

SportVis: Discovering Meaning in Sports Statistics Through Information Visualization

Andy Cox and John Stasko*

College of Computing / GVV Center
Georgia Institute of Technology

1 MOTIVATION

Suppose you want to take a traditional tabular presentation of a baseball team’s schedule (for example, the 2004 Red Sox game log [8]) and draw conclusions about specific aspects of the season or the season as a whole. How do you quickly answer questions like:

- How did the team perform in games started by Pedro Martinez? Compared to other starters?
- Did they play an unusual proportion of close games? How did they do in those games?
- Was there some kind of “shape” or pattern to their season?
- Was there some trend in the games they won or lost?

The tabular schedule typically contains ten or more attributes (e.g., date, home/away, score) for each game. For a full season, this can amount to 2,000 or more data points. Answering queries, particularly open-ended ones, from this table requires significant cognitive effort. Often, the harder problem is determining what questions to ask or what patterns to look for.

Information visualization involves creating a visual representation of abstract information and allowing manipulation of this representation to facilitate exploration and insight. Information visualization systems support visual information seeking by providing both overviews of large amounts of complex data and the means to zoom, filter, and gain details on demand [9].

SportVis uses visualization to help people discover meaning in the massive amount of statistics generated during sporting events. Unfortunately, there has been surprisingly little substantial application of information visualization to sports statistics. Presentations of general information visualization systems (e.g., [6]) and techniques (e.g., treemaps) frequently use sports data as examples. There is a dearth of tools (the visualization of a tennis match [5] is one notable exception) to help users with tasks specific to analyzing sports. Such tools could not only enhance the enjoyment of fans, but also assist media coverage and team officials decision making.

2 VISUALIZING BASEBALL

A single baseball game generates a large amount of data. One game of two teams comprises nine innings, each with multiple plate appearances. Each plate appearance has a discrete outcome and can be described by a number of attributes. We can further separate the plate appearance into individual pitches. We can aggregate the data at many levels along this hierarchy and create interactive visual representations of these aggregations.

We focus on the season and game levels in this version of the baseball SportVis. We provide two visualizations, which we call the *baseline bar display* and the *player map*. The user is always exploring data with respect to a particular team (the *focus* team). For example, when we reference the number of runs scored, we mean the runs scored by the current focus team. We use data obtained from Retrosheet [7] and Baseball Databank [4].

*e-mail: andy@{cc.gatech.edu,sportvis.com}; stasko@cc.gatech.edu

2.1 Baseline bar display

A frequent set of tasks involves evaluating the performance of a team or player over the course of one season. Each game in a season contains many attributes with respect to the focus team, including result (win or loss), location (home or away), runs scored and allowed, and starting pitcher. The goal in baseball is to win games, which is accomplished by scoring more runs than your opponent. Therefore, we want to emphasize the game outcome and runs.

The baseline bar display (Figure 1) represents games in a season by assigning several key game attributes to visual attributes. The display is divided in half by a horizontal line (the *baseline*). Each game is encoded by horizontally aligned bars, one above the baseline and one below the baseline. Each bar is colored by the outcome for the focus team (green for a win, red for a loss). The length of each bar above the baseline encodes the absolute value of the run margin (i.e., shorter bars are closer games); the length below encodes the number of runs scored by the focus team. Choosing runs scored and margin over runs allowed is a design decision—we want to emphasize the relative closeness of games. Also, we can derive runs allowed from the runs scored and run margin.



Figure 1: SportVis baseline bar display for the 2004 Boston Red Sox

The combination of visual attributes (particularly bar length and coloring) leverages preattentive processing [11] to facilitate easy detection of trends or patterns over a season. For example, the combination of the win/loss color and segmented bars emphasize sequences of games with a disproportionate amount of wins or losses. We can quickly perceive large clumps of green or red bars (long winning or losing streaks, respectively), which help spur related questions about those games. (Was a key player injured?) We can also easily identify extreme performances (e.g., large margins of victory) with runs encoded by length.

The design of the baseline bar display is inspired by the *New York Times* annual weather graphic, identified by Tufte as a “distinguished” graphical presentation of dense quantitative information [10]. The *Times* graphic encodes several variables related temperature along a time-series axis. We aim for similar information density in the baseline bar display.

Interaction distinguishes information visualization systems from

static information graphics. Dynamic queries [1] help us further explore the data. We can better understand the performance of a team under certain situations by providing filters to display only games that match these criteria. We can also sort the bars by runs (either scored or margin) to better understand of the overall distribution or the effect of luck (e.g., winning a disproportionate amount of close games). By filtering on the starting pitcher and sorting by runs scored, we can get a quick overview of the run support of a pitcher better than a single number can provide (Figure 2). These techniques allow rapid exploration of the data and continuous formulation of new questions.



Figure 2: Baseline bar display for 2004 Red Sox sorted (L to R, asc.) by runs scored and filtered for games started by Curt Schilling.

Let us consider a simple example task. Fig. 2 displays the 2004 Boston Red Sox season. We first sort the games in increasing order of runs scored by the Red Sox (note the “Sort by:” control). A dynamic query control (not shown due to space) is used to display only games started by Curt Schilling. Starting pitchers have little, if any, control over the number of runs their own team scores. so we expect the bars to be evenly distributed throughout the spectrum of runs scored. We can quickly see that the bars are more heavily distributed toward the right side, indicating that his run support was better than normal.

2.2 Player map

Baseball statistics are often categorized as counting statistics (e.g., home runs) or rate statistics (e.g., on-base percentage). Counting statistics are often context-dependent, meaning that they are not sufficient for analyzing performance. For example, knowing that a player recorded 100 base hits in a season is not enough for evaluation. If the player needed only 300 at-bats, he had a successful season; if he used 500 at-bats, he may be out of a job. Thus, we can more accurately evaluate performance using measures prorated by the number of opportunities. The hitter with a .333 batting average (i.e., a 33.3% success rate) is far more productive than one with a .200 average. To score more runs than the opponent, a team should use its most productive hitters as often as possible (ignoring defensive ability for now); a team that gives a less productive hitter a significant amount of playing time is wasting opportunities.

The player map display (Figure 3) provides a way to quickly identify ineffective allocations of playing time. The treemap design was inspired by the SmartMoney Map of the Market [12]. Each rectangle represents a player, with the area encoding the number of plate appearances. The color (a green to red spectrum encodes high to low values) of each rectangle encodes a particular rate statistic, which can be selected by the user. (Some rate statistics are more effective at predicting run scoring [2]. Though batting average is the traditional measure of a hitter, many other rate statistics are more strongly correlated with run scoring.)

This display provides an overview of the efficiency of playing time allocation. A large highly-saturated red rectangle indicates a bad offensive player receiving too many opportunities; a small green rectangle could be a candidate for more playing time. The solution is not always simple—the small green rectangle may be



Figure 3: SportVis player map for the 2004 Boston Red Sox

an injured player or the large red rectangle may be a player with superb defensive skills.

3 FUTURE WORK

This version of SportVis only scratches the surface of the potential applications of information visualization to sports statistics, especially in baseball. For example, we only look at aggregate information for any particular game. Visualizing the details of an individual game can provide insight on how teams win or lose specific games. One approach could be to use a large-scale technique such as dot-plots or information murals to display down to plate appearance level (there were nearly 185,000 plate appearances in 2004).

Most of the visualization techniques used here reveal how games were won or lost, but they are not necessarily adequate for prediction. Many performance measures (e.g., batting average) are poor predictors [2], and small sample sizes make drawing conclusions difficult. Representation of significance or uncertainty has been an ongoing research topic in information visualization [3], and such techniques would be valuable here.

REFERENCES

- [1] C. Ahlberg and B. Shneiderman. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proc. of ACM CHI '94*, pages 313–317. ACM Press, 1994.
- [2] J. Albert and J. Bennett. *Curve Ball*. Copernicus Books, 2003.
- [3] R. Amar and J. Stasko. Knowledge precepts for design and evaluation of information visualizations. *IEEE TVCG*, 11(4):432–442, July/August 2005.
- [4] Baseball Databank. <http://www.baseball-databank.org/>.
- [5] L. Jin and D. C. Banks. Tennisviewer: A browser for competition trees. *IEEE Computer Graphics and Applications*, 17(4):63–65, 1997.
- [6] R. Rao and S. K. Card. The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proc. of ACM CHI '94*, 1994.
- [7] Retrosheet. <http://retrosheet.org/>.
- [8] Retrosheet. 2004 Boston Red Sox game log. <http://retrosheet.org/boxesetc/VBOS02004.htm>.
- [9] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. 1996 IEEE Visual Languages*. IEEE Press, 1996.
- [10] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2nd edition, 2001.
- [11] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufman, 2004.
- [12] M. Wattenberg. Visualizing the stock market. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, 1999.