

# SMASH DATA

**SERVING UP INSIGHTS: A DATA-DRIVEN LOOK AT TENNIS  
PERFORMANCE AND BETTING RETURNS**

Work done by Ellen Dagher, Karine Rafla, Mohammed Al-Hussin in the scope of the EPFL  
COM-480 Course

Date: May 30, 2025

# Project Goal and Motivation

We set out to visualize over two decades of professional tennis by building an interactive website that lets users explore trends across both the WTA and ATP circuits. Rather than relying on static dashboards or basic statistics, we wanted to craft an experience where users can uncover player trajectories, betting anomalies, geographic patterns, and surface-based performance, all through smooth, animated, and exploratory visualizations.



Our plots focus on different layers of tennis insight: from ELO rating evolution and underdog upsets, to finals history and match distributions across time and space. One key visualization is a geographic map showing the number of matches played in each city, along with the top player and dominant court surface, offering a spatial narrative of tennis activity worldwide.

To guide exploration, the website is structured around a simple and intuitive choice: users first select whether they want to explore **WTA** or **ATP** data. This decision dynamically filters all visualizations, ensuring that the insights are relevant to the selected tour. Whether a user is interested in women's tennis, men's tennis, or both, the structure provides a clean entry point into their area of interest.

## Data and Technical Stack

### Datasets

We used two complementary datasets to build a comprehensive, tour-wide analysis of professional tennis from 2000 to 2024:

-  **Jeff Sackmann's tennis\_atp dataset** (via GitHub)  
A historical dataset of men's tennis matches, including rich tournament and player metadata. It contains match scores, set details, ELO-relevant stats (e.g., aces, double faults, break points), and player attributes such as nationality, height, and dominant hand. We limited our analysis to the 2000–2024 range.
-  **Ultimate Tennis Matches Dataset** (via Kaggle)  
Covers both ATP and WTA circuits from 2000–2024 with detailed match-level records and tournament data (name, surface, location, category). Crucially, it also includes **sports betting odds** from various bookmakers, enabling our underdog return analysis.

We merged, cleaned, and enriched both datasets to produce our final working data. This included calculating custom ELO ratings for both tours and computing betting returns per match.

### Tools & Technologies

- **Python, Pandas** – For merging datasets, ELO calculation, betting return derivation, and cleaning inconsistencies across sources.
- **Plotly + Leaflet** – Used for the betting scatter plot and the geographic map (integrated as HTML exports).
- **JavaScript (vanilla)** – Used for the ELO evolution timeline and the head-to-head comparison to allow tighter control over animations and layout.

- **HTML/CSS/JS** - To structure and style the website interface and manage tour/tab navigation.
- **GitHub** - For collaboration, versioning, and public deployment of our final site.

## Exploratory Data Analysis

Before finalizing the core visualizations, we conducted extensive **exploratory data analysis (EDA)** on both datasets to uncover trends, validate our hypotheses, and identify valuable features.

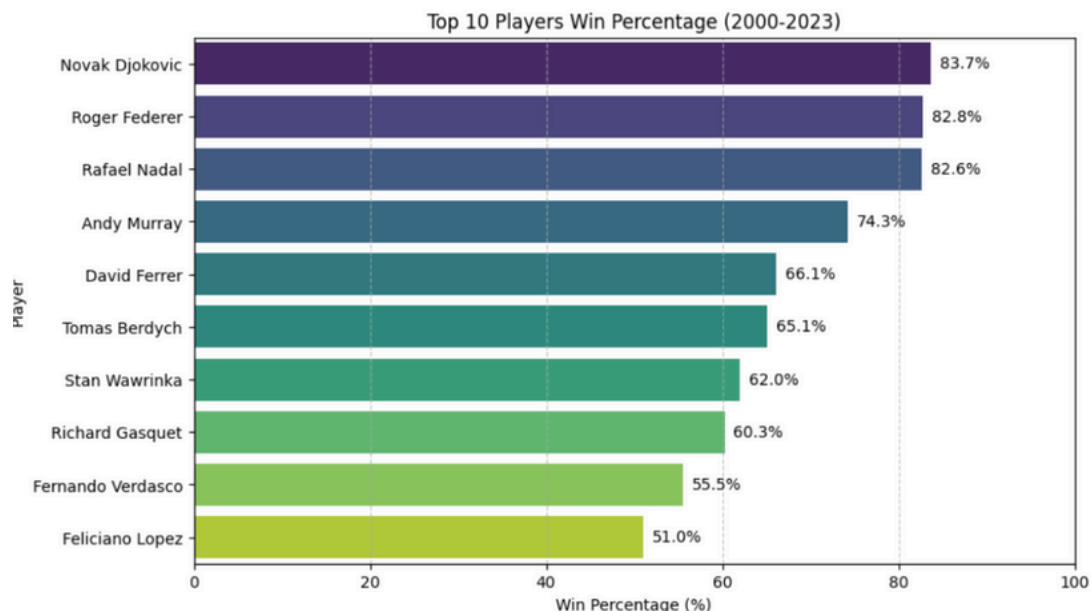
We used **Python (Pandas, Seaborn, Matplotlib)** to explore:

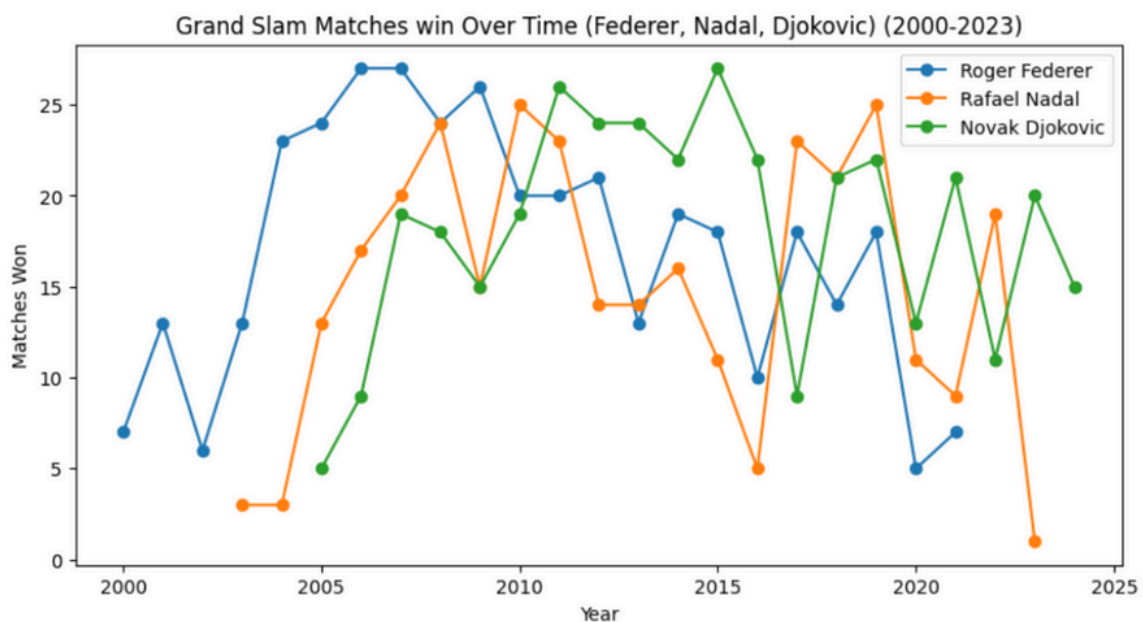
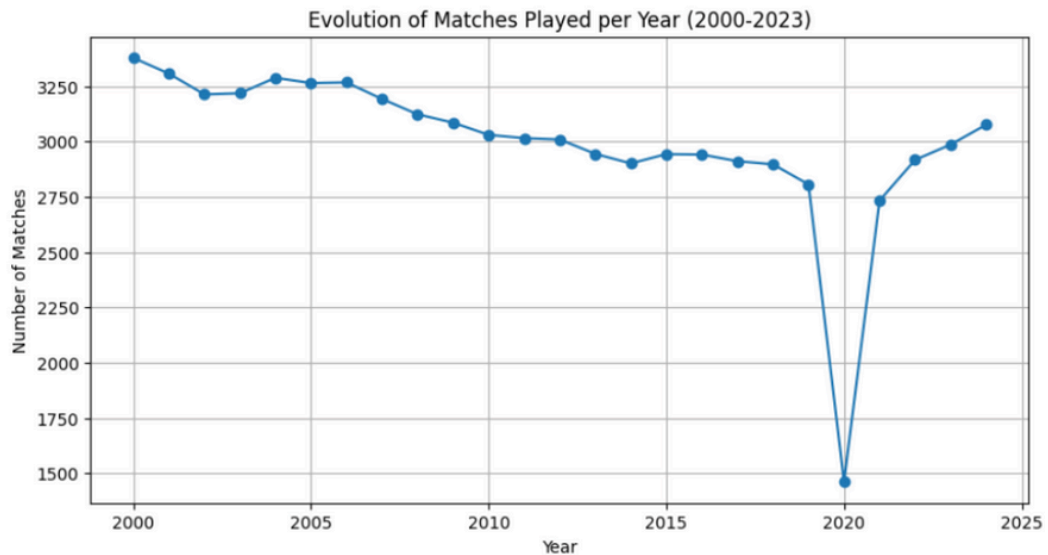
- Match distributions across surfaces, years, and player ranks
- Win percentages by court type
- Top players by match count and consistency
- The behavior of betting returns and rank differentials in upsets
- Player-level statistics like serve percentages, aces, and match durations

This step was crucial to:

- Identify useful variables for interactivity (e.g. surface, year, opponent)
- Decide which stories were **worth telling** (e.g. upsets, dominance, geography)
- Avoid cluttered or redundant plots in the final interface

Below are a few representative plots from our EDA phase:





This collaborative EDA phase shaped the entire design direction of the project — helping us go beyond basic statistics and craft visualizations that told deeper, clearer stories.

## Key Visualisations

Our visualizations are structured around three thematic exploration areas — **Bets**, **Player Evolution**, and **Around the World**, available for both **ATP** and **WTA**. Upon selecting a tour, the user is presented with the same interactive components adapted to that dataset.

Each section includes filters (e.g., surface, year, player) to allow personalized exploration. Here's a breakdown of the key visualizations:

### 1. Underdog Upsets vs. Top 10 Players

**Type:** Interactive Scatter Plot  
**Section:** *Bets*

This plot explores where tennis shocks really happen, every time a lower-ranked underdog beats a top-10 opponent. Each point represents a unique upset, where the X-axis shows the underdog's ranking and the Y-axis displays the betting return for that match.

**Interactive Features:**

- Filter by **Year**, **Surface**, and **Top-10 Opponent**
- Clickable buttons for common upset victims
- Hover tooltips reveal player names and match context

**Insights Gained:**

This visualization uncovers hidden patterns: players who are frequently upset, underdogs who consistently deliver value, and the surfaces where upsets are more common. For instance, betting return spikes are often tied to early-round hard-court matches, where newer players surprise seasoned favorites. Below we can see one of the first drafts of the bets/underdogs plot:



*Underdog Wins vs Top 10 Player for WTA*

**2. ELO Rating Timeline (2000–2024)**

**Type:** Animated Line Chart  
**Section:** *Player Evolution*

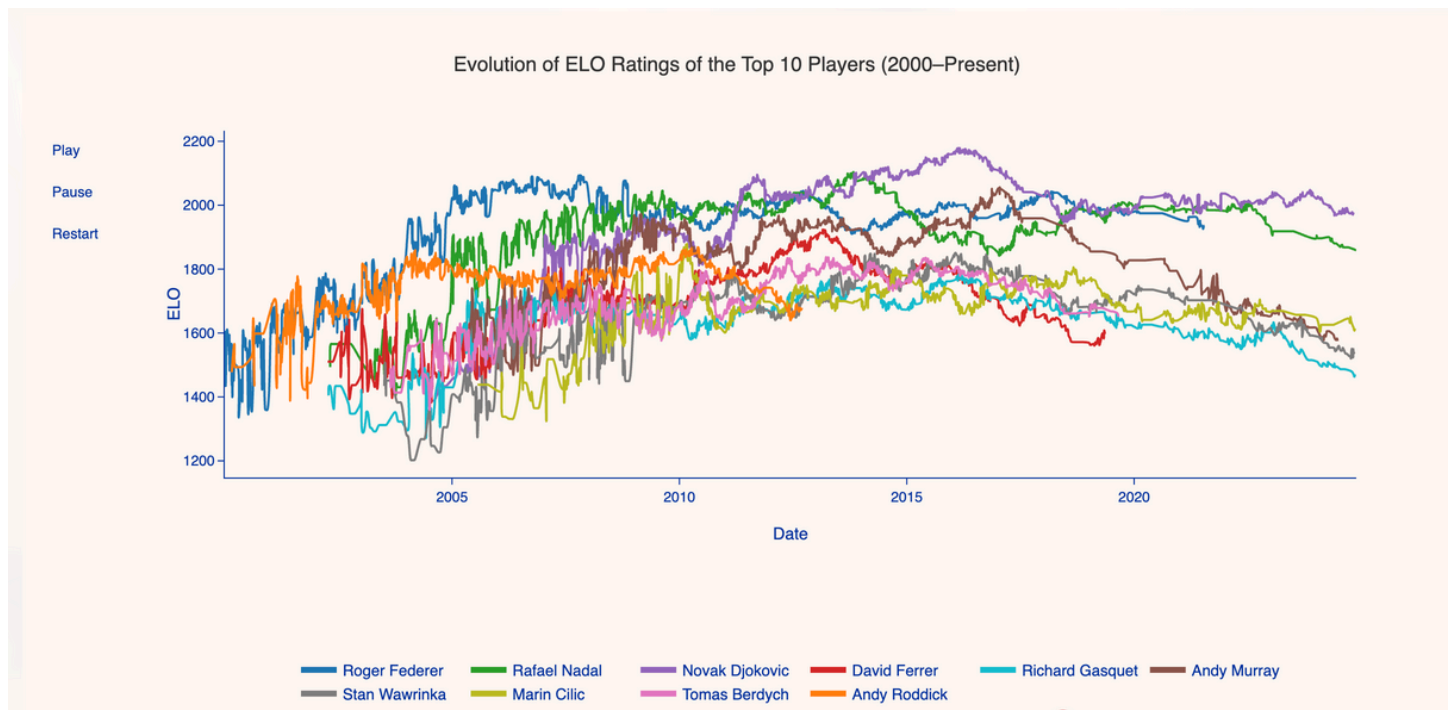
This visualization animates the evolution of ELO ratings for the top 10 players across 24 years. It captures the rise and fall of legends, the transitions between tennis eras, and the longevity of dominant figures.

### Interactive Features:

- **Play/Pause/Restart** controls for timeline animation
- **Dynamic legend** updates as new players enter the top 10
- Available for both ATP and WTA, with synchronized structure

### Insights Gained:

The official point system often fails to reflect who's truly dominating on court. That's where the ELO rating system shines: it dynamically adjusts to wins and losses based on opponent strength, rewarding quality victories and penalizing unexpected defeats, regardless of calendar structure or surface. Our animated visualization, covering 2000 to 2024, brings this to life — letting you witness the rise and fall of legends, the brief flashes of form from short-lived stars, and the enduring dominance of greats like Serena, Federer, Nadal, and Djokovic. We chose Plotly.js over D3.js to build this because it enables smooth, responsive, and interactive timeline animations with a fraction of the code. Technically, we chose **Plotly.js** over D3.js because it allows smooth timeline animations with minimal code. While D3 often requires 700+ lines for a single animated chart (see [this GitHub repo](#)), Plotly handles interactivity, transitions, and responsiveness natively. The result? A clean, intuitive, and immersive way to explore tennis history — without the overhead.



*ELO Rating Evolution for ATP*

## 3. Geographic Match Distribution Map

**Type:** Interactive Leaflet Map

**Section:** *Around the World*

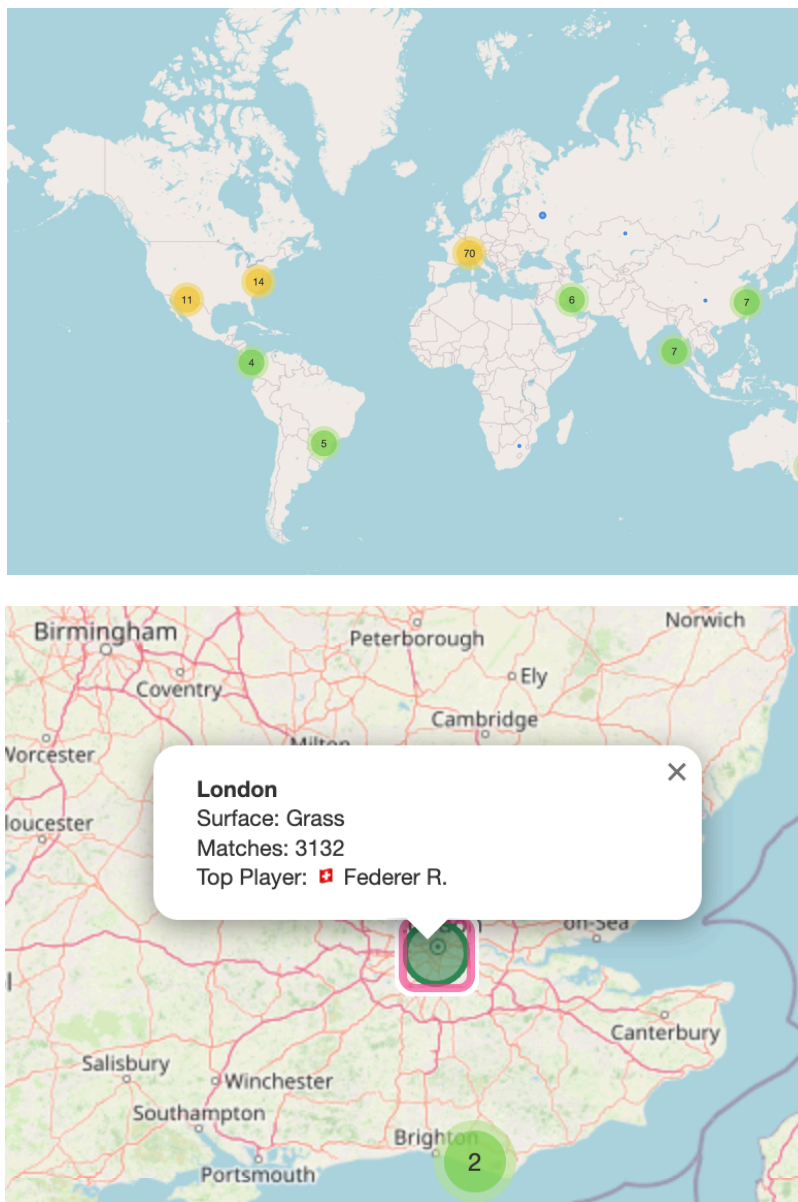
This global map shows where tennis is played, and won. Each city is marked with a bubble sized by the number of matches played there. The color represents the dominant surface type (clay, grass, hard), and the tooltip reveals the most successful player in that city.

### Interactive Features:

- Zoom/pan across continents
- Surface color legend and live hover stats
- Separate versions for ATP and WTA circuits

### Insights Gained:

This plot adds a real-world layer to tennis analytics. It shows the centrality of European cities on clay, the hard-court density in North America and Asia, and highlights key city-player relationships (e.g., Nadal in Monte Carlo). It's also useful for spotting geographical imbalances in match frequency or court diversity.



Match Distribution Map for ATP



# Design Evolution and Decisions

As we moved toward the final version, our main design philosophy was to group visualizations by **thematic relevance**. This led to a clear and engaging structure, where each tab — *Bets*, *Player Evolution*, and *Around the World* — represents a distinct angle of exploration. It also gives users the freedom to choose what they want to explore, without being overwhelmed by too many visuals on a single page.

We applied the same principle to the **tour split**: WTA and ATP data are presented in two distinct views, allowing users to dive deep into either circuit without confusion. This separation not only preserves visual clarity, but also improves performance and keeps the storytelling focused.

We consciously chose to remove traditional, obvious, or redundant statistics (such as win/loss bar charts or pie graphs). Instead, we focused on building **three core, interactive functionalities** that offer richer insight and a more enjoyable experience — even for users who aren't tennis experts. The visual design, interactivity, and filtering options were selected to make the platform accessible to casual viewers while still offering value to hardcore fans.

One significant addition made after Milestone 2 was the **Geographic Map**. Initially unplanned, we included it because it offered a spatial perspective that was missing from our other plots. It proved both useful and aesthetically engaging, allowing users to explore tennis history through city-level activity, surface types, and local player dominance — adding depth and visual variety to the site.

We also made deliberate **color choices** to reinforce clarity and user intuition. Specifically, we used **blue for ATP** and **pink for WTA** throughout the site. While these color associations are somewhat stereotypical, they proved effective in helping users immediately distinguish between the two tours. In informal testing, this color coding improved orientation and consistency — especially when switching tabs or comparing plots side-by-side. The palettes were chosen to be visually distinct but not overwhelming, reinforcing the idea that users are exploring two equally rich but separate tennis ecosystems.

We also initially developed a **Head-to-Head Comparison View** that allowed users to select two players and compare their ELO progression, match stats, and direct rivalry outcomes — designed in a duel-inspired, game-style interface. While we were excited about its potential for storytelling and user engagement, we ultimately chose not to include it in the final site. The main reason was that the component didn't deploy cleanly alongside the rest of our site architecture, and maintaining design consistency proved too difficult in the available time. Nonetheless, this prototype helped inform how we presented comparative stats and interactivity in the rest of the project.

## Technical Implementation & Challenges



From the beginning, we wanted the site to be smooth, responsive, and easy to use, while delivering advanced interactivity and storytelling. Our initial plan was to implement all visualizations using Plotly, which conveniently generates self-contained HTML files that we could embed in our main index.html.

This worked well for some visualizations , especially the Betting Scatter Plot and the Geographic Map , which integrated cleanly using Plotly’s HTML exports. However, we encountered aesthetic and layout issues with more complex, animated plots.

In particular, the ELO Evolution required more precise control over interactivity, animation pacing, and layout responsiveness. So for this, we switched to a custom JavaScript implementation. This gave us more flexibility to fine-tune animations, manage state between UI elements, and ensure consistency across both ATP and WTA versions.

While this hybrid approach increased the complexity of the codebase, it allowed us to maintain visual polish and ensure that each visualization could be optimized independently , balancing Plotly’s convenience with JavaScript’s flexibility.

One minor issue remains in the underdog betting plots: when the page first loads, the plot doesn’t always render at full width. This is a display glitch caused by the way Plotly’s static HTML handles sizing inside our embedded layout. The plot becomes fully visible as soon as the user clicks on any player name in the legend, which triggers a redraw. Since the data is intact and the fix is straightforward for users, we chose to keep the plot as-is and include a brief instruction in the interface to ensure smooth navigation.

## Peer Contributions

Our project was a highly collaborative effort, with each team member taking ownership of one major visualization component while collectively shaping the structure, design, and integration of the website. Here's a breakdown of individual responsibilities:

Team Member	Main Contributions
Ellen	Developed the <b>Underdog Upsets vs. Top 10</b> betting plots (ATP & WTA), including filters, interactivity, and styling.
Karine	Built the <b>Geographic Match Distribution Map</b> , handled surface color coding, top-player tagging, and Leaflet integration.
Mohammed	Implemented the <b>ELO Rating Timeline</b> using custom JavaScript for animation and layout control.

In addition to these individual tasks, **all three of us actively collaborated** on the structure of the website, user journey design, aesthetic decisions, and overall narrative flow. From layout tweaks to tab structuring, color choices, and tour toggles, the final product reflects a shared vision and continuous teamwork. We also **jointly conducted exploratory data analysis** at the beginning of the project to better understand the datasets, define relevant features, and shape the design of our four core visualizations.

## How to run the website locally

Run the following in your terminal or command line:

```
git clone https://github.com/com-480-data-visualization/SmashData.git
```

```
cd docs
```

```
python -m http.server 8000
```

You can now access the website locally by typing `http://0.0.0.0:8000/` into your browser.