DSAN 5650

Causal Inference for Computational Social Science
Wednesdays 6:30-9pm, Online

Prof. Jeff Jacobs
jj1088@georgetown.edu

Summer 2025, Georgetown University

Welcome to DSAN 5650: Causal Inference for Computational Social Science at Georgetown University!

The course meets on **Wednesdays from 6:30-9pm online via Zoom**

Course Staff

- Prof. Jeff Jacobs, jj1088@georgetown.edu
  - Office hours (Click to schedule): Tuesdays, 3:30-6:30pm
- TA Courtney Green, crg123@georgetown.edu
  - Office hours by appointment
- TA Wendy Hu, lh1078@georgetown.edu
  - Office hours by appointment

Course Description

This course provides students with the opportunity to take the analytical skills, machine learning algorithms, and statistical methods learned throughout their first year in the program and explore how they can be employed towards carrying out rigorous, original research in the behavioral and social sciences. With a particular emphasis on tackling the additional challenges which arise when moving from associational to causal inference, particularly when only observational (as opposed to experimental) data is available, students will become proficient in cutting-edge causal Machine Learning techniques such as propensity score matching, synthetic controls, causal program evaluation, inverse social welfare function estimation from panel data, and Double-Debiased Machine Learning.

In-class examples will cover continuous, discrete-choice, and textual data from a wide swath of social and behavioral sciences: economics, political science, sociology, anthropology, quantitative history, and digital humanities. After gaining experience through in-class labs and homework assignments focused on reproducing key findings from recent journal articles in each of these disciplines, students will spend the final weeks of the course on a final project demonstrating their ability to develop, evaluate, and test the robustness of a causal hypothesis.

Prerequisites: DSAN 5000, DSAN 5100 (DSAN 5300 recommended but not required)

Course Overview

The fundamental building block for the course is the idea of a **Data-Generating Process (DGP)**. You may have encountered this concept in passing during other DSAN courses (for example, in DSAN 5100, a phrase like "Assume X is drawn i.i.d. from a Normal distribution with mean μ and variance σ^2" is a statement characterizing the DGP of a Random Variable X), but in this course we will "zoom in" on this concept rather than treating it like a black box or a footnote to e.g. a theorem like the Law of Large Numbers.

This deep dive into DGPs is necessary for us here, since our goal in the course is **to move from *associational* statements** like "an increase of X by one unit is associated with an increase of Y by β units" to ***causal* statements** like "increasing X by one unit *causes* Y to increase by β units". As you'll see in Week 1, the tools from probability theory and statistics that you learned in DSAN 5100—Random Variables, Cumulative Distribution Functions, Conditional Probability, and so on—are necessary but **not *sufficient*** to analyze data from a causal perspective.

For example, if we use our tools from DSAN 5000 and DSAN 5100 on some dataset to discover that:

- The probability that some event E_1 occurs is $\Pr(E_1) = 0.5$, and
- The probability that E_1 occurs **conditional on** another event E_0 occurring is $\Pr(E_1 \mid E_0) = 0.75$,

we unfortunately cannot infer from these two pieces of information that the occurrence of E_0 **causes** an increase in the likelihood of E_1 occurring.

This issue (that **conditional probabilities** could not be interpreted causally) at first represented a kind of dead end for scientists interested in employing probability theory to study causal relationships... In recent decades, however, researchers have built up what amounts to an additional "layer" of modeling tools which augment the existing machinery of probability theory to address causality head-on![1]

[1] Pearl (2000) represents a key work in this field of research, as it essentially brought together different pieces of causal models into one unified, rigorous framework.

For instance, a modeling approach called **"do-Calculus"**, that we will learn in this class, *extends* the core operations and definitions of probability theory to allow such an move to deriving causality! It does this by introducing a $\mathsf{do}(\cdot)$ operator that can be applied to Random Variables like X, with e.g. $\mathsf{do}(X = 5)$ representing the event wherein someone has **intervened** in a **Data-Generating Process** to **force** the value of X to be 5.

With this operator in hand (that is, used alongside an explicit model of a DGP satisfying a set of underlying axioms which are slightly more strict than the axioms of probability theory), it turns out that we *can* make causal inferences using a very similar pair of facts! If we know that:

- The probability that some event E_1 occurs is $\Pr(E_1) = 0.5$, and
- The probability that E_1 occurs **conditional on** the event $\mathsf{do}(E_0)$ occurring is $\Pr(E_1 \mid \mathsf{do}(E_0)) = 0.75$,

now we can actually draw the inference that the occurrence of E_0 **caused** an increase in the likelihood of E_1 occurring!

This stylized comparison (between what's possible using "core" probability theory and what's possible when we augment it with additional causal modeling tools) serves as our basic motivation for the course, so that from Week 2 onwards we build upon this foundation to reach the three learning goals described in the next section!

Main Textbooks / Resources

Unlike the case for topics like calculus or statistical learning, this field is too new (and exciting! with new methods being developed month-to-month) to have a single set of "established" textbooks. Thus, the main collection of resources (books, papers, and explanatory videos) we'll draw on for this class are available on the resources page. However, there are three "core" textbooks you can draw on which best align with the topics in this course:

- Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (Morgan and Winship 2015) [PDF]
  - The book which comes closest to being an all-encompassing, single textbook for the class. It brings together different "strands" of causal modeling research (since each field—economics, bioinformatics, sociology, etc.—tends to use its own notation and vocabulary), unifying them into a single approach. The only reason we can't use it as *the* main textbook is because it hasn't been updated since 2015, and most of the assignments in this class use computational tools from 2018 onwards!
- Angrist and Pischke, *Mastering 'Metrics: The Path from Cause to Effect* (Angrist and Pischke 2014) [PDF]