Data Science Survival Skills

Exercise 0 (soft exercise)

Agenda

- Anaconda
- Virtual Environments
- Jupyter Notebooks
- Google Colab
- NumPy
- Pandas
- Matplotlib

Anaconda

- Open-source Python (and R) distribution platform for scientific computing, data science and machine learning
- Features: Package management, Cross-platform, Extensive library ecosystem, Virtual Environments, Data science tools (e.g. Jupyter Notebook, Spyder)

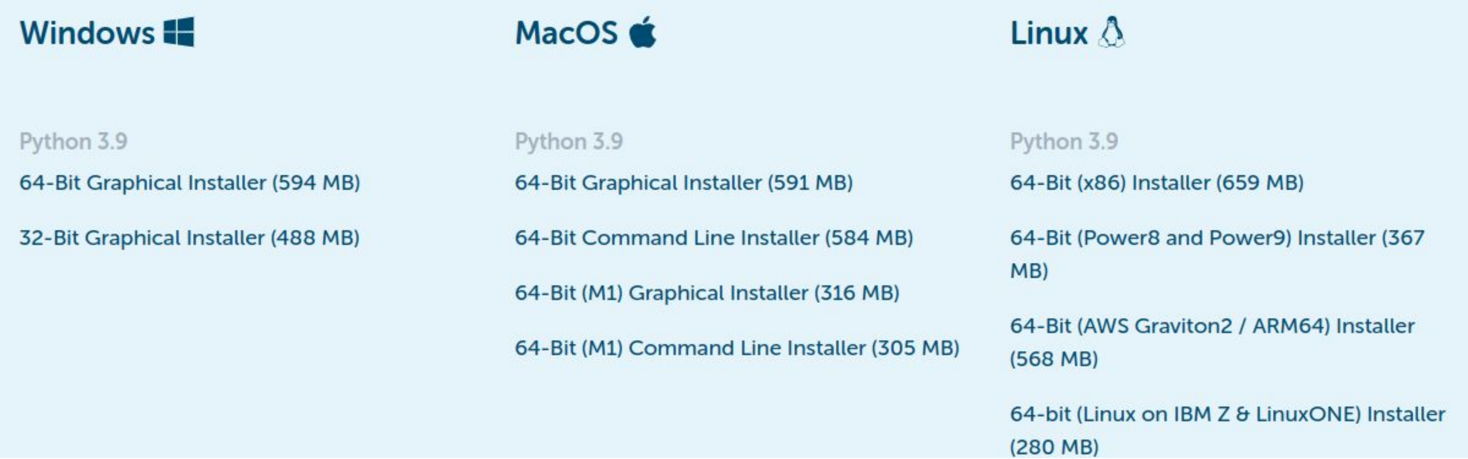
ANACONDA®

SPYDER

jupyter

Anaconda installation

- Official website: Link to Anaconda

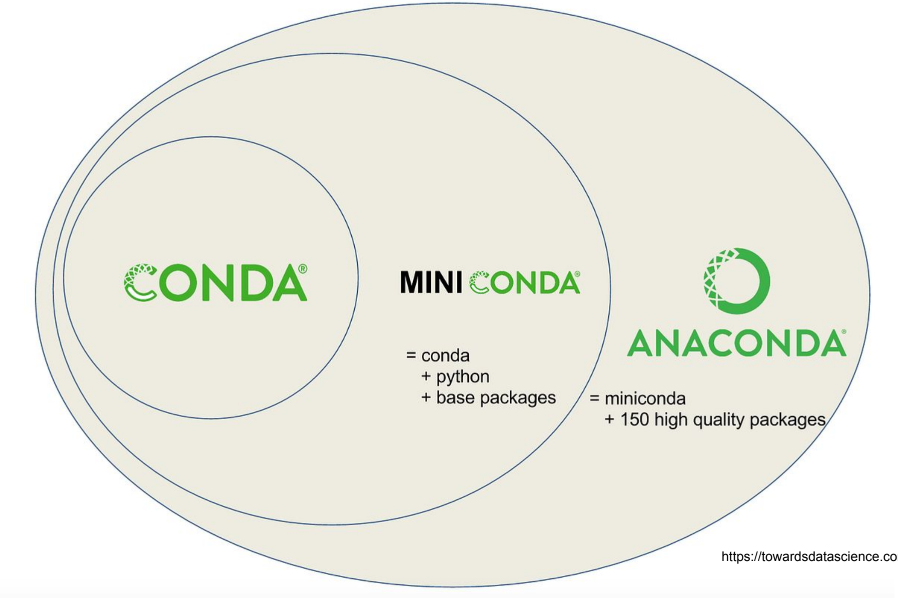


Windows	MacOS	Linux
Python 3.9	Python 3.9	Python 3.9
64-Bit Graphical Installer (594 MB)	64-Bit Graphical Installer (591 MB)	64-Bit (x86) Installer (659 MB)
32-Bit Graphical Installer (488 MB)	64-Bit Command Line Installer (584 MB)	64-Bit (Power8 and Power9) Installer (367 MB)
	64-Bit (M1) Graphical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

- Test installation with “conda –version”
- Updating Anaconda: “conda update anaconda”

Miniconda

- Minimalistic distribution of Anaconda that includes only Conda package manager and its dependencies
- Lightweight version of Anaconda
- Scriptable installation (suitable for automated or scripted installations)

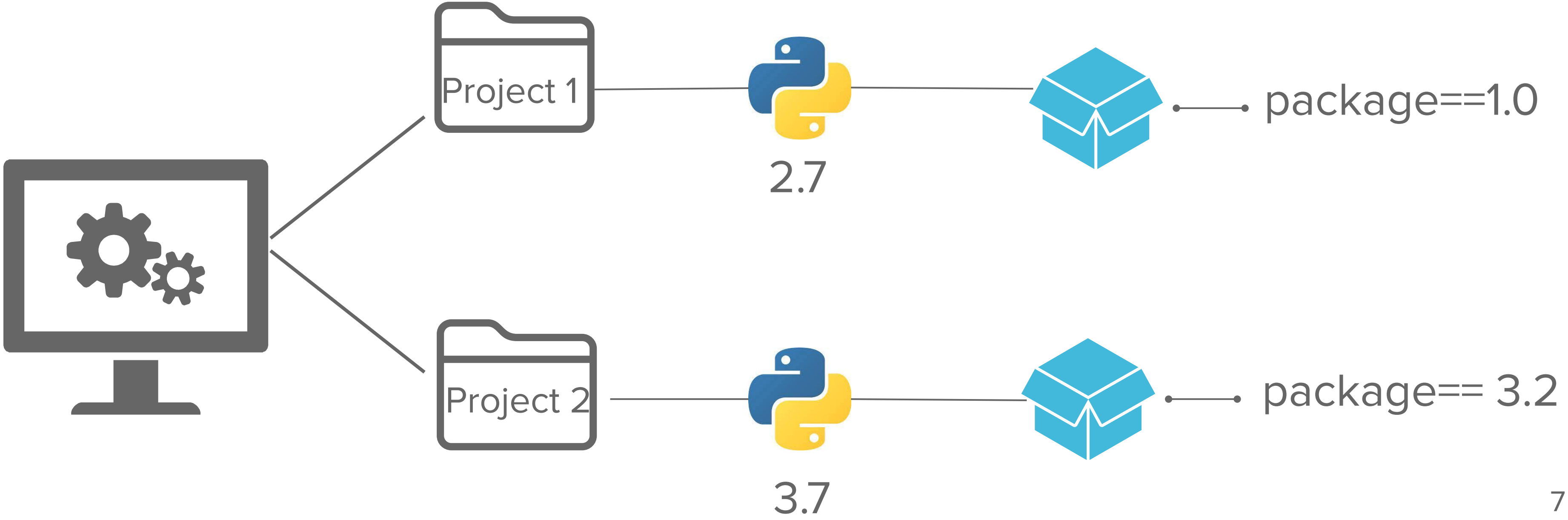


https://towardsdatascience.com/managing-project-specific-environments-with-conda-b8b50aa8be0e

1. Virtual Environments

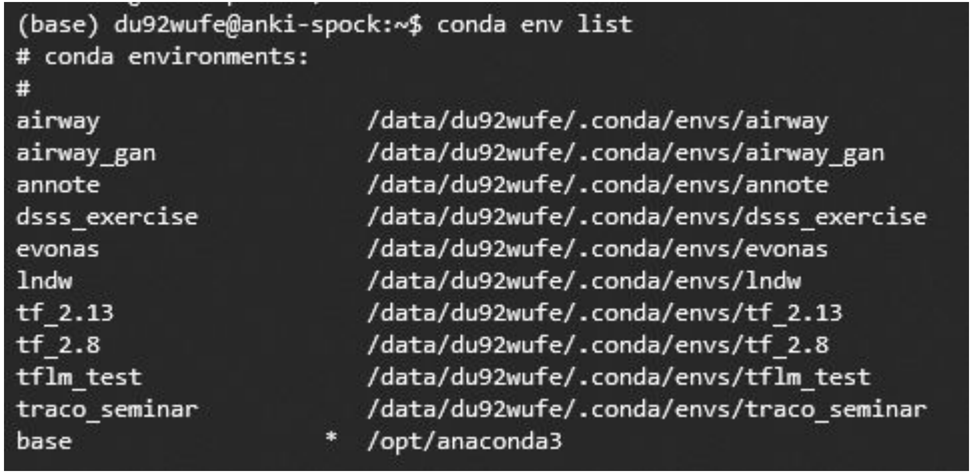
Virtualenv

- Isolated, self-contained spaces
- Manage and control the packages and dependencies for specific projects



Conda for virtualenvs

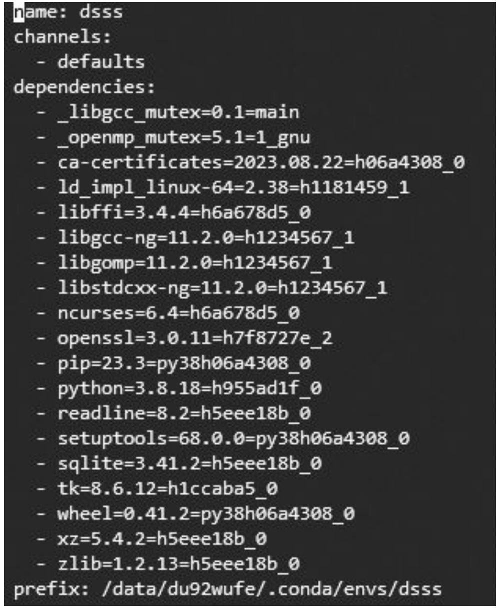
- Create: “conda create –name my_virtual_env python=3.8”
- Activate: “conda activate my_virtual_env”
- Deactivate: “conda deactivate”
- List all envs: “conda env list”
- Delete env: “conda env remove –name my_virtual_env –all”



```
(base) du92wufe@anki-spock:~$ conda env list
# conda environments:
#
airway                   /data/du92wufe/.conda/envs/airway
airway_gan               /data/du92wufe/.conda/envs/airway_gan
annote                   /data/du92wufe/.conda/envs/annote
dsss_exercise            /data/du92wufe/.conda/envs/dsss_exercise
evonas                   /data/du92wufe/.conda/envs/evonas
lndw                     /data/du92wufe/.conda/envs/lndw
tf_2.13                  /data/du92wufe/.conda/envs/tf_2.13
tf_2.8                   /data/du92wufe/.conda/envs/tf_2.8
tflm_test                /data/du92wufe/.conda/envs/tflm_test
traco_seminar            /data/du92wufe/.conda/envs/traco_seminar
base                  *  /opt/anaconda3
```

Conda for virtualenvs

- Exporting environment information ➔ environment.yml
- YAML (Yet Another Markup Language) ➔ human-readable data serialization format
- Command: “conda env export > environment.yml”



```
name: dsss
channels:
  - defaults
dependencies:
  - _libgcc_mutex=0.1=main
  - _openmp_mutex=5.1=1_gnu
  - ca-certificates=2023.08.22=h06a4308_0
  - ld_impl_linux-64=2.38=h1181459_1
  - libffi=3.4.4=h6a678d5_0
  - libgcc-ng=11.2.0=h1234567_1
  - libgomp=11.2.0=h1234567_1
  - libstdcxx-ng=11.2.0=h1234567_1
  - ncurses=6.4=h6a678d5_0
  - openssl=3.0.11=h7f8727e_2
  - pip=23.3=py38h06a4308_0
  - python=3.8.18=h955ad1f_0
  - readline=8.2=h5eee18b_0
  - setuptools=68.0.0=py38h06a4308_0
  - sqlite=3.41.2=h5eee18b_0
  - tk=8.6.12=h1ccaba5_0
  - wheel=0.41.2=py38h06a4308_0
  - xz=5.4.2=h5eee18b_0
  - zlib=1.2.13=h5eee18b_0
prefix: /data/du92wufe/.conda/envs/dsss
```

Use virtual environments!

➔ Documentation here

2. Jupyter Notebooks

Jupyter Notebooks

- Open-source project
- Interactive computing environment
- Not only for Python (supports over 40 programming languages)
- Real-time code execution
- Supports markdown (a lightweight markup language)
- Integration of data visualization libraries
- Data science: often used for data exploration, model development and visualization

3. Google Colab

Google Colaboratory

- Cloud-based free-to-use and collaborative Jupyter Notebook environment
- Runs in the cloud on Google’s servers → accessible from any device with a web browser and an internet connection (no local installations needed)
- Fully integrates with Jupyter Notebook
- Comes with a variety of pre-installed Python libraries, e.g. NumPy, Pandas, Matplotlib or scikit-learn
- Access to **Graphis Processing Units (GPUs) and Tensor Processing Units (TPUs)**, allowing to accelerate computationally intensive tasks (→ more details in lecture 1)
- Free of cost
- Supports data import/export from Google Drive and Google Sheets

4. NumPy

NumPy

- Short for “Numerical Python”
- Fundamental package for scientific computing
- Support for large, multi-dimensional arrays and matrices
- Essential tool for tasks such as data analysis, machine learning and scientific research
- Written in C programming language and highly optimized

NumPy features

- N-dimensional arrays: central feature is the “ndarray”
    - Multi-dimensional array object
    - Arrays can be of any shape and size
    - Enabling efficient storage and manipulation of large datasets
- Element-wise operations: mathematical and logical operations on entire arrays without the need of explicit loops
- Mathematical functions: statistical, linear algebra, etc.

- Documentation here

5. Pandas

Pandas

- Open-source data manipulation and analysis library
- Provides data structures and functions for efficiently working with structured data (e.g. spreadsheets, databases → more details in lecture 3)
- Two primary data structures: Series (one-dimensional array-like) and DataFrame (two-dimensional table-like)
- Can handle a wide range of data types (numerical, textual, etc.)
- Data import and export: CSV, Excel, JSON

- Documentation here

6. Matplotlib

Matplotlib

- Open-source data visualization library
- Creation of static, animated and interactive plots and graphs for data analysis
- Ability to produce publication-quality plots (➔ more details in lecture 4)
- Supports various types of plots, e.g. line plots, scatter plots, histograms, heatmaps
- Can be used interactively in a Jupyter Notebook to explore and visualize data dynamically
- Exporting plots in various formats, e.g. PNG, PDF, SVG

- Documentation here

Last slide

Next week

- **No in-person exercise!**
- We will upload a video showing how to build a computer/workstation