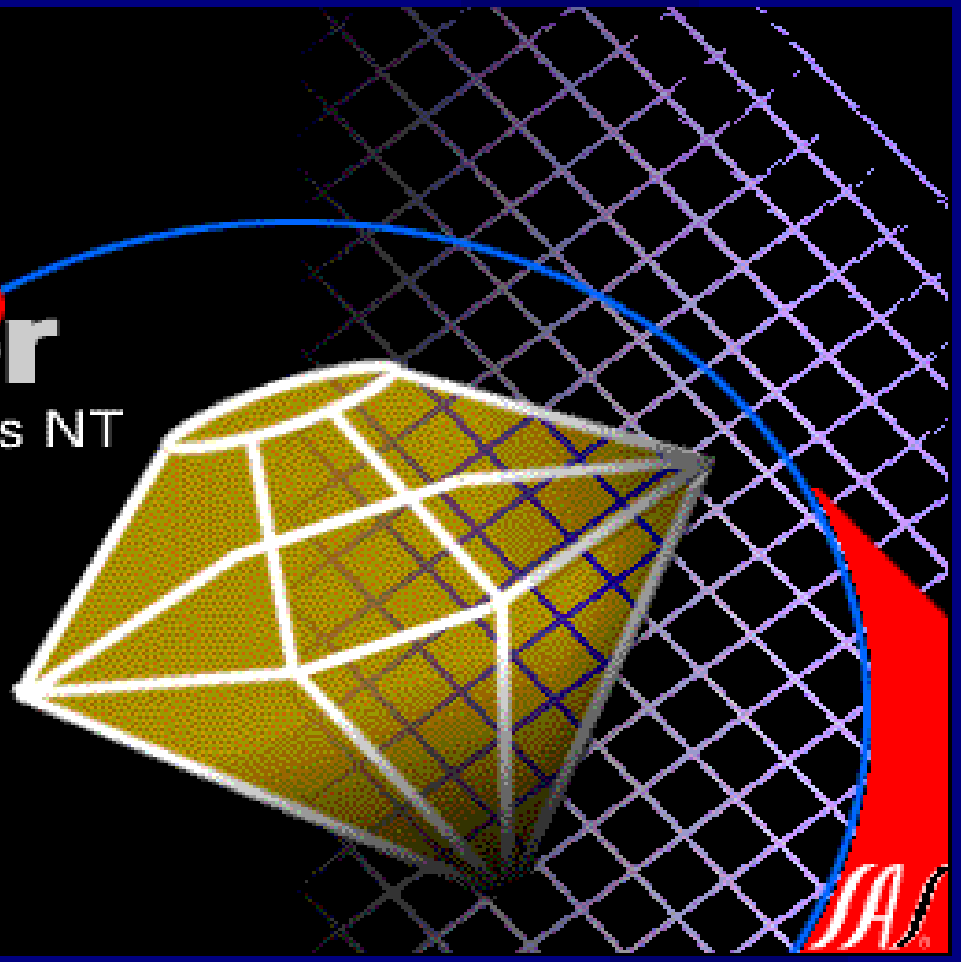
Data Mining mit der SEMMA Methodik

Enterprise Miner
For Windows NT

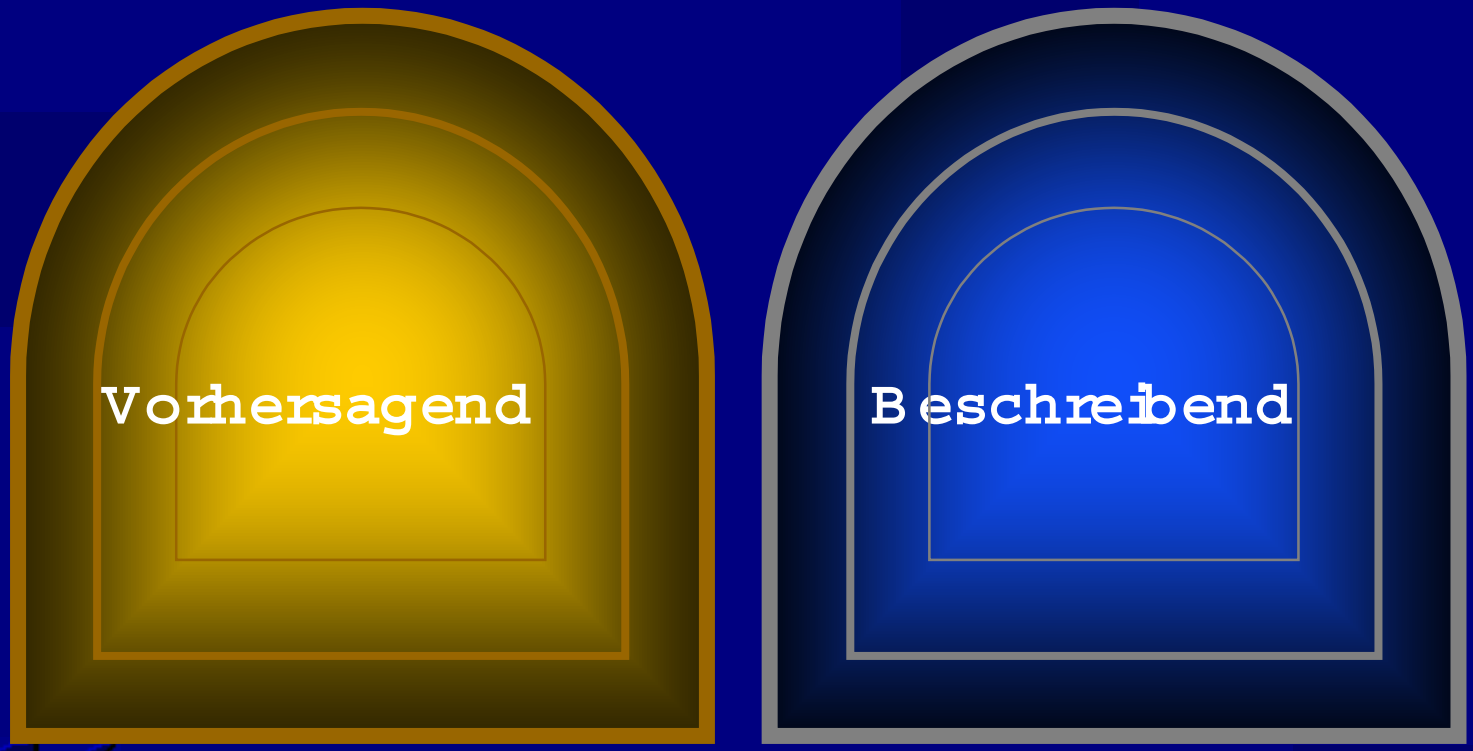


Reinhard Strüby,
SAS Institute
Stephanie Freese,
Herlitz PBS AG

Data Mining

Data Mining: Prozeß der ***Selektion, Exploration*** und ***Modellierung*** großer Datenmengen, um Information zu gewinnen und in Geschäftsvorteile umzusetzen.

2 Wege in das Datenbergwerk

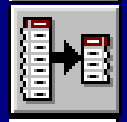


Warum Mining?

- **Wartungsoptimierung**
- **Kapazitätsplanung**
- **Ausfall-Management**
- **Kundenpflege**
- **Profilierung/Segmentierung**
- **Cross-Selling**
- **Betrugserkennung**

Data Mining Prozeß: SEMMA

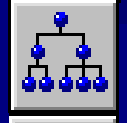
Zugriff und Aufbereitung (DW)



Sample

Explore

Modify



Model

Assess

Informationsgewinnung (BI)

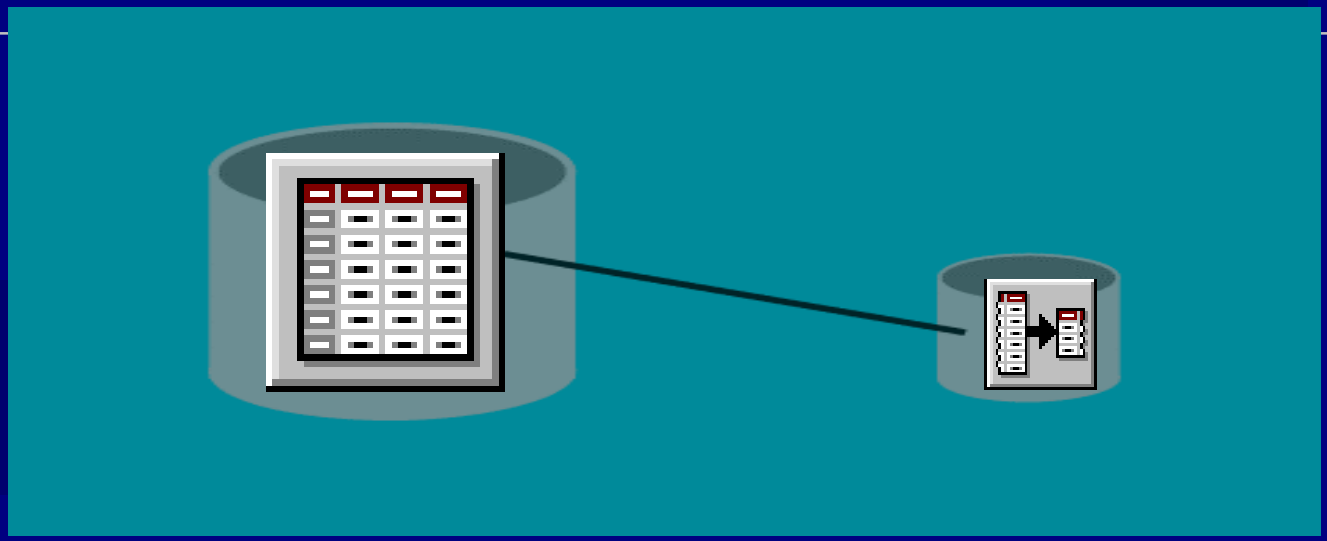
Data Warehousing Objective

Data Mining Mythen

- Data Mining ist ein Automatismus.
- Data Mining erfordert kein analytisches Wissen.
- Data Mining ermöglicht Verzicht auf Fachwissen und Datenkenntnisse.
- Data Mining Werkzeuge sind keine Statistik.

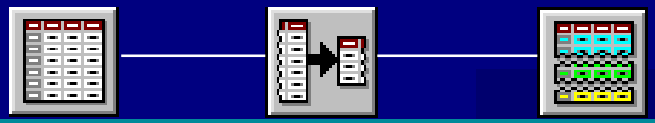
Quelle: Two Crows Corporation

Mythos



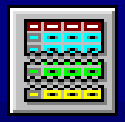
Der einzige Weg, sinnvolle Resultate zu gewinnen, ist die Nutzung aller Beobachtungen?

Sampling

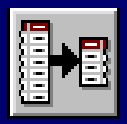


- Reduziert die Kosten der Analyse
- Erhöht die Geschwindigkeit der Analyse
- Liefert korrekte Resultate
- Bevorzugte Technik für große Files
- Empfohlen, nicht Bedingung

Sampling ist üblich



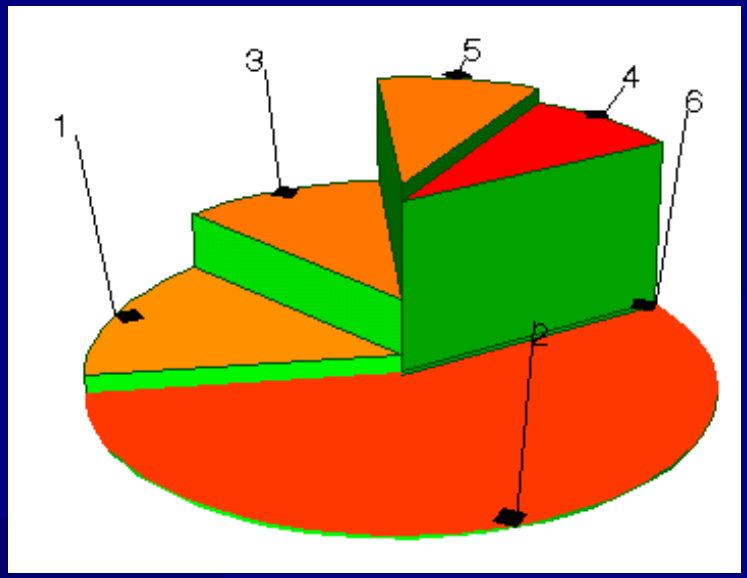
Gute Data Mining Praxis teilt die Daten in Trainings-, Test- und Validationsdateien



Modellierung seltener Ereignisse nutzt gewichtetes Sampling

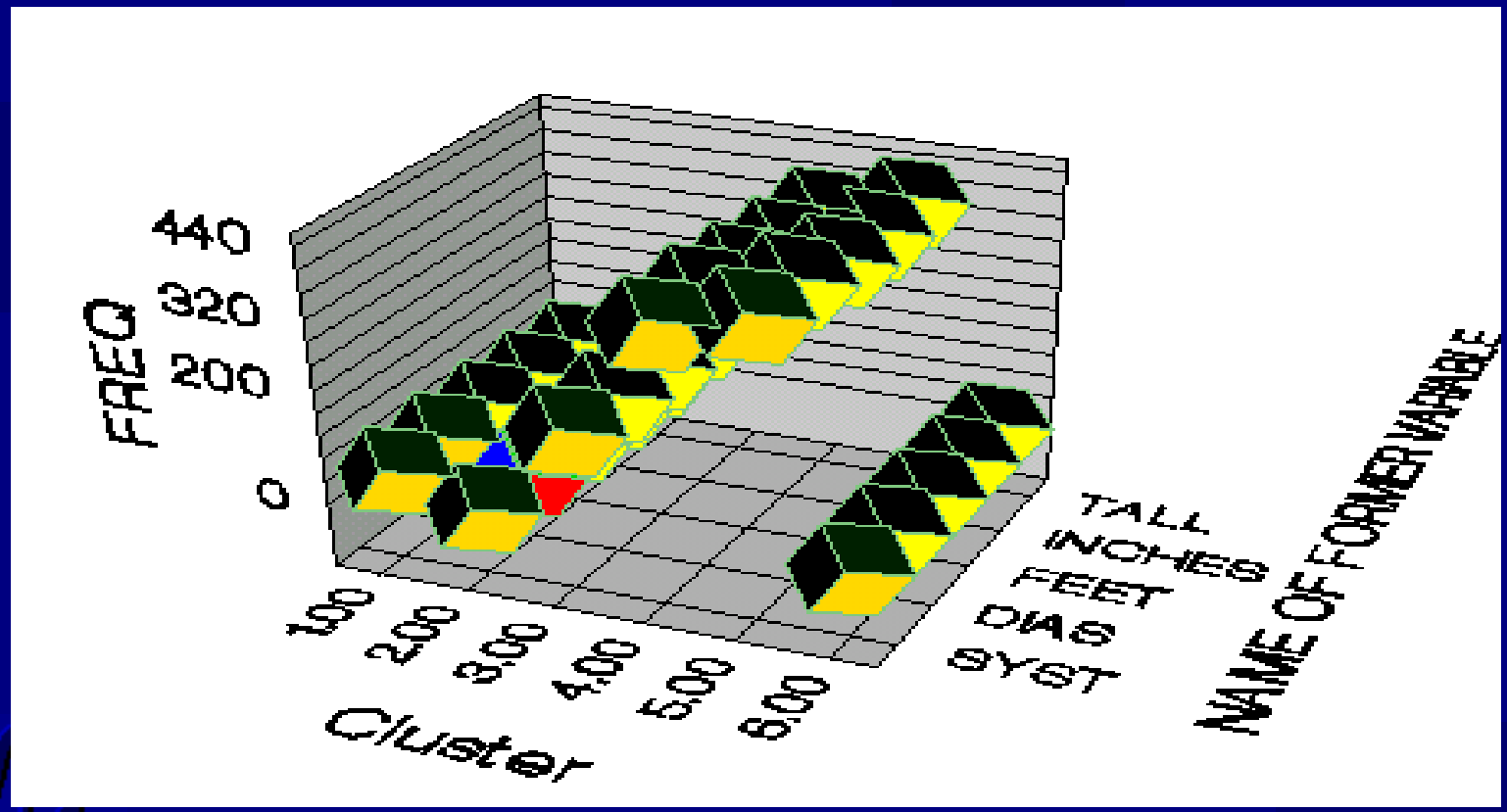
Die Datengrundlage ist in der Regel selbst eine Stichprobe aus einer Grundgesamtheit

Mythos

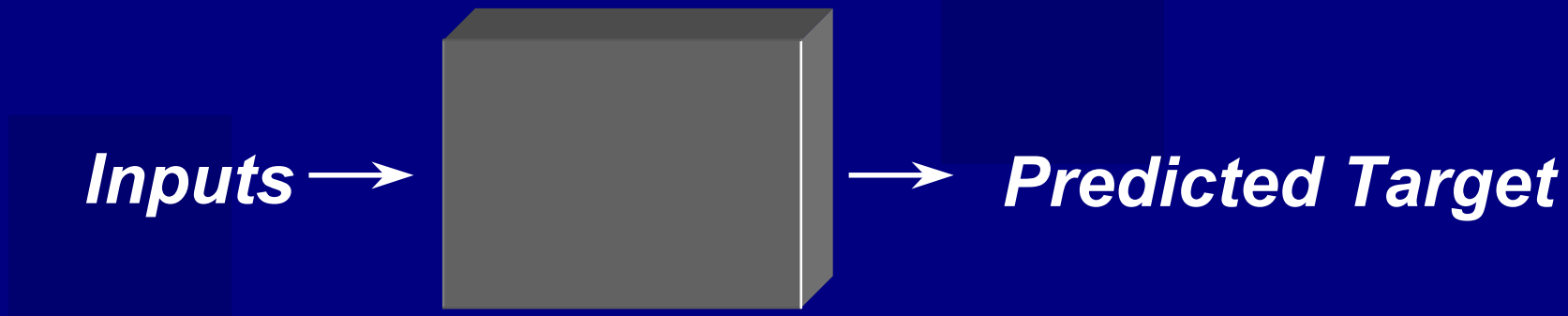


Geschäftsgrafiken sind nicht nutzbar für Data Mining?

Exploration

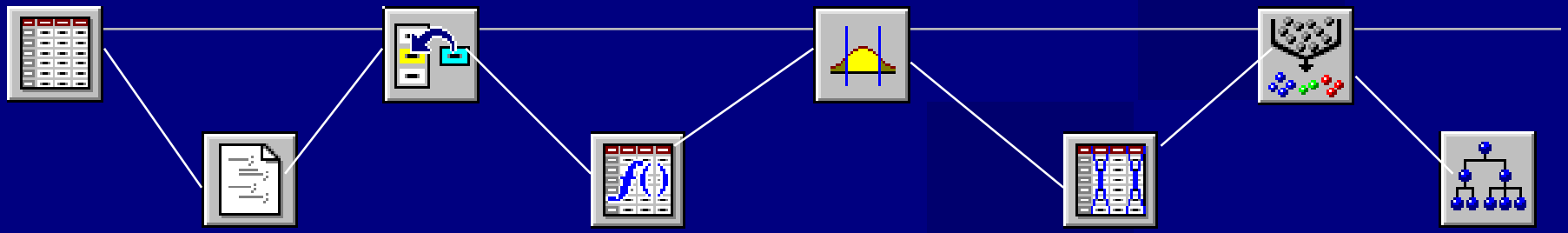


Black Box Mythos



Data Mining Software sollte beste Variablen automatisch auswählen. Warum soll ich diese Arbeit tun?

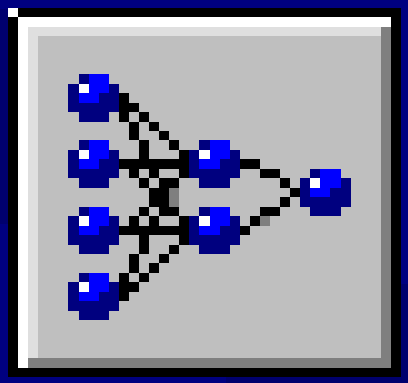
Modifikation



- Data Filtering
  - Variablen-Selektion
  - Entscheidungsbäume
- Fine-Tuning
  - Transformationen
  - Imputation

Mythos

Data Mining gleich Neural Networks?

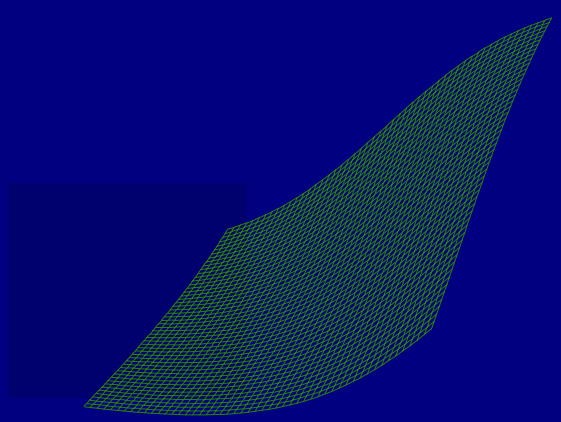


=

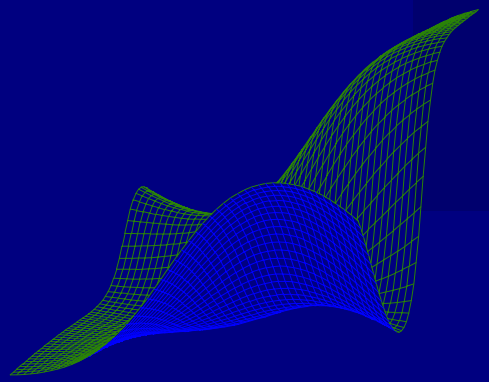
?

…und sie sind sehr schnell - korrekt?

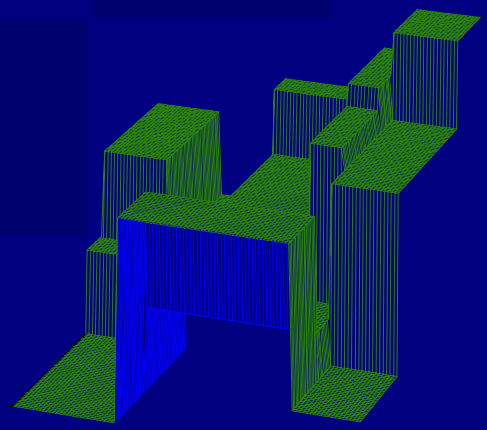
Modellierungsmethoden



Logistische Regression



Neuronale Netze



Entscheidungs-bäume

Modellierung

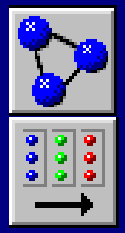
Vorhersagemodellierung im EM:

Neuronale Netze

Regressionen

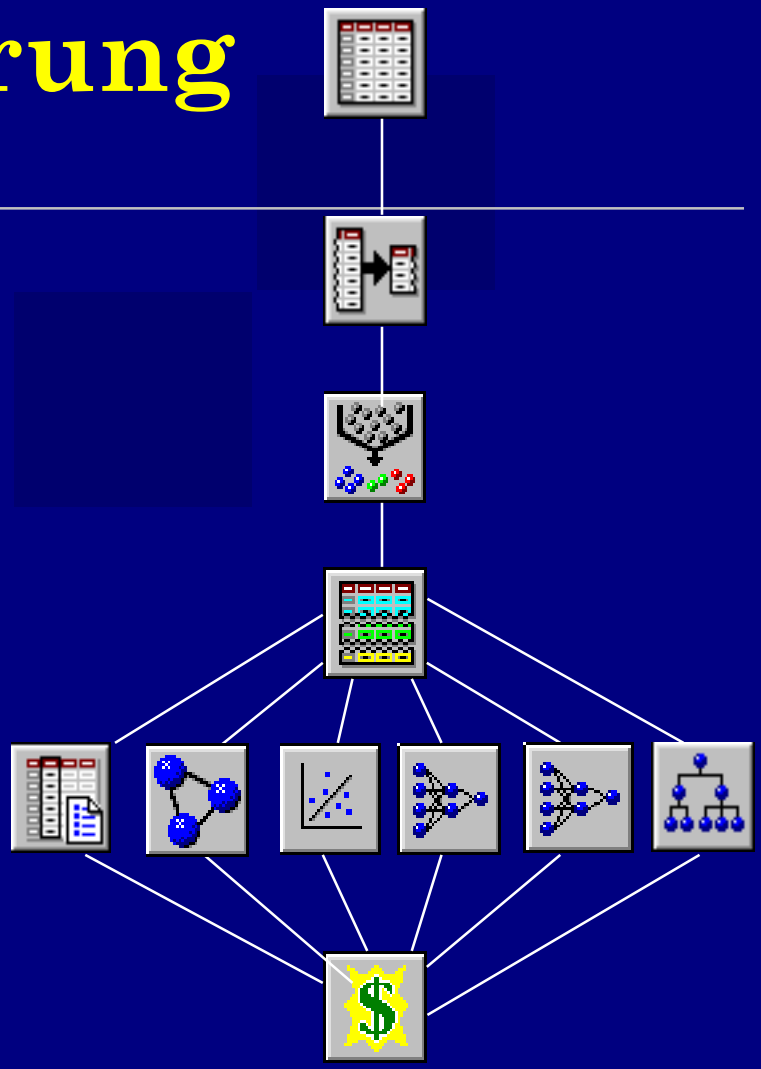
Entscheidungsbäume

Zusätzlich im EM:



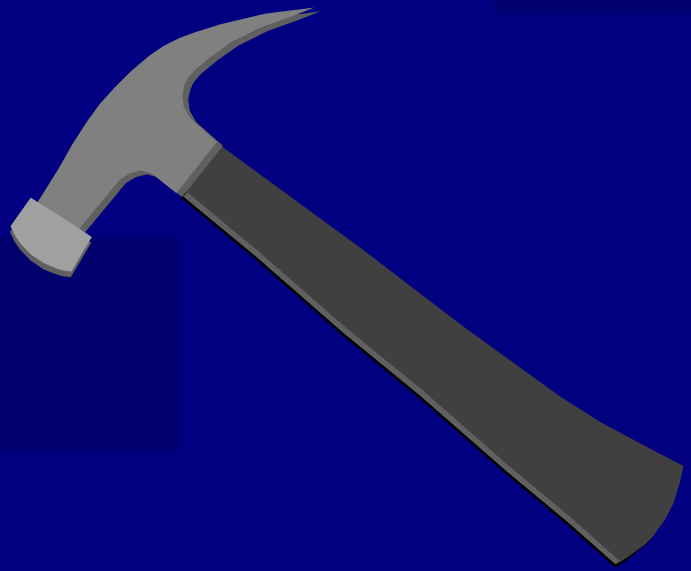
User-Defined Model

Group-Processing



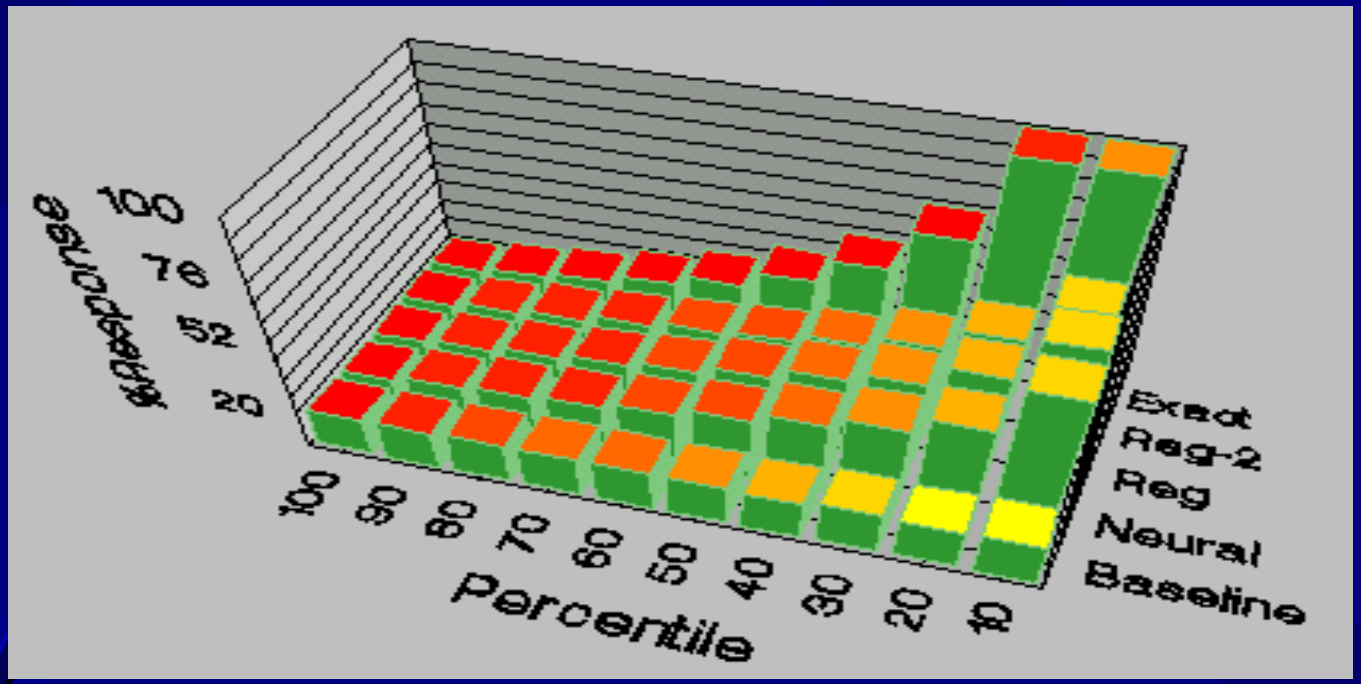
Mythos

Data Mining ist nicht iterativ?



Assess

- Güte des Modells auf den Validationsdaten?
- Welche Modellierung ist die beste?



Umsetzung der Resultate

- Scoring neuer Daten
- Darstellung der Ergebnisse im Web
- Weitergabe von Prozeßflußdiagrammen

Relations	Lift	Support(%)	Confidence(%)	Rule
1	5.08	19.70	100.00	chips
1	5.15	19.40	100.00	cigarettes
1	5.18	19.30	100.00	sausage
1	5.24	19.10	100.00	heineken
1	5.26	19.00	100.00	budweiser
1	5.46	18.30	100.00	wine_cooler
1	5.68	17.60	100.00	milk
1	7.41	13.50	100.00	salami
1	7.52	13.30	100.00	crackers
1	7.52	13.30	100.00	peanuts

SAS Enterprise Miner

- GUI für den SEMMA Prozeß
- Neue Algorithmen
- Volles Spektrum an Data Mining Techniken
- optimierte DMDB
- Code-Generator für Scoring
- Integration mit Data Warehouse Strategie und OLAP Technologie

Verbundanalyse mit dem SAS Enterprise Miner bei der Herlitz PBS AG

Heidelberg, 25./26. März 1999

Freese
CatCon/SO

Agenda

- Verbundarten und Verbundanalyse
- Durchführung der Analyse im SAS Enterprise Miner
- Auswertung der Ergebnisse

Agenda

- Verbundarten und Verbundanalyse
- Durchführung der Analyse im SAS Enterprise Miner
- Auswertung der Ergebnisse

Es wird zwischen vier Arten von Verbundwirkungen unterschieden.

- **Nachfrageverbund:**

  Identifikation von Artikeln, die in einem Einkaufsvorgang bezogen werden.

- **Bedarfsverbund:**

  Identifikation von Artikeln, die in einem komplementären Verhältnis zueinander stehen.

- **Auswahlverbund:**

  Identifikation von Artikeln, die in einem substitutionalen Verhältnis zueinander stehen.

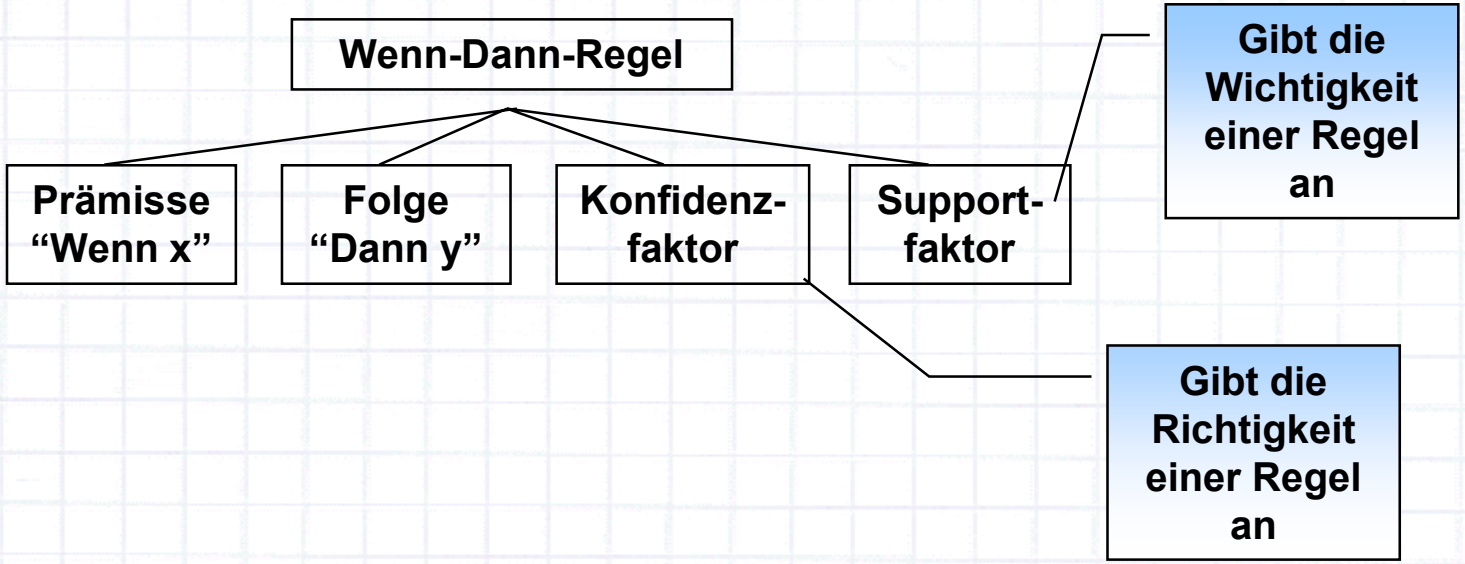
- **Akquisitionsverbund:**

  Identifikation von Artikeln, die Gegenstand einer kurzfristig angelegten absatzpolitischen Förderung sind.

Die Analyse von Verbundwirkungen ergibt einfache Regeln in Form von Wenn-Dann-Aussagen.

- **Warenkorb-/ Bondatenanalyse ⇨ Assoziationsanalyse**
  - **Fragestellung:** Welche Waren werden zusammen gekauft?
  - **Identifikation von Verbundkäufen:**

    "**Wenn** Kunden Brot und Butter kaufen, **dann** nehmen sie mit einer Wahrscheinlichkeit von 90% auch Marmelade mit."



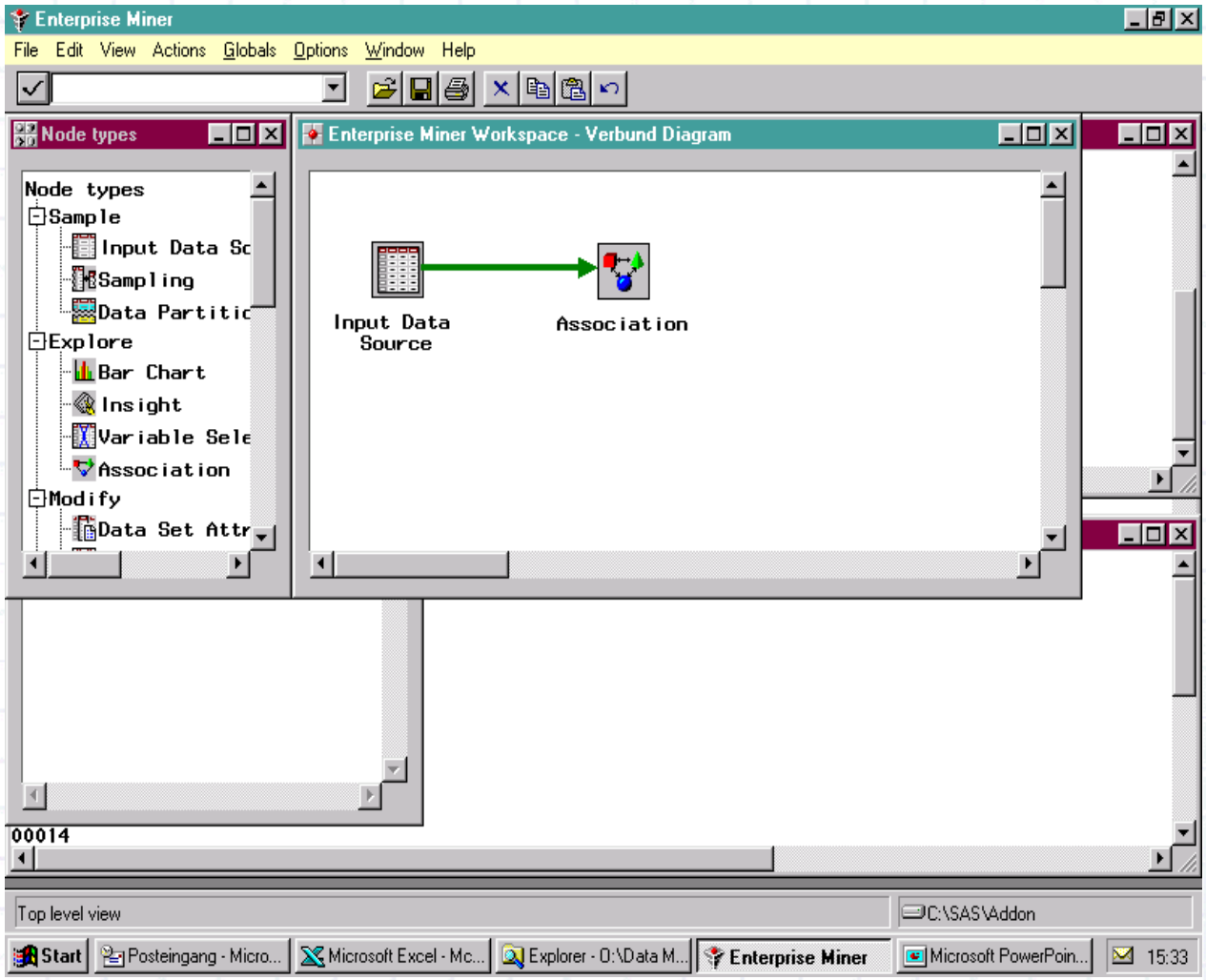
Agenda

- Verbundarten und Verbundanalyse
- Durchführung der Analyse im SAS Enterprise Miner
- Auswertung der Ergebnisse

Das Vorgehen bei der Verbundanalyse mit dem SAS Enterprise Miner ist durch folgende Punkte gekennzeichnet.

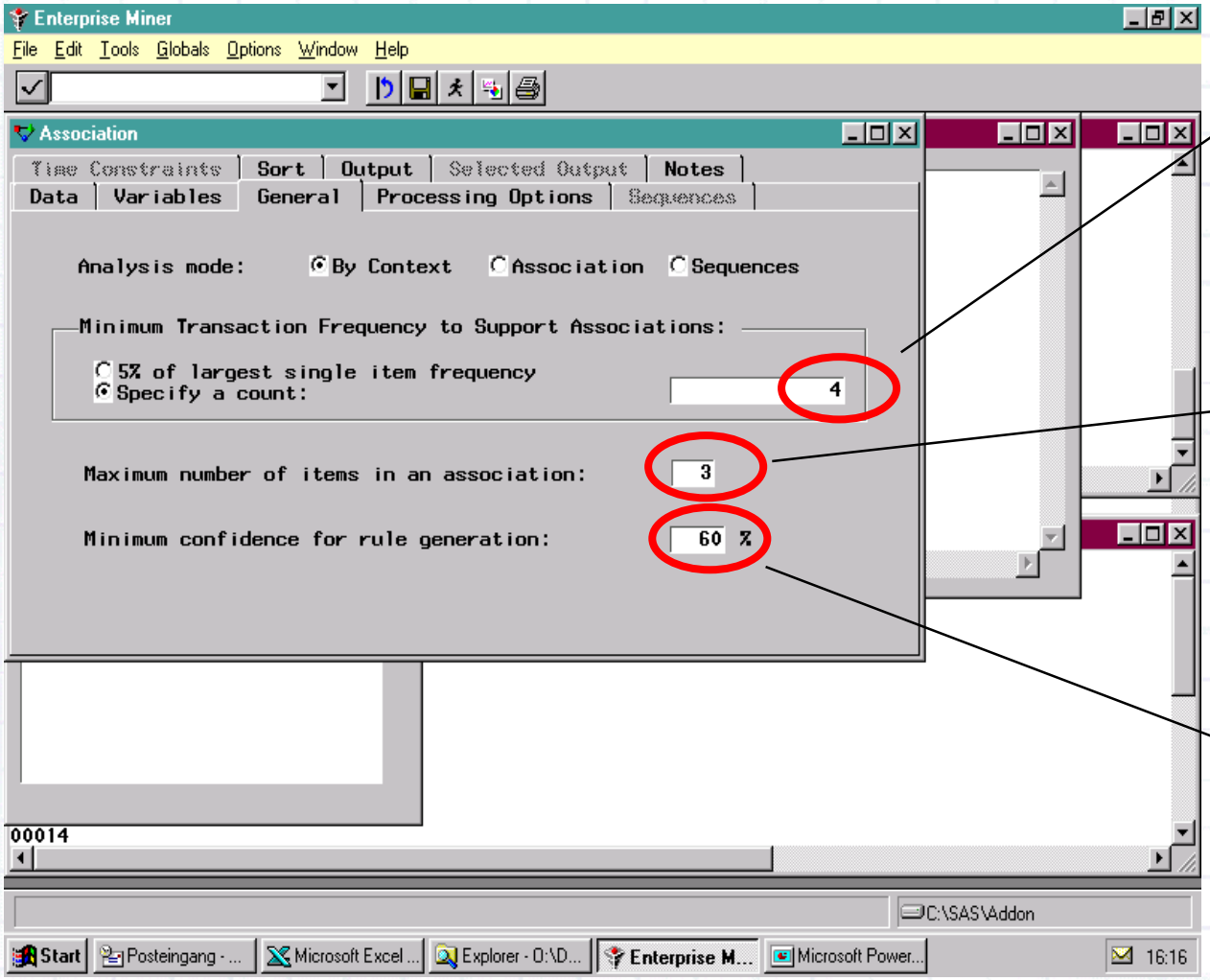
- Einlesen und Bearbeitung des Rohdatensatzes mittels SAS Programmen
- Erstellen eines Projektes und Diagramms
- Erstellen des „Process Flow Diagram“ zur Durchführung der Assoziationsregeln
- Einlesen des relevanten Datensatzes mit Hilfe des „Input Data Source“-Knotens
- Festlegen der Bedingungen für die Generierung der Regeln im „Association“-Knoten
- Durchführung der Analyse
- Visualisierung und Export der Ergebnisse zur weiteren Bearbeitung

Das Process Flow Diagram zur Durchführung von Assoziationsregeln besteht aus zwei Knoten: „Input Data Source“ und „Association“



Category Consulting

Es sind Bedingungen einzugeben, die die Regeln erfüllen sollen.



Support-Faktor

Anzahl der Artikel in der Regel

Konfidenz-Faktor

Agenda

- Verbundarten und Verbundanalyse
- Durchführung der Analyse im SAS Enterprise Miner
- Auswertung der Ergebnisse

Category Consulting

Bei der Regelgenerierung ergeben sich zu viele Regeln. Es müssen geeignete Filter eingesetzt werden, um aussagekräftige Regeln zu erhalten.

Anzahl verschiedener Artikel: 3.152
Anzahl Transaktionen (Bons): 44.704
Anzahl Artikel je Transaktion: 2,05
Untersuchungszeitraum: 12.02. - 31.08.1998

Support %	Confidence %				
	von 60 bis unter 70	von 70 bis unter 80	von 80 bis unter 90	von 90 bis 100	Summe
von 0,01 bis unter 0,03	2291	1977	867	1941	**7076**
von 0,03 bis unter 0,05	162	103	71	21	**357**
von 0,05 bis unter 0,07	34	25	10	2	**71**
von 0,07 bis unter 0,09	14	11	2		**27**
von 0,09 bis unter 0,11	14	7	3		**24**
von 0,11 bis unter 0,13	33	10	1		**44**
von 0,13 bis unter 0,15	18	8	3		**29**
von 0,15 bis 0,46	33	24	11		**68**
Summe	**2599**	**2165**	**968**	**1964**	**7696**