Basic Information

Project Title: Chess Visualization

Name/uid/email : Stephen Harman u0509544 stephen.harman@utah.edu
Garret Cervantez u0577500 u0577500@umail.utah.edu
Alex Hamrick u1126365 u1126365@utah.edu

Project Repo: https://github.com/AlexHamrick/dataviscourse-pr-chess-visualization.git

Background and Motivation

Alex: I grew up playing chess from a very young age, and I placed second in the Idaho State Chess Championship at the age of 7. After that, I quit playing and following chess until 2018 when my interest in chess was revived by the World Chess Championship. I watched all of the games between Magnus Carlsen and Fabiano Caruana, and I began playing and studying the game again.

As I have begun to re-learn about chess, I've found so many new openings, strategies, and tactics that I never knew as a kid, and I have often wondered how to best prioritize my time learning the game. As a (now) average skilled player, studying and memorizing uncommon openings has allowed me to play above my skill level by catching my opponents off guard with moves they have never seen before. This led me to question whether such strategies would continue to work at the highest level, and so I wanted to do a project to learn more about chess openings, their success rates at the highest level, and how effective becoming a specialist at a given opening would be. I also wanted to do a project that would allow us to analyze the effects of technology on chess players, as youtube and online chess engines have been instrumental in my chess revival, and I imagine similar tools have also helped those playing at the highest level.

Ultimately, I proposed a chess related project due to my own background and interest in chess, but I think our proposed project will also interest those who lack any real background with chess.

Stephen: I also grew up playing chess, but casually with friends and family. I am an intermediate player who will still play from time to time. I really enjoy chess and thought that it would be a fun project to work on.

Project Objectives

The purpose of this project is to learn more about the history of chess, understand how chess play and players have evolved in the modern era, and derive information regarding any correlations between openings, player rating, and win rate.

The first major question our project hopes to answer is: do skilled players develop expertise in one or two openings, or do they have a broader control of a large quantity of openings? Is it more effective for players to invest large amounts of time in becoming experts at a few openings, or is their time better spent studying other things? For example, Magnus Carlsen is known for playing the Sicilian Defense, but does he really play this opening a

disproportionate amount compared to other players of his caliber, and if so, does this niche expertise wind up ultimately helping or hurting him?

The next question our project hopes to answer is: what are the general trends for various common openings? Do certain openings minimize the possibility of a draw between the two players? Which openings favor black or white? If you are in a must win situation, which openings should you consider?

Our project also hopes to discover how chess players (their habits and elo) have changed over time. As chess has gotten more and more popular, we expect that the best players will reflect an overall upward trajectory in elo over time. Additionally, we want to be able to analyze how the rise of computers and chess AI have affected chess players. Now that machines surpass humans at chess, have players been able to effectively utilize new technologies to improve even further? Or has such a discovery led to an increase in drawn games as players are more able to effectively prepare for their opponents?

There are two big advantages that would come from answering the above questions. The first applies to the chess community. Understanding which openings are most effective and how specialization affects players will give players more insight as to what they should spend their time studying and could be a good indicator as to how important opening theory really is. Especially since modern openings rely on heavy memorization of hundreds or thousands of possible lines, this could indicate whether the best players specialize in just a few lines or whether they need to study every line equally.

The second benefit can be applied more generally. Understanding how play has evolved over time allows us to speculate as to how things like chess AI and technology have affected other games as well. If we find a sharp increase in skill among all high rated players in the modern era, it is likely because of the increased accessibility of high quality chess resources via the internet. Alternatively, if we see that only the best players improve, then we can hypothesize that the easy access to information allows the most skilled players to maximize their skill gap compared to other players. Insights like these can be applied more generally to other games, sports, and activities and could inspire related research in other areas.

Finally, this visualization will hopefully provide people with the ability to answer their own questions about chess and its history while also encouraging them to discover questions of their own!

Data

- Data Source
    - https://drive.google.com/file/d/0Bw0y3jV73lx_NElnLWVlNG9KNkU/edit?usp=sharing
        - WARNING: 831 MB
        - Found on this site
            - https://chess-research-project.readthedocs.io/e”n/latest/
- Data Info

- The data pgn files for GBs of chess games. Each line is a field (e.g. EventName, Elo, ECO, etc.) or a move (e.g. “1. d4 d5”). Not all fields are always present e.g. some game may not have a date.

Data Processing

The data processing is fairly substantial. Each line of the 2.86 GB file must be iterated over. From here we will create a csv where each record is now a game. Each game will have all of the following fields:

- Date
- White Name
- White Elo
- Black Name
- Black Elo
- Result
- ECO
- Number of Moves

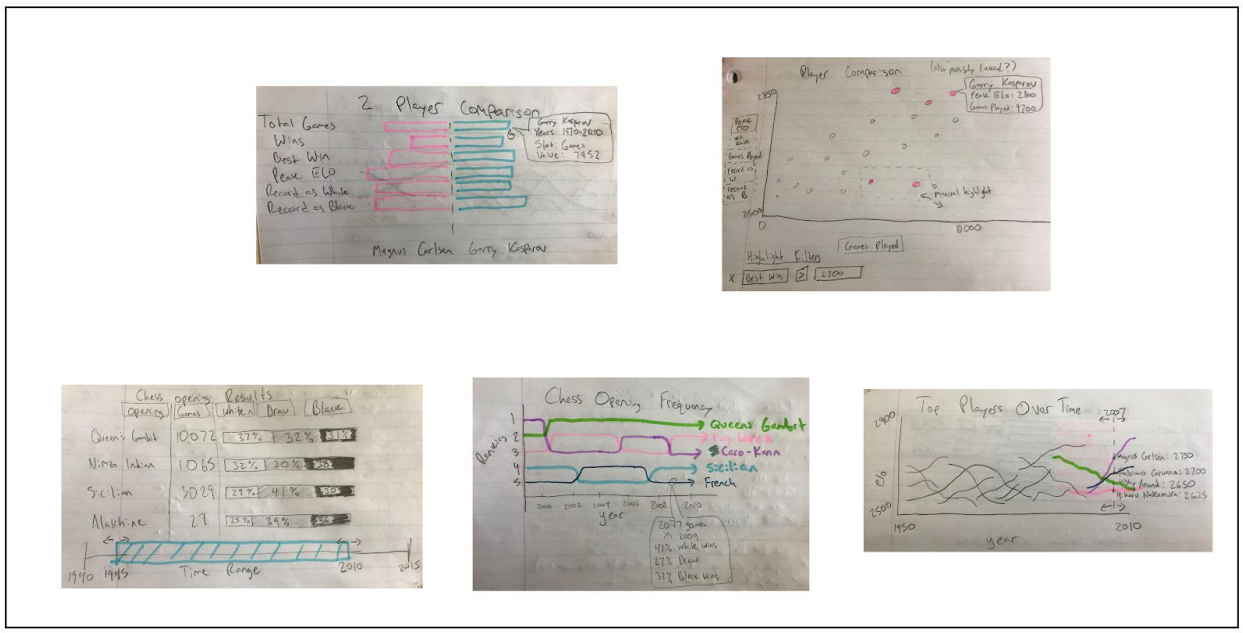
If one of these fields are not available then that game will be ignored. Let’s call this the main.csv. The main.csv will then be used to limit games by selecting only the best players from each decade. So we are going to group the main dataset by decade of game and player. Then the max elo for that decade will be taken. We will create a set of players by taking the *n* best players per decade with best defined using the max elo over that time period. From here we will then filter the main dataset to only include games where one of the players is in the best players set. Let us call this best_players.csv.

This best_players dataset is much more manageable in terms of size and format. From here we also have the ability to easily create other csvs to more easily conform to the specific visualizations we are creating and will do so as needed.

Design

Prototypes

Prototype I



A better view of each visual can be seen below. Let's discuss each visualization individually first. We will go left to right, and top to bottom.

The top row is focused on specific players and their rankings based on different stats. The top left image shows a visualization in which users can select two different players on which we have data and compare them based on a variety of factors (wins, games played, etc). This would be a nice way to display specialized information about individuals.

The top right image, on the other hand, displays all the players from a given time period and plots them based on customizable attributes (again, using wins, games played, max elo, etc). This scatterplot would allow highlighting of players who meet a certain criteria and would show more information about the individual when hovering over their data point.

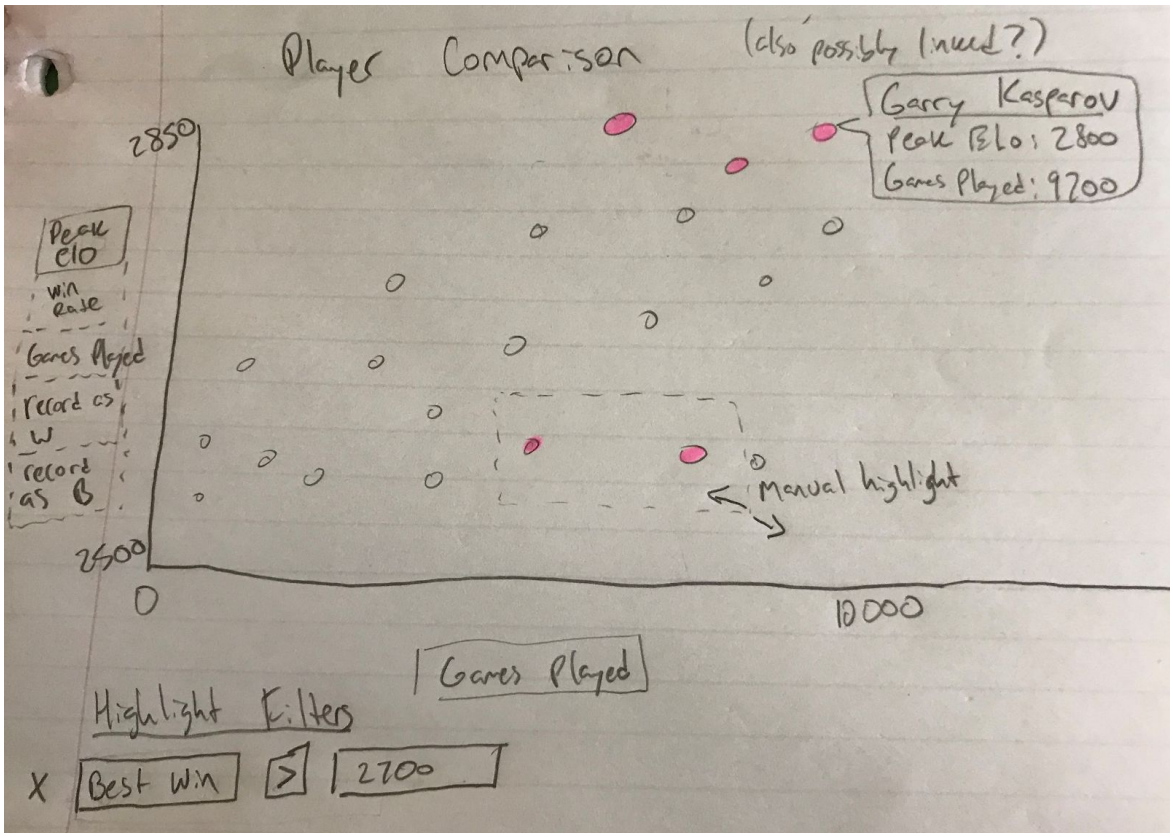
The bottom row focuses both on chess openings and general chess trends over time. The bottom right visualization lists all the played openings and shows the number of times they were played, as well as the win percentages for each player when playing this opening. Each category (games played, black win rate, white win rate, etc) would be sortable, allowing users to quickly see both the most common openings as well as the most effective openings for each color.

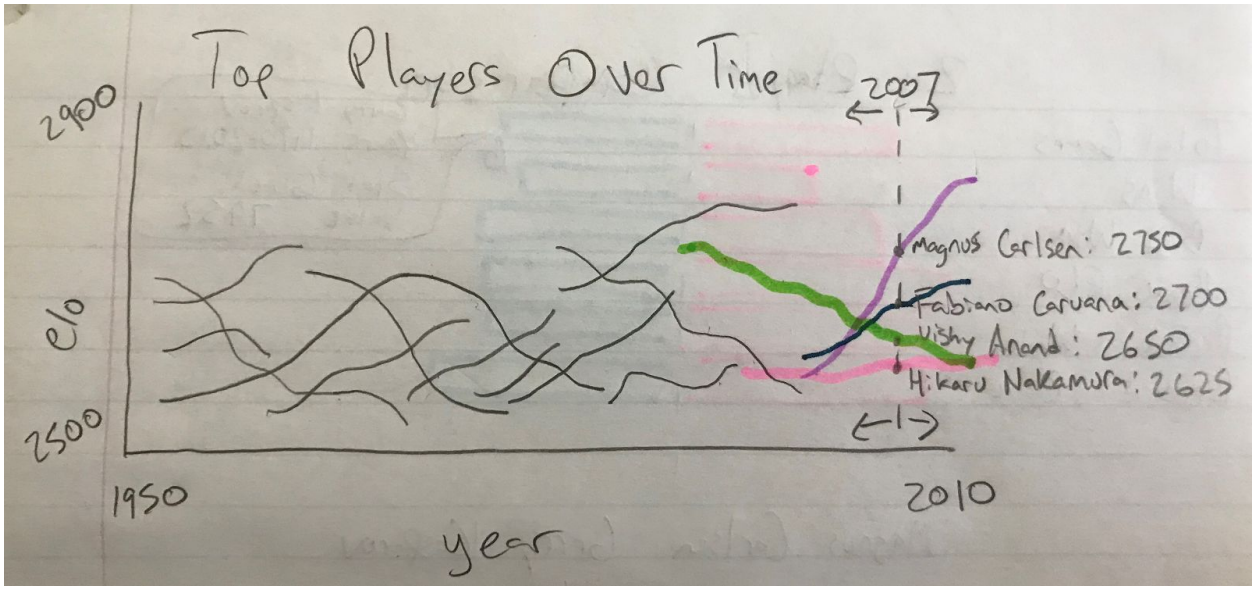
The bottom middle visualization shows the rankings of most popular chess openings over time. This makes it easy to see the trends in chess openings and gives us an idea of which openings have remained effective and which have died out over the years. Additionally, hovering over a point on one of the lines would provide additional info about the games which

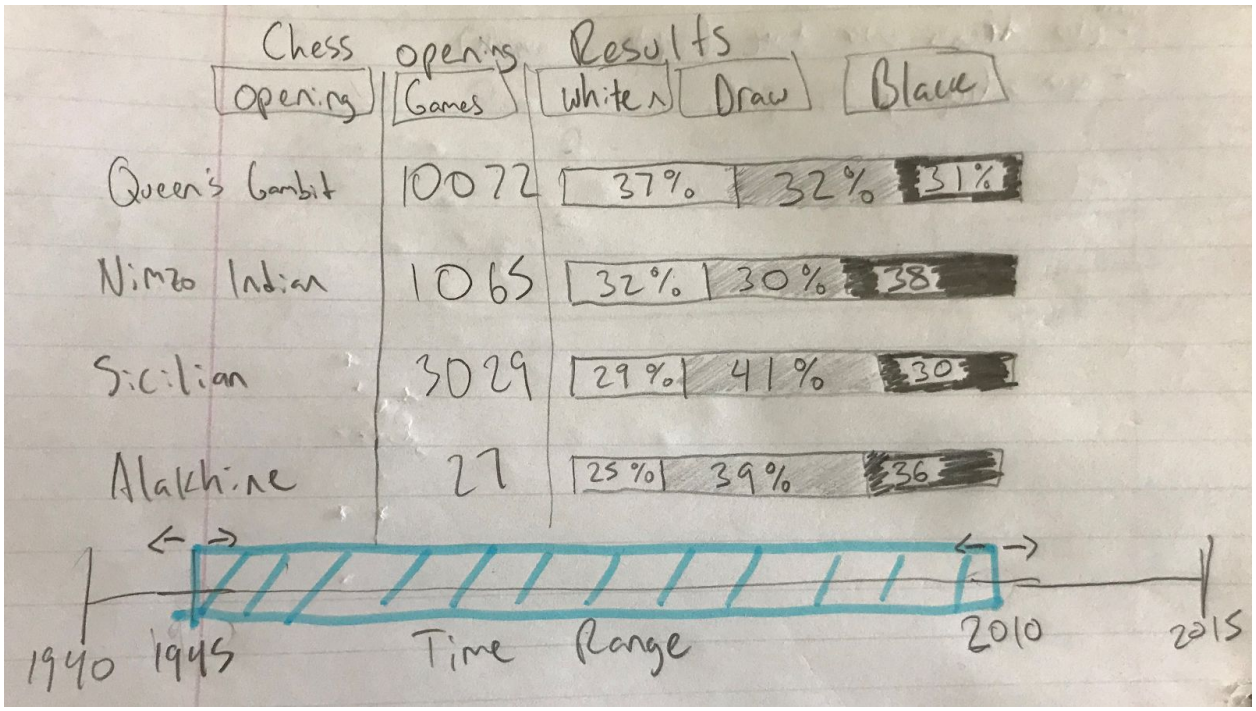
played that opening during that year, allowing users to dive deeper into the opening trends than just their frequency.

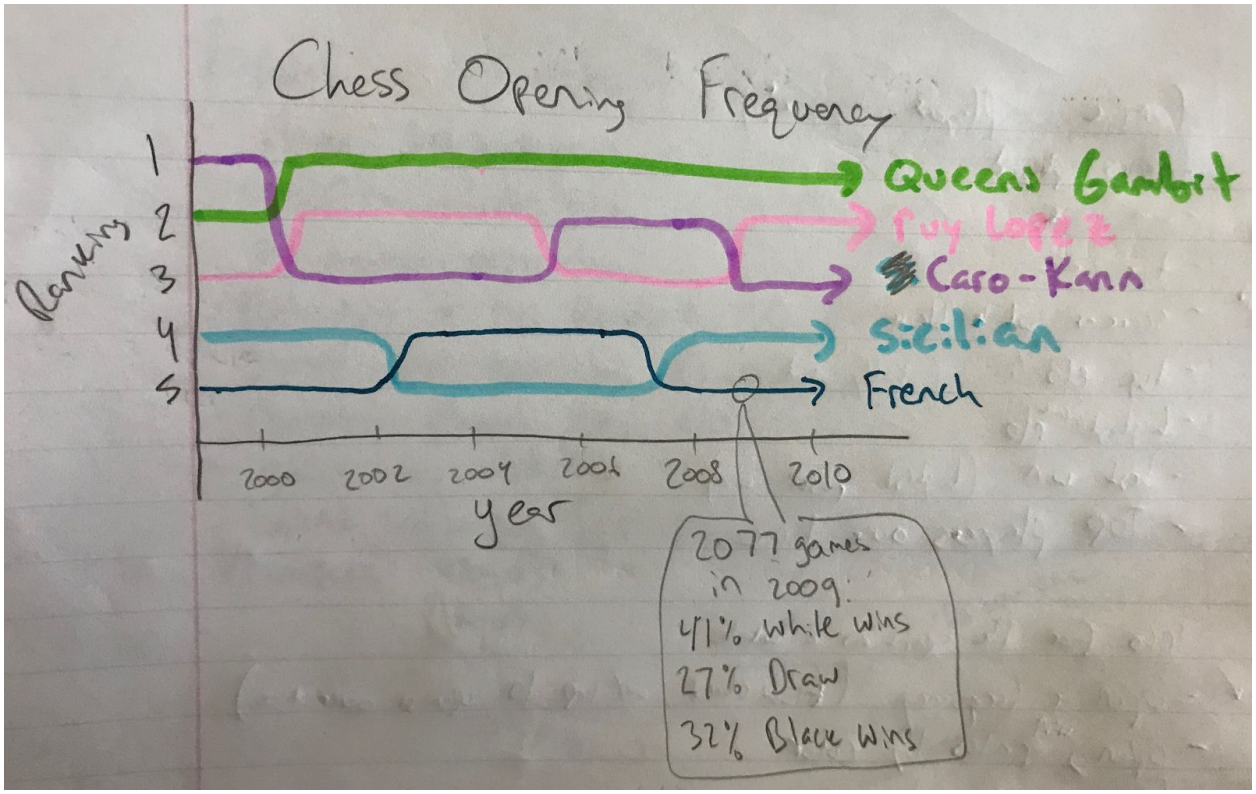
The final visualization at the bottom left shows a plot of the top chess players over time. There is a vertical, dashed line which can be manipulated to indicate the selected year. For the selected year, the top chess players for that year will be highlighted and their names will be displayed. This visualization allows users to see both the top players and player elo trends over time.


2 Player Comparison
Total Games
Wins
Best Win
Peak ELO
Record as White
Record as Black
Garry Kasparov
Years 1970-2010
Stat: Games
Value: 7952
Magnus Carlsen
Garry Kasparov

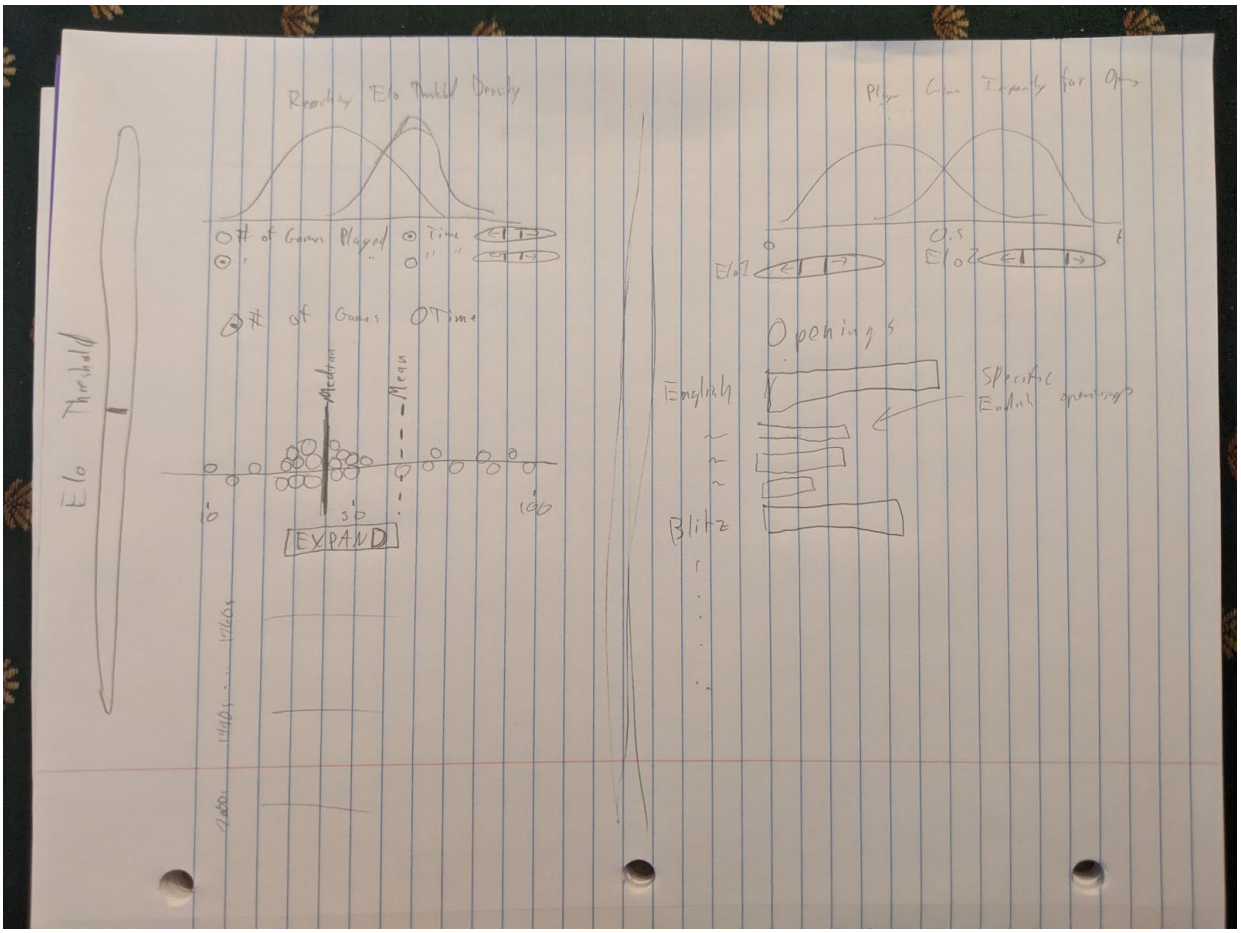

Player Comparison
(also possibly linked?)
Garry Kasparov
Peak ELO: 2800
Games Played: 9200
2850
2500
0
10000
Peak ELO
Win Rate
Games Played
Record as W
Record as B
Manual highlight
Games Played
Highlight Filters
X
Best Win
>
2700


Top Players Over Time
2007
2900
elo
2500
1950
2010
year
Magnus Carlsen: 2750
Fabiano Caruana: 2700
Vishy Anand: 2650
Hikaru Nakamura: 2625


Chess openings Results
opening
Games
White
Draw
Black
Queen's Gambit
10072
37%
32%
31%
Nimzo Indian
1065
32%
30%
38
Sicilian
3029
29%
41%
30
Alakhine
27
25%
39%
36
1940
1945
Time Range
2010
2015


Chess Opening Frequency
Ranking
1
2
3
4
5
Queens Gambit
Ruy Lopez
Caro-Kann
Sicilian
French
2000
2002
2004
2006
2008
2010
year
2077 games in 2009:
41% White Wins
27% Draw
32% Black Wins

Prototype II



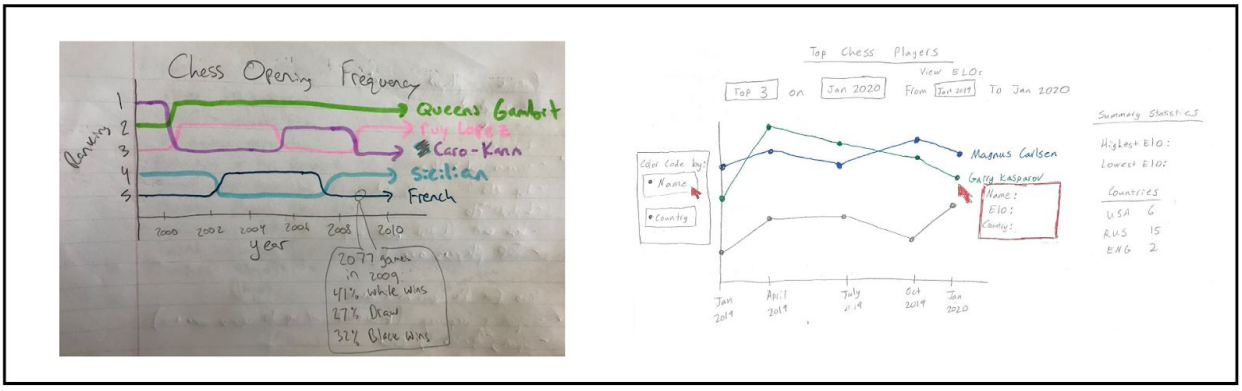
This prototype consists of 4 visualizations. Lets address each of these by their quadrants with quadrant I being located on the top right and quadrant IV being located at the bottom right. The quadrant II and quadrant III visualizations are related through an elo threshold seen on the far left. The quadrant II visualization is a density plot of either the number of games or the time it took to reach a certain elo threshold. The time and the dates are options for each density plot. There is also a filter for each density plot that limits the years you want to look at e.g. only look at the distribution from 1980 to 1990.

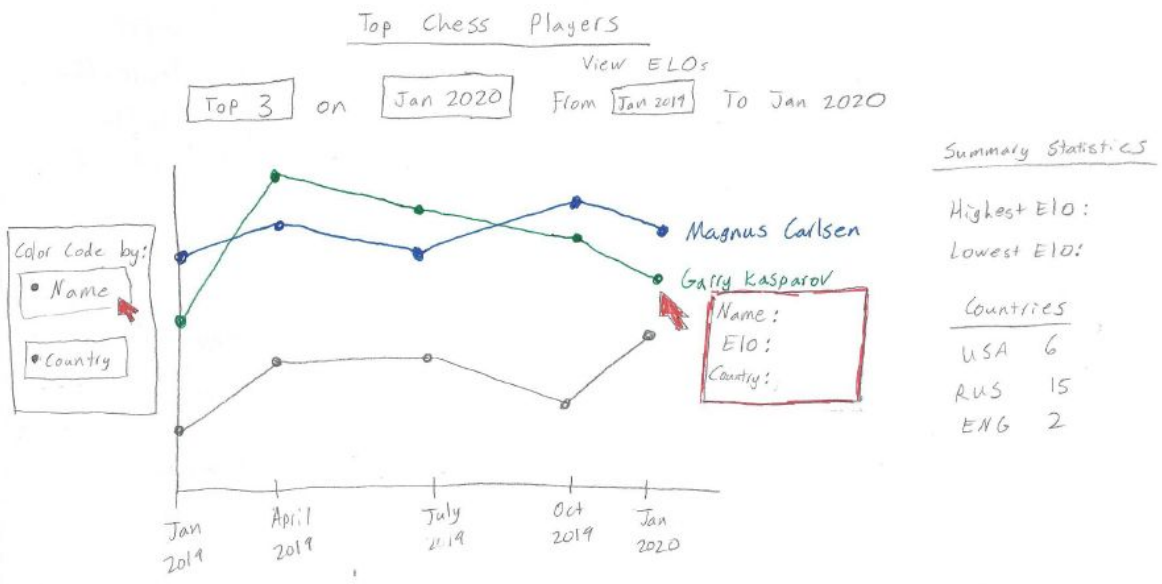
The quadrant III plot shows a beeswarm plot of how many games or how much time it takes to reach the elo threshold where each circle is a player. The beeswarm plot can then be expanded to see it grouped by the decade of their first game.

The quadrant I visualizes a density plot of the gini impurity of each player's opening moves In the most simplified terms, this shows if they used many different openings or few openings. More is done with this in the deep dive below.

The quadrant IV visualization is a bar chart that shows the number of opening moves given an opening move category (e.g. The English) then by clicking on the move a bar graph of each specific move would be shown.(e.g. Troeger defence).

Prototype III





This visualization focused on learning about the top players and openings. This provides a simple UI with information which the general chess player may be interested in. The player could choose how many of the top players to display for a given quarter. (Data source: https://ratings.fide.com/toplist.phtml). Then the range of dates can be chosen to see those players' ELO temporal progression. The top chess moves could be linked to the user's chosen dates and then insight could be gained from exploring what openings were popular at the times leading up to the players' top ranking.

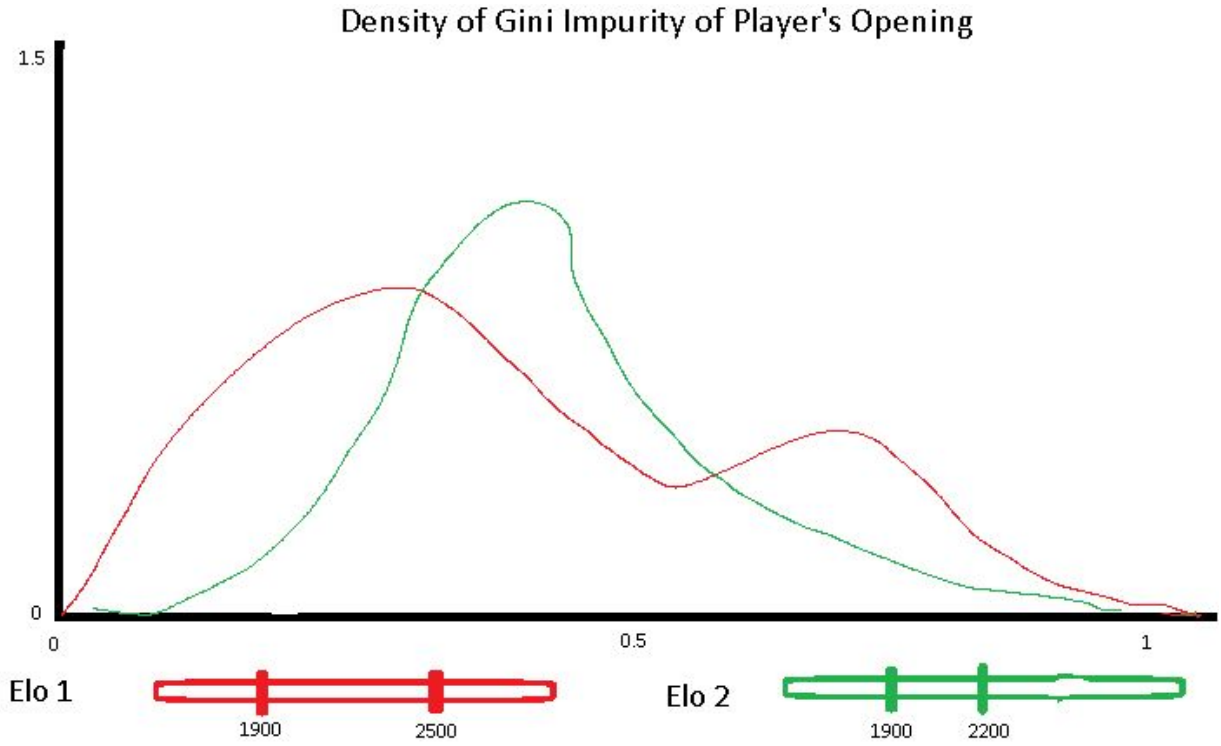
Summary statistics would be a nice addition to the right since it wouldn't clutter up the UI and it gives the opportunity to give a general summary not easily determined from the visualizations. The user could choose to have the lines color-encoded by name or by country. This would add an easy extra dimension to consider and explore alongside the ELO progression. On mouse-hover, the line could pop-out and display the player's name, exact ELO,

and country. We decided against this because we thought having a different dataset for different visualizations could be misleading to the user.

Final Design

Below are details about each visualization to be included in the final design and a layout of said designs.

Density of Gini Impurity of Player's Openings



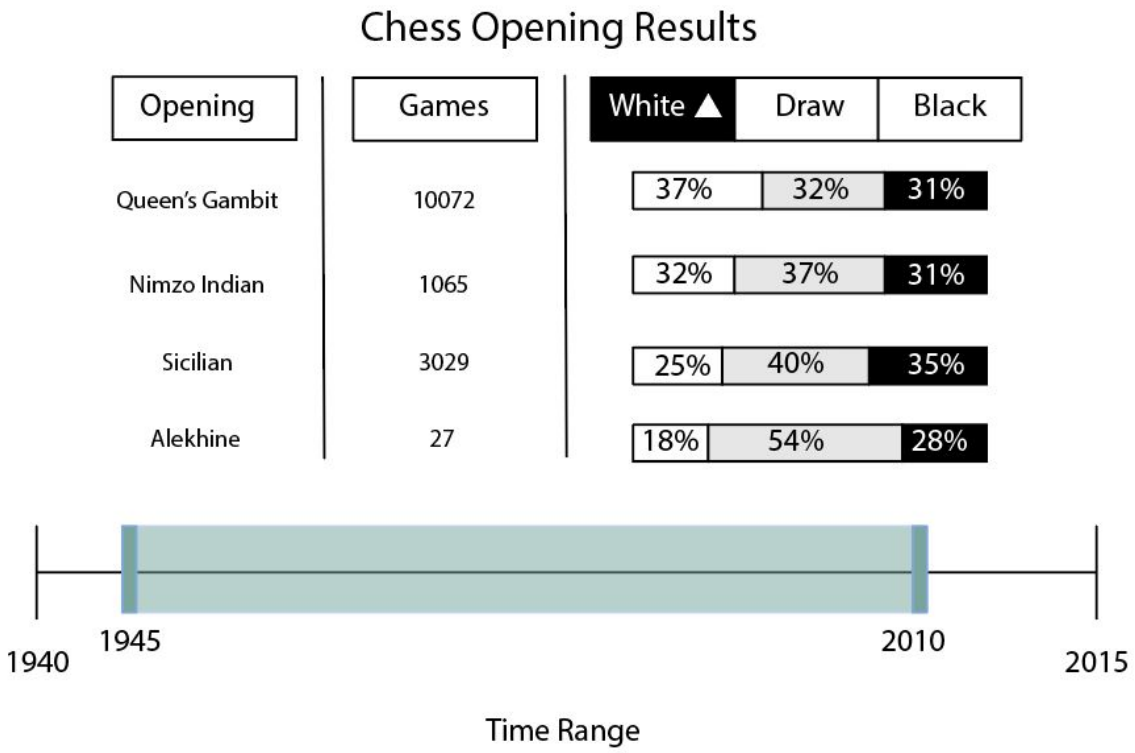
Description:

This at a density plot of the gini impurity of all players where their max elo was within a certain range. In oversimplified terms, one can think of the gini impurity as a measure of "mixedness" and by looking at the distribution of the opening moves it could suggest interesting trends. For example maybe the great players are specialists, i.e. they get really good at a specific opening, or maybe great players are generalists, i.e. they use a wide variety of openings depending on their opponent, or other related things.

- Must-have features
    - Two plots that show distributions
        - This could be either a density plot or a histogram
    - The ability to control the the range of their max elo for the distribution plots

        - Ideally this would be a slider with an upper and lower bound but could also be a “less than elo *x*” slider and “greater than elo *x*” slider.
- Optional Features
    - Kernel density options
    - Density plot
        - Instead of histogram
    - Double (lower and upper bound slider
        - instead of single lower or upper bound

Chess Opening Results



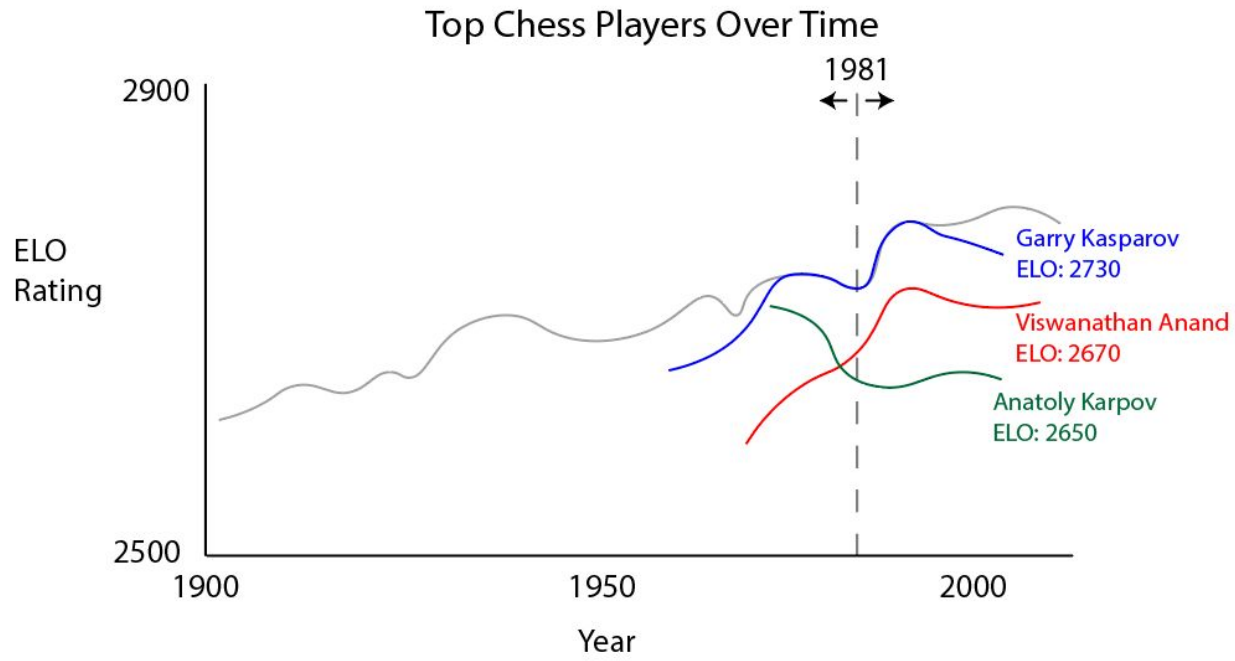
Description:

This table will show the number of games and the game result for each chess opening in the user chosen time range. The table will be sortable based upon any of its columns. This will allow the user to see how frequently the opening occurred in the games alongside the result of the games.

- Must-have features
    - Openings

    - Games
    - White, Draw, Black
    - Sorting
    - Choose Time Range
- Optional Features
    - Limiting the number of openings displayed to a custom value

Top Chess Players Over Time



Description:

This chart displays both the overall top chess ratings from every year that our data covers, as well as the career ELOs of the three highest rated players from that year. The dotted vertical line can be manipulated to change the current active year. This visualization allows users to clearly see the overall ELO trend over time, while also allowing them to get a sense for the best players over time and the various career trendlines that exist among top-tier players.

- Must-have features
    - Max elo per year line
    - Ability to manipulate current year
    - Player elo trend lines (depending on the year)
    - Labels for player name and elo in the selected year
- Optional Features
    - Toggleable number of best players (doesn't always have to be just 3)

- Additional filters on the players considered in the visualization

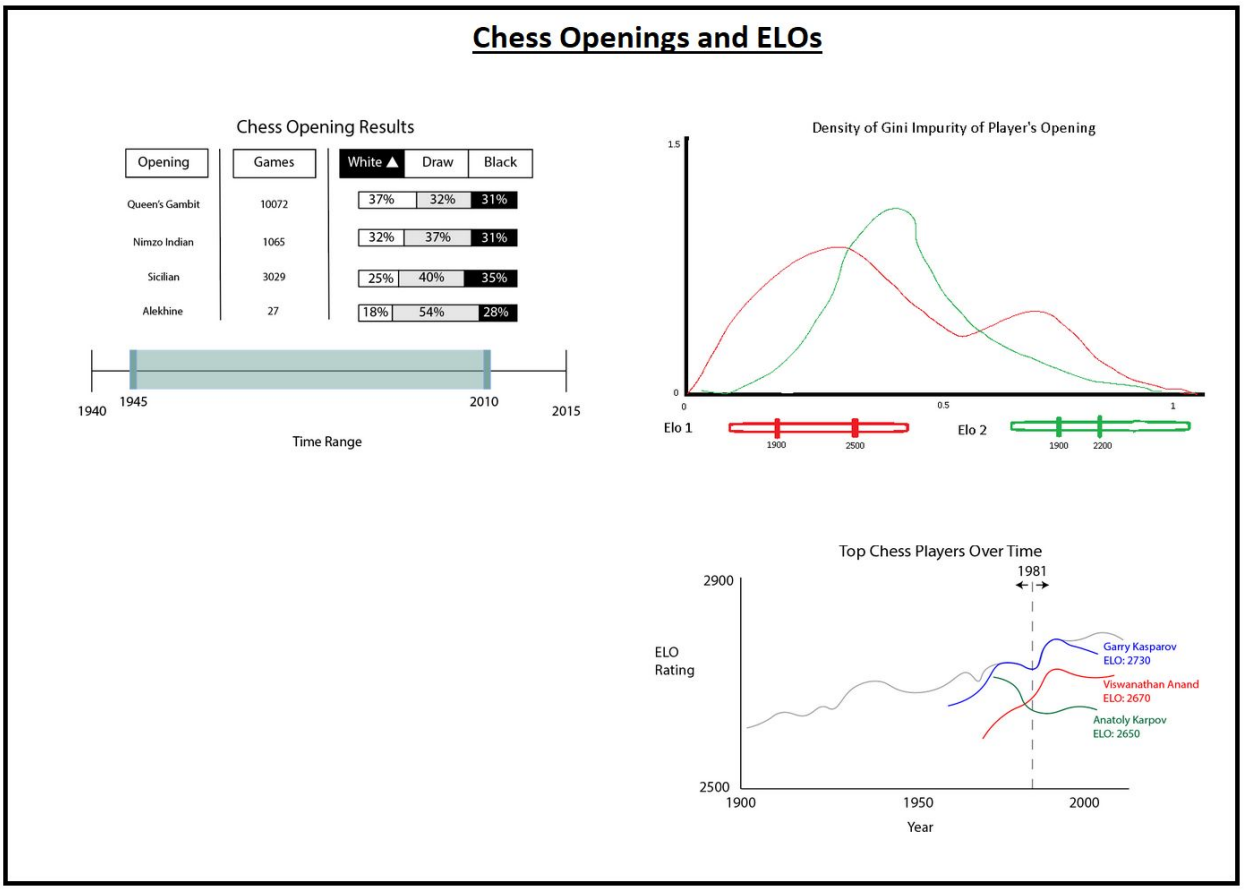
Must-Have features

- We must have the following visualizations as described above:
  - Chess Opening Results
  - Density of Gini Impurity of Player's Opening
  - Top Chess Players Over time

Optional Features

- Dark-mode-esque color theme
- Density plot of peak ELO vs how many games it took to reach that ELO

Final Design Layout



We decided on combining these three visualizations into our final design because we felt that these three provided the most interesting data to explore. The left column would contain all of

the chess openings for the chosen time range and the right half would contain The Gini Impurity and Top Chess Players visualizations. These three visualizations provide the user a great opportunity to analyze periods of time in chess and make connections between openings, gini impurity density, and the top chess players over time.

Project Schedule:

Nov 5: Project Peer Feedback

Nov 16-20: Project Review With Mentor

Dec 2: Final Project Due

Goal	Due Date
Set up structure of html (Id for left column, top right, and bottom right) (Alex and Stephen)	Nov 9
Data Pre-Processing (Garret)	Nov 9
Project Release	
Density of Gini Impurity for a fixed time range (No sliders) (Garret)	Nov 16
Top Chess Players for a single year (No slider) (Alex)	Nov 16
Chess Openings for a single year visualization with sorting (No slider) (Stephen)	Nov 16
Project Release	
Density of Gini Impurity with sliders (Garret)	Nov 23
Top Chess Players with slider (Alex)	Nov 23
Chess Openings with slider (Stephen)	Nov 23
Cleanup code and document code (All)	Nov 30
Project Release	
Project Submitted By (Due Dec 2)	Dec 1