Introduction to Python

Exercise Session 01 - ESC403.1 Introduction to Data Science

Thursday, February 25, 2021

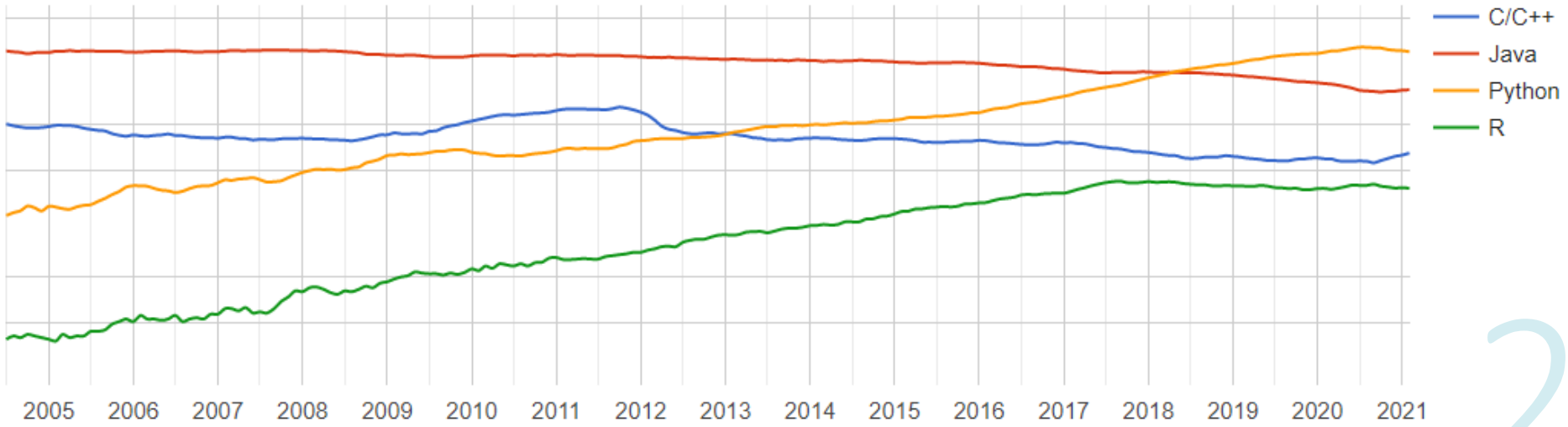
Python language

Python is **general purpose**, **object oriented** and **open source** language.

Includes a broad range of libraries which enables its **versatility** in many fields:

1. Data science
2. Neuroscience
3. Economics

Python is **extendable** since any programmer can add low level modules and it is **scalable** to support large programs.



from PopularitY of Programming Languages (https://pypl.github.io/PYPL.html)

Why Python?

- Python's code is readable and maintainable

- It is a cross – platform and cross - system language

  a developer could run the same compiled code in different platforms without recompilation

- Python offers a robust standard library

  a developer could select specific modules according to his precise needs and could add further functionality

- Supports multiple frameworks

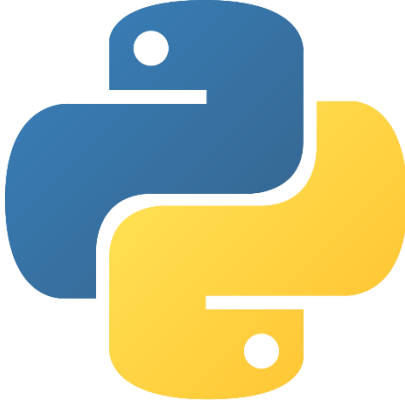
  i.e web frameworks (Django, Flask etc.)

- Simplifies complex scientific and software development applications

  the developers could take advantage of data analysis features (effective visualization)

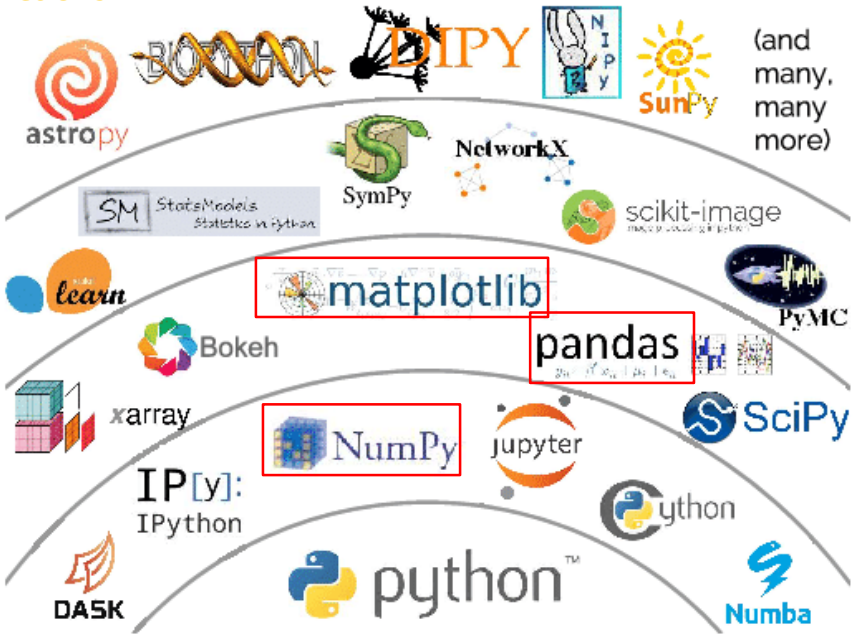
  using Python libraries written in faster languages (i.e. C), the developers analyze the data rapidly

Why Python for Data Science?



Divisions for Data Science (Prof. Feldmann – Lecture1)

1. Data Exploration and Preparation → Data access and cleaning
2. Data Representation and Transformation → Data are represented using python structures (arrays, dictionaries, tuples etc.)
3. Computing with Data → Complex numerical computations using built-in functions
4. Data Modeling → Can infer/predict statistical relationships on the data
5. Data Visualization and Presentation → Interactive visualization
6. Science about Data Science → Insightful reporting/presentation
   Big community



Jake VanderPlas's presentation at the PyCon 2017 conference

Python entails all the data science aspects

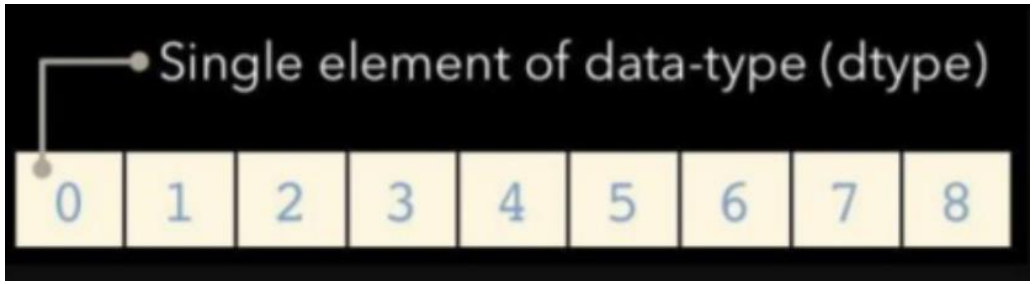
In what sense?

How is that achieved?

NumPy

What is it?

- NumPy is a Python/C library used for array-oriented computing.
- Key component: ndarray object with homogeneous data types (dtype) → operations are performed in compiled code for performance



Why NumPy?

- It is suited to many applications (image/signal processing, linear algebra etc.)
- 50x faster than Python lists
- Concise and quick computations by vectorization

How to use it?

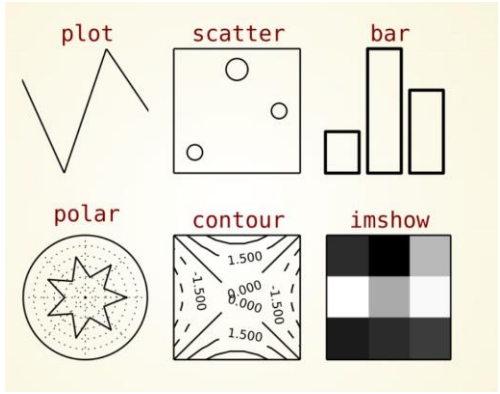
```
In [1]: import numpy as np
```

More in the hands on session …

matplotlib

What is it?

- matplotlib is a visualization library built on NumPy arrays
- includes several types of visualization styles



Why matplotlib?

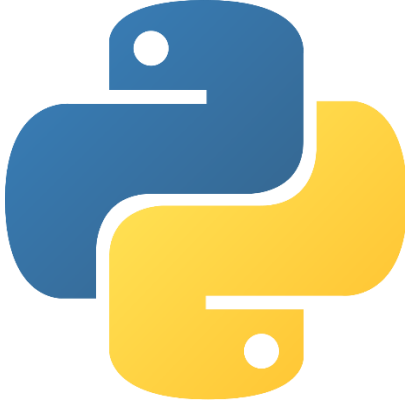
- It is compatible with many python tools and libraries allowing for advanced visualization
- offers interactive, static and animated visualizations → effective data presentation

How to use it?

```
In [1]: import matplotlib.pyplot as plt
```

More in the hands on session …

pandas



What is it?

- pandas is a data analysis library
- it comes from the abbreviation of panel data system

Data Structure	Dimensions	Description
Series	1	1D labeled homogeneous array, sizeimmutable.
Data Frames	2	General 2D labeled, size-mutable tabular structure with potentially heterogeneously typed columns.
Panel	3	General 3D labeled, size-mutable array.

Why pandas?

- it is the ideal tool for managing/cleaning/analyzing/modelling and organizing the data
- can read data from various formats including csv, txt, tsv etc.
- enables big data analysis by including SQL functionalities

How to use it?

```
In [ ]: import pandas as pd
```

More in the hands on session …

Questions?

Contact us:

Vasilis Dimakopoulos: vasileios.dimakopoulos@usz.ch
Elia Cenci: elia.cenci@uzh.ch
Darren Reed: darren.reed@uzh.ch