**Brendan Mapes**

**Master's Thesis**

**Quantitative Methods of the Social Sciences**

**Columbia University | June 30th, 2022**

# *A Machine Learning Approach to Seam Shifted Wake*

## Abstract

A machine learning based strategy is proposed to solve a missing data problem in Major League Baseball's publicly available Hawk-Eye data, specifically as it relates to pitch spin. Data related to the components of the pitched ball's unit spin vector is currently not available on a pitch-by-pitch basis but is vital to the baseball research community's understanding of a newly discovered force acting on the ball, the Seam-Shifted Wake (SSW). Random forest regression and multivariate adaptive regression splines are used to predict values for the unit spin vector on a pitch-by-pitch basis. Predicted values for the spin vector components provide necessary foundation for fully characterizing the SSW force. This work serves as a methodological foundation for further application of supervised learning techniques to the missing data problems related to the Seam-Shifted Wake force in baseball.

## Introduction

The game of baseball provides a perfect opportunity for statisticians and data scientists to apply their skills on real-world data. The nature of the game provides plenty of statistics to analyze, and Major League Baseball has made loads of data available online. Since Michael Lewis' *Moneyball* book was published in 2003, nearly every team in the MLB has made large investments in data science, looking to find competitive advantages like those *Moneyball* Oakland Athletics. Because teams like the Athletics cannot compete with the player payrolls of wealthier franchises like the New York Yankees, they have been forced to innovate elsewhere, with advanced analytics. They are forced to think outside the box, challenge the status quo, and take risks that richer teams have the luxury of avoiding. Data science is the tool helping these clubs take informed risks and challenge the status quo, all while thinking scientifically.

The research explained throughout this thesis is one application of the data science toolkit to baseball with legitimate practical relevance to MLB organizations. This work is also timely. The aerodynamic force at the heart of this research, known as Seam-Shifted Wake (SSW), was introduced as recently as 2020 and is still not entirely understood by baseball physicists [2]. This thesis will investigate and expand upon the baseball community's understanding of Seam-Shifted Wake. After demonstrating the analytical methods used to identify pitches being affected by the SSW force, it will propose a machine learning strategy to solve for components of the spin vector of a pitched baseball which is necessary for more detailed understanding of Seam-Shifted Wake's true effect on the ball.

## Background

At the start of the 2020 season, MLB switched from the radar-based technology it had been using to collect in-game data, known as Trackman, to an optical tracking technology system, known as Hawk-Eye. Hawk-Eye optical sensors provide analysts data related to pitcher and batter mechanics, pitch execution, batted-ball trajectory, and defensive metrics [9].

> "The Hawk-Eye system tracks the field of play using 12 high resolution, high-frame rate video cameras installed at each ballpark. Video from these cameras is synchronized and analyzed to detect and track ball and player movement. In addition to previously available data sets, the new platform provides real-time player pose and motion analysis by measuring multiple points on the body thirty times per second," [9].

The new tracking system provides data at a much more detailed level. It unlocked new insights in the research field of baseball aerodynamics, specifically related to pitch movement. MLB's upgrade to Hawk-Eye has finally allowed for analysts to test a list of hypotheses from MLB pitchers on why their pitches move the way they do. Pitchers are uniquely positioned to apply academic-style research to their game and get immediate feedback, and they truly benefit from the research. Unlike the in-game actions of hitters, fielders, or managers, those of pitchers are not reactionary. They initiate each play with the pitch they throw and have full control over most variables affecting pitch movement, ideally. Pitchers have performed millions of their own experiments on ball flight (every throw of their careers) and therefore have unmatched intuition on the behavior of a ball after it leaves their hand. Thanks to that intuition, they provide great hypotheses on baseball aerodynamics for physicists, engineers, and data scientists to test.

One fluid dynamics concept especially well-suited for testing after the upgrade to Hawk-Eye was the Magnus effect. It is defined as the "generation of a sidewise force on a spinning cylindrical or spherical solid immersed in a fluid when there is relative motion between the spinning body and the fluid," [9]. Prior to SSW's discovery Magnus was the only force other than gravity, drag, and pitch velocity known to act on the pitched baseball. Dr. Barton Smith of Utah State University translates the definition of Magnus into baseball terms, providing three defining characteristics of the force: 1) it pushes the ball in direction that the front is moving, 2) it acts normal to the axis of rotation, and 3) it increases with speed and rotations per minute (RPM) [5]. Until SSW's discovery, Magnus was the only force other than velocity that pitchers knew they could actively control, by adjusting how fast they spun the ball.

But the Hawk-Eye upgrade led to the discovery of a new, but similar, controllable force on the pitch called Seam-Shifted Wake (SSW). It has been affecting the movement of certain pitches all along, but now baseball researchers have a chance to quantify its effects. This gives pitchers the opportunity to use it intentionally and in turn, more effectively. Unfortunately, in the publicly available data on MLB pitchers there is an important piece of data left out that prevents full understanding of SSW's effects in this way. This work will propose a method to fill that gap in the publicly available data. Doing so will get the baseball research community closer to a level of understanding of SSW that is necessary for the science to be applied by the pitchers themselves.

## SSW basics

In 2020 Seam-Shifted Wake was introduced in the master's thesis of Andrew Smith, a student of mechanical and aerospace engineering at Utah State University under Dr. Barton Smith. Dr. Smith introduced SSW to the larger baseball community in his speech at the 2021

Sabermetrics conference. To follow is an introduction to the aerodynamics behind SSW to motivate the analysis coming later.

The term wake is generally defined as "the track left behind a moving body in a fluid" [10]. In baseball terms, it refers to the path of disturbed air left directly behind the baseball, trailing in the direction opposite of the ball's motion as shown below. These images come from Dr. Smith's original journal article on SSW, and are also available on his site, baseballaero.com.

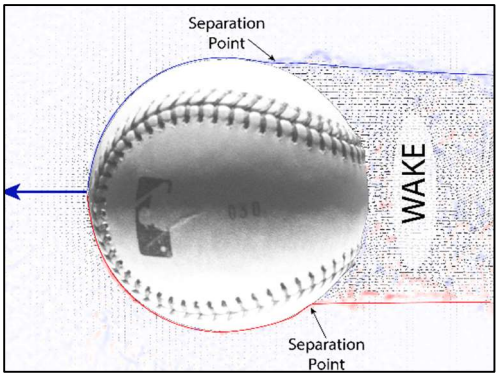



**Figure 1:** Baseball moving to the left, with the wake trailing behind the ball, starting with separation points at the top and bottom of the baseball [2].

The Magnus force can alter the two separation points noted in the diagram above, where air loses contact with the surface of the baseball and begins to form the wake. In the image below, because the ball is spinning to the right, the separation point on the top of the ball is moved backwards and the separation point on the bottom is moved forwards. This shifts the wake of the ball downward, creating an upward force on the ball opposing gravity. This is Magnus, a spin-based altering of the wake. Strong Magnus pitches, like 4-seam fastballs with high rotations per minute, feel to hitters as if they are rising as they reach the plate. In reality these pitches do not rise but only drop less than the hitter is expecting, because the Magnus force opposes, but does not overcome the force of gravity. As Dr. Smith's definition pointed out, the

Magnus force always acts in the same direction that the front of the ball is spinning. For 4-seam fastballs with pure backspin, the Magnus force is directed upwards. For 12-6 curveballs with pure topspin, the Magnus force is directed downwards. The image below shows a fastball with backspin travelling to the left, with a downwards shifted wake.

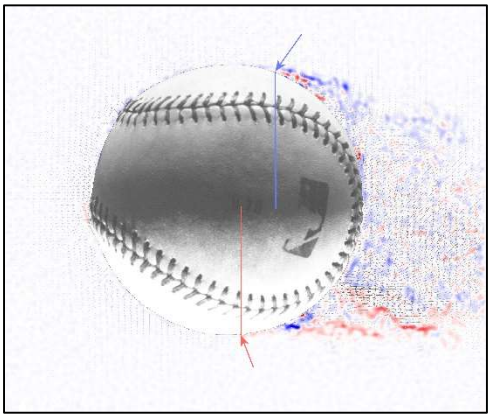


**Figure 2:** Spinning baseball with wake shifted downward due to spin (Magnus) [2].

Like Magnus, Seam-Shifted Wake alters the location of the wake to change the direction the ball travels. Unlike Magnus, SSW is not dependent on spin. In the image below a non-spinning ball is pictured in flight to the left of the page with a shifted wake. The wake is shifted upwards resulting in an SSW force directed downwards, causing the ball to drop.

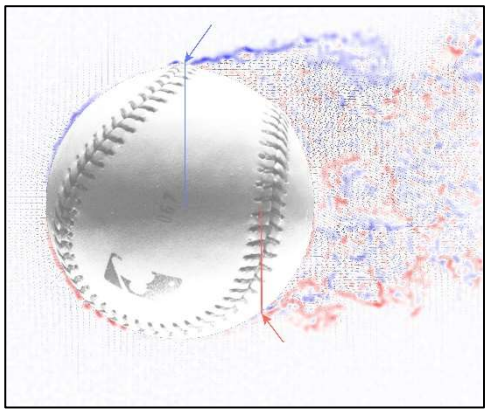


**Figure 3:** Non-spinning ball with wake shifted due to seam location (SSW) **[2].**

The SSW force only occurs when the seams of the baseball are in specific regions, where the seams meet the air above and below the baseball and can actively shift the wake. The diagram below shows the approximate regions where the seams must be for an SSW pitch, highlighted in red. If the seams do not stay consistently within these regions during the flight of a pitch, the SSW force will produce little to no movement. Additionally, if the SSW force is symmetric on both the top and bottom of the ball, the SSW force will cancel out, again producing no noticeable effect.

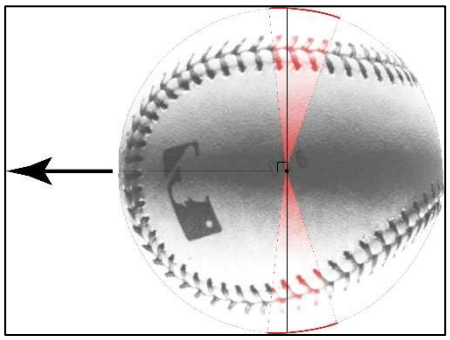


**Figure 4:** Regions (red) where seams must be located for SSW to have effect. [2].

## Identifying SSW pitches

As the initial discoverers of the SSW, Dr. Barton Smith and his students were naturally also the first to propose a method for SSW pitch identification**.** MLB's radar-based Trackman technology has long been able to estimate the spin axis of a pitched baseball with a model incorporating the known forces acting on the ball, known as the Magnus model [5]. The Magnus model estimation for spin axis of the ball is referred to in the literature as the inferred axis. But Hawk-Eye can use optical tracking to observe, rather than infer, the spin axis. Seam-Shifted Wake is the only force known at this point to be missing from those Magnus models used to infer

spin axis. So, any difference between the inferred axis and the actual axis observed by Hawk-Eye is likely due to the SSW force [2]. A scatter plot of the inferred spin axis versus the observed spin axis is therefore a useful tool in SSW pitch identification. Below is an example of such a scatterplot for 2020 Cy Young Award Winner Shane Bieber, aggregated by year and pitch type. CH refers to changeup, FF refers to 4-seam fastball, FC refers to cutter, SL refers to slider, and CU refers to curveball. For reference, the line on the plot indicates the points where inferred axis would be equal to observed axis, where the Magnus-based model is accurate, and SSW force is not suspected to be in effect. In his 2020 Cy Young year, Bieber's cutter was benefiting from the SSW force along with his slider. The plot shows that his curveball movement is dominated almost entirely by the Magnus force.

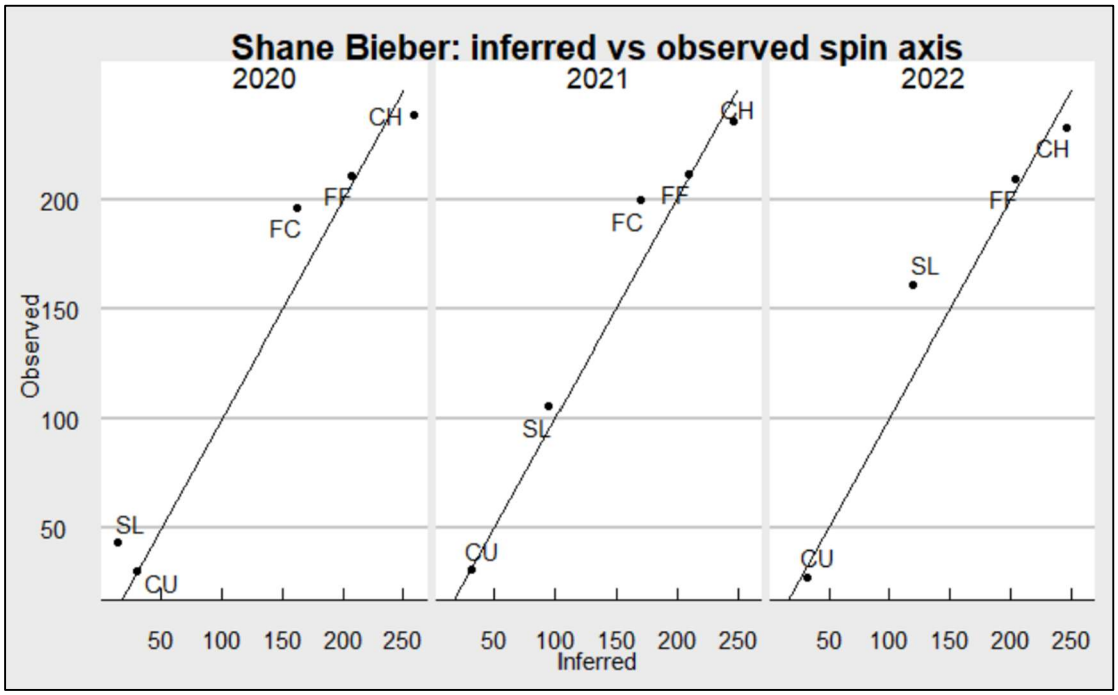



**Figure 5:** Magnus-model based "inferred" spin axis versus Hawk-Eye observed spin axis for Shane Bieber in the last three seasons, by pitch type (CH = changeup, FF = 4-seam fastball, FC = cutter, SL = slider, CU = curveball). Pitches off the line are affected by the SSW force.

Another important concept to consider when analyzing SSW pitches is active spin. Active spin is defined as "spin that contributes to the movement of a pitch," [4]. For example, a baseball spinning like a spiraling football or bullet has very little, if any, active spin. This bullet-like spin, with low active spin, is referred to as gyro. A curveball with pure topspin drops dramatically because the spin-induced Magnus force acts downwards. The ball's spin is affecting its movement. But the pitch thrown with pure gyro is not experiencing the Magnus force, and that spin is not changing that ball's movement. The gyro pitch has low active spin. High active spin values are near one and low active spin are near zero. Returning to the example of Bieber, his slider is thrown with almost pure bullet spin. On the previous plots his slider shows to be one of his primary SSW pitches, with significant difference between the observed and inferred axes every year. Dr. Smith and his team generalized this knowledge a bit further, claiming the odds of SSW influencing the pitch were much greater for those pitches thrown with gyro, or low active spin [5]. In the plot below, the generalization is reaffirmed. Axis deviation refers to the difference between the inferred and observed axes. The bell shape of the plot, centered near a deviation of zero, indicates that at high values of active spin, there tend to be lower values of deviation between the axes, or in other words, lower probability that SSW is at play.

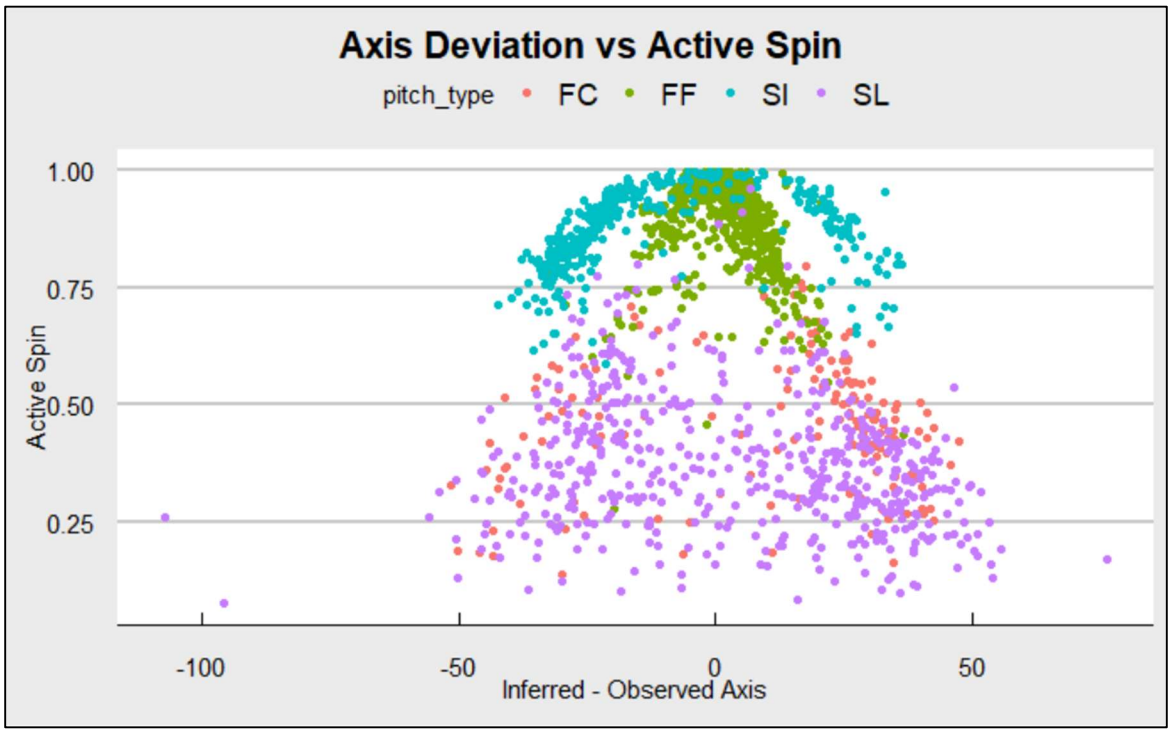



**Figure 6:** Axis deviation versus active spin, colored by pitch type, using data averaged for each pitcher in throughout 2022 season. A majority of high active spin values correspond to lower values of axis deviation, indicating SSW pitches often have low active spin.

When Dr. Smith and his students proposed analytical tools like those demonstrated above to identify pitches where this newly discovered force was at play, they laid the foundation for the rest of the baseball research community to explore SSW. Smith proved the force existed well before he had a complete understanding of exactly what that meant. He set up a laboratory to test these ideas and has proven many of his preliminary theories correct. His quantitative methods were taken a step further by Dr. Glenn Healey of UC Irvine, who proposed a method to quantify what the SSW force was actually doing to pitches in MLB games.

His preliminary findings in the article "Analyzing the Side Force on a Baseball Using Hawk-Eye Measurements" provide the foundation for the research objective of this thesis. In outlining his own paper's objective Healey writes,

"The Trackman (TM) system has been used at MLB ballparks for several years to measure the trajectory and spin vector magnitude for pitches [4]. The Hawk-Eye optical sensor was introduced as MLB's primary pitch-tracking technology in 2020. In addition to the measurements generated by the TM radar, Hawk-Eye provides information about the direction of the spin vector. We show that this additional information can be used to characterize a side force which has been theorized to result from an asymmetric flow separation, aka seam shifted wake, caused by the surface roughness of the ball. The existence of the side force is supported by differences in observed and inferred spin axes as well as by laboratory measurements," [3].

Healey begins the piece with a recap of some of the concepts discussed earlier in this paper. To summarize again, the spinning baseball in flight is under the influence of four forces, not including the newly discovered Seam-Shifted Wake. Those forces are drag, gravity, velocity applied by the pitcher, and the lift force (Magnus), as diagramed below, from Healey's piece.

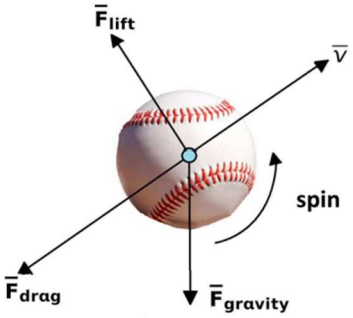



**Figure 7:** "Forces on a spinning baseball in flight" [3].

The SSW force is not included in the above diagram because the force acts outward from the page, in accordance with the right-hand rule [3]. He includes the following image to show that, with ω as the side force and the red arrow indicating the direction of spin, connecting this diagram to the previous.

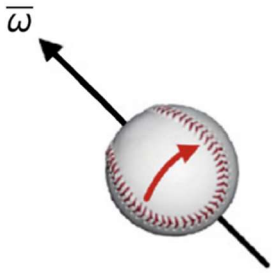



**Figure 8:** "Spin vector ω" [3].

After his brief review of baseball aerodynamics, Healey begins working towards quantifying the effect of the side force depicted above in Figure 8. To start, he develops a method to estimate the magnitude of the lift force (Magnus). The mathematical details of this specific method are not critical to the research objective of this thesis. But this was an important step in Healey's work that allowed him to proceed towards quantifying the effect of Seam-Shifted Wake. He combined large amounts of Trackman data with smaller amounts of laboratory optical data to ultimately come up with a model estimate for the Magnus force. Then he writes,

> "The Trackman (TM) system generates a nine-parameter model for each pitch in terms of the three-dimensional acceleration vector $a = (a_x, a_y, a_z)$, which is assumed constant over the pitch trajectory, and the three-dimensional velocity and position vectors for a point on the trajectory. These parameters can be used to recover the full path of the pitch from the measured release point using the equations of motion. The system also estimates the magnitude of the spin vector $|\omega|$ from the distribution of Doppler shifts. Hawk-Eye

generates the same pitch descriptors as the TM system but also provides information about the direction of the spin vector," [3].

As discussed early on in this thesis, Hawk-Eye offered a level of detail in pitch data not offered by the Trackman system. It gives additional information about the spin vector, which is especially helpful in learning more about the SSW force. More specifically:

> "Publicly available Hawk-Eye data includes the direction $\theta_\omega$ of the projection of the spin vector onto the xz-plane. If the unit spin vector is represented by $\omega = (\omega_x, \omega_y, \omega_z)$ then $\theta_\omega$ is given by $\theta_\omega = \text{atan2}(\omega_z, \omega_x)$. … $\theta_\omega$ can be used to estimate the three-dimensional spin vector $\omega$ and to separate the acceleration components. The measured $\theta_\omega$ restricts $\omega_b$ to a one-parameter family of unit vectors $\omega(\omega_y) = (\text{acos}\ \theta_\omega, \omega_y, \text{asin}\ \theta_\omega)$, [3].

To simplify, what Healey has indicated is that Hawk-Eye data on the spin vector components, specifically the averaged value for the spin vector in the y direction, $\omega_y$, can be used to estimate the x and y components of the unit spin vector, which with further mathematical work, allows for separation of the acceleration of the ball into components corresponding to lift, drag, Magnus, and Seam-Shifted Wake. The rest of Healey's work is dependent on this operation, using the averaged value for $\omega_y$ available in Hawk-Eye data to estimate the spin unit vector's other components. Once the acceleration components are separated, then some understanding of what kind of movement SSW causes is possible. Healey later presents a table showing the average inches of movement for each pitch type attributable to each force: drag, Magnus, and SSW.

| pitch type | Drag | Magnus | Side | Total |
|---|---|---|---|---|
| Changeup | 1.33 | 7.98 | 2.40 | 9.38 |
| Curve | 1.49 | 8.81 | 1.28 | 7.81 |
| Cutter | 1.20 | 4.28 | 2.04 | 5.60 |
| Four-seam | 1.18 | 9.68 | 1.16 | 10.96 |
| Split | 1.42 | 7.29 | 2.79 | 8.66 |
| Sinker | 1.20 | 9.28 | 3.42 | 11.02 |
| Slider | 1.34 | 3.35 | 1.52 | 3.51 |

**Figure 9:** "Average movement in inches due to each force by pitch type" [3].

According to these results, sinkers, splitters, changeups, and cutters have the largest amount of movement due the side force, on average across all MLB pitchers [3]. But to reiterate, these findings in Healey's work depend on the averaged value for $\omega_y$. This thesis aims to improve upon his approximation. With machine learning it is possible to predict a value for $\omega_y$ on a pitch-by-pitch basis, rather than relying on the average. To follow is an overview of the methods used to do that. First will come a description of the data sources involved.

## Data

All data used in this analysis comes from BaseballSavant, an online repository for the data collected during MLB games. The repository has data from both the Trackman and Hawk-Eye eras. For this research, two different types of data were pulled from BasballSavant, as outlined below.

The Spin Direction – Pitches Leaderboard dataset has the information on pitch spin and velocity relevant to this work. The data is aggregated by pitch type for each MLB pitcher over a given season. In other words, each row of the data set corresponds to one pitcher-pitch each year, like Shane Bieber's slider in 2020. The averaged values for each component of the unit spin vector, which were critical to Healey's research, are in this dataset. $\omega_x$, $\omega_y$, $\omega_z$ are referred to as image_spin_x, image_spin_y, and image_spin_z respectively. Other variables in the aggregated

data that are important to this analysis are pitch_type, release_speed, spin_rate, and movement_inches. Spin_rate refers the revolutions per minute of the ball, or how fast the pitcher spins the ball. Movement_inches tells how much the ball moves, in both the horizontal and vertical planes. These variables in the aggregated data were key for this analysis because they were also available in, or convertible to, variables in the second dataset used [7].

There is also the option on Savant to pull a similar dataset on an individual pitcher basis, that has pitch-by-pitch information for a selected pitcher and pitch type [8]. In this data each row corresponds to one pitch of the selected type, in each season for that pitcher. One row, for example, could represent one of the many sliders Shane Bieber threw in the 2020 season. The most important difference between the pitch-by-pitch and the aggregated data is that this is missing the variable image_spin_y, that prevents any application of Healey's methods to quantify the effect of SSW on a pitch-by-pitch basis. Throughout this work, pitch-by-pitch data was pulled for various pitches from various players, to be specified when referenced later in the work. From the pitch-by-pitch data the following variables were used: pfx_x, pfx_z, spin_axis, pitch_type, release_speed and release_spin_rate. Pfx_x and pfx_z are the pitch movement in inches, horizontally and vertically, respectively. Spin_axis refers to the observed axis of rotation of the pitch, corresponding to the observed spin axis mentioned early on in this work. Release-spin_rate corresponds directly to the spin_rate variable in the aggregated, spin direction leaderboard data [8].

These two datasets provide an opportunity to train a model that can learn from the aggregated data how to predict the image_spin_y value on a pitch-by-pitch basis, improving on Healey's approximation that relies on the average value of $\omega_y$ for all pitches of a certain type. That is the goal of this work.

## Methodology

Using machine learning to predict values for the y-component of the unit spin vector requires training and testing to occur on data that already has a value for image_spin_y present, a ground truth that is necessary for model evaluation. Fortunately, this is available in the aggregated data. Ultimately the model must work well to predict values for image_spin_y in the pitch-by-pitch data, in order to improve upon Healey's approximations. Pitch-by-pitch knowledge of the unit spin vector components will allow for a more detailed understanding of the SSW force.

Of various algorithms tested, two in particular showed promise in predicting image_spin_y in the aggregated data: random forest regression and multivariate adaptive regression splines. Closer consideration of how the predictors relate to each other and to the outcome variable image_spin_y suggests why these two were successful. Again, the predictors used to build the model were pitch velocity, spin rate, spin axis, pitch movement, pitch type, and pitch hand. It is clear the relationship between them and the outcome is not linear. It is also reasonable to assume the predictors interact with each other. Both the MARS and random forest algorithms model non-linear relationships and automatically pick up on interactions among predictors.

Models were trained using both algorithms, with results compared later. In the aggregated dataset, a few steps were taken to prepare data prior to model formulation. Flipping the sign of image_spin_y and image_spin_z for left-handed pitchers improved model performance and made these values match those of right handers in sign. These two variables also required conversion to match the spin_axis variable present in the pitch-by-pitch data for which the model was ultimately being built. That conversion called upon the equations presented in Healey's work:

$$\theta_{\omega} = \text{atan2}(\omega_z, \omega_x)$$

or equivalently,

spin_axis = atan2(image_spin_z, image_spin_x), [3].

Outliers were removed from the data, where the difference between the inferred and observed axis was greater than 50 for a given pitch, as these values were likely result of measurement error. Outliers that were removed are indicated in the plot below in red. After outlier removal, the aggregated data was filtered by pitch type and ready for model training.

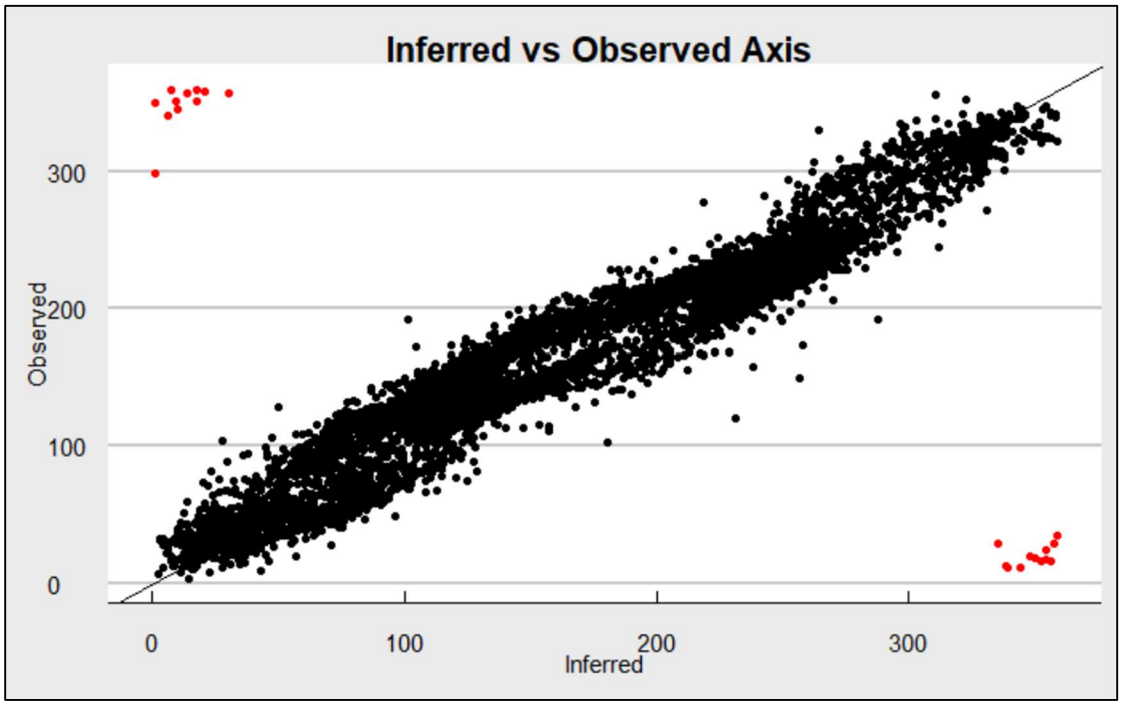



**Figure 10:** Outliers were removed from the aggregated data prior to model training. Indicated in red are the data points that were removed, where the difference between inferred and observed axis exceeded 50 degrees, suggesting potential measurement error.

Hawk-Eye spin data is only available for the 2020, 2021, and 2022 seasons. Data from each year were combined and then randomly split into a training and testing set. Models were tuned, trained, and tested individually for each pitch type, both with the random forest algorithm and the MARS algorithm.

## Results

The random forest and MARS models perform similarly across all pitch types. Shown in the table below is the Root Mean Square Error for each model when applied to predict image_spin_y in the test sets of aggregated spin data. For context, the average value of image_spin_y in the aggregated data set is about 0.55.

| Pitch Type <fctr> | RMSE (MARS) <dbl> | RMSE (RF) <dbl> |
|---|---|---|
| four-seam fastball | 0.07611633 | 0.08055057 |
| slider | 0.05193404 | 0.04495573 |
| changeup | 0.24162424 | 0.22437899 |
| curveball | 0.07205968 | 0.07324096 |
| sinker | 0.11510201 | 0.10894566 |
| cutter | 0.04275564 | 0.04487172 |
| split-finger | 0.20442328 | 0.19318823 |

**Figure 11:** Root Mean Square Error for the models corresponding to each pitch type, for both the MARS model and the random forest model. Results are mixed, but models perform similarly across all pitch types.

Referring to the works of Smith and Healey, sinkers, splitters, changeups, and cutters are the pitches most effected by SSW on average [3], so those results here are of particular interest. The models for splitters and changeups are especially poor performing, with RMSEs well over 0.15. The models for sinkers perform better but still not all that well. But the models for cutters are performing well, with RMSEs below 0.05. A reason for these differences in model

performance across pitch types is not obvious. There does seem to be some correlation between spin rate and model performance, however. On average, changeups are the pitch on this list with the lowest spin rate, followed by the split-finger. These two also have the worst model performance. Cutters on the other hand typically maintain a higher spin rate, that model performs well. These are only preliminary hypotheses.

The next challenge in this work is judging the model's performance in predicting values for image_spin_y in the pitch-by-pitch data, where there is no ground truth value for comparison. Accurate prediction in this data was ultimately the aim of this work. One strategy to build confidence in these predictions is to compare the predicted values for image_spin_y in the pitch-by-pitch data to the average value presented in the aggregated dataset. The density plot below shows such comparison for one of the games best cutter pitchers, Kenley Jansen. The plot shows the distribution of image_spin_y predictions from the MARS model, which slightly outperformed the random forest model. The dashed line represents the average value for image_spin_y presented in the aggregated data, for Jansen cutters. Ideally the peak in the plot would come centered on that line. Nevertheless, these results suggest the model is performing reasonably well.

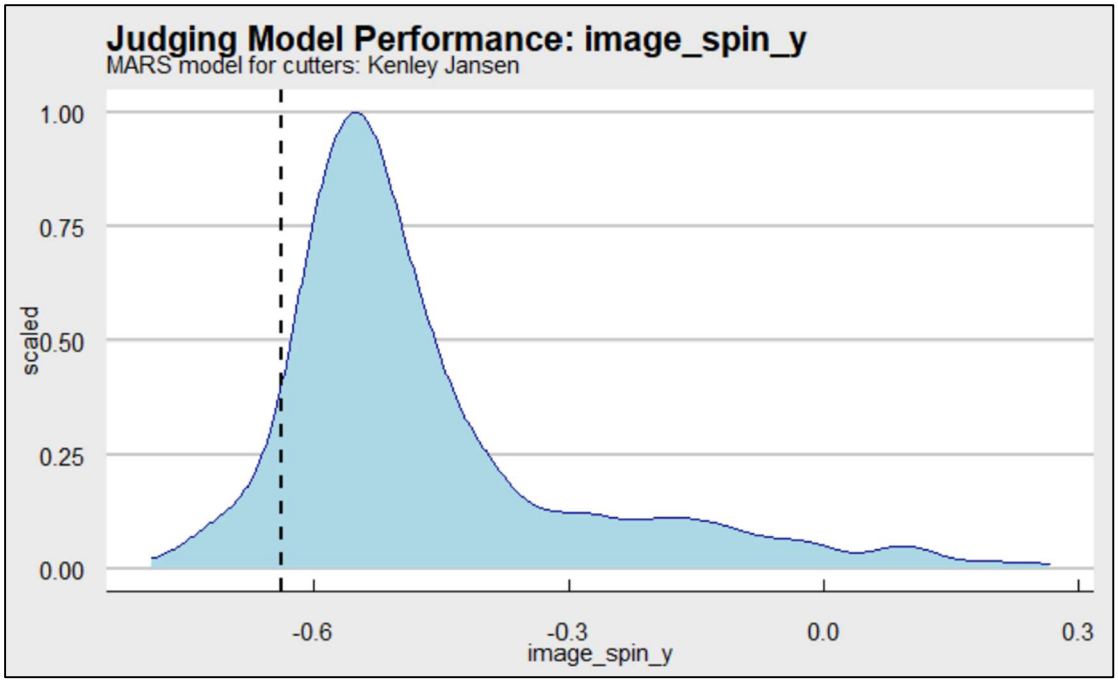



**Figure 12:** Density plot showing the distribution of estimates for pitch-by-pitch image_spin_y produced by the MARS model for cutters, specifically for Kenley Jansen cutters. The dashed line corresponds to the averaged value for image_spin_y provided in the aggregated data.

Similarly, checking the sum of the unit spin vector components can provide some feedback on how the model's performing. Ideally, image_spin_x, image_spin_y, and image_spin_z would sum to one, according to the definition of a unit vector. The density plot below shows the distribution of the summed vector components, with the dashed line at one, the correct sum for a unit vector. This distribution is more spread out, and a bit further off the mark. These results are less encouraging.

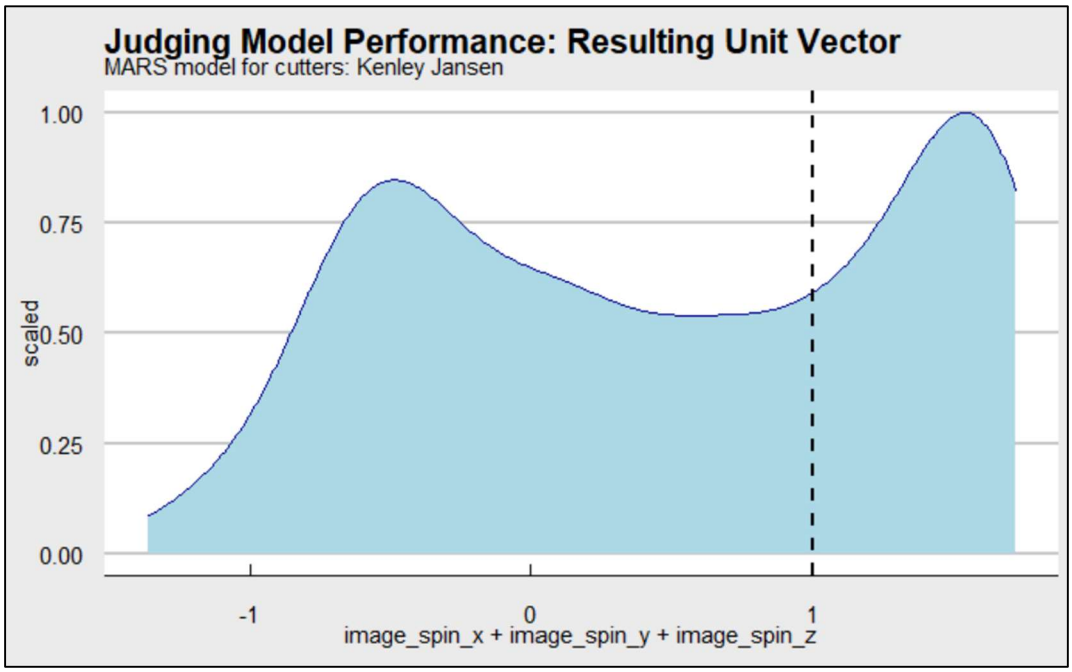
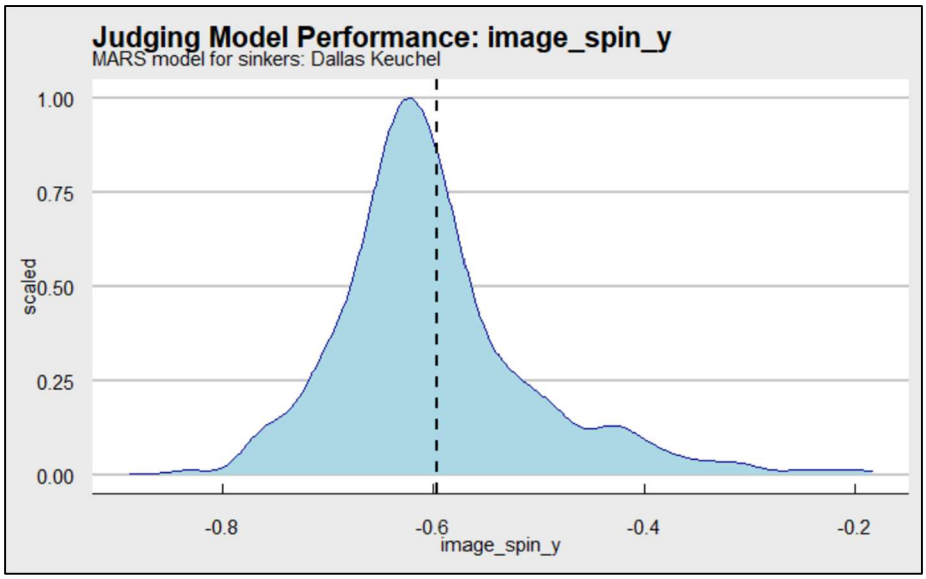
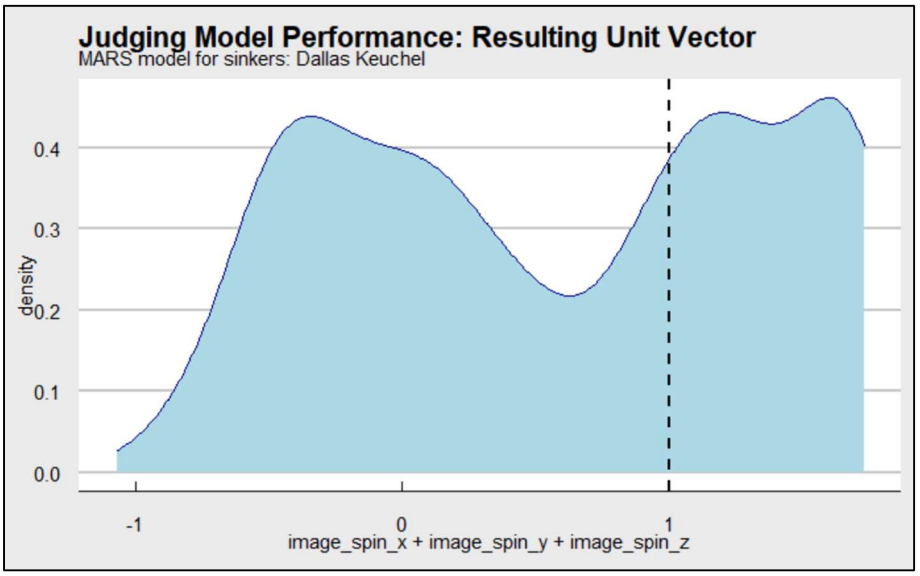


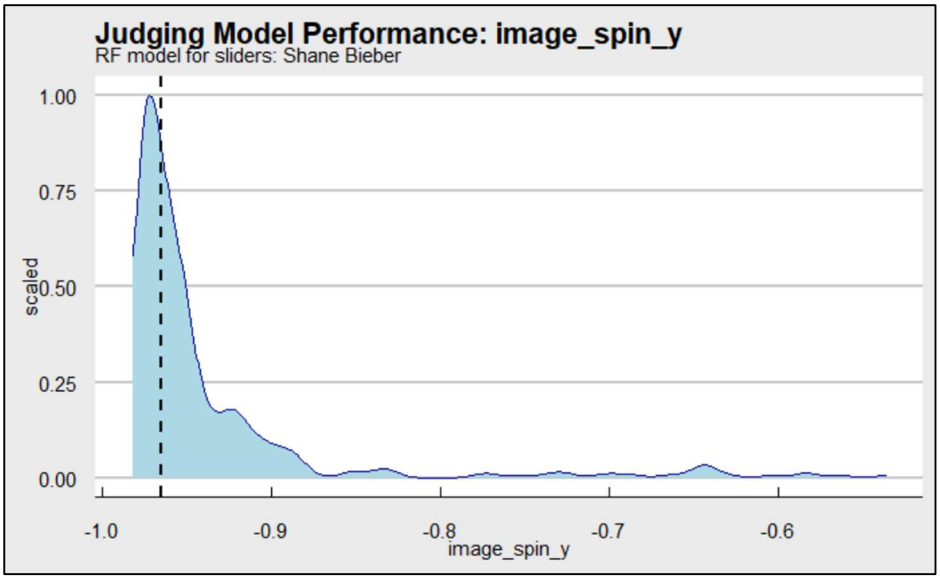

**Figure 13:** Density plot showing the distribution of the sum of the spin vector components, based off predicted values for image_spin_y provided by the MARS model, again for Kenley Jansen cutters. The sum of the unit spin vector should be equal to one by definition.

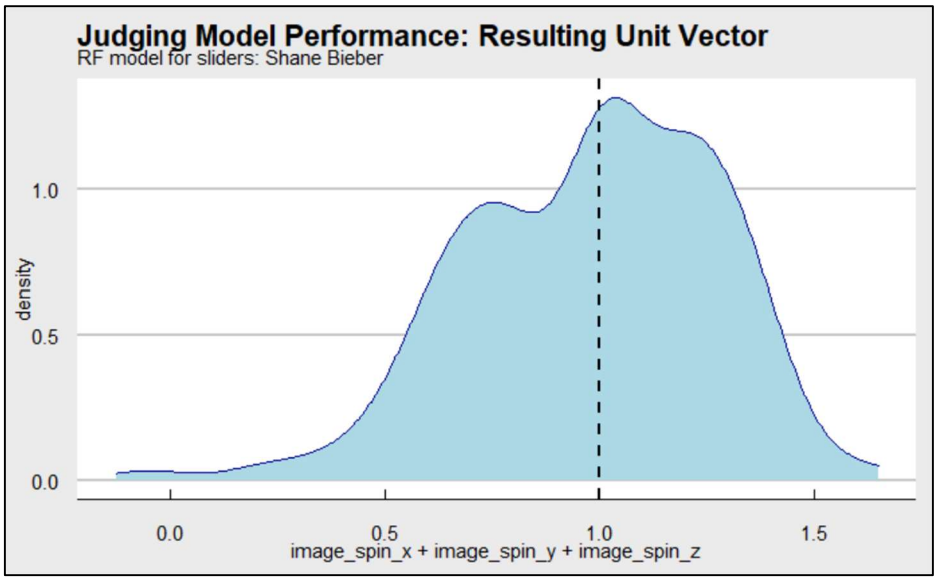
To see how the model is working on other pitch types, the same strategy can be followed for different pitchers. Below are the same two plots for Dallas Keuchel, known for his great sinker, and Shane Bieber, known for his great slider. We see similar behaviors in each of these plots. When looking at the predicted image_spin_y value alone, estimates look consistent with the value given in the aggregated data. However, when looking at the sum of the unit vector based off that predicted value of image_spin_y, estimates look less reliable. It is important to note that in the aggregated data, the components of the unit vectors do not add up to one in every row because those values are averaged. Nevertheless, the resulting predicted unit vector components would ideally be more tightly concentrated around one.




**Figures 14, 15:** Similar density plot showing the distribution of the predicted values for image_spin_y and the corresponding unit vector components. To show the model's performance on sinkers, data on Dallas Keuchel was used.




**Figures 16, 17:** Similar density plot showing the distribution of the predicted values for image_spin_y and the corresponding unit vector components. To show the model's performance on sliders, data on Shane Bieber was used.

Overall, these results indicate the model is predicting reasonably well in the pitch-by-pitch data, but potentially not to the level of accuracy required if these estimates were to be relied on to quantify the effects of SSW through Healey's proposed methods.

## Discussion

Baseball analytics research has quickly become as competitive as the game of baseball itself. The cutting-edge research in the field is kept private by the teams doing it with the goal of gaining or maintaining their competitive edge. Due to this privacy, it is difficult to gauge the potential impact of a research thesis such as this. In the publicly available research, there is still very little shared understanding on exactly what the SSW force does to a pitch. Healey seems to have made the most progress towards answering that question. Therefore, any marginal improvement on his work does have an opportunity to change the trajectory of Seam-Shifted Wake science right now.

The specific models deployed in this work are likely not predicting values for the y-component of the spin vector that are reliable or consistent enough to use to begin calculating the effect of SSW on individual pitches right away. But still, the methodology proposed here is a new and useful contribution to the small collection of publicly available SSW research. Not yet had supervised learning been applied to the publicly available spin data on Savant, to fill in the missing pieces of data not yet made public that are vital to quantifying the SSW effect. MLB may soon release the components of the spin vector on a pitch-by-pitch basis, but rather than wait for that to happen, this research lays the foundation for a machine learning based approach for imputing the values right now. Only two and a half seasons of Hawk-Eye data were available for the model training and tuning completed here. With more time and more data, these models will surely improve.

Where future research can build upon this work is in the methods for validation of the image_spin_y predictions in the pitch-by-pitch data. The validation steps taken in this work were not extremely detailed and could likely be improved. MLB could even conduct a competition as

they have before, on Kaggle, where participants compete to deploy models that predict most accurately on unseen testing data, much like the pitch-by-pitch spin data not available on Savant. Data scientists and physicists in the competition could at least see their model's true performance against the ground truth image_spin_y value on a pitch-by-pitch basis without seeing the data itself. This would allow the research community to advance their understanding of the SSW force and maintain the privacy of that data on behalf of the MLB.

## Conclusion

Here a machine learning strategy to fill in missing pieces of data in the publicly available pitch spin data has been proposed. This missing data, related specifically to the components of the unit spin vector of a pitch, are critically important to learning more about a newly discovered force acting on the ball in flight called Seam-Shifted Wake. Using random forest regression and Multivariate Adaptive Regression Splines, this method uses data on the average values for the spin vector components for pitchers' pitches for model training and testing. Those models can then be used to predict the y-component of the unit spin vector on a pitch-by-pitch basis, which can be used to calculate the corresponding x and z components. The models deployed were very accurate in predicting spin vector components for cutters in particular. This work serves as a first attempt in the publicly available research to apply supervised learning to the missing data problem in the newly formed Seam-Shifted Wake research community.

## References


**[1]** Ajeto, Michael. “We're Talking about Seam-Shifted Wake Wrong.” *Baseball Prospectus*, 17 May 2022, https://www.baseballprospectus.com/news/article/74601/climbing-the-ladder were-talking-about-seam-shifted-wake-wrong-axis-deviation/.

**[2]** A. Smith and B. Smith. “Using baseball seams to alter a pitch direction: the seam shifted wake” *Proceeedings of the Institution of Mechanical Engineers Part P: Journal of Sports Engineering and Technology*, October 2020.

**[3]** Healey, Glenn, and Lequan Wang. “Analyzing the Side Force on a Baseball Using Hawk-Eye Measurements - Vixra.” *ViXra.org*, 6 Jan. 2021, https://www.vixra.org/pdf/2101.0046v1.pdf.

**[4]** Nathan, Alan. “Pitch Movement, Spin Efficiency, and All That.” *The Hardball Times*, 27 Aug. 2018, https://tht.fangraphs.com/pitch-movement-spin-efficiency-and-all-that/.

**[5]** Smith, Barton, et al. “Not Just about Magnus Anymore.” *Baseball Prospectus*, 19 Jan. 2022, https://www.baseballprospectus.com/news/article/62912/not-just-about-magnus-anymore/.

**[6]** “Magnus Effect.” *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., https://www.britannica.com/science/Magnus-effect.

**[7]** “Statcast Active Spin Leaderboard.” *Baseballsavant.com*, https://baseballsavant.mlb.com/leaderboard/active-spin.

**[8]** “Statcast Search CSV Docuemnetion.” *Baseballsavant.com*, https://baseballsavant.mlb.com/csv-docs.

**[9]** "Sony Corporation of America Press Releases." *Sony Corporation of America Press Releases*, 28 June 2022, https://www.sony.com/content/sony/en/en_us/SCA/company news/press-releases.html/page215.page643.html.

**[10]** "Wake Definition & Meaning." *Merriam-Webster*, Merriam-Webster, https://www.merriamwebster.com/dictionary/wake.