# Data Analysis Project

**Group Members:**

Paul Harmon

Justin Gomez

**Dataset:**

The data were obtained from a repository on Kaggle.com related to the March Machine Learning Mania competition. The response variable is binary (teams either won or lost games) and there were several covariates of interest. Prior to the analysis, we re-coded variables so that the covariates were based on ratios rather than raw counts (for example, Free Throw Shooting Percentage) as well as considering differentials between winning and losing teams. The variables considered are below. Please see the dataset that is attached to this document to see the raw data.

**Raw Variables:**

* Winning Score
* Location
* Overtime

**Ratios:**

* Field Goal Percentage (Winning and Losing teams)
* Free Throw Percentage (Winning and Losing teams)
* Three-point shots made (Winning and Losing teams)

**Differentials:**

* Total Rebound Differential: (winning team rebounds – losing team rebounds)
* Assist differential
* Turnover differential
* Steals differential
* Block differential
* Personal Foul differential

**Research Questions:**

We are interested in determining the factors that influence the outcomes of college basketball games in order to predict which teams are more likely to win a matchup. The key questions of interest are as follows:

1. Does home court advantage influence the outcome of a game, after controlling for on-court performance factors?
2. Do winning teams turn the ball over less than losing teams?
3. Do any of the variables measured modify the effect of other covariates of interest? More specifically, is there an interaction between playing at home and the other covariates of interest?

The key goal of the analysis is to create a good predictive model that can, with some reasonable degree of accuracy, generate predictions as to whether a team will win or lose a game. We plan to break the dataset into training and testing set in order to assess the quality of predictions. Since the data pertain to multiple seasons, the training and testing sets are going to be chosen using a stratified random sampling method that accounts for different seasons.