DATA WAREHOUSE E KNOWLEDGE DISCOVERY

Prof. Fabio A. Schreiber
Dipartimento di Elettronica e Informazione
Politecnico di Milano
DATA WAREHOUSE (DW)

• A TECHNIQUE FOR CORRECTLY ASSEMBLING AND MANAGING DATA COMING FROM DIFFERENT SOURCES TO OBTAIN A DETAILED VIEW OF AN ECONOMIC SYSTEM

• IT IS AN
  – INTEGRATED
  – PERMANENT
  – TIME VARIANT
  – TOPIC ORIENTED
  DATA COLLECTION TO SUPPORT MANAGERIAL DECISIONS

• IT IS THE SEPARATION ELEMENT BETWEEN OLTP AND DSS WORKLOADS
ON LINE TRANSACTION PROCESSING (OLTP)

• **TYPICAL OF ELECTRONIC DATA PROCESSING (EDP) APPLICATIONS**

• **TRANSACTIONS MUST POSSESS ACID PROPERTIES**
  – STRUCTURED AND REPETITIVE
  – SHORT AND ISOLATED

• **DATA MUST BE DETAILED AND UP-TO-DATE**

• **ACCESS TO DATA IS MAINLY MADE THROUGH THE PRIMARY KEY**

• **DATABASE SIZE VARIES BETWEEN $10^2$ MBYTE AND 10 GBYTE**

• **TRANSACTION THROUGHPUT IS THE MAIN PERFORMANCE METRICS**
ON LINE ANALYTICAL PROCESSING (OLAP)

- TYPICAL OF DECISION SUPPORT SYSTEMS

- WORKLOAD IS CONSTITUTED OF VERY COMPLEX QUERIES ACCESSING MORE THAN $10^6$ RECORDS

- DATA ARE OF HISTORICAL TYPE, AGGREGATED FROM DIFFERENT SOURCES

- WAREHOUSE DIMENSIONS GO BEYOND A TBYTE

- QUERY THROUGHPUT AND RESPONSE TIME ARE THE MAIN PERFORMANCE INDEXES
DATA MODELS FOR OLAP

• They must support sophisticated analyses and computations over different hierarchical dimensions

• The most suitable logical data model is a multidimensional structure - the data cube

• The dimensions of the cube are the attributes on which searches are made (keys)

• Each dimension in turn can be a hierarchy
  – date {day - month - quarter - year}
  – product {name - type - category}
    (Land Rover - Cross-Country - Motor Vehicle)

• The cells of the cube contain the metrical values pertaining to dimensional values
LOGICAL DATA MODELS FOR OLAP
INSURANCE COMPANY EXAMPLE

DIMENSIONS
METRICAL VALUES

YEAR

AGE

POLICY NUMBER, TOTAL AMOUNT

LIFE-HEALTH
THIRD-PARTY
FIRE-THEFT
RC-CAR

POLICY NUMBER, TOTAL AMOUNT

<20 20-30 30-40 40-50 50-60 60-70 70-80 >80

<20 20-30 30-40 40-50 50-60 60-70 70-80 >80


FIRE-THEFT
THIRD-PARTY
LIFE-HEALTH
RC-CAR
OLAP OPERATIONS

• **ROLL-UP**
  – INCREASES DATA AGGREGATION LEVEL

• **DRILL-DOWN**
  – INCREASES DATA DETAIL LEVEL

• **SLICE-AND-DICE**
  – SELECTS AND PROJECTS IN ORDER TO REDUCE DATA DIMENSION

• **PIVOTING**
  – SELECTS TWO DIMENSIONS TO AGGREGATE METRICAL DATA AROUND

• **RANKING**
  – SORTS DATA FOLLOWING PREDEFINED CRITERIA

• **TRADITIONAL OPERATIONS (SELECTION, COMPUTED ATTRIBUTES, ECC.)**
OLAP OPERATIONS

DRILL-DOWN

ROLL-UP

YEAR

TYPE

AGE

MILLISECOND

TYPE

AGE

MONTH
OLAP OPERATIONS

YEAR
AGE
TYPE

YEAR
AGE
TYPE

TYPE
YEAR

PIVOTING
OLAP OPERATIONS

SLICE AND DICE
WAREHOUSE CONSTRUCTION

- DATA COME FROM DIFFERENT AND “DIRTY” SOURCES
  - LEGACY NON DOCUMENTED SYSTEMS
  - PRODUCTION SYSTEMS WITHOUT ANY INTERNAL INTEGRITY CHECK
  - EXTERNAL SOURCES WITH DOUBTFUL QUALITY FEATURES

- IT IS MANDATORY TO RESTORE DATA QUALITY IN ORDER TO COMMIT RELIABLE DECISIONS TO THEM
WAREHOUSE CONSTRUCTION

• TOOLS FOR DATA QUALITY
  – FOR MIGRATION
    • TRANSFORM AND REFORMAT DATA FROM DIFFERENT SOURCES
  – FOR SCRUBBING
    • USE DOMAIN KNOWLEDGE TO SCRUB AND HOMOGENIZE
  – FOR AUDITING
    • DISCOVER RULES AND RELATIONS AMONG DATA AND VERIFY THEIR VALIDITY

• TOOLS FOR DATA LOADING
  – VERIFY REFERENTIAL INTEGRITY VIOLATIONS
  – SORT, AGGREGATE, BUILD DERIVED DATA
  – BUILD INDEXES AND OTHER ACCESS PATHS
  – REPRESENT A HEAVY BATCH WORKLOAD
  – NEED TO ORGANIZE LOADING OPERATIONS IN A PARALLEL OR IN AN INCREMENTAL WAY
WAREHOUSE CONSTRUCTION

• INCREMENTAL LOADING
  – PERFORMED DURING UPDATING; ONLY NEW OR MODIFIED TUPLES ARE LOADED
  – CONFLICTS WITH NORMAL WORK
  – REQUIRES SHORT TRANSACTIONS (<1000 RECORD)
  – NEEDS COORDINATION TO GUARANTEE INDEXES AND DERIVED DATA CONSISTENCY

• UPDATING
  – PERIODICALLY APPLIED FOLLOWING APPLICATIONS NEEDS
  – DUPLICATION SERVER USAGE
    • DATA DISPATCHING: FOR EACH CHANGE OF THE SOURCE TABLE, TRIGGERS ARE USED TO UPDATE A SNAPSHOT LOG TABLE WHICH IS PROPAGATED THEREAFTER
    • TRANSACTION DISPATCHING: THE STANDARD LOG IS MONITORED AND THE CHANGES ON THE REPLICATED TABLES ARE TRANSFERRED TO THE DUPLICATION SERVER
DW DESIGN METHODOLOGY

• INPUT DATA ANALYSIS
  – SELECTION OF RELEVANT INFORMATION SOURCES
  – TRANSLATION INTO A REFERENCE CONCEPTUAL MODEL (E-R)
  – INFORMATION SOURCES ANALYSIS
    • IDENTIFICATION OF FACTS, MEASURES, DIMENSIONS

• INTEGRATION IN A GLOBAL CONCEPTUAL SCHEMA

• DATA WAREHOUSE DESIGN
  – CONCEPTUAL
    • AGGREGATED DATA, HISTORICAL DATA, … INTRODUCTION
  – LOGICAL

• DATA MART DESIGN (MULTIDIMENSIONAL DB)
  – IDENTIFICATION OF FACTS AND DIMENSIONS
  – E-R SCHEMA RESTRUCTURING
  – DIMENSIONAL GRAPH DERIVATION
  – TRANSLATION INTO THE LOGICAL MODEL
FREQUENT FAILURE CAUSES

• IGNORE DATA QUALITY
  – ACCURACY
  – COMPLETENESS
  – CONSISTENCY
  – TIMELINESS
  – AVAILABILITY

• NOT STORING NECESSARY DATA
  – IGNORE DATA IN EXTERNAL SOURCES
  – IGNORE “SOFT” DATA (e.g. subjective judgements)
MULTIDIMENSIONAL OLAP SERVER (MOLAP)

- **DIRECTLY IMPLEMENTS THE CUBE MODEL**
  - **MULTIDIMENSIONAL MATRIX STRUCTURES**
    - **OPTIMAL FOR DENSE STRUCTURES**
    - **SEARCH BY ADDRESS BECOMES AN ALGEBRAIC COMPUTATION**
- **HIGH AND CONSTANT QUERY PROCESSING PERFORMANCE**
  - **SPECIALIZED ACCESS METHODS**
  - **AGGREGATION AND COMPILATION ARE MADE IN ADVANCE**
- **LIMITED SCALABILITY DUE TO PREPROCESSING**
- **MORE INGENUITY REQUIRED TO THE DATA ADMINISTRATOR**
RELATIONAL OLAP SERVER (ROLAP)

- USES A STANDARD RDBMS TO IMPLEMENT THE MULTIDIMENSIONAL STRUCTURE BY MEANS OF THE GROUP\_BY OPERATION
- THE SCHEMA HAS A STAR OR A SNOWFLAKE SHAPE (IT NORMALIZES HIERARCHIES)
  - CENTRAL FACT TABLE
    - TUPLES ARE CONSTITUTED OF POINTERS (EXTERNAL KEYS) TO THE DIMENSION TABLES AND OF THE VALUES FOR THE DESCRIBED COORDINATES \( f=(k_1, ..., k_n, v_1, ..., v_m) \)
  - DIMENSION TABLES
    - STORE TUPLES WITH ATTRIBUTES PERTAINING TO EACH DIMENSION \( d_i=(k_1, a_1, ..., a_n) \)
  - FACTS CONSTELLATION
    - MANY FACT TABLES SHARE DIMENSION TABLES HAVING THE SAME STRUCTURE
STAR SCHEMA
MAIN PROBLEMS IN DW DESIGN

• LOGICAL STRUCTURES DESIGN IN ORDER TO OPTIMIZE QUERIES
  – NEED TO MINIMIZE JOINS
    • DENORMALIZATION AND DATA REPLICATION
  – TABLES DIMENSIONS REDUCTION
    • HORIZONTAL PARTITIONING
    • VERTICAL PARTITIONING BY ROWS BREAKING (USEFUL FOR DRILL-DOWN)
MAIN PROBLEMS IN DW DESIGN

- PHYSICAL STRUCTURES DESIGN
  - CHOICE OF INDEXES
  - CHOICE OF MATERIALIZED VIEWS

- VIEW AND METADATA MAINTENANCE

- REPLICATION MANAGEMENT
  - HOW AND WHEN TO UPDATE

- CONSISTENCY MANAGEMENT

- APPLICATIONS IMPLEMENTATION
KNOWLEDGE HIERARCHY

DATA

INVOICES

SALES TRENDS

MARKET RULES

STRATEGIC DECISIONS

ELEMENTS (VOLUME)

DATA

INFORMATION

KNOWLEDGE

WISDOM

VALUE ADDED

STATISTICAL PROCESSING

KNOWLEDGE DISCOVERY PROCEDURES

EXPERIENCE
KNOWLEDGE DISCOVERY AND DATA MINING

• KNOWLEDGE DISCOVERY IN DATABASES AND DATA WAREHOUSES
  – TO IDENTIFY THE MOST SIGNIFICANT INFORMATION
  – TO SHOW IT TO THE USER IN THE MOST CONVENIENT WAY

• DATA MINING
  – ALGORITHM APPLICATION TO RAW DATA IN ORDER TO EXTRACT KNOWLEDGE (RELATIONS, PATHS, …)
  – PREDICTIVE AIM (SIGNAL ANALYSIS, VOICE RECOGNITION, ECC.)
  – DESCRIPTIVE AIM (DECISION SUPPORT SYSTEMS)
KNOWLEDGE DISCOVERY PROCESS (1)

EVEN IF SPECIALIZED TOOLS ARE AVAILABLE IT REQUIRES

– A COMPETENCE IN USED TECHNIQUES
– A VERY GOOD APPLICATION DOMAIN KNOWLEDGE

SEQUENTIAL STEPS

• SELECTION
  – CHOICE OF THE SAMPLE DATA THE ANALYSIS SHALL BE FOCUSED ON

• PREPROCESSING
  – DATA SAMPLING IN ORDER TO REDUCE THEIR VOLUME
  – DATA SCRUBBING FOR ERRORS AND OMISSIONS
KNOWLEDGE DISCOVERY PROCESS (2)

• TRANSFORMATION
  – DATA TYPES HOMOGENEIZATION AND/OR CONVERSION

• DATA MINING
  – CHOICE OF THE METHOD/ALGORITHM

• INTERPRETATION AND EVALUATION
  – RETRIEVED INFORMATION FILTERING
  – POSSIBLE REFINING BY PREVIOUS STEPS REPETITION
  – SEARCH RESULTS VISUAL PRESENTATION (GRAPHICAL OR LOGICAL)
KNOWLEDGE DISCOVERY PROCESS (3)

da: G. Piatesky-Shapiro 1996
DATA MINING APPLICATIONS

• CORRESPONDENCE AND RETAIL SALES
  – WHICH “SPECIAL OFFER” SHOULD BE MADE
  – HOW TO LOCATE GOODS ON THE SHELVES

• MARKETING
  – SALES FORECASTS
  – PRODUCTS BUYING PATHS

• BANKS
  – LOANS CONTROL
  – CREDIT CARDS USAGE (AND MISUSE)

• TELECOMMUNICATIONS
  – SPECIAL FARES
DATA MINING APPLICATIONS

• ASTRONOMY AND ASTROPHYSICS
  – STARS AND GALAXIES CLASSIFICATION

• CHEMICAL AND PHARMACEUTIC RESEARCH
  – NEW COMPOUNDS DISCOVERY
  – INTERRELATIONS AMONG CHEMICAL COMPOUNDS

• MOLECULAR BIOLOGY
  – PATTERNS IN GENETICAL DATA AND IN MOLECULAR STRUCTURES

• REMOTE SURVEY AND METEOROLOGY
  – SATELLITE DATA ANALYSIS

• ECONOMIC AND DEMOGRAPHIC STATISTICS
  – CENSUS DATA ANALYSIS
DATA MINING ALGORITHMS STRUCTURE

• MODEL REPRESENTATION
  FORMALISMS TO REPRESENT AND DESCRIBE POSSIBLE PATHS

• MODEL EVALUATION
  STATISTICAL OR LOGICAL ESTIMATE OF THE CORRESPONDENCE OF A PATH TO THE SEARCH CRITERIA

• SEARCH METHOD
  – OF PARAMETERS
  – OF MODEL
    THE PARAMETERS ARE APPLIED TO MODELS BELONGING TO THE SAME FAMILY, DIFFERENTIATED BY THE REPRESENTATION TYPE, FOR QUALITY EVALUATION
WHAT KIND OF INFORMATION DO WE GET?

- **ASSOCIATIONS**
  - SET OF RULES SPECIFYING THE JOINT OCCURRENCE OF TWO (OR MORE) ELEMENTS

- **SEQUENCES**
  - POSSIBILITY OF STATING TEMPORAL SEQUENCES OF EVENTS

- **CLASSIFICATIONS**
  - GROUPING OF ELEMENTS INTO CLASSES FOLLOWING A GIVEN MODEL

- **CLUSTERS**
  - GROUPING OF ELEMENTS INTO CLASSES WHICH HAVE NOT BEEN DEFINED A-PRIORI

- **TRENDS**
  - DISCOVERY OF PECULIAR TEMPORAL PATHS HAVING A FORECASTING VALUE
DATA MINING TECHNIQUES AND MODELS

• NEURAL NETWORKS
  – VERY POWERFUL TOOL
  – LEARNING CAPABILITY
  – OPAQUE BEHAVIOUR

• INDUCTION
  – DECISION TREES (HIERARCHY OF \textit{if - then} DECISIONS)
  – RULE INDUCTION (NON HIERARCHICAL SETS DERIVED FROM DECISION TREES)
  – SELF EXPLAINING
DATA MINING TECHNIQUES AND MODELS

• RULES DISCOVERY
  – ASSOCIATION RULES (SIMULTANEOUS EVENTS $\chi \Rightarrow \gamma$)
  – SEQUENTIAL ASSOCIATIONS

• DATA VISUALIZATION
  – DATA ARE PREPARED AND GRAPHICALLY PRESENTED IN ORDER TO EVIDENTIATE POSSIBLE IRREGULARITIES OR STRANGE PATTERNS
# DATA MINING TECHNIQUES

<table>
<thead>
<tr>
<th>INFORMATION TECHNIQUES</th>
<th>ASSOCIATIONS</th>
<th>SEQUENCES</th>
<th>CLASSIFICATIONS</th>
<th>CLUSTERS</th>
<th>TRENDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEURAL NETWORKS</td>
<td></td>
<td></td>
<td>¥</td>
<td>¥</td>
<td>¥</td>
</tr>
<tr>
<td>INDUCTION</td>
<td></td>
<td></td>
<td>¥</td>
<td>¥</td>
<td>¥</td>
</tr>
<tr>
<td>RULES DISCOVERY</td>
<td>¥</td>
<td>¥</td>
<td></td>
<td></td>
<td>¥</td>
</tr>
<tr>
<td>DATA VISUALIZATION</td>
<td>¥</td>
<td>¥</td>
<td></td>
<td>¥</td>
<td></td>
</tr>
</tbody>
</table>

© Fabio A. Schreiber
THE “MARKET BASKET”

• THE BEST-KNOWN MODEL ON WHICH DATA MINING TECHNIQUES ARE APPLIED

• MAINLY, BUT NOT EXCLUSIVELY, USED FOR RETAIL SALE PROBLEMS

• THE GOAL IS TO DISCOVER RECURRENT PATTERNS IN DATA (ASSOCIATION RULES)
THE “MARKET BASKET”

\[ I = \{i_1, ..., i_k\} \quad \text{SET OF } k \text{ ELEMENTS (ITEM)} \]
\[ B = \{b_1, ..., b_n\} \quad \text{SET OF } n \text{ SUBSETS (BASKET) OF } I \]
\[ b_i \subseteq I \]

- \( I \)
  - GOODS IN A SUPERMARKET
  - WORDS IN A DICTIONARY
- \( B \)
  - A CUSTOMER’S PURCHASE
  - A DOCUMENT IN A CORPUS
- ASSOCIATION RULE \( i_1 \Rightarrow i_2 \)
  - \( i_1 \) AND \( i_2 \) SHOW TOGETHER IN AT LEAST \( s\% \) OF THE \( n \) BASKET (SUPPORT)
  - OF ALL THE BASKETS CONTAINING \( i_1 \) AT LEAST \( c\% \) CONTAIN ALSO \( i_2 \) (CONFIDENCE)
EXAMPLES OF RULES

• RULES HAVING “DIET COKE” AS CONSEQUENT
  – HELP IN DEFINING THE STRATEGIES TO INCREASE THE SALES OF A SPECIFIC PRODUCT

• RULES HAVING “GRISSINI” AMONG THEIR ANTECEDENTS
  – HELP IN UNDERSTANDING THE IMPACT OF CEASING A SPECIFIC PRODUCT SALE

• RULES HAVING “WURSTEL” AMONG THEIR ANTECEDENTS AND “MUSTARD” AS CONSEQUENT
  – EVIDENTIATE THE PRODUCT COUPLINGS IN THE ANTECEDENT INDUCING THE SALE OF THE CONSEQUENT
EXAMPLES OF RULES

- RULES CONNECTING ELEMENTS IN SHELF A TO ELEMENTS IN SHELF B
  - HELP TO AN EFFECTIVE ALLOCATION OF PRODUCTS ON THE SHELVES

- THE BEST $k$ RULES HAVING "GRISSINI" IN THE CONSEQUENT
  - IN TERMS OF CONFIDENCE
  - IN TERMS OF SUPPORT
ANY PROBLEM?

\[ c \Rightarrow \text{COFFEE IS IN THE BASKET} \]
\[ t \Rightarrow \text{TEA IS IN THE BASKET} \]

\[ \bar{c} \Rightarrow \text{NO COFFEE IN THE BASKET} \]
\[ \bar{t} \Rightarrow \text{NO TEA IN THE BASKET} \]

\[
\begin{array}{c|c|c}
  & c & \bar{c} \\
\hline
  t  & 20 & 5 & 2 \\
  \bar{t} & 70 & 5 & 75 \\
\hline
\Sigma & 90 & 10 & 100 \\
\end{array}
\]

\[ t \Rightarrow c \quad \text{IS TRUE???} \]
\[ s = 20\% \]
\[ c = \frac{P[t \land c]}{P[t]} = \frac{20}{25} = 80\% \]

PERHAPS, BUT ... THOSE WHO BUY COFFEE ANYHOW REACH 90% !!!

WARNING! A CORRELATION EXISTS BETWEEN TEA AND COFFEE

\[ r = \frac{P[t \land c]}{(P[t] \times P[c])} = 0.89 \]
DATA MINING IN A RELATIONAL ENVIRONMENT

- SQL EXTENSION BY INTEGRATING DATA MINING OPERATORS INTO THE LANGUAGE
- INTEGRATION OF AN SQL SERVER WITH A DATA MINING ENGINE
  - MINE RULE FOR ASSOCIATION RULES
  - MINE CLASSIFICATION FOR CLASSIFICATION PROBLEMS
  - MINE INTERVAL FOR DISCRETIZING CONTINUOUS ATTRIBUTES
MINE RULE: EXAMPLE

MINE RULE SimpleAssociation AS
SELECT DISTINCT 1..n item AS BODY, 1..1 item AS HEAD, SUPPORT, CONFIDENCE
FROM Purchase
GROUP BY transaction
EXTRACTING RULES WITH SUPPORT: 0.1, CONFIDENCE: 0.2

<table>
<thead>
<tr>
<th>TR</th>
<th>CLIENT</th>
<th>ITEM</th>
<th>DATE</th>
<th>PRICE</th>
<th>QUANT.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rossi</td>
<td>Ski-trousers</td>
<td>17/12</td>
<td>90.000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Rossi</td>
<td>Climbing-boots</td>
<td>17/12</td>
<td>180.000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Bianchi</td>
<td>Sport shirt</td>
<td>18/12</td>
<td>70.000</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Bianchi</td>
<td>Brown shoes</td>
<td>18/12</td>
<td>250.000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Bianchi</td>
<td>Jacket</td>
<td>18/12</td>
<td>400.000</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Rossi</td>
<td>Jacket</td>
<td>18/12</td>
<td>400.000</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Bianchi</td>
<td>Sport shirt</td>
<td>19/12</td>
<td>70.000</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Bianchi</td>
<td>Jacket</td>
<td>19/12</td>
<td>400.000</td>
<td>2</td>
</tr>
</tbody>
</table>
# MINERule: Example

<table>
<thead>
<tr>
<th>BODY</th>
<th>HEAD</th>
<th>SUPPORT</th>
<th>CONFIDENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ski-trousers</td>
<td>Climbing-boots</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Climbing-boots</td>
<td>Ski-trousers</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Sport shirt</td>
<td>Brown shoes</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>Sport shirt</td>
<td>Jacket</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Brown shoes</td>
<td>Sport shirt</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Brown shoes</td>
<td>Jacket</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Jacket</td>
<td>Sport shirt</td>
<td>0.5</td>
<td>0.66</td>
</tr>
<tr>
<td>Jacket</td>
<td>Brown shoes</td>
<td>0.25</td>
<td>0.33</td>
</tr>
<tr>
<td>Sport shirt,</td>
<td>Brown shoes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sport shirt,</td>
<td>Jacket</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Brown shoes,</td>
<td>Jacket</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brown shoes,</td>
<td>Sport shirt</td>
<td>0.25</td>
<td>1</td>
</tr>
</tbody>
</table>

© Fabio A. Schreiber
CLASSIFICATION PROBLEM

EACH ELEMENT (RECORD) OF A DATA SET IS ASSOCIATED WITH A DISTINCTIVE FEATURE CALLED CLASS, ON THE BASIS OF PREVIOUS EXPERIENCES AND OBSERVATIONS

• PHASE 1: TRAINING
  – A MODEL IS BUILT ON THE KNOWN SET (TRAINING SET) IN SUCH A WAY AS EACH CLASS IS A PARTITION OF THE SET

• PHASE 2: APPLICATION
  – THE MODEL IS USED TO CLASSIFY NEW DATA
CLASSIFICATION PROBLEM: AN EXAMPLE

- **PHASE 1: TRAINING**

  ```sql
  MINE CLASSIFICATION CarInsuranceRules AS
  SELECT DISTINCT RULES ID, *, CLASS
  FROM CarInsurance
  CLASSIFY BY Risk
  ```

- **PHASE 2: APPLICATION**

  ```sql
  MINE CLASSIFICATION TEST ClassifiedApplicants AS
  SELECT DISTINCT *, CLASS
  FROM Applicants
  USING CLASSIFICATION FROM CarInsuranceRules AS RULES
  ```
### CLASSIFICATION PROBLEM: AN EXAMPLE

<table>
<thead>
<tr>
<th>AGE</th>
<th>CAR TYPE</th>
<th>RISK</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>sports</td>
<td>high</td>
</tr>
<tr>
<td>43</td>
<td>family</td>
<td>low</td>
</tr>
<tr>
<td>68</td>
<td>family</td>
<td>low</td>
</tr>
<tr>
<td>32</td>
<td>truck</td>
<td>low</td>
</tr>
<tr>
<td>23</td>
<td>family</td>
<td>high</td>
</tr>
<tr>
<td>18</td>
<td>family</td>
<td>high</td>
</tr>
<tr>
<td>20</td>
<td>family</td>
<td>high</td>
</tr>
<tr>
<td>45</td>
<td>sports</td>
<td>high</td>
</tr>
<tr>
<td>50</td>
<td>truck</td>
<td>low</td>
</tr>
<tr>
<td>64</td>
<td>truck</td>
<td>high</td>
</tr>
<tr>
<td>46</td>
<td>family</td>
<td>low</td>
</tr>
<tr>
<td>40</td>
<td>family</td>
<td>low</td>
</tr>
</tbody>
</table>

1. **IF** Age \( \leq 23 \) **THEN** Risk **IS** High;
2. **IF** CarType = sports **THEN** Risk **IS** High;
3. **IF** CarType IN {family, truck} **AND** Age > 23 **THEN** Risk **IS** Low;
4. **DEFAULT** Risk **IS** Low

### MINE CLASSIFICATION TEST

<table>
<thead>
<tr>
<th>AGE</th>
<th>CAR TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>family</td>
</tr>
<tr>
<td>60</td>
<td>family</td>
</tr>
<tr>
<td>35</td>
<td>sports</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AGE</th>
<th>CAR TYPE</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>family</td>
<td>high</td>
</tr>
<tr>
<td>60</td>
<td>family</td>
<td>low</td>
</tr>
<tr>
<td>35</td>
<td>sports</td>
<td>high</td>
</tr>
</tbody>
</table>
DISCRETIZATION PROBLEM

ONE MUST TRANSFORM A NUMERIC ATTRIBUTE IN A CATEGORIAL ONE BY DIVIDING THE NUMERIC DOMAIN INTO A SET OF INTERVALS, EACH OF WHICH IS LABELLED WITH A CLASS LABEL, CONSTITUTING THE CATEGORIAL DOMAIN

• DISCRETIZATION METHODS
  – NON SUPERVISIONED
    ONLY THE VALUES DISTRIBUTION OF THE ATTRIBUTE IS CONSIDERED
  – SUPERVISIONED
    AN EFFORT IS MADE TO KEEP AS MUCH INFORMATION AS POSSIBLE ABOUT THE CLASSES OF A TUPLE ASSOCIATED ATTRIBUTE
DISCRETIZATION PROBLEM

SIMPLE DISCRETIZATION

– **SAME WIDTH INTERVALS**
  THE NUMERIC DOMAIN IS DIVIDED INTO A GIVEN NUMBER OF EQUAL WIDTH INTERVALS

– **SAME FREQUENCY INTERVALS**
  INTERVALS ARE NARROWER WHERE VALUES ARE DENSE AND WIDER WHERE VALUES ARE SPARSE

MINE INTERVAL SameWidthInterval AS
SELECT DISTINCT ID, LOWER, UPPER
GENERATING DiscreteCarInsurance
FROM CarInsurance
DISCRETIZE Age BY WIDTH USING 3 INTERVALS
### DISCRETIZATION PROBLEM: AN EXAMPLE

<table>
<thead>
<tr>
<th>ID</th>
<th>LOWER</th>
<th>UPPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>52</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AGE</th>
<th>CAR TYPE</th>
<th>RISK</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>sports</td>
<td>high</td>
</tr>
<tr>
<td>43</td>
<td>family</td>
<td>low</td>
</tr>
<tr>
<td>68</td>
<td>family</td>
<td>low</td>
</tr>
<tr>
<td>32</td>
<td>truck</td>
<td>low</td>
</tr>
<tr>
<td>23</td>
<td>family</td>
<td>high</td>
</tr>
<tr>
<td>18</td>
<td>family</td>
<td>high</td>
</tr>
<tr>
<td>20</td>
<td>family</td>
<td>high</td>
</tr>
<tr>
<td>45</td>
<td>sports</td>
<td>high</td>
</tr>
<tr>
<td>50</td>
<td>truck</td>
<td>low</td>
</tr>
<tr>
<td>64</td>
<td>truck</td>
<td>high</td>
</tr>
<tr>
<td>46</td>
<td>family</td>
<td>low</td>
</tr>
<tr>
<td>40</td>
<td>family</td>
<td>low</td>
</tr>
</tbody>
</table>