TOPICS IN APPLIED DATA SCIENCE FOR SOCIAL SCIENTISTS

Columbia University
GR5069, Spring 2017
Weds, 6:10PM-8:00PM
644 SEELEY W MUDD BUILDING

Instructor:	Marco Morales
Email:	mam2519@columbia.edu
Office:	270 International Affairs Building
Office Hours:	Weds 8-9PM, and by appointment
TA:	Vivian Liu
Email:	ytl2102@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:

- learn current best practices in data science that will facilitate collaboration with data scientists trained in engineering or other hard sciences,
- learn soft skills that are key to a successful interaction with business stakeholders, and
- get exposed to data science practitioners and explore real-life applications from a social science perspective.

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 CourseWorks. Additional code and materials will be available on the class GitHub repository. A Slack team for this course will be 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.

By the second session, make sure to have the latest versions of **R**, **RStudio**, and **Git** on your computer. Make sure also to have cloned the class repository to your computer.

III. Course Dynamics

The course will be alternating between two domains: **topical** and **best practices**.

i) For the **topical** domain, the course will address three topics on two sessions per topic:

   - **session 1:** we'll discuss all relevant information to the topic, be it technical, substantive context, deeper discussion on methods.

  - **session 2:** a guest speaker will demonstrate a specific application of data science on this topic from a social science perspective, delving into the code / statistics / substance of it. After the presentation, students will receive a data challenge typically based on this example, having a week to solve it.

ii) For the **best practices** domain, the format changes to applied workshops where students gain hands-on experience taking technical skills (e.g. visualizations) to the next level, or learning soft skills (e.g. presenting results to business stakeholders).

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.

We will spend the last 20-30 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.

Students who are not familiar with these tools should take advantage of the following sections to facilitate their interaction with the course:

- *Git and version control*; all class examples and code will be in a GitHub repository, be ready to make ample use of it
- *visualizations in R*; packages like ggplot, and plotly, and tools like Shiny allow for quick spin ups of dynamic and static visualizations that always come in handy
- *tidyverse tools*; the tidyverse is making strides both in non-distributed and distributed computing, might as well master it early on

IV. Course 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%): at the end of each guest presentation, students will receive a data challenge 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.

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 soft skills and best practices in Data Science? What should you expect from this course?

Week 2 - WORKSHOP: Best Practices in Data Science for Social Scientists

Coding is not enough. Structuring Data Science projects: the cookiecutter way. The (Social) Data Scientist toolkit: between ML and classical statistics. No Agile, no cry? Git on training wheels (and beyond). Slacking is a way of life.

Required Readings:

- David Donoho. 50 years of data science. Mimeo MIT, September 2015

Week 3 - TOPIC 1 — Forecasting: Introduction and substantive discussion

What is really a forecast?...really. What is forecastable (and what is not)? Ways, means and tools to forecast. Fitting models vs forecasting. Confidence intervals and prediction intervals. Cross-validation and holdout. Forecast ensembles.

Required Readings:

- J. Scott Armstrong, Kesten C. Green, and Andreas Graefe. Golden rule of forecasting: be conservative. *Journal of Business Research*, 68(8):1717–1731, 2015
- J. Scott Armstrong. Evaluating forecasting methods. In J. Scott Armstrong, editor, *Principles and Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic Publishers, Norwell, MA, 2001
- J. Scott Armstrong. Illusions in regression analysis. *International Journal of Forecasting*, 28(3):689–694, 2012

Week 4 - TOPIC 1 — Election Forecasting, invited guest: David Rothschild

Our guest for this week is David Rothschild, who is an economist at Microsoft Research, where he specializes in forecasting and public opinion and public sentiment. He is also one of the founders of PredictWise, a "website for a project that studies the collection of individual-level data for predictions, aggregation of that data into prediction, and usage in predictions", per their own description.

Required Readings:

- David Rothschild and Sharad Goel. When you hear the margin of error is plus or minus 3 percent, think 7 instead, 2016. October 5
- David Rothschild. Pollfish election 2106, 2016. URL `http://predictwise.com/blog/2016/11/pollfish-election-2016/`. November 10
- David Rothschild. Markets converge to poll, 2016. URL `http://predictwise.com/blog/2016/11/markets-converge-to-polls/`. November 17

Week 5 - WORKSHOP: data visualizations 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. Spinning up dynamic v static data visualizations.

Required Readings:

- Andrew Gelman. Why tables are really much better than graphs (with discussion). *Journal of Computational and Graphical Statistics*, 20(1):3–40, 2011
- Alberto Cairo. *The Functional Art: An Introduction to information graphics and visualization.* New Riders, San Francisco, 2013. Ch 5 -The Eye and the Visual Brain & Ch 6 -Visualizing for the Mind

Planning session: teams will define the theme for their final project and how they plan to address it.

Students will receive the first data challenge.

Week 6 - The three little algo(rithm)s: an overview

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?

Required Readings:

- Hal Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014
- Thomas Brambor, William R Clark, and Matt Golder. Understanding interaction models: Improving empirical analyses. *Political Analysis*, 14(1):63–82, 2006

Data challenge due at 6PM

Week 7 - The three little algo(rithm)s: a tale of three outputs

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 second data challenge.

Week 8 - WORKSHOP: 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.

Required Readings:

- Jonathan Nagler. Codig style and good computing practices. *PS: Political Science & Politics*, 28(3):488–492, 1995
- Hadley Wickham. *R Packages*. O'Reilly Media, Boston, MA, 2015. (ch 3)

Week 9 - Academic Holiday

Week 10 - TOPIC 2 — Text as data: Introduction and substantive discussion

It doesn't get more human than text. Beyond NLP and closer to home. The statistical and algorithmic analysis of text. Quanteda: the tidyverse of text analysis.

Data challenge due at 6PM

Week 11 - TOPIC 2 — Text as Data, invited guest: Patrick Chester

Our guest for this week is Patrick Chester who's doing his PhD at NYU. He specializes in comparative politics and political methodology, is interested in how information obtained from social media impacts protest behavior, using text to delineate delegation structures of governments, and better understanding how resource scarcity impacts international conflict.

Required Readings:

- Justin Grimmer and Brandon Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political documents. *Political Analysis*, 21 (3):267–297, 2013

Week 12 - WORKSHOP: presenting results to business stakeholders, invited guest: Jason Gilbertson

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.

Students will present a draft of their final project presentation for extensive feedback from a guest data scientist.

Week 13 - TOPIC 3 — Analyzing collective individual behavior

Analyzing behavior, beyond classification. Asking (and responding) why people behave the way they do. A longitudinal analysis of people. The time-series cross-sectional approach. The duration/event history analysis approach. Time permitting, some additional approaches (i.e. diff in diff).

- Nathaniel Beck. Time-series cross-section methods. In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, editors, *The Oxford Handbook of Political Methodology*. Oxford University Press, New York, NY, 2008
- Suzanna De Boef and Luke Keele. Taking time seriously. *American Journal of Political Science*, 52(1):184–200, 2008

Week 14 - Deep dive for individual project reviews

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 (http://gsas.columbia.edu/academic-integrity).

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 (http://gsas.columbia.edu/content/disciplinary-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.