

DATA MINING AND MACHINE LEARNING (cod. 878II) AY2025-2026 [WAI-LM]

Intro



Version 29-9-2025

Dept. of Information Engineering – University of Pisa

Email: fabrizio.ruffini@unipi.it

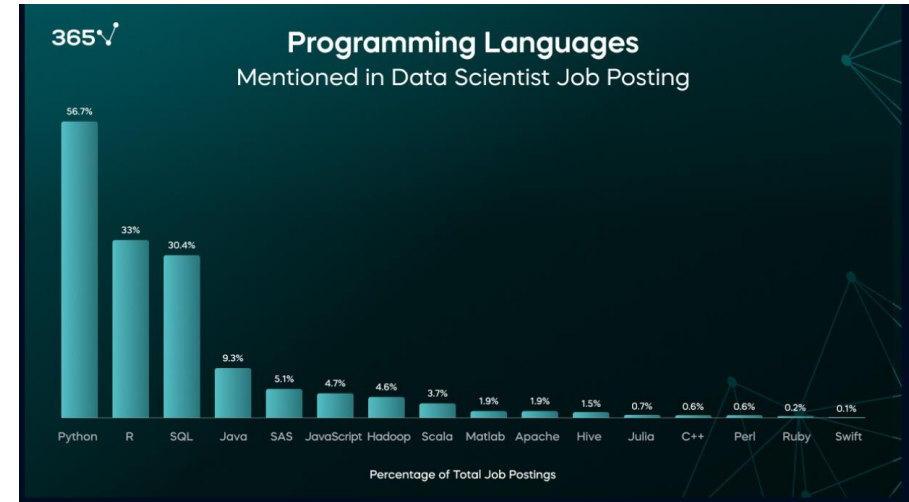
Department of Information Engineering
University of Pisa
Largo Lucio Lazzarino 1

Content of the Lecture

- Tools: python + jupyter notebook

Tools for data analysis (focus on AI/ML)

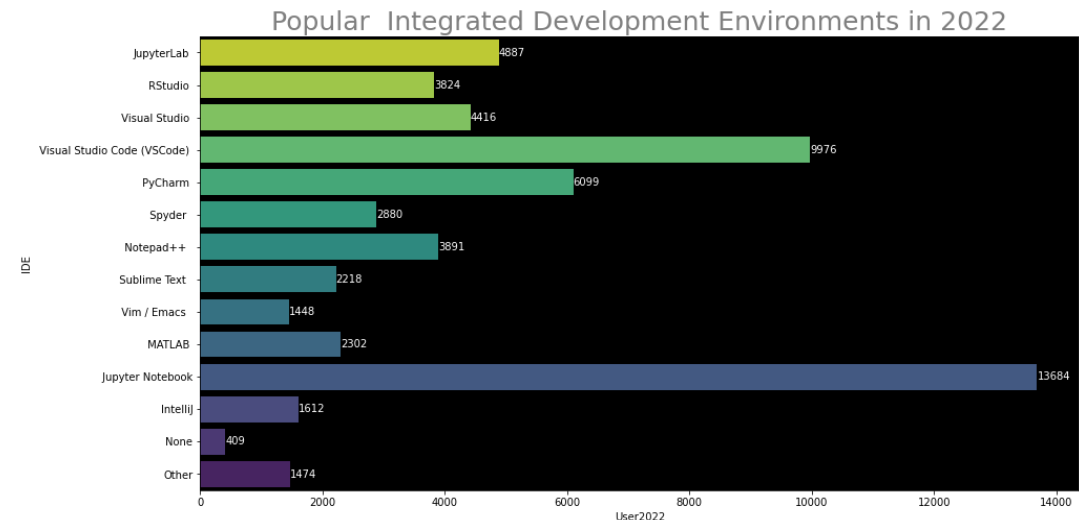
- **Python (and Jupyter nb)**
- Weka
- MatLab
- R
- ...



from
<https://365datascience.com/career-advice/data-scientist-job-market/>

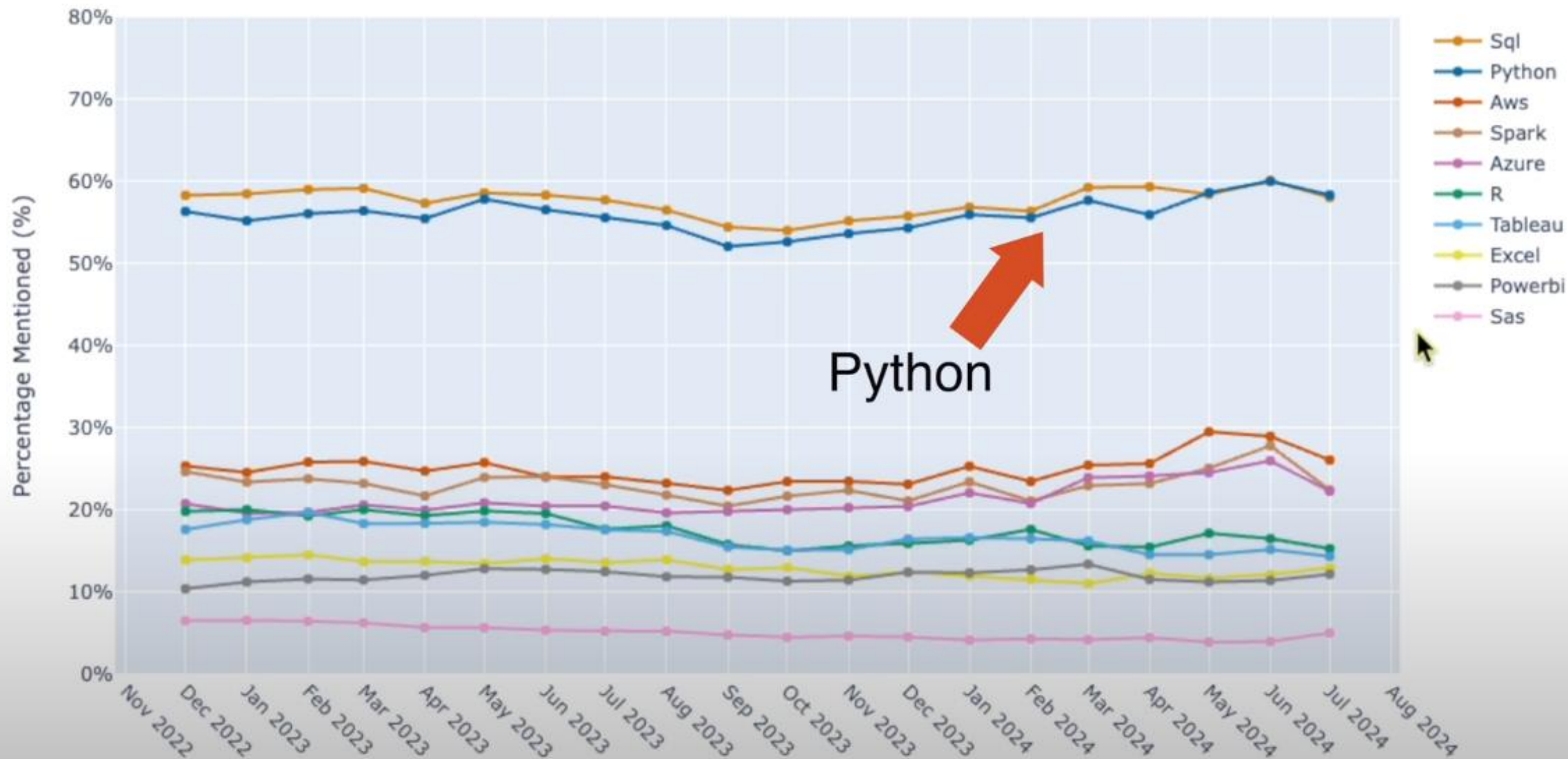
Survey:

<https://forms.office.com/e/7zrYdQiH2L>



Percentage of Data Skills Mentioned in Job Posts

Data source: <https://datanerd.tech/>



Focus on python: what



Python is a programming language created by Guido van Rossum in 1991 (name inspired by BBC TV show):

- interpreted (instructions are not directly executed by the machine),
- dynamic typed (it is interpreted at runtime by the machine itself),
- Many packages <-> large community (especially for AI activities)

Version: python 3 recommended (python 2 last release in 2010)

Focus on python: what: programming paradigms

Procedural programming paradigm

- Procedures simply contain a series of computational steps to be carried out
- Generally use reserved words that act on blocks, such as if, while, and for, to implement control flow
- Break down a programming task into a collection of variables, data structures, and subroutines

Object-oriented programming paradigm

- Break down a programming task into objects that expose behavior (methods) and data (attributes)
- Advantages:
 - natural support for modelling of real-world objects, or for reproducing abstract models
 - easy management and maintenance of large projects
 - organisation of code in the form of classes favours modularity and reuse of code

Python supports multiple progr. paradigms -> you can use different styles and approaches to problem-solving

Meet Anaconda

[Anaconda](#) is a distribution of the Python programming languages for scientific computing.

Contains most of usually used data-science packages (>250)

Package versions in Anaconda are managed by the package management system **conda** (or **miniconda**, *miniature installation of Anaconda Distribution that includes only conda, Python*).



Anaconda Repository

Our repository features over 8,000 open-source data science and machine learning packages, Anaconda-built and compiled for all major operating systems and architectures.

Focus on python: environments

Virtual environment useful to manage different projects:

1. Different versions of python (3.4, 3.8...)
2. Different packages (scikit-learn, keras...)
3. Easy-to-export for collaboration (github...)
4. ...

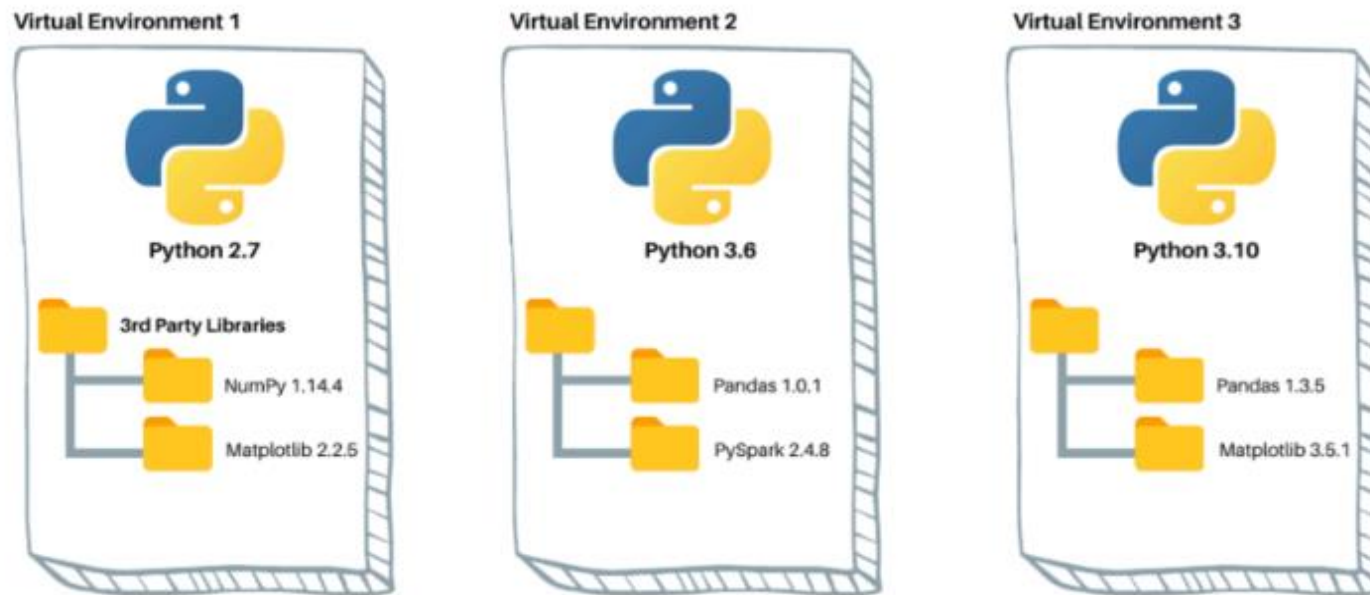


image from <https://www.dataquest.io/blog/a-complete-guide-to-python-virtual-environments/>

Focus on python: how to

Recommended Editors:

- Spyder (IDE)
- PyCharm (IDE)
- Visual studio code

During the lectures,
we will use the [Jupyter notebook](#) (comes with anaconda):

Easy to use

- On your computer
- On hosted service (eg: colab by google)

Notebooks easy to share:

- As notebooks
- As slides
- As html
- As pdf
- As .py files

Each cell can be run and produce output
cells types: text, plots, animations, "coding"

Jupyter notebook: where

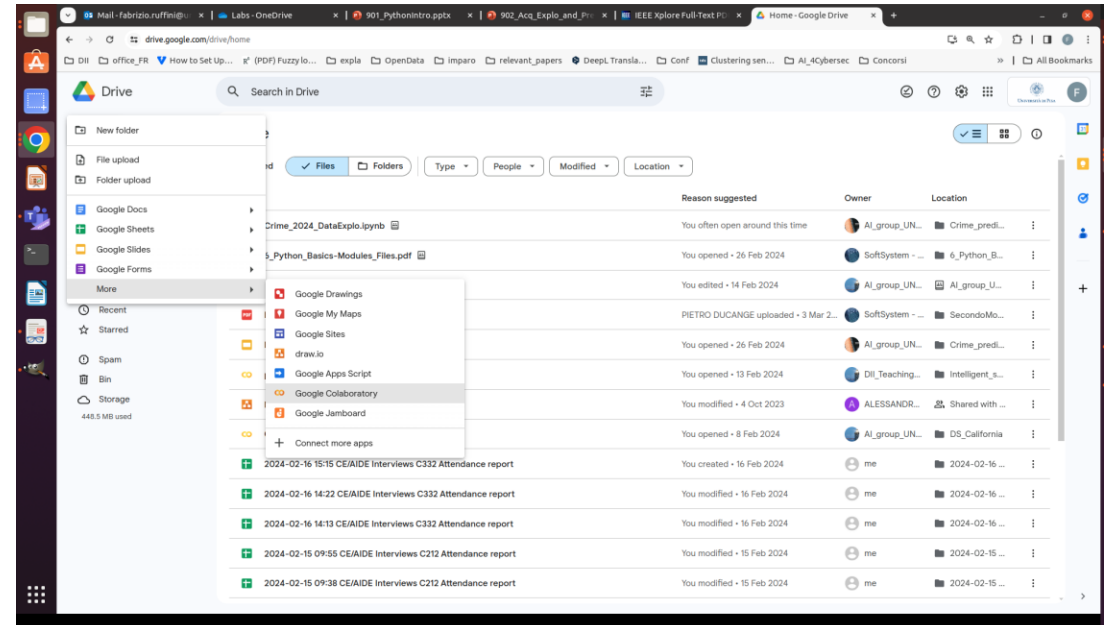
Local:

- You install, manage and control your stuff
- Resources depending on your PC

Google Colab

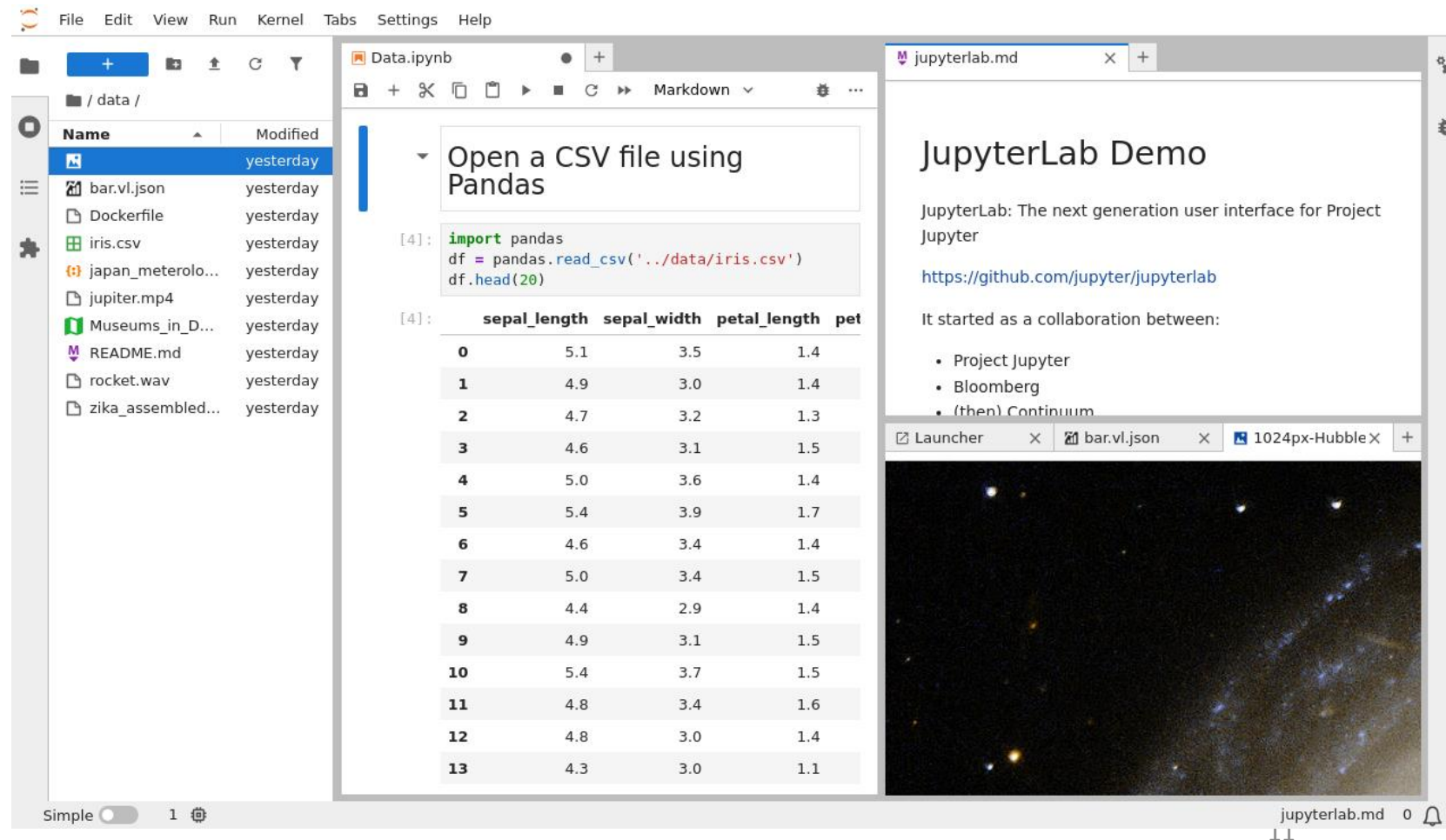
- They install, manage and control the Jupyter notebook. Many pkg already there. Files to be loaded on a mounted drive
- Access to (limited) resources. They can stop your process

<https://colab.research.google.com/>



JupyterLab

- Everything is under one roof
- Viewing CSV
- Second view for notebook
- Split view
- Rearranging cells
- Code consoles
- Preview for markdown



The screenshot displays the JupyterLab interface with three main panels:

- File Browser (Left):** Shows a directory structure with files like `bar.vl.json`, `Dockerfile`, `iris.csv`, `japan_meterolo...`, `jupiter.mp4`, `Museums_in_D...`, `README.md`, `rocket.wav`, and `zika_assembled...`. The `yesterday` folder is selected.
- Code Editor (Center):** Displays a notebook cell with the following code:

```
[4]: import pandas
df = pandas.read_csv('../data/iris.csv')
df.head(20)
```

The output shows a table of iris data with columns: `sepal_length`, `sepal_width`, `petal_length`, and `pet`.

	sepal_length	sepal_width	petal_length	pet
0	5.1	3.5	1.4	
1	4.9	3.0	1.4	
2	4.7	3.2	1.3	
3	4.6	3.1	1.5	
4	5.0	3.6	1.4	
5	5.4	3.9	1.7	
6	4.6	3.4	1.4	
7	5.0	3.4	1.5	
8	4.4	2.9	1.4	
9	4.9	3.1	1.5	
10	5.4	3.7	1.5	
11	4.8	3.4	1.6	
12	4.8	3.0	1.4	
13	4.3	3.0	1.1	
- Markdown Preview (Right):** Shows a preview of a markdown file titled `jupyterlab.md`. The content includes the title `JupyterLab Demo`, a description of JupyterLab, a link to the GitHub repository (<https://github.com/jupyter/jupyterlab>), and a list of collaborators: Project Jupyter, Bloomberg, and (then) Continuum. Below the text is a large image of a galaxy.

The bottom status bar shows the interface is in 'Simple' mode, with a tab count of 1 and a JupyterLab logo.

Focus on python: tutorials

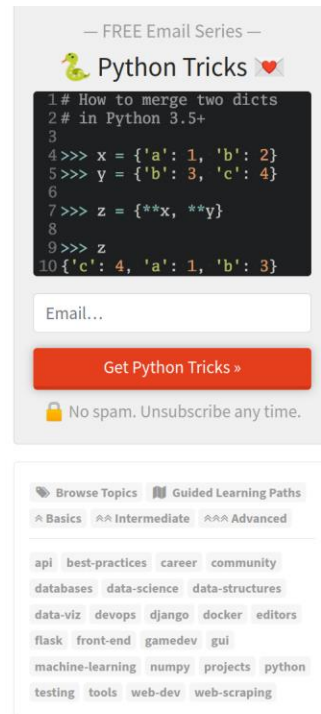
Real Python Tutorials



Python 3.12: Cool New Features for You to Try

In this tutorial, you'll learn about the new features in Python 3.12. You'll explore how the new release extends the better error messages and faster code execution found in the previous version, and you'll try out the improvements to f-strings and type variable syntax.

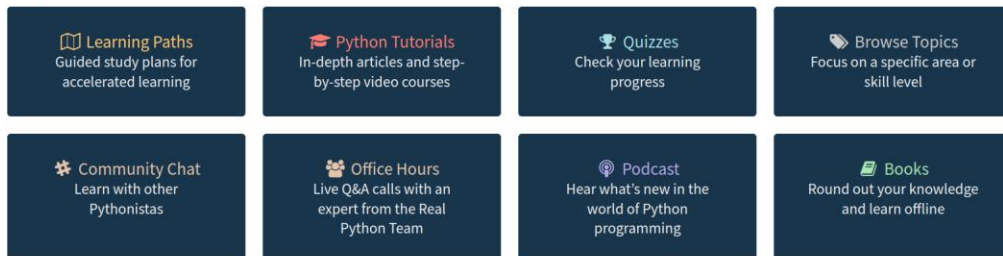
Oct 02, 2023 Intermediate python



Python tutorials:
<https://realpython.com/> :

- Basics
- Intermediate
- Advanced

Explore Real Python



... and many others available on the web




References

1. "Machine Learning and Security", C. Chio & D. Freeman (available @ biblio unipi, useful for final exam)
2. "Data Mining - Concepts and Techniques", Jiawei Han et. al, Morgan Kaufmann Publishers (fourth edition)
3. Clusit website <https://clusit.it/>


4. Python-related:
 - Python: <https://www.python.org/>
 - Anaconda: <https://www.anaconda.com/>
 - Python tutorials: <https://realpython.com/>

BACKUP

Using Anaconda on Windows

- Installation & Setup Problems
- **Not installing for "Just Me" vs. "All Users"**
 - Beginners sometimes select "All Users" without admin rights, causing installation errors.
 Fix: Choose *"Just Me"* unless you specifically need a system-wide installation.
- **Not adding Anaconda to PATH**
 - The installer warns against adding Anaconda to the system PATH. New users often do it anyway and break other Python installations.
 Fix: Don't tick the *Add to PATH* box — instead, use the **Anaconda Prompt**.
- **Conflicts with existing Python installations**
 - If Python is already installed (via python.org or Microsoft Store), Anaconda can cause version conflicts.
 Fix: Use Anaconda environments (`conda create -n myenv python=3.x`) instead of relying on the base environment.

Using Anaconda on Windows (2)

-  Usage Issues
- **Confusion between Anaconda Prompt, CMD, and PowerShell**
 - Beginners open CMD or PowerShell and type `conda`, which doesn't work unless the PATH is configured.
 - ✓ Fix: Always start with **Anaconda Prompt** (or enable Conda in PowerShell via `conda init`).
- **Navigator not launching**
 - Anaconda Navigator sometimes fails to open due to missing updates or corrupted configs.
 - ✓ Fix: Run `conda update anaconda-navigator` or delete the config files (`.conda` folder in the user directory).

Using Anaconda on Windows (3)

- Package & Jupyter Issues
- **Jupyter Notebook not found after install**
 - Sometimes Jupyter doesn't open or the command `jupyter notebook` isn't recognized.
 - ✓ Fix: Run it from **Anaconda Navigator** or reinstall with `conda install notebook`.