Data Preparation for Analytics

Gerhard Svolba (PhD)
SAS-Austria (Vienna)

SAS Presents

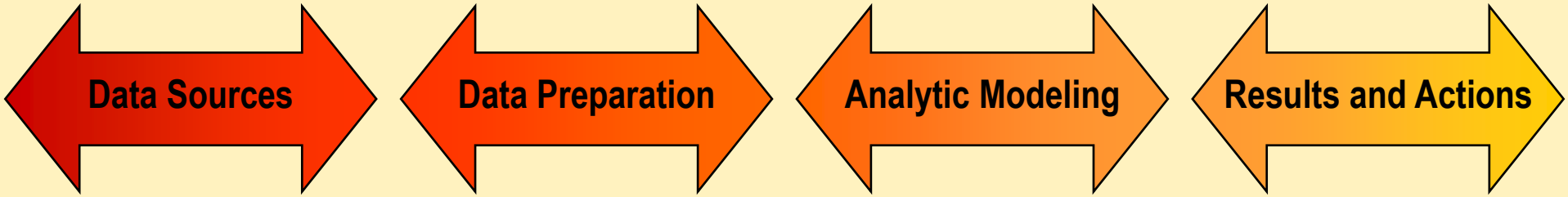
Agenda

- Analytic Data Preparation
- Data Structures for Analytic
- Tricky Data Management with the SAS® Language
- Analytic Data Management with SAS® Tools
  - SAS® Enterprise Miner
  - SAS® Forecast Server
  - SAS® ETL-Studio
- Closing Thoughts

Some Words on Data Preparation

- Is for techies
- Is boring
- Consumes up to 80 % of the project
- Is something that SAS can excellently do
- Is vital to the quality of the project

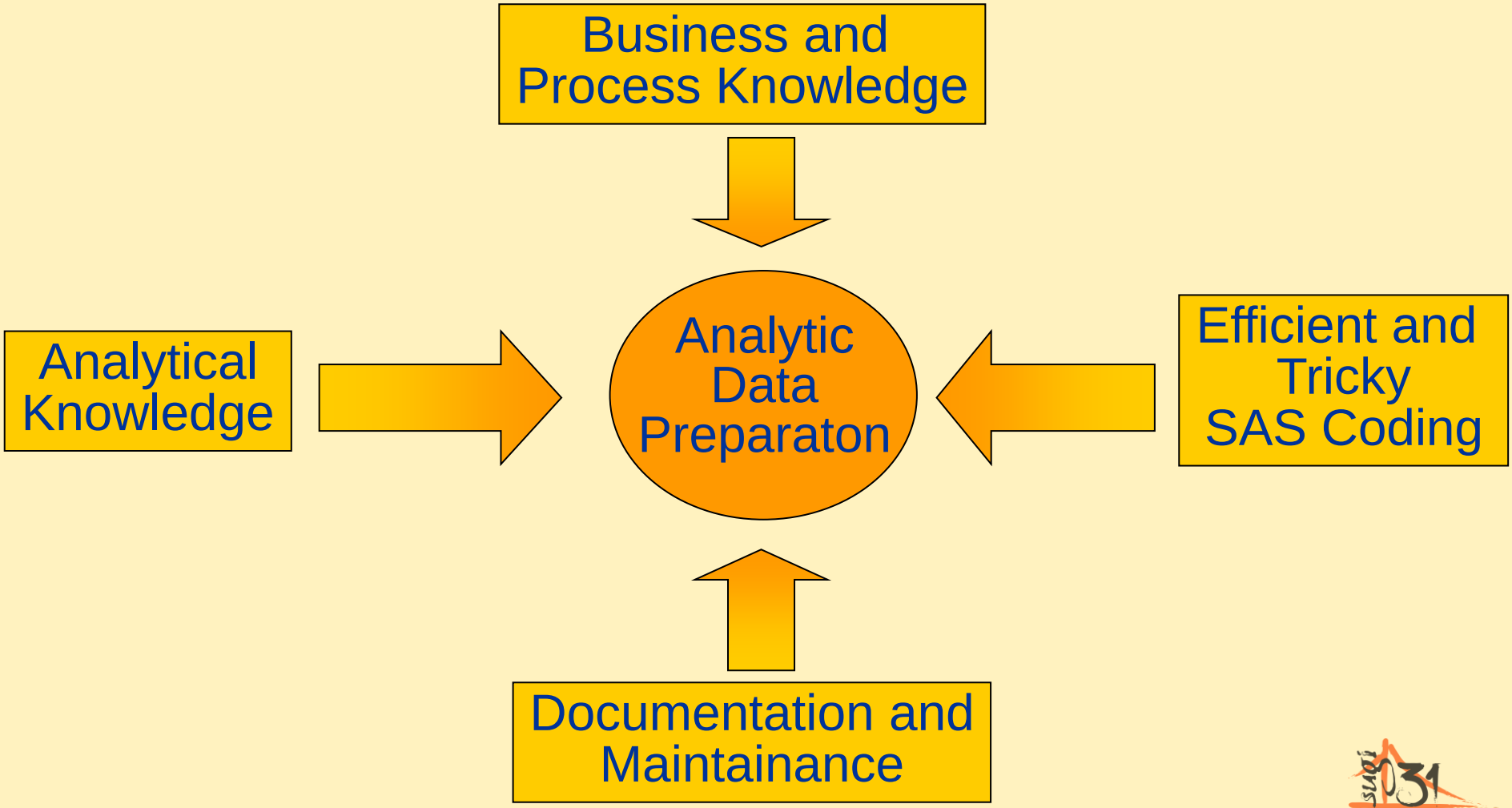
The Analysis Process: From Raw Data to Actionable Results



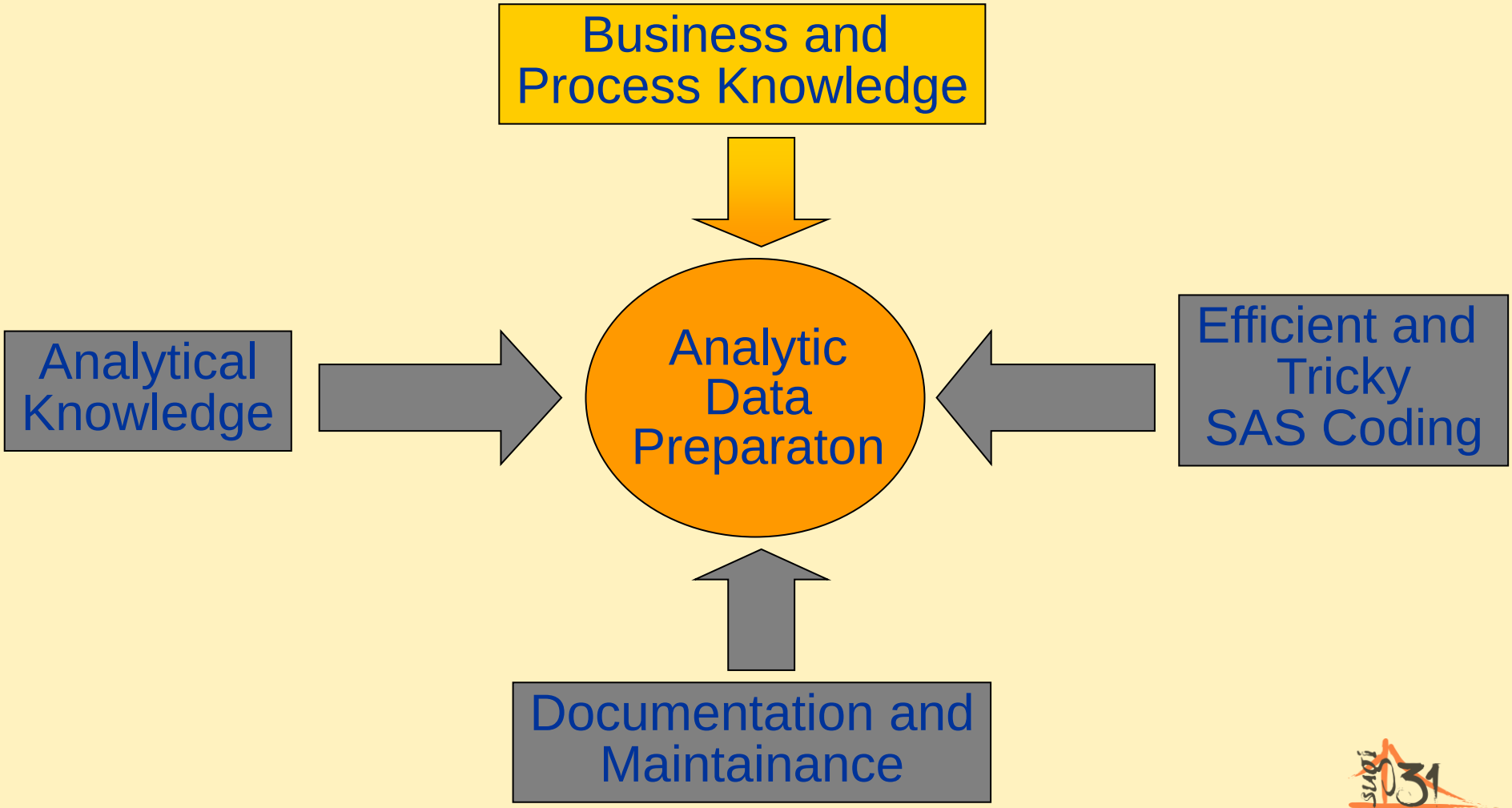
Data Sources	Data Preparation	Analytic Modeling	Results and Actions
Different Data Sources	Merges, Denormalisation	Modeling, Parameter Estimation, Tuning,	Usage of Results
Relational Models, Star Schemes	Derived Variables	Predictions, Classifications, Clustering	Profiling
	Transpositions, Aggregations		Interpretations

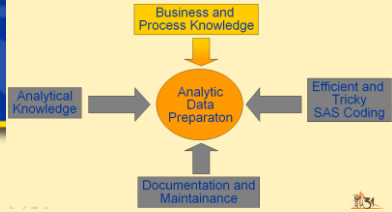
Data Availability + **Adequate Preparation** + Clever Modeling = Good Results

Four Dimensions for Analytic Data Preparation



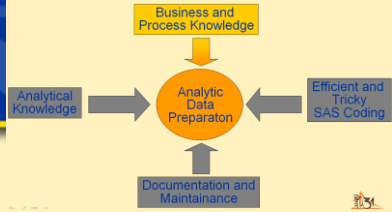
Four Dimensions for Analytic Data Preparation





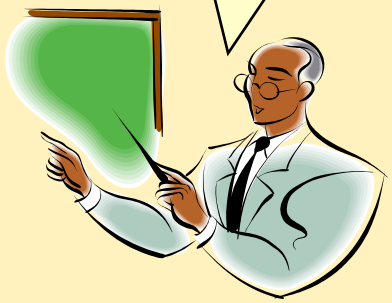
Challenging a Business Question

- „Can you calculate the probability that a customer will cancel usage of product X?“
- Questions:
  - Cancellation or decline in usage?
  - How do we treat usage of more advanced products?
  - Include non-volontary cancellation?
  - How quickly can retention measures take place?
  - Do you want to have probabilites per customer or a list of the 10.000 high risk customers or risk classes in general?



Business, Statistican and IT: The Optimal Triangle !?

I have formulated my business question. Are there any reasons, not to be back with results in 2 days.



Simon
Retention Manager

Analytic Data Preparaton

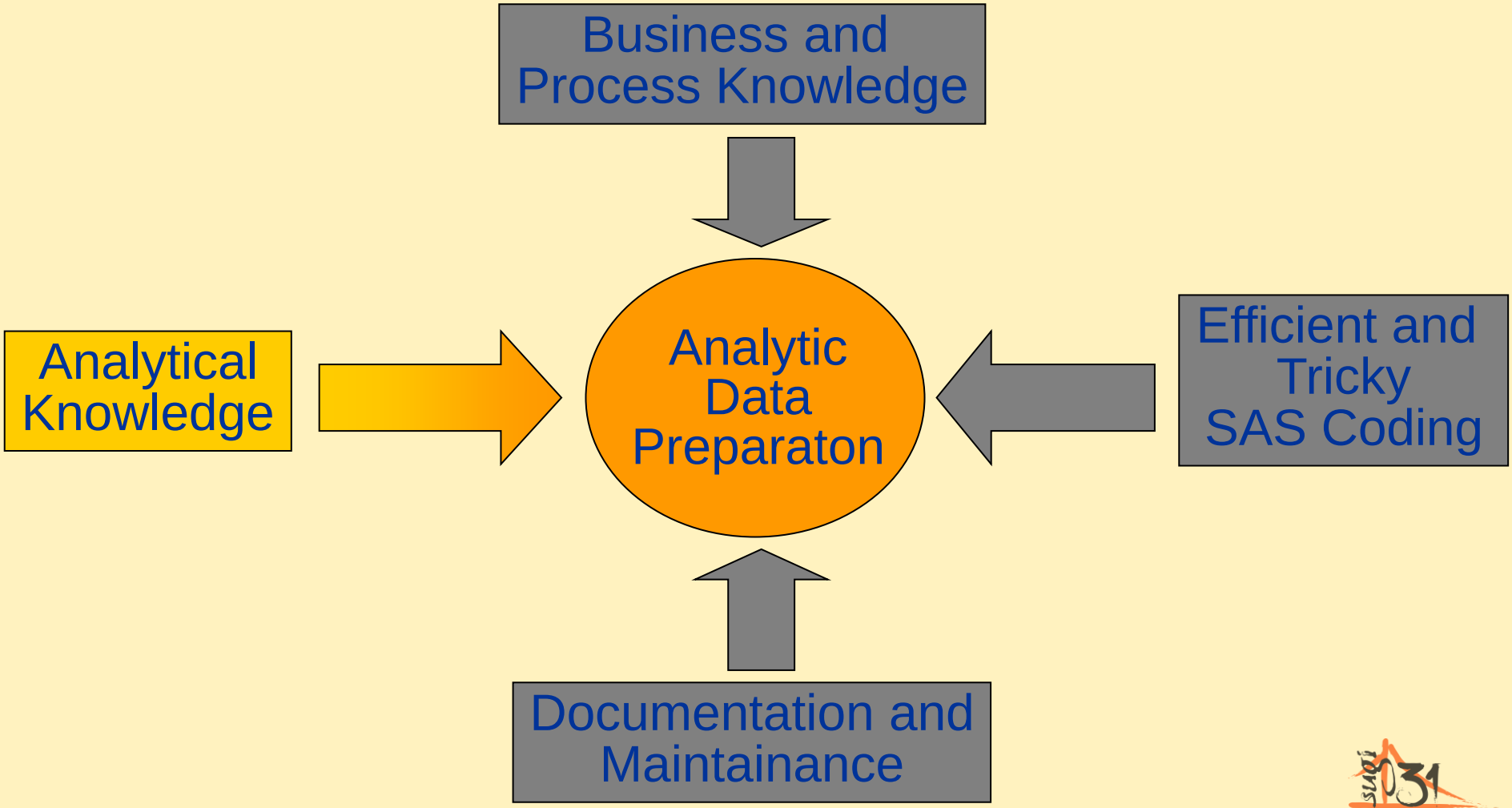
I would like to have an analytic database with a high number of attributes and an environment to perform both, data management and analysis.

Daniele
Quantitative Expert

Name me the list of attributes and derived variables that you will use in your final model and which I have to deliver periodically.

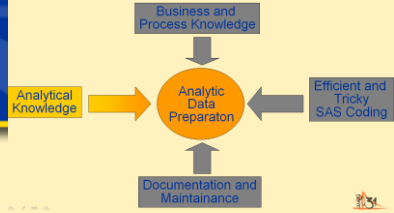
Elias
IT Expert

Four Dimensions for Analytic Data Preparation



Business Questions, Analytical Models

- Event prediction (Churn, Fraud, Delinquency, Response, …)
- Value prediction (Purchase Size, Claim Amount, …)
- Clustering (Segmentation, …)
- Market Basket Analysis (Association Analysis, …)
- Time Series Forecasting



Analysis Subjects and Multiple Observations

- *Analysis subjects* are entities that are being analyzed and the analysis results are interpreted in their context.

- *Multiple observations per analysis subject*
  - Repeated measurements over time
  - Multiple observations because of hierarchical relationships

Main Types of Data Marts

One-Row-per-Subject Data Mart

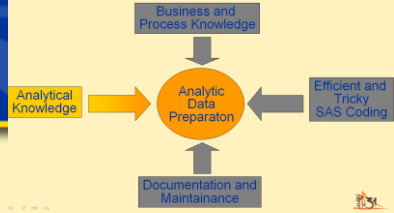
	Customer ID	Date of Birth	Age (years)	Gender	Marital Status	Academic Title	Has Title? 0/1	Branch Name	Customer Start Date	Customer Duration (months)
1	1000002	26DEC1958	44	Male	Married		0	Fil1	01JAN2000	41
2	1000005	25JUN1947	56	Male	Single	Ing.	1	Fil4	01APR1999	50
3	1000006	10DEC1945	57	Female	Married		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	corned_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	corned_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

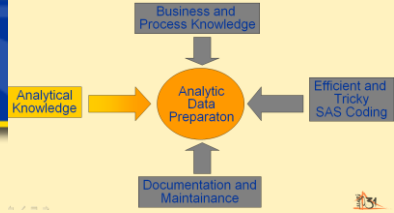
Longitudinal Data Mart

	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 Structures

	Data Mart Structure for the Analysis	
Structure of the source data: "Multiple observations per analysis subject exist?"	One-Row-per-Subject Data Mart	Multiple-Row-per-Subject Data Mart
NO		
YES		



The One-Row-Per-Subject Paradigm

Analysis Subject Master Table

ID	Birth	Sex	Region	...	...
1					
2					
3					
4					

Multiple Observations Per Analysis Subject

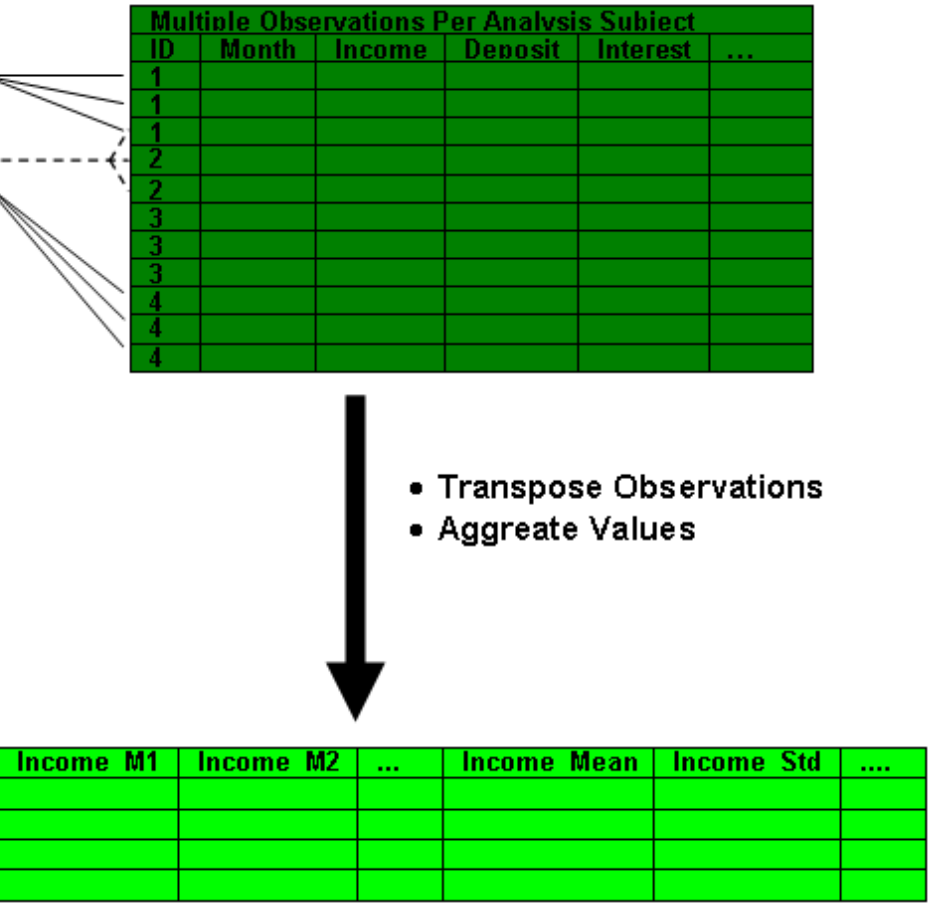
ID	Month	Income	Deposit	Interest	...
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					

- Copy Variables
- Create Derived Variables

- Transpose Observations
- Aggreate Values

ID	Birth	Sex	Region	Age	....	Income M1	Income M2	...	Income Mean	Income Std	....
1											
2											
3											
4											

The One-Row-Per-Subject Paradigm
Clever Aggregations



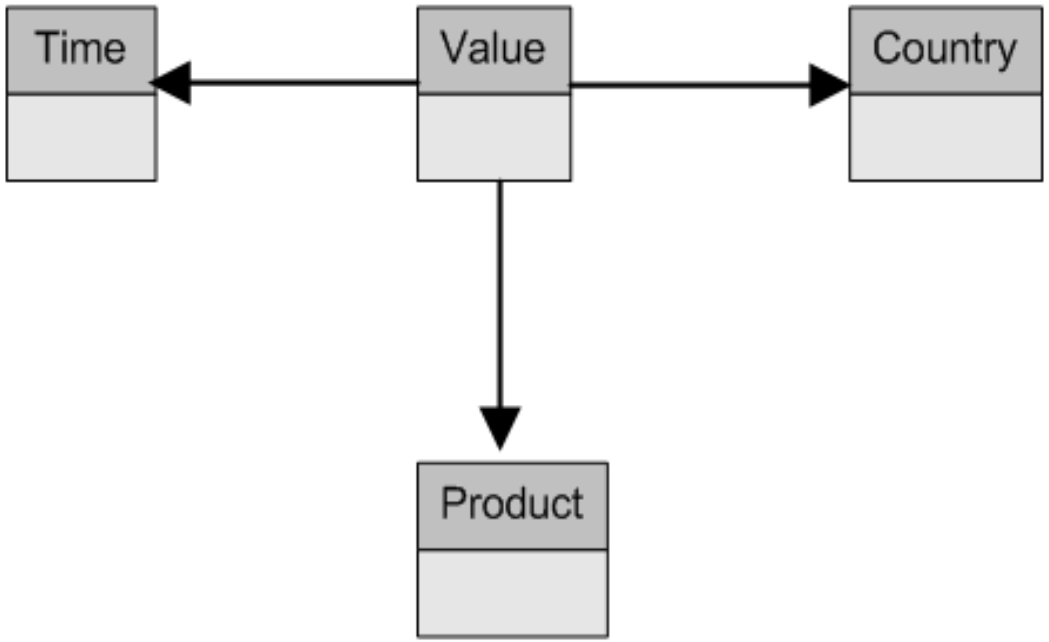
Interval Data

- Static Aggregation
- Correlation of Values
- Course over Time
- Concentration of Values

Categorical Data

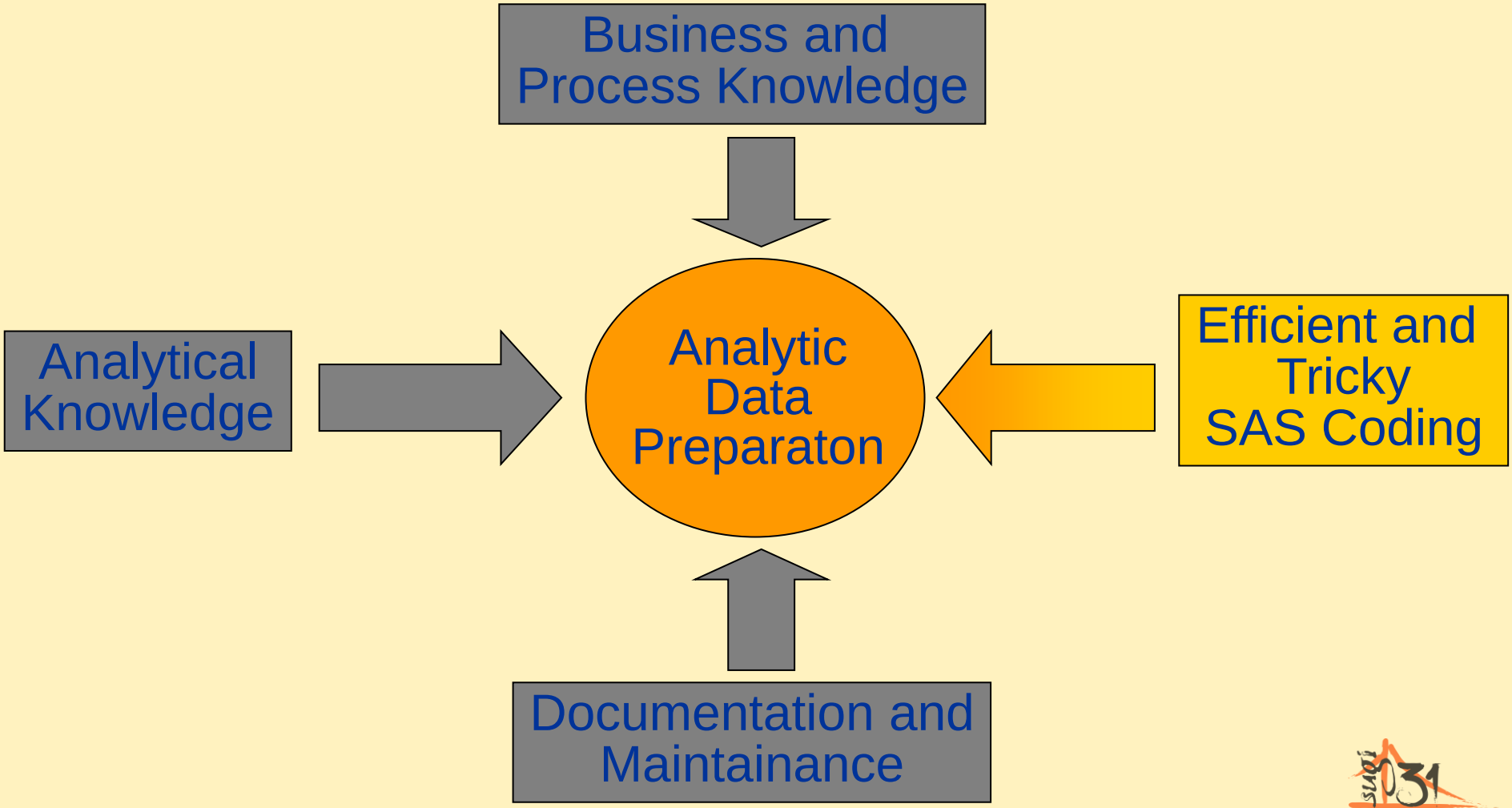
- Frequency Counts
- Concatenated Frequencies
- Total and Distinct Counts

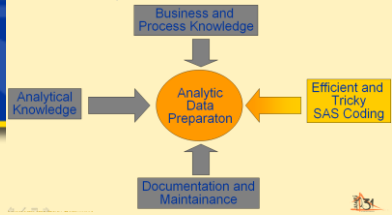
Longitudinal Data Marts



- Aggregating data at the right level
- Defining cross sectional groups
- Alignment of time values
- Creation of event indicators and input variables

Four Dimensions for Analytic Data Preparation





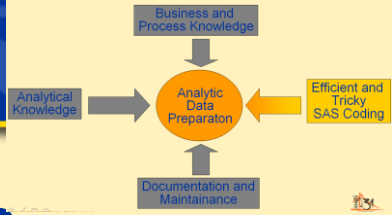
SAS Data Step vs. SQL

Transposing a table

	CUSTOMER	TIME	PRODUCT
1	0	0	hering
2	0	1	corned_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	corned_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 = sampsio.assocs(obs=21)
               OUT = assoc_tp (DROP = _name_);
 BY customer;
 ID Product;
RUN;
```

	CUSTOMER	bourbon	corned_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	.	.



Changing between longitdutinal data structures

Standard Form

←→

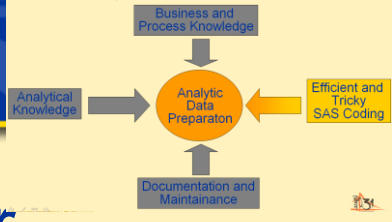
Interleaved

```
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

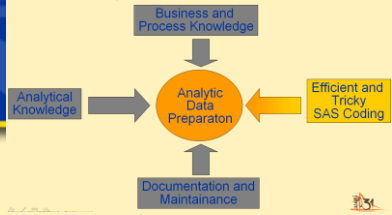
```
PROC TRANSPOSE DATA = diy intleaved
               OUT  = diy standard back
                      (drop =  name );
 BY date;
 ID Type;
 VAR value;
```



Selection of the first and last line per customer

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

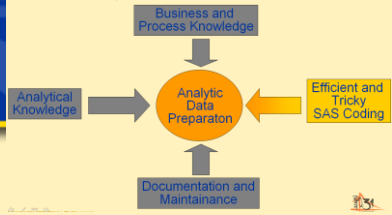
```
data customer;
 set customer;
  by CustID;
   FirstValue = First.CustID;
   LastValue  = Last.CustID;
run;
```



Creating a sequential number and summing items

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

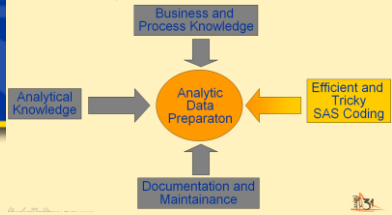
```
data customer;
 set customer;
 by CustID;
 if first.custid then do;    Purch_No = 1;
                             Cum_Poi  = Points;    end;
 else do;   Purch_No + 1;
            Cum_Poi + Points;   end;
run;
```



Copying omitted data

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

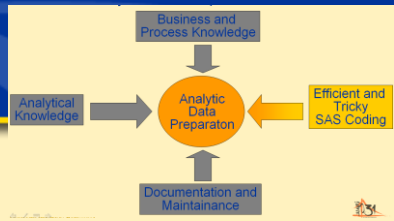
```
data customer;
 set customer;
 retain age_tmp;
 if age ne "" then age_tmp = age;
 else age = age_tmp;
run;
```



Shift down a column for k positions

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

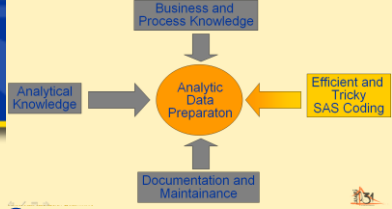
```
data measurements;
 set measurements;
  Value_PrevDay = lag(Value);
run;
```



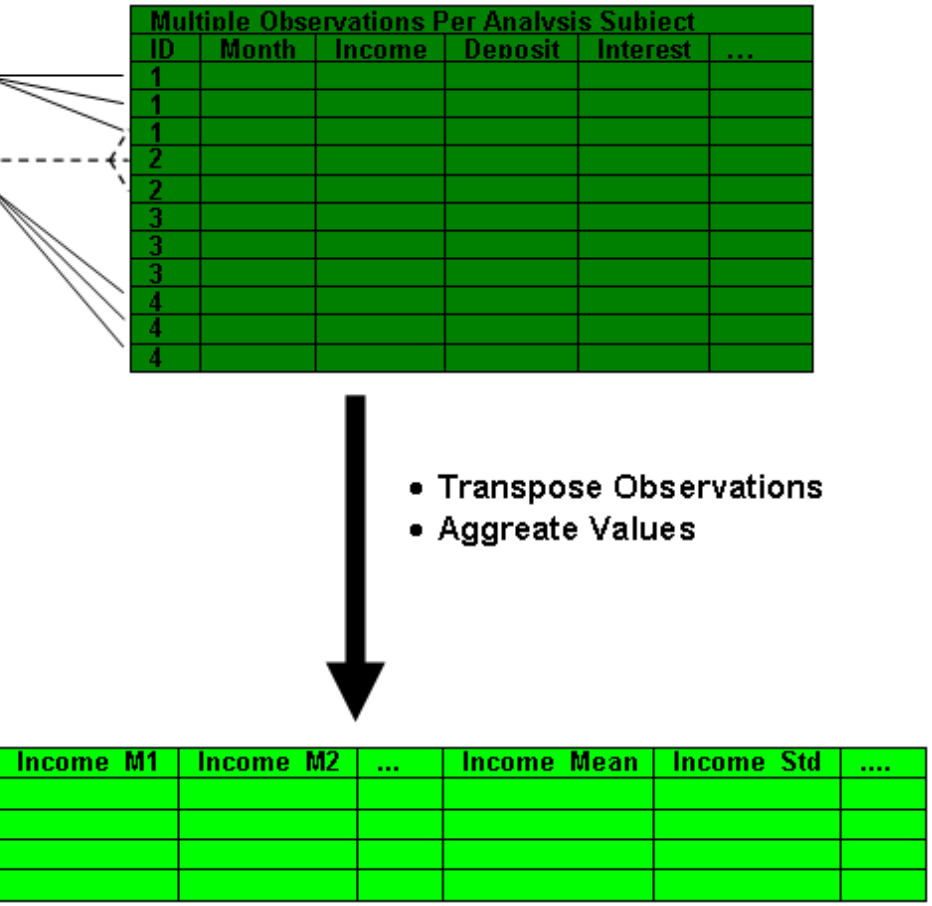
Measures for the Course over Time

	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	1.685714285	-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;
```



The One-Row-Per-Subject Paradigm
Clever Aggregations

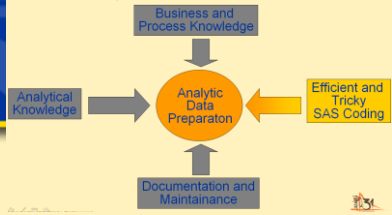


Interval Data

- Static Aggregation
- Correlation of Values
- Course over Time
- Concentration of Values

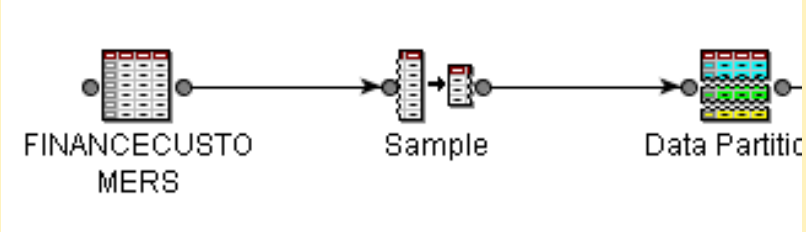
Categorical Data

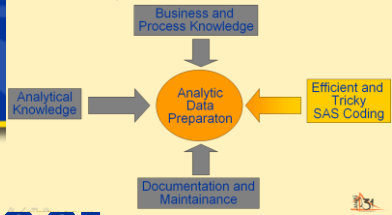
- Frequency Counts
- Concatenated Frequencies
- Total and Distinct Counts



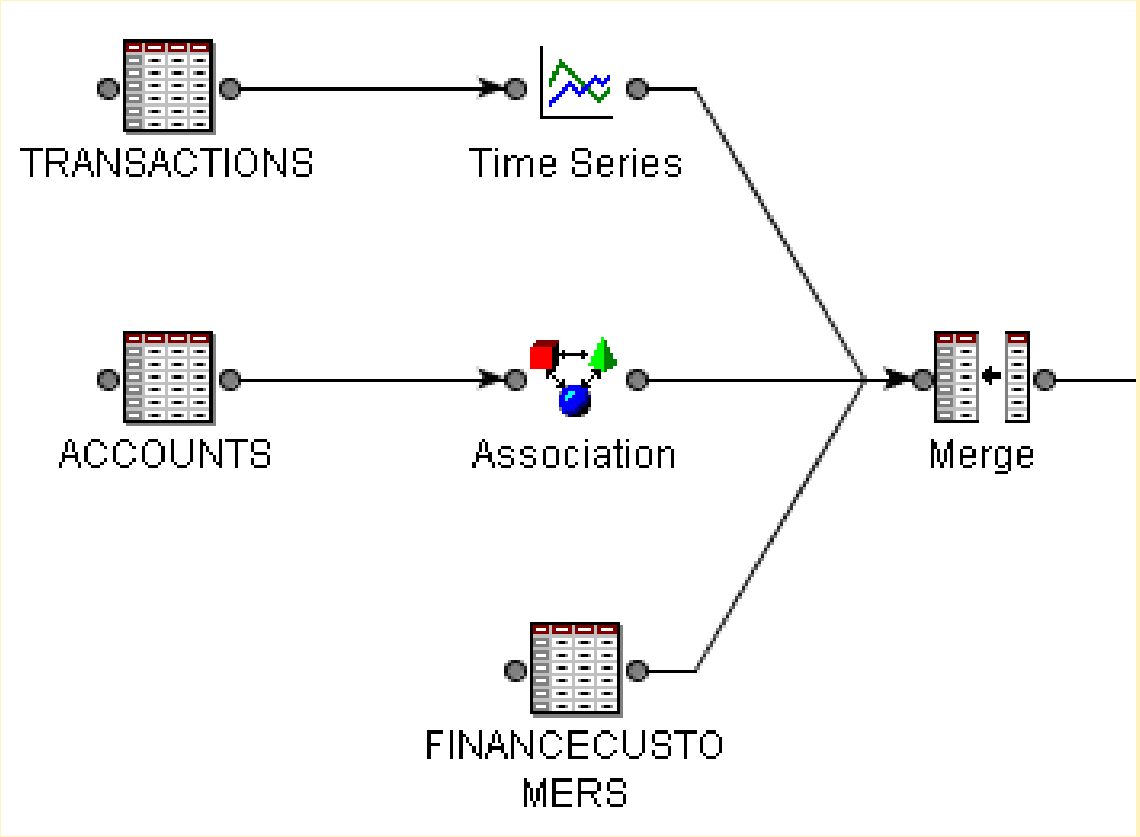
Analytic Data Preparation in SAS® Enterprise Miner

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

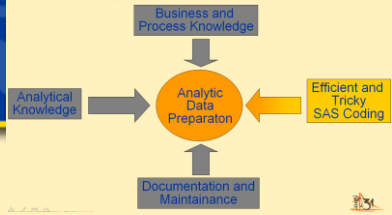




Working with Multiple-Row-per-Subject Data in SAS® Enterprise Miner

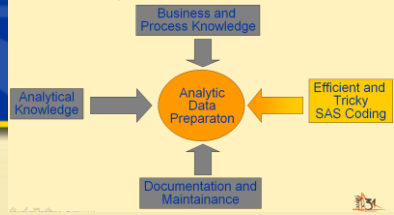


- Time Series Node
- Association Node
- Merge Node

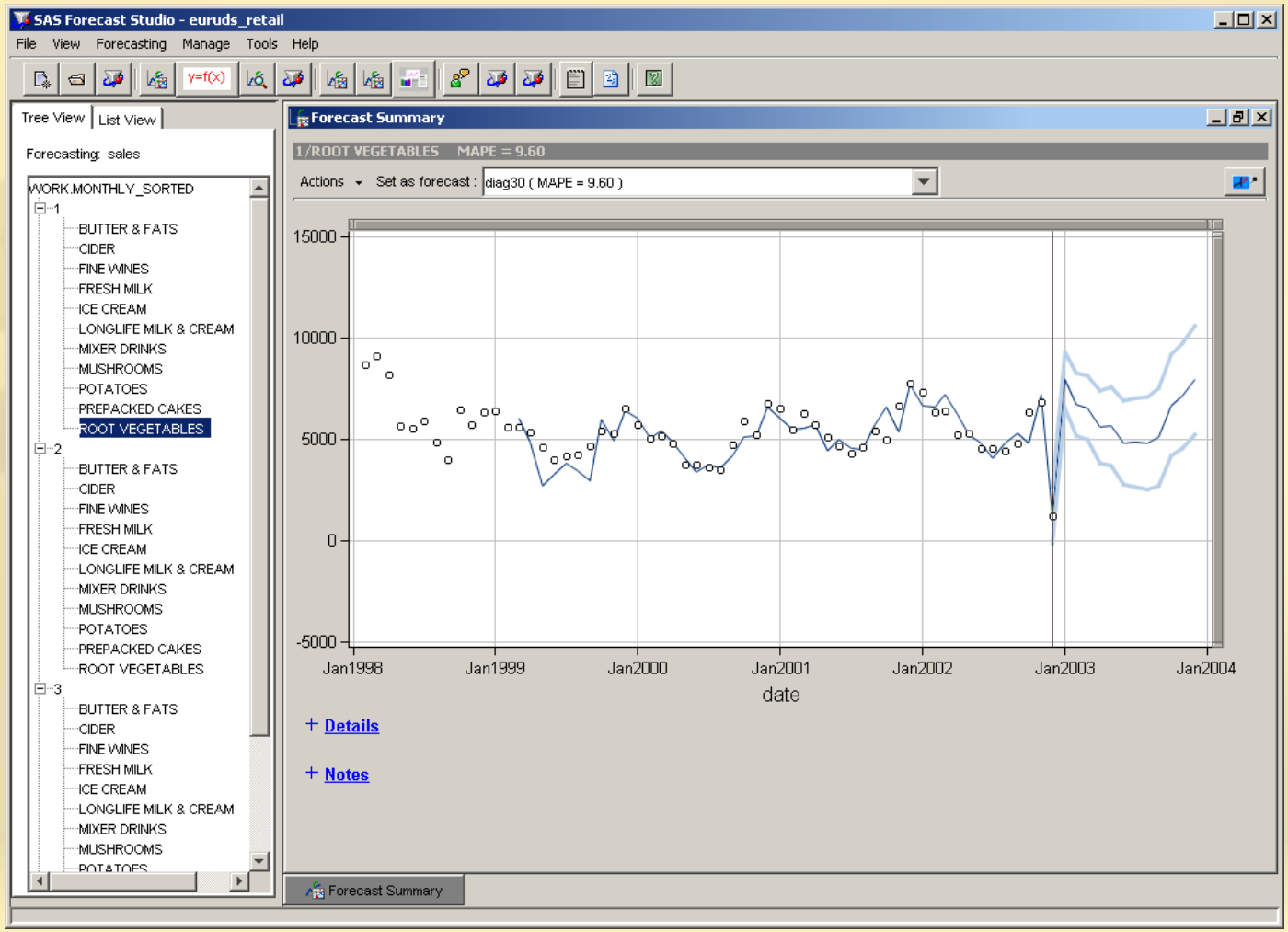


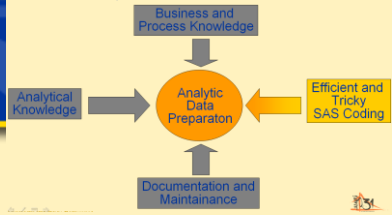
Analytic Data Preparation in SAS® Enterprise Miner - Summary

- Documentation of the data preparation process as a whole in the process flow diagram
- Definition of variables metadata that are used in the analysis
- Automatic Creation of Dummy-Variables for categorical data
- Powerful data transformations in the filter node, transform variables node and impute node.
- Possibility to perform association analysis and time-series-analysis
- Creation of score code from all nodes in SAS Enterprise Miner as SAS datastep code



SAS® Forecast Studio - Screenshot

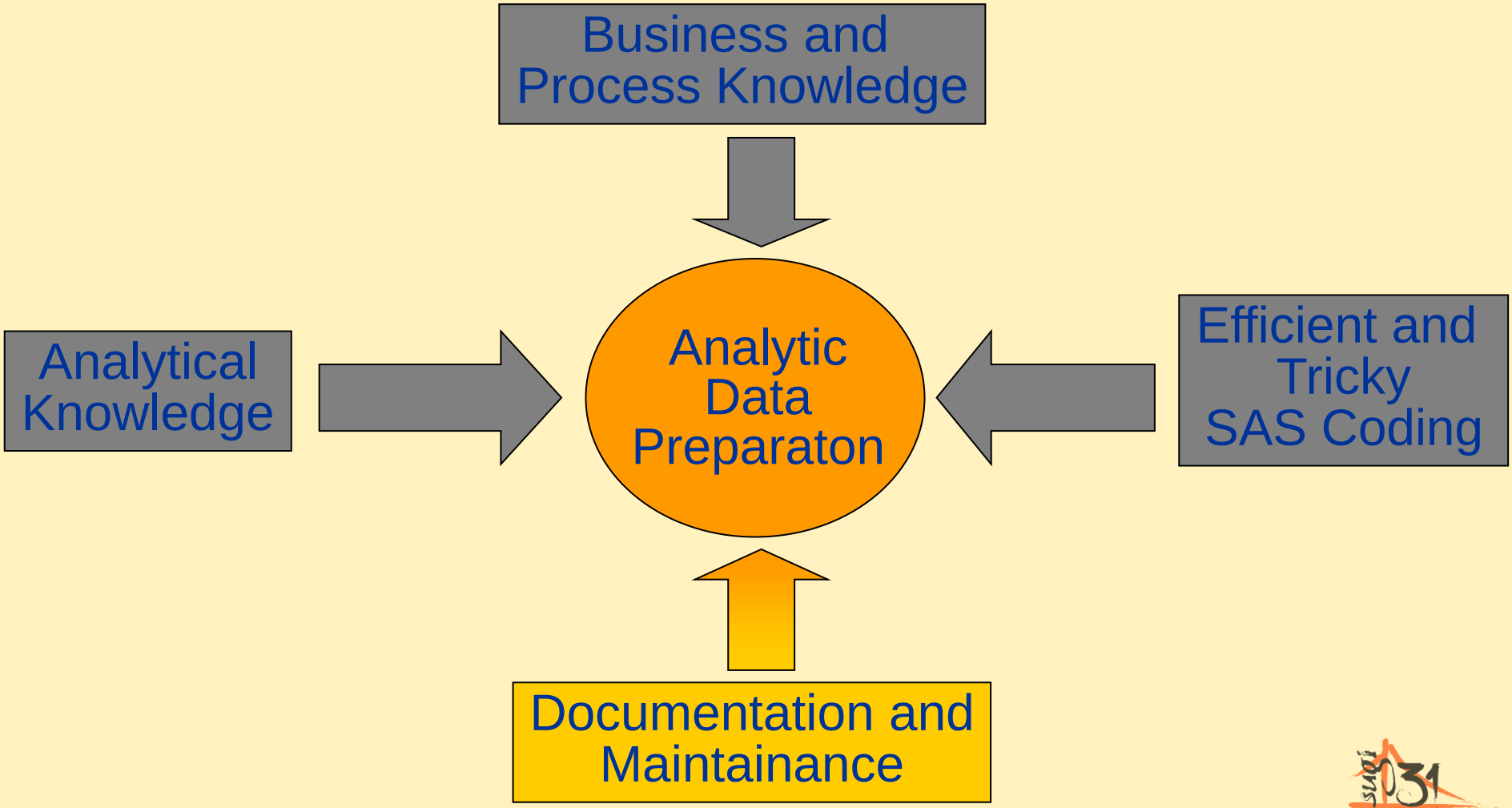


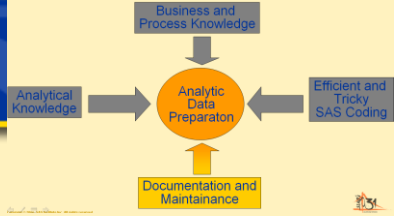


Analytic Data Preparation in SAS® Forecast Studio

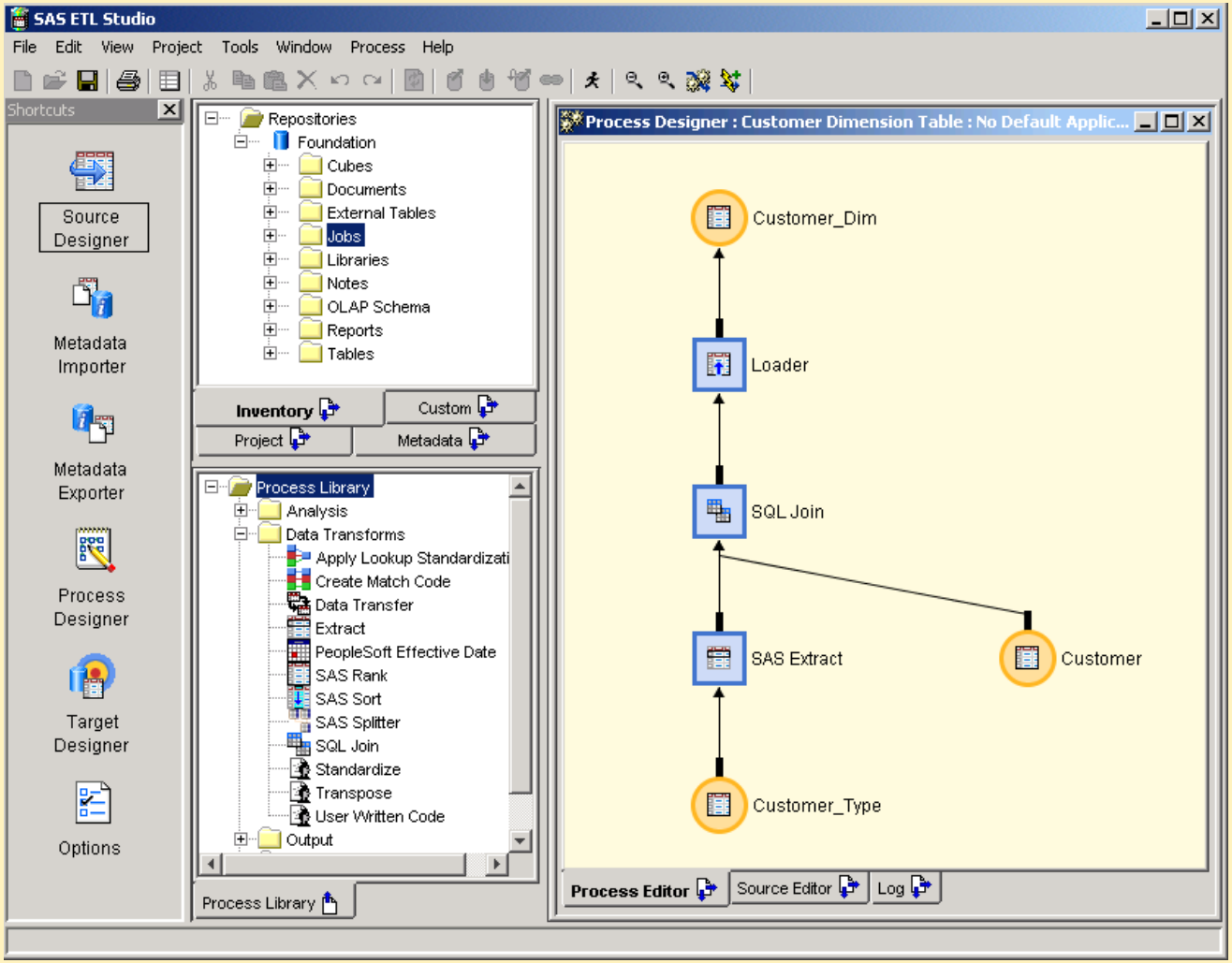
- Convert transactional data into time series data
- Treatment of missing values
- Alignment of date values in intervals
- Aggregation of data on various levels
- Handling of events
  - Allows users to create event definitions, assign events to selected series in the project and delete events
  - Users can specify event duration, shape, and recurrence options.
  - Pre-defined common events and holidays are available for inclusion in the forecasting models

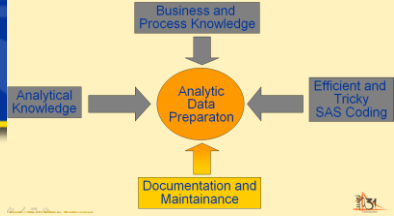
Four Dimensions for Analytic Data Preparation





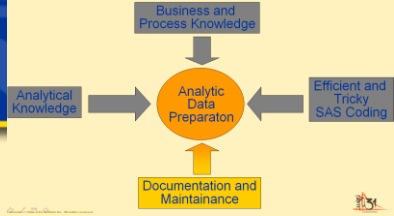
SAS® ETL Studio





SAS® ETL Studio – General Features

- Comprehensive transformation library
- Graphical user interface, drag & drop, wizard driven
- Multi-developer support: Check-in/check-out, change management
- Impact analysis
- **Documentation of the data management process**



SAS® ETL Studio – Analytic Features

- Data Mining Model Scoring
  - Register Model in SAS metadata repository
  - Apply SAS Enterprise Miner model to new data source
  - Creates the target table definition in the metadata
- Forecast Analysis Transformation
  - Allows to perform time series analysis
  - Based on HPF (high performance forecasting) procedure
  - Allows to integrate the forecast step into the data flow process in the metadata

Summary

- Analytic Data Preparation is a discipline, not a incommodious necessity
- Analytic Data Preparation is more than just coding
- SAS combines powerful data management functionality and market leading analytics in one integrated package
- SAS Tools like SAS® ETL-Studio, SAS® Enterprise Miner and SAS® Forecast Studio assist you in analytic data preparation

Recommended Reading

Data Preparation for Analytics
by Gerhard Svolba

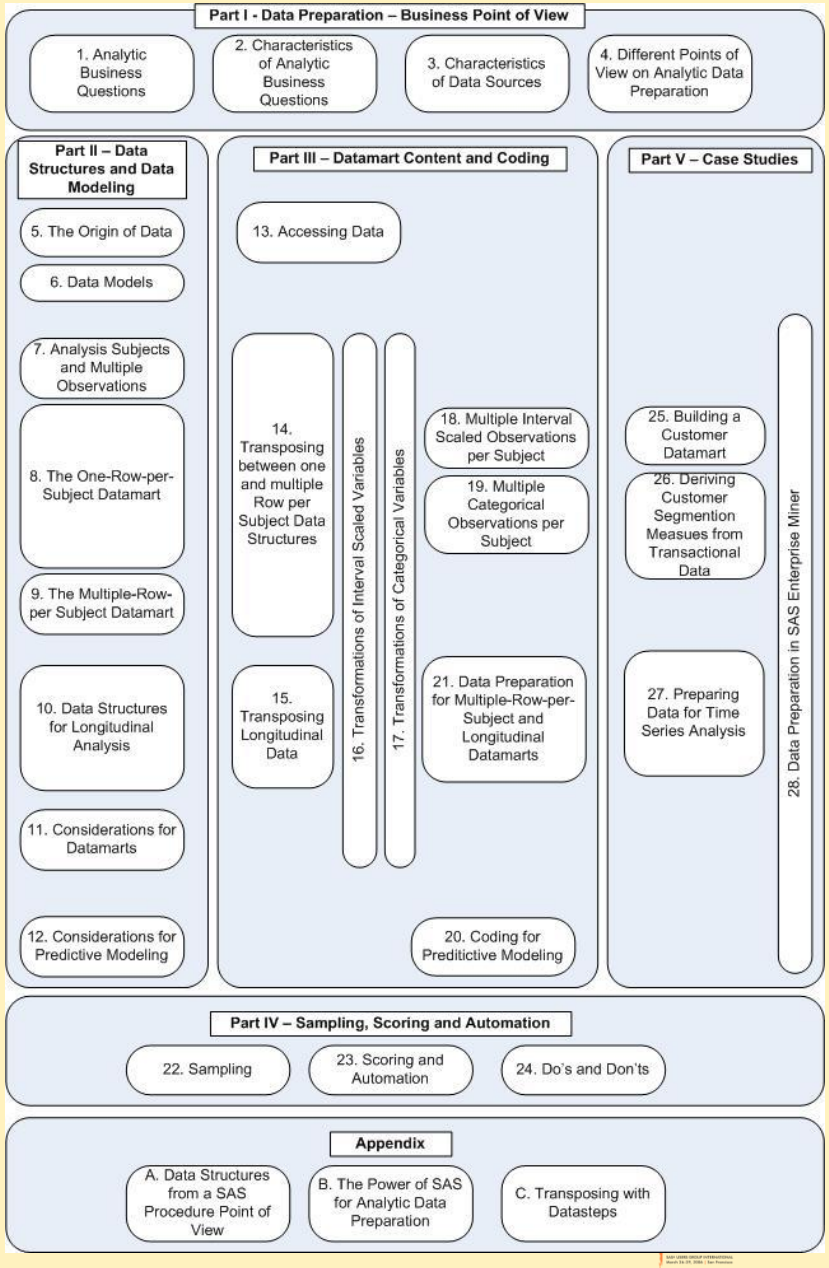
SAS-Press (#60502)

Planned publication date: Oktober 2006

Business Rationale

Concepts

Coding Examples



Questions and Contact

- Gerhard Svolba (PhD)
- Email: gerhard.svolba@aut.sas.com
- Post address:
  - SAS-Austria
  - Mariahilfer Straße 116
  - 1070 Wien
  - Austria

- Wednesday, 29th; 8:00; Room 2007
  - Paper 078-31:
  - Efficient Construction of a One-Row-per-SubjectData Mart for Data Mining.ppt