APPLIED DATA SCIENCE FOR SOCIAL SCIENTISTS

Columbia University
GR5069, Spring 2024
Tuesdays, 6:10PM-8:00PM
627 MUDD BUILDING

Instructor:	Marco Morales	
Email:	marco.morales@columbia.edu	
Office:	509E International Affairs Building	
Office Hours:	TBD, and by appointment	
Co-Instructor:	Nana Yaw Essuman	
Email:	nanayawce@gmail.com	
Office:	509E International Affairs Building	
Office Hours:	TBD, and by appointment	
TA:	Ludovico Genovese	Naveen Reddy Dyava
Email:	lg3148@columbia.edu	nd2794@columbia.edu

I. Overview

What this course is not. This is not a course in machine learning, deep learning, data visualization, applied statistics, or coding.

What this course is. This course is about the practice of Data Science in industry. It builds on the tools and techniques that you have learned in prior courses in machine learning, deep learning, data visualization, applied statistics or coding and takes them to the next level. For example,

- we will not teach you version control, but you will learn how to use it for collaboration with peers, and to create working environments and deployments
- we will not teach you how to train models and tune hyperparameters, but you will learn how to deploy models in production, how to use open source tools to facilitate model development and monitoring
- we will not teach you how to create data visualizations, but you will learn how to build frontends that use visualizations you create and automate connections to databases

This is a course to help students hit the ground running when they go into Data Science by teaching what people end up learning on the job. To do that, this course aims to:

i) teach processes and practices at the **intersection of Data Science and Data Engineering** that are central to the **data product cycle**. Data scientists typically start being exposed to Data Engineering and MLOps on the job. There's much to be gained from early exposure to concepts and practices in this field;

ii) sharpen technical skills not only at **fitting/training models**, but particularly at **building knowledge and generating insights** from the data. While this may seem obvious for a data scientist, it is not always the focus of standard training;

iii) train in **working effectively in teams** to build projects and products. Data Science is collaborative in nature and constantly evolving in **best practices** that enhance efficient workflows. Collaboration for school projects/assignments is vastly different from the **highly-structured collaboration** that happens in Data Science teams, but is not always the focus of training; and

All of these are highly valued skills in the Data Science job market, but not always considered explicitly as part of an integral Data Science curriculum.

Prerequisites: it is assumed that students have basic to intermediate knowledge of object-oriented programming — in **R** or **Python** — including experience using it for data manipulation, visualizations, and model fitting/training. Some mathematics, statistics and algebra will also be assumed.

Advisory: it is strongly suggested that this not be the first course you take in your Data Science sequence. You will get the full potential of the materials in the course if you take it after you have taken other substantive Data Science courses in your sequence.

II. Course Resources

The course will rely on a combination of curated reading materials, lectures, in-class workshops and take-home exercises that will leverage the following tools:

- There are no required textbooks for this course. Curated readings for each week's topic, as well as sample code and slides will be available in the course's **GitHub** repository. Starter code for in-class workshops and take-home exercises will be available in the course's **GitHub classroom** repository. (Please note that these are two (2) separate repositories!)
- **AWS** and **Databricks** will provide a host of tools to leverage data at scale.
- A **Slack** workspace for this course will serve as the primary means of written communication before, during and after class, where students can communicate with each other and with instructors

Instructions to get access to **GitHub classroom**, **Databricks**, **AWS**, and **Slack** will be available for registered students.

By week 2, make sure to have the latest versions of your preferred object-oriented language suite (**R** + **RStudio**, and/or **Anaconda**), and **git** installed on your computer. Sign up for a **GitHub** account if you don't have one already. Make sure also to have cloned the class repository to your computer.

III. Course Dynamics

Synchronous vs. Asynchronous Participation: This course is designed to have a combination of synchronous and asynchronous components to enhance your learning experience. It is our strong expectation that you will participate synchronously when required so that you can benefit fully from your peers and the live instruction. That said, it is completely understandable that your circumstances may make that very difficult, at least on some occasions. Please alert us when that is the case.

Expectation of Regular Participation and Utilization of Course tools: We will be monitoring student class participation and completion of assignments using the corresponding tools throughout the semester. We want to make sure that students are consistently engaged, and if that becomes difficult, that students alert us to their situations.

In preparation for each class:, you should have (i) read, thought about and be prepared to discuss all the curated readings in the course's **GitHub** repository for the week; (ii) watched the recorded video lecture in Canvas (when available); (iii) posted any questions you have on **Slack**; and (iv) completed any take-home assignments for that week.

During each live class: on average, we will devote the **first 60 mins of the class** to lectures where we will provide and contrast the Data Science and Data Engineering perspectives on each topic. We will spend the **remaining 60 minutes of the class** in hands-on workshops; have your laptop ready and be prepared to collaborate with your peers. Students will also be required to complete a number of **take-home exercises** to be submitted individually or in groups throughout the semester.

IV. Course requirements

The grade for this course will depend on the fulfillment of the following requirements:

(i) Attendance & Class Participation (20%): students are required to attend and actively participate in class exercises and discussions. Note that you will not obtain this 20% unless you actively participate on every session.

(ii) Take-home exercises (80%): students will receive exercises to enhance the learning process, and to serve as the basis for guided training

Late Submission Policy: All class assignments are expected to be submitted on the due date. For every day after the submission date, 10% of the maximum grade will be deducted from the score.

V. Course Outline

Fundamentals and Best Practices

TOPIC 1 - DATA SCIENCE AS A FUNCTION

Data Science in industry. The Data Science Shop. The Data Product Cycle. A Data Product Taxonomy. The crew that builds Data Products. Technologies and environments to build Data Products.

TOPIC 2 - VERSION CONTROL AND GITHUB

Understanding version control. Git on training wheels. GitHub and project collaboration. Some fun tricks, some necessary tricks.

TOPIC 3 - STRUCTURING YOUR WORKSPACE: DS & DE PERSPECTIVES

Portable and reproducible projects: how to do it? The Cookie-Cutter way, and why? DS projects from an engineering perspective - how and where (a cloud computing perspective).

TOPIC 4 - CODING ETIQUETTE

Not all code is created equal. Common faults when social scientists code. Getting closer to production-grade code. Appropriate messages when committing your code. Documentation and what that looks like.

TOPIC 5 - MANAGING THE PROCESS

Waterfall v Agile. Scrum vs Kanban. Minimum Viable Products (MVPs).

The Practice of Data Science

TOPIC 6 - DATA PIPELINE IN PRACTICE

Data Extraction. Data Transformation. Data aggregation. Data Storage.

TOPIC 7 - MISSING DATA AND DATA QUALITY

Missing data, implications and solutions. Data Quality: a Data Engineering Perspective. Automating data quality checks. How checks work and how to find faulty data.

TOPIC 8 - MODEL DEPLOYMENT, MODEL VERSIONING. WORKING ENVIRONMENTS (DEVELOPMENT, STAGING, PRODUCTION)

Batch models. Real-time models. Model monitoring and logging. Maintaining appropriate GitHub workflows. Model versioning.

TOPIC 9 - EXPLANATION VS PREDICTION

A common confusion: explanatory models $\neq$ predictive models. Distinguishing one from the other, strengths and limitations. Predictive models: classification and forecasting. When to use each? How to use each? When not to use them?

TOPIC 10 - MODEL EVALUATION

Evaluating models for Production. Evaluating models in Production.

TOPIC 11 - FRONTENDS AND DATA VISUALIZATIONS

Cognitive foundations of data visualization. Some best practices for data visualization. Building open-source frontends. Automating data visualizations in frontends.

TOPIC 12 - WORKFLOW COLLABORATION

Understanding how to work collaboratively within a data-driven organization, and knowing the roles and responsibilities of Data Engineers and Data Scientists.

Statement on Academic Integrity

Columbia's intellectual community relies on academic integrity and responsibility as the cornerstone of its work. Graduate students are expected to exhibit the highest level of personal and academic honesty as they engage in scholarly discourse and research. In practical terms, you must be responsible for the full and accurate attribution of the ideas of others in all of your research papers and projects; you must be honest when taking your examinations; you must always submit your own work and not that of another student, scholar, or internet source. Graduate students are responsible for knowing and correctly utilizing referencing and bibliographical guidelines. When in doubt, consult your professor. Citation and plagiarism-prevention resources can be found at the GSAS page on Academic Integrity and Responsible Conduct of Research.

Failure to observe these rules of conduct will have serious academic consequences, up to and including dismissal from the university. If a faculty member suspects a breach of academic honesty, appropriate investigative and disciplinary action will be taken following the Dean's Discipline procedures.

Statement on Disability Accommodations

If you have been certified by Disability Services (DS) to receive accommodations, please either bring your accommodation letter from DS to your professor's office hours to confirm your accommodation needs, or ask your liaison in GSAS to consult with your professor. If you believe that you may have a disability that requires accommodation, please contact **Disability Services** at 212-854-2388 or disability@columbia.edu.

Important: To request and receive an accommodation you must be certified by DS.