version 2017-11-16 / academic year 2017-18

SIOC 209: Python for Data Analysis

Instructor

- Luke Thompson, Ph.D.
- Lecturer at SIO; Research Associate at NOAA
- Office hours by appointment
- Email: luke@ucsd.edu

Meetings

- Time: Tu/Th 9:00-10:20
- Room: Sea Cave, first floor of Eckart Building, Code: XXXX
- YouTube Channel: https://www.youtube.com/channel/UCVZrIrWtcvTzYIrNx7RcDyg

Course Materials

Textbooks

- *Learn Python The Hard Way* by Zed Shaw[1] -- Step-by-step introduction to Python with no prior knowledge assumed; includes appendix Command Line Crash Course (Weeks 1-4, for students with minimal programming experience)
- *Learning Python* 3rd Edition by Mark Lutz[2] -- More traditional introduction to Python as a computer language (Weeks 1-4, for students with programming experience)
- *Python For Data Analysis* 2nd Edition by Wes McKinney[3] -- Manual focused on Pandas, the popular Python package for data analysis, by its creator (Weeks 5-10)

Command Line Resources

- Git for Windows -- BASH emulator and Git software for Windows
- Learning the Shell -- Great intro to the Unix shell
- Unix Tutorial by Julian Catchen -- From Evomics 2015 workshop in Czech Republic
- *Command Line Crash Course* by Zed Shaw -- Duplicate content of Appendix A of Shaw's *LPTHW*

- Learn the Command Line -- Code Academy

Python Resources

- MIT OpenCourseWare
- SciPy Lectures
- MatPlotLib Cheatsheet

IPython Resources from Cyrille Rossant

- IPython Interactive Computing and Visualization Cookbook
- Learning IPython for Interactive Computing and Data Visualization -- GitHub repo

Data Analysis Resources

- 10-minute Tour of Pandas by Wes McKinney -- Basic video tour
- R vs. Python for Data Analysis -- Fun cartoon to abate or fuel your biases
- *Python Scripting for Computational Science* by H. P. Langtangen -- Deeper and more mathematical treatment
- An Introduction To Applied Bioinformatics by Greg Caporaso
- A Dramatic Tour through Python's Data Visualization Landscape

Course Philosophy

1. Just like anything else, you learn Python by doing. With a few exceptions, you're not going to break your computer by trying new commands. So just try it and see what happens. Print output of commands. Print values of variables. Kick the thing until it works.
2. When you don't know how to do something, google it. You'll be amazed by the solutions you'll find to do *thing x* if you google "python thing x".
3. Learn keyboard shortcuts, as many as you can. Tab-complete in the shell and IPython/Jupyter!
4. Remember Zed's sage wisdom:
   - Practice every day.
   - Don't over-do it. Slow and steady wins the race.
   - It's alright to be totally lost at first.
   - When you get stuck, get more information.
   - Try to solve it yourself first.

Schedule Overview

Schedule is subject to change.

We will start with an introduction to the command line in Week 1, so that everyone is familiar with basic Unix commands.

Weeks 2-4 will be an introduction to programming using Python. The main text will be Shaw's *Learn Python The Hard Way*. For those with experience in a programming language other than Python, Lutz's *Learning Python* will provide a more thorough introduction to programming Python.

Also in Weeks 2-4, we will learn to use IPython and IPython Notebooks (also called Jupyter), a much richer Python experience than the Unix command line or Python interpreter.

In Weeks 5-10, we'll work through McKinney's *Python for Data Analysis*, which is all about analyzing data, doing statistics, and making pretty plots (you may find that Python can emulate much of the functionality of R and MATLAB).

Detailed Schedule

- Course material is available as .md or .ipynb files by clicking on the lesson number below.
- In addition to doing the readings, please follow along writing code (this is integral to the Shaw readings), and do any Study Drills (Shaw) and Chapter Quizzes (Lutz).

Lesson	Title	Readings	Topics	Assignment
0	Introductions and Syllabus	Obtain *Learn Python The Hard Way* (Shaw), *Python for Data Analysis* (McKinney), and *Learning Python* (Lutz)	Introductions and overview of course	--
1	Command Line and Bash	Shaw: The Hard Way Is Easier, Exercise 0, Appendix A: Command Line Crash Course	A full introduction to using the command line, the bash shell, and text editors	--
			Conda tutorial including conda	

2	Conda, IPython, and Jupyter Notebooks	Install: Miniconda 3	environments, python packages, and PIP, Python and IPython in the command line, Jupyter notebook tutorial and Python crash course	--
3	Python Basics, Strings, Printing	Shaw: Exercises 1-10; Lutz: Ch 1-7	Python scripts, error messages, printing strings and variables, strings and string operations, numbers and mathematical expressions, getting help with commands and Ipython	--
4	Taking Input, Reading and Writing Files, Functions	Shaw: Exercises 11-26; Lutz: Ch 9, 14-17	Taking input, reading files, writing files, functions	--
5	Logic, Loops, Lists, Dictionaries, and Tuples	Shaw: Exercises 27-39; Lutz: Ch 8-13	Logic and loops, lists and list comprehension, tuples, dictionaries, other types	--
6	Python and IPython Review	McKinney: Appendix: Python Language Essentials, Ch 3	Review of Python commands, IPython review -- enhanced interactive Python shells with support for data visualization, distributed and parallel computation and a browser-based notebook with support for code, text, mathematical expressions, inline plots and other rich media	--
7	Regular Expressions	Grep tutorials: Drew's Grep Tutorial, Linux Grep Tutorial; Python Regular Expressions Tutorial	Regular expression syntax, Command-line tools: `grep` , `sed` , `awk` , `perl -e` , Python examples: built-in and `re` module	--

8	Numpy, Pandas and Matplotlib Crashcourse	(no readings)	Numpy overview, Pandas overview, Matplotlib overview	--
9	Pandas Basics	McKinney: Ch 1-2, 4 (Introduction to Scientific Computing with NumPy and Pandas)	`Series` , `DataFrame` , `index` , `columns` , `dtypes` , `info` , `describe` , `read_csv` , `head` , `tail` , `loc` , `iloc` , `ix` , `to_datetime`	--
10	Pandas Advanced	McKinney: Ch 5-7 (Data Analysis with Pandas); Pandas Documentation: Indexing and Selecting Data	`concat` , `append` , `merge` , `join` , `set_option` , `stack` , `unstack` , `transpose` , dot-notation, `values` , `apply` , `lambda` , `sort_index` , `sort_values` , `to_csv` , `read_csv` , `isnull`	--
11	Plotting with Matplotlib	McKinney: Ch 8; J.R. Johansson: Matplotlib 2D and 3D plotting in Python	Matplotlib tutorial from J.R. Johansson	--
12	Plotting with Seaborn	Seaborn Tutorial	Seaborn tutorial from Michael Waskom	--
13	Pandas Time Series	McKinney: Ch 10, Pandas Documentation: Time Series and Date	Time series data in Pandas	--
14	Pandas Group Operations	McKinney: Ch 9	`groupby` , `melt` , `pivot` , `inplace=True` , `reindex`	--
15	Statistics Packages	(no readings)	Statitics capabilities of Pandas, Numpy, Scipy, and Scikit-bio	--

16	Interactive Visualization with Bokeh	Bokeh User Guide	Quickstart guide to making interactive HTML and notebook plots with Bokeh	--

Assignments

Homeworks

Programming assignments: Weekly take-home assignments will be provided when the quarter begins. These exercises will focus on analysis and visualization of scientific data.

Final Project

- Choose a data set of your own or provided in one of the texts and write a Python program (or set of Python programs or mixture of .ipynb and .py/.sh scripts) to carry out a revealing data analysis.
- Have a look at Shaw Ex43-52 and McKinney Ch10-12 for more ideas.
- You should use one or several application-specific packages, such as:
  - Data analysis and plotting: pandas, matplotlib, Seaborn
  - Bioinformatics: scikit-bio, Biopython
  - Climate science: CDMS (UV-CDAT), Iris
- You should use at least 3 user-defined functions (optional: user-defined modules and classes).

Note: There are no midterm or final exams.

---

1. Book (.pdf) and video (.mp4) no-DRM access for $29.95. ↩

2. O'Reilly Media titles are free to UCSD with Safari Books Online. A PDF of this title is also available. ↩

3. O'Reilly Media titles are free to UCSD with Safari Books Online. A PDF of the first edition of this title is also available. ↩