How the Pros Play:
An analysis of lichess.org's ranked chess games

Alex Jose, Davis Trinh, Jing Duan, and Joe Strawinski
The Data Analytics Boot Camp at The University of North Carolina at Chapel Hill

(Dated: July 13, 2019)

ABSTRACT

Chess is the most popular board game in the world. Its popularity continues to expand as it has become more accessible to play thanks to online services. Many of us have casually played a game of chess before (or at least heard of chess), but have never considered the consequences of the first move we take or the sophisticated strategies that have been developed by advanced players. Using chess match data obtained via Kaggle and Lichess, we answered several questions that had piqued our interest. From our results, we observed: the player who goes first has a slight statistical edge over the player who goes second, upsets are common however they occur much less frequently as the disparity in skill level widens, competitive matches often end in resignation rather than checkmate, matches tend to last longer when players are evenly matched, and that there are statistically favorable and unfavorable opening moves.

I. INTRODUCTION

Chess has never been more accessible to play. A large contribution towards the community's growth is it's expansion into online services. The top chess players in the world play online matches and stream themselves on sites such as YouTube and twitch.tv. Lichess.org, a popular site for online chess games currently serves over 35 million ranked matches a month.

A single chess game can be broken into three stages: Opening, Mid-game and End-game. The opening consists of the initial set of moves played by both white and black. Traditionally white makes the first move, and it is black's role to respond. The opening "sets the stage" of the game as well as the overall pace of the game. Since openings can be described in approximately 5 turns, it is a common point of study for high level chess players. Most openings in chess have been classified under the Encyclopedia of Chess Openings (or ECO).

By Mid-Game and End-Game both players are now no longer relying on previously studied positions, and instead use their general intuition and skill to pick positions that will place their opponent in *check*.

Emulation is a large part of improving in chess. Studying great chess players and analyzing common positions is common practice for mid and high level players. In this report we obtain general statistics on the online player base, and attempt to extract characteristics of game play that distinguish highly rated players. We also demonstrate how online chess can be used to analyze and improve your own games through the use of player specific analysis.

II. DATA ACQUISITION AND CLEANING

Although Lichess.org hosts fewer games than its main competitor, chess.com, it has a more accessible API. Furthermore it is possible to download every ranked game that has ever been played on the site. For general usage statistics we use a Kaggle dataset that was obtained using API calls on 1500 ranked players. This set contains approximately 20,000 ranked games. The only cleaning required for this set was removing games that were not completed.

A CSV was also generated using the set of all ranked games of May, 2015. This set was contained in a PGN format, which was read using the python-chess module.

The sets include game type, which player won, and their opening moves. The entire PGN move list is recorded as well in a standard format. Both sets contain the Elo of both white and black. Elo is a numerical ranking system that is calculated using the relative skill of a player. A higher Elo player has better, and more frequent wins. The highest ranked players in the world have Elo ratings of 2800

A wrapper for the Lichess API was used to collect all of the Lichess.org games played by chess grandmaster Zhigalko Sergei. This set contains the same information as the other sets we collected. For this set, we applied the python-chess module to parse the move list and create virtual boards. The string representing the board of was then manipulated to conform to a 2D array.

Games obtained from the raw PGNs that were missing relevant headers were removed from our analysis.

III. RESULTS AND ANALYSIS

A. General Game Statistics

Chess rules dictate that the player using the white pieces makes the first move. This results in a slight favor towards white. White wins 50% of games while Black only wins 47% of games.

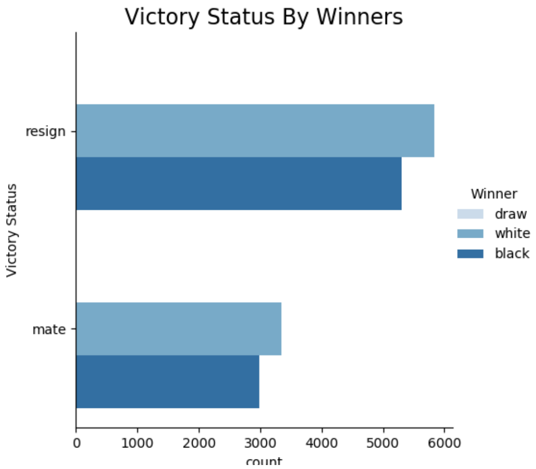


FIG. 1. The amount of winning games ending in resign or mate. This plot also shows that a majority of winning games are won by white.

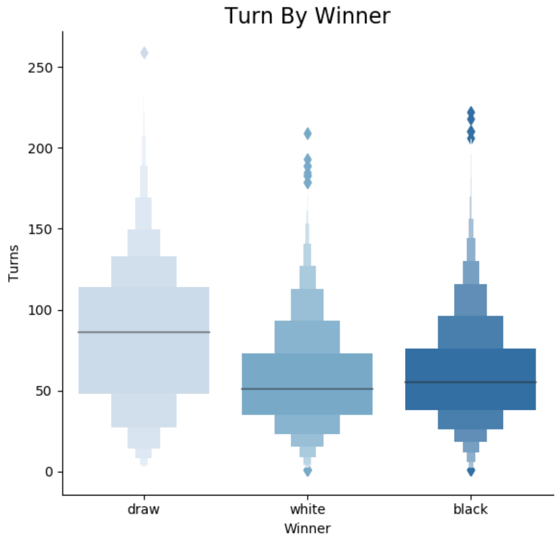


FIG. 2. A box and whisker plot with histogram and outliers.

We examined several relationships relating to the number of turns played in a game. It is worth noting that most games do not end with "check mate." Instead, most players resign a few turns before they inevitably lose (Figure 1.)

The number of turns for draws is significantly larger than turns for winning games. Intuitively this is because draws tend to occur when there are not enough pieces in play for either player to check mate their opponent. This results in a repetitive back and forth chase, that ends in a draw. In figure 2 it can be seen that games that black wins take more turns on average. This could be a reflection of black's turn disadvantage.

We analyzed the relationship between player's skill and the amount of turns in a game. Although it is not strongly correlated, when the average skill of both players is larger, games have more turns in general (Fig 3. When there is a large skill differential between the players, there is a slight trend towards games ending with fewer turns (Figure 4). These trends agree with our expectations, yet we are surprised that the trends are not pronounced.

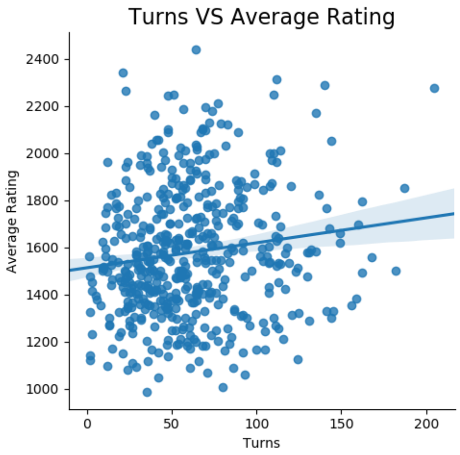


FIG. 3. Points represent individual games. Average rating is calculated as the average Elo between players.

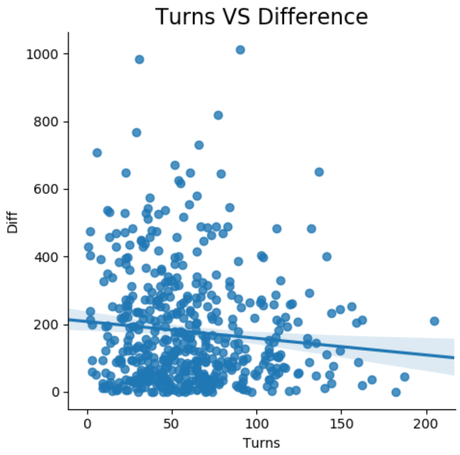


FIG. 4. The magnitude of the difference between Elo was used to approximate the skill gap. Within statistical uncertainty, games with a large skill gap end in fewer turns.

On Lichess, it is possible for users to play games that do not contribute to their ranking. We checked if these unranked games had different features compared to ranked games. Based on figure 5, the distribution of games played is mostly the same between ranked and unranked games. The most significant difference is higher ranked players experience more draws when playing unranked games. This could be indicative of players opting for more experimental techniques and less of a drive to win.

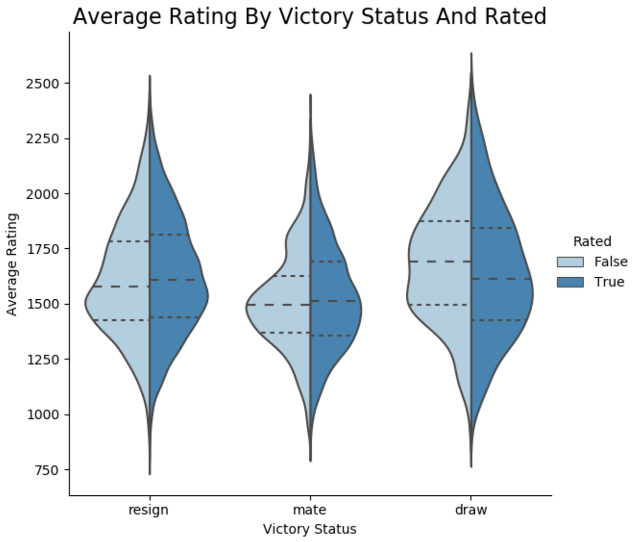


FIG. 5. A set of histograms of victory status corresponding to the average rating between players. It also shows the differences between ranked and unranked games.

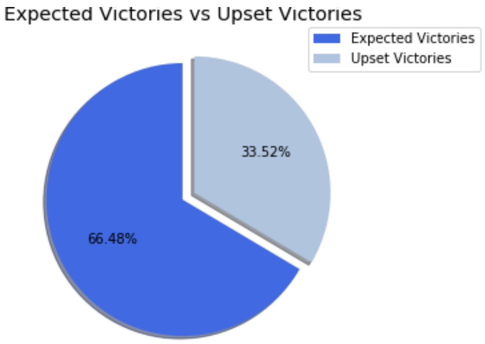


FIG. 6. Overall Upset Potential

B. Elo based trends and Upset Potential

In our investigation we looked at a 2,000,000 game sample from Lichess.org and found that about one third of the total matches ended up in an upset - defined here as a draw or loss for the higher rated Elo player Figure 6. To further elaborate on this we looked at how often upsets occur when there is a wide Elo disparity between players.

As one would intuitively expect, the frequency of upsets decreases as the skill differential between two players increases. From figure 7, we can see that the distribution of upsets is skewed to the left. A majority of upsets, by our definition, occurs when both players are evenly skilled (Elo ¡ 500).

As previously stated, white's first move advantage impacts white's overall win percentage (Figure 8). Does white still hold a slight advantage over black as skill differential increase? According to the data (Table 9), white generally retains a slight edge over black across most Elo differential groups, however the edge appears to diminish when theres a vast disparity in skill (Elo ¿ 800). Does white still hold a slight advantage over black when black is favored? Analysis of the data suggests that in matches between evenly skilled player (Elo ¡ 500), where the favored player went first, had a slightly higher win rate relative to when the favored player went second. Once again, this slight advantage diminishes as the gap between skill increases. Matches where Elo ¿ 800 saw the advantage change slightly in favor of the player who went second (Figure 10).

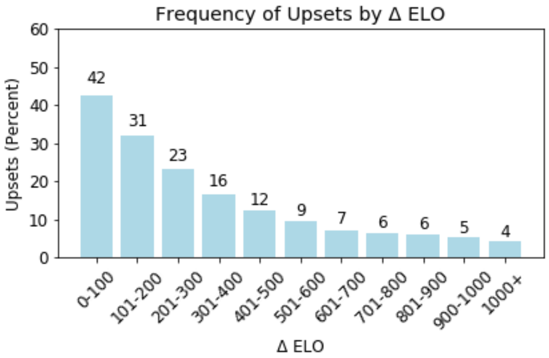


FIG. 7. Upset Potential, Binned by Elo difference

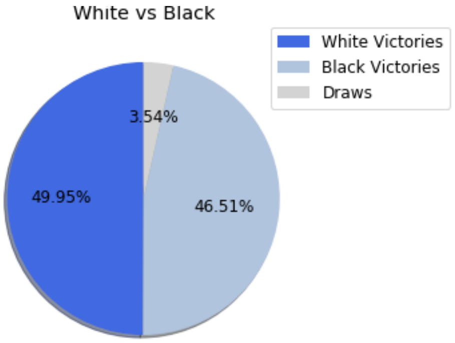
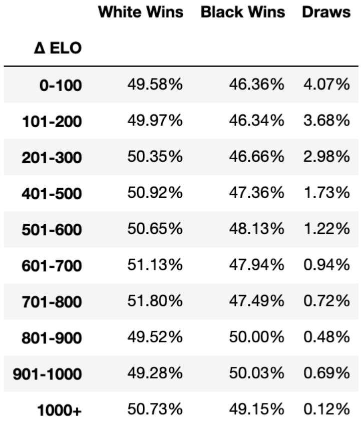


FIG. 8. Overall win percentage based off of 2 million ranked games.

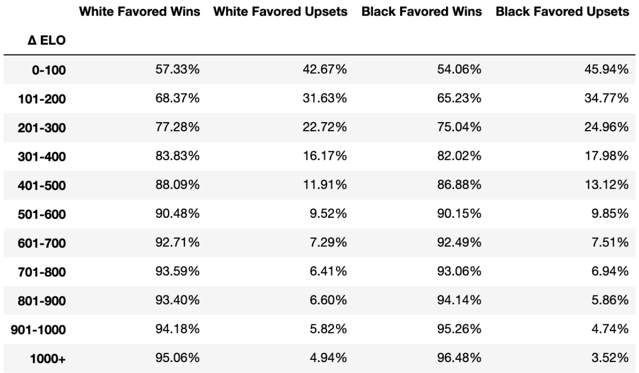
C. Opening Move Analysis

An important clarification to be made is that an opening consists of multiple moves. At what point an opening is classified as an opening is a relatively arbitrary decision. Some openings branch into variations and numbers. For this reason, when studying the effectiveness of a particular variation, we limit ourselves to only positions that occurred frequently. We calculate the "average score" of an opening by looking at white's success for individual



	White Wins	Black Wins	Draws
Δ ELO			
0-100	49.58%	46.36%	4.07%
101-200	49.97%	46.34%	3.68%
201-300	50.35%	46.66%	2.98%
401-500	50.92%	47.36%	1.73%
501-600	50.65%	48.13%	1.22%
601-700	51.13%	47.94%	0.94%
701-800	51.80%	47.49%	0.72%
801-900	49.52%	50.00%	0.48%
901-1000	49.28%	50.03%	0.69%
1000+	50.73%	49.15%	0.12%

FIG. 9. Win percentage for white and black, grouped by skill differential. White's first move advantage is less prominent when there is a large skill gap



	White Favored Wins	White Favored Upsets	Black Favored Wins	Black Favored Upsets
Δ ELO				
0-100	57.33%	42.67%	54.06%	45.94%
101-200	68.37%	31.63%	65.23%	34.77%
201-300	77.28%	22.72%	75.04%	24.96%
301-400	83.83%	16.17%	82.02%	17.98%
401-500	88.09%	11.91%	86.88%	13.12%
501-600	90.48%	9.52%	90.15%	9.85%
601-700	92.71%	7.29%	92.49%	7.51%
701-800	93.59%	6.41%	93.06%	6.94%
801-900	93.40%	6.60%	94.14%	5.86%
901-1000	94.18%	5.82%	95.26%	4.74%
1000+	95.06%	4.94%	96.48%	3.52%

FIG. 10. Percentage of wins based on which color player was favored. This statistic is analyzed for different bins of skill disparity.

games. An opening is awarded a single point if white wins, zero for a draw, and minus one for white's loss. These game scores were then averaged for groups of the the same opening.

1. *Openings with Variation*

The table 11 summarizes the most successful openings with variation for white.

The opening with the highest score for white is a the Mieses-Kotroc variation of the Scandinavian Defense. In this opening black takes white's pawn with his queen. While this gives an early piece advantage, it seems like it forces black to defend his queen early. This ultimately makes it one of the worst opening positions for black.

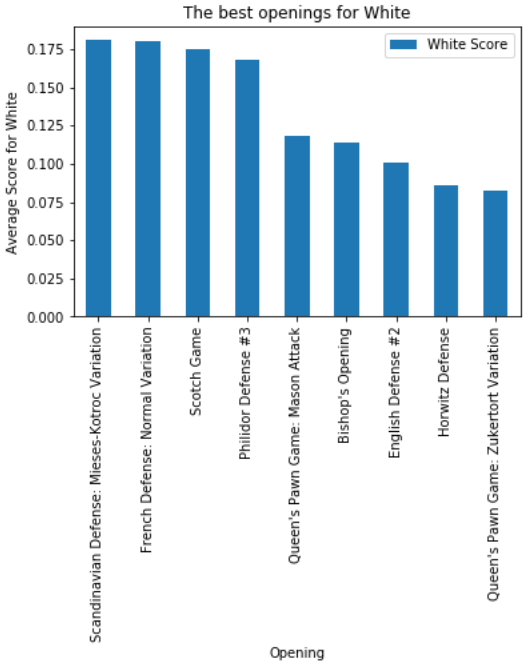


FIG. 11. The most successful openings positions for white.

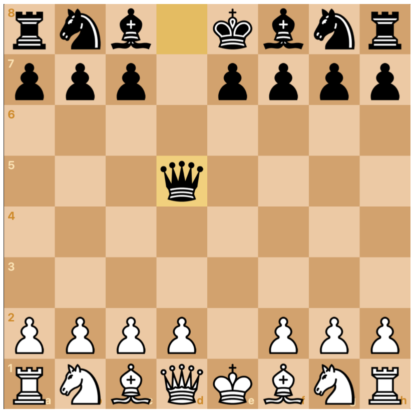


FIG. 12. The Scandanavian Defense: Mieses-Kotroc Variation

Figure 13 contains the worst scoring openings for white.

The worst position for white is the Bowdler Attack variation on a Sicilian Defense. White opens up the pawn in front of his king. Black responds making room for the queen. Then white moves their bishop to half board. This is a popular, yet disadvantages position for white. White has exposed their king and bishop, while black has the opportunity to work with their queen.

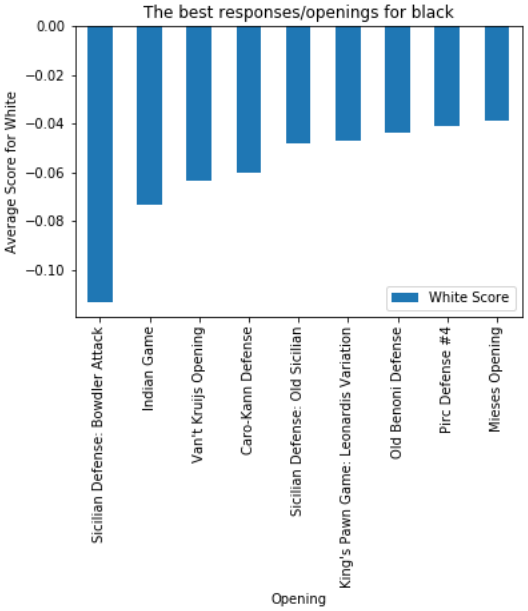


FIG. 13. The best opening positions for black

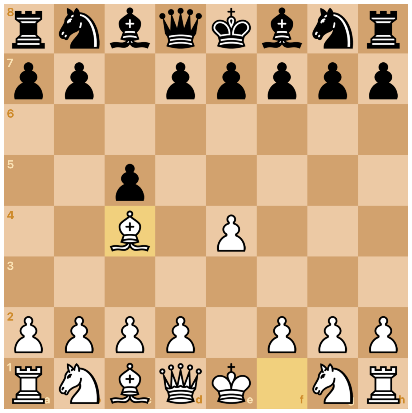


FIG. 14. Sicilian Defense Opening: Bowlder Attack Variation

2. *Opening Categories and success within Categories*

Taking a deeper look at the data, we analyzed game outcomes of variations (sub-category) in relation to the outcomes of the grouped opening (category). Our comparison of the individual variations vs. the success of the overall group within the opening category, illustrates that not all variations are created equal. It is worth noting that differences in game outcomes exist among variations and although a category of opening may indicate favorable or unfavorable success for either player based on the data, players should study and understand the differences in success for variations within an opening move in order to maximize their results. To further our investigation, we judgmentally selected (based on popularity) two openings to analyze. The Queens Gambit and Ruy Lopez are two of the most widely touted opening moves for white. The Queens Gambit includes four variations. According to the lichess.org data, all variations produced favorable outcomes for white. However, a sizable difference exists between the Queens Gambit (most favorable: 0.2908) and Queens Gambit Declined (least favorable 0.0934). These outcomes are in-line with our expectation as the Queens Gambit Declined is a response from black declining White's temporary pawn sacrifice as a defense to the Queens Gambit 1.d4. We also analyzed the Ruy Lopez, also referred to as the Spanish Opening. Among highly rated players, the Ruy Lopez is often regarded as one of the best opening moves for white. Note, it was a favorite of Bobby Fisher and more than 100 variations exists for the opening that begins with white creating a potential pin of the d-pawn or Knight. The Ruy Lopez immediately sparks an attack, putting pressure on black. However, due to the complexity and number of variations for the Ruy Lopez, a player should study and understand all the variations to maximize success. Our analysis focused on the top and bottom three variations. The least favorable outcomes for white include: - Closed Variations: Lutikov Variation - Closed Variations: Worrall Attack - Open Variations: Karpov Gambit The most favorable outcomes for white include: - Central Countergambit - Brentano Gambit - Retreat Variation

D. Player Specific Analysis: GM—Zhigalko Sergei

Sergei is ranked top 300 in the world by the World Chess Federation. His board game winning configurations were analyzed to provide insight towards top level end-game strategies. Heatmaps were generated from his winning games.

From the knights map we see that the GM favors to keep knights close the center of the board, slightly closer to his side. This position pressures a large range of squares, meanwhile protecting pieces closer to his side of the board.

Pawns on the outside tend to be developed less often. By moving the center pawns up the GM allows more key pieces to be developed: bishop, queen etc. An interesting trend here is the asymmetry. The right most pawn is moved more often than the left most pawn.

As expected, the king is often protected at the right corner of the board. It is worth noticing that the most frequent square is not the original starting square for the king. It is likely the king is moved to this square using a technique known as castling.

The queen tends to be placed in a few distinct patterns. The queen is found near the ends of the board, while also controlling the main center diagonals. His queen tends to take an extremely aggressive position on black's starting rows.

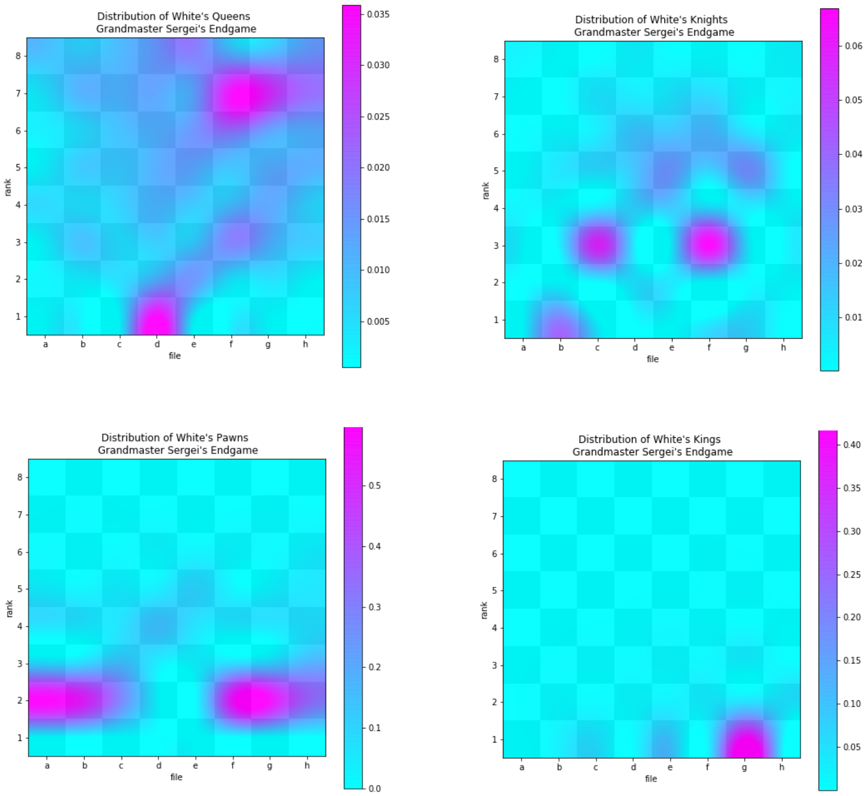


FIG. 15. Heatmaps of piece locations taken from winning board configurations.

IV. ACKNOWLEDGEMENTS AND SOURCES

Opening boards were provided by chess.com. All data was sourced from lichess.org's API. Special thanks to cyanfish's python client for lichess's API.