Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess

Reid McIlroy-Young
University of Toronto

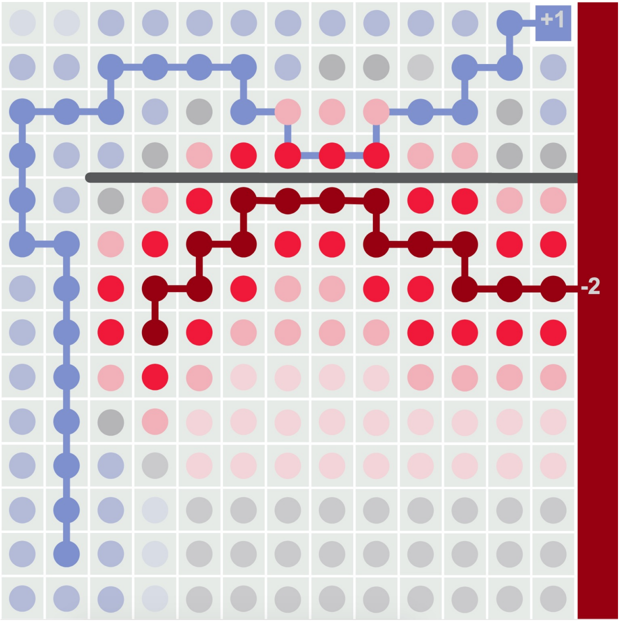
Russell Wang
University of Toronto
Now UC Berkeley

Siddhartha Sen
Microsoft Research

Jon Kleinberg
Cornell University

Ashton Anderson
University of Toronto

Motivation



Source: Comunidad de Software Libre Hackem

Modern machine learning systems can surpass human decision making

Human compatible machine learning systems are becoming more important

Motivation, cont

Existing work

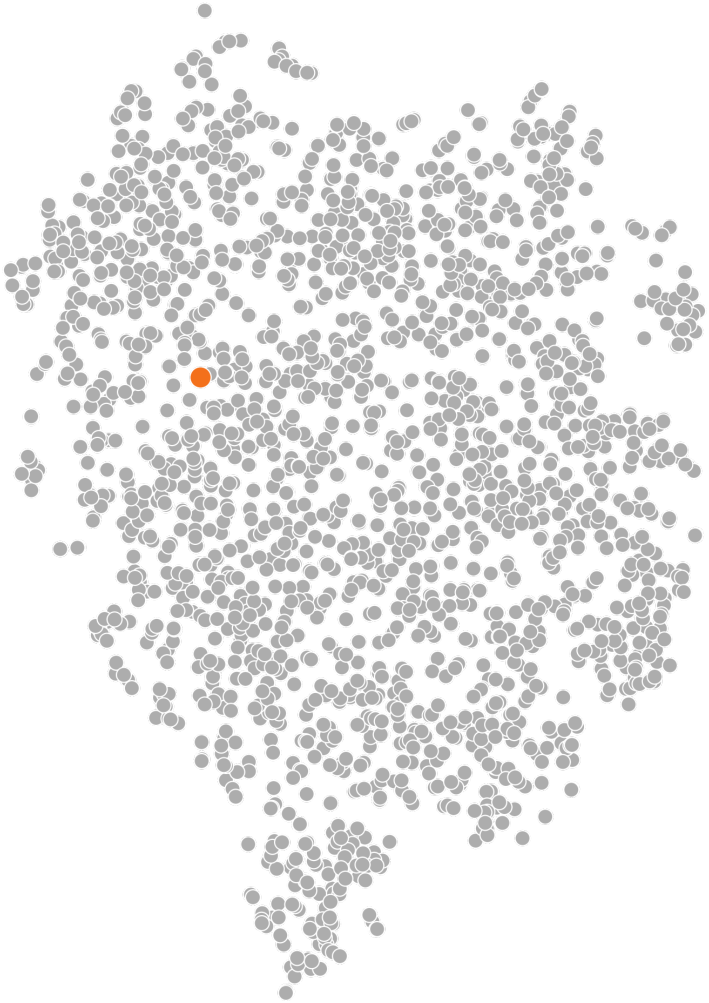
Characterize decision-making with an aggregate measure: skill, performance, age, …

What we want

AI systems that understand humans as individuals: strengths, weaknesses, style, …

Behavioral Stylometry

Task of identifying individuals simply from their decisions



Chess as a Model System

Superhuman AI

- Since 2007
- Open-source
  - *Stockfish*
  - *Leela Chess Zero*

Large Datasets

- *Lichess* – open database
- 2.6+ Billion chess games

Diverse players

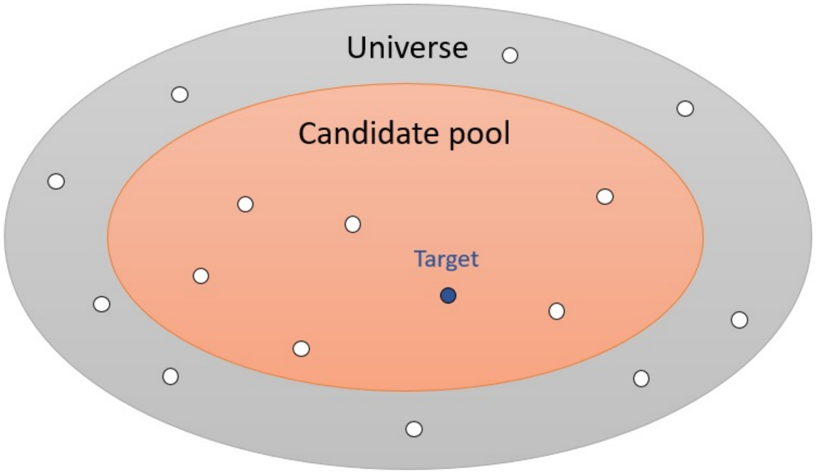
- Many countries
- Accurate skill measure (Elo)
  - Higher is stronger

Relatively Benign

- Games are public with limited PII
- Lower risk to players

Task Setup

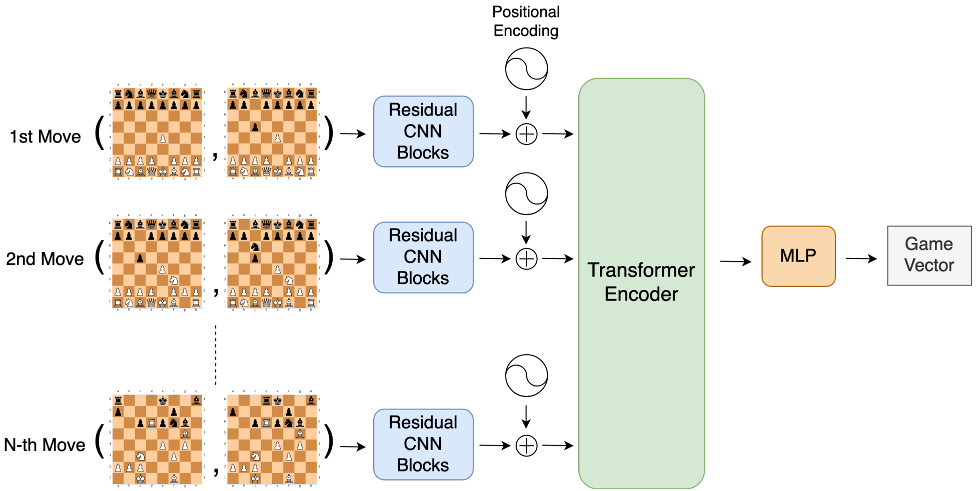
Identify a **specific** chess player from a finite of pool of candidates, using their moves during games



Given a set of moves from a *query* set (x_q) of games by an unknown **target player**, find the correct label for the player from a **candidate pool** of labelled players, from the universe of all players. Each labelled player has a *reference* set (x_r) of games.

Methods, Model Design

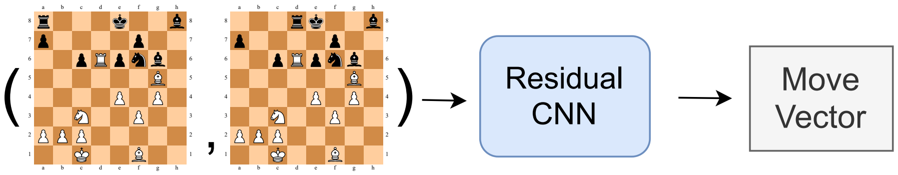
- Few shot learning approach
- Transformer takes in a chess game, outputs game embedding vector
- Training with Generalized End-to-End Loss (GE2E) loss[2]
  - Designed for speaker verification in audio

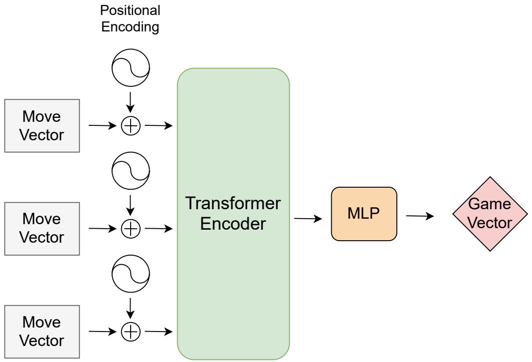


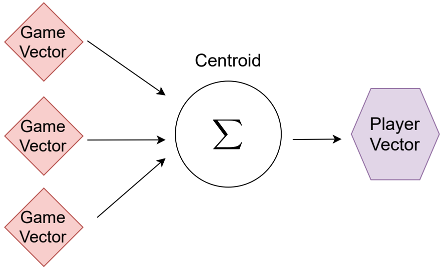
[2] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4879–4883, 2018. URL: https://wangquan.me/files/research/GE2E_ICASSP_2018.pdf

Methods, Model Pipeline

- Moves are represented as images:
  - (state before, state after)
- CNN resnet -> *move vector*
- Transformer takes all moves from game -> *game vector* (y)
- Centroid combines game vectors -> *player vector* (c)







Generalized End-to-End Loss (GE2E)

Minimize cosine distance between games by the **same player**

Maximize the distance between samples from **other players**

$$\mathbf{S}_{ji,k} = \begin{cases} w \cdot \cos(\mathbf{y}_{ji}, \mathbf{c}_j^{(-i)}) + b & \text{if} \quad k = j; \\ w \cdot \cos(\mathbf{y}_{ji}, \mathbf{c}_k) + b & \text{otherwise.} \end{cases}$$

1. Builds a similarity matrix on a batch of $N \times M$ games
   - N is number of players
   - M is number of games per player
2. For each game by each player, calculate game vector x_{ji}
   - ith game from the jth player
3. For each player compute the centroid of their games c_j
4. Then compute similarity matrix $S_{ji,k}$
   - w and b are learned scaling parameters

Generalized End-to-End Loss (GE2E), Loss Calculation

Loss per sample is then calculated as:

- ith game from the jth player

$$L(\mathbf{y}_{ji}) = -\mathbf{S}_{ji,j} + \log \sum_{k=1}^{N} \exp(\mathbf{S}_{ji,k})$$

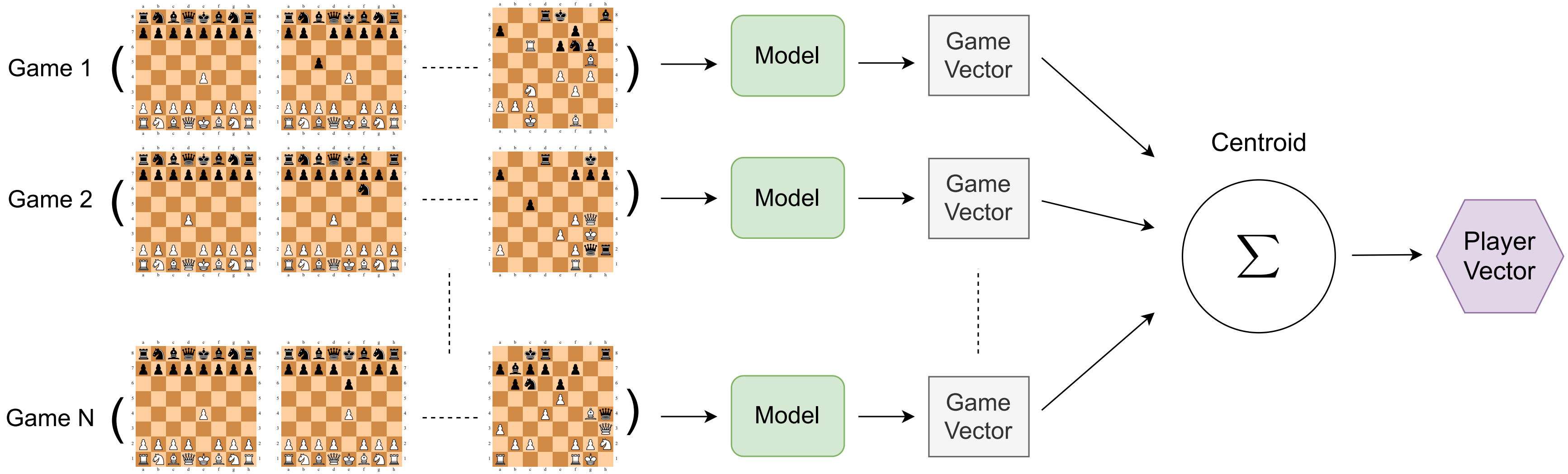
Total loss for the batch is:

$$L_{GE2E} = \sum_{j,i} L(\mathbf{y}_{ji})$$

SGD with momentum is used for optimizing

Inference

- Given query set (x_q) of a target player
- Calculate *game vectors*
- Compute centroid (*player vector*)
- Find nearest labelled player in candidate pool



Data

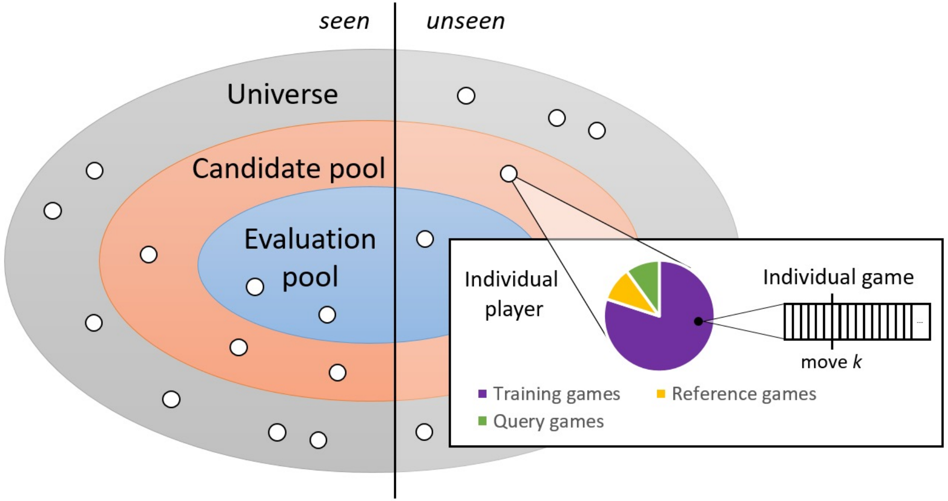
- Chess games from *Lichess*[3]
- Players with ratings between 1100 and 2000 Elo
- Blitz games (3-5 minute)
- Players with over 1000 games

- Trained on **seen** set of players
  - 63.7 million games
  - 16,181 players in total
- **Unseen** set: players never seen during training
- All results shown are on **games** the models were not trained on

[3] Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. "Learning Personalized Models of Human Behavior in Chess". URL: https://arxiv.org/abs/2008.10086

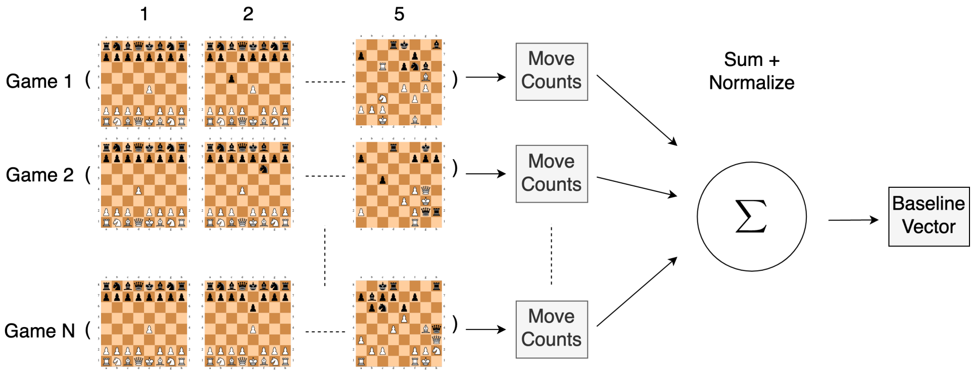
Experimental Setup

- Each target in the evaluation pool is considered separately
- **Target** players have *query set* (x_q) of 100 games
- **Candidate** players have *reference set* (x_s) of 100 games
- Focus on k=15
  - 16th and onwards moves by player
  - Mid/late game, as early is often formulaic



Baseline Model

- Sample 5 move sequence from a each game
  - k=15: 16th to 20th moves
  - k=0: 1st to 5th moves
- one-hot 4096-dimensional encoding vector for each move
- Sum game vectors, normalize
- Uses cosine distance like transformer model



Results, After Move 15 (k=15)

Test Description

- Candidate pool 2844 players
  - 2266 in **seen**, player in training
  - 578 in **unseen**, player not in training
- Only used decisions/moves after both players have made 15 actions

Accuracy (Top 1)

	Random	Baseline	Our Model
Unseen Only	0.04%	24.4%	86.0%
Unseen+ Seen	0.04%	26.8%	85.4%

Results, Whole Game (k=0)

Test Description

- Candidate pool 2844 players
  - 2266 in **seen**, player in training
  - 578 in **unseen**, player not in training
- All moves in the game are used

Accuracy (Top 1)

	Random	Baseline	Our Model
Unseen Only	0.04%	92.9%	97.9%
Unseen+ Seen	0.04%	92.9%	98.2%

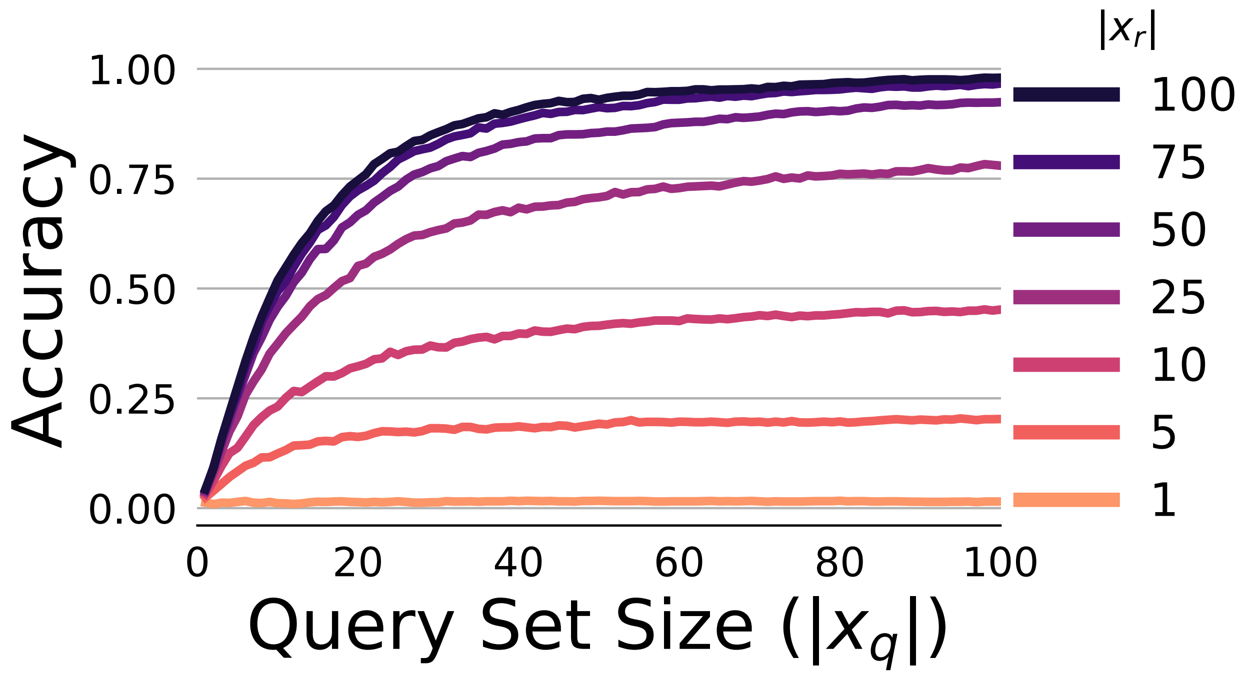
Results, Other Datasets, k=15

- ***McIlroy-Young et al.***
  - 400 players, candidate=evaluation pool
  - Personalized model as comparison
    - Requires 20k+ reference games
- **High Ranked Players**
  - Lichess and `chess.com` leaderboards
  - Candidate pool: high rank + mid rank
  - Evaluation pool: high rank
- **Large Dataset**
  - 41,184 players from Lichess
  - Candidate=evaluation pool
  - Includes 16k **seen** players

Accuracy (Top 1)

	Baseline	Personalized	Our Model
McIlroy-Young et al.	47.8%	55.2%	95.3%
High Ranked Players	2.7%		30.1%
Large Dataset	8.49%		54.0%

Increasing x_r or x_q size has diminishing returns



Ethics

Privacy

- This can be used to identify players who wish to remain anonymous
- The embedding could also reveal other information, such as gender

Generalization

- These methods may be applicable to other domains
- This work is a first look at the implications of behavioral stylometry
- The research community should develop understanding before applying these techniques to a higher stakes domain

Conclusion

Behavioral Stylometry is a novel problem	Few shot identification of chess players	Ethical Considerations require further consideration
• Design AI systems that can recognize people based on their decisions	• Transformer model that embeds players and games as vectors in a high dimensional space	• Privacy for existing players • Generalizations may cause significant concerns

Additional Information

Code	github.com/CSSLab/behavioral-stylometry
Email	reidmcy@cs.toronto.edu
Twitter	twitter.com/maiachess
Website	maiachess.com