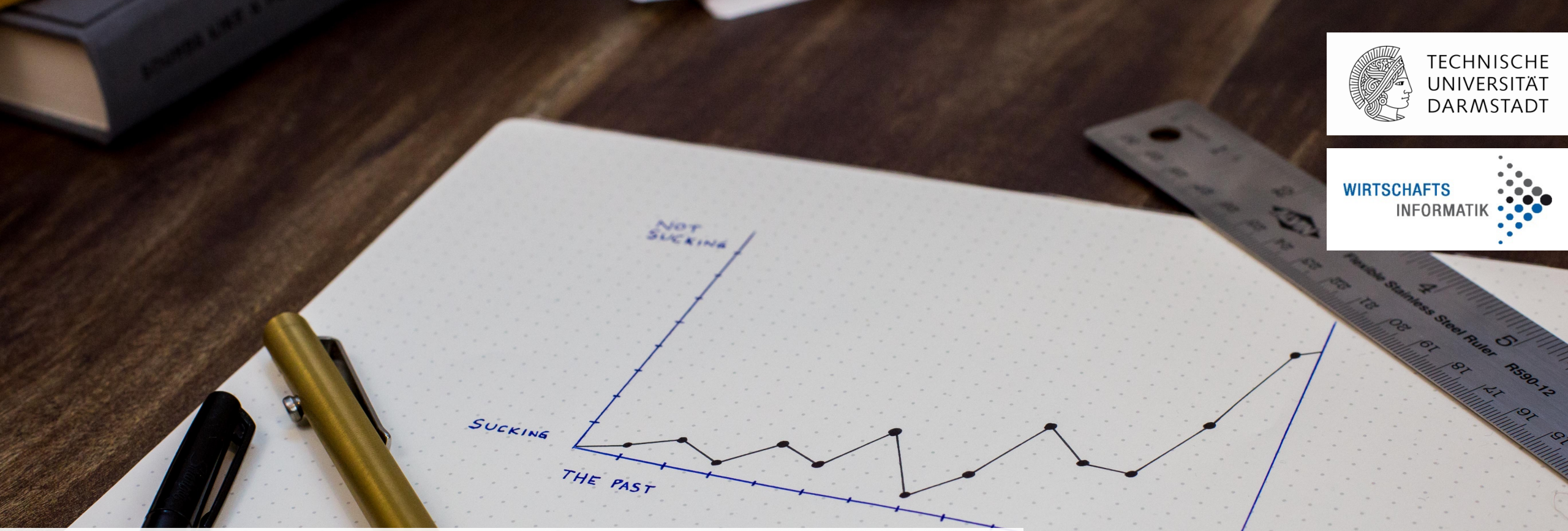


Artificial Intelligence | Basics of Algorithms & Application

Exercise 1: "Introduction to RapidMiner"

Timo Sturm & Dr. Dominik Jung
ki@is.tu-darmstadt.de
Prof. Dr. Peter Buxmann | Information Systems | Software & Digital Business
School of Business, Economics & Law
TU Darmstadt

Outline

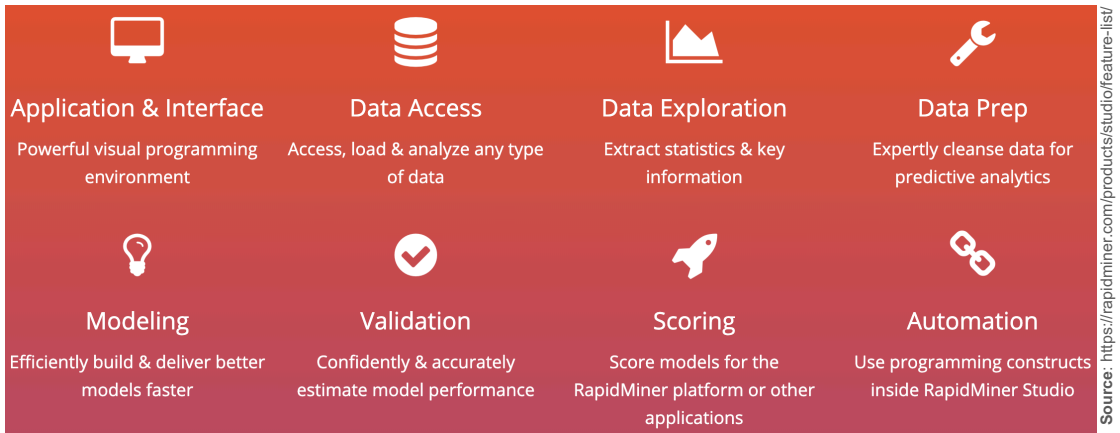
WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

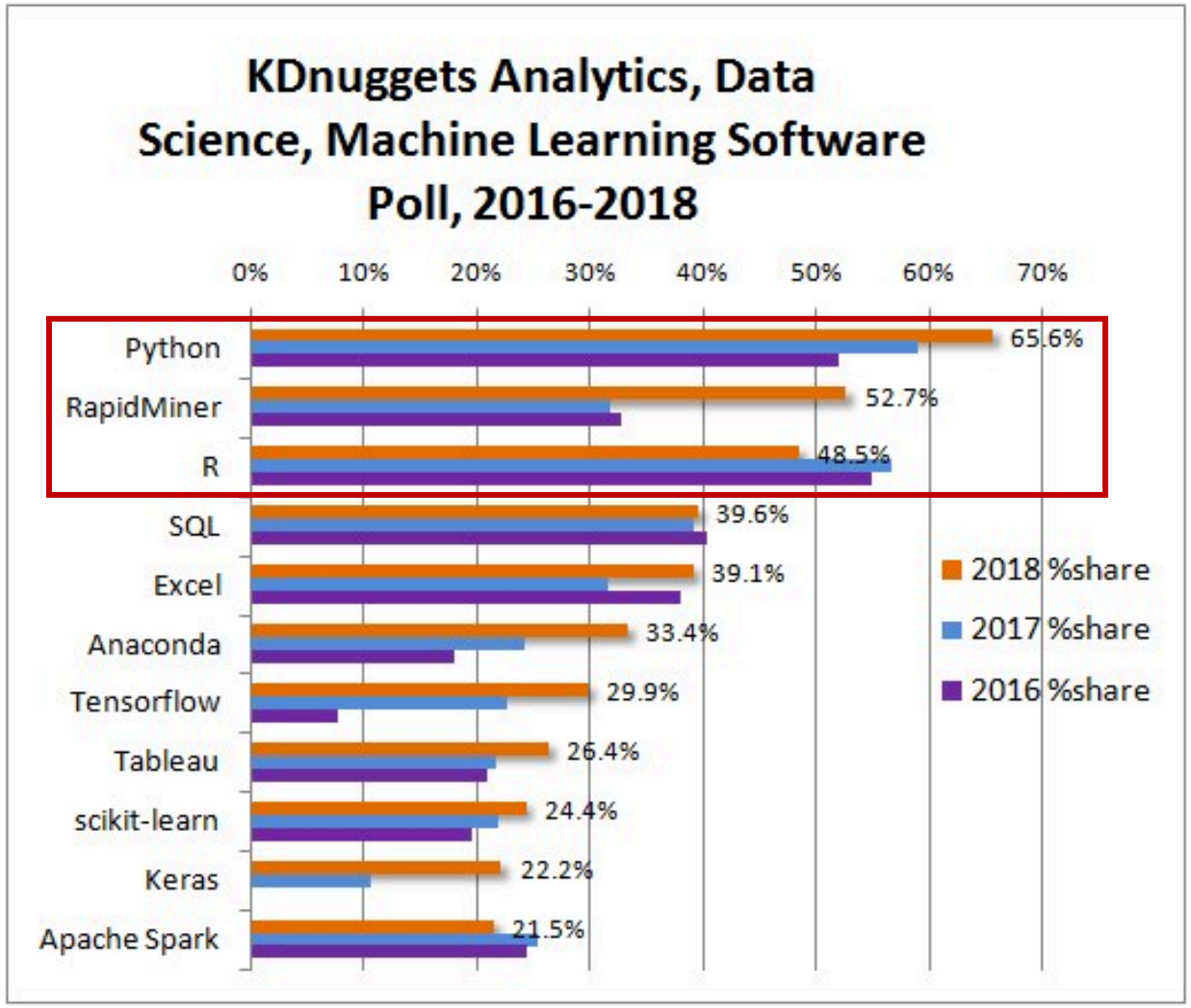
Outline

RapidMiner Overview

- Powerful data science platform for data preparation and machine learning
  - Visual design of the AI development workflow
  - Offers more than 1500 different operators
  - Especially allows for a fast and easy data exploration, data preparation, and AI prototyping
- **Origin**: Initially developed at the TU Dortmund
- **Nowadays**: Maintained by a commercial company & the open source community
- **Editions**: Open source & commercial versions exist
  - Free *Community Edition*
    (= second last version)
  - Commercial *Enterprise Edition*
    (= latest version + support services)



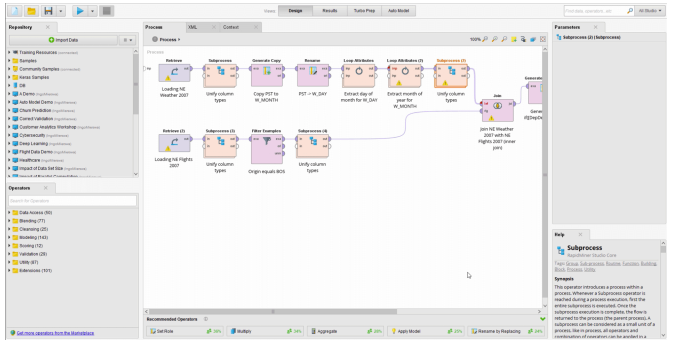
Rising Popularity of RapidMiner



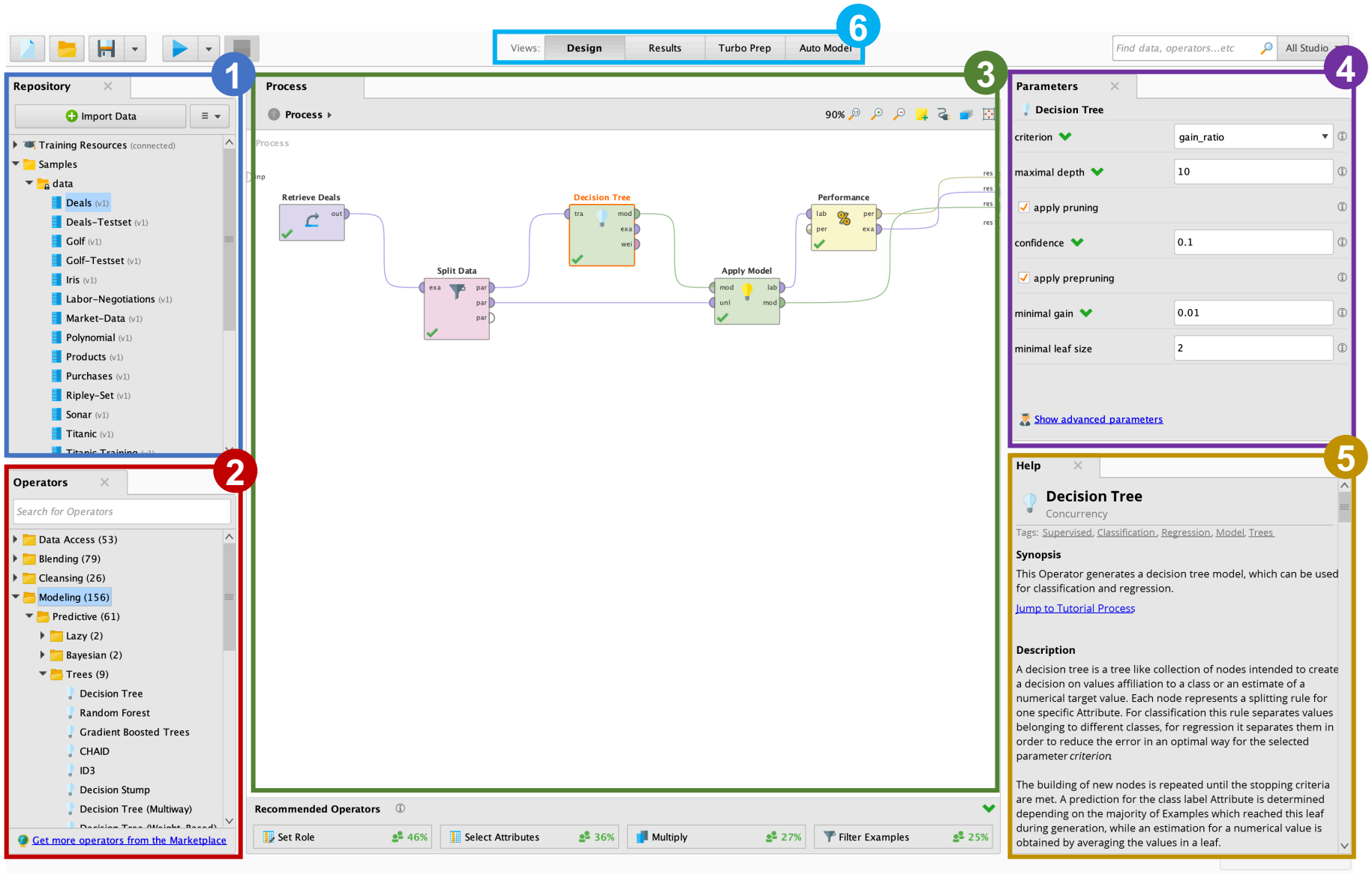
Source: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

- **Python** & **R** currently represent the **most popular programming languages** in the data science area

- **RapidMiner** has become the **leading GUI-based tool** for conducting data science projects and its popularity is increasing even further

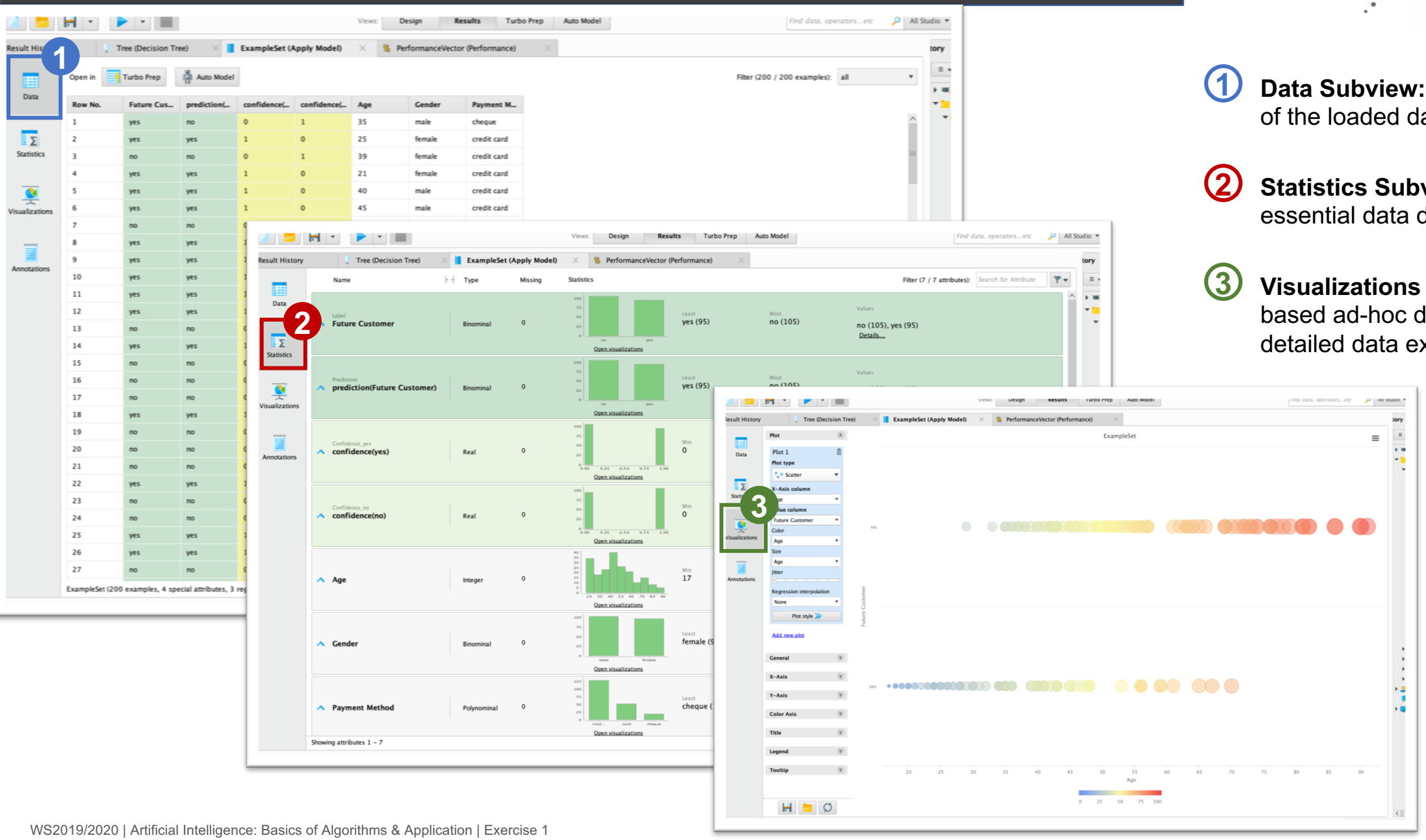


RapidMiner Studio – Main Views: The Design View



1. **Repository** holding accessible data sets & defined processes
2. **Operator catalogue** containing operators for, e.g., data preparation, model building and model evaluation
3. **Process design area** for defining especially data pipelines, model training, and model evaluation
4. Setting **parameters** for some selected operator
5. **Detailed information** w.r.t. some selected operator or other artifacts
6. Switch between the different **views**, especially the **Design & Result view**

RapidMiner Studio – Main Views: The Result View (1/2)

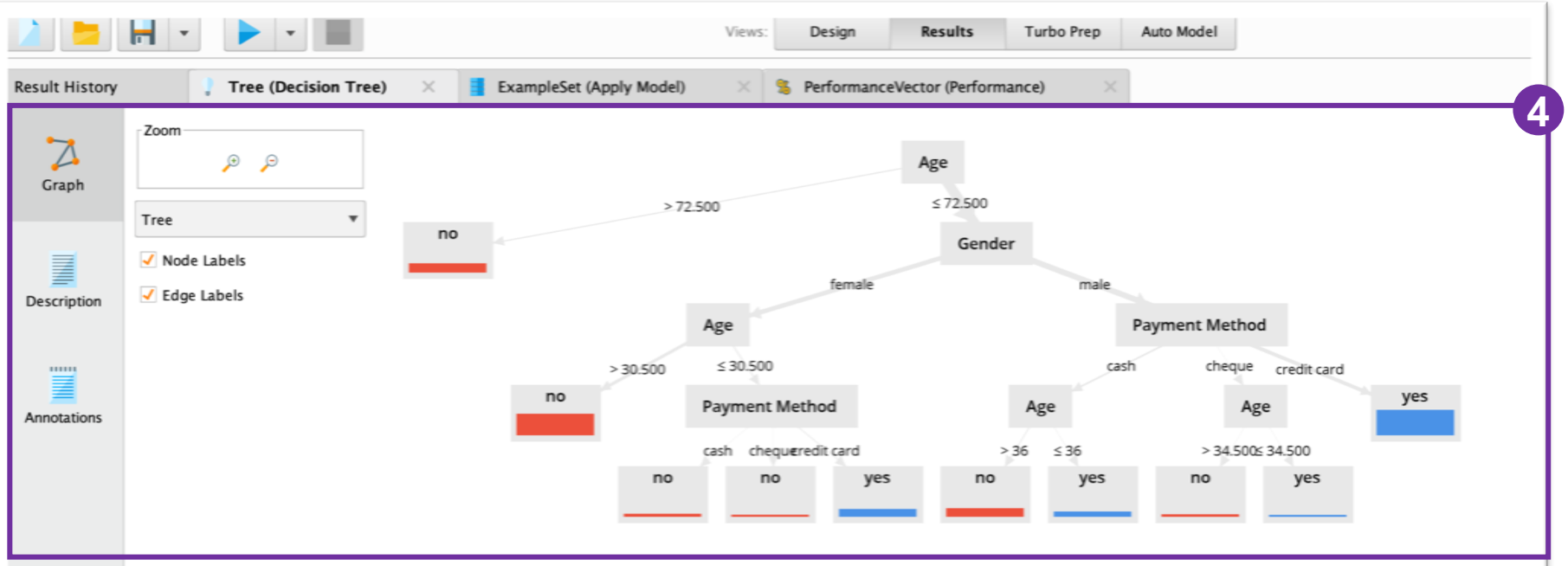


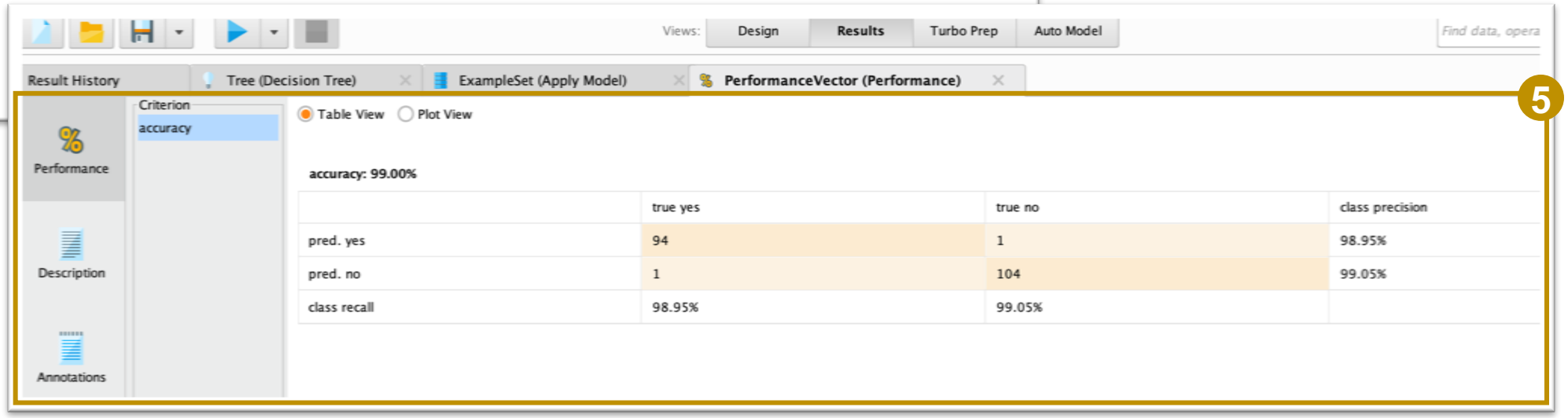
① **Data Subview:** tabular overview of the loaded data set

② **Statistics Subview:** Overview of essential data distributions

③ **Visualizations Subview:** Chart-based ad-hoc data analysis for a detailed data exploration

RapidMiner Studio – Main Views: The Result View (2/2)



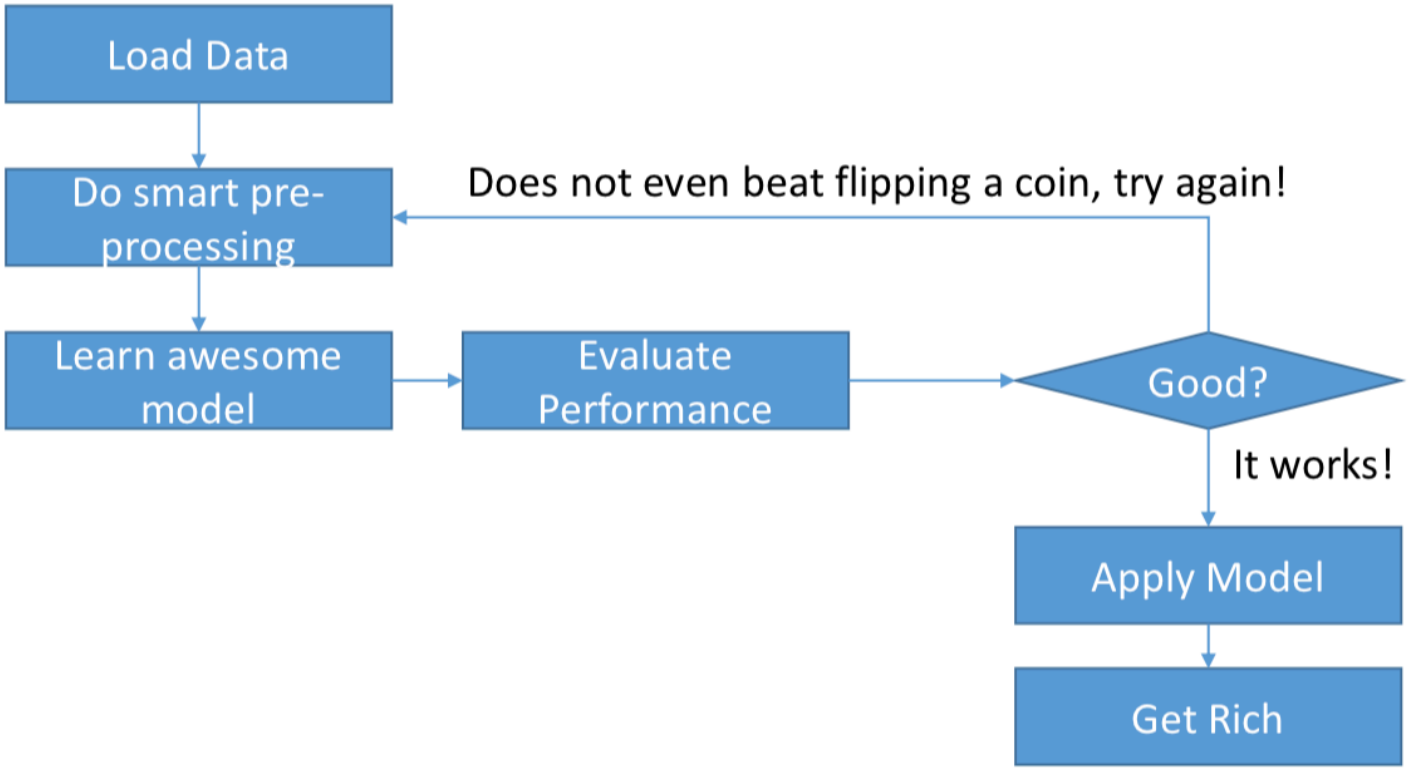


④ **Model Subview**: Visualization of the trained model (if possible)

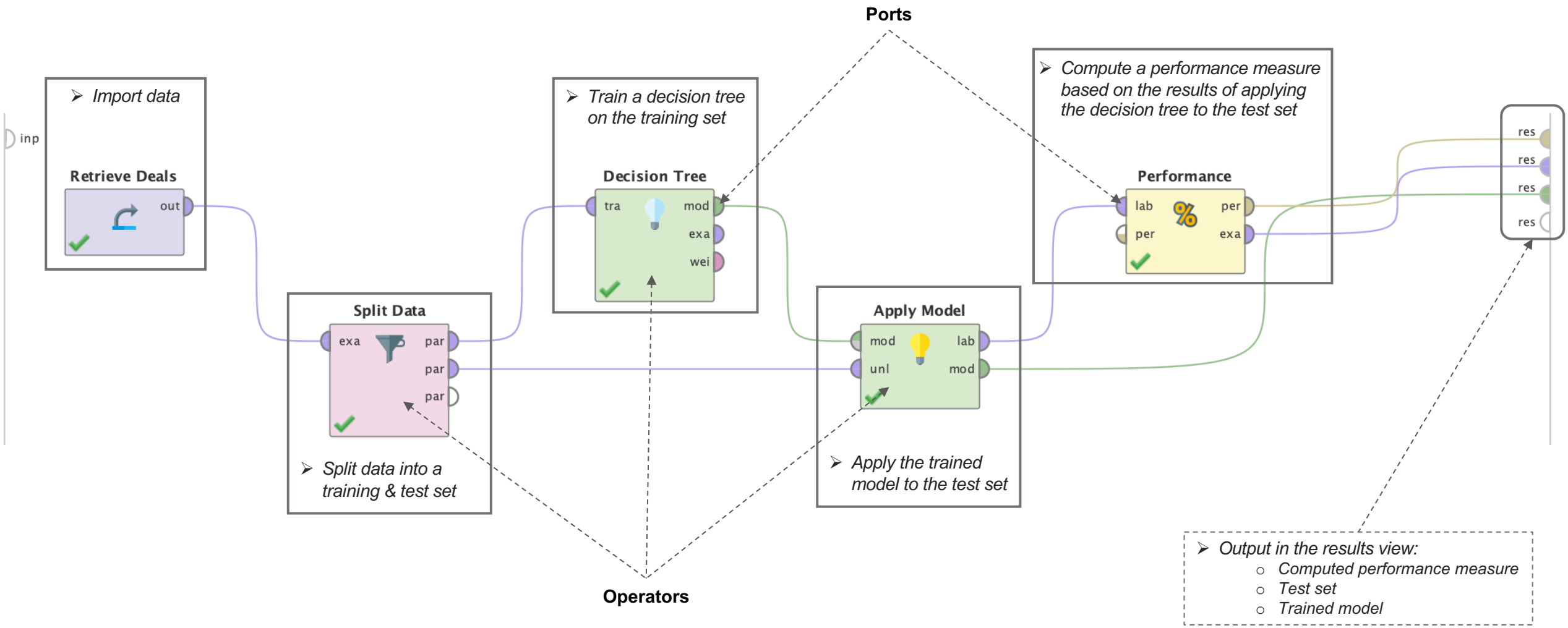
⑤ **Performance Subview:** Overview of the computed performance measure

In a Nutshell: Designing a RapidMiner Process

- **Visually design** a process for creating an AI model
- **Process** = *Data Pipeline* + *Model Development*



Processes in RapidMiner Studio: Chaining Operators



Outline

WIRTSCHAFTS INFORMATIK

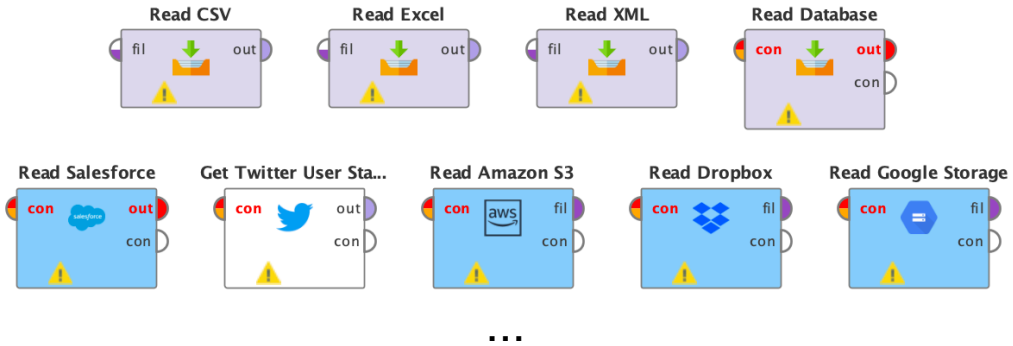
TECHNISCHE UNIVERSITÄT DARMSTADT

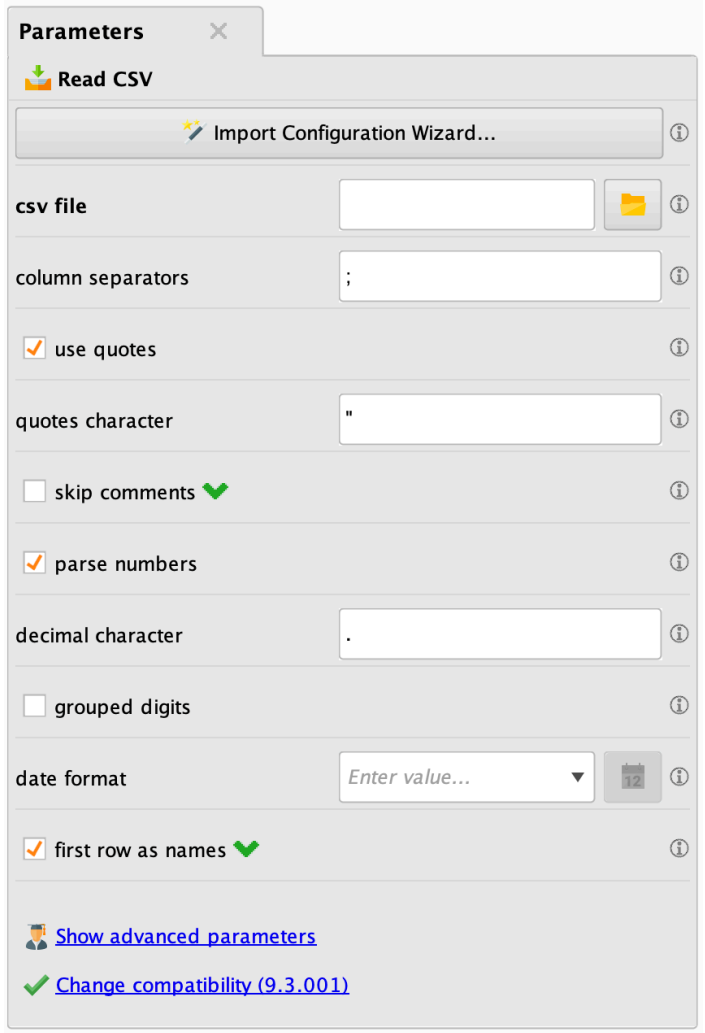
1 Introduction to RapidMiner

Operators for Data Import

- Many possibilities for importing data into RapidMiner exist
  - Read data from a **file** *(CSV, Excel, XML, ...)*
  - Read data from an **SQL database**
  - Read data from a particular **application** *(Salesforce, Twitter, ...)*
  - Read data from a **cloud storage** *(AWS, Dropbox, Google Cloud, ...)*

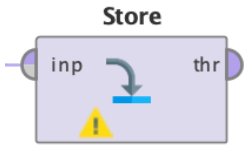
- Each operator creates an example set which contains your data
  - Records contained in your data are called "*examples*"





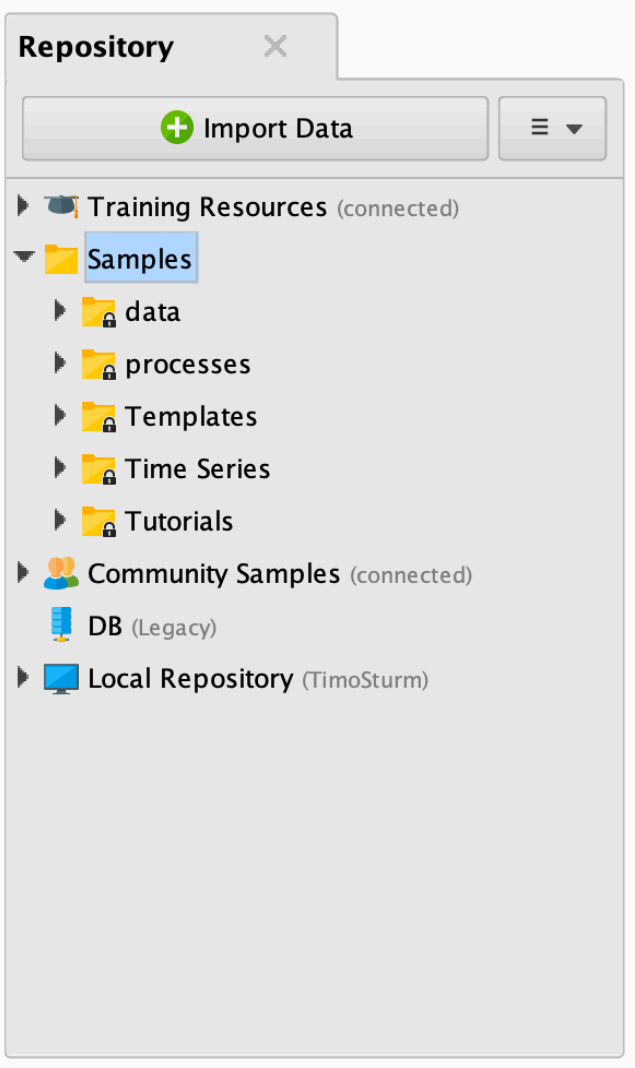
Repository

- Repository stores your data (with meta data) and your defined processes
  - Once you have imported your data and/or established a connection to a remote source, you can store the respective **example sets** in the **repository** by using the **store operator**



  - Alternatively, you can store data by following the **“Import Data” dialogue**

- By default, the repository already contains **popular samples** of example sets and **example processes** which can both be used for exploring the tool



Data in RapidMiner

- Each data set is represented as an **example set**
- Each example (= *table row*) is described by **several attributes** (= *features*)

  → **Each attribute** has:

  - a ***name***
  - a ***value type***
  - a ***role***

Customer ID	ItemsBought	ItemsReturned	ZipCode	Product
polynomin...	integer	integer	polynominal	polynominal
id	attribute	attribute	attribute	attribute
4	45	10	2	1365
5	42	18	5	2764
6	50	0	1	1343
8	13	12	4	2435
9	10	7	3	2435
10	34	17	6	2896
11	40	20	8	2869
12	40	8	2	1236
14	9	9	8	2435
15	36	7	2	1764
16	42	1	1	1547

attribute name

attribute data type

attribute role

Data in RapidMiner: Attribute Name & Value Types

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Attribute Name**: Unique name of an attribute (e.g. age, last_name, gender, ...)
- **Attribute Value Types**: Specifies the kind of data allowed for an attribute

Value Type	Description	Examples
nominal	• All kinds of text values • Includes polynomial and binomial	*boy/girl, hallo/hello/hola/salut/, ...*
binominal	• Exactly two values	*true/false, yes/no, head/tale, ...*
polynominal	• Many different string values	*red/green/blue/yellow, pig/chicken/cow, ...*
text	• Nominal data type that allows for more granular distinction (several text processing operators exist that only work with text)	*“red“/“green“/“blue“/“yellow“, “pig“/“chicken“/“cow“, ...*
file_path	• Nominal data type (rarely used) that allows for more granular distinction • Can be used to mark a column as *"only containing file paths*"	*/Users/Anonymous/Documents/VeryConfidentialData/fileName.csv; /Users/MrSampleMan/Desktop/data_final_v08.xlsx; ...*
numeric	• All kinds of number values • Includes date, time, integer, and real numbers	*3; 4.08; -1001; -0.00009; 24.12.2004 17:41; 02:38; ...*
integer	• A whole number	*42; -5; 11023; ...*
real	• A fractional number	*11.23; -0.0001; 2.37; ...*
date_time	• Both date and time	*23.12.2014 17:59; 01.01.1999 00:01; ...*
date	• Date without time	*23.12.2014; 05.04.2001; ...*
time	• Time without date	*17:59; 03:11; ...*

Data in RapidMiner: Attribute Roles

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Attribute Roles**: Determines how operators treat the different attributes
  - Label role is of utmost importance in defining the target for a prediction
  - Any attribute without a role assigned is known as a *regular* attribute

Role	Description
regular	• Attributes without a special role • Are used as input variables for learning tasks
id	• An attribute with the id role acts as an identifier for the examples • It should be unique for all examples • Different Blending Operators (Join, Union, Transpose, Pivot, ...) uses the id attribute to perform their tasks
label	• An attribute with the label role acts as a target attribute for learning operators • The label is also often called 'target variable' or 'class'
prediction	• An attribute with the prediction role is the result of an application of a learning model • The Apply Model Operator adds for example a prediction attribute to the example set • To evaluate the performance of a model, a label and a prediction attribute is necessary
cluster	• An attribute with the cluster role indicates the membership of an example set to a particular cluster • For example the k-Means Operator adds an attribute with the cluster role
weight	• An attribute with the weight role indicates the weight of the examples with regard to the label • Weights are used in learning processes to set the importance of examples • Weights can also be used to evaluate the performance of models; there they assign a severness for misclassification of single examples
batch	• An attribute with the batch role indicates the membership to a specific batch
user defined	• Any role can be assigned to an attribute by typing in the textbox • One specific user defined role cannot be assigned to more than one attribute • Attributes with user defined roles are ignored in learning processes → An attribute with a user defined role is ignored in learning processes but remains in the example set

Data in RapidMiner: Data Preparation

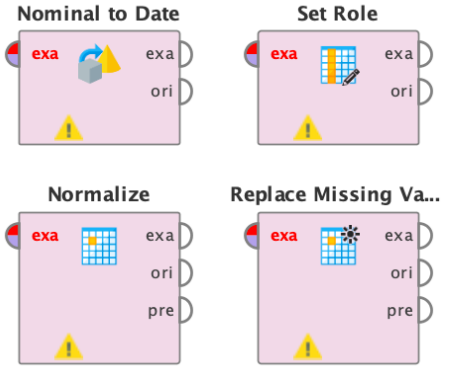
WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

- RapidMiner offers many operators for **transforming data**, e.g.:

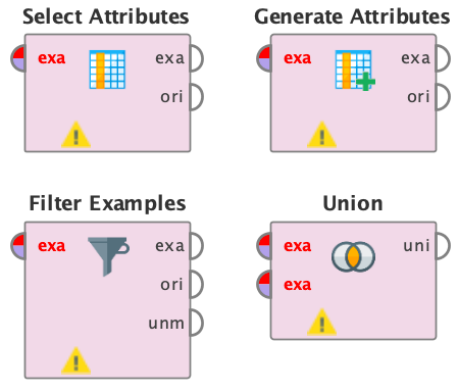
➢ Attribute Conversions & Transformations

- **Attribute Value Type Conversions**: Transform attribute from type A to type B
- **Attribute Role Conversions**: Change an attribute's role
- **Normalize**: Scale values so they fit in a specific range
- **Replace Missing Values**: Replace missing values with a specified replacement
- ...



➢ Data Set Manipulation

- **Select Attributes**: Remove attributes
- **Generate Attributes**: Create new attributes
- **Filter Examples**: Remove examples
- **Aggregation & Joins**: SQL-like functions for aggregating values and connecting data tables
- ...



Outline

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

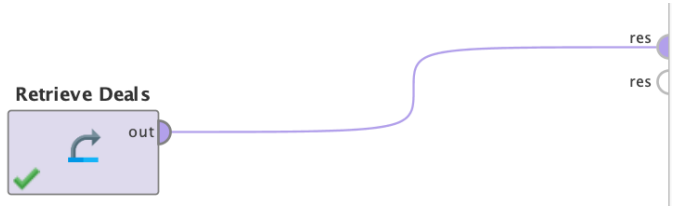
1 Introduction to RapidMiner

Data Exploration in the Result View

- The result view provides a **rich set of visualization capabilities** which allows you to explore the data as part of your data understanding phase

- Simply load data by connecting a **data retrieval operator** with a **result port**

- In the **data subview**, RapidMiner with an automatically generated **overview of each example** visualized in a **tabular format**

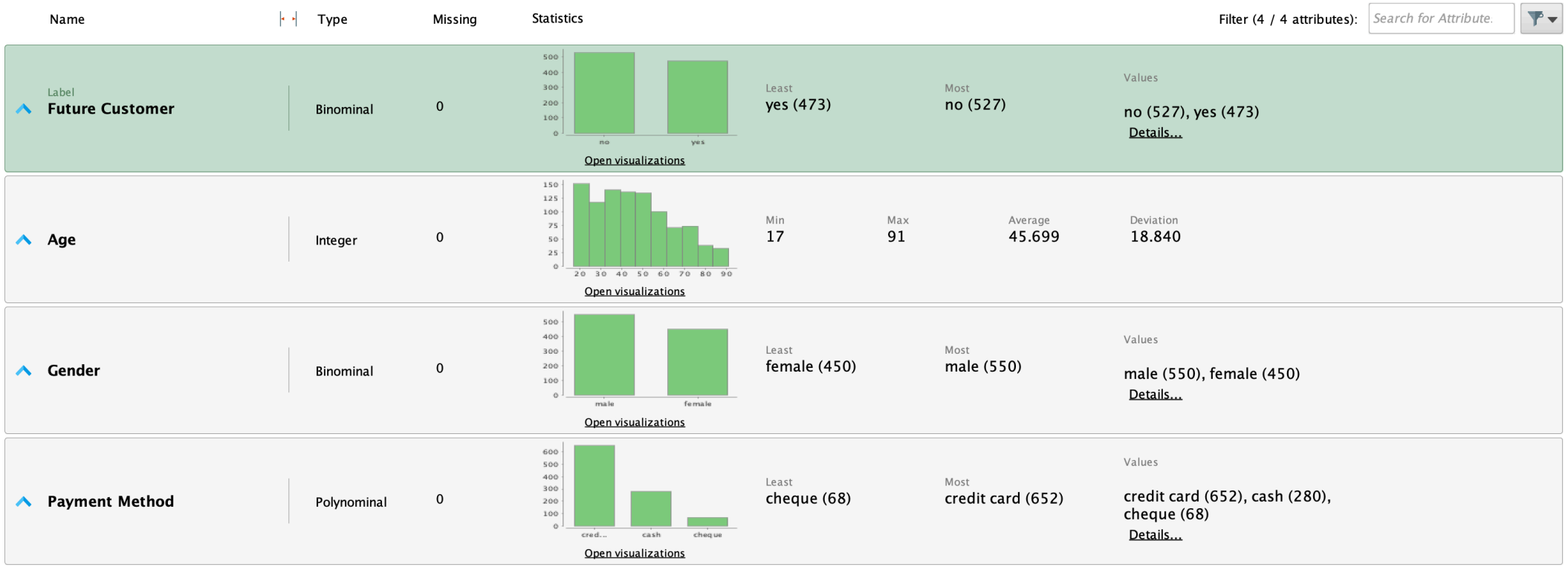


Row No.	Future Cus...	Age	Gender	Payment M...
1	yes	64	male	credit card
2	yes	35	male	cheque
3	yes	25	female	credit card
4	no	39	female	credit card
5	yes	39	male	credit card
6	no	28	female	cheque
7	yes	21	female	credit card
8	yes	48	male	credit card
9	no	70	female	credit card
10	yes	36	male	credit card
11	yes	22	male	credit card
12	no	53	female	cash
13	yes	27	male	cash
14	yes	40	male	credit card
15	yes	22	male	cash
16	no	49	female	credit card
17	no	24	female	cash
18	yes	45	male	credit card
19	yes	45	male	credit card
20	no	66	female	cash
21	no	82	female	cash
22	no	35	female	credit card
23	yes	17	female	credit card
24	no	52	male	cash
25	no	49	female	credit card
26	no	33	female	credit card
27	yes	40	male	credit card

ExampleSet (1,000 examples, 1 special attribute, 3 regular attributes)

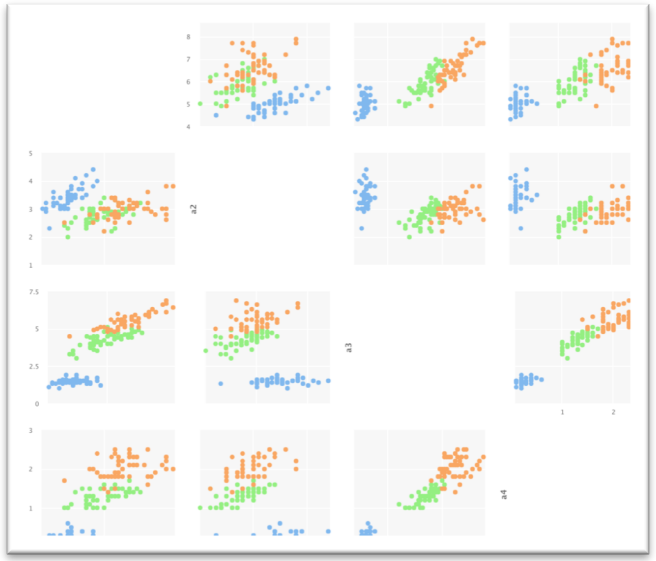
Data Exploration in the Result View

- In the **statistics subview**, RapidMiner automatically provides you with a **distribution overview** of each attribute, including the amount of missing values

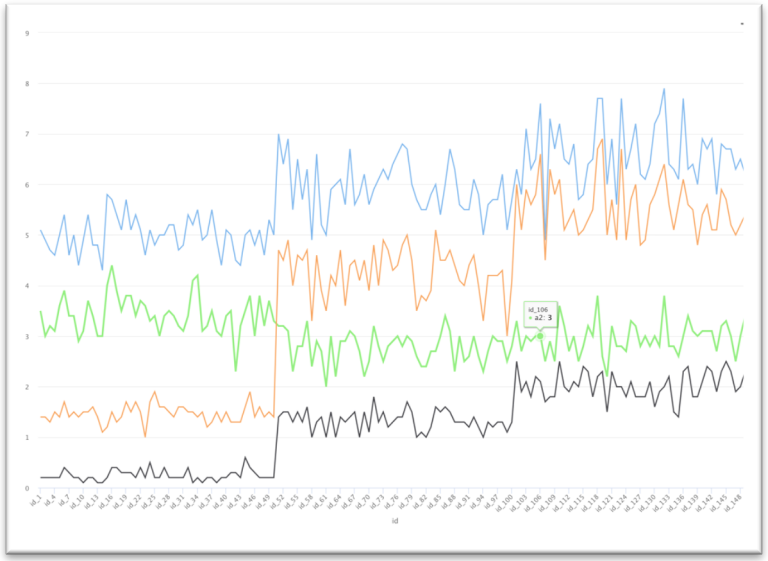


Data Exploration in the Result View

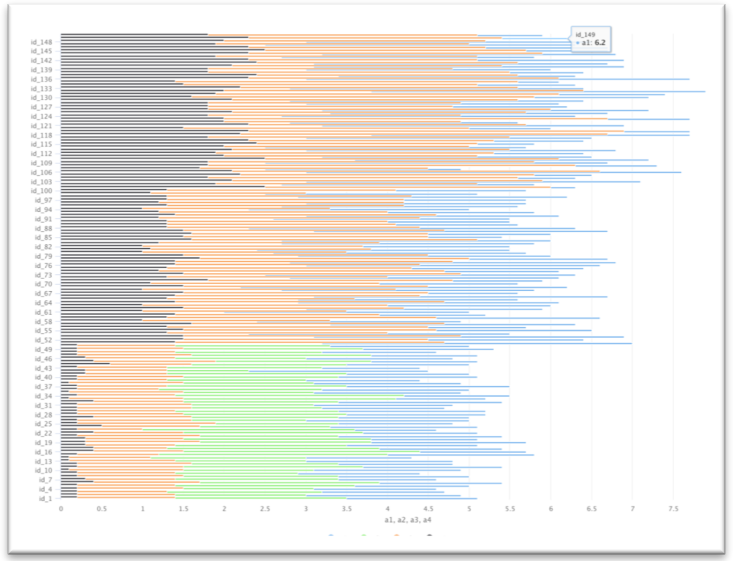
- In the **visualizations subview**, RapidMiner provides you visualization capabilities for conducting **ad-hoc analyses** to explore your data in-depth with focus on manually selected attributes
- Basically used to check whether you can observe first correlations or trends in your data set



Scatter Matrix



Line Chart

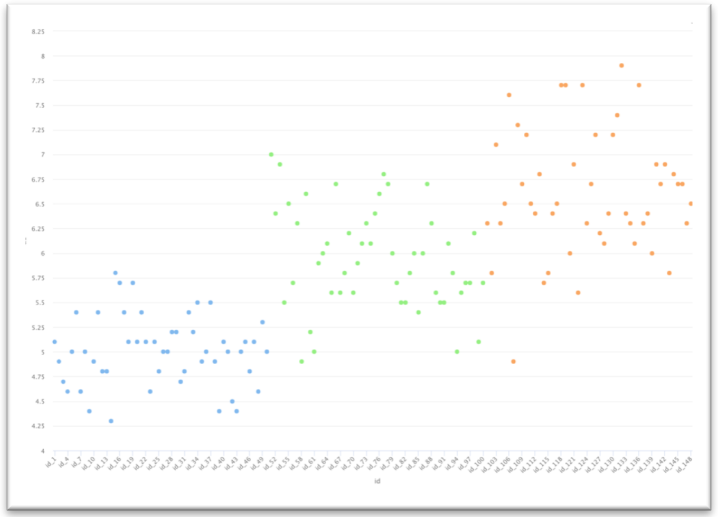


Horizontal Bar Chart

... and many more visualization possibilities!

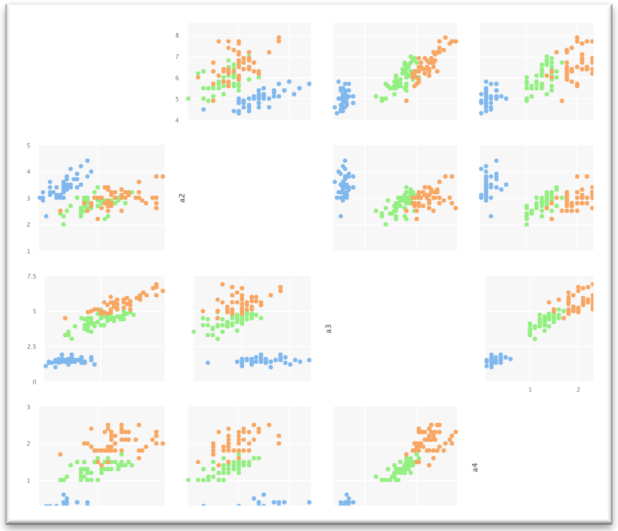
Data Exploration: Scatter Plots & Scatter Matrices

- **Scatter plots** are a useful tool to explore relations between attributes
  - Most commonly, **two-dimensional scatter plots** are used,
    but RapidMiner also allows the creation of three-dimensional scatter plots
  - The scatter plots can be extended by using different **colors**, **sizes**, or **shapes** for representing **further dimensions**



Scatter Plot

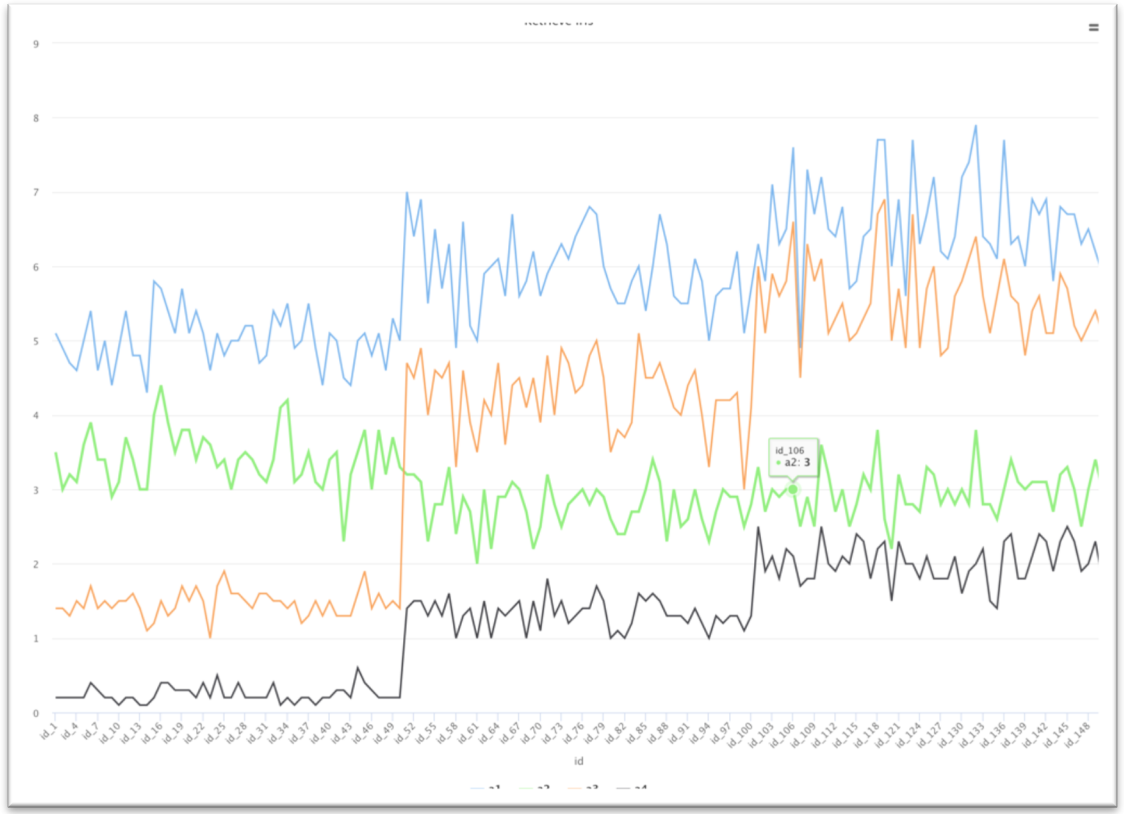
- A **scatter matrix** represents an overview of **two-dimensional scatter plots** of **all attribute pairs** contained in your data set



Scatter Matrix

Data Exploration: Line Charts

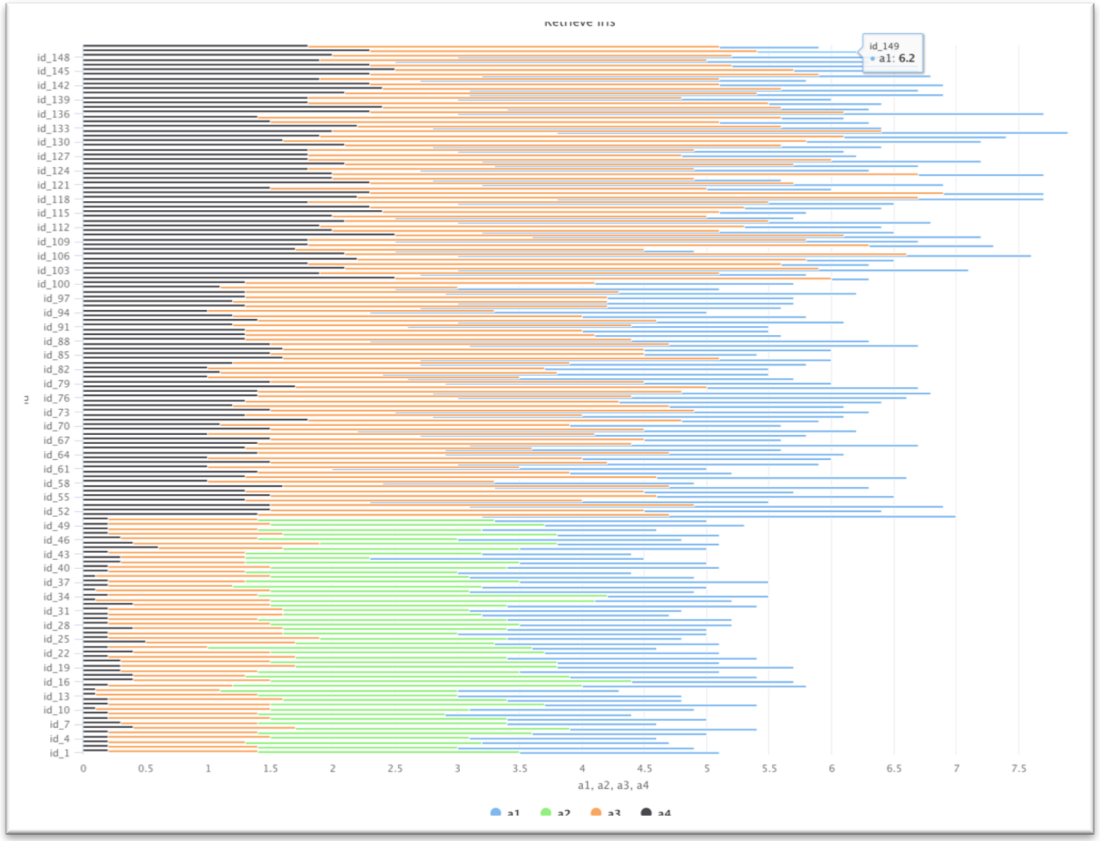
- Line charts can also be used to visualize relations between (most commonly two) attributes
- Line charts are especially helpful for visualizing developments of an attributes' values over time



Line Chart

Data Exploration: Bar Charts

- Bar charts can be used to analyze the distribution of single attributes
- Length of the bars can visualize, e.g.:
  - the amount of an attribute value in the data set
  - the numeric value of a single attribute



Horizontal Bar Chart

Outline

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Helpful Resources

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

- **RapidMiner Operator Reference Guide:** https://docs.rapidminer.com/latest/studio/operators/
- **RapidMiner Community:** https://community.rapidminer.com/
- **Book: „*Data Mining for the Masses*“:** https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf
- **Webinars & Videos offered by RapidMiner:** https://rapidminer.com/webinars-videos/
- **RapidMiner Academy:** https://academy.rapidminer.com/

Try it yourself!

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

- Install the most recent version of RapidMiner Studio and get an educational license
  - **URL**: https://rapidminer.com/get-started-educational/

- Explore the data management and exploration capabilities of RapidMiner Studio by completing the tasks of the exercise 1 sheet