



# Chess Openings

A statistical analysis

by Michael Cleversley





# TABLE OF CONTENTS



01

## Background

An overview of the study's topic

02

## Data Analysis

Data collection and manipulation

03

## Results

Confidence intervals, bootstrap, regression

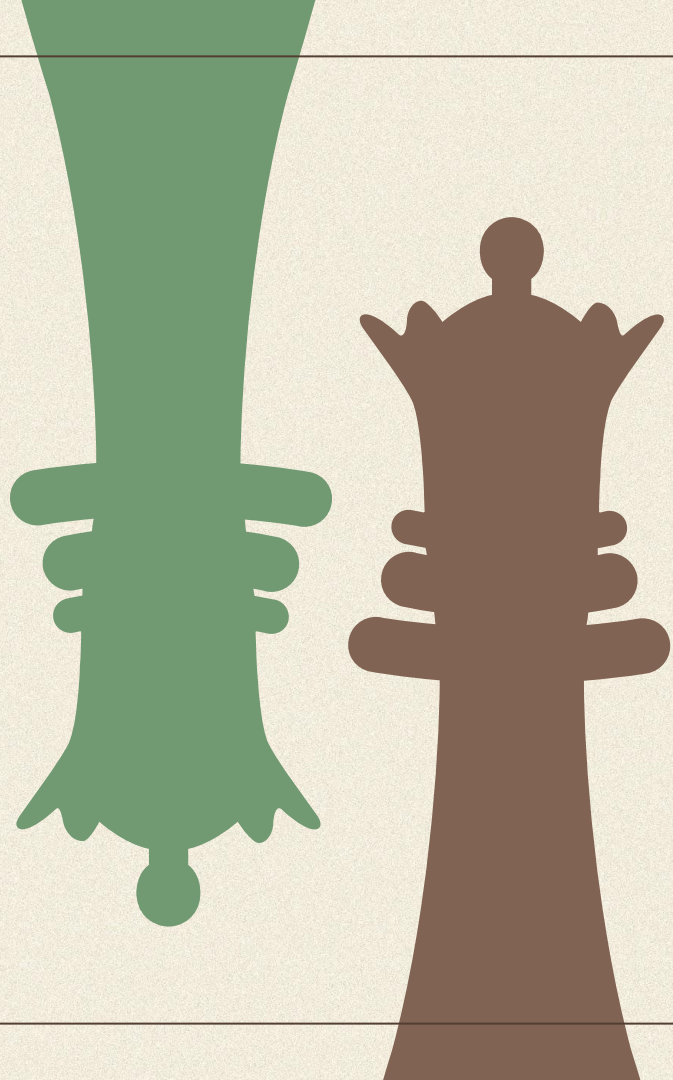
04

## Conclusions

Assumptions made, faults, extensions







# ♦ 01 ♦

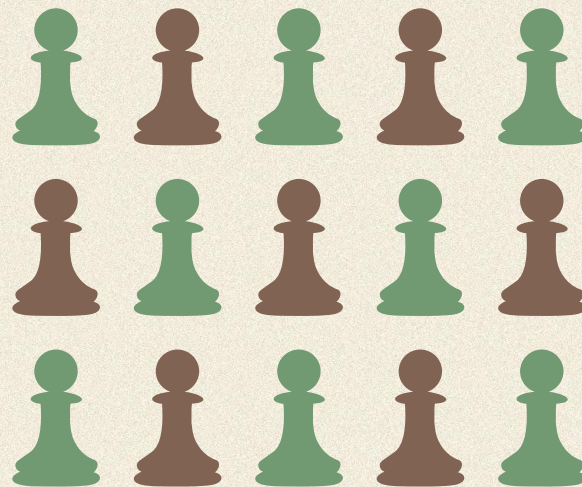
## BACKGROUND

What is chess, and what is a  
"chess opening"?



# CHESSE: A BRIEF OVERVIEW

- Chess: a very old, and very complex, boardgame
- Three phases of the game: **opening**, middlegame, and boardgame
- The opening sets the game's tone, which can take many different forms:
  - Positional
  - Tactical
  - Balanced/unbalanced
  - Drawish
  - Trappish
- Openings are defined by certain combinations of moves and have many different variations united under the theme and move order of that opening





# Opening Popularity by Move 1



**e4**

**d4**

**Nf3**

There are many, many more openings than these (and typical descriptions of openings extend beyond the first move), but these are by far the most popular.



**e4**

Often leads to sharp and quick games



**d4**

Lends to a more positional style

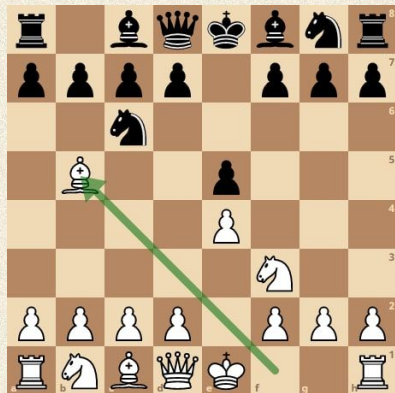


**Nf3**

Avoids the symmetrical structure of 1. e4 e5



# OPENING EXAMPLES



## Ruy Lopez

The most popular at the professional level



## Sicilian Defense

One of the most robust and most deeply studied opening



## Bongcloud Attack

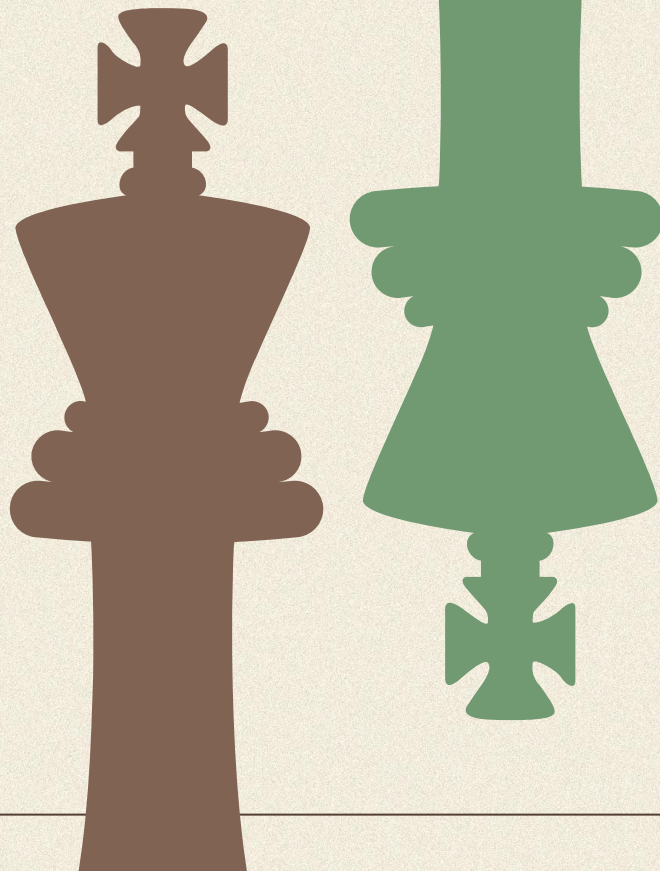
Sub-optimal, at least from a strategic standpoint



♦ 02 ♦

# DATA ANALYSIS

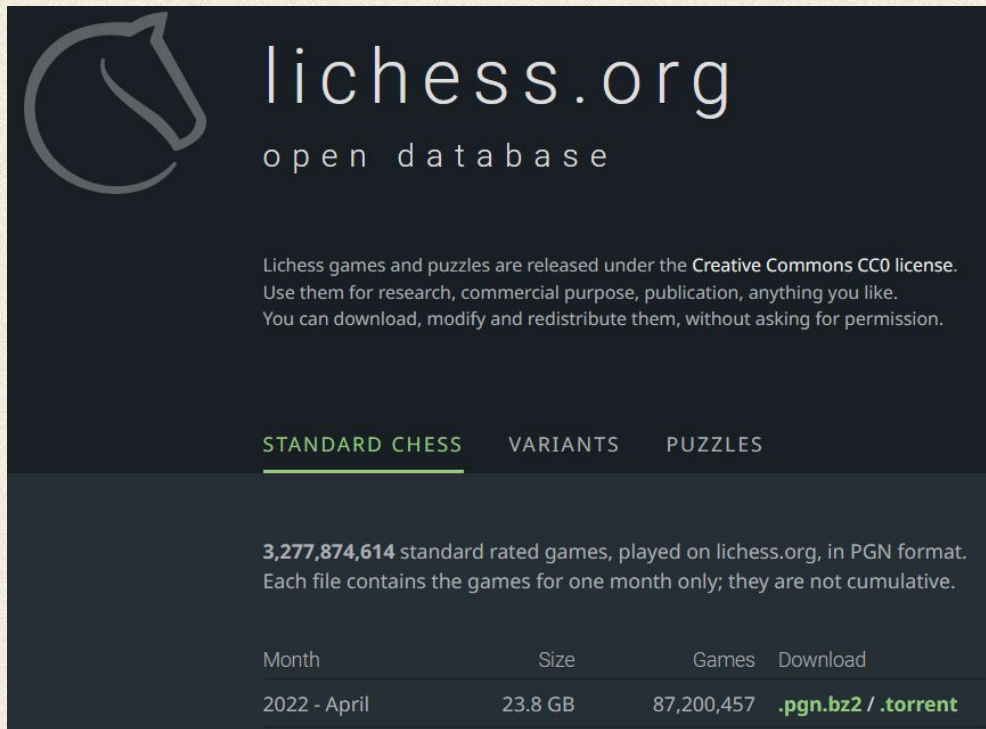
Where do I get chess data  
from, and how do I turn it  
into something useful?





# LICHESS.ORG

- lichess.org is an open-source website
- Has over 3 billion games in its database, with gigabytes of data generated each month
- Chose April 2018; ~5 GB



The screenshot shows the lichess.org website. At the top left is a horse head logo. The text 'lichess.org' is in a large, white, sans-serif font, with 'open database' in a smaller font below it. A paragraph of text states: 'Lichess games and puzzles are released under the Creative Commons CC0 license. Use them for research, commercial purpose, publication, anything you like. You can download, modify and redistribute them, without asking for permission.' Below this are three tabs: 'STANDARD CHESS' (highlighted with a green underline), 'VARIANTS', and 'PUZZLES'. A line of text reads: '3,277,874,614 standard rated games, played on lichess.org, in PGN format. Each file contains the games for one month only; they are not cumulative.' At the bottom is a table with four columns: 'Month', 'Size', 'Games', and 'Download'.

Month	Size	Games	Download
2022 - April	23.8 GB	87,200,457	<a href="#">.pgn.bz2</a> / <a href="#">.torrent</a>



# THE DATA: .pgn FILES

```
1 [Event "Rated Classical game"]
2 [Site "https://lichess.org/j1dkb5dw"]
3 [White "BFG9k"]
4 [Black "mamalak"]
5 [Result "1-0"]
6 [UTCDate "2012.12.31"]
7 [UTCTime "23:01:03"]
8 [WhiteElo "1639"]
9 [BlackElo "1403"]
10 [WhiteRatingDiff "+5"]
11 [BlackRatingDiff "-8"]
12 [ECO "C00"]
13 [Opening "French Defense: Normal Variation"]
14 [TimeControl "600+8"]
15 [Termination "Normal"]
16
17 1. e4 e6 2. d4 b6 3. a3 Bb7 4. Nc3 Nh6 5. Bxh6 gxf6 6. Be2 Qg5 7. Bg4 h5 8. Nf3 Qg6 9. Nh4 Qg5 10. Bxh5 Qxh4 11. Qf3 Kd8 12. Qxf7 Nc6 13. Qe8#
```

- Each game formatted in a text file with metadata and lists of moves
- Needed to parse through the file and extract the relevant data from the move list (notation)
- The .pgn file had millions of these games

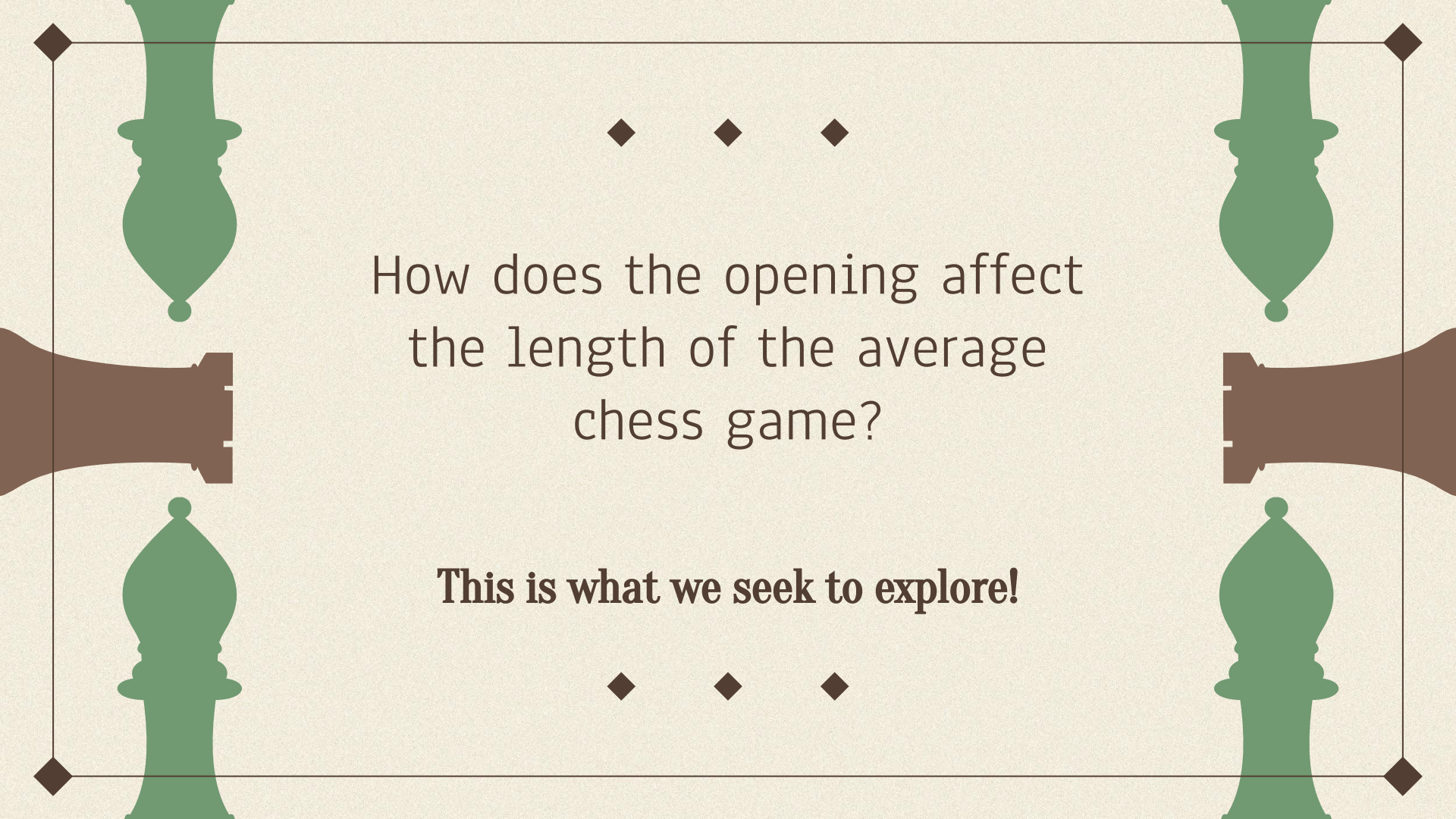


The background of the slide features a light beige chessboard with a dark brown border. Several chess pieces are visible: a dark brown king in the center, a dark brown rook on the left, a dark brown rook on the right, a green pawn at the top center, and a green queen at the bottom right. The word "DATA?" is written in a large, dark brown serif font on the left side of the board.

# DATA?

But what is *relevant* data? What are we looking for?



A decorative border surrounds the central text. It features stylized chess pieces: a green king at the top center, a brown rook on the left, a green queen at the bottom left, and a green king at the bottom right. Small dark brown diamonds are placed at the corners and along the top and bottom edges of the border.

How does the opening affect  
the length of the average  
chess game?

**This is what we seek to explore!**

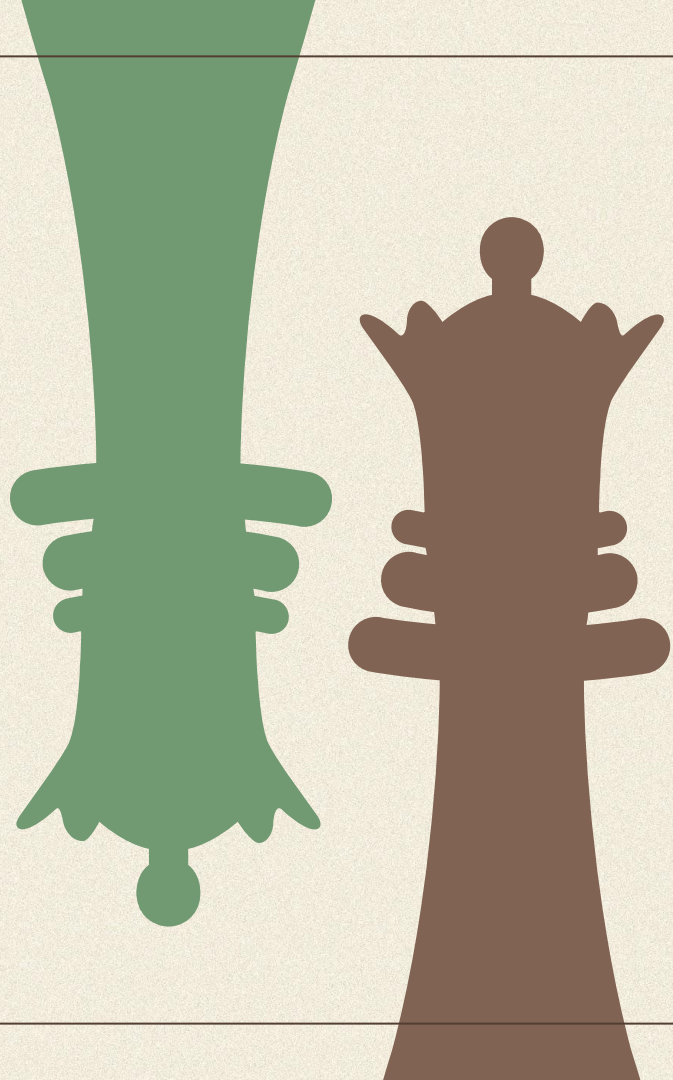


# RETURNING TO THE DATA

```
1 [Event "Rated Classical game"]
2 [Site "https://lichess.org/j1dkb5dw"]
3 [White "BFG9k"]
4 [Black "mamalak"]
5 [Result "1-0"]
6 [UTCDate "2012.12.31"]
7 [UTCTime "23:01:03"]
8 [WhiteElo "1639"]
9 [BlackElo "1403"]
10 [WhiteRatingDiff "+5"]
11 [BlackRatingDiff "-8"]
12 [ECO "C00"]
13 [Opening "French Defense: Normal Variation"]
14 [TimeControl "600+8"]
15 [Termination "Normal"]
16
17 1. e4 e6 2. d4 b6 3. a3 Bb7 4. Nc3 Nh6 5. Bxh6 gxf6 6. Be2 Qg5 7. Bg4 h5 8. Nf3 Qg6 9. Nh4 Qg5 10. Bxh5 Qxh4 11. Qf3 Kd8 12. Qxf7 Nc6 13. Qe8#
```

- Associated each opening with its average game length
- Narrowed down the scope of the openings, as each opening can have dozens of variations ("French Defense: Normal Variation" -> "French Defense")





# ◆ 03 ◆

## RESULTS


What were my findings?



The image features several stylized chess piece silhouettes in green and brown. A large green king is the central focus. To its left, a brown queen and a green rook are partially visible. Below the king, a brown knight and a brown rook are shown. The background is a light beige with a subtle grid pattern. The entire composition is framed by a thin brown border with diamond-shaped corner accents.

# 33.06

Average number of moves in April of 2018

Four small brown diamond shapes are arranged horizontally below the text.



The background features a light beige, textured surface. On the left side, there are stylized silhouettes of chess pieces: a green king in the center, a brown pawn above it, a green rook to its left, a brown queen to its left, and a brown rook below it. On the right side, there is a large brown number '3.36' and the text 'Standard deviation' below it. At the bottom right, there are four small brown diamonds arranged horizontally. The entire composition is framed by a thin brown border with diamond-shaped corner accents.

# 3.36

Standard deviation





# What openings had the smallest and greatest average number of moves?



**GREATEST:**

**Marienbad System**

A very rare and suboptimal  
sideline of the Queen's Indian  
Defense



**SMALLEST:**

**King's Pawn Opening**

The most popular "opening" -  
but only defined by one move



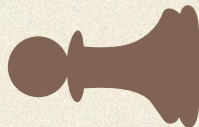
# The Stats

Marienbad System

King's Pawn

Average # moves

38.67



7.23

# of samples

83



24,647

95% C.I.

[35.41, 42.14]



[7.06, 7.40]



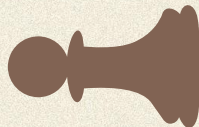


# The Stats cont.

Marienbad System

.4983

p-value



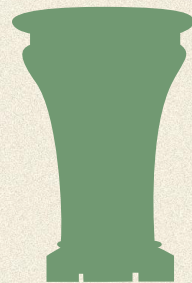
King's Pawn

.4896

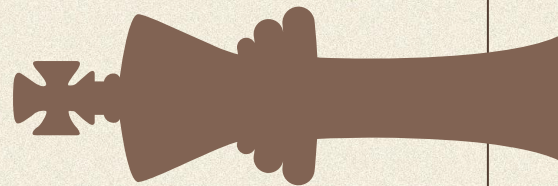
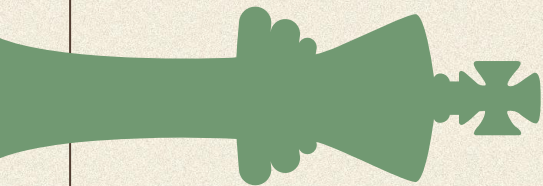




SOMETHING'S



WRONG.









## ◆ What went wrong? Why is something wrong? ◆

Two things from the stats should have stood out: the extremely low average for the King's Pawn, and the number of samples for the maximum

A low sample size for the Marienbad System (83 games vs 24,647 King's Pawn games) means the results could be extremely variable.

So what did happen? And how can we fix it?





◆ ◆ ◆

People give up.

◆ ◆ ◆

A close-up photograph of a chess game. A hand is shown dropping a white king piece onto a black king piece that is already on the board. The black king is tilted over, suggesting it has been defeated. The chessboard is visible at the bottom, and the background is dark. The text "People give up." is overlaid on the left side of the image, flanked by diamond symbols.



◆ ◆ ◆  
Or, they don't even try.  
◆ ◆ ◆





# FIXING OUR ASSUMPTIONS

In thousands of games, White played 1. e4 and the opponent never played a move. The database still logs and categorizes it.

How do we fix this? Ignore abandoned games.

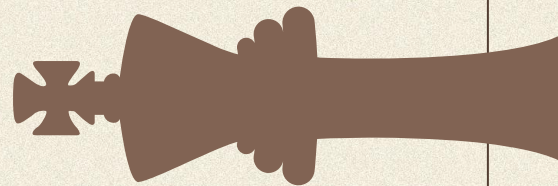
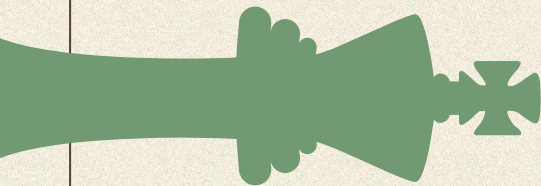
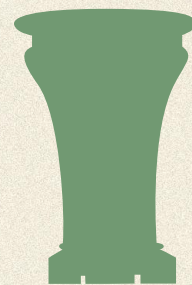
How do we fix the small sample sizes? Stipulate a minimum game count (we'll try 1,000).







◆ ◆ ◆  
Let's try  
again!  
◆ ◆ ◆






The background features several large, stylized silhouettes of chess pieces. A green king is the central figure. To its left, a brown queen and a brown rook are partially visible. To its right, a brown rook is shown. The entire scene is framed by a thin brown border with diamond-shaped corner ornaments.

# 33.16

Average number of moves in April of 2018


Four small, solid brown diamond shapes are arranged horizontally below the text.



The background features a light beige, textured surface. On the left side, there are stylized silhouettes of chess pieces: a green king in the center, a brown pawn above it, a green rook to its left, a brown knight below the king, and a brown rook to the right of the king. A thin, dark brown rectangular border with diamond-shaped corner accents frames the entire image.

# 2.86

Standard deviation

Four small, solid brown diamonds are arranged horizontally in a row, positioned below the text 'Standard deviation'.



# What openings had the smallest and greatest average number of moves?



## GREATEST: King's Indian

A hypermodern aggressive opening played at the highest levels; leads to dynamic play



## SMALLEST: King's Pawn Opening

Still the smallest average!



# The Stats

King's Indian

King's Pawn

Average # moves

38.43

21.70

# of samples

1686

7415

90% C.I.

[37.74, 39.13]

[21.28, 22.10]



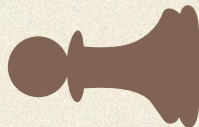
# The Stats cont.

Marienbad System

0.507

Fail to reject the null  
hypothesis!

p-value



King's Pawn

0.490

Fail to reject the null  
hypothesis!

Neither are particularly  
statistically significant!



## Simple linear regression: predicting the opening based on the first 5 moves

```
Actual is Ruy Lopez, index of 34  
Predicted is Queen's Pawn Opening, index of 30  
Actual is King's Gambit Accepted, index of 2  
Predicted is Queen's Pawn Opening, index of 30  
Actual is Scandinavian Defense, index of 4  
Predicted is Benko Gambit Accepted, index of 29  
Actual is Italian Game, index of 21  
Predicted is Queen's Pawn Opening, index of 30  
Actual is King's Indian Defense, index of 35  
Predicted is Modern Defense, index of 27
```

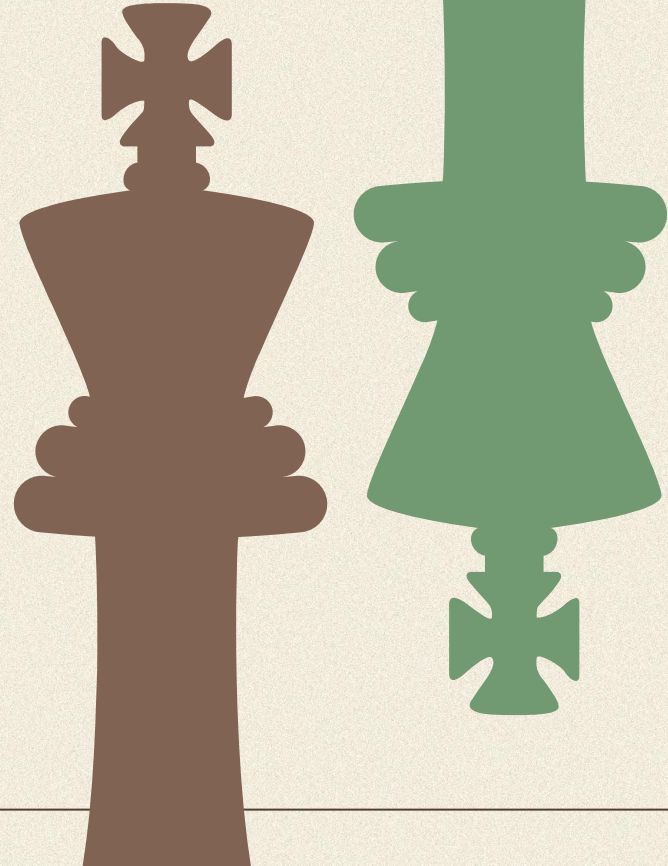
(it did not perform well - why?)



# ◊ 04 ◊

## CONCLUSION

What assumptions did I make? What would I change? How could I extend the project?





◆ ◆ ◆

Data analysis is hard.

◆ ◆ ◆

A close-up photograph of a chess game in progress. A hand is shown dropping a white king piece onto a black king piece that is already on the board. The pieces are on a checkered chessboard. The background is dark, and the lighting highlights the pieces and the hand. The text "Data analysis is hard." is overlaid on the left side of the image, flanked by diamond symbols.



# ASSUMPTIONS AND CHANGES

The dataset was accurate and unbiased - not created by machines.

Moves are representable as numbers for the sake of a linear regression model.

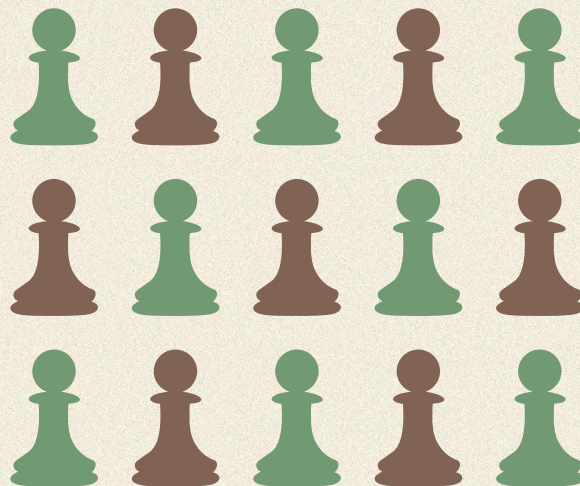
Changes: more time spent on regression, use larger dataset, incorporate machine evaluation





# PROJECT EXTENSIONS

- Larger datasets
- More robust ML models with more carefully curated data
- What else has an effect on the length of the game?
- How does the opening affect the average **engine evaluation**?





# THANKS

## DO YOU HAVE ANY QUESTIONS?



**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**

Please keep this slide for attribution