Data Mining

Metoder och tekniker

Mathias Lanner Sas Institute

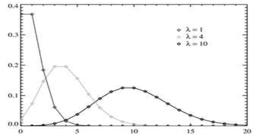
THE POWER TO KNOW®

Back in 1899…

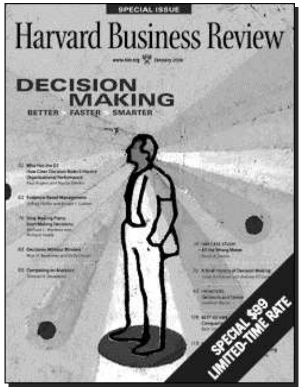
- William Sealy Gosset
  - (1876-1937)
- Also known as “Student”
  - “Student’s t-test”

Analytics & The Art Of Brewing Beer

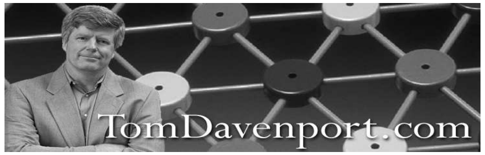
- “How much yeast to use for fermentation?”
  - Too little: incomplete fermentation
  - Too much: bitter beer
  - Concentration of yeast is difficult to measure
- Before: based on “gut feelings”
- Gosset: modeling using a probability distribution
  - “Poisson distribution” (Siméon-Denis Poisson, 1781–1840)
  - Result: more consistent products

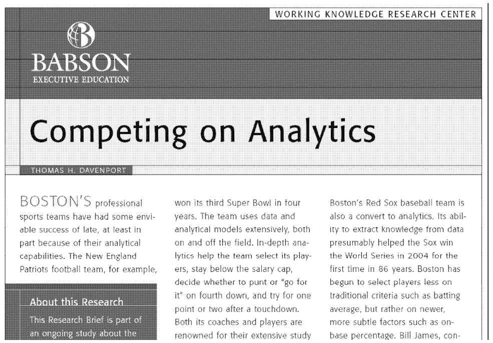


And Today…



Source: Harvard Business Review
(January 2006)





WORKING KNOWLEDGE RESEARCH CENTER

BABSON
EXECUTIVE EDUCATION

Competing on Analytics

THOMAS H. DAVENPORT

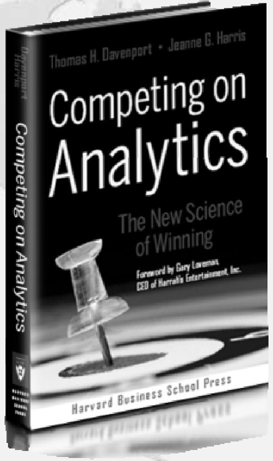
BOSTON'S professional sports teams have had some enviable success of late, at least in part because of their analytical capabilities. The New England Patriots football team, for example, won its third Super Bowl in four years. The team uses data and analytical models extensively, both on and off the field. In-depth analytics help the team select its players, stay below the salary cap, decide whether to punt or "go for it" on fourth down, and try for one point or two after a touchdown. Both its coaches and players are renowned for their extensive study

Boston's Red Sox baseball team is also a convert to analytics. Its ability to extract knowledge from data presumably helped the Sox win the World Series in 2004 for the first time in 86 years. Boston has begun to select players less on traditional criteria such as batting average, but rather on newer, more subtle factors such as on-base percentage. Bill James, con-

About this Research

This Research Brief is part of an ongoing study about the

The New "Era of Analytics"


Thomas H. Davenport • Jeanne G. Harris
Competing on Analytics
The New Science of Winning
Foreword by Gary Loveman, CEO of Harrah's Entertainment, Inc.
Harvard Business School Press

"Previous bases for competition … have been eroded … That leaves three things as the basis for competition:

- Efficient & effective execution
- Smart decision making
- Ability to wring every last drop of value from business processes

… all of which can be gained through sophisticated use of analytics."

"Competing on Analytics" (Davenport & Harris)
Harvard Business School Press
Worldwide Release: March 6, 2007

Competing on Analytics…

The idea of competing on analytics is not entirely new.

What is new is the spreading of analytical competition from individual business units to an enterprise-wide perspective.

Vad är data mining

THE POWER TO KNOW®

Vad är 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 [is] the process of efficient discovery of nonobvious valuble information from large collection of data.”*
>
> **-Berson and Smith (1997), p.565**

”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 meaningful patterns and rules.”

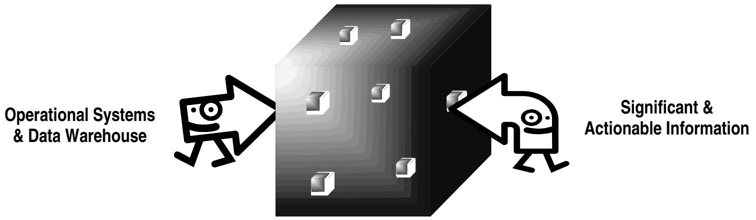
-Berry and Linoff(1997), p.5

”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 recognition 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*



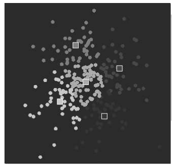
Vad handlar Data mining om?

- Data mining handlar om att finna gömda meningsfulla* och användbara** relationer mellan olika fenomen i stora data volymer
- Exempel, vi försöker upptäcka FRAUD (target) genom att använda ett antal olika indikatorer (inputs)
  - Vilket land har transaktionen gjorts i
  - Vilket belopp rör det sig om
  - Frekvens på kort användning
  - etc.
- Data mining projekt lyckas när:
  - När gömda relationer verkligen finns och är tillräckligt staka
  - Rent data, (få dubbletter bland observationerna, få missing värden, homogen kodning av nominala inputs, etc
  - Vi har en liten ide om vilka gömda relationer vi vill avslöja (affärskunskap)
  - Olika typer av färdigheter/kunskaper måste finnas i teamet (IT, verksamhetskunskap, statistik, AI)
  - Ledningsförankrat

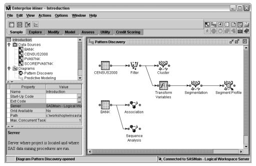
* Meningsfulla = Tolkningsbara

** Använbara = Har ett signifikant affärsvärde

Two types of analysis



Pattern Discovery



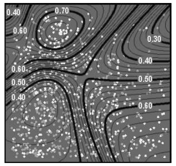
Property	Value
Name	Introduction
Start-Up Code	
Exit Code	
Server	SASMain - Logical Wo
Grid Available	No
Path	c:\workshop\winsas\a
Max. Concurrent Task	1

Server

Server where project is located and where SAS data mining procedures are run.

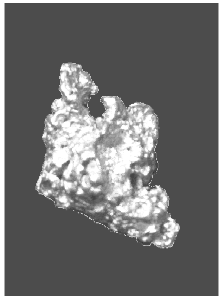
Diagram Pattern Discovery opened

Connected to SASMain - Logical Workspace Server



Predictive Modeling

Pattern Discovery



The Essence of Data Mining?

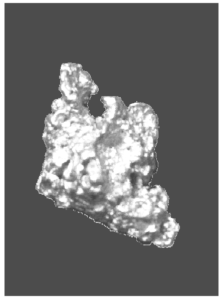
"…the discovery of interesting, unexpected, or valuable structures in large data sets."

– 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

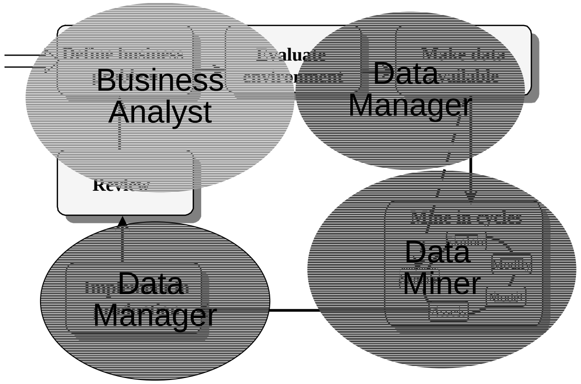
Predictive Modeling Applications

- **Database marketing**
- **Financial risk management**
- **Fraud detection**
- **Process monitoring**
- **Pattern detection**

Process - Metodik

THE POWER TO KNOW®

Project steps and rolls

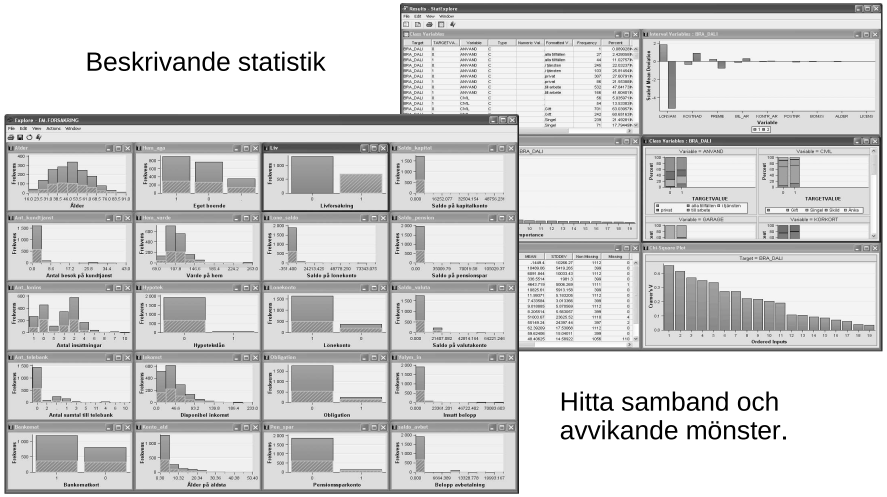


- Define business problem
- Evaluate environment and make data available
- Structured analysis
- Implement into production environments
- Review
- Team with Business, analyst and data manager

Utforska data och hitta samband

SEMMA

Beskrivande statistik

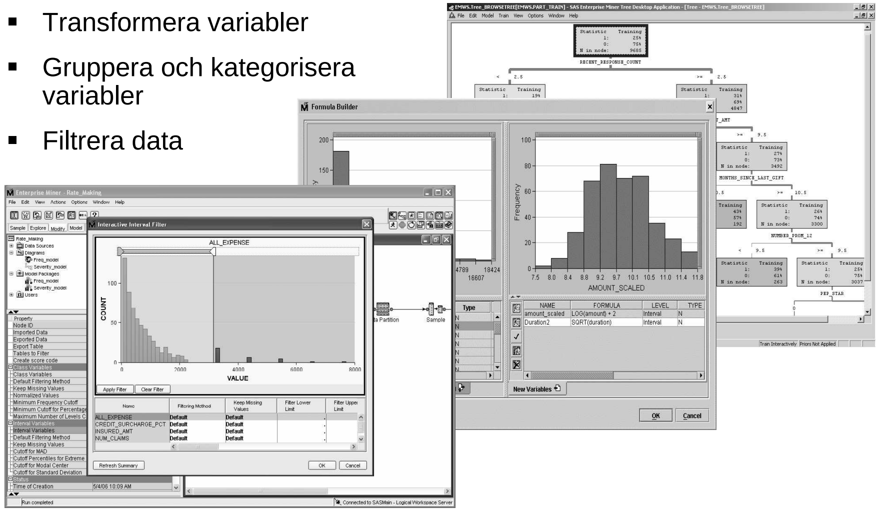


Hitta samband och avvikande mönster.

Modifiera data/variabler

SEMMA

- Transformera variabler
- Gruppera och kategorisera variabler
- Filtrera data

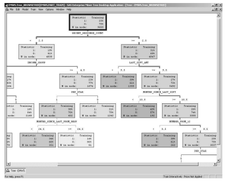


Modellering

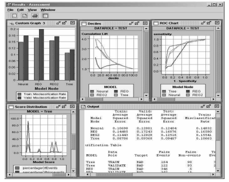
SEMMA

Prediktiv modellering

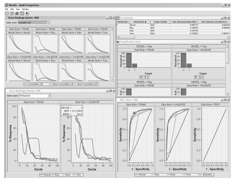
Beslutsträd



Regressionsanalys

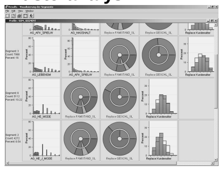


Neurala nätverk

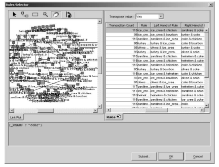


Stöd från beskrivande analys

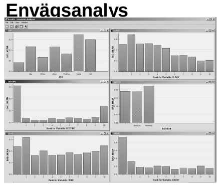
Klusteranalys



Associationsanalys



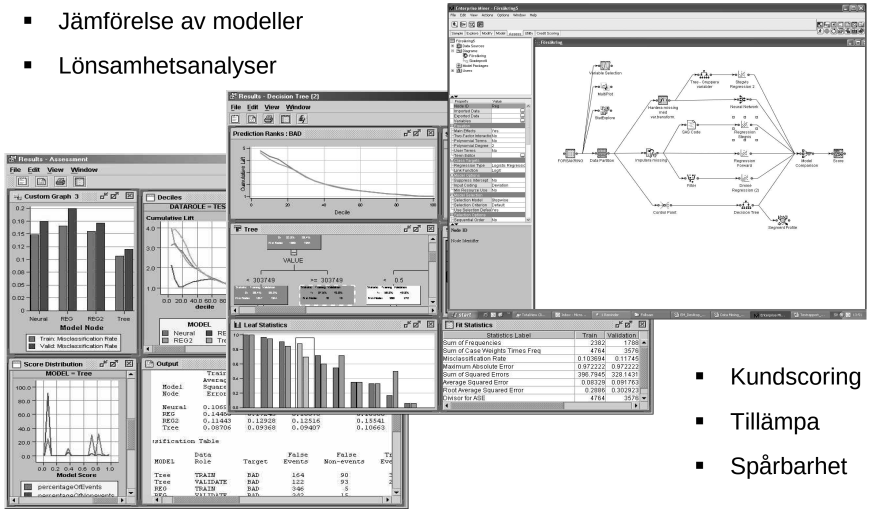
Envägsanalys



Assess – Utvärdera resultat

SEMMA

- Jämförelse av modeller
- Lönsamhetsanalyser



Fit Statistics

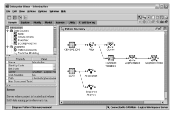
Statistics Label	Train	Validation
Sum of Frequencies	2382	1788
Sum of Case Weights Times Freq	4764	3576
Misclassification Rate	0.103694	0.11745
Maximum Absolute Error	0.972222	0.972222
Sum of Squared Errors	396.7945	328.1431
Average Squared Error	0.08329	0.091763
Root Average Squared Error	0.2886	0.302923
Divisor for ASE	4764	3576

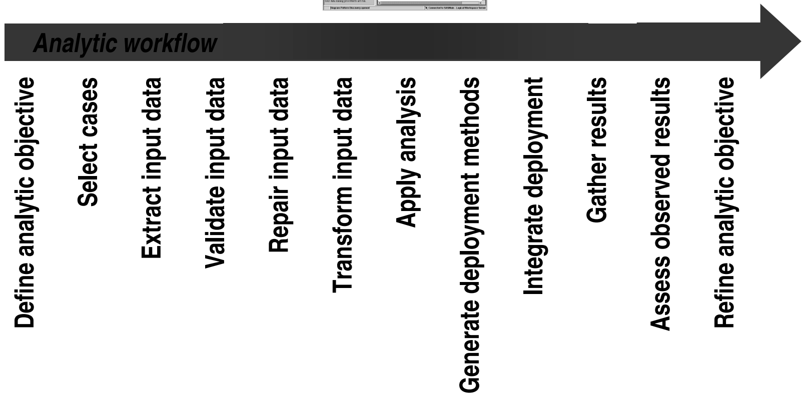
Output

```
              Train
              Averag
  Model       Square
  Node         Error

  Neural      0.1069
  REG         0.1448
  REG2        0.11443    0.12928    0.12516          0.15541
  Tree        0.08706    0.09368    0.09407          0.10663

sification Table

            Data                     False        False       Tr
MODEL       Role         Target     Events    Non-events     Eve

Tree        TRAIN         BAD        164          90
Tree        VALIDATE      BAD        122          93
REG         TRAIN         BAD        346           5
REG         VALIDATE      BAD        342          15
```

- Kundscoring
- Tillämpa
- Spårbarhet

The Analytic Workflow





Define analytic objective

Select cases

Extract input data

Validate input data

Repair input data

Transform input data

Apply analysis

Generate deployment methods

Integrate deployment

Gather results

Assess observed results

Refine analytic objective

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
  - Prediction

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
  - Prediction

Data mining

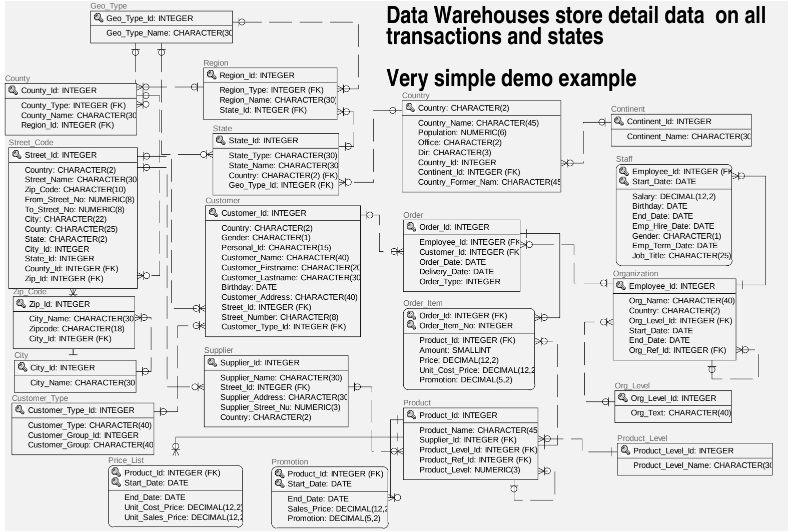
- Commercial
- Posterior Hypothesis
  - Idea after data acquisition
  - Data acquisition opportunistic
- No Experimental Design
  - Explore data
  - Create hypothesis
  - Generate query
  - Create models
- Prediction
  - Lift, Profit, Response
  - Inference

Data utmaningar

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 Warehouses store detail data on all transactions and states

Very simple demo example

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

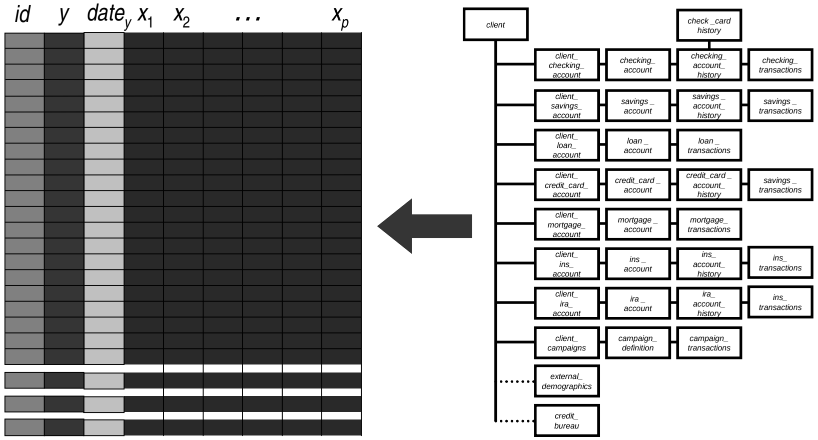
New data sources

- 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:

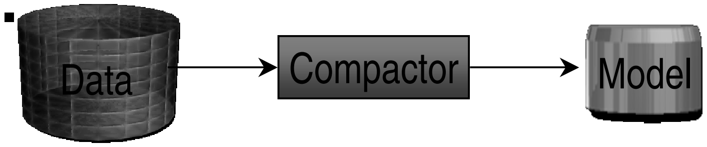
  “10 to 20 percent of the raw data used is corrupt or incomplete in some way. It is not unusual to discover that as many as half the records in a database contain some type of information that needs to be corrected”
  META Group Program Director John Ladley

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

Tekniker

Algorithms



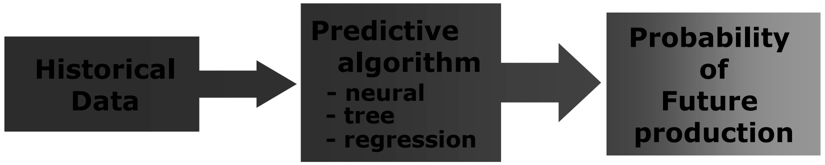
- A model is an abstraction of the data and belongs with the data
- There is nothing more in a model than what is already in the data

Algorithms

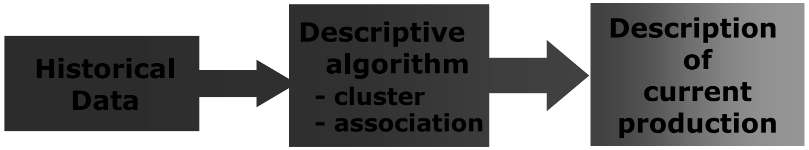
- There is no BEST algorithm per se
- Depends on
  - Nature of relationships in data
  - Data quality
  - Time available to build a model
  - Nature of model deployment
    - operational use
    - insights for business users
    - decision support etc.

Data Mining Algorithms

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

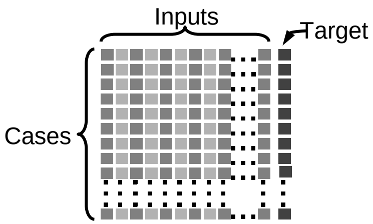


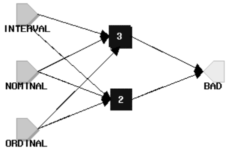
□descriptive (unsupervised)
use data on past processes to *describe* current situation



Supervised Learning

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

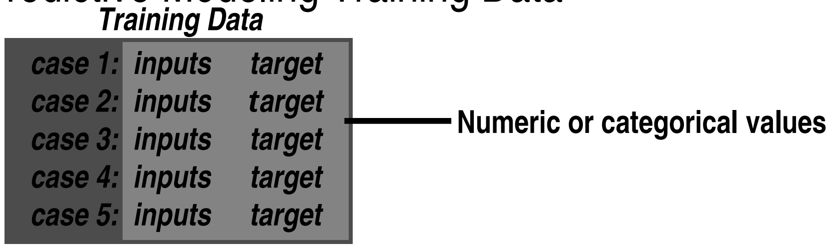




Enterprise Miner

- 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

Predictive Modeling Training Data



Predictive Modeling Score Data

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Score Data

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

Only input values known

Predictions

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Predictions

Score Data

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

Predictions

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Predictions

prediction
prediction
prediction
prediction
prediction

Score Data

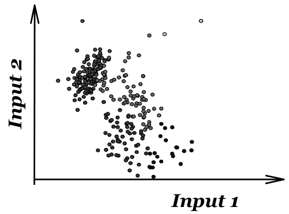
case 1:	*inputs*	*?*
case 2:	*inputs*	*?*
case 3:	*inputs*	*?*
case 4:	*inputs*	*?*
case 5:	*inputs*	*?*

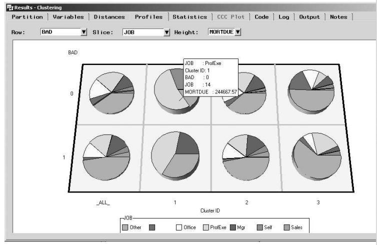
prediction
prediction
prediction
prediction
prediction

Unsupervised Learning

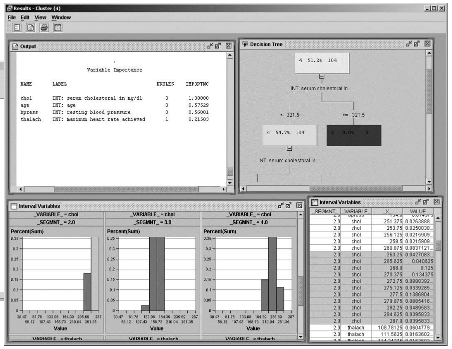
Tries to divide the data into groups such that the observations within a group have traits more similar than those assigned to different groups Enterprise Miner

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





JOB : ProfExe
Cluster ID: 1
BAD : 0
JOB : 14
MORTDUE : 244667.57



```
                    Variable Importance

NAME      LABEL                                NRULES   IMPORTNC

chol      INT: serum cholestoral in mg/dl         3      1.00000
age       INT: age                                0      0.57529
bpress    INT: resting blood pressure             0      0.56001
thalach   INT: maximum heart rate achieved        1      0.21503
```

4 51.2% 104

INT: serum cholestoral in ...

< 321.5

>= 321.5

4 54.7% 104

4 0.0% 0

INT: serum cholestoral in ...

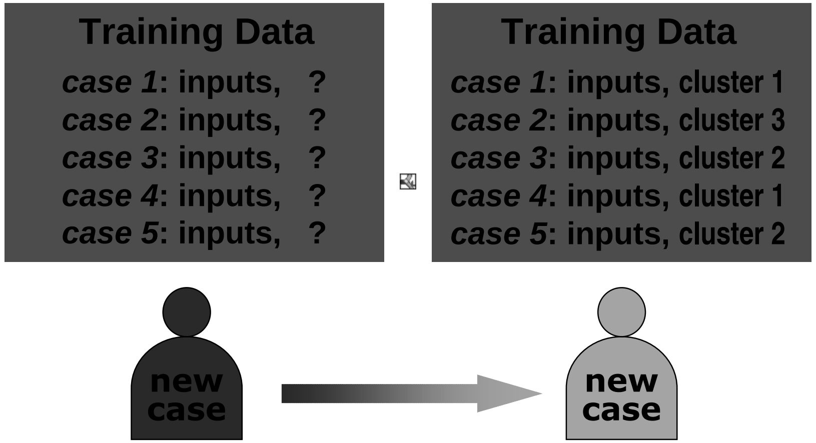
VARIABLE = chol, _SEGMNT_ = 2.0

VARIABLE = chol, _SEGMNT_ = 3.0

VARIABLE = chol, _SEGMNT_ = 4.0

SEGMNT	_VARIABLE_	_X_	VALUE
2.0	chol	251.375	0.0263888...
2.0	chol	253.75	0.0258838...
2.0	chol	256.125	0.0215909...
2.0	chol	258.5	0.0215909...
2.0	chol	260.875	0.0837121...
2.0	chol	263.25	0.0427083...
2.0	chol	265.625	0.040625
2.0	chol	268.0	0.125
2.0	chol	270.375	0.134375
2.0	chol	272.75	0.0888392...
2.0	chol	275.125	0.0339285...
2.0	chol	277.5	0.1386904...
2.0	chol	279.875	0.0885416...
2.0	chol	282.25	0.0489583...
2.0	chol	284.625	0.0395833...
2.0	chol	287.0	0.0395833...
2.0	thalach	108.78125	0.0604779...
2.0	thalach	111.5625	0.0163602...

Unsupervised Classification



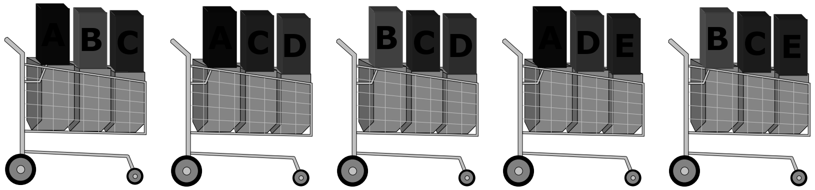
Training Data

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?

		Checking Account		
		No	Yes	
Saving Account	No	500	3,500	**4,000**
	Yes	1,000	5,000	**6,000**
				10,000

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