# Capstone Project Proposal

Government initiatives should maintain a solid connection to research and data that supports decision making. With the proper data to support decision, government entities can act with some confidence to achieve positive results from its policy implementations. Using Machine learning I would like to offer a road map for decision makers by connecting research papers and their conclusions to data as cited in their research documentation. Using NLP machine learning techniques I want to scan documents for relevant citations which support the predictions contained in the conclusions of the research document.

This project is derived from a Kaggle competition proposed by the Coleridge Initiative entitled “Show US the Data”. The competition challenges data scientists to use machine learning to “find the hidden-in-plain-sight data citations”. The initiative seeks to restore confidence in in evidenced based research so that governments can provide transparency and validation for the reasons for policy decisions.

The data sets that I have gathered to train and test my algorithm are directly from the kaggle competition. They consist of three files containing url links, document abstracts, and complete research documents. The data is mostly clean but will need to be explored for inconsistencies and missing data.

Using NLP I will seek to gather evidence of research data and connect the data to conclusions reached in the research papers. I envision that my algorithm will categorize language first as asserting a conclusion. Then connect those conclusions with data citations. Each conclusion may have a corresponding citation and a statistics.