Intro to Data Analysis in Python

Steve McLaughlin
stephen.mclaughlin@utexas.edu

NumPy carries out common mathematical functions and extends Python to work with large arrays and matrices. The other three libraries in this table use NumPy under the hood.

pandas provides a robust toolkit for working with tabular data. Much like the programming language R, it takes care of the messy stuff (e.g., dealing with text encoding problems and missing values) and provides SQL-like functions for filtering and combining data.

SciPy is a library for scientific computing and statistical analysis. The term "SciPy" also refers to the ecosystem of interoperable tools including NumPy, pandas, and matplotlib.

matplotlib (in particular, its pyplot package) creates publication-quality plots and graphs with an easy-to-use interface. Designed to mimic the format of MATLAB's plotting tools.

Install Python and the libraries above using Anaconda for OS X / Windows / Linux

- https://www.continuum.io/downloads

General overviews

- http://www.scipy.org/getting-started.html
- http://www.scipy-lectures.org

Video courses on Lynda.com

To access videos on Lynda, click "Log in," then choose "Sign in with your organization portal." Type "utexas.edu" and enter your UT EID and password.

- "Introduction to Data Analysis with Python," taught by Michele Vallisneri (link)
- "Up and Running with Python," taught by Joe Marini (link)

NumPy quickstart

- http://docs.scipy.org/doc/numpy-dev/user/quickstart.html

Lessons for new pandas users

- http://pandas.pydata.org/pandas-docs/stable/tutorials.html#lessons-for-new-pandas-users

matplotlib graph gallery

- http://matplotlib.org/gallery.html

A collection of handy code snippets

- http://chrisalbon.com