Prof. Dr. Jan Kirenz

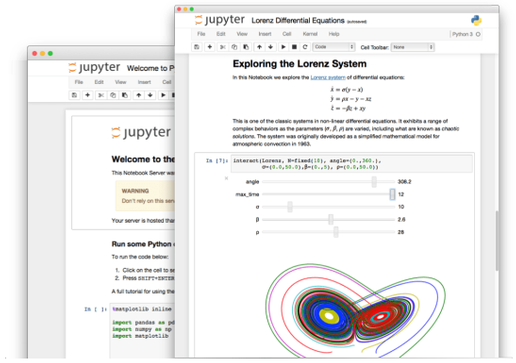
Python

Picture Credit: Google

Python

Programming Language

Python is an object-oriented language (an object is an entity that contains data along with associated metadata and/or functionality). One thing that distinguishes Python from other programming languages is that it is interpreted rather than compiled. This means that it is executed line by line which is particular useful for data analysis, as well as the creation of interactive, executable documents (VanderPlas, 2016). On top of this, there is a broad ecosystem of third-party tools and modules that offer more specialized data science functionality (like Scikit-Learn, which provides a toolkit for applying machine learning algorithms to data).



Jupyter Notebook: One important third-party tool for data science is the Jupyter Notebook, an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text (Perkel, 2018).

Anaconda is a free and open-source distribution of the Python programming language that aims to simplify package management and deployment. It already contains Jupyter Notebook and other important data science modules.

Get familiar with Colaboratory (also known as Colab), Google's free Python and Jupyter based cloud service. You don't need to install any software.

Install Anaconda (select the current version of Python 3). After installation, launch the Anaconda Navigator and start Jupyter Notebook or Jupyter Lab.

Recommended reading:

- Pandey, P. (2018). Bringing the best out of Jupyter Notebooks for Data Science. Towards Data Science
- Pandey, P. (2019). Jupyter Lab: Evolution of the Jupyter Notebook. Towards Data Science
- Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. Nature, 563(7729), p. 145.
- VanderPlas, J. (2016). Whirlwind Tour of Python. O'Reilly Media.

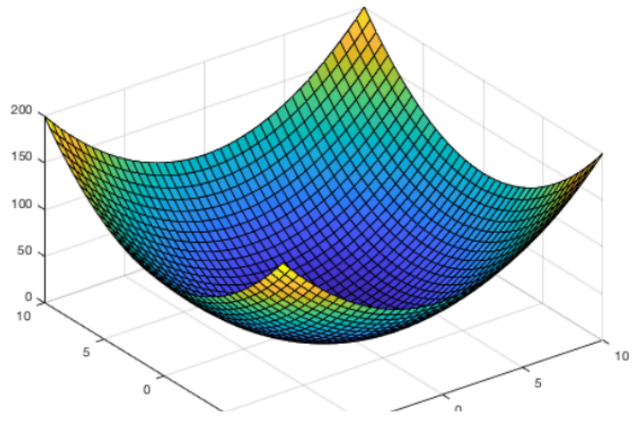
Python
Modules & Resources

There exists a broad ecosystem of third-party tools and modules that offer specialized (data science) functionality. Some of the most important for data science tasks are:

- **NumPy** provides efficient storage and computation for multidimensional data arrays.
- **Pandas** provides a DataFrame object along with a powerful set of methods to manipulate, filter, group, and transform data.
- **SciPy** contains a wide array of numerical tools such as numerical integration and interpolation.
- **Statsmodels** is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. Furthermore, you can use R-style formulas together with pandas data frames to fit your models.
- **Matplotlib** provides a useful interface for creation of publication-quality plots and figures.
- **Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **Scikit-Learn** provides a uniform toolkit for applying common machine learning algorithms to data.

Recommended reading and resources:

- Petrou, T. (2019). Minimally Sufficient Pandas. Medium
- Pandas Cheat Sheet
- Varoquaux, G. et al. (2017). Scipy Lecture Notes. Getting started with Python for science. Whitepaper



Data Science Concepts, Ideas, and Codes

- Towards Data Science
- Kaggle
- Stack Overflow

Newsletter:

- Data Science Weekly
- Data Elexir
- Data Machina

Datasets:

- Awesome Datasets