**Please provide a general description and justification for your project.**

The Data is Plural archive includes a dataset on social scientists testifying in congress. People who have very extensive experience in a certain topic could provide an excellent resource in the consultation of legislation. Because the individuals in the dataset are social scientists, one can assume that they would prioritize social benefits of whatever legislative process the congressional committee is going through. For example, an economist might make suggestions on the budget based on the financial needs of the country. The presence of social scientists in congress could have positive effects overall. Understanding what makes a solid testimony could be important for those interested in speaking to congress and making a strong impact.

**Who is the end user for this project? What problem does it solve for them? Is the problem urgent and pervasive? Is someone willing to pay for a solution?**

The end user theoretically would be those giving testimonies to congress and organizations that employ them. The problem that this solves is avoiding weak statements to congress. Testimonies can be the most memorable portions of a hearing- speeches are sometimes circulated through media and have the ability to even sway public opinion. Optimizing testimonies could save time for those delivering speeches. The most urgent problem is lack of representation of social scientists other than economists. Other social scientists can provide valuable insight on issues of structural inequity, mental health, etc. Providing social scientists with consultation on how to develop an effective testimony backed by data is crucial.

**What are the minimum requirements for this project to provide value in solving the end user's problem? How much can be accomplished in 8 weeks?**

My goal for a minimum viable product would be creating a report with all my findings based on text analysis of past congressional testimonies and the main takeaways of this analysis. At the minimum, the project can showcase some interesting observations and common themes from exploratory analysis of testimonies given by social scientists. In 8 weeks, I could spend my time producing concrete suggestions that are tied to desired outcomes for those delivering speeches to legislative bodies. For example, looking more deeply at the nature of these hearings and investigating how testimonies influence the passage of bills or other pieces of legislation could provide insight into how researchers can provide convincing arguments to their audiences.

**What data would you ideally have for this project? What data can you reasonably access? How do you reconcile any differences between these sets of resources?**

**What, if any, solutions already exist for this problem? What additional value does your solution provide?**

**How do you plan to validate the success of this project beyond measuring model performance? For example, how if your model "improves accuracy by 10%", how would you contextualize the effect of this improvement to a nontechnical stakeholder?**

I am curious about the tangible effect that these testimonies could have on actual legislative change.

Of course, the opposite could very well be true: what if their testimonies are not listened to and respected, even if their advice is beneficial? What if these scholars, academics, thinktank researchers, etc. have their own special interests they are trying to promote? What if some members of certain disciplines are given more credence than others?.

My research question is whether or not social science testimonies make an impact in congressional hearings? By impact, I decided somewhat arbitrarily on a few factors: the total presence of social scientists at hearings, the ability to get bills passed during related hearings, as well as the ability to incite change in specific committees. Using the Social Science Congressional Testimony dataset, I first looked at the most popular committees that social scientists attend. For all of time, the most popular is the Joint Economic Committee, but from 2015 onwards the most popular committee in the data was the House Committee on Financial Services. I looked at 2015 onwards because this represents the most recent sessions of congress in the data: the 114th to 116th sessions. I figured that these data would be the most relevant to my research question because I would not have to take historical factors into consideration as much. I created a similar frequency table for the witnesses who spoke and the most represented disciplines in the field.

Next, I did some NLP and sentiment analysis to understand the nature of the hearings. I used the title\_description variable to create a corpus of text, a document term matrix, and another variable calculating the sentiment value of the description using the syuzhet package. This process of sentiment analysis is concerned with “plot” development as well as sentiment by utilizing several different existing dictionaries. I felt that this was suitable and comprehensive enough to calculate the sentiments of the descriptions of the hearings because it would give an idea of the positive/negative attributes of these hearings while also giving us a sense of how “well-developed” the intention of the hearing was. Looking at the five most common committees in the data, the sentiments ranged widely from very positive to very negative across hearings. This could suggest that these hearings were of a very pressing nature and required experts to speak on the subject immediately.

Because the data were a bit limited, I retrieved some additional data using the ProPublica API for data on congress. For the most popular committee in the original dataset where most social scientists were present, HSBA, the sentiment of meetings spike in intensity in 2018 both positively and negatively. The document term matrix and word frequency table suggest that most of the hearings at the time were related to regulations on investment protection. Combining the full API dataset on congressional hearings with one of bills that were passed and a sentiment analysis reveals that the committee most likely to pass bills is the Committee on Rules. The chart created also shows that bills passed have overwhelmingly positive sentiments for their summaries.

This preliminary analysis is very limited with regards to the ability to answer my original research question. It was challenging finding ways to use the data from the government API in concert with the original dataset on testimonies when there were no related variables to join the datasets. However, through exploratory analysis, I did find some factors that do lead to “impactful” congressional hearings, most notably the sentiment and “intensity” of the hearing and/or bill description. The next steps for this project would be perhaps to find actual transcripts of the social scientists present to analyze the information shared and their persuasion tactics used. A hypothesis test can be completed on whether or not hearings with an expert present make a difference or not in the ultimate conclusion of the hearing. If we understand what makes a strong testimony, social scientists hoping to make a beneficial social impact can follow these guidelines to make sure their voices are heard in the elected body.