Read Option: Visualizing College Football Play-by-Play Data

Do-Hyoung Park
Department of Computer Science
Stanford University, Stanford, CA 94305
dhpark@stanford.edu

ABSTRACT

Football boxscores and stat sheets do an adequate job of summarizing games, in that they do the minimum of telling people when scoring plays occurred and give viewers a rough idea of how teams compare in summary statistics (such as total yards, total plays, passing yards, rushing yards, penalty yards, etc.) and how the individual players in a game contributed to the accumulation of those summary statistics. However, those traditional summary methods fall short in two ways.

Firstly, it is often difficult to parse text-based boxscores and stat sheets efficiently to get desired information about a game; Secondly, boxscores and stat sheets offer little in terms of differentiating strategic elements such as style of play (more runs vs. more passes), pace of play, explosiveness of plays, and how those elements change with time, which are elements that are often more indicative of a game's progression than simple summary statistics. In this paper, we summarize a new visualization for the outcome of football games that graphically lays out play-by-play data from college football games, spatially encoding both position and time such that viewers can rapidly deduce the strategic elements of a football game and how those elements ultimately contributed to the flow and eventual outcome of the game.

INTRODUCTION

Even with the introduction of computer science into the professional and amateur athletic community, sports leagues have been slow to adapt to the use of technology in both analytics and visualizations. However, despite initial misgivings, data analytics has come to be widely embraced in the baseball and basketball communities as a tool to supplement, not replace, experiential knowledge of the game in order to enrich the quality of both in-game and personnel decisions that are made over the course of games and seasons. However, that computer science "revolution" in sports has still ignored, for the most part, the use of data visualization in summarizing games, seasons, and player contributions despite their utility in helping people easily draw conclusions and distinguish meaningful patterns from large amounts of data. While Major League Baseball has recently started to visualize hit locations, pitch locations and fielding locations (Daren Willman at MLB.com does some fantastic work) and the National Basketball Association has started using shot charts, among other things, football hasn't been visualized too much despite the wealth of opportunity inherent in the game's design.

Firstly, football remains the one sport (among the four "major" American sports and soccer) in which an objective can be clearly defined and constrained to one spatial axis. That is, a team's success is measured by its ability to advance "down" the field along a 100-yard-long axis, with the length of each "play," in yards, clearly defined. Unlike in other "free-form" spatial sports like soccer, basketball, and hockey, it isn't crucial to see movement along the perpendicular axis to understand the results of the game, and thus, the game can actually be reduced quite easily to the idea of a ball moving along one spatial axis with time without sacrificing much knowledge.

Secondly, college football in particular has seen a divergence in strategies over the last decade or so because of the talent disparities that are inherent among teams. While some teams have the talent to play a more "traditional" mix of running the ball and passing the ball, other teams don't have talent at the "skill positions" (running back and wide receiver) and on the offensive line to make such a strategy viable. Thus, some types of teams have to resort to more extreme strategies in order to remain competitive and level the playing field against other teams that might have better talent than them. For example, the service academies (Army, Navy, and Air Force) are known for playing the "triple-option" style of football, in which they almost exclusively run the ball. Meanwhile, other teams like Washington State, Texas Tech and California, which don't recruit well on the offensive line, play a style called the "Air Raid," which relies on almost exclusively passing the ball. Some teams play slowly and methodically, knowing they can beat their opponents talent-wise without any added elements, while other teams play as quickly as possible to try and fatigue opponents into mistakes.

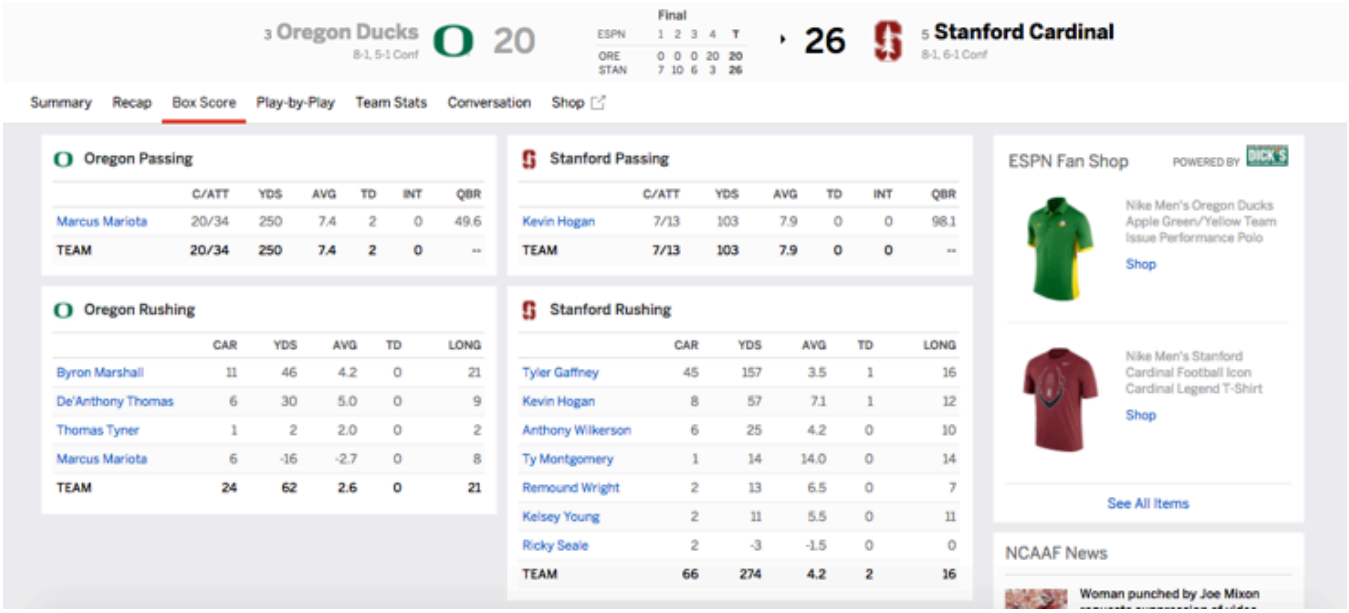

Figure 1: A view of how ESPN presents the results of the Stanford-Oregon game in 2013. The box score at the top presents how many points each team scored in each quarter, and the stat sheet underneath provides individual stats. However, they do not offer a strong idea of the strategies each team used, other than the fact that Oregon seemed to prefer throwing the ball and that Stanford heavily preferred running the ball.

It is that disparity in strategies and how they contribute to the ebbs and flows of games that are often more critical to understanding why games progress in the manner in which they do, and yet, traditional methods of summarizing games, such as box scores (which say how many points were scored by each team in each quarter) and stat sheets (which break down the summary statistics of the game by team and by player) don't offer much in the way of understanding those strategies, how they fit together, and how they contributed to the outcome. The only true way to visualize those strategies is to parse the play-by-play data from the game, but that remains difficult due to the sheer volume of plays in a game (often over 100). Furthermore, there isn't a temporal element to box scores and stat sheets – that is, they don't allow for evaluation of how teams' strategies or stat accumulations changed with time, despite time and that strategic chess match between the players and coaches on the field, as well as adjustments, playing a crucial role in determining the outcome of a game.

ESPN, Fox Sports, and other major sports media organizations continue to turn to antiquated summary methods [1] and haven't approached using visualizations to any great extent to supplement their game coverage, and given that it's not a pressing area of research for many academics, precious little progress has been made in figuring out how to visualize those "big-picture" elements of the game in a meaningful way despite their importance. However, it remains an important problem to solve because of the utility a solution would provide for viewers (who would gain a better understanding of games) and for analysts and sportswriters (who would be able to quickly understand the strategies and their connection to the game's outcome) for their analysis and game stories. Thus, our visualization hopes to bridge the gap between visualizations and the strategy underwriting the game of American football.

RELATED WORK

In terms of football, the only real visualization that occurs on a large scale is with drive charts on ESPN and StatBroadcast, which is the service that records and provides stats to many NCAA Division I athletic programs around the country. ESPN's NFL GameCast charts each play after it develops on a mock football field with a bar that has length proportional to the distance that the play covered, which assists people in visualizing the coverage of a play as the game occurs. However, in terms of using visualization to summarize a game after it ends to aid in postgame recap coverage, very few (if any) publications go the extra mile. ESPN only provides very basic bar and doughnut charts comparing teams' performances in its postgame college football wrap.

The only substantive and new visualization work in recapping football games was conducted in 2013 by Christopher G. Healey at North Carolina State University, who parsed NFL play-by-pay data from the 2012-13 season and created a visualization that represented the ball's position on the field on the horizontal axis and encoded time moving downward on the vertical axis to create a two-dimensional figure that charted a football game in both space and time [2]. Healey's concept was the basis for the visualization done in this project, because in many ways, his style of visualization fell short in achieving a clear purpose.

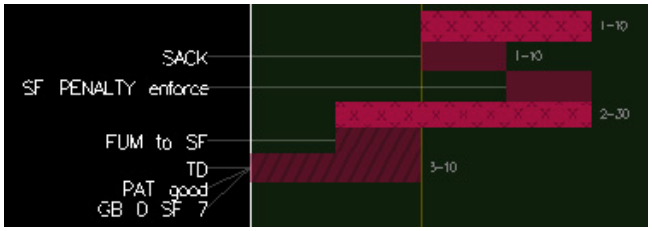

Figure 2: An example of Healey's visualization technique for NFL games. The use of textures to encode differing play types and area to encode time between plays were among several design flaws that made his visualization difficult to parse.

Healey's work felt like just doing visualization for the sake of doing one, rather than with a clear intent, and some of the variable encodings and design choices that he used reflected that. For one, he made graphs that progressed vertically from the start of the game to the finish of the game without any breaks, which made it so that only a handful of plays would fit on the screen at once. This resulted in an unwieldy visualization that made it nearly impossible to get a big-picture idea of the trends and strategies in the game because of the difficulty in comparing the plays from two different time periods, which would often require much scrolling up and down the screen to compare.

The graphs were also heavily annotated, with lots of floating text and horizontal lines disrupting the core design distracting from the core of the visualization, which should have been the bars denoting the plays. Finally, the choice to use textures (different patterns in the fills of individual bars) to denote play types and the use of area to denote the time each play took were ineffective design choices, as those elements have been shown to be less effective at conveying data to viewers than alternatives [3]. Finally, the haphazard use of gradient further detracted from the ability to draw any meaningful conclusions from the data.

We felt that Healey's work had hinted at what could have been a significant new use of visualization to help viewers better understand football games but fell short in achieving a discernable goal, and the ultimate goal of our visualization described in this paper became to improve upon Healey's work in terms of design elements and readability in order to provide a new kind of visualization that could ultimately help people draw meaningful conclusions that old tools like box scores and stat sheets would not be able to find. In particular, the idea of encoding time as a spatial variable on the vertical axis and encoding space as a variable on the horizontal axis made intuitive sense and is something that has been prevalent in other visualizations, such as the aforementioned drive charts used by several data and recap services.

METHODS

The final visualization in this project, called "Read Option," was implemented in JavaScript using the D3.js library, which is commonly used to create visualizations for use in web browsers. Data for the 2013 NCAA Division I-A (FBS) college football season was downloaded from cfbstats.com. The data set came packaged in 17 .csv files, each of which contained data for different elements of the college football season. For example, the file that was the primary source of data for this visualization was the play.csv file, which contained information for all 150,000-plus plays that occurred in the 2013 college football season, with information such as the ID of the game it occurred in, the offensive team, the defensive team, the offense's score at the time of the play, the defense's score at the time of play, the spot of the ball, the yardage to go to the goal line, and the type of play.

Another file used was the drive.csv file, which contained information about every drive from the 2013 college football season, which was used to cover any edge cases in which the result of the last play of the drive was unclear from the play.csv file. The game.csv file helped match the game ID numbers in the game.csv file to the teams that played in each game, the date of the game, and the stadium in which the game was held, while the stadium.csv file was used to reference the stadium names and locations for use in the visualization's header. Finally, the team.csv file was used to match team ID numbers from the play.csv and game.csv files to the actual names of the teams. These data files were parsed by the program and were used to create three "areas" in the visualization.

The "title area" parsed the play.csv file for the final score of the game and the team.csv file for the teams that played in the game, and used that information to express the final score of the game as a header for the visualization. The "title area" also contained a sub-header with other relevant information, such as the name of the stadium, the city in which the game was played, and the date of the game, which were read from the stadium.csv file.

The "box score area," on the top right of the visualization, scraped the play.csv file for data pertaining to how many points were scored in each quarter, and the final score of the game, and expressed that data in a table as a big-picture summary of the scoring in the game.

Finally, the "game area" was the main section of the visualization, consisting of four vertically stretched football fields, one for each quarter, that encoded ball position on the horizontal axis and time on the vertical axis. To populate this area, the program read in the play-by-play data from the play.csv file and, focusing on the distance covered by each play and the time stamp of each play, created a horizontal bar for every play. The length of each bar, representing the distance covered by the play, was calculated using the difference between the spot of the ball on consecutive plays. The horizontal position of each bar was computed by scaling the yardage gained or lost on each play to the scale of the visualization and locating the corresponding location on the small "field" in the relevant quarter in the "game area" of the visualization. The vertical position of each bar represented the time stamp of the play,

scaled such that the total vertical distance in each "quarter" represented 15 minutes of game time.

This vertical position was more difficult to compute, as the play.csv file was incomplete and only contained time stamps for a few plays in each game (timeouts, kickoffs, extra point attempts, and the first play of each drive). Because of this incomplete data, the time stamps for all of the plays without time data had to be interpolated using the closest plays earlier and later in time that had populated time stamps. A linear interpolation was done using those earlier and later plays to populate the missing time fields for most plays, resulting in most of the plays on any given drive being more or less spaced evenly in time.

That linear interpolation is obviously imperfect in accurately representing a football game, since the time between plays varies considerably in reality due to the clock continuing to run when the ball stays in bounds and the clock stopping on incomplete passes or when the ball-carrier goes out of bounds. Although that is a significant limitation in the game-accuracy of the visualization, it actually accounts for a significant positive in terms of drawing conclusions from the visualization, which will be discussed in further detail later in this paper.

Finally, the bars were colored according to the type of play they represented and the team that ran the play, with different hues encoding different play types and different colors (red and blue) encoding the two different teams. Color and hue were specifically chosen because they are easy to discern at a quick glance. It is also worth noting that in this design, time between plays was encoded with position of the bar instead of with bar area, as design studies in the past have determined that people cannot very accurately estimate areas.

Given that the primary objectives of the visualization were to give viewers a clear idea of the styles of play of each team (primarily expressed in the types of plays run), the pace of play (primary expressed in the time stamps of the plays), and how those elements evolved over time (primary expressed in differences between the quarters), it was important to highlight those differences in the variable encodings, which is why the specific choices were made to use color/hue and position to encode play type and time, as those encodings stood out well and ended up being quite easy to differentiate in the final visualization. In order to highlight the differences between the quarters, the primary structure of the visualization was to have four separate visualization areas that were horizontally aligned next to each other in order to facilitate comparisons in play styles and pace between the quarters. It was also important to lay out the different elements of the visualization in such a way that the majority of the plays in the game (if not all of them) are visible on the screen at the same time. Whether or not this is achieved is dependent on the zoom level and screen resolution of the viewer's monitor, but the zoom level can be adjusted so that the majority of the visualization is on the screen at once, which is the most effective manner in which it should be viewed.

Lots of experimentation was done to find the optimal ratios of bar heights and distances between plays relative to the scale of the visualization as a whole to ensure that the visualization was not too crowded or visually overwhelming. Special care had to be taken in such a regard because of the large number of plays in every college football game, and we wanted to present all of that information and be able to identify all of the individual plays in the game while still making sure that it wasn't too messy to identify big-picture trends and draw conclusions from the data.

RESULTS

The visualization is live at the following URL: http://stanford.edu/~dpark027/448B/project/. It consists of a HTML area at the top of the screen in which the viewer can manually select a home team and visiting team via drop-down menus, or select a random game from the 2013 season using the "Random Game" button. The visualization will update in real time when a valid matchup between a home team and visiting team (e.g. a matchup that was actually contested in the 2013 college football season) is selected by the viewer. If the selected matchup is not valid (e.g. the specified matchup between teams did not actually occur in 2013) the visualization will not change. The default loaded visualization is the Stanford-Oregon game from Nov. 7, 2013, in which No. 5 Stanford defeated No. 3 Oregon 26-20 en route to a second consecutive Pac-12 title and appearance in the 100th Rose Bowl Game. This choice was not random – apart from being the most significant home football game in Stanford history, it also featured two teams with drastically different play styles, which are readily highlighted using our visualization.

The visualization, shown in Fig. 3 below, features the "title area" in the top left, the "box score area" in the top right, and the "game area" filling up most of the space. Each of the four quarters is represented by a vertically stretched football field, in which the left end zone represents the visiting end zone (in which the home team scores) and the right end zone represents the home end zone (in which the visiting team scores). In each of the quarters, time progresses vertically from top to bottom – that is, the top of the quarter represents the clock reading 15:00 and the bottom of the quarter represents the clock reading 0:00 (the clock winds down in football, as in all major sports except soccer). Time is scaled linearly from the top to the bottom of each quarter.

The plays themselves are displayed as bars aligned on top of the fields representing each quarter. Black bars represent special teams plays (kickoffs or punts) by either team, with the start of the bar representing the ball position at the start of the play, and the end of the bar representing the ball position at the end of the return by the opposing team. Yellow bars represent penalties, with no distinguishing

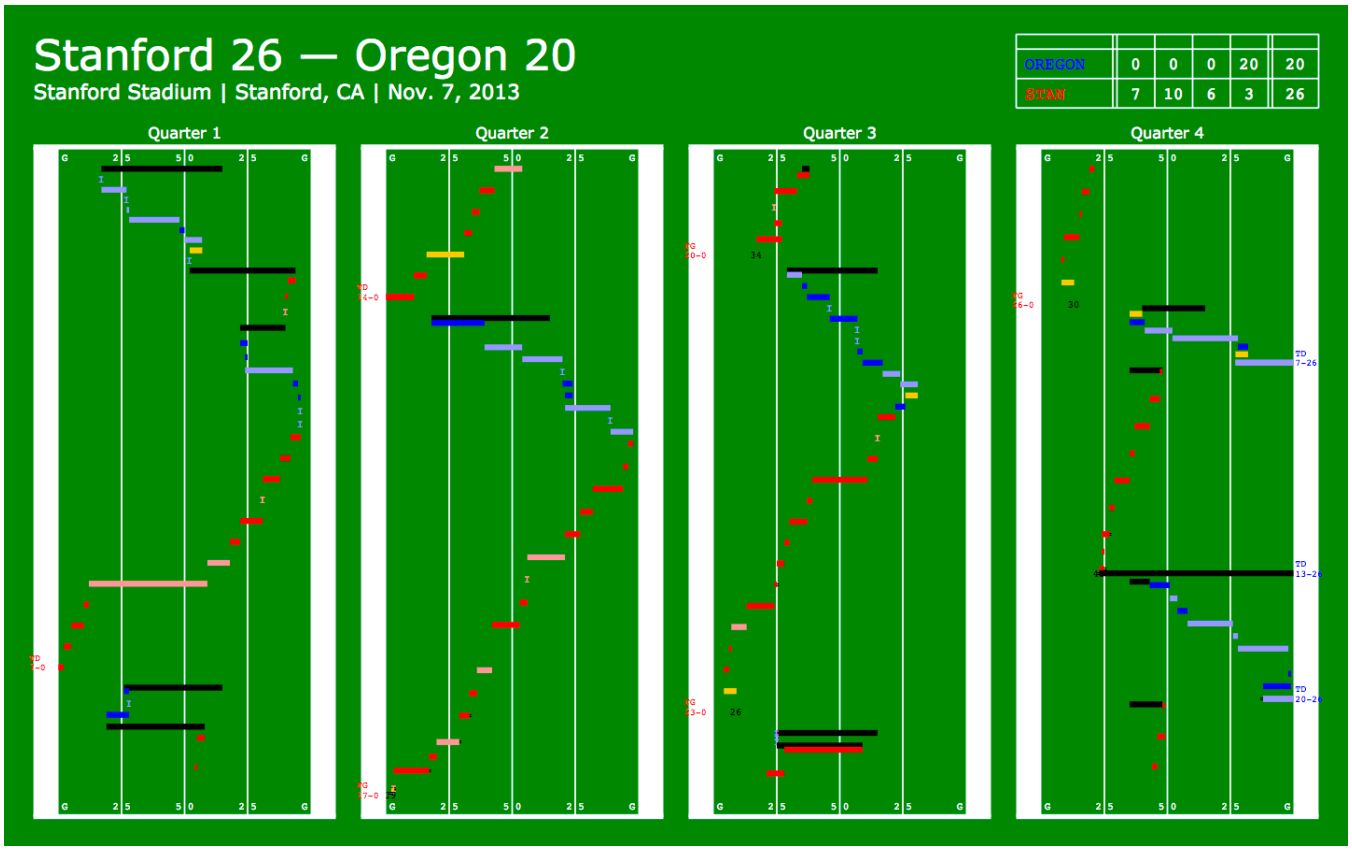

Figure 3: The complete "Read Option" visualization of the 2013 game between Stanford and Oregon. The "title area" is visible at the top left and is populated with the final score, the location, and the date. The "box score area" at the top right shows the points breakdown by quarter. The "game area" takes up the majority of the space and charts every play in the game by Stanford (red) and Oregon (blue) by quarter.

factors between penalties imposed on the home team and penalties imposed on the visiting team (it should be clear from contextual evidence which team the penalty has been imposed on, based on whether the bar goes forward or backward from the previous spot of the ball and which of the teams is on offense). Red bars represent plays by the home team (in this case, Stanford), while blue bars represent plays by the visiting team (in this case, Oregon). Darker hues (of both colors) denote running plays, while lighter hues (of both colors) denote passing plays for the corresponding team. Field goal attempts are denoted by black numbers on the field, which represent the distance of the attempted kick. Incomplete passes are denoted by a "I" on the field at the spot of the ball, and timeouts are denoted by a "=" where the ball was spotted when the timeout was called by one of the teams. Touchdowns and field goals are marked by "TD" or "FG" with the resulting game score in the margin of the corresponding end zone.

Although extensive user testing of the visualization has not yet been done to evaluate the advantages of "Read Option" over other options, the example shown in Fig. 3 makes it apparent that it is much easier to draw conclusions about styles of play, pace of play, and how those strategic elements evolved over the course of the game and contributed to the final score than it is from the ESPN summary of the game shown in Fig. 1.

With regards to style of play, Stanford is widely known in the college football world to be a team that emphasizes running the ball primarily and only passing sparingly. That's evident looking at the visualization above, in which the majority of Stanford's plays are dark red as opposed as light red, showing a very high ratio of running plays to total plays called. The evolution of that strategy as the game progresses is also evident, as Stanford has several pass plays (in light red) in the first two quarters, but only completes one pass play in the second two quarters, which means that Stanford, with the lead, elected to do nothing but run the ball late in the game in order to try and "burn clock," a common strategy in college football. In comparison, Oregon exhibits a healthy mixture of running and passing (dark blue and light blue) throughout the course of the game, which is, again, readily apparent with just a quick glance at the visualization.

Pace of play is also quickly deduced from the visualization, and becomes particularly apparent in the fourth quarter, when the distances between the red bars are particularly wide, as Stanford waits a long time between plays to try to run down the clock while ahead, while the vertical distances

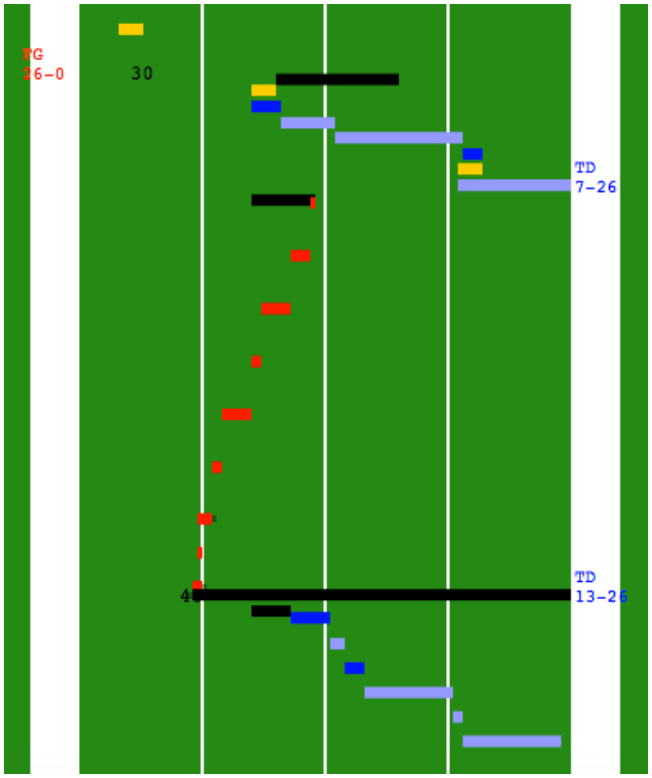

Figure 4: A close-up of a selection of plays from the fourth quarter of Stanford-Oregon, showing the noticeable difference between the large vertical distances between the red bars (corresponding to a slow pace of play by Stanford) and the small distances between the blue bars (corresponding to a rapid pace of play by Oregon).

between the blue bars is minimal as Oregon tries to snap the ball as quickly as possible to erase its deficit. One of the primary storylines after the game was the contrast between how slow and methodical Stanford's offense was, as compared to Oregon's fast-paced, balanced offense, and that is quickly deduced from this visualization but not as readily seen in ESPN's summary statistics in Fig. 1.

The limitations of using linear interpolation to fill in the empty time stamps for the majority of the plays in the game was discussed in the "methods" section, but one advantage of interpolating play times is that it sacrifices realism for a better representation of pace of play, in that it's much easier to read average pace of play for a drive from evenly spaced bars, as opposed to haphazardly placed bars that might more accurately represent the timing of the plays in the actual game but might hinder quick comprehension of how the pace of play on that drive actually compares to elsewhere in the game.

Other general game elements that can be more easily seen in the visualization include the timing of the scoring plays and how effectively each team controlled the clock (e.g. time of possession). Current game summaries traditionally list the scoring plays in chronological order and box scores give a rough temporal order of scoring plays based on what quarter they occurred in, but still don't offer much in terms of context of the play, which this visualization does offer. And in terms of controlling the clock, looking at Fig. 3, it's immediately clear that the story of the game was the fact that Stanford was able to hold onto the ball for large swaths of game time due to Oregon's defense not being able to stop Stanford's methodical rushing attack, as made clear by the prevalence of short rushing plays by Stanford that led to extensive drives that stretched for long vertical distances in all four quarters (corresponding to long times). Again, that is not an element that is obvious when looking at countable stats in box scores or stat sheets, but it is very easily deduced by looking at our visualization.

DISCUSSION

It is our hope that this paper makes clear the potential benefits of using our visualization to understand and diagram football games after they occur, as a tool to enrich and supplement the knowledge gained from reading traditional box scores and stat sheets for a more holistic idea of the strategies and styles of the two teams involved in the game and how those factors, ultimately, led to the outcome of the game. Sports-related visualizations such as this one could potentially come in useful for sportswriters writing on deadline, as this offers a way for sportswriters to not only see all of the plays in a game in a compact space, but also to see the big-picture trends of the game and how they evolved over time in ways that might not have been obvious from the play-by-play or summary data alone, which could lead to topics for potential game stories. That is, indeed, the point of visualization – to present data in a new way such that patterns are easier to visualize and become more obvious. It could also be useful to coaches in a similar way. By easily being able to see what future opponents have done in the past, coaches could quickly get a basis for what they need to focus on in a week of practice or how to call their plays during the game itself.

The other important takeaway from this type of visualization is that it does not necessarily need to spell out the conclusions that we feel that viewers *should* draw from it; rather, it simply presents the data in such a way that it allows viewers to more readily draw those conclusions for themselves. Nowhere in the visualization itself do we ever allude specifically to styles of play or pace of play; instead, with the data in front of them, viewers can take those mental leaps for themselves, or perhaps see other patterns that make more sense to them given how the data is presented. That is something rather unique to the sports world – given the different styles and perspectives that everybody brings to the sport, there often remains much more that is up for interpretation following a game, which might not necessarily be the case for most other data sets, which often have a specific trend or conclusion that can be drawn from them. Because of this, sports is a field in which it is particularly important to give as many different ways to view data as possible, to aid in the breadth and depth of the exploratory process for information hidden in the data.

FUTURE WORK

This visualization is effective in presenting all of the data from a football game that does not necessarily involve the summary statistics (total yards, rushing yards, passing yards, completion percentage, etc.), but often, as has been discussed in this paper, it is important to know both the play-by-play data and summary statistics to gain the most complete knowledge of how a game played out, meaning that it would be an improvement to integrate those summary statistics into the visualization as well moving forward. One interesting way to do this would be to allow the reader to select some summary statistic (say, total offensive yards, or rushing yards by a certain player, or something similar), and to plot it on the football fields (with the bars representing all of the plays) as a line graph with the magnitude of the statistic represented on the horizontal axis and the vertical axis representing time. If both the summary statistic and the plays were plotted in the same area, it would be easy to gain the benefits from the current benefits, as well as being able to track how the chosen summary statistic varied with time and how it was affected by the plays in the game. For example, if the chosen stat was, say, rushing yards by a certain Stanford running back in the game, it would be useful to see how quickly he gained rushing yards in the first quarter versus in the second quarter, to see how his effectiveness changed as the game progressed and how much his impact on the game changed with time. It would also be interesting to see which individual plays or drives contributed the most to his accumulation of stats, and conversely, which of his stats contributed the most to the outcome of the game.

Another interesting thought would be to track individual player usage alongside the existing visualization by shading the field a different color for the chunks of time in which a certain player was in the game, which would allow the viewer to quickly visualize how often and when a team would rotate players at a certain position, and what situations dictated those rotations. For example, it would be easy to see if a certain player entered only for third-down situations, or only in passing downs, or late in the game in so-called "garbage time." This would be of particular use to track the situational uses of wide receivers, running backs, and quarterbacks, and to track which players made the biggest impacts when they were in the game.

The final improvement specific to this visualization would be to somehow integrate online highlight videos or gameplay videos such that clicking the bar corresponding to any big play would bring up the video of that play. The idea would be that highlight videos in isolation are interesting and give a "quick hits" summary of the plays that shaped the game, but offer precious little context for each play, which the visualization would be able to do. As long as the visualization is used as an aid in recapping a game, it would also be helpful for viewers to be able to step away from the visualization and look at a play as it actually unfolded on the field to add further context to the information presented in the visualization. If this is too ambitious, even including some interactivity in the main visualization by providing information about a play by hovering over its corresponding bar (such as formations of the offense and defense, down and distance, types of routes run, etc.) would be interesting, though it would be extremely time-consuming to collect data sets that included such information, because it would have to be collected by hand while manually watching through the game videos, which is the major roadblock to something like this being feasible.

Although it would be interesting to visualize play-by-play data like this for the other major sports as well, it would be much more difficult for those sports because time could not be encoded as a spatial variable for a two-dimensional visualization because those games are played in more than one spatial dimension (as opposed to football, which sees its objective lie along only one spatial axis). It would be difficult, yet rewarding in many of the same ways that this visualization was, if somebody could figure out a good and clean way to visualize every play in a baseball or basketball game so that meaningful new conclusions can be drawn from those visualizations.

ACKNOWLEDGMENTS

A special thanks goes out to Dr. Maneesh Agrawala, Ludwig Schubert, and Peter Washington, whose teaching and guidance have made CS 448B this quarter at Stanford a rewarding, thoughtful experience. I'd also like to thank cfbstats.com for providing the data sets used for this visualization. Finally, I'd like to thank my good friend Alexa Philippou for her continued words of encouragement that helped keep me sane during finals week, and my roommate Leon Yao for putting up with my fiery football hot takes and for being amenable to random sports discussions at all hours of the day, night, and everything in between.

REFERENCES

1. Stanford silences Oregon for three quarters, holds off rally. http://www.espn.com/college-football/recap?gameId=333110024.
2. Visualizing NFL Football Games. https://www.csc2.ncsu.edu/faculty/healey/NFL_viz/.
3. Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods. *J. Am. Stat. Assoc.* (Sept. 1984), 531-554