M1b: Tools

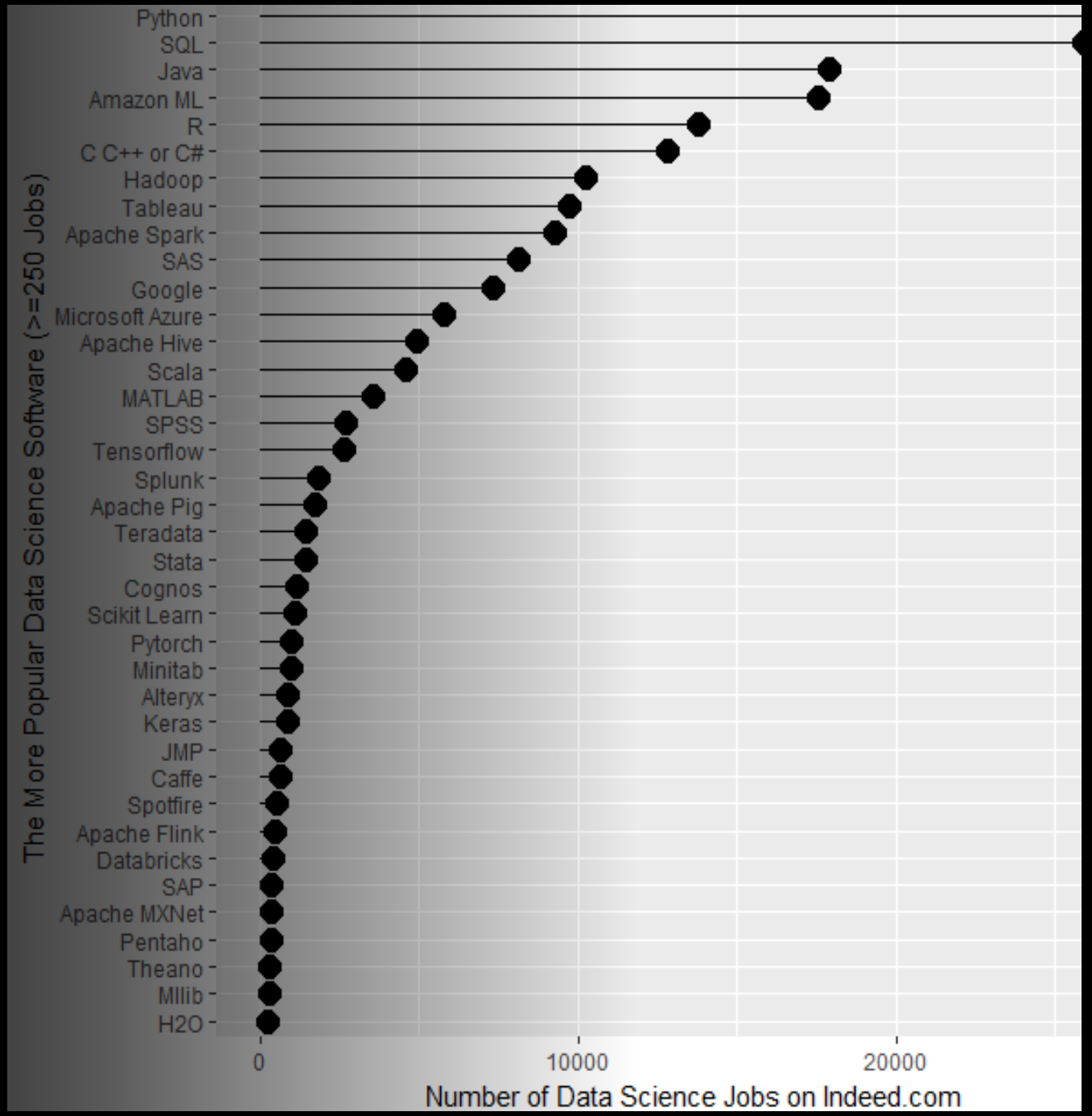
DSE 10200: Introduction to Data Science

Instructor: Michael Grossberg

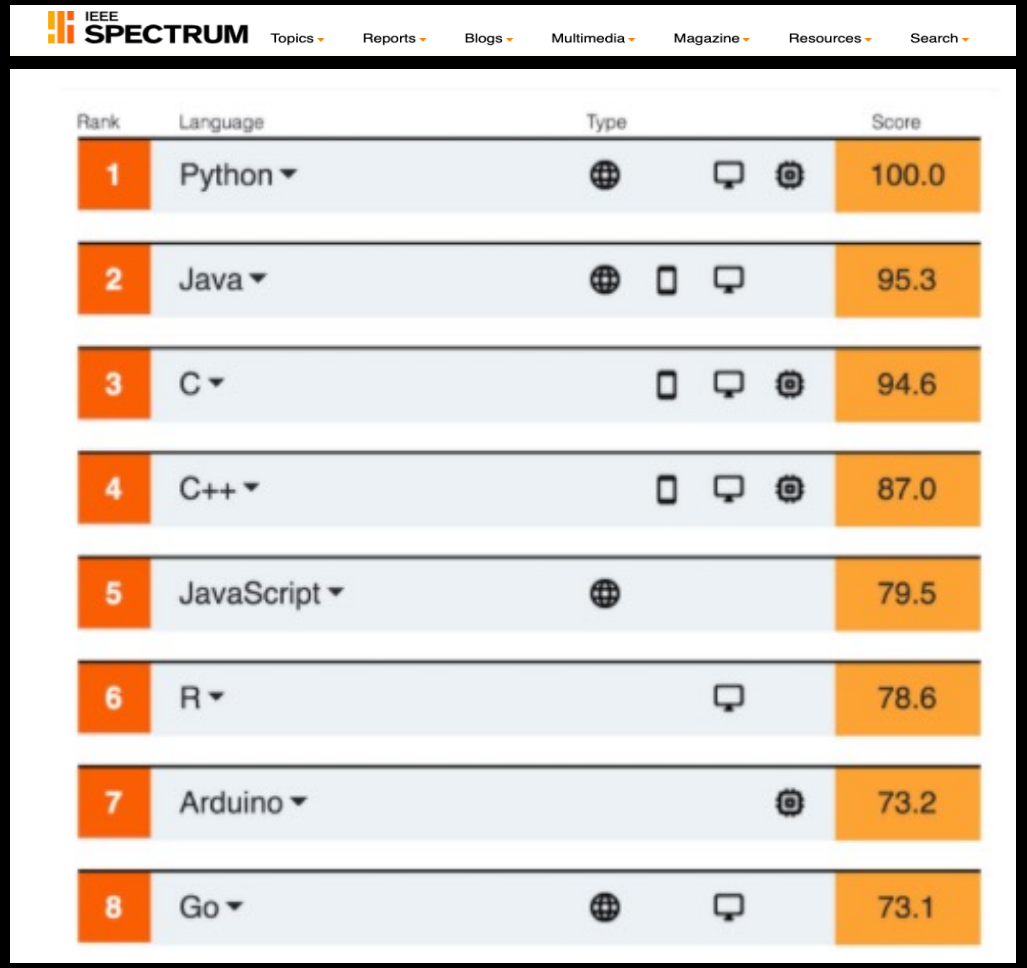
python

Why Python?

Python Most Popular Data Science Language



Python most popular programming language



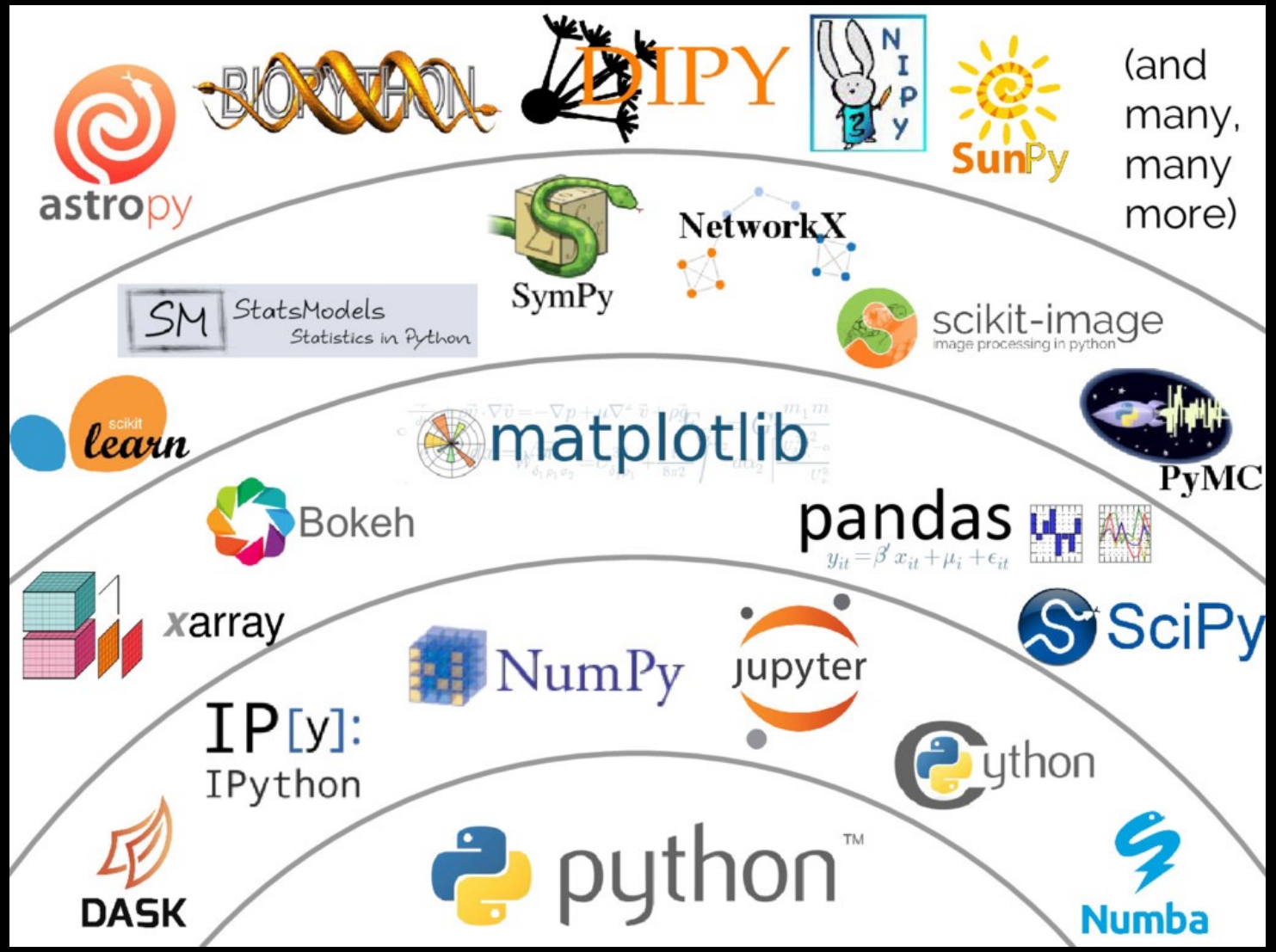
IEEE SPECTRUM Topics Reports Blogs Multimedia Magazine Resources Search

Rank	Language	Type	Score
1	Python		100.0
2	Java		95.3
3	C		94.6
4	C++		87.0
5	JavaScript		79.5
6	R		78.6
7	Arduino		73.2
8	Go		73.1

Top Programming Languages 2020

Python rules the roost, but Cobol gets a pandemic bump

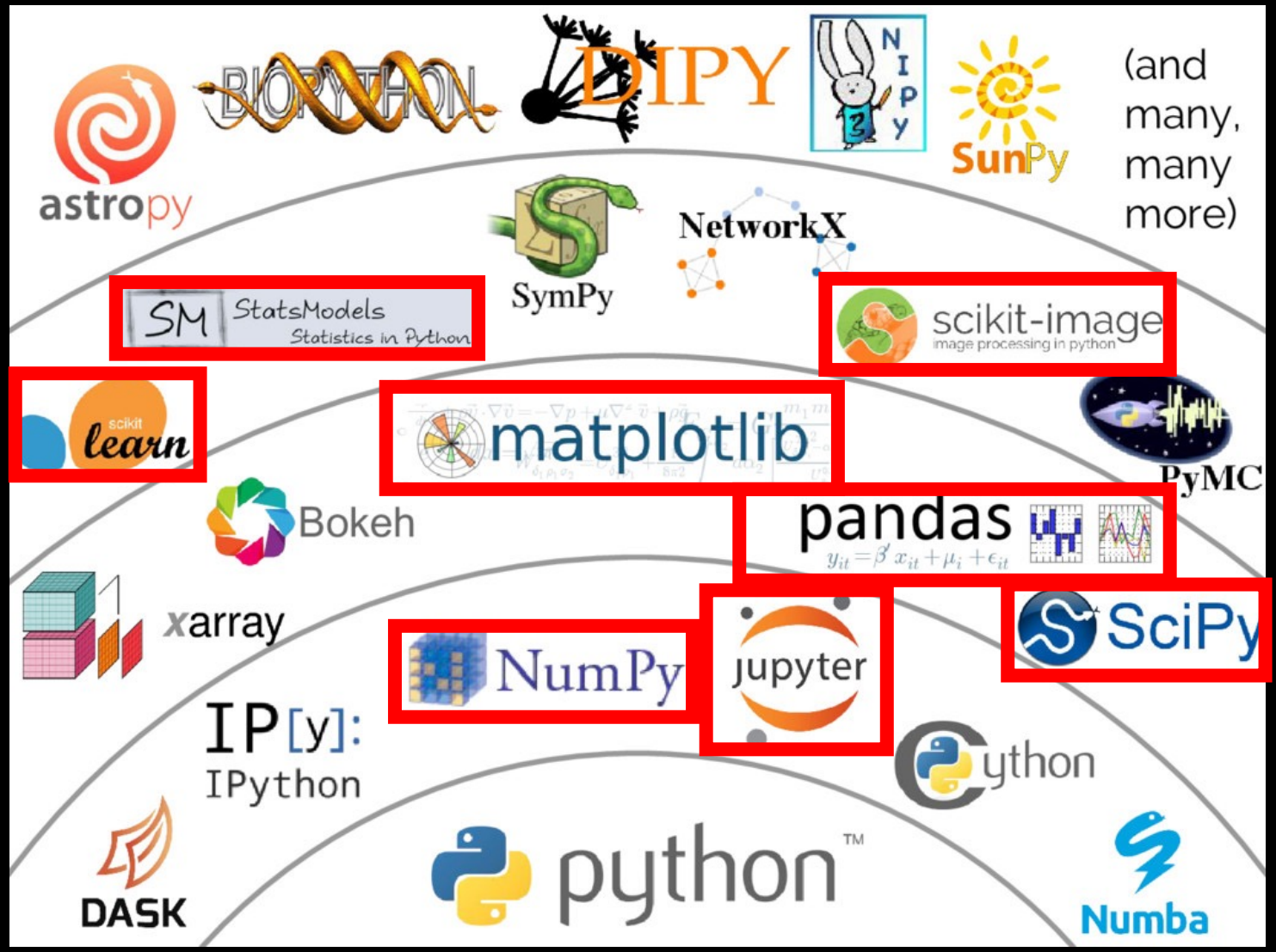
By **Stephen Cass**



Python as MATLAB Replacement

Libraries

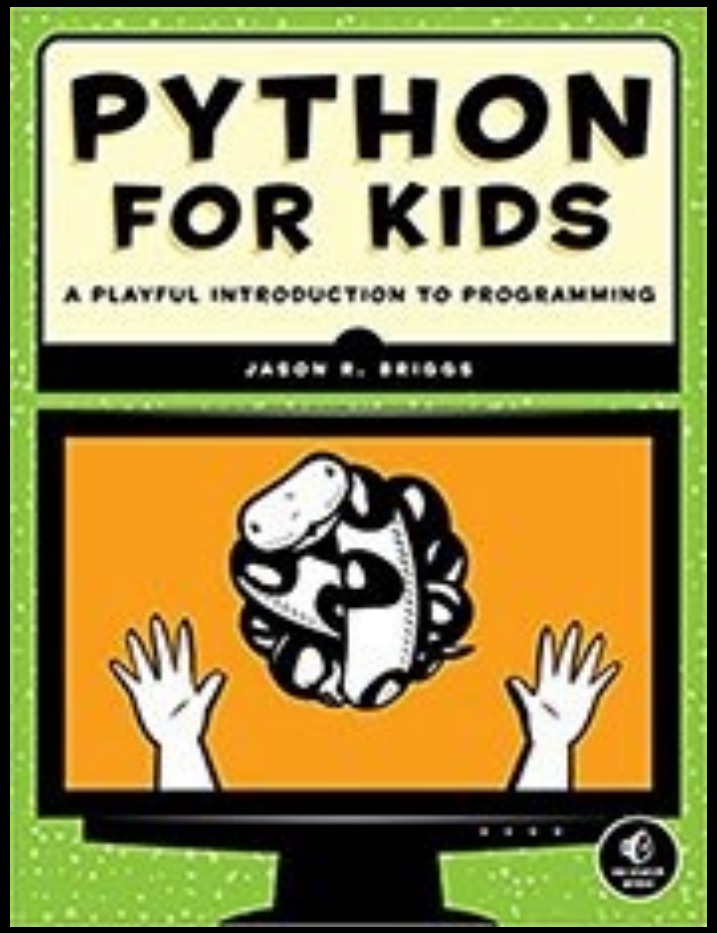
like MATLAB toolboxes



Python as MATLAB Replacement

Libraries

like MATLAB toolboxes

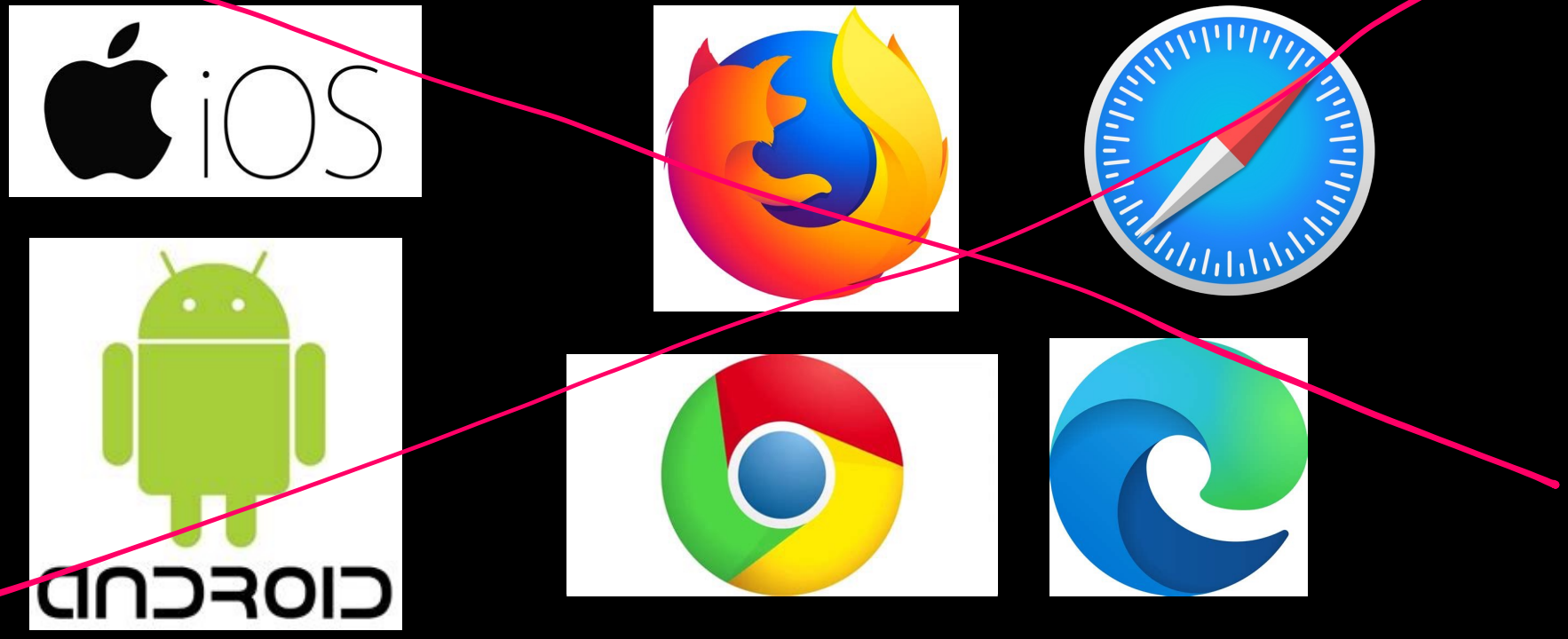


Taught in elementary, middle schools and high schools

Python Easy (easier) to learn

Is python the best language for everything?

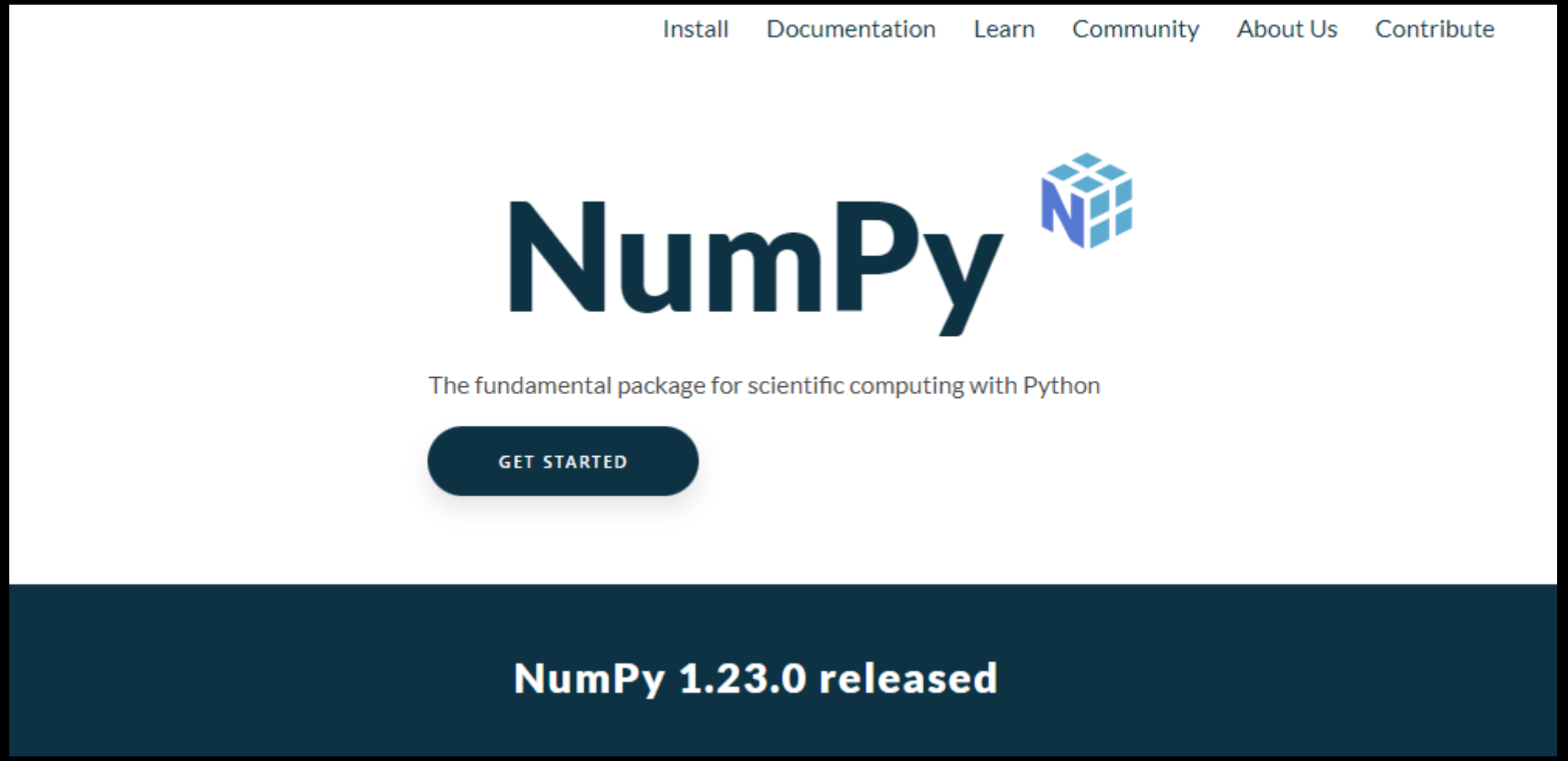
Is python the best language for everything?



Mobile and Web => JavaScript Better

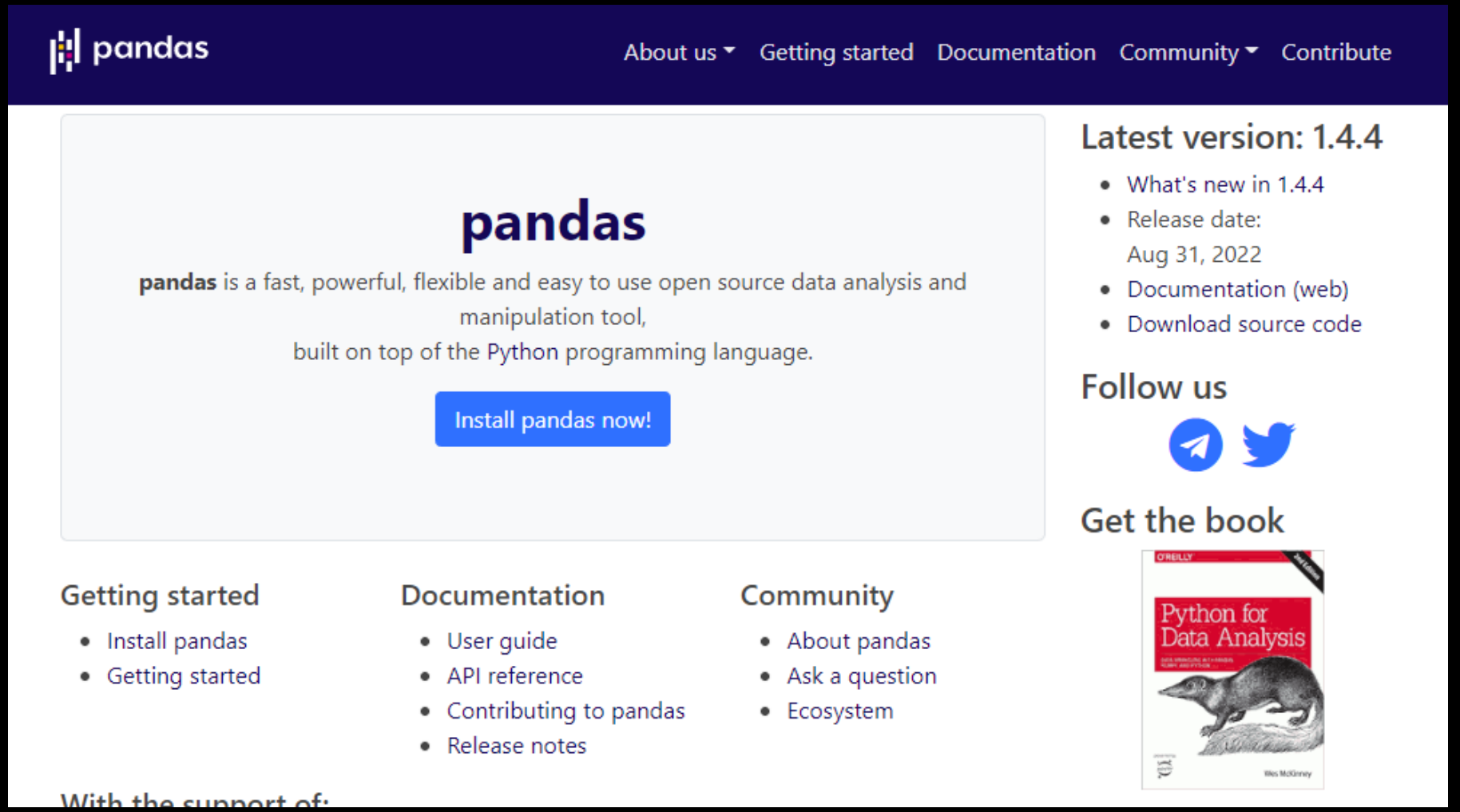
Python Pretty Good at Most Things

Math Library



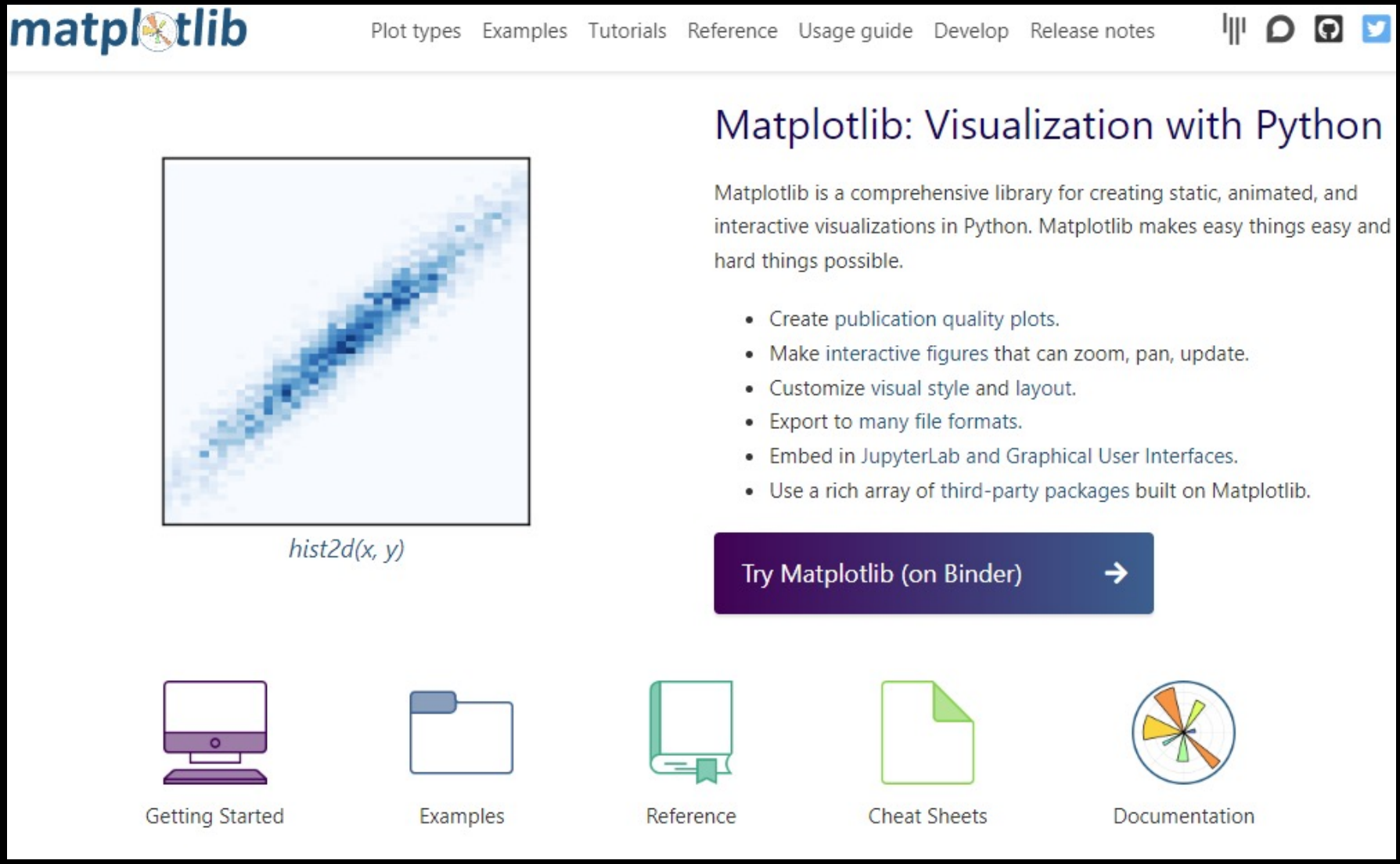
https://numpy.org/

Pandas



https://pandas.pydata.org/

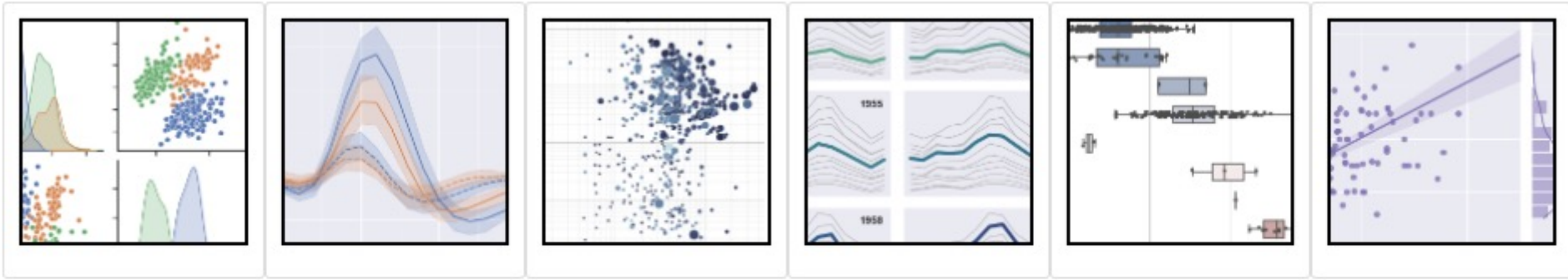
Matplotlib



https://matplotlib.org/

Seaborn

seaborn: statistical data visualization



Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see some of the things that you can do with seaborn, and then check out the tutorial or API reference to find out how.

To see the code or report a bug, please visit the GitHub repository. General

Contents

- Introduction
- Release notes
- Installing
- Example gallery
- Tutorial
- API reference

Features

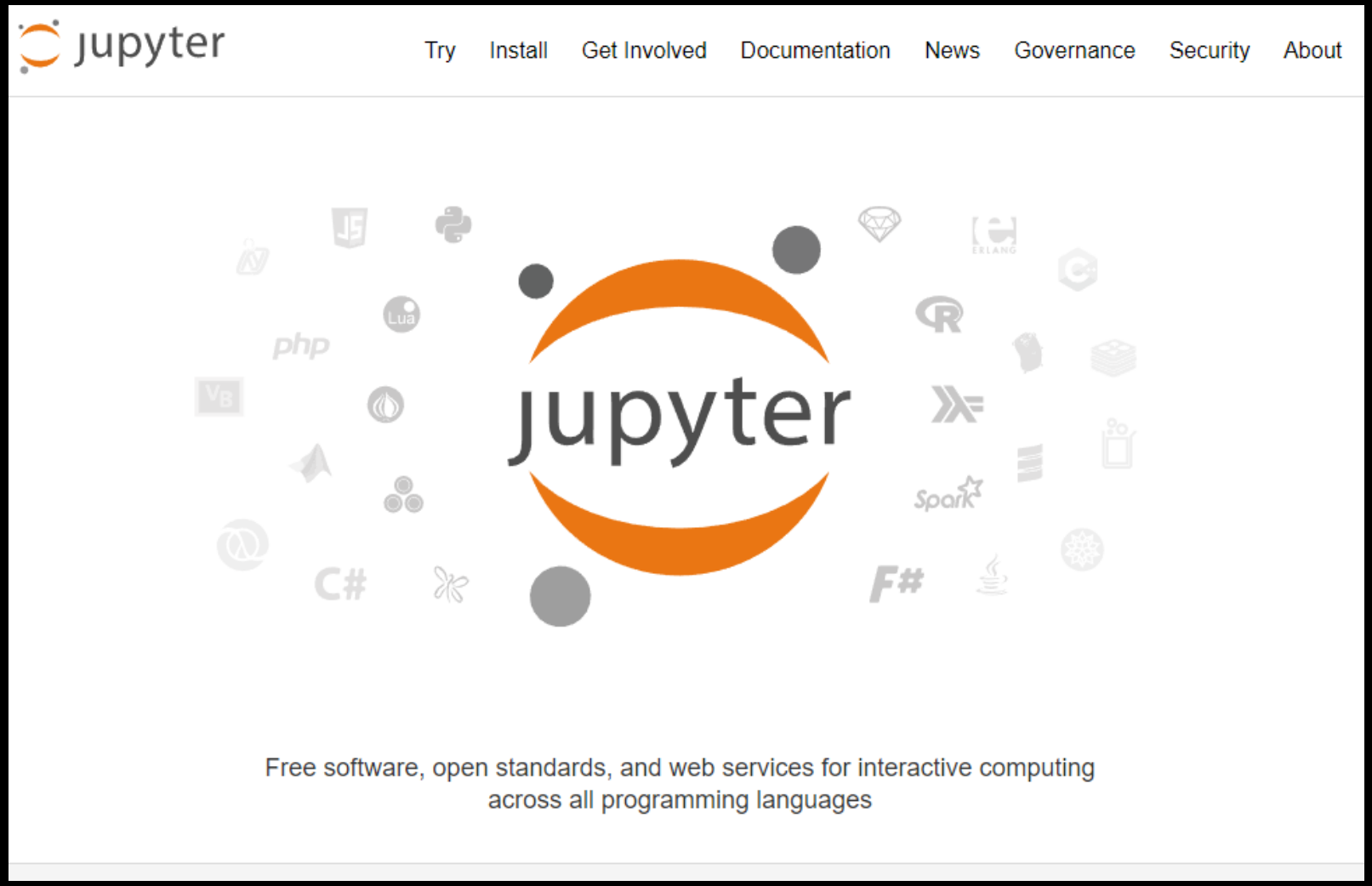
- Relational: API | Tutorial
- Distribution: API | Tutorial
- Categorical: API | Tutorial
- Regression: API | Tutorial
- Multiples: API | Tutorial
- Style: API | Tutorial
- Color: API | Tutorial

https://seaborn.pydata.org/

Sklearn

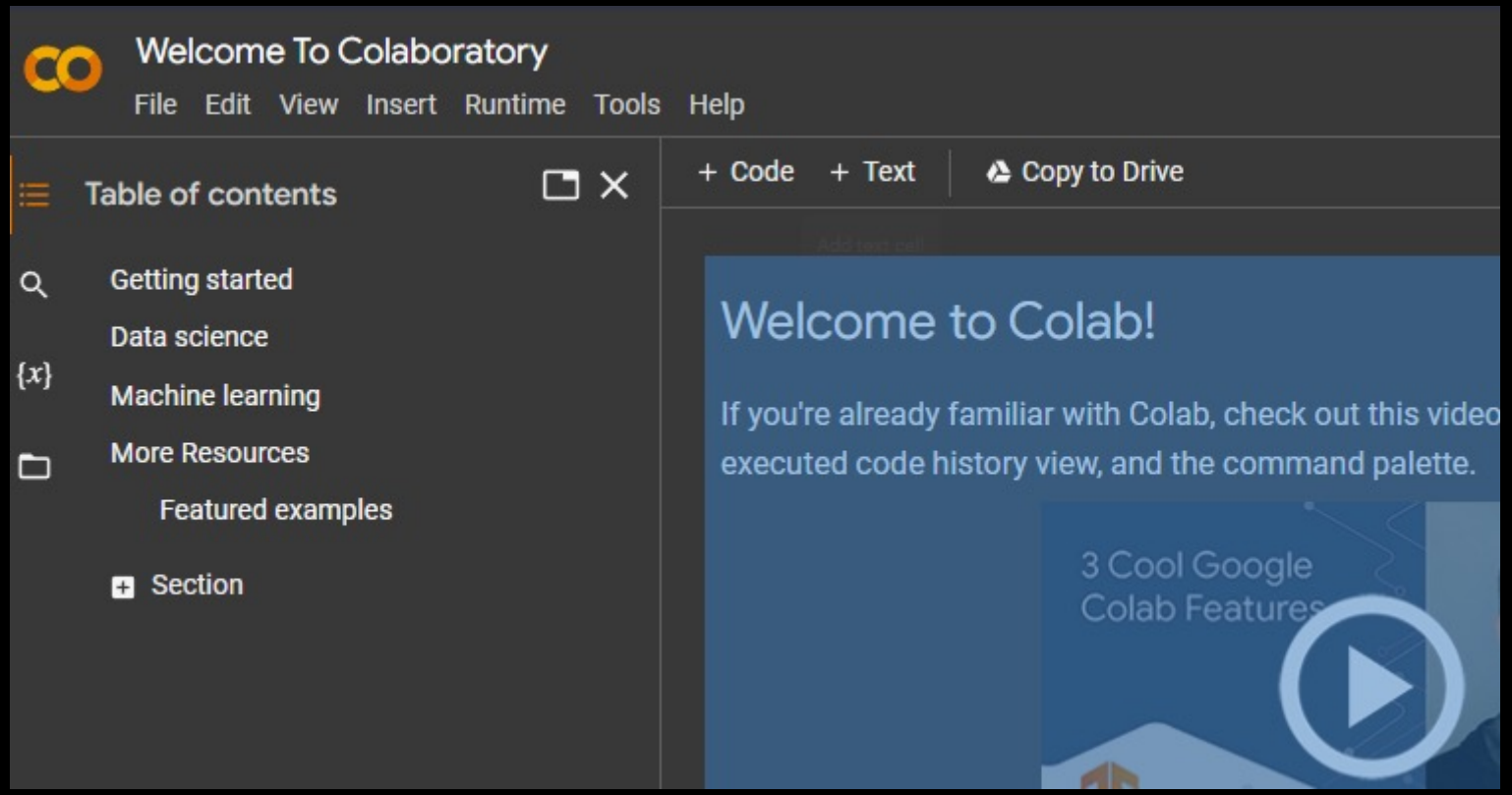

scikit learn
Install
User Guide
API
Examples
Community
More
Go
scikit-learn
Machine Learning in Python
Getting Started
Release Highlights for 1.1
GitHub
Simple and efficient tools for predictive data analysis
Accessible to everybody, and reusable in various contexts
Built on NumPy, SciPy, and matplotlib
Open source, commercially usable - BSD license
Classification
Identifying which category an object belongs to.
Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...
Regression
Predicting a continuous-valued attribute associated with an object.
Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...
Boosted Decision Tree Regression
training samples
n_estimators=1
n_estimators=300
Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...
K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross

Jupyter



https://jupyter.org/

Google Colab

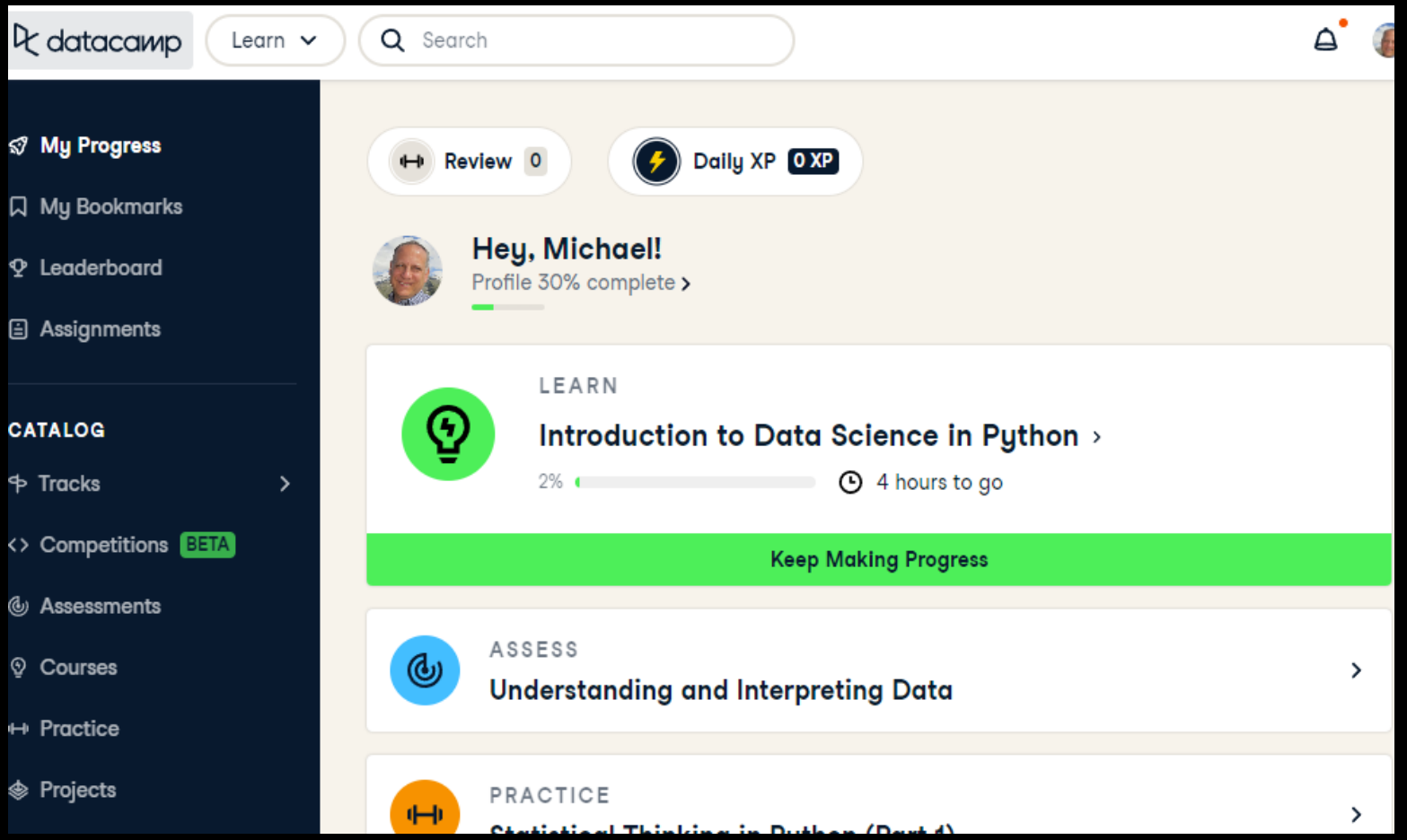


https://colab.research.google.com/

Anaconda

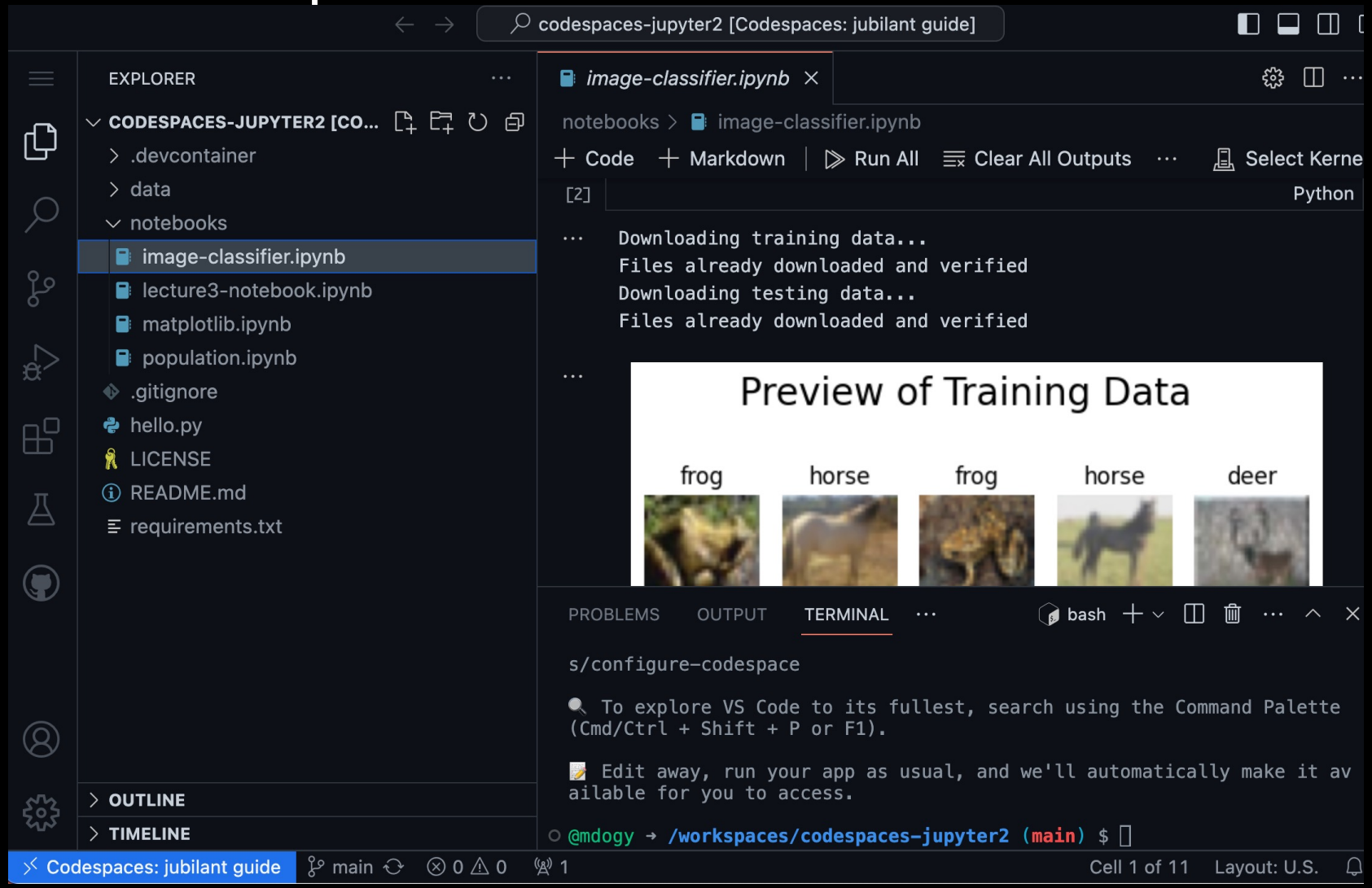
https://www.anaconda.com/

Data Camp



https://app.datacamp.com/learn

Github Codespaces



https://docs.github.com/en/codespaces/getting-started/quickstart