

Python for Environmental Analysis and Visualization
Short Title: Big Data with Python
Syllabus
BC EESC 3050
Milbank 222
Monday and Wednesday: 2:40-3:55

Brian Mailloux
404J Altschul Hall
Office hours: M, W 4-5:00 and by
appointment
bmaillou@barnard.edu
212 854 7956

COURSE OBJECTIVES

Short Description for Catalog

Big Data is changing how we interact with and understand the environment. Yet analyzing Big Data requires new tools and methods. Students will learn to use the iPython programming environment to analyze and visualize large environmental data sets. This will include both time series and spatial analyses.

Course description:

Big Data is changing how we interact with and understand the environment. Yet to analyze Big Data requires new tools and methods. Microsoft Excel is not sufficient to analyze big data and quickly becomes frustrating. We are going to analyze environmental data using Python in order to better understand the earth's system. The class will be a hands-on flipped classroom with programming occurring interactively during class and assignments designed to strengthen the methods and results of class.

The class will begin by analyzing climate data from Central Park and comparing it to different cities around the world. This analysis will enable us to better learn about climate and comparative statistics while at the same time learning the programming language Python.

The class will then utilize Python to analyze bacterial transport distances. We will learn how computer programs can quickly compute and store hundreds to millions of simulations to find optimal fitting parameters.

The class will analyze 50,000 arsenic measurements from Bangladesh. We will spatially analyze the data and learn how to sub-select large data sets and perform statistics while mapping the results.

The class will then visualize and analyze global scale data sets. We will download global scale data sets from the internet and use the data in our own analyses.

The class will then be introduced to time series data and learn how to organize a time series and then analyze a time series for trends.

Finally, the students will individually analyze a data set and present the results of the analysis. Students will learn how to clean and organize data and independently perform an analysis on the data set with the goal of testing a hypothesis or answering a specific question.

Student Learning Goals and Outcomes:

- To be able to organize and analyze large datasets.
- To be able to perform statistical analyses on large data sets
- To be able to write code in Python.
- To be able to download and analyze datasets from the internet
- To be able to plot and map large data sets.
- To develop a hypothesis and then be able to take any data set no matter how large or complicated and analyze and visualize it in order to test that hypothesis.

Lecture topics:

Week #	Date	Lecture #	Topic	Due
1	9/7	1	Class Introduction Central Park Example	
2	9/12	2	Print, Markdown, directories, lists	Code.org 1 hour Frozen
	9/14	3	Make a list and plot it	Notebook 1
3	9/19	4	More lists, Start Strings	Notebook 2
	9/21	5	For loops	Notebook 3
4	9/26	6	If/elif/else Plot LGA data	All code.org due Notebook 4
	9/28	7	Fit a line. Fit to LGA and JFK data	Notebook 5
5	10/3	8	Numpy, fit a line. GDP data	Notebook 6
	10/5	9	Arrays and Colormaps	Notebook 7
6	10/10	10	Pandas (the Python library) and Arsenic data	Notebook 8
	10/12	11	More Pandas, Multiple plots	Notebook 9
7	10/17	12	Correlations in Pandas	Notebook 10
	10/19	13	Senior thesis data? p-hacking	Notebook 11
8	10/24	14	Sediment Core Data on your own. Test a hypothesis	Read paper before arriving in class. Notebook 12
	10/26	15	Sediment Core Data on your own	
9	10/31	16	Catch up TBD	Notebook 13d
	11/2	17	CO ₂ Time Series	All Code Academy Due
	11/7		Holiday	
10	11/9	18	CO ₂ Time Series Part 2	
11	11/14	19	Mapping part 1	Notebook 14d
	11/16	20	Mapping part 2	
12	11/21	21	Netcdf, SST, Animated GIF	Notebook 15d
	11/23	22	Day before thanksgiving	

Week #	Date	Lecture #	Topic	Due
			Work on Data Set 1	
13	11/28	23	Netcdf, SST, Animated GIF	
	11/30	24	Work on Data Set 1	Notebook 16d
14	12/5	25	Work on Data Set 2	
	12/7	26	Work on Data Set 3	
15	12/12	27	Final Presentations	Final Projects Due Early before class. tba

Website

We will be using both courseworks and edblogs. Be prepared to use both. You will get all of your class content from the blog. You will upload your homework to courseworks.

Readings

Required

Each day you will be handed a packet on the days work. These packets should be put into a binder and brought to every class. They will become your book. You should add notes to them and use them. I use them all the time as I program.

Supplemental

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython Paperback – October 29, 2012 - by Wes McKinney (Author)
 Introduction to Computer Science Using Python: A Computational Problem-Solving Focus Paperback – December 25, 2012, by Charles Dierbach (Author)

Grading:

Online Tutorials 20%
 Python Notebooks for each Section 55%
 Final Presentation 15%
 Class Participation. 10%

CLASS PREMISE: This is a flipped classroom. You will learn some programming from code.org and code academy. During class time we will work on in class projects that are designed to teach about different components of Environmental Science while at the same time teaching you how to analyze and visualize large data sets and how to program. At the end of each class or section you will have an iPython Notebook that teaches you multiple concepts. Then at the end of each class you will use this knowledge for your nightly homework. The Notebooks and scripts will contain all your notes, figures, results, etc. By running and examining the Notebooks and results we will be able to quickly determine how much of the in class material you completed. You can not rush through the class work as by taking your time and learning it you learn the concepts needed for the homework and the next class. It all builds on itself.

CLASS PROJECTS:

WEB BASED TUTORIALS: In order to help you learn programming you will do web based tutorials during the first half of class. These tutorials don't always match class but will help you learn how to program. Code.org is designed for younger kids but really teaches for loops and if statements so it is really helpful. For the web based tutorials you will join the class group at each website and your progress will be tracked online by the professor. You have three due dates for the web based tutorials. Your grade will be the percent you complete by the due date. If you complete 100% of the tutorials you get 100%. So make sure you give yourself enough time and complete the tutorials. In the past some people have done the wrong tutorial. Make sure you are doing to correct tutorial and sign up for the correct class. **YOU WILL NOT GET CREDIT FOR DOING THE WRONG ONE.** Your three tutorials are

1. Code.org 1 hour with Frozen
2. Code.org 20 hour "Accelerated Course". (You do not need to do the offline parts)
3. Code academy Introduction to Python. (90% for finishing through battle ship. 100% for finishing the whole thing)

NIGHTLY PYTHON NOTEBOOKS: There will be a nightly homework based on each class. The homeworks will be posted on edblogs. If you diligently worked through the work in class you should be able to do the nightly homework. This makes coming to class a necessity as if you fall behind it will be difficult to accomplish all of the steps independently. The notebooks will consist of code, annotation, and results. The figures are presented inline in the notebooks. You need to add comments and markdown to your notebooks describing what you did. You are expected to have a figure caption and a description of the figure. Homework is due at the beginning of class time. I will use the time stamp on courseworks dropbox to determine if the homework is on time. I will post grading rubrics on courseworks. You can see exactly what I am looking for. For the notebooks, if the notebook takes two classes it will be worth twice as much in the grading. If you do not add Markdown and notes you will lose points and the amount of points will get larger with each homework. For many homeworks I will make the final 10% of the grade harder to achieve and take more time. For example you can choose to not work on the last question and get a 90. This will be clearly stated on each homework.

STACKOVERFLOW: Part of learning to program is learning to find answers on your own. This might be through trial and error, asking a friend or looking online. When learning something new I will watch a youtube video or read a tutorial. But for programming questions stack overflow is a commonly used website. People post questions and answers. It can be really helpful for figuring out bugs and issues and doing something new. Unfortunately people can also be mean. You will need to use it over the course of the semester. I can't answer all your questions and you need to learn how to learn on your own.

FINAL POSTER SESSION: Each student is going to create a poster presentation from a data set. Each student will analyze their own data set. The student will develop questions or hypotheses about the data set and will use what we learned throughout the semester to tackle their problem. The final few days of class will be analysis of the data followed by the student posters on the last day. You can analyze a dataset that you are interested in and find or I can supply you with a dataset. You can not just repeat an analysis you did for another class. Many people have used

this section to analyze their senior thesis data more in depth. Final Posters will be due before the final class. There will be a separate page describing the process and the due dates.

HANDING IN HOMEWORK: I will use the timestamp on dropbox in courseworks to check when it was posted. All homework is due at the beginning of class. Each day late you will lose 10% with a maximum loss of 50%. In order to do well in class you need to hand in each homework and get them in on time. When handing in homework you will post your notebook on dropbox in courseworks. You will name it by assignment number and name as shown on courseworks.

THANKSGIVING WEDNESDAY: I realize many people are traveling the Wednesday before Thanksgiving. I will have class that day but I will not take attendance. If you can, I would recommend coming to class that day. It will give you a chance to get ahead as we start the final projects. I will be there and willing to help you on any class material.

Class Attendance and Participation:

It is expected that every student attends every class. This is a hands-on class where we learn by coding as a group during class. If you miss a class you will fall behind and it will be difficult to catch up during that section. In addition, students will help propose answers to coding problems during class. I will have a sign in sheet that you will need to sign as you enter class and if you miss more than two classes you will lose participation points. No headphones are allowed while programming and I expect you to talk to your neighbor about coding during class.

OFFICE HOURS: I will hold regular office hours. I also come early and stay late at class. But also, I am usually on the 4th floor of Altschul either in my office or the lab. If you find me I am happy to answer questions. Also, as part of participation everyone is required to come and ask at least one question over the course of the semester. I will have a sign up sheet next to my door that you can sign once you have asked a question.

Collaboration:

When working on coding and projects it is easy to run into roadblocks and stumble. This is when talking to a classmate is very helpful. You can ask how you tackled a problem. How did you iterate over the data? What type of if statements did you use? I strongly urge you to work with and talk to your classmates. The classroom should have people speaking. You are not allowed to wear headphones during class. However, you cannot copy a classmate's code. Since everyone comments and codes slightly differently it is easy to observe when someone has copied and pasted and this is not allowed and will be considered plagiarism and reported.

Email:

I try to answer emails as fast as possible. But do not expect rapid replies late at night or on weekends. Also, if a person emails me a good question, I usually do not reply directly but send an email to the whole class through Courseworks with the question and answer.

Courseworks

Courseworks will be utilized to post all files and readings from class.

Course Requirements

If you miss the test you need a note from your class Dean in order to reschedule the exam.

Definition of Grades: All grades will be based on a scale of 100 with

A+	97.00-100	Rare performance. Reserved for exceptional achievement
A	93.00-96.99	Excellent work. Outstanding achievement.
A-	90.00-92.99	Excellent work that exceeds course expectations.
B+	87.00-89.99	Very good work. Solid achievement (expected of Barnard/Columbia undergraduates) that meets all course expectations.
B	83.00-86.99	Good work. Acceptable achievement that meets almost all course expectations.
B-	80.00-82.99	Satisfactory work. Acceptable achievement that meets major course expectations.
C+	77.00-79.99	Fair achievement just above that which is minimally acceptable
C	73.00-76.99	Fair achievement but only minimally acceptable.
C-	70.00-72.99	Barely acceptable achievement.
D	60.00-69.99	Very low performance. Unsatisfactory work. Lowest achievement to still allow for a passing grade. This grade may not be counted toward the major or minor option
F	59.99 or less	Failure.

Consult the Barnard or Columbia College Catalogs or the Registrar's Office for other information about grading, including: the definitions of other letter grades, pass/D/fail option, incompletes, and calculation of GPA.

The Honor Code and Academic Integrity

The Barnard Honor Code (established in 1912) reads:

We, the students of Barnard College, do hereby resolve to uphold the honor of the College by refraining from every form of dishonesty in our academic life. We consider it dishonest to ask for, give, or receive help in examinations, quizzes, or to use in them any papers or books in any manner not authorized by the instructor, or to present oral or written work that is not entirely our own, except in such way as may be approved by the instructor. We pledge to do all that is in our power to create a spirit of honesty and honor for its own sake.

The Honor Code governs all aspects of academic work. If a violation should arise, it will be reported to the Dean of Studies for appropriate action. Honor Board Guidelines on the procedures for implementing the Honor System and acting on charges of dishonesty can be found in the Student Handbook. Remember that Barnard students reaffirm their acceptance of the Honor Code by signing their registration form. Columbia students commit themselves to the Honor Code upon registering for a Barnard course. Because the Honor Code is not entirely specific and contains qualifications and exceptions, such as "authorized by the instructor" or "approved by the instructor", please read the attached summary of "What behaviors constitute academic dishonesty?"

Students with disabilities

Students who may need disability-related classroom accommodations are encouraged to make an appointment to see me as soon as possible. Disabled students who need test or classroom accommodations must be registered in advance with the Office of Disability Services (ODS) in 105 Hewitt. Thank you.