Introduction to Data Science

Course Design

Xiaoyang Ye
xiaoyang.ye@brown.edu

1 Course Description

> "They [data scientists/computational social scientists] must articulate how the combination of academic, industrial, and governmental collaboration and dedicated scientific infrastructure will solve important problems for society - saving lives; improving national security; enhancing economic prosperity; nurturing inclusion, diversity, equity, and access; bolstering democracy; etc." *(Lazer et al., 2020, Science)*

As "science of planning for, acquisition, management, analysis of, and inference from data" (NSF 2014), data science is emerging as a field that is revolutionizing science and industries alike. Work across nearly all domains is becoming more data driven, affecting both the jobs that are available and the skills that are required. As more data and ways of analyzing them become available, more aspects of the economy, society, and daily life will become dependent on data.

Today, the term "data scientist" typically describes a knowledge worker who is principally occupied with analyzing complex and massive data resources. However, data science spans a broader array of activities that involve applying principles for data collection, storage, integration, analysis, inference, communication, and ethics.

The goal of this undergraduate-level course is to introduce students to the field of data science. **A key goal is to give all students the ability to make good judgments, use tools responsibly and effectively, and ultimately make good decisions using data.**

This course will repeatedly engage students in the full cycle by which we learn from data and prepare students to immediately use a high-level language to explore, visualize, and pose questions about data. With project-based activities in empirical settings throughout the course, students will also acquire skills in data acquisition, modeling, management and curation, data visualization, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving.

With students from varied backgrounds and degrees of preparation, this course is committed to preparing them for success in a variety of careers. The course design and practice will emphasize on the balance of "teaching computer scientists research design" and "teaching social scientists how to code."

Extensions: This course can be extended to a graduate-level (e.g., master's, Ph.D.) course with a more algorithmic language and more intensive coding practices. The course design is also flexible to the language choice of R or Python.

2 Outline

Part I: Introduction to the Field

1. Introduction: What is data science and what does data scientists do?
2. Methods and applications: Why data science?

Part II: Foundational skills for data science

3. Getting started with R and R Studio
4. R basics: Exploring R with the {*swirl*}
5. Expanding the toolbox: Python, R Markdown, Jupyter Notebook
6. Open science and reproducibility: Github, pre-registry, Open Science Framework (OSF)

Part III: Communicating with data

7. 80% of data science is to clean data {*tidyverse*}
8. Visualized storytelling ({*ggplot*} & {*shiny*})

Part IV: Introduction to machine learning

9. Supervised learning 1: Decision tree, random forest, boosting
10. Supervised learning 2: Regression models, SVM, Naive Bayes
11. Unsupervised learning: clustering & PCA
12. Text-as-data

Part V: Data science is evolving very fast!

13. Selected topics: deep learning, NLP, automl, and more responsible AI
14. Selected topics: causal inference in data science
15. Selected topics: data science in practice (speakers from academia, industry, or government)

3 Assignments, Activities, and Projects

3.1 Participation

The course will be offering various class activities (e.g., poll, survey, post) and students are expected to complete them.

3.2 Short review/reflection paper & leading discussions

Each week, guided by a few reading questions, students will write a short paper to indicate a comprehension of the main issues characterized in the field of data science. The topics will focus on ethnics, bias, fairness, and reproducibility. Students, individually or in group of 2-3, will be responsible for a short presentation on the readings to class and set the stage for class discussion.

3.3 Weekly quizzes

In order to ensure all the students meeting the learning goals, students will take a low-stakes weekly quiz, which covers questions about the concept, theory, and/or coding skills. This quiz is designed to be low-stakes that students can retake the quiz as needed.

3.4 Labs

Students will work individually or in group for the weekly lab assignments. These assignments will offer opportunities for students to learn and practice their analytical and computing skills. Assignments are mostly submitted using Jupyter Notebook and Github. Few exceptions will be in other formats such as a webpage or a shinny App.

3.5 Notebook

One output for this course will be a self-created notebook or the class. The idea is that ultimately this notebook might prove useful as a reference for their own work. The notebook will include material on the assigned readings, summaries of class participation and lectures, documented R/Python codes.

The notebook will be written using *Jupyter Notebook* in a clear style, be well organized, and one hopes, useful to students when the course is over.

3.6 Final project

Students will work in group on a research question with data-driven solutions, give a presentation, and submit a product report. Two possible choices might be (1) that all the teams work on the same project (like Kaggle competition), or (2) that students work on projects of their choice. This will be a practicing example for students' larger-scale capstone projects in the coming semesters.

4 Reading List (Preliminary)

“Textbooks”

- Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* O’Reilly.
- R for Data Science [Github repository & Bookdown]

Additional books

- Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R (Second Edition).* New York: Springer. [Online book]
- Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge University Press.
- Hardt, M., & Recht, B. (2021). *Patterns, Predictions, and Actions: A Story About Machine Learning.* [Online book]
- *Practical Statistics for Data Scientists* [Github repository & Jupyter notebook]
- *Introduction to Machine Learning with Python* [Github repository]

Further reading

Introduction of data science

- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.
- Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3-28.
- Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science, 355*(6324), 483-485.

Predictability

- Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. arXiv:1907.07592.
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... & McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences, 117*(15), 8398-8403.
- Lepri, B., Oliver, N., & Pentland, A. (2021). Ethical machines: the human-centric use of artificial intelligence. *Iscience,* 102249.

Text-as-data

- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535-74.
- Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in highâĂŘdimensional choices: method and application to congressional speech. *Econometrica, 87*(4), 1307-1340.
- Liu, J., & Cohen, J. (2020). Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *EdWorking PaperNo, 20-239.*
- Adukia, A., Eble, A., Harrison, E., Runesha, H. B., & Szasz, T. (2021). What We Teach About Race and Gender: Representation in Images and Text of ChildrenâĂŹs Books. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2021-44).

Causal inference in data science

- Hassani, H., Huang, X., & Ghodsi, M. (2018). Big data and causality. *Annals of Data Science, 5*(2), 133-156.
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228-1242.
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685-725.
- Scholkopf, B. (2019). Causality for machine learning. *arXiv*:1911.10500.
- Liu, T., Ungar, L., & Kording, K. (2021). Quantifying causality in data science with quasi-experiments. *Nature Computational Science, 1*(1), 24-32.

Behavioral data science

- Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M., & Mainen, Z. F. (2014). Big behavioral data: Psychology, ethology and the foundations of neuroscience. *Nature Neuroscience, 17*(11), 1455-1462.
- Giest, S., & Mukherjee, I. (2018). Behavioral instruments in renewable energy and the role of big data: A policy perspective. *Energy Policy, 123*, 360-366.
- Betsch, C. (2020). How behavioural science data helps mitigate the COVID-19 crisis. *Nature Human Behaviour, 4*(5), 438-438.
- Kasy, M., & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica, 89*(1), 113-132.