Syllabus for Fall 2018
Practicum in Large-Scale Data Analysis and Processing (G5052)

Credits = 2 (P/F only)
Day/Time: TBD
Location: TBC

Gregory M. Eirich, QMSS
gme2101@columbia.edu
Office Hours in 807A IAB: TBA

Course Goals

This practicum course is meant to offer valuable training to students. Specifically, this practicum will mimic the typical conditions that students would face in an internship in a large data-intense institution. The practicum will focus on four core elements involved in most internships: (1) Developing the intuition and skills to properly scope ambiguous project ideas; (2) practicing organizing and accessing a variety of large-scale data sources and formats; (3) conducting basic and advanced analysis of big data; and (4) communicating and “productizing” results and findings from the earlier steps, in things like dashboards, reports, interactive graphics, or apps. The practicum will also give students time to reflect on their work, and how it would best translate into corporate, non-profit, start-up and other contexts.

The specific databases and datasets used will depend on instructor expertise and will shift based on needs of various partner organizations. For the Fall 2018 semester, the data will be derived primarily from NYC public data sources, and the ultimate deliverables will be aimed at the relevant agencies who are responsible for the data.

This practicum is meant for each student to develop their own ultimate individual deliverable, but there will be time set aside during each class for students to consult and share with one another.

Course Expectations

Attendance and Class Participation. Your attendance and participation are necessary at every meeting.

Intermediate Assignments. Students will have 6 large check-in assignments throughout the semester. These check-in assignments will ensure that students are making progress and any large-scale issues are being addressed before too much time has passed.

Plagiarism and Academic Dishonesty: Students must do all their work within the boundaries of acceptable academic norms. See the Academic Honesty page of the CU website regarding college policy on plagiarism and other forms of academic dishonesty -
http://www.columbia.edu/cu/history/ugrad/main/handbook/academic_honesty.html. Students found guilty of plagiarism or academic dishonesty will be subject to appropriate disciplinary action, which may include reduction of grade, a failure in the course, suspension or expulsion. This includes lab reports – if they are copied from another student, severe penalties may be applied.

**Students with Disabilities:** Students with special needs who may require classroom/assignment accommodations should make an appointment with me before or during the first week of class. You should also contact the Office of Disability Services (ODS) in Lerner Hall before the start of the course to register for these accommodations. The procedures for registering with ODS can be found at [http://health.columbia.edu/services/ods](http://health.columbia.edu/services/ods) or by calling (212) 854-2388.

**Late Assignments.** Students will lose points for handing in late assignments, at the discretion of the instructor and teaching assistant.

**Changes:** There may be adjustments in the scheduling of assignments, exams, and classrooms. Changes will be posted on Courseworks along with other announcements.
Proposed Schedule for the Practicum

Class #1 – Introduction

Part I: How to Scope an Ambiguous Project

Class #2 - What is the point of the project? Who is going to use this work? Talking to Stakeholders.
Class #3 - Proposing a “Minimally Viable Product.” Sketching out the final deliverable. Setting checkpoints.

Part II: Organizing and Accessing Large-Scale Datasets

Class #4 - Where is my data? What format is it in? How am I going to get it out? Basics of databases.
Class #5 - What other data might I need? Can I merge datasets together? On what basis?

Part III: Basic and Advanced Analysis

Class #6 - Visualizing and exploring the data. Dealing with data, measurement issues.
Class #7 - Finding associations among variables. Developing hypotheses.
Class #8 - Supervised and unsupervised data analysis techniques.
Class #9 - Wait, what’s the point of my project again? Am I on track? Do I need to recalibrate?

Part IV: Communicating and “Productizing” Results

Class #10 - How can I communicate what I have learned?
Class #11 - How can I make a tool that will help someone with their job? What are the key principles of U/X that should drive the creation of my tool?
Class #12 - How should interactivity and graphics be incorporated into my data product?
Class #13 - How can I make sure that my results are dynamic and can live beyond my time working on it?

Part V: Conclusion

Class #14 - Presenting the final products. Debriefing. What have we learned?