Data Mining
Metoder och tekniker

Mathias Lanner Sas Institute

THE POWER TO KNOW.

What is data mining

What is data mining?

”Data mining uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational databases-patterns that ordinary methods might miss.”

-Two Crows Corporation (1998),p.1

”Data Mining, as we use the term, is the exploration and analysis by automatic or semiautomatic means, of large quantities of data in order to discover meaningsful patterns and rules.”

-Berry and Linoff(1997), p.5

”Data Mining [is] the process of efficient discovery of nonobvious valuble information from large collection of data.”

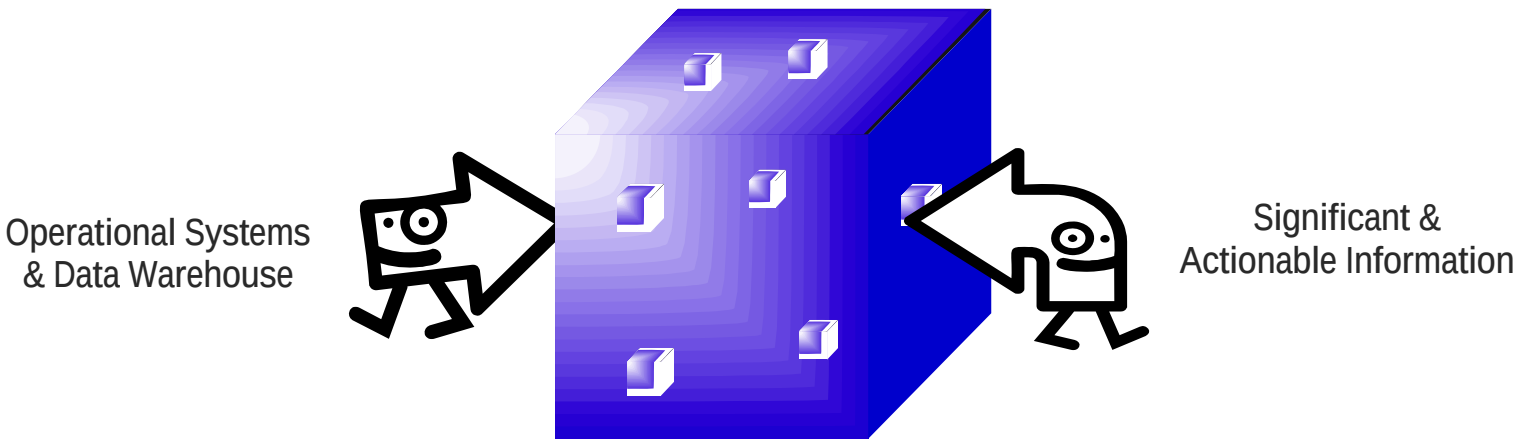
-Berson and Smith (1997), p.565

”Data Mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognation technologies as well as statistical and mathematical techniques.”

-Erick Brethnoux, Gartner Group

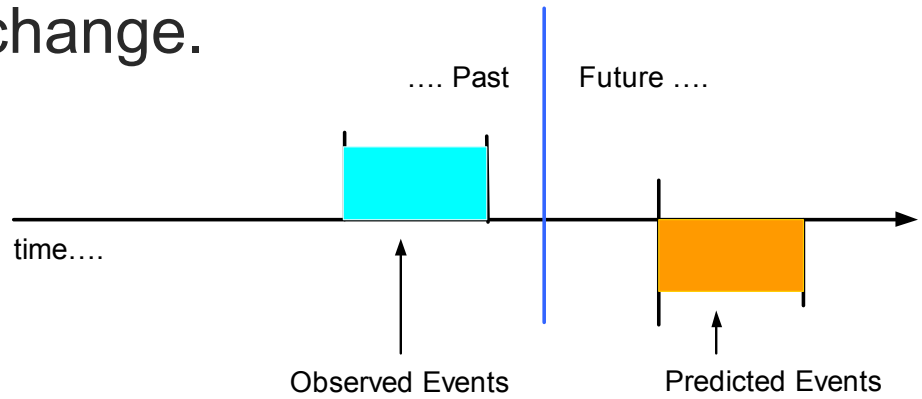
Data Mining Definition :

The *process of selecting, exploring, and modeling* large amounts of data to uncover previously unknown information for a *business advantage*

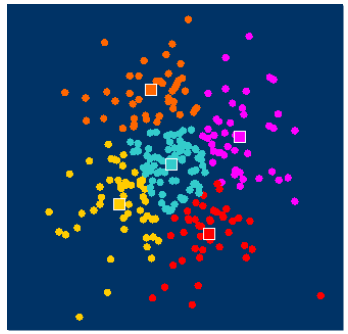


Data Mining Is:

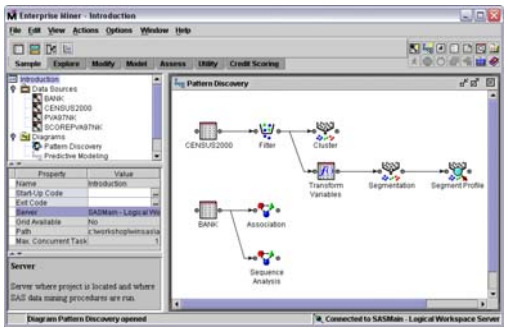
- Discovering patterns and relationships represented in data.
  - Developing models to understand and describe characteristics and activity based on these patterns.
- Using this understanding to help evaluate future options, gain insights and take decisions.
  - Deploy the results of the analysis to affect business change.

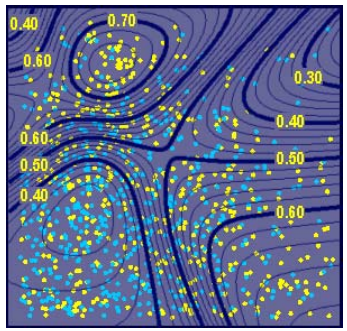


Two types of analysis



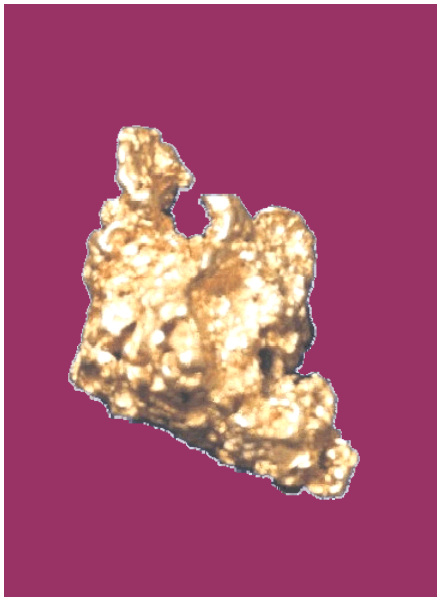
Pattern Discovery





Predictive Modeling

Pattern Discovery



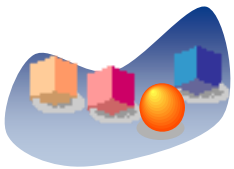
The Essence of Data Mining?

"…the discovery of interesting, unexpected, or valuable structures in large data sets."

– David Hand

Pattern Discovery Applications

Data reduction



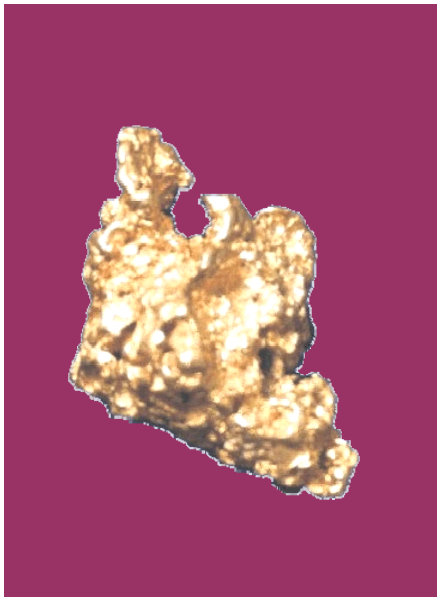
Novelty detection

Clustering

Market basket analysis

Sequence analysis

Predictive Modeling



The *Essence* of Data Mining

"Most of the big payoff [in data mining] has been in predictive modeling."

– Herb Edelstein

Data Mining
Delivering Value Across Industries

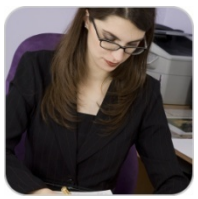
Retail	Identify the most profitable customers and the underlying reasons for their loyalty
Life Sciences	Find promising new molecular drug compounds
Manufacturing	Save on downtime by applying predictive maintenance to manufacturing sites
Financial Services	Grow customer profitability and reduce risk exposure through more accurate credit scoring
Insurance	Rate making - set more profitable rates for insurance premiums
Telco	Help retain customers (prevent churn) and identify up-sell/cross-sell opportunities for individual customers
Government	Detect and deter fraudulent behavior

Process -methodology

Successful Data Mining – Iterative & Interactive

Data Manager

- Data preparation
- Deployment services
- Report administration

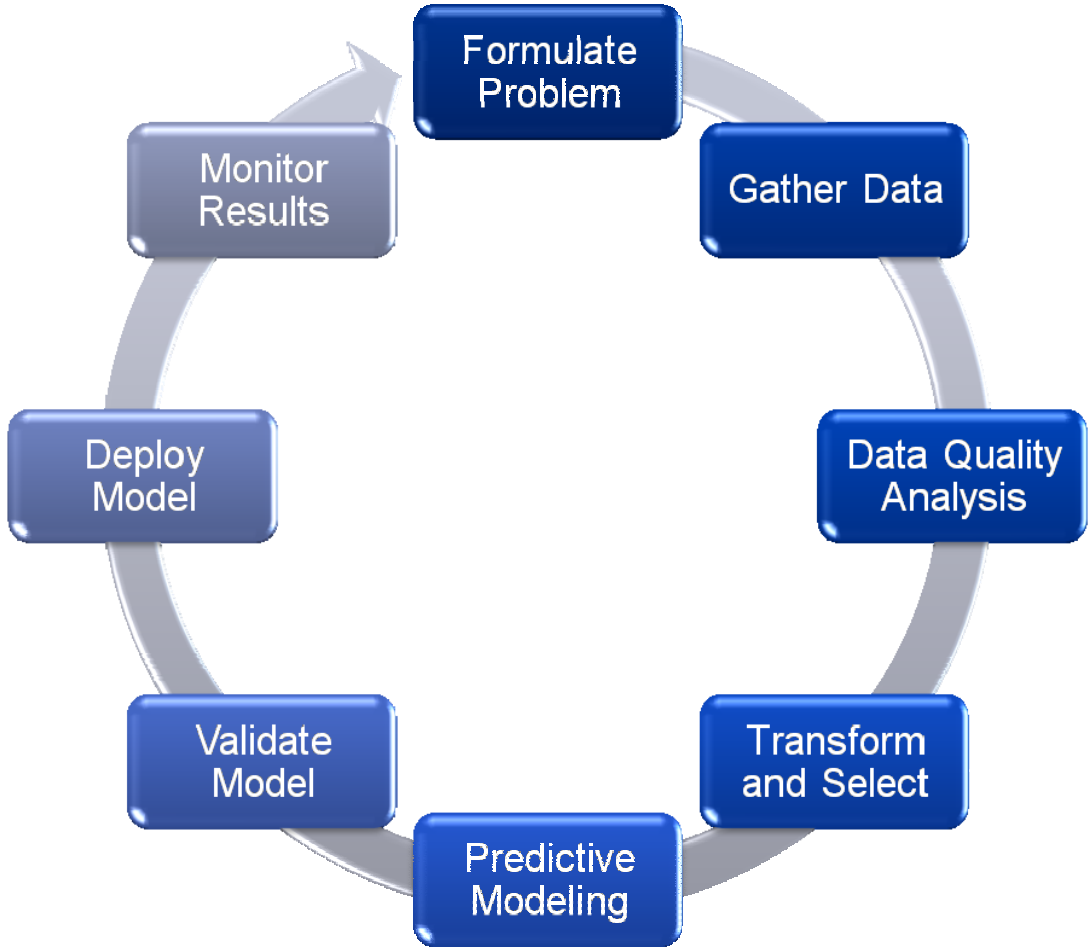


Data Miner

- Exploratory analysis
- Descriptive segmentation
- Predictive modeling

Business Manager

- Root Cause Analysis
- Domain expert
- Evaluates processes and ROI



Explore data and find relationships

SEMMA

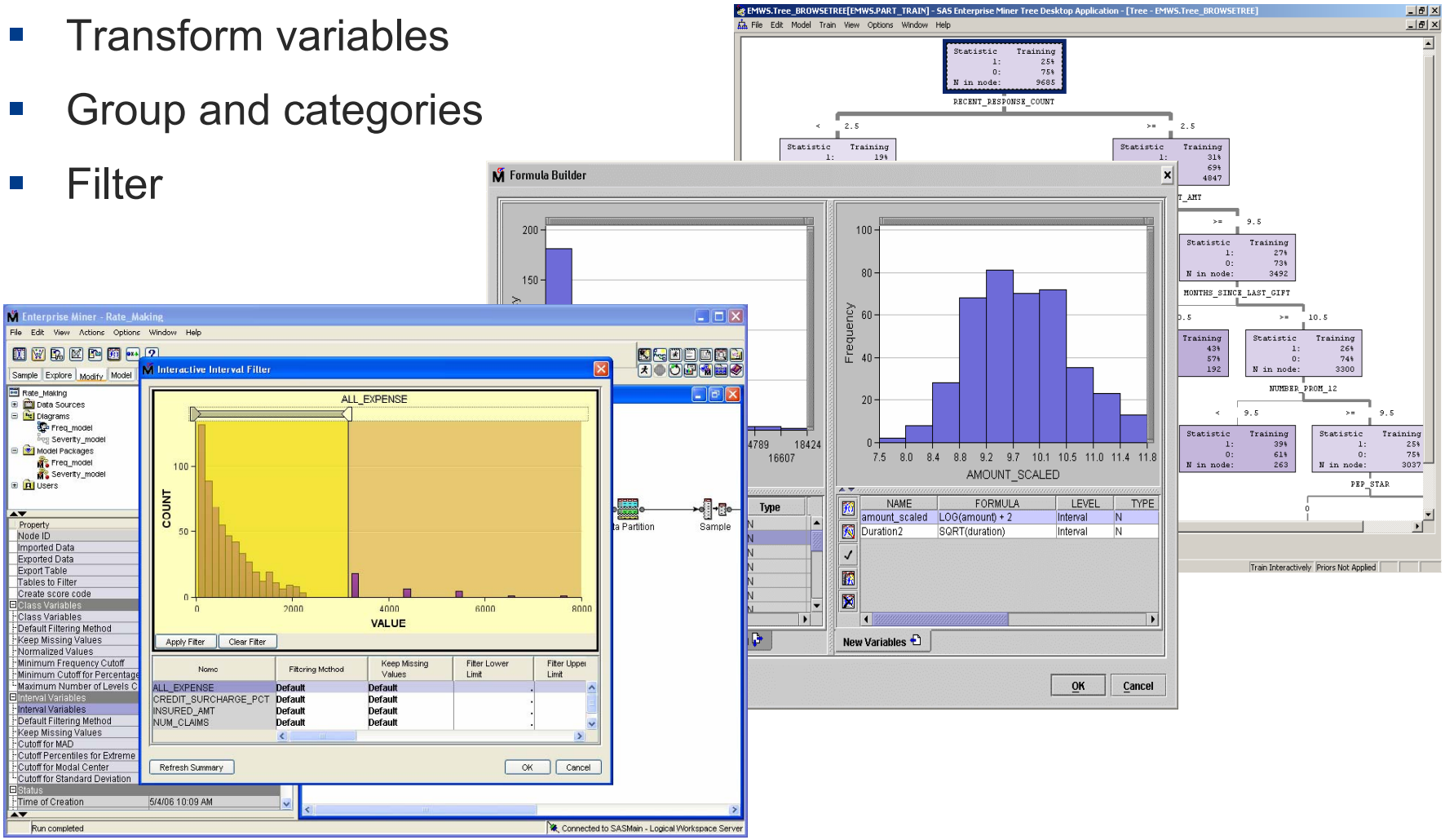
Descriptive statistics

Find patterns in the data

Modify variables/data

SEMMA

- Transform variables
- Group and categories
- Filter

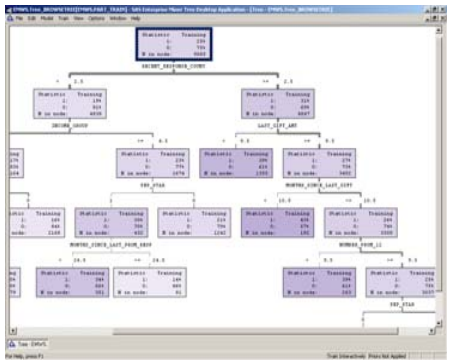


Modelling

SEMMA

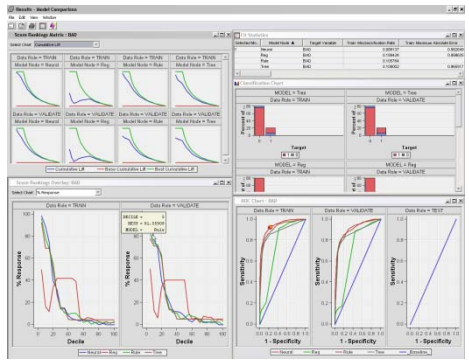
Prediktive modelling

Decision tree



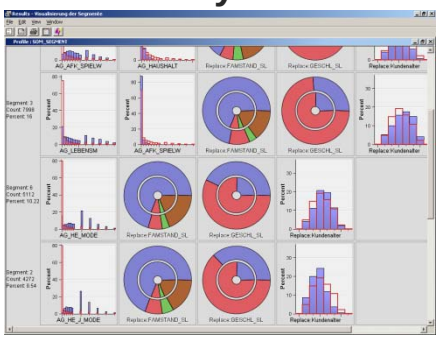
Regression

Neuralnetworks

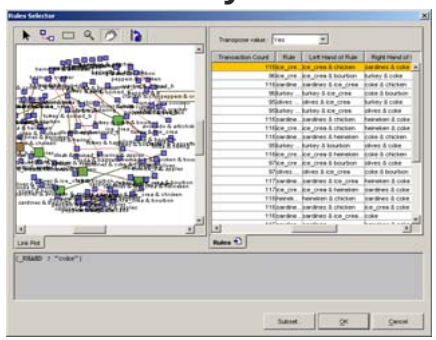


Exploring techniques

Cluster analysis



Basket analysis

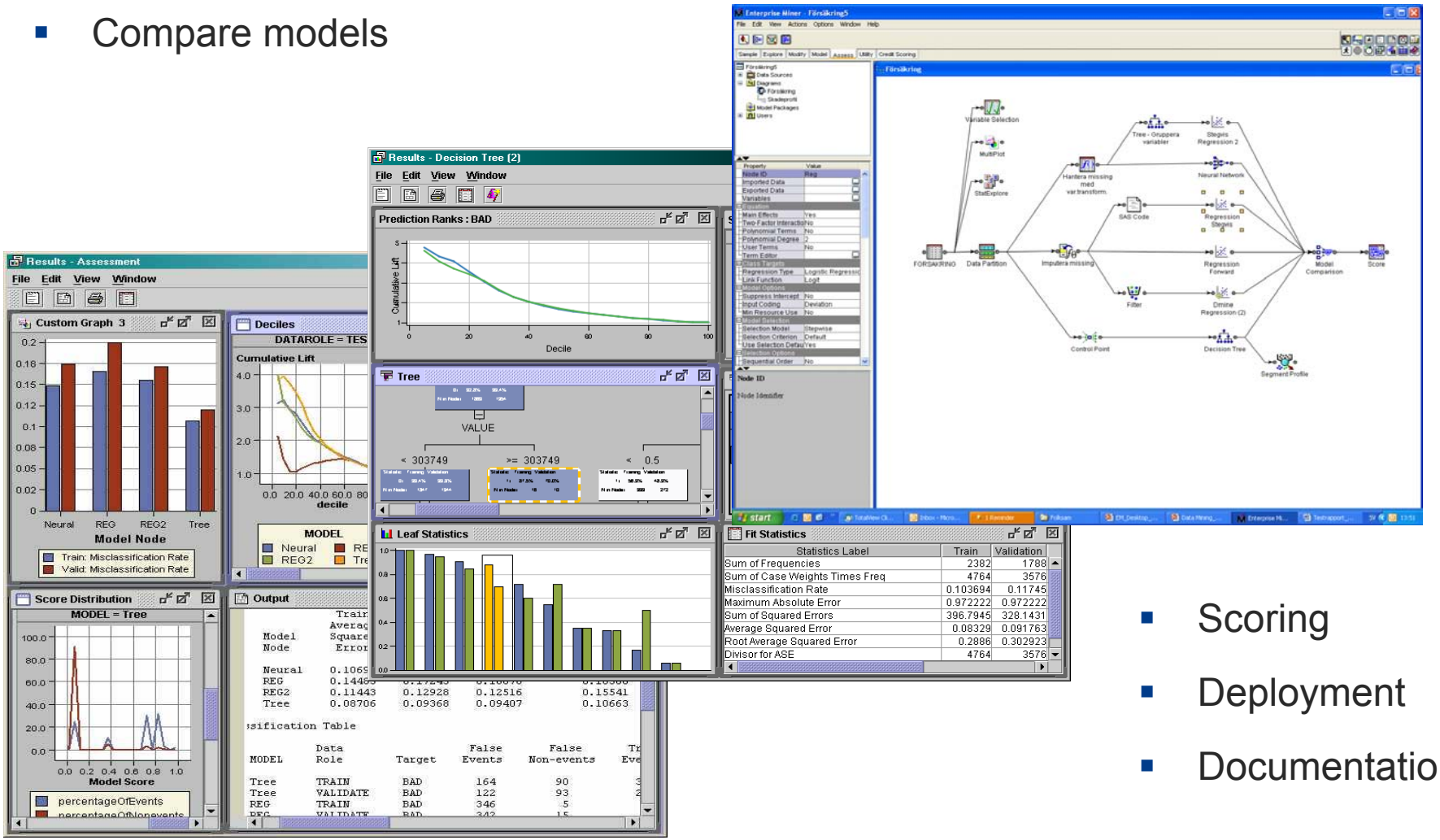


Univariate

Assess –

SEMMA

- Compare models



- Scoring
- Deployment
- Documentation

Statistics & Data mining

Statistics

- Experimental
- Prior Hypothesis
  - Idea before data acquisition
  - Data acquisition planned
- Experimental Design
  - Sampling strategies
  - Factorial designs
  - Required confidence
  - Minimize model terms
- Inference
  - Hypothesis testing
  - Inference

Data mining

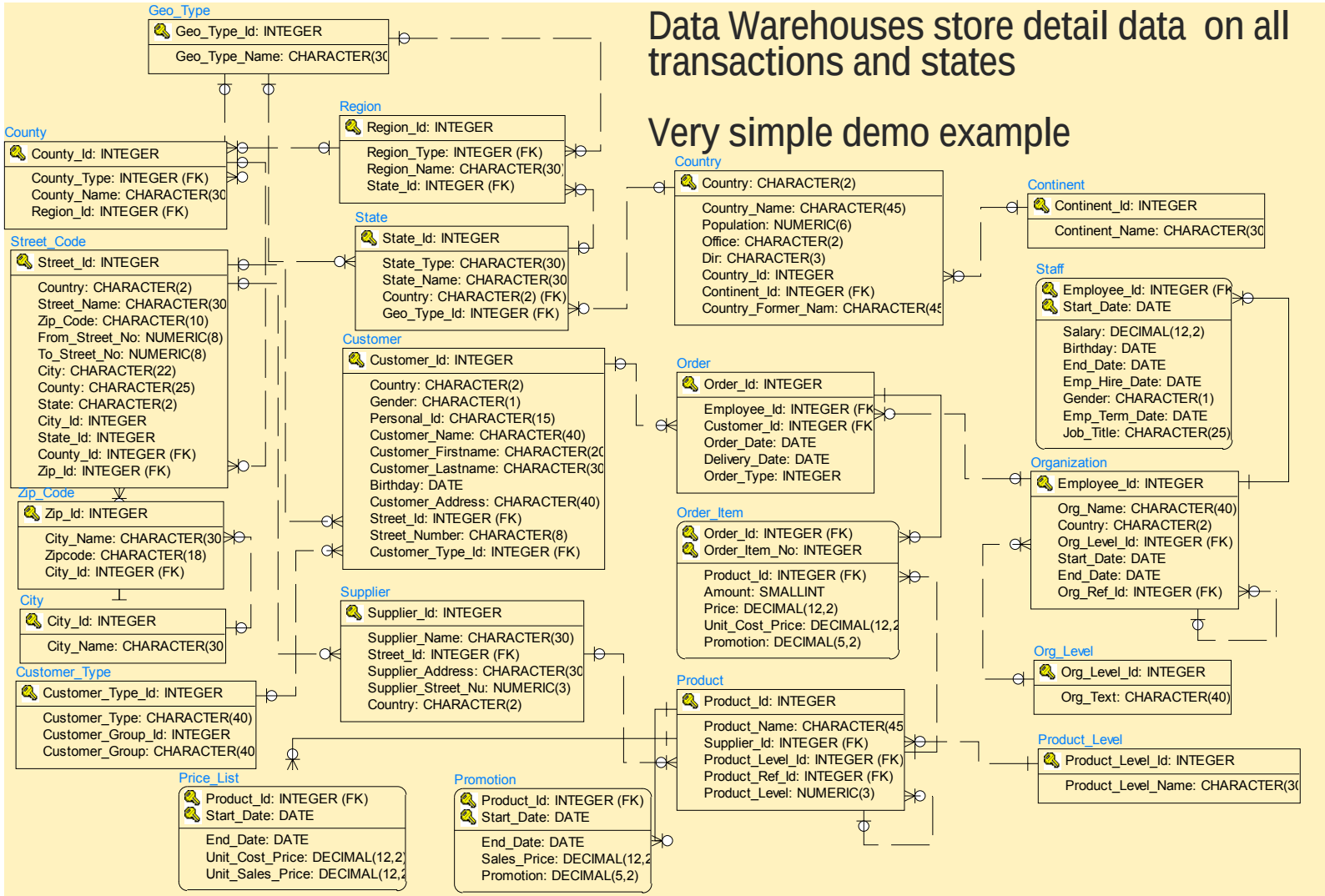
- Commercial
- Posterior Hypothesis
  - Idea after data acquisition
  - Data acquisition opportunistic
- No Experimental Design
  - Explore data
  - Create hypothesis
  - Generate query
  - Create models
- Prediction
  - Lift, Profit, Response
  - Prediction

Data Challenges

Data is collected differently

	Experimental	Opportunistic
Purpose	Research	Operational
Value	Scientific	Commercial
Generation	**Actively controlled**	**Passively observed**
Size	**Small?**	**Massive (large N and p)**
Hygiene	Clean	Dirty
State	Static	Dynamic

Where does mining data come from ?



Data is becoming wider and wider

- Used to work with a couple of dozens of variables
- Nowadays at least a couple of hundreds
  - Data from different sources
  - Derived data (differences, rations, trends etc.)
  - Data from combined algorithms (market basket analysis, combined with clustering combined with predictive modeling)
- Can become thousands
  - Pharma: micro-array data
  - Interactions

New data sources

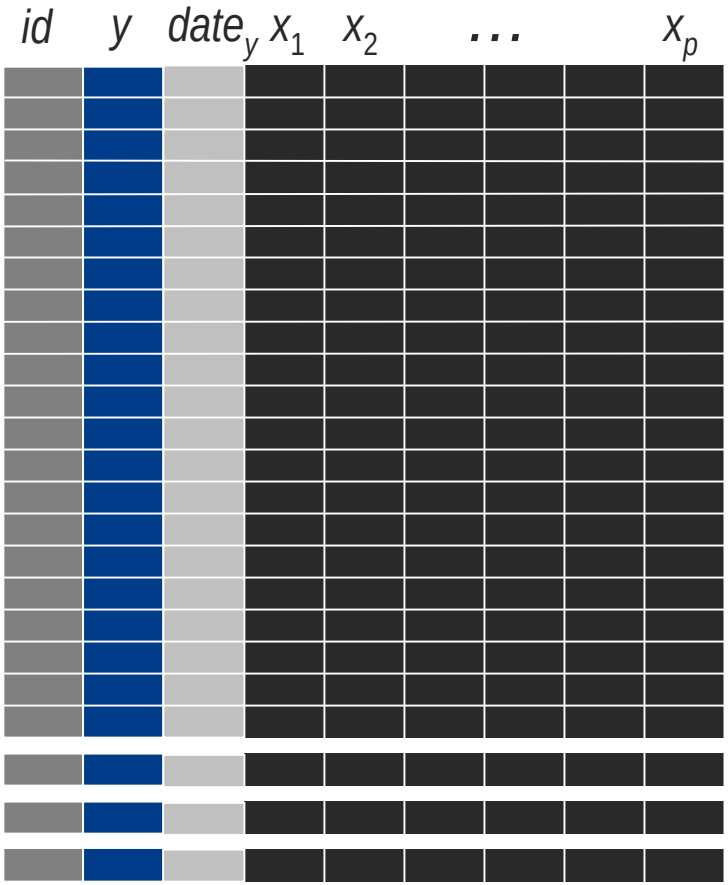
- Extreme commercial data warehouses
  - Many gigabytes of data
  - Stores may have 100,000+ SKU items
  - Sales histories for every item/basket saved
  - Rollups can produce terms >> 10,000 terms

- Digital data acquisition
  - Biometrics: microarray, mass spectrometry
  - Chip fabs: 30,000 measurements per manufacturing run.
  - ISP: every page, server, router, switch, at timepoints
    - University: 50-60 GB / day
    - Regional telecom: 6 TB / day

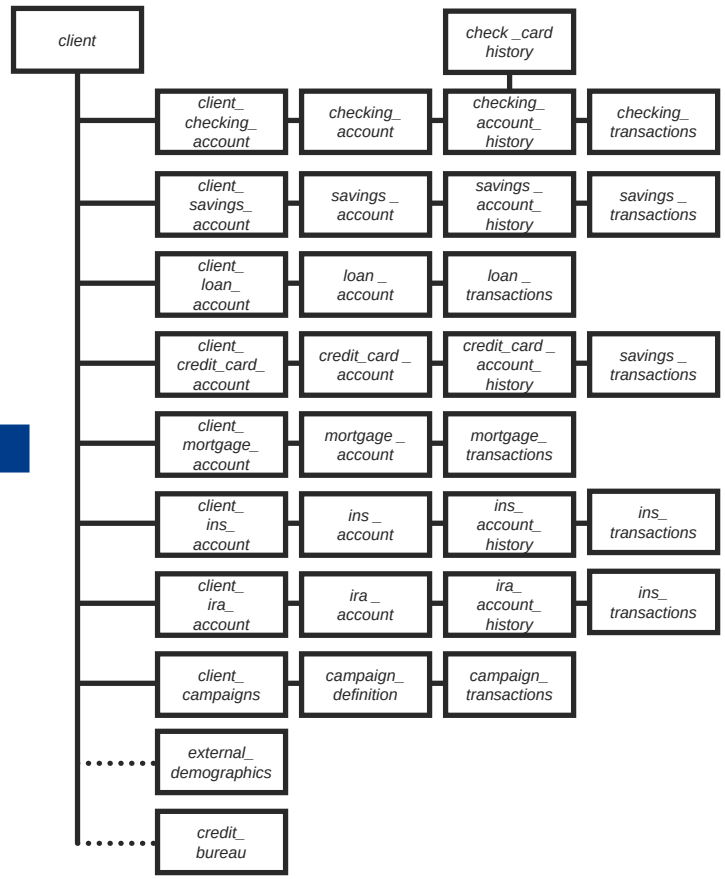
Integration

- Integrate data access and management
  - Prepare data for analytics in enterprise warehouse
    - Join tables
    - Clean data
    - Create derived variables (aggregations, ratios, trends etc.)
    - Create samples
    - Create data mining metadata (targets, inputs, rejected)

Predictive Model Development Data



Model Development Data Set



Marketing Data Warehouse

Data Quality

- Data quality
  - Data mining requires detail data
  - New level of data quality is necessary
  - Lot of time spent for data cleaning
  - Use the warehouse to correct the errors

- META Group:

  “10 to 20 percent of the raw data used is corrupt or incomplete in some way. It is not unusual to discover that as many as half the records in a database contain some type of information that needs to be corrected”
  META Group Program Director John Ladley

Data Quality

- **Intelligent methods to deal with missing values**
  - Use robust estimators for distribution
  - Predict missing values from remaining information with trees
  - Track the replaced values – add degrees of freedom for missingness
  - Use clustering for replacing missing values
  - Use algorithms that can deal with missing values automatically

Techniques

THE POWER TO KNOW.

Definition: What is a model?

- **A function assigning an outcome**
  - Prediction, classification, cluster, rule,…
  - input vars, output vars
  - SCORECODE
- **Attributes of the function**
  - Type, formula, objective, target
  - Attribute importance, ranking, p-value
  - Model quality measures: stats, gains, distributions, centroids
- **History of the function**
  - Who, what, when, how
  - Auditing, regulation, best practices
- **Purpose of the function**
  - Cross-sell, acquisition, approval, collection, churn, fraud,…
- **Status of the model**
  - Development, production, retired, champion, challenger

Algorithms

- A model is an abstraction of the data and belongs with the data
- There is nothing more in a model than what is already in the data

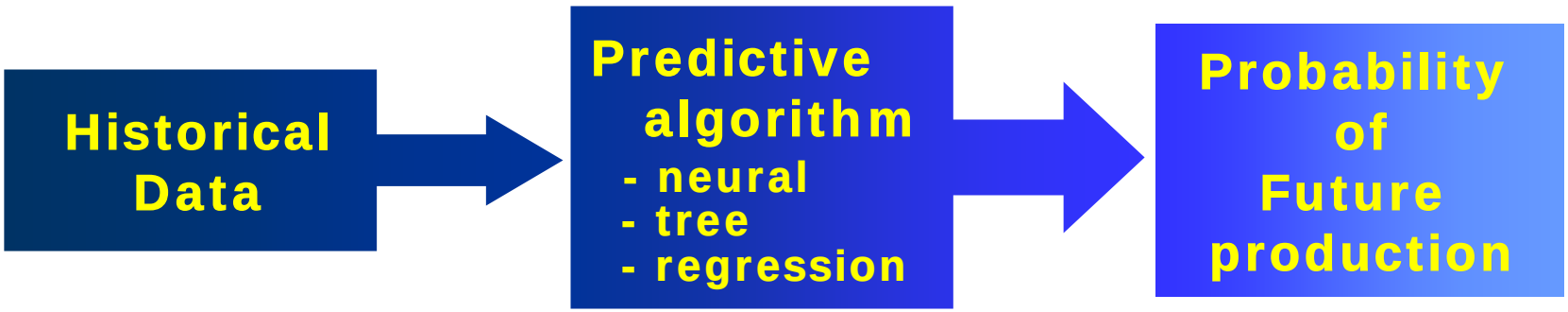
Algorithms

- There is no BEST algorithm per se
- Depends on
  - Nature of relationships in data
  - Data quality
  - Time available to build a model
  - Nature of model deployment
    - operational use
    - insights for business users
    - decision support etc.

Data Mining Algorithms

① predictive (supervised)

use data on past processes to *predict* future production

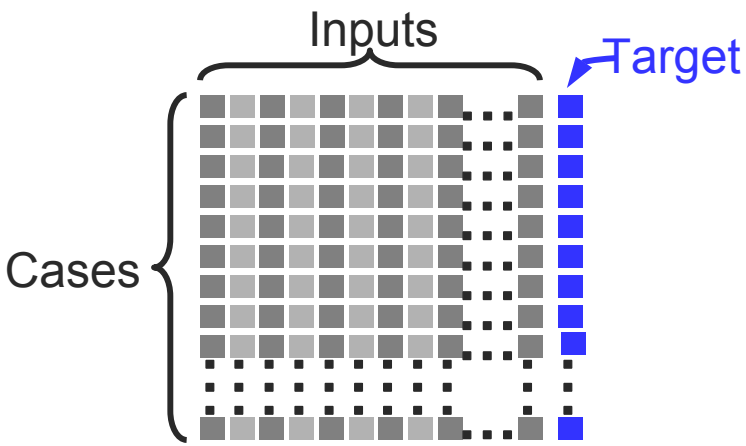


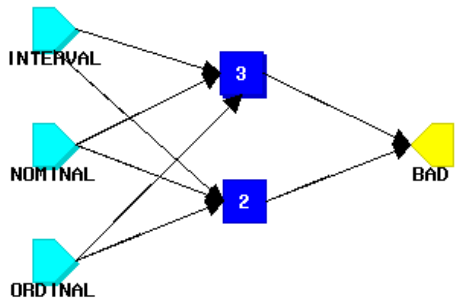
② descriptive (unsupervised)

use data on past processes to *describe* current situation

Supervised Learning

Tries to find good rules for predicting the value of a target(s) from the values of the inputs variables.





Enterprise Miner

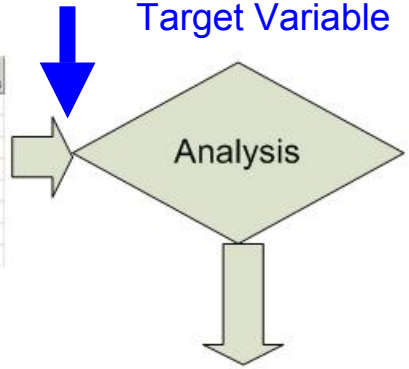
- Logistic and OLS
- Tree Classifiers
- Neural Networks
- Ensembles
- Memory Based Reasoning
- Two-stage modeling
- Fast Variable Selection
- Principal Components
- PLS Regression
- Support Vector Machine
- Gradient Bosting
- SAS/STAT

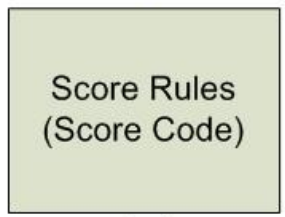
Data Mining
Model Training and Scoring

Analysis Datamart

	Identification#	Vehicle Use	Value of Vehicle	Type of Car	Age	Claim Happens
1	10005854	Private	9860	Pickup	42	Yes
2	10020811	Commerci	30460	Panel Truck	58	Yes
3	10023726	Private	16580	SUV	45	No
4	10081891	Private	20030	SUV	43	No
5	10081891	Private	24360	Sedan	43	No
6	10083752	Private	3770	Sedan	27	Yes
7	10101436	Private	11560	SUV	51	No
8	10120916	Commerci	13950	SUV	38	No

Target Variable





Scoring Datamart

	Identification#	Vehicle Use	Value of Vehicle	Type of Car	Age
619	20240150	Commerci	7050	SUV	48
620	20246469	Private	12020	Pickup	51
621	20251776	Private	10120	SUV	38
622	20254273	Commerci	16920	Van	39
623	20264511	Private	37130	Pickup	54
624	20267443	Commerci	28460	Panel Truck	51
625	20276305	Commerci	20990	Van	40
626	20289114	Commerci	9340	Pickup	49
627	20312611	Private	8960	SUV	60
628	20339617	Private	14970	Sedan	53
629	2034339	Private	26340	SUV	37

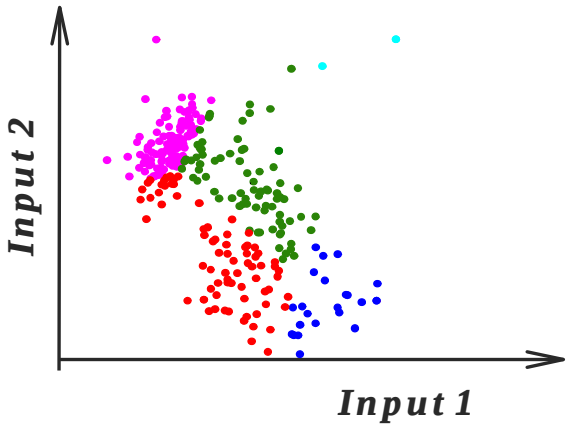
Prediction

Datamart with Scores

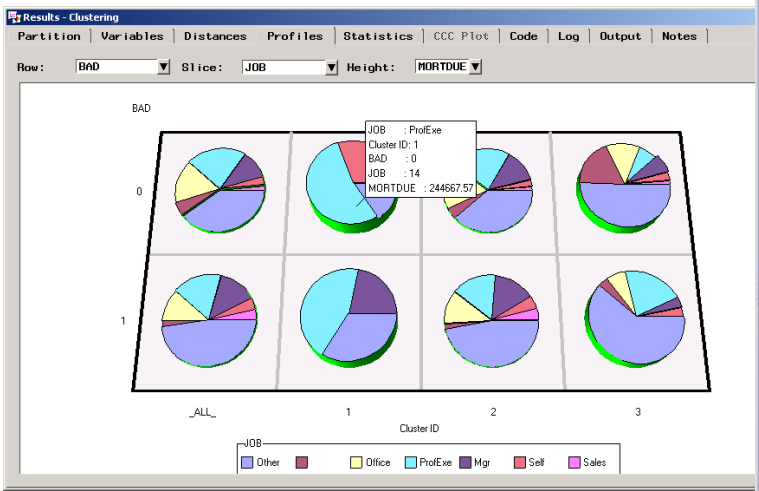
	Identification#	Vehicle Use	Value of Vehicle	Type of Car	Age	Claim_Prob
619	20240150	Commerci	7050	SUV	48	0.0412
620	20246469	Private	12020	Pickup	51	0.2186
621	20251776	Private	10120	SUV	38	0.8300
622	20254273	Commerci	16920	Van	39	0.3449
623	20264511	Private	37130	Pickup	54	0.0568
624	20267443	Commerci	28460	Panel Truck	51	0.2325
625	20276305	Commerci	20990	Van	40	0.0530
626	20289114	Commerci	9340	Pickup	49	0.5778
627	20312611	Private	8960	SUV	60	0.8300
628	20339617	Private	14970	Sedan	53	0.1315
629	2034339	Private	26340	SUV	37	0.2017

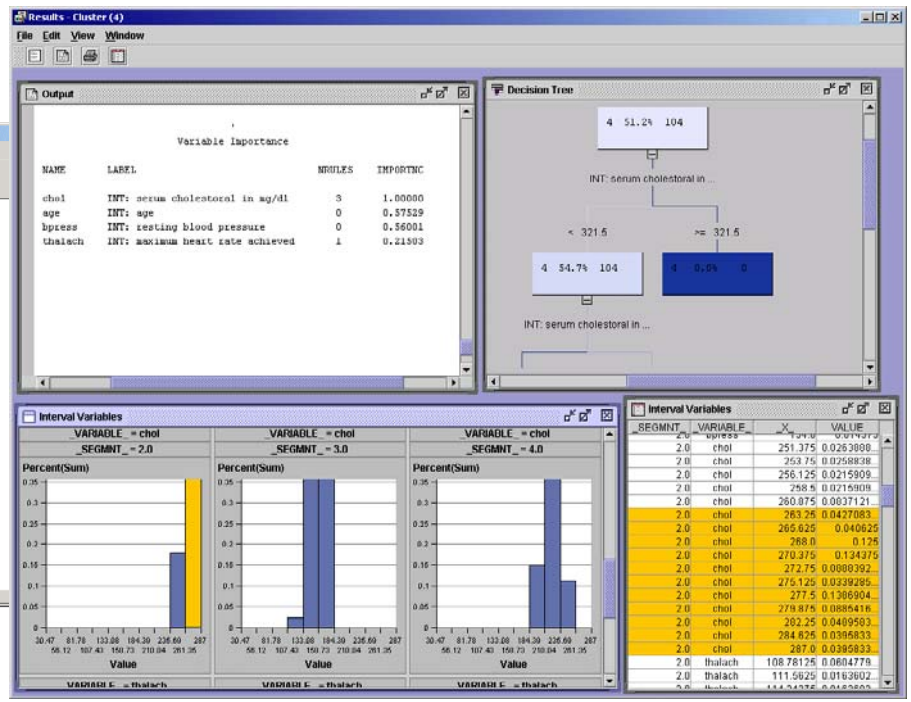
Unsupervised Learning

Tries to divide the data into groups such that the observations within a group have traits more similar than those assigned to different groups Enterprise Miner



- k-Means
- SOM/Kohonen Networks
- Rule Builder
- SAS/STAT





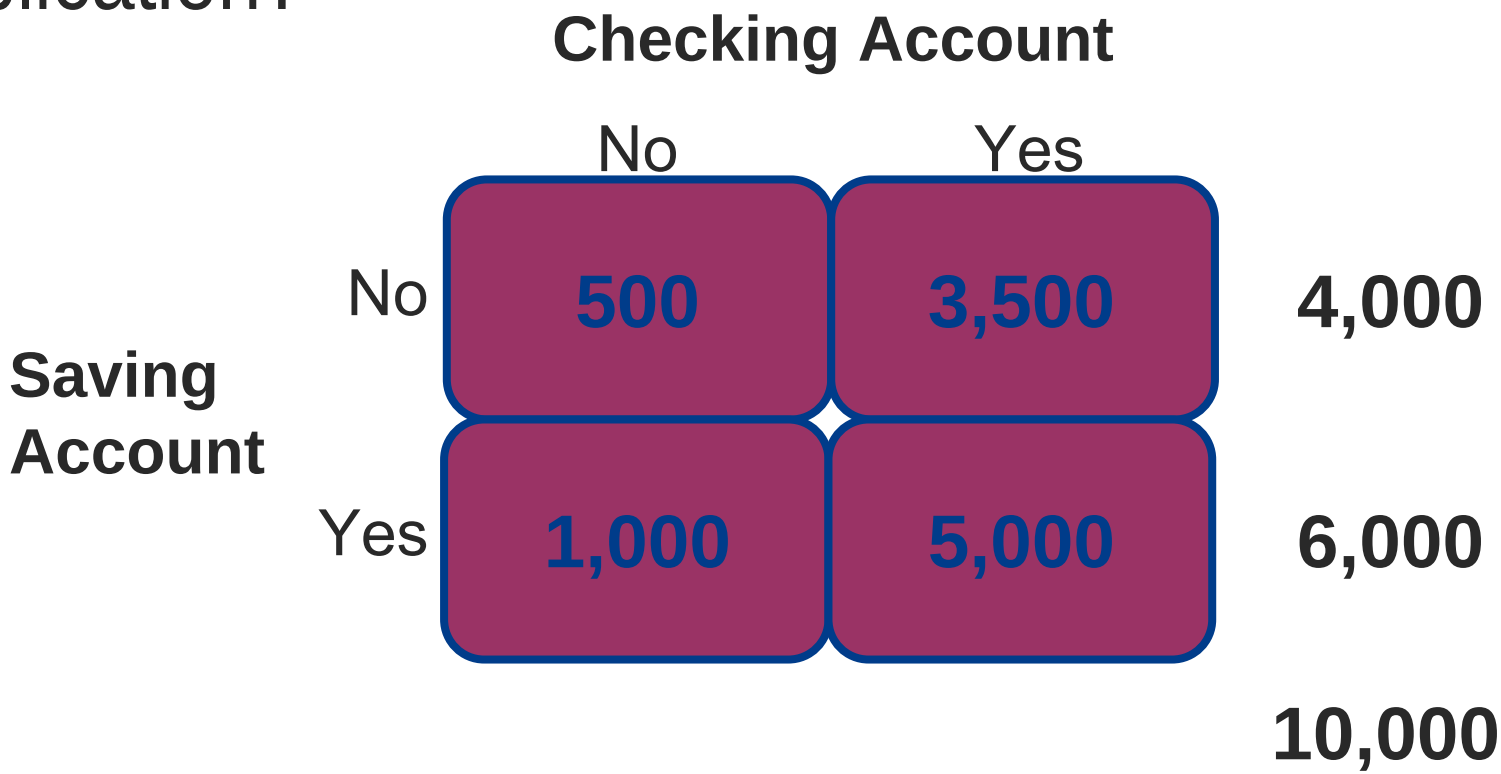
Unsupervised Classification

Training Data		Training Data
case 1: inputs, ?		*case 1*: inputs, cluster 1
case 2: inputs, ?		*case 2*: inputs, cluster 3
case 3: inputs, ?		*case 3*: inputs, cluster 2
case 4: inputs, ?		*case 4*: inputs, cluster 1
case 5: inputs, ?		*case 5*: inputs, cluster 2

Market Basket Analysis

Rule	Support	Confidence
A ⇒ D	2/5	2/3
C ⇒ A	2/5	2/4
A ⇒ C	2/5	2/3
B & C ⇒ D	1/5	1/3

Implication?



Support(SVG ⇒ CK) = 50%
Confidence(SVG ⇒ CK) = 83%
Expected Confidence(SVG ⇒ CK) = 85%
Lift(SVG ⇒ CK) = 0.83/0.85 < 1