# Understanding Your Problem
# Unifying Data Science 2023

Nick Eubank

March 11, 2023

**Due March 28th, 8am**

Now that you are familiar with how to answer Causal Questions, the time has come for you to answer a Causal Question of your own!

As we've discussed at length in this class, however, we do data science to solve problems. So before specifying a causal question you wish to answer, you will need to specify a *problem* you wish to help address.

Once you've specified a problem you wish to address, the next step will be to target your efforts by specifying and answering at least three (and potentially more than three) Exploratory Questions. Your Exploratory Questions should be designed to motivate and focus your subsequent efforts to answer a Causal Question.

I hope you're able to have fun with this exercise. It is rare in school that we get to invest in answering precisely the questions we find exciting, and I hope you will see this as an opportunity to invest in learning about (and help address) a problem you care about personally. Moreover, as we've discussed before, this is a potential portfolio piece—one unique to your team—you can discuss in interviews and post on your personal website.

## Your Assignment

Your first assignment consists of two parts:

1. You must identify a problem, then pose *and answer* a set of Exploratory Questions that help you better understand the problem space.
2. Motivated by those results, you must propose a Causal Question you wish to answer.

### Part 1: Identifying and Understand a Problem

You must specify the **problem** your team wishes to address, along with at least three (but potentially more than three) Exploratory Questions that you wish to answer. These questions need not require complicated analyses to answer—I'm more than happy for the answers to these questions to entail simple summary statistics, and they may even be summary statistics you find that someone else has computed. But it is a *canonical* issue with data scientists that they are prone to jumping to complicated analyses or regressions without actually getting to know

their problem space, and this assignment is meant to give you practice in starting with a wider perspective.

We've discussed what Exploratory Questions may look like previously, but here are a few examples of problems and associated Exploratory Questions:

- *Problem:* Too many people are killed in car accidents.
  *Exploratory Questions:*
  - What share of car-related fatalities is due to car-pedestrian, single-car, or multiple-car accidents?
  - What share of car-related fatalities occurs on freeways as opposed to in cities?
  - What share of car-related fatalities involves a driver under the influence of drugs or alcohol?
- *Problem:* Many states are adding bureaucratic hurdles to getting social services, but the effect of these hurdles is unclear, both in terms of their effect on reducing fraud and on deterring entitled recipients from getting aid.
  *Exploratory Questions:*
  - What states have changed their rules around social service provision (helpful if we want to do a pre-post analysis or a difference-in-difference analysis)?
  - Are there a lot of people who are entitled to social services who don't receive them? (Do we know?)
  - What social service programs that have had bureaucratic hurdles imposed serve the largest populations?
- *Problem:* Police shootings involving people with mental health issues are much too common, and it's not clear the police are appropriately trained to deal with people dealing with mental health crises.
  *Exploratory Questions:*
  - What states have changed their rules around social service provision (helpful if we want to do a pre-post analysis or a difference-in-difference analysis)?
  - Are there a lot of people who are entitled to social services who don't receive them? (Do we know?)
  - What social service programs that have had bureaucratic hurdles imposed serve the largest populations?

Note that while the problems in these examples are all "big" problems—in the sense of being societally important questions—your problem need not be of this nature. Past teams have done projects trying to figure out how to optimally train in tennis (by looking at whether playing more tennis improves or hinders subsequent tournament performance), how to improve AirBnB host profits (by looking at whether "super host" status improves AirBnB host revenues above and beyond the effect of just having the features that make one eligible to be a super-host), and how to minimize cell-phone user churn.

With that said, be aware that you will need to find data about your problem, and that *is* often easier to do when wrestling with public problems; companies tend to be quite protective of their commercially-sensitive data.

**Deliverables:**

- A clear statement of the problem your team is interested in addressing.

- A list of Exploratory Questions you set out to answer and, for each, an explanation for why you think answering the question may be useful.
- Quantitative answers to your Exploratory Questions. Again, these will most likely be simple summary statistics, and while **at least one answer must have been generated from your own calculations**, you may also use data and statistics from secondary sources as answers to your other Exploratory Questions.

### Part 2: Causal Project Backwards Design

The second deliverable required for this first assignment is a plan for answering a Causal Question. For this portion of the assignment, please follow the template provided below.

To make this more concrete, a collection of past *Unifying Data Science* projects can be found on Sakai in the *Past UDS Projects* folder in *Resources*. These projects were not designed with the *exact* prompt you are receiving, so you will find some differences (I'm explicitly adding Exploratory Questions this year, for example). Nevertheless, they should give you a good sense of how students have approached this in the past.

**Deliverables:**

- A Backwards Design plan for answering a Causal Question motivated by the answers you generated to your Exploratory Questions.

**Backwards Design Template**

# 1 Topic:

*What is your project about? What problem are you seeking to solve, or in which domain do you think you can contribute meaningfully?*

# 2 Project Question

*What specific question are you seeking to answer with this project? For this project, this must be a* ***causal*** *question.*

# 3 How Will Answering This Problem Help Address Your Problem?

*Your answer to this question should be informed by your answers to your Exploratory Questions.*

# 4 Ideal Experiment

*If you were a god, what experiment would you run to answer your question? Define both your treatment variable and your outcome of interest.*

# 5 Pick a Study Context

*Where can you get data that (a) measures your outcome variable, and (b) includes variation in your treatment variable?*

# 6 Project Design

*Given the context you want to study (and data you can find), what design do you think would be feasible?*

# 7 Model Results

*One of the hardest parts of developing a good data science project is developing a question that is actually answerable. Perhaps the best way to figure out if your question is answerable is to see if you can imagine what an answer to your question would look like. Below, draw the graph, regression table, etc. that you would consider to be an answer to your question. Then draw it again, so you have a model result for if treatment has an effect and a model result for if your treatment does not have an effect. (If the answer to your question is continuous, not discrete (like: what is the effect of health insurance on life expectancy), draw it for high values (high inequality) and low values (low inequality)).*

**Result if your hypothesis is true** **Result if your hypothesis is false**

# 8 Final Variables Required

*Now that you've specified what an answer to your question looks like, what data do you need to generate that answer?*

*For each variable, define both the variable you need* ***and*** *the population for which you need the variables to be defined.*

*You don't have to be too specific ("I need variable 7 from the NHGIS 2019 census 1% sample release") – just define it in the most general terms that are still specific enough to meet your needs (e.g. I need income data for a nationally representative sample of US citizens from both before and after 2012).*

# 9 Data Sources

*Finally, given the variables you need for your analysis, what actual data sources do you think will have the data you need?*

*In specifying the datasets you need, if you list more than one* ***also*** *indicate how you think you can relate these datasets (i.e. if you're gonna merge them, what variables do you think those datasets will provide that will allow you merge them? There's no use saying "I'll merge this political survey with medical records of who has received bad care" if the political survey doesn't provide identifying information you can use to link survey respondents to medical records, even if you have both the survey and medical records!)*