APPLIED DATA SCIENCE FOR SOCIAL SCIENTISTS

Columbia University
GR5069, Spring 2022
Mon, 10:10AM-12:00PM (EST)
TBD

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:	TBD
Email:	

I. Overview

In his now classic Venn diagram, Drew Conway described **Data Science** as sitting at the intersection of **good hacking skills**, **math / statistics knowledge**, and **substantive expertise**. Standard quantitative training in the social sciences supplies a fluid combination of all three, but tailored to **understanding human behavior**, and to **explaining why things happen the way they do**. Social scientists are, thus, a particular kind of data scientist.

This course is a collection of topics that fill very specific gaps identified over the years on what a social scientist should know at minimum when entering Data Science, and what a data scientist should be skilled at to add immediate value to their teams.

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 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 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 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

iv) enhance **soft skills** that are key to a successful interaction with business stakeholders. The most important — and often neglected — activity of a data scientist is to obtain expert knowledge from and communicate with non-technical audiences. The greatest insight/project/product is moot if no one outside the Data Science team understands it or its value.

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 estimation. Some mathematics, statistics and algebra will also be assumed.

II. Course Resources

The course will rely on a combination of curated reading materials, recorded lectures, live video-conferencing weekly sessions, 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 Educate classroom** and **Databricks Community** will provide a host of free 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 Community**, **AWS Educate classroom**, **Slack** and **FlipGrid** will be available on **Canvas** for registered students.

By week 2, make sure to have the latest versions of **R**, **RStudio**, **Anaconda**, and **git** installed on your computer. Sign up for a **GitHub** account if you don't have one already. **Atom** is also recommended to simplify your interaction with git and GitHub. Make sure also to have cloned the class repository to your computer.

III. Course Dynamics

Synchronous Participation vs. Asynchronous Participation: This course is designed to have a combination of synchronous and asynchronous participation 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. On those occassions, the synchronous portions can be done asynchronously as well. Likewise, assignments and some forms of participation can also be done asynchronously.

Expectation of Regular Participation and Utilization of Course tools: We will be monitoring student 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; (iii) posted any questions you have on **Slack**; and (iv) completed any take-home assignments for that week.

During each live class: we will devote the **first 30 mins of the class** to answer questions live, clarify concepts, and answer related questions where appropriate. We will spend the **remaining 90 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 throughout the semester.

Students will be working in teams on a **Course Project** to be completed and presented at the end of the course. We will spend the first 10 minutes of each class in a standup-like session to assess weekly progress and to connect expertise in the group to address unsolved problems. Teams must be prepared *every week* to provide evidence of their progress.

IV. Course requirements

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

(i) Class participation (10%): students are required to actively participate in class exercises and discussions. Note that you will not obtain this 10% unless you actively participate on every session.

(ii) Take-home exercises (30%): students will receive exercises to enhance the learning process, and to serve as the basis for guided training

(iii) Course project (40%): students will be randomly assembled into teams to work on a project that can be addressed with originality using Data Science tools. Development of this project will start on week 2.

(iv) Project presentation (20%): students will bring together all they have learned throughout the course in a 10-20 minute presentation of their project.

Note that more than half of your grade will depend on your final project. Note also that *how* you present your project is almost as important as the project itself.

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

WEEK 1 - OVERVIEW: DS & DE PERSPECTIVES

Introduction to data science and data engineering. Levels of engagement of data engineering on data science processes.

WEEK 2 - WORKSHOP: VERSION CONTROL AND GITHUB

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

WEEK 3 - SETTING UP PROJECTS: 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).

WEEK 4 - WORKSHOP: 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.

Applications

WEEK 5 - WORKSHOP: DATA PIPELINE IN PRACTICE

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

WEEK 6 - MISSING DATA AND DATA QUALITY

Automating data quality checks. How checks work and how to find faulty data. Missing data, implications and solutions.

WEEK 7 - MODEL DEPLOYMENT, MODEL VERSIONING. WORKING ENVIRONMENTS (DEV, STAGING, PROD)

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

WEEK 8 - ACADEMIC HOLIDAY

WEEK 9 - WORKSHOP: INTERACTIVE WORKING SESSION

Working in a collaborative environment. Working on a classmate's GitHub repo

WEEK 10 - 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?

WEEK 11 - MODEL EVALUATION

Evaluating models for Production. Evaluating models in Production.

Data Science Outputs

WEEK 12 - DATA VISUALIZATIONS

Cognitive foundations of data visualization. Building appropriate visualizations for a specific purpose and audience. Storytelling with data. Some best practices for data visualization. Data visualization for non-technical audiences.

WEEK 13 - WORKFLOW COLLABORATION

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

WEEK 14 - PRESENTING RESULTS TO BUSINESS STAKEHOLDERS

Make sure you are answering the right question! Storytelling with decks. Elements for a successful "deck". Who's afraid of the big bad graph? One-liners are good headliners. What to include and for whom.

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.