ATP Player Evaluation

Shruthi Harve · Week 4 · MSDS456 Assignment 1

OVERVIEW

Through this analysis, I deduce that a player's winning percentage is highly linked to **his ability to create breakpoint opportunities**, which is likely a result of his strong tennis play. Each player has different strengths – for instance John Isner wins based on his consistently competitive serve. When assessing the strength of a player's serve, **finishing a service game in fewer points is a stronger indicator of win probability than having a high percentage of inbound first serves**. It is also crucial to analyze data by surface type, as the game play drastically changes based on the court.

DATA

I used the ATP tour-level main draw data by match from 2009-2019 in order to focus my analysis on recent history. This dataset contains performance statistics for each player by match, as well as rankings and general demographic data.

PART 1 – PYTHAGOREAN WIN FORMULA METRIC

The Pythagorean Wins Formula is commonly used to evaluate a team's performance in other sports. However, it is not standardized within tennis. I analyzed the various performance metrics of a tennis player from 2009-2018 and the ability for it to predict a player's average win probability. **Appendix A** shows the correlation between these variables after being put through the Pythagorean formula[1], the accuracy of each metric compared to actual win percentage, as well as the RMSE. I deduced that the most insightful metric to use is either "Breakpoints Won" or "Breakpoint

[1] I assumed "2" as the exponent for the Pythagorean Wins Formula for all variables to prioritize identifying the key input variable for tennis.

Opportunities” due to the distribution accuracy as well as the lower RMSE values. I then replicated the same analysis on only 2019 data (**Appendix B**), which shows “Breakpoint Opportunities” with high accuracy and high precision, compared to “Breakpoints Won” with low precision. Since “Breakpoint Opportunities” was consistently strong, I suggest for future analysis using a player’s ability to create breakpoint opportunities for himself as the key indicator for win probability.

$$Win\ \% = \frac{(BP\ Opportunities\ Created)^2}{(BP\ Opportunities\ Created)^2 + (BP\ Opportunities\ Given)^2}$$

This conclusion differs from that of Stephanie Kovalchik in her 2015 Pythagorean analysis, where she deemed “Breakpoints Won” as the key variable. What her analysis fails to consider is the chance of matches that lead to straight tie breaks, or perhaps the opponent’s ability to serve an ace, denying the player a chance to convert that break opportunity. Therefore, when bringing it back to the context of the game, using “Breakpoint Opportunities” makes more sense. A competitive player creates his own opportunities. The mindset for both players changes on that high-pressure point. Whether he is able to convert his opportunity to a win partially depends on his opponent’s ability to save the breakpoint (i.e. strong service, fighting spirit, ball placement strategy).

Therefore, I recommend using “Breakpoint Opportunities” over “Breakpoints Won” for win calculations because: (1) it is more indicative of a player’s competitiveness, (2) there are more or the same number of break opportunities as break wins in a match, where players have multiple chances to break serve within a game, and (3) sometimes sets end in a tiebreak, thereby rendering breakpoint wins as unavailable for that set, whereas break opportunities may still be present.

PART 2 – PLAYER COMPARISON

As we learned from the previous section, it is crucial for players to create break point opportunities on the opponent’s service. This stems from what every tennis player learns during training: since the server has total control of how a point starts, he must master his serve. Serving out wide vs. down the center, with spin, at a certain speed, etc. are all factors that a player can learn to serve competitively. But no matter what, a player needs to serve inbounds to have any chance at winning the point. Therefore, I was curious to see how a consistently inbounds 1st serve stacks up against other metrics.

I filtered the dataset to players who have played at least 42 matches (in the 75th percentile) to smooth out averages and to account for any outliers. Doing so showed John Isner (USA) with the most consistently inbound 1st serves, followed by Dusan Lajovic (SRB) and Pablo Carreno Busta (SPN). The first plot in **Appendix C** shows these three players benchmarked against the top ranked player Novak Djokovic indicated in dark blue. This scatterplot shows us that: (1) there is little correlation between the two variables, (2) Lajovic and Carreno Busta rank far below Isner, despite their high inbound 1st serve percentages, and (3) although Djokovic has the 8th best inbound 1st serve percentage, the top 5 ranked players cluster around 64-67% inbounds, which is comparable to many other lower ranking players.

In order to better understand different play styles, I compared inbound 1st serve percentage against four other metrics:

1. <u>% 1st Serves Won</u> – were players able to win the serves they got in? This is indicative of their ability to set up points for winning shots.
2. <u>Ace % Service Points</u> – aces are automatic winners. A high percentage of aces can indicate that their serves are more competitive (i.e. speed, ball placement, spin).
3. <u>Breakpoint Opportunities Allowed</u> – how many breakpoint opportunities have they allowed on their serve? A low percentage can indicate a more confident & solid service play.
4. <u>Service Points per Service Game</u> – how quickly can a player close out his serve? A smaller number indicates unequal playing strength.

The strongest relationship between Win Percentage and these 4 variables is with Service Points per Game, at -74% correlation (**Appendix D**). At least 4 points are played per game; an average closer to 4 points indicates that either the player's service game is outstanding (commonly wins all 4 points on serve) or horrific (commonly loses all 4 points on serve). Given the professional level of play, I assume a lower average here indicates a player's ability to use his serve to set himself up for winners. What is interesting about the Speed of Service Games scatter plot in **Appendix E** is that Isner is performing similarly to top 3 ranked players, also in the bottom right cluster:

#	PLAYER	AVG RANK	AVG # SV PT
1	Djokovic	1.1	5.879
2	Nadal	1.9	5.883
3	Federer	3.6	5.917
4	Isner	12.5	5.970
5	Bourista Agut	17.8	5.983

This suggests that these players focus on strategically using their service to cater to their strengths – whether it be their volleying skills, ball placement, or the service itself.

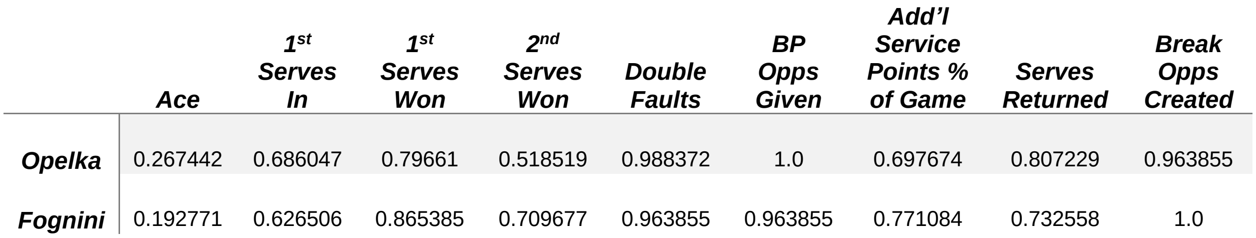
What is clear is that Isner is the strongest server in the ATP. Looking at the other 3 scatter plots in **Appendix E**, He has the highest Ace percentage, a competitively high 1st Serve Win percentage, and one of the lowest proportions of Breakpoint Chances Given to his opponent. Additionally, his stats for these metrics are significantly removed from the pack, indicating that he has an outstanding service play style.

PART 3 – PLAYER REPLACEMENT

The Davis Cup is a team-based tennis tournament grouped by country. In 2019, USA lost during the Round Robin stage (1st round), where they played 3 matches against both, Canada and Italy. Although they won 2-1 against Italy, they lost 1-2 against Canada. Considering Reilly Opelka lost both his matches, I wonder if John Isner could have outperformed Opelka's stats and perhaps led USA to the next round.

For this analysis, I looked specifically at the Opelka vs. Fognini matchup, where Opelka only won the second set in a tie break (Davis Cup). **Appendix F** shows a heatmap of the player's performance statistics from 2009-2019, where values were standardized so that 1.0 is "best" and 0.0 is "worst". What is interesting is that Opelka's statistics are either on par with or much better than Fognini's. For instance, Opelka wins a greater percentage of his service points, possibly partially contributed by his higher ace percentage. Additionally, Opelka's statistics look comparable to Isner's, with the greatest variance of 7% in 1st Service In %. That being said, Opelka's win percentage is significantly lower than both, Isner's and Fognini's.

Looking at Opelka's and Fognini's statistics from the 2019 Davis Cup match, Fognini served a significantly high proportion of aces compared to his average 4%. This could either be because his aces have improved and/or Opelka was struggling to return. The data is currently not available to confirm if this is due to Fognini's winners or Opelka's errors.



	Ace	*1st Serves In*	*1st Serves Won*	*2nd Serves Won*	*Double Faults*	*BP Opps Given*	*Add'l Service Points % of Game*	*Serves Returned*	*Break Opps Created*
Opelka	0.267442	0.686047	0.79661	0.518519	0.988372	1.0	0.697674	0.807229	0.963855
Fognini	0.192771	0.626506	0.865385	0.709677	0.963855	0.963855	0.771084	0.732558	1.0

PART 4 – IMPACT OF SURFACE TYPE

The game play changes significantly on different surface types. For instance, clay & grass make the ball bounce much slower than hard court, and the surface can be slippery, making it difficult for a player to get proper footing. I looked at the same average stats for Isner, Opelka and Fognini across the surfaces (**Appendix G**). What stands out right away is that Opelka has a 20% win percentage on clay, compared to his 55% on hard court.

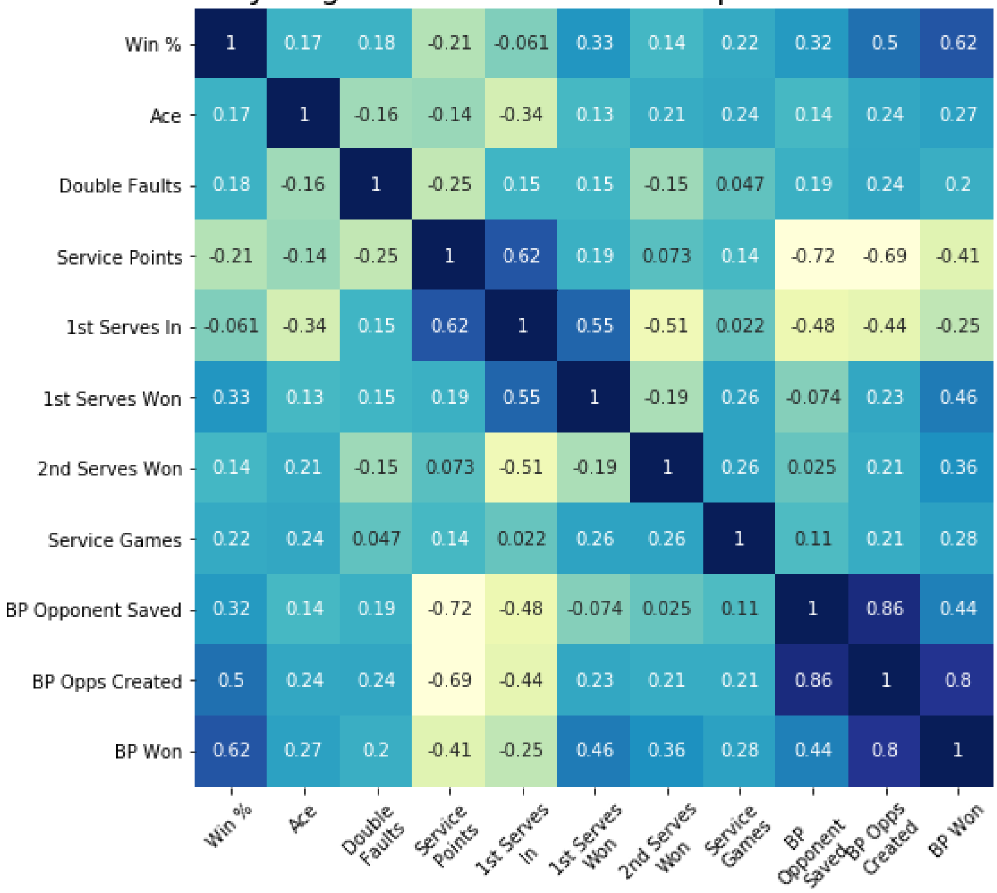
In this by-surface breakout of the stats, we start to see some differentiating performance stats between Isner and Opelka, namely on grass courts. If this match was played on grass instead of hard court, I would have recommended Team USA play Isner over Opelka; Opelka has a significantly lower ace percentage on his serve on grass and clay than on hard court, perhaps due to the slower bounce, making returns more achievable. Conversely, Isner's and Fognini's stats are relatively consistent across the surfaces. This leads me to suspect that Opelka lacks the experience to be competitive an all surfaces. Both, Isner and Fognini are over 10 years older than him, providing them with enough experience to improve their game play all round. Opelka, on the other hand, only seems to perform well on hard court, which he was trained on, but struggles on clay and grass.

Taking this a step further, I am curious about the mental maturity of a young player like Opelka. A solo sport like tennis requires high levels of mental stamina on top of physical training. Although there are sometimes passionate outbursts on court, the sport often calls for a level of poise, especially off-court. In a November 2019 interview, Opelka speaks candidly about his opinion against the structure of the new ATP Cup. While he makes good points about how the new tournament favors already higher-ranked players, he uses demeaning and prideful language, which is not on brand for tennis players. This could be a result of his immaturity, as well as his disappointment in his double-losses in the recently passed Davis Cup. This leads me to question if Opelka's low win percentage are partially attributable to an inability to focus his mind during high-pressure points. This analysis requires a further deep-dive into point-by-point play.

APPENDIX A

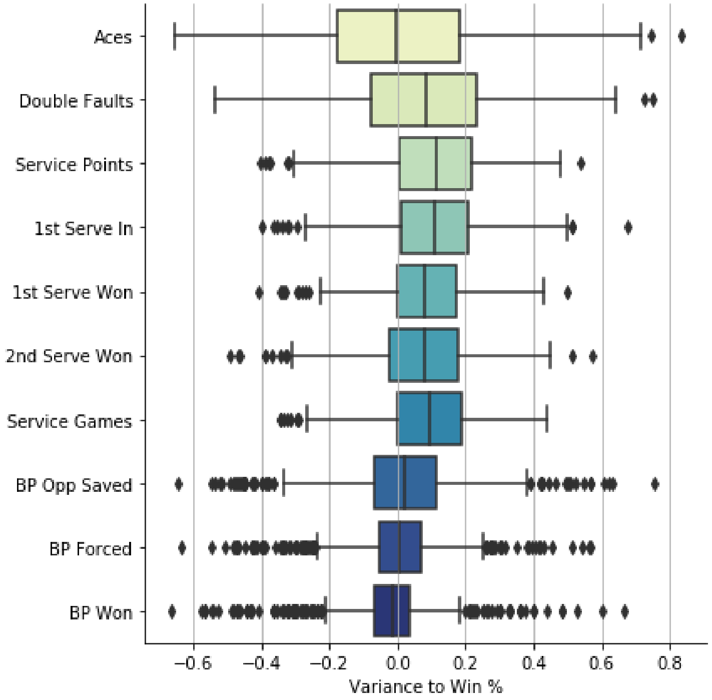
2009-2018 DATA

Pythagorean Win Formula Output Correlation



	Win %	Ace	Double Faults	Service Points	1st Serves In	1st Serves Won	2nd Serves Won	Service Games	BP Opponent Saved	BP Opps Created	BP Won
Win %	1	0.17	0.18	-0.21	-0.061	0.33	0.14	0.22	0.32	0.5	0.62
Ace	0.17	1	-0.16	-0.14	-0.34	0.13	0.21	0.24	0.14	0.24	0.27
Double Faults	0.18	-0.16	1	-0.25	0.15	0.15	-0.15	0.047	0.19	0.24	0.2
Service Points	-0.21	-0.14	-0.25	1	0.62	0.19	0.073	0.14	-0.72	-0.69	-0.41
1st Serves In	-0.061	-0.34	0.15	0.62	1	0.55	-0.51	0.022	-0.48	-0.44	-0.25
1st Serves Won	0.33	0.13	0.15	0.19	0.55	1	-0.19	0.26	-0.074	0.23	0.46
2nd Serves Won	0.14	0.21	-0.15	0.073	-0.51	-0.19	1	0.26	0.025	0.21	0.36
Service Games	0.22	0.24	0.047	0.14	0.022	0.26	0.26	1	0.11	0.21	0.28
BP Opponent Saved	0.32	0.14	0.19	-0.72	-0.48	-0.074	0.025	0.11	1	0.86	0.44
BP Opps Created	0.5	0.24	0.24	-0.69	-0.44	0.23	0.21	0.21	0.86	1	0.8
BP Won	0.62	0.27	0.2	-0.41	-0.25	0.46	0.36	0.28	0.44	0.8	1

Pythagorean Calc Accuracy with 2 Exponent

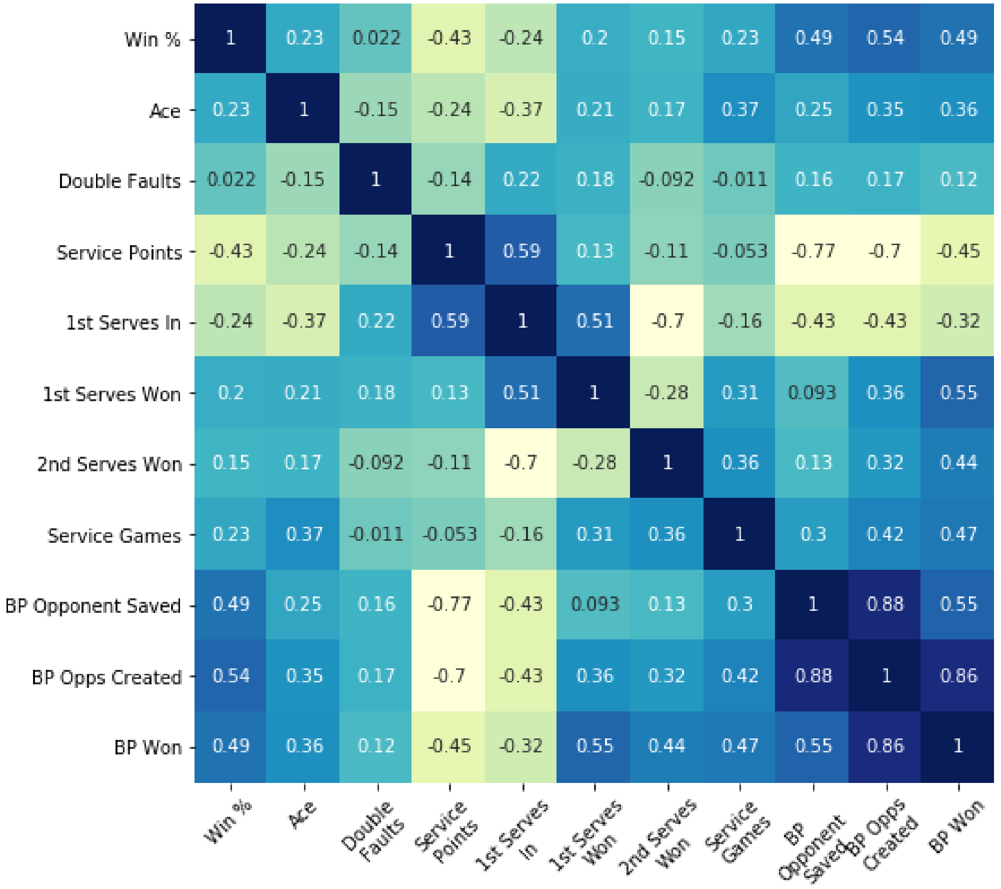


VARIABLE	RMSE
BP WON	0.135148
BP OPPORTUNITIES	0.135609
1ST SERVE WON	0.139073
SERVICE GAME	0.149366
2ND SERVE WON	0.156177
BP OPPONENT SAVED	0.167718
1ST SERVE IN	0.168254
SERVICE POINT	0.168706
DOUBLE FAULT	0.210352
ACE	0.227905

APPENDIX B

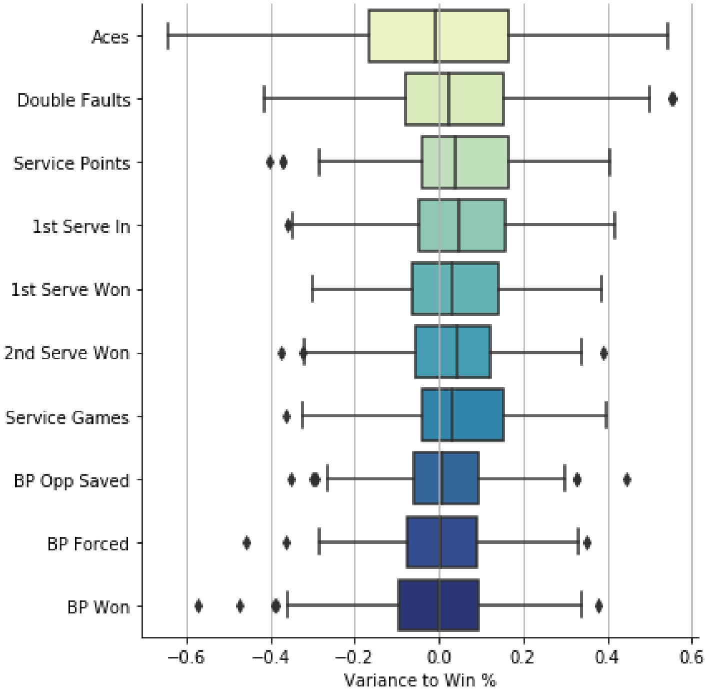
2019 DATA

2019 Pythagorean Win Formula Output Correlation



	Win %	Ace	Double Faults	Service Points	1st Serves In	1st Serves Won	2nd Serves Won	Service Games	BP Opponent Saved	BP Opps Created	BP Won
Win %	1	0.23	0.022	-0.43	-0.24	0.2	0.15	0.23	0.49	0.54	0.49
Ace	0.23	1	-0.15	-0.24	-0.37	0.21	0.17	0.37	0.25	0.35	0.36
Double Faults	0.022	-0.15	1	-0.14	0.22	0.18	-0.092	-0.011	0.16	0.17	0.12
Service Points	-0.43	-0.24	-0.14	1	0.59	0.13	-0.11	-0.053	-0.77	-0.7	-0.45
1st Serves In	-0.24	-0.37	0.22	0.59	1	0.51	-0.7	-0.16	-0.43	-0.43	-0.32
1st Serves Won	0.2	0.21	0.18	0.13	0.51	1	-0.28	0.31	0.093	0.36	0.55
2nd Serves Won	0.15	0.17	-0.092	-0.11	-0.7	-0.28	1	0.36	0.13	0.32	0.44
Service Games	0.23	0.37	-0.011	-0.053	-0.16	0.31	0.36	1	0.3	0.42	0.47
BP Opponent Saved	0.49	0.25	0.16	-0.77	-0.43	0.093	0.13	0.3	1	0.88	0.55
BP Opps Created	0.54	0.35	0.17	-0.7	-0.43	0.36	0.32	0.42	0.88	1	0.86
BP Won	0.49	0.36	0.12	-0.45	-0.32	0.55	0.44	0.47	0.55	0.86	1

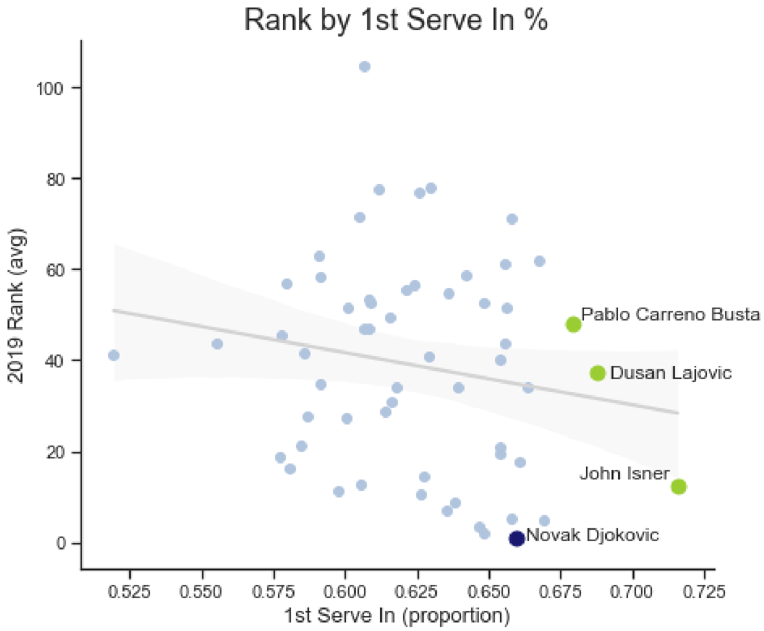
2019 Pythagorean Calc Accuracy with 2 Exponent

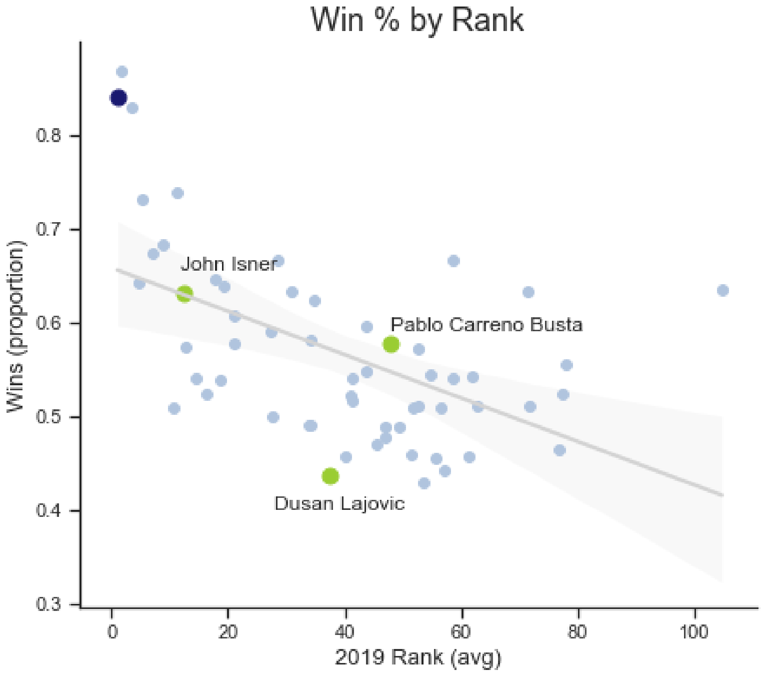


VARIABLE	RMSE
BP OPPORTUNITIES	0.114401
BP OPPONENT SAVED	0.118605
1^{ST} SERVE WON	0.126751
SERVICE GAME	0.129255
2^{ND} SERVE WON	0.131420
BP WON	0.136100
SERVICE POINT	0.141080
1^{ST} SERVE IN	0.143155
DOUBLE FAULT	0.185764
ACE	0.20619

APPENDIX C

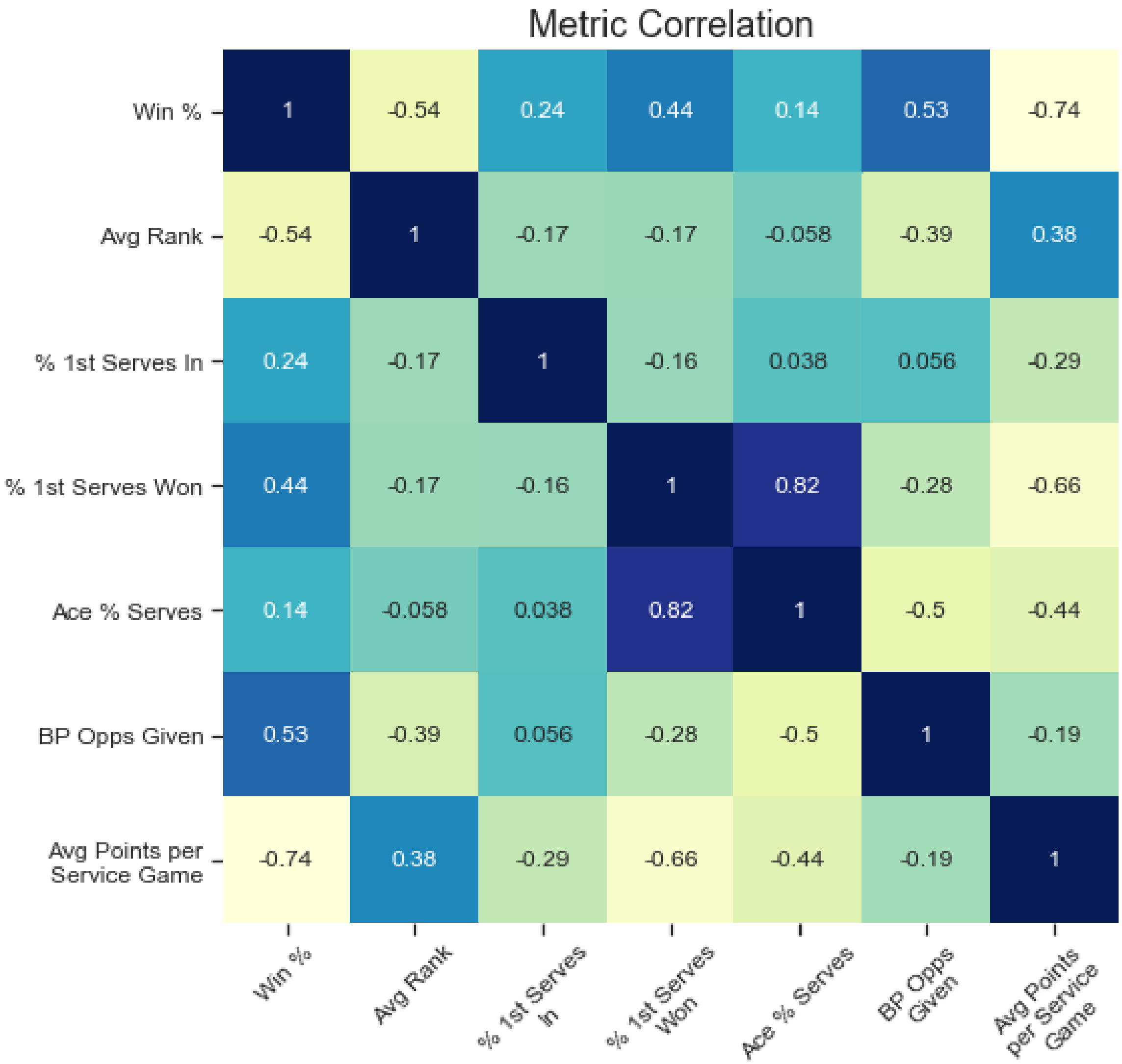
2019 DATA





APPENDIX D

2019 DATA

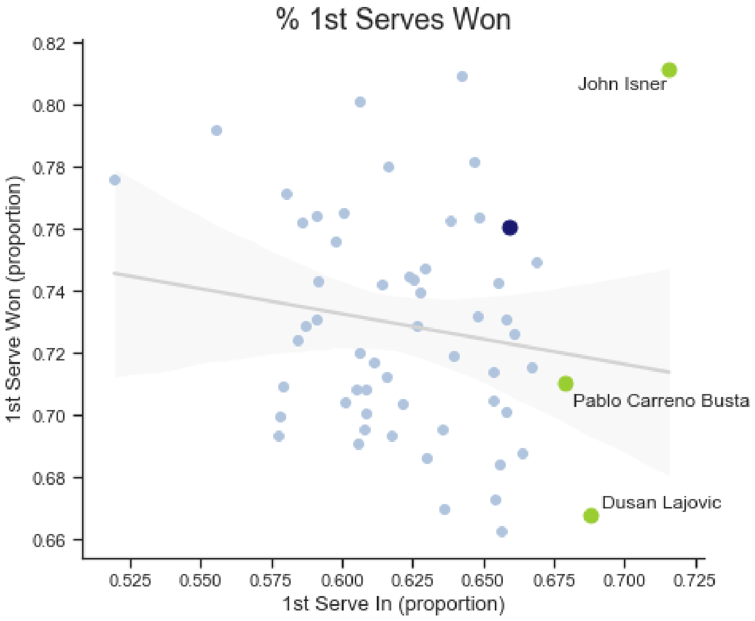


Metric Correlation

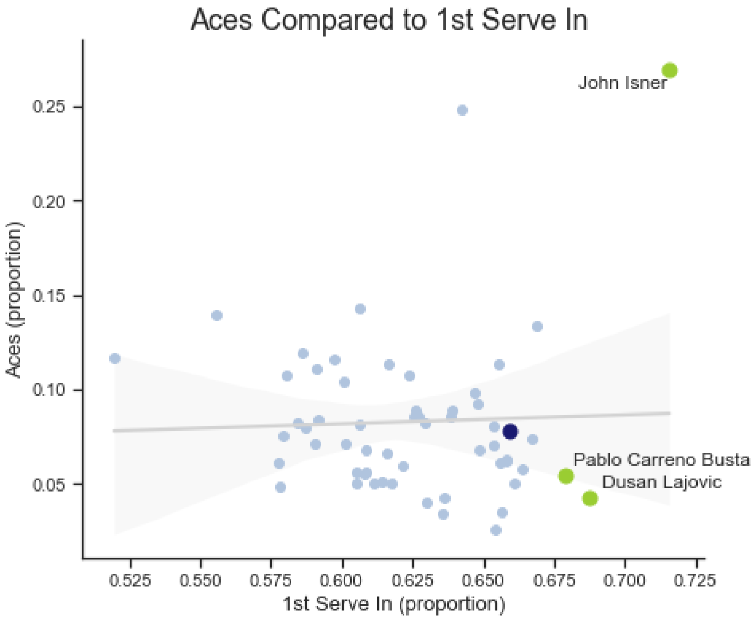
	Win %	Avg Rank	% 1st Serves In	% 1st Serves Won	Ace % Serves	BP Opps Given	Avg Points per Service Game
Win %	1	-0.54	0.24	0.44	0.14	0.53	-0.74
Avg Rank	-0.54	1	-0.17	-0.17	-0.058	-0.39	0.38
% 1st Serves In	0.24	-0.17	1	-0.16	0.038	0.056	-0.29
% 1st Serves Won	0.44	-0.17	-0.16	1	0.82	-0.28	-0.66
Ace % Serves	0.14	-0.058	0.038	0.82	1	-0.5	-0.44
BP Opps Given	0.53	-0.39	0.056	-0.28	-0.5	1	-0.19
Avg Points per Service Game	-0.74	0.38	-0.29	-0.66	-0.44	-0.19	1

APPENDIX E

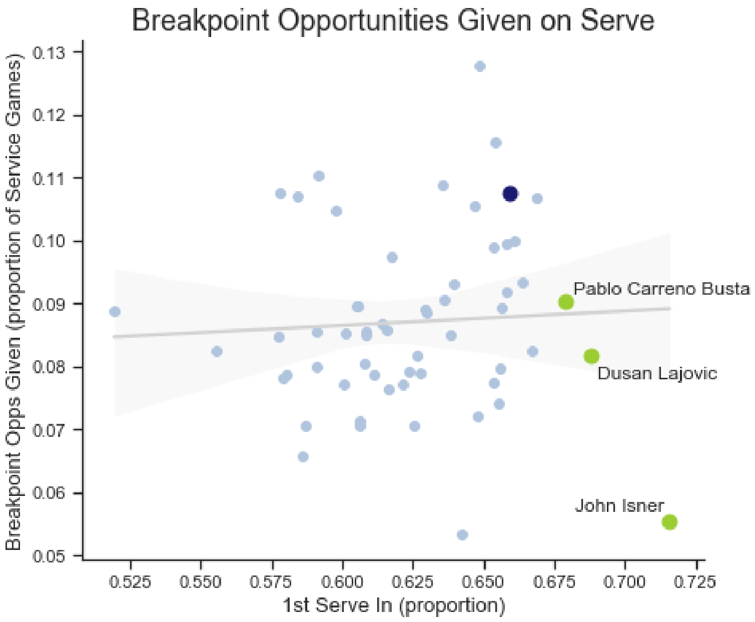
2019 DATA



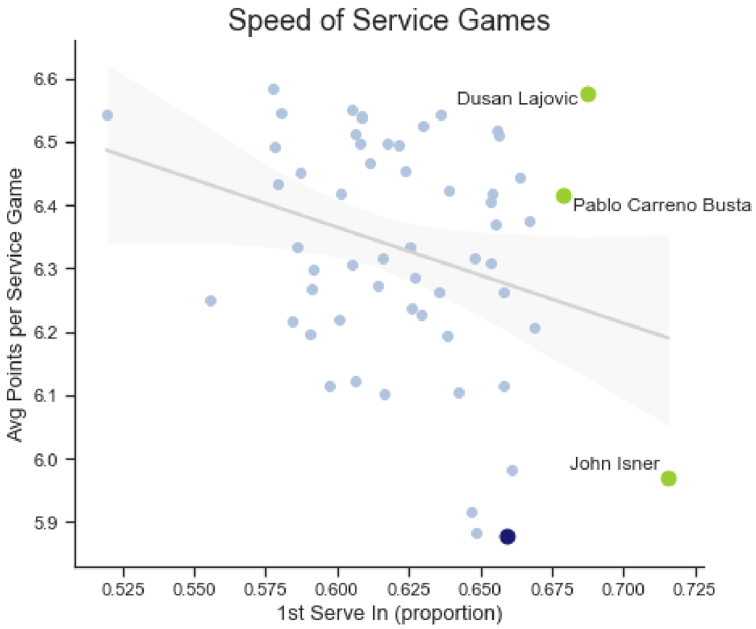
% 1st Serves Won



Aces Compared to 1st Serve In



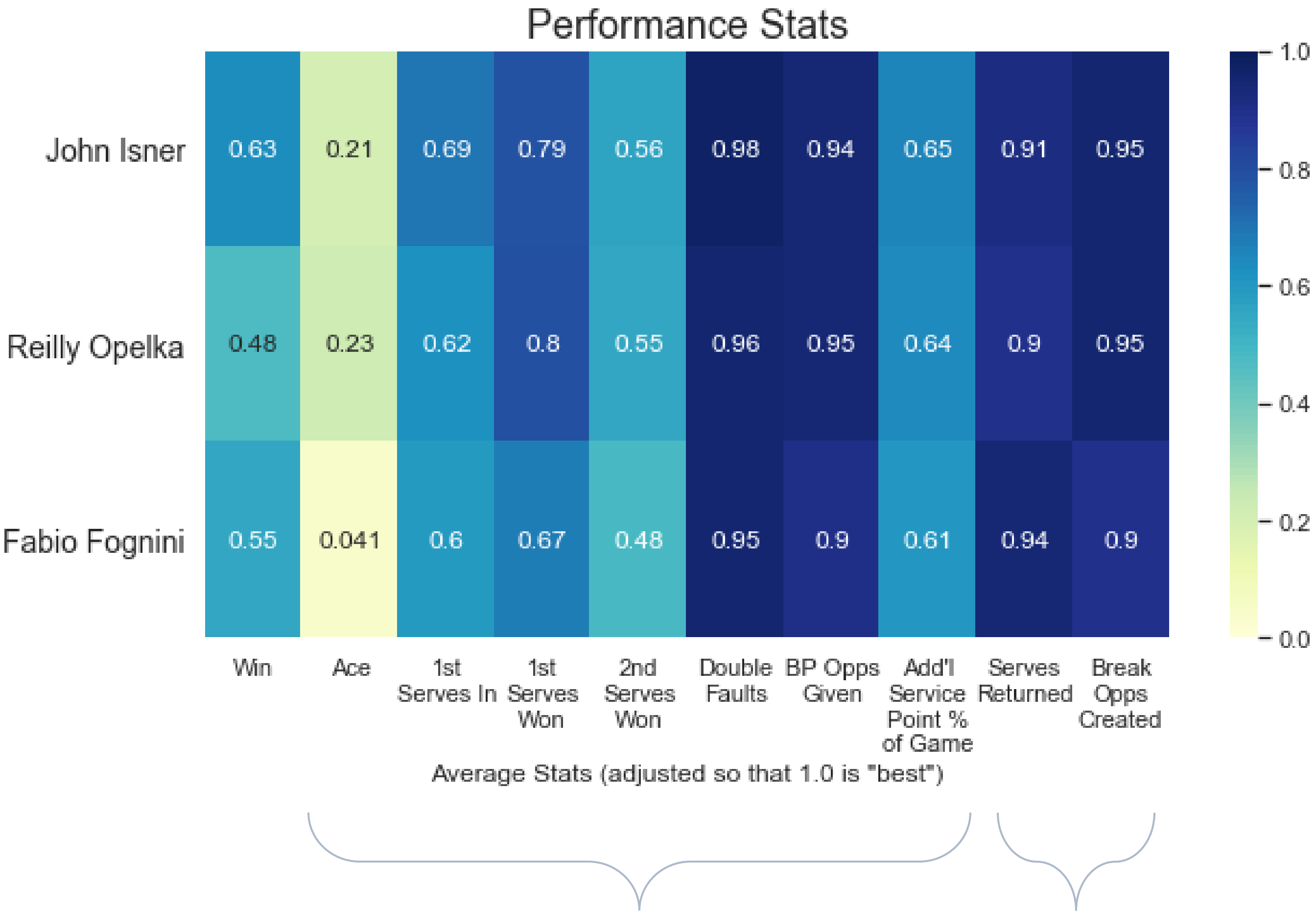
Breakpoint Opportunities Given on Serve



Speed of Service Games

APPENDIX F

2009-2019 DATA

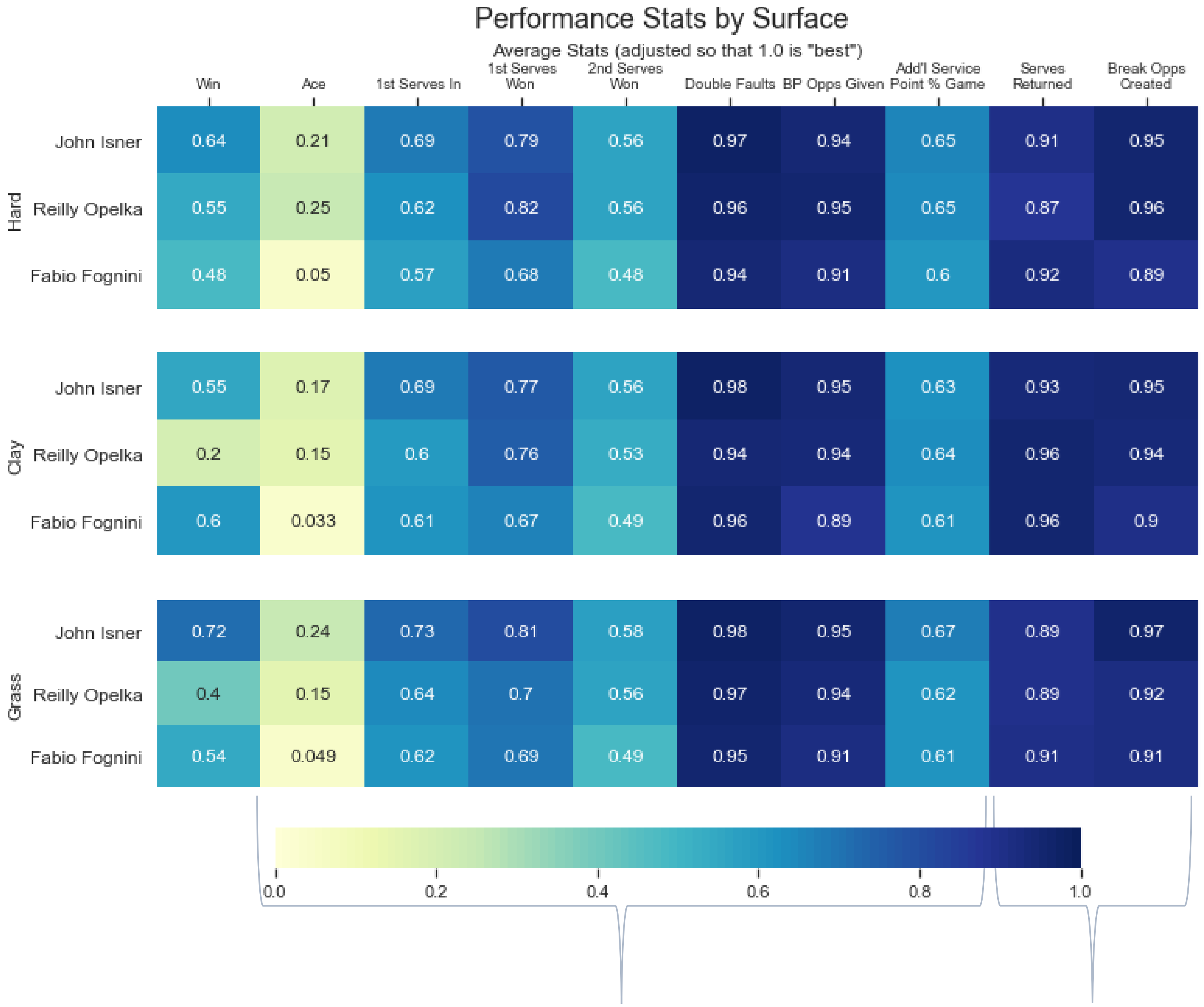


ON SERVE

ON RETURN

APPENDIX G

2009-2019 DATA



Performance Stats by Surface

Average Stats (adjusted so that 1.0 is "best")

		Win	Ace	1st Serves In	1st Serves Won	2nd Serves Won	Double Faults	BP Opps Given	Add'l Service Point % Game	Serves Returned	Break Opps Created
Hard	John Isner	0.64	0.21	0.69	0.79	0.56	0.97	0.94	0.65	0.91	0.95
	Reilly Opelka	0.55	0.25	0.62	0.82	0.56	0.96	0.95	0.65	0.87	0.96
	Fabio Fognini	0.48	0.05	0.57	0.68	0.48	0.94	0.91	0.6	0.92	0.89
Clay	John Isner	0.55	0.17	0.69	0.77	0.56	0.98	0.95	0.63	0.93	0.95
	Reilly Opelka	0.2	0.15	0.6	0.76	0.53	0.94	0.94	0.64	0.96	0.94
	Fabio Fognini	0.6	0.033	0.61	0.67	0.49	0.96	0.89	0.61	0.96	0.9
Grass	John Isner	0.72	0.24	0.73	0.81	0.58	0.98	0.95	0.67	0.89	0.97
	Reilly Opelka	0.4	0.15	0.64	0.7	0.56	0.97	0.94	0.62	0.89	0.92
	Fabio Fognini	0.54	0.049	0.62	0.69	0.49	0.95	0.91	0.61	0.91	0.91

ON SERVE

ON RETURN

REFERENCES

Herman, M. (2019, November 21). American Opelka calls ATP Cup 'pathetic'. Retrieved August 9, 2020, from https://www.reuters.com/article/us-tennis-daviscup-opelka/american-opelka-calls-atp-cup-pathetic-idUSKBN1XV2CS

ITF Licensing (UK) Ltd. (n.d.). Davis Cup - Finals 2019. Retrieved August 9, 2020, from https://www.daviscup.com/en/draws-results/tie.aspx?id=M-DC-2019-FLS-F-M-USA-ITA-01

Kovalchik, S. (2015, September 26). Converting Clutch into Wins - A Pythagorean Model for Tennis. Retrieved August 9, 2020, from http://on-the-t.com/2015/09/26/converting-clutch-into-wins/

Sackmann, J. (n.d.). JeffSackmann Github. Retrieved August 9, 2020, from https://github.com/JeffSackmann