Neuerungen im Enterprise Miner 5.2 & Text Miner 2.3

Ulrich Reincke, SAS Deutschland

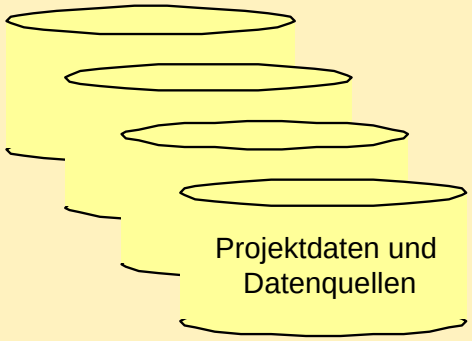
Agenda

Der Neue Enterprise Miner 5.2

Der Neue Text Miner 2.3

Diskussion

Data Mining bisher:

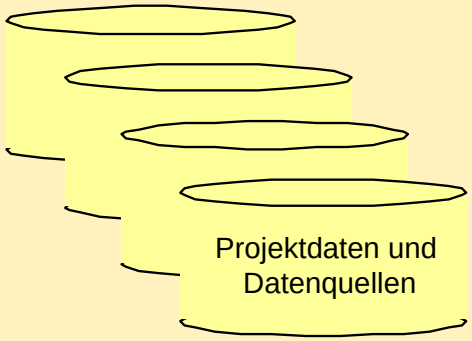


Business Analyst
(Modellentwicklung)

IT-Administration
(Datenbereitstellung
Model Deployment)

Fachanwender
(Modellanwendung)

Data Mining bisher:



Business Analyst
(Modellentwicklung)

Besprechung

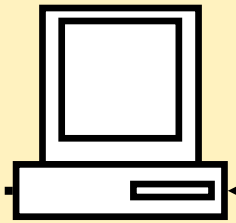
IT-Administration
(Datenbereitstellung
Model Deployment)

Besprechung

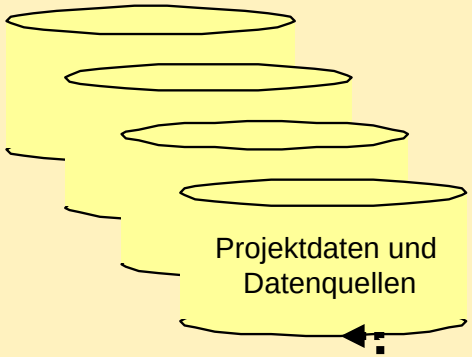
Fachanwender
(Modellanwendung)

Besprechung

Data Mining bisher:



Enterprise Miner



Business Analyst
(Modellentwicklung)

Besprechung

IT-Administration
(Datenbereitstellung
Model Deployment)

Besprechung

Fachanwender
(Modellanwendung)

Besprechung

Das Bermuda Dreieck von Data Mining Projekten

Daten

Projektdaten und Datenquellen

Business Analyst
(Modellentwicklung)

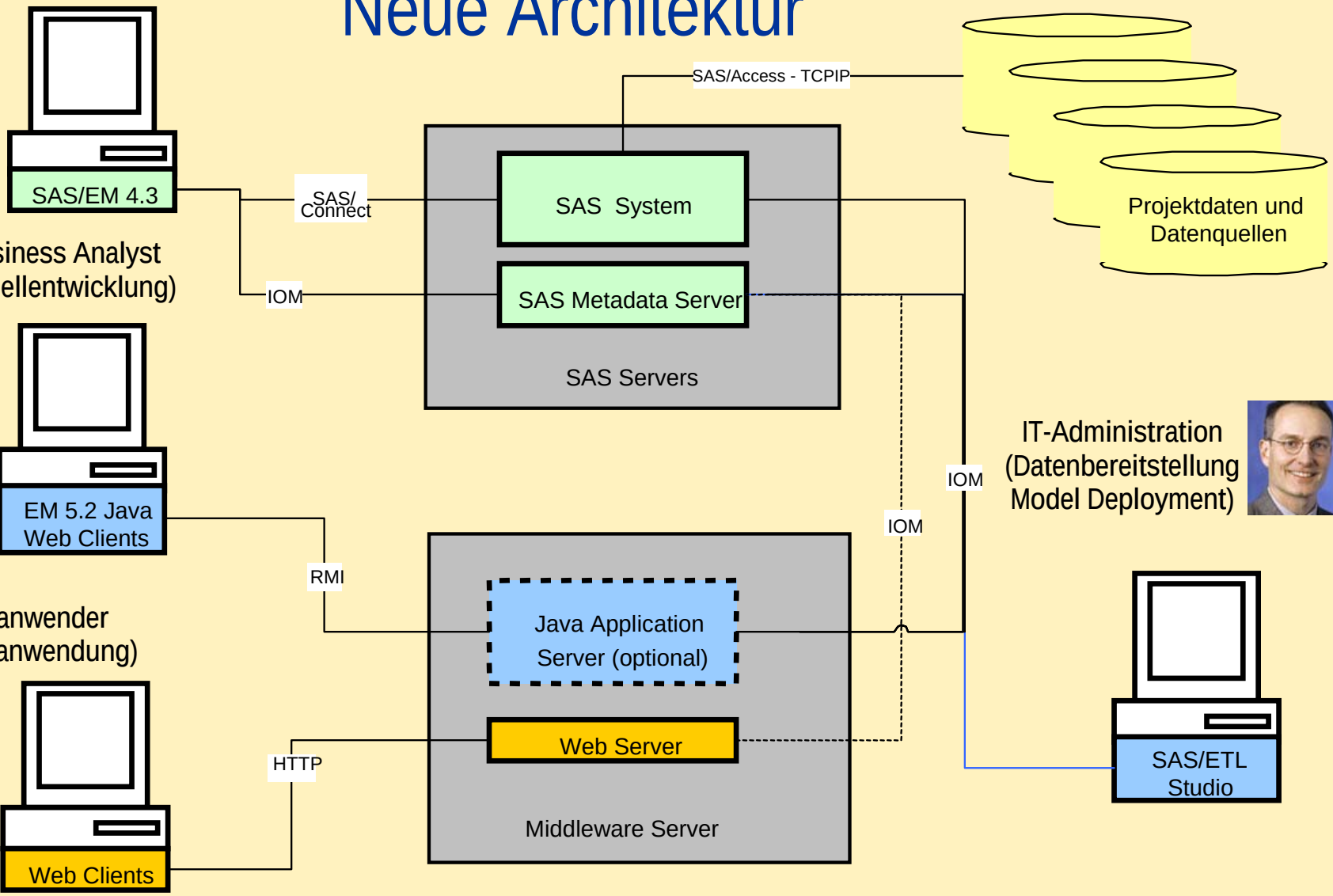
IT-Administration
(Datenbereitstellung
Model Deployment)

Fachanwender
(Modellanwendung)

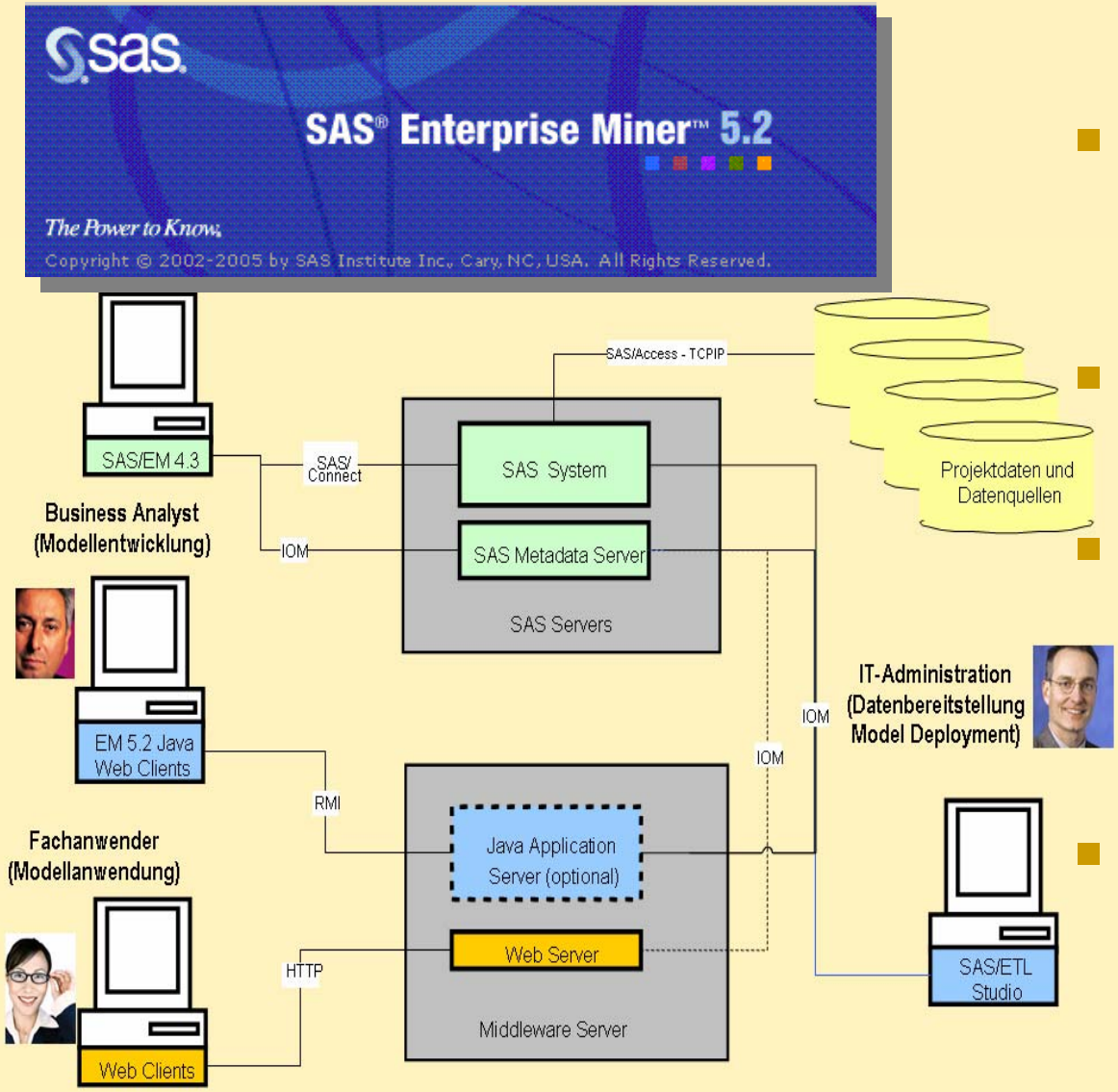
Ergebnisanwendung

Menschen

Neue Architektur



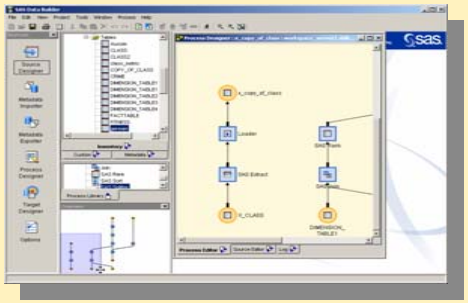
Skalierbare N-Tier-Architektur



- Neu gestaltete Java-basierte Oberfläche
- Unterstützt n Tier Architektur mit Java Application Server
- Performance & Skalierbarkeit
- Ermöglicht Web-basiertes gemeinsames Arbeiten an Data Mining Projekten
- Durchgehendes Metadatenkonzept

Einheitliches Metadaten-Konzept von SAS

Datenmanagement im ETL Studio



1. Registrierung der Daten für Modellierung

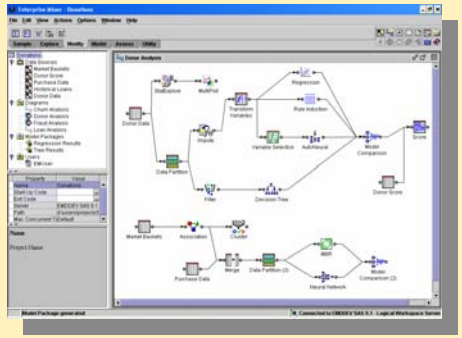
4. Scoring im Batch-Modus

Metadaten-Server

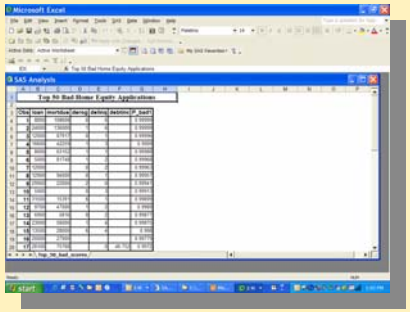
2. Verwendung der Daten für die Modellierung

3. Registrieren als Results Package

Modellentwicklung im Enterprise Miner



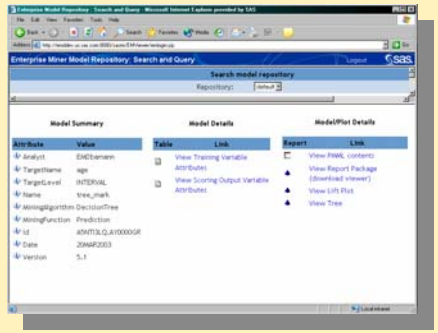
Modellanwendung (z.B. mit MS Office Add-In)



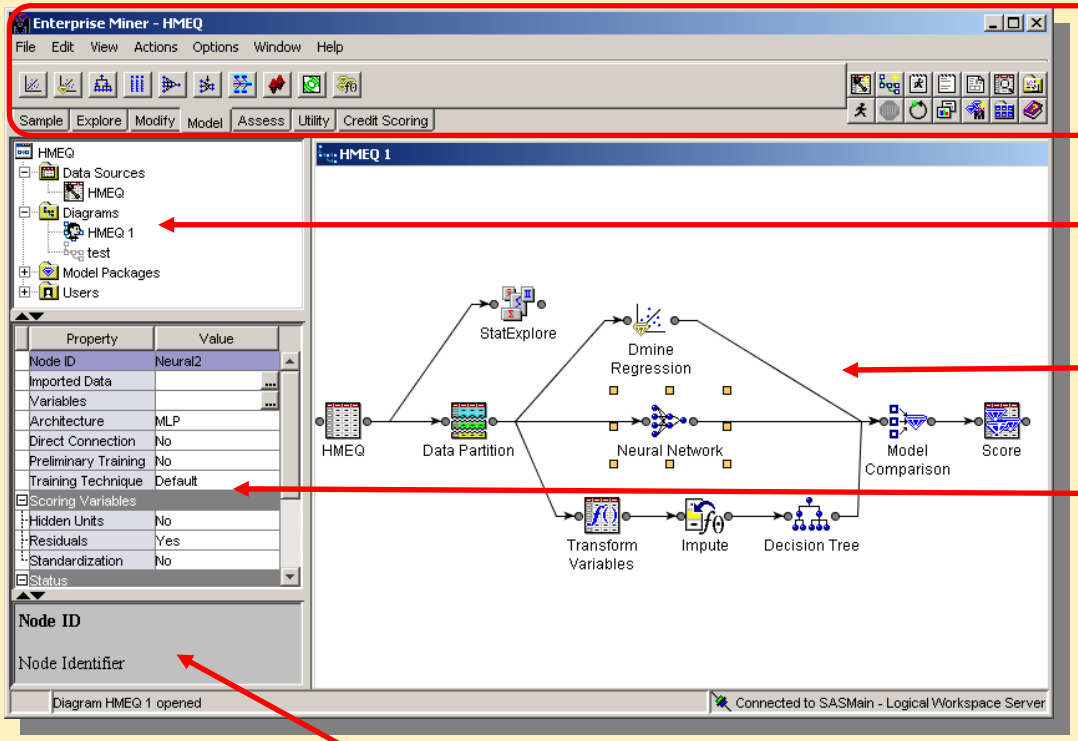
5a. Anzeige der Scoring-Ergebnisse

5b. Verteilung von Reports und Anzeige im Web

Modell-Management (Model Repository Viewer)



Neue Java-basierte Benutzeroberfläche



Neue gestaltete Menü- und Symbolleisten

Projektverwaltung und Datenquellen-Definition

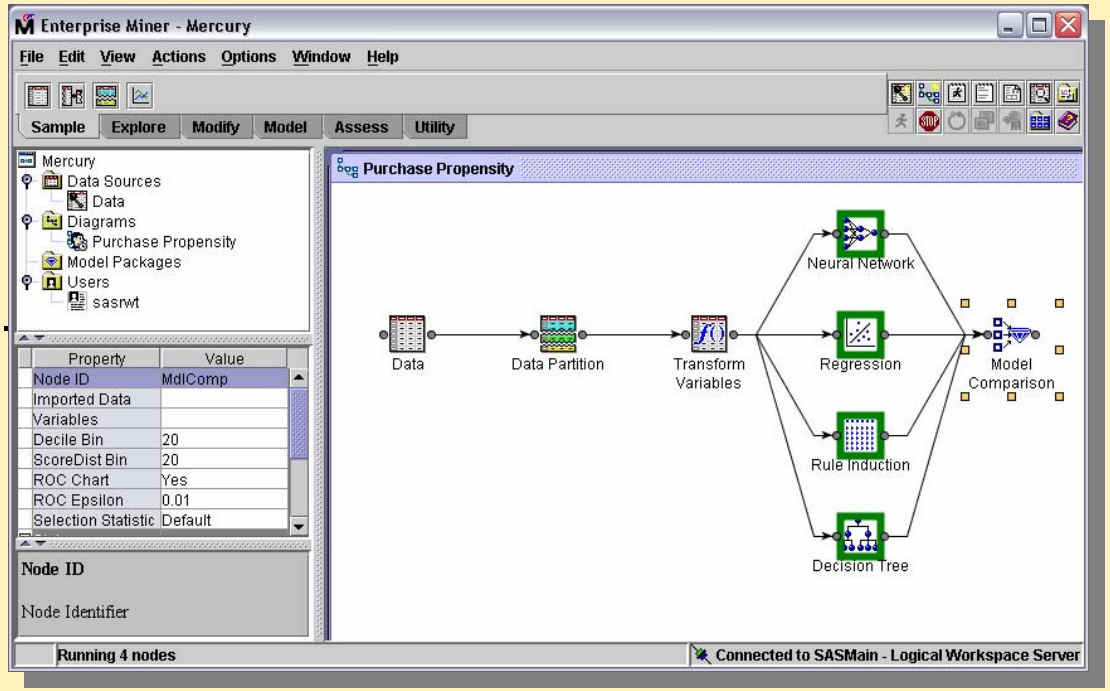
Prozessflussdiagramm als Arbeitsbereich

Eigenschaften für einzelne Objekte (Knoten, Daten...)

Kurzer Hilfetext zur Erläuterung

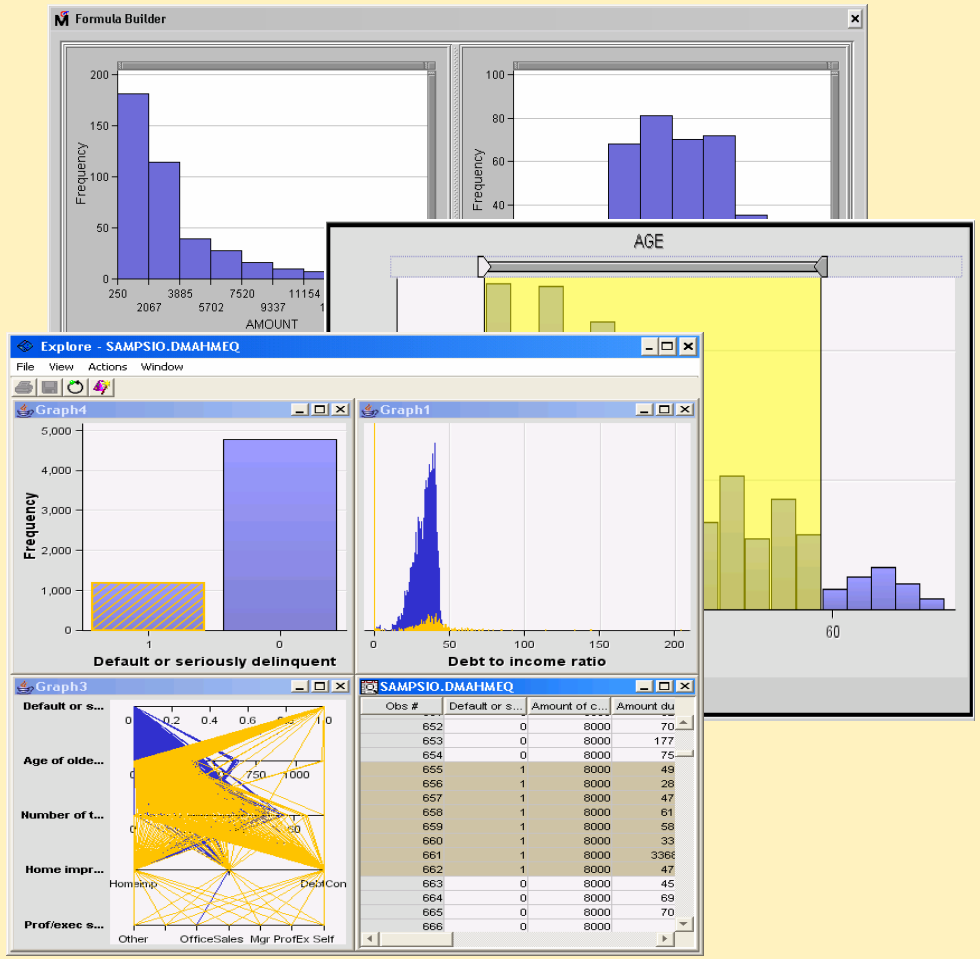
Verbesserte Performance und Skalierbarkeit

- Server-seitige Projektverwaltung
- Paralleles Verarbeiten mehrerer Modelle
- Unterstützung von GRID-Computing
- Algorithmen mit Multi-Threading
- Asynchrones Modell-Training
- Sauberer Abbruch der Verarbeitung
- Disconnect/Reconnect aus lfd. Verarbeitung

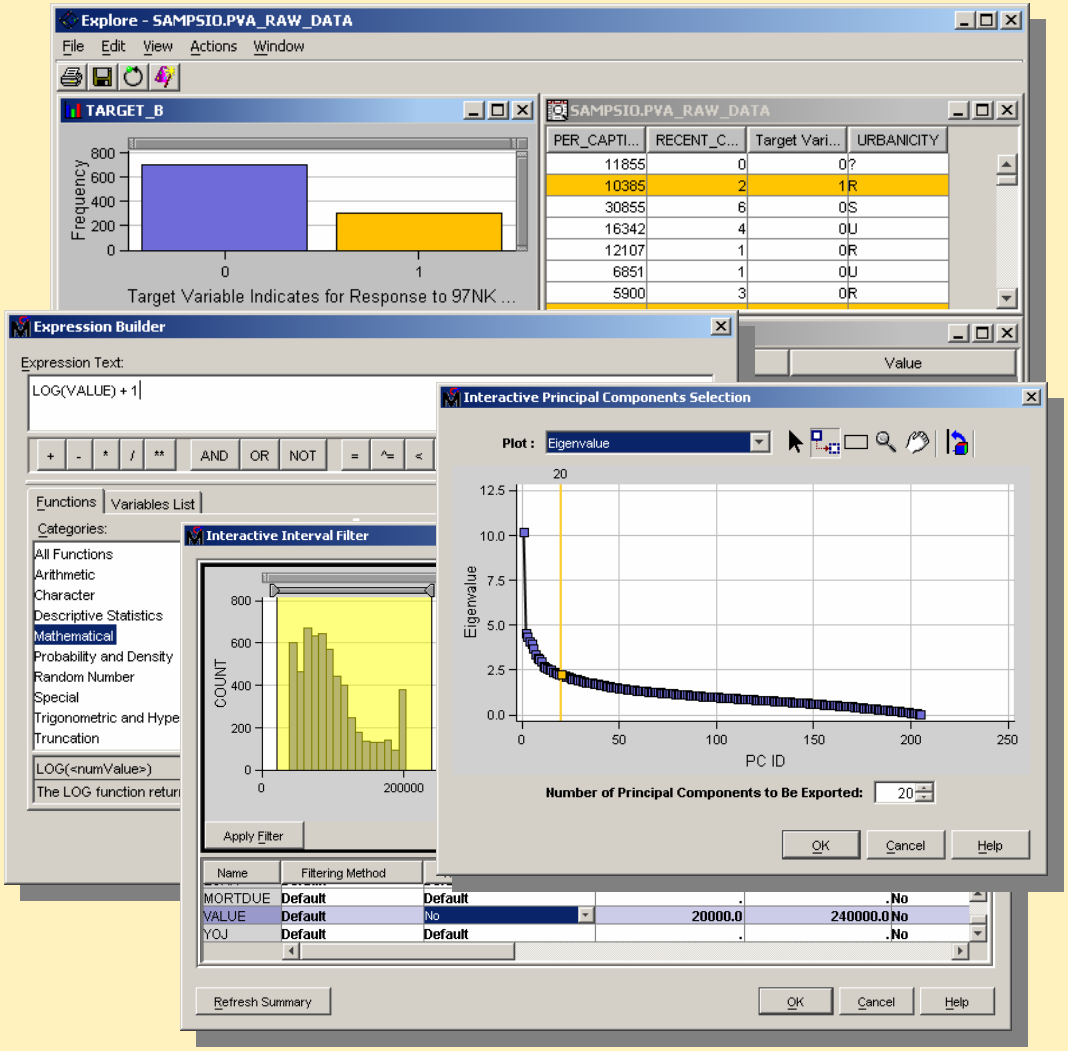


Datenexploration und Datenmodifikation

- Partitionieren von Daten (Training/Validation/Test)
- Deskriptive Statistik und explorative Grafiken
- Werteersetzung, Behandlung fehlender Werte
- Interaktive Variablentrans-formationen
- Variablenauswahl und Datenfilter
- Zusammenführen von Dateien und Stichproben
- Definieren von Entschei-dungsprofilen für Zielgrößen
- ...

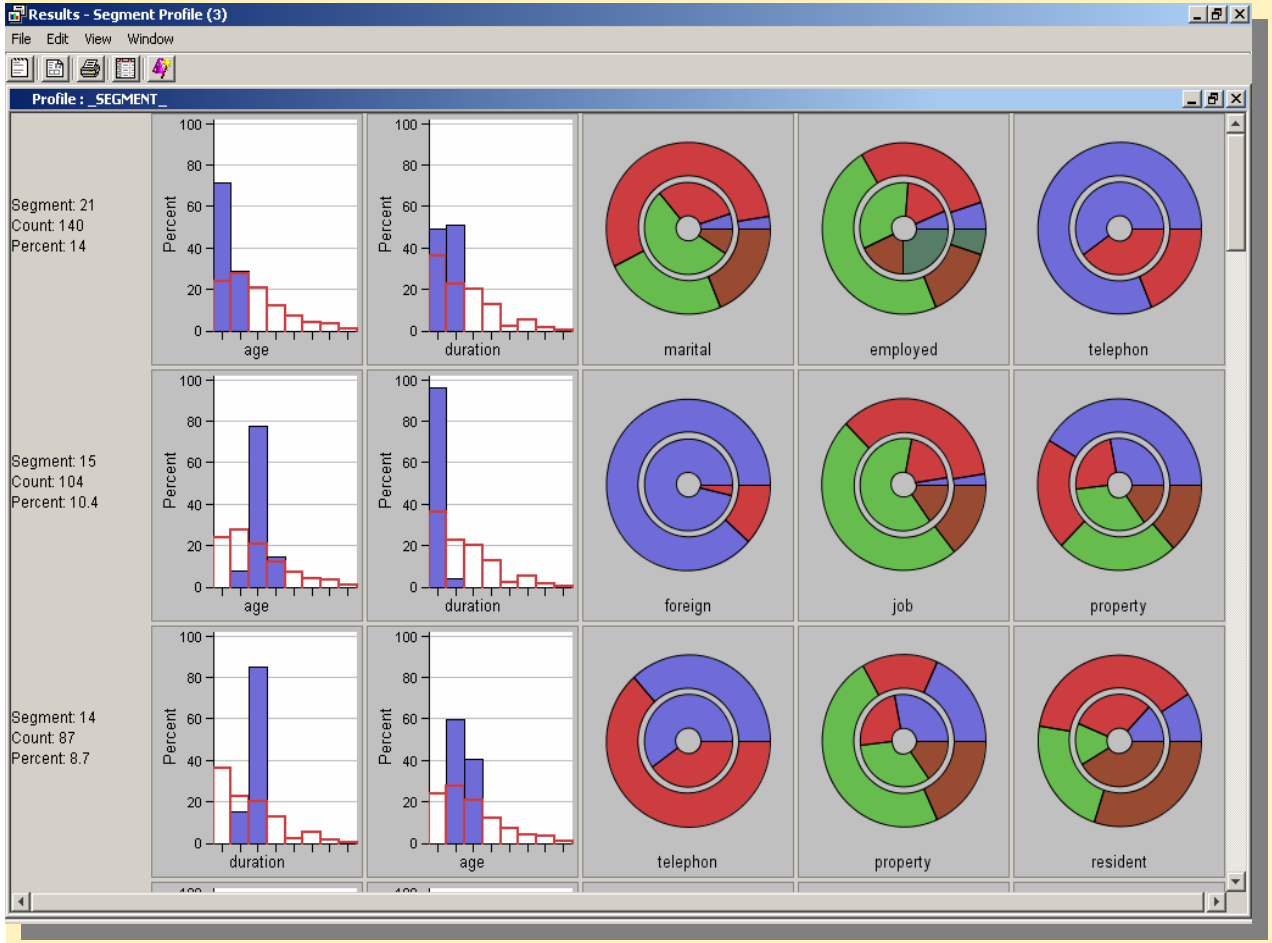


Interaktive Steuerungsmöglichkeiten



- Dynamisch verknüpfte Grafik- und Tabelleninhalte
- Flexible Grafiken über neuen Java Graphics Wizard
- Editor für benutzerdefinierte Variablentransformationen
- Benutzerdefinierte Bereichs-selektion im Filter-Knoten
- Auswahl beizubehaltender Hauptkomponenten in PCA

Neuer Segment Profiler für Visualisierung von Clusterergebnissen



Ermöglicht leichtere Interpretation von Segmentierungen

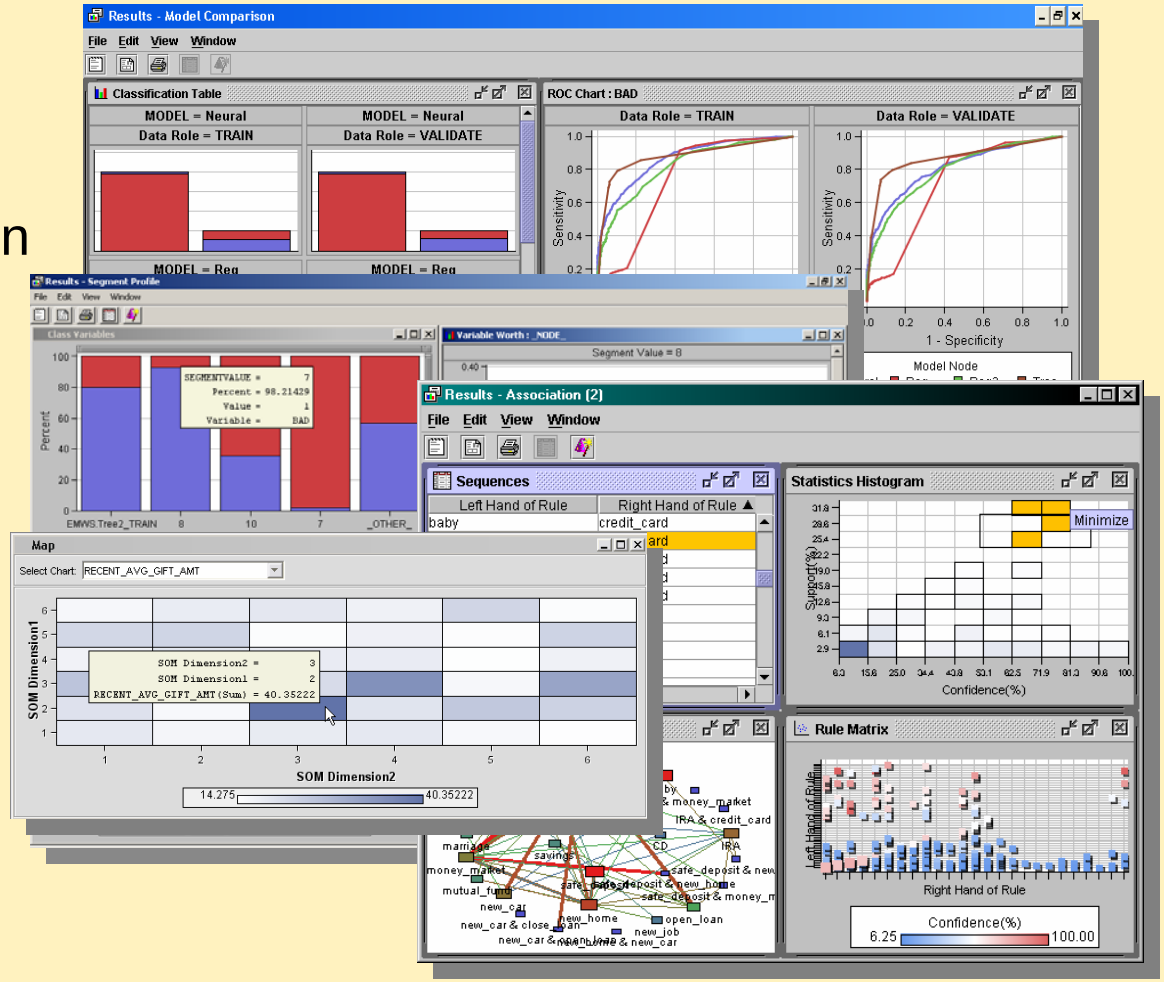
Ermittlung der Wichtigkeit (Worth) von Merkmalen für einzelne Segmente

Merkmalsprofile der jeweiligen Segmente im Vergleich zum Total

- Histogramme für kontinuierliche Variablen
- Kreisdiagramme für kategoriale Variablen

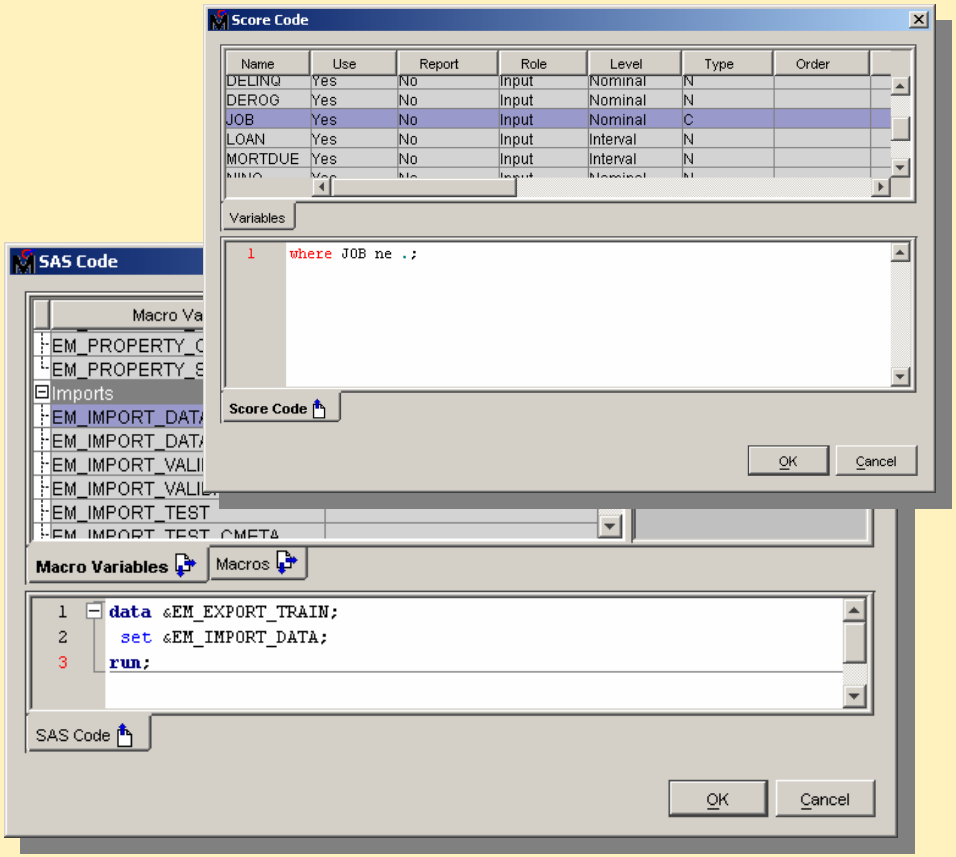
Modellbewertung und Vergleich

- Lift-Diagramme
- ROC-Kurven
- Modellgüte-Statistiken
- Iterationsverläufe
- Segmentprofile
- Link Graph
- Regelmatrix
- Topologische Karte
- …



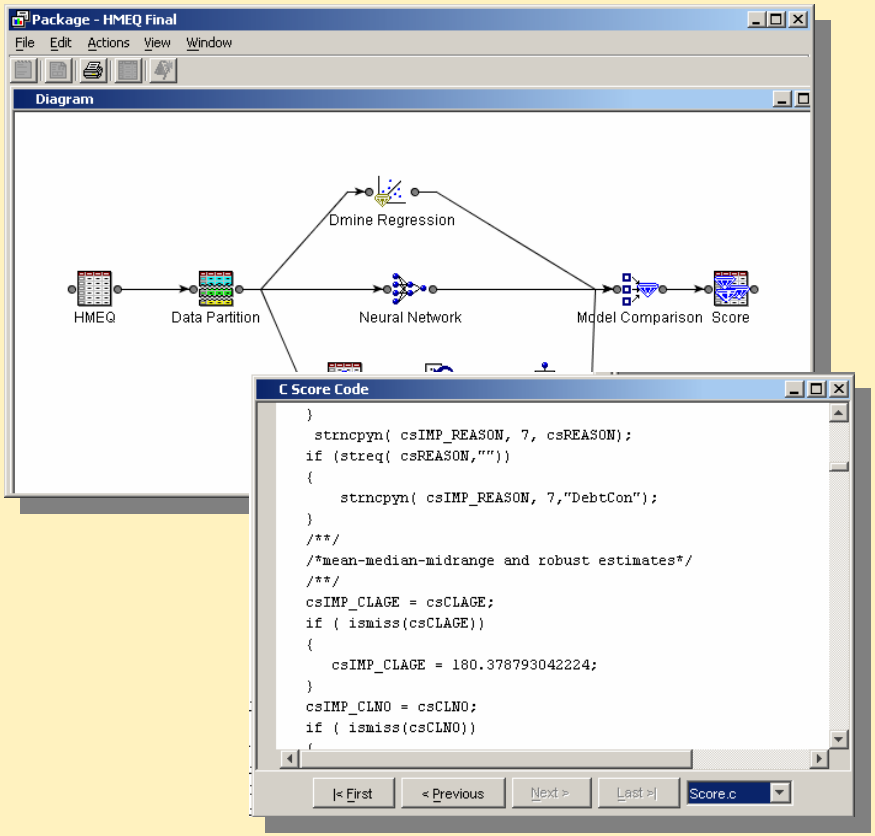
Erweiterbarkeit durch eigenen SAS Code

- Code-Knoten für Zugriff auf SAS Prozeduren
- Referenzieren von Makro-variablen per Drag & Drop
- Ermöglicht benutzerdefinierte Funktionalität für
  - Reporting
  - Datentransformationen
  - Analytische Verfahren
- Generieren eigener Knoten-erweiterungen



Score Code für Modellanwendung

- Export von Score Code nach SAS, C, Java und PMML
- Berücksichtigung sämtlicher Verarbeitungsschritte im Code
- Registrieren als Model Package im SAS Metadaten-Server
- Score-Code-Ausführung über Plug-In in SAS ETL Studio



Agenda

Der Neue Text Miner 2.3

Sprachunterstützung für 8 Sprachen

- Englisch
- Französisch
- Deutsch
- Niederländisch
- Italienisch
- Portugiesisch
- Spanisch
- Schwedisch

Macros zur Textbereinigung

```
%textsyn(docds=EMWS.TEXT2_DOCUMENTS,
         termds= EMWS.TEXT2_TERMS,
         outds= EMWS.TEXT2_OUT_T,
         synds=sasuser.autosyns,
         dict=data.engdict,
         textvar=DESCRIPTION);
```

EXAMPLE1	EXAMPLE2	TERM	PARENT	CAT...	CHILDNDO...	NUMDOCS	MINSPED
REPOSITIONED LEFT SIDE !!A.!! C. _ H		a.	a		1.0	97.0	-0.0
... DAMPER NOISE REPLACE SHOCK !!A		absorbe	absorber		1.0	5.0	4.0
... REPL. R. RR. SHOCK !!ABSORBR!!		absorbr	absorber		1.0	5.0	7.0
... THE LF FRT WHEN !!ACC!! LYBED THE	RADIO INOP !!ACC!! SOCK	acc	ac		2.0	20.0	12.0
!!ACCELERATOR!! ...	WHEN PRESSING !!ACCE	accelerator	acceleration		4.0	5.0	13.0
!!ACCELORATOR!! STICKS.		accelorator	accelerator		1.0	4.0	9.0
... SHIFT AND VIBRATES ON !!ACEL!! FRC		acel	accel		1.0	9.0	6.0
... REPL !!ACTUATER!! UNIT. RETEST.		actuater	actuator		1.0	6.0	12.0
REAR HATCH RATTLES _ !!ADJD!! HATC		adjd	adj		1.0	38.0	10.0
... UP MOLYKOATED FRONT PADS !!ADJS	... DROVE SEVERAL TIMES	adjsuted	adjusted		2.0	23.0	6.0
REPLACE SEAT !!ADJUSTER!!		adjuster	adjusted		1.0	23.0	12.0
... 50MPH TIRES OUT B !!AL!!ANCE BALA		al	all		1.0	128.0	12.0
... NEC TO SUBLET FOR !!ALIG!!	... SUBLET FOR 4 WHEEL	alig	align		3.0	154.0	8.0
STEERING WHEEL NOT CENTERED, !!A		aligh	alig		1.0	3.0	8.0
... _ _ _ NOT !!ALIGHNED!! AND END CA		alighned	aligned		1.0	7.0	7.0
!!ALRAM!! REMOTE INOP, REPLACED TF		alram	alarm		1.0	7.0	10.0
LEAK FROM THE !!ALXE!! BOOT		alxe	axle		1.0	5.0	12.0
SPLR GRNSH CMNG !!APRT!! _ ROK SI		aprt	apart		1.0	12.0	12.0
WHEEL !!BALAN!!CE4 WHEN DRIVING 6		balan	balance		1.0	15.0	14.0
... _ DR. , RAD. !!BALANC!!		balanc	balance		1.0	15.0	5.0
... TO TRAILER HITCH WIRING, !!BARE!! V		bare	bar		1.0	3.0	10.0
... CAR WONT START CK !!BAT!!. FAILED		bat	batt		1.0	4.0	8.0

Nutzungsmöglichkeiten von Perl Regular Expressions über den Code Knoten

Perl Regular Expression Quick Reference 1.03

N.B.: this quick reference is just that - some of the explanations have been simplified. For the authoritative documentation, see the latest edition of *Programming Perl* or `perldoc perlre`.

Specific characters:

`\t`	A tab character
`\n`	A newline character (OS neutral)
`\r`	A carriage return character
`\f`	A form feed character
`\cX`	Control character `CTRL-X`
`\NNN`	Octal code for character *NNN*

Metacharacters:

The following 12 characters need to be escaped with a backslash - "\" - because by default, they mean something special.

`\ | ( ) [ { ^ $ * + ? .`

`.`	Match one character
`|`	Alternation
`( )`	Group and capture
`[ ]`	Define character class
`\`	Modify the meaning of the next char.

Anchors:

`^`	Match at the beginning of a string (or line)
`$`	Match at the end of a string (or line)
`\b`	Match at a 'word' boundary
`\B`	Match at not a 'word' boundary

Quantifiers:

These quantifiers apply to the proceeding *atom*.

`*`	Match 0 or more times
`+`	Match 1 or more times
`?`	March 0 or 1 times
`{N}`	Match exactly *N* times
`{N,}`	Match at least *N* times
`{N,M}`	Match at least *N* but not more than *M* times

By default, quantifiers are "*greedy*". They attempt to match as many characters as possible. In order to make them match as few characters as possible, follow them with a question mark "?".

Character class metacharacters:

`^`	If the first character of a class, negates that class
`-`	Unless first or last character of a class, used for a range

Character class shortcuts:

`\d`	`[0-9]`	A digit
`\D`	`[^0-9]`	A non-digit
`\s`	`[ \t\n\r\f]`	A whitespace char.
`\S`	`[^ \t\n\r\f]`	A non-whitespace char.
`\w`	`[a-zA-Z0-9_]`	A 'word' char.
`\W`	`[^a-zA-Z0-9_]`	A 'non-word' char.

These shortcuts can be used either on their own, or within a character class.

Metaquote & case translations:

`\Q`	Quote (de-meta) characters until `\E`
`\U`	Uppercase characters until `\E`
`\L`	Lowercase characters until `\E`

Special variables:

`$``	The characters to the left of the match
`$&`	The characters that matched
`$'`	The characters to the right of the match
`\N`	The characters captured by the N^{th} set of parentheses (if on the match side)
`$N`	The characters captured by the N^{th} set of parentheses (if not on the match side)

Modifiers:

These modifiers apply to the entire pattern

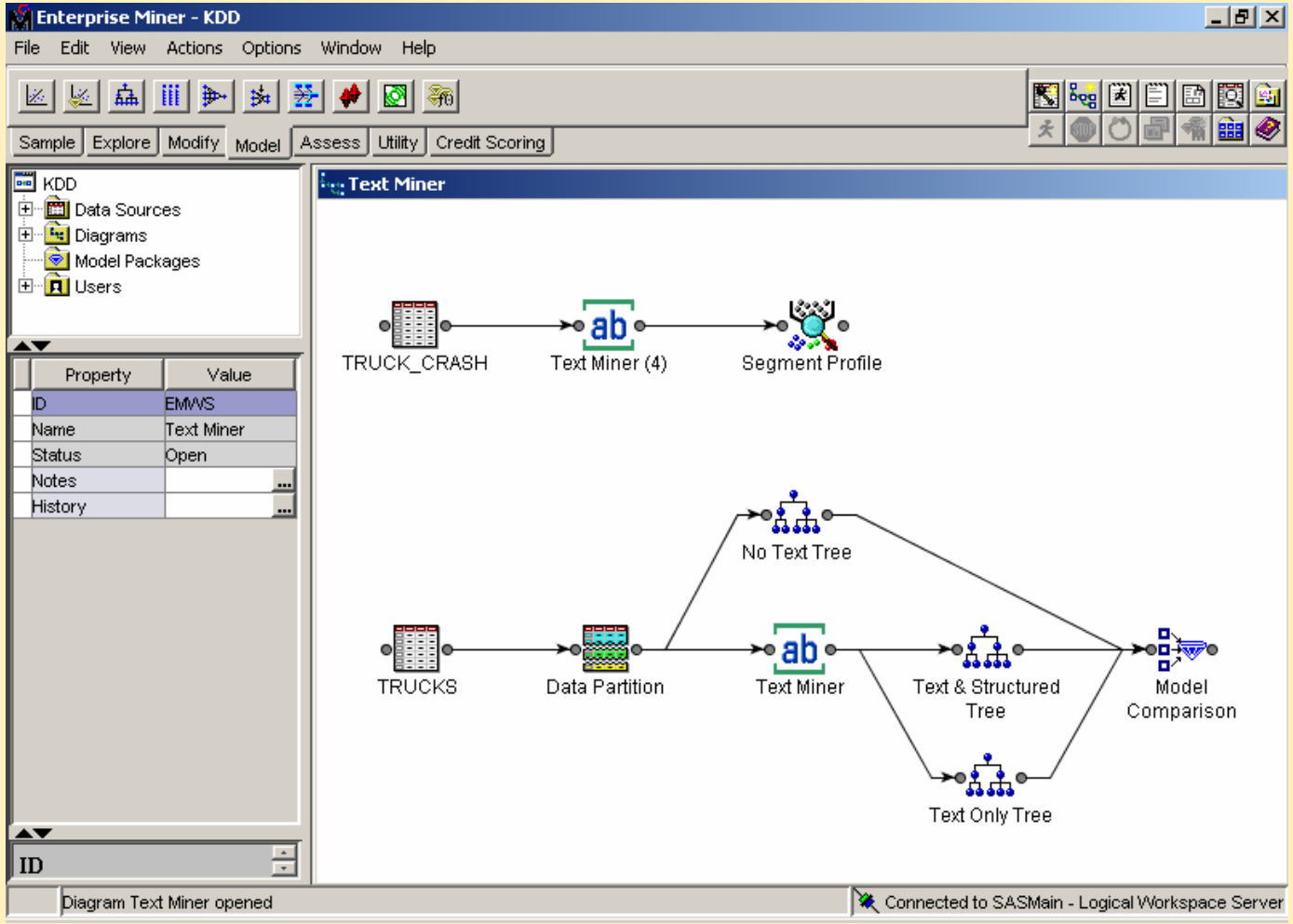
`/i`	Ignore case
`/g`	Match globally (all)
`/m`	Let ^ and $ match next to embedded `\n`
`/s`	Let . match `\n`
`/x`	Ignore most whitespace and allow comments
`/e`	Evaluate right hand side of `s///` as an expression

All except `/e` apply to both `m//` and `s///`.

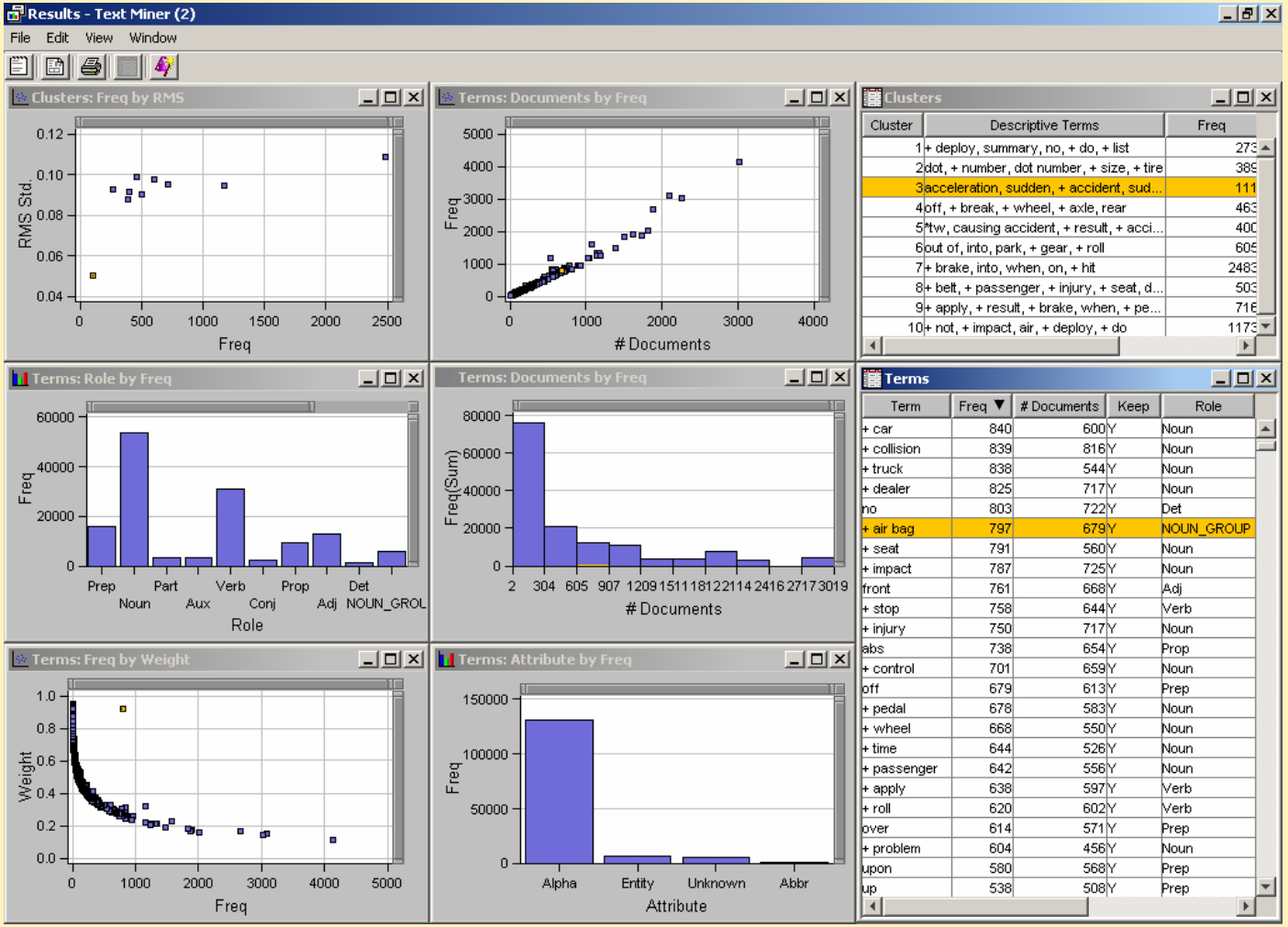
Binding operators:

`=~`	True if the regex matches
`!~`	True if the regex doesn't match

Integration in das Prozessflussdiagramm der Enterprise Miner Oberfläche



Text Miner Static Results Browser

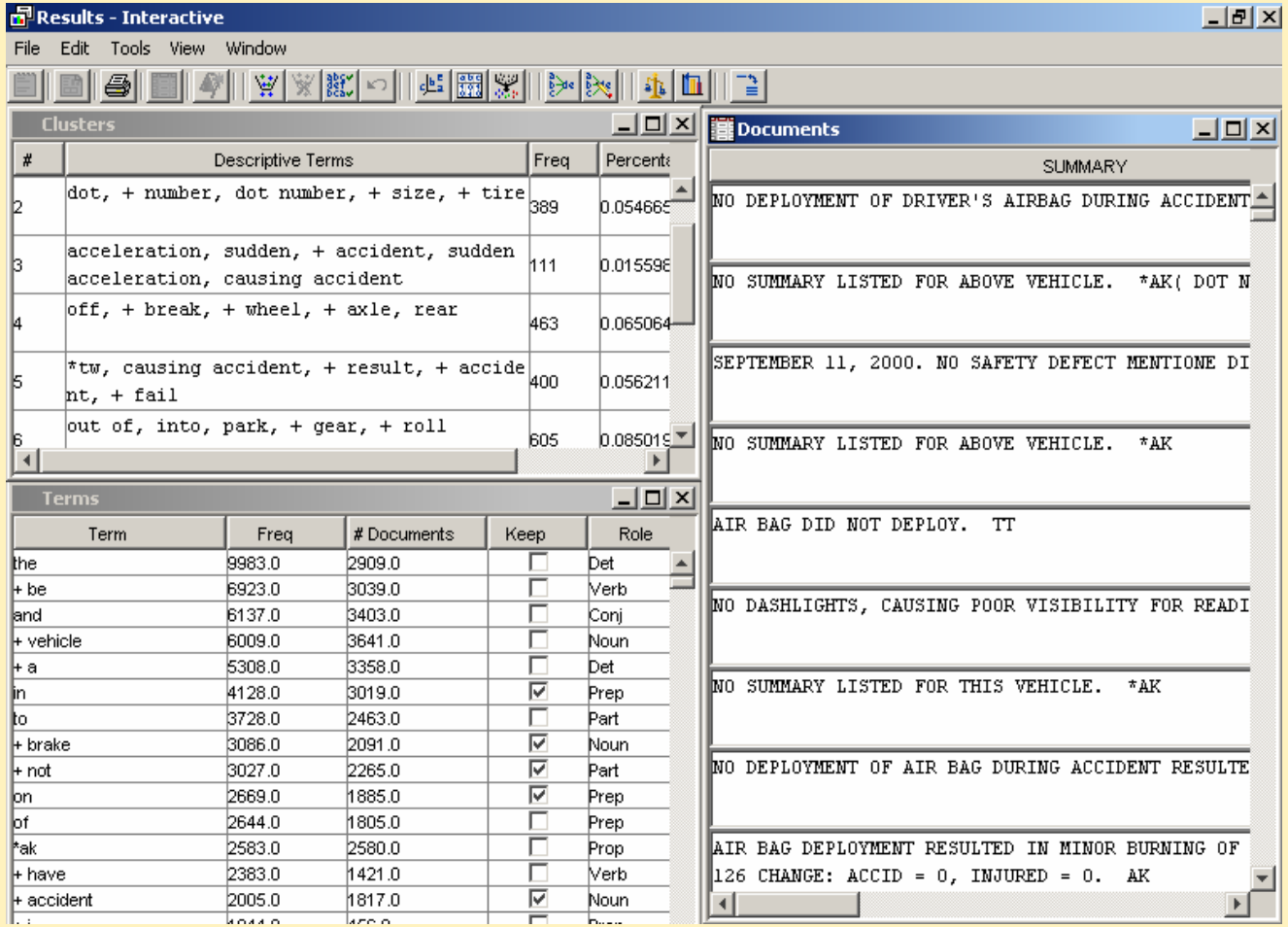


Interaktives Ergebnisfenster

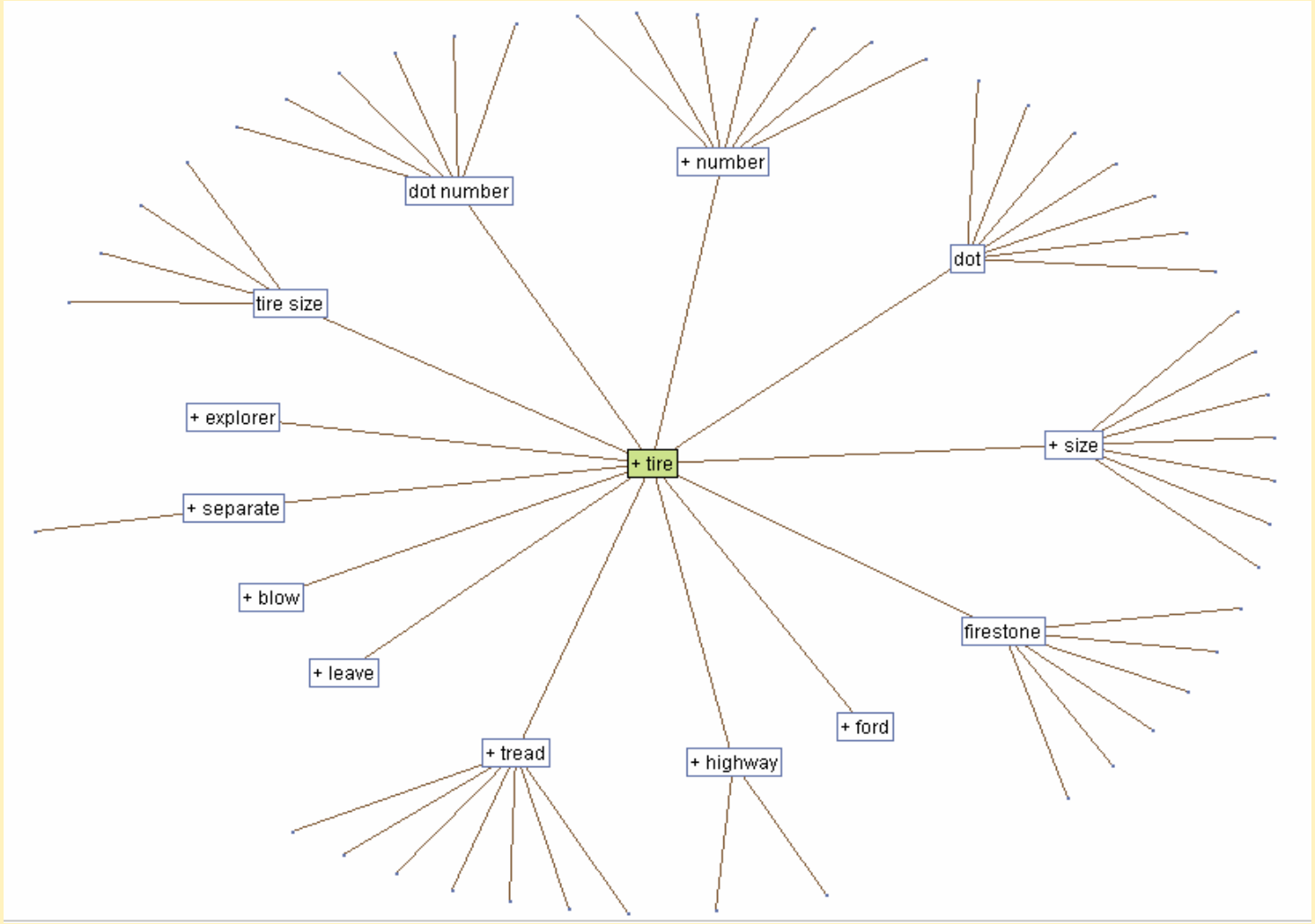
Clusters Fenster

Terms Fenster

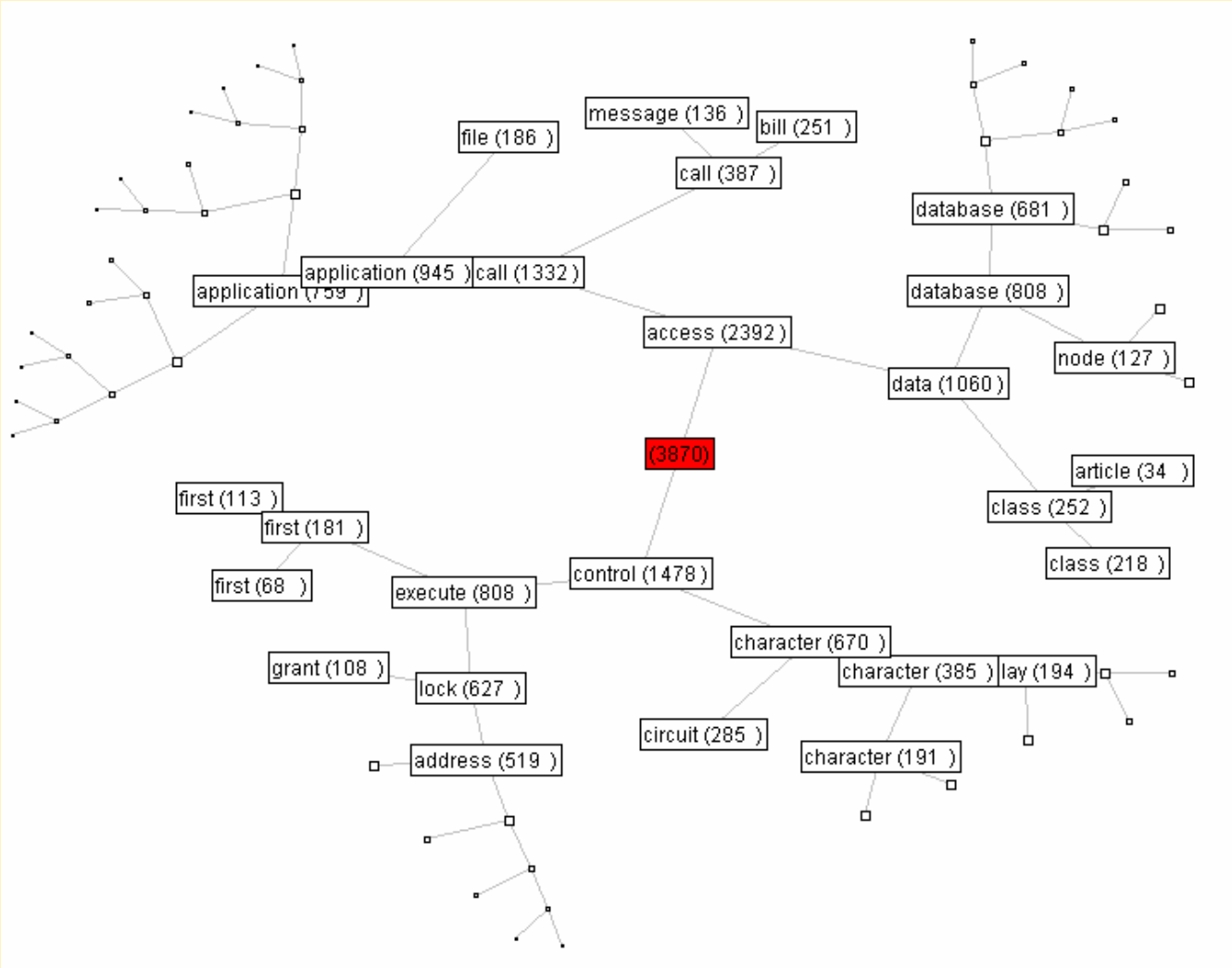
Dokument Fenster



Konzept Linking: Gefundene Beziehung zwischen dem Begriff “tire” und anderen Begriffen



Hierarchisches Dokument Clustering



Zusammenfassung

- Enterprise Miner 5.2
  - Einheitliches Metadaten-Konzept
  - Score Code in SAS, C, Java und PMML
  - Bequeme Verwaltung (Model Packages und Model Repository Viewer)
  - Batch Jobs zur Automatisierung von Data Mining-Abläufen
  - Viele neue Analytische Funktionalitäten
  - Performance
- Text Miner 2.3
  - Nahtlose Integration in den Enterprise Miner Prozessfluss
  - Unterstützung von 8 Sprachen
  - Textbereinigung, Perl Regular Expressions
  - Konzept Linking
  - Hierarchisches Dokument Clustering mit Java Tree

Diskussion