

---

ISyE 2028 – Basic Statistical Methods - Fall 2015  
Bonus Project: “Big” Data Analytics  
Proposal (or Final Report)

---

Zachary Cole  
ISyE 2028 C  
902865943  
[zcole7@gatech.edu](mailto:zcole7@gatech.edu)

## **Problem Description**

I would like to determine which quantitative data about college basketball teams best predicts a higher seed, as well as greater success in the NCAA Tournament. After the college basketball regular season ends, a committee analyzes the success of each team and selects the best teams for a 68-team tournament to determine the national champion. The tournament, known as March Madness, is one of the most exciting events in American sports. One of the things that makes March Madness so exciting is that it is so difficult to predict, largely due to the high number of teams, many of which are relatively unknown to most fans.

## **Intended Result**

It may be easier to predict the outcome of some games by determining which factors are “overrated” or “underrated” by the selection committee. An example of what I mean by this is the Ratings Percentage Index (RPI). For many years the RPI has been a large factor used by the committee in seeding teams. However, it has been criticized as not being a very good factor in determining how good a team is. If I find statistically that the RPI truly does not predict a team’s success in the tournament, a reasonable assumption may be that a team that appears to be seeded too high, but has a strong RPI, may underperform in the tournament (if they earned their high seed due to a high RPI, and RPI doesn’t mean much, this team is likely overrated). Hopefully I will find which statistics actually affect a team’s success, and which statistics simply seem like they would, either to casual fans or to the selection committee.

## **Methodology**

I intend to analyze data from NCAA Tournaments since 1985 (before 1985, the tournament was 32 teams – it went to 64 in 1985, and has since increased to 68, but the results from the 64-team tournaments are still relevant and give me a larger sample size). I will use Python to scrape statistical information such as season record, conference tournament result, points per game and points allowed per game, from internet sources such as sports-reference.com and kenpom.com. From there, I will use R and Excel to analyze the data and present it in a meaningful way.