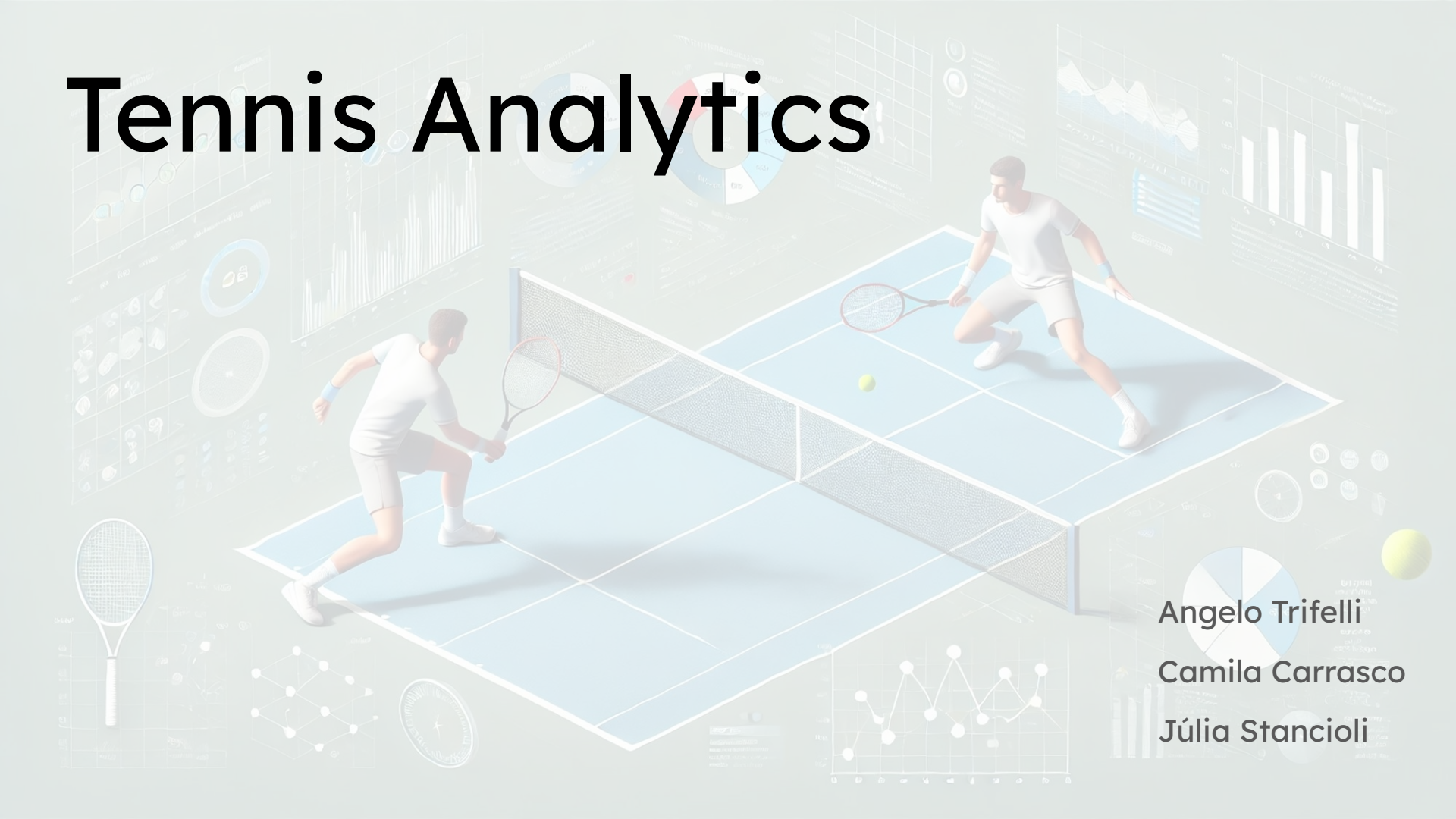
Tennis Analytics



Angelo Trifelli

Camila Carrasco

Júlia Stancioli

Project goals

- This project aims to analyze the individual performance of the top-20 tennis players of the last decade (2010 - 2019)
- We want to provide insights about which attributes most distinguish different high-performance players and what major changes and trends happened in the sport over the last 10 years
- Finally, we want to offer insightful information about the ATP circuit for tennis analysts, coaches, journalists and sport enthusiasts

Data structure

- In the original database, each dataset represents a year in the ATP circuit, where the rows are matches and the columns are match attributes (duration, score, round, …) or player attributes (age, ranking, break points faced, double faults, …)
- We applied several data processing steps to produce one dataset for each of the top-20 selected players, for seasons 2010 - 2019
- The final datasets only represent information about our targeted players, regardless of the result of the match (winning or losing)

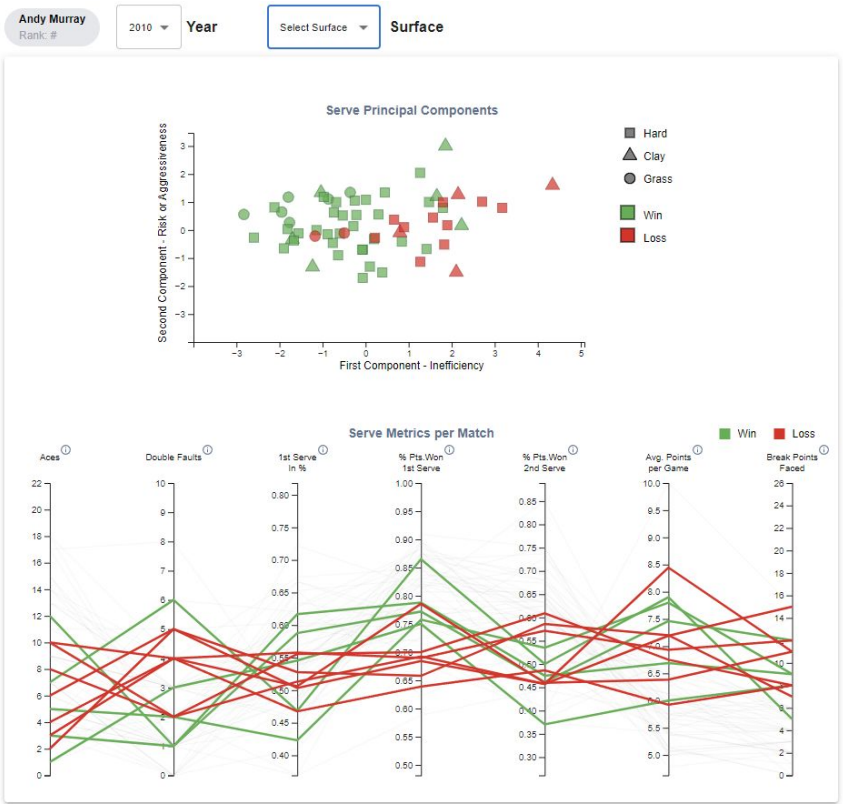
Dimensionality Reduction - PCA

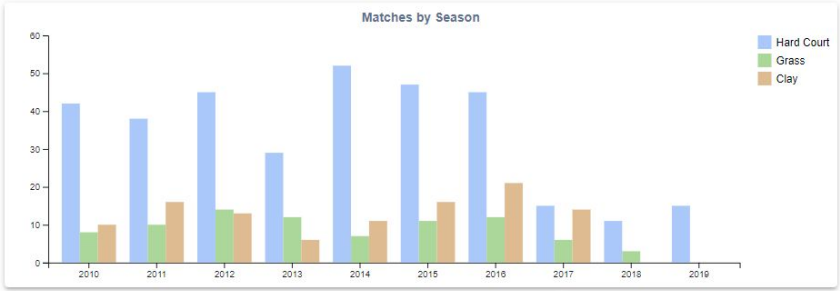
- We applied the PCA to generate an interpretable 2-dimensional representation of 7 serve attributes:
  - aces
  - double faults
  - percentage of legal 1st serves
  - percentage of points won with 1st serve
  - percentage of points won with 2nd serve
  - average number of points played per service game
  - number of break points faced
- The first principal component represents the inefficiency of the serve, while the second represents the risk or aggressiveness of the serve

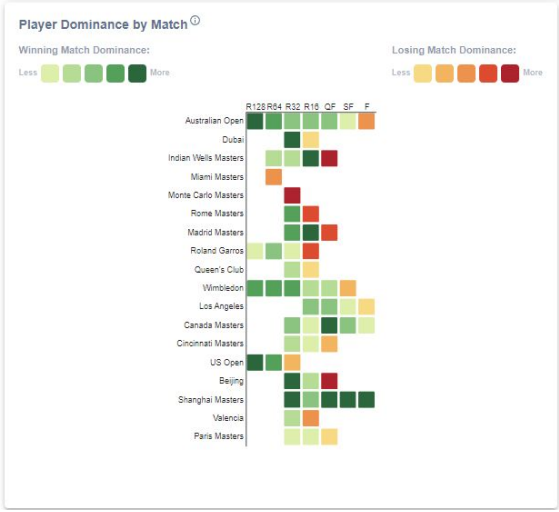
Challenges

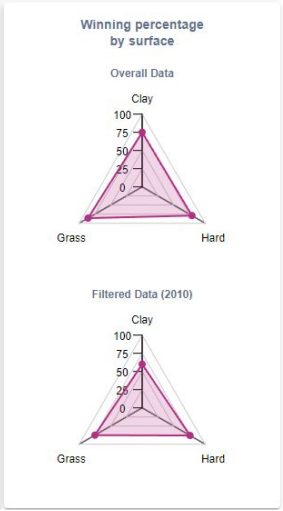
- The original dataset was too large to be used in the application, demanding several transformations:
  - Application of filters (reducing time period, choosing specific columns)
  - Selection of a subset of players
  - Attributes of one single player per instance, instead of winner and loser simultaneously
- To select the top-20 players of the decade, we sorted all players by the total match wins and the average win rate over that time span
- After applying the PCA, we had to carefully examine the loadings associates with each feature to understand what characteristic was being represented by each component

Overview


Andy Murray
Rank: #
2010
Year
Select Surface
Surface
Serve Principal Components
Second Component - Risk or Aggressiveness
First Component - Inefficiency
Hard
Clay
Grass
Win
Loss
Serve Metrics per Match
Win
Loss
Aces
Double Faults
1st Serve In %
% Pts.Won 1st Serve
% Pts.Won 2nd Serve
Avg. Points per Game
Break Points Faced


Matches by Season
Hard Court
Grass
Clay
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019


Player Dominance by Match
Winning Match Dominance:
Losing Match Dominance:
Less
More
R128 R64 R32 R16 QF SF F
Australian Open
Dubai
Indian Wells Masters
Miami Masters
Monte Carlo Masters
Rome Masters
Madrid Masters
Roland Garros
Queen's Club
Wimbledon
Los Angeles
Canada Masters
Cincinnati Masters
US Open
Beijing
Shanghai Masters
Valencia
Paris Masters


Winning percentage by surface
Overall Data
Clay
Grass
Hard
Filtered Data (2010)
Clay
Grass
Hard

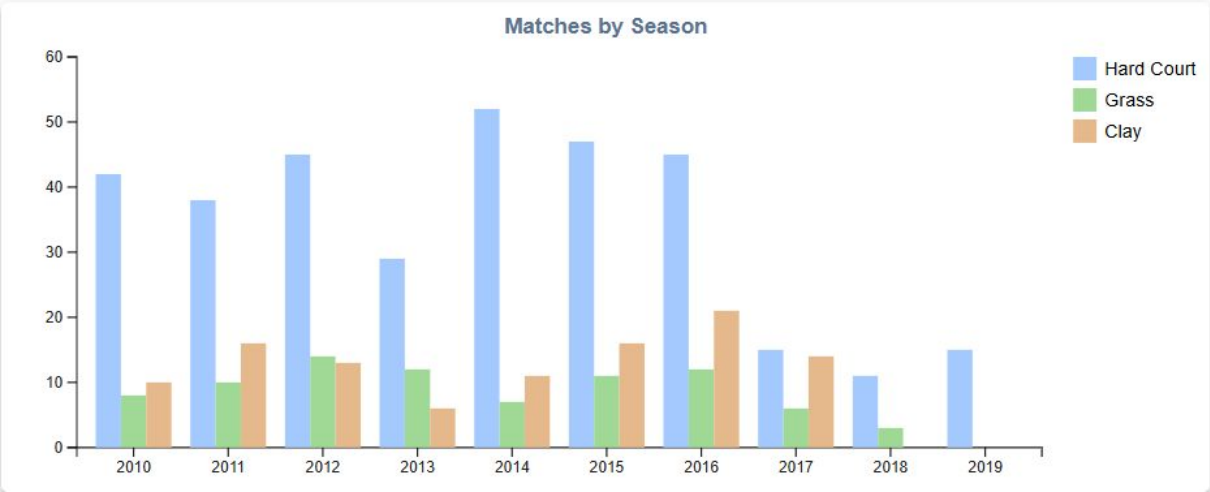
Technologies

- **Python**: scripting programming language used to pre-process the data, apply the dimensionality reduction technique and generate the required files used by the application.

- **React**: javascript library used to build the structure of the frontend
  - Reusable UI components.
  - Support of simple user interactions.
  - Seamless data integration.

- **D3.js:** javascript library used to create every chart of the dashboard
  - Real-time data updates.
  - Seamless integration with React.
  - Complete control over styling, chart interactions and transformations.

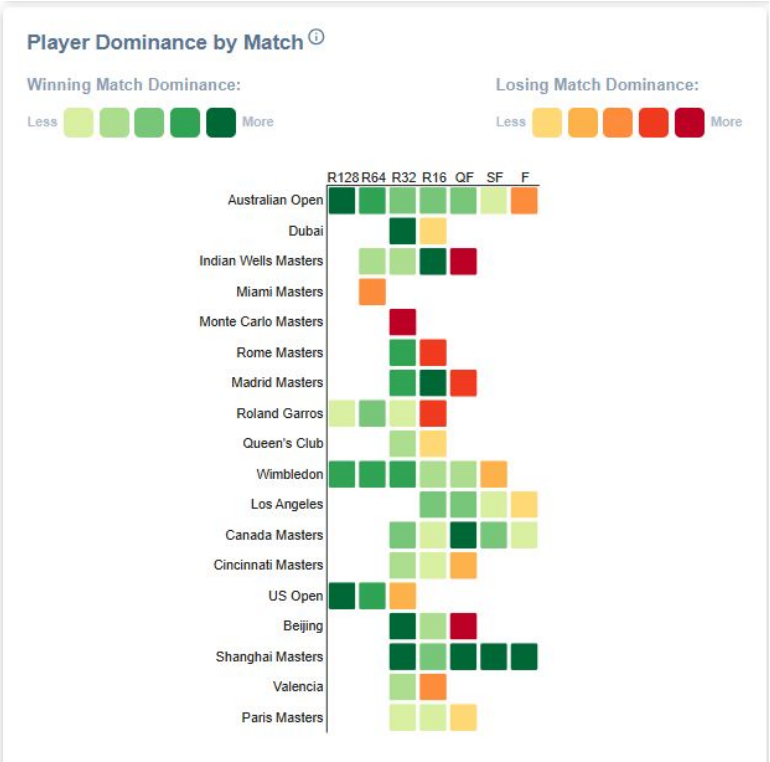
Visualizations



- In this chart, we represent the amount of matches played by the selected player on each season, by surface

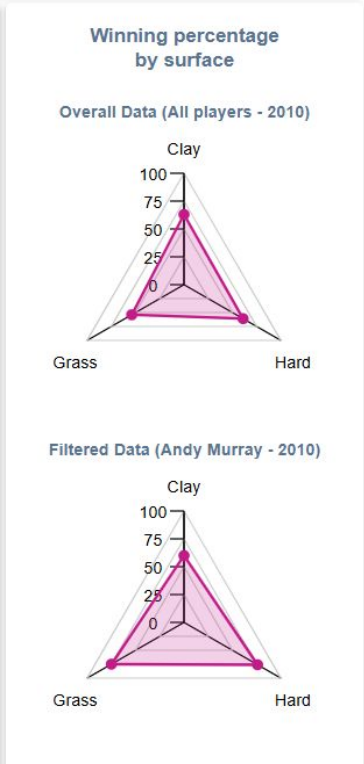
- This graph is useful to understand how full was the tournament calendar for the player and the difference in match volume between the 3 surfaces

Visualizations



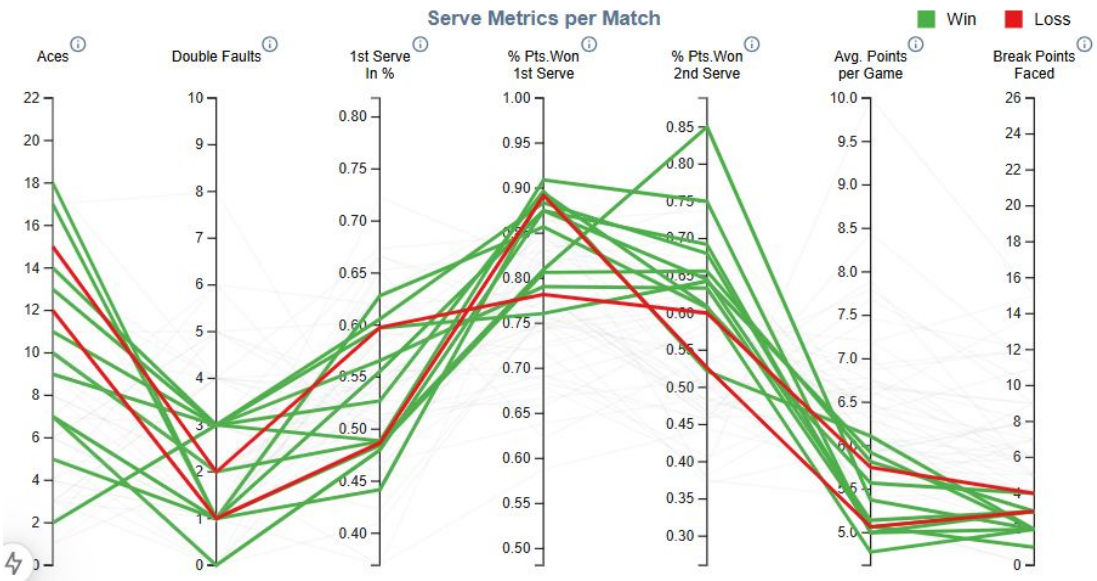
- In this chart, we represent the dominance of the player throughout the rounds of every tournament played in the season
- The dominance is estimated by the ratio between the games won by the player and the total number of games played in the match

Visualizations



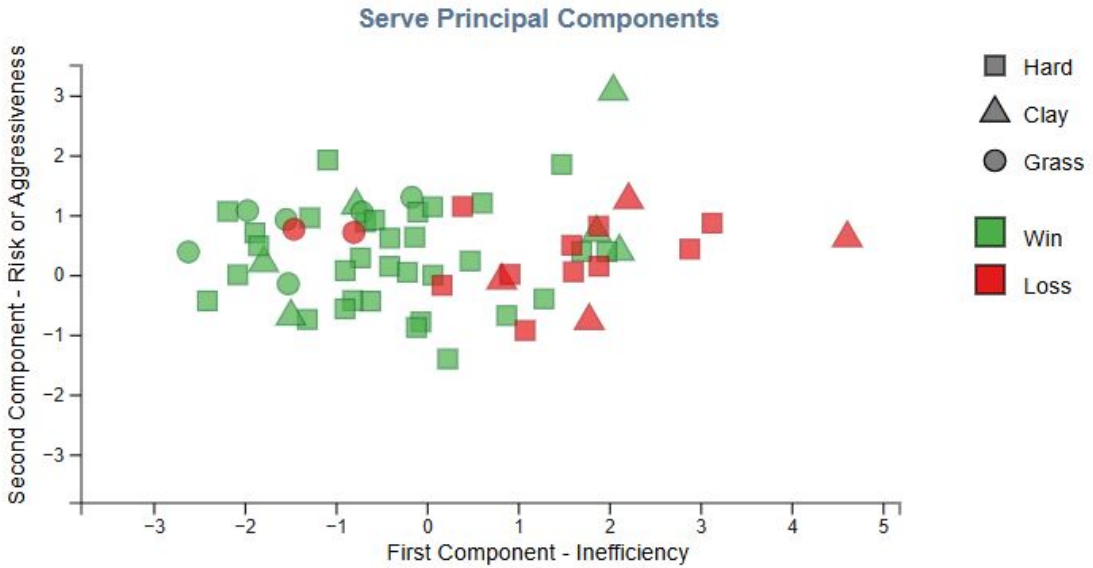
- In the first radar chart, we represent the average winning percentage of all players in a given season, for each surface.
- In the second radar chart, we also represent the winning percentage, but considering only the selected player in the given season.
- These charts can be used to identify the different player profiles: some playing styles are better suited for a given surface than others

Visualizations



- In this chart, we represent the most important serve attributes assessed in official matches
- The serve quality is often associated with the result of the match

Visualizations



- In this chart, we represent the 2 principal components extracted from the serve attributes using the PCA method: serve inefficiency and risk/aggressiveness
- Using this graph, we can assess how much different kinds of serve influence the outcome of the match
- We can also analyze how the serve varies among the surfaces