

# Data Preparation for Analytics

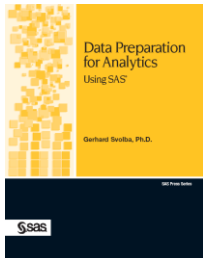
Dr. Gerhard Svolba  
SAS-Österreich, Wien

KSFE  
Aachen, 29. Februar 2008

**THE  
POWER  
TO KNOW®**

# Agenda

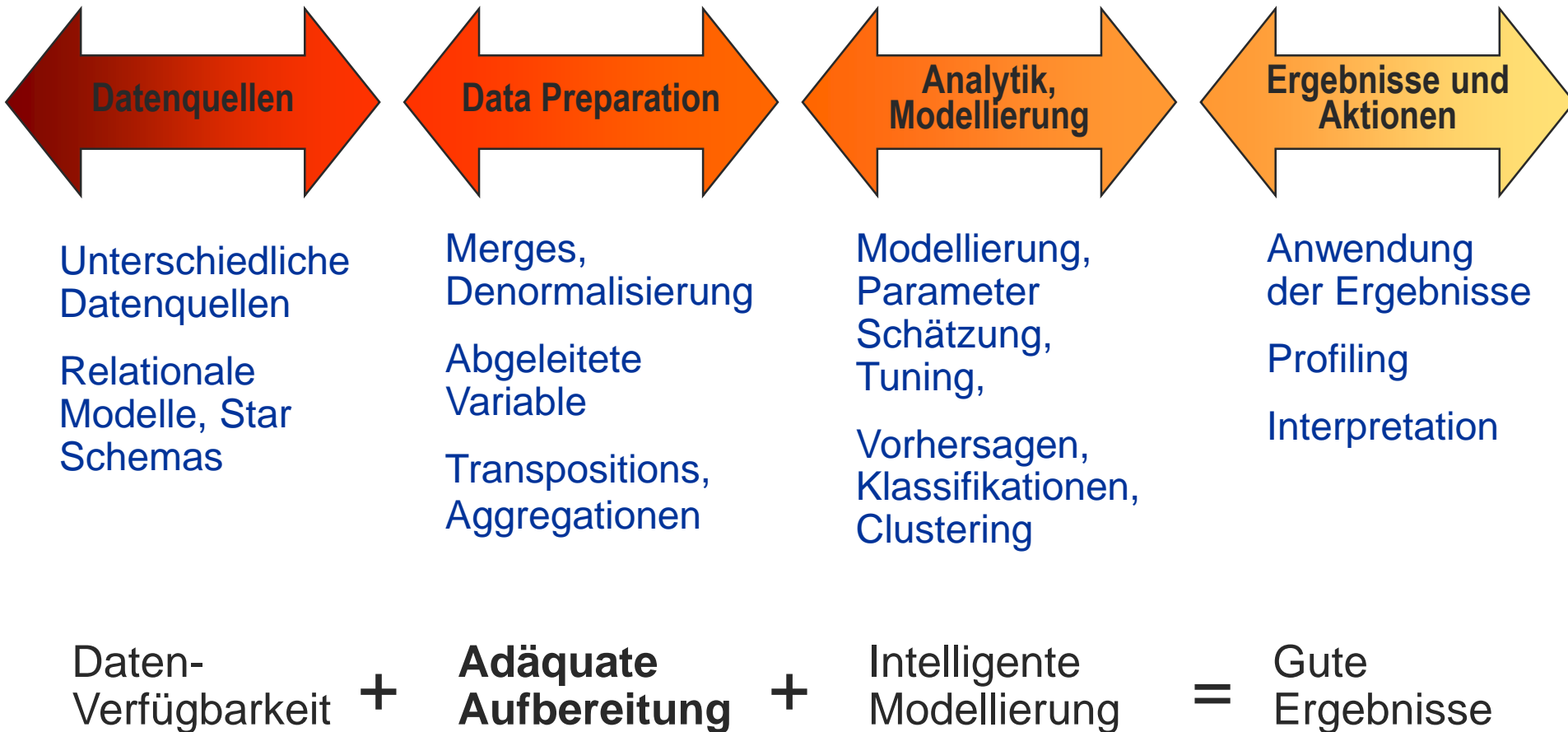
- Data Preparation for Analytics –  
Fachliche Überlegungen
- Datenstrukturen für Analytik
- Fallbeispiel – Aufbau eines  
Analysedatenbestands für Data Mining
- Datenmanagement mit der SAS® Language und  
mit SAS® Tools
- Zusammenfassung



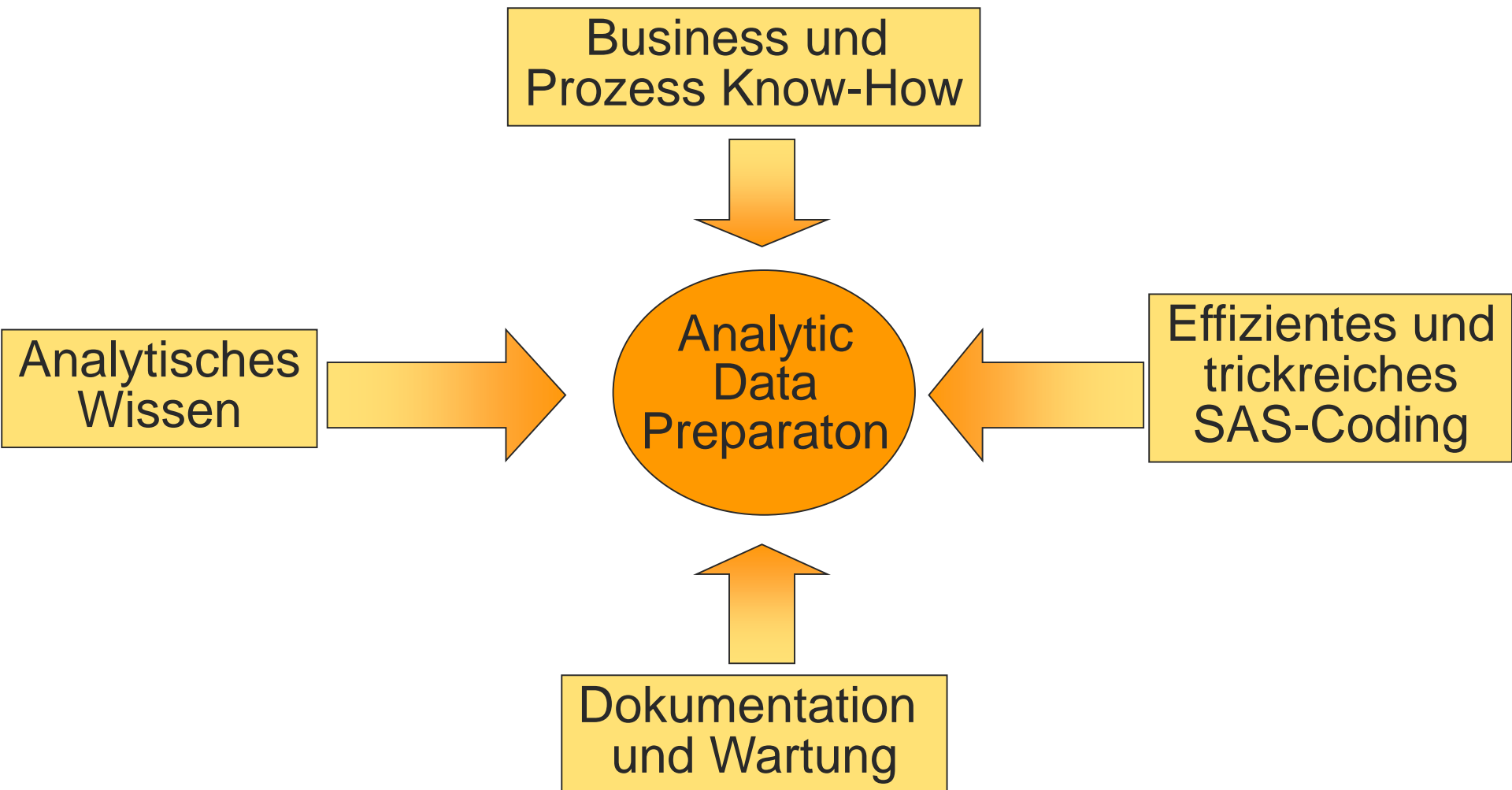
# Einige Worte über Data Preparation

- Ist für Techniker
- Ist nur Programmierung
- Ist fad
- Verbraucht bis zu 80 % des Projektaufwands
- War bisher nicht im Fokus der Data Mining Literatur
- Ist etwas, das SAS perfekt kann
- Ist enorm wichtig für den Projekterfolg

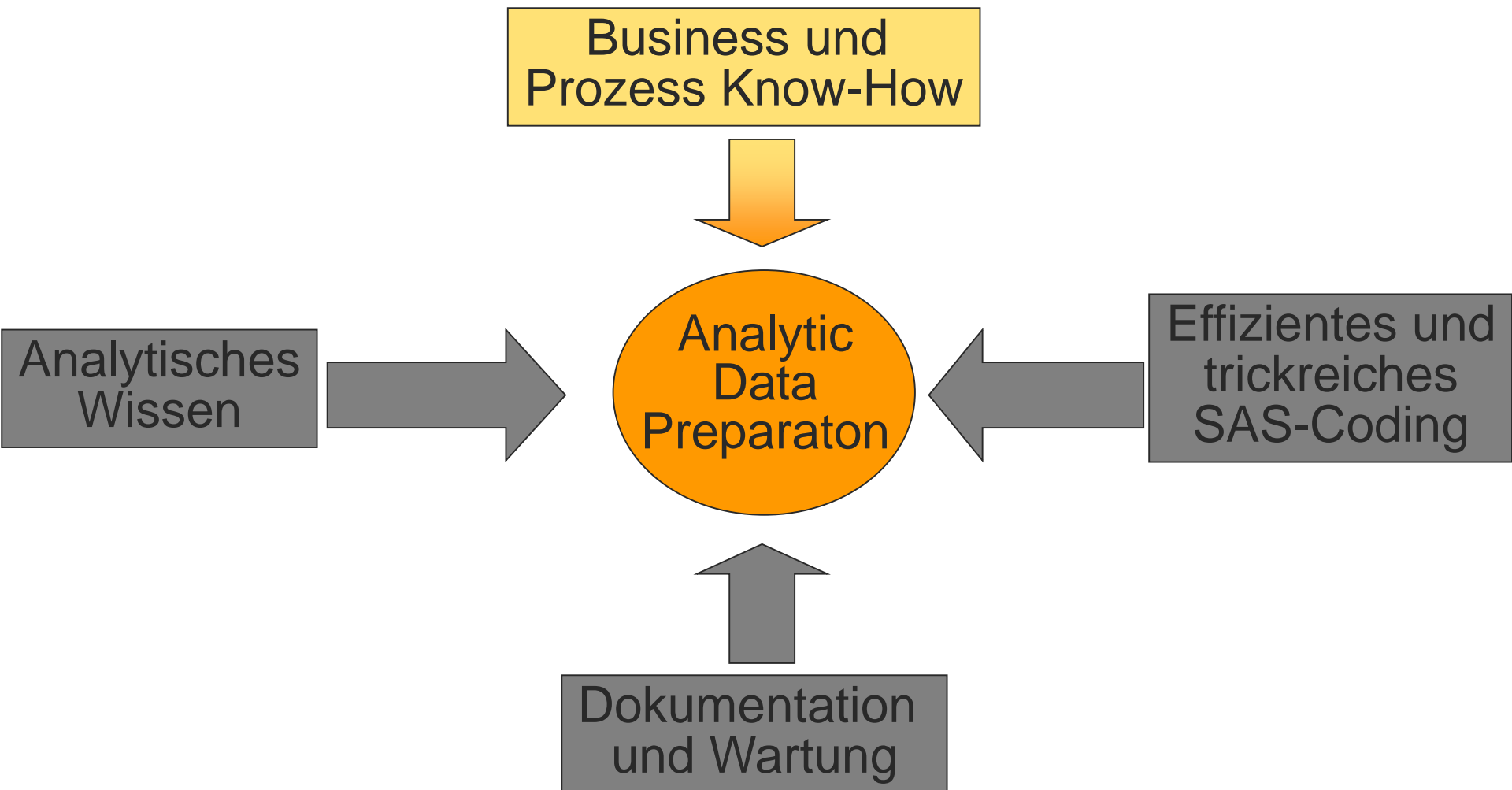
# Der Analyseprozess: Von Rohdaten zu verwertbaren Ergebnissen



# Vier Dimensionen für Data Preparation



# Vier Dimensionen für Data Preparation



# Hinterfragen einer Business-Fragestellung

- Kannst Du mir die Wahrscheinlichkeit ausrechnen, daß ein Kunde Produkt X kündigt?
- Fragen
  - Kündigung oder Nutzungsreduktion?
  - Wie behandeln wir die Nutzung von höherwertigen Produkten?
  - Schliessen wir unfreiwillige Kündigungen ein?
  - Wie schnell können Kundenbindungsmaßnahmen durchgeführt werden?
  - Wird eine Kündigungswahrscheinlichkeit pro Kunde benötigt oder eine Liste der 10.000 gefährdetsten Kunden?

# Fachabteilung, Statistiker und IT: Das optimale Dreieck !?

Ich habe meine Fragen  
formuliert. Gibt es  
irgendeinen Grund, dass  
ich die Antworten nicht  
binnen 2 Tagen bei mir  
habe?



**Simon**  
Retention Manager

**Analytic  
Data  
Preparation**

Ich hätte gerne eine analytische  
Datenbank mit einer grossen  
Anzahl an Attributen und eine  
Umgebung, in der es möglich ist,  
sowohl Datenmanagement und  
Analyse durchzuführen.



**Daniele**  
Analytikerin

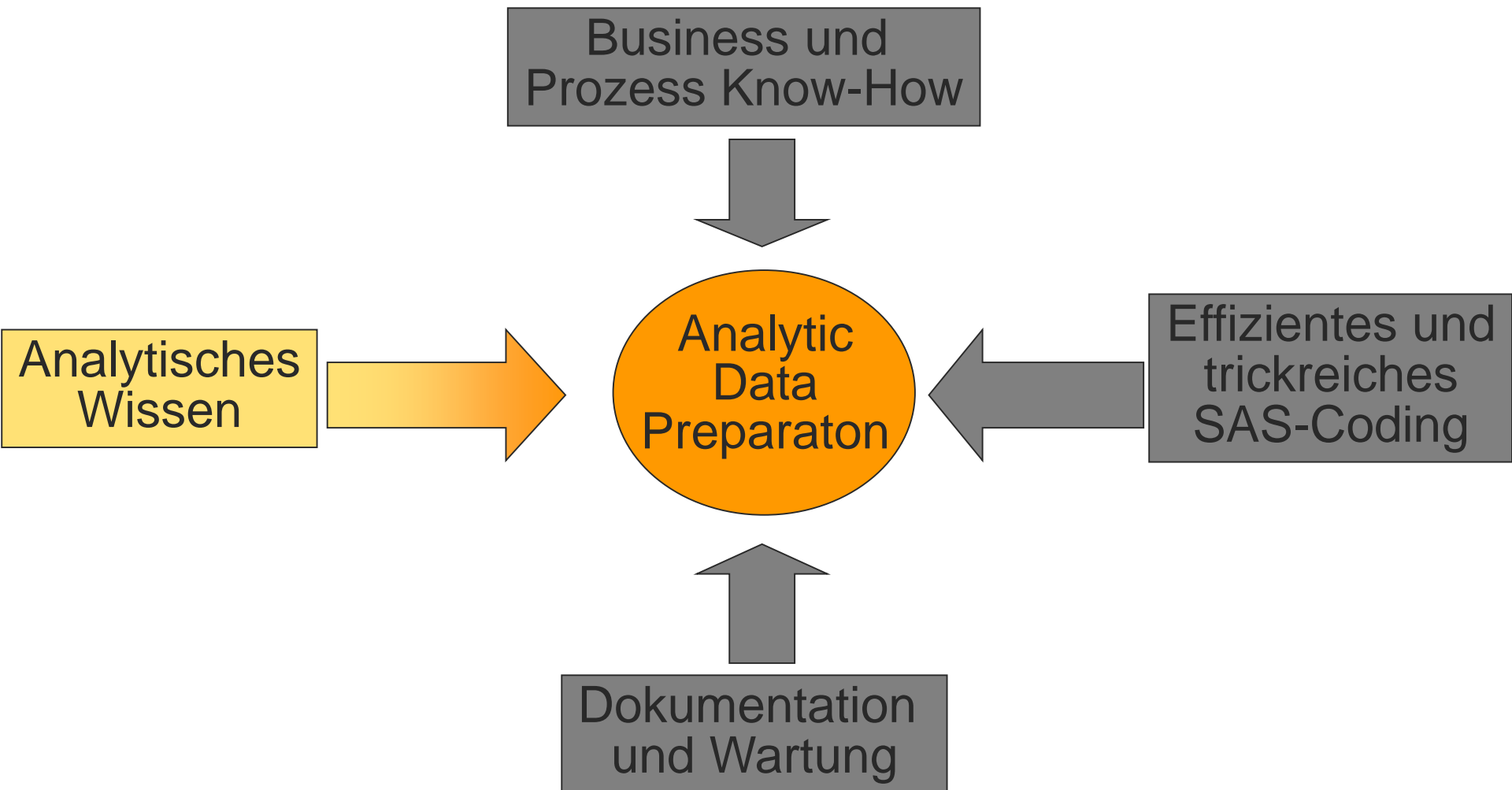
Nenne mir die Liste der  
Attribute und abgeleiteten  
Variablen die Du im Modell  
verwenden wirst und die  
ich regelmässig liefern  
soll.



**Elias**  
IT Experte



# Vier Dimensionen für Data Preparation



# Fragestellung, Analytische Modelle

- Ereignisvorhersage  
(Churn, Betrug, Rückzahlung, Antwort, ...)
- Vorhersage von Werten  
(Kaufhöhe, Schadensbetrag, ...)
- Clustering (Segmentation, ...)
- Market Basket Analysis (Assoziationsanalyse, ...)
- Zeitreihenanalyse

# Analysis Subjects und mehrfache Beobachtungen

- *Analysis Subjects* sind die zentralen Elemente der Analyse in deren Kontext die Ergebnisse der Analyse interpretiert werden.
  
- *Mehrfache Beobachtungen pro Analysis Subject*
  - Wiederholte Beobachtungen über die Zeit
  - Mehrfache Beobachtungen aufgrund hierarchischer Abhängigkeiten

# Haupt-Typen von Datamarts

## One-Row-per-Subject Data Mart

	Customer ID	Date of Birth	Age (years)	Gender	Marital Status	Academic Title	Has Title? 0/1	Branch Name	Customer Start Date	Customer Duration (months)
1	1000002	26DEC1958	44	Male	Married		0	Fil1	01JAN2000	41
2	1000005	25JUN1947	56	Male	Single	Ing.	1	Fil4	01APR1999	50
3	1000006	10DEC1945	57	Female	Married		0	Fil4	01SEP1996	81
4	1000007	02JUN1934	69	Male	Married		0	Fil1	01SEP1997	69
5	1000008	15DEC1957	45	Male	Single	Dr.	1	Fil3	01JAN1996	89
6	1000009	11MAR1959	44	Male	Single		0	Fil2	01JUL2001	23
7	1000014	23AUG1952	51	Male	Single		0	Fil4	01MAY1996	85
8	1000015	12MAY1959	44	Male	Single		0	Fil2	01FEB1999	52
9	1000016	11FEB1967	36	Male	Married		0	Fil2	01FEB2001	28

## Multiple-Row-per-Subject Data Mart

	CUSTOMER	TIME	PRODUCT
1	0	0	hering
2	0	1	comed_b
3	0	2	olives
4	0	3	ham
5	0	4	turkey
6	0	5	bourbon
7	0	6	ice_crea
8	1	0	baguette
9	1	1	soda
10	1	2	hering
11	1	3	cracker
12	1	4	heineken
13	1	5	olives
14	1	6	comed_b
15	2	0	avocado
16	2	1	cracker
17	2	2	artichok
18	2	3	heineken
19	2	4	ham
20	2	5	turkey
21	2	6	sardines

## Longitudinal Data Mart

	Date	ELECTRO	GARDENING	TOOLS
1	15/08/05	15725	13913	9441
2	16/08/05	15120	16315	9922
3	17/08/05	16631	18996	11345
4	19/08/05	18080	16325	9326
5	20/08/05	15604	14690	9108
6	21/08/05	14518	14388	9371
7	22/08/05	13048	15249	8390
8	23/08/05	13857	13974	10982
9	24/08/05	14869	15704	12104
10	26/08/05	12262	13836	8112
11	27/08/05	15011	13438	8599
12	28/08/05	13612	12625	8389
13	29/08/05	11546	13566	8249
14	30/08/05	21352	16918	13337
15	31/08/05	22900	20813	14099
16	02/09/05	15333	15626	8896
17	03/09/05	13156	13306	8082
18	04/09/05	19294	16361	16267
19	05/09/05	15917	15587	15539

# Data Mart Strukturen

	Data Mart Structure für die Analyse	
Struktur der Quelldaten: “Existieren mehrfache Beobachtungen pro analysis subject?”	One-Row-per-Subject Data Mart	Multiple-Row-per-Subject Data Mart
NEIN		
JA		

# Terminologie

	ID	TIME	WEIGHT
1	1	1	77
2	1	2	79
3	1	3	83
4	2	1	62
5	2	2	58
6	2	3	59
7	3	1	99
8	3	2	97
9	3	3	92

- Multiple-rows-per-subject data mart
- Univariate data set
- LONG data set

	ID	weight1	weight2	weight3
1	1	77	79	83
2	2	62	58	59
3	3	99	97	92

- One-row-per-subject data set
- Multivariate data set
- WIDE data set

# Beispiel

	ID	Drug	Depleted	Histamine	Measurement
1	1	Morphine	N	0.04	0
2	1	Morphine	N	0.2	1
3	1	Morphine	N	0.1	3
4	1	Morphine	N	0.08	5
5	2	Morphine	N	0.02	0
6	2	Morphine	N	0.06	1
7	2	Morphine	N	0.02	3
8	2	Morphine	N	0.02	5
9	3	Morphine	N	0.07	0
10	3	Morphine	N	1.4	1
11	3	Morphine	N	0.48	3
12	3	Morphine	N	0.24	5
13	4	Morphine	N	0.17	0
14	4	Morphine	N	0.57	1
15	4	Morphine	N	0.35	3
16	4	Morphine	N	0.24	5
17	5	Morphine	Y	0.1	0
18	5	Morphine	Y	0.09	1
19	5	Morphine	Y	0.13	3
20	5	Morphine	Y	0.14	5

```
%MAKEWIDE (DATA=dogs_long,
            OUT=dogs_wide,
            ID=id,
            COPY=drug depleted,
            VAR=Histamine,
            TIME=Measurement);
```

	ID	Drug	Depleted	Histamine0	Histamine1	Histamine3	Histamine5
1	1	Morphine	N	0.04	0.2	0.1	0.08
2	2	Morphine	N	0.02	0.06	0.02	0.02
3	3	Morphine	N	0.07	1.4	0.48	0.24
4	4	Morphine	N	0.17	0.57	0.35	0.24
5	5	Morphine	Y	0.1	0.09	0.13	0.14

# Exercise

	ID	Drug	Depleted	Histamine0	Histamine1	Histamine3	Histamine5
1	1	Morphine	N	0.04	0.2	0.1	0.08
2	2	Morphine	N	0.02	0.06	0.02	0.02
3	3	Morphine	N	0.07	1.4	0.48	0.24
4	4	Morphine	N	0.17	0.57	0.35	0.24
5	5	Morphine	Y	0.1	0.09	0.13	0.14

```
%MAKELONG(DATA=dogs_wide,
           OUT=Dogs_long_from_wide,
           ID=id,
           COPY=drug Depleted,
           ROOT=Histamine,
           MEASUREMENT=Measurement);
```

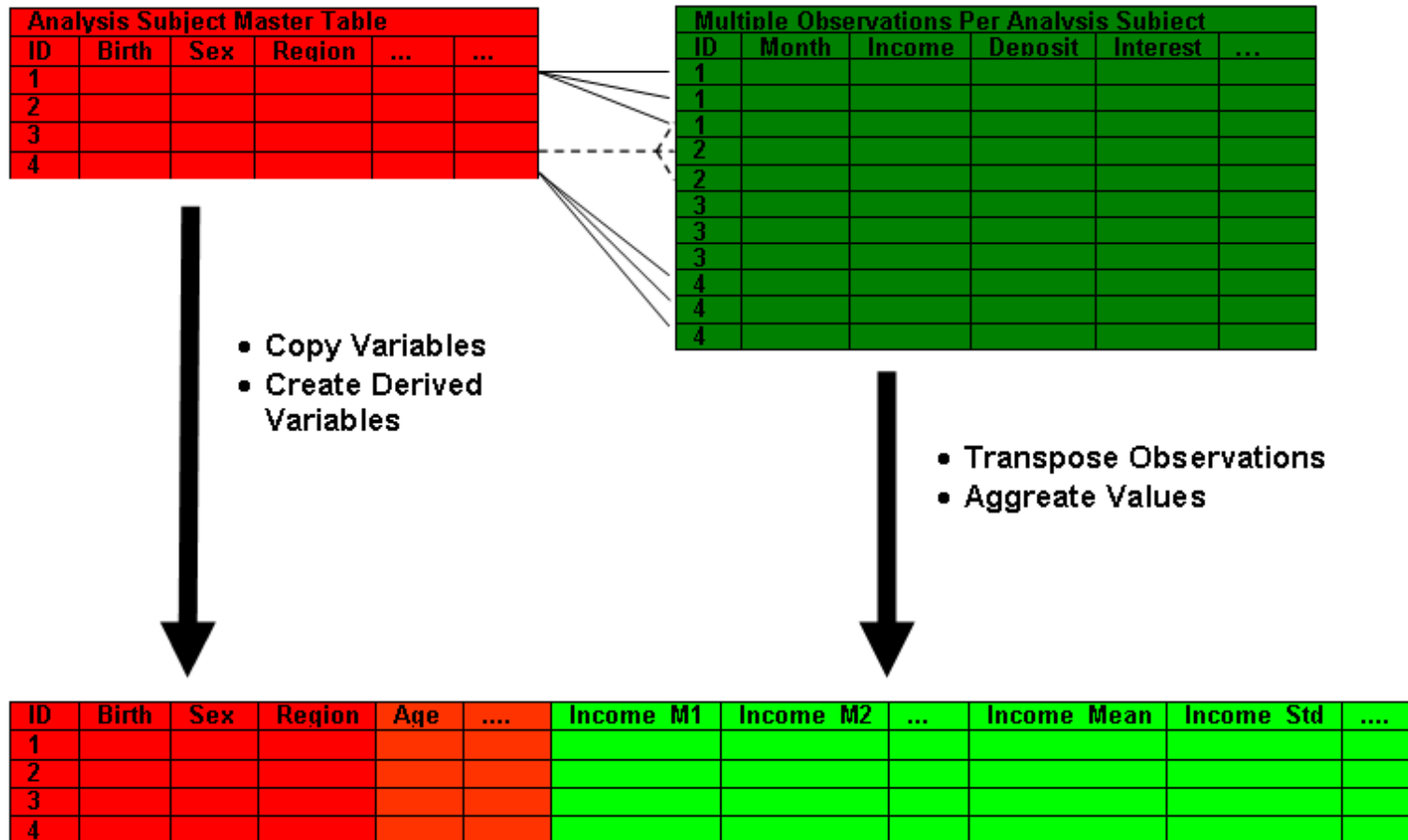
	ID	Drug	Depleted	Histamine	Measurement
1	1	Morphine	N	0.04	0
2	1	Morphine	N	0.2	1
3	1	Morphine	N	0.1	3
4	1	Morphine	N	0.08	5
5	2	Morphine	N	0.02	0
6	2	Morphine	N	0.06	1
7	2	Morphine	N	0.02	3
8	2	Morphine	N	0.02	5
9	3	Morphine	N	0.07	0
10	3	Morphine	N	1.4	1
11	3	Morphine	N	0.48	3
12	3	Morphine	N	0.24	5
13	4	Morphine	N	0.17	0
14	4	Morphine	N	0.57	1
15	4	Morphine	N	0.35	3
16	4	Morphine	N	0.24	5
17	5	Morphine	Y	0.1	0
18	5	Morphine	Y	0.09	1
19	5	Morphine	Y	0.13	3
20	5	Morphine	Y	0.14	5



# Der One-Row-Per-Subject Data Mart

- Wird von vielen statistischen Verfahren benötigt
  - Regressionsanalyse, Neuronale Netzwerke, Entscheidungsbäume, Survivalanalyse, Clusteranalyse, ...
- Prominenteste Data Mart Struktur in Data Mining
  - Ereignisvorhersage  
(Churn, Betrug, Rückzahlung, Antwort, ...)
  - Vorhersage von Werten  
(Kaufhöhe, Schadensbetrag, ...)
  - Clustering (Segmentation, ...)

# Das One-Row-Per-Subject Paradigma

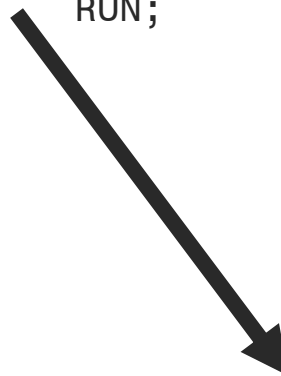


# Transposing der Daten nach „One-Row-Per-Subject“

	ID	TIME	WEIGHT
1	1	1	77
2	1	2	79
3	1	3	83
4	2	1	62
5	2	2	58
6	2	3	59
7	3	1	99
8	3	2	97
9	3	3	92

```
PROC TRANSPOSE DATA = long
                OUT = wide(DROP = _name_)
                PREFIX = weight;

    BY id ;
    VAR weight;
    ID time;
RUN;
```



	ID	weight1	weight2	weight3
1	1	77	79	83
2	2	62	58	59
3	3	99	97	92

# Intelligente und trickreiche Aggregationen

Multiple Observations Per Analysis Subject

ID	Month	Income	Deposit	Interest	...
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					

- Transpose Observations
- Aggregate Values

Income M1	Income M2	...	Income Mean	Income Std	....

## Interval Daten

- Statische Aggregationen
- Korrelation von Werten
- Verläufe über die Zeit
- Konzentration von Werten

## Kategorielle Daten

- Häufigkeitszählungen
- „Concatenated Frequencies“
- Total und Distinct Counts

# Korrelation von Werten

	CustID	Month	Usage
1	1	1	52
2	1	2	54
3	1	3	58
4	1	4	47
5	1	5	38
6	1	6	22
7	2	1	22
8	2	2	24
9	2	3	30
10	2	4	28
11	2	5	31
12	2	6	30

Wie korrelieren die monatlichen Werte pro Kunde mit dem Gesamtmittelwert pro Monat?

	CustID	Usage
1	1	0.26
2	2	-0.81
3	3	0.64
4	4	0.45
5	5	0.09
6	6	-0.17
7	7	0.21
8	8	0.18
9	9	.
10	10	0.72

# Kenngrößen für den Zeitverlauf

	CustID	M1	M2	M3	M4	M5	M6	LongTerm	ShortTerm	LongShortInd
1	1	52	54	58	47	38	22	-5.971428571	-16	--
2	2	22	24	30	28	31	30	1.6857142857	-1	+=
3	3	100	120	110	115	100	95	-2.285714286	-5	--
4	4	43	43	43	.	42	41	-0.395348837	-1	==
5	5	20	29	35	39	28	44	3.4571428571	16	++
6	6	16	24	18	25	30	24	1.8571428571	-6	+-
7	7	80	70	60	50	60	70	-2.571428571	10	-+
8	8	90	95	80	100	100	90	1	-10	=-
9	9	47	47	47	47	47	47	0	0	==
10	10	50	52	0	50	0	52	-2.742857143	52	-+

```

PROC REG DATA = longitud NOPRINT
    OUTEST=Est_LongTerm(KEEP = CustID month
                        RENAME = (month=LongTerm));

MODEL usage = month;
BY CustID;
RUN;

PROC REG DATA = longitud NOPRINT
    OUTEST=Est_ShortTerm(KEEP = CustID month
                        RENAME = (month=ShortTerm));

MODEL usage = month;
BY CustID;
WHERE month in (5 6);
RUN;

```

# Konzentration von Werten

	CustID	ContractID	Usage1
1	1	1	20
2	1	2	40
3	1	3	60
4	1	4	5
5	1	5	2
6	1	6	1
7	2	1	10
8	2	2	10
9	2	3	12
10	2	4	11
11	3	1	40
12	3	2	30
13	3	3	30
14	3	4	10
15	3	5	5
16	4	1	4
17	5	1	1
18	5	2	2
19	5	3	3
20	6	1	1
21	6	2	2
22	6	3	3
23	6	4	4

Konzentration =  
Anteil der Summe der Top 50 %  
Sub-Hierarchien/  
Gesamtsummen über alle  
Sub-Hierarchien

	CustID	usage1_conc
1	1	0.94
2	2	0.53
3	3	0.74
4	4	1.00
5	5	0.67
6	6	0.70

# Kategorische Variable: Häufigkeitszählungen

## Quelldaten

	Cust_id	Account_id	Account_type
1	1	1	SAVING
2	1	2	CHECKING
3	1	3	SAVING
4	1	4	LOAN
5	2	5	CHECKING
6	2	6	SAVING2
7	3	7	LOAN
8	3	8	MORTGAGE
9	3	9	SAVING
10	3	10	CHECKING
11	4	11	CHECKING
12	5	12	LOAN
13	5	13	SAVING
14	5	14	CHECKING
15	5	15	SAVING2
16	5	16	SPECIAL
17	5	17	SAVING
18	5	18	SAVING

## Absolute und relative Häufigkeiten

	Cust_id	CHECKING	LOAN	SAVING	OTHERS	Checking_rel	loan_rel	saving_rel	others_rel
1	1	1	1	2	0	25	25	50	0
2	2	1	0	1	0	50	0	50	0
3	3	1	1	1	1	25	25	25	25
4	4	1	0	0	0	100	0	0	0
5	5	1	1	4	1	14	14	57	14

## Counts und Distinct Counts

	Cust_id	Nr_Account	Distinct_Count	Distinct_Prop	OnlyDistinctAccounts	Possible_Prop	AllPossibleAccounts
1	1	4	3	75.0	0	75.0	0
2	2	2	2	100.0	1	50.0	0
3	3	4	4	100.0	1	100.0	1
4	4	1	1	100.0	1	25.0	0
5	5	7	4	57.1	0	100.0	1

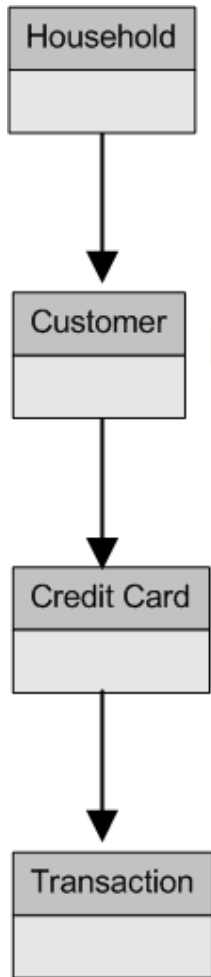


# Kategorische Variable: Concatenated Frequencies

	Cust_id	CHECKING	LOAN	SAVING	OTHERS	Checking_rel	loan_rel	saving_rel	others_rel
1	1	1	1	2	0	25	25	50	0
2	2	1	0	1	0	50	0	50	0
3	3	1	1	1	1	25	25	25	25
4	4	1	0	0	0	100	0	0	0
5	5	1	1	4	1	14	14	57	14

Account_ RowPct	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0_100_0_0	12832	30.61	12832	30.61
100_0_0_0	9509	22.69	22341	53.30
50_0_0_50	4898	11.69	27239	64.98
33_0_0_67	1772	4.23	29011	69.21
0_0_100_0	1684	4.02	30695	73.23
67_0_0_33	1426	3.40	32121	76.63
0_0_50_50	861	2.05	32982	78.69
50_0_50_0	681	1.62	33663	80.31

# Hierarchies: Aggregating Up, Copying Down



# Haupt-Typen von Datamarts

## One-Row-per-Subject Data Mart

	Customer ID	Date of Birth	Age (years)	Gender	Marital Status	Academic Title	Has Title? 0/1	Branch Name	Customer Start Date	Customer Duration (months)
1	1000002	26DEC1958	44	Male	Married		0	Fil1	01JAN2000	41
2	1000005	25JUN1947	56	Male	Single	Ing.	1	Fil4	01APR1999	50
3	1000006	10DEC1945	57	Female	Married		0	Fil4	01SEP1996	81
4	1000007	02JUN1934	69	Male	Married		0	Fil1	01SEP1997	69
5	1000008	15DEC1957	45	Male	Single	Dr.	1	Fil3	01JAN1996	89
6	1000009	11MAR1959	44	Male	Single		0	Fil2	01JUL2001	23
7	1000014	23AUG1952	51	Male	Single		0	Fil4	01MAY1996	85
8	1000015	12MAY1959	44	Male	Single		0	Fil2	01FEB1999	52
9	1000016	11FEB1967	36	Male	Married		0	Fil2	01FEB2001	28

## Multiple-Row-per-Subject Data Mart

	CUSTOMER	TIME	PRODUCT
1	0	0	hering
2	0	1	comed_b
3	0	2	olives
4	0	3	ham
5	0	4	turkey
6	0	5	bourbon
7	0	6	ice_crea
8	1	0	baguette
9	1	1	soda
10	1	2	hering
11	1	3	cracker
12	1	4	heineken
13	1	5	olives
14	1	6	comed_b
15	2	0	avocado
16	2	1	cracker
17	2	2	artichok
18	2	3	heineken
19	2	4	ham
20	2	5	turkey
21	2	6	sardines

## Longitudinal Data Mart

	Date	ELECTRO	GARDENING	TOOLS
1	15/08/05	15725	13913	9441
2	16/08/05	15120	16315	9922
3	17/08/05	16631	18996	11345
4	19/08/05	18080	16325	9326
5	20/08/05	15604	14690	9108
6	21/08/05	14518	14388	9371
7	22/08/05	13048	15249	8390
8	23/08/05	13857	13974	10982
9	24/08/05	14869	15704	12104
10	26/08/05	12262	13836	8112
11	27/08/05	15011	13438	8599
12	28/08/05	13612	12625	8389
13	29/08/05	11546	13566	8249
14	30/08/05	21352	16918	13337
15	31/08/05	22900	20813	14099
16	02/09/05	15333	15626	8896
17	03/09/05	13156	13306	8082
18	04/09/05	19294	16361	16267
19	05/09/05	15917	15587	15539

# Types of Multiple-Rows-per-Subject Data Mart

- no numeration of the multiple rows
- an ordinal numerator variable or sequence variable
- an interval scaled variable such as a time variable (time series data) → LONGITUDINAL DATA MARTS

	CUSTOMER	PRODUCT	Segment
1	213	baguette	SILVER
2	213	hering	SILVER
3	213	avocado	SILVER
4	213	artichok	SILVER
5	213	heineken	SILVER
6	213	chicken	SILVER
7	213	coke	SILVER
8	217	baguette	GOLD
9	217	hering	GOLD
10	217	avocado	GOLD
11	217	artichok	GOLD
12	217	heineken	GOLD
13	217	apples	GOLD
14	217	peppers	GOLD
15	221	soda	SILVER
16	221	olives	SILVER
17	221	bourbon	SILVER
18	221	cracker	SILVER
19	221	heineken	SILVER
20	221	turkey	SILVER
21	221	steak	SILVER

	ID	Name	Key	Value
1	1	Alice	Sex	F
2	1	Alice	Age	13
3	1	Alice	Height	56.5
4	1	Alice	Weight	84
5	2	Barbara	Sex	F
6	2	Barbara	Age	13
7	2	Barbara	Height	65.3
8	2	Barbara	Weight	98
9	3	Carol	Sex	F
10	3	Carol	Age	14
11	3	Carol	Height	62.8
12	3	Carol	Weight	102.5
13	4	Jane	Sex	F
14	4	Jane	Age	12
15	4	Jane	Height	59.8
16	4	Jane	Weight	84.5

	CUSTOMER	TIME	PRODUCT
1	0	0	hering
2	0	1	comed_b
3	0	2	olives
4	0	3	ham
5	0	4	turkey
6	0	5	bourbon
7	0	6	ice_crea
8	1	0	baguette
9	1	1	soda
10	1	2	hering
11	1	3	cracker
12	1	4	heineken
13	1	5	olives
14	1	6	comed_b

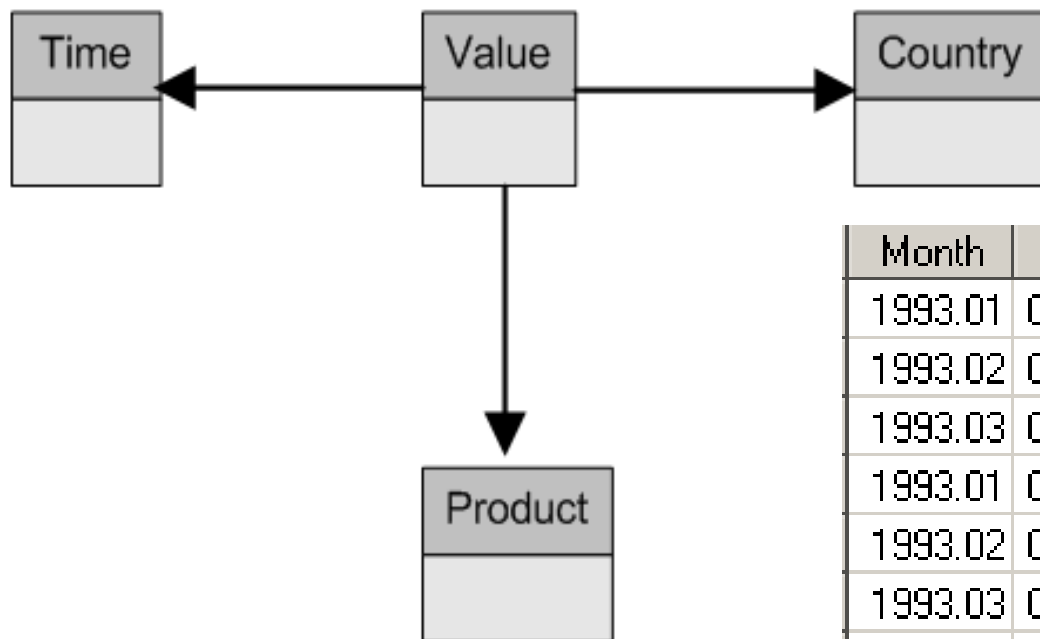
- The standard form for longitudinal data

	Period	Actual Sales	Predicted Sales
1	1993.01	\$29,813.00	\$32,385.00
2	1993.02	\$29,584.00	\$29,163.00
3	1993.03	\$29,873.00	\$31,818.00
4	1993.04	\$30,581.00	\$27,429.00
5	1993.05	\$31,617.00	\$30,263.00
6	1993.06	\$33,605.00	\$27,634.00
7	1993.07	\$33,578.00	\$33,220.00
8	1993.08	\$31,160.00	\$28,874.00
9	1993.09	\$28,696.00	\$28,470.00
10	1993.10	\$31,355.00	\$30,262.00
11	1993.11	\$27,659.00	\$31,434.00
12	1993.12	\$31,956.00	\$29,259.00

- The inter-leaved longitudinal data mart

	Period	Variable	Value
1	1993.01	ACTUAL	\$29,813.00
2	1993.01	PREDICT	\$32,385.00
3	1993.02	ACTUAL	\$29,584.00
4	1993.02	PREDICT	\$29,163.00
5	1993.03	ACTUAL	\$29,873.00
6	1993.03	PREDICT	\$31,818.00
7	1993.04	ACTUAL	\$30,581.00
8	1993.04	PREDICT	\$27,429.00
9	1993.05	ACTUAL	\$31,617.00
10	1993.05	PREDICT	\$30,263.00
11	1993.06	ACTUAL	\$33,605.00
12	1993.06	PREDICT	\$27,634.00
13	1993.07	ACTUAL	\$33,578.00
14	1993.07	PREDICT	\$33,220.00
15	1993.08	ACTUAL	\$31,160.00
16	1993.08	PREDICT	\$28,874.00
17	1993.09	ACTUAL	\$28,696.00
18	1993.09	PREDICT	\$28,470.00
19	1993.10	ACTUAL	\$31,355.00
20	1993.10	PREDICT	\$30,262.00
21	1993.11	ACTUAL	\$27,659.00
22	1993.11	PREDICT	\$31,434.00
23	1993.12	ACTUAL	\$31,956.00
24	1993.12	PREDICT	\$29,259.00

# The cross-sectional dimension data mart

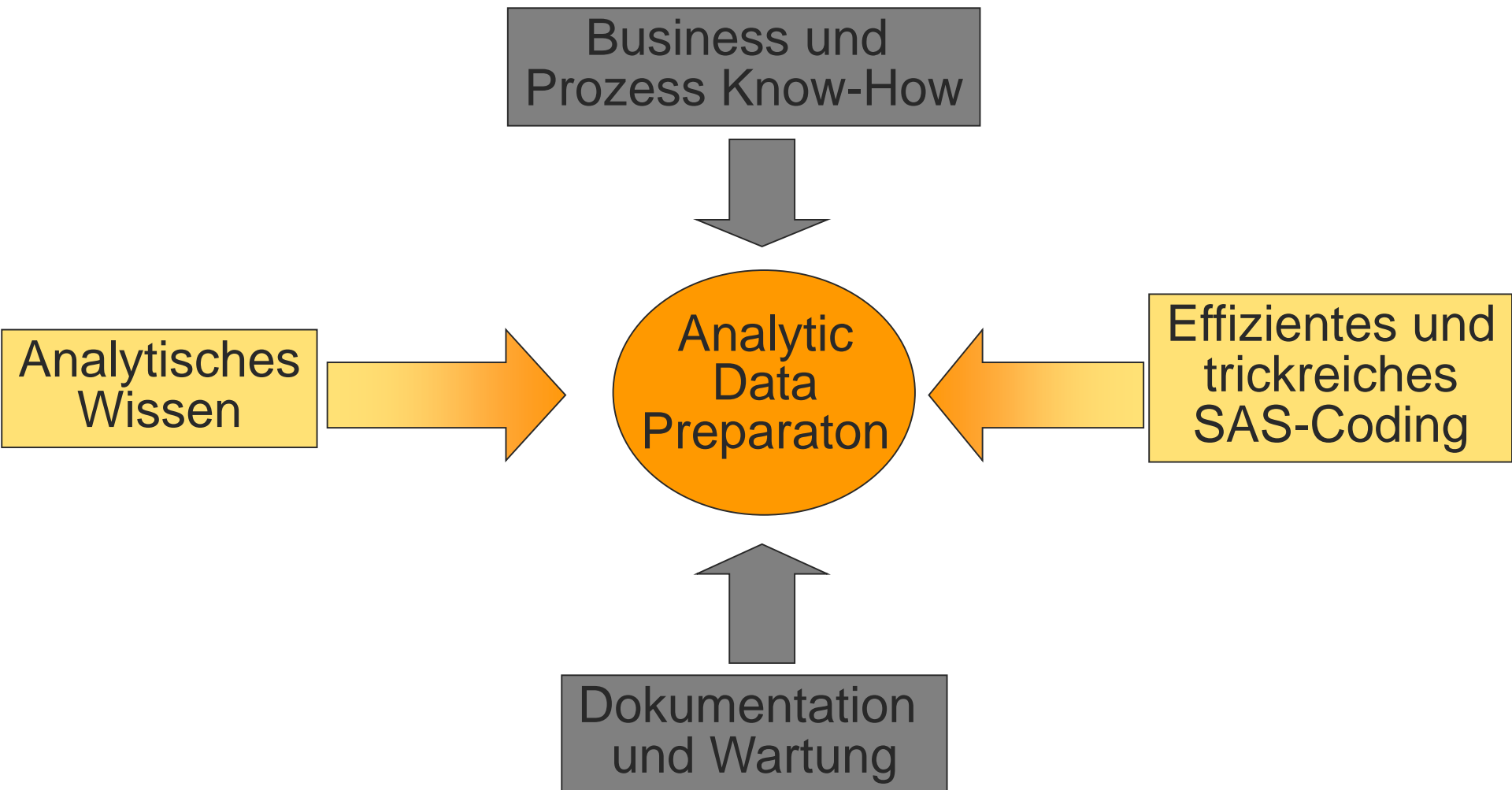


Month	Country	Product	Actual Sales
1993.01	CANADA	BED	\$856.00
1993.02	CANADA	BED	\$1,581.00
1993.03	CANADA	BED	\$1,900.00
1993.01	CANADA	SOFA	\$1,953.00
1993.02	CANADA	SOFA	\$2,483.00
1993.03	CANADA	SOFA	\$2,495.00
1993.01	GERMANY	BED	\$1,875.00
1993.02	GERMANY	BED	\$1,929.00
1993.03	GERMANY	BED	\$1,222.00
1993.01	GERMANY	SOFA	\$3,723.00
1993.02	GERMANY	SOFA	\$2,393.00
1993.03	GERMANY	SOFA	\$2,799.00

# Longitudinal Data Marts

- Aggregation von Daten am richtigen Level
  - Transactional Data
  - Finest Granularity
  - Most Appropriate Aggregation Level
- Definition von Cross Sectional Groups
- Ausrichten von Zeit-Werten
- Erzeugen von Event-Indikatoren und Input-Variablen

# Vier Dimensionen für Data Preparation

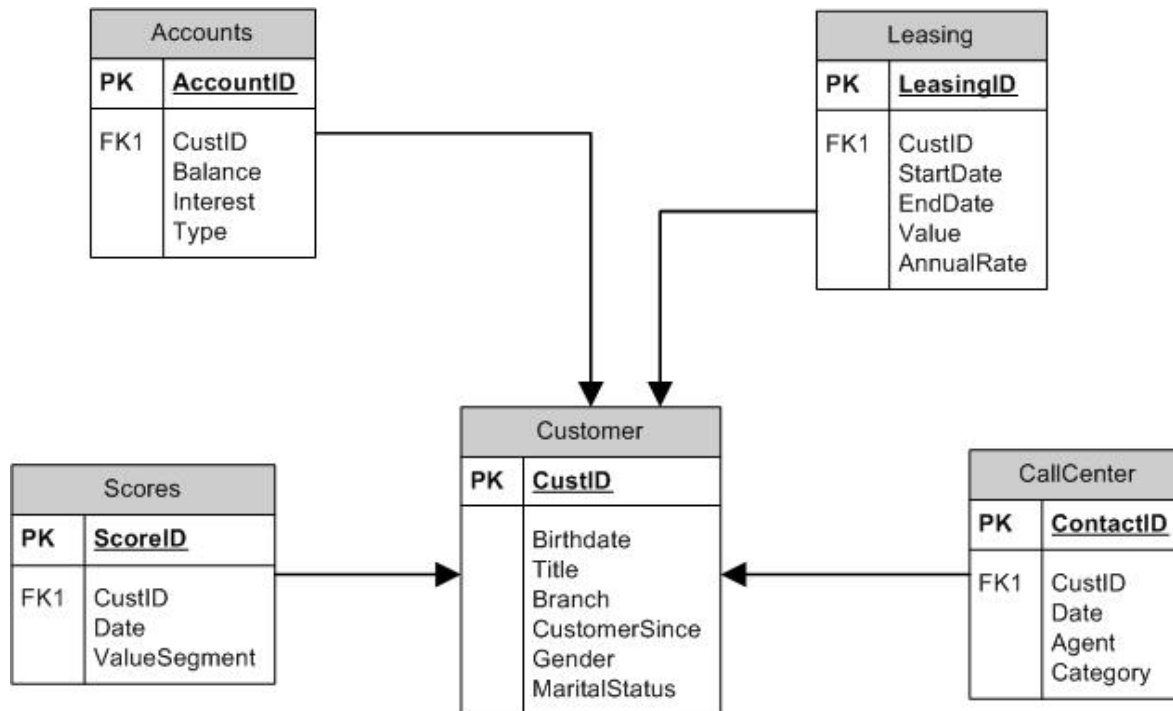




# Case Study: Fragestellung

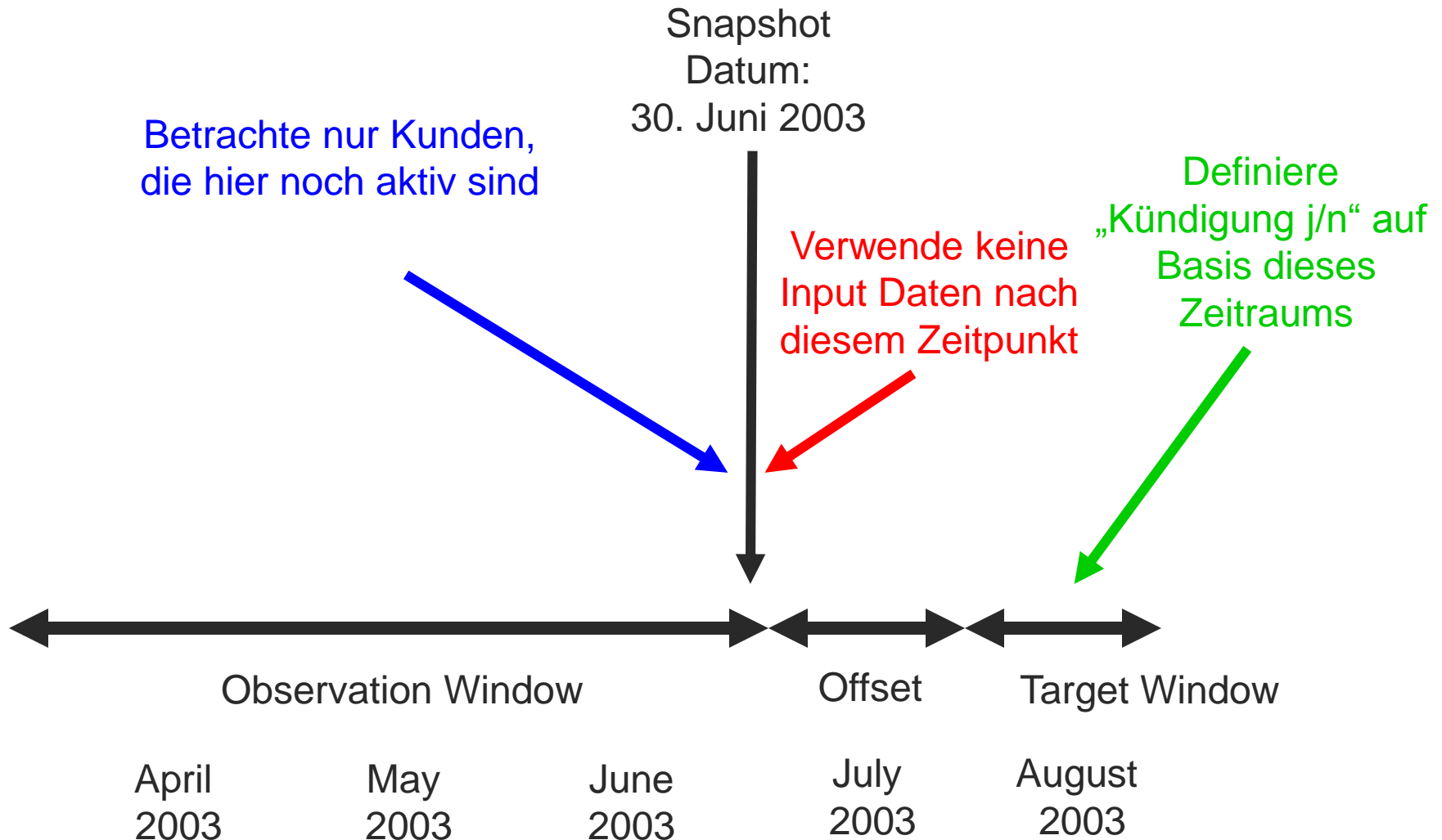
- Vorhersage der Kunden, die eine hohe Wahrscheinlichkeit haben, das Unternehmen zu verlassen
- Ableiten der Zielvariable „Cancellation YES/NO“ von der monatlichen Wertsegment-Historie (Eintrag „8. LOST“)
- Aufgabe: Erzeugen eines one-row-per-subject Data Mart für Data Mining Analyse

# Case Study: Daten und Datenmodell



- *Customer Daten:* Kundenstammdaten und demographische Daten
- *Account Daten :* Kontodaten (Monatsaggregate)
- *Leasing Daten :* Daten über Leasingverträge
- *Call Center Daten :* Call Center Kontakte
- *Score Daten:* Monatliche Wertsegment Daten

# Überlegungen für Predictive Modeling



# Verwenden von Daten aus der Callcenter Tabelle

	CustID	ContactID	Date	Agent	Category
1	1000008	1	19JUL2003:00:00:00	58	Telebanking
2	1000014	2	08APR2003:00:00:00	94	Complaint
3	1000014	3	02MAR2003:00:00:00	56	Complaint
4	1000018	4	12JUN2003:00:00:00	28	Telebanking
5	1000028	5	23FEB2003:00:00:00	36	Telebanking
6	1000034	6	20MAR2003:00:00:00	24	Telebanking
7	1000035	7	24MAY2003:00:00:00	21	Telebanking
8	1000035	8	25JUN2003:00:00:00	81	Telebanking
9	1000037	9	06JAN2003:00:00:00	32	Complaint
10	1000039	10	26JUN2003:00:00:00	70	Complaint
11	1000040	11	28APR2003:00:00:00	31	Complaint
12	1000040	12	19MAY2003:00:00:00	68	Complaint
13	1000041	13	18JUL2003:00:00:00	12	Telebanking
14	1000050	14	04JUL2003:00:00:00	99	Telebanking

```
%let snapdate = '30JUN2003'd;
PROC FREQ DATA = callcenter NOPRINT;
  TABLE CustID / OUT = CallCenterComplaints
    (DROP = Percent RENAME =
      (Count = Complaints));
  WHERE Category = 'Complaint' and
    datepart(date) <= &snapdate;
RUN;
```

# Verwenden von Daten aus der SCORES-Table

	CustID	ScoreID	Date	ValueSegment
1	1000002	1000001	01JAN2003	3. BRONCE
2	1000002	1000002	01FEB2003	2. SILBER
3	1000002	1000003	01MAR2003	1. GOLD
4	1000002	1000004	01APR2003	3. BRONCE
5	1000002	1000005	01MAY2003	2. SILBER
6	1000002	1000006	01JUN2003	2. SILBER
7	1000005	1000007	01JAN2003	2. SILBER
8	1000005	1000008	01FEB2003	3. BRONCE
	1000005	1000009	01MAR2003	1. GOLD
	1000005	1000010	01APR2003	1. GOLD
	1000005	1000011	01MAY2003	3. BRONCE
	1000005	1000012	01JUN2003	3. BRONCE
	1000006	1000013	01JAN2003	2. SILBER
	1000006	1000014	01FEB2003	1. GOLD
	1000006	1000015	01MAR2003	3. BRONCE
	1000006	1000016	01APR2003	1. GOLD
	1000006	1000017	01MAY2003	3. BRONCE
	1000006	1000018	01JUN2003	3. BRONCE

```

%let snapdate = '30JUN2003'd;
DATA ScoreFuture(RENAME = (ValueSegment =
                           FutureValueSegment))
  ScoreActual
  ScoreLastMonth(RENAME = (ValueSegment =
                           LastValueSegment));

SET Scores;
DATE = INTNX('MONTH',Date,0,'END');
DROP Date;
IF Date = &snapdate THEN OUTPUT ScoreActual;
ELSE IF Date = INTNX('MONTH',&snapdate,-1)
      THEN OUTPUT ScoreLastMonth;
ELSE IF Date = INTNX('MONTH',&snapdate,2)
      THEN OUTPUT ScoreFuture;
  
```



```
DATA CustomerMart;
ATTRIB /* Customer Baseline */
CustID          FORMAT    = 8.          LABEL = "Customer ID"
Birthdate       FORMAT    = DATE9.     LABEL = "Date of Birth"
Alter           FORMAT    = 8.          LABEL = "Age (years) "
Gender          FORMAT    = $6.         LABEL = "Gender"
MaritalStatus   FORMAT    = $10.        LABEL = "Marital Status"
Title           FORMAT    = $10.        LABEL = "Academic Title"
HasTitle        FORMAT    = 8.          LABEL = "Has Title? 0/1"
Branch          FORMAT    = $5.         LABEL = "Branch Name";
MERGE Customer (IN = InCustomer)
      AccountSum (IN = InAccounts)
      AccountTypes
      LeasingSum (IN = InLeasing)
      CallCenterContacts (IN = InCallCenter)
      CallCenterComplaints
      ScoreFuture
      ScoreActual
      ScoreLastMonth;
BY CustID;
IF InCustomer;
```

```

/* Customer Baseline */
HasTitle = (Title ne "");
Alter = (&Snapdate-Birthdate)/365.25;
CustomerMonths = (&Snapdate- CustomerSince)/(365.25/12);
/* Accounts */
HasAccounts = InAccounts;
LoanPct = Loan / BalanceSum * 100;
SavingAccountPct = SavingAccount / BalanceSum * 100;
FundsPct = Funds / BalanceSum * 100;
/* Leasing */
HasLeasing = InLeasing;
/* Call Center */
HasCallCenter = InCallCenter;
ComplaintPct = Complaints / Calls *100;
/* Value Segment */
Cancel = (FutureValueSegment = '8. LOST');
ChangeValueSegment = (ValueSegment = LastValueSegment);
RUN;

```



# Screenshots des Ergebnis- Data Mart

	Customer ID	Date of Birth	Age (years)	Gender	Marital Status	Academic Title	Has Title? 0/1	Branch Name	Customer Start Date	Customer Duration (months)
1	1000002	26DEC1958	44	Male	Married		0	Fil1	01JAN2000	41
2	1000005	25JUN1947	56	Male	Single	Ing.	1	Fil4	01APR1999	50
3	1000006	10DEC1945	57	Female	Married		0	Fil4	01SEP1996	81
4	1000007	02JUN1934	69	Male	Married		0	Fil1	01SEP1997	69
5	1000008	15DEC1957	45	Male	Single	Dr.	1	Fil3	01JAN1996	89
6	1000009	11MAR1959	44	Male	Single		0	Fil2	01JUL2001	23
7	1000014	23AUG1952	51	Male	Single		0	Fil4	01MAY1996	85
8	1000015	12MAY1959	44	Male	Single		0	Fil2	01FEB1999	52
9	1000016	11FEB1967	36	Male	Married		0	Fil2	01FEB2001	28

	Customer ID	Customer has any accounts	Number of Accounts	All Accounts Balance Sum	Average Interest	Loan Balance Sum	Saving Account Balance Sum	Funds Balance Sum	Loan Balance Proportion	Saving Account Balance Proportion	Funds Balance Proportion	Customer has any leasing contract	Number of leasing contracts	Totals leasing value	Total annual leasingrate
1	1000002	1	2	3100.84	5.0	1550.42	1550.42	0.00	50.00	50.00	0.00	1	1	521763.0	254.69
2	1000005	1	1	3775.31	6.0	0.00	3775.31	0.00	0.00	100.00	0.00	1	1	855215.0	232.52
3	1000006	1	1	2376.43	2.0	0.00	0.00	2376.43	0.00	0.00	100.00	1	1	560362.0	167.37
4	1000007	1	2	3625.44	5.0	0.00	1812.72	1812.72	0.00	50.00	50.00	1	2	1735708	168.75
5	1000008	1	1	3350.65	2.0	0.00	0.00	3350.65	0.00	0.00	100.00	1	1	5276.00	109.15
6	1000009	1	3	3575.46	4.0	1191.82	0.00	1191.82	33.33	0.00	33.33	1	2	591963.0	170.14
7	1000014	1	2	3000.92	4.5	0.00	3000.92	0.00	0.00	100.00	0.00	1	1	564728.0	92.51
8	1000015	1	1	2801.09	5.0	0.00	2801.09	0.00	0.00	100.00	0.00	1	1	393984.0	189.54
9	1000016	1	2	3325.66	1.0	0.00	1662.83	0.00	0.00	50.00	0.00	0	0	0.00	0.00

	Customer ID	Customer has any call center contact	Number of call center contacts	Number of complaints	Percentage of complaints	Currenty Value Segment	Last Value Segment	Change in Value Segment	Customer cancelled
1	1000002	0	0	0		2. SILBER	2. SILBER	1.00	0
2	1000005	0	0	0		3. BRONCE	3. BRONCE	1.00	0
3	1000006	0	0	0		3. BRONCE	3. BRONCE	1.00	0
4	1000007	0	0	0		2. SILBER	1. GOLD	0.00	0
5	1000008	1	1	0	0.00	2. SILBER	2. SILBER	1.00	0
6	1000009	0	0	0		3. BRONCE	4. LEAD	0.00	0
7	1000014	1	2	2	100.00	3. BRONCE	1. GOLD	0.00	0
8	1000015	0	0	0		3. BRONCE	4. LEAD	0.00	0
9	1000016	0	0	0		2. SILBER	2. SILBER	1.00	0



# Paper von SUGI 31 (2006 in San Francisco)

- <http://www2.sas.com/proceedings/sugi31/078-31.pdf>
- Oder
- [http://sascommunity.org/wiki/Gerhard\\_Svolba](http://sascommunity.org/wiki/Gerhard_Svolba)

# Die Möglichkeiten und Stärke der SAS Language für Data Preparation

# SAS Data Step vs. SQL

## Transponieren einer Tabelle

	CUSTOMER	TIME	PRODUCT
	0	0	hering
2	0	1	comed_b
3	0	2	olives
4	0	3	ham
5	0	4	turkey
6	0	5	bourbon
7	0	6	ice_crea
8	1	0	baguette
9	1	1	soda
10	1	2	hering
11	1	3	cracker
12	1	4	heineken
13	1	5	olives
14	1	6	comed_b
15	2	0	avocado
16	2	1	cracker
17	2	2	artichok
18	2	3	heineken
19	2	4	ham
20	2	5	turkey
21	2	6	sardines

```
PROC TRANSPOSE DATA = sampsis.assoc(obs=21)
                OUT = assoc_tp (DROP = _name_);

    BY customer;
    ID Product;
RUN;
```

	CUSTOMER	bourbon	comed_b	ham	hering	ice_crea	olives	turkey	baguette
1	0	1	1	1	1	1	1	1	.
2	1	.	1	.	1	.	1	.	1
3	2	.	.	1	.	.	.	1	.
4	3	1	.	1	.	1	1	1	.
5	4	.	1	.	1	.	1	1	.
6	5	.	.	1	.	1	.	.	.
7	6	1	.	.	.	1	1	1	.
8	7	1	1	.	.	1	.	.	1
9	8	1	.	.	.	.	1	.	1
10	9	1	1	.	1	.	.	.	.
11	10	.	.	.	.	.	.	1	1
12	11	.	1	.	1	.	.	.	1
13	12	.	1	.	1	.	1	.	.

# Wechseln zwischen longitudinalen Datenstrukturen


Standard  
Form



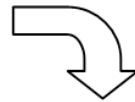
Interleaved

```
PROC TRANSPOSE DATA = diy standard
OUT = diy intleaved
(rename = (Col1 = Value))
NAME = Type;


BY date;
RUN;
```



Date	Quantity	Volume
15/08/05	7321	39079
16/08/05	7926	41357
17/08/05	9507	46972
19/08/05	8607	43731
20/08/05	8034	39402
21/08/05	7775	38277
22/08/05	7723	36687
23/08/05	7413	38813
24/08/05	8229	42677
26/08/05	6914	34210
27/08/05	7419	37048
28/08/05	6730	34626
29/08/05	7228	33361
30/08/05	9444	51607
31/08/05	10830	57812




	Date	Type	Value
1	15/08/05	Quantity	7321
2	15/08/05	Volume	39079
3	16/08/05	Quantity	7926
4	16/08/05	Volume	41357
5	17/08/05	Quantity	9507
6	17/08/05	Volume	46972
7	19/08/05	Quantity	8607
8	19/08/05	Volume	43731
9	20/08/05	Quantity	8034
10	20/08/05	Volume	39402
11	21/08/05	Quantity	7775
12	21/08/05	Volume	38277
13	22/08/05	Quantity	7723
14	22/08/05	Volume	36687
15	23/08/05	Quantity	7413
16	23/08/05	Volume	38813



```
PROC TRANSPOSE DATA = diy intleaved
OUT = diy standard back
(drop = name );

BY date;
ID Type;
VAR value;
```



# Selektion der ersten und letzten Beobachtung pro Kunde

CustID	Month	Value
1	7	45
1	8	34
1	9	5
2	7	34
2	8	32
2	9	44
3	7	56
3	8	54
3	9	32

```
data customer;
  set customer;
  by CustID;
  FirstValue = First.CustID;
  LastValue  = Last.CustID;
run;
```

# Erzeugen einer Sequenz-Variable und kumulativer Summen

CustID	Date	Points
1	10.03.2004	45
1	04.04.2004	10
1	20.04.2004	20
1	16.05.2004	18
1	01.06.2004	5
2	01.02.2004	10
2	19.03.2004	30
3	05.08.2004	4
3	16.08.2004	16
3	31.08.2004	12
3	10.09.2004	20

```
data customer;
set customer;
by CustID;
if first.custid then do;   Purch_No = 1;
                           Cum_Poi  = Points;   end;

else do;   Purch_No + 1;
           Cum_Poi + Points;   end;

run;
```

# Kopieren von fehlenden Daten

CustID	Age	Gender	Month	Value
1	26	m	7	45
			8	34
			9	5
2	37	w	7	34
			8	32
			9	44
3	46	m	7	56
			8	54
			9	32

```

data customer;
  set customer;
  retain age_tmp;
  if age ne "" then age_tmp = age;
  else age = age_tmp;
run;

```

# Verschieben einer Spalte um k-Beobachtungen abwärts

Date	Value
01.01.2004	45
01.02.2004	34
01.03.2004	5
01.04.2004	34
01.05.2004	32
01.06.2004	44

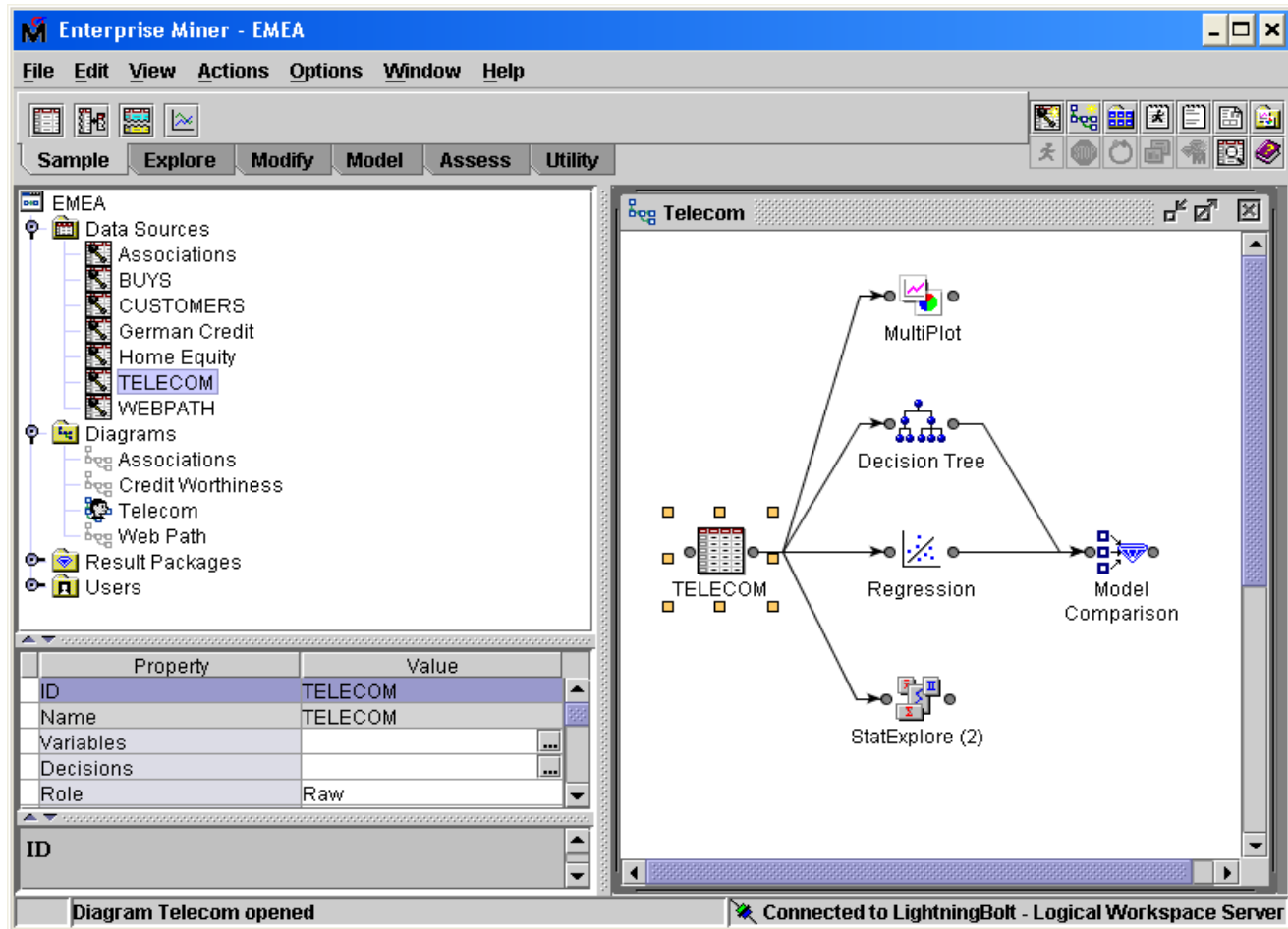
```
data measurements;
  set measurements;
  Value_PrevDay = lag(Value);
run;
```



# Datenmanagement mit SAS® Tools

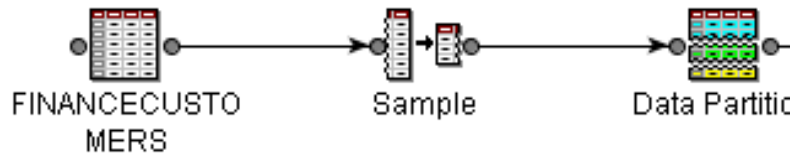
- SAS®Enterprise Miner
- SAS®Forecast Studio
- SAS®Data Integration Studio

# Analytische Datenaufbereitung im SAS® Enterprise Miner

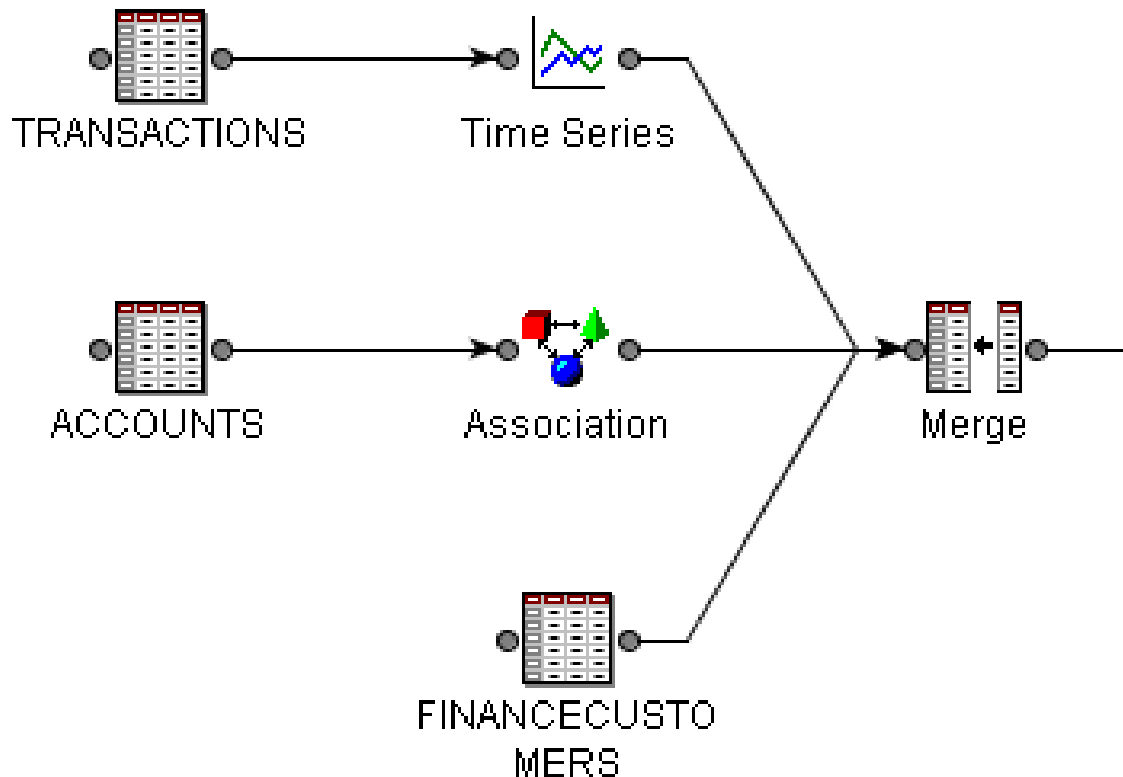


# Analytische Datenaufbereitung im SAS® Enterprise Miner

- Input Data Source Node
- Sample Node
- Data Partition Node
- Metadata Node
- Filter Node
- Transform Variables Node
- Impute Node
- SAS Code Node
- Principal Components Node



# Arbeiten mit Multiple-Row-per-Subject Daten im SAS® Enterprise Miner

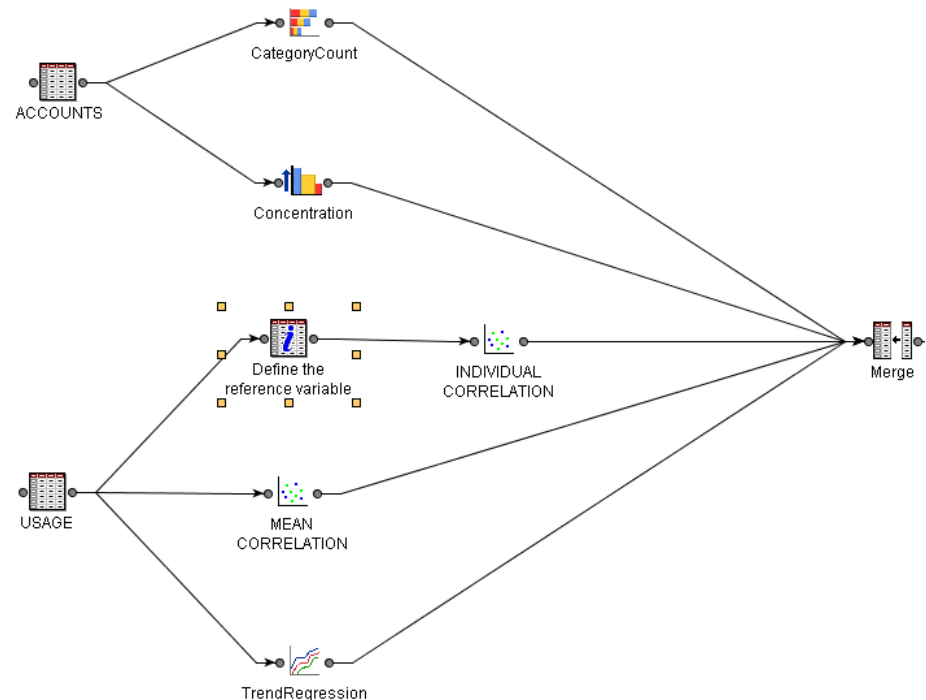
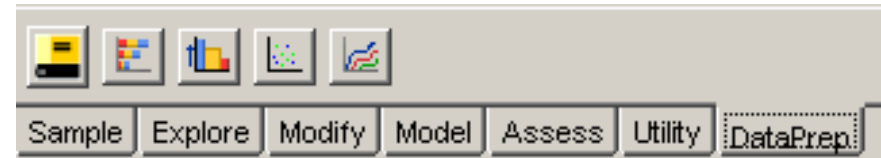


- Time Series Node
- Association Node
- Merge Node

# Extending SAS®Enterprise Miner with Data Preparation for Analytics Nodes

## Extension Nodes

- Correlation
- TrendRegression
- Concentration
- CategoryCount



# Analytische Datenaufbereitung im SAS® Enterprise Miner - Zusammenfassung

- Dokumentation des Datenaufbereitungsprozesses als Ganzes im Prozessflussdiagramm
- Definition der Metadaten, die in der Analyse verwendet werden
- Automatisches Erzeugen von Dummy-Variablen für kategorielle Daten
- Mächtige Datentransformationen
  - Im filter node, transform variables node und impute node
  - Möglichkeit Assoziationsanalyse und Zeitreihenanalyse durchzuführen
- Erzeugen von Score Code von allen Nodes im SAS Enterprise Miner als SAS Datastep Code

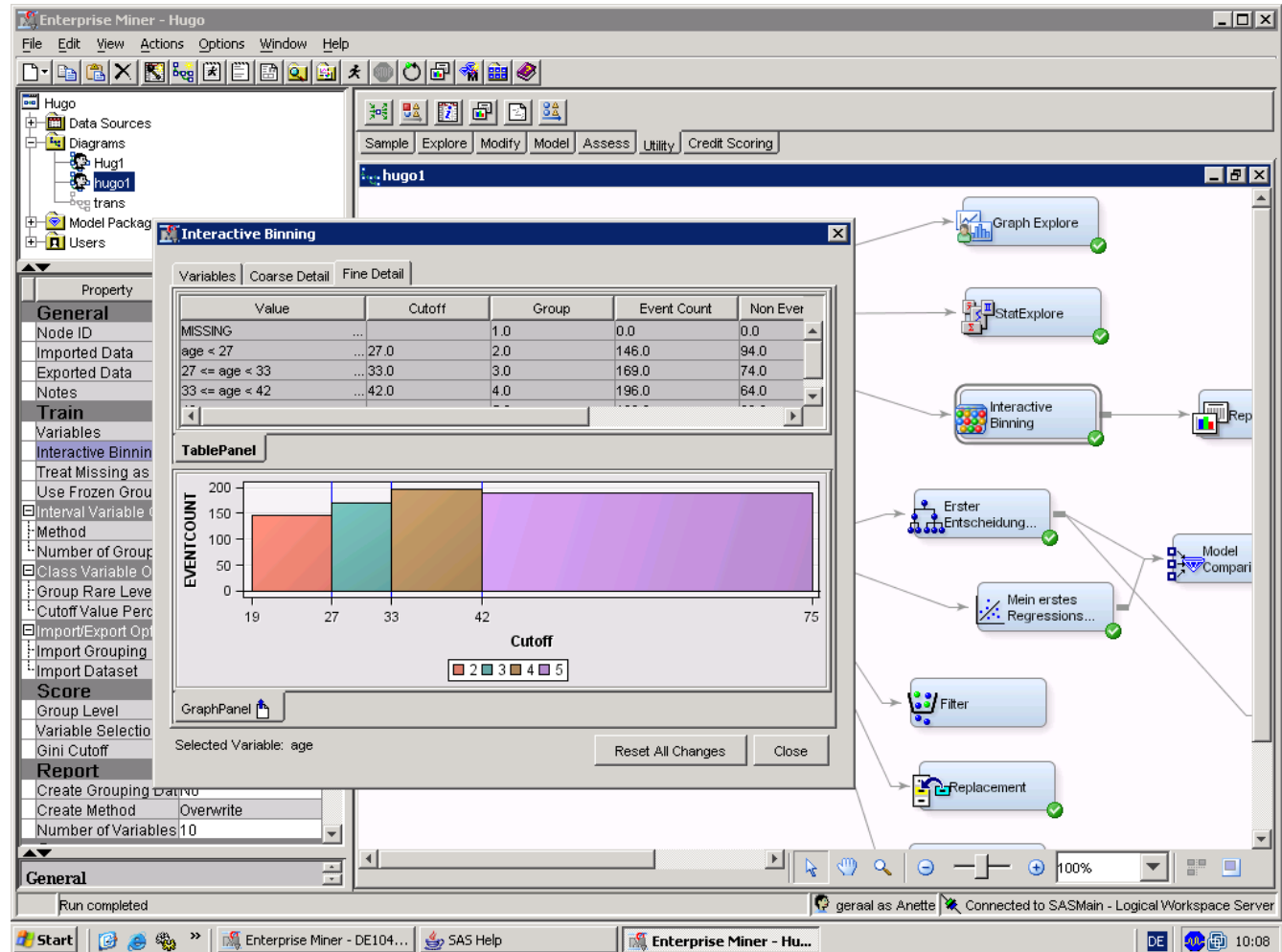
# Enterprise Miner 5.3: Neue Benutzeroberfläche

The screenshot displays the SAS Enterprise Miner 5.3 interface. On the left, a tree view shows the project structure: Hugo, Data Sources, Diagrams (hugo1), Model Packages, and Users. Below this is a 'Property' panel with a table of settings for the 'hugo1' diagram.

Property	Value
<b>General</b>	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interaction	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
<b>Class Targets</b>	
Regression Type	Logistic Regression
Link Function	Logit
<b>Model Options</b>	
Suppress Intercept	No
Input Coding	Deviation
<b>Model Selection</b>	
Selection Model	Backward
Selection Criterion	Default
Use Selection Default	Yes
Selection Options	...
<b>Optimization Options</b>	
Technique	Default
Default Optimization	Yes

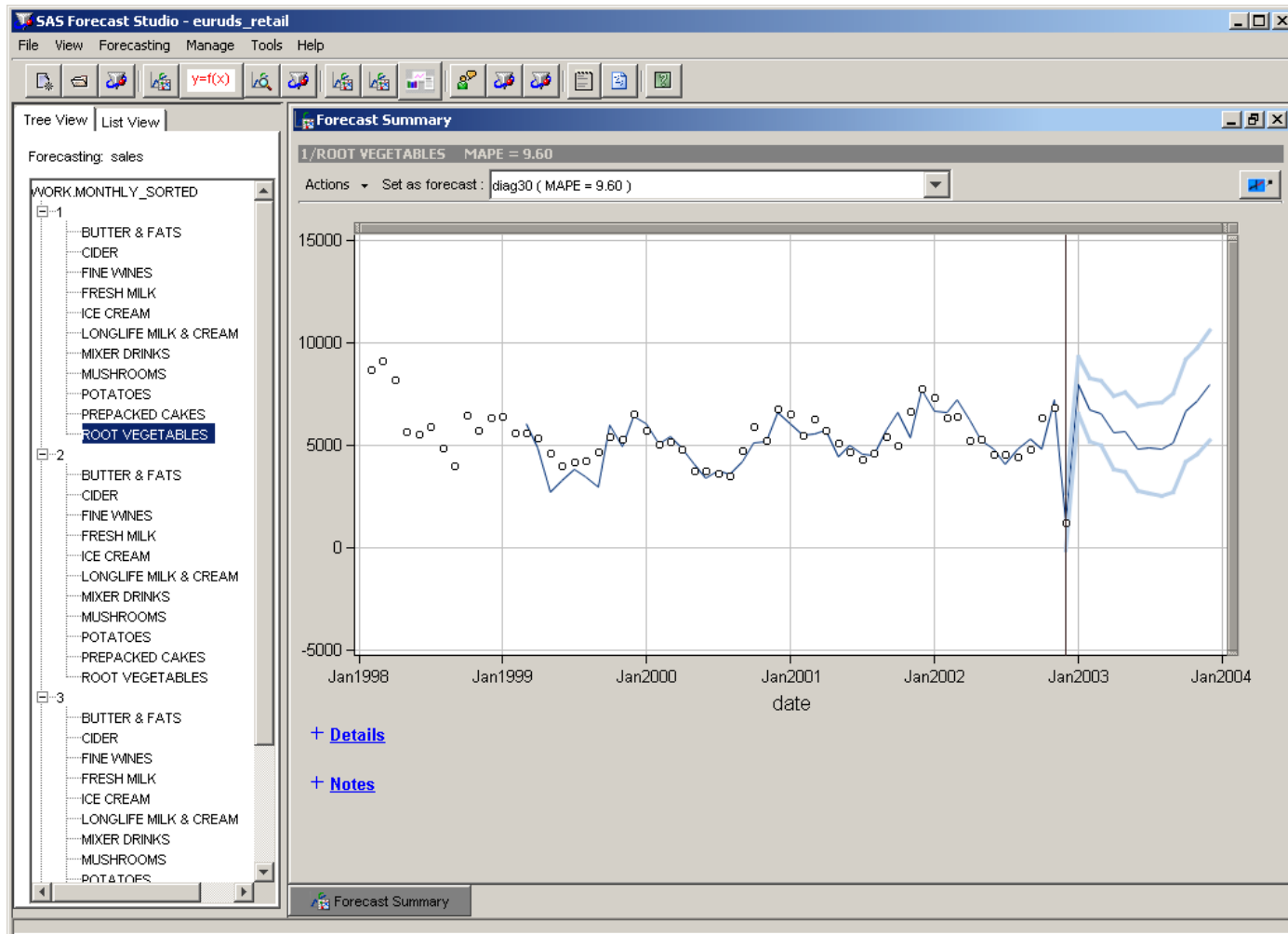
The main workspace shows a workflow diagram for 'hugo1'. It starts with a 'German Credit' data source, which branches into 'Graph Explore', 'StatExplore', and 'Interactive Binning'. 'Data Partition' is connected to 'German Credit' and further branches into 'Erster Entscheidung...', 'Mein erstes Regressions...', 'Filter', and 'Replacement'. 'Erster Entscheidung...' and 'Mein erstes Regressions...' both lead to 'Model Compari...'. The interface includes a menu bar (File, Edit, View, Actions, Options, Window, Help), a toolbar, and a status bar at the bottom showing 'Diagram hugo1 opened' and 'Connected to SASMain - Logical Workspace Server'.

# Enterprise Miner 5.3: Neuer Knoten für (Interactive) Binning





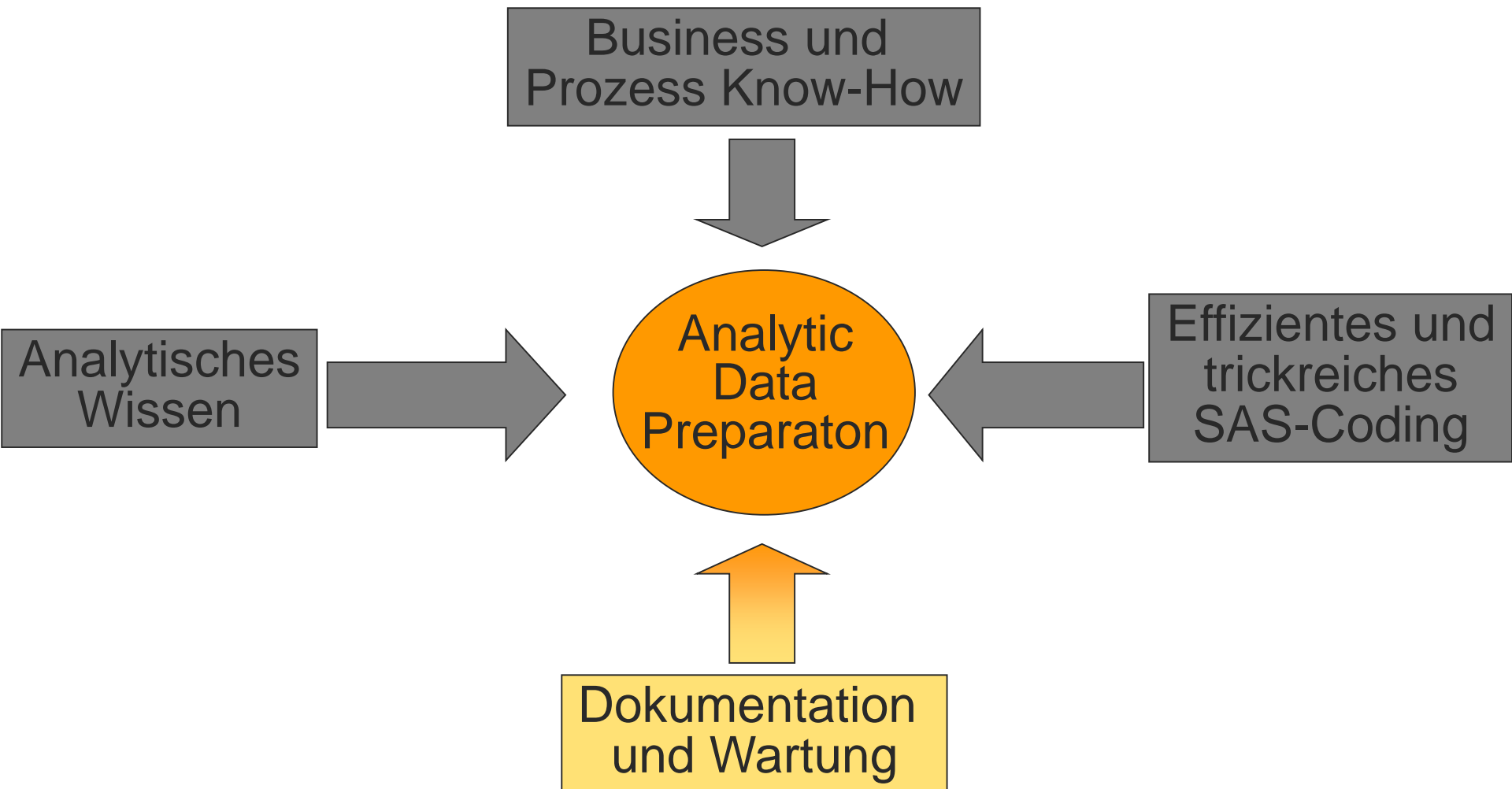
# Analytic Data Preparation im SAS® Forecast Studio



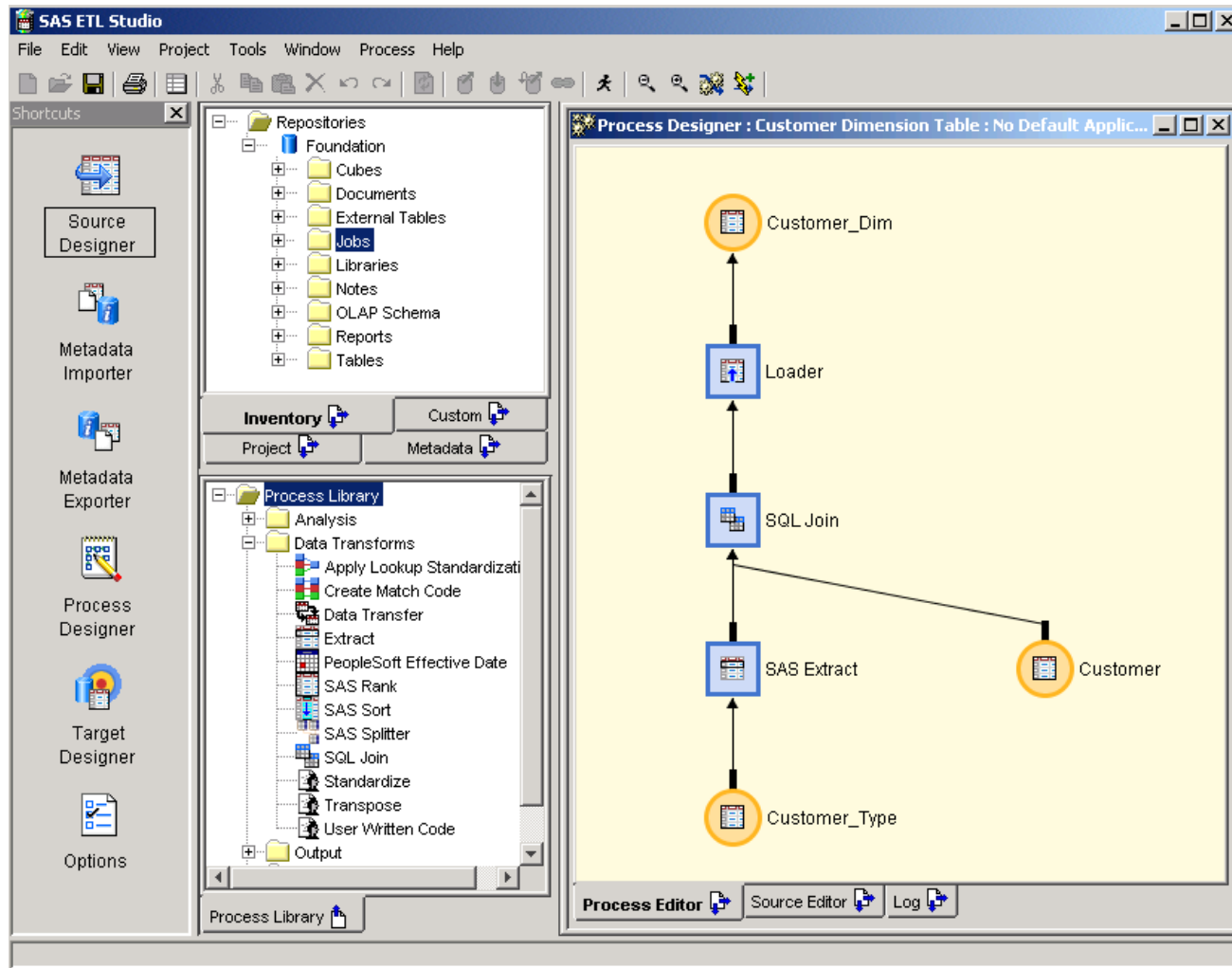
# Analytic Data Preparation im SAS® Forecast Studio

- Konvertieren von transaktionellen Daten in Zeitreihendaten
- Behandeln von fehlenden Werten
- Ausrichten von Datumswerten in Zeitintervallen
- Aggregation von Daten auf unterschiedlichen Stufen
- Behandlung von “Ereignissen”
  - Ermöglicht die Definition von Ereignissen, sowie die Zuweisung von Ereignissen zu bestimmten Zeitreihen
  - Benutzer kann Event-Dauer, Form und Wiederholungsoptionen setzen

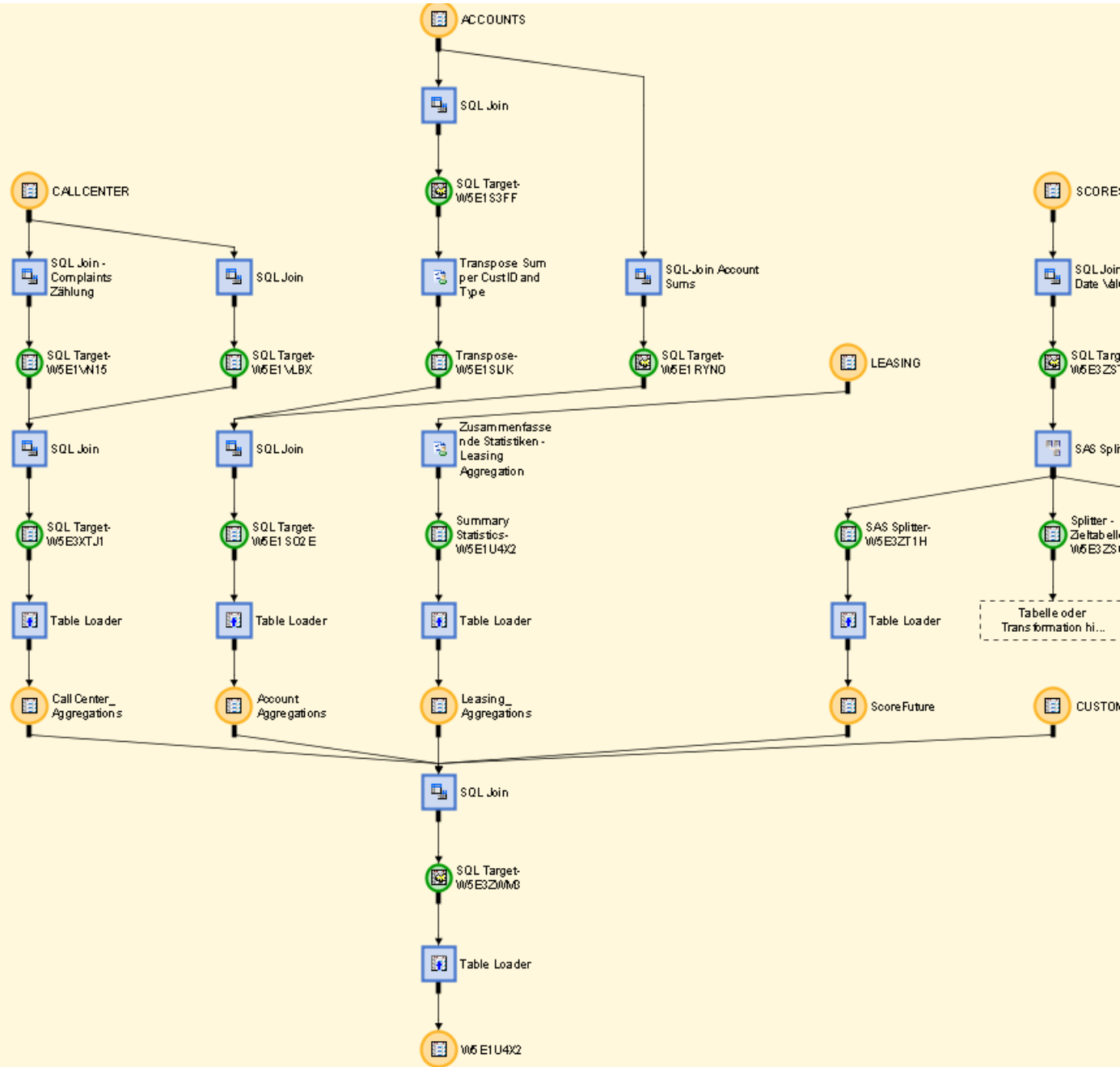
# Vier Dimensionen für Data Preparation



# SAS® Data Integration Studio



# SAS® Data Integration Studio



# SAS® Data Integration Studio – Allgemeine Eigenschaften

- Umfassende Transformations Bibliothek
- Graphisches User-Interface; drag & drop, Wizards
- Multi-developer support: Check-in/check-out, Change Management
- Impact analysis
- **Dokumentation des Datenmanagement-Prozesses**

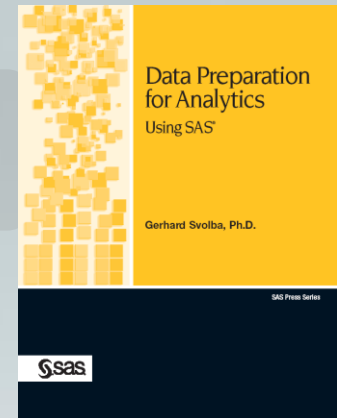
# SAS® DI Studio – Analytische Merkmale

- Data Mining Model Scoring
  - Registrieren von Modellen im SAS Metadata Repository
  - Anwenden eines SAS Enterprise Miner Modells auf neue Daten
  - Erzeugt die Zieltabellen-Definition in den Metadaten
- Forecast Analysis Transformationen
  - Erlaubt die Durchführung von Zeitreihenanalysen
  - Basiert auf den HPF (high performance forecasting) Procedures
  - Ermöglicht die Integration des Forecasting-Schritts in den Datenfluss-Prozess in den Metadaten

# Summary

- Analytic Data Preparation ist eine Disziplin und nicht eine lästige Notwendigkeit
- Analytic Data Preparation ist mehr als nur Coding
- Das One-Row-Per-Subject Paradigma
  - Zentrale Rolle in Data Mining und Predictive Modeling
  - Gib Dich nicht mit einfachen Transpositions, Summen oder Mittelwerte zufrieden.
  - Intelligente Aggregationen können der Schlüssel zu guten Modellen sein
- Predictive Modeling und historische Daten:  
Welche Daten dürfen für das Modell verwenden?
- SAS kombiniert mächtiges Datenmanagement und marktführende Analytik in einem Paket
- SAS Tools wie SAS® ETL-Studio, SAS® Enterprise Miner and SAS® Forecast Studio unterstützen bei der analytischen Datenaufbereitung





## „Recommended Reading“

### Data Preparation for Analytics

by Gerhard Svolba

SAS-Press (#60502)

---

Business Rationale  
Concepts  
Coding Examples

**THE  
POWER  
TO KNOW®**

***"It is exciting to see a book completely devoted to data preparation in SAS."***

Jin Li

Statistician

Capital One Financial Services

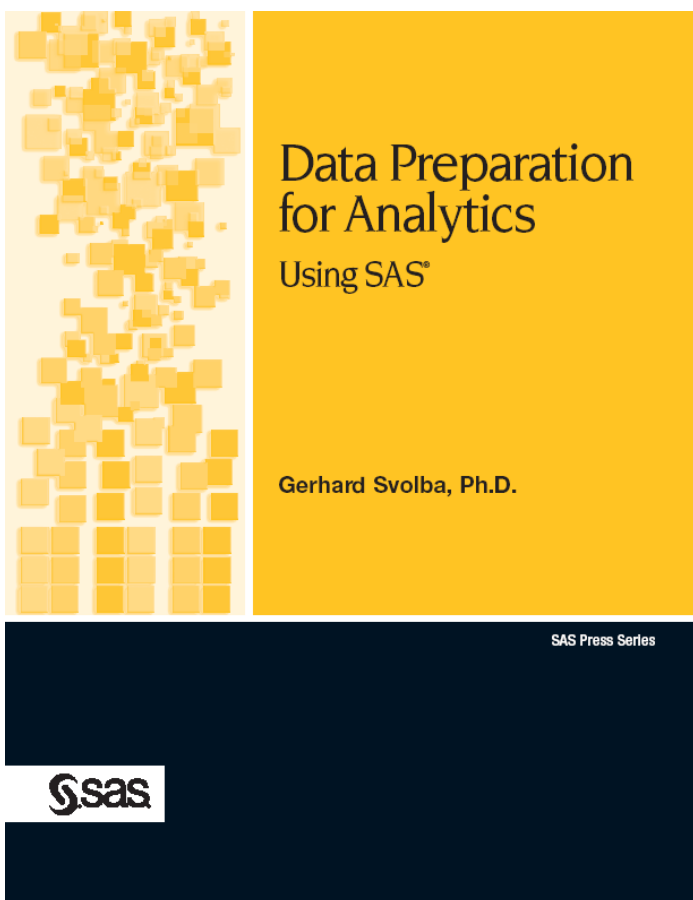
***This is a must read for anyone, who prepares data for data mining and analytics. Not only you receive ideas and suggestions for important derived variables for your datamart, you also get a lot of insight on the business rationale behind the scenes. The author does a great job in explaining you step by step the world of data preparation for data mining and shows a lot of example code and macros.***

Christine Hallwirth

MHE & Partners

***Gerhard Svolba's book "Data Preparation for Analytics" is an "owner's manual" for data miners, business analysts and all who prefer to be in charge of and responsible for their own datasets. This book is for those who are not afraid of data, who understand data, and for whom rolling up their sleeves and getting their hands into data is an integral part of analytics and predictive modeling.***

Lessia Shajenko  
(American Bank)



- Günter Schmölz
- *This book is a must for those, who are preparing data for datamining and those who want to be efficient and effective in preparing any meaningful analyses. It provides a lot of examples how to transform variables and data. I have got practical experience with SAS for more than 10 years, but this book gave me many new approaches, ideas and solutions. This book is very convenient and enjoyable to use, because it often starts with basics, so it is easy to understand and it makes it easy to get into it. Additionally it leads to more sophisticated solutions. This book deals exactly with those topics which are central for advanced datamining.*
- Friedrich Bauernberger
- *Data Preparation, zweifellos der wichtigste Teil eines jeden Data Mining Projekts, wird meist zu wenig Aufmerksamkeit zu Teil. Dieses Buch nimmt sich genau dieser Thematik an und bietet selbst für einen erfahrenen SAS-User Tipps & Tricks um effizienter und schneller Daten aufzubereiten.  
Ein Buch aus der Praxis - ich konnte einige Bereich gleich in aktuelle Projekte übernehmen. Bleibt als einzige Kritik dass ich dieses Buch bereits vor 10 Jahren benötigt hätte.*

# SAS Training vom 6. – 8. Mai 2008 in Heidelberg

## Building Analytic Datamarts

- **Im Kurs werden dabei folgende Kenntnisse vermittelt**
  - Verständnis des Ökosystems für analytische Datenaufbereitung
  - Kenntnis der häufig verwendeten analytischen Datenstrukturen und deren Eignung für bestimmte analytische Fragestellungen
  - Leitfaden für die Herangehensweise bei der Erstellung von relevanten abgeleiteten Variablen
  - Tipps & Tricks zur effizienten SAS Programmierung bei der Erstellung analytischer Datenbestände
- **Der Kurs gliedert sich in folgende Themenblöcke**
  - Das Ökosystem für analytische Datenaufbereitung: Fachliche Anforderungen und Prozesse, Personen und Datenquellen
  - Datenquellen und Datenmodelle
  - Datenstrukturen für analytische Datenbestände und deren Umstrukturierung
  - Datentransformationen für analytische Fragestellungen
  - Erstellen eines “one-row-per-subject data marts”
  - Datenaufbereitung für Predictive Modeling
  - Datenaufbereitung für Zeitreihenanalysen
  - Automatisierung und der Einsatz von SAS®Tools wie SAS®Enterprise Miner und SAS®Data Integration Studio
  - Scoring, Sampling und andere Datenaufbereitungsthemen
- **Die Zielgruppe**
  - Der Kurs richtet sich an jene, die Datenbestände für Statistik, Data Mining und Zeitreihenanalyse aufbereiten; die Daten für die Erstellung dieser Datenbestände zur Verfügung stellen; die Analysen auf Basis von Datenbeständen durchführen. Die typischen Kursteilnehmer sind u.a. User von SAS®Enterprise Miner, SAS®STAT, SAS®ETS, SAS®Forecast Server, SAS®Base und SAS®Data Integration Studio, sowie jene, die in relationalen Datenbanken analytische Datenbestände mit SQL aufbereiten.
- **Die Voraussetzungen**
  - Voraussetzung für den Kurs sind grundlegende Kenntnisse in der SAS Programmierung. Erfahrungen in den Bereichen Statistik und Data Mining sind empfehlenswert.
- **Dauer**
  - 3 Tage

# Questions and Contact

- Gerhard Svolba (PhD)
- Email: [gerhard.svolba@aut.sas.com](mailto:gerhard.svolba@aut.sas.com)
- [http://sascommunity.org/wiki/Gerhard\\_Svolba](http://sascommunity.org/wiki/Gerhard_Svolba)
  - <http://www.sas.com/apps/pubscat/bookdetails.jsp?pc=60502>
  - [http://support.sas.com/publishing/bbu/companion\\_site/60502.html](http://support.sas.com/publishing/bbu/companion_site/60502.html)
  - <http://www2.sas.com/proceedings/sugi31/078-31.pdf>

