Capstone 2 Project proposal:

**Problem Statement:** Tennis is an extremely competitive sport that involves physical skill, mental capability, and logic. But what exactly determines the winner of a tennis match? There are a multitude of different statistics in a tennis match, including number of aces, number of break points, number of double faults, and many more. What effect do these factors have on the winner of a tennis match? This is the question that this project will explore. More generally, the problem to be solved is determining what statistics or areas are most important in determining the winner of a tennis match. In addition, machine learning models will be created to predict the outcome (win/loss) of tennis matches. There are many possible clients who could benefit from this project. Most obvious would be the tennis players and coaches. Based on a predictive model, they could practice/teach the areas of the game that were found to be most important in the model. Other possible clients include tennis bettors for obvious reasons; namely, to determine which player to bet on. Finally another possible client could be sponsors, such as Nike or Wilson.

**Data Wrangling:** Kaggle has provided a dataset including detailed information of ATP (Association of Tennis Professionals, Men) matches from 2000 to 2019. The individual match information provided includes number of aces, number of double faults, 1st serve in percentage, break points saved, and more. While the data tables from each individual year only include a winner ID and loser ID, Kaggle has also provided a separate table including information (country, birth date, etc.) about the different players and can be referenced according to the IDs in the individual match tables. There are a total of 21 csv files: one csv file for matches in each year from 2000-2019 and one csv file for players. The first step was to combine the 20 csv files containing match data into a single dataframe. This dataframe had 59,430 entries and 32 columns. However, each entry in this dataframe represented a single match, including statistics for both the winner and the loser in the same row. From this dataframe, I created a new dataframe where each row represented either a winner or a loser. Additionally, I added a column named ‘outcome’ where ‘1’ represents a win and ‘0’ represents a loss. There was some missing data where when one statistic was missing, all the statistics were missing. These rows with missing statistics were dropped from the dataframe. Several columns were dropped as well. Columns with irrelevant information, including tourney\_id and tourney\_name, were dropped. Columns that were the same regardless of winner or loser, including surface, score, round and minutes, were dropped. Finally, the best\_of column was dropped after eliminating all matches that were a best of 5 sets and keeping only the matches that were a best of 3 sets. Next, I supplemented my matches dataframe with player data, including the birthdate, name and country. From the birthdate, I extracted the age at the time of the tournament.

**Model Building:**

Construct various machine learning model (KNN, support vector machine, logistic regression, Random forest) to determine accuracy scores and find best model.