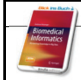


Andreas Holzinger
VO 709.049 Medical Informatics
18.11.2015 11:15-12:45

Lecture 06
Multimedia Data Mining and Knowledge Discovery

a.holzinger@tugraz.at
Tutor: markus.plass@student.tugraz.at
http://hci-kdd.org/biomedical-informatics-big-data

Schedule of the course

Keywords of the 6th Lecture

- Artificial neural networks
- Bayesian network
- Curse of dimensionality
- Deep Learning
- Data Mining
- Knowledge Discovery in medical data
- Medical text mining
- Model based clinical decision making
- Supervised learning
- Support Vector Machines (SVM)
- Unsupervised learning

Advance Organizer (1/2)

- **Artificial neural network (ANN)** = a computational adaptive model (inspired by biological neural networks), consisting of interconnected groups of artificial neurons; processes information using a connectionist approach.
- **Association rule learning** = a set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases used for data mining;
- **Classification** = a set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized; these techniques are often described as supervised learning because of the existence of a training set; they stand in contrast to cluster analysis, a type of unsupervised learning; used e.g. for data mining;
- **Cluster analysis** = statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance; a type of unsupervised learning because training data are not used - in contrast to classification; used for data mining.
- **Data mining** = a set of techniques to extract patterns from large data by combining methods from statistics and machine learning with database management (e.g. association rule learning, cluster analysis, classification, regression, etc.);
- **Knowledge Discovery (KD)** = process of identifying valid, novel, useful and understandable patterns out of large volumes of data

Advance Organizer (2/2)

- **Deep Learning** = class of machine learning algorithms using layers of non-linear processing units for feature extraction (remember: features are key for learning and understanding) - learning representations from data;
- **Knowledge Extraction** = is the creation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images) sources;
- **Multimedia** = several data of different modalities are processed at the same time, i.e. encompassing audio data (sound, speech), image data (b/w and colour images), video data (time-aligned sequences of images), electronic ink (sequences of time aligned 2D and 3D coordinates of a stylus, pen, data gloves etc.)
- **Principal Component Analysis (PCA)** = statistical technique for finding patterns in high-dimensional data;
- **Supervised learning** = inferring a function from supervised training data on the basis of training data which consist of a set of training examples, the input objects (typically vectors) and a desired output value (also called the supervisory signal).
- **Supervised learning algorithm** = analyzes the training data and produces an inferred function, called a classifier (if the output is discrete) or a regression function (if the output is continuous); the algorithm generalizes from the training data to unseen situations.
- **Support vector machine (SVM)** = concept for a set of related supervised learning methods to analyze data and recognize patterns, used for classification and regression analysis.
- **Unsupervised learning** = establishes clusters in data, where the class labels of training data is unknown.

Glossary

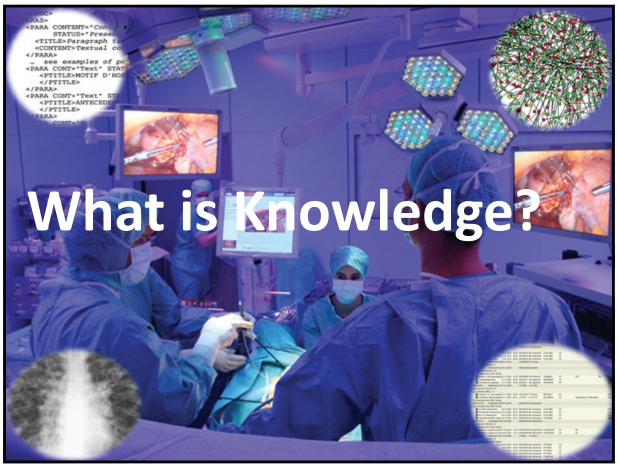
- ANN - artificial neural network
- ANN = Artificial Neural Network
- ANOVA = Analysis of Variance
- AUC - area under the curve
- CDT = Clinical Decision Tree
- DM = Data Mining
- KDD = Knowledge Discovery from Data(bases)
- MDM = Multimedia Data Mining
- MELD - model for end-stage liver disease
- MM = Multimedia
- NLP = Natural Language Processing
- ROC - receiver-operating characteristic
- SVM = Support Vector Machine

Learning Goals: At the end of this 6th lecture you …

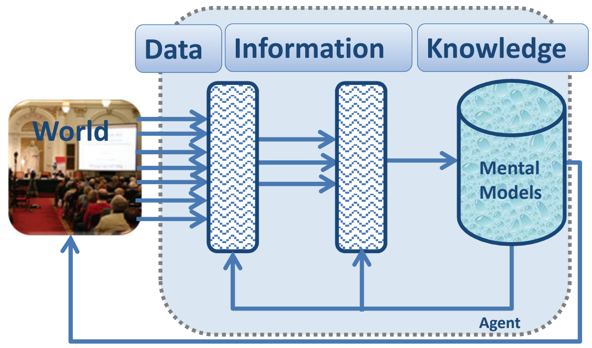
- … are aware of the importance of gaining **knowledge from (big) data**;
- … know the differences between Data Mining and Knowledge Discovery;
- … understand the basic process of knowledge discovery from data(bases) (KDD-chain);
- … have an overview on some data mining algorithms used in biomedical informatics;
- … have seen some examples of data mining applied in the biomedical domain;

Slide 6-1: Key Challenges

- 1) Cross-disciplinary cooperation with domain experts
- 2) Data-driven challenges including
  - a) Massive data sets;
  - b) Heterogeneous Data;
  - c) Streaming Data (e.g. from sensor nets, Multimedia, etc.);
  - d) Graph Data (e.g. Protein Network data, etc.);
  - e) Data restrictions (accessibility, privacy, safety, security, legal restrictions, fair use, etc.);
- 3) **Context** - Data Mining in a particular context
- 4) Interpretability
- 5) Computational Resources
- 6) Benchmarking against Gold-Standards
- 8) Embedded data mining

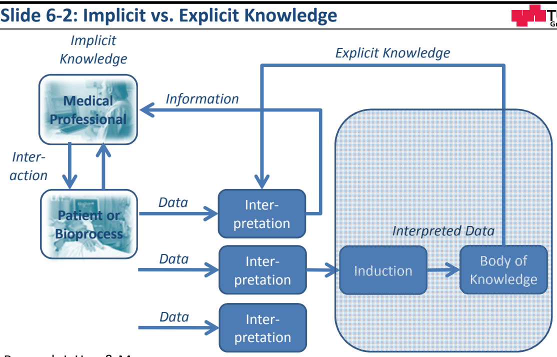


Adaptive Agent in a physical world



Knowledge := a set of expectations

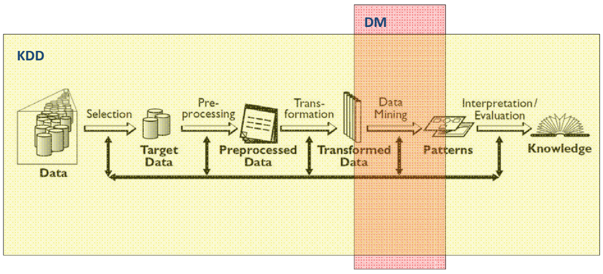
Slide 6-2: Implicit vs. Explicit Knowledge



Bemmel, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*

What is the difference between Knowledge Discovery and Data Mining?

Slide 6-3: The classic differentiation between DM and KDD



- KDD = Knowledge Discovery and Data Mining
- DM = Data Mining

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39, 11, 27-34.*

http://hci-kdd.org/

HCI-KDD

http://hci-kdd.org

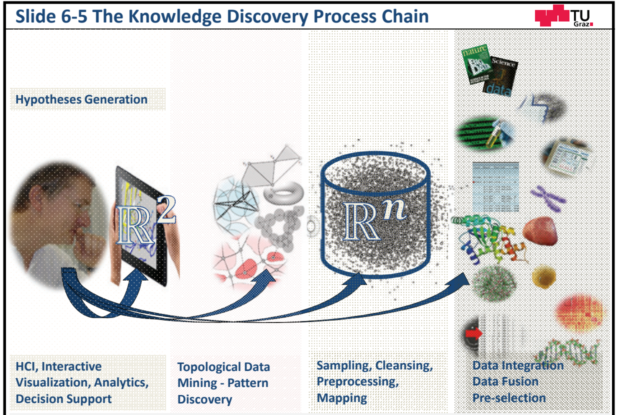
Work areas & Research Topics of the Holzinger Group

Interactive	Data Mining	Knowledge Discovery		
Data Visualization	Mining Algorithms	Data Mapping	Prepro-cessing	Data Fusion
HCI	GDM	Graph-based Data Mining		KDD
	EDM	Entropy-based Data Mining		
	TDM	Topological Data Mining		

Privacy, Data Protection, Safety and Security

Slide 6-4 Interactive Knowledge Discovery and Data Mining

Slide 6-5 The Knowledge Discovery Process Chain



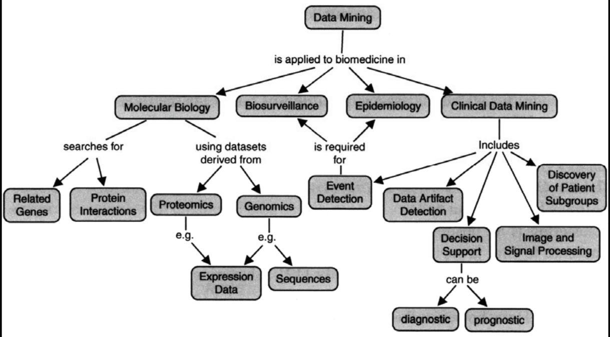
Privacy, Data Protection, Data Security, Data Safety

Slide 6-6: Definitions

- Data mining is the set of methods and techniques for exploring and analyzing (big) data sets;
- in an automatic or semi-automatic way, in order to find certain unknown or hidden rules, associations or tendencies;
- relevant essentials of the useful information while reducing the quantity of data;
- **descriptive (or exploratory)** techniques are designed to bring out information that is present but buried in a mass of data;
- **predictive (or explanatory)** techniques are designed to extrapolate new information based on the present information;

Tufféry, S. (2011) Overview of Data Mining. *Data Mining and Statistics for Decision Making (Wiley Series in Computational Statistics). New York, John Wiley & Sons, Ltd, 1-24.*

Slide 6-7 Data Mining in Biomedical Engineering

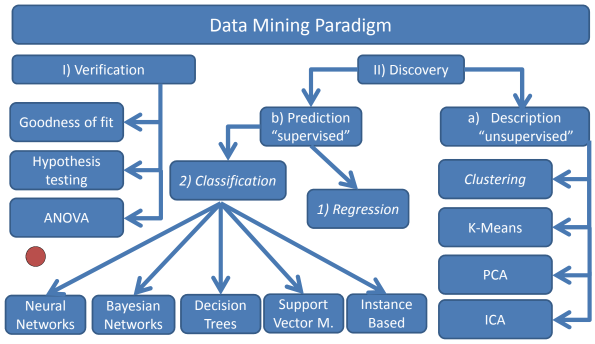


Suh, S. C., Gurupur, V. P. & Tanik, M. M. (2011) *Biomedical Engineering: Health Care Systems, Technology and Techniques. New York, Springer.*

Slide 6-8 Typical Data Mining Tasks

- Clustering: assigning a set of objects into groups
- Classification: predicting an item class, identifying to which set of categories a new observation belongs
- Associations: finding for example that A & B & C occur frequently together
- Visualization: to facilitate human cognition
- Deviation Detection: finding changes
- Anomaly Detection: finding anomalies
- Estimation: predicting a continuous value
- Link Analysis: finding relationships
- Forecasting: predicting a trend

Slide 6-9 Taxonomy of Data Mining Methods



Maimon, O. & Rokach, L. (Eds.) (2010) *Data Mining and Knowledge Discovery Handbook. Second Edition, New York, Dordrecht, Heidelberg, London, Springer.*

Data Mining, Knowledge Discovery, Machine Learning

Machine learning

- Machine learning is NOT a well defined field: it refers to a broad range of various **algorithms** within a feature space; hence:
- **Features** are key to machine learning and knowledge discovery!
- Tom Mitchell: A scientific field is best defined by the **central questions it studies.**
- ML seeks to answer the question *"How can we build computer systems that **automatically** improve with experience, and what are the fundamental laws that govern all learning processes?"*

Machine Learning is the most growing technical field …

- Progress in ML is driven by the ongoing explosion in the availability of online data and at the same time low-cost computation.

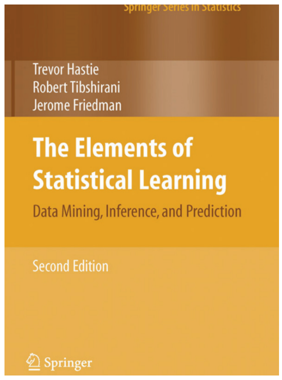
Jordan, M. I. & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. Science, 349, (6245), 255-260.

What is a best practice example of Machine Learning

Dietterich, T. G. & Horvitz, E. J. 2015. Rise of concerns about AI: reflections and directions. Commun. ACM, 58, (10), 38-40.

Machine Learning and Statistics are closely related

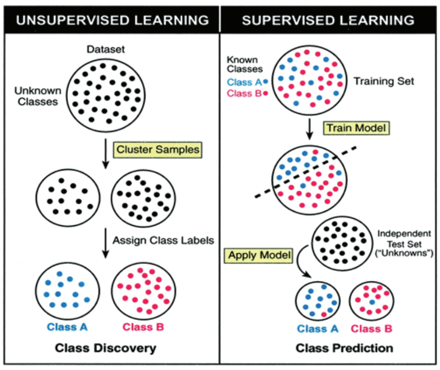
- Machine Learning is the development of algorithms which can **learn from data**
- Machine Learning has a pre-history in **statistical learning,** which is the application of statistical models and the assessment of **uncertainty**



Slide 6-10 Two main issues: Unsupervised vs. Supervised L.

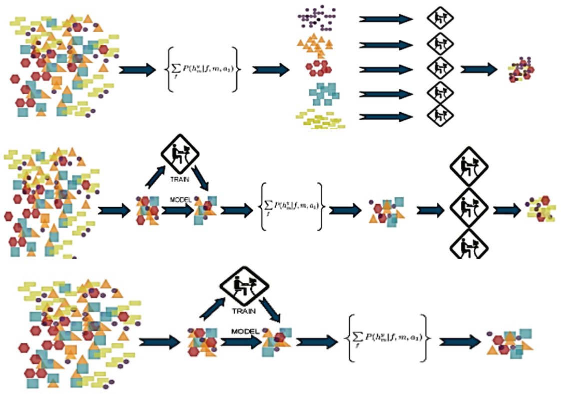
- **Unsupervised learning (e.g. clustering)**
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of clusters in the data;
- **Supervised learning (e.g. classification)**
  - Supervision = the training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations;
  - New data is classified based on the training set

Slide 6-11 Example Unsupervised vs. supervised learning

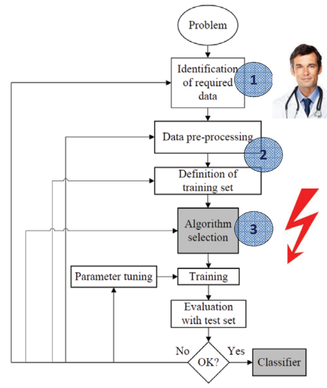


Ramaswamy, S. & Golub, T. R. (2002) DNA Microarrays in Clinical Oncology. *Journal of Clinical Oncology, 20, 7, 1932-1941.*

Slide 6-12 Unsupervised > Supervised > Semi-Supervised



Slide 6-13: Supervised Learning Process



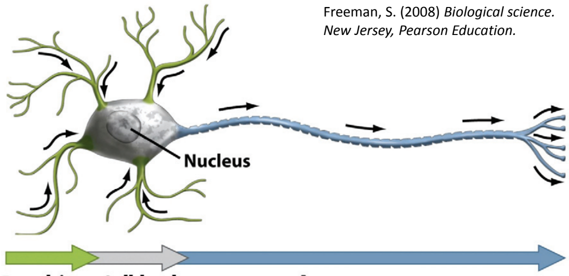
Kotsiantis, S. B. (2007) Supervised machine learning: A review of classification techniques. *Informatica, 31, 249-268.*

Slide 6-14 Example for supervised learning: ANN

Mikula, S., Trotts, I., Stone, J. M. & Jones, E. G. (2007) Internet-enabled high-resolution brain mapping and virtual microscopy. *NeuroImage, 35, 1, 9-15. http://brainmaps.org*

Slide 6-15: Neuron - Information flow through

Freeman, S. (2008) *Biological science. New Jersey, Pearson Education.*

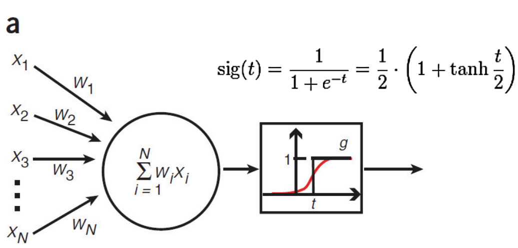


Dendrites Collect electrical signals

Cell body Integrates incoming signals and generates outgoing signal to axon

Axon Passes electrical signals to dendrites of another cell or to an effector cell

Slide 6-16: Perceptron - Artificial Neural Network ANN 1/4

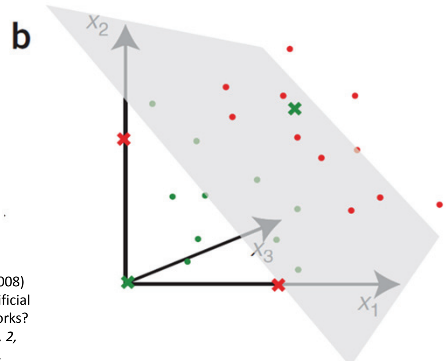


$$\mathrm{sig}(t) = \frac{1}{1+e^{-t}} = \frac{1}{2} \cdot \left(1 + \tanh \frac{t}{2}\right)$$

$$\sum_{i=1}^{N} w_i x_i = w_1 x_1 + w_2 x_2 + ... + w_N x_N$$

Krogh, A. (2008) What are artificial neural networks? *Nature, 26, 2, 195-197.*

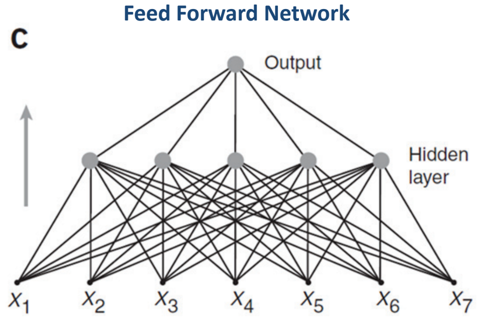
Slide 6-17 Classification Problem in Hyperplane - ANN 2/4



Krogh, A. (2008) What are artificial neural networks? *Nature, 26, 2, 195-197.*

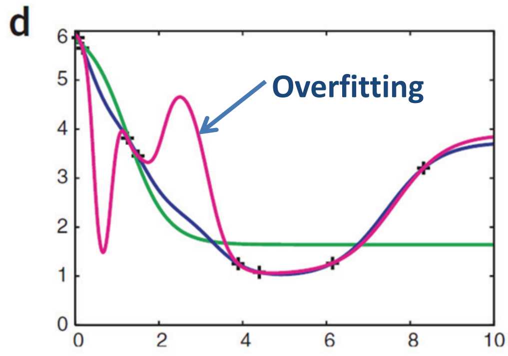
Slide 6-18 Multi-Layer Perceptron - ANN 3/4

Feed Forward Network



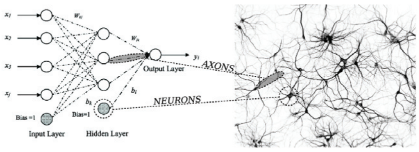
Krogh, A. (2008) What are artificial neural networks? *Nature, 26, 2, 195-197.*

Slide 6-19: Danger of Over-fitting – ANN 4/4



Krogh, A. (2008) What are artificial neural networks? *Nature, 26, 2, 195-197.*

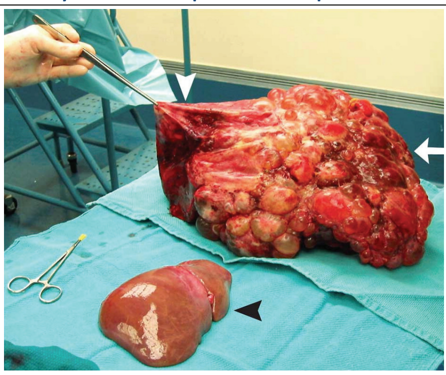
Slide 6-20: Neural Networks in Biomedical Engineering



- Many applications, many other networks, for example:
- Hopfield networks,
- Boltzmann machines,
- Kohonen nets,
- Unsupervised networks, …

Biomedical Examples

Slide 6-21: Risky Medical Example: Liver transplantation



Wall, W. J. (2007) Liver Transplantation for Polycystic Liver Disease. *New England Journal of Medicine, 356, 15, 1560-1560.*

Slide 6-22: Model for End-Stage Liver Disease (MELD)

TABLE 3. Three-Month Death Rates

MELD	≤9	10-19	20-29	30-39	≥40
Hospitalized	4% (6/148)	27% (28/103)	76% (16/21)	83% (5/6)	100% (4/4)
Ambulatory noncholestatic	2% (5/213)	5.6% (14/248)	50% (15/30)	—	—
Ambulatory PBC	1% (3/308)	13% (2/16)	0% (0/2)	—	—
Historical	8% (55/711)	26% (90/344)	56% (47/84)	66% (23/35)	100% (5/5)

N = 2,278

TABLE 2. Validity of MELD in Predicting Mortality

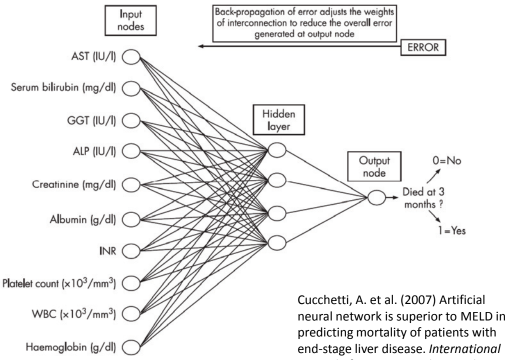
	Hospitalized N = 282	Ambulatory Noncholestatic N = 491	Ambulatory PBC N = 326	Historical N = 1179
1-Week mortality	0.95 (0.88-1.00)	0.80 (0.67-0.94)	—	0.84 (0.78-0.89)
3-Month mortality	0.87 (0.82-0.92)	0.80 (0.69-0.90)	0.87 (0.71-1.00)	0.78 (0.74-0.81)
1-Year mortality	0.85 (0.80-0.90)*	0.78 (0.70-0.85)	0.87 (0.80-0.93)	0.73 (0.69-0.76)†

NOTE. Values reported are the concordance statistic (95% CI).
* N = 257; 25 patients lost to follow-up.
† N = 1,108; 71 patients lost to follow-up.

Consists of serum bilirubin and creatinine levels, International Normalized Ratio (INR) for prothrombin time, and etiology of liver disease.

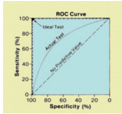
Kamath, P. S., Wiesner, R. H., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., D'Amico, G., Dickson, E. R. & Kim, W. (2001) A model to predict survival in patients with end-stage liver disease (MELD). *Hepatology, 33, 2, 464-470.*

Slide 6-23: ANN application Example: Liver transplantation



Cucchetti, A. et al. (2007) Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease. *International Journal of Gastroenterology and Hepatology - GUT, 56, 2, 253.*

Slide 6-24: Diagnostic accuracy of the ANN



	Internal cohort (Bologna)			External cohort (London)		
Variable	Survivors (n = 213)	Dead (n = 38)	p Value	Survivors (n = 120)	Dead (n = 17)	p Value
AST (IU/l)	93.7 (68.6)	116.6 (105.8)	0.207	84.2 (65.6)	116.1 (101.9)	0.253
Total bilirubin (mg/dl)	3.4 (2.9)	11.7 (9.6)	0.001	4.1 (5.4)	9.0 (8.5)	0.015
GGT (IU/l)	77.3 (63.9)	43.1 (25.4)	0.001	186.7 (43.3)	81.0 (79.5)	0.007
ALP (IU/l)	357.1 (219.0)	302.9 (110.1)	0.137	281.1 (152.5)	224.7 (204.9)	0.469
Creatinine (mg/dl)	0.92 (0.21)	1.02 (0.29)	0.011	1.28 (1.23)	1.79 (1.64)	0.044
Albumin (g/dl)	3.1 (0.4)	2.9 (0.4)	0.016	3.1 (0.6)	2.5 (0.8)	0.001
INR	1.57 (0.31)	2.11 (0.54)	0.001	1.21 (0.37)	1.68 (0.83)	0.017
Platelet count (×10³/mm³)	74.7 (53.8)	64.6 (37.8)	0.267	149.1 (135.3)	112.7 (103.5)	0.294
WBC (×10³/mm³)	4.3 (1.6)	5.5 (2.2)	0.002	6.1 (5.1)	7.5 (6.3)	0.236
Haemoglobin (g/dl)	12.1 (5.7)	10.2 (1.3)	0.038	11.1 (1.9)	10.6 (2.4)	0.229

ALP, alkaline phosphatase; AST, aspartate aminotransferase; GGT, γ-glutamyl-transpeptidase; INR, international normalised ratio; WBC, white cell. Continuous variables are reported as mean (SD).

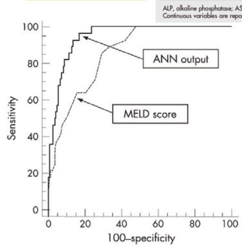


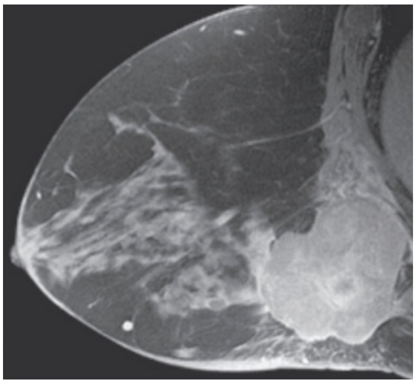
Table 3 Receiver-operating characteristic analysis of the artificial neural network in comparison with the model of the end-stage liver disease performance

	ANN		MELD score		
	AUC	95% CI	AUC	95% CI	p Value*
Training/cross validation	0.98	0.94 to 0.99	0.86	0.80 to 0.91	0.002
Internal validation	0.95	0.86 to 0.99	0.85	0.74 to 0.96	0.032
External cohort	0.96	0.91 to 0.98	0.86	0.79 to 0.91	0.044

ANN, artificial neural networks; AUC, area under the curve; MELD, model for end-stage liver disease.
*p values of each ANN subgroup compared with the MELD score of the same subgroup (Hanley–McNeil method).

Cucchetti, A. et al. (2007) Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease. *International Journal of Gastroenterology and Hepatology - GUT, 56, 2, 253.*

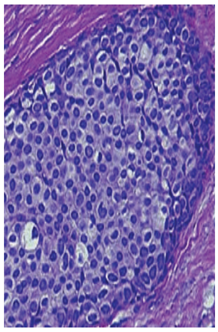
Slide 6-25: Another Clinical Case Example



Overmoyer, B. A., Lee, J. M. & Lerwill, M. F. (2011) Case 17-2011 A 49-Year-Old Woman with a Mass in the Breast and Overlying Skin Changes. *New England Journal of Medicine, 364, 23, 2246-2254.*

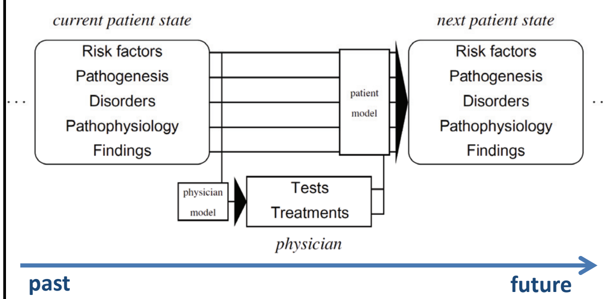
Slide 6-26: Important in Clinical practice -> prognosis !

- = the prediction of the future course of a disease conditional on the patient's history and a projected treatment strategy
- Danger: probable Information !
- Therefore valid prognostic models can be of great benefit for clinical decision making and of great value to the patient, e.g., for notification and quality of-life decisions



Knaus, W. A., Wagner, D. P. & Lynn, J. (1991) Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science, 254, 5030, 389.*

Slide 6-27 Model-based Clinical Decision Making Strategy



van Gerven, M. A. J., Taal, B. G. & Lucas, P. J. F. (2008) Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics, 41, 4, 515-529.*

Remember from Lecture 2 …

Note: Probable Information p (x) !

Slide 6-28: Bayesian Network (BN) - Definition

- is a **probabilistic model**, consisting of two parts:
- 1) a dependency structure and
- 2) local probability models.

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i \mid Pa(x_i))$$

Where $Pa(x_i)$ are the parents of x_i

BN inherently model the **uncertainty in the data.** They are a successful marriage between probability theory and graph theory; allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data. Furthermore this model allows different strategies to integrate two data sources.

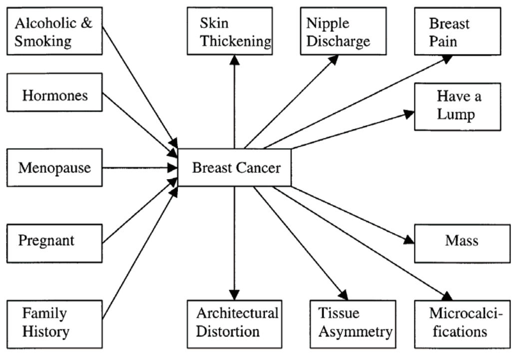
Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, Morgan Kaufmann.*

Slide 6-29: Example: Breast cancer - Probability Table

Category	Node description	State description
Diagnosis	Breast cancer	Present, absent.
Clinical history	Habit of drinking alcoholic beverages and smoking	Yes, no.
	Taking female hormones	Yes, no.
	Have gone through menopause	Yes, no.
	Have ever been pregnant	Yes, no.
	Family member has breast cancer	Yes, no.
Physical findings	Nipple discharge	Yes, no.
	Skin thickening	Yes, no.
	Breast pain	Yes, no.
	Have a lump(s)	Yes, no.
Mammographic findings	Architectural distortion	Present, absent.
	Mass	Score from one to three, score from four to five, absent
	Microcalcification cluster	Score from one to three, score from four to five, absent
	Asymmetry	Present, absent.

Wang, X. H., et al. (1999) Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics, 54, 2, 115-126.*

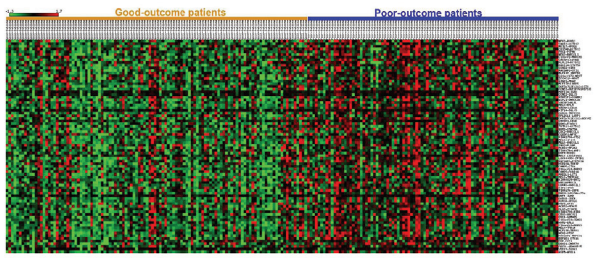
Slide 6-30 Breast cancer – big picture – state of 1999



Wang, X. H., et al. (1999) Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics, 54, 2, 115-126.*

Slide 6-31: 10 years later: Integration of microarray data

- Integrating microarray data from multiple studies to increase sample size;
- = approach to the development of more robust prognostic tests

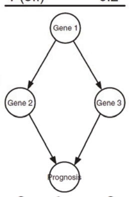


Xu, L., Tan, A., Winslow, R. & Geman, D. (2008) Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics, 9, 1, 125-139.*

Slide 6-32 Example: BN with four binary variables

Gene 1	
P(on)	0.8
P(off)	0.2

Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4



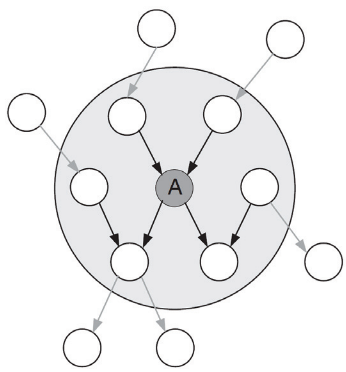
Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Prognosis	Gene 2 on Gene 3 on	Gene 2 on Gene 3 off	Gene 2 off Gene 2 on	Gene 2 off Gene 3 off
P(good)	0.6	0.1	0.9	0.5
P(poor)	0.4	0.9	0.1	0.5

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*

Slide 6-33 Concept Markov-Blanket

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*



Slide 6-34: Dependency Structure -> first step (1/2)

- First the structure is learned using a search strategy.
- Since the number of possible structures increases super exponentially with the number of variables,
- the well-known greedy search algorithm K2 can be used in combination with the Bayesian Dirichlet (BD) scoring metric:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right]$$

N_{ijk} … number of cases in the data set D
having variable i in state k associated with the j-th instantiation of its parents in current structure S.
n is the total number of variables.

Slide 6-35: Dependency Structure – first step (2/2)

- Next, N_{ij} is calculated by summing over all states of a variable:
- $N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \cdot N'_{ijk}$ and N'_{ij} have similar meanings but refer to prior knowledge for the parameters.
- When no knowledge is available they are estimated using $N_{ijk} = N/(r_i q_i)$
- with N the equivalent sample size,
- r_i the number of states of variable i and
- q_i the number of instantiations of the parents of variable i.
- $\Gamma(.)$ corresponds to the gamma distribution.
- Finally $p(S)$ is the prior probability of the structure.
- $p(S)$ is calculated by:
- $p(S) = \prod_{i=1}^{n} \prod_{l_i=1}^{p_i} p(l_i \longrightarrow x_i) \prod_{m_i=1}^{o_i} p(m_i x_i)$
- with p_i the number of parents of variable x_i and o_i all the variables that are not a parent of x_i.
- Next, $p(a \longrightarrow b)$ is the probability that there is an edge from a to b while $p(ab)$ is the inverse, i.e. the probability that there is no edge from a to b

Slide 6-36: Parameter learning -> second step

- Estimating the parameters of the local probability models corresponding with the dependency structure.
- CPTs are used to model these local probability models.
- For each variable and instantiation of its parents there exists a CPT that consists of a set of parameters.
- Each set of parameters was given a uniform Dirichlet prior:

$$\mathrm{p}(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1}, \ldots, N'_{ijk}, \ldots, N'_{ijr_i})$$

Note: With θ_{ij} a parameter set where i refers to the variable and j to the j-th instantiation of the parents in the current structure. θ_{ij} contains a probability for every value of the variable x_i given the current instantiation of the parents. Dir corresponds to the Dirichlet distribution with $(N'_{ij1}, \ldots, N'_{ijr_i})$ as parameters of this Dirichlet distribution. Parameter learning then consists of updating these Dirichlet priors with data. This is straightforward because the multinomial distribution that is used to model the data, and the Dirichlet distribution that models the prior, are conjugate distributions. This results in a Dirichlet posterior over the parameter set:

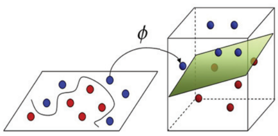
$$\mathrm{p}(\theta_{ij}|D, S) = Dir(\theta_{ij}|N'_{ij1} + N_{ij1}, \ldots, N'_{ijk} + N_{ijk}, \ldots, N'_{ijr_i} + N_{ijr_i})$$

with N_{ijk} defined as before.

Are there alternatives to such network approaches?

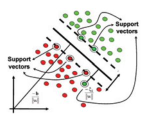
Slide 6-37: Support Vector Machine SVM – (Vapnik, 1992)

- Uses a nonlinear mapping to transform the original data (input space) into a higher dimension (feature space)

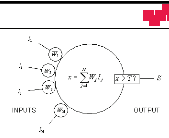


- = classification method for both linear and nonlinear data;
- Within the new dimension, it searches for the linear optimal separating **hyperplane** (i.e., “decision boundary”);
- By nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated with a hyperplane;
- The SVM finds this hyperplane by using **support vectors** (these are the “essential” training tuples) and **margins** (defined by the support vectors);

Slide 6-38 SVM vs. ANN

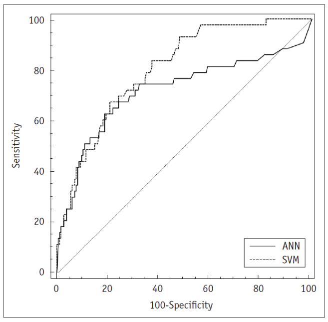


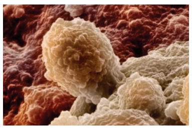
- **SVM**
  - Deterministic algorithm
  - Nice generalization properties
  - Hard to learn – learned in batch mode using quadratic programming techniques
  - Using kernels can learn very complex functions



- **ANN**
  - Nondeterministic algorithm
  - Generalizes well but doesn’t have strong mathematical foundation
  - Can easily be learned in incremental fashion
  - To learn complex functions—use multilayer perceptron (nontrivial)

Slide 6-39 Clinical use: SVM are more accurate than ANN





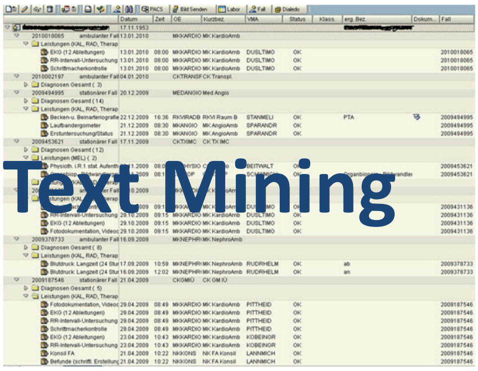
Kim, S. Y., Moon, S. K., Jung, D. C., Hwang, S. I., Sung, C. K., Cho, J. Y., Kim, S. H., Lee, J. & Lee, H. J. (2011) Pre-Operative Prediction of Advanced Prostatic Cancer Using Clinical Decision Support Systems: Accuracy Comparison between Support Vector Machine and Artificial Neural Network. *Korean J Radiol, 12, **5**, 588-594.*

Slide 6-40 The 10 top data mining algorithms

- **C4.5** = for generation of decision trees used for classification, (statistical classifier, Quinlan (1993));
- **k-means** = a simple iterative method for partition of a given dataset into a user-specified number of clusters, k (Lloyd (1957));
- **Apriori** = for finding frequent item sets using candidate generation (Agrawal & Srikant (1994));
- **EM** = Expectation–Maximization algorithm for finding maximum likelihood estimates of parameters in models (Dempster et al. (1977));
- **PageRank** = a search ranking algorithm using hyperlinks on the Web (Brin & Page (1998));
- **Adaptive Boost** = one of the most important ensemble methods (Freund & Shapire (1995));
- **k-Nearest Neighbor** = a method for classifying objects based on closest training examples in the feature space (Fix & Hodges (1951));
- **Naive Bayes** = can be trained very efficiently in a supervised learning setting (Domingos & Pazzani (1997));
- **CART** = Classification And Regression Trees as predictive model mapping observations about items to conclusions about the goal (Breiman et al 1984);
- **SVM** = *support vector machines offer one of the most robust and accurate methods among all well-known algorithms (Vapnik (1995));*

Wu et al. (2008) Top 10 algorithms in data mining. *Knowledge & Information Systems, 14, 1, 1-37*.

Still a big problem … Text



Kreuzthaler, M., Bloice, M. D., Faulstich, L., Simonic, K. M. & Holzinger, A. (2011) A Comparison of Different Retrieval Strategies Working on Medical Free Texts. *Journal of Universal Computer Science, 17, 7, 1109-1133.*

Slide 6-41: Selection of Semantic Methods

- Latent Semantic Analysis (LSA)
- Probabilistic latent semantic analysis (PLSA)
- Latent Dirichlet allocation (LDA)
- Hierarchical Latent Dirichlet Allocation (hLDA)
- Semantic Vector Space Model (SVSM)
- Latent semantic mapping (LSM)
- Principal component analysis (PCA)

Holzinger, A., Schantl, J., Schroettner, M., Seifert, C. & Verspoor, K. 2014. Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges. In: Lecture Notes in Computer Science LNCS 8401. Berlin Heidelberg: Springer pp. 271-300.

Future Outlook

Slide 6-44 Understanding Natural Language …

- … is a grand challenge in future computing
- … most of information in the hospital is unstructured and based on natural language
- … masses of information is not easily processable by humans
- … legacy approaches have all failed; "searching" not the right approach- Search is a way to gather information – but **not to answer questions**
- A new approach is needed, leveraging **content analytics** and natural language processing [1]

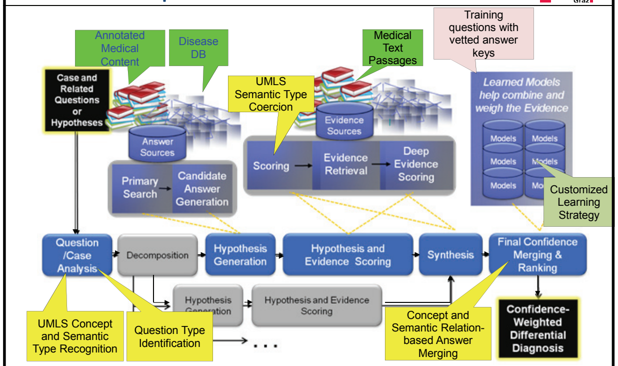
[1] Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A. & Hofmann-Wellenhof, R. 2013. Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an assistive technology in the biomedical domain. *Springer Lecture Notes in Computer Science LNCS 7947.* Heidelberg, Berlin, New York: Springer, 13-24.

Slide 6-45: Watson – a Workload Optimized System

- 90 x IBM Power 750 servers
- 2880 POWER7 cores
- POWER7 3.55 GHz chip
- 500 GB per sec on-chip bandwidth
- 10 Gb Ethernet network
- 15 Terabytes of memory
- 20 Terabytes of disk, clustered
- Can operate at 80 Teraflops
- Runs IBM DeepQA software
- Scales out with and searches vast amounts of unstructured information with UIMA & Hadoop open source components
- Linux provides a scalable, open platform, optimized to exploit POWER7 performance
- 10 racks include servers, networking, shared disk system, cluster controllers

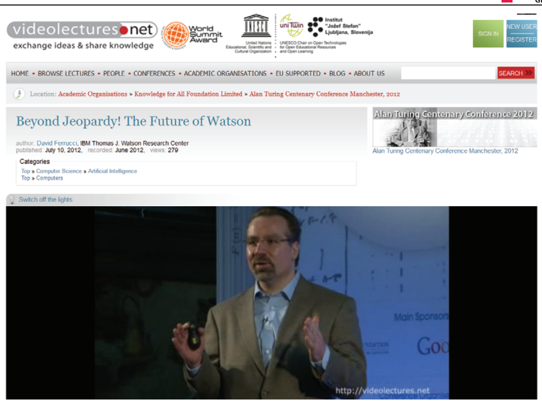
Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E. & Prager, J. 2010. Building Watson: An overview of the DeepQA project. *AI magazine, 31, (3), 59-79.*

Slide 6-46: Example: IBM Watson for Healthcare

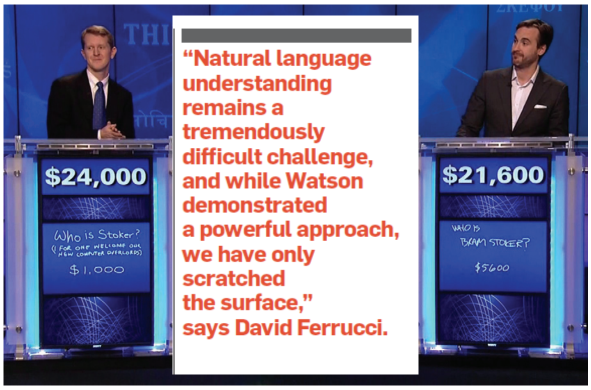


Ferrucci, D., Levas, A., Bagchi, S., Gondek, D. & Mueller, E. T. 2013. Watson: Beyond Jeopardy! *Artificial Intelligence, 199–200, (0), 93-105.*

Slide 6-47 IBM Watson at work



Much open work to do …



Thank You!

Future Outlook: Image data mining open research issues

Task	Approach/Application	Issue
Preprocessing	Pixel, region and grid level	Global vs. local level features
Feature extraction and transformation	Color histograms [135], color moments [134], color sets [130], shape descriptors, texture descriptors, edges	Sensitive to the parameters e.g., number of bins, bin boundaries, selection of feature vectors (Fourier etc.)
Image classification	GMM [30], SVM [141], Bayesian classifier [140] to classify image	Large training data needed for GMM, SVM kernel functions, etc.
Image clustering	K-means in preprocessing stage to identify patterns [63]	Unknown number of clusters
Image association	A-priori based association between structures and functions of human brain [91]	Scalability issue in terms of number of candidate patterns generated

Bhatt, C. & Kankanhalli, M. (2011) Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications, 51, 1, 35-76.*

Future Outlook: Video data mining open research issues

Task	Approach/Application	Issue
Preprocessing	Shot level, frame level, scene level, clip level	Depends on video structure and application type
Feature extraction and transformation	Image features at frame level as well as motion descriptors, camera metadata like motion [161], date, place, time etc.	Motion calculation is computationally expensive
Video classification	Human body motion recognition [18], goal detection [161], gestures etc.	Not enough training data to learn events of interest, domain knowledge dependence
Video clustering	Clustering of shots [152, 153], segments [100] for video, for indexing etc.	Scalability for large video clusters is an issue
Video association	A-priori based association finding sequence of events in movies [127]	Finding semantic event boundaries and temporal distance thresholds

Bhatt, C. & Kankanhalli, M. (2011) Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications, 51, 1, 35-76.*

Future Outlook: Audio data mining open research issues

Task	Approach/Application	Issue
Preprocessing	Phoneme level, word level, window of fixed sizes	Domain dependence e.g., phoneme level if foreign terms, segmenting out silence, noise etc.
Feature extraction and transformation	Pause rate, zero crossing rate, Mel frequency cepstral coefficients [19], bandwidth, spectral centroid, frequency spectrum	Sensitive to the parameters e.g., number of bins, frequency resolution, smoothing factors etc.
Audio classification	HMM [24], GMM [67], SVM [46], Bayesian classifier to do segmentation and classification of speech, music etc.	Model based approaches have problem to work well in real time as more number of iterations and lot of data are needed for training.
Audio clustering	Clustering speaker gender/speech segment of same speaker [92]	Large clusters are sometimes biased

Bhatt, C. & Kankanhalli, M. (2011) Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications, 51, 1, 35-76.*

Future Outlook: Text data mining open research issues

Task	Approach/Application	Issue
Preprocessing	Part of speech tagging, stemming, stop word removal, text chunking etc.	Resolving ambiguity is main problem, single word different meaning in different context
Feature extraction and transformation	Identification of important keywords using TF, IDF [117] etc.	Different sizes and contexts of the documents
Text classification	K-nearest neighbor classification, decision tress, naïve bayes classifier [34]	Final decisions depends on relatively few terms
Text clustering	Bi section k means clustering, co clustering [34]	Finding good distance measures

Bhatt & Kankanhalli (2011)

Future Outlook: Multimodal data mining open research issues

Task	Approach/Application	Issue
Preprocessing	Treat multiple streams separately [99] or consider multiple streams as one [161]	Multimodal data stream synchronization
Feature transformation	Metadata fusion [146]	External knowledge fusion
Multimodal classification	Pre filtering: heuristic rule based [16], SVM classifier based [14], sub space based [129]; fusion: late [99], early [161] and meta [81, 82]	Class imbalance, multimodal classifier fusion
Multimodal clustering	Mixed media graph [103], clustering [7, 42, 148], EEML [47]	Cross modal correlation discovery
Multimodal association	Generalized sequence pattern mining [90, 126]	Automatic identification of temporal structures

Bhatt, C. & Kankanhalli, M. (2011) Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications, 51, 1, 35-76.*

Sample Questions (1)

- What is the main goal in knowledge discovery?
- Describe the classical process of KDD!
- How do we define a data object?
- What are the most important data mining tasks?
- What is the difference between supervised and unsupervised learning?
- What is the difference between classification and numeric prediction?
- Why is cluster analysis in medicine important?
- How is data mining defined?
- Describe the taxonomy of data mining paradigms!
- What is a neural network and a neuron?
- What is a artificial neural network?
- How does an ANN work?

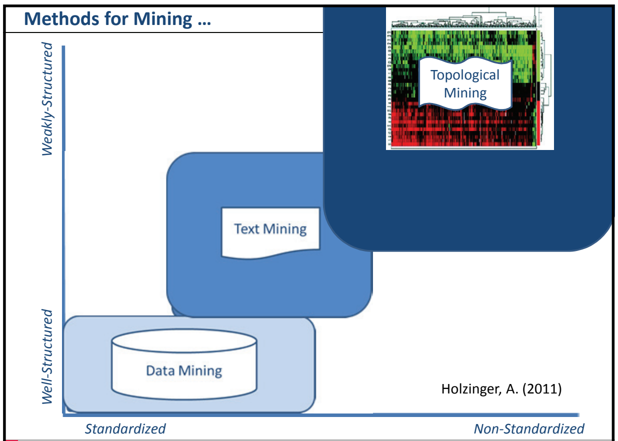
- Provide an example on the use of ANN for medical decision making!
- What can you infer from a Receiver Operating Characteristic (ROC)?
- How can you rate the diagnostic accuracy of an ANN?
- What is model based clinical decision making?
- What is very important in clinical patient management?
- What is a Bayesian Network (BN Definition)?
- Why do we need a Markov-Blanket?
- What is the principal function of a Support Vector Machine?
- How would you describe the differences between ANN and SVM?
- Why is text mining in the medical domain practice so difficult?
- Just name some important semantic methods for NLP!
- What is the typical system architecture of a NLP System?

Some Useful Links (1)

- http://www.cs.waikato.ac.nz/~ml/weka/
- http://www.youtube.com/watch?v=eDYOH9q2QdA&feature=list_related&playnext=1&list=SPA45CD9BACEC8B923 (intro to neural networks)
- http://lcn.epfl.ch/tutorial/english/index.html (Neural networks simulation)
- http://leenissen.dk/fann/wp/ (Fast Artificial Neural Network Library)
- http://www.neurosolutions.com/ (End-user friendly Neural Network Development Environment)
- http://www.support-vector-machines.org
- http://www.kernel-machines.org

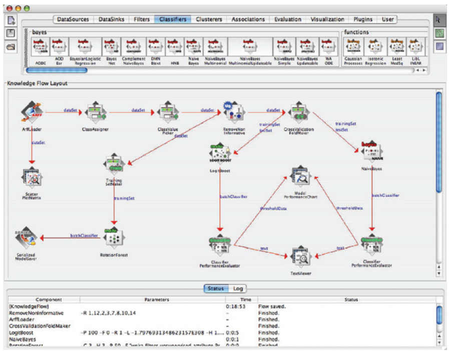
A. Holzinger 709.049 79/118 Med. Informatics L06

Methods for Mining …



A. Holzinger 709.049 80/118 Med. Informatics L06

Appendix: WEKA Tool – The knowledge flow interface



Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009) The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11, 1, 10-18.*

A. Holzinger 709.049 81/118 Med. Informatics L06

Appendix: Criteria for selecting a data mining system

- Scalability
  - Row (or database size) scalability
  - Column (or dimension) scalability
  - Curse of dimensionality: it is much more challenging to make a system column scalable that row scalable
- Availability of Visualization tools
  - Visualization categories: data visualization, mining result visualization, mining process visualization, and visual data mining
- Data mining query language and graphical user interface
- Easy-to-use and high-quality graphical user interfaces are essential for end-user guided, highly interactive data mining systems

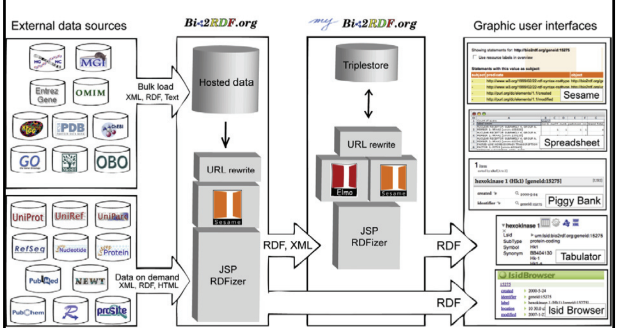
A. Holzinger 709.049 82/118 Med. Informatics L06

Appendix: Examples of professional data mining systems

- Mirosoft SQLServer 2005 (Integrated DB and OLAP with mining Support OLEDB for DM standard)
- SAS Enterprise Miner (A variety of statistical analysis tools, Data warehouse tools and multiple data mining algorithms)
- IBM Intelligent Miner (A wide range of data mining algorithms; Scalable mining algorithms; Toolkits: neural network algorithms, statistical methods, data preparation, and data visualization tools; Tight integration with IBM's DB2 relational database
- SGI MineSet (advanced statistics and visualization tools)
- Clementine (SPSS – integrated data mining development environment – multiple data mining algorithms)

A. Holzinger 709.049 83/118 Med. Informatics L06

Appendix: Example Bio2RDF Architecture

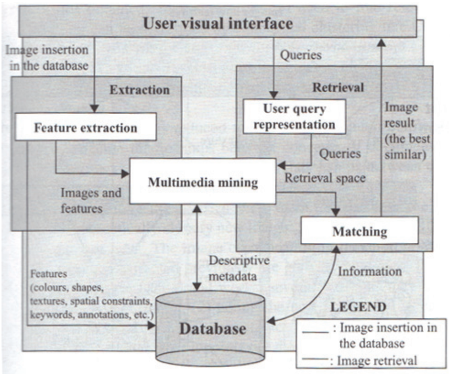


Belleau, F. et al. (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics, 41, 5, 706-716.*

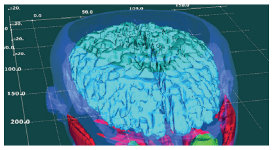
A. Holzinger 709.049 84/118 Med. Informatics L06

Appendix: Sample Architecture for a visual retrieval system

Bouet, M. & Aufaure, M.-A. (2007) New Image Retrieval Principle: Image Mining and Visual Ontology. In: Petrushin, V. A. & Khan, L. (Eds.) *Multimedia Data Mining and Knowledge Discovery.* London, Springer.



Some Data Mining wording first



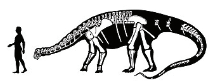
- Data sets are made up of data objects.
- Each data object represents an entity.
- Data objects are described by attributes.
- We want to mine samples (aka examples, instances, data points, objects, tuples) out of databases
- Database rows → data objects; columns → attributes

Schedule

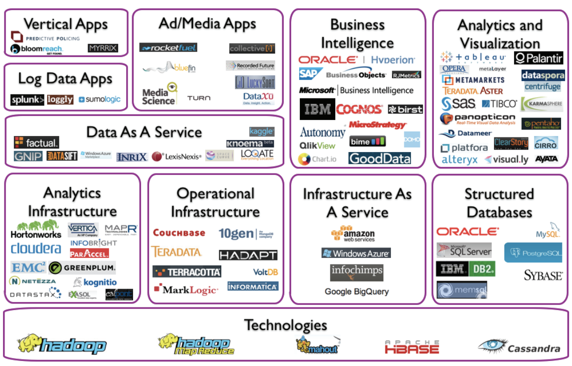
- 1. Intro: Computer Science meets Life Sciences, challenges, future directions
- 2. Back to the future: Fundamentals of Data, Information and Knowledge
- **3. Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)**
- 4. Biomedical Databases: Acquisition, Storage, Information Retrieval and Use
- 5. Semi structured and weakly structured data (structural homologies)
- 6. Multimedia Data Mining and Knowledge Discovery
- 7. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
- 8. Biomedical Decision Making: Reasoning and Decision Support
- 9. Intelligent Information Visualization and Visual Analytics
- 10. Biomedical Information Systems and Medical Knowledge Management
- 11. Biomedical Data: Privacy, Safety and Security
- 12. Methodology for Info Systems: System Design, Usability & Evaluation

Knowledge Discovery is Making Data Useful

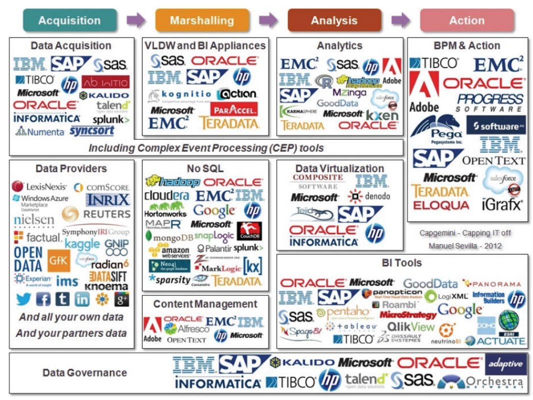
- **Masses of data –**
- **Graph-based data may contain information about design principles and/or the evolutionary history of complex systems**
- **As in Paleontology: Discovery from past data**


Big Data Players



Feinleib (2012), http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/



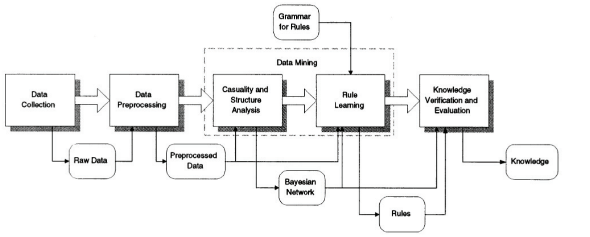
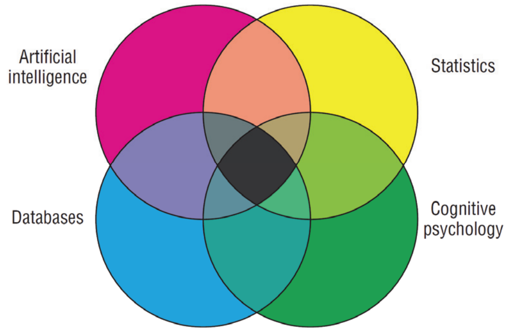


Fig. 1. The knowledge discovery process.

Shun Ngan, P., Leung Wong, M., Lam, W., Leung, K. S. & Cheng, J. C. Y. 1999. Medical data mining using evolutionary computation. *Artificial Intelligence in Medicine, 16, (1), 73-96.*

The fields of KDD (Interaction is missing!)



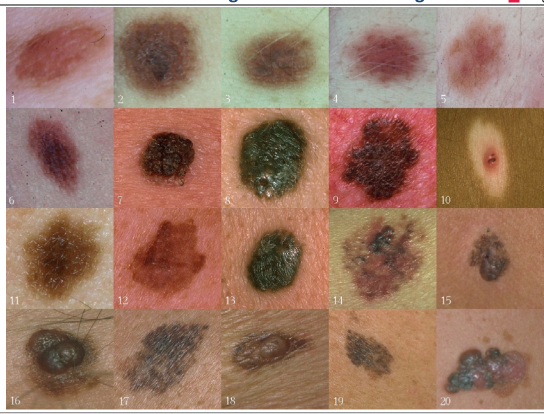
Pazzani, M. J. (2000) Knowledge discovery from data? *Intelligent Systems and their Applications, IEEE, 15, 2, 10-12.*

Slide 6-12 Example Classification vs. Numeric Prediction

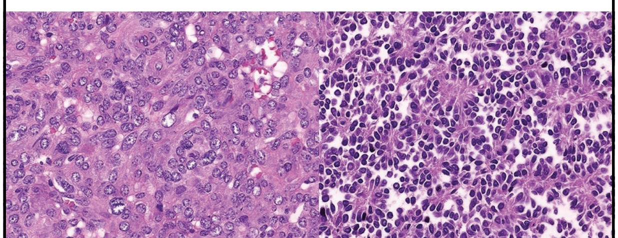
- **Medical Decision:** is a tumor **malign or benign?**
- **Classification** = predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- **Numeric Prediction** = models continuous-valued functions, i.e., predicts unknown or missing values

Image Source: John Nguyen (2010)

Slide 6-13 Which one is malign – which one is benign?

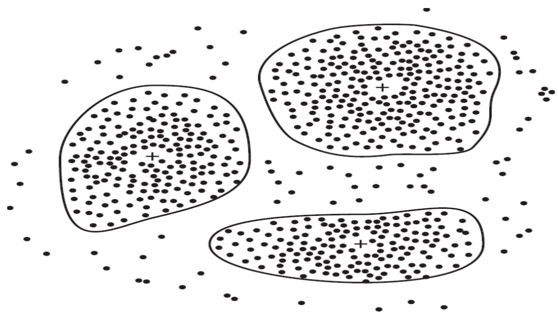


Slide 6-14 Example: Benign vs. Malign – molecular level



webpathology.com Am J Surg Pathol 2002 March; 26(3):320-7

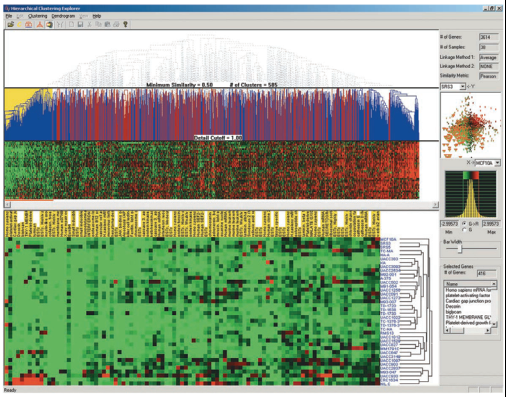
Slide 6-15 Cluster analysis



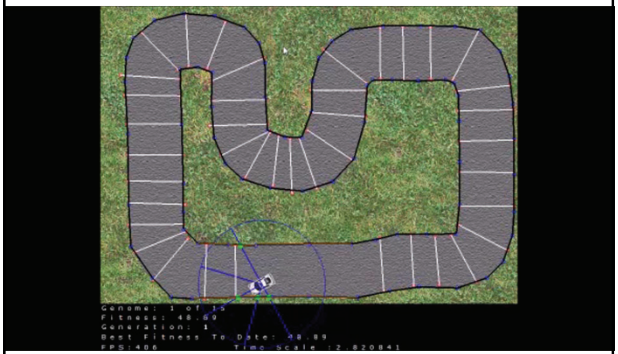
Han, J., Kamber, M. & Pei, J. (2011) *Data Mining: Concepts and Techniques (Third Edition). The Morgan Kaufmann Series in Data Management Systems. San Francisco, Morgan Kaufmann Publishers.*

Slide 6-16 H. Clustering Explorer: melanoma gene expression

Seo, J. & Shneiderman, B. (2002) Interactively exploring hierarchical clustering results in gene identification. *Computer, 35, 7, 80-86.*



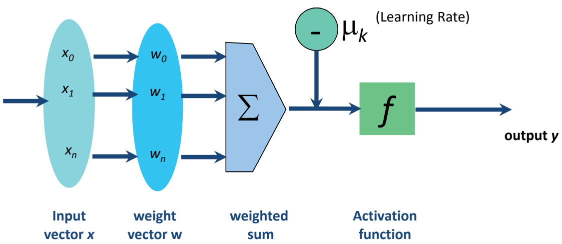
ANN Demo: Learning Process



http://www.youtube.com/watch?v=0Str0Rdkxxo

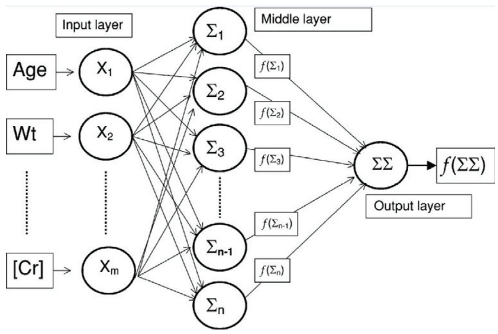
Slide 6-24 Artificial Neural Network ANN

- **The *n*-dimensional input vector *x* is mapped into variable *y* by means of the scalar product and a nonlinear function mapping**



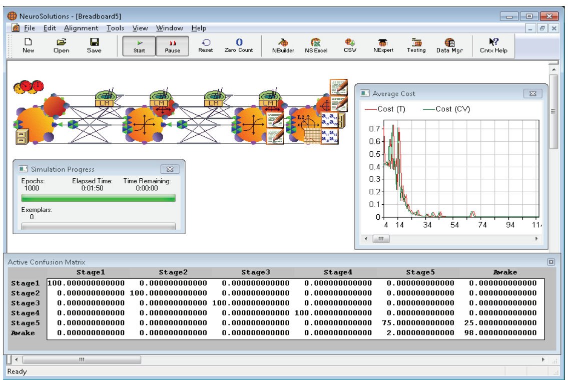
Han, J., Kamber, M. & Pei, J. (2011) *Data Mining: Concepts and Techniques (Third Edition). The Morgan Kaufmann Series in Data Management Systems. San Francisco, Morgan Kaufmann Publishers.*

Slide 6-25 Typical ANN architecture

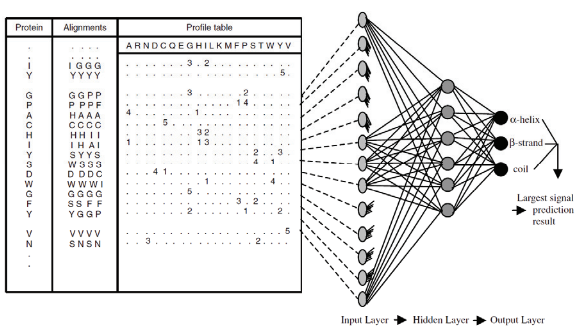


Tangri, N., Ansell, D. & Naimark, D. (2008) Predicting technique survival in peritoneal dialysis patients: comparing artificial neural networks and logistic regression. *Nephrology Dialysis Transplantation, 23, 9, 2972.*

The Tool of the Liver Transplantation Example

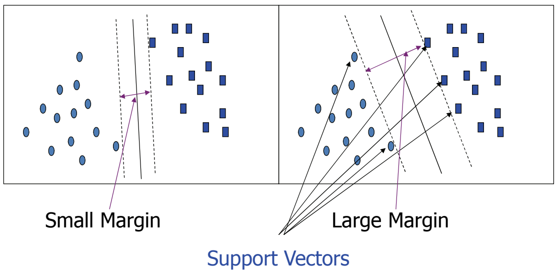


http://www.neurosolutions.com



Hu, X. & Pan, Y. 2007. *Knowledge discovery in bioinformatics: techniques, methods, and applications, Hoboken (NJ), Wiley.*

Support Vectors



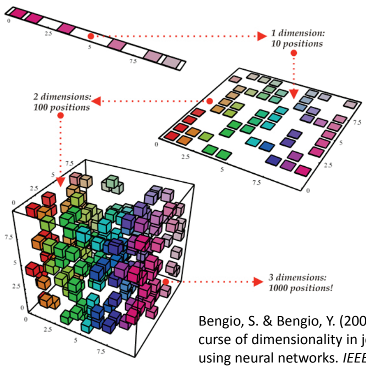
Han, J., Kamber, M. & Pei, J. (2011) Data Mining: Concepts and Techniques (Third Edition). The Morgan Kaufmann Series in Data Management Systems. San Francisco, Morgan Kaufmann Publishers.

Support Vector Machine Demonstration

Video removed due to file size

SVM with a polynominal Kernel – Visualization by Udi Aharoni

Remember: The Curse of Dimensionality



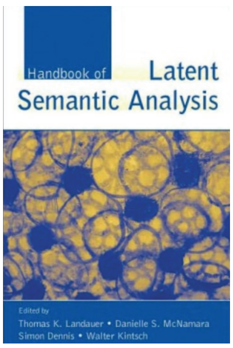
Bengio, S. & Bengio, Y. (2000) Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks, 11, 3, 550-557.*

SVM advantages in high dimensional data

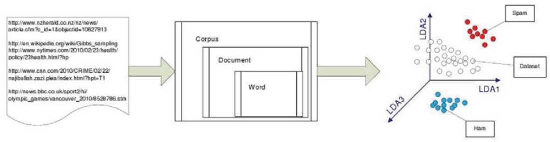
- The complexity of trained classifier is characterized by the # of support vectors – *not the dimensionality of the data*
- The support vectors are the essential or critical training examples —they lie closest to the decision boundary
- If all other training examples are removed and the training is repeated, the same separating hyperplane would be found
- The number of support vectors found can be used to compute an (upper) bound on the expected error rate of the SVM classifier, which is independent of the data dimensionality
- Thus, an SVM with a small number of support vectors can have good generalization, even when the dimensionality of the data is high

Han, J., Kamber, M. & Pei, J. (2011) Data Mining: Concepts and Techniques (Third Edition). The Morgan Kaufmann Series in Data Management Systems. San Francisco, Morgan Kaufmann Publishers.

Slide 6-41: Latent Semantic Analysis (LSA)

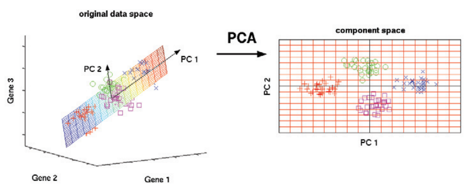


Slide 6-42: Latent Dirichlet allocation (LDA)



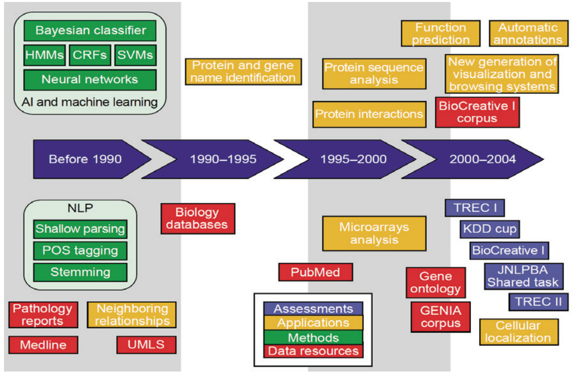
http://www.csmining.org/index.php/latent-dirichlet-allocation.html

Slide 6-43: Principal component analysis (PCA)

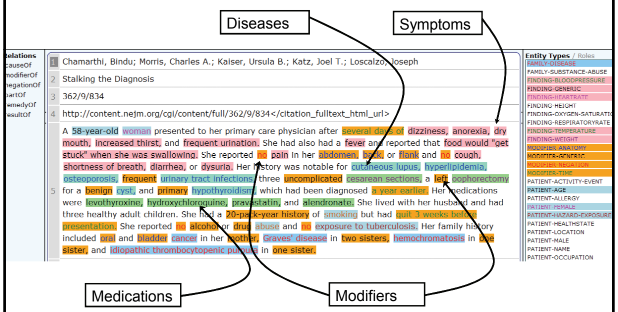


http://www.nlpca.org/pca_principal_component_analysis.html

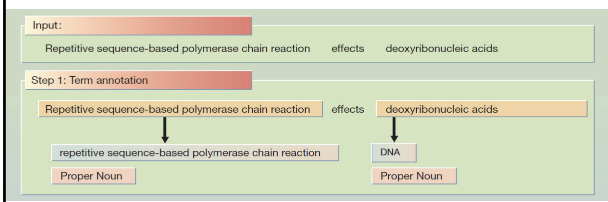
History of using NLP in biomedicine and molecular biology



Krallinger, M., Erhardt, R. A. A. & Valencia, A. (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today, 10, 6, 439-445.*

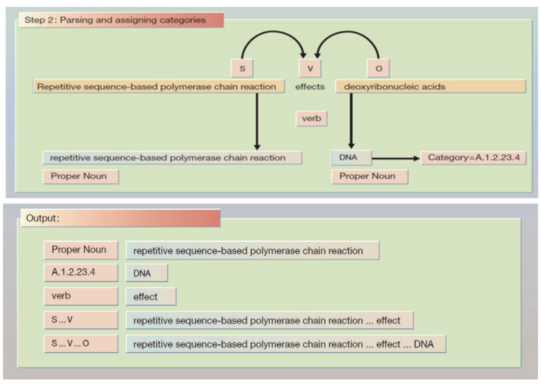


Example: Medical Text Mining process (1)



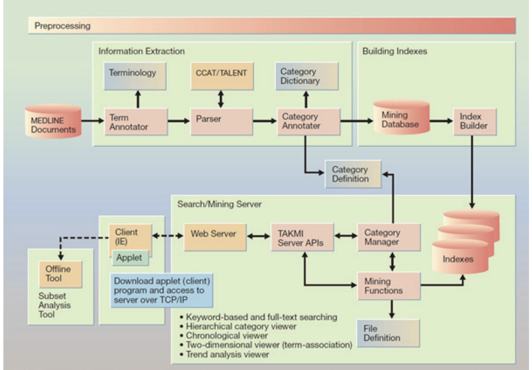
Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. & Takeda, K. (2004) A text-mining system for knowledge discovery from biomedical documents. *IBM SYSTEMS JOURNAL, 43, 3, 516-533.*

Example: Medical Text Mining process (2)



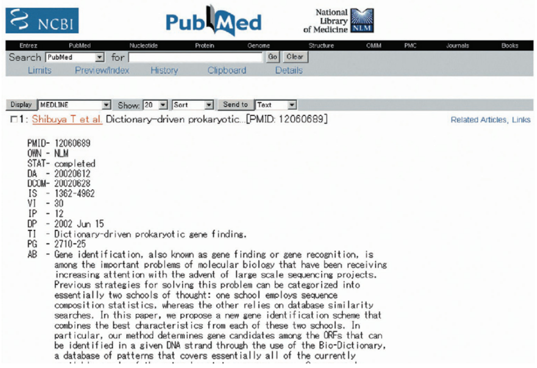
Uramoto, N. et al. (2004)

Example: Medical Text Mining architecture (3)



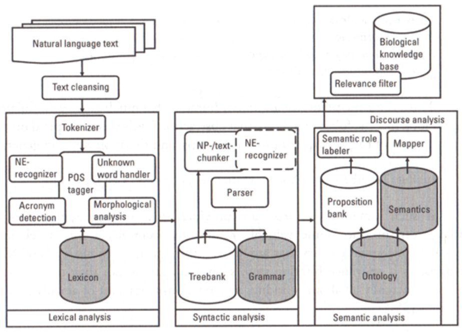
Uramoto et al. (2004)

Example: Medical Text Mining results (4)



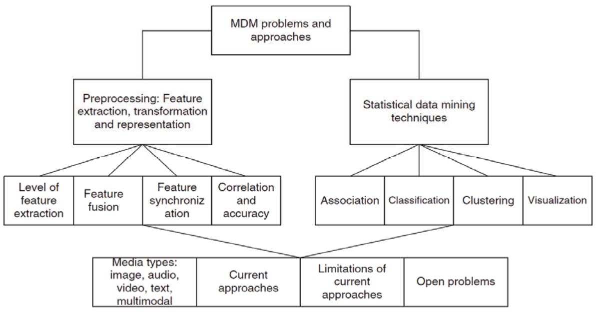
Uramoto et al. (2004)

NLP system architecture for text mining in biology



Hahn, U. & Wermter, J. (2006) Levels of Natural Language Processing for Text Mining. In: Ananiadou & McNaught (Eds.) *Text Mining for Biology and Biomedicine. Boston, Artech, 13-41.*

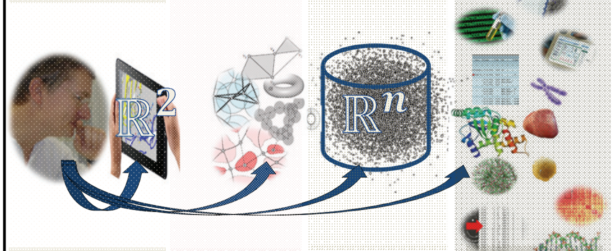
Summary: Multimedia Data Mining Overview



Bhatt, C. & Kankanhalli, M. (2011) Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications, 51, 1, 35-76.*

Slide 6-5 The Knowledge Discovery Process Chain

Hypotheses Generation



HCI, Interactive Visualization, Analytics, Decision Support

Topological Data Mining - Pattern Discovery

Sampling, Cleansing, Preprocessing, Mapping

Data Integration Data Fusion Pre-selection

Privacy, Data Protection, Data Security, Data Safety