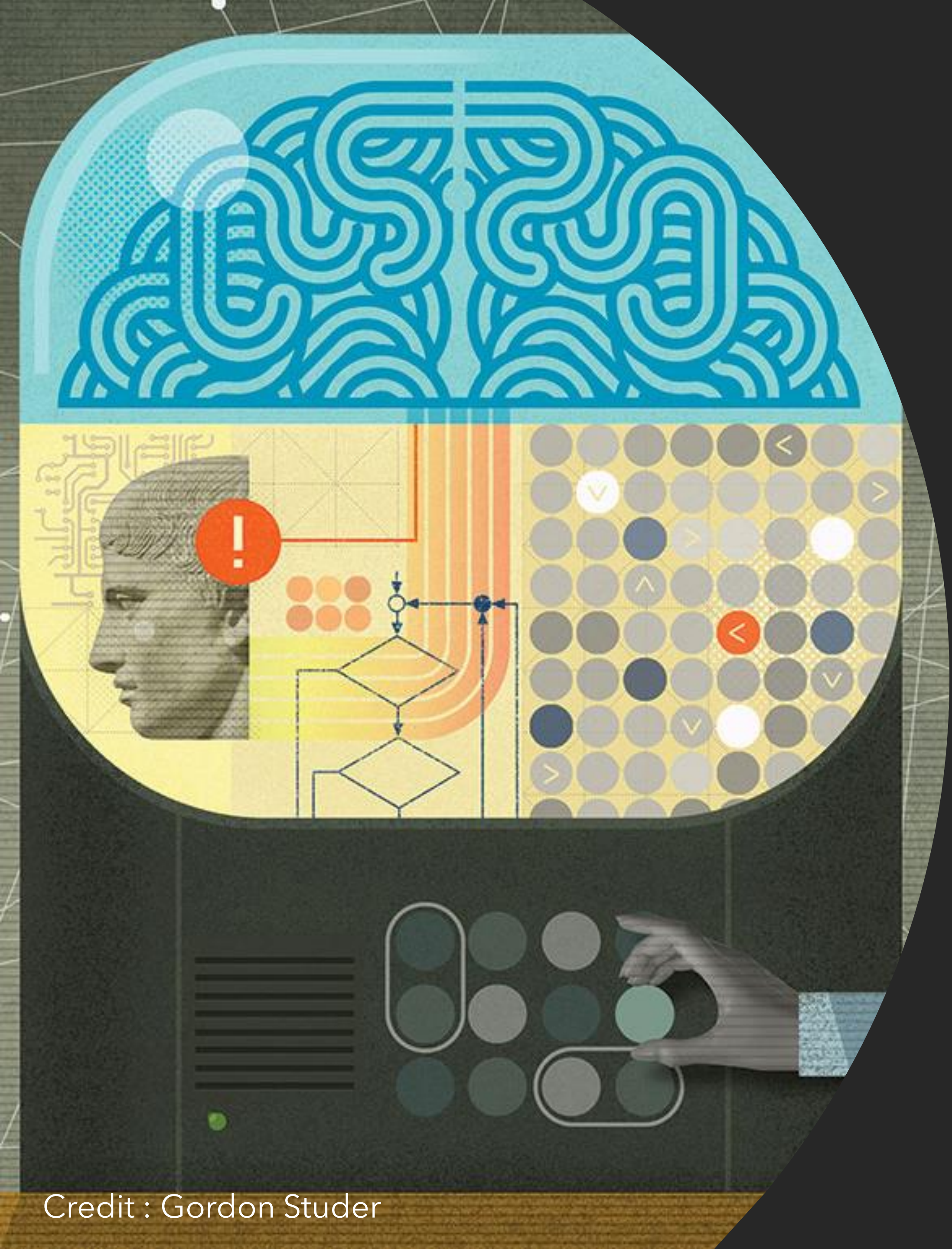
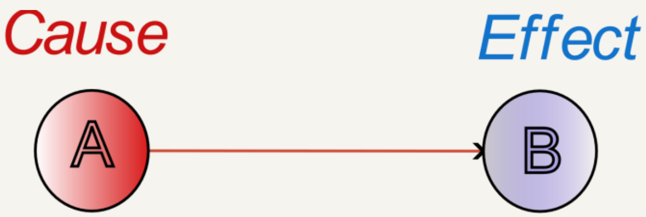


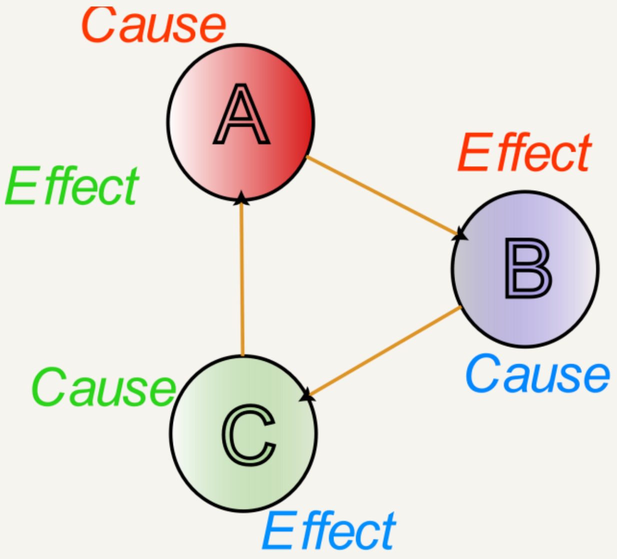
Credit : Gordon Studer

Causal AI

Simarjit Dhillon

Causal Reasoning

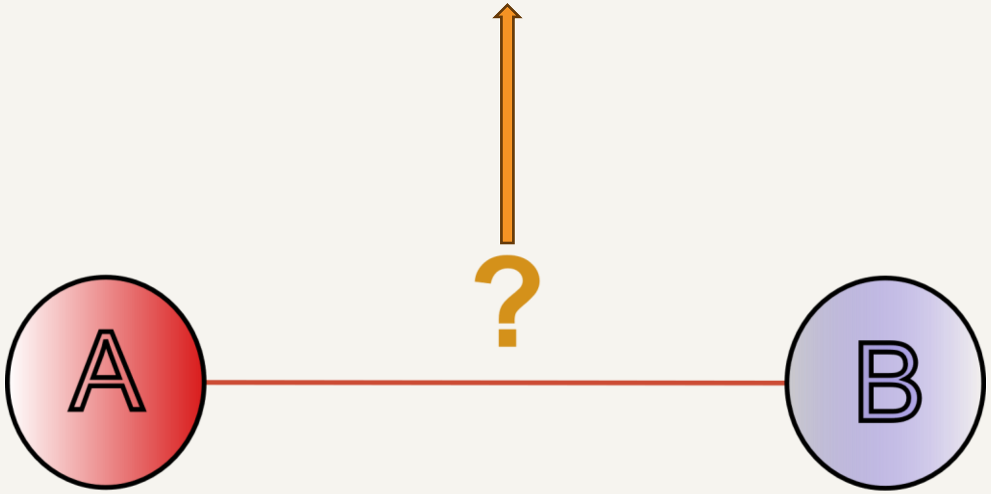




Causal Inference

Null Hypothesis studies :

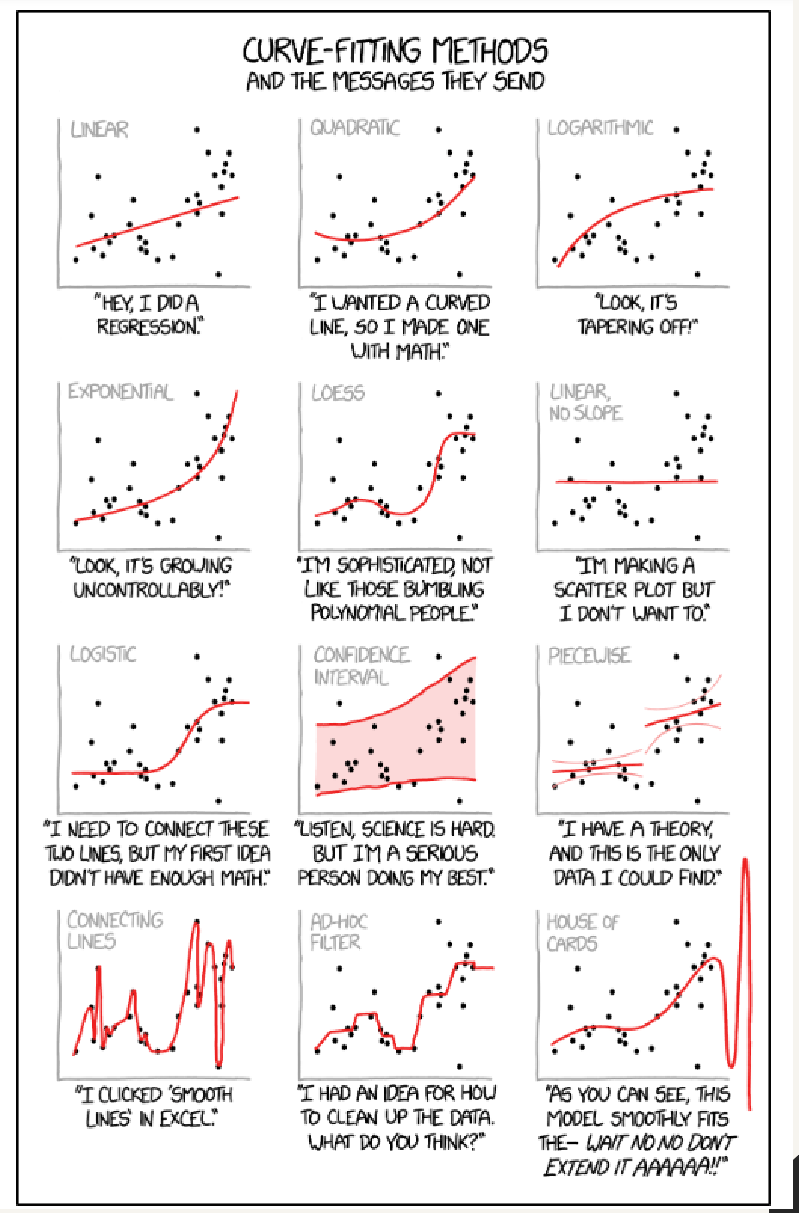
- Statistical inferencing
- Bayesian Analysis



Correlation ≠Causation

And (other) problems with traditional models

- Training issues
  - Over/Underfitting
  - Not enough data
  - Inaccuracy -> Usually needs more data
  - BIG models/networks
- **Lack of reasoning**
  - By extension lack of transparency
  - Leads to difficulty making General AI.
- ***Bias***


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Credit : https://cims.nyu.edu/~brenden/courses/labincp/labs/LabReg-IntroToRegression.html

Hypothetical Scenario : College Admission [2]

Ideal

Marker Reference :

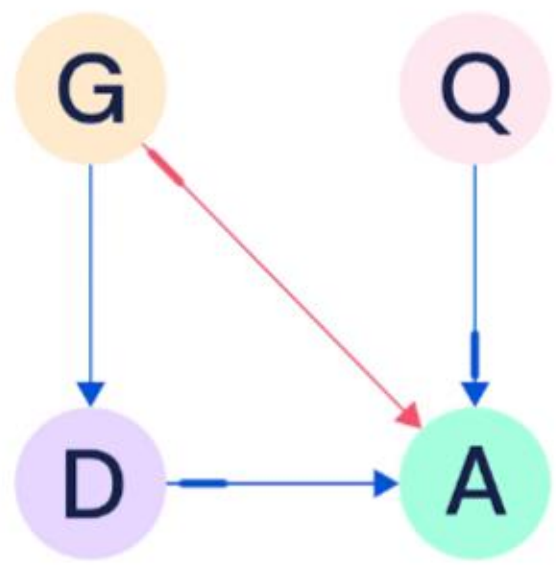
Q : Qualifications

G : Gender

D : Departments

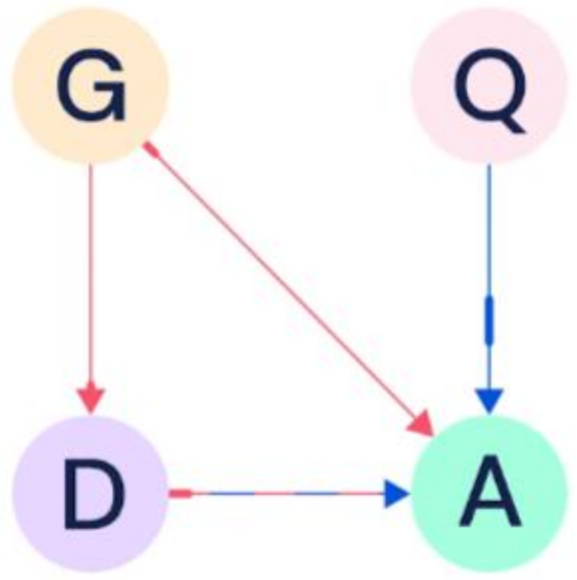
A : Admission

Indirect Bias



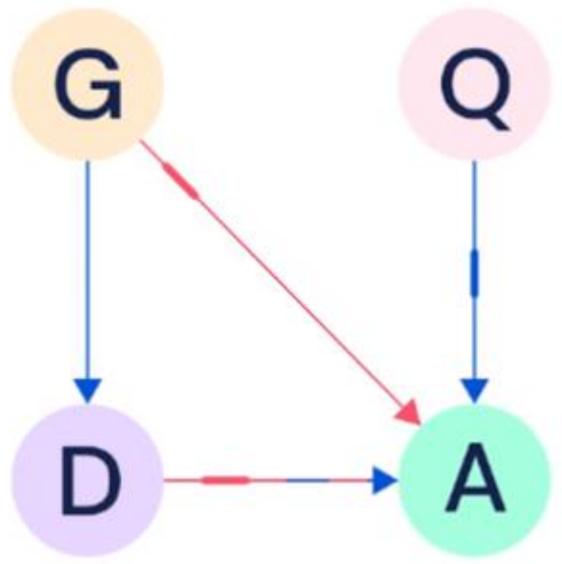
Voluntarily Choose Departments

Direct Bias



Due to systemic biases

Direct Bias



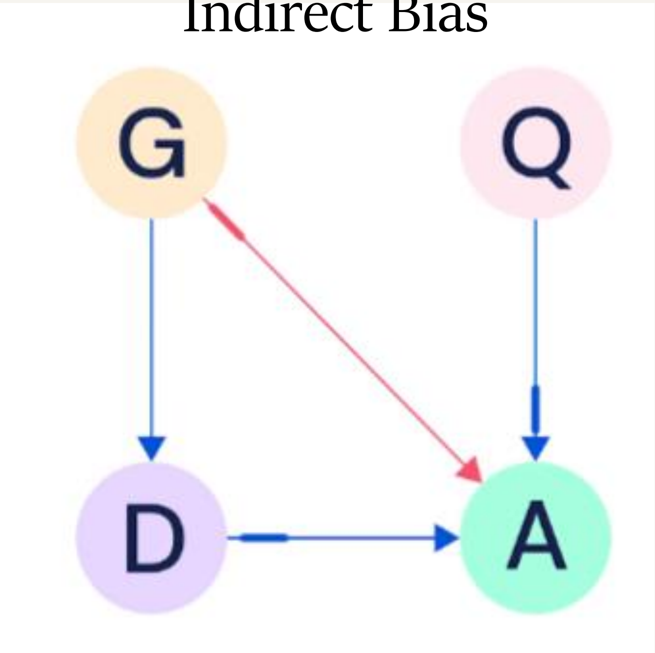
School lowers admission rate of the departments that women choose

Scenario Contd [2].

- School trains an AI model based on historical admission data.
- You want to judge if the model is fair or unfair.

- *Statistical Parity (50-50 M/F ratio)*
- *Equal False positives(FP) and False negatives (FN) for both*

Indirect Bias



Voluntarily Choose Departments

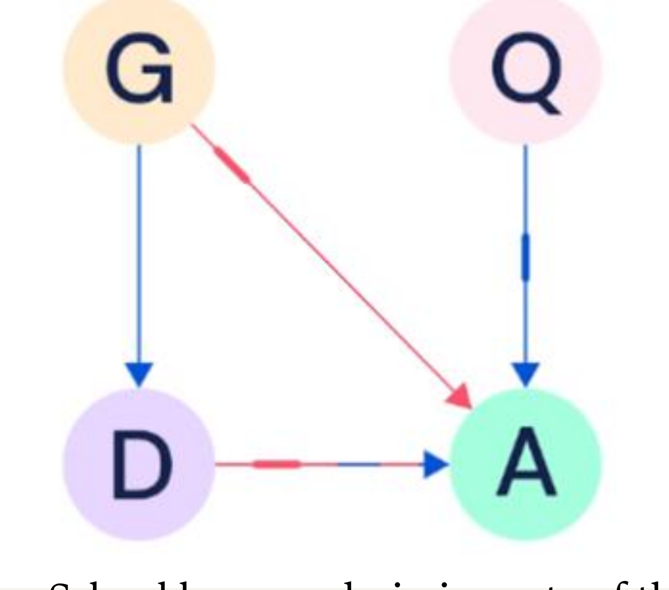
Uneven Statistical Parity (50-50 M/F ratio)

But Even FP , EP for both

- Might show uneven parity
- But got there through FAIR means!

If you use both methods, how do you know which scenario are you in ?

Direct Bias



School lowers admission rate of the departments that women choose

Equal False positives and False negatives for both

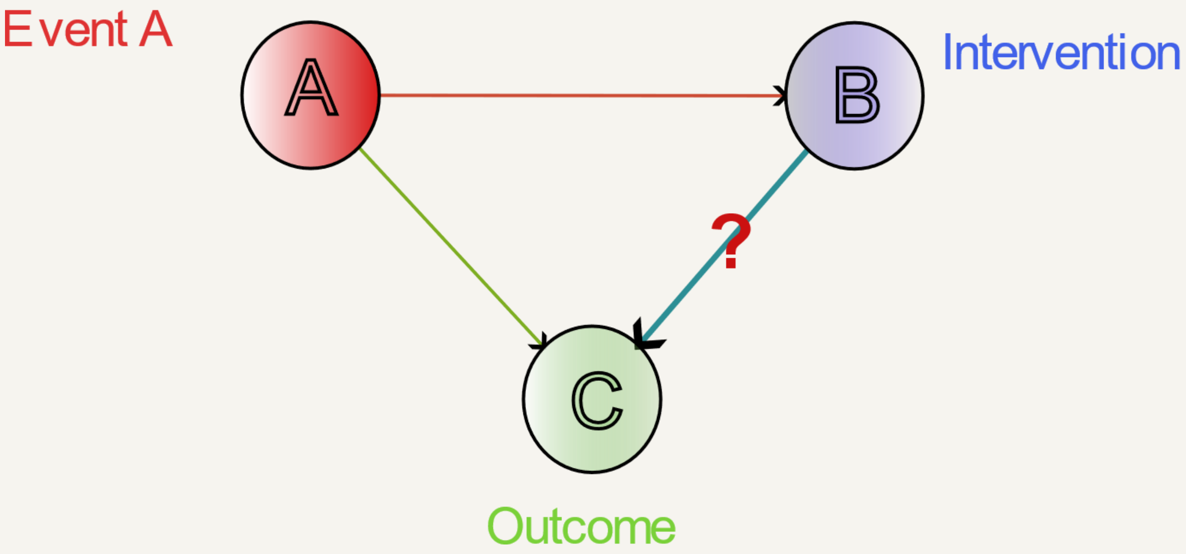
But Uneven Statistical parity

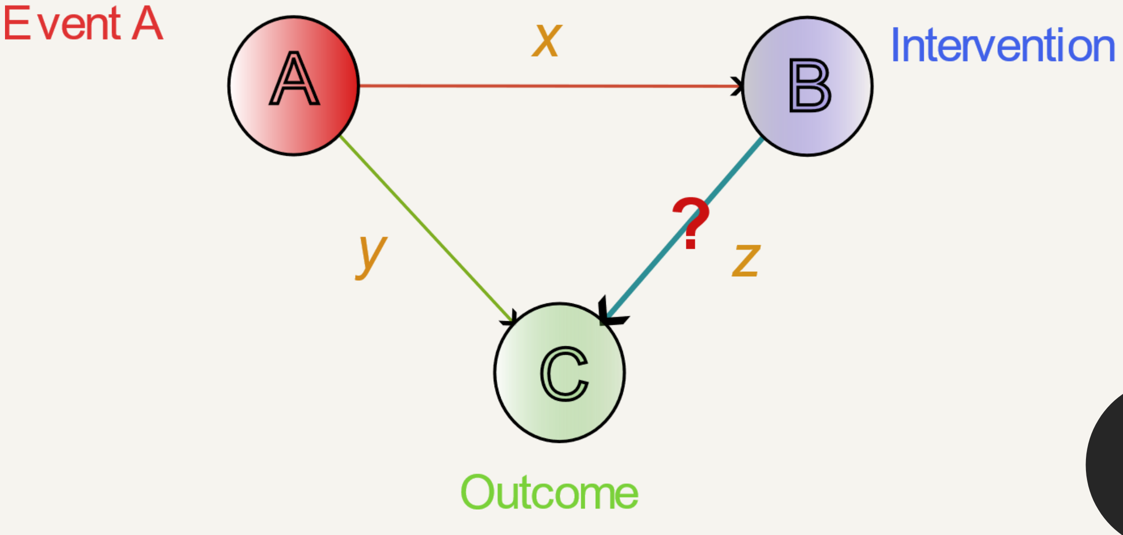
- Might show even FP and EP for both
- But got there through UNFAIR means!

How does Causal Inference work ?

Showing causality A causes B (A -> B):
1. A comes temporally before B
2. P(B | A) > P(B | not A)
3. Nothing else causes A -> B

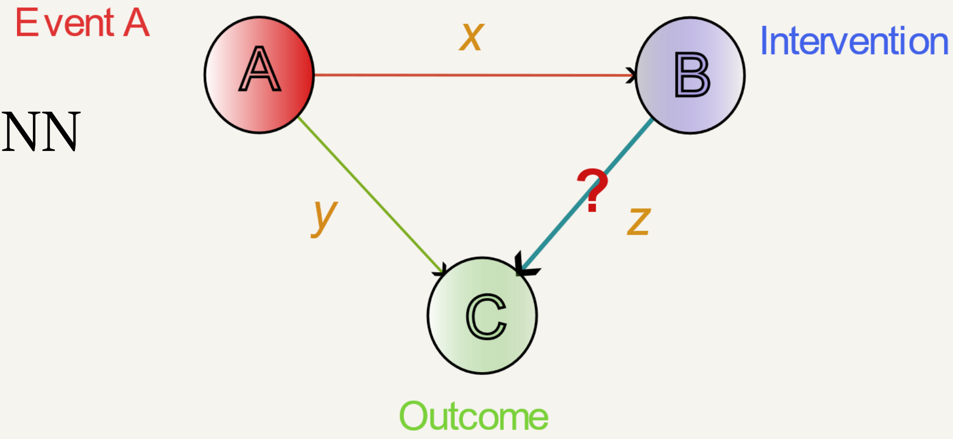
- Create Causal Influence Diagrams (CIDs) / Causal Network Graphs
  - Done through identifying Intervention variables
  - Tweak these intervention variables to see effect on outcome variables/nodes
  - Get weight data

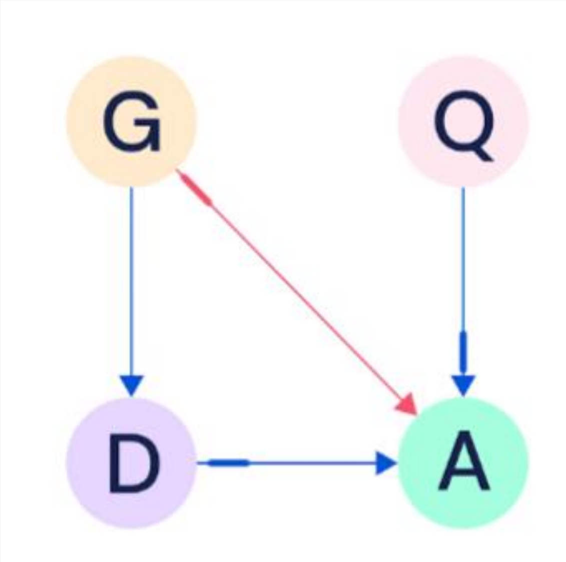




Advantages of Causal Inference in AI

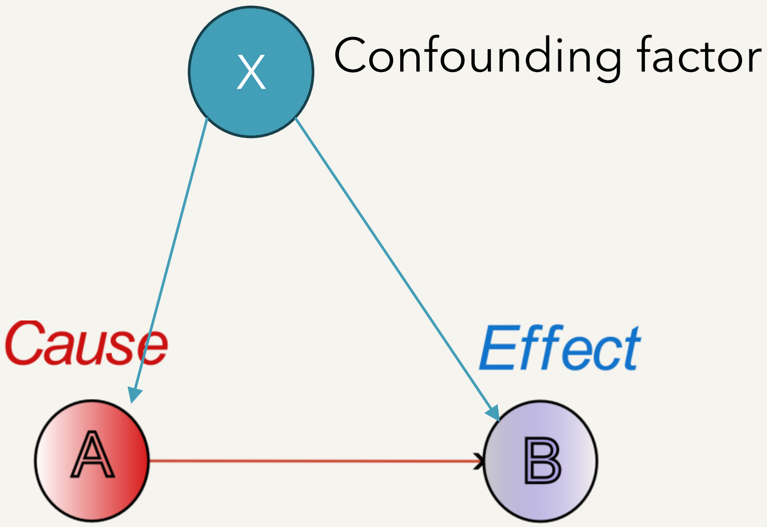
- Models are more efficient to train
  - Causal networks are much smaller than typical NN, CNN models
- Capable of reasoning
  - By extension transparency
  - Can congregate multiple causal models together for more General AI
- Easy to identify biases
- Can target individuals and comment on fairness of treatment of individuals
  - Using counterfactuals





Disadvantages of Causal Inference

- Very high complexity !
  - Functions for finding causal relations can be very complex – (showing causality is difficult)
- Ignorability
- Issues with training data!
  - What if we don’t have proper data points for counterfactuals?



Work Done

This is still a very theoretical / research heavy field so few publications have been made and even little actual application:

1. Causal AI is being researched for use in self sustaining AI-native wireless networks and digital twins [1].
2. GPT model trained with causal inference in Othello (board game) was able to use an internal representation of a board state, when interventions were made (even outside of trained board states) it was able to update its predictions [4].
3. British scientists used a causal AI technique called counterfactual event attribution in the potential outcomes framework to determine whether human-produced greenhouse gas emissions were an underlying cause of the deadly European heatwave of 2003 [5].

Finally : The Future of Causal AI

- Google DeepMind Conference Paper in 2024 states “By establishing a formal connection between causality and generalization, our results show that causal world models are a necessary ingredient for robust and general AI” [4]
- Causal models are likely going to have to be integrated with LLMs for general viability and trust in professional field.
- Without the ability for reasoning and response to decisions made by the model (theoretically only possible through Causal models for now) , AI regulation and widespread adoption will be a huge challenge.

Questions?

- References :

1. https://arxiv.org/pdf/2309.13223
2. https://deepmind.google/discover/blog/causal-bayesian-networks-a-flexible-tool-to-enable-fairer-machine-learning/
3. https://deepmind.google/discover/blog/discovering-when-an-agent-is-present-in-a-system/
4. https://openreview.net/pdf?id=pOoKI3ouv1
5. https://ssir.org/articles/entry/the_case_for_causal_ai