What is Data?

http://www.kdnuggets.com/datasets/index.html

	S*tudies*	*Education*	*Works*	*Income (D)*
1	**Poor**	**SPM**	**Poor**	**None**
2	**Poor**	**SPM**	**Good**	**Low**
3	**Moderate**	**SPM**	**Poor**	**Low**
4	**Moderate**	**Diploma**	**Poor**	**Low**
5	**Poor**	**SPM**	**Poor**	**None**
6	**Moderate**	**Diploma**	**Poor**	**Low**
7	**Good**	**MSC**	**Good**	**Medium**
:				
99	**Poor**	**SPM**	**Good**	**Low**
100	**Moderate**	**Diploma**	**Poor**	**Low**

Knowledge

studies(Poor) AND work(Poor) => income(None)

studies(Poor) AND work(Good) => income(Low)

education(Diploma) => income(Low)

education(MSc) => income(Medium) OR income(High)

studies(Mod) => income(Low)

studies(Good) => income(Medium) OR income(High)

education(SPM) AND work(Good) => income(Low)

Data Mining: Definition

Extraction of knowledge from data. Exploration and analysis of large quantities of data to discover meaningful pattern from data.

Motivation

1. Huge amounts of data
2. Important need for turning data into useful information
3. Fast growing amount of data, collected and stored in large and numerous databases exceeded the human ability for comprehension without powerful tools.

Evolution of Database Technology

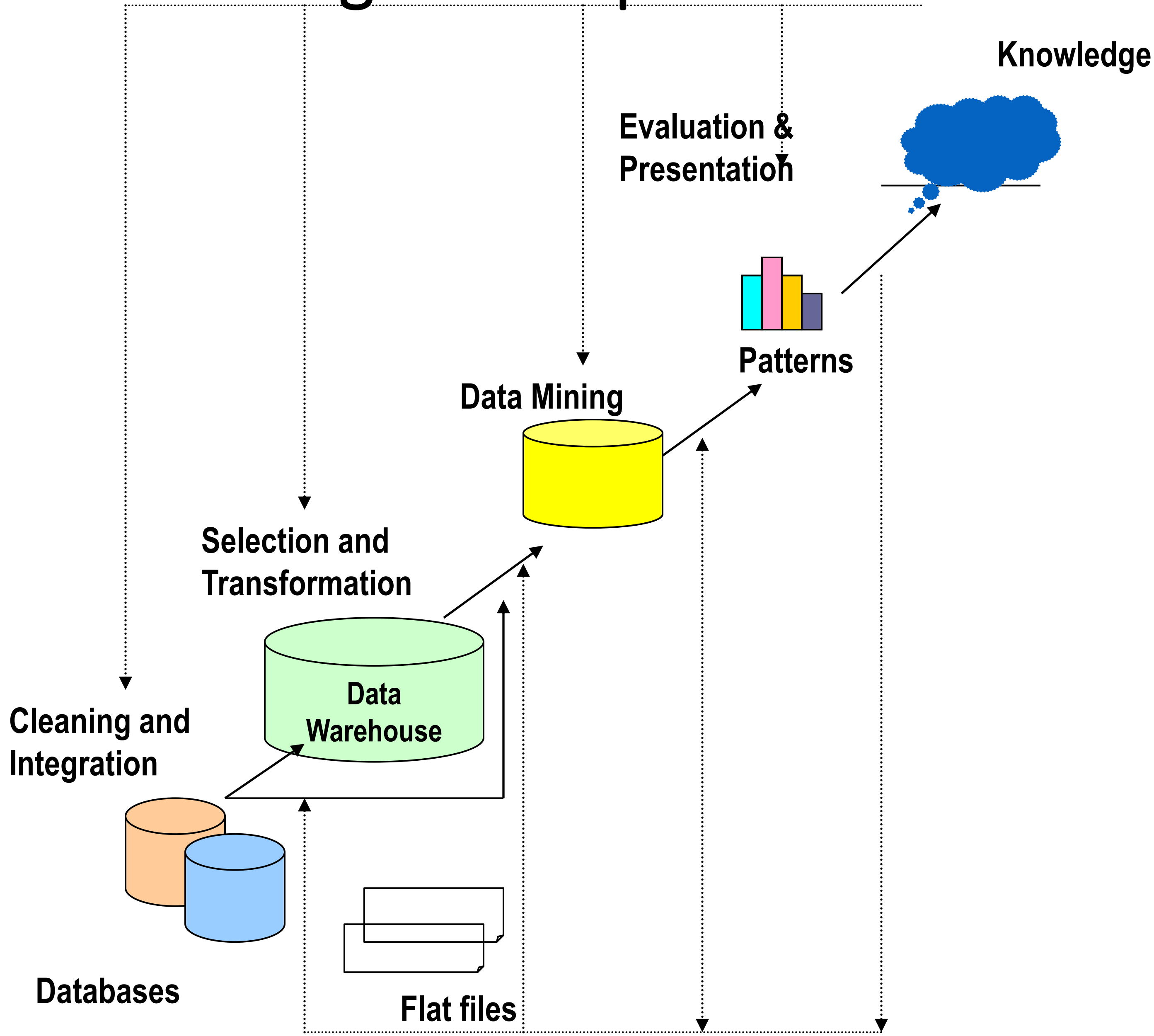
YEAR	TECHNOLOGY
1960s:	Data collection, database creation, IMS and network DBMS
1970s:	Relational data model, relational DBMS implementation
1980s:	RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
	Application-oriented DBMS (spatial, scientific, engineering, etc.)
1990s:	Data mining, data warehousing, multimedia databases, and Web databases
2000s	Stream data management and mining
	Data mining with a variety of applications
	Web technology and global information systems

The Evolution of Data Mining

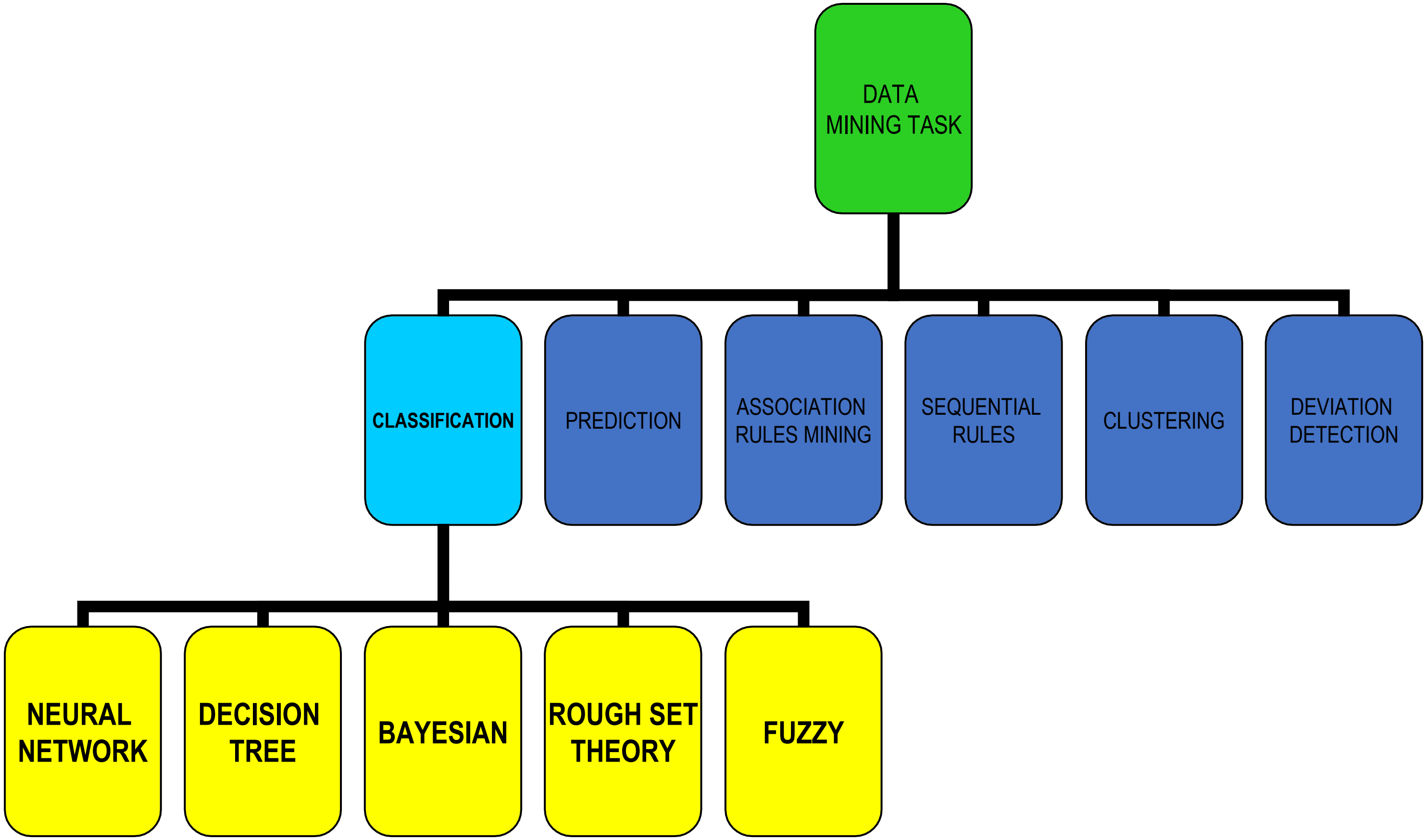
- Data mining is a natural development of the increased use of computerized databases to store data and provide answers to business analysts.

Evolutionary Step	**Business Question**
Data Collection (1960s)	"What was my total revenue in the last five years?
Data Access (1980s)	"What were unit sales in Selangor last March?
Data Warehousing and Decision Support	"What were unit sales in Selangor last March? Drill down to Kuala Lumpur.
Data Mining	"What's likely to happen to Selangor unit sales next month?Why?

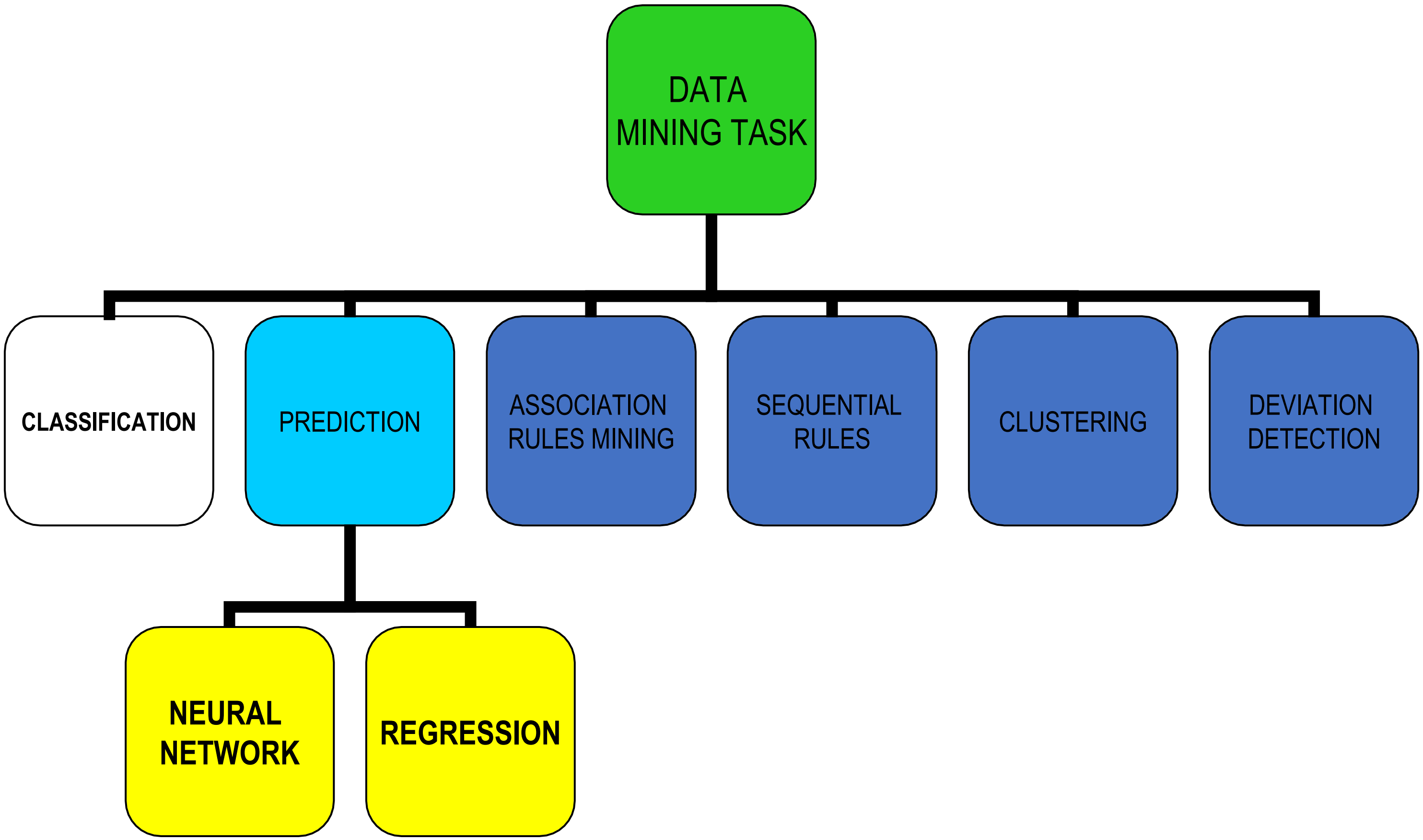
Data Mining as a Step of KDD



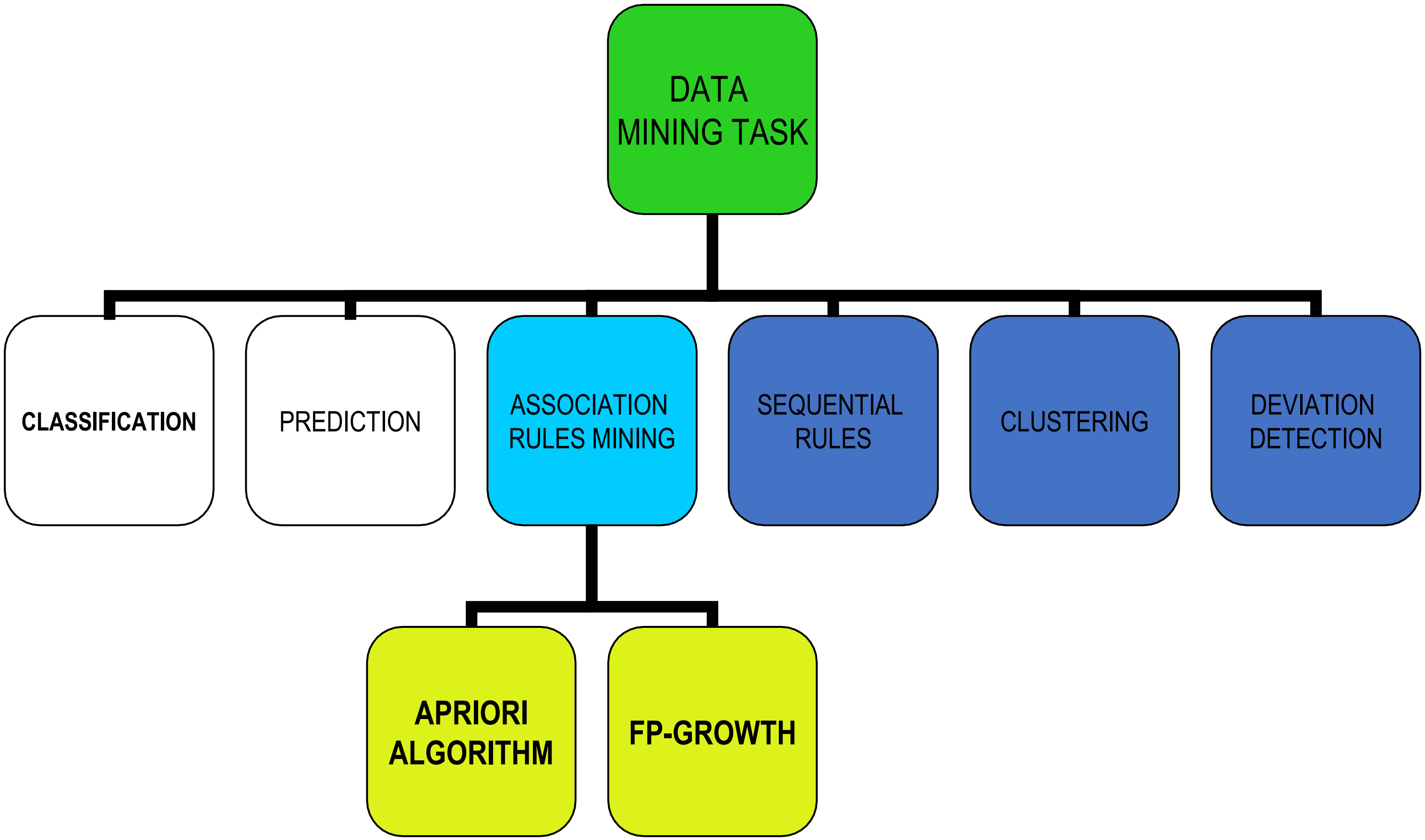
Data Mining Task &Techniques



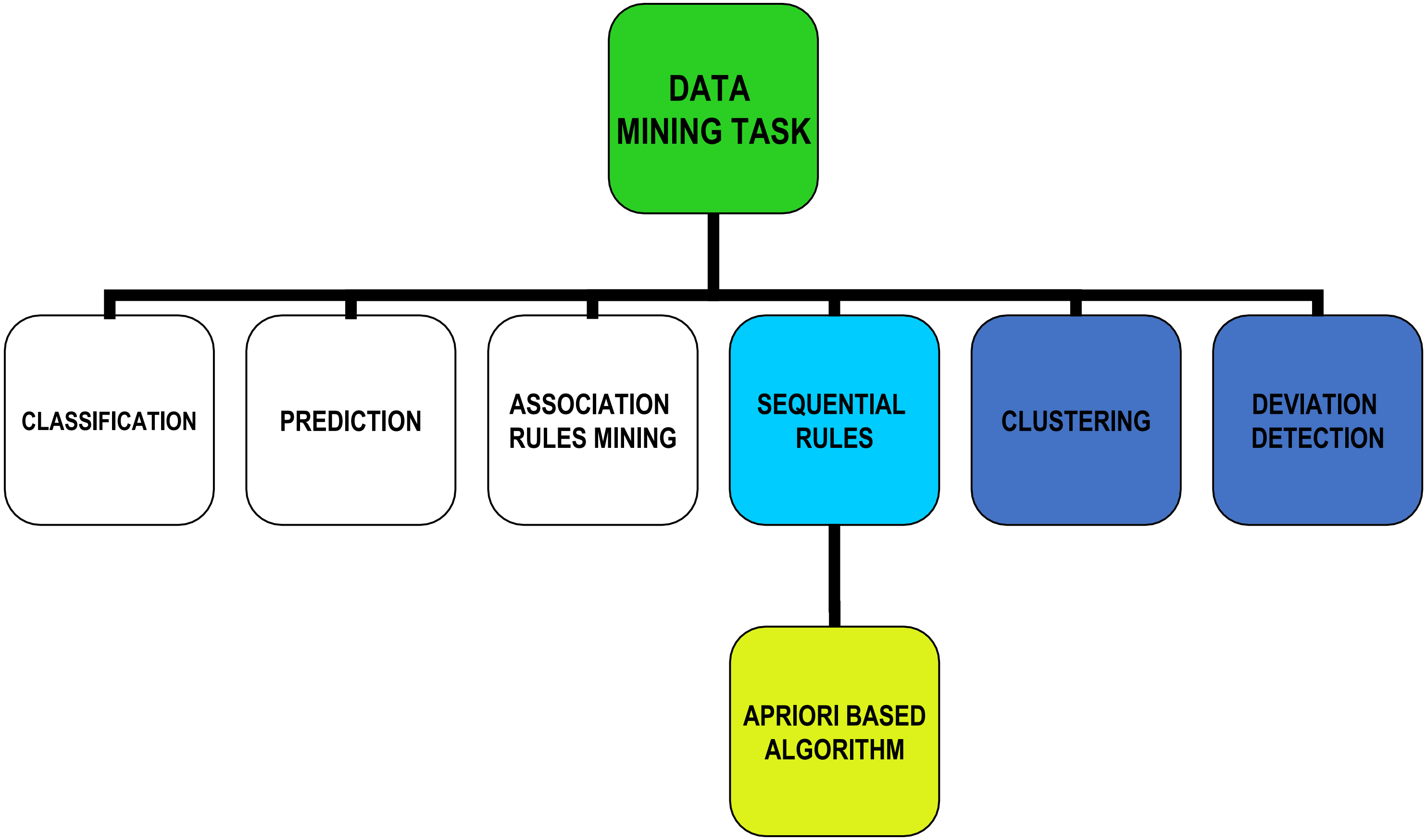
Data Mining Task &Techniques



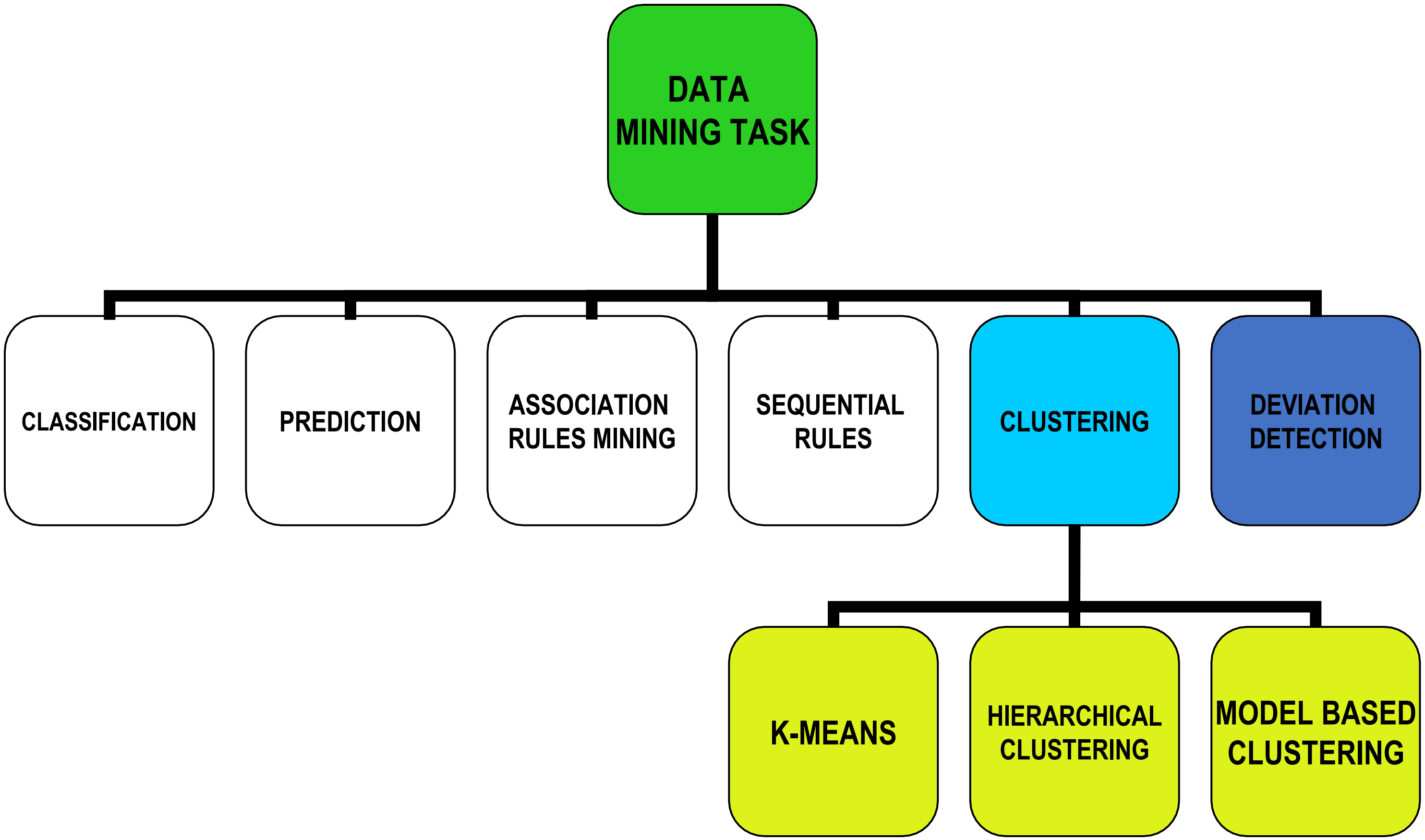
Data Mining Task &Techniques



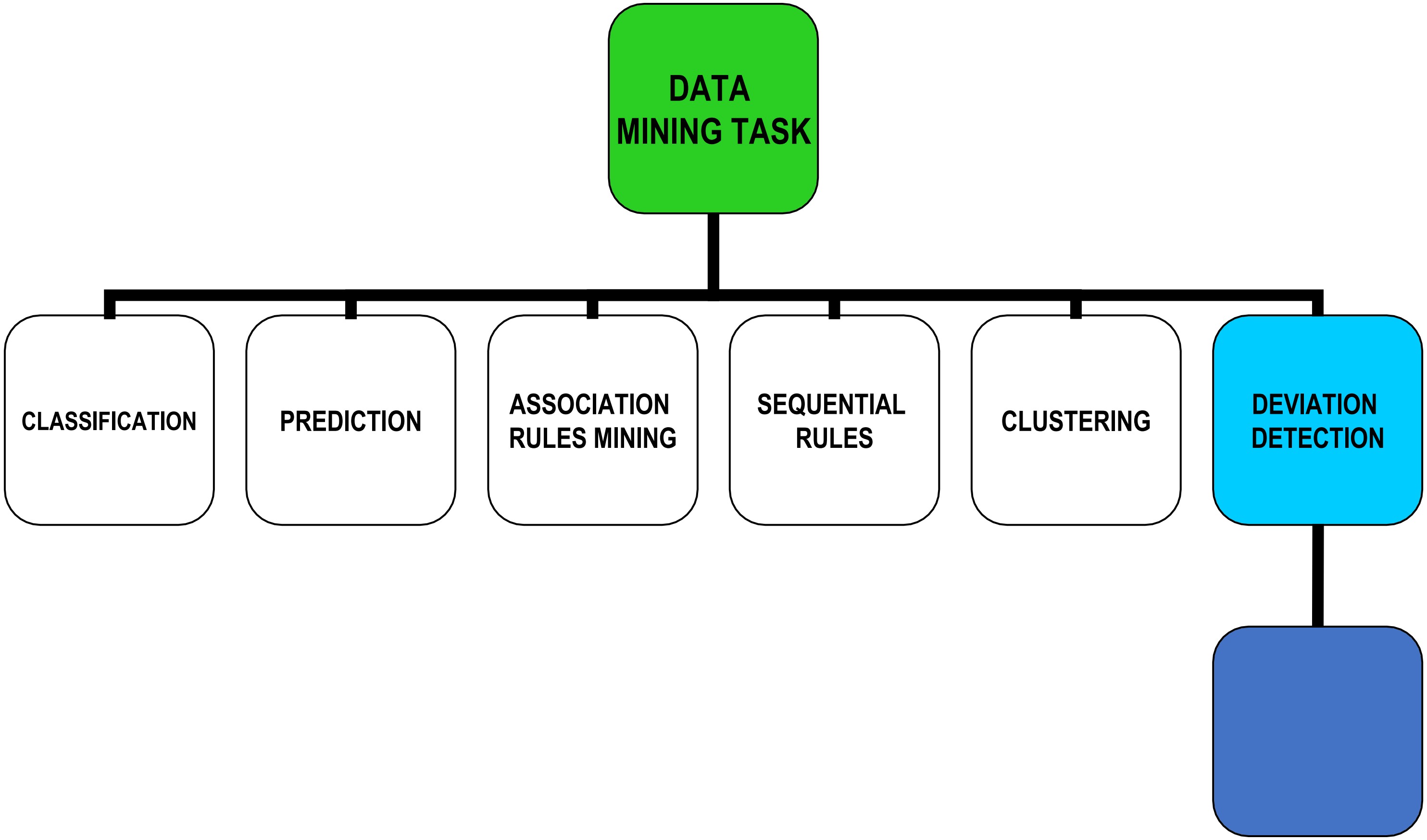
Data Mining Task &Techniques



Data Mining Task &Techniques



Data Mining Task &Techniques



Classification of Data Mining Systems

Kinds of DB

Relational
Data warehouse
Transactional DB
Advanced DB system
Flat files
WWW

Kinds of Knowledge

Classification
Association
Clustering
Prediction
Sequential
:

Techniques used

DB oriented techniques
Statistic
Machine learning
Pattern recognition
Neural Network
Rough Set etc

Application adapted

Finance
Marketing
Medical
Stock
Telecommunication, etc

DATA MINING MULTIDICIPLINES

- Database Technology
- Statistic
- Machine Learning
- High Performance Computing
- Information Science
- Information Retrieval
- Visualisation
- Business Intelligence
- Soft Computing
- Pattern Recognition

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
  - Stream data mining
  - DNA and bio-data analysis

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

- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association

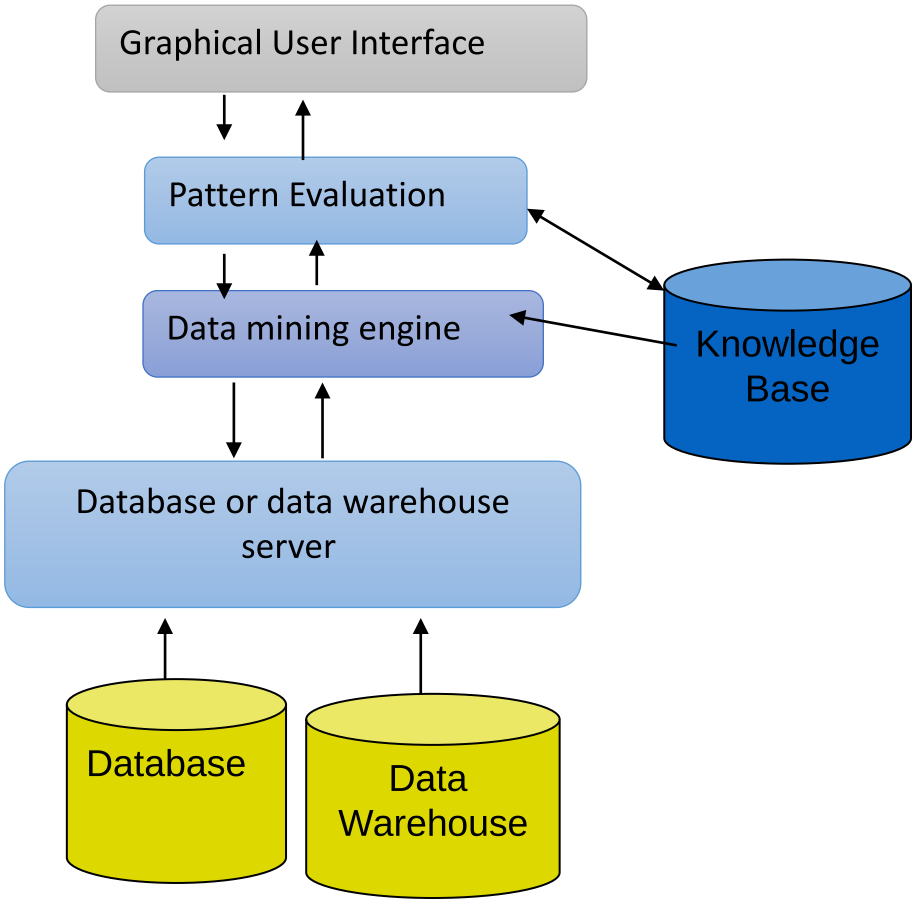
Market Analysis and Management

- Customer profiling
  - What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers
- Provision of summary information
  - multidimensional summary reports
  - statistical summary information (data central tendency and variation)

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - <u>Auto insurance</u>: ring of collisions
  - <u>Money laundering:</u> suspicious monetary transactions
  - <u>Medical insurance</u>
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests

Data Mining Engine



- DM system consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis, evolution and deviation analysis.

EXAMPLE
Building a DM Model

Data Mining algorithm

TRAINING DATA

	Studies	*Education*	*Works*	*Income (D)*
1	Poor	SPM	Poor	None
2	Poor	SPM	Good	Low
3	Moderate	SPM	Poor	Low
4	Moderate	Diploma	Poor	Low
5	Poor	SPM	Poor	None
6	Moderate	Diploma	Poor	Low
7	Good	MSC	Good	Medium
:				
99	Poor	SPM	Good	Low
100	Moderate	Diploma	Poor	Low

DM MODEL

Classification
Rules/ Classifier

1. If studies="poor" and work="poor" then Income="poor"
2. If studies="good" and work="poor" then Income="low"
3. If studies="good" then Income="good"
4. ..
5. ..

:
N

EXAMPLE
Applying a DM model

TEST DATA

	Studies	*Education*	*Works*	*Income (D)*
1	Poor	SPM	Poor	?
2	Poor	SPM	Good	?
3	Moderate	SPM	Poor	?
4	Moderate	Diploma	Poor	?
5	Poor	SPM	Poor	?
6	Moderate	Diploma	Poor	?
7	Good	MSC	Good	?
:				
m	Poor	SPM	Good	?
m+n	Moderate	Diploma	Poor	?

DATA MINING MODEL

Classification Rules/ Classifier

1. If studies="poor" and work="poor" then Income="poor"
2. If studies="good" and work="poor" then Income="low"
3. If studies="good" then Income="good"
4. ..
5. ..

:

N

DECISION MAKING

DATASET

	Studies	*Education*	*Works*	*Income (D)*
1	Poor	SPM	Poor	None
2	Poor	SPM	Good	Low
3	Moderate	SPM	Poor	Low
4	Moderate	Diploma	Poor	Low
5	Poor	SPM	Poor	None
6	Moderate	Diploma	Poor	Low
7	Good	MSC	Good	Medium
:				
99	Poor	SPM	Good	Low
100	Moderate	Diploma	Poor	Low

PATTERN/KNOWLEDGE/RULES

studies(Poor) AND work(Poor) => income(None)

studies(Poor) AND work(Good) => income(Low)

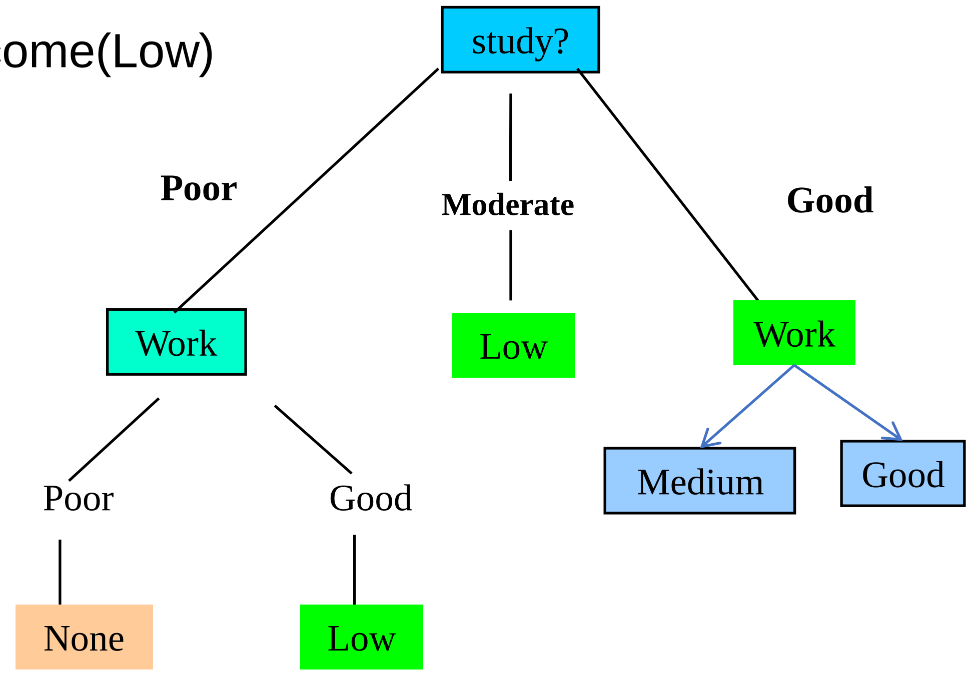
education(Diploma) => income(Low)

education(MSc) => income(Medium) OR income(High)

studies(Mod) => income(Low)

studies(Good) => income(Medium) OR income(High)

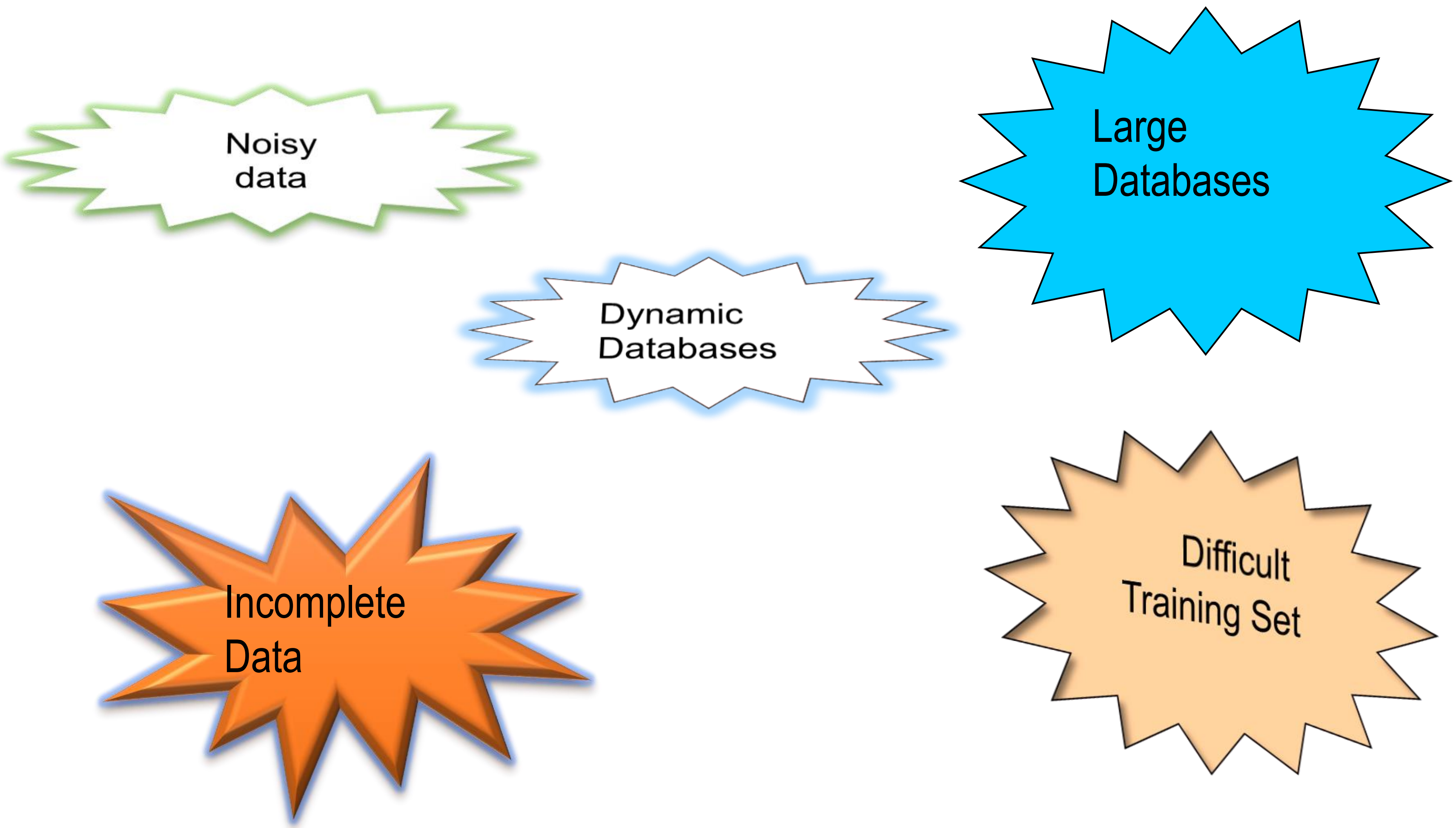
education(SPM) AND work(Good) => income(Low)



Comparing DATA MINING MODELS

- Predictive Accuracy
- Speed
- Robustness
- Scalability
- Interpretability

Data Mining : Problems and Challenges



Performance Issues

Cost of the Learning Set

-number of examples necessary for training

-cost of assuring the good accuracy

Performance Issues

Time and Memory Constraint

-time complexity of the learning phase

-time taken for evaluation

-time it takes to reach a certain level of accuracy

Performance Issues

Predictive Ability

-to be able to predict the correct decision towards the test or unseen data

-involve the generation of rules

-measuring the quality or accuracy of rules

SUMMARY

- Data mining can best be described as a business intelligence (BI) technology that has various techniques to extract comprehensible, hidden and useful information from a population of data.

- BI technology makes it possible to discover hidden trends and patterns in large amounts of data.

- The output of a data mining exercise can take the form of patterns, trends or rules that are implicit in the data.

- Through data mining and the new knowledge it provides, individuals are able to leverage the data to create new opportunities or value for their organizations

Current Technology in Data Mining and Business Intelligence

Data Analysis and Data Mining:

- Exploratory and automated data analysis
- Knowledge-based analysis
- Statistical pattern recognition
- Data mining algorithms and processes
- Classification, projection, regression, optimization clustering
- Information extraction and retrieval
- Multivariate data visualization

Applications and Tools:

- Visualization tools
- Applications (e.g. commerce, engineering, finance, manufacturing, science)
- Human-computer interaction in intelligence data analysis
- Business intelligence and data analysis systems and tools

SUMMARY

Data Mining Methods

- decision trees, classification, association, clustering, attributes, statistical modeling, Bayesian classification, k-nearest neighbors, CART. Extensive use of SPSS' Clementine data mining suite.

Applied Data Mining

- Statistical model building and deployment. Model choice. Visualization, report writing, graphical presentation. Extensive use of data mining software.

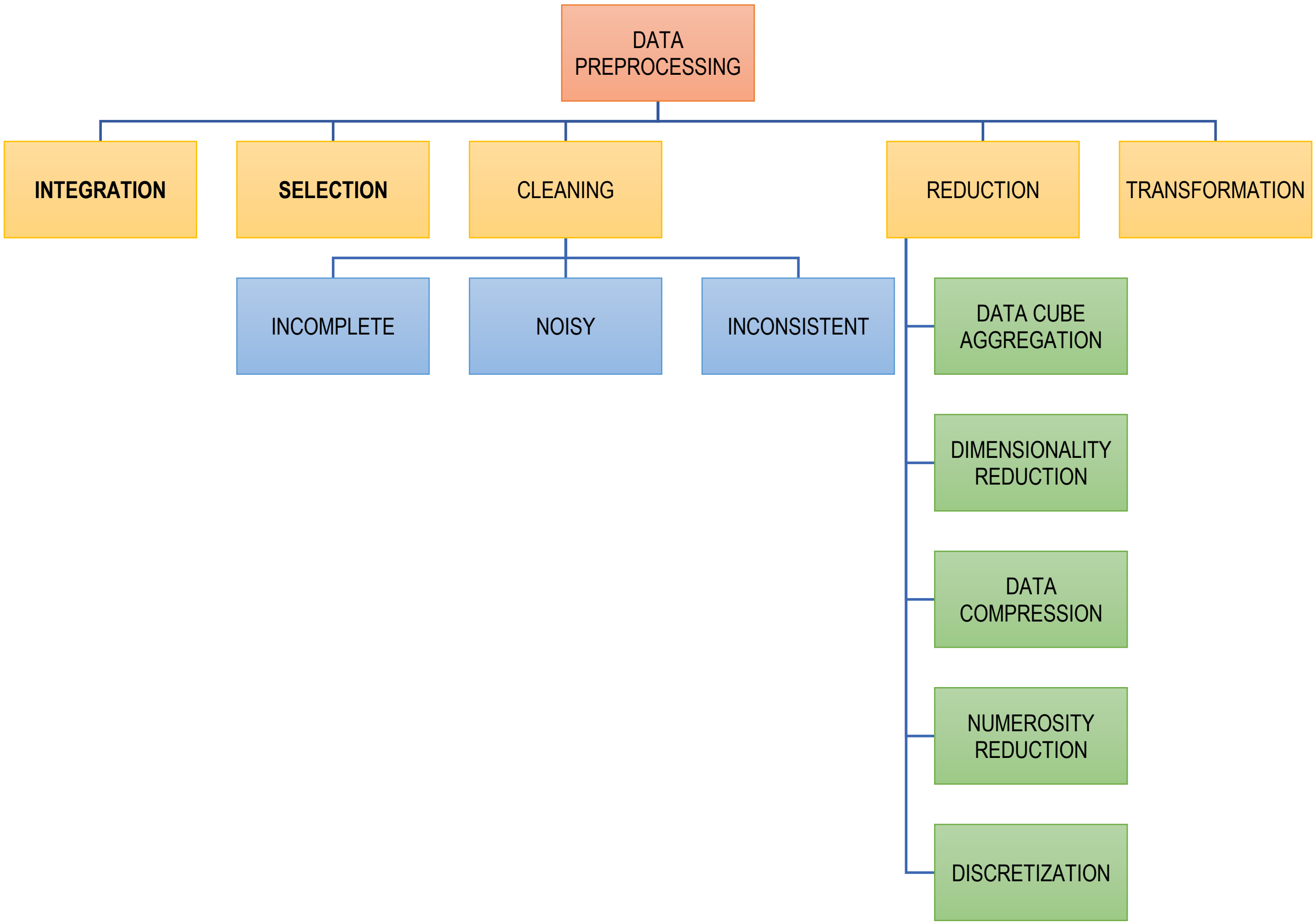
Advanced Methods in Data Mining

- Text data mining, text classification, naive Bayes, the EM algorithm, optimization, visualization, genetic algorithms, data augmentation, Markov-chain Monte Carlo techniques, knowledge extraction. Extensive use of data mining software.

Overview

- An important issue for data warehousing and data mining
- real world data tend to be incomplete, noisy and inconsistent
- includes
  - data cleaning
  - data integration
  - data transformation
  - data reduction

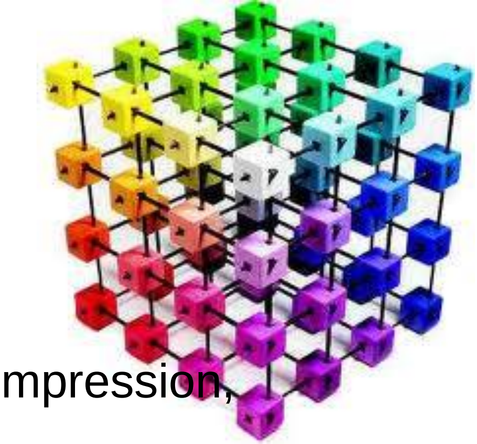
Data PREPROCESSING



Overview

- Data integration
  - combines data from multiple sources to form a coherent data store.
  - Metadata, correlation analysis, data conflict detection and resolution of semantic heterogeneity contribute towards smooth data integration.
- Data cleaning
  - fill in missing values
  - smooth noisy data
  - identify outliers
  - correct data inconsistency

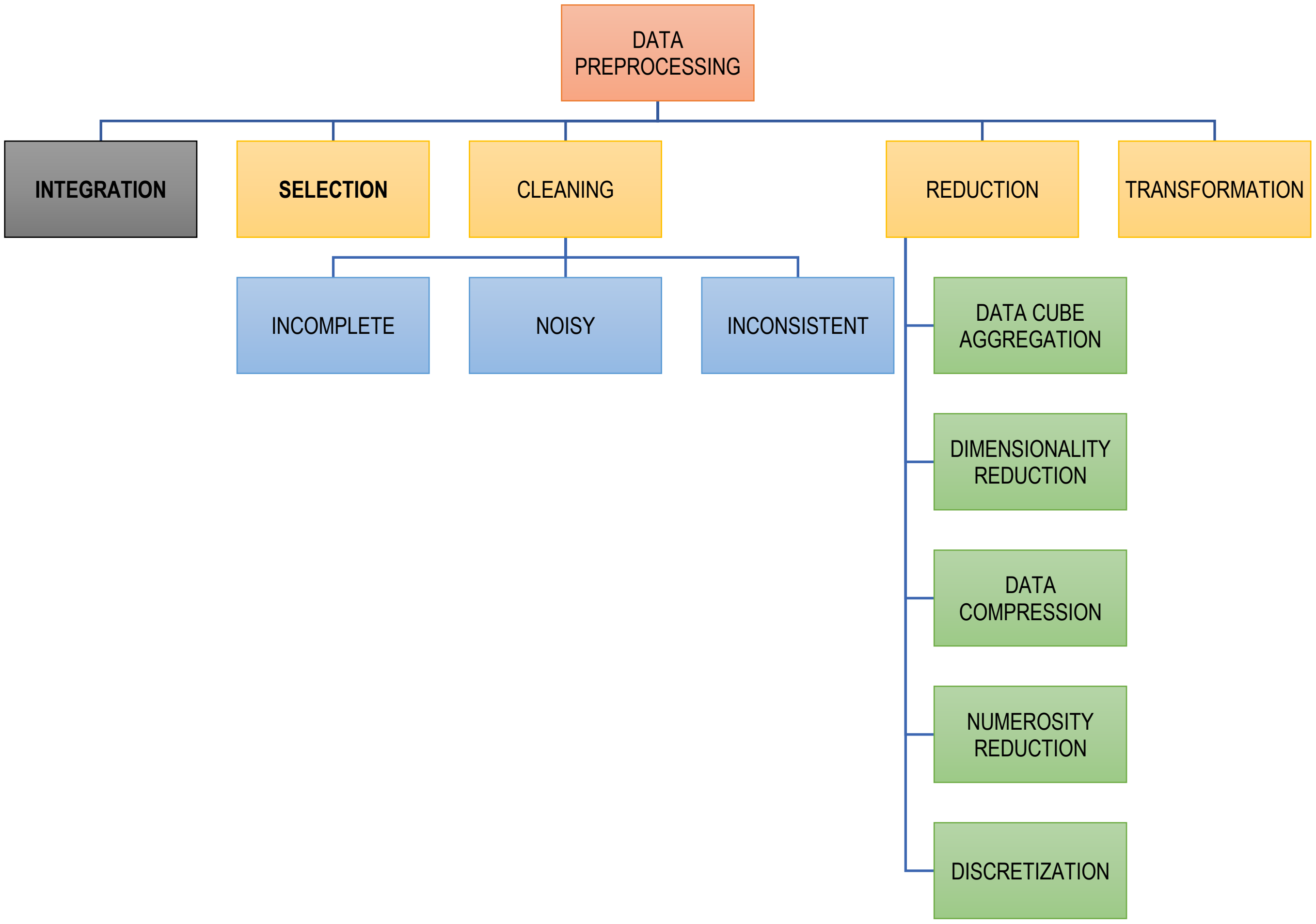
Overview



- Data reduction
  - data cube aggregation, dimension reduction, data compression, numerosity reduction and discretization.
  - Used to obtain a reduced representation of the data while minimizing the loss of information content.

- Data transformation
  - convert the data into appropriate forms for mining.
  - E.g. attribute data maybe normalized to fall between a small range such as 0.0 to 1.0

Data PREPROCESSING



Data Integration

- Data integration
  - combines data from multiple sources into a coherent data store e.g. data warehouse
  - sources may include multiple database, data cubes or flat files
  - Issues in data integration
    - schema integration
    - redundancy
    - detection and resolution of data value conflicts

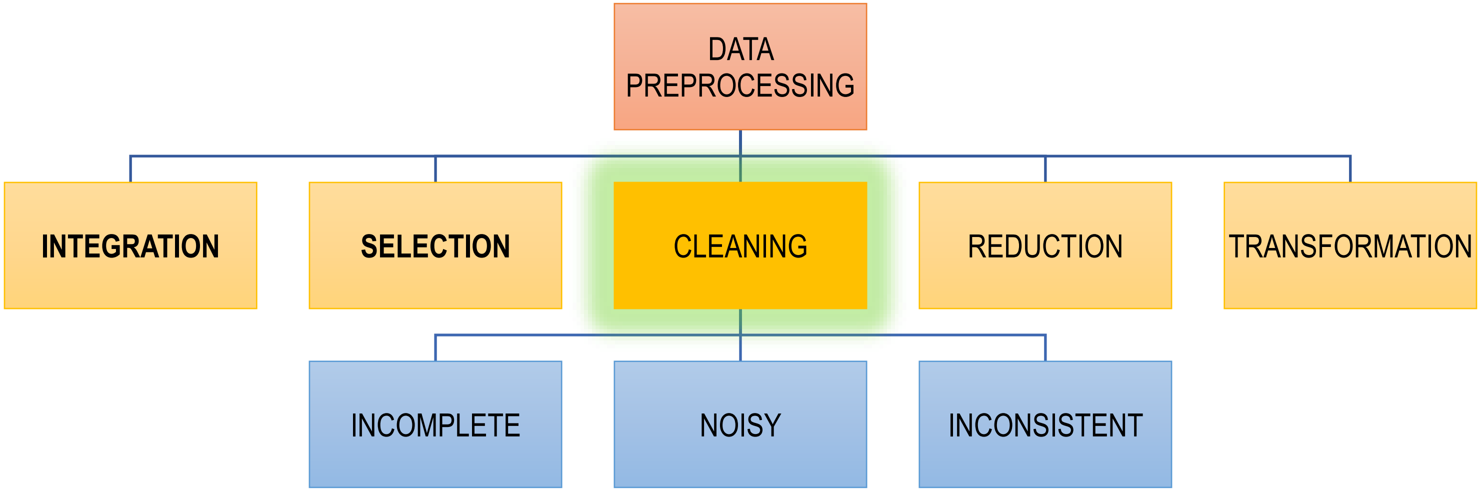
Data Integration

- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units

Data Integration

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlation analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data PREPROCESSING



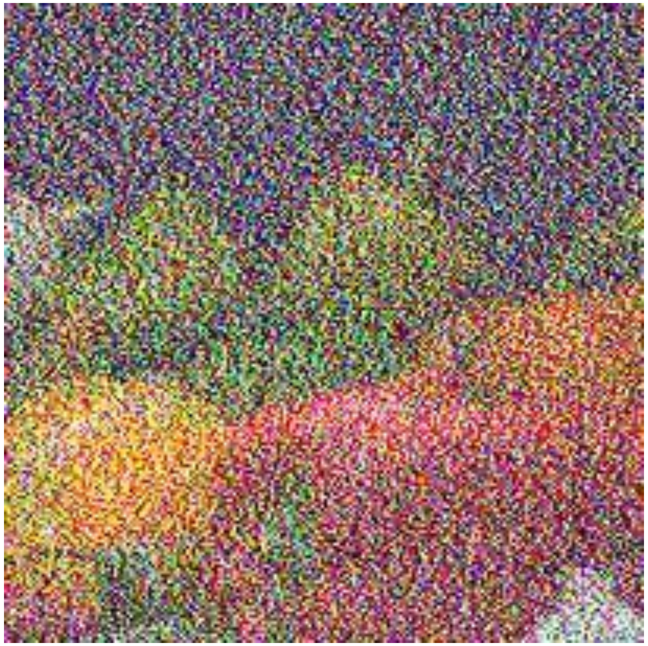
Data Cleaning : Missing Values

- Method of filling the missing values
  - Ignore the tuple
  - Fill in the missing value manually
  - Use a global constant
  - Use the attribute mean
  - Use the attribute mean for all samples belonging to the same class
  - Use the most probable value

Data Cleaning: Noisy Data

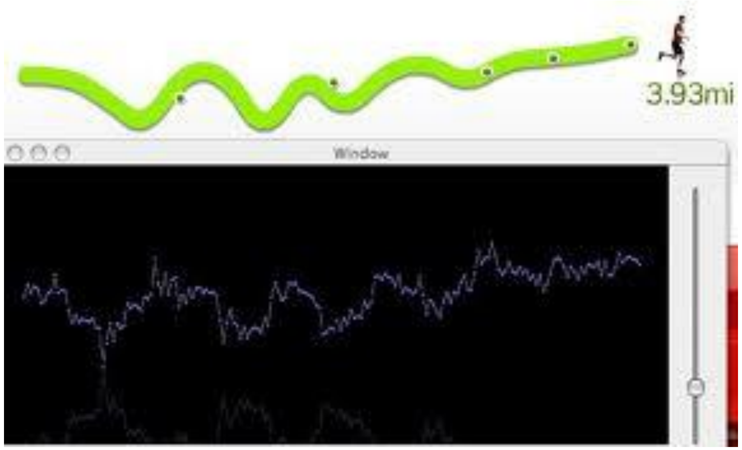
- Noise - random error or variance in a measured variable

- smooth out the data to remove the noise



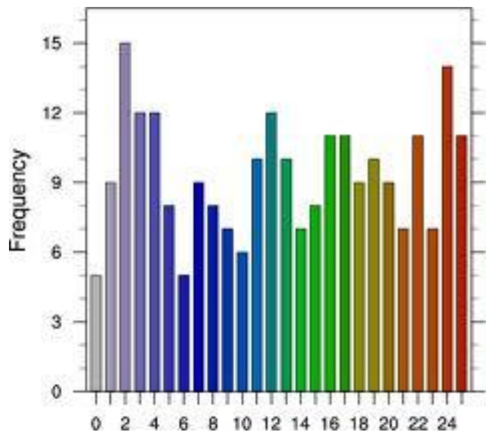
Data Cleaning: Noisy Data

- Data Smoothing Techniques
- Binning
  - smooth a sorted data value by consulting its neighborhood
  - the sorted values are distributed into a number of buckets or bins
    - smoothing by bin means
    - smoothing by bin medians
    - smoothing by bin boundaries



Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well.

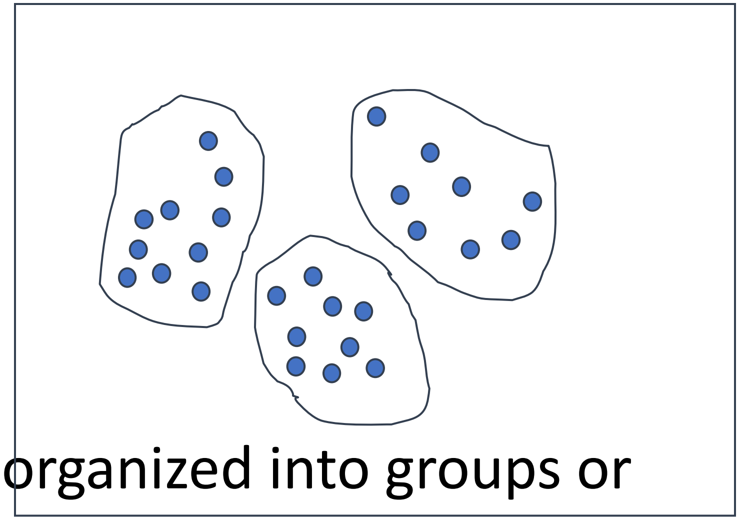


- Equal-depth (frequency) partitioning:
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

Binning Methods for Data Smoothing

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

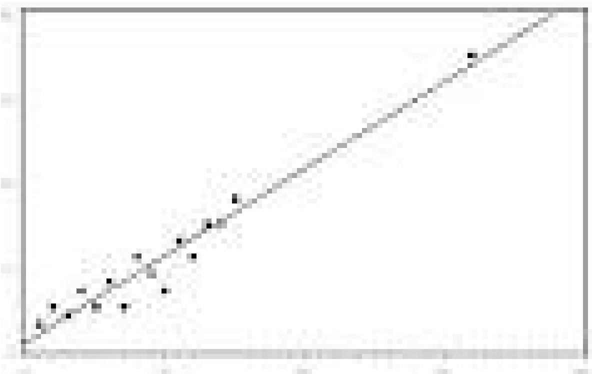
Handling outliers



- Clustering

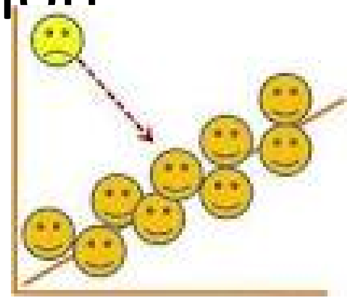
Outliers may be detected by

clustering, where similar values are organized into groups or clusters.



- Regression

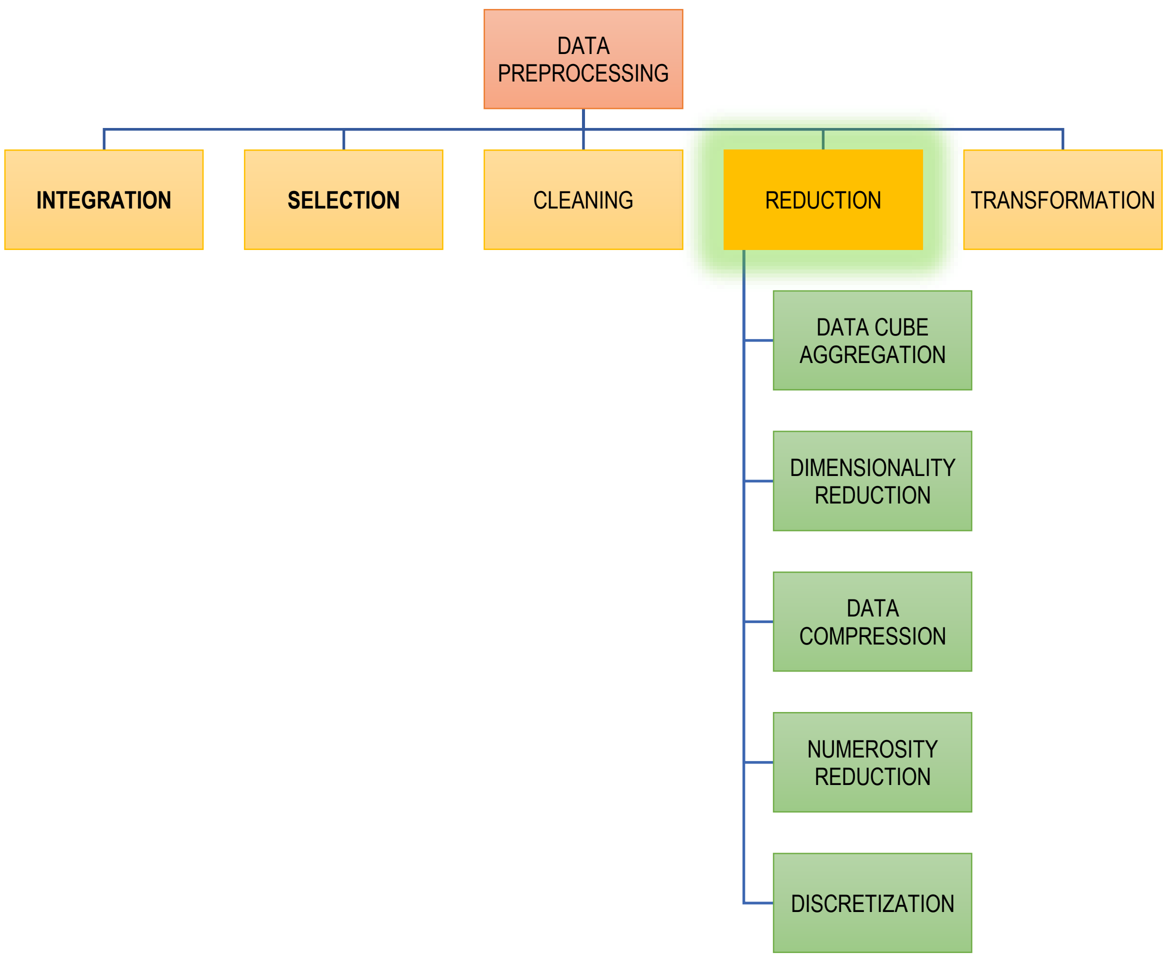
- Combined computer and human inspection



Data Cleaning : Inconsistent Data

- Can be corrected manually using external references
- Source of inconsistency
  - error made at data entry, can be corrected using paper trace

Data PREPROCESSING



Data Reduction

- To obtain a reduced representation of the data set that is
  - much smaller in volume
  - but closely maintains the integrity of the original data
  - mining on the reduced dataset should be more efficient yet produce the same analytical results.

Data Cube Aggregation

- The lowest level of a data cube
  - the aggregated data for an individual entity of interest
  - e.g., a customer in a phone calling data warehouse.
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

Data Cube Aggregation

Sales data for company *AllElectronics* for 1997 - 1999

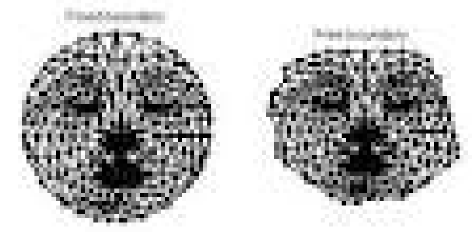
Year = 1999

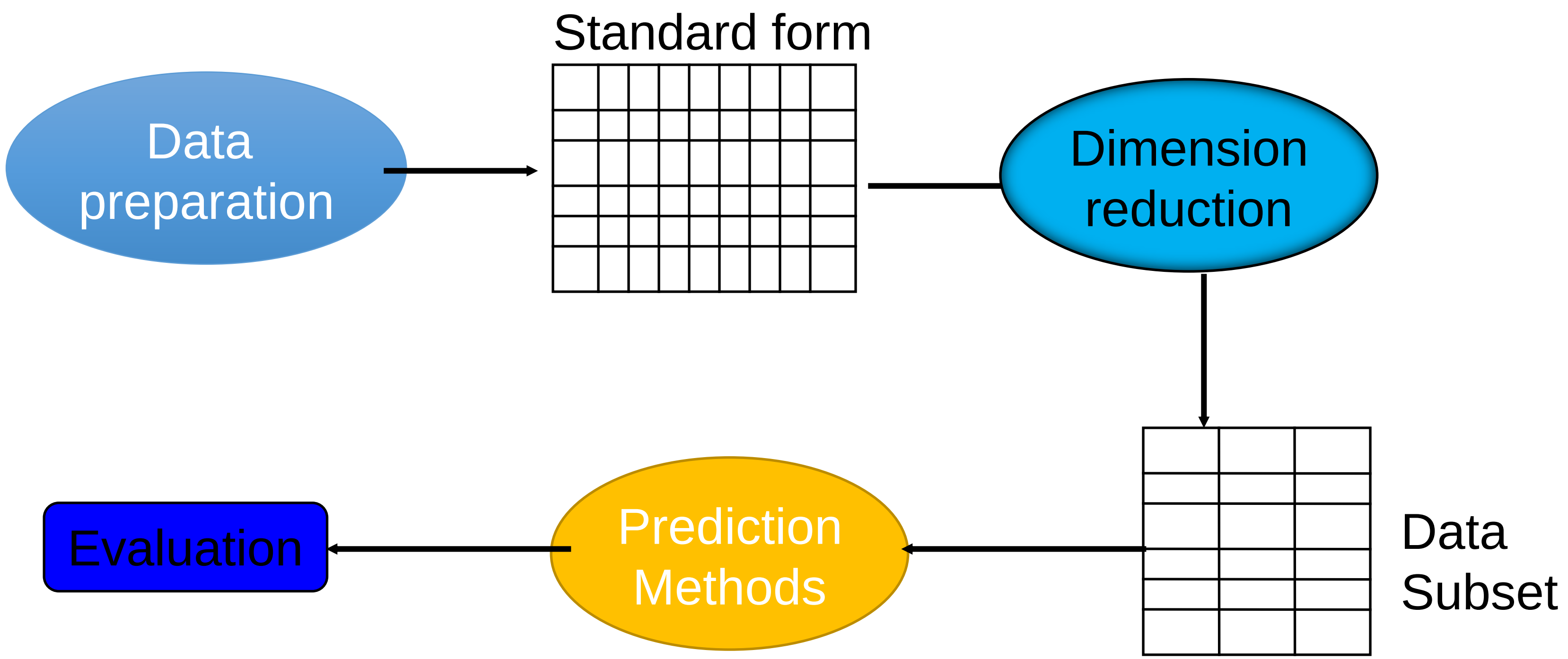
Year = 1998

Year = 1997	
Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	**Sales**
1997	$1,568,000
1998	$2,356,000
1999	$3,594,000

Dimensionality Reduction

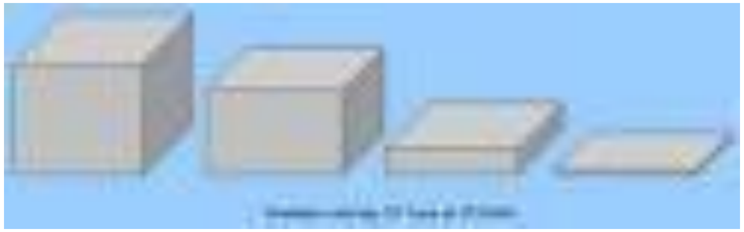




The role of dimension reduction in Data Mining

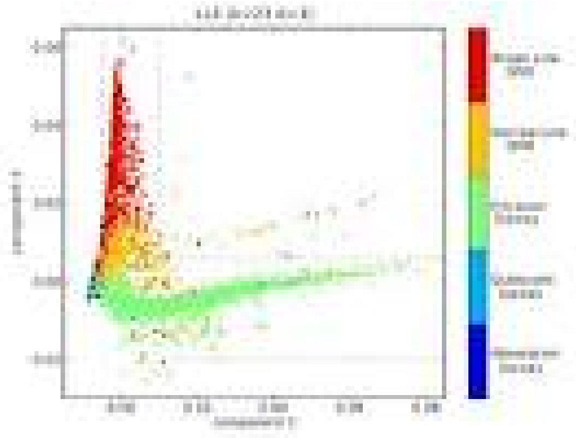
Dimensionality Reduction

- Data sets for analysis may contain hundreds of attributes that may be irrelevant to the mining task or redundant
- Dimensionality reduction reduces the dataset size by removing such attributes among them



Dimensionality Reduction

- How can we find a good subset of the original attributes??
- attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.



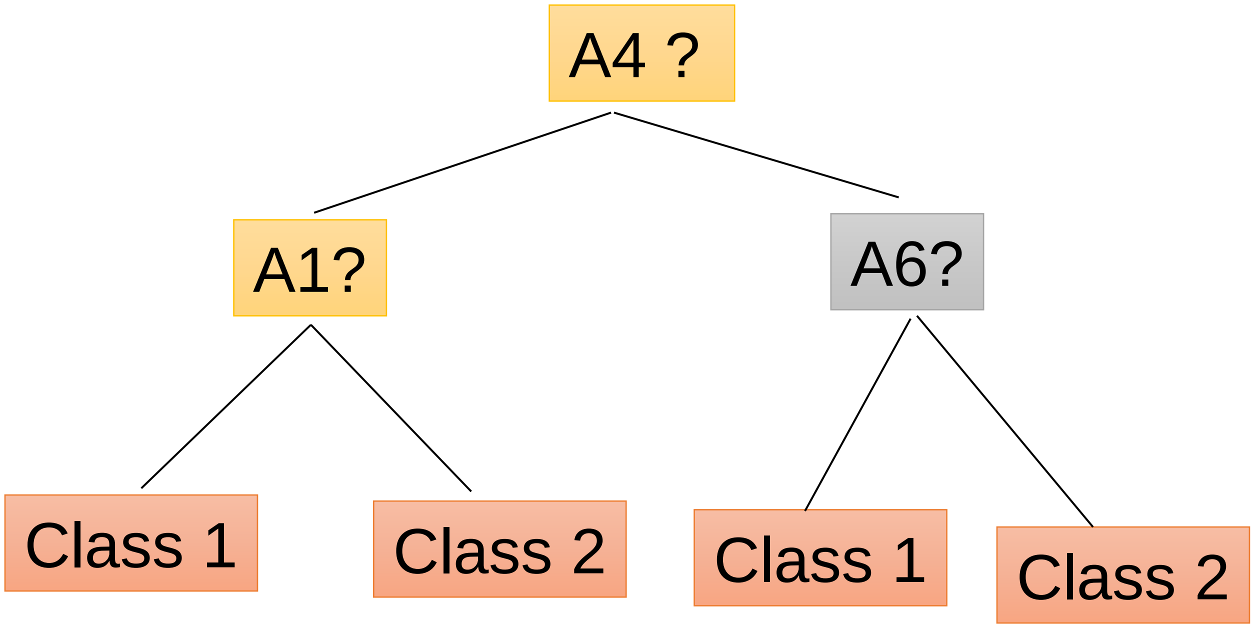
Dimensionality Reduction (Techniques)

- Attribute subset selection techniques
  - **Forward selection**
    - start with empty set of attributes,
    - the best of the original attributes is determined and added to the set.
    - At each subsequent iteration or step, the best of the remaining original attributes is added to the set.
  - **Stepwise backward elimination**
    - starts with the full set of attributes
    - At each step, it removes the worst attribute remaining in the set.
  - **Combination of forward selection and backward elimination**
    - the procedure combines and selects the best attribute and removes the worst from among the remaining attributes

Attribute subset selection techniques

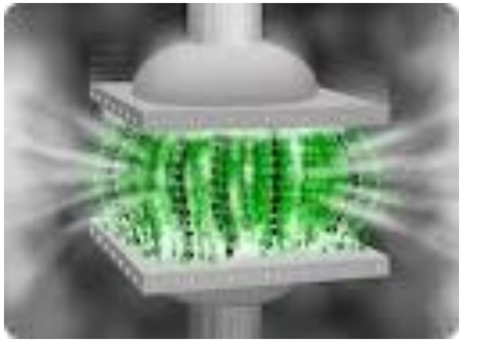
Decision tree induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



Reduced attribute set: {A1, A4, A6}

Data Compression



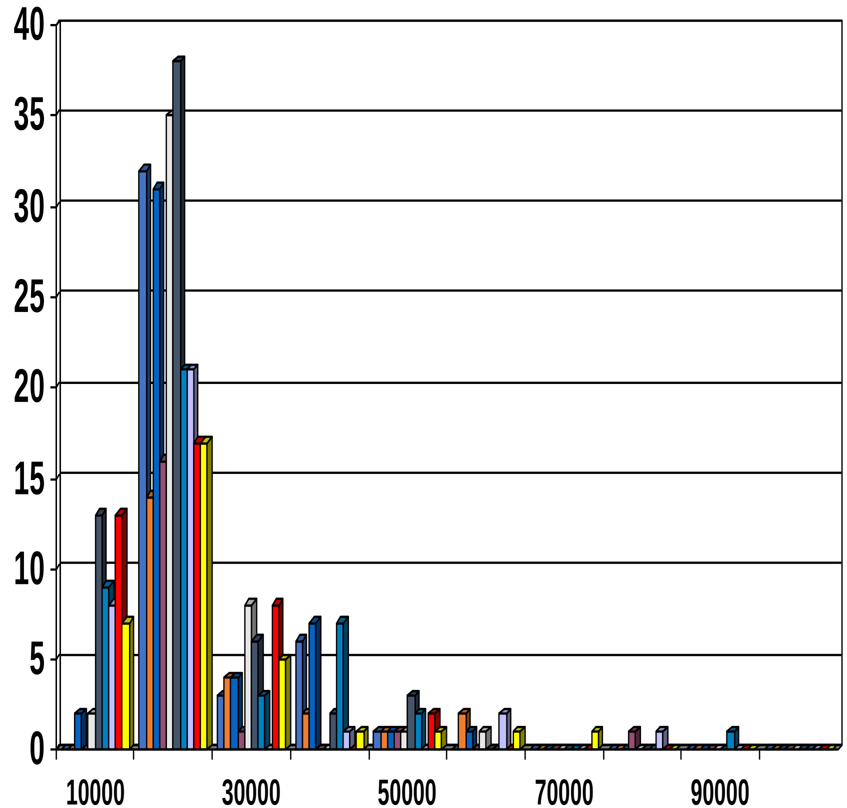
- Apply data encoding or transformation to obtain a reduced or compressed representation of the original data
- lossless
  - although typically lossless, they allow only limited manipulation of data.
- Two methods of lossy data compression
  - Wavelet Transforms
  - Principle Component Analysis

Numerosity Reduction

- Numerosity reduction technique can be applied to reduce the data volume by choosing alternative, smaller forms of data representation
- techniques
  - Regression and Log-Linear Models
  - Histograms
  - Clustering
  - Sampling

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems. (pg 126)

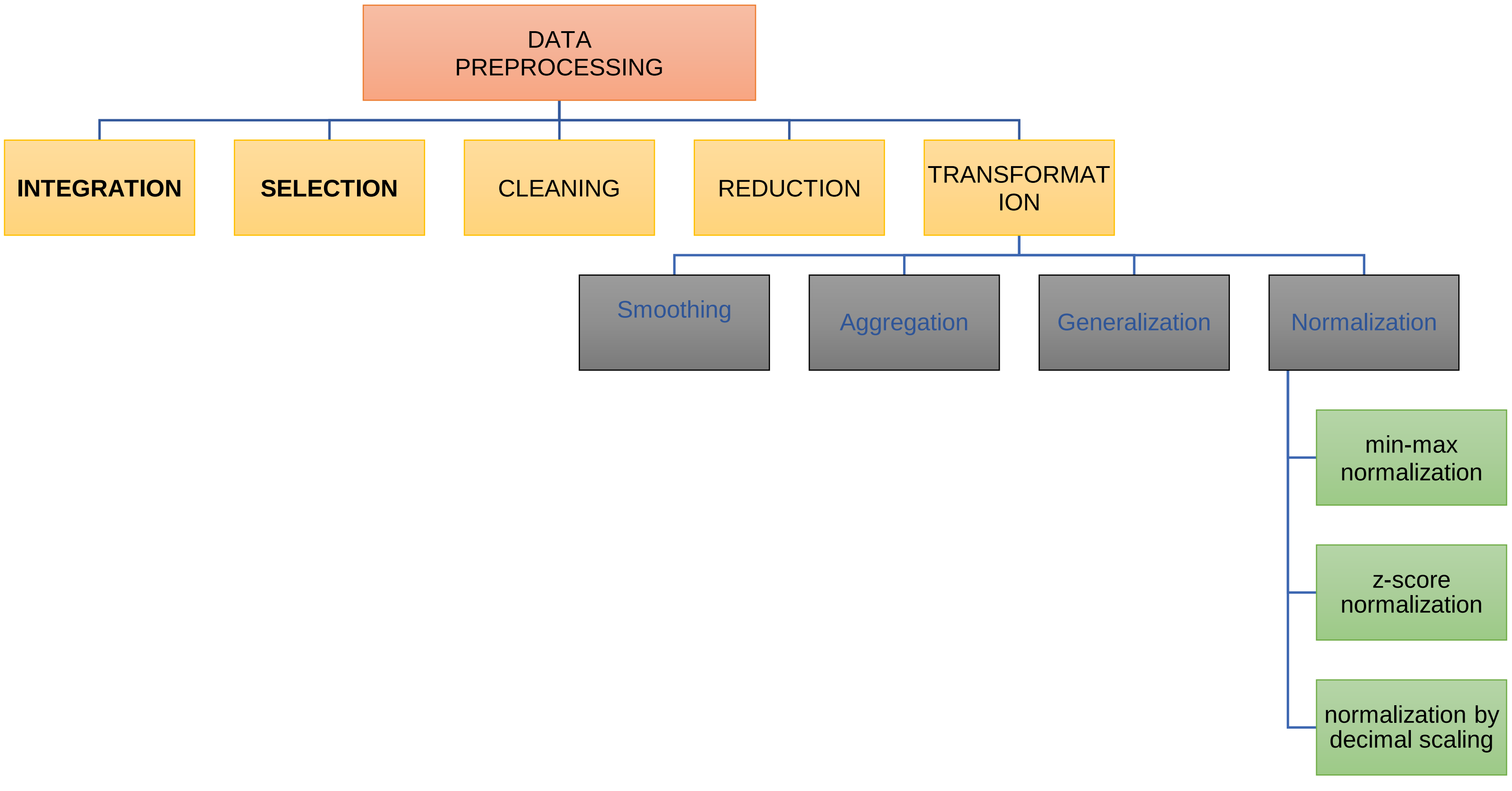


Discretization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers

- Discretization:
  - divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Data PREPROCESSING



Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|)<1

RAW DATA

Male	Typical angina	145	233	TRUE	LV hypertrophy	150	No	2.3	Downsloping	0	Fixed defect	0	No
Male	Asymptomatic	160	286	FALSE	LV hypertrophy	108	Yes	1.5	Flat	3	Normal	2	Yes
Male	Asymptomatic	120	229	FALSE	LV hypertrophy	129	Yes	2.6	Flat	2	Reversable defect	1	Yes
Male	Non-anginal pain	130	250	FALSE	Normal	187	No	3.5	Downsloping	0	Normal	0	No
Female	Atypical angina	130	204	FALSE	LV hypertrophy	172	No	1.4	Upsloping	0	Normal	0	No
Male	Atypical angina	120	236	FALSE	Normal	178	No	0.8	Upsloping	0	Normal	0	No
Female	Asymptomatic	140	268	FALSE	LV hypertrophy	160	No	3.6	Downsloping	2	Normal	3	Yes
Female	Asymptomatic	120	354	FALSE	Normal	163	Yes	0.6	Upsloping	0	Normal	0	No
Male	Asymptomatic	130	254	FALSE	LV hypertrophy	147	No	1.4	Flat	1	Reversable defect	2	Yes
Male	Asymptomatic	140	203	TRUE	LV hypertrophy	155	Yes	3.1	Downsloping	0	Reversable defect	1	Yes
Male	Asymptomatic	140	192	FALSE	Normal	148	No	0.4	Flat	0	Fixed defect	0	No
Female	Atypical angina	140	294	FALSE	LV hypertrophy	153	No	1.3	Flat	0	Normal	0	No
Male	Non-anginal pain	130	256	TRUE	LV hypertrophy	142	Yes	0.6	Flat	1	Fixed defect	2	Yes
Male	Atypical angina	120	263	FALSE	Normal	173	No	0	Upsloping	0	Reversable defect	0	No
Male	Non-anginal pain	172	199	TRUE	Normal	162	No	0.5	Upsloping	0	Reversable defect	0	No
Male	Non-anginal pain	150	168	FALSE	Normal	174	No	1.6	Upsloping	0	Normal	0	No
Male	Atypical angina	110	229	FALSE	Normal	168	No	1	Downsloping	0	Reversable defect	1	Yes
Male	Asymptomatic	140	239	FALSE	Normal	160	No	1.2	Upsloping	0	Normal	0	No
Female	Non-anginal pain	130	275	FALSE	Normal	139	No	0.2	Upsloping	0	Normal	0	No
Male	Atypical angina	130	266	FALSE	Normal	171	No	0.6	Upsloping	0	Normal	0	No
Male	Typical angina	110	211	FALSE	LV hypertrophy	144	Yes	1.8	Flat	0	Normal	0	No
Female	Typical angina	150	283	TRUE	LV hypertrophy	162	No	1	Upsloping	0	Normal	0	No
Male	Atypical angina	120	284	FALSE	LV hypertrophy	160	No	1.8	Flat	0	Normal	1	Yes
Male	Non-anginal pain	132	224	FALSE	LV hypertrophy	173	No	3.2	Upsloping	2	Reversable defect	3	Yes
Male	Asymptomatic	130	206	FALSE	LV hypertrophy	132	Yes	2.4	Flat	2	Reversable defect	4	Yes

CLEANED DATA

Male	Typical angina	145	233	TRUE	LV hypertrophy	150	No	2.3	Downsloping	0	Fixed defect	0	No
Male	Asymptomatic	160	286	FALSE	LV hypertrophy	108	Yes	1.5	Flat	3	Normal	2	Yes
Male	Asymptomatic	120	229	FALSE	LV hypertrophy	129	Yes	2.6	Flat	2	Reversable defect	1	Yes
Male	Non-anginal pain	130	250	FALSE	Normal	187	No	3.5	Downsloping	0	Normal	0	No
Female	Atypical angina	130	204	FALSE	LV hypertrophy	172	No	1.4	Upsloping	0	Normal	0	No
Male	Atypical angina	120	236	FALSE	Normal	178	No	0.8	Upsloping	0	Normal	0	No
Female	Asymptomatic	140	268	FALSE	LV hypertrophy	160	No	3.6	Downsloping	2	Normal	3	Yes
Female	Asymptomatic	120	354	FALSE	Normal	163	Yes	0.6	Upsloping	0	Normal	0	No
Male	Asymptomatic	130	254	FALSE	LV hypertrophy	147	No	1.4	Flat	1	Reversable defect	2	Yes
Male	Asymptomatic	140	203	TRUE	LV hypertrophy	155	Yes	3.1	Downsloping	0	Reversable defect	1	Yes
Male	Asymptomatic	140	192	FALSE	Normal	148	No	0.4	Flat	0	Fixed defect	0	No
Female	Atypical angina	140	294	FALSE	LV hypertrophy	153	No	1.3	Flat	0	Normal	0	No
Male	Non-anginal pain	130	256	TRUE	LV hypertrophy	142	Yes	0.6	Flat	1	Fixed defect	2	Yes
Male	Atypical angina	120	263	FALSE	Normal	173	No	0	Upsloping	0	Reversable defect	0	No
Male	Non-anginal pain	172	199	TRUE	Normal	162	No	0.5	Upsloping	0	Reversable defect	0	No
Male	Non-anginal pain	150	168	FALSE	Normal	174	No	1.6	Upsloping	0	Normal	0	No
Male	Atypical angina	110	229	FALSE	Normal	168	No	1	Downsloping	0	Reversable defect	1	Yes
Male	Asymptomatic	140	239	FALSE	Normal	160	No	1.2	Upsloping	0	Normal	0	No
Female	Non-anginal pain	130	275	FALSE	Normal	139	No	0.2	Upsloping	0	Normal	0	No
Male	Atypical angina	130	266	FALSE	Normal	171	No	0.6	Upsloping	0	Normal	0	No
Male	Typical angina	110	211	FALSE	LV hypertrophy	144	Yes	1.8	Flat	0	Normal	0	No
Female	Typical angina	150	283	TRUE	LV hypertrophy	162	No	1	Upsloping	0	Normal	0	No
Male	Atypical angina	120	284	FALSE	LV hypertrophy	160	No	1.8	Flat	0	Normal	1	Yes
Male	Non-anginal pain	132	224	FALSE	LV hypertrophy	173	No	3.2	Upsloping	2	Reversable defect	3	Yes
Male	Asymptomatic	130	206	FALSE	LV hypertrophy	132	Yes	2.4	Flat	2	Reversable defect	4	Yes

Summary

- Data preparation is a big issue for both warehousing and mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot methods have been developed but still an active area of research