Syllabus for Capstone Project Course

Course Name:	Capstone Project and Presentation in Data Science		
Course Number:	DS-GA-1006 (3 credits)		
Year of the Curriculum:	Fall semester, second year		
Lecture:	Tues	5:10pm - 7:00pm	5 Washington Pl., Rm 101
Recitation:	Thurs	6:15pm - 7:05pm	5 Washington Pl., Rm 101
Course Director:	David K. A. Mordecai		mordecai (at) cims.nyu.edu
Co-instructor:	Michael O'Neil		oneil (at) cims.nyu.edu
Teaching assistants:	Kentaro Hanaki		kh1715 (at) nyu.edu
	Rahel Jhirad		rahel.jhirad (at) nyu.edu
Course email:	cds_capstone (at) cims.nyu.edu (Reaches all faculty team members)		
Course webpage:	http://cds.nyu.edu/ds-ga-1006-fall-2014/		

Course Description

The purpose of the Capstone Project is for the students to apply theoretical knowledge acquired during the Data Science program to a project involving actual data in a realistic setting. During the project, students engage in the entire process of solving a real-world data science project, from collecting and processing actual data to applying suitable and appropriate analytic methods to the problem. Both the problem statements for the project assignments and the datasets originate from real-world domains similar to those that students might typically encounter within industry, government, non-governmental organizations (NGOs), or academic research.

Depending on the project's complexity, students will work individually or in small teams on a problem statement, typically specified by a faculty, industry, or governmental sponsor. The sponsor will usually be responsible for supplying the relevant data set. Research groups (both from within, as well as external to NYU) may propose projects. A list of possible projects will be posted early in the semester so students can align themselves with problems statements corresponding to their individual interests. Pending approval by the Course Director, students are free to design their own problem statement and construct their own data set. As the project and problem statements warrant, students may be permitted to organize into teams of two to three participants. Teams larger than three will be considered for approval on a case-by-case basis. Each project team will be supervised by the Course Director (in some cases with a relevant faculty advisor) and advised by a Project Coach assigned from the academic, governmental, NGO or industry sponsor. The final problem statements and the composition of the teams will be approved by the Course Director.

Illustrative project examples

- A large insurance company has an anonymized dataset of worker compensation claims. The insurance claims dataset incorporates claimant demographics, claims payments, etc. A team comprised of capstone students, advised by the instructor in conjunction with a technical coach from the company, employ the dataset to develop and implement an analytic solution to reduce workplace injuries using software tools studied in previous courses.

- A professor from the Department of Politics has a dataset consisting of *tweets* from individuals, each labelled with some indication of the political party affiliation of the individual. Students use text classification methods studied in class to build infrastructure that can predict party affiliation and voting behavior.
- An astronomy professor has access to image data from the *Kepler* satellite of several stars outside of the solar system over a long period of time. Using time series analysis and Bayesian modelling, the team of capstone project students develops a method for statistically discovering the existence of new exoplanets based on the light intensity fluctuations of these stars.

Contact hours

The course will consist of one weekly lecture of 150 minutes, and one recitation (attendance required) of 50 minutes, where more focus will be paid to specific topics and one-on-one interaction with the course teaching assistants will be possible. The recitation section may also include workshops, seminars, or supervised research activities. Each instructor and teaching assistant will hold two office hours.

Course Aims

- Students will demonstrate an ability to handle a problem in data science from the point of problem definition through delivery of a solution. In doing so, they will demonstrate proficiency in collecting and processing real-world data, in designing the best methods to solve the problem, in implementing a solution, and quantifying the robustness and accuracy of their model.
- Students will demonstrate competence in presenting material by delivering two presentations: a proposal on how to approach the problem and their final solution.
- Students will learn how to work in small teams with at least one other student on their project.
- Students will write a report on their project for evaluation by the instructor(s) in consultation with the project advisors. The report will be structured as a typical research paper, and hence will include three main sections:
    a. Motivation, problem definition, and existing approaches
    b. Proposed solution and details of implementation
    c. Results, conclusion, and directions for future work

Prerequisites

Successful completion of the following courses within the Data Science Masters program curriculum:

- Introduction to Data Science
- Statistical and Mathematical Methods
- Machine Learning and computational statistics
- Big Data

Explicit permission of the course director is also sufficient, assuming that the student has previously completed similar course work or gained experience in hands-on projects.

Weekly topics

The weekly topics will change from year to year, please see the associated course schedule for details. Midway through the semester, students will be responsible for a brief presentation of their progress, which should include any roadblocks or difficulties that they've encountered. An end of semester final presentation will take place during the last week of classes.

Description of Project Requirements

- Demonstrate ability to carry out a data science project from end to end.
- Demonstrate proficiency in preparation and walk through of a presentation.

- Demonstrate ability to carry out a literature search and summarize the state of the art.
- Demonstrate ability to translate the project objects into a realistic work plan that draws on multiple people.
- Demonstrate ability to design and implement required software using tools such as R, MatLab, Torch, and traditional programming languages such as C, C++, Java.
- Demonstrate ability to professionally present the project plan and results.

Course evaluation

Students will complete an anonymous survey. The tabulated results will be reviewed by the instructor, the director of the program, and chair of the home department of the instructor. Issues will be identified and managed to successful remediation.

Grading

A letter grade will be given based on achievement on the project requirements listed above and class participation.

Bibliography and other resources

There is no required textbook for the course. Course instructors can recommend various references (texts and journal articles) particular to topics of interest.

Academic Integrity Policy

The course conforms to NYU's policy on academic integrity for students:
www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/academic-integrity-for-students-at-nyu.html
Students are responsible for reading and adhering to this policy.