Toolboxes for Data Scientists

- Python
- R
- Matlab / Octave

Fundamental Python Libraries for Data Scientists

1. Numeric and Scientific Computation: NumPy and SciPy
2. Machine Learning in Python: SCIKIT-Learn
3. Python Data Analysis: PANDAS

Data Science Ecosystem Installation

- All in one bundle Anaconda

Getting Started

1. After installing Anaconda launch jupyter notebook from windows start menu
2. On linux run jupyter notebook from terminal
3. It will launch browser displaying jupyter homepage
4. To start a new notebook press New -> Notebooks -> Python 3
5. A blank notebook Untitled will be created
6. Click "Untitled" to rename and save the notebook
7. import tool boxes by adding the following lines in first cell

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

8. To execute a single cell press the `Run` button or click on Cell -> Run or press Ctrl + Enter

The DataFrame data structure

- key data structure in Pandas is DataFrame object.

- a tabular structure with rows and columns(can be seen as flexible spreadsheet).
- rows have specific index to access them, which can be any name or value
- columns are called “Series”, a special type of data
- Following code can be used to create a dataframe

```
data = {'year': [2010, 2011, 2012, 2010, 2011, 2012, 2010, 2011, 2012],
        'team': ['FCBarcelona', 'FCBarcelona', 'FCBarcelona', 'RMadrid',
                         'RMadrid', 'RMadrid', 'ValenciaCF',
                 'ValenciaCF', 'ValenciaCF'],
        'wins':   [30, 28, 32, 29, 32, 26, 21, 17, 19],
        'draws':  [6, 7, 4, 5, 4, 7, 8, 10, 8],
        'losses': [2, 3, 2, 4, 2, 5, 9, 11, 11]
        }
football = pd.DataFrame(
    data, columns=['year', 'team', 'wins', 'draws', 'losses'])
football
```