Brown University, Economics 1660, Spring 2016

Big Data

Instructor: **Daniel Björkegren**
danbjork@brown.edu
Robinson Hall 303D
Office Hours: Th 3-3.55pm
or by appointment

TA: **Simon Freyaldenhoven**
simon_freyaldenhoven@brown.edu
Office Hours: M 11am-12pm
W 9-10am
Robinson Hall Basement

The spread of information technology has lead 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 004, 015, 017, or 019). Knowledge of econometrics and programming is assumed.

Class Meetings are Tuesdays from 4.00-6.30pm in Barus and Holley Room 158.

Mobile phones, laptops, and electronic tablets should be turned off and may not be used in class. (I may announce special working sessions where laptops may be used.)

Class readings will consist of research papers, which will be made available on the course Canvas website, supplemented by textbook readings from

- *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.

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 midnight 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. 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%

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.

Accommodations

If you need academic accommodations because of a documented disability, please let me know during office hours. You may also speak with Student and Employee Accessibility Services at 401-863-9588 to discuss the process for requesting accommodations.

Class Schedule (tentative)

Class	Date	Topic	Reference HTF	Assignment Due (by **midnight, day before class**)
1	2/2	Introduction and Clustering	14.3	
2	2/9	Visualization		PS 1 – Clustering
3	2/16	Trees	9.2	PS 2 – Visualization
	2/23	*No class – Long Weekend*		
4	3/1	Measurement		PS 3 – Trees **and** PS 0 – Programming
5	3/8	Bias / variance tradeoff	7	PS 4 – Feature Generation and Fit I
6	3/15	Regularization	3	PS 5 – Fit II
7	3/22	Prediction, causality, and A/B testing		PS 6 – Regularization
	3/29	*No class – Spring Break*		
8	4/5	Structural Models: Identification		PS 7 – Causality and Optimal Pricing
9	4/12	Structural Models: Strategic Interaction		PS 8 – Optimal Pricing Competition*
10	4/19	Topics: Networks		
11	4/26	TBD		
	5/10	*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

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. Something to Talk About: Social Spillovers in Movie Consumption. Journal of Political Economy. Forthcoming.

Digital Exhaust

Björkegren, D., & Grissen, D. (2015). Behavior Revealed in Mobile Phone Usage Predicts Loan 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 16 June 2016