# Pattern Mining Hex-map Trajectory Cell Paths: Approach Specification

## Data Manipulation:

### Background

In professional basketball, a variety of real-time strategies are implemented over the course of a game. Considering the competent environment, the identification and analysis of these strategies has drawn a great deal of attention. With the advent of advanced analytics, a variety of intriguing applications for machine learning to perform this pattern mining and analysis have emerged. One such interesting application is deep learning systems that study player trajectories. They do this by highlighting moments, or actions, of interest and training the model on hundreds of images representing the paths the players took during the slice of time. In this analysis, we seek to replicate a similar pipeline, with a few novel variations, to perform pattern mining and analysis of dribble-handoff actions using NBA SportVU positional player data, hex-map plotting, and k-means clustering.

### Candidate Selection/Classification

Already performed was a long series of extraction steps by which the dribble-handoff action (DHO) was identified within the data set. These began with preprocessing steps to combine the positional data with event descriptions that provide greater context to the moments, players involved, times things happened, and the resulting shot attempts and scores. The data was the scanned for consistency, a set of rules were created to identify potential DHO candidates, and candidates were generated for 15 regular season games, each with at least 4 quarters and come containing as many as 2 additional 5-minute overtimes. These candidates were then manually labeled by watching game recordings, with an average recall around 89% for the rule-based algorithm.

These candidates were then converted into a postgres database where feature vectors describing the characteristics of each action were created. Candidates were split into ‘approach’ and ‘execution’ sections that correspond to the time before and after the ‘screen’ (the moment the two offensive players are closest during the action). A total of 50 unique features were then generated ranging from: positions of offensive players and the ball at each moment, relative distances, average speeds, slopes and intercepts of trajectories, number of players past half court, whether it was an inbound pass, the offset into the play and the game, and the player archetypes of the players. There was in total 1850 candidates created and labeled through this process.

The training set was then split 80/20 and trained on a variety of machine learning models including: support vector machine, decision tree, gaussian naïve bayes, multi-layer perceptron, and keras neural network. The best performing classifier used to identify the final training set for this project was the keras neural network, with an f1 score of 95.9%.

### Trajectory Mapping

Once the positive actions were identified from the neural network, hexbins can be created and plotted for the candidates. These were created by converting the raw coordinate data into cell locations and plotting them using a colormap based of the density of points to encode the momentum of the players. All coordinates were converted to one half of the court, to preserve the patterns present by both the home and away teams. Three sets of images will be created: offensive players, defensive players, the ball. Some additional work will need to be done to identify the defensive players and map their trajectories. Ultimately the dataset will consist of these labeled hexbins images.

## Data Analysis:

### K-means Clustering

K-means clustering is a deep learning machine learning algorithm used to study the similarities in datasets. It is particularly useful for identifying groups of like instances within unknown data, which makes it a particularly good fit for this project. By analyzing the trajectory images created using the classification pipeline, we can identify variants in the DHO strategies as they are performed in game by players. The bulk of the machine learning for this project will be configuring the images, data, and model to extract the most meaningful and interesting clusters from the deep learning module.

### Cluster Maps

Once clusters have been identified, we can represent them in different ways. For example, we can take an individual player, and see how many times in the dataset they were involved in a particular set of clusters. We could do a similar thing for a team or player architype. These mapping are cluster profiles and can tell us a great deal about the kinds of strategies and decisions players, coaches, teams, and organizations make.

Another interesting plotting tool for clusters is radar maps. A radar map is a 2-dimensional plot that renders a polygon whose vertices are the totals for the different clusters. These maps themselves are a visual embedding of pattern structures, and can be compared across teams, players and even quarters or halves of a game or the season.

### Pattern Evaluation

Ultimately, the points of identifying strategy is to optimize it, so any pattern mining ultimately requires a means of evaluation. When it comes to actions in the NBA, this is not a straightforward problem. Although it is not the outright goal of this project, the following is some consideration for how quality metric frameworks might be layered onto this pipeline.

Some quality metrics evaluate basketball actions based off how many points they generate on average, which is good for whole plays, but not as good when looking at smaller independent actions that might be part of a larger sophisticated effort or dynamic set. As a result, some point-per-possession model might provide some insight, but the more insightful project might be investigating the space creation because of these actions.

There is a concept in the literature and NBA vernacular referred to as court realty. This is the idea that not only is space valuable in basketball, relative to the small dimensions of play, but that much like real estate, the most important feature is location. As a result, a space evaluation model that values the type of space created would be a powerful analytic layer to this pipeline. The space evaluation using plots of players in a confined space can be done in a variety of ways. One interesting approach from the literature is Voronoi diagrams, in which polygons are drawn between each of the players on the court according to their proximity to other players. These polygons are then assigned to one team or the other depending on occupant. A total amount of space created, then, can be calculated by taking the sum of the areas of each team’s polygons. These Voronoi diagrams could then be layered with offensive players shooting maps that indicate their offensive proficiency from different areas on the court as well as the defensive effectiveness of the defender. Such a final metric would then be capable of capturing the value created by an action by incorporating the quality of space and opportunity it created. This metric could then be used to rank and analyze the various cluster profiles of teams, players, and action variants.