# CS-GY 6053/CS-UY 3943 - Foundations of Data Science
# Final Project

**Brief Summary**
You will be working in groups of 3 members to complete a causal inference-based data analysis using concepts learned during the semester in this course. This project will be evaluated based on the following 3 components:

| Component | Due Date | Percent of Final Grade |
|---|---|---|
| Proposal | November 6th, 2024, 11:59 PM EST | 5% |
| Data Communication/ Screencast | December 4th, 2024, 11:59 PM EST | 10% |
| Write-up | December 4th, 2024, 11:59 PM EST | 25% |

**Core Requirements**
While the example problems considered in this course have come from a variety of domains and have utilized different models, a fundamental approach has been pursued each time:

1) Identify a question/estimand
2) Describe a scientific/causal model
3) Define a corresponding statistical model
4) Validate the model on simulated data with known parameter values
5) Analyze the real data
6) Describe how you would expand the project based on what you have learned

You will conduct this 6 step process on a real dataset of your choosing. There are obvious limitations to how in-depth a project can be when one only has part of a semester to complete the project. Therefore, the expectation is not that your project will be a finished product for a company or lead to publishable results. However, your group will need to complete each of the 6 steps described above.

There are a few additional requirements:

- The project must involve causal inference
- The causal model must include at least 3 variables including a treatment variable, an outcome variable, and one potential confound
    - For the purposes of this project **DO NOT** define a causal model that considers more than 3 variables. Properly modeling 3 variables may require more work than you anticipate and it is better to execute a smaller project well than to shoot

for a larger project that you don't have enough time to develop properly. View this project as your initial modeling effort. Step 6 above allows you to think about and describe how you would further develop the model.

- Project code and analysis will be submitted on Gradescope
- A code walkthrough in the form of a screencast is required
    - Only group members' voices are required (no need to include videos of yourselves presenting your work)
    - Each group member must be involved in the screencast
        - Introduce yourself when you begin presenting
    - Screencast should be a maximum of 15 minutes in length

**Choosing a Project**

The basic consideration for any project is a question of interest. A music streaming service might ask why certain customers cancel subscriptions after 6 months of using the service. A computational biologist might be interested in understanding why certain individuals tend to have severe reactions to a medication while others suffer no ill effects. A public health administrator might want to understand what messaging best encourages individuals to return for a second dose of a vaccine sequence. The number of questions that can be answered (assuming data exists to help provide an answer) are endless.

You may not have any burning questions that you would like to tackle. If you do, this is a great opportunity to put what you have learned in this course into practice in order to investigate and, hopefully, provide some answers to a question of interest using data. If not, here are some possible data sources for a causal inference project:

Twins Dataset:
https://github.com/AMLab-Amsterdam/CEVAE/blob/master/datasets/TWINS/ReadmeTwins

National Study of Learning Mindsets:
https://github.com/grf-labs/grf/tree/master/experiments/acic18

Jobs Dataset:
https://paperswithcode.com/dataset/jobs

ChatGPT Advice:
https://github.com/petezh/ChatGPT-Advice

ICLR Reviews:
https://cogcomp.github.io/iclr_database/

News-Tweet Dataset:
https://github.com/bywords/NTPairs
**Warning: There may be a good deal of data engineering work needed to get useable data for the News-Tweet data**

Various R datasets:
https://cran.r-project.org/web/packages/causaldata/causaldata.pdf
Note: If you are unfamiliar with R and need help getting access to the data, ask for assistance

These are some suggestions if you have no idea where to begin in identifying a dataset. Keep in mind that these datasets are often related to work that has already been published. Do not simply re-do the analysis that was completed by someone else. If you want to find inspiration from a previous analysis, feel free. But be sure to document what source you are using for that inspiration as well how your analysis builds on/differs from what was done previously.

If you already have your own question and want to use a data source other than one mentioned above, go ahead. This project is your opportunity to tackle a problem of interest and use the knowledge that you have gained from this course to do so.

**Evaluation**
Expectations for each component of the project are listed below

*Proposal Requirements:*
- 1 page (maximum)
- A clear question to be answered/estimand identified
- Identification and description of the data to be used
    - Each group must work on a unique dataset
    - Once a dataset has been selected, add it along with your group members names to this Google Doc to ensure no other group selects the same dataset
- Causal model in the form of a directed acyclic graph (DAG)
    - Ok if it changes in the final submission
- A proposed statistical model
    - Ok if it changes in the final submission
- Submitted in PDF format on Gradescope
- Feedback will be provided within 5 days after proposal due date

*Screencast Requirements:*
- Approximately 15 minutes
    - Exceeding this limit by more than 2 minutes will result in a grading penalty
- **Each group member must speak during the presentation**
- Required information
    - Clear explanation of the question/estimand investigated
    - Description of the data and the variables included in the causal model

    - Any scaling/transformations performed on the data
    - Brief discussion of the statistical model used and how it was validated
    - Answer to the question/summary of what was learned

*Deliverable Requirements:*

- At a minimum, a well-organized and detailed Jupyter Notebook (ipynb and pdf versions)

**Grading Guidelines**

The following list of questions will be used when grading the project. If your submission will result in the questions below being answered affirmatively, you can expect to receive a good grade on the project.

- Submission
    - Has the group uploaded a Jupyter Notebook containing code for the project?
    - Has the group uploaded a PDF version of the writeup?
    - Has the group uploaded the data or identified the location where the data can be found?
- Introduction
    - Is a description of the data provided? write up
    - Is a clear question/estimand identified? write up
- Causal Model
    - Is a causal model described (preferably with a DAG)? write up
    - Are the assumed causal relationships between variables clearly described? writeup
    - Are causal model variables clearly labeled? writeup
- Statistical model
    - Are the priors used in the statistical model(s) justified? *This justification can be the result of prior predictive simulation, using information from outside of the dataset, or some combination of both.* justification is writeup / Prior Predictive Simulation is Code
    - Is the choice of distribution for the outcome variable reasonable given the observed data? writeup
    - Does the statistical model match the assumed causal model such that the question of interest can be addressed? writeup
    - Is each statistical model used described (preferably using mathematical notation) in its entirety? writeup
    - Is confound properly handled statistically? writeup
    - Has the statistical model been evaluated on simulated data where the parameter values are pre-defined providing evidence that the statistical model can estimate the parameter values? Code
- Posterior Model
    - Is the correct computational model defined based on the statistical model? Code
    - Is the underlying code for the computational model provided? Code
    - Is the underlying code organized so that it can be clearly understood? Code
    - Are the built-in diagnostics for the posterior samples utilized to assess the quality of the sampling of the posterior distribution? Code

- Posterior Predictive Checks
    - Is there an attempt to reconcile the posterior estimates to the observed data sample visually? Code
    - Is there a discussion of the results of the posterior predictive check(s) that examines how well the posterior approximation fits the observed data? Writeup
- Discussion
    - Is there a discussion of what was learned from the model that addresses the question under consideration? writeup
    - Is the confounding variable explicitly addressed in the discussion? writeup
    - Are the conclusions reached supported by evidence from model results? writeup
- Screencast
    - Was a screencast properly submitted?
    - Is the audio clear?
    - Does the audio match the results that are shown on the screen?
    - Is the question that is being investigated clearly described?
    - Does each group member clearly identify themselves when their part of the presentation begins?
    - Is proper terminology used when describing the model and results?
    - Does the length of the screencast respect the time limit?

Please ask questions about requirements/expectations early and often. Our class's discussion forum will be used to answer final project questions as they arise.

**Good Luck!!**