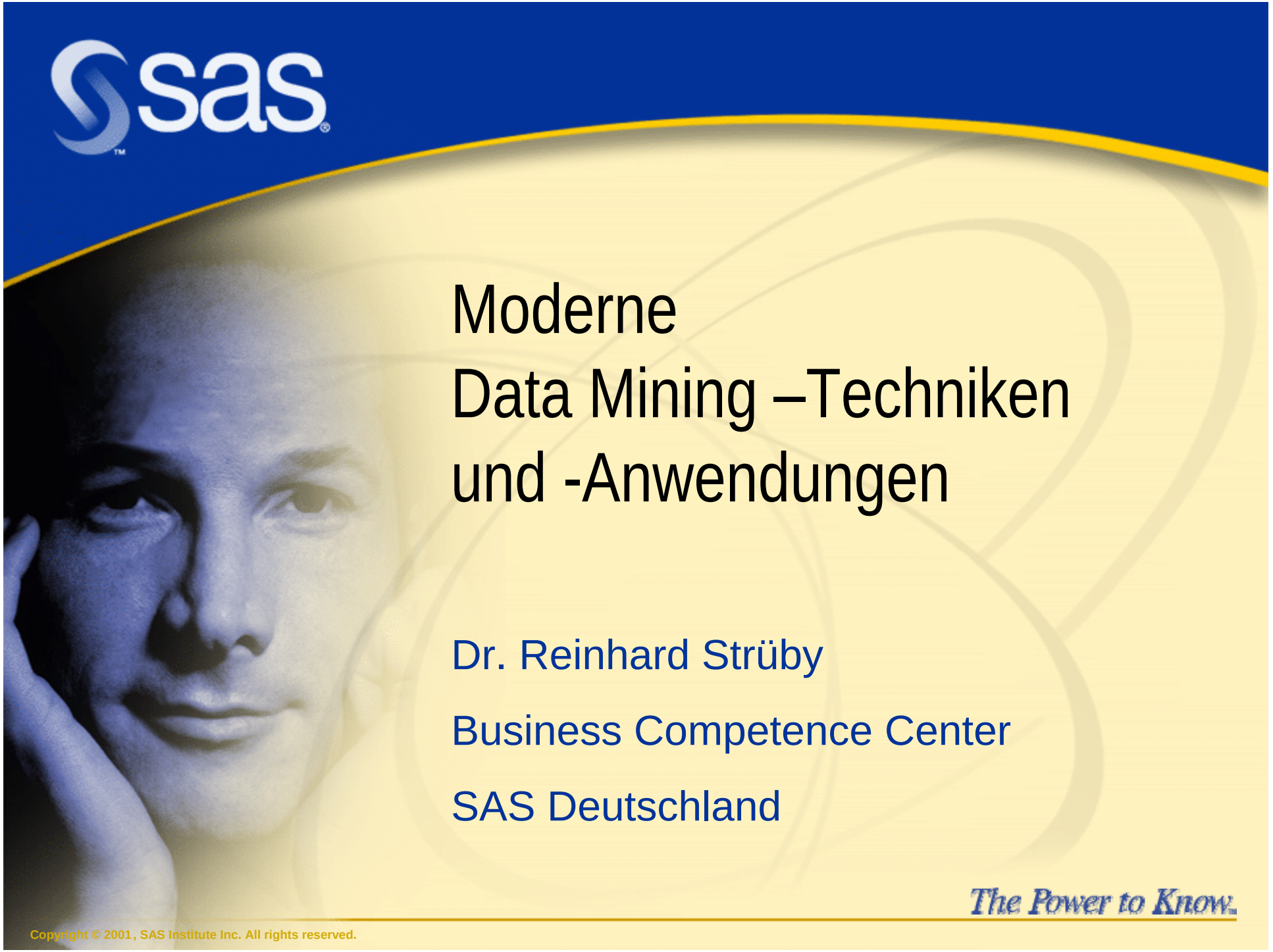


Moderne Data Mining –Techniken und -Anwendungen

Dr. Reinhard Strüby

Business Competence Center

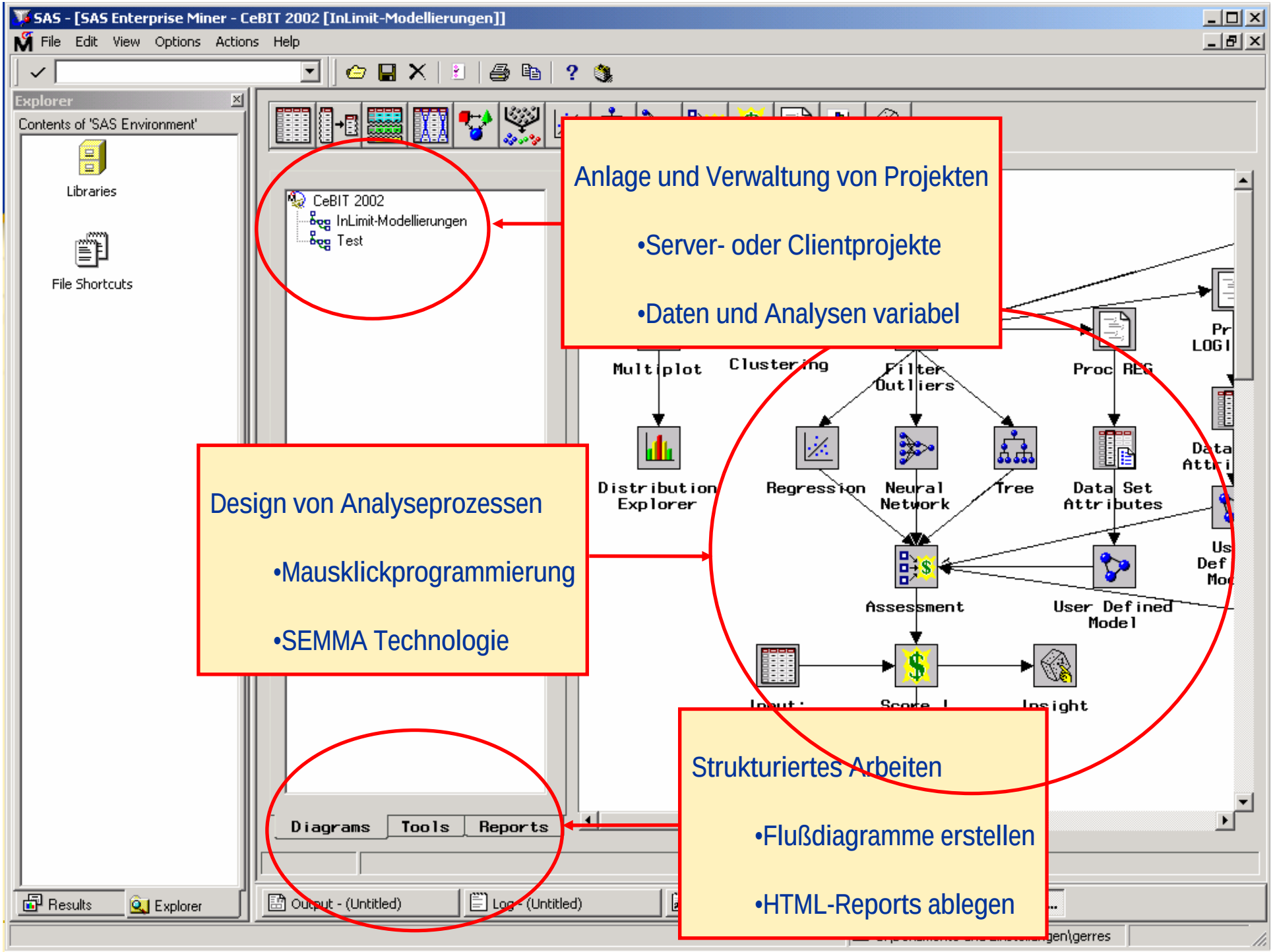
SAS Deutschland

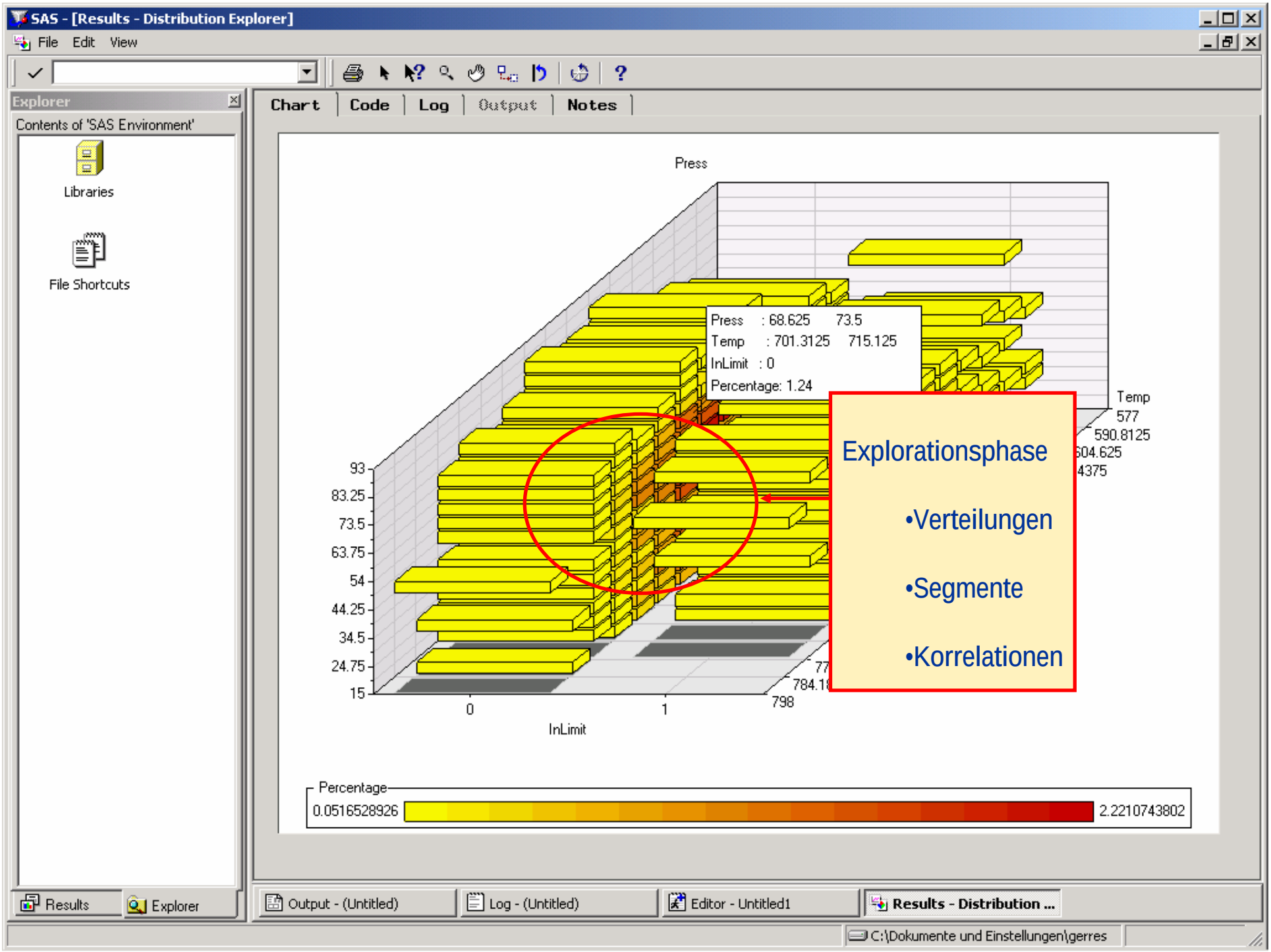
The Power to Know.

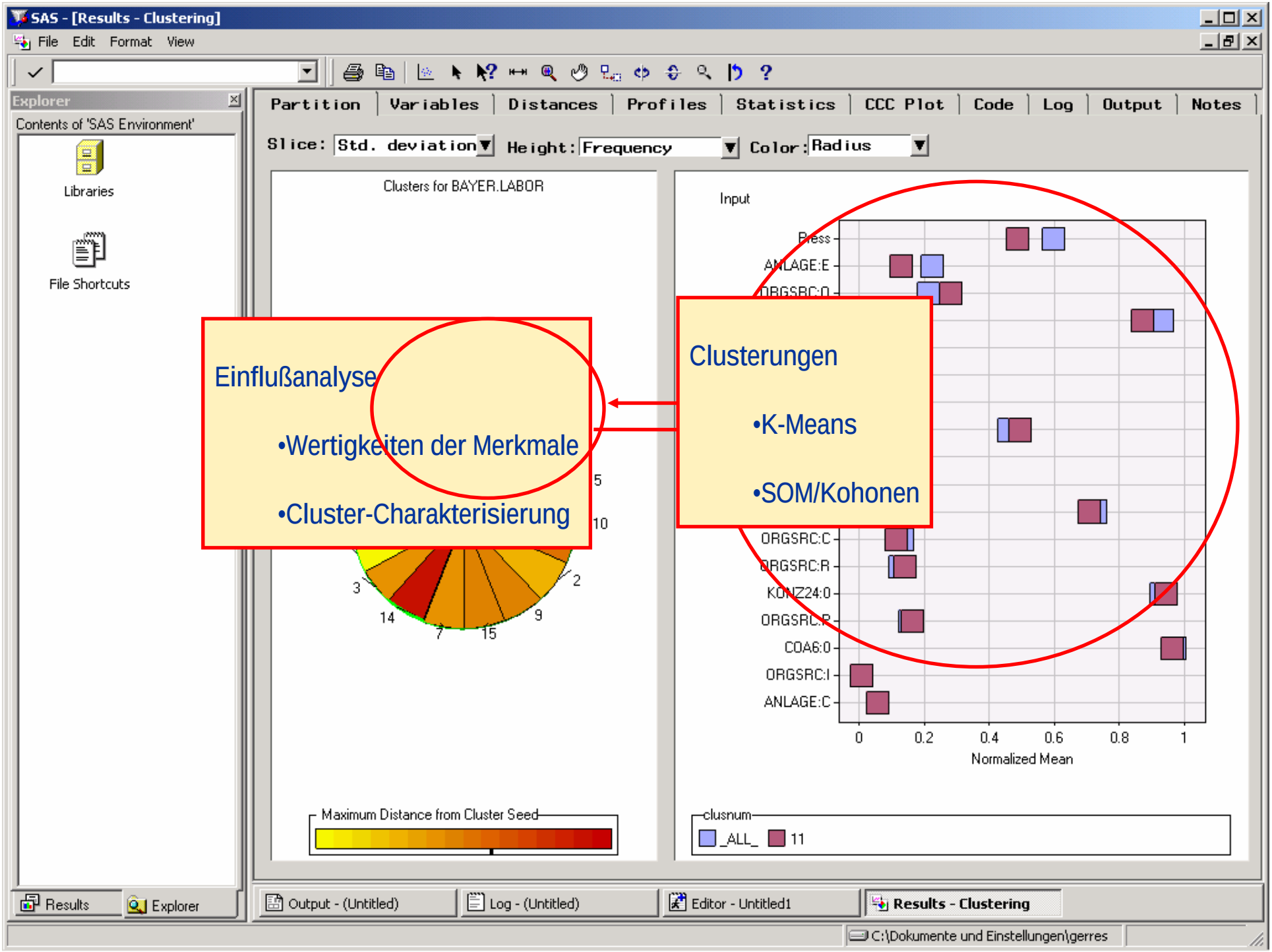
SAS Data Mining: SEMMA Technologie

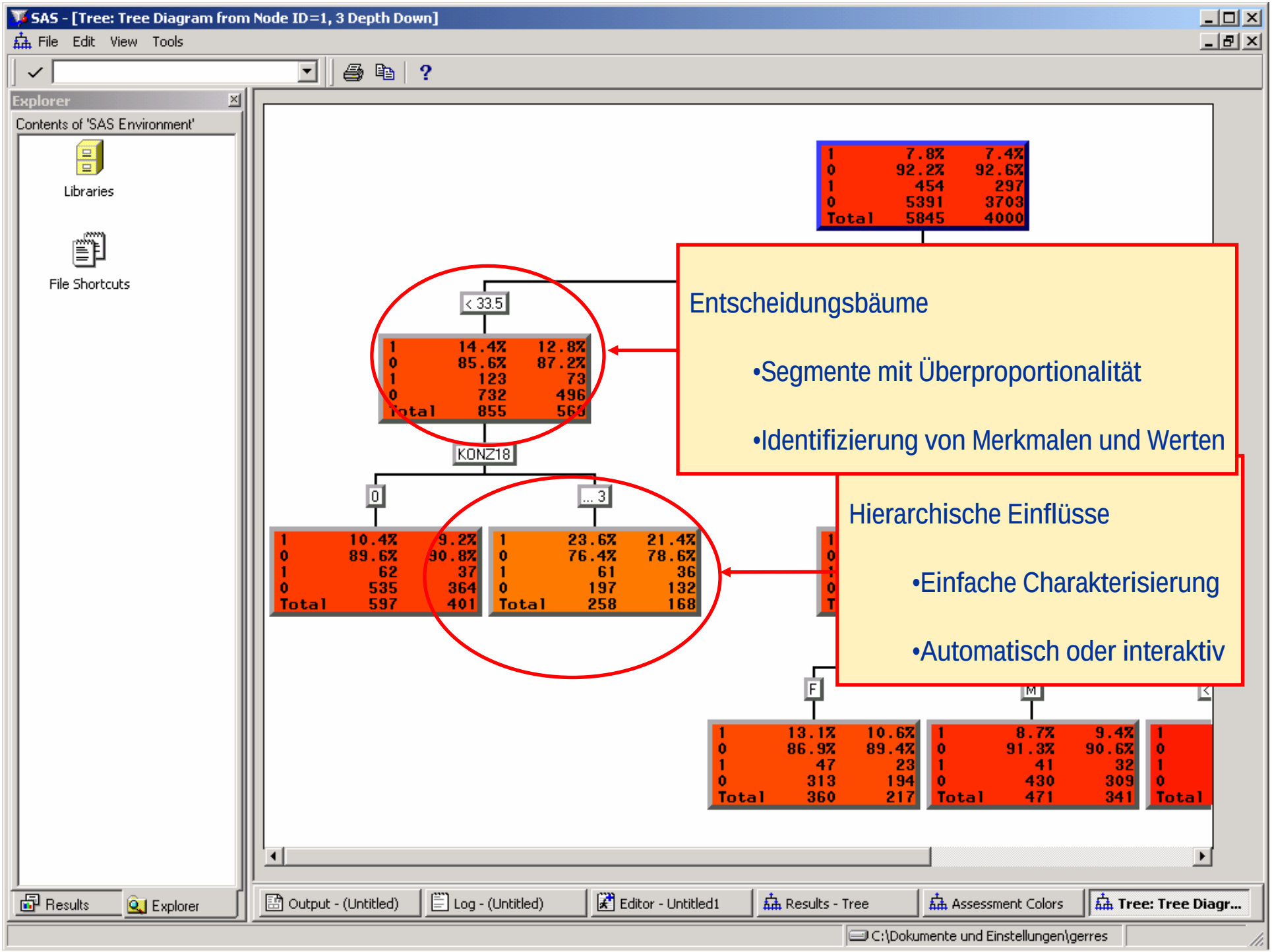
- **S**tichproben
  - Zufällig, geschichtet
- **E**xploration
  - Grafisch, interaktiv
- **M**odifikationen
  - Ausreißer, Transformationen, Imputationen
- **M**odellierungen
  - Regressionen, Entscheidungsbäume, Neuronale Netze
- **A**ssessment
  - Liftcharts, Profitcharts

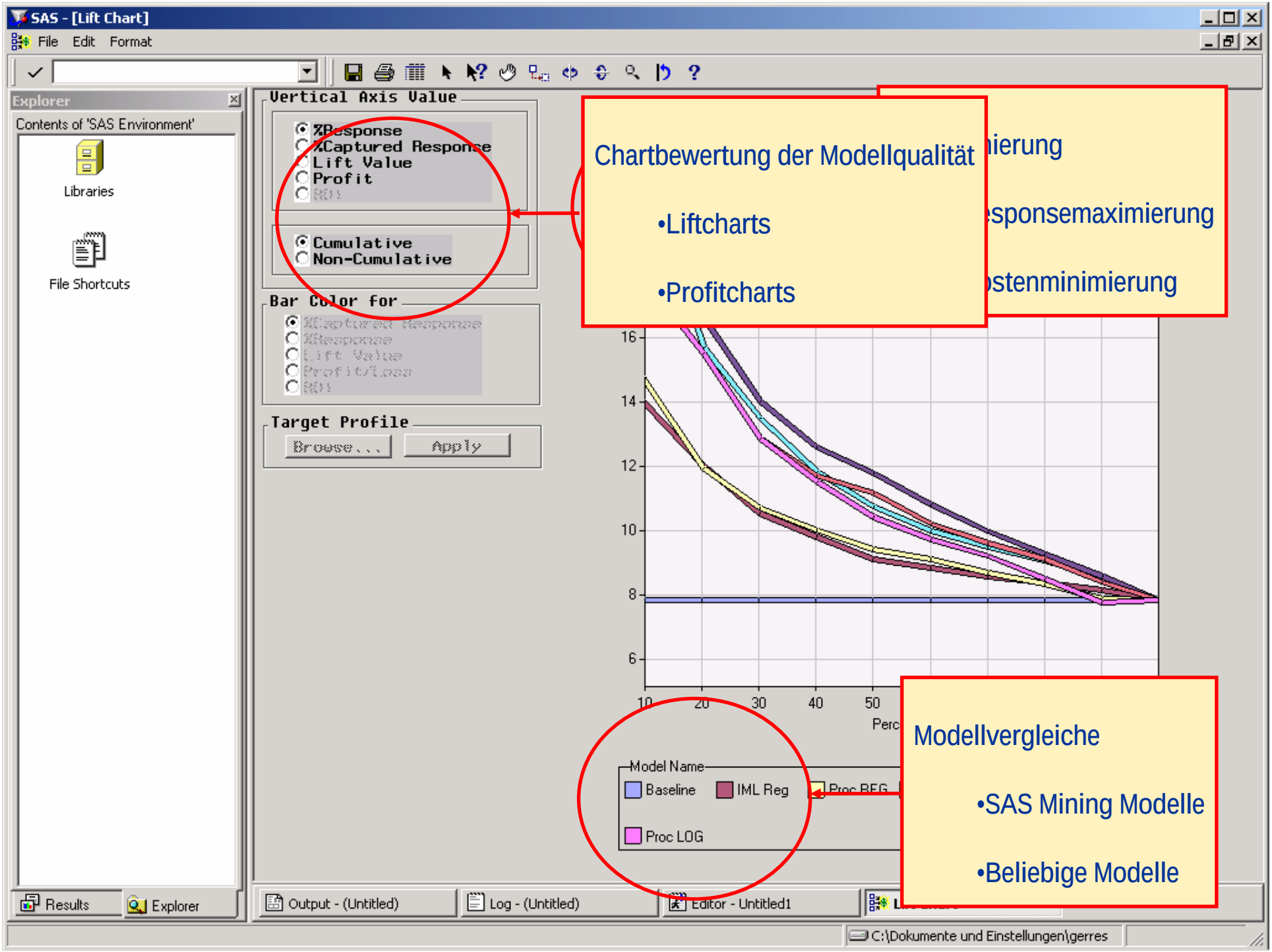
The Power to Know.

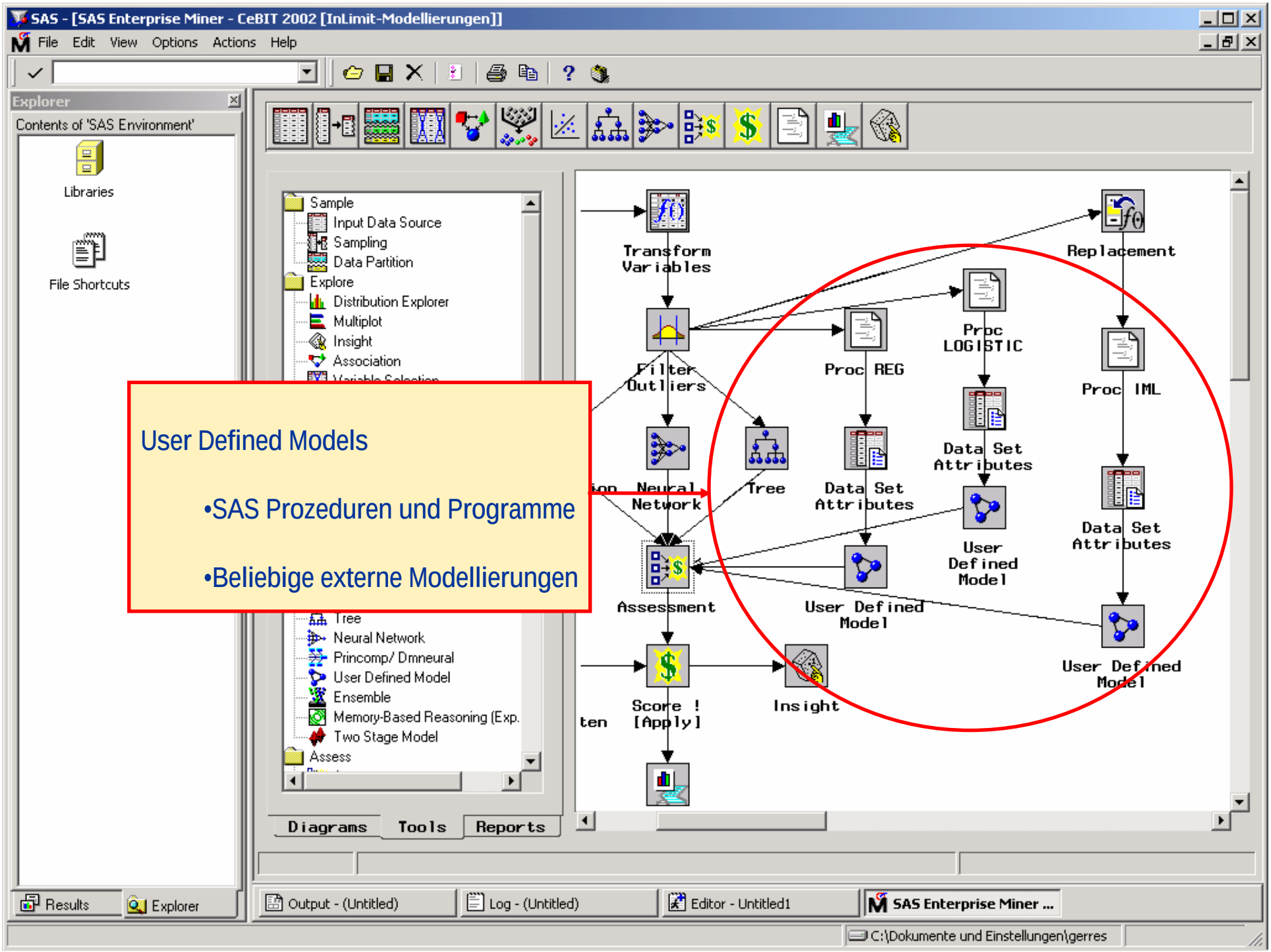

SAS - [SAS Enterprise Miner - CeBIT 2002 [InLimit-Modellierungen]]
File Edit View Options Actions Help
Explorer
Contents of 'SAS Environment'
Libraries
File Shortcuts
CeBIT 2002
InLimit-Modellierungen
Test
Anlage und Verwaltung von Projekten
•Server- oder Clientprojekte
•Daten und Analysen variabel
Multiplot
Clustering
Filter Outliers
Proc REG
Distribution Explorer
Regression
Neural Network
Tree
Data Set Attributes
Assessment
User Defined Model
Insight
Design von Analyseprozessen
•Mausklickprogrammierung
•SEMMA Technologie
Diagrams
Tools
Reports
Strukturiertes Arbeiten
•Flußdiagramme erstellen
•HTML-Reports ablegen
Results
Explorer
Output - (Untitled)
Log - (Untitled)

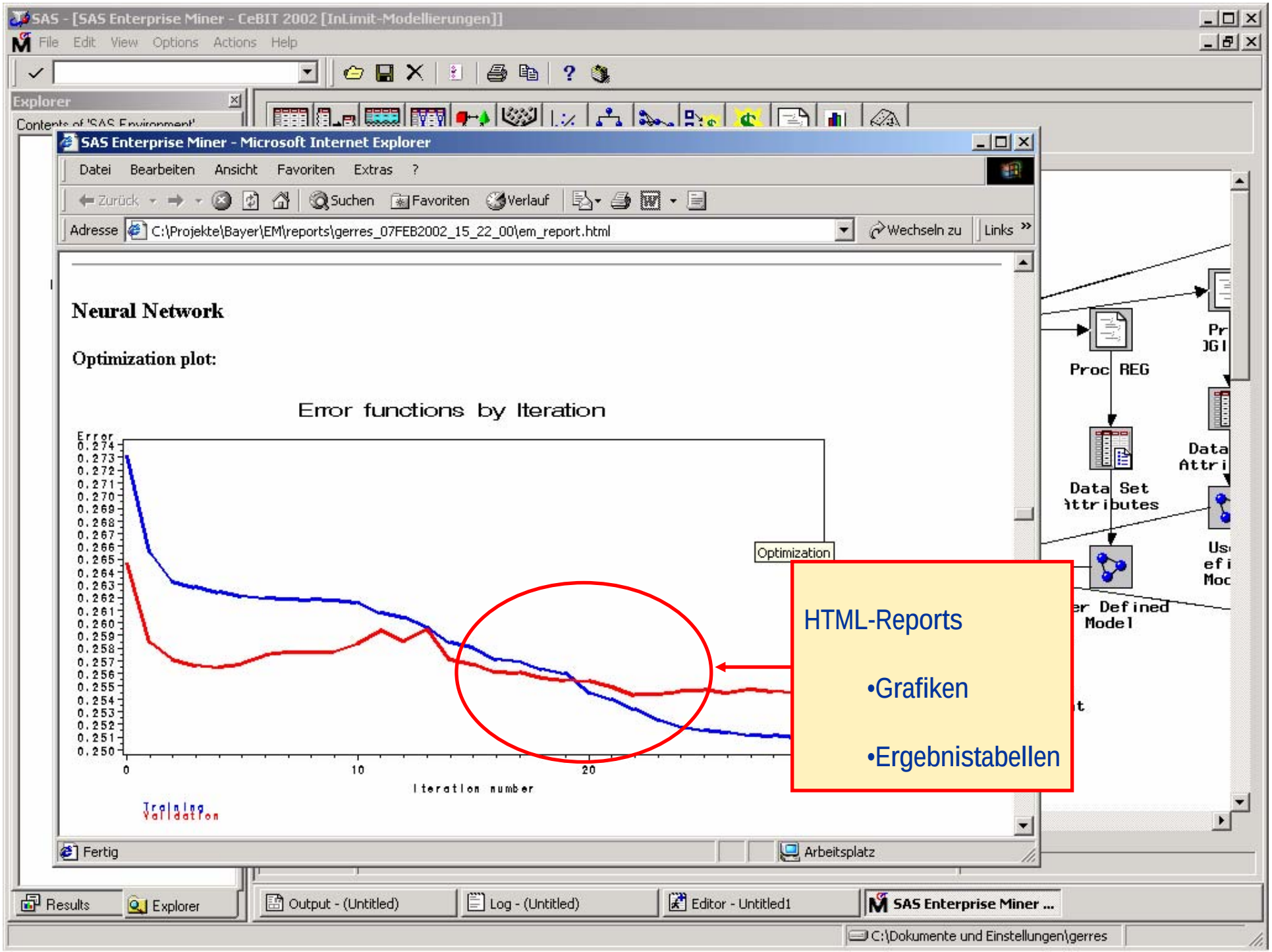

SAS - [Results - Distribution Explorer]
File Edit View
Explorer
Contents of 'SAS Environment'
Libraries
File Shortcuts
Chart Code Log Output Notes
Press
Press : 68.625 73.5
Temp : 701.3125 715.125
InLimit : 0
Percentage: 1.24
Temp
577
590.8125
604.625
4375
93
83.25
73.5
63.75
54
44.25
34.5
24.75
15
0
1
InLimit
798
Explorationsphase
•Verteilungen
•Segmente
•Korrelationen
Percentage
0.0516528926
2.2210743802
Results Explorer
Output - (Untitled)
Log - (Untitled)
Editor - Untitled1
Results - Distribution ...
C:\Dokumente und Einstellungen\gerres

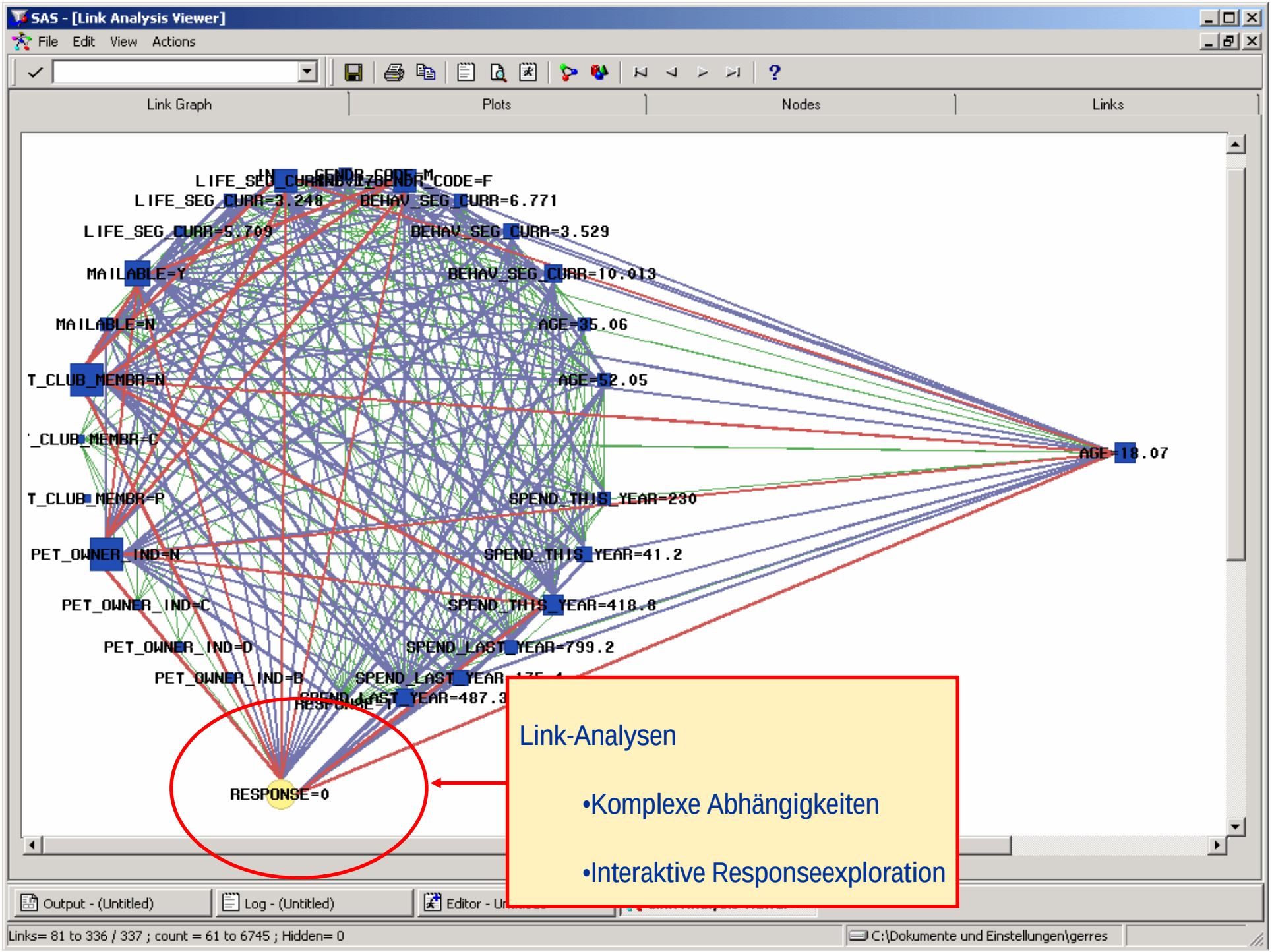

SAS - [Results - Clustering]
File Edit Format View
Explorer
Contents of 'SAS Environment'
Libraries
File Shortcuts
Partition Variables Distances Profiles Statistics CCC Plot Code Log Output Notes
Slice: Std. deviation
Height: Frequency
Color: Radius
Clusters for BAYER.LABOR
Input
Press
ANLAGE:E
ORGSRC:C
ORGSRC:R
COA6:0
ORGSRC:I
ANLAGE:C
0 0.2 0.4 0.6 0.8 1
Normalized Mean
Maximum Distance from Cluster Seed
clusnum
ALL 11
Einflußanalyse
•Wertigkeiten der Merkmale
•Cluster-Charakterisierung
Clusterungen
•K-Means
•SOM/Kohonen
Results Explorer
Output - (Untitled) Log - (Untitled) Editor - Untitled1 Results - Clustering
C:\Dokumente und Einstellungen\gerres


SAS - [Tree: Tree Diagram from Node ID=1, 3 Depth Down]
File Edit View Tools
Explorer
Contents of 'SAS Environment'
Libraries
File Shortcuts
< 33.5
KONZ18
0
... 3
F
M
Entscheidungsbäume
•Segmente mit Überproportionalität
•Identifizierung von Merkmalen und Werten
Hierarchische Einflüsse
•Einfache Charakterisierung
•Automatisch oder interaktiv
Results
Explorer
Output - (Untitled)
Log - (Untitled)
Editor - Untitled1
Results - Tree
Assessment Colors
Tree: Tree Diagr...
C:\Dokumente und Einstellungen\gerres


SAS - [Lift Chart]
File Edit Format
Explorer
Contents of 'SAS Environment'
Libraries
File Shortcuts
Vertical Axis Value
%Response
%Captured Response
Lift Value
Profit
Cumulative
Non-Cumulative
Bar Color for
Target Profile
Browse...
Apply
Chartbewertung der Modellqualität
•Liftcharts
•Profitcharts
ierung
esponsemaximierung
stenminimierung
16
14
12
10
8
6
10
20
30
40
50
Perc
Model Name
Baseline
IML Reg
Proc REG
Proc LOG
Modellvergleiche
•SAS Mining Modelle
•Beliebige Modelle
Results
Explorer
Output - (Untitled)
Log - (Untitled)
Editor - Untitled1
C:\Dokumente und Einstellungen\gerres


SAS - [SAS Enterprise Miner - CeBIT 2002 [InLimit-Modellierungen]]
File Edit View Options Actions Help
Explorer
Contents of 'SAS Environment'
Libraries
File Shortcuts
Sample
Input Data Source
Sampling
Data Partition
Explore
Distribution Explorer
Multiplot
Insight
Association
Tree
Neural Network
Princomp/ Dmneural
User Defined Model
Ensemble
Memory-Based Reasoning (Exp.
Two Stage Model
Assess
Diagrams
Tools
Reports
Transform Variables
Filter Outliers
Neural Network
Tree
Assessment
Score ! [Apply]
Insight
Proc REG
Data Set Attributes
User Defined Model
Proc LOGISTIC
Data Set Attributes
User Defined Model
Replacement
Proc IML
Data Set Attributes
User Defined Model
User Defined Models
•SAS Prozeduren und Programme
•Beliebige externe Modellierungen
Results
Explorer
Output - (Untitled)
Log - (Untitled)
Editor - Untitled1
SAS Enterprise Miner ...
C:\Dokumente und Einstellungen\gerres


SAS - [SAS Enterprise Miner - CeBIT 2002 [InLimit-Modellierungen]]
File Edit View Options Actions Help
Explorer
SAS Enterprise Miner - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten Verlauf
Adresse C:\Projekte\Bayer\EM\reports\gerres_07FEB2002_15_22_00\em_report.html
Wechseln zu Links
Neural Network
Optimization plot:
Error functions by Iteration
Error
0.274
0.273
0.272
0.271
0.270
0.269
0.268
0.267
0.266
0.265
0.264
0.263
0.262
0.261
0.260
0.259
0.258
0.257
0.256
0.255
0.254
0.253
0.252
0.251
0.250
0
10
20
Iteration number
Training
Validation
Optimization
HTML-Reports
•Grafiken
•Ergebnistabellen
Proc REG
Data Set Attributes
Defined Model
Fertig
Arbeitsplatz
Results
Explorer
Output - (Untitled)
Log - (Untitled)
Editor - Untitled1
SAS Enterprise Miner ...
C:\Dokumente und Einstellungen\gerres



Link-Analysen

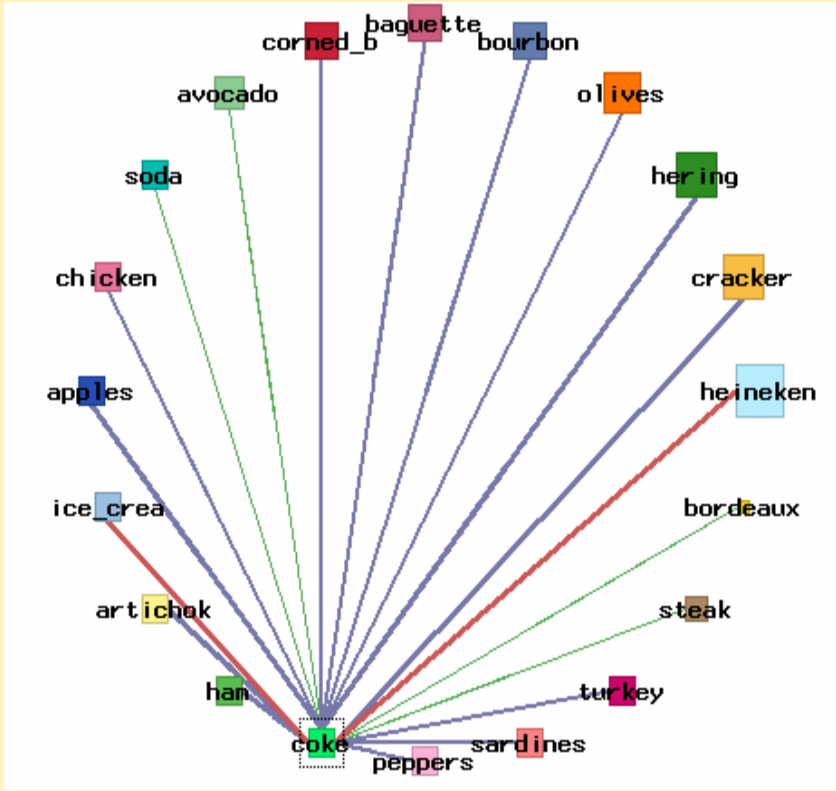
- Komplexe Abhängigkeiten
- Interaktive Responseexploration

Assoziationen: Warenkorb

Assoziationen

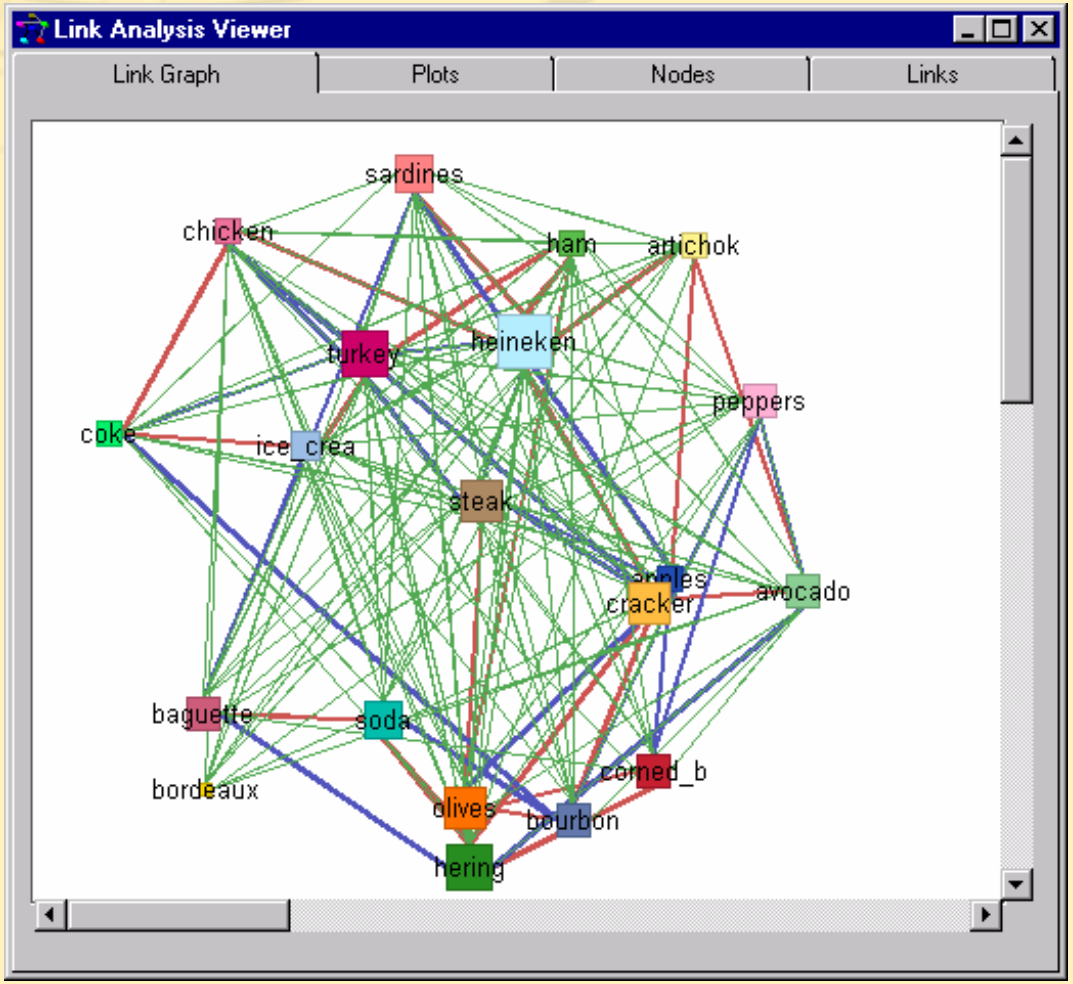
	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	2.38	21.98	74.32	220.00	coke ==> ice_crea
2	2	2.38	21.98	70.29	220.00	ice_crea ==> coke
3	2	1.91	21.08	58.13	211.00	avocado ==> artichok
4	2	1.91	21.08	69.18	211.00	artichok ==> avocado
5	2	1.68	14.69	49.66	147.00	sardines ==> coke
6	2	1.68	14.69	49.66	147.00	coke ==> sardines
7	2	1.66	11.79	51.98	118.00	steak ==> apples
8	2	1.66	11.79	37.58	118.00	apples ==> steak
9	2	1.65	22.08	78.09	221.00	turkey ==> olives
10	2	1.65	22.08	46.72	221.00	olives ==> turkey
11	2	1.63	15.08	51.01	151.00	sardines ==> ice_crea
12	2	1.63	15.08	48.24	151.00	ice_crea ==> sardines
13	2	1.62	25.07	78.93	251.00	soda ==> cracker
14	2	1.62	25.07	51.43	251.00	cracker ==> soda
15	2	1.55	13.39	47.35	134.00	turkey ==> ham

Link Analysis



The Power to Know.

Warenkorb-Assoziationen (MDS-Plot)



- Visualisierung der Beziehungen in einem Warenkorb
- Farben und Linienstärken entsprechend Konfidenz und Lift
- Größe der Knoten proportional zum Verkaufsertrag

The Power to Know.

Text Mining

Knowledge

- Aufdecken und Verwenden von Wissen in Dokumenten
- Mustererkennung
- Beziehungen zwischen Dokumenten und Begriffen
- Transformation von Text in quantitative Informationen

The Power to Know.

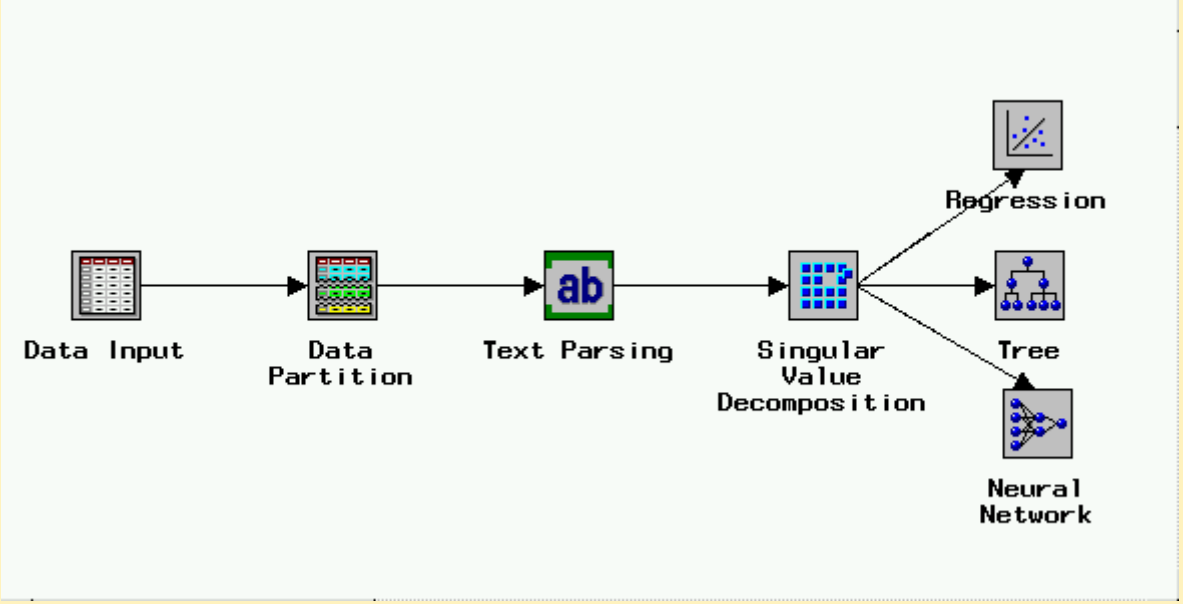
Text Mining

- Ähnlichkeiten von Dokumenten
  - Gemeinsamkeiten in Verwendung von Wortkombinationen
  - Ähnlichkeiten von Wörtern
    - basierend auf ihrem Vorkommen in ähnlichen Dokumenten
- Verwendung externer Quellen
  - Navigations-Sequenzen bei Webnutzung
  - Akzeptanz von Themen für spezielle Kundengruppen
  - Kundencluster aus analogem Navigationsverhalten

Text Mining Anwendungen

- Klassifikation und Clusterung
  - Email-Filterung
  - Mail & Help Desk Routing
  - Kategorisierung in Wissensdatenbanken
- Prognose
  - Aktienkurs aus gegenwärtigen Geschäftsberichten
  - Kundenzufriedenheit
  - Servicekosten anhand von Service Log Daten in Textform

Text Mining Prozeß



The Power to Know.

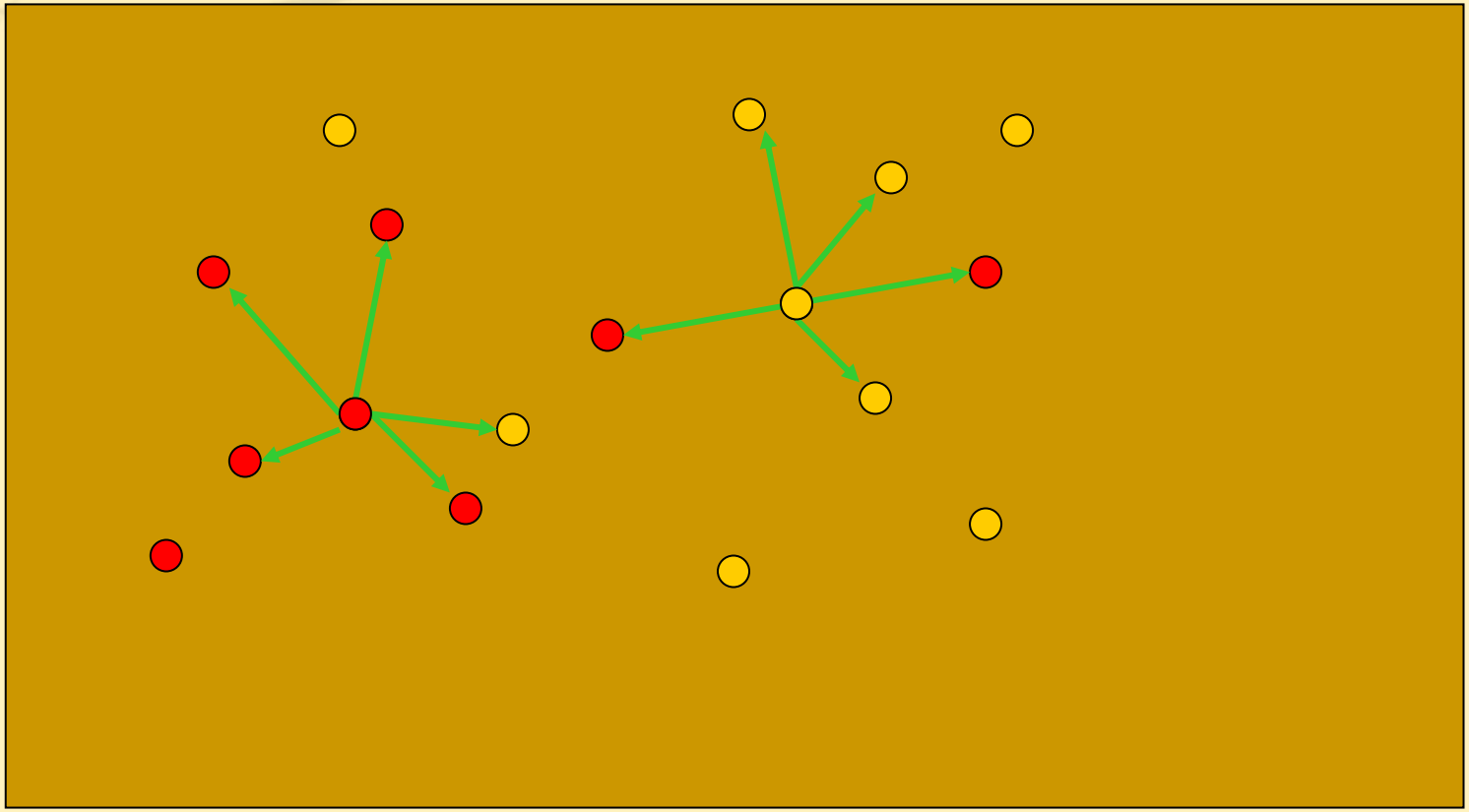
Singular Value Decomposition (SVD)

- Dimensionsreduzierung
  - Informationsmatrix
- Latente Semantik Analyse
  - Zusammenhang zwischen Begriffen und Dokumenten
  - Dokumente mit unterschiedlichen Wortmustern aber ähnlicher Bedeutung

Memory-Based Reasoning

- **“When I see a bird that walks like a duck, swims like a duck and quacks like a duck – I call that bird a duck.” Richard C. Cushing**
- *k*-Nearest Neighbor Algorithmus
- Startprobe: Basisprognose durch Ähnlichkeiten
- Weitere Variable bekannt
  - Nachführung der Probe
  - Präzisierung der Prognose
- Ähnlichkeiten zu iterativen Segmentierungen

Memory-Based Reasoning



The Power to Know.

Memory-Based Reasoning

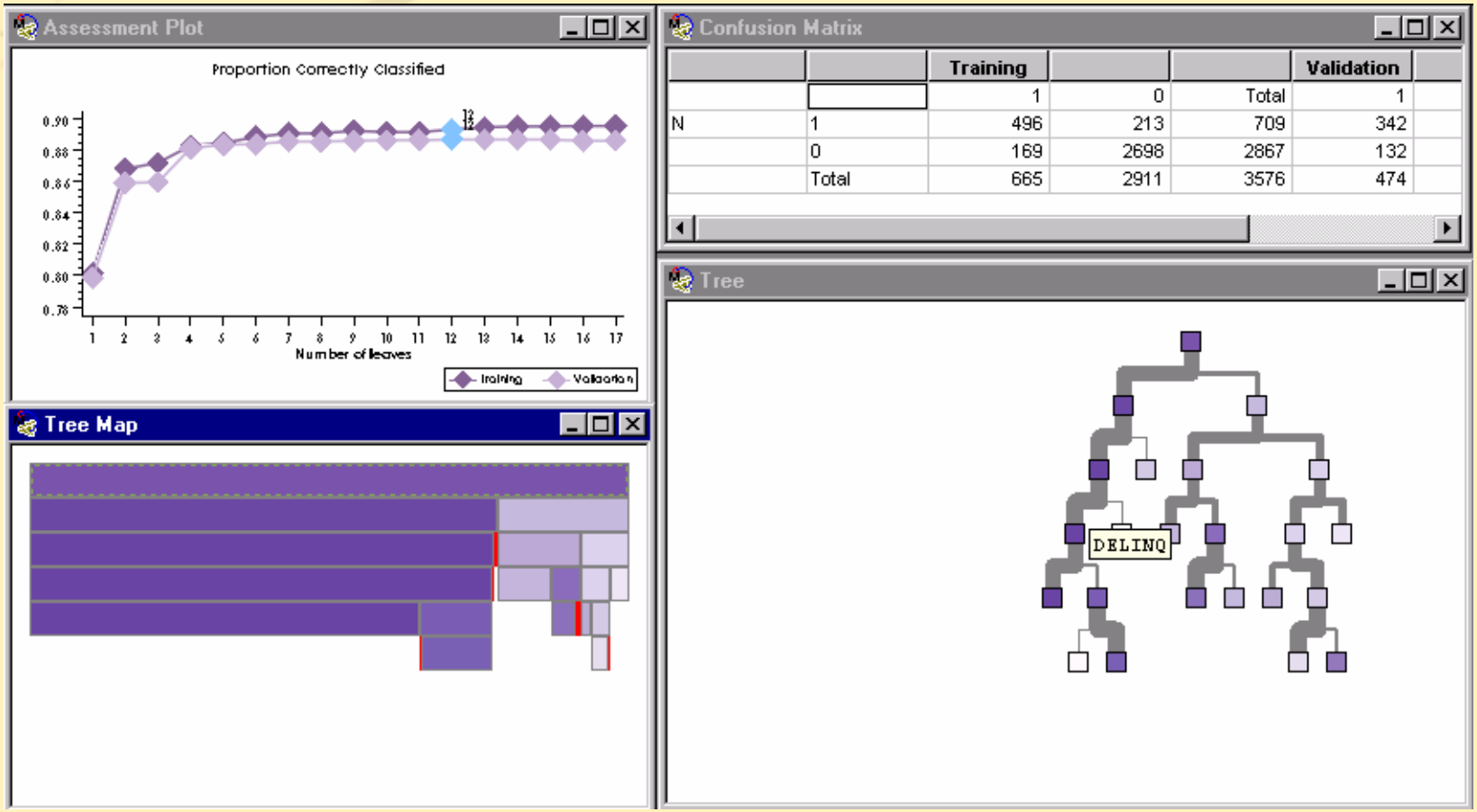
Observation ID	Target : Purchase (Y/N)	Observation Ranking Based on the Distance to the Probe (1: closest 5: farthest)
7	Y	3
12	N	2
35	Y	5
108	Y	1
334	N	4

k	Observation ID of Nearest Neighbors	Target Value of Nearest Neighbors	Posterior Probabilities of the Probe
1	108	Y	prob(Y) = 100% prob(N) = 0%
2	108, 12	Y, N	prob(Y) = 50% prob(N) = 50%
3	108, 12, 7	Y, N, Y	prob(Y) = 67% prob(N) = 33%
4	108, 12, 7, 334	Y, N, Y, N	prob(Y) = 50%; prob(N) = 50%
5	108, 12, 7, 334, 35	Y, N, Y, N, Y	prob(Y) = 60% prob(N) = 40%

Memory-Based Reasoning im Vergleich

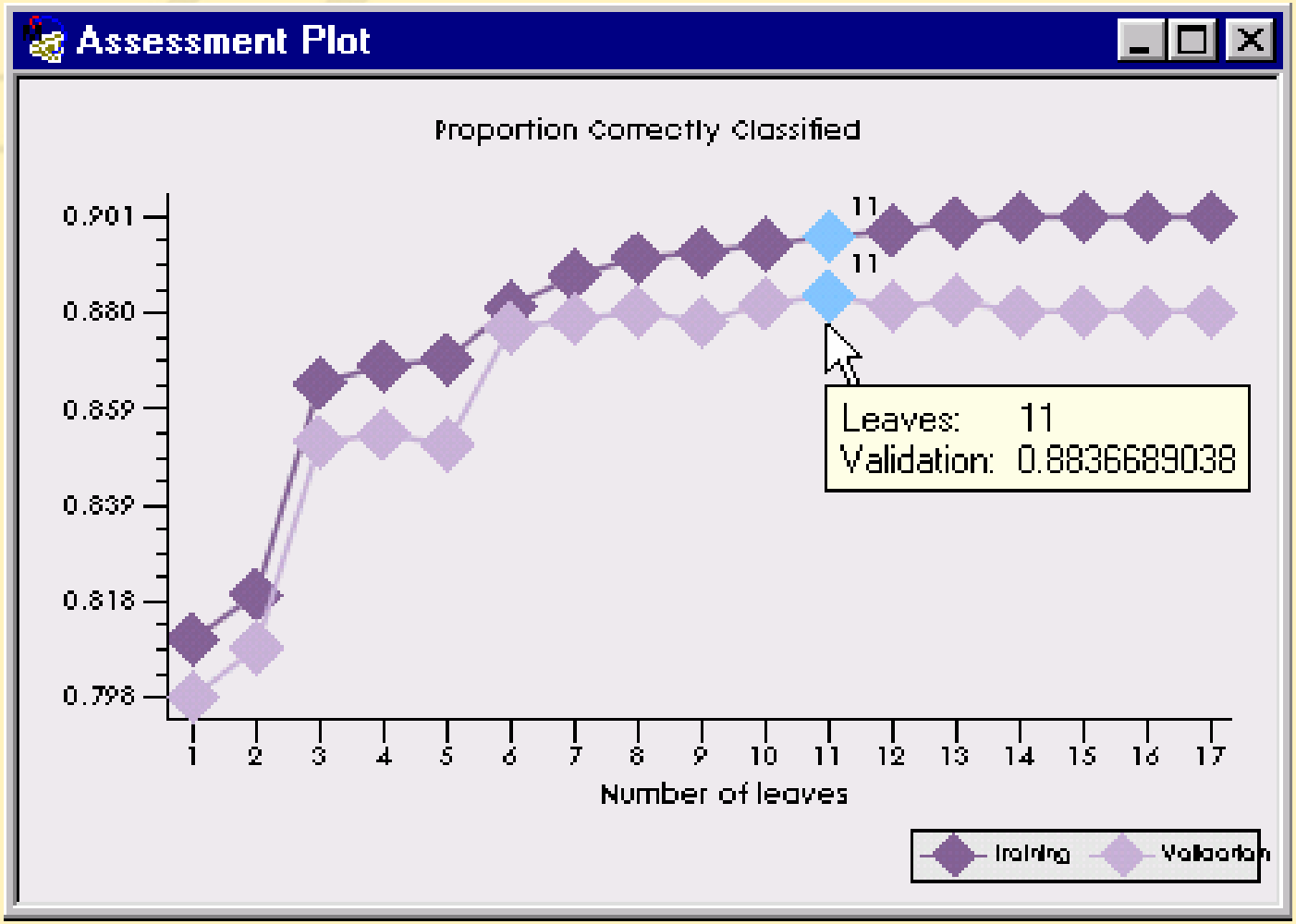
- Alternative zu anderen prediktiven Modellierungstechniken
  - Vorteilhaft bei erkennbaren Strukturen und Mustern
- Modellanpassung mit zeitlicher Dimension
- Einfache Interpretation
- Hoher Speicherdedarf
- Aufwendige Modellnutzung
  - Keine Formel, kein SAS Base Code
  - Rechenintensive Abstandskalkulationen

MFC Tree Browser



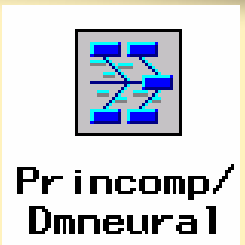
The Power to Know.

MFC Tree Assessment Plot



The Power to Know.

Princomp / Dmneural



- Additive nichtlineare Modelle
- Wenn Kollinearitäten Modellgüte einschränken
- Hauptkomponenten mit höchstem Bestimmtheitsmaß auf Zielgrößen
- Bucket-Hauptkomponenten als Input-Variablen einer Neuronalen-Netz-Modellierung
- Binäre oder intervallskalierte Zielvariable

NEURAL oder DMNEURAL ?

- **NEURAL** für kleinere Datenmengen
  - Feingliedrige Modellierung komplexer nichtlinearer Zusammenhänge
  - Zeitintensiv

- **DMNEURAL** für große Datenmengen
  - Gröbere Modellierung
  - Hohe Performance und numerische Stabilität

The Power to Know.

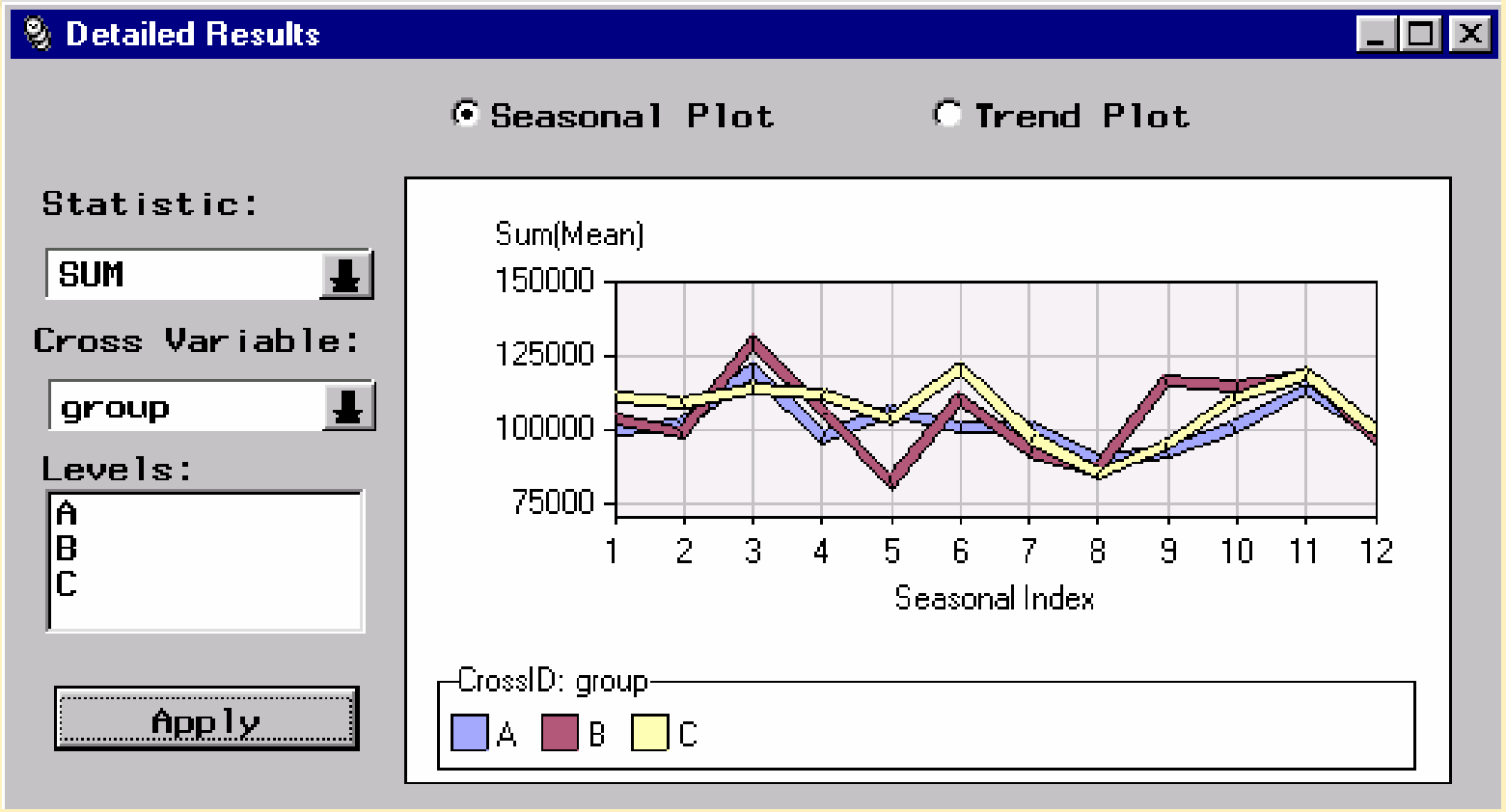
Time Series Node

- Operative Transaktionsdaten in Zeitreihen überführen
- Trendberechnungen
- Saisonale Effekte aufdecken

The Power to Know.

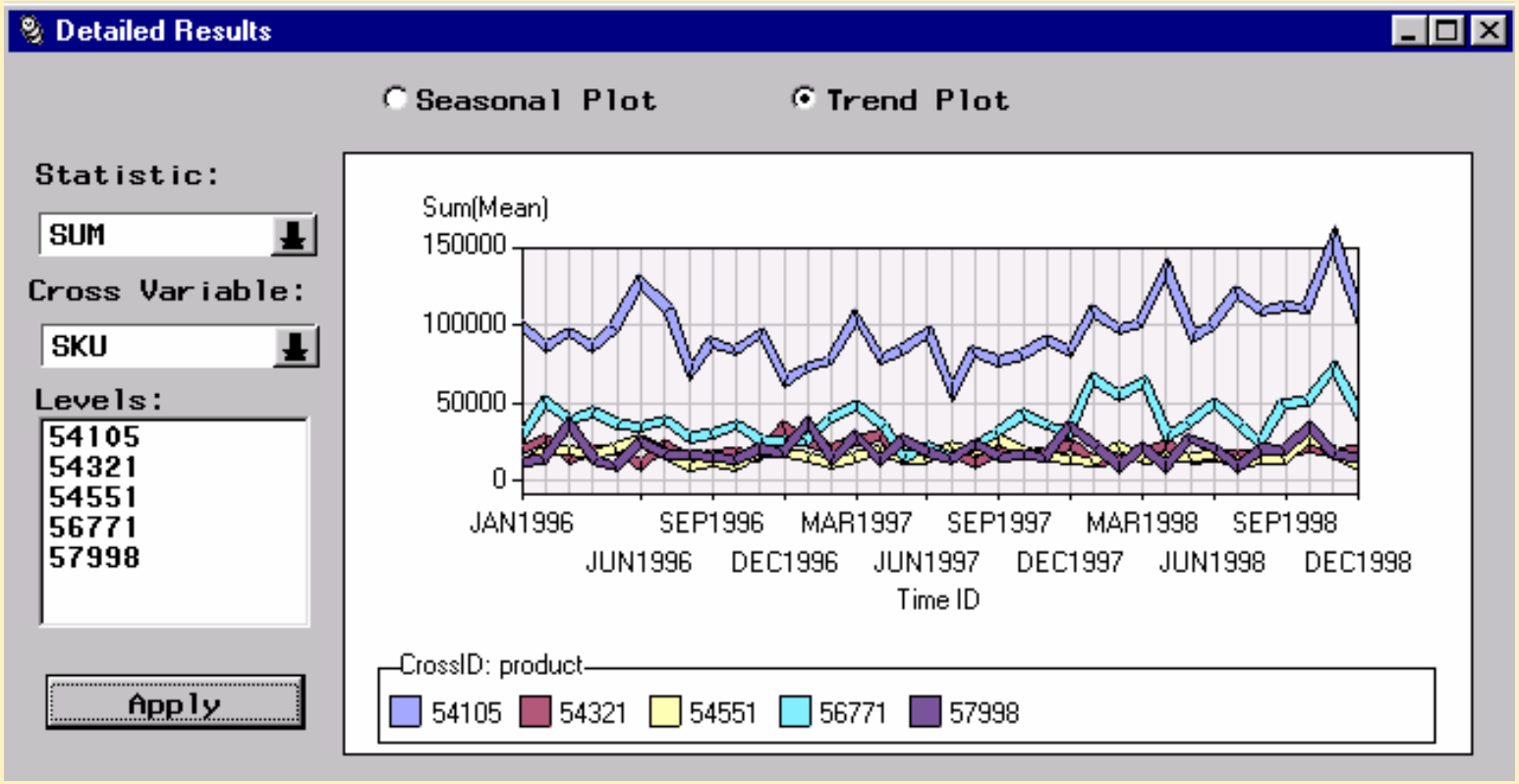
Seasonal Plot

Mittlere Zielwerte für jede Saison



Trend Plot

Mittlere Zielwerte für jedes Intervall des Zyklus



The Power to Know.

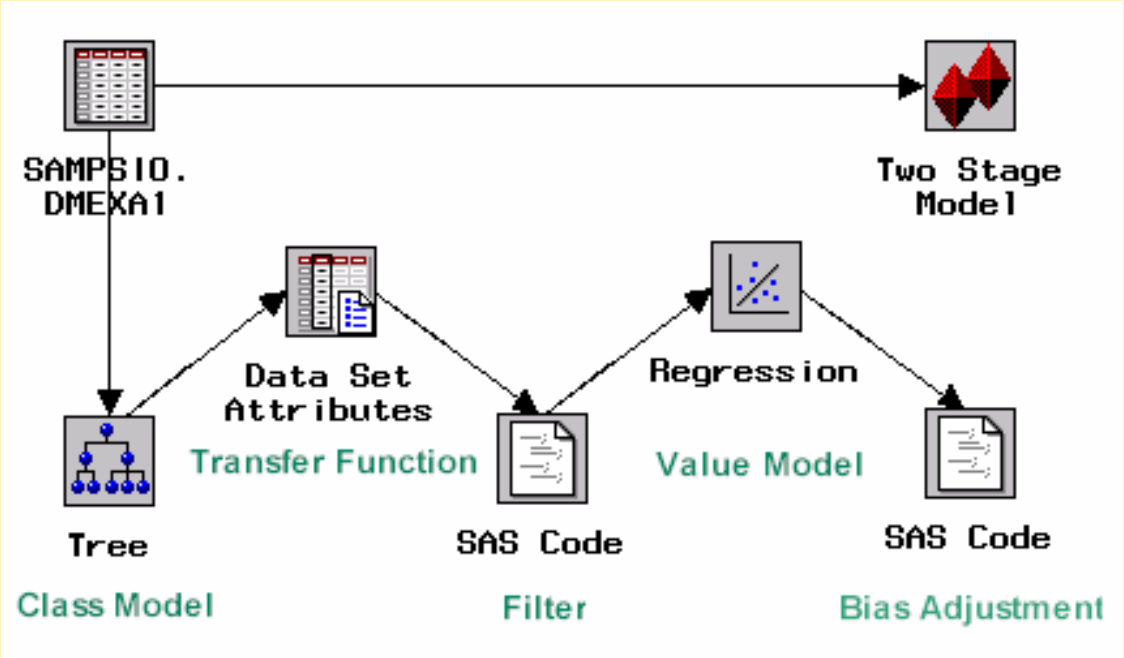
Zweistufenmodelle

- Kombinierte Vorhersage von nominaler und intervallskalierter Zielgröße
- Beispiel: Kundenbewertung
  - Kunde kauft: ja/nein
  - Wenn Kunde kauft: Umsatzprognose

Zweistufenmodelle

- Automatische Modellwahl für beide Teilmodelle
- Modellverknüpfung kann gesteuert werden
  - Transfer-Funktionen
  - Filter-Optionen
- Intervall-Modell wird korrigiert durch Posteriori-Wahrscheinlichkeiten des Class-Modells
- Erzeugter Scorecode kombiniert beide Modelle
- Assessment-Plots für das zusammengesetzte Modell

Erhebliche Vereinfachung



The Power to Know.

Typische Anwendungen für Mehrstufenmodelle

- Marketing Automation / Kampagnenmanagement
  - Basisselektionen
  - Auswahl der Kommunikationskanäle
  - Cell Splits
  - Testkampagnen
  - Responsebewertungen
  - Optimierte mehrstufige Kampagnen

Score Code Deployment
C*Score und J*Score

- Automatische Übersetzung
  - Input: SAS Datastep Score Code generiert durch Enterprise Miner (kein manueller Code)
  - Ouput: Score Code in C und Java Syntax

- Erleichterte Verwendung eines EM Model Score Code außerhalb der SAS Umgebung
  - Erfordert kein SAS System zur Laufzeit
  - Erlaubt Einbettung in externe Anwendungen

Engine Diagnostic Tools bei der LH-Technik

- Fehlerfrüherkennung
- Verbrauchsoptimierter Motorbetrieb
- Vermeidung von Triebwerksstandläufen
- Optimierung des Arbeitsumfangs bei der Triebwerksüberholung

The Power to Know.

Corrective Maintenance Action

ACMS

Data Link

Action Order

Test Cell

MS
Eng.
T/C
Cust.
OEM

On Request Output:
-Input Data
-Trends
-Alert Messages

automatic
Alert Messages

IBM Mainframe Frankfurt
ACM Ground System
Performance Analysis
- Aircraft
- Engine
- APU

Trend Recognition

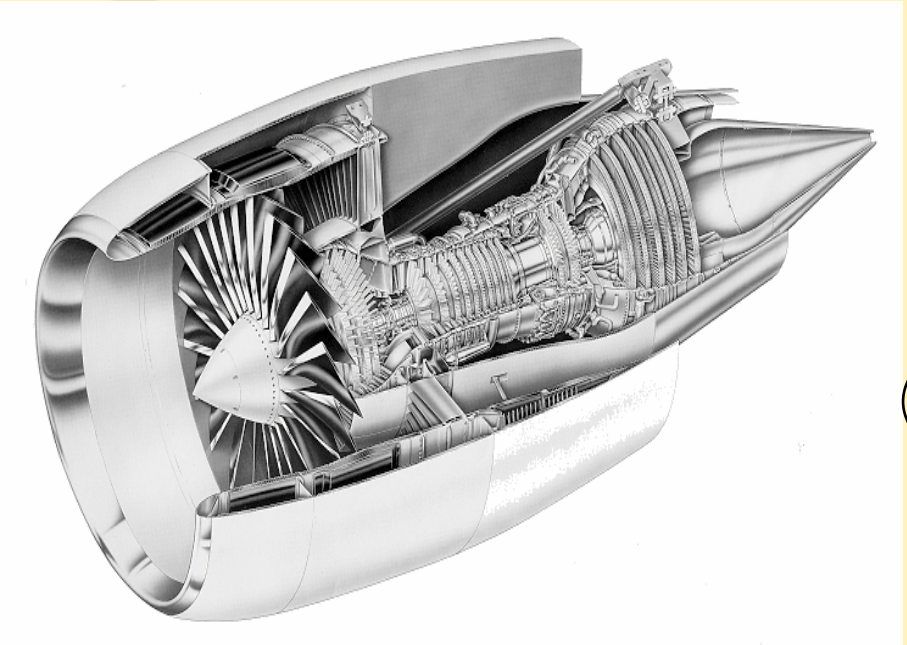
The Power to Know.

Engine Condition Monitoring - Funktionen

SLOATL
EGT MARGIN

OIL SYSTEM
MONITORING

EARLY
FAILURE
DETECTION



SOAP

COLD
FAN TRIM
BALANCE

CONTROLS
MONITORING

TEST CELL

MPA

DERATE
STATISTICS

The Power to Know.

Weiterentwicklung des Instandhaltungssystems

- Erhöhung der Sicherheitsreserven
- Verringerung des Kraftstoffverbrauchs
- Minimierung der Umweltbelastung
- Verringerung der Wartungskosten

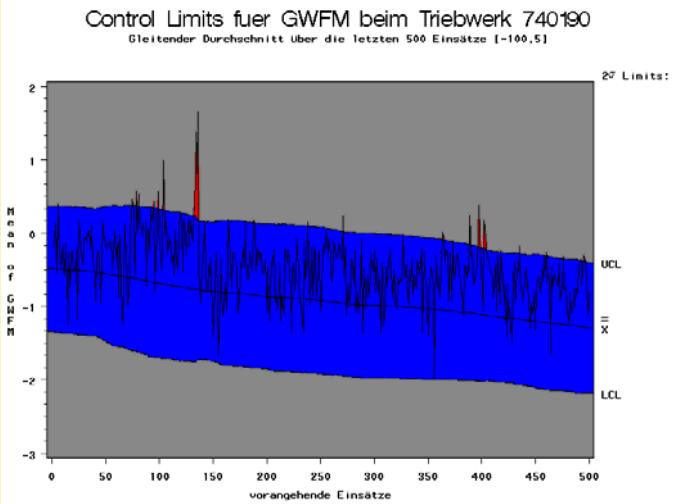
The Power to Know.

Analyseziele

- Erweiterung bisheriger Analysen durch komplexere nichtlineare Modelle mit Wechselwirkungen
- Selektion wesentlicher Variablen aus wachsender Zahl von Meßgrößen
- Erkennen von Zusammenhängen innerhalb von Schadensgruppen

Ergebnisse

- Präzisere Analyseergebnisse durch Kombination mehrerer signifikanter Triebwerksparameter
- Reduzierung redundanter Warnmeldungen
- Erhöhte Flexibilität des Systems

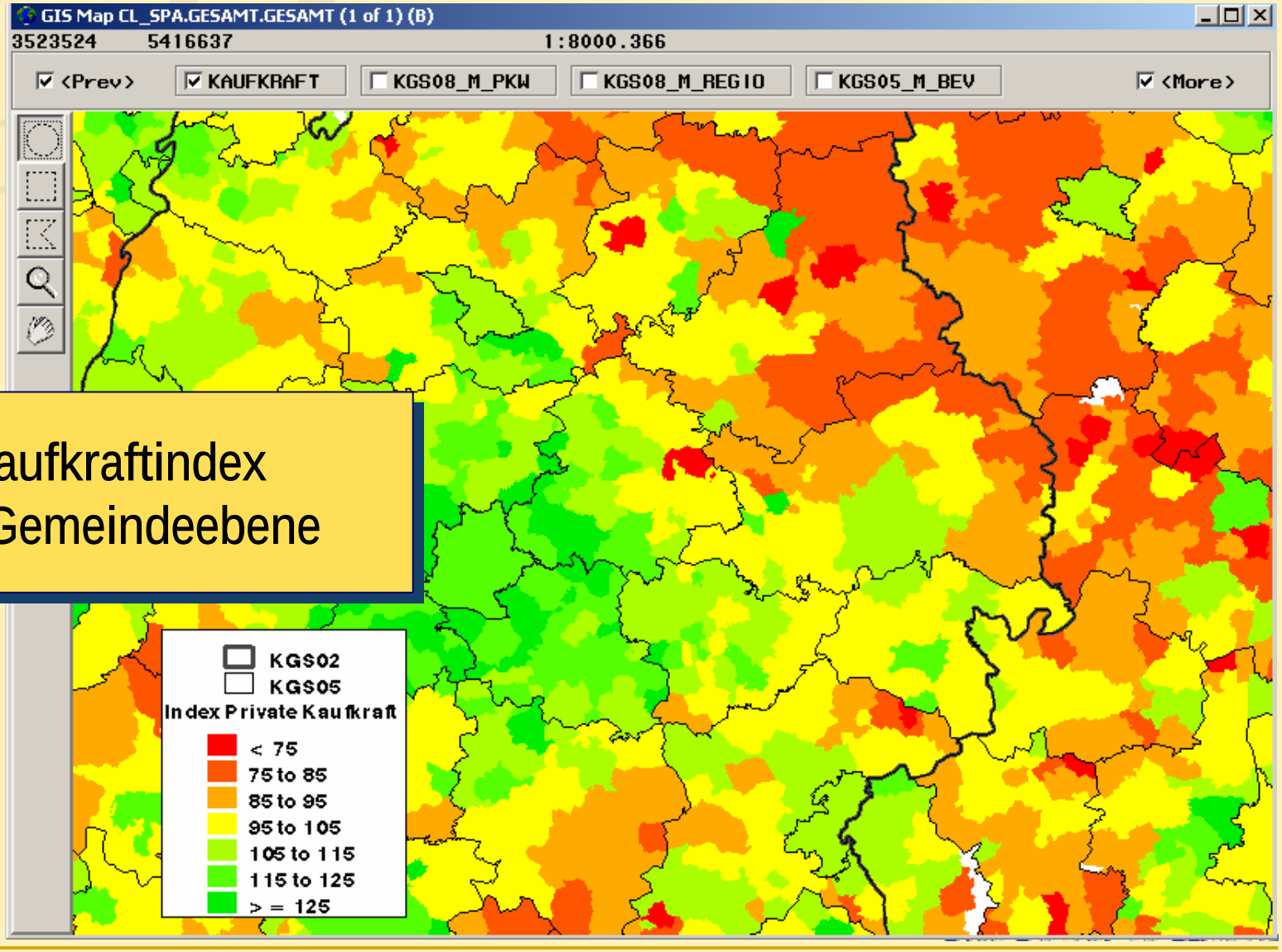


The Power to Know.

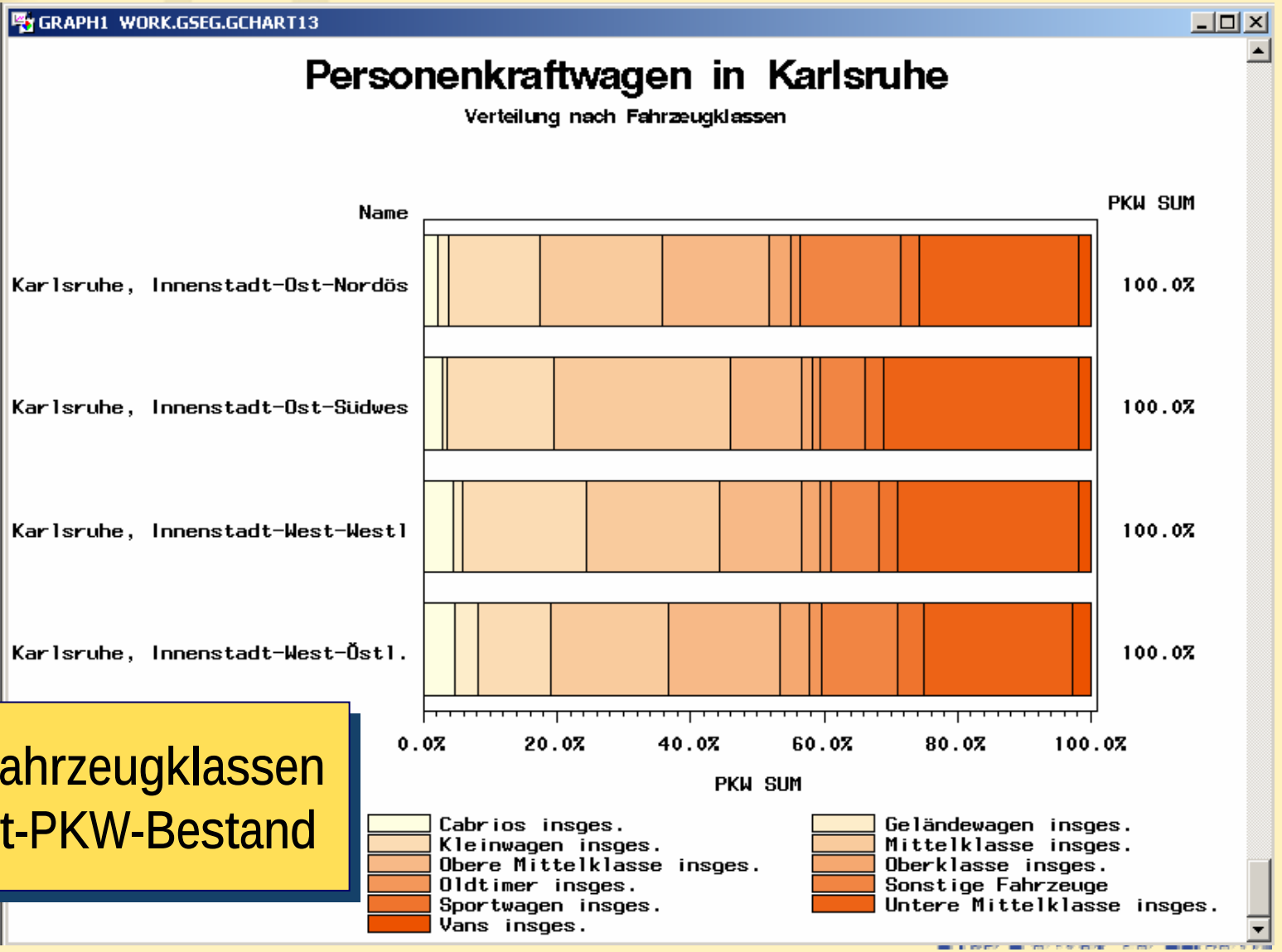
Data Mining in der Genforschung

- Proc ALLELE
  - Analyse genetischer Markierungsdaten
- Proc CASECONTROL
  - Tests auf Relationen zwischen Markierung und Erkrankung
- Proc FAMILY
  - Transmission/Disequilibrium-Tests (TDT)
- Proc HAPLOTYPE
  - MLE für Haplotype-Häufigkeiten
- Proc PSMOOTH
  - Glättungsmethoden für multiples Testen

The Power to Know.

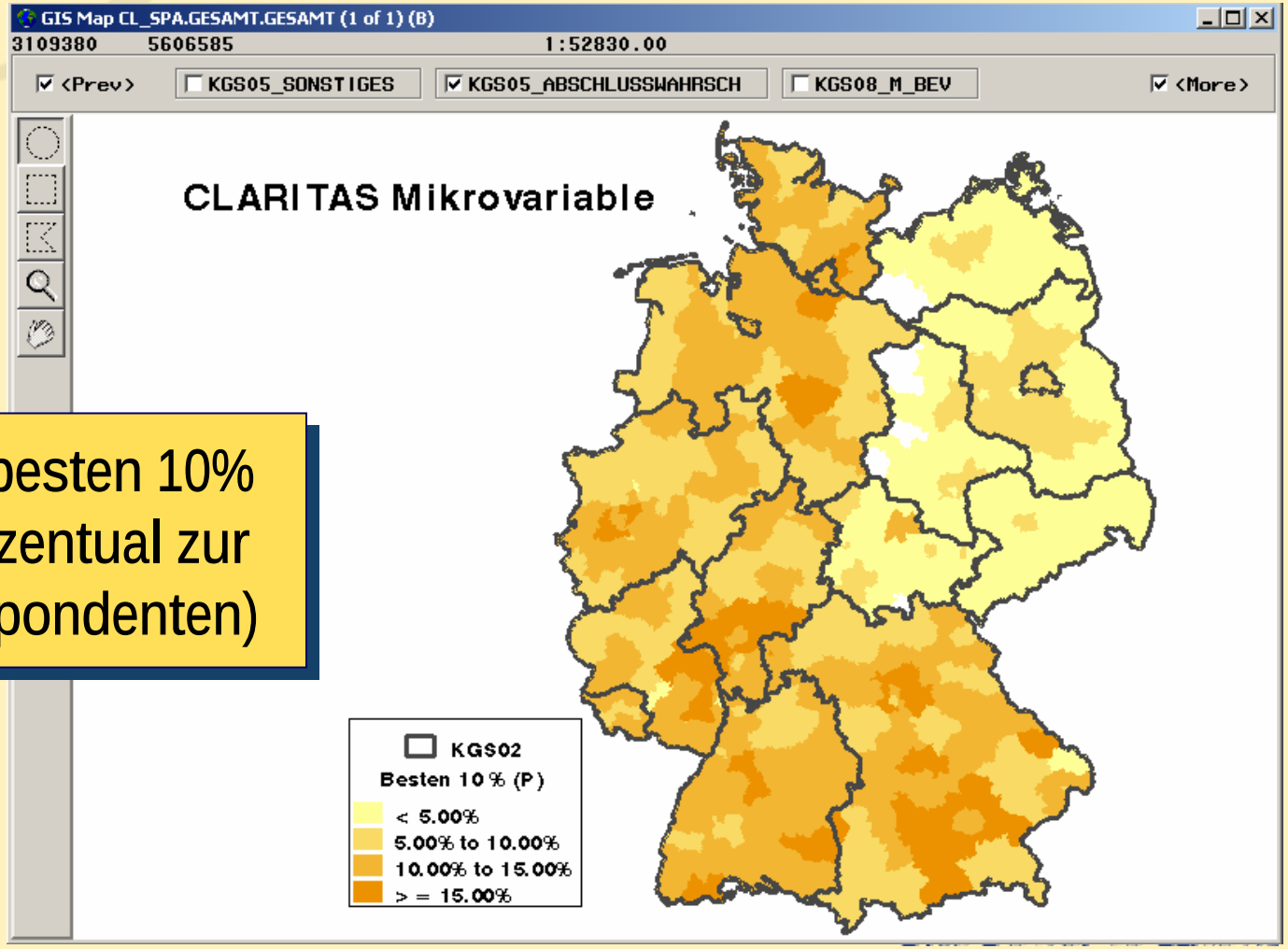


Kaufkraftindex auf Gemeindeebene

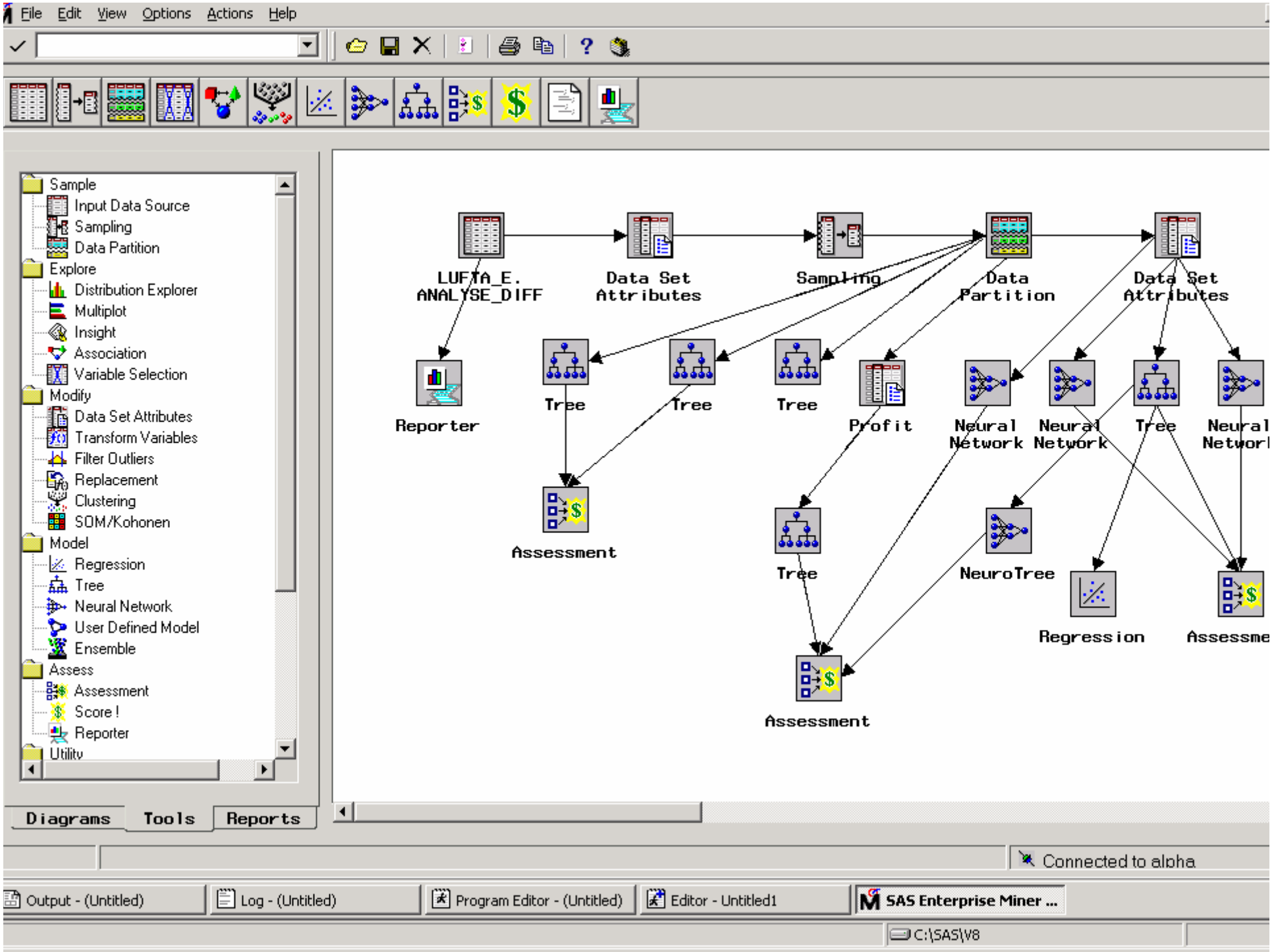


Anteil der Fahrzeugklassen am Gesamt-PKW-Bestand

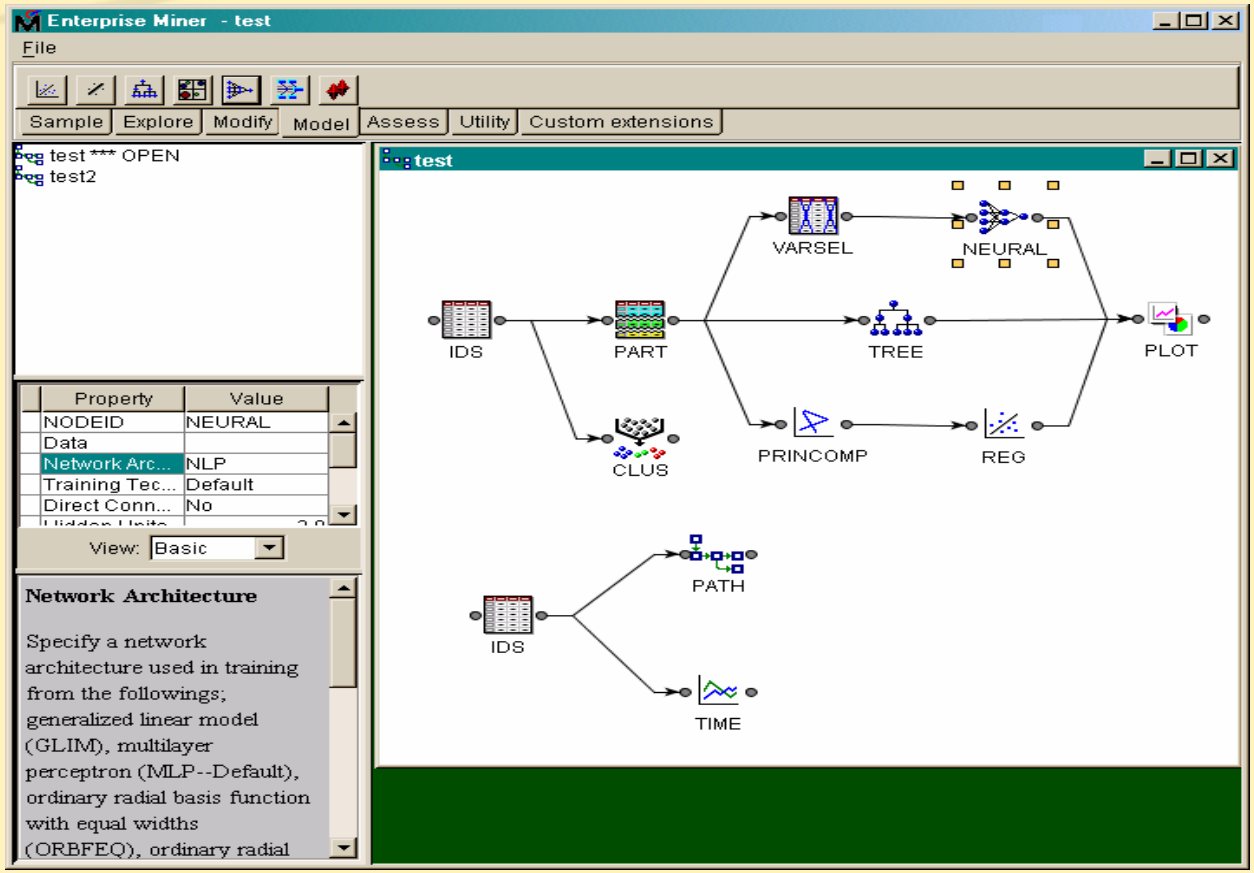
Abschlusswahrscheinlichkeit Lebensversicherung über € 100.000



Verteilung der besten 10% (Verteilung prozentual zur Anzahl der Respondenten)


File Edit View Options Actions Help
Sample
Input Data Source
Sampling
Data Partition
Explore
Distribution Explorer
Multiplot
Insight
Association
Variable Selection
Modify
Data Set Attributes
Transform Variables
Filter Outliers
Replacement
Clustering
SOM/Kohonen
Model
Regression
Tree
Neural Network
User Defined Model
Ensemble
Assess
Assessment
Score !
Reporter
Utility
Diagrams
Tools
Reports
LUFTA_E.
ANALYSE_DIFF
Data Set
Attributes
Sampling
Data
Partition
Data Set
Attributes
Reporter
Tree
Tree
Tree
Profit
Neural
Network
Neural
Network
Tree
Neural
Networ
Assessment
Tree
NeuroTree
Regression
Assessme
Assessment
Connected to alpha
Output - (Untitled)
Log - (Untitled)
Program Editor - (Untitled)
Editor - Untitled1
SAS Enterprise Miner ...
C:\SAS\V8

SAS Enterprise Miner Version 5.0 (SAS Version 9)



The Power to Know.

The Power to Know.