Analyzing the Modern NFL Running Back

Mike Gallaspy
Kevin Kannappan

ABSTRACT:

Running Backs (RBs) in the National Football League (NFL) have historically been key components to a given team's championship quest. In order to understand the value of an RB to a team, statistics such as yards gained and touchdowns scored are the primary tools. We believe those to be outcome statistics, as opposed to explaining *how* these RBs are successful. Provided a data set from the NFL's proprietary statistics called "Next Gen Stats", we seek to determine how valuable RBs are to their teams today. Specifically, the data set contains spatial and vector information at the time of the handoff (when the RB received the football) and the yards gained on the play. We have taken the data and produced an interactive dashboard using R's Shiny programming. We found that teams that average more yardage appear to average faster RB speeds, most of the top RB rushing teams have been playoff teams, top speed is generally reached at the center of the field, and certain RBs tend to only run in certain paths.

INTRODUCTION:

The National Football League (NFL) has generally known to have become a "passing" league and has valued offenses catered to throwing the football instead of running it. Validating this claim, in the past 5 years, only 9 Running Backs (RBs) have been selected in the first round of the NFL draft compared to 17 Quarterbacks (QBs) and 17 Wide Receivers (WRs)[1]. Top-tier RBs have traditionally been valued for how many yards they accrue whilst running (rushing) and how many touchdowns they scored. For example, last year's Pro Bowl (All Star) selections either were in the top 5 rushing or in scoring[2].

For a passionate fan and researchers alike, both of these statistics compiled come short on too much information. While we know the outcome (yards gained and score yes/no), we do not understand *how* these outcomes have been achieved. Which way did the RB run? How fast were they going? Are the best RBs the fastest ones too? We wish to understand *why* certain teams are more effective at running the football and *why* certain RBs are getting the most yardage.

Our goal is to gain a richer understanding of both the teams that employ the RBs and the RBs themselves. We developed a series of interactive visualizations to help pair outcomes with inputs and compare teams and players. Additionally, the tool is able to visualize the information spatially so that we can understand the critical points on the field to rushing. The tool that we designed should be an immediate upgrade to fans and researchers that want more than just yards and touchdowns.

DATASET:

The NFL has implemented a series of wearable and stadium sensors designed around gathering more detailed data on each play. The resulting statistics from those sensors has been compiled and branded as the NFL's "Next Gen Stats." Normally, this data is proprietary and reserved for NFL personnel only. However, to generate more interest in NFL statistical analysis, the league has shared some of its data in Kaggle competitions, notably the Kaggle Data Bowl[3]. The objective of this competition was to predict yardage gained on run plays through the 2017 to part of the 2019 season. Some of the data that they included to be used as features were the spatial positioning (X,Y), orientation, direction, speed and acceleration of the players at the time of the handoff. Unless we were to embark on a large-scale computer visioning project, we believe that this is the only publicly available dataset that contains such spatial information.

The data contained the spatial information and movement of all players at the time of a running play. Considering that the movement of players aside from the runner (for our purposes) was out of scope for the project, our core data processing revolved around identifying the running player on each play and then dropping irrelevant data (data on other players). We adjusted the format to a row for each running play, spatial/movement/other information on the running player and the resulting yardage. Since the direction of play changes throughout the game, the data was normalized by considering the initial orientation of the RB, and mirroring spatial attributes around the center of the field so that the direction of play was consistent for all rows.

Lastly, we leveraged the yards gained, initial positioning, and direction information to calculate the end-point of the run (in X,Y). The end-point coordinates were leveraged so that we could create rushing "vectors" following the path of the play. As we noted above, we were disappointed to find that the data did not include more time-steps in the play. So, while we know for certain the onset direction of

the run, we would have no way of determining whether or not the player "cut-back" (pivoted direction) to the other side of the field. Therefore, with the absence of that information, our plotted vectors are considered to be estimates.

TASKS:

Leveraging the data that we obtained and engineered on NFL rushes over the course of the past few seasons, we created three interactive visualizations to help us gain a richer understanding of RBs. Specifically, we will use the data to help us in answering the following questions on teams: Which teams have the most productive (yards) RBs? Which teams are the fastest (quickest)? How have teams performed on different downs/seasons? How have teams performed against different amounts of defenders in the box (close proximity to the run play) or by lining up in certain offensive formations? With regard to spatial information, where do RBs run with the most speed on the field? What are the different distributions of run direction and speed for each team? RB? Hence, these visualizations will be leveraged as a tool for different audiences (fans, researchers, team personnel, etc.) to learn more detailed information on the modern NFL RB and rushing in the NFL in general.

We began our analyses with the intention of executing the complete project in d3.js. Upon further investigation into the rigors of the interactivity and the complexity of our tasks, we pivoted away from the tool. In short, it was difficult for us to develop the dashboard feel that we wanted as a result – more of a data analysis tool or product as opposed to some visualizations on their own. We settled on using R Studio's Shiny language (integrating some d3.js) which is built largely upon base R and were able to effectively develop three visualizations to answer these questions. First, we created a series of interactive bar graphs that would allow the user to explore different RB metrics at a team level. Next, we developed an interactive heat map, overlaid on a football field, that navigated the relationship between speed and origin position. Lastly, we created an interactive line plot, overlaid on a football field, that drew line segments for each running play in the data set. We believe all audiences mentioned above may find value in all three of the visualizations developed.

SOLUTIONS:

Summary View:

The objective of the summary view interactive bar chart (Figure 1) was for the user to quickly understand how teams compare across different rushing metrics. First, the user is asked to select whether or not to generate a stacked bar chart with different aggregation levels (None, Season, Down) of a given metric (Speed, Acceleration, and Yards). From there, drop down filters are implemented to allow a user to focus on a team or a specific offensive formation and sliders to specify which seasons to include and how many defenders are in close proximity to the play. Lastly, the user specifies the aggregation type (Average, Max, Total).

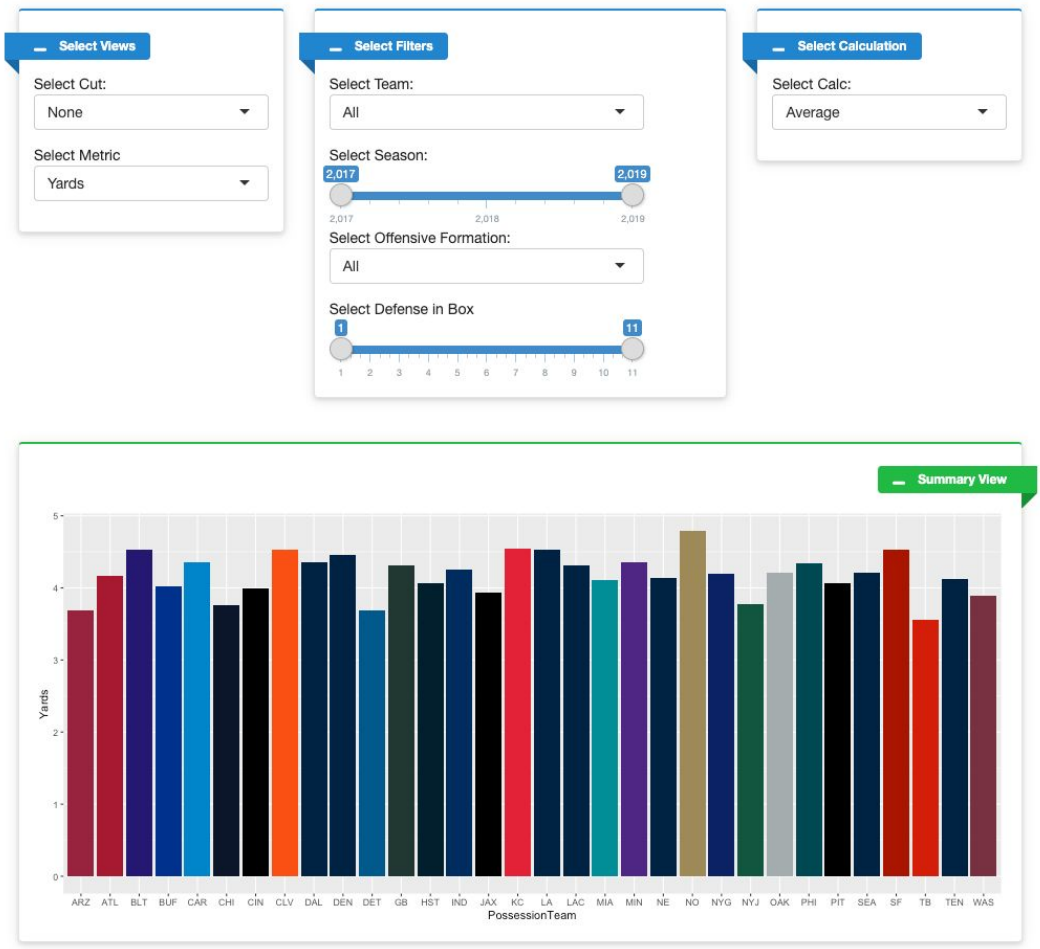

Figure 1: Summary View Interactive Visualization

Both the horizontal and vertical position channels are leveraged to create categorical separation between teams and to specify the magnitude of the selected metric, respectively. Area (2D size) is used effectively to compare the metrics. Color is used as additional encoding to separate teams and to maintain consistency throughout different metric selection. When a user specifies an additional cut, an additional color channel (saturation and luminosity) is added to the plot to determine scaled differences (season and down) across different aggregations. This visualization achieves the following tasks:

1. Compare team rushing efficiency: which team had the most yards, the most expected yards, and the longest rushing play
2. Compare team movement statistics: which team had the fastest/quickest runners, and on average, the fastest/quickest runners
3. Determine if there are any trends across seasons or downs in both (1) and (2)
4. Test hypotheses on different teams using offensive formations or against different defensive alignments

Field Heatmap:

The goal of the field heatmap visualization (Figure 2) was to aid in understanding the spatial dependencies of RB and play characteristics. The visualization breaks the field up into a rectangular grid with a configurable number of bins, and uses a dropdown menu to select a value to visualize (speed, acceleration or direction of RB, or yards gained by the play), an aggregation modality (mean, median, or max), as well as a slider that allows one to filter the data by plays that gained a specified minimum and maximum amount of yards.

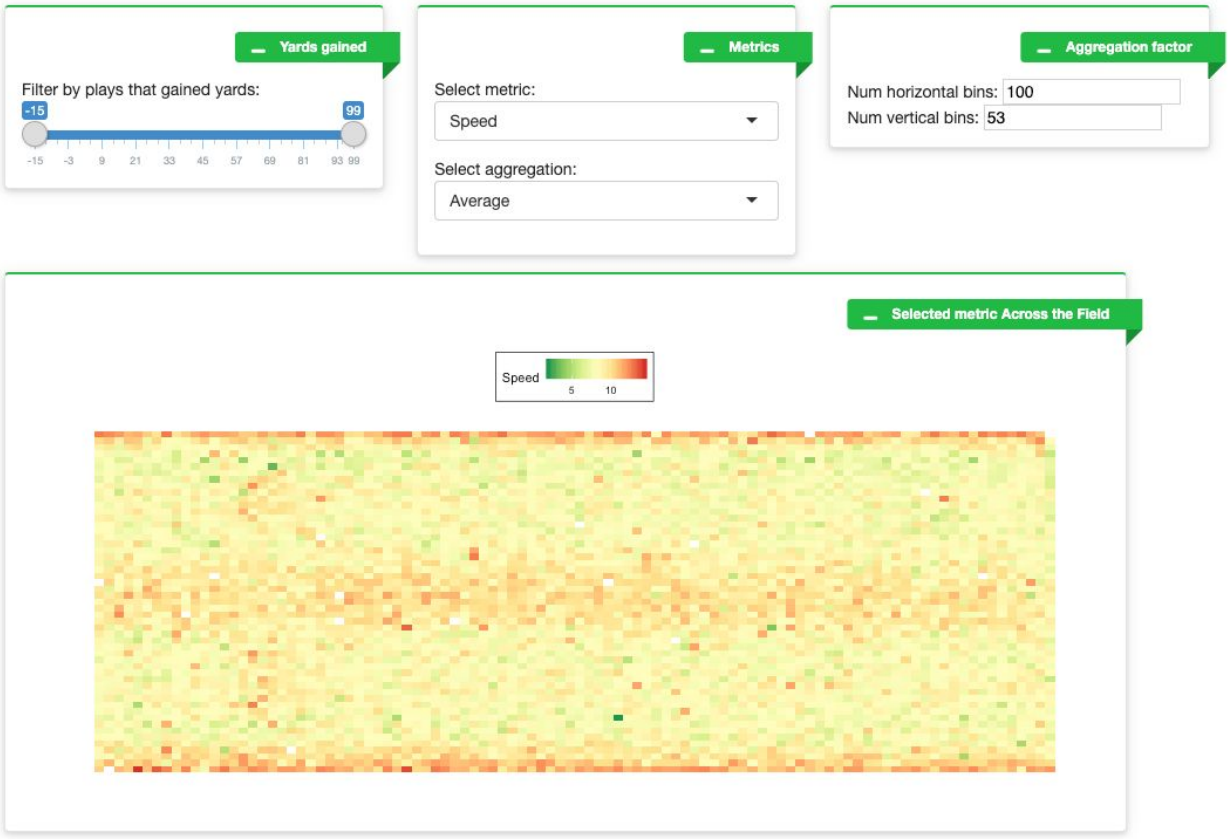

Figure 2: Field Heatmap Interactive Visualization

The position channel is used to naturally indicate the position of an RB at the beginning of a play. The selected characteristic (e.g. speed of RB) is aggregated across the spatial bin, and encoded into the color channels (hue and luminosity) of the bin. When the user visualizes RB direction (Figure 5) arrow glyphs are drawn on top of the bins that redundantly encode the direction in the angle of the arrow. This visualization enables a rich set of tasks, including:

1. Observing trends at locations on the field. For example one can note that RBs achieve higher speeds in general along the centers and edges of the short axis of the field.
2. Understanding spatial aggregation of characteristics. By adjusting the number of bins, one can see for example that RBs are fastest on average further away from the end zone, but achieve the highest maximum speeds close to the end zone.

3. Investigating hypotheses about successful versus unsuccessful plays, which can be explored by filtering by plays that lost yards versus plays that gained yards.

Distribution of Runs

The objective of the distribution of runs interactive line chart (Figure 3) was for the user to visually explore actual running plays by the team that they played for. First, the user is asked to select a team. From there, the user may decide to use a drop down to look at a specific player and use a slider to specify which season to limit the run.

Figure 3: Distribution of Runs Interactive Visualization

The position channel is used to naturally indicate the position of an RB at the beginning and ends of a rushing play. The length (size 1D) and angle channels are able to capture trends and differences of different the runs, as yards gained is on a common scale between the runs. Color is then used as a channel to determine categorical differences between the different RBs. Lastly, area (size 2D) is used to capture the speed of the runner on a common scale (MPH). This visualization enables a rich set of tasks, including:

1. Observing the different play-types that an RB rushes. In Figure 3, we can see that Alfred Morris appears to be used in short yardage situations, right near the end zone, whereas Tevin Coleman appears to primarily be running in between the tackles in the center of the field.

2. Determine the most valuable RB on a given team. By looking at how many times an RB rushes, and if they rush on all parts of the field, we know that the given RB is valuable to the team.

RESULTS:

Summary View:

The visualization does a nice job allowing the user to explore differences in metrics across teams through the different channels that the visualization employs. The position channel is able to separate the teams and determine differences in metrics. Area fills the rectangular bar bins and length is also implemented to help separate those differences. While the color hue encoding is challenging as there are 32 teams in the NFL, having primary colors associated with each team makes it easier to navigate. Lastly, by toggling the cut aggregation drop-down, color saturation and luminosity channels can help to differentiate metrics across downs and seasons. In Figure 4 below, we can easily determine which teams have a heavy emphasis on running the football (Baltimore, San Francisco, and New Orleans).

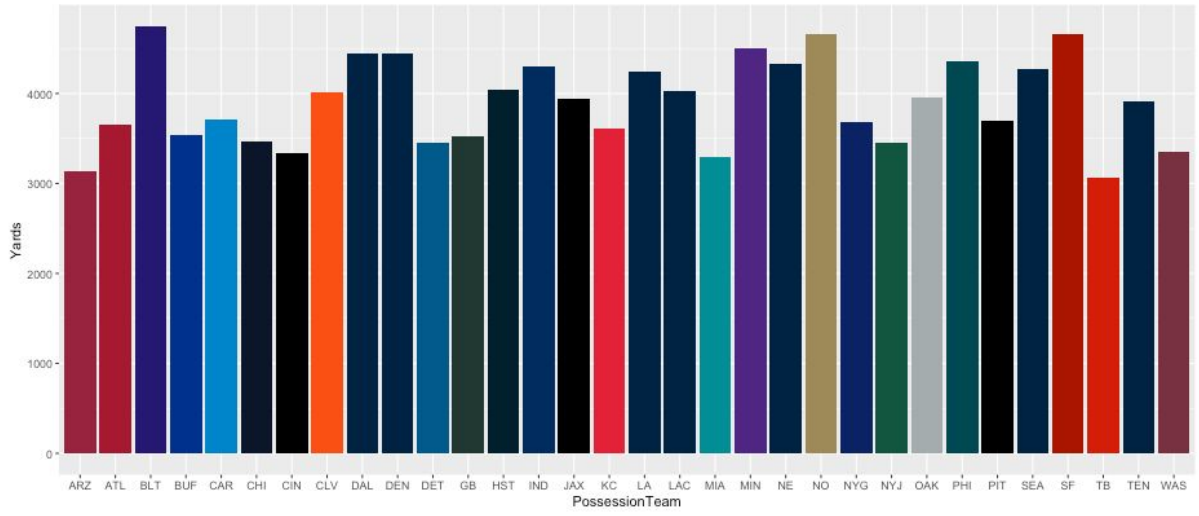

Figure 4: Visualization of Total Yards Gained by Team

Leveraging the interactivity of the graph allows for narrative building: specifically, a user could observe the trends between speed, acceleration and yards. For example, by toggling speed after looking at yards, one may assume that San Francisco relies on faster RBs and thus may get more yardage due to consistently faster RBs.

Field Heatmap:

The visualization is very effective at exhibiting spatial dependencies, since the position channel is used to encode the position of an RB on the field in such a natural way that is difficult to misinterpret. The color hue and luminosity channels are used to fill in the rectangular

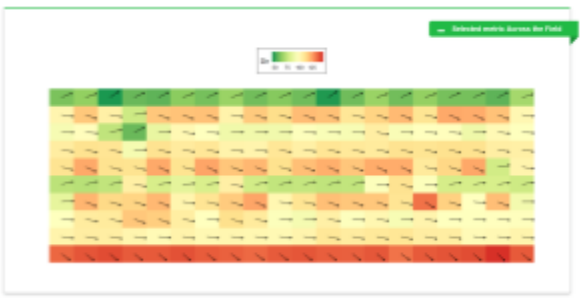

Figure 5: Visualization of direction on field heatmap.

regions, encoding the range of values visualized, and are effective at identifying trends since nearby values have similar colors. For example in (Figure 2) one can see at a glance where the fastest "lanes" in the field are. When the user chooses to visualize RB direction, the redundant arrow glyphs emphasize the physical meaning of the quantity, and the color channels facilitate quick comparison.

The high degree of interactivity makes the visualization very versatile. For example, a coarse trend can be seen in Figure 5 that RBs at the edges of the short axis tend to be moving even closer to the edge, whereas RBs in the center regions of the field tend to move along the center. Adjusting the number of bins allows one to easily spot outliers - for example by visualizing yards gained and setting the number of bins to be large, one can pinpoint precisely where anomalously successful plays occurred. A user who carefully studies the visualization may be able to form and explore sophisticated hypotheses.

Distribution of Runs:

Similar to the field heatmap, this visualization is also very effective at exhibiting spatial dependencies, since the position channel is used to encode the position of an RB at the start and end of the run. The length and angle of the lines are able to accurately project a theorized "rush path." The color hue channel is used to determine categorical differences between players and the area channel is able to adequately encode the speed on a given run.

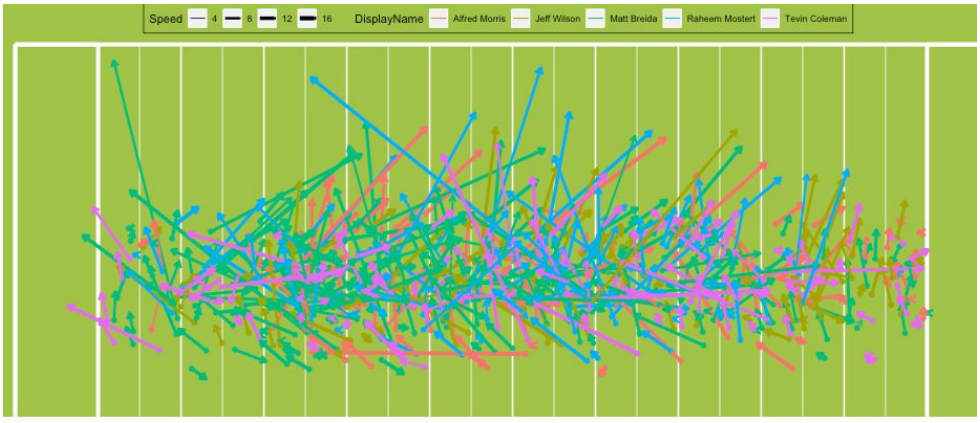

Figure 6: Distribution of SF Runs from 2018-mid 2019

The interactivity of the plot can help to zero in on individual player tendencies. Specifically, we can understand how different RBs accrue the yardage that they do. In Figure 6, Mostert seems to gain a lot of yardage on lateral runs toward the sidelines whereas Colvin seems to rush towards the middle of the field. Morris seems to only rush for a short amount of yardage with low speed, leaving the interpretation that he is used in short-yardage situations.

Across all three visualizations, we experienced difficulty in trying to implement them in d3.js to the point where we abandoned that work and repurposed it into a different programmatic solution. We believe d3 is a great tool for extremely custom visuals, but other than that, too much overhead to implement. Lastly, in the third visualization in particular, without more data on specific time steps for each run, we note that it is difficult to see more unique trends in the different runs. Additionally, if an RB has a lot of attempts, the plot can become convoluted with too much color. Above all, we feel that fans, researchers, and team personnel will find our tool extremely valuable.

Sources List

1. http://www.drafthistory.com/index.php/positions/wr
2. https://www.pro-football-reference.com/years/2019/rushing.htm
3. https://www.kaggle.com/c/nfl-big-data-bowl-2020/