Decoding Tennis Dynamics: Analyzing How Strategic Shot Selection Evolves Over The Years

Rally Lin (rl326), Derek He (dh356), Michael Li (mll80), Lucas Sasaya (lts28)

Part 1: Introduction and Research Questions

In competitive tennis, strategy is crucial to success as players seek to capitalize on their strengths and exploit their opponents' weaknesses. Recently, computational and data science methodologies have aided in finding these strengths and weaknesses, with professional players incorporating data scientists to watch and assess player performance and innovate new playing patterns in the realm of tennis strategy. Our project, "Decoding Tennis Dynamics," aims to use data science to generate valuable insights about how specific strategic elements in tennis, like if the utilization of a topspin forehand or an underspin slice evolves over the years.

Our **research questions** are as follows:

> *How does tennis shot selection evolve over the years in terms of frequency for men's matches, with specific shots such as forehand, backhand, serves, etc.? Furthermore, how has the server win percentage in service games changed over the years, especially when considering the percentage of serves that land as a first serve compared to a second serve?*

Our investigation is relevant, as understanding the impact of shot selection on match outcomes and serve returns is relevant to players, coaches, and fans. We hope the results of our investigation will enhance our understanding of the game and potentially influence future training methodologies for tennis players.

Part 2: Data Sources

Main Data: Match Charting Project (MCP), charting-m-stats-Overview.csv

In our project, we used four datasets. These data sources were all collected by Jeff Sackmann's Tennis Match Charting Project on Github. For our first source, we include the overview to evaluate a player's performance. The overview represents either player's performance in a specific set or overall in the match, describing the serving, returning, and shot-making statistics to gain insights into their strategies and effectiveness during the match. We used the overview to analyze players' general performance comprehensively, which we used to transition to explore the following data sets on a more detailed basis.

The second source we used is charting-m-points-2009.csv. This dataset focuses on such the series of points played between two players in a given tournament, as number of points, set scores, game scores, current point score, tiebreak indicators, serving and returning players, shot types, additional notes, shot characteristics (e.g., serve, in/out), and point outcomes, focusing on the period from 1960-2009. Notably, the dataset includes indicators for shot types, shot characteristics (e.g., isAce, isUnforced/isForced), and rally information, providing a comprehensive record for match analysis. which is also the information we used with charting-m-points-2010s.csv and charting-m-points-2020s.csv.

Although this isn't a dataset we explicitly derived from the match charting project, we used the dataset named 'sufficient_points_by_year' to aggregate tennis match data across different years. This dataset provides average statistics of numeric and boolean columns grouped by year, allowing us to observe trends and patterns in player performance over time. By filtering for years with a sufficient

number of data points (more than 300 in this case), we ensure robustness in our analysis and focus on years with significant data coverage. The specific csv files can also be found in 🗀 data .

Part 3. Modules

Module 03: Visualization: We used the visualization module for the presentation of our data and the discovery of trends before and after statistical inference to provide a clear representation of our results. The concepts we used include Seaborn's sns.lineplot() to plot line plots, the pie chart to visualize the percentage of winners overall for a shot type, and a distribution plot to visualize the trend between specific shot selections used in a match over the years. Each shot was visualized in its line plot. For example, in one plot, we have '% A Serve is an Ace' from 1960 to 2020, where a point represents the percentage that a serve is an ace across men's matches for a given year.

Module 04: Data Wrangling: We used the data wrangling module because we needed to reformat and refine the data that was initially collected from GitHub. In the csv files, there are numerous matches from different tournaments and players. We wrangled our data using boolean masks, string strip, and boolean indexing to accurately translate and filter data between games and players for exploratory data analysis in the data gathering and cleaning stage. In addition to our data gathering and cleaning stage, we used regex to filter for the last shot having a star, thus capturing the winning shot sequence during the data analysis stage to create our pie chart. When creating specific columns like aces, we used .notna to remove missing values.

Module 06: Combining Data: We used the combining data module during the data investigation stage to combine data from different years. The concepts we used include pd.concat to merge data from the three csv files charting-m-points-to-2009, charting-m-points-to-2010s, and charting-m-points-to-2020s to create a master data frame that we used to find relationships between shot selection over the years. Additionally, we grouped the data in the merged data frame into certain relationships like year and match size to plot the correct shot selection on the y-axis. The concepts we used include pandas.groupby('year') and ('rally_winning_shot') to split our data by the years and specific winning shots.

Part 4. Results and Method

We began our investigation by accessing data from JeffSackmann's Match Charting Project (MCP), particularly the charting-m-points-to-2009.csv, charting-m-points-2010s.csv, charting-m-points-2020s.csv, and charting-m-stats-Overview.csv files. First, we read the overview file and create a new column 'year' by selecting the first four indices of 'match_id' to get a better idea of the data set. After reading the former three files using Pandas, we merged the three data frames into a single merged data frame containing data for over 50 years of chartered matches called points_df.

Using the merged data frame, points_df, we create a new column containing the year of all matches. We use the new column to create a new data frame that groups by year and calculates the mean of numerical and boolean columns for each year. We count the number of data points per year and save it into a new column and use it to filter our data set to only include years with sufficient data (> 300). Using the data, we create 2x2 grid subplots, specifying format and title with .subplots. We call sns.lineplot to create the four graphs in Figure 1. Each subgraph represents the percentage of a specific serve across men's matches over the years. For example, in the green line plot, the leftmost bullet point means that the percentage of

server wins of the service game for 1960 across matches was approximately 0.625 or 62.5%. Looking at the line plots, there is a positive trend of Server Win, Ace Serves, and 1st Serve In percentages from 1960 to 2020. However, there is a slightly negative trend in 2nd Serve In percentages. In all four graphs, there are low percentages for all serves between the years 1970 - 1980. Although we know how server win percentage has changed over the years, we wanted a more nuanced look at how specific winning shot frequencies have evolved over the years.

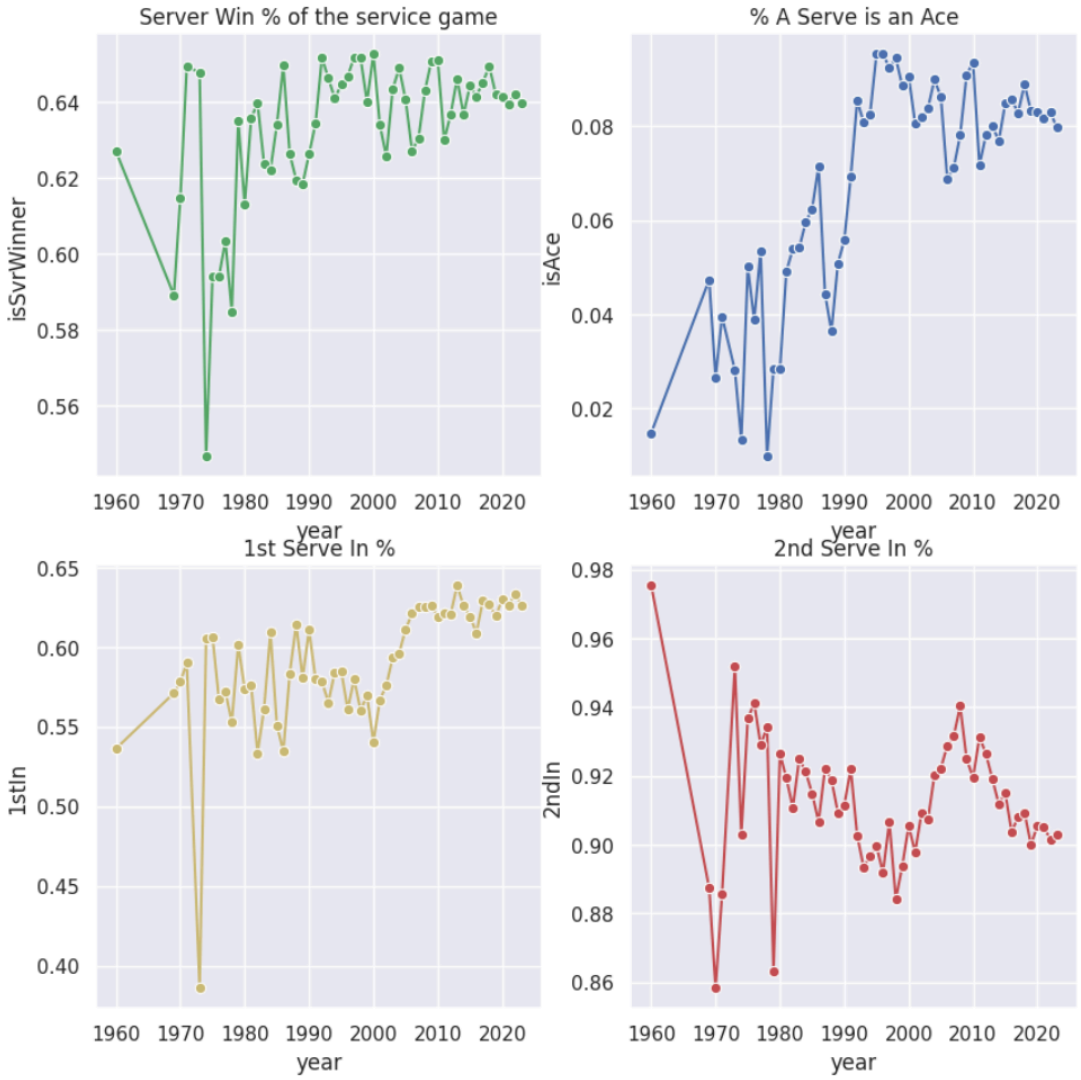

Figure 1: Serve Information

To do this, we created a new function that extracts the winning shot from a rally. We used regex to extract only the last rally sequence and placed the extracted values into a new column. Using the detailed point-by-point data provided in our merged data frame and filtering, we extracted the shot ID (consisting of a character representing the shot type and an integer representing the shot direction) of winning points. After filtering the blank winning shots out using boolean masks, we created a threshold of 0.02 to filter out less frequent categories (less than 2%) from the winning shots. Using this filtered data, we were able to create a pie chart (Figure 2) to show the proportion of each shot type among all winning shots.

From the pie chart, it is clear that groundstrokes represent the majority of winning shots. This makes sense as groundstrokes are one of the core fundamental shots in tennis. After groundstrokes, volleys make up around 21.2% (combined forehand and backhand volleys) of winning shot types in our data, with overhead smashes following at 9.2%. To discover how winning shot type has evolved over the years rather than as a percentage for all years, we wanted a more detailed look into the change in percentages over time.

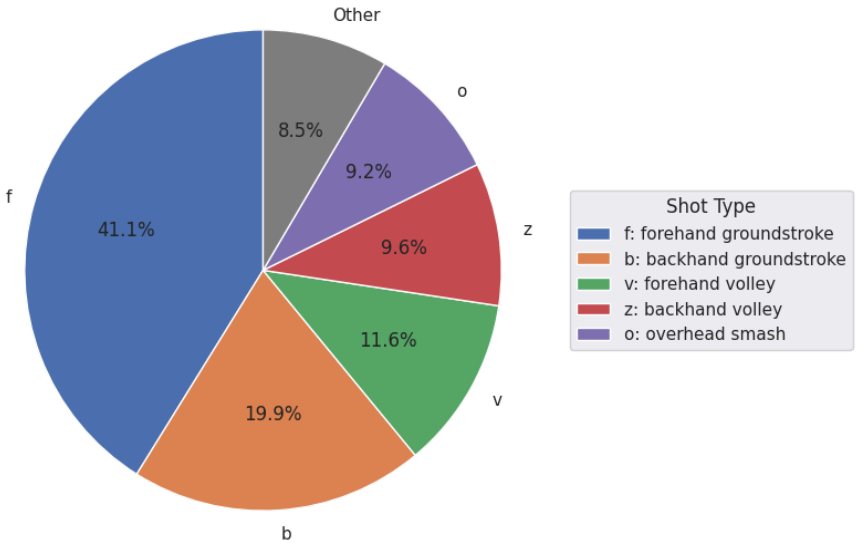

Figure 2: Distribution of Winning Shot Type for All Years

To generate a yearly distribution of winning shot types, we generated a stacked-bar-chart to visualize the fraction of winning shot types for all matches in each year of play (Figure 3). We grouped the filtered data frame by year and type of winning shot and counted the occurrences of each combination. Then, we reshape the data using .unstack to make years the index and the winning shot as the column, filling in missing values with 0. We calculated a fractional distribution by dividing the yearly distribution by the sum of shots for each year, giving us the proportion of each shot type for each year. Using the fractional distribution, we select the top five shot types and label shots that fall outside the top five as ‘other’. Using the new fractional distribution, we generate a stacked-bar-chart with a specified order parameter for our legend, which includes: forehand groundstroke, backhand groundstroke, forehand volley, backhand volley, overhead smash, and then other.

Figure 3 represents the annual variation in tennis winning shot types over several decades. This stacked bar chart emphasizes how different shot types have grown and shrunk in frequency of usage within the sport, with each bar segmented to show the proportion of each winning shot type per year. As can be seen from the plot, the fraction of winning shots that were forehand groundstroke gradually increased over time, decreasing the relative prevalence of other winning shot types. While some shot types remained relatively the same, like backhand groundstroke, other shot types, most notably forehand volley and backhand volley, decreased in proportion by nearly 50% from 1980 to 2020. As shown, the forehand groundstroke still remains the most popular shot over time, even doubling in size over the years.

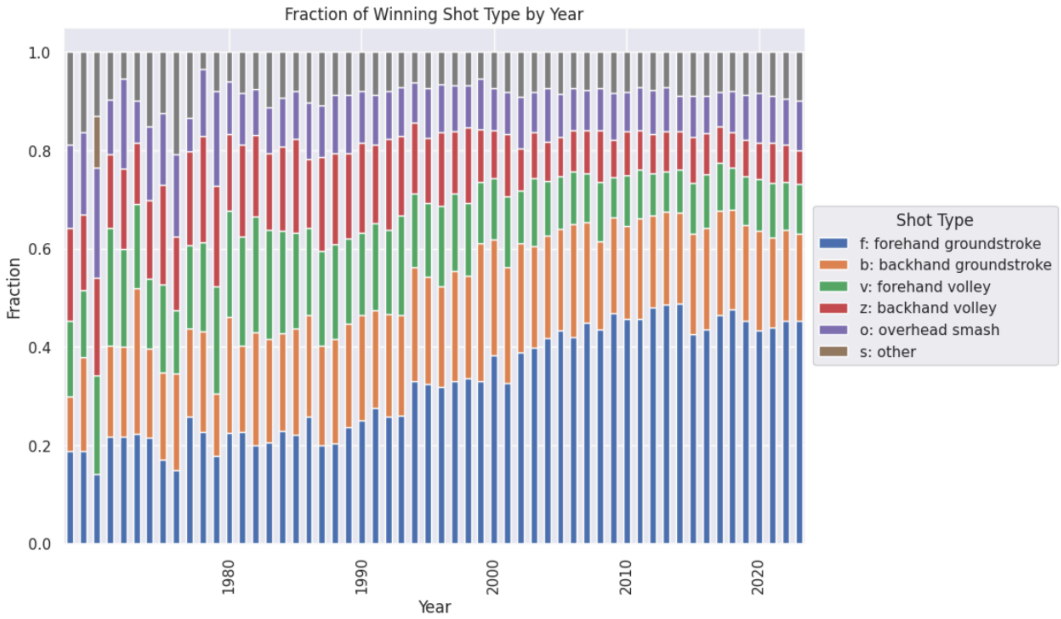

Figure 3: Distribution of Winning Shot Type by Year

Lastly, we attempted to create a machine learning model to predict the winner of a point from a segment of the shot data. The task of our model was to: *"within a (short) rally that ended within three or four hits, predict the winner of a point given shot data from the first two hits."*

Using regular expressions on the shot strings, we filtered the dataset for rallies that ended in three or four hits (either the returner won on the serve return, or the server won on the following hit). Then, we extracted the data for the type and direction (represented by a two character string) of the first and second hits. Finally, we constructed a `LogisticRegression` model and trained it to make predictions for the winner of a point based on the first and second shots.

Running our model, we obtained an accuracy score of `0.61`. This is higher than the 0.57 accuracy that we would obtain on a dummy model which predicts that the server (the more frequent of the two classes) wins every point. The model returns `True` when it predicts the server is the winner of a point, and `False` when it predicts the returner is the winner of a point. As can be seen from the confusion matrix visualized in Figure 4, the model tends to significantly overpredict that the server will win the point.

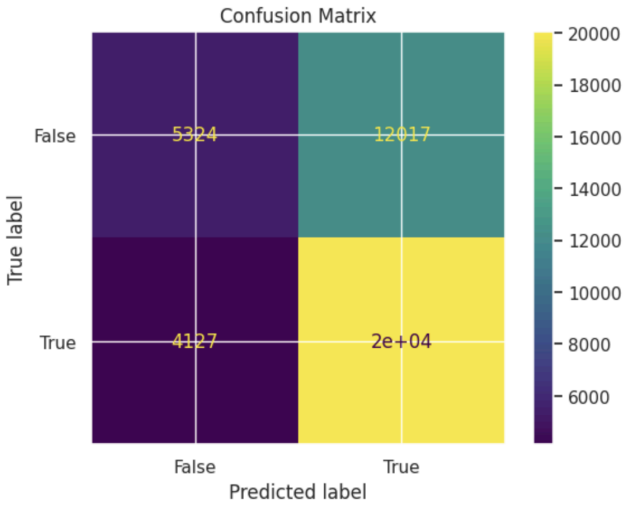

Figure 4: Confusion Matrix for Point Prediction Model

The low overall accuracy of our model highlights the difficulty of creating an accurate predictive model upon complex shot data. Despite this, the model was able to take advantage of some basic relationships between the first few shots and the point winner on short rallies.

The full implementation of our project can be accessed in this Colab notebook.

Part 5: Limitations and Future Work

While this dataset is extensive, it is not exhaustive of all professional matches and may be biased towards certain players or tournaments that are 'better documented' than others. This could skew our understanding of shot selection trends and server win percentages if certain styles or strategies are overrepresented. Furthermore, some data files were formatted with unclear titles and inconsistent labeling across different files. This inconsistency leads to difficulties when parsing the data. Another limitation is the historical coverage of our data. Although it spans from 1960 to 2020, the frequency and detail of data points vary significantly across different periods. Earlier decades might be underrepresented, affecting the robustness of our long-term trend analysis. Additionally, the accuracy of manual charting in the Match Charting Project could introduce errors. Since there were many contributors to Sackmann's charting project, they could have recorded data inaccurately and could have led to inconsistencies in how shots are recorded, particularly in the subjective categorization of unforced errors or shot types. Ethically, our analysis might contribute to reinforcing certain playing styles or strategies at the expense of others, potentially influencing coaching and player development in a direction that favors data-dominant interpretations of performance.

If there was a future continuation of our current project, we should include the women to also draw descriptive conclusions that reflect the nuanced differences and similarities in gameplay strategy between genders. We could also keep track of individual players over the course of their careers to observe how their strategies evolve with age, experience, and how changes in racket technology could still impact strategic decisions in the game. As far as how our results could improve, we could enhance the granularity of our data analysis by incorporating more detailed match contexts, such as player fatigue levels, weather conditions including wind or sun, and surface type for each match.

Regarding any potential informed audience members, it would be useful to interview any of the Duke Men's players/coaches to get their current viewpoints of how they win matches. Given the large sample size of the matches that were played this season, we could improve by drawing on specific shots or strategies that worked for them when it came to winning points. Furthermore, it would be insightful for a coach to include his perspectives on the training procedures required to compete at the highest level.

Part 6: Conclusion

Firstly, we observed a general increase in the efficiency and effectiveness of serve-related metrics. The percentage of server wins, ace serves, and first serves in, all show a positive trend from 1960 to 2020, as demonstrated in our line plots in Figure 1. This suggests that serves have become a more dominant factor in winning points, reflecting perhaps improvements in player fitness, serve technique, and strategic use of the serve. However, there is a noticeable dip in the efficiency of second serves in, which exhibited a slight negative trend during the same period. This could indicate a shift in strategic emphasis towards more powerful, riskier first serves. In our analysis of winning shot types in Figure 2, groundstrokes emerged as the predominant shots leading to point wins, which is consistent with the fundamental nature of these shots in tennis. Our deeper exploration into the yearly changes in winning shot types in Figure 3 highlighted how the use of forehand groundstrokes as a winning shot has at least doubled over the years while the use of both forehand and backhand volleys has decreased, reflecting a shift towards baseline play and away from net play for the modern game.