What I will be implementing for this thesis:

1. Simple python library for reading and writing from csv and excel files. Similar software exists but it is not as robust as is needed for this project, many libraries either only read from files or write to files, and ineffectively at even that. ([OpenPyXL](https://openpyxl.readthedocs.io/en/stable/) is my leaping point).
2. Library for parsing natural language questions into simplified logical queries.
   1. This is the most traditionally *academic* part of my project.
   2. There are already numerous papers using a range of approaches for this problem (a few cited below) –– deep learning, knowledge graphs and so on. However, none of these libraries are designed to process queries as vague and open ended as the type expected in this project. Most libraries deal with simple queries (“Where was the Olympics held in 2010?” or “What was the average age of participants?”). Thus, I will test multiple of these different network architectures to find a baseline for our problem, and then I will build in domain specific logic for handling econometric questions as well as retrain the tabular QA networks on dataset’s specific to the econometric queries I expect.
   3. I will be able to evaluate this work in the same way one might evaluate any new machine learning approaches -- I will compare the f1 scores with previous work on a new testing dataset.
3. Python library of advanced econometric analysis techniques. [PyEcon](https://pyecon.org/) already provides an easy to use API for econometric analysis, however it is far too basic for our needs. We will need to greatly expand the depth and specificity of the library -- adding methods for performing unsupervised learning as well as more advanced statistical analysis.
   1. Sample questions our API should be able to answer:
      1. “What is the ‘strongest’ instrument to use in this dataset for a regression of X on Y?”
      2. “How many lags is optimal in a standard time series regression on Y, using an Akaike Information Criterion?”
      3. “In an ADL time series model, which variables exhibit Granger causality on Y?”
4. Simple GUI and CLI for interacting with the finished project.