CS37300:
Data Mining and Machine Learning

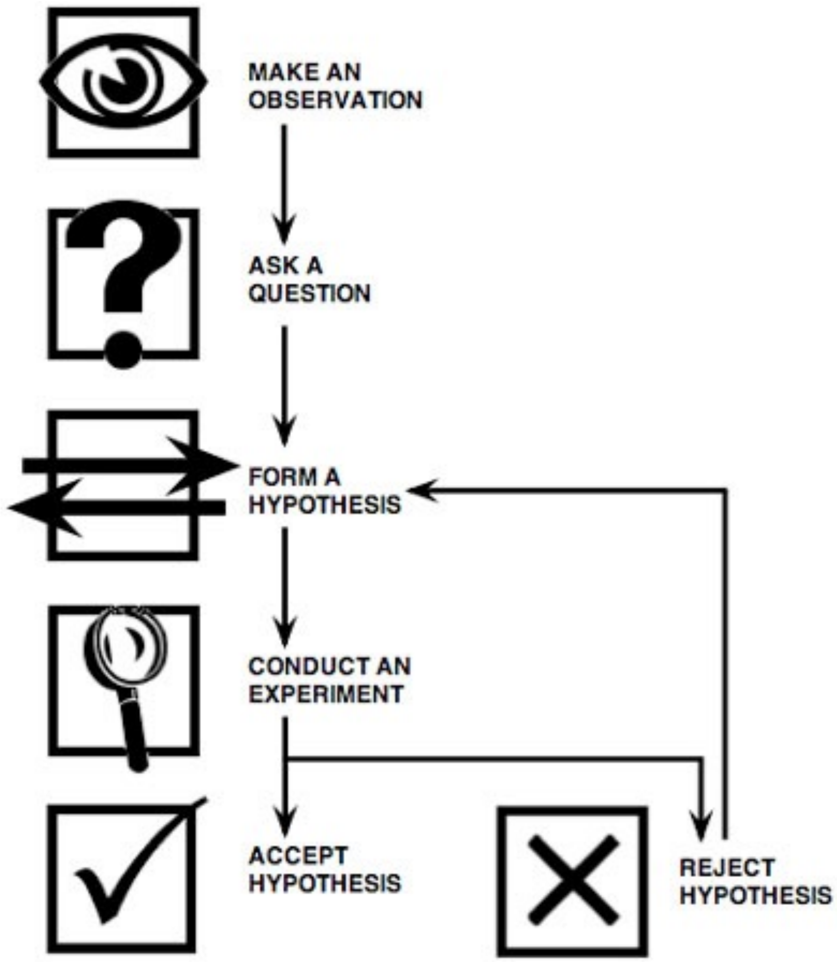
Hypothesis Testing

Profs. Tianyi Zhang and Rajiv Khanna

6 Sep 2023

Hypothesis testing

Hypothesis Testing



What is a hypothesis?

- **Hypotheses** are tentative statements of the expected relationships between two or more variables
  - **Inductive** hypotheses are formed through inductively reasoning from many specific observations to tentative explanations (*bottom-up*)
  - **Deductive** hypotheses are formed through deductively reasoning implications of theory (*top-down*)
- Reasons for using hypotheses
  - Provides a useful framework for organizing and summarizing results and conclusions
  - Provides focus and directs research investigation

PURDUE UNIVERSITY

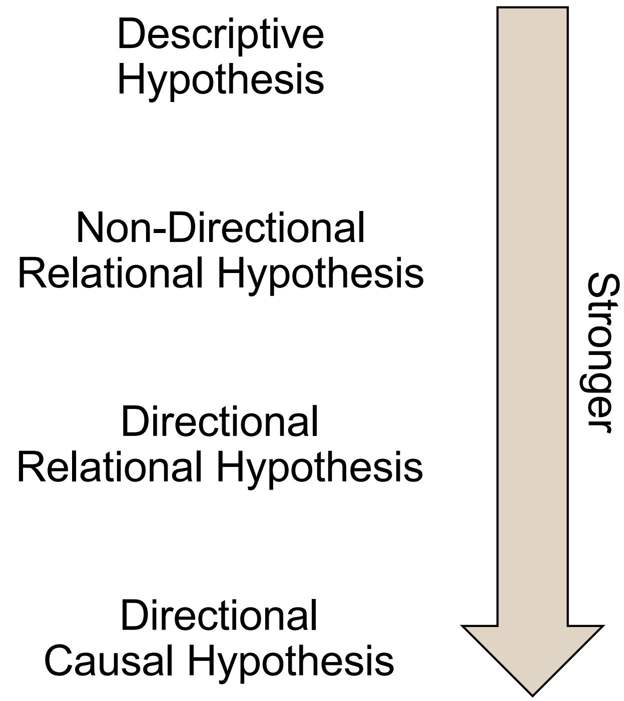
Types of hypotheses

Broad categories

- **Descriptive**: propositions that describe a characteristic of an object
- **Relational**: propositions that describe the relationship between 2+ variables
- **Causal**: propositions that describe the effect of one variable on another

Specific characteristics

- **Non-directional**: an differential outcome is anticipated but the specific nature of it is not known (e.g., the tuning parameter will affect algorithm performance)
- **Directional**: a specific outcome is anticipated (e.g., the use of pruning will increase accuracy of models compared to no pruning)



From claims to testable hypotheses

- Israel COVID cases breakdown Aug 17, 2021

Age	Total	Vax %		Severe Cases		Score function
Conditional	Individuals (approx.)	Not Vax	Fully Vax	Not Vax per 100k	Fully Vax per 100k	Conditional Efficacy
[12,15]	650,000	62.1%	29.9%	0.3	0.0	100.0%
[16,19]	600,000	21.9%	73.5%	1.6	0.0	100.0%
[20,29]	1,200,000	20.5%	76.2%	1.5	0.0	100.0%
[30,39]	1,050,000	16.2%	80.9%	6.2	0.2	96.8%
[40,49]	900,000	13.2%	84.4%	16.5	1.0	94.2%
[50,59]	750,000	10.0%	88.0%	40.2	2.9	93.2%
[60,69]	550,000	8.8%	89.8%	76.6	8.7	89.8%
[70,79]	350,000	4.2%	94.6%	190.1	19.8	90.6%
[80,89]	120,000	5.6%	92.6%	252.3	47.9	84.0%
90+	50,000	6.1%	90.5%	510.9	38.6	93.0%

PURDUE UNIVERSITY From claims to testable hypotheses

Claim: Vaccine efficacy is higher for 90+ y/o individuals than for [80,89] y/o individuals

Building the hypothesis:

- **Step 1:** Express data as random variables (joinly). E.g.:
  - A age
  - SV severe vax per 100k
  - SU severe unvax per 100k
  - $\mathrm{Y} = \mathrm{SU}/(\mathrm{SU} + SV)$ observed vaccine efficacy

From claims to testable hypotheses

Claim: Vaccine efficacy is higher for 90+ y/o individuals than for [80,89] y/o individuals

Building the hypothesis:

- **Step 2:** Restate claim as a hypothesis about the relationship between the random variables, e.g.,
  - Hypothesis: $E[Y|A > 90] > E[Y|80 \leq A \leq 89]$
- **Step 3:** Determine type of hypothesis (and consider whether you can make it stronger)

From claims to testable hypotheses

- **Claim**: Vaccine efficacy is higher for 90+ y/o individuals than for [80,89] y/o individuals
- **Types of hypotheses**:
  - *Descriptive*: Efficacy values vary (i.e., Y varies).
  - *Non-directional relational*: Y varies based on age (i.e., A and Y are associated)
  - *Directional-relational*: A > 90 folks have higher efficacy (i.e., A > 90 is associated with smaller Y)
  - *Causal-relational*: A > 90 folks have higher vaccine efficacy because these are monitored more closely in nursing homes and get medical interventions earlier before they get too sick

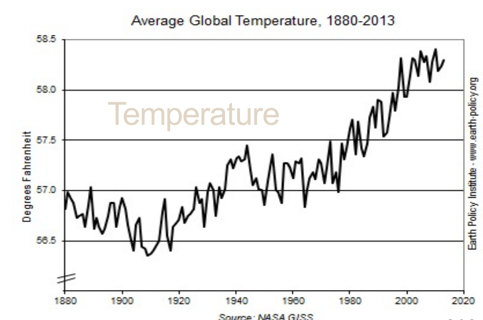
Is Aspirin effective in reducing cancer risk?

- Here, we are looking at causal effects…
- Data
  - A person represented by random variable $X \in$ {{Age}, {Sick, Not Sick}, …}
    - Recruit people: x_{john} , x_{mary} , x_{eve} , x_{adam} , …
    - Medicine to take: $T \in$ {aspirin, placebo}
- Hypothesis testing by experimentation
  - Force ½ (randomly chosen) of the people to take aspirin : $Y|X, \textbf{do}(T = \text{aspirin}) \in \{1 \text{ - Cancer in 1yr}, 0 \text{ - No Cancer in 1yr}\}$
    - Here, the **do**() notation means forcing T to be something (intervention)
  - Force remaining ½ to NOT take aspirin: $Y|X, do(T=placebo)$
    - Hypothesis: $E[Y|X, do(T=aspirin)] < E[Y|X, do(T=placebo)]$

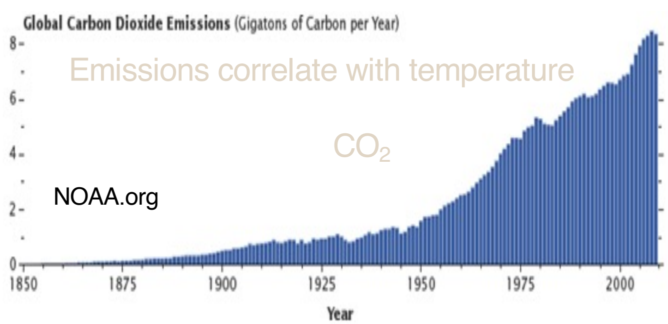
Directional
Causal Hypothesis

Example:

- CLAIM 1: The temperature of the planet is rising and the increase is **due** to human activities such as fossil fuel use and deforestation.
- Which kind of data could support such claim?

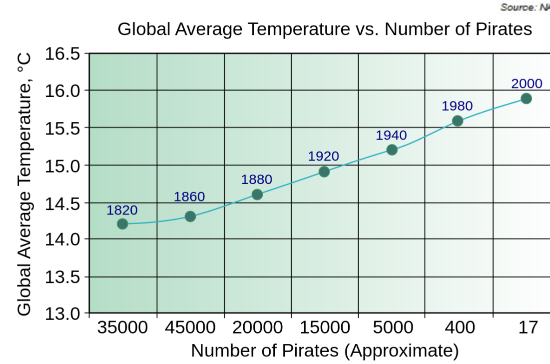


NAS



Not enough

Why?



Directional Relational Hypothesis

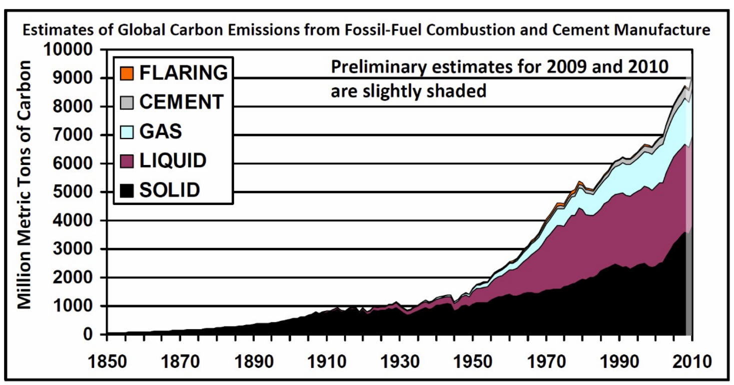
- CLAIM 2: The temperature of the planet is rising with increased human activity

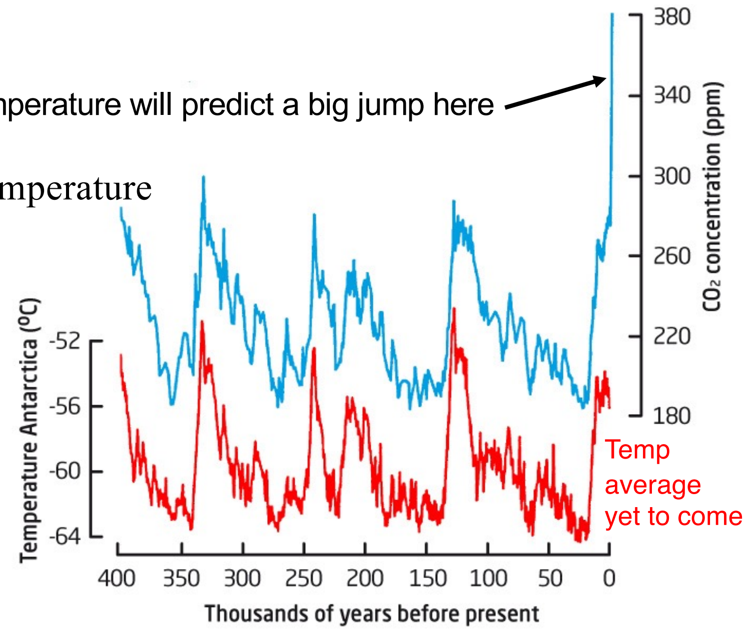
Model trained on CO2 data to predict Antartica's temperature will predict a big jump here

$f(x_{CO2})$ = Temperature

CO$_2$ jump related to human activity

$g(x_{fossil\text{-}fuel})$ = Atmospheric CO$_2$ – Natural CO$_2$





Directional Relational Hypothesis

Causal Claims without Experiments are Difficult

- CLAIM 1: The temperature of the planet is rising and the increase is **DUE** to human activities such as fossil fuel use.
  - How would you test it?
  - How it is tested:
    - Climate models (we know how climate works)
      - We know how much energy the sun outputs
      - We know how much energy the planet radiates back into space
      - We know where the energy goes inside the planet
      - https://earthobservatory.nasa.gov/features/EnergyBalance
    - Historic *natural experiment* events:
      "Coal-burning in Siberia after volcanic eruption led to climate change 250 million years ago"

Directional
Causal Hypothesis

PURDUE
UNIVERSITY®

- Biases and need for sound experimental design

Neglecting base rates

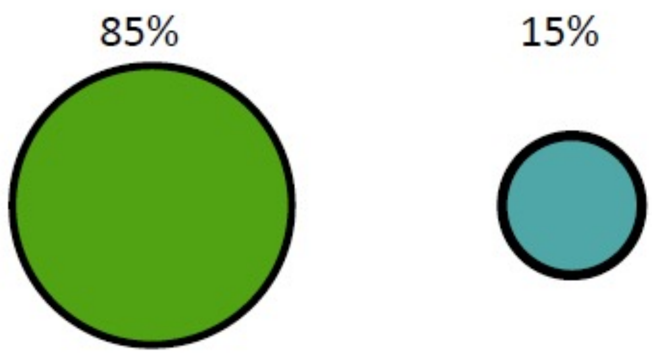
- Taxi-cab problem *(Tversky & Kahneman '72)*
  - 85% of the cabs are Green
  - 15% of the cabs are Blue
  - An accident eyewitness reports a Blue cab
  - But she is wrong 20% of the time.
- What is the probability that the cab is Blue?
  - Participants tend to overestimate probability, most answer 80%
  - They ignore baseline prior probability of blue cabs.

PURDUE UNIVERSITY

More on neglecting base rates

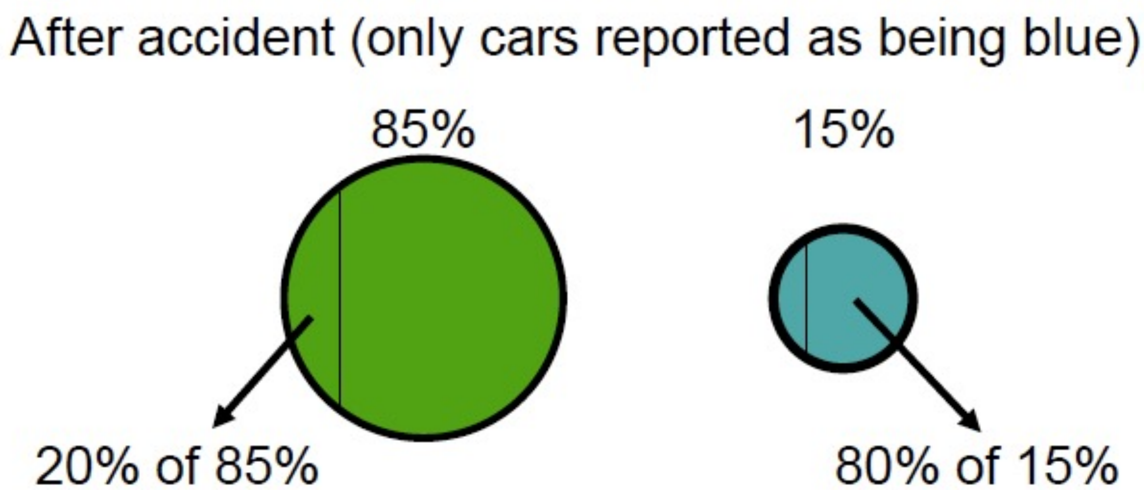
A priori (beforehand)

$$P(green) = 0.85$$
$$P(blue) = 0.15$$



$$P(seeBlue|blue) = 0.80$$
$$P(seeBlue|green) = 0.20$$

More on neglecting base rates



More on neglecting base rates

- How to compute probability

$$
\begin{aligned}
P(blue|seeBlue) &= \frac{P(blue \wedge seeBlue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue|blue)P(blue) + P(seeBlue|green)P(green)} \\
&= \frac{0.80 \cdot 0.15}{(0.80 \cdot 0.15) + (0.20 \cdot 0.85)} \\
&= 0.41
\end{aligned}
$$

Most people answered 80%

Heuristics and biases

- Tversky & Kahneman, psychologists, propose that people often do not follow rules of probability when making decisions
- Instead, decision making may be based on heuristics
  - Lowers cognitive load but may lead to systematic errors and biases
- Examples:
  - Observation bias
    - Confirmation bias
  - Availability heuristic
  - Representativeness heuristic
  - Conjunction fallacy (we will not cover this)
  - Numerosity heuristic (we will not cover this)

Arthritis study *(Redelmeier & Tversky '96)*

- Common belief:
  - **Arthritis pain is associated with changes in weather**
- Experiment:
  - Followed 18 arthritis patients for 15 months
  - 2 x per month assessed: (1) pain and joint tenderness, and (2) weather
- Results:
  - No correlation between pain/tenderness and weather
  - Patients saw correlation that did not exist... why?

Arthritis study (cont)

- Patients noticed when bad weather and pain co-occurred, but failed to notice when they didn't.
  - Better memory for times that bad weather and pain co-occurred.
  - Worse memory for times when bad weather and pain did not co-occur
- **Confirmation bias**: People often seek information that **confirms** rather than disconfirms their original hypothesis

PURDUE UNIVERSITY

Estimating probabilities *(Tversky & Kahneman '73/'74)*

- *Question*: Is the letter **R** more likely to be the 1st or 3rd letter in English words?
- *Results*: Most said **R** more probable as 1st letter
- *Reality*: **R** appears much more often as the 3rd letter, but it's easier to think of words where **R** is the 1st letter

Estimating probabilities (cont)

- *Question*: Which causes more deaths in developed countries? (a) traffic accidents or (b) stomach cancer
- *Typical guess*: traffic accident = 4X stomach cancer
- *Actual*: 45,000 traffic, 95,000 stomach cancer deaths in US
- Ratio of newspaper reports on each subject: 137 (traffic fatality) to 1 (stomach cancer death)

- **Availability heuristic**: Tendency for people to make judgments of frequency on basis of how easily examples come to mind

Base Rate Study *(Kahneman & Tversky '73)*

- Participants told that for a set of 100 people are either:
  - 30% engineers/70% lawyers, or
  - 70% engineers/30% lawyers
- *Given*: A description of a person Jack, which is representative of a prototypical engineer (e.g., likes carpentry and mathematical puzzles, careful, conservative)
- *Question*: Is Jack more likely to be a lawyer or engineer?
- *Results*: Participants in the 30% condition judged Jack just as likely to be an engineer as participants in the 70% condition.

Base rate study (cont)

- People use the representative heuristic to make inferences...
  - Inferences is based solely on similarity of target to category members
  - Base rates (70%-30%) are ignored
- ...rather than using formal statistical rules to make inferences
  - Inferences should be based on similarity of target to category members AND base rates (70%-30%)

- **Representative heuristic**: categorizations made on the basis of similarity between instance and category members

Gambler's fallacy

- *Gambler's fallacy*: belief that if deviations from expected behavior are observed in repeated independent trials, then future deviations in the opposite direction are then more likely
- This is an example of the **representativeness heuristic**—where the probability of an event is judged by its similarity to the population from which sample is drawn
- The sequence "H T H T T H" is seen as more representative of a prototypical coin sequence. Why?
  - When people are asked to make up random sequences, they tend to make the proportion of H and T closer to 50% than would be expected by random chance
  - People believe that short sequences should be representative of longer ones

Base Rate and Predictive Modeling

- Suppose we had developed a classifier to predict if a student would fail CS37300
  - The classifier is correct 90% of the time
- Would you have used this to decide if you are ready to take CS37300?

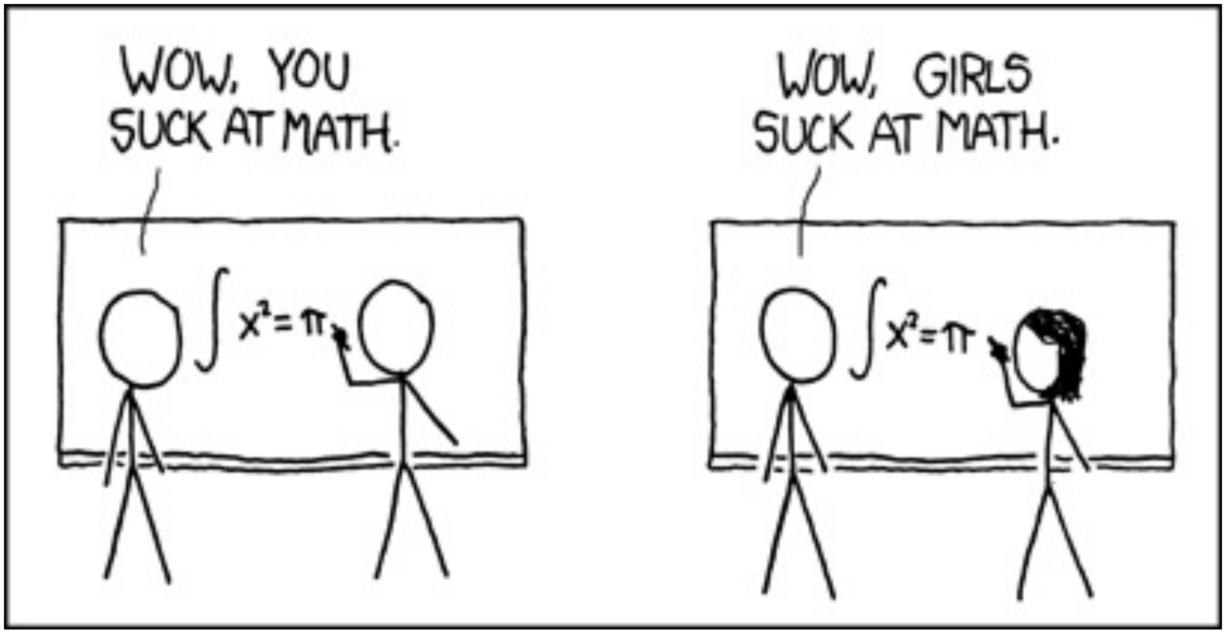
PURDUE UNIVERSITY

Interpretation of these findings

- People do not use proper statistical/probabilistic reasoning... instead people use heuristics which can **bias** decisions
- Heuristics can often be very effective (and efficient) for social inferences and decision-making
- ... but be aware that **heuristics can bias** results from **exploratory data analysis and other modeling efforts**

Heuristics and bad generalizations

- Source: https://xkcd.com/385/



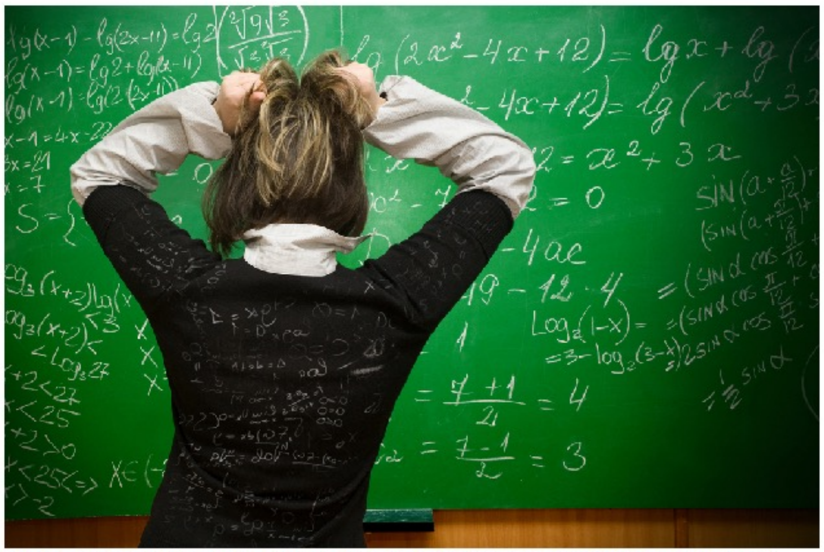
- Revisit again when discussing interpretability

POLITICS
→ *Climate Change, Climate Desk, Guns, Science*

Science Confirms: Politics Wrecks Your Ability to Do Math

Farewell, Enlightenment: New research suggests that people even solve math problems differently if their political ideology is at stake.

—*By* ***Chris Mooney*** | Wed Sep. 4, 2013 12:59 PM EDT



A new study finds that even how you solve a difficult math problem can depend on your politics. *AlenKadr/Shutterstock*

Everybody knows that our political views can sometimes get in the way of thinking clearly. But perhaps we don't realize how bad the problem actually is. According to a new psychology paper, our political passions can even undermine our very basic reasoning skills. More specifically, the study finds that people who are otherwise very good at math may totally flunk a problem that they would otherwise probably be able to solve, simply because giving the right answer goes against their political beliefs.

Kahan et al. (2013) "Motivated numeracy and enlightened self-government." *Social Science Research Network*.

1000+ participants were asked about their political views and also asked a series of questions to gauge their mathematical reasoning ability.

Participants were then asked to solve a fairly difficult problem that involved interpreting the results of a (fake) scientific study.

One group was given a problem involving the effectiveness of a new **skin cream**. The other group was given a mathematically similar problem, but the data involved the effectiveness of a **gun control** measure.

What was the result?
Highly numerate people were <u>more susceptible</u> to letting politics skew their reasoning than were those with less mathematical ability.