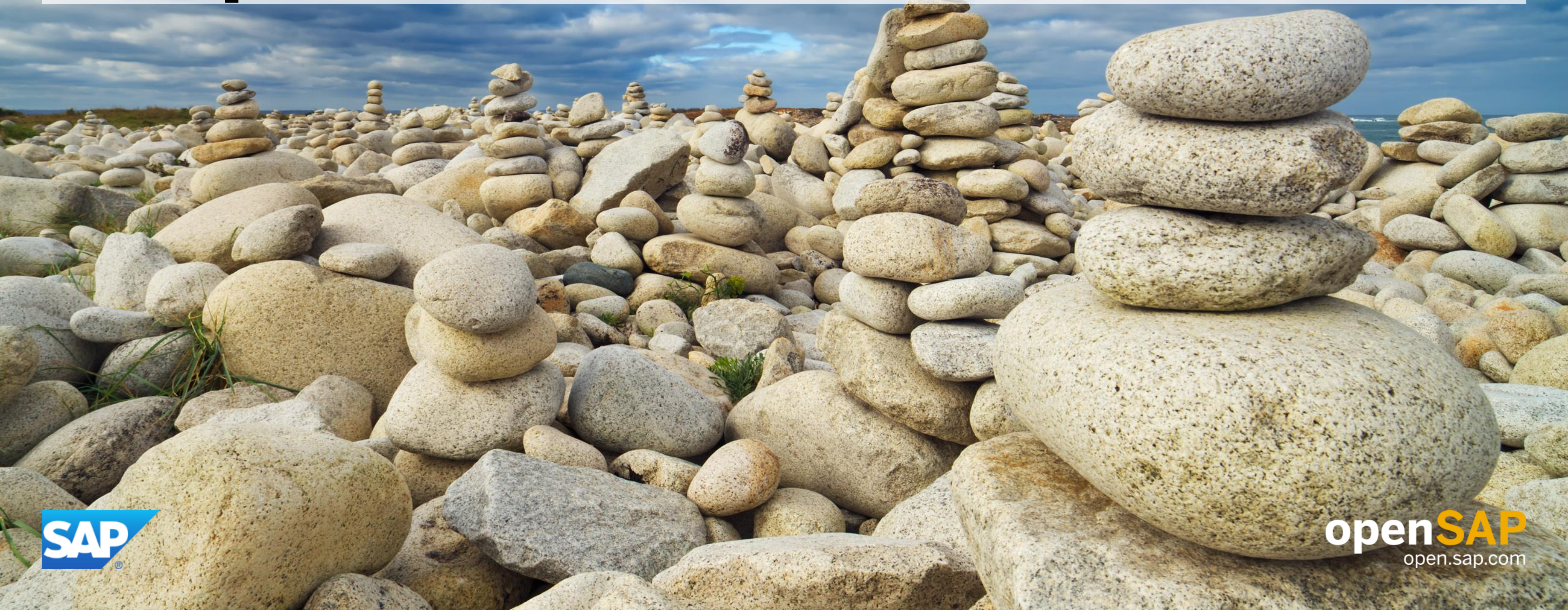
Week 2 Unit 1: **Data Preparation Phase – Overview**



openSAP
open.sap.com

Data Preparation Phase – Overview

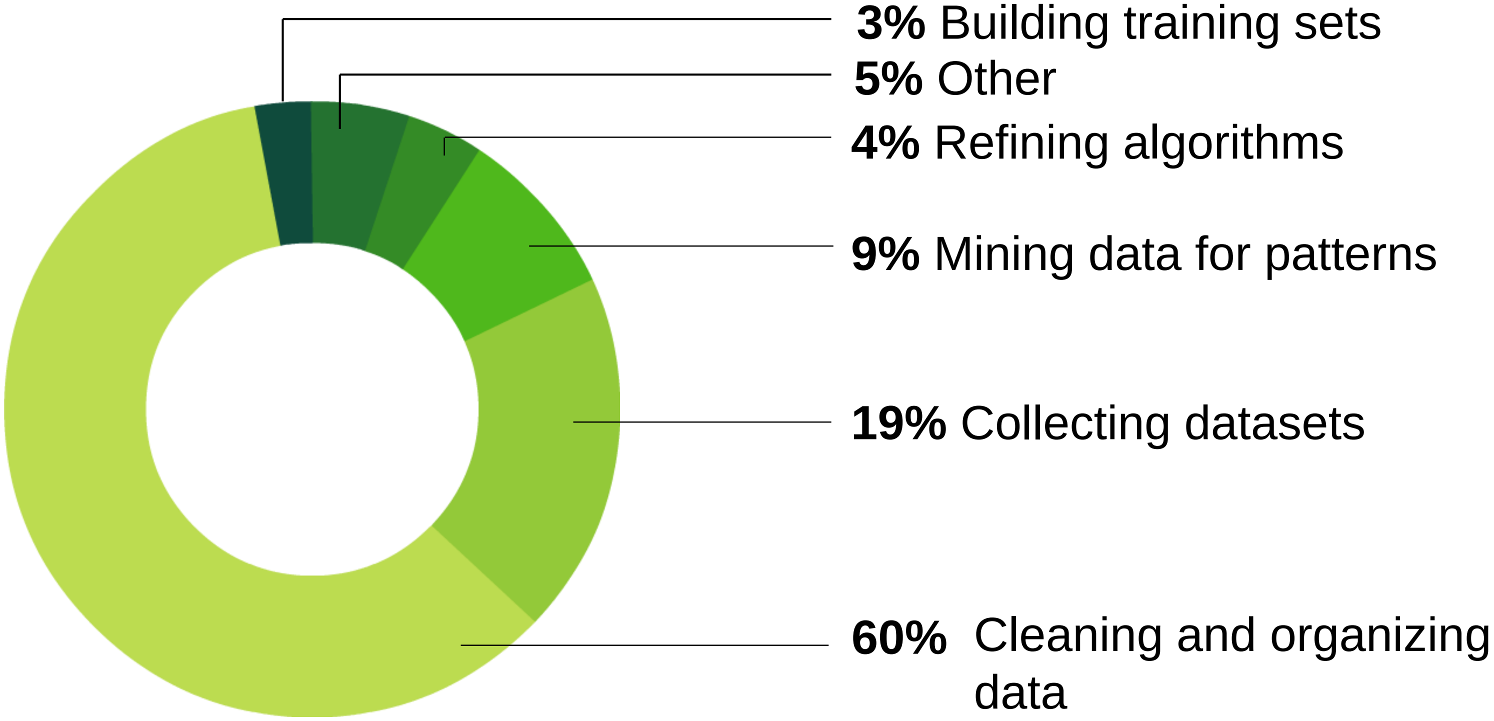
Introduction to data preparation

The chart below shows that 3 out of every 5 data scientists spend the most time during their working day cleaning and organizing data.

New York Times article reported that data scientists spend from **50% to 80%** of their time mired in the more mundane task of collecting and preparing unruly digital data before it can be explored for useful nuggets.

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. New York Times. STEVE LOHR. AUG. 17, 2014

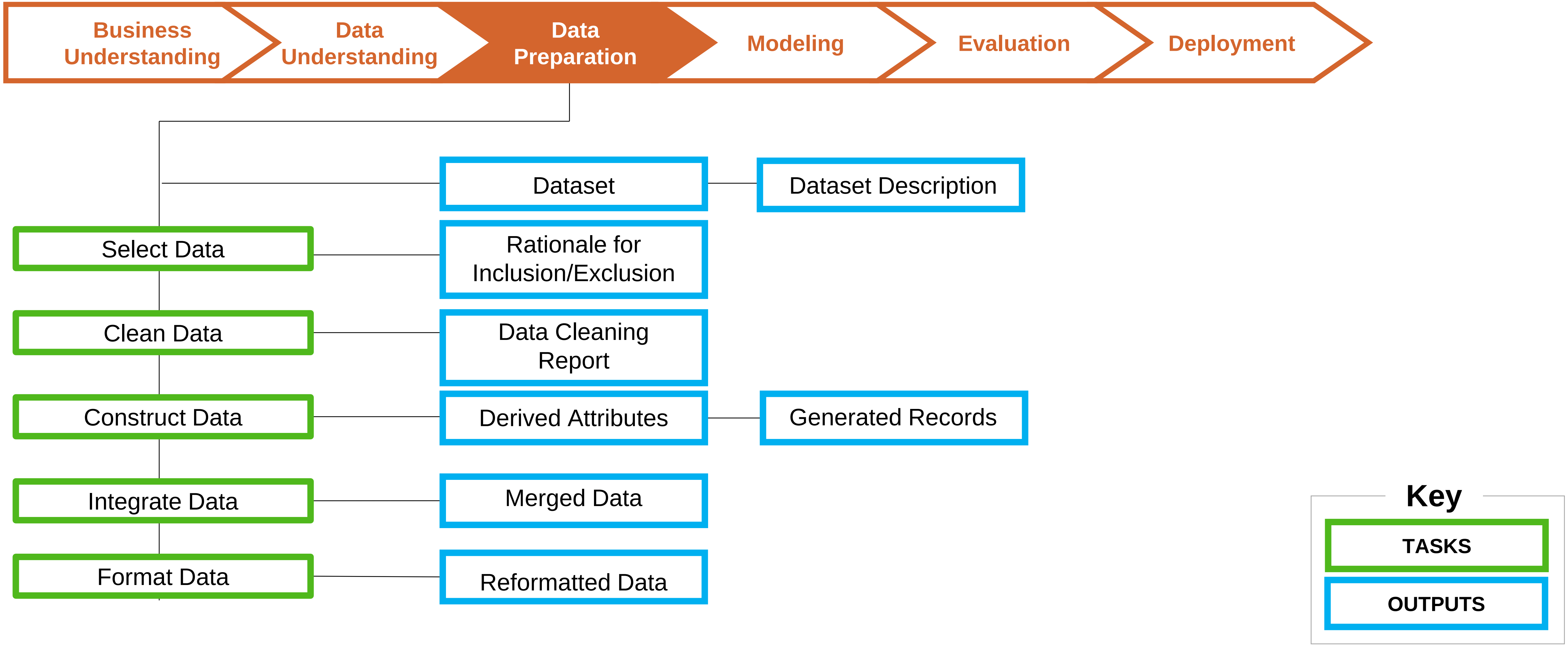
What data scientists spend the most time doing



CrowdFlower Data Science Report 2016

Data Preparation Phase – Overview

CRISP-DM – Phase 3: Data Preparation



Data Preparation Phase – Overview

Phase 3: Outputs

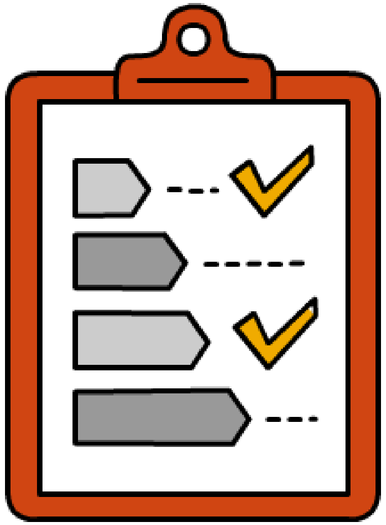
- **Dataset**
  - This is the dataset (or datasets) produced by the *Data Preparation* phase, which will be used for modeling or the major analysis work of the project.
- **Dataset description**
  - Describe the dataset (or datasets) that will be used for the modeling or the major analysis work of the project.

Data Preparation Phase – Overview

Phase 3.1: Select Data

- **Task**
  - Decide on the data to be used for analysis.
  - Criteria include relevance to the data mining goals and quality and technical constraints such as limits on data volume or data types.
  - Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

- **Output – Rationale for inclusion/exclusion**
  - List the data to be included/excluded and the reasons for these decisions.



Data Preparation Phase – Overview

Phase 3.2: Clean Data

- **Task**
  - Raise the data quality to the level required by the selected analysis techniques.
  - This may involve selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.

- **Output – Data cleaning report**
  - Describe what decisions and actions were taken to address the data quality problems reported during the *Verify Data Quality* task of the *Data Understanding* phase.

Data Preparation Phase – Overview

Phase 3.3: Construct Data

- **Task**
  - This task includes constructive data preparation operations such as the production of derived attributes, entire new records, or transformed values for existing attributes.
- **Output – Derived attributes**
  - Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. Examples: *area = length * width.*
- **Output – Generated records**
  - Describe the creation of completely new records.

Data Preparation Phase – Overview

Phase 3.4: Integrate Data

- **Task**
  - These are methods whereby information is combined from multiple tables or records to create new records or values.

- **Output – Merged data**
  - Merging tables refers to joining together two or more tables that have different information about the same objects.
  - Merged data also covers aggregations.

Data Preparation Phase – Overview

Phase 3.5: Format Data

- **Task**
  - Formatting transformations refer to primarily *syntactic* modifications made to the data that do not change its meaning, but might be required by the modeling tool.

- **Output – Reformatted data**
  - Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Thank you

Contact information:

open@sap.com

openSAP
open.sap.com

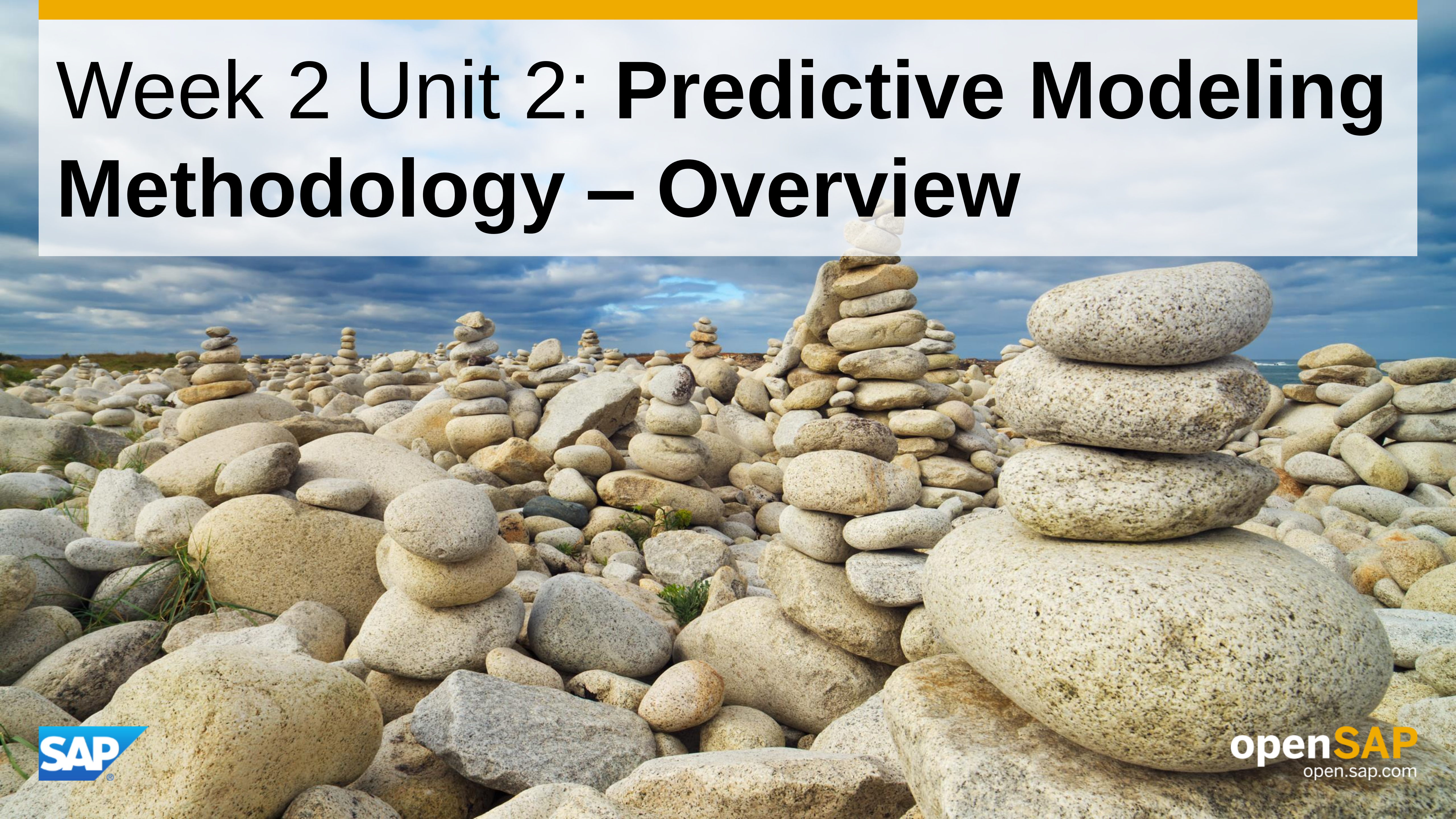
Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors.

National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP SE or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP SE or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.

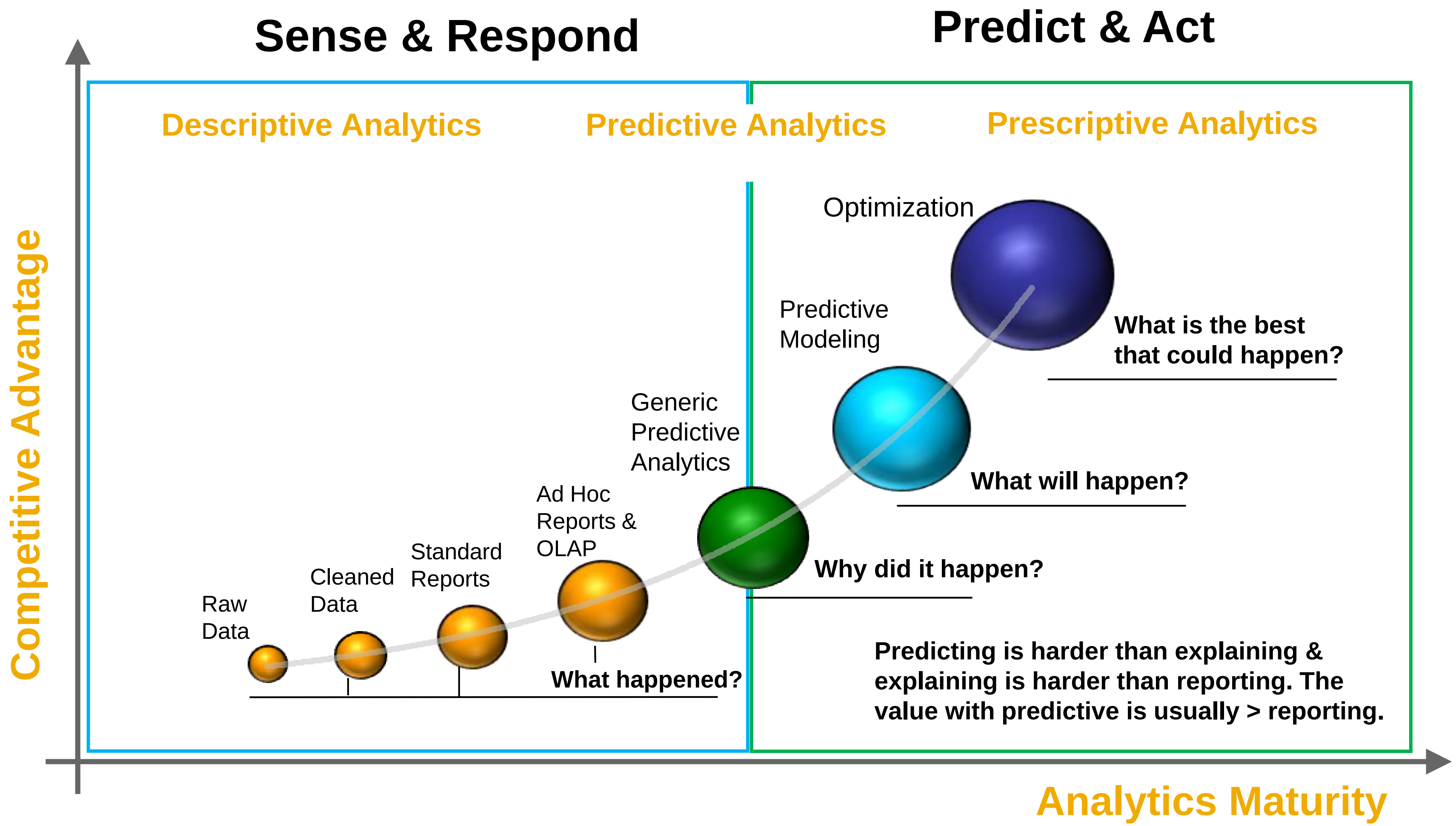
Week 2 Unit 2: **Predictive Modeling Methodology – Overview**



openSAP
open.sap.com

Predictive Modeling Methodology – Overview

Introduction



Predicting is harder than explaining & explaining is harder than reporting. The value with predictive is usually > reporting.

The key is unlocking data to move decision making from sense & respond to predict & act

Predictive Modeling Methodology – Overview

Use predictive analytics to solve a variety of business challenges

SALES & MARKETING

- Churn Reduction
- Customer Acquisition
- Lead Scoring
- Product Recommendation
- Campaign Optimization
- Customer Segmentation
- Next Best Offer/Action

OPERATIONS

- Predictive Maintenance
- Load Forecasting
- Inventory/Demand Optimization
- Product Recommendation
- Price Optimization
- Manufacturing Process Optimization
- Quality Management
- Yield Management

FRAUD & RISK

- Fraud and Abuse Detection
- Claims Analysis
- Collection and Delinquency
- Credit Scoring
- Operational Risk Modeling
- Crime Threat
- Revenue and Loss Analysis

FINANACE & HR

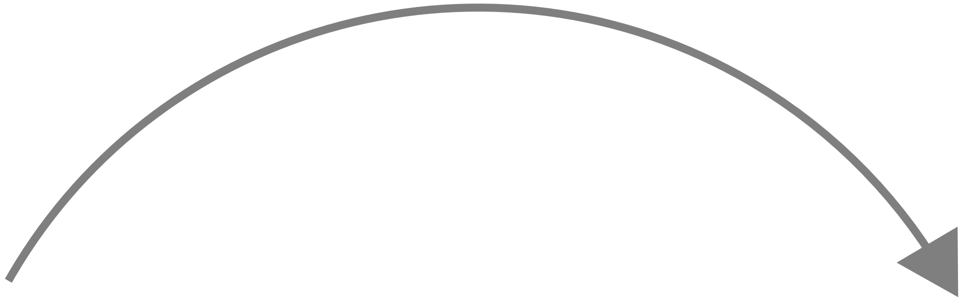
- Cash Flow and Forecasting
- Budgeting Simulation
- Profitability and Margin Analysis
- Financial Risk Modeling
- Employee Retention Modeling
- Succession Planning

OTHER SECTORS

- Life Sciences
- Healthcare
- Media
- Higher Education
- Public Sector / Social Sciences
- Construction and Mining
- Travel and Hospitality
- Big Data and IoT

Predictive Modeling Methodology – Overview

Build and Apply



Model Build

(the Learning Phase)

Predictive models are built or “trained” on historic data with a known outcome.

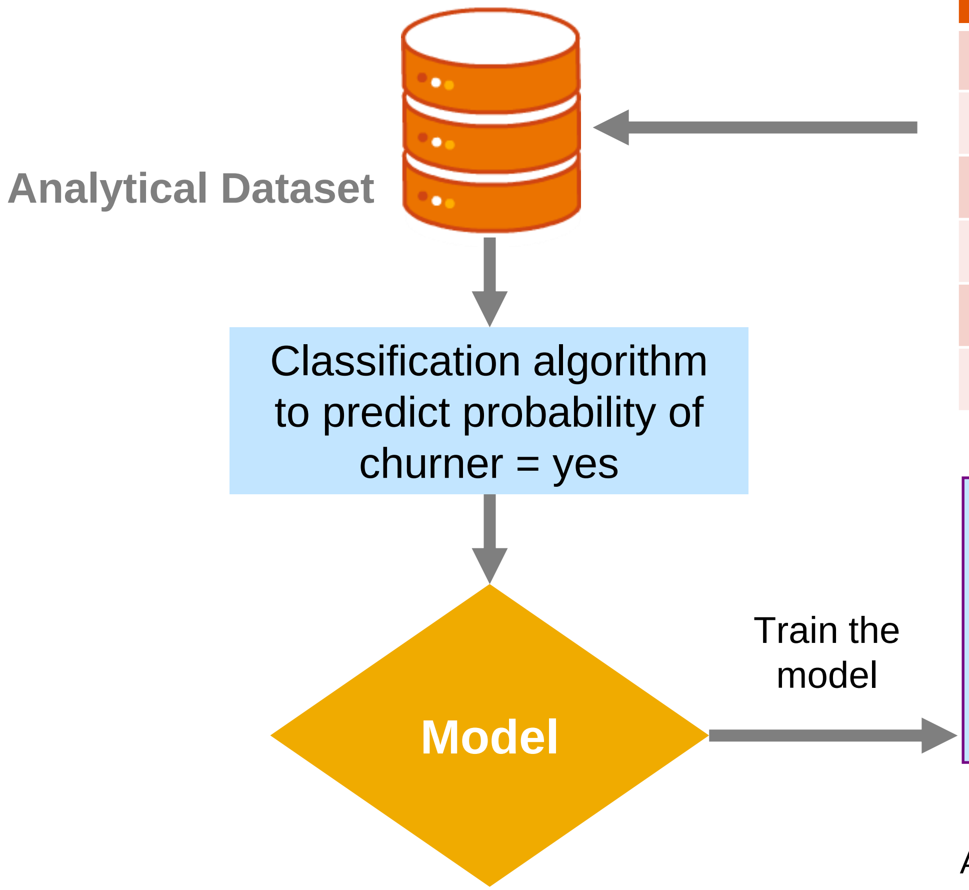
Model Apply

(the Applying Phase)

Once the model has been built, it is applied onto new, more recent data, which has an unknown outcome (because the outcome is in the future).

Predictive Modeling Methodology – Overview

Building the model – Learning phase



Analytical Dataset

Classification algorithm to predict probability of churner = yes

Model

Train the model

Explanatory Variables

Target

Name	City	Age	Churner
Mike	Miami	42	**yes**
Jerry	New York	32	**no**
Bryan	Orlando	18	**no**
Patricia	Miami	45	**yes**
Elodie	Phoenix	35	**no**
Remy	Chicago	72	**yes**

When we train the model, the outcome is known

Estimation

Validation

Create a data sub-sample called “Estimation” which we use to build the model, and a hold-out sub-sample , called “Validation” to test the model.

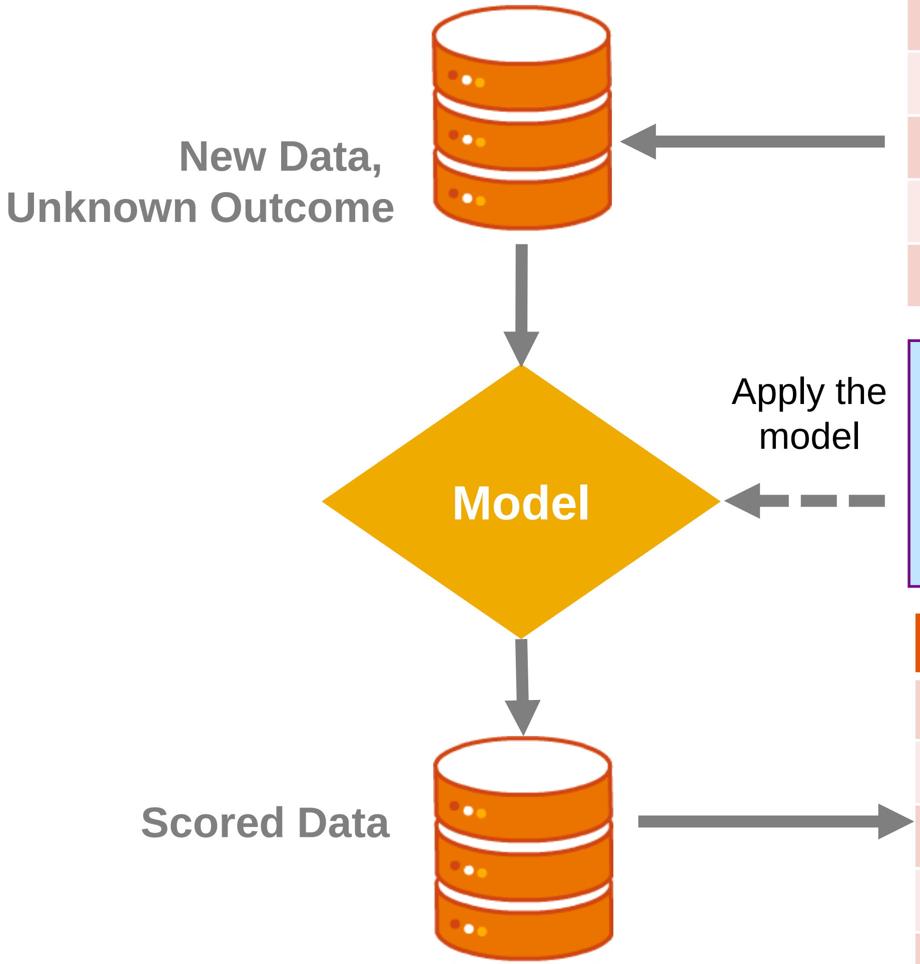
IF city= ‘Miami’ → Score = +0.7
IF city= ‘Orlando’ → Score = +0.2
IF age > 42 → Score = +0.05*age + 0.06
IF age <= 42 → Score = +0.01*age + 0.02
…..

Produce a “**scorecard**”

Add up each component score to give an overall score for each customer. This will equate to their churn probability.

Predictive Modeling Methodology – Overview

Using the model – Applying phase



Name	City	Age	Churner
Marine	Miami	45	**?**
Julien	Miami	52	**?**
Fred	Orlando	20	**?**
Michelle	Boston	34	**?**
Nicolas	Phoenix	90	**?**

Recent data, with customers who have not yet made a decision to churn or remain.

IF city= ‘Miami’ → Score = +0.7
IF city= ‘Orlando’ → Score = +0.2
IF age > 42 → Score = +age*0.05 + 0.06
IF age <= 42 → Score = +age*0.01 + 0.02
…..

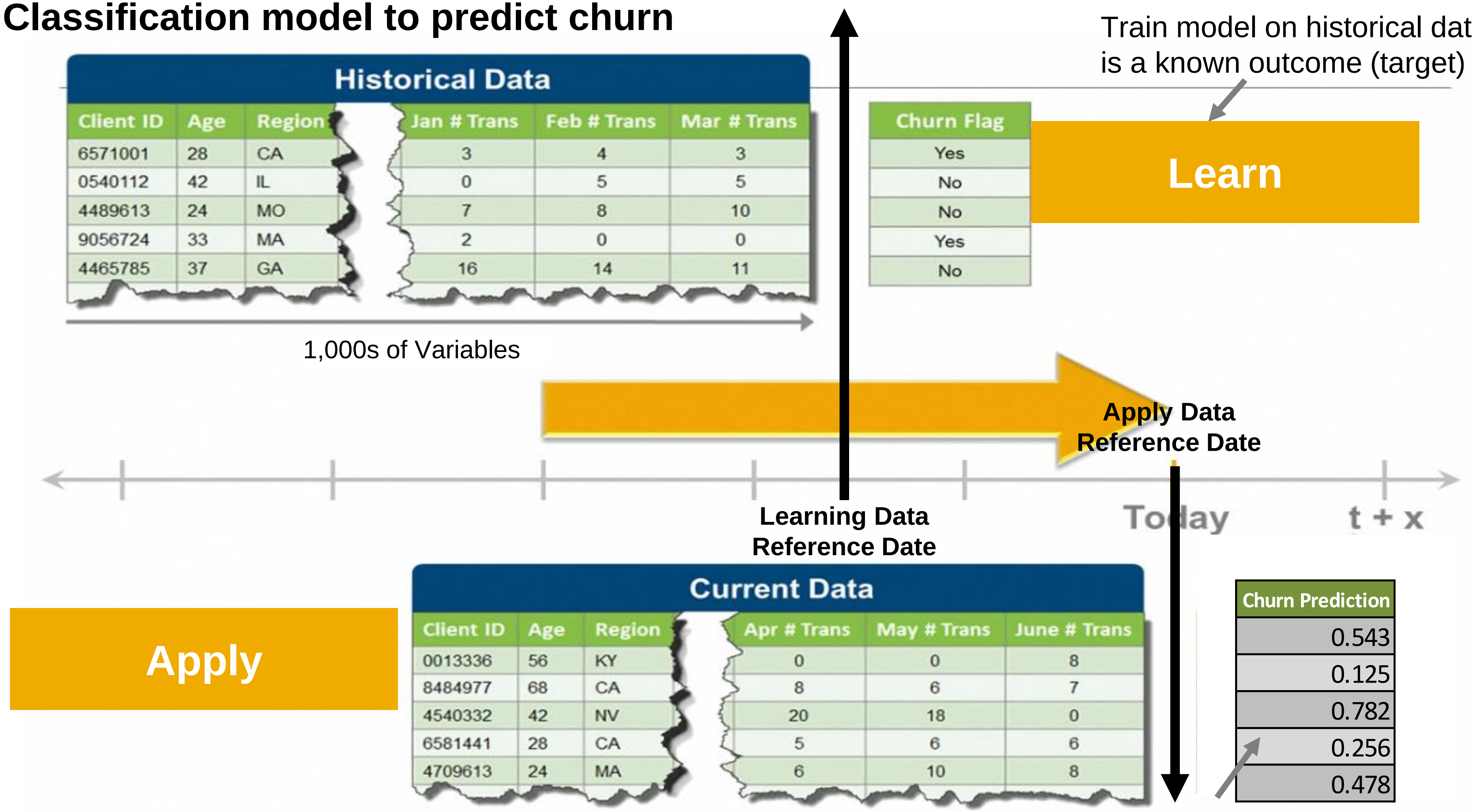
“Apply” the model onto new data to calculate the overall “score” or “probability” for each customer.

Name	City	Age	Score
Marine	Miami	45	**0.8**
Julien	Miami	52	**0.9**
Fred	Orlando	20	**0.6**
Michelle	Boston	34	**0.5**
Nicolas	Phoenix	90	**0.4**

Predictive Modeling Methodology – Overview

Moving data through time

Example: Classification model to predict churn



Train model on historical data where there is a known outcome (target)

Historical Data

Client ID	Age	Region	Jan # Trans	Feb # Trans	Mar # Trans
6571001	28	CA	3	4	3
0540112	42	IL	0	5	5
4489613	24	MO	7	8	10
9056724	33	MA	2	0	0
4465785	37	GA	16	14	11

Churn Flag
Yes
No
No
Yes
No

1,000s of Variables

Current Data

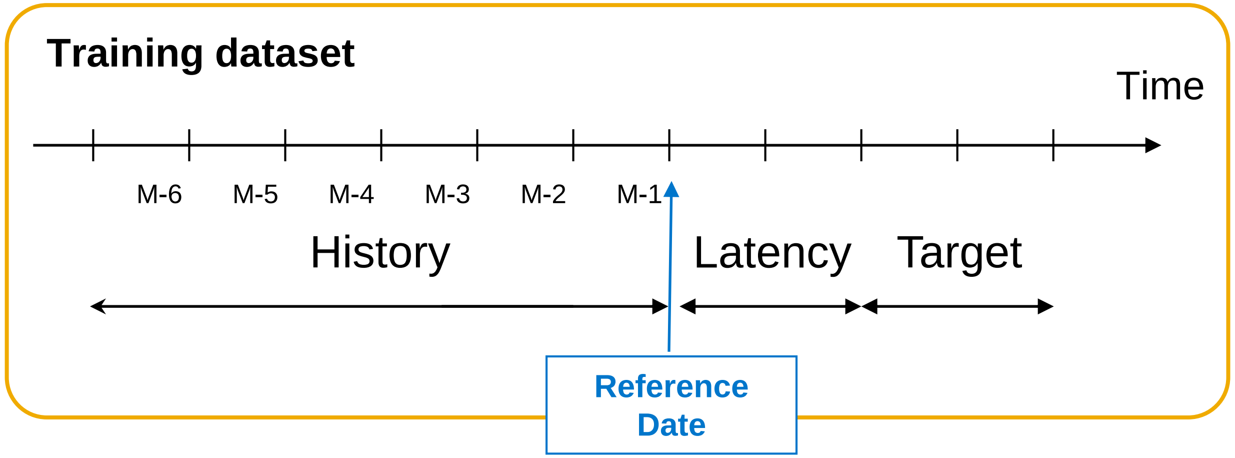
Client ID	Age	Region	Apr # Trans	May # Trans	June # Trans
0013336	56	KY	0	0	8
8484977	68	CA	8	6	7
4540332	42	NV	20	18	0
6581441	28	CA	5	6	6
4709613	24	MA	6	10	8

Churn Prediction
0.543
0.125
0.782
0.256
0.478

Apply the model on current data where we do not know the outcome. The model predicts the outcome probability for each client ID.

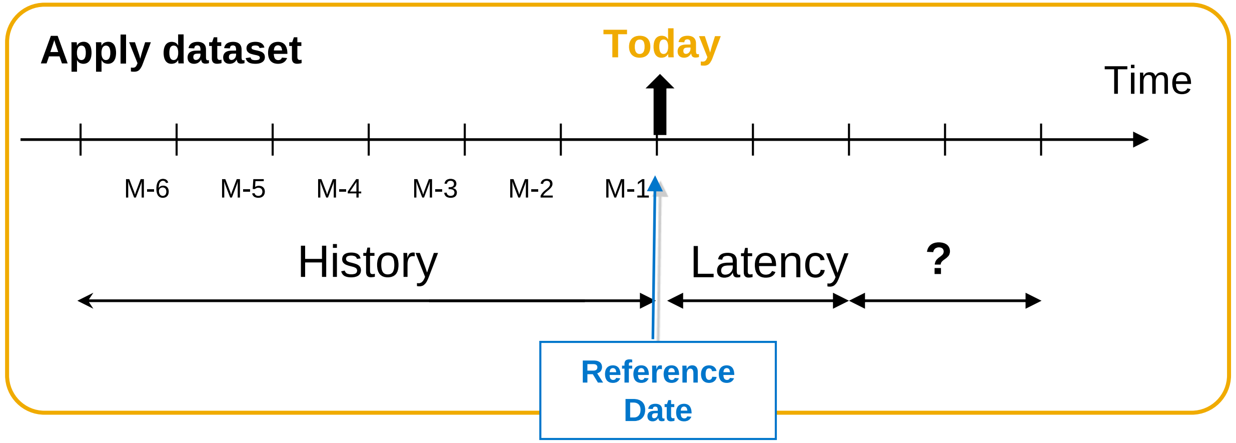
Predictive Modeling Methodology – Overview

Dataset timeframes



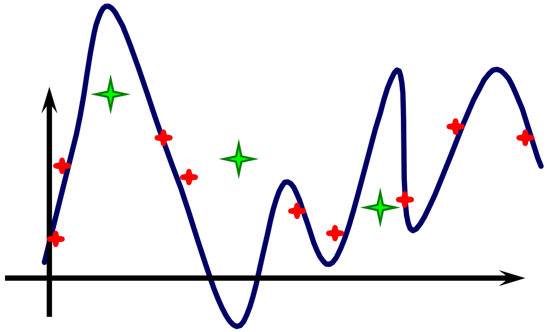
For the training, the target must be known. It has occurred before today (and after the reference date).

For the application, the target is in the future and is therefore unknown.



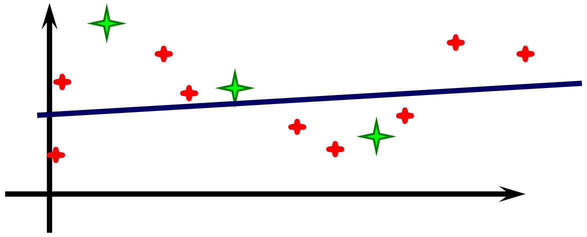
Predictive Modeling Methodology – Overview

Model fitting



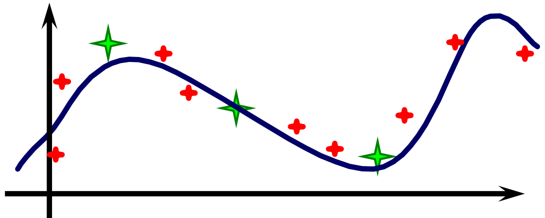
Over-Fit Model/Low Robustness

(No Training Error, High Test Error)



Under-Fit Model/High Robustness

(High Training Error = High Test Error)



Robust Model

(Low Training Error ≈ Low Test Error)

∼ Model Built
✚ Known Data
✚ New Data

Thank you

Contact information:

open@sap.com

© 2016 SAP SE or an SAP affiliate company. All rights reserved.

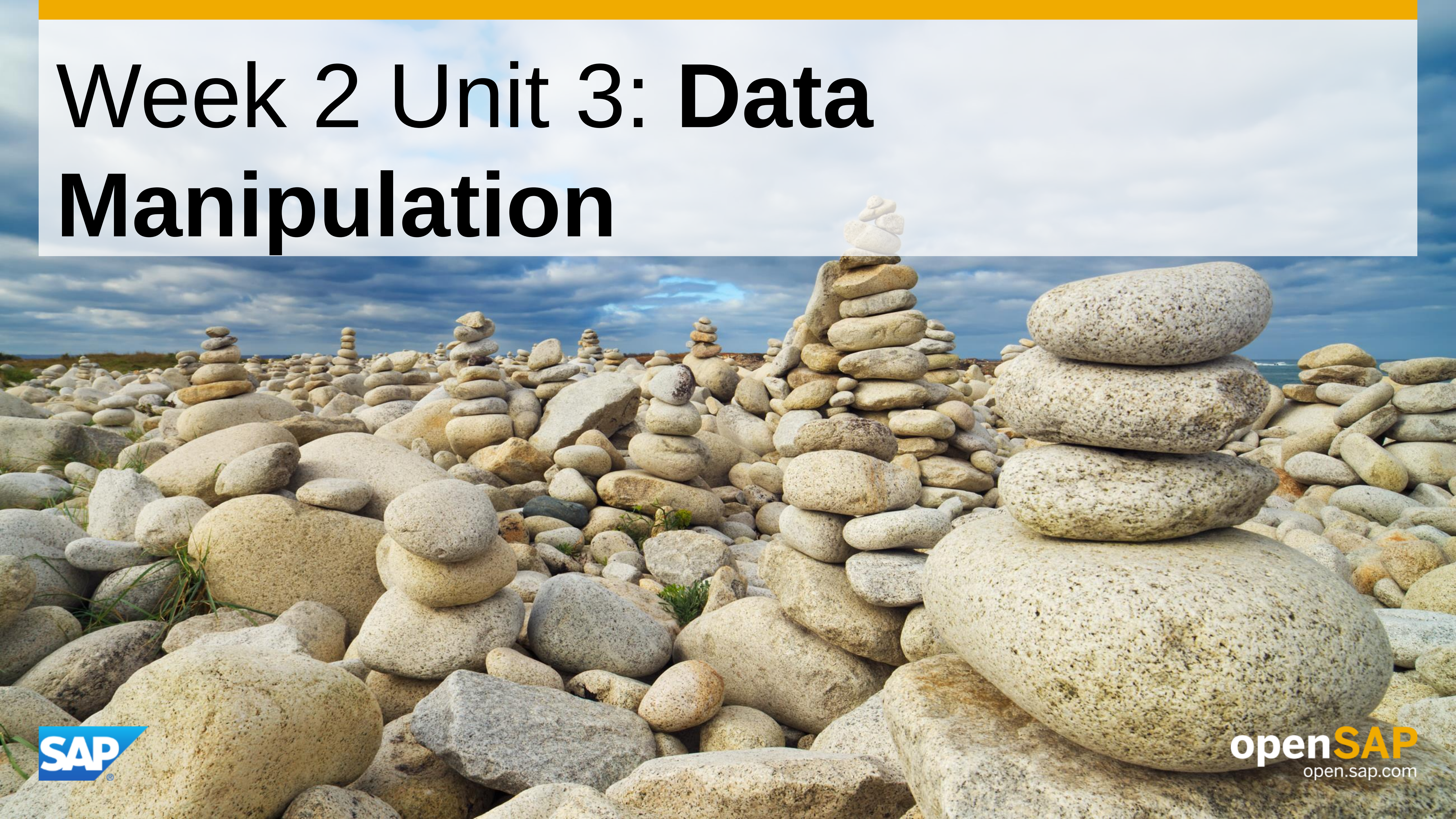
Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors.

National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP SE or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP SE or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.

Week 2 Unit 3: **Data Manipulation**

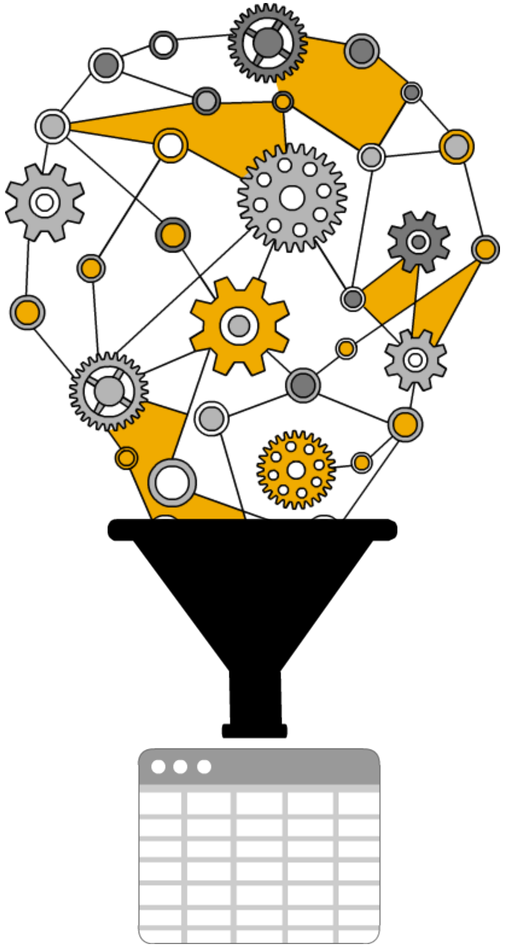


open.sap.com

Data Manipulation

Introduction

- Most data mining activities will require the data to be “prepared” before the analysis is undertaken.
- Data manipulation is often driven by domain knowledge.
- This is a process where database tables are merged and aggregated, new variables and transformations created in order to try and improve model quality, IF/THEN conditions created, and filters applied, etc.



Data Manipulation

Entity

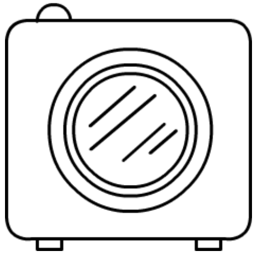
- The first step is to identify the “entity” for the analysis.
  - An entity is the object targeted by the planned analytical task.
  - It may be a customer, a product, or a store, etc., and is usually identified by a unique identifier.
  - The entity defines the granularity of the analysis.

Items of significance to an enterprise are data entities

Sale

Customer

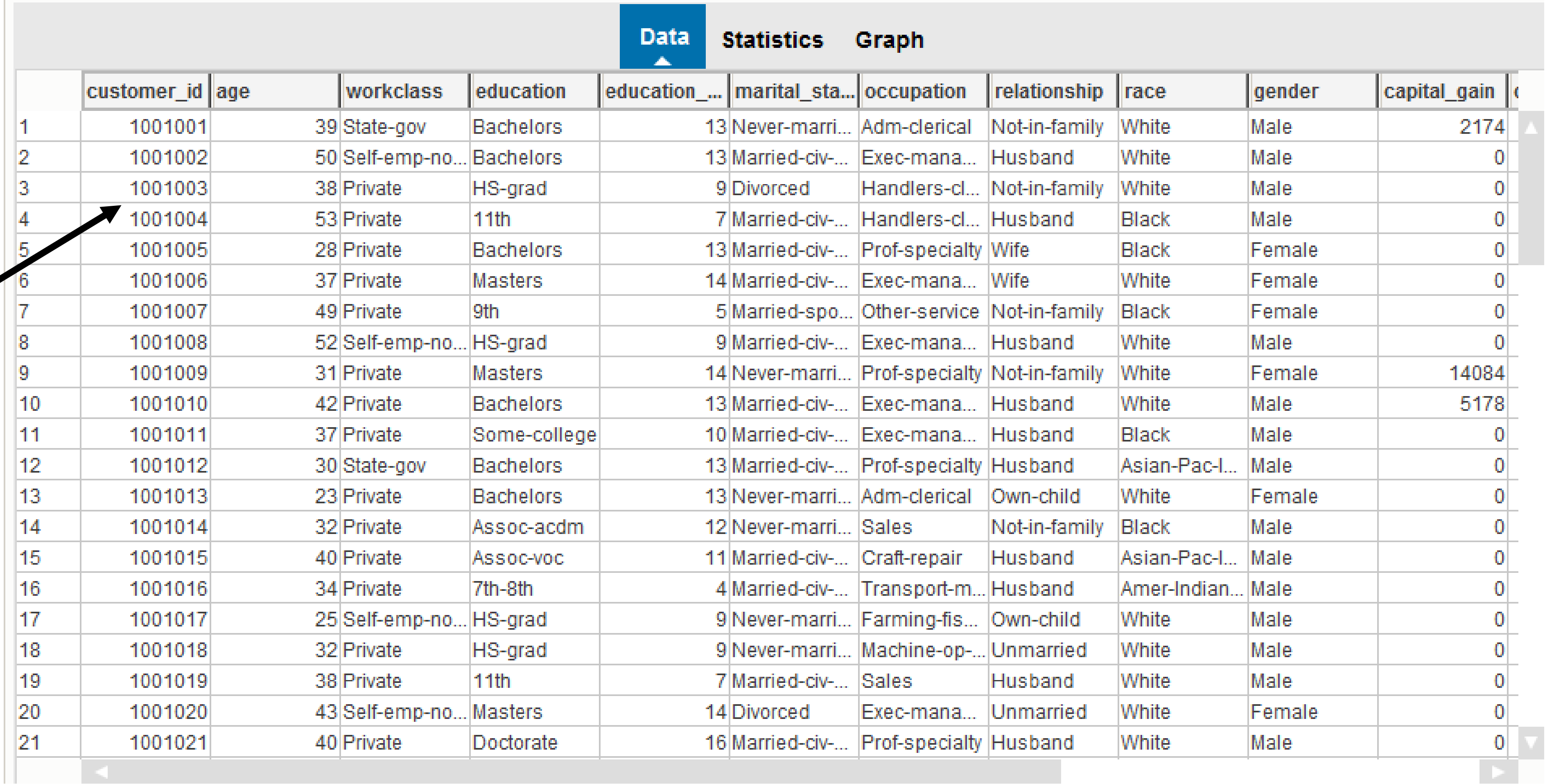
Material



Product

Data Manipulation

Analytical record



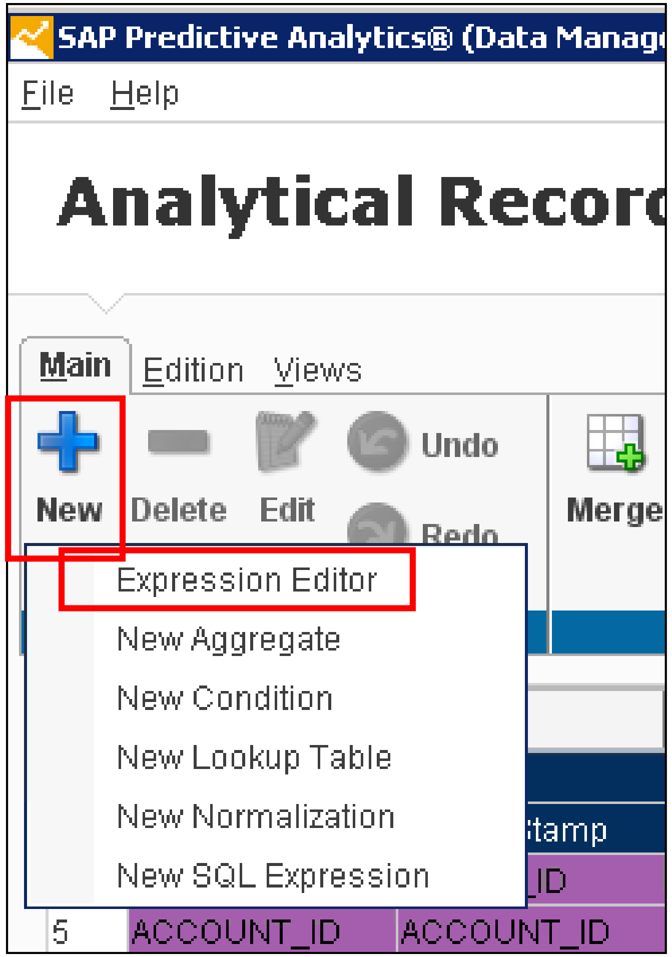
	customer_id	age	workclass	education	education_...	marital_sta...	occupation	relationship	race	gender	capital_gain
1	1001001	39	State-gov	Bachelors	13	Never-marri...	Adm-clerical	Not-in-family	White	Male	2174
2	1001002	50	Self-emp-no...	Bachelors	13	Married-civ-...	Exec-mana...	Husband	White	Male	0
3	1001003	38	Private	HS-grad	9	Divorced	Handlers-cl...	Not-in-family	White	Male	0
4	1001004	53	Private	11th	7	Married-civ-...	Handlers-cl...	Husband	Black	Male	0
5	1001005	28	Private	Bachelors	13	Married-civ-...	Prof-specialty	Wife	Black	Female	0
6	1001006	37	Private	Masters	14	Married-civ-...	Exec-mana...	Wife	White	Female	0
7	1001007	49	Private	9th	5	Married-spo...	Other-service	Not-in-family	Black	Female	0
8	1001008	52	Self-emp-no...	HS-grad	9	Married-civ-...	Exec-mana...	Husband	White	Male	0
9	1001009	31	Private	Masters	14	Never-marri...	Prof-specialty	Not-in-family	White	Female	14084
10	1001010	42	Private	Bachelors	13	Married-civ-...	Exec-mana...	Husband	White	Male	5178
11	1001011	37	Private	Some-college	10	Married-civ-...	Exec-mana...	Husband	Black	Male	0
12	1001012	30	State-gov	Bachelors	13	Married-civ-...	Prof-specialty	Husband	Asian-Pac-I...	Male	0
13	1001013	23	Private	Bachelors	13	Never-marri...	Adm-clerical	Own-child	White	Female	0
14	1001014	32	Private	Assoc-acdm	12	Never-marri...	Sales	Not-in-family	Black	Male	0
15	1001015	40	Private	Assoc-voc	11	Married-civ-...	Craft-repair	Husband	Asian-Pac-I...	Male	0
16	1001016	34	Private	7th-8th	4	Married-civ-...	Transport-m...	Husband	Amer-Indian...	Male	0
17	1001017	25	Self-emp-no...	HS-grad	9	Never-marri...	Farming-fis...	Own-child	White	Male	0
18	1001018	32	Private	HS-grad	9	Never-marri...	Machine-op-...	Unmarried	White	Male	0
19	1001019	38	Private	11th	7	Married-civ-...	Sales	Husband	White	Male	0
20	1001020	43	Self-emp-no...	Masters	14	Divorced	Exec-mana...	Unmarried	White	Female	0
21	1001021	40	Private	Doctorate	16	Married-civ-...	Prof-specialty	Husband	White	Male	0

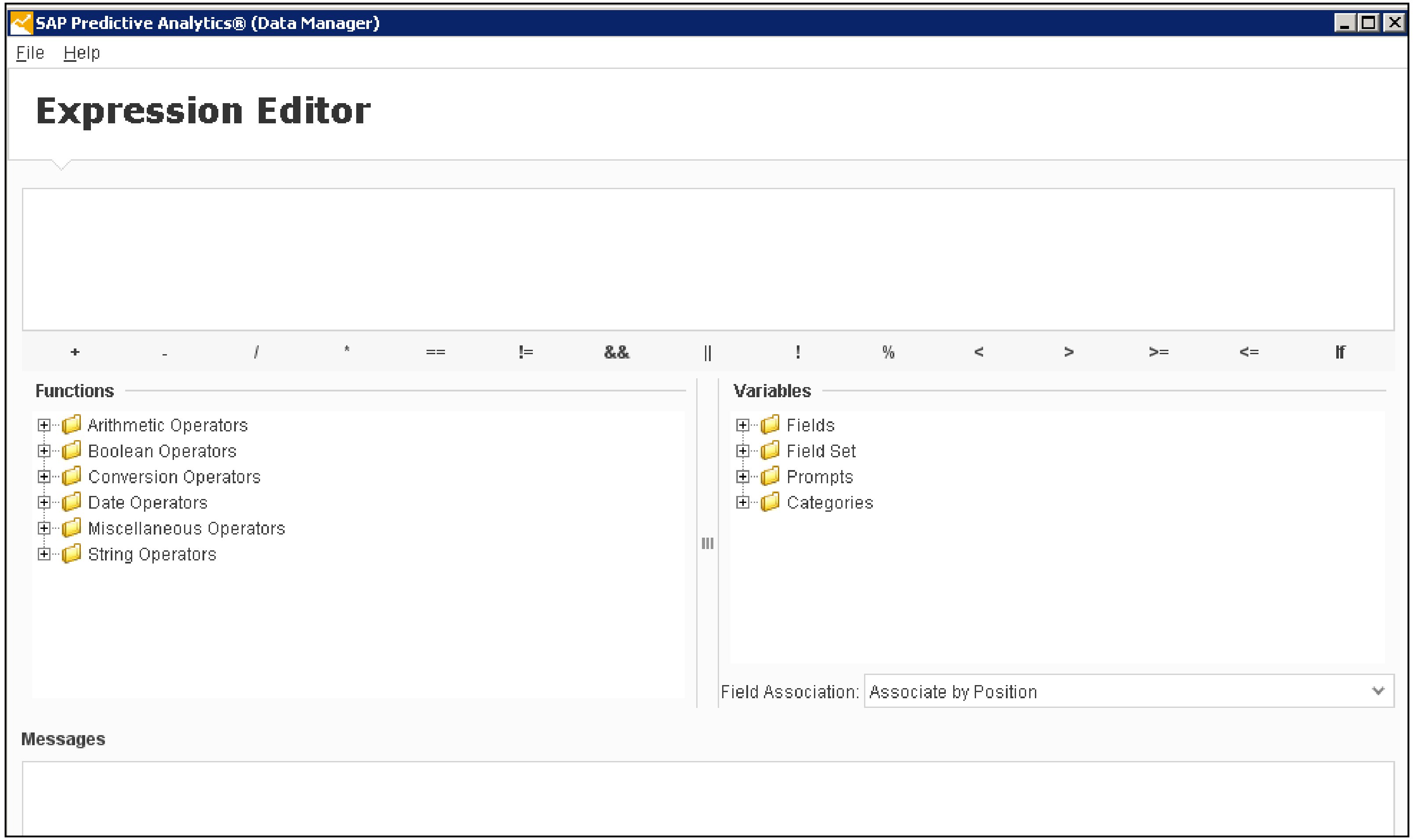
Entity is customer_id

The analytical record is a 360º view of each entity, collecting all of the static and dynamic data together that can be used to define the entity.

Data Manipulation

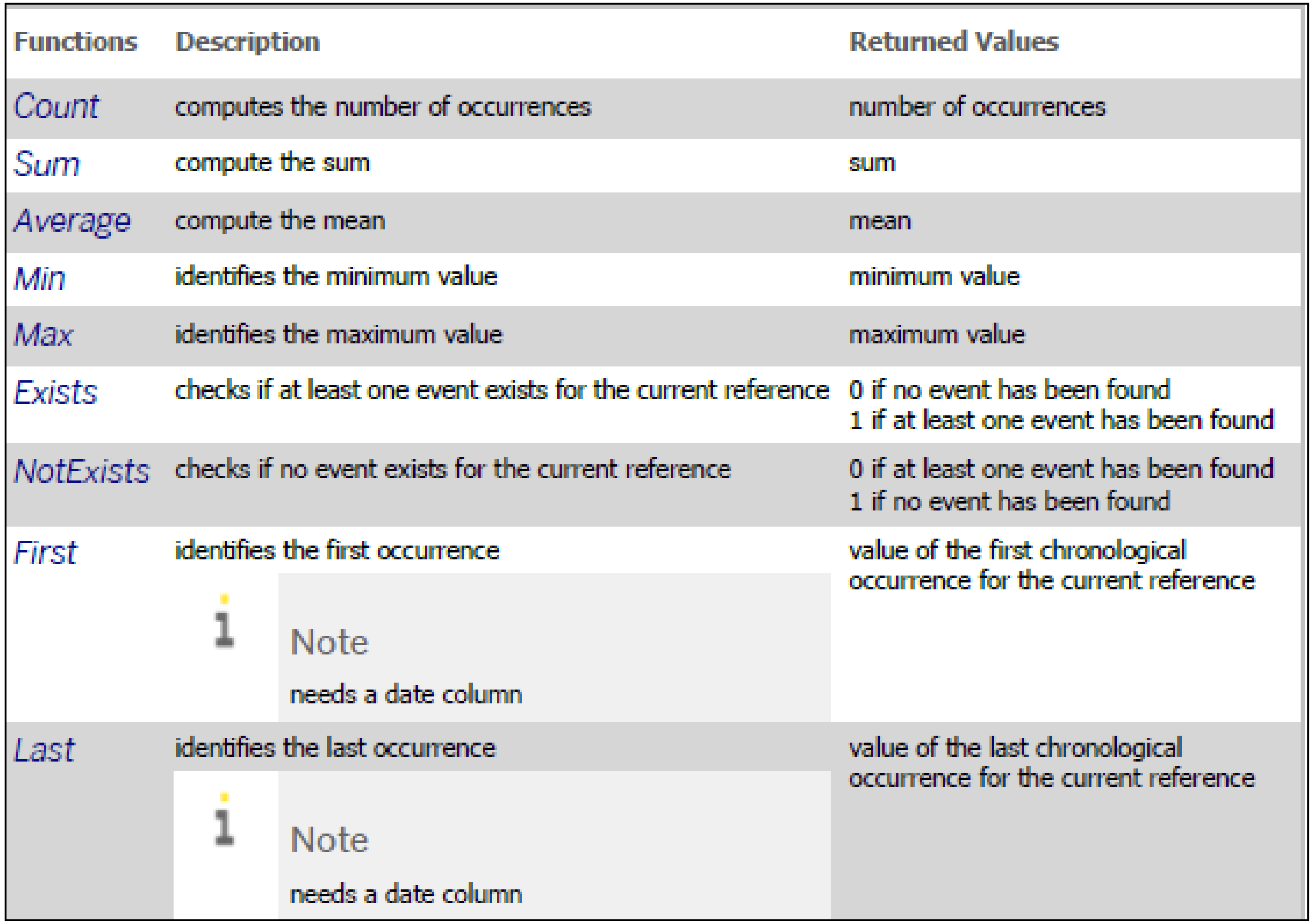
Creating new data transformations in SAP Predictive Analytics Data Manager





Data Manipulation

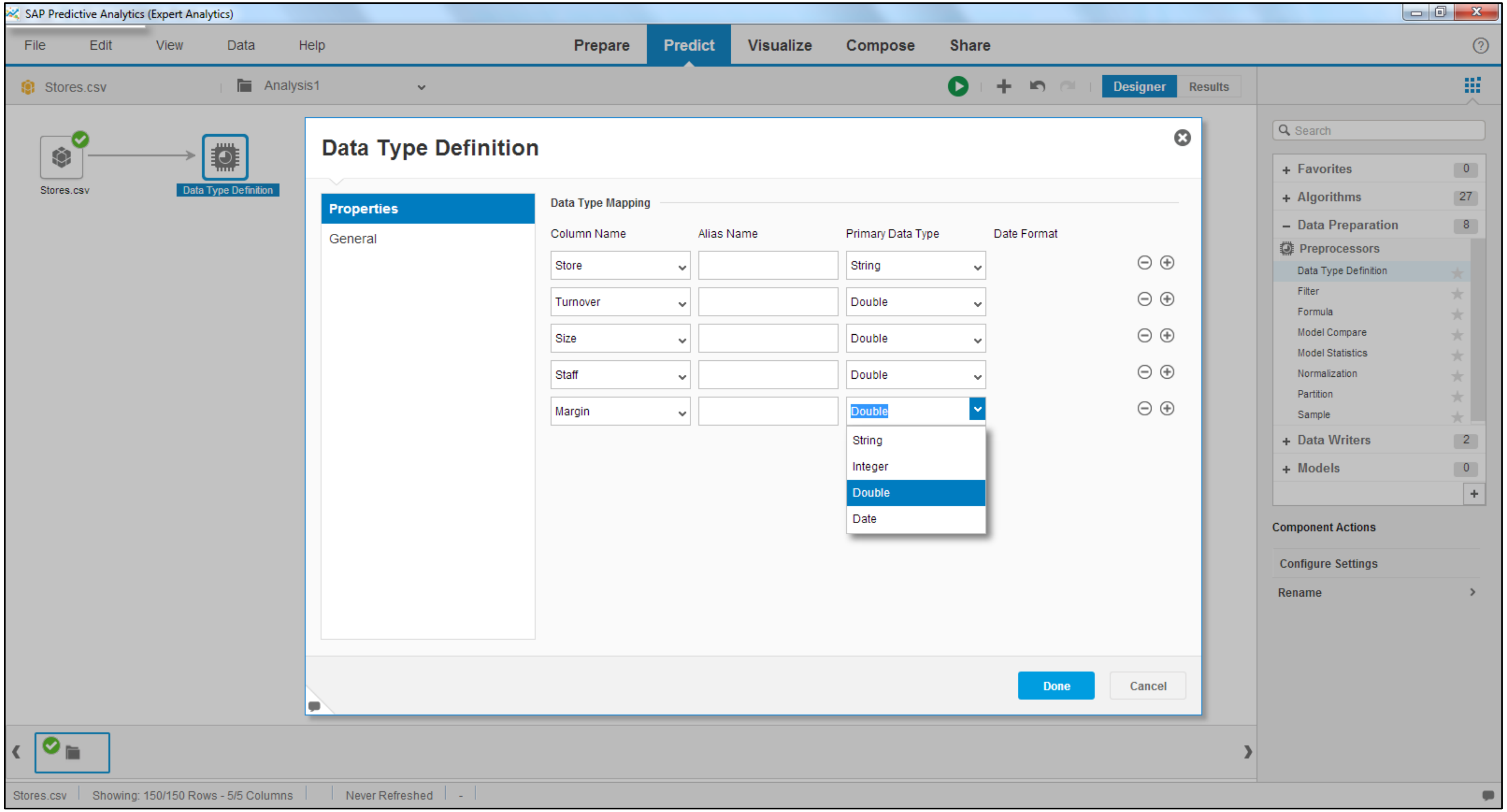
Creating data aggregations



Functions	Description	Returned Values
Count	computes the number of occurrences	number of occurrences
Sum	compute the sum	sum
Average	compute the mean	mean
Min	identifies the minimum value	minimum value
Max	identifies the maximum value	maximum value
Exists	checks if at least one event exists for the current reference	0 if no event has been found 1 if at least one event has been found
NotExists	checks if no event exists for the current reference	0 if at least one event has been found 1 if no event has been found
First	identifies the first occurrence Note needs a date column	value of the first chronological occurrence for the current reference
Last	identifies the last occurrence Note needs a date column	value of the last chronological occurrence for the current reference

Data Manipulation

Converting data types



Data Type Definition

Properties

General

Data Type Mapping

Column Name	Alias Name	Primary Data Type	Date Format
Store		String	
Turnover		Double	
Size		Double	
Staff		Double	
Margin		Double	

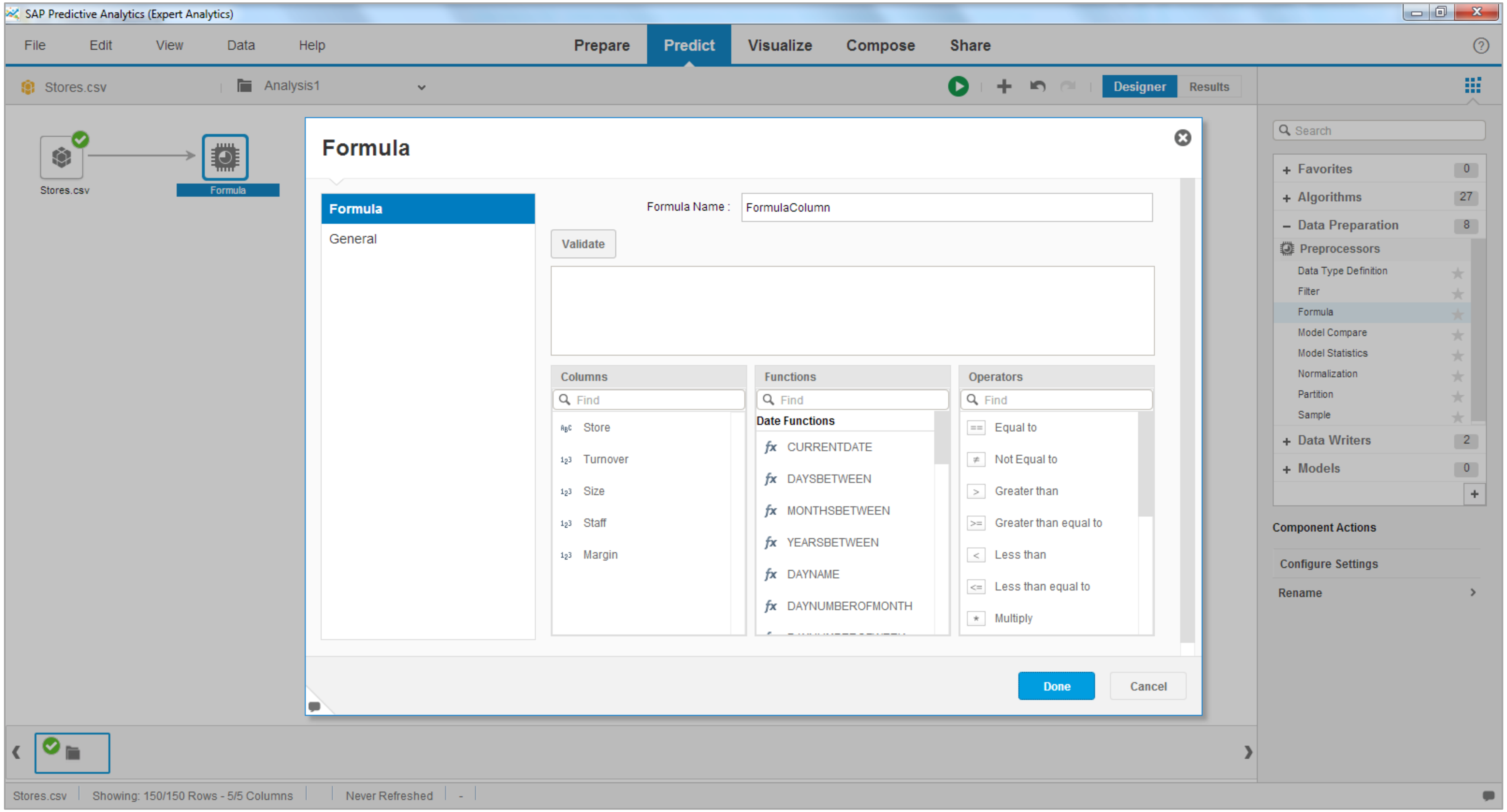
String
Integer
Double
Date

Done

Cancel

Data Manipulation

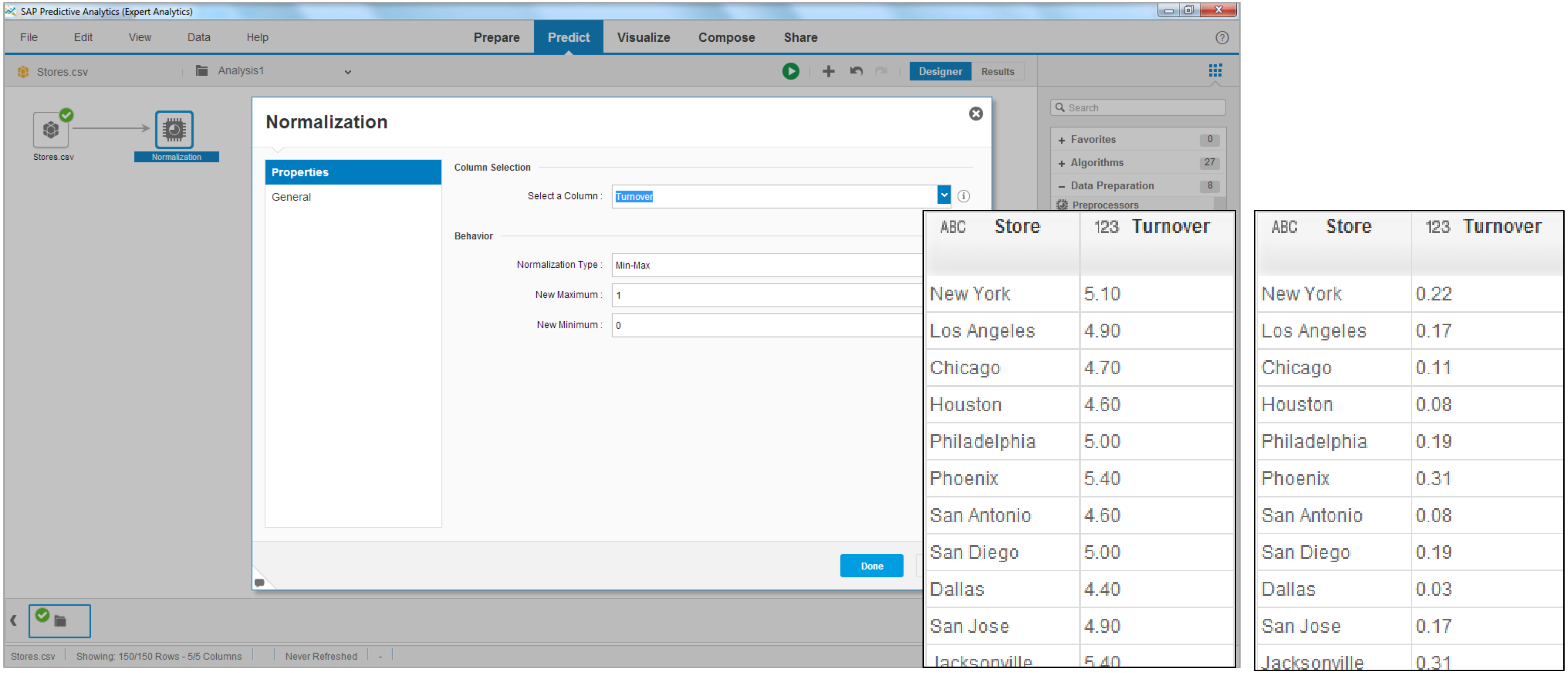
Forming new variables



Formula Name : FormulaColumn

Data Manipulation

Transforming variables



Normalization

Column Selection

Select a Column : Turnover

Behavior

Normalization Type : Min-Max

New Maximum : 1

New Minimum : 0

ABC Store	123 Turnover
New York	5.10
Los Angeles	4.90
Chicago	4.70
Houston	4.60
Philadelphia	5.00
Phoenix	5.40
San Antonio	4.60
San Diego	5.00
Dallas	4.40
San Jose	4.90
Jacksonville	5.40

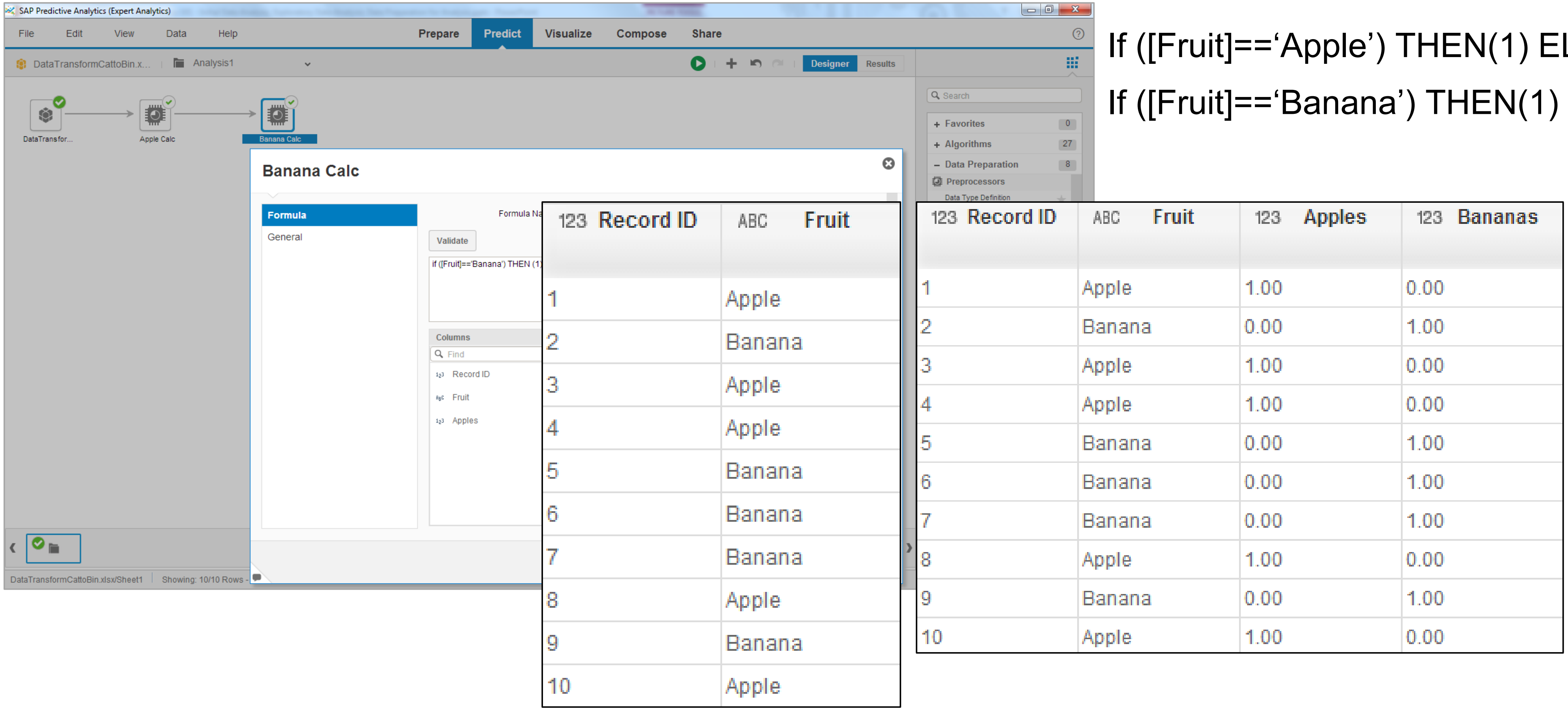
ABC Store	123 Turnover
New York	0.22
Los Angeles	0.17
Chicago	0.11
Houston	0.08
Philadelphia	0.19
Phoenix	0.31
San Antonio	0.08
San Diego	0.19
Dallas	0.03
San Jose	0.17
Jacksonville	0.31

Data Manipulation

Transforming variables

If ([Fruit]==‘Apple’) THEN(1) ELSE (0)

If ([Fruit]==‘Banana’) THEN(1) ELSE (0)



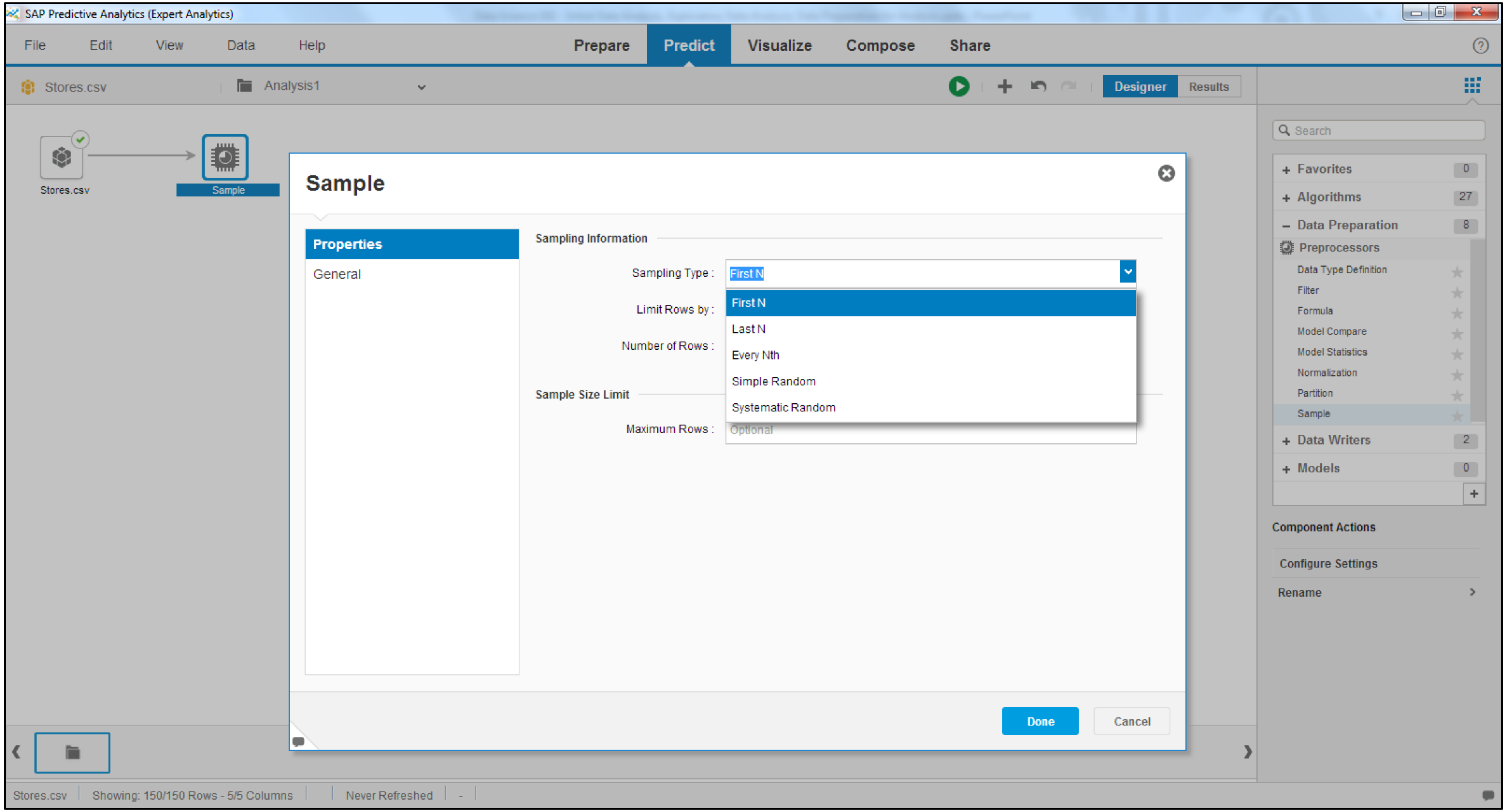
if ([Fruit]=='Banana') THEN (1

123 Record ID	ABC Fruit
1	Apple
2	Banana
3	Apple
4	Apple
5	Banana
6	Banana
7	Banana
8	Apple
9	Banana
10	Apple

123 Record ID	ABC Fruit	123 Apples	123 Bananas
1	Apple	1.00	0.00
2	Banana	0.00	1.00
3	Apple	1.00	0.00
4	Apple	1.00	0.00
5	Banana	0.00	1.00
6	Banana	0.00	1.00
7	Banana	0.00	1.00
8	Apple	1.00	0.00
9	Banana	0.00	1.00
10	Apple	1.00	0.00

Data Manipulation

Sampling data



Sample

Sampling Information

Sampling Type : First N

Limit Rows by :

Number of Rows :

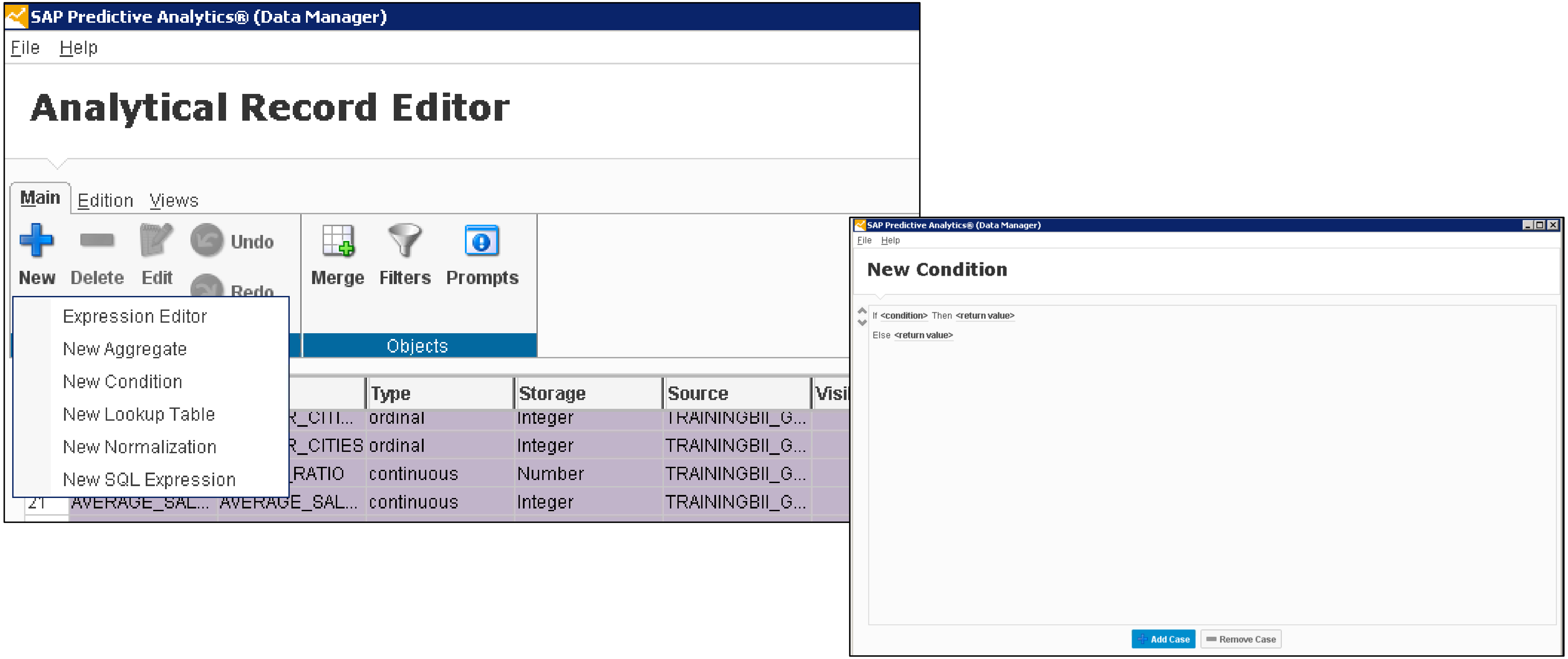
Sample Size Limit

Maximum Rows : Optional

Stores.csv | Showing: 150/150 Rows - 5/5 Columns | Never Refreshed | -

Data Manipulation

Conditions



		Type	Storage	Source	Visi
	R_CITI...	ordinal	Integer	TRAININGBII_G...	
	R_CITIES	ordinal	Integer	TRAININGBII_G...	
	_RATIO	continuous	Number	TRAININGBII_G...	
21 AVERAGE_SAL...	AVERAGE_SAL...	continuous	Integer	TRAININGBII_G...	

If <condition> Then <return value>

Else <return value>

Data Manipulation

Documentation

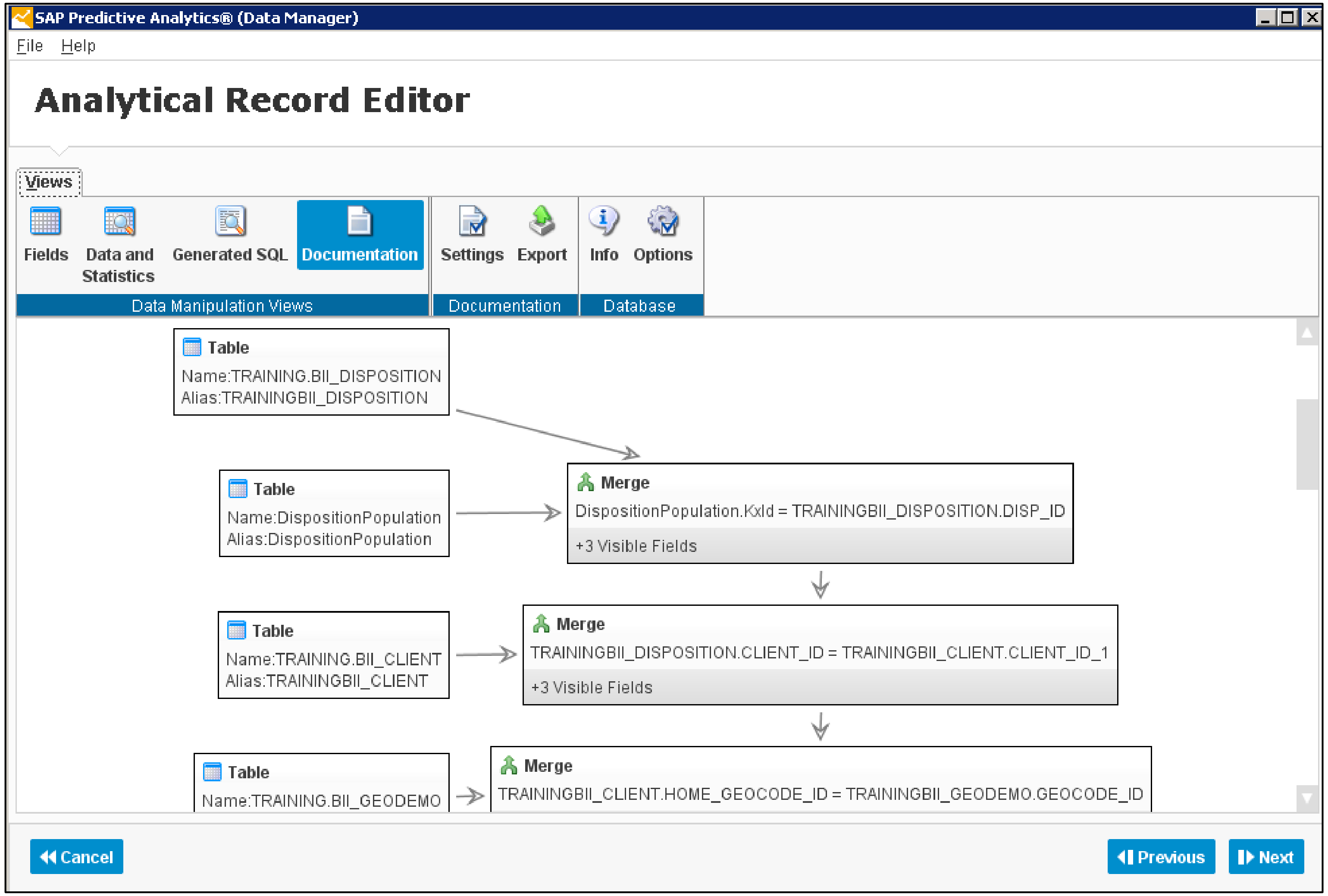


Table
Name:TRAINING.BII_DISPOSITION
Alias:TRAININGBII_DISPOSITION

Table
Name:DispositionPopulation
Alias:DispositionPopulation

Merge
DispositionPopulation.KxId = TRAININGBII_DISPOSITION.DISP_ID
+3 Visible Fields

Table
Name:TRAINING.BII_CLIENT
Alias:TRAININGBII_CLIENT

Merge
TRAININGBII_DISPOSITION.CLIENT_ID = TRAININGBII_CLIENT.CLIENT_ID_1
+3 Visible Fields

Table
Name:TRAINING.BII_GEODEMO

Merge
TRAININGBII_CLIENT.HOME_GEOCODE_ID = TRAININGBII_GEODEMO.GEOCODE_ID

Thank you

Contact information:

open@sap.com

© 2016 SAP SE or an SAP affiliate company. All rights reserved.

Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors.

National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP SE or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP SE or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.

Week 2 Unit 4: **Selecting Data – Variable and Feature Selection**

open.sap.com

Selecting Data – Variable and Feature Selection

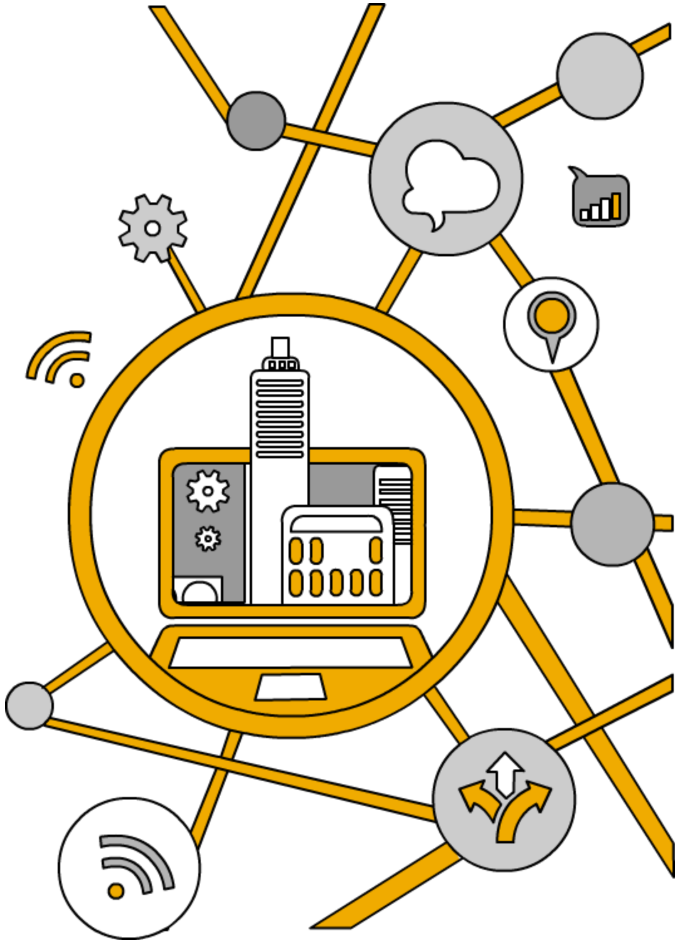
Introduction

- Feature or variable selection is the process of selecting a subset of relevant explanatory variables or predictors for use in data science model construction.
- It is also known as variable selection, attribute selection, or variable subset selection.
- Often, data contains many features that are either *redundant* or *irrelevant*, and can be removed without incurring much loss of information.
- Remember that domain knowledge can be the best selection criterion of all!!

Selecting Data – Variable and Feature Selection

Traditional approaches to variable selection

- Traditional approaches to selecting the variables to go into a model can be very time consuming, especially when there are 1000s of variables to analyze.
- The most popular form of feature selection is **stepwise regression**. This is an algorithm that adds the best feature (or deletes the worst feature) in a series of iterative steps. The main control issue is deciding when to stop the algorithm.
- Other automated selection processes are **backward elimination** and **forward selection**.



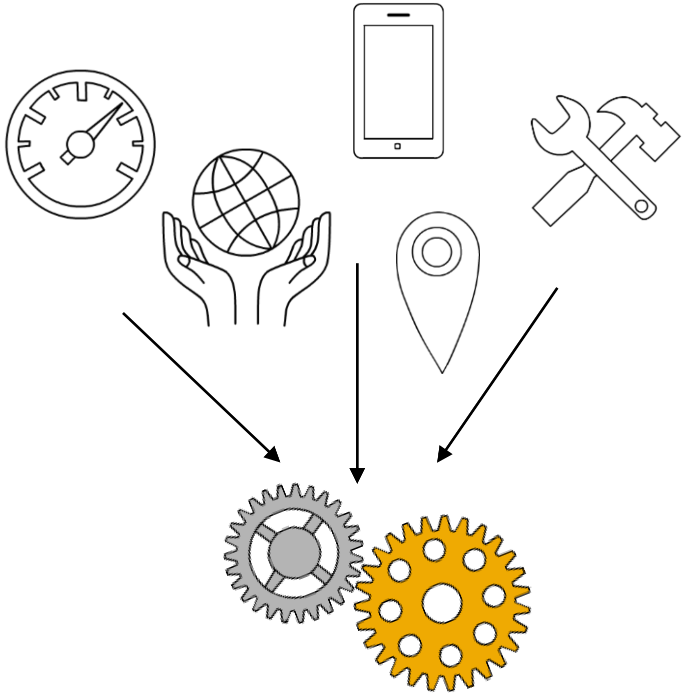
Selecting Data – Variable and Feature Selection

Backward elimination

1. Backward elimination starts with all candidate features.

2. Test the deletion of each feature using the chosen model comparison criterion, deleting the feature (if any) that improves the model the most by being deleted.

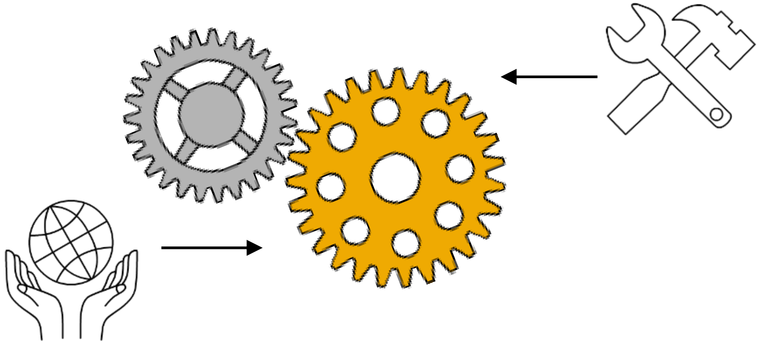
3. Repeat this process until no further improvement is possible.



Selecting Data – Variable and Feature Selection

Forward selection

1. Forward selection starts with no features in the model.
2. Test the addition of each feature using the chosen model comparison criterion.
3. Add the feature (if any) that improves the model the most.
4. Repeat this process until no other feature additions improve the model.



Selecting Data – Variable and Feature Selection

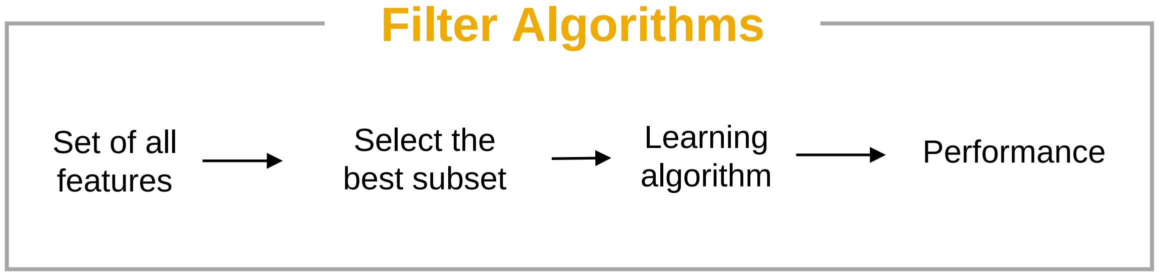
Stepwise regression

- This is a combination of backward elimination and forward selection.
- At each stage in the process, after a new variable is added, a test is made to check if some variables can be deleted without appreciably increasing the error.
- The procedure terminates when the measure is (locally) maximized, or when the available improvement falls below some critical value.
- One of the main issues with stepwise regression is that it is prone to overfitting the data. However, this problem can be mitigated if the criterion for adding (or deleting) a variable is stiff enough.

Selecting Data – Variable and Feature Selection

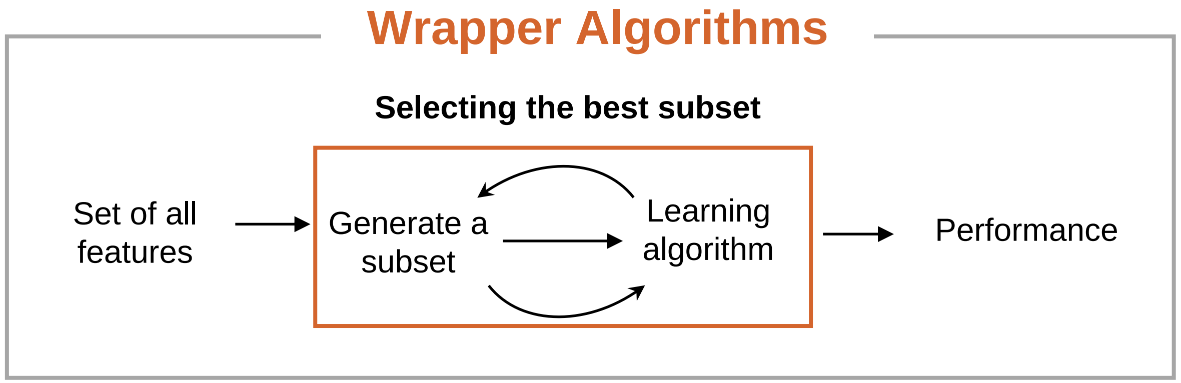
Modern approaches to variable selection

Filter Algorithms



Set of all features → Select the best subset → Learning algorithm → Performance

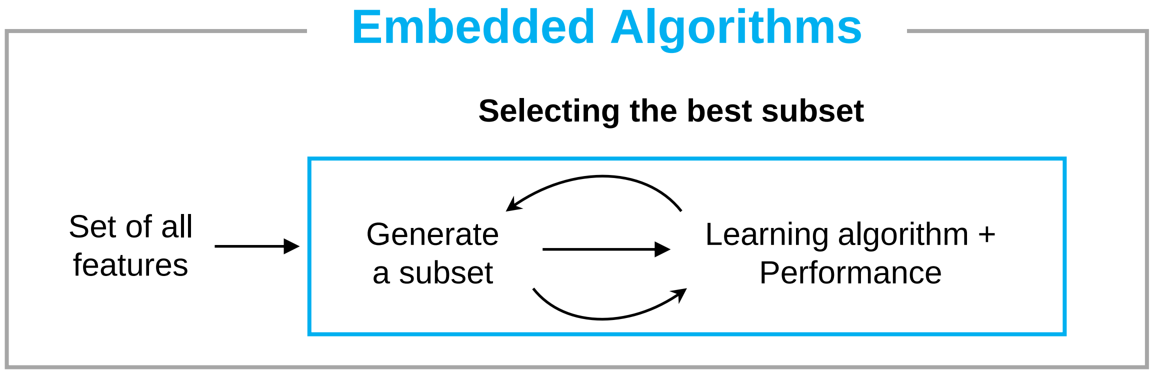
Wrapper Algorithms



Selecting the best subset

Set of all features → Generate a subset ⇄ Learning algorithm → Performance

Embedded Algorithms



Selecting the best subset

Set of all features → Generate a subset ⇄ Learning algorithm + Performance

Thank you

Contact information:

open@sap.com

open.sap.com

© 2016 SAP SE or an SAP affiliate company. All rights reserved.

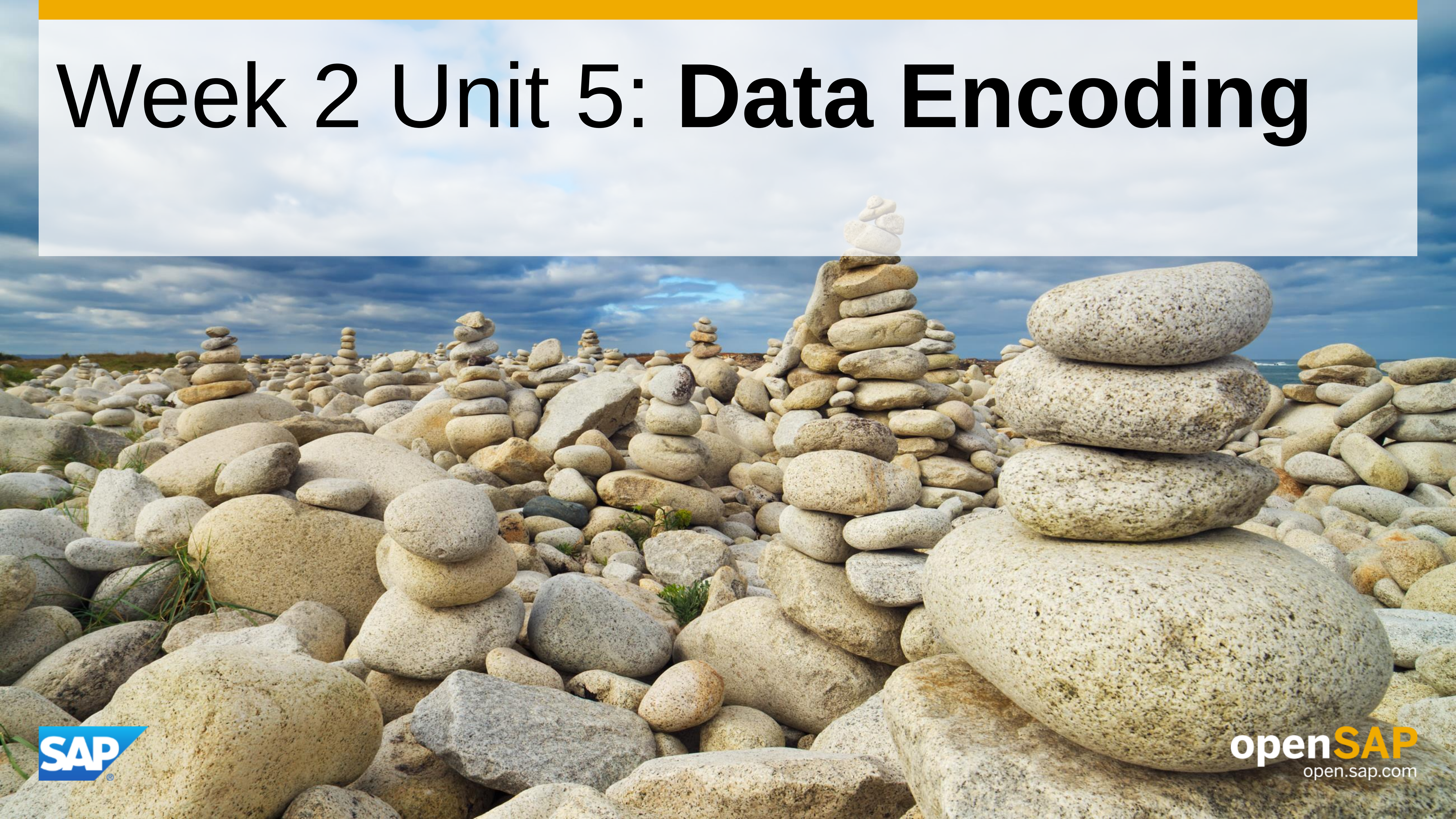
Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors.

National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP SE or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP SE or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.

Week 2 Unit 5: **Data Encoding**



Data Encoding

Introduction

- Data encoding is an essential part of the data preparation process.

- The data encoding process prepares missing values in the data, deals with outliers, and creates data bins or bands to transform raw data into a “mineable” source of information.

Data Encoding

Nominal variable

- A nominal variable is a discrete (categorical), qualitative variable that characterizes, describes, or names an element of a population.

Examples:

- Hair color (brown, blond, ginger…)
- Make of car (Mercedes, Ford….)
- Gender (male, female)
- Postal (ZIP) code
- Residence city (London, New York, Paris…)

Note: The order of the categories does not matter

Data Encoding

Ordinal variable

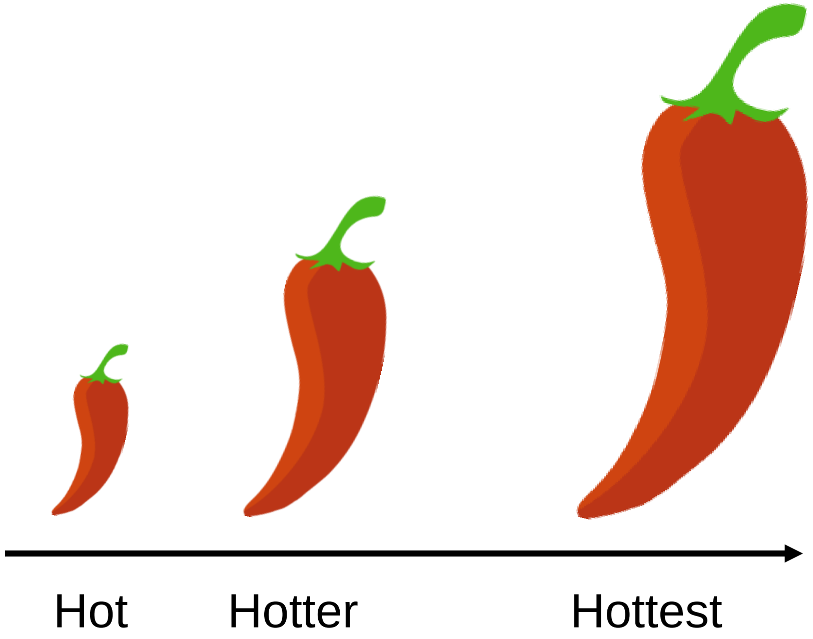
- An ordinal variable is a discrete (categorical), qualitative variable that has order.

Examples:

- Gold, silver, bronze
- Satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)
- Pain level (mild, moderate, severe)

Note: The order of the categories does matter

The “Hot” Scale



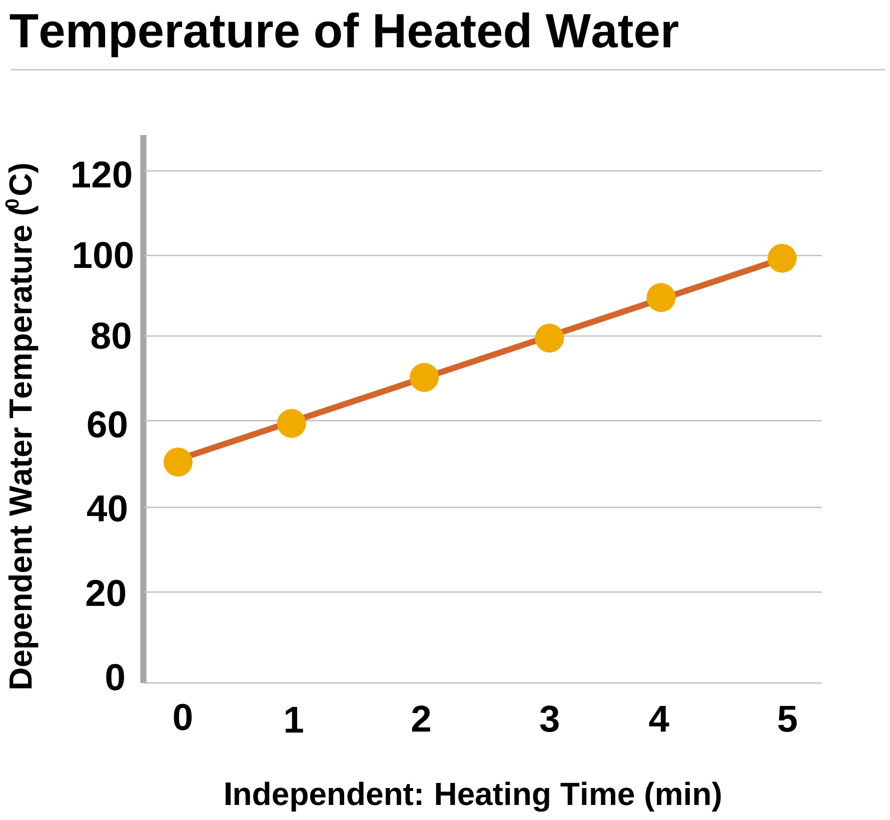
Data Encoding

Continuous variable

- A continuous variable is a quantitative variable.
- It is a real number that can take any value (with fractions/ decimal places) between two specific numbers.
- It accommodates all basic arithmetic operations (addition, subtraction, multiplication, and division).

Examples:

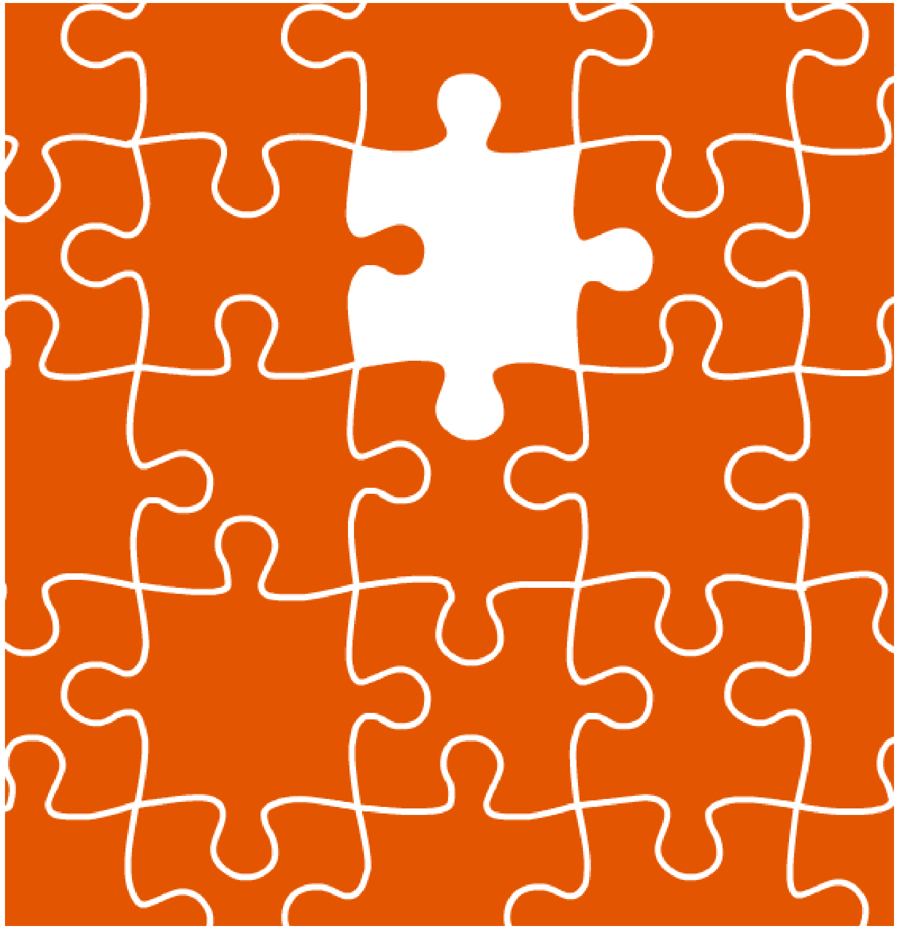
- Income
- Age (years)
- Running time (minutes)
- Bank account balance ($)
- Distance (miles)
- Any ratio or calculated value
- This includes most business data



Data Encoding

Missing values

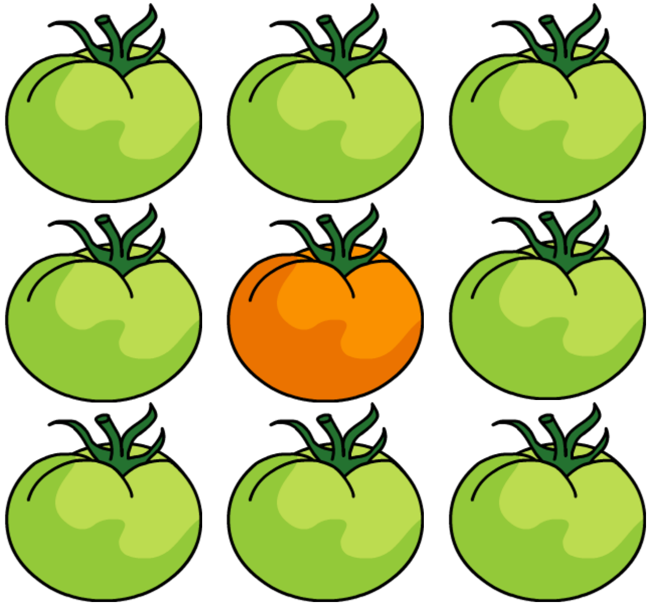
- A missing value is an empty cell in your dataset.
- Missing values in a dataset can be due to error or because they are simply not available.
- They can be removed from the dataset, estimated, or kept.
- The analysis could also be stopped so that further investigation of the reason for missing values can be undertaken.



Data Encoding

Outliers

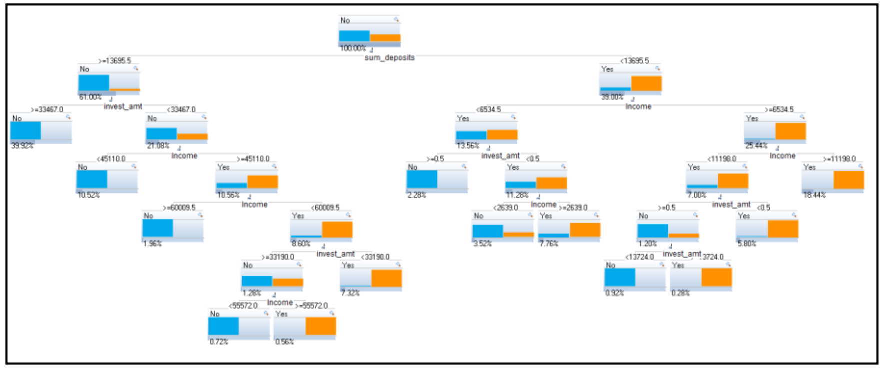
- For a continuous variable – An outlier is a single or low-frequency occurrence of the value of a variable that is far from the mean as well as the majority of other values for that variable.
- For a categorical variable (nominal or ordinal) – An outlier is a single or very low-frequency occurrence of a category of a variable.

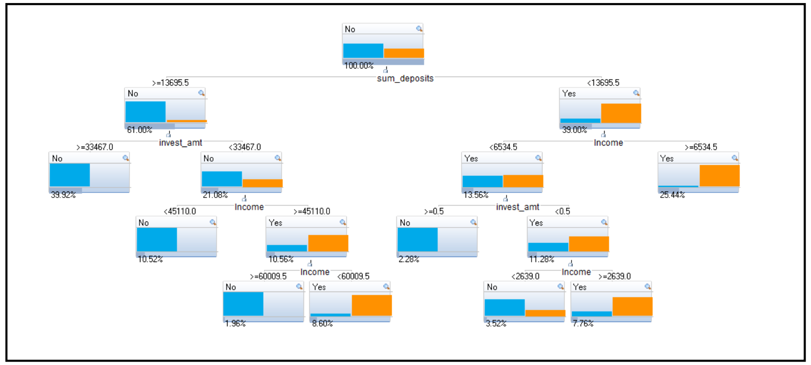


Data Encoding

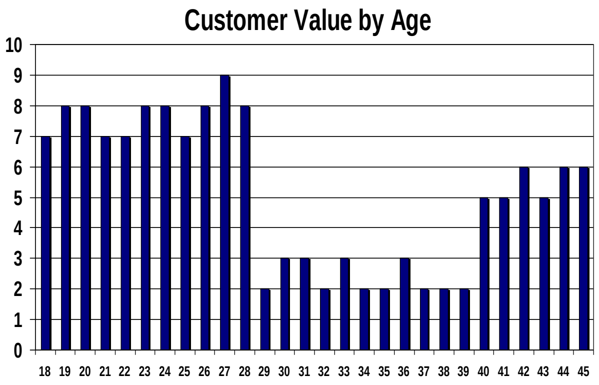
Binning

1. A decision tree before and after binning

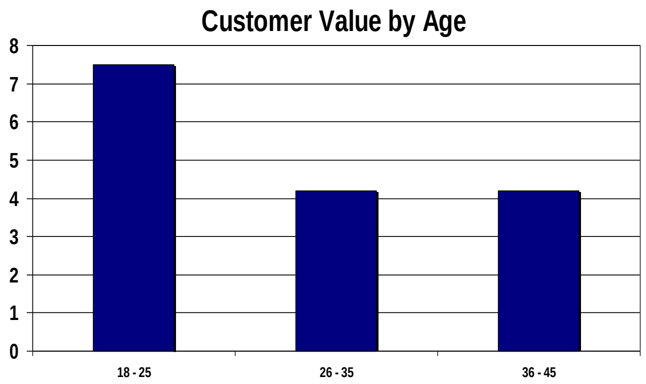




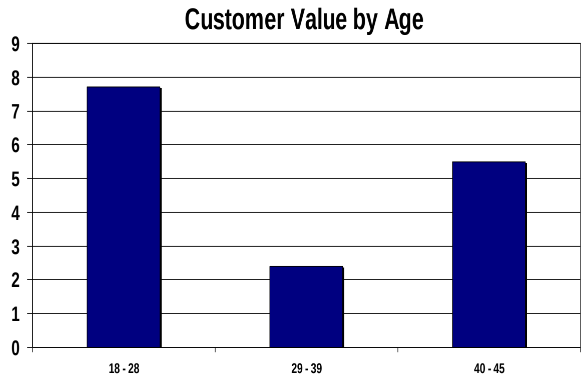
2. Continuous variable binning – variable “AGE”, no binning, manual binning, SAP Predictive Analytics (Automated) binning



Raw data – No binning



Manual binning



Using SAP Predictive Analytics (Automated) to automatically find the best discrimination

Thank you

Contact information:

open@sap.com

openSAP
open.sap.com

© 2016 SAP SE or an SAP affiliate company. All rights reserved.

Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors.

National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP SE or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP SE or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.