# Capstone 1 Project Proposal: MLB

By: Sinead O'Connor

1. **The Problem:**
   What is the probability that the home team will win a regular season baseball game?

2. **The Client:**
   Billions of dollars are bet on sports each year in America. Given this and the fact that are 2,430 regular season Major League Baseball games each year, it could be very useful to sports betting companies to have an accurate model to predict the outcomes of baseball games. They can use the results to set betting odds.

   In determining which factors are the most important in winning a baseball game, the model could also help MLB managers and coaches make informed decisions about building a winning team, for example if it's more important for them to have good starting pitchers or good hitters.

3. **The Data:**
   I will be using Game Log data sets from Retrosheet.org. The game logs contain information on the starting players of each team as well as statistics defensive, offensive, and pitching statistics for both the home team and the away team. I will import game log files from the website using urllib request.
   https://www.retrosheet.org/gamelogs/

   I will also use starting pitching logs from baseballmusings.com, which has statistics about the starting pitchers for each game such as Innings Pitched, and number of walks. Since starting pitchers are considered such an important part of the game. I will import the starting pitcher log tables by web scraping using BeautifulSoup.
   https://www.baseballmusings.com/ChooseStarterTeam.html

4. **Solving the Problem:**
   I plan to build several classification models to predict the home team's probability of winning. I will use metrics from the home team and the visiting team such as starting pitchers' ERAs and batters' statistics. I plan to train the data on nine seasons and test it on one season.

5. **The Deliverables:**
   As deliverables, I will submit all jupyter notebooks with the code as well as a slide deck with the final presentation.