Data mining Theory and examples

Mathias Lanner, SAS Institute Sverige

Agenda

- Introduction to data mining
- Process –methodology
- Data Challenges
- Modelling Techniques

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

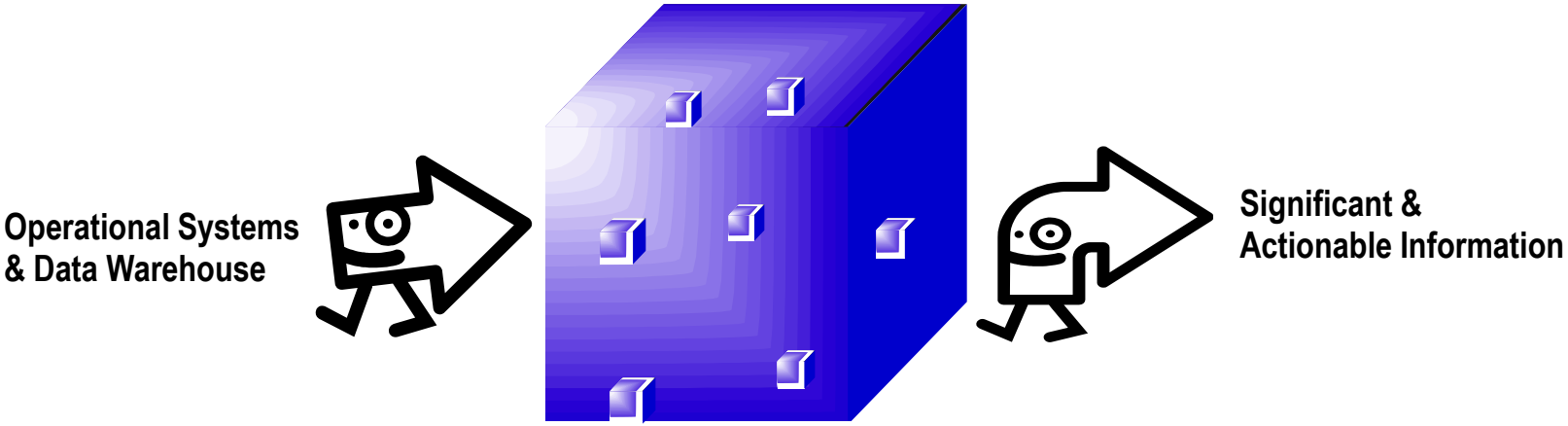
”Data Mining, as we use the term, is the exploration and analysis by automatic or semiautomatic means, of large quantities of data in order to discover meaningsful patterns and rules.”
-Berry and Linoff(1997), p.5

”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

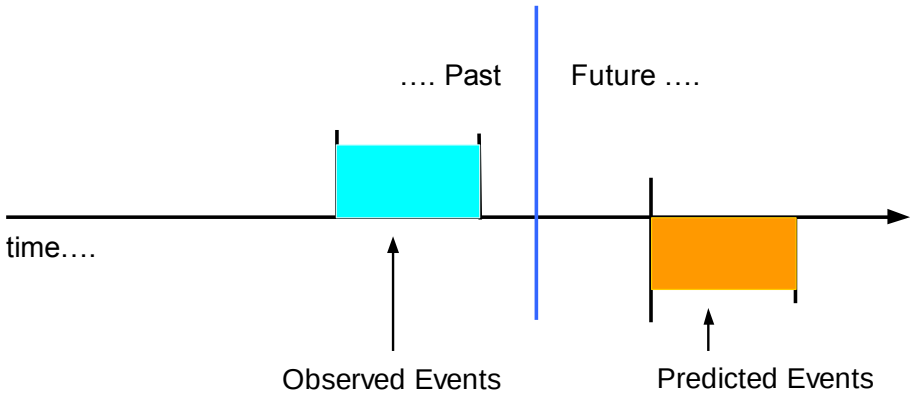
Data Mining Definition :

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

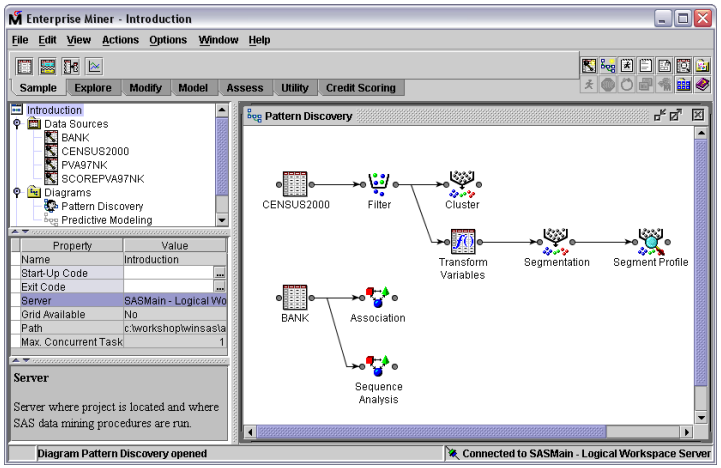


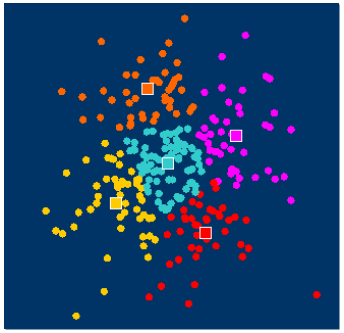
Data Mining Is:

- 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.

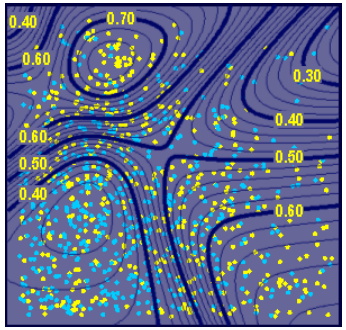


Two types of analysis





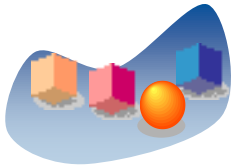
Pattern Discovery



Predictive Modeling

Pattern Discovery Applications

Data reduction



Novelty detection

Clustering

Market basket analysis

Sequence analysis

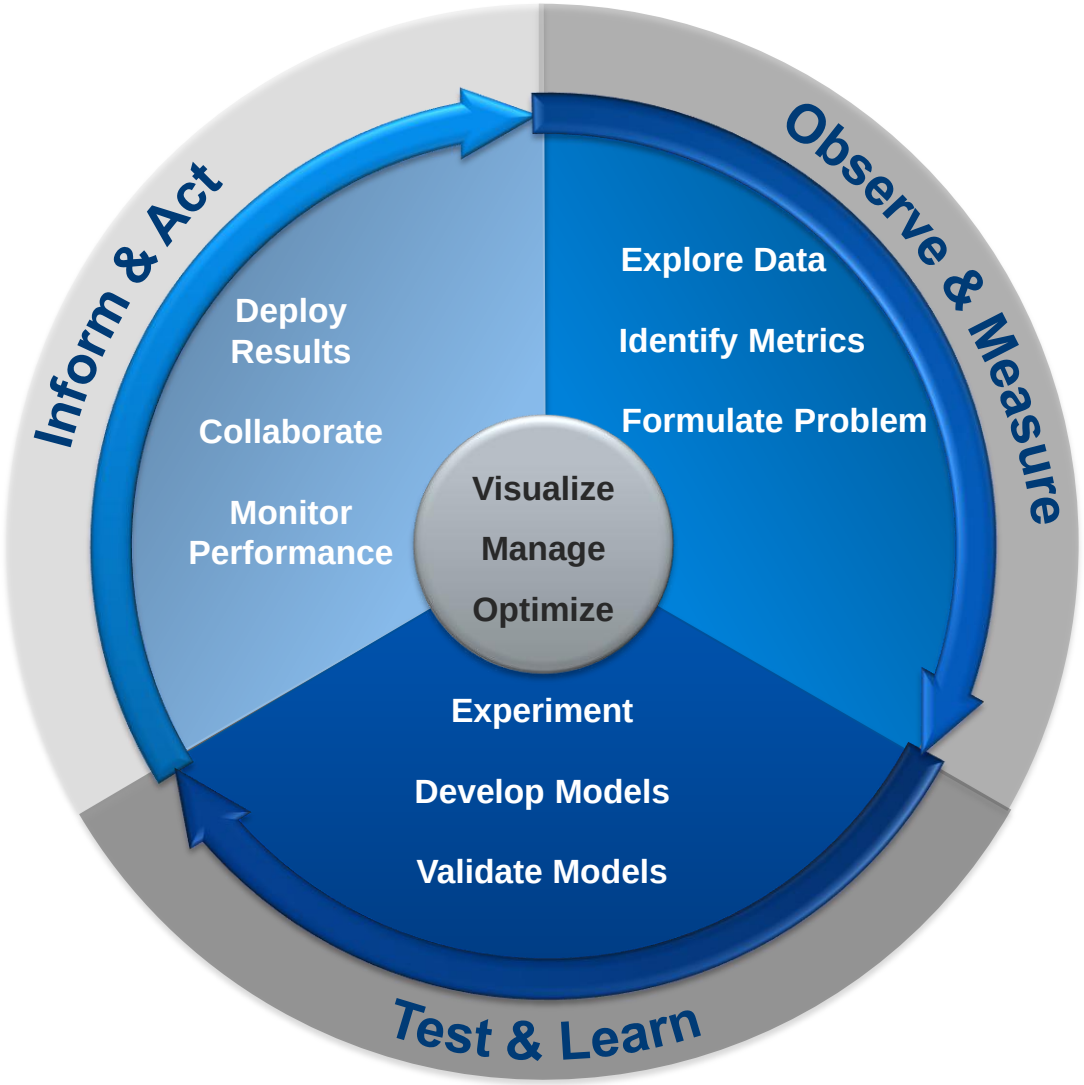
Cross-Industry Data Mining Applications
Customer Analytics

Application	What is Predicted?	Driven Business Decision
Profiling and Segmentation	Customer's behaviors and needs by segment	How to create better-targeted product/service offers?
Cross-sell and Up-Sell	Identify what will customer's buy?	Which product/service to recommend?
Acquisition and Retention	Customer's preferences and purchase patterns	How to grow and maintain valuable customers?
Campaign Management	Evaluate the success of customer communications	How to direct right offer to right person at the right time?
Profitability and Life-time Value	Understand the drivers of future value (margin and retention)	Identify economically valuable channels/demographics and incremental benefits?

Industry Specific Data Mining Applications

Application	What is Predicted?	Driven Business Decision
Credit Scoring *(Banking)*	Measure credit worthiness of new and existing set of customers	How to assess and control risk within existing (or new) consumer portfolios?
Market Basket Analysis *(Retail)*	Which products are likely to purchased together?	How to increase sales with cross-sell/up-sell, loyalty programs, promotions?
Asset Maintenance *(Utilities, Mfg., Oil & Gas)*	Identify real drivers of asset or equipment failure	How to minimize operational disruptions and maintenance costs?
Health & Condition Mgmt. *(Health Insurance)*	Identify patients at risk of a chronic illness & offer treatment program	How can we reduce healthcare costs and satisfy patients?
Fraud Mgmt. (Govt., *Insurance, Banks)*	Detect unknown fraud cases and future risks	How to decrease fraud losses and lower false positives?
Drug Discovery *(Life Science)*	Find compounds that have desirable effects & detect drug behavior during trials	How to bring drugs quickly and effectively to the marketplace?

Successful Analytics – Iterative and Interactive



Successful Data Mining – Iterative & Interactive

Data Manager

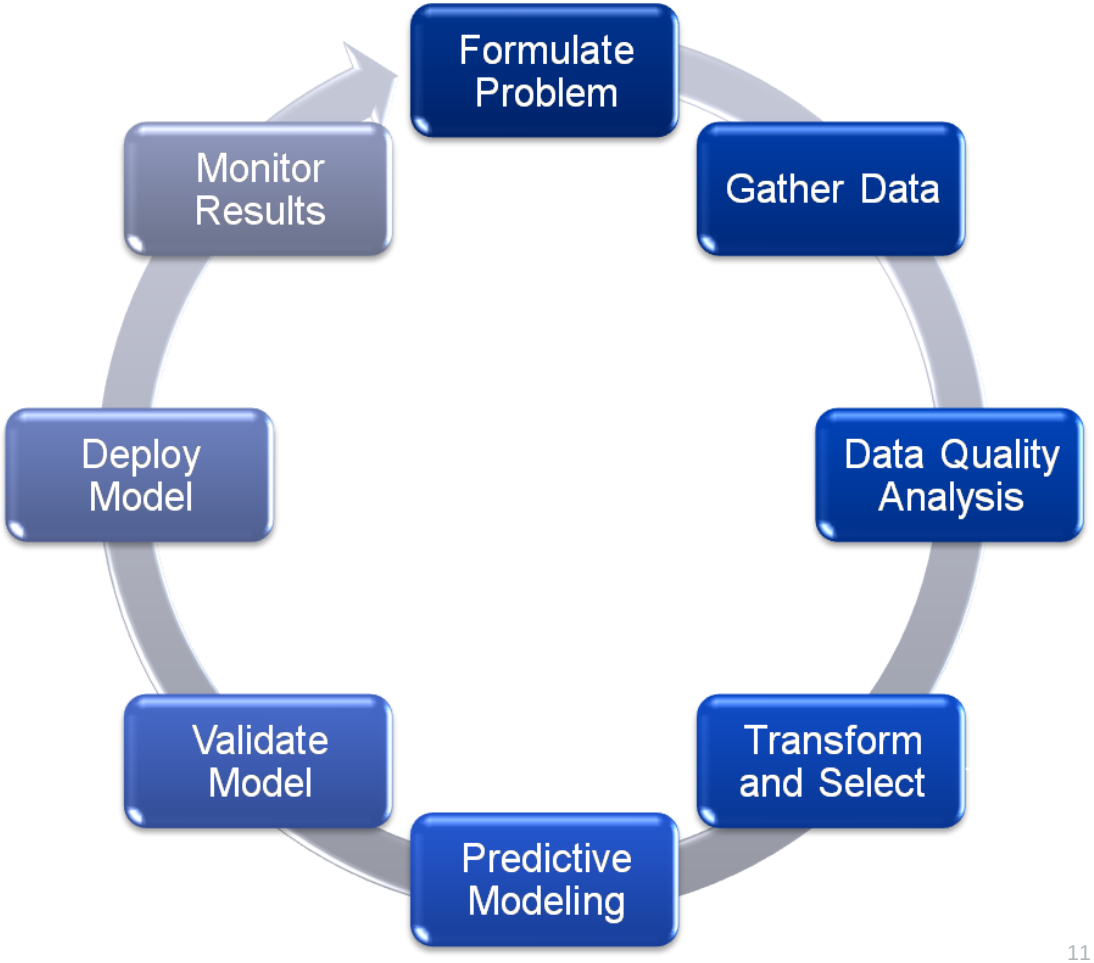
- Data preparation
- Deployment services
- Report administration

Data Miner

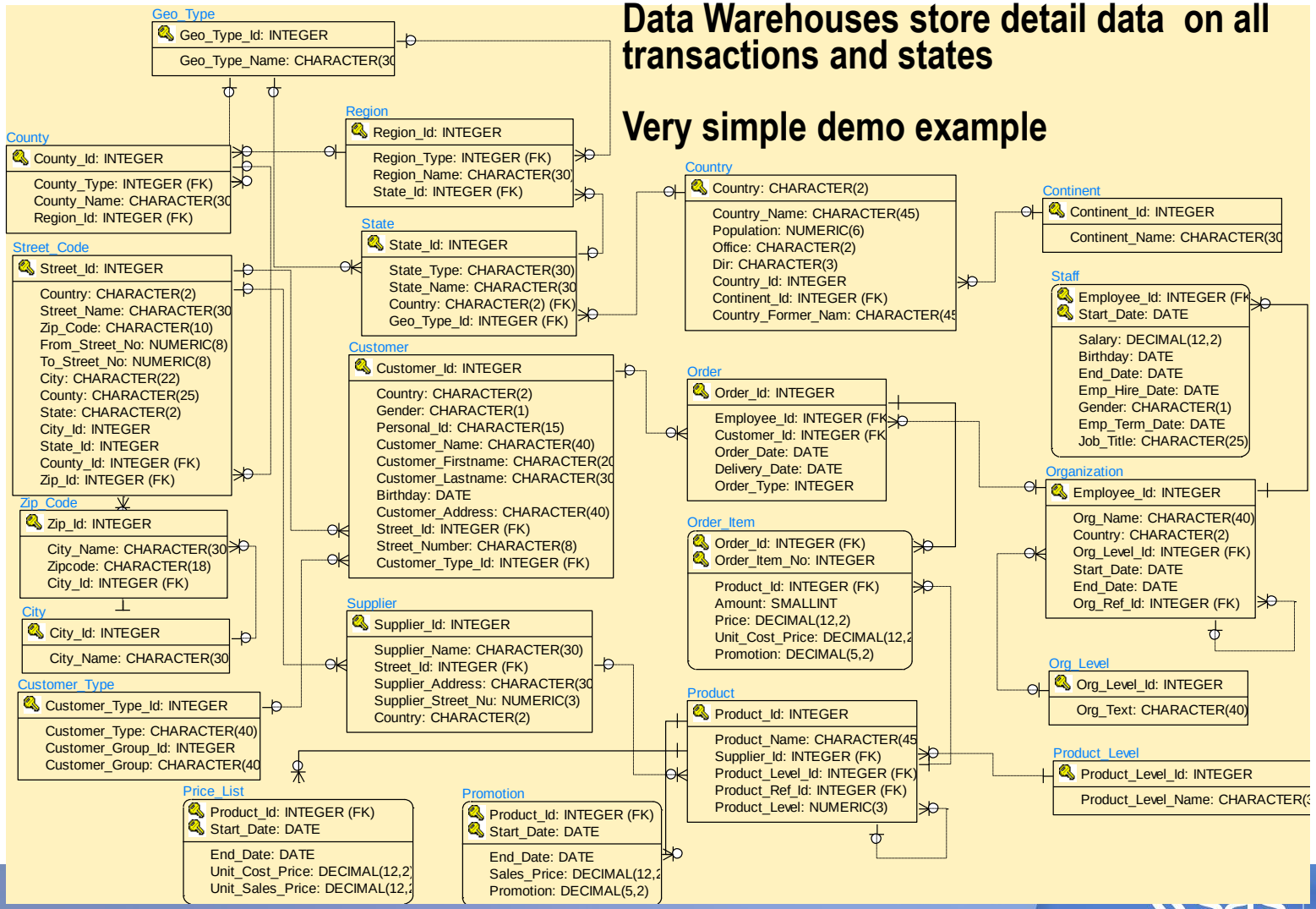
- Exploratory analysis
- Descriptive segmentation
- Predictive modeling

Business Manager

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



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

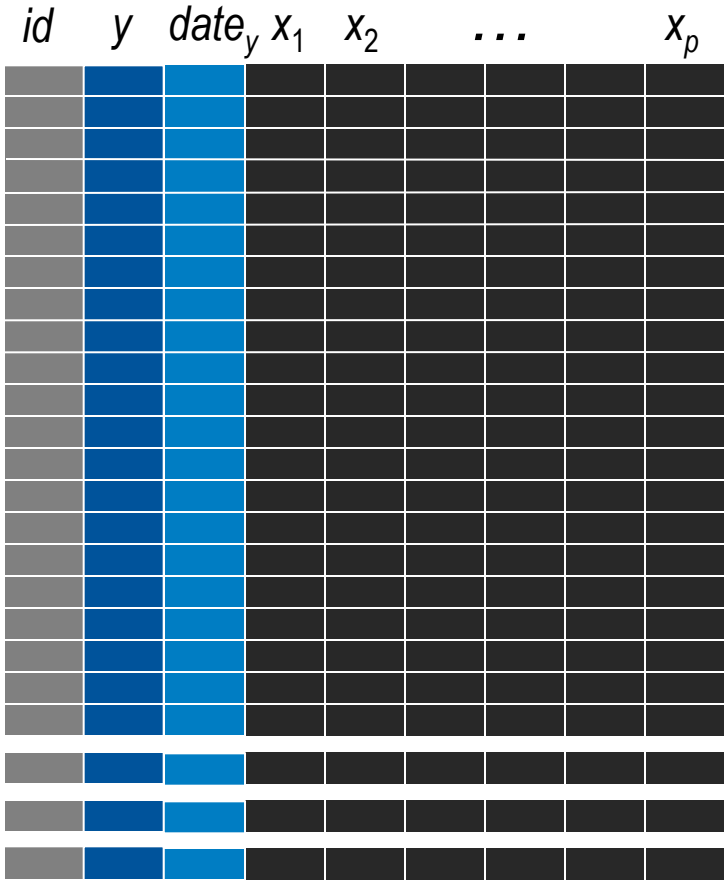
- Extreme commercial data warehouses
  - Many gigabytes of data
  - Stores may have 100,000+ SKU items
  - Sales histories for every item/basket saved

- 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
  - Social Media

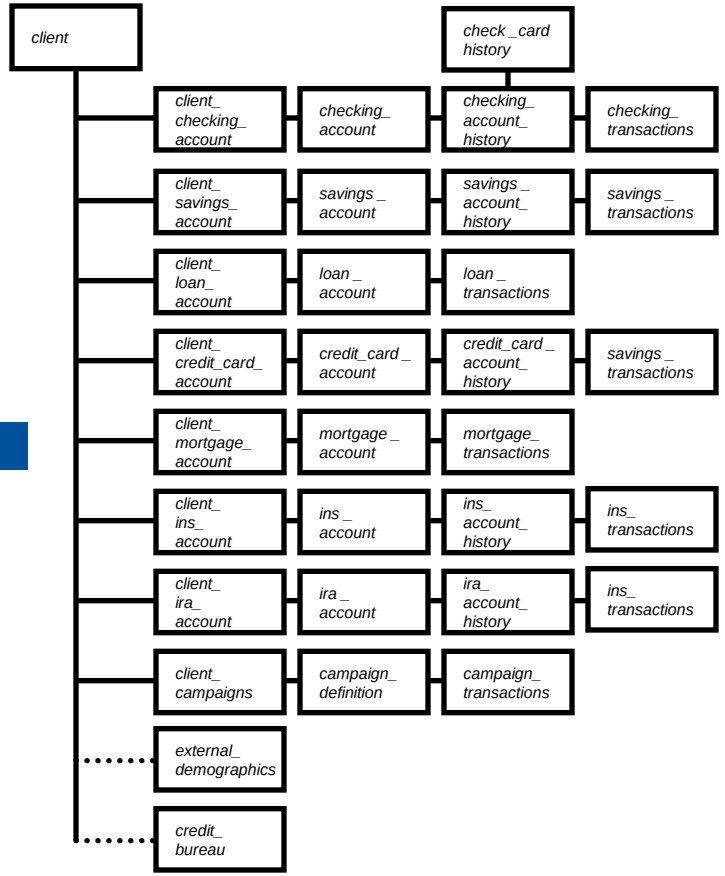
Integration

- 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



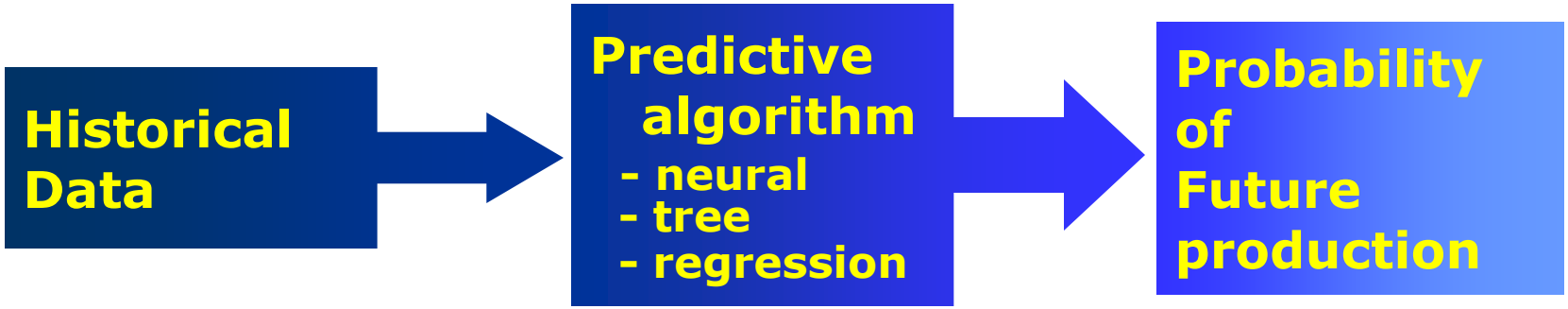
Model Development Data Set



Marketing Data Warehouse

Data Mining Algorithms

① predictive (supervised)
use data on past processes to *predict* future production

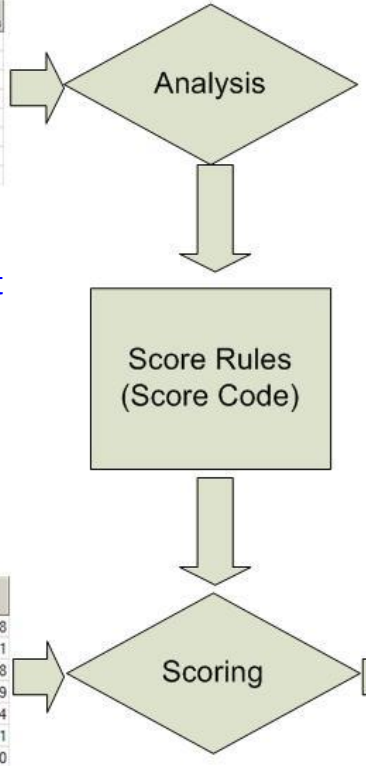


② descriptive (unsupervised)
use data on past processes to *describe* current situation

Predictive modeling 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



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

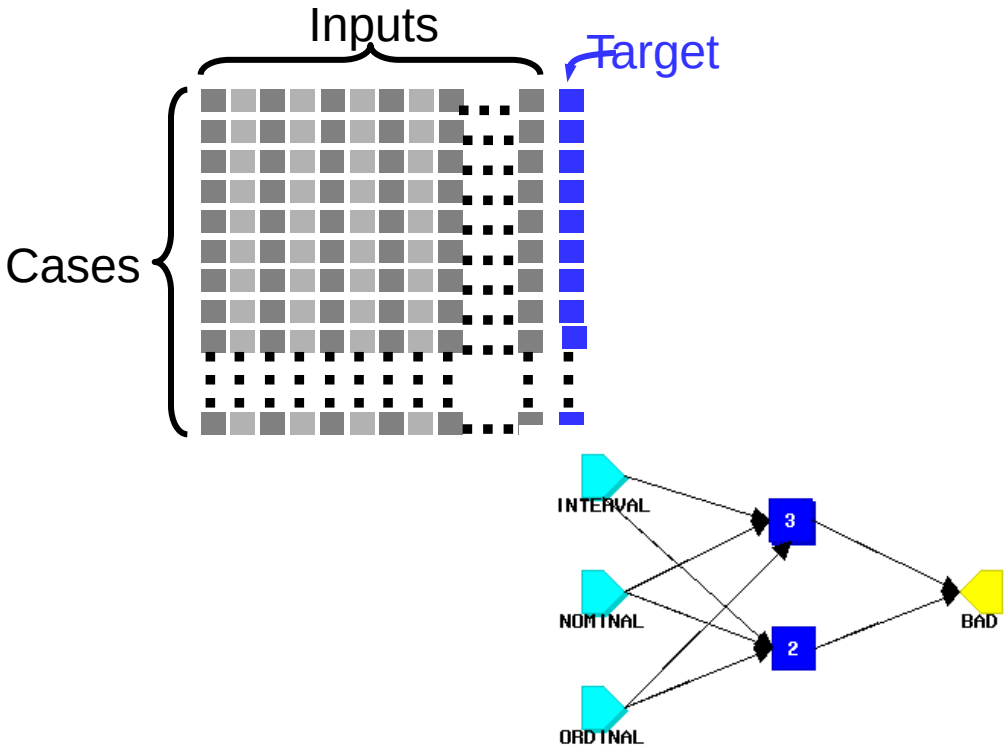
	Identification#	Vehicle Use	Value of Vehicle	Type of Car	Age	Claim_Prob
619	20240150	Commerci	7050	SUV	48	0.0412
620	20246469	Private	12020	Pickup	51	0.2186
621	20251776	Private	10120	SUV	38	0.8300
622	20254273	Commerci	16920	Van	39	0.3449
623	20264511	Private	37130	Pickup	54	0.0568
624	20267443	Commerci	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

Algorithms

- There is no BEST algorithm
- 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.

Supervised Learning

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

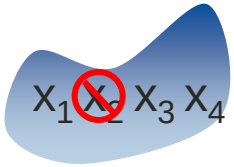


Enterprise Miner

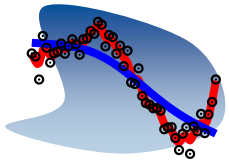
- Logistic and OLS
- Tree Classifiers
- Neural Networks
- Ensembles
- Memory Based Reasoning
- Two-stage modeling
- Fast Variable Selection
- Principal Components
- PLS Regression
- Support Vector Machine
- Gradient Bosting
- SAS/STAT

Predictive Modeling Essentials

Predict new cases



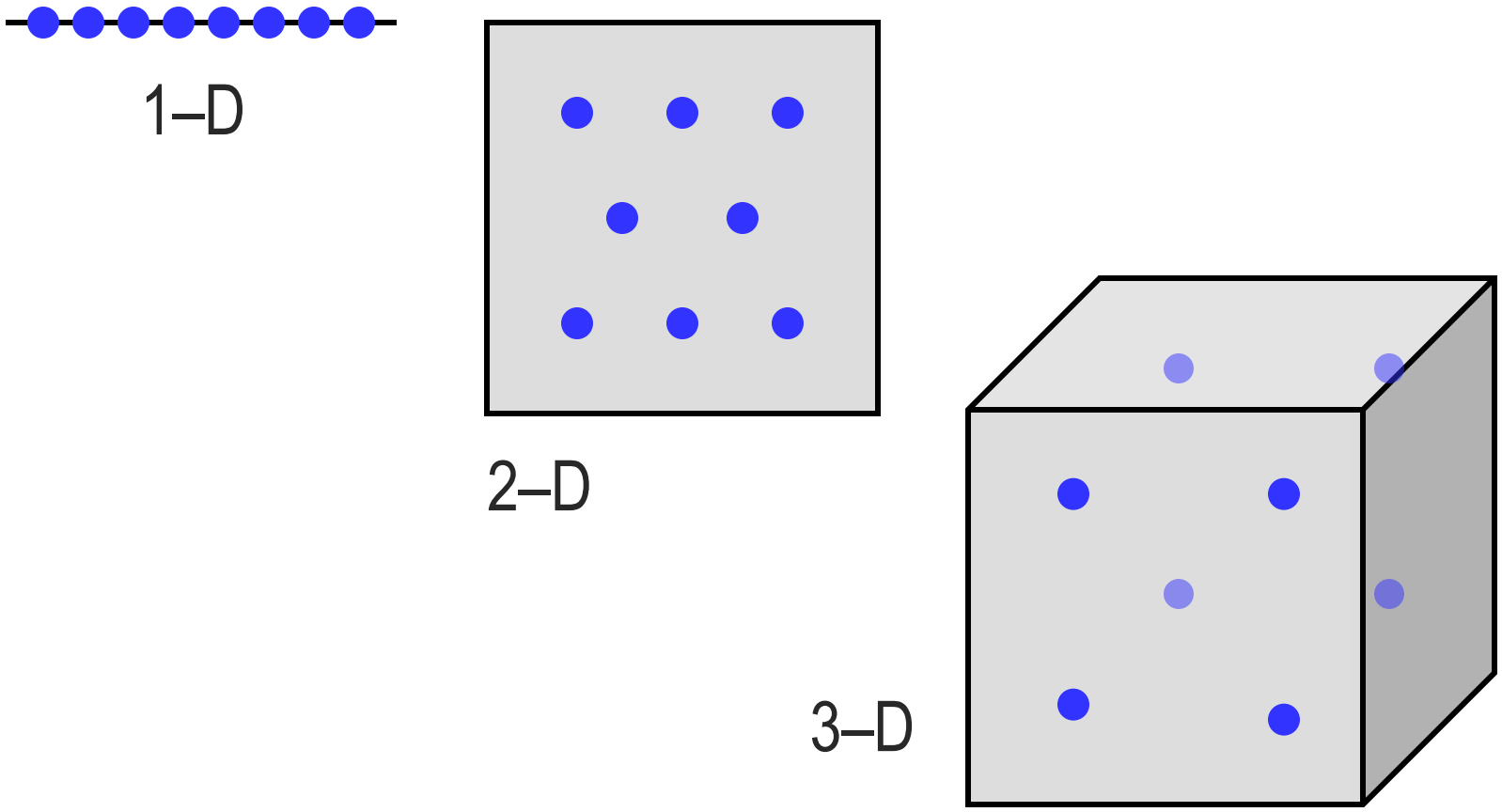
Select useful inputs



Optimize complexity

21

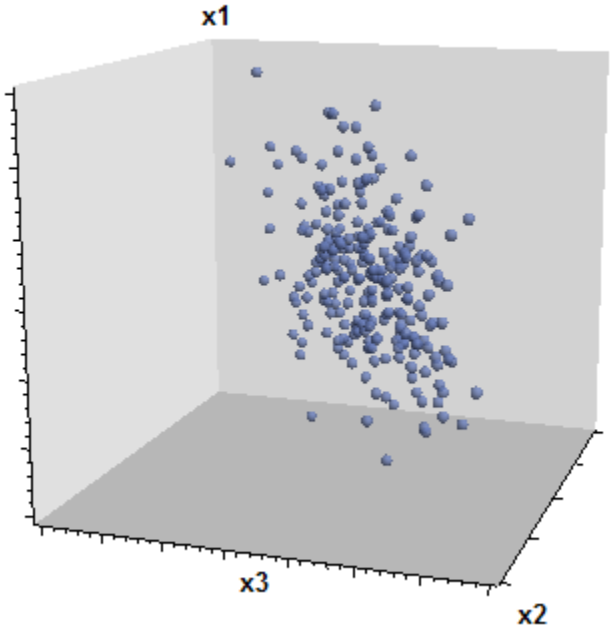
Curse of Dimensionality



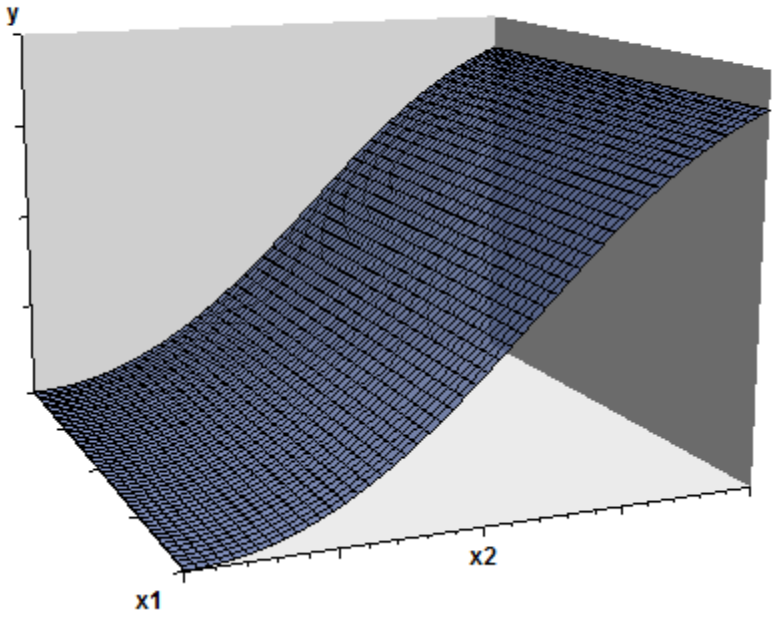
22

Input Selection

Redundancy

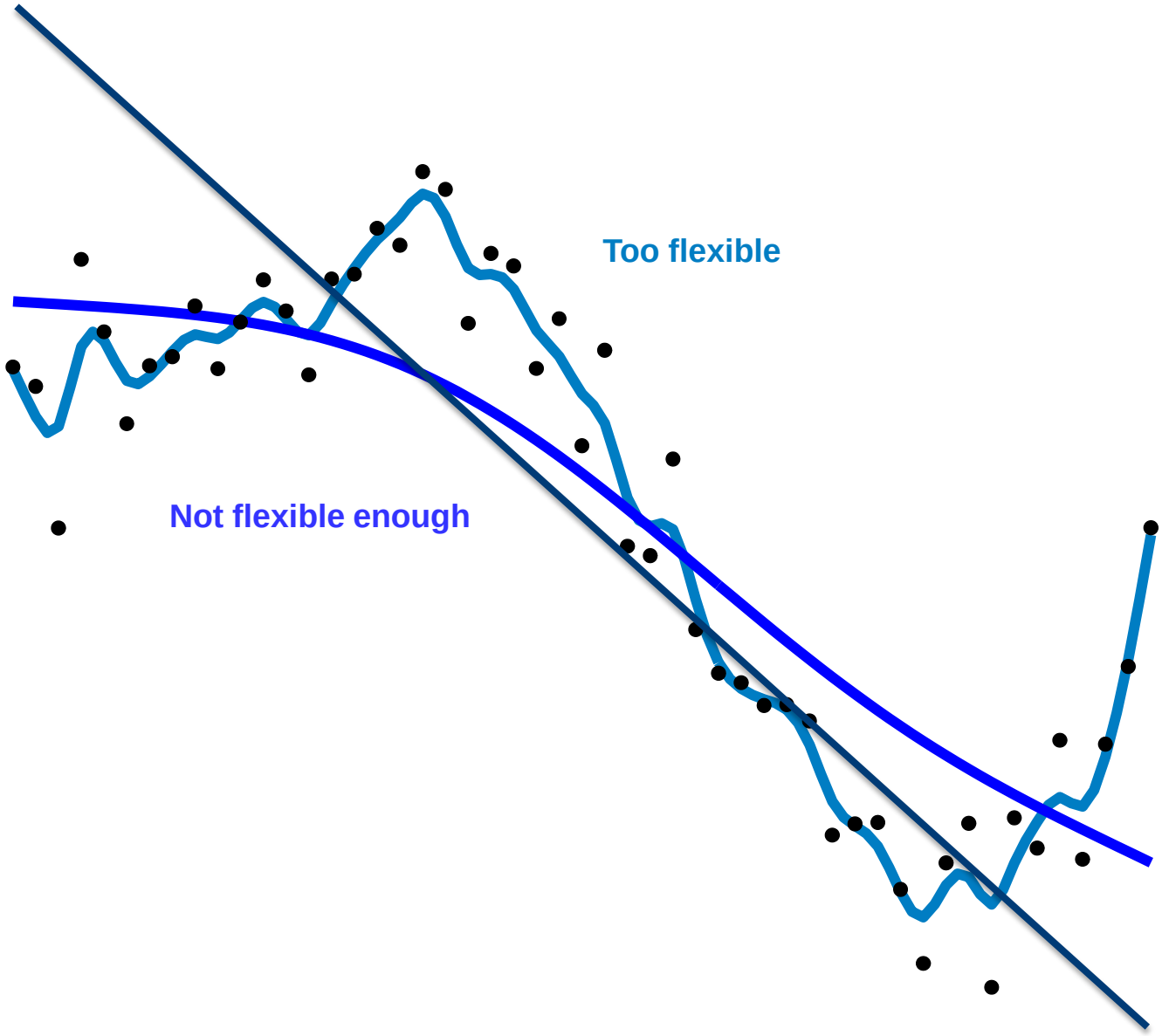


Irrelevancy



23

Model Complexity



24

Data Splitting

25

Assessment Types

The Model Comparison tool provides

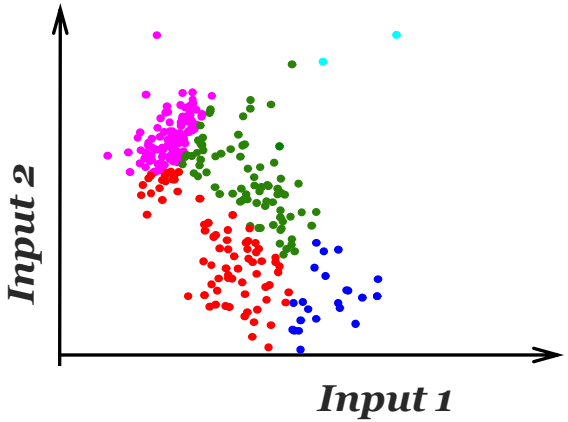
Summary statistics

Statistical graphics

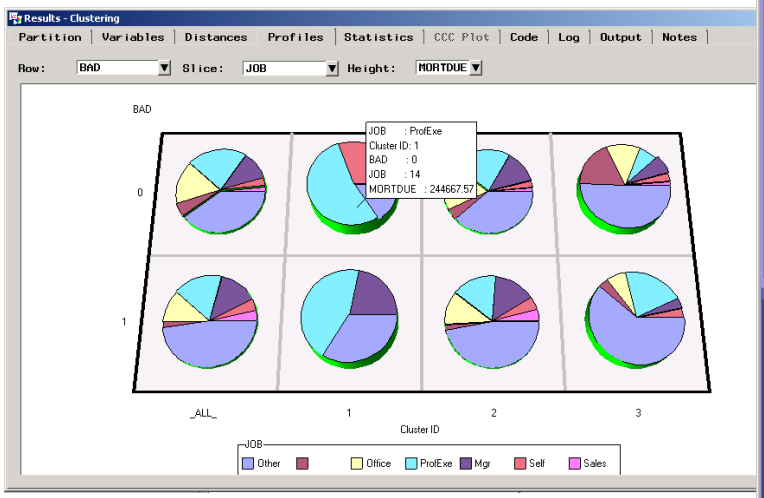
26

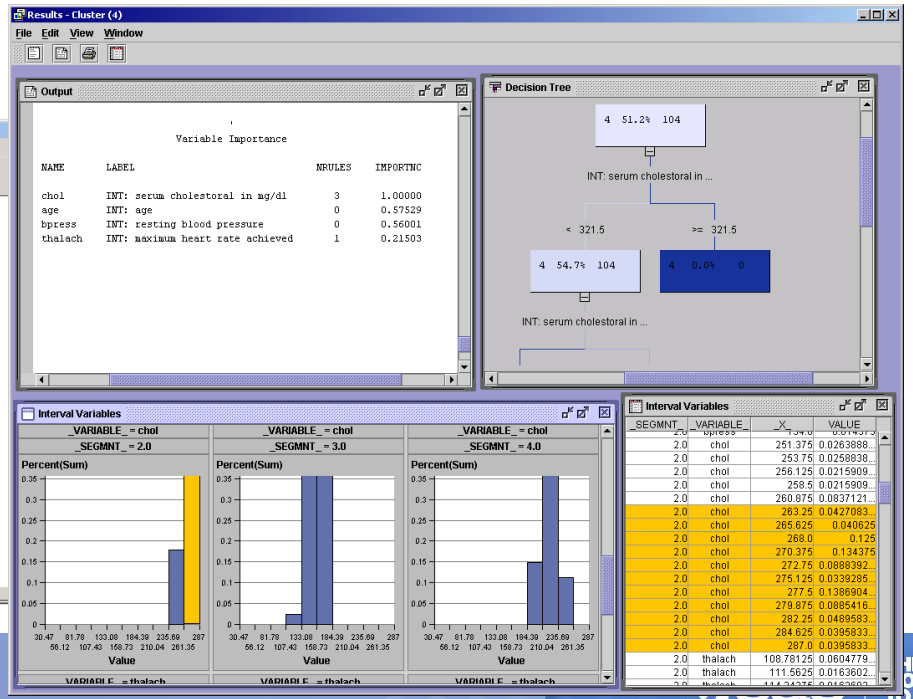
Unsupervised Learning

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

case 1: inputs, ?
case 2: inputs, ?
case 3: inputs, ?
case 4: inputs, ?
case 5: inputs, ?

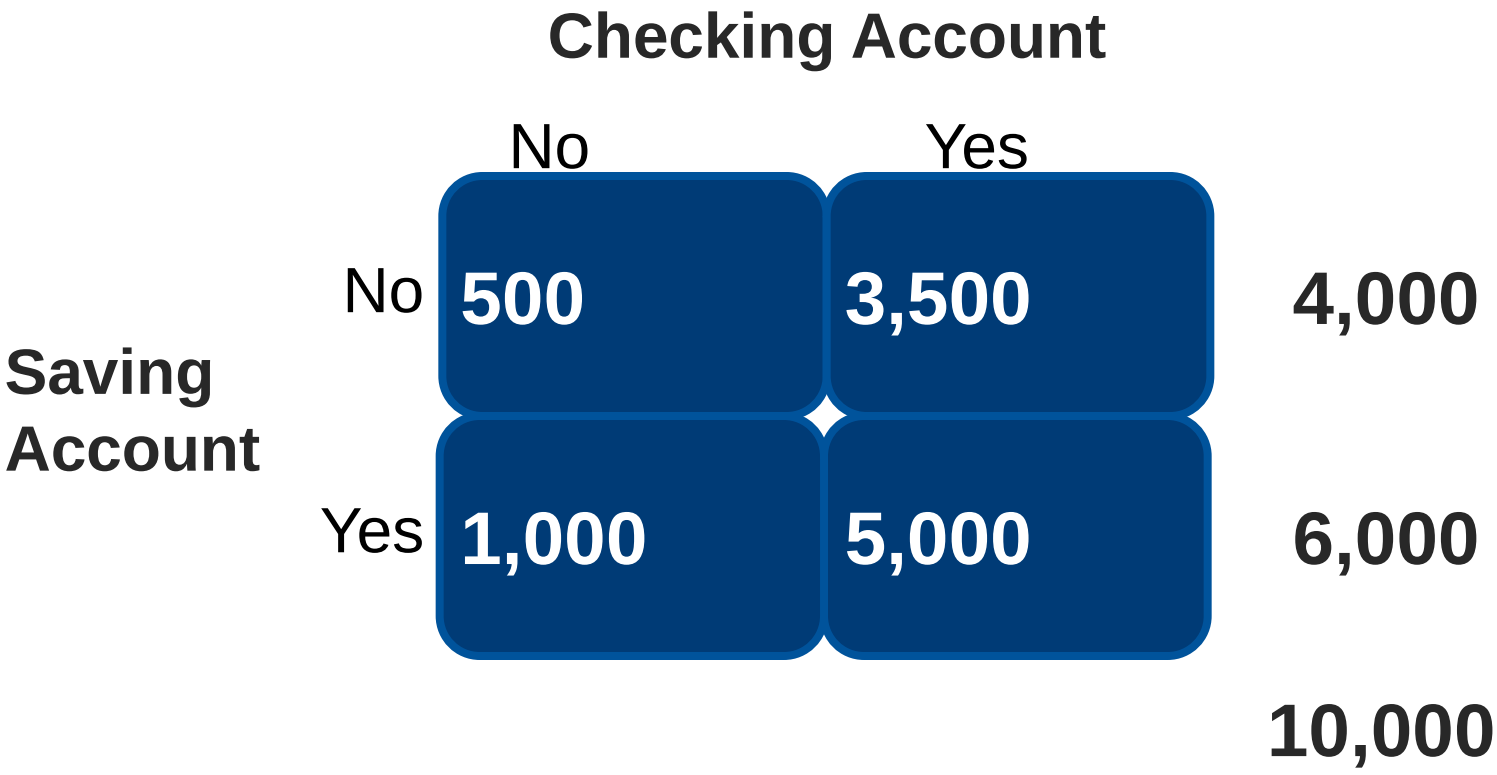
Training Data

case 1: inputs, cluster 1
case 2: inputs, cluster 3
case 3: inputs, cluster 2
case 4: inputs, cluster 1
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