**Gabriel Stengel**

**Advisor: Jaswinder Singh**

**Princeton Computer Science**

**WORKING TITLE:**

Econometric AI: Flexible Statistical Querying on Unseen Data

**GOALS AND OVERVIEW**

*I) Project Outline.*

I will develop an automated statistical querying tool that can be flexibly applied to a variety of econometric problems. On the front end, the engine will use machine learning natural language processing (NLP) algorithms to interpret natural language queries. On the back end, the econometric libraries I develop will use machine learning inferential capabilities to allow the engine to handle open-ended statistical queries. Additionally, I will implement unsupervised learning techniques in order to generate brief data summaries for users –– eliminating the need for very basic data analysis and queries.

*II) Goals.*

The engine should improve on conventional dataset analysis in two complementary ways. First, the user’s line of questioning can be open-ended. (Rather than hand-selecting variables and then using a computer to analyze a hand-selected economic relationship between them, she can use the toolkit to assist both her choice of variables and her choice of economic relationship.) Second, the manner of questioning is itself more flexible, in that it permits natural language inquiry. It will not require specific programming syntax. The end result is that users of the final product need not be well-practiced statistical coders. One could do serious and mature economic data analysis with no programming experience and only basic knowledge of the statistical tools customarily used by empirical economists.

Additionally, the project should prove fruitful outside of the fairly niche domain of academic, economic analysis. The work should lay the groundwork for a general purpose Question-Answering system that allows anyone to perform basic machine learning analysis on any dataset. NLP question answering on tabular data is a studied field, but by baking in greater statistical domain knowledge I hope to augment the usefulness of such systems.

*II) Empirical verification and Evaluation.*

Once the toolkit is ready for use, I will demonstrate its practicality by using it to

suggest improvements to existing economic studies. One of the initial motivations for this project was to help my friends and peers in the Economics department improve upon their independent work.

For example:

Last year my friend and partner in this endeavor, John Willet, studied the effect of congenital afflictions on depression incidence in American adolescents. All studies involving depression incidence struggle with the selection bias implicit in data related to qualitative, fully self-reported conditions like depression. John dealt with the issue like many researchers before him – by using an existing symptom inventory test as a benchmark to hand-select questions from the questionnaire that “seemed” related to depression. Then he used an ordinary set of linear regressions to check his constructed indicator against the actual depression indicators in the questionnaire.

The toolkit I build this year will make it possible to automate the selection of

questions in the questionnaire that are most predictive of depression, making the

resultant indicator more predictive and robust. Thus, we will be able to look back at John’s junior work and see what analysis he might’ve done if a tool like mine had been available to him at the time.