Data Visualization: The Crossroads Between Computer Science Curriculum and Baseball

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Matthew Pezolt

Spring, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Rosanne Vrugtman, Department of Computer Science

Data Visualization: The Crossroads Between Computer Science Curriculum and Baseball

CS4491 Capstone Report, 2022

Matthew Pezolt
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
mjp7ss@virginia.edu

Abstract

USA Baseball, the national body for United States representation in international baseball competition, wanted to further improve its player identification capabilities through enhanced data collection and visualization methods. I designed a web-based database/interface that re-imagined the way USA Talent Identifiers were able to analyze athletes' quantitative performances. I developed the interface using python and its library of data analysis packages, as well as its web-based interface package called "Dash". These packages allowed me to recreate mainstream data analysis tools already common in the baseball industry as well as the freedom to further expand and generate new visualizations where I felt popular analytical tools were failing to paint the entire picture.

The project evolved and became a new staple for USA Baseball's data evaluation process for prospective athletes. It served as a great means to compare players in the national team selection process and also served as an advanced preparation tool for evaluating competition in international tournaments. The project is still undergoing further enhancements as I think of new ideas for expanding with the calculation of more traditional baseball statistics. I will be deploying and testing this database with the University's baseball program this season to further evaluate the long-term impact the project can have on player development at the collegiate level.

1 Introduction

"This team is on a mission. Not only to win, but to leave no doubt that we deserve to win." – Veronica Alvarez, Women's National Team Manager

Every year USA Baseball fields National Teams that go on to play in international competitions. These teams include the 12U, 15U, 18U, Collegiate, and Women's National Teams that all face off against the best of the best from other nations. These events may not have the global magnitude of the soccer World Cup, but there is a level of nationalism that develops during these events: A desire to prove that America is the premiere baseball power in the World at every level.

In order to fulfil this desire and satisfy the nationalism that stems from this environment, there is a vested interest in each team fielding the best possible roster. At each level there are entire years' worth of talent identification events and training camps all over the country in hopes of putting the best team on the field for international competition. There is a need for a standardized tool for comparing and evaluating players that not all members of the roster selection process have seen in person at every event. This is the root of why the development of a web-based database was so important, allowing all players to be evenly evaluated under a more

standardized talent identification criteria, leading to better overall rosters representing USA in international competition.

2. Related Work

The baseball industry is filled with leading-edge research in data-analysis, especially at the professional level. The multi-billion-dollar industry has 30 MLB organizations racing against each other to find some form of competitive advantage. According to Cox (2022), the front-line analytics in baseball today is centered around Machine Learning and the ability to accurately predict what was previously guessed at [1]. Based on conversations I have had with industry professionals; the research is being applied to pitcher delivery mechanics to find "tips" or "tells" indicating the type of pitch that is coming.

This leading-edge research in machine learning is paired with more conventional pitching/hitting movement and result analysis to form one of the fastest growing fields of research and innovation in the world. While the availability of machine learning resources is limited to mainly professional clubs, Verducci (2022) posits that the reach of this information gathering technology such as Trackman, Rapsodo, and slow-motion cameras has gone beyond the professional domain in recent years [2].

Data literacy and analysis has become a hot topic in the amateur baseball world as well. High school programs and private instructors have adopted the technology to begin the metric optimization process at incredibly young ages. Hayhurst (2022) makes the point that College programs (especially large Division I programs) are using quantitative analysis to drive player development and in-game decision making in ways that have never been seen before at the amateur level [3]. Accompanying this rise in data availability, is the need for efficient analysis. Programs have resorted to outsourcing their data processing and analysis to platforms such as BaseballCloud (2022), paying thousands of dollars for the resources the platform provides [4].

This data boom has infiltrated amateur baseball so heavily that it would be foolish for an organization such as USA Baseball to not take advantage of the benefits the data has to offer. My project provided a way for USA Baseball to process the tens of thousands of data points collected throughout the evaluation process without having to spend the thousands of dollars that outsourcing the analysis would cost.

3. Process Design

The backbone of my web-based database was built using Dash, Python's framework equivalent to R's ShinyApp, which was how USA Baseball hosted their existing database. I chose to use Dash because it offers an extensive library of documentation and has plenty of useful tutorials online. This was my first dive into web interface design and learning how to establish the interface was an early challenge that was made much easier with the availability of the online resources for the framework. The framework also includes packages for page layout and formatting that are parallel to HTML. Being able to incorporate HTML website design with traditional python code made the aesthetic design of the interface much easier, as seen in Figure 1.

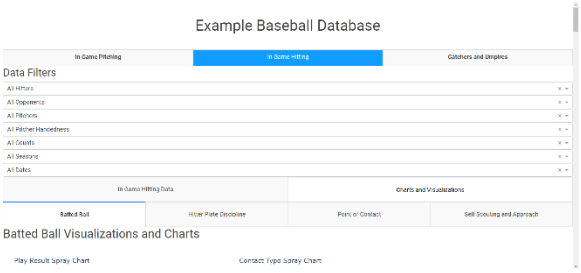

Figure 1: Web Interface Skeleton

The project was entirely self-motivated. At the beginning of the Summer, I was told that I had complete creative control over data analysis. As a result of this I was able to fully immerse myself in the development process. Using raw data collected from games and talent evaluation events throughout the summer, I was able to really dive into the meanings of metrics and brainstorm ways to visualize that data, leading to useful and meaningful graphs and evaluations.

I visualized the data using the plotly package in Python (also available in R). I chose this because it was the recommended library to combine with Dash and because I had experience using it through my course work at the University of Virginia. Plotly allows for the creation of interactive graphs that can be styled according to common baseball visualizations found in databases such as BaseballCloud (2022) [4]. It also allowed me to create basic traces that make the data more understandable, such as being able to include an outline of a field on spray charts (see figure 2) so viewers understand hit locations under the context of a field's dimensions as well as a strike zone so viewers can assess balls/strikes while looking at the pitch locations.

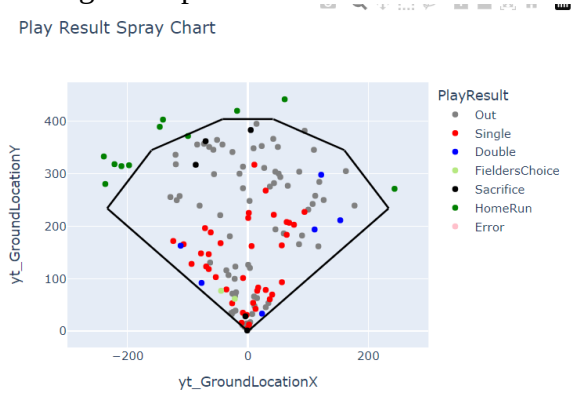

Figure 2: Example of Spray Chart with Field Trace

4. Results

The project was a very fulfilling experience over the course of the Summer. I was able to roll out my first edition of the database. From there, it was used in the player analysis process during roster evaluation for the 15U National Team that won a gold medal in the 15U World Cup, our 16U/17U National Team Development Program, and our 13U/14U Athlete Development Program. It was also used in the post team evaluation of the Collegiate National Team.

I have brought the database with me here to UVA this year and am utilizing it with our Baseball Program to create a larger emphasis on data driven player development. The goal is to best apply our data to player development for the purpose of winning games. I also am taking feedback from coaches here and undertaking an independent study surrounding baseball data visualization to take this database to the next level and maximize the results seen on the field next spring.

5. Conclusion

This project offered a new means of standardizing the player evaluation process for USA Baseball. The ability to compare player performance after the fact with concrete data has given USA Baseball a valuable means of determining rosters for international competition. As a result, we have seen fantastic success, with the 12U, 15U, and 18U teams winning their respective World Cups; the Collegiate National Team winning a bronze medal in their international tournament; and the Women's National Team winning their international friendly series vs Canada.

6. Future Work

As the state of baseball analytics is an ever-evolving field, this project remains a work in progress. From a USA Baseball perspective, I am looking to incorporate more than just the in-game performance data. I hope to incorporate practice data, as

well as our PDP Assessment data (a series of tests gauging things from reaction time, jumping ability, speed, and information processing that USA Baseball athletes are put through). I hope this information can also be useful in the roster construction process here in the future. I have also brought this project with me to UVA for the baseball program here. I hope to continue its in-game data analysis development and help apply the findings to our data driven player development here as well.

References

[1] – Cox. 2022. How AI and Machine Learning are Revolutionizing Baseball. (May 2022). Retrieved September 22, 2022. https://www.cox.com/residential/articles/ai-machine-learning-baseball.html

[2] – Verducci, T. 2022. From Trackman to Edgertronic to Rapsodo, the Tech Boom Is Fundamentally Altering Baseball. (March 2019). Retrieved September 22, 2022 from https://www.si.com/mlb/2019/03/29/technology-revolution-baseball-trackman-edgertronic-rapsodo

[3] – Hayhurst, C. 2022. EdTech: Data Analytics Helps College Coaches and Athletes Optimize Training and Performance. (August 2019). Retrieved September 22, 2022 from https://edtechmagazine.com/higher/article/2019/08/data-analytics-helps-college-coaches-and-athletes-optimize-training-and-performance

[4] BaseballCloud. 2022. BaseballCloud: New Age Solutions for Old School Problems. Retrieved from https://www.baseballcloud.com/