TOPICS IN APPLIED DATA SCIENCE FOR SOCIAL SCIENTISTS

Columbia University
GR5069, Spring 2018
Weds, 6:10PM-8:00PM
270B INTERNATIONAL AFFAIRS BUILDING

Instructor:	Marco Morales
Email:	mam2519@columbia.edu
Office:	270 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**. By training, 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 not intended to teach you how to code, create visualizations, or estimate models. It presumes you have learned that in other classes. This course is intended to take you to the next level in becoming a data scientist. Therefore you will:

- sharpen your technical skills and better align them with common business use cases and expectations,
- learn current best practices in data science that will facilitate collaboration with data scientists trained in engineering or other hard sciences, and
- learn soft skills that are key to a successful interaction with business stakeholders.

All of these are highly valued skills in the data science job market, but seldom considered as part of an integral training for data scientists.

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, econometrics 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*. OReilly Media, Boston, MA, 2017
- Hadley Wickham. *Advanced R*. Taylor & Francis Group, Boca Raton, Fl, 2014
- Winston Chang. *R Graphics Cookbook*. O'Reilly Media, Boston, MA, 2013
- Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY, second edition edition, 2016
- Drew Conway and John Myles White. *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started*. O'Reilly Media, Boston, MA, 2012

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. But 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 **Slack**.

By the third session, make sure to have the latest versions of **R**, **RStudio**, and **Git** on your computer. Also, create a **GitHub** account to use throughout the course. 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:

Class participation (20%): students are expected to have read all required readings before class and actively participate in class discussion. This course is designed to be applied with in-class discussion.

Data challenges (30%): students will receive data challenges where they will push the envelope on the code that has been provided as an example.

Final project (30%): students will be randomly assembled into teams, and each team will choose a social science problem that can be addressed with originality using Data Science tools.

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.

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 business 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

The final component will be collaboration between students to generate a product by the end of the course. Data Science is collaborative in nature. No single individual possesses all knowledge and tools to solve every problem, so problems are tackled in groups that leverage everyone's knowledge and expertise. Hence, the final project will be carried out in groups of at least 2, depending on class size. Make sure to take advantage of **Slack** for group collaboration.

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? Slacking as a way of life. Project collaboration and version control.

WEEK 3: WORKSHOP - 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 (and beyond). GitHub and project collaboration. Some fun tricks, some necessary tricks.

Students will receive data challenge #1

WEEK 5: WORKSHOP - DATA VISUALIZATION FOR TECHNICAL AND NON-TECHNICAL AUDIENCES

Understand your data (then conquer)! Low hanging fruit with data explorations. What constitutes a good data visualization? Storytelling with data. Some best practices for data visualization. Data visualization for non-technical audiences.

Data challenge #1 due at 6PM

WEEK 6: 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 7: 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 8: 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?

Students will receive the data challenge #2

WEEK 9: ACADEMIC HOLIDAY

WEEK 10: GUEST TALK - CODING ETIQUETTE FOR SOCIAL SCIENTISTS

Coding as a social activity. Not all code is created equal. Pseudo-code v code. Defensive coding. Common faults when social scientists code. Getting closer to production-grade code.

Data challenge #2 due at 6PM

WEEK 11: TOPIC - CONDITIONAL RELATIONSHIPS IN THE DATA

Not all relationships are linear. Not all non-linear relationships can be uncovered in ML. Modeling and interpreting conditional relationships. Actionable insights from conditioning variables.

Students will receive the data challenge #3

WEEK 12: GUEST TALK - DATA SCIENCE FOR GOOD

Data scientists working on relevant social science projects will share their experiences.

Data challenge #3 due at 6PM

WEEK 13: 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.

WEEK 14: DEEP DIVE FOR TEAM PROJECT REVIEW

Prior to the final presentation before a panel of data scientists, we'll engage in a deep review of projects, presentations. Teams will receive feedback for a successful presentation of their project.

WEEK 15: PROJECT PRESENTATIONS

Bringing together all skills and best practices acquired in class, students will engage in a 10-20 minute presentation of their final project before business stakeholders. Students should be prepared to answer questions, and be ready to translate complex concepts/models to easy-to-understand arguments that can be understood by non-technical audiences.

Project presentations will de done before a panel of industry practitioners, for a taste of real world experience.

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 (https://gsas.columbia.edu/student-guide/research/academic-integrity-and-responsible-conduct-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 Dean's Discipline procedures (https://gsas.columbia.edu/student-guide/policy-handbook/deans-discipline).

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.