ECON368/DCS368: Big Data and Economics

Kyle Coombs he/him/his

Fall 2023

E-mail: kcoombs@bates.edu
Web: **kylecoombs.com**
Office Hours: T, 3-4pm, W 9-10am (Zoom or in-person)
Class Hours: T/Th 9:30-10:50am
Office: PGill 276
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).[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 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.

Windows only: Windows Subsystem for Linux

Windows users will need some form of command line interface tool. My recommendation is Windows Subsystem for Linux (`https://learn.microsoft.com/en-us/windows/wsl/install`). This allows you to run Linux within your Windows machine, so you can more easily implement tips you find online for software installation, etc.

[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.

LaTeX software

TeX Live:
A LaTeX software distribution that is compatible with the LaTeX Workshop extension in Visual Studio Code. Installation instructions can be found here: https://www.tug.org/texlive/. Use the "easy install" option for your operating system.

Overleaf:
I also recommend that you create an Overleaf account. Overleaf is a useful tool for learning LaTex, which you will use to write your final. You can create an account here https://www.overleaf.com/register. Note, you can sync your Overleaf account with your GitHub account using instructions here https://www.overleaf.com/learn/how-to/Git_integration. This is a premium feature at present, so I do not require it. You can also sync it with Dropbox, which is similarly a premium feature https://www.overleaf.com/learn/how-to/Dropbox_Synchronization. (Student plans cost $89 per year.)

Recommended but not required:

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.

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 a few quality of life packages for R that are distributed via *Anaconda* or *PIP*. https://docs.anaconda.com/free/anaconda/install/index.html
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. https://github.com/randy3k/radian

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. *LaTeX Workshop* by James Yu – https://marketplace.visualstudio.com/items?itemName=James-Yu.latex-workshop
3. *GitHub Classroom* by GitHub – https://marketplace.visualstudio.com/items?itemName=GitHub.classroom&ssr=false#overview

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 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, 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)[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.

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.

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

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
  - **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 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 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:

- "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."

[3] Large data files do not need to be uploaded. Code to download or instructions to access will suffice.

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 short-term), 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.

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. 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:** 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]**, tin, Wickham, 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], api
- **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: **McDermott and Rubin [2023a]** crs, Lovelace et al. [2019]
- **Application:** Neighborhoods and Mobility
  - Readings: **Chetty et al. [2018]**
- ***Problem Set 2 due Friday at Midnight***

Week 5, 10/03 - 10/05: Functions & Parallel programming

- **Skills:** Functions
  - Readings: McDermott and Rubin [2023b], McDermott and Rubin [2023c] Wickham [2023], tid
- **Skill:** Parallel Programming
  - Readings: **McDermott and Rubin [2023d]**, Eddelbuettel [2020], McDermott and Rubin [2023d]
- ***Problem Set 5 due Friday at Midnight***

Week 6, 10/10 - 10/12: Regression review & Causal Inference

- **Skills:** OLS, IV, Potential Outcomes
  - Readings: Huntington-Klein [2021] Chapter 13, Cunningham [2021] Chapter 4
- **Application:** Returns to Education and College Proximity
  - Readings: **Card [1993]**

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:** Causal Effects of Neighborhoods
  - Reading: **Chetty and Hendren [2018]**, **Bergman et al. [2019]**, **Chetty et al. [2016]**
- ***Problem Set 3 due Friday at Midnight***

Week 8, 10/31 - 11/02: Regression Discontinuity Design

- **Skills:** RDD, McCrary Test, fuzzy RDD
  - Readings: Cunningham [2021] Chapter 6, Huntington-Klein [2021] Chapter 20
- **Applications:** College wage premia, Peru's Mining *Mita*, class sizes
  - Readings: **Dell [2010]**, **Zimmerman [2014]**, **Angrist and Lavy [1999]**, Chetty et al. [2023]
- ***Problem Set 4 due Friday at Midnight***

Week 9, 11/07 - 11/09: Machine Learning I

- **Skills:** Decision Trees
  - Readings: **Athey and Imbens [2019]**, **Varian [2014]**, Mullainathan and Spiess [2017], Kleinberg et al. [2015]
- **Application:** Bias and Judicial Decisions
  - Readings: **Kleinberg et al. [2018]**, Bertrand and Mullainathan [2004], Simonsohn
- ***Data Description due 11/17 at Midnight***

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

- **Skills:** Regression penalization methods, Causal Forests
  - Readings: **Athey and Imbens [2019]**, Varian [2014], Mullainathan and Spiess [2017], Kleinberg et al. [2015]
- **Application:** Summer Jobs and At-Risk Youth
  - Readings: **Davis and Heller [2017]**, 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 I

- **Skills:** Regular expressions, WordClouds, sentiment analysis
  - Readings: **Gentzkow et al. [2019]**
- **Application:** Google Flu Trends, Racial Animus and Elections
  - Reading: **Lazer et al. [2014]**, **Stephens-Davidowitz [2014]**, Ginsberg et al. [2009]

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

- **Skills:** Topics modeling, LLMs, AI
  - Reading: **Ash and Hansen [2023]**
- **Application:** EJMR, Temperature and Twitter
  - Readings: **Wu [2018]**, Moore et al. [2019]
- ***Problem Set 6 due Friday at Midnight***

Final Project due 12/11 at Midnight

What is missing?

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

References

Raj Chetty, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. The opportunity atlas: Mapping the childhood roots of social mobility. 10 2018. doi: 10.3386/W25147. URL https://www.nber.org/papers/w25147.

Raj Chetty, John Friedman, and Nathaniel Hendren. The opportunity atlas mapping the childhood roots of social mobility, 2020.

Liran Einav and Jonathan Levin. Economics in the age of big data. *Science*, 346:1–7, 2014. URL https://www.science.org.

Matthew M Gentzkow Jesse Shapiro, Chicago Booth, Matthew Gentzkow, and Jesse M Shapiro. Code and data for the social sciences: A practitioner's guide, 2014. URL http://faculty.chicagobooth.edu/matthew.gentzkow/research/CodeAndData.pdf,.

Grant McDermott. Docker lecture, 2022. URL https://raw.githack.com/uo-ec607/lectures/master/13-docker/13-docker.html#1.

Nick Huntington-Klein, 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*, 59:944–960, 7 2021. ISSN 1465-7295. doi: 10.1111/ECIN.12992. URL https://onlinelibrary.wiley.com/doi/full/10.1111/ecin.12992https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12992https://onlinelibrary.wiley.com/doi/10.1111/ecin.12992.

tinyverse. URL https://www.tinyverse.org/.

Hadley Wickham. Journal of statistical software tidy data. URL http://www.jstatsoft.org/.

Uri Simonsohn. [95] groundhog: Addressing the threat that r poses to reproducible research - data colada, 2021. URL https://datacolada.org/95.

Uri Simonsohn. [100] groundhog 2.0: Further addressing the threat r poses to reproducible research - data colada, 2022. URL http://datacolada.org/100.

Raj Chetty and John N. Friedman. A practical method to reduce privacy loss when disclosing statistics based on small samples. *AEA Papers and Proceedings*, 109:414–20, 5 2019. ISSN 2574-0768. doi: 10.1257/PANDP.20191109.

John M. Abowd and Ian M. Schmutte. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109:171–202, 1 2019. ISSN 0002-8282. doi: 10.1257/AER.20170627.

An introduction to apis | zapier guides. URL https://zapier.com/resources/guides/apis.

Edward L. Glaeser, Hyunjin Kim, and Michael Luca. Nowcasting gentrification: Using yelp data to quantify neighborhood change. *AEA Papers and Proceedings*, 108:77–82, 5 2018. ISSN 2574-0768. doi: 10.1257/PANDP.20181034. URL https://doi.org/10.1257/pandp.20181034.

Edward L. Glaeser, Hyunjin Kim, and Michael Luca. Nowcasting the local economy: Using yelp data to measure economic activity. 11 2017. doi: 10.3386/W24010. URL https://www.nber.org/papers/w24010.

Grant McDermott and Ed Rubin. Spatial analysis, 2023a. URL https://grantmcdermott.com/ds4e/spatial-analysis.html.

Overview of coordinate reference systems (crs) in r. URL http://spatialreference.org/.

Robin Lovelace, Jakub Nowosad, and Jannes Muenchow. *Chapter 2 Geographic data in R | Geocomputation with R*. 2019. ISBN 9780203730058. URL https://r.geocompx.org/spatial-class.html.

Grant McDermott and Ed Rubin. Functions: Introductory concepts, 2023b. URL https://grantmcdermott.com/ds4e/funcs-intro.html.

Grant McDermott and Ed Rubin. Functions: Advanced concepts, 2023c. URL https://grantmcdermott.com/ds4e/funcs-adv.html.

Hadley Wickham. *Metaprogramming*. 2023. URL https://adv-r.hadley.nz/metaprogramming.html.

Tidy eval helpers — tidyeval • ggplot2. URL https://ggplot2.tidyverse.org/reference/tidyeval.html.

Grant McDermott and Ed Rubin. Parallel programming, 2023d. URL https://grantmcdermott.com/ds4e/parallel.html.

Dirk Eddelbuettel. Parallel computing with r: A brief review. 2020. URL https://arxiv.org/abs/1912.11144.

Nick Huntington-Klein. *The Effect: An introduction to research and causality*, volume Chapman and Hall. 2021. URL https://theeffectbook.net/ch-StatisticalAdjustment.html.

Scott Cunningham. *Causal Inference: The Mixtape*, volume Yale Press. 2021. URL https://mixtape.scunning.com/04-potential_outcomes.

David Card. Using geographic variation in college proximity to estimate the return to schooling. 10 1993. doi: 10.3386/W4483. URL https://www.nber.org/papers/w4483.

Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, 133:1107–1162, 8 2018. ISSN 0033-5533. doi: 10.1093/QJE/QJY007. URL https://dx.doi.org/10.1093/qje/qjy007.

Peter Bergman, Raj Chetty, Deluca Stefanie, Nathaniel Hendren, Lawrence F. Katz, and Christopher Palmer. Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice, 8 2019. URL https://www.nber.org/papers/w26164.

Raj Chetty, 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*, 106:855–902, 4 2016. ISSN 0002-8282. doi: 10.1257/AER.20150572.

Melissa Dell. The persistent effects of peru's mining mita. *Econometrica*, 78:1863–1903, 2010. doi: 10.3982/ECTA8121. URL https://scholar.harvard.edu/files/dell/files/ecta8121_0.pdf.

Seth D. Zimmerman. The returns to college admission for academically marginal students. *Journal of Labor Economics*, 32:711–754, 10 2014. ISSN 0734306X. doi: 10.1086/676661. URL https://www.journals.uchicago.edu/doi/full/10.1086/676661.

Joshua D. Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114:533–575, 5 1999. ISSN 0033-5533. doi: 10.1162/003355399556061. URL https://dx.doi.org/10.1162/003355399556061.

Raj Chetty, David J. Deming, and John N. Friedman. Diversifying society's leaders? the causal effects of admission to highly selective private colleges. 7 2023. URL https://www.nber.org/papers/w31492.

Susan Athey and Guido W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 2019.

Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28:3–28, 2014. ISSN 0895-3309. doi: 10.1257/JEP.28.2.3. URL http://dx.doi.org/10.1257/jep.28.2.3.

Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31:87–106, 2017. doi: 10.1257/jep.31.2.87. URL https://doi.org/10.1257/jep.31.2.87.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105:491–95, 5 2015. ISSN 0002-8282. doi: 10.1257/AER.P20151023.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133:237–293, 2 2018. ISSN 0033-5533. doi: 10.1093/QJE/QJX032. URL https://dx-doi-org.lprx.bates.edu/10.1093/qje/qjx032.

Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94:991–1013, 8 2004. ISSN 0002-8282. doi: 10.1257/0002828042002561.

Uri Simonsohn. [51] greg vs. jamal: Why didn't bertrand and mullainathan (2004) replicate? - data colada. URL https://datacolada.org/51.

Jonathan M.V. Davis and Sara B. Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107:546–50, 5 2017. ISSN 0002-8282. doi: 10.1257/AER.P20171000.

Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo. Streetscore - predicting the perceived safety of one million streetscapes, 2014.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Source: Journal of Economic Literature*, 57:535–574, 2019. doi: 10.2307/26787457.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343:1203–1205, 3 2014. ISSN 10959203. doi: 10.1126/SCIENCE.1248506/SUPPL_FILE/1248506.LAZER.SM.REVISION1.PDF. URL https://www.science.org/doi/10.1126/science.1248506.

Seth Stephens-Davidowitz. The cost of racial animus on a black candidate: Evidence using google search data. *Journal of Public Economics*, 118:26–40, 10 2014. ISSN 0047-2727. doi: 10.1016/J.JPUBECO.2014.04.010.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature 2009 457:7232*, 457:1012–1014, 2 2009. ISSN 1476-4687. doi: 10.1038/nature07634. URL https://www.nature.com/articles/nature07634.

Elliott Ash and Stephen Hansen. Text algorithms in economics. *https://doi.org/10.1146/annurev-economics-082222-074352*, 15, 7 2023. ISSN 1941-1383. doi: 10.1146/ANNUREV-ECONOMICS-082222-074352. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-economics-082222-074352.

Alice H. Wu. Gendered language on the economics job market rumors forum. *AEA Papers and Proceedings*, 108:175–79, 5 2018. ISSN 2574-0768. doi: 10.1257/PANDP.20181101.

Frances C. Moore, Nick Obradovich, Flavio Lehner, and Patrick Baylis. Rapidly declining remarkability of temperature anomalies may obscure public perception of climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 116:4905–4910, 2019. ISSN 10916490. doi: 10.1073/PNAS.1816541116/-/DCSUPPLEMENTAL.