Project 2 Proposal

Michael Langmayr (ml8vp), Jonathan Gomez (jag2j),   
Yihnew Eshetu (yte9pc), Nathan England (nle4bz)

# Objective and Starting Questions

We will use data provided by the Department of Education to answer understand emerging trends and outcomes in collegiate education. To begin our data exploratory work, we pose several initial questions in four categories, which we plan to narrow down to one or two primary questions. Regardless of the final questions elected, our questions will seek to relate important social issues, economic or demographic, to the data.

## Category: Earning potential

We want to investigate the debt and/or opportunity cost incurred by college.

* **Question**: How many years does debt take to pay off?  
  *Example model*: Year ~ program\_enrolled + college\_group + program\_cost
* **Question**: At what year after graduation do cumulative earnings exceed median wage?  
  *Example model*: (C - C(year)) ~ ...
* **Question**: Can we predict the mean earning power S based on program factors?  
  *Example model*: S ~ program + college + …

## Category: Completion

Can we predict student completion?

* **Question**: The completion of an average student?  
  *Example model:* completion ~ income\_bracket + work\_part\_time + first\_generation + ...
* **Question**: The yearly completion rate of a program?

## Category: Enrollment numbers

How is the amount of enrollment in a program changing over time?

## Category: Enrollment numbers

How is enrollment changing by demographic and/or major?

* **Question**: Does the make-up of first-generation students change over time?
* **Question**: How many people by gender are enrolling in STEM over time?

# About the Data

We are using data collected by the U.S Department of Education through the College Scoreboard project (<https://collegescorecard.ed.gov/data/>). The College Scoreboard project was designed to provide reliable and unbiased information about college performance.

Armed with this accessible and accurate information, students and their families are able to make more informed decisions, and measure college outcomes and the causal impact of each postsecondary institution. This project initially started in 1996 and still provides transparent information for all undergraduate degree institutions to students and families to this current date. Many of the data points are made possible by federal reporting from institutions, federal financial aid, and tax information.

# Data Cleansing

Before we can start the data needs to be filtered and harmonized. At the first glimpse we’ll be checking for typographical errors and remove them. Afterwards we’ll need to filter the data down to what we actually need to answer the questions we posed. Out of the box the dataset was one large table with over 200 columns. For our purposes we can narrow this down to 10-20 columns that we can use as response variable or predictors.

The dataset has very cryptic column names such as “CIP26BACHL” which stands for “Bachelor's degree in Biological And Biomedical Sciences.” Those abbreviations are explained in a separate yaml file and we’re planning to make those names more self-explanatory before we feed them into the model.

Lastly, once we have the data in the right format, we’re going to look for outlier values and remove them to increase the accuracy of our model.

# Methods of Analysis

We anticipate this project to use several modeling methods learned from the course including multiple linear regression and logistic regression.

Multiple linear regression will be used to answer questions with quantitative outcomes (response variables). These questions include the objective questions pertaining to earning potential and enrollment numbers.

Logistic regression will be used to answer questions with binary outcomes (response variables). The questions include the objective questions pertaining to completion (i.e predicting whether or not a student is likely to complete their program).

Specific modeling techniques we plan on utilizing include (but are not limited to):

* model selection through the use of automatic search procedures
* model validation by assessing model fit, coefficient interpretation, and utilization of training and test sets
* Performing model diagnostics and implementing remedial measures through the use of residuals analysis, looking at leverage and measures of influence, and transforming the response and/or predictor variables when necessary