The data I will be using is as many chess games as possible - and combining as much data as I can from different ELO (chess rating) clusters, labeled as the following:

- Novice (500-800)
- Beginner (800-1099)
- Intermediate (1100-1399)
- Intermediate2 (1400-1699)
- Advanced (1700-1999)
- Expert (2000-2299)
- Master(2300+)

Getting enough for each level is simple, as long as they all have the same key requirements: the ELOs of each player, the time each player has for the game, and a string of all the moves of the game, as recorded in simple chess script.

One such example is a chess games database from Kaggle:
https://www.kaggle.com/datasnaek/chess

**Data Provenance, Data Biography:**

1. Who collected the data?
    a. Lichess, the 2nd most popular website to play chess (Chess.com)
    b. Should be noted that lichess ratings are known to be different than typical chess ratings, and there are conversions made, but its rough and it changes at each level
2. How did they collect it?
    a. Lichess API
    b. They were the most recent games taken from users from the top ~100 teams on Lichess.
3. For what purpose?
    a. Same reason as me, to analyze chess games
4. How is it used? By whom? What are its impacts? On whom?
    a. They have not used it on their own - they made the dataset for others to use
5. What are the known limitations?
    a. Some missing data
        i. Nothing to impact
    b. Data is “old” - 4 years ago
        i. Found a https://database.lichess.org/ but a bit harder to parse through as of now

**Ownership of the data, including licensing issues**

- All games are based on the collection of any given user's game history. In Lichess’s terms and service, as well as their main settings, players can disable using their data for public use.
- Data has a CC0: Public Domain License

**No worries about Bias or Censoring**