Data Mining & Business Intelligence: Practical Tools

Motaz K. Saad
msaad@iugaza.edu
Dept. of CS – College of IT
Oct. 2007

Outline

- Why Data Mining?
- What is Data Mining?
- Data Mining Tools
  - Open Source Tools: WEKA, Rapid Miner
  - Non-Open Source Tools: SPSS, MS SQL 2005 Analytics Service (SSAS), Oracle Data Miner (ODM)

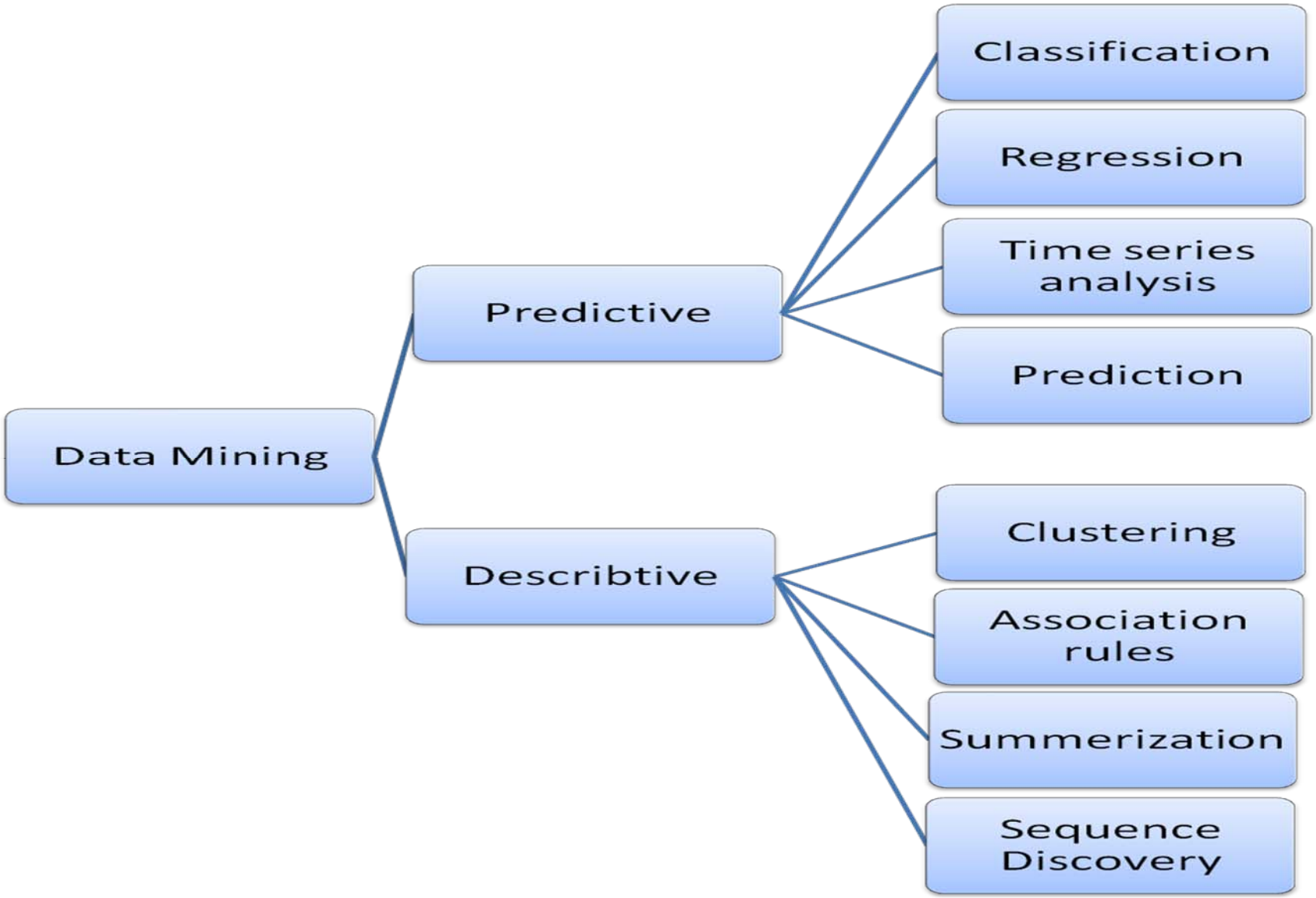
Why Data Mining?!

- The Explosive Growth of Data (from gigabytes to terabytes). The explosion of data caused by:
- Automated data collection tools, database systems, Web, computerized society.
- Abundant data available from various sources:
  - Business: Web, e-commerce, transactions, stocks …
  - Science: Remote sensing, bioinformatics, scientific simulation …
  - Society and everyone: news, digital cameras …
- From the previous discussion we conclude that:
- ***"We are drowning in data, but starving for knowledge!"***
- *"Necessity is the mother of invention"* **Data mining is: Automated analysis of massive data sets.**

What is Data Mining?

- Data mining (knowledge discovery from data) is Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.
- Alternative names for Data Mining: Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Data Mining Functionality (Tasks)



Data Mining Functionality (Tasks)

- ***Classification*** maps data into predefined groups or classes: Supervised learning, Pattern recognition, Prediction
- ***Clustering*** groups similar data together into clusters: Unsupervised learning, Segmentation, Partitioning

Data Mining Functionality (Tasks)

- ***Link Analysis*** uncovers relationships among data: Association Rules, Sequential Analysis determines sequential patterns.
- ***Outlier analysis***
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)

Why Data Mining?—Potential Applications

- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Bioinformatics and bio-data analysis

Example 1: Market Analysis and Management

- Where does the data come from?—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

Example 1: Market Analysis and Management (Cont.)

- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - Identify the best products for different customers
  - Predict what factors will attract new customers

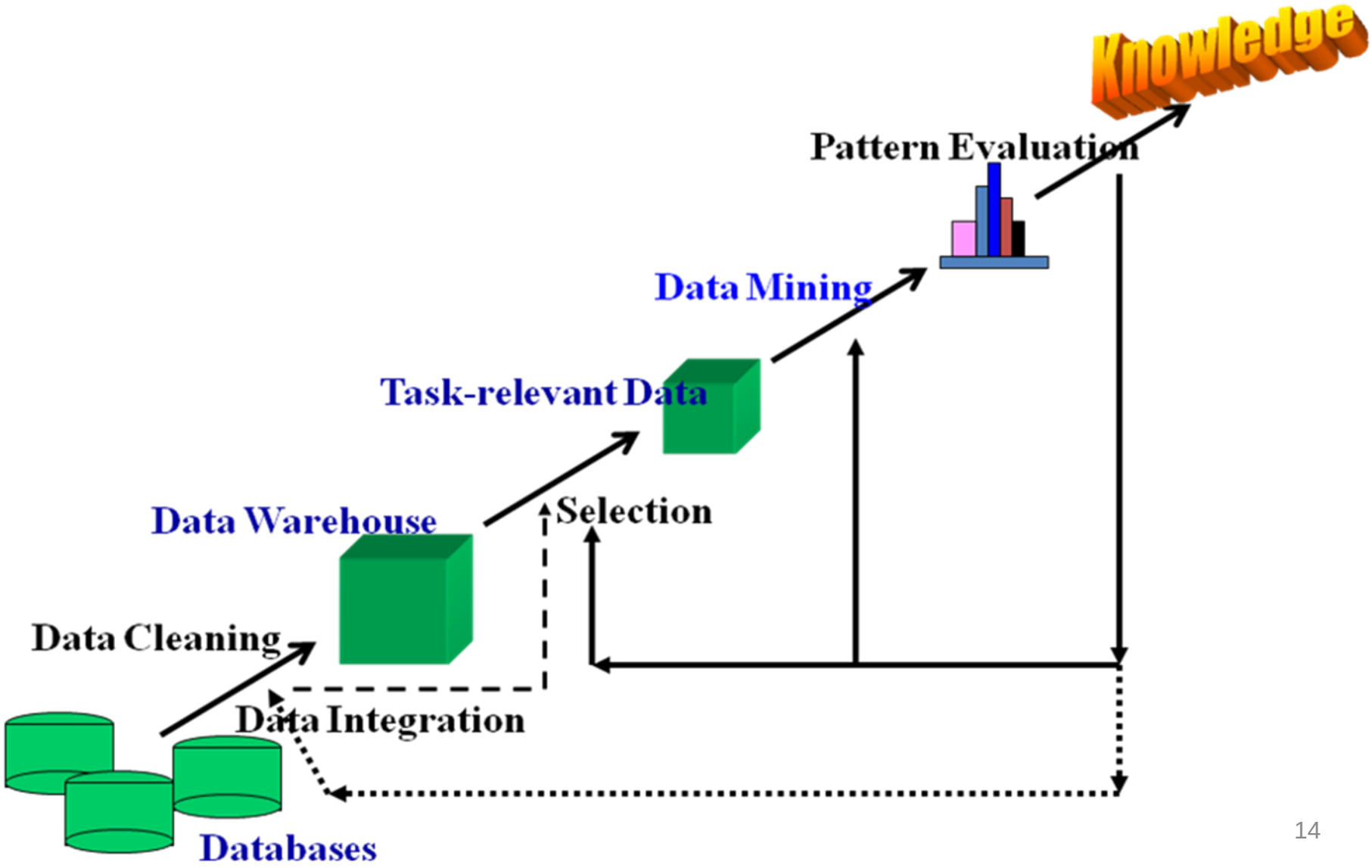
Example 2: Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: detect a group of people who stage accidents to collect on insurance.
  - Money laundering: detect suspicious monetary transactions
  - Medical insurance
    - Professional patients: detect professional patients and ring of doctors and ring of references.
    - Unnecessary or correlated screening tests

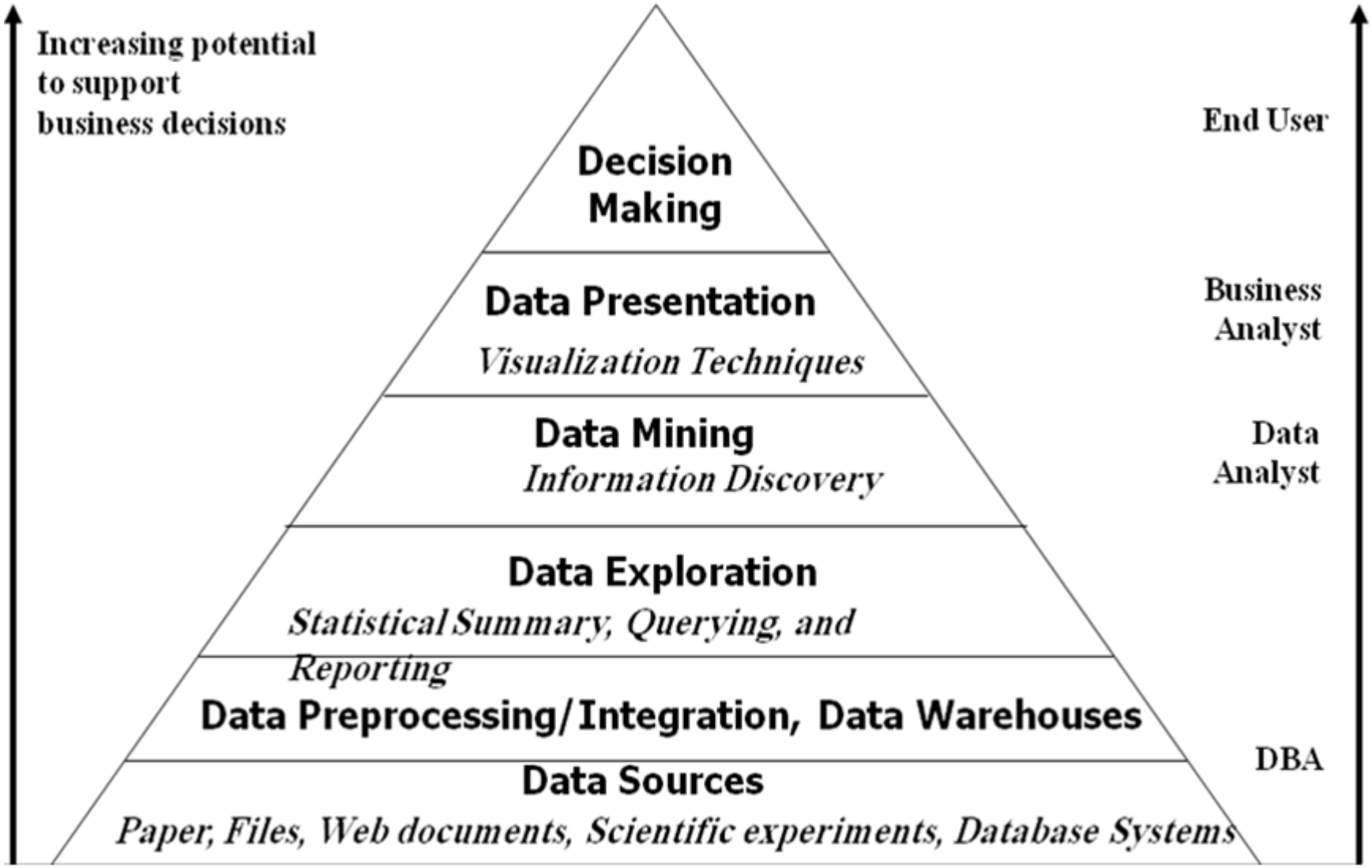
Example 2: Fraud Detection & Mining Unusual Patterns (CONT.)

- Telecommunications: phone-call fraud
  - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
- Retail industry
  - Analysts estimate that 38% of retail shrink is due to dishonest employees

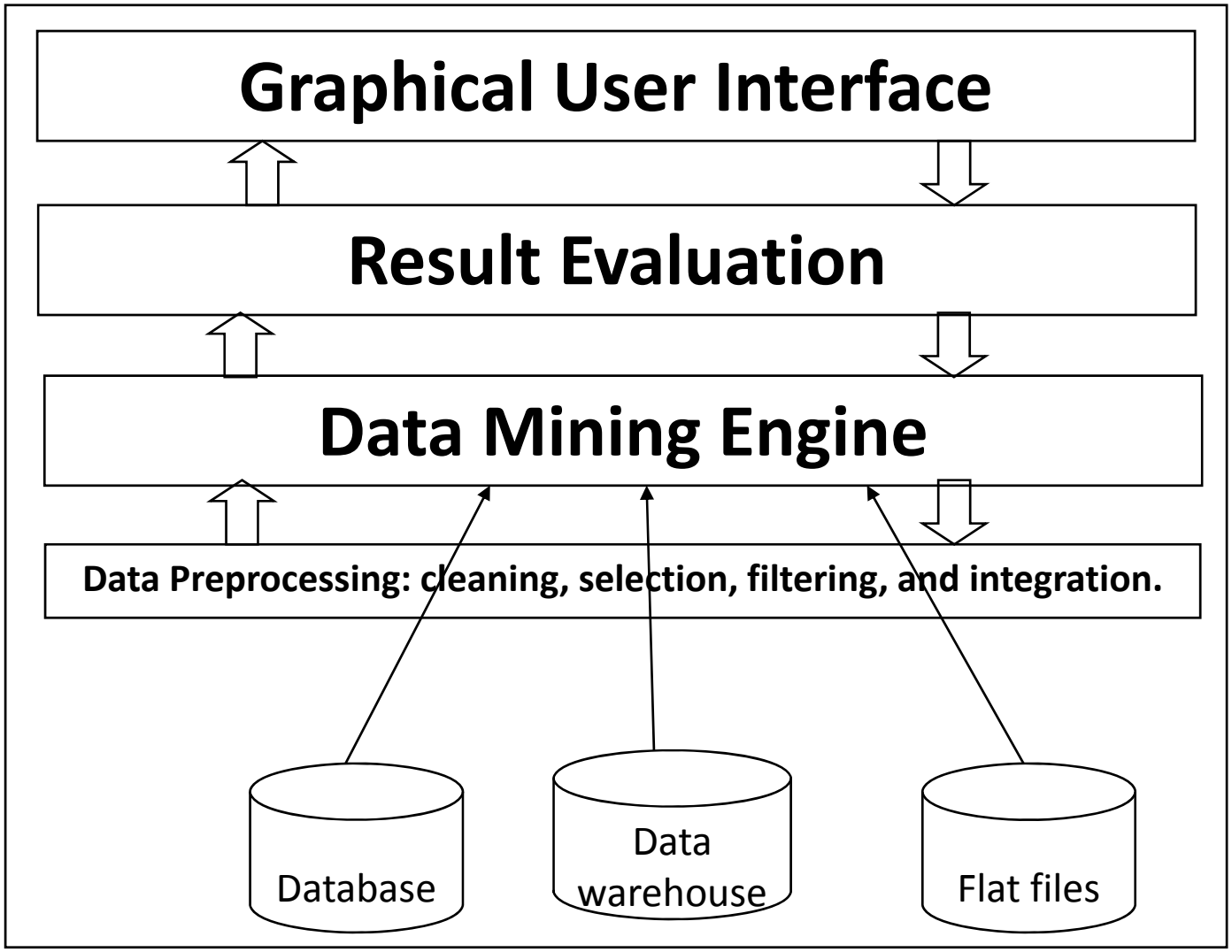
Knowledge Discovery (KDD) Process



Data Mining (DM), Decision Support System (DSS) and Business Intelligence (BI)



Architecture: Typical Data Mining System



Database vs data mining processing

- **Query:** Well defined (SQL)
- **Data:** Operational data
- **Output:** Precise Subset of database

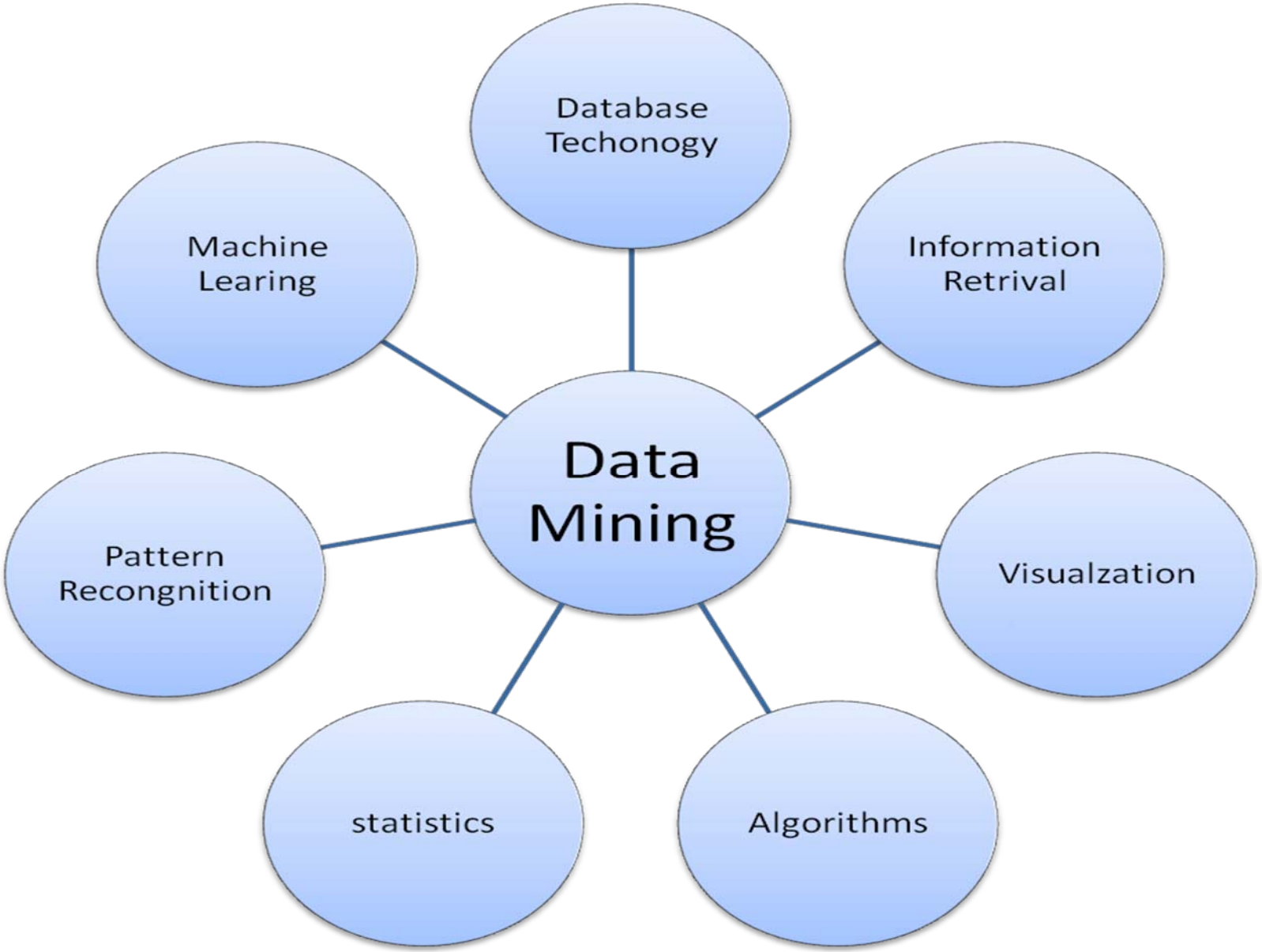
- **Query:** Poorly defined, No precise query language: Data Mining Query Language (DMQL)
- **Data:** Not operational data (warehouses)
- **Output:** Fuzzy, not a subset of database (**knowledge**)

Database vs. data mining processing: An example

- Find all credit applicants with last name of Smith.
- Identify customers who have purchased more than $10,000 in the last month.
- Find all customers who have purchased milk

- Find all credit applicants who are poor credit risks. (Classification).
- Identify customers with similar buying habits. (Clustering).
- Find all items which are frequently purchased with milk. (Association rules).

Data Mining: Confluence of Multiple Disciplines



Some Data Mining tools

- Open Source Data Mining Tools
  - Weka
  - Rapid Miner
- Non-open Source Data Mining Tools
  - SPSS
  - MS SQL 2005 Analytics Service (SSAS)
  - Oracle Data Miner (ODM)

Weka: The Bird

WEKA

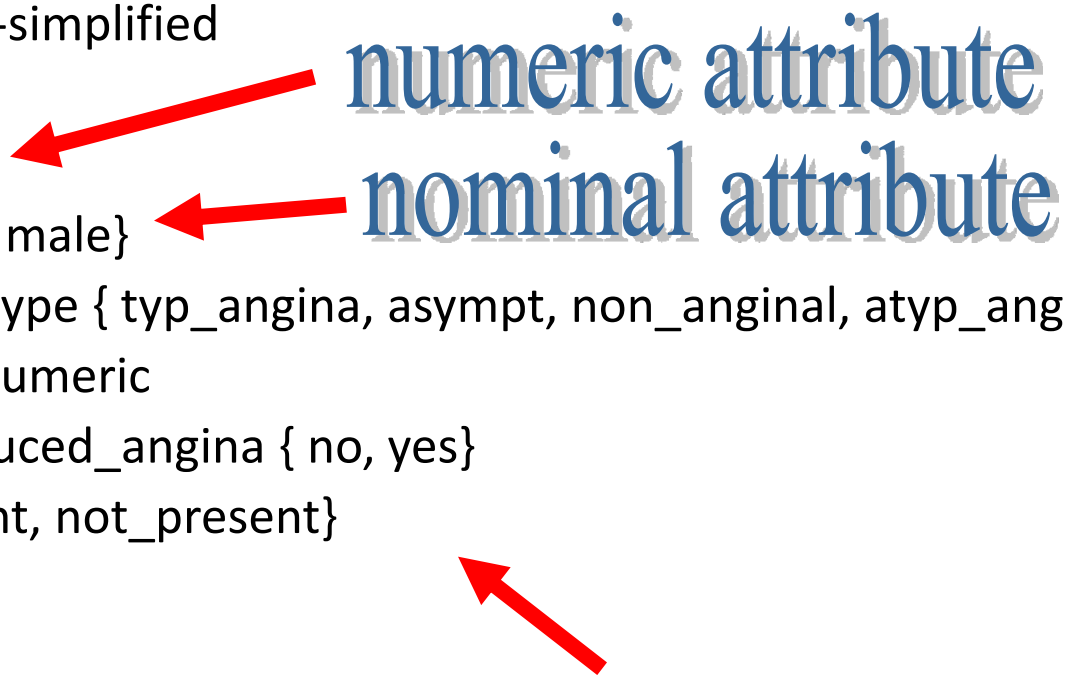
- Machine learning/data mining software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications
- Complements “Data Mining” by Witten & Frank
- Main features:
  - Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  - Graphical user interfaces (incl. data visualization)
  - Environment for comparing learning algorithms

WEKA deals with “flat” files

@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...



Weka Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …

Weka Explorer: building “classifiers”

- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes’ nets, …
- “Meta”-classifiers include:
  - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, …

Weka Explorer: clustering data

- WEKA contains “clusterers” for finding groups of similar instances in a dataset
- Implemented schemes are:
  - *k*-Means, EM, Cobweb, *X*-means, FarthestFirst
- Clusters can be visualized and compared to “true” clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution

Weka Explorer: finding associations

- WEKA contains an implementation of the Apriori algorithm for learning association rules
  - Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
  - milk, butter $\Rightarrow$ bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all rules that have a given minimum support and exceed a given confidence

Weka Explorer: attribute selection

- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
  - A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
  - An evaluation method: correlation-based, wrapper, information gain, chi-squared, …
- Very flexible: WEKA allows (almost) arbitrary combinations of these two

Weka Explorer: data visualization

- Visualization very useful in practice: e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
  - To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- “Jitter” option to deal with nominal attributes (and to detect “hidden” data points)
- “Zoom-in” function

Weka: Performing experiments

- Experimenter makes it easy to compare the performance of different learning schemes
- For classification and regression problems
- Results can be written into file or database
- Evaluation options: cross-validation, learning curve, hold-out
- Can also iterate over different parameter settings
- Significance-testing built in!

Weka The Knowledge Flow GUI

- New graphical user interface for WEKA
- Java-Beans-based interface for setting up and running machine learning experiments
- Data sources, classifiers, etc. are beans and can be connected graphically
- Data “flows” through components: e.g.,

  “data source” -> “filter” -> “classifier” -> “evaluator”
- Layouts can be saved and loaded again later

Weka: try it yourself!

- WEKA is available at

  http://www.cs.waikato.ac.nz/ml/weka

- Also has a list of projects based on WEKA
- WEKA contributors:

  Abdelaziz Mahoui, Alexander K. Seewald, Ashraf M. Kibriya, Bernhard Pfahringer , Brent Martin, Peter Flach, Eibe Frank ,Gabi Schmidberger ,Ian H. Witten , J. Lindgren, Janice Boughton, Jason Wells, Len Trigg, Lucio de Souza Coelho, Malcolm Ware, Mark Hall ,Remco Bouckaert , Richard Kirkby, Shane Butler, Shane Legg, Stuart Inglis, Sylvain Roy, Tony Voyle, Xin Xu, Yong Wang, Zhihai Wang

Rapid Miner

- Rapid Miner (formerly YALE “Yet Another Learning Environment”)
- The world-leading open-source system for knowledge discovery and data mining.
- It is available as a stand-alone application for data analysis and as a data mining engine which can be integrated into own products.

Rapid Miner (Cont.)

- It is available in different flavours:
  - A free open-source version licensed under the GPL, a free version with an improved user interface.
  - Under a developer license (OEM) which allows the integration of RapidMiner as a powerful library even into proprietary products. Enhance your products with adaptability and innovative analytical features.

Rapid Miner: Features

- Freely available **open-source** knowledge discovery environment
- 100% pure **Java** (runs on every major platform and operating system)
- KD processes are modeled as simple **operator trees** which is both intuitive and powerful
- Operator trees or subtrees can be saved as **building blocks** for later re-use

Rapid Miner: Features

- Internal **XML** representation ensures standardized interchange format of data mining experiments
- Simple scripting language allowing for automatic **large-scale** experiments
- **Multi-layered data view concept** ensures efficient and transparent data handling

Rapid Miner: Features

- Flexibility in using RapidMiner:
  - **Graphical user interface** (GUI) for interactive prototyping
  - **Command line mode** (batch mode) for automated large-scale applications
  - **Java API** (application programming interface) to ease usage of RapidMiner from your own programs
- Simple plugin and extension mechanisms, a broad variety of **plugins** already exists and you can easily add your own

Rapid Miner: Features

- powerful plotting facility offering a large set of sophisticated **high-dimensional visualization** techniques for data and models
- **more than 400** machine learning, evaluation, in- and output, pre- and post-processing, and visualization operators plus numerous meta optimization schemes

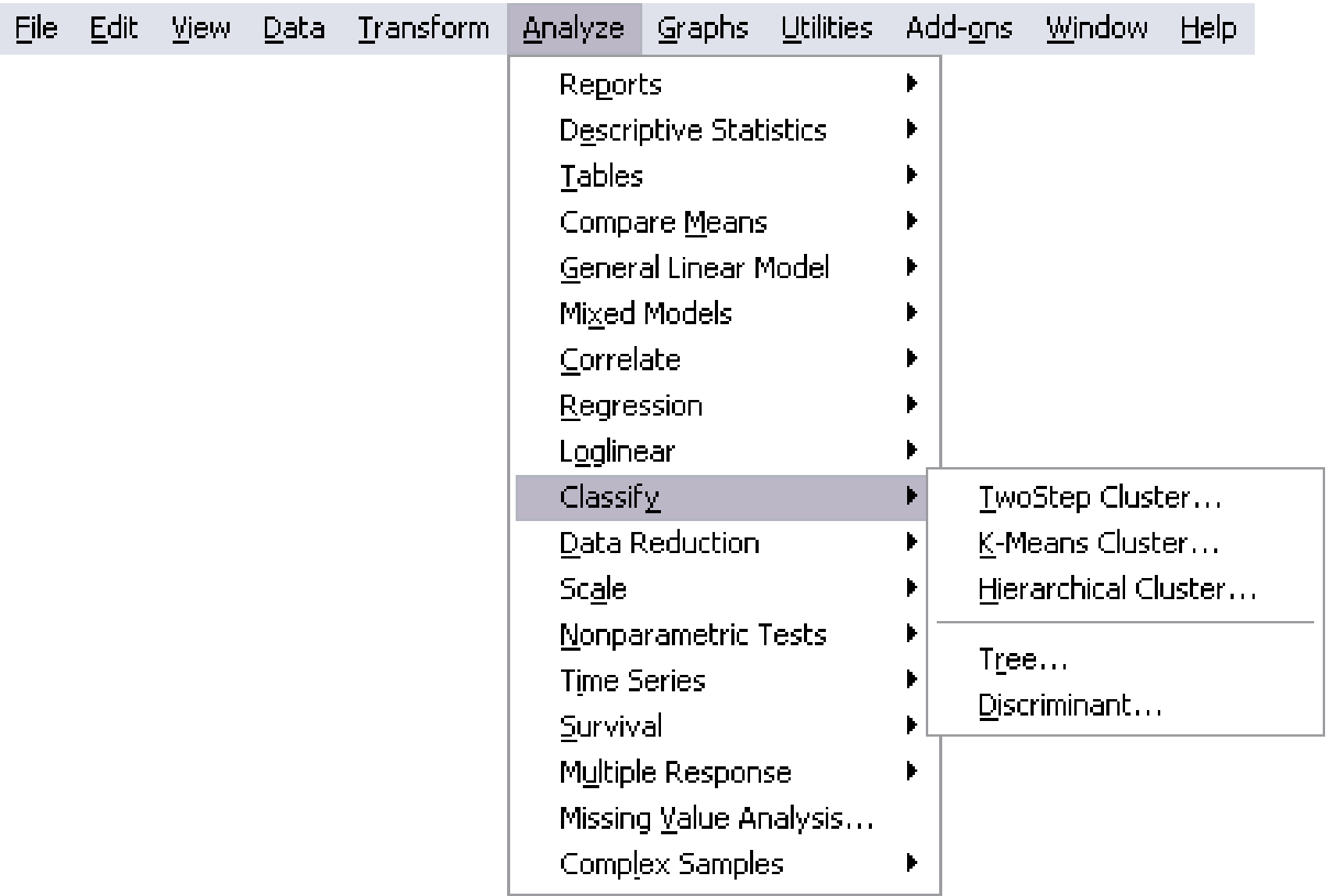
Rapid Miner: Features

- Machine learning library **WEKA** fully integrated
- RapidMiner was successfully applied on a wide range of applications where its rapid prototyping abilities demonstrated their usefulness, including **text mining**, **multimedia mining**, **feature engineering**, **data stream mining** and **tracking drifting concepts**, development of **ensemble methods**, and **distributed data mining**.

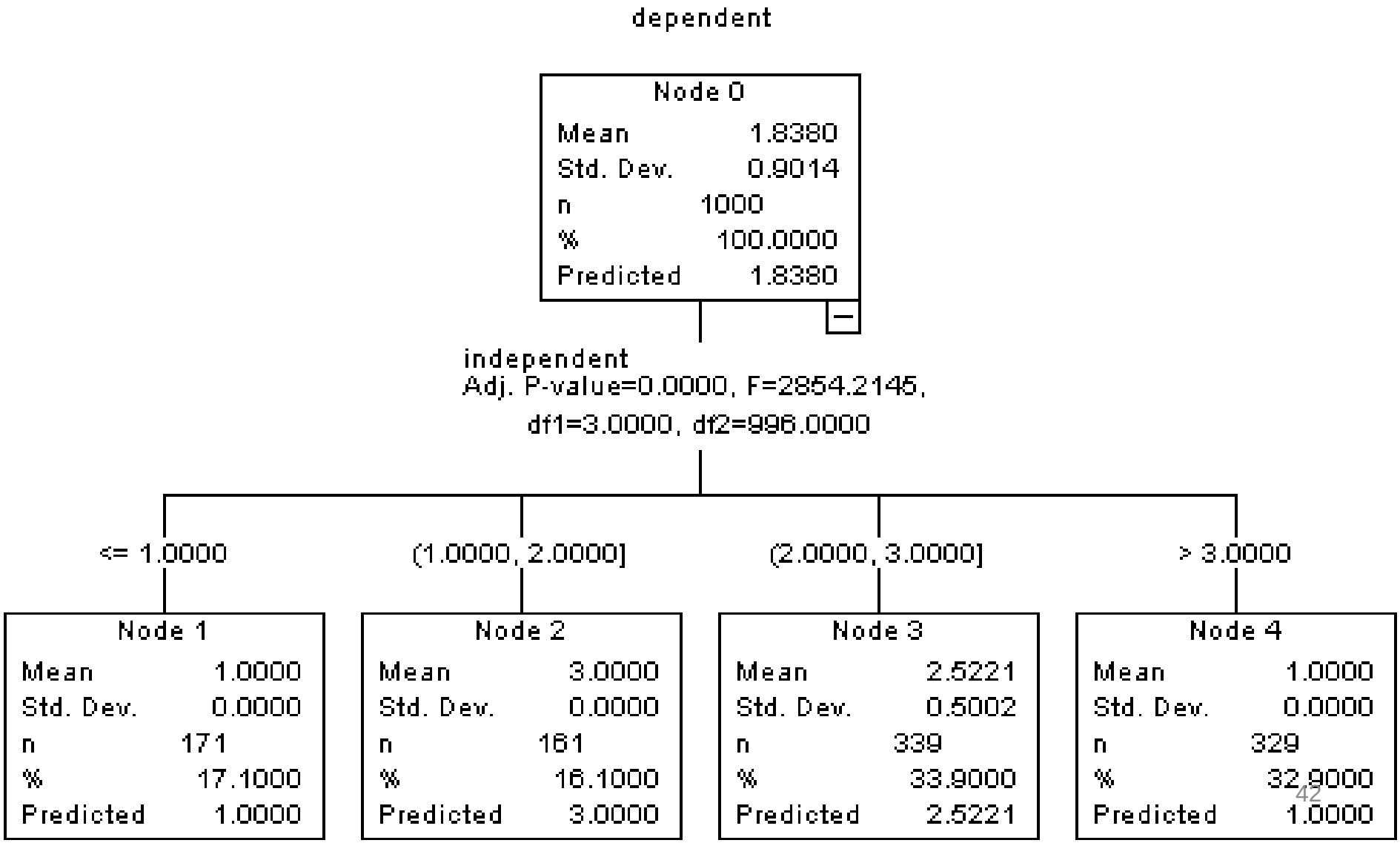
SPSS

- Analytical Software at **SPSS**.com.
- Specializing in **data mining**, customer relationship management, business intelligence and **data** analysis.

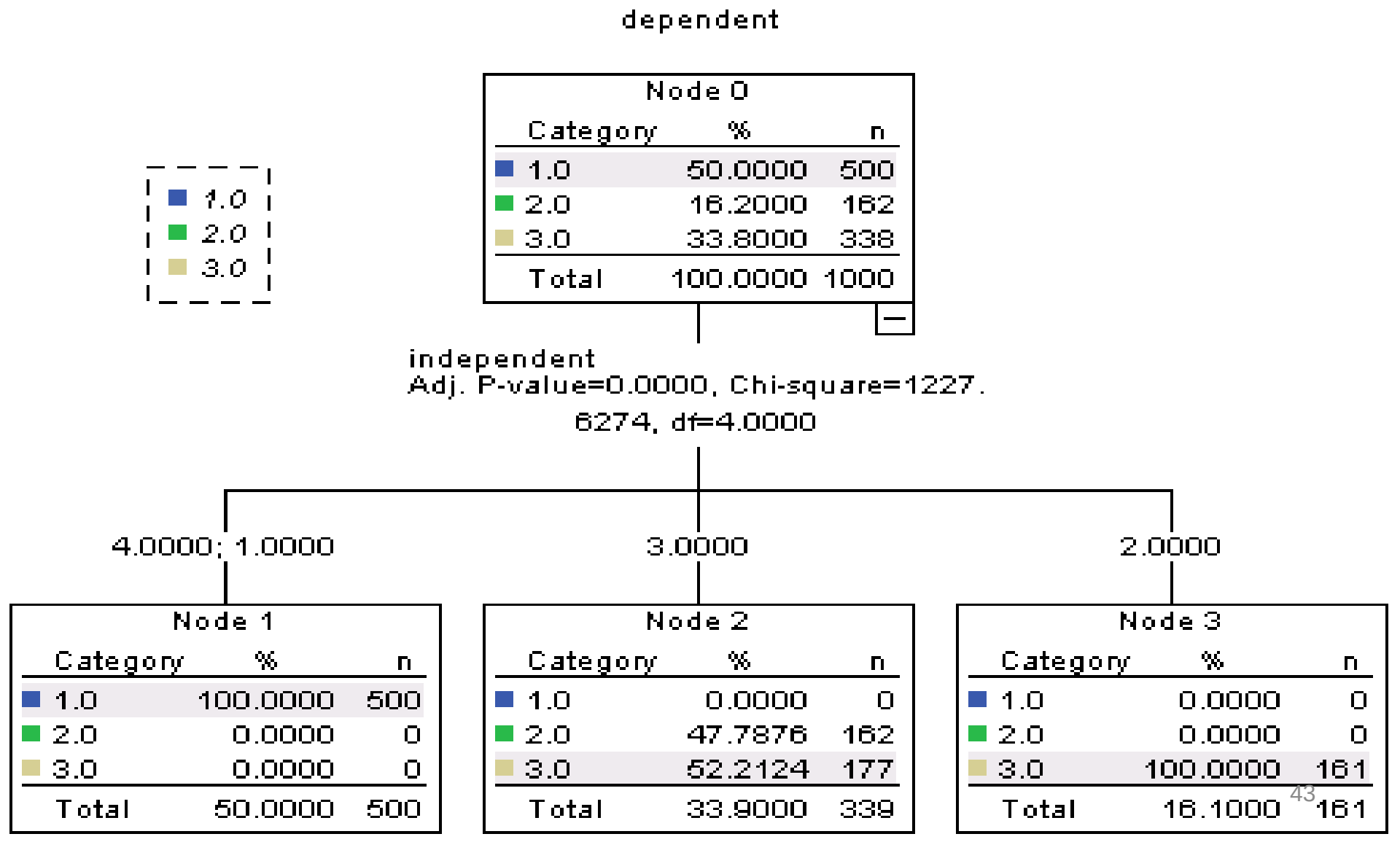
SPSS Classification tree



SPSS Classification tree (CONt.)

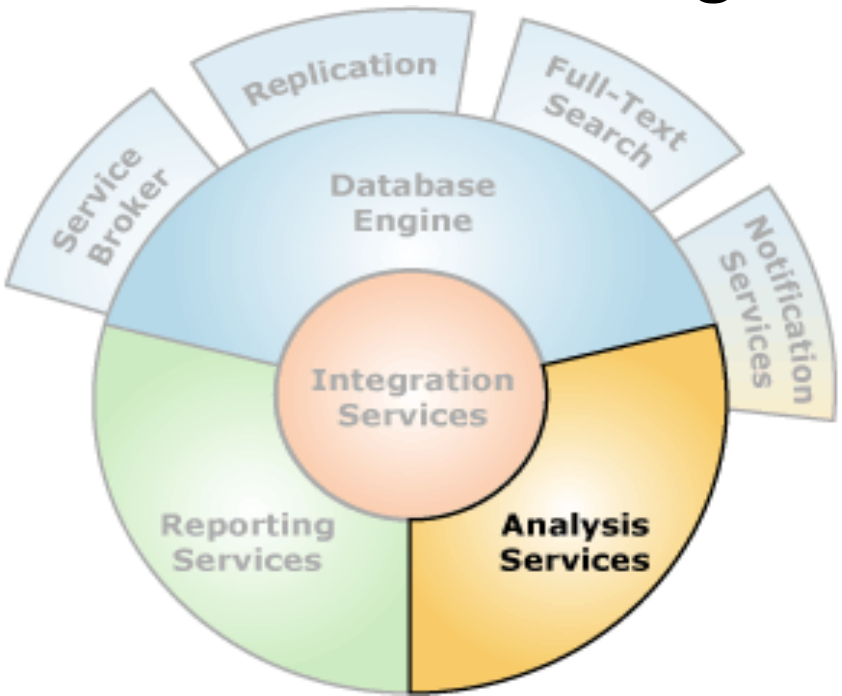


SPSS Classification tree (CONt.)



MS SQL 2005 Analytics Service (SSAS)

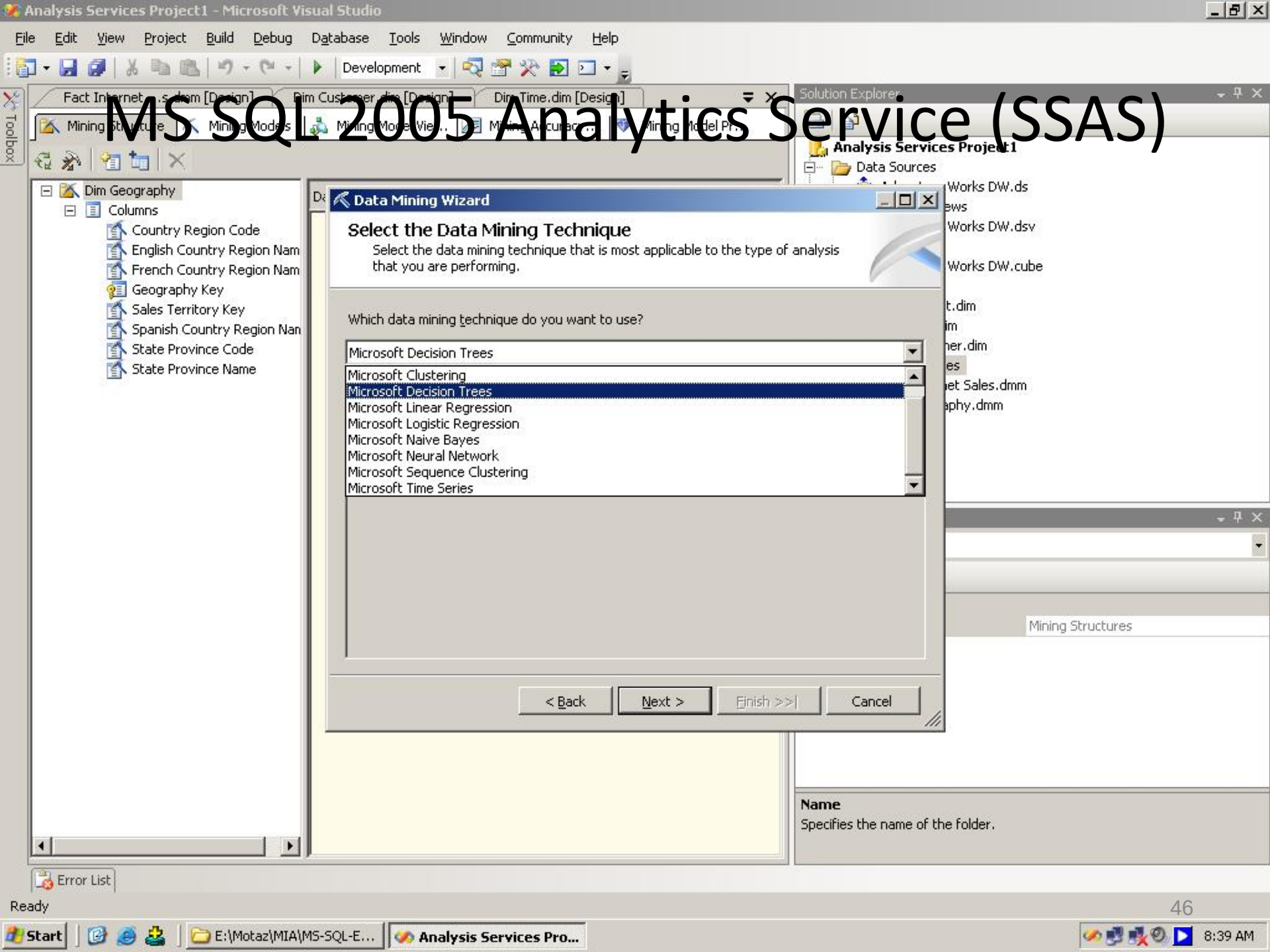
- Microsoft SQL Server 2005 Analysis Services (SSAS) delivers online analytical processing (OLAP) and data mining functionality for business intelligence applications.



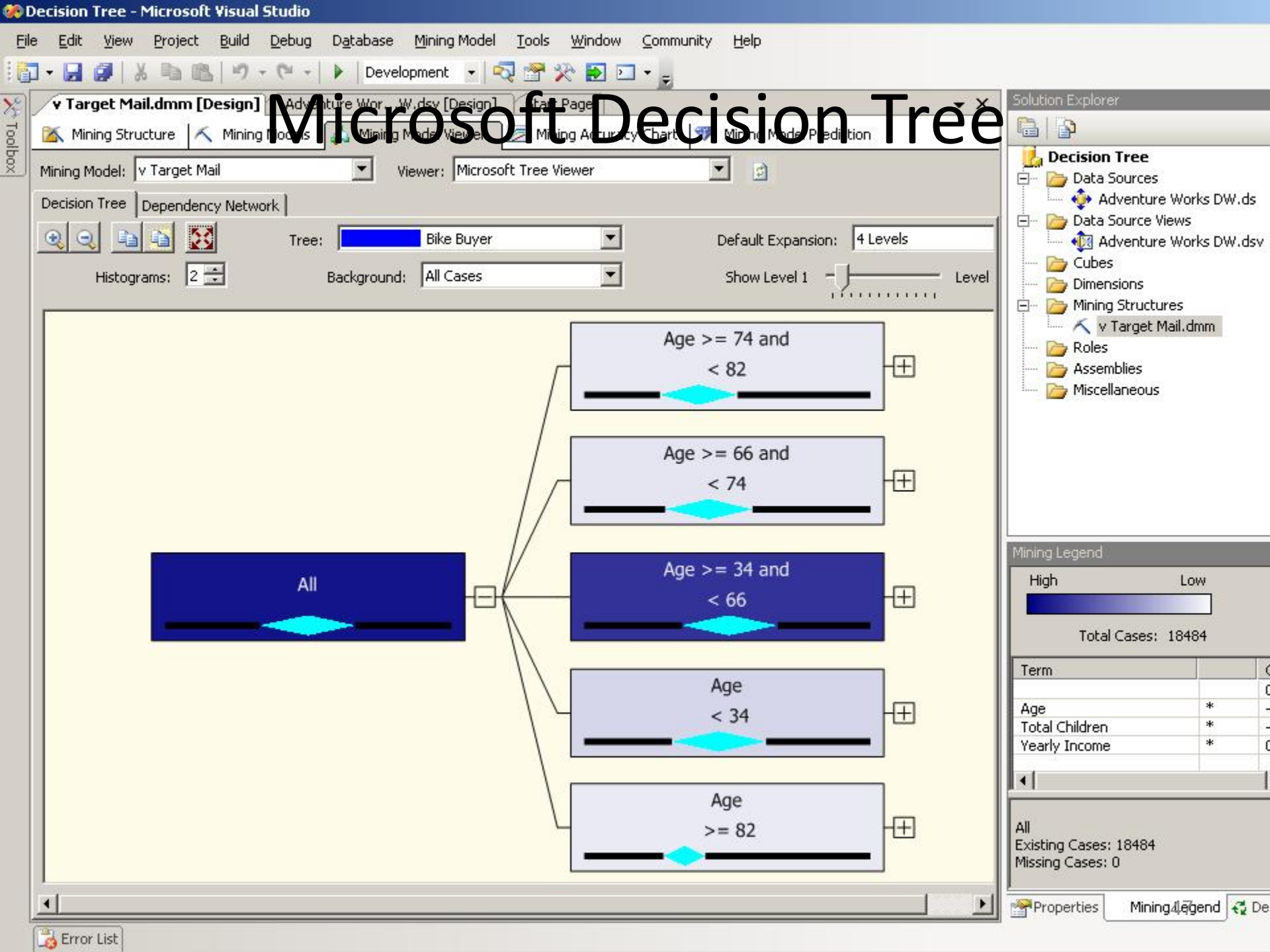
MS SQL 2005 Analytics Service (SSAS)

- Analysis Services supports OLAP by letting you design, create, and manage multidimensional structures that contain data aggregated from other data sources, such as relational databases.
- For data mining applications, Analysis Services lets you design, create, and visualize data mining models that are constructed from other data sources by using a wide variety of industry-standard data mining algorithms.

MS SQL 2005 Analytics Service (SSAS)



Microsoft Decision Tree



Oracle Data Miner

- Oracle Data Mining (ODM)
- An option to Oracle Database 10*g* Enterprise Edition
- Enables you to produce actionable predictive information and build integrated business intelligence applications.
- Using data mining functionality embedded in Oracle Database 10*g*, you can find patterns and insights hidden in your data.

Oracle Data Miner

- Application developers and integrators can quickly automate the distribution of new business intelligence—predictions, patterns and discoveries—throughout your organization.
- Oracle Data Mining enables business decision makers, data analysts, integrators, and IT to extract greater value from corporate data resulting in better informed business decisions that address a wide range of business problems.

Oracle Data Miner

Oracle Data Miner - View : MINING_DATA_BUILD_V

File View Data Activity Tools Help

Navigator

conn
Mining Acti
Anoma
Assoc
Attribu
Classif
Cluste
Featur
Regres
Data Sourc
CTXSY
DMSY
DMUSE
Vi

Activity Tasks

Name

Activities Server

New Activity Wizard - Step 1 of 5: Model Type

Select Mining Activity Type

Choose a model function type and algorithm. Review the descriptions to be sure you have picked the most appropriate selections. Click the Help button for additional details.

Function Type: Anomaly Detection

Algorithm:

Description:

- Anomaly Detection
- Association Rules
- Attribute Importance
- Classification
- Clustering
- Feature Extraction
- Regression

- Maximum prediction accuracy that avoids overfit.
- Supports sparse transactional data.
- Supports text data.

Usage:

Standard binary supervised classification algorithms require the presence of both positive and negative examples (counterexamples) of a target class. Anomaly Detection requires only the presence of examples of a single target class. In outlier detection, typical examples in a distribution are separated from the atypical (outlier) examples.

Help < Back Next > Finish Cancel

101534	M	38	Married	Argentina	E: 90,000 - 1...	Assoc-A	Other	3	4
101535	F	33	NeverM	United State...	L: 300,000 a...	Bach.	Prof.	2	4
101536	M	59	Married	United State...	E: 90,000 - 1...	10th	Crafts	3	6
101537	M	31	Married	United State...	E: 90,000 - 1...	< Bach.	Exec.	3	3
101538	M	39	Married	Germany	I: 170,000 - 1...	Assoc-A	Crafts	3	6

Start | Oracle Data Miner ... | Clustering.JPG - Paint | Clustering Models.txt... | My Documents | E:\Motaz\MIA\ODM | 8:37 AM

Oracle Data Miner: Decision Tree

Result Viewer: MINING_DATA_B25773_DT

File Publish Help

Tree | Results | Build Settings | Task

Nodes ☐ Show Leaves Only — Show Levels: 3

Node ID	Predicate	Predicted Value	Confidence	Cases	Support
⊟0	true	M	0.6955	890	1.0000
⊟1	HOUSEHOLD_SIZE is in { 3 6-8 }	M	0.9882	422	0.4742
5	AGE <= 26.5	M	0.8148	27	0.0303
6	AGE > 26.5	M	1.0000	395	0.4438
⊟2	HOUSEHOLD_SIZE is in { 1 2 4-5 9+ }	F	0.5684	468	0.5258
⊟3	OCCUPATION is in { Armed-F Crafts Farming Handl...	M	0.8495	93	0.1045
7	AGE <= 27.5	M	1.0000	31	0.0348
8	AGE > 27.5	M	0.7742	62	0.0697
⊟4	OCCUPATION is in { ? Cleric. Exec. House-s Machin...	F	0.6720	375	0.4213
9	HOUSEHOLD_SIZE is in { 1 2 }	F	0.5512	254	0.2854
10	HOUSEHOLD_SIZE is in { 4-5 9+ }	F	0.9256	121	0.1360

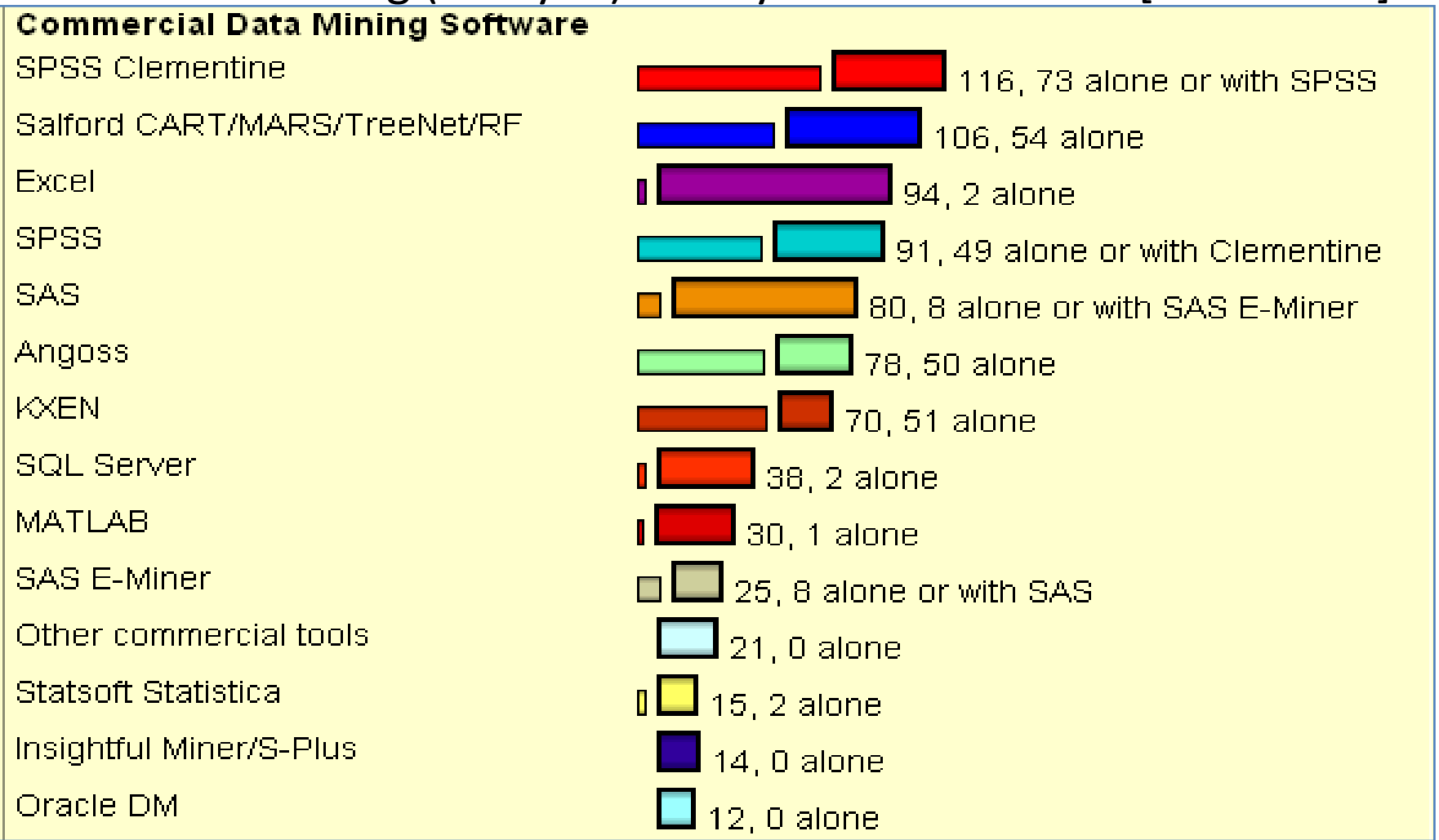
Predicted Target Value: M
Support: 0.0303
Confidence 0.8148
Cases: 27
Level: 2

Split Rules: (•) Full Rule ( ) Surrogate

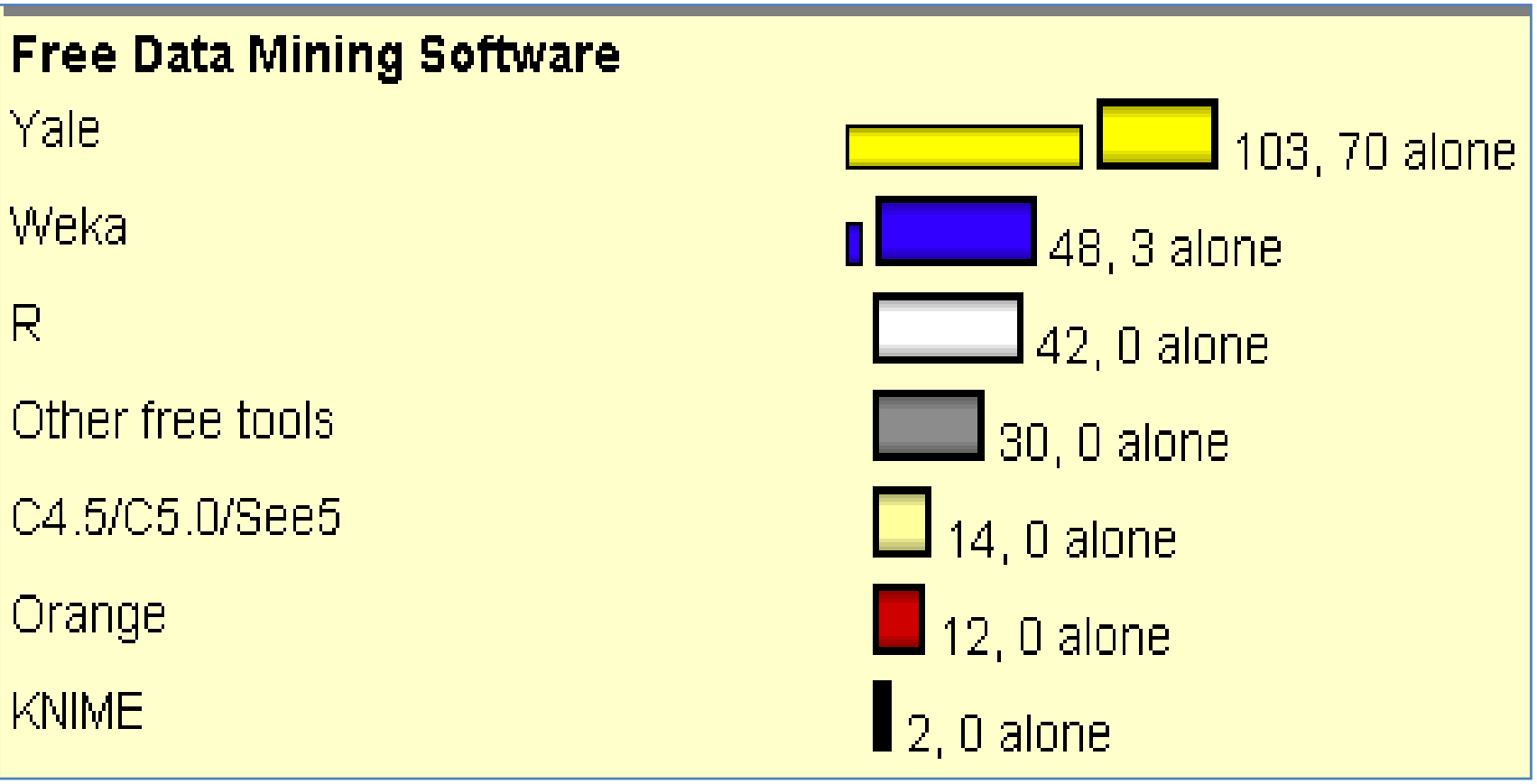
AGE <= 26.5 AND
HOUSEHOLD_SIZE is in { 3 6-8 }

Dnuggets.com Poll : Data Mining / Analytic Software Tools (May 2007)

- Data Mining (Analytic) tools you used in 2007: [534 voters]

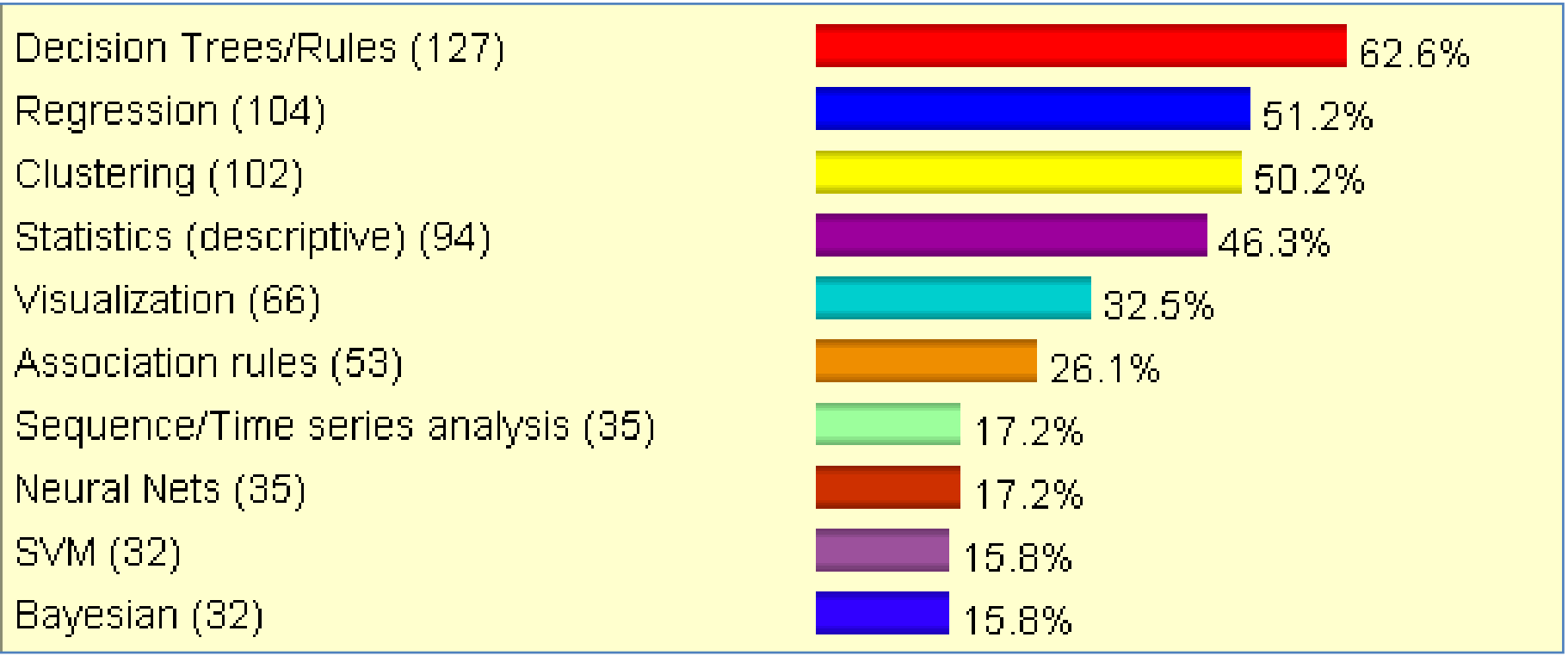


Dnuggets.com Poll : Data Mining / Analytic open source Software Tools (May 2007)



Kdnuggets.com Poll : Data Mining Methods (Mar 2007)

- **Data mining/analytic methods you used frequently in the past 12 months: [203 voters]**



Thank you: Q & A ?!