Euclidean – suqare root of quadratic distances of each dimension
- The nodes output is non-linear function of how dimension is close to the input is: the closer the input, the stronger the output.

Radial basis function network

- „Radial“ refers to the fact that all inputs of same distance from node‘s position produce same output
- In two dimensions they produc circle, in 3D- sphere
- RBF nodes are in hidden layer and also have transfer functions
- Instead of S-shape (as in MLP) these are bell-shape Gaussians (multidimensional normal curve)
- Unlike MLP the RBF does not have weights associated with connections between input and hidden layers

Probabilistic NN



Probabilistic Neural Network model

This type of network copies every training case to the hidden layer of the network, where the Gaussian kernel-based estimation is further applied. The output layer is then reduced, by making estimations from each hidden unit.

The training is extremely fast, as it just copies the training cases after their normalization to the network. But this procedure tends to make the neural network very large, therefore this makes them slow to execute.

Probabilistic Neural Network model

During the testing stage the Probabilistic Neural Network model requires a number of operations approximately proportional to the square of the number of training cases, therefore for the large number of cases the total duration of creating model becomes similar to the other network types that are usually described as being far slower to train (e.g. multilayer perceptrons).

If the prior probabilities (of class distribution) are known and different from the frequency distribution of the training set, they can be incorporated in training of the network model, otherwise the distribution is described by frequency (StatSoft Inc.).

Memory-Based Reasoning MBR

- MBR belong to the class of tasks- Nearest neighbour techniques
- MBR results are based on analogous situations in past
- Application:
  - Collaborative filtering (not only similarity among neighbours but also their preferences), customer response to offer
  - Text mining approach
  - Acoustic engineering: mobile app Shazam which identifies songs from snippets captured in mobile phone
  - Fraud detection (similarity to known cases)

Memory-Based Reasoning MBR

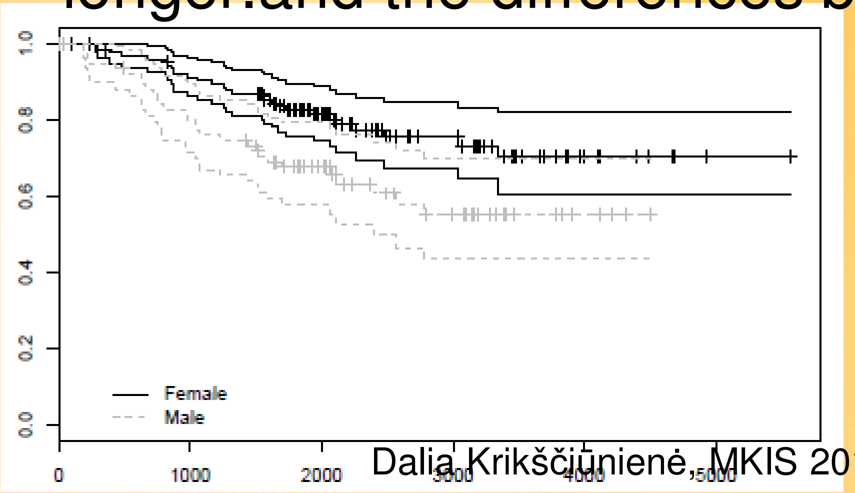
- MBR uses data as it is. Unlike other DM techniques it does not care of data formats
- Main components: distance function between two records and combination function (combine results from several neighbors and give result)
- Ability to adapt- add new categories
- Does not need long training, e.g. for Shazam app new songs are added on daily basis and app just works
- Disadvantage- method requires larga sample data base. Classifying new record needs processing all historizal records

Survival analysis

- It means time-to-event analysis. It tells when to start worrying about customers doing something important
- It identifies which factors are most correlated with the event
- Survival curves provide snapshots of customers and their life cycles, it takes care of very important facet of customer behaviour- tenure.
  - When customer is likely to leave
  - .. Or migrate to other customer segment
  - Compound effect of other factors to tenure

Survival analysis

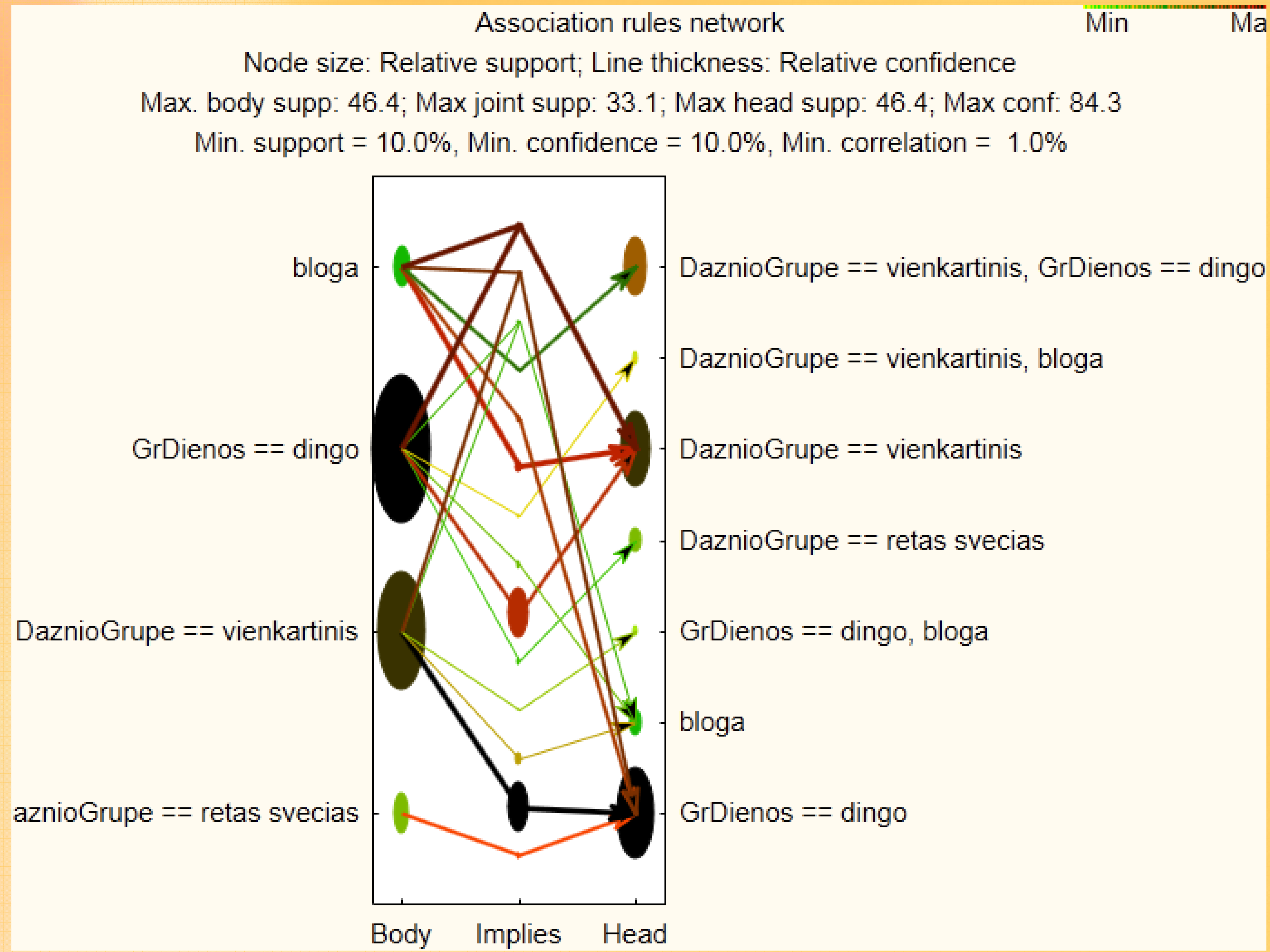
- Survival curve plotting: proportion of customers that are expected to survive up to particular point in tenute, based of historical info, how long customers survived in past : starts at 100%, decreases
- Graph procedures: Cox proportional hazards regression model. It shows how many customers are here after some time (e.g. 2000 days). Likelihood that they will stay longer.and the differences between two groups



Association rules

- They allow analysts and researchers to uncover hidden patterns in large data sets, such as "customers who order product *A* often also order product *B* or *C*" or "employees who said positive things about initiative *X* also frequently complain about issue *Y* but are happy with issue *Z*."
- Supports all common types of variables or formats in which categories, items, or transactions are recorded:**Categorical Variables, Multiple Response Variables, Multiple Dichotomies.** *STATISTICA Association Rules* (e.g., information regarding purchases of consumer items)

Association rules



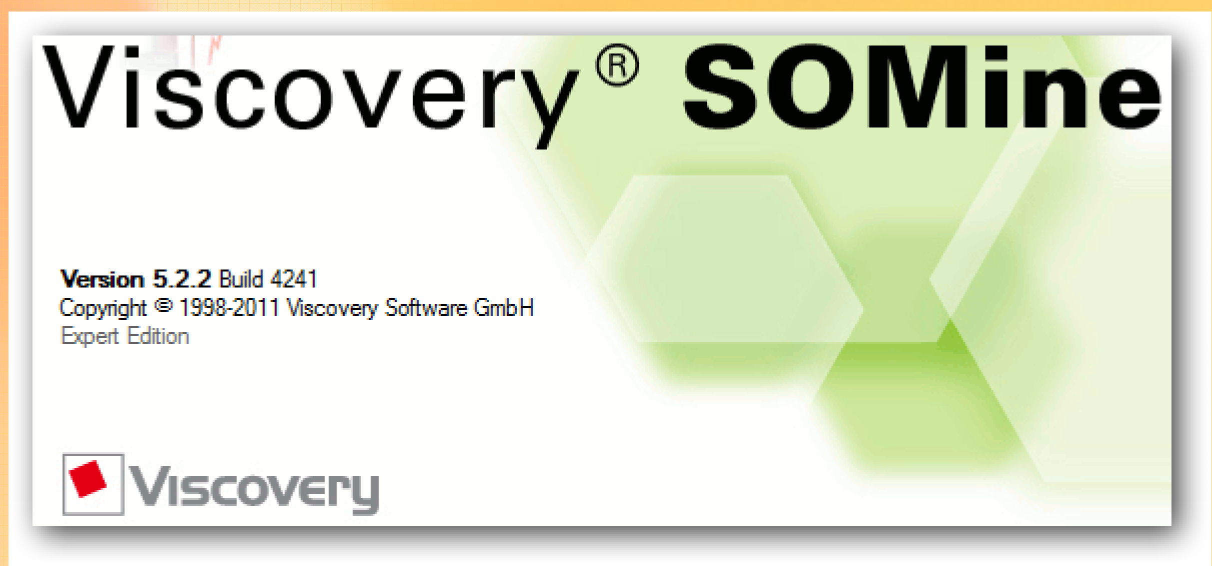
SOM – self organizing maps

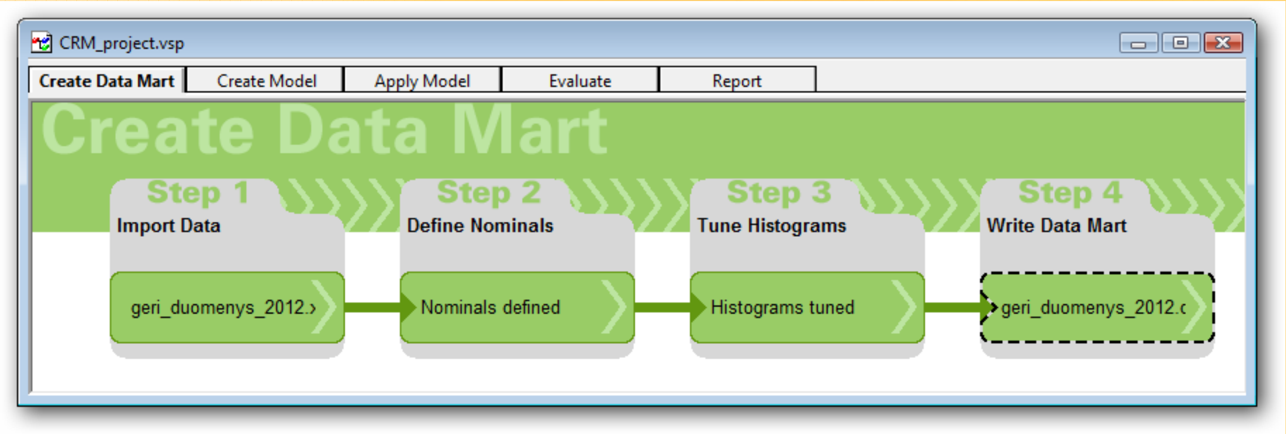
- A **self-organizing map** (**SOM**) or **self-organizing feature map** (**SOFM**) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space.

SOM – self organizing maps

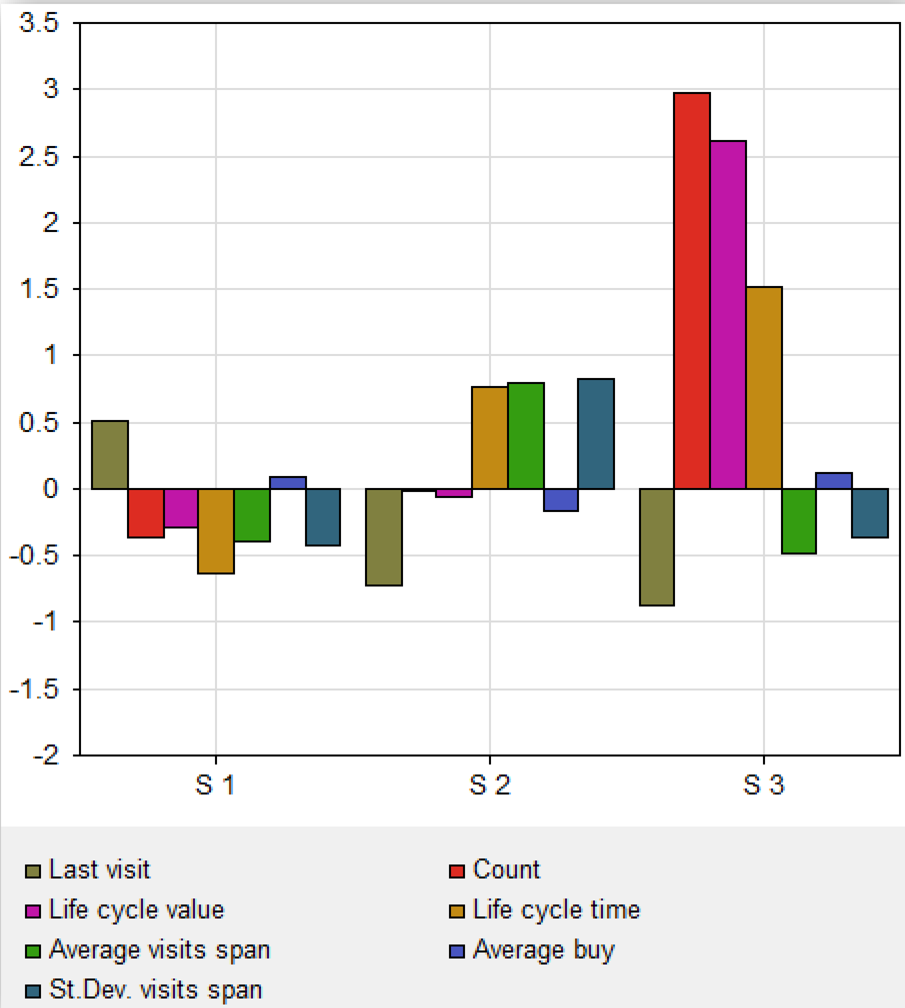
- For data mining purposes, it has become a standard to approximate the SOM by a two-dimensional hexagonal grid. The “nodes” on the grid are associated so-called “reference vectors” which point to distinct regions in the original data space. Starting with sets of numerical, multivariate data, these reference vectors on the grid gradually adapt to the intrinsic shape of the data distribution, whereby the reference vectors of neighbored nodes point to adjacent regions in the data space. Thus the order on the grid reflects the neighborhood within the data, such that data distribution features can be read directly from the emerging landscape on the grid.

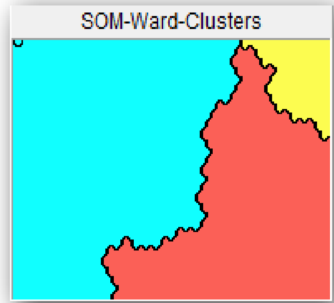
SOM – self organizing maps

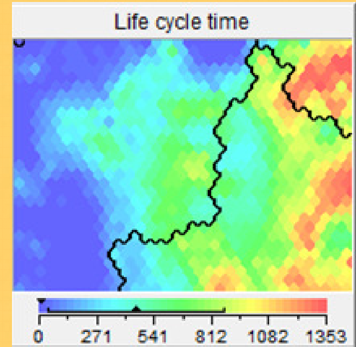




SOM – self organizing maps: cluster differences, influence of single variable to cluster separation



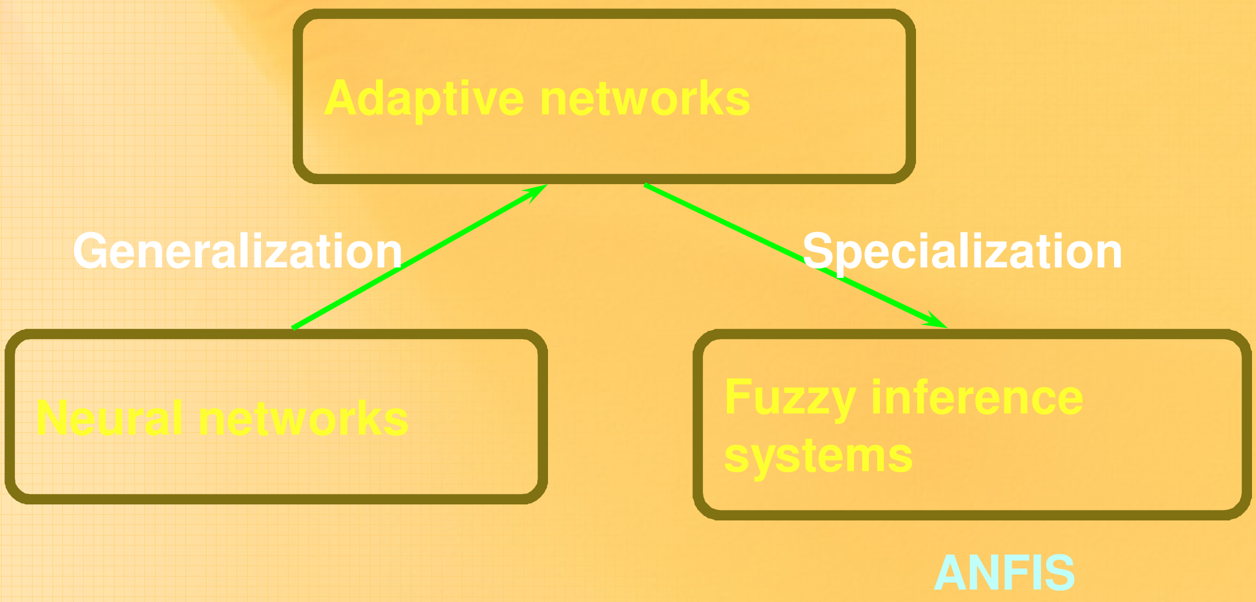




Attribute	Mean	Std. Deviati...	Minimum	Maximum
Last visit	107	156	1	649
Count	21.25	7.46	10.00	43.00
Life cycle value	24553	14066	3938	73582
Life cycle time	1092	287	149	1392
Average visits s...	55.8	22.1	13.9	115.8
Average buy	1196	552	164	3005
St.Dev. visits span	80.1	41.5	15.4	218.2

Fuzzy inference

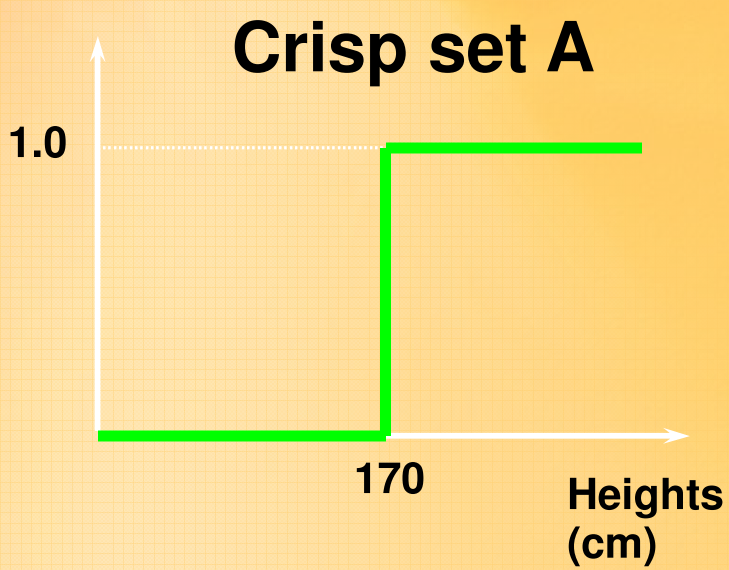
•Basic approach of ANFIS

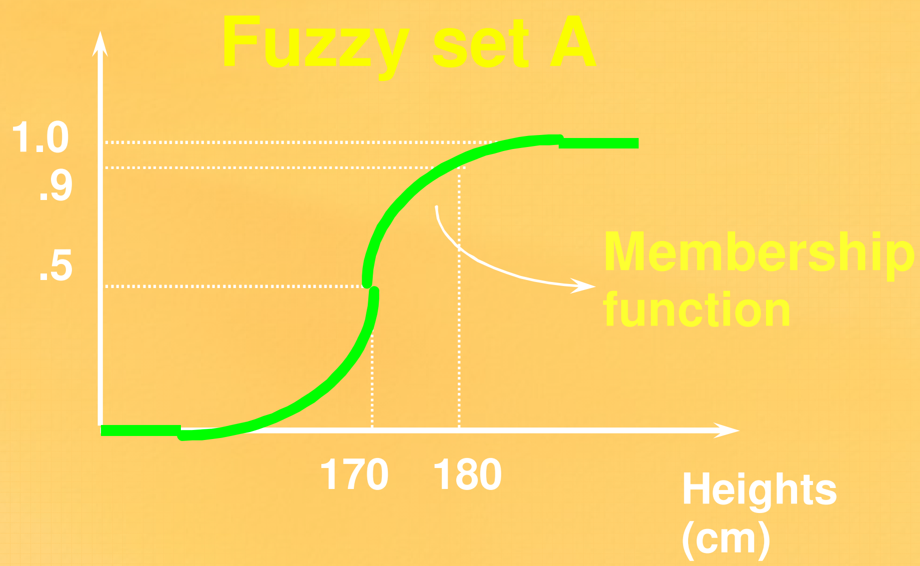


Fuzzy Sets

- Sets with fuzzy boundaries

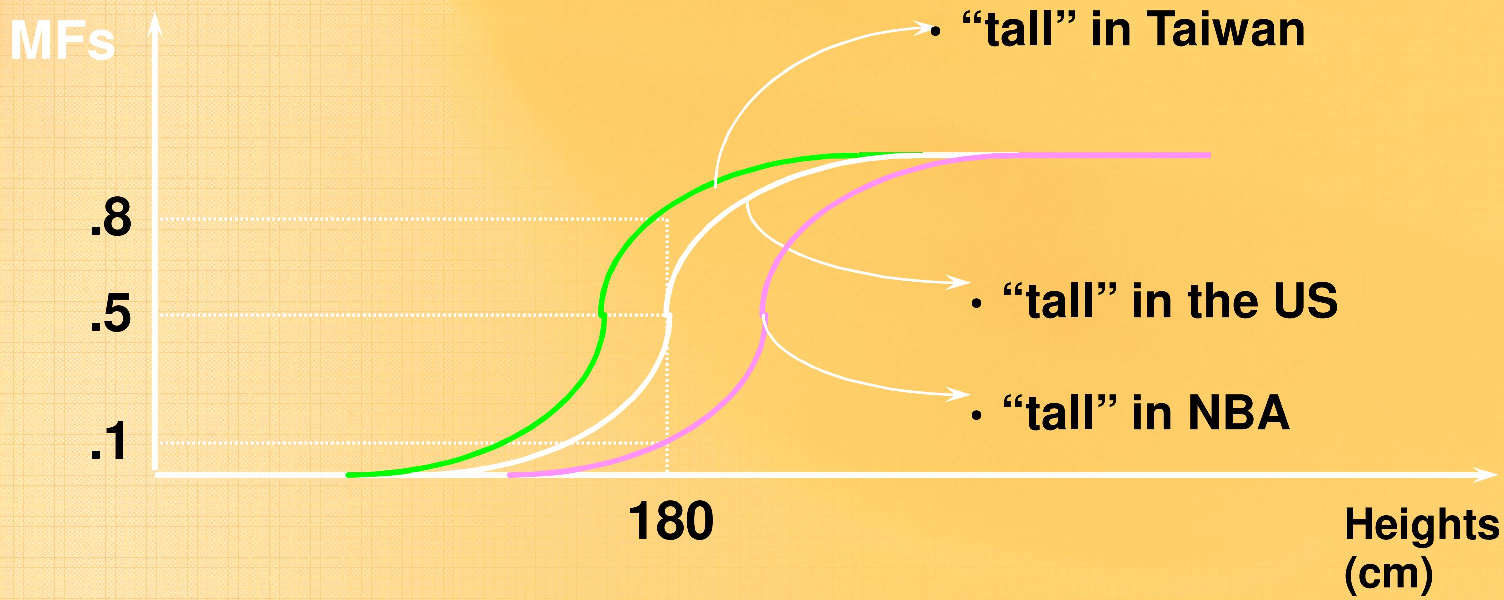
A = Set of tall people





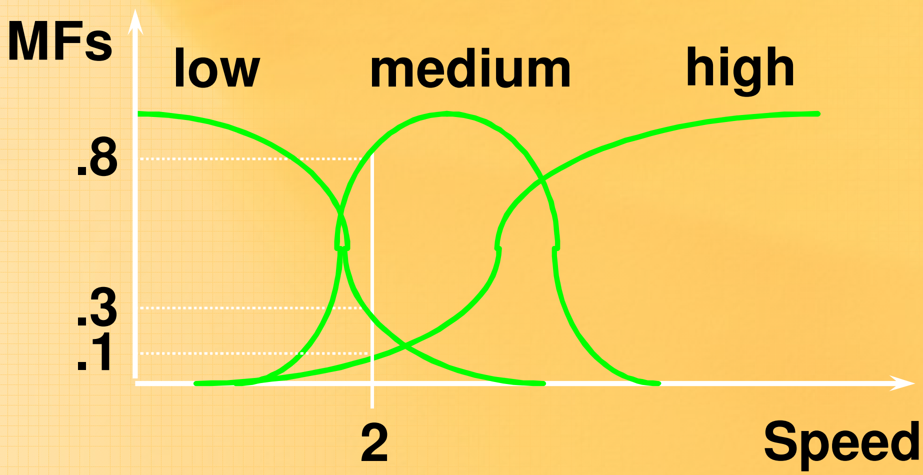
Membership Functions (MFs)

- Subjective measures
- Not probability functions



Fuzzy Inference System (FIS)

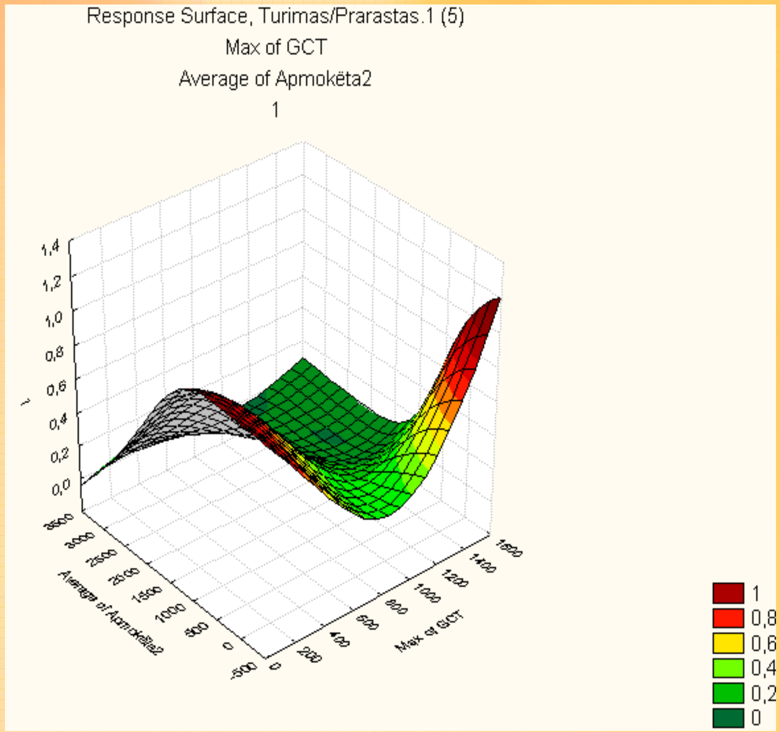
If speed is low then resistance = 2
If speed is medium then resistance = 4*speed
If speed is high then resistance = 8*speed

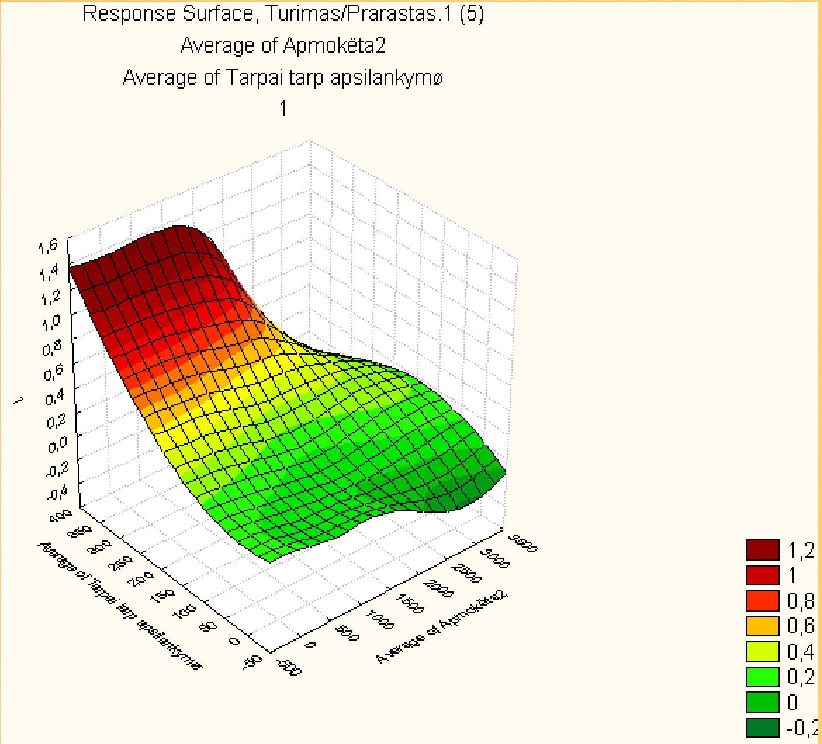


Rule 1: w1 = .3; r1 = 2
Rule 2: w2 = .8; r2 = 4*2
Rule 3: w3 = .1; r3 = 8*2

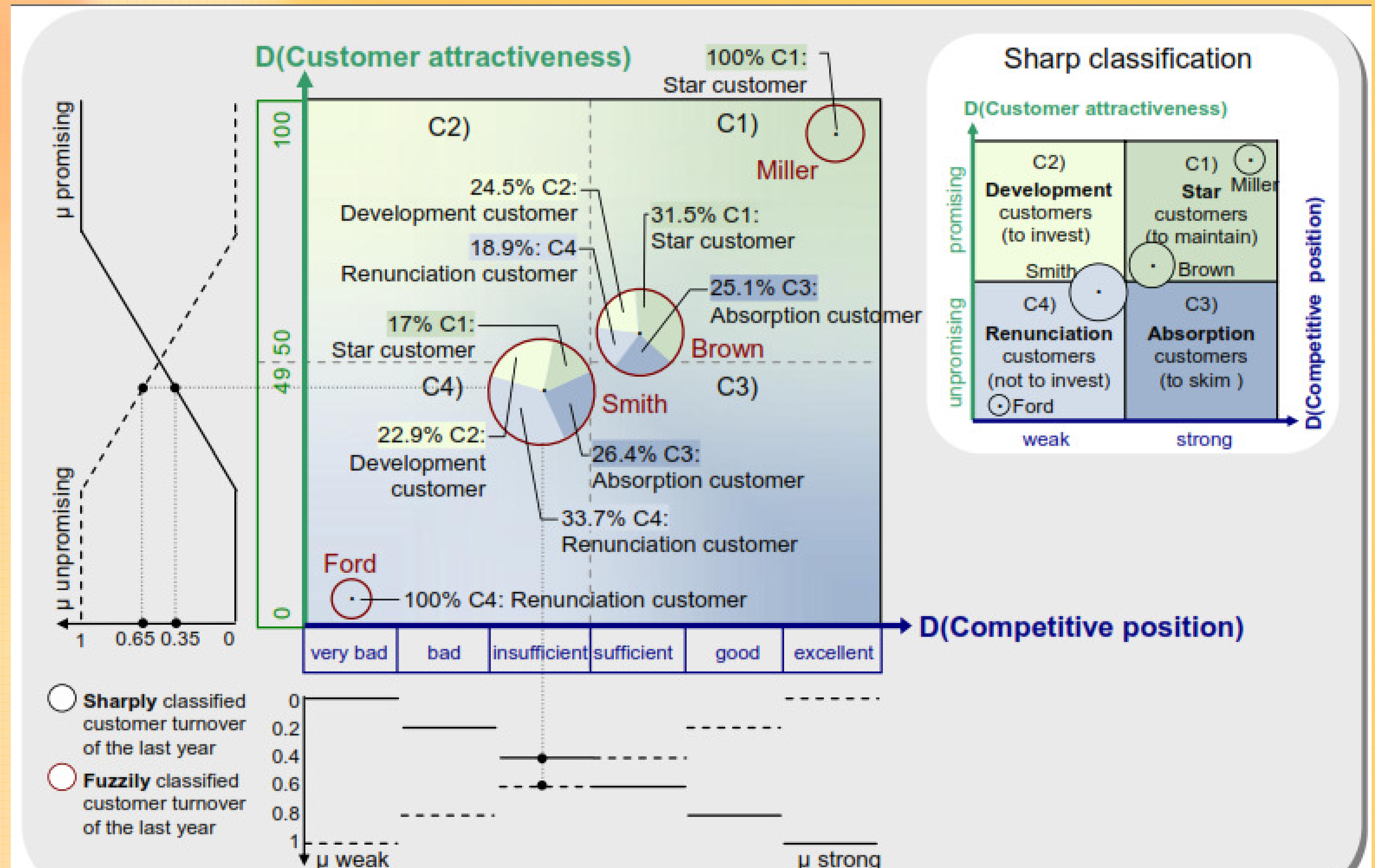
Resistance = Σ(wi*ri) / Σwi
= 7.12

Fuzzy inference: surface diagrams for relationship among variables

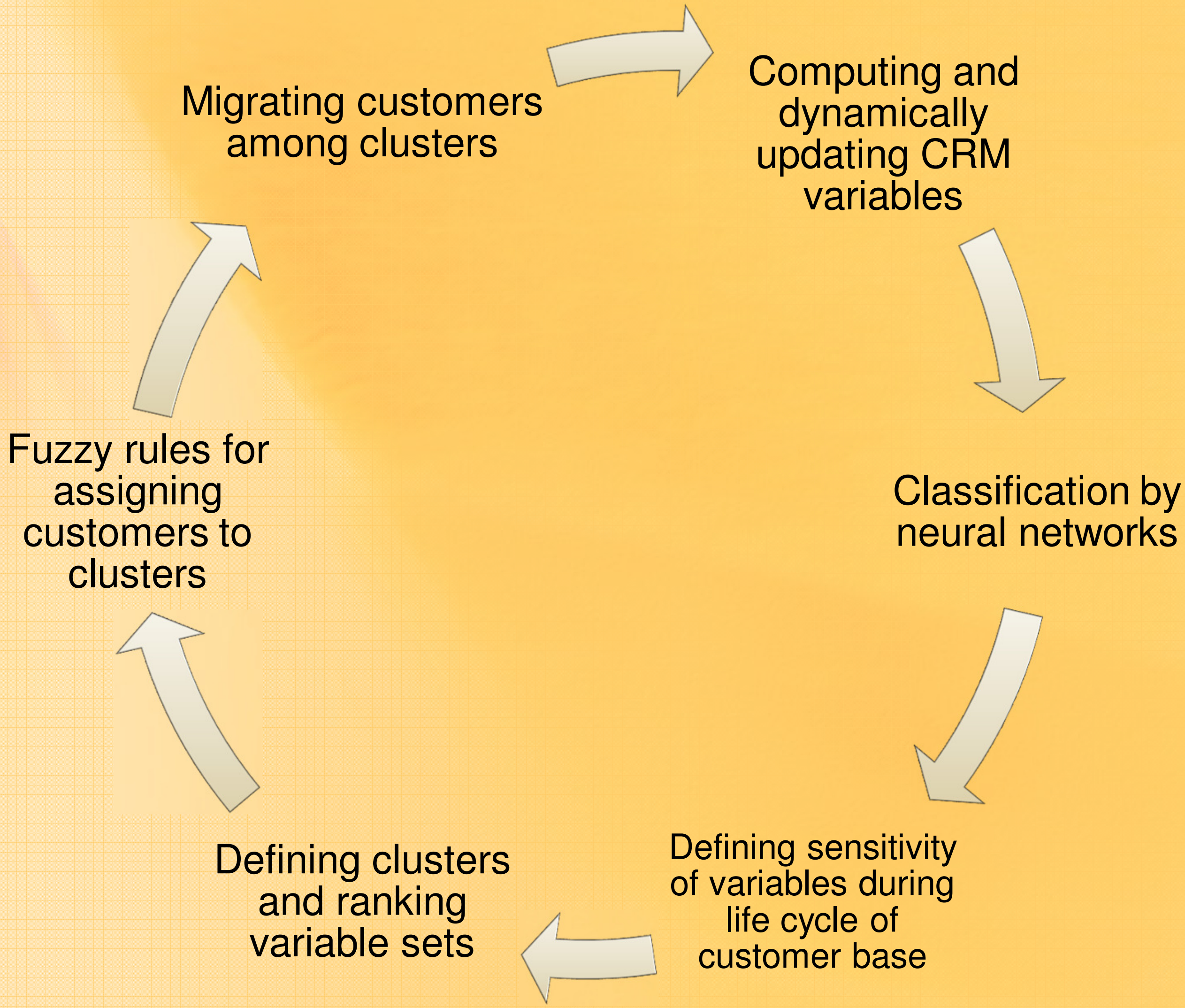




Fuzzy methods for marketing



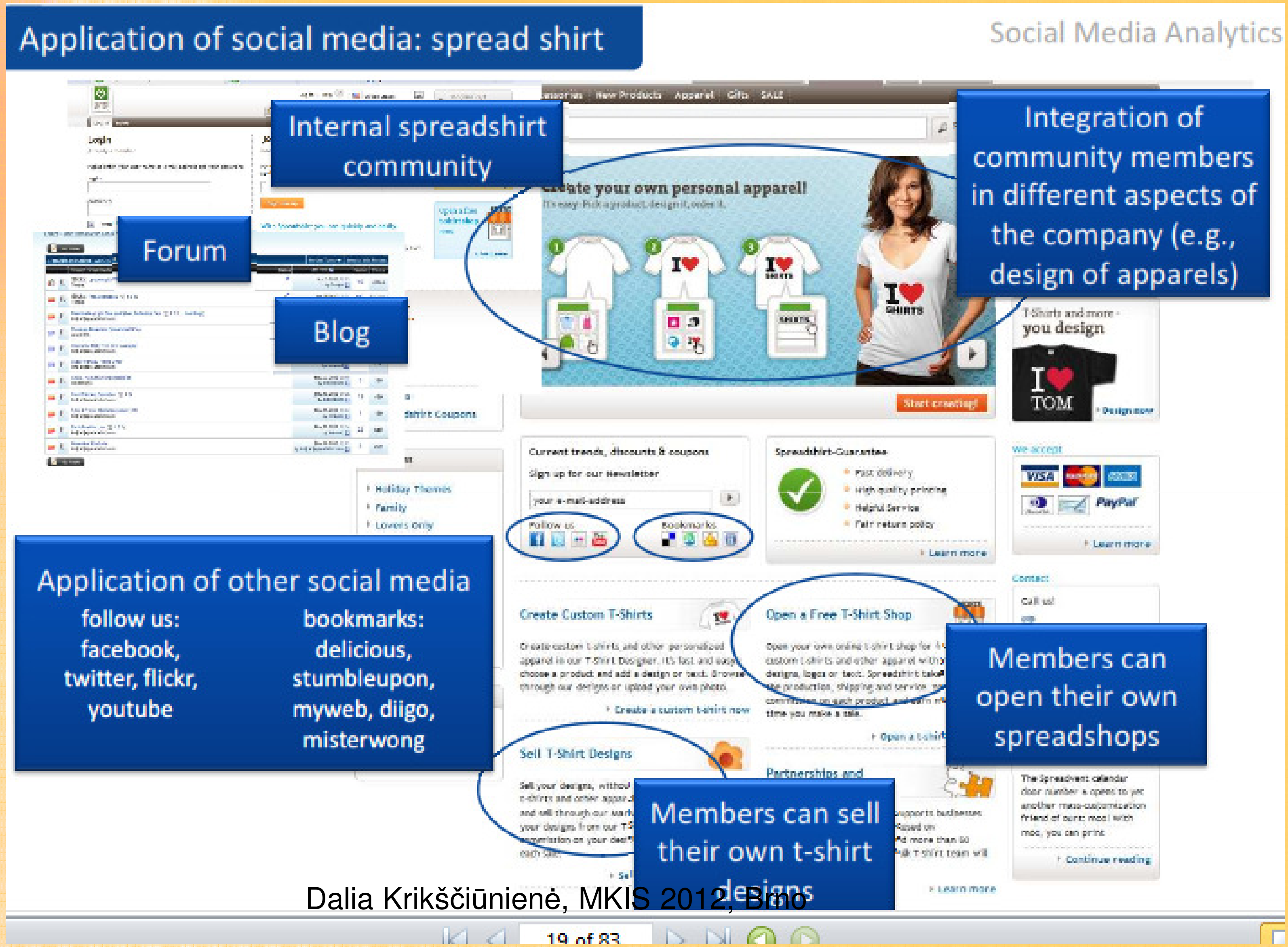
Combining methods for exploring customer performance



Web data mining

- Indicators for evaluation
- Opinion mining
- Text mining approaches and process
- Static analytic
- Dynamic analytic
- Sentiment analysis
- Classification
- Social network generation for analysis
- Social network analysis approach

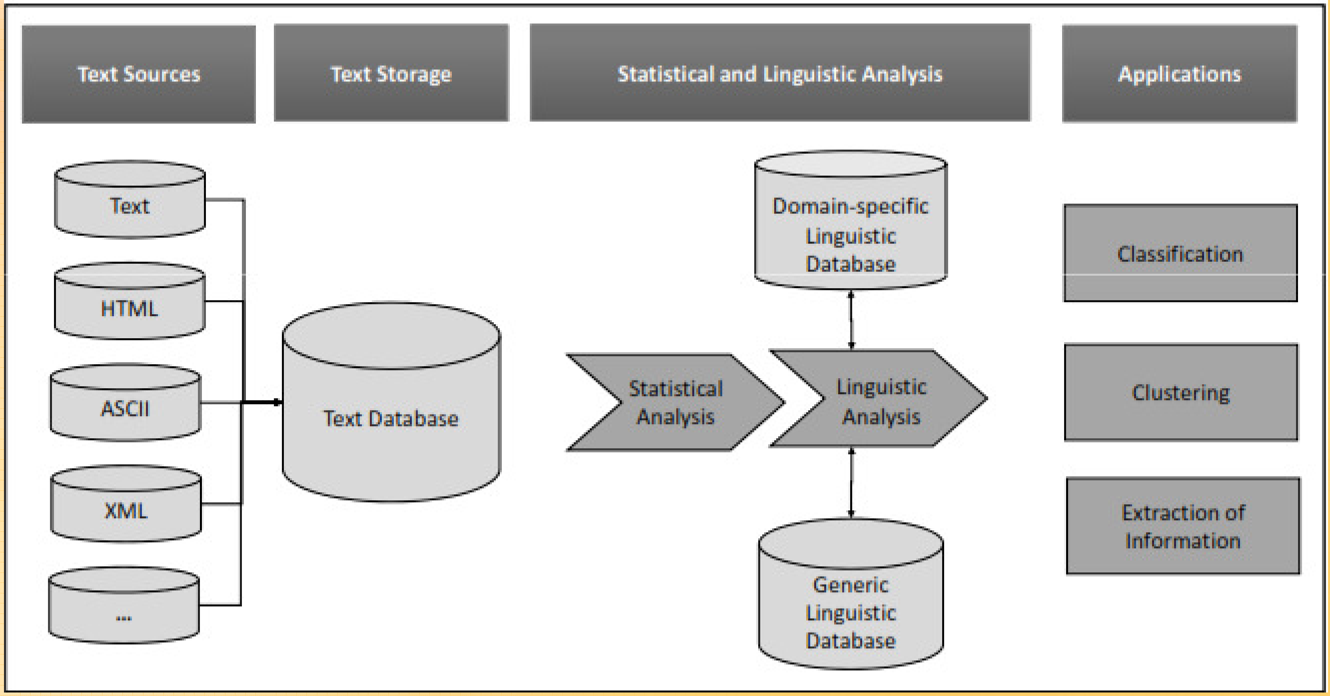
Social media analytics



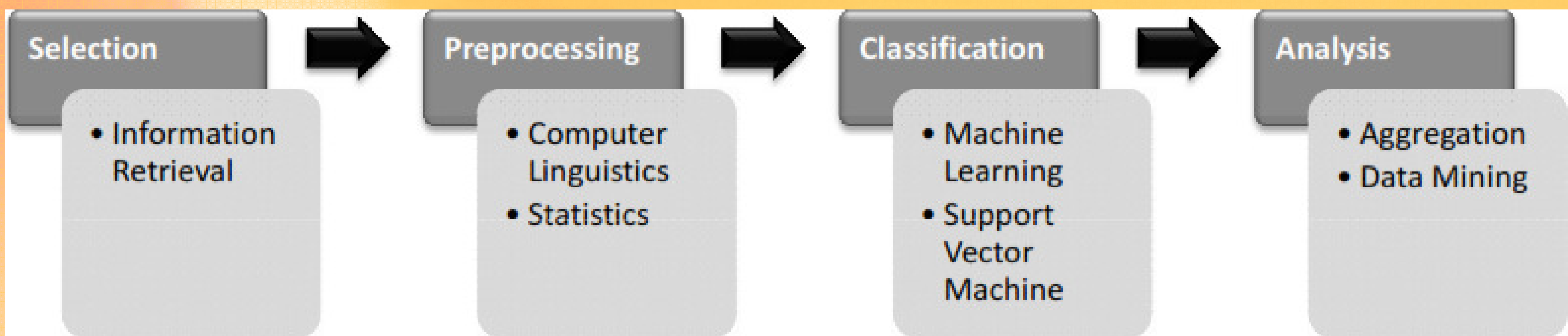
Analytic types in social media: Opinion mining

Analytic types in social media: text mining

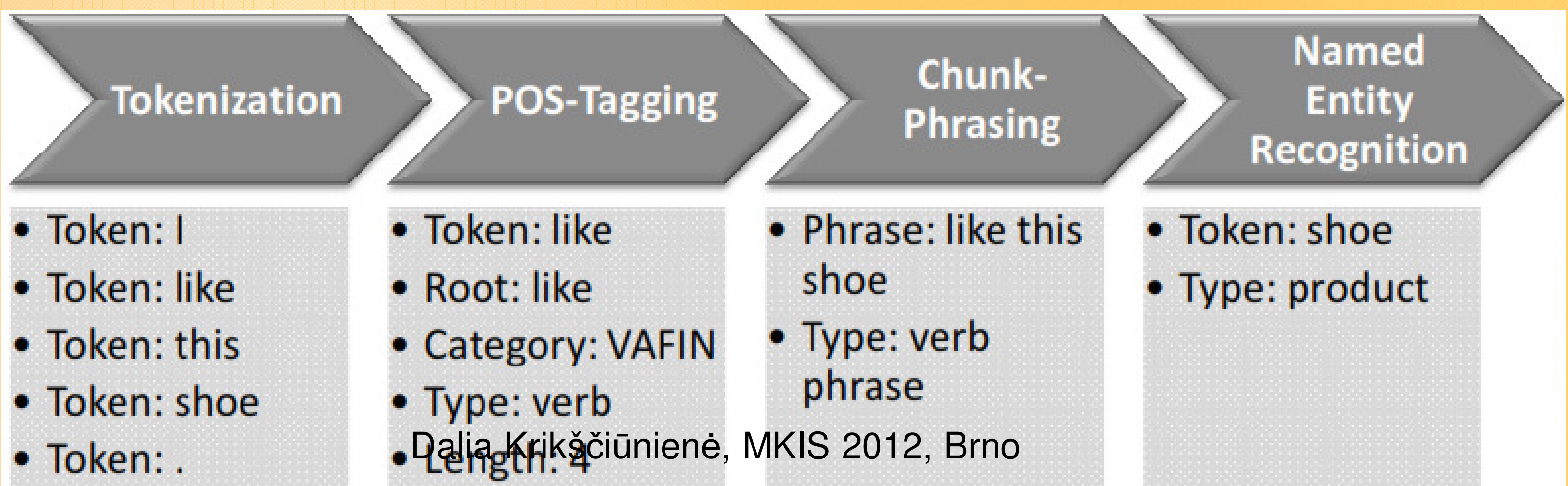
Text Mining aims at discovery and extraction of relevant information and knowledge from unstructured text, e.g. semantics of content or relationships of authors.



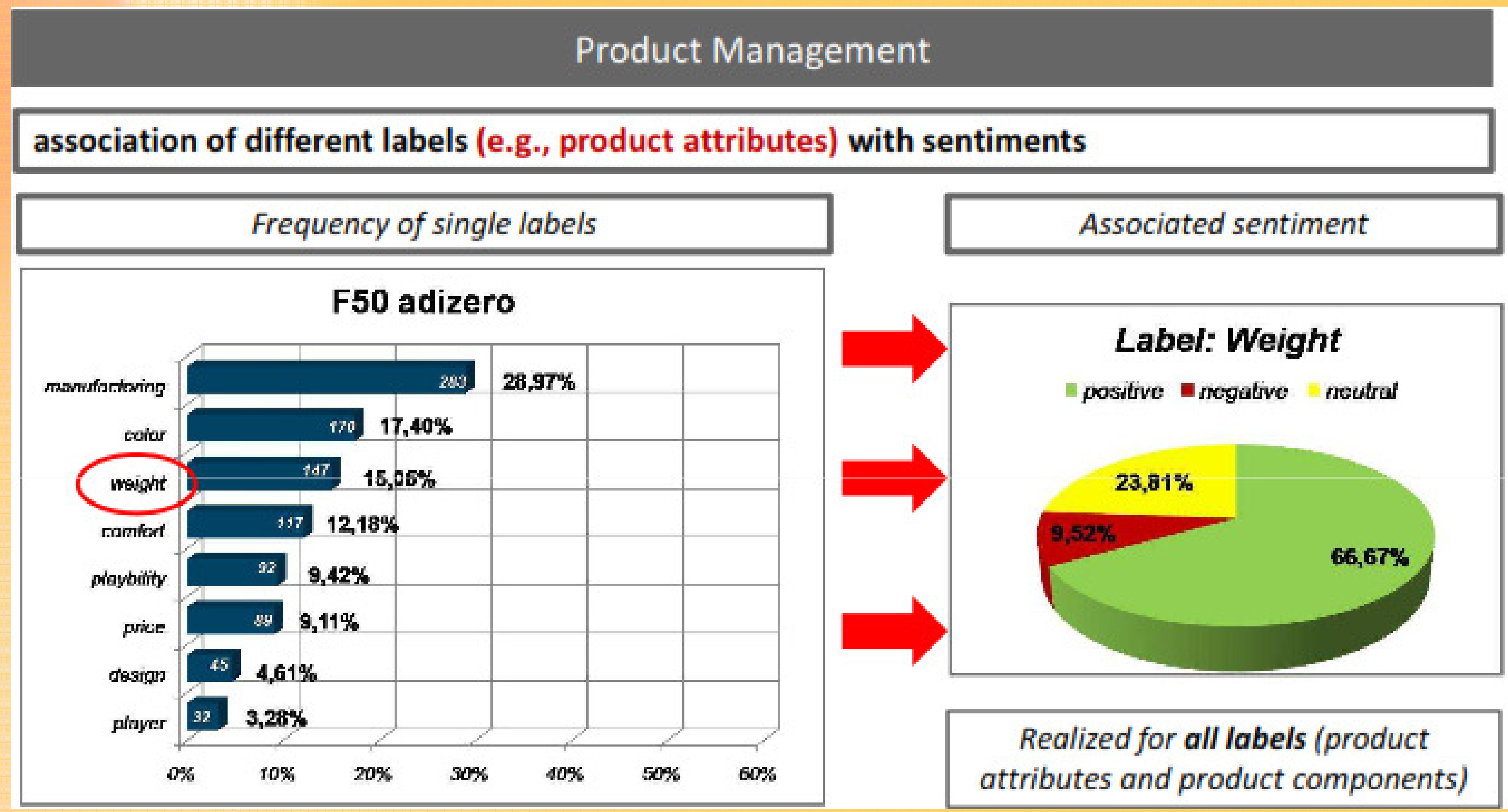
Mining process



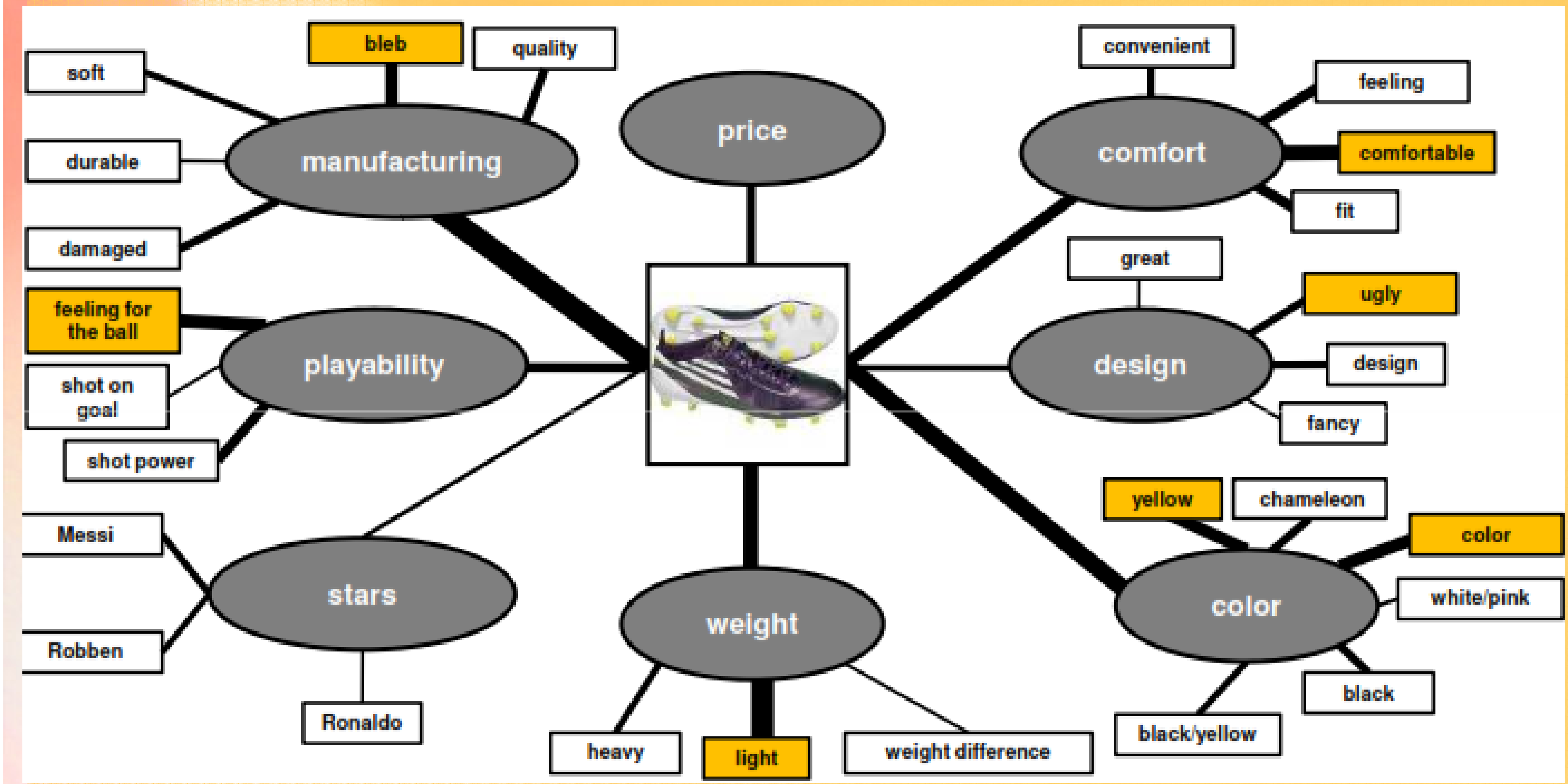
- Example "I like this shoe"



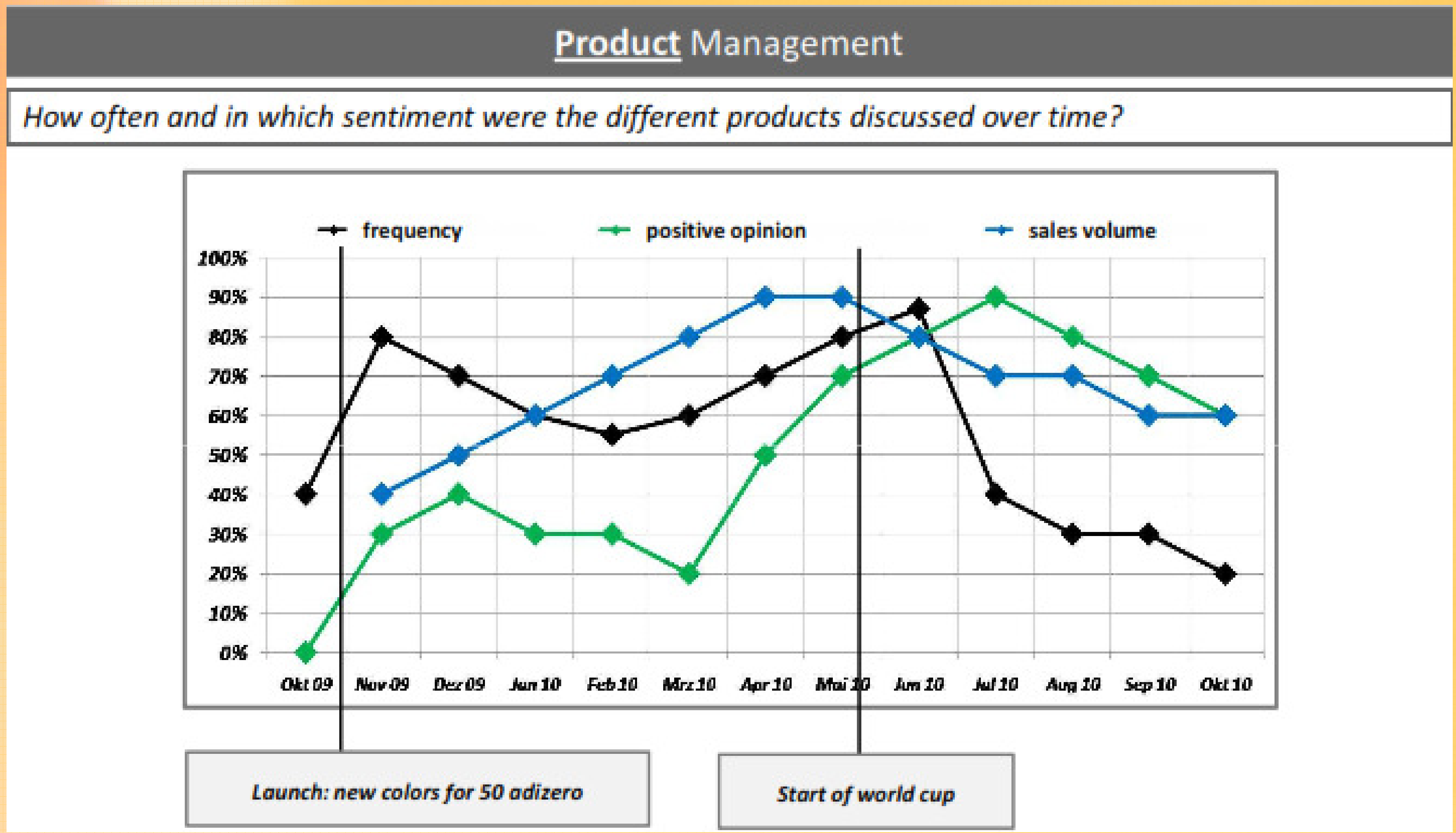
Static analytics (reporting, pivoting)



Static analytics (reporting, pivoting)



Dynamic analytics



Sentiment classification (text)

Player Management

Analysis of single words within the label (e.g., skills) and associated sentiments

Positive	
Single Words	**Frequency**
goal	33
ball	17
pass	12
technique	11
shoot	04

SKILLS

neutral — (34) 13,18%

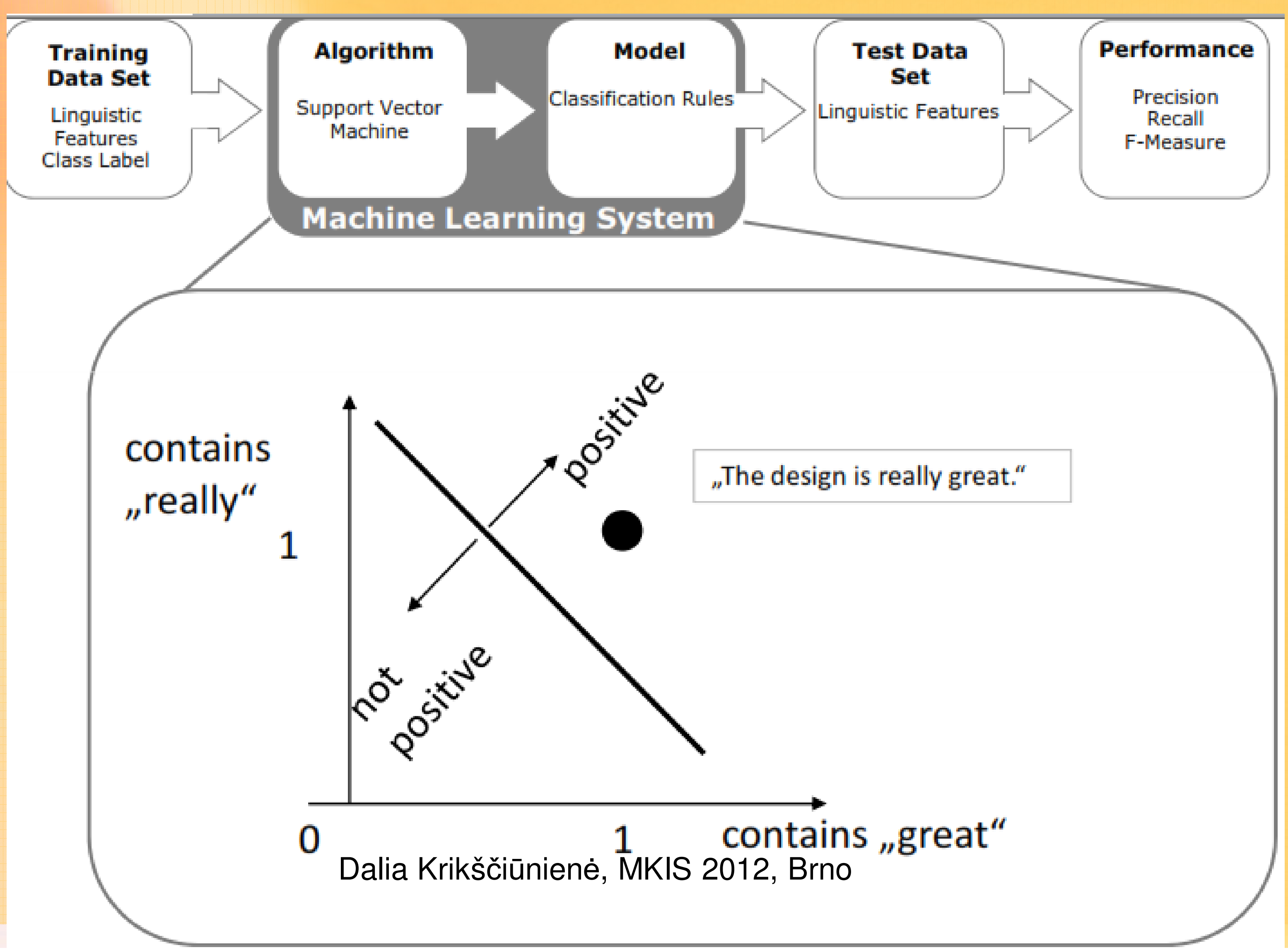
negative — (88) 34,11%
ball
goal
technical
dribbling
shot
pass
bad pass

positive — (136) 52,71%
goal
ball
pass
technique
shoot

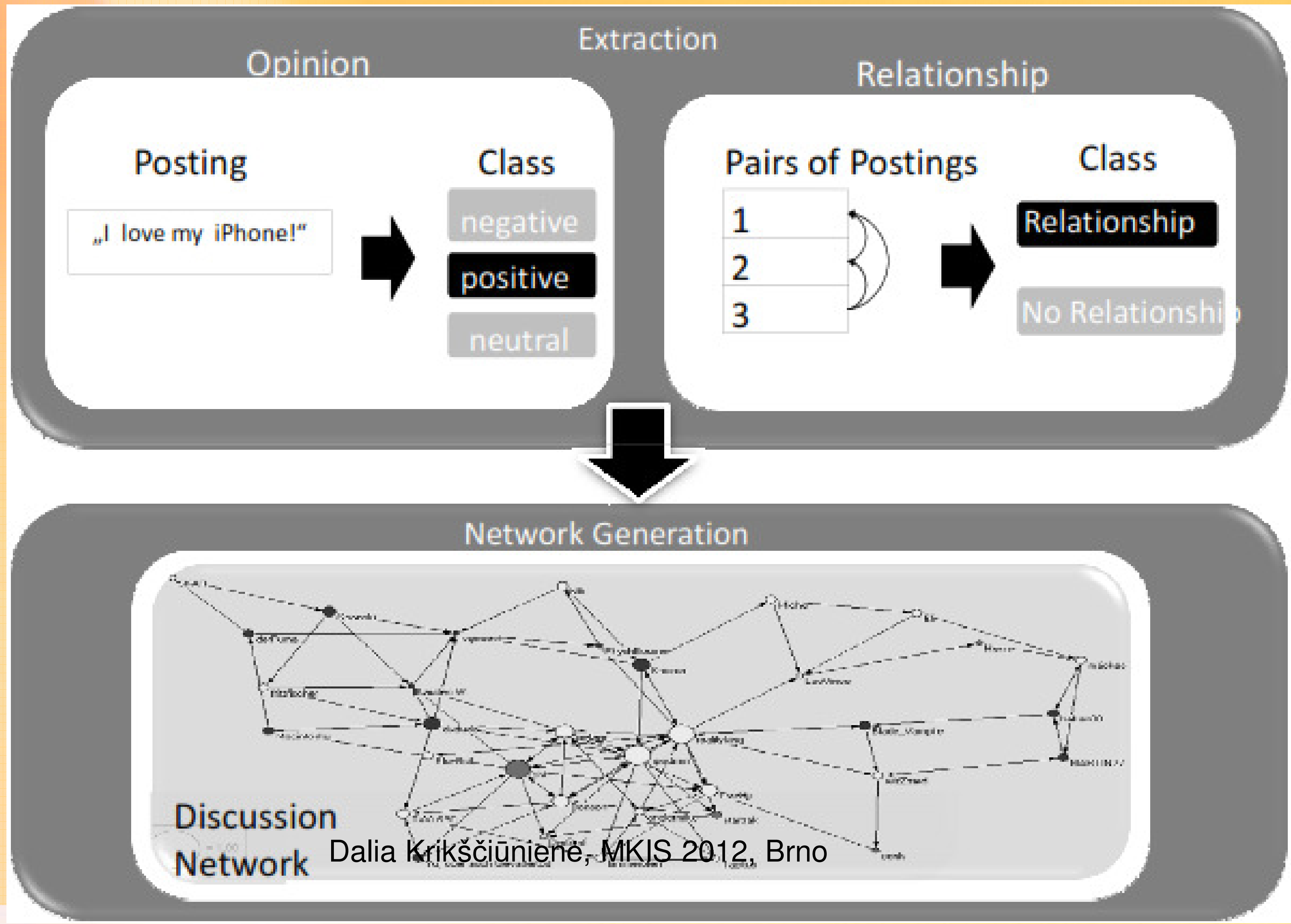
negative	
Single Words	**Frequency**
ball	14
goal	11
technical	08
dribbling	07
shot	06
pass	05
bad pass	03

37

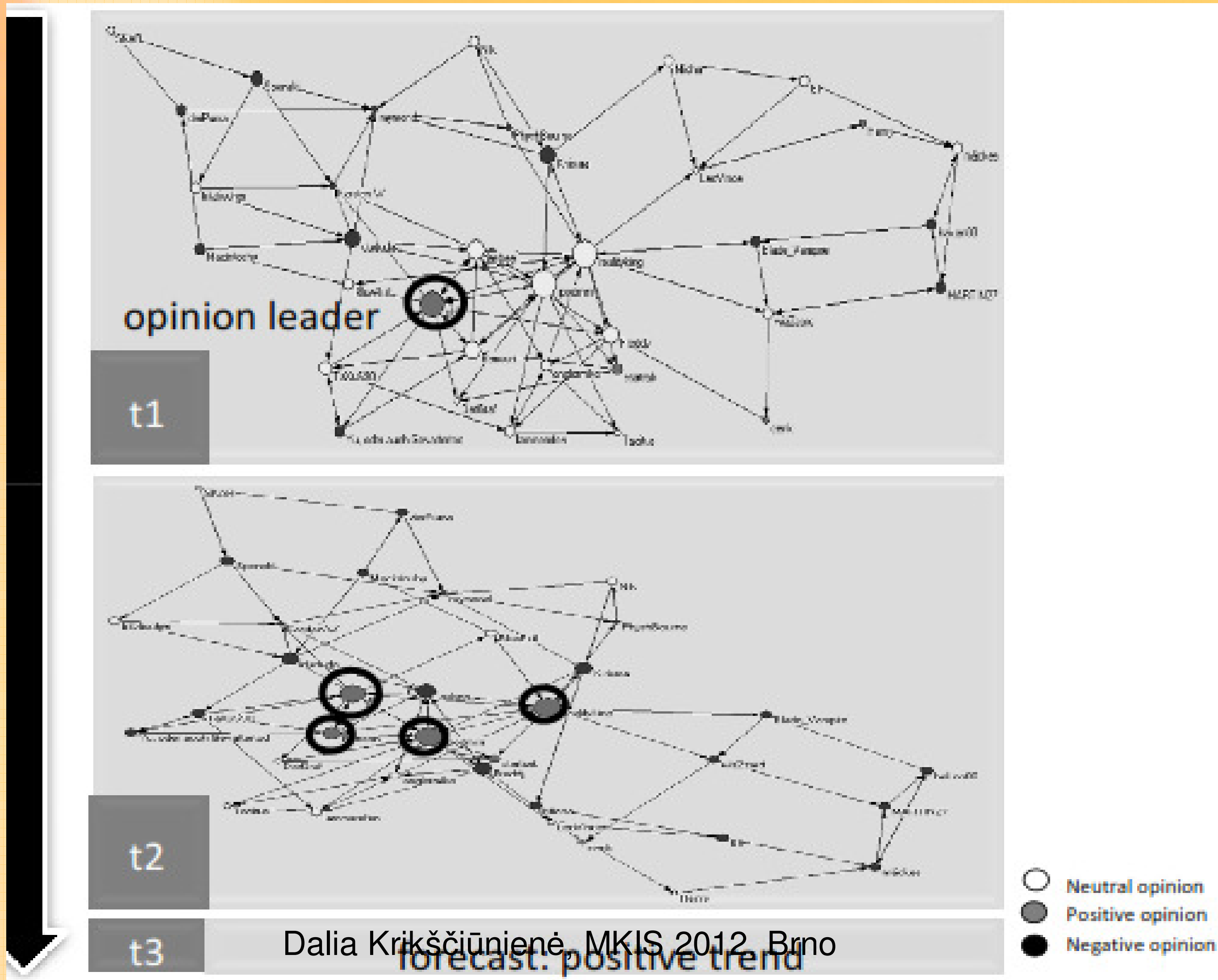
Classification (support vector machine SVM)



Social network generation for analysis



Social network analysis approach



Assignment 2

Tools &software: Sugar CRM, MS Excel pivot module, *Statistica* advanced models, *Viscovery SoMine*

2nd team assignment and lab work training:

- *Operational CRM* (Sugar CRM)
- *Analytical CRM* (CRM performance analysis by applying business intelligence approaches (pivoting, visualization) and computational intelligence methods (neural networks, fuzzy rules, Kohonen self organizing networks)

Assignment 2 – Task description

- The data file for analysis CRM_data_for_analysis.xls
- The task description is in file 2_assign_CRM_task.pdf
- The outcome – report, Excel data file and Statistica workbook file.
- https://inet.muni.cz/app/soft/licence

Literature

Berry, M.,J.A., Linoff, G.S. (2011), "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management", (3rd ed.), Indianapolis: Wiley Publishing, Inc.

(Electronic Version): StatSoft, Inc. (2012). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/

(Printed Version): Hill, T. & Lewicki, P. (2007). STATISTICS: Methods and Applications. StatSoft, Tulsa, OK.

Sugar CRM Implementation http://www.optimuscrm.com/index.php?lang=en

Statsoft: the creators of Statistica http://www.statsoft.com

Viscovery Somine http://www.viscovery.net/