

Datenanalyse mit Python: Ein Einstieg

ditact 2018

Salzburg, 27.-28. August

Christine Koppelt

Course Materials

<https://github.com/cko/ditact2018-data> (<https://github.com/cko/ditact2018-data>).

About me

- Senior Consultant at INNOQ since 2011
- Software development since 2007
- Diplom Mathematikerin (FH)
- Current Focus: Microservices, Devops, Data Engineering

About you

- What is your name?
- What got you interested in this course?
- What do you already know about Python and data analysis?
- Grab a Post-It and write down one question you would like to have answered during the course

Agenda Monday

- Welcome & Intro
- Getting used to Conda & Jupyter
- Quick Python repetition
- Getting started with numpy & pandas
- Descriptive statistics

Agenda Tuesday

- Combining data, cleaning data
- Plotting & visualization
- Time series
- Linear Regression

Data Analysis

- Goal: Discovering useful information, supporting informing conclusions and decision-making based on data by using statistical methods
- How
 - Data Loading
 - Data Cleaning & Preparation
 - Exploring & Visualization
 - Modeling
 - Interpreting

Conda Overview

Conda Basic Facts

- Package management and environment mangement system
- Supports Windows, MacOS, Linux
- Supports Python, R, Scala, Java, ...
- Huge software repository: <https://repo.continuum.io/pkgs/>
(<https://repo.continuum.io/pkgs/>).
- Maintained by Anaconda Inc.

Anaconda vs Miniconda

- Anaconda
 - Complete Python and R distribution
 - Includes conda (the package and environment management program)
 - Includes 100+ scientific Python and R packages
- Miniconda
 - Lightweight version of Anaconda
 - Contains only Python and conda and a few packages

Command Line - Package Manager

- List all installed packages: `conda list`
- List all available packages: `conda search`
- Search package with all version by name: `conda search panda*`
- Search online: <https://anaconda.org/>(<https://anaconda.org/>)

Command Line - Package Manager

- Install a package: `conda install pandas=0.23.4`
- Install another Python version: `conda install python=3.7.0`
- Update all packages: `conda update --all`

Command Line - Environment Manager

- List environments: `conda env list`
- Create an environment: `conda create myenv`
- Activate an environment: `source activate myenv`
- Show info about the environment: `conda info`
- Deactivate the environment: `source deactivate`

Cheatsheet

- [https://conda.io/docs/ downloads/conda-cheatsheet.pdf](https://conda.io/docs/downloads/conda-cheatsheet.pdf)
([https://conda.io/docs/ downloads/conda-cheatsheet.pdf](https://conda.io/docs/downloads/conda-cheatsheet.pdf)).

Jupyter Overview

Project Jupyter

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

<https://jupyter.org/>

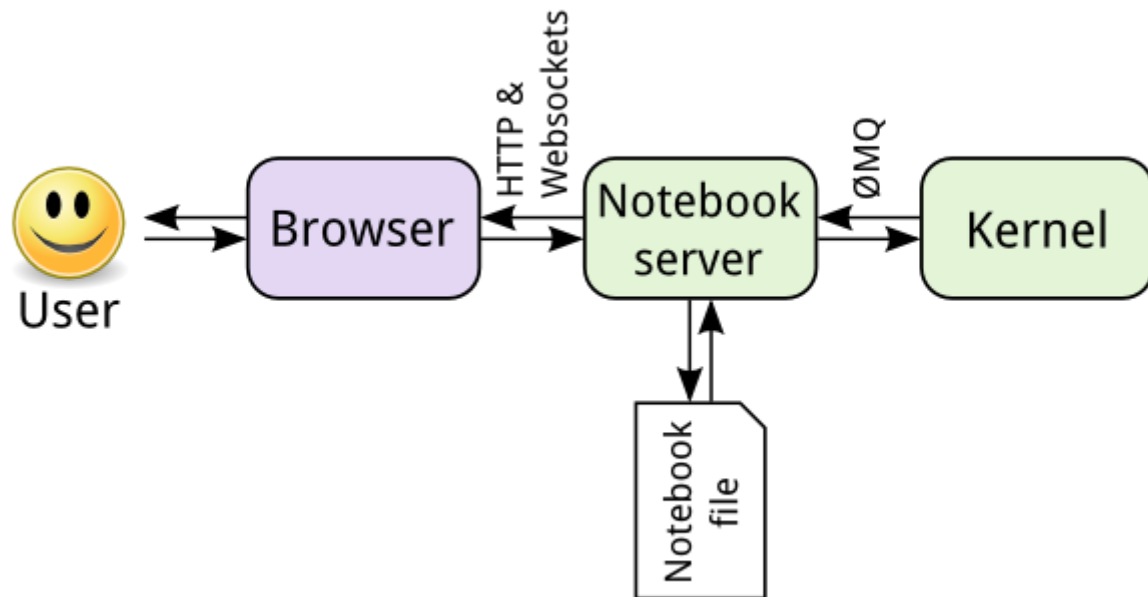
- Origin: iPython, iPython Notebook
- Open source, BSD license
- Started in 2014 by Fernando Pérez, assistant professor in the Department of Statistics at UC Berkeley
- Supported by Microsoft, Google and several foundations
- Very popular in the data analysis / data science / machine learning space

Jupyter Ecosystem

- Supports ~50 languages: Python, R, Julia, Scala, ...
- Similar software: MATLAB, Mathematica, R Studio, Tableau, PowerBI, Excel
- ipywidgets, interactive
- nbviewer
- nbconvert
- RISE, nbpresent
- latex, rst export
- Hub

Demo

Architecture



Use Cases

- Data analysis, data exploration, machine learning
- Data query tool (for debugging or for support)
- Python in the browser
- Publishing and sharing
- Presentations
- Not: software development

Run Cells

- Run and stay at current cell: `Ctrl+Enter`
- Run and advance to next cell: `Shift+Enter`
- Run all cells in a notebook -> Menu

Manage Cells

- Switch between command and edit mode: Enter, ESC/Ctrl+M
- In command mode:
 - Delete cell: dd
 - Add cell before a or after b current cell
 - Copy cell: c + v
 - Change cell type: markdown m, code y, raw r

Exercise 1 - Conda & Jupyter

Goals:

- Have a working Jupyter environment ready
- Getting familiar with Conda & Jupyter

Tasks:

- Install the command line tool `curl` via conda and use it to download a data file:
`curl https://raw.githubusercontent.com/cko/ditact2018-data/master/data/cafe.csv --output cafe.csv`
- Update `requirements.txt` via conda
- Create a Jupyter notebook file, create some code cells, write some Python code, like `print('Hello world')`, and execute it
- Create a markdown cell
- Try some shortcuts:
 - Execute a cell: `Ctrl+Enter` and `Shift+Enter`
 - Create a cell before a or after b
 - Copy c and paste v a cell
- Print your current working directory