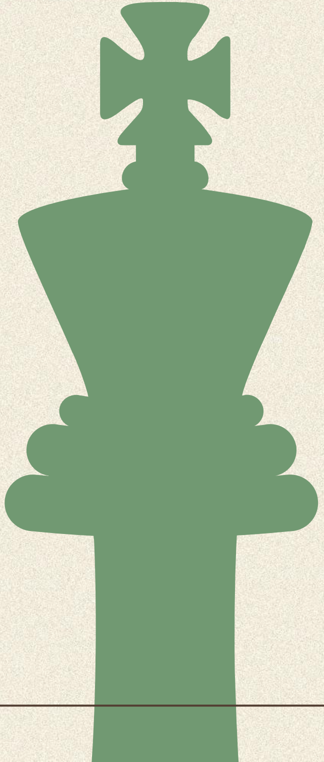
# Chess Openings

A statistical analysis

by Michael Cleversley



# TABLE OF CONTENTS

## 01 Background

An overview of the study’s topic

## 02 Data Analysis

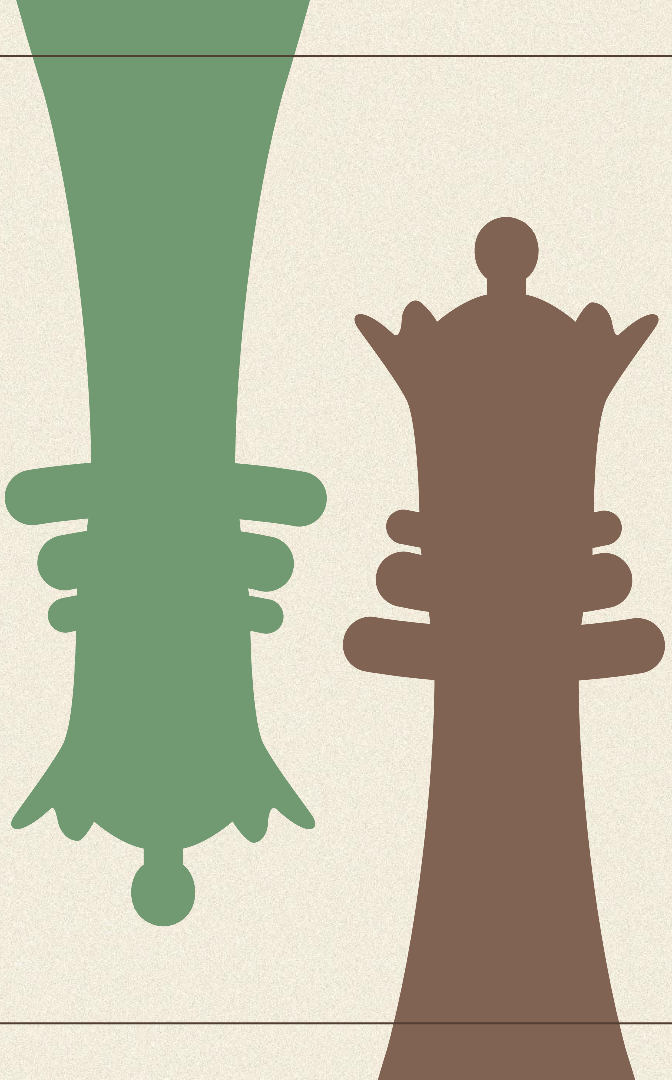
Data collection and manipulation

## 03 Results

Confidence intervals, bootstrap, regression

## 04 Conclusions

Assumptions made, faults, extensions



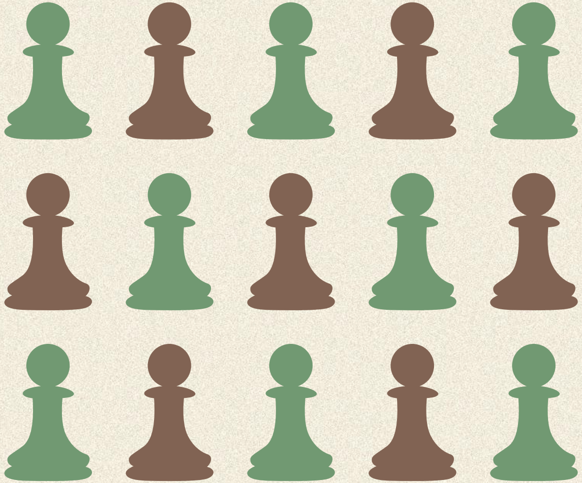
# 01

# BACKGROUND

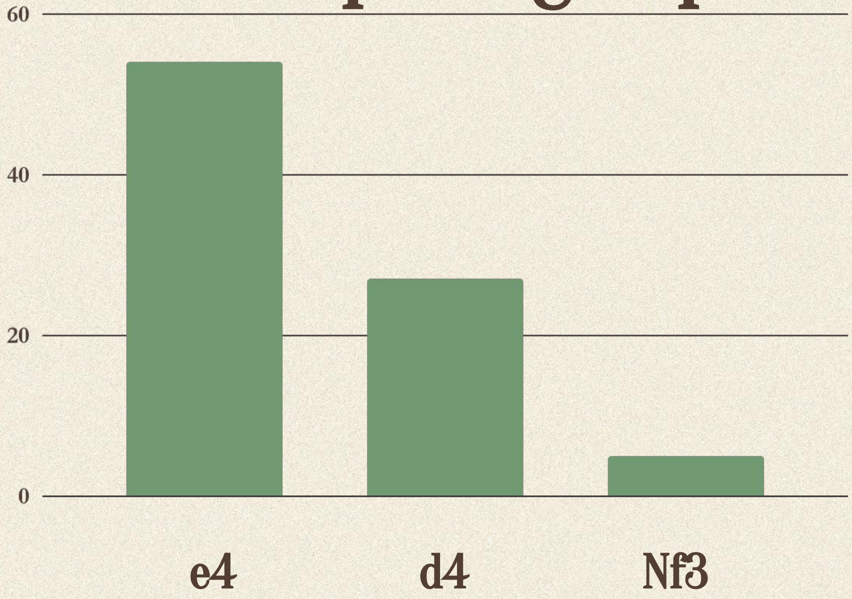
What is chess, and what is a “chess opening”?

# CHESS: 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



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

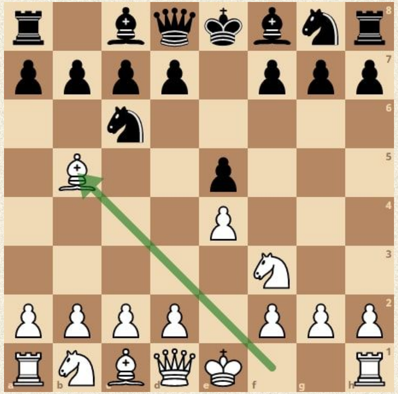
## d4

Lends to a more positional style

## Nf3

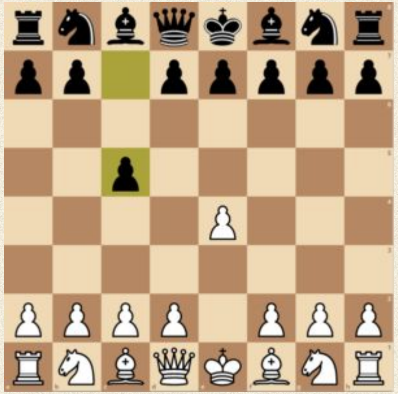
Avoids the symmetrical structure of 1. e4 e5

# OPENING EXAMPLES



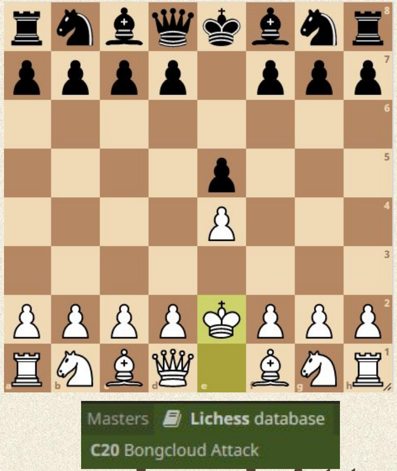
## Ruy Lopez

The most popular at the professional level



## Sicilian Defense

One of the most robust and most deeply studied opening



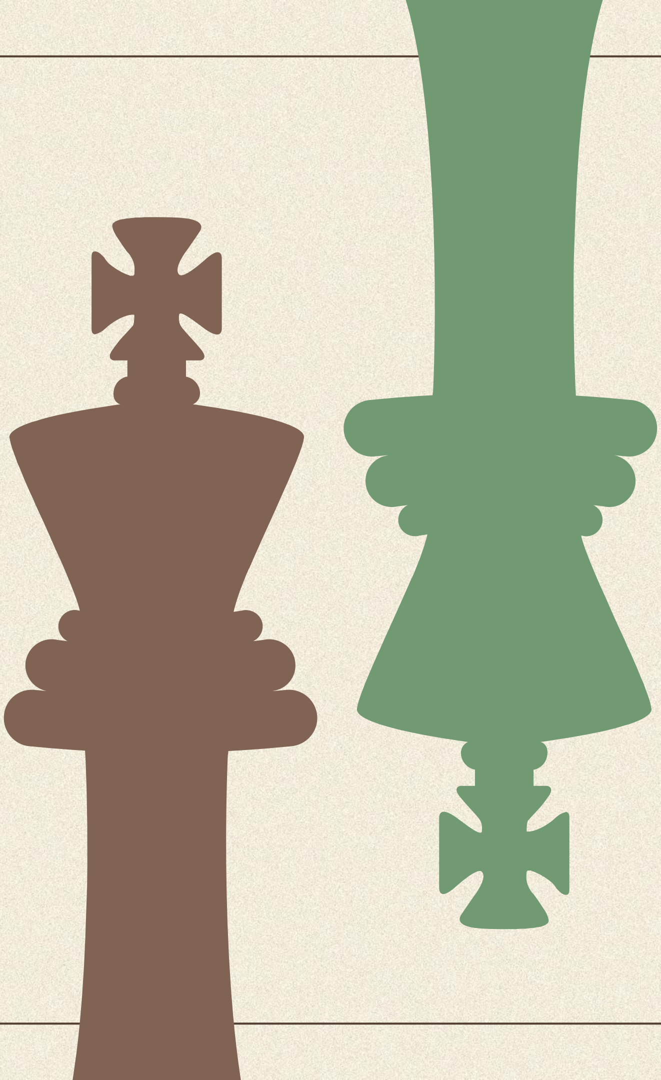
## Bongcloud Attack

Sub-optimal, at least from a strategic standpoint

# 02

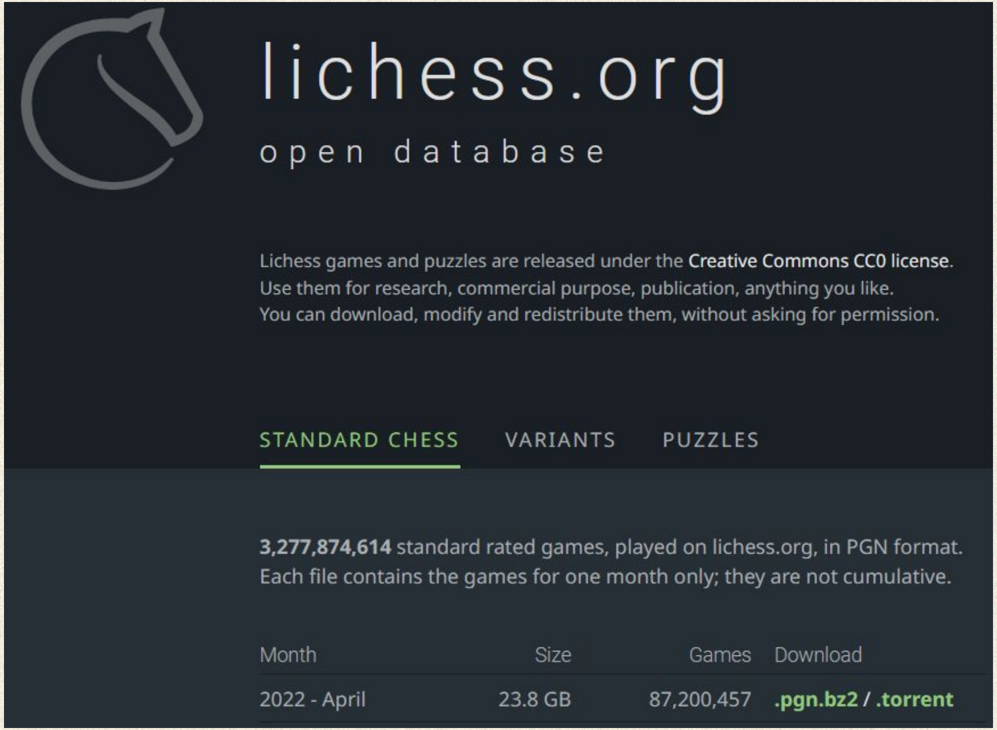
# DATA ANALYSIS

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



lichess.org

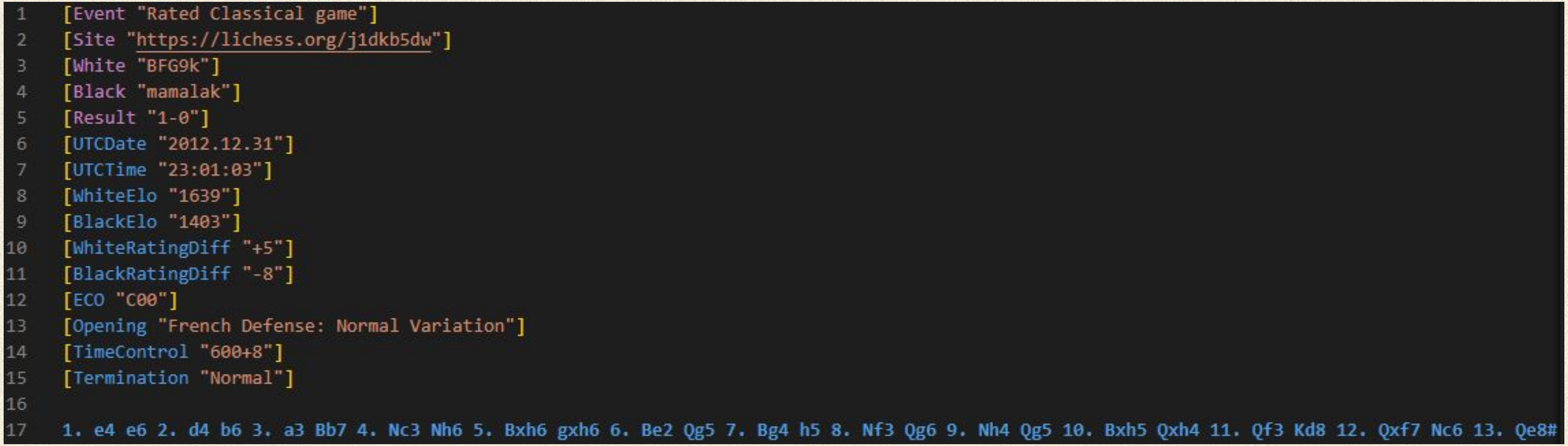
open database

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.

**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.

| Month | Size | Games | Download |
|---|---|---|---|
| 2022 - April | 23.8 GB | 87,200,457 | **.pgn.bz2 / .torrent** |

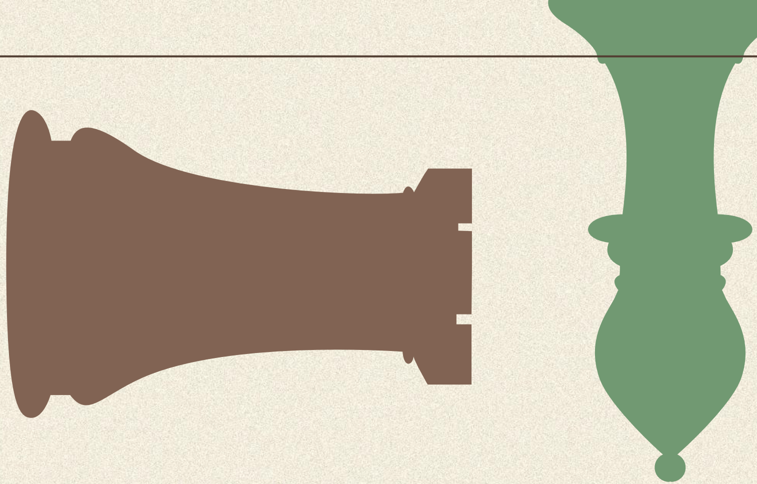
# THE DATA: .pgn FILES

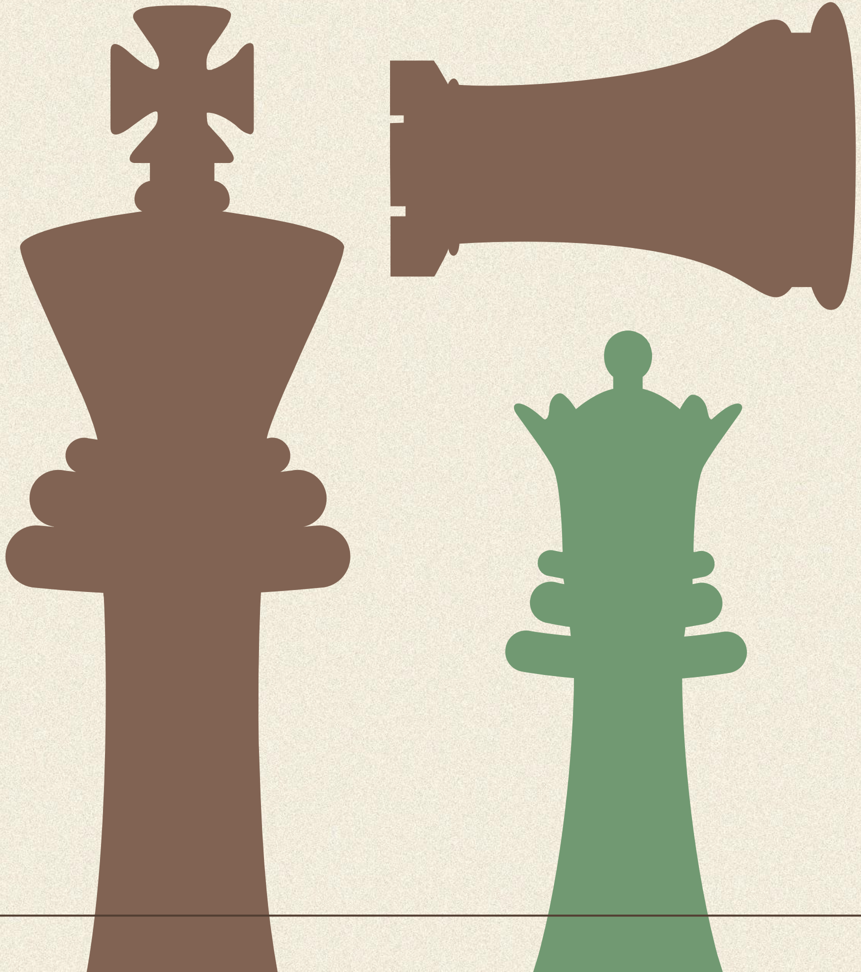


```
[Event "Rated Classical game"]
[Site "https://lichess.org/j1dkb5dw"]
[White "BFG9k"]
[Black "mamalak"]
[Result "1-0"]
[UTCDate "2012.12.31"]
[UTCTime "23:01:03"]
[WhiteElo "1639"]
[BlackElo "1403"]
[WhiteRatingDiff "+5"]
[BlackRatingDiff "-8"]
[ECO "C00"]
[Opening "French Defense: Normal Variation"]
[TimeControl "600+8"]
[Termination "Normal"]

1. e4 e6 2. d4 b6 3. a3 Bb7 4. Nc3 Nh6 5. Bxh6 gxh6 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





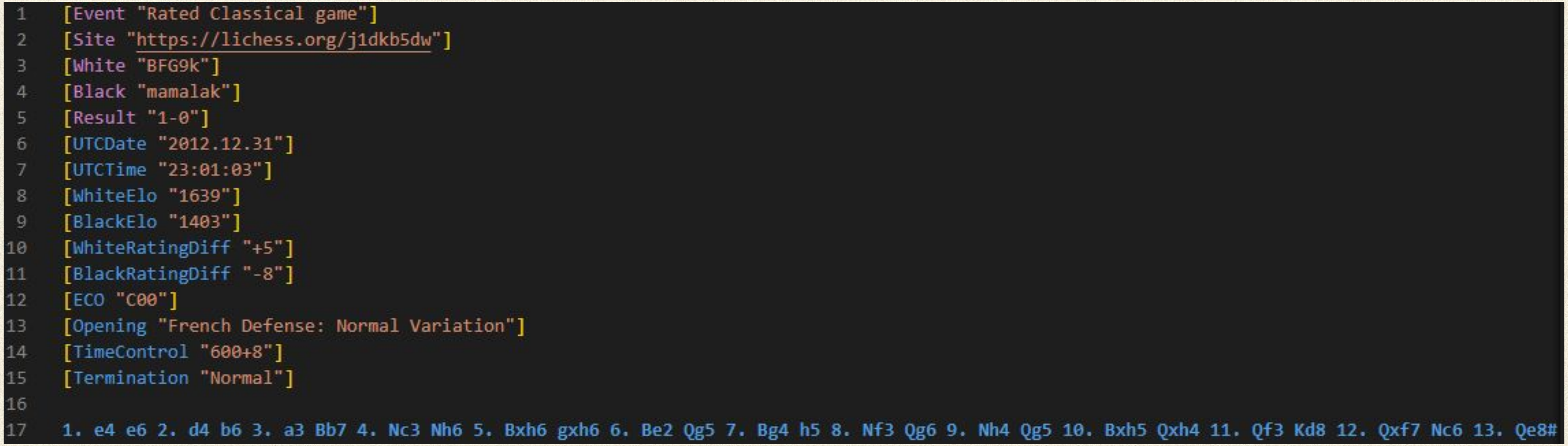
# DATA?

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

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

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

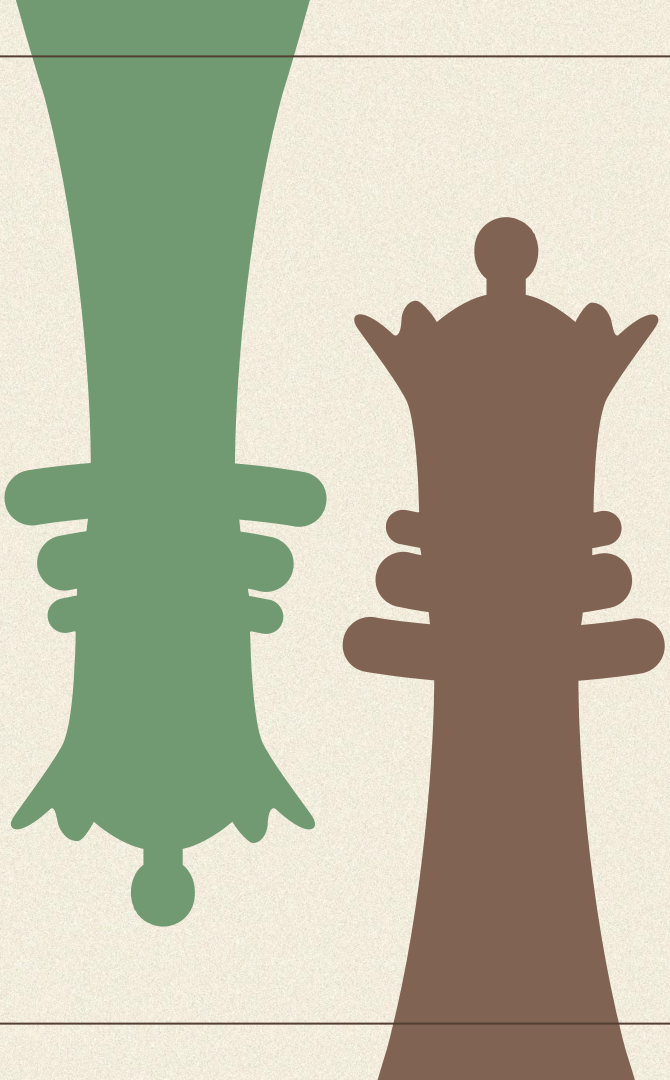
# RETURNING TO THE DATA



```
[Event "Rated Classical game"]
[Site "https://lichess.org/j1dkb5dw"]
[White "BFG9k"]
[Black "mamalak"]
[Result "1-0"]
[UTCDate "2012.12.31"]
[UTCTime "23:01:03"]
[WhiteElo "1639"]
[BlackElo "1403"]
[WhiteRatingDiff "+5"]
[BlackRatingDiff "-8"]
[ECO "C00"]
[Opening "French Defense: Normal Variation"]
[TimeControl "600+8"]
[Termination "Normal"]

1. e4 e6 2. d4 b6 3. a3 Bb7 4. Nc3 Nh6 5. Bxh6 gxh6 6. Be2 Qg5 7. Bg4 h5 8. Nf3 Qg6 9. Nh4 Qg5 10. Bxh5 Qxh4 11. Qf3 Kd8 12. Qxf7 Nc6 13. Qe8#
```

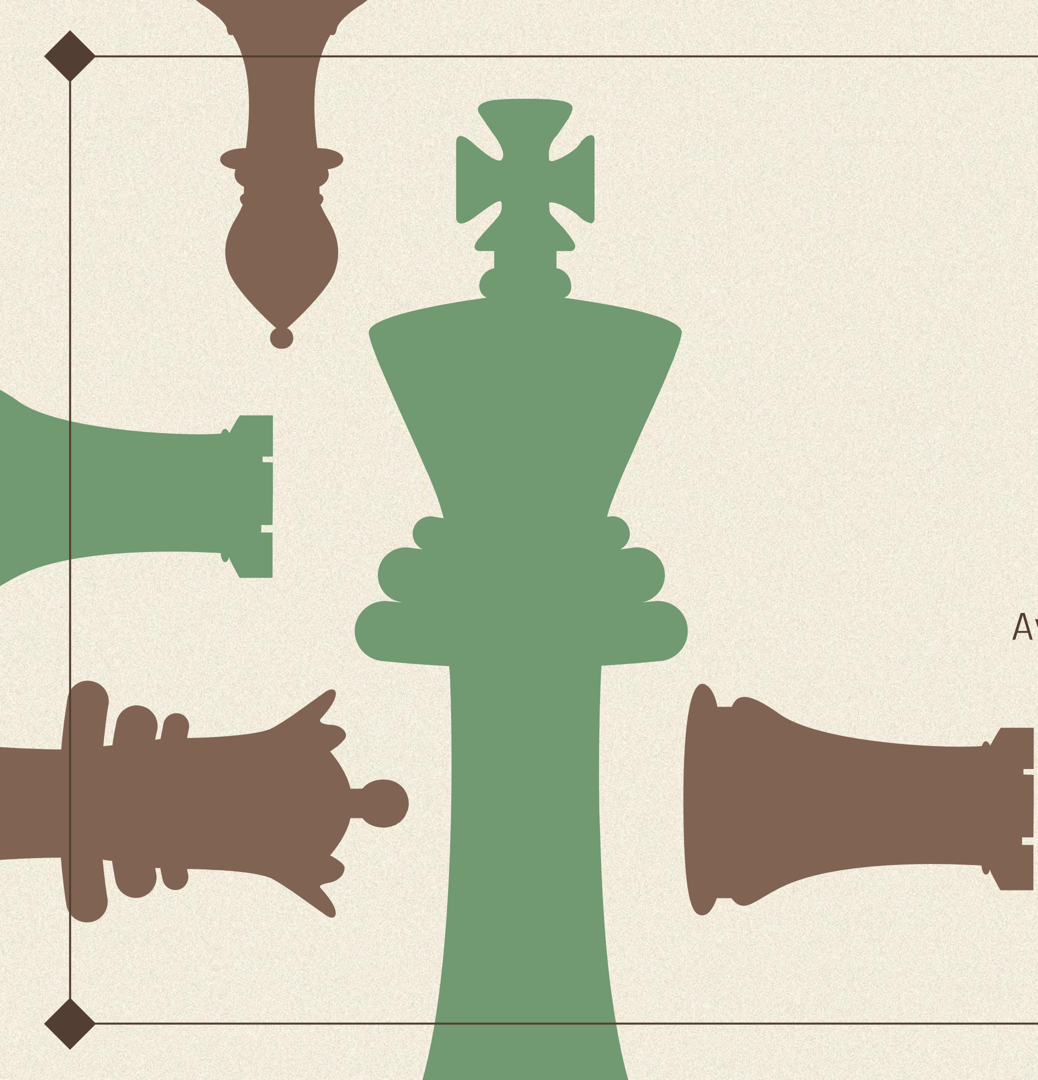
- Associated each opening with its average game length
- Narrowed down the scope of the openings, as each opening can have dozens of variations (“French Defense: Normal Variation” -> “French Defense”)



# 03

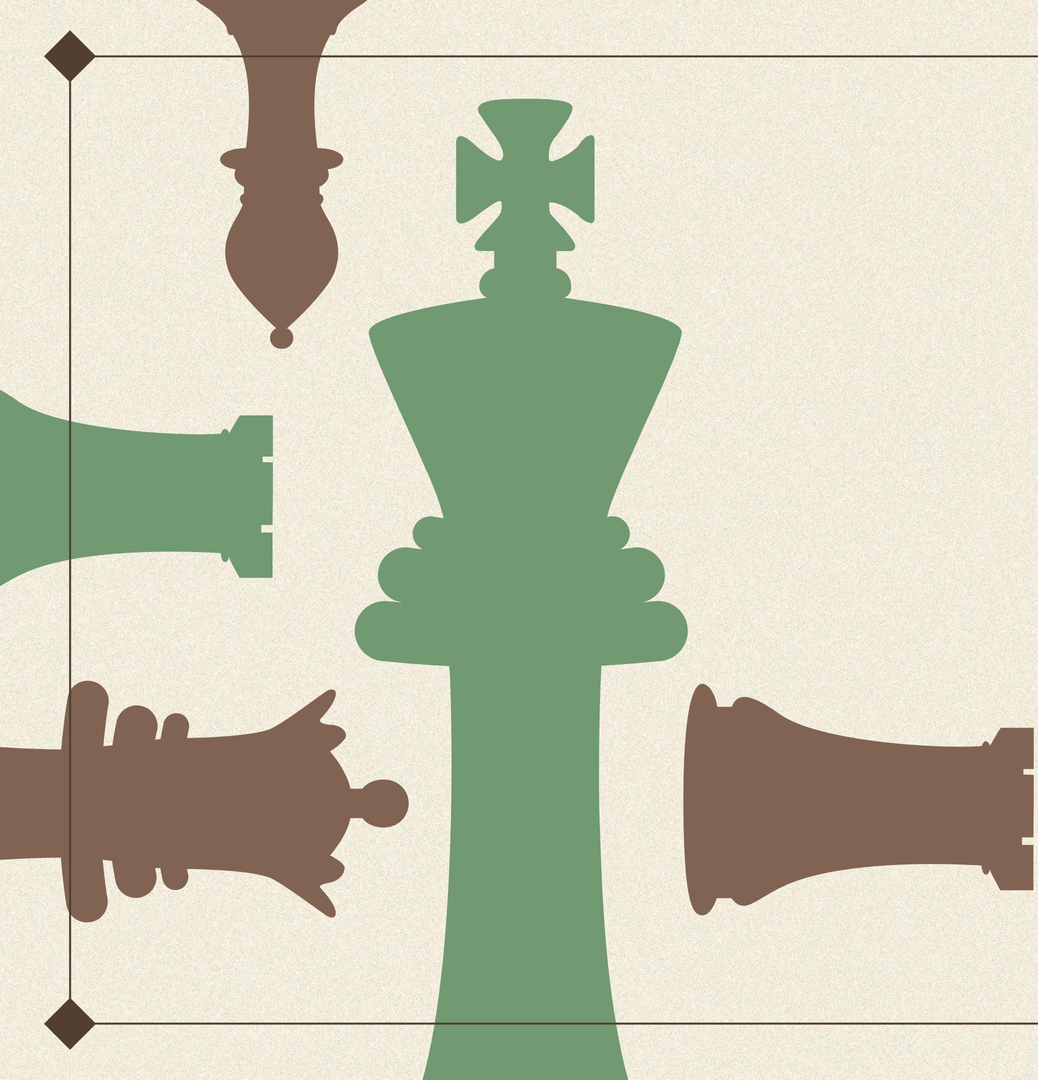
# RESULTS

What were my findings?



# 33.06

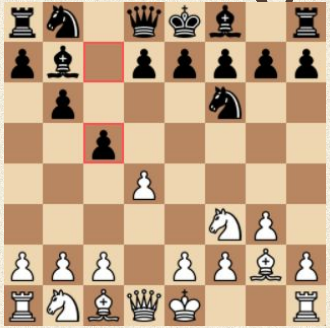
Average number of moves in April of 2018



# 3.36

Standard deviation

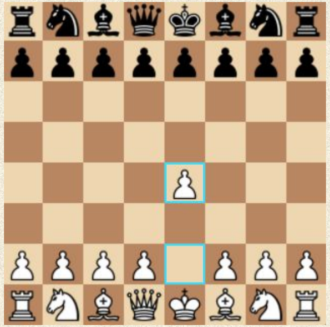
# What openings had the smallest and greatest average number of moves?



## 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 |
|---|---|---|
| 38.67 | Average # moves | 7.23 |
| 83 | # of samples | 24,647 |
| [35.41, 42.14] | 95% C.I. | [7.06, 7.40] |

# The Stats cont.

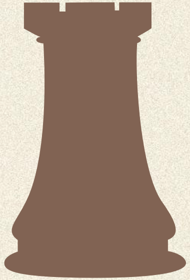
Marienbad System

King's Pawn

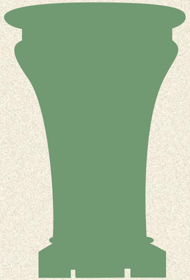
p-value

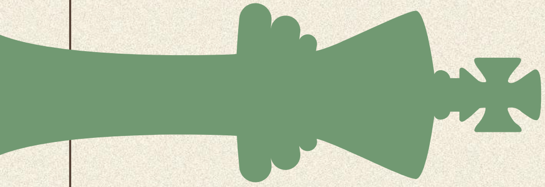
.4983

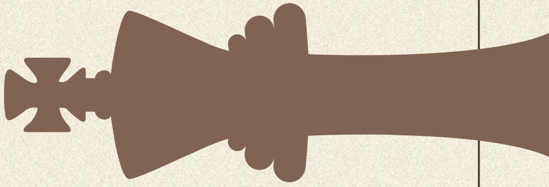
.4896



# 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.

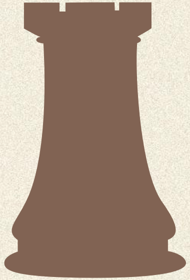
## Or, they don’t even try.

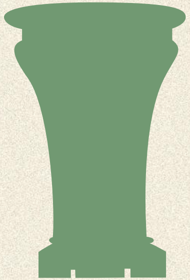
# FIXING OUR ASSUMPTIONS

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

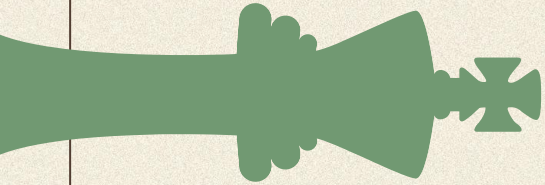
How do we fix this? Ignore abandoned games.

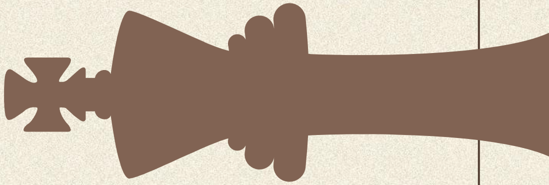
How do we fix the small sample sizes? Stipulate a minimum game count (we'll try 1,000).

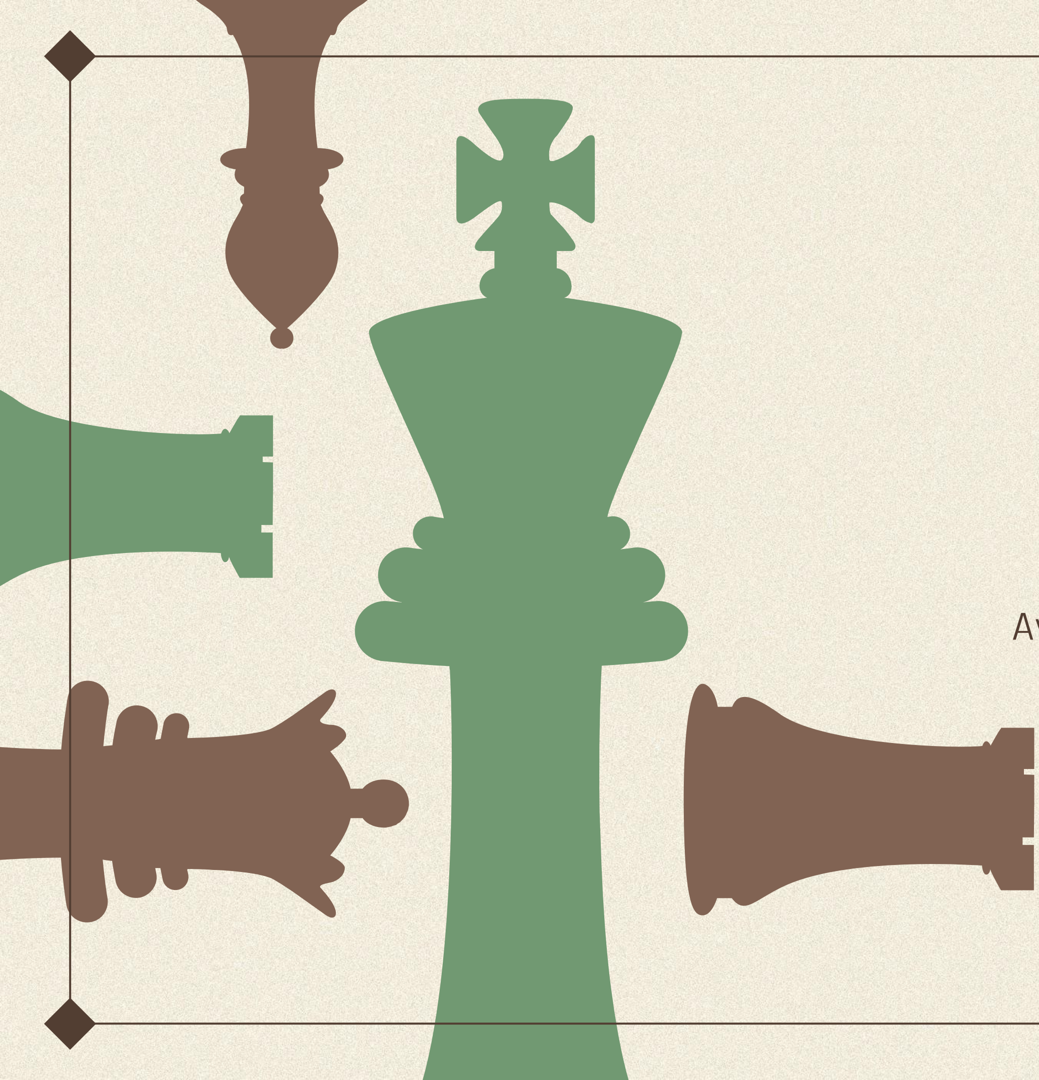




# Let’s try again!

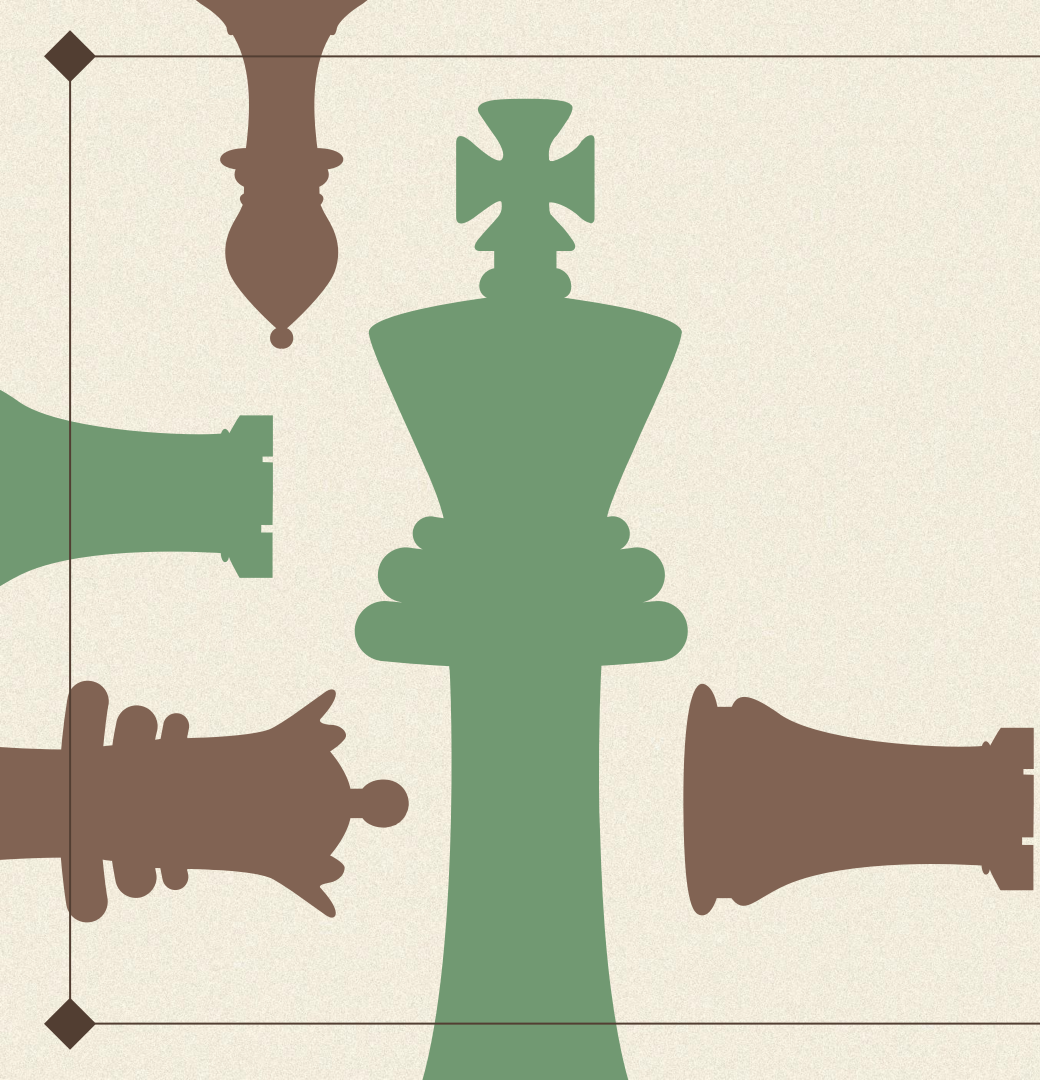






# 33.16

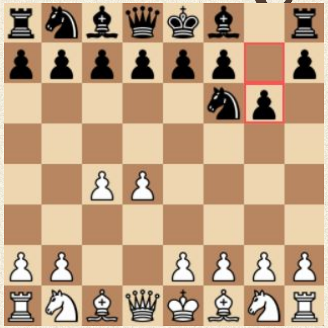
Average number of moves in April of 2018



# 2.86

Standard deviation

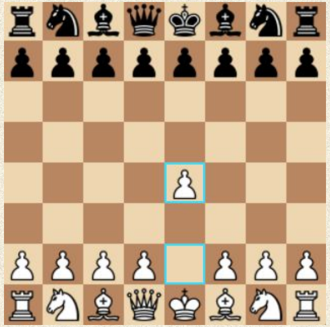
# What openings had the smallest and greatest average number of moves?



## 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 |
|---|---|---|
| 38.43 | Average # moves | 21.70 |
| 1686 | # of samples | 7415 |
| [37.74, 39.13] | 90% C.I. | [21.28, 22.10] |

# The Stats cont.

## Marienbad System

0.507

Fail to reject the null hypothesis!

p-value

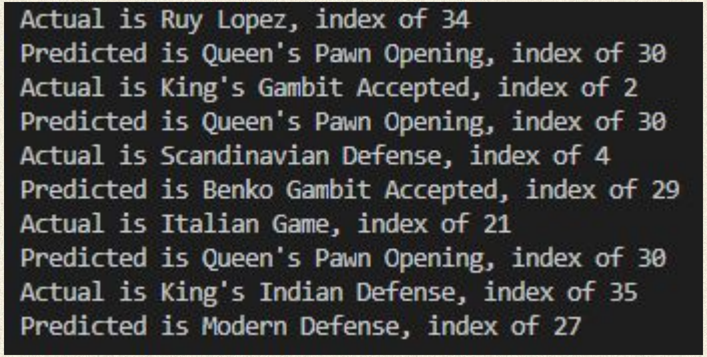
## King’s Pawn

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



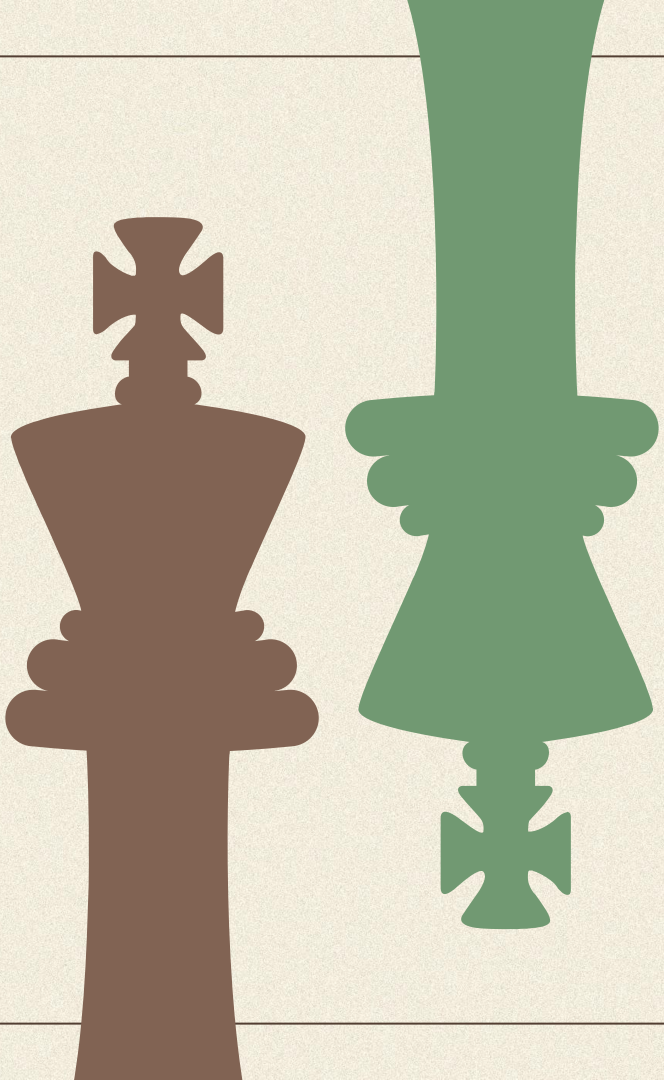
```
Actual is Ruy Lopez, index of 34
Predicted is Queen's Pawn Opening, index of 30
Actual is King's Gambit Accepted, index of 2
Predicted is Queen's Pawn Opening, index of 30
Actual is Scandinavian Defense, index of 4
Predicted is Benko Gambit Accepted, index of 29
Actual is Italian Game, index of 21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

# ASSUMPTIONS AND CHANGES

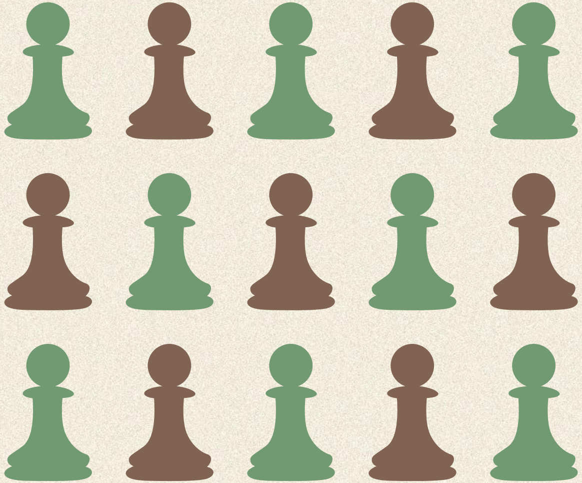
The dataset was accurate and unbiased - not created by machines.

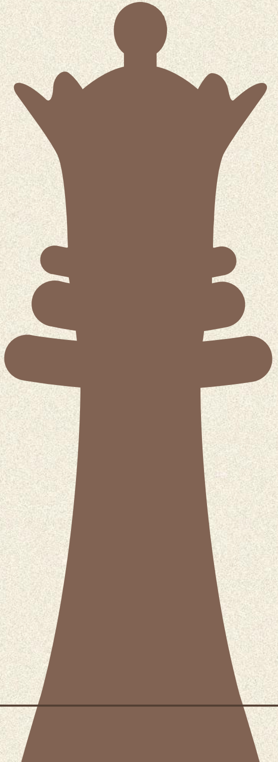
Moves are representable as numbers for the sake of a linear regression model.

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

# PROJECT EXTENSIONS

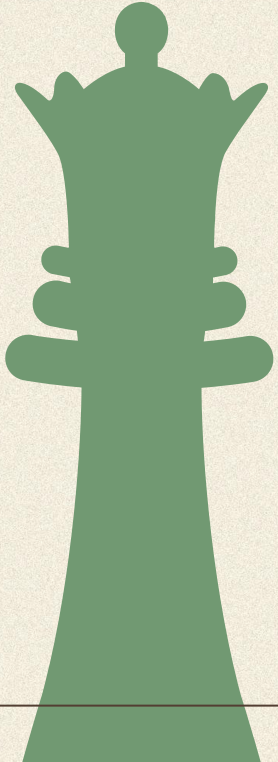
- Larger datasets
- More robust ML models with more carefully curated data
- What else has an effect on the length of the game?
- How does the opening affect the average **engine evaluation**?





# 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**