# Cluster Mining Patterns from Hex-Path Trajectory Images

The goal of this project is to derive a pipeline for building descripting trajectory images of basketball action using hex-bins and pass those along to a visual deep learning module used to cluster and derive patterns and variants in player movements. Such a pipeline would be capable of taking raw data representing many individual player movements for a given basketball action and provide an analysis of the strategies present in the data set. Below is a summary of the dataset, AI mechanism used in the pipeline and the projected value such a pipeline would bring.

## Data Description

The data itself is contained in a relationship postgres database and contains identifying data for the teams and players involved in the basketball games, as well as highly precise 2-dimensional coordinates for all players present on the court, captured at a rate of 25 times per second throughout the duration of play. This dataset is derived from two sources: SportVU live-tracking data and event-annotations from the NBA. The first dataset is the raw coordinate data obtained using video capture devices and live games. The second is a set of high level descriptions of events in games with information like the teams in the game, the score at the time, the quarter and timestamp the event occurred, identifying information for the players involved, and a text description of what happened. A fair amount of pre-processing has already been done to identify the plays that these coordinates belong to, as well as to isolate and trim them to focus on players directly involved and remove outside noise. In addition, a manual-labeling process has already been performed, identifying which actions in the dataset are instance of dribble-hand-offs, a common play-action performed in the NBA. The focus of this project will involve the trajectories for the positively-labeled actions, and a series of steps leveraging python libraries for transforming the data and creating the images, an additional contribution of the project.

## AI Mechanisms

The main target of this project is the unsupervised learning algorithm k-means, specifically the implementation offered by the OpenCV python library. K-means is a form of clustering that uses Euclidean distances based on similarity metrics to isolate patterns and variants into ‘clusters. This is often used with images for image recognition but can be applied in a more abstract manner. By allowing the dataset itself to define the patterns and variant, we are essentially looking for emerging trends in basketball strategy. The clusters can then be labeled using domain specific terminology, allowing for the layering of higher levels of analytics and performance metrics.

To optimize the clustering process, a certain level of data abstraction is required. The raw dataset itself contains highly precise coordinate data and rendering images of those exact paths would results in a highly diverse dataset that is difficult to cluster. Often, cases such as this require the use of a convolutional neural network (CNN) to ‘blur’ out the precise details, leaving a more generalized interpretation of the trajectory path that is more easily comparable. This results in tighter clusters and better analysis. By representing the trajectories in a hex-bin path, we seek to explore an alternative to a CNN in prepping images for k-means clustering.

## Projected Value

Sports analytics is obsessed with pattern detection and evaluation, and a pipeline like that proposed here would be capable of providing a new means for obtaining and examining those patterns. The practical application as far as coaching, development and strategy is clear, in that you have better and more contextual metrics to evaluate player decisions and actions. The process itself, however, could be abstracted and applied to any field with highly precise trajectory data and a desire to analyze patterns. For example, if an airline had records of all of its flights out of Atlanta to Denver in the last year, they could create a hex-map of that data to abstract the precise data and cluster those images to identify patterns and variants in flights that could then be mapped back to key concerns like flight times and fuel costs. As a result, this project seeks to not only offer an unsupervised learning approach to sport strategy analysis, but also a more abstract process by which any data set of trajectory coordinates can be mined for patterns.