# Capstone Proposal

**Background:**

For my capstone project, I am planning on doing two things.

1. Create a supervised learning model that can predict topics in a given text as a multivariate classification problem.
2. Create an unsupervised cluster ensemble model that can help visualize topic clusterings.

Currently, I am working under Prof. Dr. Donald Sull at MIT Sloan School of Management on a research project that aims to quantitatively analyze corporate culture. Additionally, I work for (and will soon become the CTO of) his company, CultureX, which provides in-depth corporate consulting. The goal of my capstone is to further aid the investigation into Corporate culture using some other off-the-shelf tools (as all of our tools are proprietary). The data I speak of can be referred to as “the CultureX data”, and refers to the data being processed at both CultureX and MIT.

**What is the problem you are attempting to solve?**

We have a corpus of text from [glassdoor.com](http://glassdoor.com) that has been processed using our proprietary software. This methodology has proven very effective, but we have been curious for many months what kind of results we would get using Deep Learning, or other black-box models. Additionally, we want to investigate whether the clustering of our topics is close to what some other clustering models would produce. *It is important to note that some models like LSA and LDA have been applied with horrible results.*

**How is your solution valuable?**

This solution will be able to provide information as to whether our custom built approach is significantly better in its predictive ability, disregarding explanatory ability. If I find that my models both cluster topics very well as well as predict with a high rate of accuracy, that can tell us a lot about our analysis. So far our approach has vastly outperformed any other technique for topic modeling but would be interesting to see how much latent information the neural network can capture. We would use these results to either boost our confidence in our approach or make us go back to the drawing board.

**What is your data source and how will you access it?**

My data source (“the CultureX data”) will be a matrix containing pre-processed text (removed punctuation, spell-check, etc.) and *N*-columns, where each column represents a topic, and the cell contains 0 or 1, for whether or not the text talks about that topic. We have a vast amount of data, roughly 4GB+ of text data from company employees. To access this, I will just load the file from my local computer. To run my models, I will use an AWS AMI to run my processes in parallel.

**What techniques from the course do you anticipate using?**

I anticipate using much of the knowledge gained from my Neural Network Specialization, as well as my NLP specialization. Additionally, I will make reference to and use techniques from a paper by Justin Grimmer (Stanford) & Gary King (Harvard) to make an ensemble of cluster models. I will use many of the model evaluation techniques from the coursework, as well as some application specific ones that I find on my own.

**What do you anticipate to be the biggest challenge you’ll face?**

The biggest challenge I think I will face is the implantation of the cluster ensemble. The paper is very thorough, but I don’t know if time will permit me to fully implement. I might reach out to the authors and ask for the source code (because I can’t find it anywhere).

Additionally, I think evaluating the multivariate classifier might prove difficult, as well as building the language model (using LSTM and GRU cells).