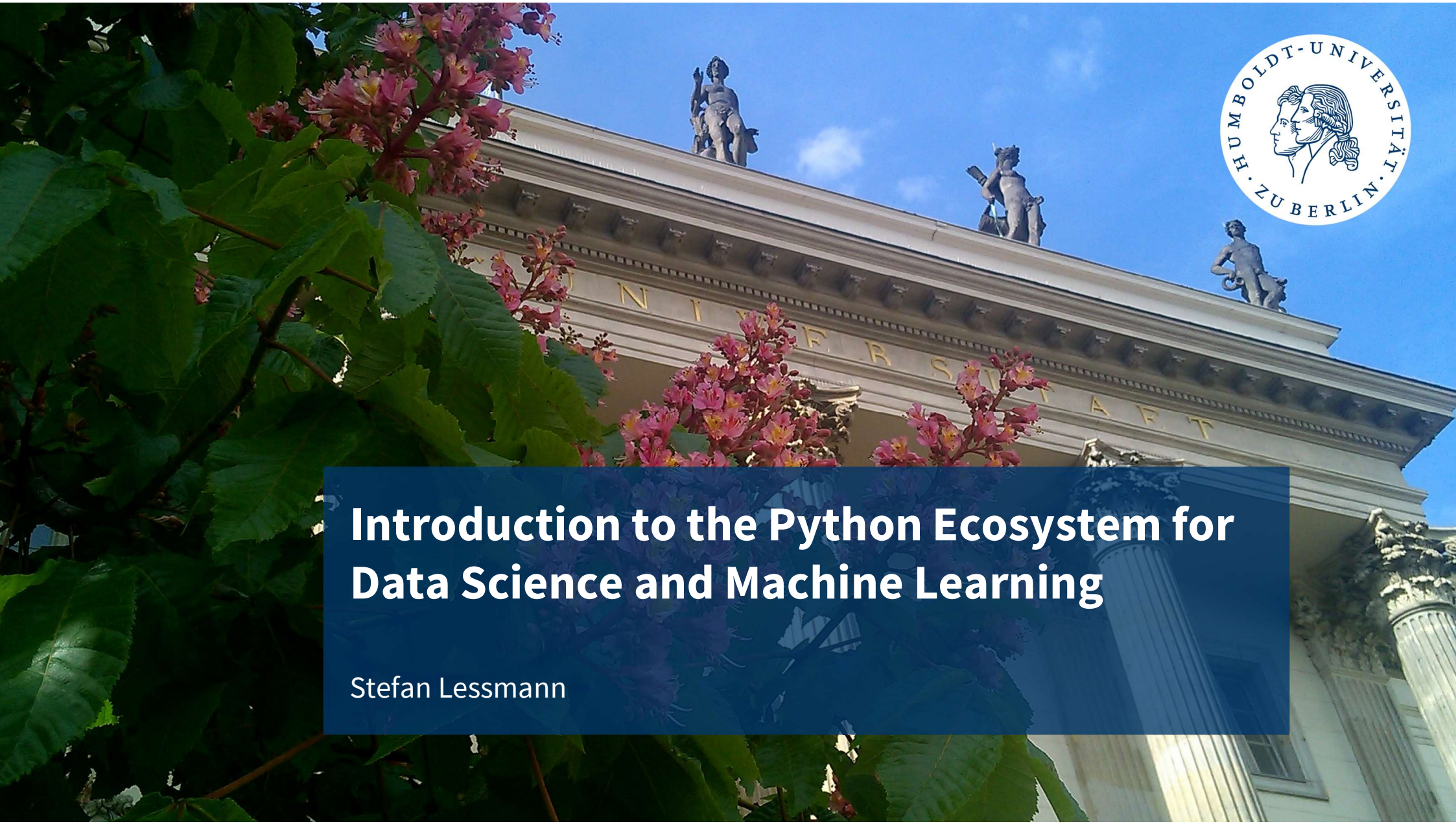
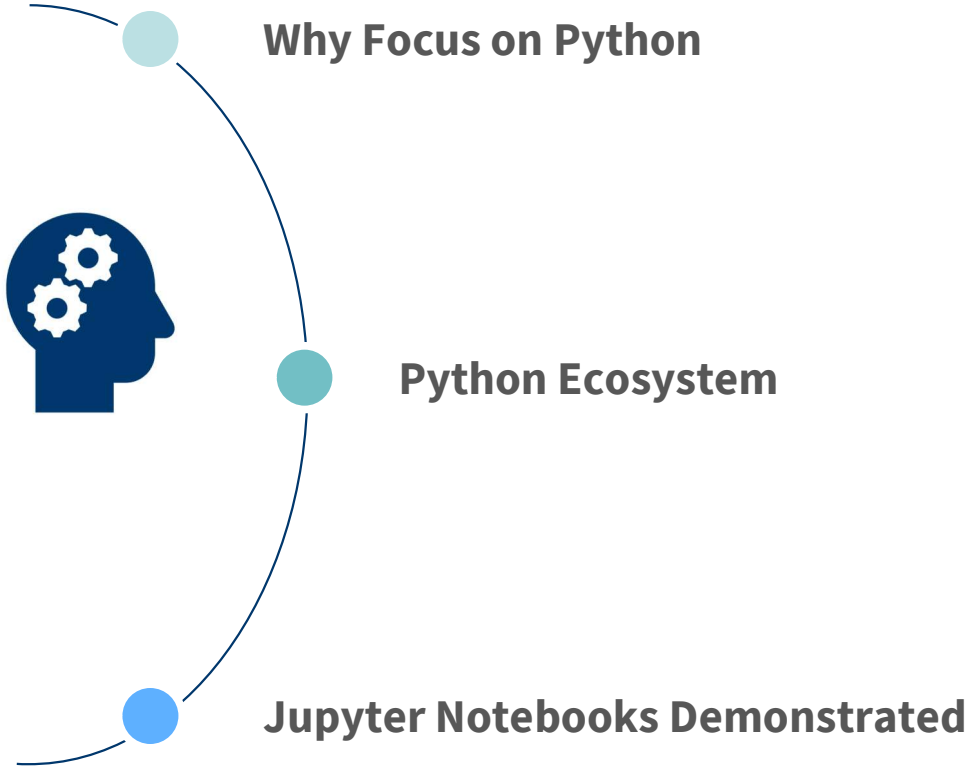


Introduction to the Python Ecosystem for Data Science and Machine Learning

Stefan Lessmann

Agenda

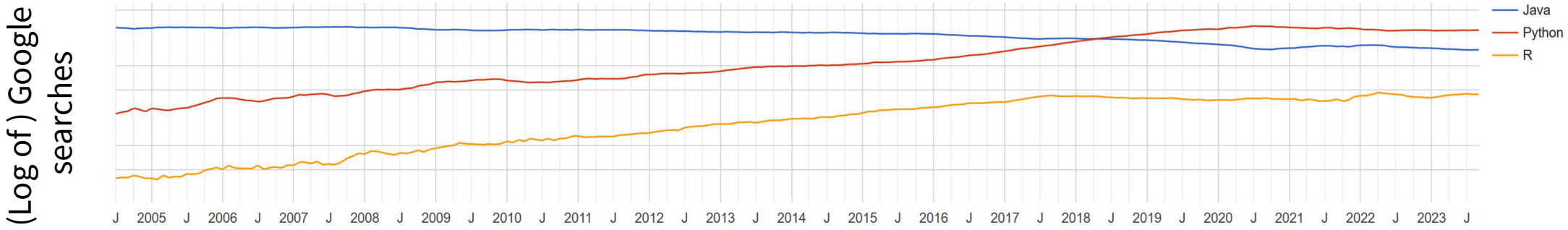


- Why Focus on Python
- Python Ecosystem
- Jupyter Notebooks Demonstrated

Why Focus on Python

Popularity of Programming Languages

PYPL PopularitY of Programming Language index

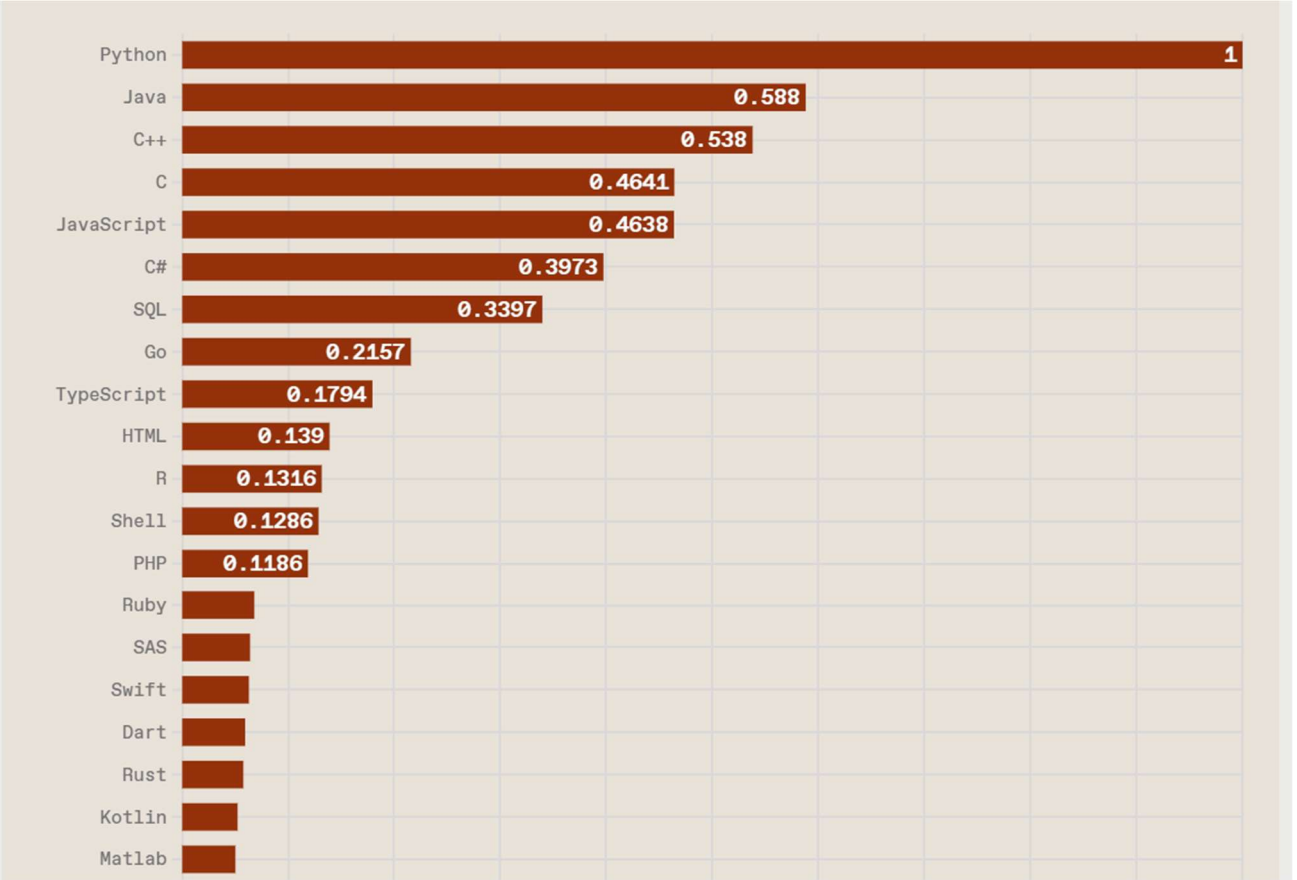


Source: https://pypl.github.io/PYPL.html

Popularity of Programming Languages

The Top Programming Languages 2023 - IEEE Spectrum

- **Possible reasons**
  - Powerful libraries for various purposes
  - Huge (supportive) developer community
  - Powerful tools to raise efficiency
  - Used by many enterprises
  - Extensively used in education and research
- **Speculation**
  - GenAI effect
  - Popularity of languages with high “web-visibility” likely to increase



Source: https://spectrum.ieee.org/the-top-programming-languages-2023

Future of Coding

Perhaps we don’t need Python in a low/no code world?

Andrej Karpathy

I like to train deep neural nets on large datasets

- **But we still need analytical thinking**
- **Also, LLM + Coding far more useful than LLM alone**
- **Some interesting opinions**
  - Michael Spencer on LinkedIn: What is the Future of Coding? | LinkedIn
  - Andrew Ng in The Batch, Oct. 4 2023: AI's New Power Couple, Movie Industry Limits AI, YouTube Goes Generative, More Web Data = More Bias

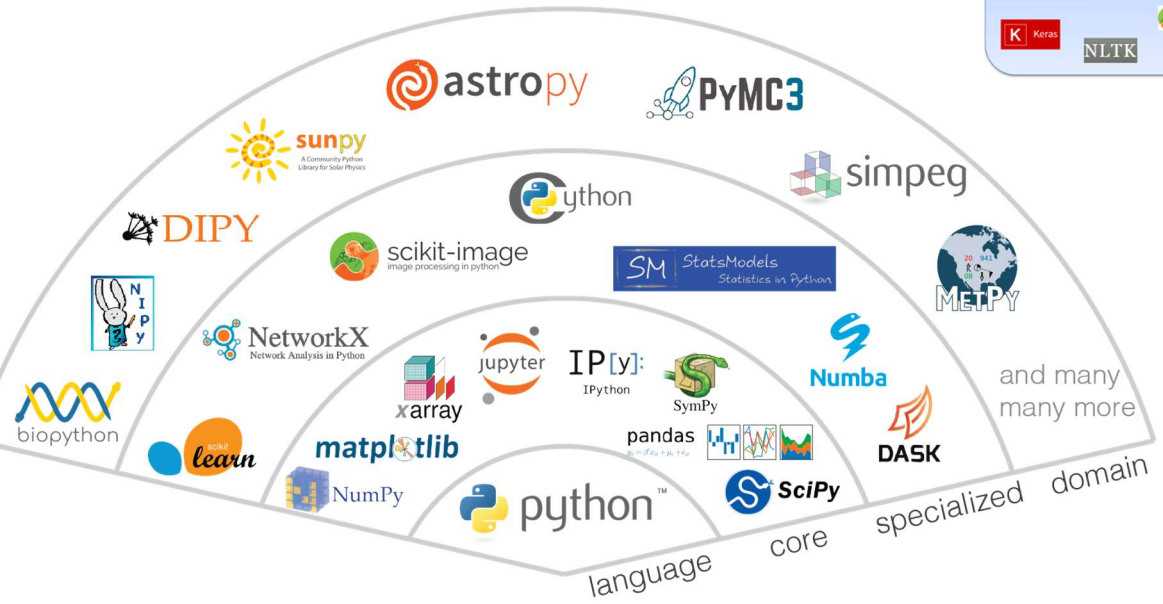
https://twitter.com/karpathy/status/1617979122625712128

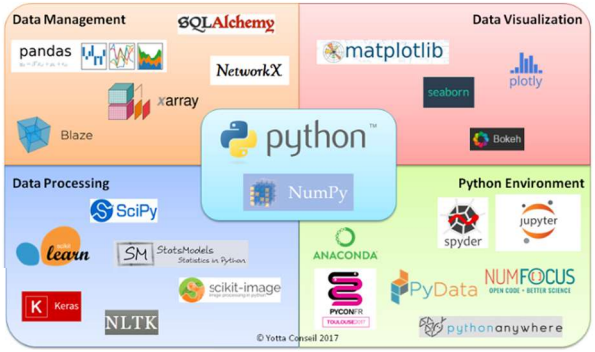
https://openai.com/blog/openai-codex

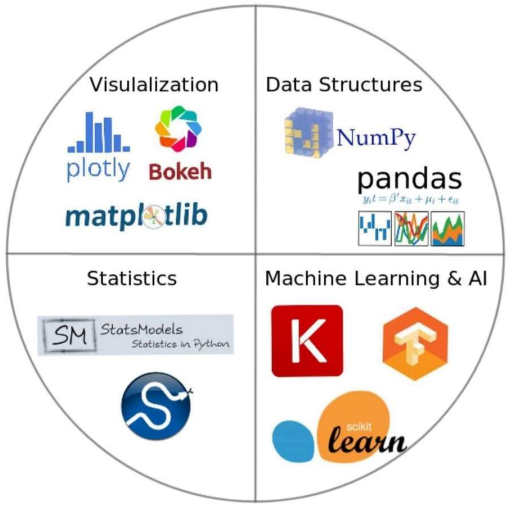
Sources:
(15) What is the Future of Coding? | LinkedIn
https://karpathy.ai/
https://nira.com/github-copilot/
AI's New Power Couple, Movie Industry Limits AI, YouTube Goes Generative, More Web Data = More Bias (deeplearning.ai)

Python Ecosystem

The Python Ecosystem







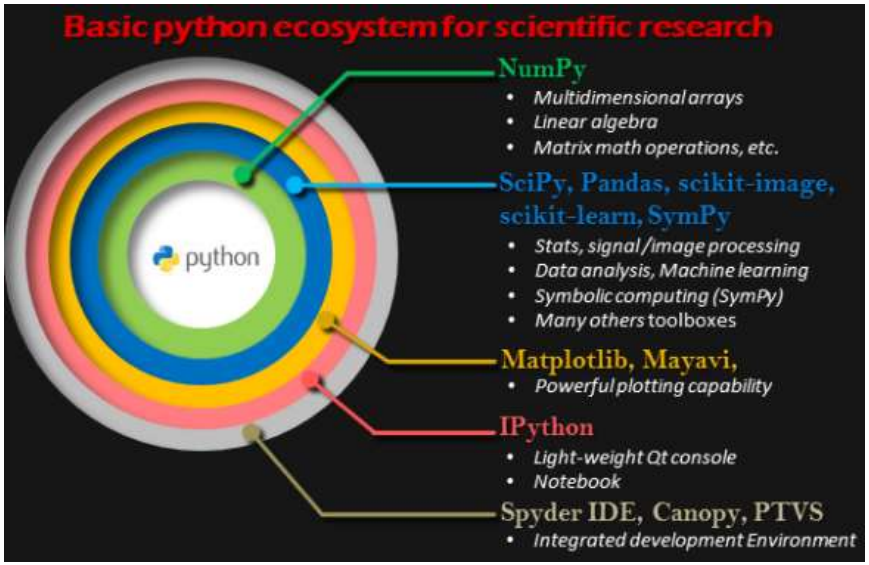
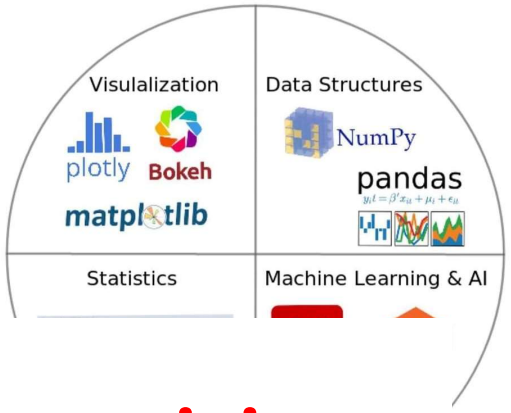


Image sources (left to right):
https://jupytearth.org/jupyter-resources/introduction/ecosystem.html
https://atrebas.github.io/post/2019-01-15-2018-learning/
https://www.facebook.com/megatekictacademy/photos/a.399385480230645/2266338440201997/?type=3
https://indranilsinharoy.com/2013/01/06/python-for-scientific-computing-a-collection-of-resources/

The Python Ecosystem



I know this looks very complicated, and to be honest, it is

complicated. But don‘t be overwhelmed!

We will introduce tools / technologies slowly and selectivly.

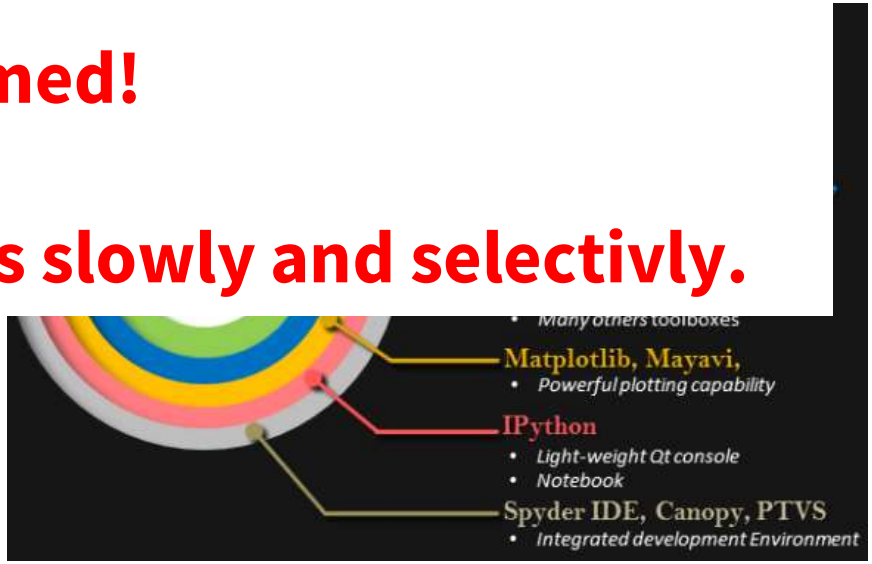


Image sources (left to right):
https://jupytearth.org/jupyter-resources/introduction/ecosystem.html
https://atrebas.github.io/post/2019-01-15-2018-learning/
https://www.facebook.com/megatekictacademy/photos/a.399385480230645/2266338440201997/?type=3
https://indranilsinharoy.com/2013/01/06/python-for-scientific-computing-a-collection-of-resources/

The Python Ecosystem
Why Python is so popular

- **Programming language is the core**
  - Defined syntax, set of instructions, data types, etc.
  - Tools to translate Python code into machine readable format
  - Just like any other programming language
- **Auxiliary layers make Python power-full and the first choice for data science**
  - Working with arrays (NumPy)
  - Visualization (Matplotlib, seaborn, …)
  - Working with (relational) data (Pandas)
  - ML/DL algorithms (sklearn, tensorflow, Pytorch)
  - Environment for creating computational essay (i.e., notebooks)

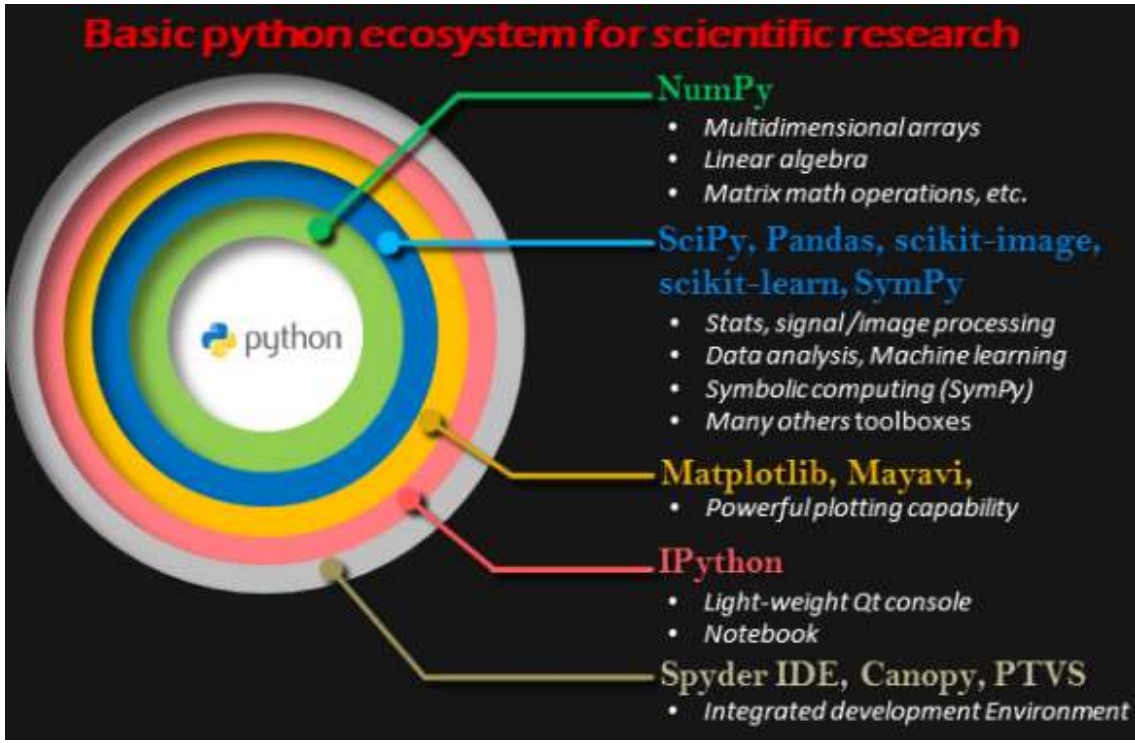


Image source:
https://indranilsinharoy.com/2013/01/06/python-for-scientific-computing-a-collection-of-resources/

And what About…

Indeed, we see many similarities between R and Python in terms of their features.

Yet, Python has an important advantage over R when it comes to running code in production.

Jupyter (IPython) Notebooks

Very similar to R Markdown (should you know it)

- **Environment that integrates (Markup) Text and Python codes**
  - Basic functionality to format and organize text
  - Functionality to execute programming codes
  - Code output is directly integrated into your notebook
- **Use cases**
  - MANY, but typically in education and research
    - Exercises in a lecture: you receive a notebook with verbal task descriptions and write the programming codes to perform these tasks
    - You write a seminar/bachelor thesis and develop a (or multiple) notebook(s) for the empirical experiments
    - You write a blog post about a research paper, new ML algorithm, etc.
  - Prototyping
- **Notebooks are not meant to write code for production**

Many Other Essential or At Least Useful Tools to Master

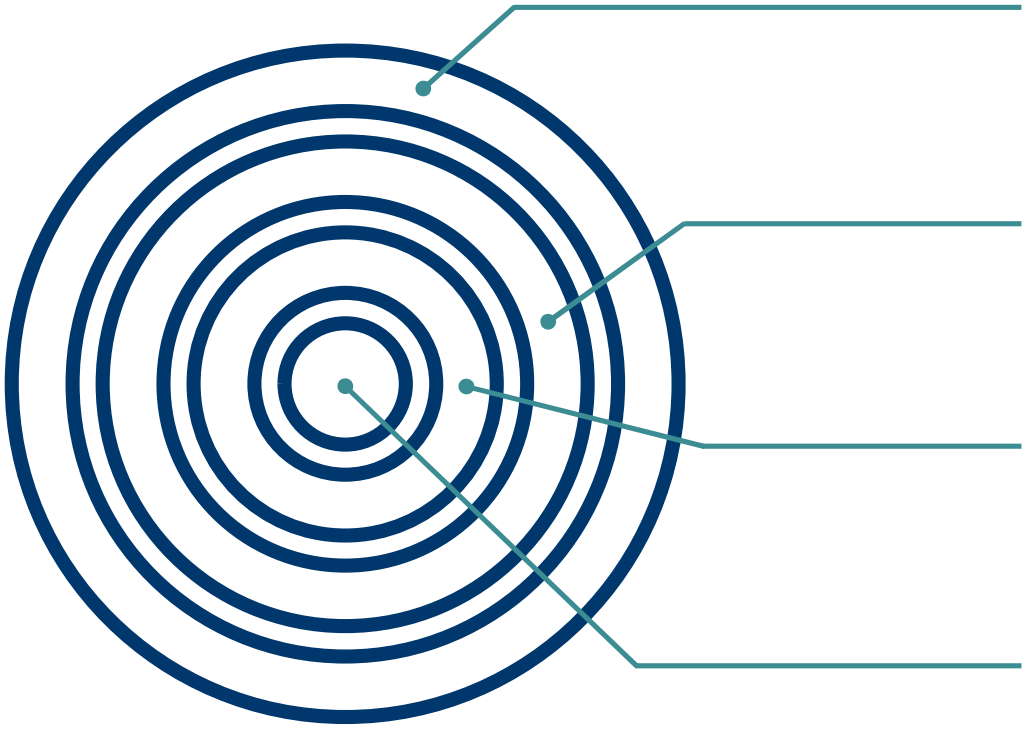
Take this as a Glossary. You will come across these terms. Then look here!

- **Virtual environments**
  - A sandbox for individual projects
  - Libraries and Python itself get updated from time to time
    - Installing newer version of e.g. a library may break code you once write
    - Conflicts between version X of library A and version Y of library B are also possible
- **Package Managers**
  - Tools to manage virtual environments
    - Create, environments, install libraries to an environment, update libraries in an environment, etc.
    - Check dependencies between libraries
  - Common choices for Python include *conda* and *pip*
- **Python distributions**
  - Pre-packaged set of Python + a set of libraries that are often used together in specific contexts
  - Common choice for machine learning context is *Anaconda*
- **Other tools:** code version control, collaboration, project management, deployment, etc.

Jupyter Notebooks Demonstrated

Jupyter Notebooks vs Python?

Notebooks are a part of the Python data science ecosystem. They are a front-end tool and facilitate both, the writing of code and the presentation of results.



Front-end tools

- Jupyter, IPython notebooks
- Other IDEs (e.g., Spider, VS Code, PyCharm, …)

Libraries for specialized purposes

- Visualization, machine Learning, NLP, …
- Matplotlib, seaborn, sklearn

Libraries for scientific computing

- Matrices, tensors, data containers, etc.
- NumPy, Pandas

Python programming language

- Python interpreter
- Core libraries

Ways to Use and Interact with Notebooks

Many choices... which is best for you?

Create a local environment

- Install required software (all free) on your computer
- Full flexibility but will cost you some time

- **Option 1: Anaconda distribution**
  - You download Anaconda (https://www.anaconda.com/)
  - This gives you almost all you need
  - You work directly with Jupyter
- **Option 2: Integrated development environment (IDE)**
  - Proper – heavyweight – programming tool (e.g., Eclipse)
  - Popular choices for Python programming include Visual Studio Code, PyCharm, and Spider
  - These tools integrate with Jupyter and facilitate writing Jupyter notebooks

Use a cloud solution

- No need to install anything. Only need a web-browser. Codes run on server.
- Upload of data sets, demo notebooks, etc. can be cumbersome

- **Option 1: Google Colab** (https://colab.research.google.com/)
  - You need a Google account. Upload of resources will then work best via GDrive
  - Simplest solution, but you depend on Google
  - Other options are available (Kaggle, Amazon, …) but have no general advantages IMHO
- **Option 2: HUB JupyterHub** (https://jupyterhub.cms.hu-berlin.de/)
  - Most privacy friendly solution
  - You have access using your HU Account
  - Upload of resources a bit cumbersome
  - Reliability not yet at 100%
    - Might be down for maintenance during tutorial
    - Might struggle with high load in large courses
    - WiFi connectivity at HUB WiWi can also be an issue

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742
Fax. +49.30.2093.5741

stefan.lessmann@hu-berlin.de
https://www.linkedin.com/in/stefanlessmann/

www.hu-berlin.de