CSE 199 Data, AI, and Society Lecture 5

Data Science Examples and Foibles

Kenneth W. Regan

CSE199, Fall 2026

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: $\text{Run_Likelihood} = a + b{\cdot}(\text{Pitch_Count})$.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: Run_Likelihood $= a + b\cdot$(Pitch_Count).
- [show graphs from FanGraphs article, bottom of page. And this.]

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: Run_Likelihood $= a + b\cdot$(Pitch_Count).
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: Run_Likelihood $= a + b\cdot$(Pitch_Count).
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: Run_Likelihood $= a + b\cdot$(Pitch_Count).
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.**Slope** of **regression line**.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: Run_Likelihood $= a + b\cdot$(Pitch_Count).
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.**Slope** of **regression line**.
- Strength of **correlation** is measured by R—or its square, R^2.

1. Linear Models

Represent a targeted **dependent variable** as a *linear* function of one or more **independent variables**. Schematically:

$$\begin{aligned} Y &= a + bX \qquad \text{or} \\ Z &= a + bX + cY \qquad \text{(etc.)} \end{aligned}$$

- E.g. for a baseball pitcher: Run_Likelihood $= a + b\cdot$(Pitch_Count).
- [show graphs from FanGraphs article, bottom of page. And this.]
- Goal is to fit the **coefficients** $a, b, \ldots$ by **linear regression** on available data for the modeled variables.
- Magnitude of b gives strength of effect.**Slope** of **regression line**.
- Strength of **correlation** is measured by R—or its square, R^2.
- But whether this amounts to **causation** may remain problematic.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level. (Note NFElo, Nate Silver.)

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level. (Note NFElo, Nate Silver.)

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level. (Note NFElo, Nate Silver.)

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level. (Note NFElo, Nate Silver.)

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**
- Difference matters: **200 Elo** $\approx$ **75%** favorite, **70 Elo** $\approx$ **60%**.

I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level. (Note NFElo, Nate Silver.)

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**
- Difference matters: **200 Elo** $\approx$ **75%** favorite, **70 Elo** $\approx$ **60%**.

I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move. Variant: **EV** includes moves of equal-optimal value, if any.

Chess Example

My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level. (Note NFElo, Nate Silver.)

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**
- Difference matters: **200 Elo** $\approx$ **75%** favorite, **70 Elo** $\approx$ **60%**.

I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move. Variant: **EV** includes moves of equal-optimal value, if any.
- The **average error** per move, measured as **ACPL** for "average centipawn loss."

Chess Example

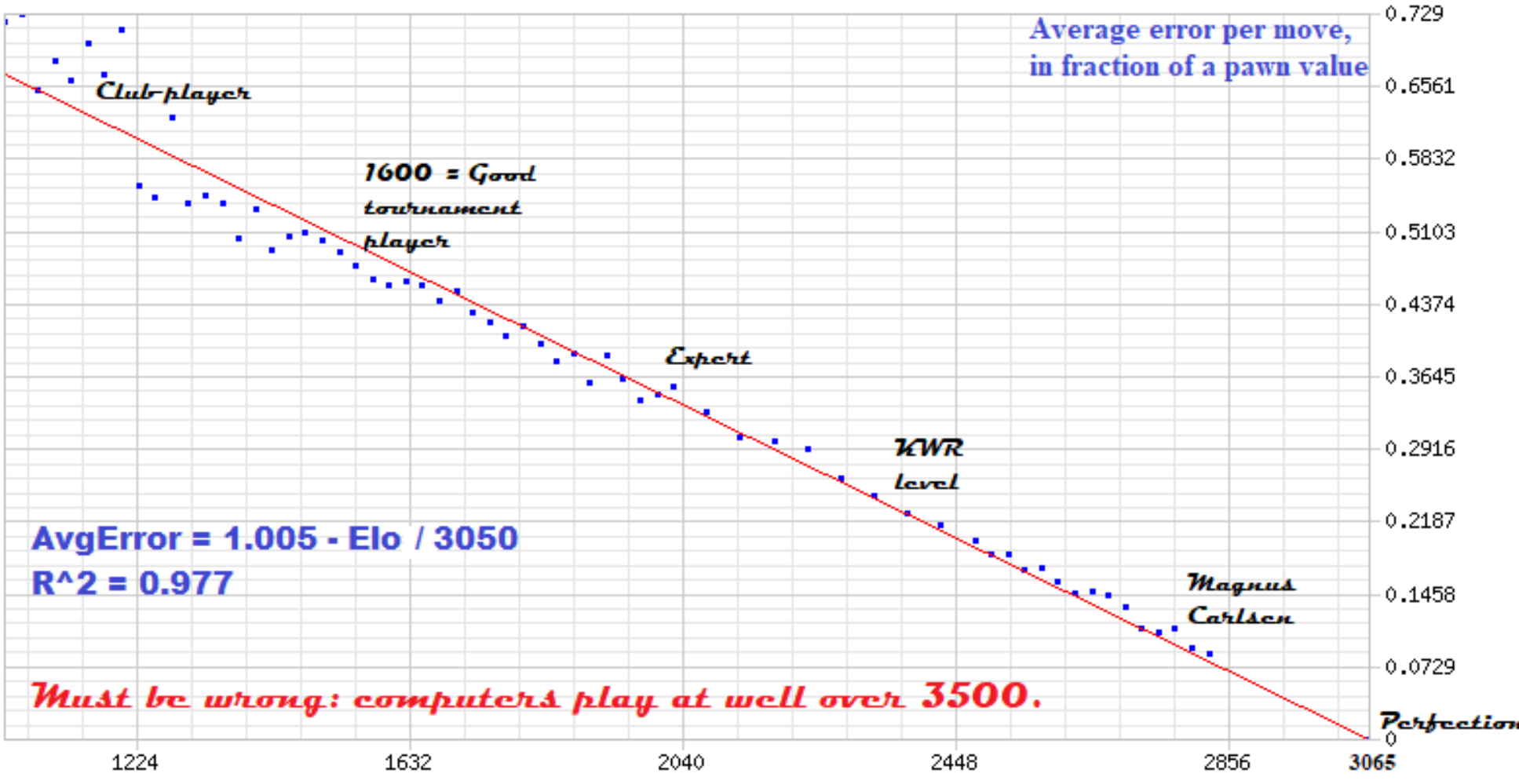
My chess cheating work starts with a player P's **Elo rating** as the main independent variable giving P's skill level. (Note NFElo, Nate Silver.)

- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**
- Difference matters: **200 Elo** $\approx$ **75%** favorite, **70 Elo** $\approx$ **60%**.

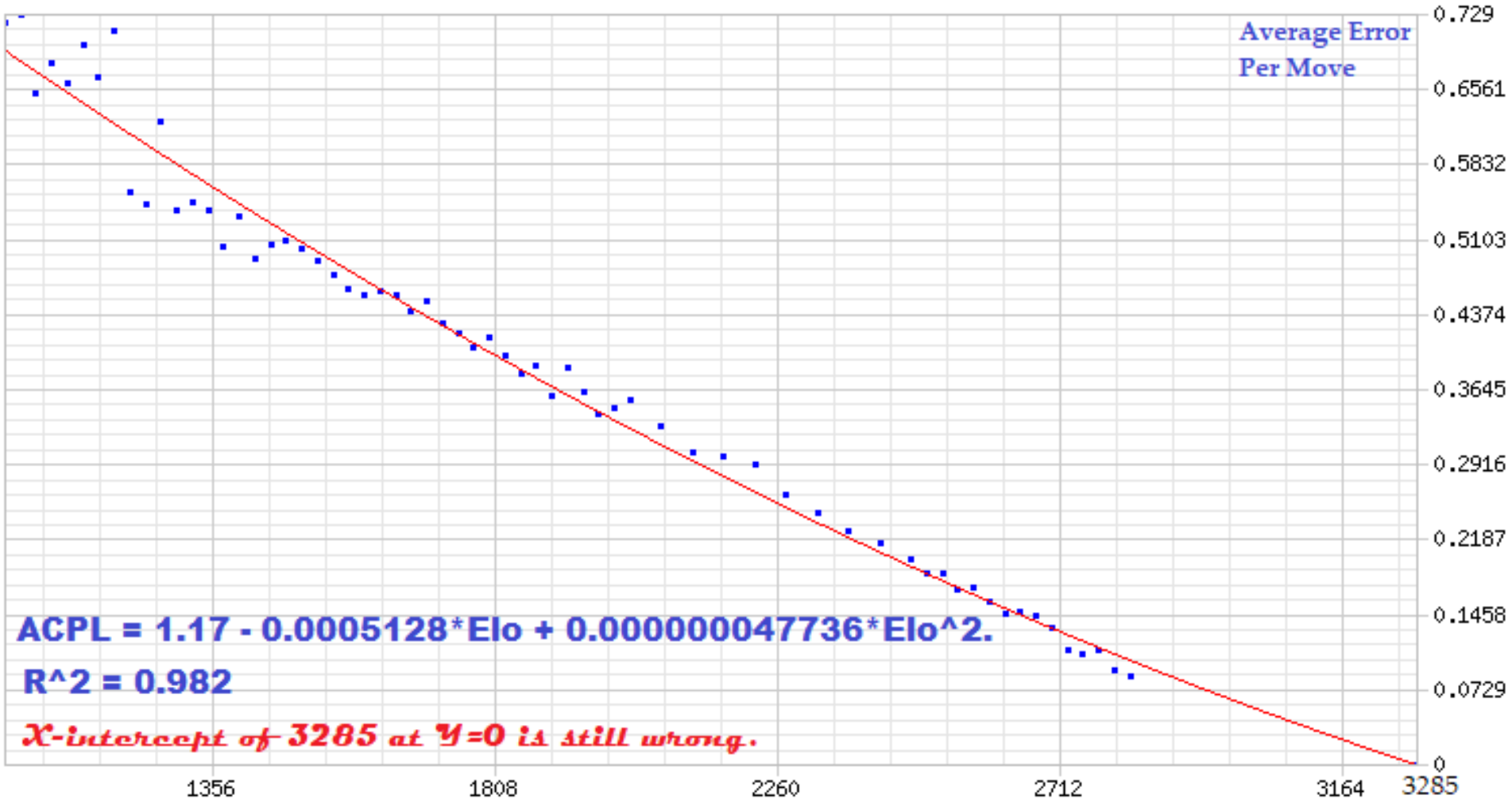
I use one or more of the following as **proxy** variables for possible illicit computer use (which I can't observe directly):

- The frequency **T1** of P playing an engine's 1st-listed move. Variant: **EV** includes moves of equal-optimal value, if any.
- The **average error** per move, measured as **ACPL** for "average centipawn loss." My **ASD** variant *scales* loss according to the overall advantage.

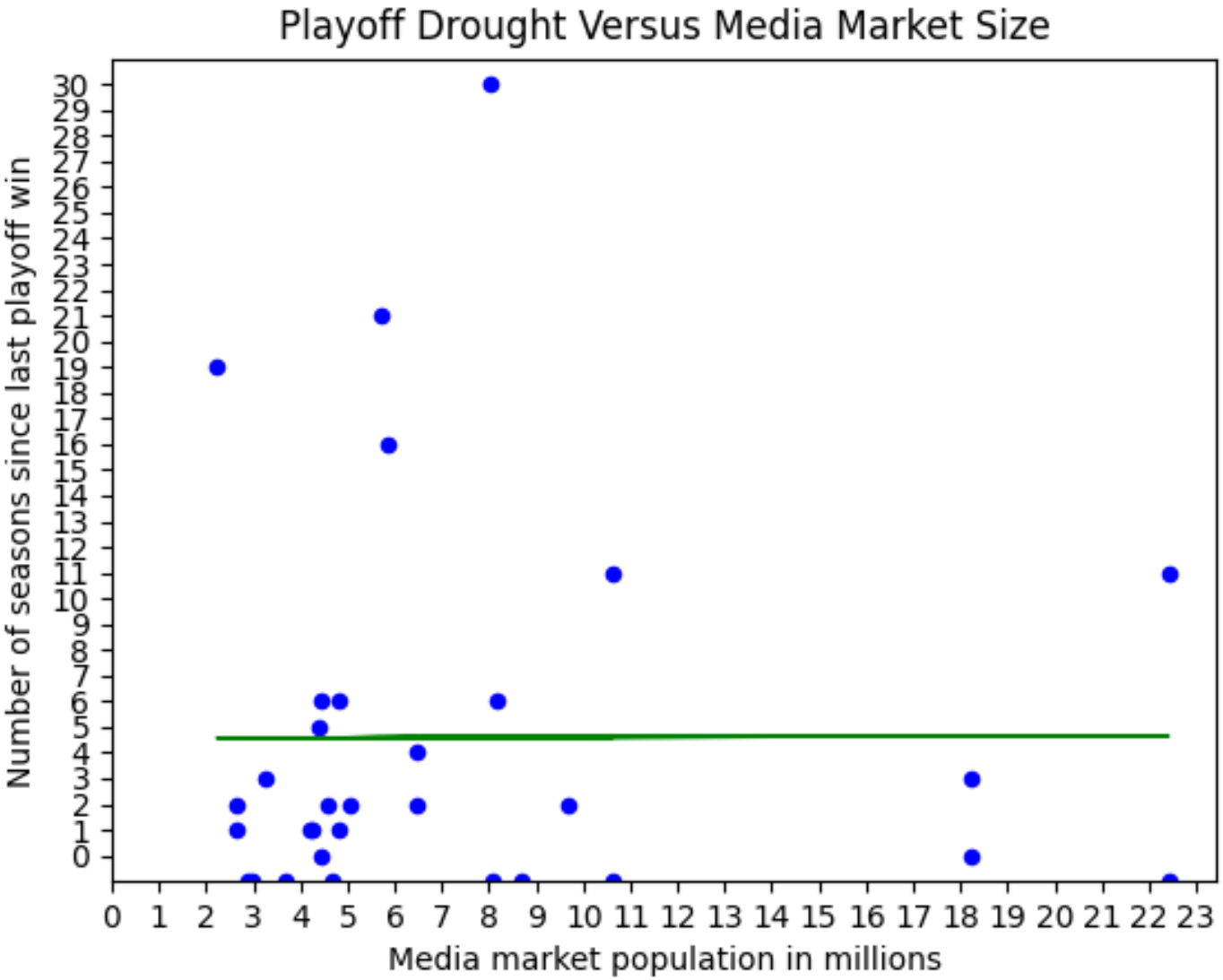
Linear Model: ACPL $= a + b \cdot$Elo Rating



Quadratic Fit—Only Marginally Better



A Desired Null Result? (data left in 2023)



Issues With Linear Models

Issues With Linear Models

- Do we have all the relevant input/independent factors?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
  - If **one player's** performance is way above the line, = cheating?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
  - If **one player's** performance is way above the line, = cheating?
  - What if ey had really easy positions?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
  - If **one player's** performance is way above the line, = cheating?
  - What if ey had really easy positions? I.e., does inference *scale down?*

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
  - If **one player's** performance is way above the line, = cheating?
  - What if ey had really easy positions? I.e., does inference *scale down?*
  - Can you use the regression line to **predict** on (i.e., *bet on*) a given game turn whether P will play the move the computer likes?

Issues With Linear Models

- Do we have all the relevant input/independent factors?
- Can we intepret **cause-and-effect**, or merely correlation?
- Are there hidden factors that might be **biasing** results?
- Does our output variable really signify what we want it to?
- E.g., does ACPL = the **skill** of a chess *performance*? Does ACPL=0 mean perfect play?
  - The corresponding Elo rating is only **3065** in the linear model, **3285** in the quadratic fit—both well short of perfection.
  - Hence must be more to chess **skill**—ACPL is at most **accuracy**.
- Regression by itself says little about what **judgments** it supports:
  - If **one player's** performance is way above the line, = cheating?
  - What if ey had really easy positions? I.e., does inference *scale down?*
  - Can you use the regression line to **predict** on (i.e., *bet on*) a given game turn whether P will play the move the computer likes?
- Need a **probabilistic model** that individuates game positions.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)
  - Much like chess ratings R—but student aptitudes aren't published.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)
  - Much like chess ratings R—but student aptitudes aren't published.
  - Nevertheless, judging the difficulty/likelihoods of questions is needed to estimate the grading curve and so design a **fair** exam.

2. Probabilistic Modeling

Working Definition: The practice of assigning probabilities p_i to unknown outcomes i and then reasoning and acting based on those probabilities being correct. **Examples:**

- Gauging opponent's likely choices in Rock-Paper-Scissors.
- Weather forecasts: $p =$ local chance of rain, etc.
- For each question on a test, the likelihoods p_i that students in the class will get it right.
  - There is a hidden variable here: the **aptitudes** or the students. (Including how much they've done HWs and studied.)
  - Much like chess ratings R—but student aptitudes aren't published.
  - Nevertheless, judging the difficulty/likelihoods of questions is needed to estimate the grading curve and so design a **fair** exam.
  - It helps to be *confident* that the class won't just bomb your exam.

Grading Probabilistic Forecasters

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.
- Who was the more-accurate forecaster *on these outcomes*, **A** or **B**?

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.
- Who was the more-accurate forecaster *on these outcomes*, **A** or **B**?
- From "hedge-fund perspective," it was B: $14/18 = 77.8\%$.

Grading Probabilistic Forecasters

- Suppose you are playing an NFL survivor pool.
- You rely on two odds-giving experts: **Forecaster A** and **Forecaster B**.
- You pick only games where they say a team has **75–80%** chance of winning. Hard to get stronger odds, plus you have to pick a different team each week.
- You follow picks by **Forecaster A** and are right all 18 weeks—you win!
- Whereas with Forecaster B you would have lost **4** times.
- Who was the more-accurate forecaster *on these outcomes*, **A** or **B**?
- From "hedge-fund perspective," it was B: $14/18 = 77.8\%$.
- My chess model's probability forecasts are similarly accurate within $\sim 5\%$.

3. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$

3. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.

3. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.

3. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

3. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.

3. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.

3. Predictive Analytics

Working Definition: (*I* require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.
- Homes being close together does not affect the expectation but does widen the confidence interval.

3. Predictive Analytics

Working Definition: (I require bullet 4) Means that the model:

1. Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
2. Assigns to each m_j a probability p_j.
3. Projects risk/reward quantities associated to the outcomes.
4. Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.
- Homes being close together does not affect the expectation but does widen the confidence interval.

In my model, the m_j are possible moves in chess positions.

Chess and Tests—With Partial Credits (Or LLMs?)

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

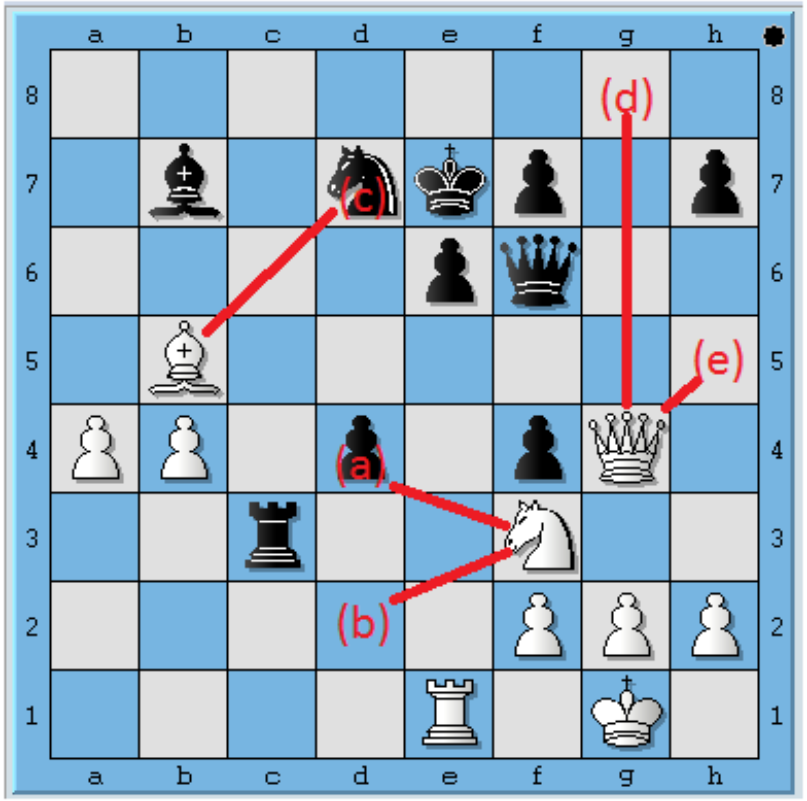
(c) proliferation . . blighted

(d) destruction . . disputed

(e) disappearance . . frustrated

(source: itunes.apple.com)

=



Chess and Tests—With Partial Credits (Or LLMs?)

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

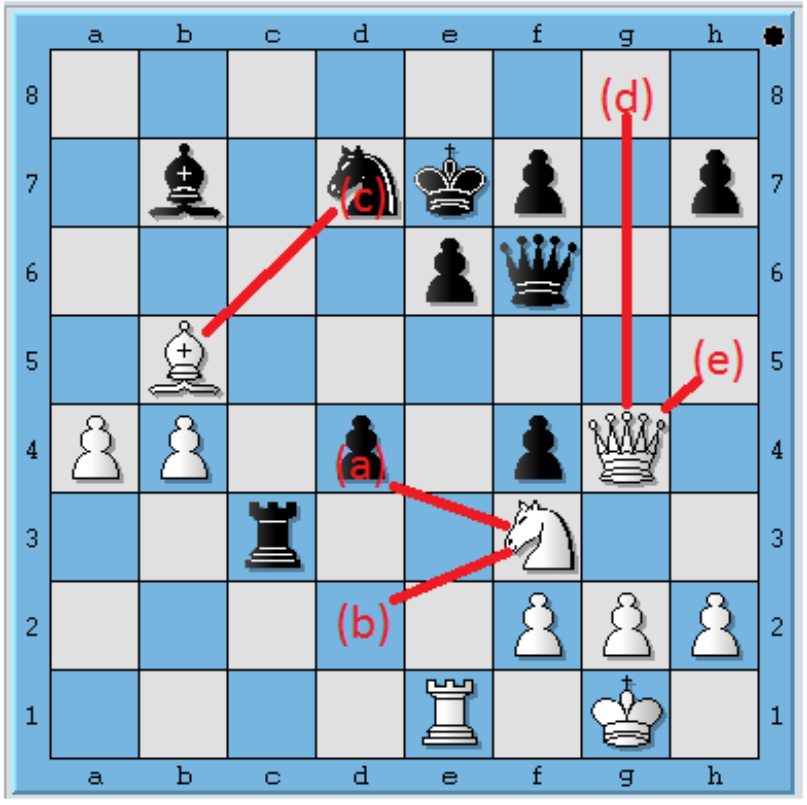
(c) proliferation . . blighted

(d) destruction . . disputed

(e) disappearance . . frustrated

(source: itunes.apple.com)

=



Here (b,c) are **equal-optimal** choices, (a) is bad, but (d) and (e) are reasonable—worth part credit.

Chess and Tests—With Partial Credits (Or LLMs?)

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

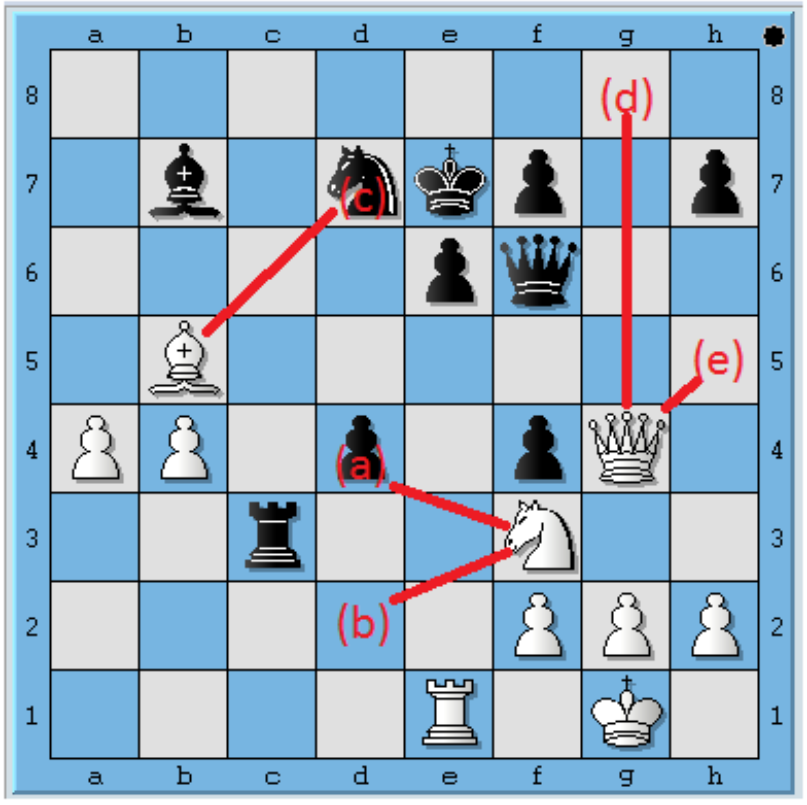
(c) proliferation . . blighted

(d) destruction . . disputed

(e) disappearance . . frustrated

(source: itunes.apple.com)

=



Here (b,c) are **equal-optimal** choices, (a) is bad, but (d) and (e) are reasonable—worth part credit.

4. Preference Aggregation: Voting, Ranking, Rating

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

4. Preference Aggregation: Voting, Ranking, Rating

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:

4. Preference Aggregation: Voting, Ranking, Rating

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.

4. Preference Aggregation: Voting, Ranking, Rating

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.

4. Preference Aggregation: Voting, Ranking, Rating

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.
- **Kenneth Arrow, 1950s:** No way to fix except to allow minority rule, dictatorship, or weighting votes unequally via **ratings**.

4. Preference Aggregation: Voting, Ranking, Rating

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.
- **Kenneth Arrow, 1950s:** No way to fix except to allow minority rule, dictatorship, or weighting votes unequally via **ratings**.
- A Likert scale converts strong/weak preferences to grainy numbers.

4. Preference Aggregation: Voting, Ranking, Rating

Thorny Question:

> How shall we combine N individual votes into one result? Is there always a clear result? Or at least a clear best way to get a result?

- Becomes gnarly with 3 or more candidates. Condorcet's Paradox:
  - Suppose voter V1 prefers $A > B > C$, voter V2 prefers $B > C > A$, and voter V3 prefers $C > A > B$.
  - Then 2/3 prefer A to B, 2/3 prefer B to C, and 2/3 prefer C to A.
- **Kenneth Arrow, 1950s:** No way to fix except to allow minority rule, dictatorship, or weighting votes unequally via **ratings**.
- A Likert scale converts strong/weak preferences to grainy numbers.
- Rating numbers convert into **rankings**. E.g. chess players by Elo.

Polling

- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.

Polling

- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random.

Polling

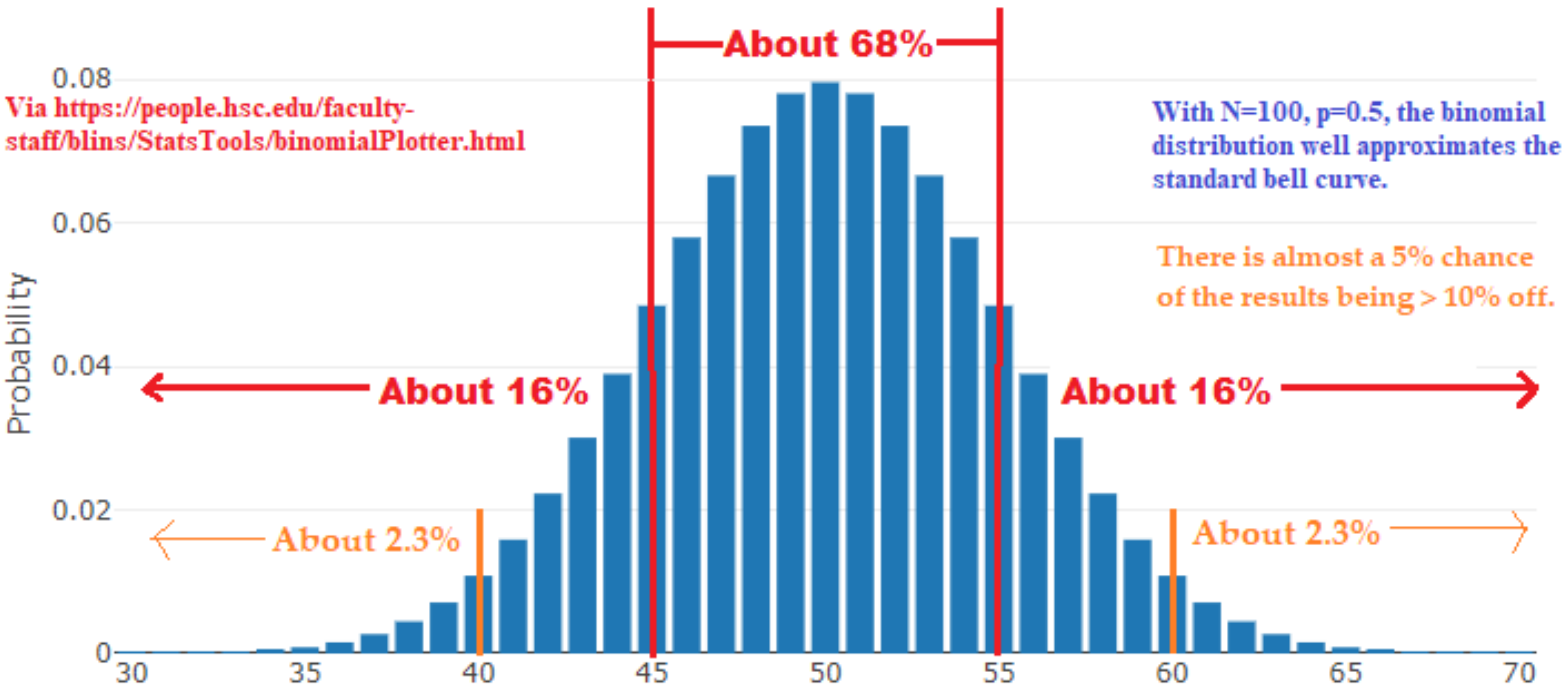
- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random. Like flipping a fair coin $N = 100$ times.

Polling

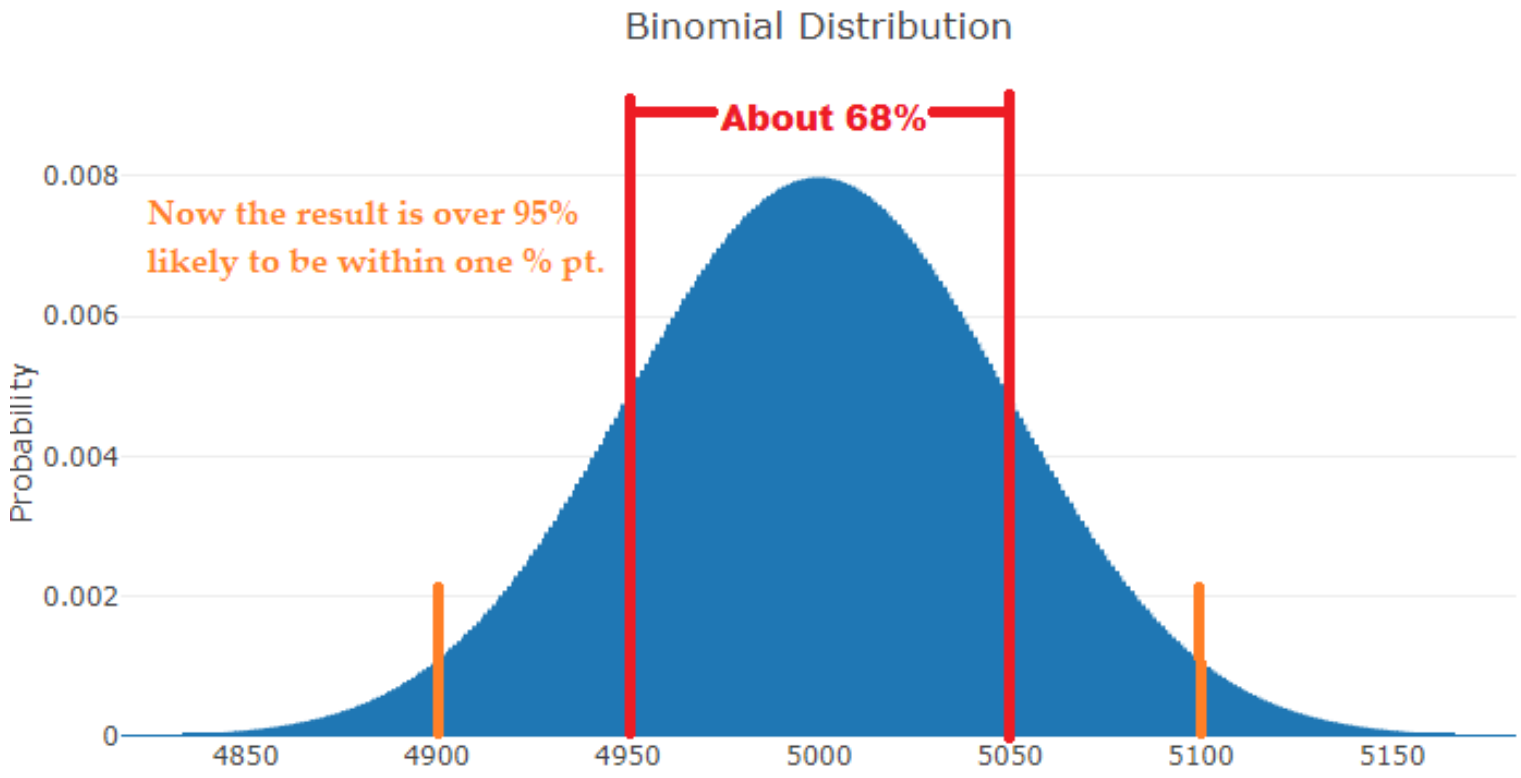
- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random. Like flipping a fair coin $N = 100$ times.
- **Mean** $\mu = 50$ and **std. deviation** $\sigma = \sqrt{Np(1-p)} = \sqrt{25} = 5$.

Polling

- Suppose the **ground truth** is that 1 million people favor A and 1 million favor B. A **50-50** election.
- Poll 100 people at random. Like flipping a fair coin $N = 100$ times.
- **Mean** $\mu = 50$ and **std. deviation** $\sigma = \sqrt{Np(1-p)} = \sqrt{25} = 5$.
- Almost **one-third** chance poll results will be $< 45\%$ or $> 55\%$.

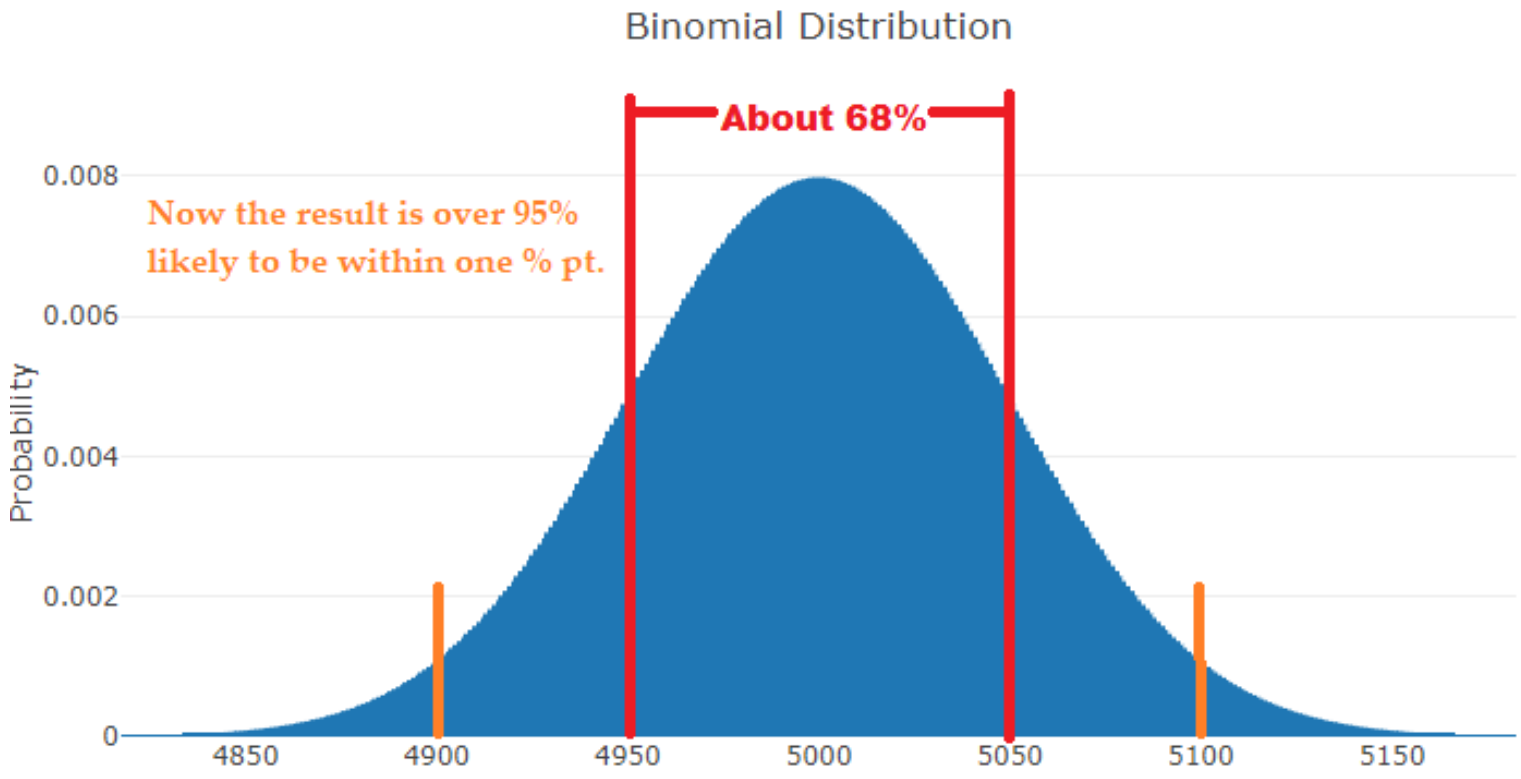


Try Polling 10,000 People Instead



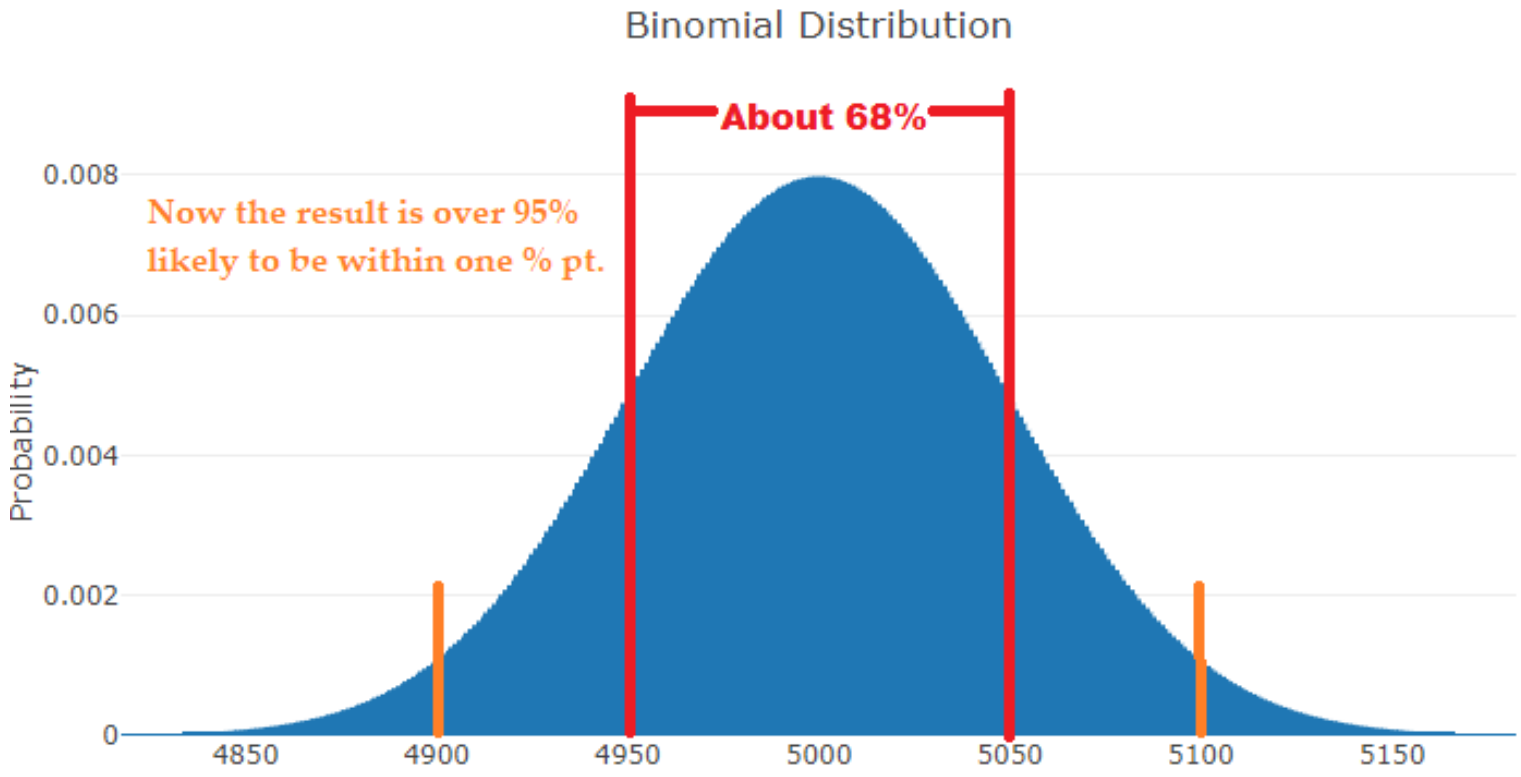
- Better—but polling 10,000 *es caro*!

Try Polling 10,000 People Instead



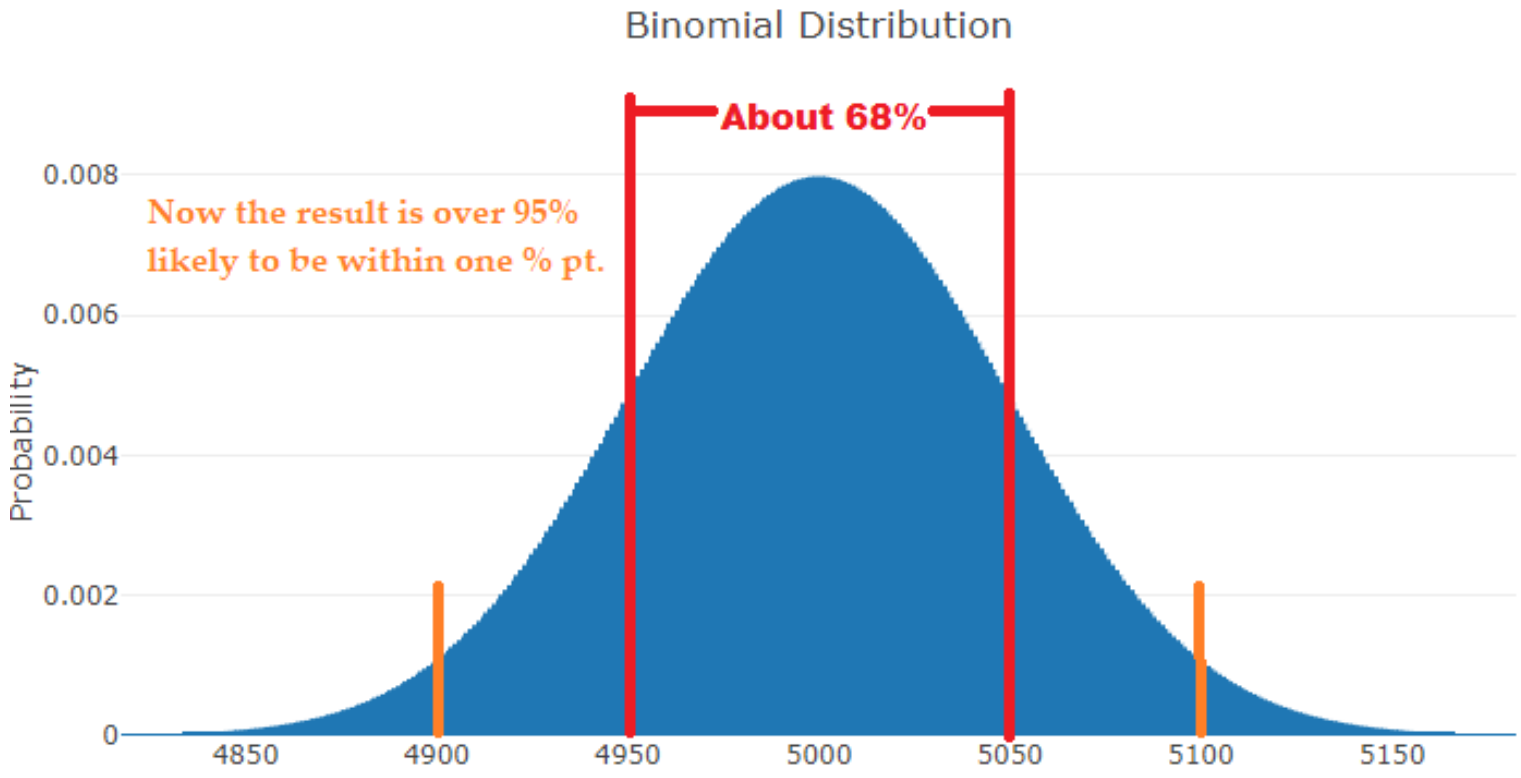
- Better—but polling 10,000 *es caro*! Still $\pm 1\%$ **margin of error**.

Try Polling 10,000 People Instead



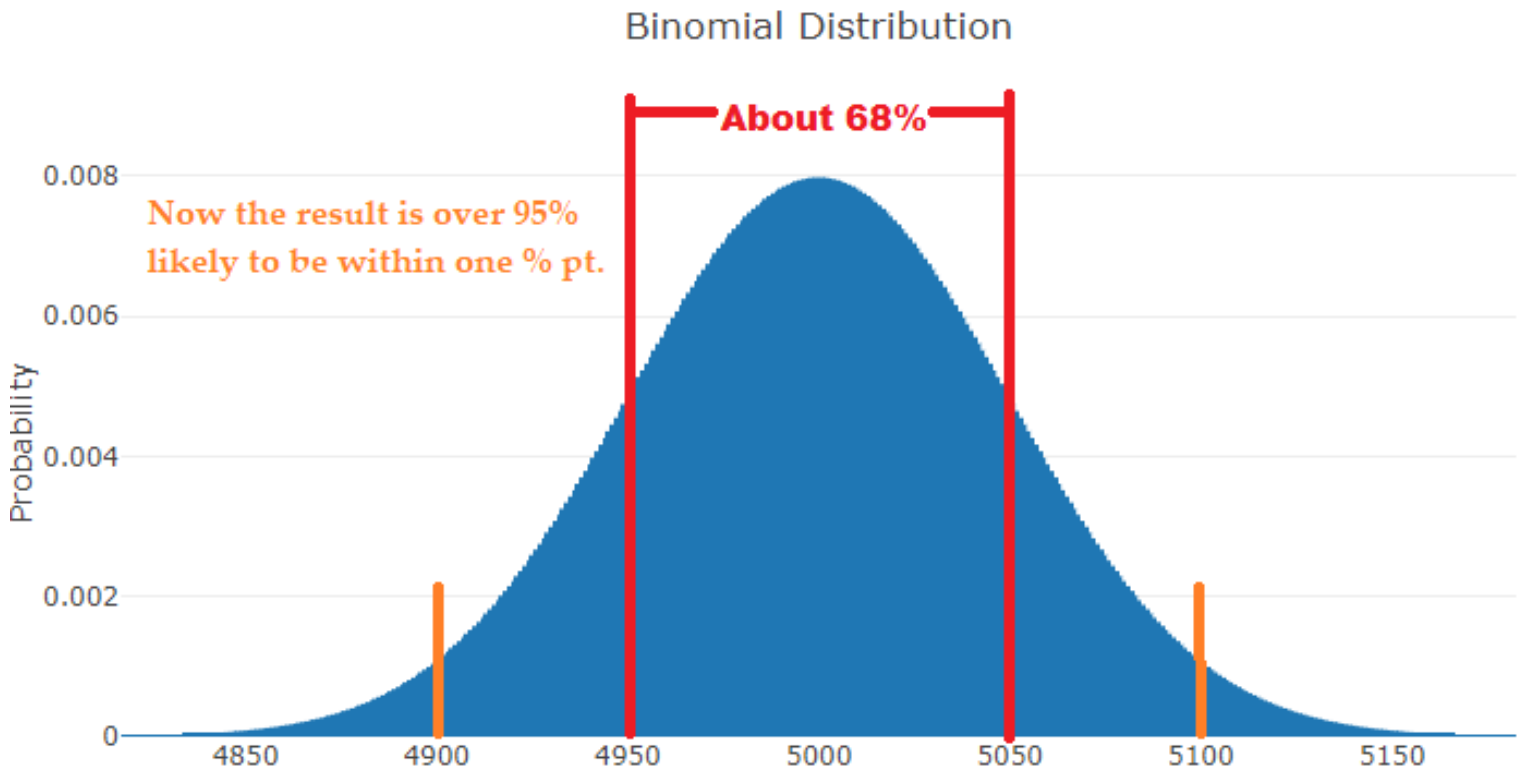
- Better—but polling 10,000 *es caro*! Still ±1% **margin of error**.
- Compromise: poll 1,024. Then $\sigma = \sqrt{Np(1-p)} = \sqrt{256} = 16$.

Try Polling 10,000 People Instead



- Better—but polling 10,000 *es caro*! Still ±1% **margin of error**.
- Compromise: poll 1,024. Then $\sigma = \sqrt{Np(1-p)} = \sqrt{256} = 16$.
- Gives $\pm 2\sigma/1024 = \pm\frac{1}{32} \approx \pm 3.1\%$ margin of "95%" confidence.

Try Polling 10,000 People Instead



- Better—but polling 10,000 *es caro*! Still ±1% **margin of error**.
- Compromise: poll 1,024. Then $\sigma = \sqrt{Np(1-p)} = \sqrt{256} = 16$.
- Gives $\pm 2\sigma/1024 = \pm\frac{1}{32} \approx \pm 3.1\%$ margin of "95%" confidence.
- So results 47%-to-53% count as "statistically tied" (yuck).

Precision, Accuracy, and a "Murphy's Law"

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both. For a poll, increasing N improves precision, **but** subject to this "law of diminishing returns":

> Precision improves only in proportion to $\sqrt{N}$, whereas inaccuracy from *skew* scales as N.

Thus using 100× more people brought only 10× more precision, but would keep percentage error—which is $\frac{skew}{N}$—at the same rate.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both. For a poll, increasing N improves precision, **but** subject to this "law of diminishing returns":

> Precision improves only in proportion to $\sqrt{N}$, whereas inaccuracy from *skew* scales as N.

Thus using 100× more people brought only 10× more precision, but would keep percentage error—which is $\frac{skew}{N}$—at the same rate. Your 10× narrower confidence intervals would give you misplaced confidence in a wrong result.

Precision, Accuracy, and a "Murphy's Law"

- **Accuracy** means how close your projection is to the truth.
- **Precision** means how narrow is your range of uncertainty.
- Terms often lumped together but are completely separate. Pictures.
- You want to improve both. For a poll, increasing N improves precision, **but** subject to this "law of diminishing returns":

> Precision improves only in proportion to $\sqrt{N}$, whereas inaccuracy from *skew* scales as N.

Thus using 100× more people brought only 10× more precision, but would keep percentage error—which is $\frac{skew}{N}$—at the same rate. Your 10× narrower confidence intervals would give you misplaced confidence in a wrong result. **(Your HW will emphasize detecting possible sources of bias/inaccuracy and how to manage them.)**

Interpretation and Poll Aggregation

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic. New Site.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic. New Site.
- I "poll' chess tournaments where proposition A = player underperforms eir Elo rating, B = the player overperforms it.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic. New Site.
- I "poll' chess tournaments where proposition A = player underperforms eir Elo rating, B = the player overperforms it.
- Can both detect and rule out large-scale cheating.

Interpretation and Poll Aggregation

- If you get a result of 52%, similar math presuming $p = 0.52$ gives a "95%" confidence interval of about 49% to 55%.
- Since the interval nips under 50%, a "normal polling error" could mean you are really behind.
- Or you could be in 55% "landslide" territory.
- If you had $4x$ as many polls, you'd cut your error margins in half...
- **Poll Aggregation** does this. RealClearPolitics was first in 2002, but it was Nate Silver's high accuracy in 2008 when he helmed FiveThirtyEight that made this seem like magic. New Site.
- I "poll' chess tournaments where proposition A = player underperforms eir Elo rating, B = the player overperforms it.
- Can both detect and rule out large-scale cheating.
- Aggregating tournaments checks my formulas for accuracy and bias.

Extra Slides In Case of Time (on sixth day, e.g.)

Some Hard Data Challenges (still—?)

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.
  - Super-Recognizers.com

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.
  - Super-Recognizers.com
- Scene analysis in greater generality.

Some Hard Data Challenges (still—?)

- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding phrases.
- Election status (might not be well-defined).
- Identifying faces conclusively.
  - Apple iPhone X has bet on it.
  - Scotland Yard employs special humans to examine photos.
  - Super-Recognizers.com
- Scene analysis in greater generality.
- General anomaly alert systems.

The Replication Crisis

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study...

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.
  - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.
  - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.
- Just like if I focused on one high player in the Chess Olympiad—ignoring that there were almost 1,000 other players.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.
  - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.
- Just like if I focused on one high player in the Chess Olympiad—ignoring that there were almost 1,000 other players.
- When others try to replicate the study, the ground truth proves out.

The Replication Crisis

- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.
  - The ground truth is "no effect"—analogous to our 50-50 election.
  - But $\sim$ our poll analogy, one study randomly gets results outside the margin of error, i.e., "$> 2\sigma$."
  - This is the academic threshold to publish, so they do.
  - The others who get "$< 2\sigma$" (or even $< -2\sigma$ or other forms of "no effect") stay silent.
- Just like if I focused on one high player in the Chess Olympiad—ignoring that there were almost 1,000 other players.
- When others try to replicate the study, the ground truth proves out.
- Can happen with 50 different big ideas, too (see this).

Study Size Matters

- Bookending this is that human-subject studies tend to be small.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
  - Get higher N from less-targeted situations.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
  - Get higher N from less-targeted situations.
  - Such as chess—in real competitions rather than simulations (such as your "Prisoner's Dilemma" activity).

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
  - Get higher N from less-targeted situations.
  - Such as chess—in real competitions rather than simulations (such as your "Prisoner's Dilemma" activity).
- Mining social media is a major example.

Study Size Matters

- Bookending this is that human-subject studies tend to be small.
- Landmark studies by Kahneman et al. were only $N \sim 100$ people.
- Even some of his famous book Thinking Fast and Slow has come under a cloud.
- Possible to get closed-world 95% or 99% confidence...
- ...but beyond that, the "Murphy's Law" that precision grows only as $\sqrt{N}$ while skew grows as N kicks in.
- Premise of my own Kahneman obit:
  - Get higher N from less-targeted situations.
  - Such as chess—in real competitions rather than simulations (such as your "Prisoner's Dilemma" activity).
- Mining social media is a major example.
- Can we make a tight enough relation between our measurements M and the results Y we are trying to capture?

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.
- Fair Design of Auctions—employed every millisecond on the Web.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.
- Fair Design of Auctions—employed every millisecond on the Web.
- If time allows, show this GLL article about fairness in sampling and ranking.

Algorithmic Fairness

- Is a big area now, drawing many researchers from my own field of Computational Complexity Theory.
- Some of the math grew out of avoiding deadlock among distributed processes.
- Fair Design of Auctions—employed every millisecond on the Web.
- If time allows, show this GLL article about fairness in sampling and ranking.
- Main contentious issues come from AI Fairness.

Bias in Datasets

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.
- Article, "Racial Bias in Computer-Aided Diagnosis" (CalPoly 2023)

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.
- Article, "Racial Bias in Computer-Aided Diagnosis" (CalPoly 2023)
- Another "Murphy's Law" situation: Minorities have **higher variance** in trained diagnoses simply because they comprise a minority in the datasets.

Bias in Datasets

- Article, "How our data encodes systematic racism" in Dec. 2020.
- Example from article: "Google Image search results for 'healthy skin' show only light-skinned women."
- Same search in late 2024 gives results that are still horribly...**sexist!**
- (And maybe under-represents East- and South-Asians.)
- Amazon resumé filter: Trained on mostly male hiring data, perpetuated the same.
- The **reductive** math of training methods picks out **key words**.
- Article, "Racial Bias in Computer-Aided Diagnosis" (CalPoly 2023)
- Another "Murphy's Law" situation: Minorities have **higher variance** in trained diagnoses simply because they comprise a minority in the datasets.
- Fix only by having **ample** data from all groups.

Bias-Removal Dilemma

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Sep. 2024 Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Sep. 2024 Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.
- Shows up for the men too. *Should I remove this bias?*

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Sep. 2024 Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.
- Shows up for the men too. *Should I remove this bias?*
- (Note my model's high accuracy and precision overall.)

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Sep. 2024 Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.
- Shows up for the men too. *Should I remove this bias?*
- (Note my model's high accuracy and precision overall.)
- Doing so would make my model more predictive...

Bias-Removal Dilemma

- Often a tradeoff between **fairness** and **predictive accuracy**.
- Depends on whether quality metrics focus on **output** or **input**.
- Another aspect is **model neutrality** (of input at least).
- **Revisit question of affirmatively de-biasing Google search.**
- Distinguish (a) bias from the model being **skewed** away from what should be true outputs given the inputs, from (b) true outputs that are biased in a **normative** sense by missing a target of fairness.
- Removing skew bias improves accuracy and can preserve neutrality.
- Removing normative bias may reduce accuracy and neutrality.
- Sep. 2024 Women's Olympiad Chess Example: Skewed outcomes on Captures, Advancing Moves, and Knight Moves.
- Shows up for the men too. *Should I remove this bias?*
- (Note my model's high accuracy and precision overall.)
- Doing so would make my model more predictive...but less neutral.

Summarizing Internet and Society and Data

Summarizing Internet and Society and Data

- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.

Summarizing Internet and Society and Data

- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.
- They will hence comprise more of our Distributed Mind and Being.

Summarizing Internet and Society and Data

- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.
- They will hence comprise more of our Distributed Mind and Being.
- This millennium's advances will open exciting possibilities for many of you.

Summarizing Internet and Society and Data

- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.
- They will hence comprise more of our Distributed Mind and Being.
- This millennium's advances will open exciting possibilities for many of you.
- We have, however, tried to promote both *ethical responsibility* and *self-skepticism* in them.

Summarizing Internet and Society and Data

- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.
- They will hence comprise more of our Distributed Mind and Being.
- This millennium's advances will open exciting possibilities for many of you.
- We have, however, tried to promote both *ethical responsibility* and *self-skepticism* in them. Skeptical inquiry is a UB brand.

Summarizing Internet and Society and Data

- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.
- They will hence comprise more of our Distributed Mind and Being.
- This millennium's advances will open exciting possibilities for many of you.
- We have, however, tried to promote both *ethical responsibility* and *self-skepticism* in them. Skeptical inquiry is a UB brand. It does not mean denialism.

Summarizing Internet and Society and Data

- One Upshot: Mathematical patterns and laws will increasingly govern our interactions and understandings on the Internet.
- They will hence comprise more of our Distributed Mind and Being.
- This millennium's advances will open exciting possibilities for many of you.
- We have, however, tried to promote both *ethical responsibility* and *self-skepticism* in them. Skeptical inquiry is a UB brand. It does not mean denialism.
- All best with the rest of term and your time at UB!