Fall 2023 DS-GA 1006 Capstone Project and Presentation
Syllabus

Description

This required course for the MS in Data Science should be taken in the second year of study. The purpose of the capstone course is to make the theoretical knowledge acquired by the students operational in realistic settings. While working on the capstone project, students see through the entire process of solving a real-world problem: from collecting and processing real-world data, to designing the best method to solve the problem, and implementing a solution. The problems and datasets come from real-world settings identical to what the student would encounter in industry, government, or academic research.

Students will work in small teams, guided by the course instructors and project mentors on a real-world data science research project. This course will encourage smaller group interactions between the instructors, project partners, and students, and among the students. The goal of this course is to model the real-world environment of working on complex research and development projects, with course instructors serving as advisors to the student teams. A large choice of projects will be available before the start of the course. Projects must have instructor approval (we will scrutinize the mentorship plan, data description and the description of metrics for success).

Throughout this course, we place a lot of emphasis on *mentorship*. In addition to the project mentors, each group will be assigned a member of the instructor team as the capstone co-mentor to ensure and track progress and troubleshoot if necessary.

Regular meetings with the project mentors and instructional mentors are a required part of this course, at least once every two weeks.

Instructors

- Brian McFee, Assistant Professor of Music Technology and Data Science, brian.mcfee@nyu.edu
- Jacopo Cirrone, Data Science / Colton Faculty Fellow, cirrone@courant.nyu.edu
- Elisha Cohen, Data Science Faculty Fellow, ec1302@nyu.edu
- Saadia Gabriel, Data Science Faculty Fellow, sg8390@nyu.edu

Project Administrator and Communication:

Please contact the academic admin team for any logistical questions surrounding the Capstone course: ds-capstone@nyu.edu

Use Brightspace for discussions with peers and troubleshooting help.

NYU Brightspace should be used for contacting the capstone course team.

- Please begin the subject with: [Capstone]
- Please cc the academic admin team on all emails pertaining to class: ds-capstone@nyu.edu

Prerequisites

This course is only open to MSDS students. The first year sequence of courses is a prerequisite:

- Introduction to Data Science
- Probability and Statistics

- Machine Learning
- Big Data

Structure

The capstone course consists of the following elements:

- Capstone Project: A large selection of projects will be distributed before the first day of class. You are expected to work in teams of 2-4 with your peers on one project for the entire semester. Working on the project constitutes the bulk of your time for this course.
- Mentorship: You will meet weekly with your capstone mentor and regularly (approximately every other week) with an assigned member of the instructional team. These meetings are required. Some of the class and lab time will be allotted to these meetings, but there will be scheduling flexibility and a flexible mix of in-person and virtual meetings.
- Guest lectures: 3-4 guest lecturers will provide their insights on working in Data Science in academia and industry.
- Mid-term presentation: You will orally present your project and initial results to your peers and the instructional team. This is a required in-person meeting.
- Final Report: You will write a final paper-style report which will be evaluated by your capstone mentors and the instructional team.
- Final poster presentation: In the last week of classes, you will present a poster on your project to the instructional team, mentors and your peers. This is a required in-person meeting. We might try to move this to an afternoon/evening spot to accommodate full-time working students.

Logistics

Lecture period: Mondays 4:55pm-6:35pm Location: GCASL C-95

Lab: Tuesdays 7:10pm-8:00pm Location: GCASL C-95

Assembling teams:

We will circulate a list of carefully chosen pre-approved projects in advance before the first course meeting (**Sep 11**). We recommend that you use Brightspace to find like-minded team mates by **Sep 15**. The instructors will then assign any remaining students that are not able to find partners.

Project assignment:

Teams will bid on pre-approved projects via a google form [link to be provided at a later date]. Each team will choose at least three projects in writing. The bid should include:

- An explanation of the reasoning for working on the project, and
- Team qualifications as they pertain to the project.

Data: You might be given access to proprietary data in many of the projects. Please do not share the data with anyone outside your team.

Code: Each team will create and use a private shared repository (e.g., GitHub) for their project. All code and project descriptions should be kept there.

Note on industry projects: You will notice that there are two types of projects, those coming from industry, and those coming from academic institutions (mostly NYU). Please know that when you work with our industry partners, a priori your work and your code belongs to you. It might be that some industry projects have company specific terms, which should be indicated in the project description. Note that you are in no way

required to sign anything you are not comfortable with, and of course you can also choose a non-industry project. Please approach your instructors if anything is unclear.

A note on attendance

Students will spend a significant fraction of face time on meetings with mentors and faculty advisers. These meetings can be either remote (on Zoom) or in-person. We require all full-time students to attend the guest lectures, mid-term presentations and final presentation in-person.

We are aware that part-time students might have conflicts with work-hours. We ask those students to get in touch with the instructor team to work out a remote solution before Sep 15. However, we note that the mid-term presentation (the one where your group presents) and the final poster presentation are required in-person meetings for all students (if public health guidelines allow). We will aim to move the final poster presentation to a later hour to hopefully facilitate attendance, but cannot guarantee this at this point.

Midterm presentations

In early November, every group has 4min to present their project. We will send a schedule a couple of days before. Please submit 4 Google slides (as per the provided template).

You will also be required to provide feedback on all other projects being presented on the same day you are presenting. This feedback will be anonymized and provided to each group.

Assessment

We will take the heterogeneity of projects into account for assessment and grading. The following will contribute to your final grade:

- Assessment provided by the capstone mentor
- Assessment by the capstone instructor
- Assessment of interim progress (meetings with capstone instructors, short summaries submitted by the teams)
- Midterm presentation
- Final Report
- Final Poster Presentation

The weighting of these components will be as follows:

- Mentorship/Instructor meeting attendance and assessment by mentors: 20%
- Midterm and final presentations and assessment by instructors: 30%
- Final report: 50%

Note that, formal assessment aside, one of the rewarding aspects of this course is that some projects become academic papers published in conferences or journals, and others become seeds for industrial prototypes.

Resources and References

The following references might be of use:

- MIT course: "Missing semester of CS" https://missing.csail.mit.edu/
- Reproducible deep learning course: https://www.sscardapane.it/teaching/reproducibledl/
- High-performance computing (HPC) at NYU https://sites.google.com/nyu.edu/nyu-hpc/hpc-systems