Utilizing Parallel Coordinates to Analyze Collegiate American Football Statistics

Sean Owens, sas364

I. INTRODUCTION

American football is currently greatly lacking in meaningful statistical analysis visualizations. This paper will present a solution this problem. It will begin by briefly reviewing the current state of sports statistical analysis and visualization. Then, the problem to be solved will be stated. The paper will then outline the data to be used, target audience, and user tasks to be performed by the visualization solution. A review of the design decisions made during the development of the visualization will be included along with a description of the implementation methods used to complete the design. Finally, a future work section will outline improvements that can be made to the current solution as well as present an alternative visualization concept.

II. RELATED WORK

While sports statistics have been recorded for a long time, the widespread use of those statistics to analyze performance is a relatively recent development. Initially, statistics were analyzed to determine the potential performance of individual athletes. An entire field of mathematics, Sabermetrics, was developed to analyze baseball performance [1]. Now, the area of sports analytics is rapidly advancing. Recently, information visualization techniques have been developed to aid in analysis. Initially, scatter plots and bar charts were used to analyze potential correlations, such as this baseball analysis by Cox and Stasko [2]. More recently, there has been a push towards visualizations that correlate statistical data with a spatial location. An example of this kind of visualization is the Snapshot hockey visualization presented by Pileggi et al [3].

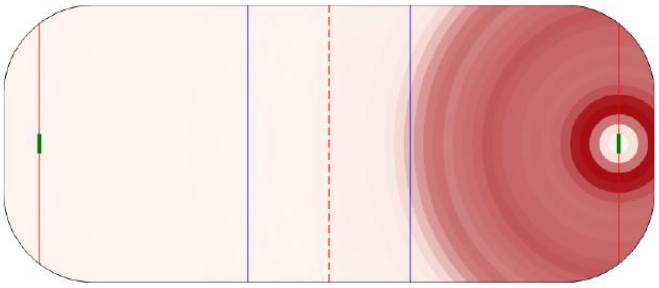

Fig. 1. Hockey rink used as a spatial substrate. Used in the Snapshot visualization [3].

This trend, however, cannot be easily translated to American football. The difference between American football and other sports such as hockey, football (known as soccer in the United States), and basketball, is that American football only has one important spatial dimension. Therefore, there are very few American football visualizations that utilize the field as a spatial substrate, as is done in other sports. An example of a visualization that does attempt this spatial mapping is ESPN's College Football Gamecast, shown in Figure 2. Another concept for a spatially mapped visualization is presented in the future work section of this paper.

Fig. 2. ESPN's Gamecast showing a single drive overlaid on a football field. Image from: http://digitalvideospace.blogspot.com/2012/10/second-screen-and-college-football.html

III. PROBLEM

While sports visualization as a whole is becoming a much larger area of interest, there are very few visualizations that specifically address the sport of American football. This is due in part to American football's lack of a simple two-dimensional mapping, creating a difficult environment to display statistical data on. There are an even fewer number of visualizations for collegiate level football. Furthermore, almost all of these visualizations only display live, single-game data which is inadequate when looking for performance trends that only manifest over hundreds of games.

College football is a multi-billion dollar industry with many schools having football programs worth over $100 million [4]. As the elite programs win more prestigious games, they are awarded larger payouts and endorsement deals. This increased

worth translates to better facilities and ultimately better recruiting which in turn leads to more wins in the future. Therefore, in a time when the best teams are getting better and the worst teams are getting worse, any legal competitive advantage is crucial.

A key area that can provide this advantage is statistical analysis. This visualization looks to provide an easy-to-use snapshot of the statistical domain of college football. The visualization will compare teams based on the number of wins they have achieved in a season. This visualization will attempt to show statistically what facets of the game are most important to focus on given the ultimate goal of maximizing the number of wins in a season.

IV. Data

The data used for this visualization is season-long football data for 31 statistical categories for every team in NCAA Division 1, FBS Schools (the highest level of collegiate competition) over seven seasons. The list of FBS Schools and an enumeration of the statistical categories used can be found in the Appendix. There are 12 official statistics recorded by the NCAA. 11 of these statistics are used in the visualization along with 20 other unofficial categories that are also provided by the NCAA statistics archives. There are many more statistical categories that could be included in the visualization; however, the 31 categories used were chosen to give an overview of the statistical domain.

The data used was compiled by the NCAA (National Collegiate Athletic Association). The stats for the current year can be found at the NCAA's website [5]. Previous years' data can be found at the same website by modifying the URL. The years used for the visualization are 2005 through 2011. The main reason that these years were chosen is that 2005 was the first year that data for each statistical category chosen is available for every team. This range allows the user to overview a large data set without having to deal with null data points.

V. Audience

This visualization targets any football statistical analysts who would like to explore the college football statistical domain. The term analyst can be used to define multiple levels of users. On the amateur level, there are average fans and amateur analysts who wish to track a team's performance over time. On the professional level, the main users are professional analysts who can work for independent companies, individual schools, or league administrations. Other users on the professional level are coaches and trainers who can use the visualization to focus their practices on deficient areas.

VI. Visualization Tasks

There are three main user tasks that this visualization serves:

1. Observe which statistics predict wins.

This is the main user task. While the data is plotted on various category axes, the comparison between each data path is based on the number of wins associated with each. Therefore, a user would perform this task to identify patterns in the data that indicate whether a particular statistic correlates to a higher win total. Furthermore, the user may wish to perform the inverse of this operation and search for statistical categories that have no correlation to win percentage, indicating that improving this aspect of a team's performance will not produce an increase in the team's ability to win games.

2. Compare n-win teams vs. m-win teams.

A secondary task provided by this visualization is the ability to compare teams that have a certain number of wins against teams of a different number of wins. The visualization grants the user the ability to see at a finer level the differences between individual numbers of wins. Figure 3 shows an example of this task. The visualization shows that, while a statistic may show that teams in the 10+ win range (top line) have a much higher statistical average than the other ranges in a certain category (Figure 3(a)), the individual win separation does not exhibit the same level of correlation to win percentage (Figure 3(b)). The limitation of this task is set by the granularity of the view levels (i.e. teams with win counts in the 5-9 range cannot be compared to teams in the 0-4 range). This limitation is addressed in the future work section.

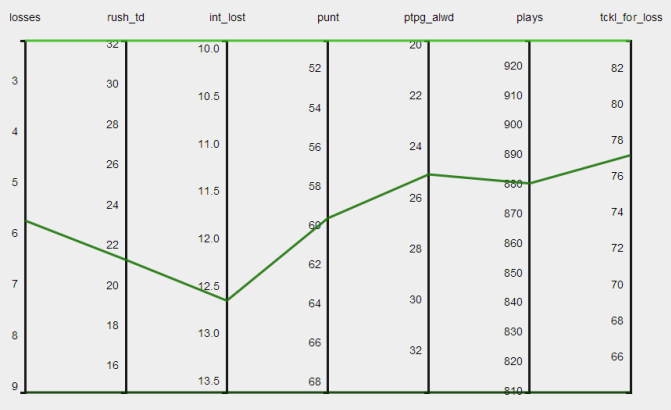

Fig. 3(a). Data paths for the three different win ranges.

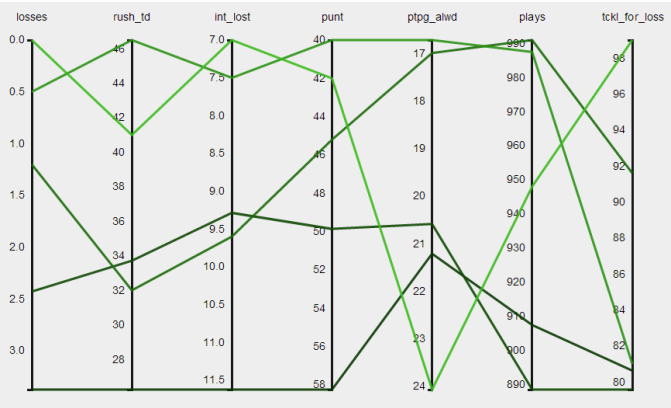

Fig. 3(b). Data paths for wins in the 10+ range.

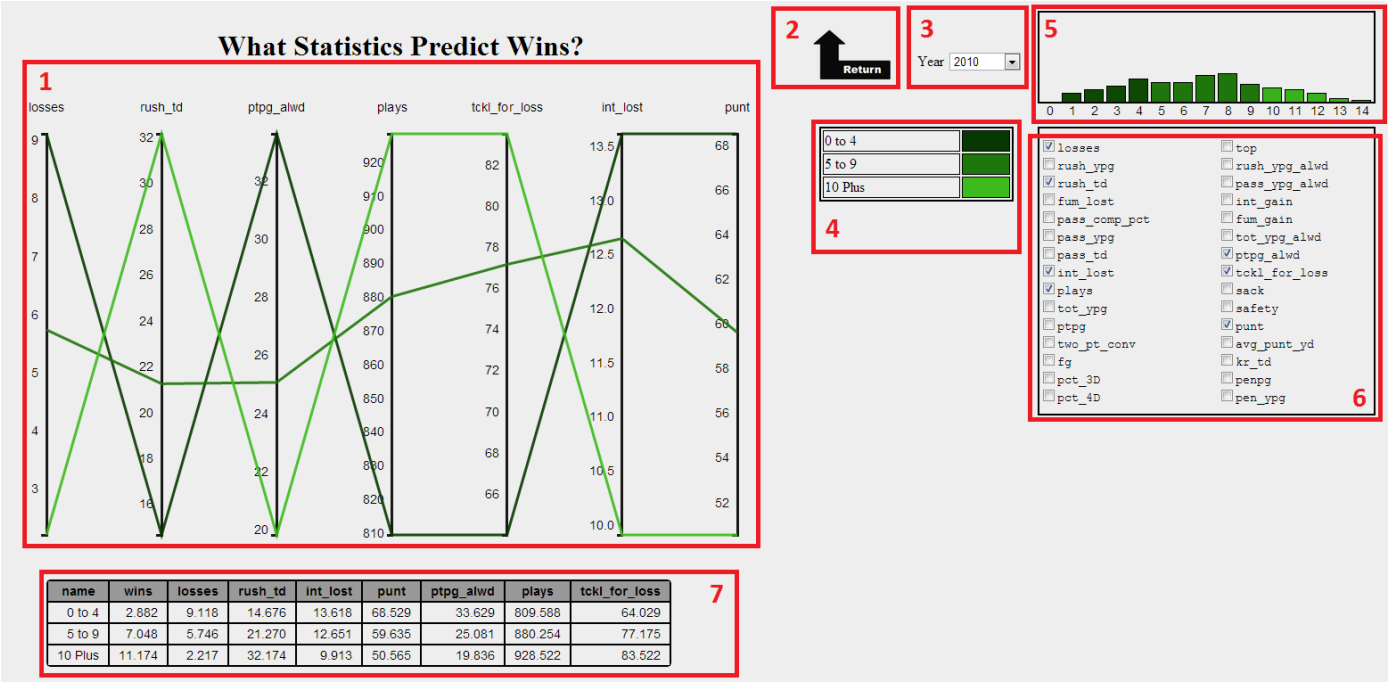

Fig. 4. An overview of the visualization highlighting the seven design zones.

3. Compare team a vs. team b.

A third user task is the ability to compare individual teams. This task is designed for fans of a specific team or for an analyst who is looking for a particular statistical pattern that was previously observed subjectively. An example of this would be if a team seemed to win often because of a particular statistical feature, the analyst could investigate whether that team's performance is an anomaly or a possible hidden "key to victory." Again, this task is limited by the selection granularity of the current visualization. However, this task is further limited by the number of teams represented in the data. It is clear that over 100 data paths would create a highly congested graph.

VII. Design

A. Color selection

Two color schemes were chosen for this visualization. The first color scheme is the main color scale used in the parallel coordinates chart, legend, and win bar chart. This scale is a linear scale along the chroma-lightness plane in the HCL color space.

Fig. 5. Color scale used in the visualization.

Using the color scale shown in Figure 5, lower values are tied to darker/desaturated colors while higher values are tied to lighter/saturated colors. Because the visualization has to distinguish up to fifteen different interval values, a green hue was chosen for its higher response than the blue or red channels.

The second color scheme chosen was the selection color. These colors are used in the parallel coordinate chart, win bar chart and data table to distinguish the current element selection from the others in the zone. For this reason, a highly saturated blue hue was chosen as the selection color, and a highly desaturated gray was chosen for the deselected objects. This choice allows the selected object to “stand out” from the other elements in the zone. Furthermore, a blue hue was chosen for the selection color because it is in the opposite color channel (Yellow/Blue) from the green (Red/Green) color scale used elsewhere in the visualization.

B. Visualization Design

There are seven zones of the visualization that will be described below. The locations of the zones are shown in Figure 4.

1) Parallel Coordinates Chart

The parallel coordinates chart is the main zone of the visualization. The base structure of the chart is a standard parallel coordinate system. The categories selected by the users are converted to individual axes that are arranged next to one another horizontally. After the axes and scales have been determined each data element to be plotted defines a path that connects to each axis at that elements value. This allows for multiple, multivariate data elements to be compared simultaneously.

For this visualization, the user is comparing elements based on their number of wins. Therefore, an additional visual variable was needed to indicate this value. Each path being plotted is colored according to their win count using the color scale described in the previous section.

A key feature of the parallel coordinate chart is that the user can very easily determine if two statistical categories are correlated, uncorrelated, or inversely correlated. In order to aid in this use of this feature, the user has the capability to flip an individual axis so that its range is up-side down relative to the other axes. This ability allows for much easier detection of anomalous and/or uncorrelated categories. Moreover, the user has the ability to slide and reorder the categories as desired. This gives the user the control necessary to analyze different sets of statistics on demand.

Another key feature of this section is the changing of detail level. Because there are fifteen different possible values for the win variable, displaying all of the data at once on a single chart is highly ineffective. Therefore, this visualization relies on different view levels to display all of the data. There are three view levels. The first level is an overview level that groups wins into three ranges (0-4, 5-9, 10+). The second view level displays a single range from the previous view with each path being a separate win value (i.e. 0,1,2,3,4). The third view level displays all of the teams with a specific win value as individual paths. Because of this separation of view levels, it is important that the user not lose their "virtual orientation." So, an animation is included to transition from one view state to the next, allowing the user to keep track of which data subset they are entering/leaving.

2) Return Button

The return button was included to give the user the ability to return to a previous level of view. For instance, if the user were to select to view all of the teams with six wins, pressing the return button would allow the user to return to a view of combined data paths, each corresponding to a specific number of wins between five and nine.

Two options were considered for this functionality. The first was a scrolling behavior that would function similar to a zoom control found in many interactive mapping programs. This option was considered for its intuitiveness. The second option was the button feature which was ultimately chosen. The decision against the scroll option was made because of its requirement for the user to interpret each view as a zoomed in view of a previous view. Because no frame of reference was provided elsewhere in the visualization to orient the user to their current view level, this would ultimately be an unintuitive method for navigation.

3) Year Selector

The year selector is a simple feature to allow the user to choose different data sets (one for each year) or a combination of all of the data sets.

There were two options for this zone. The first option was a selector. The second option was a set of independently selectable objects (i.e. checkboxes). The first option was ultimately chosen due to the implementation methods described in the next section, specifically the decision to create static combinations of the data and only allow the user to view a single year or all of them together.

4) Legend

The legend is the key structure used to communicate the current view to the user. Because of this, the legend will update every time the view changes. The legend, therefore, serves three purposes. The first is to tell the user what data is being shown. The second purpose is to connect the data names to the paths on the parallel coordinates chart. This connection is accomplished with a simple color chart. The final function of the legend is to allow the user to brush individual teams in the highest detail view. This functionality allows the user to highlight individual teams for quick location on the chart because at this level of detail it is possible for there to be over twenty paths on the chart at a given time. An example of this selection is shown in Figure 6.

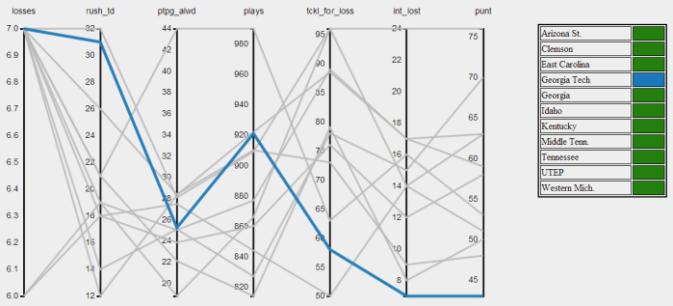

Fig. 6. Individual win view with one team selected.

5) Win Bar Chart

The win bar chart was included to give the user an overview of the data set currently selected. The chart is a histogram showing how many teams have each number of wins in the given data set. Improvements to this zone are included in the future work section.

6) Category Selector

The category selector zone allows the user to select or deselect different statistical categories. Upon selection the category will be added as another axis to the parallel coordinate chart as well as another column in the data table. The category selector is vital to this visualization. It allows the user to compare different statistics to find trends.

7) Data Table

The data table was a zone that was included to give the user the ability to view the actual values presented in the parallel coordinates chart. As with the legend, the data table must update with every view change to accommodate new data sets, new path objects, or new category sets. Because the number of categories and/or data elements can become large, brushing was included in the table. Figure 7 shows how this allows the user to highlight a single data element and have its entire row be highlighted as well as highlight a statistic category and have its entire column be highlighted.

name	wins	losses	rush_td	int_lost	punt	plays
Akron	1.000	11.000	10.000	14.000	81.000	742.000
Bowling Green	1.000	10.000	16.000	20.000	67.000	820.000
Memphis	1.000	11.000	5.000	13.000	80.000	718.000
New Mexico	1.000	11.000	10.000	15.000	87.000	805.000
San Jose St.	1.000	12.000	7.000	17.000	84.000	788.000

Fig. 7. Data table with single row highlighted on hover.

VIII. Implementation

A. Data Structures

There are three main data structures used in this program:

1. The first structure contains all of the raw data for each team individually. It is a three-dimensional structure. The first dimension is an array with an element for each year of data and an extra element to store the combined data for all years. The second dimension is a list of objects, each corresponding to a team. Finally, the third dimension is an associative array that relates data to the statistical categories given by the visualization.
2. The second structure contains the combined data for each set of teams that has a given number of wins. Similar to the previous structure, this structure's first dimension is an array corresponding to the year. The second dimension is an array whose indices equate to the number of wins for all of the teams each contains. The third dimension is a list of team objects similar to that in the previous structure. The fourth dimension is an associative array that holds the data for each team.
3. The third structure holds all of the combined data used by the visualization in the highest view level. This structure is virtually the same as the first structure with the exception that the objects in the second dimension do not correspond to individual teams but to an aggregated group of teams all falling within the same range of wins (i.e. 0-4,5-9,or 10+).

B. Program Flow

There are five main portions of this program:

1. Data load

In this step, the data is loaded from each .csv file. Because data reads are asynchronous in D3, a count variable is used to insure that all files are read before the visualization is run. Once the files are read, the next step is called.

2. Visualization setup

In the set up step, the objects needed by the visualization are created. The data structures defined above are populated, the color and category scales are defined, and the return button, year selector and visualization window are defined.

3. Chart setup

In the chart setup step, the parallel coordinates chart is constructed. This consisted of defining each of the axes and their scales. Also, the drag behavior allowing the axes to be reordered is defined in this function. Furthermore, the win overview bar chart and the category selection pane are defined in this step.

4. Plot

In the plot step, the data paths are added to the chart and the legend and data table are populated.

5. Update

The update step is the largest step. This step handles all of the interaction with the visualization. A function is defined to handle changes to the view level, year, or category list. The fundamental purpose of each change function is to update the active data set, remove old data from the screen and make a new call to the plot step (and the chart step in the case of a category list change).

C. Programming Resources

This visualization was written in JavaScript. All data linking and charts were created using the D3 JavaScript library [6]. The colors chosen for the color scale, team selection, and table highlighting were all generated using the "HCL Picker" designed by Tristen Brown [7]. Page structure and formatting were created using HTML and CSS respectively.

IX. Future Work

A. Visualization Improvements

There are several areas of improvement for this visualization. Overall, more coordination between the zones would be very beneficial. Examples of this would be to allow the user to select views from the win bar chart zone and be able to brush elements in the parallel coordinates chart, legend, and data table and have those selections be reflected in the other two zones.

Another area for improvement would be to add more statistical categories to the visualization. This visualization was designed to give an overview of the statistical domain. Therefore, many statistical categories were omitted to simply save on development time and complexity, but could easily be included.

Another modification that would benefit the user would be the ability to dynamically select the win ranges. This would allow the user to compare individual win values that cannot currently be compared (i.e. 4 and 5). Furthermore, allowing the user to define the year range being used would give them added control of the data set being analyzed.

Finally, a major modification that could be explored for effectiveness is the inclusion of multiple views. This would be useful if the user wished to compare one year to another. While this would not directly aid the user in the tasks defined, it could allow him/her to find anomalous data sets to be removed before further analysis.

B. Spatial Football Visualization

This concept visualization provides a theoretical approach to visualizing American football using the spatial layout inherent to the sport.

As stated previously, American football differs from other sports in that it only has one meaningful dimension. This visualization utilizes this fact and displays the data on a one-dimensional graph. The graph is a line that is segmented into 100 increments, each representing a yard line on the football field.

This visualization looks at individual play data. Specifically three different values: starting location, ending location, and play type. Each play is represented by an arc that begins at the starting location and ends at the ending location. The height of the arc is determined by the length of the play and the color of the arc is determined by the type of play. Plays that advance the ball in the positive direction are shown on top the line

while plays that result in a loss of yardage are shown under the line. A concept drawing of this visualization is shown in Figure 8.

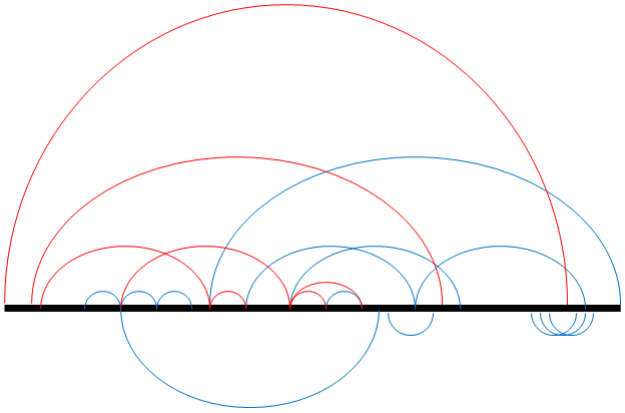

Fig. 8. Concept drawing of spatial based football visualization.

In Figure 8, red arcs represent passes and blue arcs represent runs. There are many features that could be included with a visualization such as this. Some examples are: brushing based on play type or distance; filtering based on play metadata such as down, time during game, or location; filtering based on team, year, game, etc.; altering the range displayed by the base line, allowing the user to focus on a particular section of the field. Furthermore, visualization techniques such as edge bundling may prove useful as the number of arcs increases.

APPENDIX

List of statistic categories used in the visualization:

- Wins
- Losses
- Rushing Yards/Game
- Rushing TDs
- Fumbles Lost
- Pass Completion Pct.
- Passing Yards/Game
- Passing TDs
- Interceptions Lost
- Plays
- Total Yards/Game
- Points/Game
- Two-Point Conversions
- Field Goals
- Third Down Pct.
- Fourth Down Pct.
- Time of Possession
- Rushing Yards/Game Allowed
- Passing Yards/Game Allowed
- Interceptions Gained
- Fumbles Gained
- Total Yards/Game Allowed
- Points/Game Allowed
- Tackles for Loss
- Sacks
- Safeties
- Punts
- Average Punt Yards
- Kick Return TDs
- Penalties/Game
- Penalty Yards/Game

List of Division I FBS Schools (years of data inclusion in parentheses):

1. Akron (2005 - 2011)
2. Alabama (2005 - 2011)
3. UAB (2005 - 2011)
4. Arizona St. (2005 - 2011)
5. Arizona (2005 - 2011)
6. Arkansas St. (2005 - 2011)
7. Arkansas (2005 - 2011)
8. Auburn (2005 - 2011)
9. Ball St. (2005 - 2011)
10. Baylor (2005 - 2011)
11. Boise St. (2005 - 2011)
12. Boston College (2005 - 2011)
13. Bowling Green (2005 - 2011)
14. BYU (2005 - 2011)
15. Buffalo (2005 - 2011)
16. Fresno St. (2005 - 2011)
17. California (2005 - 2011)
18. UCLA (2005 - 2011)
19. UCF (2005 - 2011)
20. Central Mich. (2005 - 2011)
21. Cincinnati (2005 - 2011)
22. Clemson (2005 - 2011)
23. Colorado St. (2005 - 2011)
24. Colorado (2005 - 2011)
25. Connecticut (2005 - 2011)
26. Duke (2005 - 2011)
27. East Carolina (2005 - 2011)
28. Eastern Mich. (2005 - 2011)
29. Fla. Atlantic (2005 - 2011)
30. FIU (2005 - 2011)
31. Florida St. (2005 - 2011)
32. Florida (2005 - 2011)
33. Georgia Tech (2005 - 2011)
34. Georgia (2005 - 2011)
35. Hawaii (2005 - 2011)
36. Houston (2005 - 2011)
37. Idaho (2005 - 2011)
38. Illinois (2005 - 2011)
39. Indiana (2005 - 2011)
40. Iowa St. (2005 - 2011)
41. Iowa (2005 - 2011)
42. Kansas St. (2005 - 2011)
43. Kansas (2005 - 2011)
44. Kent St. (2005 - 2011)
45. Kentucky (2005 - 2011)

46. LSU (2005 - 2011)
47. Louisiana Tech (2005 - 2011)
48. Louisville (2005 - 2011)
49. Marshall (2005 - 2011)
50. Maryland (2005 - 2011)
51. Memphis (2005 - 2011)
52. Miami (OH) (2005 - 2011)
53. Miami (FL) (2005 - 2011)
54. Michigan St. (2005 - 2011)
55. Michigan (2005 - 2011)
56. Middle Tenn. (2005 - 2011)
57. Minnesota (2005 - 2011)
58. Mississippi St. (2005 - 2011)
59. Ole Miss (2005 - 2011)
60. Missouri (2005 - 2011)
61. North Carolina (2005 - 2011)
62. Nebraska (2005 - 2011)
63. UNLV (2005 - 2011)
64. Nevada (2005 - 2011)
65. New Mexico St. (2005 - 2011)
66. New Mexico (2005 - 2011)
67. North Carolina St. (2005 - 2011)
68. North Texas (2005 - 2011)
69. La.-Monroe (2005 - 2011)
70. Northern Ill. (2005 - 2011)
71. Northwestern (2005 - 2011)
72. Notre Dame (2005 - 2011)
73. Ohio St. (2005 - 2011)
74. Ohio (2005 - 2011)
75. Oklahoma St. (2005 - 2011)
76. Oklahoma (2005 - 2011)
77. Oregon St. (2005 - 2011)
78. Oregon (2005 - 2011)
79. Penn St. (2005 - 2011)
80. Pittsburgh (2005 - 2011)
81. Purdue (2005 - 2011)
82. Rice (2005 - 2011)
83. Rutgers (2005 - 2011)
84. San Diego St. (2005 - 2011)
85. San Jose St. (2005 - 2011)
86. South Carolina (2005 - 2011)
87. South Fla. (2005 - 2011)
88. Southern California (2005 - 2011)
89. SMU (2005 - 2011)
90. Southern Miss. (2005 - 2011)
91. Texas St. (2011)
92. La.-Lafayette (2005 - 2011)
93. Stanford (2005 - 2011)
94. Syracuse (2005 - 2011)
95. Temple (2005 - 2011)
96. Tennessee (2005 - 2011)
97. Texas A&M (2005 - 2011)
98. TCU (2005 - 2011)
99. Texas Tech (2005 - 2011)
100. Texas (2005 - 2011)
101. UTEP (2005 - 2011)
102. UTSA (2011)
103. Toledo (2005 - 2011)
104. Troy (2005 - 2011)
105. Tulane (2005 - 2011)
106. Tulsa (2005 - 2011)
107. Air Force (2005 - 2011)
108. Army (2005 - 2011)
109. Navy (2005 - 2011)
110. Utah St. (2005 - 2011)
111. Utah (2005 - 2011)
112. Vanderbilt (2005 - 2011)
113. Virginia Tech (2005 - 2011)
114. Virginia (2005 - 2011)
115. Wake Forest (2005 - 2011)
116. Washington St. (2005 - 2011)
117. Washington (2005 - 2011)
118. West Virginia (2005 - 2011)
119. Western Ky. (2007 - 2011)
120. Western Mich. (2005 - 2011)
121. Wisconsin (2005 - 2011)
122. Wyoming (2005 - 2011)Wyoming (2005 - 2011)

REFERENCES

[1] Jim Albert. "An introduction to sabermetrics." *Bowling Green State University (http://www-math.bgsu.edu/~albert/papers/saber. html)* (1997).

[2] Andy Cox and John Stasko. "Sportsvis: Discovering meaning in sports statistics through information visualization." *Compendium of Symposium on Information Visualization*. 2006.

[3] Hannah Pileggi, et al. "SnapShot: Visualization to Propel Ice Hockey Analytics." *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012): 2819-2828.

[4] Chris Smith. "College Football's Most Valuable Teams." *Forbes*. Forbes Magazine, 22 Dec. 2011. Web. 10 Dec. 2012. <http://www.forbes.com/sites/chrissmith/2011/12/22/college-footballs-most-valuable-teams/>.

[5] http://web1.ncaa.org/mfb/mainpage.jsp

[6] http://d3js.org/

[7] http://tristen.ca/hcl-picker/