Darren Lund
Dr. Jarvis
Oct. 20 2017

# Data Science Project Proposal

This project will be a solo effort to study the predictability of the NCAA March Madness Tournament. Using data scraped from kenpom.com and espn.com, I will analyze team, and individual performance in an effort to predict the score that team would achieve in a certain game, against a specific opponent. This will attempt to answer the questions of whether or not the outcome of a game can be determined by seasonal statistics, and how one or multiple cinderella teams can change the overall tournament.

Specific techniques I plan on using to accomplish this include using Bernoulli distributions to predict successful scoring for individual players to provide a more accurate overall team performance. I also intend to use the Central Limit Theorem by performing a multitude of simulations per game and taking the mean as the final score to get more accurate results. The reason why I believe this method may be better than others, such as the PageRank algorithm, is because while the PageRank algorithm may be able to reflect which team overall is better, each game has unique factors that may effect the overall result. A team that is better may still lose, as evidenced by the existence of cinderella teams and stories.

In order to measure my success, I intend to run the simulations on previous tournaments, as well as participate in this year's tournament and observing the number of correct predictions, both in who won, what the average score was for each team, and average number of points contributed by each player.

Some data will come from kenpom.com, and the rest will come from espn.com. It will be a combination of overall team statistics [from pages such as https://kenpom.com/team.php?team=Gonzaga, available only to subscribers (which I've already purchased)] and individual player statistics (from pages such ashttp://www.espn.com/mens-college-basketball/team/stats/_/id/222/villanova-wildcats). I plan on scraping the data directly from these sites.