Statistic Methods in Data Mining

Introduction

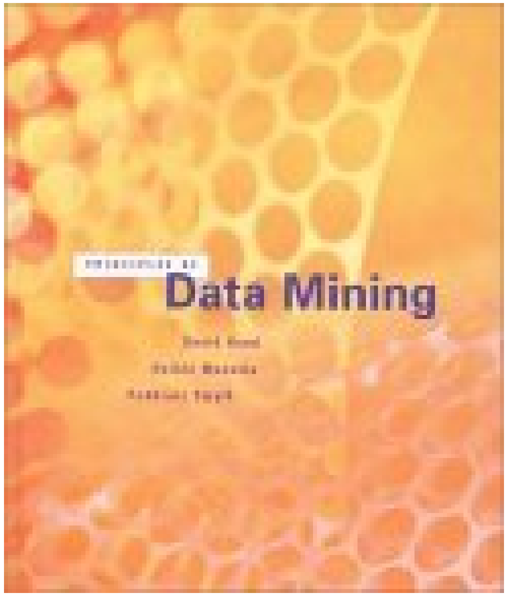
Professor Dr. Gholamreza Nakhaeizadeh

content

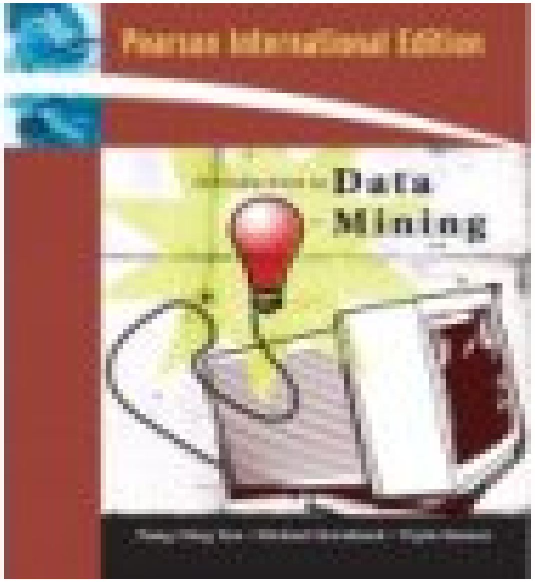
Introduction

- Literature used
- Why Data Mining?
- Examples of large databases
- What is Data Mining?
- Interdisciplinary aspects of Data Mining
- Other issues in recent data analysis: Web Mining, Text Mining
- Typical Data Mining Systems
- Examples of Data Mining Tools
- Comparison of Data Mining Tools
- History of Data Mining, Data Mining: Data Mining rapid development
- Some European funded projects
- Scientific Networking and partnership
- Conferences and Journals on Data Mining
- Further References

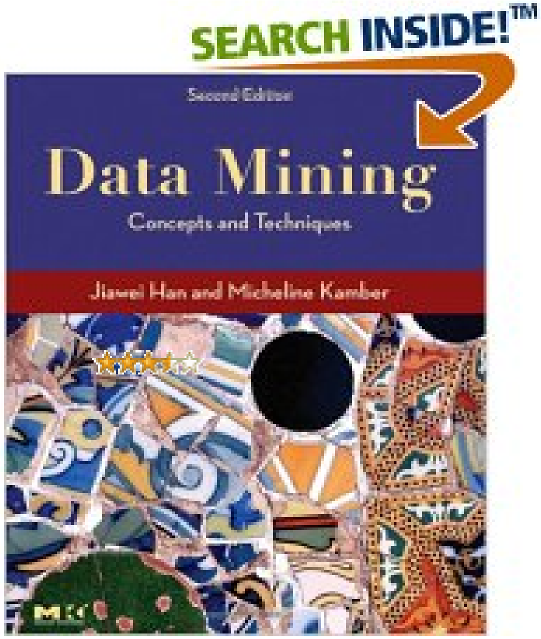
Literatur used (1)



Principles of Data Mining
David J. Hand, Heikki Mannila,
Padhraic Smyth



Pang-Ning Tan,
Michael Steinbach,
Vipin Kumar



Jiawei Han and
Micheline Kamber

Literature Used (2)

http://cse.stanford.edu/class/sophomore-college/projects-00/neural-networks/

http://www.cs.cmu.edu/~awm/tutorials

http://www.crisp-dm.org/CRISPwP-0800.pdf

http://en.wikipedia.org/wiki/Feedforward_neural_network

http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Feedback%20networks

http://www.dmreview.com/

http://www.planet-source-code.com/vb/scripts/ShowCode.asp?lngWId=5&txtCodeId=378

http://download-uk.oracle.com/docs/html/B13915_02/i_olap_chapter.htm#BABCBDFA

http://download-uk.oracle.com/docs/html/B13915_02/i_rel_chapter.htm#BABGFCFG

http://training.inet.com/OLAP/home.htm

http://www.doc.gold.ac.uk/~mas01ds/cis338/index.html

http://wwwmaths.anu.edu.au/~steve/pdcn.pdf

www.kdnuggets.com

The Data Warehouse Toolkit by Ralph Kimball (John Wiley and Sons, 1996)

Building the Data Warehouse by William Inmon (John Wiley and Sons, 1996)

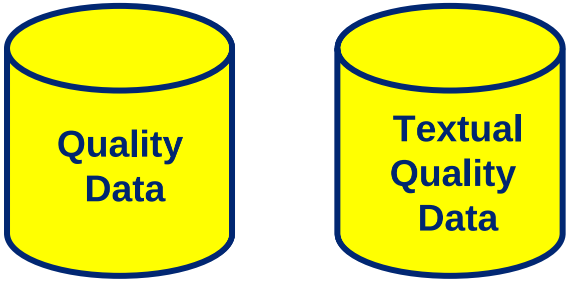
Why Data Mining ? (1)

Huge volume of data, specially, in large companies available:

Product and process data

- **Supplier data**
- **Development data**
- **Production data**
- **Sales data**
- **After sales data**
- **Customer data**
- **Finance data**
- **Employee data**
- **.........**

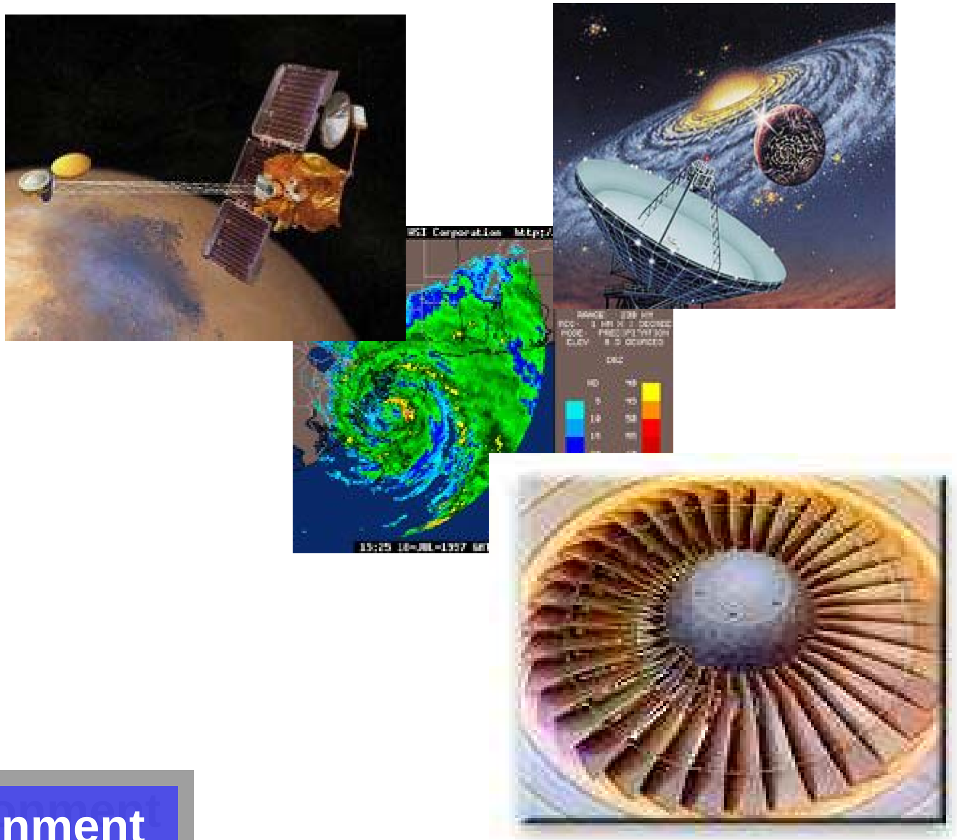
Two examples from Automotive Industry



Data Mining: From high volume data to high value Information

Why Data Mining (2)

- Remote sensor satellite data
- Telescope data
- weather data
- Scientific simulations
- ……



generate terabytes of data in a short time

An interdisciplinary analysis environment is necessary

Data Mining: From high volume data to high value Information

Examples of large databases

Source: http://www.alltop10.net/top-10-list/others/not-specified/top-10-largest-databases-in-the-world.html

World Data Centre for Climate

If you had a 35 million euro super computer lying around what would you use it for? The stock market? Building your own internet? Try extensive climate research – if there's a machine out there that has the answer for global warming, this one might be it. Operated by the Max Planck Institute for Meteorology and German Climate Computing Centre, The World Data Centre for Climate (WDCC) is the largest database in the world. The WDCC boasts 220 terabytes of data readily accessible on the web including information on climate research and anticipated climatic trends, as well as 110 terabytes (or 24,500 DVD's) worth of climate simulation data. To top it off, six petabytes worth of additional information are stored on magnetic tapes for easy access. How much data is six petabyte you ask? Try 3 times the amount of ALL the U.S. academic research libraries contents combined.“

By the Numbers

- 220 terabytes of web data
- 6 petabytes of additional data

Examples of large databases

YouTube

Source: http://www.alltop10.net/top-10-list/others/not-specified/top-10-largest-databases-in-the-world.html

After less than two years of operation YouTube has amassed the largest video library
(and subsequently one of the largest databases) in the world.
YouTube currently boasts a user base that watches more than 100 million clips per
day accounting for more than 60% of all videos watched online.
In August of 2006, the Wall Street Journal projected YouTube's database to the sound
of 45 terabytes of videos. While that figure doesn't sound terribly high relative to the amount
of data available on the internet, YouTube has been experiencing a period of substantial
growth (more than 65,000 new videos per day) since that figures publication, meaning that
YouTube's database size has potentially more than doubled in the last 5 months.
Estimating the size of YouTube's database is particularly difficult due to the varying
sizes and lengths of each video. However if one were truly ambitious (and a bit forgiving)
we could project that the YouTube database will expect to grow as much as 20 terabytes
of data in the next month.
Given: 65,000 videos per day X 30 days per month = 1,950,000 videos per month;
1 terabyte = 1,048,576 megabytes. If we assume that each video has a size of
1MB, YouTube would expect to grow 1.86 terabytes next month. Similarly, i
f we assume that each video has a size of 10MB, YouTube would
expect to grow 18.6 terabytes next month.

By the Numbers

100 million videos watched per day
65,000 videos added each day
60% of all videos watched online
At least 45 terabytes of videos

What is Data Mining ?

One of the most used definition (Fayyad et al 1996):

Knowledge Discovery in Databases (KDD) is a **process** that aims at finding
valid,
useful,
novel
and
understandable
patterns in data

KDD and Data Mining:

- KDD comes originally from AI
- Data Mining is a part of KDD
- In the praxis KDD and Data Mining are used as synonyms

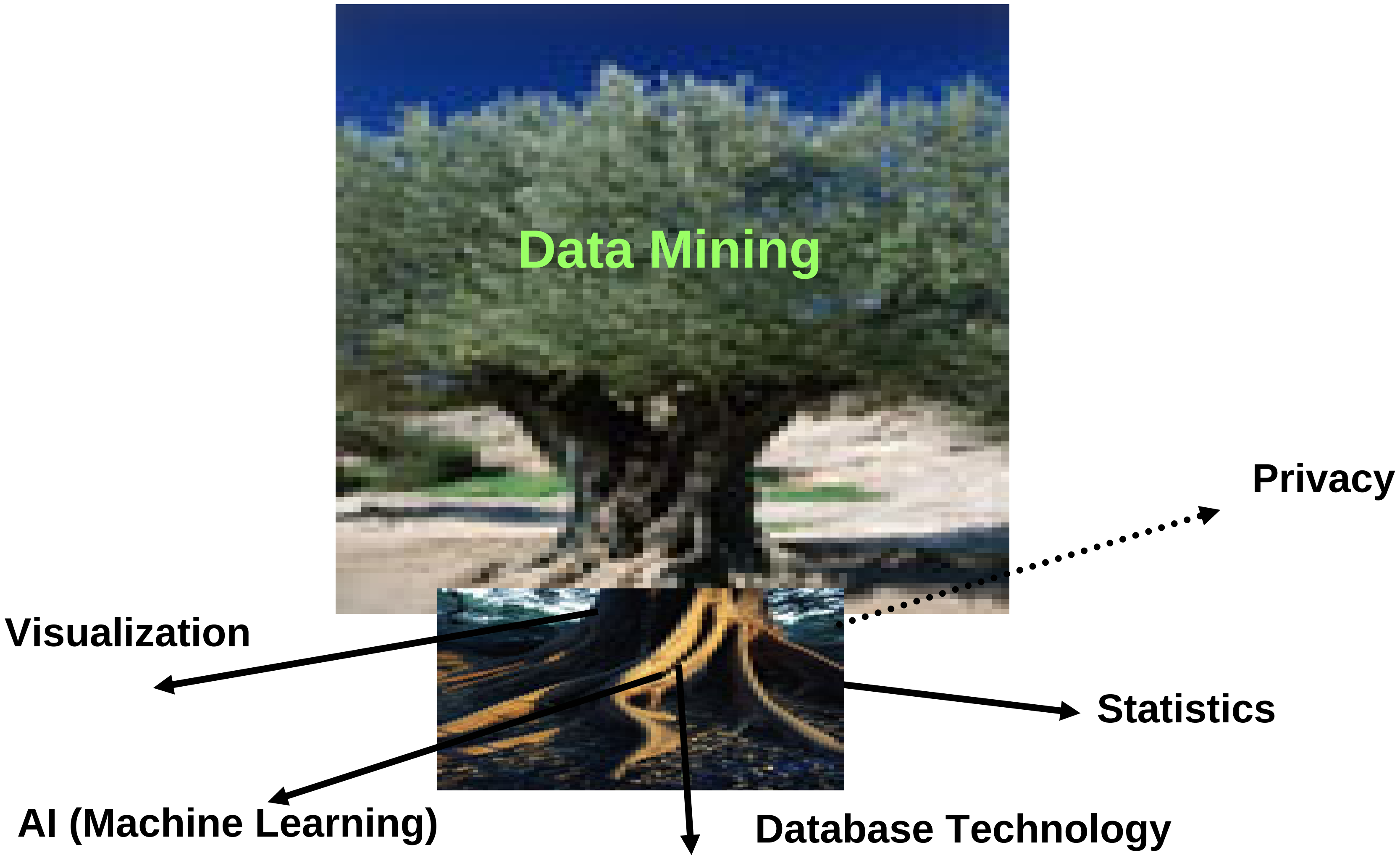
Is a model the same as a pattern?

- Y= 2 + 3X (Generality)
- If country= Iran then carpet export= high (Locality)

Implicit and explicit patterns

Understandable pattern: Rules
Non-understandable: Trained artificial neural networks (ANN)

Interdisciplinary aspects of Data Mining



Data Mining

Privacy

Visualization

Statistics

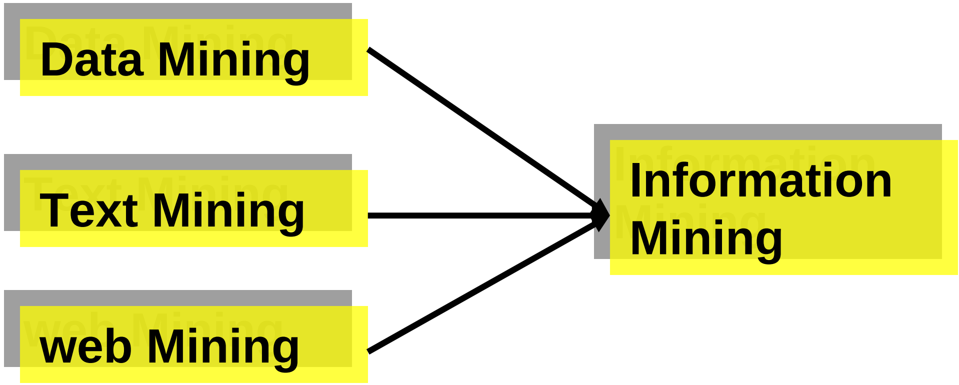
AI (Machine Learning)

Database Technology

Other issues in recent data analysis

- Text Mining
- web Ming

Application of Data Mining Methods to text and web driven data



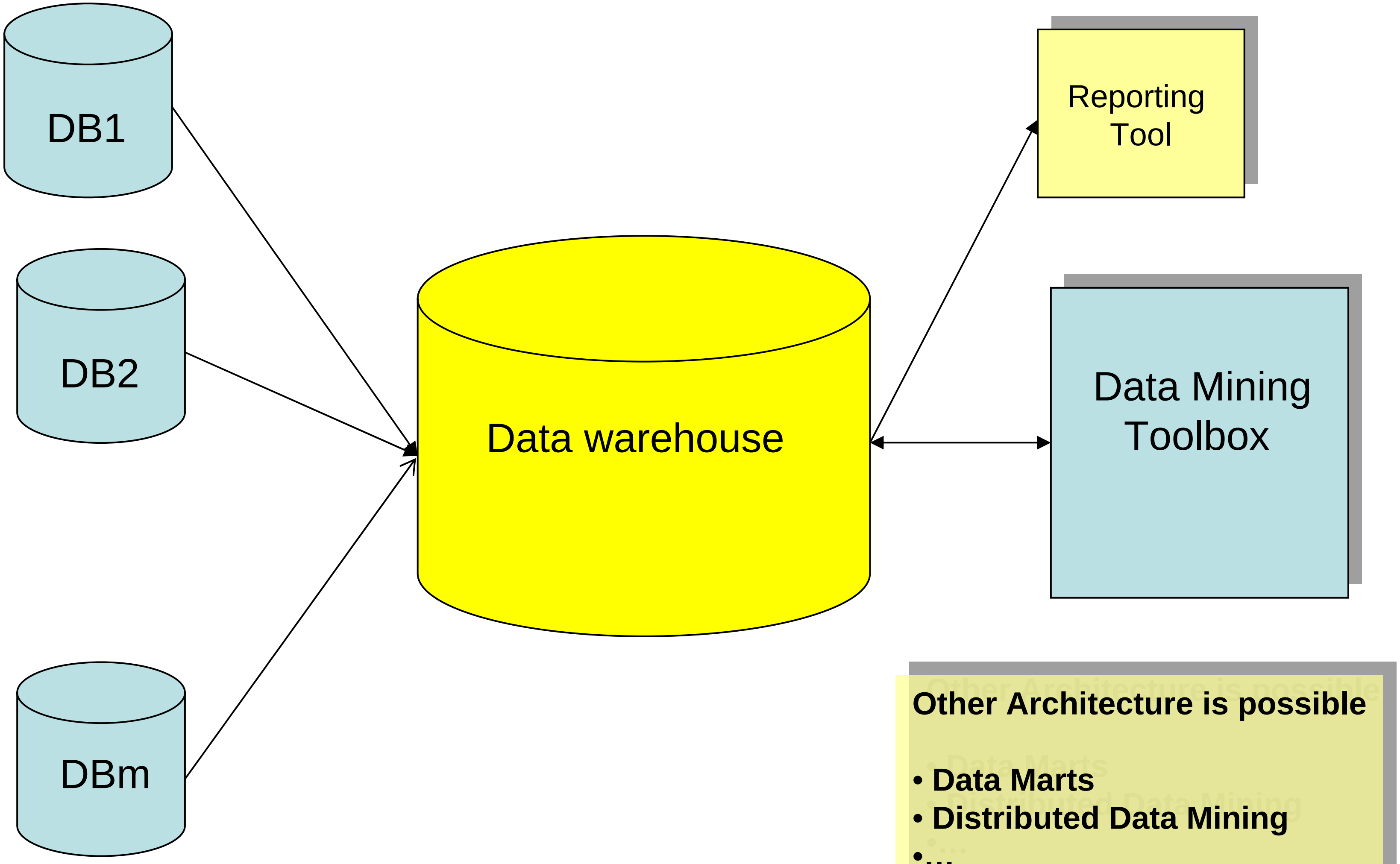
Data Mining

Text Mining

web Mining

Information Mining

Typical Data Mining Systems



Other Architecture is possible

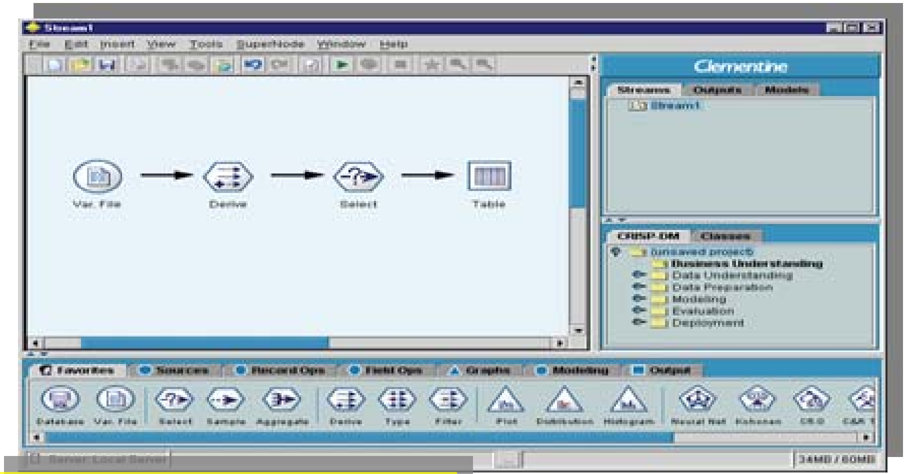
- **Data Marts**
- **Distributed Data Mining**
- **…**

Other Architecture is possible

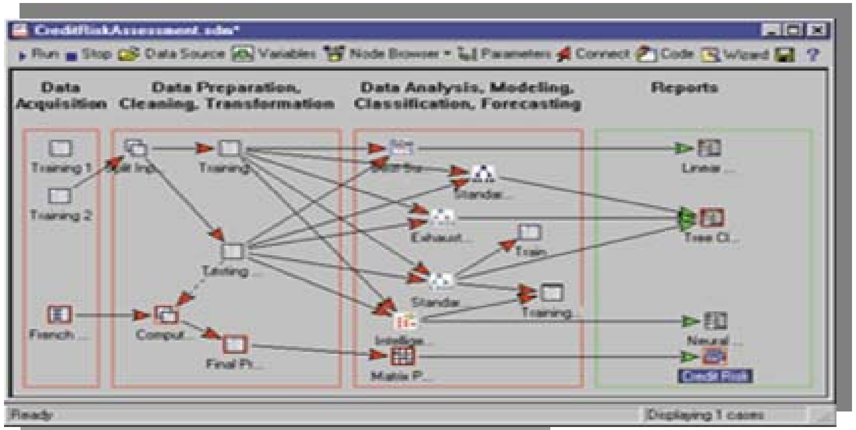
• Data Marts
• Distributed Data Mining
• …

Examples of Data Mining Tools (1)

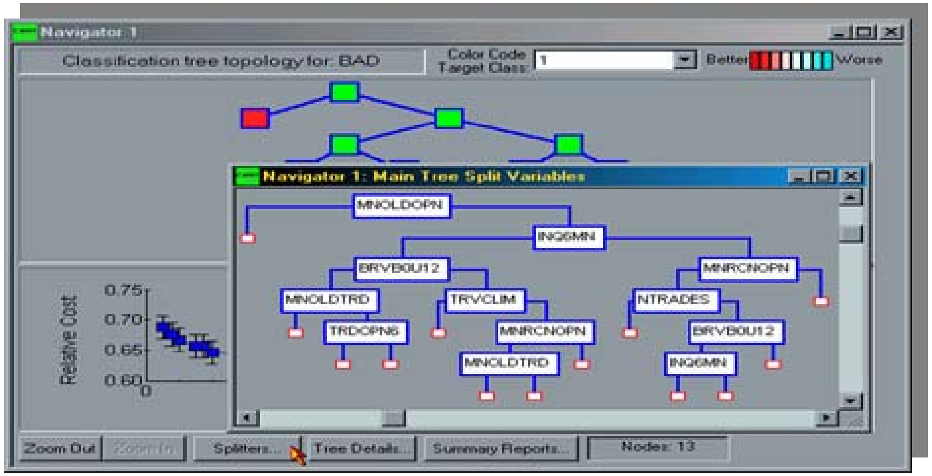
SPSS Clementine


Stream1
File Edit Insert View Tools SuperNode Window Help
Var. File
Derive
Select
Table
Clementine
Streams Outputs Models
Stream1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram Neural Net Kohonen C5.0 C&R T
Server: Local Server
34MB / 60MB

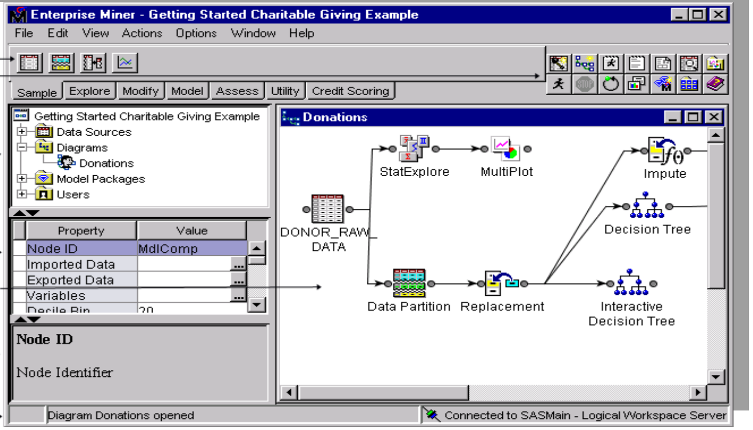
Statistica Data Miner


CreditRiskAssessment.sdm*
Run Stop Data Source Variables Node Browser Parameters Connect Code Wizard
Data Acquisition
Data Preparation, Cleaning, Transformation
Data Analysis, Modeling, Classification, Forecasting
Reports
Training 1
Training 2
French
Training
Testing
Comput
Final Pr
Standar
Exhaust
Train
Training
Intellige
Matrix P
Linear
Tree Cl
Neural
Credit Risk
Ready
Displaying 1 cases

CART



SAS Enterprise Miner


Enterprise Miner - Getting Started Charitable Giving Example
File Edit View Actions Options Window Help
Sample Explore Modify Model Assess Utility Credit Scoring
Getting Started Charitable Giving Example
Data Sources
Diagrams
Donations
Model Packages
Users
Donations
StatExplore
MultiPlot
Impute
DONOR_RAW_DATA
Decision Tree
Data Partition
Replacement
Interactive Decision Tree

Property	Value
Node ID	MdlComp
Imported Data	
Exported Data	
Variables	
Decile Bin	20

Node ID

Node Identifier

Diagram Donations opened

Connected to SASMain - Logical Workspace Server

Toolbar

Toolbar Shortcut Buttons

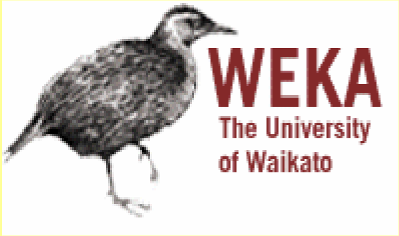
Project Panel

Properties Panel

Diagram Workspace

Help Panel

Status Bar

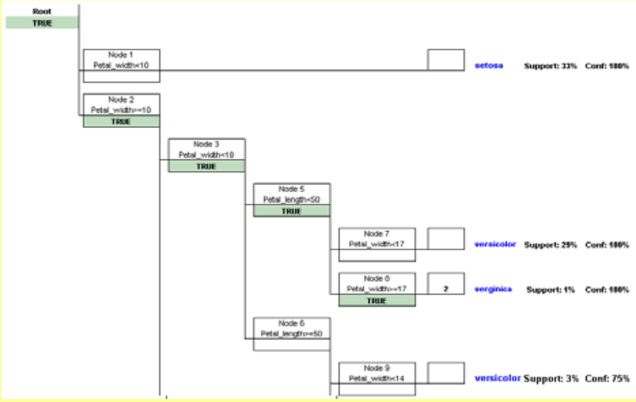


(open source)

Ian witten, Frank Eibe: Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)

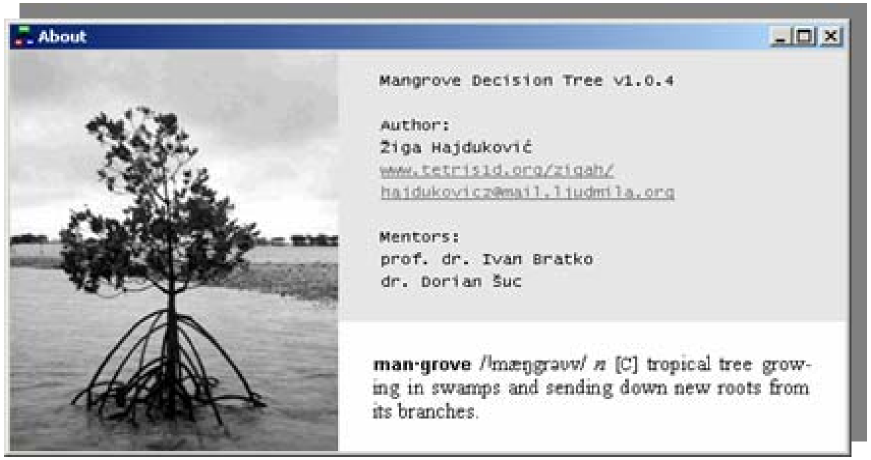
http://www.cs.waikato.ac.nz/ml/weka/

Excel based classification tree tool



http://www.geocities.com/adotsaha/CTree/CtreeinExcel.html

Mangrove Decision Tree



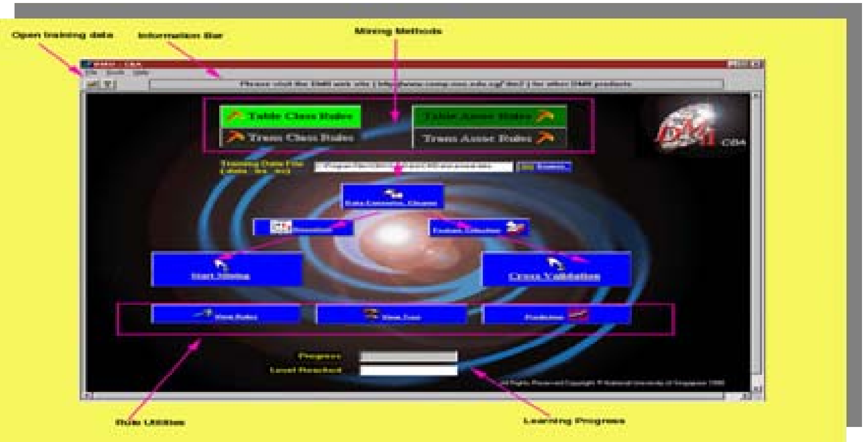
Mangrove Decision Tree v1.0.4

Author:
Žiga Hajduković
www.tetrisid.org/zigah/
hajdukovicz@mail.ljudmila.org

Mentors:
prof. dr. Ivan Bratko
dr. Dorian Šuc

man·grove /ˈmæŋɡrəʊv/ *n* [C] tropical tree growing in swamps and sending down new roots from its branches.

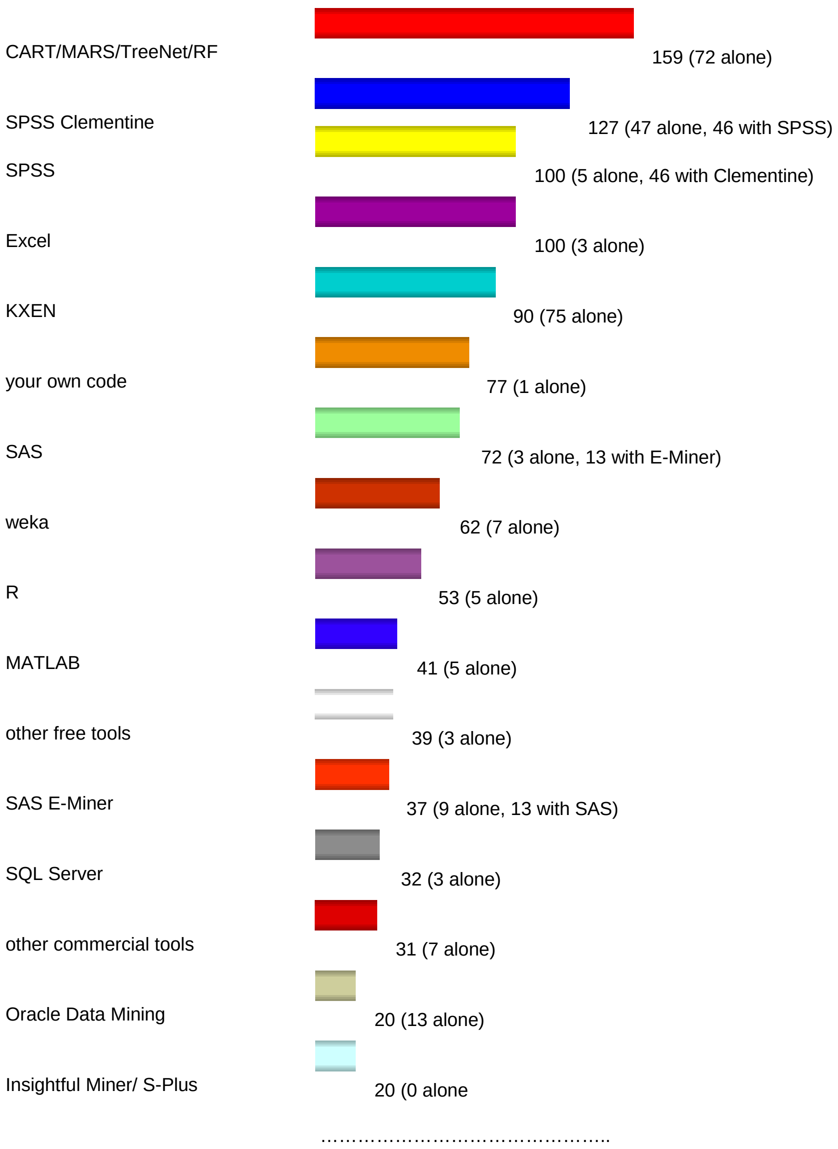
CBA: Classification Based on Association Rules



KDnuggets : Polls : Data Mining (Analytic) Tools (May 2006)

Source: www.kdnuggets.com

Data mining/analytic tools you used in 2006: [561 voters]



Comparison of Data Mining Tools

KDD-98:
A Comparison of Leading Data Mining Tools

John F. Elder IV & Dean W. Abbott
Elder Research

Fourth International Conference
on Knowledge Discovery & Data Mining

Friday, August 28, 1998
New York, New York

Comparison of Data Mining Tools

Source: http://web.cs.wpi.edu/~ruiz/KDDRG/dm_tools.html

[Knowledge Discovery and Data Mining Research Group KDDRG]()

Project on

Comparing Data Mining Tools and Systems

COMPARING THE EFFECTIVENESS OF MINESET AND INTELLIGENT MINER IN KNOWLEDGE EXTRACTION

Project Members

- **Faculty:** Carolina Ruiz, Matt Ward.
- Students: Chris Martino.

Project Description

The primary goal of this project was to compare two commercial data mining packages: IBM's Intelligent Miner and SGI's MineSet, using association rules and decision trees as a basis. The main factors evaluated were ease of use, overall performance, and the presentation of results. To accomplish this, both packages were used to mine identical data sets and the results were compared.

Some European funded Projects

- StatLog
- CRISP-DM
- INRECA
- MetaL
- READ
- Data Mining Grid

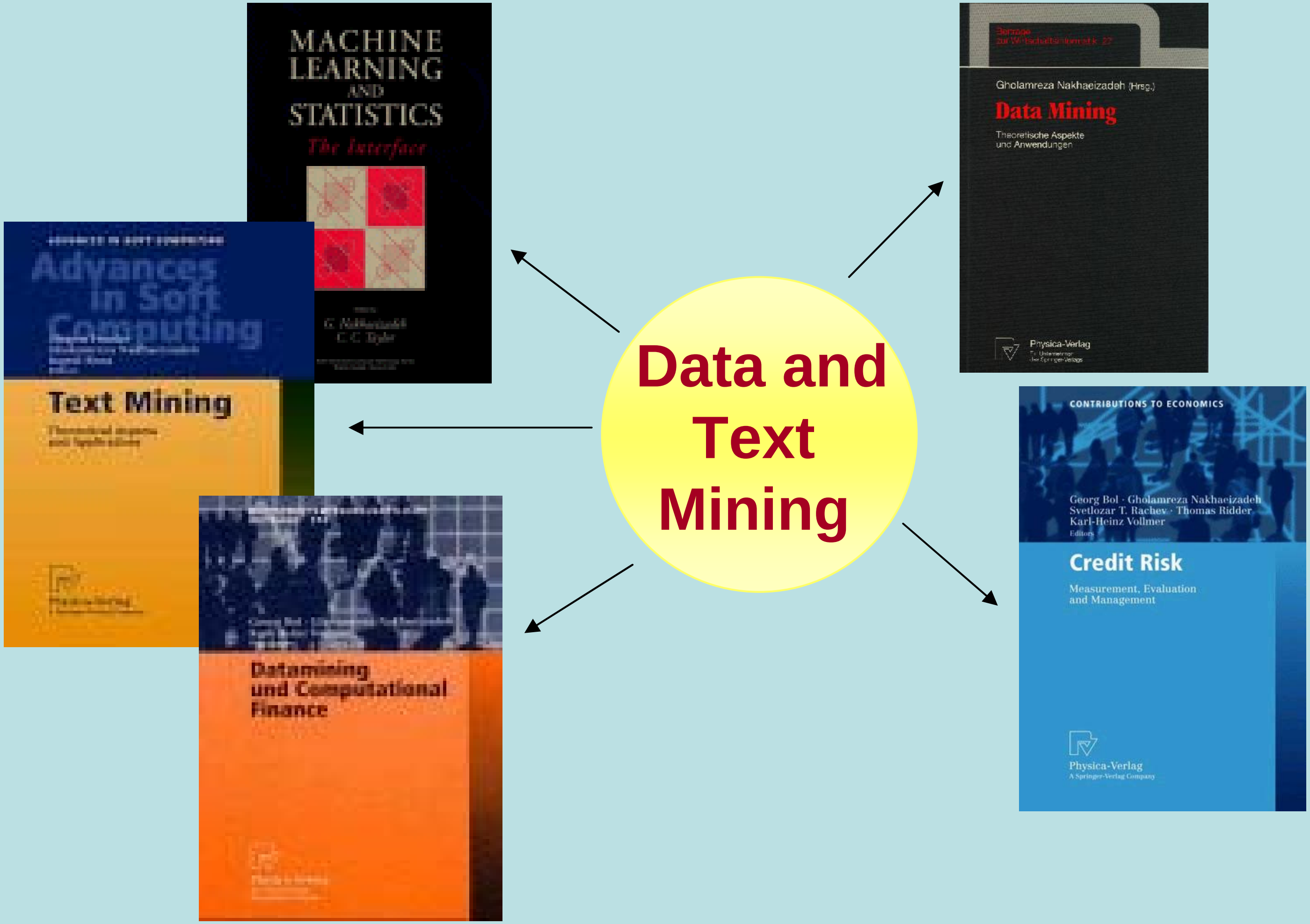
Scientific Networking

1994-2001 European Network of Excellence in Machine Learning

2002-2005 European Network of Excellence in Knowledge Discovery

Since 2005 Ubiquitous Knowledge Discovery

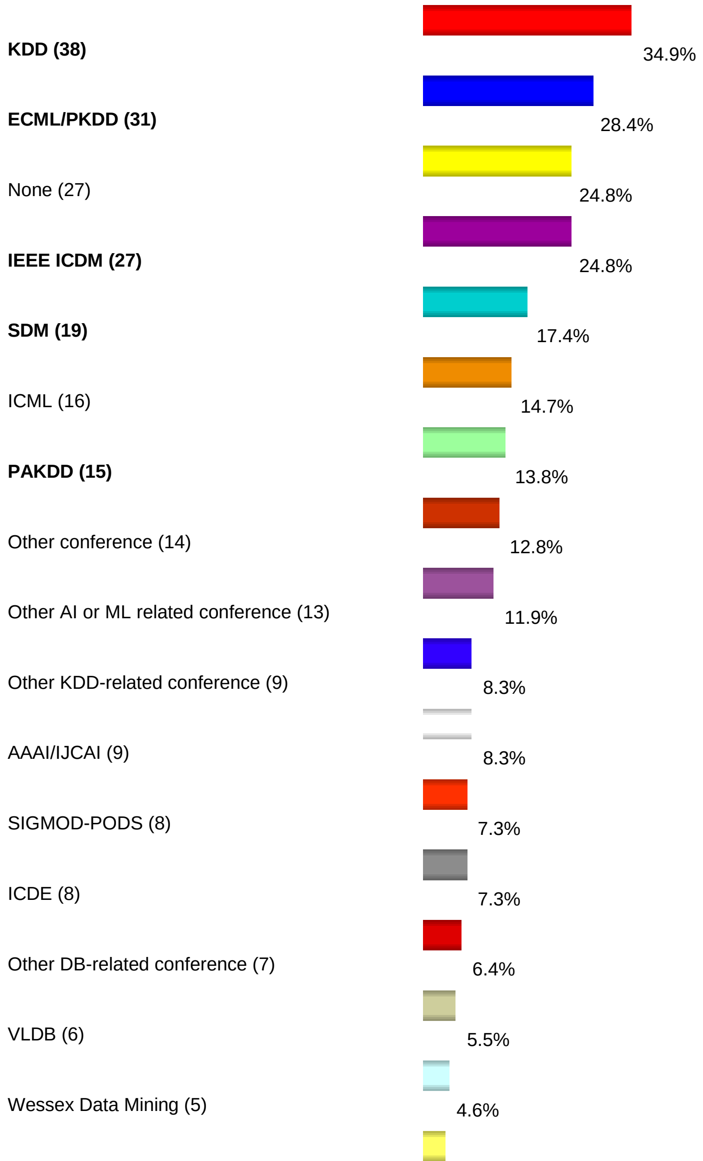
Selected Books


Data and Text Mining
MACHINE LEARNING AND STATISTICS
The Interface
G. Nakhaeizadeh
C. C. Taylor
Advances in Soft Computing
Text Mining
Theoretical Aspects and Applications
Physica-Verlag
Datamining und Computational Finance
Physica-Verlag
Gholamreza Nakhaeizadeh (Hrsg.)
Data Mining
Theoretische Aspekte und Anwendungen
Physica-Verlag
CONTRIBUTIONS TO ECONOMICS
Georg Bol · Gholamreza Nakhaeizadeh
Svetlozar T. Rachev · Thomas Ridder
Karl-Heinz Vollmer
Editors
Credit Risk
Measurement, Evaluation and Management
Physica-Verlag
A Springer-Verlag Company

KDnuggets : Polls : Conferences papers were submitted to (Feb 2008)

Source: wwww.kdnuggets.com

To which conferences did you submit a paper in the last 2 years: [109 voters total]



Conferences

- KDD
- PKDD-ECML
- SIAM-Data Mining
- ICDM,
- PAKDD
- ICML
- ……

Journals

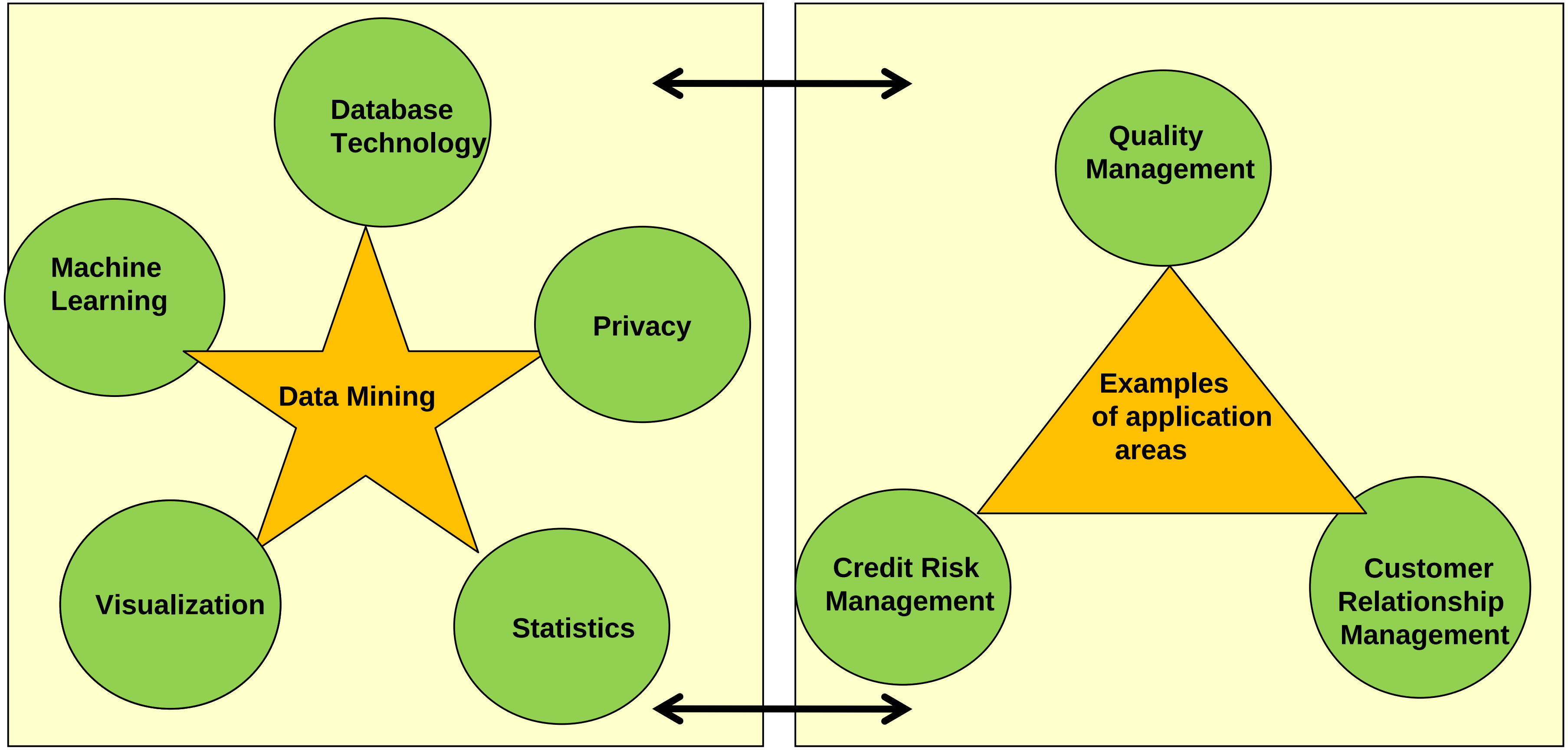
- ACM Transactions on KDD (New)
- IEEE Transactions On Knowledge and Data Engineering
- KDD Explorations
- Data Mining and Knowledge Discovery
- Machine Learning
- …

Further References

- Michael Berry & Gordon Linoff, Mastering Data Mining, John wiley & Sons, 2000.
- Patricia Cerrito, Introduction to Data Mining Using SAS Enterprise Miner, ISBN: 978-1-59047-829-5, SAS Press, 2006.
- K. Cios, w. Pedrycz, R. Swiniarski, L. Kurgan, Data Mining: A Knowledge Discovery Approach, Springer, ISBN: 978-0-387-33333-5, 2007.
- Margaret Dunham, Data Mining Introductory and Advanced Topics, ISBN: 0130888923, Prentice Hall, 2003.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996 (order on-line from Amazon.com or from MIT Press).
- Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, 2nd edition, Morgan Kaufmann, ISBN 1558609016, 2006.
-
- David J. Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining , MIT Press, Fall 2000
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Verlag, 2001.
- Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, ISBN: 0471228524, wiley-IEEE Press, 2002.
- Daniel T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, ISBN: 0471666572, John wiley, 2004 (see also companion site for Larose book).
- Glenn J. Myatt, Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, John wiley, ISBN: 0-470-07471-X, November 2006.
- Olivia Parr Rud, Data Mining Cookbook, modeling data for marketing, risk, and CRM. wiley, 2001.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison wesley (May, 2005). Hardcover: 769 pages. ISBN: 0321321367
- Ripley, B.D. (1996) Pattern Recognition and Neural Networks, Cambridge: Cambridge University Press.
- Sholom M. weiss and Nitin Indurkhya, Predictive Data Mining: A Practical Guide, Morgan Kaufmann, 1997
- Graham williams, Data Mining Desktop Survival Guide, on-line book (PDF).
- Ian witten and Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufman, ISBN 1558605525, 1999.
- Ian witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, ISBN 0120884070, 2005.

Examples of Data Mining applications in industry and commerce

Optimal structure of a Data Mining Team



Success Factors of DM-Applications

KDD-95 panel on Commercial KDD Applications: The "Secret" Ingredients for Success

Sunday, August 20, 1:30 -- 2:30 pm, Palais Des Congres, Montreal, Canada Position statements of:

- Tej Anand, AT&T GIS
- Dr. Gholamreza Nakhaeizadeh, Daimler-Benz
- Evangelos Simoudis, IBM, co-chair
- Gregory Piatetsky-Shapiro, GTE Laboratories, co-chair
- Ralphe wiggins, statement Harvesting
- Kamran Parsaye, statement Discovery
- Mario Schkolnick, SGI

Source: http://www-aig.jpl.nasa.gov/public/kdd95/KDD95-Panels.html

Success Parameters of Data Mining Solutions

□ Clear defined goals

□ Importance of the business problem

□ Management attention and support

□ Competence of the Data Mining team

□ Data availability and quality

□ Close cooperation between the Data Mining team and the end-users

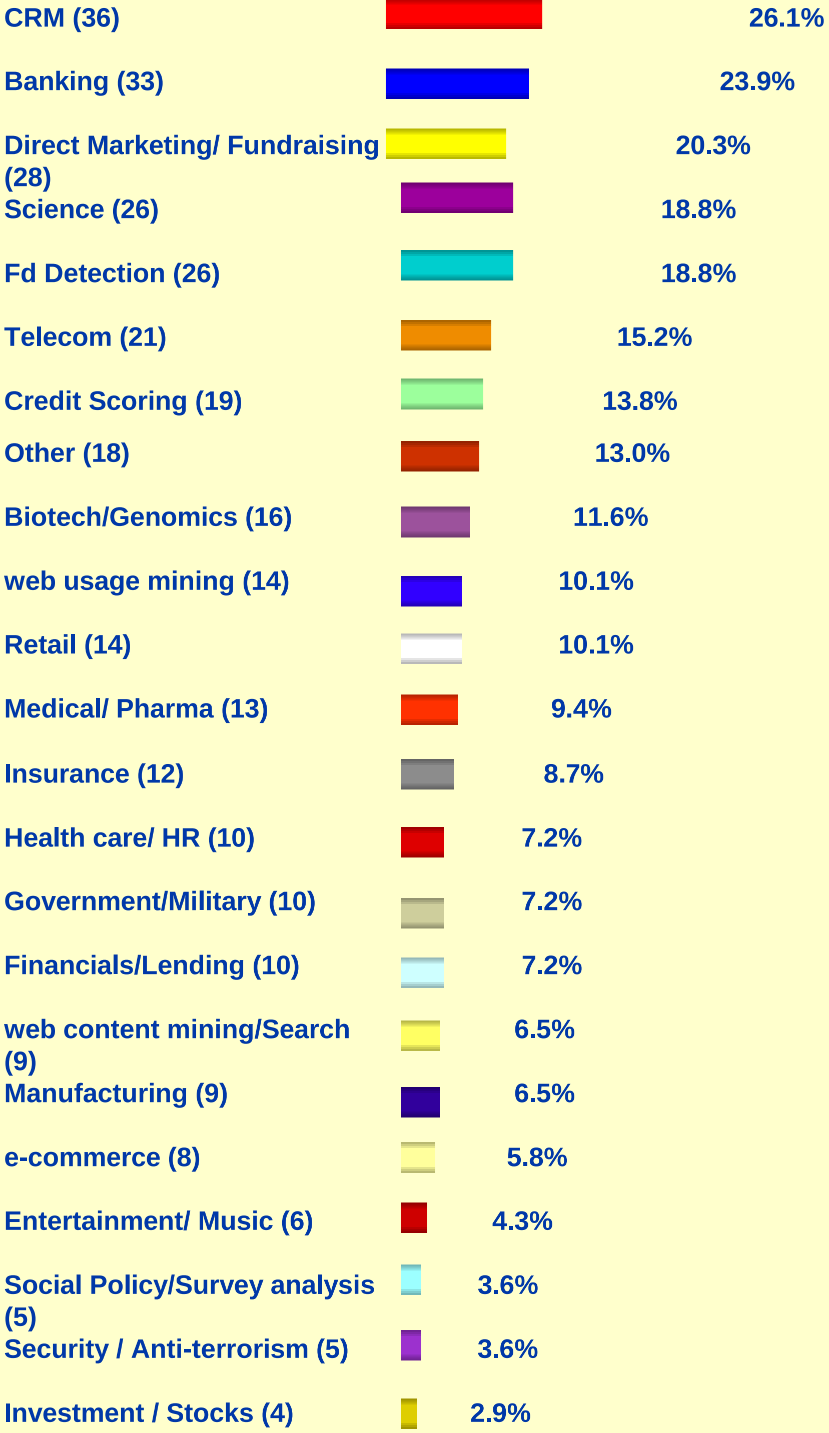
□ Integration of the Data Mining Solution in the daily business process of the users

□ Other parameters (Please describe briefly) where you applied data mining in the past 12 months

Source:

http://www.kdnuggets.com/polls/2007 data_mining_applications.htm

Potential Applications



Predictive Modeling

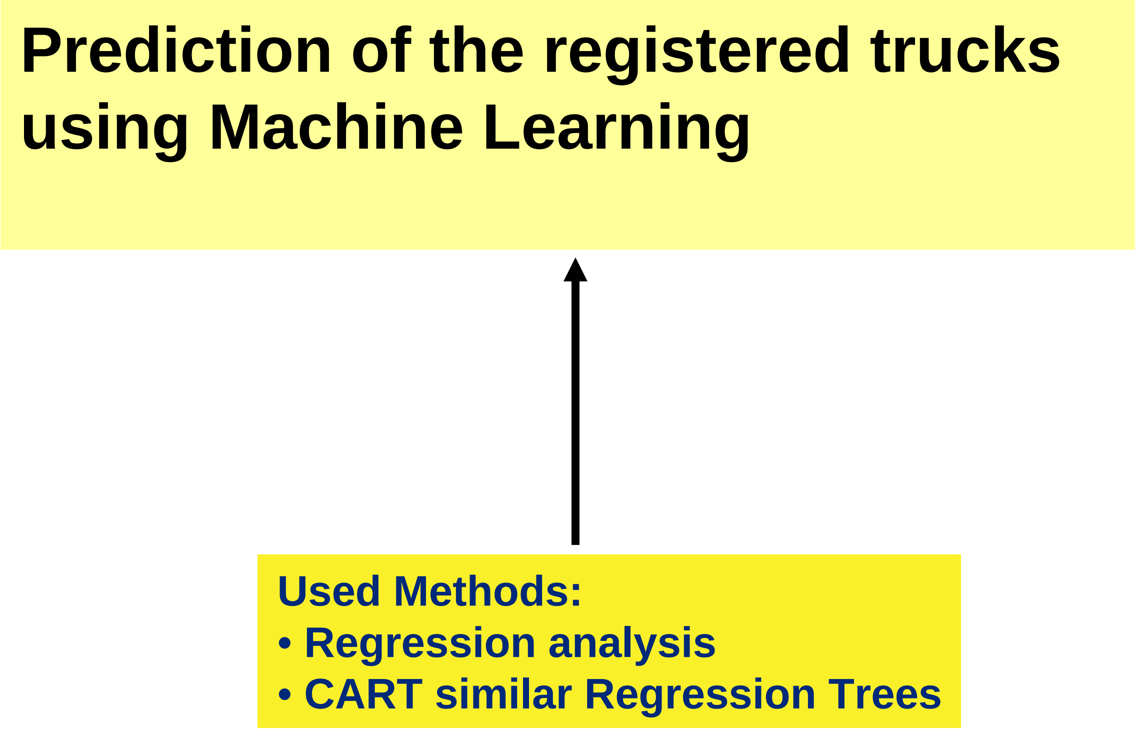
Predictive Modeling as an important component of CRM

work of statisticians such as Fisher in thirties in the area Discriminant analysis

Time series-referred and other prognosis procedures, 1950+

New impulse by DATA Mining 1989+

Application in Business & Banking (1)

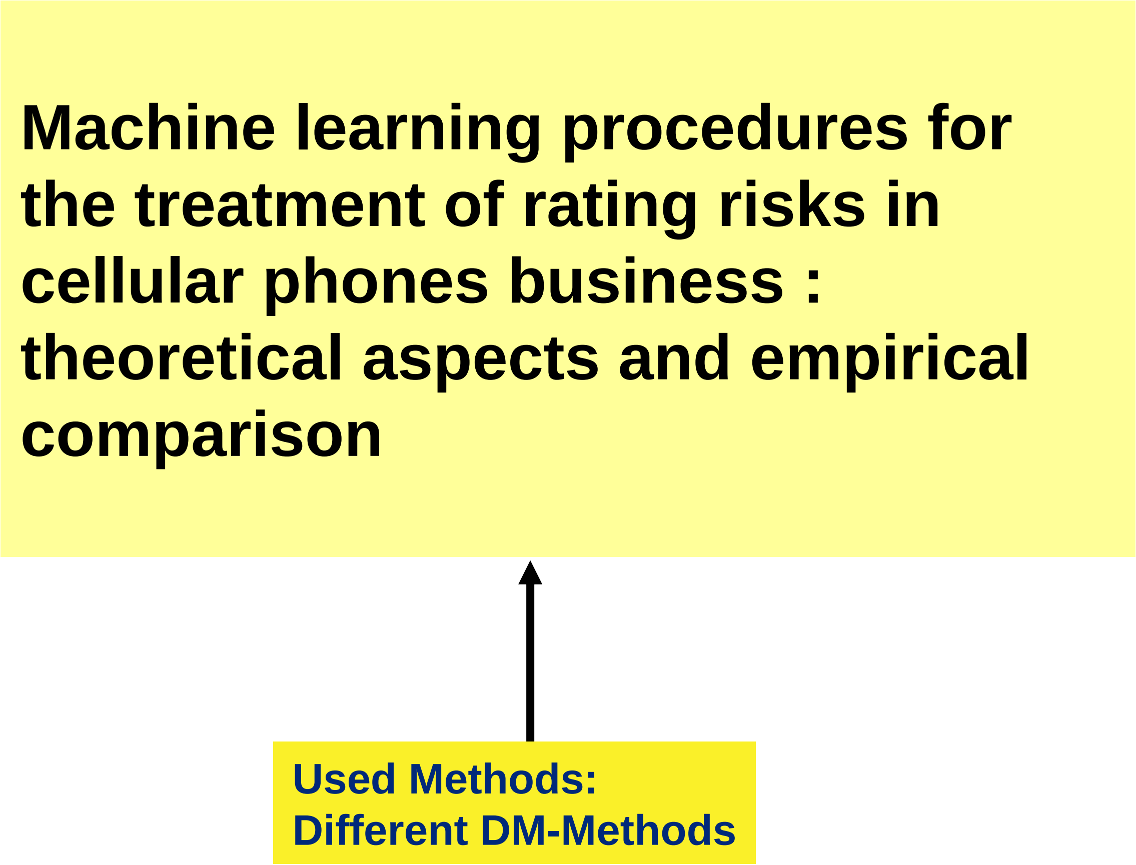


Prediction of the registered trucks using Machine Learning

Used Methods:

- Regression analysis
- CART similar Regression Trees

Application in Business & Banking (2)



Machine learning procedures for the treatment of rating risks in cellular phones business : theoretical aspects and empirical comparison

Used Methods:
Different DM-Methods

Application in Business & Banking (3)

Customer Value:

Value Oriented Customers Acquisition in the Automotive Industry

Prediction of options Order using ANN and statistical Methods

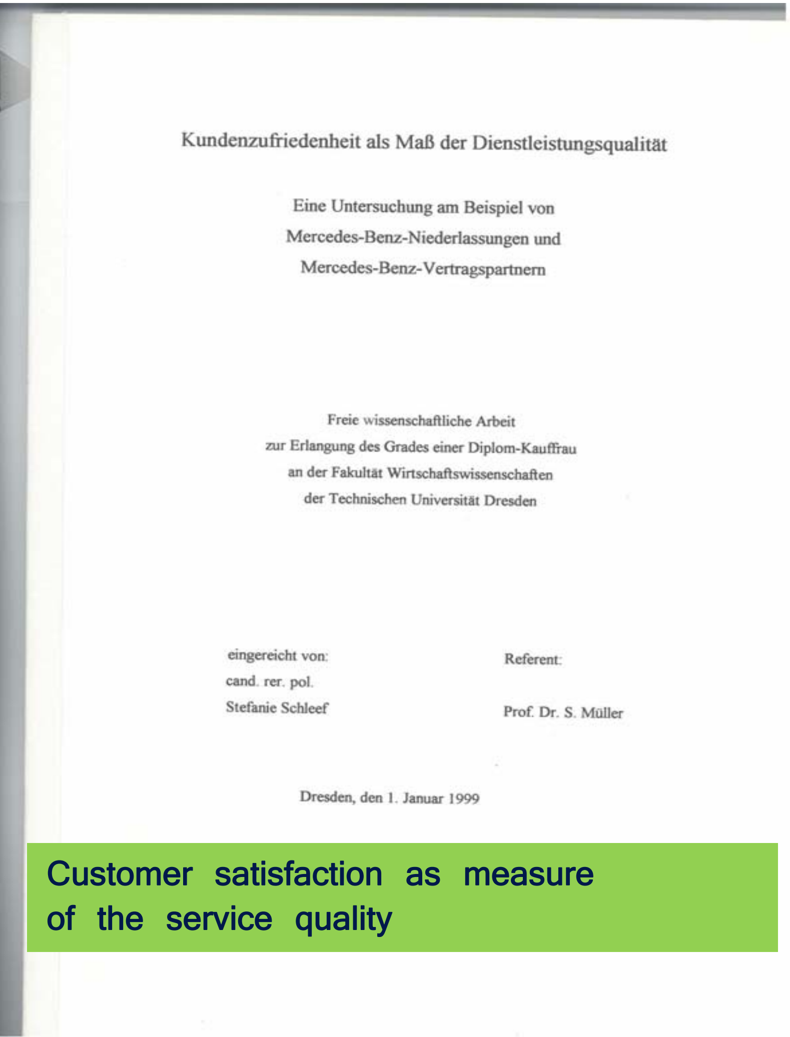
Application in Business & Banking (4)

WAPS: a Data Mining Support Environment for the Planning of Warranty and Goodwill Costs in the Automobile Industry

Used Methods:
Regression analysis

http://portal.acm.org/citation.cfm?id=312307

Application in Business & Banking (5)



Kundenzufriedenheit als Maß der Dienstleistungsqualität

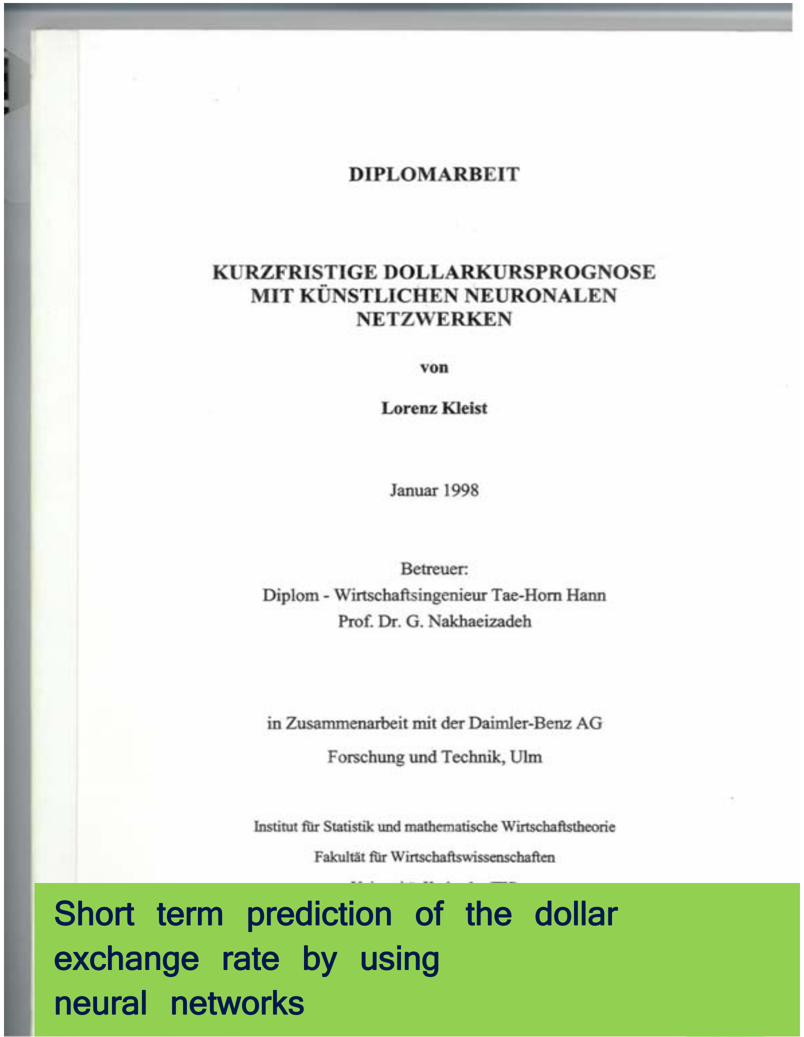
Eine Untersuchung am Beispiel von
Mercedes-Benz-Niederlassungen und
Mercedes-Benz-Vertragspartnern

Freie wissenschaftliche Arbeit
zur Erlangung des Grades einer Diplom-Kauffrau
an der Fakultät Wirtschaftswissenschaften
der Technischen Universität Dresden

eingereicht von:	Referent:
cand. rer. pol.	
Stefanie Schleef	Prof. Dr. S. Müller

Dresden, den 1. Januar 1999

Customer satisfaction as measure of the service quality



DIPLOMARBEIT

KURZFRISTIGE DOLLARKURSPROGNOSE
MIT KÜNSTLICHEN NEURONALEN
NETZWERKEN

von

Lorenz Kleist

Januar 1998

Betreuer:
Diplom - Wirtschaftsingenieur Tae-Horn Hann
Prof. Dr. G. Nakhaeizadeh

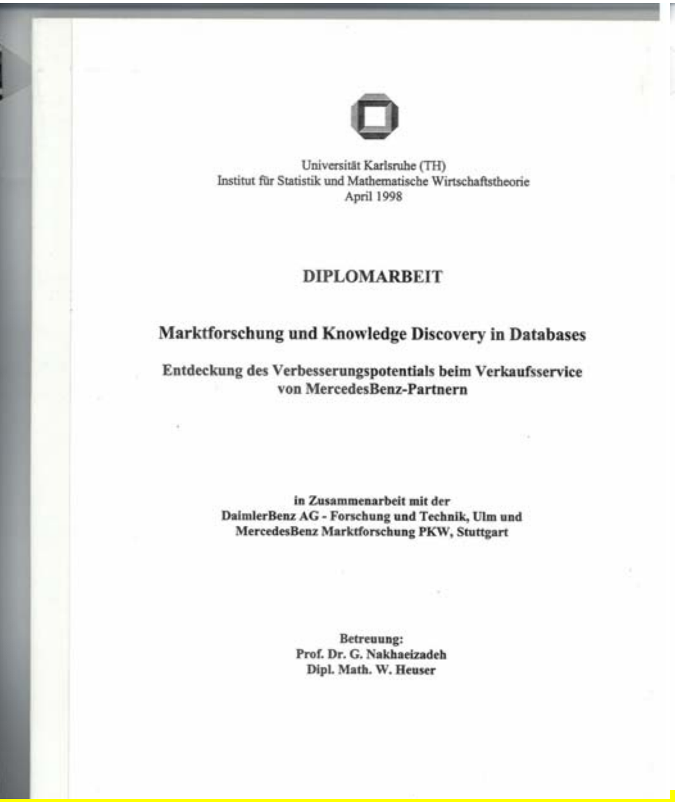
in Zusammenarbeit mit der Daimler-Benz AG

Forschung und Technik, Ulm

Institut für Statistik und mathematische Wirtschaftstheorie
Fakultät für Wirtschaftswissenschaften

Short term prediction of the dollar exchange rate by using neural networks

Application in Business & Banking (6)



Universität Karlsruhe (TH)
Institut für Statistik und Mathematische Wirtschaftstheorie
April 1998

DIPLOMARBEIT

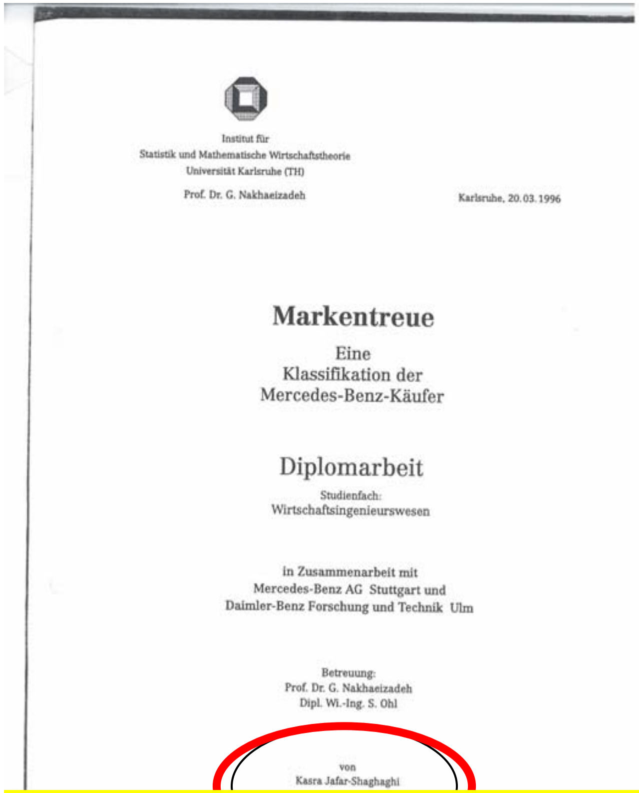
Marktforschung und Knowledge Discovery in Databases

Entdeckung des Verbesserungspotentials beim Verkaufsservice von MercedesBenz-Partnern

in Zusammenarbeit mit der
DaimlerBenz AG - Forschung und Technik, Ulm und
MercedesBenz Marktforschung PKW, Stuttgart

Betreuung:
Prof. Dr. G. Nakhaeizadeh
Dipl. Math. W. Heuser

Market Research and Knowledge Discovery in Databases



Institut für
Statistik und Mathematische Wirtschaftstheorie
Universität Karlsruhe (TH)

Prof. Dr. G. Nakhaeizadeh

Karlsruhe, 20.03.1996

Markentreue

Eine
Klassifikation der
Mercedes-Benz-Käufer

Diplomarbeit

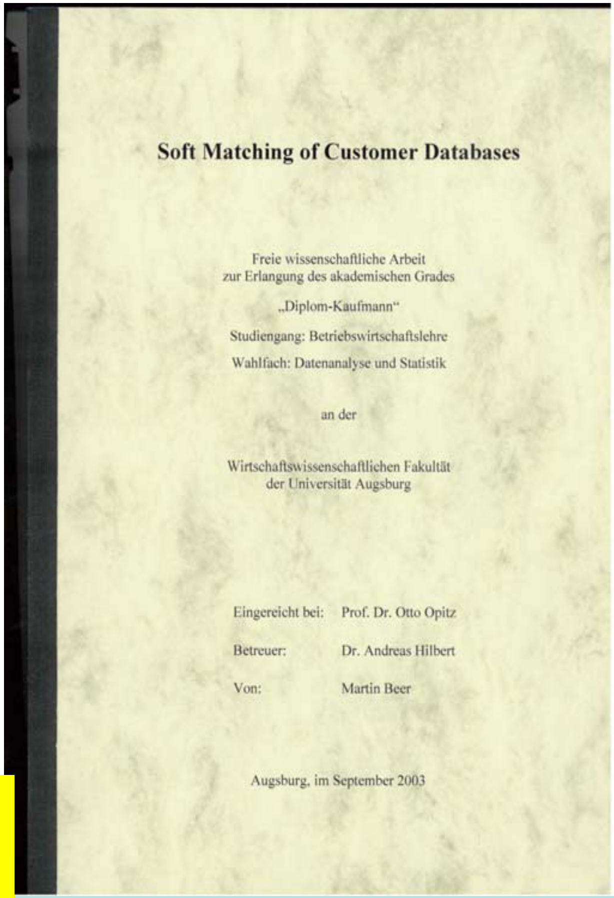
Studienfach:
Wirtschaftsingenieurswesen

in Zusammenarbeit mit
Mercedes-Benz AG Stuttgart und
Daimler-Benz Forschung und Technik Ulm

Betreuung:
Prof. Dr. G. Nakhaeizadeh
Dipl. Wi.-Ing. S. Ohl

von
Kasra Jafar-Shaghaghi

Brand Loyalty
A classification
of Mercedes-Benz Buyers



Soft Matching of Customer Databases

Freie wissenschaftliche Arbeit
zur Erlangung des akademischen Grades

„Diplom-Kaufmann“

Studiengang: Betriebswirtschaftslehre

Wahlfach: Datenanalyse und Statistik

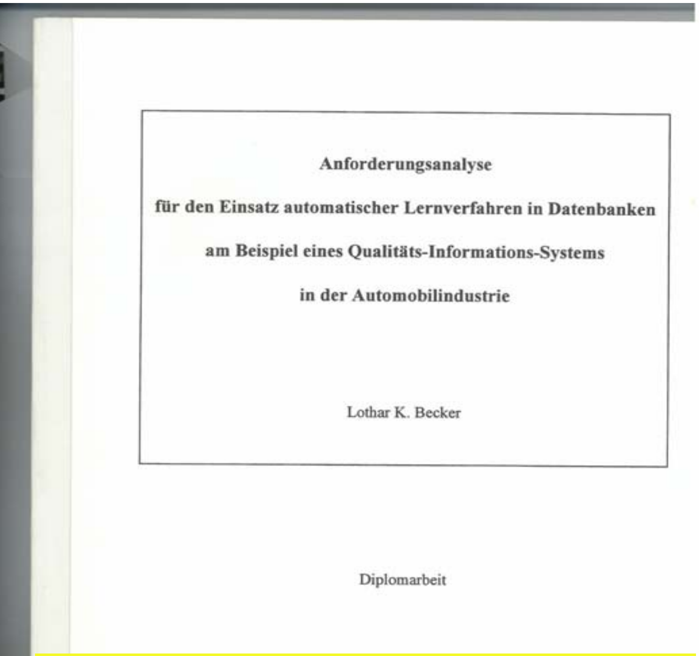
an der

Wirtschaftswissenschaftlichen Fakultät
der Universität Augsburg

Eingereicht bei:	Prof. Dr. Otto Opitz
Betreuer:	Dr. Andreas Hilbert
Von:	Martin Beer

Augsburg, im September 2003

Application in Business & Banking (7)



Anforderungsanalyse

für den Einsatz automatischer Lernverfahren in Datenbanken

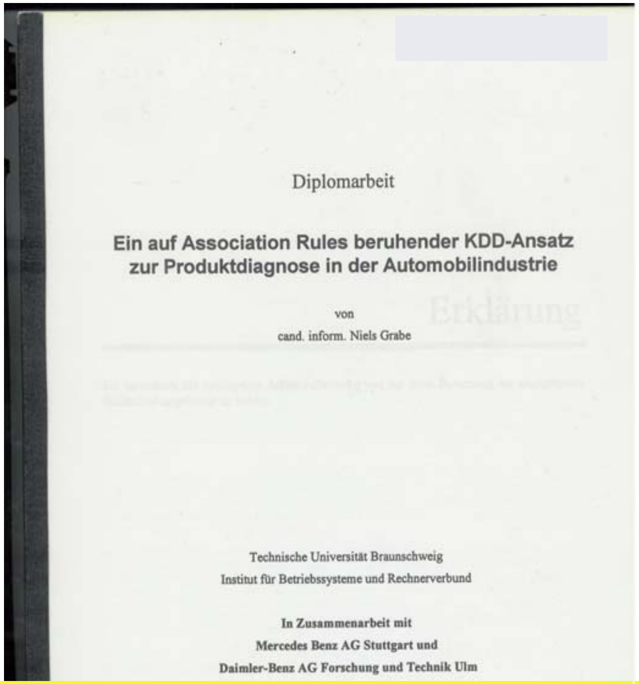
am Beispiel eines Qualitäts-Informations-Systems

in der Automobilindustrie

Lothar K. Becker

Diplomarbeit

Requirement analysis for the application of automatic learning procedures in data bases by the example of a quality information system



Diplomarbeit

Ein auf Association Rules beruhender KDD-Ansatz
zur Produktdiagnose in der Automobilindustrie

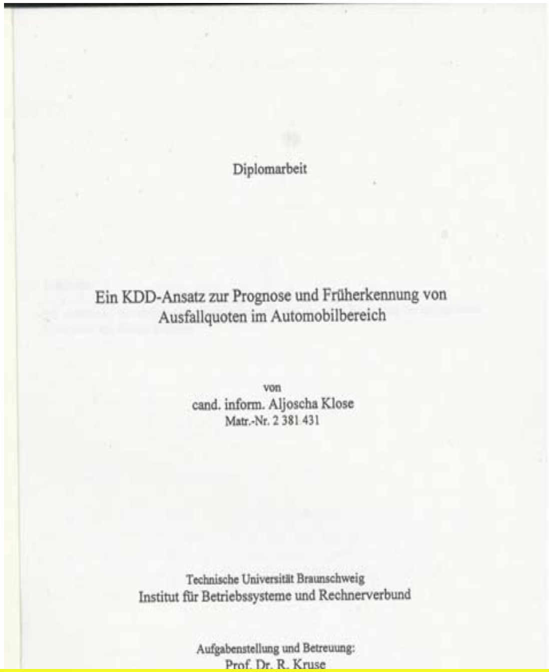
von

cand. inform. Niels Grabe

Technische Universität Braunschweig
Institut für Betriebssysteme und Rechnerverbund

In Zusammenarbeit mit
Mercedes Benz AG Stuttgart und
Daimler-Benz AG Forschung und Technik Ulm

A KDD-approach based on association rules for the product diagnosis in the automobile industry



Diplomarbeit

Ein KDD-Ansatz zur Prognose und Früherkennung von
Ausfallquoten im Automobilbereich

von

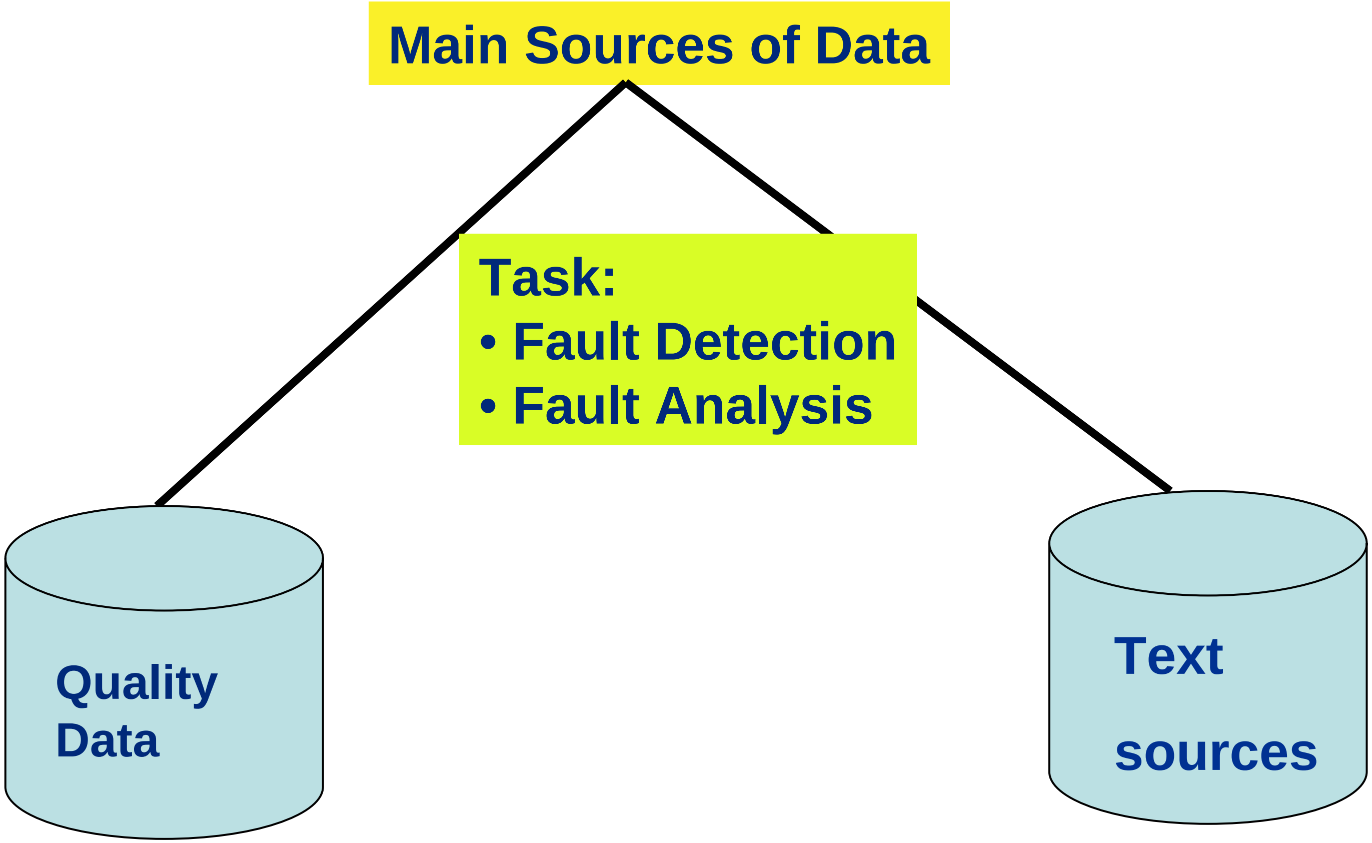
cand. inform. Aljoscha Klose
Matr.-Nr. 2 381 431

Technische Universität Braunschweig
Institut für Betriebssysteme und Rechnerverbund

Aufgabenstellung und Betreuung:
Prof. Dr. R. Kruse

A KDD-approach for prediction and early detection of failure rates in the automobile industry

Data Mining in Qualty Management (1)



Main Sources of Data

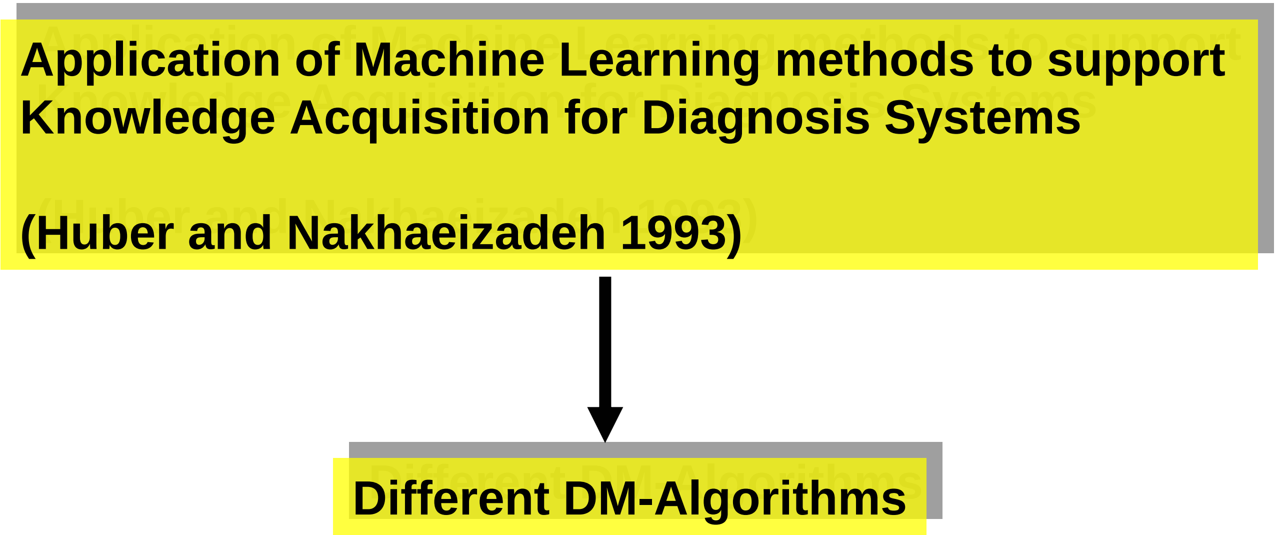
Task:
- Fault Detection
- Fault Analysis

Quality Data

Text sources

Data Mining in Qualty Management (2)

Application in Diagnostics



Application of Machine Learning methods to support Knowledge Acquisition for Diagnosis Systems

(Huber and Nakhaeizadeh 1993)

Different DM-Algorithms

Data Mining in Qualty Management (3)

- So far the DaimlerChrysler engineers responsible for the worldwide testing program for 60 F-Cell vehicles have collected a lot of information
- In fact, around one terabyte of data is currently stored on the server they use for work related to the project
- This huge amount of data has been collected since testing began more than one year ago - and it continues to grow gigabyte by gigabyte every day the customer-operated vehicles are on the road.

- The data log in the F-Cell is truly a black box
- The device, which is mounted behind the COMAND system in the center console, saves some 60 parameter values several times per second.
- If something unusual happens to the powertrain during a trip, the device will begin to store up to 600 parameter values at the same speed.

Source: www.daimlerchrysler.com