Data Processing for Machine Learning on Patents
Albert Cai (EAS, WH 2018)
Advisor: Anoop Menon

My project this summer was to explore questions about innovation, such as "How do innovations affect others?" and "What types of innovations tend to be the most influential?" To do this, we analyzed patent data from the United States Patent and Trademark Office (USPTO). On their website, the USPTO has full-text data from every patent granted since 1976 and every patent application since 2001. As there have been millions of patents granted in that time, each of which can have thousands of words, we used machine learning to analyze the massive amounts of data. As a general overview, we had three main steps. First, we extracted the patent data from the USPTO into a workable format. Then, we processed that data into a different form. Lastly, this transformed data was fed into machine learning algorithms that predicted which patents were the most important.

My part of the project dealt mainly with the first two steps—extracting and processing the data. After extracting the data from online, I then worked with the group to determine how the data should be processed. For example, one simple metric we used to judge patent quality was forward citations, which is the number of patents that cite the current patent. To do this, we used the parsed data for future patent citations and counted the number of times each patent showed up. Then, in one experiment, we used the abstract text of patents and tried predicting the number of forward citations each patent would receive. Our machine learning algorithm still needs to be fine-tuned, but we reached preliminary results in which our machine was better than randomly guessing.

I learned about different facets of data science from this project, none of which I had used before. On one hand, I learned the technical skills involved with Python, SQL, and XML. Equally as important were the conceptual skills that I learned, like cleaning, reshaping, and interpreting the data. One of the most difficult challenges of this project was checking whether or not the hundreds of gigabytes of data was correct. It was impossible to manually check that every entry was correct,

so I learned to randomly sample different edge cases in the data to check that it was generated correctly. Luckily, the forward citation metric is searchable on the USPTO website, so I was able to check whether random patents had the correct number. I also checked edge cases—I was surprised to see that some patents received over 1,000 forward citations! However, some other metrics we generated were not as simple to check, and I had to learn to balance checking everything perfectly with how much time it would take.

On a broader scale, this project fit excellently into my education. As I'm majoring in Systems Science Engineering and Management, I was able to combine the two into a project that involved both. Working with the data used applied the skills I've learned through engineering, and understanding our results used a more general understanding of business.