CS171 Final Project - Process Book

Michael Woo, George Zhang, Jeffrey Zhao

http://gzhang01.github.io/CS171-Final-Project/

Overview

Over the past decade, sports data analytics has grown substantially. Driven by the development of innovative ways of tracking and collecting data as well as new methods of data storage and visualization, the field has changed the way players, coaches, and fans view sports. The LA Times released an article last month upon Kobe Bryant's retirement plotting every single shot he ever took in his career[1]. Visualizations such as these can provide fans with not only a look into the life and career of a particular athlete, but larger trends as well in a way that would not have been possible before the rise of sports analytics.

We wanted to look at some of these larger trends. In particular, our goal was to determine what makes a basketball team successful in today's National Basketball League. This question was motivated by the ongoing success of certain teams, such as the Golden State Warriors[2], and certain players, such as Stephen Curry. We wanted to determine what specifically allows them to perform at such a high level year after year.

To answer this question, we took a three-pronged approach. Our first task was to determine some overall trends within the NBA. We wanted to know the strategies teams are using in the NBA today and how they have changed over the past decade or so, with the rationale being that successful strategies are more likely to become adopted league-wide. Our second task would be to examine whether or not successful teams employed these strategies. For this part, we took the Golden State Warriors as a case study. Our third task was to determine whether successful players also utilized such strategies, with Stephen Curry being our case study.

Visualization 1 - The Rise of the Three Point Shot

The goal of the first visualization is to determine overall league-wide trends. We believed that successful strategies were more likely to be adopted by teams. Using data from Basketball Reference[3], we obtained a data set that included league averages in many statistics, such as shooting numbers and percentages; number of blocks, steals, and assists; and the pace of the game, among others. We immediately decided some stats were not relevant[4] while others were a bit too complicated[5]. We wanted a statistic that was easy to understand but also could vary largely between teams.

The number of three point shots made and attempted fit our criteria perfectly. We noticed that in the 1980s, very few three point shots were being attempted. There were on average only 2-3 per game. In the 2015-16 season, however, an average of 24.1 three point shots were attempted per game. We were interested in this large change in three point shots attempted, as

---

1 http://graphics.latimes.com/kobe-every-shot-ever/

2 The Golden State Warriors, incidentally, had a record-breaking season in 2015-16. Analyzing their success was a significant motivator in our project choice.

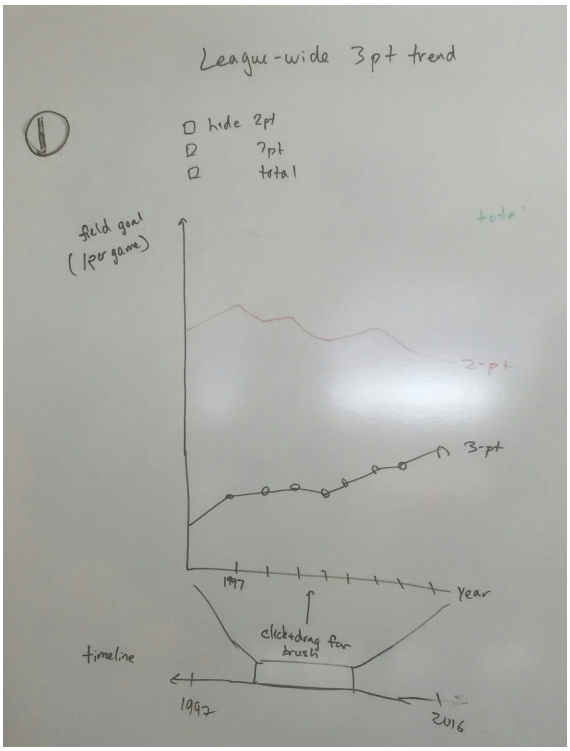
3 http://www.basketball-reference.com/leagues/NBA_stats.html

4 The increase in average height over the past 60 years and the decrease in average age over the past 20, while interesting, probably was not contributing to the success of certain teams over others.

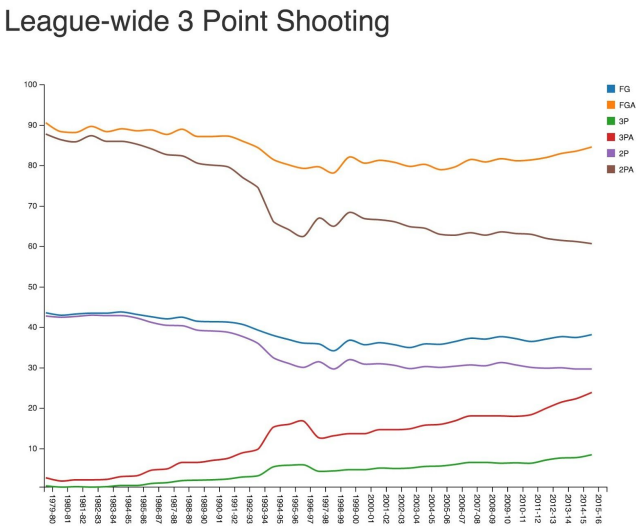
5 e.g. offensive rating

it suggested a shift in the overall strategy of the game. We also noticed that despite an average of 24.1 shots attempted per game, only 8.5 three point shots were made per game on average. This low shooting percentage from beyond the three point line made us realize that success in three point shots could very well separate good and bad teams.

Thus our first visualization attempted to show the trends that we found. Our first sketch is shown to the right. We initially only looked at the number of shots attempted. We believed that the total number of shots attempted (green line) probably has not changed significantly over time[6]. Thus to explain the increase in the number of three point shots attempted (black line), we expected the number of two point shots attempted (red line) to have decreased. To show this data, we decided a simple line chart would be ideal, since we are looking at data over time. We also wanted to be able to toggle the visibility of certain lines. For example, if a user clicked the “hide total” button at the top, we would hide the total shots line. This would allow us to scale the y-axis to see the other lines better. The original plan also included a timeline at the bottom that allowed the user to select certain years to look at. We did not quite know the scale of the x-axis at the time, and so we thought that an option to zoom in on certain periods of time would be useful.



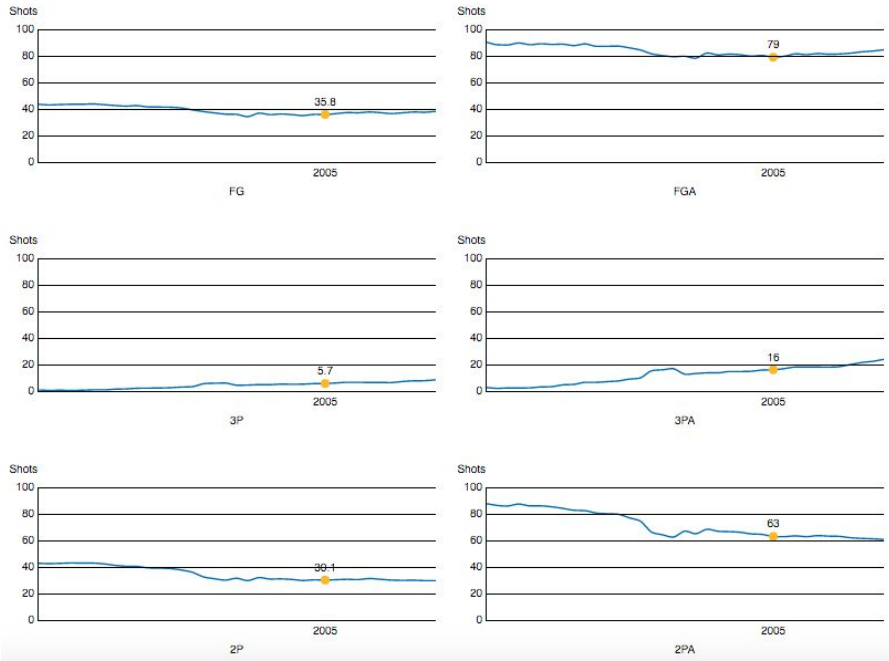
Our initial visualization of the data looked like the graph to the right. Here we graphed total number of shots taken (orange) and made (blue), number of two point shots taken (brown) and made (purple), and number of three point shots taken (red) and made (green). We were relatively happy with the results, as we saw that neither total shots taken or made changed too much, while the number of three point shots increased as the number of two point shots decreased[7]. However, we were not particularly pleased with the way the data is presented in this visualization. The most



[6] After all, the length of a basketball game has not changed during this span.

[7] We did notice the slight bump in the three point trends from 1994-1997. Turns out that during these years, the NBA experimented with moving the three point line closer into the basket, which naturally would cause players to attempt three point shots more.

significant issue is that it looks like there are three sets of two lines, but the upper two sets do not go together. Thus we wanted to separate the graphs and pair them up so that the relationship between the lines make more sense.



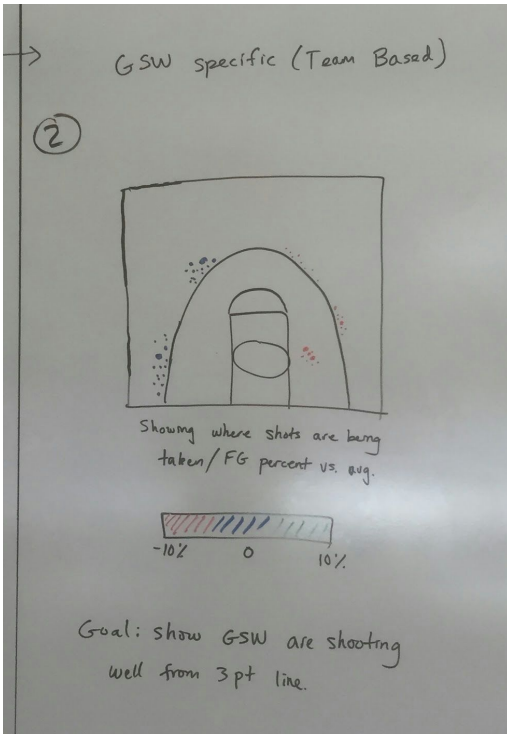
Season	1979-80	2015-16	
Field Goals	43.6	38.2	-14.14%
Field Goals Attempted	90.6	84.6	-7.09%
3-Pt Field Goals	0.8	8.5	962.50%
3-Pt Field Goals Attempted	2.8	23.9	753.57%
2-Pt Field Goals	42.8	29.7	-44.11%
2-Pt Field Goals Attempted	87.8	60.7	-44.65%

The above was the next iteration of our design. We put the shots made in graphs on the left and put the shots attempted in graphs on the right. As a result, we can compare shots made vs. attempted by looking across the row and different types of shots by looking down columns. A slightly earlier version did not have the horizontal lines. While the lines themselves may be considered “chart junk” according to Tufte’s principles, we decided that the lines help orient the user when comparing between rows. Mousing over any graph provides the specific year and value (shots / game) in all graphs, so it is very easy to compare these values across years as well. Finally, we have a comparison table to the right of the graphs showing values for two years and the percentage change between the years. Initially, they are set to the beginning and end of the data set (1979-80 and 2015-16 season respectively), but clicking on the graphs replaces the years in alternating fashion. Thus it is possible to see the percentage change between any two given seasons. We feel that this representation presented all the data we wanted to convey: namely change in trends over time (both from the graphs and the change table) and the data given a year (hover info).

Visualization 2 - Player Shot Chart

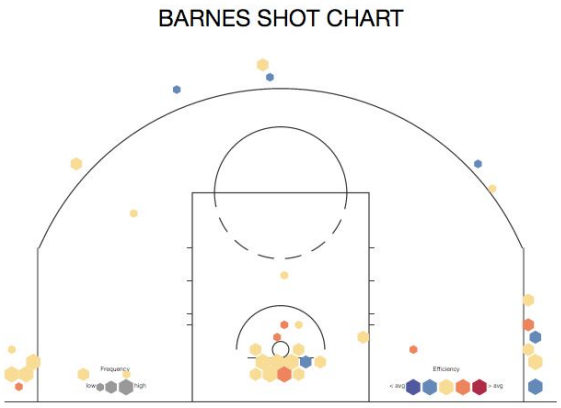
After determining that three point shots could be a factor in determining the success of a team, we wanted to look at where on the court players were taking their shots. We took data for the Golden State Warriors from NBA.com. This data provided shot locations, which we could then plot on the court. Our goal with this visualization was to show that a successful team like the

Golden State Warriors both shoots from beyond the three point line often and is very efficient from beyond the three point line.



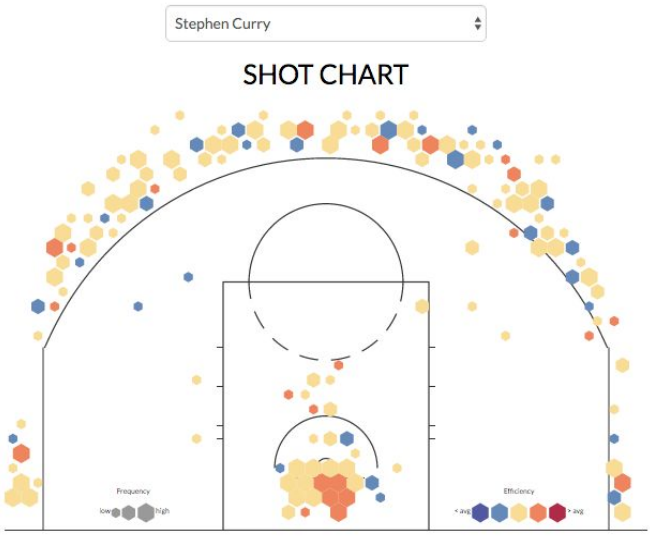
Our first sketch looked like the one to the right. We were initially only considering showing team data. We would plot the aggregate shot data on a single graph, where position encodes the relative position on the court the shot was taken from, color encodes the shooting efficiency as compared to the league mean, and size of dot encodes the number of shots taken from that location. We were hoping to see that the Golden State Warriors had many dots in the green (high shooting efficiency) outside the three point line, which would suggest that they were taking and making many three point shots.

When we pitched this idea to our section friends, they suggested that we make the visualization more interactive. As a result, we produced the sketch shown below. The first major change we made was to aggregate location data into hexagonal shapes. We decided that this would allow us to focus less on plotting every single shot and more on bigger picture trends. The second thing we added was a drop down list allowing the user to select different players on the Warriors. Since different players have different playing styles, we decided that the ability to see shot charts for different players would be useful in determining who tends to shoot from where. Breaking down by player also allowed the data from this visualization to flow into our third one a bit more nicely, which we will discuss below. We also decided to add a tooltip which would appear upon hovering over one of the hexagonal regions. The tooltip would then show data about the region, such as shooting efficiency for the player and average league shooting efficiency. Finally, we decided to change the color scheme to that commonly utilized by heat maps, as we believed the coloring would be more intuitive for the user.



As we collected our data, we needed a quick and easy way of visualizing it. We decided to see if there were any existing tools that would allow us to easily plot a basketball court structure in d3. We ended up finding an open source library[8] that allowed us to do just this. In addition, we were able to plot some preliminary points to see how things looked. The result was the plot to the right. As we can see, the hexagonal structure and color coding[9] was relatively simple to create, and we were relatively happy with the result we got. Ultimately, we used this to explore the data as well as produce the visualizations.

Now that we could see the data in a useful way, we worked on collecting more. Unfortunately, this process proved difficult and time-consuming. NBA.com does not allow for easy access to its data in convenient formats, so we could not simply export a CSV like we could with Basketball Reference. NBA.com does have an API, which we used to scrape the data[10]. However, scraping player shot location data required us to input the player ID every time, so data collection for this visualization turned out to be very time-consuming. We collected all the data we needed from the Golden State Warriors, but we had trouble getting league average data based on shot location. Thus instead of comparing shooting efficiency to league average, we decided to compare it to the player's own average. We decided that this was not a huge trade-off, as we can still see where players are shooting well from compared to other shots they made, which is also valuable data.



The final iteration is shown to the right. We have an option to select players from a dropdown at the top. Unfortunately, the

[8] https://github.com/virajsanghvi/d3.basketball-shot-chart
[9] Note: colors represent standard deviations from the mean
[10] Sample scraping URL: http://stats.nba.com/stats/shotchartdetail?Period=0&VsConference=&LeagueID=00&LastNGames=0&TeamID=0&Position=&Location=&Outcome=&ContextMeasure=FGA&DateFrom=&StartPeriod=&DateTo=&OpponentTeamID=0&ContextFilter=&RangeType=&Season=2015-16&AheadBehind=&PlayerID=2760&EndRange=&VsDivision=&PointDiff=&RookieYear=&GameSegment=&Month=0&ClutchTime=&StartRange=&EndPeriod=&SeasonType=Regular+Season&SeasonSegment=&GameID=

tooltip plan did not work out. The library we used took the data in formatted it, and we found that we could not access the data that was bound to each of the hexagons. However, we decided that since we did not have league average data anyway, a tooltip would not actually be able to provide much more data that what was already shown visually, so we decided to stick with what we had.

Ultimately, we were satisfied by the result we found. Players such as Stephen Curry and Klay Thompson had lots of shots from beyond the three point line, and they were generally colored as yellow and above. Considering that Curry and Thompson are relatively good players compared to the NBA average, average shooting efficiency and above for them suggested a good shooting efficiency in general from beyond the three. Thus our visualization told the story we were hoping to tell.

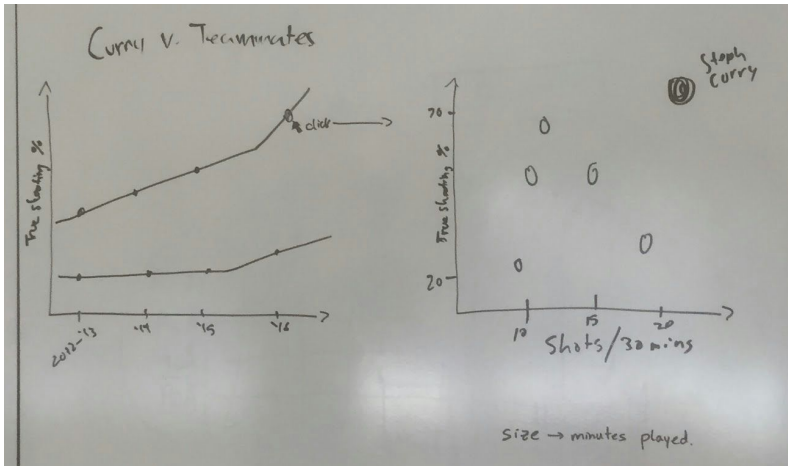
Visualization 3 - True Shooting Percentage

In this visualization, we wanted to look at another measure of how well a player is shooting: the true shooting percentage. The true shooting percentage is calculated as follows:

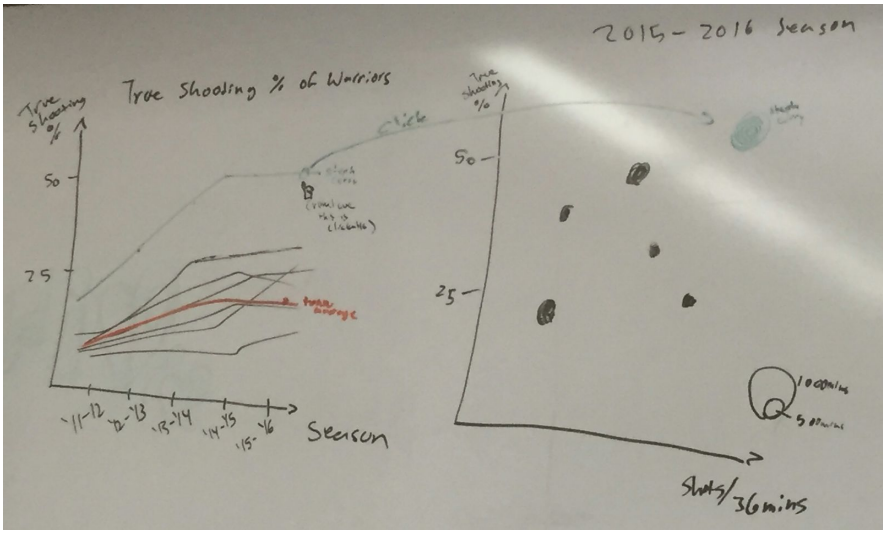
$$TSP = PTS / (2 * TSA)$$
$$TSA = FGA + 0.44 * FTA$$

Thus the true shooting percentage is a function of PTS (points) and TSA (true shooting attempts), with TSA a function of FGA (field goals attempted) and FTA (free throws attempted). The general intuition is that three-point shots are weighted more in the true shooting percentage, since three-point shots net more points, whereas free throws are weighted less, both in terms of points and in terms of shots attempted. Thus players who have a high true shooting percentage tend to be better players both in terms of shooting overall, but especially in terms of shooting three point field goals.

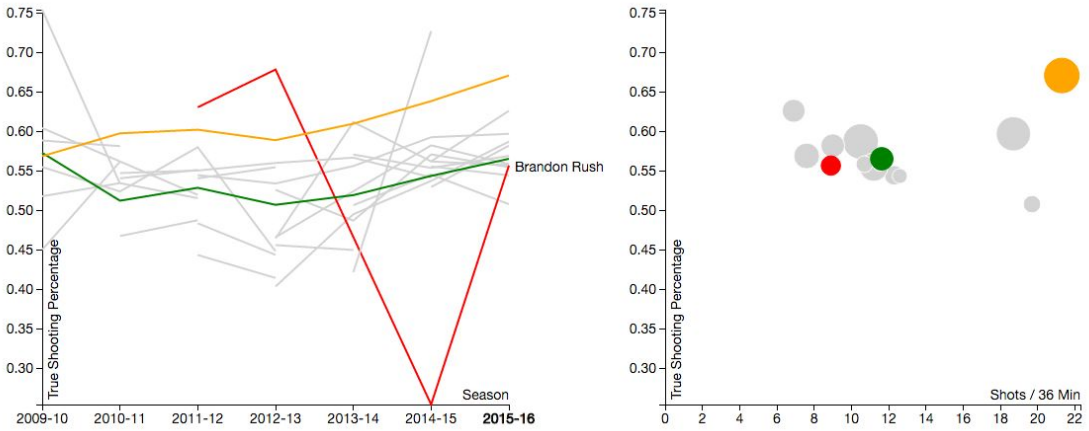
Given this, we expected players who were successful in the league both overall and in terms of three point shots, such as Stephen Curry, to have a higher true shooting percentage than other players. Thus we decided to take a look at this statistic in more detail. The first iteration of our visualization is shown below. We wanted to have a line chart showing Curry's TSP over his seasons and the TSP of his teammates over the same time span. This would allow us to visualize how well Curry does compared to other good players on his team. In addition, we wanted a linked scatter plot that



would plot TSP vs. shots / 36 minutes for Curry and his teammates when the user selects a given year. Shots / 36 min was a stat we cared about, since shooting well given lots of shots is more significant than shooting well in only a handful of shots. In addition, we had the scatter plot encode total minutes played in the size of the circle, again because playing well for longer is more significant than playing well for a few minutes. We felt that the combination of these two plots would nicely compare players based on their TSP and overall importance to the team.



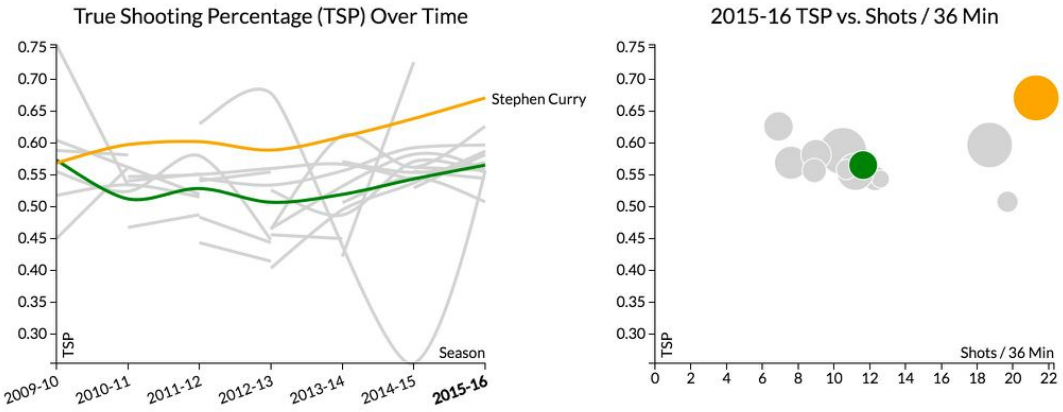
Feedback from section prompted us to make the origin of the dots from the scatter plot more obvious. One of the concerns that arose was that the scatter plot data seemed disconnected from the line graph data. As a result, we decided to plot more player's data in the line graph, but use a light gray color for most of it. We would have Curry's data be one color and the average be another so that they both pop out when compared to the other data. Clicking a year now would prompt a transition from one graph to the other, which would make the origin of the data clear. We kept many of the encodings we had before because we still wanted to convey all the information. However, we decided that within the context, we would explain the terms we use better, since there was some confusion from test users who did not have much knowledge about basketball.



The first iteration of the visualization looked like the above. The player data was taken from Basketball Reference[11]. We had lots of data from various seasons, so the first task was to

[11] e.g. http://www.basketball-reference.com/players/c/curryst01.html

extract only the data we needed. Getting the first implementation involved merely adding lines and points to the svg using d3. There were no significant transitions; the data merely appeared on the scatterplot. However, we did implement hovering in this iteration. Hovering over a line or circle would link that player across the graphs, allowing the user to see the trend over the years, the data for that specific year, and the player name.



2015-16 Season

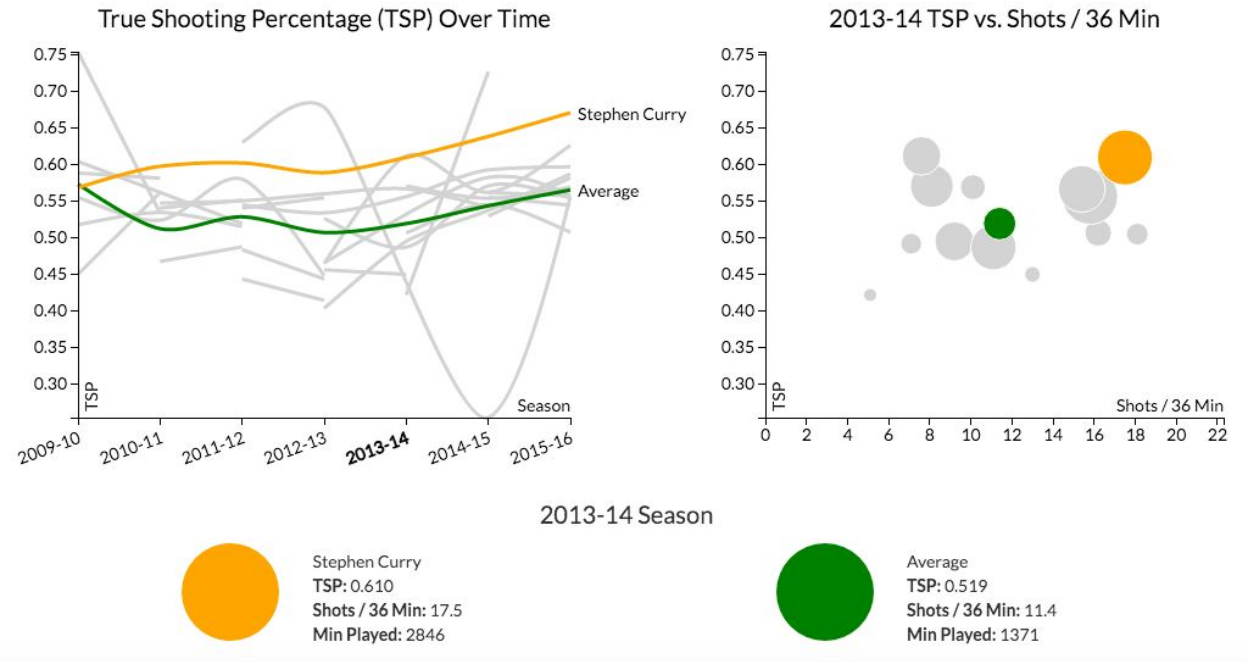
Name	Stephen Curry	Average
TSP	0.671	0.565
Shots / 36 Min	21.3	11.6
Min Played	2700	1399

The two big features of the second iteration (shown above) were more information and fancier transitions. We decided that since the goal was to compare Curry vs. his teammates, we would create a table listing some stats. Thus the chart lists TSP, shots / 36 minutes, and minutes played for Curry and the average player on the Golden State Warriors. Clicking on a specific circle changed the player being compared to Curry, so the rightmost column can show different data points. We also implemented a more intuitive transition between years. Selecting a different year from the x-axis of the line graph will cause the points corresponding to data in that year to flow from their places on the line graph to their correct place and size on the scatter plot. While this occurs, the old data fades and ultimately disappears once the new data is in the right place[12]. The final little change we made was to make the lines into curves. We decided this had a better aesthetic appeal.

During one of our final discussions, we decided that the table did not fit particularly well with the rest of the visualization. In addition, we wanted to make the connection between the scatterplot and the data in the table more clear. Thus we decided to take advantage of the fact that the dots are already color-coded to draw the connection (see below). We decided that by using similar colors like this, we could connect the scatter plot data with the detailed information below more effectively. In addition, we lose the borders of the table, making the information more compact

[12] This transition is hard to describe in words and harder to capture in a series of screenshots. Please check out the final version of our project to see the transition in action.

and maximizing our data-ink ratio. The use of whitespace also gives the visualization a cleaner look, and transitions between player comparisons when clicking circles on the scatterplot no longer shifted the whole table columns[13] but merely changed words. We felt that this design was much simpler and effective, and so our final iteration took on this form.



Conclusion

Over the course of the past several weeks, our project and goals have changed as we explored the data set. When we first started out, we did not have a specific story we were trying to tell with our visualizations[14]. We knew we wanted to work with basketball data, especially in light of the Warrior's historic season, but exactly what we were trying to show with the data was still unknown. As a result, our first sketch ideas were relatively vague and focused on looking at player data in various ways.

After presenting our proposal in studio, however, our TF Benjy noted that we talked about wanting to look at trends in the NBA, but did not have a visualization that looked at data over time. This prompted us to look for historic data, which immediately led to the conception of visualization 1. As described above, the analysis of historic data allowed us to find trends in three point shots, which led to the creation of the two other visualizations. Thus our final product largely arose from peer feedback and data exploration, which are two critical steps of the overall design process.

---

[13] We were using bootstrap tables in the previous version. While bootstrap does a great job with static content (e.g. good spacing between columns), dynamic updates caused the column sizes to change, so the word positions would shift slightly. The result was a "jumpy" sort of transition, which we did not like.

[14] See appendix A: Initial Project Plan

In terms of implementation, each of our group members headed one visualization. We all experienced roughly the same design process: data collection, initial data visualization, feedback and improvement, final product. The initial sketches were done together early on, and we took the opportunity to coordinate the between the visualizations so that we could provide a compelling story overall. We met weekly to discuss our progress and provide feedback for each other on what could be improved. Feedback from peers in studio[15] and from friends also helped improve our visualizations; for example, we tried to make interactions as intuitive as possible[16], largely in response to the fact that our beta-testers did not click on several interactive objects unless prompted.

Given this workflow, most of our work could be done independently, and indeed, most of the coding of the visualizations were done separately. However, we found that the iterations of working and discussing that we had were very useful. Our process gave us weekly deadlines to meet while also ensuring that the time we scheduled to meet would be efficient and productive.

Our final product is something that we are proud of and feel address the questions we initially set out to answer. Our overall goal was to determine what makes a team successful in today's NBA. By analyzing the overall trends over the past 40 years and looking at specific case studies in the past season, we feel we have an answer to this question -- namely success in three-point shooting. In addition, we believe that the visualizations we produced not only accurately portray the data we used to develop this answer but also allows the reader to follow the same process we used to arrive at the answer.

Of course, nothing is ever perfect, and given more time and data we would have loved to be able to explore this field more. We have merely found a correlation between three-point shots and success in the NBA; much more research would be needed to determine whether a causative relationship exists. Additional iterations of feedback / improvements would indubitably lead to more elegant and intuitive visualizations as well. However, given the time horizons we had, we are happy with our final product and believe that it conveys our story in a clear and engaging way.

[15] See appendix B: HE Report

[16] e.g. changing the cursor to a hand pointer / changing color of text upon hovering over clickable items

A. Initial Project Plan

Introduction
Data analytics in sports has grown rapidly over the past decade. Team data, such as win-loss records, and individual data have contributed to changes in the way teams play and construct their rosters. Our goal is to provide compelling visualizations for viewing this data with the objective of determining what makes a team successful. We will focus on basketball, specifically the Golden State Warriors, for our analysis.

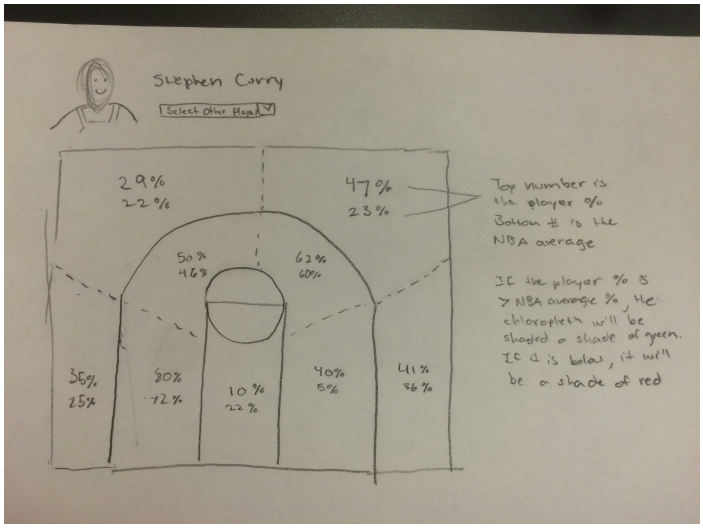
The Golden State Warriors in the 2015-2016 NBA Season are on track to break the all time single season win record and are arguably playing some of the best basketball in the last couple of decades. Our visualizations will use rich data sources to try and explain what exactly makes the Warriors different from other teams in the league and attempt to explain some factors of their success.

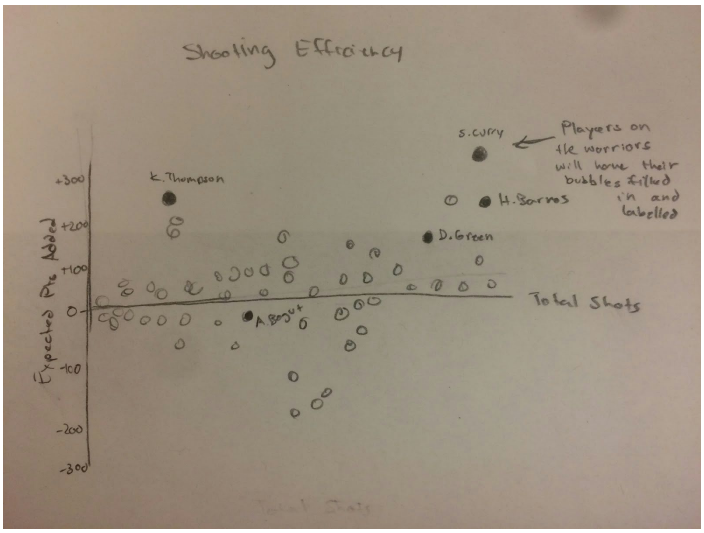
Data
Our data will largely come from stats.nba.com. An additional data source might be basketball-reference.com. The NBA has detailed stats on both players and teams. Since we will be focusing on the Golden State Warriors, we will mostly take advantage of their diverse player stats, though if we successfully complete the visualizations ahead of schedule, the team data available provides an opportunity to compare data between teams. A few particular aspects of the data we might look at include field goal percentage, points per game, rebounds per game, and PER (Player Efficiency Rating).

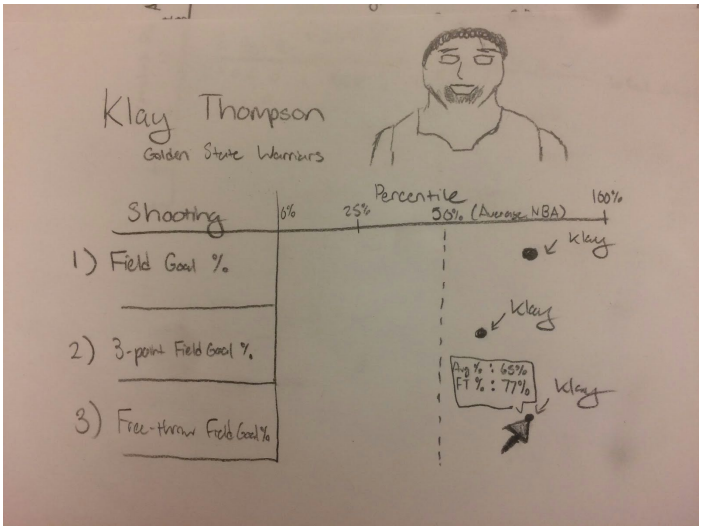
Goals
- What's successful in the NBA
- Find out what kind of style in the NBA is trending?
- Who are the star players in each position according to the data?
- How might coaches / teams respond to new styles of play?

Tasks
- Obtaining data / finding additional sources
- Coming up with an overall storyline for the website
- Writing content that contextualizes visualizations
- Design/CSS for the website outside of the visualizations
- Design/implementation of visualizations
- Putting website on production

Visualization Mockups







B. HE Report

Two Questions:

1. Where in the court do the Golden State Warriors take the most shots from?
2. Which players in the NBA have the highest true shooting percentage relative to shots/30 mins?

Lauren

1. To find where in the court the Golden State Warriors take the most shots from, I would look at the second visualization. In the second visualization, I would look for places on the court where there was a higher concentration of light blue dots. The high concentration of light blue dots would show me the place from where the Golden State Warriors take more than the average number of shots compared to other teams.
    a. However, I'm not sure if that will tell me where they take the *most* shots from. From what you've drafted out, it looks as if the Golden State Warriors take more three-pointer shots than the average NBA team, but I can't tell whether that's the place from where they take the most shots overall. Maybe they actually take more regular shots than three pointers, but I can't tell that from the information in the graph. If you truly want viewers to be able to see where the Golden State Warriors take the most shots from, not just where they take more shots from than the average team, this is third severity (unless I'm missing something). If you're fine with viewers merely seeing that the GSW take more three pointers than other teams, no need to change.
    b. Secondly, is there any ability to sort this graphic by player? It would be interesting to see how many of the Golden State Warriors' comparatively high three-pointer ratio comes solely from him. This is severity one.
    c. This visualization merely shows that the Golden State Warriors take a lot of three pointers, but it doesn't show how many of those shots are made. For all a non-basketball fan knows, perhaps the GSW don't make much of any of those shots. Adding in some dimension that shows how successful these shots are would be very interesting. This is severity one.
2. For this, I would look at the second scatterplot in the third visualization. The answer to this seems quite straightforward. Because the scatterplot's x-axis represents players' shots per 30 minutes and the y-axis depicts players' highest true shooting percentages, I would look for the upper right-most dot in the scatterplot. In this instance, that player appears to be Steph Curry.
    a. This graph is very straightforward, which I like, but I can't tell whether there is interactivity beyond a basic tooltip. Adding more interactivity would be a severity one.
    b. Secondly, the labels on the axes could be written in a way that would be more straightforward to non-avid basketball fans like myself. This is severity one.

Risham Dhillon

1. I would go to the second visualization of the shots taken from different parts of the basketball. I would look for areas on the court that are more blue because that would indicate that this was the area that the GSW were shooting the best from.
    a. One of the problems that I see here is that sometimes with color scales it might become a little bit difficult to be easily able to judge relative quantities (talking about the field goal percentage you include here). This is severity 1.
    b. It might be interesting to make this visualization more interactive maybe (maybe you can filter by team member and see where that team member takes his best shots from, etc etc). This seems to be severity 2.
    c. Also, it might be helpful to have like text labels on the basketball field to help someone understand. I think this would be severity 1.
2. Steph Curry has the highest true shooting percentage it seems. To do this, I went to the third visualization.
    a. I don't really understand what the first graph in part 3 is trying to do -- maybe it's because I don't understand basketball terminology that well. Some explainer text would help. This is severity 1.
3. I was just thinking that it might be useful to have a javascript tutorial that might go through your website or give a quick overview of different basketball terminology (severity 2).
4. Another thing that might be useful would be some main points or story line that is clearly labeled that tells you what the main takeaways you want to be are (severity 2).

Jocelyn

1. I would look at the second vis and look for areas with highest density of dots -- since the dots represent the shots, the area with highest density of dots would be the area where the largest number of attempts happen. If I want to take success rate into account, I would then look at the color of the dots.
    a. If you can explain basketball terms such as shooting average that would be really helpful!
    b. It would also be cool to see changes from year to year, if you have data from past years. It would show improvement/changes in strategy, etc (Severity 1)
2. Looking at the 3rd visualization, it seems like Steph Curry has the highest true shooting percentage relative to shots/30 mins.
    a. Perhaps have a little pop up box (like a javascript tutorial) that explains the terms when the visualization first appears
    b. Would the graph to the right (scatterplot) replace the line graph to the left when you click, or would it show up on the side?
    c. What does the "size -> minutes played" note mean? How is that related to the shooting % presented in the scatter plot (Severity 2.5)
    d. What is the purpose of the scatterplot -- to show correlation? To prove that Steph Curry is a better player? How is this useful for users? (since the line graph already shows the Steph Curry is a better player than his teammates)

3. Overall
    a. Perhaps more interaction ideas can be included beyond the basic levels

Jamie

1. To answer this question, I would scroll to the second visualization, which details shots by location on the court. I would look at the color scale to determine how to interpret the viz.
    a. Based on the color scale provided, I know that the greener ones are +10% but I don't really know what that means. Is that 10% more than the league average?
    b. I would like to be able to brush over a region to see the total stats
    c. This vis is GSW-specific but I would like to know how other teams compare - if other high-performing teams have followed a similar trend/pattern? How the other teams that have similar patterns have been doing in recent seasons?
    d. Also you say that the goal is to show that they are shooting "well" from the 3pt line. Does this mean that you are trying to prove that they are actually making most of the shots taken from the 3pt line? Is there information on how many shots are taken, vs how many are actually made? Maybe some kind of shading or texture to indicate this other dimensions.
2. To determine the true shooting average, I would scroll down to the third visualization, which has interesting details about the tsa.
    a. I don't understand the first graph (on the left) -- what are the two lines? The title, "Curry vs Teammates" makes me think that one line is Curry and the other is… the average of all of his teammates? The sum of teammates?
    b. What is the league average?
    c. What is the significance of the true shooting average?
    d. Any explanations for why Curry's is so exemplary?
    e. How will I know that I should click on the line chart?
    f. Why is it important that I know the time played, as visualized in the size of the bubbles?
    g. Can I see the bubble chart over multiple seasons? Like maybe allow for brushing over the line chart and then overlay the bubble charts for multiple seasons?
3. Overall
    a. I get that this is clearly geared towards an audience that already likes and/or is interested in bball but it would be helpful to include some examples or definitions of bball shots, clips of games, etc.
    b. Interesting nyt piece on steph curry's artistry that may help guide your narrative a little bit or just provide good images, interesting facts, perspectives etc
    c. there are also obviously a lot of other articles on curry...