Data and Society Parts IV+V Mix

Kenneth W. Regan

(Includes material from Kenny A. Joseph and some other past CSE199 units.)

CSE199, Fall 2024

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

- 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.
- But note this Oct. 6 NYT Upshot article by Nate Cohn on *skew*.
- Silver was non-renewed after ABC bought 538 and has his own 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.

Optimization and Simulation and General Metrics

- Computing ratings and rankings and odds tries to **optimize** for *future accuracy*.
- Often hard to solve analytically, but can be done by **simulating** the model.
- Silver and 538 do this. Simulated. Not simulated.
- My chess metrics derive from a Gaussian model and simulations *conform* to it.
- Simulations are especially useful for systems that **don't** have express models.
- Search Engine Optimization (SEO) is a big area.
- **Can both simulate and solve the random walk on the Net Graph.**
- Last, let's talk briefly about Sentiment Analysis.

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, “Vehicle defect discovery from social media.”
- Often simply tells whether a page exudes happiness/contentment or sadness/anger.
- Can we use it to predict elections? Brexit 2016, Canada 2015, USA 2016 (paper), USA 2016 (BrandsEye).
- (show Python 3 Trinkets web app and activity code.)
- Simplistic idea: if the electorate is (un-)happy that’s (bad) good news for the incumbent.
- “Joy” is an express term of the Harris-Walz campaign. Does it show?

Part VI: AI (part)

Alan Turing: Besides his WWII work on the Enigma machine (featured in the movie *The Imitation Game*) and **Turing Machine** theory of computation in his 1936-38 PhD thesis under Alonzo Church, he is considered the **founder** of Artificial Intelligence.

The **Church-Turing Thesis** is primarily stated in terms of the class of *computable functions*, but here is Turing's angle:

Anything that human beings can consistently deduce or classify can also be achieved by computers acting alone.

The **Turing Test** involves computers trying to be indistinguishable from humans in ordinary life communications and transactions.

Turing All the Possibilities

TP: If it is easy for humans then it will soon be easy for computers.

Defied by a CAPTCHA: "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"

- Vision tasks hard for computers but easy for us. Fortmeyer Tweet
- Too breakable? RECAPTCHA uses a different principle.

Also defies the logical **contrapositive** of Turing's Principle:

If it is really hard for computers then it should be hard for humans.

What we fear when worrying that AI will take away our jobs is:

Stuff that is hard for humans but easy for computers.

The logical **converse** of Turing's Principle acts as a brake, however:

If X is hard for humans—insofar as we can't consistently agree on answers—then X is hard for computers too.

Some Hard Data Challenges (based on the converse principle)

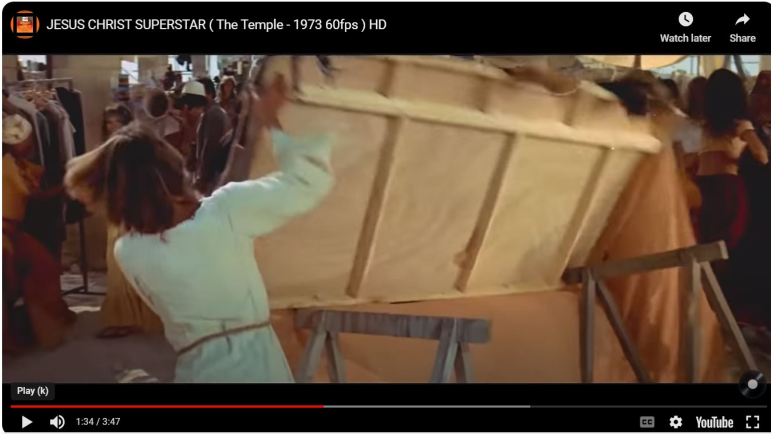
- Inferring people's opinions and beliefs based on text alone. **Stance Classification**
  - How to do it when grammar and intent may differ?
  - Example: "[*that*—] you didn't build that" video. Article by me.
- Reliable automatic translation.
  - Google Translate data-mines known translations for corresponding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

(Chat)GPT, DALL-E, LaMDA, Etc.

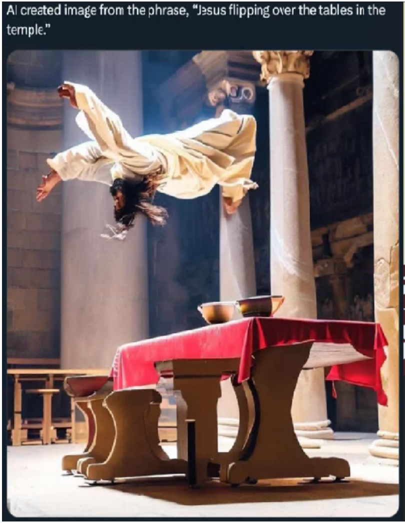
- If you state a topic in brief prose, **GPT-x** composes an essay on it.
- Or even a whole newspaper article.
- **DALL-E** (play on Salvador Dali and the WALL-E movie robot) will create a graphic image in a specified style.
- Examples verging on my professional areas.
- **LaMDA** = Language Model for Dialogue Applications. Claimed by one engineer to evoke human-level *sentience* in conversations.
- A big step up from 1960s "ELIZA." **New (11/28/22)**: ChatGPT.
- Main paradigm of their operation is "find the next word" or "best next visual element."
- But subject to **hallucinations** and other foibles—some shown by me here and here and here.

AI Art Adventure

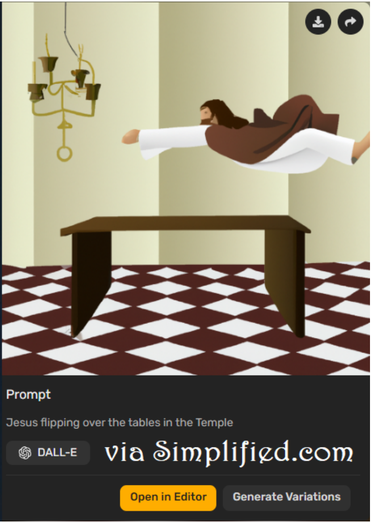
"Jesus flipping over the tables in the Temple." From the movie *Jesus Christ Superstar*—then try it on Cutout or NightCafe or Simplified:



Two Results—one famous, one mine



AI created image from the phrase, "Jesus flipping over the tables in the temple."



Prompt

Jesus flipping over the tables in the Temple

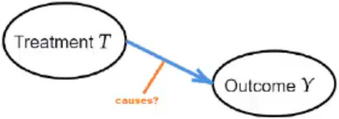
via Simplified.com

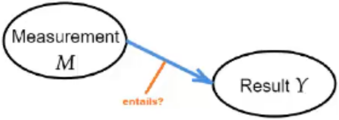
ChatGPT Is Made of Us ("Pogo" Quote)

- *We* are the training data for ChatGPT and other Large Language Models (LLMs).
- (Up to date only thru 2021, however.)
- Example: Writing a Limerick (in Latvian!). [show]
- Does ChatGPT know the inner experience of writing poetry (in Latvian), or is it only shuffling symbols that imitate how poetry (in Latvian) has been written in the past?
- This updates and focuses the "Chinese Room" Argument.
- Given that ChatGPT has already processed the data and rules to write grammatical and cogent Latvian, a minimal threshold on the way to *sentience*, IMHO, is that a non-Latvian speaker like myself, giving examples of high-quality limericks in English and with no further Latvian data of any kind, should be able to get it to write superb limericks in Latvian.
- (But possibly I already pushed it to the limits of its current data.)

Part V: Societal Computing and Fairness

Studies in medicine, psychology, and other sciences have enabled us to gauge significant causes and effects. Two typical notations for the objects of these studies:





- Often Y is a binary choice: does a desired outcome happen? does the result go one way or the other way?
- The math for determining whether there is a significant causal relationship then resembles a simple poll.
- For a targeted value Y, the study's findings can be phrased as whether Y is significantly ahead.
- I.e., is Y beyond the *margin of error* for the **null hypothesis** of no causation?

The Replication Crisis

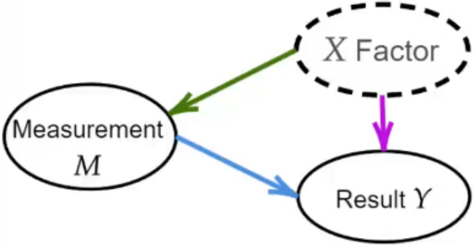
- Means that findings of significant causality in one published paper are not found when another team re-creates the study.
- A Simple Mechanism:
  - Something like Covid brings fresh Big Ideas in medicine and psychology (etc.).
  - More than 50 of the world's major institutions launch a study... privately.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

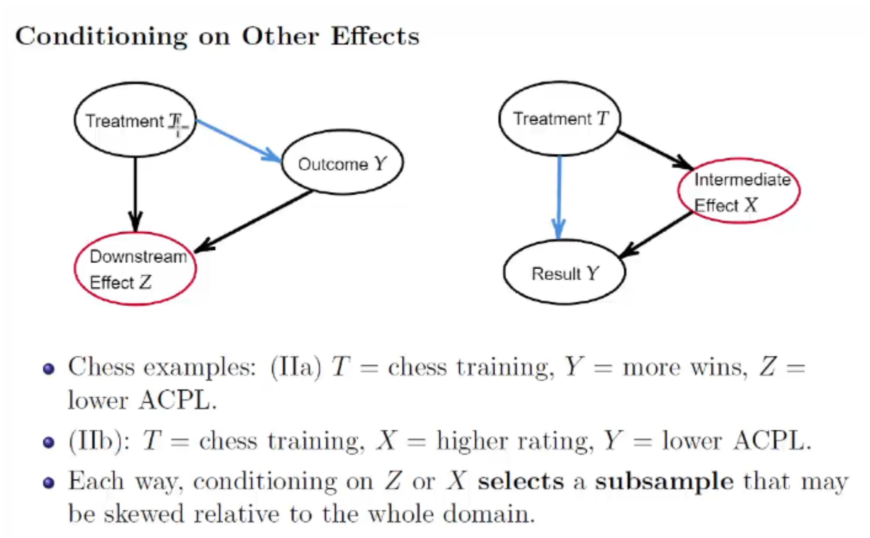
Other Study Design Flaws To Beware

Confounding Factor(s):



- Possible example: X = a scandal, such as in North Carolina.
- Can both *stimulate* M (such as "heat") while *inhibiting* Y (such as "Challenger Wins").
- Even if impact is positive on both M and Y, X can dominate, drown out, or otherwise skew the effect we are trying to analyze.

Selection Bias From Conditioning

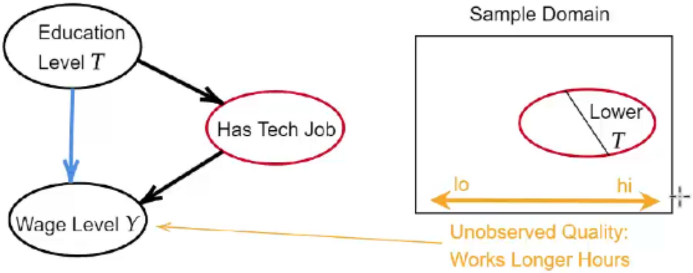


Conditioning on Other Effects

- Chess examples: (IIa) T = chess training, Y = more wins, Z = lower ACPL.
- (IIb): T = chess training, X = higher rating, Y = lower ACPL.
- Each way, conditioning on Z or X **selects** a **subsample** that may be skewed relative to the whole domain.

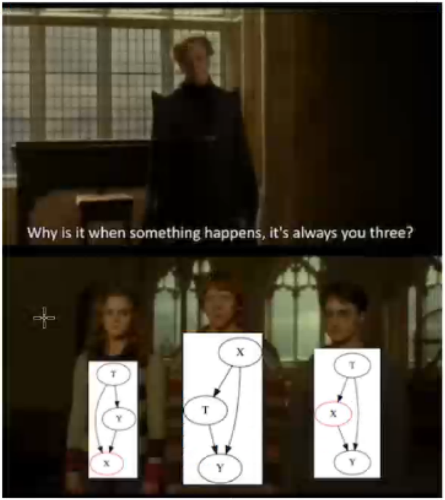
Non-Chess Example (adapted from here

Suppose we are doing a large-scale study of the effect of education on wages, but decide to condition on people having tech jobs:



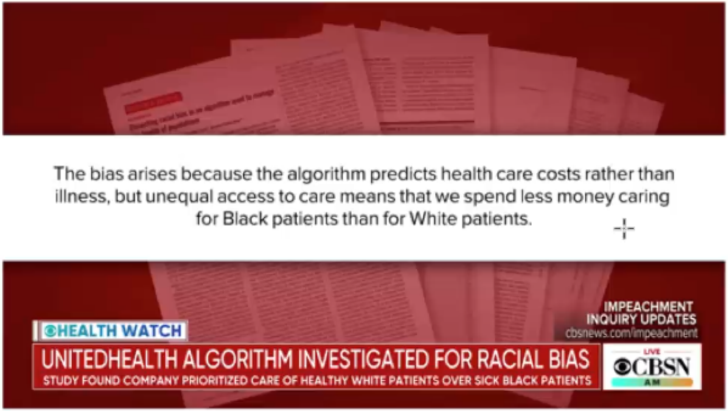
- Subsample from conditioned variable is skewed.
- (Source says "white-collar jobs" rather than "tech jobs.")
- Can also happen from choices of unrepresentative proxy variables.

Harry Potter Meme (also from here



Example of Bias From Proxy Variable (K. Joseph)

Here the variable Y' = health care costs used for Y = level of illness did implicit conditioning. Video.



The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients.

HEALTH WATCH
UNITEDHEALTH ALGORITHM INVESTIGATED FOR RACIAL BIAS
STUDY FOUND COMPANY PRIORITIZED CARE OF HEALTHY WHITE PATIENTS OVER SICK BLACK PATIENTS
IMPEACHMENT INQUIRY UPDATES
cbsnews.com/impeachment
LIVE
CBSN
AM

Some Notorious Inferences and Model Decisions

- **Target**ing ads at a pregnant teen: article.
- Amazon often recommends to me the book *Quantum Algorithms Via Linear Algebra*. Problem is—I co-wrote it. Nice to hear…
- Bond and CDO (Collateralized Debt Obligation) ratings before the 2008 crash.
- Book *Weapons of Math Destruction*, by Cathy O'Neill. Thesis: Mathematical models fossilize biases in data from remote history and skewed prior sources.
- Book Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, by Seth Stephens-Davidowitz. Thesis: Formal survey responses are inconsistent with opinions from the same populations mined on social media.
- Insofar as we are the training data for the Internet, the latter has baked in tangible amounts of racism and sexism.