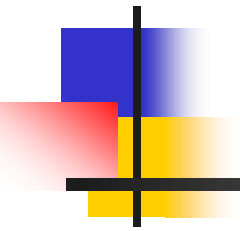


Data Mining Tools

Petr Berka
(berka@vse.cz)

University of Economics Prague

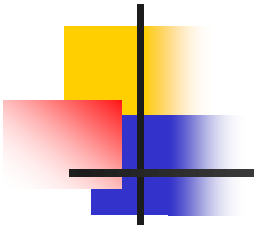
Seminar S1

Data Mining Tools and measurement data analysis

Vysoká škola manažmentu v Trenčíne
International Workshop on Knowledge Management
IWKM 2018

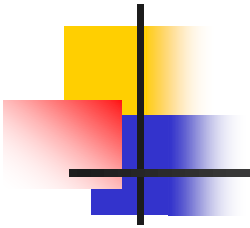
October, 18 – 19
Bratislava 2018

Data Mining, Data Analytics, Data Science



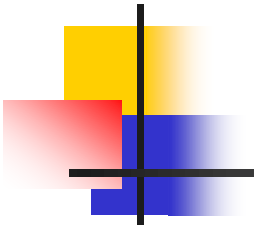
- **Data mining** is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
- **Data analytics** is the discovery, interpretation, and communication of meaningful patterns in data.
- **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining.

(wikipedia)



Types of Data Mining Tools

- Data mining suites
- Programming tools
- Cloud solutions

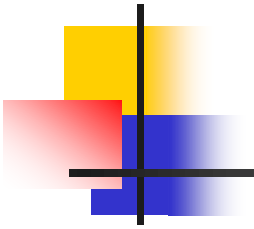


Data Mining Suites

Stand-alone tools that implement a number of data mining and data pre-processing algorithms

- Commercial
- Free/Open Source

www.kdnuggets.com lists about 90 commercial and about 30 free tools



Data Mining Suites Features

Commercial

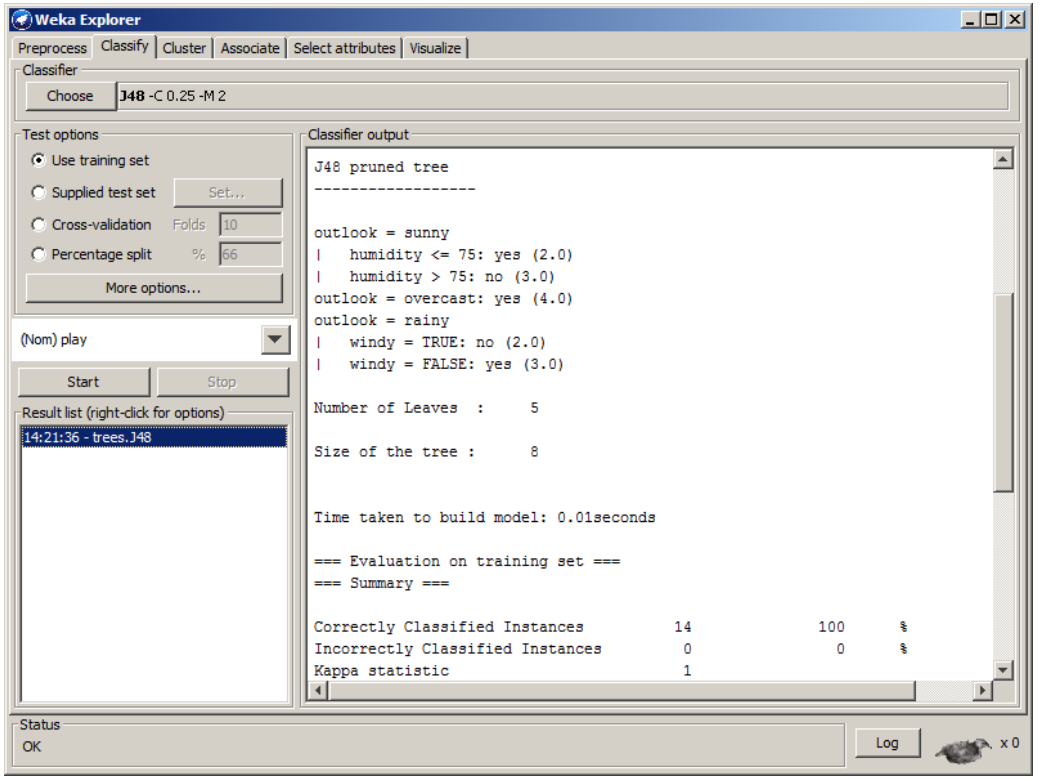
- No control of implementation
- Full control of installation
- Full control of data
- Full access to support

Free

- No/limited control of implementation
- Full control of installation
- Full control of data
- No/limited access to support

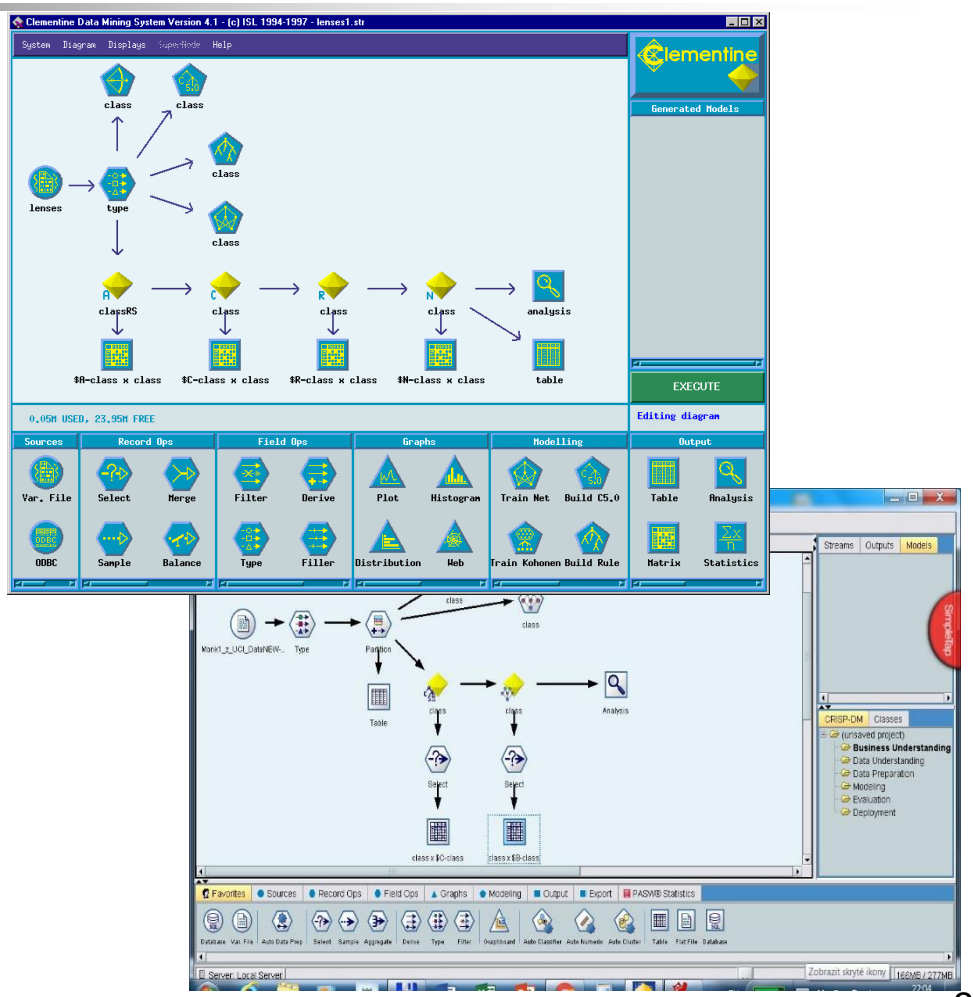
Weka (www.cs.waikato.ac.nz/ml/weka/)

- A collection of machine learning algorithms for data mining tasks from the University of Waikato, NZ. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.
- Development started in 1997.



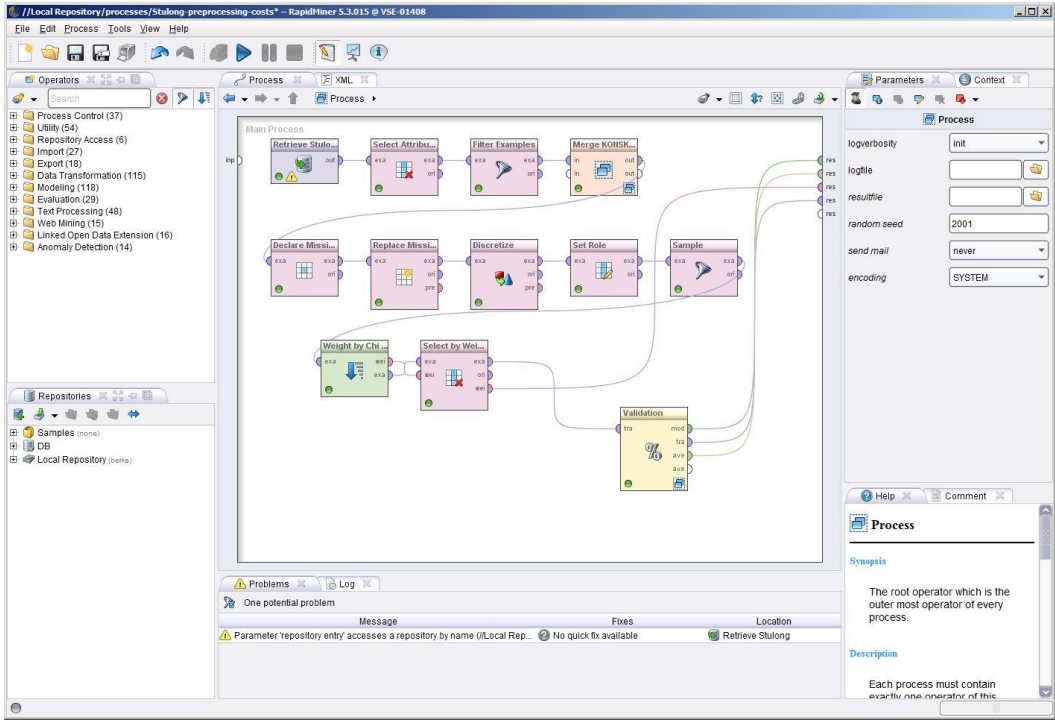
IBM SPSS Modeler (www.ibm.com)

- IBM SPSS Modeler, originally named Clementine (developed by ISL in collaboration with Sussex University in 1994), is a commercial DM system. Clementine introduces a visual interface that allows to use statistical and DM algorithms in an intuitive way without programming.



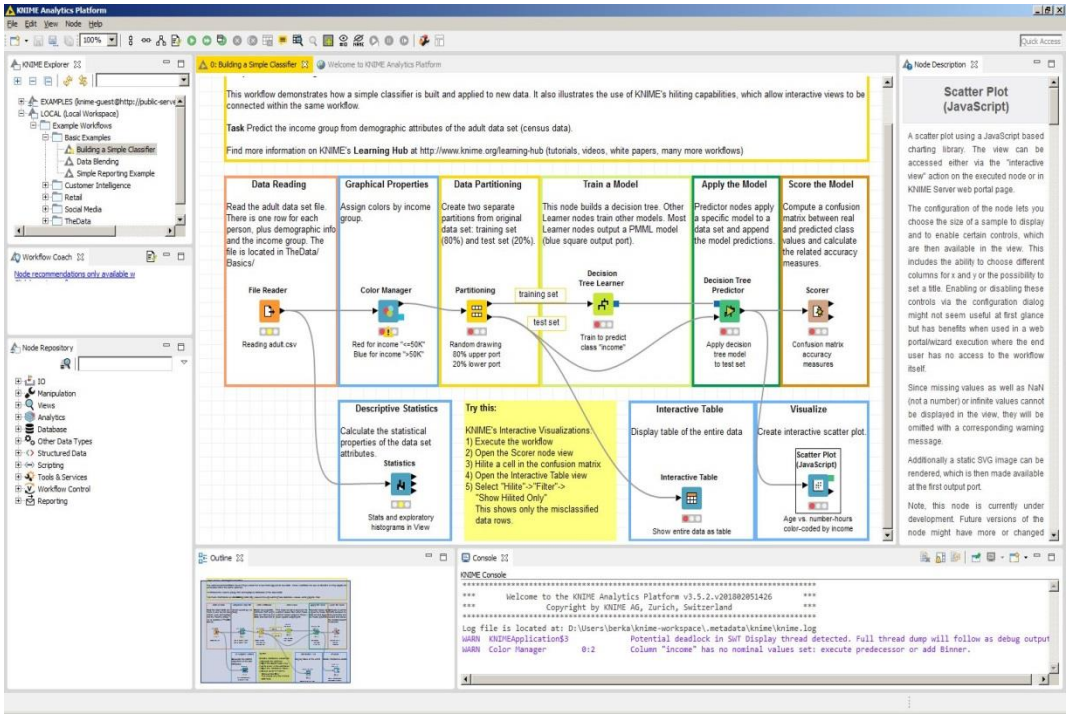
Rapid Miner (www.rapidminer.com)

- Data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics.
- Free version limited to 10 000 rows in data table.
- The development of the system started in 2001 at the University of Dortmund (under the name YALE).



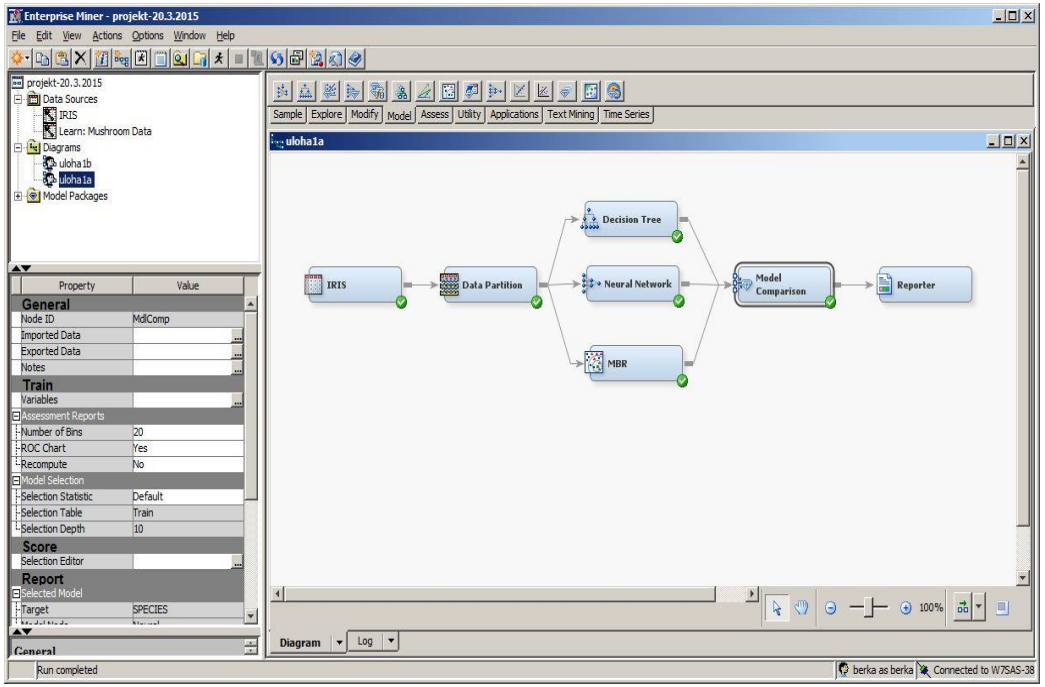
KNIME (www.knime.com)

- KNIME the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform.
- The development of the system started in 2004 at University of Konstanz.

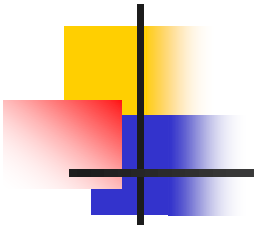


SAS Enterprise Miner (www.sas.com)

- SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis of them.
- The original focus of SAS was on statistical data analysis (SAS Institute started as a project at North Carolina State University). SAS Enterprise Miner, as a stand-alone data mining tool, was released in 1999.



Gartner Magic Quadrant 2018



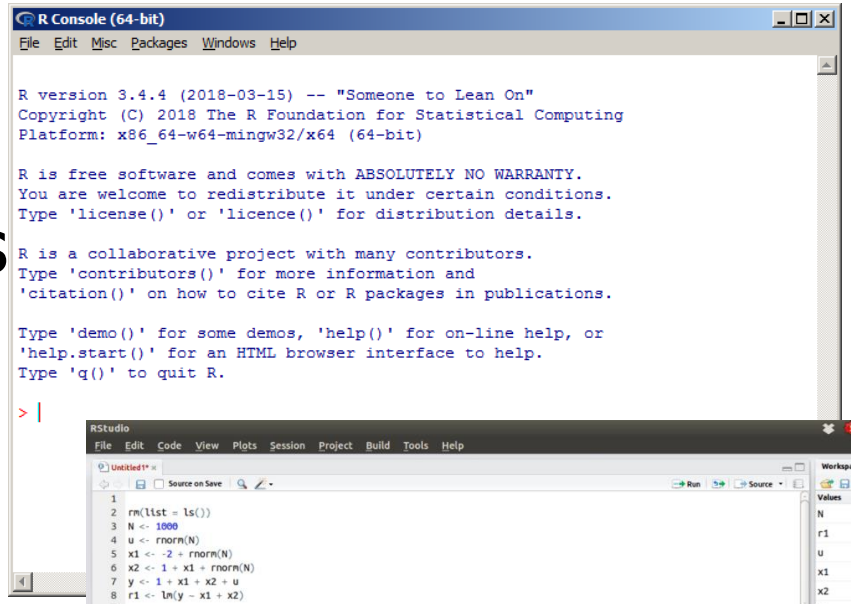
Programming Tools Features

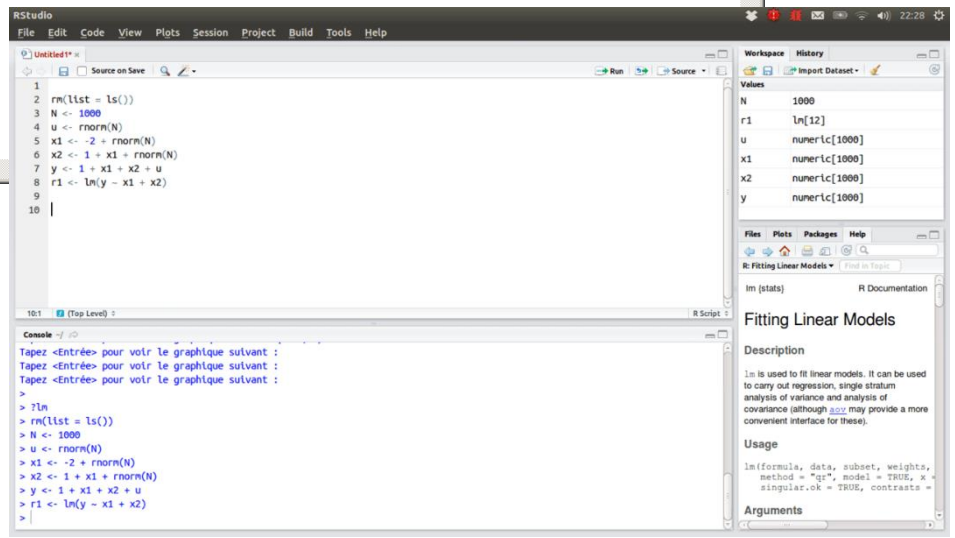
Do it (almost) yourself solutions

- Full control of implementation
- Full control of installation
- Full control of data
- No/limited access to support

(www.r-project.org)

- R is a programming environment for data analysis and graphics (widely used for statistical data analysis).
- RStudio makes R easier to use. It includes a code editor, debugging and visualization tools.

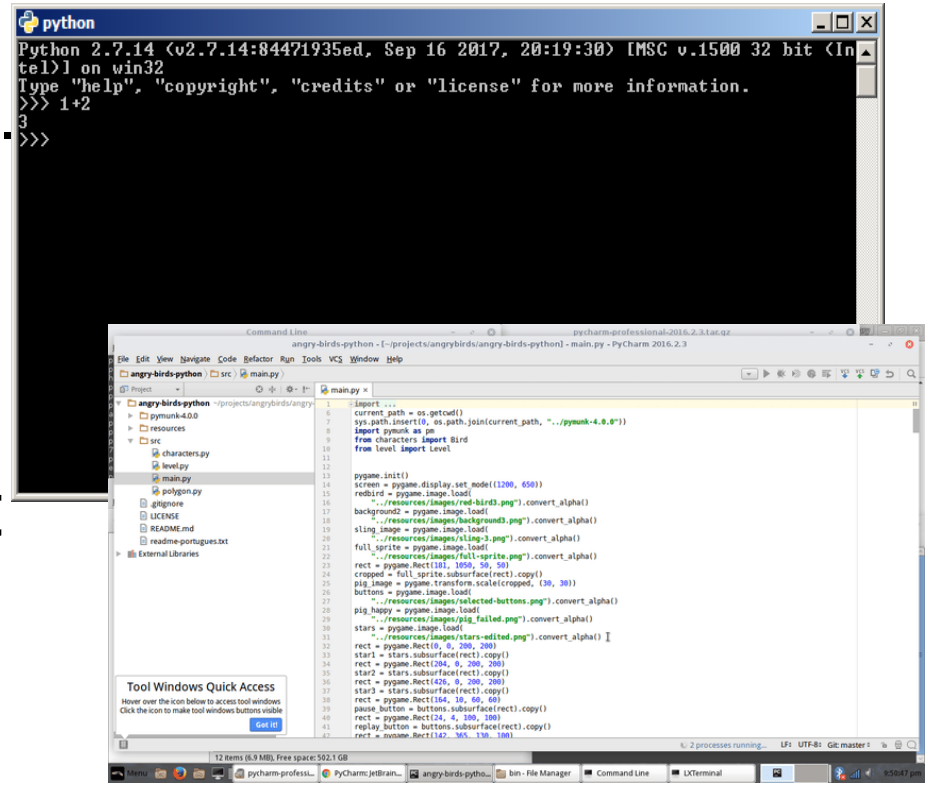




Rstudio

python™ (www.python.org) …

- Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language.
- A number of IDE's (Integrated Development Environment) exist to support programming in Python.



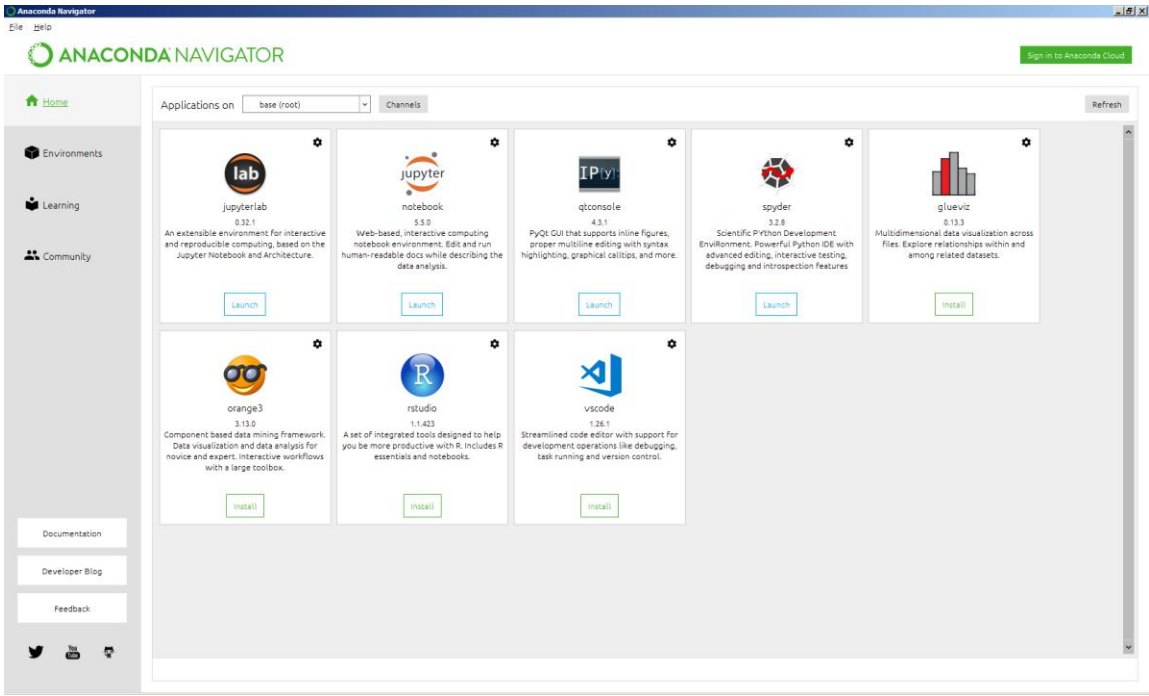
pycharm

… + libraries

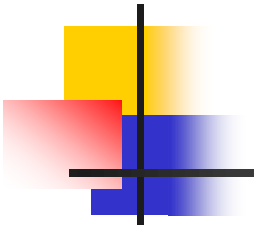
- Scikit-learn (scikit-learn.org) is a free software machine learning library.
- TensorFlow (www.tensorflow.org) is an open source software library for high performance numerical computation including machine learning (neural networks).
- XGBoost (http://dmlc.cs.washington.edu/xgboost.html) is an open source software library which provides a gradient boosting framework.
- Keras (keras.io) is an open source neural network library.

ANACONDA (www.anaconda.com)

- A free and open source distribution of the Python and R programming languages for data science and machine learning related applications (industry standard for developing, testing and training on a single machine).



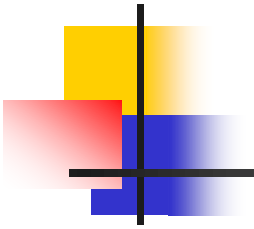
Anaconda Navigator



Cloud Solutions

Cloud computing refers to both the applications delivered as services over the Internet and the hardware and system software in the data centers that provide those services (Armbrust et al., 2009)

- PaaS – platform as a service
- IaaS – infrastructure as a service
- SaaS – software as a service
  - MLaaS – machine learning as a service



Cloud Solution Features

~ Systems

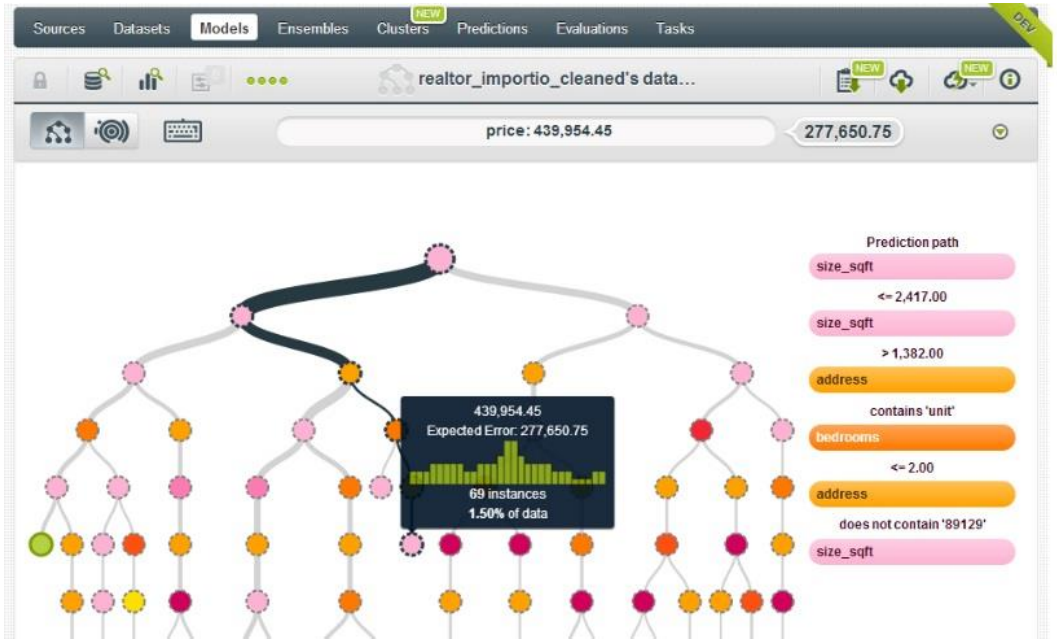
- No control of implementation
- No control of installation
- Limited control of data
- Limited access to support

~ Programming tools

- Full control of implementation
- No control of installation
- Limited control of data
- Limited access to support

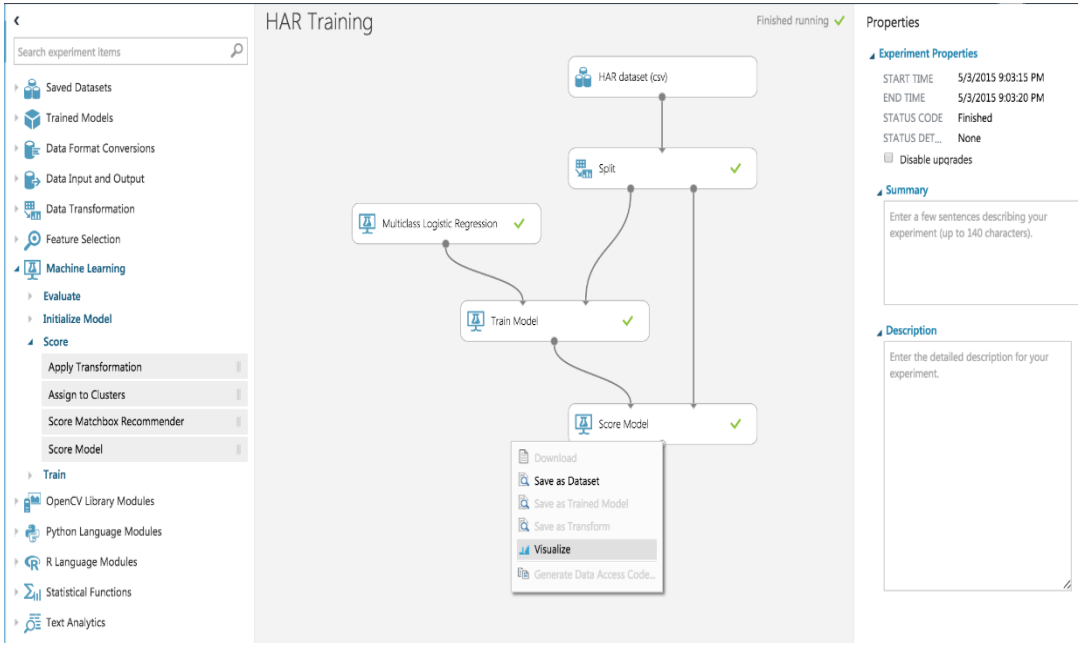
BigML (bigml.com)

- BigML is a Machine Learning service that offers an easy-to-use interface to import data and get predictions out of them.
- Implemented methods are decision trees and clustering.



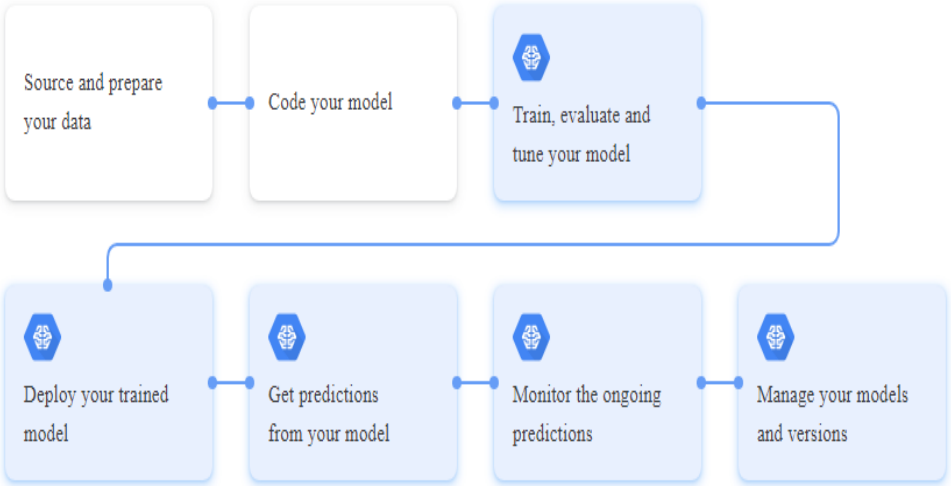
MS AzureML (studio.azureml.net)

- MS Azure is a cloud computing service created by Microsoft for building, testing, deploying, and managing applications.
- AzureML is a cloud-based predictive analytics service (decision trees, decision forrest, SVM, neural networks, logistic regression, clustering).



Google Cloud ML Engine (https://cloud.google.com/ml-engine/)

- Google Cloud Machine Learning (ML) Engine is a managed service that enables developers and data scientists to build and bring machine learning models to production.
- Cloud ML Engine supports Scikit-learn, TensorFlow and XGBoost.

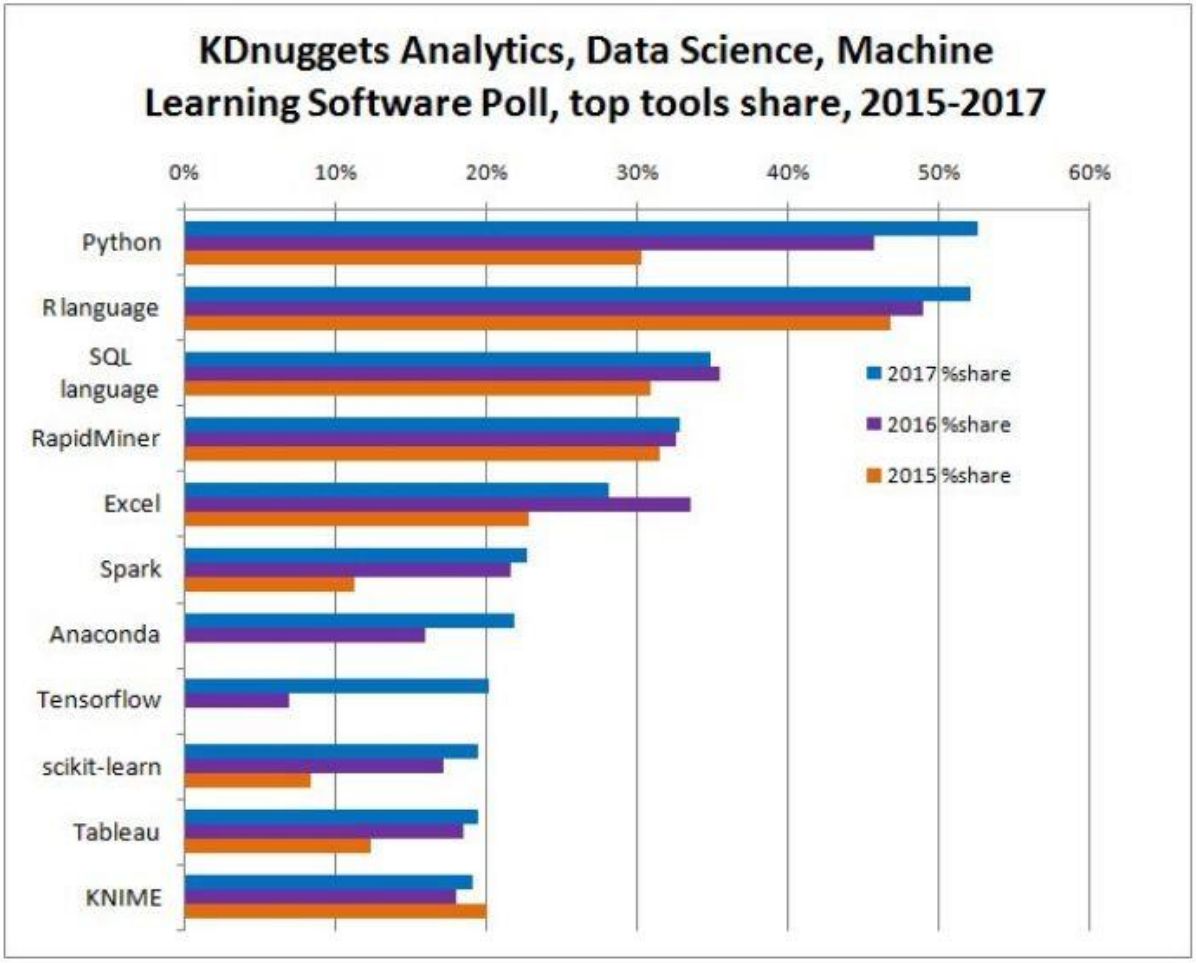


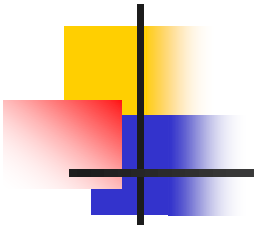
ML workflow

Suitability of tools

	Education in data mining	Research in data mining	Large apps	Small apps
Commercial suites	**Not much**	**No**	**Yes**	**Partially**
Free/open source suites	**Yes**	**Partially**	**Not much**	**Yes**
Cloud solutions	**Partially**	**No**	**Partially**	**Yes**
Programming tools	**Yes**	**Yes**	**Partially**	**Yes**

Popularity of tools





Conclusion

- There is a number of systems and programming tools for data mining suitable for different types of users
- The picture is even broader as I didn’t discuss tools related to Big Data concept

Thank you