NYU TANDON SCHOOL OF ENGINEERING

NYU School of Engineering, Computer Science and Engineering September 2019
CS 6053 Fall 2019

CS-GY 6053: Foundations of Data Science

About the Course and Goals:
This course offers students a practical, hands-on introduction to the growing field of "Data Science," and will equip them with the fundamental quantitative and computational analytics used. As data-access and data-driven methods become the norm in modern business and research environments, there is a growing demand for individuals who are able to derive meaningful insight from large, unruly data in a variety of domains. Simultaneously, students who are experienced in these approaches can also bring value to a variety of domains. The emphasis primarily is on understanding the fundamental concepts and applications of data science. We will cover several algorithms though this is not an algorithms course, nor a course in machine learning or computational theory. Our aim rather is to present fundamental algorithms within the context of a larger data mining and decision-making processes and use data in the wild to solve domain-area problems. In other words, this is more of a course about learning to think, rather than learning to code or use a certain package. A mix of skills, from data munging to machine learning and econometrics to effective visualization and communication will be comprehensively covered. Through a combination of data-intensive exercises and guest lectures by experts in the field, this course provides an overview of key concepts, skills, and technologies used by data scientists. The course materials, assignments and project will all be prepared using the python programming language.

Course Prerequisites: Machine Learning (CS 6923) is useful but can be taken concurrently. Students should be comfortable with basic mathematical and statistical skills in addition to having good programming skills; strong knowledge of statistics and linear algebra will be helpful. The main consideration is that the assignments and project are time-consuming, so be prepared to spend time and attention on this course.

Time and Location: Tu 3.20 PM - 5.50 PM 2 Metrotech, 8th floor, room 820

Instructor: Prof. Rumi Chunara, email: rumi.chunara@nyu.edu, office: 1106 in 370 Jay St.
Please make an appointment (see below).

Office Hours: Th 10:00-11:00 in 10.007, or by appointment (send email). I will also typically be available after class to chat.

Course Webpage: via NYU Classes

Textbooks: The following books are recommended as supplements and can be good references if you want to review or go further into detail on any topics.

Python reference: VanderPlas, Jake. "Python Data Science Handbook." (2016). O'Reilly Media, Inc.

Statistics reference: Bruce, Peter and Bruce, Andrew "Practical Statistics for Data Scientists." (2016). O'Reilly Media, Inc.

Statistical learning reference: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R.* New York: Springer (available online for free).

TA/Grader: Alex (Cheng) Shi, location 370 Jay St. 8th floor TA space. Time: TBA (or arrange by email). We will also use the NYU Courses forum for course questions and discussion. More details below and in class.

Grading Policy: Course assignments (30%), quizzes (25%), project (35%), participation (10%).

Assignments: There will be several assignments involving programming and written components. Every assignment has to be done individually by each student. Discussions between students and help in using the programming environment are permitted and encouraged (see below). However, no code or solutions may be copied! The assignments must be completed in Python. Homework will be due Tuesday morning at 9am. There is no late policy for assignments; if you think you should be exempt from the class policy, notifications should be communicated via #4 below.

Quizzes: There will be a series of very short quizzes at the beginning of selected classes covering material to-date. The lowest quiz mark will be dropped. Quiz marks will not be excused.

Project: The goal of this project is for the student to gain experience in understanding a substantive problem/question, acquiring data relevant to the problem/question, and applying machine learning techniques in an effort to address the problem/question. The project will be performed in teams, details to be announced. Students will be responsible for selecting a domain area problem (some example problems will be suggested) and finding the appropriate data. The grade for the project will be divided between problem and background research, proposed solutions (short presentation), and implementation. More details will be provided in class.

Course Outline: General topics covered; full timeline will be posted on the course website.

• Data handling/missing data approaches • Statistics recap • Introduction to Python • Time series analysis • Machine learning: Regression • Supervised machine learning, predictive models	• Guest lectures from domain area and industry experts • Model evaluation and tradeoffs • Visualization • Practical use of language models • Social network analysis • Ethics and data • Unsupervised learning

Policies and Expectations: Below are a few expectations and policies for the class.

1. It is expected that you make an attempt to learn the material in this course and participate in class.
   I expect you to try solving each assignment on your own. However, when being stuck on a problem, I **encourage** you to collaborate with other students in the class, subject to the following rules:

   - Code should be commented, and any submission should include all information such that someone else can reproduce what you did from what you submit.

   - You may discuss a problem with any student in this class and work together on solving it. This can involve brainstorming and verbally discussing the problem, going together through possible solutions, but should **not** involve one student telling another a complete solution.

   - Once you solve the homework, you must **write up your solutions and code on your own**, without looking at other people's write-ups, online solutions or giving your write-up to others.

   - In your solution for each problem, you must **write down the names** of any person with whom you discussed it. This will not affect your grade.

2. You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including having all electronic devices turned off and put away for the duration of the class (unless you are taking notes or working through examples or otherwise directed) and refraining from chatting or doing other work or reading during class.

3. Do not consult solution manuals or other people's solutions from similar courses.

4. There is no late policy (i.e. I will not accept any late submissions). If you have extenuating circumstances (such as a death in the family) that you believe would warrant exemption from the class policy, Deanna Rayment, deanna.rayment@nyu.edu, is the Coordinator of Student Advocacy, Compliance and Student Affairs and handles excused absences. She is located in 5 MTC, LC240C and can assist you should it become necessary.

5. If you are student with a disability who is requesting accommodations, please contact New York University's Moses Center for Students with Disabilities (CSD) at 212-998-4980 or mosescsd@nyu.edu. You must be registered with CSD to receive accommodations. Information about the Moses Center can be found at www.nyu.edu/csd. The Moses Center is located at 726 Broadway on the 3rd floor.

5. Start on the assignments early! Assignments may be more substantial than they appear on first sight. But others may be easier, one you figure out the necessary parts. But they all need some planning, so a last-minute approach is highly discouraged. Also, the TA or I cannot guarantee availability to answer questions right before a deadline.

6. Use the class forum. Your engagement on the forum will contribute to your class participation grade. Also, as above, since it's difficult to guarantee available to answer every email in a timely manner, we plan to check the forum regularly.

7. Again, do not copy code or any other work. There are tools that will discover this!

8. Students who need to should consult the Moses center for support and full accommodation will be made in class as necessary.

9. Everyone in the class will be held to the same standards.

10. This syllabus is subject to change. Updated information will be posted on the course website and/or discussed in class.