Universiteit van Amsterdam

Scientific Data Analysis

What drives match outcomes in professional tennis?

Efe Aras, Stijn Jongbloed, Sebas van Waard
and Liam Gatersleben

19 december 2025

Contents

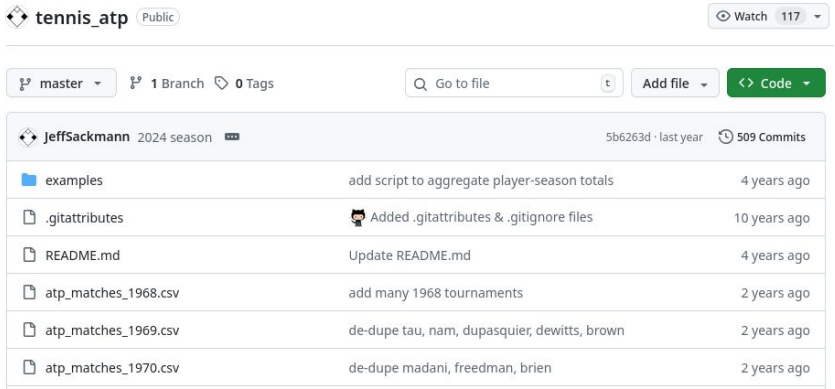
1. Motivation

- Knowledge for its own sake, it seemed interesting.
- Interesting from a sports science perspective, what should a player focus on in training.
- Gambling, both from the perspective of the bookmaker as well as gamblers.

2. Data description

Dataset

- ATP Tennis dataset
- Size & Scope: ~80k matches from 1991-2024
- Features: player attributes, rankings, surface, and recent performance
- Match-level data (player 1 vs player 2)
- Includes matches from both player perspectives to support logistic regression modeling
- Focus on the main tier of male singles.



	tourney_name	surface	winner_id	winner_hand	winner_ht	winner_ioc	winner_age	best_of	minutes	winner_rank	winner_rank_points	loser_rank
1963	Washington	Hard	111805	R	175	KOR	27	3	123	175	343	89
1964	Washington	Hard	133430	L	185	CAN	25.2	3	88	139	450	76
1965	Washington	Hard	200670	R	183	USA	25.6	3	99	143	436	109
1966	Washington	Hard	106331	R	183	AUS	30.1	3	97	94	653	114
1967	Washington	Hard	105430	R	175	MDA	34.7	3	133	152	411	62
1968	Paris Olympics	Clay	104925	R	188	SRB	37.1	3		2	8460	
1969	Paris Olympics	Clay	104745	L	185	ESP	38.1	3		161	380	86
1970	Paris Olympics	Clay	136440	L	180	GER	30.2	3		70	776	177
1971	Paris Olympics	Clay	208286	R	185	ITA	23.4	3		45	1155	23
1972	Paris Olympics	Clay	202104	R	170	ARG	23.5	3		18	2250	77
1973	Paris Olympics	Clay	133975	R	183	LBN	29.4	3		170	353	110
1974	Paris Olympics	Clay	105554	R	175	GBR	34.1	3		58	844	382
1975	Paris Olympics	Clay	126774	R	193	GRE	25.9	3		11	3705	88
1976	Paris Olympics	Clay	100644	R	198	GER	27.2	3		4	6845	72

3. Hypotheses Questions

What factors drive match outcome in professional tennis?

Based on these factors, can we can we accurately predict match outcome?

Player Characteristics

- Height
- Age
- Dominant Hand

Performance Metrics

- Ranking
- Surface-specific Win Rate
- Current Win Streak

Match Context

- Match Surface

Playing Style

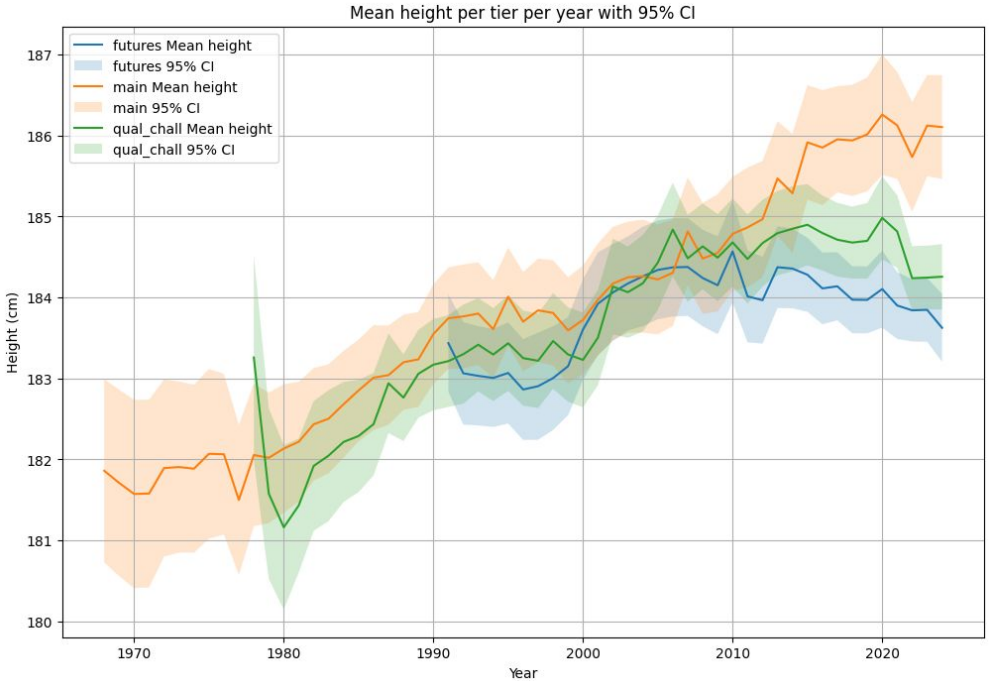
- Fast
- Balanced
- Endurance

4. Analysis: Height

-
- Cleaning to be done, impossible lengths, unsure about implications for the rest of the data.
- Testing idea, if there is a advantage for certain heights there will be a difference in heights between the 3 tiers.

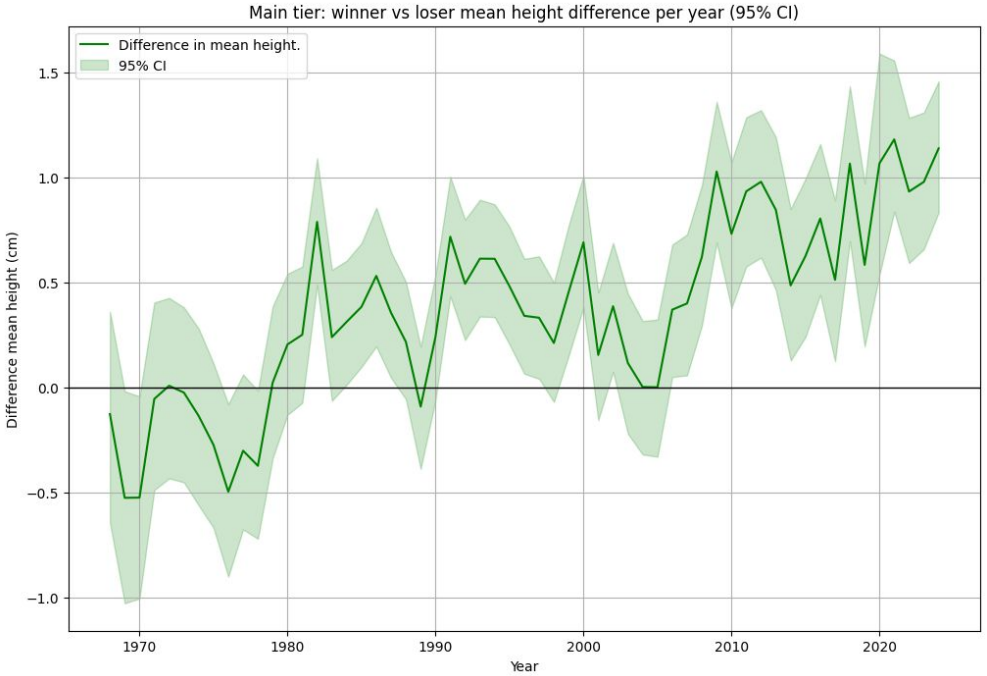
4. Analysis: Height

- **H0:** Player height is not associated with probability of winning
- Mean of unique player lengths per year per tier.
- 95% CI to see if results are significant.
- Student's t-distribution, height normally distributed, individual heights within group independent.
- Historically no significant difference between tiers, recently the main tier had statistically significant taller players.
- No statistic tests performed, but clearly taller than world average.
- So what about within the main tier?

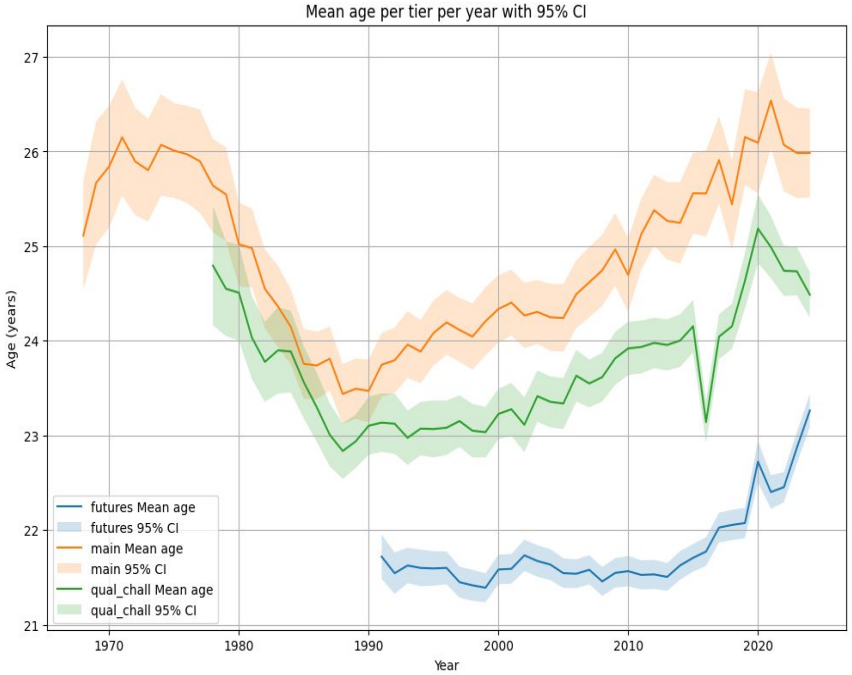


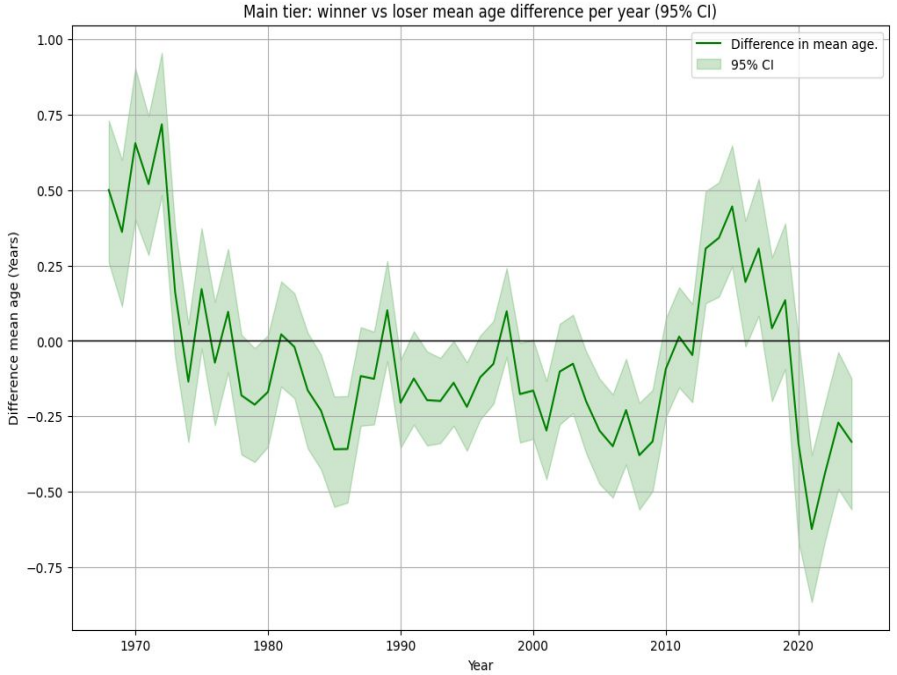
4. Analysis: Height

- Mean difference of height between winners and losers computed.
- 95% CI once again.
- Heights across matches not independent, so bootstrapping had to be used.
- Especially recently, winners within the main tier have been taller on average to a significant degree.
- Overall conclusion, being taller does give a player an advantage.
- Null hypothesis rejected.



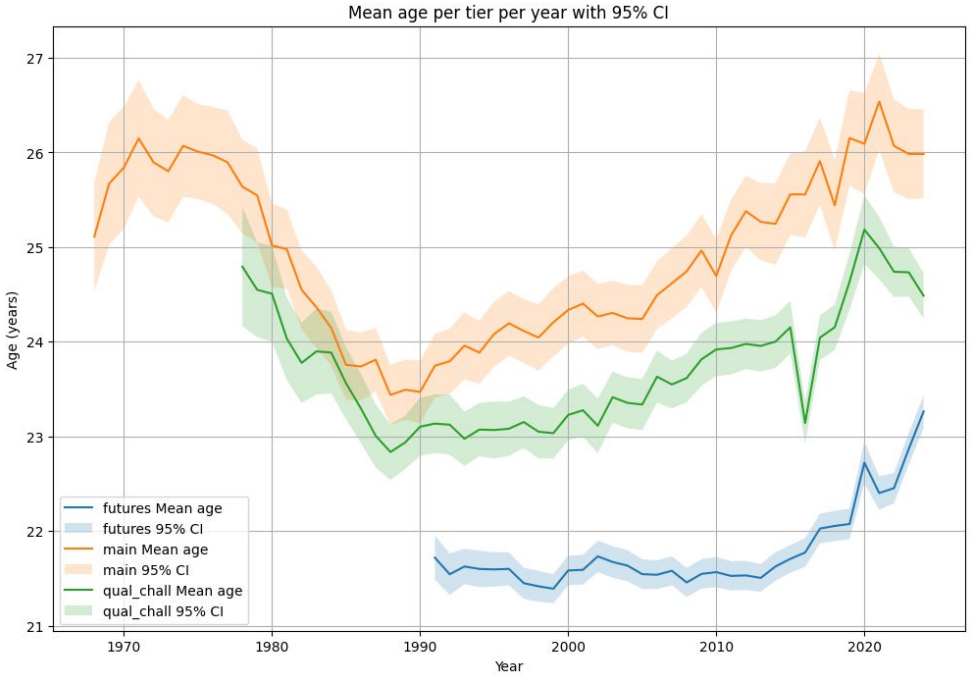
4. Analysis: Age





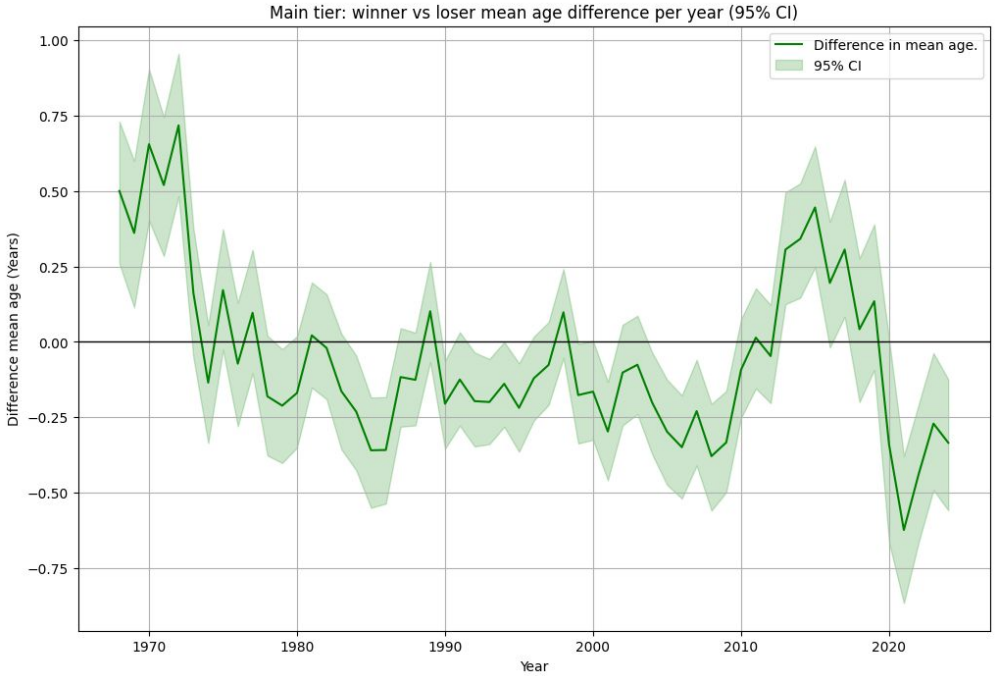
4. Analysis: Age

- Same as height, mean of unique player ages per year.
- Sorted ages, so youngest player age of the year.
- CI again through Student's t-distribution. Age less normal than height, but with the smallest group being 240 players it still works.
- Mean age clearly increases per tier
- The mean age of the main tier hovers around 26 in recent years, suggesting that that is the prime age for top tennis players.



4. Analysis: Age

- Like with heights:
  average_winner_age - average_loser_age
- Also like heights, CI computed using bootstrapping because the data is no longer independent.
- Unlike heights, less clear trents. While statistically significant within some years, those years contradict other years.

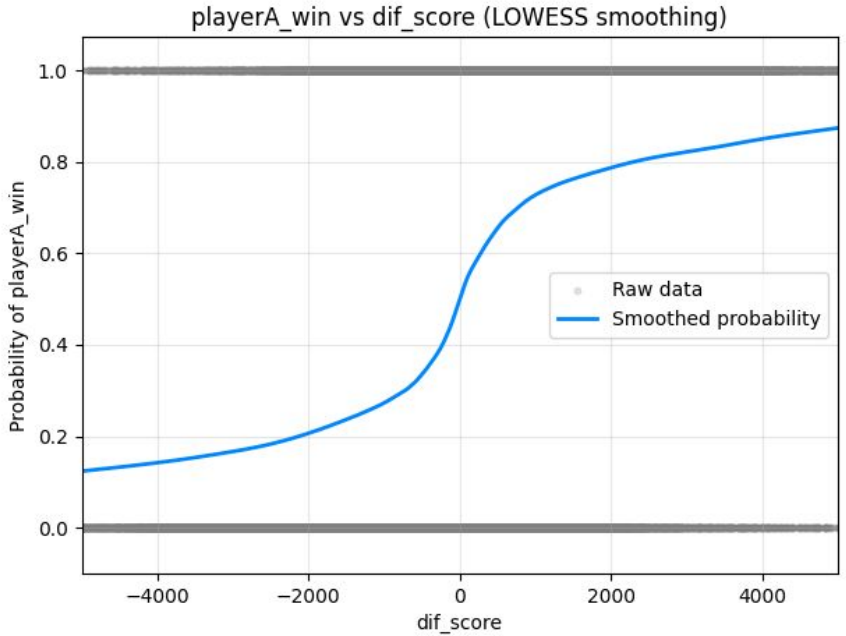


4. Analysis: Ranking

Does the difference in the players rankings have a significant effect on their win rate?

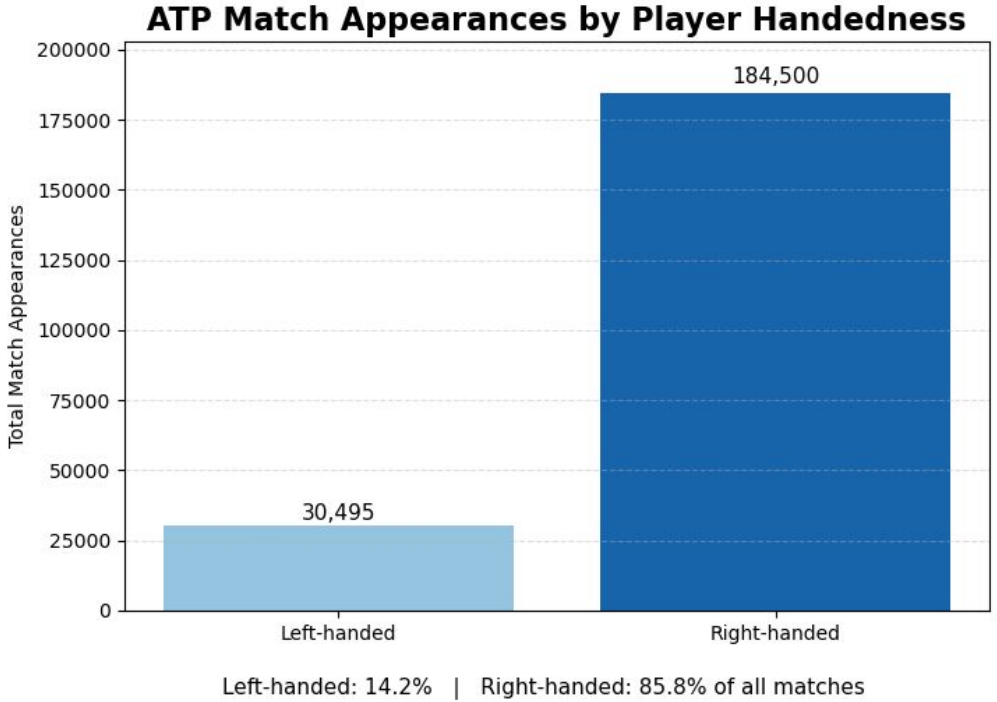
H0: There is no effect of absolute difference in ranking of the players on match outcome

- Every player has an ATP ranking based on their performance
- Take the absolute difference
- Statistically significant:
  - $p < 0.001$



4. Analysis: Dominant Hand

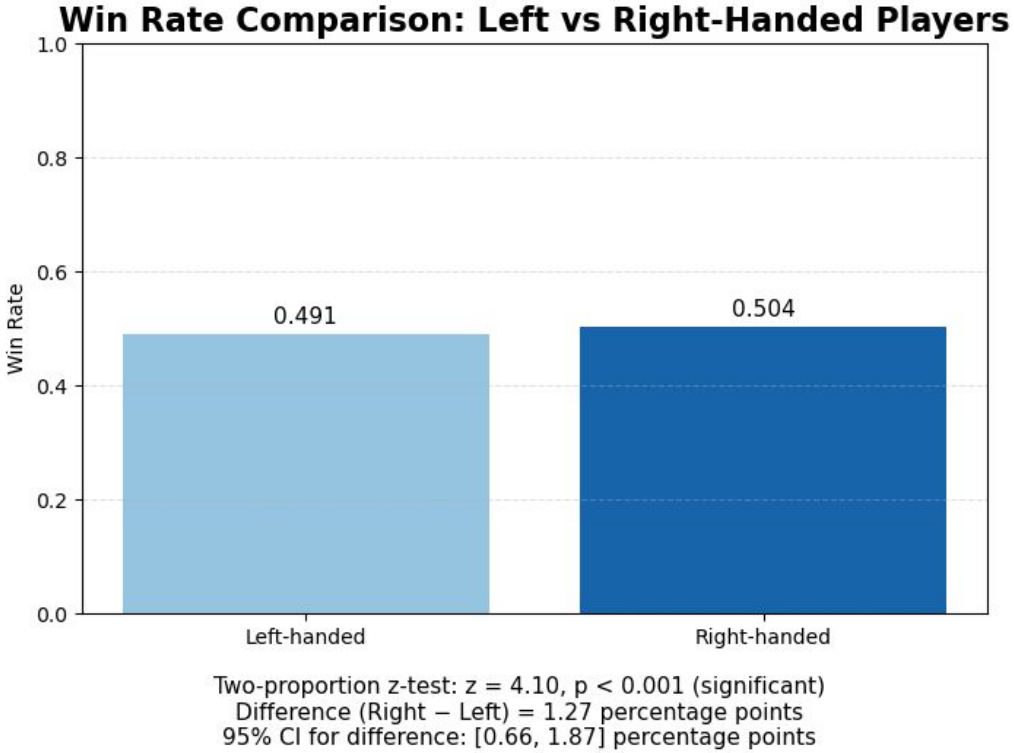
H0: A player's dominant hand has no significant effect on their win rate.



4. Analysis: Dominant Hand

Analysis:

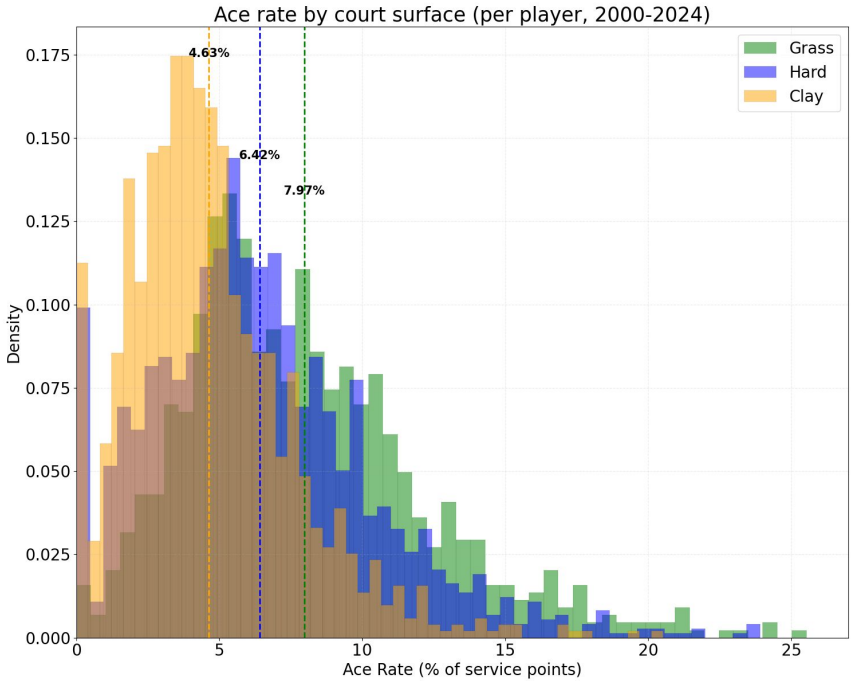
- Very large sample size (< 200k)
- Two-proportion z-test
- 1.27 percentage points difference
- 95% CI difference [0.66, 1.87]
- H0 rejected



4. Analysis: surface ace rates

H0: The type of surface has no significant effect on the ace rates

- Ace rates by surface
  - Aggregating per player
- Plot indeed shows significant differences
  - Clay - slow surface
  - Hard - medium fast surface
  - Grass - fast surface
- Statistical test ANOVA shows
  - p-value: 2.10717e-90
  - H0 rejected



4. Analysis: surface win rate

H0: Surface-specific win rate has no significant effect on match outcome

- Calculated surface win rate on past matches before current
- Early matches and new players can be noisy, with many 0% or 100% win rates
  - Applied laplace smoothing to reduce this noise
  - Use matches from years prior to the training data to initialize win rates, so players don't all start as "new"
- Logistic regression: p-value < 0.0001

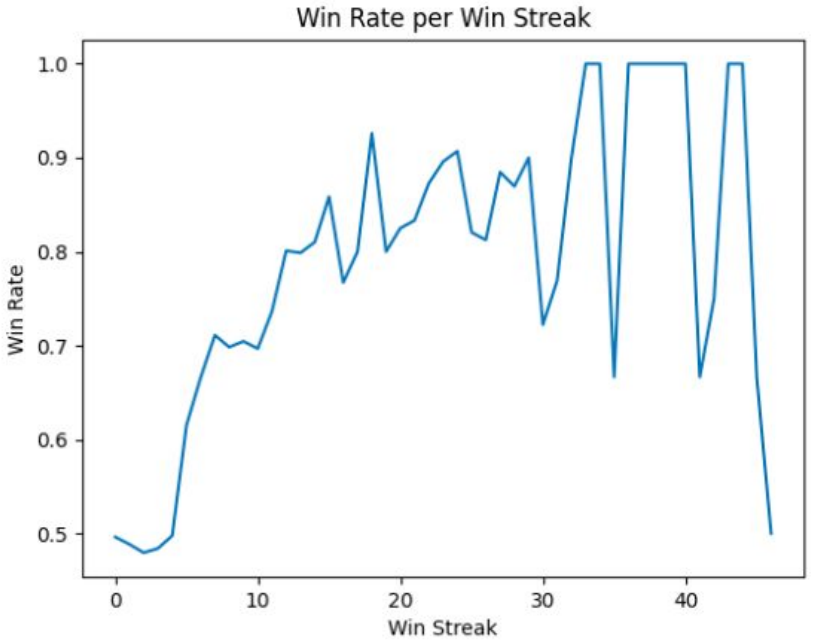
Surface	LogReg model	Accuracy	Odds per 10% wr
Hard	Rank diff	59.9%	
Hard	Rank/wr diff	60.8%	1.26x
Clay	Rank diff	58.1%	
Clay	Rank/wr diff	59.1%	1.20x
Grass	Rank diff	57.7%	
Grass	Rank/wr diff	58.3%	1.19x

4. Analysis: Win Streak

Does the players win streak of previous matches have significant effect on the win rate of their next game?

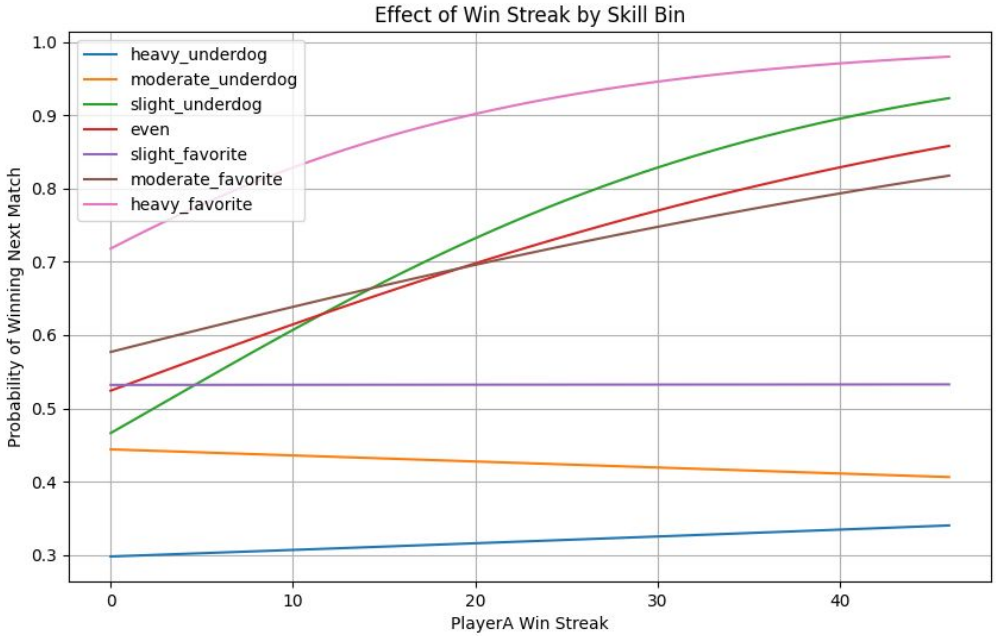
H0: There is no significant effect of winstreak on the match outcome

- Logistic Regression
- Initial significant ($p < 0.001$), but small effect on model accuracy
- However there does seem to be effect when looking at win-rate



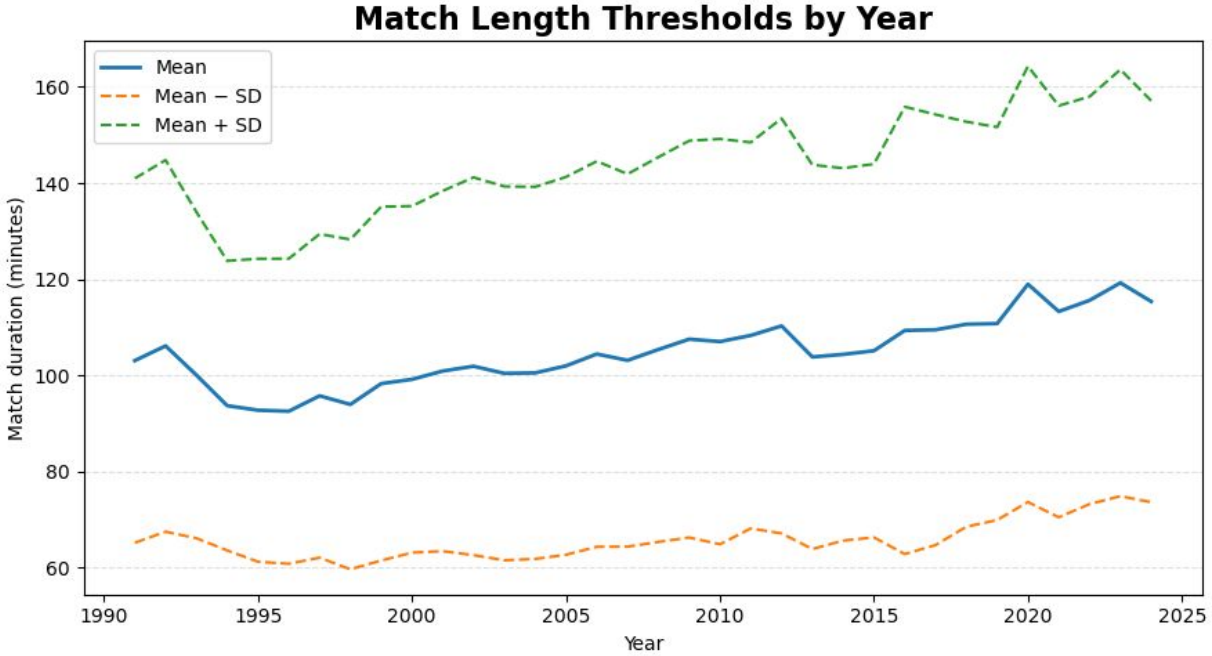
4. Analysis: Win Streak

- The players that are high ranking are the players that can achieve high win-streaks
- Maybe there is a difference in effect of win-streak when different ranked players are matched up
- Bin matches based on relative rank



4. Analysis: Player Archetype

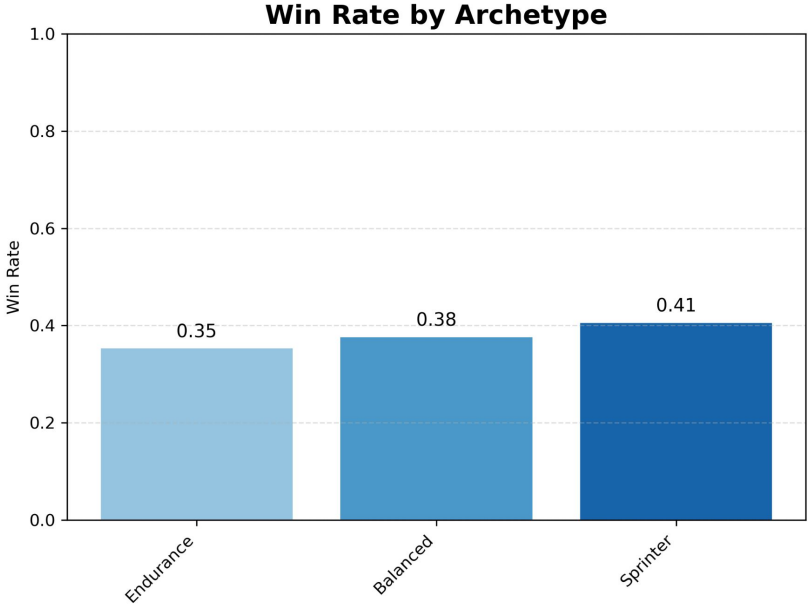
Dividing players into different playstyles.



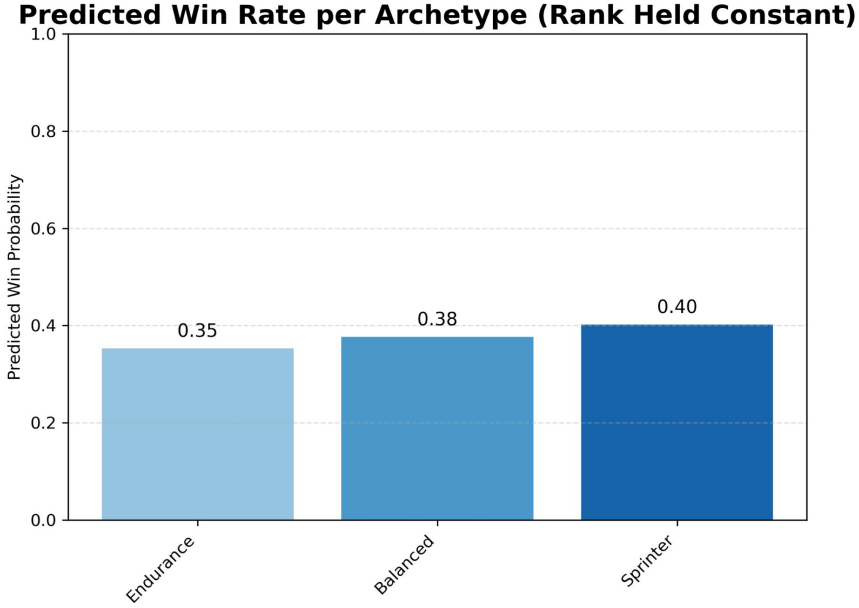
4. Analysis: Player Archetype

player_id	player_archetype	opponent_id	opponent_archetype	won
100284	Endurance	100923	Endurance	1
100284	Endurance	101086	Endurance	1
100284	Endurance	101381	Endurance	1
100284	Endurance	101774	Balanced	1

4. Analysis: Player Archetype



Chi-square test: $p < 0.001$ (significant)

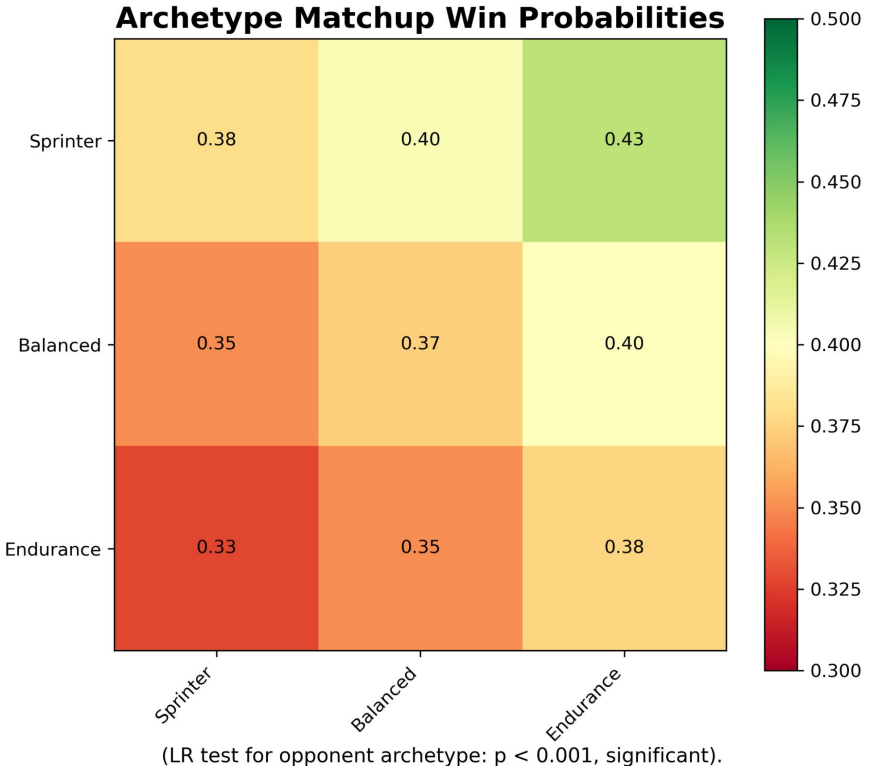


Logit (rank-controlled), archetype effect (LR test): $p < 0.001$ (significant)

4. Analysis: Player Archetype

H0: The archetype (endurance, balanced, sprinter) of a player has significant impact on their win rate.

Accepted



(LR test for opponent archetype: $p < 0.001$, significant).

5. Prediction Model

- **Patsy formulas.** Builds the factors in an equation which is inserted into a sigmoid for prediction.
- Example: *result ~ I(p1_age - p2_age)*.
- Match outcomes split for results, both a win and a lost row.
- Trained on the main tier from 1991 to 2021, tested on 2022 to 2024.
- 67.4% accuracy, or correct predictions on the test set. Could be improved with more time to finetune the formula.
- Accuracy all or nothing, log_loss expresses degree to which the prediction was correct.
- Even amount of wins and losses, so the baseline is a coin toss. The test set contains 13782 entries. A binomial test revealed that the p-value approaches 0.

Formula	Accuracy	Log Loss
Basic model + win-streak	67.39%	0.593
Basic model	65.81%	0.612
Rel. ranking model (baseline)	61.3%	0.669

6. Conclusion

Questions?