TOPICS IN APPLIED DATA SCIENCE FOR SOCIAL SCIENTISTS

Columbia University
GR5069, Spring 2019
Weds, 6:10PM-8:00PM
270B INTERNATIONAL AFFAIRS BUILDING

Instructor:	Marco Morales
Email:	marco.morales@columbia.edu
Office:	509E International Affairs Building
Office Hours:	Weds 8-9PM, and by appointment
TA:	Ummugul Bezirhan
Email:	ub2126@tc.columbia.edu

I. Overview

In his now classic Venn diagram, Drew Conway described *Data Science* as sitting at the intersection between **good hacking skills**, **math and statistics knowledge**, and **substantive expertise**. As a result of normal instruction, social scientists possess a fluid combination of all three but also bring an additional layer to the mix. We have acquired slightly different training, skills and expertise tailored to **understand human behavior**, and to explain **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 know to hit the ground running and add immediate value to their teams.

To do that, this course aims to:

- sharpen you 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,
- 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 DS teams, but is not always the focus of training,

- learn 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; and
- sharpen and 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 **R**, including experience using it for data manipulation, visualizations, and model estimation. Some mathematics, statistics and algebra will also be assumed.

II. Course Resources

There are no required textbooks for this course, but you might find these to be very useful resources for the course and later in your careers:

- Hadley Wickham and Garret Grolemund. *R for Data Science*. O'Reilly Media, Boston, MA, 2017
- Hadley Wickham. *Advanced R*. Taylor & Francis Group, Boca Raton, Fl, 2014
- Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY, second edition edition, 2016
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, NY, 2013
- Scott E. Page. *The Model Thinker: What You Need to Know to Make Data Work for You*. Basic Books, New York, NY, 2018

All required readings will be available on **Canvas**. Additional code and materials will be available on the class **GitHub** repository. A **Slack** workspace for this course is available with multiple channels to facilitate team collaboration within the class.

For the vast majority of coding problems you will encounter, it is very likely that someone has faced a similar one (and solved it!) in the past. Take advantage of collective wisdom and make extensive use of resources like **Stack Overflow** where you will most likely be able to find an answer and a snippet of code that goes with it.

Be mindful of the ***15-minute rule***: if you haven't been able to figure out an answer in 15 minutes, reach out to the person next to you or post your question in the appropriate **Slack** channel.

By the second session, make sure to have the latest versions of **R**, **RStudio**, and **Git** installed on your computer. If you don't have one already, create a **GitHub** account to use throughout the course. **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 requirements

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

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

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

(iii) Final project (30%): students will be randomly assembled into teams of at least 3 and choose a social science problem 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 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.

IV. Course Dynamics

The course will be alternating between three domains:

i) **topical lectures** will discuss specific topics in depth to underscore specific technical aspects or applications

ii) **applied workshops** will focus on hands-on learning for students to learn interactively

iii) **guest talks** will feature an external data scientist to lecture on specific topics or applications from their everyday domain and perspective

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 adress unsolved problems. Teams must be prepared *every week* to provide evidence of their progress.

V. Course Outline

WEEK 1: INTRODUCTION

Introduction. What this course is (and what it is not). Course overview. What's different about Social Scientists doing Data Science? Why focus on these skills and best practices in Data Science? What should you expect from this course?

WEEK 2: TOPIC - BEST PRACTICES IN DATA SCIENCE FOR SOCIAL SCIENTISTS

Coding is not enough. The (Social) Data Scientist toolkit: between ML and classical statistics. No Agile, no cry? Project collaboration and version control.

WEEK 3: TOPIC - SETTING UP A DATA SCIENCE PROJECT

Reproducible projects. Portable projects. Structuring Data Science projects: the cookiecutter way.

WEEK 4: WORKSHOP - VERSION CONTROL AND GITHUB

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

WEEK 5: TOPIC - DATA PRODUCT CYCLE (GUEST: NANA YAW ESSUMAN)

A view from Data Engineering. Getting data. Storing data. Accessing data. Transforming data (real-time or batch?). Utilizing data in models. Storing model outputs. Monitoring data quality. Defining an "end" to the data cycle. Optimization. Collaboration with Data Engineers at each step of the cycle. How to ask the right questions?

WEEK 6: WORKSHOP - DATA VISUALIZATION: BEYOND PRETTY GRAPHS

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 7: WORKSHOP - CODING ETIQUETTE FOR SOCIAL SCIENTISTS

Not all code is created equal. Common faults when social scientists code. Getting closer to production-grade code. Appropriate messages when committing your code.

WEEK 8: TOPIC - MISSING DATA

Why is missing data important? How to think of missing data. Model-based data imputation. Algorithm-based data imputation.

WEEK 9: ACADEMIC HOLIDAY

WEEK 10: WORKSHOP - WORKFLOW COLLABORATION (GUEST: NANA YAW ESSUMAN)

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

WEEK 11: TOPIC - 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 12: TOPIC - 3 ALGORITHMS I

The best kept secret: a good insight always trumps perfection. A peak under the hood on your most used algorithms for analysis: linear regression, logistic regression, random forest. Why, when and for what?

WEEK 13: TOPIC - 3 ALGORITHMS II

How much mileage can you extract from each algorithm? Exploring what you can get in practice. How far, how fast, how good?

WEEK 14: TOPIC - CONDITIONAL RELATIONSHIPS IN THE DATA

When the values of one variable depend on the values of other variable(s). Parametric and non-parametric modeling of conditional relationships in the data. Interpreting conditional relationships. Actionable insights from conditioning variables.

WEEK 15: WORKSHOP - 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.