March 17-19, 2021

BIOF085: Introduction to Data Science using Python

Introduction

This is a 3-day workshop on using Python for Data Science. Over the three days, we will introduce the Python scripting language and also the major packages used in Python for data science. This is quite a condensed introduction, so we will move at a pretty fast clip. After this workshop, you should know the following skills:

1. Use Python through an integrated development environment (we will introduce two: Spyder, Jupyter Notebooks) and/or a shell terminal
2. Gain an understanding of how to use freely available Python packages in your environment
3. Load data into Python
4. Clean, manipulate and munge raw data to make it amenable for analysis
5. Visualize data through statistical graphs and some interactive graphs
6. Run basic statistics and regression models on data
7. Run basic machine learning models on data
8. Have a high-level idea of how to apply Python to various data science problems, including some bioinformatics problems

Instruction Information

Instructors: Gaby Gerlach, Roshni Bhatt

We can be contacted on the Slack Channel (see below and announcement)

Note: The content of this course was designed and created by Dr. Abhijit Dasgupta who has been doing bioinformatics for 20 years and teaching it for almost as long. It is his voice and computer in the recorded screen casts, and we thank him for the provided material.

Format

This is an online workshop, which seems an oxymoron. We will use hybrid (in-person and online) materials during this workshop, to help instructor-student engagement as well as provide a modicum of independent, self-paced learning. We expect that you will participate fully in this workshop over the 3 days. In particular, all in-person sessions (see schedule) will assume that you have completed the asynchronous material and progress checks assigned for the previous time period. Our engagement during this workshop will take several forms:

- **Class materials:** All materials, including screencasts, slides, videos, textual handouts and assignments/progress checks, will be available on the FAES Canvas Site under this class's site.

- **In-person engagement:** We will have several sessions of in-person engagement through Zoom. The links will be posted in Canvas, and Zoom can also be accessed through Canvas
- **Communications:** We will have a dedicated **Slack channel, biof085-students.slack.com**, for this class. **Please check the Announcements on Canvas for the link.** We will be monitoring this Slack channel from 1 week before the workshop starts to 2 days after the workshop ends. Please promptly join this channel.

Software

We will be using Python 3 for this workshop, since Python 2 is no longer being maintained. In particular, we will be using the Python 3.8 distribution provided by Anaconda. This distribution comes "batteries included" for all the data science work we'll be doing, including all the requisite packages. This distribution is available for Windows, MacOS and Linux.

The Anaconda Python distribution is available at https://www.anaconda.com/products/individual (Links to an external site.). Scroll down to the bottom of the page and download the Python 3.8 Graphical Installer appropriate for your operating system. You should install the **64-Bit Graphical Installer** unless you have a really old computer.

If you have installation issues, please come to the Day 0 session, and if that is not possible ping Gaby on Slack **no later than 9pm the day before class starts** so we can help you with any issues. **If you are not at the day 0 session, please complete the day 0 section of canvas before day 1.** The Anaconda distribution is robust and should install effortlessly on all operating systems, in our experience, but you never know.

March 17-19, 2021

Schedule

Day	Time	Format	Topic	Instructor	Resource
0	5:30pm-6:30pm	In-person via zoom	Logistics	Gaby and Roshni	
1	9am – 11am	In-person via zoom	Why python?	Gaby	00_python_primer
			A python primer		
	11 am - noon	In-person via zoom	Python tools for data science	Roshni	01_python_tools
	1pm-2pm	In-person via zoom	Python tools for data science	Gaby	02_python_pandas
			Data wrangling, cleaning, summarizing		
	2pm -5:00pm	Asynchronous material	Data Munging	Slack	Day 1 independent work videos
2	9:30am –10:30am	In-person via zoom	Data Visualization/Stats	Roshni	Day 2 introduction
	10:30am-12:30pm	Asynchronous material	Data Visualization	Slack	Day 2 independent work videos
	1:30pm - 2:30pm	In-person via zoom	Statistical Analysis pipeline	Roshni	04_stat_intro
	2:30pm-4:00pm	Asynchronous material	Statistical Analysis using Python	Slack	stats_python_practice
	4:00pm-4:30pm	In-person via zoom	Q&A	Roshni	
3	9am –10am	In-person via zoom	Data analytics: Machine Learning	Gaby	05_python_learning
	10am – 12:30	Asynchronous material	Data analytics: Machine Learning	Slack	Day 3 independent work videos
	1:30pm -2:30pm	In-person via zoom	String manipulation, RegEx, intro to bioinformatics	Gaby	06_python_appl
	2:30-4:00pm	Asynchronous material	Use of skills learned in ML/bio application	Slack	Day 3 Project - genomics_project
	4:00 – 4:30 pm	In-person via zoom	Explanation of tools, Q&A, miscellaneous	Gaby	06_python_appl

Asynchronous material

Asynchronous material will be a combination of screencasts, videos, slides and textual material, as well as progress checks. The idea will be to learn for 15 minutes, then complete a progress check, and then move to the next module. We also have quick progress checks during the in-person sessions as well.

Office hours

We won't have any office hours per se, but we will be available at the end of each day on Zoom, and throughout the 3 days of the workshop on Slack. We will try to respond promptly to all queries, and will either reply individually or, if the question is of general interest and would fill a gap in the materials, on the general slack channel.

FAQs

Q: Do I need to know programming or Python to do this workshop?

A: You will not need to know programming or Python to do this workshop. However, some familiarity with general programming concepts or some experience with a scripted language like R, SAS, Stata or Java would be very useful to pick up concepts, due to the pace of the workshop. We'll try to explain the basic concepts thoroughly, but this workshop is designed with the assumption of attendees having computer literacy.

Q: Do I need to know biology or genomics or the like to do this workshop?

A: The material in this workshop is mainly domain-agnostic, except for the section on bioinformatics. We will cover bioinformatics at a high level and show examples, so you do not need a bioinformatics background to do this workshop.

Q: Do you expect me to finish the asynchronous sections before coming to the Q & A session?

A: **Yes, absolutely!** However, we recognize that people work at different paces and if it is not possible for you to finish the work it is alright. We will have a hard stop each evening at 5pm for so that both you and we can rest and recharge for the following day, and you have some opportunity to digest material. Keeping up with the material will allow you to get the most out of the workshop as days 2 and 3 build on the first day.

Q: How long are you available each day?

A: We will be available from 8:30 am to 6 pm each day for questions/comments/explanations on Slack. If there are questions/comments sent after 6pm we'll answer them based on time availability that evening, but will definitely address them before the next day's session. For the two days after the workshop, we will answer all queries, but maybe not as immediately; you can expect responses around lunch time and at the end of the day.