Brown University, Economics 1660, Fall 2020

Big Data

Instructor: **Daniel Björkegren**
danbjork@brown.edu
Office Hours:
Thursdays 2-3pm ET
Sign up in advance at
http://danbjork.youcanbook.me
zoom link will be emailed

TA: **José Andrés Jurado**
jose_jurado_vadillo@brown.edu
Office Hours:
Mondays 9-9.45 am ET
Wednesdays 11-11.45am ET
Fridays 5-5.45pm ET
https://brown.zoom.us/j/9080429911

If these hours do not work for you, you may email to ask about other potential meeting times.

The spread of information technology has led to the generation of vast amounts of data on human behavior. This course explores ways to use this data to better understand the societies in which we live. The course weaves together methods from machine learning (OLS, LASSO, trees) and economics (reduced form causal inference, economic theory, structural modeling) to work on real world problems. We will use these problems as a backdrop to weigh the importance of causality, precision, and computational efficiency.

Best practice for the methods we will be exploring are still evolving, so this course will be experimental both in terms of content and pedagogy.

Prerequisites are Econ 1110 or 1130; Econ 1629 or 1630; and an introductory computer science course (CS 040, 112, 150, 170, or 190). Knowledge of econometrics and programming is assumed.

The class includes a synchronous online meeting: Tuesdays 4-6.20pm US Eastern Time
As well as asynchronous content.

Class readings will consist of research papers, which will be made available on the course Canvas website, supplemented by textbook readings from the following books (optional):

- *The Elements of Statistical Learning*, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (HTF). An electronic version is available free from http://statweb.stanford.edu/~tibs/ElemStatLearn/ . You can also access it via SpringerLink, where a soft cover version can be purchased for $25.
- *Causal Inference in Statistics: A Primer*, by Judea Pearl, Madelyn Glymour, and Nicholas Jewell. (PGJ)

Problem Sets

Problem sets are designed to help you understand the methods developed in this course. We will use real world problems (such as those faced by businesses, nonprofits, and governments) to prompt the methods. You'll be asked to develop the methods from scratch, which will involve mathematical proofs and coding algorithms in Python and R. You'll be asked to link this mathematical understanding back to the real world setting, by weighing different theoretical approaches and interpreting statistics in light of policy choices.

Some problem sets will require additional interaction outside of class, including the optimal pricing competition (marked with * on the syllabus). More details will be given closer to the date.

Problem sets must be turned in online through Canvas by 3pm the day before class. Include your writeup as a PDF or Word document. Since we will often discuss the results of the problem set in class, late submissions will not be accepted.

It can be helpful to learn these concepts in a group, so you may work on the problem sets in a small group (up to 3), except where noted otherwise. However, the purpose of the problem sets would be defeated if you obtained answers but not understanding from your group. For this reason, answers must be written up individually, in your own words, unless directed otherwise. Please write the names of your study group member(s) on your problem set. Duplicate answers will be penalized as if the assignment were not submitted at all, and subject to disciplinary review. Problem sets will be graded based on process and observed effort in obtaining a solution, and not simply whether the answers are 'correct.'

Final Project

The course will culminate with a final project that allows you to apply the methods you've learned to a selected topic. Projects will involve programming, statistical interpretation, and conveying your findings in a short written document. It is recommended that you work in a group of up to 3 people for each project. Each group can submit one assignment jointly. In the final week of the course each group will have the opportunity to present their project. These presentations will allow the rest of the class to learn about your work as well as provide ideas and feedback. A final written document will be due during reading period.

Grading

Coursework will consist of the following, comprising the following portion of the final grade:

Problem sets	50% total (each weighted equally)
Final project	30%
Class participation	20%

<u>Problem Set Standards</u>

Problem sets are designed to be challenging, and will be graded by this rubric:

	Excellent	Good	Poor
Engagement (40 points)	Deep engagement with the problems. Attempts to conclusively answer deeper questions (puzzles) that arise. When problems cannot be solved with textbook approaches, generates new approaches to solve them. Comments on the desirability of these approaches. (35-40 points)	Good engagement with the problems. Attempts to solve deeper questions that arise, but does not go out of its way to do so. (20-34 points)	Does not engage with all problems, or misses the presence of deeper questions. (19 or below)
Understanding (40 points)	Writing displays a clear understanding of the problem, the methods, and associated limitations. (35-40 points)	Writing is unclear or does not demonstrate full understanding of the problem/methods/limitations. (20-34 points)	Writing is unclear or demonstrates substantial misconceptions. (19 or below)
Writing Basics (20 points)	Well written: clear and understandable. (18-20)	Writing is understandable. (13-17)	Does not follow proper structure, includes errors, or poorly written. (12 or below)

Note that *problem set grades do not directly depend on getting a 'correct' answer*, though arriving at an incorrect answer may be a symptom of insufficient engagement. Bonus points may be awarded for other insights.

Submitting academic work that uses others' ideas, words, research, or images without proper attribution and documentation is a violation of Brown's academic code.

Over 12 class sessions, students will spend 2.5 hours per week in class (30 hours total). Homework and reading will take approximately 15 hours per week (180 hours total). In addition, students should expect to spend at least 20 hours on the final project. Total: 230 hours.

Regrade Policy

Requests for reconsideration of grades are not encouraged, and will be accepted only in writing, with a clear statement of what has been misgraded, within one week of receiving the graded assignment. Please submit your full assignment so grading on all questions can be reconsidered.

Office Hours

Office hours are a great learning opportunity. Please come to my and the TA's office hours with questions on the material covered in class, comments on the course, or if you want to talk about anything in economics. Please do not use either my or the TA's office hours to talk about grades.

Class Recording and Distribution of Materials

I would like to record our discussion because some students may have poor internet connections, or have health issues. This means that we will record all classes, and make them available upon request to students that are enrolled but cannot be present. If you have questions or concerns about this protocol, please contact me so that we can talk through those to also ensure your full participation in this course. Lectures and other course materials are copyrighted. Students are prohibited from reproducing, making copies, publicly displaying, selling, or otherwise distributing the recordings or transcripts of the materials. The only exception is that students with disabilities may have the right to record for their private use if that method is determined to be a reasonable accommodation by Student Accessibility Services. Disregard of the University's copyright policy and federal copyright law is a Student Code of Conduct violation.

Technology

This course will use the following technological platforms: Zoom, Google Docs, Canvas, and Piazza. I am committed to ensuring access to online course resources by students. If you have any concerns or questions about access or the privacy of any of these platforms, please reach out to me. The IT Service Center (https://it.brown.edu/get-help) provides many IT Services including remote assistance, phones, tickets, and chat. Please also see the Online and Hybrid Learning Student Guide.

Course Expenses

If your Brown undergraduate financial aid package includes the Book/Course Material Support Pilot Program (BCMS), concerns or questions about the cost of books and course materials for this or any other Brown course (including RISD courses via cross-registration) can be addressed to bcms@brown.edu. For all other concerns related to non-tuition course-related expenses, whether or not your Brown undergraduate financial aid package includes BCMS, please visit the

Academic Emergency Fund in E-GAP (within the umbrella of "E-Gap Funds" in UFunds) to determine options for financing these costs, while ensuring your privacy.

Accommodations

Brown University is committed to full inclusion of all students. Please inform me early in the term if you may require accommodations or modification of any of course procedures. You may speak with me after class, during office hours, or by appointment. If you need accommodations around online learning or in classroom accommodations, please be sure to reach out to Student Accessibility Services (SAS) for their assistance (seas@brown.edu, 401-863-9588). Students in need of short-term academic advice or support can contact one of the academic deans in the College.

Class Schedule (tentative)

Class	Date	Topic	Reference	Assignment Due (by **3pm, day before class**)
1	9/15	Introduction and Clustering	HTF 14.3	
2	9/22	Visualization		PS 1 – Clustering
3	9/29	Trees	HTF 9.2	PS 2 – Visualization
4	10/6	Measurement		PS 3 – Trees **and** PS 0 – Programming
5	10/13	Fit	HTF 7	PS 4 – Measurement and Fit
6	10/20	Regularization	HTF 3	PS 5 – Generalizing Fit
7	10/27	Prediction and Causality	PGJ 3 and 4	PS 6 – Regularization and Loss
	11/3	*No class – Election Day*		
8	11/10	Structural Models		PS 7 – Causality and Optimal Pricing
9	11/17	Structure and Strategic Interaction		PS 8 – Optimal Pricing Competition*
10	11/24	Project Presentations / TBD		
11	12/1	Topics: Manipulation and Welfare		
	12/11	*No class*		**Final Project**

*: problem set will require additional interaction outside of class

Resources

If you'd like a refresher, a Python tutorial is available here:
https://www.codecademy.com/learn/python

Exploring Data

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, Conn.: Graphics Pr.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn: Graphics Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd edition). Cheshire, Conn: Graphics Pr.

Hadley Wickham. 'ggplot2.' http://link.springer.com.revproxy.brown.edu/book/10.1007/978-0-387-98141-3

'Awesome Interactive Journalism.' https://github.com/wbkd/awesome-interactive-journalism

Implementing Tufte graphs in R: http://motioninsocial.com/tufte/

Machine Learning

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Commun. ACM*, *55*(10), 78–87. http://doi.org/10.1145/2347736.2347755

Prediction

Berk, R., & Bleich, J. (2013). Forecasts of Violence to Inform Sentencing Decisions. *Journal of Quantitative Criminology*, *30*(1), 79–96. http://doi.org/10.1007/s10940-013-9195-0

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, *105*(5), 491–495. http://doi.org/10.1257/aer.p20151023

Perry, W. L., McInnis, B., Price, C. C., Smith, S., & Hollywood, J. S. (2013). *Predictive Policing*. RAND. Retrieved from http://www.rand.org/pubs/research_reports/RR233.html

Applications

Kang, J. S., Kuznetsova, P., Choi, Y., & Luca, M. (2013). Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. Retrieved from http://www.hbs.edu/faculty/Pages/item.aspx?num=45649

Gilchrist DS, Sands EG. (2015). Something to Talk About: Social Spillovers in Movie Consumption. Journal of Political Economy.

Digital Exhaust

Björkegren, D., & Grissen, D. (2015). Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment. *Working Paper*.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. http://doi.org/10.1126/science.aac4420

Zheng, Y.-T., Yan, S., Zha, Z.-J., Li, Y., Zhou, X., Chua, T.-S., & Jain, R. (2013). GPSView: A Scenic Driving Route Planner. *ACM Trans. Multimedia Comput. Commun. Appl.*, *9*(1), 3:1–3:18. http://doi.org/10.1145/2422956.2422959

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. http://doi.org/10.1126/science.1248506

Comprehensibility

Cowen, T. (2013). *Average Is Over: Powering America Beyond the Age of the Great Stagnation*. New York, New York: Dutton.

Kleinberg, J., & Mullainathan, S. (2015). We Built Them, But We Don't Understand Them. *Edge*. Retrieved from http://edge.org/response-detail/26192

Data Set Sources

Datahub: https://datahub.io
UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/

City Open Data Census: http://us-city.census.okfn.org/
Police Open Data Census: https://codeforamerica.github.io/PoliceOpenDataCensus/

Kiva: http://build.kiva.org/
Yelp: http://www.yelp.com/dataset_challenge
Wikipedia Clickstream: https://datahub.io/dataset/wikipedia-clickstream/resource/be85cc68-d1e6-4134-804a-fd36b94dbb82

Million Song Dataset: http://labrosa.ee.columbia.edu/millionsong/

Version 13 September 2020