ATP Tennis Players Analytics

Trifelli Angelo
Sapienza University of Rome
trifelli.1920939@studenti.uniroma1.it
https://github.com/Angelo-Trifelli

Stancioli Júlia
Sapienza University of Rome
stancioli.2175608@studenti.uniroma1.it
https://github.com/julia-stancioli-sapienza

Carrasco Camila
Sapienza University of Rome
carrasco.2177282@studenti.uniroma1.it
https://github.com/Cam1703

Abstract

Tennis is a rich and complex sport where numerous variables can influence the outcome of a match. For this reason, understanding tennis players performance requires sophisticated analysis of complex, multi-dimensional data spanning different surfaces, tournaments, and playing conditions. This paper presents a comprehensive visualization dashboard designed to analyze tennis player performance data, with a particular focus on serve metrics and surface-specific outcomes. The tool offers an interactive interface featuring principal component analysis of serve characteristics, parallel coordinates for match metrics, seasonal distribution analysis, and surface-specific performance indicators. Through data visualization techniques, users can explore insights about player adaptation across surfaces, serving patterns, and tournament-specific performance trends.

Keywords: Visualization Tool · Visual Analytics · Tennis · Sports · Player Performance

1 Introduction

Professional tennis has undergone a dramatic transformation in recent decades, evolving from a sport primarily guided by intuition and experience to one increasingly driven by data-informed decision making. The emergence of sophisticated tracking systems, high-speed cameras, and advanced statistical analysis has generated unprecedented volumes of performance data. However, the true challenge lies not in data collection but in transforming this wealth of information into actionable insights that can enhance player performance, inform strategic decisions, and deepen our understanding of the game [1].

Visual analytics has emerged as a crucial tool in bridging the gap between raw tennis data and actionable insights. By leveraging interactive visualization techniques, analysts can reveal patterns and relationships that might be obscured in traditional statistical analysis, enable intuitive exploration of complex, multi-dimensional data, facilitate communication of insights to players, coaches, and other stakeholders, and support data-driven decision making in training and competition strategies.

The tennis_atp GitHub repository [2] provides significant amounts of data about ATP tennis matches spanning several decades. This comprehensive dataset includes multiple CSV files categorized by match type (singles, doubles, challengers, qualifiers) and contains key features such as tournament data, player statistics, and match details. For the purposes of this paper, we focus on data from male top players for the years 2010-2019, allowing us to explore the evolution of the game over a crucial decade that saw significant changes in playing styles and strategies.

We present a novel interactive visualization framework designed to transform complex match data into clear, actionable insights. Our proposed tool enables users to analyze match patterns, player performance, and strategic trends through a series of interconnected, interactive visualizations. The framework focuses on:

1. Dynamic representation of player performance metrics across multiple dimensions
2. Comparative analysis of playing styles and strategies across different surfaces
3. Pattern recognition in head-to-head matchups
4. Temporal analysis of strategic trends and gameplay evolution

Our visualization framework distinguishes itself through its ability to integrate multiple data streams into coherent, interactive displays that support both high-level pattern recognition and detailed statistical analysis. The approach makes complex statistical relationships accessible to a broad range of users, from professional coaches to tennis enthusiasts.

In the following sections, we present the methodology behind our visualization framework's development, demonstrate its key features through case studies, and discuss its potential applications in professional tennis analysis and training. We also examine how this approach can contribute to our understanding of the sport's tactical evolution and help predict future trends in playing styles and strategies.

2 Related Works

Many tennis visualization systems have been developed over the last years, some of them requiring advanced tools for collecting the data required for the analysis *i.e* ball tracking technologies and player tracking.

TenniVis [3] is a visualization tool that requires only a single video camera to manually collect match data and analyze in detail every single set played in a match. The user is able to filter the data in several ways and also view a short video clip for each point scored, but the interaction and the analysis is limited to a single match at a time and doesn't allow for comparison of multiple matches or historical analysis.

HotShots [4] provides a visualization tool that displays a court view and allows the user to analyze each individual match in detail by emphasising a stroke-by-stroke analysis. The user can visualize how each individual stroke has been performed on the tennis court, its trajectory and the outcome.

CourtTime [5] is an interactive visualization tool designed for analyzing individual tennis matches. In particular, the tool provides the ability to analyse player positioning and shot patterns over time, allowing insights and patterns to be discovered within a particular match based on the player's play style.

All the mentioned tools offer the possibility of studying tennis matches and individual player performances in depths. In particular, they all use different datasets depending on the specific needs of the visualisation tool used. Moreover, none of them offer the possibility to perform a historical analysis of individual player performances or a simple global comparison between multiple matches. We believe that both of these aspects are crucial for studying the evolution of tennis matches over the past years and discovering additional insights for tennis players.

3 Methodology

The proposed visualization tool is a React-native application providing a dashboard that allows users to analyze and gain insights for each individual player. In what follows, we will describe all the steps involved in data preprocessing, the selected dimensionality reduction technique and the visualization and interaction techniques that have been employed.

3.1 Data Pre-Processing

In this section, we explain the steps involved in data processing, including the dimensionality reduction technique. The original dataset is formed by csv files divided by year, where each file contains data from all the matches played in the ATP circuit throughout the season. For each match, we have some general data of the match (tournament, round, surface, duration, score) as well as columns representing attributes of both players, such as: ranking, aces, double faults, break points faced, and others. Firstly, we selected only the seasons between 2010 and 2019; also, we removed tournaments smaller than ATP250 and played in formats different than usual (such as the Laver Cup and the Davis Cup). At this point, we had over 25.000 tuples with 51 columns each. To make the data feasible for our application, we applied two other processing steps.

First, in order to reduce the number of columns, we transformed the dataset to contain only the performance attributes of one player at a time. Since the original dataset represented both players attributes simultaneously, the player attributes were duplicated (each one for the winner and the loser). Then, we also had to reduce the set of players to be shown in the dashboard. For that, we ordered the players by the number of matches won and the percentage of matches won in the selected time frame and selected the top-20 players. The final dataset corresponds to all matches played between 2010 and 2019 by these top-20 players, whether the player has won or lost the match.

For the dimensionality reduction technique, we chose the Principal Component Analysis (PCA) to generate a resumed representation of the serve attributes. We chose PCA over t-SNE and MDS because of the interpretability of the results: we wanted to understand which characteristics of the serve could be found by combining the most famous serve attributes. The serve metrics used in the PCA method are:

- Number of aces
- Number of double faults
- Percentage of legal first serves
- Percentage of points won playing with the first serve
- Percentage of points won playing with the second serve
- Average number of points played in the service games
- Number of break points

Serve metric	First Component	Second Component
Aces	-0.14	0.65
Double faults	0.26	0.60
% 1st serve in	-0.16	-0.23
% 1st serve win	-0.46	0.36
% 2nd serve win	-0.39	0.07
Avg. points per game	0.49	0.06
Break points faced	0.53	0.11

Table 1: *Loading values for 1st and 2nd components*

After executing the PCA method, we got the loading results showed on table 1. On the first component, we have positive values for the attributes (on descending magnitude):

- Number of break points faced
- Average number of points played per service game
- Number of double faults

and negative values for:

- Percentage of points won playing with the first serve
- Percentage of points won playing with the second serve
- Percentage of legal first serves
- Number of aces

From this grouping of features, we can interpretate the first component as a measure of the inefficiency of the serve (since negative values express serve attributes that contribute to the player winning the point and positive values represent attributes related to the player having trouble in the service games).

For the second component, also showed on table 1, the only attribute with a negative loading value is the percentage of legal first serves. For the positive loading variables, the most relevant ones are:

- Number of aces
- Number of double faults
- Percentage of points won playing with the first serve

If we think carefully about this attributes, we can associate this component with the risk or the level of aggressiveness of the serve. When a player is more aggressive on the serve, he can increase the number of aces and amount of points won with the serve. However, with a greater risk, he can also have a higher number of double faults. Hence, the two principal components found using the serve metrics represent a measure of serve inefficiency (first component) and risk aggressiveness (second component).

3.2 Visualizations

In the following section, we will offer a detailed explanation of the visualization and interaction techniques employed in the application. All the discussed graphs are interactive, since they offer the user the ability to select and filter the data on display, with the result of the interaction being reflected on the other graphs. It's important to notice that the data represented in each chart refers only to the currently selected player. The user can use a dedicated button to open a menu that allows to switch between different players and change the data that will be displayed in each chart.

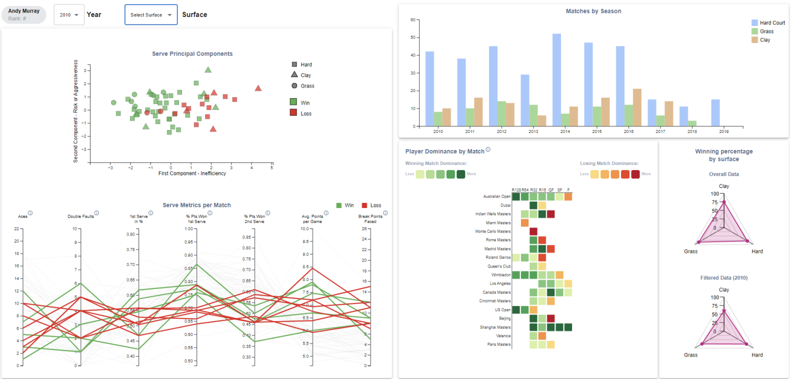

Figure 1: Project Dashboard

3.2.1 Scatter Plot Chart

The scatter plot chart contains data points resulting from the application of the PCA dimensionality reduction technique. Each data point represents a match played by the player in the selected year, it can be represented with three different shapes (one for each surface) and the color of the point describes the outcome of the corresponding match (green for win, red for loss). The user can select data points by brushing specific parts of the chart and the selected points will be highlighted also on the other charts.

3.2.2 Parallel Coordinates Chart

The parallel coordinates chart allows to visualize some categorical attributes related to the **serve metrics** of the currently selected player, in order to find insights about the strenghts and weaknesses caused by the surface but also to find insights related to matches won/lost with unexpected levels of dominance. Each line corresponds to an individual match played and the color of the line describes the outcome of the corresponding match (green for win, red for loss). The chart is updated according to the selected filters (year and surface) and any matches are highlighted in another charts. The user is able to interact with this chart by brushing onto the axes and highlight some specific matches. The categorical attributes that are shown in this chart are the following: *aces*, *double faults*, *1st Serve in* % (percentage of 1st serve attempts that were inside the service box), % *Pts.Won 1st Serve* (percentage of points won when the player hit a 1st serve), % *Pts.Won 2nd Serve* (percentage of points won when the player hit a 2nd serve), *Avg. Points per Game* (average number of points played on the player's service games) and *Break Points Faced* (number of opportunities for the opponent to break the player's serve).

3.2.3 Bar Chart

The bar chart allows to visualize the total amount of matches played by the selected player for each season. In particular, for each year the chart displays the total amount of matches played divided by surface: *Hard*, *Grass* and *Clay*. Everytime the user selects a different player, the chart is recomputed with the new data of the new selected player.

3.2.4 Heatmap Chart

The heatmap chart allows to analyze every tournament played by the selected player on a specific year and, in particular, it allows to analyze the level of **dominance** for every single match played. Each row of the heatmap corresponds to a tournament played in the currently selected year and each column corresponds to a stage of the tournament: Round of 128 (*R128*), Round of 64 (*R64*), Round of 32 (*R32*), Round of 16 (*R16*), Quarter Finals (*QF*), Semi-Finals (*SF*) and Final (*F*). Each cell of the heatmap will eventually correspond to a match played in a specific tournament and in a specific stage of such tournament. The heatmap may have some empty cells in some rows, either because the player lost a match and did not advance to the next stage of the tournament or because the tournament did not have all the stages mentioned (for example, some tournaments start at the Round of 64)

We define as **dominance** the percentage of games won by the selected player in a specific match. A high level of dominance will be associated with a high number of games won by the player. This metric is essential to recognize the *skill gap* between players and allows to identify individual matches that may be interesting to analyze. The chart offers two different color palettes: one for winning matches (Green palette) and one for losing matches (Orange/Red palette). The user is able to click on a specific cell to highlight the selected match also on the other charts of the dashboard.

3.2.5 Radar Chart

The radar chart allows to visualize the performance of the selected player for each surface: *Hard*, *Grass* and *Clay*. In particular, the dashboard offers two different radar charts: the first radar shows the average of the winning percentage by surface for all players selected for analysis. The second radar shows only the winning percentage by surface of the currently selected player. Both charts are updated according to the selected year.

4 Discovered Insights

Let's go through the representations in the dashboard and find which interesting information can be extracted. We must always keep in mind that the dashboard focuses on a single player, so we will give some examples of how the same graph can vary among different players.

4.1 Injuries and preferred surface

Firstly, we can analyze the bar chart representing the number of matches played in each surface throughout the seasons. In the following figures 2, 3 and 4, we show the bar chart for the players Juan Martin Del Potro, David Ferrer and Novak Djokovic.

When looking at these graphs, we can identify seasons with a sudden drop of the amount of matches played compared to other seasons. This cases represent seasons in which the players had some serious injury:

- Del Potro - 2010, 2014 and 2015: wrist injury
- Del Potro - 2019: knee injury
- Djokovic - 2017: elbow injury

In the case of David Ferrer, we can see a gradual decrease in the number of matches played, associated with the decline of this career and eventual retirement in 2019.

Another behaviour that can be identified in this graph is the player's preference towards the different surfaces. Due to the overall structure of the ATP circuit, there are more tournaments played on hard courts, followed by clay and than grass. Naturally, this structure impacts the number of matches played on each surface by the players. However, we can see that David Ferrer played almost as many matches on clay as on hard court in most seasons, which indicates that Ferrer preferred to play clay tournaments. On the other hand, Del Potro plays much more on hard courts than on clay, even when compared to Djokovic. Hence, we can see that Djokovic balances between the three surfaces similarly to the general structure of the circuit, whereas Del Potro favours hard court tournaments.

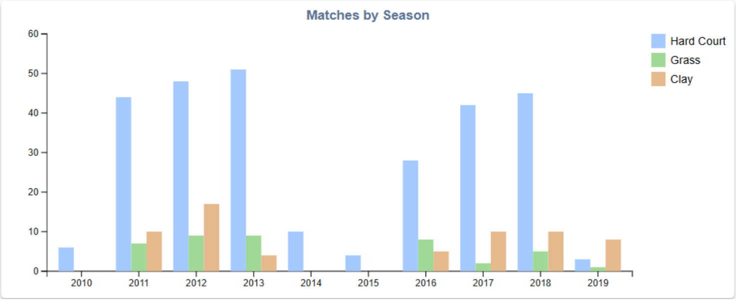

Figure 2: Juan Martin Del Potro

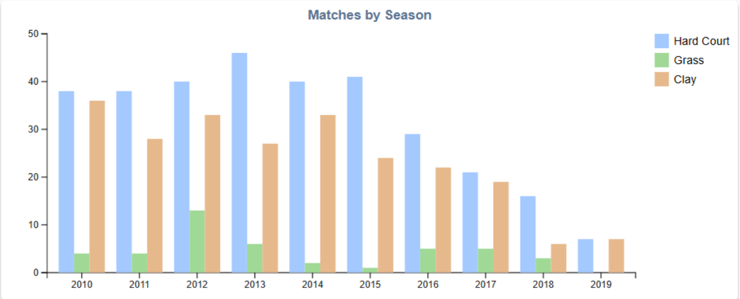

Figure 3: David Ferrer

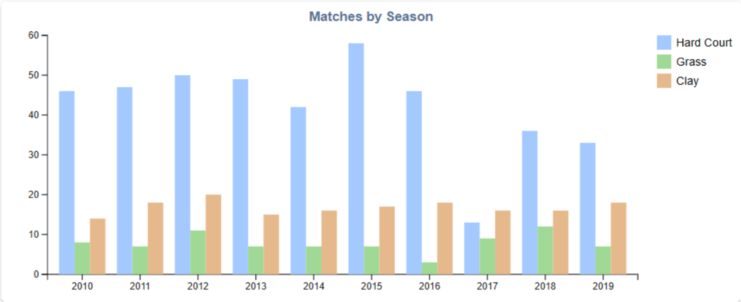

Figure 4: Novak Djokovic

4.2 Serve inefficiency impact on match dominance

A very common question among the tennis community is ”how much does the quality of the serve impact the outcome of the match?”. To assess this question, we can look at the match dominance graph associated with the serve principal components graph. As discussed on section 3.1, the first component represents the serve inefficiency (as the x-axis increases in the graph, the worse is the serve). On figure 5 we show three segmentations of this graph for the hard-court matches played by John Isner in 2010, increasing the serve inefficiency on each segment showed. Below each principal components graph, we have the associated dominance graph for the matches selected. We can see that, as the serve inefficiency increases, the dominance of matches won decreases and the amount of matches lost increases. So, for John Isner, we can say that the quality of the serve is very important for the outcome of the match, as shown by the dominance graph. As different players have different qualities, each player should be assessed individually to understand the intensity of this association.

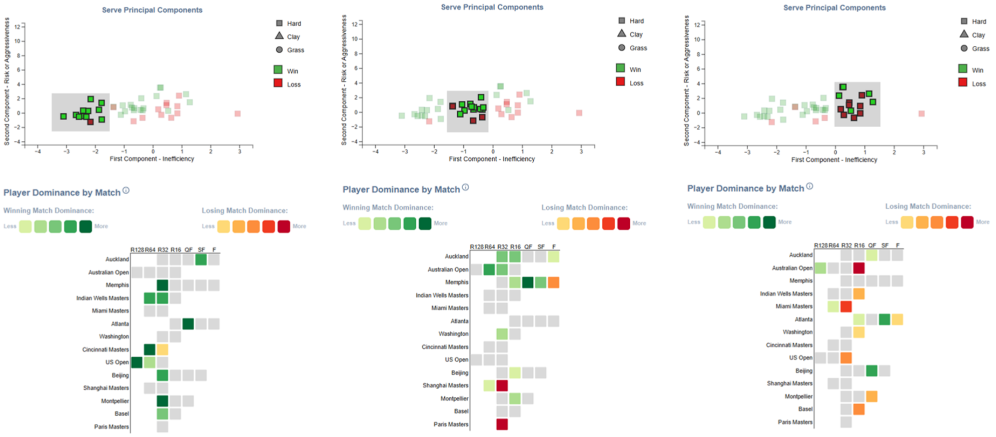

Figure 5: Serve Principal Components and Match Dominance - John Isner - 2010 - Hard Court

4.3 Aggressiveness level influenced by the tournament stage

Another interesting relation that we can observe is the level of aggressiveness depending on the tournament stage. In particular, inside the scatter plot we can notice that most of the points associated with a high level of aggressiveness are usually associated with matches played during the early stages of a tournament. On the other hand, the majority of matches played during the final stages of a tournament (Semi-Final/Final) tend to have a lower level of aggressiveness. For example, as we can see in figure 6 the matches with higher level of aggressiveness played by Andy Murray in 2012 are all related to early tournament stages.

This shows that players tend to be more cautious in the final stages of a tournament, preferring to avoid aggressive serves and sacrificing potential aces to reduce the amount of double faults and secure a win. This behaviour can also be observed for matches played on a surface for which a player has performed particularly well during the year. In fact, as we can see on the figure 7 the radar chart shows that Rafael Nadal's performance (win-rate) on clay was higher than average, but even on a surface where the player is particularly skilled, aggressiveness was not affected and all finals are associated with negative values on the scatter plot.

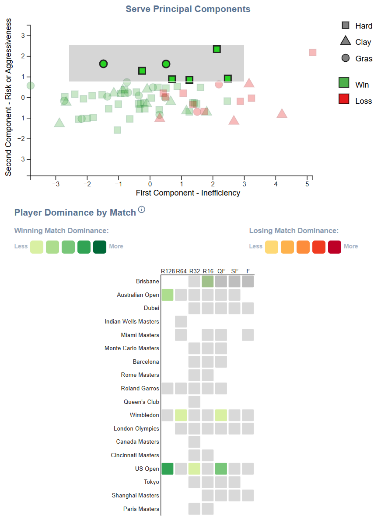

Figure 6: Andy Murray - 2012

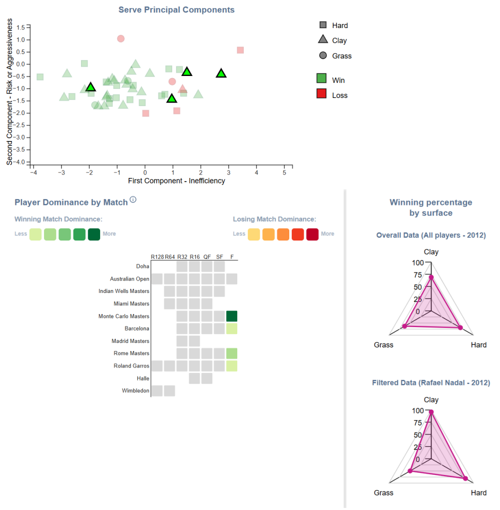

Figure 7: Rafael Nadal - Clay finals in 2012

5 Conclusions

The employed visualization tool proved that a historical global analysis is essential for capturing key insights about tennis players and their playstyle. In particular, it showed how different matches can be related with each other based on several serve metrics attributes and how a player's playstyle and decisions can be influenced. It showed that a player's good performance on a given surface is not exclusively correlated with a high level of dominance but also with the ability to control the match and secure victory.

This tool is not without limitations and future research can focus on addressing them such as considering the whole *tennis_atp* dataset and analyze the performances also on different types of matches (for example, double matches). Moreover, the tool could be incorporated also with a player-by-player comparison in order to allow a more refined analysis of tennis matches and gain additional insights.

References

[1] Liu, Z. (2023). The evolution of tennis from traditional sport to modern phenomenon. *Physical Education and Sport Through the Centuries*, 10(2), 78–92. https://doi.org/10.5937/spes2302078L

[2] Sackmann, J. (n.d.). *tennis_atp* [Dataset]. GitHub. Retrieved on 9 February 2025, from https://github.com/JeffSackmann/tennis_atp

[3] T. Polk, J. Yang, Y. Hu and Y. Zhao, "TenniVis: Visualization for Tennis Match Analysis," in IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 2339-2348, 31 Dec. 2014, doi: 10.1109/TVCG.2014.2346445

[4] Arvind Srinivasan, Abhinav Kannan, Niklas Elmqvist, "HOTSHOTS: Visualizing Stroke-by-Stroke Tennis Data", https://hotshots-v2.vercel.app/.

[5] T. Polk, D. Jäckle, J. Häußler and J. Yang, "CourtTime: Generating Actionable Insights into Tennis Matches Using Visual Analytics," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 397-406, Jan. 2020, doi: 10.1109/TVCG.2019.2934243