Compact Course in Data Mining

Applications

Professor Dr. Gholamreza Nakhaeizadeh

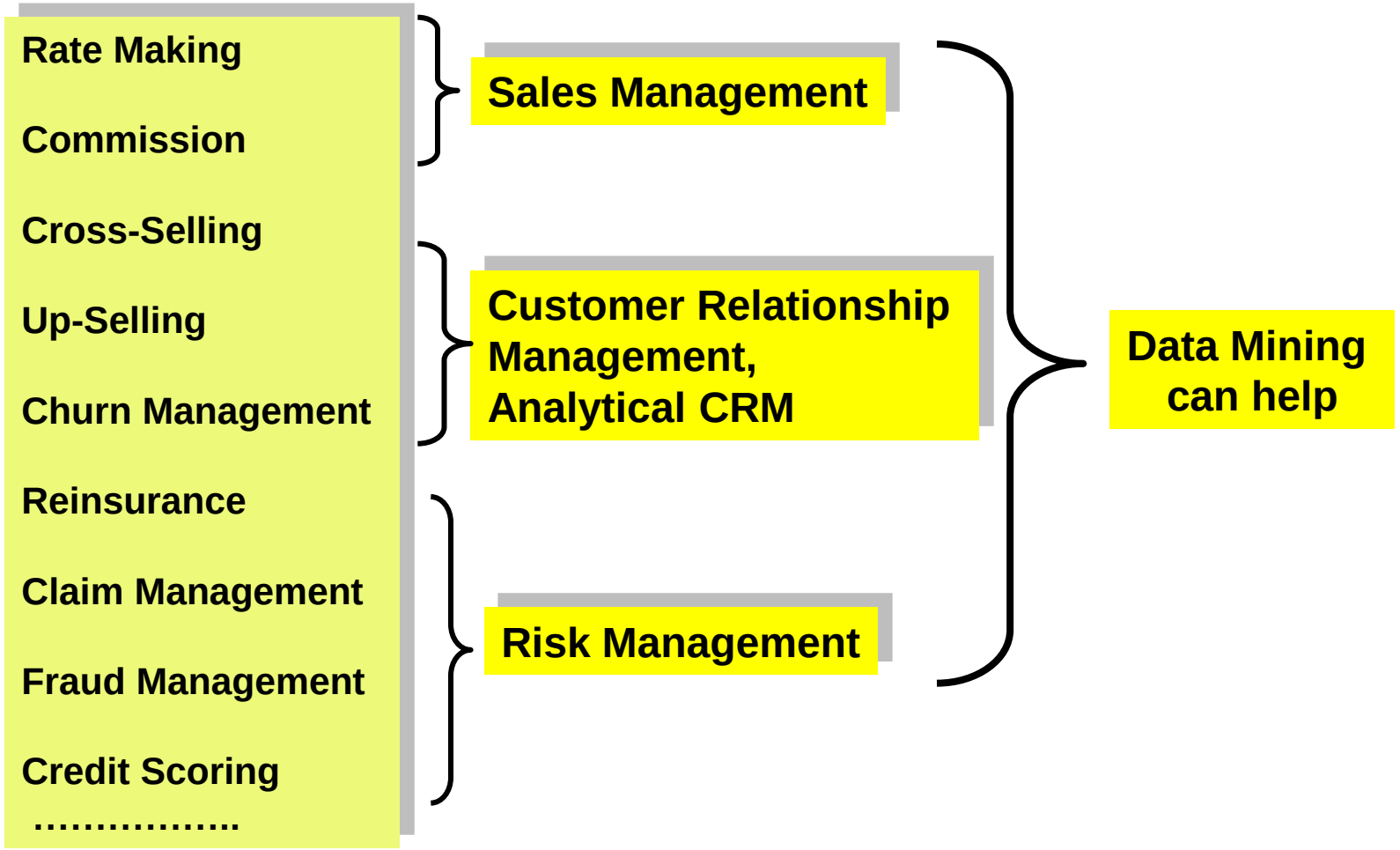
Content

General Aspects

- Application of Data Mining in:
  - Financial risk management
  - Financial forecasting
  - Fraud detection
  - Customer Relationship management
  - Money Laundering
  - …….
- Case Studies

Working with Data Mining Tool

Data Mining Applications

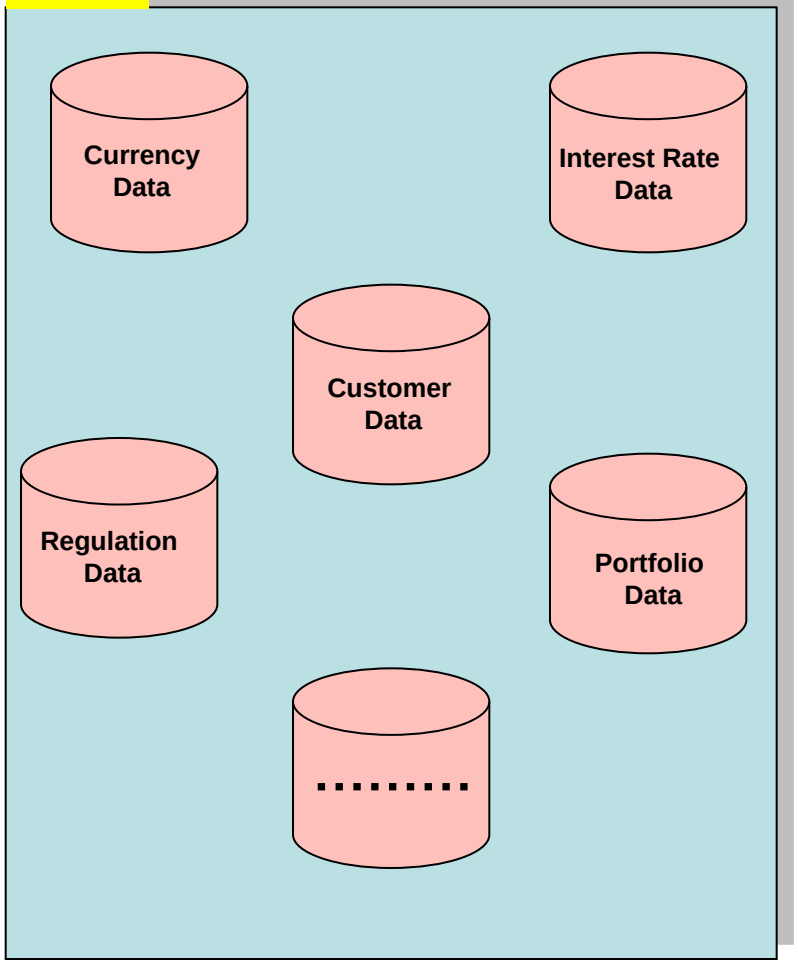


Why Data Mining in Banking ?

Business Issues

- Credit Risk
- Market Risk
- Controlling
- Trading
- Portfolio Manag.
- Investm. Manag.
- CRM
- Regulations& Compliance
- ....

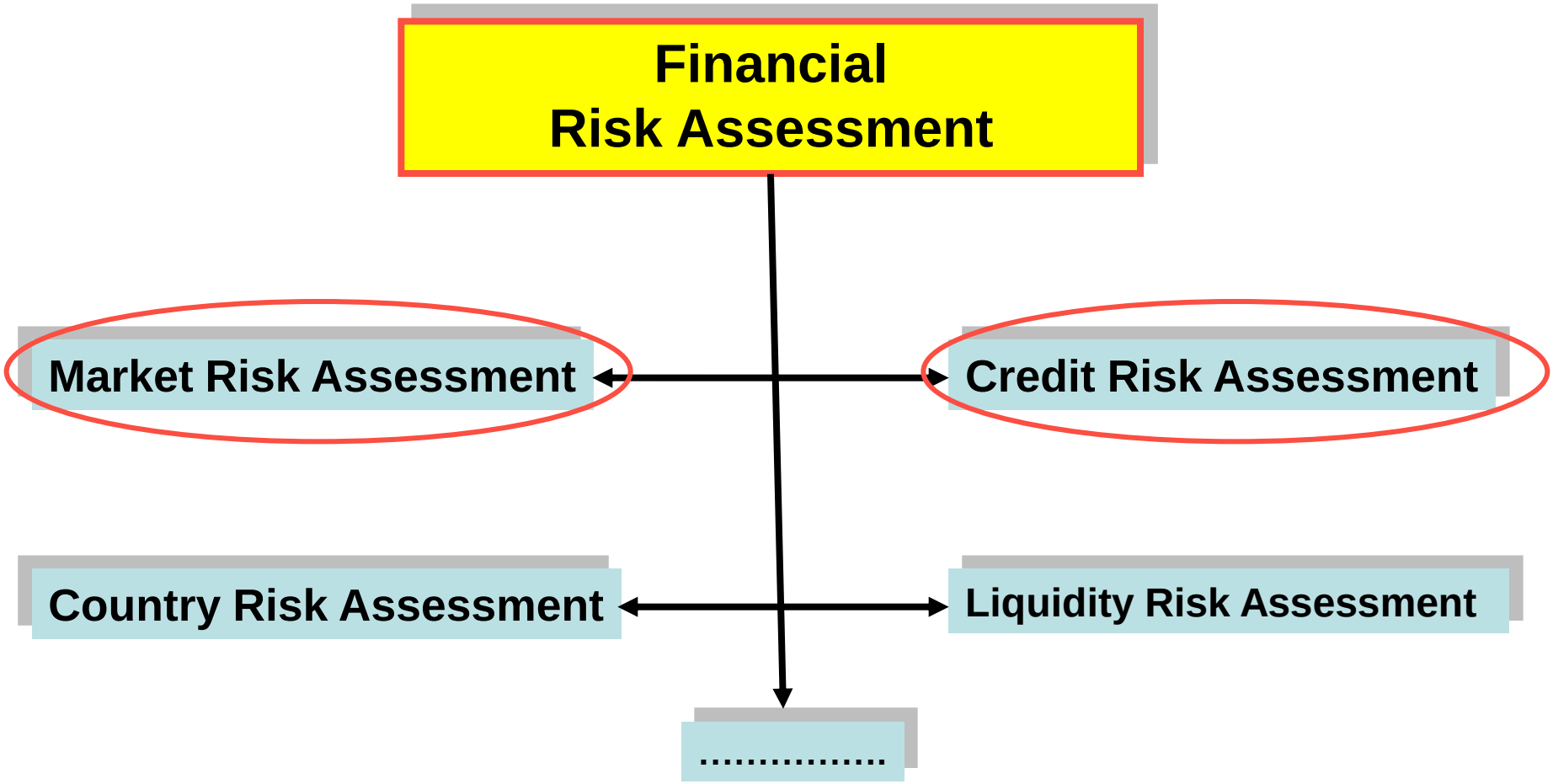
Data



DM- Applications

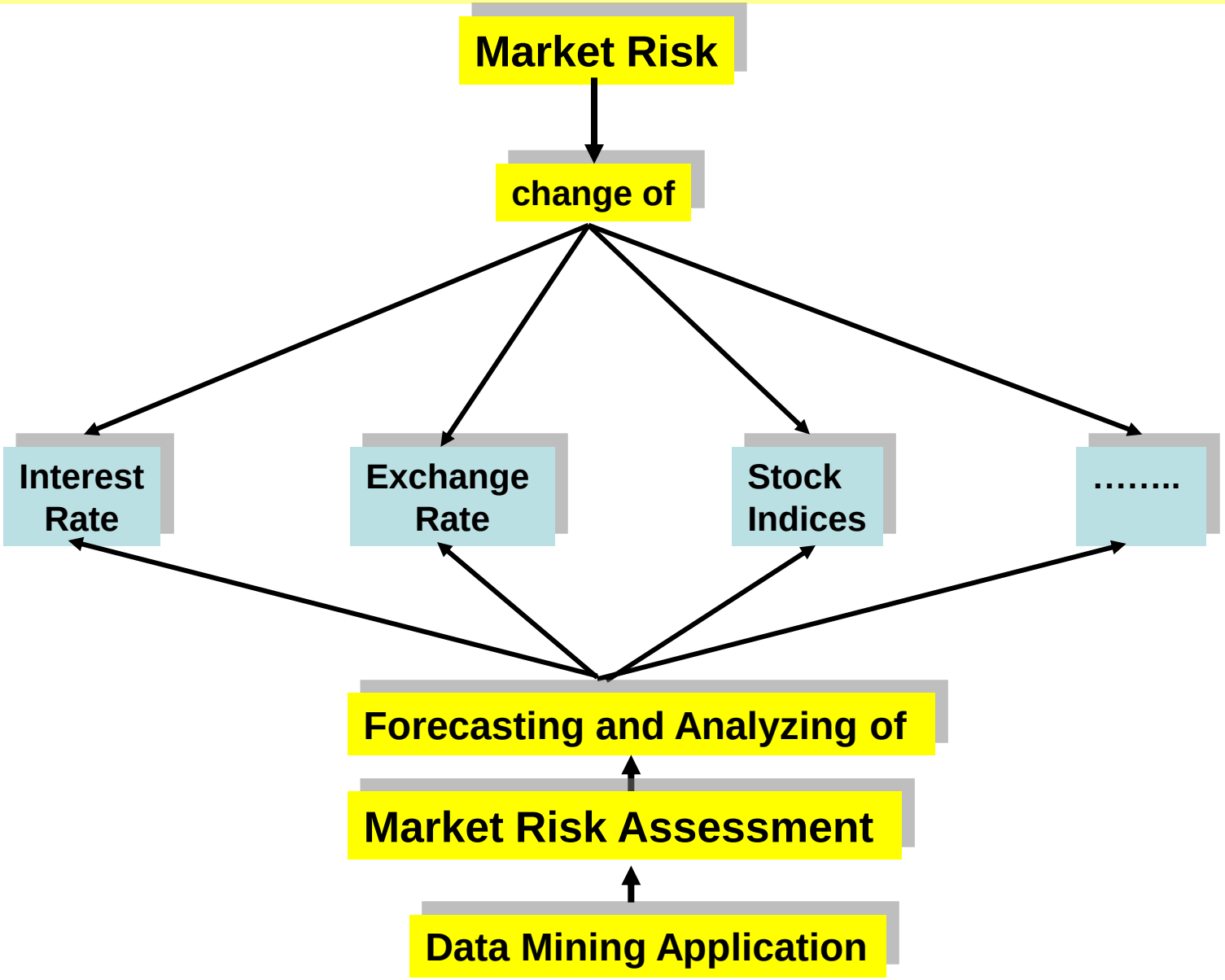
- Credit Scoring
- Market Forecasting
- Cross-Selling
- Up-Selling
- Churn Management
- Fraud Detection
- ....

Business Issues: Financial Risk Assessment

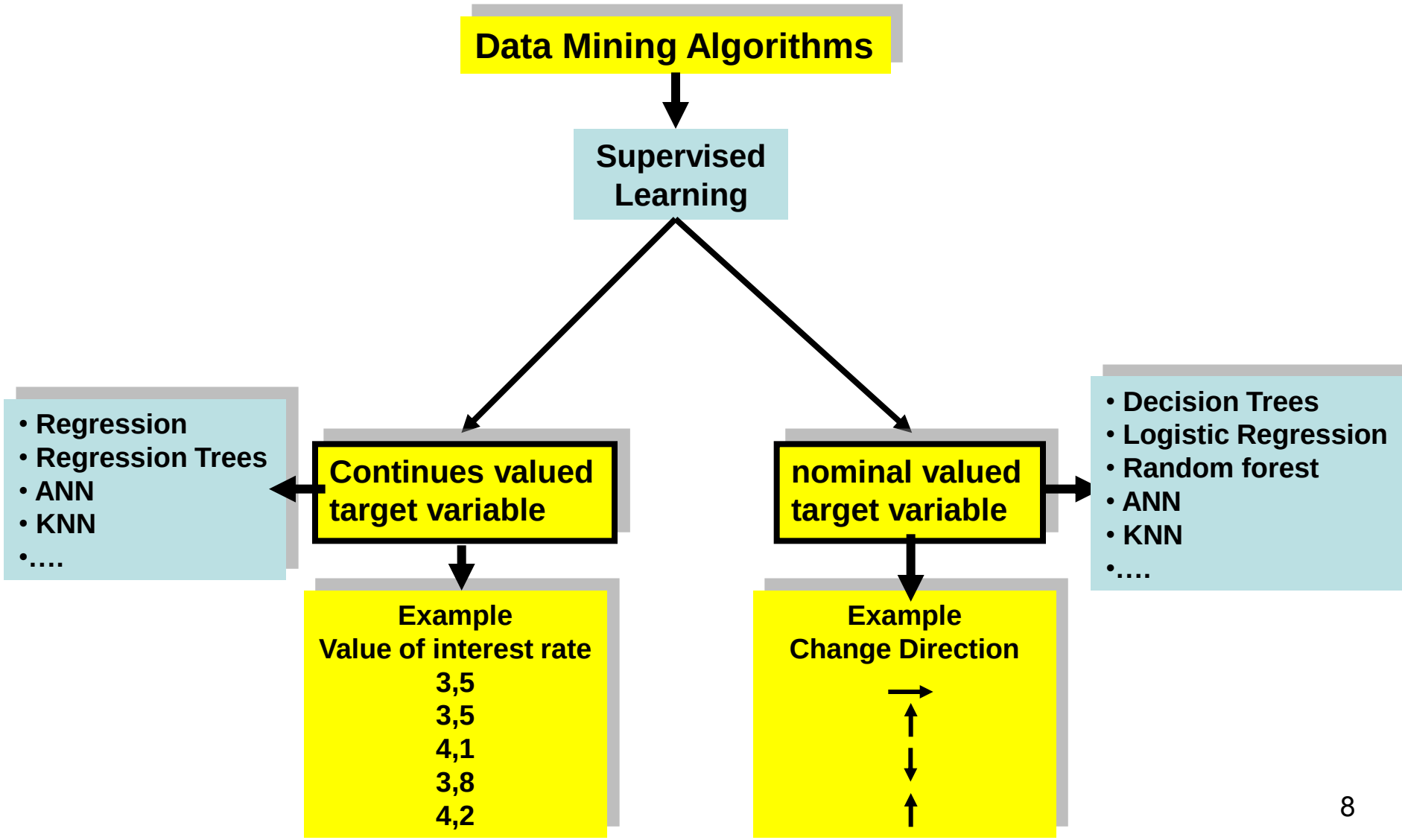


Application of Data Mining in Market Risk Management

Business Issues



Application of Data Mining in Market Risk



Market Risk, Practical Work: Forecasting of Interest Rate

Business Goal: Building a Prediction and Concept Description System

RapidddMiner: InterestRate

Load from Datasets: InterestRate_Train and Interestrate_Test (comparison between regression and ANN)

Y= GDP
CO= Total Personal Consumption
I= Total Gross Private Investment
G= Government Purchases of Goods and Services
R= Interest rate
YD= Disposal Income

Market Risk, Demo: Forecasting of „Deutsche Aktien Index“ (DAX)

Business Goal: Building a Prediction and Concept Description System

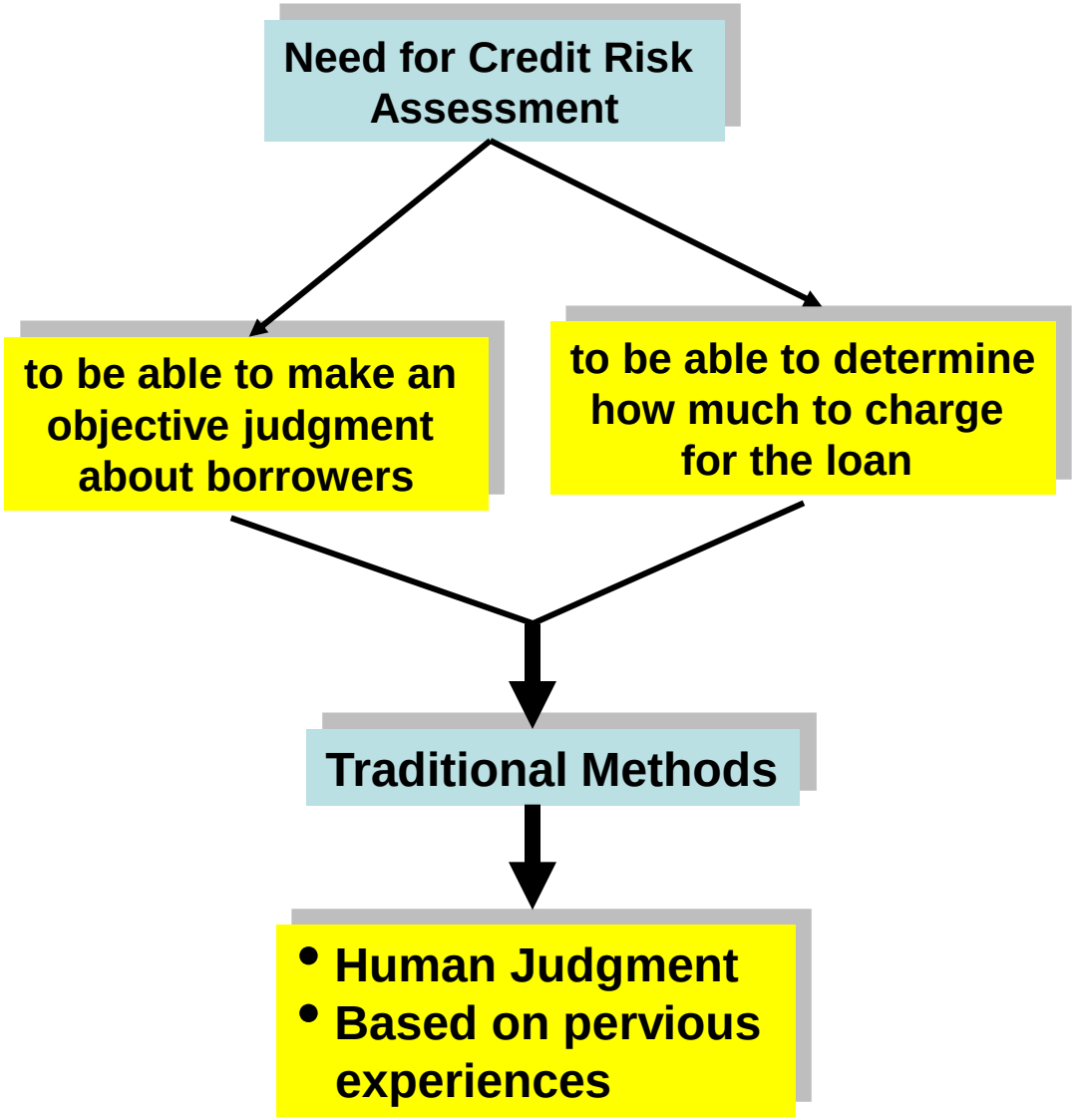
RapidddMiner: GermanStocks (comparison between regression and ANN)

Use: Split !

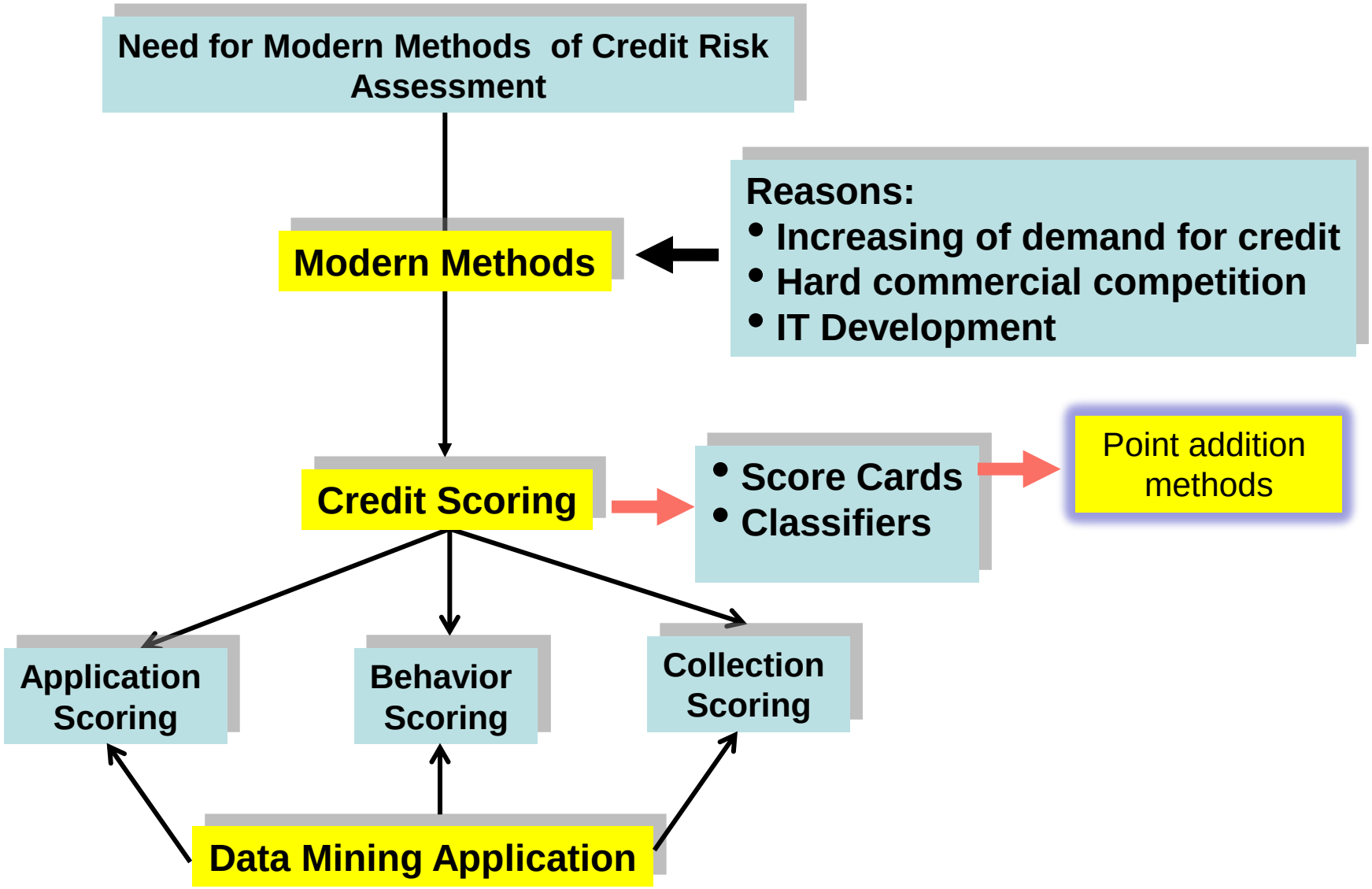
	Description of Variables
	Interest Rate
bmw	**BMW-Stock Price**
mru	**Münchner Rückv.-Stock Price**
rwe	**RWE-Stock Price**
vow	**VW-Stock Price**
kar	**Karstadt-Stock Price**
sie	**Siemens-Stock Price**
bas	**BASF-Stock Price**
index	**Index of Dax**
time	**Number of the days**

Application of Data Mining in Credit Risk Management

Business Issue: Credit Risk Assesment

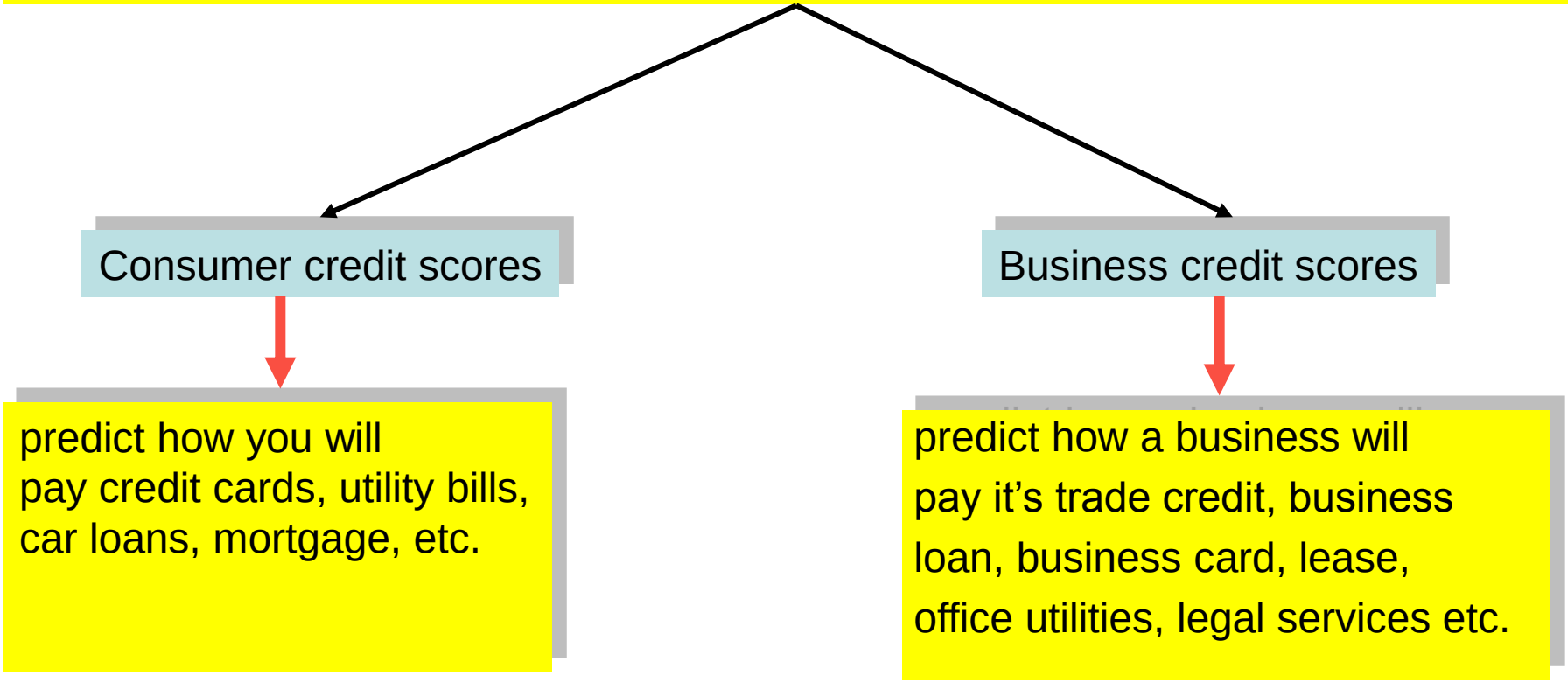


Business Issue: Credit Risk Assesment



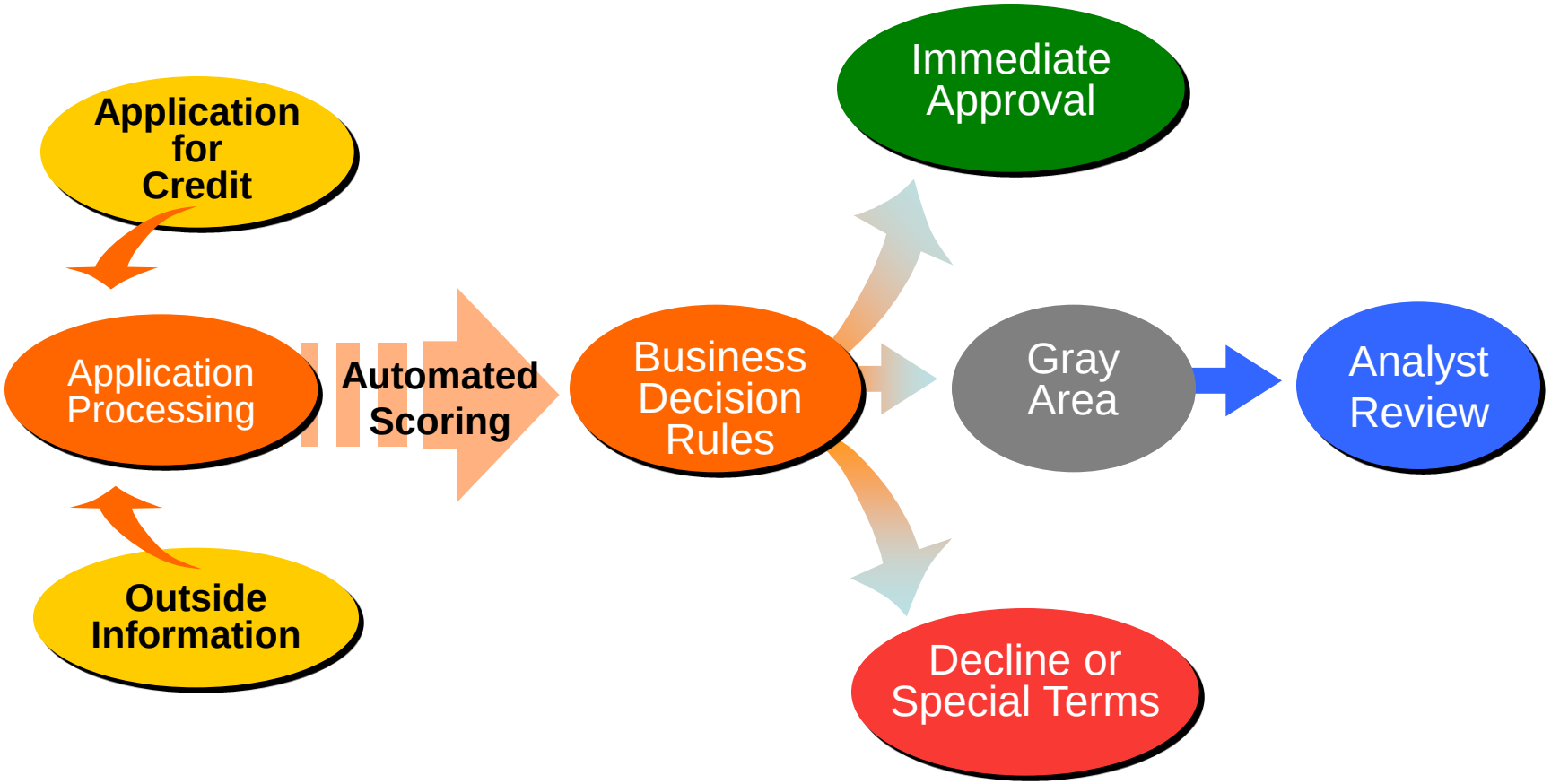
Business Issue: Credit Scoring

Credit scoring: involves all methods and techniques to predict the Creditworthiness of individual consumers or business companies



Source: ***www.biia.com/library/FCFPForumOCt2006.ppt***

Credit Risk Assesment: Credit approval process



Source: *www.biia.com/library/FCFPForumOCt2006.ppt*

Credit Scoring

Risk is one of the factor of the credit approval

- The main aim of credit business is normally profit maximizing
- Not always a bad risk is unprofitable and vice versa

Examples:

- Low risk borrowers paying on time are not profitable
  (Interest can no be charged for delay)
- High risk applicants can be profitable
  (Interest can be charged for delay)
- Credits may be profitable even they default

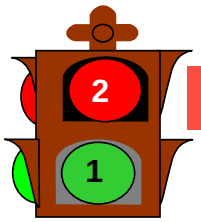
Profitability of a credit depends on several factors:
- Interest rate
- Collection cost
- …

Credit Risk Assesment

Business Issue: Credit Scoring

Number of classes

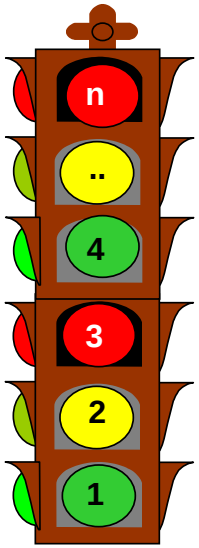
two classes



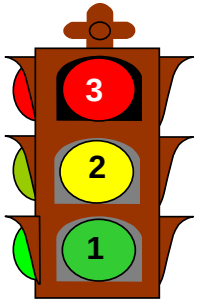
Bad Risk

Good Risk

more classes



three classes



High Risk

Medium Risk

Low Risk

Variables used in credit scoring model (Business Credit Scoring)

Financial

- Receivables
- Payables
- Cash
- Dividends
- Current Liabilities
- Current Assets
- Working Capital
- Net Worth

Credit

- Payment History
- Trade Experiences
- Bank Loans
- Secured Financing
- Public Filings
- Previous Bankruptcy
- Trends
- Condition Assessment

Business Demographics

- Line of Business
- Size (employee, sales)
- Years in Business
- Business History
- Suits, Liens, Judgments
- HQ/Branch/Single Loc.
- Location
- Special Events

Source: *www.biia.com/library/FCFPForumOCt2006.ppt*

Variables used in credit scoring model (Consumer Credit Scoring)

- Own / Rent
- Years at address
- Occupation
- Years on job
- Dept St / Major CC
- Bank reference
- Debt ratio
- No. of recent inquiries
- Years in file
- # Rev trades outstanding
- % Credit line utilization
- ……

The role of Credit Scoring in Insurance

Why Insurers Use Credit Information in Insurance Underwriting ?

1. There is a ***strong correlation*** between ***credit standing*** and ***loss ratios***

2. There is a ***distinct*** and ***consistent decline*** in relative ***loss ratios*** (which are a function of both claim frequency and cost) as ***credit*** standing ***improves***.

3. The relationship between credit standing and relative loss ratios is statistically ***irrefutable***.

Source: http://www.iii.org/media/presentations/credit/

Why Insurers Use Credit Information in Insurance Underwriting ? (continues)

Personal Responsibility

- Responsibility is a personality trait that carries over into many aspects of a person's life
- It is intuitive and reasonable to believe that the responsibility required to prudently manage one's finances is associated with other types of responsible and prudent behaviors, for example:
  - Proper maintenance of homes and automobiles
  - Safe operation of cars

Stability

- It is intuitive and reasonable to believe that financially stable individual are likely to exhibit stability in many other aspects of their lives.

- **Regarding the points mentioned, it is reasonable to use Credit Scores of Customers as additional information in insurance business**

Source: http://www.iii.org/media/presentations/credit/

Business Issue: Credit Risk Assesment

Credit bureaus in USA

Zahlen - Daten - Fakten

SCHUFA Holding AG

- Established: 1927
- Hauptsitz: Wiesbaden
- Staff 2007: 768

meineSCHUFA.de

Experian
A world of insight

EQUIFAX

INNOVIS®

TransUnion.

KREDİ KAYIT BÜROSU A.Ş.

Credit Bureau of Turkey

Credit Risk Assesment: Example

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

Application of:

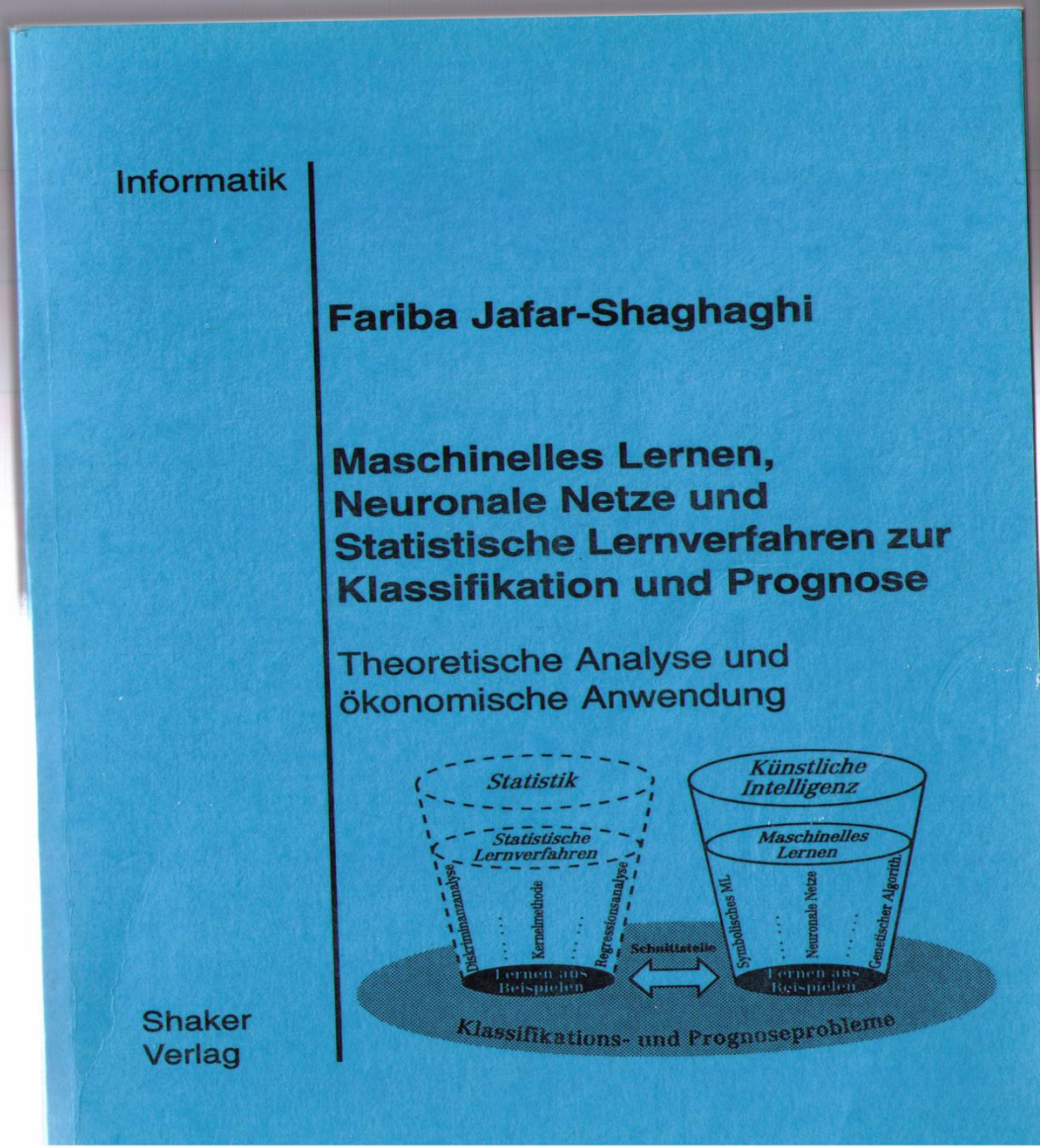
- About 40000 observations
- Statistical methods
- Decision Trees
- ANN
- Dynamic Data Mining

Credit Risk Assesment: Example

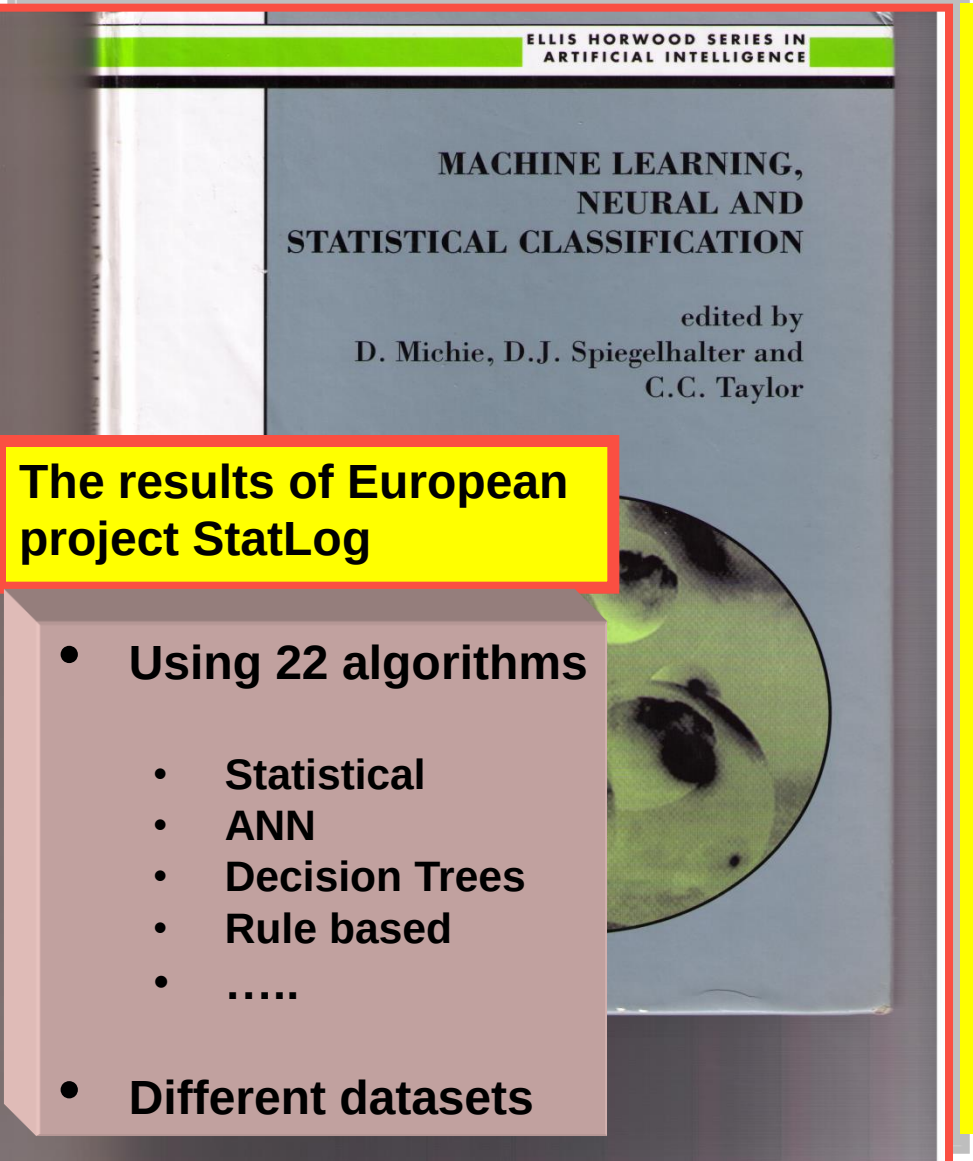
Machine learning, neural networks and statistical learning procedures for the classification and prediction

Application of:

- Different datasets from 101 to 8900 observations
- Statistical methods
- Decision Trees
- ANN
- K Nearest Neighbors



Credit Risk Assesment: Example



The results of European project StatLog

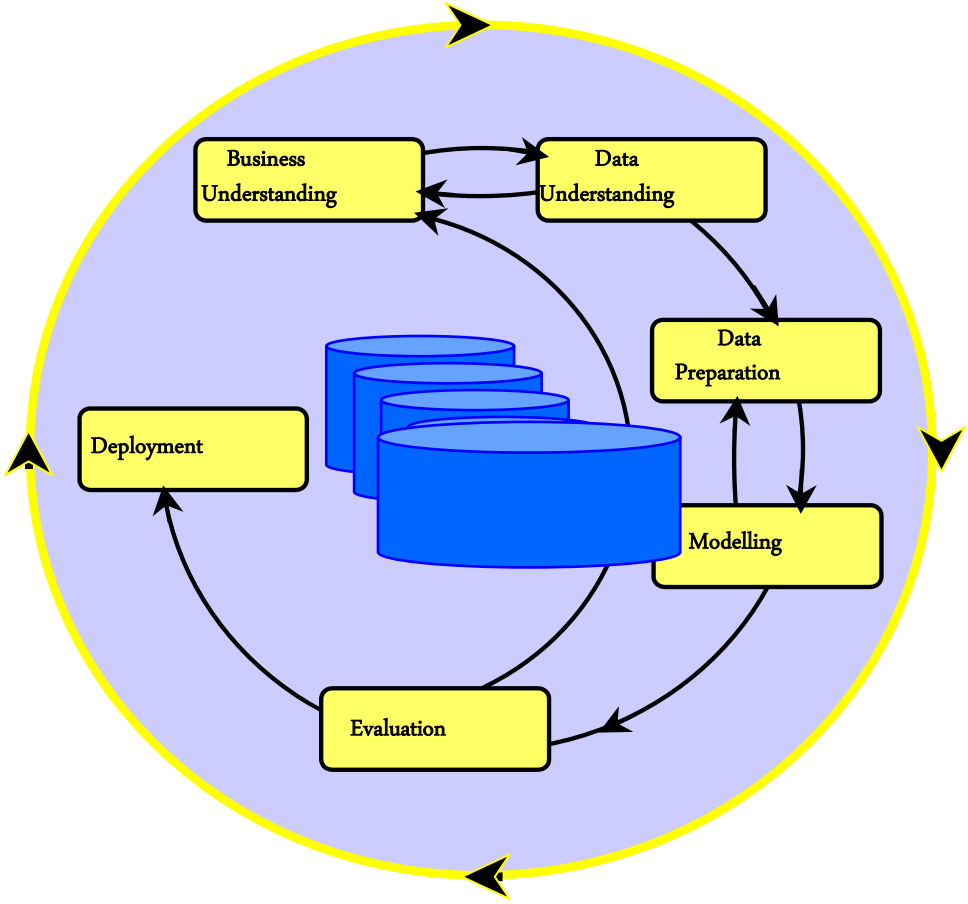
- Using 22 algorithms
  - Statistical
  - ANN
  - Decision Trees
  - Rule based
  - …..
- Different datasets

Error rates for Credit Datasets ordered by their average rank over the datasets.

credit	Cr.Aus	Cr.Man
CAL5	0.131	0.023
DIPOL92	0.141	0.020
Logdisc	0.141	0.030
SMART	0.158	0.020
C4.5	0.155	0.022
IndCART	0.152	0.025
Bprop	0.154	0.023
Discrim	0.141	0.033
RBF	0.145	0.031
Baytree	0.171	0.028
iTrule °	0.137	0.046
AC2	0.181	0.030
k-NN	0.181	0.031
Naivebay	0.151	0.043
CASTLE	0.148	0.047
ALLOC80	0.201	0.031
CART	0.145	
NewID	0.181	0.033
CN2	0.204	0.032
LVQ	0.197	0.040
Kobonen		0.043
Quadisc	0.207	0.050
Default	0.440	0.050

Data Mining Process

CRISP-DM



Own Experience

Credit Risk Assesment: Example, collection scoring

Optimization of collection efforts in automobile financing
A KDD supported Environment

F. Artiles1, H. Jeromin[1], H. Kauderer[2], G. Nakhaeizadeh[2]
[1]debis Financial Services, Credit Risk Management, Berlin
[2]DaimlerChrysler AG, Research and Technology 3, Ulm
{harald.kauderer, rheza.nakhaeizadeh}@daimlerchrysler.com

Summary

This contribution describes a project in the domain of risk management at the financial services subsidiary of DaimlerChrysler. The project is based on an innovative KDD approach to optimize the collection efforts in vehicle financing business. In contrary to the conventional "credit scoring" – which is very often associated with checking customer's creditability when he applies for a credit – in this project we follow the credit life cycle down the line and focus the problem of dealing with delinquent customers in a current credit portfolio. The developed Data Mining solution is implemented in the end-user workflow environment and in daily operation.

Credit Risk Assesment: Example, collection scoring

Business understanding

Decreasing of the profit margins in vehicle financing business was observed

↓

more effort in developing appropriate "collection" strategies was necessary

↓

"collections" cover all the instruments and activities carried out to deal with delinquent customers

Credit Risk Assesment: Example, collection scoring

Business understanding

Business objective

to achieve a risk-adjusted allocation of human resources in a phone collections department

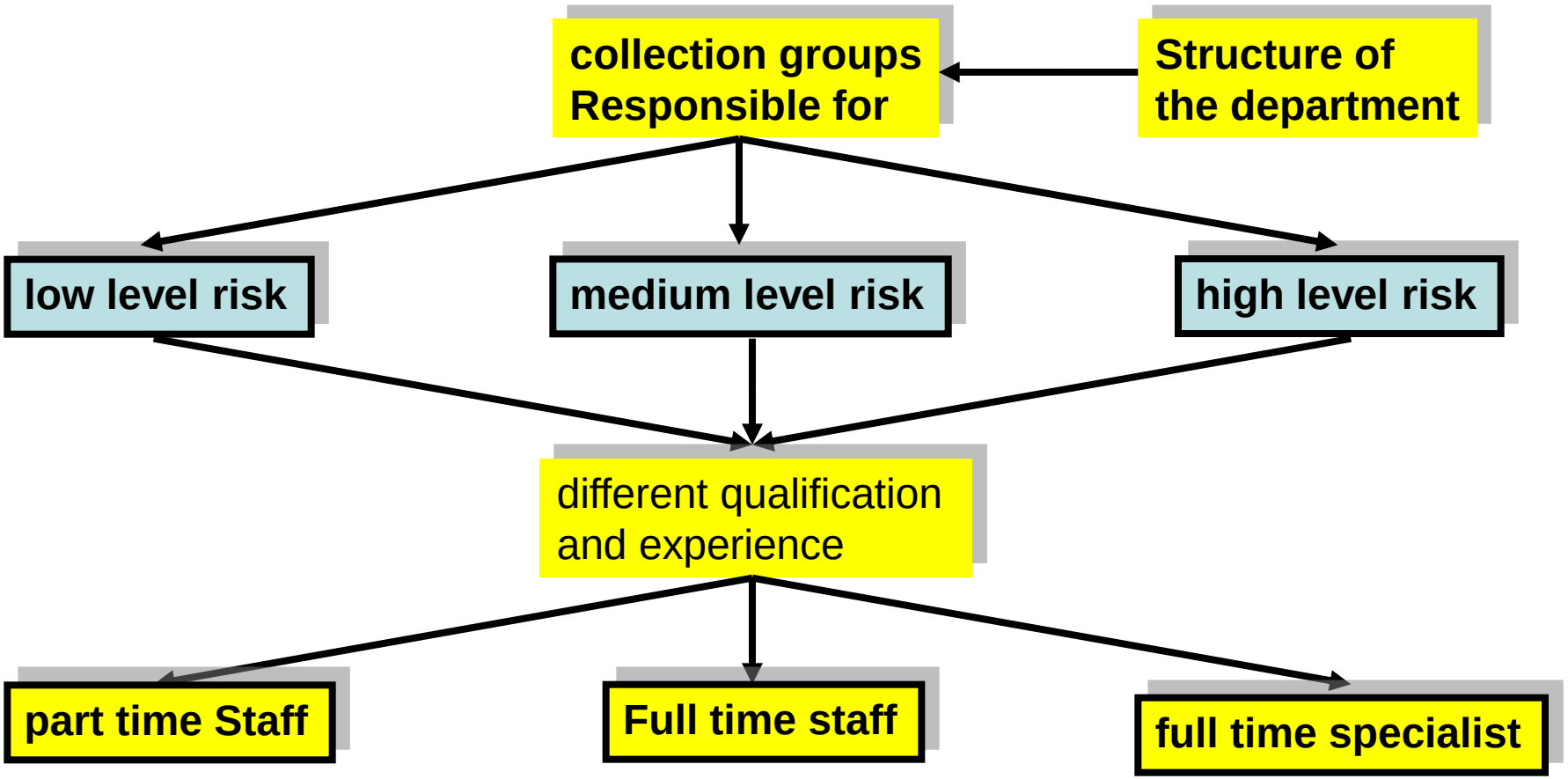
Strategies to "collect the money"

- sending a letter
- doing phone calls
- repossession of the vehicles

The selection of the "right" strategy depends on the individual risk position of each customer

Credit Risk Assesment: Example, collection scoring

Business understanding



Credit Risk Assesment: Example, collection scoring

Data Mining Task

Estimation of the individual risks of today's delinquent customers

nominal target variable classification task

Data understanding

Source of the data

the accounting system at the end-user side covers customer information:

- on the contract (amount financed, monthly payment, term, etc.)
- on the vehicle (make, type, model, etc.)
- on the current and past state of delinquencies

Data quality

- extremely high
- no missing values
- very few outliers

Credit Risk Assesment: Example, collection scoring

Data Preparation

Definition of the target variable

Definition was originally used:

assign each customer based on its *current* state of "days delinquent" to the different collection groups

If	**0 <days_delinquent_today <= 30**	**then**	**Level_1_collections (low risk)**
else	**if 30 < days_delinquent_today <= 60**	**then**	**Level_2_collections (medium risk)**
Else	**60 < days_delinquent_today**	**then**	**Level_3_collections (high risk)**

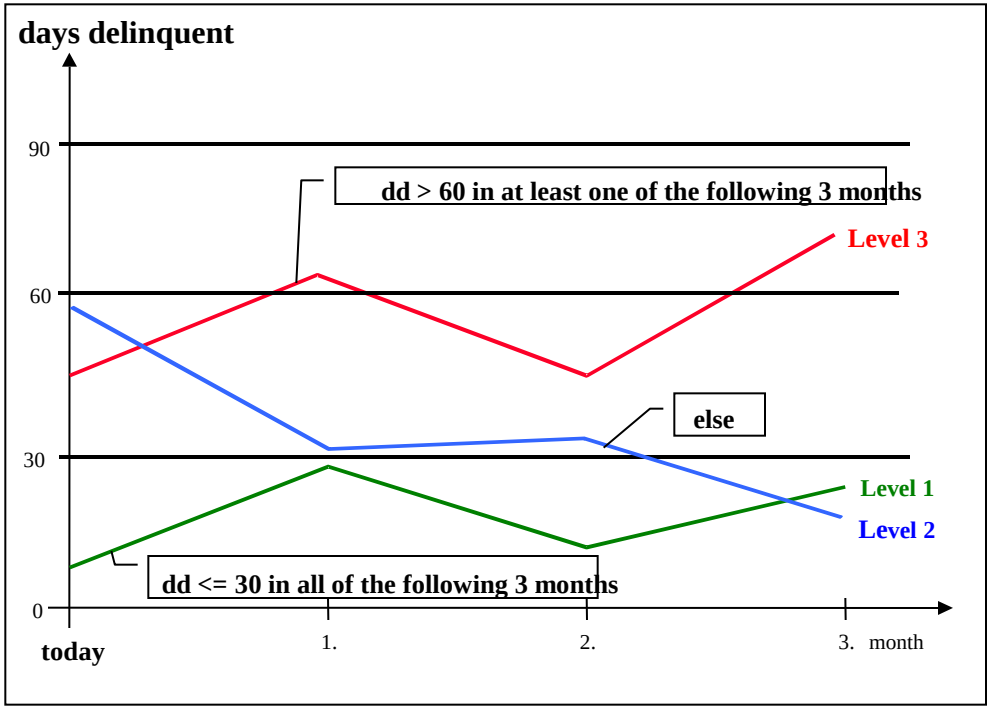
customer's payment ability and behavior is perfectly correlated with his *current* state of delinquency

assumption

sub-optimal solution due to past experience

Credit Risk Assesment: Example, collection scoring

Data Preparation (continues)



New definition of the target variable

Consideration the future behavior of current delinquent customers in definition of the target variable

Level 1: Today's delinquent customer has to stay under 30 days delinquency threshold in all of the following three months.

Level 2: else.

Level 3: Today's delinquent customer violates the 60 days delinquency threshold at least in one of the following three months.

Credit Risk Assesment: Example, collection scoring

Model generation and selection

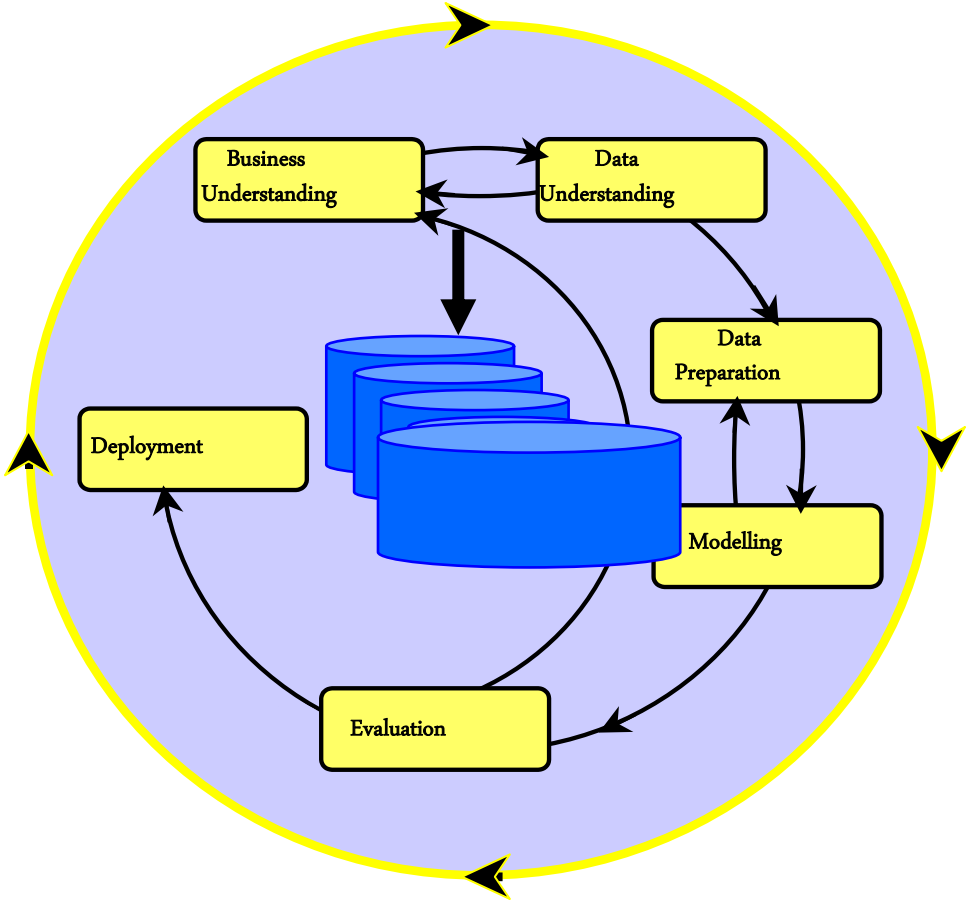
- Using the algorithms and C5.0 and M6.1 and ANN and by variation of the parameters of these algorithms several models were generated and evaluated

- As result of the evaluation process it was agreed to implement the regression tree model (M6.1) into the workflow of phone collections in the end-users department

Business Issue: Credit Risk Assesment

Data Mining Process

CRISP-DM



Credit Scoring Case Study

Business understanding

Business goal:
Profit maximizing in consumer credit business

→ **Quickly and accurately assess risk by identification of risky new credit applicants**

Importance:

Identification a bad risk credit applicant as good → **Can cause credit loss**

Identification a good risk credit applicant as bad → **Can have opportunity cost**

Credit Scoring, Business Understanding

Interaction with the end users (experts)

- **Organization of the meetings and workshops**
- **Assess situation**

The roll and responsibilities of the end users (experts):

- acquisition and classification of data necessary for Data Mining
- valuation of the merits of generated rules
- these responsibilities require the expert to do what an expert does best: to exercise his or her expertise
- This seems a more natural than the traditional knowledge acquisition process

Credit Scoring Case Study

Data understanding

Target Variable: Nominal

Historical and current labeled information about 1000 bank customers are available

Data Mining Task: Classification

Number of attributes: 20, most of them nominal

Example of attribute used:

- **Maturity of credit**
- **Volume of credit**
- **Gender**
- **Age**
- **Occupation**

…

Data Mining Algorithms
Artificial Neural Networks (ANN)
Decision Trees
…

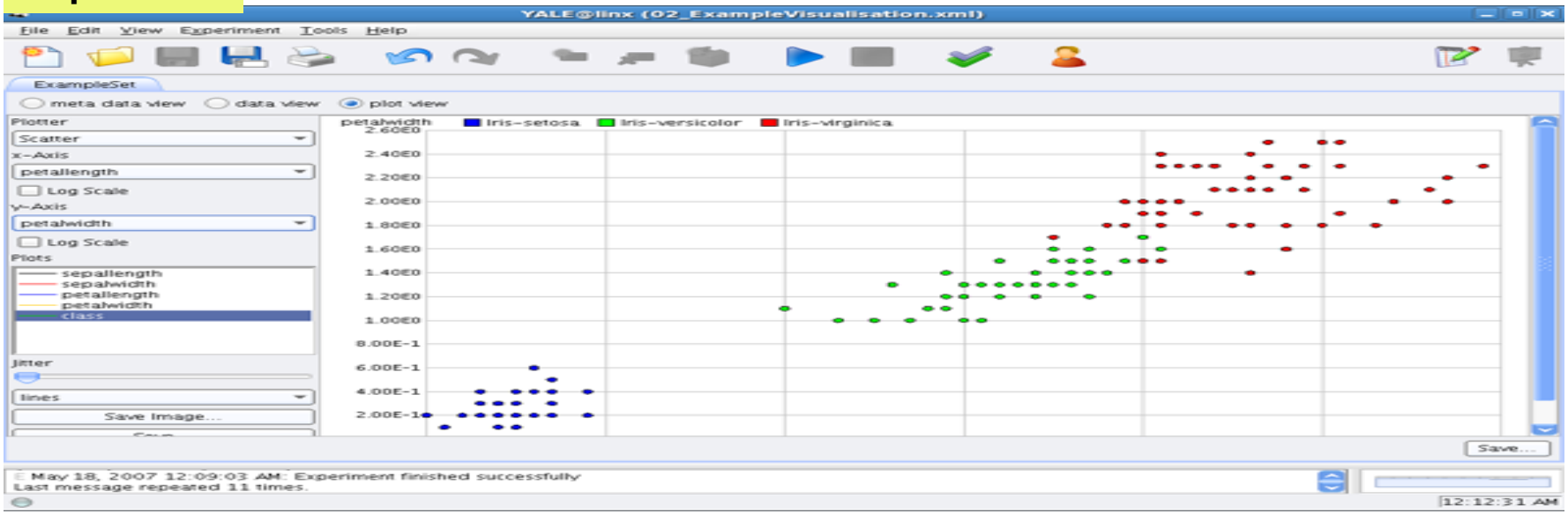
Understandability is important

ANN is no optimal algorithm

DT, Rule based

Credit Scoring Case Study

RapidMiner



Load German_CreditTr.aml

Comparison of different algorithms

German_CreditDT.xml

Load German_CreditANN.xml

Application of Data Mining in Customer Relationship Management

Business Issue: Customer Relationship Management (CRM) in Banking

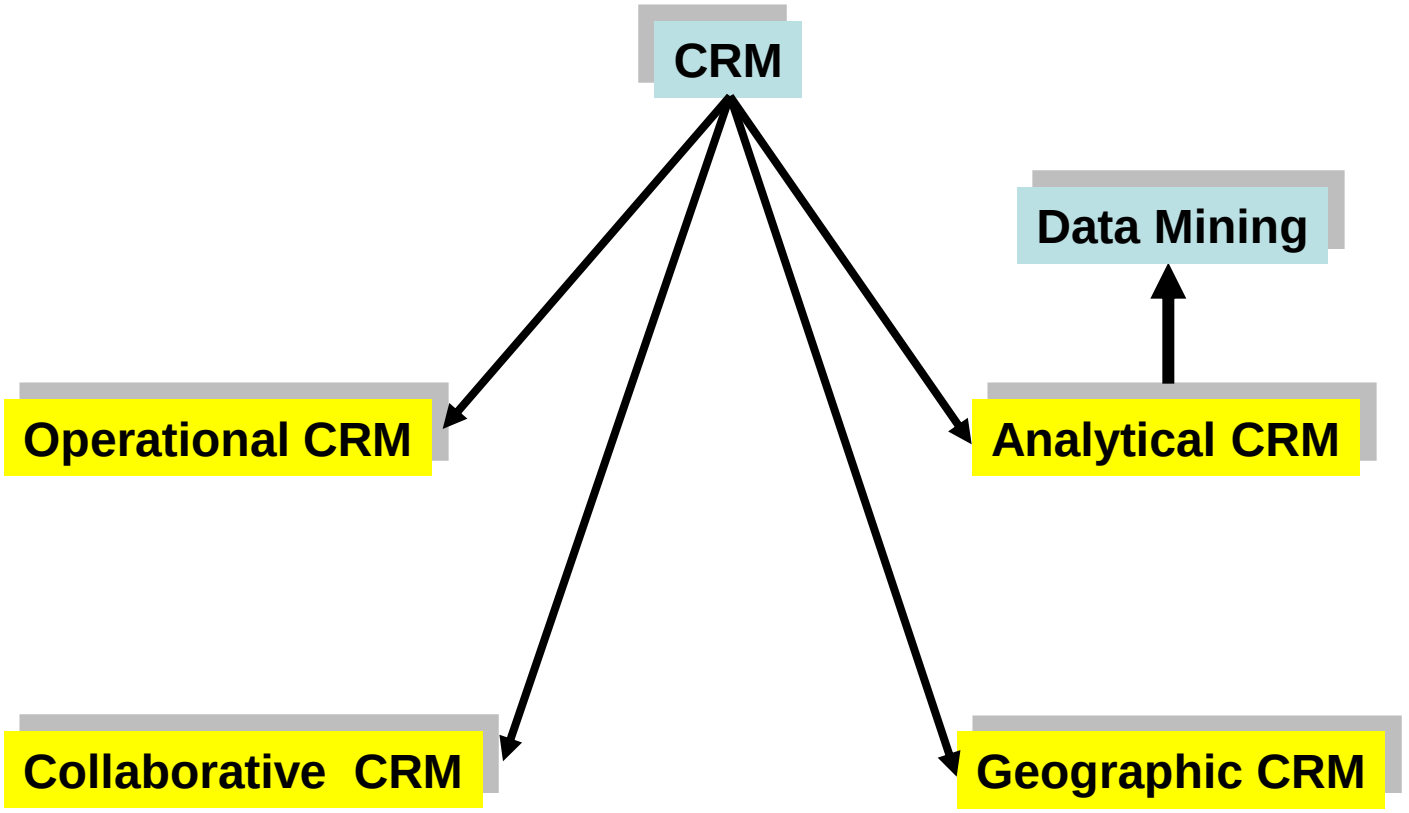
Definition: CRM consists of the processes a company uses to track and organize its contacts with its current and prospective customers

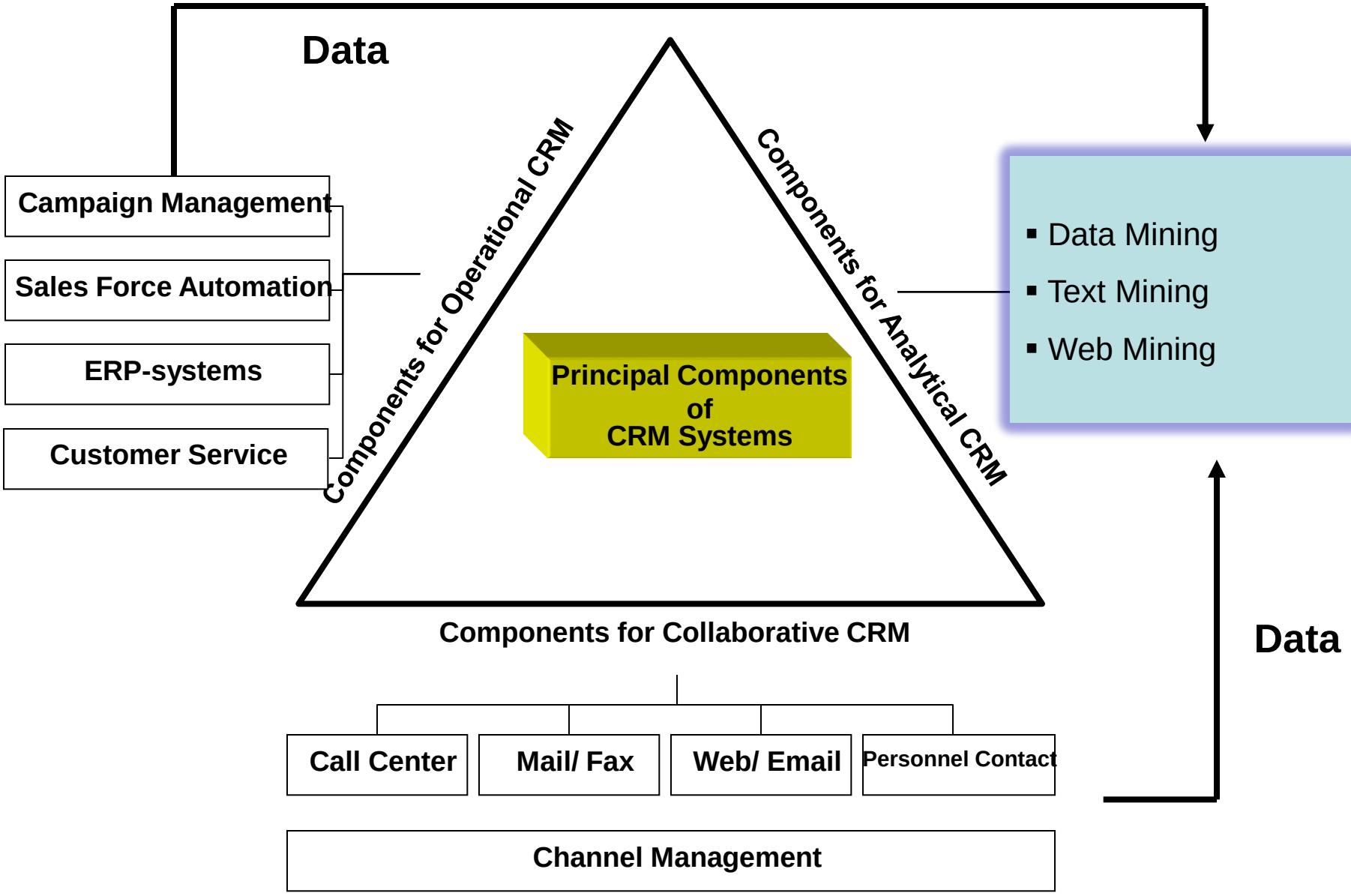
Source:http://en.wikipedia.org/wiki/Customer_relationship_management

CRM Goals

- Customer retention and brand loyalty
  (it is more difficult to gain a new customer than to keep one)
- Reduction of costs of operation
- Identifying potential customers
- Providing 360-degree view of the customer

Business Issue: CRM in Banking

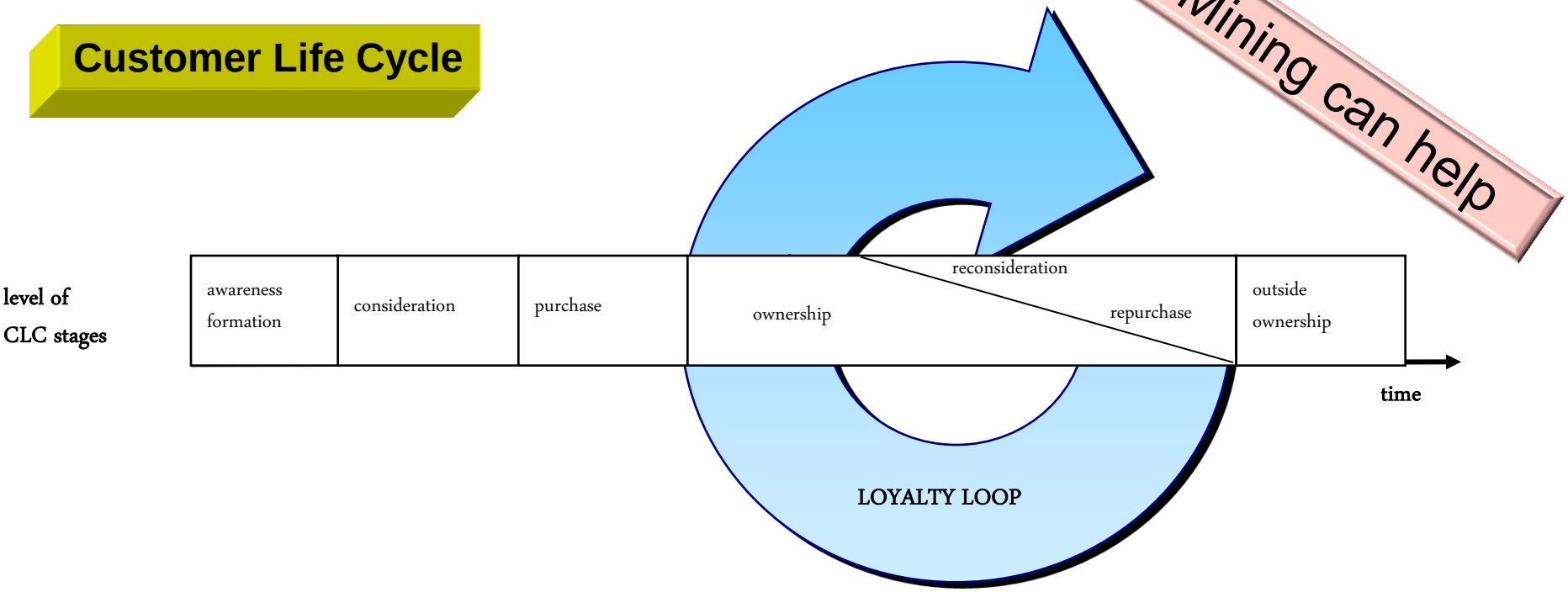




Analytical CRM

Data Mining can help

Customer Life Cycle



acquisition program | loyalty program | acquisition program

Typical aCRM tasks

- data collection
- predictive modeling
- Response analysis
-..

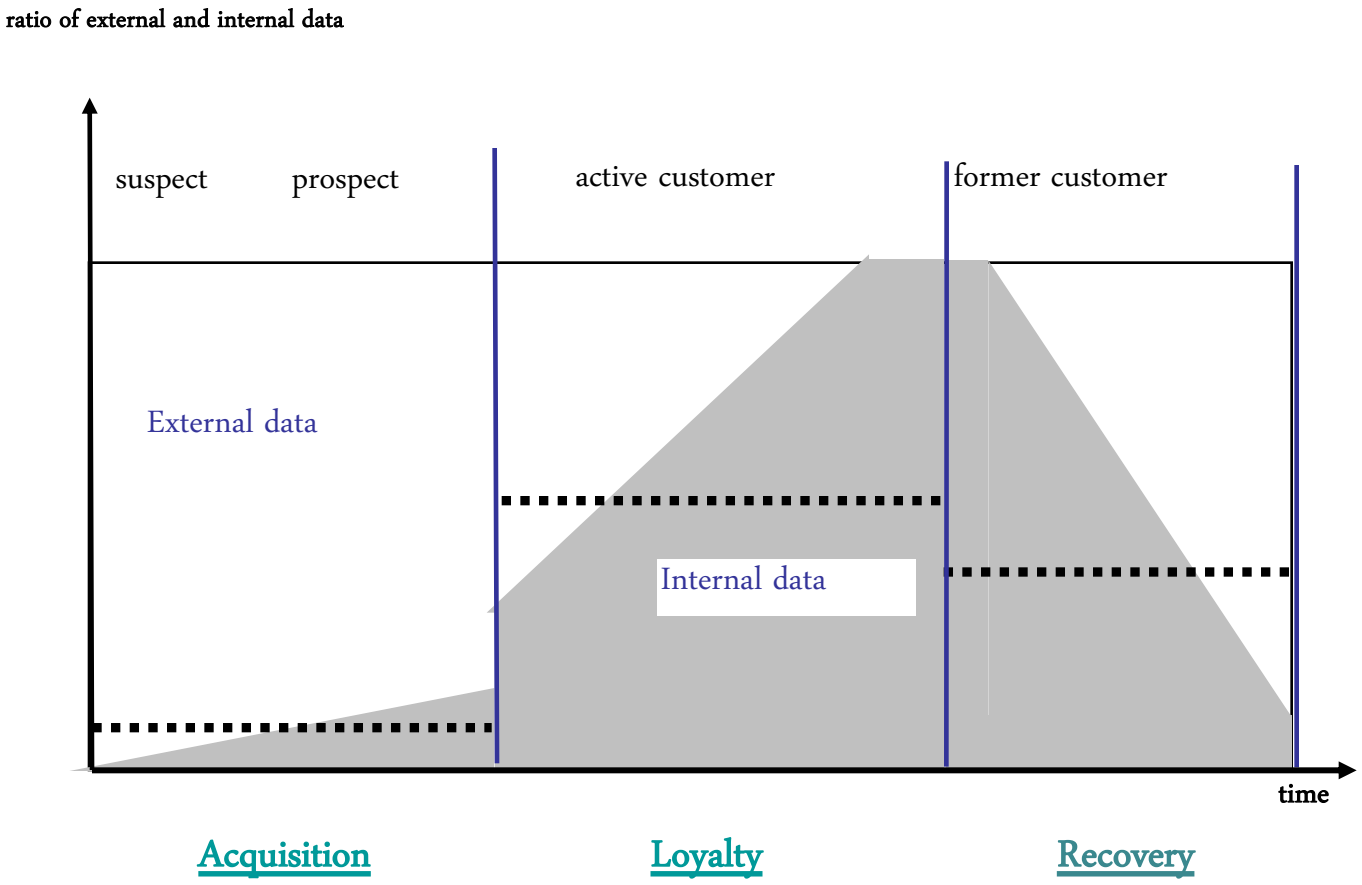
- customer segmentation
- cross/up-selling analysis
- customer value analysis
-..

- churn analysis
- modeling for recovery
- response analysis
-..

Source: http://www.prudsys.de/Service/Downloads/bin/DMC2003_arndt-daimlerchrysler.pdf

Analytical CRM: Data Situation

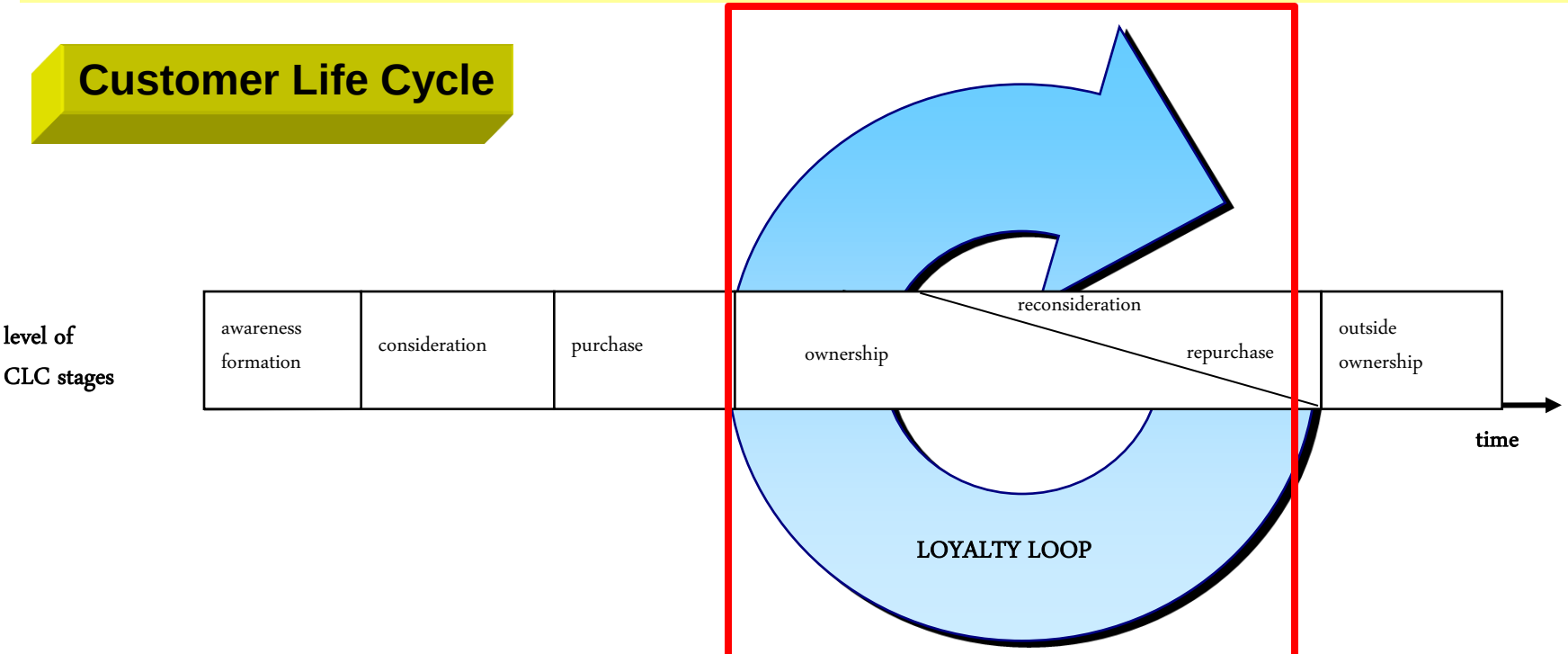
Data Situation Along the Customer Life Cycle



Source: http://www.prudsys.de/Service/Downloads/bin/DMC2003_arndt-daimlerchrysler.pdf

Analytical CRM, Cross-Selling, Up-Selling

Customer Life Cycle



acquisition program

loyalty program

acquisition program

Typical aCRM tasks

- data collection
- predictive modeling
- Response analysis
-..

- customer segmentation
- cross/up-selling analysis
- customer value analysis
-..

churn analysis
modeling for recovery
response analysis
..

Source: http://www.prudsys.de/Service/Downloads/bin/DMC2003_arndt-daimlerchrysler.pdf

Cross-Selling, Up-Selling

Cross-Selling

Selling an additional product or service to an existing customer

Up-Selling

selling something that is more profitable or otherwise preferable for the seller instead of, or in addition to the original sale

Increasing the company's success

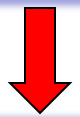
Source: http://en.wikipedia.org/wiki/Upselling

Cross-Selling, Up-Selling

Crucial questions in cross- and up-selling:

- Who should be contacted ?
- Which channel should be used ?
- Which product should be offered ?
- When ?

Analyzing of current customers behavior



- Which products use my clients at present and for how long?
- How customers consume can be increased?
- Which additional products should I provide customers?
- Which products or services are often purchased together?

Appropriate algorithms for Cross-Selling and Up-Selling

- Which additional products should I provide customers?

Product 1?

Product 2 ?

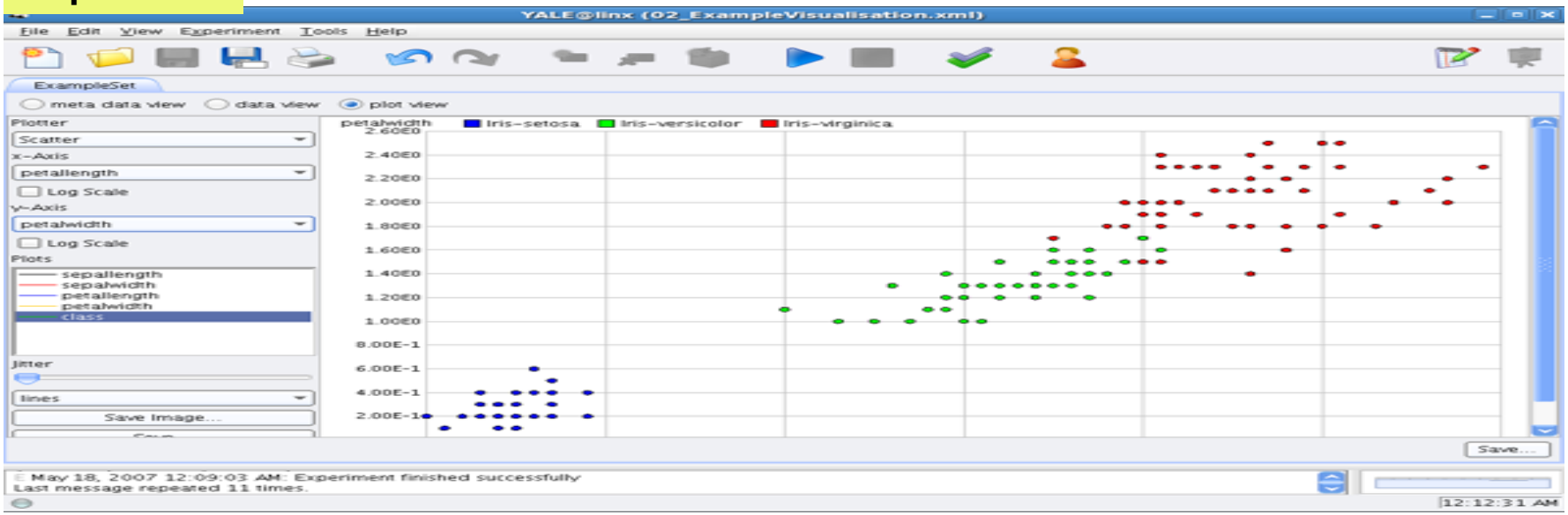
…….

Classification Task

- DT
- ANN
- KNN
- …..

Practical Work

RapidMiner



Utility→generate → upselling

1. Compare the performance of KNN with other algorithms using Xvalidation

Appropriate algorithms for Cross-Selling and Up-Selling

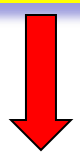
- Which products or services are often purchased together?

(Product 1, Product 2) → (product3)

(product 4) → (product2)

…….

Unsupervised Learning

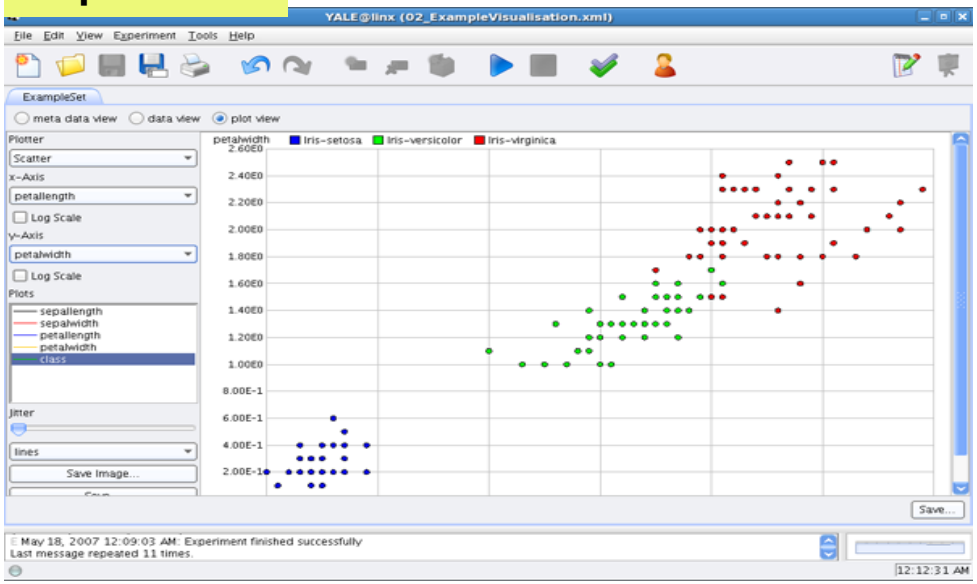


Association Mining

Application Examples: Data Mining in CRM Cross Selling

Working with DM-Tool Rapid Miner

RapidMiner

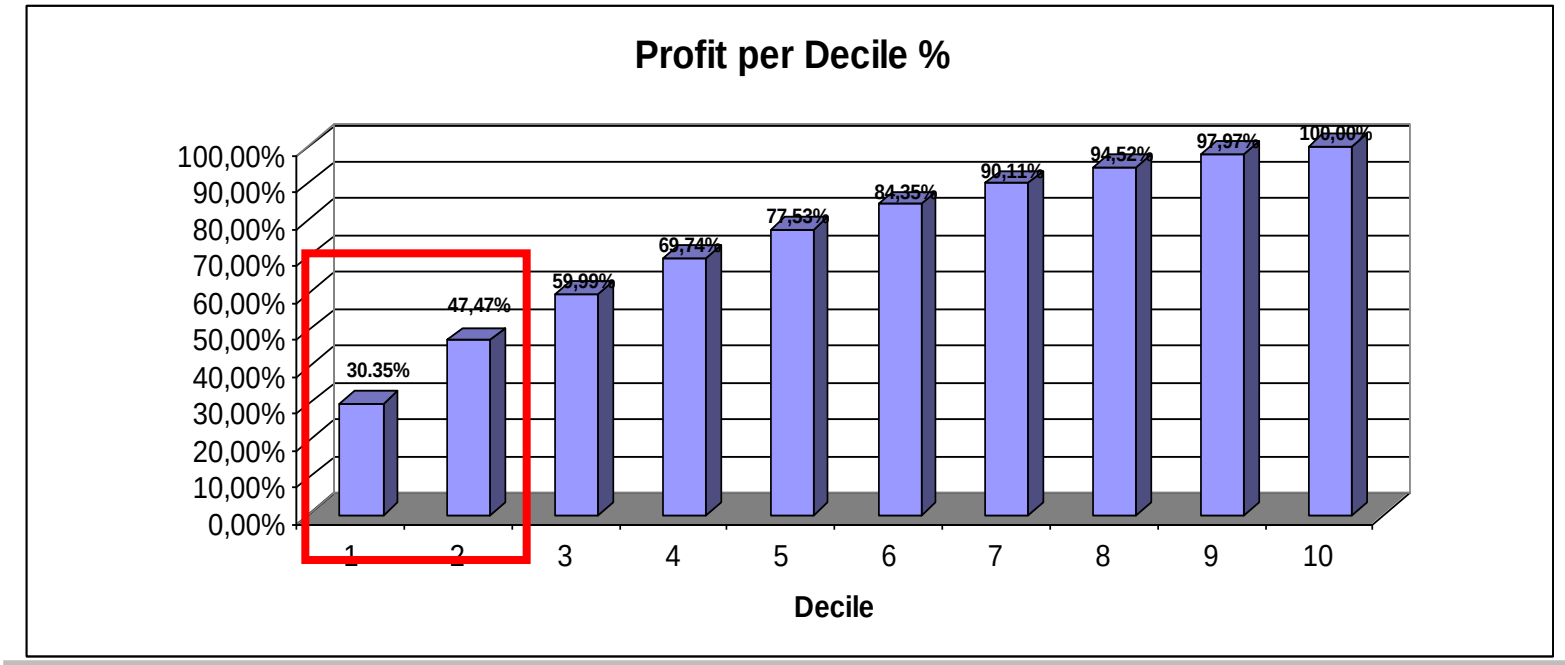


Open in Workspace: Bookstore.xml

Analytical CRM in Banking: Customer Value

Importance

Customer value analysis makes differentiation in individual level possible. Therefore, it is pre condition for profitable growth



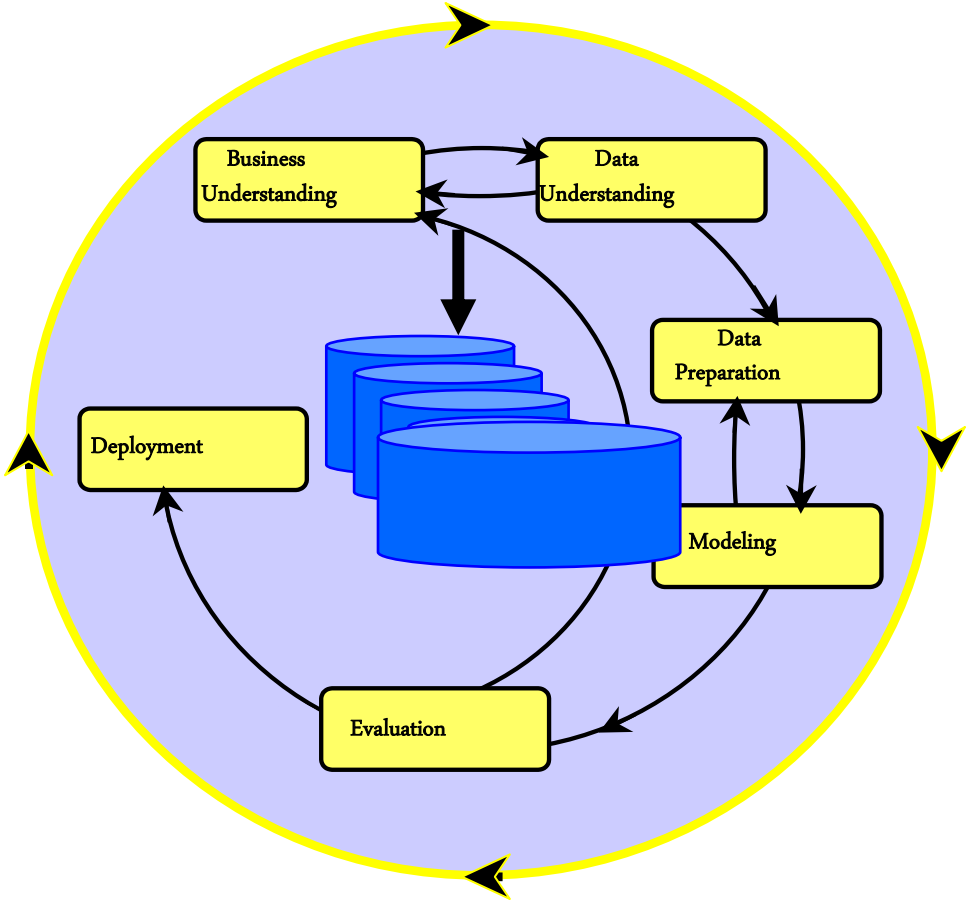
The 20% of customers generate 47.5 % of profit

(based on a database containing information on 1.6 Million private customers)

Source: Dr. Dirk Arndt "value oriented customer management", personal correspondence

Data Mining Process

CRISP-DM



Application Examples: Data Mining in CRM Churn Analysis

Customer Retention: Churn Analysis

Business Understanding

Business Problem

- Customer Churn is one of the important issue in highly competitive business

- Customer Churn describes the number or percent of the customers who cut their relationship with the organization

Churn Analysis, Business Understanding

Goals :

- **Identifying the customers who are at risk of leaving with a certain probability**
- **Determining whether the effort is worth to retain such customers**

- **How much is being lost because of customer churn ?**
- **What is the scale of the efforts that would be appropriate for retention campaign?**

Churn Analysis, Business Goals

Data Mining system

Construction a system that can

- help analyzing of the customers the churn's reasons
- as an early warning system, identify the customers who are likely to cut their relationship with the organization

Data Mining task

- Prediction or Classification (Supervised Leaning)
- Understandability of the results is important

Churn Analysis, Data Understanding

Which algorithms (s) could be used ?

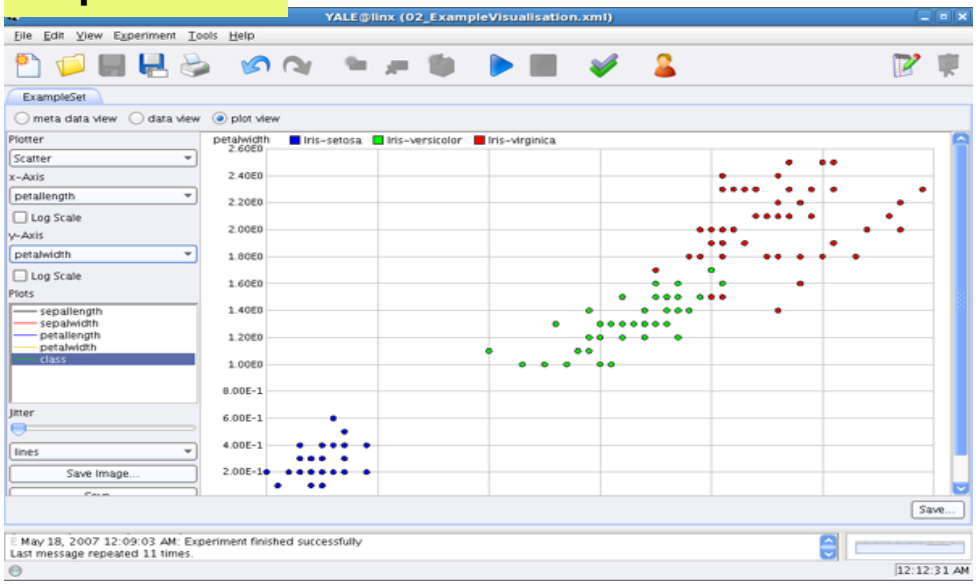
We need algorithms :

- **Which algorithm can handle nominal-valued classes ?**
- **Which algorithm can deliver understandable rules ?**
- **Which are the candidates ?**

- **Decision Trees**
- **Rule Based Methods**
- **……**

Working with DM-Tool Rapid Miner

RapidMiner



Open in Workspace: Churn.xml

Case Study - LG Telecom

Problem

Customer Churn Rate
⇒ **14%** (1999)

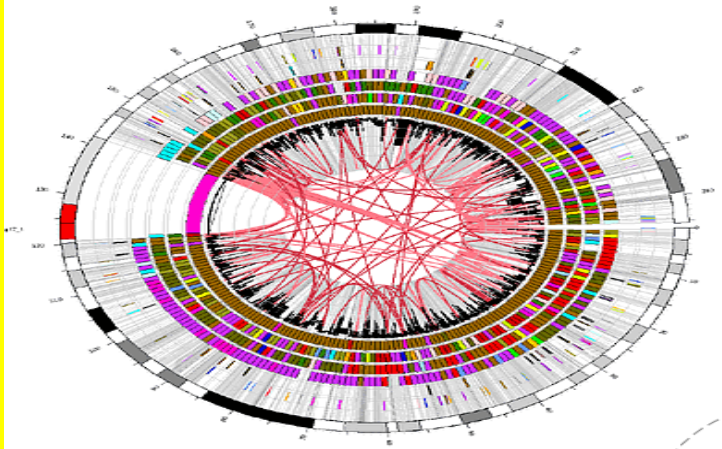
CRM

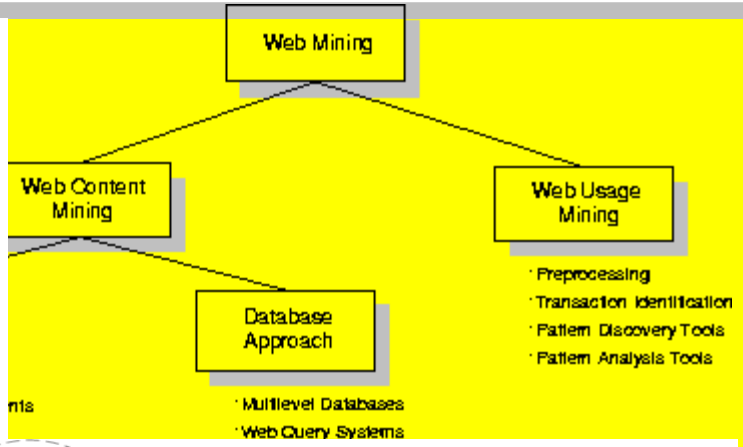
1. Constructing Data Warehouse
2. Establishing CRM Strategy
3. Constructing Data Mining System
4. Constructing CRM Marketing DB
5. Constructing E-mail Marketing System
6. Expanding E-CRM

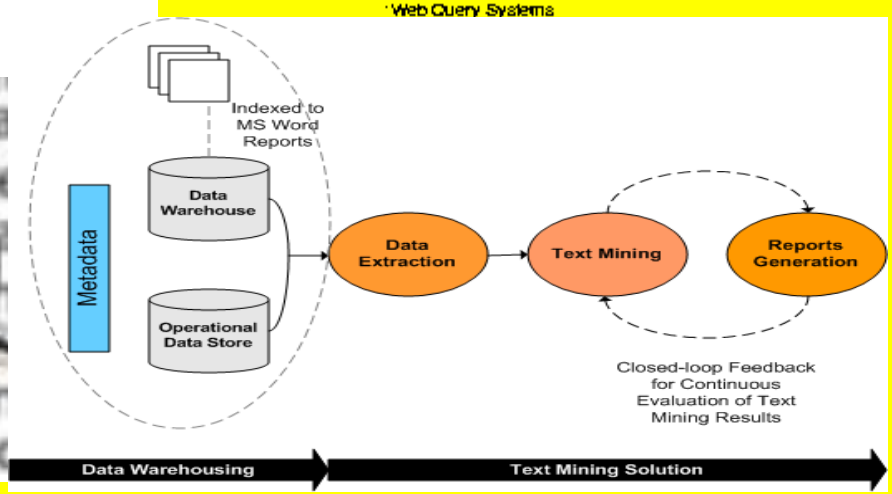
Result

Earn 50 billion won
Customer churn Rate
⇒ Decreased by 15%
(Preventing Secession of 350 Thousands Customers During 2001-2002)

Source: mis.postech.ac.kr/board/downadd.asp?wp_id=4325







Web and Text Mining in CRM

Web Mining

Definition:

Web Mining is the process of discovering useful and previously unknown patterns from the Web data

Web Mining Categories:

- **Web Content Mining**
- **Web Structure Mining**
- **Web Usage Mining**

Web Content Mining

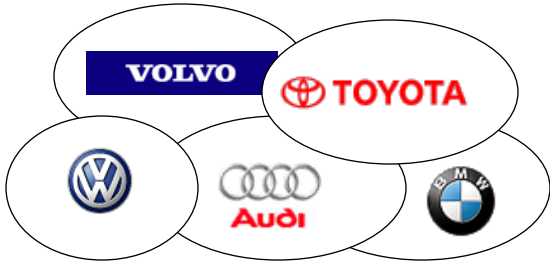
Relation to Text Mining

Examples:

- Forums analysis
- E-Mail Mining
- “Looking for others”
- Content Identification
- ……

Web Content Mining, looking for others

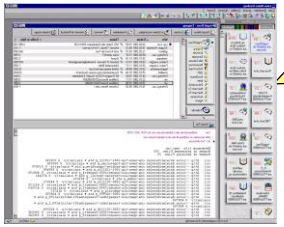
Browser-based Search Tool



Interesting Documents

Result

Foreign Web Server of competitors or suppliers or partners or ...



E-Mail:

Retrieved URLs from foreign servers related to interesting documents

Web Usage Mining

Using Web Mining to optimize the CRM-process

Examples:

- **Car configuration**
- **Recommendation systems**
- **Optimal placement of advertisements**
- **……**

Application of Data Mining in Fraud Detection

Fraud Detection, General Aspects

Corruption

- Bribery
- Embezzlement
- Fraud
- Extortion
- Favouritism
- Nepotism

is a criminal deception or the use of false representations to gain an unjust advantage. It covers both bribery and embezzlement

Fight Against Fraud

Prevention

- Can't be perfect
- Inconvenient
- Expensive

Detection

Identifying fraud as soon as it occurred

Fraud Detection, Genral Aspects

- Internal Fraud
- External Fraud

Types of Fraud

- Credit Card Fraud
- Money Laundering
- Insurance Fraud
- Telecommunication Fraud
- Computer intrusion
- ……..

detection

Data Mining can help

Fraud Detection systems „are used to catch bad guys doing bad things“

Important: Fraud Detection is a continual developing process, because patterns of fraud are dynamic and change over the time

Fraud Detection, why Data Mining ?

Why data mining is needed in Fraud Detection ?

Huge volume of Data; example:

- Over 1.59 Mrd. Visa cards in circulation
- 6800 transactions per second (peaks)
- 20000 members banks
- Millions of merchants

(Source: http://www.rgrossman.com/talks/grossman-iciq-07-v4.pdf)

Data Mining can help

- Performance Challenge
- Storage Challenge

→ Fast and efficient algorithms

→ Modern databases technology

Fraud Detection, Importance

Extent of internal fraud

- A recent survey by KPMG Peat Marwick found that nearly 60 percent of all small business owners reported that their companies have experienced some type of internal financial fraud within their own Employee.

- More than 75 percent of companies surveyed had actually been the victim of employee fraud within the previous 12-month perio

Source: http://www.nfib.com/object/2991852.html

Fraud Detection, Credit Card Fraud

Extent

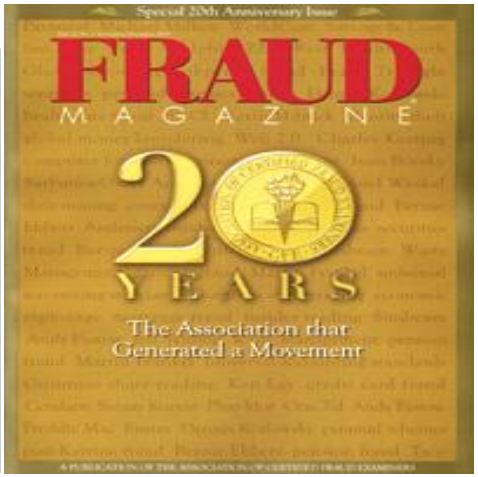
The Washington Post

21. July 2002

"**Credit card fraud costs the industry about a billion dollars a year**, or 7 cents out of every $100 spent on plastic. But that is down significantly from its peak about a decade ago, Sorrentino says, in large part because of powerful technology that can recognize unusual spending patterns."

Fraud Detection, IT Impacts

Impact of IT on Fraud perpetration



- Internal fraud is as old as business
- Internal fraud coupled with IT-savvy is a killer combination
- Since the introduction of the first commercial computer (UNIVAC, on this date in 1951) computers have been used to make the fraudster's job easier

Examples:

- Generating of bogus invoices and paying them to bogus companies
- Nigerian 419 scam
- Most large organizations swap millions into overnight instruments to take advantage of the best interest rates only to swap them back into their working accounts during the day. Skimming a piece of that transaction could be simple.

http://blogs.zdnet.com/threatchaos/?p=341

Fraud Detection, Example Health Insurance

Extent

In the U.S. alone, health care fraud is estimated to cost somewhere between $80 billion and $170 billion per year, according to the National Health Care Anti-Fraud Association.

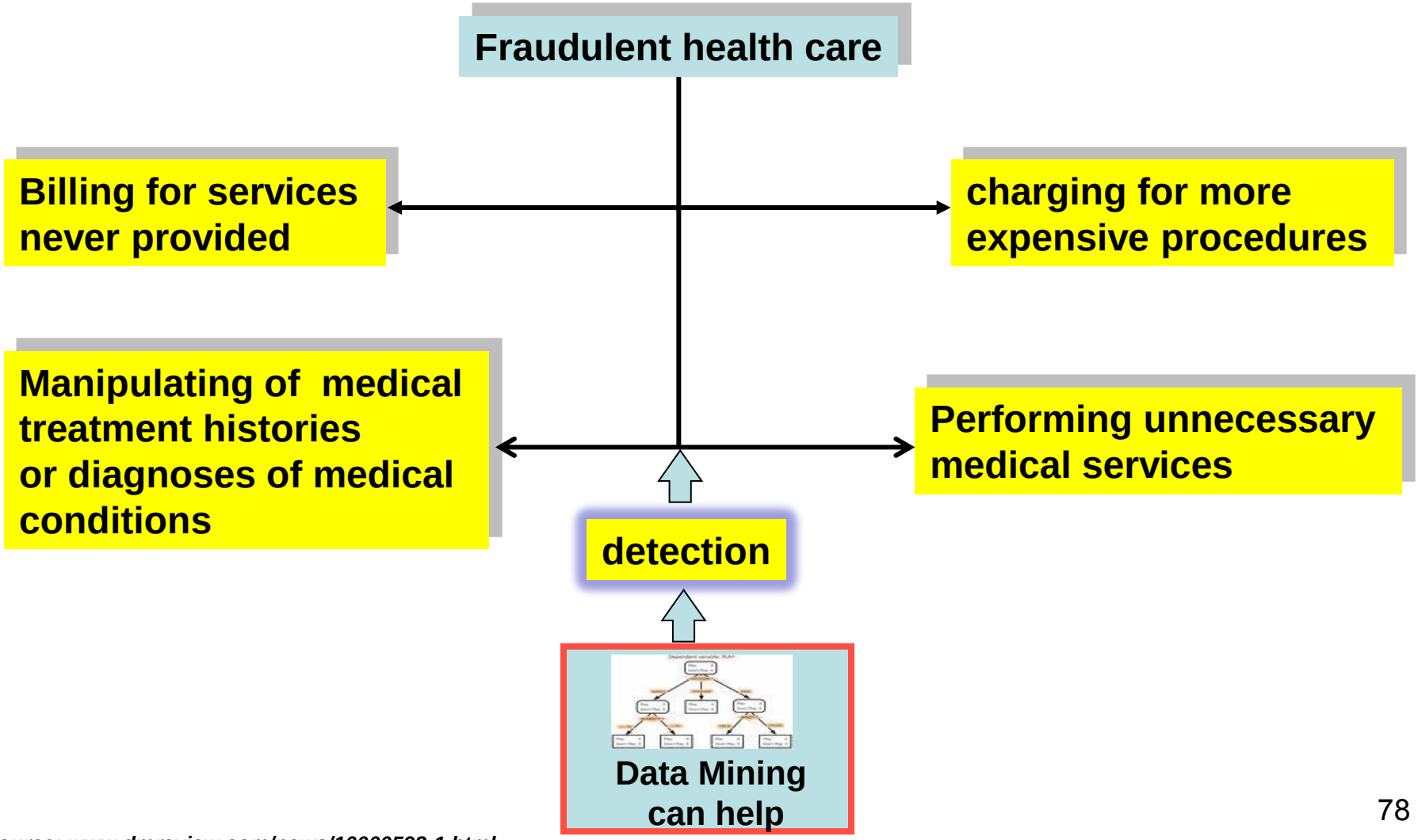
Health Care Fraud:

- **is criminal activity**
- **affects taxpayers, government agencies and health insurance firms**
- **affects human health**

→ **Alone this fact makes fraud detection necessary**

Source: www.dmreview.com/news/10000592-1.html

Fraud Detection, Example Health Insurance



Source: www.dmreview.com/news/10000592-1.html

Fraud Detection, Data Mining Methods*

There are a lot of data mining methods can be used :
Common Characteristic of Data Mining Models used in FD

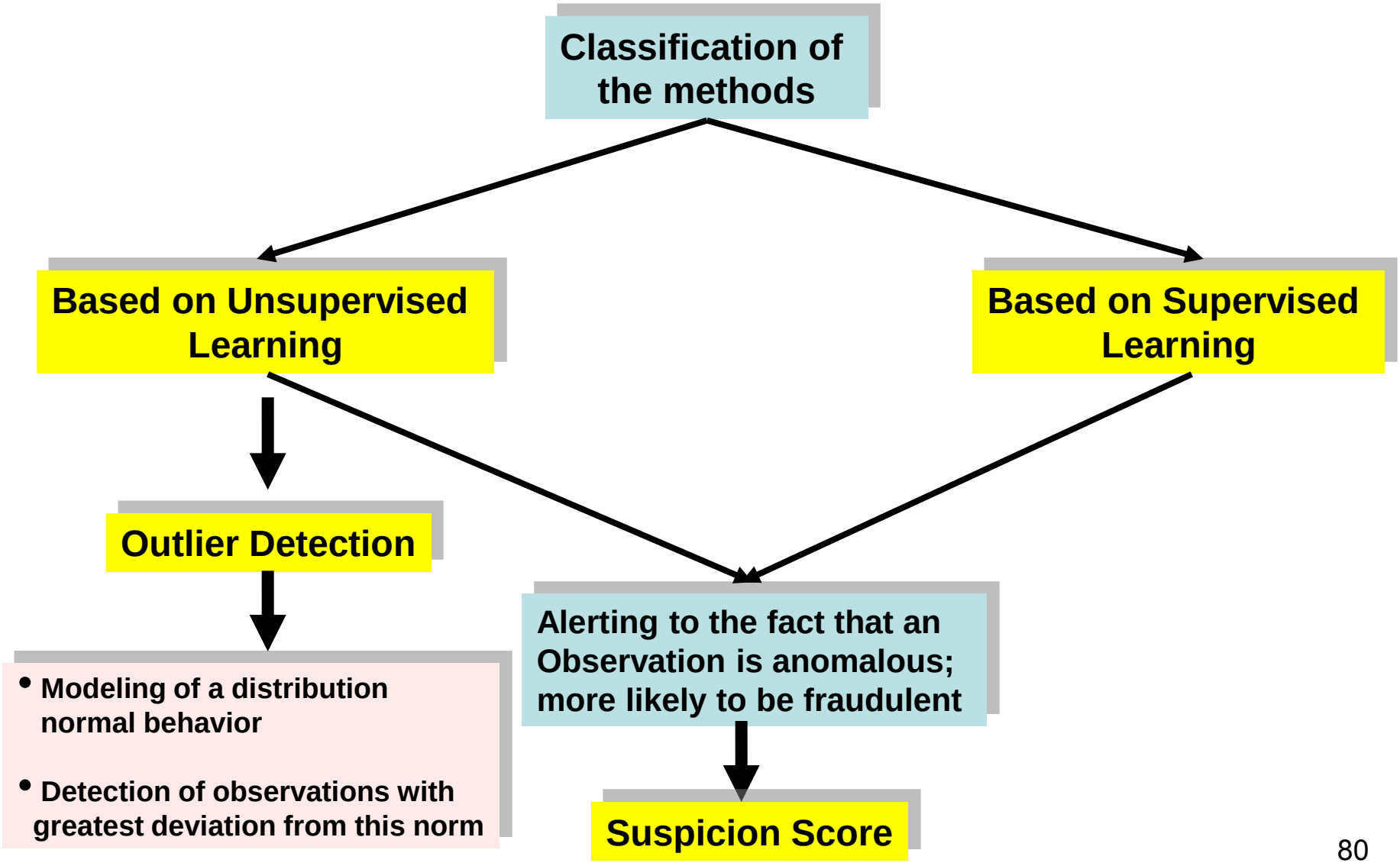
They are based on Comparing the observed Data with their expected values

Expected values can be:

- Numerical summaries of some aspect of behavior
- Simple graphical summaries showing
- Multivariate behavior profiles based on past behavior
  Example: the way of an account has been used in the past

* based on : http://metalab.uniten.edu.my/~abdrahim/ntl/Statistical%20Fraud%20Detection%20A%20Review.pdf

Fraud Detection, Data Mining Methods



Fraud Detection, Suspicion Scores

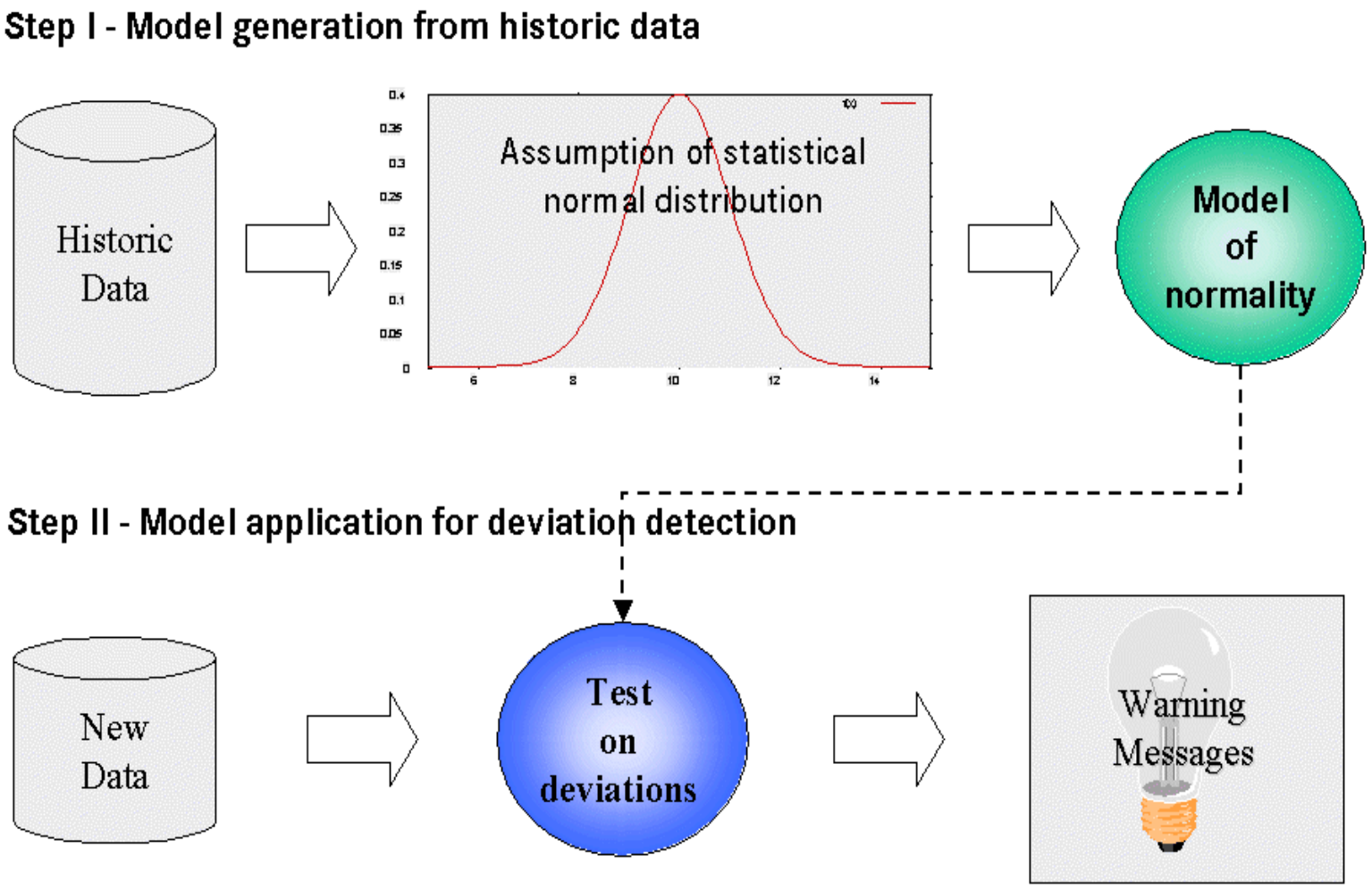
observation	ordered Suspicion score
O_1	S_1
O_2	S_2
O_3	S_3
.....	.....
.....	.....

- Compromise between the cost of detecting and saving reached
- Problematic of fraud publicity
- Damaging the customer relation in the case of false positive

Regarding analyzing cost, more attention should be paid to observations with highest scores

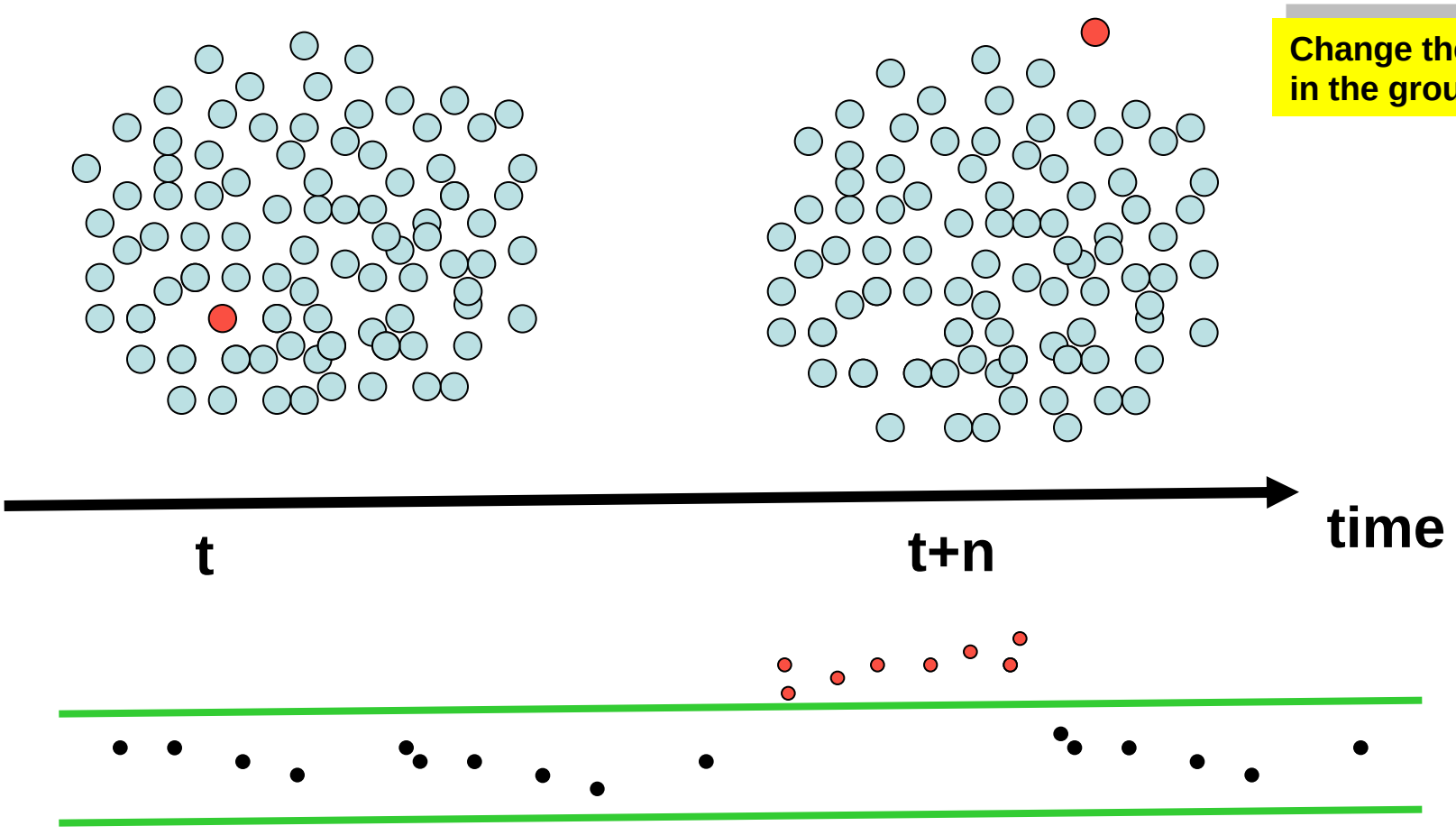
Difficult to find case studies together with the used data

Example



Source: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining
Pages: 432 - 437

Example, observation the expenditure and number of transactions

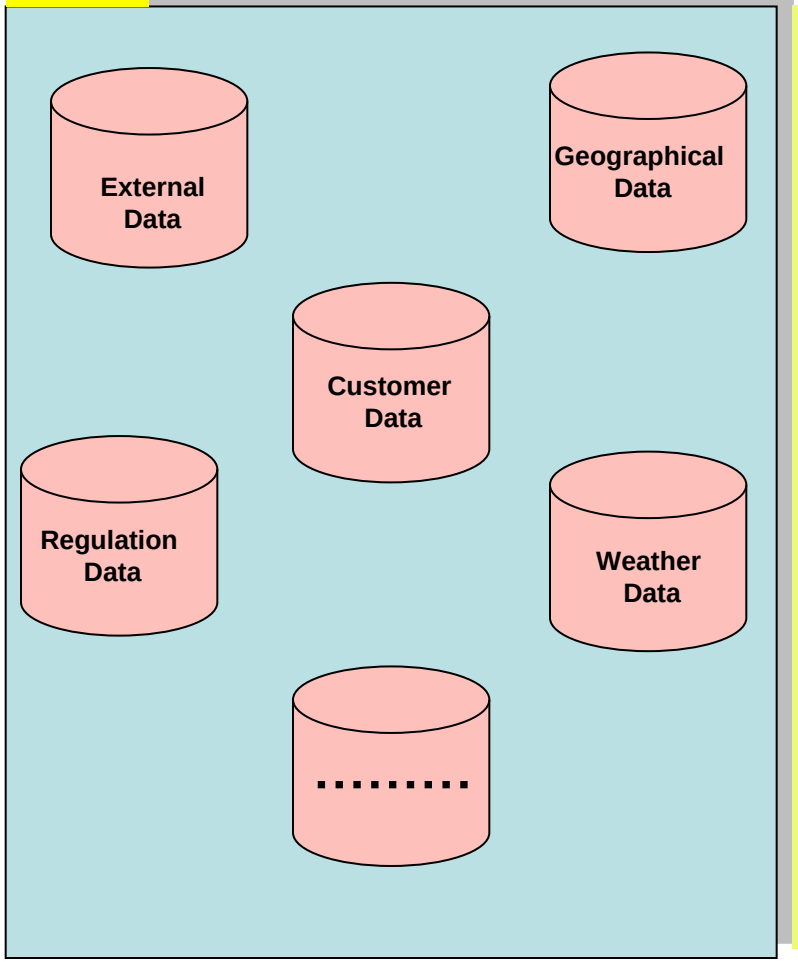


Why Data Mining in Insurance Industry ?

Business Issues

- Health
- Fire
- Building
- Transport
- Motor
- Holiday
- Accident
- Legal expenses
- Unemployment
- ….

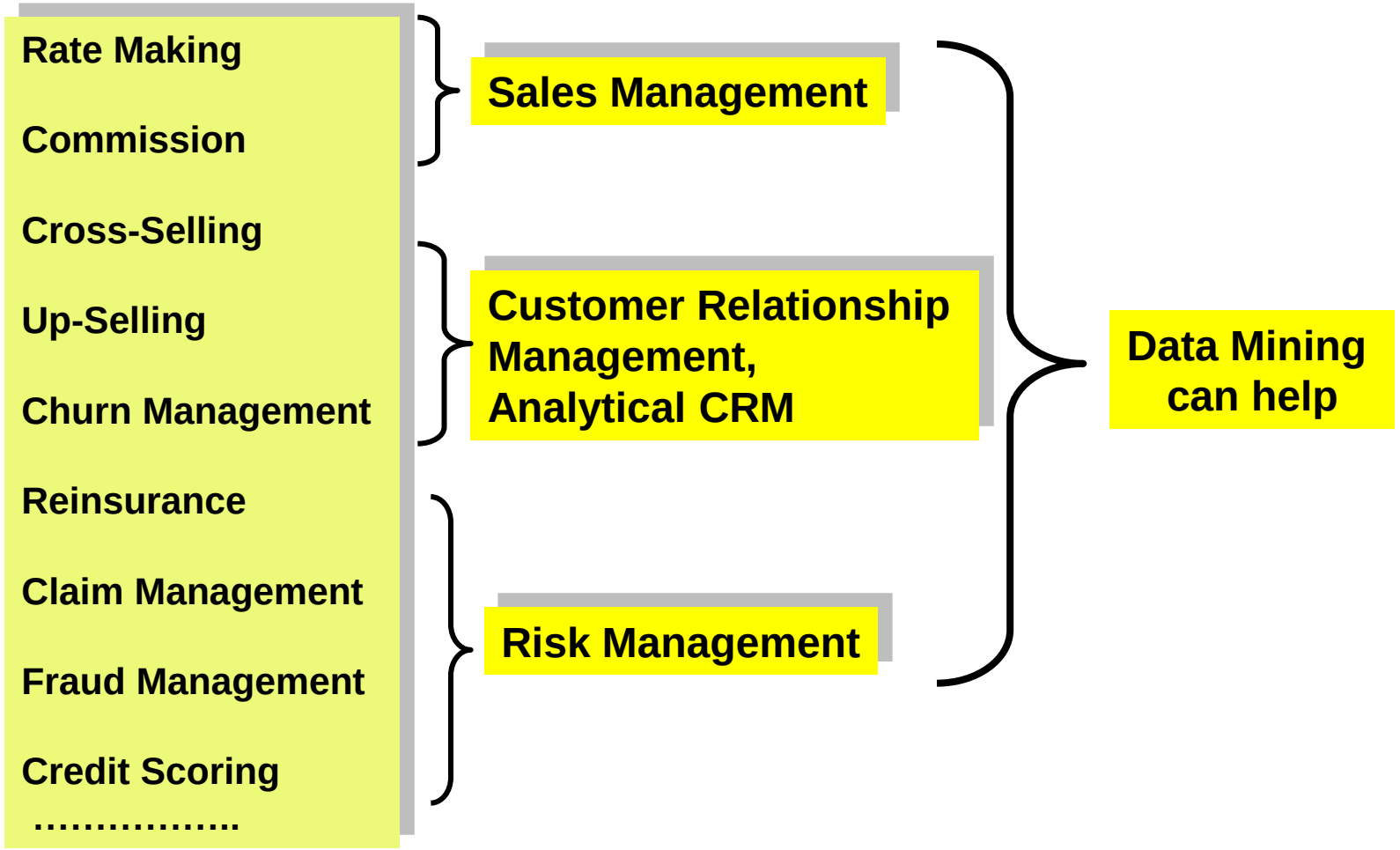
Data



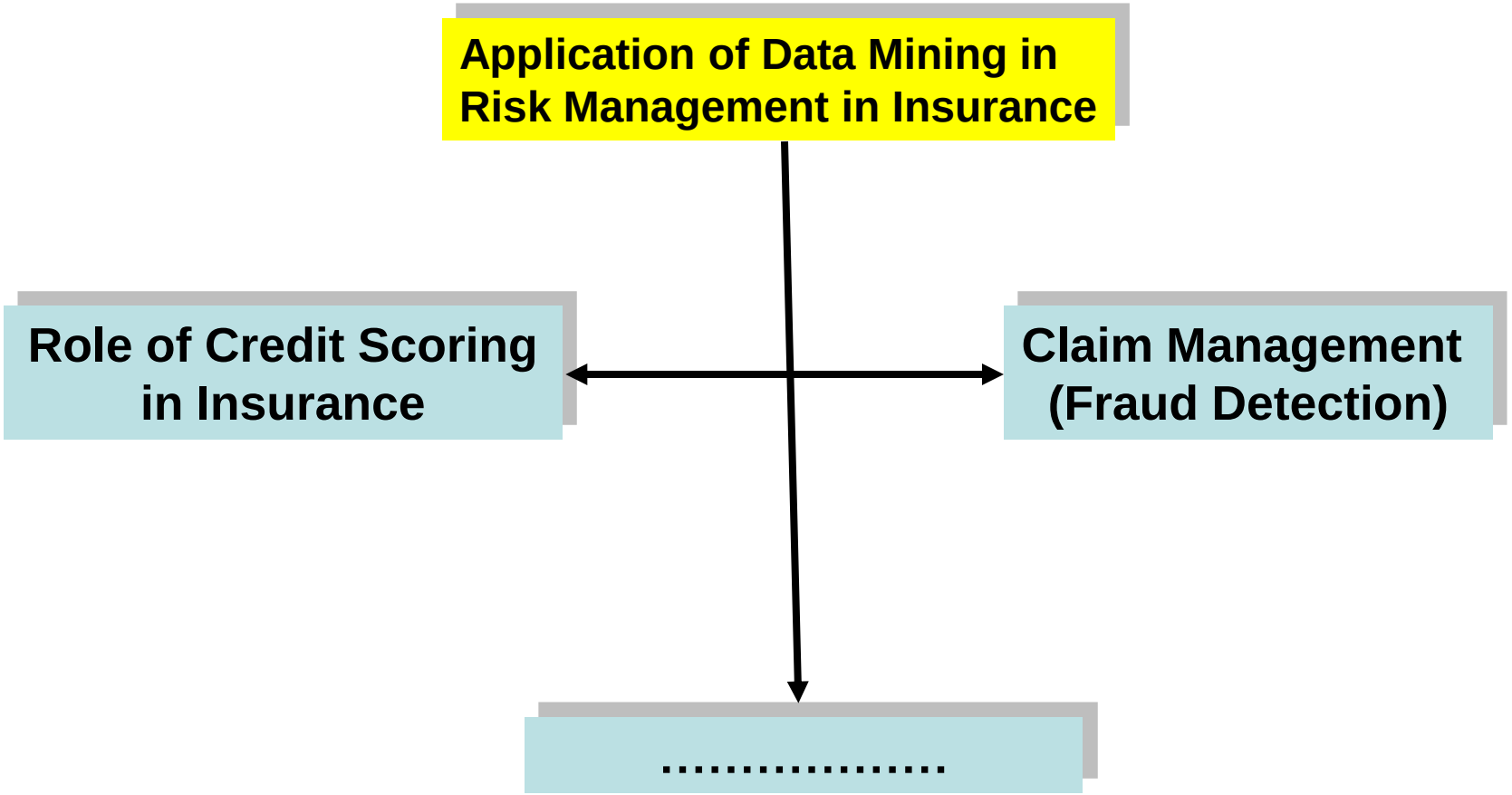
DM-Applications

- Rate Making
- Commission
- Cross-Selling
- Up-Selling
- Churn Management
- Reinsurance
- Claim Management
- Fraud Management
- Credit Scoring
- ……..

Data Mining Applications



Risk Management in Insurance



Application of Data mining in Insurance

Case Study : Fraudulent Claims

Project: Worker's compensation insurance

Methods: Supervised Learning

Tools: CART, Logistic Regression

Data : More than 90 predictors about claim (30 variables) , claimant (35 variable), injury or disease (25 variables)

Goals:

- Identifying most important predictors (Attribute selection)
- Classifying of serious and non serious claims

Reported in : http://web.maths.unsw.edu.au/~inge/statlearn/kolyshkina.pdf

Application of Data mining in Insurance

Case Study: Fraudulent Claims (continues)

Findings and Results

The CART® methodology proved superior to the methodology of logistic regression for this Problem.

Predictors identified by CART as. significant:

- some were expected to be so an the Basis of previous experience and analysis, for example, injury details and age

- some others were unexpected such as language skills of the claimant

Prediction of serious claims. The Model:
— classifies correctly about 80% of all serious claims
— targets 30% of all claims as "likely to be serious". Of these about half turn out to be serious.

Reported in : http://web.maths.unsw.edu.au/~inge/statlearn/kolyshkina.pdf

Application of Data mining in Insurance

Case Study: Claim Management

Data Mining for a health insurance company

Methods: Unsupervised and Supervised Learning

Tools: CART, MARS, hybrid model

Data : More than 300 predictors including:

- **Demographic variables (age, gender,..)**
- **Socio-economic and geographic variable (area of residence,..)**
- **Membership and product details (duration of the membership..)**
- **Claim history and medical diagnosis**
- **Other Variables: distribution channel, payment methods etc.**

Goal:
- **Identifying most important predictors (Attribute selection)**
- **Forecasting hospital claim cost**

Reported in : http://web.maths.unsw.edu.au/~inge/statlearn/kolyshkina.pdf

Application of Data mining in Insurance

Case Study : Claim Management

Findings and results:

- Predictors of the highest importance for overall hospital cost were age of the member, gender, number of hospital episodes and hospital days in the previous years, the type of cover and soeio-economic characteristics of the member.
- Other important predictors included duration of membership, family status of the member, the type of cover that the member had in the previous year, previous medical history and the number of physiotherapy services received by the member in the previous year.
- The fact that the number of ancillary services (physiotherapy) affected hospital claims cost was a particularly interesting finding.

Reported in : http://web.maths.unsw.edu.au/~inge/statlearn/kolyshkina.pdf

Application Examples:
Data Mining in Location Finding

Find the best location for an organization by using Data Mining

Business Understanding

Business Problem:

We want to find the best place for an organization

- **Customer Satisfaction**
- **more revenue**
- **……**

Business Goals

Find a place in which the highest sales is reachable

Transfer the business goals to Data Mining goals

Data Mining Goals

- Find an algorithm that forecast precisely the reachable sales of different locations

- **Target Variable is the sales volume of each organization and a continues-valued attribute**

- **Other input attributes are**
  - **Competition: the number of direct market competitors within a two-mile radius**
  - **Population: the number of people living within a three- mile radius**
  - **Income: the average household income of population measured in variable P**

- **Target variable and all attributes are continuous-valued**

→ **prediction task**

↓

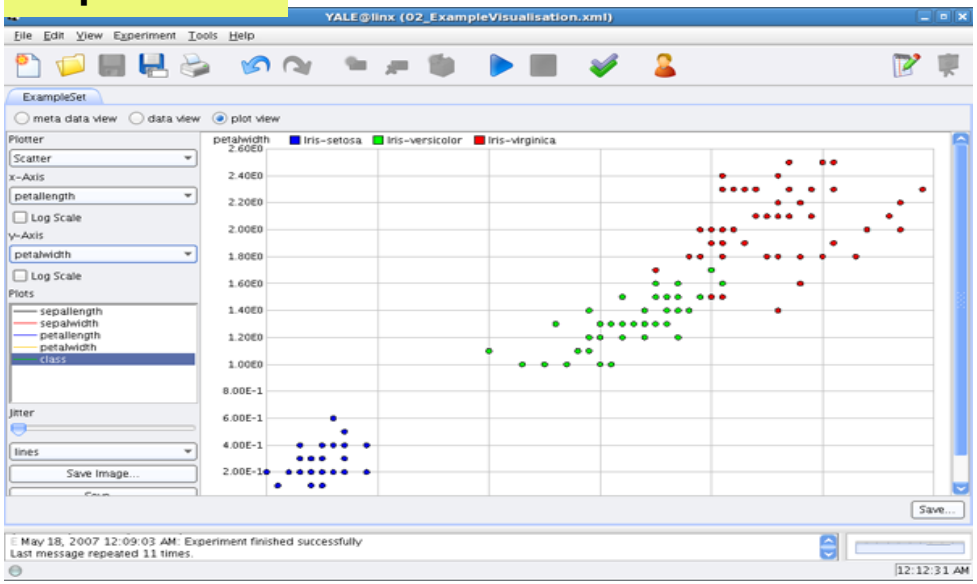
Which algorithms can solve the problem ?

↓

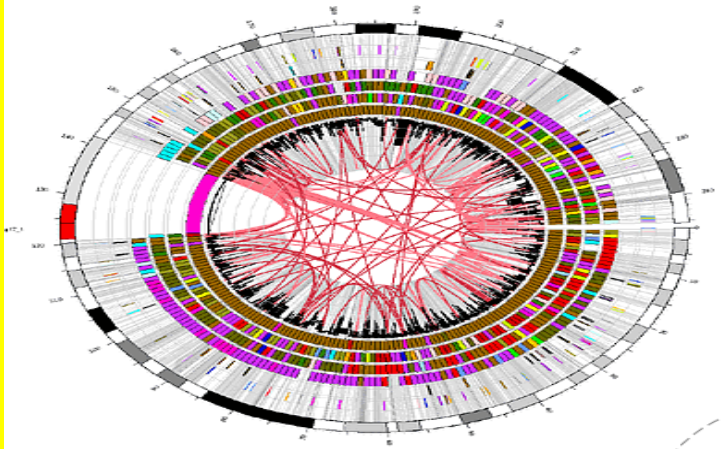
- **Regression**
- **Neural Networks**
- **Regression Trees**
- **…….**

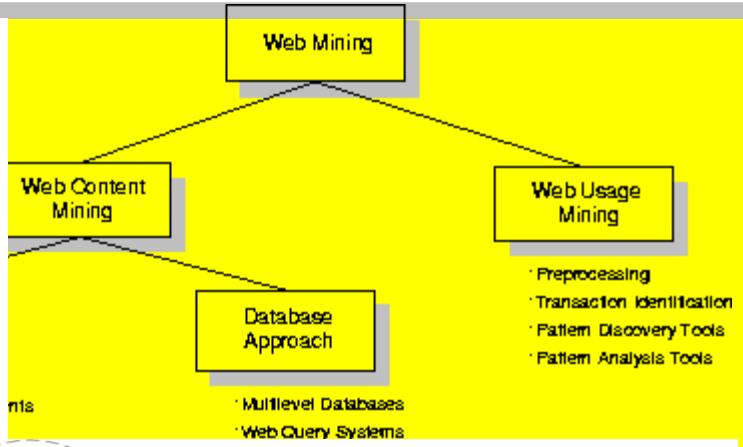
Working with DM-Tool Rapid Miner

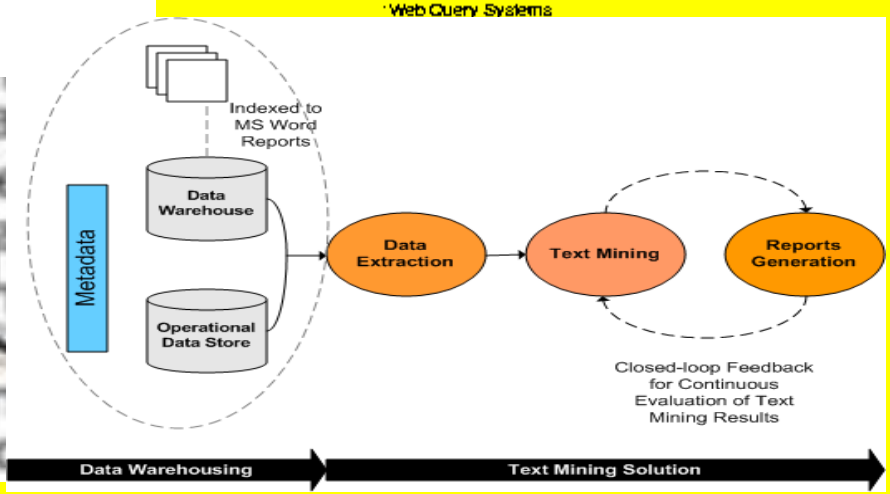
RapidMiner



Open in Workspace: Location.xml







Web and Text Mining in CRM

Web Mining

Definition:

Web Mining is the process of discovering useful and previously unknown patterns from the Web data

Web Mining Categories:

- **Web Content Mining**
- **Web Structure Mining**
- **Web Usage Mining**

Web Content Mining

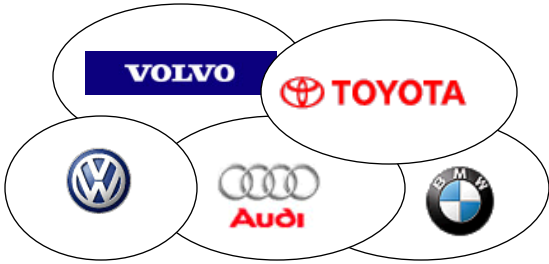
Relation to Text Mining

Examples:

- Forums analysis
- E-Mail Mining
- “Looking for others”
- Content Identification
- ……

Web Content Mining, looking for others

Browser-based Search Tool



Interesting Documents

Result

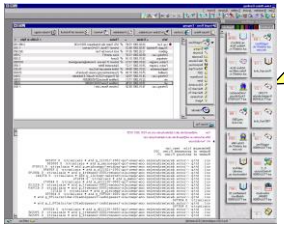
Foreign Web Server

of competitors

or suppliers

or partners

or ...



E-Mail:

Retrieved URLs from foreign servers

related to interesting documents

Web Usage Mining

Using Web Mining to optimize the CRM-process

Examples:

- **Car configuration**
- **Recommendation systems**
- **Optimal placement of advertisements**
- **……**

Application Examples:
Data Mining in Fraud Detection
Deviation Detection in Warranty Cost Statements

Case Study, REVI-MINER

REVI-MINER, a KDD Environment for Deviation Detection and Analysis of Warranty and Goodwill Cost Statements in the Automotive Industry

E. Hotz, W. Heuser, U. Grimmer,G. Nakhaeizadeh, M. Wieczorek

Abstract

In order to get the refund of vehicle repair costs, workshops of DaimlerChrysler AG worldwide regularly submit the warranty and goodwill cost statements to the central warranty department in Germany. These statements should be examined for validity and correctness, which is a very complex task for the warranty cost controllers.

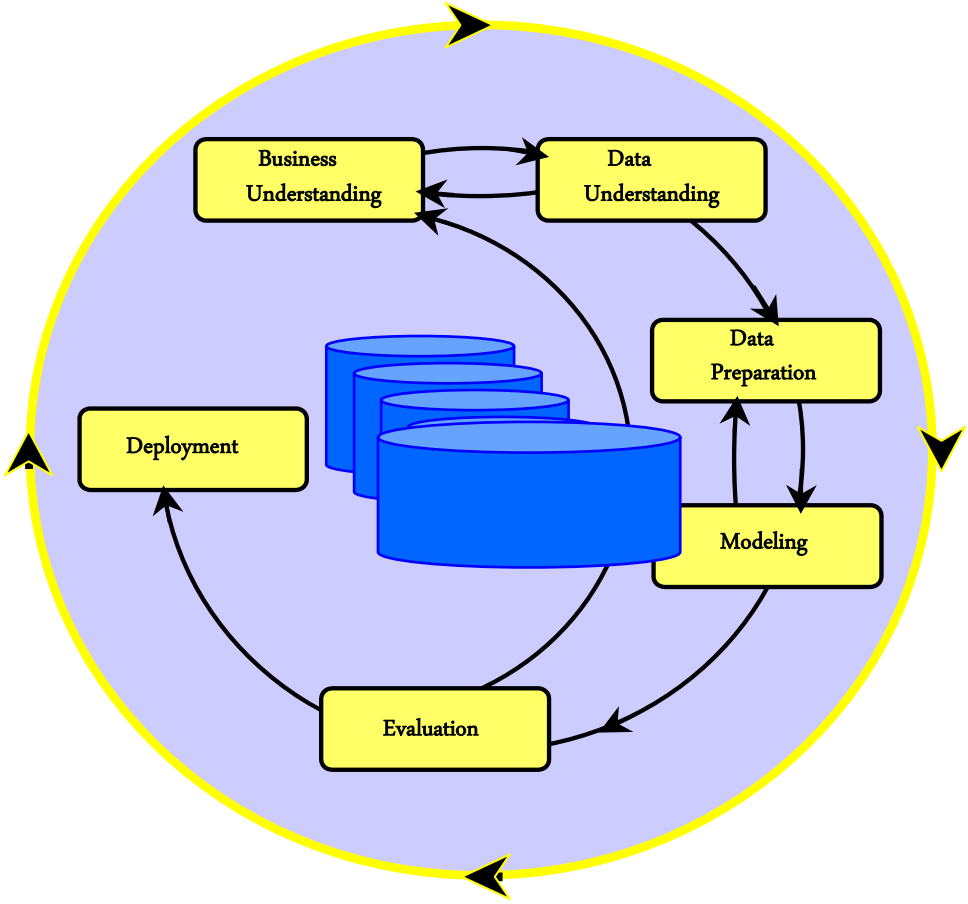
REVI-MINER is a KDD-environment which supports the detection and analysis of deviations in warranty and goodwill cost statements. The system is developed within a cooperation between DaimlerChrysler Research & Technology and the direction Global Service and Parts (GSP) and is based upon the CRISP-DM methodology as a widely accepted process model for the solution of Data Mining problems.

Furthermore we have implemented different approaches based on Machine Learning and Statistics that can be used for data cleaning in the preprocessing phase. The applied Data Mining models are developed by using a statistical deviation detection approach. The tool supports the controller within his task to audit the authorized workshops.

Source: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining
Pages: 432 - 437

Data Mining Process

CRISP-DM



Own Experience

Case Study, REVI-MINER

Business Understanding

- Refunding of vehicle repair costs
- workshops worldwide regularly submit the warranty and goodwill cost statements to the central warranty department in Germany
- These statements should be examined for validity and correctness
- This is a very complex task for the warranty cost controllers

Problem complexity

- **increasing complexity of the product structure:**

  - **different vehicle business divisions (passenger cars, trucks, transporters, busses, …**
  - **about 150 vehicle series with several body versions and combustion types**
  - **more than twenty production plants**

- **different warranty and goodwill policy for different sales markets and repair areas**

Case Study, REVI-MINER

Old Audit System

- The old Audit System was a standard system and had the following shortcomings
  - Inflexible, not very purposeful , time-consuming
  - The report generated by the system was a very complicated hardcopy table which had to be processed with difficulty manually.

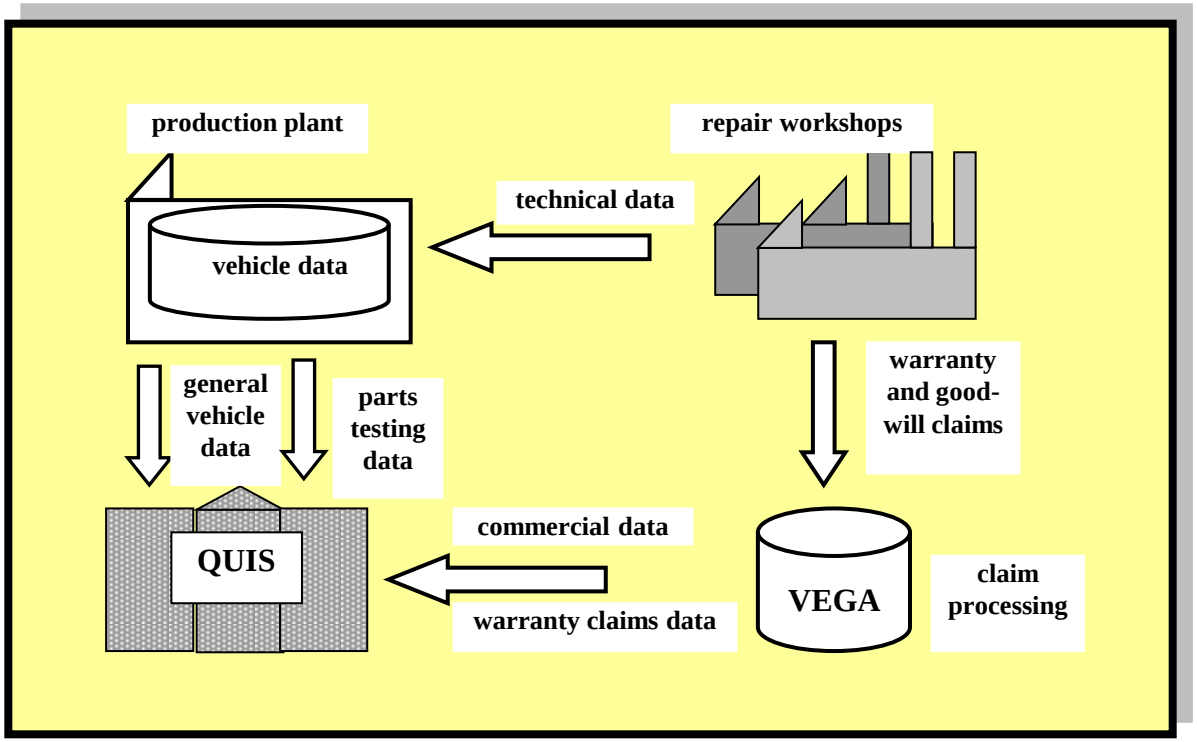
Business goal: Developing an audit system allows for:

- periodic auditing of workshops within shortening time intervals
- fast detection of possibly available abnormalities in the warranty cost statements, analyzing their trend and determining which workshop is responsible for these trends
- avoidance of false alarms by indicating fraudulent activities that really justify the controlling of the workshops
- choice from a wide range of parameters while initiating an audit report
- visualization of the results

Case Study, REVI-MINER

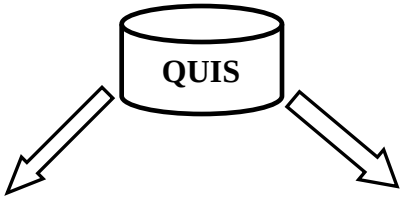
Data understanding

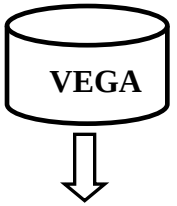
The available historical data about warranty and goodwill costs is a part of the database **QUIS (QUality Information System) that can be considered as a kind of data warehouse containing information on produced** vehicles and their repairs



Case Study, REVI-MINER

Data Preparation





query 1: general vehicle data

- VIN (vehicle ID number)
- date of production
- motor type
- continent
- country

⇒ new vehicle series
⇒ new motor types for existing vehicle series

query 2: data on repair

- date of production
- date of first permission
- date of repair
- date of credit note
- VIN (vehicle ID number)
- **dealer number (workshop) ⇒ repair area**
- total cost
- material cost
- unit cost
- incidental cost
- **...**

query 1: workshop organization

- workshop address
- repair authorization for the different vehicle business divisions
- affiliation to special workshop subgroups
  - ➢branch offices
- workshop (dealer) number
  - ➢trade partners
  - ➢representatives

Case Study, REVI-MINER

Data Preparation (continues)

Data Cleaning

To check the quality of data the following approaches are developed

- **Descriptive statistic approach: Stored (historic) data has been described by descriptive statistics. The descriptions have been compared to values known from the documentation, or other sources than (accuracy Check)**

- **a statistical prototype based on normal distribution assumption (Outlier Detection)**

- **Application of GritBot* developed by Ross Quinlan (Outlier Detection)**

* Quinlan, R., GritBot – An informal tutorial, http://www.rulequest.com/gritbot-unix.html, 2000

Case Study, REVI-MINER

Deviation Analysis

Criteria chosen for deviation analysis

Discussions with the end users showed that the needed criteria to identify and analyze deviations in warranty and goodwill data should cover the main cost types (damage types)

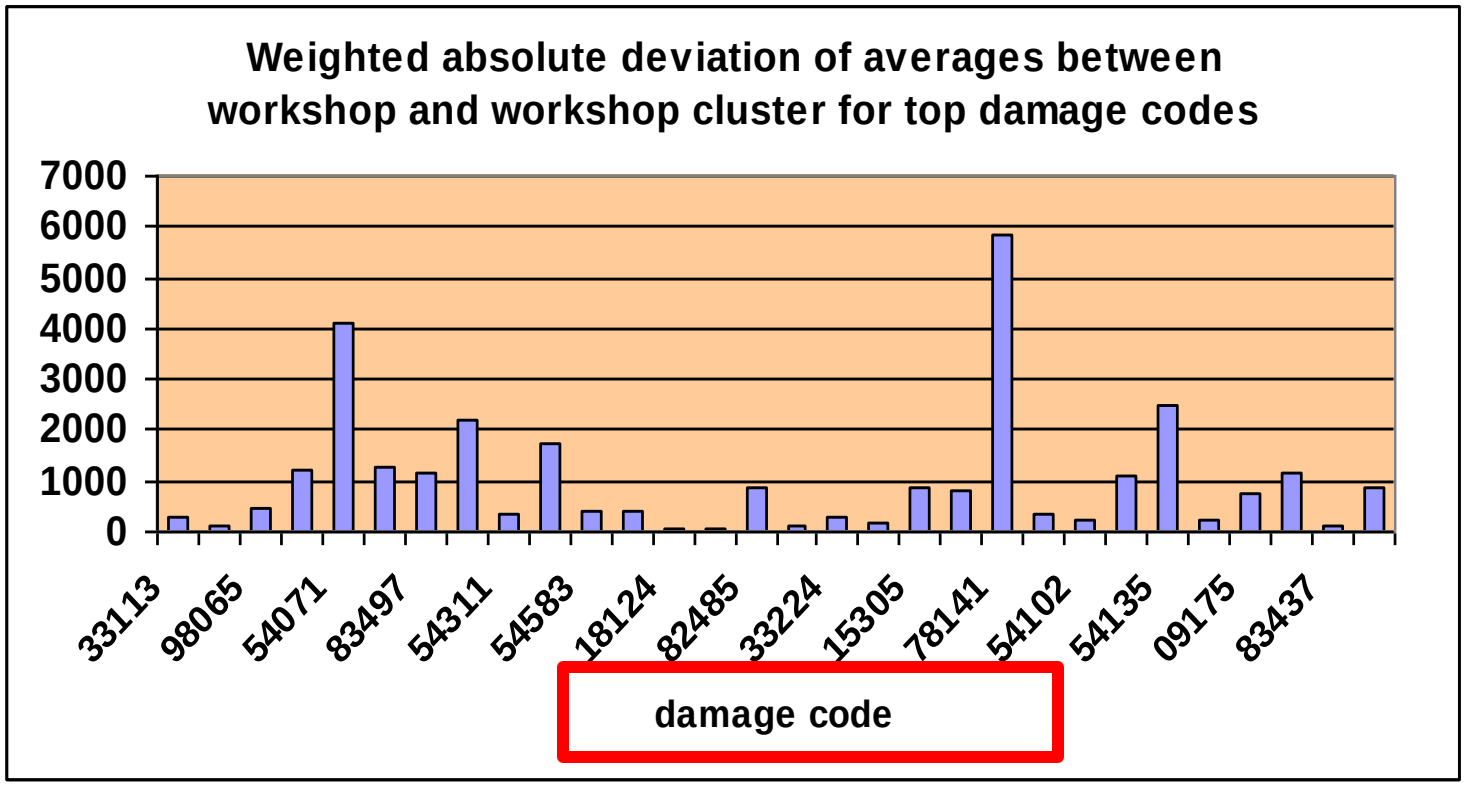
- total cost (total number of repairs)
- labor cost (number of working hours)
- cost for repair material (number of repairs with deployment of repair material)
- cost for exchange of vehicle aggregates, e.g. gear unit, air conditioner unit, motor unit (number of repairs with deployment of aggregates)

All criteria by cost and damage types must be calculated for each damage code on the chosen level of damage code aggregation (2-digit, 5-digit or 7-digit damage code) for each workshop.

Case Study, REVI-MINER

Deviation Analysis , some results

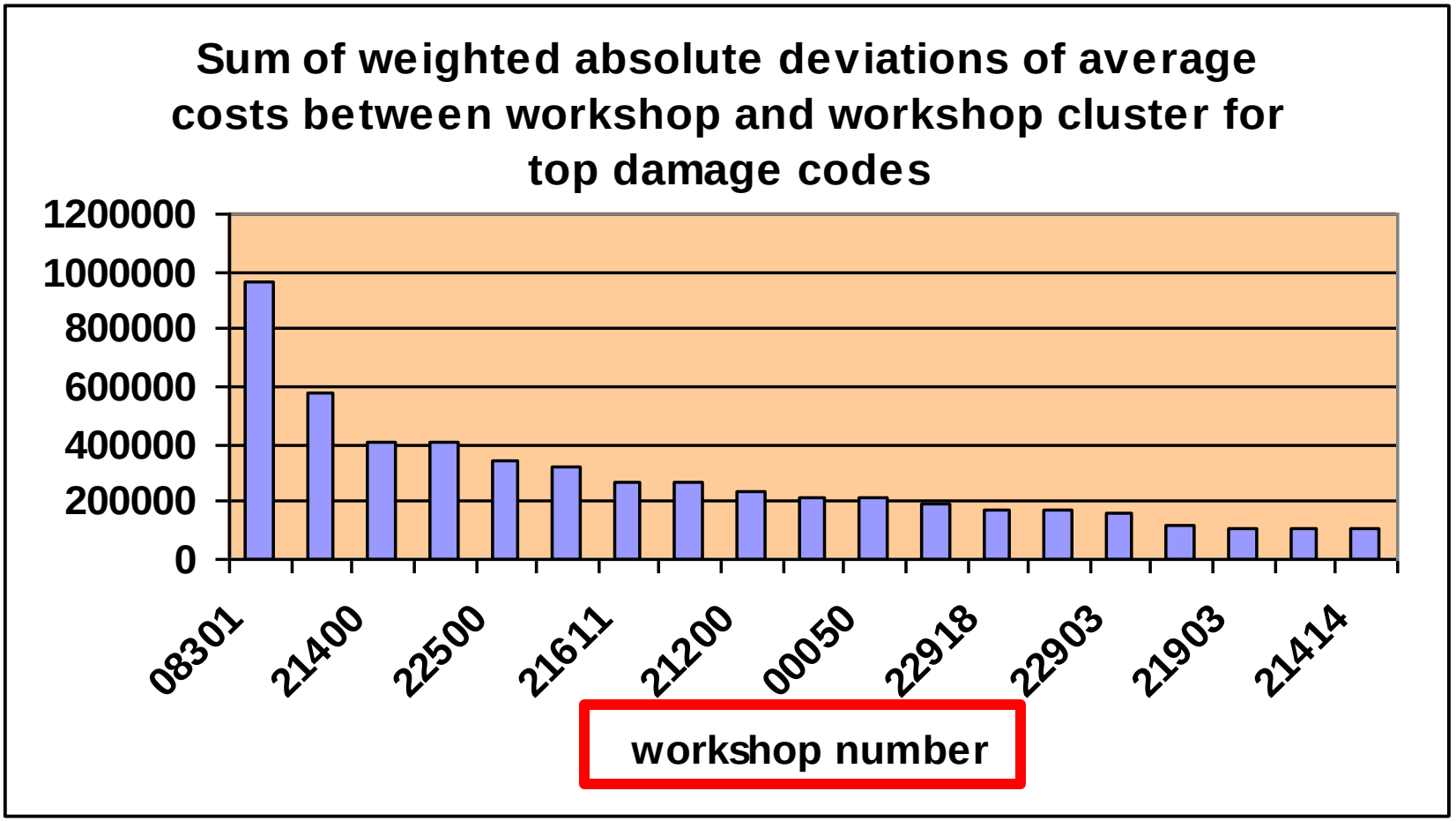
Situation of one special workshop in comparison to others in the same cluster, for each top damage



Case Study, REVI-MINER

Deviation Analysis , some results

Situation of each workshop in comparison to others in the same cluster, for some top damages



Case Study, REVI-MINER

Deployment

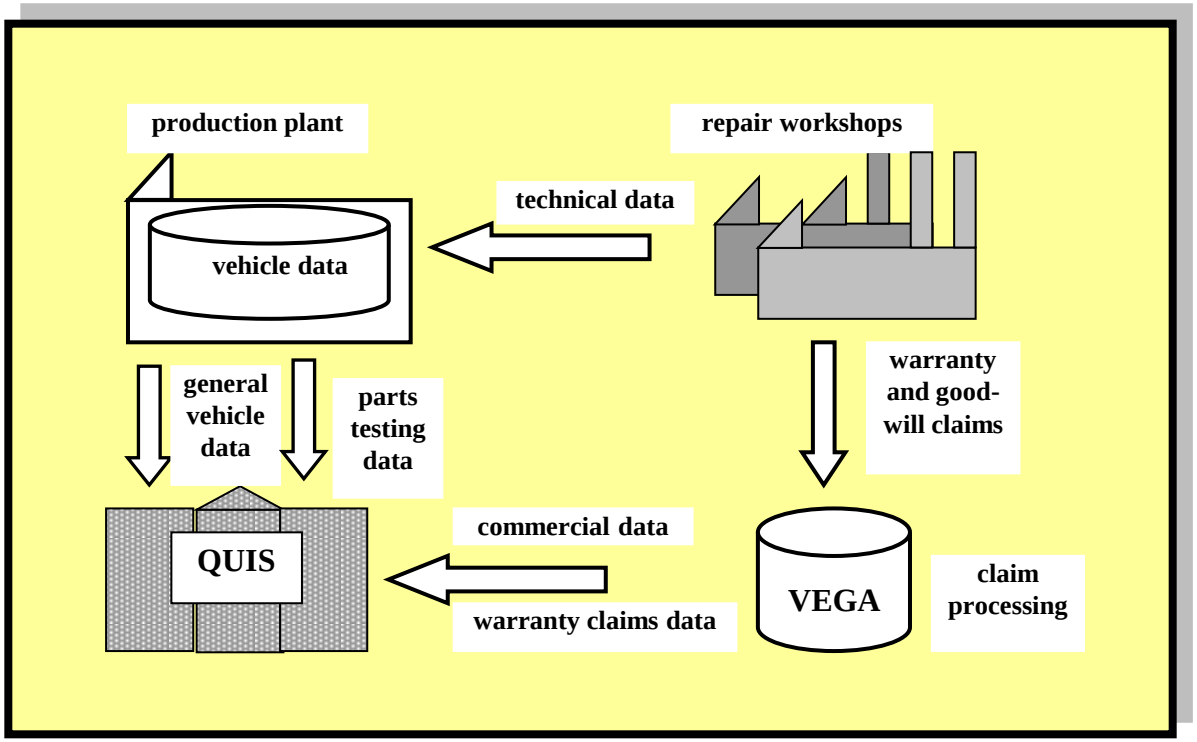
- The Data Mining tool REVI-MINER has been supporting the controlling efforts to detect and avoid fraudulent activities within the workshop organization
- Its functionality covered the essential phases of a Data Mining process and provides a user interface with easily manageable menus based upon VISUAL BASIC forms
- REVI-MINER provides the methods for a fast, efficient and meaningful analysis of the warranty and goodwill data for workshops thus giving the experts of the revision department crucial hints upon possibly fraudulent activities

Application Examples: Quality Management by Using Data Mining

Project 1: An Early Warning System for Vehicle Related Quality

Data understanding

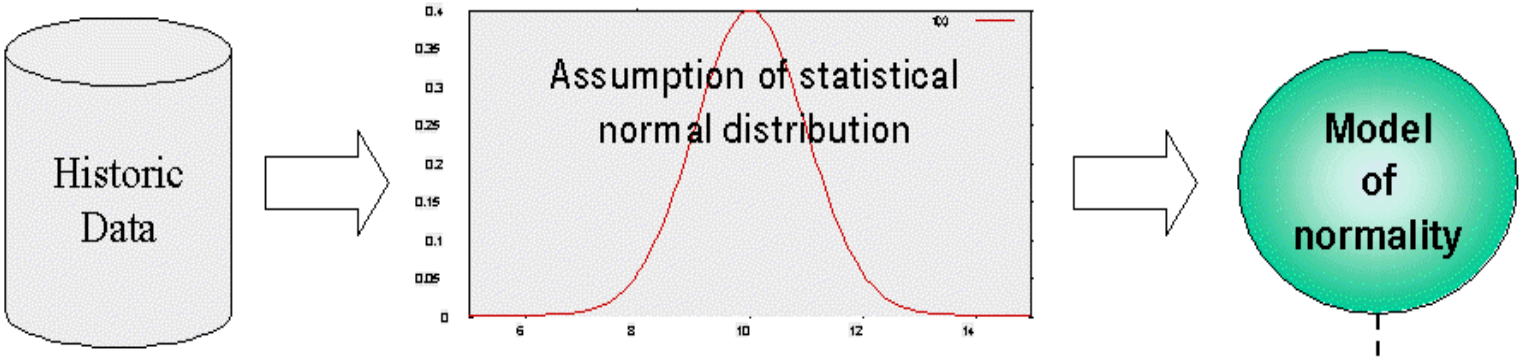
The available historical data the database **QUIS (QUality Information System) that can be considered as a kind of data warehouse containing information on produced** vehicles and their repairs



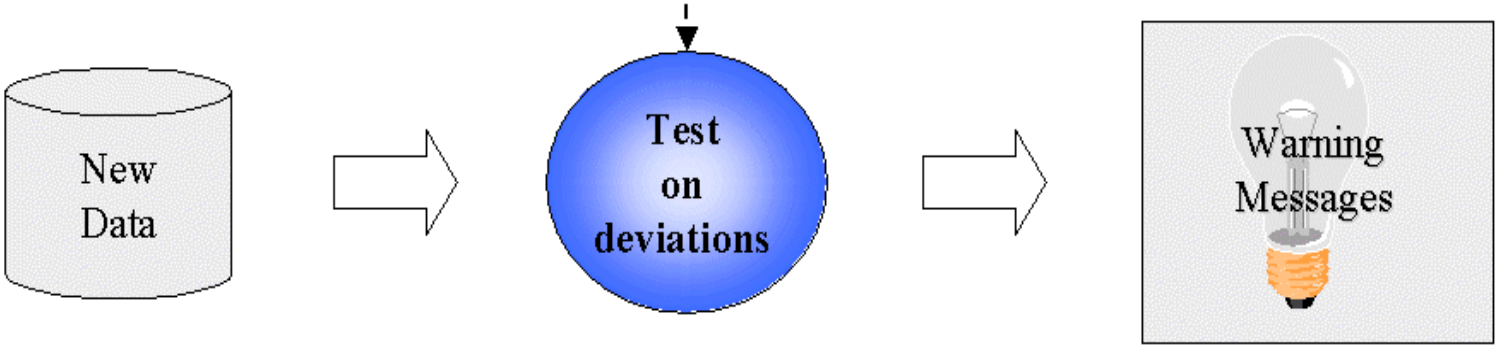
Basic idea of the early warning system

optimized by grid computing

Step I - Model generation from historic data



Step II - Model application for deviation detection



Project 2 : Fault Analysis

Film presentation