Data Science Capstone & Ethics (ENGI E4800), Fall 2019

Instructors: Sining Chen and Adam Kelleher

Capstone committee: Sining Chen, Adam Kelleher, Eleni Drinea and Tian Zheng

Course overview

This course provides a unique opportunity for students in the MS in Data Science program to apply their knowledge of the foundations, theory and methods of data science to address data driven problems in industry, government and the non-profit sector. The course activities focus on a semester-long project sponsored by an affiliate company or a Columbia faculty member. The project synthesizes the statistical, computational, engineering and social challenges involved in solving complex real-world problems.

Projects (Industry and Faculty-Sponsored)

There are two types of projects. One type is sponsored by **industry** (or government/non-profit). This semester's industry sponsors include: Amazon, Bloomberg, Capital One, DiDi, Goldman Sachs, Johnson & Johnson, J.P.Morgan, KPMG, Moelis, Neoway, MSR NYC, Ralph Lauren, Unilever, Vanguard. Most of these projects are mentored by a mentor from the sponsor, and *supervised* by the course instructor or a Columbia faculty member.

The other type is offered by Columbia **faculty**. These projects are mentored by the professor (or postdoc) who proposed the project, and supervised by a course instructor.

Mentors, Supervisors, Instructors

Mentors' responsibilities include: set goals and milestones for the team, meet with the students weekly or biweekly, provide knowledge and guidance, monitor the progress, and provide the instructors with detailed assessment of the team and team members' work as a basis for grading.

Supervisors' responsibilities include: facilitate all aspects of the project, meet with and provide additional guidance to the students as needed, resolve issues.

Instructors: provide course level information, coordination and supervision. Give final grades based on participation/reports/posters, mentor/supervisors' assessments, self and peer assessments.

Course Format

Besides informational sessions when announced, there will be no class meetings. Instead, every team will be holding 30-60 min weekly or bi-weekly meetings with the mentor, and supervisor if applicable. Meetings can be in person or online.

Self-Grouping

4-5 students will work together as a team on each project. Students will be asked to form *groups* of sizes 1 to 5 before the start of the semester. Groups will submit project preferences as one unit (regardless of size). Note that groups of size less than 5 will likely be merged by the instructors to form a *team* of size 4-5 at the time of project assignment.

Team Forming and Project Assignment

During the first week of class, students will be provided information on all of the projects. Each Group (sizes 1-5) will submit preferences via google forms by the end of the week. Each group should select Top 5 projects(projects the group would LOVE to be assigned to) and Bottom 5 projects (those the group prefer NOT to be assigned to).

At the beginning of the 2nd week, groups will be merged into Teams and will be assigned to projects by the instructors based on interests, background, experiences and learning opportunities.

Note that there is NO guarantee that everyone on a team will get a project that s/he LOVEs. The instructors will do everything we can to not assign students to their Bottom 5 choices.

Team Work

Students are expected to meet often as a team in order for project and required deliverables to take shape. It is up to the team to manage their tasks and time effectively. Each team should use a Github repository to organize and communicate about their work. Each team should add their course instructor, and possibly industry mentor and/or faculty supervisor, to their Github repository.

Some past successful mode of collaboration also include:
--Selection of a team captain (with or without help from mentor/instructor/supervisor)
--Use of Slack (mentor, supervisors' participation are optional)
--Weekly progress log/meeting minute in a shared folder (rotate loggers). A sample weekly progress log/minute is provided in the appendix.

Typical project phases (may vary depending on the particular project)

Phase 1: (Weeks 1 & 2) Background and problem definition
Phase 2: (Weeks 3 & 4) Data collection, wrangling and cleaning
Phase 3: (Weeks 5, 6 & 7) Exploratory Data Analysis
Phase 4: (Weeks 8, 9 & 10) Coding prototypes of algorithms and models
Phase 5: (Weeks 11 & 12) Data visualization and reporting
Phase 6: (Week 13) Productionizing any models or algorithms, if applicable

Grades

At a high-level, the student's letter grade will be determined by the quality of their work, the written reports, the final poster presentation, professionalism, adjusted by the self/peer-assessment results. It is very likely that members of the same team do not receive the same grades.

1. **Weekly/Bi-weekly meetings and participation**

   Meetings and participation will account for 30% of the final grade. Evaluation will be based on clear and concise thinking towards achieving the research goals of the project, quality of discussions, participation and professionalism. In particular, students should attend **all** weekly team meetings in person or online, unless there is a medical or family emergency. (An interview is not an acceptable excuse.) Students should always be on time for meetings: students will be assigned letter grades individually, not as a team, and individual grades may be lowered if a student misses meetings or comes late or leaves early without prior approval.

   Further, team members (i.e., students) are responsible for (a) meeting all deadlines, including those set by the instructor and/or team mentor even if they do not appear in the syllabus, and (b) maintaining adequate progress throughout the semester. Teams should keep in mind that their team mentors will discuss with the instructors the quality of their work. Teams are encouraged to post informal summaries of weekly or bi-weekly discussions on the course piazza, to keep all parties informed.

   Other components of participation include: activity on github repos(for example, organizing the repo, actual code commits, resolving issues & answering comments, etc); attendance of team hacking sessions, and more.

2. **Progress Reports**

   Students are expected to write two original *progress* reports during the semester and one *final* report.

- The first *progress* report aims at synthesizing phases 1-3 of the project and is due on **Friday, October 18**.
- The second *progress* report aims at synthesizing phases 4-6 of the project and is due on **Monday, November 25**.
- The *final* report is a culmination of how the project synthesizes the statistical, computational, social and ethical challenges involved in solving complex real-world problems. The two *progress* reports will be essential in pulling together the *final* report, which is due on **Friday, December 20**.

Expected length of each report: 6-12 pages. More information about expectations will be communicated on courseworks closer to the deadlines. Late reports will not be accepted without prior discussion with the instructors. Students will be asked to explain their individual contributions to the work. Each progress report will account for 20% of the final grade, while the final report will account for 15% of the final grade.

3. Final poster session

Students will summarize and present their work in a poster presentation on Tuesday, December 10, 4-8pm. **Attendance is mandatory**; failure to attend will result in failing the course. The final poster session will account for 15% of the final grade. Industry affiliates and DSI faculty will all be invited to attend the poster session.

4. Ethics

Every Friday from 1:00pm-2:30pm, we will hold a “data for good” seminar (room TBA) as an optional part of this course. Internal and external speakers will present a topic related to data ethics, applications for societal good and advancing the state-of-the-art in data science. The final report will contain a discussion of "ethical challenges" and the seminar series will help the students prepare for that part of the report.

5. Self/Peer assessment

To promote fairness and collaboration among team members, we ask the students to fill out a self/peer-assessment form for the team at the end of the semester. The mentors will be asked to fill out a similar form for each member. The instructors will summarize the results and factor the assessments into the final grade. Please note that the students should not feel that they should score high on all items to receive a good grade. The best scenario is when each member is engaged and brings his/her unique strength to the team.

Key dates

- **Sep 6:** Course overview and Q&A with instructors on Friday, September 6, 1-2pm.
- **Sep 3-8:** Students review the projects and enter their preferences on surveymonkey.
- **Sep 9-10:** Team and project assignments by instructors.
- **Sep 10-11:** Introductory emails to teams, mentors and supervisors.
- **Sep 11 onwards:** Introductory meetings with teams, team mentors and instructors.
- **Oct 18**: First *progress* report due
- **Nov 25:** Second *progress* report due
- **Dec 10:** Final poster session (**4-8pm**, **mandatory attendance)**
- **Dec 20:** *Final* report due