Beyond the Numbers: A Visual Display of Advanced NBA Defensive Analytics

Stephen Pelkofer
Georgia Institute of Technology

Richard Reasons
Georgia Institute of Technology

Matt Luppi
Georgia Institute of Technology

Hyun Jin lee
Georgia Institute of Technology

Tim Kim
Georgia Institute of Technology

1 THE PROPOSAL

A new wave of analytics have provided ways for fans of all sports to become more engaged with their favorite games. Unfortunately for fans, analytics have mainly focused on the offensive aspects of the game. Our team's aim is to provide an interactive application that will allow fans of the National Basketball Association (NBA) to explore, visualize, and assess the *defensive* performance of the league's teams and players.

2 SURVEY

As the burgeoning interest in sports analytics continues to grow, so too does the amount of new statistics, metrics, and models aimed at satisfying this collective appetite. However, as Franks et al. convincingly argue in their proposed set of "meta-metrics" [1], not all new sports metrics are constructive. Many lack, discrimination, stability or substantial independence from existing metrics. In our survey of the existing mainstream and academic analyses of defensive performance, we share in their criticisms and formalize a few of our own. These criticisms take two main forms. First, most conventional sources have over privileged offense-related metrics and have been slow to move beyond static, box-score style statistics. Secondly, early academic efforts have succeeded in incorporating modern computing and analysis techniques but have failed to reach or resonate with a general audience.

2.1 Offense-Oriented

The NBA has a very active, analytical following that pores over team performance, player statistics, and salary cap implications. Websites such as NBA Advanced Stats, Cleaning the Glass, and FiveThirtyEight are the major purveyors of basketball related statistics. Though these websites have begun to evolve and introduce more sophisticated analysis, much of what they provide are conventional, offense-oriented statistics in a timeworn tabular format.

The historically heavy skew towards offensive statistics is a fact that has not gone unnoticed. Several early research papers cite this as a principal motivation for their work [2] [3]. Others have tried to expand the canon of traditional defensive statistics by exploring new metrics like rebounding efficiency [4], the three dimensions of rebounding [5] or the effectiveness of forcing weak side dribble[6]. We also reviewed research papers that eschewed defensive statics altogether in favor of defensive systems and strategy [7]. Despite these initial efforts, the reality has been slow to change. There is still a strong need to correct the offensive bias and to provide contemporary analysis that better reflects the true duality of the sport.

2.2 Academic Obscurity

One of the major recent developments propelling a new class of NBA statistics is the availability of play by play and player tracking data. Every NBA basketball arena now deploys advanced camera systems that track X, Y positioning coordinates of players and X, Y, Z positioning of the ball over 72,000 times per game [8]. The torrent of data enabled by this technology has led to a wave of new analyses that are able to incorporate critical but previously missing spatial dimensions. Some of the very first analyses to process player tracking data were able to enumerate previously inferred relationships, like the precise relationship between shot success and shot distance or shot success and opponent proximity [9].

Former Harvard researchers Kirk Goldsberry and Aledxander Franks were at the forefront of this new era of analysis. They published and contributed to several early papers that were among the first to incorporate and contextualize spatial data [10] [11] [12]. More recent analysis has taken several of the spatio-temporal techniques initially developed on basketball data sets and has begun applying them to team-based invasion sports in general [13]. While these papers have mined insight by applying some of the most sophisticated quantitative methods to these new spatially enriched datasets, such as expectation-maximization algorithms, Metropolis-Hastings, and singular-value decomposition (SVD), very little of this work has been able to crossover into the mainstream.

2.3 Moving Beyond The Box-Score

To date most of the advances in defensive analytics on popularized media platforms, such as FiveThirtyEight, are composite, all-in-one type scoring systems. For example, FiveThirtyEight developed a proprietary system called Defensive Rating Accounting for Yielding Minimal Openness by Nearest Defender (DRAYMOND) [14] while ESPN utilizes it's own Defensive Rating Plus Minus (DRPM) [15] in order to distill defensive performance into a singular score. Though these metrics show an appreciation for defense, there are several serious drawbacks to this type of approach. First, it's very likely that the process of compressing the breadth of defensive performance into a singular statistic results in high amounts of distortion. Secondly, it obscures, or perhaps fails to reveal, the different underlying dimensions and relationships that encompass good defense. Lastly, despite their aims these metrics are ultimately another box-score style statistic best situated in a table and lacking the expressiveness of a strong visual presentation.

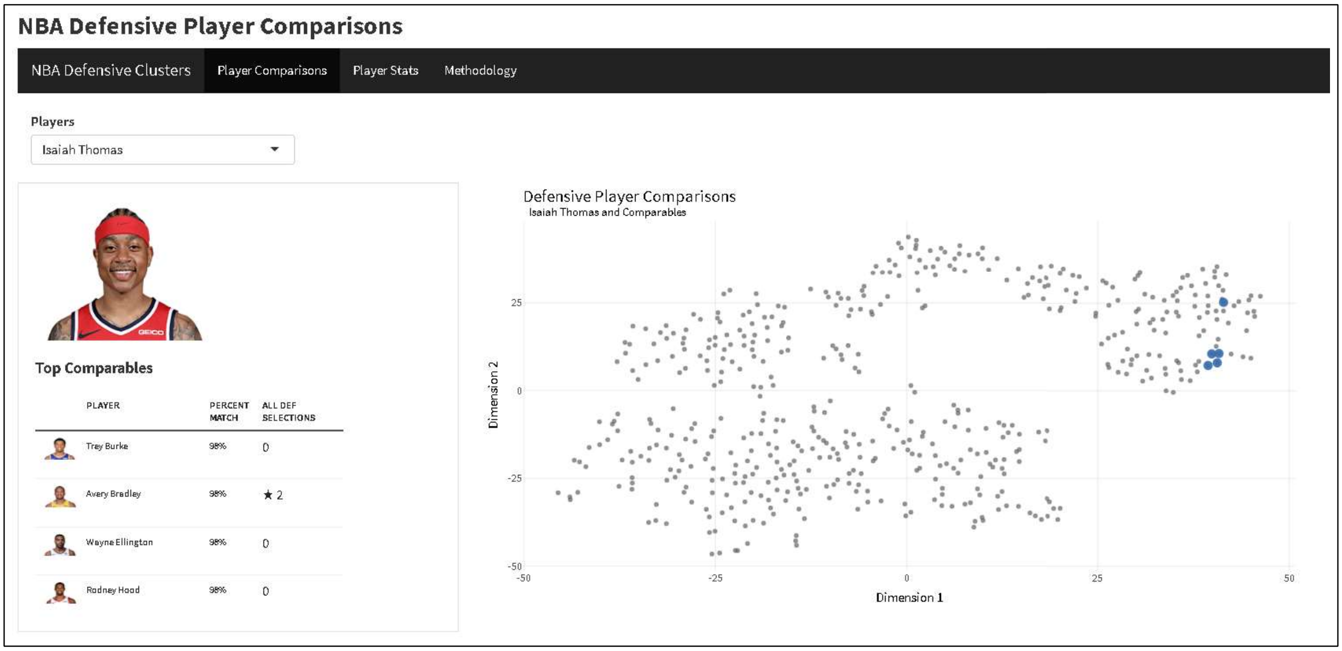
However, in our review of the existing literature, we have come across several examples that strike a powerful balance between analytical rigor and accessible results. While the Sloan research paper Quantifying Shot Quality in the NBA [16] primarily focuses on offense, it contains many effective examples of visualizations that are intuitive yet informationally dense. Similarly, CourtVision: New Visual and Spatial Analytics [17] and Counterpoints: Advanced Defensive Metrics for NBA Basketball [18] expose general tendencies and spatial behaviors that can only be conveyed and interpreted visually. Since sports research and analytics remains a growing but nascent field, we also expanded our survey to include general research on topics such as visualization techniques for high-dimensional data. This search proved beneficial in exposing us to some previously unknown approaches such UMAP and t-SNE dimensionality reduction techniques for enhanced cluster separation and visualization [19] [20].

Ultimately, we find ourselves very supportive of the arguments Goldsberry puts forward in his recent book Sprawlball [21]. We believe that the future of defensive analytics in the NBA will increasingly rely on spatial data for quantification and visualizations for communication. It is the goal of our project to advance this paradigm forward, however slightly, by blending advanced quantitative analysis with engaging and interactive visualizations.

3 PROPOSED METHOD

3.1 Core Design Rationale

The design of our application was permanently shaped by the decision to have the experience be as user-driven as possible. In contrast to many of the popularized websites that display stat lines for passive consumption, we wanted our application to empower user-led discovery of the data. By structuring the app so that all of its defensive comparisons and visualizations are directed by the initial selection, or any subsequent selections thereafter, the end-user is in complete control of the experience. This construction allows them to develop and test their own hypotheses, mine their own insights, and explore previously unknown or underappreciated relationships in the data. The app is ultimately designed to be a platform and a tool that provides a low-friction means to a user-directed end and not an end in itself.



Players: Isaiah Thomas

Top Comparables

PLAYER	PERCENT MATCH	ALL DEF SELECTIONS
Trey Burke	98%	0
Avery Bradley	98%	★ 2
Wayne Ellington	98%	0
Rodney Hood	98%	0

Figure 1: Landing Page

3.2 Discriminators and Innovations

Our proposed method was informed by our survey of related works and guided by our interests as fans and citizen-analysts of the National Basketball Association. In developing our application, we intended to address several of the shortcomings that were identified in our review, including a lack of intuitive or visual interpretability, an over emphasis on box-score statistics, and a propensity to reduce defensive performance into single, all-in-one metrics. In order to differentiate our application from existing sources, we have outlined what we deem to be unique innovations specific to our approach.

Our first critical advantage is our custom data set. To compile this set of features, we scraped and collated data from multiple different online sources. The application benefits from the diversity of this data set which takes into account hundreds of features, including shot contests, rebounds, box-outs, and allows us to incorporate adjustments for time played or pace of individual games played. This data set appears to be significantly more exhaustive than those referenced in any of the analyses we reviewed and allows us to normalize the raw statistics where needed.

The breadth of this data set also enables some of our other key innovations. As noted in the survey, sports metrics often have a high degree of absolute correlation stemming from a common trait or physical feature such as speed or height. To prevent priviledging any facet of defensive which stems from a single factor but is heavily expressed and represented across several defensive metrics we performed principal component analysis on our feature set. By performing this dimensionality reduction we can ensure we're maximizing the amount of independent information accounted for in our model and not inadvertently overweighting any one facet of defense.

Our research also revealed that nearly all the popularized sites have focused on developing scoring systems that result in an ordinal rating of players. These systems rely on weightings that are preset, opaque, and unchanging despite a reasonably wide-range of defensive preferences amongst both fans and teams. To integrate rather than ignore these preferences we chose to build our analysis around a clustering model. By leveraging this sort of machine learning algorithm, we can cluster players with similar defensive performances and playstyle. This allows the end-user to select players whose defensive-play they admire and discoverer other players who exhibit the most similarity. We believe this sort of approach provides a more engaging experience for the audience by incorporating the type of subjective preferences that enrich fandom while still using computational analysis to provide objective and data-backed results.

3.3 User Interface

Our app's user interface is generated through R's shiny package. Its convenient html wrapper functions, built-in reactive programming library, and straightforward server scripts allow for rapid development in a single language while still supporting high levels of interactivity. Upon launching the app, the user arrives at the designated landing page as shown by Figure 1. From this home page, the user can select any desired player from the dropdown box situated in the top left-hand corner. The player selection will render all of the visualizations and graphics specific to that player, including those contained on the app's other pages. In addition to player selection, the landing page displays four players with the highest similarity based on our clustering algorithm – identifying which, if any, of the most similar players have received All-Defensive team accolades.

After the user has selected a player from the dropdown, the players individual statistics are generated on the second tab, Player Stats. The Player Stats page outputs metrics that were derived from the player tracking data - mainly how effective was the player in defending field goals relative to other players in the league. This is displayed in a shot zone output as depicted in Figure 2, as well as in a scatter plot that provides data by shot distance range.

3.4 Implementation and Design

Guided by our core design principle that the app and its experience should be user-led, we quickly gravitated towards an approach that centered around clustering analysis. By using a clustering algorithm we could leave the data relatively intact, letting it speak for itself rather than biasing it or encumbering it with our own assumptions and weightings. To prepare the data for clustering we first aggregated and merged our various data sets by player and then normalized the feature set to represent a per 36 minute basis. Normalizing the data in this manner greatly reduced the collinearity between our chosen features, which are mostly defensive statistics, and total defensive minutes played. Without this normalization, clusters were predominantly determined by minutes played instead of defensive ability. Next we performed principal component analysis on the data set to dampen the effects of any intercorrelation between related statistics and to and extract as much of the independent structure of the data as possible. Though PCA causes us to lose some interpretability directly traceable to the original features we believe this to be a critical step due to the high correlation between certain statistics. Additionally, it's a necessary step to facilitate the visualization of the data in a lower dimensional space in order to illustrate clustering or distance between certain players. While such visualizations would technically be possible by limiting variable selection to just two or three features, we believe the results would be neither interesting nor informative with such restrictive constraints. For this reason, we opted to sacrifice interpretability to have a fuller model with more complete player comparisons.

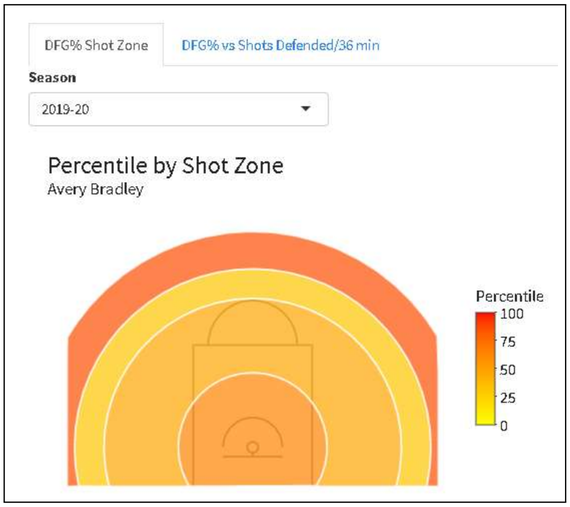


Figure 2: Defended Field Goal Percentage (DFG Pct) By Shot zone

For each active player we used a k-nearest neighbors clustering technique to compute the most similar defenders as identified by the shortest distance between all pairs of players. To perform both the PCA and nearest neighbors algorithms, we utilized Python's popular machine learning library scikit-learn. With the pairwise distance calculated between all 513 players active in the

2019-2020 season we transformed the results to make them easy to return and render based on end-user selections. To help visualize the clustering amongst players we used t-SNE to cast our 50 principal components into two dimensions to render on a scatterplot. As would be expected, the relationships present in the full dataset cannot be perfectly represented in fewer dimensions. The consequence of this result is that the four most comparable players as determined by the full set of components are not always the four closest neighbors in the two-component projection. However the two-component projection provides a reliable vicinity for the cluster, illustrating both its location and helping to identify other players within the same neighborhood but who fall outside of the top four.

4 EXPERIMENTS AND MODEL EVALUATION

Since our clustering analysis was an unsupervised machine learning technique, evaluating its performance is not a straightforward or trivial task. Unlike supervised examples our data set contains no ground truth class assignments or objective performance criteria by which to evaluate it. Therefore our repeated calibrating and iterating were largely guided by our domain knowledge in both the sport and the field of analytics.

In our first iteration of the clustering model and it's reduced, two-dimensional graph it was evident that minutes played was exerting a dominant linear influence on the model. To neutralize the affect of time on the court, we normalized the statistics to a per 36-minute basis. This helped reverse the magnitude of linearization in our plotted data. The impacts of this change could also be observed in the explained variance of our first principal component which fell nearly fivefold from 53% to 11% of the total variance. Paradoxically, we interpreted this reduction to be favorable as the linear relationship between minutes played and stat line contributions is so conspicuous that its inclusion would have overshadowed subtler and more interesting relationships.

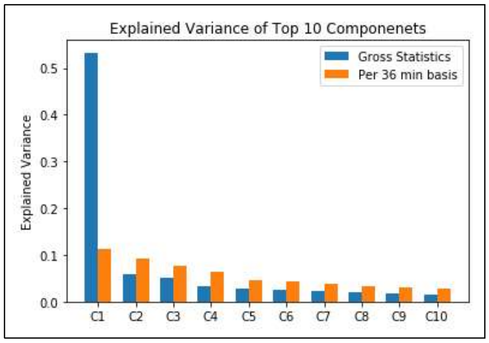


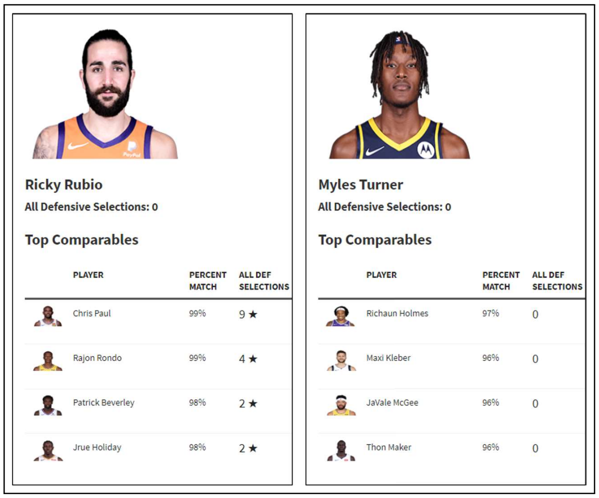
Figure 3: Explained Variance Before and After Normalizing on a Per 36 Min Basis

We also had to rigorously test and ensure that the t-SNE process was not fundamentally altering the clustering structure of our data. Otherwise our k-nearest neighbors results for player comparisons and our two-dimensional plot would be discordant. To confirm that the t-SNE analysis was not distorting the global structure of the data we conducted several tests using agglomerative clustering. These tests included running the agglomerative clustering before and after t-SNE as well as feeding the t-SNE algorithm the clustering assignments as a feature. In comparing the different plotted outputs it was clear that performing t-SNE was helping to minimize crowding while keeping the core cluster structure intact.

Analyzing the results of the player comparison model in a quantitative manner was difficult because most of the defensive data available to us was used in the generation of the model itself. Two data sets were were able to use were All-Defensive team nominations- a yearly poll of NBA writers and broadcasters of the best defensive players during the previous season- and player positions. Our model often compared players with All-Defensive team nominations to other players with All-Defensive team nominations, indicating the model accurately quantifies the relationships viewers identify on the court. Patrick Beverley, a point guard who has made 2 All-Defensive teams, is a good example: all of his comparisons were other point guards and his top 2 closest comparisons have been nominated for 4 and 1 All-Defensive teams, respectively. His third closest comparison, Elfrid Payton, may end up making an All-Defensive team in the future- he has been playing basketball professionally for half as long as Patrick Beverley.

5 DISCUSSION AND LIMITATIONS

We spent a significant amount of time exploring the data and the results in the same mode as our envisioned end-users. Using our collective knowledge of the sport and familiarity with certain players we tried to evaluate the overall sensibility of comparisons for many different players. One good case study is Ricky Rubio, a player who has been "snubbed" from selection onto an All-NBA Defensive team despite being an above average defender [22]. He has garnered votes for selection onto to the team from 2014-2019, but has yet to exceed the necessary amount of votes required for selection. Despite only utilizing player statistics, our methodology concurs with this popular opinion. Rubio's top player comparisons have all been selected to multiple All-NBA Defensive teams, validating that our analysis clusters players effectively.



Ricky Rubio

All Defensive Selections: 0

Top Comparables

PLAYER	PERCENT MATCH	ALL DEF SELECTIONS
Chris Paul	99%	9 ★
Rajon Rondo	99%	4 ★
Patrick Beverley	98%	2 ★
Jrue Holiday	98%	2 ★

Myles Turner

All Defensive Selections: 0

Top Comparables

PLAYER	PERCENT MATCH	ALL DEF SELECTIONS
Richaun Holmes	97%	0
Maxi Kleber	96%	0
JaVale McGee	96%	0
Thon Maker	96%	0

Figure 4: Examples of Higher and Lower Quality Comparisons

Conversely, NBA center, Myles Turner, is another player often in the conversation for the All-NBA Defensive Team but yet to break-through [23]. The eye test and general consensus of his level of defensive play would suggest that his player comparisons would be players with selections to All-NBA Defensive teams. However, his top comparable players have not only never been selected to an All-NBA Defense team, but they are primarily bench players. Additionally, his player comparisons are all centers, which indicates that in this particular instance our methodology has resulted in clustering based more on the statistics associated with a position than the level of defensive performance within that position.

In addition to positive sensitivity, complications also arise from having to normalize the data cross a per 36 minute basis. While we stand by this decision, it does discount the skill involved in sustaining high-level performance across increased durations of court time. Higher skilled players who spend more time on the court are likely to have some sort of fatigue related drag on their statistics when calculated on a per 36 minute basis. Though 4 rebounds in 10 minutes of game time and 12 rebounds in 30 minutes equate to the same per 36 minute average, the latter is presumably harder as it requires sustaining this level of performance versus achieving it in well-rested bursts. Normalizing on a per 36 minute basis also provided less accurate comparisons for NBA players who play very few minutes. Some players only received playtime at the very end of lopsided games in which the final outcome is a foregone conclusion. These players were placed into game situations that are not reflective of the normal flow of an NBA game and are difficult to compare with players in more typical situations.

6 CONCLUSION

Unfortunately, as a result of the coronavirus pandemic the 2019-2020 NBA season has been suspended. Despite this freeze in nearly all sporting events world-wide, we believe the market for sports analytics amongst both teams and fans alike will immediately resume its rapid growth in the years to come. It's also our opinion that compelling digital visualizations will be the likely catalyst and substrate for much of this evolution. The integration of geospatial analysis, modern computing, and sensor processing and it's adaptation to sports is well underway and likely to change how sports are played, strategized, and even appreciated by fans for years to come. Through our different strengths we contributed to this project in a manner that was varied but equal – initiating our own incremental step towards what we believe will be the future of sports analytics.

REFERENCES

[1] Cervone Bornn Franks, D'Amour. Meta-analytics: tools for understanding the statistical properties of sports metrics. *Journal of Quantitative Analysis in Sports*, 3.3, 2007.

[2] Pelton Rosenbaum Kubatko, Oliver. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 12.4, 2017.

[3] Kirk Goldsberry and Eric Weiss. The dwight effect: A new ensemble of interior defense analytics for the nba. *Sports Aptitude, LLC*, pages 1–11, 2013.

[4] Slobodan Simovic. Analysis of influence of basketball game-related statistics on final result based on differences at the 2017 fiba asia cup. *Asian Journal of Sports Medicine*, 10.1, 2019.

[5] Jeff Su Sheldon Kwok Tal Levy Adam Wexler Noel Hollingsworth Rajiv Maheswaran, Yu-Han Chang. The three dimensions of rebounding. *MIT Sloan Sports Analytics Conference*, 2014.

[6] James T. Bartholomew and David A. Collier. The benefits of forcing offensive basketball players to their weak side. *Journal of Multidisciplinary Research*, 4:19–27, 2012.

[7] Miguel Ángel Gómez Jesús Salado Alejandro Álvarez, Enrique Ortega. Study of the defensive performance indicators in peak performance basketball. *Sports Psychology Magazine*, 18, 2009.

[8] Mark Wilson. Moneyball 2.0: How missile tracking cameras are remaking the nba. https://www.fastcompany.com/1670059/moneyball-20-how-missile-tracking-cameras-are-remaking-the-nba.

[9] S. Weil. The importance of being open: What optical tracking data can say about nba field goal shooting. *MIT Sloan Sports Analytics Conference*, 2011.

[10] Alexander Franks. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9.1:94–121, 2015.

[11] Bornn Goldsberry D'Amour, Cervone. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. *MIT Sloan Sports Analytics Conference*, 2014.

[12] Andrew Miller, Luke Bornn, Ryan Adams, and Kirk Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. *31st International Conference on Machine Learning*, 2014.

[13] Joachim Gudmundsson and Michael Horton. Spatio-temporal analysis of team sports – a survey. *ACM Computing Surveys*, 50, 2016.

[14] Fivethirtyeight: A better way to evaluate nba defense. https://fivethirtyeight.com/features/a-better-way-to-evaluate-nba-defense/.

[15] Espn: The next big thing: real plus-minus. https://www.espn.com/nba/story/_/id/10740818/introducing-real-plus-minus.

[16] Jeff Su Sheldon Kwok Tal Levy Adam Wexler Kevin Squire Yu-Han Chang, Rajiv Maheswaran. Quantifying shot quality in the nba. *MIT Sloan Sports Analytics Conference, year = 2014*.

[17] Kirk Goldsberry. Courtvision: New visual and spatial analytics for the nba. 2012.

[18] Counterpoints : Advanced defensive metrics for nba. *MIT Sloan Sports Analytics Conference*, 2015.

[19] Etienne Becht. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 31, 2019.

[20] Laurens van der Maaten and Geoffrey Hinto. Visualizing data using t-sne. *Journal of machine learning research*, pages 2579–2605, 2008.

[21] Kirk Goldsberry. Sprawlball. sprawlball: A visual tour of the new era of the nba. houghton mifflin harcourt. 2019.

[22] Ben Beecken. Timberwolves' ricky rubio left off of all-nba defensive teams. https://dunkingwithwolves.com/2020/04/14/minnesota-timberwolves-josh-okogies-path-to-improvement/.

[23] Ben Gibson. Did myles turner deserve the all-defensive second-team spot over joel embiid? https://8points9seconds.com/2019/05/25/myles-turner-pacers-joel-embiid/.