PPOL 564-01
Data Science I: Foundations
Fall 2019

Instructor

Professor: Eric Dunford

- **Office**: 404 Old North
- **Office Hours**: Tuesdays 3pm to 5pm
- **Email**: eric.dunford@georgetown.edu
- **Pronouns**: He/Him

Teaching Assistant: Andrea Quevedo

- **Office Hours**: by appointment
- **Email**: aq38@georgetown.edu
- **Pronouns**: She/Her

Class Website: www.ericdunford.com/ppol564

Course Description

This first course in the core data science sequence introduces students to the programming and mathematical concepts that underpin statistical learning. The aim of the course is to provide students with the foundations necessary to grasp the concepts and algorithms encountered in Data Science II and III. Students will cover linear algebra with a focus on linear regression and dimension reduction; multivariate calculus with an emphasis on optimization algorithms, specifically gradient descent; and probability theory with an emphasis on simulation and sampling. Throughout the course, students will be introduced to the fundamentals of programming and manipulating data with Python. Students will work in Jupyter notebooks and use Git/GitHub to submit coding assignments, developing literate programming and reproducible research skills they will use throughout the program.

The course covers four main topics. Topic 1 focuses on building competency with programming and wrangling data in Python with an emphasis on producing readable, reproducible, and well-documented code. Students will build competency with version control using Git/Github. The skills acquired in Topic 1 are applied throughout the course both in the form of coding "discussions" and applied assignment. Topic 2 delves into linear algebra. Students will gain a vital intuition of many of the mathematical properties that inform statistical learning

and data decomposition. Topic 3 covers uni- and multivariate calculus with an emphasis on optimization where students will gain an intuition of computational optimization using Python. Finally, Topic 4 briefly covers probability theory with a focus on computational simulation, Bayes rule, and sampling.

Time and location

Classes will be held on **Mondays** and **Wednesdays** from ***2:00pm to 3:15pm*** in **Car Barn 205**:

- September 4, 9, 11, 16, 18, 23, 25, 30
- October 2, 7, 9, 16, 21, 23, 28, 30
- November 4, 6, 11, 13, 18, 20, 25
- December 2, 4, 9, 20

  Note: Class will meet later in the day on September 23rd.

Recitation will be held most weeks on Mondays from *5:00pm to 5:50pm* in Car Barn 203. Attendance is required. A complete schedule of recitation dates can be found on the class website (see "Lab Schedule" tab @ http://ericdunford.com/ppol564/#course_outline).

Holidays/Breaks/Away (No class):

- August 28 (Professor is Away)
- September 2 (Labor Day)
- October 14 (Mid-Semester Holiday)
- November 27 (Thanksgiving Recess)

Course Objectives

The course aims to provide students with the following competencies:

- General understanding of python's programming syntax and data structures.
- Competency of manipulating and exploring data with Pandas.
- Understanding of key mathematical concepts that underpin machine learning and statistical estimation.
- Applied experience programming and debugging in Python.
- Competency using version control (Git/Github).

Required Materials

Readings: We will rely primarily on the following text for this course.

- **A Mathematics Course for Political and Social Research**. Moore, Will H., and David A. Siegel. Princeton University Press, 2013.
- Additional readings will be posted for each class day and can be found on the course website. Most reading material is open source and available via a link on the reading list, otherwise it can be found on CANVAS.
- All class slides and lecture notes will be available on the class website.

Class Website: A class website (www.ericdunford.com/ppol564) will be used throughout the course and should be checked on a regular basis for lecture materials and required readings.

Class Slack Channel: The class also has a dedicated slack channel (ppol564-foundations.slack.com). The channel serves as an open forum to discuss, collaborate, pose problems/questions, and offer solutions. Students are encouraged to pose any questions they have there as this will provide the professor and TA the means of answering the question so that all can see the response. If you're unfamiliar with, please consult the following start-up tutorial (https://get.slack.help/hc/en-us/articles/218080037-Getting-started-for-new-members). Please follow the ***invite link*** to be added to the Slack channel.

Canvas: A Canvas site (http://canvas.georgetown.edu) will be used throughout the course and should be checked on a regular basis for announcements, readings, and assignments. All readings and assignments will be posted on Canvas; they will not be distributed in class or by e-mail. Support for Canvas is available at (202) 687-4949

NOTE: Students are encouraged to run lecture code on their own machines. If you do not have access to a laptop on which you can install `python3` and an `Anaconda` distribution, please contact the professor and/or TA for assistance. Only `python3` will be used in this course.

Course Requirements

Assignment	Percentage of Grade
Participation	10%
Coding Discussions	15%
Problem sets	35%
Final	40%

Preparation and Participation (10%): It is imperative that you arrive to class prepared for lecture and any hands-on activities during the lab. As a result, 10% of each student's grade will be based on class participation. See "Attendance/Participation" in the Course

Policies section for more details.

Coding Discussions (15%): Most weeks there will be a coding problem/prompt/dataset to explore pushed to the class Github repository. Students will be required to submit an original response to the prompt. Each student must then respond to another student's coding solution in the form of a *substantive* contribution to the existing code or by debugging an issue. ***All edits should be pushed to the master branch***. Submission and response will be assigned `3` points each week. A point will be awarded for (1) submitting on time, (2) responding on time, and (3) the quality of the post & response. "Quality" is defined as an *original*, well-commented, and clear solution/analysis (i.e. the solution follows the guidelines required for each coding assignment entry - see below). Students are encouraged to generate a response before looking at the responses of their peers. If a student appears to have copied an answer of another student, we'll examine the time stamp and only award a quality point to the first entry of the timeline of responses that appear duplicative.

All submission must be posted by Friday by 11:59PM. All coding responses to another student's push must be pushed by Sunday 11:59PM. It is vital that all students submit their answers in a timely fashion in order for others to respond and for us to discuss outcomes in recitation. The schedule for the coding assignments are listed below.

No.	Coding Discussion Week	Date Assigned
1	Week 2	September 11
2	Week 3	September 18
3	Week 5	October 2
4	Week 7	October 16
5	Week 8	October 23
6	Week 9	October 30
7	Week 11	November 13

The goal of the coding discussions is to apply a concept learned during the week in a way that helps build a greater level of programming fluency. Programming skills are honed through active usage and repetition. Learning to read other people's code and detecting issues is vital to successful collaborations in applied work. The point is not to be "right" necessarily but rather to try, learn, and collaborate. As such, when receiving suggested edit requests

Problem Sets (35%): There will be five assignments throughout the course of the semester. These assignments will take the form of questions that the student must complete independently or with assigned partners (See Course Policies "Homework Partner"). The goal of the assignment is to reinforce the student's comprehension of the materials covered in each section. All assignments will be posted Wednesday afternoon by 5:00PM on the class CANVAS site for the weeks marked on the syllabus. Problem sets are due on the date and time posted on Canvas and must be submitted on Canvas. Generally, a week will be allotted to complete each assignment. Late assignments will be penalized a letter grade for every day they are overdue.

The assignments will be in the form of a Jupyter Notebook (.ipynb). Student's must submit

completed assignments as Jupyter notebooks (.ipynb) with all coding chunk run (so that the output is visible upon opening the notebook). All assignment submissions must adhere to the following guidelines:

- (i) all code must run;
- (ii) solutions should be readable
  - Code should be thoroughly commented (the Professor/TA should be able to understand the codes purpose by reading the comment),
  - Coding solutions should be broken up into individual code chunks in Jupyter notebooks, not clumped together into one large code chunk (See examples in class or reach out to the TA/Professor if this is unclear),
  - Each student defined function must contain a doc string explaining what the function does, each input argument, and what the function returns;
- (iii) Commentary, responses, and/or solutions should all be written in Markdown and should contain no grammatical or spelling errors;
- (iv) All mathematical formulas should be written in LaTex;

The follow schedule lays out when each assignment will be assigned.

Assignment	Date Assigned
No. 1	September 25
No. 2	October 9
No. 3	November 6
No. 4	November 20
No. 5	December 4

Final (40%): We will have one final for this course at the end of the semester. The final seeks to test each student's understanding and comprehension of the main topics of the course. Students will be given two hours to complete the final. Details on the final will be announced in class as we approach the end of the term. ***No practice exam will be offered*** (See "No Practice Exams" in Course Policies). The final will be held on December 20 from 12:30PM to 2:30PM.

Grading:

Course grades will be determined according to the following scale:

Letter	Range
A	95% – 100%
A-	91% – 94%
B+	87% – 90%

Letter	Range
B	84% – 86%
B-	80% – 83%
C	70% – 79%
F	< 70%

Managing the Workload: How to Succeed in this Course

- **Come Prepared.**
  - Do the readings. Think about the readings on their own terms, but also in terms of how the concepts apply to things you're interested in.
  - It is expected that students bring their computers to class to partake in computational activities or play with coding being discussed in class. Moreover, students should have all relevant software up and running on their machines.
- **Ask Questions.**
  - Formulating a question helps you engage with the material much more deeply. If you have a question, it's almost certain that others do too; asking a question will not only help yourself, but you will help others. Most importantly, asking questions helps keep the class on track. If there are lots of questions, we'll slow down and get things figured out. If there are few questions, we'll charge ahead.
- **Collaborate.**
  - Utilize **the class slack channel** to pose any questions, insights, coding problems and concerns. The channel will offer an open forum to communicate, collaborate, and collectively problem solve.
- **Start homework early.**
  - Sometimes the data doesn't cooperate, or there is an error in your code that will take you awhile to figure out and debug. You don't want to find this out at 11pm the night the homework is due. Also, the more you are doing homeworks on time, the more you will be able to follow the lectures.
- **Try doing it the hard way.**
  - A core factor in the success of a data scientist is being able to explain how an algorithm or analysis was constructed, not just use software. In this class, where possible, build from scratch rather than an overly convenient library. This will allow you to become more creative down the line.

Course Policies

Attendance/Participation

Participation is required in this course. Participation can be decomposed into attendance (class and recitation), engagement, and completing the class assignments. Specifically, I define "engagement" as:

- Asking questions and participating in class (no zombies)
- Paying attention to the professor during lecture
  - not looking at your computer screen for extended periods of time
  - never looking at your phone during class
- No side conversations during lecture

I reserve the right to deduct attendance points from students who are not participating as expected.

Attendance will be taken daily. A sheet of paper will be made available at the front of the room at the start of every class. Students must write their name on the paper. The **paper will be removed 5 minutes after the start of class**. Students who walk in late after that point will not have an opportunity to write their name and will be considered absent. This log will be used, in part, to calculate the attendance grade.

If absent, each student is responsible to make up the materials missed during a lecture on their own. All lecture notes will be posted on the class website. Thus, students who missed a lecture should reach out to their peers in the class for lecture notes. It is not the responsibility of the Professor/TA to fill absentee students in on any missed content

Communication

- For private questions concerning the class, email is the preferred method of communication. All email messages must originate from your Georgetown University email account(s). Please use a professional salutation, proper spelling and grammar, and patience in waiting for a response. The professor reserves the right to not respond to emails that are drafted inappropriately. ***Please email the professor and the TA directly rather than through the Canvas messaging system.*** Emails sent through CANVAS will be ignored.
- For general, class-relevant questions, `Slack` is the preferred method of communication. Please use the general or the relevant channel for these questions.
- I will respond to all emails/slack questions within 24 hours of being sent during a weekday. I will not respond to emails/slack sent late Friday (after 5PM) or during the weekend until Monday (9AM). Please plan accordingly if you have questions regarding current or upcoming assignments. Please address the professor and TA by their last name unless stated otherwise.

Electronic Devices

The use of laptops, tablets, or other mobile devices is permitted *only for class-related work*. Audio and video recording is not allowed unless prior approval is given by the professor. Please mute all electronic devices during class.

Assignments and Late Work

Assignments should be clear, legible, and submitted in the required format. Writing assignments will be graded on the basis of content, logic, analysis, mechanics, organization, and research. Due dates for all assignments will be posted on Canvas and are non-negotiable. Exceptions to this policy will be made only under extremely unusual circumstances and will require valid documentation from the student. ***Late problem sets will be penalized a letter grade per day.***

Proof of Diligent Debugging

When reaching out to the professor or teaching assistant regarding a technical question, error, or issue you ***must*** demonstrate that you made a good faith effort to debugging/isolate your problem prior to reaching out. In as concise a way as possible, send a record of what you tried to do. ***The professor/TA is a resource of last resort***. As software is continually being refined in data science and new approaches continually emerge and changing, learning how to frame your question and find a similar solution online is a key tool for success in this domain. If you make a diligent effort beforehand to solve your problem, we will do the same in trying to help you figure out a solution.

No Practice Exams

Most data analytic jobs now require that you perform a task or take a test. There is no practice test for these exams. You know what you know, and you're either able to implement it or not. Moreover, life rarely offers us a chance to practice before being tested. As such, there are no practice exams in my class. Students will be given a study guide for exams that broadly outlines topics that might be on the exams. Students are encouraged to talk with the professor and TA during office hours about the exam, but no example questions will be offered.

Class Seating

Students should (and must) sit beside a different student each time the class meets. The aim is to facilitate diverse interactions. If all or some students fail to follow this procedure, then the professor reserves the right to generate a random seating assignment each class. Failure to comply with the class seating policy will result in a deduction in participation points.

Homework Partner

Students will be randomly paired with a partner for each assignment. Student may work with that partner when completing the assignment. The student may not work with anyone else on the assignment. Students are not required to work with their partner but should. Failure to comply (e.g. by working with someone other than your partner) will result in a 5 point reduction n the total score of the assignment. The list of random pairing can be found on CANVAS with the assignment details. The point of this policy is to facilitate diverse student collaboration.

Use of Class Materials

Increasingly, with the proliferation of certain websites, questions about the ownership of course materials have arisen (and Georgetown is actively working on policies to address these concerns). I consider my syllabus, lectures, handouts, problem sets, and problem set answers to be my intellectual property. I respectfully request that you refrain from sharing my materials in any electronic (or paper) format. You are welcome to record my lectures for your own use, but they should not be posted anywhere. Sharing notes, on an occasional basis, with others in the class is fine as long as they are not posted.

Academic Resource Center/Disability Support

If you believe you have a disability, then you should contact the Academic Resource Center (arc@georgetown.edu) for further information. The Center is located in the Leavey Center, Suite 335 (202-687-8354). The Academic Resource Center is the campus office responsible for reviewing documentation provided by students with disabilities and for determining reasonable accommodations in accordance with the Americans with Disabilities Act (ASA) and University policies. For more information, go to http://academicsupport.georgetown.edu/disability/.

Important Academic Policies and Academic Integrity

McCourt School students are expected to uphold the academic policies set forth by Georgetown University and the Graduate School of Arts and Sciences. Students should therefore familiarize themselves with all the rules, regulations, and procedures relevant to their pursuit of a Graduate School degree. The policies are located at: http://grad.georgetown.edu/academics/policies/.

Plagiarism

Plagiarism is the intentional or unintentional presentation of another person's idea or product as one's own. Plagiarism includes, but is not limited to the following: copying verbatim all or part of someone else's written work; using phrases, charts, figures, illustrations, code, or

mathematical / scientific solutions without citing the source; paraphrasing ideas, conclusions, or research without citing the source; and using all or part of a literary plot, poem, film, musical score, or other artistic product without attributing the work to its creator. In technology, plagiarism is the verbatim use of a code chunk from a peer or third party website to complete an assignment questions. Students can avoid unintentional plagiarism by following carefully accepted scholarly practices. Students who plagiarize will receive a 0 on the plagiarized assignment and may fail the course, if deemed necessary.

Provosts Policy Accommodating Students Religious Observances

Georgetown University promotes respect for all religions. Any student who is unable to attend classes or to participate in any examination, presentation, or assignment on a given day because of the observance of a major religious holiday (see below) or related travel shall be excused and provided with the opportunity to make up, without unreasonable burden, any work that has been missed for this reason and shall not in any other way be penalized for the absence or rescheduled work. Students will remain responsible for all assigned work. Students should notify professors in writing at the beginning of the semester of religious observances that conflict with their classes. The Office of the Provost, in consultation with Campus Ministry and the Registrar, will publish, before classes begin for a given term, a list of major religious holidays likely to affect Georgetown students. The Provost and the Main Campus Executive Faculty encourage faculty to accommodate students whose bona fide religious observances in other ways impede normal participation in a course. Students who cannot be accommodated should discuss the matter with an advising dean.

Statement on Sexual Misconduct

Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. However, university policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

Associate Director
Jen Schweer, MA, LPC
Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley
Trauma Specialist
Counseling and Psychiatric Services (CAPS)

(202) 687-6985
els54@georgetown.edu

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu.

Course Calendar

Week	Date	Topic
1	Sept. 4	Reproducibility
2	Sept. 9	Version Control
2	Sept. 11	Python Notebooks
3	Sept. 16	Data Types in Python
3	Sept. 18	Control Sequences, Iteration, and Functions
4	Sept. 23	Comprehensions and Generators
4	Sept. 25	Numpy
5	Sept. 30	Data Wrangling with Pandas (part 1)
5	Oct. 2	Data Wrangling with Pandas (part 2)
6	Oct. 7	Exploratory Data Analysis
6	Oct. 9	Vectors
7	Oct. 16	Trigonometry of Vectors
8	Oct. 21	Matrix Transformations
8	Oct. 23	Matrix Operations and Inversions
9	Oct. 28	Linear Regression
9	Oct. 30	Eigen Decompositions
10	Nov. 4	Decompositions in Practice
10	Nov. 6	Differentiation
11	Nov. 11	Optimizing Univariate Functions
11	Nov. 13	Optimizing Multivariate Functions
12	Nov. 18	Gradient Decent
12	Nov. 20	Constrained Optimization and Regularization
13	Nov. 25	Probability
14	Dec. 2	Bayes Rule & Naive Bayes Algorithm
14	Dec. 4	Simulation and MCMC Sampling
15	Dec. 9	Wrap Up
Final	Dec. 20	12:30pm-2:30pm - Location TBA

IMPORTANT: This syllabus is subject to change and may be amended throughout the course to reflect any changes deemed necessary by the professor. Any changes will be announced in-class or on Canvas.