Skill Assessment Versus Prediction in Game Play and Cheating Detection

Kenneth W. Regan
University at Buffalo (SUNY)

Oxford University, 3 November 2015

Skill Assessment “Versus” Prediction

- Skill Assessment: how well people did.

Skill Assessment “Versus” Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.

Skill Assessment “Versus” Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?

Skill Assessment “Versus” Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- Meta: Is this performance really by this person?

Skill Assessment “Versus” Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- Meta: Is this performance really by this person?
- Chess cheating detection needs both and more.

Cycling Analogy

- E-Doping means cheating with computer assistance.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately. *versus:*

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately. *versus:*
3. "Person X has hematocrit $> 50\%$."

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately. *versus:*
3. "Person X has hematocrit $> 50\%$."
4. "Person X made moves highly similar to Code Patch Y."

Why Chess?

- Long history, worldwide competitions.

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise (except for time taken on each move?).

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise (except for time taken on each move?).
- Computers play much better than best humans, which is **awful!**

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise (except for time taken on each move?).
- Computers play much better than best humans, which is **great!** since we can generate huge amounts of authoritative analysis data.

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise (except for time taken on each move?).
- Computers play much better than best humans, which is **great!** since we can generate huge amounts of authoritative analysis data.
- Chess—much more than Go for instance—lends itself to robust numerical evaluation.

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise (except for time taken on each move?).
- Computers play much better than best humans, which is **great!** since we can generate huge amounts of authoritative analysis data.
- Chess—much more than Go for instance—lends itself to robust numerical evaluation.
- Chess move options are *discrete*, hence closer to applications like *multiple-choice tests*.

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise (except for time taken on each move?).
- Computers play much better than best humans, which is **great!** since we can generate huge amounts of authoritative analysis data.
- Chess—much more than Go for instance—lends itself to robust numerical evaluation.
- Chess move options are *discrete*, hence closer to applications like *multiple-choice tests*.
- Both chess and online games foster notions of **difficulty**.

Why Chess?

- Long history, worldwide competitions.
- Game data readily and publicly available.
- Game data is precise (except for time taken on each move?).
- Computers play much better than best humans, which is **great!** since we can generate huge amounts of authoritative analysis data.
- Chess—much more than Go for instance—lends itself to robust numerical evaluation.
- Chess move options are *discrete*, hence closer to applications like *multiple-choice tests*.
- Both chess and online games foster notions of **difficulty**.
- Chess seems better for notions of **depth**.

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number.

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. “I’m a 2370.”

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. “I’m a 2370.”
- Number has no absolute meaning—only rating differences matter.

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. "I'm a 2370."
- Number has no absolute meaning—only rating differences matter.
- Difference of 200 $\approx$ 75% expectation for higher player,

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. "I'm a 2370."
- Number has no absolute meaning—only rating differences matter.
- Difference of 200 $\approx$ 75% expectation for higher player,
- Predictive content: your rating is the current best estimate of how you will perform in the next tournament.

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. "I'm a 2370."
- Number has no absolute meaning—only rating differences matter.
- Difference of 200 $\approx$ 75% expectation for higher player,
- Predictive content: your rating is the current best estimate of how you will perform in the next tournament.
- **TPR**: Tournament Performance Rating.

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. "I'm a 2370."
- Number has no absolute meaning—only rating differences matter.
- Difference of 200 $\approx$ 75% expectation for higher player,
- Predictive content: your rating is the current best estimate of how you will perform in the next tournament.
- **TPR**: Tournament Performance Rating.
- Rating and TPR based only on results of games and ratings of opponents.

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. "I'm a 2370."
- Number has no absolute meaning—only rating differences matter.
- Difference of 200 $\approx$ 75% expectation for higher player,
- Predictive content: your rating is the current best estimate of how you will perform in the next tournament.
- **TPR**: Tournament Performance Rating.
- Rating and TPR based only on results of games and ratings of opponents.
- Indeed relatively few games: 100 in a year is a lot for pro and amateur alike.

Chess Ratings: The (Original) Elo System

- Skill Assessment in One Number. "I'm a 2370."
- Number has no absolute meaning—only rating differences matter.
- Difference of 200 $\approx$ 75% expectation for higher player,
- Predictive content: your rating is the current best estimate of how you will perform in the next tournament.
- **TPR**: Tournament Performance Rating.
- Rating and TPR based only on results of games and ratings of opponents.
- Indeed relatively few games: 100 in a year is a lot for pro and amateur alike. Compare to 1,200 being a common need for a good election poll.

Elo Rating Examples

- Bobby Fischer hit **2800** on the US Chess Federation's Elo tabulation, **2785** on the FIDE list in July 1972.

Elo Rating Examples

- Bobby Fischer hit **2800** on the US Chess Federation's Elo tabulation, **2785** on the FIDE list in July 1972.
- Current world champion Magnus Carlsen broke Garry Kasparov's record of 2851, reached 2882 a year ago. Now 2850.

Elo Rating Examples

- Bobby Fischer hit **2800** on the US Chess Federation's Elo tabulation, **2785** on the FIDE list in July 1972.
- Current world champion Magnus Carlsen broke Garry Kasparov's record of 2851, reached 2882 a year ago. Now 2850.
- Current world #42 has 2702, world #100 has 2653.

Elo Rating Examples

- Bobby Fischer hit **2800** on the US Chess Federation's Elo tabulation, **2785** on the FIDE list in July 1972.
- Current world champion Magnus Carlsen broke Garry Kasparov's record of 2851, reached 2882 a year ago. Now 2850.
- Current world #42 has 2702, world #100 has 2653.
- Formal "Master" designation for USCF is 2200; "FIDE Master" is a formal *title* (IMHO) more typical of 2300.

Elo Rating Examples

- Bobby Fischer hit **2800** on the US Chess Federation's Elo tabulation, **2785** on the FIDE list in July 1972.
- Current world champion Magnus Carlsen broke Garry Kasparov's record of 2851, reached 2882 a year ago. Now 2850.
- Current world #42 has 2702, world #100 has 2653.
- Formal "Master" designation for USCF is 2200; "FIDE Master" is a formal *title* (IMHO) more typical of 2300. Likewise "International Master" $\approx$ 2400, *Grandmaster* $\approx$ 2500, "strong GM" $\approx$ 2600.

Elo Rating Examples

- Bobby Fischer hit **2800** on the US Chess Federation's Elo tabulation, **2785** on the FIDE list in July 1972.
- Current world champion Magnus Carlsen broke Garry Kasparov's record of 2851, reached 2882 a year ago. Now 2850.
- Current world #42 has 2702, world #100 has 2653.
- Formal "Master" designation for USCF is 2200; "FIDE Master" is a formal *title* (IMHO) more typical of 2300. Likewise "International Master" $\approx$ 2400, *Grandmaster* $\approx$ 2500, "strong GM" $\approx$ 2600.
- USCF uses 2000–2199 = "Expert," 1800–1999 = "Class A," 1600–1799 = "Class B" and so on.

Elo Ratings, continued

- Adult beginner typically 600, tournament/club "novice" 1200; scholastics go down below 100.

Elo Ratings, continued

- Adult beginner typically 600, tournament/club “novice” 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a “Class Unit.”

Elo Ratings, continued

- Adult beginner typically 600, tournament/club "novice" 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a "Class Unit." Number of class units from beginner to champion = game's Human Depth.

Elo Ratings, continued

- Adult beginner typically 600, tournament/club "novice" 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a "Class Unit." Number of class units from beginner to champion = game's Human Depth.
- From 600 to 2800 gives chess a human depth of 11. 8×8 checkers estimated at 10, backgammon and bridge similarly.

Elo Ratings, continued

- Adult beginner typically 600, tournament/club "novice" 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a "Class Unit." Number of class units from beginner to champion = game's Human Depth.
- From 600 to 2800 gives chess a human depth of 11. 8×8 checkers estimated at 10, backgammon and bridge similarly.
- Shogi (Japanese chess) at 14, Go at least above 20, maybe 25?

Elo Ratings, continued

- Adult beginner typically 600, tournament/club “novice” 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a “Class Unit.” Number of class units from beginner to champion = game’s Human Depth.
- From 600 to 2800 gives chess a human depth of 11. 8×8 checkers estimated at 10, backgammon and bridge similarly.
- Shogi (Japanese chess) at 14, Go at least above 20, maybe 25?
- Chess computer programs (called *engines*) on desktop PC hardware reach **3200** on all rating lists, **3380** on CCRL.

Elo Ratings, continued

- Adult beginner typically 600, tournament/club "novice" 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a "Class Unit." Number of class units from beginner to champion = game's Human Depth.
- From 600 to 2800 gives chess a human depth of 11. 8×8 checkers estimated at 10, backgammon and bridge similarly.
- Shogi (Japanese chess) at 14, Go at least above 20, maybe 25?
- Chess computer programs (called *engines*) on desktop PC hardware reach **3200** on all rating lists, **3380** on CCRL.
- Computers at least even at Shogi, knocking on door at Go?

Elo Ratings, continued

- Adult beginner typically 600, tournament/club "novice" 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a "Class Unit." Number of class units from beginner to champion = game's Human Depth.
- From 600 to 2800 gives chess a human depth of 11. 8×8 checkers estimated at 10, backgammon and bridge similarly.
- Shogi (Japanese chess) at 14, Go at least above 20, maybe 25?
- Chess computer programs (called *engines*) on desktop PC hardware reach **3200** on all rating lists, **3380** on CCRL.
- Computers at least even at Shogi, knocking on door at Go? "Moore's Law" of Games?

Elo Ratings, continued

- Adult beginner typically 600, tournament/club "novice" 1200; scholastics go down below 100.
- László Mérő formalized the 75%-gap as a "Class Unit." Number of class units from beginner to champion = game's Human Depth.
- From 600 to 2800 gives chess a human depth of 11. 8×8 checkers estimated at 10, backgammon and bridge similarly.
- Shogi (Japanese chess) at 14, Go at least above 20, maybe 25?
- Chess computer programs (called *engines*) on desktop PC hardware reach **3200** on all rating lists, **3380** on CCRL.
- Computers at least even at Shogi, knocking on door at Go? "Moore's Law" of Games? Beyond chess ceiling of **3500**??

Idea of “Intrinsic Performance Ratings” (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.

- Your 50–100 games will have 1,200—2,400 relevant moves. (I standardly exclude turns 1–8 and positions where one side has an overwhelming advantage.)

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.
- Your 50–100 games will have 1,200—2,400 relevant moves. (I standardly exclude turns 1–8 and positions where one side has an overwhelming advantage.)
- From just 200–300 moves in a tournament, error bars are high, $2\sigma = \pm 200$–300 typical.

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.
- Your 50–100 games will have 1,200—2,400 relevant moves. (I standardly exclude turns 1–8 and positions where one side has an overwhelming advantage.)
- From just 200–300 moves in a tournament, error bars are high, $2\sigma = \pm 200$–300 typical.
- **Deep Blue** played 2850–2900 in each of the matches against Garry Kasparov, while GK played

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.
- Your 50–100 games will have 1,200—2,400 relevant moves. (I standardly exclude turns 1–8 and positions where one side has an overwhelming advantage.)
- From just 200–300 moves in a tournament, error bars are high, $2\sigma = \pm 200$–300 typical.
- **Deep Blue** played 2850–2900 in each of the matches against Garry Kasparov, while GK played **under 2600**.

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.
- Your 50–100 games will have 1,200—2,400 relevant moves. (I standardly exclude turns 1–8 and positions where one side has an overwhelming advantage.)
- From just 200–300 moves in a tournament, error bars are high, $2\sigma = \pm 200$–300 typical.
- **Deep Blue** played 2850–2900 in each of the matches against Garry Kasparov, while GK played **under 2600**. But **$\pm$225–300**.

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.
- Your 50–100 games will have 1,200—2,400 relevant moves. (I standardly exclude turns 1–8 and positions where one side has an overwhelming advantage.)
- From just 200–300 moves in a tournament, error bars are high, $2\sigma = \pm 200$–300 typical.
- **Deep Blue** played 2850–2900 in each of the matches against Garry Kasparov, while GK played **under 2600**. But **±225–300**.
- Can pinpoint current quality of rapidly improving player.

Idea of "Intrinsic Performance Ratings" (IPRs)

- Primarily Skill Assessment; IPR for one event or series only.
- Based only on quality of your own move decisions. Results, opponents not involved.
- Your 50–100 games will have 1,200—2,400 relevant moves. (I standardly exclude turns 1–8 and positions where one side has an overwhelming advantage.)
- From just 200–300 moves in a tournament, error bars are high, $2\sigma = \pm 200$–300 typical.
- **Deep Blue** played 2850–2900 in each of the matches against Garry Kasparov, while GK played **under 2600**. But **±225–300**.
- Can pinpoint current quality of rapidly improving player.
- "Match Elo" versus "Hidden Rating" at *League of Legends*.

Case Example: April 2015

- The “San Sebastian Open”—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles...

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles. . .
- But my full cheating test showed only a "1.3-sigma" deviation,

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles. . .
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles...
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.
- Was dead lost against Epishin, lucked out also in previous round,

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles...
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.
- Was dead lost against Epishin, lucked out also in previous round,
- World #2 Fabiano Caruana had sensational 7-win streak against the top last Sept.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles...
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.
- Was dead lost against Epishin, lucked out also in previous round,
- World #2 Fabiano Caruana had sensational 7-win streak against the top last Sept.—but his IPR was "only" 2900 while his opponents played under 2600.

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
    - How often did the favored horses win in a racing week?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
    - How often did the favored horses win in a racing week?
    - Do basketball teams average "covering their spread"?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds “even out” over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
  - How often did the favored horses win in a racing week?
  - Do basketball teams average “covering their spread”?
  - How often did Player X make the move favored by an engine?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
  - How often did the favored horses win in a racing week?
  - Do basketball teams average "covering their spread"?
  - How often did Player X make the move favored by an engine?
  - How does his/her "Average Error" compare?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
  - How often did the favored horses win in a racing week?
  - Do basketball teams average "covering their spread"?
  - How often did Player X make the move favored by an engine?
  - How does his/her "Average Error" compare?
- Also project *standard deviation* and *confidence intervals*.

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \ldots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \ldots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E
4. Defines *fallible agent* $P(s, c, \ldots)$.

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \dots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E
4. Defines *fallible agent* $P(s, c, \dots)$.
5. Main Output: Probabilities $p_{t,i}$ for $P(s, c, \dots)$ to select option i at time t.

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \dots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E
4. Defines *fallible agent* $P(s, c, \dots)$.
5. Main Output: Probabilities $p_{t,i}$ for $P(s, c, \dots)$ to select option i at time t.
6. Derived Outputs:
   - Aggregate statistics: *move-match* **MM**, *equal-top value* **EV**, *average scaled difference* **ASD**, . . .
   - Projected confidence intervals: Bernoulli Trials + $|T|$-adjustment.
   - IPRs similarly reflect errors from the regression.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \ldots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \ldots)$.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i$$

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i = \exp(-\left(\frac{\delta(v_1, v_i)}{s}\right)^c),$$

- Here $\delta(v_1, v_i)$ scales $v_1 - v_i$ in regard to $|v_1|$.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i = \exp(-\left(\frac{\delta(v_1, v_i)}{s}\right)^c),$$

- Here $\delta(v_1, v_i)$ scales $v_1 - v_i$ in regard to $|v_1|$.
- Ratio not difference on LHS so x_i on RHS has 0-to-1 scale.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i)$ = the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i = \exp(-\left(\frac{\delta(v_1, v_i)}{s}\right)^c),$$

- Here $\delta(v_1, v_i)$ scales $v_1 - v_i$ in regard to $|v_1|$.
- Ratio not difference on LHS so x_i on RHS has 0-to-1 scale.
- Given $(x_1, \dots, x_i, \dots, x_\ell)$, fit subject to $\sum_i p_i = 1$ to find p_1. Other p_i follow by $p_i = h^{-1}(h(p_1)(1 - x_i))$.

The Data (Before August 2015)

- Over 3 million moves of 50-PV data: $>$ 250 GB.

The Data (Before August 2015)

- Over 3 million moves of 50-PV data: $>$ 250 GB.
- Over 40 million moves of Single-PV data: $>$ 50 GB

The Data (Before August 2015)

- Over 3 million moves of 50-PV data: $>$ 250 GB.
- Over 40 million moves of Single-PV data: $>$ 50 GB
- $=$ 150 million pages of text data at 2k/page.

The Data (Before August 2015)

- Over 3 million moves of 50-PV data: $>$ 250 GB.
- Over 40 million moves of Single-PV data: $>$ 50 GB
- $=$ 150 million pages of text data at 2k/page.
- All this was taken on two quad-core home-style PC's plus a laptop.

The Data (Before August 2015)

- Over 3 million moves of 50-PV data: $>$ 250 GB.
- Over 40 million moves of Single-PV data: $>$ 50 GB
- $=$ 150 million pages of text data at 2k/page.
- All this was taken on two quad-core home-style PC's plus a laptop. Is this "Big Data"?

The Data (Before August 2015)

- Over 3 million moves of 50-PV data: $>$ 250 GB.
- Over 40 million moves of Single-PV data: $>$ 50 GB
- = 150 million pages of text data at 2k/page.
- All this was taken on two quad-core home-style PC's plus a laptop. Is this "Big Data"?

The Data (Before August 2015)

- Over 3 million moves of 50-PV data: $>$ 250 GB.
- Over 40 million moves of Single-PV data: $>$ 50 GB
- $=$ 150 million pages of text data at 2k/page.
- All this was taken on two quad-core home-style PC's plus a laptop. Is this "Big Data"? New sets being taken with **UB CCR** cluster.

Fitting and Fighting Parameters

- For each Elo level E training set, find $(s, c, \dots)$ giving best fit.

Fitting and Fighting Parameters

- For each Elo level E training set, find $(s, c, \dots)$ giving best fit.
- **Or** do Bayesian update to infer parameter(s) that best explain data [Haworth, later work joint with me and G. Di Fatta].

Fitting and Fighting Parameters

- For each Elo level E training set, find $(s, c, \dots)$ giving best fit.
- **Or** do Bayesian update to infer parameter(s) that best explain data [Haworth, later work joint with me and G. Di Fatta].
- In frequentist view, can use many different fitting methods...
  - Can compare methods...
  - Whole separate topic...

Fitting and Fighting Parameters

- For each Elo level E training set, find $(s, c, \dots)$ giving best fit.
- **Or** do Bayesian update to infer parameter(s) that best explain data [Haworth, later work joint with me and G. Di Fatta].
- In frequentist view, can use many different fitting methods...
  - Can compare methods...
  - Whole separate topic...
  - Max-Likelihood does relatively poorly.
- Often s and c trade off markedly, but $E' \sim e(s, c)$ condenses into one Elo.
- Strong linear fit—suggests Elo mainly influenced by error.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.
- Is it scientific?

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.
- Is it scientific?
- With my formulas, yes—using *your* games in *real* tournaments.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.
- Is it scientific?
- With my formulas, yes—using *your* games in *real* tournaments.
- Goal is **natural** scoring and distribution evaluation for multiple-choice tests, especially with partial-credit answers.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.
- Is it scientific?
- With my formulas, yes—using *your* games in *real* tournaments.
- Goal is **natural** scoring and distribution evaluation for multiple-choice tests, especially with partial-credit answers.
- Connect to parameters in **Item-Response Theory** (IRT) test-taking models.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.
- Is it scientific?
- With my formulas, yes—using *your* games in *real* tournaments.
- Goal is **natural** scoring and distribution evaluation for multiple-choice tests, especially with partial-credit answers.
- Connect to parameters in **Item-Response Theory** (IRT) test-taking models. IRT does both skill and prediction.

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.
  - Good retreating moves are harder to find(?)

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.
  - Good retreating moves are harder to find(?)
  - Planning tendency may show in repeated moves of same piece.

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.
  - Good retreating moves are harder to find(?)
  - Planning tendency may show in repeated moves of same piece.
- Prediction Idea 2: Use Player-Specific Information ("profiling").
  - Regress previous games by player.

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.
  - Good retreating moves are harder to find(?)
  - Planning tendency may show in repeated moves of same piece.
- Prediction Idea 2: Use Player-Specific Information ("profiling").
  - Regress previous games by player.
  - Style is more "positional"?

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.
  - Good retreating moves are harder to find(?)
  - Planning tendency may show in repeated moves of same piece.
- Prediction Idea 2: Use Player-Specific Information (“profiling”).
  - Regress previous games by player.
  - Style is more “positional”? “tactical”?

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.
  - Good retreating moves are harder to find(?)
  - Planning tendency may show in repeated moves of same piece.
- Prediction Idea 2: Use Player-Specific Information (“profiling”).
  - Regress previous games by player.
  - Style is more “positional”? “tactical”?
- Drawbacks: loss of neutrality and portability.

Separating Skill Assessment and Prediction [BHR]

- Thus far using same formulas for both.
- Linchpin: Use **best-available** computer move values for assessment.
- Prediction Idea 1: Use chess-specific features.
  - Good retreating moves are harder to find(?)
  - Planning tendency may show in repeated moves of same piece.
- Prediction Idea 2: Use Player-Specific Information ("profiling").
  - Regress previous games by player.
  - Style is more "positional"? "tactical"?
- Drawbacks: loss of neutrality and portability.
- Can we find more properties in the raw numerical data?

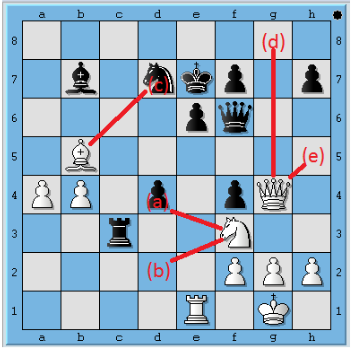
Example of "Swing" over Increasing Depths

The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

- (a) vigor . . corroborated
- (b) feebleness . . dashed
- (c) proliferation . . blighted
- (d) destruction . . disputed
- (e) disappearance . . frustrated

(source: itunes.apple.com)

=



Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.

An “ESP Test”

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.
- Conditioned on one of the two moves having been played, let us invite humans to guess **which move is listed first by the program**.

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.
- Conditioned on one of the two moves having been played, let us invite humans to guess **which move is listed first by the program**.
- The values are identical to the engine: it would not matter to the quality of the output which one the engine listed first. The values give no human reason to prefer one over the other.

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.
- Conditioned on one of the two moves having been played, let us invite humans to guess **which move is listed first by the program**.
- The values are identical to the engine: it would not matter to the quality of the output which one the engine listed first. The values give no human reason to prefer one over the other.
- So this is a kind of ESP test.

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.
- Conditioned on one of the two moves having been played, let us invite humans to guess **which move is listed first by the program**.
- The values are identical to the engine: it would not matter to the quality of the output which one the engine listed first. The values give no human reason to prefer one over the other.
- So this is a kind of ESP test. *How well do humans perform on it?*

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.
- Conditioned on one of the two moves having been played, let us invite humans to guess **which move is listed first by the program**.
- The values are identical to the engine: it would not matter to the quality of the output which one the engine listed first. The values give no human reason to prefer one over the other.
- So this is a kind of ESP test. *How well do humans perform on it?*
- Dick's Dean at Princeton: PEAR—Princeton Engineering Anomalies Research.

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.
- Conditioned on one of the two moves having been played, let us invite humans to guess **which move is listed first by the program**.
- The values are identical to the engine: it would not matter to the quality of the output which one the engine listed first. The values give no human reason to prefer one over the other.
- So this is a kind of ESP test. *How well do humans perform on it?*
- Dick's Dean at Princeton: PEAR—Princeton Engineering Anomalies Research.
- PEAR did 10,000s–100,000s of trials, trying to judge significance of deviations like 50.1% or even 50.01%.

An "ESP Test"

- In 8%–10% of positions, engine gives the top two moves the same value. Values are discrete up to 1 **centipawn**.
- More often *some* pair of moves in the top 10 (say) will end up tied.
- Conditioned on one of the two moves having been played, let us invite humans to guess **which move is listed first by the program**.
- The values are identical to the engine: it would not matter to the quality of the output which one the engine listed first. The values give no human reason to prefer one over the other.
- So this is a kind of ESP test. *How well do humans perform on it?*
- Dick's Dean at Princeton: PEAR—Princeton Engineering Anomalies Research.
- PEAR did 10,000s–100,000s of trials, trying to judge significance of deviations like 50.1% or even 50.01%.
- How well do real humans perform on *my* ESP test??

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.
3. **0.03**, the higher move is played 60–61% of the time.

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.
3. **0.03**, the higher move is played 60–61% of the time.
4. **0.00**, the higher move is played

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.
3. **0.03**, the higher move is played 60–61% of the time.
4. **0.00**, the higher move is played 57-59% of the time.

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.
3. **0.03**, the higher move is played 60–61% of the time.
4. **0.00**, the higher move is played 57-59% of the time.

- Last is not a typo—see post "**When is a Law Natural?**"

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.
3. **0.03**, the higher move is played 60–61% of the time.
4. **0.00**, the higher move is played 57-59% of the time.

- Last is not a typo—see post "**When is a Law Natural?**"
- Similar 58%-42% split seen for any pair of tied moves. What can explain it?

Sensitivity—Still the Slime Mold Story?

Conditioned on one of the top two moves being played, if their values (Rybka 3, depth 13) differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.
3. **0.03**, the higher move is played 60–61% of the time.
4. **0.00**, the higher move is played 57-59% of the time.

- Last is not a typo—see post "**When is a Law Natural?**"
- Similar 58%-42% split seen for any pair of tied moves. What can explain it?
- Relation to slime molds and other "semi-Brownian" systems?

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation:

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.
- By stability, lower move can become 1st only with *strictly higher* value.

Measuring “Swing” and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.
- By stability, lower move can become 1st only with *strictly higher* value.
- Lead moves tend to have been higher at lower depths. Lower move “swings up.”

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.
- By stability, lower move can become 1st only with *strictly higher* value.
- Lead moves tend to have been higher at lower depths. Lower move "swings up."
- Formulate numerical measure of swing "up" and "down" (a trap).

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.
- By stability, lower move can become 1st only with *strictly higher* value.
- Lead moves tend to have been higher at lower depths. Lower move "swings up."
- Formulate numerical measure of swing "up" and "down" (a trap).
- When best move swings up **4.0–5.0** versus **0.0–1.0**, players rated 2700+ find it only **30%** versus **70%**.

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.
- By stability, lower move can become 1st only with *strictly higher* value.
- Lead moves tend to have been higher at lower depths. Lower move "swings up."
- Formulate numerical measure of swing "up" and "down" (a trap).
- When best move swings up **4.0–5.0** versus **0.0–1.0**, players rated 2700+ find it only **30%** versus **70%**.
- Goal is to develop a **Challenge Quotient** based on how much trappy play a player sets for the opponent

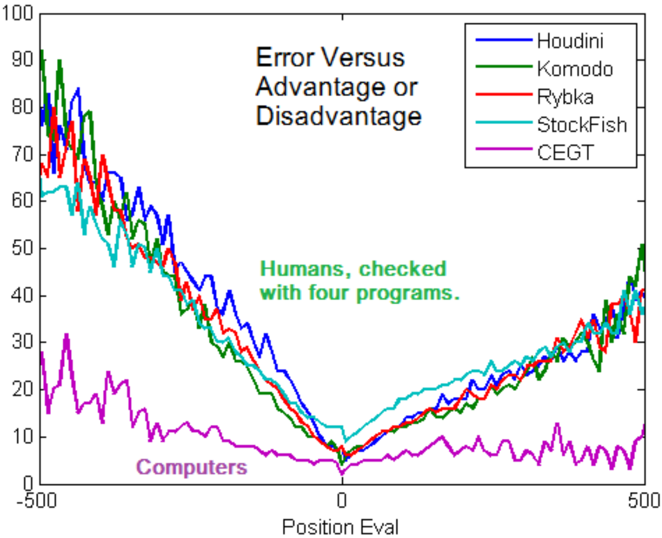
Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.
- By stability, lower move can become 1st only with *strictly higher* value.
- Lead moves tend to have been higher at lower depths. Lower move "swings up."
- Formulate numerical measure of swing "up" and "down" (a trap).
- When best move swings up **4.0–5.0** versus **0.0–1.0**, players rated 2700+ find it only **30%** versus **70%**.
- Goal is to develop a **Challenge Quotient** based on how much trappy play a player sets for the opponent—and emself.

Measuring "Swing" and Complexity and Difficulty

- Non-Parapsychological Explanation: *Stable* Library Sorting.
- Chess engines sort moves from last depth to schedule next round of search.
- By stability, lower move can become 1st only with *strictly higher* value.
- Lead moves tend to have been higher at lower depths. Lower move "swings up."
- Formulate numerical measure of swing "up" and "down" (a trap).
- When best move swings up **4.0–5.0** versus **0.0–1.0**, players rated 2700+ find it only **30%** versus **70%**.
- Goal is to develop a **Challenge Quotient** based on how much trappy play a player sets for the opponent—and emself.
- Separates *performance* and *prediction* in the model.

Human Versus Computer Phenomena



The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{c}{c+x}dx$$

balances evals well for Rybka, with c very near 1.0.

The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{c}{c+x}dx$$

balances evals well for Rybka, with c very near 1.0.

- A mix of three factors?

The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{c}{c+x}\,dx$$

balances evals well for Rybka, with c very near 1.0.

- A mix of three factors?

(A) Human perception of value as proportional to stakes, *per* Ariely-Kahneman-Tversky.

The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{c}{c+x}dx$$

balances evals well for Rybka, with c very near 1.0.

- A mix of three factors?

(A) Human perception of value as proportional to stakes, *per* Ariely-Kahneman-Tversky.

(B) Rationally playing less *catenaccio* when marginal impact of evaluation on win probability is minimal. (Leo Stedile, working under Mark Braverman)

The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{c}{c+x}dx$$

balances evals well for Rybka, with c very near 1.0.

- A mix of three factors?

(A) Human perception of value as proportional to stakes, *per* Ariely-Kahneman-Tversky.

(B) Rationally playing less *catenaccio* when marginal impact of evaluation on win probability is minimal. (Leo Stedile, working under Mark Braverman)

(C) Greater volatility intrinsic to chess as game progresses.

A. Perception Proportional to Benefit

How strongly do you perceive a difference of 10 dollars, if:

- You are buying lunch and a drink in a pub.
- You are buying dinner in a restaurant.
- You are buying an I-pad.
- You are buying a car.

For the car, maybe you don't care. In other cases, would you be equally thrifty?

> *If you spend the way you play chess, you care maybe* $4\times$ *as much in the pub!*

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.
- *How to test apart from cause A?*

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.
- *How to test apart from cause A?*
- Expect eval-error curve to shift in games between unequally-rated players.

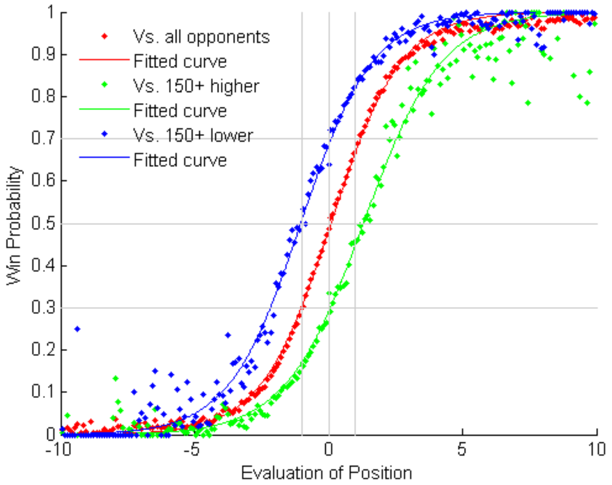
B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

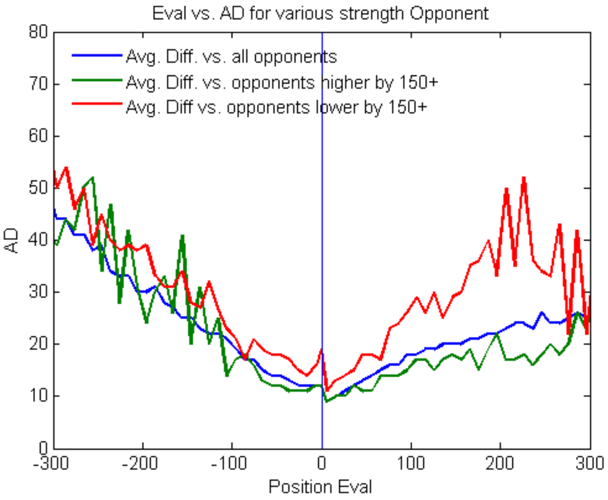
$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.
- *How to test apart from cause A?*
- Expect eval-error curve to shift in games between unequally-rated players.
- Results so far show no shift—

Human Versus Computer Phenomena



Eval-Error Curve With Unequal Players



Some IPRs—Historical and Current

- Carlsen:
  - **2985** at London 2011 (Kramnik 2857, Aronian 2838).
- Kasparov:
  - Was playing **2860** to Karpov's **2760** when 1984-85 match aborted.

Some IPRs—Historical and Current

- Carlsen:
  - **2985** at London 2011 (Kramnik 2857, Aronian 2838).
- Kasparov:
  - Was playing **2860** to Karpov's **2760** when 1984-85 match aborted.
  - Both over **2800** in 1986, Kasparov **2905**.

Some IPRs—Historical and Current

- Carlsen:
  - **2985** at London 2011 (Kramnik 2857, Aronian 2838).
- Kasparov:
  - Was playing **2860** to Karpov's **2760** when 1984-85 match aborted.
  - Both over **2800** in 1986, Kasparov **2905**.
  - Both under **2675** in New York-Lyon match 1990.

Some IPRs—Historical and Current

- Carlsen:
  - **2985** at London 2011 (Kramnik 2857, Aronian 2838).
- Kasparov:
  - Was playing **2860** to Karpov's **2760** when 1984-85 match aborted.
  - Both over **2800** in 1986, Kasparov **2905**.
  - Both under **2675** in New York-Lyon match 1990.
- Bobby Fischer:
  - **2920** over all 3 Candidates' Matches in 1971.
  - **2650** vs. Spassky in 1972 (Spassky 2645).
  - **2725** vs. Spassky in 1992 (Spassky 2660).
- Hou Yifan: **2970** vs. Humpy Koneru (2685) in Nov. 2011.
- Paul Morphy: **2345** in 59 most impt. games, 2125 vs. Anderssen.
- Capablanca: **2935** at New York 1927.
- Alekhine: **2810** in 1927 WC match over Capa (2730).

Computer and Freestyle IPRs

Analyzed Ratings of Computer Engine Grand Tournament (on commodity PCs) and PAL/CSS Freestyle in 2007–08, plus the Thoresen Chess Engines Competition (16-core) Nov–Dec. 2013.

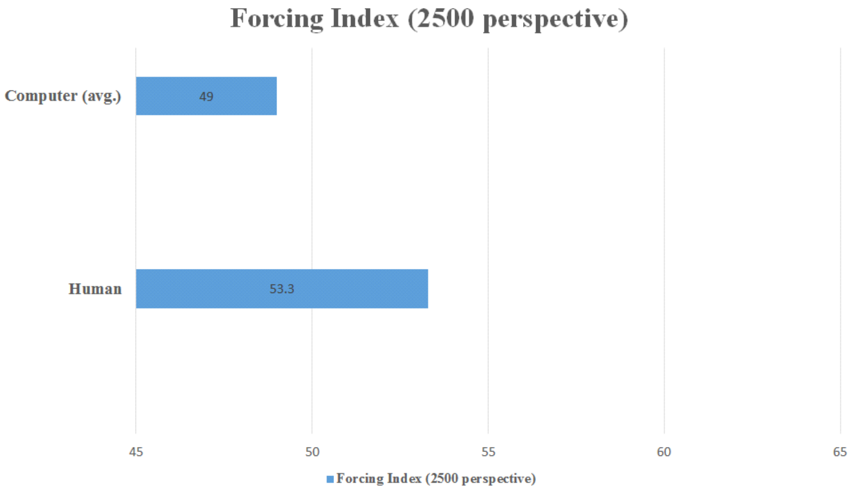
Event	Rating	2σ range	#gm	#moves
CEGT g1,50	3009	2962–3056	42	4,212
CEGT g25,26	2963	2921–3006	42	5,277
PAL/CSS 5ch	3102	3051–3153	45	3,352
PAL/CSS 6ch	3086	3038–3134	45	3,065
PAL/CSS 8ch	3128	3083–3174	39	3,057
TCEC 2013	3083	3062–3105	90	11,024

Computer and Freestyle IPRs—To Move 60

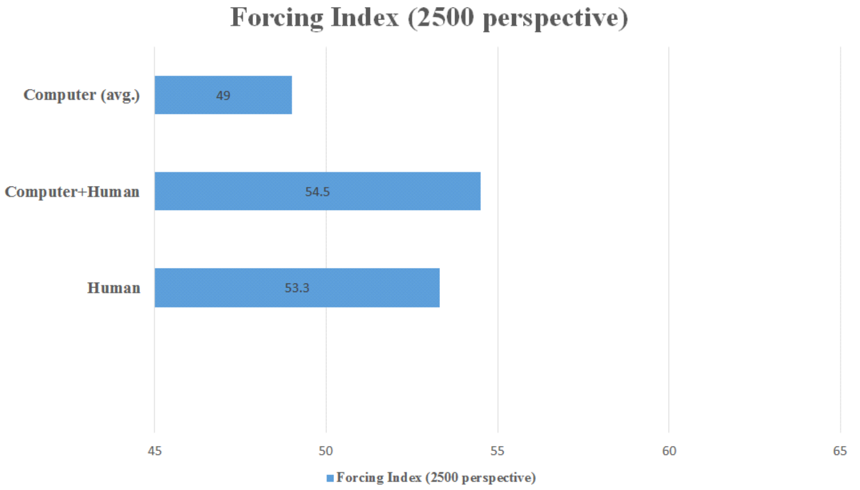
Computer games can go very long in dead drawn positions. TCEC uses a cutoff but CEGT did not. Human-led games tend to climax (well) before Move 60. This comparison halves the difference to CEGT, otherwise similar:

Sample set	Rating	2σ range	#gm	#moves
CEGT all	2985	2954–3016	84	9,489
PAL/CSS all	3106	3078–3133	129	9,474
TCEC 2013	3083	3062–3105	90	11,024
CEGT to60	3056	3023–3088	84	7,010
PAL/CSS to60	3112	3084–3141	129	8,744
TCEC to60	3096	3072–3120	90	8,184

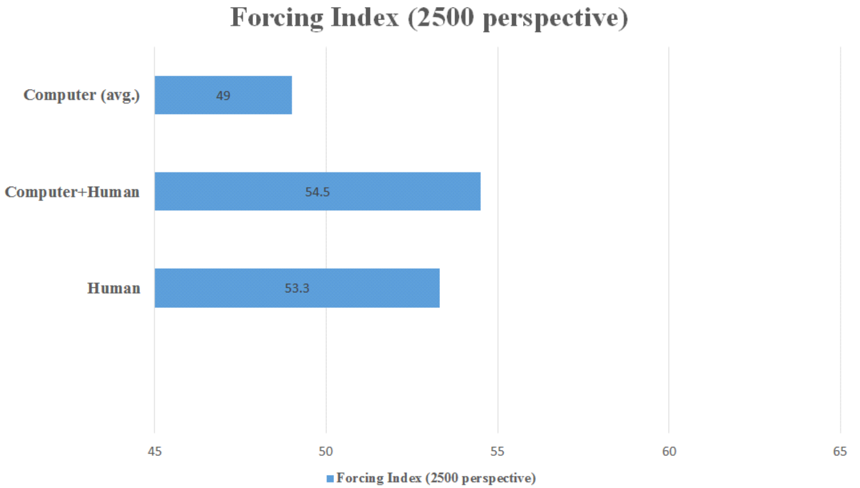
Degrees of Forcing Play



Add Human-Computer Tandems

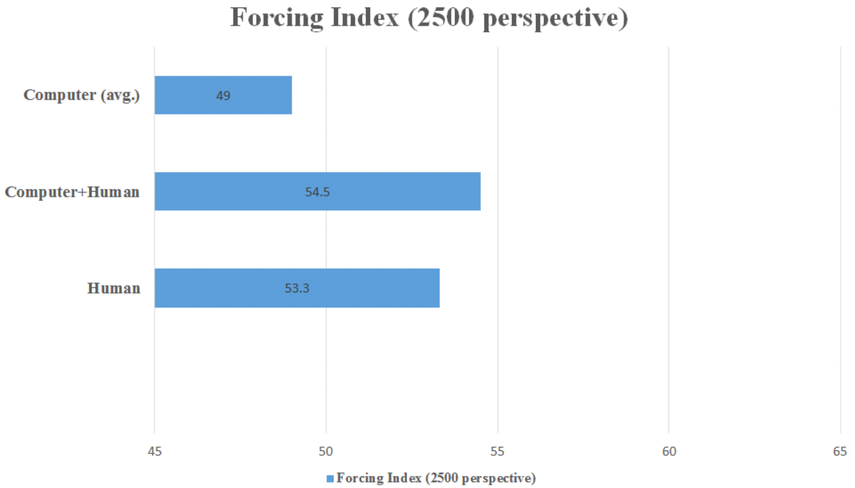


Add Human-Computer Tandems



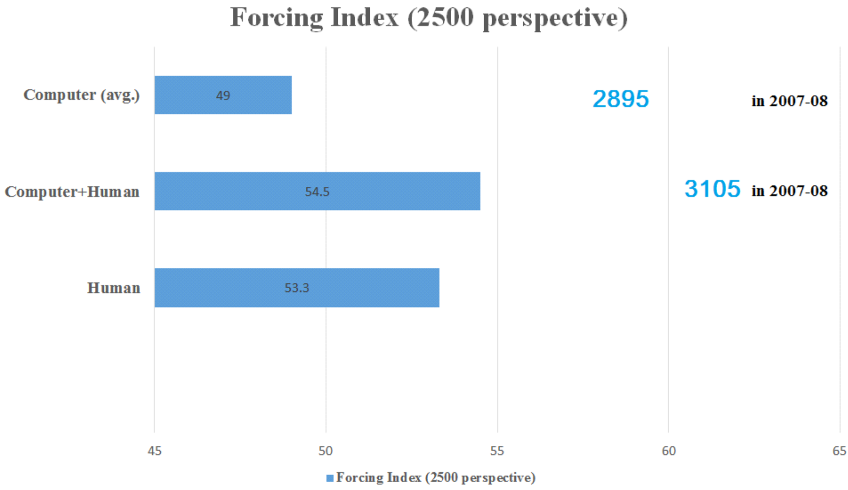
Evidently the humans called the shots.

Add Human-Computer Tandems



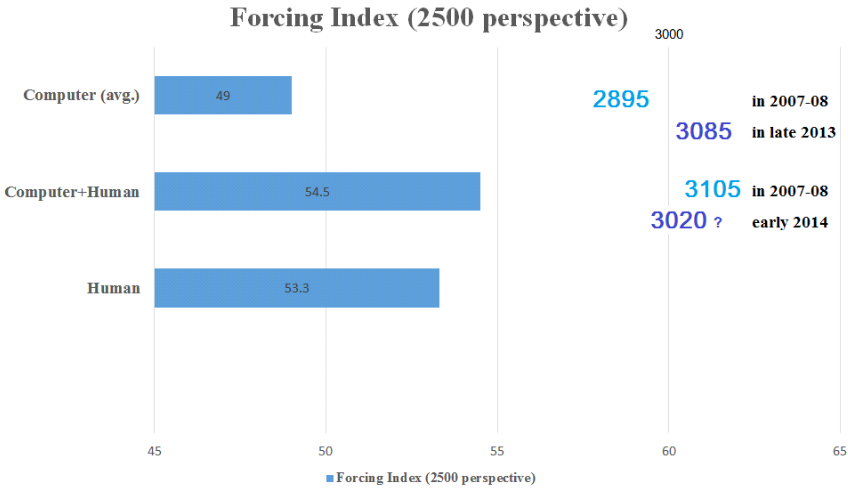
Evidently the humans called the shots. But how did they play?

2007–08 Freestyle Performance



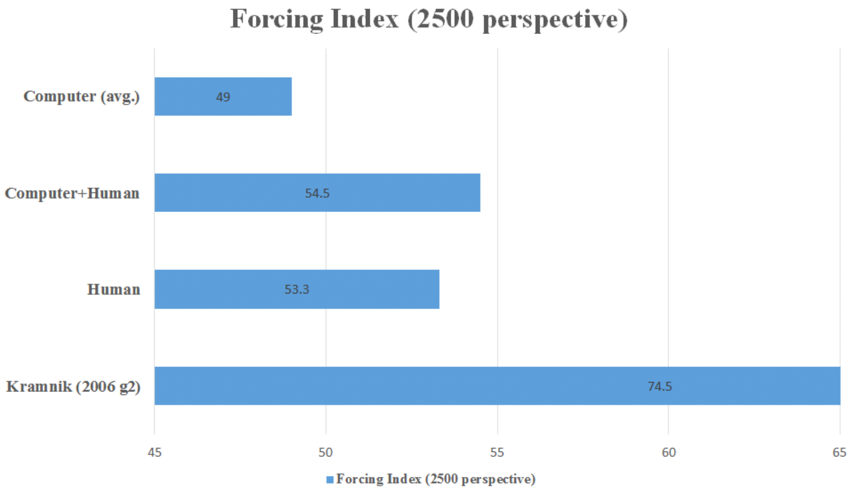
Adding 210 Elo was significant. Forcing but good teamwork.

2014 Freestyle Tournament Performance



Tandems had marginally better W-L, but quality not clear...

Add Topalov Forcing Kramnik



Last bar goes way off the chart

Is There Room to Grow?

- In *chess*, alas some hints of "no."
- If (randomizing) **3200**-level programs can score **10%** against **any** strategy, then **no** strategy can ever exceed **3550**.
- In 2010–2014 many more games between players rated **under 1600** and between **2800+** became available.
- Analysis in my model shows a **linear** relationship between rating and my Average Scaled Difference **ASD** statistic clear down to **1200** level.
- The **y-intercept** of the line is consistently near **3370**.
- But Komodo and Stockfish on 4-core PCs are rated **over 3370** on CCRL. How can this be?
  - Well, CCRL uses a 40 moves in 40 minutes time control. Other lists use other times and show ratings still in the 3100s.
- Best explanation: IPR correlates 85–90% with ASD and 10–15% with move-matching—which has y-intercept near **4500**.

Solution and Opportunities

- Hence my model projects a ceiling around **3500-3550**.
- Still not much room to grow... in chess that is.
- This may already explain the diminishing returns from adding humans... in chess.
- But the larger marriage of **Shallow but Broad** to **Deep but Narrow** that was theoretically driving the gains still has potential.
- Revisit trying to "humanize" chess programs?
- Complexity theory classifies chess as "Hard to Parallelize."
- Whether **chess endgame tables** are **"Associatively Compressible"** is an indicator worth pursuing.
- Model has many other applications: study human performance under distraction; design multiple choice tests to standards of difficulty; extend *intrinsic* Elo quality measures to other domains.

Conclusions

- Lots more potential for research and connections...
- Can use support—infrastructure, student helpers...
  - Run data with other engines Houdini, Stockfish, Komodo....
  - Run more tournaments.
  - Run to higher depths—how much does that matter?
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.

Conclusions

- Lots more potential for research and connections...
- Can use support—infrastructure, student helpers...
    - Run data with other engines Houdini, Stockfish, Komodo....
    - Run more tournaments.
    - Run to higher depths—how much does that matter?
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.

Conclusions

- Lots more potential for research and connections...
- Can use support—infrastructure, student helpers...
  - Run data with other engines Houdini, Stockfish, Komodo....
  - Run more tournaments.
  - Run to higher depths—how much does that matter?
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.
- Learn more about human decision making.

Conclusions

- Lots more potential for research and connections...
- Can use support—infrastructure, student helpers...
  - Run data with other engines Houdini, Stockfish, Komodo....
  - Run more tournaments.
  - Run to higher depths—how much does that matter?
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.
- Learn more about human decision making.
- Thus the Turing Tour comes back to the human mind.

Conclusions

- Lots more potential for research and connections...
- Can use support—infrastructure, student helpers...
    - Run data with other engines Houdini, Stockfish, Komodo....
    - Run more tournaments.
    - Run to higher depths—how much does that matter?
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.
- Learn more about human decision making.
- Thus the Turing Tour comes back to the human mind.
- **Thank you very much for the invitation.**