KSFE 1. Konferenz der SAS Benutzer in Forschung und Entwicklung

SAS Institute

Data Mining

-

Marketing-Schlagwort oder ernstzunehmende Innovation?

Hans-Peter Höschel,
SAS Institute, Heidelberg

Datamining als Marketing-Schlagwort

- **Wunsch: grosse Datenmengen auswerten, Tera- und Gigabytes, mit "modernen" Methoden**
- **Gigantomanie von technischen Softwareanbietern und Fachjournalisten**
- **Seriöse Datamining Anbieter bieten**

  =>

  Optimales Aufwand-Nutzen-Verhältnis bei verschiedenen Anwendungssituationen

  - » durch abgestuft leistungsfähige Algorithmen
  - » seriöse Aufwand-Nutzen Kalkulation

Typische Datamining Fragen

- **Datamining-Werbung**
  - » "In 20% der Fälle, bei denen ein spezieller Markentoaster verkauft wurde, kauften die Kunden auch passende Küchenhandschuhe und Tischdecken."
- **Datamining-Wunsch:** **Datamining beantwortet uns auf Knopfdruck heute die Fragen, die wir morgen stellen wollten.**
- **Datamining-Realität**
  - » seit vielen Jahren funktionsfähige Teil- und Speziallösungen und nunmehr
  - » einige neuere Ansätze durch leistungsfähigere DV und spezialisierte Algorithmen

Datamining - Realität praktisch einsetzbare Software

➢ Rationaler Kern beim Datamining

» bekannte und neue Algorithmen möglichst einfach vom Endanwender bei grossen Datenmengen anwenden.

➢ Tendenzen aktueller Datamining Lösungen

» 1. Abweichungsanalyse

> “intelligente” SQL-Algorithmen = “Einfaches” Datamining:

» 2. Gruppieren ohne Zielvariable

> Visuell interaktiv & automatisch (letzteres bekannt als Clustern)

» 3. Ursache-Wirkungsanalyse mit Zielvariable

> Automatische Response Analyse

> Entscheidungsbäume & Segmentation

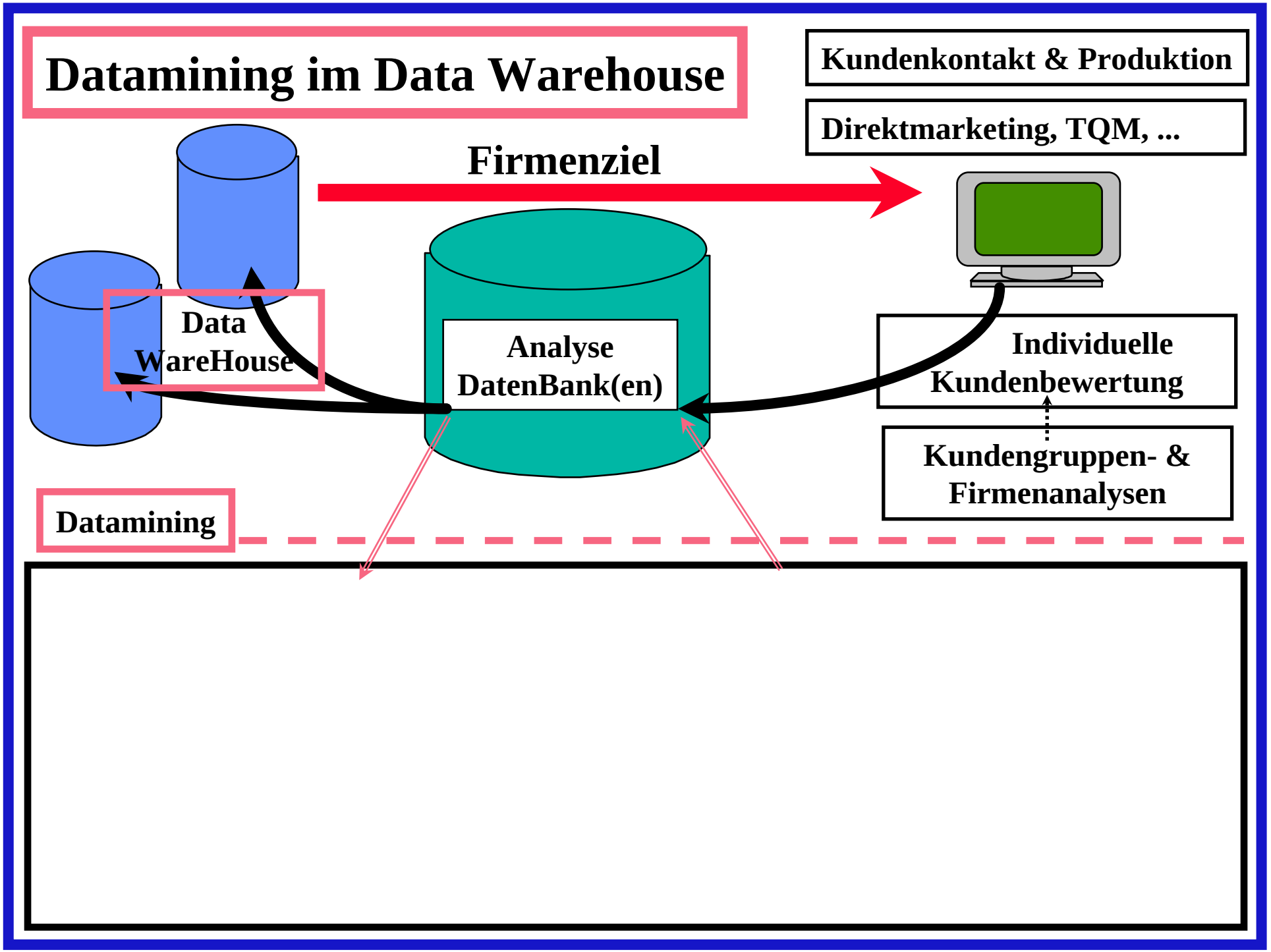
> Neuronale Netze

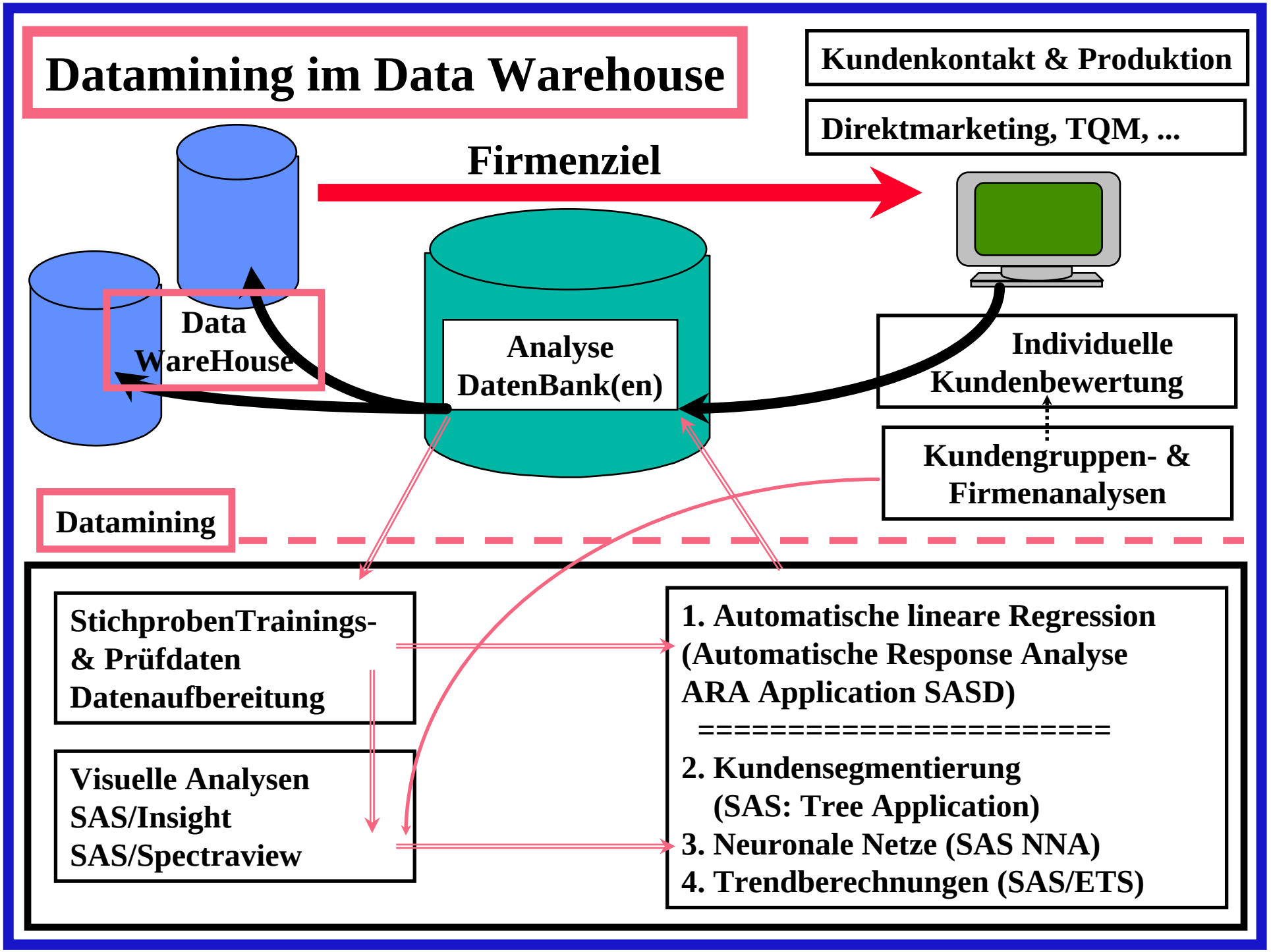
Datamining Anwendungsfelder in Marketing und Produktion

- **Anwendungsgebiete**
  - » Database(d)-Marketing
    - > Kundenklassifikation Basis: Verkaufs- u. sozio-ökonomische Daten
    - > Kaufwahrscheinlichkeiten für bestimmte Produkte
  - » Produkt-Design, TQM - Total Quality Management
    - > bedarfsgerechte Entwicklung neuer Produkte
    - > Qualitätskontrolle Abweichungsanalyse
  - » Controlling
    - > Abweichungsanalyse
- **Endanwender: Kenntnisse Datenanalyse notwendig. Aber Vorteil:**
- **Keine Annahmen über Zufallsverteilungen**

Wie funktioniert Datamining ?

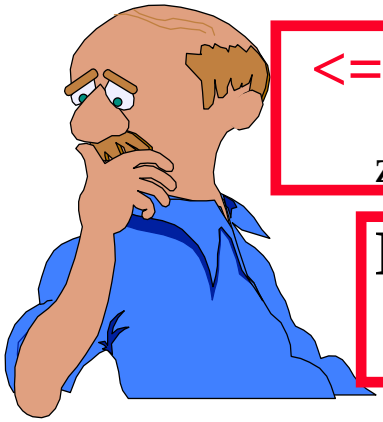
- **Vorarbeiten: Daten bereitstellen**
  - » Extraktion, Prüfung, Korrektur, Selektion, Transformation = über **80% des Gesamtaufwandes** in grossen Projekten
- **Datamining:**
  - » Abweichungsanalysen
  - » Klassifikation: Zuordnung von Daten zu Klassen
  - » Clustering: Bildung von Klassen ähnlicher Daten
  - » Entdecken von Abhängigkeiten und Trends
- **Umsetzen in Aktionen Marketing und Produktion**
- **Gesamt-Aufwand: erheblich, aber er rentiert sich.**


Datamining im Data Warehouse
Kundenkontakt & Produktion
Direktmarketing, TQM, ...
Firmenziel
Data
WareHouse
Analyse
DatenBank(en)
Individuelle
Kundenbewertung
Kundengruppen- &
Firmenanalysen
Datamining


Datamining im Data Warehouse
Kundenkontakt & Produktion
Direktmarketing, TQM, ...
Firmenziel
Data
WareHouse
Analyse
DatenBank(en)
Individuelle
Kundenbewertung
Kundengruppen- &
Firmenanalysen
Datamining
StichprobenTrainings-
& Prüfdaten
Datenaufbereitung
Visuelle Analysen
SAS/Insight
SAS/Spectraview
1. Automatische lineare Regression
(Automatische Response Analyse
ARA Application SASD)
==========
2. Kundensegmentierung
(SAS: Tree Application)
3. Neuronale Netze (SAS NNA)
4. Trendberechnungen (SAS/ETS)

Datamining Schritt 1: Konzept und Ziele

Kundenanalyse mit Punktebewertung - Beispiel Marketingaktion Testversand

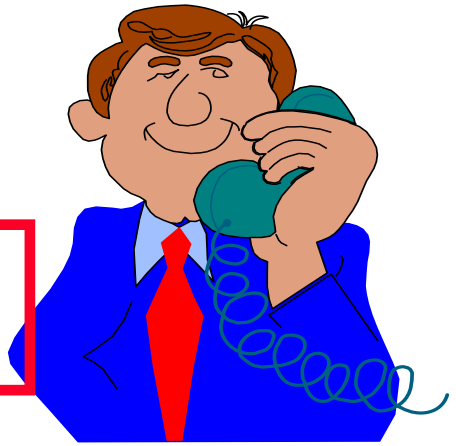


<= 0 Punkte = antwortet nicht

z.B. 49500 Kunden

Kunde antwortet: 1000 Punkte =>

z.B. 500 Kunden



Welche Faktoren ??? (beeinflussen das Kundenverhalten?)

oder besser: Aus welchen Daten kann man das Kundenverhalten vorausberechnen?

Kundendaten z.B.: Alter, Geschlecht, Beruf, Bildung, Kinderzahl, Wohnungsgrösse, Umsätze in verschieden Warengruppen, Umsätze zu bestimmten Zeiten, ...

S

Nutzen durch Datenanalyse & Datamining

- **Katalogwerbung 1 Million Kunden**
- **Werbebrief 10 DM. Antworten erbringen ca. 1000 DM Deckungsbeitrag. Die Antwortquote steige von 1% auf 2% bei Selektion der 10% "besten" Kunden.**
- **Gewinn ohne Selektion: 1 Mio Werbebriefe: Kosten = 10 Mio DM. Antworten 1% = 10.000Kunden=> *1000DM => Ertrag = 10 Mio DM. Gewinn = 0 DM**
- **mit Selektion durch Datenanalyse: 10% von 1 Mio = 100000 Werbebriefe: Kosten= 1 MioDM. Antwort 2%= 2000 => * 1000DM=> Ertrag= 2 Mio. Gewinn = 1 Mio**

Gewinn = 1 Mio DM

S

SEMMA - die SAS® Data Mining Methodik

Stichproben

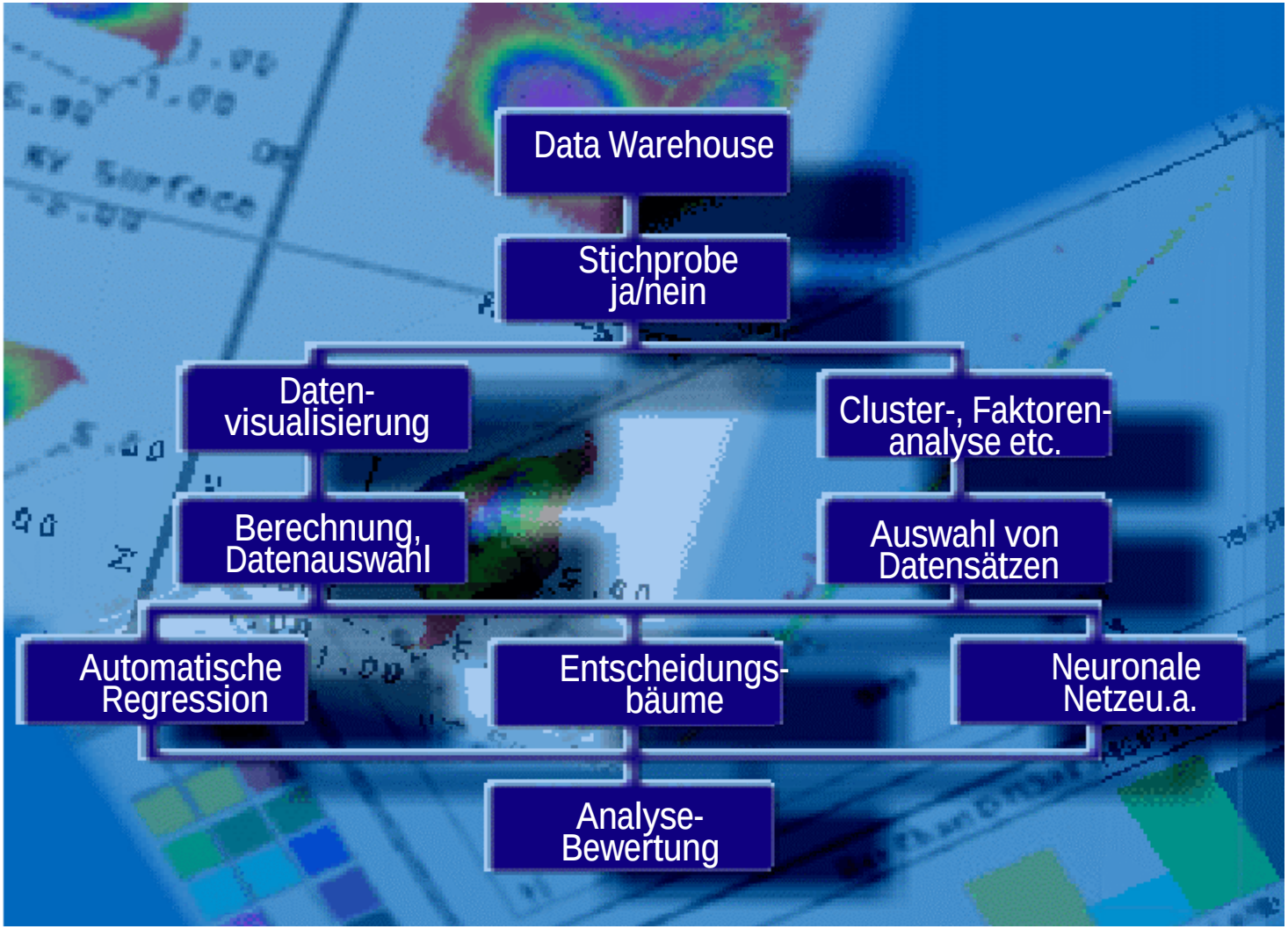
Exploration

Modifikation

Modellierung

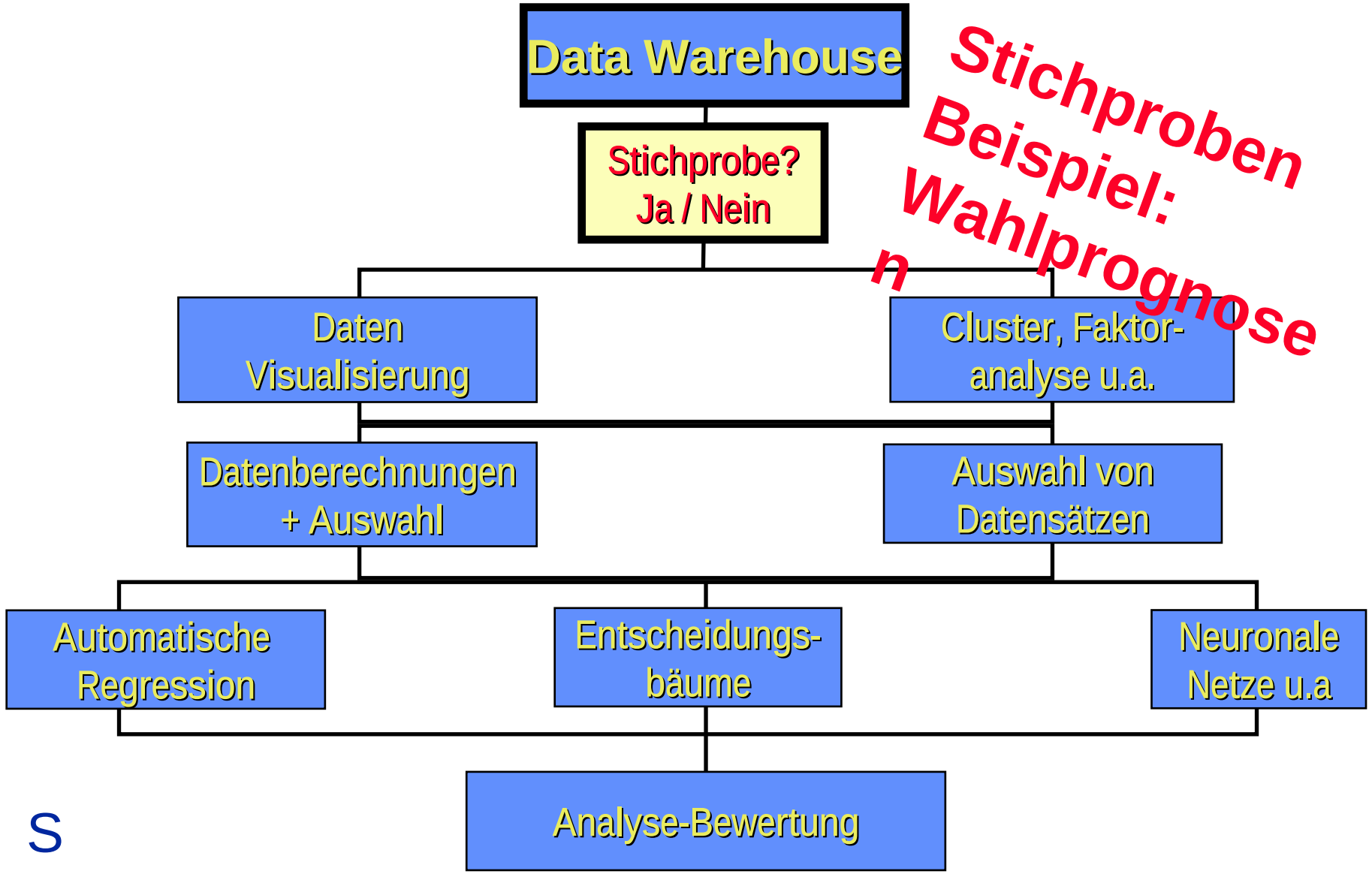
Auswertung

S



SEMMA - die Datamining Technologie 1

STICHPROBEN



Stichproben Beispiel: Wahlprognose n

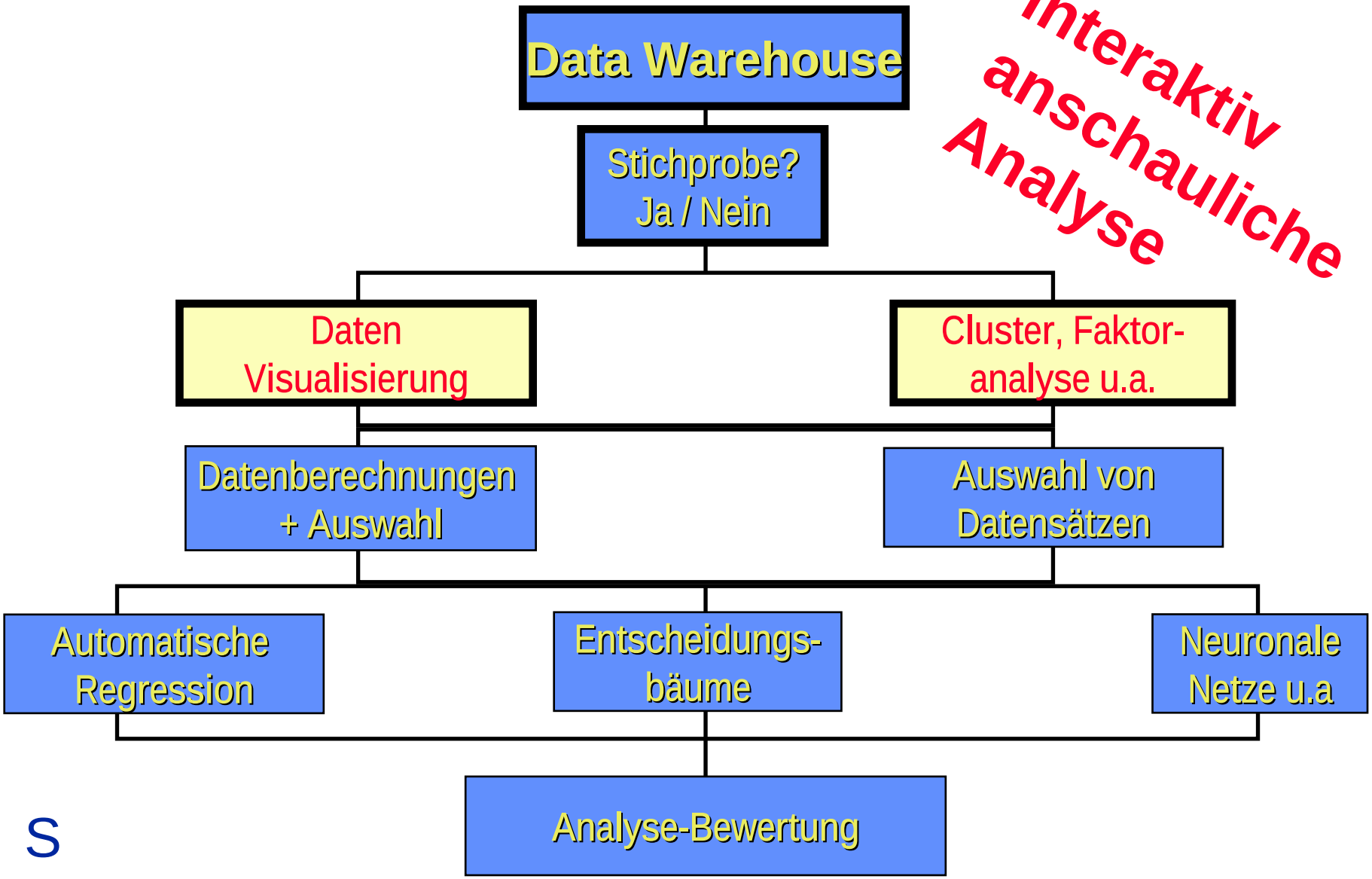
S

SAS Institute

SEMMA- die Datamining Technologie 2

EXPLORATION

interaktiv anschauliche Analyse

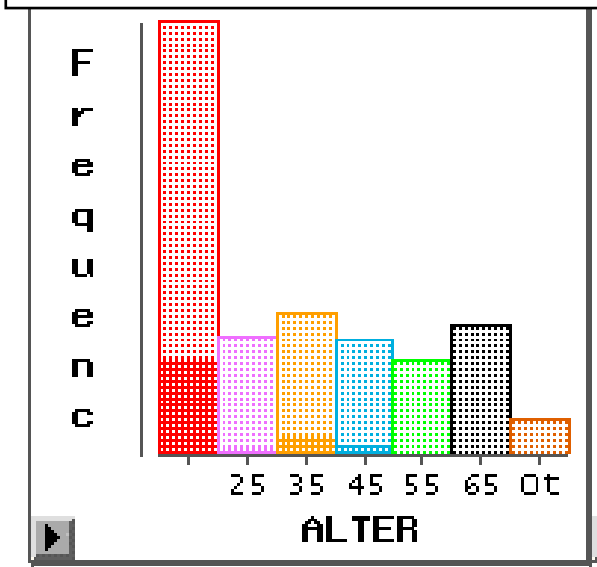


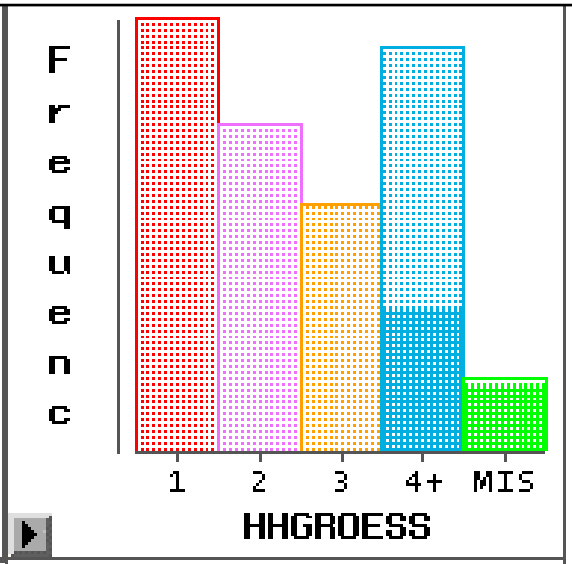
S

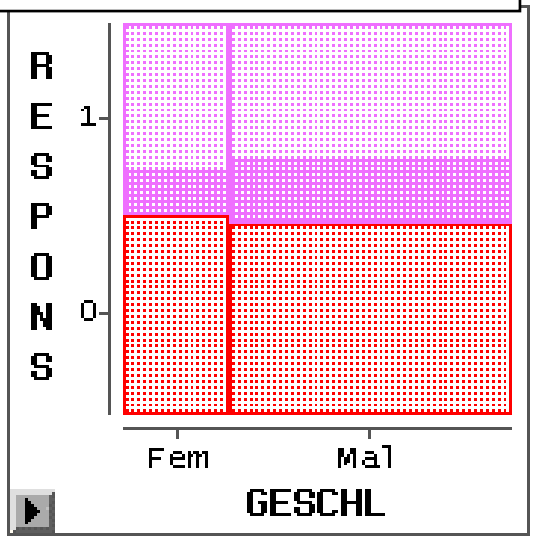
SAS Institute

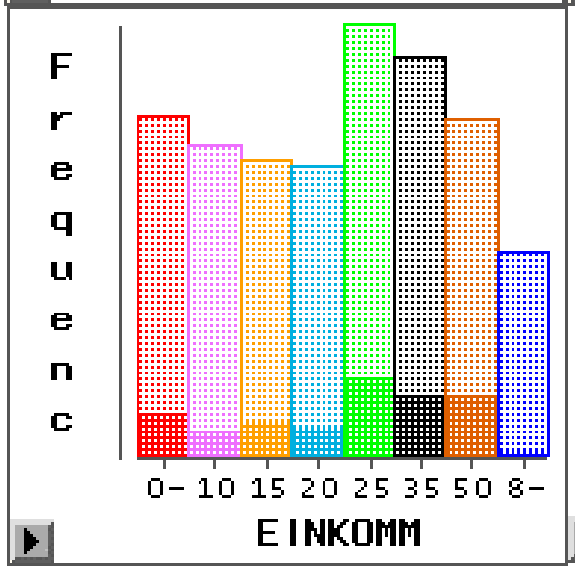
Datamining Exploration: Visuelle Analyse

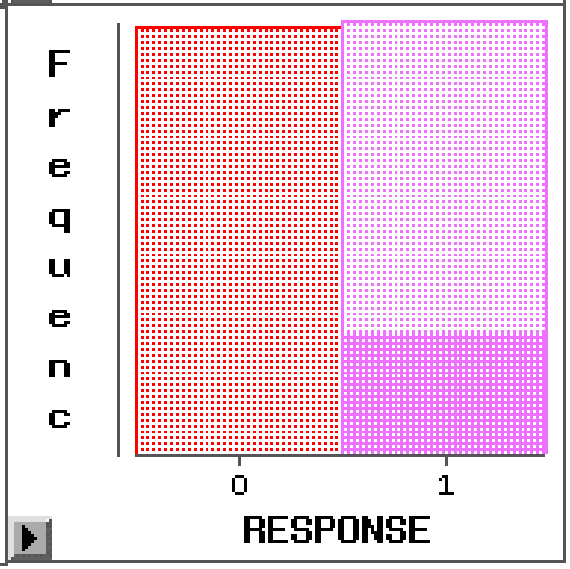
Databased Marketing Testaktion

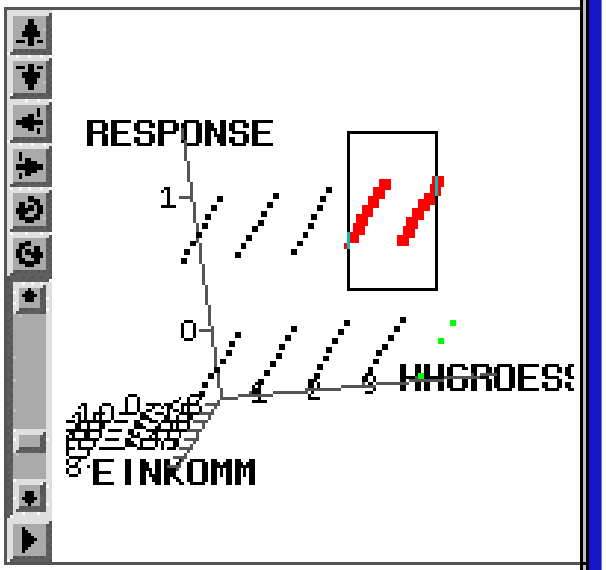






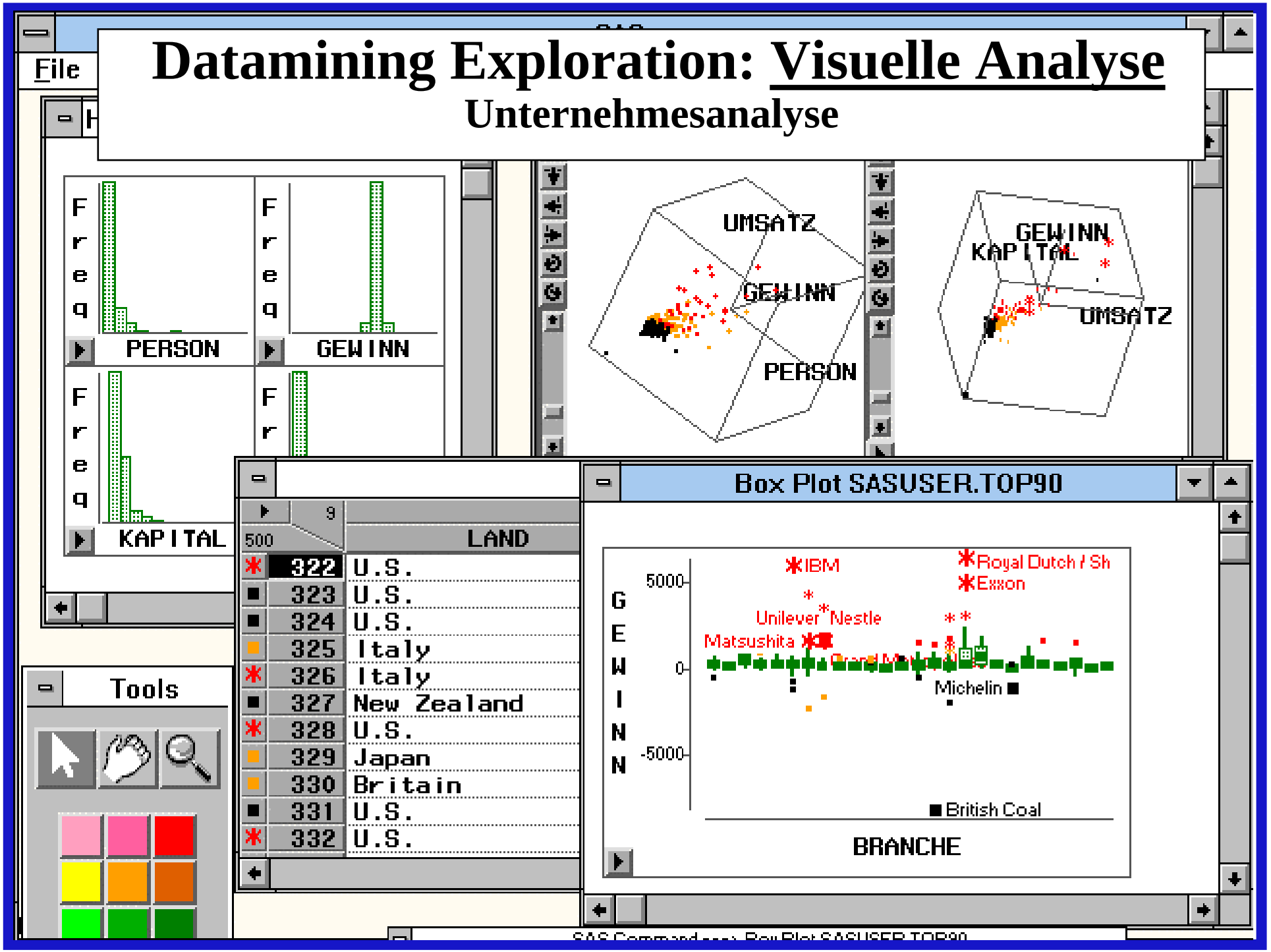






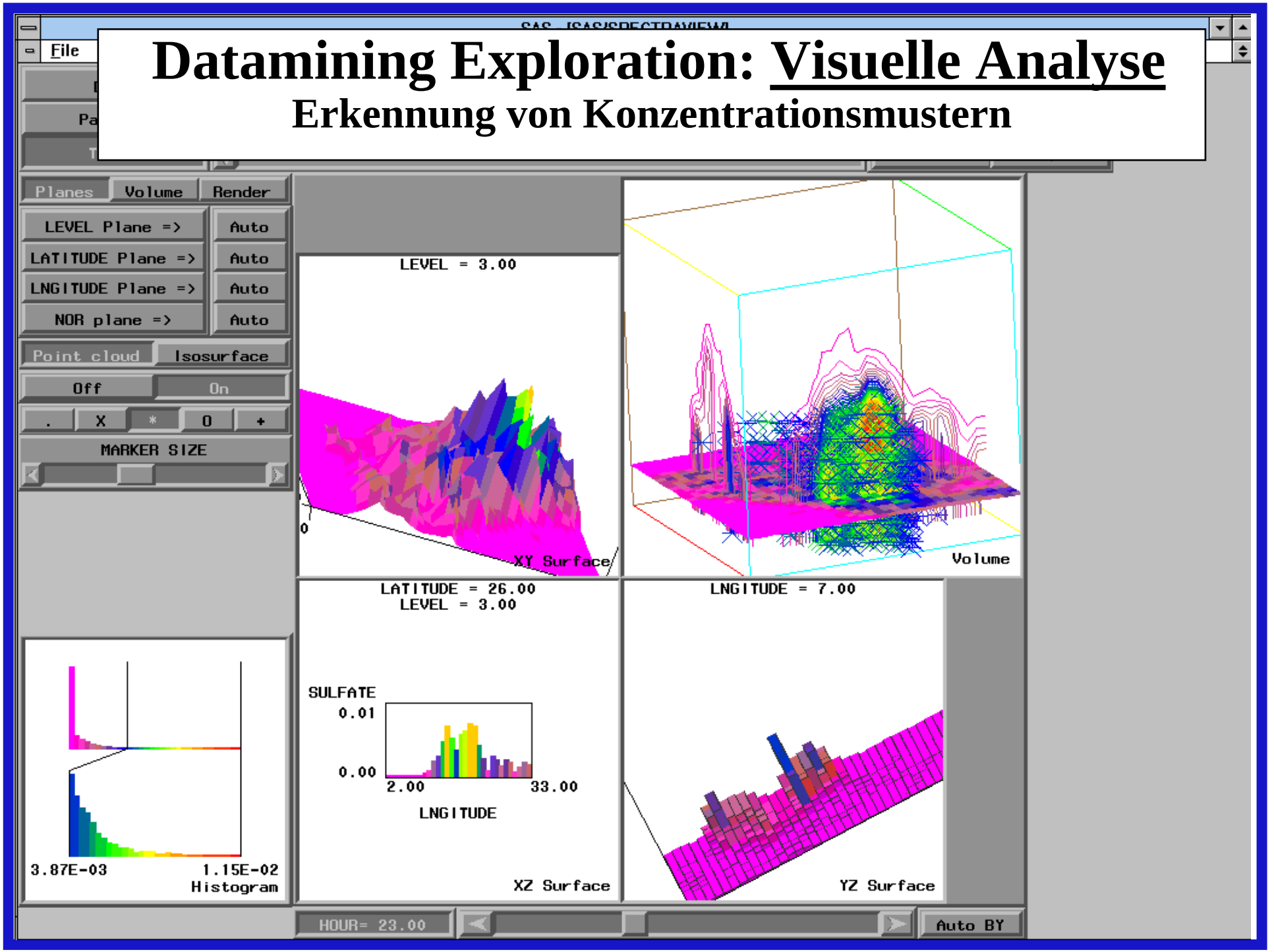
Datamining Exploration: Visuelle Analyse

Unternehmesanalyse



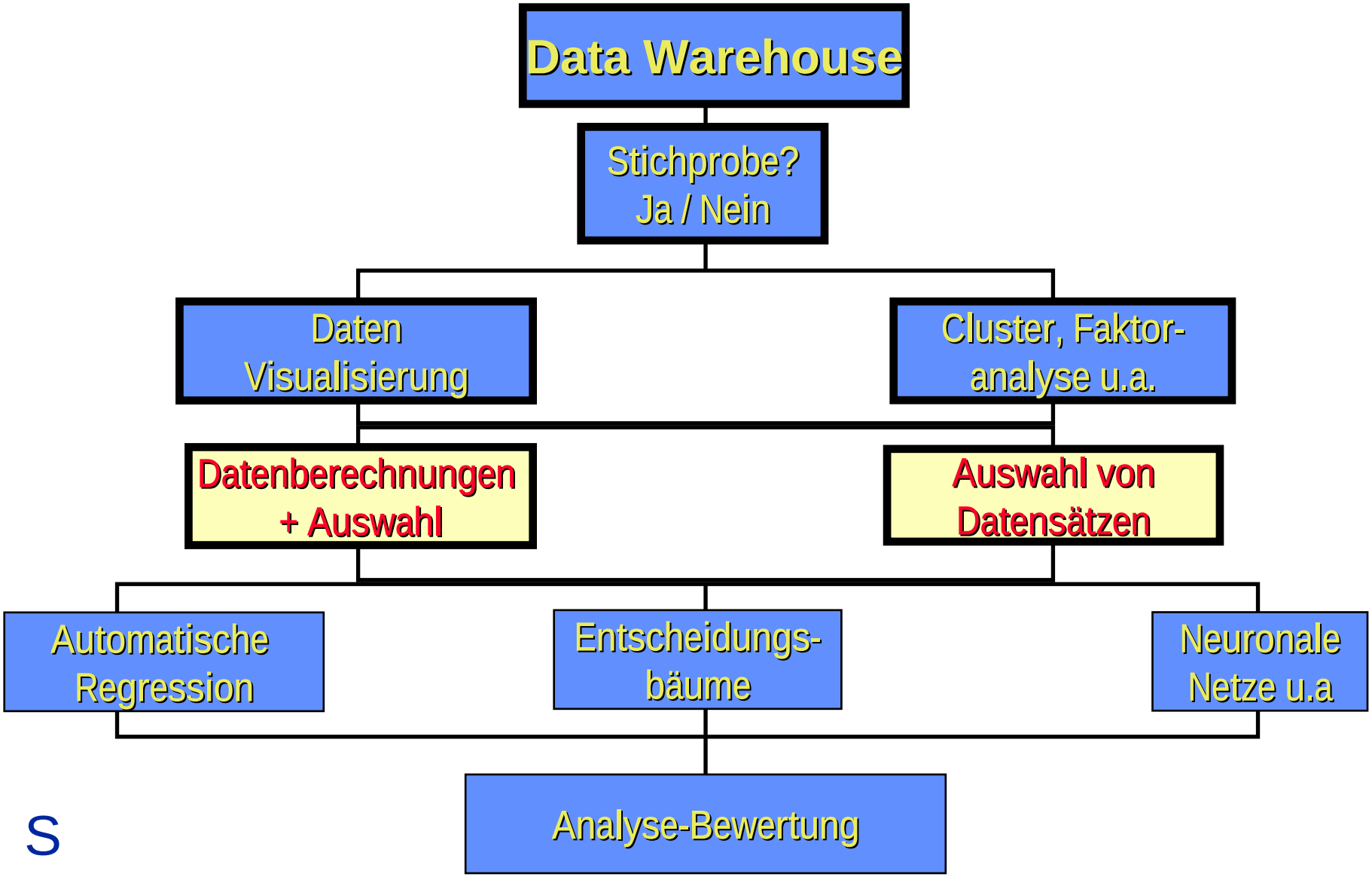
Datamining Exploration: Visuelle Analyse

Erkennung von Konzentrationsmustern



SEMMA- die Datamining Technologie 3

MANIPULATION

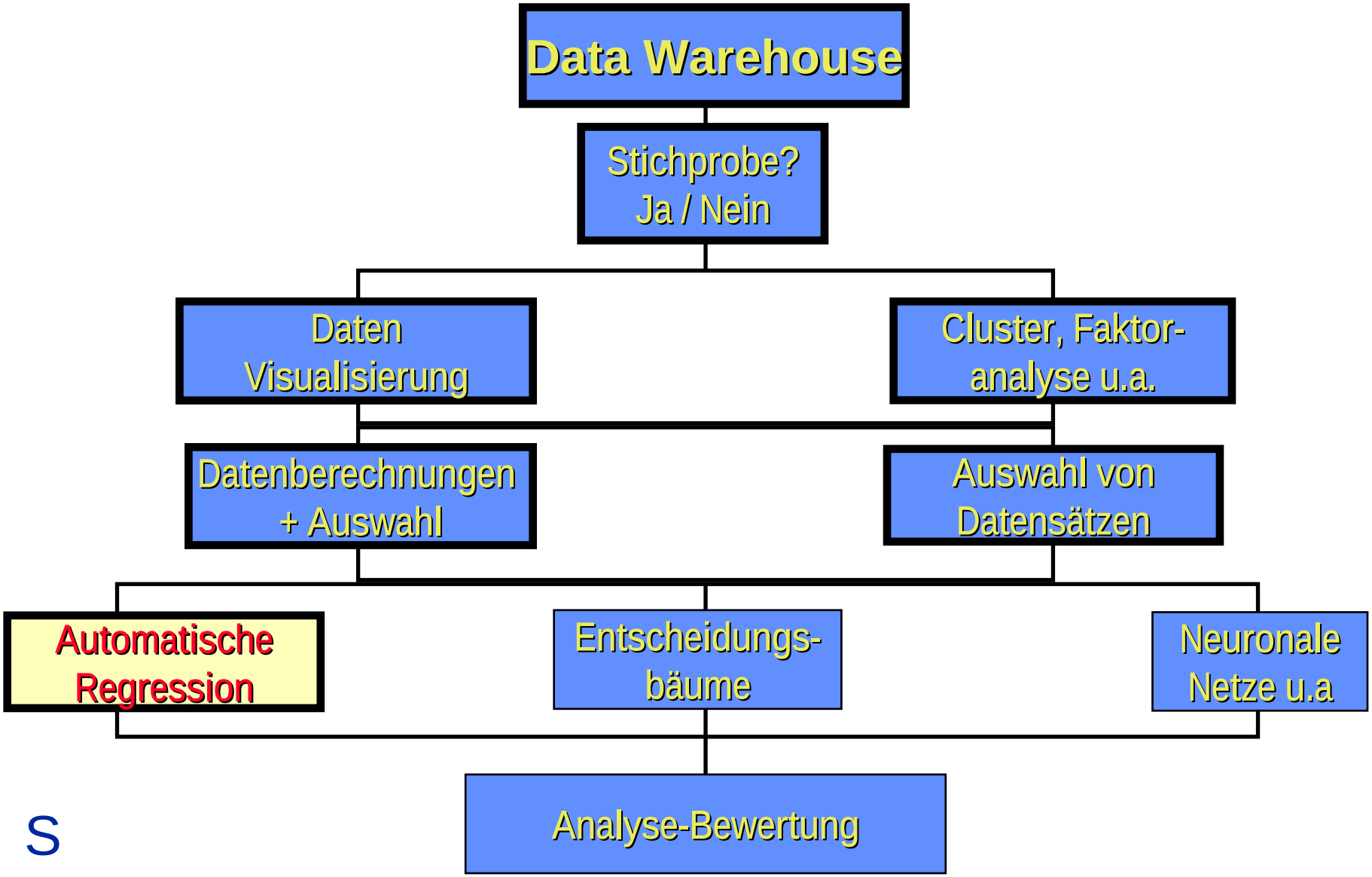


S

SAS Institute

SEMMA- die Datamining Technologie 4

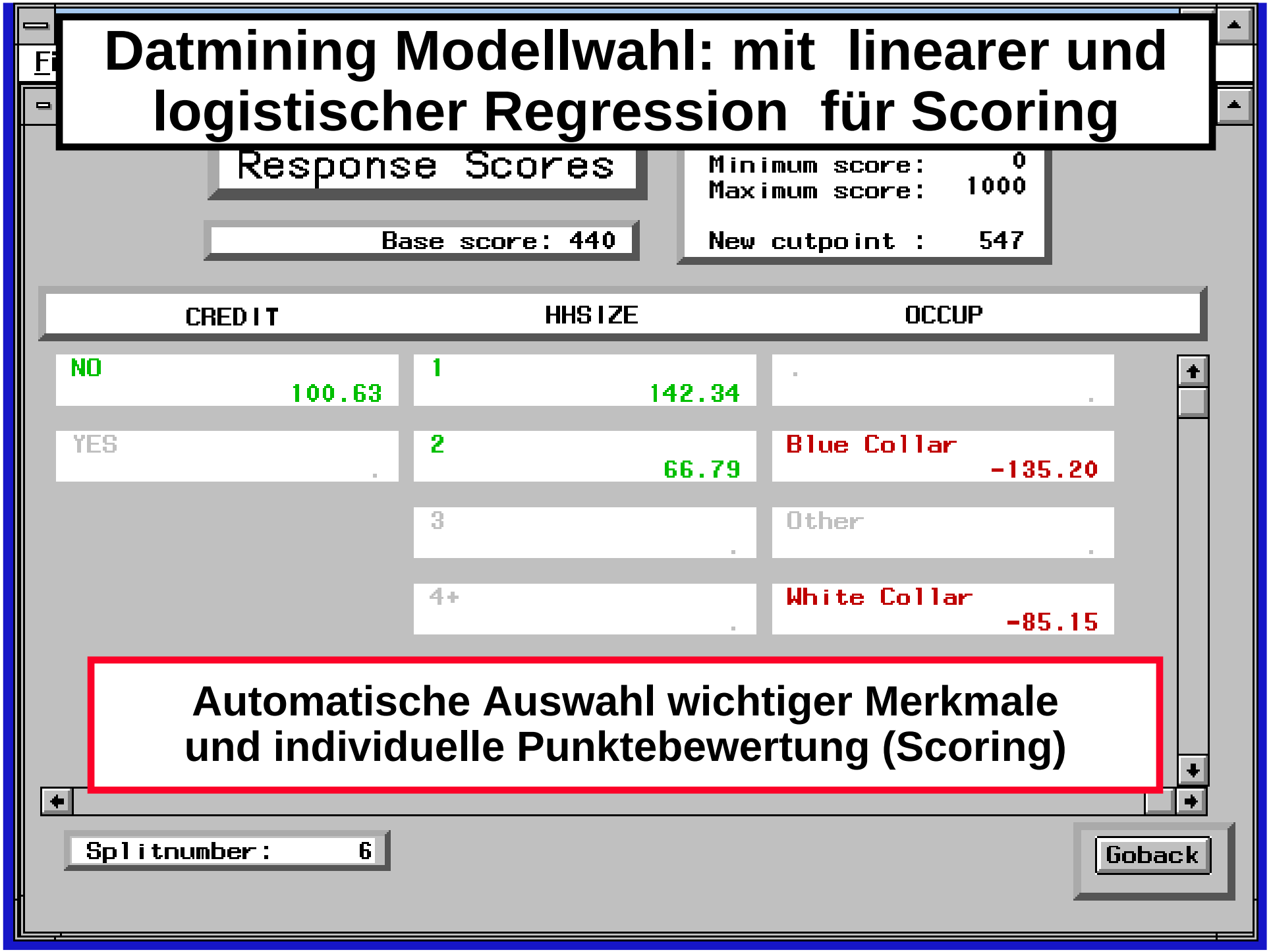
MODELLWAHL 1. Automatische Regression



S

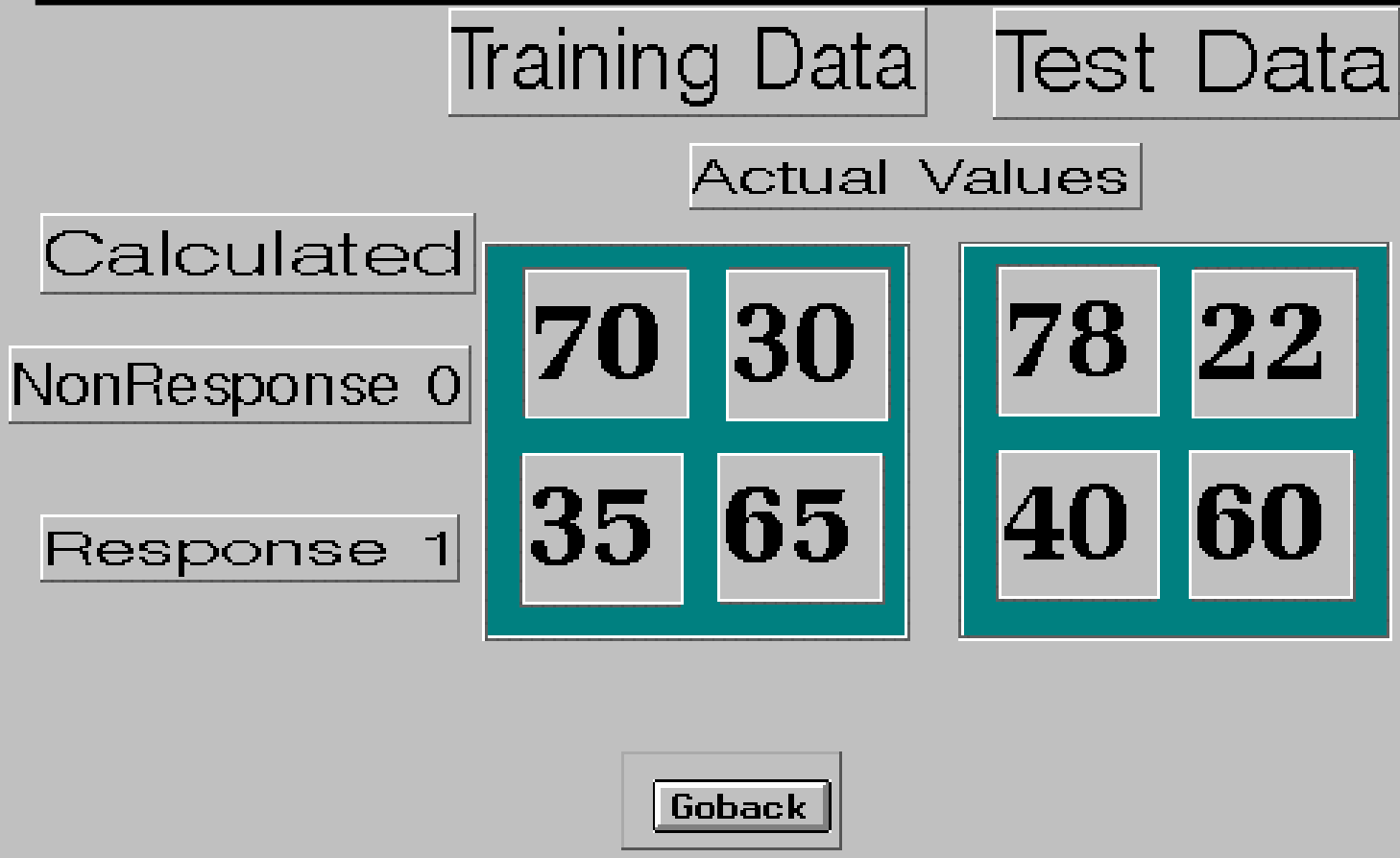
SAS Institute

Datmining Modellwahl: mit linearer und logistischer Regression für Scoring



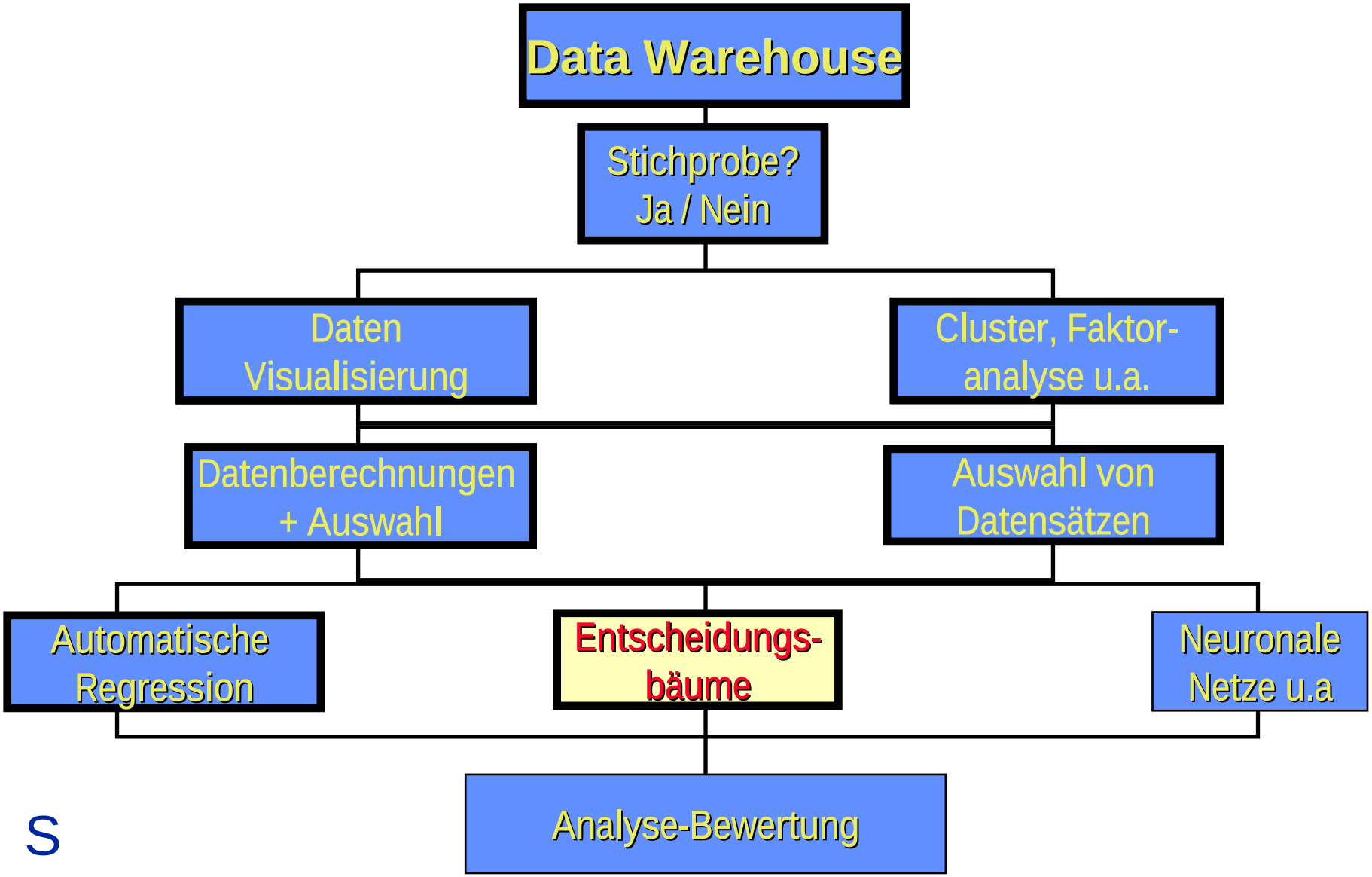
Automatische Auswahl wichtiger Merkmale und individuelle Punktebewertung (Scoring)

Gütematrix aller Datamining-Verfahren mit Zielvariable: Regression, Entscheidungsbäume, Neuronale Netze



SEMMA- die Datamining Technologie 4

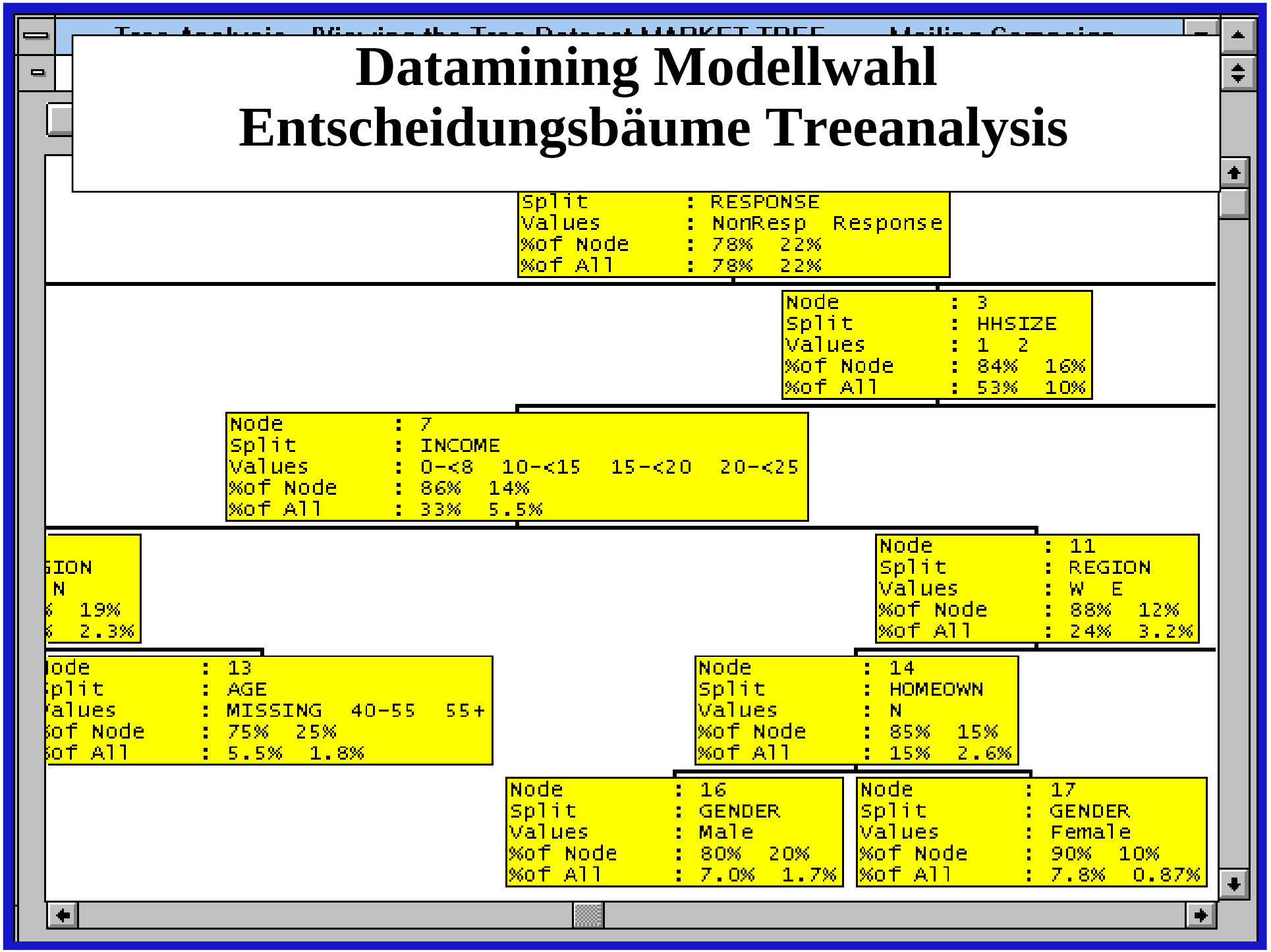
MODELLING 2. Entscheidungsbäume



S

SAS Institute

Datamining Modellwahl
Entscheidungsbäume Treeanalysis



Datamining Gewinnoptimierung durch Kundensegmentierung

Datasets

AIRPLUS.TESTMAIL
AIRPLUS.TESTSAMP
MARKET.MAILSAMP
MARKET.MAILSHOT

Dependent Variable

RESPONSE

Node	Cum Total (Popn)	Cum %of Total (Popn)	Index (Cum
11	85	1.21%	454
12	157	2.23%	419
6	490	6.95%	361
19	898	12.73%	295
13	1005	14.25%	287
9	1511	21.43%	249
	1672		
	1989	28.20%	218
	2207	31.30%	208
	2443	34.64%	197
	2983	42.30%	178
	3197	45.33%	171
	3442	48.81%	164

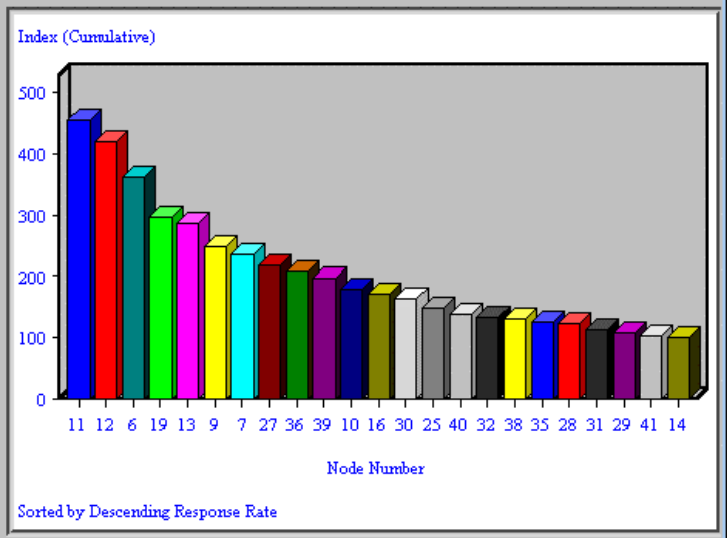
Produce Graph

Tree Analysis

File Globals Options Window Help

Tree Analysis for AIRPLUS.TESTMAIL

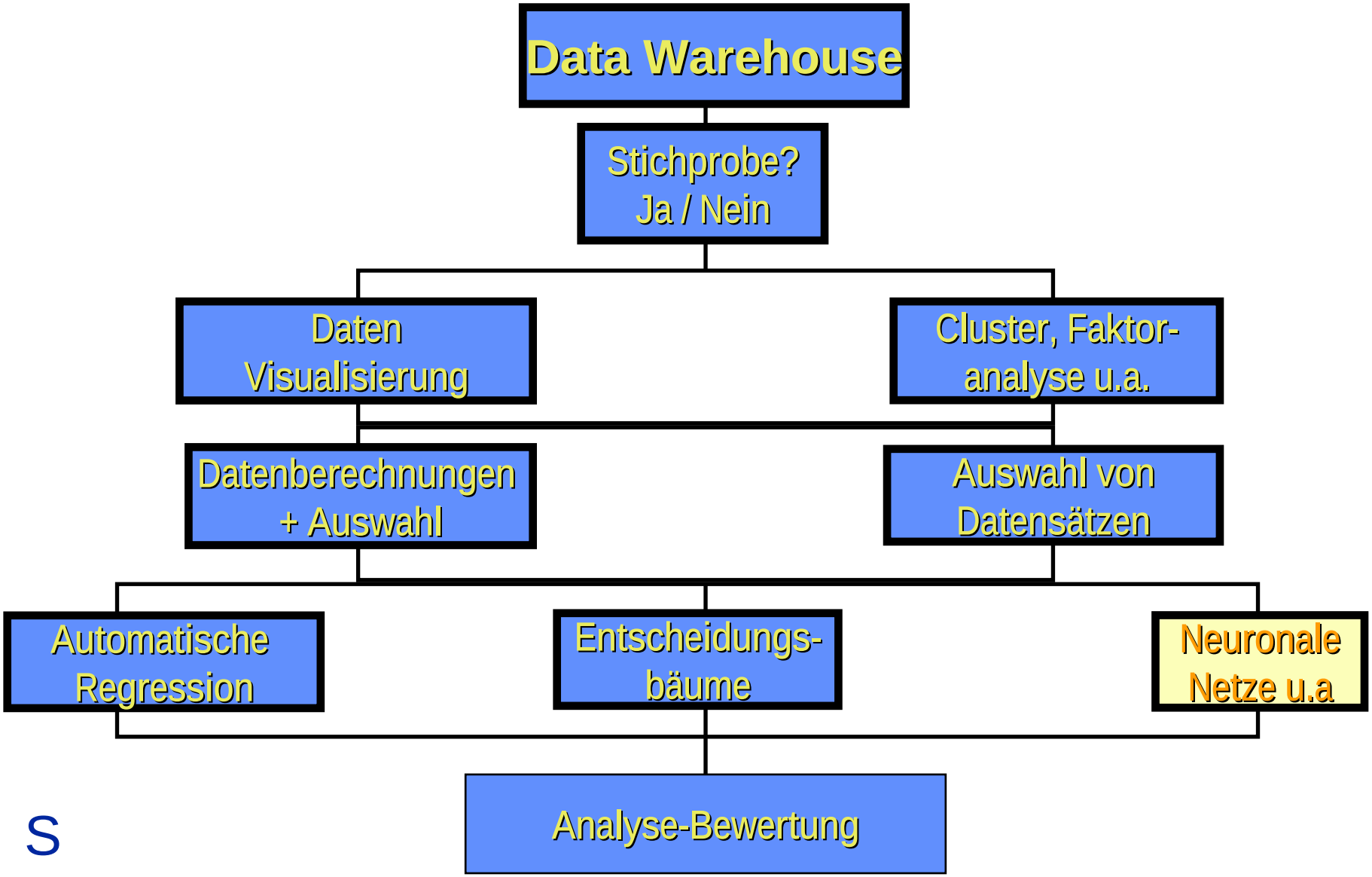
Analysis Variable
Node Count
Node Response Rat
Cum Total (Popn)
Cum %of Total (Pc
Index (Cumulative
Index (Node)
Node Number



Goback

SEMMA- die Datamining Technologie 4

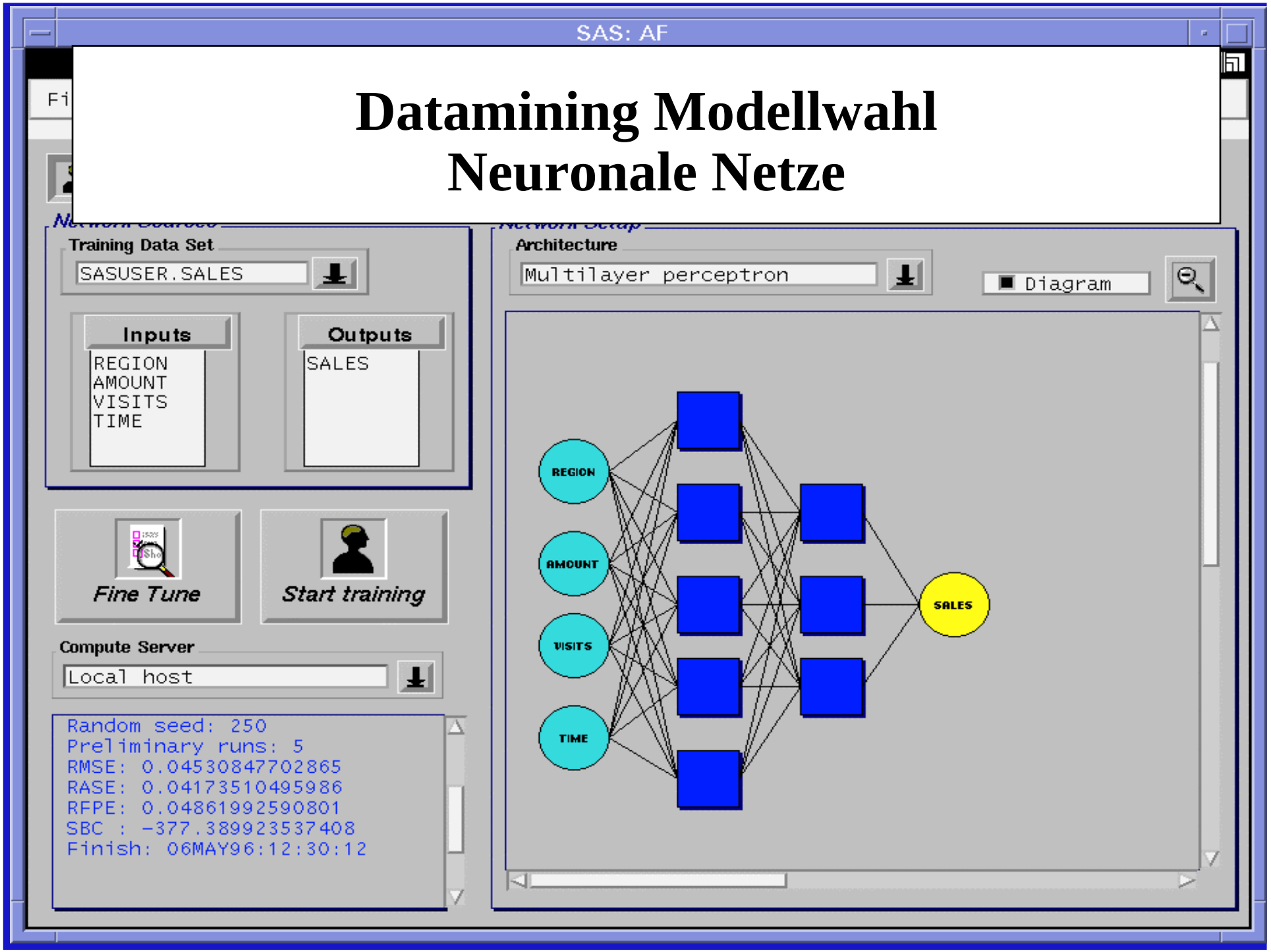
MODELLING: 3. Neuronale Netze



S

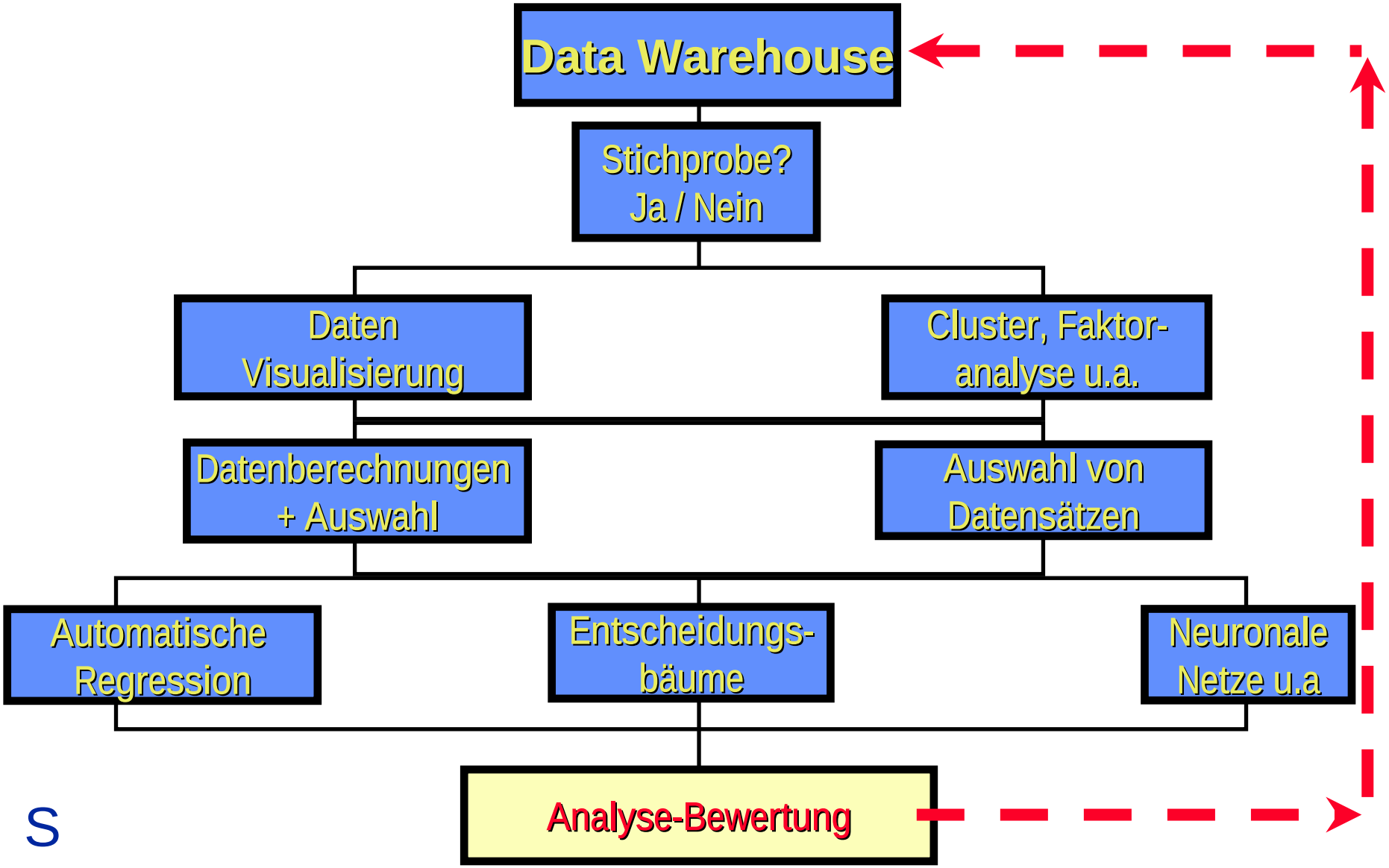
SAS Institute

Datamining Modellwahl
Neuronale Netze



SEMMA- die Datamining Technologie 5

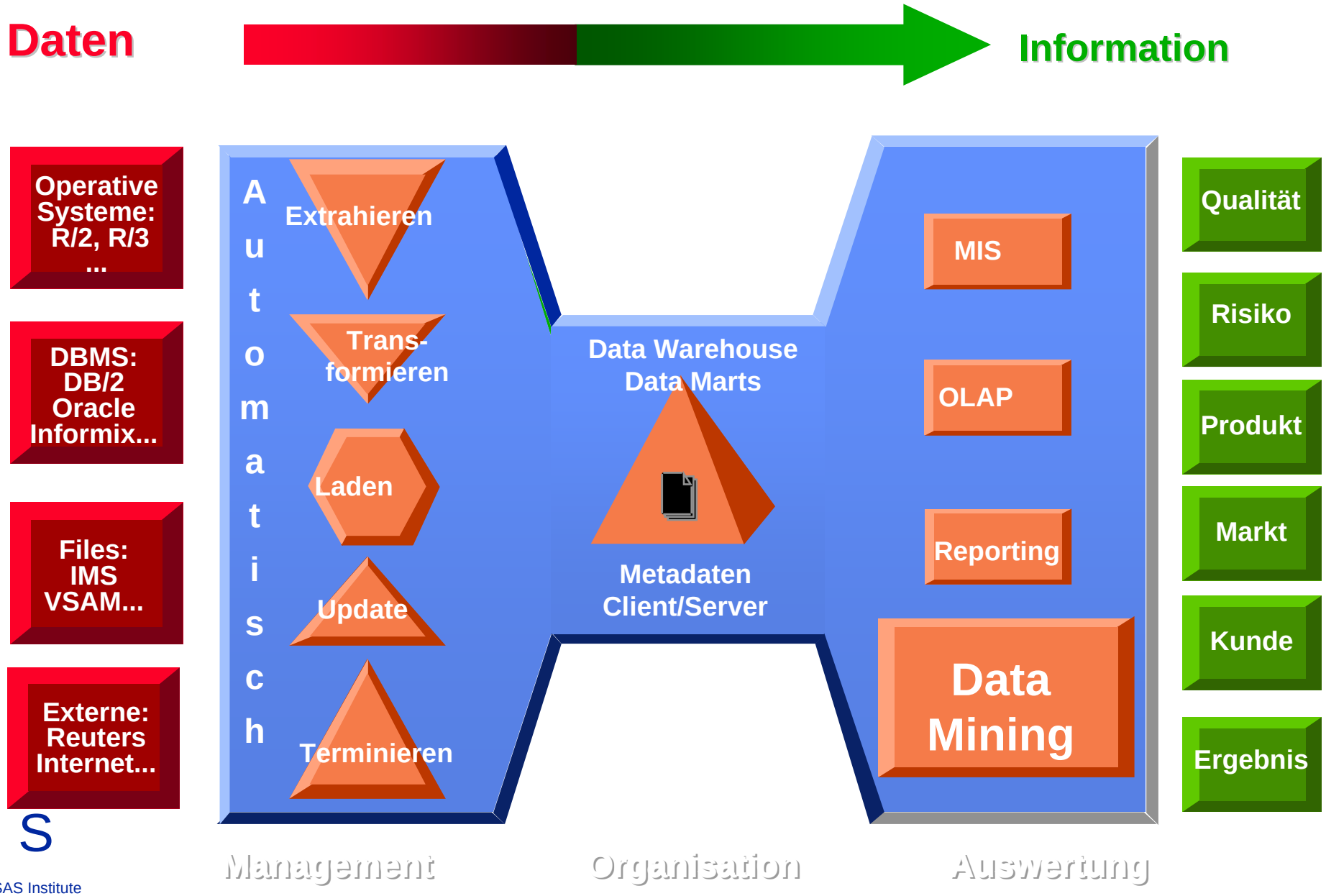
ANALYSE-BEWERTUNG



S

SAS Institute

Datamining im offenen Data Warehouse



SAS Institute