**Proposed project**:

My proposed project is a data mining project that scrapes data from one website and compares it to another and reports the differences.

There is a website called, HGMD (https://www.qiagenbioinformatics.com/products/human-gene-mutation-database/), which has a list of all published mutations in every gene. The user enters the specific gene and the website shows a list of all of the published mutations, which contains the variant and position. However, some of these publications are from 20 years ago where they might not be considered mutations today, now that we have more information. Genetic information is constantly being updated as more research is being conducted. To check if the published mutation is still considered a mutation, there is another website, exac (http://exac.broadinstitute.org), which has a list of all known variants for all known genes that are not mutations. The user can enter the specific gene and get a list of all the variants and the position. Then the variants and positions are compared from each website, if the two match, then the variant is no longer considered a mutation. Currently, geneticists need to perform this analysis manually by visually comparing the websites. The reason I want to do this project is because my wife and her colleges need to do this manually and her bioinformatics department is too busy at this time to find a solution. My hope is to find a solution and present it to her team and eventually get a job in their bioinformatics group.

To perform this project I plan to have the user enter the URL’s with the gene info into R. Then R will scrape the websites and put the data into data frames. I will compare the data frames and print the differences. I am not sure of the details now, I will need to research the best method to do this (use R and compare or make a database with the info and use SQL). I want to also see if it is possible to enter the gene into R, as a text input, and have R search the database for the gene info then scrape the website, I am not sure if it is possible and I will need to research it.

This project is dependent on if it is possible to use R to scrape the information from the websites, I am not sure if it is possible with R and I am currently trying to figure it out. Also, I will need permission to use the HGMD website, it is very expensive to get a licenses and I will ask the company for a 2 month trial for a graduate research project. If they give me a trial then I can do it, if not, it is not feasible to pursue this project.

**Backup project:**

Can heart disease be predicted with machine learning? I plan to train a logistic regression model for the data by using the glm() function. This will let me review the p-values where I can assess the data for good indications of heart disease. I will also look at the model deviance and test set performance. Once I start the project, I will see if I can apply lasso to the data set and assess the results for regularization. The data set I will be using is the Statlog (Heart) Data Set from the UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29>.

The reason I want to purse a topic on heart disease is because my father just had a stent put in his heart two weeks ago, he had 70% blockage in his left ventricle. Then last week, he was going in and out of atrial fibrillation (AFib) and now he is being treated with beta blockers. I am interested in seeing if and how this can be prevented in the future.