

# ECON368/DCS368: Big Data in Economics

Kyle Coombs he/him/his

Fall 2023

E-mail: [kcoombs@bates.edu](mailto:kcoombs@bates.edu)

Office Hours: T, 3-4pm, W 9-10am (Zoom or in-person)

Office: PGill 276

Web: [kylecoombs.com](http://kylecoombs.com)

Class Hours: T/Th 9:30-10:50am

Class Room: Carnegie 339

Course Website: <https://github.com/ECON368-fall2023-big-data-and-economics>

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*.

## Course Objectives

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

1. Organize empirical projects that are replicable, reproducible, and collaborative using good programming practices
2. Collect and clean big or novel datasets using APIs, web scraping, and other methods
3. Use Big Data to generate key insights about economic opportunity, inequality, and racial discrimination

4. Understand the differences between prediction, causality, and description, and when to apply each
5. Explain what data science is, and how Big Data differs from other types of data

## 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 the software requirements for this course are open-source and/or free. Please aim to have **R** and **Rstudio** installed by the start of our first lecture. Other installation will be a part of Problem Set 0. 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, 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](#)).<sup>1</sup> 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 sign up 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:

---

<sup>1</sup>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.

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

## 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/tools that we will be using down the line.

### Recommended but not required:

1. *GitHub Copilot* by GitHub – <https://marketplace.visualstudio.com/items?itemName=GitHub.copilot>
2. *ChatGPT - Genie AI* by Genie AI – <https://marketplace.visualstudio.com/items?itemName=genieai.chatgpt-vscode>
3. *Anaconda* or *PIP* – largely used for Python installations, there are a few quality of life packages for R that are distributed via *Anaconda* or *PIP*.
4. *Radian* – Radian allows you to use RStudio similar to how you would RStudio. You will be able to run code directly into a terminal with Ctrl+Enter, but also have access to GitHub CoPilot coding assistance.

### Operating system-specific recommendations:

- **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 it's 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, each week's lectures will be posted by the preceding Sunday 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)<sup>2</sup>
- **“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.

Taking a step back, one of the goals of this course (and most Data Science courses) 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.

## Grading Policy

The course will have seven coding problem sets (50%), weekly short student presentations (10%) and a final project (40%). The final project will be a short research paper on a topic of your choice.

- 50% of your grade will be determined by the top 5 out of 7 problem sets
- 10% of your grade will be determined by 1-2 short class presentations
- 40% of your grade will be determined by a final project due 12/11
  - 5% of your grade will be determined by a project proposal due 9/22

---

<sup>2</sup>FWIW, Jake VanderPlas’s **“Python Data Science Handbook”** is excellent option for anyone looking for a Python equivalent.

- 5% of your grade will be determined by a literature review due 10/17
- 5% of your grade will be determined by a data description & analysis due 11/17
- 5% of your grade will be determined by a peer review due 12/1
- 10% of your grade will be determined by your code
- 10% 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 roughly 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). You can sign up at <https://github.com/ECON368-fall2023-big-data-and-economics/presentations>.

Presentations will cover a technical skill (e.g. minimally reproducible examples), analytic method (e.g. regression discontinuity design), or a key reading (e.g. an applied research paper). A successful presentation of a skill or method will introduce it, explain how it works, and provide a use case. A successful presentation of a key reading will introduce the research question, explain the methodology, and summarize the key findings.

## Final Project

The final project will be a research paper 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, analysis, and discussion of results. The report should be written in  $\LaTeX$  or RMarkdown. All code and other materials required<sup>3</sup> to reproduce results should be submitted in a GitHub repository.

During the semester, there will be four 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 data description, which will be due 11/17. The fourth will be a peer review due 12/1. These are to help you stay on track for the final project, which is due on 12/11.

For ideas on datasets that are “big,” consider <https://www.kaggle.com/datasets> or this guide put together by Christine Murray: <https://libguides.bates.edu/ECON368>.

## Peer review

After you submit your data description & analysis, I will randomly assign each of you a peer’s draft up to this point. You will have two weeks to provide constructive feedback on the project in the form of a short paper and a meeting during class.

Examples of constructive feedback:

---

<sup>3</sup>Large data files do not need to be uploaded. Code to download or instructions to access will suffice.

- “I believe your empirical design does not account for variation by county and time. Consider adding fixed effects.”
- “This project aligns closely with this other literature. Check out read Author X (YYYY).”
- “Your regression discontinuity design may suffer from manipulating at the cutoff. Consider implementing a McCrary test.”
- “It would be helpful to present your results in terms of expected earnings increases. That would make your main outcome easier to interpret.”

Not constructive feedback:

- “I noticed three typos.”
- “This idea is bad, and you should feel bad.”
- “This is not actually economics.”
- “Your mother was a hamster and your father smells of elderberry.”

## Course Policies

### During Class

We will be doing active coding projects during class, so please bring your personal laptops. 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 breakfast/lunch in class as the classes are typically active.

### Artificial Intelligence

I encourage each of you to make use of artificial intelligence-driven digital assistants, like ChatGPT and Github CoPilot. These tools are not a substitute for your own ingenuity, but instead a complement as they are incredibly useful for tasks like coding or proofreading. Please note during assignments whether and where you used ChatGPT, as you would cite your (human) sources.

### 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.

### Policies on Incomplete Grades and Late Assignments

**Grace Period Days:** Everyone will receive three “grace period” days to turn in work after the due date. Late assignments will be accepted for no penalty if turned in within the “grace period.” You can use these whenever you wish, but once you use them, they are gone.

After the “grace period,” the instructor, department, or college must authorize an extension. If not authorized, 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.

**End of course:** If an extension beyond the “grace period” is not authorized by the instructor, department, or college, 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.

### 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](mailto: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. Bolded readings are “key” readings.

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

- **Skills:** Installation of *R*, *VSCode*, etc.
- **Application:** Opportunity Atlas basics
  - Readings: **Chetty et al. (2018)**, Chetty et al. (2020), Einav and Levin (2014)
- *Problem Set 0 due Sunday 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
  - Readings: Shapiro et al. (2014), McDermott (2022)
- **Application:** Hidden Decisions of Researchers, Data Colada
  - Readings: **Huntington-Klein et al. (2021)**, Simonsohn (2021), Simonsohn (2022)
- *Problem Set 1 due Friday at Midnight*

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

- **Skills:** APIs, scraping, hashing, differential privacy
  - Readings: **Chetty and Friedman (2019)**, Abowd and Schmutte (2019)
- **Application:** Nowcasting Gentrification using Yelp Data
- *Project Proposal due 9/22 at Midnight*
  - Readings: **Glaeser et al. (2018)**, Glaeser et al. (2017)

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

- **Skills:** Map projections, shapefiles, *sf*
  - Reading:
- **Application:** Neighborhoods and Mobility
  - Readings: **Chetty et al. (2018)**,
- *Problem Set 2 due Friday at Midnight*

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

- **Skills:** OLS, IV, Potential Outcomes
  - Readings: Huntington-Klein (2021) Chapter 13, Cunningham (2021) Chapter 4
- **Application:** Causal Effects of Neighborhoods
  - Readings: **Chetty et al. (2016)**

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

- **Skills:** Kernels, RDD, McCrary Test, fuzzy RDD
  - Readings: Cunningham (2021) Chapter 6, Huntington-Klein (2021) Chapter 20
- **Application:** Class sizes and test scores
  - Readings: **Angrist and Lavy (1999)**, Chetty et al. (2011)
- *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*

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

- **Skills:** Frisch-Waugh-Lovell Theorem, Event Studies, *fixest*
  - Huntington-Klein (2021) Chapters 16-18, Cunningham (2021) Chapters 8, 9
- **Application:** Moving to Opportunity
  - Reading: **Chetty and Hendren (2018)**
- *Problem Set 4 due Friday at Midnight*

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

- **Skills:** Audit studies, correspondence studies
  - Readings: TK
- **Application:** Resume studies and Racial Discrimination, Creating Moves to Opportunity
  - Readings: Bertrand and Mullainathan (2004)

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

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

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

- **Skills:** Regression penalization methods, Causal Forests
  - Readings: Athey and Imbens (2019)
- **Application:** Summer Jobs and At-Risk Youth
  - Readings: , Naik et al. (2014)
- ***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
  - Readings: Gentzkow et al. (2019)
- **Application:** Twitter, inherited bias
  - Readings:

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

- **Skills:** Topics modeling, LLMs, AI
  - Reading: Ash and Hansen (2023)
- **Application:** EJMR, climate and attitude
  - Reading: TK
- ***Problem Set 6 due Friday at Midnight***

**Final Project due 12/11 at Midnight**

### **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

## References

- Abowd, John M. and Ian M. Schmutte**, “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices,” *American Economic Review*, 1 2019, 109, 171–202.
- Angrist, Joshua D. and Victor Lavy**, “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement,” *The Quarterly Journal of Economics*, 5 1999, 114, 533–575.
- Ash, Elliott and Stephen Hansen**, “Text Algorithms in Economics,” <https://doi.org/10.1146/annurev-economics-082222-074352>, 7 2023, 15.
- Athey, Susan and Guido W. Imbens**, “Machine Learning Methods That Economists Should Know About,” *Annual Review of Economics*, 2019, 11.
- Bertrand, Marianne and Sendhil Mullainathan**, “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination,” *American Economic Review*, 8 2004, 94, 991–1013.
- Chetty, Raj and John N. Friedman**, “A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples,” *AEA Papers and Proceedings*, 5 2019, 109, 414–20.
- **and Nathaniel Hendren**, “The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects,” *The Quarterly Journal of Economics*, 8 2018, 133, 1107–1162.
- **, John Friedman, and Nathaniel Hendren**, “The Opportunity Atlas Mapping the Childhood Roots of Social Mobility,” 2020.
- **, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter**, “The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility,” 10 2018.
- **, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan**, “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star,” *The Quarterly Journal of Economics*, 11 2011, 126, 1593–1660.
- **, Nathaniel Hendren, and Lawrence F. Katz**, “The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment,” *American Economic Review*, 4 2016, 106, 855–902.
- Cunningham, Scott**, *Causal Inference: The Mixtape*, Vol. Yale Press. 2021.
- Einav, Liran and Jonathan Levin**, “Economics in the age of big data,” *Science*, 2014, 346, 1–7.
- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy**, “Text as Data,” *Source: Journal of Economic Literature*, 2019, 57, 535–574.
- Glaeser, Edward L., Hyunjin Kim, and Michael Luca**, “Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity,” 11 2017.
- **, — , and —**, “Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change,” *AEA Papers and Proceedings*, 5 2018, 108, 77–82.

**Huntington-Klein, Nick**, *The Effect: An introduction to research and causality*, Vol. Chapman and Hall 2021.

—, **Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky**, “The influence of hidden researcher decisions in applied microeconomics,” *Economic Inquiry*, 7 2021, 59, 944–960.

**McDermott, Grant**, “Docker Lecture,” 2022.

**Naik, Nikhil, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo**, “Streetscore - Predicting the Perceived Safety of One Million Streetscapes,” 2014.

**Shapiro, Matthew M Gentzkow Jesse, Chicago Booth, Matthew Gentzkow, and Jesse M Shapiro**, “Code and Data for the Social Sciences: A Practitioner’s Guide,” 2014.

**Simonsohn, Uri**, “[95] Groundhog: Addressing The Threat That R Poses To Reproducible Research - Data Colada,” 2021.

—, “[100] Groundhog 2.0: Further addressing the threat R poses to reproducible research - Data Colada,” 2022.