Syllabus: Data Science for Social Impact (DSSI)

Course information

Course numbers: DATA SCI 154, EARTHSYS 153, ECON 163, MS&E 134, POLISCI 154, SOC 127, COMM 140X, PUBLPOL 155
Class times: Tuesday and Thursday 3:00 to 4:20PM in Bishop Auditorium. Friday classes are at 10:30, 12:30, and 1:30. Students may choose which Friday class they prefer to attend each week.
Canvas site: https://canvas.stanford.edu/courses/209797
Prerequisites:

- Experience programming in R or python, or willingness to learn very quickly on your own.
- A basic statistics or data science course, such as DATASCI 112, ECON 102 or 108, CS 129, EARTHSYS 140, HUMBIO 88, POLISCI 150A, or STATS 60 or 101, SOC 180B, or MS&E 125.

Optional fun class gatherings:

- April 3, 2025: 6:30-7:45PM at Professor Allcott's home (730 Frenchman's Road, Stanford)
- April 29, 2025: 6:30-8PM at Professor Allcott's home
- May 29, 2025: 6:30-8:30 PM at Professor Allcott's home

Teaching team

Professors:

- Hunt Allcott. Website: allcott.stanford.edu.
  - My schedule is very variable, so I do not have standing weekly office hours. Right after class is often a good time to talk!
- Mallory Nobles. Website: https://datasciencemajor.stanford.edu/mallory-nobles
  - Fridays 3:00-4:00, CoDA 138. Happy to also meet at other times over Zoom or in person - please email to arrange!

Please call us "Professor Allcott" and "Professor Nobles." After you graduate, please call us by our first names.

Teaching Assistants:

- Nick Scott-Hearn. Email: nicksh@stanford.edu

- Patricio Ortiz. Email: patricio.ortiz@stanford.edu
- Jason Canaday. Email: jcanaday@stanford.edu

Office hours:
- Fridays, 3-5 PM, CoDa, Room B06
- Mondays, 3-5 PM, Sequoia Hall, Fishbowl Room (220)

Course overview

You have some experience coding in R or Python. You've taken a class or two in basic stats or data science. But what's next? How can you use data science skills to make the world a better place?

If you're asking those questions, then Data Science for Social Impact is for you.

In this class, you'll work in four areas where data are being used to make the world better: health care, education, detecting discrimination, and clean energy technologies. You'll work with data from hospitals, schools, police departments, and electric utilities. You'll apply causal inference, prediction, and optimization techniques to help businesses, governments, and other organizations make better decisions. You'll see the challenges that arise when analyzing real data – for example, when some data are missing, or when the randomized experiment gets implemented wrong. You'll get ideas for an impactful and meaningful senior thesis, summer internship, and future career.

Concretely, you'll have weekly problem sets involving data analysis in Python. You'll learn and apply techniques like fixed effects regression, difference-in-differences, instrumental variables, regularized regression, random forests, causal forests, and optimization. Class sessions will feature active learning, discussions, and small-group case studies. You should only enroll if you expect to attend regularly and complete the problem sets on time.

Learning goals

In this class, students will:
1. practice implementing important data science methodologies such as OLS, instrumental variables, difference-in-differences, regularized regression, random forests, and linear optimization;
2. work through practical data analysis challenges such as missing data, missing labels, feature leakage, and imperfect compliance and treatment-control imbalance in experiments;

3. develop initial subject-matter expertise in health care, education, criminal justice, and clean energy technologies; and
4. build new connections and friendships with a community of people interested in using data science to help make the world a better place.

Course materials

Poll Everywhere: Sign up for an account at polleverywhere.com.

Computer: Bring your laptop computer to class every day to answer polls and work with data. You'll need to be connected to WiFi and have Poll Everywhere working when class starts.

It is easy to distract ourselves with messages, news, and other things on our smartphones and computers. During each class time, you should close all apps and windows other than those that we are using for class. Your smartphone should be put away during class except if you are using it for polls.

Data analysis software: We will use python, via Google Colab. You will need a Google account to use Colab. See here for Colab instructions.

Textbook: There is no required textbook. All readings will be available online or through Canvas. We will regularly use the following references:

- Allcott, Hunt (2025). Empirical Environmental Economics Lecture Notes.
- Cunningham, Scott (2023). Causal Inference: The Mixtape.

Grading and deadlines

This is a five-unit course, which means that you can expect up to 15 hours of work per week, including class time. Grading will not be on a curve; if everyone does well, everyone will get a good grade. The table below summarizes the work and contribution to the overall course grade. All deadlines listed below are at 11:59PM Pacific Time. There will be no final exam.

Category	% of grade	Notes
Class participation	10%	● Polls will be part of your grade starting the second week of class. ● There will typically be multiple polls per day. Often there is no right

Category	% of grade	Notes
(measured through PollEverywhere and Google Forms responses)		answer; in these cases you get credit for any response. If there is a right answer, you'll get more credit for the right response. ● Sometimes the software doesn't work well, or you have to miss class for an interview or competition or because you are sick. Thus, we will drop the two class days on which you have the lowest scores. ● We will typically have polls at the very beginning of class.
Problem sets	90%	● There will be up to nine problem sets, due each Monday (at 11:59PM Pacific Time). ● A fundamental (and fun) part of the class is to discuss the problem sets on Tuesdays in class. Thus, (i) no credit can be awarded for submissions received after class begins, and (ii) assignment extensions are not possible and would constitute what Stanford calls a "fundamental alteration of the course." ● Late submissions will receive partial credit that declines linearly from 100% if submitted on time to 0% if submitted at the beginning of class on Tuesday, i.e. 100% – (hours late) / 15. ● The one exception is that any students who join the class late can submit any problem sets they missed by the Monday after they join. ● When computing grades, we will give only 25% of the usual weight to the one problem set on which you have the lowest grade.

Course policies

1. **Honor Code**: The teaching team and students follow the Stanford Honor Code.
2. **Common syllabus**: We follow the policies in Stanford's economics common syllabus (including the policy on correcting any errors in grading).
3. **Attendance**: We hope to make the class time useful and fun. We expect that everyone will attend class (including Fridays) except in exceptional circumstances, such as an interview, a competition, or because you are sick.
4. **Discussions:** Every day in class, we'll discuss the analyses you're working on. During these discussions, we'll randomly select students to share their thinking with the class. We do random selection instead of asking for volunteers because it helps engage everyone, not just a few people. We'll be nice when we call on you, and it's totally OK if you don't have the right answer.
5. **Emergencies:** The teaching team wants to do everything we can to accommodate diverse personal situations. The teaching team can extend assignment deadlines or modify course requirements in exceptional situations, such as the death of a

family member or a medical emergency. Please email the TAs if such a situation arises.
In any given quarter, many of us will have other personal events that are less severe than these two examples. It can be hard to assess the severity of these situations in a way that is fair to everyone, so we usually don't extend deadlines or otherwise modify course requirements. However, the teaching team stands ready to support you and help you catch up on the materials.

6. **Accessibility:** Stanford is committed to providing equal educational opportunities for disabled students. Disabled students are a valued and essential part of the Stanford community. We welcome you to our class.
If you experience disability, please register with the Office of Accessible Education (OAE). Professional staff will evaluate your needs, support appropriate and reasonable accommodations, and prepare an Academic Accommodation Letter for faculty. To get started, or to re-initiate services, please visit oae.stanford.edu.
If you already have an Academic Accommodation Letter, we invite you to share your letter with us. Academic Accommodation Letters should be shared at the earliest possible opportunity so we may partner with you and OAE to identify any barriers to access and inclusion that might be encountered in your experience of this course.
(You probably know that the above is standard language, but we agree wholeheartedly, and please let us know how we can help.)
7. **AI.** We encourage you to use AI tools ChatGPT and GitHub Copilot to help develop your code. However, remember that our success in an economy with AI depends on both (i) using AI well and (ii) having cognitive and contextual ability that AI doesn't have. (ii) requires learning and applying foundational concepts, even if ChatGPT could do your analyses for you.
8. **Group work:** We encourage you to do the problem sets in groups and discuss the final projects with others. However, all work you submit must be your own individual work.
9. **Questions and discussion board**: We'll use the Discussions page on Canvas. If you have questions about the problem sets, please post them there instead of emailing us. Many others may have the same questions, so your questions benefit everyone!

Schedule and readings

Every week has a different topic but the same structure: The weekly class cadence is:

- **Second half of Tuesday class**: overview of new week. Using lecture format (with some polls and think-pair-shares), the instructors give background on this week's application area and technical methodology.
- **Thursday class**: methodology and programming session. Students learn the methodology and apply it to this week's dataset in a session that weaves lecture, group discussion, and programming.
- **Friday classes**: the TAs lead hands-on sessions focused on implementing the problem set.
- **Monday night**: analysis writeup due at 11:59PM.
- **First half of Tuesday class**: problem set recap. The instructors lead a class discussion of the problem set, including both analysis decisions and real-world implications.

Weekly schedule

Week	Tuesday Date	Area	Methodology	Application
1	4-1-2025	Health	Regression	Effect of social media on mental health
2	4-8-2025	Health	IV, randomized encouragement	Oregon Health Insurance Experiment
3	4-15-2025	Health	Regularized regression	Predicting mortality
4	4-22-2025	Criminal justice	Diff-in-diff	Discrimination and Ban the Box
5	4-29-2025	Criminal justice	Algorithmic fairness	Predicting recidivism
6	5-6-2025	Education	Optimization	School choice
7	5-13-2025	Education		School start times
8	5-20-2025	Energy	Advanced diff-in-diff	Behavioral energy efficiency
9	5-27-2025	Energy	Fixed effects	Renewable energy and carbon neutrality
10	6-3-2025			[Class ends after June 3.]

Part I: Health

Week 1: How Does Social Media Use Affect Mental Health?

The goals for this week are:
1. practice working in Colab and implementing OLS regression;
2. be able to explain the Rubin causal model, correlation vs. causation, and the expected sign of omitted variable bias;
3. be able to explain mental health trends over the past few decades; and
4. begin to get to know your classmates and get excited about the course.

Readings:
- Allcott E3 lecture notes: Chapter 3: Linear Regression.
    - Optional: Sections 4.1-4.3: Rubin Causal Model
- Mixtape: Probability and Regression Review, Sections 2.13, 2.25, and 2.26.
- Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow (2020). "The Welfare Effects of Social Media." *American Economic Review*, Vol. 110, No. 3 (March), pages 629-676.
    - *Focus on the introduction and Section IV.C ("Effects on Subjective Well-Being"). Feel free to lightly skim the rest.*
    - Optional: listen to Audioloom podcast here
- Optional:
    - New York Times review of Jonathan Haidt's book, The Anxious Generation
    - Professor Allcott's Washington Post op-ed about social media and mental health.

Lecture slides:
- ClassOverview.pdf
- SocialMedia_1.pdf
- SocialMedia_2.pdf
- SocialMedia_Recap.pdf

Problem set:
- In this week's problem set, you'll estimate the effects of social media use on subjective well-being using data from a randomized experiment.
- Colab starter code is here, using data here.

Week 2: What Are the Effects of Health Insurance?
The goals for this week are:
1. practice implementing instrumental variables regressions;
2. be able to explain the math and substance of the assumptions under which instrumental variables deliver causal estimates;
3. be able to explain basic issues in the economics of health insurance; and
4. and get to know several new classmates that you haven't worked with yet.

Readings:

- Allcott E3 lecture notes: Chapter 11: Instrumental Variables
- Mixtape: Instrumental Variables, Sections 7.2 and 7.3
- Finkelstein, Amy, et al. (2012). "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics*, Vol. 127, No. 3 (August), pages 1057-1106.
- Optional background:
    - Online article on randomized encouragement experimental designs
    - MIT article on MIT professor Amy Finkelstein and the Oregon Health Insurance Experiment
    - Forbes article on Finkelstein's health insurance proposal with Stanford professor Liran Einav

Lecture slides:

- HealthInsurance_1.pdf
- HealthInsurance_2.pdf
- HealthInsurance_Recap.pdf

Problem set:

- In this week's problem set, you'll estimate the effects of health insurance on health and other outcomes using data from the Oregon Health Insurance Experiment.
- Colab starter code is here, using data here.

Week 3: How Well Can We Predict Mortality?

The goals for this week are:

1. practice implementing regularized regression models;
2. be able to explain key concepts including prediction vs. parameter estimation, overfitting, and regularization;
3. determine which variables predict mortality and the extent to which mortality is predictable with insurance data; and
4. continue to get to know new classmates.

Readings:

- James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor (2023). An Introduction to Statistical Learning.
    - Sections 2.2.1 and 2.2.2: Measuring the Quality of Fit, and The Bias-Variance Tradeoff

- Intro to Section 6: pages 229-230
- Section 6.2. *(We recommend skipping "Bayesian Interpretation of Ridge Regression and the Lasso.")*

- Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer (2018). "Predictive modeling of U.S. health care spending in late life." *Science*, Vol. 360, No. 6396, pages 1462–1465.
- Optional: Mullainathan, Sendhil, and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*, Vol. 31, No. 2, pages 87–106.
    - *Pages 87-98 are the most relevant for class.*

Reading questions:
1. Explain briefly what "overfitting" means.
2. Explain briefly how regularized regression addresses overfitting.
3. Explain briefly why the lasso yields sparse models, while the ridge does not.
4. Explain briefly how to find the optimal λ parameter.

Lecture slides:
- HealthPrediction_1.pdf
- HealthPrediction_2.pdf
- HealthPrediction_Recap.pdf

Problem set:
- In this week's problem set, you'll predict mortality using synthetic Medicare claims data.
- Colab starter code is here, using data here: New Link.

Part II: Criminal Justice

Week 4: Does "Ban the Box" Reduce Discrimination?

The goals for this week are:
1. implement a standard difference in differences regression;
2. be able to explain key diff-in-diff concepts such as parallel trends;
3. be able to explain and evaluate statistical tests for discrimination;
4. continue to get to know classmates.

Readings:
- Agan, Amanda, and Sonja Starr (2018). "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment." *Quarterly Journal of Economics*, Vol. 133, No. 1 (February): pages 191-235.
    - Read the intro, then skim Sections II-IV.C, stopping at the end of page 216.

- Allcott, Hunt (2025). Empirical Environmental Economics Lecture Notes. Chapter 9: Difference-in-Differences.
- Optional:
    - Lang, Kevin, and Ariella Kahn-Lang Spitzer (2020). "Race Discrimination: An Economic Perspective." *Journal of Economic Perspectives*, Vol. 34, No. 2 (Spring): pages 68-89.
    - Alexander, Michelle (2010). "The New Jim Crow."

Reading questions:
1. Briefly describe another example of difference-in-differences other than the ones mentioned above. Specifically, describe (i) the research question, (ii) the outcome Y, (iii) treatment T, and (iv) the required parallel trends assumption.
2. Briefly explain why the "parallel leads" test is a suggestive but not direct test of "parallel trends."

Lecture slides:
- BantheBox_1.pdf
- BantheBox_2.pdf
- BantheBox_Recap.pdf

Problem set:
- In this week's problem set, you'll estimate employment discrimination and how that changed after a Ban the Box policy took effect.
- Colab starter code is here, using data here.

Week 5: Can We Make Efficient and Fair Prediction Algorithms for Criminal Justice?
The goals for this week are:
1. get better at regularized regression;
2. explain alternative notions of algorithmic fairness;
3. quantitatively evaluate the fairness of different algorithms; and
4. continue to get to know classmates.

Readings:
- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016). "Machine Bias." *ProPublica*, May 23.
- Chouldechova, Alexandra (2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data*, Vol. 5, No 2.
    - *Read only through page 7, stopping at the beginning of the "Assessing Impact" section.*
- Optional:

- Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel (2016). "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *Washington Post*, October 17.
  - *If you cannot access washingtonpost.com, a pdf version (unfortunately without key graphics) is here.*
- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainthan, and Ashesh Rambachan (2018). "Algorithmic Fairness." *American Economic Review, Papers and Proceedings*. 108 (May): pages 22–27.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018). "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, Vol. 133, No. 1 (February): pages 237–293.
  - *The intro and Section II.A are most relevant.*

Reading questions:

1. Concisely state the following definitions of fairness discussed in Chouldechova (2017): calibration, error rate balance, and statistical parity.
2. Given these three definitions and/or others, precisely state the definition of algorithmic fairness that you think is most important.

Lecture slides:

- Fairness_1.pdf
- Fairness_2.pdf
- Fairness_Recap.pdf

Problem set:

- In this week's problem set, you'll predict rearrest among defendants released on bail, using data from New York.
- Colab starter code is here, using data here.

Part III: Education

Week 6: Education

The goals for this week are:

1. practice implementing a linear programming optimization model;
2. be able to explain key optimization concepts such as decisions, objectives and constraints in the context of linear programs;
3. explore how school choice algorithms impact outcomes and educational equity; and
4. continue to get to know classmates.

Readings:

- Allman, Maxwell, et al. "Designing school choice for diversity in the San Francisco Unified School District." *Proceedings of the 23rd ACM Conference on Economics and Computation. 2022.*
- "San Francisco is changing its school assignment system. This data shows why" San Francisco Chronicle
- Optional:
    - Reardon, Sean F., et al. "Is separate still unequal? New evidence on school segregation and racial academic achievement gaps." (2019).
    - Abdulkadiroğlu, Atila, and Tayfun Sönmez. "School choice: A mechanism design approach." *American economic review* 93.3 (2003): 729-747.
    - We can draw school zones to make classrooms less segregated. This is how well your district does. Vox
    - Monarrez, Tomás E. "School attendance boundaries and the segregation of public schools in the United States." *American Economic Journal: Applied Economics* 15.3 (2023): 210-237.

Reading questions:

1. Briefly describe the three policies (zones, zones + reserves, priorities) modeled in Allman, Maxwell, et al.
2. Briefly describe how the choice model in Allman, Maxwell et al. predicts students' school preferences.

Lecture slides:

- Lecture 1
- Lecture 2, class co-lab (not for HW)

Problem set:

- In this week's problem set, you will implement a slight reformulation of the school choice problem used by San Francisco's Unified School District.
- To complete this week's problem set, please answer the questions in this colab notebook.
    - Due to data sharing limitations, the format of this week's problem set is different from other weeks.
    - Please make a copy of the colab notebook, and complete your responses directly in your copy of the notebook. Please turn in a PDF of your completed notebook on Canvas.
- You will need several files to run the notebook: `opt_inputs.pickle`, `table2.csv`, `table3.csv`, and `assignment.py`

Week 7: What can we learn from Boston's experience in moving from problem to policy to practice when considering new bus routes and school start times?

The goals for this week are:

1. get better at formulating linear programming optimization models;
2. Explore strategies for moving from problem to policy to practice; and
3. continue to get to know classmates.

Readings:

- Optional:
  - Bertsimas, Dimitris, Arthur Delarue, and Sebastien Martin. "Optimizing schools' start time and bus routes." *Proceedings of the National Academy of Sciences* 116.13 (2019): 5943-5948

Lecture slides:

- Lecture 1
- Lecture 2

Problem set:

- There is no problem set this week. In exchange for a free week from problem sets, you are expected to fully participate in Thursday's class where we will do an exercise to explore moving from problem to policy to practice. **Thursday's class will be held in Lathrop 018.**

Part IV: Clean Energy

Week 8: How Effective is Behavioral Energy Efficiency?

The goals for this week are:

1. be able to explain and implement key panel data techniques including fixed effects and clustered standard errors;
2. implement diff-in-diff / event study designs in panel data, including evaluating parallel leads;
3. evaluate one type of energy efficiency program and be able to explain how this fits into the broader clean technology landscape; and
4. continue to get to know classmates.

Readings:

- Allcott E3 lecture notes: Chapter 7: Panel Data
- Fleming, Kevin Charles (2012). "Who's Saving Energy in Your Neighborhood?" *Pacific Standard* magazine (March 5).
- Allcott, Hunt, and Todd Rogers (2014). "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review*, Vol. 104, No. 10 (October), pages 3003-3037.
  - *Read the introduction, Section I, and Section III. Skim the rest as you like.*

- Optional: Allcott, Hunt, and Michael Greenstone (2012). "Is There an Energy Efficiency Gap?" *Journal of Economic Perspectives*, Vol. 26, No. 1 (Winter), pages 3-28.

Reading questions:

1. Describe a situation where $\hat{\beta}_1^{POLS}$ is biased and $\hat{\beta}_1^{FE}$ is unbiased. Specifically, describe (i) the research question, (ii) the outcome and explanatory variables, and (iii) the practical reasons for bias or lack thereof.
2. Describe a situation where $\hat{\beta}_1^{POLS}$ is unbiased and $\hat{\beta}_1^{FE}$ is biased. Specifically, describe (i) the research question, (ii) the outcome and explanatory variables, and (iii) the practical reasons for bias or lack thereof.
3. Why might household fixed effects not make much difference when using diff-in-diff to estimate the Opower treatment effect?

Lecture slides:

- EnergyEfficiency_1.pdf
- EnergyEfficiency_2.pdf
- EnergyEfficiency_Recap.pdf

Problem set:

- In this week's problem set, you'll estimate the effects of behavioral energy efficiency programs using data from a randomized experiment.
- Colab starter code is here, using household data here and energy use data here.
- We'll build this slide deck in class for the recap.

Week 9: How Do Data Centers and Renewables Affect CO2 Emissions?

The goals for this week are:

1. thoughtfully implement fixed effects estimators;
2. carefully articulate how to think about additionality and causal inference in a more nuanced setting;
3. quantitatively estimate short-run marginal emission rates from electricity generation; and
4. practice presentation skills.

Readings:

- Penrod, Emma (2024). "How can more companies use 24/7 clean energy? Hourly matching is 'not the hard part,' analysts say." *UtilityDive* (June 13).
- Dunkle Werner, Karl, and Arik Levinson (2025). "Greenhouse Gases Resulting from Grid-Connected Electricity Demand." Working paper (April).
    - *Read the introduction and lightly skim the rest*

- Novan, Kevin (2015). "Valuing the Wind: Renewable Energy Policies and Air Pollution Avoided." *American Economic Journal: Economic Policy*, Vol. 7, No. 3 (August), pages 291–326.
    - *Read the introduction and lightly skim the rest*
- As needed, review Allcott E3 lecture notes: Chapter 7: Panel Data

Reading questions:

1. Drawing on Dunkle Werner and Levinson (2025) and your expertise in thinking about causality, precisely describe a situation in which an investment in a data center plus renewable energy would be carbon-neutral, and another situation in which it would not.
2. Describe a regression you could run to estimate the effect of additional electricity demand on CO_2 emissions in Texas. What controls would you include to avoid omitted variable bias?

Lecture slides:

- GridEmissions_1.pdf
- GridEmissions_2.pdf
- GridEmissions_Recap.pdf

Problem set:

- In this week's problem set, you'll estimate the effects of data centers and renewable energy generation on CO_2 emissions using data from Texas.
- Colab starter code is here, using data here.

Additional resources

- Writing tutors from the Hume Center for Writing and Speaking, to sharpen your essay-writing skills
- Academic coaches from the Center for Teaching and Learning, to help you manage your time and work effectively
- Study halls, organized by the Center for Teaching and Learning, to work and learn in quiet companionship with other students
- Study Tips and Tools, from the Center for Teaching and Learning
- Undergraduate Advising Directors, Academic Advising
- Well-Being services, including well-being coaches, Vaden Health Center
- Subject Matter Tutoring, Center for Teaching and Learning
- Tutoring for Student Athletes, AARC
- The Flourish newsletter on health and wellness

- First Generation and/or Low-Income Student Success Center, including financial support