Enhancing Sports Performance and Strategy

Through Big Data Analytics:

A Case Study of STATCAST in MLB Baseball Teams

Yen Chun Lin

American University

ITEC 670 Database and Big Data

Professor Dr. Frank Armour

April 29 ,2024

Outline

1. **Abstract**
2. **Introduction**
    A. Big data
    B. STATCAST system
    C. Objective
3. **The Challenges and Benefits of Big Data Analysis in Sports**
    A. Challenges of big data analysis in sports
    B. Benefits of big data analysis in sports
4. **Data collection**
5. **The application of STATCAST**
    A. The application of STATCAST for general managers
    B. The application of STATCAST for coaches and players
6. **Example Use Case**
    A. The functions and query system of STATCAST
    B. Example: Batting Information
    C. Example: Pitching Information
7. **Conclusion**
8. **References**

Abstract

The application of big data analytics in sports offers significant benefits for athletes aiming to optimize their performance. The analysis yields a range of statistical insights, including the visualization of player performance and the identification of standout performers. By leveraging these analysis results, team managers gain a reliable basis for identifying and targeting suitable players to enhance team composition and strategic competitiveness.

Moreover, big data analytics have the capacity to unveil the strengths and weaknesses of players, furnishing invaluable insights for both athletes and coaches to optimize individual growth and bolster overall team dynamics.

This study undertakes an investigation into the potential advantages of advanced data analytics in the realm of sports, specifically concentrating on the baseball domain and the utilization of STATCAST technology. The objective is to elucidate how these analytics contribute to the enhancement of sports performance and strategy within Major League Baseball (MLB) teams. Through a thorough analysis of extensive datasets, STATCAST yields a plethora of invaluable information pertinent to player performance assessment and strategic decision-making.

Introduction

Big data:

Big data involves a diverse range of data types, including structured, semi-structured, and unstructured data, which organizations collect, analyze, and explore to extract valuable information and insights. Big data finds wide application in machine learning projects, predictive modeling endeavors, and diverse advanced analytics applications (Hashemi-Pour et al., 2024). The three foundational attributes commonly associated with big data, commonly referred to as the three V's, include volume, velocity, and variety. The volume of data is particularly significant, often representing vast quantities of information. In relation to variety, big data encompasses diverse types of data, such as unstructured, structured, and semi-structured data, which exist across heterogeneous environments and data formats. Additionally, high velocity refers to the rapid pace at which data is generated, collected, and processed within big data systems. (Tiao, 2024).

STATCAST system:

STATCAST is an advanced automated tool designed to analyze player movements and athletic capabilities in Major League Baseball (MLB) with high speed and precision. It epitomizes a sophisticated tracking technology that facilitates the

comprehensive collection and analysis of voluminous baseball data. Its operation involves a set of cameras and radar systems installed in each MLB stadium, enabling meticulous monitoring, and recording of every on-field movement. STATCAST systematically tracks and quantifies various facets of gameplay, encompassing pitching metrics (velocity, spin rate), hitting dynamics (exit velocity, launch angle), running statistics (sprint speed, base-to-base times), and fielding attributes (catch probability, catcher pop time). Importantly, while these granular data points are accrued at a play-by-play level, they concurrently underpin the development of longitudinal player statistics, such as batters' hard-hit rates or fielders' Outs Above Average, across multiple seasons (*Statcast: Glossary*).

Objective:

The objective of this study is to investigate the potential of big data analysis within the context of baseball. In contrast to traditional baseball statistics, which typically provide basic metrics such as the count of home runs or batting averages for players, big data analysis provides a more comprehensive understanding of each play or hit. For instance, big data analysis can unveil nuanced batting or pitching insights for individual players, such as a players' capability at connecting with high balls within the strike zone or a pitcher's ability to consistently induce swings and misses with their curveball. This information not only aids players in devising strategies to

defeat their opponents but also furnishes invaluable insights for general managers throughout various stages such as the draft, trade market, and free agency.

The Challenges and Benefits of Big Data Analysis in Sports

Challenges of big data analysis in sports:

High dimensionality and substantial sample size are hallmark features of big data, precipitating three prominent challenges. Firstly, the accumulation of noise, the emergence of false correlations, and unintentional uniformity might be caused by the increased dimensionality. Secondly, the convergence of high dimensionality and large sample size can introduce issues such as significant computational costs and algorithmic instability. Thirdly, the expansive datasets in big data are frequently sourced from diverse origins at disparate time intervals, employing different technologies. These challenges might lead to problems such as experimental disparities, statistical biases, and the need to provide more flexible and robust methodologies (Fan et al., 2014).

Benefits of big data analysis in sports:

Sports competitions typically yield substantial volumes of player-related data. Leveraging big data technology enables the conversion of this data into actionable insights for athletes. Athletes can improve their performance by obtaining better game

plans or training methods through the information derived from the analysis results of big data. The escalating demand for sports statistics has positioned big data as the premier technology for sports analytics, thereby enhancing the sports domain to a heightened plane (Dmonte & Dmello, 2017).

Data Collection

STATCAST captures data points in baseball such as pitching, hitting, defensive rating, and running speed using camera and radar systems installed in all 30 MLB stadiums. Nearly seven terabytes of data are recorded by the system during each game (Wittenberg, 2021). When a batter swings the bat, the camera and radar system are triggered to record relevant information regarding the bat, including details such as exit velocity and launch angle. After capturing this information, the system proceeds to store the collected data for further application. Teams have the capability to access the dataset through the application programming interface (API) owned by Major League Baseball (MLB) (Wittenberg, 2021).

pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	batter	pitcher	events	description
SI	2020-09-18T00:00:00.000+0000	91	1.59	5.02	Sherriff, Ryan	600524	595411	field_out	hit_into_play
SI	2020-09-18T00:00:00.000+0000	90.8	1.57	5	Sherriff, Ryan	600524	595411	NaN	foul
SI	2020-09-18T00:00:00.000+0000	91.2	1.8	4.95	Sherriff, Ryan	600524	595411	NaN	ball
SI	2020-09-18T00:00:00.000+0000	91.4	1.83	4.81	Sherriff, Ryan	600524	595411	NaN	ball
SI	2020-09-18T00:00:00.000+0000	91	1.69	4.93	Sherriff, Ryan	600524	595411	NaN	called_strike
SI	2020-09-18T00:00:00.000+0000	90.5	1.74	4.84	Sherriff, Ryan	669720	595411	field_out	hit_into_play
SI	2020-09-18T00:00:00.000+0000	91.8	1.7	4.96	Sherriff, Ryan	669720	595411	NaN	called_strike
SI	2020-09-18T00:00:00.000+0000	89.7	1.6	4.95	Sherriff, Ryan	578428	595411	field_out	hit_into_play
SI	2020-09-18T00:00:00.000+0000	89.8	1.61	5.01	Sherriff, Ryan	578428	595411	NaN	called_strike
FF	2020-09-18T00:00:00.000+0000	95	2.9	5.38	Scott, Tanner	664040	656945	field_out	hit_into_play

Figure 1. Sample of data collected by STATCAST (Wittenberg, 2021)

The application of STATCAST

The film "Moneyball" serves as a notable illustration of the application of big data analysis in sports. It utilized data analysis techniques to identify undervalued players and strategically assemble a competitive lineup while adhering to a constrained budget. This will serve as an excellent illustration of how the STATCAST system could provide significant benefits to MLB teams.

The application of STATCAST for general managers:

The primary objective for general managers of each team is to construct an optimal roster that can achieve peak performance throughout the season. They are tasked with identifying potential strengths and weaknesses within their roster and aiming for a balanced offense-defense team composition. Additionally, they must assess whether trading players would enhance their team's chances of winning. In this context, STATCAST could serve as a valuable and indispensable tool. General managers can leverage STATCAST's extensive data analyses and visualization graphs to determine which players are most suitable for their team and to identify the most optimal contracts for these players.

The application of STATCAST for coaches and players:

Furthermore, STATCAST can prove beneficial for both players and coaches. For instance, pitchers can utilize STATCAST data to discern that their curveballs are more prone to resulting in home runs compared to other pitch types. This insight can prompt players to prioritize enhancing their pitching performance with curveballs during training sessions.

Similarly, STATCAST might reveal that the weakness of the opponent lies in left-handed batters struggling against forkballs delivered by left-handed pitchers. In response, coaches could strategize by deploying a left-handed pitcher as the starting pitcher to exploit this vulnerability and secure a strong performance in the early innings of the game.

Example Use Case

The functions and query system of STATCAST:

One of the advantages of STATCAST for baseball is its capacity to visually represent complex data in coherent patterns. Figure 2 exemplifies the diverse analysis formats that STATCAST offers. Information pertaining to players' batting, pitching, defense, and overall team travel schedule is accessible on the website.

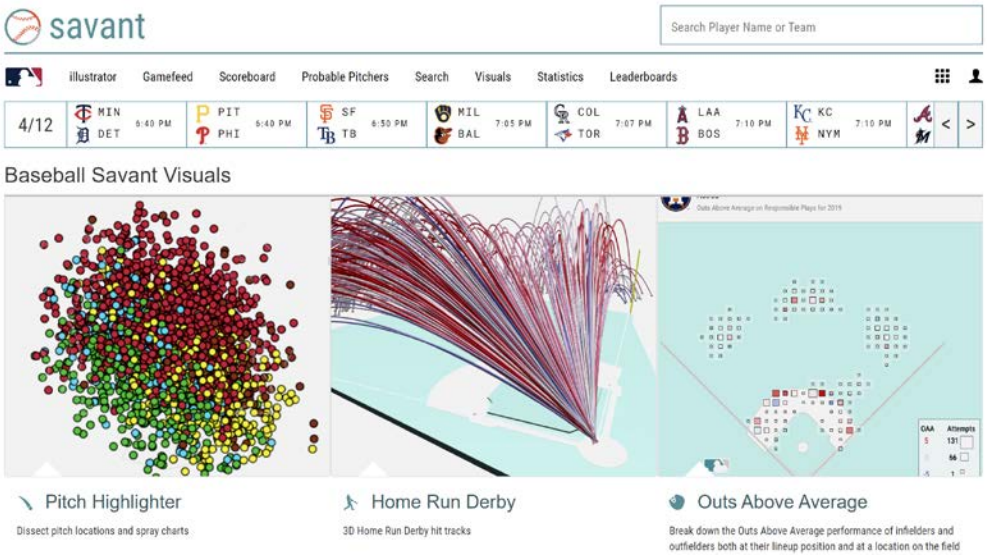


Figure 2. Example of visual graphs on the MLB STATCAST website (baseballsavant.mlb.com).

Additionally, STATCAST provides a robust query system for users. Through this query system, users can filter and explore information related to different players and their performances. Figure 3 demonstrates the query system available within the STATCAST platform, showcasing how users can utilize it to refine the information they are seeking.

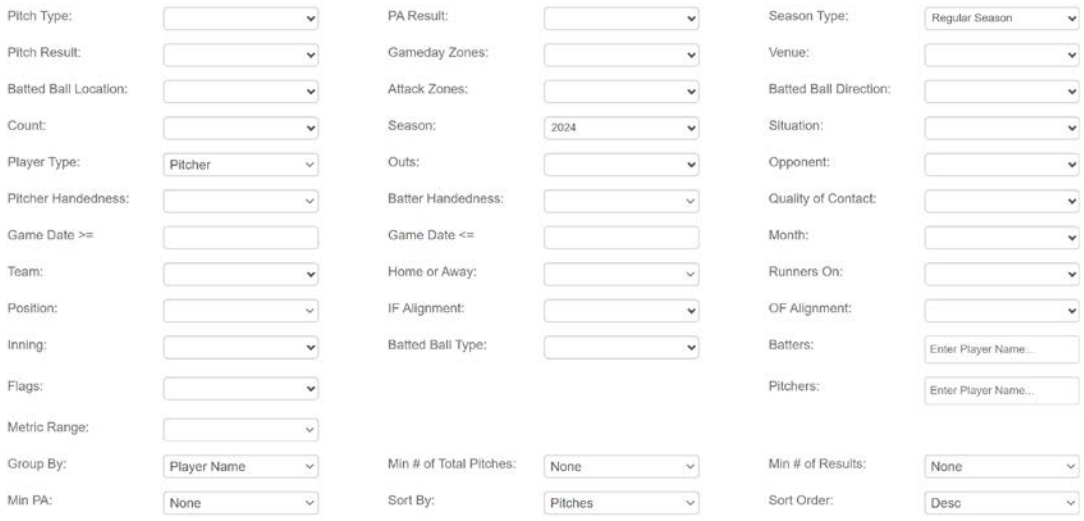
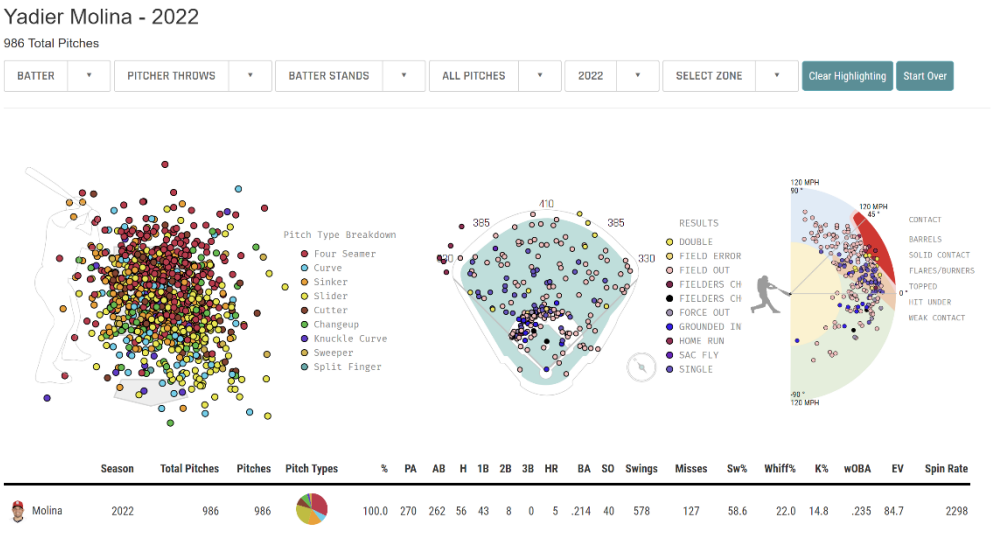


Figure 3. Query system in STATCST from MLB website (baseballsavant.mlb.com).

Example - Batting Information:

Upon executing a query, users gain the ability to filter and explore the data. Figure 4 exemplifies the utilization of the "Pitch Highlighter" function to gather insights into Yadier Molina's batting performance against different pitch types. To refine the results further, users have the option to select from various filter metrics. For example, Figure 5 illustrates the process of filtering Molina's batting performance against fastballs thrown by right-handed pitchers during the 2022 season. The analysis results unveiled by STATCAST enable both players and coaches to discern that Molina exhibited superior batting performance when confronting sliders compared to curveballs during the 2022 season. This insight holds potential significance in crafting an optimal training regimen and directing efforts towards addressing weaknesses. Conversely, it also furnishes opponents with valuable insights to formulate game strategies when competing against the player.



	Season	Total Pitches	Pitches	Pitch Types	%	PA	AB	H	1B	2B	3B	HR	BA	SO	Swings	Misses	Sw%	Whiff%	K%	wOBA	EV	Spin Rate
Molina	2022	986	986		100.0	270	262	56	43	8	0	5	.214	40	578	127	58.6	22.0	14.8	.235	84.7	2298

Figure 4. Example of Yadier Molina's batting performance against different pitch types in the 2022 MLB season (baseballsavant.mlb.com).

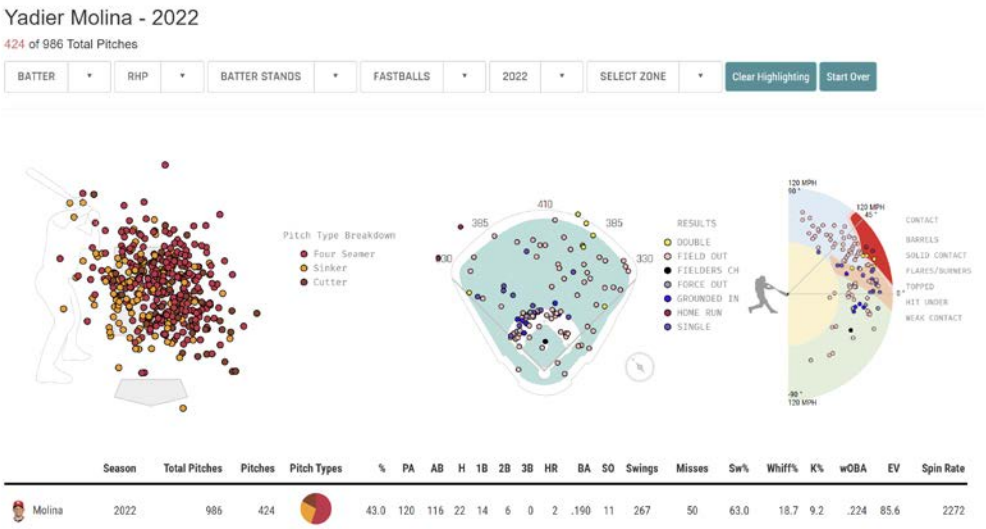


Figure 5. Example of Yadier Molina’s batting performance against fastballs thrown by right-handed pitchers during the 2022 MLB season (baseballsavant.mlb.com).

Example - Pitching Information:

By employing the "3D Pitch Tracks" function, users can attain comprehensive insights into pitchers' pitching details. STATCAST provides fundamental pitching information. By utilizing filter metrics, users can obtain more detailed information for each pitch thrown by the pitchers. For instance, details such as the pitch type, the speed and spin for each pitch, and the outcome of every pitch can be elucidated using the filter metrics. Upon executing a query, users can access detailed information for every pitch. Figure 6 delineates the pitching data of Adam Wainwright, enabling users to apply filter metrics to select specific game data and acquire desired information.

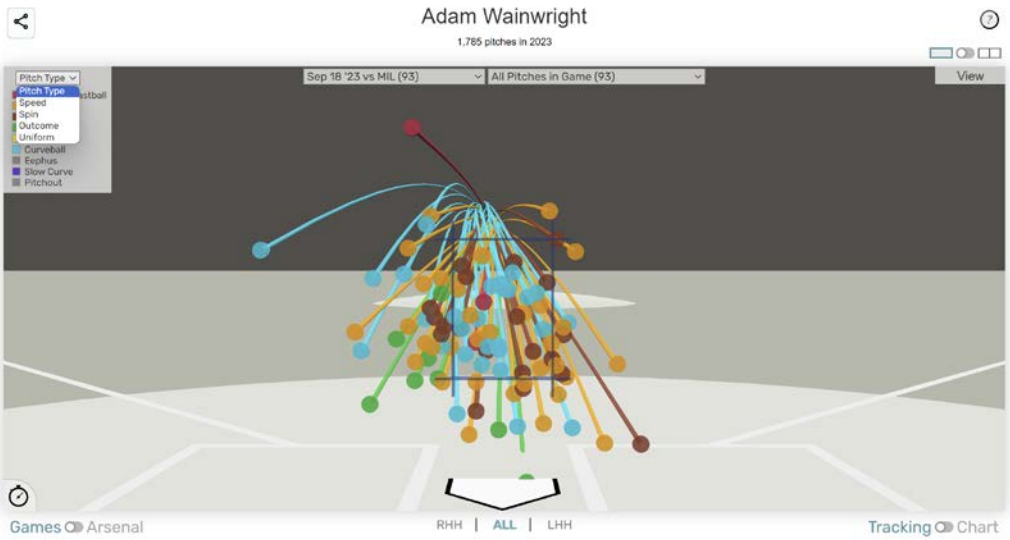


Figure 6. Example of Adam Wainwright's pitching performance (baseballsavant.mlb.com).

Conclusion

The application of big data analysis offers numerous advantages to sports, and STATCAST exemplifies this application within Major League Baseball. STATCAST represents an advanced analysis system that surpasses traditional baseball datasets in the scope and depth of information it provides. Users can access sophisticated data derived from big data analysis, and visual graphs facilitate the comprehension of complex datasets, aiding in the interpretation of information.

General managers can utilize big data to make informed decisions regarding drafting, trading, and free agency, thereby enhancing team strength and performance. Coaches can develop more effective game plans to achieve greater success within a season, while players can leverage analysis results to identify specific areas for improvement.

The application of big data analysis in sports has extended beyond its focus on player performance, significantly altering the operational strategies of organizations within the sports industry. For instance, to effectively manage the expanding customer database and increase fan attendance at games, organizations leverage big data analysis tools to develop sophisticated strategies aimed at engaging the audience and boosting ticket sales. This integration of big data analytics has reshaped business practices, enabling organizations to optimize their operations and enhance fan experiences (Watanabe et al., 2021).

These benefits highlight the significance of big data analysis in sports, contributing to the continual enhancement of the sports industry through the provision of potent tools for advancement.

References

Dmonte, R., & Dmello, A. (2017, January 1). Big Data in sports. https://www.ijert.org/research/big-data-in-sports-IJERTV6IS010289.pdf

Fan, J., Han, F., & Liu, H. (2014, February 5). Challenges of Big Data analysis. Academic.oup.com. https://academic.oup.com/nsr/article/1/2/293/1397586

Hashemi-Pour, C., Botelho, B., & Bigelow, S. J. (2024, March 21). What is Big Data and why is it important?: Definition from TechTarget. Data Management. https://www.techtarget.com/searchdatamanagement/definition/big-data

Kaur, A., Kaur, R., & Jagdev, G. (2021, March 31). Analyzing and exploring the impact of Big Data Analytics in sports sector - SN computer science. SpringerLink. https://link.springer.com/article/10.1007/s42979-021-00575-y

Lage, M., Ono, J. P., Cervone, D., Chiang, J., Dietrich, C., & Silva, C. (2016, September 29). StatCast Dashboard: Exploration of Spatiotemporal Baseball Data. IEEE. https://ieeexplore.ieee.org/abstract/document/7579419

Statcast: Glossary. MLB.com. (n.d.-a). https://www.mlb.com/glossary/statcast

Tiao, S. (2024, March 11). What is Big Data?. Oracle. https://www.oracle.com/big-data/what-is-big-data/

Watanabe, N. M., Shapiro, S., & Drayer, J. (2021, April 20). Big Data and Analytics in Sport Management. Human Kinetics. https://journals.humankinetics.com/view/journals/jsm/35/3/article-p197.xml?content=fulltext

Wittenberg, M. (2021, October 28). Moneyball 2.0: Real-time decision making with MLB's Statcast Data. Databricks. https://www.databricks.com/blog/2021/10/28/moneyball-2-0-real-time-decision-making-with-mlbs-statcast-data.html