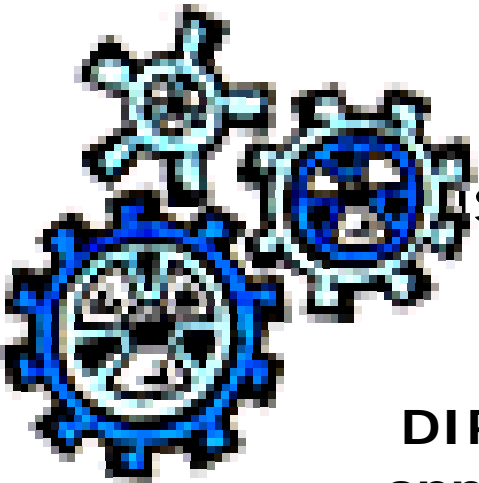


Data Mining

(Analisi dei dati ed estrazione di conoscenza)
Informatica per l'Economia e l'Azienda



Fosca Giannotti

Pisa KDD Lab, ISTI-CNR & Univ. Pisa

<http://www-kdd.isti.cnr.it/>

DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2006/2007

Data Mining

(Analisi dei dati ed estrazione di conoscenza)

⌘ Acronimo: DM (Data Mining)

⌘ Orario: Lunedì 14-16 aula D1, Giovedì 11-13 aula C

⌘ Docenti:

⊠ Fosca Giannotti, ISTI-CNR, fosca.giannotti@isti.cnr.it

⊠ Mirco Nanni, ISTI-CNR, mirco.nanni@isti.cnr.it

⌘ Ricevimento:

⊠ Giannotti: mercoledì 15-17, ISTI, Area Ricerca CNR, località San Cataldo, Pisa (prenotazione per e-mail)



Data Mining

⌘ Riferimenti bibliografici

- ⑩ Pang-Ning Tan, Michael Steinbach, Vipin Kumar, **Introduction to DATA MINING**, Addison Wesley, ISBN 0-321-32136-7, 2006
- ⑩ Jiawei Han, Micheline Kamber, **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, 2000
http://www.mkp.com/books_catalog/catalog.asp?ISBN=1-55860-489-8
- ⑩ U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (editors). **Advances in Knowledge discovery and data mining**, MIT Press, 1996.
- ⑩ Barry Linoff **Data Mining Techniques for Marketing Sales and Customer Support**, John Wiles & Sons, 2002

⌘ I lucidi utilizzati nelle lezioni saranno resi disponibili attraverso il sito web del corso:

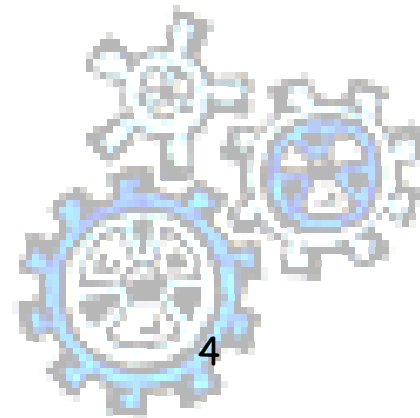
<http://www-kdd.isti.cnr.it/>



Questionario

⌘ Tipi di Lauree di primo livello:

- ☒ Laurea in Informatica: 3 Siena (WEA), 10 Pisa (Wea), Cagliari (Wea), Bari (Wea)
- ☒ Laurea in Informatica Umanistica: 4



Questionario

⌘ Messaggio e-mail fosca.giannotti@isti.cnr.it con
subject: Corso Data Mining

⌘ Contenuto

- Nome e Cognome.....
- e-mail:.....
- Corso di laurea, provenienza :.....
- Corsi di basi di dati:
-



Contenuti del corso Data Mining

Analisi dei Dati ed Estrazione di conoscenza

- ☒ Una parte preliminare dove si introducono i concetti essenziali del processo di estrazione della conoscenza: studio e preparazione dei dati, forme dei dati, misure e similarità dei dati
- ☒ Una parte centrale dove si introducono le principali tecniche di datamining (regole associative, classificazione e clustering). Di queste tecniche si studieranno gli aspetti formali e implementativi;
- ☒ Una parte più metodologica dove: si visiteranno alcune casi di studio nell'ambito del marketing, del supporto alla gestione clienti e dell'evasione fiscale.
- ☒ Una parte del corso si concentrerà sulle dimensioni e le metodologie di sviluppo del processo di sul processo di estrazione di conoscenza.
- ☒ L'ultima parte del corso ha l'obiettivo di introdurre gli aspetti di privacy ed etici inerenti all'utilizzo di tecniche inferenza sui dati e dei quali l'analista deve essere a conoscenza



Contenuti del corso

⌘ Introduzione e Concetti Basici (6 ore)

- Le applicazioni
- Il processo di knowledge discovery

⌘ Il processo di estrazione della conoscenza

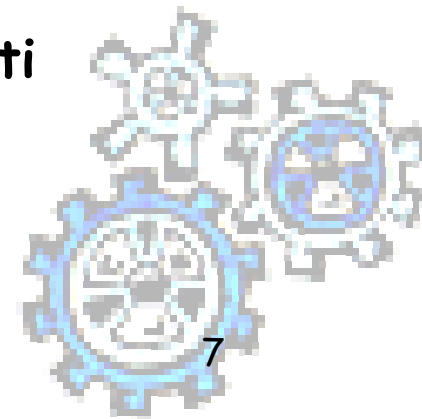
- Le fasi iniziali: preparazione e pulizia dei dati

⌘ Introduzione alle tecniche di base (6 + 2 Ore)

- Regole Associative
- Alberi di decisione
- Clustering

⌘ Algoritmi di Base (4 + 2 ore)

- Regole associative: algoritmo Apriori e varianti
- Alberi di Decisione: C4.5
- Clustering: K-Means



Contenuti del corso

⌘ Il processo di estrazione della conoscenza

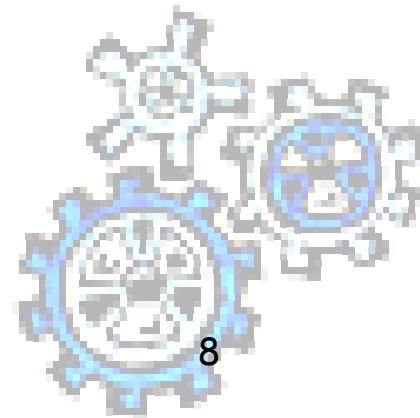
- Le dimensioni del processo
- Esempi di estrazione (Evasione fiscale, Business Intelligence)

⌘ Interpretazione e valutazione della qualità della conoscenza estratta

- Rassegna di strumenti commerciali
- Uno standard metodologico: CRISP
- Alcuni casi di studio: Customer Segmentation, Basket Marketing Analysis, GeoMarketing...

⌘ Aspetti di privacy (2 ore)

⌘ Contenuti del libro



Modalità di valutazione

⌘ Verifica Intermedia 40%

⌘ Progetto (Analisi dei dati): 60%

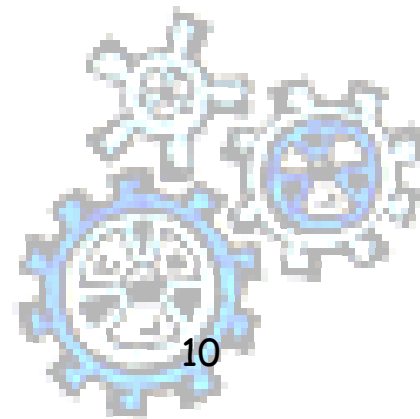
☑️ Progetti: Si dovranno fare gruppi da due. Gli studenti di un gruppo riceveranno lo stesso voto. La divisione del lavoro è loro responsabilità. I progetti, corredati di relazione, debbono essere presentati oralmente e se necessario con dimostrazione.

☑️ I lucidi e le relazioni debbono essere rese disponibili in PDF, PPT o HTML.



Seminar 1 outline

- ⌘ Motivations
- ⌘ Application Areas
- ⌘ KDD Decisional Context
- ⌘ KDD Process
- ⌘ Architecture of a KDD system
- ⌘ The KDD steps in short
- ⌘ 4 Examples in short



Evolution of Database Technology: from data management to data analysis

⌘ 1960s:

☒ Data collection, database creation, IMS and network DBMS.

⌘ 1970s:

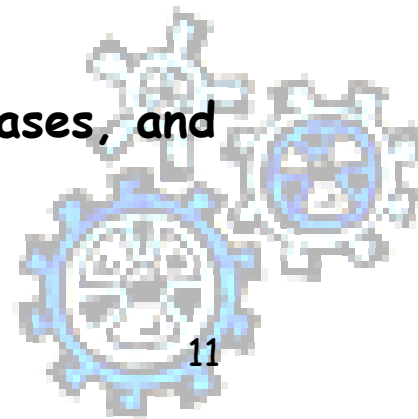
☒ Relational data model, relational DBMS implementation.

⌘ 1980s:

☒ RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.).

⌘ 1990s:

☒ Data mining and data warehousing, multimedia databases, and Web technology.



Why Data Mining

⌘ Increased Availability of Huge Amounts of Data

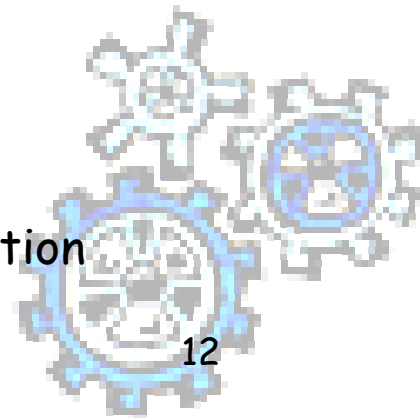
- ⊗ point-of-sale customer data
- ⊗ digitization of text, images, video, voice, etc.
- ⊗ World Wide Web and Online collections

⌘ Data Too Large or Complex for Classical or Manual Analysis

- ⊗ number of records in millions or billions
- ⊗ high dimensional data (too many fields/features/attributes)
- ⊗ often too sparse for rudimentary observations
- ⊗ high rate of growth (e.g., through logging or automatic data collection)
- ⊗ heterogeneous data sources

⌘ Business Necessity

- ⊗ e-commerce
- ⊗ high degree of competition
- ⊗ personalization, customer loyalty, market segmentation



Motivations

“Necessity is the Mother of Invention”

⌘ *Data explosion problem:*

- ⊞ Automated data collection tools, mature database technology and internet lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.

⌘ *We are drowning in information, but starving for knowledge!* (John Naisbett)

⌘ *Data warehousing and data mining :*

- ⊞ On-line analytical processing
- ⊞ Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases.

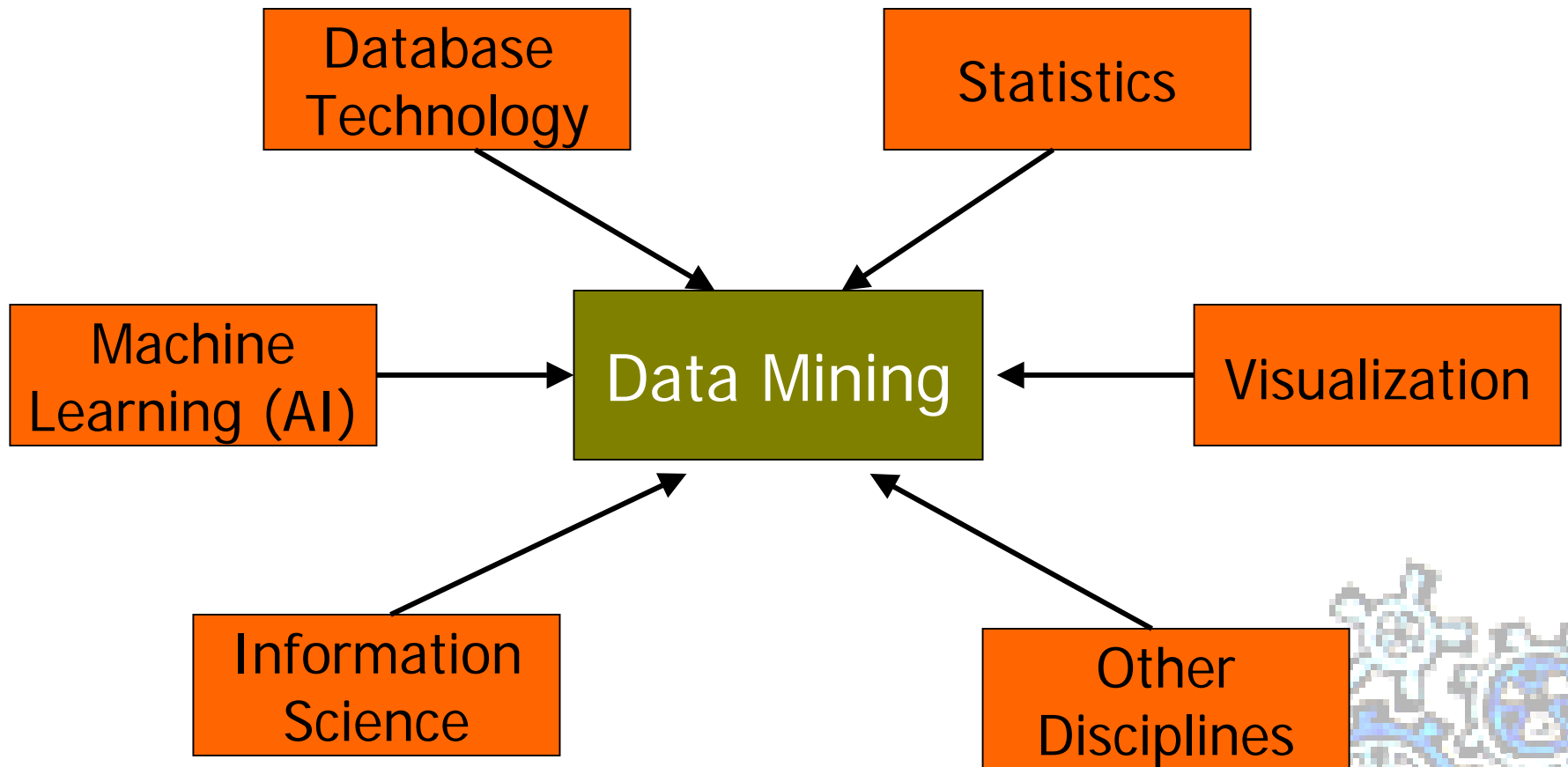


Motivations for DM

- ⌘ Abundance of business and industry data
- ⌘ Competitive focus - Knowledge Management
- ⌘ Inexpensive, powerful computing engines
- ⌘ Strong theoretical/mathematical foundations
 - ⊞ machine learning & artificial intelligence
 - ⊞ statistics
 - ⊞ database management systems



Data Mining: Confluence of Multiple Disciplines



Sources of Data

⌘ Business Transactions

- ☒ widespread use of bar codes => storage of millions of transactions daily (e.g., Walmart: 2000 stores => 20M transactions per day)
- ☒ most important problem: effective use of the data in a reasonable time frame for competitive decision-making
- ☒ e-commerce data

⌘ Scientific Data

- ☒ data generated through multitude of experiments and observations
- ☒ examples, geological data, satellite imaging data, NASA earth observations
- ☒ rate of data collection far exceeds the speed by which we analyze the data

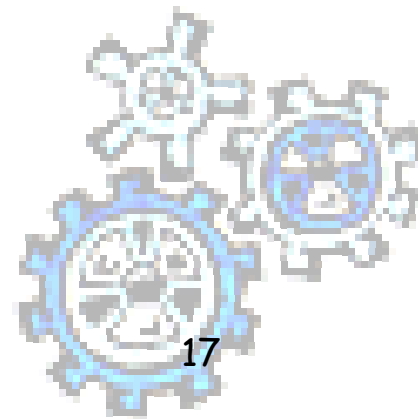
Sources of Data

⌘ Financial Data

- ☒ company information
- ☒ economic data (GNP, price indexes, etc.)
- ☒ stock markets

⌘ Personal / Statistical Data

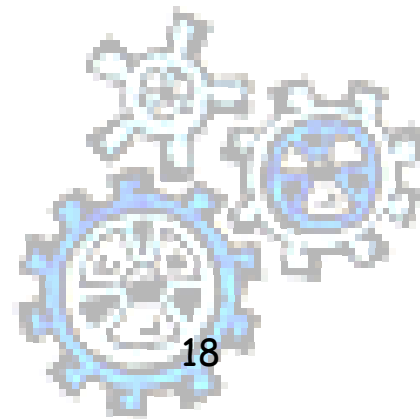
- ☒ government census
- ☒ medical histories
- ☒ customer profiles
- ☒ demographic data
- ☒ data and statistics about sports and athletes



Sources of Data

⌘ World Wide Web and Online Repositories

- ☒ email, news, messages
- ☒ Web documents, images, video, etc.
- ☒ link structure of of the hypertext from millions of Web sites
- ☒ Web usage data (from server logs, network traffic, and user registrations)
- ☒ online databases, and digital libraries



Classes of applications

⌘ Database analysis and decision support

☒ Market analysis

- target marketing, customer relation management, market basket analysis, cross selling, market segmentation.

☒ Risk analysis

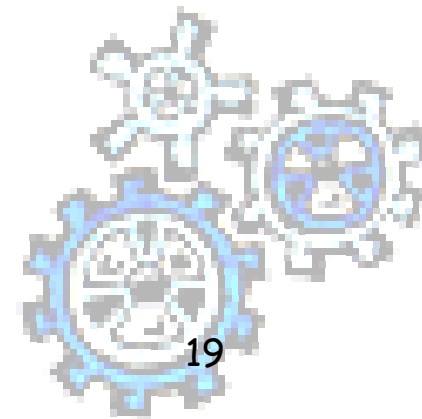
- Forecasting, customer retention, improved underwriting, quality control, competitive analysis.

☒ Fraud detection

⌘ New Applications from New sources of data

☒ Text (news group, email, documents)

☒ Web analysis and intelligent search



Market Analysis

⌘ Where are the data sources for analysis?

- ☑ Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies.

⌘ Target marketing

- ☑ Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.

⌘ Determine customer purchasing patterns over time

- ☑ Conversion of single to a joint bank account: marriage, etc.

⌘ Cross-market analysis

- ☑ Associations/co-relations between product sales
- ☑ Prediction based on the association information.



Market Analysis (2)

⌘ Customer profiling

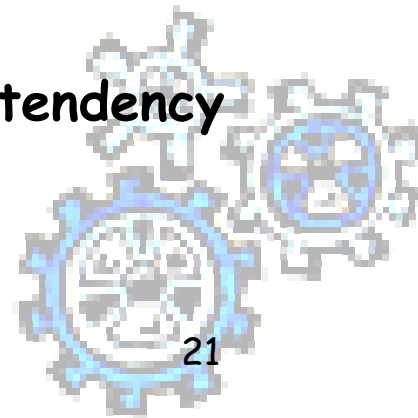
- ☑ data mining can tell you what types of customers buy what products (clustering or classification).

⌘ Identifying customer requirements

- ☑ identifying the best products for different customers
- ☑ use prediction to find what factors will attract new customers

⌘ Summary information

- ☑ various multidimensional summary reports;
- ☑ statistical summary information (data central tendency and variation)



Risk Analysis

⌘ Finance planning and asset evaluation:

- ☑ cash flow analysis and prediction
- ☑ contingent claim analysis to evaluate assets
- ☑ trend analysis

⌘ Resource planning:

- ☑ summarize and compare the resources and spending

⌘ Competition:

- ☑ monitor competitors and market directions (CI: *competitive intelligence*).
- ☑ group customers into classes and class-based pricing procedures
- ☑ set pricing strategy in a highly competitive market



Fraud Detection

⌘ Applications:

- ☑ widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.

⌘ Approach:

- ☑ use historical data to build models of fraudulent behavior and use data mining to help identify similar instances.

⌘ Examples:

- ☑ auto insurance: detect a group of people who stage accidents to collect on insurance
- ☑ money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
- ☑ medical insurance: detect professional patients and ring of doctors and ring of references



Fraud Detection (2)

⌘ More examples:

☒ *Detecting inappropriate medical treatment:*

- ☒ Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (save Australian \$1m/yr).

☒ *Detecting telephone fraud:*

- ☒ Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.

- ☒ **Retail:** Analysts estimate that 38% of retail shrink is due to dishonest employees.



Other applications

⌘ Sports

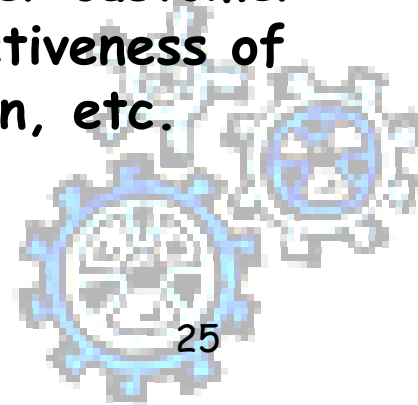
- ☒ IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat.

⌘ Astronomy

- ☒ JPL and the Palomar Observatory discovered 22 quasars with the help of data mining

⌘ Internet Web Surf-Aid

- ☒ IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

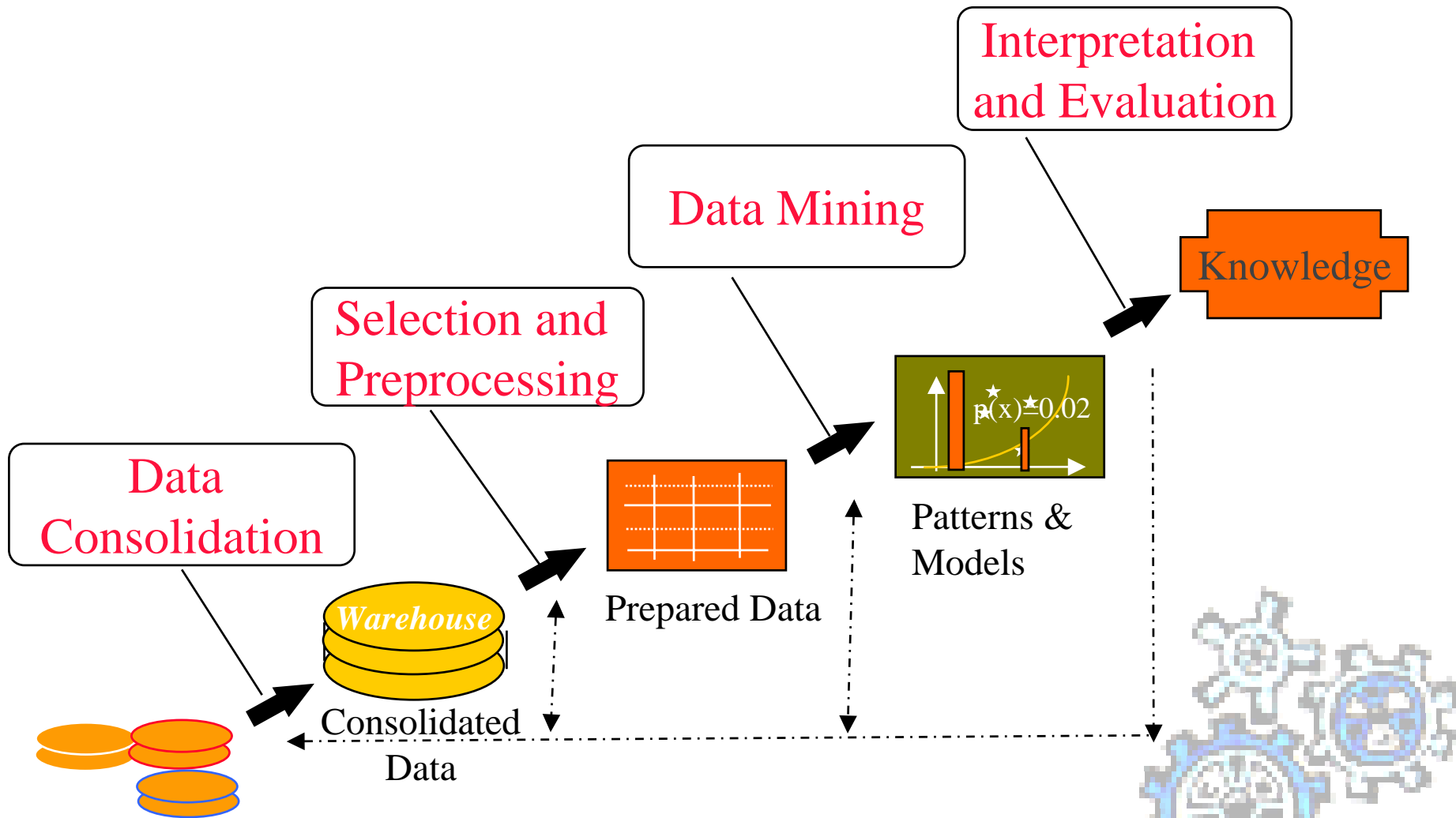


What is Knowledge Discovery in Databases (KDD)? A process!

- ⌘ The selection and processing of data for:
 - ⊞ the identification of **novel**, accurate, and **useful** patterns, and
 - ⊞ the modeling of real-world phenomena.
- ⌘ **Data mining** is a major component of the KDD process - automated discovery of patterns and the development of predictive and explanatory models.



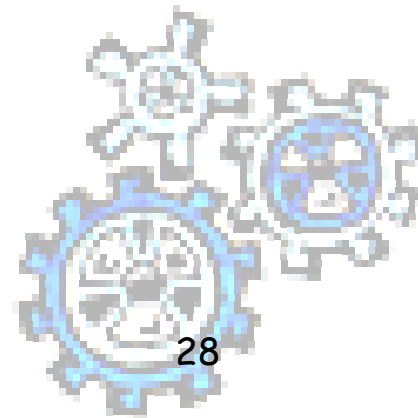
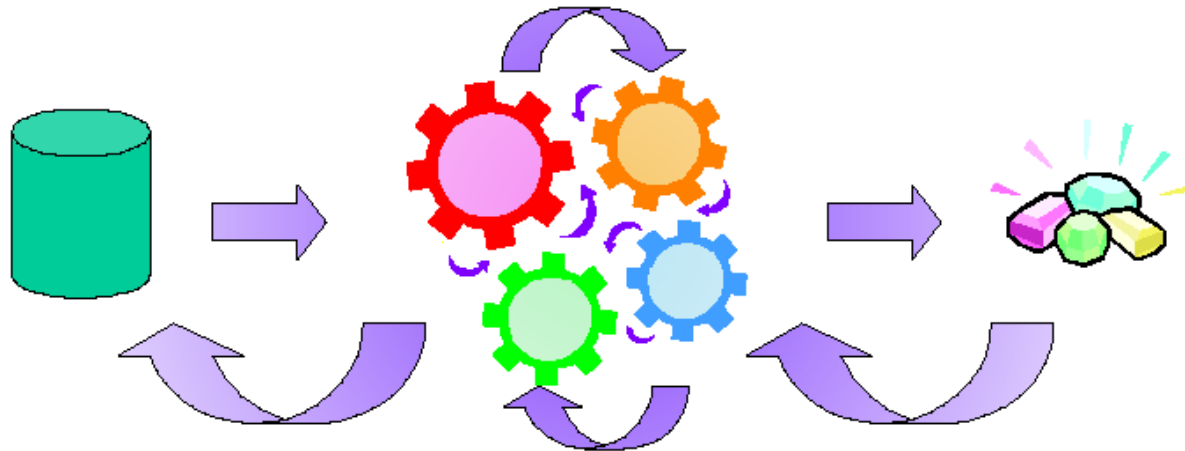
The KDD process



The KDD Process in Practice

⌘ KDD is an Iterative Process

☒ art + engineering rather than science



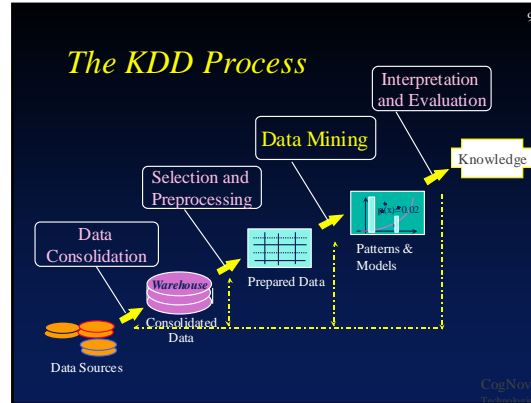
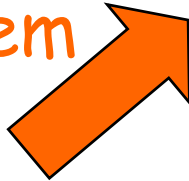
The steps of the KDD process

- ⌘ Learning the application domain:
 - ⊞ relevant prior knowledge and goals of application
- ⌘ **Data consolidation**: Creating a target data set
- ⌘ **Selection and Preprocessing**
 - ⊞ *Data cleaning* : (may take 60% of effort!)
 - ⊞ *Data reduction and projection*:
 - ⊞ find useful features, dimensionality/variable reduction, invariant representation.
- ⌘ Choosing functions of data mining
 - ⊞ summarization, classification, regression, association, clustering.
- ⌘ Choosing the mining algorithm(s)
- ⌘ **Data mining**: search for patterns of interest
- ⌘ **Interpretation and evaluation**: analysis of results.
 - ⊞ *visualization, transformation, removing redundant patterns, ...*
- ⌘ Use of discovered knowledge



The virtuous cycle

Problem



Knowledge



Identify
Problem or
Opportunity

Act on
Knowledge

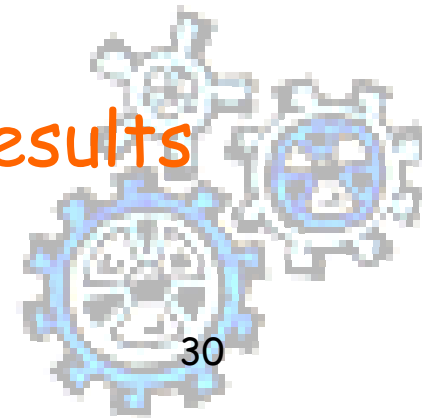
Strategy



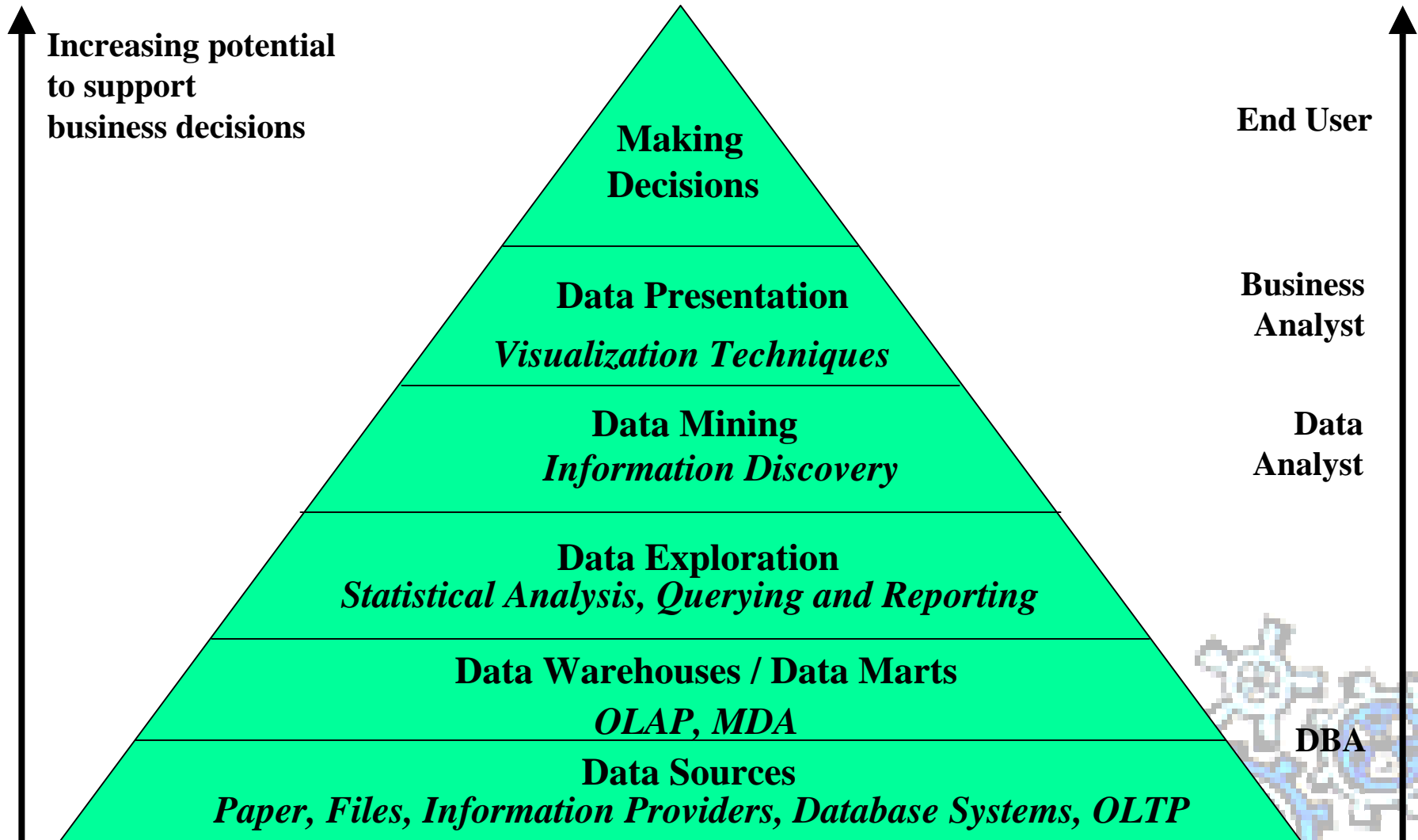
Measure effect
of Action



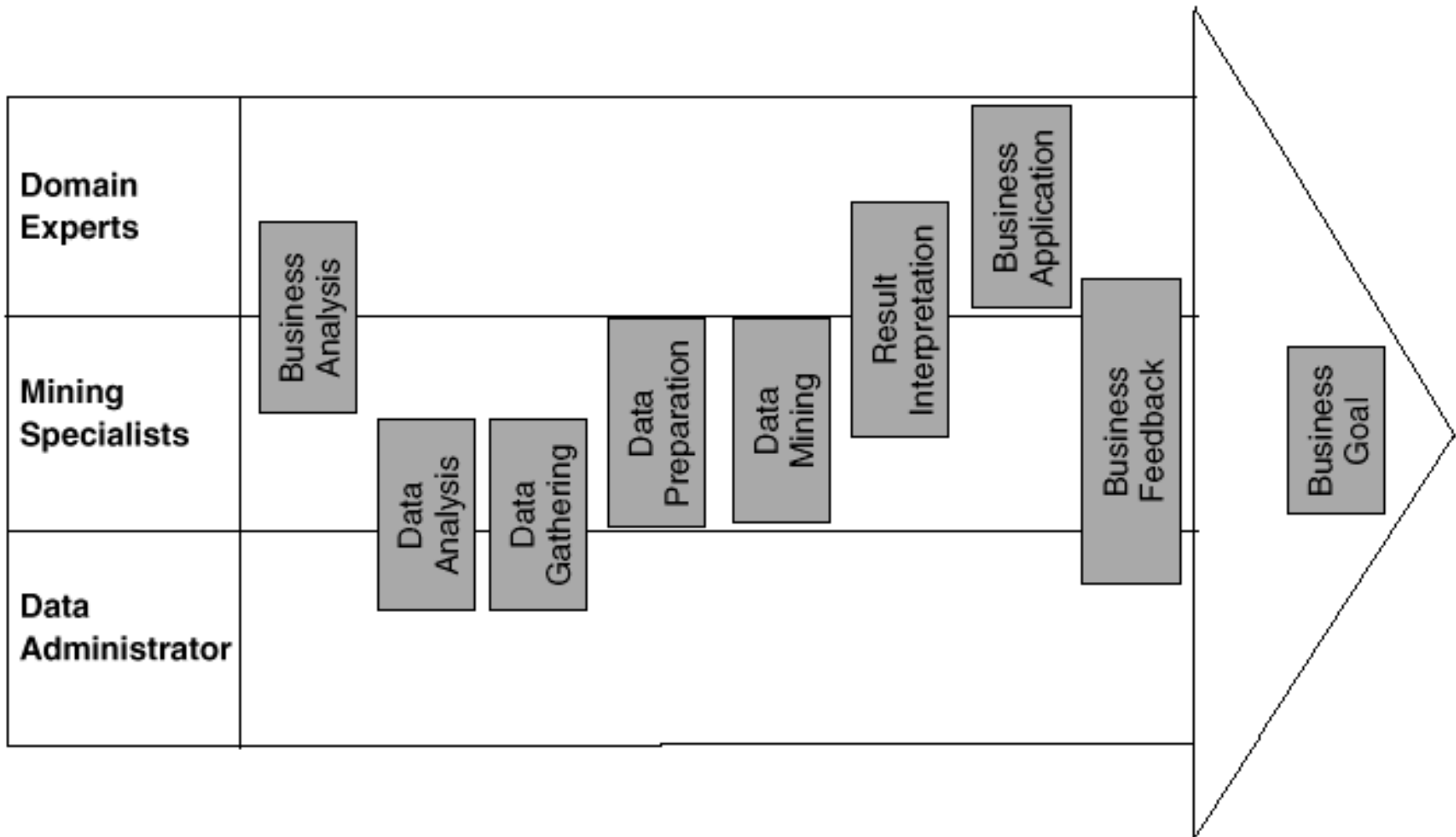
Results



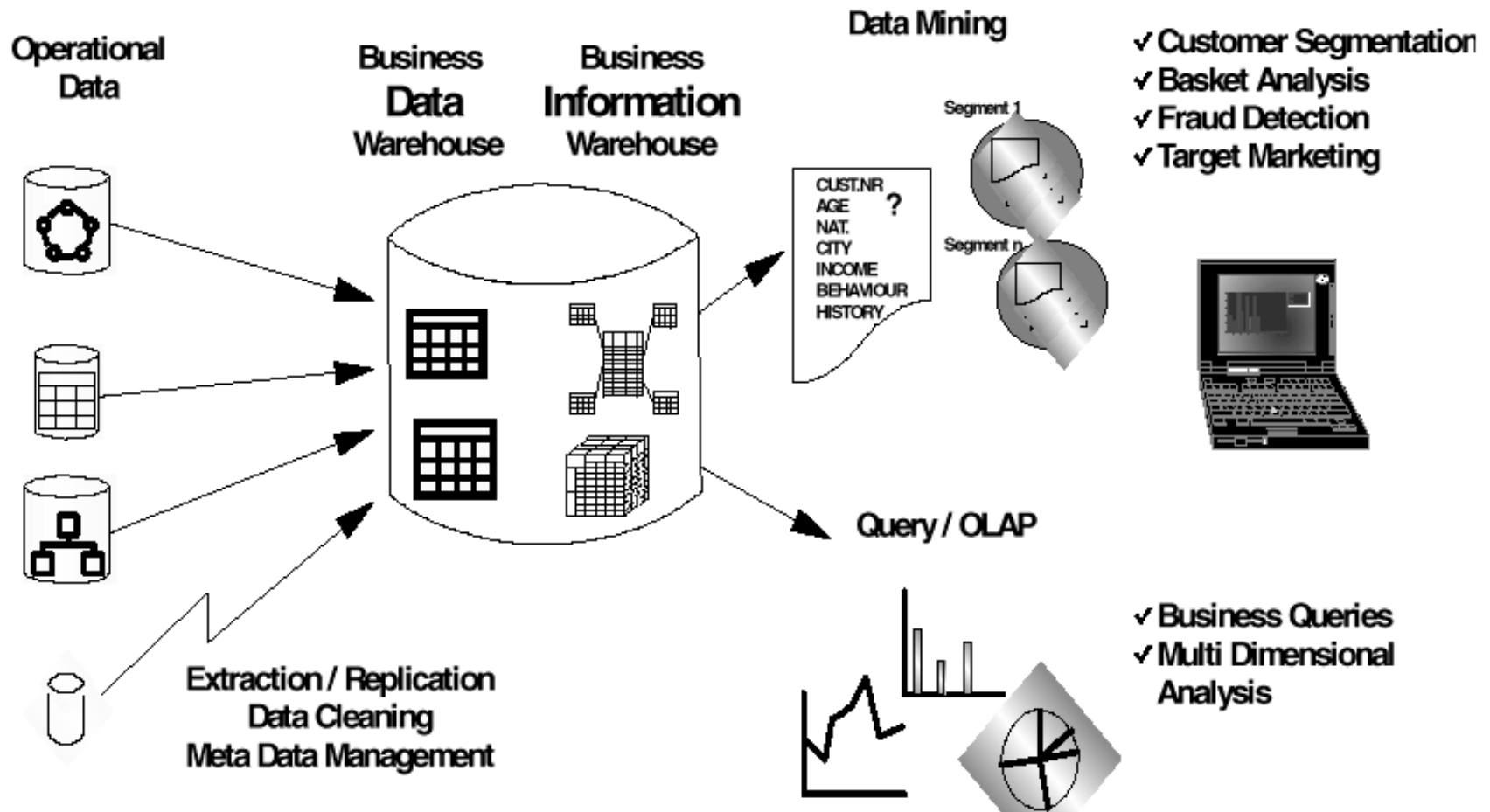
Data mining and business intelligence



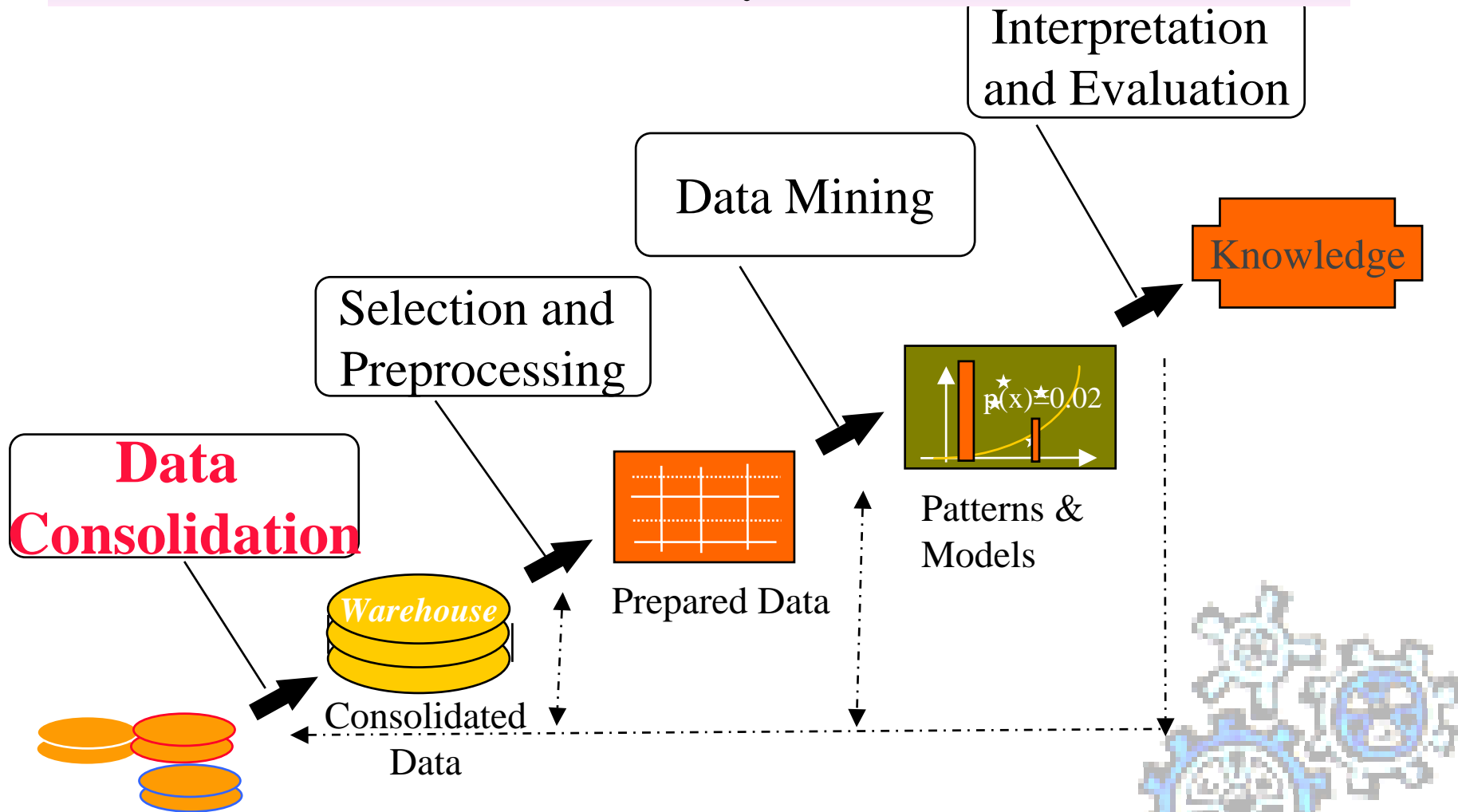
Roles in the KDD process



A business intelligence environment




The KDD process

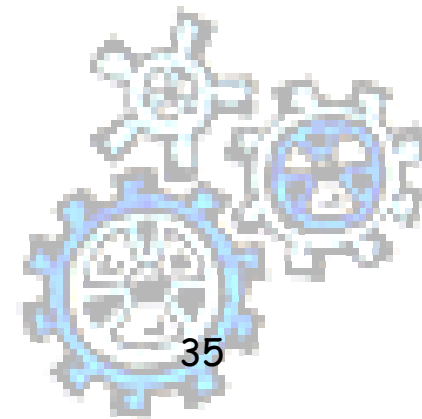


Data consolidation and preparation

Garbage in

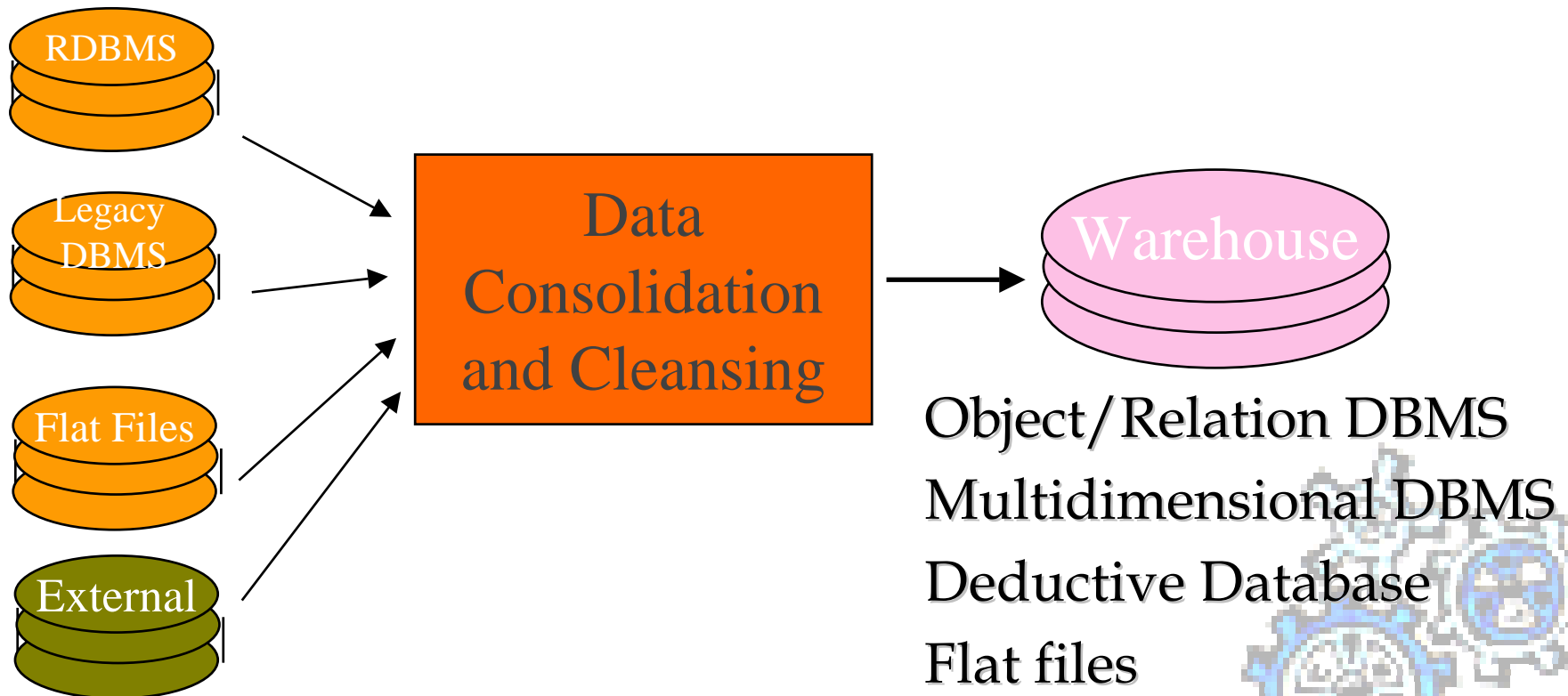
Garbage out

- ⌘ The quality of results  relates directly to quality of the data
- ⌘ 50%-70% of KDD process effort is spent on data consolidation and preparation
- ⌘ Major justification for a corporate data warehouse



Data consolidation

From data sources to consolidated data repository

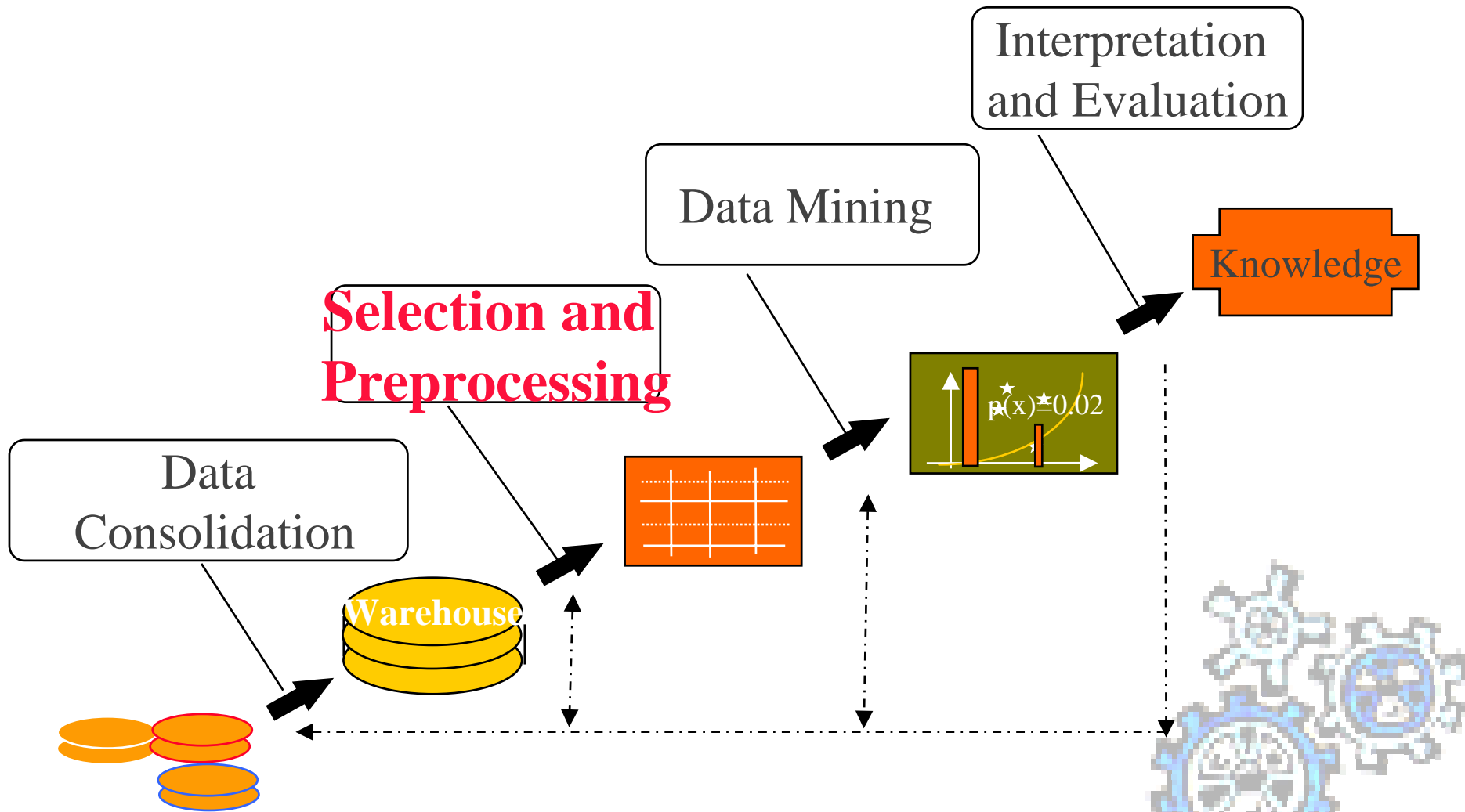


Data consolidation

- ⌘ Determine preliminary list of attributes
- ⌘ Consolidate data into working database
 - ☒ Internal and External sources
- ⌘ Eliminate or estimate missing values
- ⌘ Remove *outliers* (obvious exceptions)
- ⌘ Determine prior probabilities of categories and deal with *volume bias*

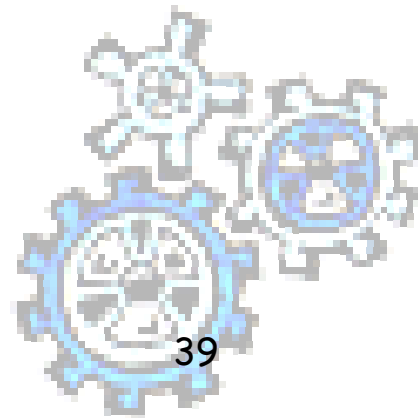


The KDD process

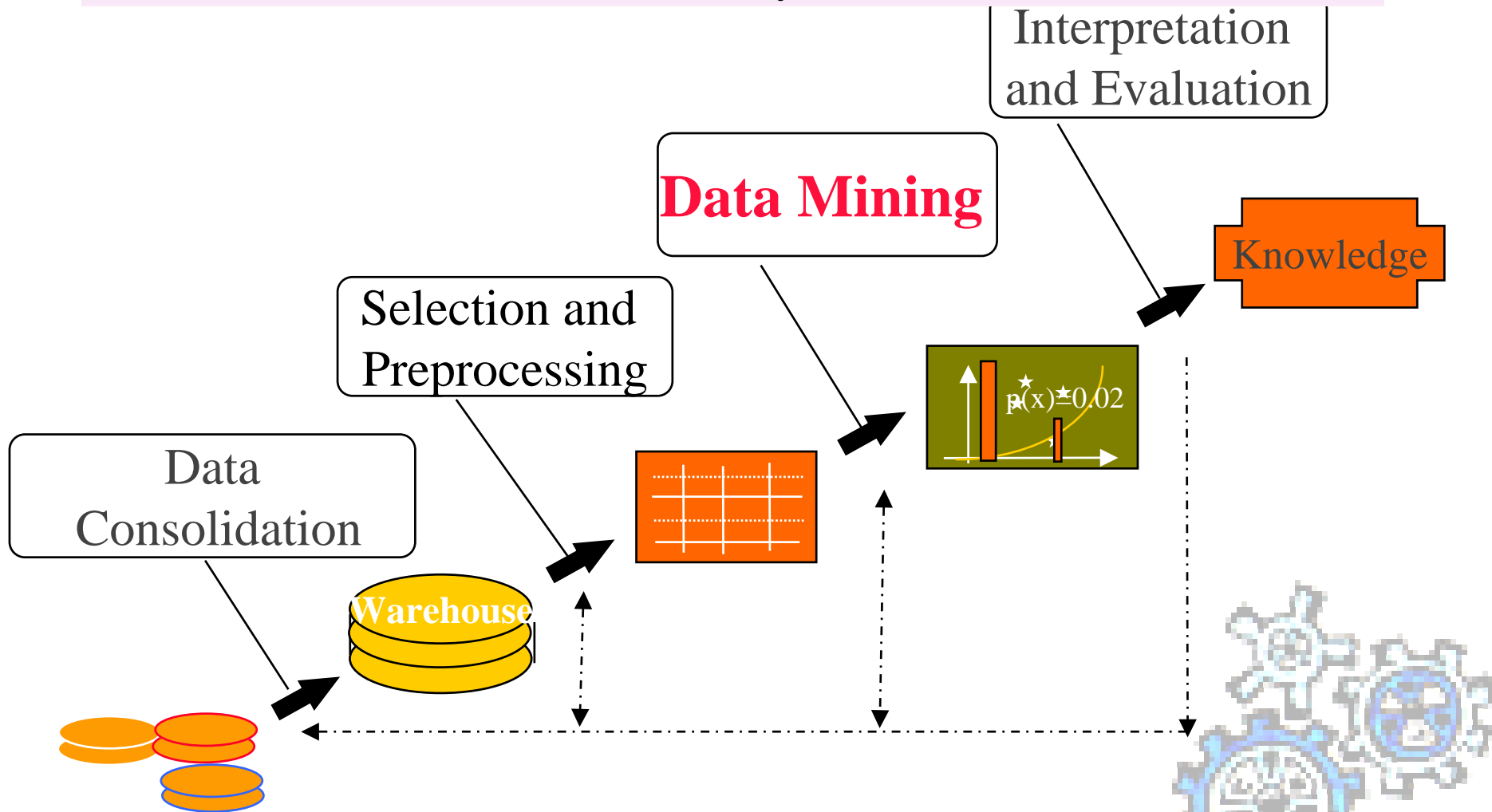


Data selection and preprocessing

- ⌘ Generate a set of examples
 - ⊞ choose sampling method
 - ⊞ consider sample complexity
 - ⊞ deal with volume bias issues
- ⌘ Reduce attribute dimensionality
 - ⊞ remove redundant and/or correlating attributes
 - ⊞ combine attributes (sum, multiply, difference)
- ⌘ Reduce attribute value ranges
 - ⊞ group symbolic discrete values
 - ⊞ quantify continuous numeric values
- ⌘ Transform data
 - ⊞ de-correlate and normalize values
 - ⊞ map time-series data to static representation
- ⌘ **OLAP and visualization tools play key role**



The KDD process



Data mining tasks and methods

⌘ Directed Knowledge Discovery

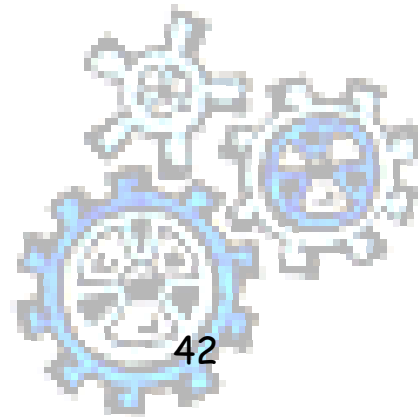
- ☑ Purpose: Explain value of some field in terms of all the others (goal-oriented)
- ☑ Method: select the target field based on some hypothesis about the data; ask the algorithm to tell us how to predict or classify new instances
- ☑ Examples:
 - ☑ what products show increased sale when cream cheese is discounted
 - ☑ which banner ad to use on a web page for a given user coming to the site



Data mining tasks and methods

⌘ Undirected Knowledge Discovery (Explorative Methods)

- ⊞ Purpose: Find patterns in the data that may be interesting (no target specified)
- ⊞ Method: clustering, association rules (affinity grouping)
- ⊞ Examples:
 - ⊞ which products in the catalog often sell together
 - ⊞ market segmentation (groups of customers/users with similar characteristics)



Data Mining Models

⌘ Automated Exploration/Discovery

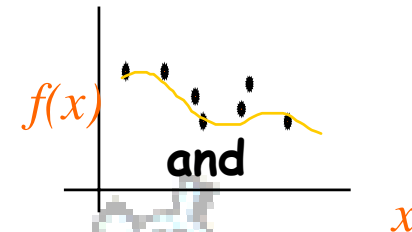
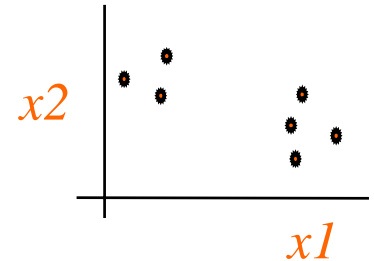
- ⊞ *e.g.*.. **discovering new market segments**
- ⊞ clustering analysis

⌘ Prediction/Classification

- ⊞ *e.g.*.. **forecasting gross sales given current factors**
- ⊞ regression, neural networks, genetic algorithms, decision trees

⌘ Explanation/Description

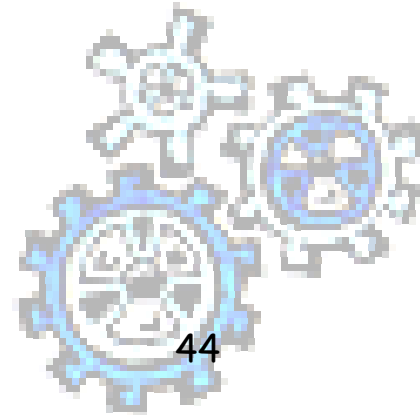
- ⊞ *e.g.*.. **characterizing customers by demographics**
purchase history
- ⊞ decision trees, association rules



if age > 35
and income < \$35k
then ...

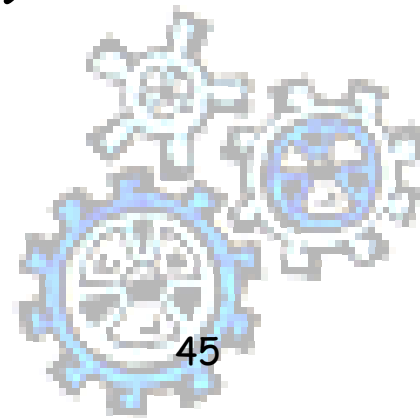
Automated exploration and discovery

- ⌘ **Clustering:** partitioning a set of data into a set of classes, called *clusters*, whose members share some interesting common properties.
- ⌘ **Distance-based numerical clustering**
 - ⊞ metric grouping of examples (K-NN)
 - ⊞ graphical visualization can be used
- ⌘ **Bayesian clustering**
 - ⊞ search for the number of classes which result in best fit of a probability distribution to the data
 - ⊞ AutoClass (NASA) one of best examples



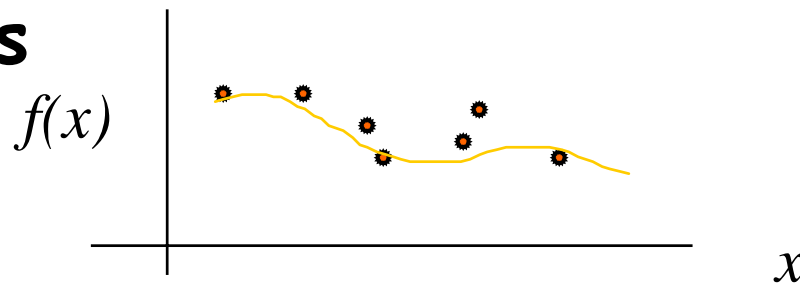
Prediction and classification

- ⌘ **Learning** a predictive model
- ⌘ **Classification** of a new case/sample
- ⌘ **Many methods:**
 - ⊞ Artificial neural networks
 - ⊞ Inductive decision tree and rule systems
 - ⊞ Genetic algorithms
 - ⊞ Nearest neighbor clustering algorithms
 - ⊞ Statistical (parametric, and non-parametric)



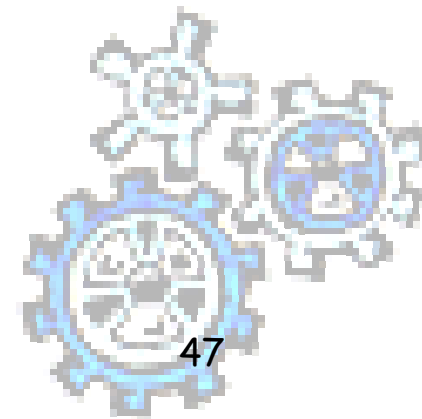
Generalization and regression

- ⌘ The objective of learning is to achieve good *generalization* to new unseen cases.
- ⌘ Generalization can be defined as a mathematical *interpolation* or *regression* over a set of training points
- ⌘ Models can be validated with a previously unseen test set or using cross-validation methods



Classification and prediction

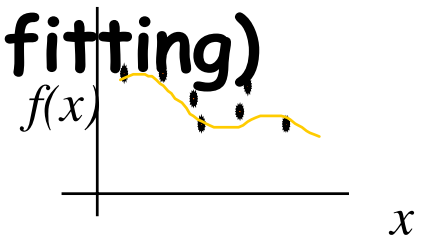
- ☒ Classify data based on the values of a *target* attribute, e.g., classify countries based on climate, or classify cars based on gas mileage.
- ☒ Use obtained model to predict some unknown or missing attribute values based on other information.



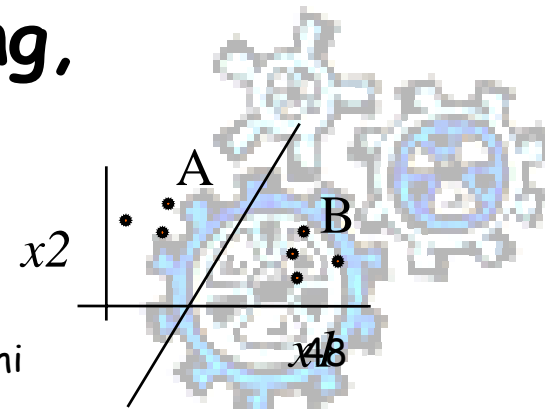
inductive modeling = learning

Objective: *Develop a general model or hypothesis from specific examples*

⌘ **Function approximation (curve fitting)**

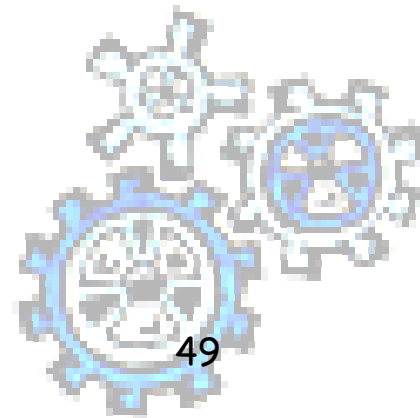


⌘ **Classification (concept learning, pattern recognition)**



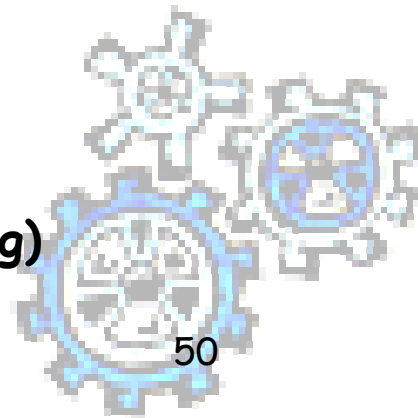
Explanation and description (MOBASHER RULES)

- ⌘ Learn a generalized hypothesis (model) from selected data
- ⌘ Description/Interpretation of model provides new knowledge
- ⌘ Affinity Grouping
- ⌘ Methods:
 - ⌘ Inductive decision tree and rule systems
 - ⌘ Association rule systems
 - ⌘ Link Analysis
 - ⌘ ...



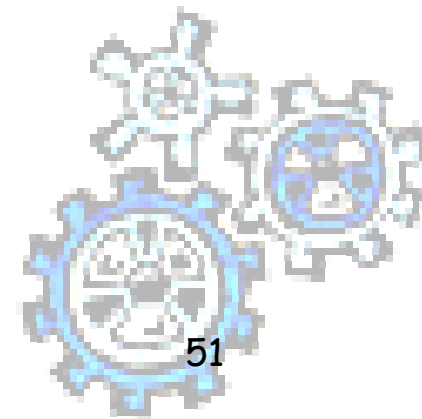
Affinity Grouping

- ⌘ Determine what items often go together (usually in transactional databases)
- ⌘ Often Referred to as Market Basket Analysis
 - ⊞ used in retail for planning arrangement on shelves
 - ⊞ used for identifying cross-selling opportunities
 - ⊞ “should” be used to determine best link structure for a Web site
- ⌘ Examples
 - ⊞ people who buy milk and beer also tend to buy diapers
 - ⊞ people who access pages A and B are likely to place an online order
- ⌘ Suitable data mining tools
 - ⊞ association rule discovery
 - ⊞ clustering
 - ⊞ Nearest Neighbor analysis (memory-based reasoning)



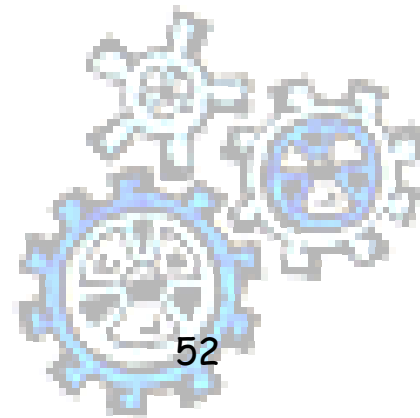
Exception/deviation detection

- ⌘ Generate a model of normal activity
- ⌘ Deviation from model causes alert
- ⌘ Methods:
 - ⊞ Artificial neural networks
 - ⊞ Inductive decision tree and rule systems
 - ⊞ Statistical methods
 - ⊞ Visualization tools

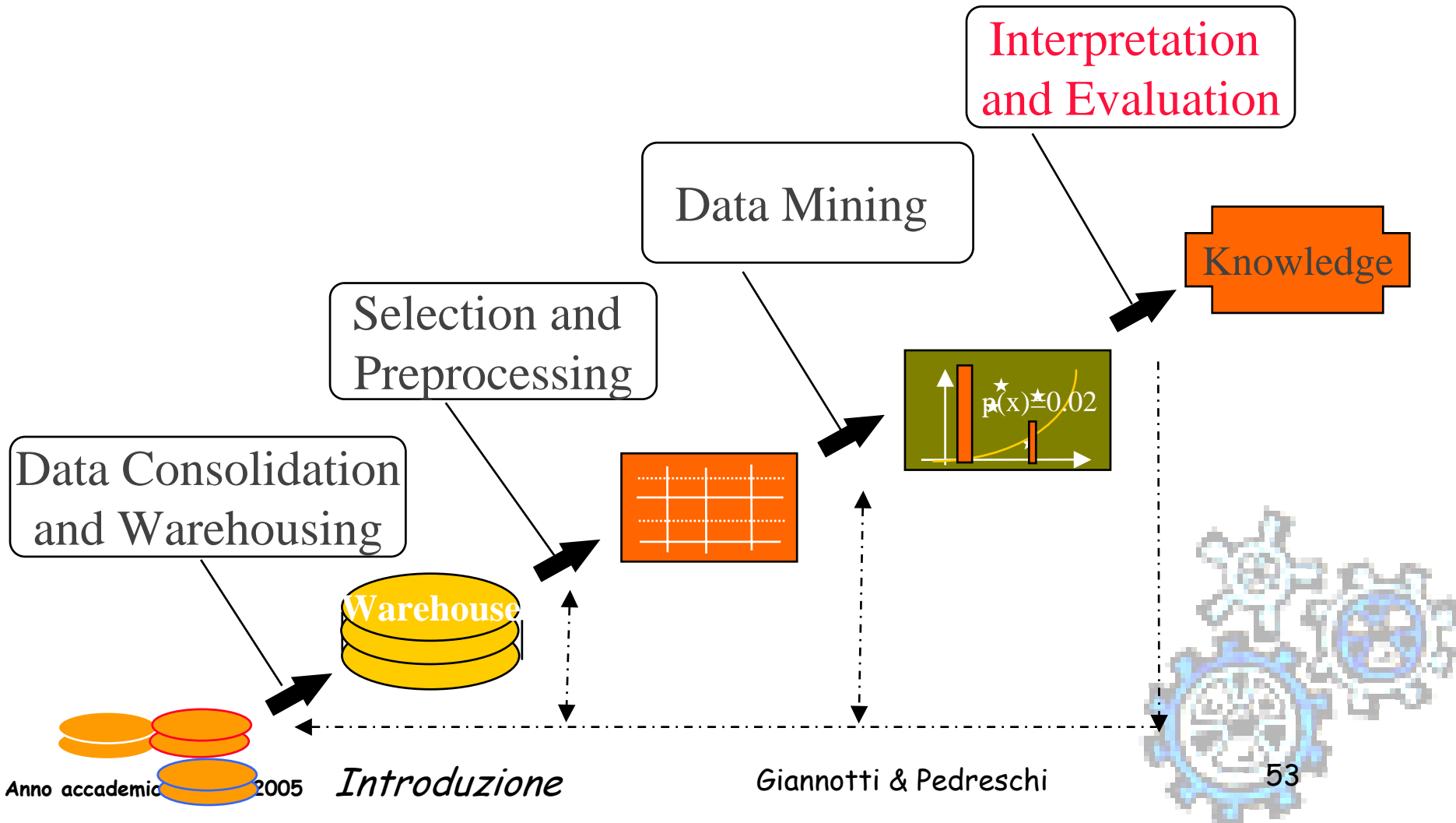


Outlier and exception data analysis

- ⌘ Time-series analysis (trend and deviation):
 - ☒ Trend and deviation analysis: regression, sequential pattern, similar sequences, trend and deviation, e.g., stock analysis.
 - ☒ Similarity-based pattern-directed analysis
 - ☒ Full vs. partial periodicity analysis
- ⌘ Other pattern-directed or statistical analysis



The KDD process



Are all the discovered pattern interesting?

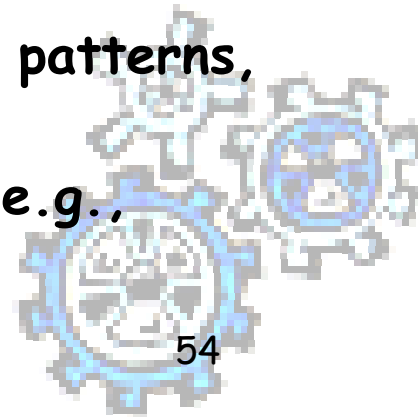
⌘ A data mining system/query may generate thousands of patterns, not all of them are interesting.

⌘ Interestingness measures:

- ⊞ easily understood by humans
- ⊞ valid on new or test data with some degree of certainty.
- ⊞ potentially useful
- ⊞ novel, or validates some hypothesis that a user seeks to confirm

⌘ Objective vs. subjective interestingness measures

- ⊞ Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
- ⊞ Subjective: based on user's beliefs in the data, e.g., unexpectedness, novelty, etc.



Interpretation and evaluation

Evaluation

- ⌘ Statistical validation and significance testing
- ⌘ Qualitative review by experts in the field
- ⌘ Pilot surveys to evaluate model accuracy

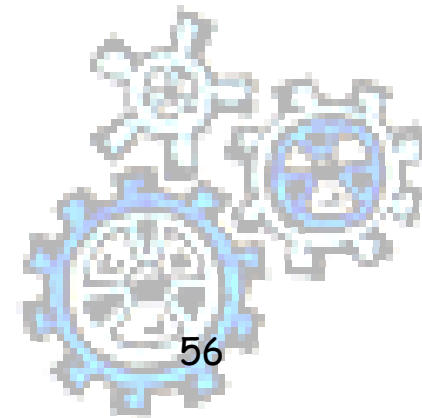
Interpretation

- ⌘ Inductive tree and rule models can be read directly
- ⌘ Clustering results can be graphed and tabled
- ⌘ Code can be automatically generated by some systems (IDTs, Regression models)



Seminar 1 - Bibliography

- ⌘ Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000
- ⌘ Micheael, J. A. Berry, Gordon S. Linoff, Mastering Data Mining, Wiley, 2000
- ⌘ Klosgen, Zytkow, Handbook of Data Mining, Oxford, 2001



Seminar 1 - Bibliography

Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000
http://www.mkp.com/books_catalog/catalog.asp?ISBN=1-55860-489-8

- ⑩ U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (editors). Advances in Knowledge discovery and data mining, MIT Press, 1996.
- ⑩ David J. Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, MIT Press, 2001.
- ⑩ S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann, ISBN 1-55860-754-4, 2002



Examples of DM projects

Competitive Intelligence
Fraud Detection,
Health care,
Traffic Accident Analysis,
Moviegoers database: a simple
example at work

L'Oreal, a case-study on competitive intelligence:

Source: DM@CINECA

<http://open.cineca.it/datamining/dmCineca/>

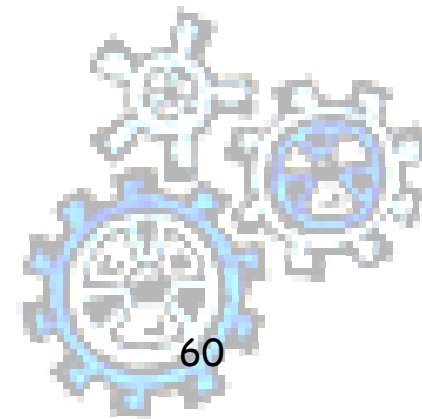
A small example

⌘ Domain: **technology watch** - a.k.a. competitive intelligence

- ☑ Which are the emergent technologies?
- ☑ Which competitors are investing on them?
- ☑ In which area are my competitors active?
- ☑ Which area will my competitor drop in the near future?

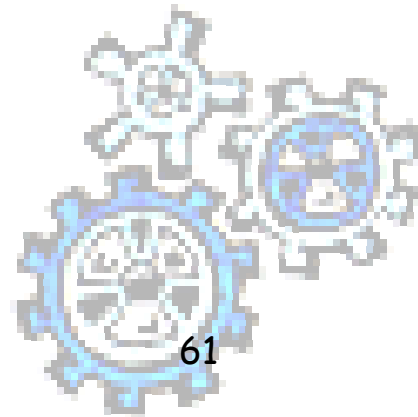
⌘ Source of data:

- ☑ public (on-line) databases



The Derwent database

- ⌘ Contains all **patents** filed worldwide in last 10 years
- ⌘ Searching this database by keywords may yield thousands of documents
- ⌘ Derwent documents are semi-structured: many long text fields
- ⌘ **Goal**: analyze Derwent documents to build a model of competitors' strategy



Structure of Derwent documents

Raccolta dei Documenti

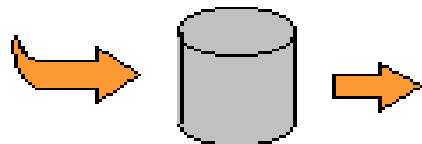
esempio di documento brevettuale



1/3881 - (C) Derwent Info 1994

AN: 94-364398 [45]

TI: Television with function for enlarging picture by variation of deflection frequency - has microprocessor for controlling system synchronous signal output, horizontal and vertical frequency drive circuit, sync. signal counter, signal detector.



DC: W03

PA: (GLDS) GOLDSTAR CO LTD

IN: O.KEITH

NP: 1

PR: 88KR-011143 880831

IC: H04N-005/262; C08J-005/18; G11B-005/704

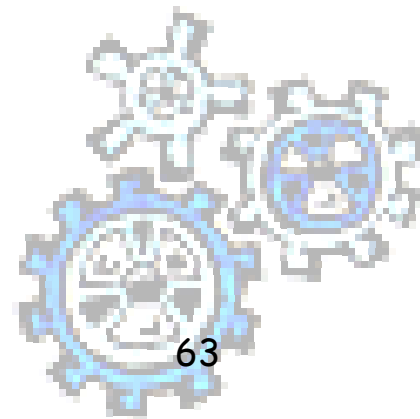
PN: KR940043 B1 940120 DW9445

AB: abstract

Example dataset

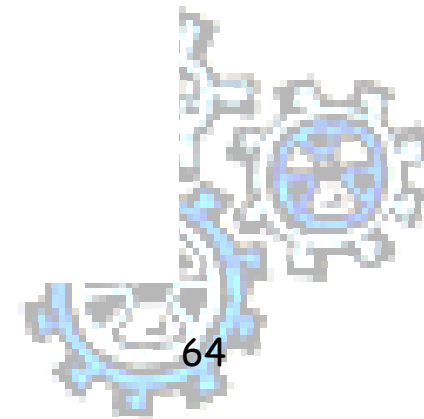
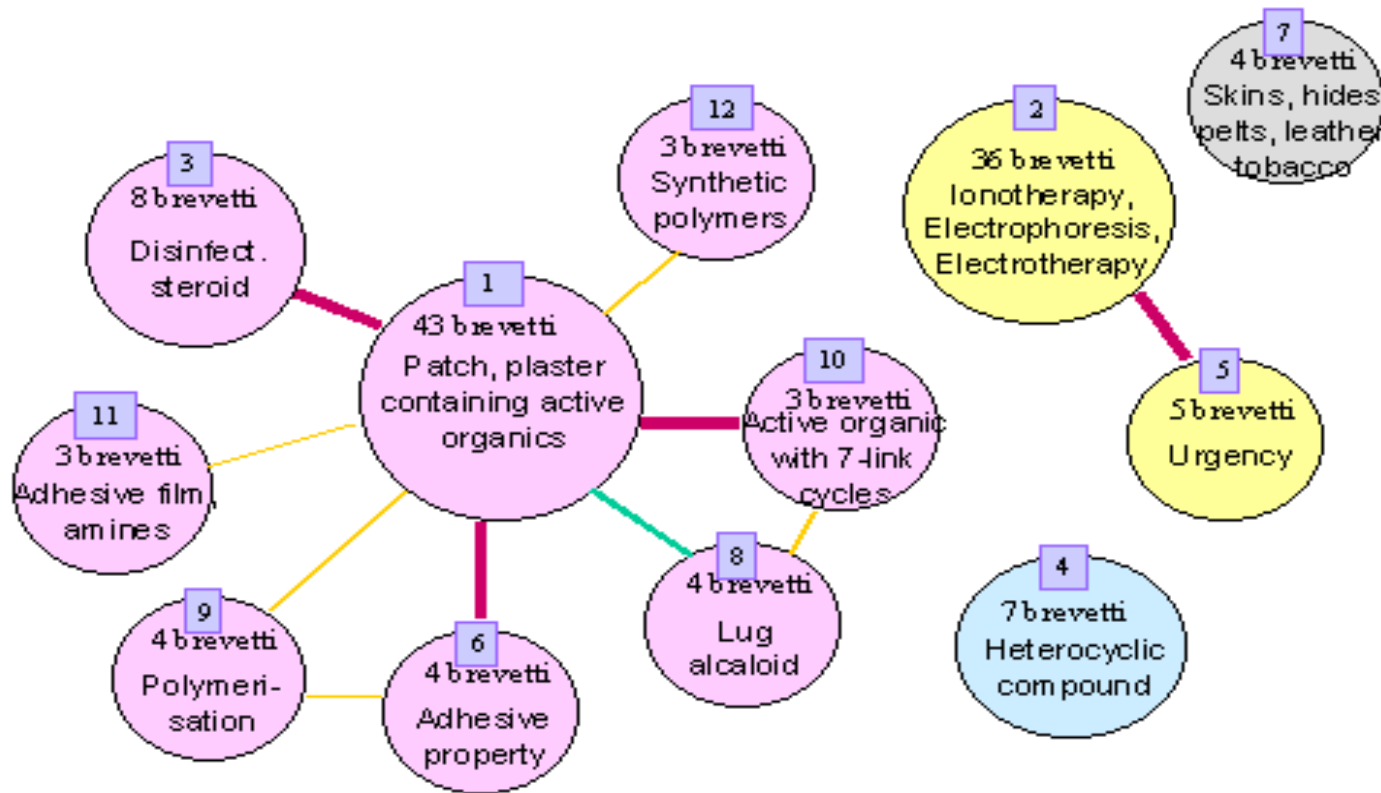
⌘ Patents in the area: patch technology (cerotto medicale)

- ⊞ 105 companies from 12 countries
- ⊞ 94 classification codes
- ⊞ 52 Derwent codes



Clustering output

Patch technology- *mappa dei clusters*



Zoom on cluster 2

Patch technology- *descrizione del cluster n.2*

Classificazione Internazionale:

A61N-001/30 Electrotherapy; Appliances of electrical power by contact electrodes; Ionotherapy or electrophoresis devices
A61M-037/00 Therapeutic patch

Classificazione Derwent:

S05 Electromedical
P34 Health, Electrotherapy

Società proprietarie:



DRUG DELIVERY SYST 42%



BASF AG 36%



KOREA RES INST CHEM 16%



MEDTRONIC INC 6%

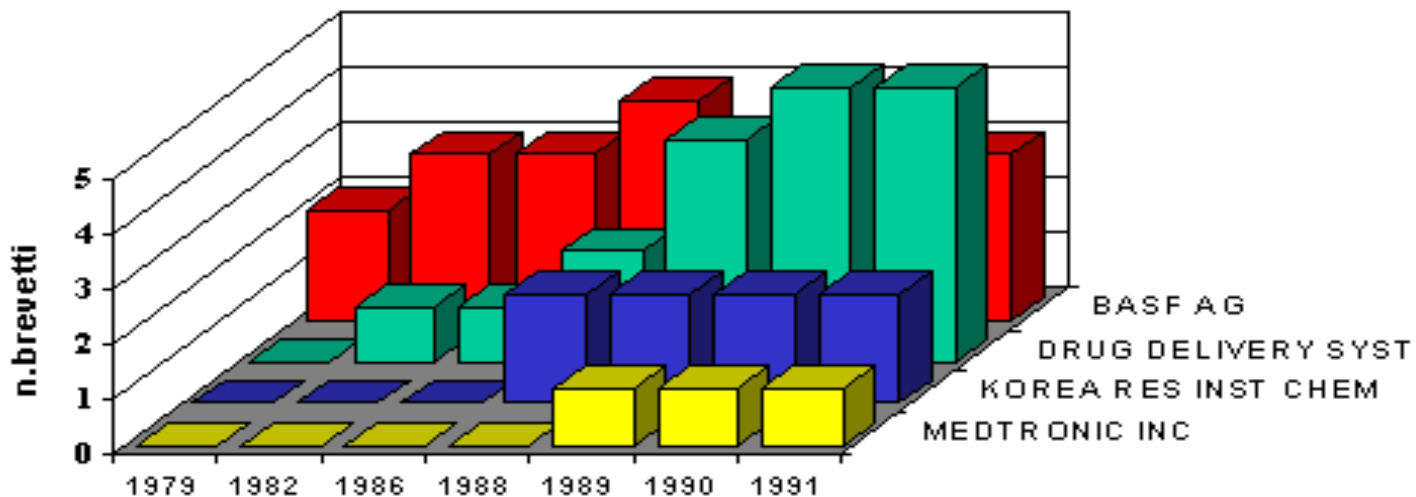
Anno n. brevetti

1979	2
1982	4
1986	4
1988	8
1989	10
1990	11
1991	11

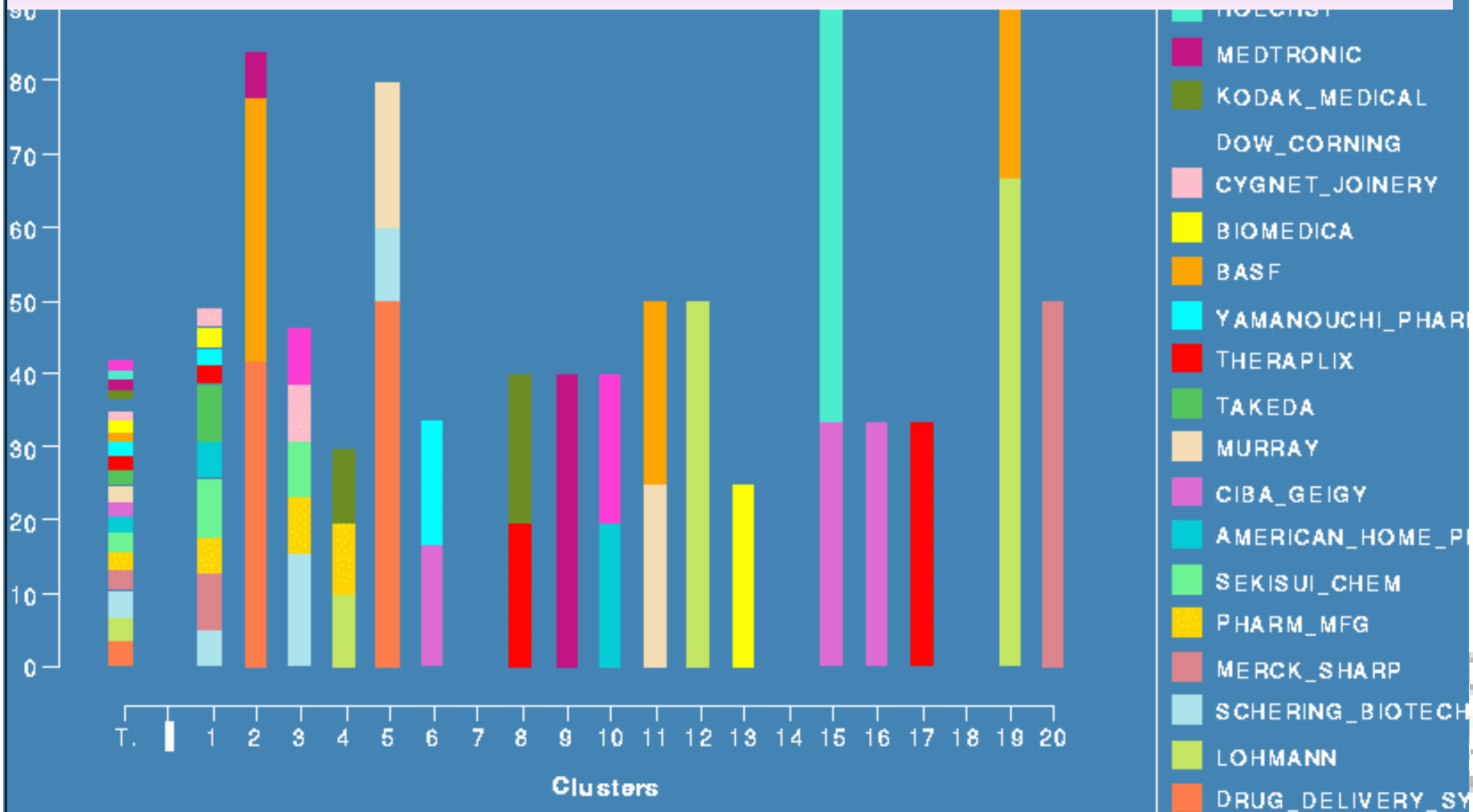


Zoom on cluster 2 - profiling competitors

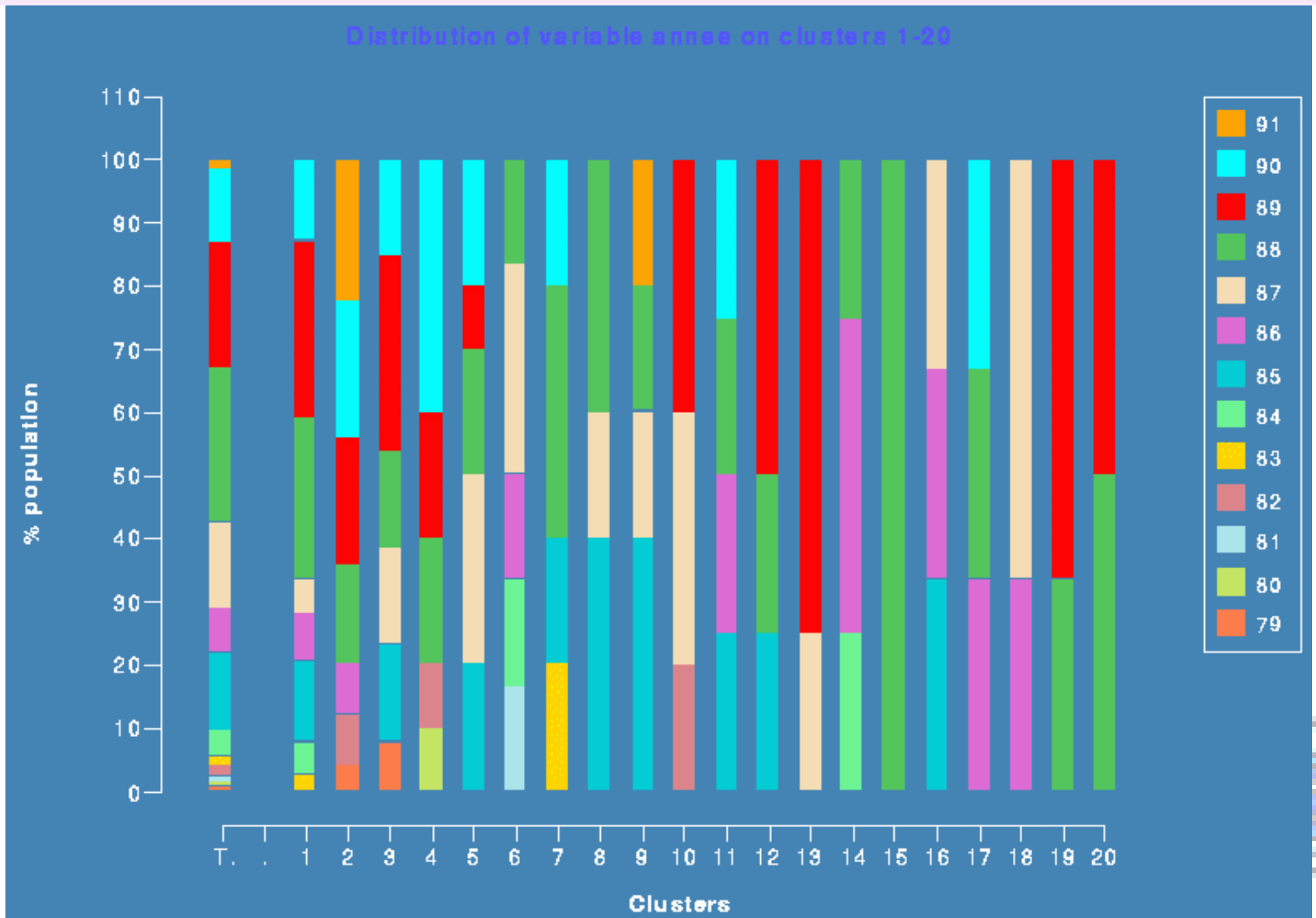
Patch technology- cluster n.2 -
attività della concorrenza nel tempo



Activity of competitors in the clusters



Temporal analysis of clusters

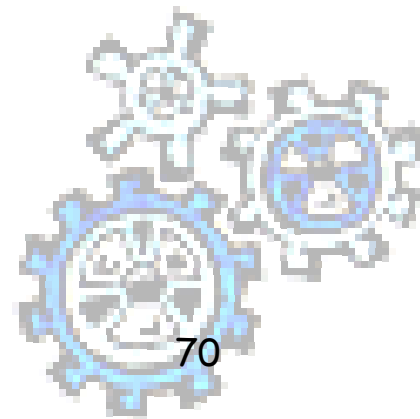


Fraud detection and audit planning

Source: Ministero delle Finanze
Progetto Sogei, KDD Lab. Pisa

Fraud detection

- ⌘ A major task in fraud detection is constructing *models* of fraudulent behavior, for:
 - ☒ preventing future frauds (*on-line* fraud detection)
 - ☒ **discovering past frauds** (*a posteriori* fraud detection)
- ⌘ **analyze historical audit data to plan effective future audits**

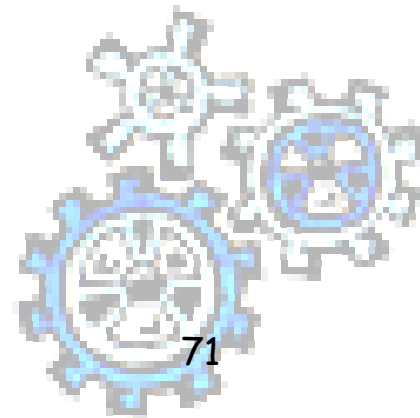


Audit planning

⌘ Need to face a trade-off between conflicting issues:

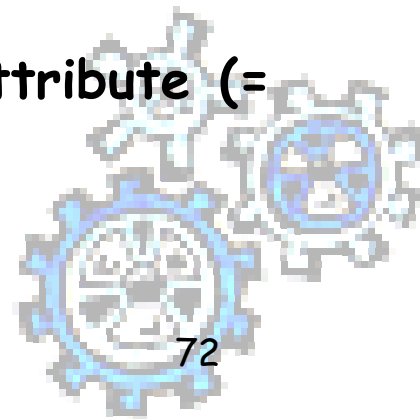
☑ *maximize audit benefits*: select subjects to be audited to maximize the recovery of evaded tax

☑ *minimize audit costs*: select subjects to be audited to minimize the resources needed to carry out the audits.

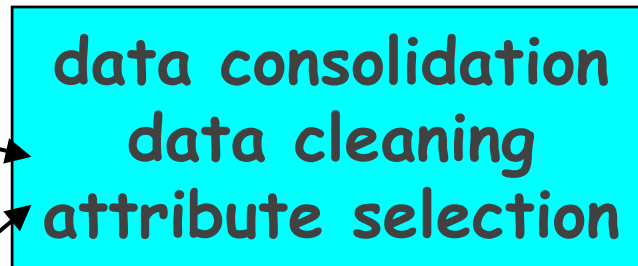
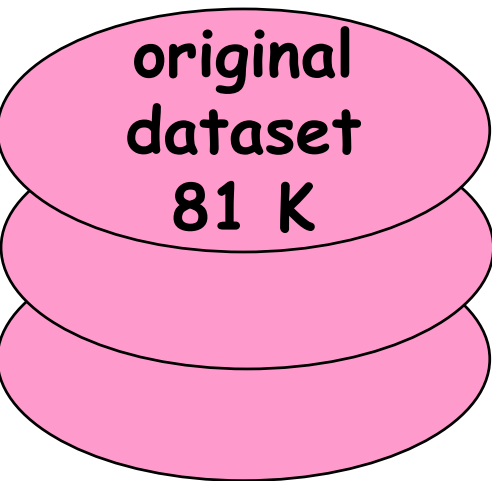


Available data sources

- ⌘ Dataset: **tax declarations**, concerning a targeted class of Italian **companies**, integrated with other sources:
 - ☑ social benefits to employees, official budget documents, electricity and telephone bills.
- ⌘ Size: **80 K** tuples, 175 numeric attributes.
- ⌘ A subset of **4 K** tuples corresponds to the **audited** companies:
 - ☑ outcome of audits recorded as the **recovery** attribute (= **amount of evaded tax ascertained**)



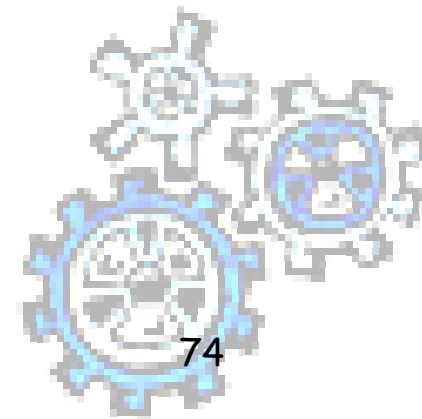
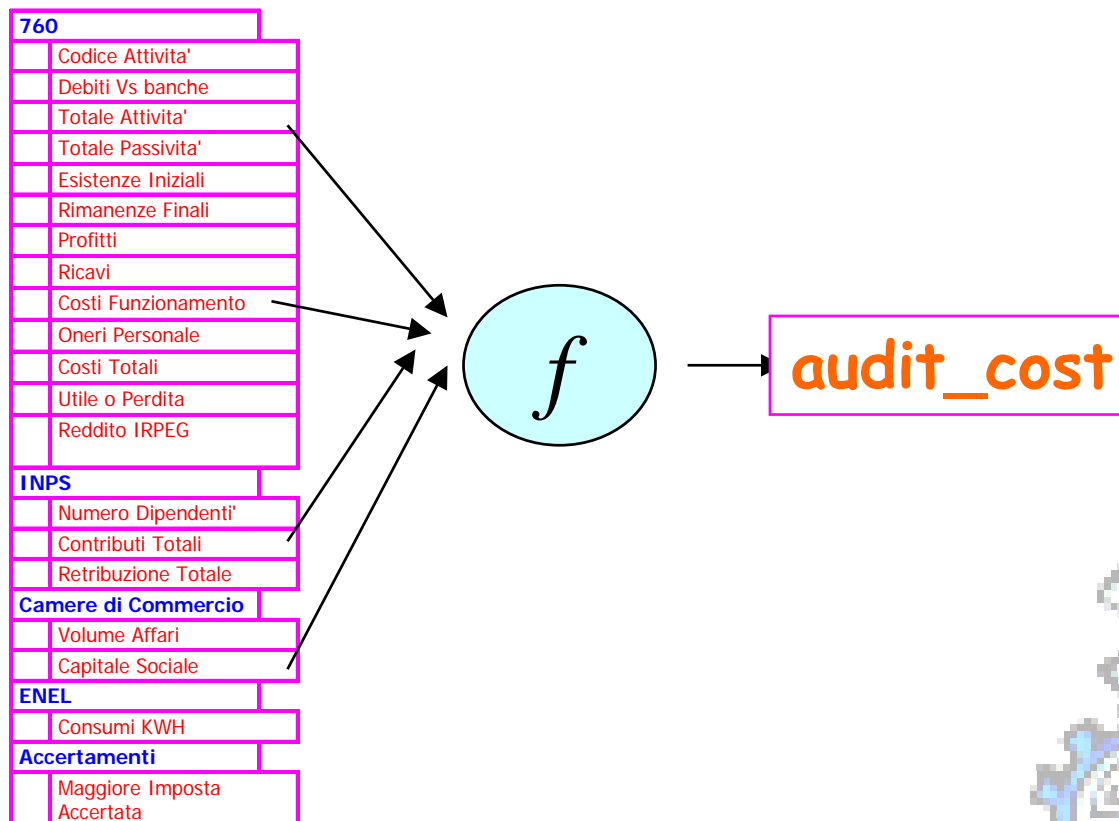
Data preparation



TAX DECLARATION	
	Codice Attivita'
	Debiti Vs banche
	Totale Attivita'
	Totale Passivita'
	Esistenze Iniziali
	Rimanenze Finali
	Profitti
	Ricavi
	Costi Funzionamento
	Oneri Personale
	Costi Totali
	Utile o Perdita
	Reddito IRPEG
SOCIAL BENEFITS	
	Numero Dipendenti'
	Contributi Totali
	Retribuzione Totale
OFFICIAL BUDGET	
	Volume Affari
	Capitale Sociale
ELECTRICITY BILLS	
	Consumi KWH
AUDIT	
	Recovery

Cost model

⌘ A derived attribute **audit_cost** is defined as a function of other attributes



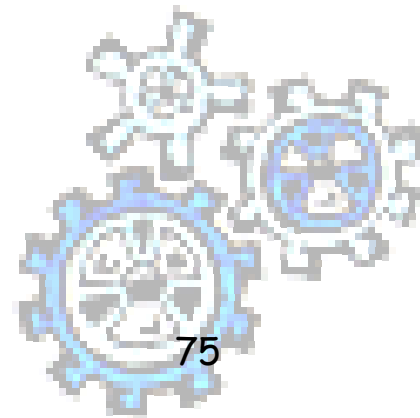
Cost model and the target variable

⌘ recovery of an audit after the audit cost
 $\text{actual_recovery} = \text{recovery} - \text{audit_cost}$

⌘ target variable (class label) of our analysis is set as the **Class of Actual Recovery (c.a.r.)**:

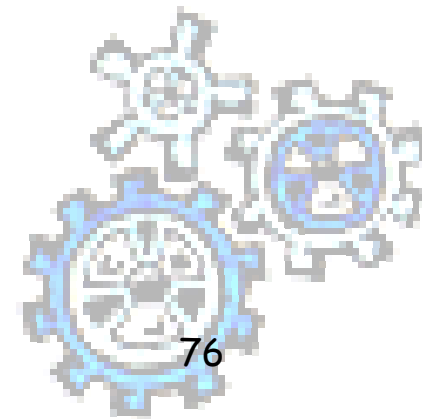
⌘ $c.a.r. =$

<i>negative</i>	<i>if</i>	$\text{actual_recovery} \leq 0$
<i>positive</i>	<i>if</i>	$\text{actual_recovery} > 0.$



Quality assessment indicators

- ⌘ The obtained classifiers are evaluated according to several **indicators**, or metrics
- ⌘ **Domain-independent** indicators
 - ☒ confusion matrix
 - ☒ misclassification rate
- ⌘ **Domain-dependent** indicators
 - ☒ audit #
 - ☒ actual recovery
 - ☒ profitability
 - ☒ relevance



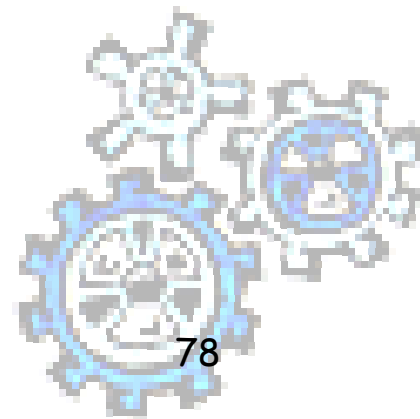
Domain-dependent quality indicators

- ⌘ **audit #** (of a given classifier): number of tuples classified as positive =
 $\# (FP \cup TP)$
- ⌘ **actual recovery**: total amount of actual recovery for all tuples classified as positive
- ⌘ **profitability**: average actual recovery per audit
- ⌘ **relevance**: ratio between profitability and misclassification rate



The REAL case

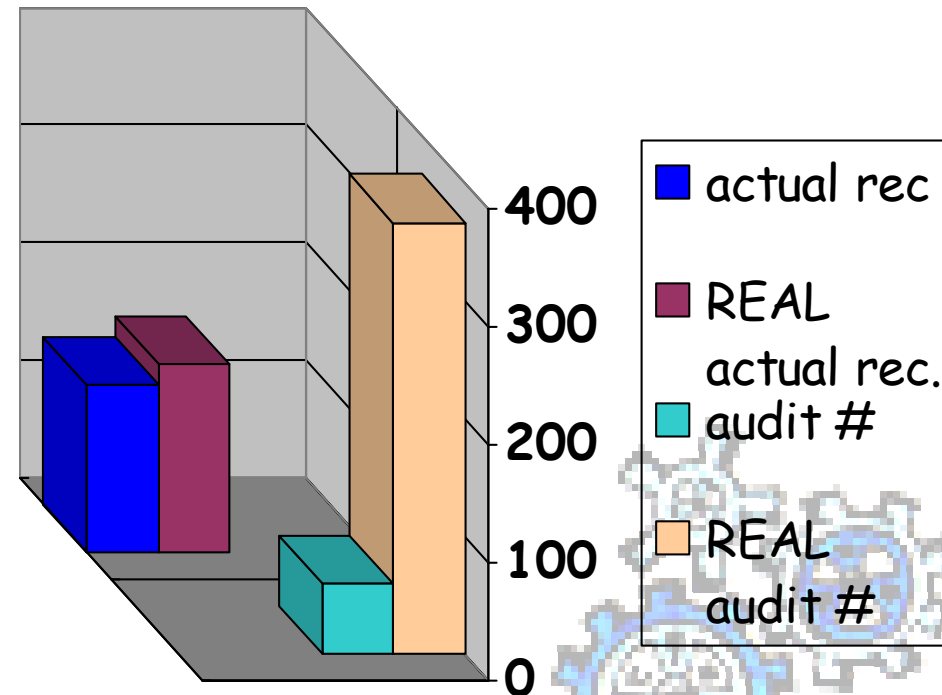
- ⌘ Classifiers can be compared with the REAL case, consisting of the whole test-set:
- ⌘ audit # (REAL) = 366
- ⌘ actual recovery(REAL) = 159.6 M euro



Model evaluation: classifier 1 (min FP)

- no replication in training-set (unbalance towards negative)
- 10-trees adaptive boosting

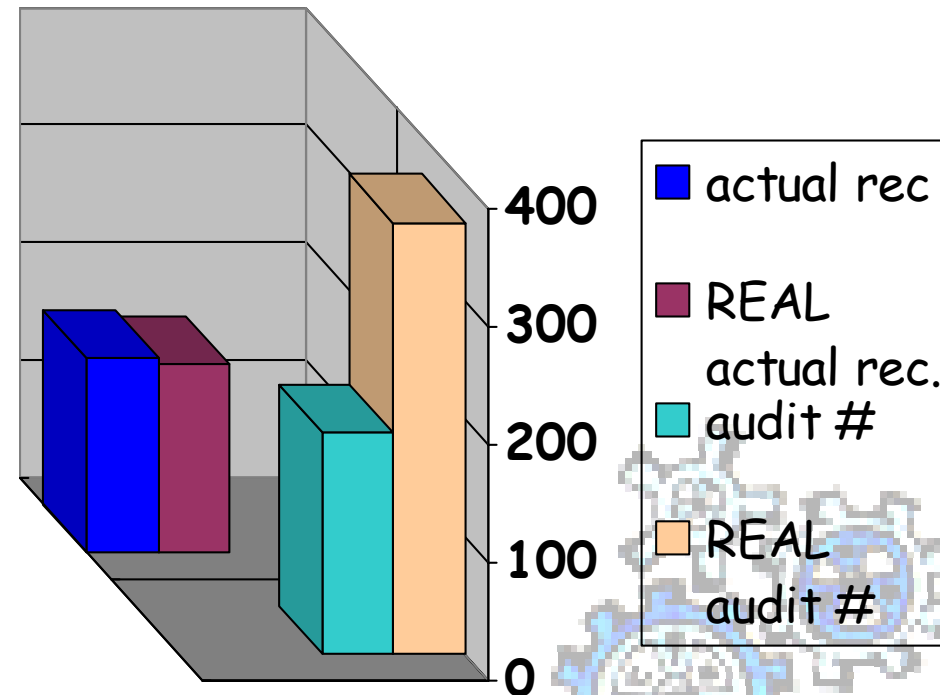
- misc. rate* = 22%
- audit #* = 59 (11 FP)
- actual rec.* = 141.7 Meuro
- profitability* = 2.401



Model evaluation: classifier 2 (min FN)

- ⊞ replication in training-set (balanced neg/pos)
- ⊞ misc. weights (trade 3 FP for 1 FN)
- ⊞ 3-trees adaptive boosting

- ⊞ *misc. rate* = 34%
- ⊞ *audit #* = 188 (98 FN)
- ⊞ *actual rec.* = 165.2 Meuro
- ⊞ *profitability* = 0.878

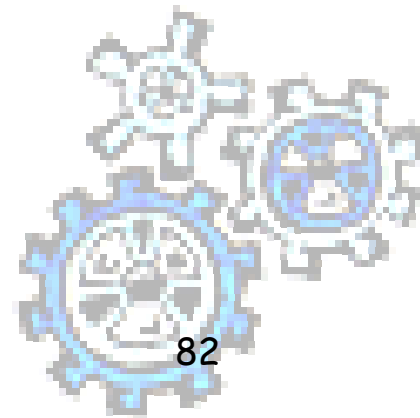


Atherosclerosis prevention study

2nd Department of Medicine, 1st Faculty of
Medicine of Charles University and Charles
University Hospital, U nemocnice 2, Prague
2 (head. Prof. M. Aschermann, MD, SDr,
FESC)

Atherosclerosis prevention study:

- ⌘ The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.
- ⌘ Used for Discovery Challenge at PKDD 00-02-03-04



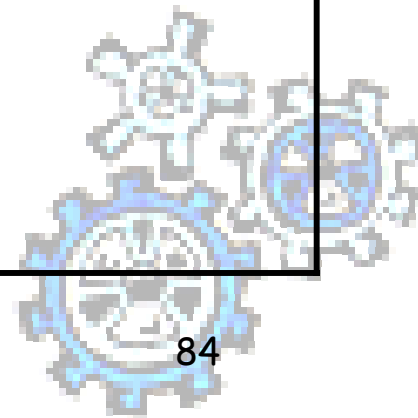
Atherosclerosis prevention study:

- ⌘ Study on 1400 middle-aged men at Czech hospitals
 - ☑ Measurements concern development of cardiovascular disease and other health data in a series of exams
- ⌘ The aim of this analysis is to look for associations between medical characteristics of patients and death causes.
- ⌘ Four tables
 - ☑ Entry and subsequent exams, questionnaire responses, deaths



The input data

Data from Entry and Exams		
General characteristics	Examinations	habits
Marital status	Chest pain	Alcohol
Transport to a job	Breathlessness	Liquors
Physical activity in a job	Cholesterol	Beer 10
Activity after a job	Urine	Beer 12
Education	Subscapular	Wine
Responsibility	Triceps	Smoking
Age		Former smoker
Weight		Duration of smoking
Height		Tea
		Sugar
		Coffee

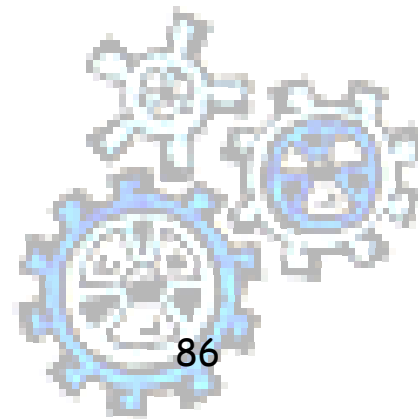
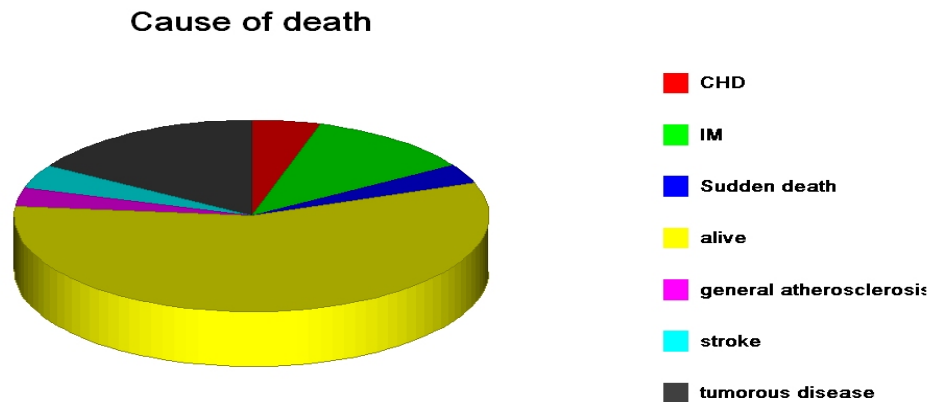


The input data

DEATH CAUSE	PATIENTS	%
myocardial infarction	80	20.6
coronary heart disease	33	8.5
stroke	30	7.7
other causes	79	20.3
sudden death	23	5.9
unknown	8	2.0
tumorous disease	114	29.3
general atherosclerosis	22	5.7
TOTAL	389	100.0

Data selection

- ⌘ When joining “Entry” and “Death” tables we implicitly create a new attribute “Cause of death”, which is set to “alive” for subjects present in the “Entry” table but not in the “Death” table.
- ⌘ We have only 389 subjects in death table.



The prepared data

Patient	General characteristics		Examinations		Habits		Cause of death
	Activity after work	Education	Chest pain	...	Alcohol	
1	moderate activity	university	not present		no		Stroke
2	great activity		not ischaemic		occasionally		myocardial infarction
3	he mainly sits		other pains		regularly		tumorous disease
.....	alive
389	he mainly sits		other pains		regularly		tumorous disease

Descriptive Analysis/ Subgroup Discovery / Association Rules

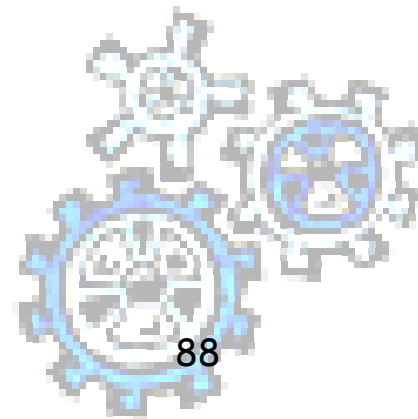
Are there strong relations concerning death cause?

General characteristics (?) \Rightarrow Death cause (?)

Examinations (?) \Rightarrow Death cause (?)

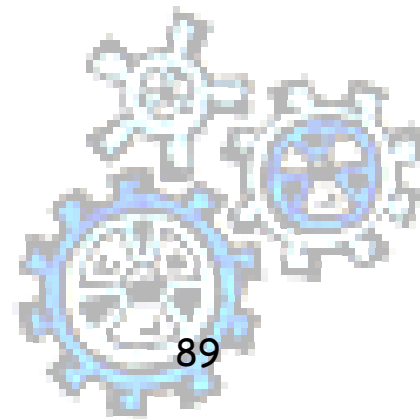
Habits (?) \Rightarrow Death cause (?)

Combinations (?) \Rightarrow Death cause (?)



Example of extracted rules

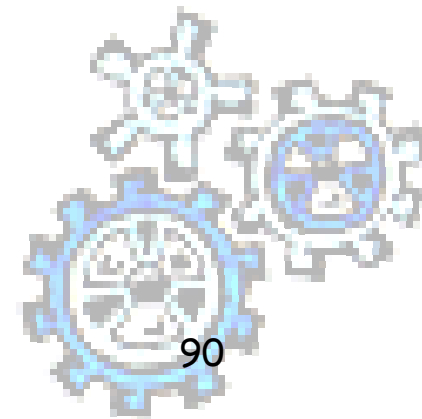
- ⌘ **Education(university) & Height<176-180>**
⇒ Death cause (tumouros disease), 16 ; 0.62
- ⌘ **It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.**



Example of extracted rules

⌘ **Physical activity in work(he mainly sits) & Height<176-180> \Rightarrow Death cause (tumouros disease), 24; 0.52**

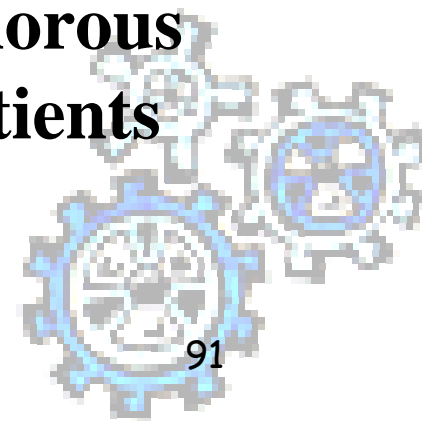
⌘ **It means that on tumorous disease have died 24 i.e. 52% of patients that mainly sit in the work and whose height is 176-180 cm.**



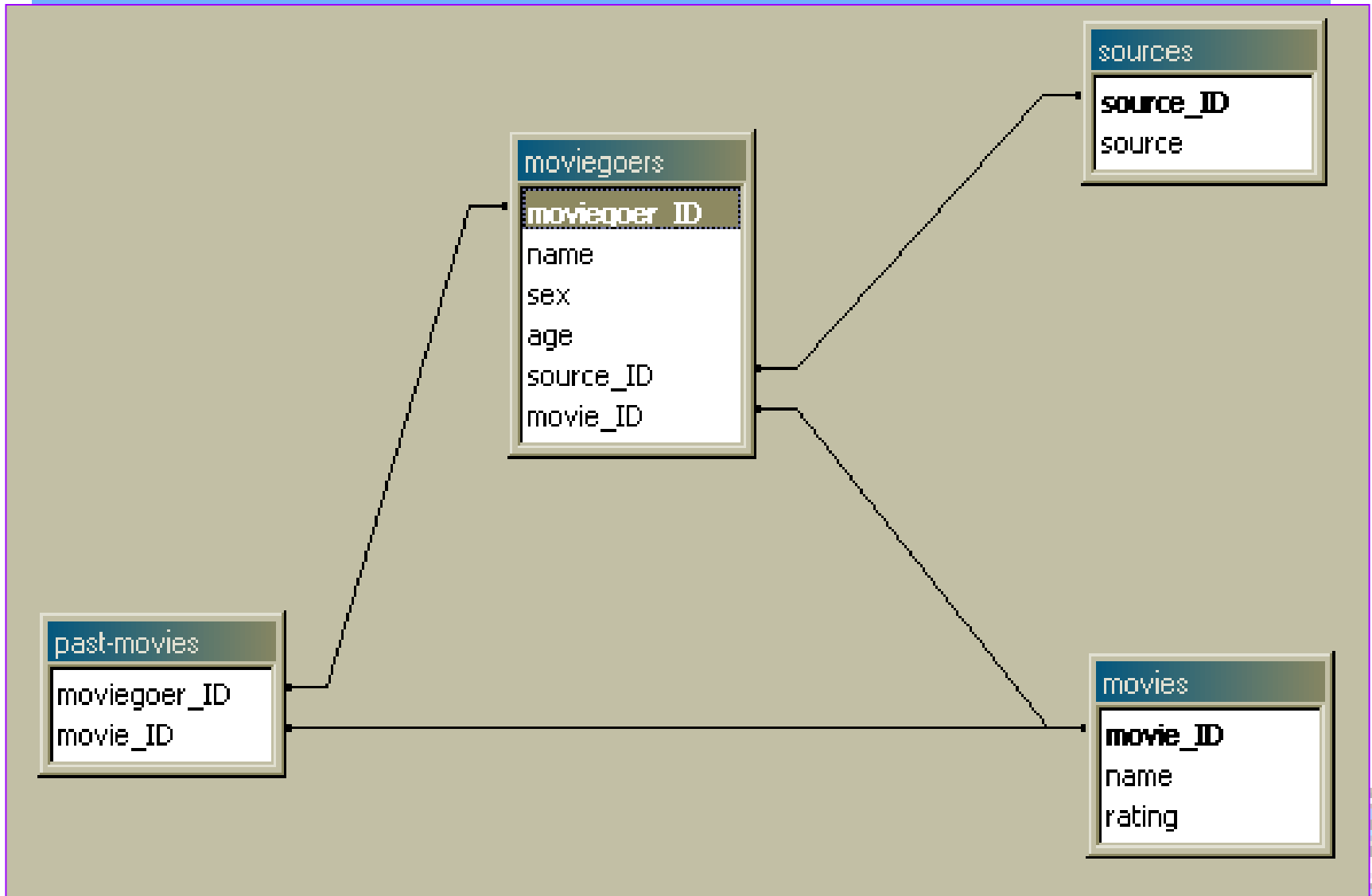
Example of extracted rules

⌘ **Education(university) & Height<176-180>**
⇒ Death cause (tumorous disease),
16; 0.62; +1.1;

⌘ **the relative frequency of patients who died on tumorous disease among patients with university education and with height 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients**



Moviegoer Database :

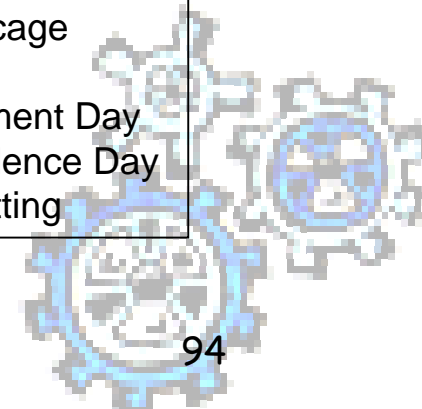


```

SELECT moviegoers.name, moviegoers.sex, moviegoers.age,
       sources.source, movies.name
FROM movies, sources, moviegoers
WHERE sources.source_ID = moviegoers.source_ID AND
      movies.movie_ID = moviegoers.movie_ID
ORDER BY moviegoers.name;

```

moviegoers.name	sex	age	source	movies.name
Amy	f	27	Oberlin	Independence Day
Andrew	m	25	Oberlin	12 Monkeys
Andy	m	34	Oberlin	The Birdcage
Anne	f	30	Oberlin	Trainspotting
Ansje	f	25	Oberlin	I Shot Andy Warhol
Beth	f	30	Oberlin	Chain Reaction
Bob	m	51	Pinewoods	Schindler's List
Brian	m	23	Oberlin	Super Cop
Candy	f	29	Oberlin	Eddie
Cara	f	25	Oberlin	Phenomenon
Cathy	f	39	Mt. Auburn	The Birdcage
Charles	m	25	Oberlin	Kingpin
Curt	m	30	MRJ	T2 Judgment Day
David	m	40	MRJ	Independence Day
Erica	f	23	Mt. Auburn	Trainspotting



Example: Moviegoer Database

⌘ Classification

- ☑ determine sex based on age, source, and movies seen
- ☑ determine source based on sex, age, and movies seen
- ☑ determine most recent movie based on past movies, age, sex, and source

⌘ Estimation

- ☑ for predict, need a continuous variable (e.g., "age")
- ☑ predict age as a function of source, sex, and past movies
- ☑ if we had a "rating" field for each moviegoer, we could predict the rating a new moviegoer gives to a movie based on age, sex, past movies, etc.



Example: Moviegoer Database

⌘ Clustering

- ☒ find groupings of movies that are often seen by the same people
- ☒ find groupings of people that tend to see the same movies
- ☒ clustering might reveal relationships that are not necessarily recorded in the data (e.g., we may find a cluster that is dominated by people with young children; or a cluster of movies that correspond to a particular genre)



Example: Moviegoer Database

⌘ Association Rules

- ☑ market basket analysis (MBA): "which movies go together?"
- ☑ need to create "transactions" for each moviegoer containing movies seen by that moviegoer:

name	TID	Transaction
Amy	001	{Independence Day, Trainspotting}
Andrew	002	{12 Monkeys, The Birdcage, Trainspotting, Phenomenon}
☑ Andy	003	{Super Cop, Independence Day, Kingpin}
Anne	004	{Trainspotting, Schindler's List}
...

ting" }

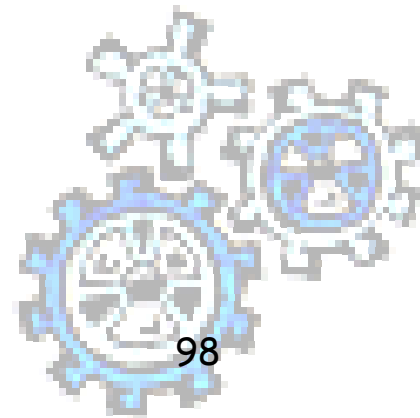
"f" }



Example: Moviegoer Database

⌘ Sequence Analysis

- ☑ similar to MBA, but order in which items appear in the pattern is important
- ☑ e.g., people who rent "The Birdcage" during a visit tend to rent "Trainspotting" in the next visit.



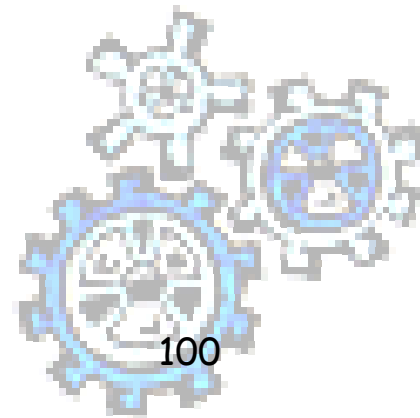
On the road to knowledge: mining 21 years of UK traffic accident reports

Peter Flach et al.

Silnet Network of Excellence

Mining traffic accident reports

- ⌘ The Hampshire County Council (UK) wanted to obtain a better insight into how the characteristics of traffic accidents may have changed over the past 20 years as a result of improvements in highway design and in vehicle design.
- ⌘ The database, contained police traffic accident reports for all UK accidents that happened in the period 1979-1999.



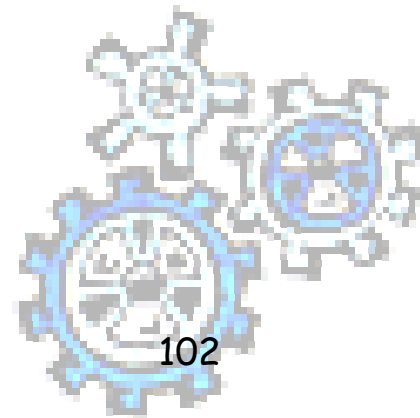
Business Understanding

- ⌘ Understanding of road safety in order to reduce the occurrences and severity of accidents.
 - ⊗ influence of road surface condition;
 - ⊗ influence of skidding;
 - ⊗ influence of location (for example: junction approach);
 - ⊗ and influence of street lighting.
- ⌘ trend analysis: long-term overall trends, regional trends, urban trends, and rural trends.
- ⌘ the comparison of different kinds of locations is interesting: for example, rural versus metropolitan versus suburban.



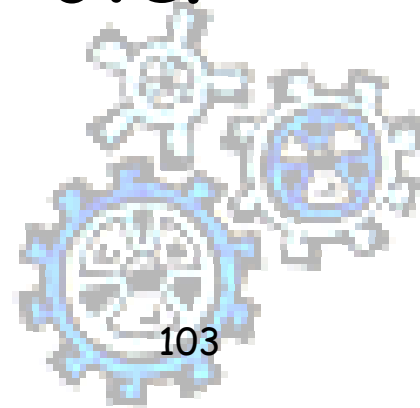
Data understanding

- ⌘ Low data quality. Many attribute values were missing or recorded as unknown.
- ⌘ Different maps were created to investigate the effect of several parameters like accident severity and accident date.



Modelling

- ⌘ The aim of this effort was to find interesting associations between road number, conditions (e.g., weather, and light) and serious or fatal accidents.
- ⌘ Certain localities had been selected and performed the analysis only over the years 1998 and 1999.



Extracted rule

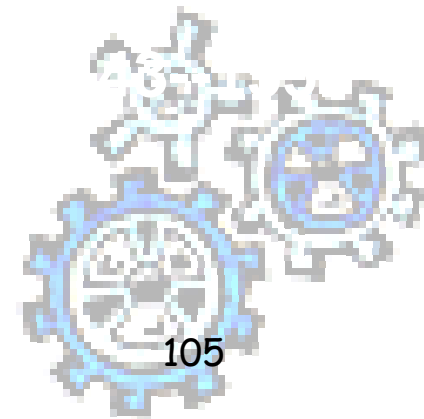
	FATAL	NonFATAL	TOTAL
Road=V61 AND Weather=1	15	141	156
NOT(Road=V61 AND Weather=1)	147	5056	5203

- ⌘ The relative frequency of fatal accidents among all accidents in the locality was 3%.
- ⌘ The relative frequency of fatal accidents on the road (V61) under fine weather with no winds was 9.6% – more than 3 times greater.



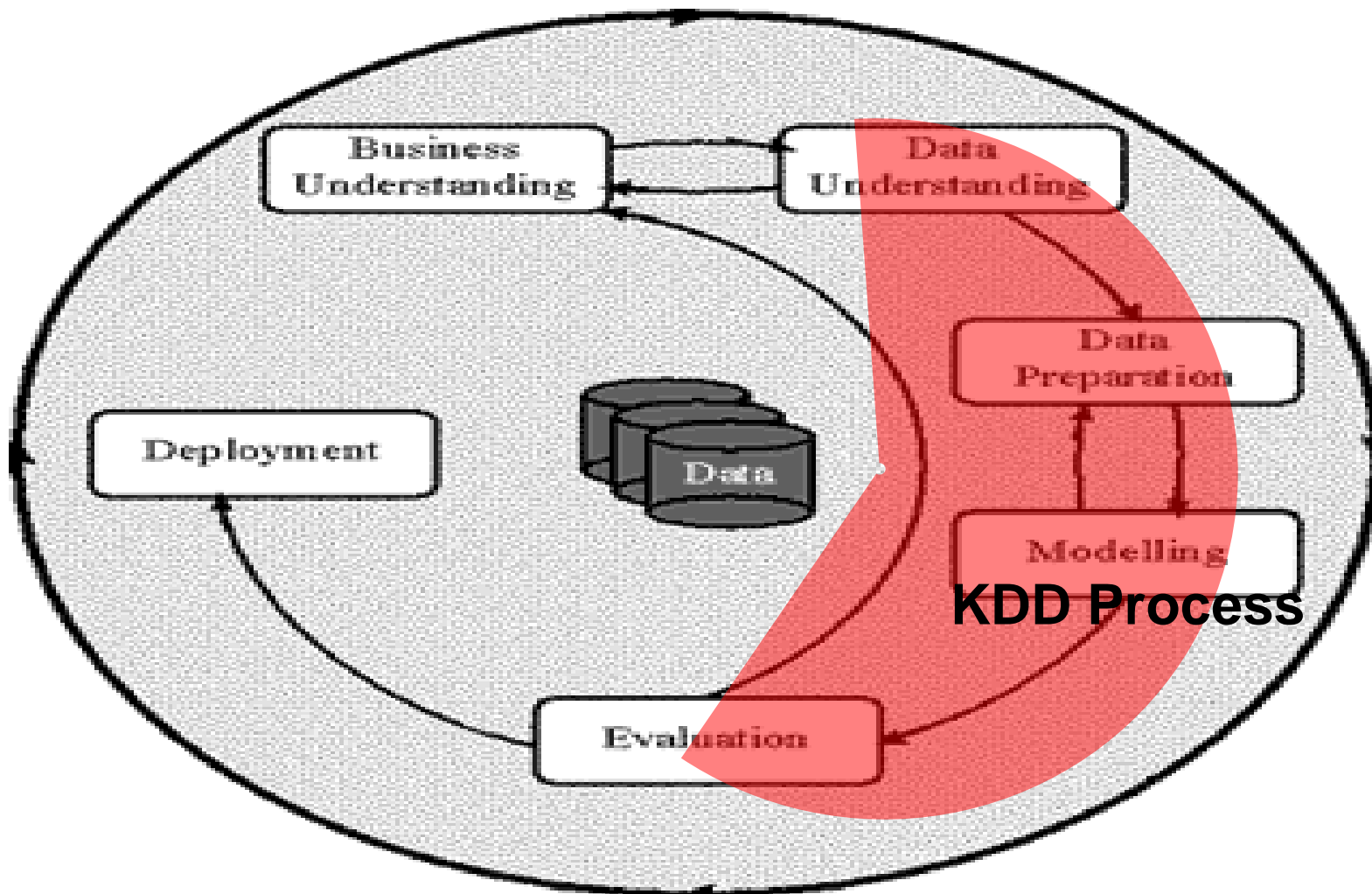
Seminar 1 Case studies - Bibliography

- ⌘ Artif. Intell. Med. 2001 Jun;22(3), Special Issue on Data Mining in Medicine
- ⌘ Klosgen, Zytkow, Handbook of Data Mining, Oxford, 2001
- ⌘ Micheael, J. A. Berry, Gordon S. Linoff, Mastering Data Mining, Wiley, 2000
- ⌘ ECML/PKDD2004 Discovery Challenge
homepage[<http://lisp.vse.cz/challenge/ecmlpkdd2004/>]
- ⌘ On the road to knowledge: mining 21 years of UK traffic accident reports, in *Data Mining and Decision Support: Aspects of Integration and Collaboration*, pages 143--155. Kluwer Academic Publishers, January 2003



How to develop a Data Mining Project?

CRISP-DM: The life cycle of a data mining project



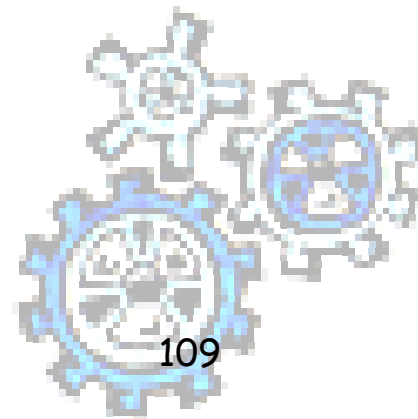
Business understanding

- ⌘ Understanding the project objectives and requirements from a business perspective.
- ⌘ then converting this knowledge into a data mining problem definition and a preliminary plan.
 - ⊞ Determine the Business Objectives
 - ⊞ Determine Data requirements for Business Objectives
 - ⊞ Translate Business questions into Data Mining Objective



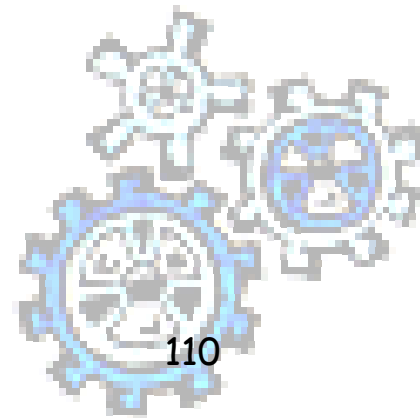
Data understanding

- ⌘ Data understanding: characterize data available for modelling. Provide assessment and verification for data.



Modeling

- ⌘ In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
- ⌘ Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data.
- ⌘ Therefore, stepping back to the data preparation phase is often necessary.



Evaluation

- ⌘ At this stage in the project you have built a model (or models) that appears to have high quality from a data analysis perspective.
- ⌘ Evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.
- ⌘ A key objective is to determine if there is some important business issue that has not been sufficiently considered.



Deployment

- ⌘ The knowledge gained will need to be organized and presented in a way that the customer can use it.
- ⌘ It often involves applying “live” models within an organization’s decision making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases.



Deployment

- ⌘ It can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.
- ⌘ In many cases it is the customer, not the data analyst, who carries out the deployment steps.

