Data Mining

Lecture 14
Revision

http://www.informatik.uni-hamburg.de/WTM/

Why Data Mining?

Trends Leading to
Data Flood:

- Bank, telecom, other business transactions ...
- Scientific data: astronomy, biology, etc.
- Web, text, and e-commerce

From Data to Knowledge

Medical Data by Dr. X, Tokyo Med. & Dent. Univ., 38:

10, M, 0, 10, 10, 0, 0, 0, SUBACUTE, 37, 2, 1, 0,15,-,-, 6000, 2, 0, abnormal, abnormal,-, 2852, 2148, 712, 97, 49, F,-,multiple,,2137, negative, n, n, ABSCESS,***VIRUS***

12, M, 0, 5, 5, 0, 0, 0, ACUTE, 38.5, 2, 1, 0,15, -,-, 10700,4,0,normal, abnormal, +, 1080, 680, 400, 71, 59, F,-,ABPC+CZX,, 70, negative, n, n, n, BACTERIA, ***BACTERIA***

15, M, 0, 3, 2, 3, 0, 0, ACUTE, 39.3, 3, 1, 0,15, -, -, 6000, 0,0, normal, abnormal, +, 1124, 622, 502, 47, 63, F, -,FMOX+AMK, , 48, negative, n, n, n, BACTE(E), ***BACTERIA***

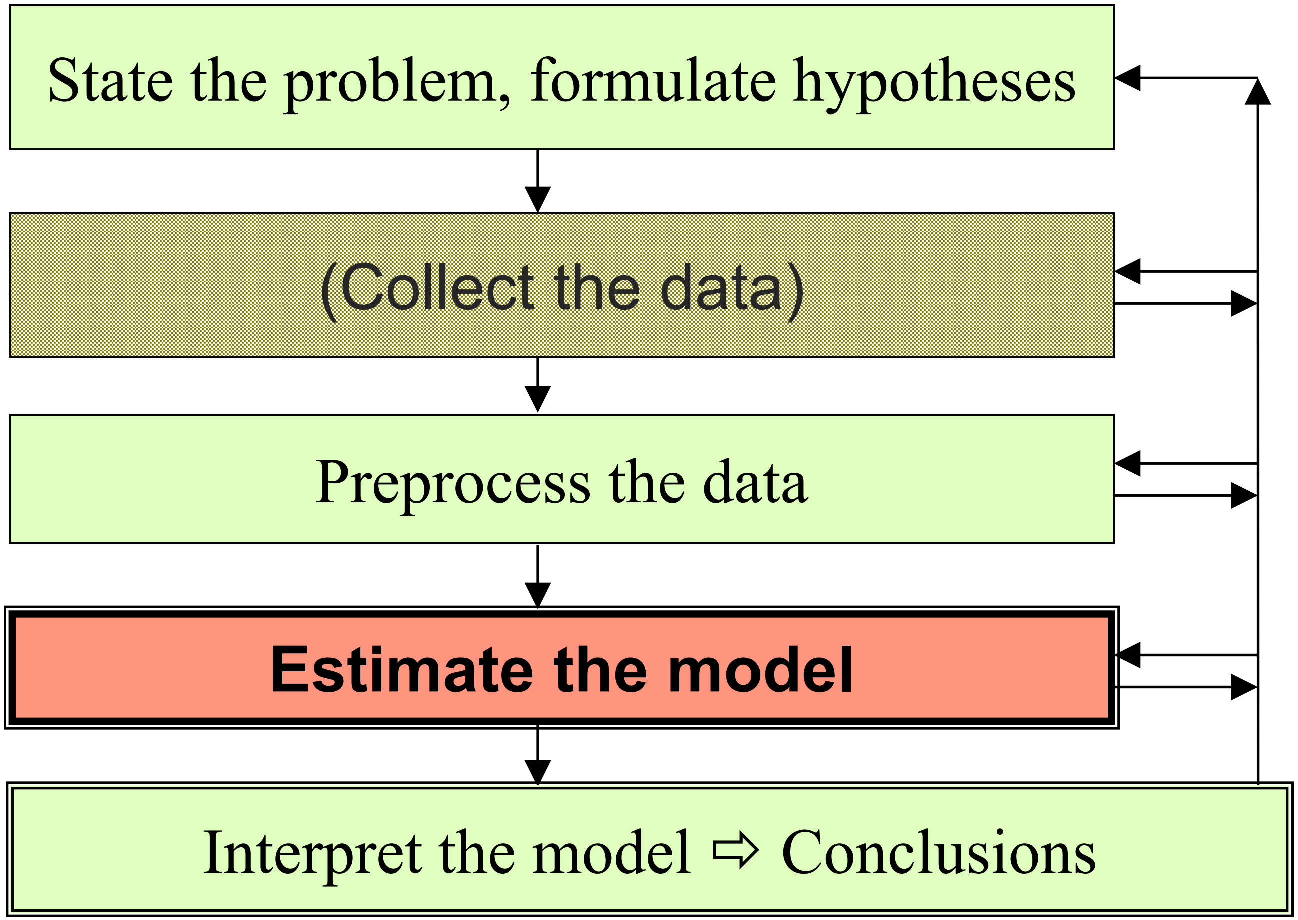
16, M, 0, 32, 32, 0, 0, 0, SUBACUTE, 38, 2, 0, 0, 15, -, +, 12600, 4, 0,abnormal, abnormal, +, 41, 39, 2, 44, 57, F, -, ABPC+CZX, ?, ? ,negative, ?, n, n, ABSCESS,***VIRUS***

Numerical attribute Categorical attribute Missing values Class labels

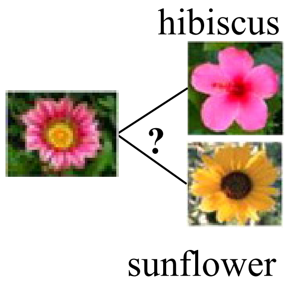
IF cell_poly <= 220 AND Risk = n AND Loc_dat = + AND Nausea > 15
THEN Prediction = VIRUS [87,5%]

Predictive accuracy

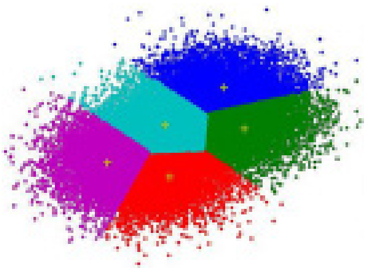
Data Mining as a simplified Process



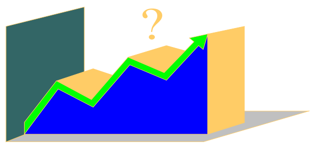
Primary Tasks of Data Mining I



- ***Classification***:
  - Find the description of several predefined classes
  - Classify a data item into one



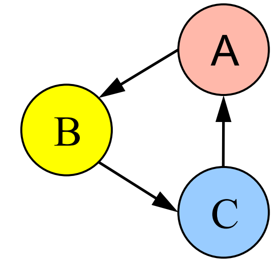
- ***Clustering***:
  - Identify a finite set of categories
  - … or clusters to describe the data



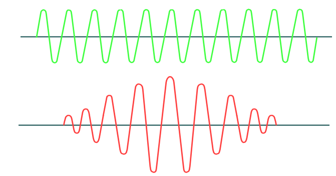
- ***Regression***:
  - Maps a data item to a real-valued prediction variable

Primary Tasks of Data Mining II

- ***Dependency modeling***:
  - Find a model that describes significant dependencies between variables



- ***Deviation*** and ***change detection***:
  - Discover the most significant changes in the data



- ***Summarization***:
  - Find a compact description for a subset of data

?

Data Mining

Lecture 2
From Data to Visualisation

http://www.informatik.uni-hamburg.de/WTM/

Attribute Types Overview

- Many types of data, e.g., numerical, text, graph, Web, image

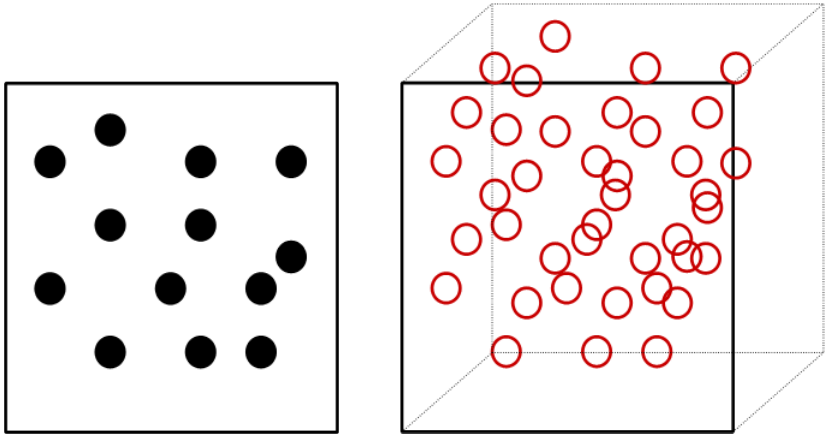
Type	Description	Examples	Operations
Nominal	Uses a label or name to distinguish objects	ZIP-Code, ID, Gender	= or !=
Ordinal	Uses values to provide the ordering of objects.	Opinion, grades	< or >
Interval	Uses units of measurements, but the origin is arbitrary.	Celsius, Fahrenheit, calendar dates	+ or -
Ratio	Uses units of measurement, the origin is not arbitrary.	Kelvin, length, counts, age, income	+, -, *, /

Curse of Dimensionality

- The size of a data set yielding the same density of data points in k-dimensional space, increases ***exponentially*** with dimensions

to achieve the same density of n points in k dimensions, we need n^k data points

Same density of data:



Low dimensions

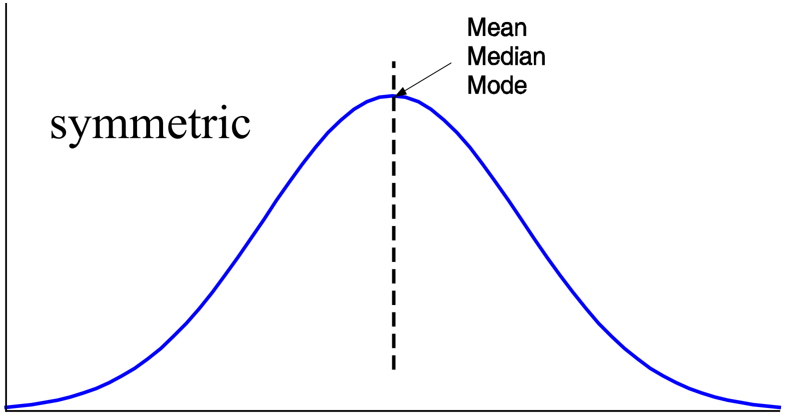
k dimensions

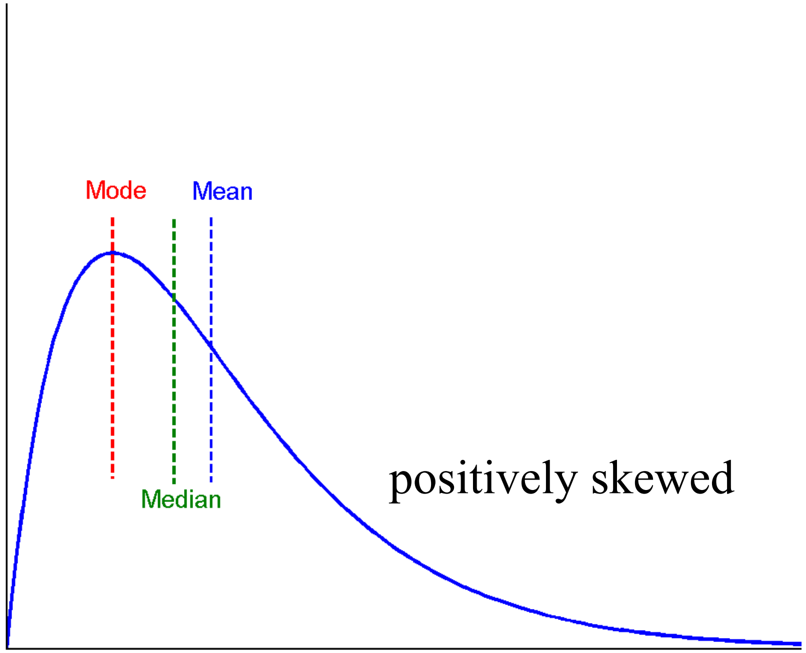
- **Example**
  - k = 1
    $\rightarrow$ n = 100 samples
  - k = 5
    $\rightarrow$ n = $100^5 = 10^{10}$ samples

Gain Insight into Data

- Statistical data ***description***: central tendency
  - Median, mean and mode; symmetric, positively and negatively skewed data
  - Quartiles and standard deviation

- Graphical displays and data ***visualization***





Data Matrix and Dissimilarity Matrix

- ***Data matrix***
  - n data points with p dimensions

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- ***Dissimilarity matrix***
  - n data points, but registers only the distance
  - A triangular matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

- Often used: ***Minkowski distance***

Data Mining

Lecture 3
Preprocessing Methods

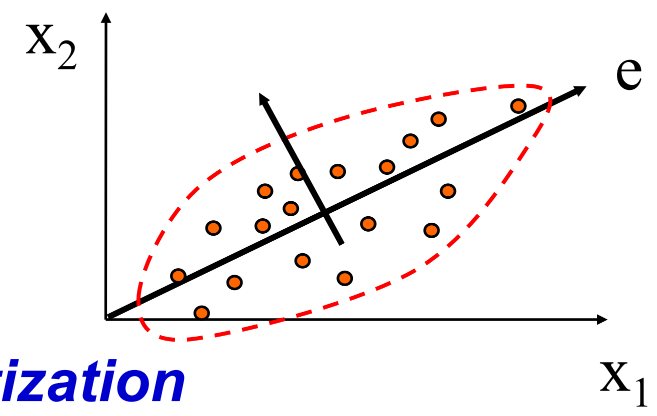
http://www.informatik.uni-hamburg.de/WTM/

Preprocessing Methods

- Data ***quality***: accuracy, completeness, consistency, timeliness, believability, interpretability
- Data cleaning: e.g. missing/noisy values, outliers
- Data ***integration*** from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Data ***reduction***
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression

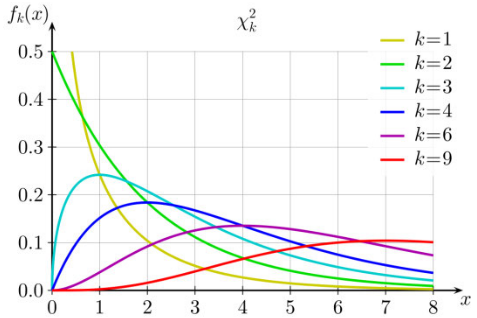


- Data ***transformation*** and data ***discretization***
  - Normalization

Correlation Analysis (nominal Data)

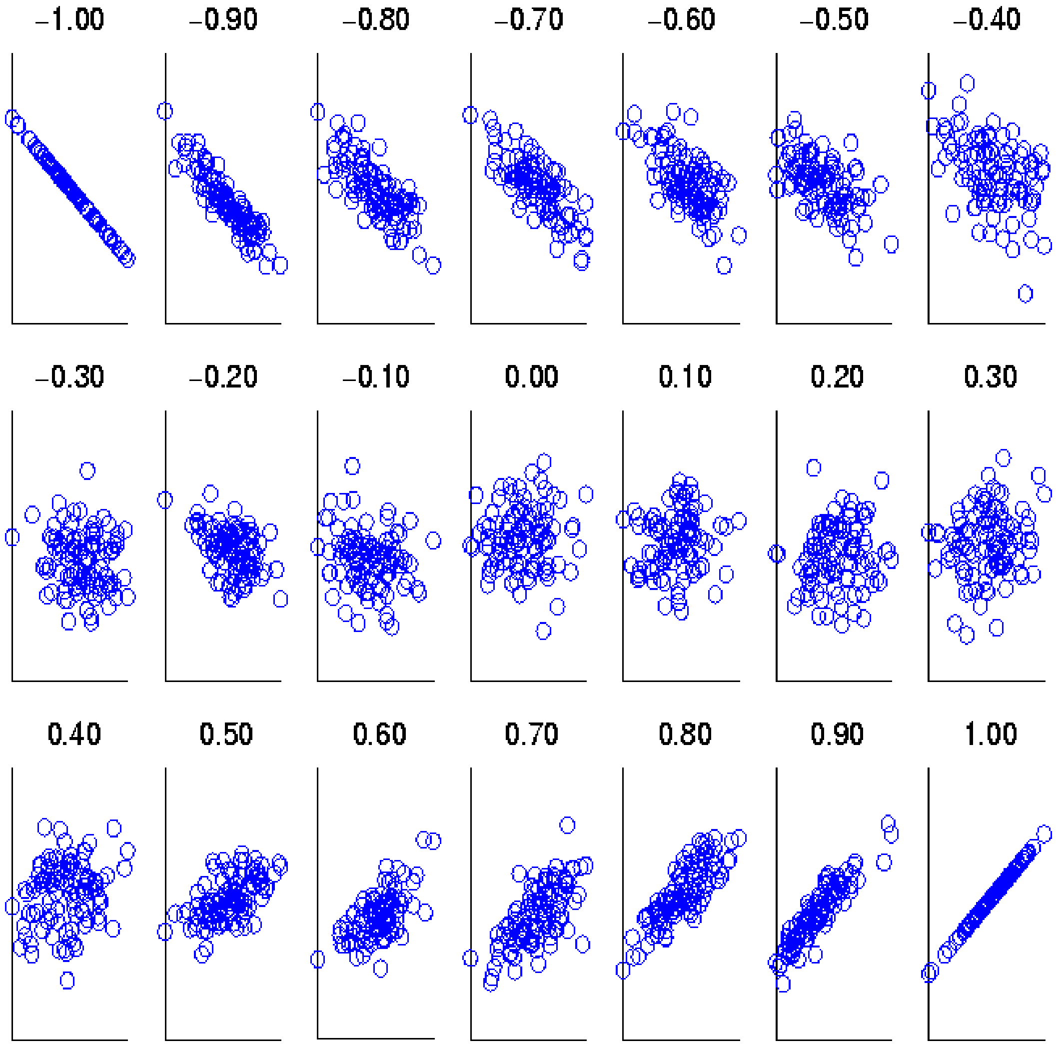
- ***X^2 (chi-square) test***

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$



- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

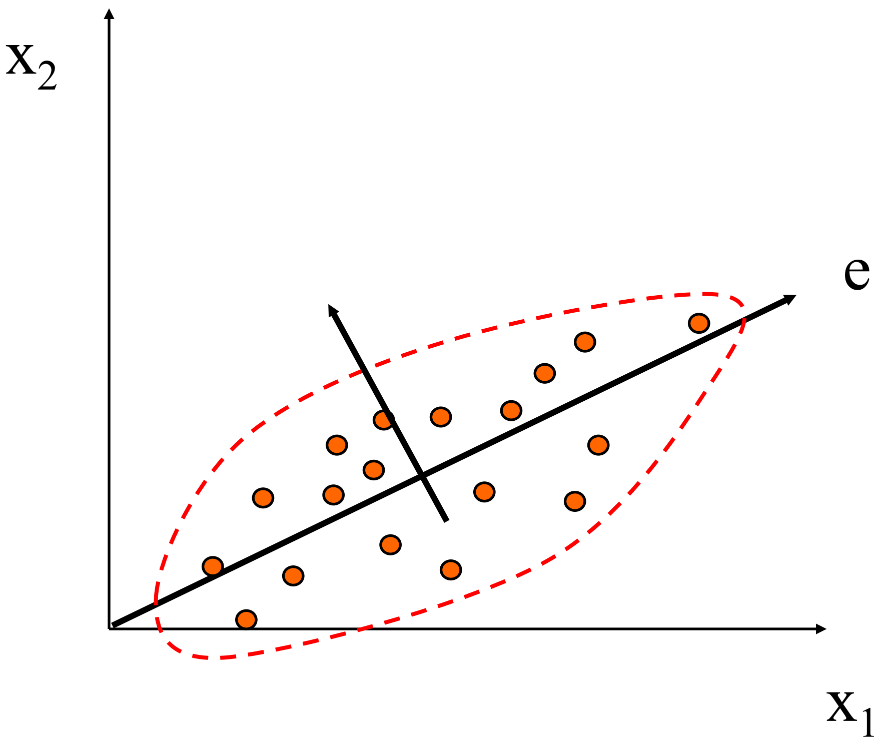
Visually evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



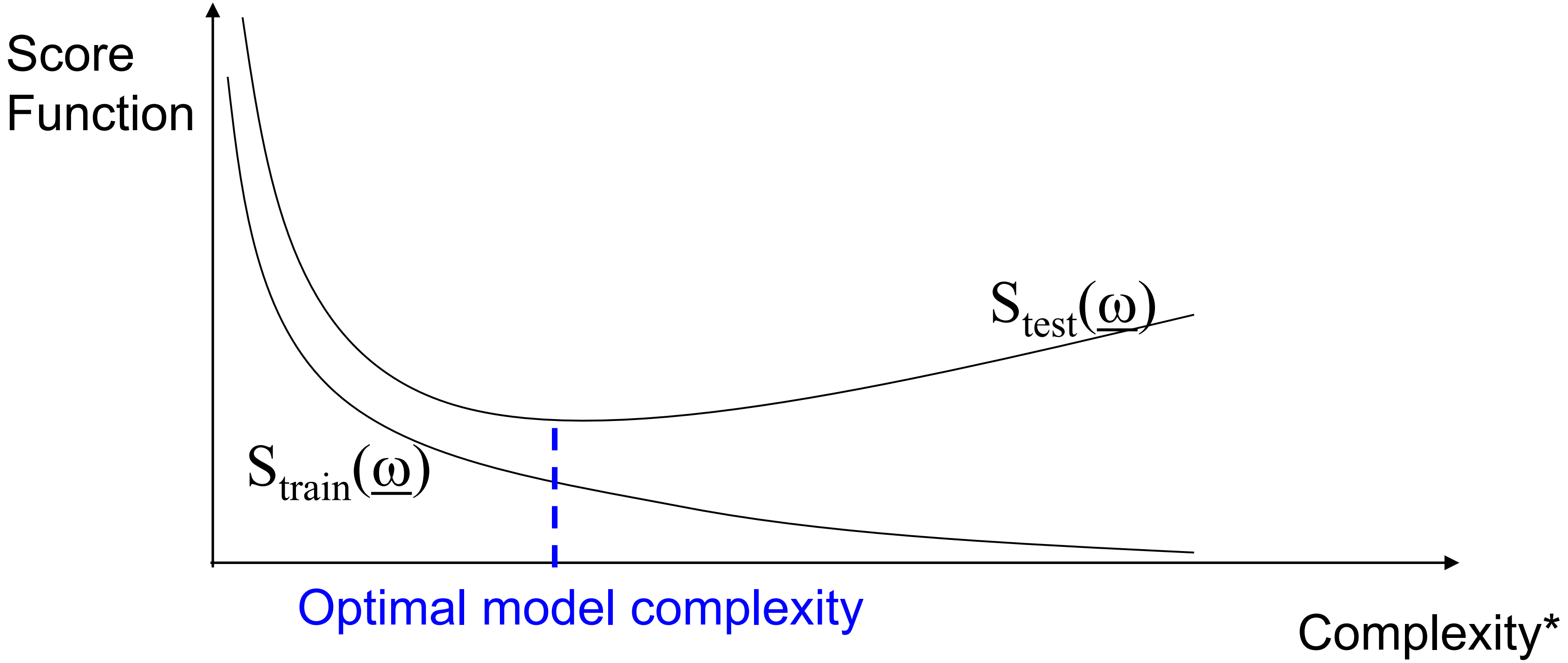
Data Mining

Lecture 4

Learning from Data towards Data Warehouses

http://www.informatik.uni-hamburg.de/WTM/

Complexity and Generalization



- *Complexity = degrees of freedom in the model

  e.g. number of variables
- cf. Vapnik Chervonenkis dimension

The Confusion Matrix

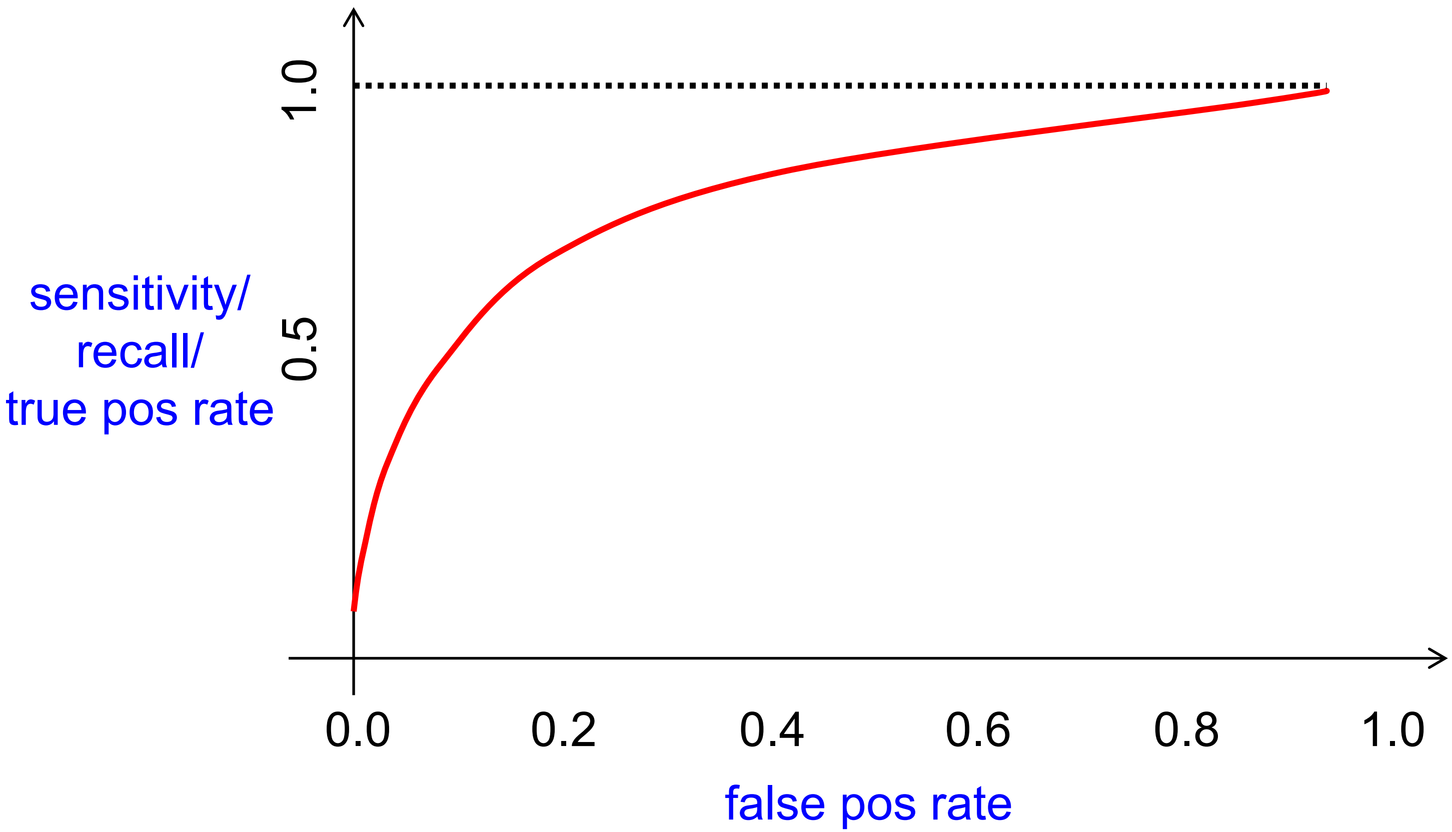
Actual **Predicted**	Class 1	Class 2
Class 1	A: True Positive	B: False Positive
Class 2	C: False Negative	D: True Negative

- Evaluation metrics:
  - **Accuracy** A = (A+D)/(A+B+C+D)
  - True positive rate TPr = A/(A+C) = 1- false negative rate = Sensitivity
  - False positive rate FPr = B/(B+D) = 1- true negative rate
  - Specificity SP = 1 - FPr
  - **Recall** R = A/(A+C)
  - **Precision** P = A/(A+B)
  - **F-score** F = 2 P R / (P+R)

different in Kantardzic book!

- Use evaluation metrics for ***model selection*** via Holdout method; random subsampling; Cross-validation; Bootstrap

Receiver Operating Characteristic (ROC)



- measures overall model performance

The Apriori Algorithm – an Example

$Sup_{min} = 2$

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan for count

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

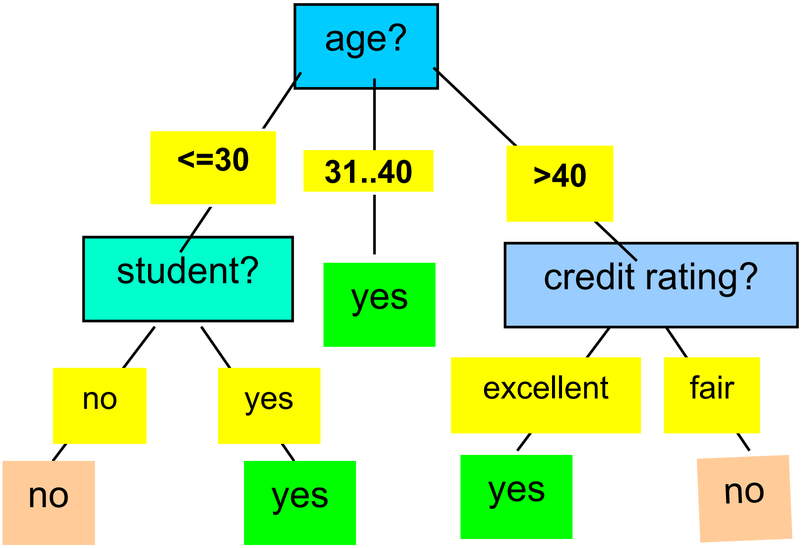
Data Mining

Lecture 5
Decision Trees and Classification

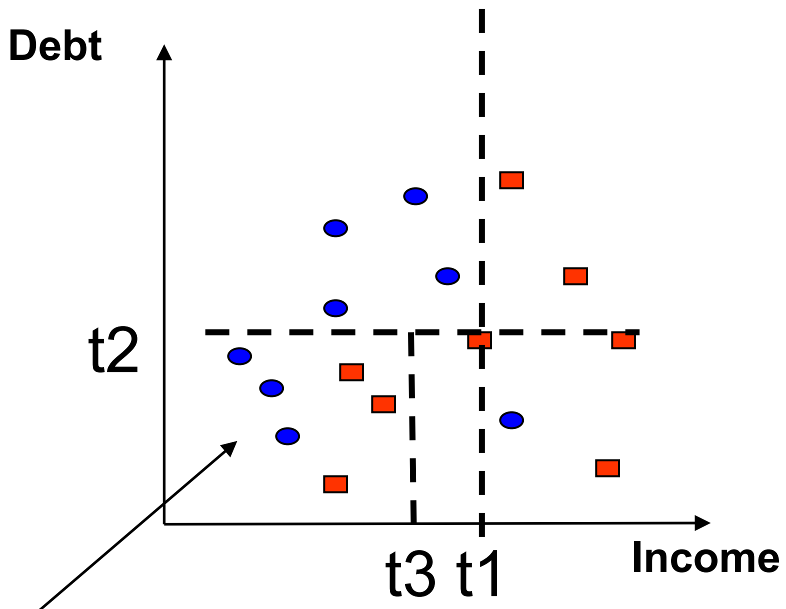
http://www.informatik.uni-hamburg.de/WTM/

Decision Trees and Classification

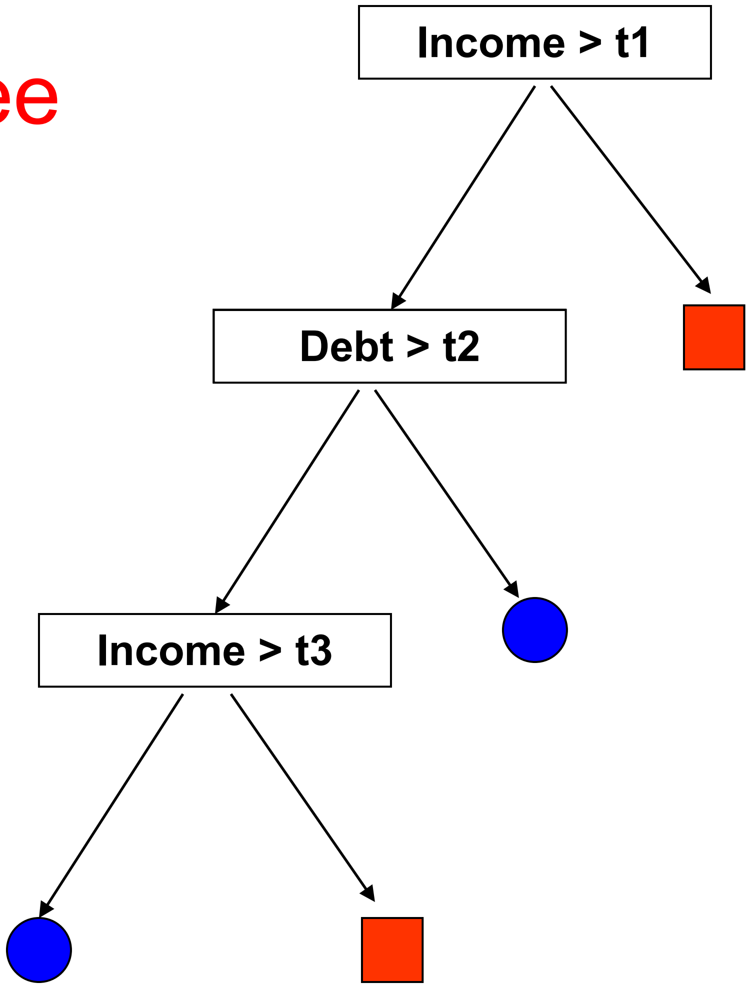
- Classification – a Two-Step Process
  - Model construction
  - Model usage
- Decision Tree Induction
  - Supervised learning
  - Rule extraction
- Overfitting and its avoidance
  - Tree Prepruning
  - Tree Postpruning



Decision Tree



boundaries are piecewise linear and axis-parallel



Decision Trees handle high-dim space and missing values, are easy to implement (no geometry), may yield intuitive rules, discover important rule first

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- ***Information*** (entropy) to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- ***Information needed*** (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- ***Information gained*** by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

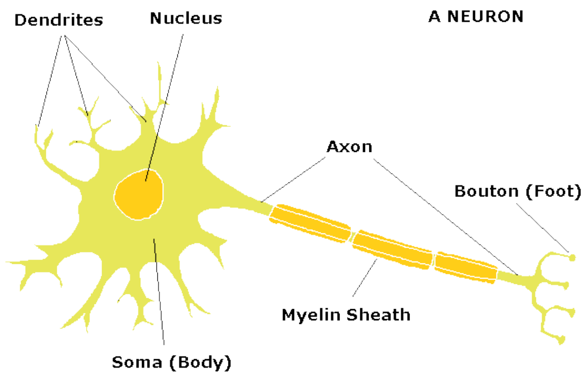
Data Mining

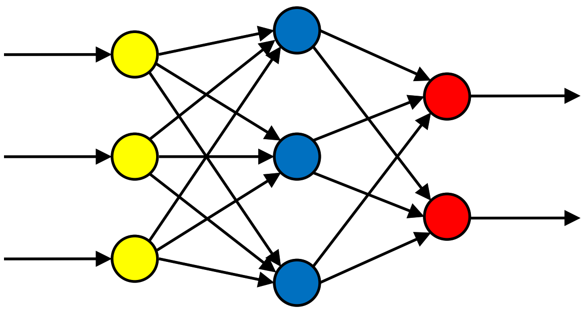
Lecture 6
Classification with Supervised Neural Networks

http://www.informatik.uni-hamburg.de/WTM/

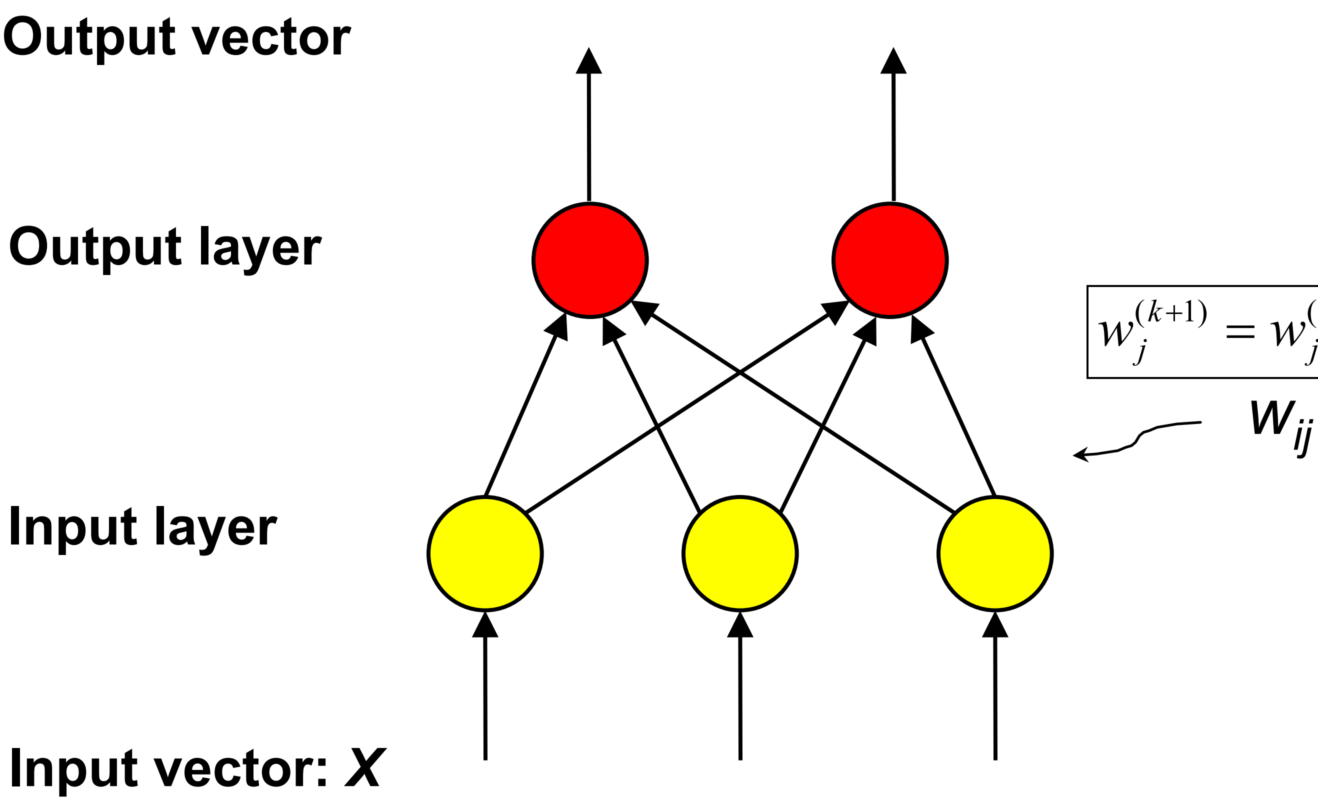
Classification with Supervised Neural Networks

- A neural network: A set of connected input/output units where each connection has a weight
- The network ***learns by adjusting the weights*** so it can predict the correct class label of the input tuples
- ***"connectionist learning"***



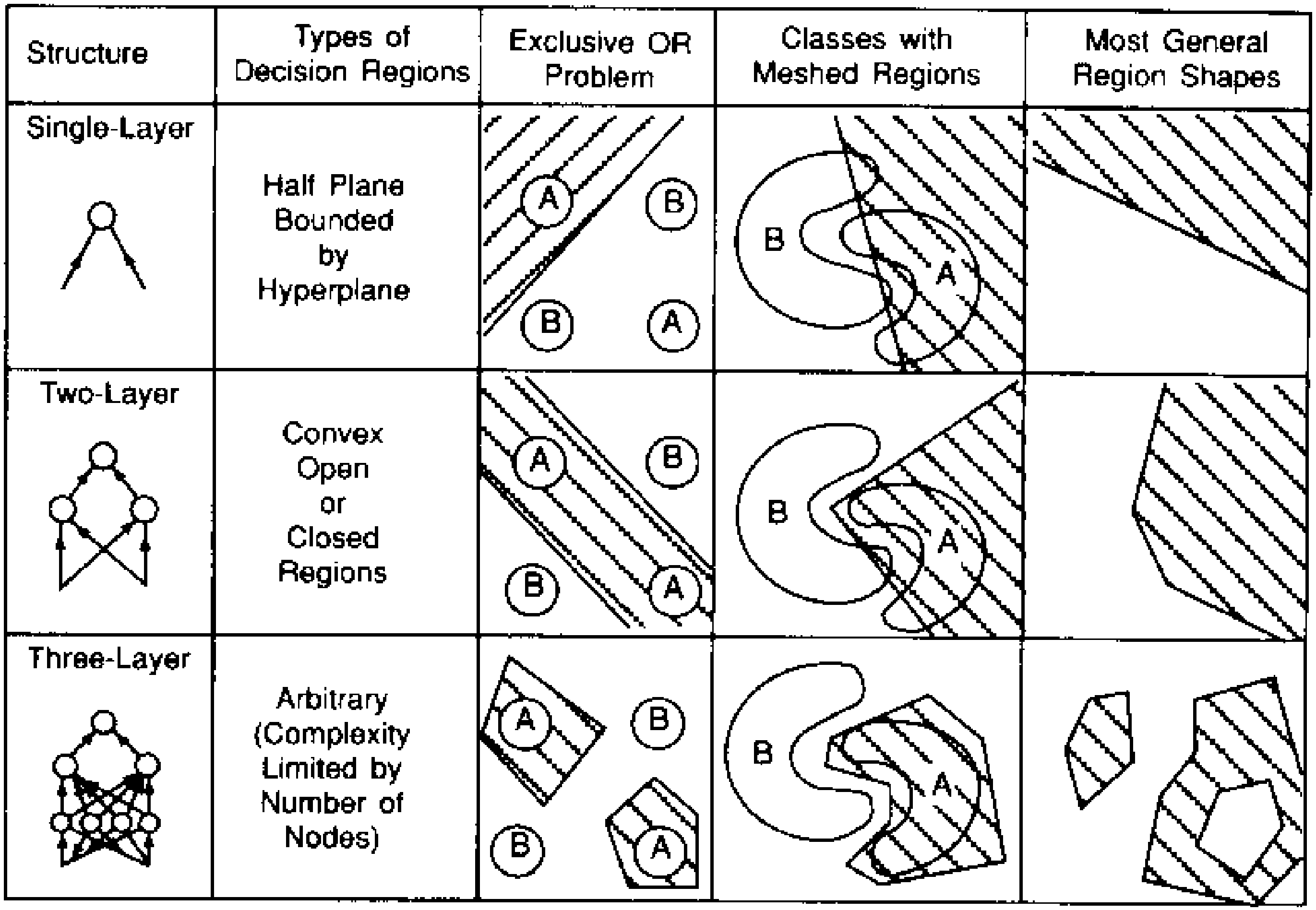


Perceptron Network



$$\boxed{w_j^{(k+1)} = w_j^{(k)} + \lambda(y_i - \hat{y}_i^{(k)})x_{ij}}$$

Decision Boundaries (Lippmann)



Data Mining

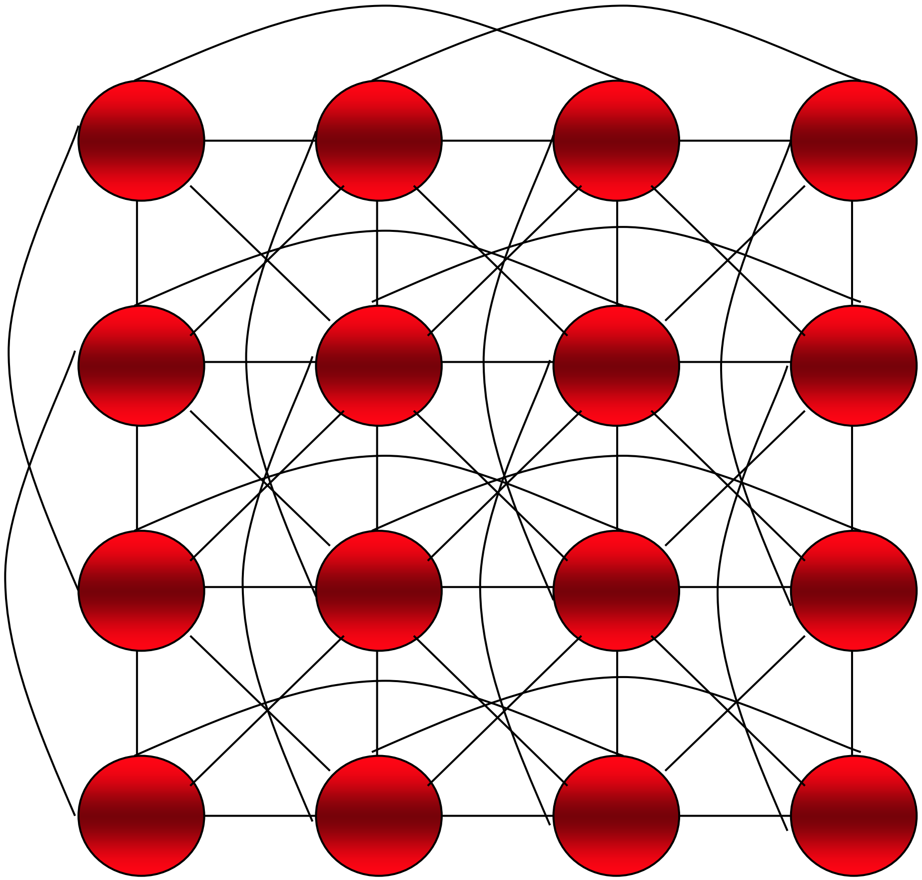
Lecture 7
Associative Networks and Recurrent Classification

http://www.informatik.uni-hamburg.de/WTM/

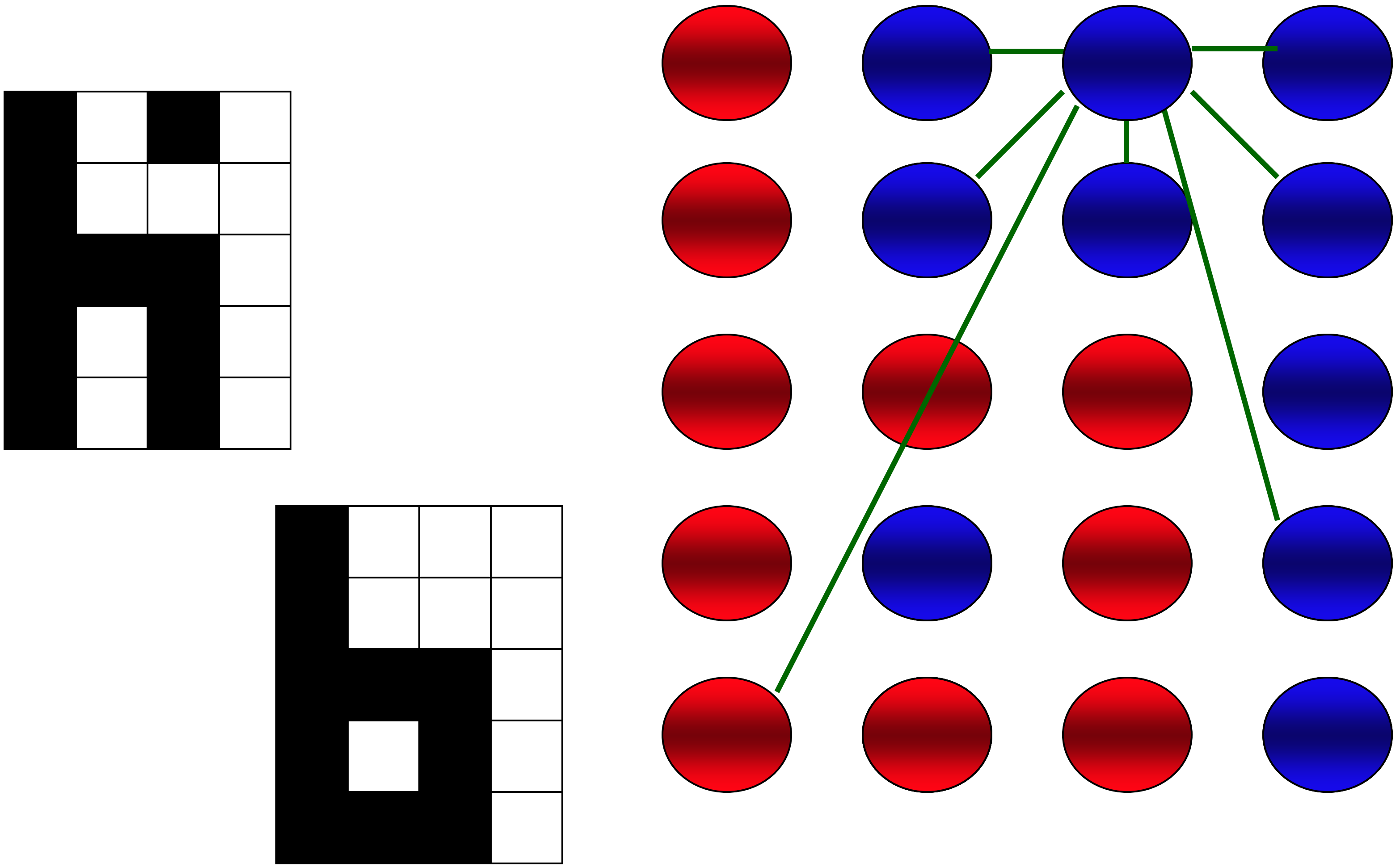
The Hopfield Network

- All connected to every other neuron
- Synchronous or random update

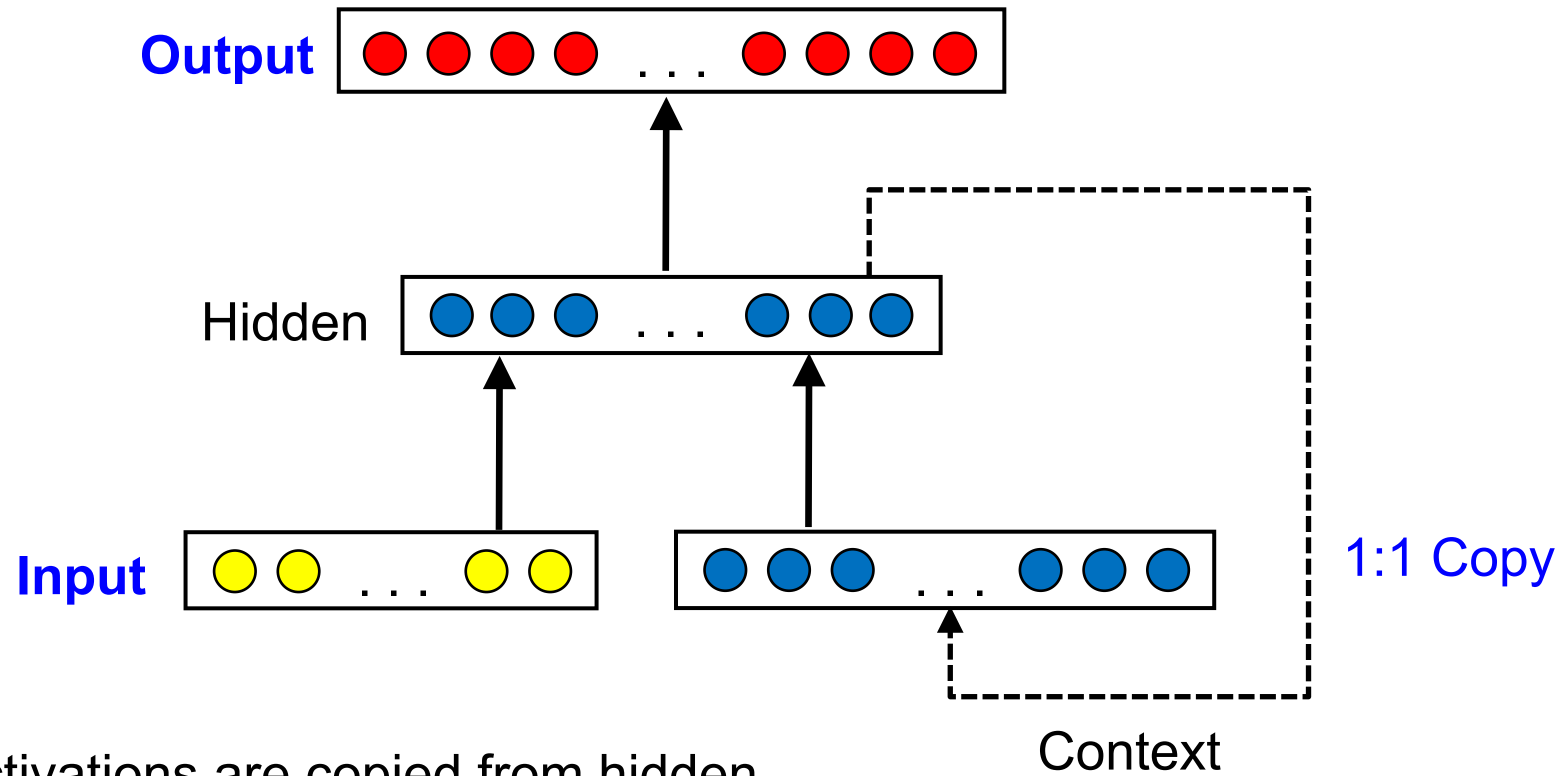
$$s_i = \text{sign}\left(\sum_{j=1}^{n} w_{ij} s_j\right)$$



Using the Memory



Simple recurrent network (SRN)



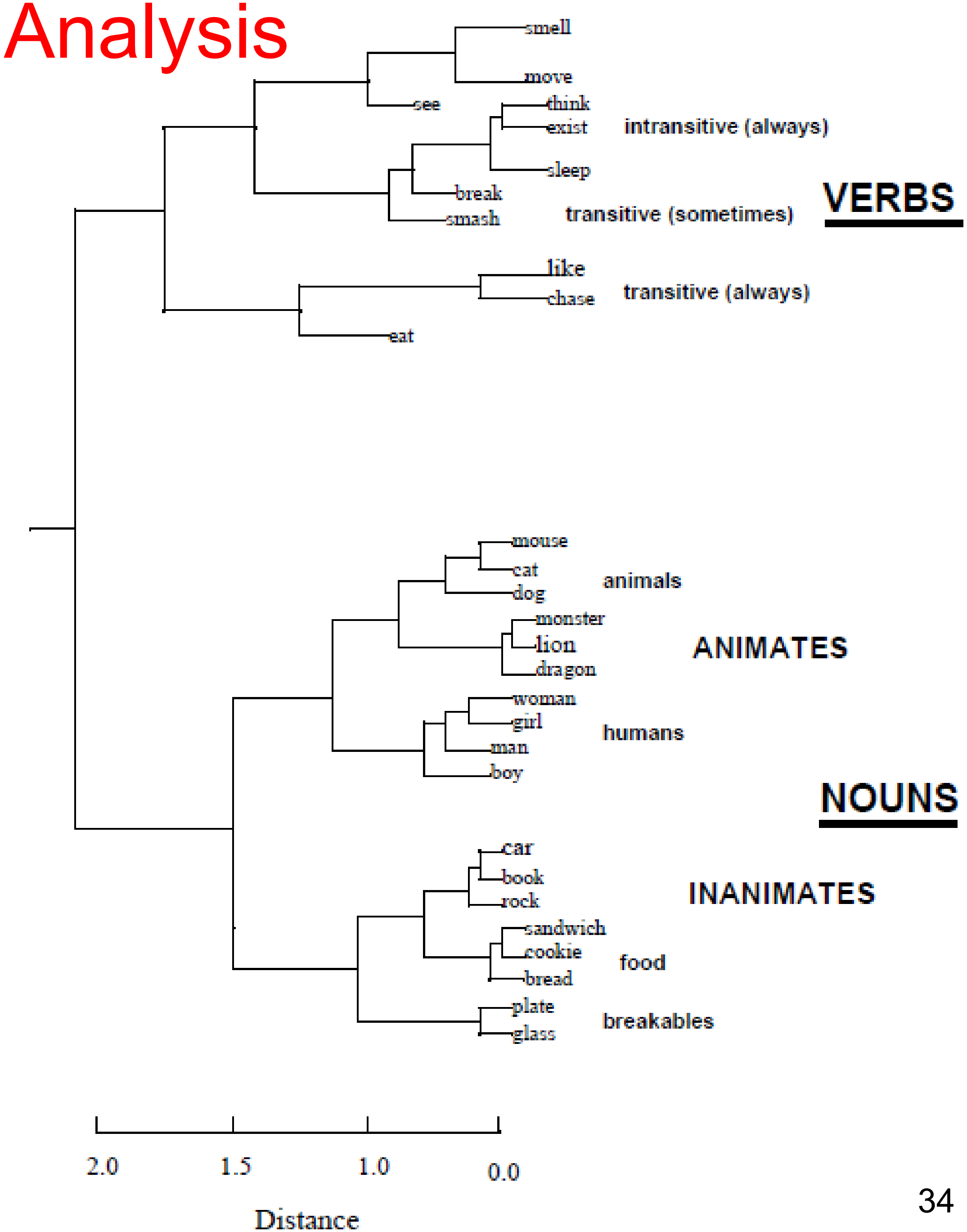
- Activations are copied from hidden layer to context layer
- Straight lines represent trainable connections

Example Prediction

Input: x_1 x_2 x_3 .... x_t

Output: x_2 x_3 x_4 .... x_{t+1}

Hierarchical Cluster Analysis of Hidden Layers



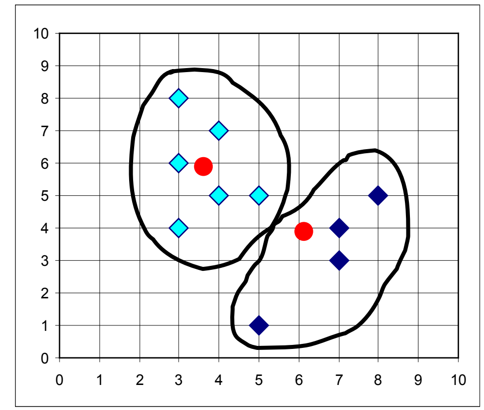
Data Mining

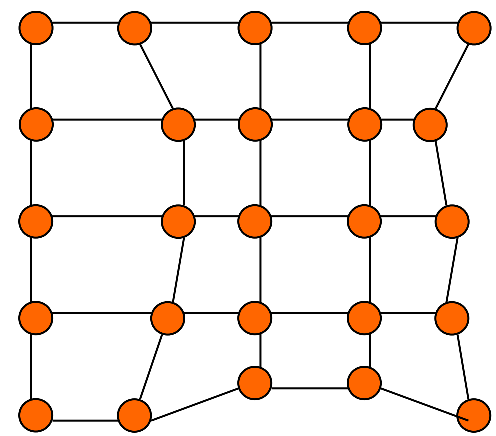
Lecture 8
Clustering and Selforganizing Networks

http://www.informatik.uni-hamburg.de/WTM/

Clustering and Selforganizing Networks

- ***Cluster analysis*** groups objects based on their ***similarity***
- Measure of similarity can be computed for ***various types of data***
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- ***Outlier detection*** and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches





K-means and SOM: `Cost Functions´

- K-means:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2$$

- SOM:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \sum_{j}^{k} h(|i - j|)(p - m_j)^2$$

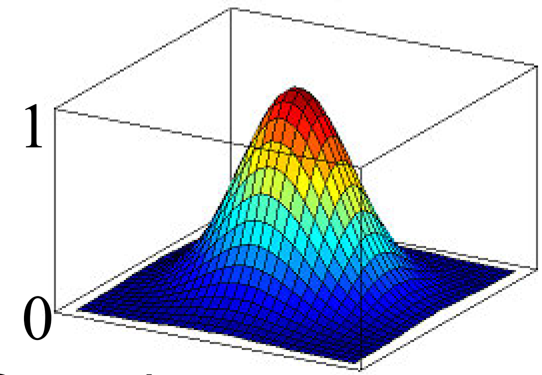
neighbourhood activation function h

Neighborhood Function Preserves Topology

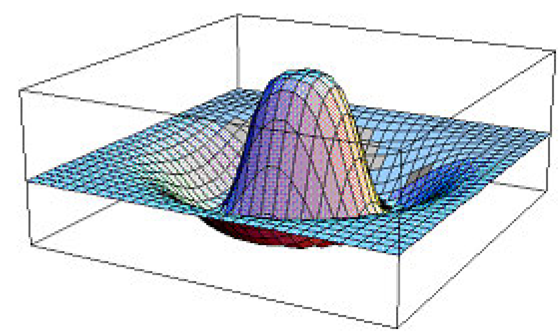
- The neighborhood function $h(n_b,t)$ determines the degree of weight vector change of the neighbors

$$w_j^T \leftarrow w_j^T + \eta(t) \cdot \boxed{h(n_b,t)} \cdot \left(x - w_j^T\right)$$

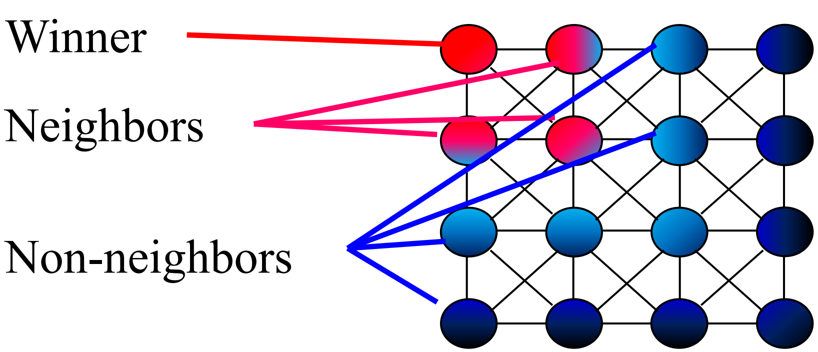
  - Mostly: Gaussian function
    rarely: Mexican Hat function
  - Width decreases during training
    ($\rightarrow$ implicit decrease of learning rate)
  - *May* decrease to zero ($\rightarrow$ k-means)



Gaussian
(not normalized)



Mexican Hat
(Difference of Gaussian)

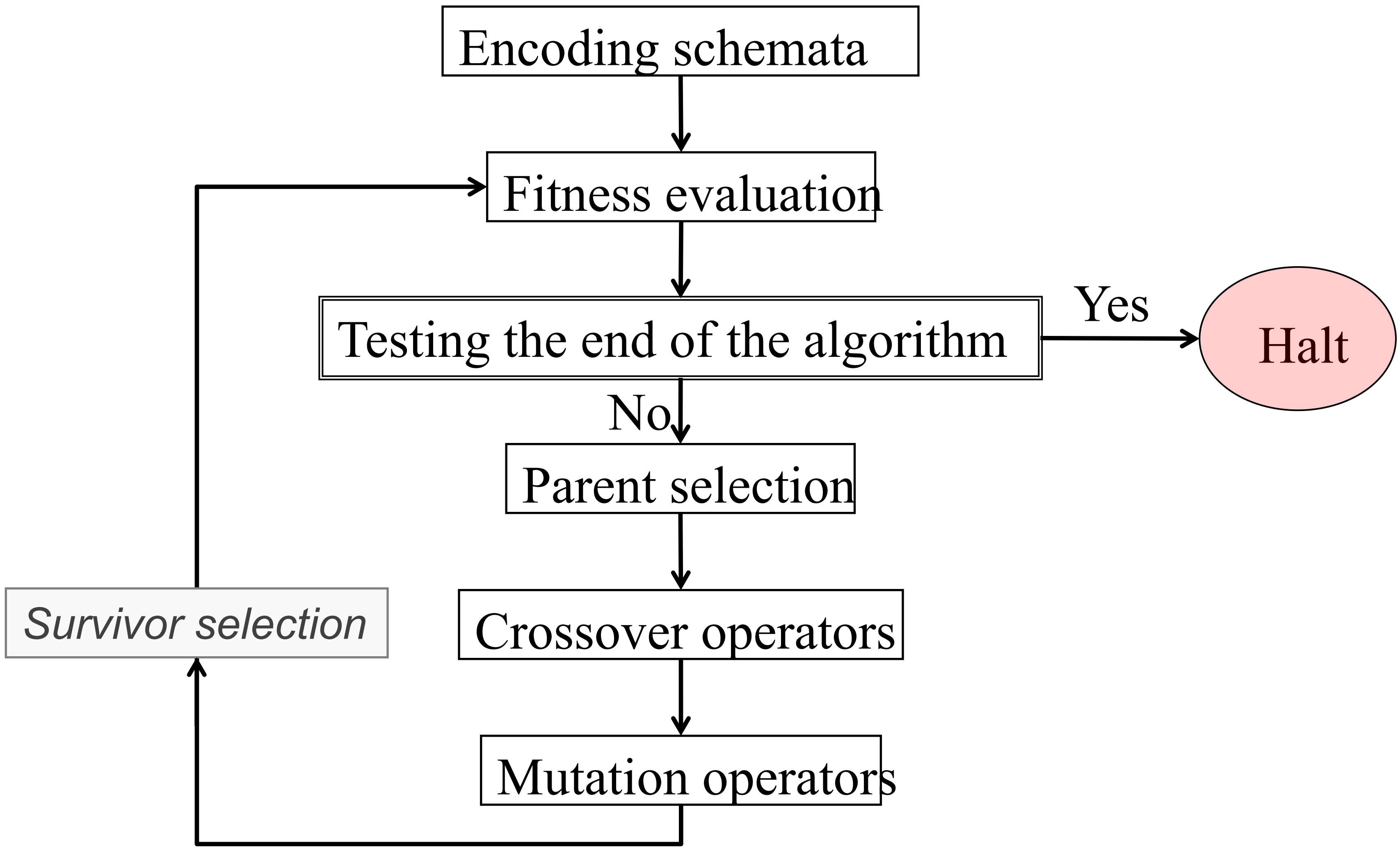


Data Mining

Lecture 9
Genetic and fuzzy mining

http://www.informatik.uni-hamburg.de/WTM/

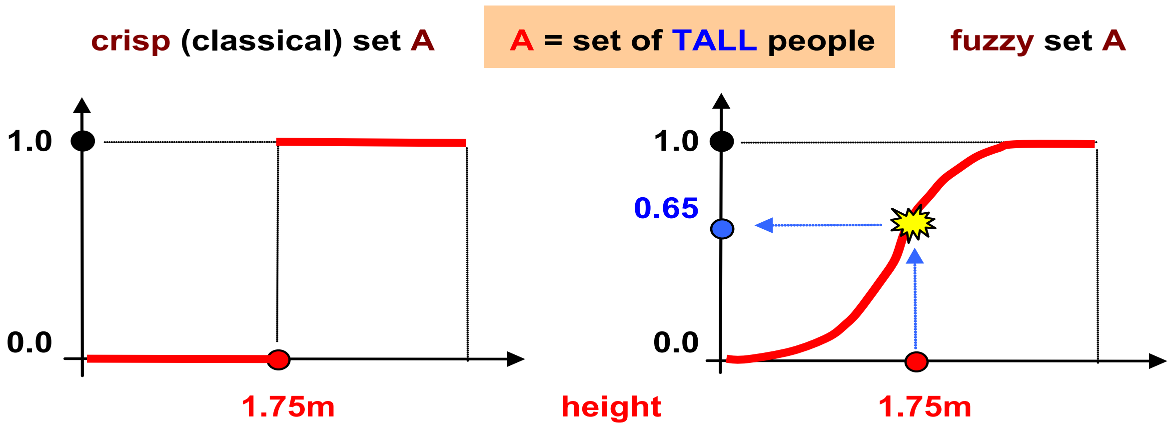
Major Phases of a Genetic Algorithm



Fuzzy Logic

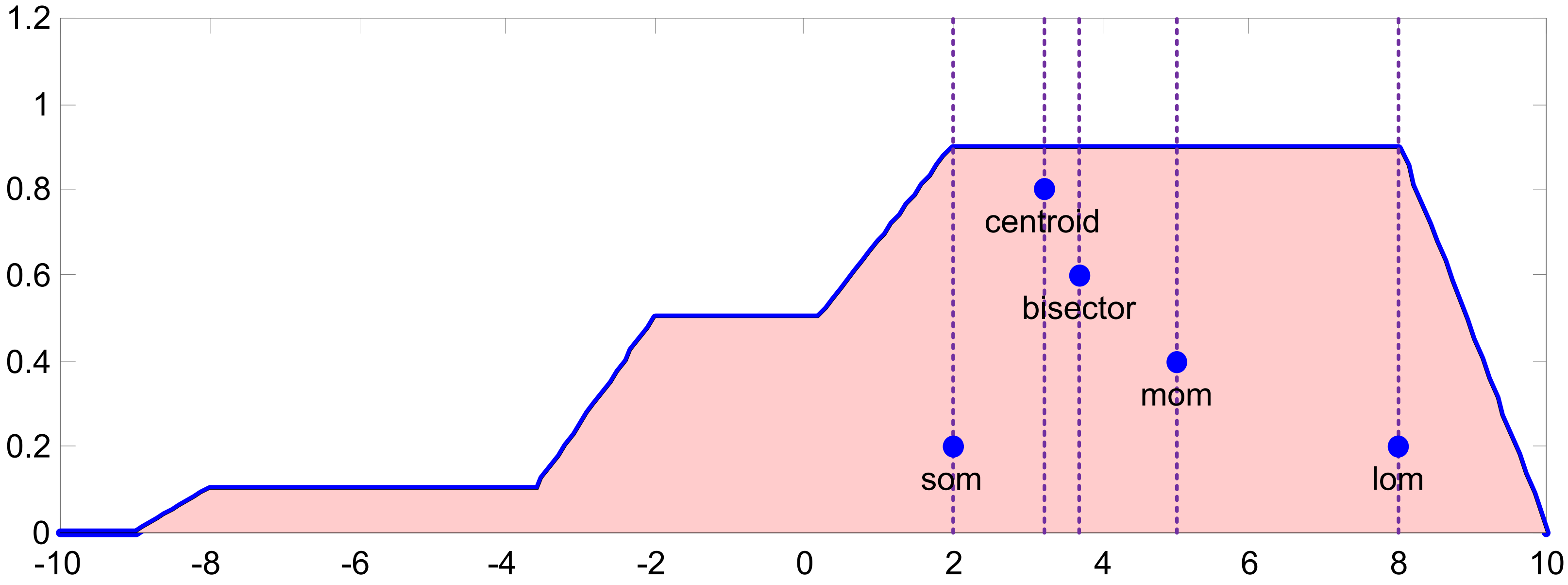
- Fuzzy logic:
  - Describes ***imprecision*** or ***vagueness***
  - Values in the range of [0,1]

- Fuzzy Set A is a universal set U determined by a membership function $\mu_A(x)$ that assigns to each element $x \in U$ a number A(x) in the unit interval [0,1]

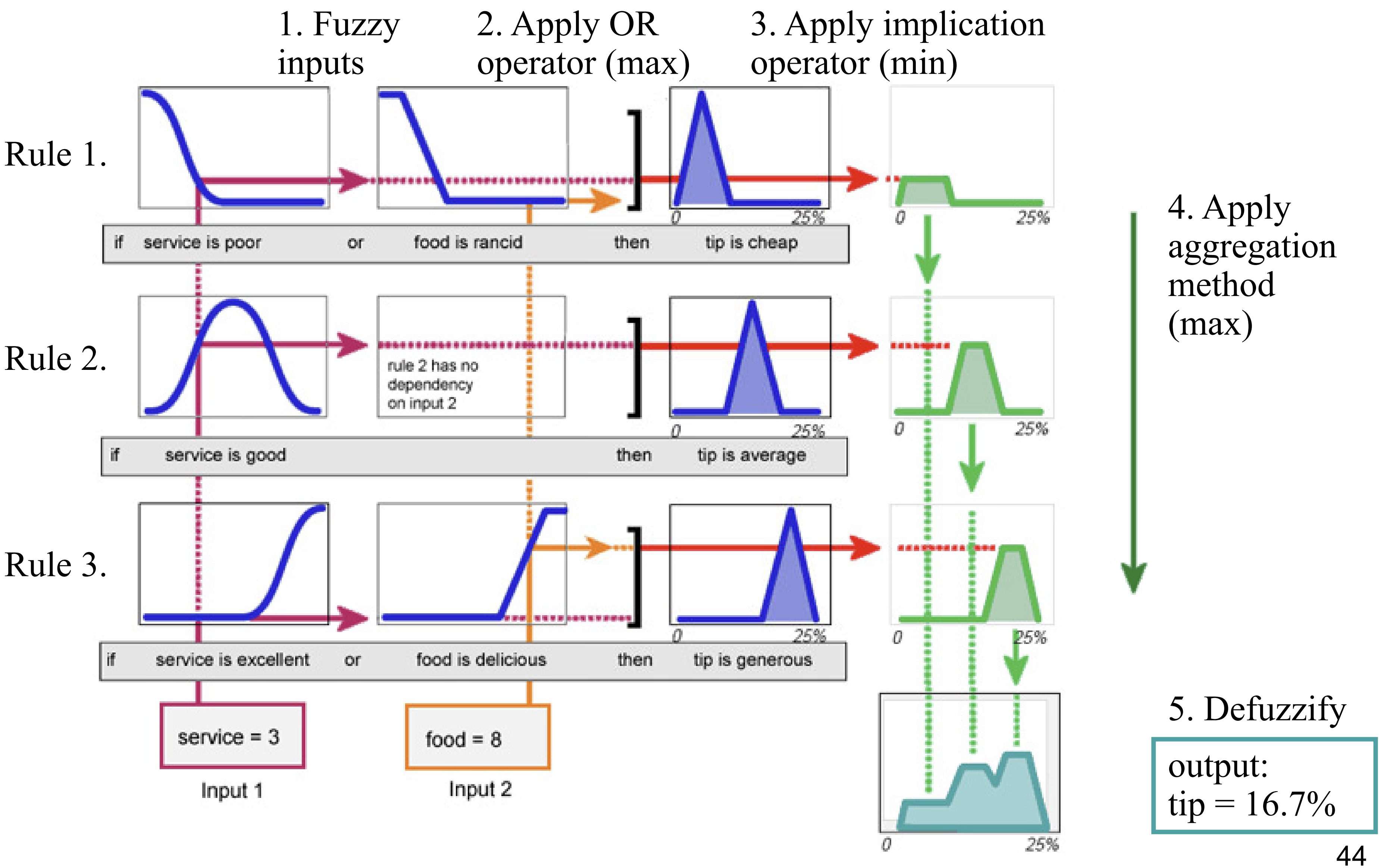


Defuzzification Methods

- Transforms fuzzy output of the inference engine to crisp output using membership functions analogous to the fuzzifier
- Commonly used techniques:
  - ***centroid*** of area
  - ***bisector*** of area
  - ***mom***: mean of maximum
  - ***som***: smallest of maximum
  - ***lom***: largest of maximum
  - …



Fuzzy Inferencing: Mamdani's Method



Data Mining

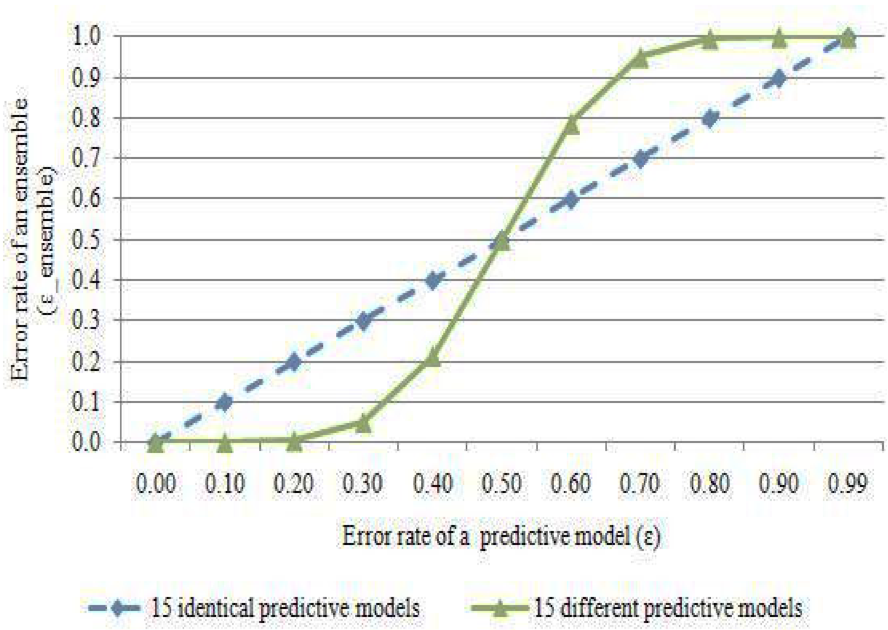
Lecture 10
Ensemble Learning

http://www.informatik.uni-hamburg.de/WTM/

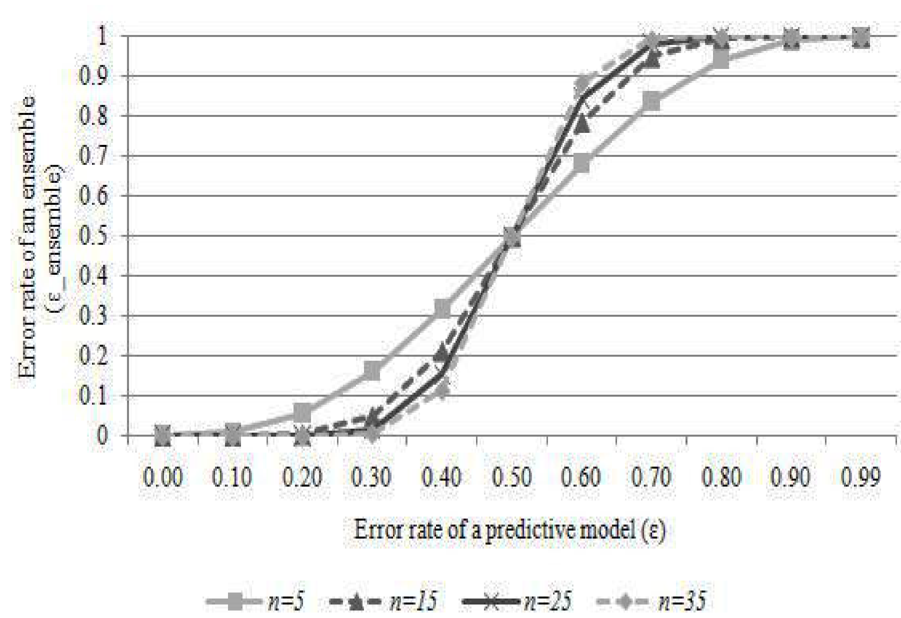
Ensembles Give Better Results

- Majority vote of *n*=15 classifiers, error rate each ε=0.3:

$$\varepsilon_{ensemble} = \sum_{i=8}^{15} \binom{15}{i} \cdot \varepsilon^{i} (1-\varepsilon)^{15-i} = 0.05$$



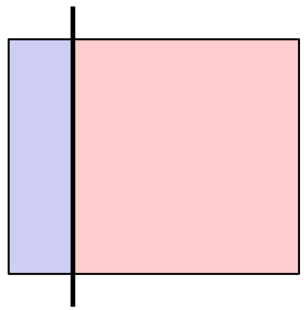
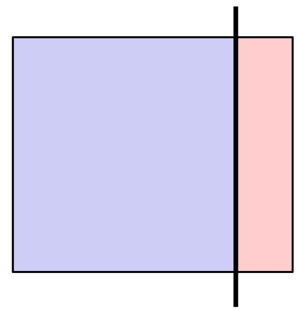
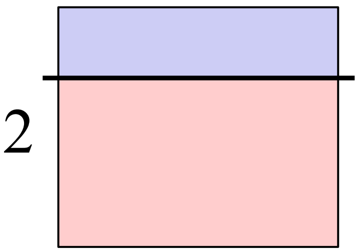
(a) Identical predictive models vs. different predictive models in an ensemble

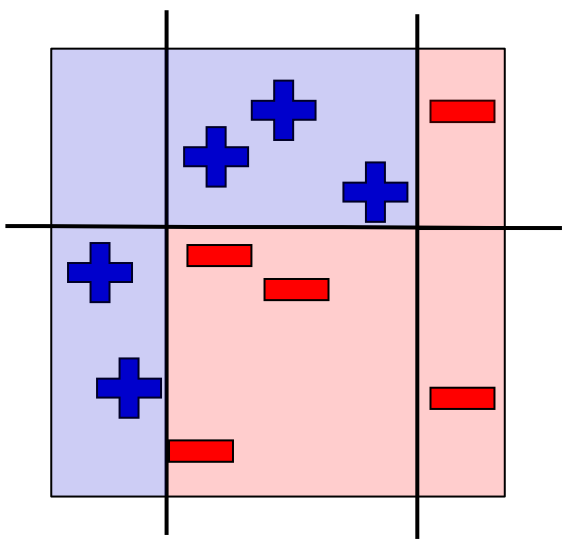


(b) The different number of predictive models in an ensemble

AdaBoost

- Final classifier:

$H_{final} = sign\Big(0.42$  $+0.65$  $+0.92$  $\Big)$

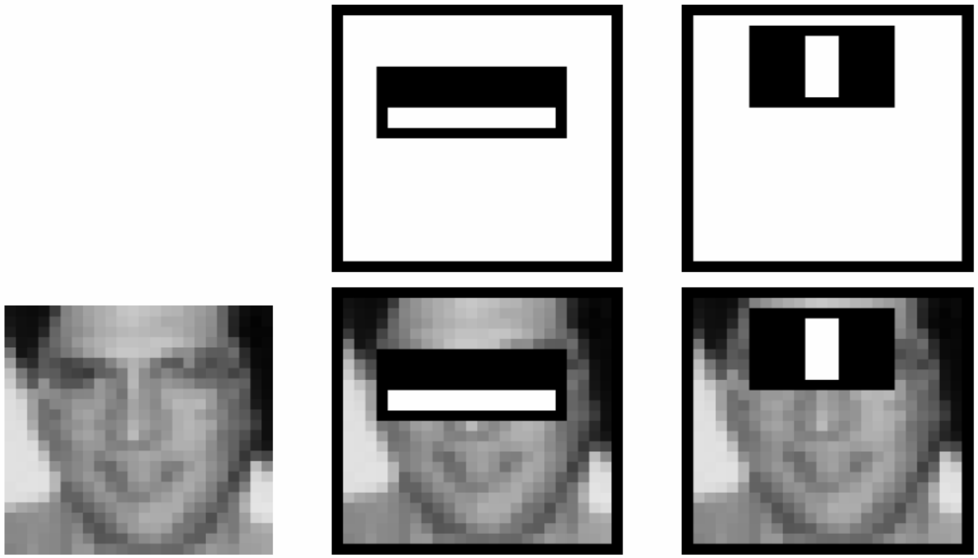
$=$ 

Many variants of AdaBoost exist depending on:

- how to set the weights ε of the data during ***learning***
- how to set the weights α to combine the hypotheses for ***classification***

Boosting for Face Detection

- First two features (weak classifiers) selected by boosting:



- This feature combination can yield 100% detection rate, however, while also finding many of false positives

Data Mining

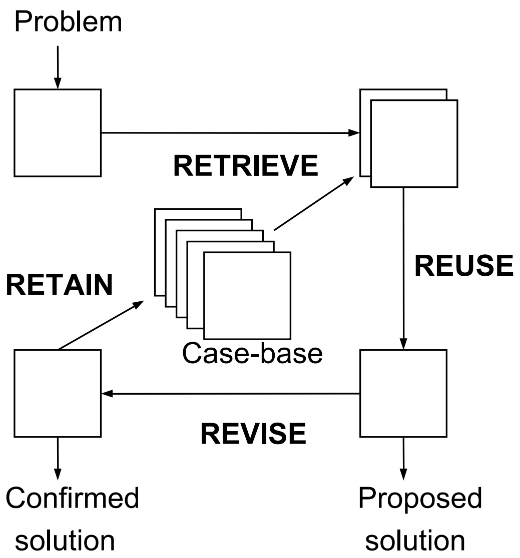
Lecture 11

Mining Structure from Graphs and High-Dimensional Data

http://www.informatik.uni-hamburg.de/WTM/

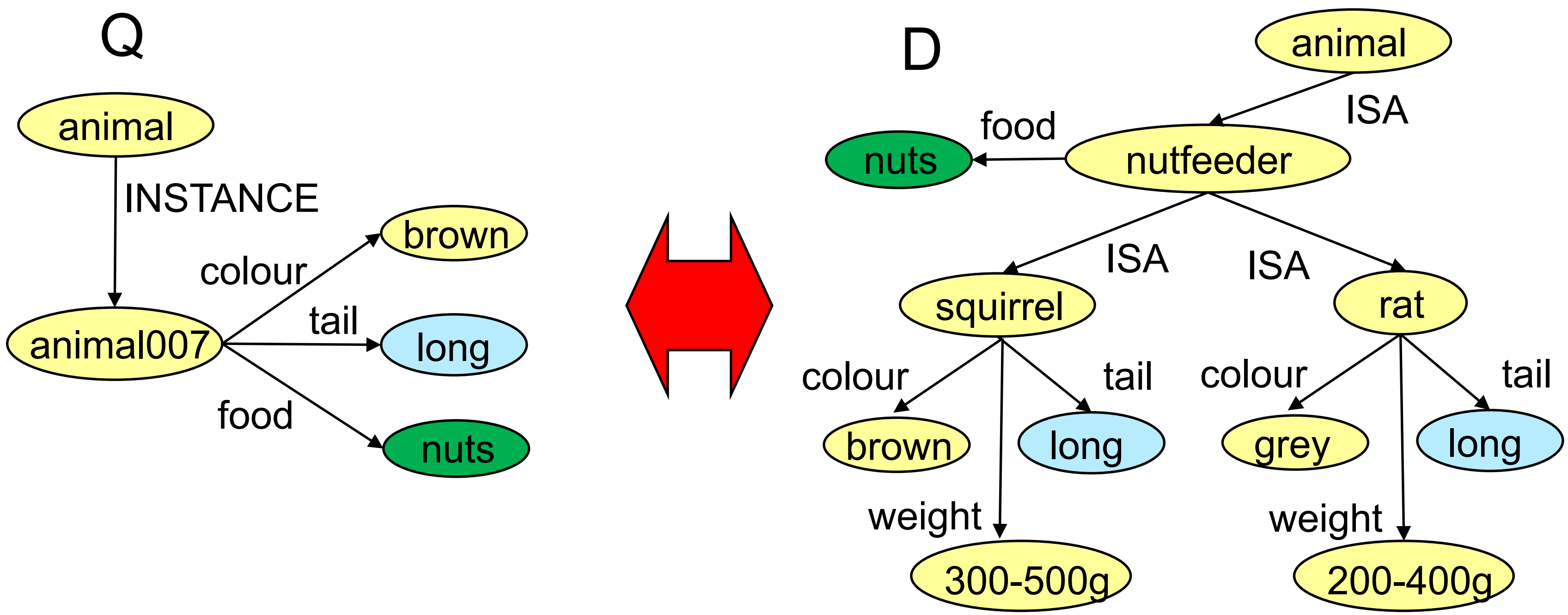
Case based Reasoning

- Provides an automated method for ***storing experience*** and reusing it to ***make decisions*** in the future
- Index vocabulary for most important features
- Applications:
  - Medicine (diagnosis)
  - Law (precedence)
  - Financial and Management (prediction)
  - Oil drilling (risk assessment)



Semantic Networks

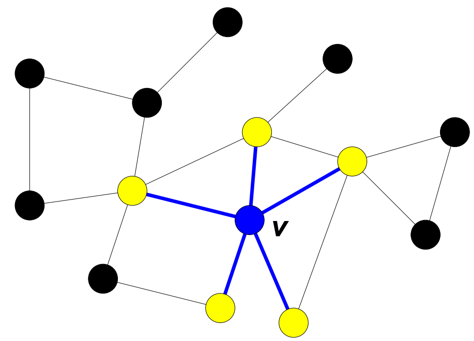
- Represents ***domain specific*** knowledge
- Models ***concepts*** & ***inheritance*** relations , e.g. INSTANCE and ISA



- Classification by relational matching of query object Q to database D

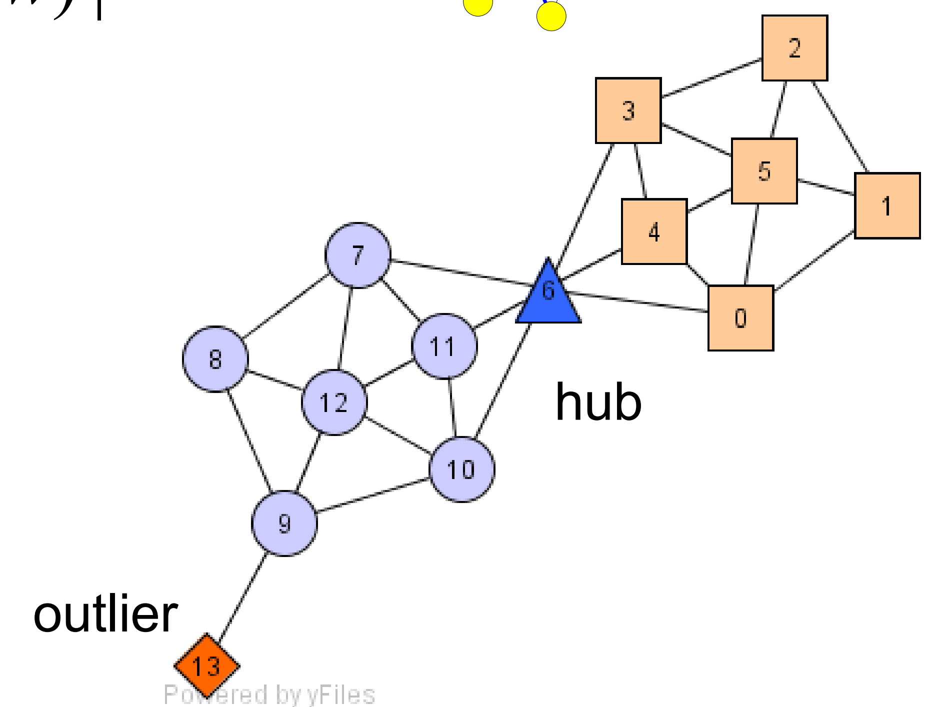
Structure Similarity

- The desired features tend to be captured by a measure we call Structural Similarity



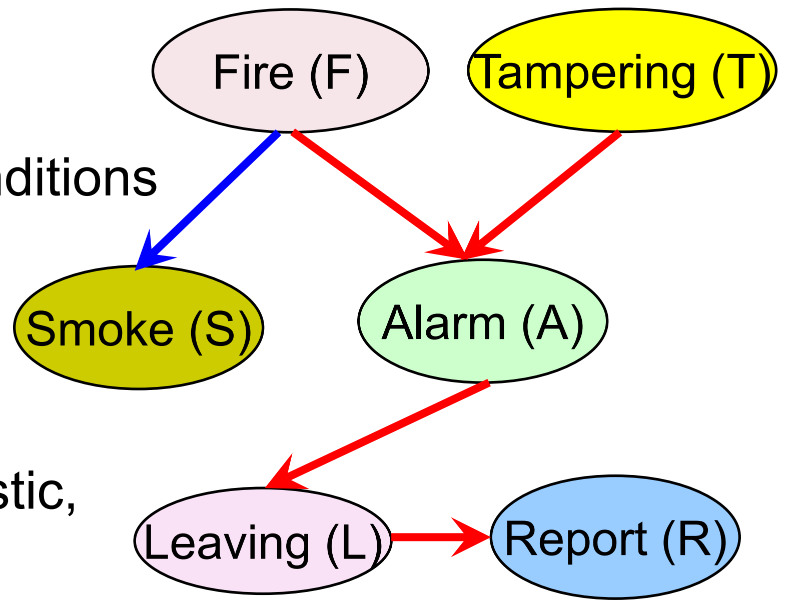
$$\sigma(v,w)=\frac{|\Gamma(v)\bigcap\Gamma(w)|}{\sqrt{|\Gamma(v)|\cdot|\Gamma(w)|}}$$

- Structural similarity is large for members of a clique and small for hubs and outliers



Bayes Networks

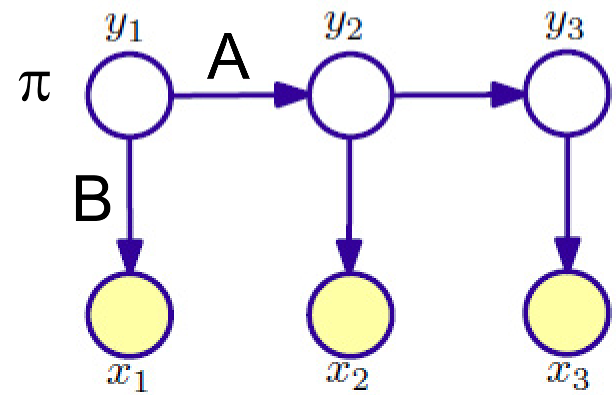
- Bayes Theory, Bayes Theorem
  - Determine ***likelihood*** for certain conditions
  - Compute ***joint probability***
- Bayesian Networks
  - Directed acyclic graph
  - Different types of reasoning: diagnostic, predictive, inter-causal, or combined
- ***Conditional Probability Tables*** for each possible combination of parents



Fire	$\Theta_{s=T\|f}$
True	.90
False	.01

Fire	Tampering	$\Theta_{a=T\|f,t}$
True	True	.5
True	False	.99
False	True	.85
False	False	.0001

Hidden Markov Models



- Model λ:(A, B, π)
  - A: State-transition matrix
  - B: Symbol-emission matrix
  - π: initial state probability vector
  - describes transition- and emission probabilities
- ***Markov property***: next state depends only on current state
- Only emissions are observable, but unknown which state produced them (so: states are ***hidden***)
- Can do:
  - Given HMM & observation sequence $\rightarrow$ infer state sequence
  - Given HMM $\rightarrow$ how probable is a state sequence
  - Given observation sequence(s) $\rightarrow$ learn HMM

Data Mining

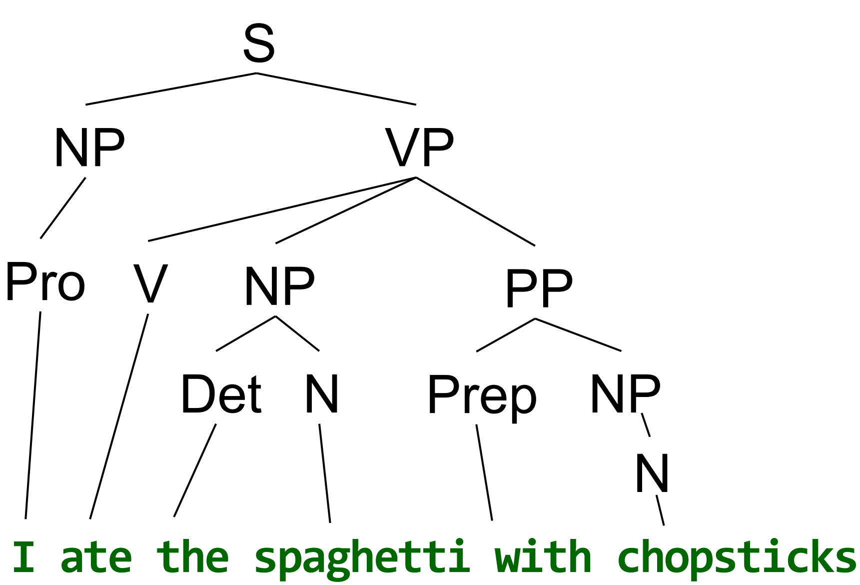
Lecture 12
Text Mining

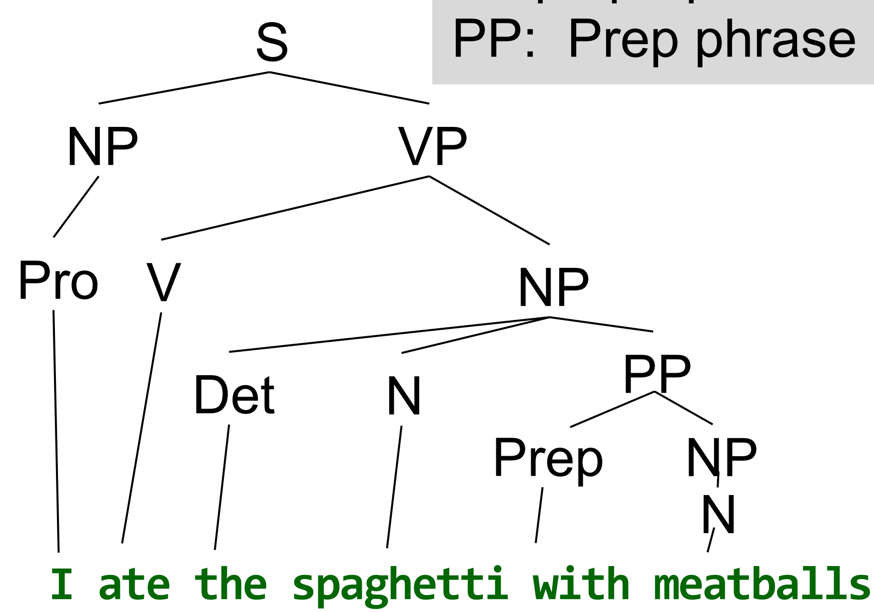
http://www.informatik.uni-hamburg.de/WTM/

Natural Language Processing

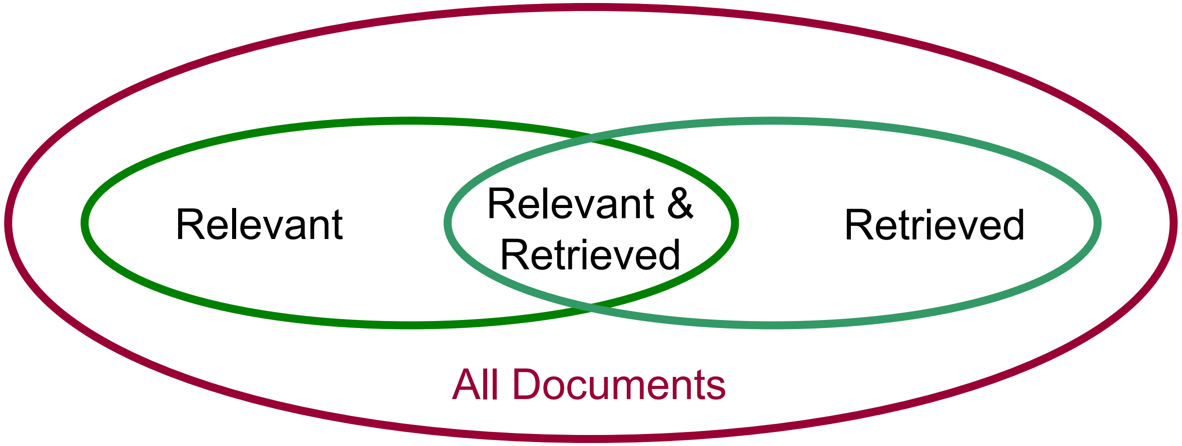
- Lexicon, Word sense ***disambiguation***
- Part-of-Speech ***tagging***
- Produce the correct ***syntactic parse tree*** for a sentence

S: sentence
NP: noun phrase
VP: verb phrase
N: noun
V: verb
Pro: pronoun
Det: determinant
Prep: preposition
PP: Prep phrase





Basic Measures for Text Retrieval



- ***Precision***: the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- ***Recall***: the percentage of documents that are relevant to the query and were, in fact, retrieved

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

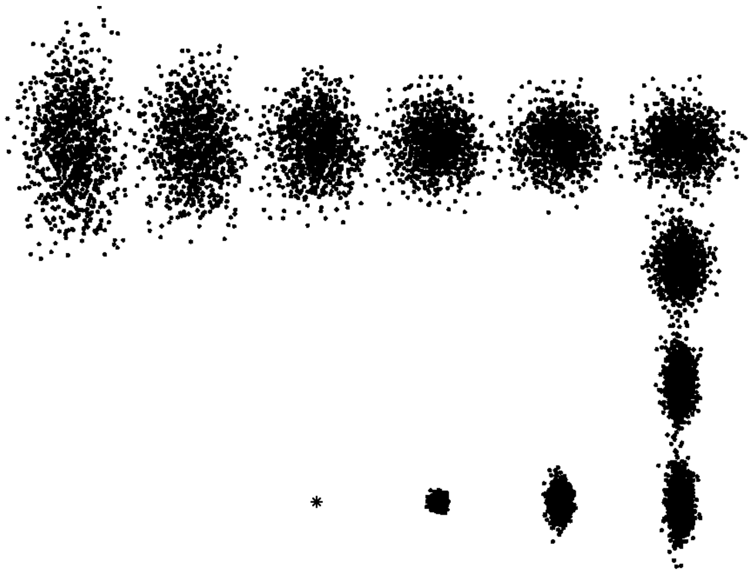
Data Mining

Lecture 13
Hybrid Systems and Current Topics in Data Mining

http://www.informatik.uni-hamburg.de/WTM/

Particle Filter Algorithm

1. Initialise particles randomly
2. For N steps do
   1. For all particles p do
      1. If number of particles < threshold: Resample
      2. Update particles
      3. Change weights depending on observation
      4. Normalise weights



- Weight of particle = Level of certainty

Modelling Uncertainty in Data

- Difficult to know noise
  - Particle *P* usually modelled with ***Gaussian noise*** with mean μ and variance σ:

$$P(\mathbf{x}_{m_i}^{t+1} \mid \mathbf{x}_s, \mathbf{z}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mathbf{z}-\mu)^2}{2\sigma^2}}$$

Position $\mathbf{x}$ is a vector over coordinates *x* & *y*, and the angle *θ*

Position of particle i at next time step

Position of robot

Estimated tracker measurement

Gaussian white noise

  - Quality of estimate depending on used variances
    - Could be fixed…
    - …or dynamic over the position:

$$\sigma^{t+1} = h(\mathbf{z}, \sigma) = \operatorname{asin}\left(\sigma \big/ \sqrt{dx^2 + dy^2}\right)$$

Integration into Hybrid Systems

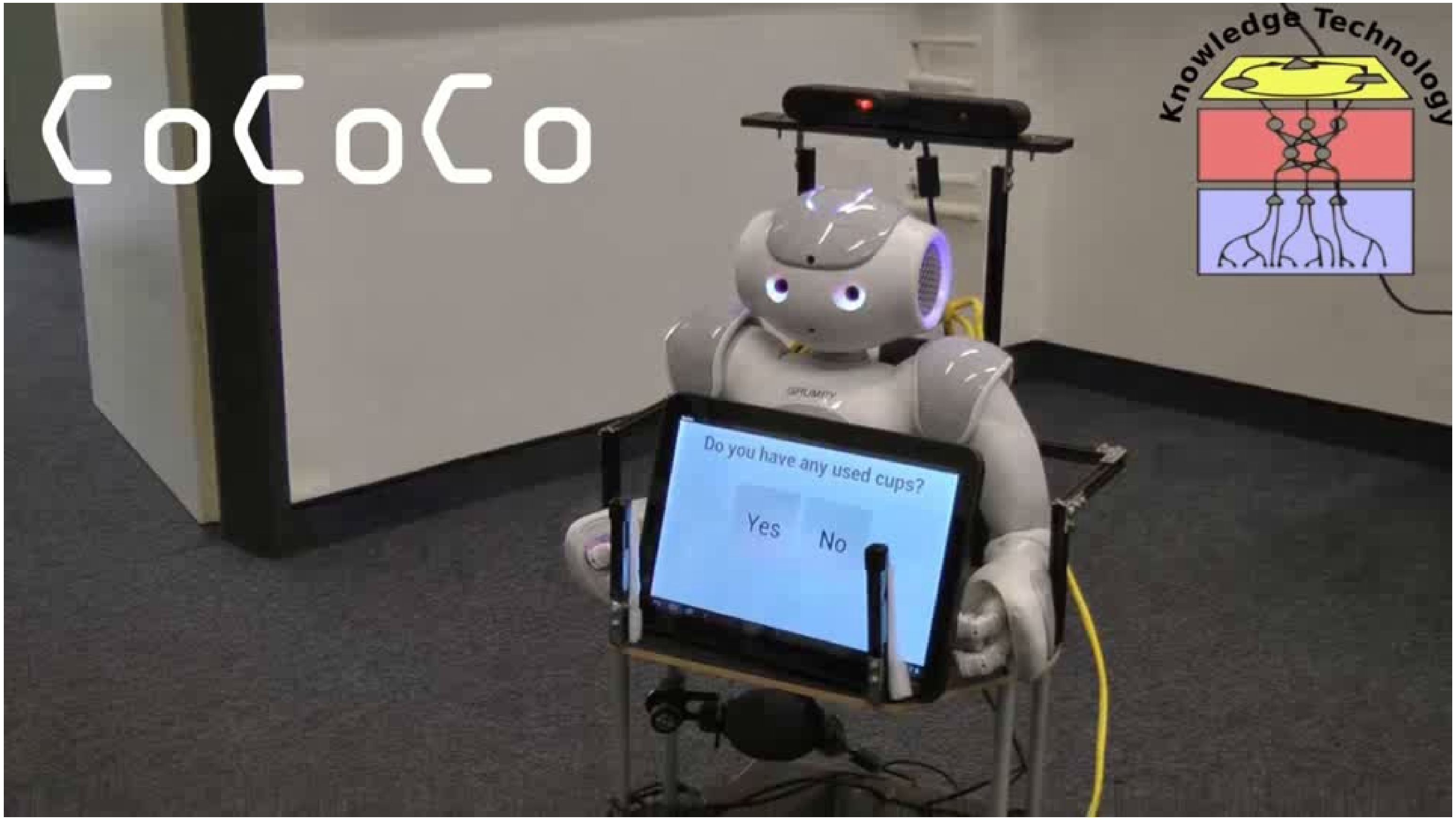
	Neural/Statistical/ Sub-symbolic	Symbolic/Structural/ Rule-based
Knowledge format	Numbers, Connections	Rules, Trees, Structure
Representation	Distributed	Local
Computational elements	Numerical associations Weights Thresholds	Premises, Conclusions Rule strength Predicates
Processing	Continuous activations	Discrete symbols
Cognitive level	Low	High
Basic units	Neurons	Rules
Manipulated by…	Continuous math	Logic
Representation	Compact but distributed	Verbose (→ brittle)

- Hybrid systems combine both properties

Data Mining Klausur

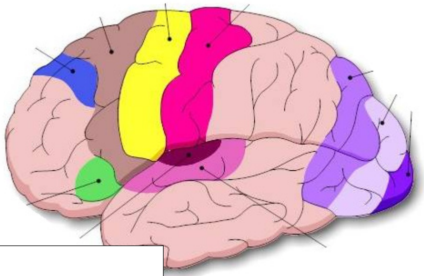
- Wann?
  - 1.Termin: 15.07.2014
  - 2.Termin: 29.09.2014 (Nachschreibeklausur)
- Wo?
  - Von-Melle-Park 6, Hörsaal Phil B (15.7.), Phil C (29.9.)
- Wann?
  - Beginn Klausur: 9:30 Uhr, Einlass: 9:00 Uhr
  - Ende Klausur: 11:30 Uhr
- **Hinweis: Personalausweis mitbringen!**
- **Mobiltelefone sind während der Klausur auszuschalten**

Data Mining in a recent Hybrid System



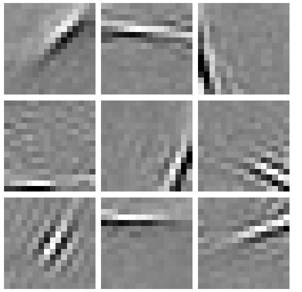
MSc Project Human-Robot Interaction WS2013/2014

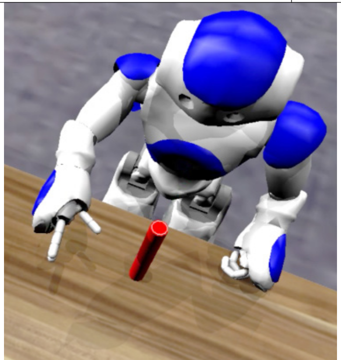
WTM for the Winter Semester (1) …

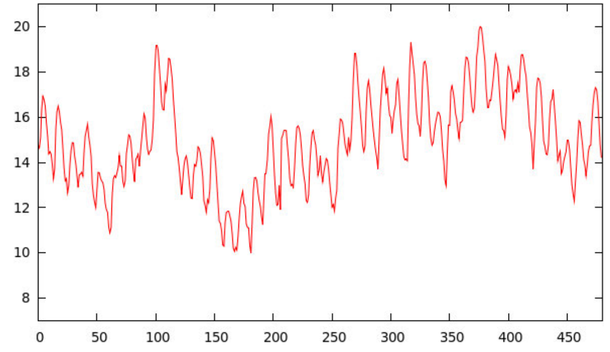


- BSc Practicum: Neural Networks

Methods	Feedforward networks	Recurrent networks
Unsupervised learning	Self-organizing maps, generative models	Hopfield network, Boltzmann machine
Supervised learning	Multi-layer perceptron (MLP)	Elman network
Reinforcement learning	Actor-critic, SARSA, CACLA	

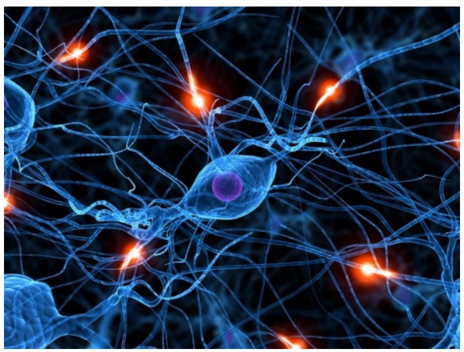


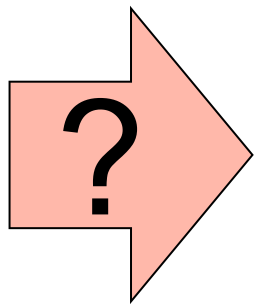


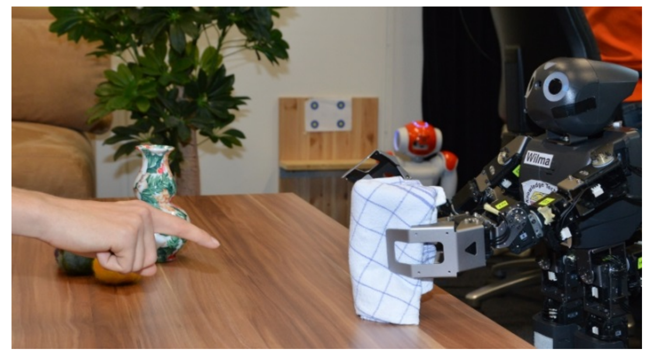


WTM for the Winter Semester (2) …

- BSc Project: Neural Networks for Robots
  - How do we get a robot to behave intelligently?
  - Humans are controlled by a complex neural network







How can neural networks be modelled?

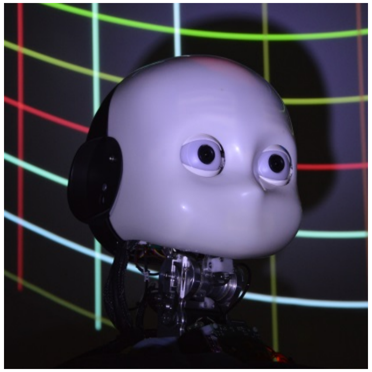
How do I design networks to show certain behaviour?

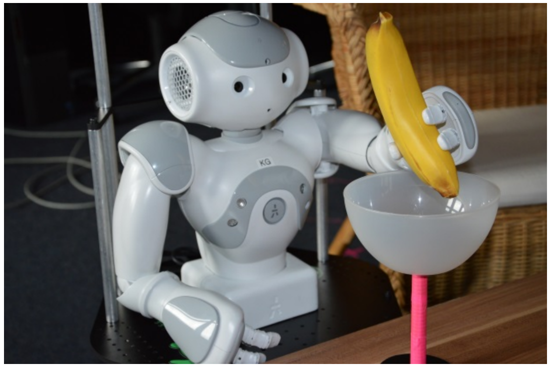
How do I integrate NNs in a robot?

  - **Aim of the project: Create neural network controllers that get our robot to do something intelligent!**

... some Outlook for the Master (1) ...

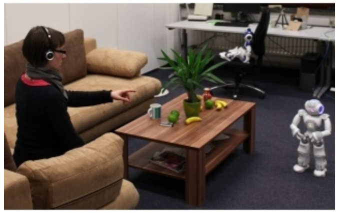
- L+S: Bio-inspired Artificial Intelligence
  - Adaptation, learning, development, evolution!
  - Learn about the nature and human!
  - Learn about brain and mind!
  - Experience how to build intelligent systems and robots!

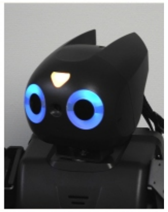




... some Outlook for the Master (2) ...

- MSc Project: Human-Robot Interaction
  - Challenge: Robotic device capable ***of interacting with people*** as naturally as we interact with each other
  - Approach: solve a ***simple task*** in a ***complex environment***, e.g. “Serve coffee!”
  - Inspiration: RoboCup@home tasks
  - Chance: Follow up on award-winning ideas and environments of the recent student groups




… and Topics for later BSc or MSc Projects

- Check for current offers:
  http://www.informatik.uni-hamburg.de/WTM/teaching/suggested_topics_titles.shtml

- Of course, feel free to discuss your own ideas with us

- Or contact your WTM tutors:
  heinrich@informatik.uni-hamburg.de
  jirak@informatik.uni-hamburg.de
  weber@informatik.uni-hamburg.de

- *Additional*: Oberseminar Knowledge Technology
  http://www.informatik.uni-hamburg.de/WTM/teaching/