

SDSC 2001 Python for Data Science

Lecture 1

Dr. Xinyue Li

Sep. 2nd 2022

Course Overview

- SDSC 2001 Python for Data Science
- Lectures and tutorials: F 9:00-11:50

- Instructor: Xinyue LI, PhD

xinyueli@cityu.edu.hk

- Teaching Assistants: Cheng CAO
Collin SAKAL
Wei ZHANG

chengcao3-c@my.cityu.edu.hk

csakal2-c@my.cityu.edu.hk

wzhang472-c@my.cityu.edu.hk

Course Overview

- Instructor: Xinyue LI, PhD

xinyueli@cityu.edu.hk

- Teaching Assistants: Cheng CAO
Collin SAKAL
Wei ZHANG

chengcao3-c@my.cityu.edu.hk

csakal2-c@my.cityu.edu.hk

wzhang472-c@my.cityu.edu.hk

- Office Hours: Cheng CAO Wed 2-3pm
 Collin SAKAL Fri 8-9am
 Wei ZHANG Wed 3-4pm

Starting from the 5th week

Course Overview

- Course objective:

This course provides students with extensive exposures to the use of Python specifically for data science. Topics include Python language fundamentals, data analysis using Python libraries, applied machine learning in Python, and the practice of scientific computing. The students will acquire hands-on experience using Python and the popular packages related to data manipulation, processing and analysis, with the minimal theoretical background in methodological aspects. The students are expected to develop Python codes independently from scratch in professional ways for elementary algorithms in data sciences.

Course Overview

- Basics of Python programming language:
 - Installation and setup of Python and Python Packages (shell and interactive environment); IPython and Jupyter Notebook;
 - Basics of Python language; data structures in Python; file operations; functions in Python;
 - Data processing; profiling and timing; version control skills (github)
- Introduction to Python for scientific computing, data processing and plotting:
 - Scientific computing with NumPy (multidimensional array, indexing and slicing, array functions, pseudo-random generators);
 - Data manipulation with Pandas (series, dataframe, selection, filtering, sorting, ranking);
 - Visualization with Matplotlib
- Python for elementary machine learning (Scikit-Learn):
 - Pre-processing
 - Feature extraction
 - Simple linear regression analysis
 - Classification analysis
 - Decision tree and random forest
 - K nearest neighbors
 - Model selection

Course Overview

- Grading structure
 - Assignments: 25%
 - Course project: 25%
 - Examination: 50%
- Note:
- For assignments, relevant python codes, outputs, and interpretations, if any, shall be provided. Late submission is not accepted unless the student obtains the instructor's consent one day in advance. You may discuss problems with fellow students, but you must write up your own solutions and your own codes.
- The course project involves real data analysis. The objective is to perform data exploration and data analysis using techniques learned in class. All students must work independently and CANNOT discuss the project with fellow students.
- *For a student to pass the course, at least 30% of the maximum mark for the course project must be obtained.

Course Overview

- Grading structure
 - Assignments: 25%
 - Course project: 25%
 - Examination: 50%
- Note:
- Class attendance and participation

Class Overview

- Environment set-up

Environment Set-up

Installation and Set-up of Python

- Python: a programming language
 - <https://www.python.org/>
- Python is an interpreted, high-level, general-purpose programming language.
- Python's design philosophy emphasizes code readability.
- Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Anaconda Distribution of Python

- Anaconda: a data science platform for python
 - a distribution of the Python and R programming languages for data science and machine learning
 - <https://www.anaconda.com/>
- Thousands of packages
 - Analysis
 - Visualization
 - Modelling

Differences between Python and Anaconda?



- Anacondas are native to South America while pythons are naturally found in Asian and African tropics.
- Comparatively, anaconda is heavier, but python is longer.
- Python is more agile than anaconda is.
- The colouration patterns are organised and arranged to an order in anaconda but not in pythons.
- Anaconda is a good swimmer and found around water more often than not, whereas python prefers to perch on trees and dry habitats.
- Python is a selective feeder while anaconda is a general predator.
- Pythons are more popular among humans as a pet, but anacondas are not commonly reared as pets.

Installation and Set-up of Python

- Python: a programming language
 - <https://www.python.org/>
- Anaconda: a data science platform for python
 - a distribution of the Python and R programming languages for data science and machine learning
 - <https://www.anaconda.com/>

Installation and Set-up of Python

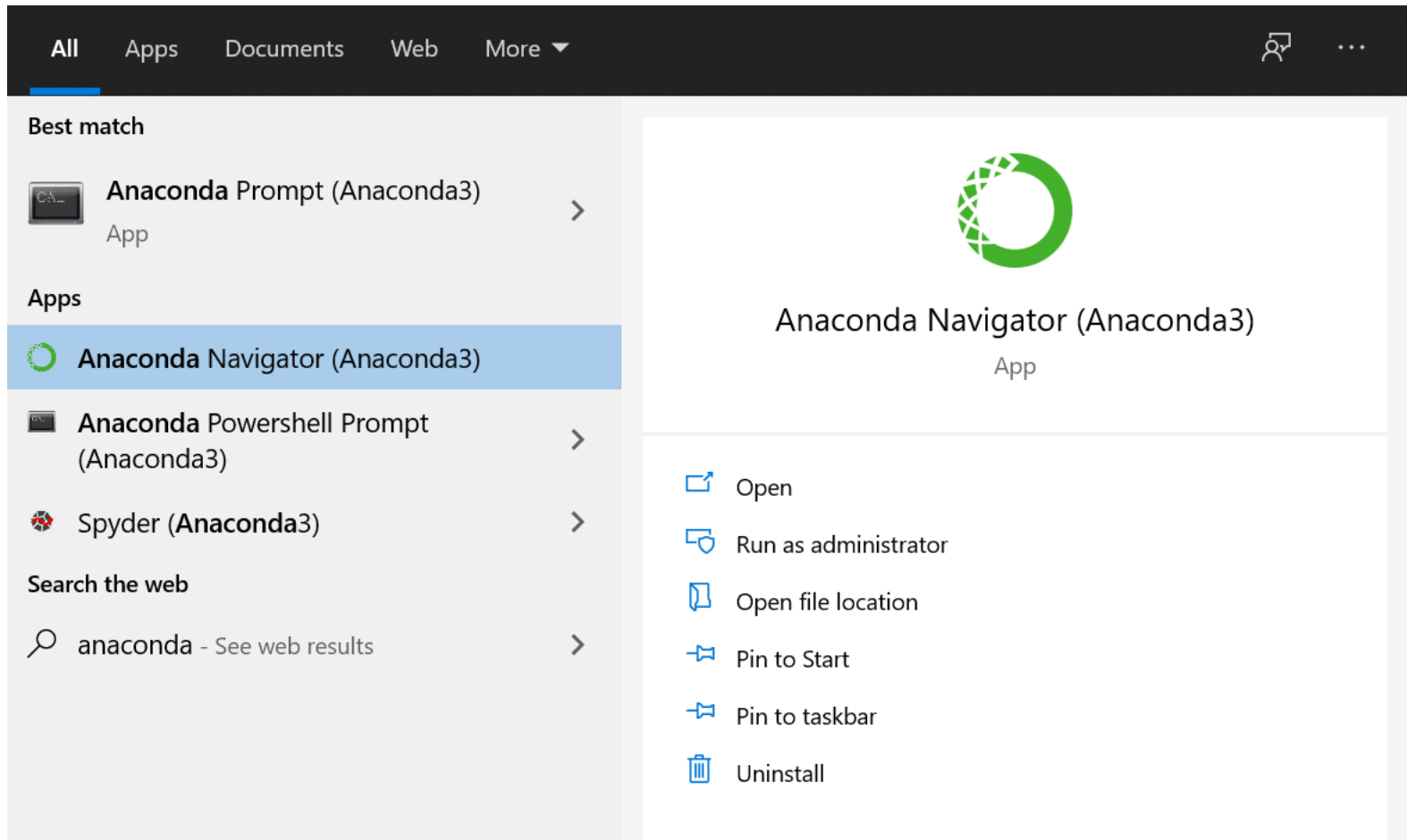
- Go to Anaconda: <https://www.anaconda.com/products/individual>
- Python 3.9
- Choose the version for your system

Anaconda Installers

Windows 	MacOS 	Linux 
Python 3.9 64-Bit Graphical Installer (594 MB) 32-Bit Graphical Installer (488 MB)	Python 3.9 64-Bit Graphical Installer (591 MB) 64-Bit Command Line Installer (584 MB) 64-Bit (M1) Graphical Installer (316 MB) 64-Bit (M1) Command Line Installer (305 MB)	Python 3.9 64-Bit (x86) Installer (659 MB) 64-Bit (Power8 and Power9) Installer (367 MB) 64-Bit (AWS Graviton2 / ARM64) Installer (568 MB) 64-bit (Linux on IBM Z & LinuxONE) Installer (280 MB)

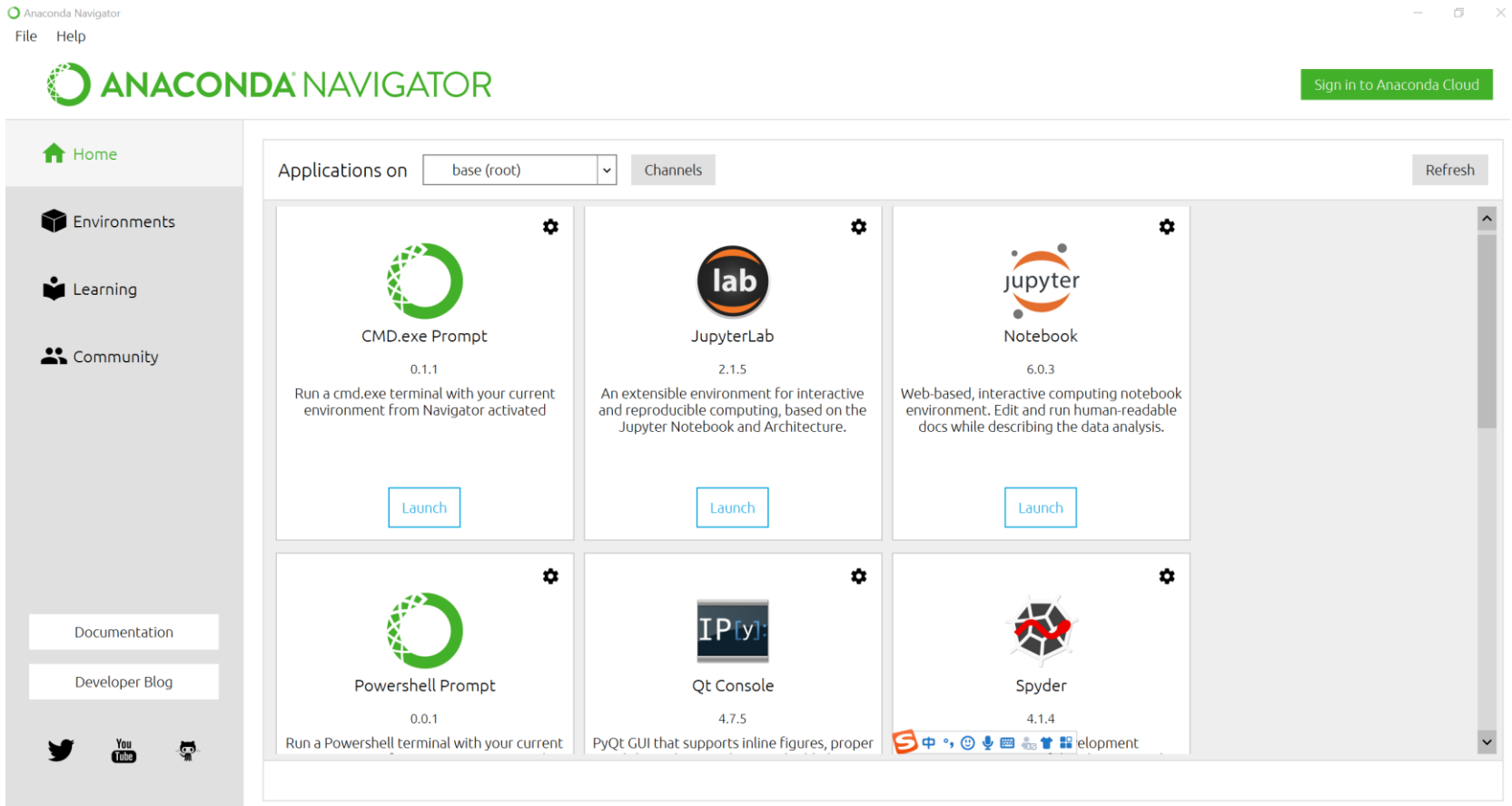
Installation and Set-up of Python

- Anaconda Navigator



Installation and Set-up of Python

- Anaconda Navigator



Installation and Set-up of Python

- Anaconda Navigator
- Download/upgrade/downgrade packages
 - Manage packages easily
- Create separate working environments
 - With different packages/versions of packages installed
- Documents for learning
- Python community
 - Exchange ideas
 - Solve problems

Installation and Management of Packages

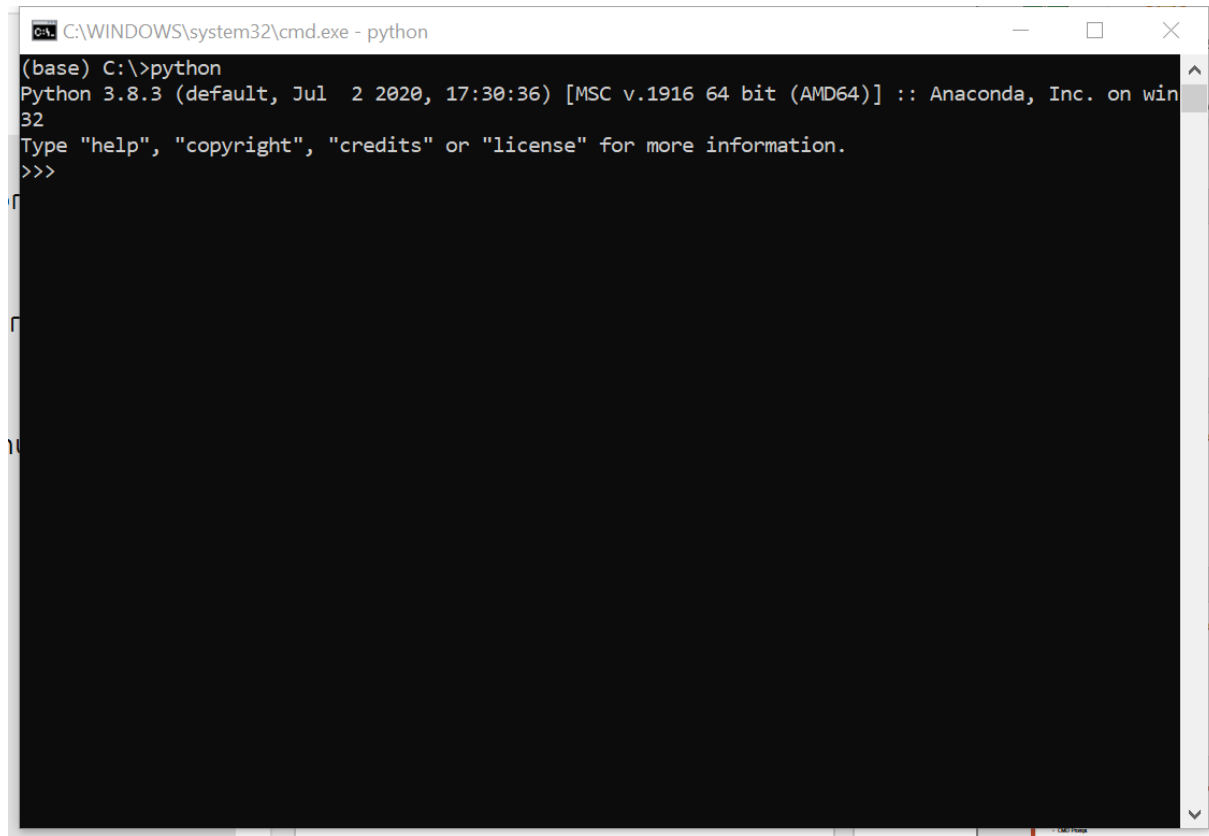
- Install packages
- Manage version control

Installation and Management of Packages

- Manage environments
- Create separate working environments
 - With different packages/versions of packages installed

Python: Command Line

- CMD Prompt



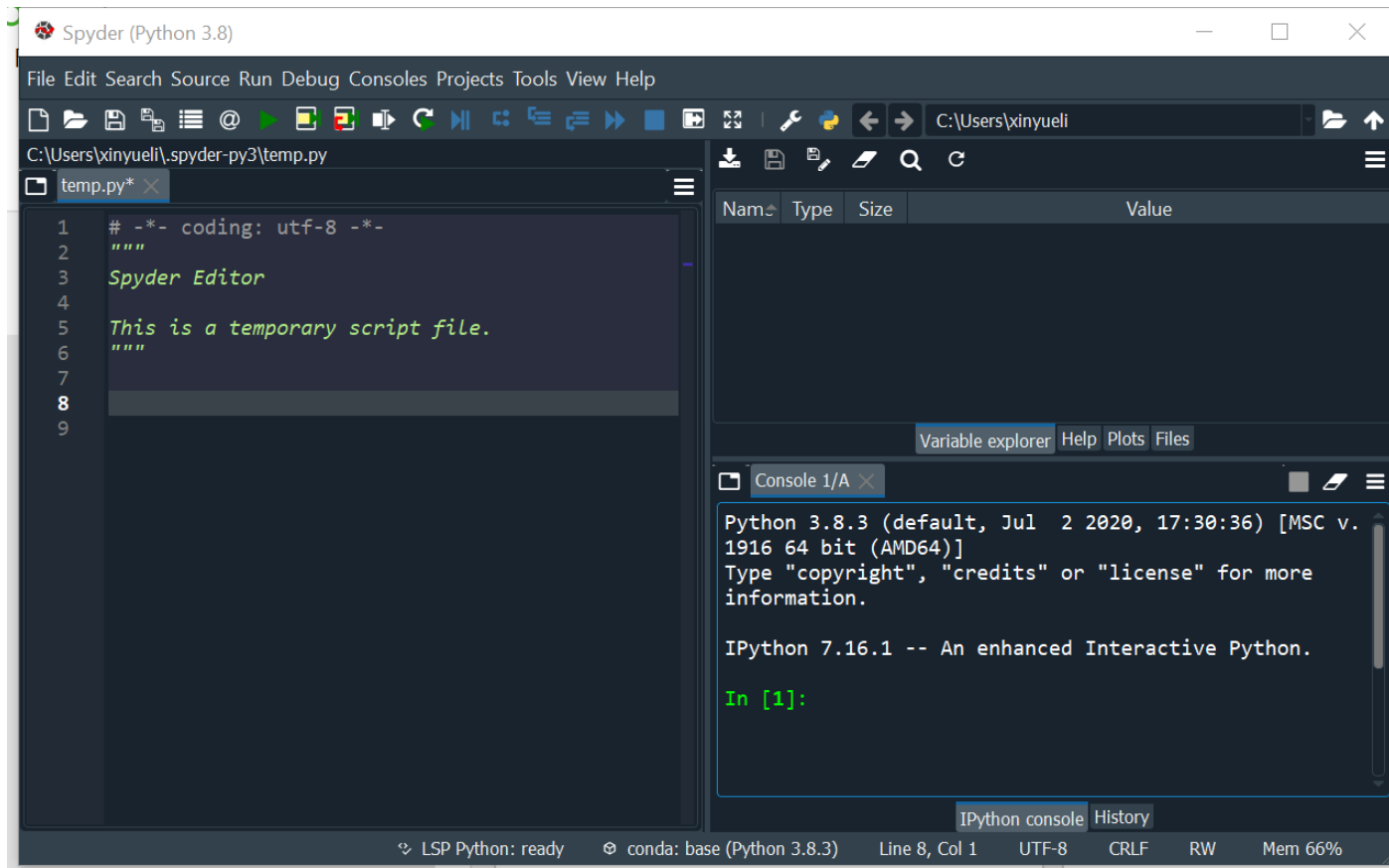
A screenshot of a Windows Command Prompt window. The title bar reads "C:\WINDOWS\system32\cmd.exe - python". The command prompt shows the following text:
(base) C:\>python
Python 3.8.3 (default, Jul 2 2020, 17:30:36) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>

Python: Use of Spyder

- Spyder
 - An IDE, or Integrated Development Environment, for Python
- IDEs increase programmer productivity by combining common activities of writing software into a single application
 - Editing source code
 - Syntax highlighting
 - autocomplete
 - Building executables
 - Automatically build an executable file
 - Debugging
 - Examine different variables and inspect codes in a deliberate way
 - Provide hints while coding to prevent errors

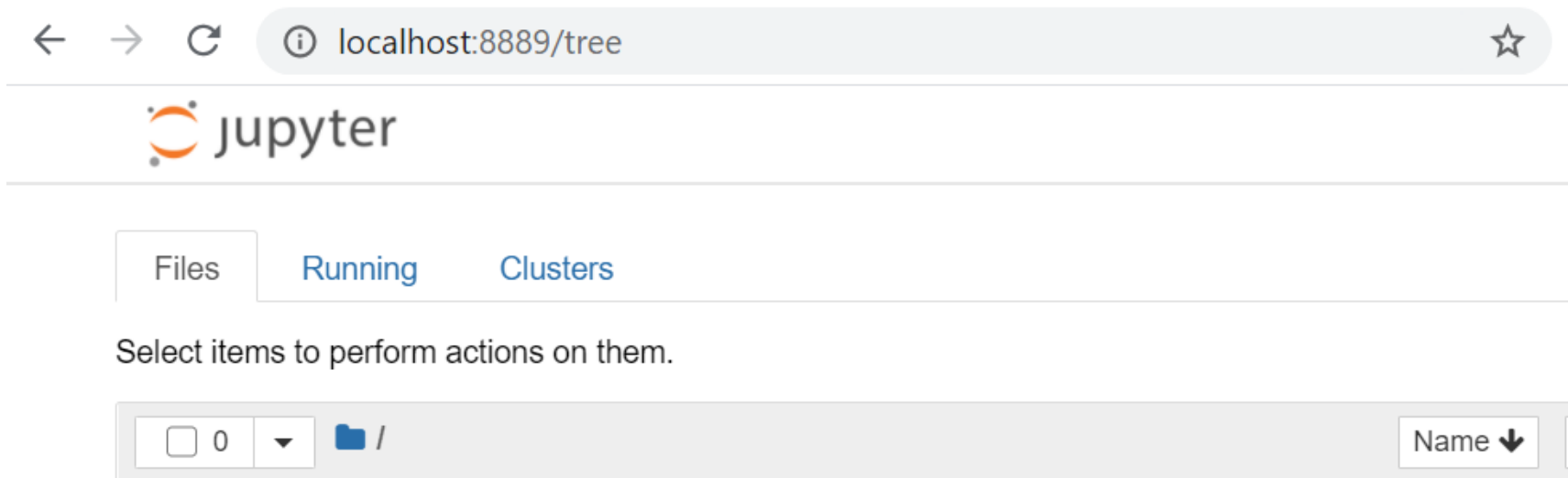
Python: Use of Spyder

- Spyder
 - An IDE, or Integrated Development Environment, for Python



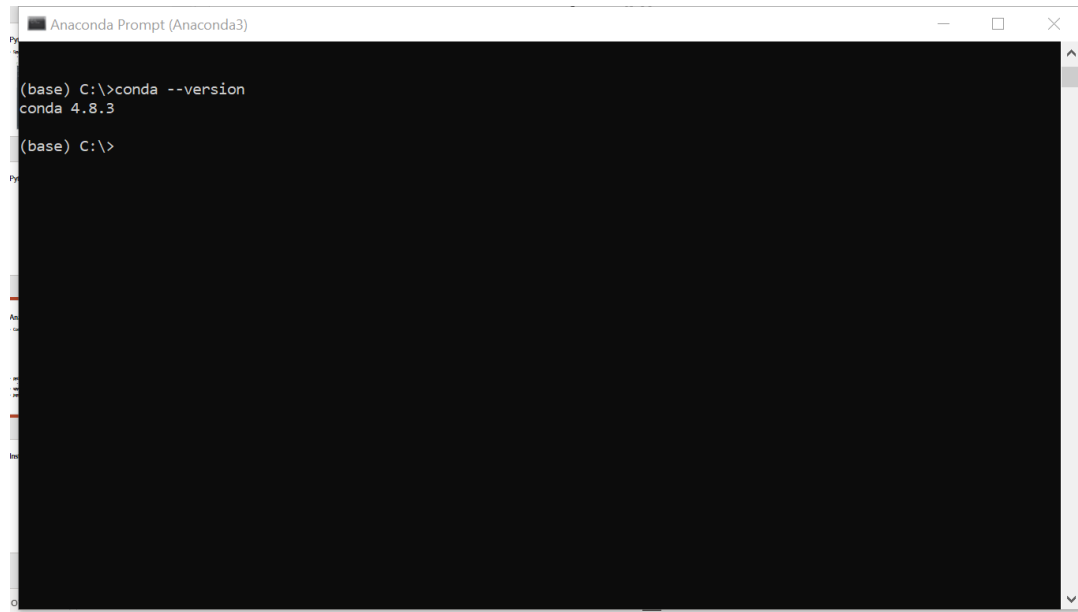
Python: Jupyter Notebooks

- Jupyter notebooks
 - An open-source web application
 - Create and share documents that contain live code, equations, visualizations and narrative text



Anaconda Prompt

- Conda cheat sheet: [link](#)



```
Py
Anaconda Prompt (Anaconda3)
(base) C:\>conda --version
conda 4.8.3
(base) C:\>
```

- python
 - exit()
- spyder
- jupyter-notebook