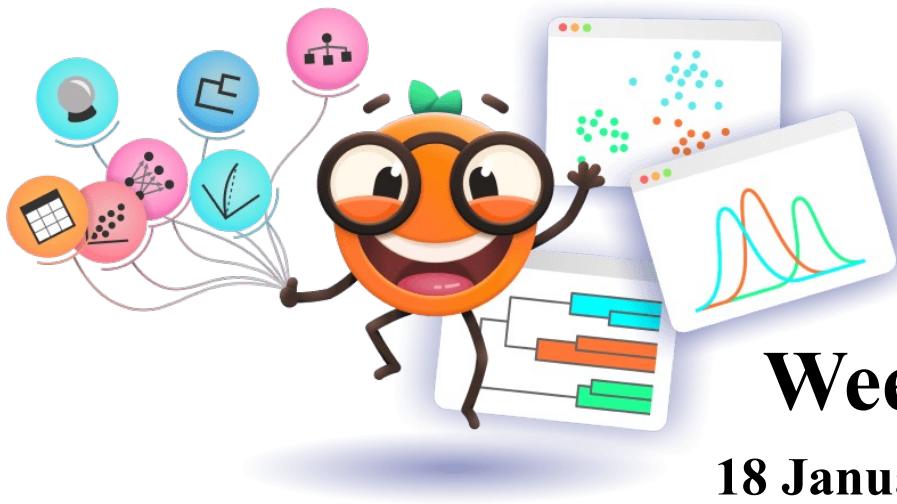




# Data Science

# Introduction to Orange3



## Training Series

### Week 1

18 Januari 2024

Orange3

Prepared by I Made Murwantara, Universitas Pelita Harapan

©IMM 2024



# Overview

1. What is Orange3
2. Resources and Installation
3. Visual Modelling
4. Dataset
5. Basic Module & Library
6. Add-on
7. First Try



# What is Orange3

- University of Ljubljana, Slovenia
- <https://orange.biolab.si/>
- Orange Widgets and Canvas are based on Qt,
  - which is distributed under GPL 3.0 (as well as LGPL 2.1, see <https://www.qt.io/licensing/>)
- Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) [Orange: Data Mining Toolbox in Python](#), *Journal of Machine Learning Research* 14(Aug): 2349–2353.
- A visual programming tools for Data Science



# Resources

- <https://orangedatamining.com>
- Windows & MacOS, Linux (via pip)
- Global Discussion via  
datascience.stackexchange.com
- Documents: <http://file.biolab.si/notes/>
- Blog: <https://orangedatamining.com/blog/>



# Installation

<https://orangedatamining.com/download/>

## Windows

### Standalone installer (default)

↳ [Orange3-3.36.2-Miniconda-x86\\_64.exe \(64 bit\)](#)

Can be used without administrative privileges.

## pip

Orange can also be installed from the Python Package Index. You may need additional system packages provided by your distribution.

```
pip install orange3
```

## Installing from Source

Clone our repository from [GitHub](#) or download the [source code tarball](#). Then follow the instructions in [README.md](#)

To run Orange Canvas run

```
python -m Orange.canvas
```



# Add-On

1. Orange3-Associate
2. Orange3-Bioinformatics
3. Orange3-Educational
4. Orange3-Explain
5. Orange3-Fairness
6. Orange3-Geo
7. Orange3-ImageAnalytics
8. Orange3-Network
9. Orange3-Prototypes
10. Orange3-SingleCell
11. Orange-Spectroscopy
12. Orange3-Survival-Analysis
13. Orange3-Text
14. Orange3-Textable
15. Orange3-Timeseries
16. Orange3-WorldHappiness

Installer

Filter... Add more...

Name	Version	Action
- Orange3	3.35.0 < 3.36.2	
<input type="checkbox"/> Associate	1.3.0	
<input checked="" type="checkbox"/> Bioinformatics	4.8.0 < 4.8.1	
<input type="checkbox"/> Educational	0.7.2	
<input type="checkbox"/> Explain	0.6.8	
<input type="checkbox"/> Fairness	0.1.7	
- Geo	0.4.0 < 0.4.1	
<input type="checkbox"/> Image Analytics	0.12.2	
<input checked="" type="checkbox"/> Network	1.8.0	
<input type="checkbox"/> Prototypes	0.21.1	
<input checked="" type="checkbox"/> Single Cell	1.5.0	
<input checked="" type="checkbox"/> Spectroscopy	0.6.10 < 0.6.11	
<input checked="" type="checkbox"/> Survival Analysis	0.5.1 < 0.6.0	
<input checked="" type="checkbox"/> Text	1.14.0 < 1.15.0	
<input type="checkbox"/> Textable	3.1.11	
<input checked="" type="checkbox"/> Timeseries	0.6.0 < 0.6.1	
<input type="checkbox"/> World Happiness	0.1.9	



# Screenshot – Dataset

Load & Edit  
data in the file  
widget

File

(File: ecoli.tab) ... Reload

URL:

Info

336 instance(s), 7 feature(s), 1 meta attribute(s)  
Classification; discrete class with 8 values.

Columns (Double click to edit)

1	mcg	C	numeric	feature
2	gvh	C	nominal	feature
3	lip	D	string datetime	feature 0.48, 1.00
4	chg	D	nominal	feature 0.50, 1.00
5	aac	C	numeric	feature
6	alm1	C	numeric	feature
7	alm2	C	numeric	feature
8	localization site	D	nominal	target cp, im, imL, imS, imU, om, omL, pp
9	name	S	string	meta

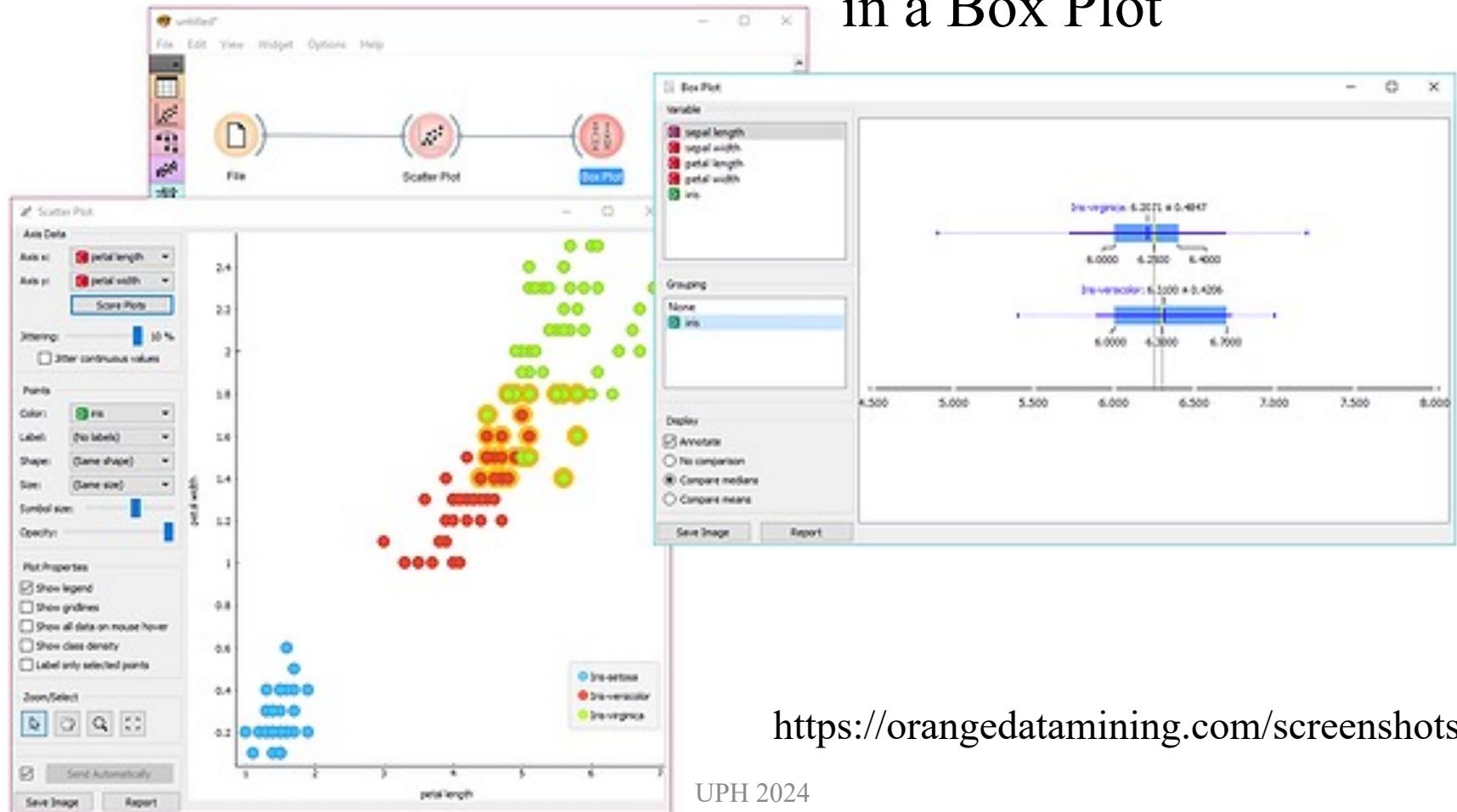
Browse documentation data sets Report Apply

<https://orangedatamining.com/screenshots/>



# Screenshot – Interactive Visualisation

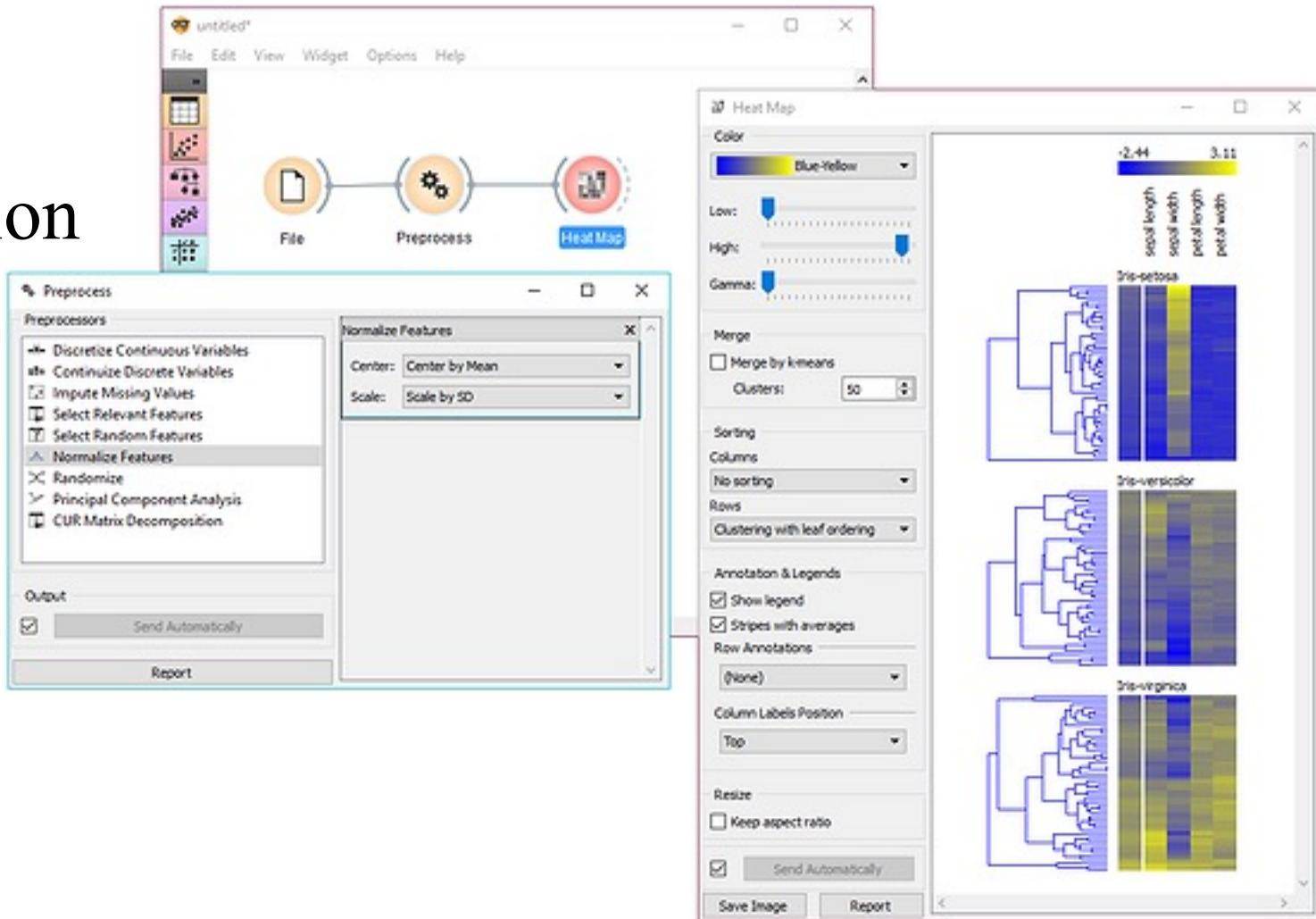
Data Selection in  
Scatter Plot visualized  
in a Box Plot





# Screenshot – Analytic Visualisation

## Heatmap Visualisation





# Help

Tree

Name: Tree

Parameters

- ✓ Induce binary tree
- ✓ Min. number of instances in leaves: 2
- ✓ Do not split subsets smaller than: 5
- ✓ Limit the maximal tree depth to: 100

Classification

- ✓ Stop when majority reaches [%]: 95

Apply Automatically

Tree

A tree algorithm with forward pruning.

Inputs

- Data: input dataset
- Preprocessor: preprocessing method(s)

Outputs

- Learner: decision tree learning algorithm
- Model: trained model

**Tree** is a simple algorithm that splits the data into nodes by class purity (information gain for categorical and MSE for numeric target variable). It is a precursor to [Random Forest](#). Tree in Orange is designed in-house and can handle both categorical and numeric datasets.

It can also be used for both classification and regression tasks.

Tree

Name: Tree

①

Parameters

②

- ✓ Induce binary tree
- ✓ Min. number of instances in leaves: 2
- ✓ Do not split subsets smaller than: 5
- ✓ Limit the maximal tree depth to: 100



# Example Workflows

**Example Workflows**

## File and Data Table

The basic data mining units in Orange are called widgets. There are widgets for reading the data, preprocessing, visualization, clustering, classification and others. Widgets communicate through channels. Data mining workflow is thus a collection of widgets and communication channels.

In this workflow, there is a File widget that reads the data. File widget communicates this data to Data Table widget that shows the data spreadsheet. Notice how the output of the file widget is connected to the input of the Data Table widget.

**Path:** /Applications/Orange.app/Contents/Frameworks/Python...anvas/workflows/110-file-and-data-table-widget.ows

A File widget. Double click to open it and select the dataset file.

A Data Table widget. Double click the icon to see the data in a spreadsheet.

The output of the Data Table to send out any data (rows) that are selected to the widget.

This output is not used, hence dashed line. You can add another Data Table by clicking on its icon from the toolbox on the left, connect the output of Data Table to the input of new Data Table (1) and check if the selected data from Data Table is indeed sent to the downstream widget. This demo works best if both widgets are open, that is, their windows displayed.

The output of the File widget.

The input of the Data Table widget.

The communication channel. It passes the dataset from the File widget to the Data Table.

File and Data Table

Interactive Visualizations

Visualization of Data Subsets

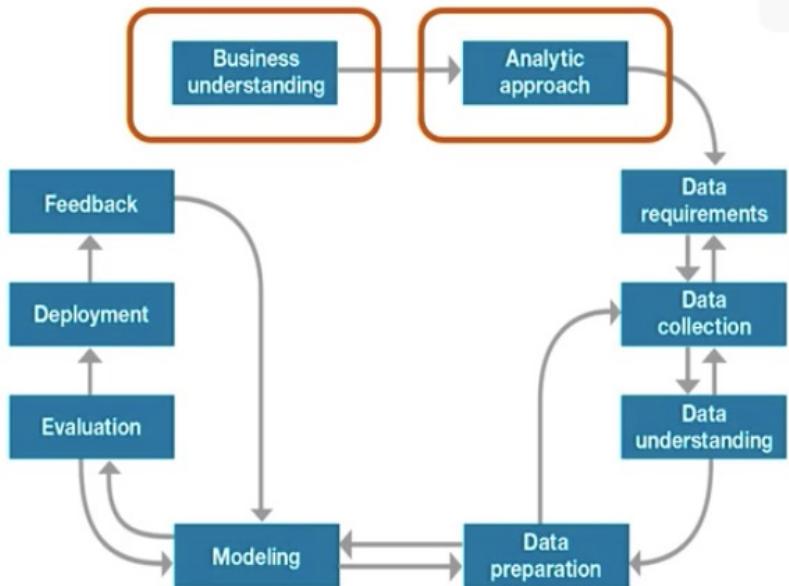
Classification Tree

Principal Component Analysis

Hierarchical Clustering



# Data Science Methodology



## Business understanding

- What is the problem that you are trying to solve?*

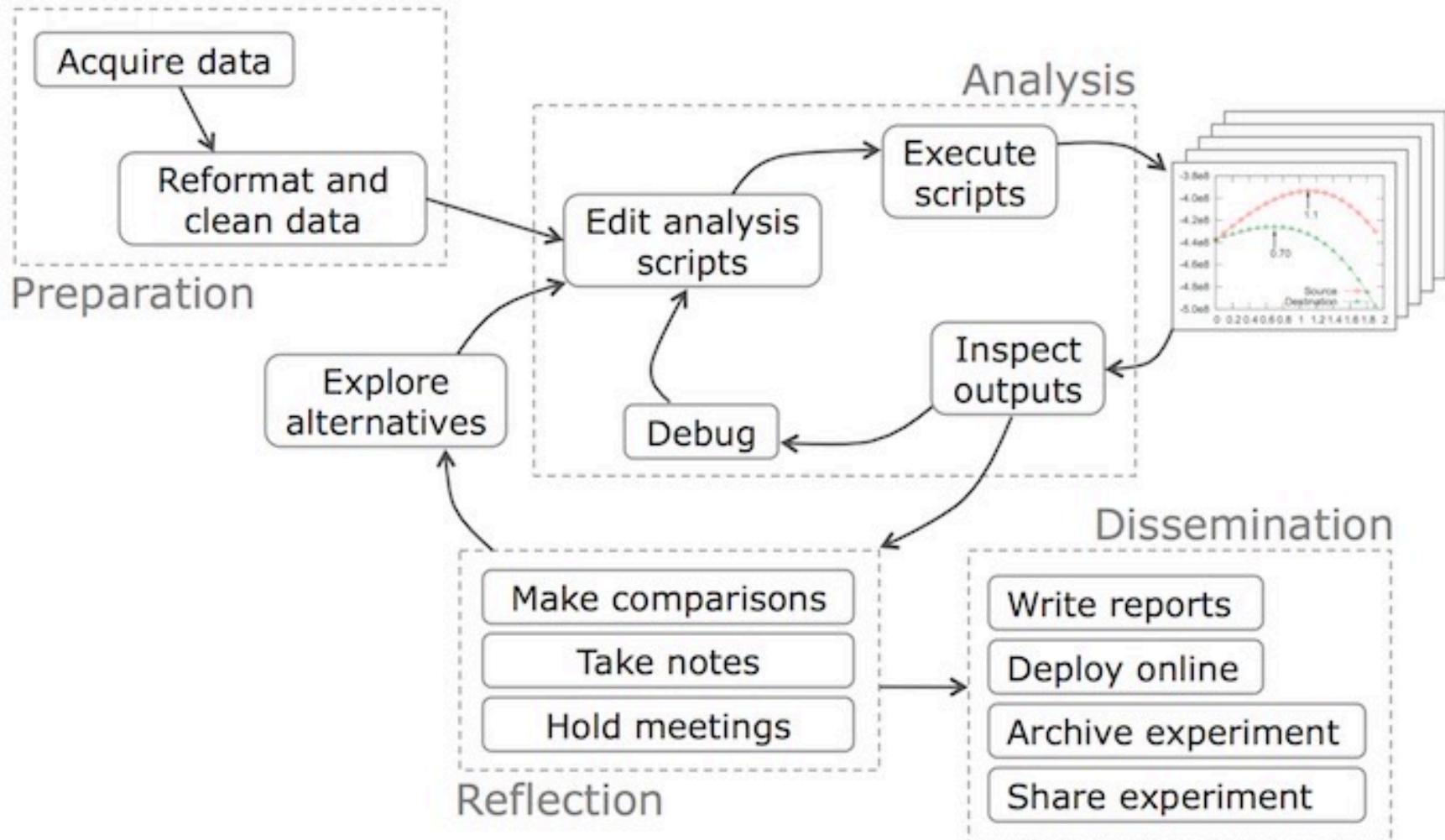


## Analytic approach

- How can you use data to answer the question?*



# Data Science Workflow



<https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext>



# Data Engineering

**Data**

File	CSV File Import	Datasets	SQL Table
Data Table	Paint Data	Data Info	Rank
Edit Domain	Color	Feature Statistics	Save Data

**Transform**

Data Sampler	Select Columns	Select Rows	Transpose
Single Cell Preprocess	Merge Data	Concatena...	Select by Data Index
Unique	Aggregate Columns	Group by	Pivot Table
Apply Domain	Preprocess	Impute	Continuize
Discretize	Randomize	Purge Domain	Melt
Formula	Create Class	Create Instance	Python Script

Python Script

Editor

```
def python_script():
    import numpy as np
    from Orange.data import Table, Domain, ContinuousVariable,
    DiscreteVariable
    domain = Domain([ContinuousVariable("age"),
                    ContinuousVariable("height"),
                    DiscreteVariable("gender", values=["M", "F"]))]
```

Library

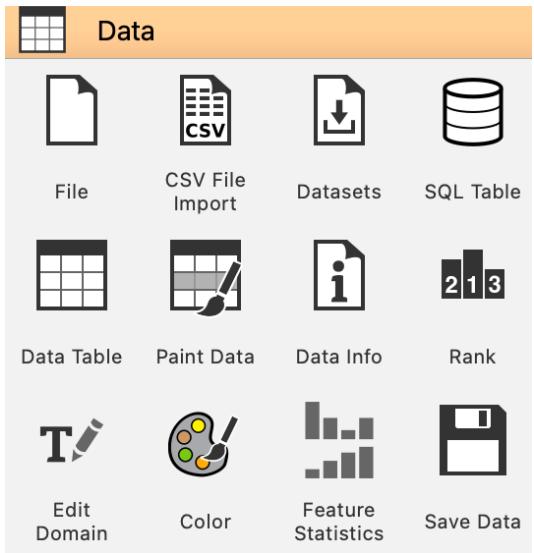
Table from numpy

Console

```
Python 3.8.12 (v3.8.12:b28285d7ec, Mar 23 2022, 18:22:40)
[Clang 13.0.0 (clang-1300.0.29.30)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
(ipythonConsole)
>>>
```



# Data - File



File

File: datasets/iris.tab

URL: [https://github.com/Opensourcefordatascience/Data-sets/raw/master/blood\\_pressure.csv](https://github.com/Opensourcefordatascience/Data-sets/raw/master/blood_pressure.csv)

File Type

Automatically detect type

Info

120 instances  
5 features (no missing values)  
Data has no target variable.  
0 meta attributes

Columns (Double click to edit)

Name	Type	Role	Values
1 patient	N numeric	feature	
2 sex	C categorical	feature	Female, Male
3 agegrp	C categorical	feature	30-45, 46-59, 60+
4 bp_before	N numeric	feature	
5 bp_after	N numeric	feature	

Source

File: datasets/iris.tab

URL:

File Type

✓ Automatically detect type

- Basket file (\*.basket \*.bsk)
- Comma-separated values (\*.csv \*.csv.gz \*.gz \*.csv.bz2 \*.bz2 \*.csv.xz \*.xz)
- Microsoft Excel 97-2004 spreadsheet (\*.xls)
- Microsoft Excel spreadsheet (\*.xlsx)
- Pickled Orange data (\*.pkl \*.pickle \*.pk.gz \*.pickle.gz \*.gz \*.pkl.bz2 \*.pickle.bz2 \*.bz2 \*.pkl.xz \*.pickle.xz \*.xz)
- Tab-separated values (\*.tab \*.tsv \*.tab.gz \*.tsv.gz \*.gz \*.tab.bz2 \*.tsv.bz2 \*.bz2 \*.tab.xz \*.tsv.xz \*.xz)
- Agilent Mosaic Image (\*.dmt)
- Agilent Mosaic Image (IFG) (\*.dmt)
- Agilent Mosaic Tile-by-tile (\*.dmt)
- Agilent Single Tile Image (\*.dat)
- Agilent Single Tile Image (IFG) (\*.seq)
- Envi (\*.hdr)
- Envi hdr or STXM hdr+xim files (\*.hdr)
- Galactic SPC format (\*.spc \*.SPC)
- Gwyddion Simple Field (\*.gsf)
- HDF5 file @HERMRES/SOLEIL (\*.hdf5)
- HDF5 file @ROCK(hyperspectral imaging)/SOLEIL (\*.h5)
- Hyperspectral map ASCII (\*.xyz)
- Matlab (\*.mat)
- NXS HDF5 file @O8/Diamond Light Source (\*.nxs)
- NeaSPEC (\*.nea \*.txt)
- NeaSPEC raw files (\*.gsf)
- OPUS Spectrum (\*.0\* \*.1\* \*.2\* \*.3\* \*.4\* \*.5\* \*.6\* \*.7\* \*.8\* \*.9\*)
- Omnic map (\*.map)
- PTIR Studio file (\*.ptir)
- Renishaw WiRE WDF reader (\*.wdf \*.WDF)
- SPA (\*.spa \*.SPA \*.srs)
- STXM/NEXAFS .hdr+.xim files (\*.hdr)
- Spectra ASCII (\*.dat \*.dpt \*.xy \*.csv)
- Spectra ASCII or Agilent Single Tile Image (\*.dat)
- XAS ascii spectrum from ROCK (\*.txt)



# Data - CSV

Encoding: Unicode (UTF-8)

Cell delimiter: Comma

Quote character: "

Number separators: Grouping: Decimal:

Column type:

	1	2	3	4	5
1	patient	sex	agegrp	bp_before	bp_after
2	1	Male	30-45	143	153
3	2	Male	30-45	163	170
4	3	Male	30-45	153	168
5	4	Male	30-45	153	142
6	5	Male	30-45	146	141
7	6	Male	30-45	150	147
8	7	Male	30-45	148	133
9	8	Male	30-45	153	141
10	9	Male	30-45	153	131
11	10	Male	30-45	158	125
12	11	Male	30-45	149	164
13	12	Male	30-45	173	159
14	13	Male	30-45	165	135
15	14	Male	30-45	145	150

Reset Restore Defaults

Column type: Auto

	1	3	4	5
1	patient	agegrp	bp_before	bp_after
2	1	30-45	143	153
3	2	30-45	163	170
4	3	Male	30-45	153
5	4	Male	30-45	168
6	5	Male	30-45	142
7	6	Male	30-45	146
8	7	Male	30-45	141
9	8	Male	30-45	133
10	9	Male	30-45	158
11	10	Male	30-45	125
12	11	Male	30-45	149
13	12	Male	30-45	173
14	13	Male	30-45	165
15	14	Male	30-45	150





# Datasets

Title	Size	Instances	Variables	Target	Tags
BBC3	2.6 MB	1407	3	C categorical	text, classification, news
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	C categorical	biology
Smoking effect on B lymphocytes	1.8 MB	79	3000	C categorical	genomics
HDI	45.2 KB	188	53		economy, geo
ParlaMint	1.7 MB	1000	17	C categorical	text, classification, time, politics
SentiNews	5.0 MB	2000	7	C categorical	text, sentiment
TKI resistance	1.2 MB	280	467	C categorical	spectral
Abalone	187.5 KB	4177	8	N numeric	biology
Adult	4.1 MB	32561	15	C categorical	economy, fairness
Roman Amphorae	23.7 KB	164	16	C categorical	archaeology, image analytics
Attrition - Predict	838 bytes	3	18	C categorical	economy, synthetic, education
Attrition - Train	182.2 KB	1470	18	C categorical	economy, synthetic
Auto MPG	17.3 KB	398	9	N numeric	
Bank Marketing	466.1 KB	4119	20	C categorical	economy
Banking Crises	31.3 KB	211	73		time, economy
Bank Note	11.0 KB	67	2	C categorical	classification, time

## Description

**Breast Cancer and Docetaxel Treatment** (2006), from [NCBI](#)

Breast cancer core biopsies taken from patients found to be resistant (greater than 25% residual tumor volume) or sensitive (less than 25% residual tumor volume) to docetaxel treatment.

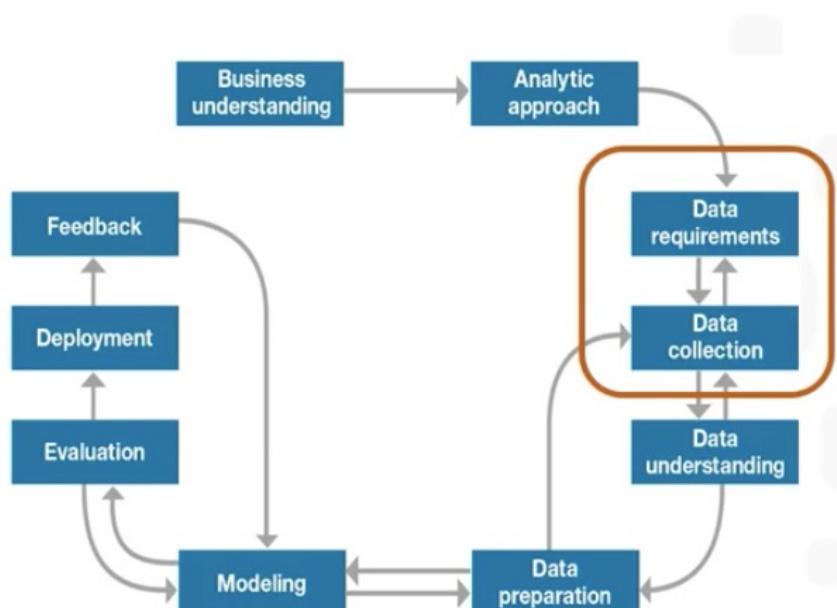
## References

Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG et al. (2005) Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. *J Clin Oncol*, 23(6): 1169-77.



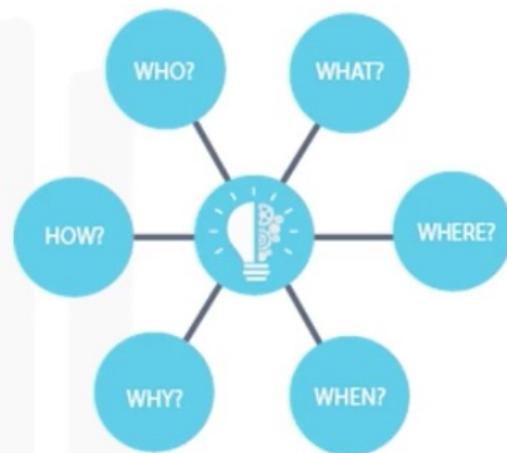
# Methodology

## From Requirements to Collection



### Data Requirements

- *What are data requirements?*



### Data Collection

- *What occurs during data collection?*



# Data Engineering (Cont.)

**Dataset Flow Diagram:**

```

graph LR
    Datasets -- Data --> DataTable[Data Table]
    DataTable -- Data --> MatchingData[Matching Data]
    MatchingData -- Data --> SelectRows[Select Rows]
    SelectRows -- Data --> DataTable1[Data Table (1)]
  
```

**Data Table (Top Window):**

	car	buying	maint	doors	persons	lug_boot	safety
1	unacc	vhigh	vhigh	2	2	small	low
2	unacc	vhigh	vhigh	2	2	small	med
3	unacc	vhigh	vhigh	2	2	small	high
4	unacc	vhigh	vhigh	2	2	med	low
5	unacc	vhigh	vhigh	2	2	med	med
6	unacc	vhigh	vhigh	2	2	med	high
7	unacc	vhigh	vhigh	2	2	big	low
8	unacc	vhigh	vhigh	2	2	big	med
9	unacc	vhigh	vhigh	2	2	big	high

**Data Table (1) (Bottom Window):**

	car	buying	maint	doors	persons	lug_boot	safety
1	unacc	vhigh	high	2	2	small	low
2	unacc	vhigh	high	2	2	small	med
3	unacc	vhigh	high	2	2	small	high
4	unacc	vhigh	high	2	2	med	low
5	unacc	vhigh	high	2	2	med	med
6	unacc	vhigh	high	2	2	med	high
7	unacc	vhigh	high	2	2	big	low

**Datasets View:**

Title	Size	Instances	Variables	Target	Tags
Baker's Yeast	95.7 KB	186	81	categorical	biology
Bank Marketing	466.1 KB	4119	20	categorical	economy
Banking Crises	31.3 KB	211	73		time, economy
Bone Healing	11.6 KB	37	0	categorical	image analytics, biology
Bone marrow mononuclear cells with AML (sample)	334.0 KB	1000	1004	categorical	biology
Breast Cancer	18.4 KB	286	10	categorical	biology
Breast Cancer Wisconsin	34.9 KB	683	10	categorical	biology
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	categorical	biology
COMPAS Analysis	2.7 MB	7214	52	categorical	criminal justice, fairness
Car Evaluation	50.7 KB	1728	6	categorical	synthetic
Climate of European cities	4.4 KB	41	16		geography
Conferences	2.3 KB	42	5		
Congressional Voting Records	17.8 KB	435	17	categorical	politics
Course Grades	9.2 KB	16	7		synthetic, education
Cyber Security Breaches	225.0 KB	1055	10		security, time, geo

**Description:**  
**Car Evaluation (1999), from UCI ML Repository**  
 This is a synthetic data set derived from a simple hierarchical decision model to demonstrate decision support system DEX (see Bohanec and Rajkovic). The decision model included six attributes, including buying and maintenance price, the number of passengers, size of the luggage boot, and evaluated the utility of the car from a buyer's perspective. All attributes were discrete, having from three to four values. The data set provides car's utility for all possible combinations of attribute values. The data set was originally created to showcase the ability of machine learning by function decomposition to recreate the hierarchy of the decision model.

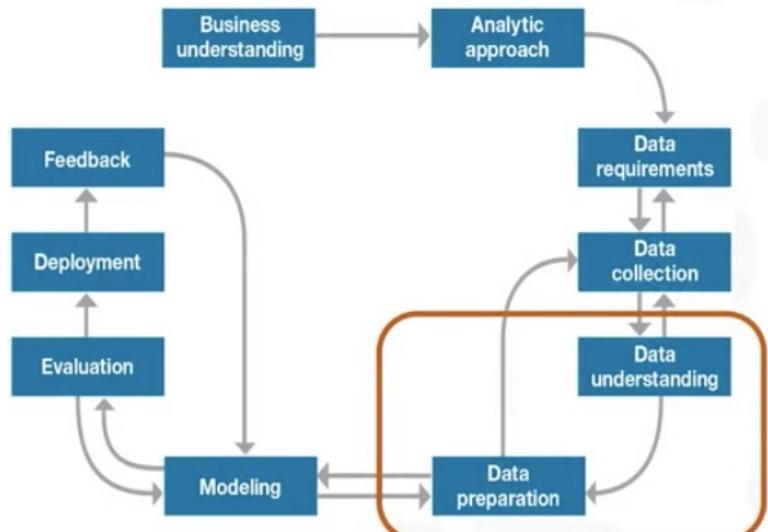
**References:**  
 Bohanec M, Rajkovic V (1988) Knowledge acquisition and explanation for multi-attribute decision making. In 8th International Workshop on Expert Systems and their Applications, Avignon, France, pages 59–78.  
 Zupan B, Bohanec M, Demsar J, Bratko I (1999) Learning by discovering concept hierarchies. Artificial Intelligence 109: 211–242.

- Interactive
- Visual
- Simplified



# Data Understanding

## From Understanding to Preparation



### Data understanding

- What does it mean to “prepare” or “clean” data?



### Data preparation

- What are ways in which data is prepared?



# Data Engineering (Cont.)

The figure illustrates a data engineering workflow using three main components: Data Info, Feature Statistics, and Data Sampler.

**Data Info:** Shows details about the "car" dataset, including 1728 rows, 7 columns, 6 categorical features, and 4 classes. It also displays a summary of additional attributes.

**Feature Statistics:** A table showing feature distributions and statistics for each column. The columns are Name, Distribution, Mean, Mode, Median, Dispersion, and Min.

Name	Distribution	Mean	Mode	Median	Dispersion	Min.
buying	Stacked bar chart showing distribution across categories (blue, red, green, orange).		high			1.39
maint	Stacked bar chart showing distribution across categories (blue, red, green, orange).		high			1.39
doors	Stacked bar chart showing distribution across categories (blue, red, green, orange).		2			1.39
persons	Stacked bar chart showing distribution across categories (blue, red, green, orange).		2			1.1
lug_boot	Stacked bar chart showing distribution across categories (blue, red, green, orange).		big			1.1
safety	Stacked bar chart showing distribution across categories (blue, red, green, orange).		high			1.1
car	Stacked bar chart showing distribution across categories (blue, red, green, orange). The bars are much shorter than the other features, indicating lower dispersion.	unacc				0.836

**Data Sampler:** A configuration panel for sampling data from the dataset. It includes options for Sampling Type (Fixed proportion of data, Fixed sample size, Cross validation, Bootstrap), Options (Replicable (deterministic) sampling, Stratify sample (when possible)), and a Sample Data button.



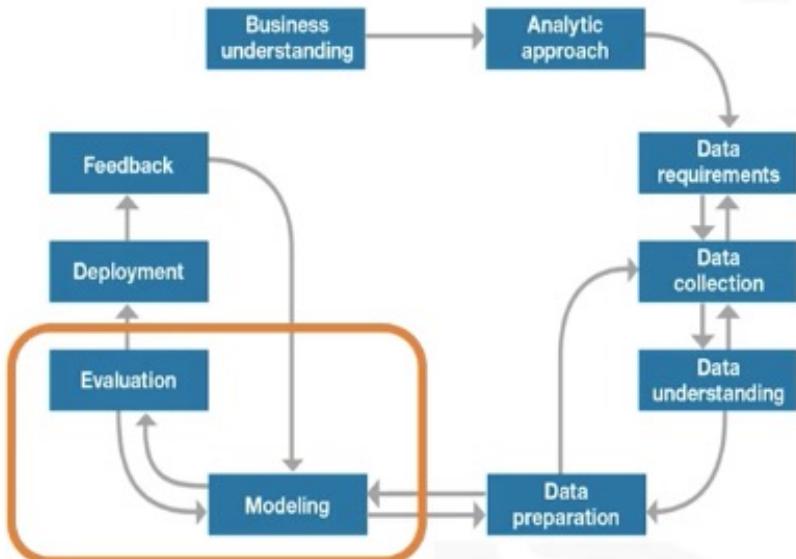
# Data Engineering (Cont.)

The image shows the KNIME Data Engineering interface with four main panels:

- Transform Panel:** On the left, it contains various data manipulation icons grouped under categories like Data Sampler, Select Columns, Merge Data, Unique, Apply Domain, Discretize, Formula, etc.
- Preprocess Panel:** Shows the "Preprocessors" list with "Impute Missing Values" selected. It includes options for Average/Most frequent, Replace with random value, or Remove rows with missing values. The "Output" section has a checked "Send Automatically" option.
- Formula Panel:** Displays variable definitions for "size". A formula is defined: "size := "low" if sepal\_length < 6 else "mid" if sepal\_length < 7 else "high"".
- Create Instance Panel:** Allows creating instances for variables like CRIM, ZN, INDUS, CHAS, NOX, RM, and AGE. It includes sliders for continuous values and dropdowns for categorical values. A checkbox for "Append this instance to input data" is checked.

# Modeling

## From Modeling to Evaluation



### Modeling

- In what way can the data be visualized to get to the answer that is required?*



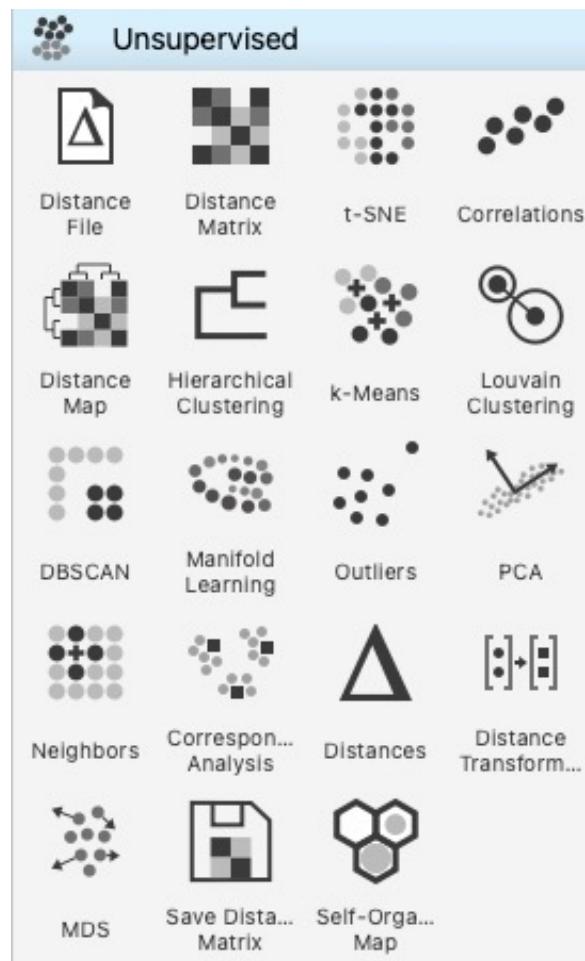
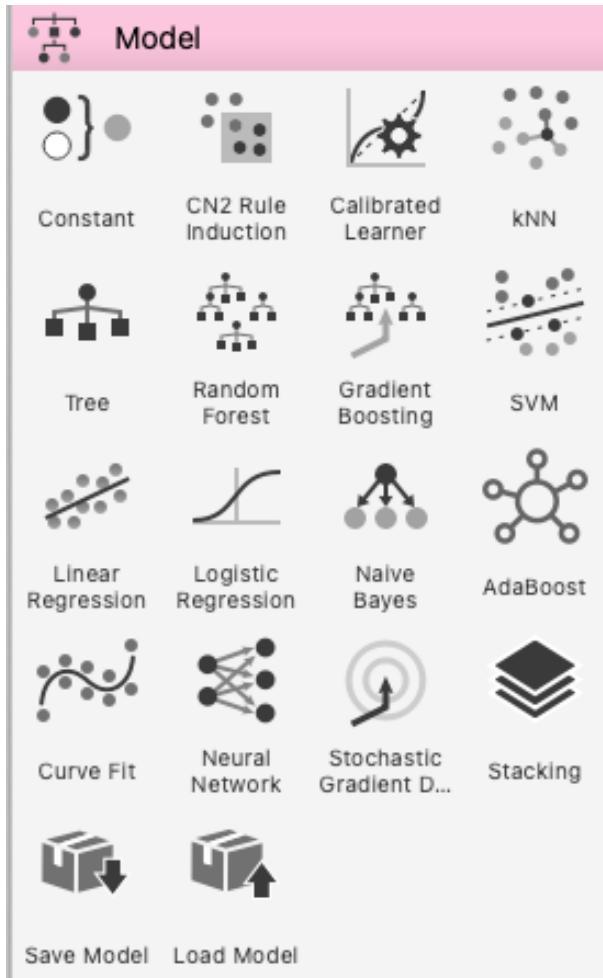
### Evaluation

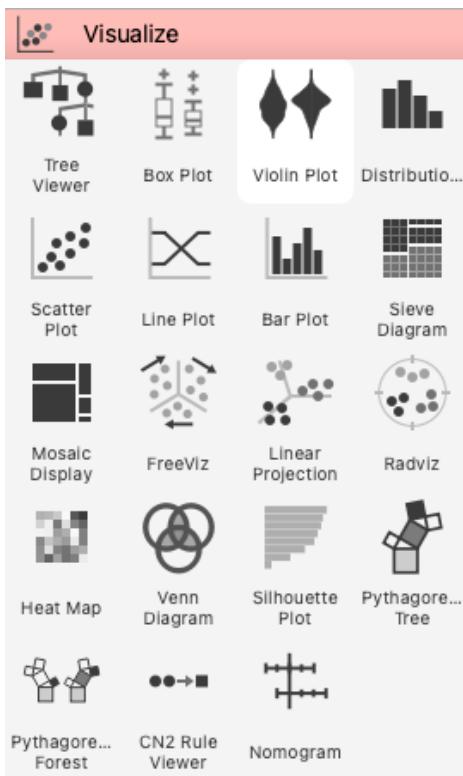
- Does the model used really answer the initial question or does it need to be adjusted?*



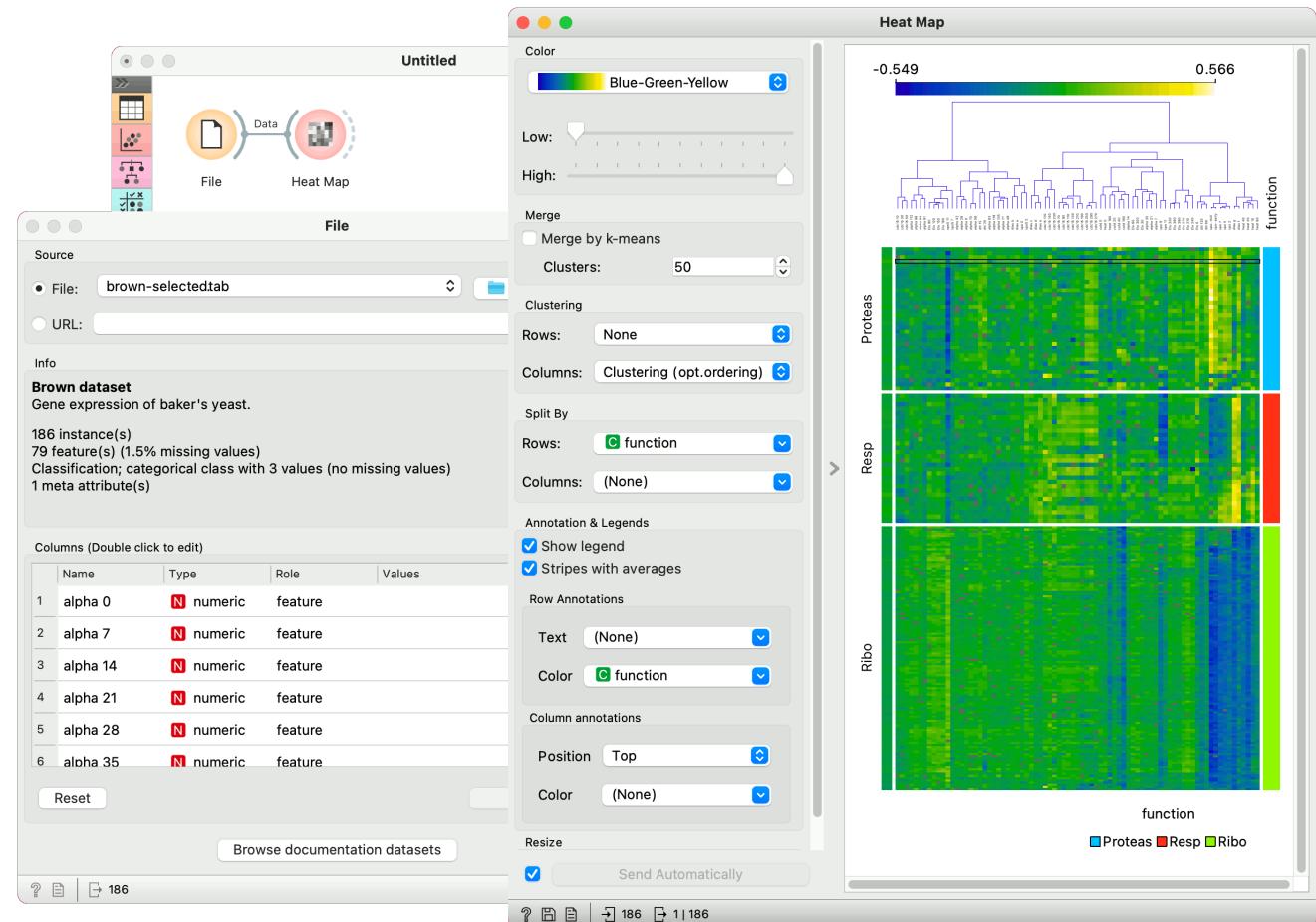
# Modeling in Orange3

## Supervised      UnSupervised



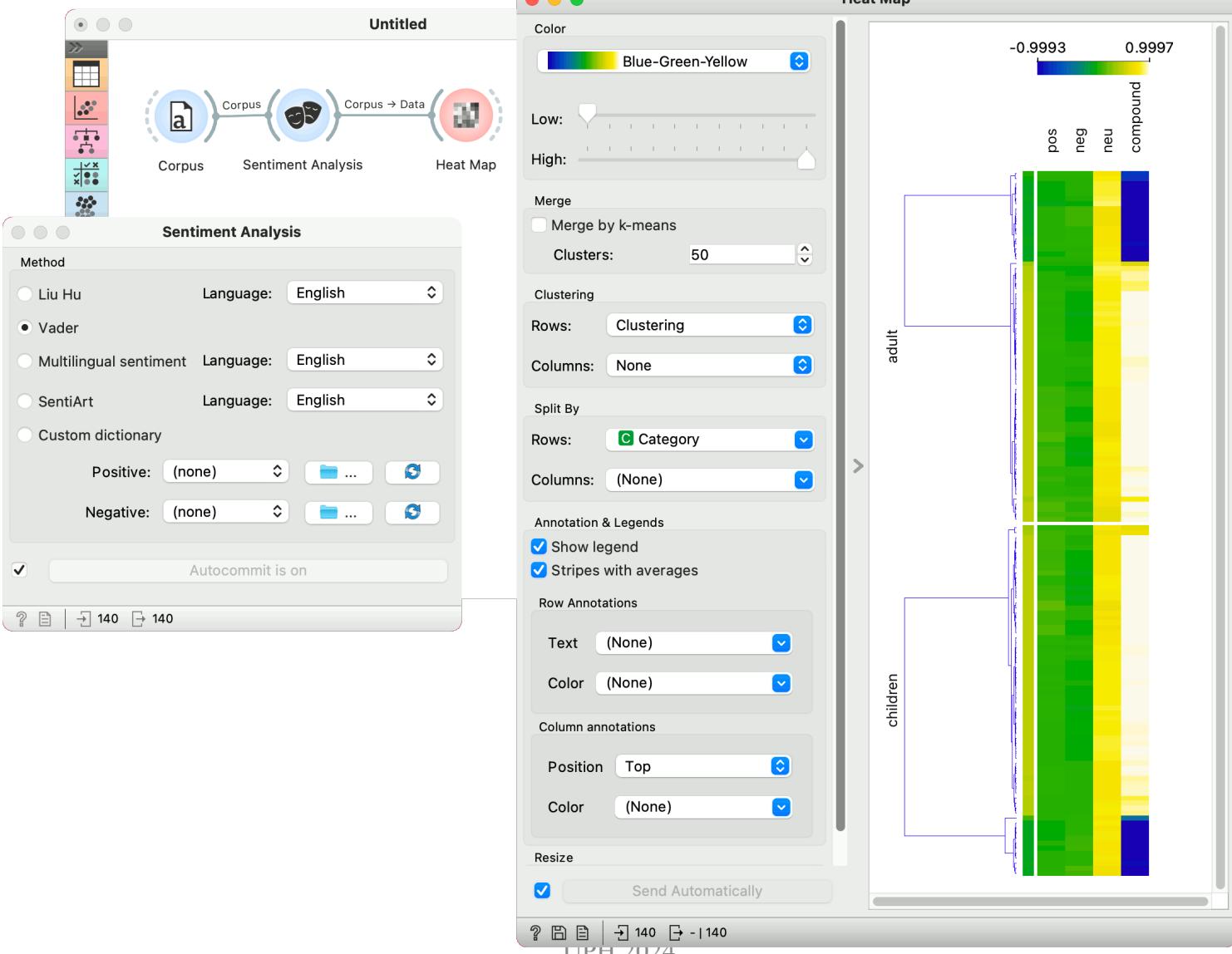


# Visualization



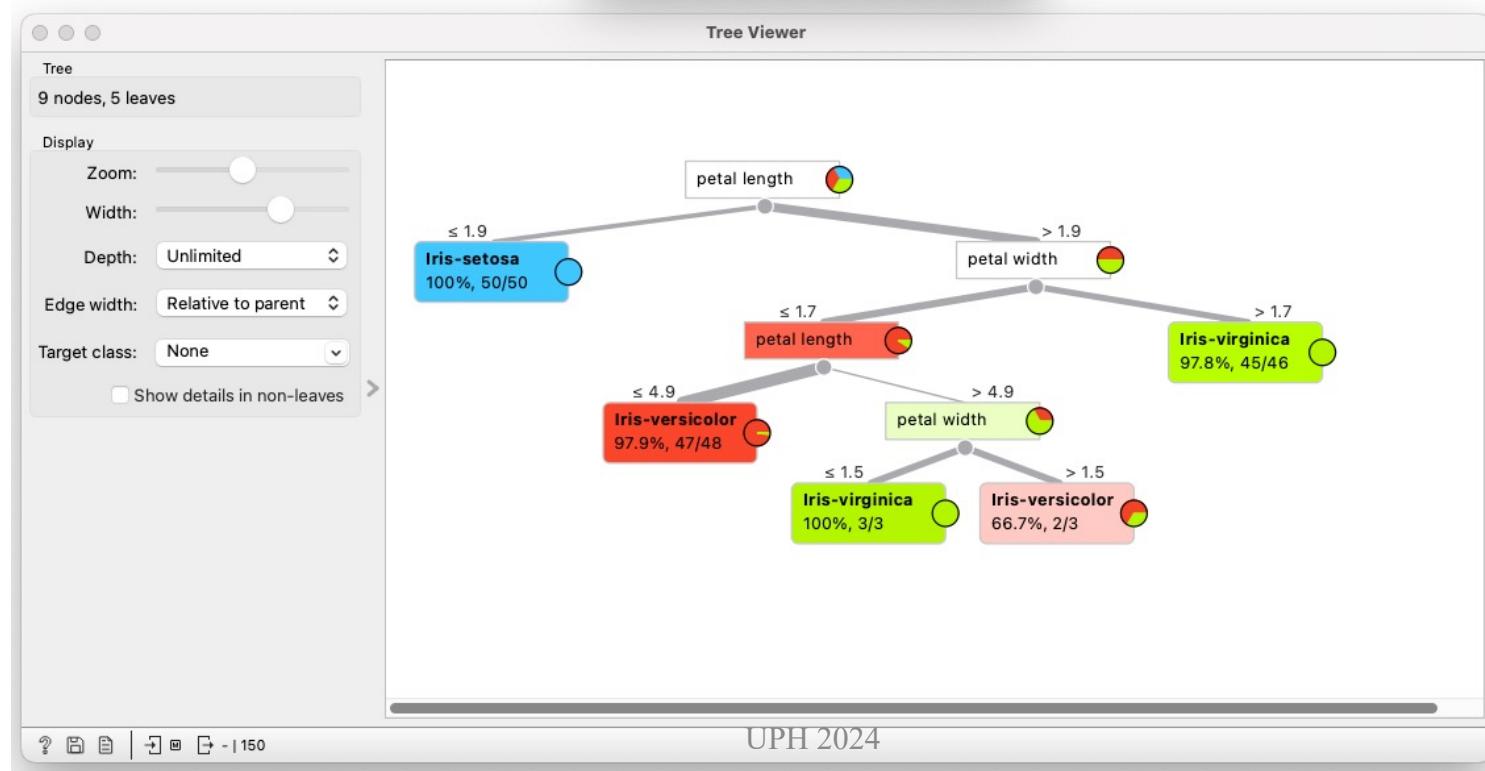


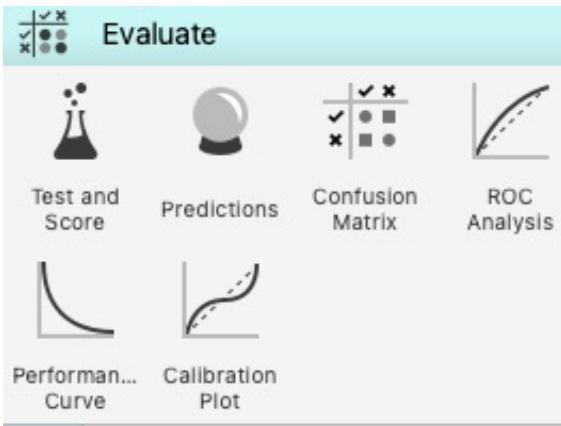
# Visualization



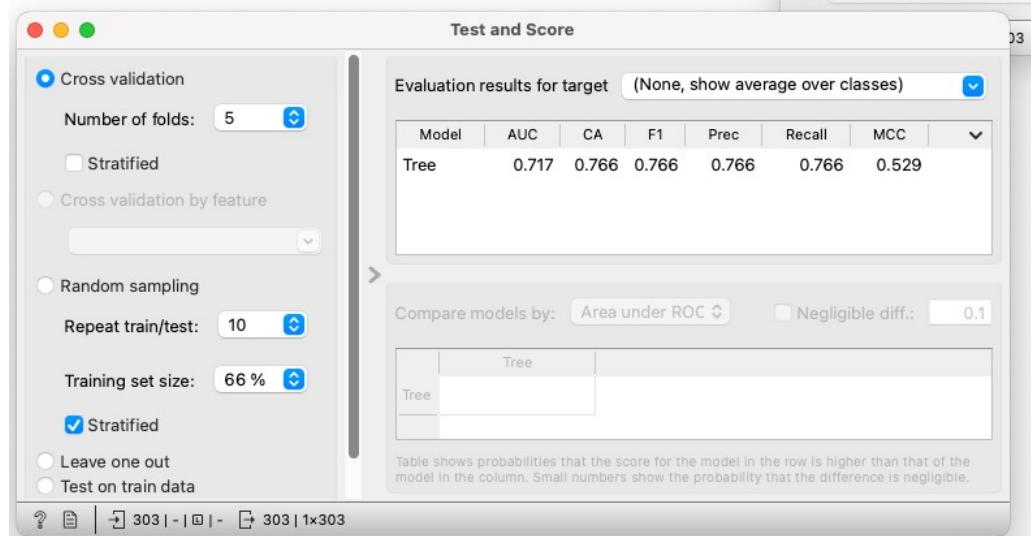
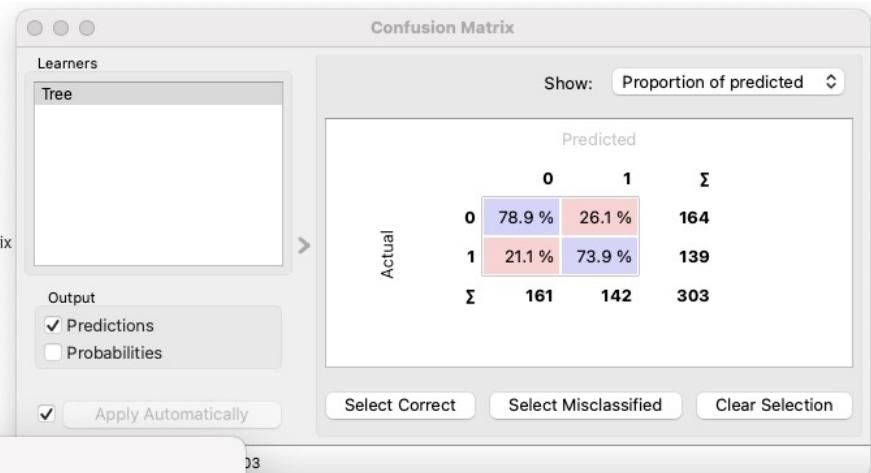
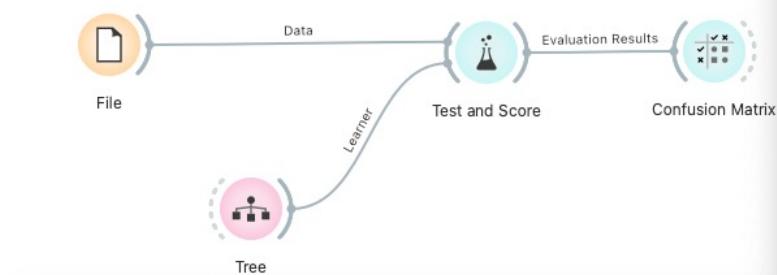


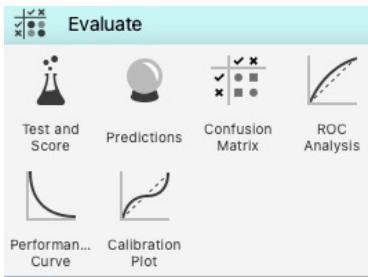
# Tree



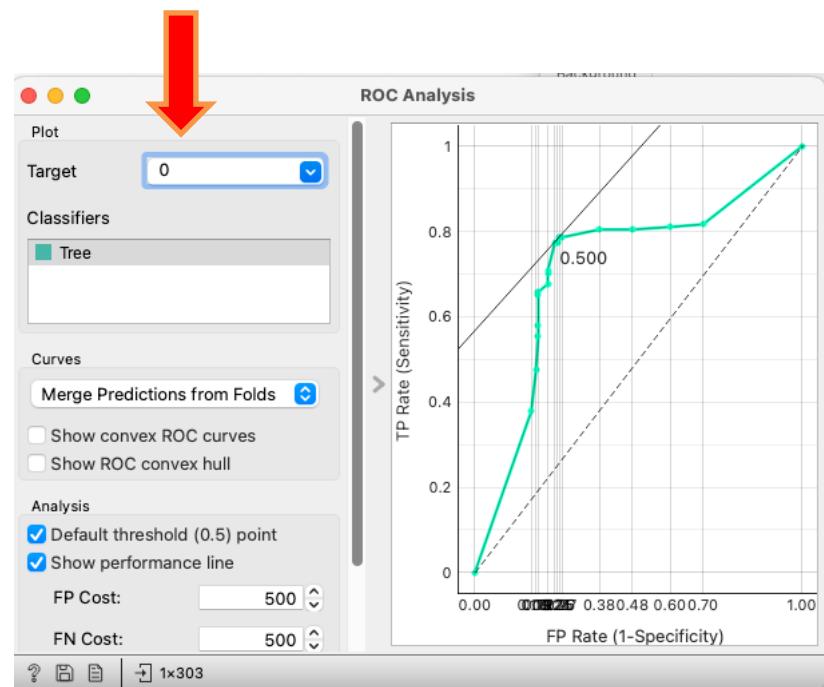
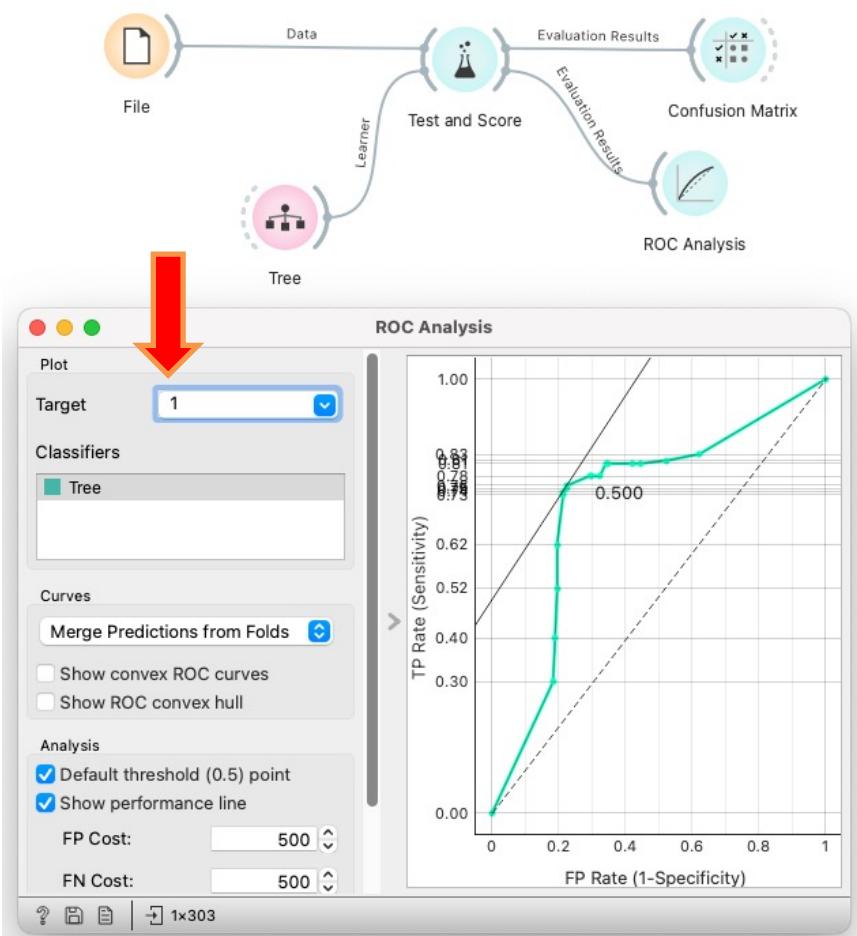


# Evaluate





# Evaluate

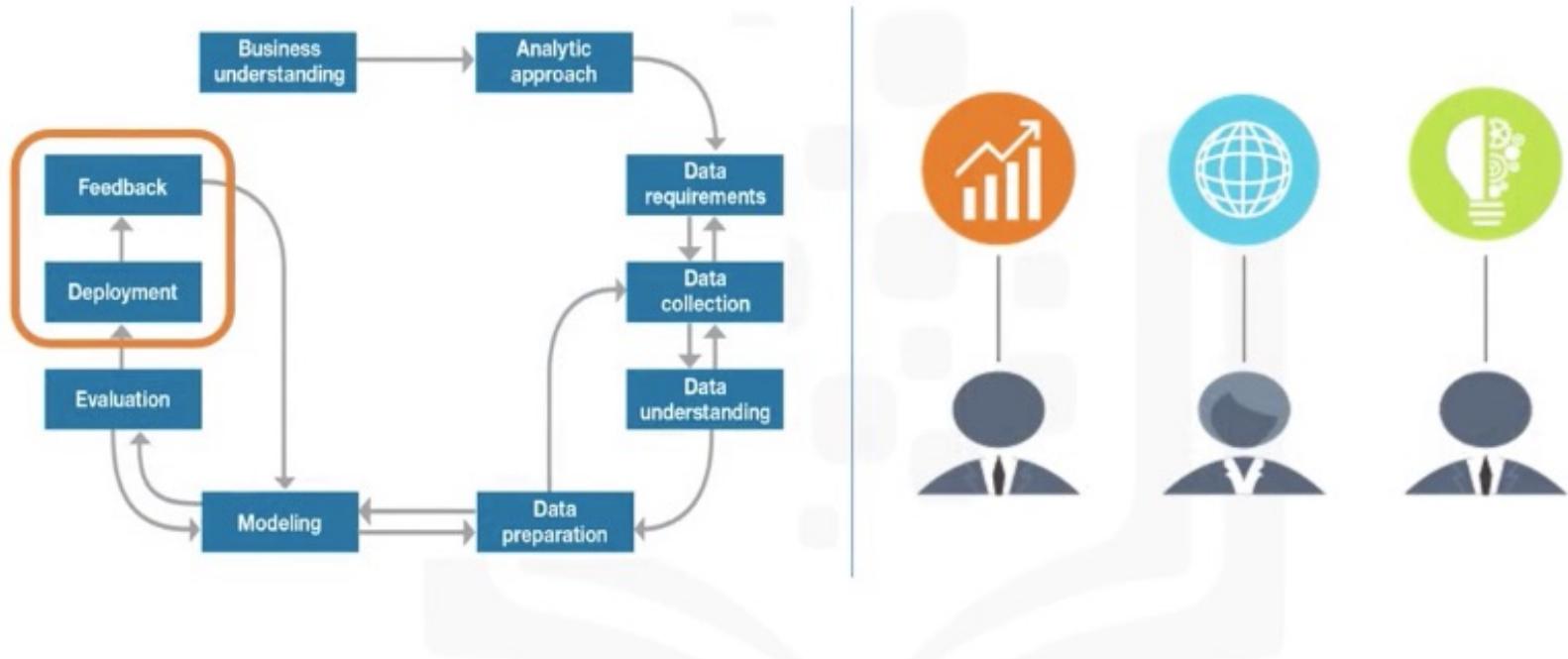


Interactive



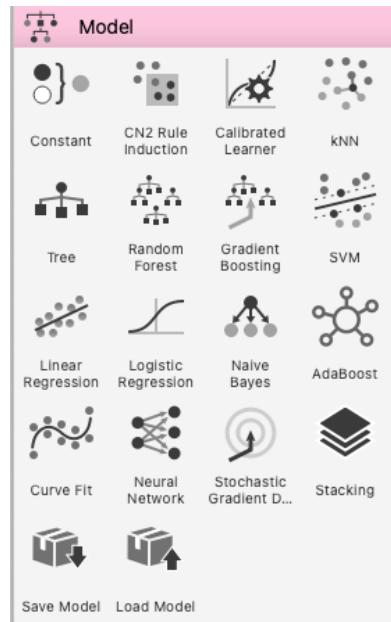
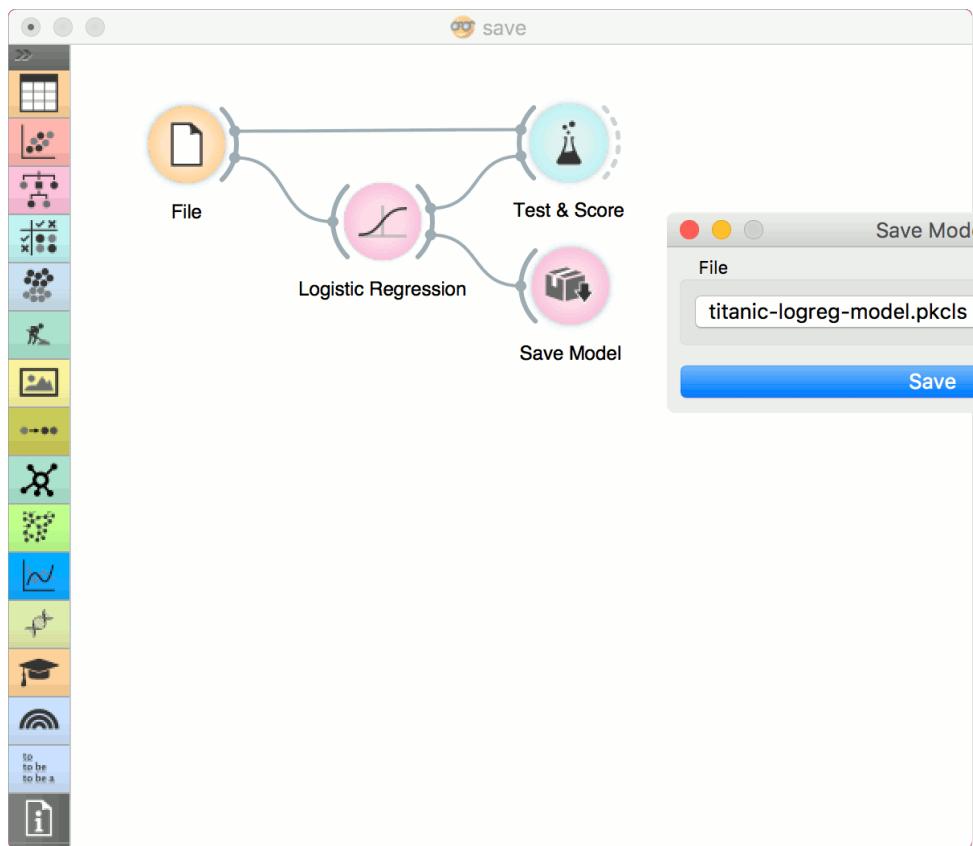
# Deployment

## From Deployment to Feedback



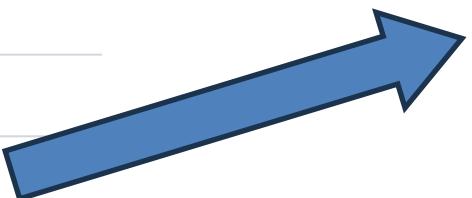
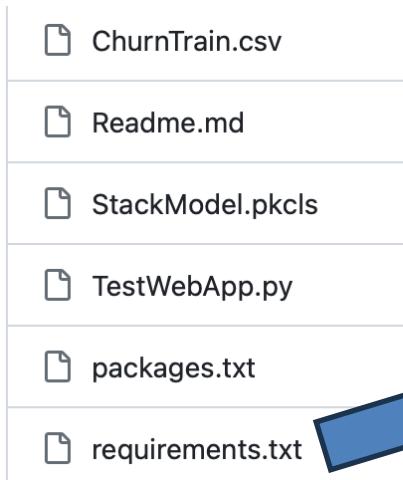
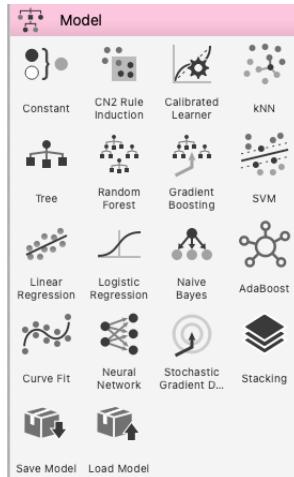
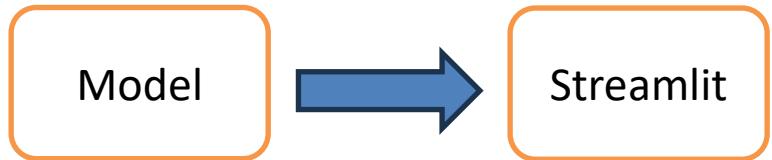


# Deployment





# Deployment



```
1 streamlit==1.1.0  
2 pandas==1.3.4  
3 numpy==1.20.3  
4 Orange3==3.30.2  
5 xgboost==1.5.0
```

[This Photo](#) by Unknown Author is licensed under [CC BY-NC](#)

Example:

<https://github.com/TonciG/Telecom-Customer-Churn---Web-App>



# Any Question ?