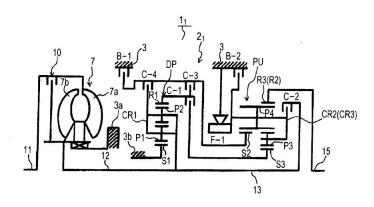USPTO patent documents contain drawings, of which contain figures and parts noted as number labels but not with text. The text is contained separately, and is what's needed in order to understand the patent. This creates friction for people being able to learn about how patents work.

By extracting figure locations and titles, part label locations and text; each document should allow for an interactive interface to allow for easier reading. In Figure 2 below you can see that the patent image has numbers, these are the numbers that will be extracted. The description will be on a separate page, of which each description will be matched with the corresponding figure number. This will increase the speed at which the document can be read.

FIG. 2

This could be useful in a few different ways; one could be for your average laymen trying to understand a patent, it creates a way that they don't have to try and find the separate patent part explanation, it could also help a patent clerk. If they need to perform a comparative analysis on a handful of similar patents this would increase their ability to find what they need more quickly. Based on the giant back log of the current US patent system, even if this solution gave a 5% increase, it would help.

The USPTO provides 10 TB of image data in bulk download or api access. I will be using a portion of this data.

I anticipate on using the overall concepts of machine learning; using a test & training set. The libraries I plan on using are scikit image and possibly opencv. These were not used in the course, but I believe that they're a close step to what has been in the course.

The biggest challenges that I expect are going to be working with large amounts of data, as well as the learning process for scikit-image and being able to identify the figures at a high rate of success.

If this project does idea doesn't work closely enough with the course material, then I can add on NLP. When looking at the descriptions of the patents, they can often be both intentionally and unintentionally overly filled with technical jargon. This creates another difficult place to be able to read the documents for the layman, including myself.