**NBA Hot Hand Theory - Milestone #3**

Group: Zainab Doctor, Hai-Tsang Huang, Zecai Liang, Bailey Liao

**1. Project Approach and Cleaning the Data**

The variables we are interested in using are Game\_ID, shot\_number, period or game\_clock, shot\_clock, shot\_distance, shot\_result, close\_defender\_distance, FGM and player\_ID. We will clean the data eliminating shots with shot\_clock as NaN or game\_clock under 2 seconds, with the assumption that these may be last-second attempts during which the ball is thrown up. Next, we will use Game\_ID and player\_ID and shot\_number to calculate the shots per game per player, and exclude games where players only attempt 4 shots or fewer per game based on previous hot hand literature using 4 as a minimum streak number. This number may increase.

We plan to build our initial regression model of shot likelihood based on the shot\_distance and close\_defender distance. We need to develop our approach a little bit more as we explore the data.

Other features we would like to look at are whether players are more likely to get “hot” when playing at home versus away, with the hypothesis that they know their home court better. Additionally, if players are stratified based on their average points per game across a season (ie 0-10, 10-20, 20-30+ points per game), are higher tier players more likely to get hot? Are there mid-tier players that get "hot" by having nights that look like those of upper tier players? Another aspect that would be interesting to explore is using Defender\_ID to investigate whether certain defenders are better at stopping or breaking streaks.