NetGain: A Tennis Point Prediction Tool

Jackson Weil, Chris White, Thomas Latawiec, Mia Patrikios, and Matthew Phan

I. Objective

The objective of our project is to turn a vast amount of tennis shot data into digestible information that can aid a player, coach, or fan in predicting the opposition's behavior and giving a strategic edge to the player using our insights.

A. Exploratory Data Analysis

The first and most basic goal is aggregating and plotting the data in a visual way. For example, we could show a heatmap with likelihoods of the next shot based on the previous shot. We could show the likelihood of a certain player serving to a certain spot. In general, we should start our analysis with the simplest means of analyzing the data: counting it.

B. Adding Complexity: Machine Learning

While counting occurrences is valuable, there is much more to be learned about players' strategies from the vast amount of data we have. We plan to use one simple machine learning method such as decision trees and one more advanced method that takes the temporal aspect of a tennis point into account like an RNN or LSTM.

C. Cherry on Top: Reinforcement Learning

To further exploit the temporal nature of each tennis point and to match the game-oriented nature of the data, we hope to train a reinforcement learning agent to either predict an opponent's next move or advise the protagonist's best move.

D. Communicating Results

These insights are only valuable if they can be explored and understood. A stretch goal for our project is to build an app where curious tennis lovers can choose a method described above and visualize its results in a dynamic manner. As this is a stretch goal, the fallback is simply making a series of key static visualizations and presenting them in document or presentation form.

II. Motivation

Many tennis coaches would tell their players to have a short memory. Hit a bad shot? Forget it - focus on the next one. Hit a good shot? Don't rest on your laurels - stay hungry. While this is prudent advice, much can be learned from having a long memory of tennis points. How often does player X serve out wide? How likely is player Y to win the point when hitting a volley? These insights are hard to make in the heat of a match, but they are the kind of insights that could help a coach decide which direction to steer a player and turn a good player into a great player.

Tennis lends itself well to tabular data, as each shot is distinct, there are only two players (or four) to consider, and the court is divided into a few distinct areas. By using a large quantity of relatively simple data, we believe we can draw conclusions about how a player behaves.

A final bit of motivation comes from the fact that tennis has been surging in popularity since the outbreak of COVID [2]. It's one of the few sports in which all players tend to keep pretty far apart!

III. Data

A. Source of Data

The data for this project comes from the SCORE Sports Data Repository. This site hosts large datasets for over 20 different sports. Our focus will be on the Grand Slam Tennis Shot-Level Data.

This dataset is based on Jeff Sackmann's Tennis Match Charting Project. It is designed to be suitable for assessing player decision making.

B. Format of Data

The dataset is separated into 8 files with the same format. There are 4 major tennis tournaments around the world each year: the Australian Open, Roland Garros, Wimbledon, and the US Open. As there is a men's league (ATP or "Association of Tennis Professionals") and a women's league (WTA or "Women's Tennis Association") playing in each of these tournaments, there are 8 separate data files.

The data come from matches from there major tournaments, with a bias toward including more popular players and more critical matches (i.e. closer to the championship match of each tournament). Each row of data represents a single tennis shot. See Table I to see how many shots were collected from each tournament.

There are 17 columns in each of the 8 datasets. The following 6 columns describe match level data: Date, Tournament, Round, Player 1, Player 2, and Point (count of points in the match).

There are also 4 columns containing point level data: Shot, Serve, ServingPlayer, and WinningPlayer.

The finest layer of detail is an individual shot. Each shot is described by the remaining 7 columns: ShotHand, ShotType, ShotDirection, ServeDirection, ShotDepth, OutcomeType, and ErrorType.

This is the dataset. A sequence of rows related to the same point can tell the story of short-term strategy between two players. A sequence of points in a match can provide insight into the long-term strategy between two players. Our hope

TABLE I
SHOT COUNTS RECORDED IN ATP AND WTA GRAND SLAM MATCHES [1].

ATP Shot Counts	
Tournament	**Shots**
Australian Open	635,064
Roland Garros	480,872
US Open	587,466
Wimbledon	457,591
WTA Shot Counts	
Tournament	**Shots**
Australian Open	291,544
Roland Garros	201,414
US Open	237,832
Wimbledon	190,204

is to gain some understanding of these strategies with the mathematical techniques mentioned in this proposal.

IV. RESPONSIBILITIES

A. App Development

While housing our analysis tools and figures in a clean web application is not an absolute must for the project, it makes sense to work on this goal from the start. It would be a huge benefit in terms of learning and wow-factor. For that reason, Chris will take the lead on app development and Jackson will help.

B. EDA

Jackson, Thomas, Mia, and Matthew will all start doing exploratory data analysis on their own before comparing results and choosing some favorite insights or figures to guide the rest of the project. In this way, most of the team will become very familiar with the dataset, which will help for the machine learning that follows.

C. Machine Learning

Jackson and Thomas will attempt to build an RNN model to predict the next shot in a tennis point, while Mia and Matthew will attempt to build an LSTM model for the same purpose. After a prototype of each is compared, the best model (by measure of prediction accuracy and understandability) will continue to be refined by the two who built it.

D. Reinforcement Learning

The two who do not continue refining the ML model, along with Chris, will work on the reinforcement learning agent.

E. Documenting Results

Jackson will lead the organization of the final paper, but everyone will contribute based on what they worked on.

V. MILESTONE TIME-LINE

Early October

The first two weeks of October will be spent starting the app development and doing exploratory data analysis. This should wrap up on October 16.

Late October

The last two weeks of October will be spent building the prototype LSTM and RNN models for temporal tennis point prediction. Additionally, the work done in the EDA stage will be fused into the web app.

Early November

The first two weeks of November will be spent tuning one of the two machine learning models and building a reinforcement learning prototype.

Late November

The last two weeks of November will be spent documenting results, touching up the machine learning and reinforcement learning models, polishing the app, and preparing to present our findings.

VI. EXPECTED OUTCOME

Mandatory

We expect to present a machine learning model that can help tennis lovers to predict what will happen next in a tennis point given what has happened in the past.

Stretch Goals

We hope to deliver a reinforcement learning agent (with a similar purpose as defined above) and a web app that makes our tools usable in a dynamic and visually appealing way.

VII. ATTRIBUTION

OpenAI's ChatGPT was used to create Table 1 (all numbers were checked against those recorded by the data providers). ChatGPT was also used to generate both bibliographic entries and for some minor formatting syntax, namely the hyperlinks.

REFERENCES

[1] OpenAI, "ChatGPT (Oct 2025 version)," San Francisco, CA: OpenAI. Available: `https://chat.openai.com/`

[2] USTA, "U.S. tennis participation surges to new high of 25.7 million players following five consecutive years of growth," USTA, Feb. 19, 2025. Available: `https://www.usta.com/en/home/stay-current/national/u-s-tennis-participation-surges-to-new-high-of-25-million-html` (accessed Oct. 2025).