# Patterns of Play

**Predicting tennis match outcomes and player styles**

Dmitri Tarasov
Hunter Hobbs
Nivetha Kesavan
Pratik Revankar

# Questions we sought to answer:

Apply data mining techniques to Association of Tennis Professionals (ATP) data gathered over the past ~100 years and attempt to infer:

- clustered player styles
- hypothetical match win predictions
- player rivalries

# Data Preparation

- Considered eight datasets
- Analyzed sparsity of data object features using pandas dataframe aggregations and visualizations in jupyter notebook
- Leveraged domain knowledge to trim irrelevant features
- Merged useful datasets on player ID

# Tools used:

NumPy





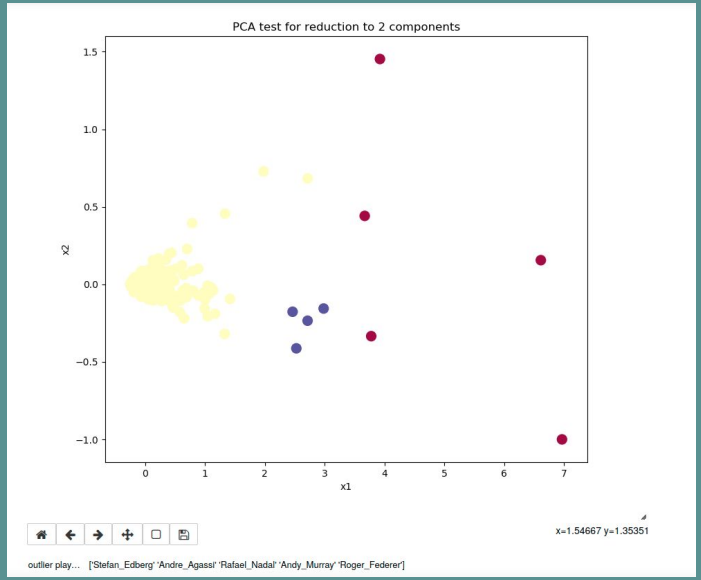
# Top Rivalries

1. Aggregated player match data to highlight top rivalries in the ATP tour for the years - 2000 to 2019.
   Djokovic-Nadal had the top rivalry with 52 total matches played and Djokovic leading 28/24
2. Other metrics for the rival pairs explored were - number of tournament finals reached, number of tournaments won over the years, Grand Slam (GS) wins, and court surface dominance

| Player1 | Player2 | Players | H2H | Total | Year | Decade |
|---|---|---|---|---|---|---|
| Djokovic N. | Nadal R. | Djokovic-Nadal | 28 / 24 | 52 | 2019 | 2010s |
| Djokovic N. | Federer R. | Djokovic-Federer | 26 / 21 | 47 | 2019 | 2010s |
| Federer R. | Nadal R. | Federer-Nadal | 17 / 24 | 41 | 2019 | 2010s |
| Djokovic N. | Murray A. | Djokovic-Murray | 25 / 10 | 35 | 2017 | 2010s |
| Ferrer D. | Nadal R. | Ferrer-Nadal | 6 / 26 | 32 | 2019 | 2010s |

# Player Styles

1. Used DBSCAN algorithm on match statistics and player skills.
2. The show clusters between average and outstanding players.
3. Highlights *contextual outliers* as players who seem to outperform the rest based on selected player style and skill set .




Novak Djokovic, Kiki Bertens, Juan Martin, and Lleyton Hewitt seem to be related

# Match Prediction

1. Data Preprocessing involved handling missing and noisy data, data integration, data cleaning and scaling.
2. KNN, SVM, AdaBoost, Decision Trees and XGBoost were trained individually to compare the performance on the holdout test set.

# Match Prediction (contd)

**Table 1: Model performance on ATP test set**

| Model | Accuracy | Recall | Precision |
|---|---|---|---|
| Decision Tree | 0.779690 | 0.765328 | 0.785620 |
| SVM | **0.795179** | 0.814896 | 0.784703 |
| AdaBoosting | 0.790982 | 0.789474 | 0.792641 |
| KNN | 0.778474 | 0.788930 | 0.772862 |
| XGB | 0.792934 | 0.794636 | 0.792008 |

# Knowledge Gained

- The Big 3 (Federer, Nadal, Djokovic) have been dominating the sport, with each holding individual records at major tournaments and Grand Slams, and having the most competitive rivalries on different court surfaces.
- Clustering on different combinations of player/match features required domain knowledge to correlate to a player style.
- Features like ace, double faults, breakpoints are more useful to determine the outcome of the match than features like surface, draw size.

# Knowledge Application

- Valuable insights for professional training to analyse and increase competitiveness.
- Insights for Pre-Match outcome prediction and online betting.
- Targeted marketing based on potential upcoming matches between top rivalries.

# References


1. Clarke, S. R. and Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. International Transactions in Operational Research, 7(6):585–594
2. Klaassen, F. J. and Magnus, J. R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. Journal of the American Statistical Association, 96(454):500–509
3. Y. Liu. Random walks in tennis Missouri Journal of Mathematical Sciences, 13 (3) (2001)
4. Newton, P. K. and Keller, J. B. (2005). Probability of winning at tennis i. theory and data. Studies in applied Mathematics, 114(3):241–269
5. William J. Knottenbelt, Demetris Spanias, Agnieszka M. Madurska : A common-opponent stochastic model for predicting the outcome of professional tennis matches, Computers & Mathematics with Applications, Volume 64, Issue 12, 2012, Pages 3820-3827, ISSN 0898-1221
6. McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. International Journal of Forecasting, 27(2):619–630.
7. Zhang, S (2019) Modelling ATP Tennis as a Network
8. Leeuw, Arie-Willem & Hoekstra, Aldo & Meerhoff, Rens & Knobbe, Arno. (2019). Tactical Analyses in Professional Tennis.