Hadoop

- Is not a database!
- Is open-source software for reliable, scalable, distributed computing
- Is not a database!
- Allows the inclusion of database-like applications:
  - HIVE
  - HDFS
- Is not a database!
- Allows the inclusion of various other applications:
  - Mahout (machine learning)
  - Pig (sql-like data-flow language)
  - Spark (fast and general compute engine for Hadoop data)
  - Zookeeper (service for distributed applications)
Introduction

- RapidMiner offers broad setup of functionalities in various field of the data mining process
- But sometimes offers not the most efficient implementation in terms of resources
  - Time
  - RAM
  - CPU
- General source for alternatives:
  - http://www.kdnuggets.com
  - http://www.google.com
Overview

- Java
  - Weka
  - Elki

- Python
  - Scikit-learn
  - Orange

- R
  - Native
  - Rattle

- C
  - VFML
Weka

- General:
  - Current Version: 3.7 (including package manager for new packages)
  - Java based
  - Includes GUI
  - Two main working scenarios:
    - Explorer: Load data, test various algorithms
    - Experimenter: Compare different setups against each other
    - KnowledgeFlow Environment: Similar to RapidMiner UI, but more complex to use
  - Most algorithms can be included into RapidMiner via Weka extension

- Resources:
  - Official Page: http://weka.sourceforge.net/packageMetaData
  - Download: https://sourceforge.net/projects/weka/
Weka – Getting Started

- Download weka from the download page & install
- Run Weka [Version]
- Choose
  - Explorer or
  - Experimenter
  - KnowledgeFlow Env

Explore new datasets running clustering, classification and association analysis on the fly. Allows the visualization of data and results.

Setup experiments using different model learning approaches, different configurations and different datasets.

Visually setup a data mining pipeline using the KnowledgeFlow Environment.
Elki

- General:
  - Current Version: 0.7.1 (Feb 16)
  - Java based
  - Includes GUI
  - Major Functionalities:
    - Clustering
    - Outlier Detection
  - Getting-Started:
    - Download the elki.jar
    - Run `java -jar elki.jar`

- Resources:
  - [http://elki.dbs.ifi.lmu.de/](http://elki.dbs.ifi.lmu.de/)
Python

- General:
  - Linux: install via cmd
  - Windows: install directly or install “IDE” e.g. WinPython

- Resources:
  - Python: https://www.python.org/
  - WinPython: https://winpython.github.io/
  - iPython: https://ipython.org/
  - ML and Data Analysis package Anaconda: https://www.continuum.io/downloads
  - Python(x,y): http://python-xy.github.io/
Scikit-learn

- **General:**
  - Python based
  - No tool, just a library
  - Install as library or in scientific bundle like python(x,y) for windows
  - Functionalities:
    - Clustering
    - Ensembles
    - Preprocessing
    - Classification & Regression

- **Resources:**
Scikit-learn – Example

- Run cross-validation with linear regression and print plot
Orange

- **General:**
  - Python based
  - GUI with process/data flow design
  - Functionalities:
    - Visualization
    - Clustering & Unsupervised pre-processing
    - Classification & Regression
    - Evaluation
  - Operators are executed on the fly

- **Resources:**
  - Official Page: [http://orange.biolab.si/](http://orange.biolab.si/)
  - Install: [http://orange.biolab.si/download/](http://orange.biolab.si/download/)
  - Documentation: [http://orange.biolab.si/docs/](http://orange.biolab.si/docs/)
Orange – Getting Started

- Download and install tool
- **Run Orange Canvas**
Native R

- General:
  - No tool, but only libraries
  - Code an run native via cmd or use RStudio (package support etc.)
  - A large set of different available library for various tasks
  - Mostly necessary to look into the documentation of the packages

- Resources:
  - Introduction to ML with R: https://www.datacamp.com/community/tutorials/machine-learning-in-r
  - RStudio: https://www.rstudio.com/
Rattle

- General:
  - Python based
  - Wrapper around RGtk2
  - Includes GUI (similar to Weka)
  - Automatically loads missing packages
  - Getting Started:
    - RStudio (or similar):
      - `install.packages("RGtk2")`
      - `install.packages("rattle")`
      - `rattle()`
    - Find Help/FAQ on the official page

- Resources:
Very Fast Machine Learning (VFML)

- **General:**
  - C based
  - No tool, just a library
  - Functionalities:
    - Preprocessing
    - Clustering
    - Classification
    - Folding
  - Documentation is not too comprehensive

- **Resources:**