ECON 5253: Data Science for Economists (Spring 2021)

	Tyler Ransom
Email	ransom@ou.edu
Office	322 CCD1
Office Hours	Tu 10:00-11:00am, Th 12:30-1:30pm
GitHub	tyleransom

- **Meeting day/time**: T,Th 3:15-4:30pm, Farzaneh Hall, Room 150
- Office hours also available by appointment
- This course takes inspiration and extensively borrows materials from similar courses taught by Jason DeBacker (U of South Carolina, Rick Evans (Rice U), and Grant McDermott (U of Oregon). Thanks to them for providing a framework for using GitHub as a class collaboration tool and for insights into teaching programming skills.

Course description

Data science is a rapidly developing field that combines the recent Big Data revolution with ever-developing statistical algorithms to inform business and policy decisions. Nearly every company you've heard of uses data science to optimize its services: Netflix uses it to recommend new programs to its viewers, Amazon uses it to determine how much it should charge for its Prime services. This class will provide students with an overview of the data science workflow, from collecting raw data to drawing a set of insights from which a decision maker can make informed decisions. Along the way we will broadly cover a variety of advances in data collection, data storage, visualization, machine learning and econometrics topics, as well as teaching and reinforcing good programming practices. The primary goal of this course is to provide you, the student, with a set of skills that will allow you to compete for a data science job.

Course Objectives and Learning Outcomes

By the end of the course, students should be able to do the following:

1. Explain the data science workflow from start to finish
2. Be able to collect data from online sources via APIs or scraping

3. Describe similarities and differences between econometrics and machine learning
4. Explain what data science is, and how Big Data differs from other types of data
5. Demonstrate good programming practices by writing code that can allow for easy collaboration with others
6. Understand the differences between prediction and causality, and the cases in which each is useful

In this course students, through lecture and application, will learn about: * Good programming practices, including how to write code collaboratively with others * Software to increase research productivity including: * LaTeX/Markdown * git * Software to collect & clean data, and estimate statistical models: * R * Julia * Python * Software to manage big data sets: * SQL * RDDs (Resilient Distributed Datasets) — Spark, Hadoop * How to access and utilize cluster computing resources * SSH (Secure Shell) * SFTP (Secure File Transfer Protocol) * SLURM (Simple Linux Utility for Resource Management) * Methods to gather and handle data including: * Costs and benefits of different data structures * Using APIs * Web scraping * Best practices for cleaning and visualizing data * Computational methods to: * Optimize and find roots of functions * Perform Monte Carlo simulations * Run computations in parallel using multiple processors (time permitting) * Basics for modeling different types of data * Machine learning basics: * Supervised vs. unsupervised learning * The five "tribes" of machine learning: how they are interconnected, and how they differ * Machine learning vs. econometrics: prediction vs. causality * Evaluating model performance * Using economic models to inform policy decisions * Computing structural models

Grades

Grades will be based on the categories listed below with the corresponding weights.

Component	Percent
Class Participation	10%
Problem Sets	35%
Exam & Quizzes	20%
Final project	35%
Total points	**100%**

Final grades will be assigned according to the standard cutoffs (90%+ for an A,

80%-89.99% for a B, etc.).

- **Participation:**
  - An important part of learning is face-to-face interaction. Thus, some of your grade will depend on attendance and active participation in class meetings.
- **Problem sets:** will be assigned approximately weekly throughout the semester.
  - You must write and submit your own computer code, although I encourage you to collaborate with your fellow students. I **DO NOT** want to see a bunch of copies of identical code. I **DO** want to see each of you learning how to code these problems so that you could do it on your own.
  - Problem set solutions, both written and code portions, will be turned in via a pull request from your private GitHub.com repository which is a fork of the class master repository on my account. (You will need to set up a GitHub account if you do not already have one.)
  - Written solutions must be submitted as PDF documents or Jupyter Notebooks.
  - Problem sets will be due on the day listed in the Daily Course Schedule section of this syllabus (see below) unless otherwise specified. Late problem sets will not receive any credit. Partially completed problem sets will receive partial credit.
- **Exam & Quizzes:**
  - We may periodically have in-class quizzes as low-stakes ways to get feedback
  - There will be a written final exam, but no midterm
- **Final Project:**
  - Collect data on and analyze a research question of your choosing, using methods taught in this course
  - Write up a ~10 page (12pt font, double spaced, excluding References, Figures, and Tables) summary of your findings, including discussion about what prior studies of the same topic have found, as well as citations to prior studies
  - Turn in the written summary report and a GitHub repository containing all materials required to reproduce the results
  - Summary report should be written in LaTeX or RMarkdown and turned in as a PDF (source code for the summary report should also be included in your GitHub repository)
  - An example of what the final product should look like is here, with LaTeX source code here and BibTeX source code here.
  - A detailed rubric for the final project is here

Communication

- I will always be available via email, and via Zoom during office hours. Zoom link for office hours will be posted on Canvas.
- Additionally, I have set up a Gitter community (see the badge at the top of this document) where I am hoping you can chat with each other about programming or other questions you have regarding the course. I will also be a participant in that community.

Daily Course Schedule

(Will be continuously updated throughout the semester)

Date	Day	Topic
Jan 26	T	What is data science / big data / why is it important? (Slides)
Jan 28	Th	Git, GitHub, computing environment, and Coding best practices (Notes) and Slides by Gran
Feb 2	T	Linux command line (Grant McDermott's slides), SSH, accessing OSCER (Notes); Git Tutor
Feb 4	Th	Overview of Data Scientists' tools (Notes)
Feb 9	T	Using data: data types, storage (Notes)
Feb 11	Th	Big Data: SQL (Notes) & RDDs (link); running jobs on the OSCER cluster
Feb 16	T	Sampling & storing Big Data (Notes)
Feb 18	Th	Web scraping/APIs to gather data (Notes; Grant McDermott's Lecture Notes; Ethics in Web
Feb 23	T	Web scraping/APIs to gather data (Notes); Grant McDermott's Lecture Notes
Feb 25	Th	Intro to Julia (Julia notes; Ivan Rudik's programming notes; Julia's "Learning Julia" page)
Mar 2	T	`ggplot2` (Basics; Kieran Healy's book)
Mar 4	Th	No class (Prof. Ransom seminar)
Mar 9	T	Getting to know your data: descriptive statistics, cleaning, tips, tricks, transformations, visua
Mar 11	Th	Modeling continuous and discrete variables (Notes) HTML slides); Simple R script
Mar 16	T	Using JuMP to optimize cool stuff [Jupyter Notebook; Julia Code] (in previous years: Linear
Mar 18	Th	Introduction to optimization (Notes)
Mar 23	T	Writing and optimizing functions in R, Python, and Julia (Notes)
Mar 25	Th	Writing and optimizing functions in R, Python, and Julia (Notes)
Mar 30	T	Debugging strategies and simulations (Notes)
Apr 1	Th	Intro to Machine Learning (Notes)
Apr 6	T	No class (University holiday)

Date	Day	Topic
Apr 8	Th	Supervised ML: Regularization, measuring model fit, tuning with cross-validation, the elastic
Apr 13	Tu	Supervised ML: The 5 Tribes of Machine Learning (Notes)
Apr 15	Th	Unsupervised ML: Clustering (Notes)
Apr 20	T	Unsupervised ML: Dimensionality reduction and reinforcement learning (Notes)
Apr 22	Th	Machine learning vs. econometrics (Notes)
Apr 27	T	Discrete choice modeling: static discrete choice (Slides)
Apr 29	Th	Final Project presentations (Rubric)
May 4	T	Final Project presentations (Rubric)
May 6	Th	Final Project presentations (Rubric)
May 4	M	Final Exam (in class, 1:30-3:30pm)

Helpful Links

- QuantEcon
- Notes on Machine Learning & Artificial Intelligence by Chris Albon
- R data wrangling cheatsheet
- R tidyverse
- Julia vs. Python for Data Science
- Machine Learning "Mind Map"
- JP Morgan massive overview of Big Data & Machine Learning
- Why it's becoming increasingly more difficult to learn to program

Books

- The Master Algorithm (Amazon link)
- Julia for Data Science (Amazon link)
- R for Data Science (Free PDF)
- Data Science at the Command Line (Free eBook)

University Policies

Religious Observance

It is the policy of the University to excuse the absences of students that result from religious observances and to reschedule examinations and additional required classwork that may fall on religious holidays, without penalty.

Reasonable Accommodation Policy

If a student requires an accommodation based on disability, the student should meet with me in my office during the first week of the semester. Student responsibility primarily rests with informing faculty at the beginning of the semester and in providing authorized documentation through designated administrative channels. The Disability Resource Center is located in the University Community Center at 730 College Avenue (405-325-3852).

Academic Integrity:

I do not tolerate academic misconduct, and neither does the University of Oklahoma. I will not hesitate to fail students who do not fully comply with the University's academic misconduct policy. If you find yourself contemplating cheating, plagiarism, or other forms of academic misconduct, please come see me first. Help is available if you are struggling. I want everyone in the class to try their best and to do their own work. Please be advised that I reserve the right to utilize anti-plagiarism resources such as TurnItIn when grading assignments.

Title IX Resources and Reporting Requirement

For any concerns regarding gender-based discrimination, sexual harassment, sexual assault, dating/domestic violence, or stalking, the University offers a variety of resources. To learn more or to report an incident, please contact the Sexual Misconduct Office at (405) 325-2215 (8 to 5, M-F) or smo@ou.edu. Incidents can also be reported confidentially to OU Advocates at (405) 615-0013 (phones are answered 24 hours a day, 7 days a week). Also, please be advised that a professor/GA/TA is required to report instances of sexual harassment, sexual assault, or discrimination to the Sexual Misconduct Office. Inquiries regarding non-discrimination policies may be directed to: Bobby J. Mason, University Equal Opportunity Officer and Title IX Coordinator at (405) 325-3546 or bjm@ou.edu. For more information, visit http://www.ou.edu/eoo.html.

Adjustments for Pregnancy/Childbirth Related Issues

Should you need modifications or adjustments to your course requirements because of documented pregnancy-related or childbirth-related issues, please contact your professor or the Disability Resource Center at (405) 325-3852 as soon as possible. Also, see http://www.ou.edu/eoo/faqs/pregnancy-faqs.html for answers to commonly asked questions.

Reasonable Accommodations for Students with Disabilities

If a student requires an accommodation based on disability, the student should meet with me in my office during the first week of the semester. Student responsibility primarily rests with informing faculty at the beginning of the semester and in providing authorized documentation through designated administrative channels. The Disability Resource Center is located in Goddard Hall (405-325-3852).