Francesco Bonchi – KDD Lab
ISTI-C.N.R.
Email: francesco.bonchi@isti.cnr.it
Web: http://www-kdd.isti.cnr.it/~bonchi/

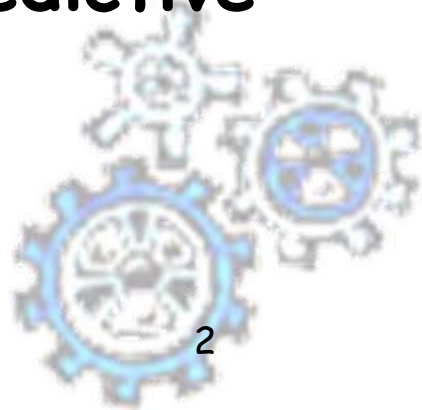
Contenuti della lezione (26/02/07)

- The KDD Process
- Professional figures in the process
- The CRISP-DM Model
- Examples of KDD projects

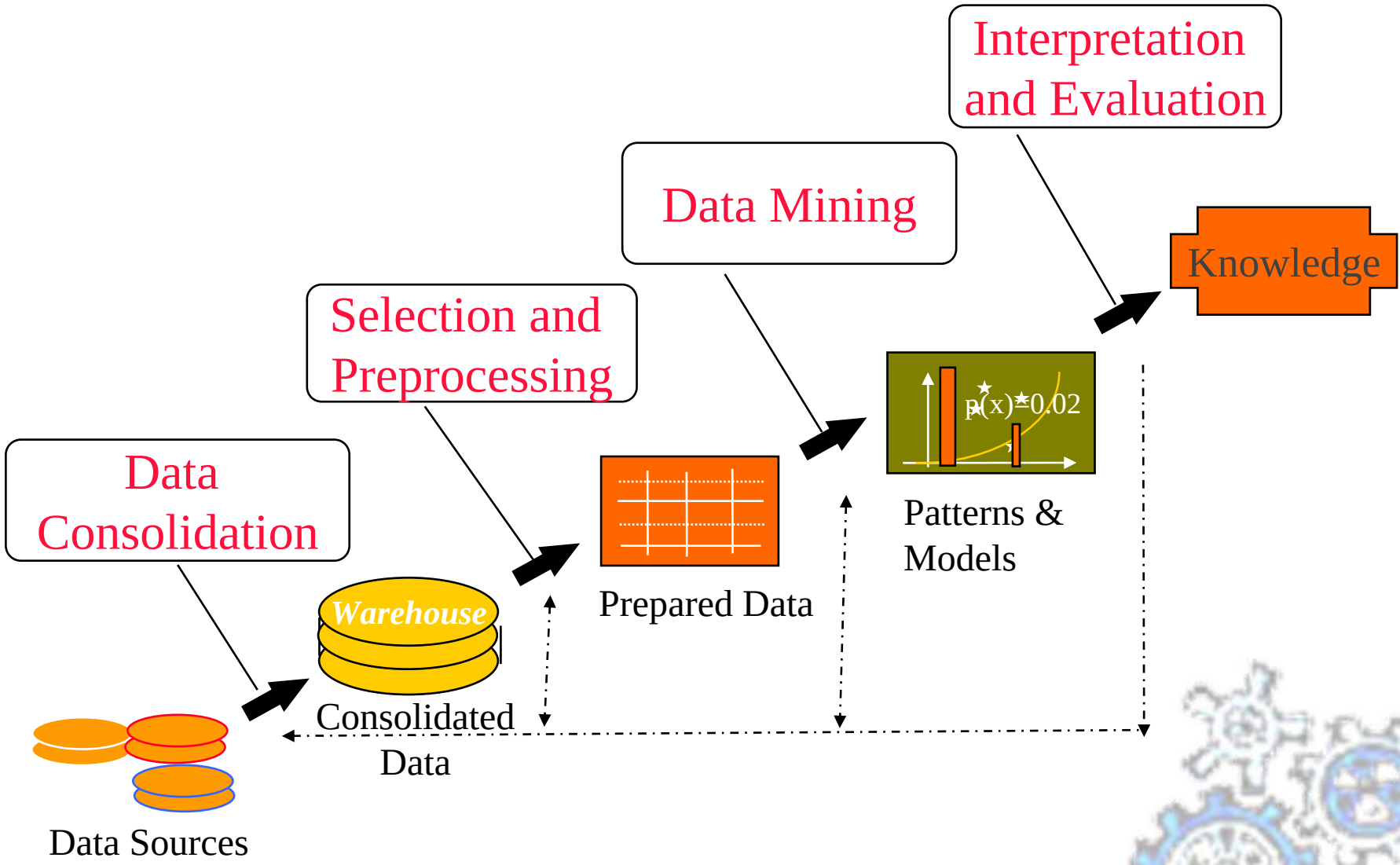
What is Knowledge Discovery in Databases (KDD)? A process!

- 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.

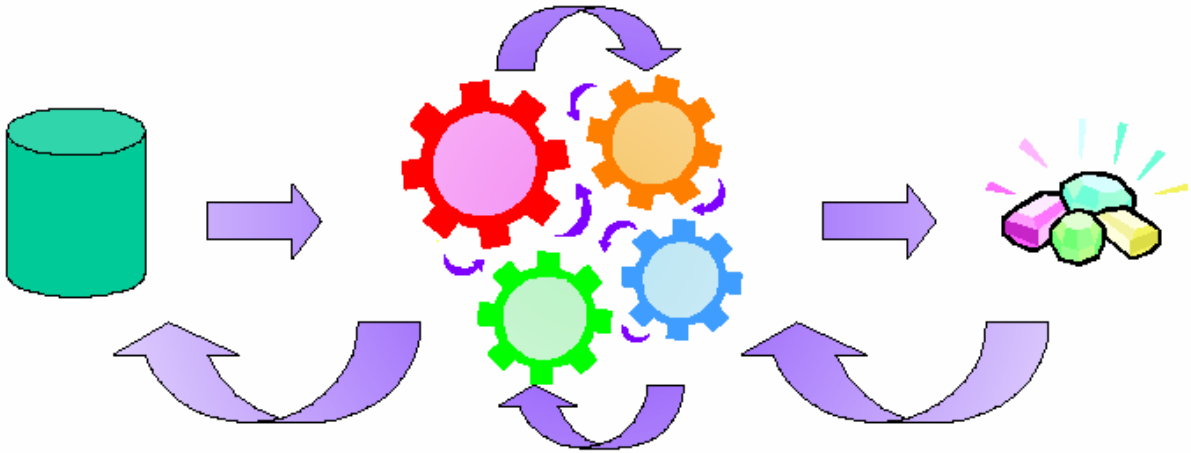


The KDD process



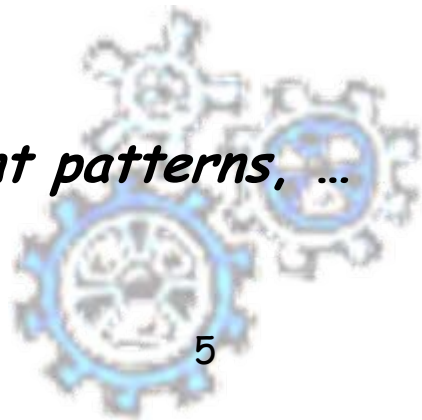
The KDD Process in Practice

- KDD is an Iterative Process
  - art + engineering rather than science

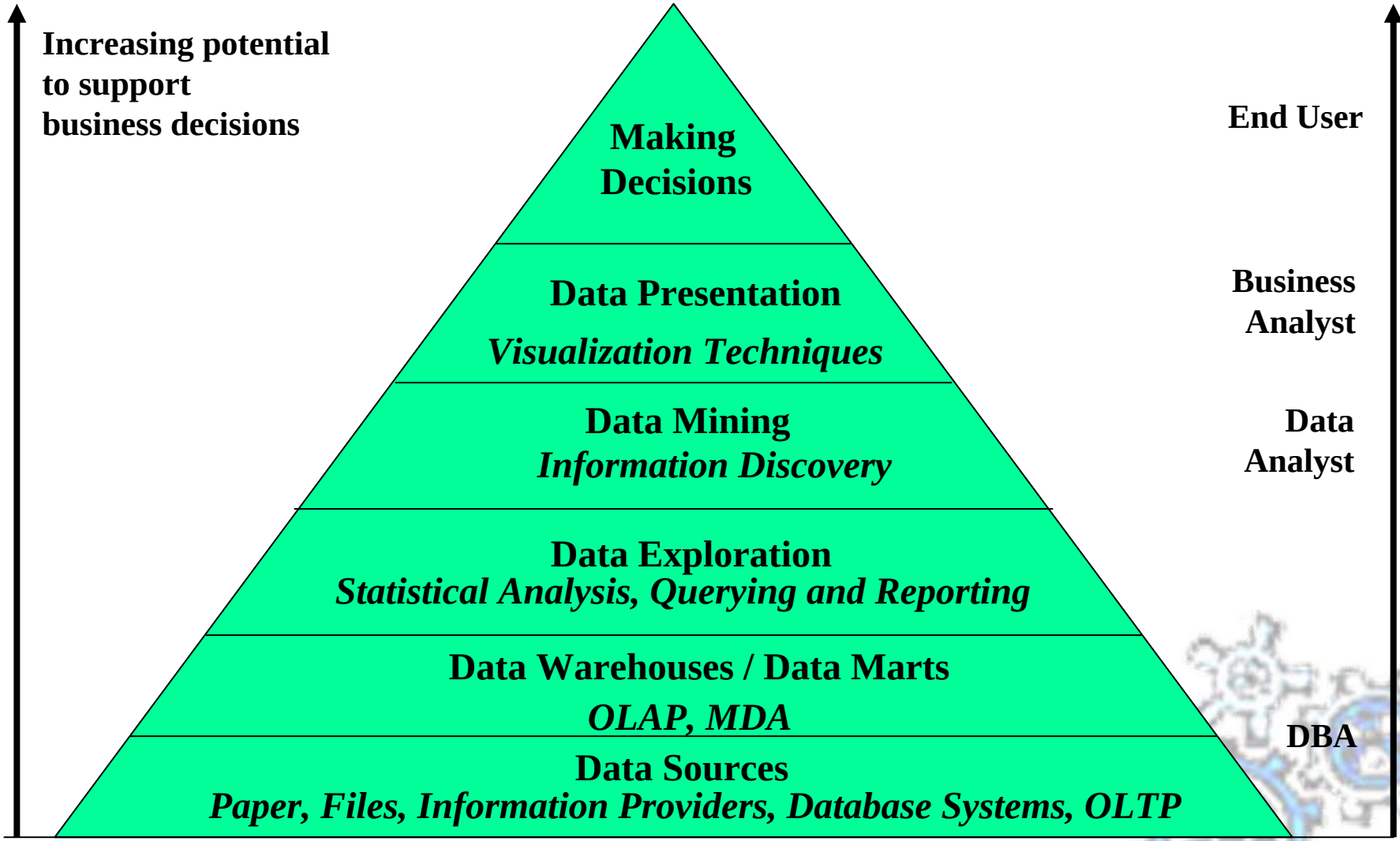


The steps of the KDD process

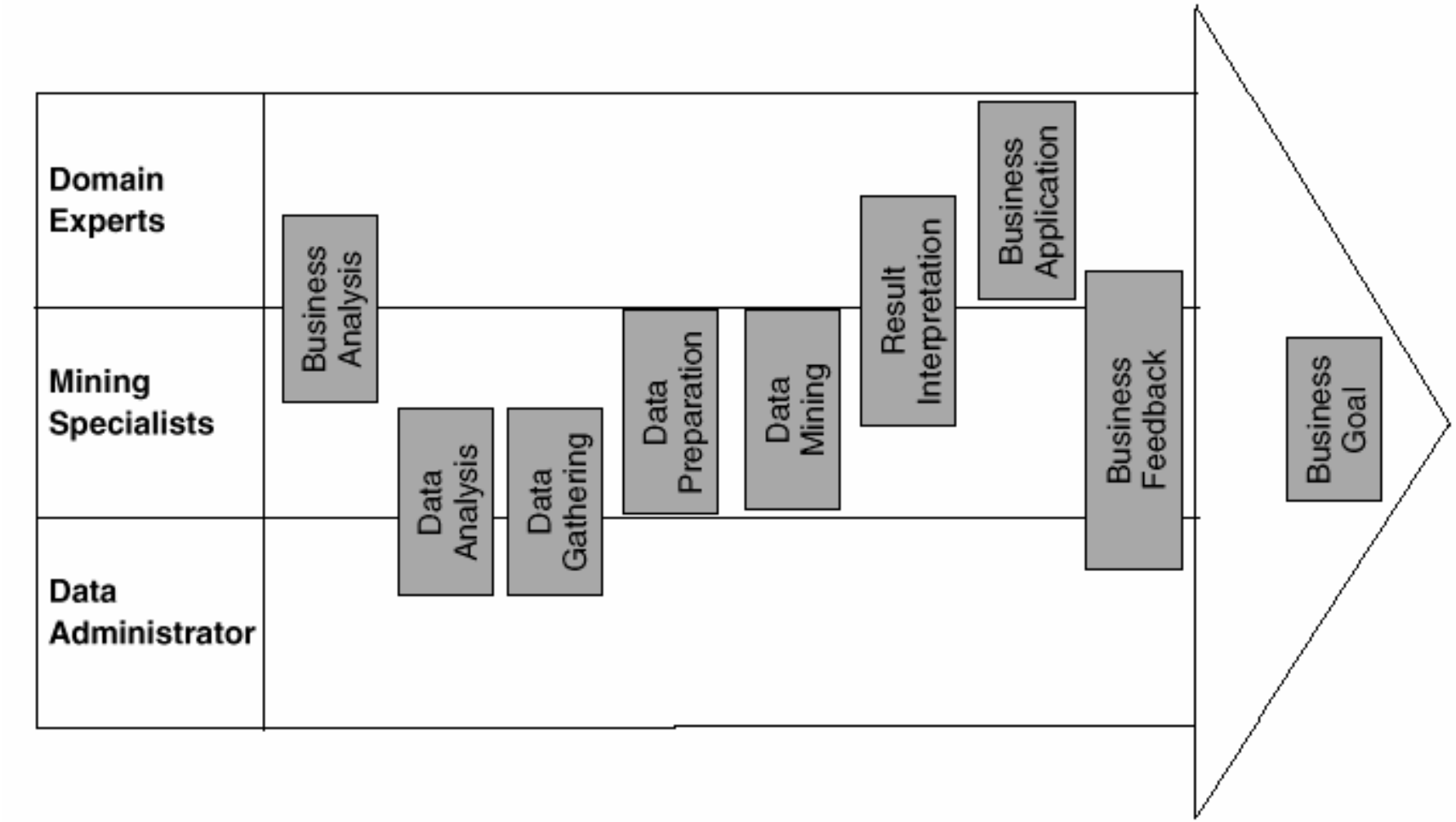
- Learning the application domain:
  - relevant prior knowledge and goals of application
- Data consolidation: Creating a target data set
- Selection and Preprocessing
  - *Data cleaning* : (may take 60% of effort!)
  - *Data reduction and projection*:
    - find useful features, dimensionality/variable 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



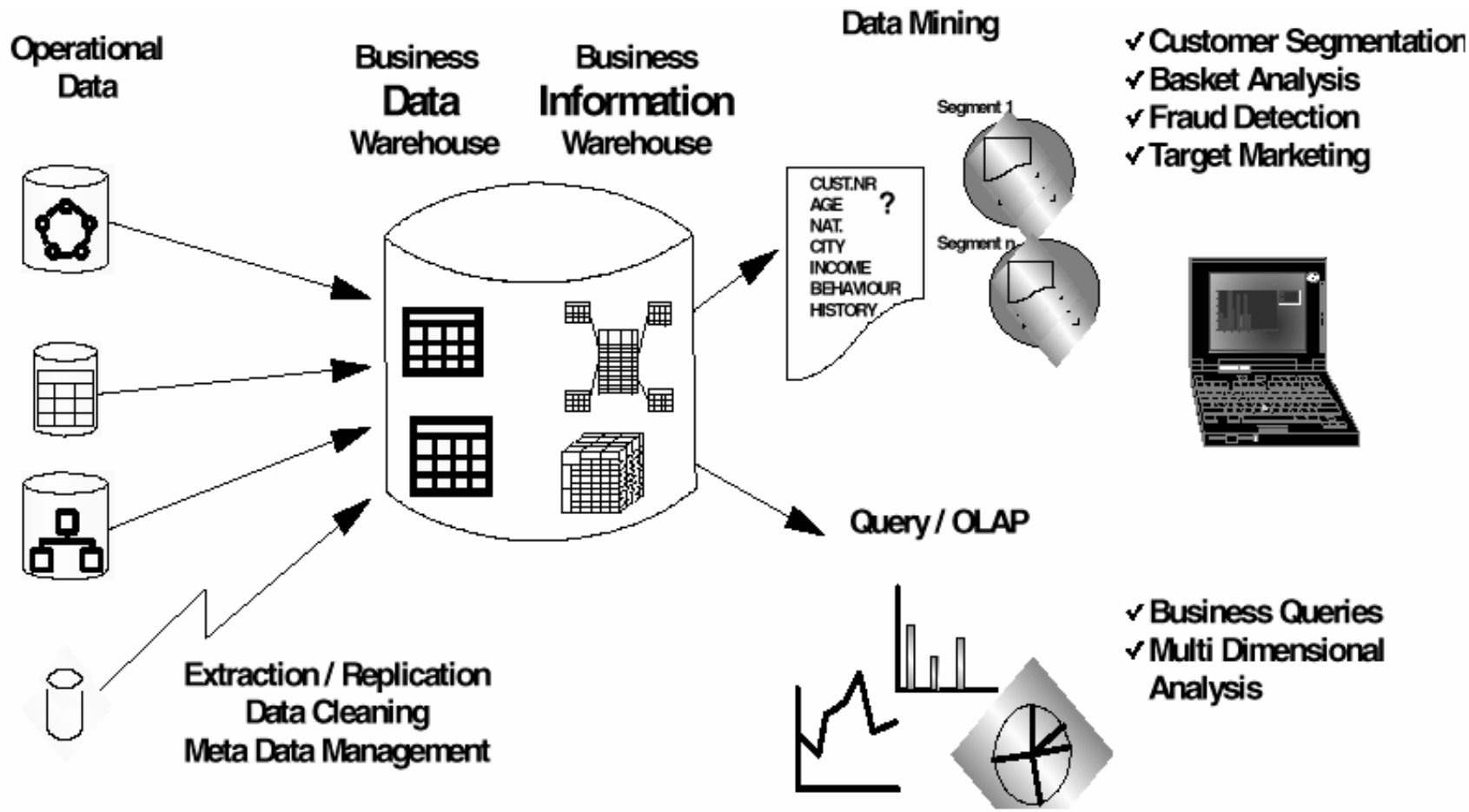
Data mining and business intelligence



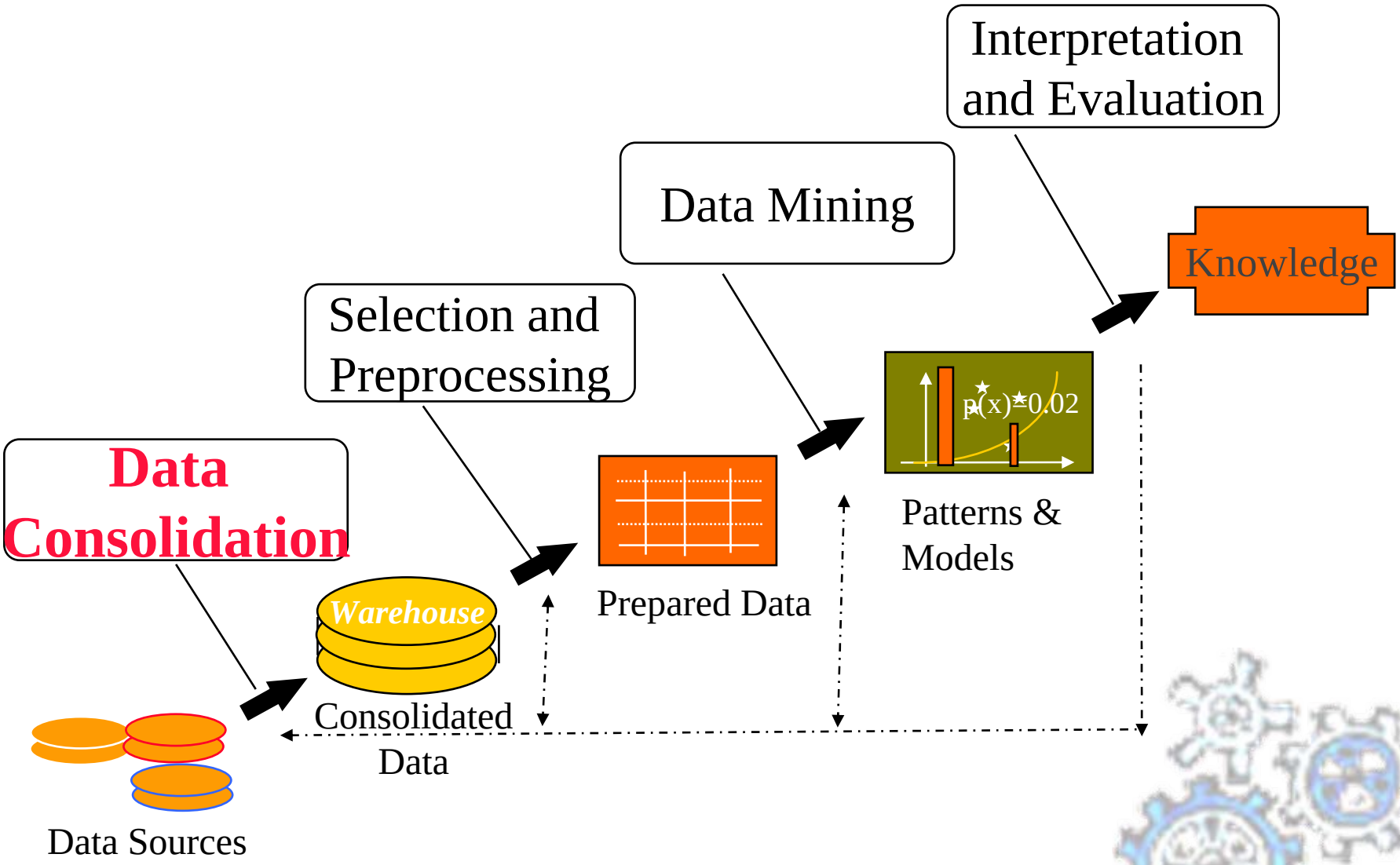
Roles in the KDD process



A business intelligence environment



The KDD process



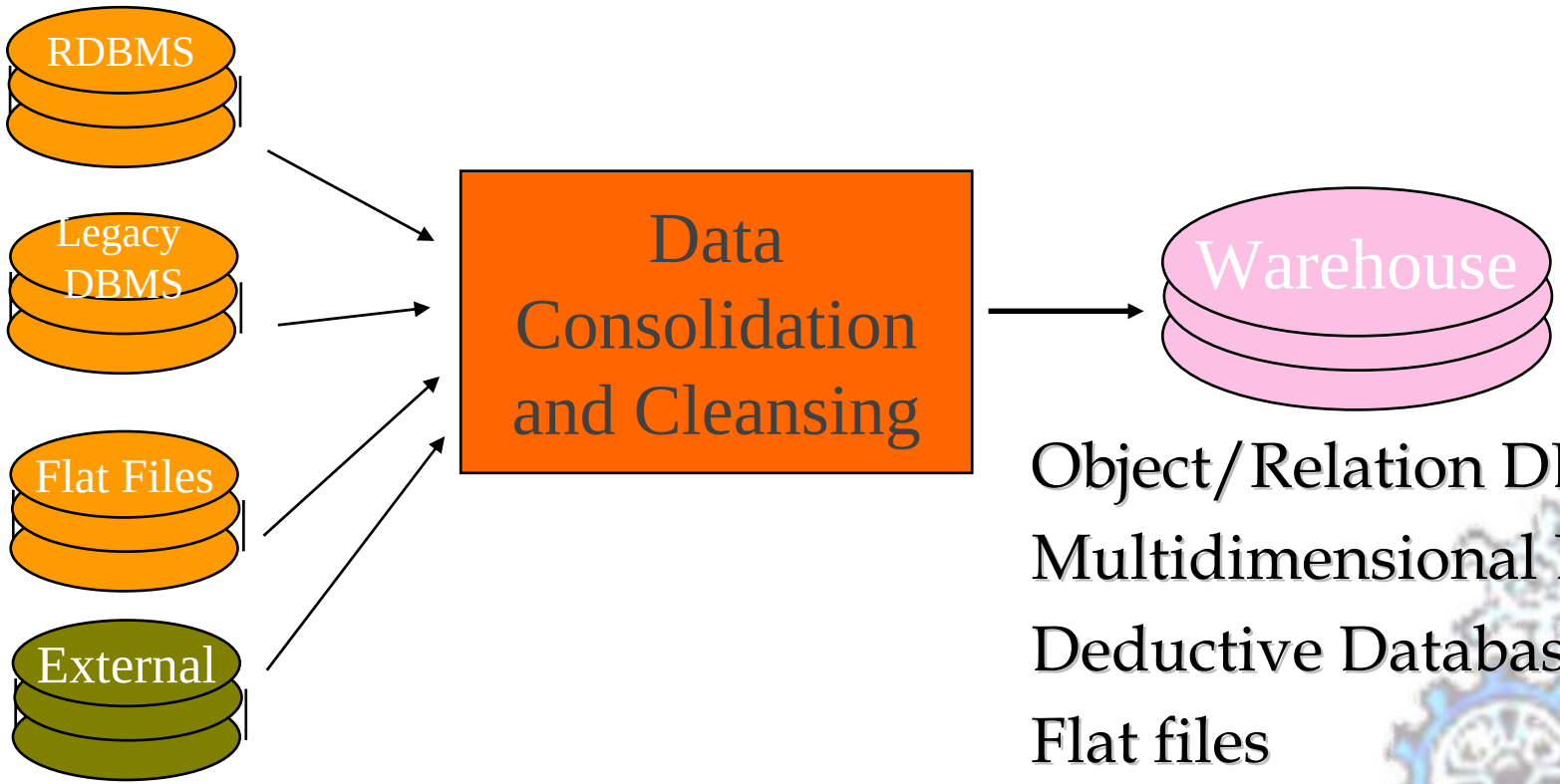
Data consolidation and preparation

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

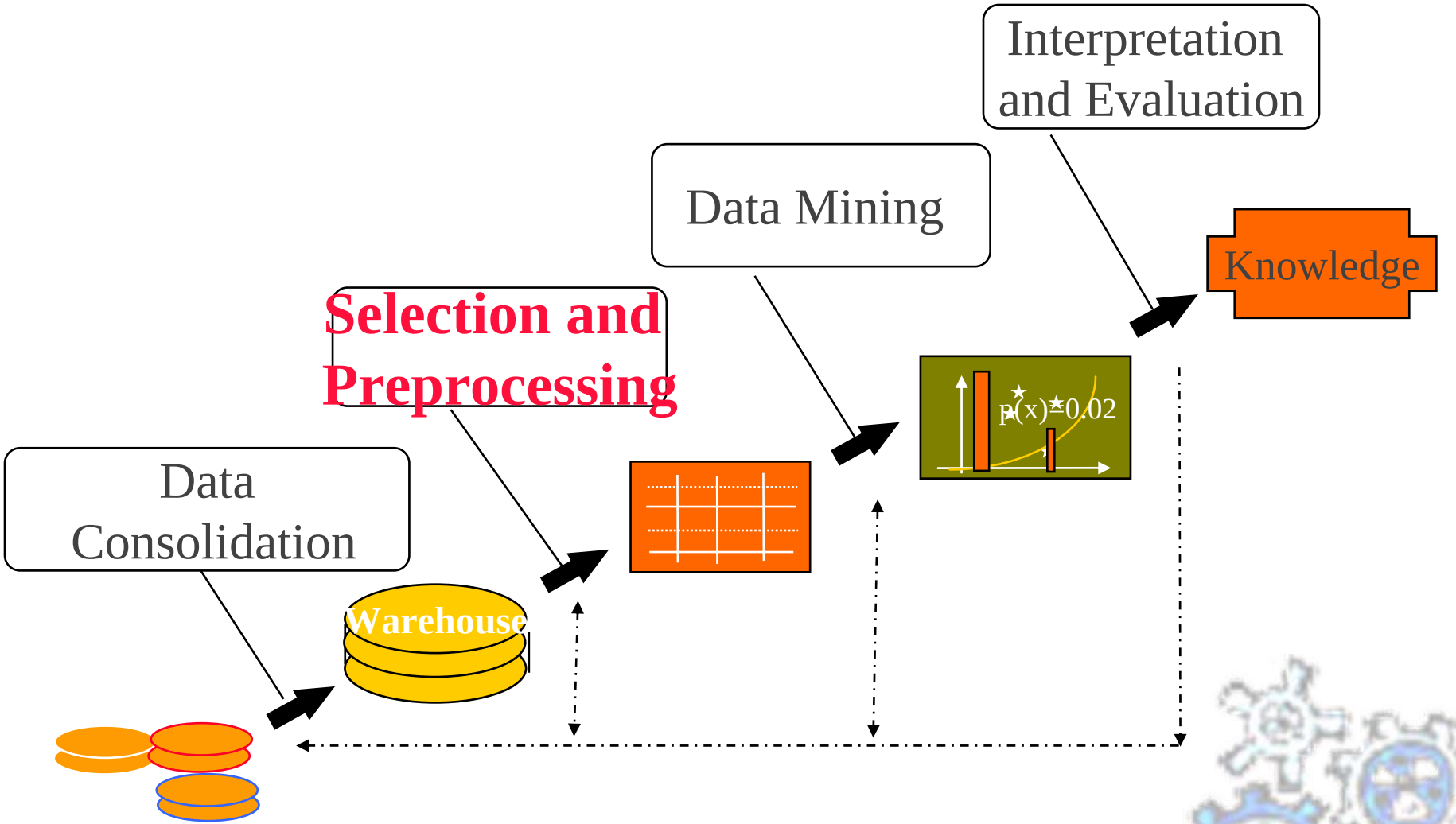


Object/Relation DBMS
Multidimensional DBMS
Deductive Database
Flat files

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*

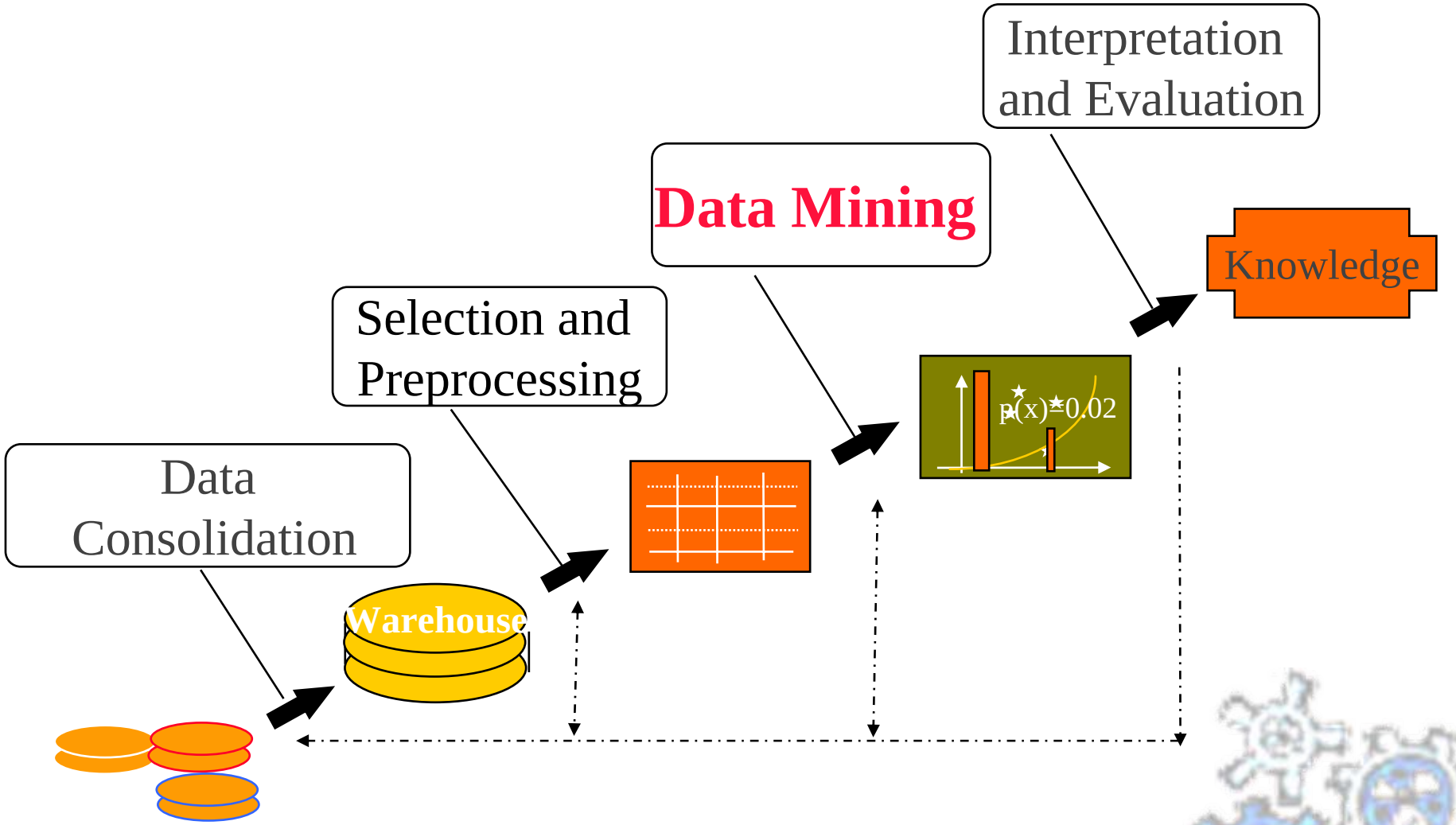
The KDD process



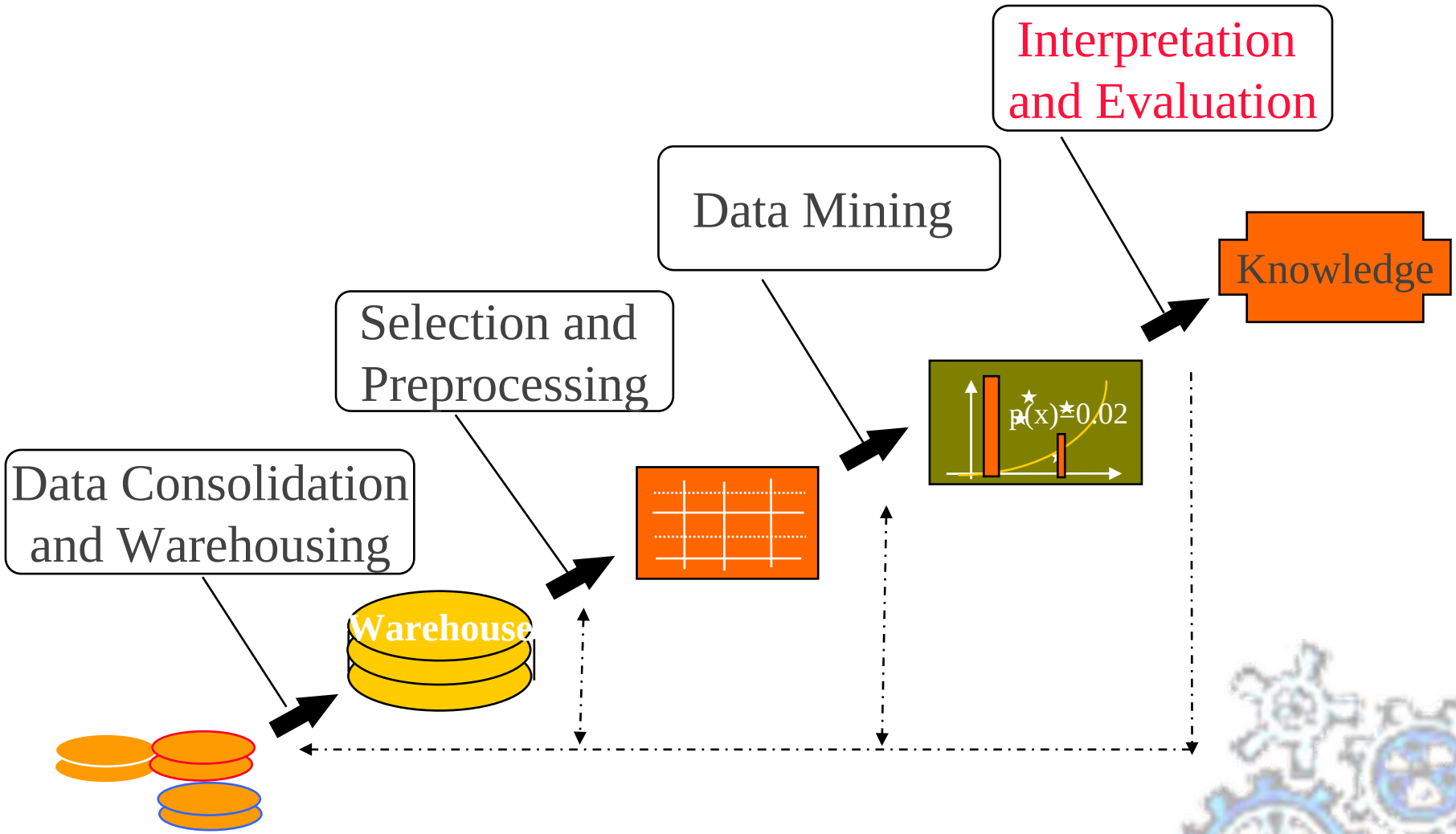
Data selection and preprocessing

- 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



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

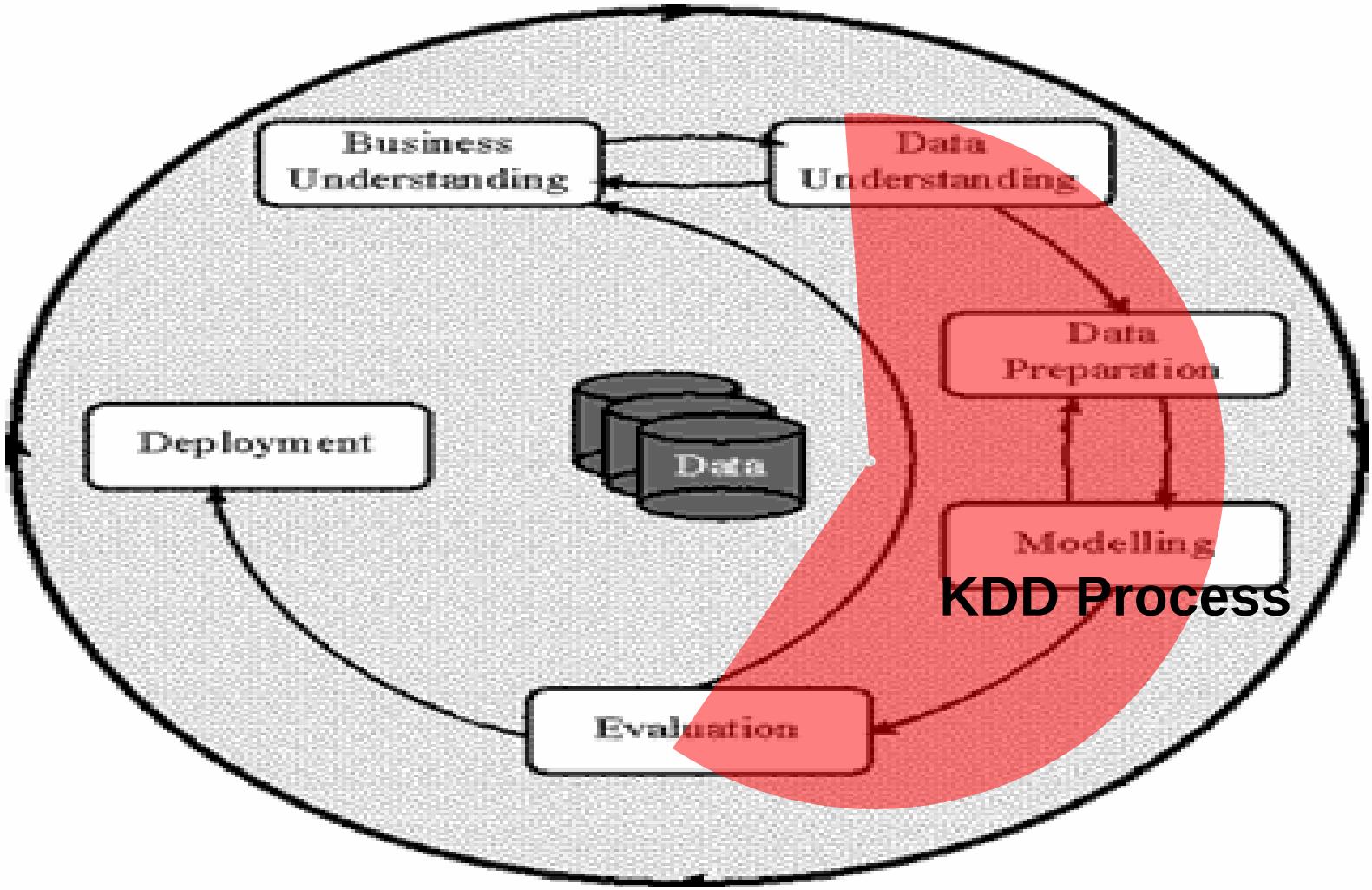
Evaluation

- 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)

CRISP-DM: The life cicle 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.

Examples of KDD projects

Competitive Intelligence
Fraud Detection,
Traffic Accident Analysis,

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

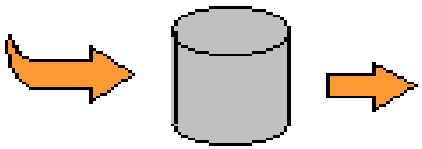
The Derwent database

- Contains all patents filed worldwide in last 10 years
- Searching this database by keywords may yield thousands of documents
- Derwent document are semi-structured: many long text fields

- Goal: analyze Derwent document 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

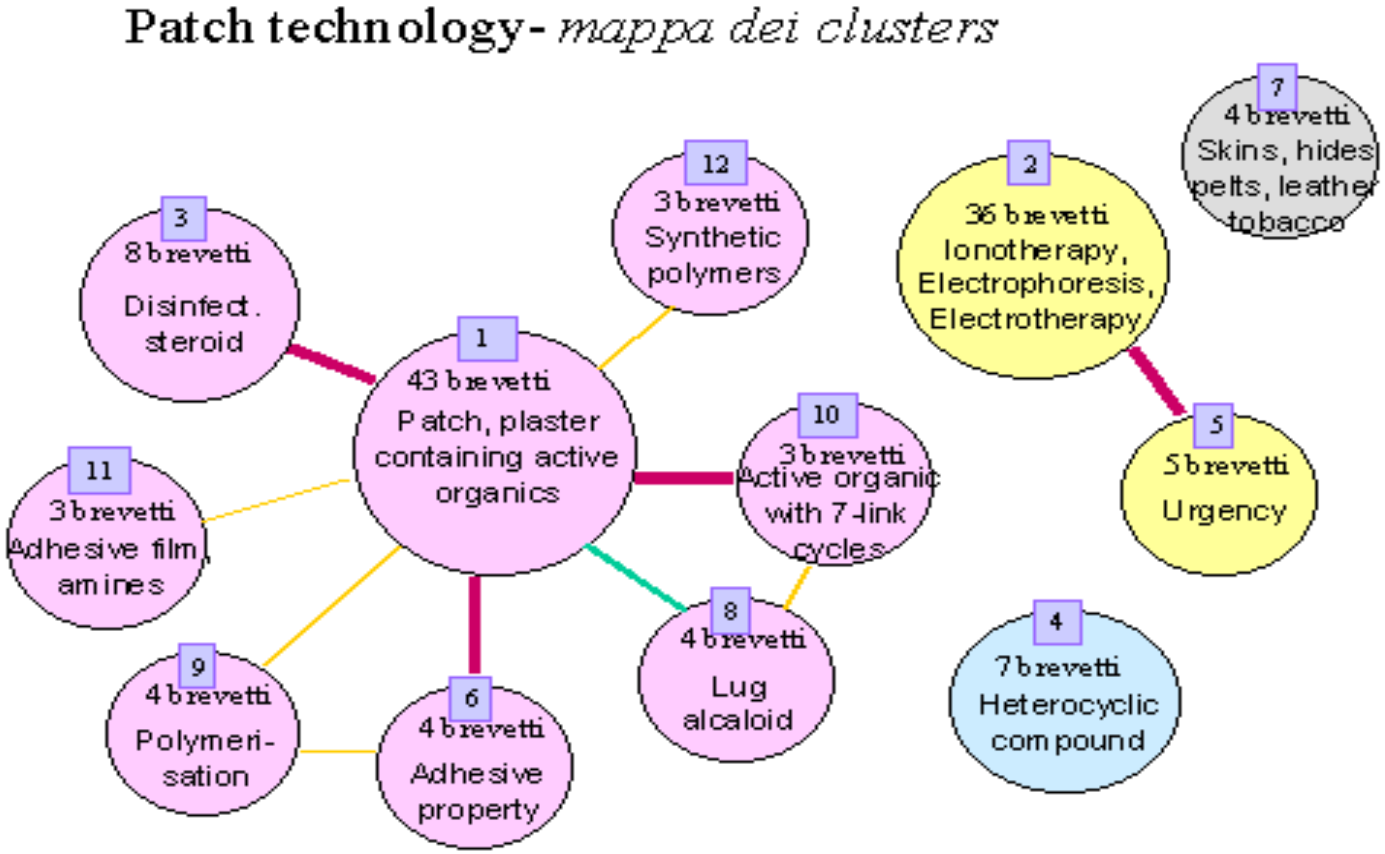
PN: KR940043 B1 940120 DW9445

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



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 electrophorese devices
A61M-037/00 Therapeutic patch

Classificazione Derwent:

S05 Electromedical
P34 Health, Electrotherapy

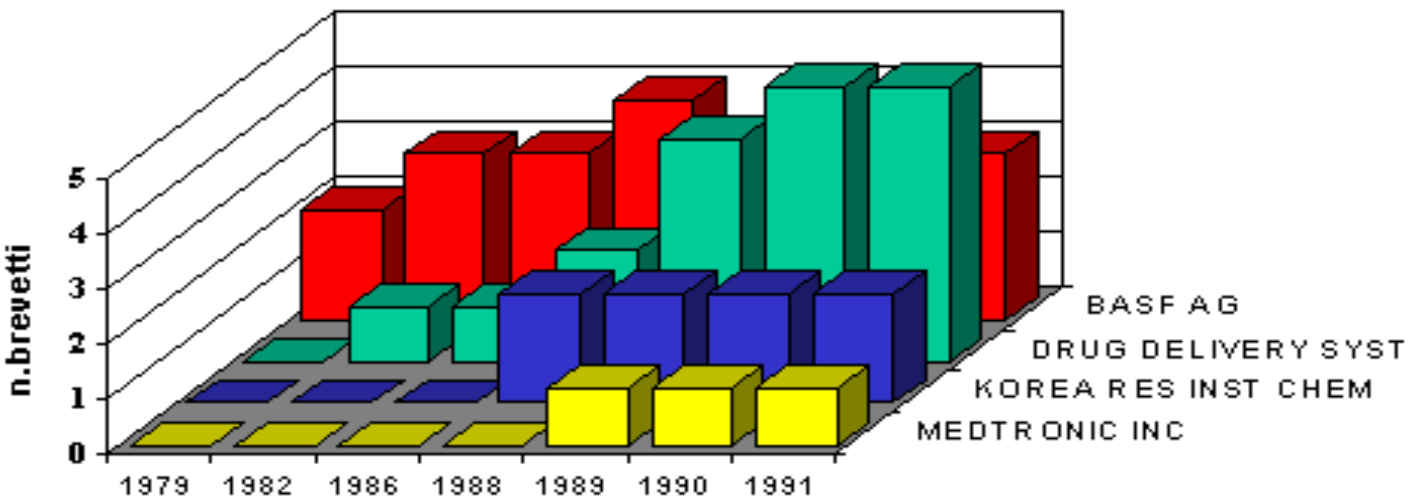
Società proprietarie:

DRUG DELIVERY SYST	42%
BASF AG	36%
KOREA RES INST CHEM	16%
MEDTRONIC INC	6%

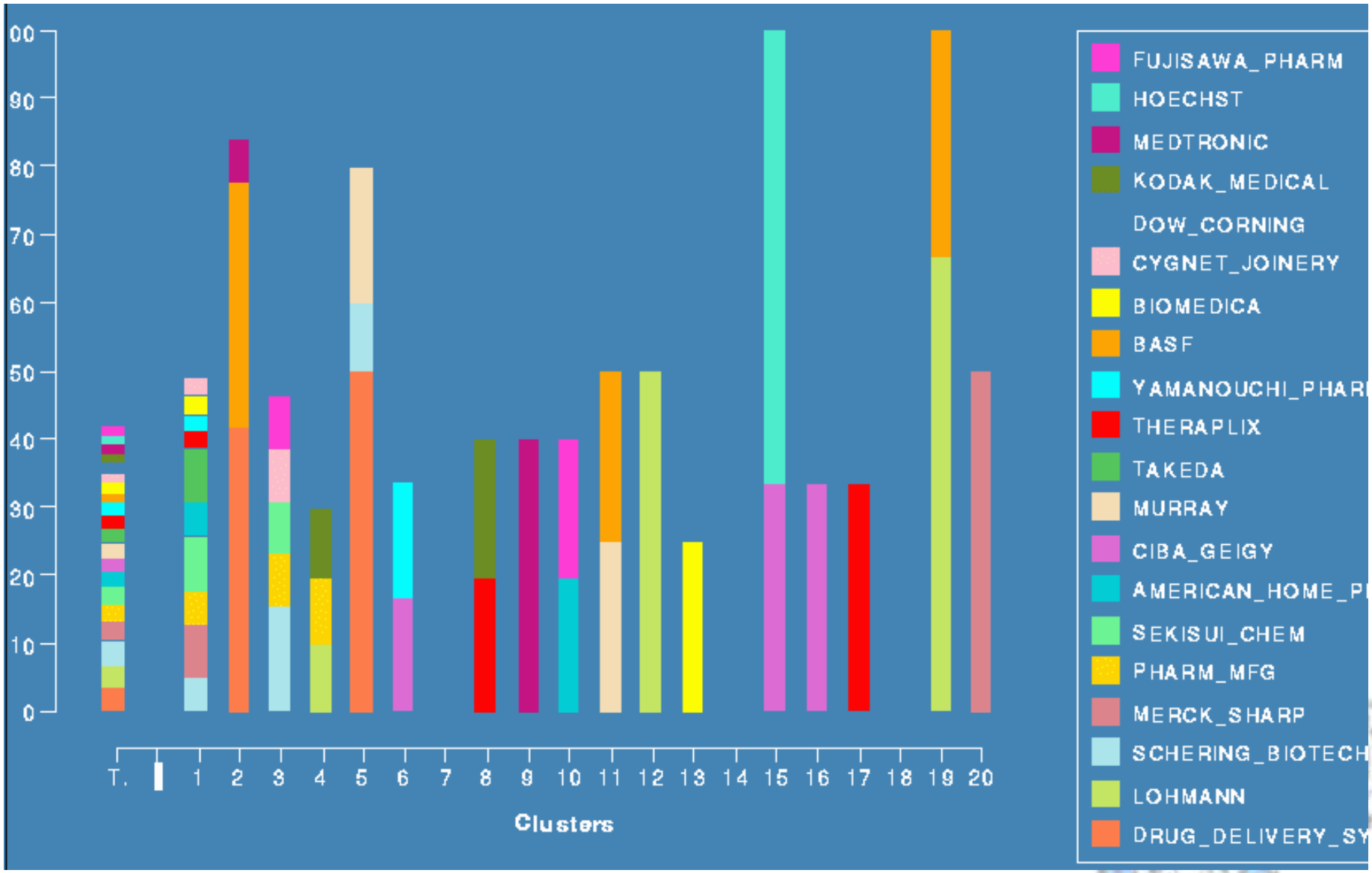
Anno	n. brevetti
1979	2
1982	4
1986	4
1988	8
1989	10
1990	11
1991	11

Zoom on cluster 2 - profiling competitors

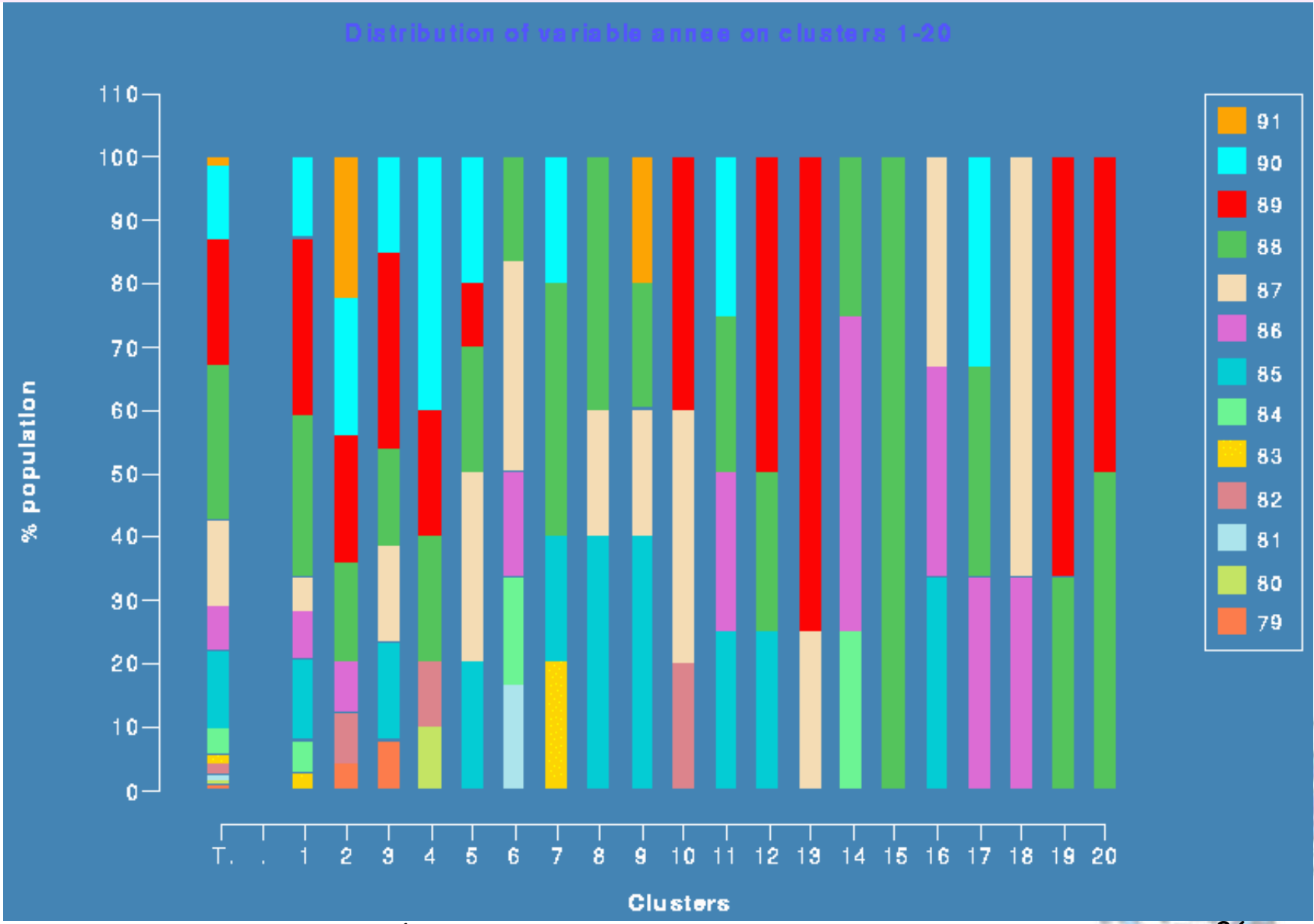
Patch technology- *cluster n.2 -*
attività della concorrenza nel tempo



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

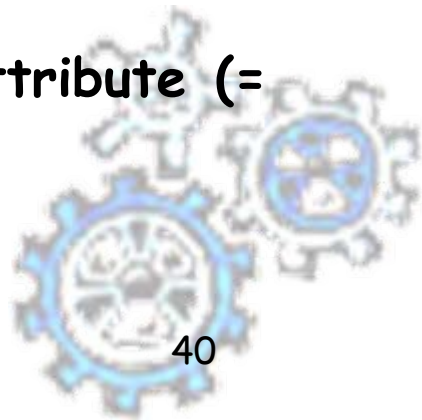
- A major task in fraud detection is constructing *models* of fraudulent behavior, for:
  - preventing future frauds (*on-line* fraud detection)
  - discovering past frauds (*a posteriori* fraud detection)
- analyze historical audit data to plan effective future audits

Audit planning

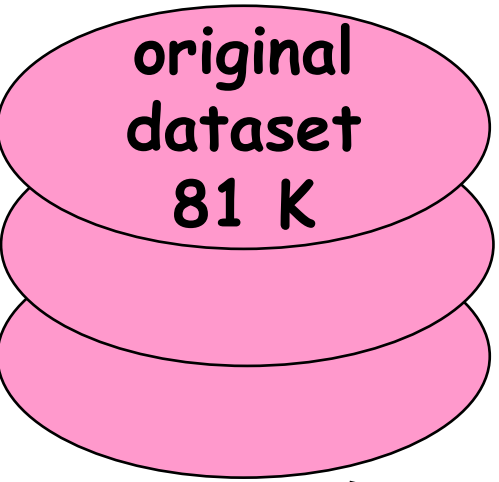
- Need to face a trade-off between conflicting issues:
  - *maximize audit benefits*: select subjects to be audited to maximize the recovery of evaded tax
  - *minimize audit costs*: select subjects to be audited to minimize the resources needed to carry out the audits.

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*)



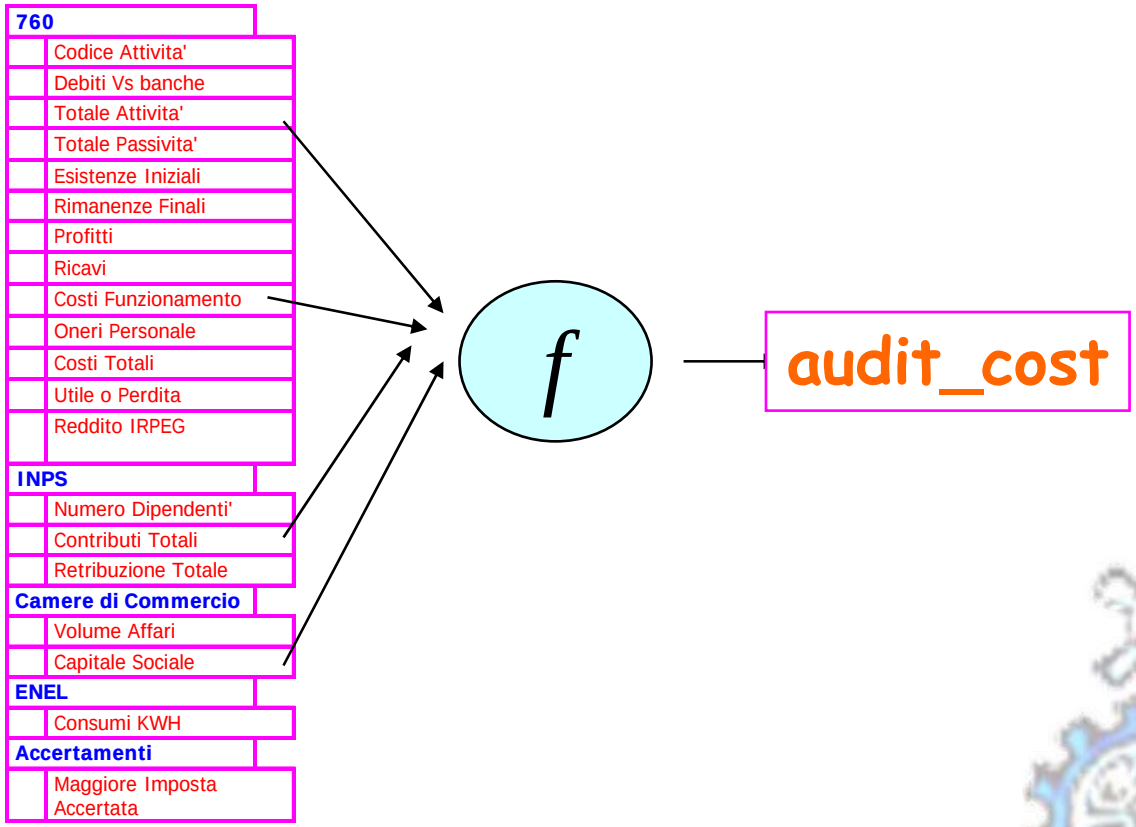
Data preparation



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
  actual_recovery = recovery - audit_cost

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

- $$c.a.r. = \begin{cases} negative & if \quad actual\_recovery \leq 0 \\ positive & if \quad actual\_recovery > 0. \end{cases}$$

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-independent quality indicators

- **confusion matrix** (of a given classifier)

negative	positive	← classified as
TN	FP	actual class negative
FN	TP	actual class positive

TN (TP): true negative (positive) tuples
FN (FP): false negative (positive) tuples

misclassification rate =
(FN ∪ FP) / # test-set

Domain-dependent quality indicators

- audit # (of a given classifier): number of tuples classified as positive =
  # (FP ∪ 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

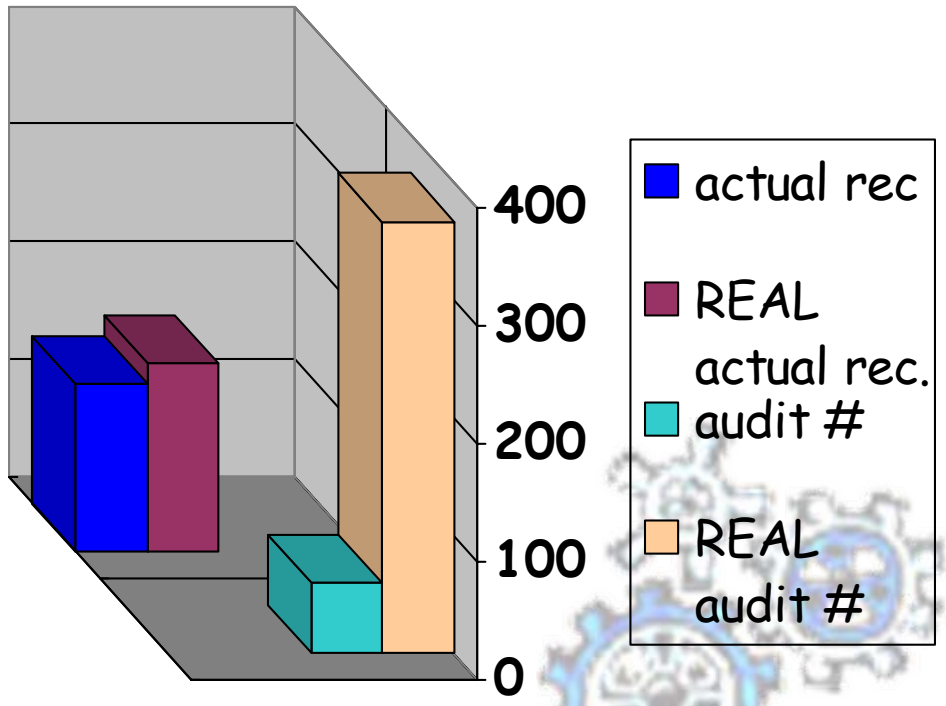
- Classifiers can be compared with the REAL case, consisting of the whole test-set:

- 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

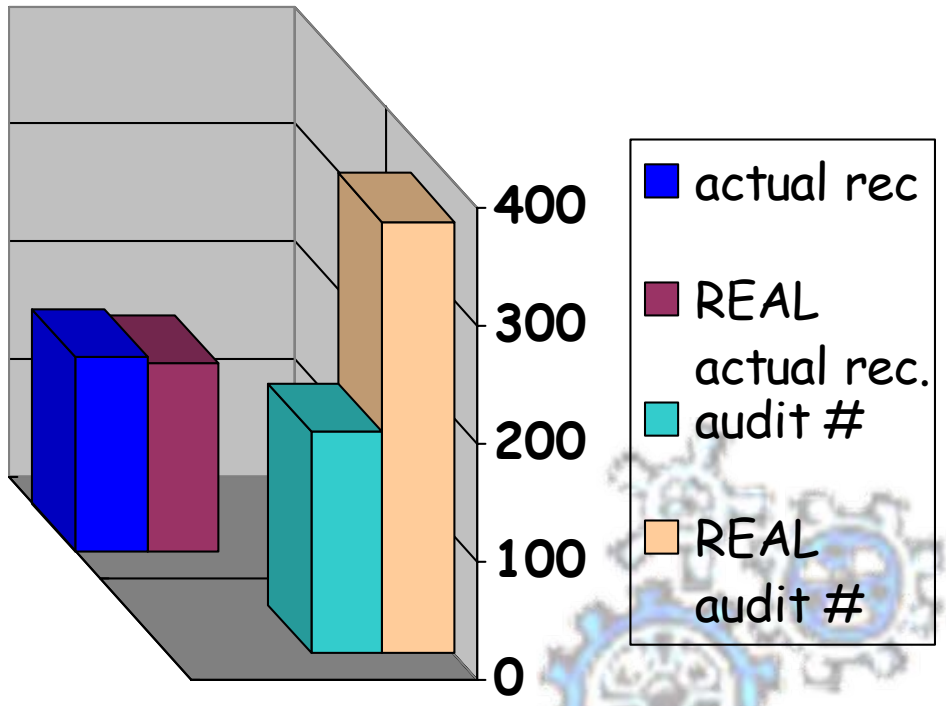
- *misc. rate* = 22%
- *audit #* = 59 (11 FP)
- *actual rec.*= 141.7 Meuro
- *profitability* = 2.401



Model evaluation: classifier 2 (min FN)

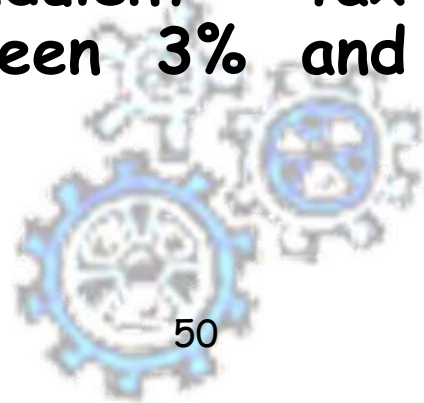
- replication in training-set (balanced neg/pos)
- misc. weights (trade 3 FP for 1 FN)
- 3-trees adaptive boosting

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



Fraud detection and audit planning

- 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)
- Focus on a posteriori FD: analyze historical audit data to plan effective future audits
- Audit planning is a key factor, e.g. in the fiscal and insurance domain:
  - tax evasion (from incorrect/fraudulent tax declarations) estimated in Italy between 3% and 10% of GNP



Case study

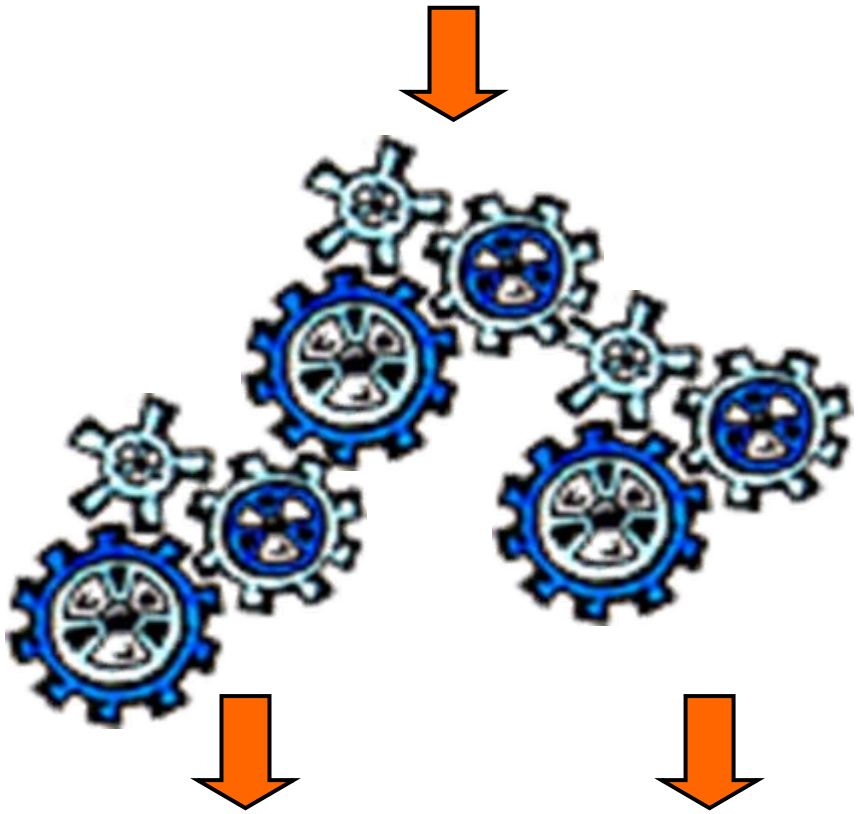
- Conducted by our Pisa KDD Lab (Bonchi et al 99)
- A data mining project at the Italian Ministry of Finance, with the aim of assessing:
  - the potential of a KDD process oriented to planning audit strategies;
  - a methodology which supports this process;
  - an integrated logic-based environment which supports its development.

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.
- Is there a KDD methodology which may be tuned according to these options?
- How extracted knowledge may be combined with domain knowledge to obtain useful audit models?

Autofocus data mining

policy options, business rules



fine parameter tuning of mining tools

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:

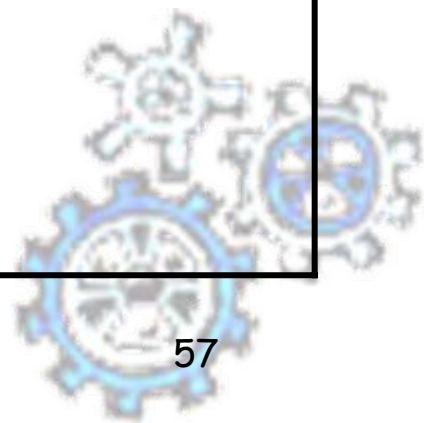
- The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.
- Used for Discovery Challenge at PKDD 00-02-03-04

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	Breathlesness	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

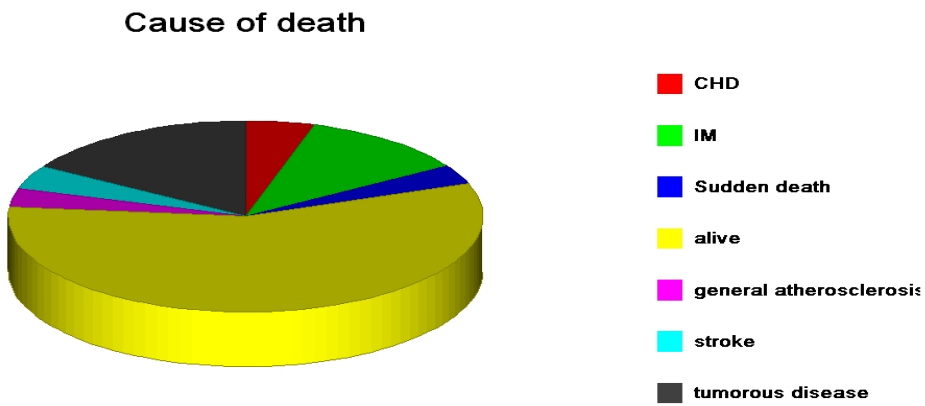


The input data

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 implicitely 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 (?) ⇒ Death cause (?)

Examinations (?) ⇒ Death cause (?)

Habits (?) ⇒ Death cause (?)

Combinations (?) ⇒ Death cause (?)

Example of extracted rules

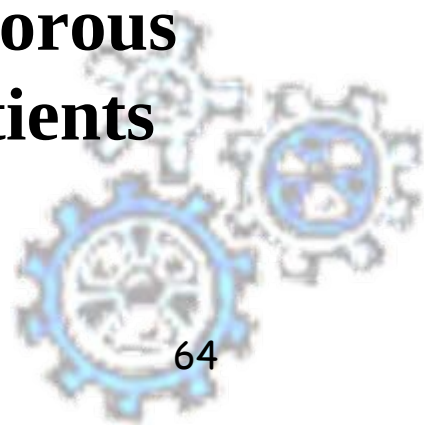
- **Education(university) & Height<176-180> ⇒Death cause (tumouros disease), *16 ; 0.62***
- **It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.**

Example of extracted rules

- **Physical activity in work(he mainly sits) & Height<176-180> ⇒ 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 (tumouros 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**



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	Non FATAL	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