SIO 209: Python for Data Analysis

Instructor

- Luke Thompson, Ph.D.
- Bioinformatics Researcher, Knight Lab, School of Medicine, UCSD
- luke@ucsd.edu
- Office: BRF2 1131 (office hours by appointment)

Meetings

- Time: Tu/Th 1:00-2:20
- Room: Sea Cave, first floor of Eckart Building, Code: TBD

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 1-5, for students with minimal programming experience)
- *Learning Python* Third Edition by Mark Lutz[2] -- More traditional introduction to Python as a computer language (Weeks 1-5, for students with programming experience)
- *Python For Data Analysis* by Wes McKinney[3] -- Manual focused on pandas, the popular Python package for data analysis, by its creator (Weeks 6-10)

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

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

Schedule

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

For Weeks 2-4, the course will split into two tracks:

- For those new in part or whole to programming, the beginner track will work through Shaw's *Learn Python The Hard Way*, after which you will have solid practical knowledge of computer programming and Python.
- For those with experience in a programming language other than Python, the advanced track will work from Lutz's *Learning Python*, which is a thorough introduction to programming Python. Students on the advanced track are still encouraged to follow along with Shaw's *Learn Python The Hard Way*.

With everyone up to speed by Week 5, we'll install and learn to use IPython and the IPython Notebook, a much richer Python experience than the Unix command line or Python interpreter. In Weeks 6-10, 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).

For your final project, you will choose a data set of your own and write a Python program to carry out a relevant data analysis.

Assignments

Shaw: Watch the videos, read the textbook, and follow along writing code in your own text files. Also do any Study Drills and append them to your code text files.

Lutz: Complete all Chapter Quizzes, explain your answer in your own words, and save your answers as a text file.

McKinney: You will receive separate exercises TBD (to be determined). These exercises will focus on analysis and visualization of scientific data.

See the Reading List for chapter listings.

Week 0 (9/22): Introduction and Syllabus

- Beginner: Get copies of Shaw's Learn Python The Hard Way and McKinney's Python for Data Analysis (you can also get Lutz's Learning Python if you like a more traditional and formal treatment)
- Advanced: Get copies of Lutz's Learning Python, Shaw's Learn Python The Hard Way, and McKinney's Python for Data Analysis (ideally you will follow both the Beginner and Advanced track)

Week 1 (9/27, 9/29): Computer Setup and Introduction to Python and Command Line

- Beginner: Shaw The Hard Way Is Easier, Ex0, Appendix A: Command Line Crash Course
- Advanced: Advanced Command Line Tutorial by Luke Thompson
- All: Make sure you have Python 2.7 installed on your machine. Open a terminal window and type "python --version". It should print "Python 2.7.X".

Week 2 (10/4, 10/6): Python Basics, Strings, Printing

- Beginner: Shaw Ex1-10
- Advanced: Lutz Ch1-7

Week 3 (10/11, 10/13): Taking Input, Reading and Writing Files, Functions

- Beginner: Shaw Ex11-26
- Advanced: Lutz Ch9,14-17

Week 4 (10/18, 10/20): Logic, Loops, Lists, Dictionaries, and Tuples

- Beginner: Shaw Ex27-39
- Advanced: Lutz Ch8-13

Week 5 (10/25, 10/27): Regular Expressions, IPython and IPython Notebook

- Python Regular Expressions Tutorial
- McKinney Appendix: Python Language Essentials (review)
- McKinney Ch3

Week 6 (11/1, 11/3): Scientifc Computing with NumPy and Pandas

- McKinney Ch1-2,4
- `numpy.ipynb`
- `lesson01.ipynb`
- `lesson02.ipynb`
- `lesson03.ipynb`

Week 7+ (11/8, 11/10, 11/15): Data Analysis with Pandas

- McKinney Ch5-7
- Pandas Documentation: Indexing and Selecting Data
- `pandas_dtype.ipynb`
- `pandas_scripps_pier.ipynb`
- `lesson04_sio209.ipynb`
- `lesson05_sio209.ipynb`
- `lesson06_sio209.ipynb`
- `answers_pandas_homework1.ipynb`
- `pandas_advanced.ipynb`

Week 8 (11/17): Plotting with MatPlotLib and Seaborn

- McKinney Ch8
- Seaborn Documentation
- `seaborn_tutorial_1_controlling_figure_aesthetics.ipynb`
- `seaborn_tutorial_2_color_palettes.ipynb`
- `seaborn_tutorial_3_visualizing_the_distribution_of_a_dataset.ipynb`
- `seaborn_tutorial_4_visualizing_linear_relationships.ipynb`
- `seaborn_tutorial_5_plotting_with_categorical_data.ipynb`
- `seaborn_tutorial_6_plotting_on_data_aware_grids.ipynb`

Week 9 (11/22): Interactive Visualization with Bokeh

- NBViewer Tutorial for JupyterHub and GitHub
- Bokeh IPython Notebooks
- Last Day: Five 10-Minute Tutorials from Knight Lab Members

Week 10 (11/29, 12/1): Time Series (and Data Aggregation, Group Operations)

- McKinney Ch9-10
- Pandas Documentation: Time Series and Date
- 12/3: Guest speakers!

Finals Week

- Final project due by end of finals week

Bonus Material: Modules, Classes, and Object-Oriented Programming

- Beginner: Shaw Ex40-42
- Advanced: Lutz Ch18-26

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 .py and .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).

---

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 this title is also available. ↩