PYCON SETTE - 15/04/16

ERIC BONFADINI
@ERICBONFADINI

INTRODUCTION TO ORANGE DATA MINING

AGENDA

- About me
- What is data mining
- Orange Data Mining
- Versions
- Demo: Canvas vs Scripting
- Resources
- Q&A

ABOUT ME

- Eric Bonfadini (@ericbonfadini)
- CTO @ Deus Technology
- Numpy, Pandas & Matplotlib user, interested in data

COMPUTERS HAVE PROMISED US A FOUNTAIN OF WISDOM BUT DELIVERED A FLOOD OF DATA

W. J. Frawley et al. (1991)

WHAT IS DATA MINING

- Involves: databases, statistics, high performance computing, machine learning, visualization, mathematics, etc.
- Goal: analyzing data and converting it into useful information
- Solution to common problems: classification, regression, clustering, etc.

WHAT IS DATA MINING

- Examples:
  - Given outlook, temperature, humidity, and windy as features, decide if it’s possible to play tennis or not
  - Given attributes like age, sex, cholesterol level, smoker, heart rate, etc decide if the patient has a heart disease
  - Analyse customers behaviour in order to find tastes and recommend some articles

WHAT IS DATA MINING

ORANGE DATA MINING

- Developed by Bioinformatics Lab at University of Ljubljana, Slovenia, in collaboration with open source community
- Provides data visualisation and data analysis for novice and expert, through interactive workflows
- Large widget toolbox and several add-ons
- Possibility to use it programmatically o via GUI (Orange canvas, PyQT)
- Open source project (GPL license)

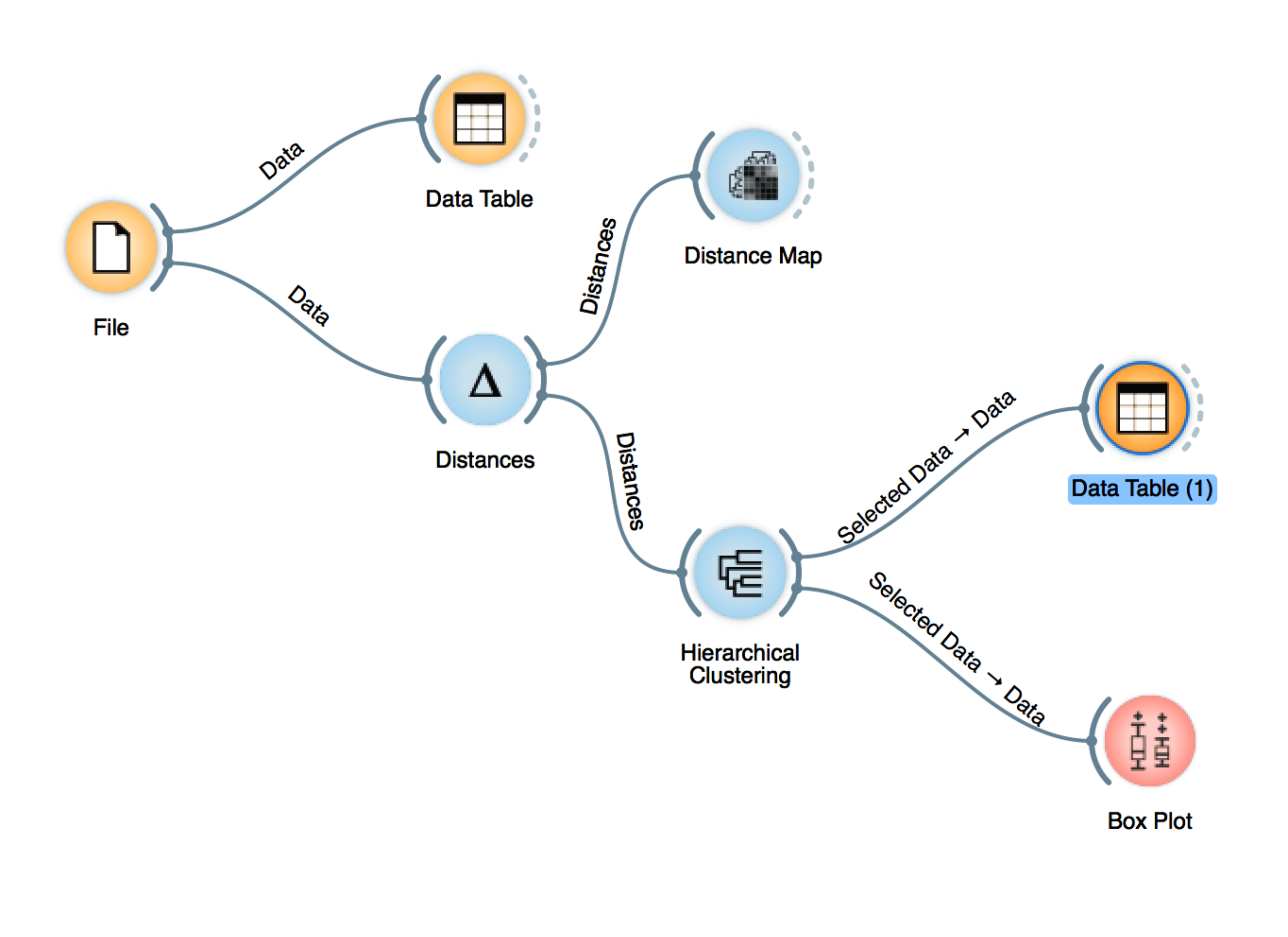
VERSIONS

- Orange 2 (https://github.com/biolab/orange)
  - Legacy version, currently marked as stable
  - Installation from source or binaries available for Windows/MacOS
  - ML proprietary algorithms written in C++, with wrappers in Python 2

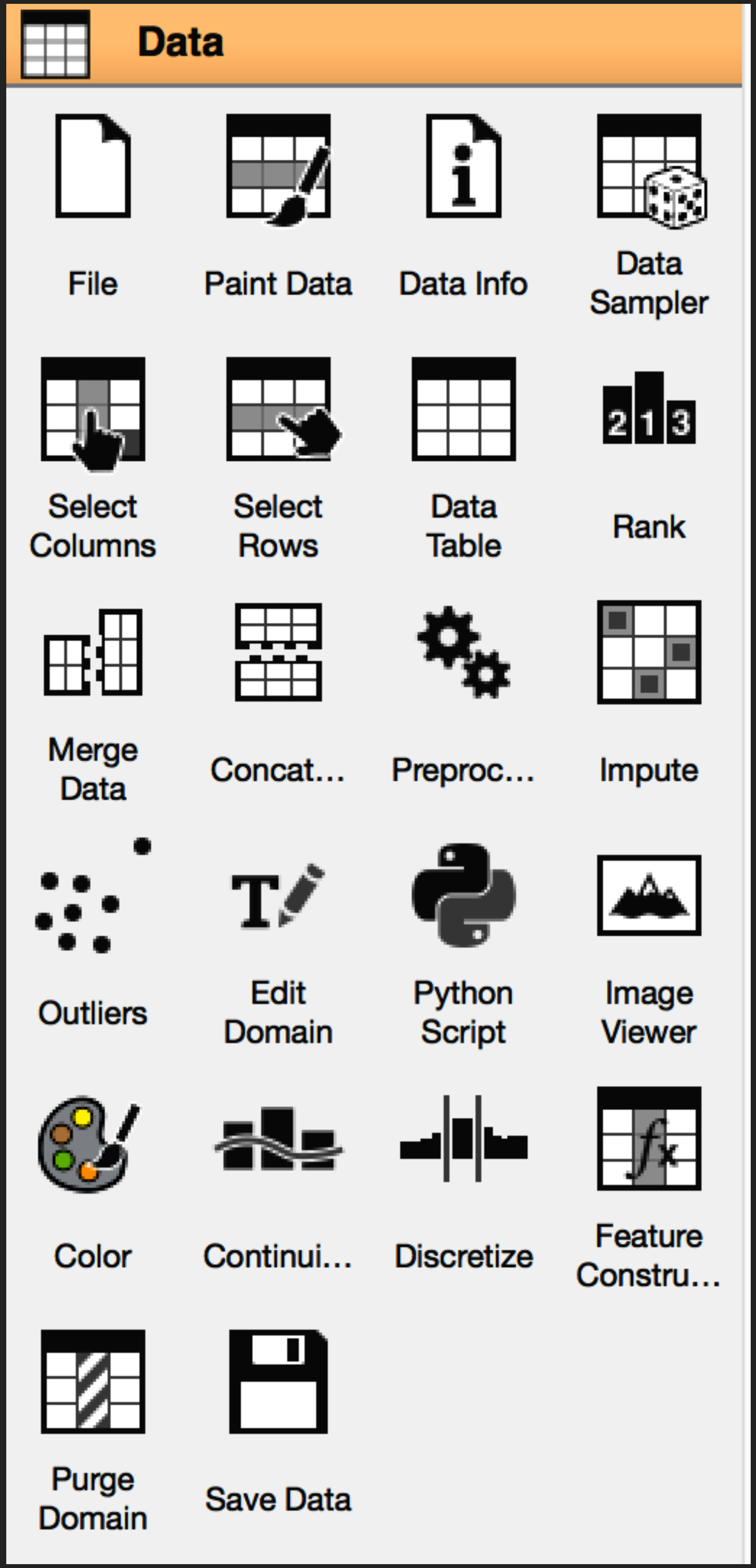
VERSIONS

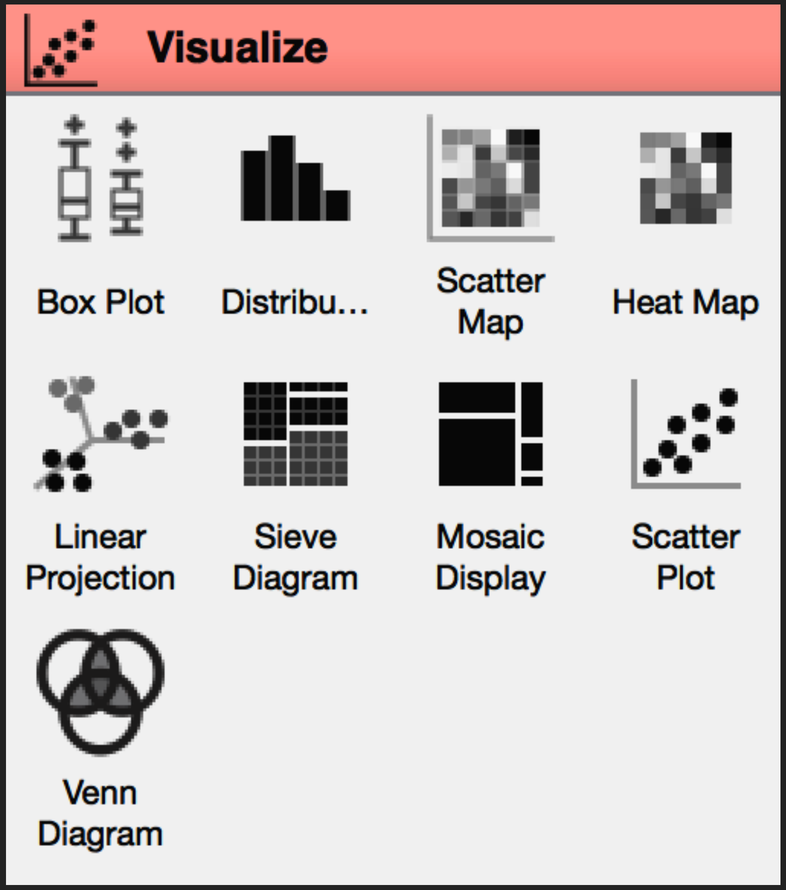
- Orange 3 (https://github.com/biolab/orange3)
  - Newer version, currently marked as development
  - Installation from source or binaries available for Windows/MacOS
  - Written completely in Python 3, ML algorithms are mostly wrappers of scikit-learn ones
  - 3 developers full time + ~10 part time + community contributions

CANVAS

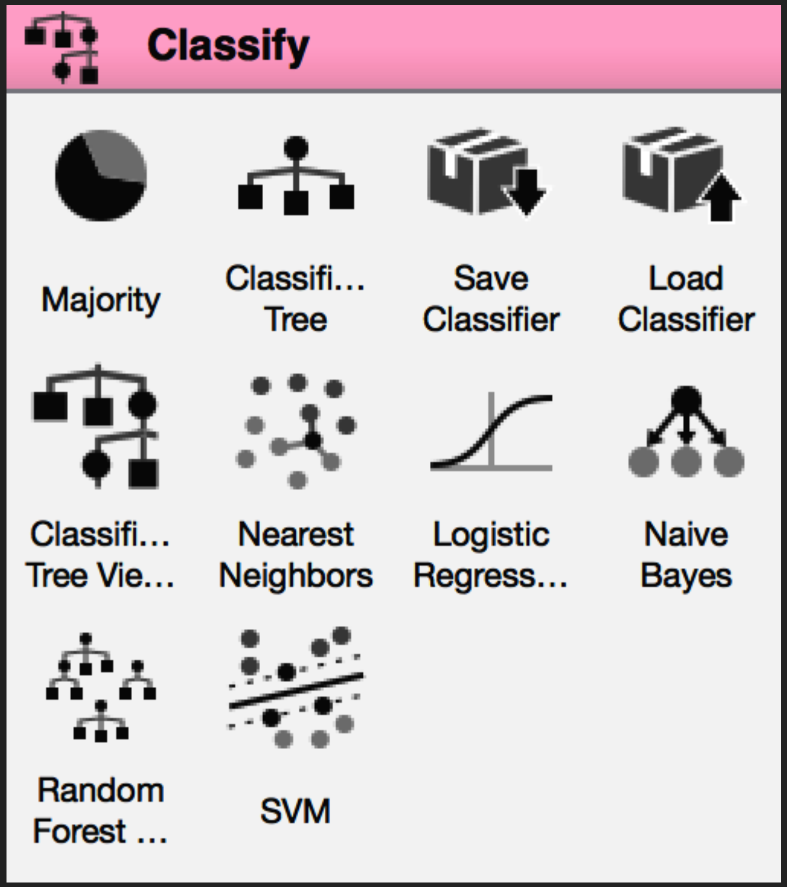

File
Data
Data Table
Data
Distances
Distances
Distance Map
Distances
Hierarchical Clustering
Selected Data → Data
Data Table (1)
Selected Data → Data
Box Plot

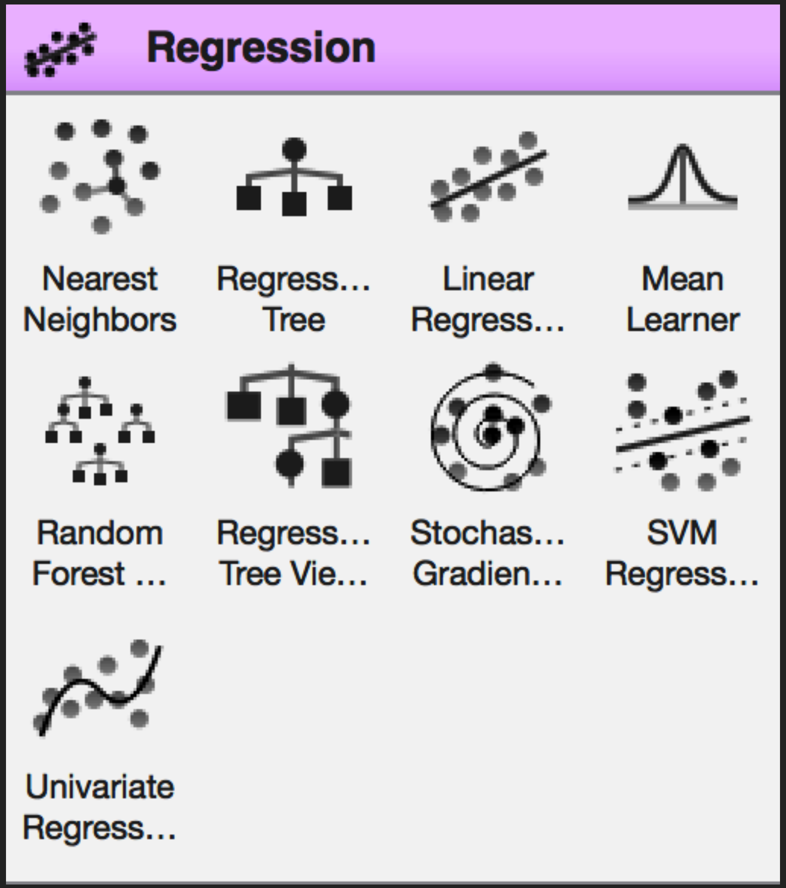
CANVAS

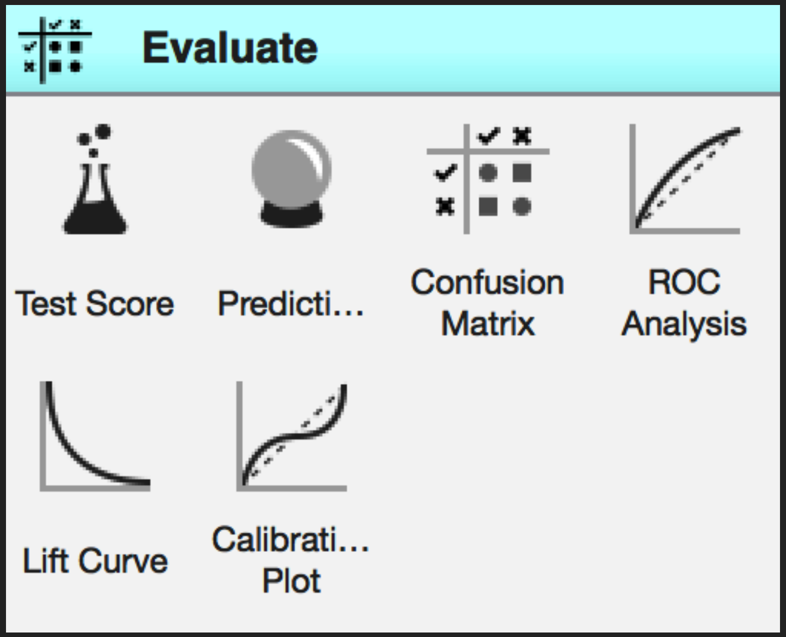

Data
File
Paint Data
Data Info
Data Sampler
Select Columns
Select Rows
Data Table
Rank
Merge Data
Concat…
Preproc…
Impute
Outliers
Edit Domain
Python Script
Image Viewer
Color
Continui…
Discretize
Feature Constru…
Purge Domain
Save Data

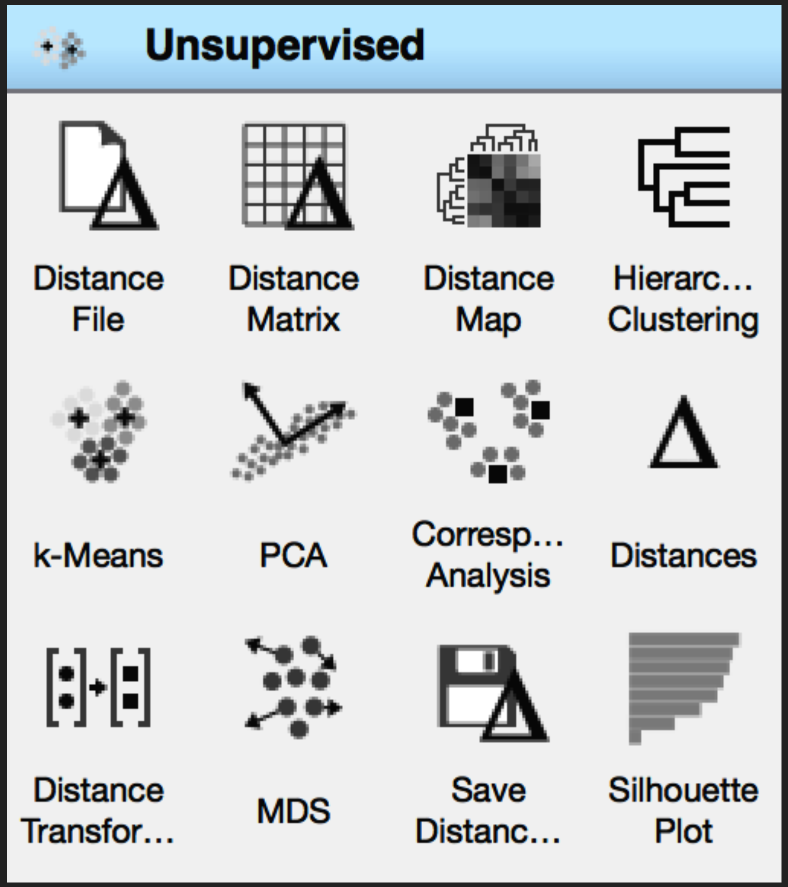

Visualize
Box Plot
Distribu…
Scatter Map
Heat Map
Linear Projection
Sieve Diagram
Mosaic Display
Scatter Plot
Venn Diagram

CANVAS


Classify
Majority
Classifi… Tree
Save Classifier
Load Classifier
Classifi… Tree Vie…
Nearest Neighbors
Logistic Regress…
Naive Bayes
Random Forest …
SVM


Regression
Nearest Neighbors
Regress… Tree
Linear Regress…
Mean Learner
Random Forest …
Regress… Tree Vie…
Stochas… Gradien…
SVM Regress…
Univariate Regress…


Evaluate
Test Score
Predicti…
Confusion Matrix
ROC Analysis
Lift Curve
Calibrati… Plot


Unsupervised
Distance File
Distance Matrix
Distance Map
Hierarc… Clustering
k-Means
PCA
Corresp… Analysis
Distances
Distance Transfor…
MDS
Save Distanc…
Silhouette Plot

DEMO: CANVAS VS SCRIPTING

- Iris: a classic multivariate data set introduced by Ronald Fisher in 1936
- 150 samples from three species of Iris (Iris setosa, Iris virginica and Iris versicolor)
- Four features: the length and the width of the sepals and petals, in centimetres

SHOW ME THE CODE!

RESOURCES

- Scripting reference (http://docs.orange.biolab.si/reference/rst/)
- Tutorial (http://docs.orange.biolab.si/3/data-mining-library/)
- Blog (http://blog.biolab.si/)
- YouTube channel (https://www.youtube.com/channel/UClKKWBe2SCAEyv7ZNGhIe4g)
- Twitter (@OrangeDataMiner)

THANK YOU!