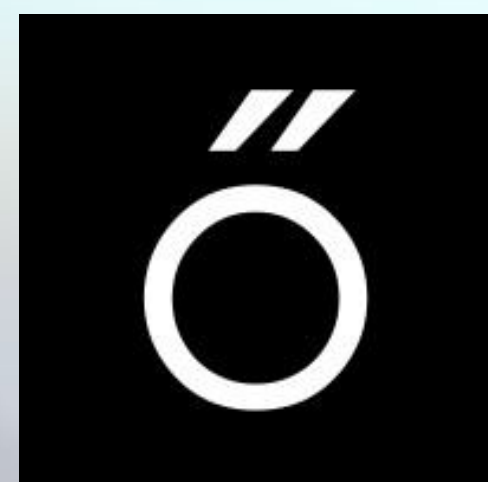




# Guess the Elo



*Investigating correlation between chess player game performance and rating.*



**The Erdős Institute**

Helping PhDs get and create jobs they  
love at every stage of their career.

**Dorian Soergel  
Foivos Chnaras  
Lang Song**

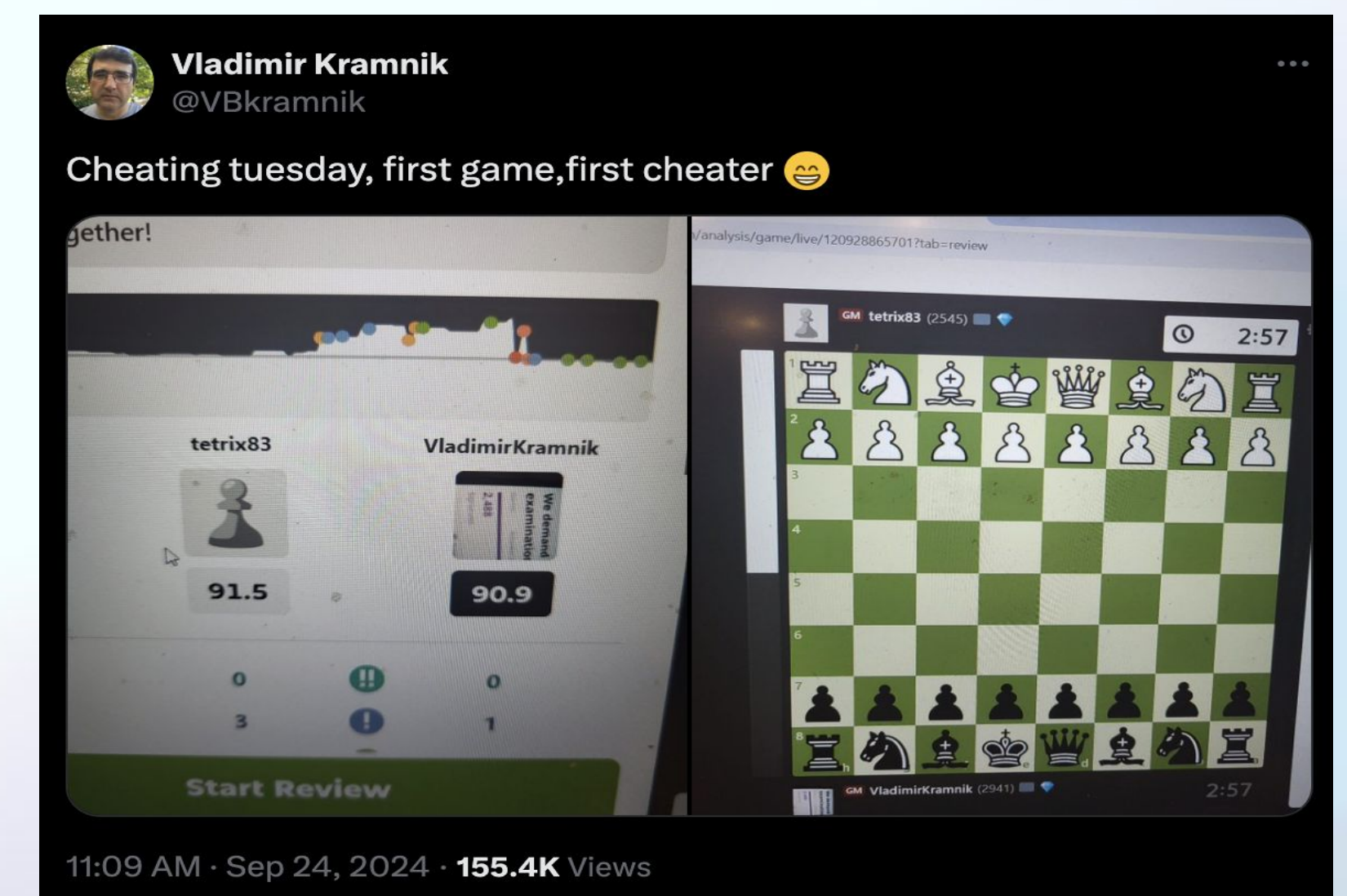
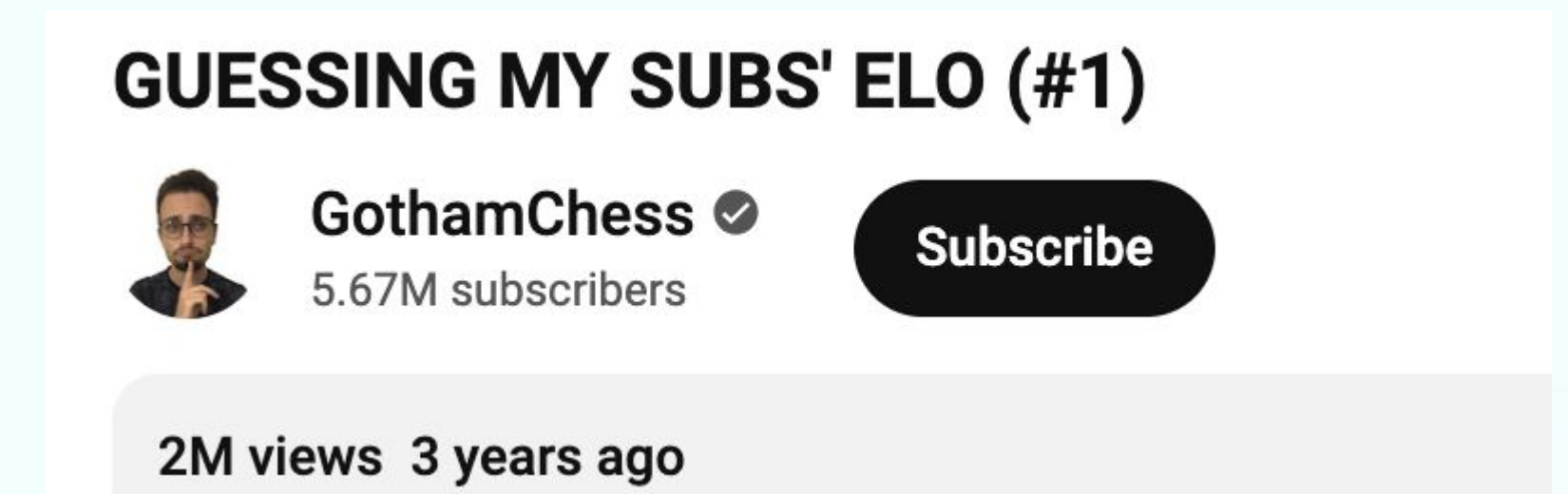
# Background & Motivation

## What is Chess Elo?

- ❑ The Elo rating system measures a chess player's skill level based on game results against other players.
- ❑ Rating Ranges (OTB):
  - ❑ Beginner: = 1000
  - ❑ Club Player: 1200 - 2000
  - ❑ Master: 2200+
  - ❑ Grandmaster: 2500+

## Why Guess the Elo?

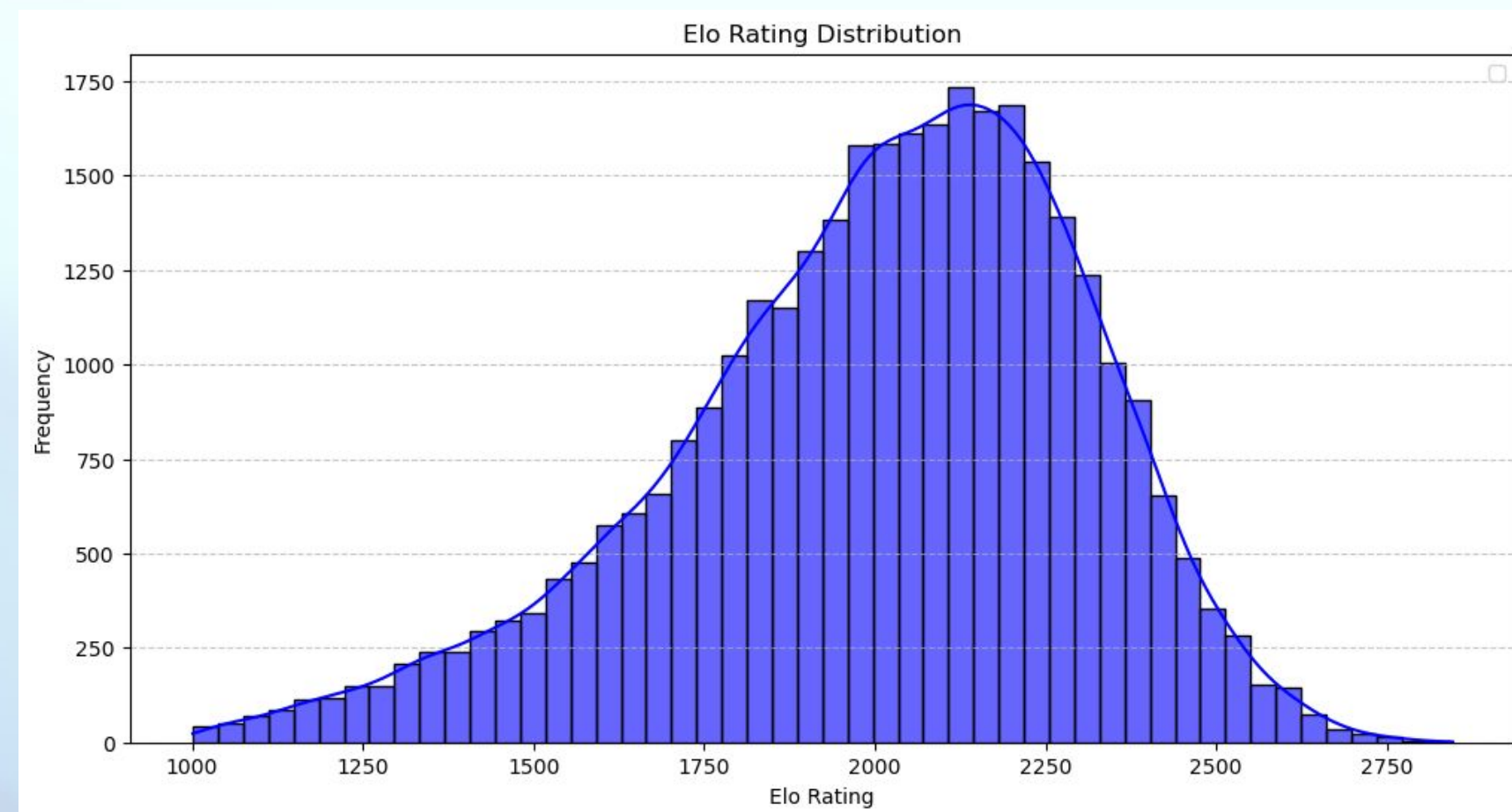
- ❑ Entertainment & Engagement
- ❑ Anti-Cheating Insights
  - ❑ Investigating whether player strength can be reliably inferred from performance at one or a few games



# Dataset

## Data Sources

- ❑ Source: from The Week in Chess (TWIC) Archive
- ❑ Over-the-board (OTB) tournament games from around the world
- ❑ Time Range: June 2012 - September 2024
- ❑ Size: Over 1 million games
- ❑ Format: PGN files, recording moves in chess notation
- ❑ Player Elo Range: 1000 - 2830



```
[Event "Cez Trophy 2012"]
[Site "Prague CZE"]
[Date "2012.06.20"]
[Round "1"]
[White "Navara,D"]
[Black "Svidler,P"]
[Result "1/2-1/2"]
[WhiteTitle "GM"]
[BlackTitle "GM"]
[WhiteElo "2706"]
[BlackElo "2741"]
[ECO "B38"]
[Opening "Sicilian"]
[Variation "accelerated fianchetto, Maroczy bind, 6.Be3"]
[WhiteFideId "309095"]
[BlackFideId "4102142"]
[EventDate "2012.06.20"]

1. Nf3 g6 2. e4 c5 3. c4 Bg7 4. d4 cxd4 5. Nxd4 Nc6 6. Be3 Nf6 7. Nc3 0-0 8. Be2
d6 9. 0-0 Nxd4 10. Bxd4 Bd7 11. Qd2 Bc6 12. f3 a5 13. b3 Nd7 14. Be3 Nc5 15.
Rab1 Qb6 16. Rfc1 Rfc8 17. Rc2 h5 18. Bf1 Kh7 19. g3 Qd8 20. Bh3 e6 21. Rd1 Be5
22. Nb5 Qf8 23. Qe2 Rd8 24. Bg5 Rd7 25. Nd4 f5 26. Nxc6 bxc6 27. Be3 Qe7 28. Bg2
Bg7 29. Rcd2 Rad8 30. Bxc5 dxc5 31. exf5 Bd4+ 32. Kf1 exf5 33. Qxe7+ Rxe7 34. f4
Rd6 35. Re2 Rxe2 1/2-1/2

[Event "Cez Trophy 2012"]
[Site "Prague CZE"]
[Date "2012.06.21"]
[Round "2"]
[White "Svidler,P"]
[Black "Navara,D"]
[Result "1-0"]
[WhiteTitle "GM"]
[BlackTitle "GM"]
[WhiteElo "2741"]
[BlackElo "2706"]
[ECO "C65"]
[Opening "Ruy Lopez"]
[Variation "Berlin defence"]
[WhiteFideId "4102142"]
[BlackFideId "309095"]
[EventDate "2012.06.20"]

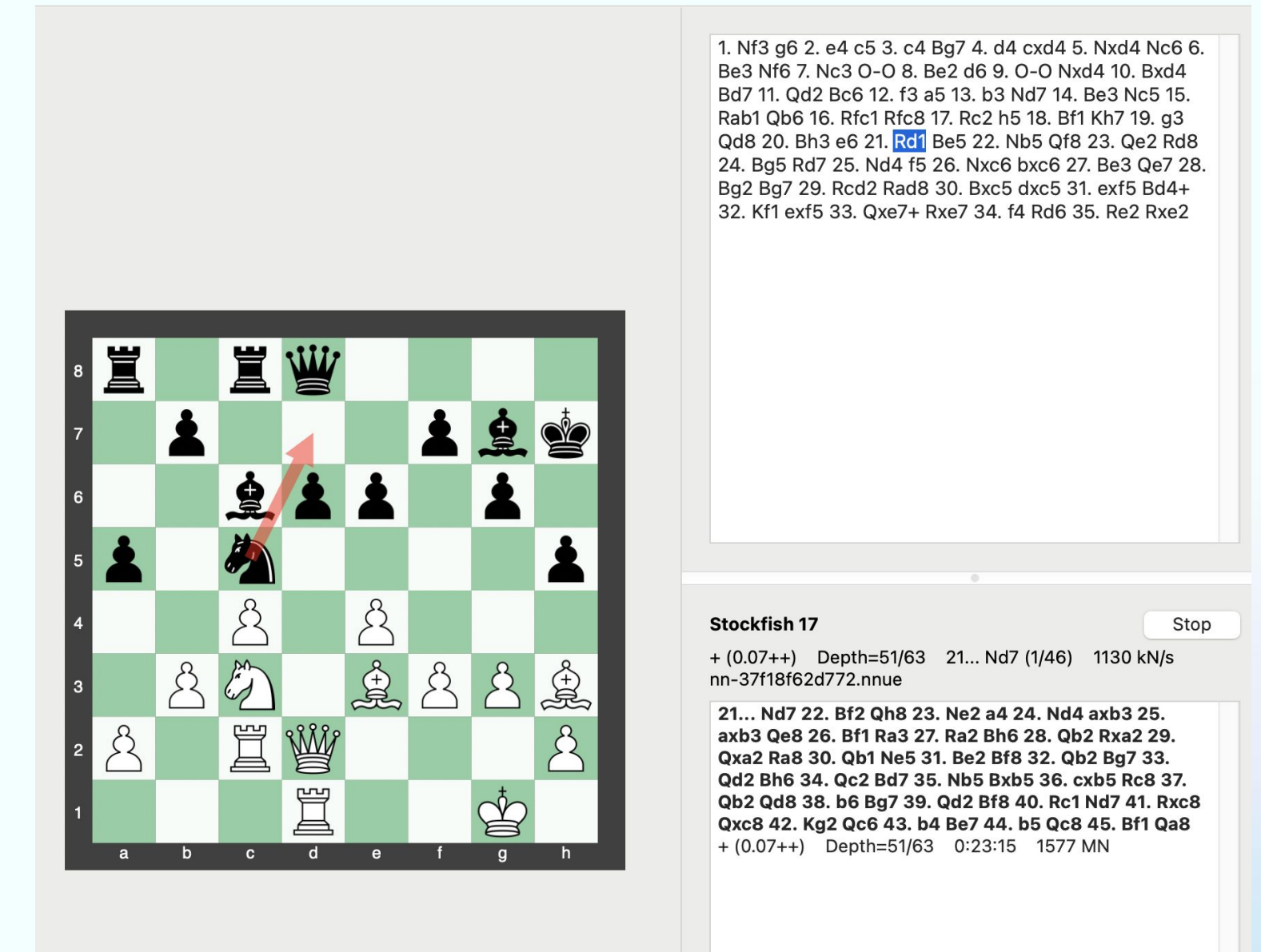
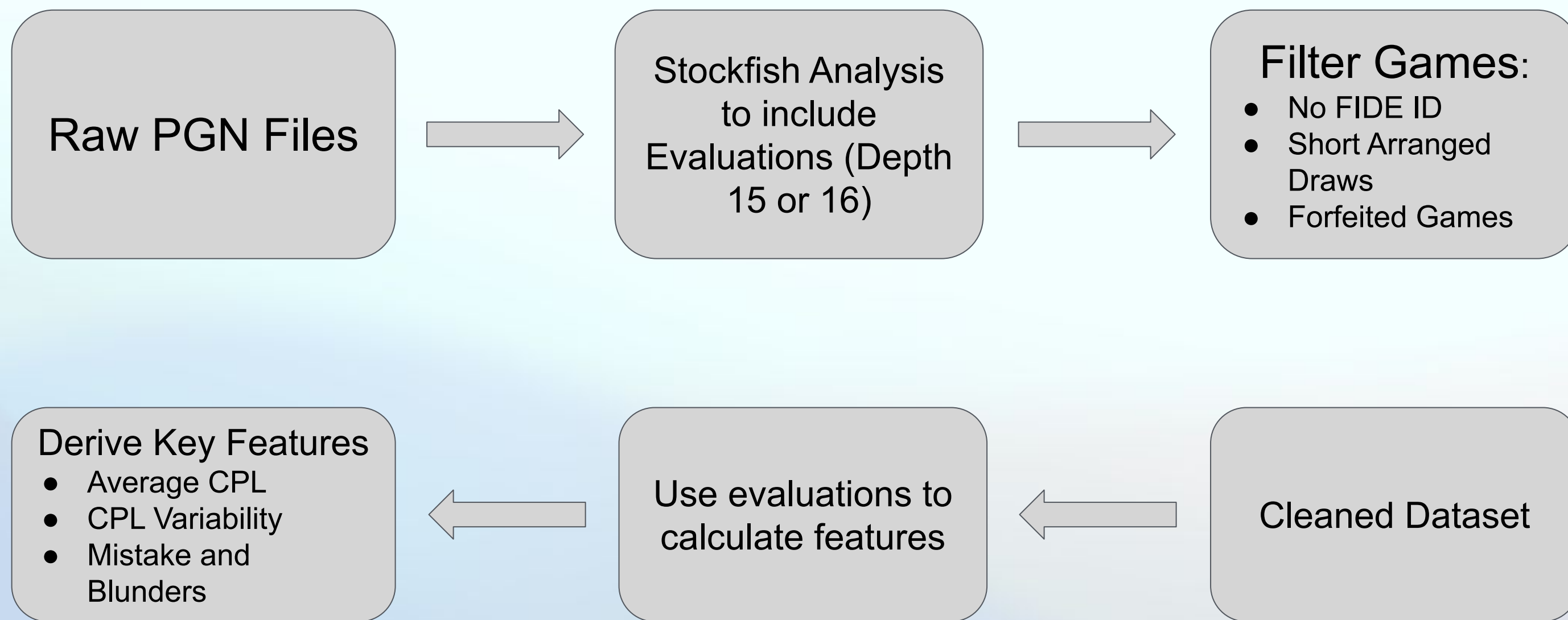
1. e4 e5 2. Nf3 Nc6 3. Bb5 Nf6 4. d3 Bc5 5. c3 0-0 6. 0-0 d6 7. Nbd2 a6 8. Bxc6
bxc6 9. Re1 Re8 10. h3 Bb6 11. Nf1 h6 12. Ng3 Be6 13. Be3 Bxe3 14. Rxe3 c5 15.
Qc2 Nd7 16. d4 cxd4 17. cxd4 exd4 18. Nxd4 a5 19. b3 Nc5 20. Rd1 Bd7 21. e5 dxe5
22. Nf3 Nb7 23. Nxe5 Nd6 24. a4 Be6 25. Qc3 Qg5 26. Nf3 Qd8 27. Nd4 Kh8 28. Nh5
Qg5 29. Nxe6 Rxe6 30. Rxe6 fxe6 31. Qxc7 Qxh5 32. Qxd6 Qe2 33. Qd3 Qxd3 34. Rxd3
Rc8 35. Re3 Rc1+ 36. Kh2 Rc2 37. f3 Rc6 38. Re5 Ra6 39. Kg3 Kg8 40. Kf4 Kf7 41.
Rb5 Kf6 42. h4 Ra8 43. g4 Ra7 44. Ke4 g5 45. hxg5+ hxg5 46. Kd4 Rd7+ 47. Kc4 Ra7
48. b4 axb4 49. Kxb4 Ke7 50. Rxc5 Kd6 51. a5 Rf7 52. a6 Kc6 53. Ra5 Rf4+ 54. Kc3
Rxf3+ 55. Kd2 1-0
```

# Data Processing

## Preprocessing and Calculate CPL

**Evaluation:** Numerical score assessing a position's favorability.

**Centipawn Loss (CPL):** The difference in evaluation between two consecutive moves.



|     |                        |
|-----|------------------------|
| ○   | Gukesh D               |
| 4   | inaccuracies           |
| 2   | mistakes               |
| 0   | blunders               |
| 31  | Average centipawn loss |
| 91% | Accuracy <i>i</i>      |
| ●   | Ding, Liren            |
| 1   | inaccuracy             |
| 0   | mistakes               |
| 0   | blunders               |
| 12  | Average centipawn loss |
| 97% | Accuracy <i>i</i>      |

# Lichess.org

- ❑ One of the two big online chess platforms
- ❑ Open API (code is in Scala)
- ❑ They have a way of computing **Total Accuracy** per game

Black: 2506



Result:  
0-1

White: 1410

● ?  
4 inaccuracies  
0 mistakes

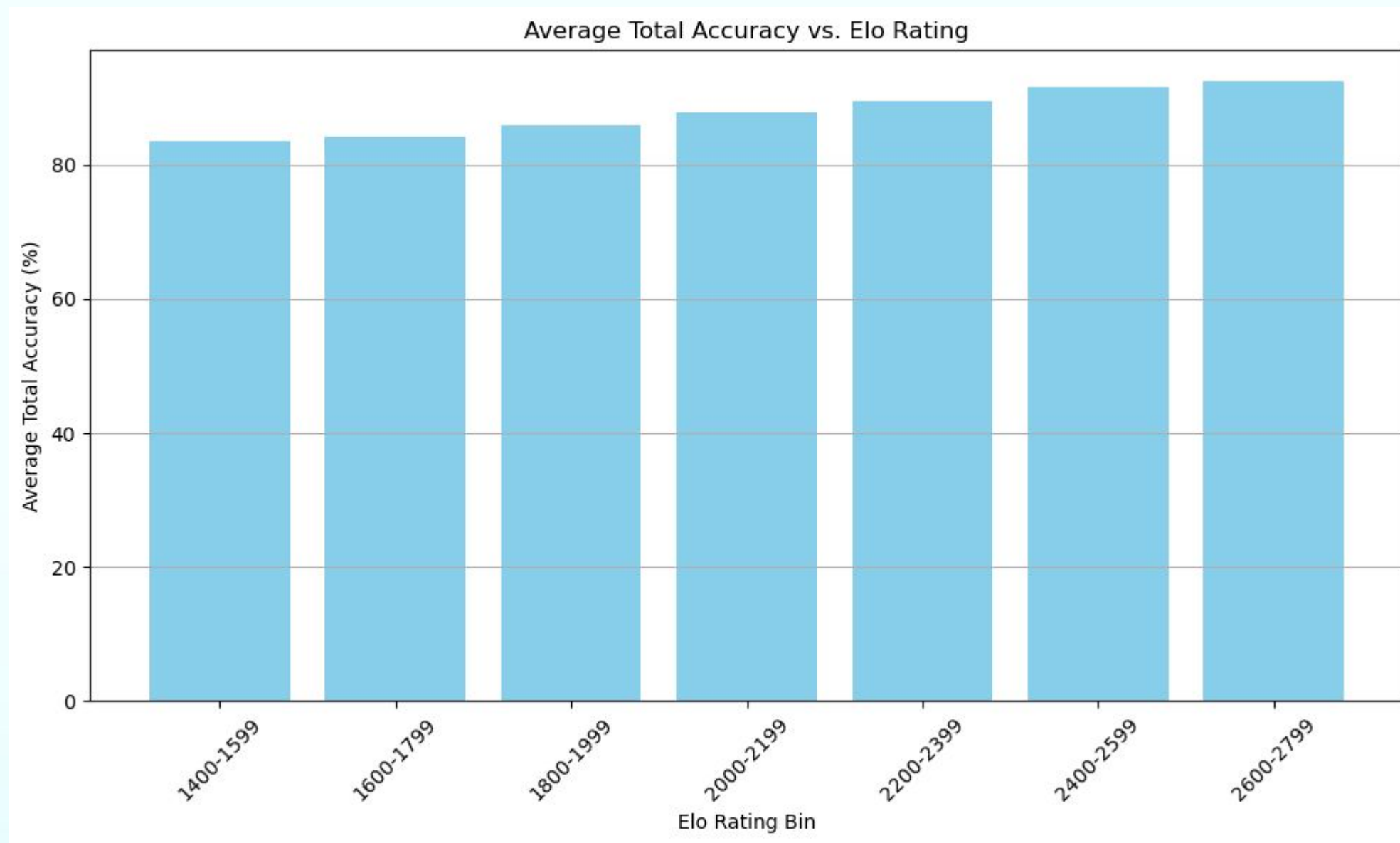
28 Average centipawn loss  
91% Accuracy ⓘ

▶ LEARN FROM YOUR MISTAKES

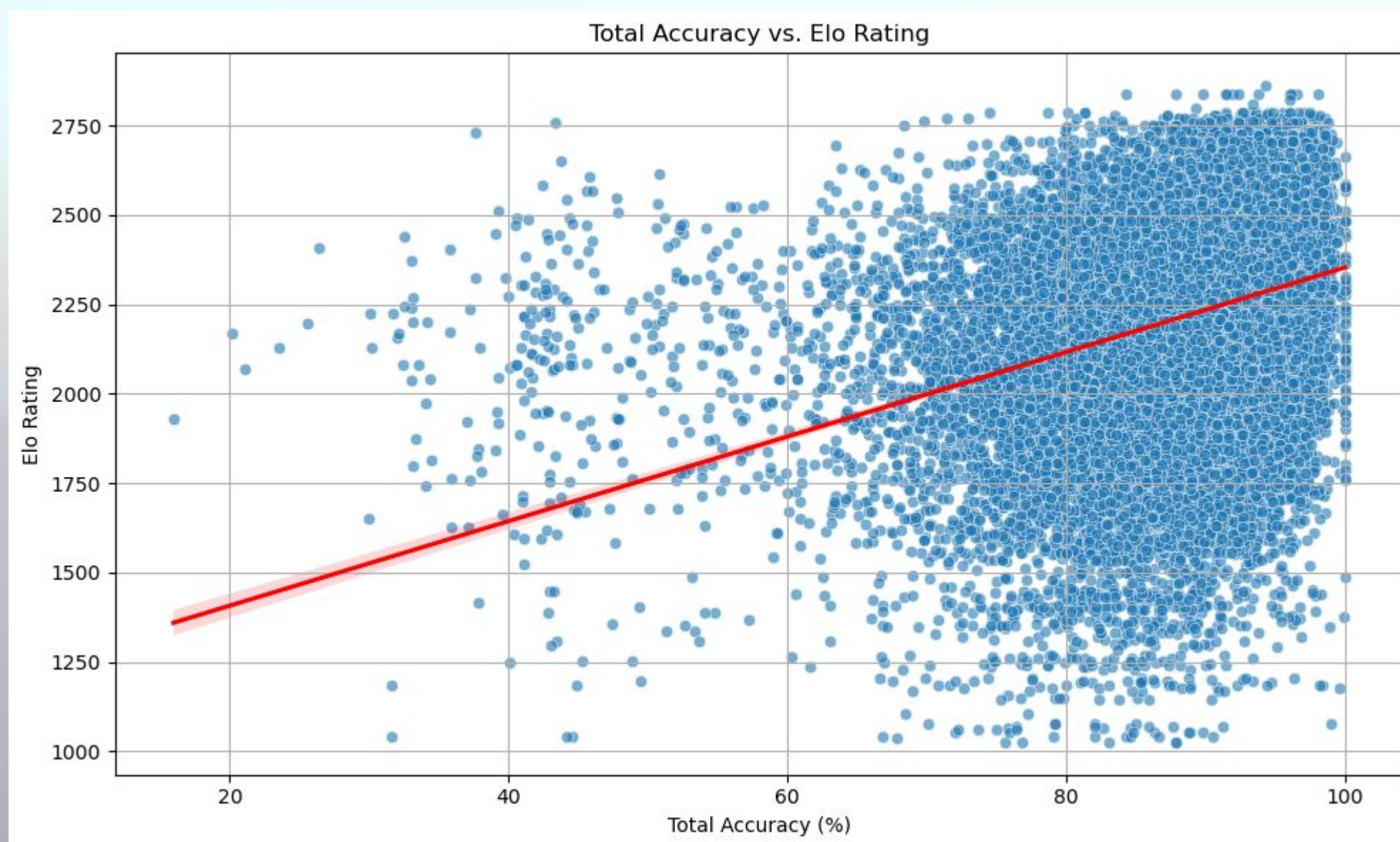
○ ?  
1 inaccuracy  
0 mistakes

14 Average centipawn loss  
96% Accuracy ⓘ

# Regressing Lichess' Accuracy over Elo



| Model             | RMSE | R <sup>2</sup> | Within ±300 Elo (%) |
|-------------------|------|----------------|---------------------|
| Linear Regression | 261  | 0.13           | 77.8                |
| Random Forest     | 274  | 0.04           | 74.9                |
| Baseline          | 281  | 0              | 72.9                |



- ❑ The correlation between total accuracy and Elo is 0.33
- ❑ There is no strong correlation between Total\_Accuracy and Elo!!

# Building New Features

- ❑ **Winning Chance:** We count how many games had at least one position with evaluation in the interval  $(x, x+0.2)$ .  
Then **Winning Chance = % games won**

- ❑ **Winning Chance Loss (WCL):**  
We compute the difference between two consecutive winning chances.

❑ **Benefit:**

| Evaluation                       | CPL | WCL   |
|----------------------------------|-----|-------|
| Move 10: +7.00<br>Move 11: +5.00 | 2   | ~4%   |
| Move 10: +2.00<br>Move 11: 0     | 2   | ~27 % |

- ❑ **Average Winning Chance Loss (AWCL):**  
The average of WCL per player per game

- ❑ **Classifying mistakes:**  
Separate WCL in intervals of 5%.  
Measure number of mistakes per interval per game:

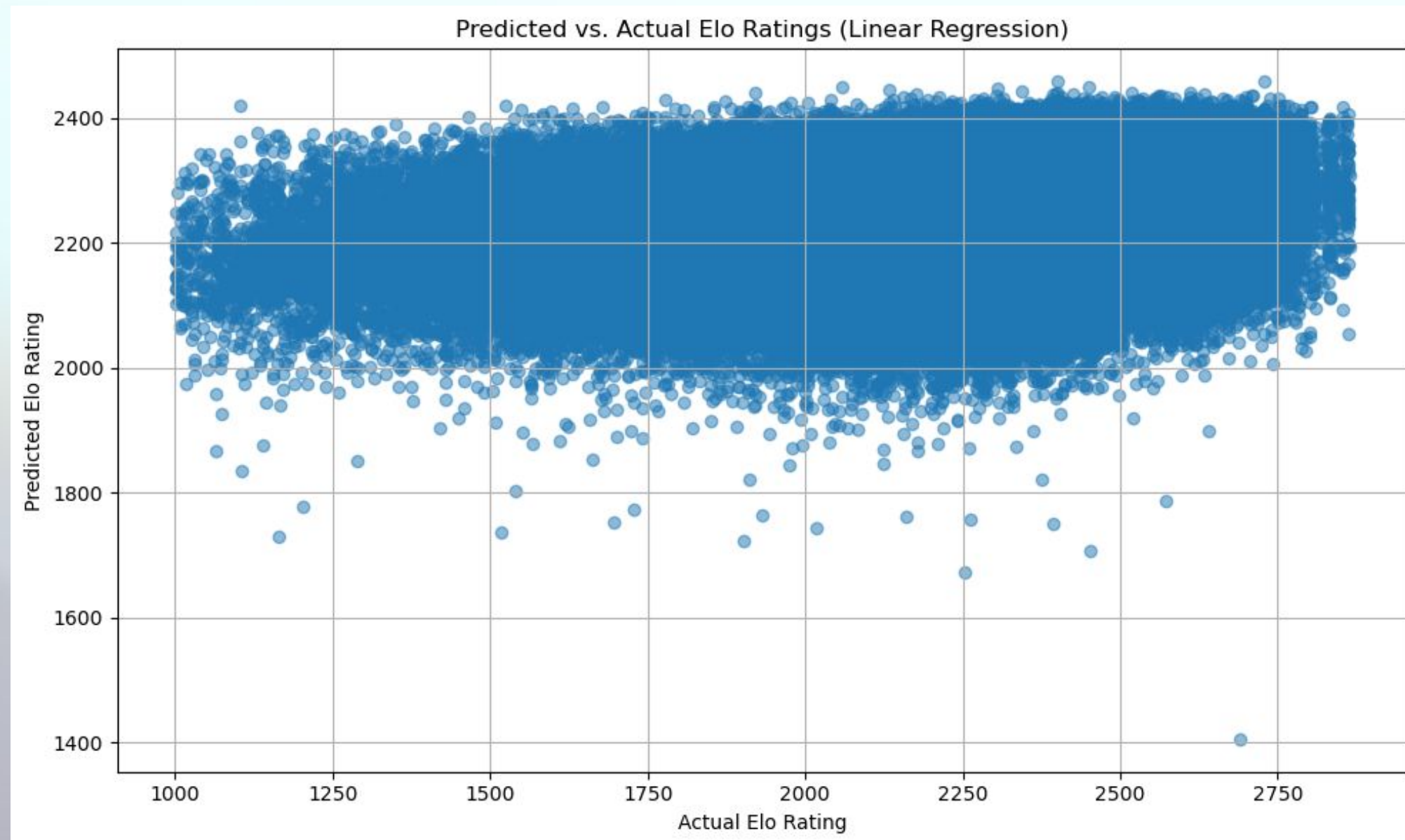
| (5,10] | (10,15] | (15,20] | (20,25] | (25,30] | (35,40] | (40,50] | (50,60] | (60,70] | (70,100] |
|--------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| 4      | 4       | 0       | 1       | 0       | 0       | 0       | 0       | 0       | 0        |

- ❑ **Room for improvement:** Develop better features

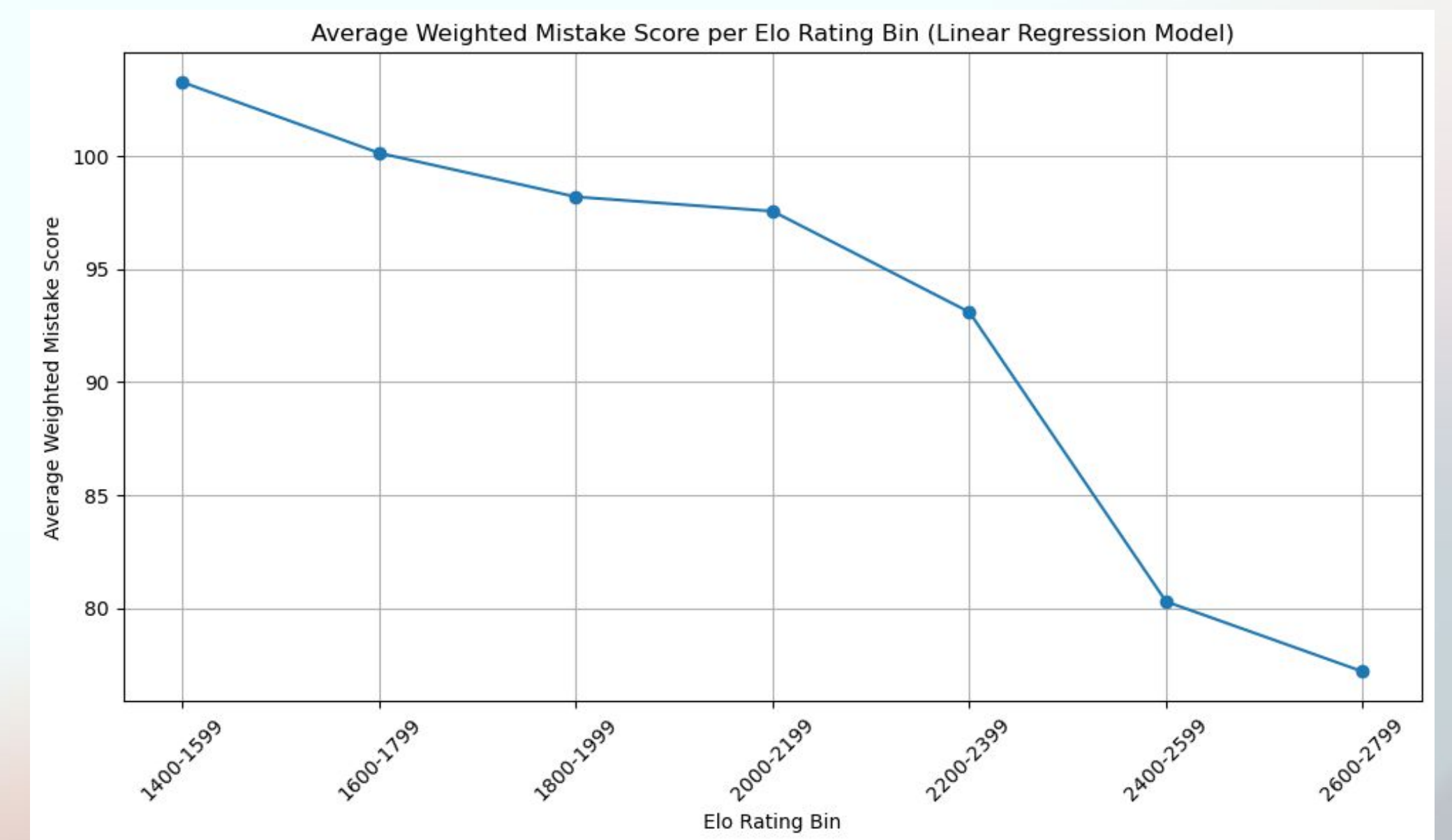
# Results

❑ Trying to predict Elo out of a single game proved difficult.

| Model             | RMSE | R <sup>2</sup> Score | Within ±300 Elo (%) |
|-------------------|------|----------------------|---------------------|
| Linear Regression | 257  | 0.16                 | 78.0                |
| Random Forest     | 274  | 0.04                 | 74.9                |
| Baseline          | 281  | 0.                   | 72.1                |

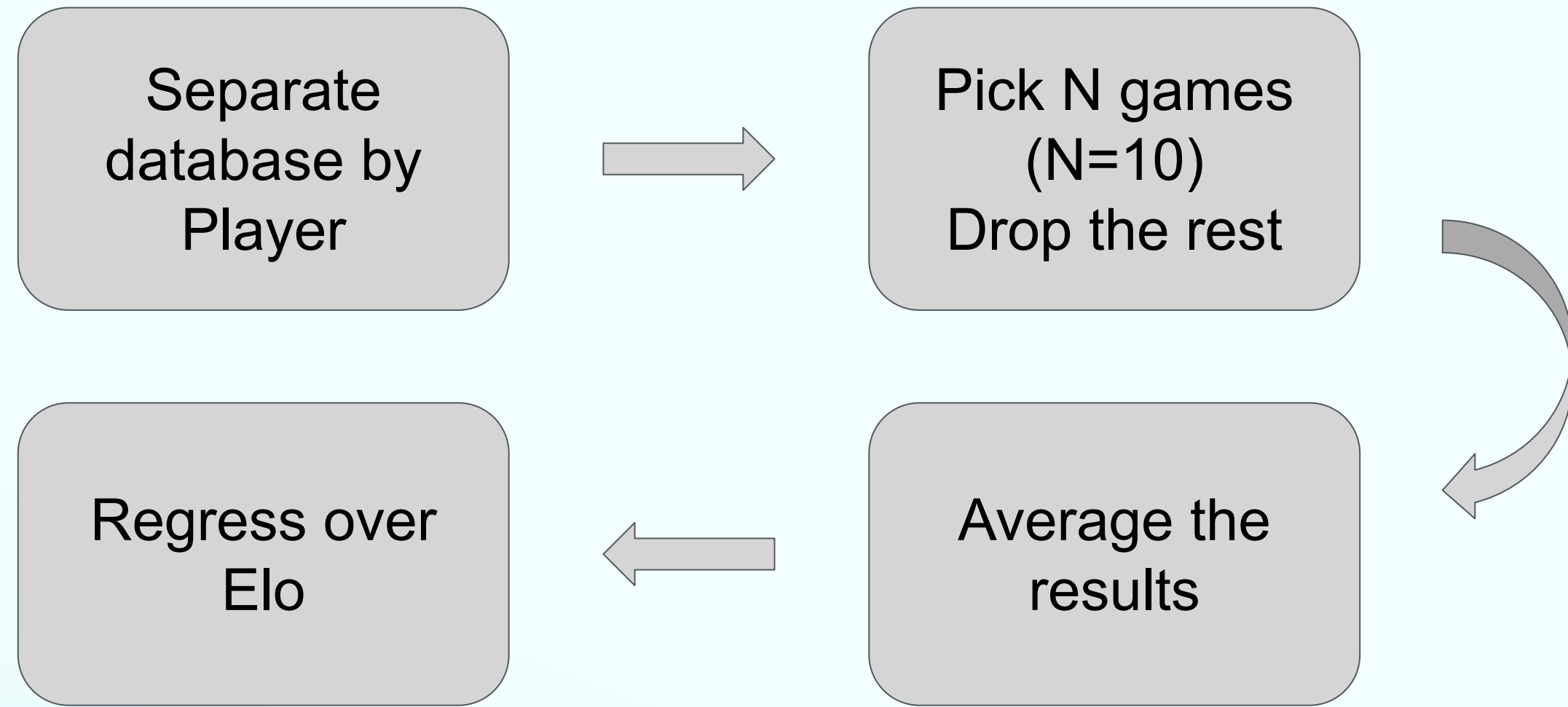


❑ On Average, there seems to be a clear correlation

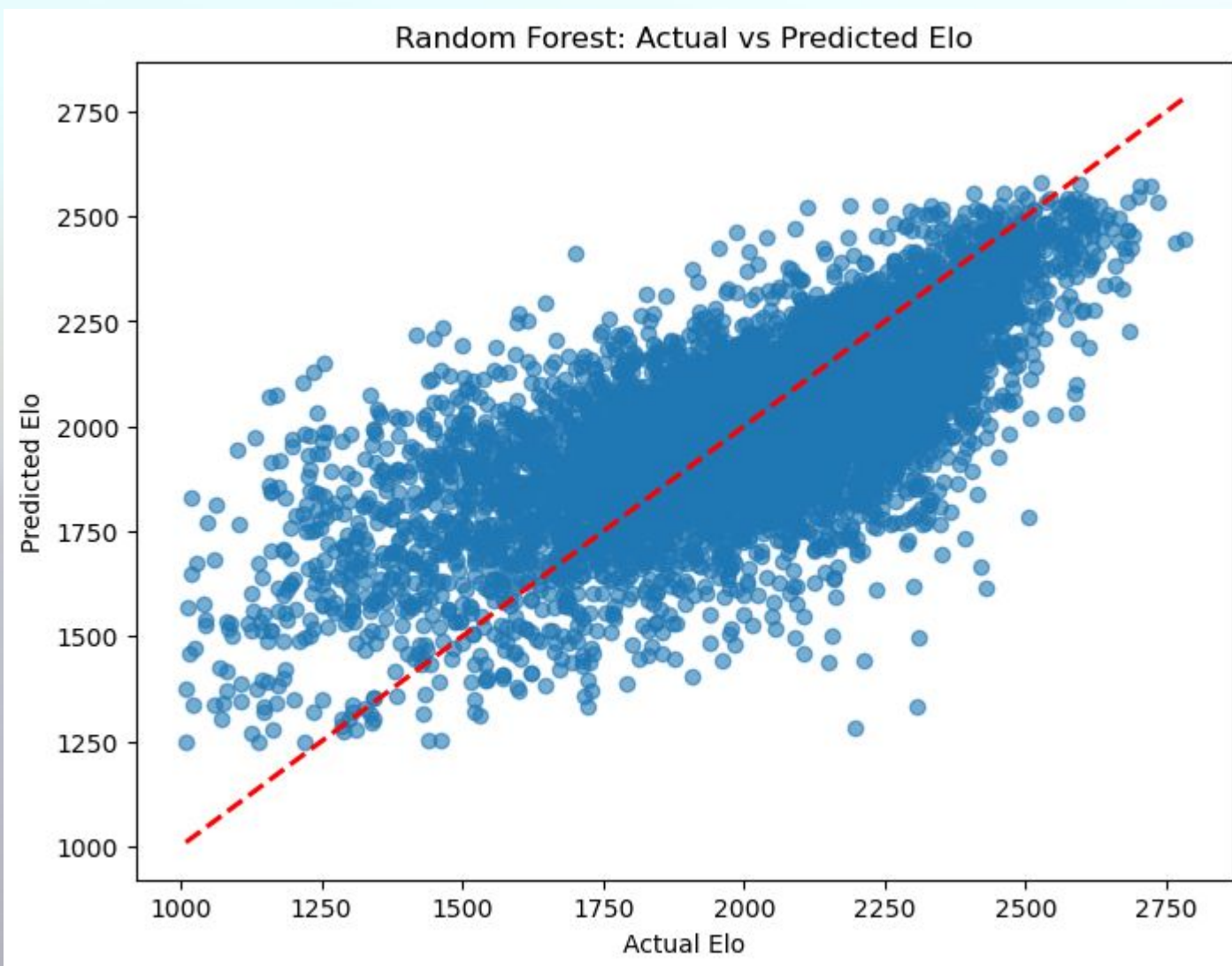
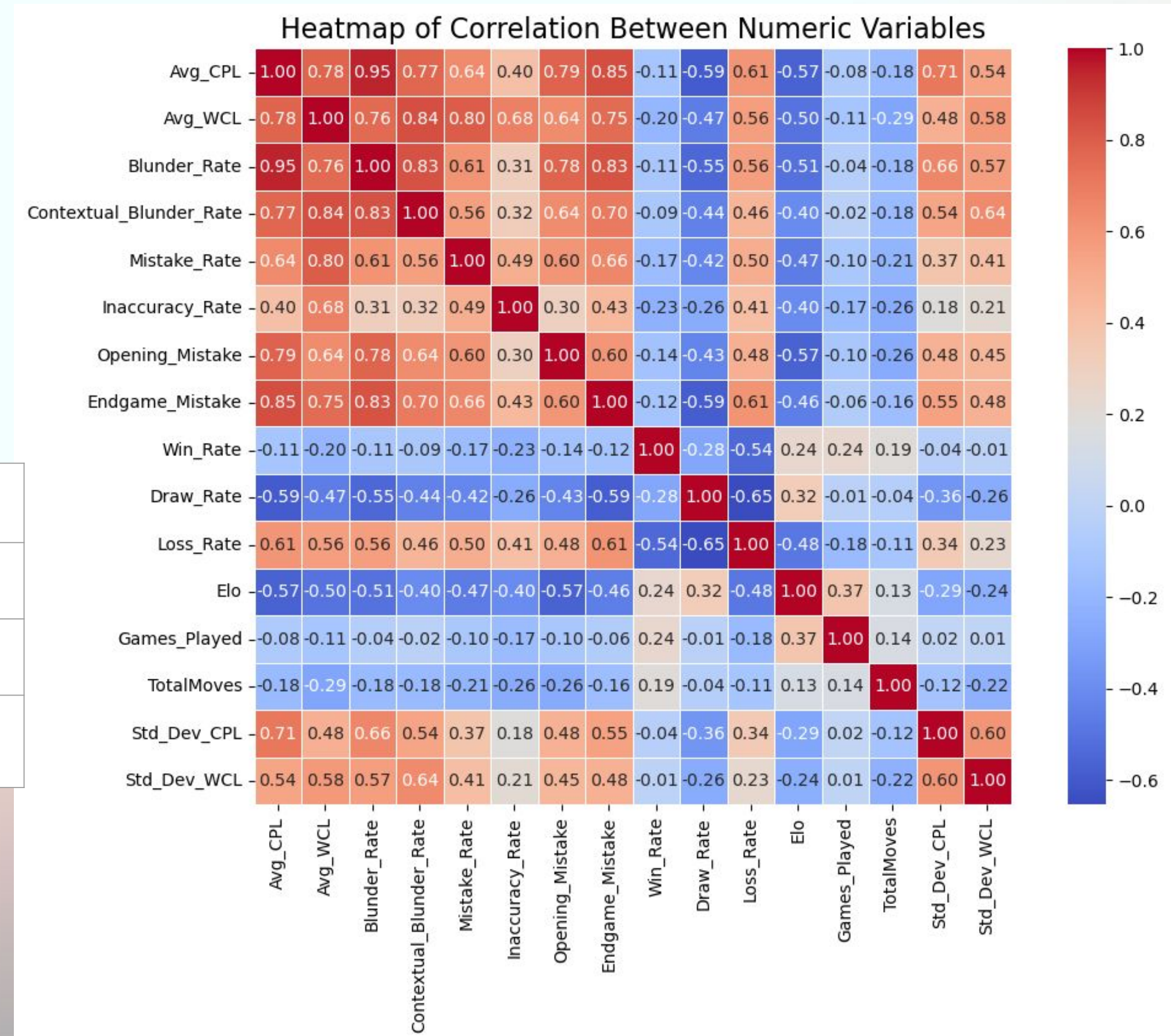




# Predictions using multiple games



- Original database: ~ 1 mil games
- Final database: ~ 20k data points



| Model                   | MAE | RMSE | R <sup>2</sup> Score |
|-------------------------|-----|------|----------------------|
| Linear Regression       | 180 | 231  | 0.41                 |
| Ridge Regression        | 180 | 231  | 0.41                 |
| Random Forest Regressor | 167 | 216  | 0.49                 |

# Conclusions

- ❑ For a single game, we found no strong correlation between game performance and Elo
- ❑ Analysis of a single game cannot be used as an indicator of cheating with our current model
- ❑ Using multiple games, results were a lot more accurate.

# Future Work

- 1. Higher Depth Analysis:** use depth 20 in stockfish for more accurate evaluations
- 2. Expand Feature Set:**
  - Use multiple games of every player (requires bigger database)
  - Move Ranking: ranking of the move played (e.g. Top-3, Top-5 move)
  - Position Complexity
  - Opening, Endgame Mistakes
- 3. Explore Deep Learning Approaches**
  - Leela Chess Zero: Investigate the potential of neural network-based chess engines for evaluation accuracy
  - Sequence Models: Apply models like RNNs to learn patterns from move sequences
- 4. Broaden Applications**
  - Cheating Detection
  - Expand to other strategy board games or card games
  - Build Web Applications for Elo Prediction