Data Mining a.a. 2008/09

Introduzione

Data Mining

- Acronimo: DM
- Orario:
  - Martedì 14-16 aula C1
  - Giovedi 11-13 aula D1
- 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**
  - **Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to DATA MINING, Addison Wesley, ISBN 0-321-32136-7, 2006**
  - 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
  - U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (editors). Advances in Knowledge discovery and data mining, MIT Press, 1996.
  - 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 wiki del corso: http://www.cli.di.unipi.it/doku/doku.php/dm**

Questionario

- **Lauree specialistiche:**
  - **Laurea Inf. Per l'Economia e l'Azienda**
    - 1 anno 13
    - 2 anno 10
  - **Informatica Umanistica**
    - 1 anno 4
    - 2 anno 2
- **Tipi di Lauree di primo livello (provenienza):**
  - **Laurea in Informatica: 14 (Pisa), 1 (Firenze), 1 (messina), 1 (Bari), 1 (Camerino), 1 (Cagliari)**
  - **Economia (1), Statistica Economica (Siena). Marketing**

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 in dettaglio

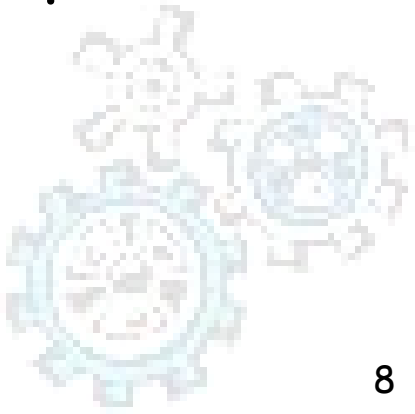
- Introduzione e Concetti Basici (2 ore)
  - Le applicazioni
  - Il processo di knowledge discovery
  - Esempi di estrazione (Evasione fiscale, Business Intelligence)
- Il processo di estrazione della conoscenza (4 ore)
  - Le fasi iniziali: preparazione e pulizia dei dati
- Introduzione alle tecniche di base (8+4 Ore)
  - Regole Associative
  - Alberi di decisione
  - Clustering
- Algoritmi di Base (8 + 4 ore)
  - Regole associative: algoritmo Apriori e varianti
  - Alberi di Decisione: C4.5
  - Clustering: K-Means

Contenuti del corso ADEC

- Interpretazione e valutazione della qualità della conoscenza estratta (6 ore)
  - Rassegna di strumenti commerciali
  - Uno standard metodologico: CRISP
  - Alcuni casi di studio: Custumer Segmentation, Basket Marketing Analysis, GeoMarketing…
- Presentazione progetti e seminari (2 ore)

Modalità di valutazione

- Verifica SCRITTA 50% SENZA LIBRI
  - O in due parti: intermedia e finale o complessiva
- Progetto (Analisi dei dati): 50%
  - 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.



Data Mining

Fosca Giannotti and Mirco Nanni

Pisa KDD Lab, ISTI-CNR & Univ. Pisa

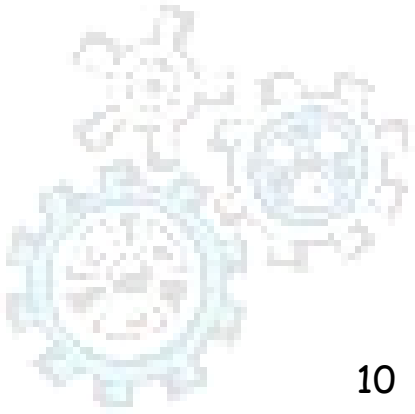
http://www-kdd.isti.cnr.it/

DIPARTIMENTO DI INFORMATICA - Università di Pisa

anno accademico 2007/2008

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

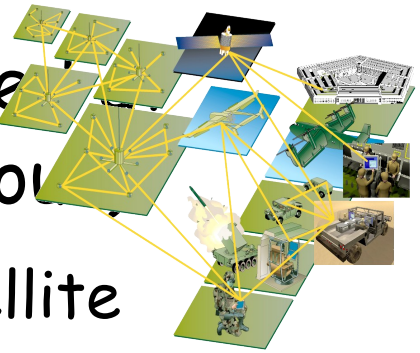
Why Mine Data? Commercial Viewpoint

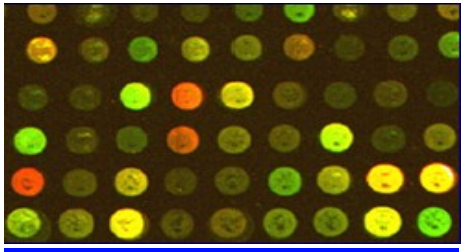
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

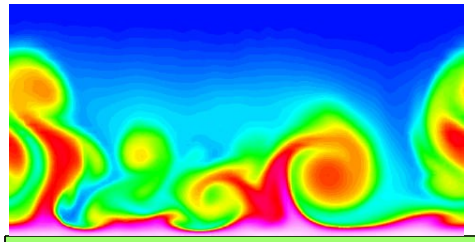
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

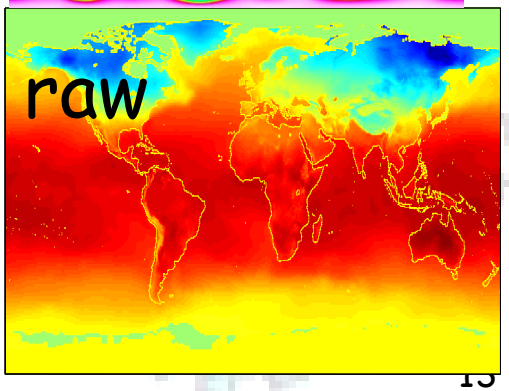
Why Mine Data? Scientific Viewpoint

- Data collected and store
  enormous speeds (GB/hou
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data



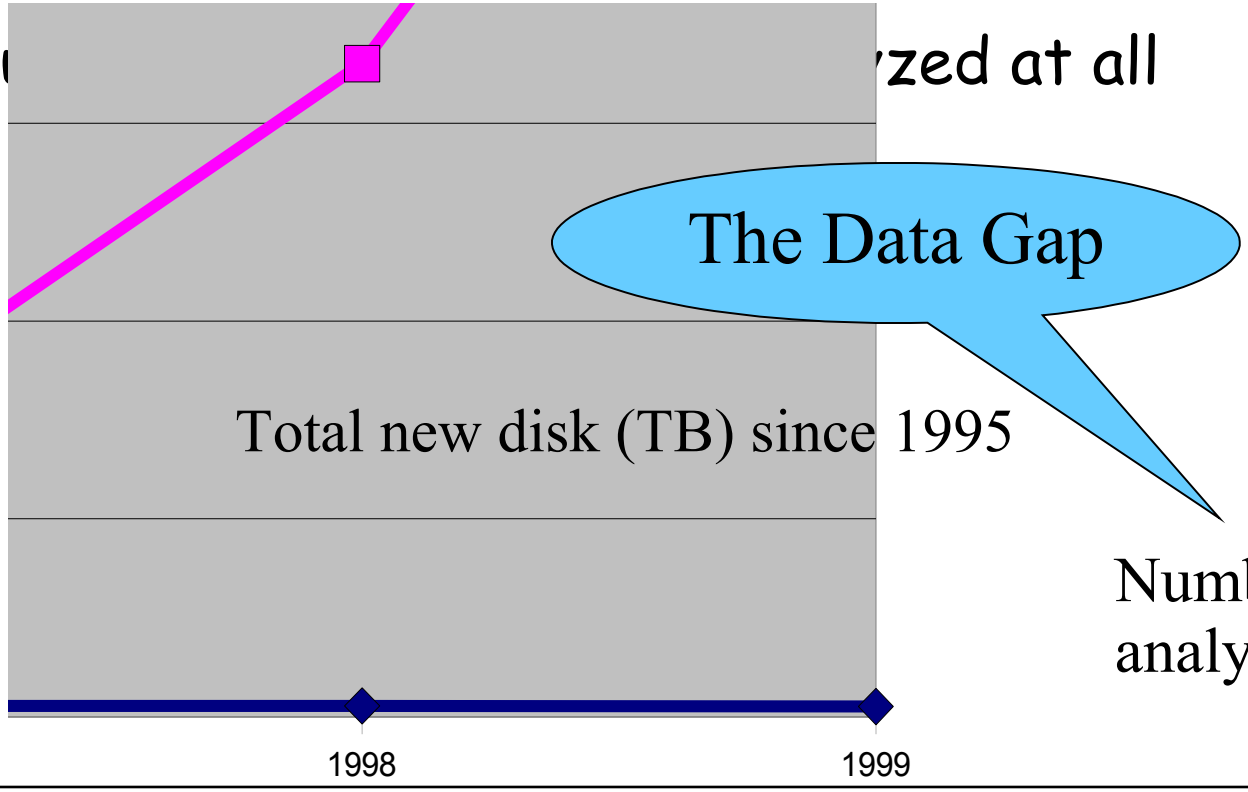






Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Mu zed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Motivations

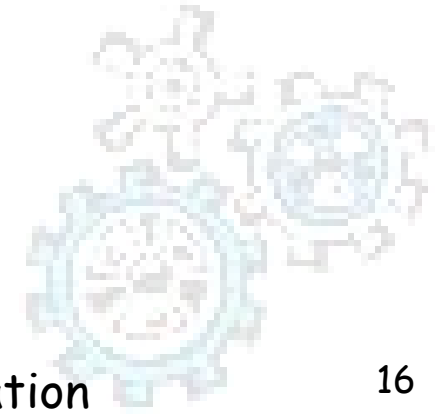
"Necessity is the Mother of Invention"

- *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.

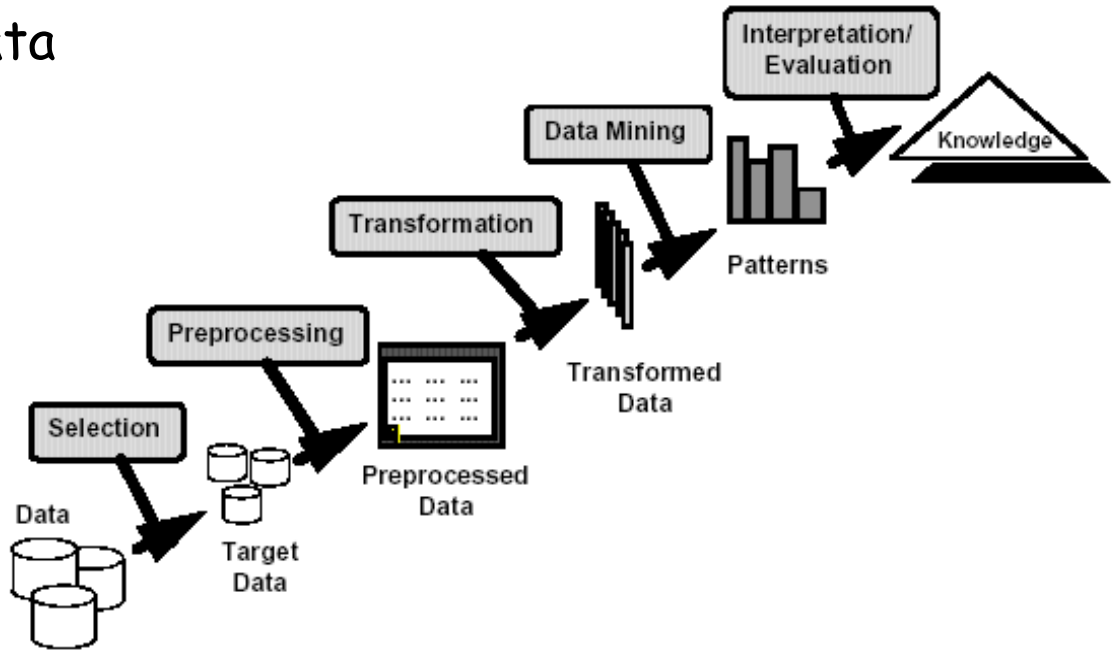
Why Data Mining

- **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



What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

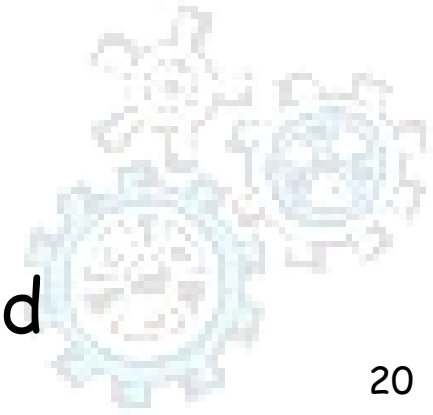
- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

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

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

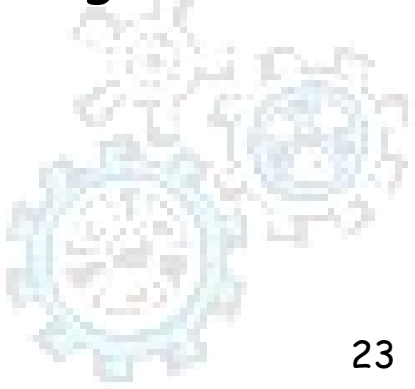
- **World Wide Web and Online Repositories**
  - email, news, messages
  - Web documents, images, video, etc.
  - link structure of of the hypertext from millions of Web sites
  - Web usage data (from server logs, network traffic, and user registrations)
  - online databases, and digital libraries

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**

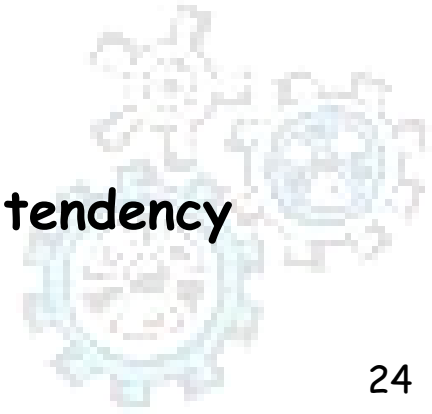
Market Analysis

- 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

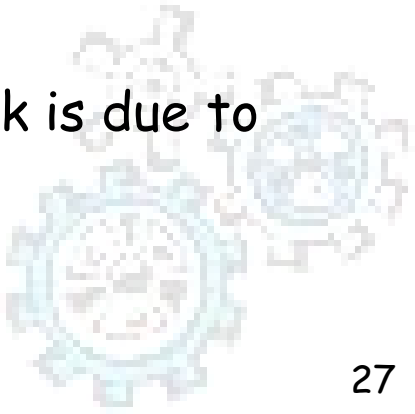
- **Finance planning and asset evaluation:**
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - trend analysis
- **Resource planning:**
  - summarize and compare the resources and spending
- **Competition:**
  - monitor competitors and market directions (CI: *competitive 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

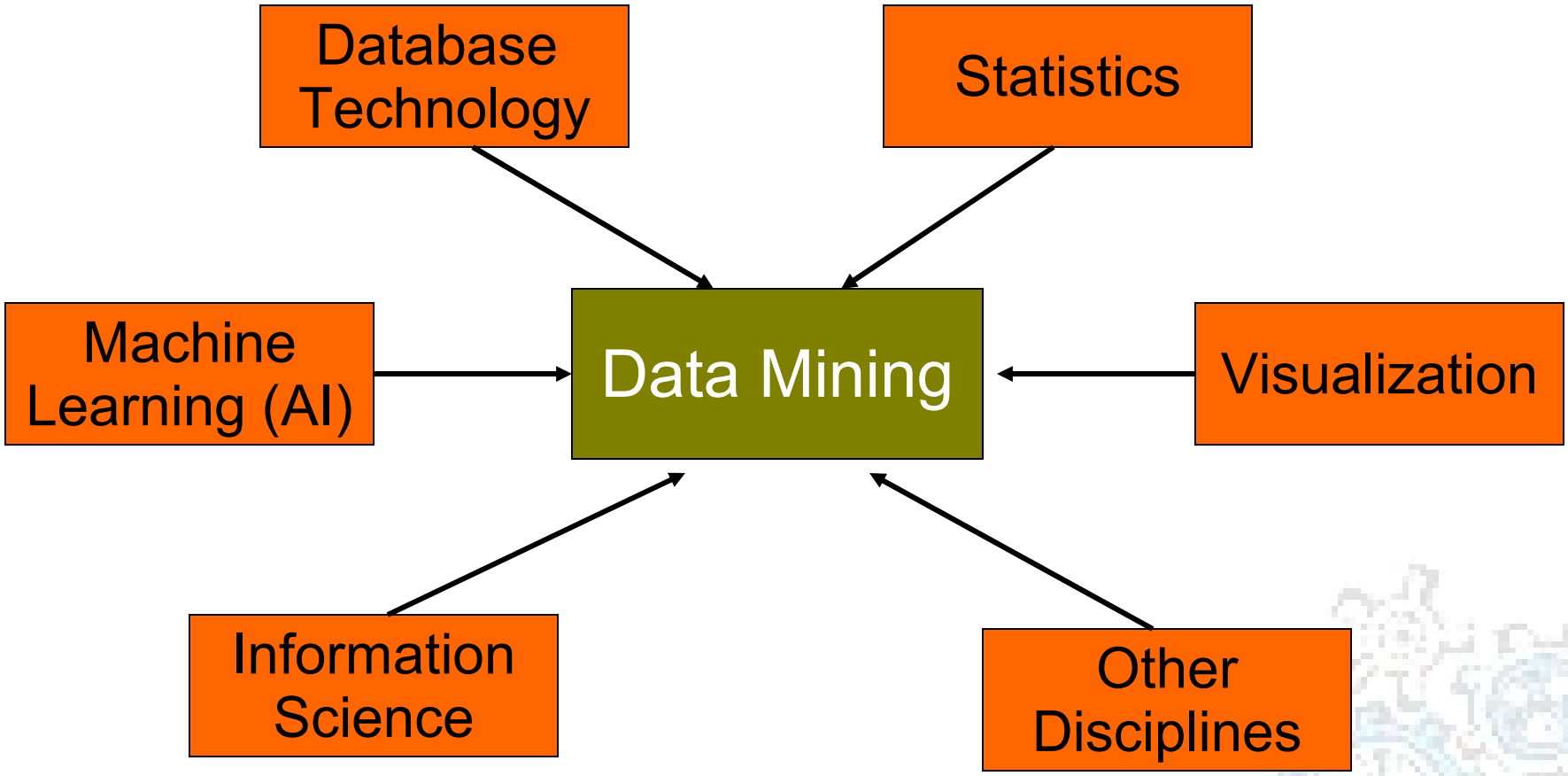
- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat.
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid 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!

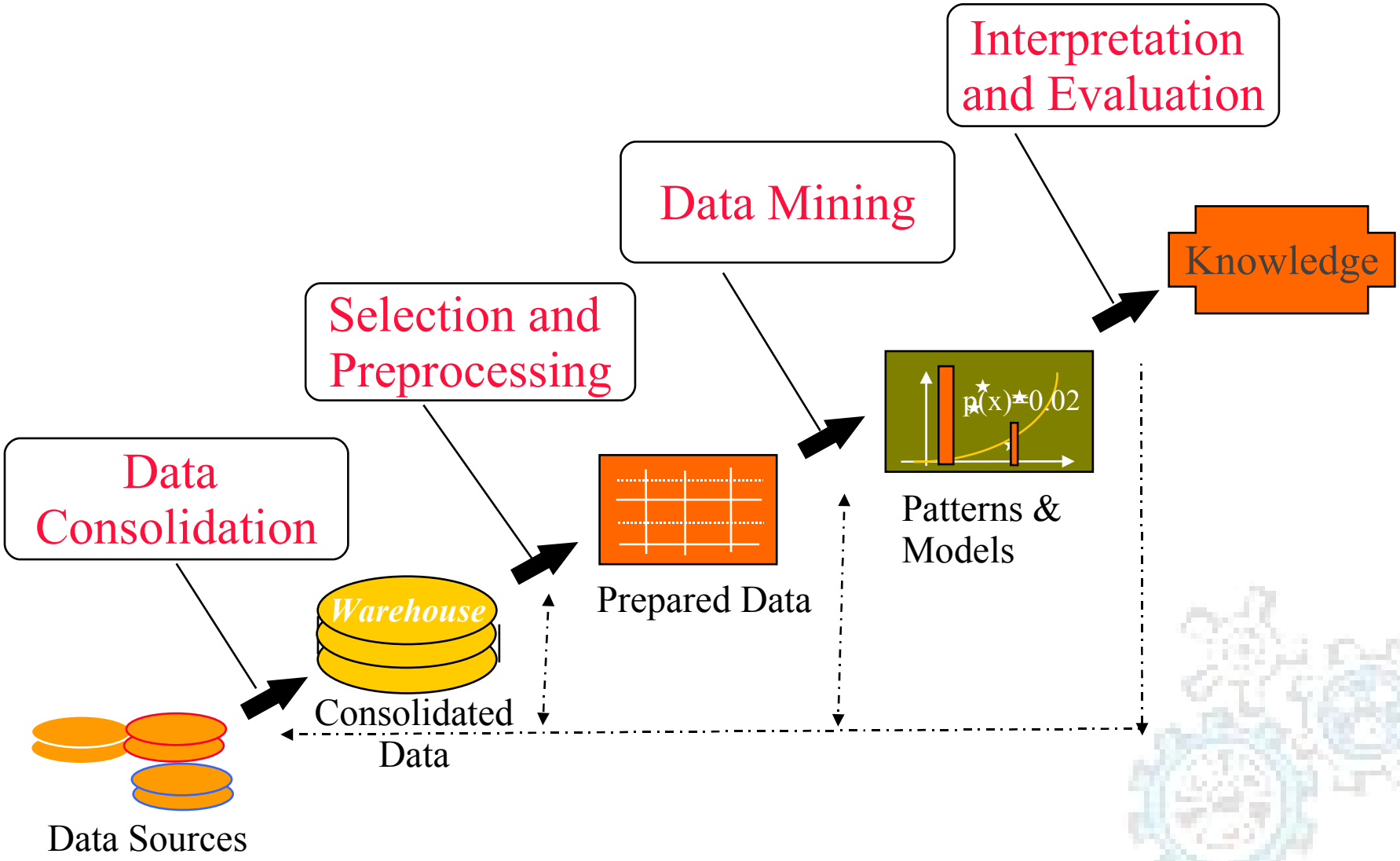
- The selection and processing of data for:
  - the identification of novel, accurate, and useful patterns, and
  - the modeling of real-world phenomena.

- Data mining is a major component of the KDD process - automated discovery of patterns and the development of predictive and explanatory models.

Data Mining: Confluence of Multiple Disciplines

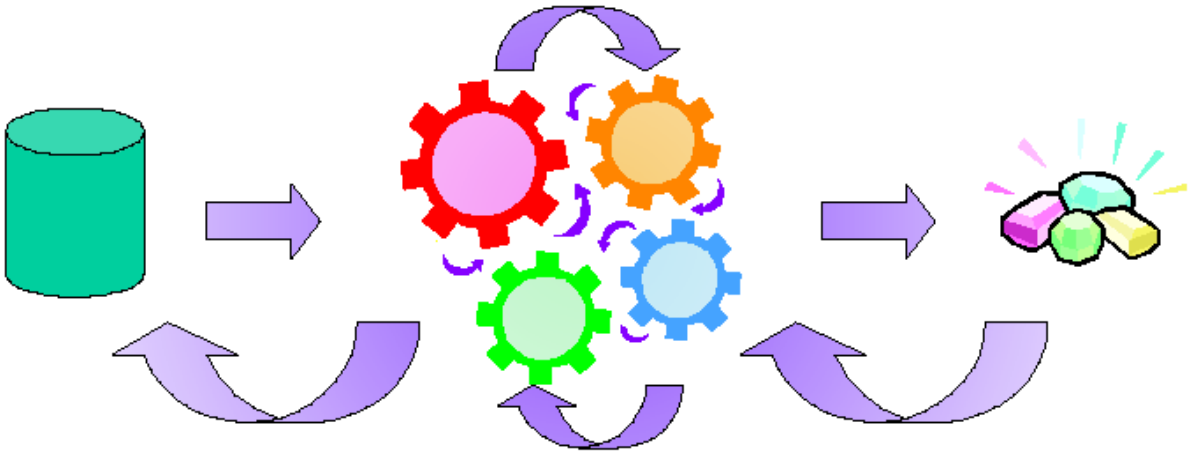


The KDD process



The KDD Process in Practice

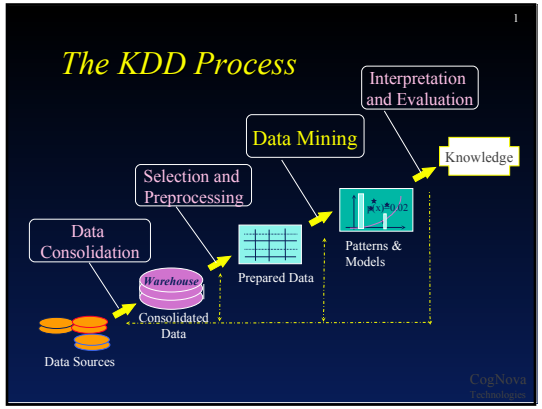
- KDD is an Iterative Process
  - art + engineering rather than science



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, remove redundant patterns*

The *virtuous cycle*



Problem

Knowledge

Identify Problem or Opportunity

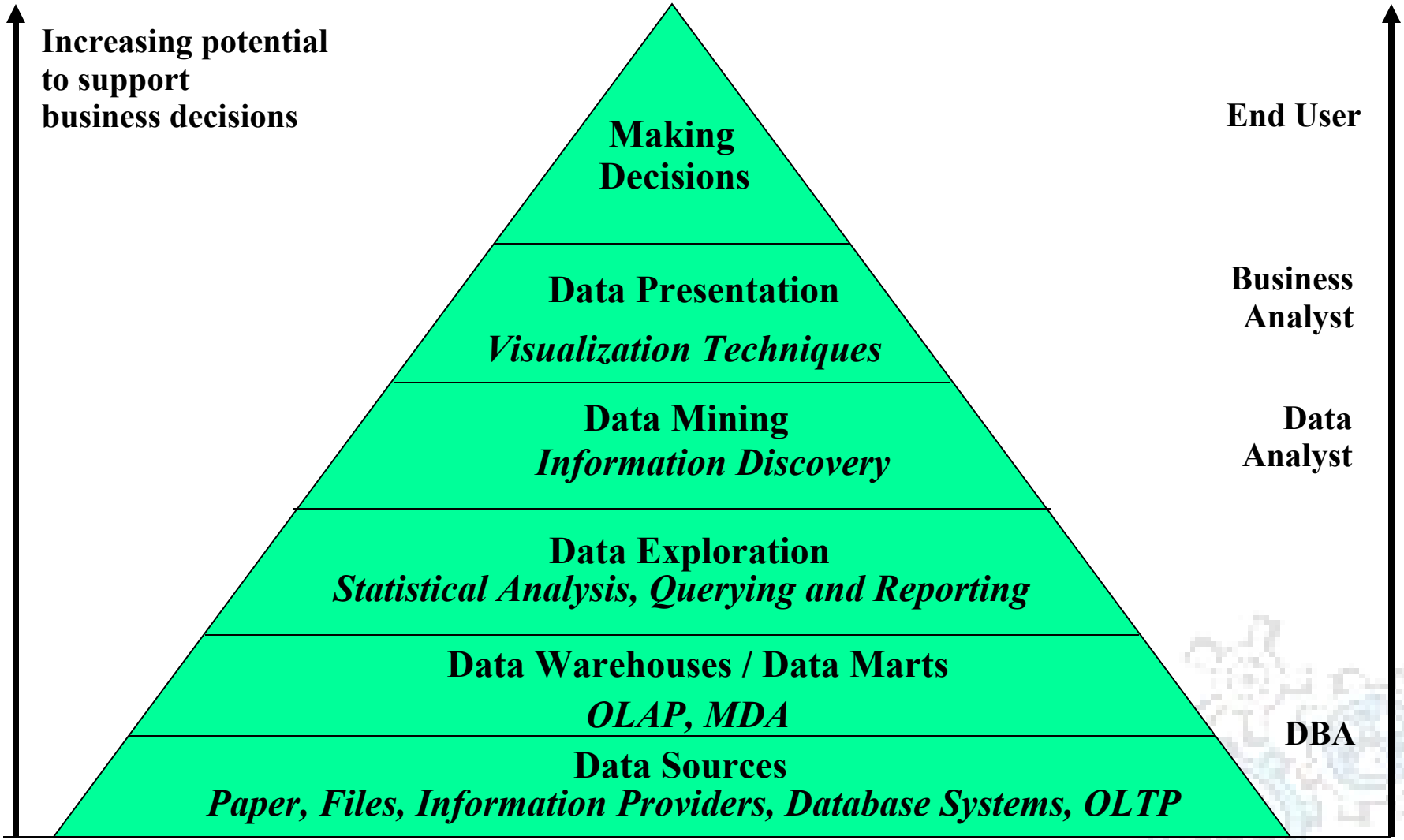
Act on Knowledge

Strategy

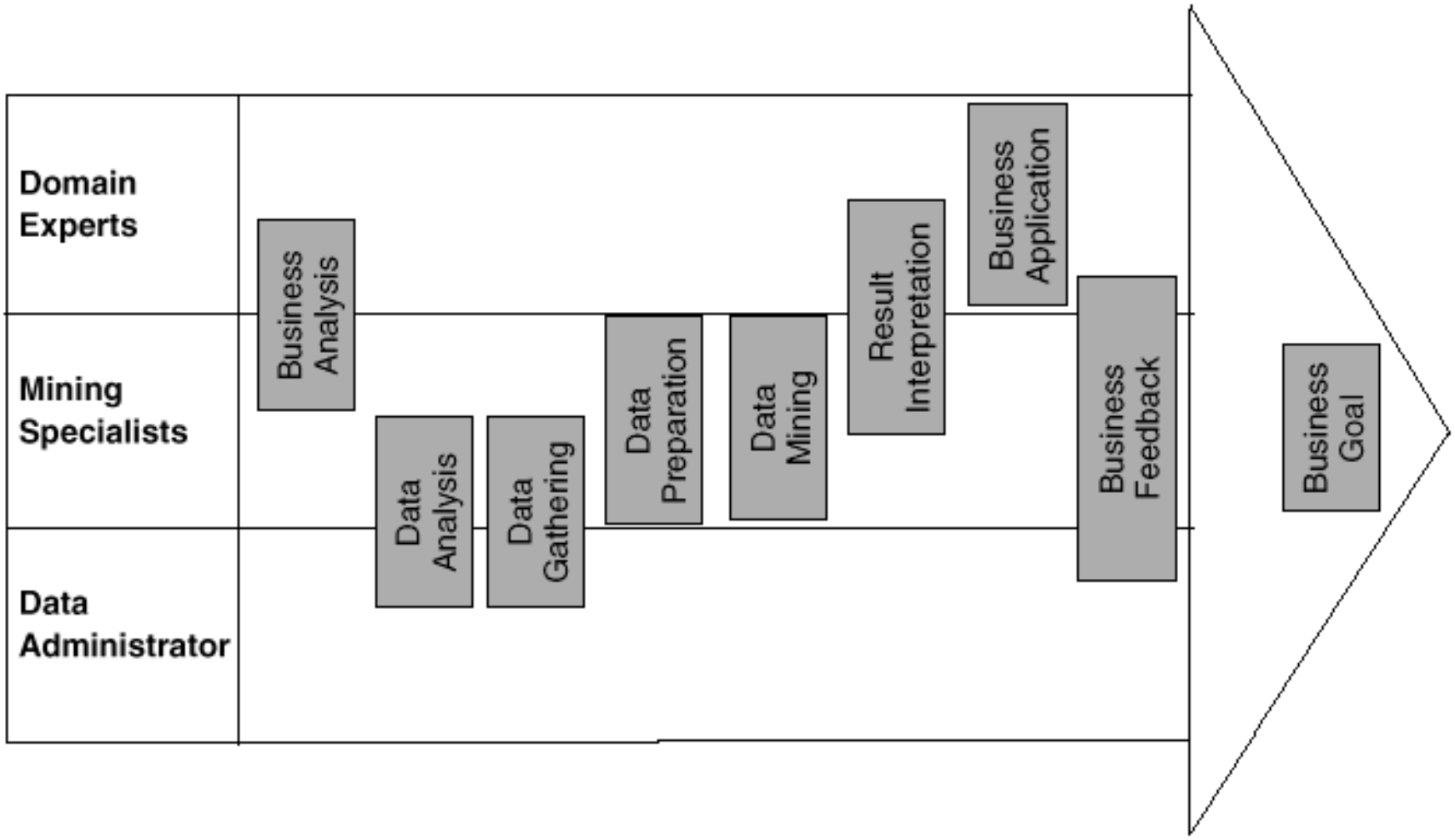
Measure effect of Action

Results

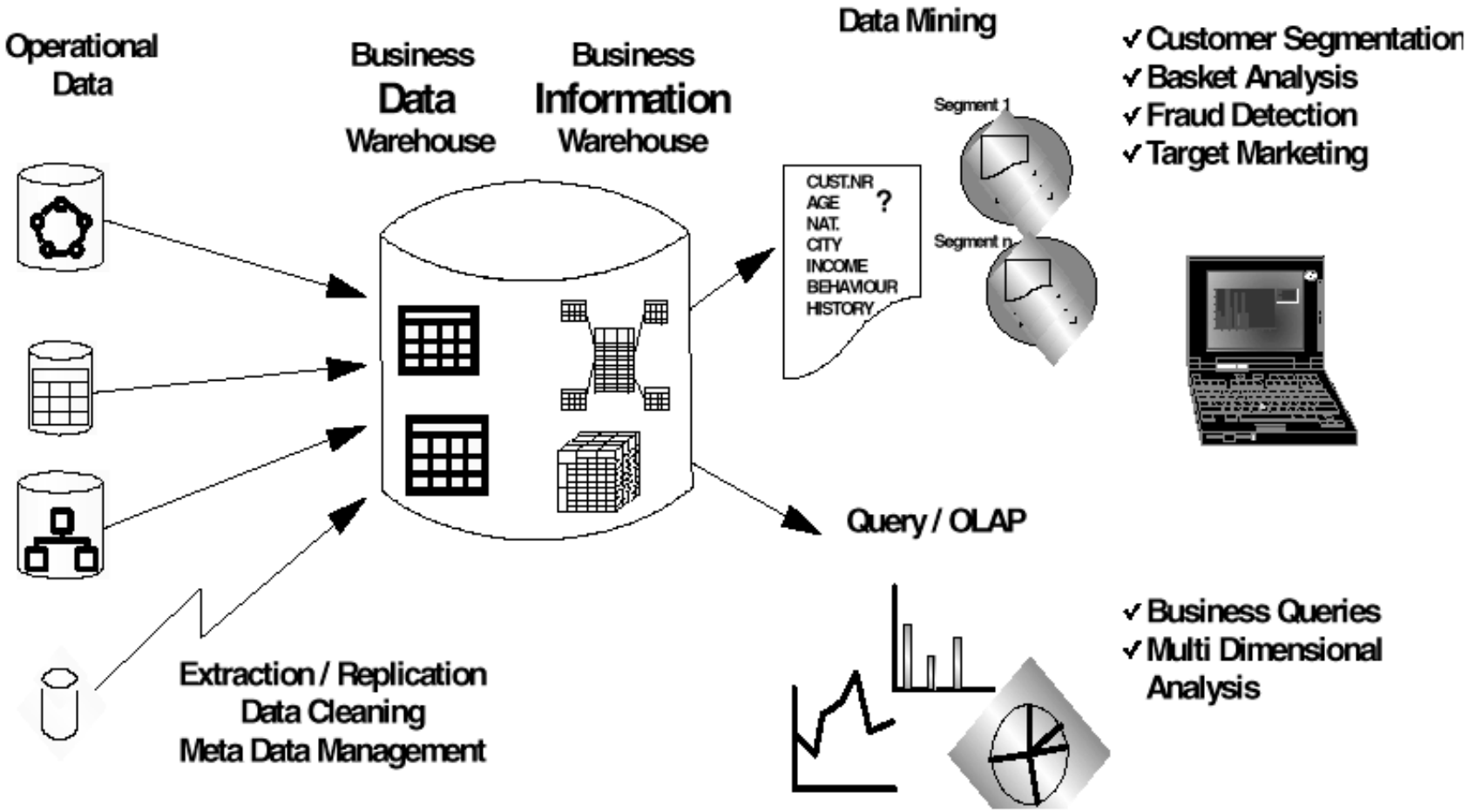
Data mining and business intelligence



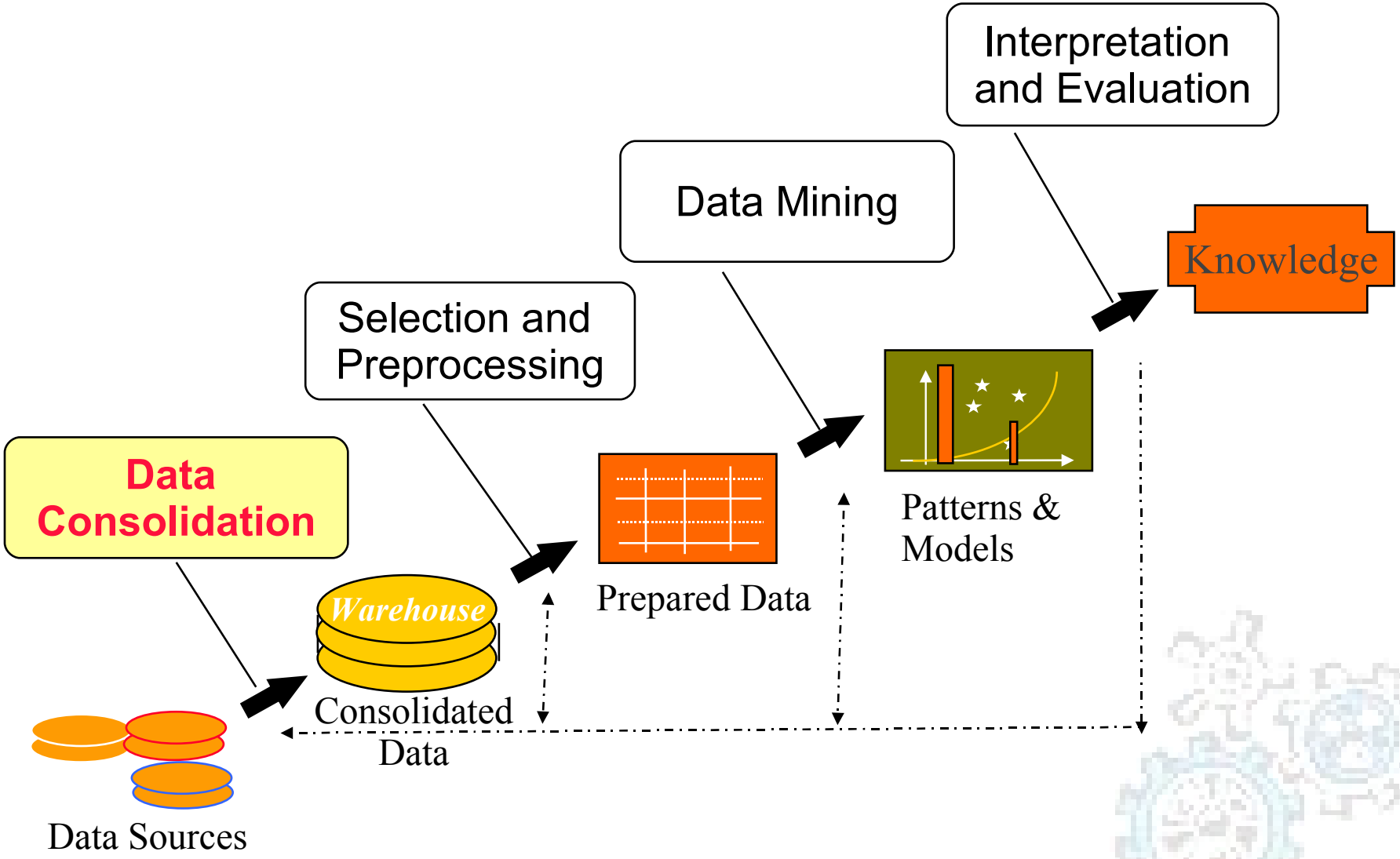
Roles in the KDD process



A business intelligence environment



The KDD process



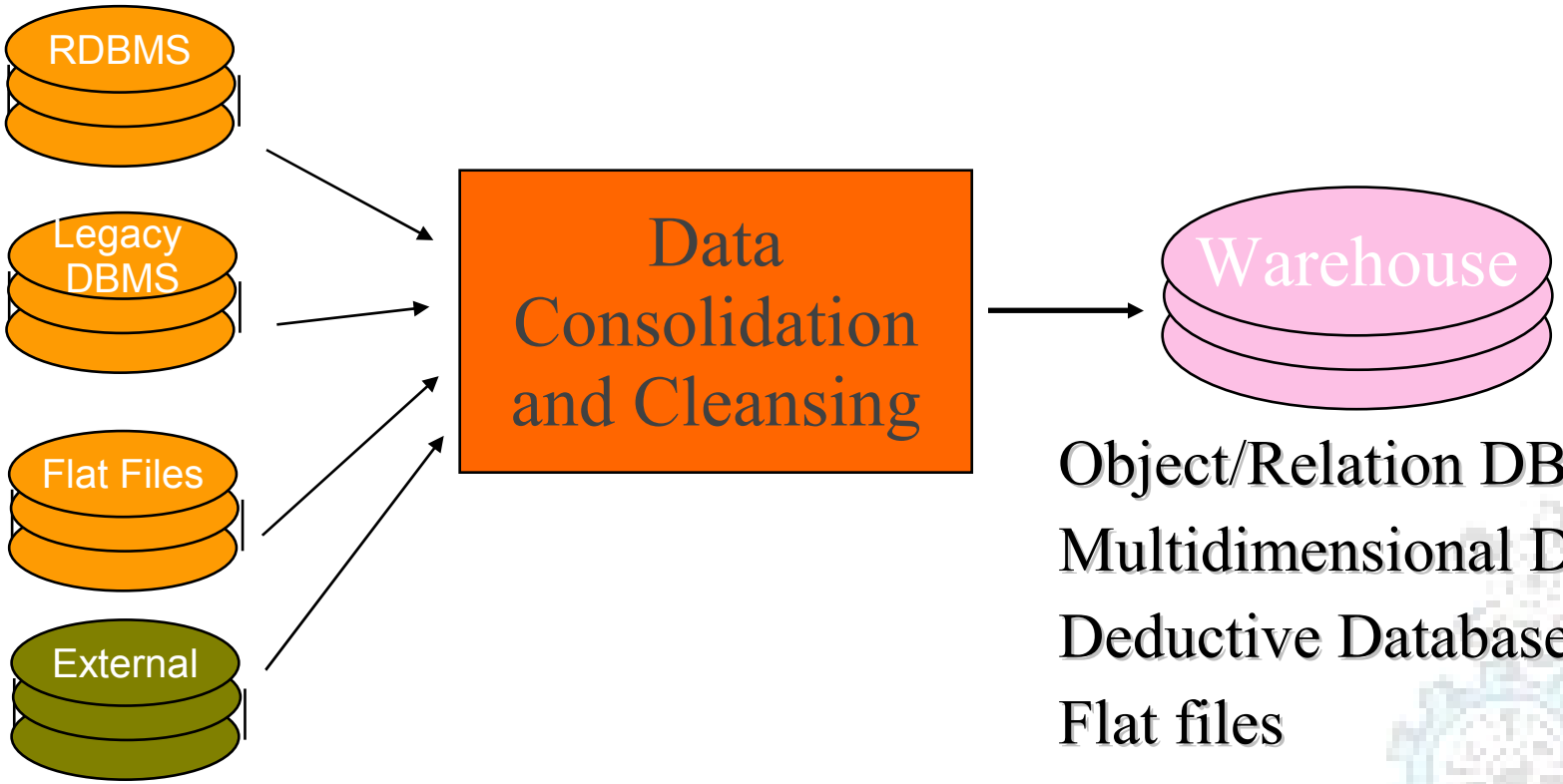
Data consolidation and preparation

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

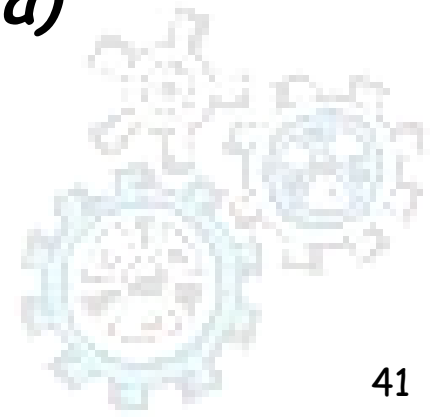
From data sources to consolidated data repository



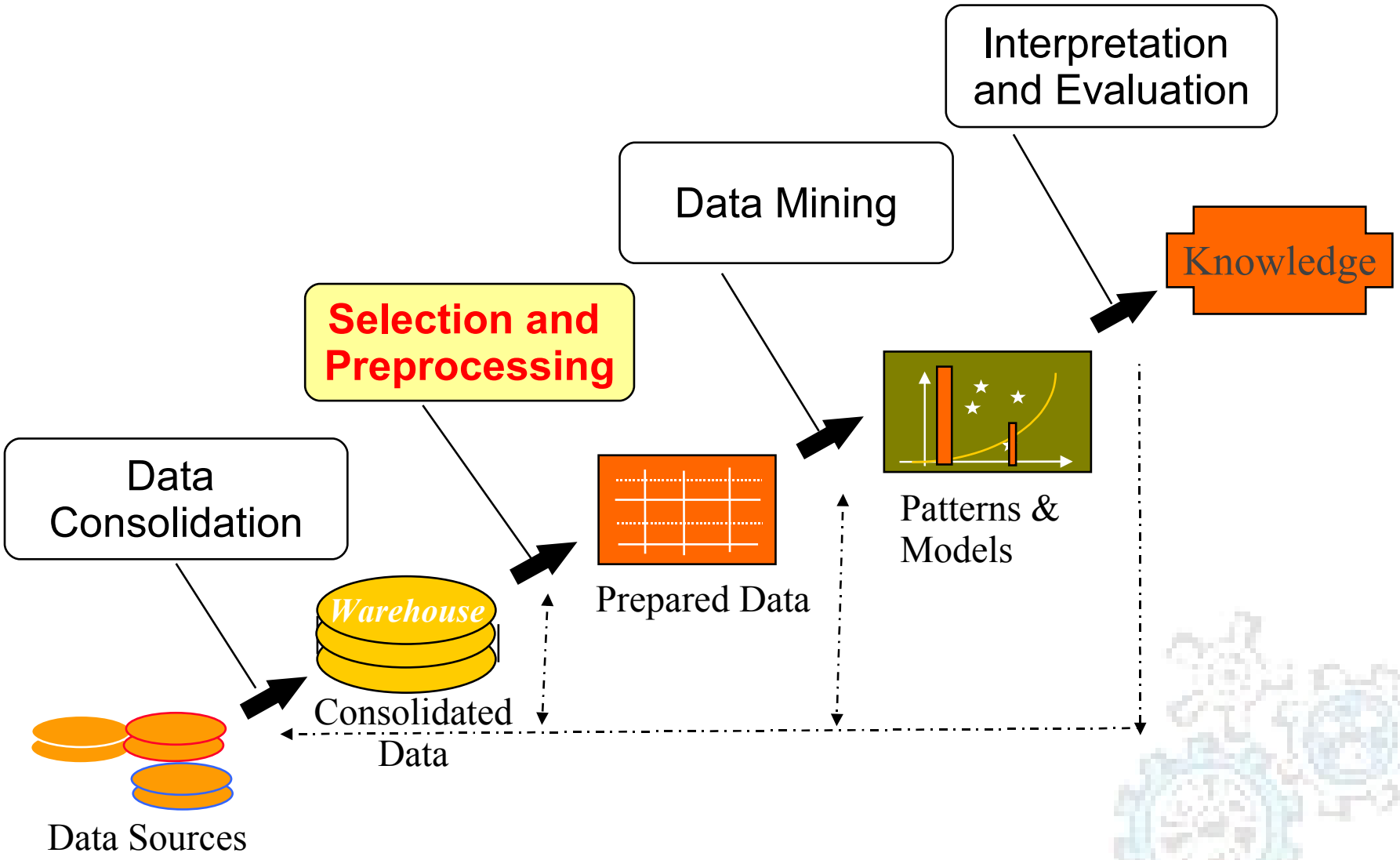
Object/Relation DBMS
Multidimensional DBMS
Deductive Database
Flat files

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 (=unbalanced data)*

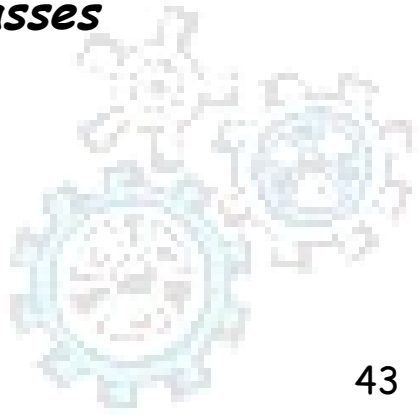


The KDD process

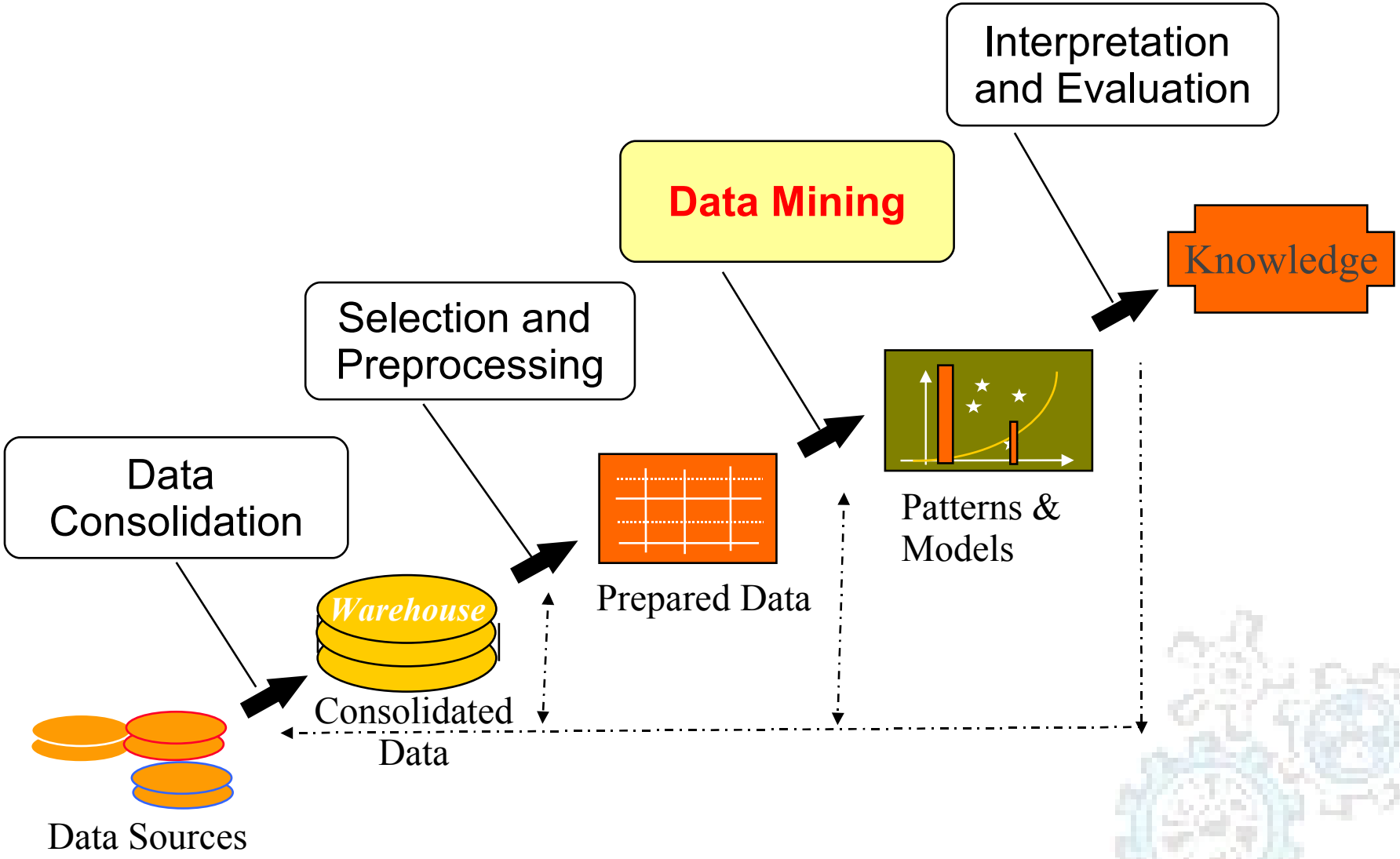


Data selection and preprocessing

- 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
  - group continuous numeric values into discrete *classes*
- Transform data
  - de-correlate and normalize values
  - map time-series data to static representation
- OLAP and visualization tools play key role



The KDD process

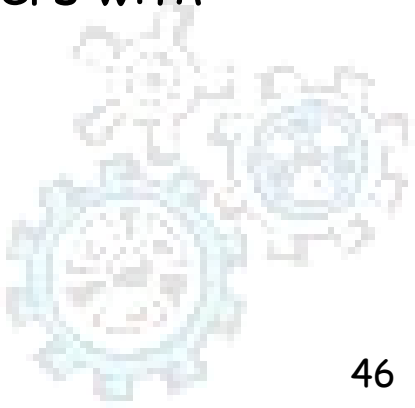


Data mining tasks and methods

- **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:
    - Which conditions (sun, humidity, wind) will lead to a weather suitable to play tennis?
    - Which customers will try to fraud the company?

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 Tasks

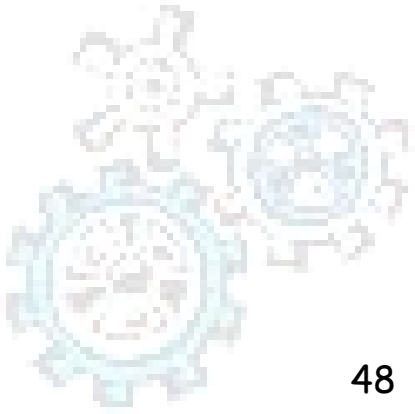
- Prediction Methods ↔ Directed KD
  - Use some variables to predict unknown or future values of other variables.
- Description Methods ↔ Undirected KD
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Etc.

covered by our course



Classification

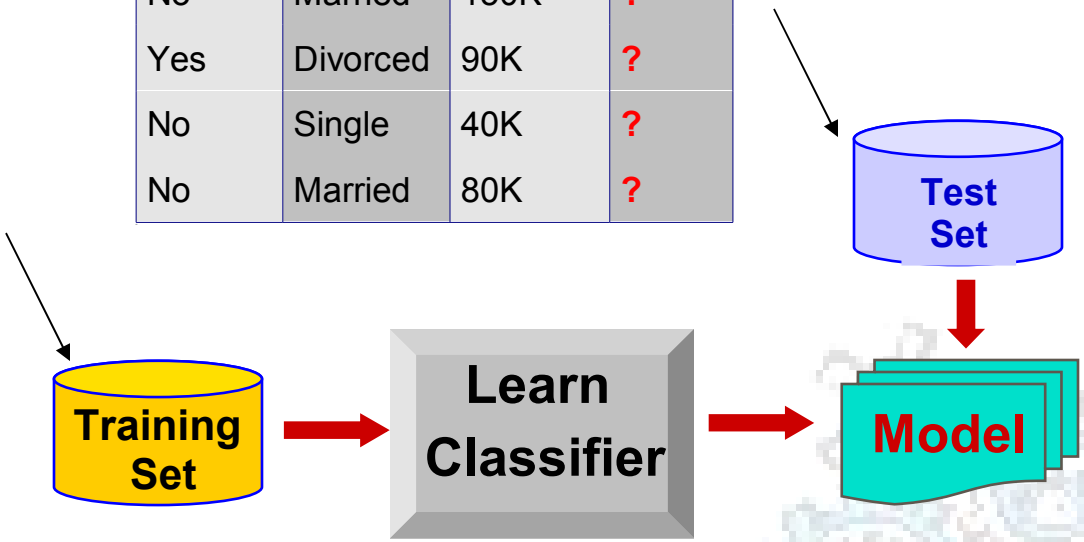
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the model accuracy
  - Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

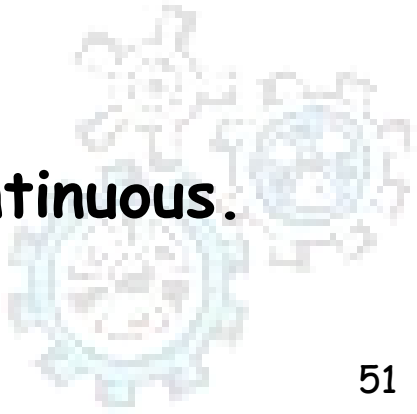
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.

- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

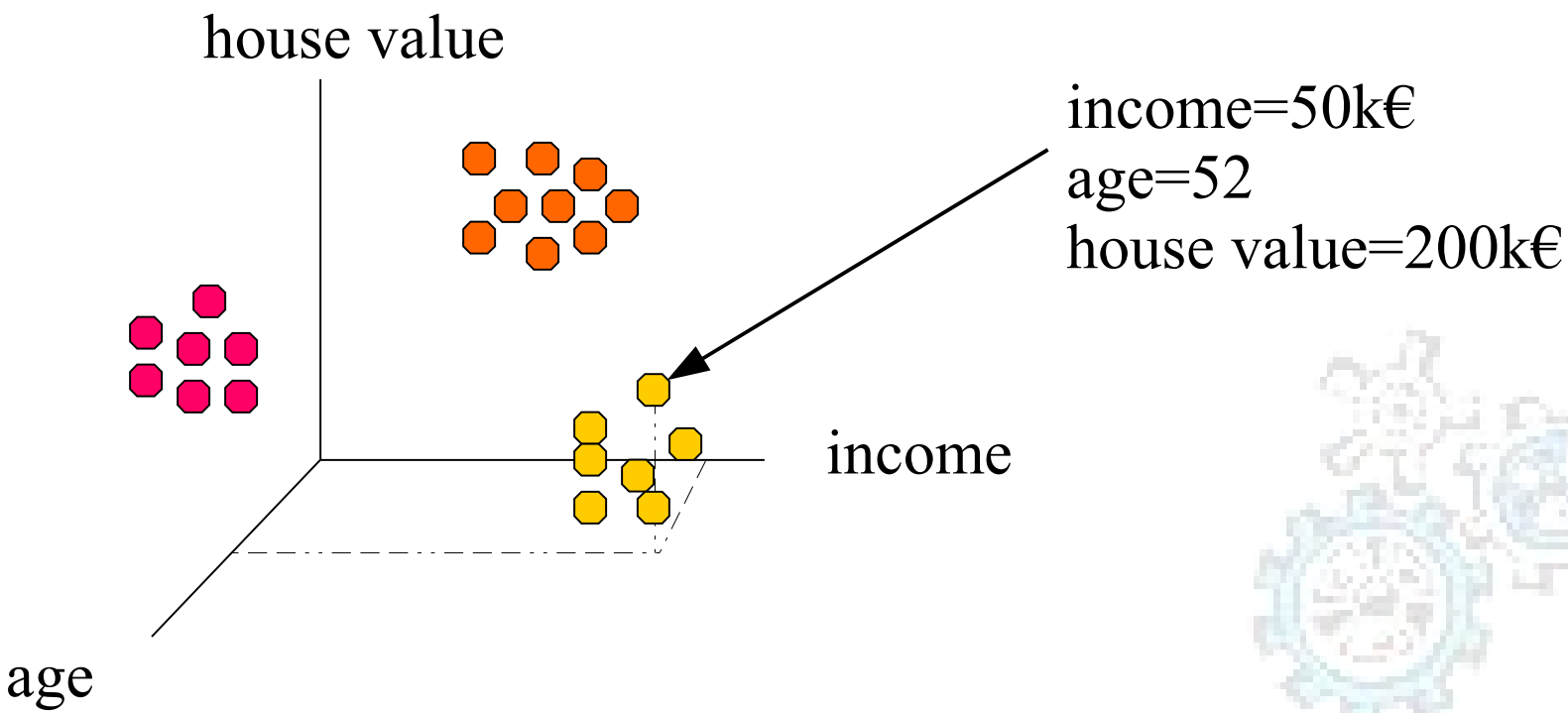


Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

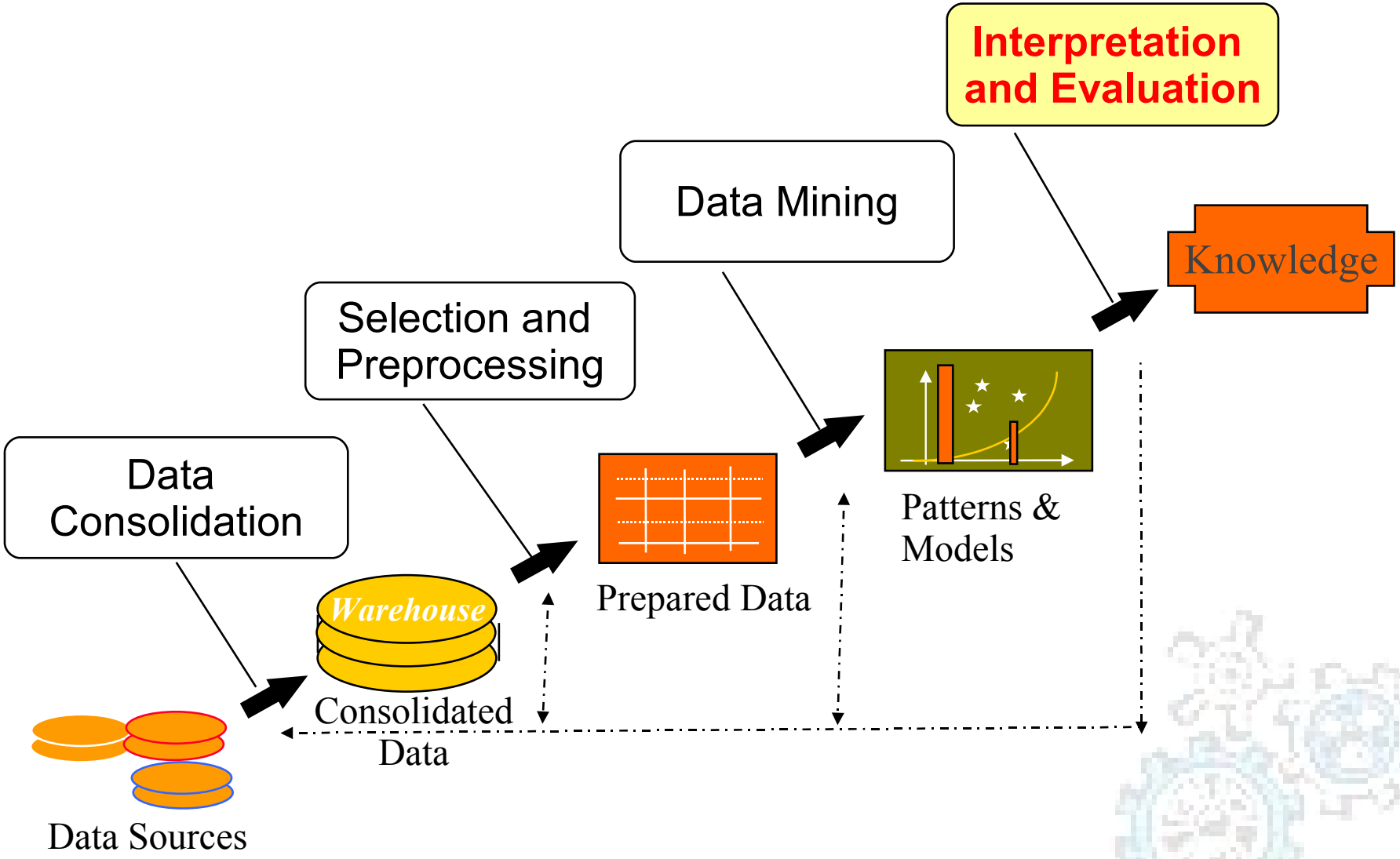
Intercluster distances are maximized



Association Rules
(or 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

The KDD process



Interpretation and evaluation

- Not all generated models are interesting
  - Need of *interestingness measures*
    - Objective vs. subjective

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)

Examples of DM projects

Competitive Intelligence

Fraud Detection,

Health care,

Traffic Accident Analysis,

Moviegoers database: a simple example at work

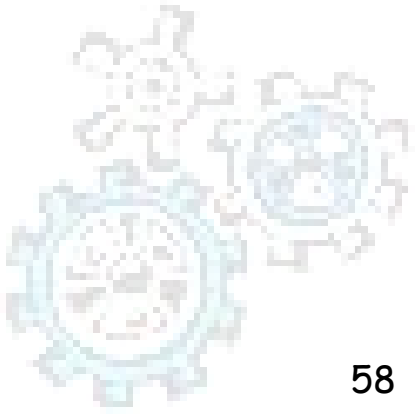
L'Oreal, a case-study on competitive intelligence:

Source: DM@CINECA

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

A small example

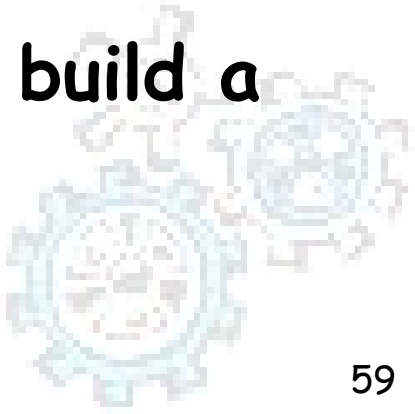
- Domain: technology watch - a.k.a. competitive intelligence
  - Which are the emergent technologies?
  - Which competitors are investing on them?
  - In which area are my competitors active?
  - Which area will my competitor drop in the near future?
- Source of data:
  - public (on-line) databases



The Derwent database

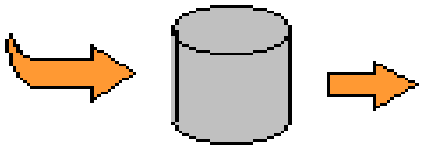
- Contains all patents filed worldwide in last 10 years
- Searching this database by keywords may yield thousands of documents
- Derwent documents are semi-structured: many long text fields

- 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

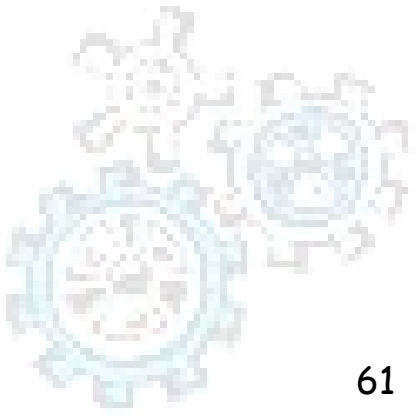
IC: H04N-005/262; C08J-005/18; G11B-005/704

PN: KR940043 B1 940120 DW9445

AB: ..... abstract .....

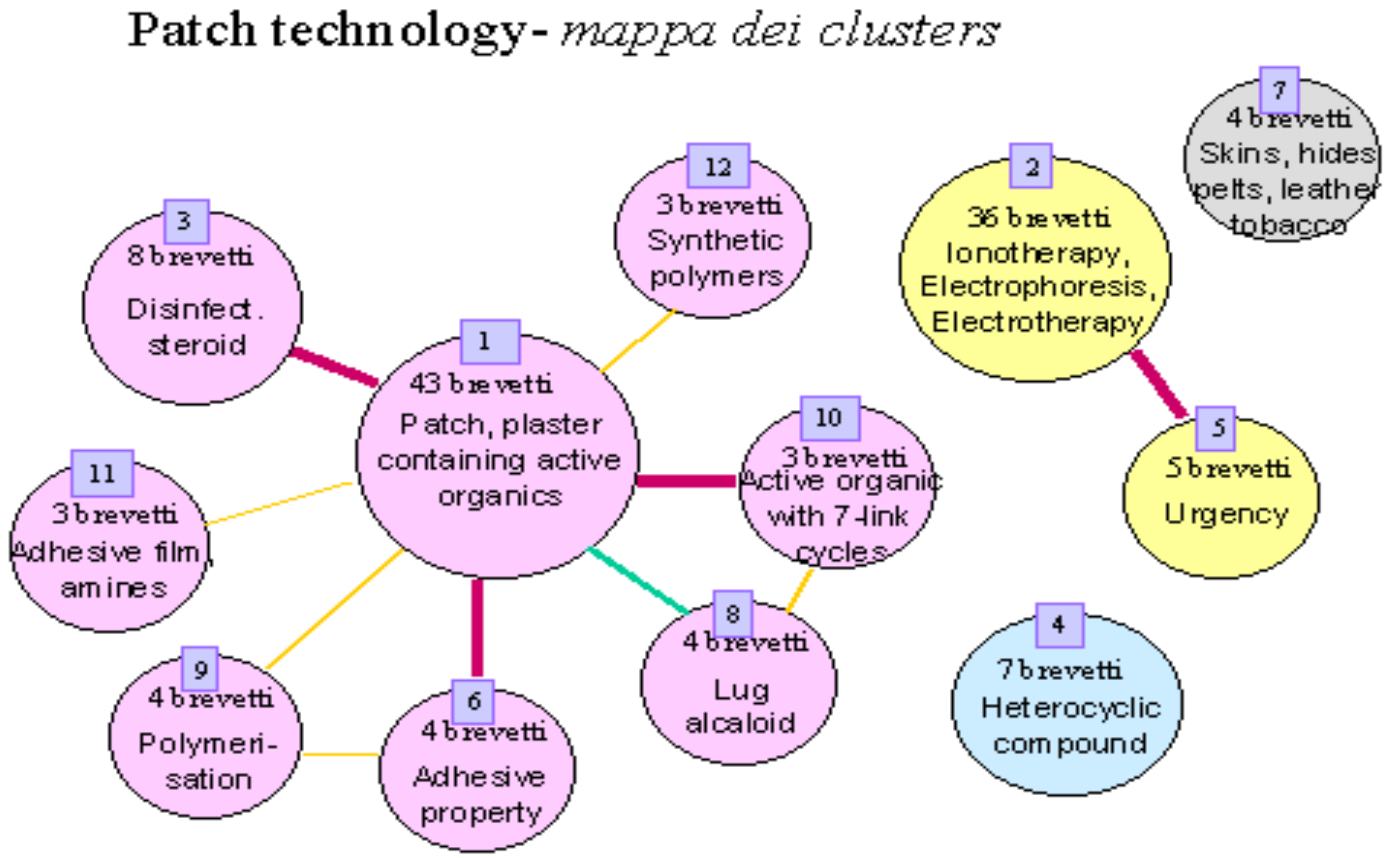
Example dataset

- Patents in the area: patch technology (cerotto medicale)
  - 105 companies from 12 countries
  - 94 classification codes
  - 52 Derwent codes



Clustering output

- Clusters patents with similar sets of keywords in the same group
- Groups are linked if they share some keywords



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

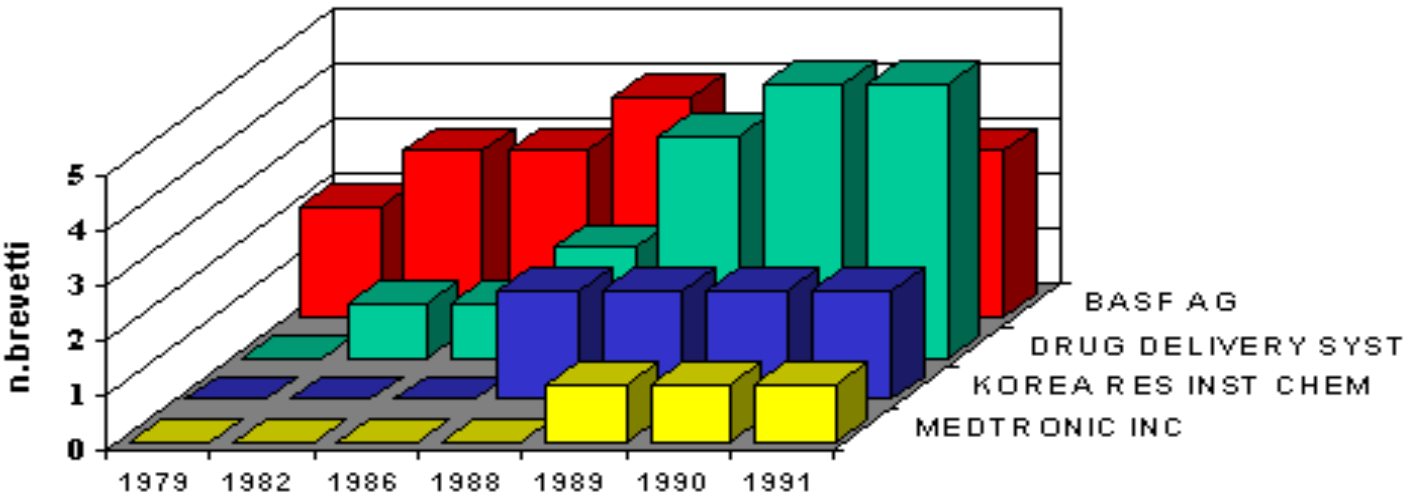
Società proprietarie:

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

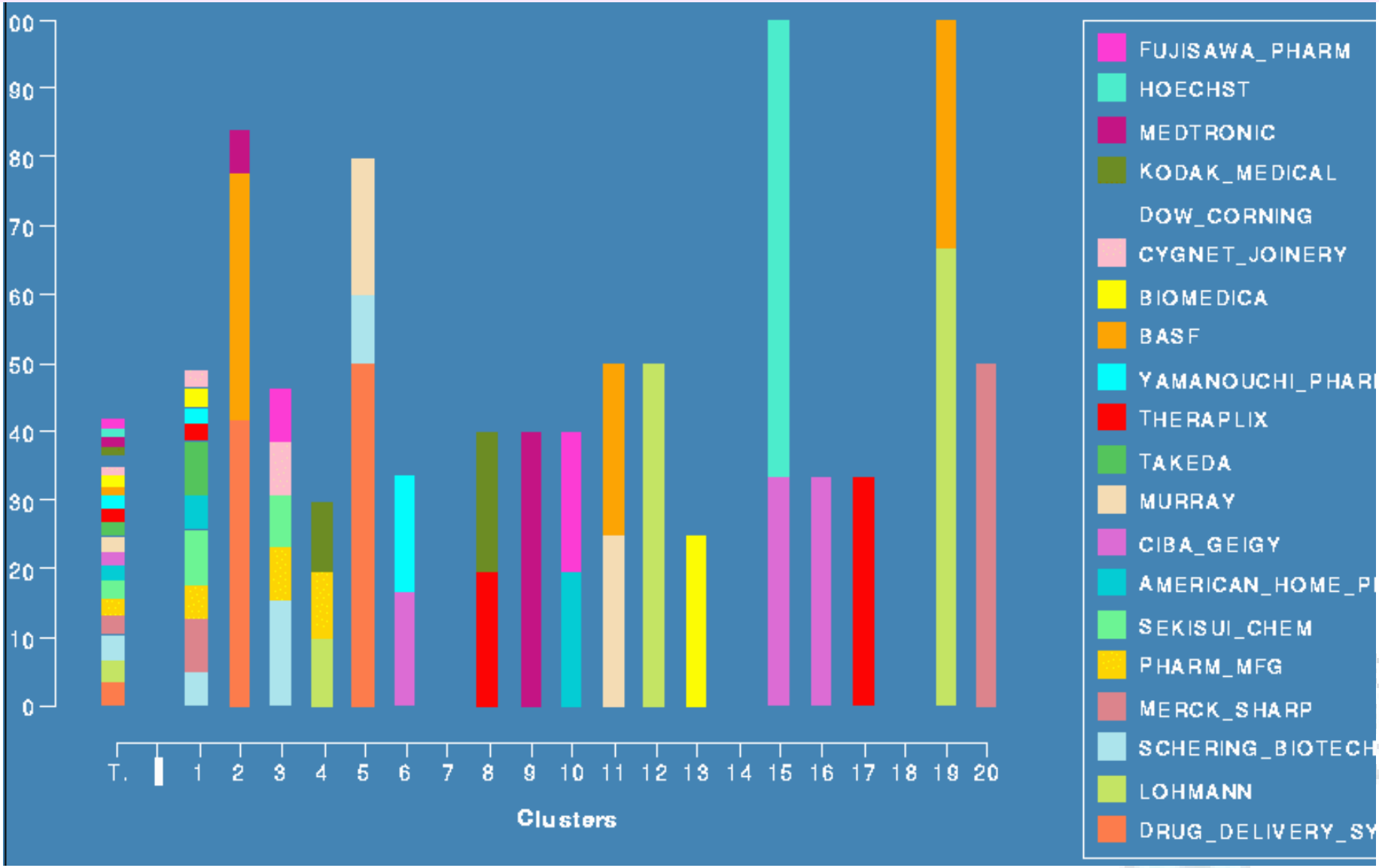
Anno	n. brevetti
1979	2
1982	4
1986	4
1988	8
1989	10
1990	11
1991	11

Zoom on cluster 2 - profiling competitors

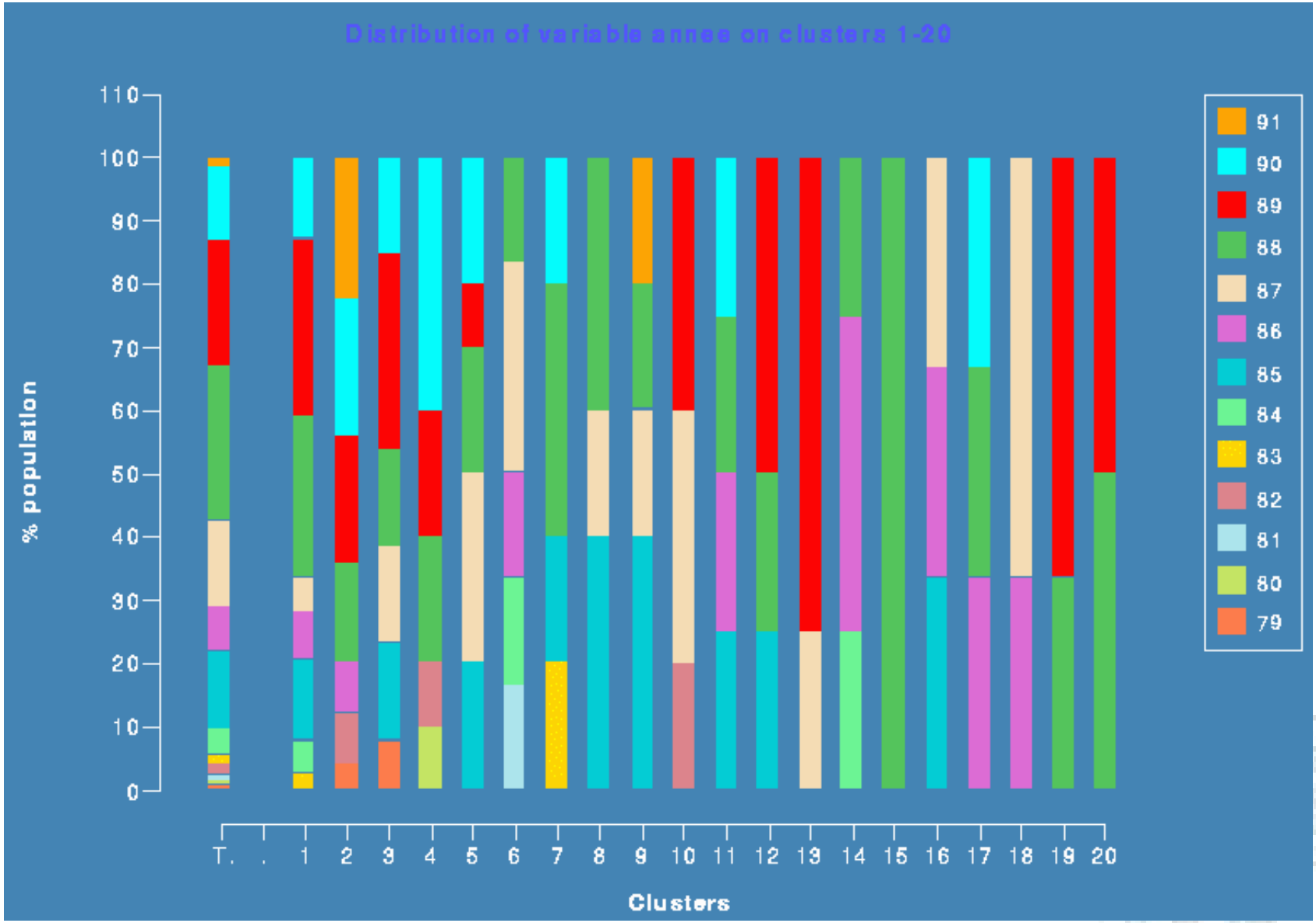
Patch technology- *cluster n.2 -*
attività della concorrenza nel tempo



Activity of competitors in the clusters



Temporal analysis of clusters



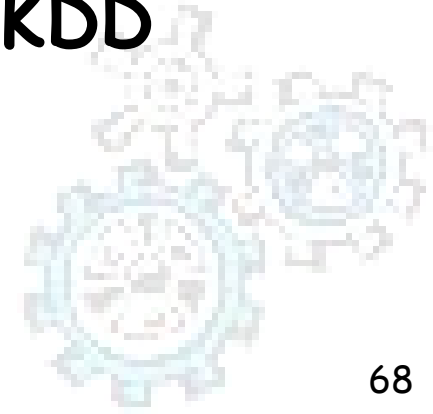
Atherosclerosis prevention study

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:

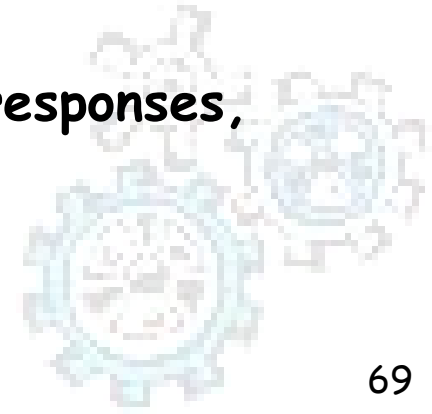
- The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.

- 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

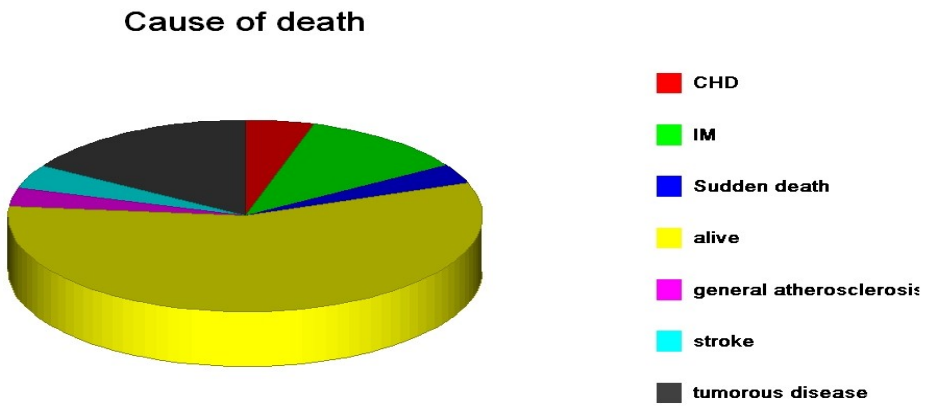
Data from Entry and Exams		
General characteristics	**Examinations**	**habits**
Marital status	Chest pain	Alcohol
Transport to a job	Breathlesness	Liquors
Physical activity in a job	Cholesterol	Beer 10
Activity after a job	Urine	Beer 12
Education	Subscapular	Wine
Responsibility	Triceps	Smoking
Age		Former 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

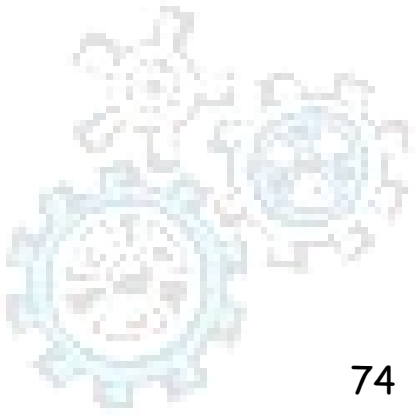
Are there strong relations concerning death cause?

General characteristics (?) ⇒ Death cause (?)

Examinations (?) ⇒ Death cause (?)

Habits (?) ⇒ Death cause (?)

Combinations (?) ⇒ Death cause (?)



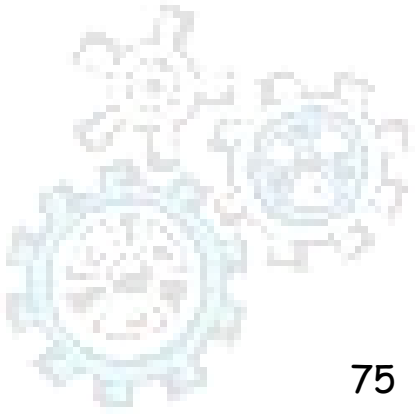
Example of extracted rules

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**