Skill Assessment Versus Prediction in Game Play and Cheating Detection

Kenneth W. Regan
University at Buffalo (SUNY)

Union College GAMES course, 20 May 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 really this person's performance?

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 really this person's performance?
- 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.

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.
- Current world #47 has 2700, world #100 has 2654.

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.
- Current world #47 has 2700, world #100 has 2654.
- 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.
- Current world #47 has 2700, world #100 has 2654.
- 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.
- Current world #47 has 2700, world #100 has 2654.
- 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"—and the number of class units from beginner to world champion as the Human Depth of a game.

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"—and the number of class units from beginner to world champion as the Human Depth of a game.
- From 600 to 2800 gives chess a human depth of 11. Our 8×8 checkers was 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"—and the number of class units from beginner to world champion as the Human Depth of a game.
- From 600 to 2800 gives chess a human depth of 11. Our 8×8 checkers was 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"—and the number of class units from beginner to world champion as the Human Depth of a game.
- From 600 to 2800 gives chess a human depth of 11. Our 8×8 checkers was 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 almost **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"—and the number of class units from beginner to world champion as the Human Depth of a game.
- From 600 to 2800 gives chess a human depth of 11. Our 8×8 checkers was 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 almost **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"—and the number of class units from beginner to world champion as the Human Depth of a game.
- From 600 to 2800 gives chess a human depth of 11. Our 8×8 checkers was 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 almost **3200** on all rating lists, **3380** on CCRL.
- Computers at least even at Shogi, knocking on door at Go? "Moore's Law" of Games?

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.)
- Though in a typical 9-game international event this struggles to go over 200; in a "weekend Swiss" event, less.

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.)
- Though in a typical 9-game international event this struggles to go over 200; in a "weekend Swiss" event, less.
- Can pinpoint current quality of rapidly improving player, when the Elo rating may "lag."

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.)
- Though in a typical 9-game international event this struggles to go over 200; in a "weekend Swiss" event, less.
- Can pinpoint current quality of rapidly improving player, when the Elo rating may "lag." No "K-Factor."

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.)
- Though in a typical 9-game international event this struggles to go over 200; in a "weekend Swiss" event, less.
- Can pinpoint current quality of rapidly improving player, when the Elo rating may "lag." No "K-Factor."
- "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 of decision-making situations t.
   Chess game turns

Context: Decision-Making Model at Chess

1. Domain: A set 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 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 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 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* $P(s, c, \ldots)$.
5. Main Output: Probabilities $p_{t,i}$ for $P(s, c, \ldots)$ to select option i at time t.

Context: Decision-Making Model at Chess

1. Domain: A set 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* $P(s, c, \ldots)$.
5. Main Output: Probabilities $p_{t,i}$ for $P(s, c, \ldots)$ to select option i at time t.
6. Derived Outputs:
   - Aggregate statistics: *move-match* MM, *average error* AE, ...
   - Projected confidence intervals for those statistics.
   - "Intrinsic Performance Ratings" (IPR's).

The talk then moved to webpages and Q&A...

[At this point I showed webpages linked from my professional homepage and my "Fidelity" site, articles on the "Gödel's Lost Letter" weblog, and diagrams from a paper and another talk on my site. I also showed some recent news, including a Siberian teenager caught with a hidden camera having been installed above her habitual table and an earpiece through which an accomplice fed her analyzed moves. This opened out into some questions and answers, during which I touched on some points included in slides that follow, which are selected from other talks on my site.

I did, however, forget to return to the issue of *profiling* a specific player (specifically as an element of prediction), which I had mentioned at the beginning, and the issue of chess-specific psychological factors such as good *retreating* moves being harder to find, which I have not yet fully researched.]

How the Model Operates

1. Use analysis data and parameters $s, c, \ldots$ to compute "perceived inferiorities" $x_i \in [0.0, 1.0]$ of each of N possible moves. Let $a_i = 1 - x_i$.

$$(x_1 = 0.0 \leq x_2 \leq x_3 \leq \cdots \leq x_N) \equiv (a_1 = 1.0 \geq a_2 \geq \cdots \geq a_N \approx 0)$$

How the Model Operates

1. Use analysis data and parameters $s, c, \ldots$ to compute "perceived inferiorities" $x_i \in [0.0, 1.0]$ of each of N possible moves. Let $a_i = 1 - x_i$.

$$(x_1 = 0.0 \leq x_2 \leq x_3 \leq \cdots \leq x_N) \equiv (a_1 = 1.0 \geq a_2 \geq \cdots \geq a_N \approx 0)$$

2. For a fixed function h, solve $\frac{h(p_i)}{h(p_1)} = a_i$ subject to $\sum_{i=1}^{N} p_i = 1$.

How the Model Operates

1. Use analysis data and parameters $s, c, \ldots$ to compute "perceived inferiorities" $x_i \in [0.0, 1.0]$ of each of N possible moves. Let $a_i = 1 - x_i$.

$$(x_1 = 0.0 \leq x_2 \leq x_3 \leq \cdots \leq x_N) \equiv (a_1 = 1.0 \geq a_2 \geq \cdots \geq a_N \approx 0)$$

2. For a fixed function h, solve $\frac{h(p_i)}{h(p_1)} = a_i$ subject to $\sum_{i=1}^{N} p_i = 1$.
3. It suffices to compute p_1; then $p_i = h^{-1}(a_i h(p_1))$ is relatively easy.

How the Model Operates

1. Use analysis data and parameters $s, c, \ldots$ to compute "perceived inferiorities" $x_i \in [0.0, 1.0]$ of each of N possible moves. Let $a_i = 1 - x_i$.

$$(x_1 = 0.0 \leq x_2 \leq x_3 \leq \cdots \leq x_N) \equiv (a_1 = 1.0 \geq a_2 \geq \cdots \geq a_N \approx 0)$$

2. For a fixed function h, solve $\frac{h(p_i)}{h(p_1)} = a_i$ subject to $\sum_{i=1}^{N} p_i = 1$.
3. It suffices to compute p_1; then $p_i = h^{-1}(a_i h(p_1))$ is relatively easy.
4. Model uses $a_i = e^{-(\frac{\delta_i}{s})^c}$, where δ_i is the *scaled* difference in value between the best move and the i-th best move. Also fairly cheap.

How the Model Operates

1. Use analysis data and parameters $s, c, \ldots$ to compute "perceived inferiorities" $x_i \in [0.0, 1.0]$ of each of N possible moves. Let $a_i = 1 - x_i$.

$$(x_1 = 0.0 \leq x_2 \leq x_3 \leq \cdots \leq x_N) \equiv (a_1 = 1.0 \geq a_2 \geq \cdots \geq a_N \approx 0)$$

2. For a fixed function h, solve $\frac{h(p_i)}{h(p_1)} = a_i$ subject to $\sum_{i=1}^{N} p_i = 1$.
3. It suffices to compute p_1; then $p_i = h^{-1}(a_i h(p_1))$ is relatively easy.
4. Model uses $a_i = e^{-(\frac{\delta_i}{s})^c}$, where δ_i is the *scaled* difference in value between the best move and the i-th best move. Also fairly cheap.
5. The model is **trained** by regression to find the best-fit parameters $s, c, \ldots$ on designated sets of games by players of various Elo levels.

How the Model Operates

1. Use analysis data and parameters $s, c, \ldots$ to compute "perceived inferiorities" $x_i \in [0.0, 1.0]$ of each of N possible moves. Let $a_i = 1 - x_i$.

$$(x_1 = 0.0 \leq x_2 \leq x_3 \leq \cdots \leq x_N) \equiv (a_1 = 1.0 \geq a_2 \geq \cdots \geq a_N \approx 0)$$

2. For a fixed function h, solve $\frac{h(p_i)}{h(p_1)} = a_i$ subject to $\sum_{i=1}^{N} p_i = 1$.
3. It suffices to compute p_1; then $p_i = h^{-1}(a_i h(p_1))$ is relatively easy.
4. Model uses $a_i = e^{-(\frac{\delta_i}{s})^c}$, where δ_i is the *scaled* difference in value between the best move and the i-th best move. Also fairly cheap.
5. The model is **trained** by regression to find the best-fit parameters $s, c, \ldots$ on designated sets of games by players of various Elo levels.
6. The same regression on one player's games yields his/her $s, c, \ldots$ and corresponding **IPR**; the cheating test starts with the $s, c, \ldots$ for the player's posterior rating.

Main Principle and Schematic Equation

The probability $\Pr(m_i \mid s, c, \dots)$ depends on the value of move m_i *in relation to the values of other moves*.

- Too Simple:

$$\Pr(m_i \mid s, c, \dots) \sim g(s, c, val(m_i)).$$

Doesn't take values of the other moves into account.

Main Principle and Schematic Equation

The probability $\Pr(m_i \mid s, c, \dots)$ depends on the value of move m_i *in relation to the values of other moves*.

- Too Simple:

$$\Pr(m_i \mid s, c, \dots) \sim g(s, c, val(m_i)).$$

  Doesn't take values of the other moves into account.

- Cogent answer—let m_1 be the engine's top-valued move:

$$\frac{\Pr(m_i)}{\Pr(m_1)} \sim g(s, c, val(m_1) - val(m_i)).$$

  That and $\sum_i \Pr(m_i) = 1$ minimally give the Main Principle.

Main Principle and Schematic Equation

The probability $\Pr(m_i \mid s, c, \dots)$ depends on the value of move m_i *in relation to the values of other moves*.

- Too Simple:

$$\Pr(m_i \mid s, c, \dots) \sim g(s, c, val(m_i)).$$

  Doesn't take values of the other moves into account.

- Cogent answer—let m_1 be the engine's top-valued move:

$$\frac{\Pr(m_i)}{\Pr(m_1)} \sim g(s, c, val(m_1) - val(m_i)).$$

  That and $\sum_i \Pr(m_i) = 1$ minimally give the Main Principle.

- *Much* Better answer (best?): Use $\frac{\log(1/\Pr(m_1))}{\log(1/\Pr(m_i))}$ on LHS.

Main Principle and Schematic Equation

The probability $\Pr(m_i \mid s, c, \dots)$ depends on the value of move m_i *in relation to the values of other moves*.

- Too Simple:

$$\Pr(m_i \mid s, c, \dots) \sim g(s, c, val(m_i)).$$

  Doesn't take values of the other moves into account.

- Cogent answer—let m_1 be the engine's top-valued move:

$$\frac{\Pr(m_i)}{\Pr(m_1)} \sim g(s, c, val(m_1) - val(m_i)).$$

  That and $\sum_i \Pr(m_i) = 1$ minimally give the Main Principle.

- *Much* Better answer (best?): Use $\frac{\log(1/\Pr(m_1))}{\log(1/\Pr(m_i))}$ on LHS.
- Needs Multi-PV analysis—already beyond Guid-Bratko work.
- Single-PV data on **millions** of moves shows other improvements.

The Data

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

The Data

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

The Data

- 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 taken on two quad-core home-style PC's plus a laptop. Is this "Big Data"?

The Data

- 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 taken on two quad-core home-style PC's plus a laptop. Is this "Big Data"?

“Big-Data” Aspects

"Big-Data" Aspects

1. Synthesis of two different kinds of data.
   - Single-PV data acts as scientific control for Multi-PV data.
   - Covers almost entire history of chess.
   - Shows large-scale regularities.

"Big-Data" Aspects

1. Synthesis of two different kinds of data.
   - Single-PV data acts as scientific control for Multi-PV data.
   - Covers almost entire history of chess.
   - Shows large-scale regularities.
2. Model design decisions based on large data.
   - Logarithmic scaling law
   - "58%-42% Law" for probability of equal-value moves
   - Choice of fitting methods

"Big-Data" Aspects

1. Synthesis of two different kinds of data.
   - Single-PV data acts as scientific control for Multi-PV data.
   - Covers almost entire history of chess.
   - Shows large-scale regularities.
2. Model design decisions based on large data.
   - Logarithmic scaling law
   - "58%-42% Law" for probability of equal-value moves
   - Choice of fitting methods
3. Scientific discovery beyond original intent of model.
   - Human tendencies (different from machine tendencies?)
   - Follow simple laws...

Better, and Best?

Need a general function f and a function $\delta(i)$ giving a *scaled-down* difference in value from m_1 to m_i.

$$\frac{f(\Pr_E(m_i))}{f(\Pr_E(m_1))} = g(E, \delta(i)).$$

Implemented with $f = \log$ and log-log scaling, as guided by the data.

Best model? Let *weights* w_d at different *engine depths* d reflect a player's depth of calculation. Apply above equation to evals at each depth d to define $\Pr_E(m_i, d)$. Then define:

$$\Pr_E(m_i) = \sum_d w_d \cdot \Pr_E(m_i, d).$$

This accounts for moves that *swing* in value and idea that weaker players prefer weaker moves. **In Process Now.**

Why Desire Probabilities?

- Allows to *predict* the # N of agreements with any sequence of moves m_*^t over game turns t, not just computer's first choices:

$$N = \sum_t \Pr_E(m_*^t).$$

- **and** it gives **confidence intervals** for N.
- Also predicts *aggregate error* (AE, scaled) by

$$e = \sum_t \sum_i \delta(i) \cdot \Pr_E(m_i^t).$$

  Comparing e with the *actual* error e' by a player over the same turns leads to a "virtual Elo rating" E' for those moves.
- IPR $\equiv$ "Intrinsic Performance Rating."

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.

Judgment By Your Peers

Training Sets: Multi-PV analyze games with both players rated:

- 2690–2710, in 2006–2009 and 1991–1994
- 2590–2610, "" "", extended to 2580–2620 in 1976–1979
- 2490–2510, all three times
- 2390–2410, (lower sets have over 20,000 moves)
- 2290–2310, (all sets elim. moves 1–8, moves in repetitions,
- 2190–2210, (and moves with one side $>$ 3 pawns ahead)
- Down to 1590–1610 for years 2006–2009 only.
- 2600-level set done for **all years since 1971**.

Training the Parameters

- Formula $g(E;\delta)$ is really

$$g(s,c;\delta) = \frac{1}{e^{x^c}} \quad \text{where} \quad x = \frac{\delta}{s}.$$

- s for *Sensitivity*: smaller $s \equiv$ better ability to sense small differences in value.
- c for *Consistency*: higher c reduces probability of high-δ moves (i.e., blunders).
- Full model (in progress) adds parameter d for depth of calculation.

Training the Parameters

- Formula $g(E;\delta)$ is really

$$g(s,c;\delta) = \frac{1}{e^{x^c}} \quad \text{where} \quad x = \frac{\delta}{s}.$$

- s for *Sensitivity*: smaller $s \equiv$ better ability to sense small differences in value.
- c for *Consistency*: higher c reduces probability of high-δ moves (i.e., blunders).
- Full model (in progress) adds parameter d for depth of calculation.
- Needs large-scale approximation to handle 15–20x data increase and tuning conversions between different chess engines (all in progress).

Fitting and Fighting Parameters

- For each Elo E training set, find (s, c) giving best fit.
- Can use many different fitting methods...
  - Can compare methods...
  - Whole separate topic...

Fitting and Fighting Parameters

- For each Elo E training set, find (s, c) giving best fit.
- Can use many different fitting methods...
  - Can compare methods...
  - Whole separate topic...
  - Max-Likelihood does *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.

Some IPRs—Historical and Current

- Magnus Carlsen:
  - 2983 at London 2011 (Kramnik 2857, Aronian 2838, Nakamura only 2452).
  - 2855 at Biel 2012.
- Bobby Fischer:
  - 2921 over all 3 Candidates' Matches in 1971.
  - 2650 vs. Spassky in 1972 (Spassky 2643).
  - 2724 vs. Spassky in 1992 (Spassky 2659).
- Hou Yifan: 2971 vs. Humpy Honeru (2683) in Nov. 2011.
- Paul Morphy: 2344 in 59 most impt. games, 2124 vs. Anderssen.
- Capablanca: 2936 at New York 1927.
- Alekhine: 2812 in 1927 WC match over Capa (2730).

Results and Implications for Human Thinking

1. Sensitivity to small changes in the value of moves.
2. Degrees of sensitivity to changes in value at different depths of search.
3. Tangibly greater error in positions where one side has even a slight advantage.
4. Natural variability in performance, which we argue is intrinsic and unavoidable.
5. Correspondences with results in item-response theory and psychometric test scoring.
6. Quality of human-computer teams compared to computers or humans playing separately.

1. 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.

1. 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.

1. 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.

1. 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.

1. 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 "When is a Law Natural?"

1. 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 "When is a Law Natural?"
- Stockfish versions round evals to nearest 0.04 or 0.02.

1. 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 "When is a Law Natural?"
- Stockfish versions round evals to nearest 0.04 or 0.02.
- Relation to slime molds and other "semi-Brownian" systems?

2. Depth-of-Phenomenon Effects (ongoing)

- Tied-top law extends to 3, 4, tied moves in similar 58% ratio of choice to the next.

2. Depth-of-Phenomenon Effects (ongoing)

- Tied-top law extends to 3, 4, tied moves in similar 58% ratio of choice to the next.
- Lead moves tend to have been higher at lower depths. Does this explain it?

2. Depth-of-Phenomenon Effects (ongoing)

- Tied-top law extends to 3, 4, tied moves in similar 58% ratio of choice to the next.
- Lead moves tend to have been higher at lower depths. Does this explain it?
- How less likely to be found is a move whose value **"Swings Up"** only at high depth, compared to one having the same value at all depths?

2. Depth-of-Phenomenon Effects (ongoing)

- Tied-top law extends to 3, 4, tied moves in similar 58% ratio of choice to the next.
- Lead moves tend to have been higher at lower depths. Does this explain it?
- How less likely to be found is a move whose value **"Swings Up"** only at high depth, compared to one having the same value at all depths?
- How more likely to be played is a **"Swing Down"** move—a *trap*?

2. Depth-of-Phenomenon Effects (ongoing)

- Tied-top law extends to 3, 4, tied moves in similar 58% ratio of choice to the next.
- Lead moves tend to have been higher at lower depths. Does this explain it?
- How less likely to be found is a move whose value **"Swings Up"** only at high depth, compared to one having the same value at all depths?
- How more likely to be played is a **"Swing Down"** move—a *trap*?
- Goal is to develop a **Challenge Quotient** based on how much trappy play a player sets for the opponent

2. Depth-of-Phenomenon Effects (ongoing)

- Tied-top law extends to 3, 4, tied moves in similar 58% ratio of choice to the next.
- Lead moves tend to have been higher at lower depths. Does this explain it?
- How less likely to be found is a move whose value **"Swings Up"** only at high depth, compared to one having the same value at all depths?
- How more likely to be played is a **"Swing Down"** move—a *trap*?
- Goal is to develop a **Challenge Quotient** based on how much trappy play a player sets for the opponent—and emself.

2. Depth-of-Phenomenon Effects (ongoing)

- Tied-top law extends to 3, 4, tied moves in similar 58% ratio of choice to the next.
- Lead moves tend to have been higher at lower depths. Does this explain it?
- How less likely to be found is a move whose value **"Swings Up"** only at high depth, compared to one having the same value at all depths?
- How more likely to be played is a **"Swing Down"** move—a *trap*?
- 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.

3. The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

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

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

3. The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

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

  balances evals well for Rybka, with c very near 1.0.
- A mix of three factors?

3. The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

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

  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.

3. The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

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

  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)

3. The Imbalance-Error Phenomenon

- [show data]
- The *metric correction*

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

  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 reval-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 reval-error curve to shift in games between unequally-rated players.
- *Will need many such games*, if not prevented by cause C.

C. Similar Phenomenon in Computer-Played Games

- [show data from new "Computer and Freestyle Study."]

C. Similar Phenomenon in Computer-Played Games

- [show data from new "Computer and Freestyle Study."]
- [Segue to item 6. in outline.]

4. Is Savielly Tartakover Right?

The winner is the player who makes the next-to-last blunder.

- We like to think chess is about Deep Strategy.
- This helps, but is it statistically dominated by blunders?
- Recent Examples:
  - USA-Russia and USA-China matches at 2012 Olympiad.
  - Gelfand-Anand 2012 Rapid playoff.
- My Average Error (AE) stat shows a tight linear fit to Elo rating.
- Full investigation will need ANOVA (analysis of variance).

5. Variance in Performance, and Motivation?

- Let's say I am 2400 facing 2600 player.
- My expectation is 25%. Maybe:
  - 60% win for stronger player.
  - 30% draw.
  - 10% chance of win for me.
- In **12-game match**, maybe under 1% chance of winning if we are random.
- But my model's intrinsic error bars are often 200 points wide over 9–12 games.
- Suggests to take event not game as the unit.
- How can we be motivated for events?

7. Procrastination...

- (Show graph of AE climbing to Move 40, then falling.)
- Aug. 2012 *New In Chess*, Kramnik-Grischuk, Moscow Tal Mem.
  - King's Indian: 12. Bf3!? then 13. Bg2 N (novelty)
  - "Grischuk was already in some time pressure."
- IPR for Astana World Blitz (cat. 19, 2715) 2135.
- IPR for Amber 2010+2011 (cat. 20+21): 2545.
- Can players be coached to play like the young Anand?

8. Human Skill Increasing Over Time?

- In 1970s, two 2700+ players: Fischer and Karpov. In 1981: none!
- Sep. 2012 list, 44 2700+ players. Rating Inflation?
- My results:
- 1976–1979 vs. 1991–1994 vs. 2006–2009: Little or no difference in IPR at all rating levels.
- 2600 level, 1971–present:
  - Can argue 30-pt. IPR difference between 1980's and now.
  - Difference measured at 16 pts. using 4-yr. moving averages, 10-year blocks.
  - Explainable by faster time controls, no adjournments?
- Single-PV AE stat in **all** Cat 11+ RRs since 1971 hints at mild **deflation**.
- Moves 17–32 show similar results. Hence not just due to better opening prep?
- Increasing skill consistent with Olympics results.

9. Are We Reliable?

- One blunder in 200 moves can "ruin" a tournament.
- But we were reliable 99.5% of the time.
- Exponential $g(s, c)$ curve fits better than inverse-poly ones.
- Contrary to my "Black Swan" expectation.
- But we are even more reliable if we can use a computer...
- (show PAL/CSS Freestyle stats if time).

10. Not Just About Chess?

- *Only chess aspect of entire work is the evaluations coming from chess engines.*
- No special chess-knowledge, no "style" (except as reflected in fitted s, c, d).
- General Problem: **Converting Utilities Into Probabilities** for *colordarkredfallible agents.*
- Framework applies to multiple-choice tests, now prevalent in online courses.
- Alternative to current psychometric measures?
- Issue: Idea of "best move" at chess is the same for all human players, but "best move" in sports may depend on natural talent.

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.**