Introduction to Data Mining

José Hernández-Orallo

Dpto. de Sistemas Informáticos y Computación

Universidad Politécnica de Valencia, Spain

jorallo@dsic.upv.es

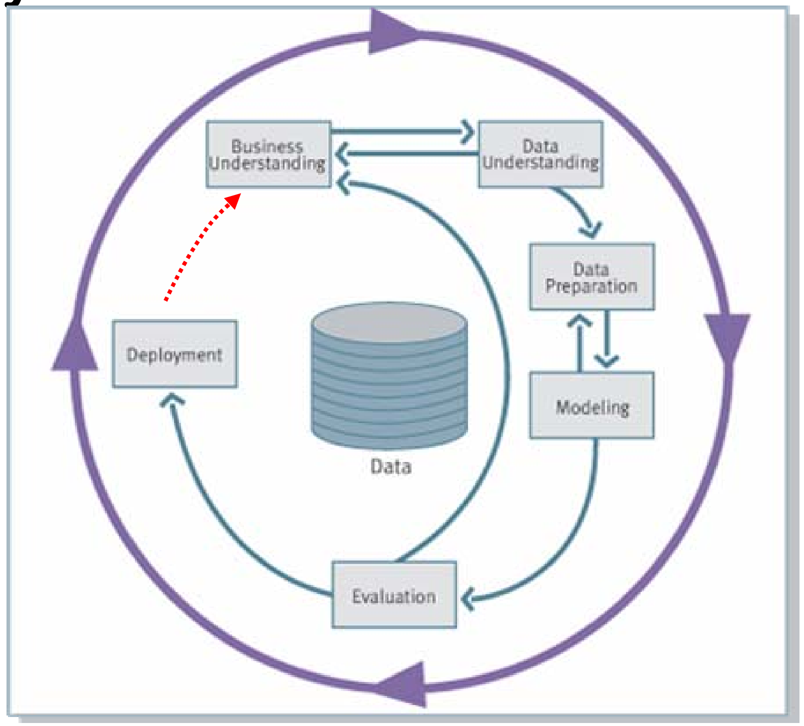
Roma, 14-15th May 2009

Outline

- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

CRISP-DM Methodology

- **CRISP-DM (www.crisp-dm.org) (*CRoss-Industry Standard Process for Data Mining*)**
  - **A company consortium (initially under the funding of the European Commission), which includes SPSS, NCR and DaimlerChrysler.**



CRISP-DM Methodology

- Business Understanding:
  - Understand the project goals and requirements from a business perspective. Substages:
    - establishment of business objectives (initial context, objectives and success criteria),
    - evaluation of the situation (resource inventory, requirements, assumptions and constraints, risks and contingences, terminology and costs and benefits),
    - establishment of the data mining objectives (data mining objectives and success criteria) and,
    - generation of the project plan (project plan and initial evaluation of tools and techniques).

CRISP-DM Methodology

- Data understanding:
  - Collect and familiarise with data, identify the data quality problems and see the first potentialities or data subsets which might be interesting to analyse (according the business objectives from the previous stage). Substages:
    - initial data gathering (gathering report),
    - data description (description report),
    - data exploration (exploration report) and
    - data quality verification (quality information).

CRISP-DM Methodology

- Data preparation:
  - The goal of this stage is to obtain the “minable view”. Here we find: integration, selection, cleansing and transformation. Substages:
    - data selection (inclusion/exclusion reasons),
    - data cleansing (data cleansing report),
    - data construction (derived attributes, generated records),
    - data integration (mixed data) and
    - data formatting (reformatted data).

CRISP-DM Methodology

- Data modelling:
  - It is the application of modelling techniques or data mining to the previous minable views. Substages:
    - selection of the modelling technique (modelling technique, modelling assumptions),
    - evaluation design (test design),
    - model construction (chosen parameters, models, model description) and
    - model evaluation (model measures, revision of the chosen parameters).

CRISP-DM Methodology

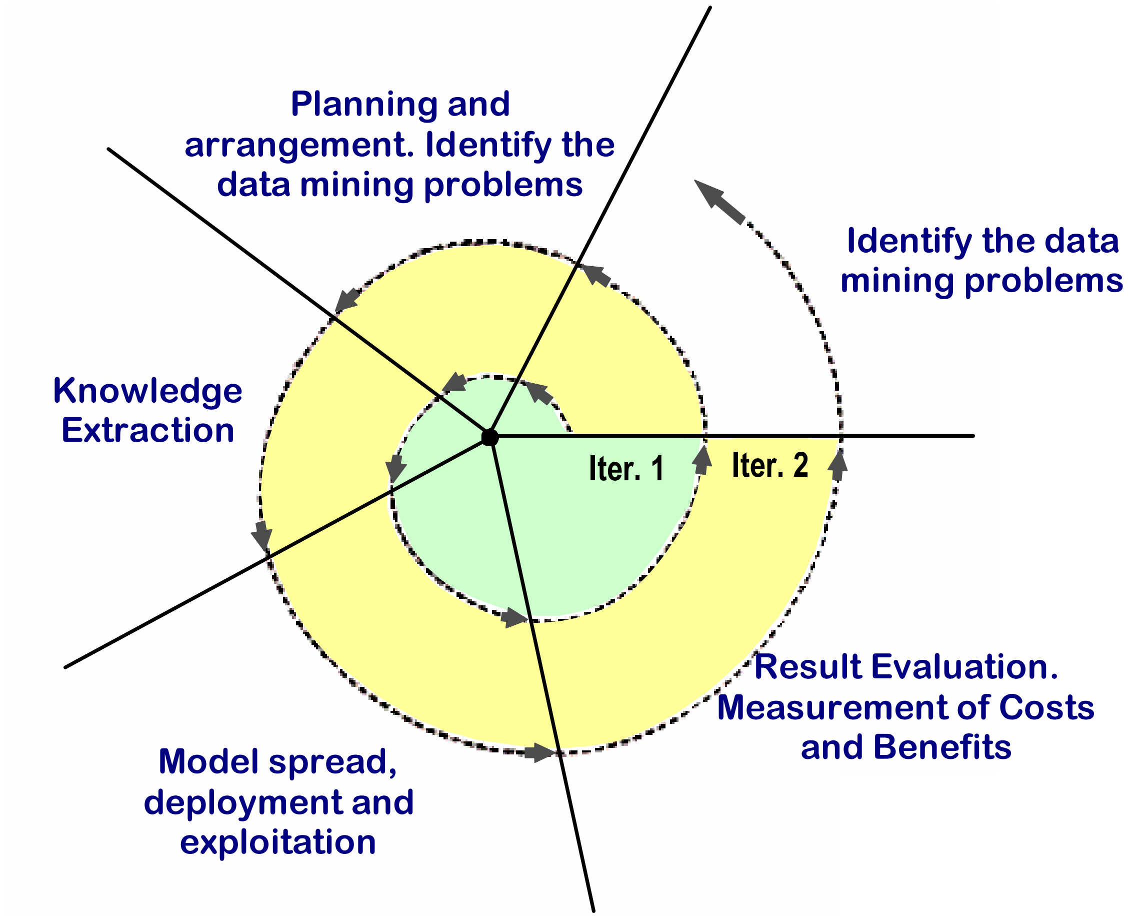
- **Evaluation:**
  - **It is necessary to evaluate (from the view point of the goal) the models of the previous stage. In other words, if the model is useful to answer some the business requirements. Substages:**
    - **result evaluation (evaluation of the data mining results, approved models),**
    - **revise the process (process revision) and,**
    - **establishment of the following steps (list of possible actions, decisions).**

CRISP-DM Methodology

- **Deployment:**
  - **The idea is to exploit the potential of the extracted models, integrate them in the decision-making processes of the organisation, spread reports about the extracted knowledge, etc. Substages:**
    - **deployment planning (deployment plan),**
    - **monitoring and maintenance planning (monitoring and maintenance plan),**
    - **generation of the final report (final report, final presentation) and,**
    - **project revision (documentation of the experience).**

CRISP-DM Methodology

- **Progressive implementation on an organisation:**



Tools

Clementine

PRW

Enterprise Miner

DataMIND

GainSmarts

Intelligent Miner for data

MineSet

Mi

Model 1

DARWIN

ModelQuest Enterprise™

pcOLPARS

CART™ for Windows®

wizwhy® For windows 95 / NT

NeuroShell® 2

S-PLUS

COGNOS® Tools That Build Business®

See5

Tools

- Links to Commercial and non-commercial DM Software:
  - http://www.kdnuggets.com/software/index.html
- Free:
  - WEKA (http://www.cs.waikato.ac.nz/~ml/weka/) (Witten & Frank 1999, 2006)
  - Rproject: free tool for statistical analysis (http://www.R-project.org/)

Tools

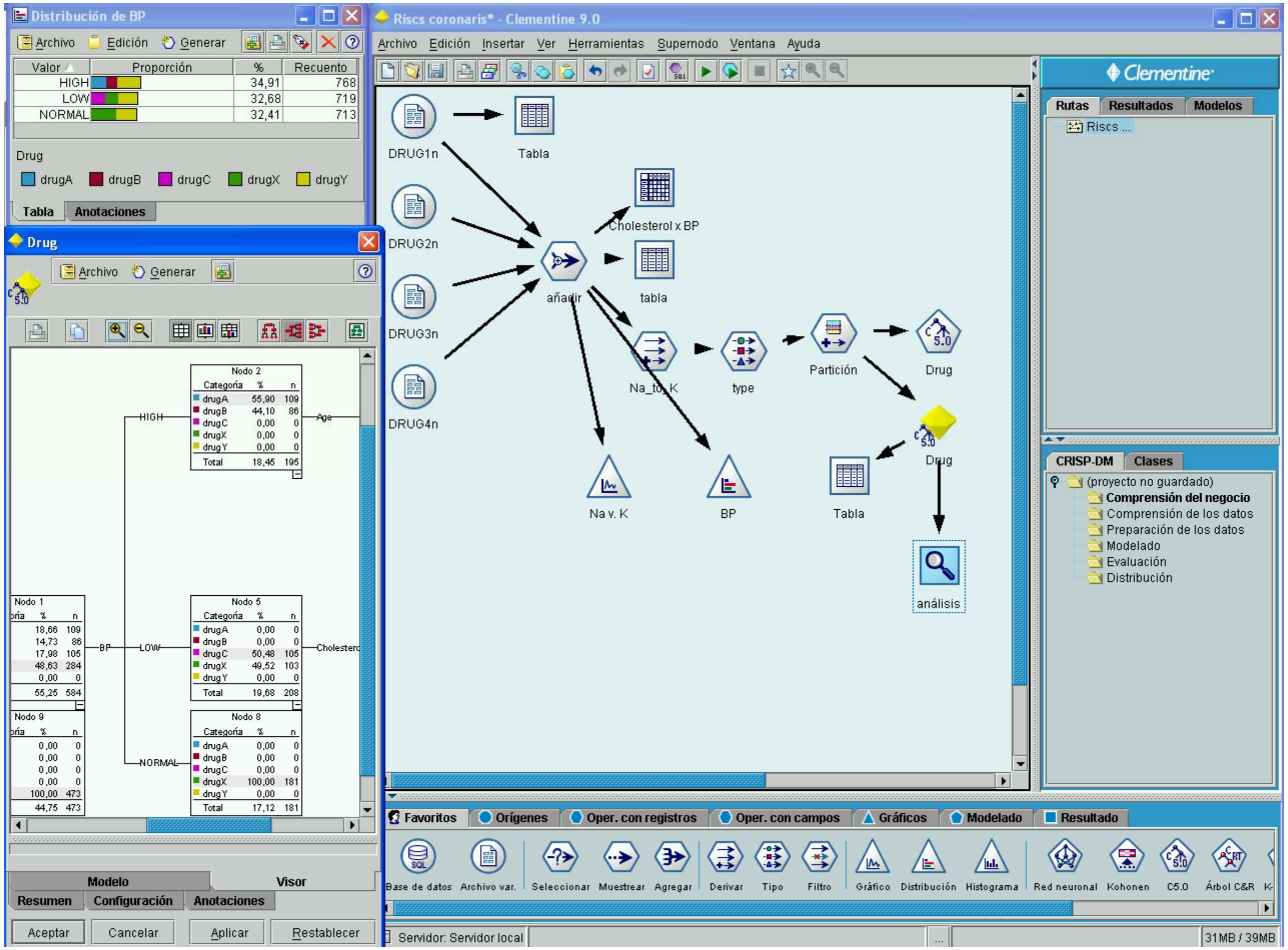
EXAMPLE: Clementine

www.spss.com

- **Tool that includes:**
  - Several data sources (ASCII, XLS and many DBMS through ODBC).
  - Visual interface.
  - Several data mining techniques: neural networks, decision trees, rules, a priori, regression, …
  - Data processing (pick & mix, combination and separation).
  - Report and batch facilities.

Tools

- **EXAMPLE: Clementine (www.spss.com)**



Tools

EXAMPLE: Clementine

- Drug study

  - A number of hospital patients suffer a pathology which can be treated with a wide range of drugs.
  - 5 different drugs are available. Patients respond differently to these drug.

  - Problem:

Which drug is the most appropriate one for a new patient?

Tools

EXAMPLE: Clementine.

First step: DATA ACCESS:

- **Read the data: e.g. a textfile with delimiters.**
- **The fields are named:**

age	age
sex	gender
BP	blood pressure (High, Normal, Low)
Cholesterol	cholesterol (Normal, High)
Na	Sodium concentration in blood.
K	Potasium concentration in blood.
drug	drug to which the patient reacted satisfactorily.

The attributes/variables can be combined:

E.g. A new attribute (Na/K), can be added.

Tools

EXAMPLE: Clementine

Second Step: Familiarisation with the data. We visualise the records:

table1

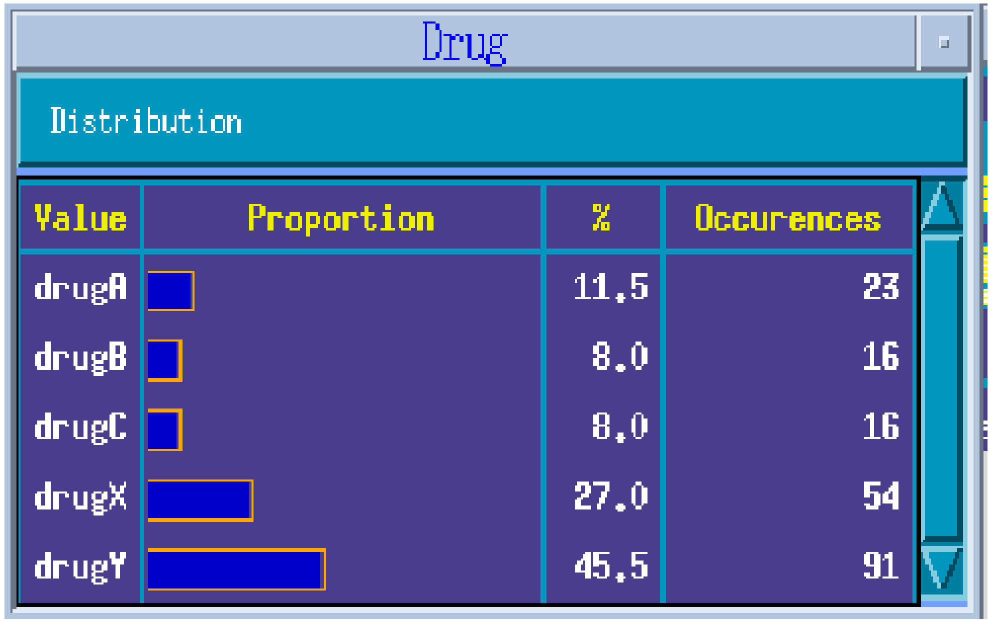
Table

Age	Sex	BP	Cholestrol	Na	K	Drug	Na_to_K
23	F	HIGH	HIGH	0.79	0.03	drugY	25.35
47	M	LOW	HIGH	0.74	0.06	drugC	13.09
47	M	LOW	HIGH	0.7	0.07	drugC	10.11
28	F	NORMAL	HIGH	0.56	0.07	drugX	7.8
61	F	LOW	HIGH	0.56	0.03	drugY	18.04
22	F	NORMAL	HIGH	0.68	0.08	drugX	8.61
49	F	NORMAL	HIGH	0.79	0.05	drugY	16.28
41	M	LOW	HIGH	0.77	0.07	drugC	11.04
60	M	NORMAL	HIGH	0.78	0.05	drugY	15.17
43	M	LOW	NORMAL	0.53	0.03	drugY	19.37

Tools

EXAMPLE: Clementine

- Allows field selection and filtering.
- Can show graphically some data properties. E.g. :
  Which is the proportion of cases which reacted well to the drug?

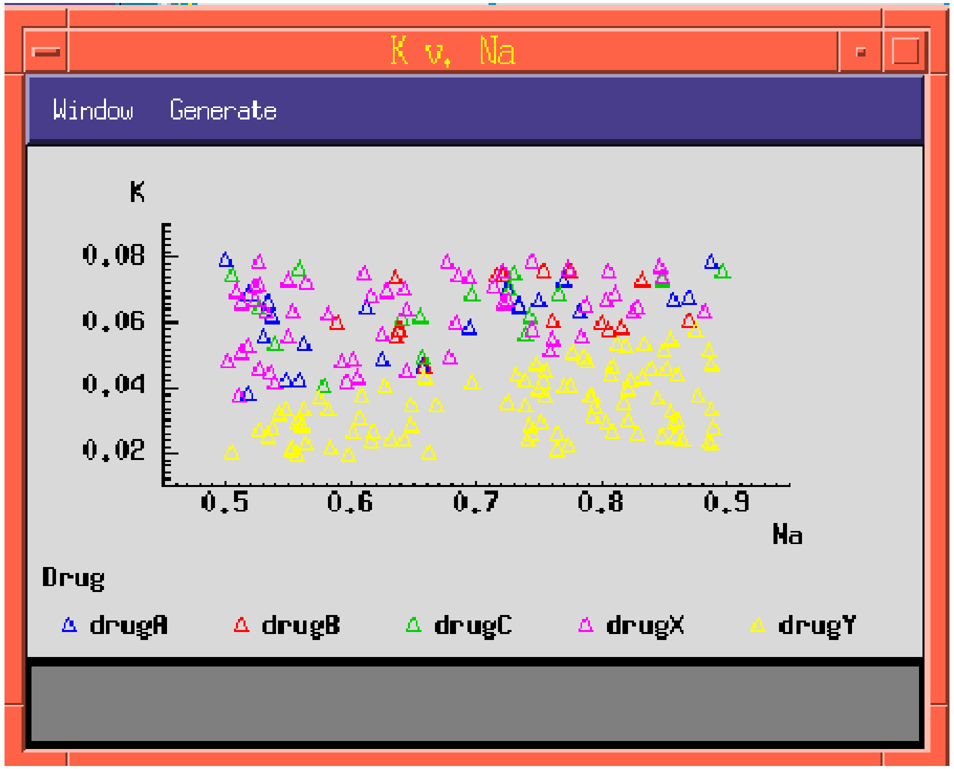


Tools

EXAMPLE: Clementine

- Can find relations. E.g:

The relation between sodium and potasium is shown in a plot.



We observe an apparently random distribution (except from drug *Y*)

Tools

EXAMPLE: Clementine

- We can clearly observe that the patients with high Na/K quotient respond better to drug Y.
- But we want a classification model for every new patient, i.e.:

Which is the best drug for each patient?

Third step: Model construction

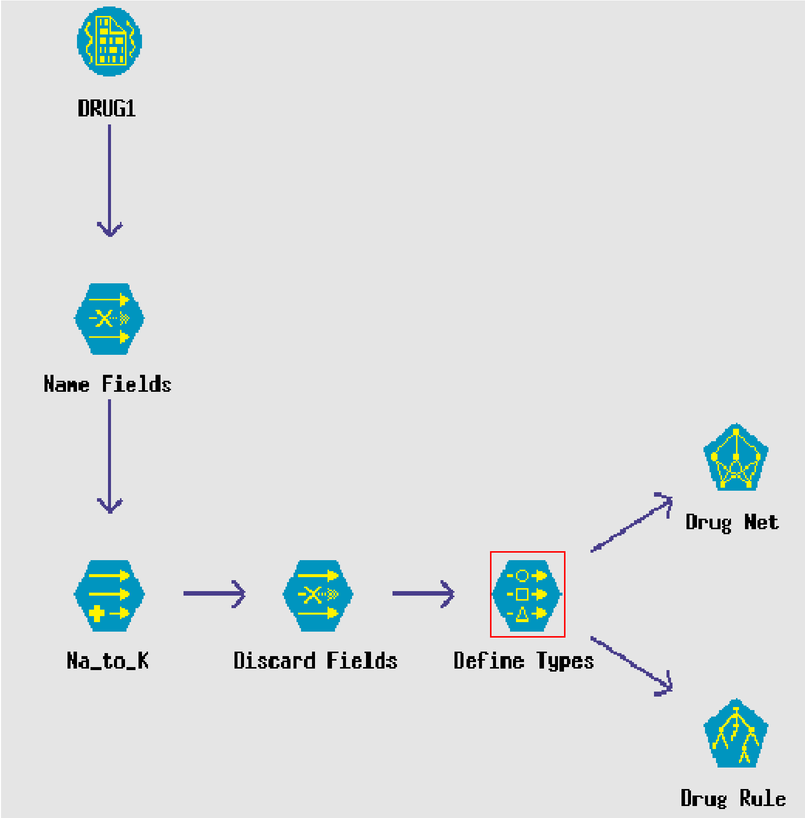
Tasks performed in Clementine:

- Filter non-desired (irrelevant) attributes.
- Type the fields.
- Construct models (rules, decision trees, neural networks, …)

Tools

EXAMPLE: Clementine

This process is performed and graphically visualised in Clementine:



From 2,000 examples the models are trained.

Tools

EXAMPLE: Clementine

Models can be browsed:

```
Rule   Folding   Select   Generate   View

Na_to_K < 16.084
    BP HIGH
        Age < 46
            Cholestrol HIGH -> drugA
            Cholestrol NORMAL
        Age >= 46
            Age < 60
            Age >= 60
    BP LOW
        Cholestrol HIGH
            Na_to_K < 15.013 -> drugC
            Na_to_K >= 15.013 -> drugY
        Cholestrol NORMAL -> drugX
    BP NORMAL
        Na_to_K < 14.884 -> drugX
        Na_to_K >= 14.884 -> drugY
Na_to_K >= 16.084 -> drugY
```

The rules extend the same criterion which was discovered previously, i.e., drug *Y* for the patient with high Na/K ratio. But it also gives rules for the rest.

Tools

EXAMPLE: SAS ENTERPRISE MINER (EM)

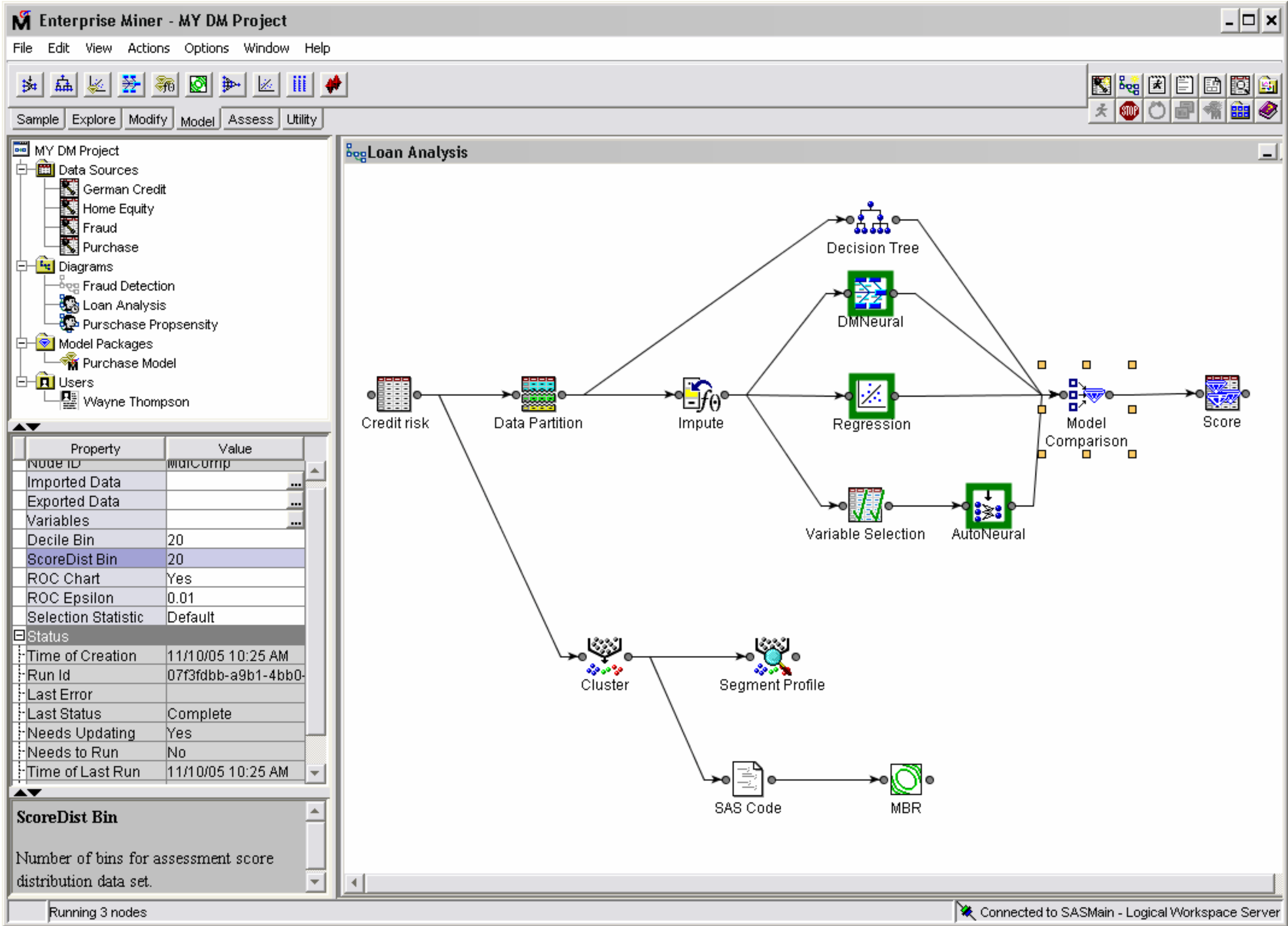
- Suite that includes:
  - Database connection (through ODBC and SAS datasets).
  - Sampling and inclusion of derived variables.
  - Data evaluation through dataset split into: training, validation (in case) and test.
  - Different data mining techniques: decision trees, regression, neural network, clustering, ...
  - Model comparisons.
  - Model conversion into SAS code.
  - Graphical interface.
- Also includes tools for all the process flow: the stages can be repeated, modified and stored.

Tools

EXAMPLE:

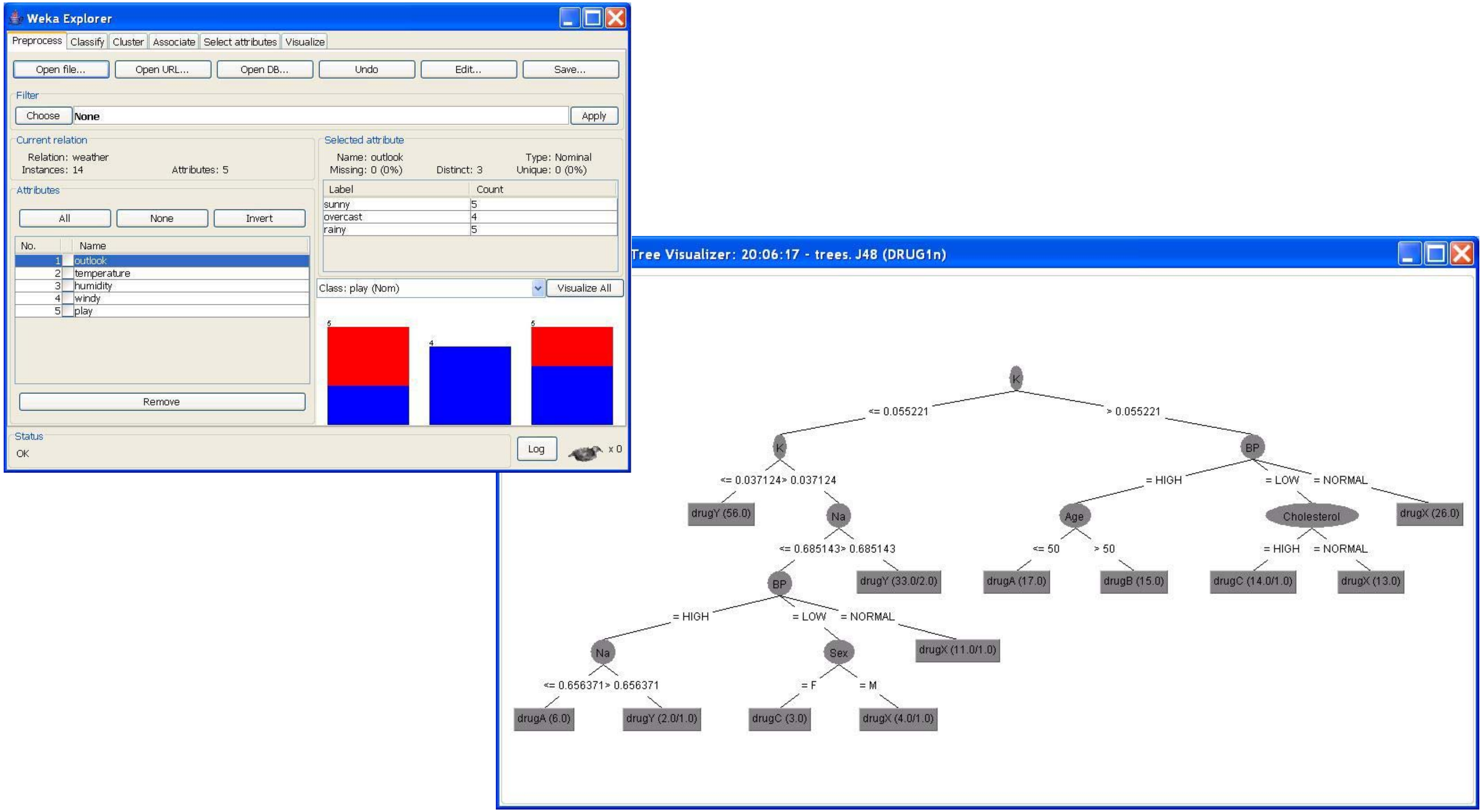
SAS ENTERPRISE MINER (EM)

(process flow, KDD)



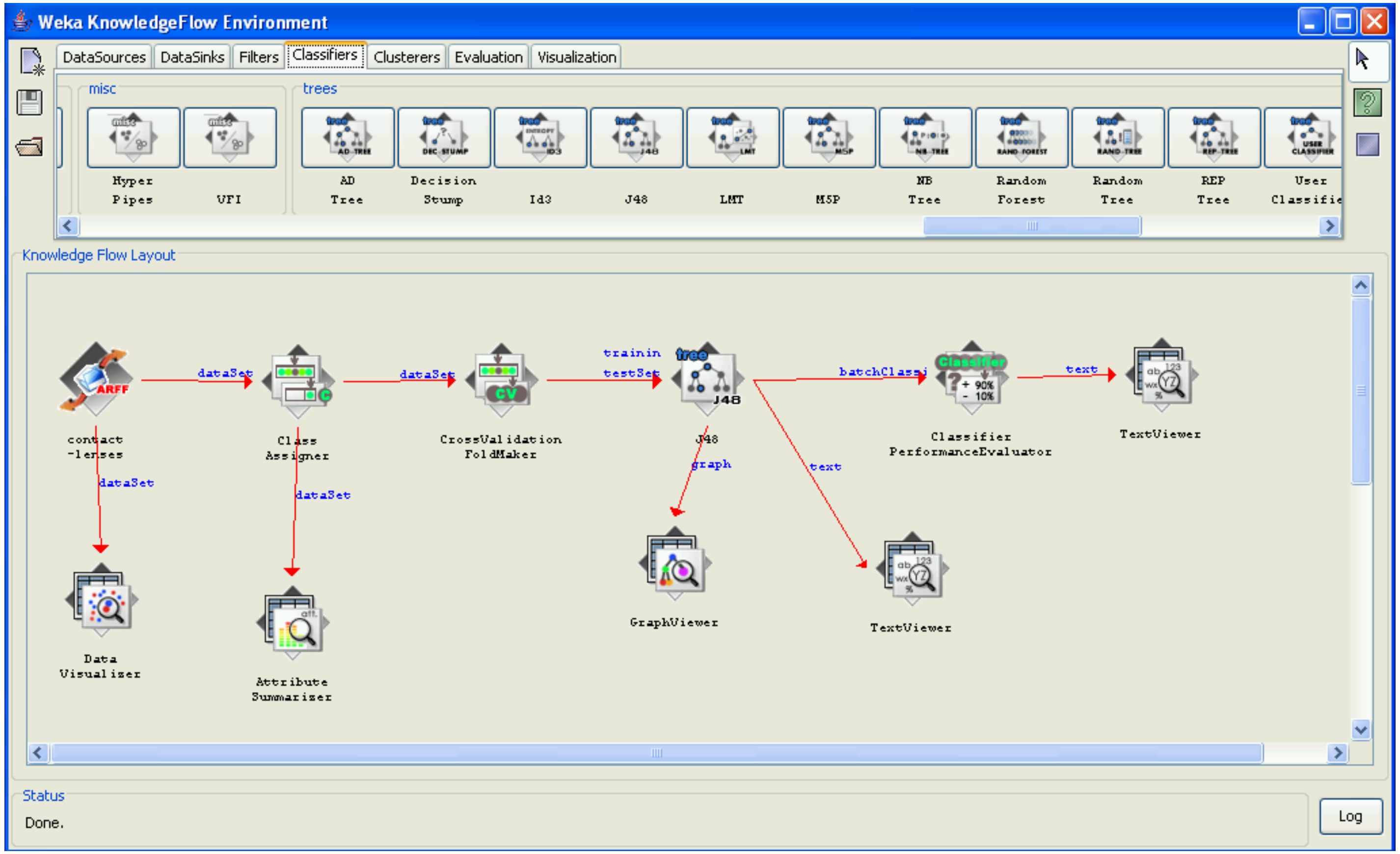
Tools

Weka, University of Waikato, NZ. (cs.waikato.ac.nz)



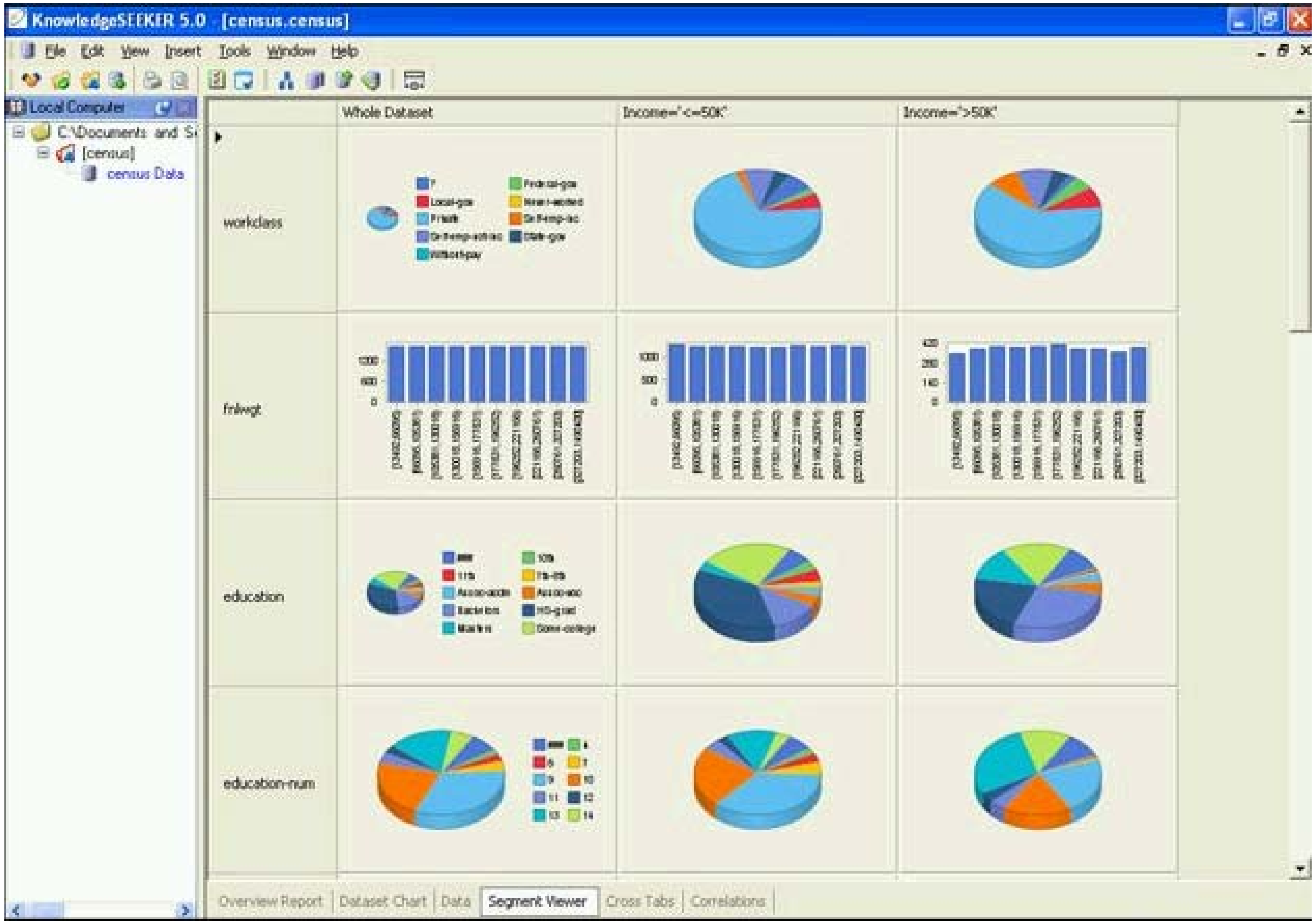
Tools

Weka, University of Waikato, NZ. (cs.waikato.ac.nz)



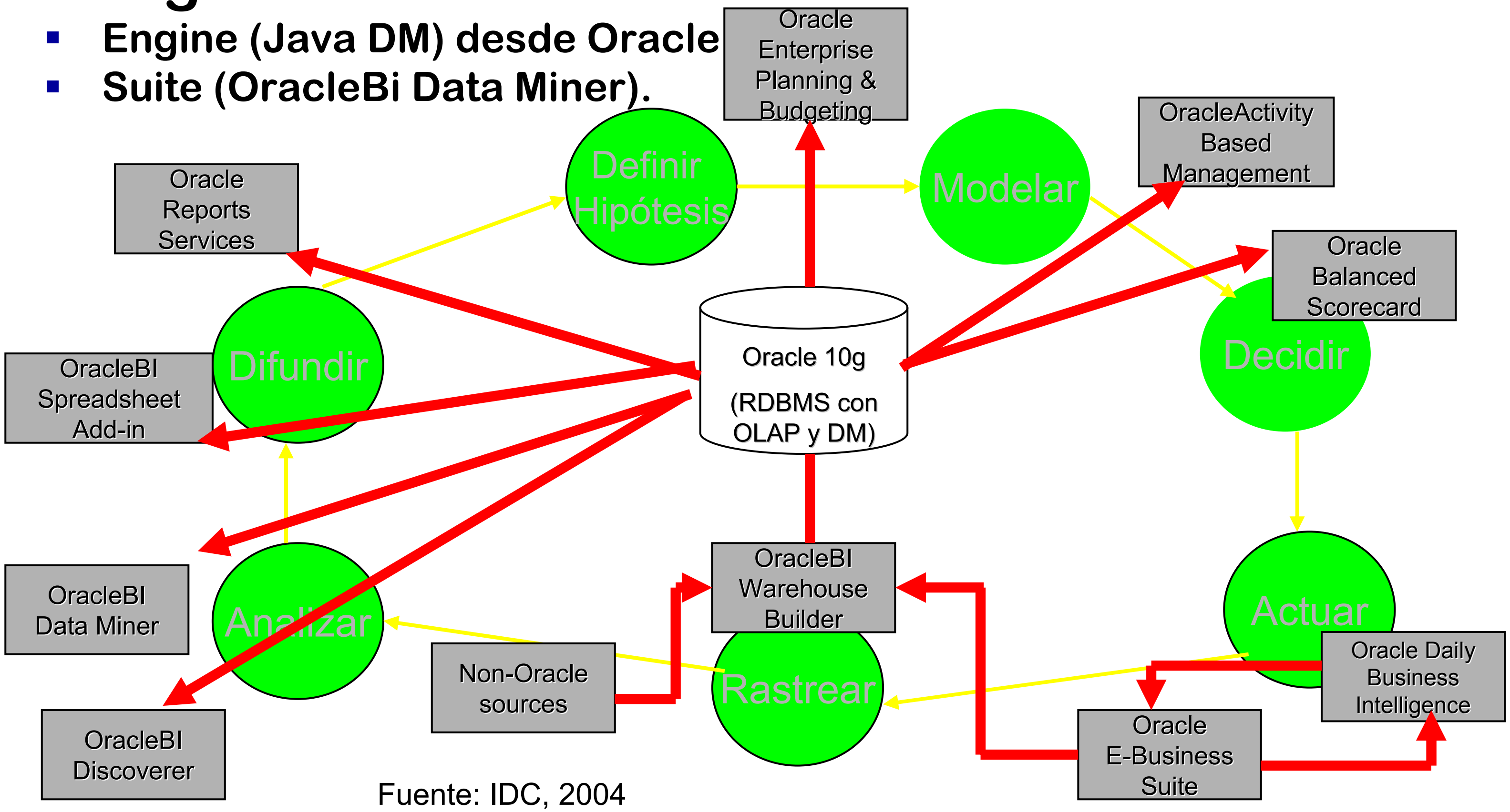
Tools

Angoss Knowledge Seeker:



Tools

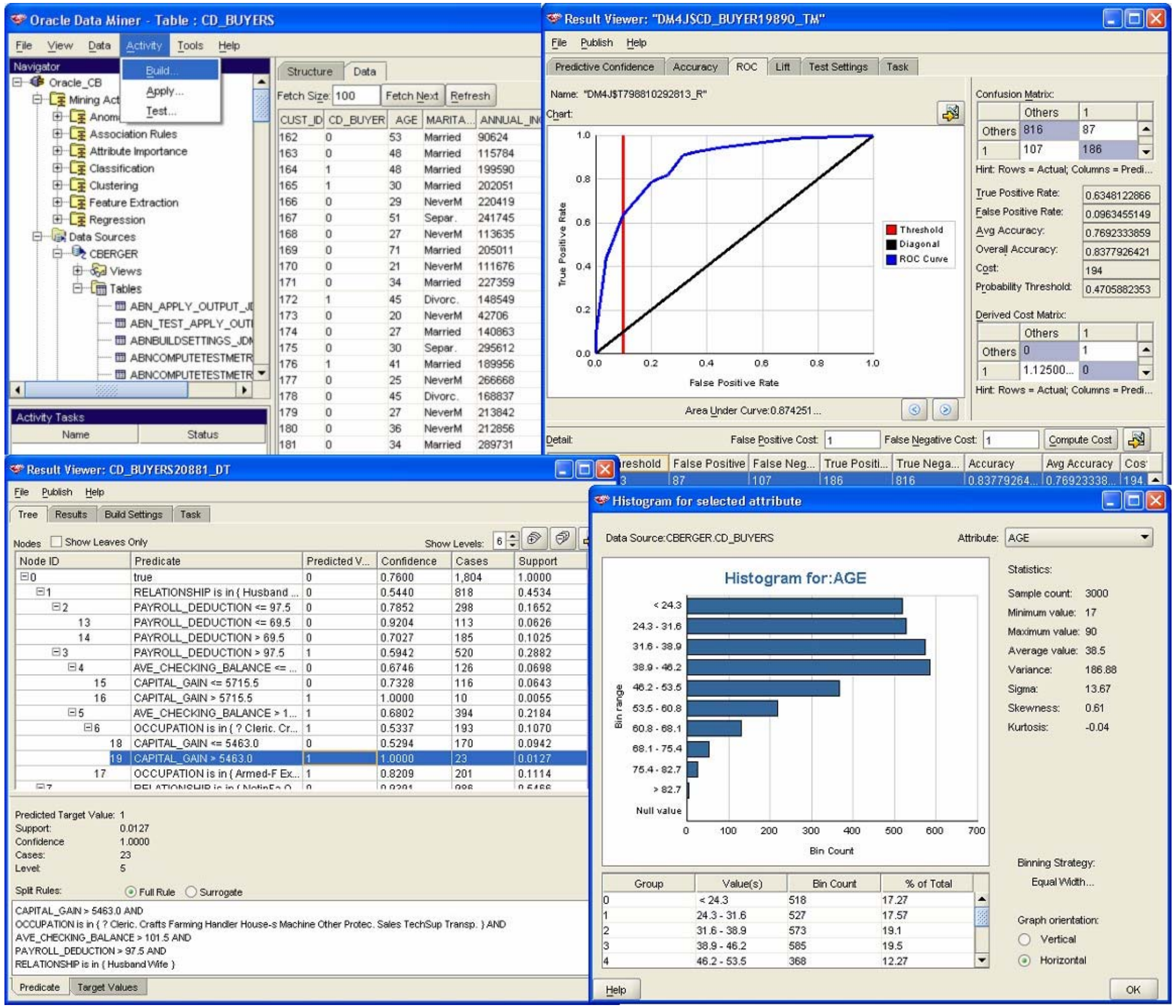
- **Oracle: Tools "Business Intelligence" and "Data Mining"**
  - **Engine (Java DM) desde Oracle**
  - **Suite (OracleBi Data Miner).**



Fuente: IDC, 2004

Tools

OracleBI Data Miner



Tools

MS SQL SERVER: Analysis Services

- **OLAP Services in SQL Server 97 was extended in SQL Server 2000 with DM features. This was called "Analysis Services". Much more techniques included in the new SQL Server (2005).**
- **In SQL Server 2007, three different interfaces available, extended DMX language.**

- **It is based on the "OLE DB for Data Mining": an extension of the DB access protocol: OLE DB.**
- **Implements an SQL extension which works with DMM (Data Mining Model) .**

Mining Non-structured Data

- *Web Mining* refers to the “global process of discovering information and knowledge which can be potentially useful and which is previously unknown from data on the web”. (Etzioni 1996)

- Web Mining combines goals and techniques from different areas:
  - Information Retrieval (IR)
  - Natural Language Processing (NLP)
  - Data Mining (DM)
  - Databases (DB)
  - WWW research
  - Agent Technology

- There are several kinds of web mining:
  - *web content mining.*
  - *web structure mining.*
  - *web use mining.*

To know more... Some pointers

- General resources:
  - www.kdnuggets.com
- Associations:
  - ACM SIGKDD (and the journal: “explorations”)
  - http://www.sigkdd.org/explorations/issue.php?issue=current
- Some books:
  - Berry M.J.A.; Linoff, G.S. “Mastering Data Mining” Wiley 2000.
  - Berthold, M.; Hand, D.J. (ed) “Intelligent Data Analysis. An Introduction” Second Edition, Springer 2002.
  - Dunham, M.H. “Data Mining. Introductory and Advanced Topics” Prentice Hall, 2003.
  - Han, Jiawei; Micheline Kamber “Data Mining: Concepts and Techniques” Morgan Kaufmann, April 2000.
  - Witten, I.H.; Frank, E. "Tools for Data Mining", Morgan Kaufmann, 2005.