Combining NBA Spatiotemporal data with player statistics to create a coaching assistant using Machine Learning.

Final Year Project Preliminary Report

Supervisor: Jamie Ward

Cleon Grant

Table of Contents

[1. Introduction 2](#_Toc1708018)

[1.1Purpose 2](#_Toc1708019)

[1.2 Motivation 3](#_Toc1708020)

[2. Aims and objectives 3](#_Toc1708021)

[2.1 Aims 3](#_Toc1708022)

[2.2 Objectives 4](#_Toc1708023)

[3. Methods 4](#_Toc1708024)

[4. Project plan 5](#_Toc1708025)

[4.1 Requirement Analysis and Feasibility 5](#_Toc1708026)

[4.2 Planning 6](#_Toc1708027)

[4.3 Design 7](#_Toc1708028)

[4.4 Develop 7](#_Toc1708029)

[4.5 Testing 7](#_Toc1708030)

[5. Progress to date 9](#_Toc1708031)

[6. Planned work 9](#_Toc1708032)

[7. Conclusion 9](#_Toc1708033)

[Appendices 10](#_Toc1708034)

[References 13](#_Toc1708035)

# Introduction

## Purpose

The purpose of this report is to outline how I plan to use Machine Learning to create a coaching assistant for NBA teams using the spatiotemporal data provided by analytics companies that have partnered with the NBA (1). The model will respond to game information by suggesting tactics and providing generated statistics for meaningful analysis.

Spatiotemporal data simply refers to data containing space and the time that relates to it. As NBA players play is a fixed sized court, positions can be tracked and mapped easily. Events can be conveyed combining positional and timed data. It could be a minor event such as dribbling the ball from the baseline to half-court, or even someone slam dunking the ball.

With the above information, we can find analyse things such as, who hold the ball the most. We can even find out which attacking formation yields the most points. Combining this with team and player statistics from the NBA, we can learn some powerful things about a team’s performance like who is the best player to give the ball to depending on an offensive formation a chosen team are in.

Taking this further, we can add Machine Learning into the mix, allowing us to suggest offensive or defensive positioning depending on the players on the court, for example. As the NBA is full of very rich data, the possibility and scope can be very expansive. All this information is useful to coaches that may realise this information.

## 1.2 Motivation

Surprising to most, coaches tend to not have the most say in how a team plays as a whole. This is could be due to players having more power as they may be keeping the franchise afloat, or get on with their team more than the coaches do. Consequently, this results in coaches not getting to implement the system they want to.

Coaching brings many challenges for players as well as the coaches themselves. Many are judged on their in-game decision making and utilisation of players. Predominantly in the NBA, teams even lose their head coaches after intense derby matches against rival teams after a rough performance.

Furthermore, the pressure to perform well for fans, the franchise board and player agents are extremely high. Therefore, my motivation was to help those teams that are struggling to succeed due to one or more of the above issues. The outcome of this project should give coaches the ability to make more informed decisions in and out-of-game. They will be able to quickly identify their key players, the best tactics for their squad and hopefully have a more effective team.

I have been aware of how much stats American sports and the NBA kept track of, and thus had the original idea to have a video feed analyse a game. However, due to time constraints and the discovery of the use of Spatiotemporal pattern recognition in sports (2)(3), I was then on the hunt for data. I have managed to contact SportVU, the company in charge of tracking the aforementioned information for the NBA.

# Aims and objectives

## 2.1 Aims

I plan to create something that can bridge the gap of imperfect knowledge across the NBA. Coaches will be able to tell how effective their team play with their current play style.

To achieve this, I will have to explore the relationships between player movement and player statistics. I have created several objectives to accomplish throughout the project. Upon competition of each modularised task, I aim to build on an expansive project that will ultimately help identify complex gameplay such as the type of defensive positioning the team are playing with and how a team can overcome it.

## 2.2 Objectives

Below is a list of objectives in my project:

* Player statistics will be extracted from the official NBA website, holding statistical information on every player in the NBA over a specific period of time i.e 2014-16.
* Movement data will be collected from the official source that captures this information every game, SportVU over the same period of time
* The data from both sets will be synthesised to reveal correlations between player ability and player and ball mobility.
* Construct an algorithm that can visualise spatiotemporal data, for feedback to users
* Devise a Machine Learning Algorithm that can classify moments in a chosen game, whether a player is shooting, dribbling, passing through the spatiotemporal data that I’ve collected.
* Produce further models to classify opponent defensive formation
* Create a model to decide what form of attack strategy to use in-game dependant on many variables such as players on the court, positioning and so on.
* Create an algorithm to analyse the gameplay in useful ways for a coach, such as how much a how effective a player is with how long they’ve held the ball and completed stats within that time.
* Measure success with coaching suggestions. How good are the suggestions compared to coach’s intuition?

# Methods

As this project is very expandable, there are plenty of steps towards getting the base of the project working well, before adding on the extra bits.

Behind every good model is a good data set. For this project to be successful, the data that is collected is key. I would need information such as x and y location of the ball, players on each team at any time, what team they’re on, their full name, their role on the team and so on. This information is provided for artificial intelligence analysis which has already been split into training and testing by SportVU (4) and has 2 datasets available for different things. The data set holds over 8 million events of basketball matches over 2002/3 to 2016/17 seasons.

Once the data is collected and interpreted, I would need to visualise the data, to firstly understand myself what is going on. Although this adds developer bias, I do have a background in basketball and believe I can distinguish what is happening accurately. I could also cross reference with watching the actual match if I’m unsure. Later in this report, I will outline how I plan to use the coach’s feedback mitigate this bias and measure the success and accuracy of the software.

After I have a visualisation of the data, I would need to create a classification model that will primarily pick out aspects of basketball and detect them from the data. For example, if the trajectory of the ball travels from one person to another and possession changes from one teammate to another, it will be considered as a pass to that player. This can be cross-referenced with player stats for that game for better accuracy. I will be programming in Python with some help from anaconda libraries to aid with classification such as Pandas and Sklearn. I will also use NumPy to help with data analysis and allow me to plot data to make graphs and charts.

Next up would be to create an algorithm that will take the classifications of the data and make sense of them. I would do this by taking into account the offensive and defensive positioning of the players. In basketball, there are many different variations on a play and therefore and depending on the root of that variation, should a counter play be involved. For example, against a 2-3 zone defence, a typical response would to either stretch the defence out of position by various passes or involve your shooters more as they are more likely to be left open. In order to construct something like this, I would use a Hidden Markov Model to help make decisions based on the game situation. (5)

To expand further on the project, I would then go so far as to take into consideration the available players on the court. I would consider that players all-time stats, season stats and in-game stats. I would then use that data to then suggest who would be good for a certain time a game. This would require me to use either NBAstats.com data to get all the player information up until 2016. For the 2017 data, I would have to request it or use a web crawler.

With so much rich data provided by the NBA, the amount that can be analysed, predicted and implied is virtually limitless. Therefore, I plan to consider the project fully completed after the above method to give this project a suitable time scope. However, that does not mean that it will end there, as there will be room to improve the algorithms and as well as build on top of the project.

# 4. Project plan

For this project, I have decided to follow the systems life cycle. This means that my development plan is separated into at least 5 steps to ensure that I take the correct approach in development. This is the model that companies use to ensure that their solutions are made correctly for their clients without missing important pieces of information.

### 4.1 Requirement Analysis and Feasibility

For the first stage, I need to discern whether if my product is unique, and/or required by the user. SportVU, as well as some independent developers, are the only people that have started on projects similar for purposes of an article or personal interest. Therefore, it is evident that something like this project either doesn’t exist or is currently in development.

Feasibility looks at if the project is affordable and possible within given time constraints. This project's feasibility will not require any money to be spent making it entirely and therefore is affordable. Regarding time, I have decided to take modularise the project, meaning that I will complete simple portions of my project and have a minimal viable product, with all the basic features. I will then use that to build up to what I hope to end up with mentioned in the section above.

### 4.2 Planning

To plan effectively, I decided to write down all the tasks that make up the completed project. I then picked the most important tasks that make up the core of the programme and made a path between them on what I would work on and in what order. After, I assigned those core tasks deadlines, for a minimal viable product.

With all of this planning, I was able to convert this process into a Gantt chart courtesy of Zoho.eu. I made these core tasks, as well as the other deadlines such as the write-ups, deliverables and prototypes; milestones. Everything else became a small task to be completed.

This really helped me break down my project, so I could see what required the most attention and what required the least. (See Appendix A)

The critical flow of my project is as follows:

**>** Retrieve Data

**>** Analyse/Visualise

**>** Classify one game aspect (e.g shooting)

**>** Run analysis for that single aspect (e.g When a good/bad shot occurs depending on play)

**>** Classify if chosen aspect should be happening at any chosen point in game

Following this work flow will result in having the most minimalistic version of my project. From here I can expand the project to detect for other more complex aspects of the sport, such as screening. In which the cycle of development will restart after classifying another game aspect.

### 4.3 Design

Designing the interface was very simple, as this software doesn’t need to have the most captivating user interface or experience. You would be able to select a game to visualise with the analysis and options on the left. The reason for the simple looking UI is this isn’t planned to become a commercial product. If one day in the future I want to make it commercial and perhaps a universal coaching tool, then the user interface and experience will be revised. (See Appendix B)

### 4.4 Develop

For development, I quickly decided to use Python and Anaconda. This is because Python is very easy to write and test small tasks with. Anaconda is very powerful for machine learning as it allows for complex libraries to be included which makes creating models far easier and quicker.

After some research and meetings with my supervisor, I found that using NumPy, Pandas and Sklearn will get me closer to my goal.

As mentioned in my critical path, I need to first retrieve data and analyse it. To do so, I would use Pandas and NumPy to quickly complete these tasks.

Sklearn will handle most of the classification tasks as it’s a library that holds different machine learning algorithms that I can call upon on datasets. I can also tweak them to work better for what I need to do. All graphing and analysis will be left up to Sklearn and NumPy.

### 4.5 Testing

The testing stage of this project is definitely the most important part as it is how I can truly validate all of the statistics, outputs and bias. As a result, I have decided to spend the most time on this portion of project. The testing phase will be split into two main parts: Software Testing and User Testing.

#### 4.5.1 Software testing

Here is where I will test the software in two different ways.

White box testing looks at the inner workings of the project. I will be testing that my code works the way it’s intended. This could be through making unit tests as well as running through the code with a debugger. (the IDE I’m using, PyCharm, has a very powerful debugging tool that works well for white box testing) The aim of this is to make sure the code I have written firstly works and does what it was written to do. I will test this during the development stage naturally, however it is important to do it after as well to make sure your outputs are actually from the correct lines of code.

Black box testing looks at the software as a whole. This does not involve looking at the code at all, only at the output. As it would be hard to objectively say that the program is outputting the correct numbers, I have decided to write some test statements. If my program completes all or most of the statements that I know that I have successfully created a program that does what I set out to do. Below is an example of what the black box testing truth table would look like.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Number | Summary | Steps | Actual Result | Expected Result |
| 1 | Select a team to start a visualisation | >Start program  >Choose a season  >Choose a team  >Select a match | Works as expected | A visualisation is displayed from that match from tip off. | |

#### 4.5.2 User Testing

The user testing will just as important as it will truly help me understand how viable this project will be. There are several aspects of this that need to be tested extensively, therefore, I will have to ask for the participation of many coaches, where I will have them try certain portions of the software.

For example, I can make some look at a defensive position from previous games and get them to suggest what sort of offensive plays they would run. I would then compare the results to my software. As I am currently unsure of how many coaches’ input I can obtain, I will not have a limit or a set number of participants. This, however, will be ideal as there are plenty of things I would like to test, and it would be generally better to get an even spread across all areas.

I will make sure that all consenting participants fill out a form beforehand as well as a questionnaire after to collect data regarding the accuracy of the software. The questions will be similar to the test cases in the black box testing. This difference would that the answers will be coming from and matched with those very familiar with the game.

Unfortunately, I currently don’t think I would be able to get end-user feedback from any coach’s from within the NBA as that would improve the accuracy of my project further as they are the intended target.

# Progress to date

To date, I have been making small steps forward towards the final project. I have been following the software life cycle as well as my critical path to get my initial product.

So far I have a written plan detailing where and how I would complete the steps above as well as some of the academic papers and blog posts I have to read regarding spatiotemporal pattern recognition in sports (2)(3)(5)(6).

I have also made a Gantt chart to give myself deadlines for each thing in the critical path. This will help me organise my time before bigger deadlines come such as the final report draft or the delivery date for the project.

I have also started playing around with a sample dataset I found on GitHub(8). This data is in exactly the same format as the data from SportVU’s dataset just smaller sitting at 30,000 entries. This is stored in a JSON format, which I convert into a dictionary into Python and then into a Panda Dataframe. This will then allow me to start picking out information to continue down my critical path with. Once this works fine, I will introduce the larger dataset that has over 8 million entries. (See Appendix C)

I am currently following a few blogs where people have visualised sports data in other languages or interpreted this data as there are a lot of nested dictionaries and no labels. This is where I am at currently and have set myself to be done after next week.

# Planned work

The next step after visualising the data would be to personally interpret it. This is to figure out how to classify aspects of the game as well as to see what would be the easiest to figure out. I will keep choosing the easiest to make sure I have time with the rest of the project. Once I can see how to categorise certain movement, I can begin to define it and write some code to allow it to be detected during the run of a game.

I have realised some small challenges with classifying which is why I want to personally see the visualisation. As the amount of information is limited, a ball travelling towards the basket could be a shot, but it could also be a player near the basket. It could even be a shot that has missed or pass that went to the opponent. All these cases can really affect the output and would need to be as accurate as possible. This is also where I would start introducing Sklearn into my code.

# Conclusion

The NBA is full of rich, player by player statistical information as well as real-time spatiotemporal data. Combining those together gives you rich information to be analysed and utilised to aid coaches by giving a second opinion in those tough decisions during a close, heated game.

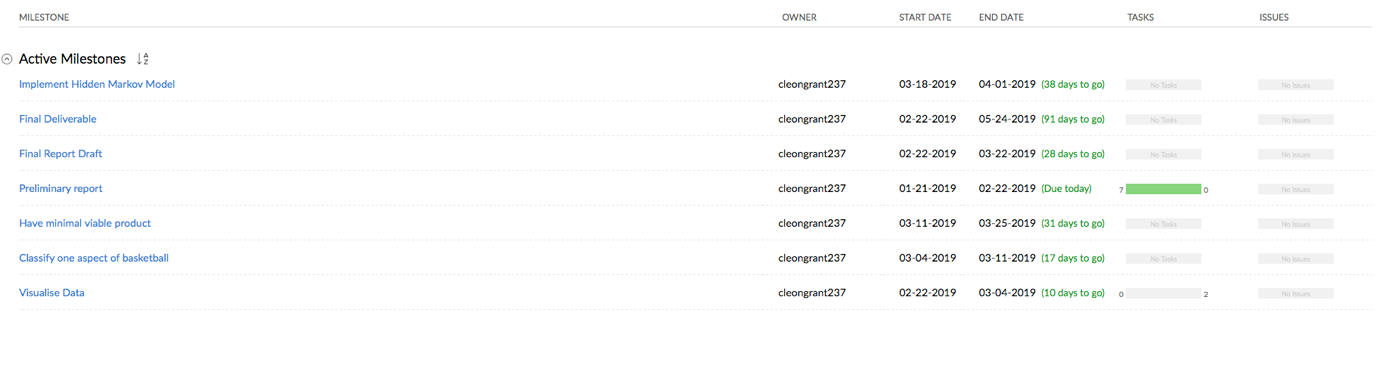
Correlating this information time signatures with where things are on the court, we can know; who has the ball, what players are on the court, and at what time events in the game are happening. It could be a minor event such as dribbling the ball from the baseline to half-court, or even someone slam dunking the ball.

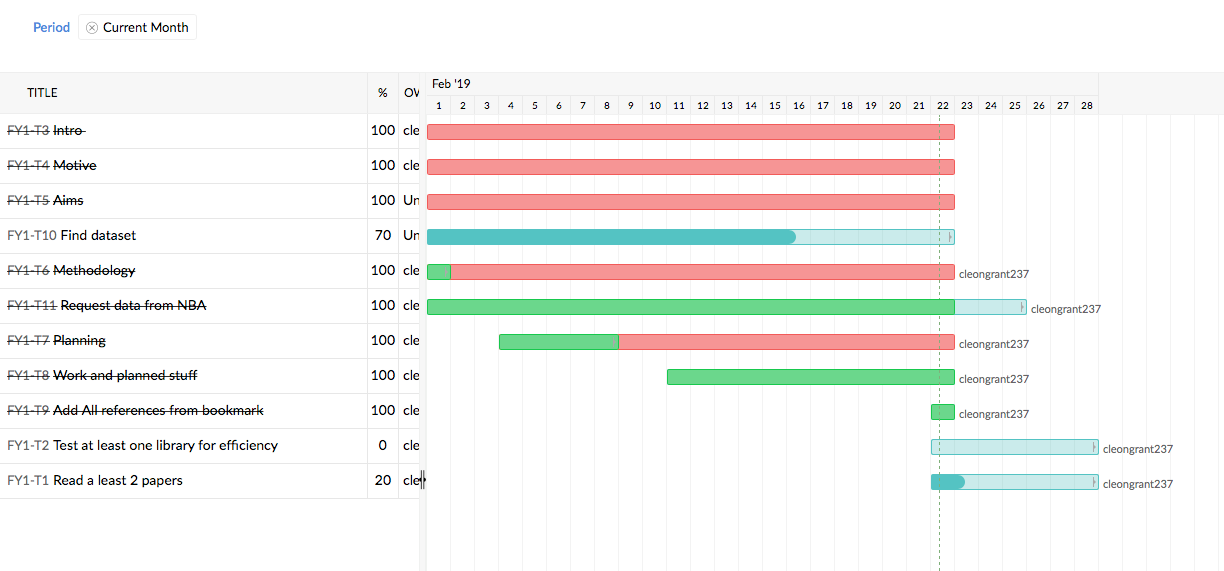
So far I have a detailed plan that will give me a minimal product that I can build upon through iterating the critical path. I will be using Python and Anaconda libraries to complete this project and use coaches to validate the outputs.

# Appendices

Appendix A

This is a snapshot of my gantt chart and milestones. I have made dates that I can try to stick by which will help me organise my project





Appendix B

This area will let the user switch between different analytics or game aspects

This is a very early version of how this will be designed.

Switch team

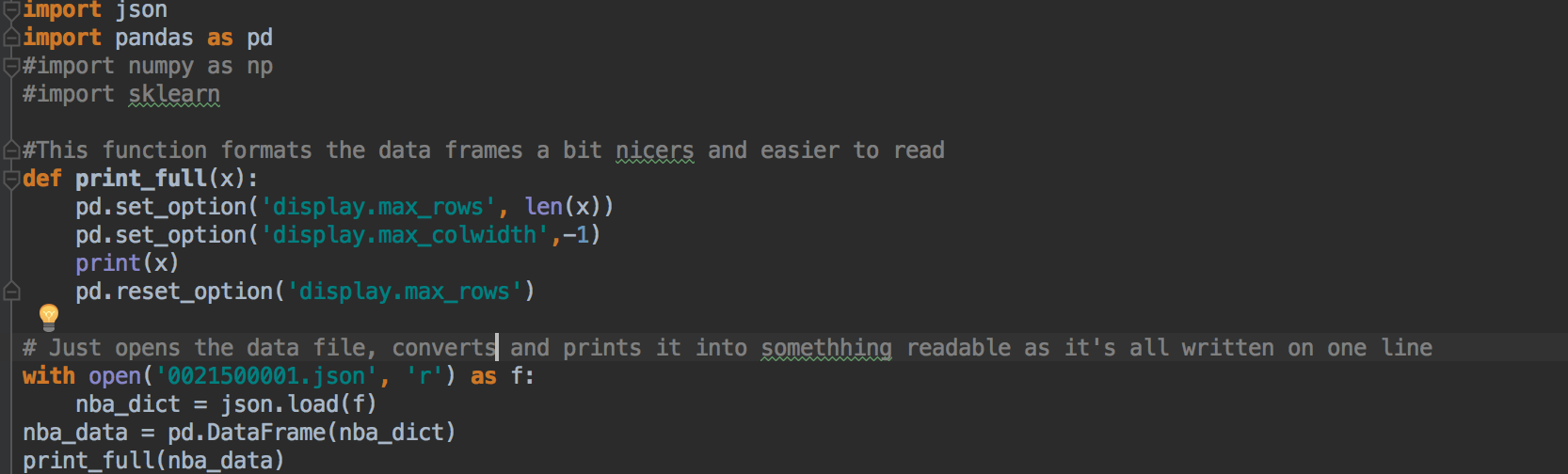
This is where the visualisation of the basketball will be shown. It can be paused, and resumed, rewound and fast forwarded.

This is where the outputs would be displayed for each player, depending on the option chosen at the top of this section

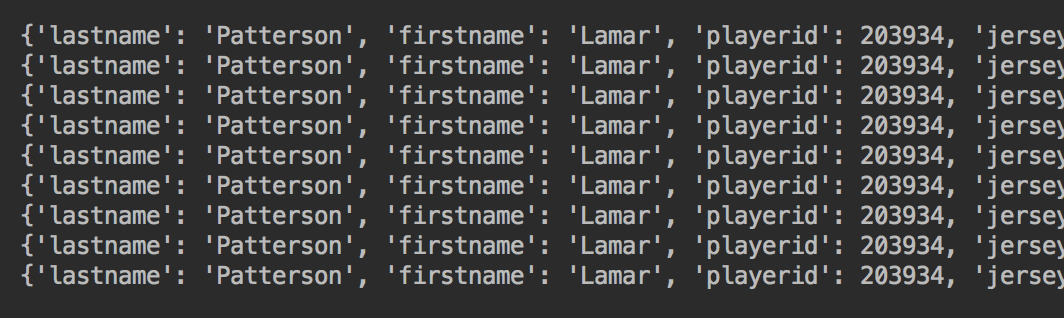
This is where the information about a player would display, one side would be information about each player on the team

Appendix C

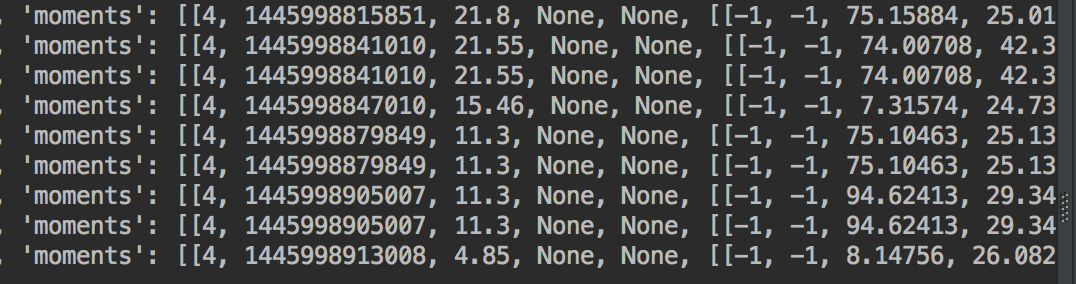
Below is a snippet of code that I am using to display the data.



This is some of the output. It is split up into events and moments. Moments hold the multiple snapshots that SportsVU track



This is the movement data. As there are no labels, I would need to figure out what each position in the dictionary and arrays do.



# References

1. NBA News, Highlights and Videos | Sky Sports [Internet]. Sky Sports. 2019 [cited 22 February 2019]. Available from: http://www.nba.com/

2. Chen C, Liu T, Wang Y, Chu H, Tang N, Lio H. Spatio-Temporal Learning of Basketball Offensive Strategies [Internet]. Taipei: Inst. of Information Science Academia Sinica; 2019 [cited 22 February 2019]. Available from: https://people.cs.nctu.edu.tw/~yushuen/data/BasketballTactic15.pdf

3. Gudmundsson J, Horton M. Spatio-Temporal Analysis of Team Sports. ACM Computing Surveys. 2017;50(2):1-34 Available from: http://materials.dagstuhl.de/files/17/17171/17171.JoachimGudmundsson.Slides.pdf.

4. SportVU-Basketball [Internet]. STATS. 2019 [cited 22 February 2019]. Available from: https://www.stats.com/sportvu-basketball/

5. Bashir F, Khokhar A, Schonfeld D. Object Trajectory-Based Activity Classification and Recognition Using Hidden Markov Models. IEEE Transactions on Image Processing [Internet]. 2007 [cited 22 February 2019];16(7):1912-1919. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.1720&rep=rep1&type=pdf

6. Keshri S, Oh M, Zhang S, Iyengar G. Automatic Event Detection in Basketball. Presentation presented at; 2017; IEOR Department, Columbia University. Available from: http://nessis.org/nessis17/Keshri.pdf

7. [Internet]. Building NBA Defenses using the convex hull. 2015 [cited 22 February 2019]. Available from: https://squared2020.com/2015/11/08/building-nba-defenses-using-the-convex-hull/

8. Shah R. R Scripts for analyzing NBA sportvu motion data. 2019. [cited 22 February 2019] Available from: https://github.com/rajshah4/NBA\_SportVu