ECON368/DCS368: Big Data in Economics

Kyle Coombs he/him/his

Fall 2023

E-mail: kcoombs@bates.edu
Web: kylecoombs.com
Office Hours: W 9-12am (Zoom or in-person)
Class Hours: T/Th 9:30-10:50am
Office: ???
Class Room: ???

Course Website: TBD

OH Link: https://calendar.app.google/XF36Ujpg9NcJbSD58

Note

This syllabus contains a rough outline of the course and may change in the future. If you have any questions, you should check with me.

Course Description

Economics is at the forefront of developing statistical methods for analyzing data collected from uncontrolled sources. Since econometrics addresses challenges in estimation such as sample selection bias and treatment effects identification, the discipline is well-suited for the analysis of large and unsystematically collected datasets. This course introduces statistical (machine) learning methods, which have been developed for analyzing such datasets but which have only recently been implemented in economic research. We will cover a variety of topics including data collection, data management, data description, causal inference, and data visualization. The course also explores how econometrics and statistical learning methods cross-fertilize and can be used to advance knowledge in the numerous domains where large volumes of data are rapidly accumulating. We will also cover the ethics of data collection and analysis. The course will be taught in *R*.

Required Materials

Course notes, assignments, extra readings, recordings, and all other materials are available on the GitHub Classroom page. *The notes are adapted from Grant McDermott's course at the University of Oregon, Tyler Ransom's course at the University of Oklahoma, and Raj Chetty's course at Harvard University.*

Software requirements

All of the software requirements for this course are open-source and/or free. Please aim to have everything installed by the start of our first lecture. I will be available for installation troubleshooting during the first week of the semester. If you want a detailed tutorial on how to achieve a perfect working setup, I can think of no finer guide than Jenny Bryan *et al.*'s http://happygitwithr.com/ (see esp. sections 4 – 15).

R and RStudio

We will mainly be using the statistical programming language ***R*** (download here). Please make sure that you install the **RStudio IDE** too (download here).

Git and GitHub Classroom and GitHub CoPilot

We will also make extensive use of the **Git** version control system (follow the OS-specific installation instructions here). Once you have installed Git, please create an account on **GitHub** (here) and register for an education discount to get unlimited private repos (here).[1] Now is probably a good time to tell you that I am going to run the course through GitHub Classroom. You will receive an email invitation to the course repo with instructions in due time, but suffice it to say that this is how we'll submit assignments, provide feedback, receive grades, etc.

You will also need to signup for a **GitHub CoPilot** using the instructions here https://docs.github.com/en/copilot/quickstart. GitHub CoPilot is an AI tool that will help you write code. It is not perfect, but it is very useful. It is free for students, faculty, or maintainers of open source projects. It helped me write this syllabus.

Visual Studio Code

We will largely be using Visual Studio Code, or VSCode, for coding with *R*. VSCode is free and open-source, and is available for Windows, Mac, and Linux. You can download it at https://code.visualstudio.com/download. Once you have installed VSCode, you will need to install a variety of extensions. We will cover installations during the problem set (or as they become necessary), but here is a list:

1. The *R* extension by REditorSupport – https://code.visualstudio.com/docs/languages/r
2. *GitHub Copilot* by GitHub – https://marketplace.visualstudio.com/items?itemName=GitHub.copilot
3. *LaTeX Workshop* by James Yu – https://marketplace.visualstudio.com/items?itemName=James-Yu.latex-workshop

Recommended but not required extensions:

1. *ChatGPT - Genie AI* by Genie AI – https://marketplace.visualstudio.com/items?itemName=genieai.chatgpt-vscode

[1] GitHub recently announced unlimited free private repos for everyone. However, you are limited to three collaborators per private repo, so the education discount still makes sense.

Other

You are ready to start this course once you have installed R, RStudio, and Git (as well as created an account on GitHub). The last thing I want you to do for now is make sure that your system is configured to handle some additional packages that we will be using down the line. This varies by operating system:

- **Linux:** You should be good to go.
- **Mac:** Install the Homebrew package manager. I also recommend that you make sure your C++ toolchain is configured/open. Don't worry, it's simpler than it sounds. Just download the macOS Rtools installer and follow the instructions.
- **Windows:** Install Rtools. While its not essential, I also recommend that you install the Chocolatey package manager for Windows. Furthermore, please install the Windows Subsystem for Linux (WSL) and the Ubuntu distribution. Instructions here.

I will provide instructions for any further software requirements as the need arises; i.e. when we get to the relevant lecture. On that note, the lectures have all been posted ahead of time on the course website. Each lecture lists all the *R* packages and external libraries (if relevant) required for a particular class. I'll try to remind you, but my expectation is that you will look at these requirements and ensure that you have them installed *before* we start class.

Textbook and other readings

There's no set textbook for this course. I'll draw on readings from select *free* sources as needed listed below. You don't *need* to buy any of these (excellent) books to complete the course. But I can eagerly recommend leafing through at least one or two of them. Each of these books is freely available online if you can't afford a hard copy:

- "**Causal Inference: The Mixtape**" (Scott Cunningham)
- "**The Effect**" (Nick Huntington-Klein)
- "**Data Visualization: A practical introduction**" (Kieran Healy)
- "***R* for Data Science**" (Garrett Grolemund and Hadley Wickham)[2]
- "**Advanced *R***" (Hadley Wickham)
- "**Geocomputation with *R***" (Robin Lovelace, Jakub Nowosad and Jannes Muenchow)
- "**Spatial Data Science**" (Edzer Pebesma and Roger Bivand)
- "**An Introduction to Statistical Learning**" (Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani)
- Etc.

[2]FWIW, Jake VanderPlas's "**Python Data Science Handbook**" is excellent option for anyone looking for a Python equivalent.

Taking a step back, one of the goals of this course is to make you aware of the incredible array of instruction material that is freely available online. I also want to encourage you to be entrepreneurial. In that spirit, many of the lectures will follow a tutorial on someone's blog tutorial, or involve reproducing an existing study with open source tools. Each lecture will come with a set of recommended readings, which I expect you to at least look over before class.

Prerequisites

Prerequisites: ECON 255 and ECON 260 or ECON 270 The course assumes background in econometrics and statistics.

Teaching Assistant

There is no teaching assistant for this course.

Course Objectives

After this course is done, you should know how to:

1. Organize big data projects to be replicable and reproducible
2. Demonstrate good programming practices by writing code that can allow for easy collaboration with and replication by others
3. Collect and clean big or novel datasets using APIs, web scraping, and other methods
4. Make key insights into economic opportunity, racial discrimination, and economic mobility using big data
5. Understand the differences between prediction, causality, and description, and when to apply each
6. Explain what data science is, and how Big Data differs from other types of data

Grading Policy

The course will have a biweekly written problem sets (50%), short student presentations (10%) and a final project (40%). The final project will be a short research paper on a topic of your choice. The project will be graded based on a written report.

- **50%** of your grade will be determined by problem sets
- **10%** of your grade will be determined by short class presentations.
- **40%** of your grade will be determined by a final project
  - **5%** of your grade will be determined by a project proposal due TK
  - **5%** of your grade will be determined by a literature review due TK

-
- **5%** of your grade will be determined by a data description due TK
- **10%** of your grade will be determined by your code
- **15%** of your grade will be determined by a written summary of results

Problem Sets

All problem sets will be posted on GitHub Classroom and will be due in exactly two weeks. Problem sets should be turned in on GitHub Classroom. Working in groups on problem sets is not forbidden, but every student has to submit individual solutions in his/her own words.

Short Class Presentations

Almost every lecture will begin with a short student presentation. These should last 5-10 minutes and cover a prescribed topic (either assigned or approved ahead of time).

Tuesday presentations will involve presenting a technical skill or analytic method that will be used during the week. A successful presentation will introduce the skill or method, explain how it works, and provide an example of how it can be used.

Thursday presentations will involve presenting a key paper for the lecture that applies the technical skill or analytic method. A successful presentation will introduce the paper, explain the methodology, and summarize the key findings.

Final Project

The final project will be a research proposal on a topic of your choice, which uses methods taught in this course. The project will be graded based on a 5-10 page written report and your code. The report should include a literature review, a description of the data, a description of the methods, and a description of the expected results. The report should be written in LaTeXor RMarkdown. All code and other materials required[3] to reproduce results should be submitted in a GitHub repository.

During the semester, there will be three assignments related to the final project. The first will be a project proposal, which will be due 9/21. The second will be a review of the relevant literature, which will be due 10/17. The third will be a description of your data, which will be due 11/17. These are to help you stay on track for the final project, which is due on 12/11.

[3]Large datafiles do not need to be uploaded. Code to download or instructions to access work.

Course Policies

During Class

I understand that the electronic recording of notes will be important for class and so computers will be allowed in class. Please refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course. Eating and drinking are allowed in class, but please refrain from it affecting the course. Try not to eat your lunch in class as the classes are typically active.

Attendance Policy

For complete attendance and excused absence policies, please see **https://www.bates.edu/dof/course-attendance-policy-guideline-for-absences/**. Attendance is expected in all lectures. Valid excuses for absence will be accepted before class. In extenuating circumstances, valid excuses with proof will be accepted after class. For every class missed the participation grade will be dropped 1 point.

Policies on Incomplete Grades and Late Assignments

If an extended deadline is not authorized by the instructor or department, an unfinished incomplete grade will automatically change to an F after either (a) the end of the next regular semester in which the student is enrolled (not including summer sessions), or (b) the end of 12 months if the student is not enrolled, whichever is shorter. Incompletes that change to F will count as an attempted course on transcripts. The burden of fulfilling an incomplete grade is the responsibility of the student. The college policy on incomplete grades is located at TK.k

Late assignments will be accepted for no penalty if a valid excuse is communicated to the instructor before the deadline. After the deadline, assignments will be accepted for a 50% deduction to the score up to 2 days after the deadline. After this any assignments handed in will be given 0.

Academic Integrity and Honesty

Students are required to comply with the Bates policy on academic integrity in the Code of Student Conduct at https://www.bates.edu/student-conduct-community-standards/student-conduct/code-of-student-conduct/. Don't cheat. Don't be that person. Yes, you. You know exactly what I'm talking about. See https://www.bates.edu/student-conduct-community-standards/student-conduct/academic-integrity-policy/ for a detailed explanation of academic integrity.

Accommodations by Zoom

I prefer that all of you attend lecture in person, but I understand that there are sometimes unavoidable conflicts. As such, the course will have an option to tune in via Zoom for those with an excused absence related to health, family, or other unavoidable conflicts/emergencies. If you have a reason you need to attend a lecture via Zoom, please get in touch to explain the situation. If you do not get in touch and attend a lecture via Zoom without approval, I will consider it an absence. Approval can be given after the fact, but I prefer to know about hybrid attendance ahead of time. Several of you have been in touch about this option already and do not need to seek further approval.

Accommodations for Disabilities

Reasonable accommodations will be made for students with verifiable disabilities. In order to take advantage of available accommodations, students must register with the Office of Accessible Education and Student Support (AESS) in Ladd Library G35. For more information on Bates' policy on working with students with disabilities, please see the AESS webpage on Requesting Services (https://www.bates.edu/accessible-education-student-support/requesting-services/how-to-register-for-accommodations/).

Non-Discrimination Policy Bates College provides equality of opportunity in education and employment for all students and employees. Accordingly, Bates College affirms its commitment to maintain a work environment for all employees and an academic environment for all students that is free from all forms of discrimination.

Discrimination based on race, color, religion, creed, sex, national origin, age, disability, veteran status, or sexual orientation is a violation of state and federal law and/or Bates College policy and will not be tolerated. Harassment

of any person (either in the form of quid pro quo or creation of a hostile environment) based on race, color, religion, creed, sex, national origin, age, disability, veteran status, or sexual orientation also is a violation of state and federal law and/or Bates College policy and will not be tolerated. Retaliation against any person who complains about discrimination is also prohibited. Bates's policies and regulations covering discrimination, harassment, and retaliation may be accessed at https://www.bates.edu/here-to-help/policies/equal-opportunity-policy/. Any person who feels that he or she has been the subject of prohibited discrimination, harassment, or retaliation should contact the Director of Title IX & Civil Rights Compliance and Title IX Coordinator, Gwen Lexow, at titleix@bates.edu or https://www.bates.edu/here-to-help/make-a-report/.

Accommodations for Families

If you are a parent or guardian of a child and you are unable to attend class and care for that child for class one day, please be in touch in case you need further accommodations. You are invited to attend the lecture via Zoom or watch it asynchronously if that will make it easier to not miss course material.

Tentative schedule and weekly learning goals

The schedule is tentative and subject to change. Each week, I will cover a specific of topic. On Tuesday, we will cover relevant data science skills. On Thursday, we will apply those skills to a specific application.

Week 1, 09/05 - 09/07: Introduction to Big Data

- **Skills:** Installation of *R*, *VSCode*, etc.
- **Application:** Opportunity Atlas basics
- ***Problem Set 0 due Friday at Midnight***

Week 2, 09/12 - 09/14: Coding workflow, staying organized, and version control

- **Skills:** Functions, folder structure, Git(Hub), minimally reproducible examples, Docker
- **Application:** Hidden Decisions of Researchers, Data Colada
- ***Problem Set 1 due Friday at Midnight***

Week 3, 09/19 - 09/21: Gathering Data, Ethics, and Privacy

- **Skills:** APIs, scraping, hashing, differential privacy
- **Application:** Nowcasting Gentrification using Yelp Data, EJMR
- ***Project Proposal due 9/22 at Midnight***

Week 4, 09/26 - 09/28: Spatial Analysis

- **Skills:** Map projections, shapefiles, *sf*
- **Application:** Neighborhoods and Mobility
- ***Problem Set 2 due Friday at Midnight***

Week 5, 10/03 - 10/05: Regression review & Causal Inference

- **Skills:** OLS, IV, Potential Outcomes
- **Application:** Causal Effects of Neighborhoods

Week 6, 10/10 - 10/12: Regression Discontinuity Design

- **Skills:** Kernels, RDD, McCrary Test, fuzzy RDD
- **Application:** Class sizes and test scores
- ***Problem Set 3 due Friday at Midnight***

Fall Recess, 10/17 - 10/19: Databases on Tuesday, then rest!

- **Skills:** SQL
- ***Literature Review due 10/17 at midnight***[4]

Week 7, 10/24 - 10/26: Panel data and two-way fixed effects

- **Skills:** Frisch-Waugh-Lovell Theorem, Event Studies, *fixest*
- **Application:** Moving to Opportunity
- ***Problem Set 4 due Friday at Midnight***

Week 8, 10/31 - 11/02: Field Experiments

- **Skills:** Audit studies, correspondence studies
- **Application:** Resume studies and Racial Discrimination, Creating Moves to Opportunity

Week 9, 11/07 - 11/09: Functions & Parallel programming

- **Skills:** Functions, parallel programming, cluster computing, and cloud computing
- **Application:** Judicial decisions
- ***Problem Set 5 due Friday at Midnight***

Week 10, 11/14 - 11/16: Machine Learning

- **Skills:** Regression penalization methods, Causal Forests
- **Application:** Summer Jobs and At-Risk Youth
- ***Data Description due 11/17 at Midnight***

Thanksgiving Recess, 11/21 - 11/23: Gobble, gobble!

Week 11, 11/28 - 11/30: Text analysis

- **Skills:** Regular expressions, WordClouds, sentiment analysis
- **Application:** Twitter, inherited biases?

Week 12, 12/05 - 12/07: Text analysis

- **Skills:** topics modeling, LLMs, AI
- **Application:** EJMR, climate and attitude
- ***Problem Set 6 due Friday at Midnight***

Final Project due 12/11 at Midnight

[4] Can change as needed.

What is missing?

- Data types, data storage
- Command line interface
- Optimization, vectorization
- Cluster computing
- Prediction and Machine Learning
- Cross-validation
- Supervised vs. unsupervised ML
- Bayesian ML

Want to add: section on protection of identity, privacy, and ethics. Hashing, etc. Maybe drop Field Experiments? Drop databases?