Introduction

The advent of big data has fundamentally transformed the sport of baseball. What began as a data-driven movement spearheaded by amateur statisticians to quantify the value of players and strategies has evolved into a full-fledged analytics revolution embraced across professional baseball organizations. This rapid penetration of data-centric processes in baseball has been enabled by advances in sensor technologies, computation, and database infrastructure, as elucidated extensively in the literature.

Several papers discuss how high-resolution optical recognition systems and radar technologies like Statcast now capture millions of data points per game on factors ranging from pitch speed and trajectory to fielder positioning and sprint velocity (Bai & Bai, 2021; Mizels et al., 2022). Integrated with traditional stats and injury indicators, these vast datasets are leveraged using sophisticated machine learning algorithms to extract nuanced performance insights and predict outcomes like career longevity or injury susceptibility (Dmonte & Dmello, 2017). Healey (2017) was also using Statcast data from an earlier season to prove that past results classified by specific weights based on criteria such as launch angle of hits were more predictive of future success for offensive performers than standard outcome-based statistics.

These analytic techniques have demonstrated tangible impacts on team strategy, player evaluations, roster optimization, and fan engagement initiatives (Watanabe et al., 2021). By quantifying the multidimensional attributes of players, managers can precisely assess talent and optimize lineups to extract maximum tactical advantage - seen in Oakland A's strategic roster building approach highlighted in Moneyball (Dmonte & Dmello, 2017). Granular physical motion indicators allow targeted biomechanical interventions and training adjustments to improve technique and reduce injury risk (Mizels et al., 2022). Descriptive analytics reports even enhance fan experience by providing detailed interactive infographics on player performance. Oh, Han, and Kim (2021) further discussed fan engagement techniques, looking to find correlations between fan attendance and enjoyment with the stadium and gameday experience.

However, ethical questions around usage transparency, predictive model interpretability, and data privacy safeguards have also surfaced, requiring careful regulatory oversight (Watanabe et al., 2021). Overall, the literature extensively covers how baseball has vigorously embraced analytics, demonstrating the transformative potential when cutting-edge technology, statistical rigor, and strategic decision-making intersect. But optimally harnessing big data innovations to elevate performance, engagement, and safety while mitigating risks remains an ongoing pursuit for baseball stakeholders.

“Data analytics and performance: The moderating role of intuition-based HR management in major league baseball” by Kim et al (2021) had the specific objective of evaluating data analytics and performance in the light of the moderating role of intuition-based Human Resource (HR) management in Major League Baseball (MLB). Kim et al (2021, p.204) has identified trends in decreased spectrum of social simplicity or available strategies and diminishing mechanisms of value creation or causal clarity that are directly caused by proportionate increase in reliance on data-driven decisions in highly competitive and specialized industries which have contributed into corresponding directly proportionate diminishing of positive effects of social capital due to increasing use of data-driven decision making during deployment of human resources.

“A multidisciplinary perspective on publicly available sports data in the era of big data: a scoping review of the literature on Major League Baseball” by Huang and Hsu, published in 2021 had specific objective of applying multidisciplinary perspective on publicly available sports data to understand development of data-driven baseball

research that satisfies application domains of big data maturity model. The research paper can provide insights into suitability of multidisciplinary perspectives, and especially biomechanical data in MLB.

Challenges

Scalability

A major challenge in sports analytics is dealing with the immense scale of data being generated. As discussed in the Watanabe et al. (2021) paper, sports events like the Super Bowl can generate millions of social media posts. Collecting, storing, and analyzing such a high volume of unstructured data presents scalability issues. Traditional data management and analysis methods struggle with this volume. Scaling up the computational infrastructure to handle large volumes in a cost-effective manner is an ongoing challenge.

Heterogeneity

The variety and heterogeneity of data in sports analytics also poses challenges. As highlighted in the Bai and Bai (2021) review, sports data contains diverse modalities like videos, sensor data, text, tabular statistics etc. Integrating such multimodal data and making sense of unstructured formats like video and text adds to the complexity. Lack of standardized formats and semantics makes combining different data sources difficult. Oh, Han, and Kim (2021) similarly discussed the difficulties of combining both manually-collected interviews and unstructured data from social media platforms into a coherent data narrative. Kim et al (2021) also bring up that we discover evidence of significant fan preference heterogeneity across Major League Baseball markets, toward both home and visiting club quality, using a generalized linear mixed model.

Velocity

The speed at which sports data is generated also creates big data challenges. As discussed in the Dmonte and Dmello (2017) paper, real-time IoT sensors and tracking technologies produce streaming data at high velocity during sporting events. Healey (2017) also discussed velocity of data collection via MLB's proprietary sensors, and also mentioned how velocity would get much higher than the 2014 data set used. The analytics has to keep pace with this volume and velocity of generation. Building scalable and flexible data pipelines to ingest and process high-velocity streaming data is an open challenge.

Veracity

Ensuring the quality and veracity of sports data is not easy given the diversity of sources. As the Mizels et al. (2022) review highlights, even basic metrics like errors in baseball have human judgment involved. Ensuring the reliability and accuracy of performance data from various providers is an ongoing concern. Cleaning noisy multimodal data at scale remains challenging.

Privacy

The Watanabe et al. (2021) paper highlights privacy as an emerging challenge with the proliferation of personal and biometric data. Protecting athlete privacy, preventing leaks of health and performance data, and ensuring ethical use of analytics is of utmost importance. Appropriate de-identification, access control, and governance frameworks need to be implemented.

Model Interpretability

As machine learning is increasingly used for prediction and decision-making in sports, explaining model predictions and ensuring transparency is vital. Sports teams need to trust and interpret the underlying logic. Lack of model interpretability and auditability remains a challenge as highlighted in the Bai and Bai (2021) review. Adoption of

explainable AI techniques is an active area of research.Computing and cloud technologies make firms for example MLB's to perceive big data analytics as part of artificial intelligence and machine learning which is consistent with application of biomechanical data on sensing. Researchers put claims that managers of MLB's have potential to apply artificial intelligence which includes scope of current research on biomechanical data which is consistent with application of artificial intelligence identified by Tambe, Cappeli and Yakubovich (2019).

Approaches

Data Collection

Sensor-based tracking: The Mizels et al. (2022) and Bai & Bai (2021) papers discuss the use of high-resolution optical recognition and radar systems like Statcast to collect millions of data points on factors like spin rate, launch angle, sprint speed etc. during games. Healey (2017) similarly used sensor data from Statcast and official, proprietary MLB systems.
Injury databases: Mizels et al. (2022) describe the Health Injury Tracking System (HITS) created by MLB to systematically record injury epidemiology data and return-to-play timelines.Kim et al. (2021) mentions about the data-driven injury where one essential component of data-driven injury prevention in sports, including baseball, is the gathering of biometric data.
Wearable IMUs: Emerging wearable technology using inertial measurement units are proposed for prospectively gathering biomechanical data on factors like torque and arm speed during pitching (Mizels et al., 2022).
Web scraping: The Watanabe et al. (2021) paper mentions scraping routines to collect large volumes of textual data from public websites like Twitter and Yelp for analysis. Oh, Han, and Kim (2021) also performed data scraping against social media sites including Twitter and Facebook for analysis.
Surveys: Dmonte & Dmello (2017) state that data collection in sports relies on inputs from various sensors, cameras, trackers, and surveys. Oh, Han, and Kim (2021) used focus groups interviews to gather initial data for their subject.

Data Storage and Management

Relational databases: Dmonte & Dmello (2017) discuss the use of column-oriented, in-memory relational database systems like SAP HANA by major leagues to handle massive structured and unstructured data.David Marshall (2021) similarly used a relational database.
Data warehouses: The NBA statistics site described in Dmonte & Dmello (2017) stores decades of basketball data in a centralized data warehouse. Oh, Han, and Kim (2021) also used a data warehouse for scraped data from social networks.
Distributed storage: Bai & Bai (2021) mention Hadoop-based distributed storage to scale sports data platforms.

Data Analysis

Statistical analysis: Bai & Bai (2021) discuss the use of statistical techniques like hypothesis testing, clustering, regression modeling etc. to find patterns. Healey (2017) used Bayesian analysis to determine correlations
Machine learning: Mizels et al. (2022) describe the use of ML algorithms to predict injury risk and location based on performance data.
Artificial Intelligence: Kim et al (2021) mentions the important domains where AI is significantly influencing MLB, influencing the sport's future and ushering in a new era of baseball.

Social network analysis: Watanabe et al. (2021) mention analyzing interactions between individuals in sports communities using network analysis techniques. Yu-Chia (2021) also used analyses of social network interactions. Kim et al (2021) says that the Big sports data has led to the expansion of both the Internet and sports. Analysts can often extract vast amounts of data for use by the media, fans, athletes, and organizations for all major sports. These initiatives usually work in tandem with leading technology companies that have come to understand the tremendous advantages of sports analytics.
Visualization: Dmonte & Dmello (2017) state that data visualization tools are important for exploring and deriving insights from sports data.
Correlation: Huang and Hsu (2021) describes the Correlation analysis, which is a tool used could examine the relationship between a pitcher's ERA and strikeout rate.

Project Proposal

The recent studies by Mizels et al., Zhongbo Bai and Xiaomei Bai, Ruth Dmonte and Asher Dmello, and Watanabe et al. show that data analysis in baseball has changed a lot, giving teams and fans more information from the vast amount of data collected during games. Our project, "Baseball Analytics: A Deep Dive into Baseball Data Using an Integrated Tech Stack," aims to take this analysis further by using big datasets from MLB and other sources. This will help give a better understanding of the game through data.

Our project also addresses the challenges mentioned in these studies, like making sure the data is accurate, organizing the data well, and dealing with different types of data from various sources. By using a combination of tech tools like Kafka, AWS, and FastAPI, we plan to overcome these challenges and provide real-time and after-game analysis. This will help teams, coaches, and fans get more insight into the game, helping with strategy, player performance review, and making the game more enjoyable for fans. Through this project, we aim to add to the growing field of sports data analysis, especially in baseball, and provide useful tools for better game analysis and enjoyment.

Sources Cited(Papers reviewed)

Bai, Z., & Bai, X. (2021). Sports big data: Management, analysis, applications, and challenges. *Complexity*, *2021*, 1–11. https://doi.org/10.1155/2021/6676297

Dmonte, R., & Dmello, A. (2017). Big Data in Sports : Leverage Big Data in Sports: An Insight using SAP HANA. *International Journal of Engineering Research & Technology (IJERT)*, *6*(1), 380–383.

Healey, G. (2017). Learning, visualizing, and assessing a model for the intrinsic value of a batted ball. *IEEE Access*, *5*, 13811–13822. https://doi.org/10.1109/access.2017.2728663

Mizels, J., Erickson, B., & Chalmers, P. (2022). Current state of data and analytics research in baseball. *Current Reviews in Musculoskeletal Medicine*, *15*(4), 283–290. https://doi.org/10.1007/s12178-022-09763-6

Oh, S. W., Han, J., & Kim, G. H. (2021). A study of investigating on multi-environmental factors of professional baseball stadium: Using big data analysis. *The Korean Journal of Physical Education*, *60*(4), 145–158. https://doi.org/10.23949/kjpe.2021.7.60.4.11

Watanabe, N. M., Shapiro, S., & Drayer, J. (2021). Big Data and Analytics in Sport Management. *Journal of Sport Management*, *35*(3), 197–202. https://doi.org/10.1123/jsm.2021-0067

Jaemin Kim, Clay Dibrell, Ellen Kraft, David Marshall (2021). Data analytics and performance: The moderating role of intuition-based HR management in major league baseball. Journal of Business Research, Volume 122, Issue 1, Pages 204-216. Doi: https://doi.org/10.1016/j.jbusres.2020.08.057

Huang, Jyh-How and Hsu, Yu-Chia (2021). "A Multidisciplinary Perspective on Publicly Available Sports Data in the Era of Big Data: A Scoping Review of the Literature on Major League Baseball". SAGE Open, vol. 11(4), pages 1-20, 21582440211, November