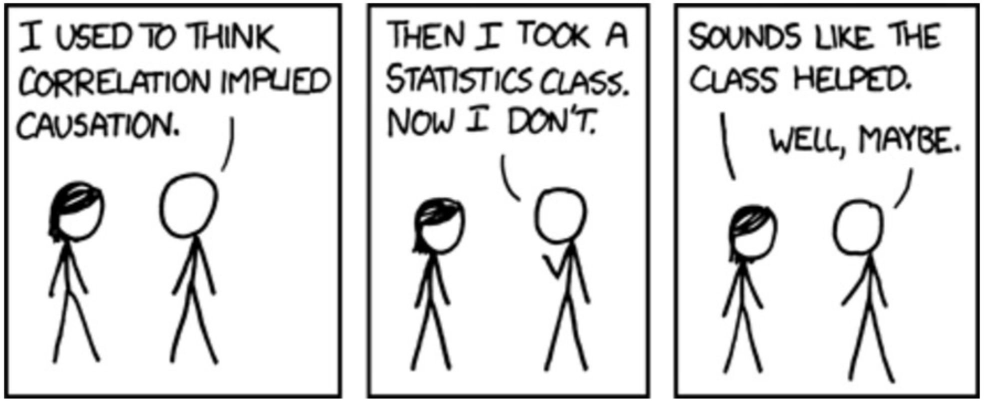
Causal Inference in Data Analysis with Applications to Fairness and Explanations

Sudeepa Roy

DUKE
COMPUTER SCIENCE

Babak Salimi

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE



This Tutorial…

An overview of "Causal Inference" concepts and methods

Applications of causal inference on "Responsible Data Science" (fairness & explainability)

Some recent research on causal inference (biased toward our work)

Outline

Introduction

Pearl's Graphical Causal Model

Rubin's Potential Outcome Framework

Briefly: some recent research on causal inference techniques (scalability & relational)

Sudeepa

about 1:15 hours

30 mins

Causal Fairness

Causal Explainability

Babak

about 1:15 hours

Outline

Sudeepa

Introduction

Pearl's Graphical Causal Model

Rubin's Potential Outcome Framework

Briefly: some recent research on causal inference techniques (scalability & relational)

Babak

Causal Fairness

Causal Explainability

“Causal Analysis” is Important

Clinical Trial

Does quitting smoking reduce insurance premium?

The New York Times

Opinion

OP-ED CONTRIBUTOR

Social Programs That Work

Sam Island

By Ron Haskins

Dec. 31, 2014

At 24 mostly rural locations in Florida, Wyman’s Teen Outreach Program works with 6,000 ninth graders a year to promote healthy behaviors, life skills and a sense of purpose. Evaluat program, which is based on a nine-month curriculur helped reduce teen pregnancies and lowered the risl suspension and dropout.

At 160 elementary schools in low-income communities in California, Colorado, Maryland, New York, Oklahoma, South Carolina, Texas, Washington and the District of Columbia, a program called Reading Partners pairs volunteer tutors with children for twice-weekly 45-minute sessions. An evaluation of the schools across three states by the research firm d substantial improvements in reading skills.

Do reading sessions by volunteers help improve reading skills of children?

performance level are grouped together and receive daily, 90-minute reading classes, as well as one-on-one tutoring and cooperative learning activities. We know it works because a study that randomly assigned 41 schools across 11 states to an experimental or control group found impr including comprehension, in students in tl Most of the students were black or Hispa families. Success for All was awarde double its network of schools over fi to improve effectiveness in new sites

Social Studies

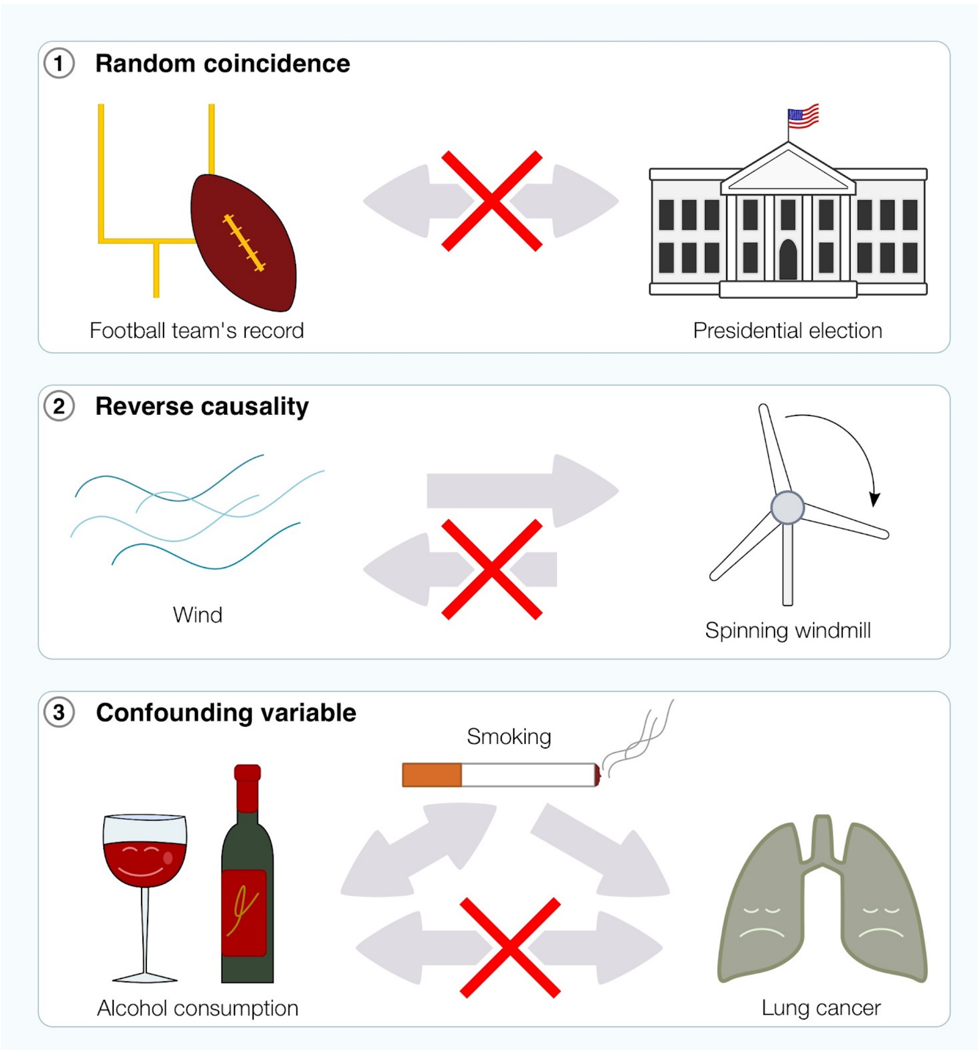
In Lancaster County, Pa., the Nurse-Family Partnership serves 175 low-income, first-time moms. Nurses start visiting the mothers

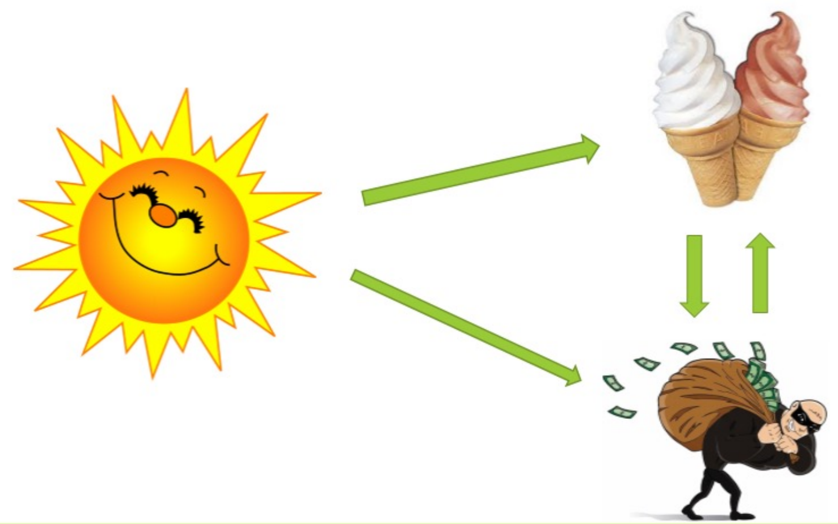
Do home-visits of expecting mothers by nurses help children’s well being later?

planning future pregnancies — and on life skills. Typically, 20 to 30 visits are involved. Three randomized controlled trials have shown that the program has major impacts that last at least until the child is 15. The mothers who participated were less likely to abuse or neglect their kids, and more likely to be working, and their kids were more likely to be healthy and ready for school.

https://www.nytimes.com/2015/01/01/opinion/social-programs-that-work.html

Correlation ≠ Causation





https://sitn.hms.harvard.edu/flash/2021/when-correlation-does-not-imply-causation-why-your-gut-microbes-may-not-yet-be-a-silver-bullet-to-all-your-problems/

What is Causality?

Causality: A (really) long history

Aristotle
(384-322 BC)
Metaphysics / Four Causes

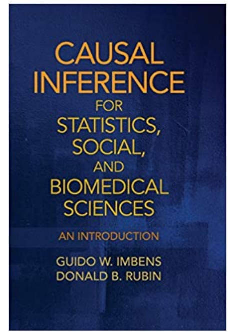
David Hume
(1738)
A Treatise of Human Nature

Karl Pearson
(1911)
The Grammar of Science, 3rd ed.

Jerzy Neyman
(1923)
Master's thesis: On the Application of Probability Theory to Agricultural Experiments. Essay on Principles

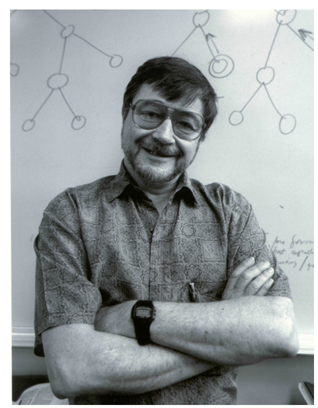
- "*We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.*" — Aristotle
- "..before we can accept [any cause of a progressive change] as a factor we must have not only shown its plausibility but if possible have demonstrated its quantitative ability" - Pearson
- "*…Thus we remember to have seen that species of object we call Flame, and to have felt that species of sensation we call Heat. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one Cause and the other Effect, and infer the existence of the one from that of the other.*" -- Hume

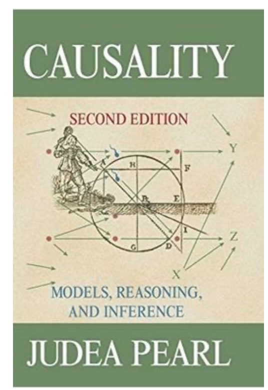
Two Popular *Formal* Causal Models

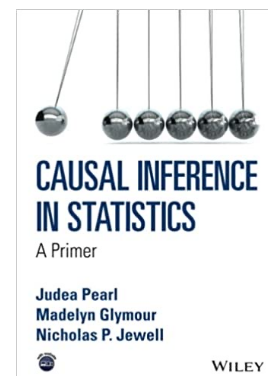


Both are used in research and practice in recent times

Neyman-Rubin Potential Outcome Model (1974 - …..)
(Statistics)







Pearl's Graphical Causal Model (1985, 1999 - ….)
(AI)

Gold standard of causal inference: Controlled Trial

Clinical Trial

The New York Times

Opinion

OP-ED CONTRIBUTOR

Social Programs That Work

Sam Island

By Ron Haskins

Dec. 31, 2014

At 24 mostly rural locations in Florida, Wyman's Teen Outreach Program works with 6,000 ninth graders a year to promote healthy behaviors, life skills and a sense of purpose. Evaluat program, which is based on a nine-month curriculur helped reduce teen pregnancies and lowered the risl suspension and dropout.

At 160 elementary schools in low-income communities in California, Colorado, Maryland, New York, Oklahoma, South Carolina, Texas, Washington and the District of Columbia, a program called Reading Partners pairs volunteer tutors with children for twice-weekly 45-minute sessions. An evaluation of the program in 19 schools across three states by the research firm M.D.R.C. found substantial improvements in reading skills.

Success for All, a comprehensive schoolwide reform progra primarily for high-poverty elementary schools, emphasizes detection and prevention of reading problems before they become serious. Students of various ages who read at the same performance level are grouped together and receive daily, 90-minute reading classes, as well as one-on-one tutoring and cooperative learning activities. We know it works because a study that randomly assigned 41 schools across 11 states to an experimental or control group found impr including comprehension, in students in tl Most of the students were black or Hispar families. Success for All was awarded $50 double its network of schools over five yea to improve effectiveness in new sites.

Social Studies

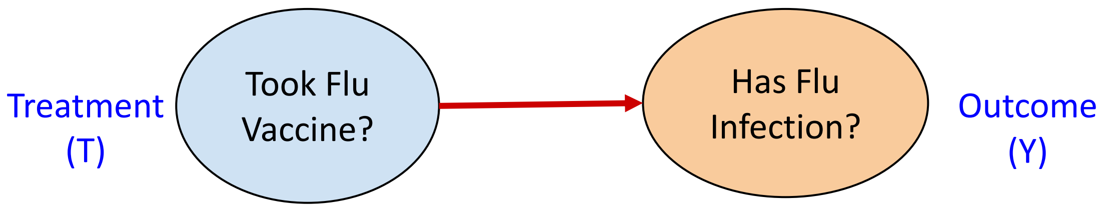
In Lancaster County, Pa., the Nurse-Family Partnership serves 175 low-income, first-time moms. Nurses start visiting the mothers before birth and continue, with diminishing frequency, until the child is 2. The nurses are trained to form a close relationship with the mother and advise her on prenatal health and child-rearing issues — including smoking and drinking during pregnancy and planning future pregnancies — and on life skills. Typically, 20 to 30 visits are involved. Three randomized controlled trials have shown that the program has major impacts that last at least until the child is 15. The mothers who participated were less likely to abuse or neglect their kids, and more likely to be working, and their kids were more likely to be healthy and ready for school.

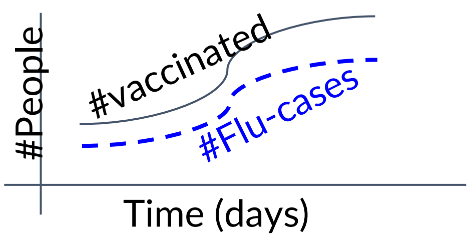
https://www.nytimes.com/2015/01/01/opinion/social-programs-that-work.html

How does controlled trial help?

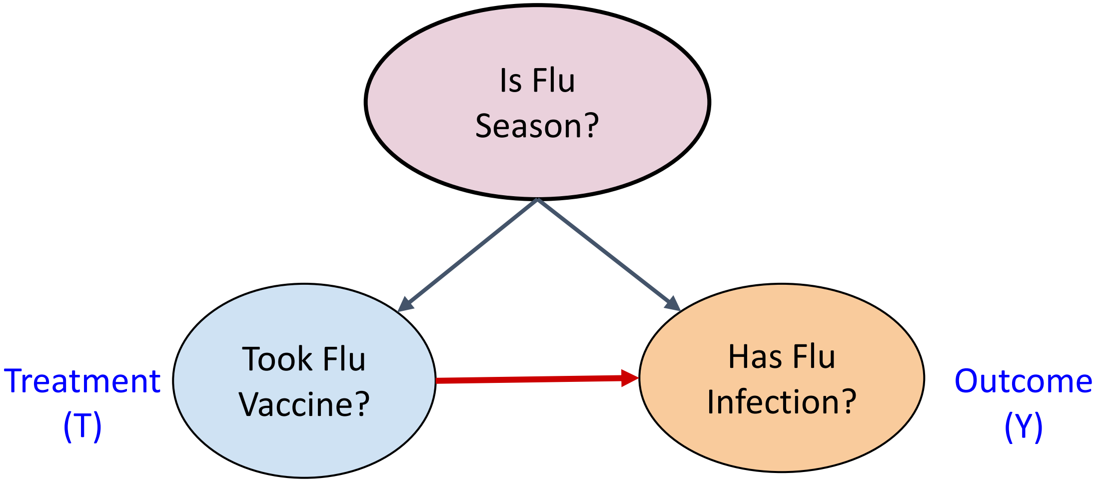
… we will see using
Neyman-Rubin Potential Outcome Framework

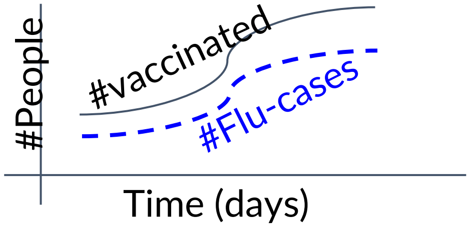
Treatment (T) & Outcome (Y)





(again) Correlation vs. Causation



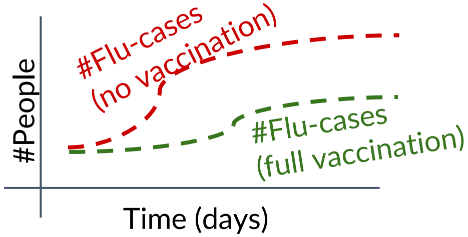


Positive Correlation

During a Flu Season -

- More Flu Infection
- More Flu Vaccination

→ Doesn't Imply Vaccines causes Flu!

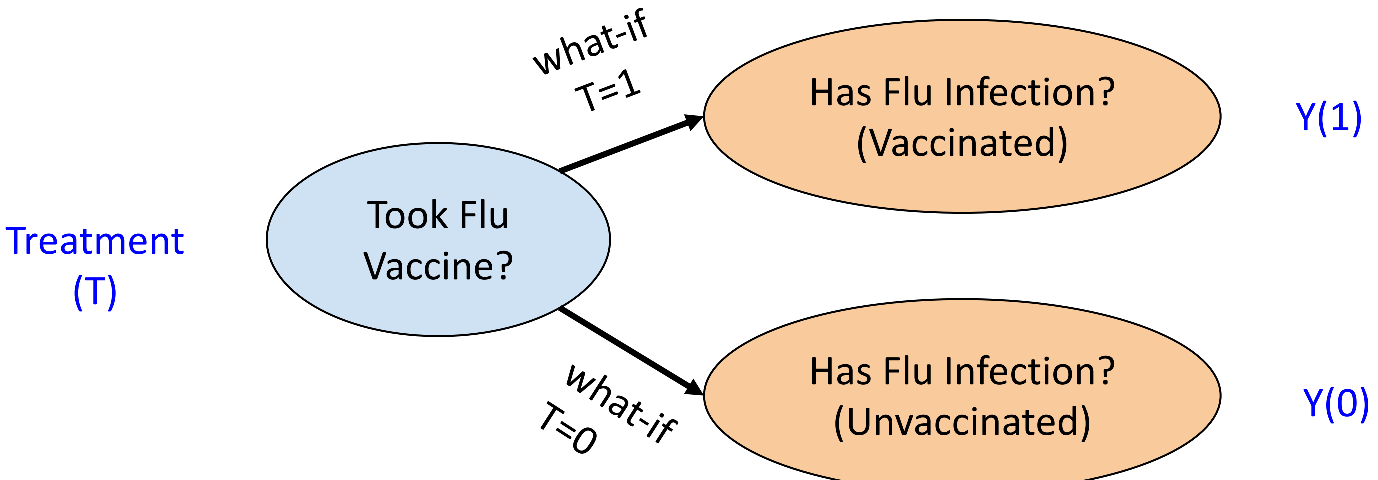


Causation (intervention)

During a Flu Season -

- What-if no-one was vaccinated?
  - Will the number of cases be more?
- What-if everyone was vaccinate?
  - Will the number of cases be small?

Goal: Average Treatment Effect



"Potential" Outcomes (only one is observed)

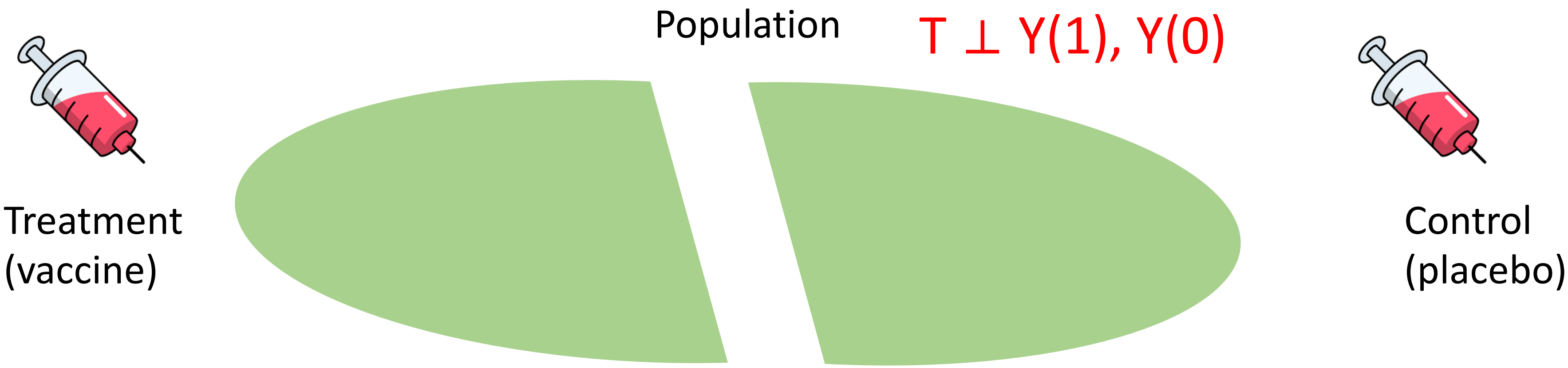
Average Treatment Effect (ATE) = **E**[Y(1) - Y(0)]

Controlled Trial = Randomized Experiments

Population

[Rubin, ‘74 ‘05]

Controlled Trial = Randomized Experiments



[Rubin, ‘74 ‘05]

Controlled Trial = Randomized Experiments

Population

$T \perp Y(1), Y(0)$

Treatment
(vaccine)

Control
(placebo)

Can be estimated from experimental observed data

Average Treatment Effect (ATE) = $\mathbf{E}[Y(1) - Y(0)]$

$= E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0]$

[Rubin, '74 '05]

Controlled Trial = Randomized Experiments

Treatment
(vaccine)

Population $T \perp Y(1), Y(0)$

Control
(placebo)

Can be estimated from experimental observed data

Average Treatment Effect (ATE) = $\mathbf{E}[Y(1) - Y(0)]$

$= E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0]$

Always Possible?

[Rubin, '74 '05]

Randomized Experiments not always feasible

1. Infeasibility or high cost
   - e.g., how allocation of government funding in different research areas will affect the number of academic jobs in these areas

2. Ethical reasons
   - e.g., effect of availability to better resources during childhood on higher education in the future

3. Prohibitive delay
   - e.g., effect of childhood cholesterol on teen obesity

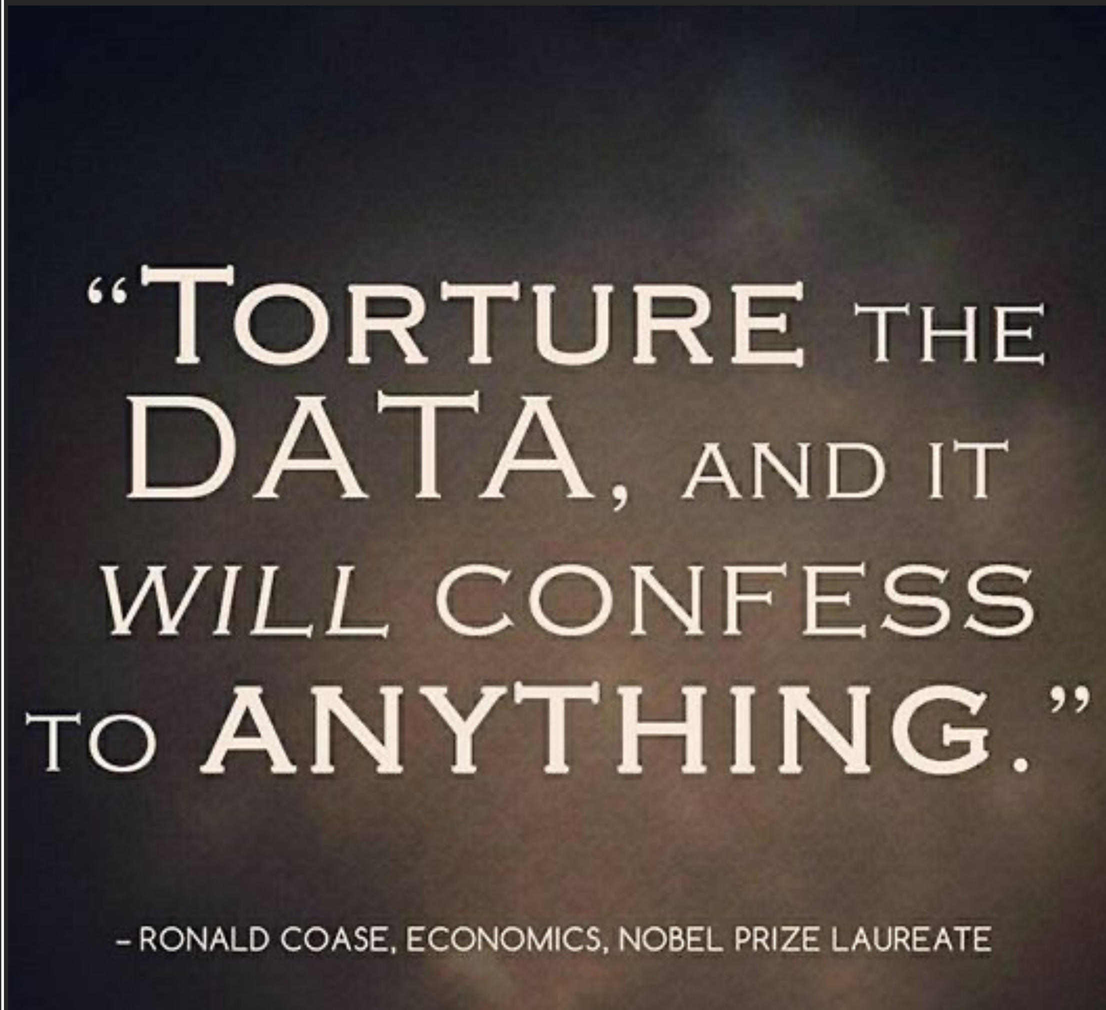
4. Experiments can be on a small population, may have a large variance, or may not be possible for the targeted population

Controlled trials not always possible

Solution: Infer Causality from "Observed Data" (Observational Studies)

Possible by both causal models: Pearl's and Rubin's

So we want to infer causality from observed data!


"TORTURE THE DATA, AND IT WILL CONFESS TO ANYTHING."
– RONALD COASE, ECONOMICS, NOBEL PRIZE LAUREATE


Simpson Paradox

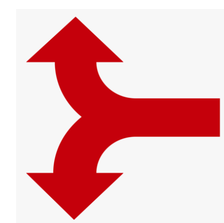
Q: Does "Gender" affect admission decision?

	Admitted	Total	%
Male	27	100	27%
Female	60	200	30%

Department A

	Admitted	Total	%
Male	150	200	75%
Female	78	100	78%

Department B



	Admitted	Total	%
Male	177	300	59%
Female	138	300	46%

Total

Which version do we report? What variables to condition on?

Focus for the first half of the tutorial!

How can we do
sound causal analysis
from
observed data

Outline

Sudeepa

- Introduction
- **Pearl's Graphical Causal Model**
- Rubin's Potential Outcome Framework
- Briefly: some recent research on causal inference techniques (scalability & relational)

Babak

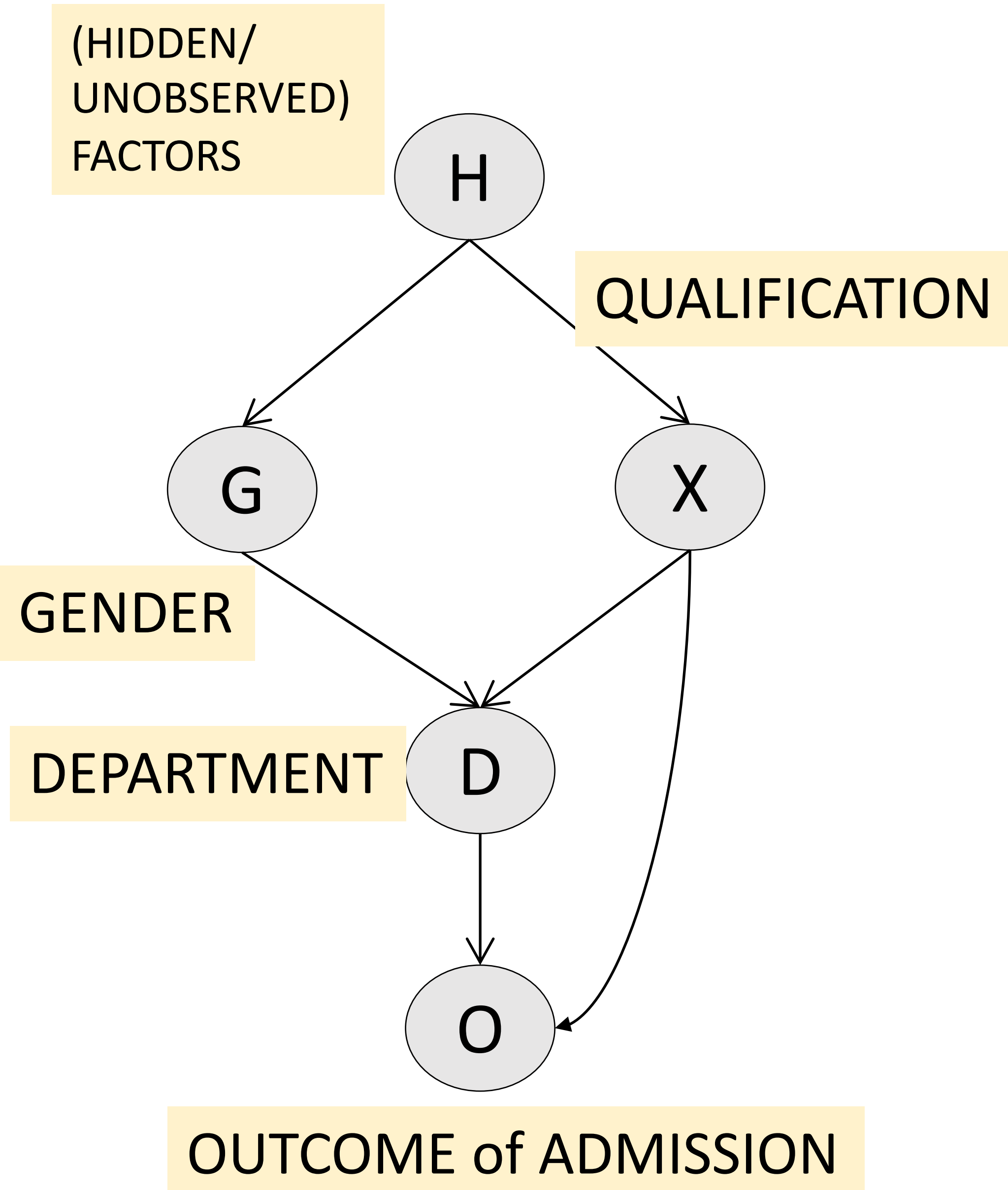
- Causal Fairness
- Causal Explainability

Review: Probability

- $\Pr(X = x)$
- $\Pr(AB) = \Pr(A \wedge B)$
- $\Pr(A|B) = \Pr(A \wedge B) / \Pr(B)$
- $\Pr(A \mid B) = \Pr(B \mid A) \Pr(A) / \Pr(B)$ --- Bayes' Rule
- If A and B are independent
  - $\Pr(A \wedge B) = \Pr(A)\Pr(B)$
  - $\Pr(A|B) = \Pr(A)$

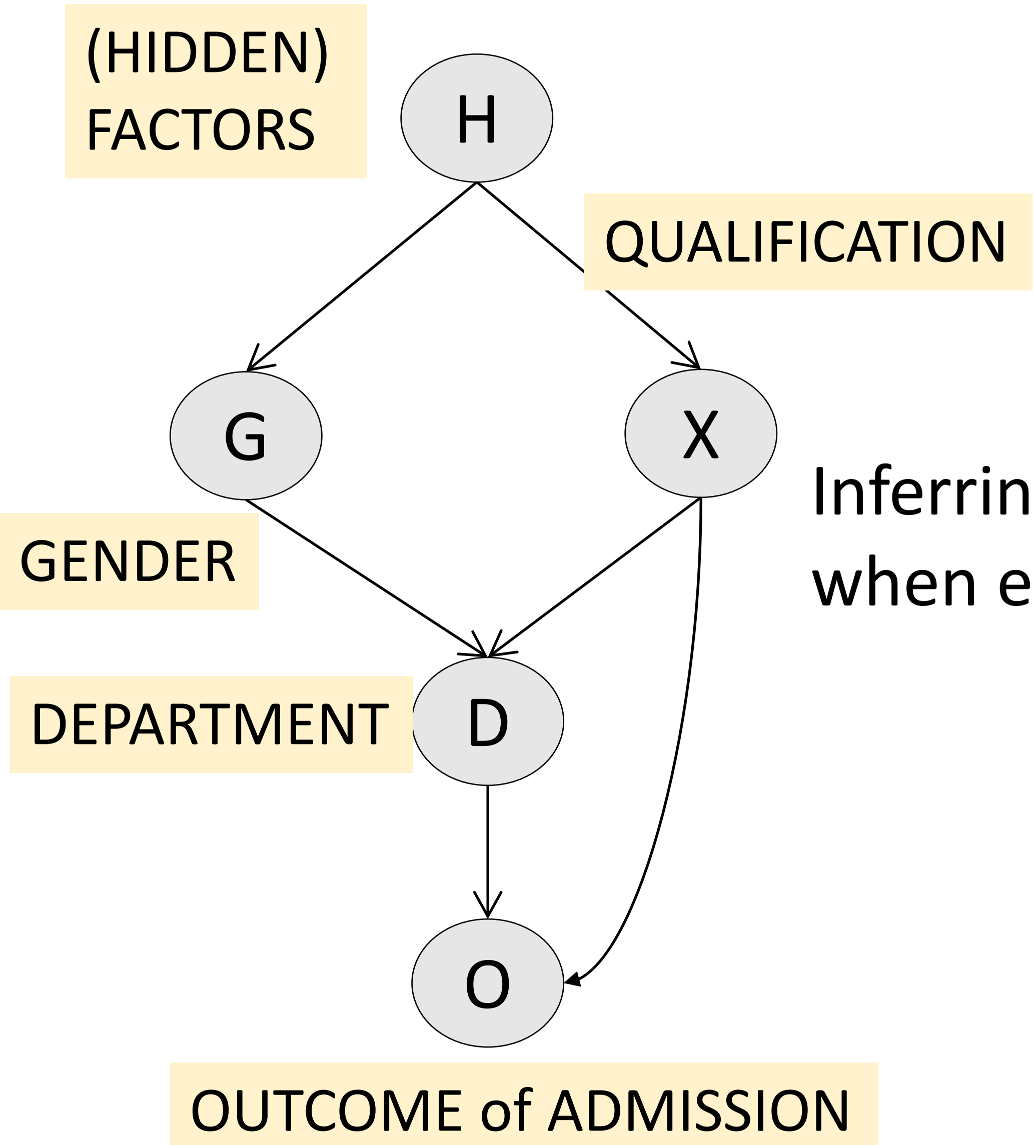
Default assumption

Review: Directed Acyclic Graphs



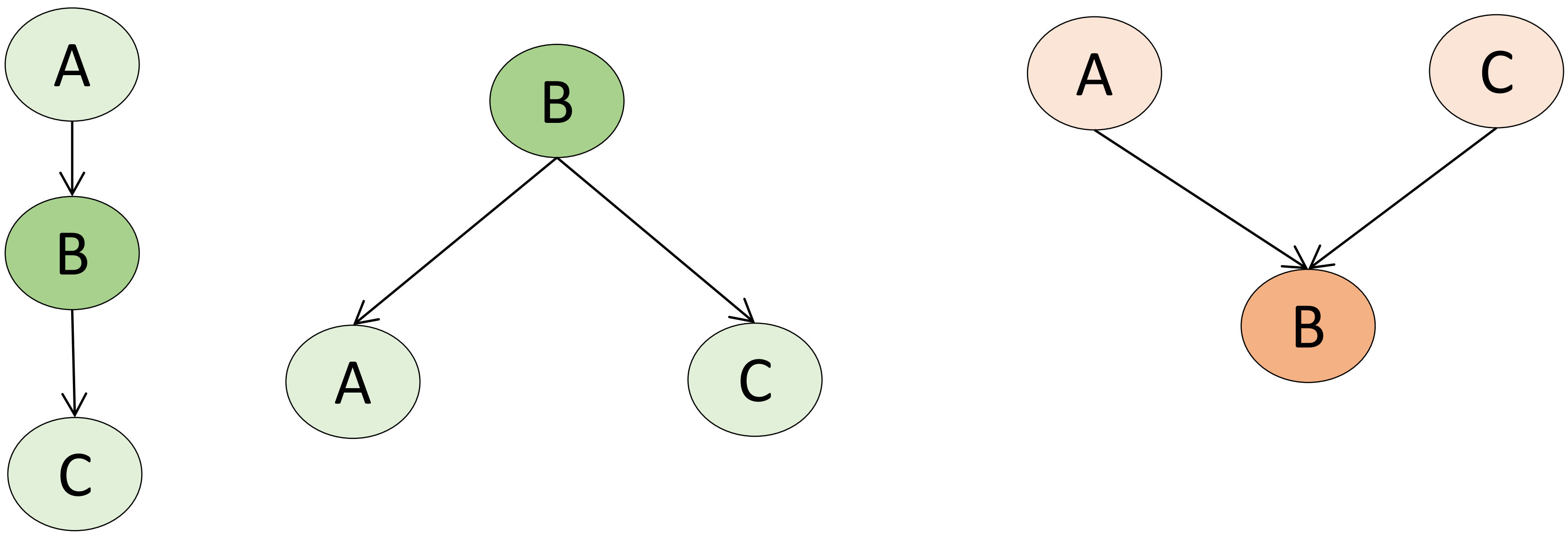
- Parent
  - H is a parent of X
- Child
  - X is a child of H
- Ancestor
  - H is an ancestor of D
- Descendant
  - D is a descendant of H
- Path (directed & undirected)
  - Directed: H → X → D → O
  - Undirected: X – D – G – H

Next: Concepts from (Directed) Graphical Models



Inferring conditional independence when events are nodes

Chain, Fork, Collider

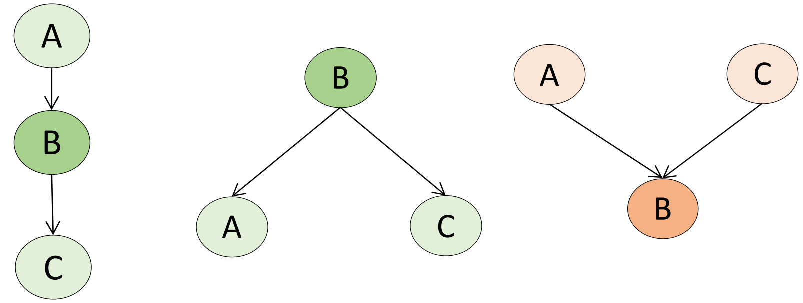


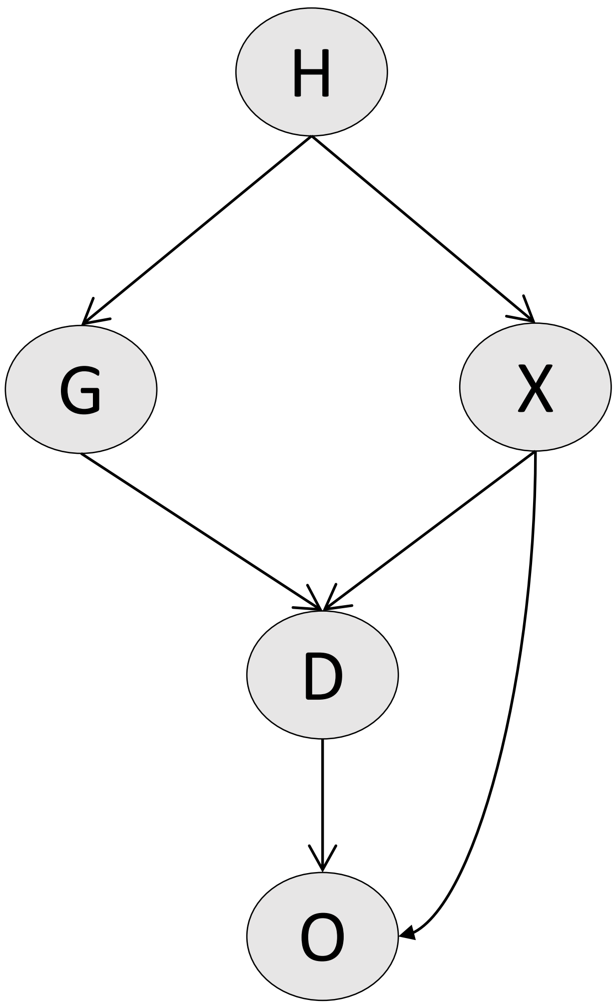
Chain Fork Collider

A & C are correlated
A & C are independent conditioned on B

A & C are independent
A & C are correlated conditioned on B
or any descendant of B
(B “explains away” A & C)

Blocking a path

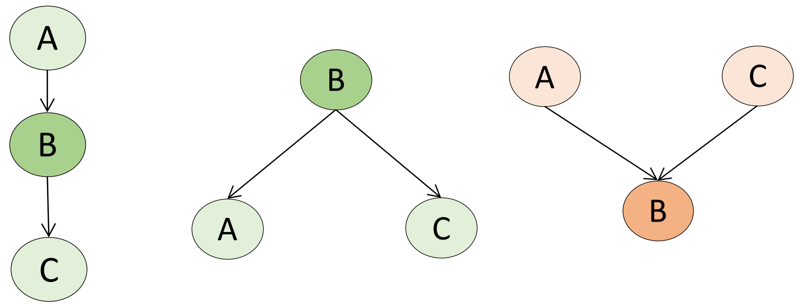


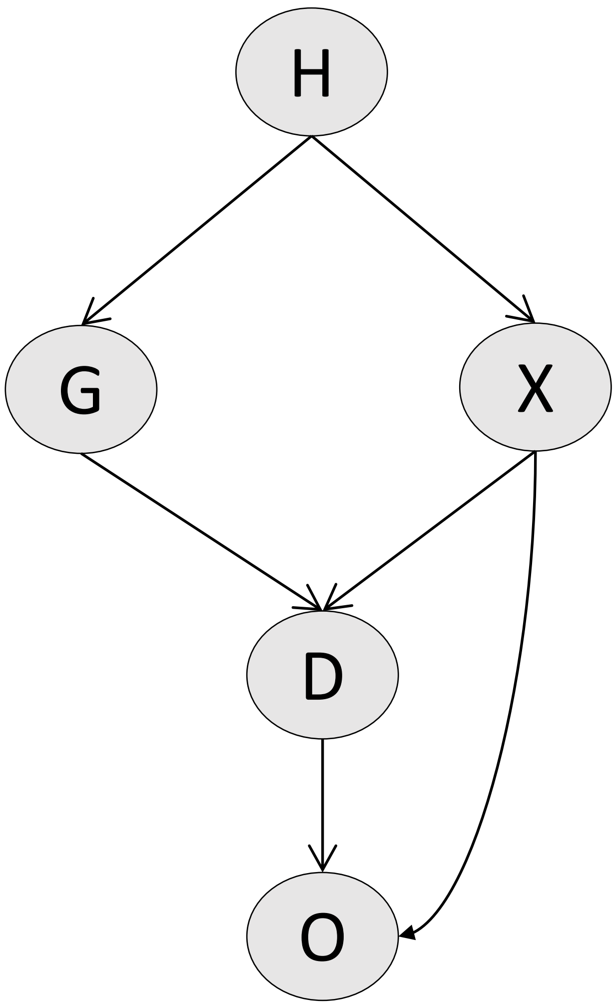


A path p is blocked by a set of nodes Z if

- P contains a chain of the form A→B→C, or a fork of the form A←B→C such that B∈Z,
  or
- p contains a collider node B of the form A → B ← C such that neither B nor any descendants of B is in Z.

Blocking a path



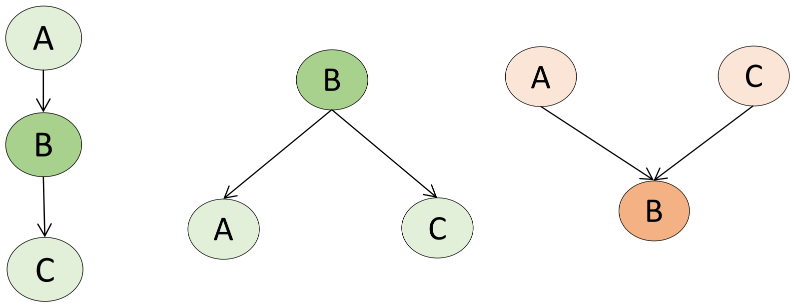


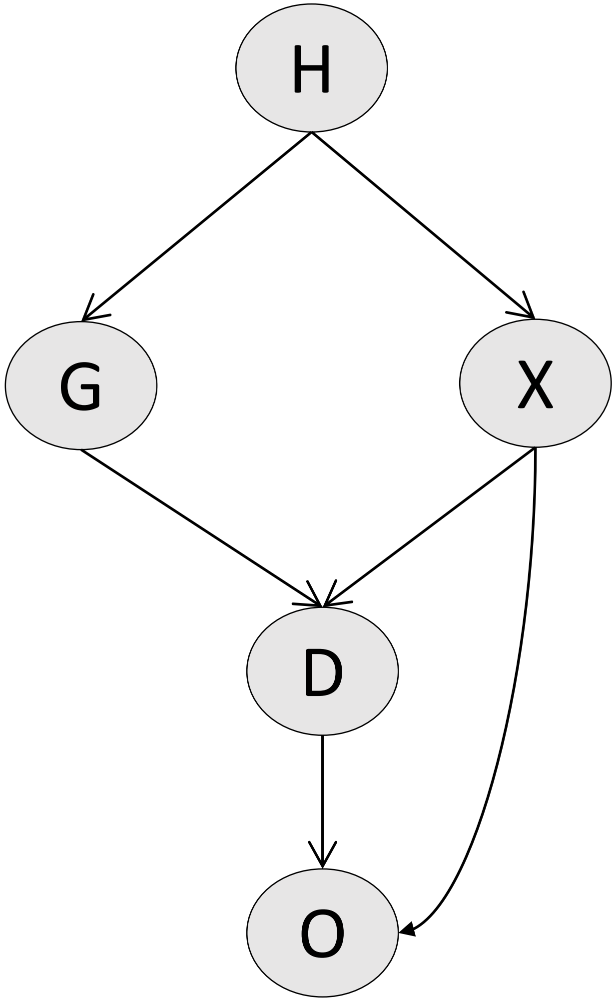
A path p is blocked by a set of nodes Z if

- P contains a chain of the form A→B→C, or a fork of the form A←B→C such that B∈Z,
  or
- p contains a collider node B of the form A → B ← C such that neither B nor any descendants of B is in Z.

X blocks the path H – X – D – O
G blocks the path D – G – H – X
D unblocks the path H – G – D – X
{DG} blocks the path H – G – D – X

d-Separation

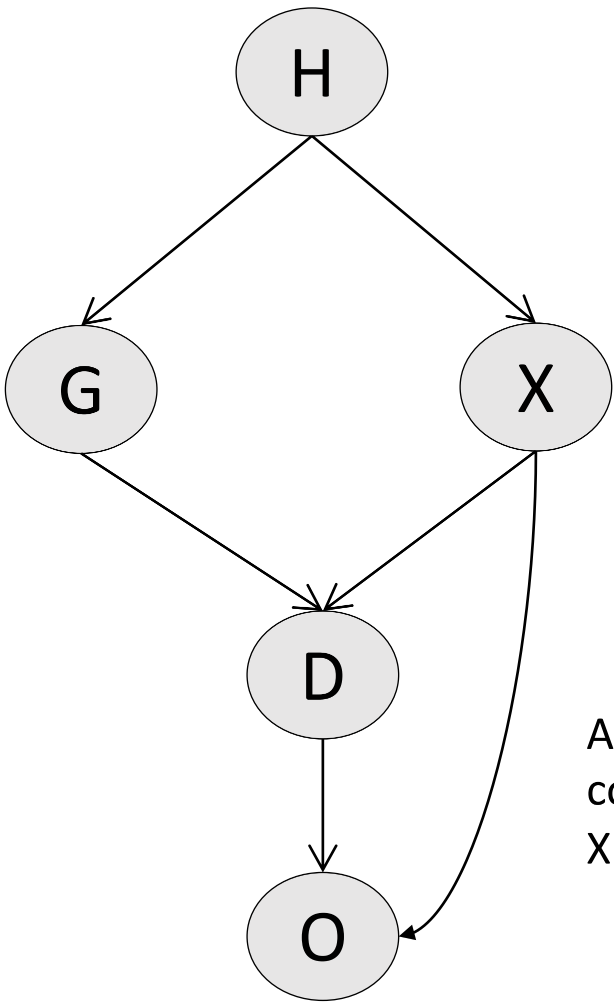




If a set of nodes Z blocks every path between two nodes X and Y, then X and Y are d-separated conditioned on Z

H and D are d-separated by {XG}
G and X are d-separated by {H}
G and X are NOT d-separated by {HD}

d-Separation and Conditional Independence



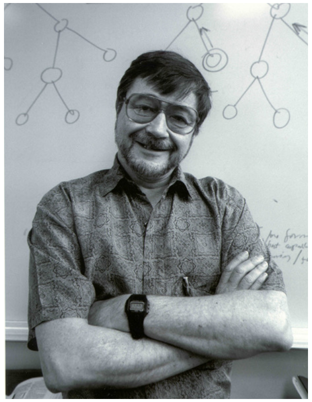
If a set of nodes Z blocks every path between two nodes X and Y, then X and Y are d-separated conditioned on Z

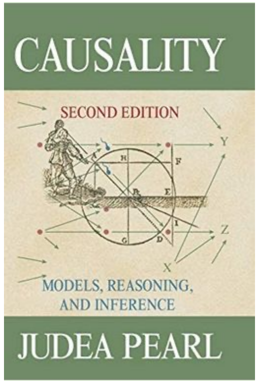
H and D are d-separated by {XG}
G and X are d-separated by {H}
G and X are NOT d-separated by {HD}

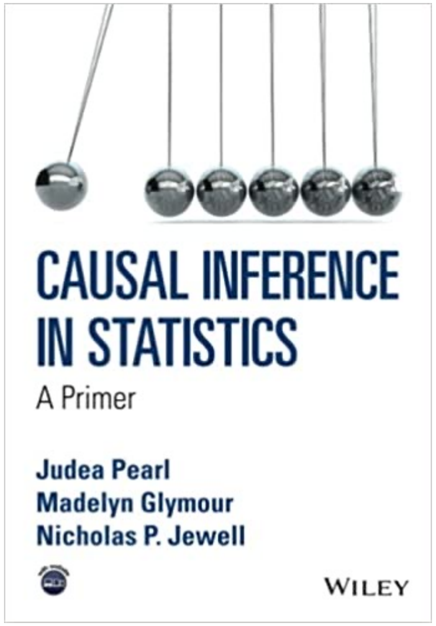
A probability distribution Pr and DAG G are Markov compatible: if X and Y are d-separated conditioned on Z, then X and Y are also conditionally independent given Z in Pr

H and D are conditionally independent given {XG}

Special case: Independence in Bayesian Network:
A node is conditionally independent of its non-descendants given its parents







Main reference used here

Structural, Graphical, and Probabilistic Causal Models

Structural Causal Model

- $M = \langle U, V, F \rangle$
  - a set of observable or endogenous variables V that are inside the model,
  - a set of noise or exogenous variables U that are outside of the model, and
  - a set of structural equations F, one F_X for each endogenous variable $X \in V$ The structural equations assign every endogenous variable a value based on other endogenous and exogenous variables.
  - $F_X : Dom(Pa_V(X)) \times Dom(Pa_U(X)) \rightarrow Dom(X)$

Endogenous parents of X

Exogenous parents of X

Domain

Structural Causal Model as a Graphical Causal Model

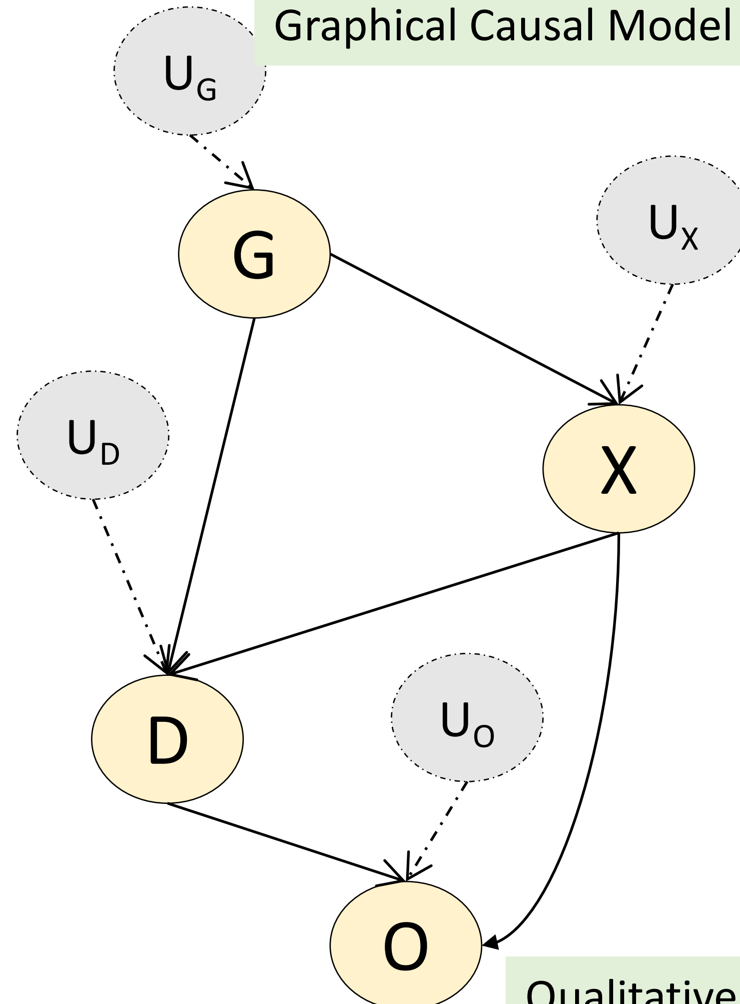
- $M = \langle U, V, F \rangle$
- Endogenous (observable) variables $V = \{G, X, D, O\}$
- Exogenous (noise) variables $U = \{U_G, U_X, U_D, U_o\}$
- Structural equations F:

$$\{G = F_G(U_G),$$
$$X = F_X(U_X, G),$$
$$D = F_D(U_D, G, X),$$
$$O = F_O(U_O, X, D)\}$$

Can be linear, exp, …

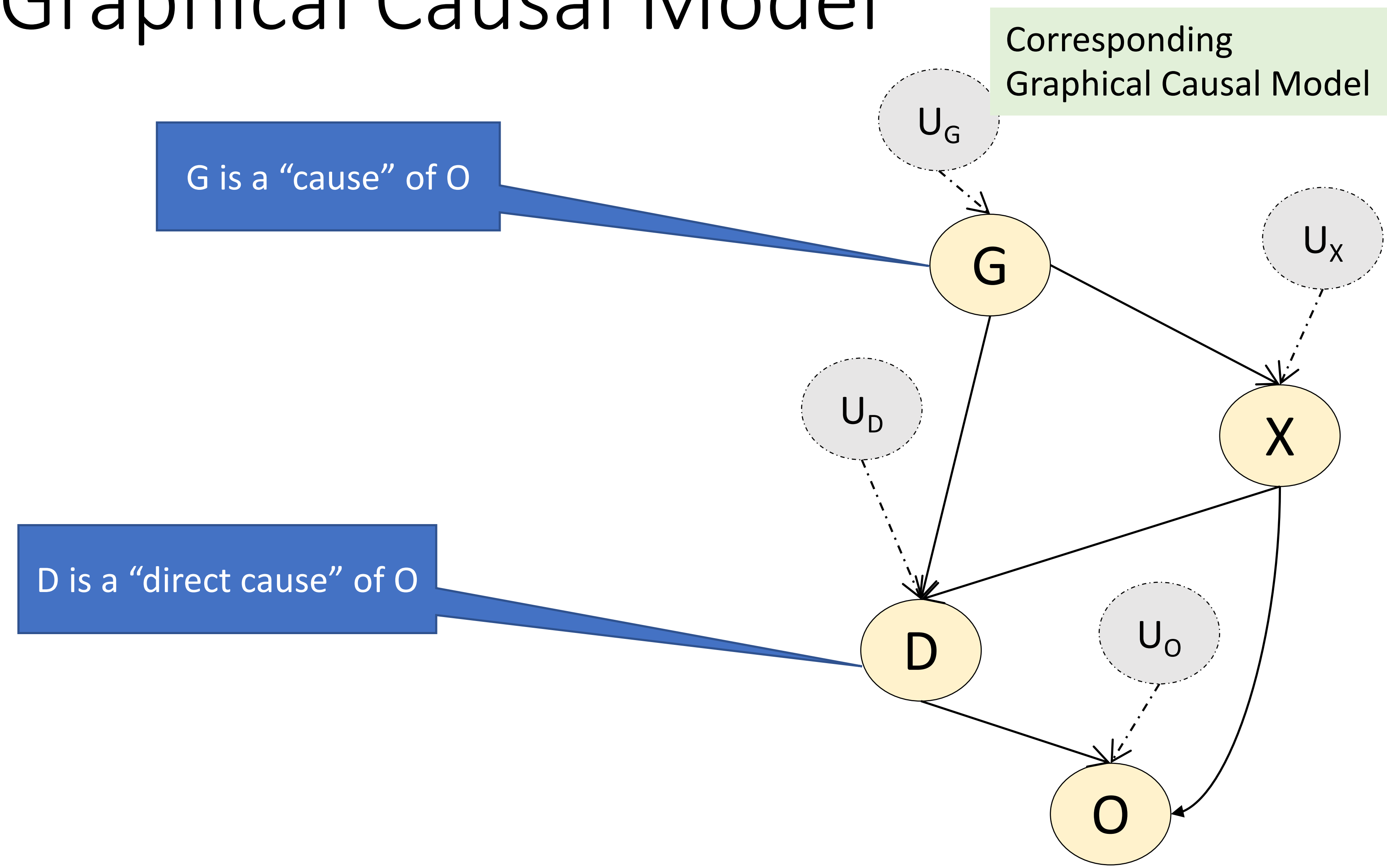
Quantitative

Corresponding Graphical Causal Model



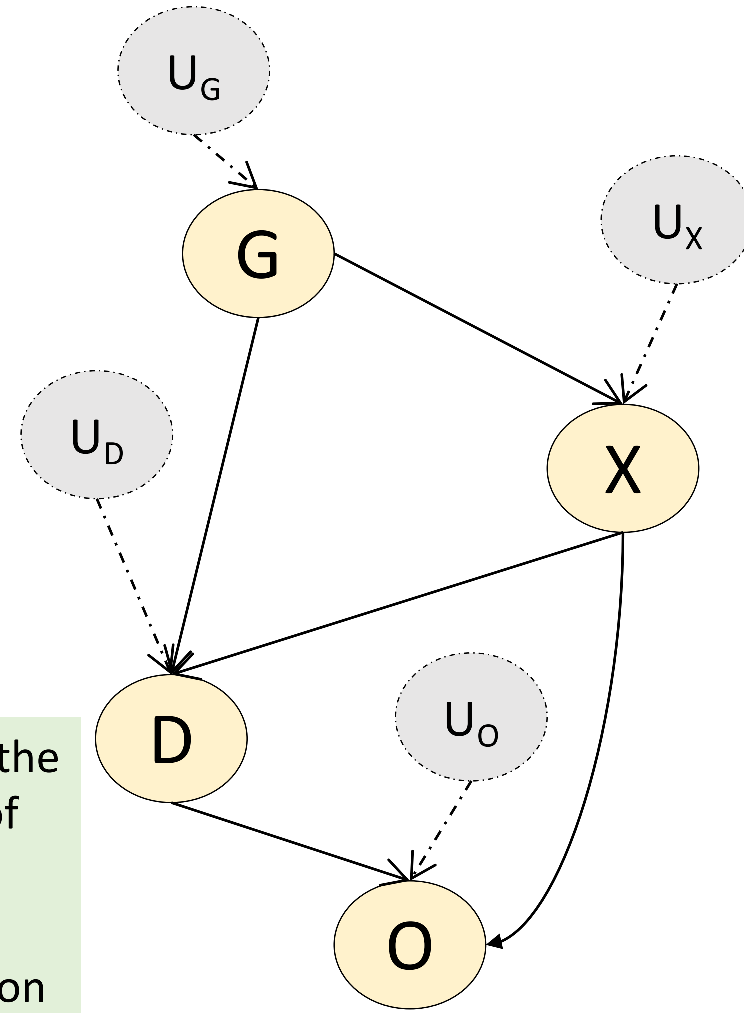
Qualitative

Structural Causal Model as a Graphical Causal Model



Structural/Graphical Causal Model to Probabilistic Model

- <M, Pr>
- M is a Structural Causal Model
- Pr is the Probability distribution
  - Satisfies Causal Markov Condition
  - Conditional independence in directed graphical models
- $Pr(X_1, X_2, \ldots) = \prod_i Pr(X_i \mid Pa(X_i))$



If we knew the values of the exogenous variables and the structural equations in F, we exactly know the values of endogenous V

But not in practice – so assume a probability distribution Pr(U = u), which gives a Pr distribution on V

Model for “Intervention” and “Counterfactuals”

Intervention (do-operators) and Counterfactuals

Intervention:

Change the reality by setting X to x: or $X \leftarrow x$

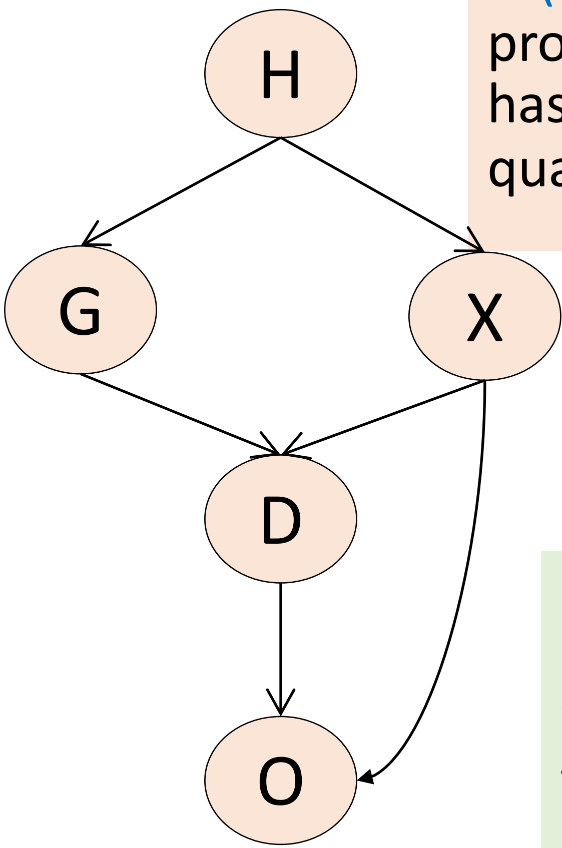
- Modeled by do-operator
- $\Pr(Y = y \mid do(X = x))$

Counterfactuals:

- "If X was set to x, what would have been the value of Y"
- $Y_{X=x}$ (or Y_x) $= y$

do-operators vs. conditional probabilities

Conditioning

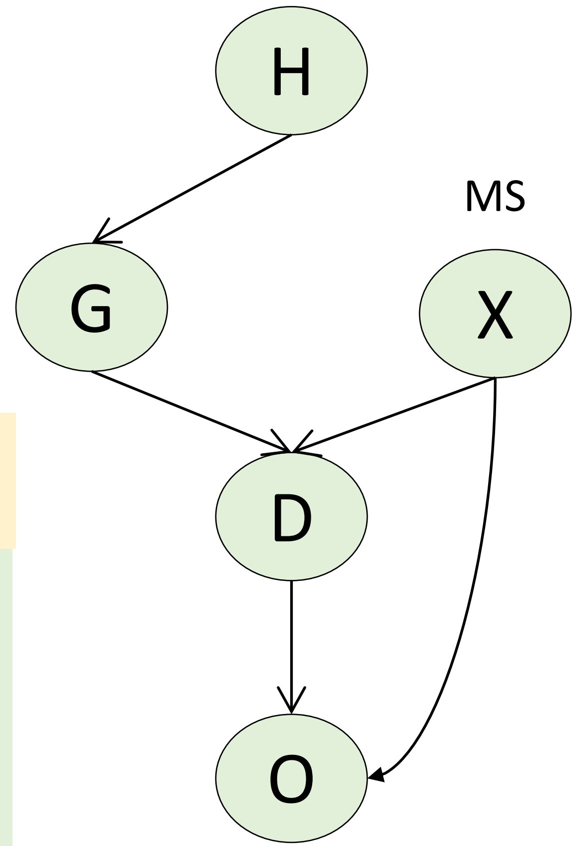


Pr(O = yes | X = MS):
probability of admission if it has been “observed” that the qualification is MS degree

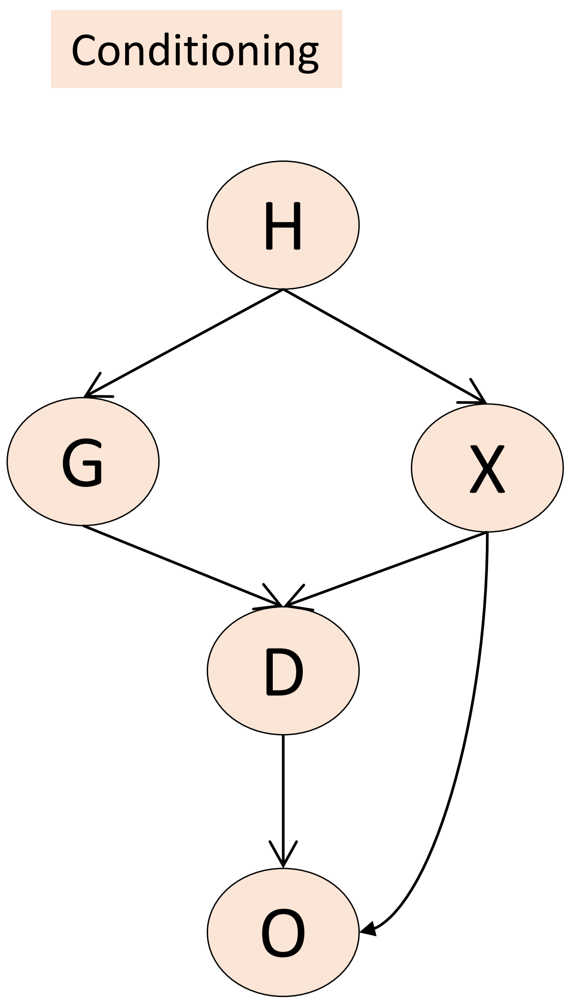
X = f(H) changes to X = MS

Pr(O = yes | do(X = MS)):
probability of admission if we have an intervention on the world and “force” the degree to be MS

Intervention



Observational Study by Pearl's Model



Goal:

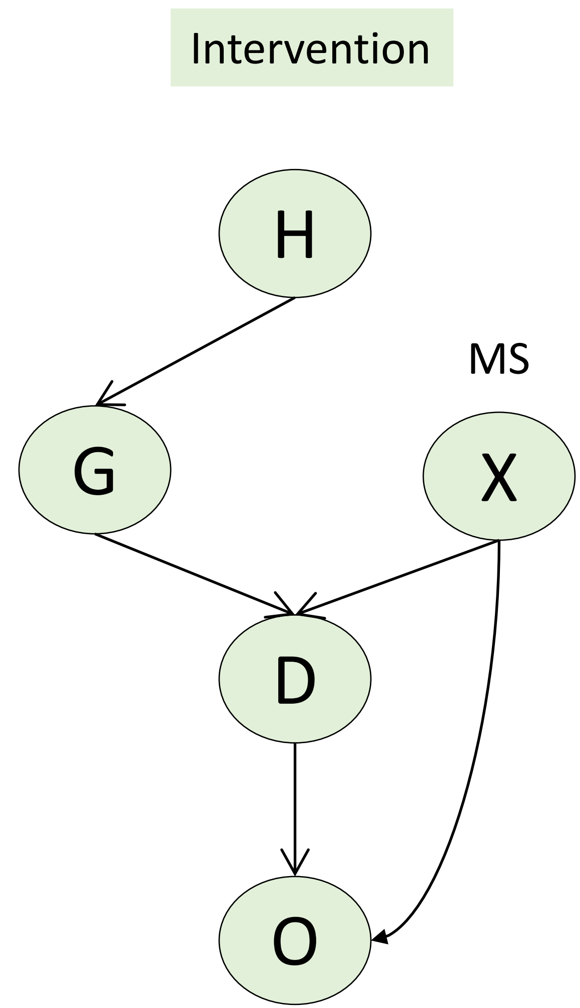
Express causal relationship as do-operators

to

conditional probabilities

- Need

1. A valid causal DAG
2. Graph Surgery
3. Observed Data



Identification

Estimation

Estimating Causal Effect

- Treatment X, Outcome Y
- Goal is to estimate causal effect $Pr(Y = y \mid do(X = x))$

- A set Z of variables is called admissible covariates for estimating the above causal effect if

$$Pr(Y = y \mid do(X = x)) = \sum_z Pr(Y = y \mid X = x, Z = z) \; Pr(Z = z)$$

In short

$$Pr(y \mid do(X = x)) = \sum_z Pr(y \mid x, z) \; Pr(z)$$

We are adjusting for Z here

Quantities of Interest

- (Recall) Average Treatment Effect ATE $E[Y(1) – Y(0)]$

Similar quantities from Pearl's work:

- Average Causal Effect:

$$\Pr(Y = y \mid do(X = x)) - \Pr(Y = y \mid do(X = x'))$$

- $E[Y \mid do(x)] = \sum_y y \Pr(y \mid do(x))$

$$= \sum_y y \sum_z \Pr(y \mid x, z) \Pr(z)$$

$$= \sum_z \sum_y y \Pr(y \mid x, z) \Pr(z)$$

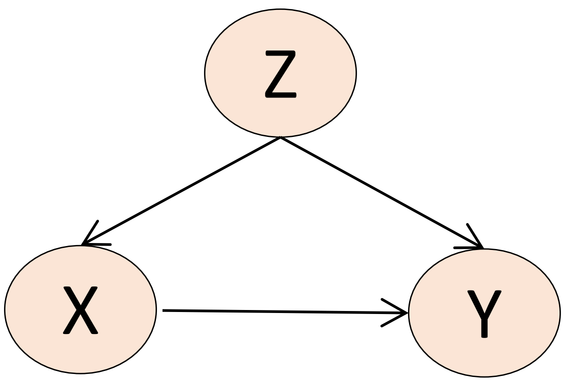
$$= \sum_z \Pr(z) E(y \mid x, z)$$

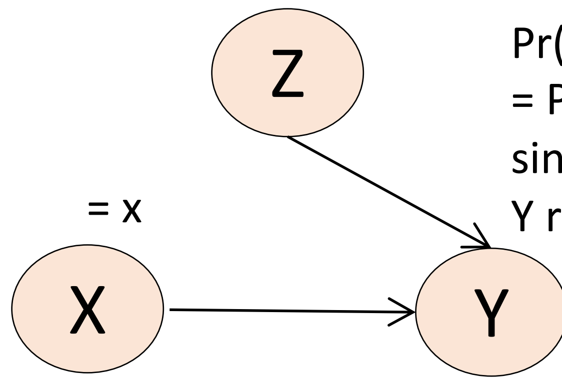
- Average difference: $E(Y \mid do(x)) - E(Y \mid do(x'))$

Example: Adjustment Formula

(a)

$Pr(Z = z) = Pr_m(Z = z)$, since Z is not affected by removing the arrow to X





$Pr(Y = y \mid X = x, Z = z)$ (b)
$= Pr_m(Y = y \mid X = x, Z = z)$
since the process of generating Y remains the same in both

$Pr(Y = y \mid do(X = x)) = Pr_m(Y = y \mid X = x)$. = Probability in the manipulated model

(c) Z and X are d-separated in the modified model, hence independent:

$Pr_m(Z = z \mid X = x) = Pr_m(Z = z)$

$Pr(Y = y \mid do(X = x))$
$= Pr_m(Y = y \mid X = x)$
$= \sum_z Pr_m(Y = y \mid X = x, Z = z)\, Pr_m(Z = z \mid X = x)$
$= \sum_z Pr_m(Y = y \mid X = x, Z = z)\, Pr_m(Z = z)$ ---- by (c)
$= \sum_z Pr(Y = y \mid X = x, Z = z)\, Pr(Z = z)$ ---- by (a), (b)

Now can be estimated from the observed data!

How do we find admissible covariates?

Sufficient Condition

Back-door Criterion

- Graphical Causal Model G, Treatment X, Outcome Y
- A subset of variables Z is admissible for estimating causal effect from X to Y if it satisfies two conditions

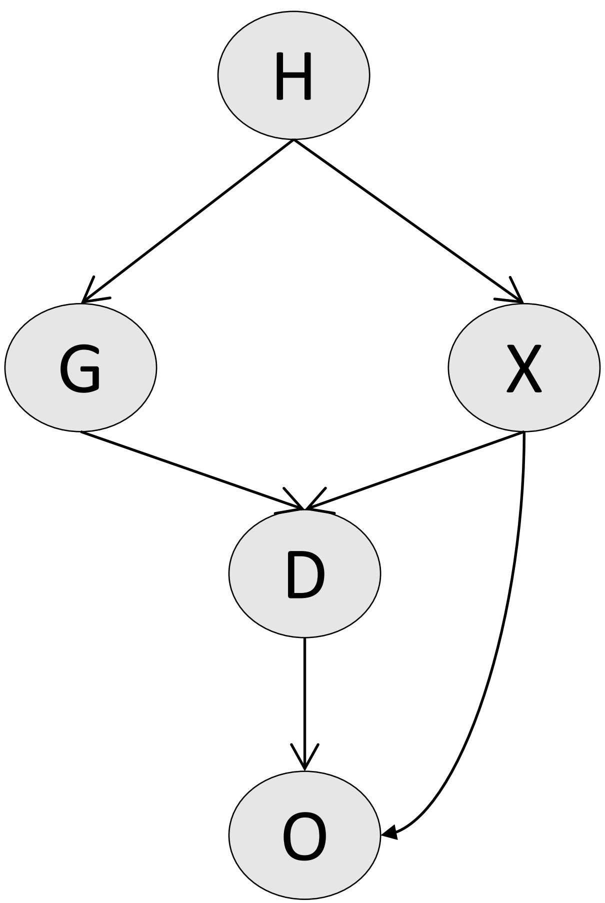
1. No element of Z is a descendant of X
2. Z "blocks" all paths between X and Y that end with an arrow pointing to X (back door paths from X to Y)

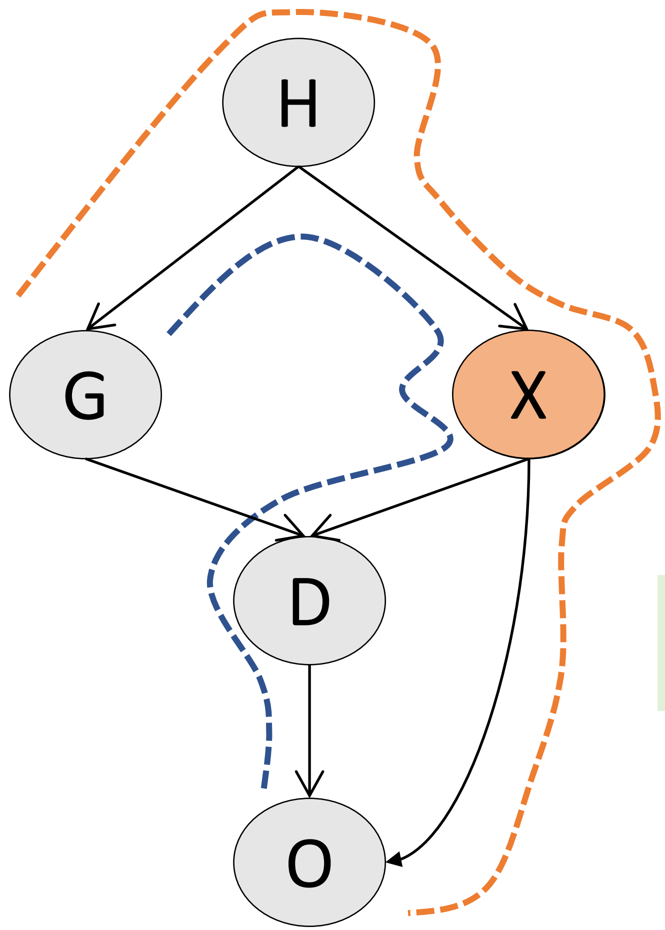
Then

$Pr(Y = y \mid do(X = x)) = \sum_z Pr(Y = y \mid X = x, Z = z)\, Pr(Z = z)$

Example: Back door criterion

- Treatment G, Outcome O
- X satisfies the backdoor criterion



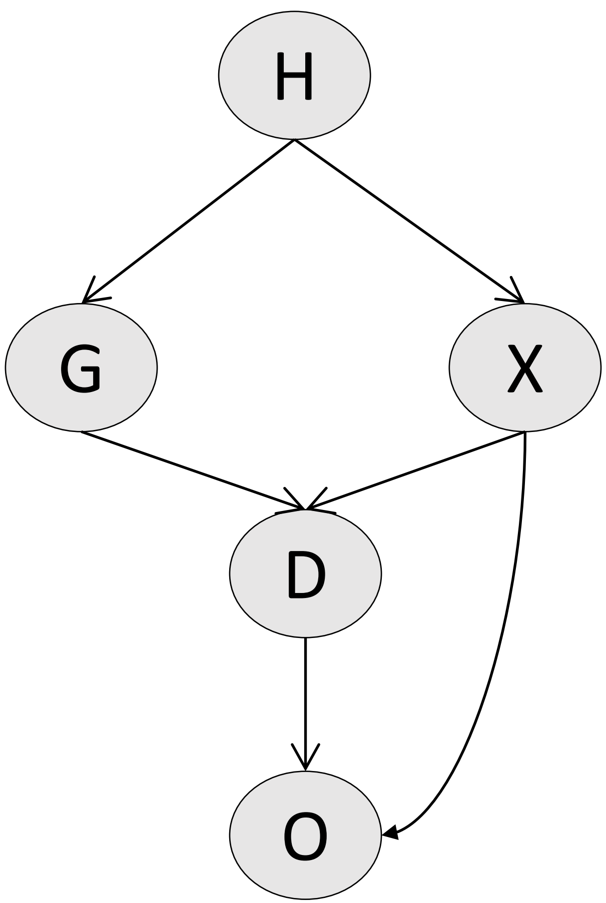


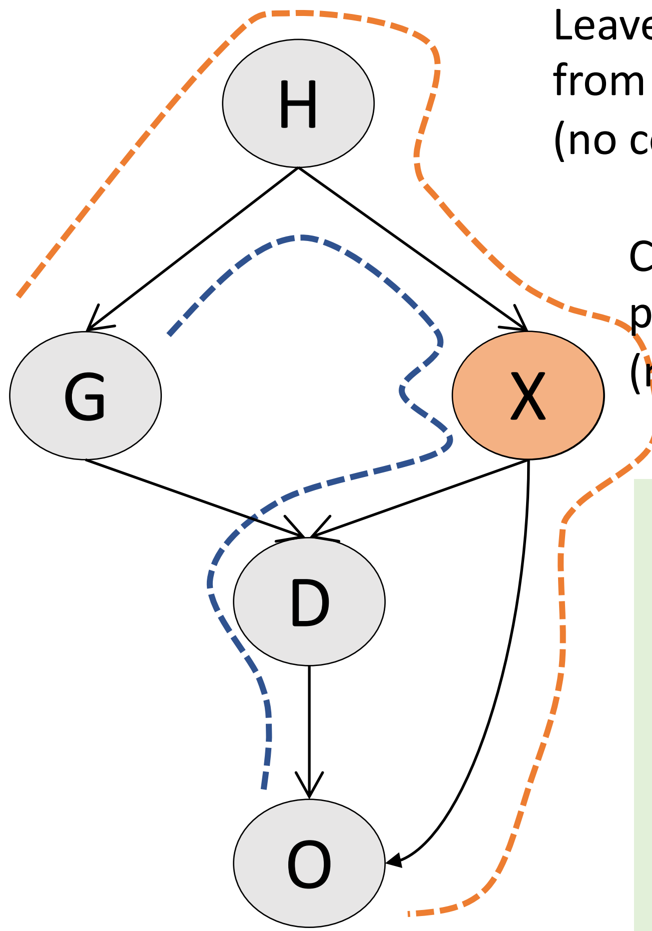
Two backdoor Paths between G and O

Both are blocked by X

Example: Back door criterion

- Treatment G, Outcome O
- X satisfies the backdoor criterion





Block all spurious paths between G & O (not causal)

Leave all directed paths from G to O unperturbed (no condn. on desc. of G)

Create no spurious paths between G & O (no condn. on collider)

Parents of G always satisfies backdoor

but they may be unmeasured like H

Then look for another variable like X

Counterfactuals

- "If X was set to x, what would have been the value of Y"

  $Y_{X=x}$ (or Y_x)
- An "if" statement where the if-portion is not true (counterfactual or hypothetical or retrospective estimate)
- $E(Y \mid do(X = x))$: predicts the effect of intervention

- $E(Y \mid do(X = x)) = E(Y_{X=x})$
- $E(Y \mid do(X = x), Z = z) = E(Y_{X=x} \mid Z = z)$

When all variables are from the same world: the modified distribution created by do(X = x)

$E(Y_{X=1} \mid Y_{X=1} = y')$: About two different worlds – cannot be expressed as do-operator or intervention, need counterfactuals & structural equations

More on counterfactuals when Babak covers causal XAI

Outline

Introduction

Pearl's Graphical Causal Model

Rubin's Potential Outcome Framework

Briefly: some recent research on causal inference techniques (scalability & relational)

Causal Fairness

Causal Explainability

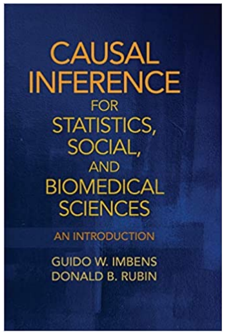
Sudeepa

Babak

Potential Outcome Framework

- Referred to as Neyman-Rubin's model or Rubin's model
  - First proposed in Neyman's Ph.D. thesis (1923)
  - A model for "Randomized Experiments" by Fisher (1920s-30s)
  - Further developed by Rubin (1978) and others

- Establish a causal relationship between a potential cause (treatment) and its effect (outcome)



Potential Outcome Model: Applications

Widely used in

- Medicine
  - Christakis and Iwashyna 2003; Rubin 1997
- Economics
  - Abadie and Imbens 2006; Galiani, Gertler, and Schargrodsky 2005; Dehejia and Wahba 2002, 1999
- Political science
  - Bowers and Hansen 2005; Imai 2005; Sekhon 2004
- Sociology
  - Morgan and Harding 2006; Diprete and Engelhardt 2004; Winship and Morgan 1999; Smith 1997
- Law
  - Rubin 2001

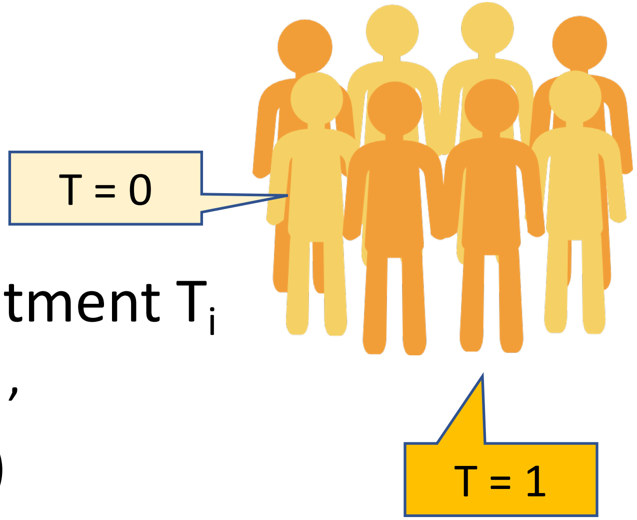
References in [Sekhon 2007]

Units

- N "units"
  - physical objects at particular points in time
  - e.g., individual people, one person at different points of time, plots of lands

Units	Covariates	Treatment assignment	Potential Outcome: Treatment	Potential Outcome: Control	Unit-level causal effects	Summary of causal effects
1	X_1	T_1	Y_{11}	Y_{01}	Y_{11} - Y_{01}	
2	X_2	T_2	Y_{12}	Y_{02}	$Y_{12} - Y_{02}$	$E[Y_1 - Y_0]$
…						
N	X_n	T_N	Y_{1N}	Y_{0N}	$Y_{1N} - Y_{0N}$	

Treatment and Control



- Each unit i can be exposed or not to a treatment T_i
  - e.g., individuals taking an Aspirin vs. placebo ,
- "Active Treatment" or "Treatment" (T_i = 1)
  - if exposed
- "Control Treatment" or "Control" (T_i = 0)
  - if not exposed

Units	Covariates	Treatment assignment	Potential Outcome: Treatment	Potential Outcome: Control	Unit-level causal effects	Summary of causal effects
1	X_1	T_1	Y_{11}	Y_{01}	Y_{11} - Y_{01}	$E[Y_1 - Y_0]$
2	X_2	T_2	Y_{12}	Y_{02}	$Y_{12} - Y_{02}$	
…						
N	X_n	T_N	Y_{1N}	Y_{0N}	$Y_{1N} - Y_{0N}$	

Covariates

- Variables that take their values before the treatment assignment
- Cannot be affected by the treatment
  - e.g., pre-aspirin headache pain, gender, blood-pressure

Units	Covariates	Treatment assignment	Potential Outcome: Treatment	Potential Outcome: Control	Unit-level causal effects	Summary of causal effects
1	X_1	T_1	Y_{11}	Y_{01}	$Y_{11} - Y_{01}$	$E[Y_1 - Y_0]$
2	X_2	T_2	Y_{12}	Y_{02}	$Y_{12} - Y_{02}$	
…						
N	X_n	T_N	Y_{1N}	Y_{0N}	$Y_{1N} - Y_{0N}$	

Potential Outcome

- $Y(1)= Y_1$ (for treatment, $T_i = 1$)
- $Y(0) = Y_0$ (for control, $T_i = 0$)
- for i-th unit : Y_{1i} and Y_{0i}
- Observed outcome $Y = T_i Y_{1i} + (1 - T_i) Y_{0i}$

Units	Covariates	Treatment assignment	Potential Outcome: Treatment	Potential Outcome: Control	Unit-level causal effects	Summary of causal effects
1	X_1	T_1	Y_{11}	Y_{01}	$Y_{11} - Y_{01}$	$E[Y_1 – Y_0]$
2	X_2	T_2	Y_{12}	Y_{02}	$Y_{12} – Y_{02}$	
…						
N	X_n	T_N	Y_{1N}	Y_{0N}	$Y_{1N} – Y_{0N}$	

Unit-level causal effect

- The comparisons of Y_{1i} and Y_{0i}
  - difference or ratio
  - Typically, Y_{1i} - Y_{0i}
- For any unit i, only one of them can be observed
  - we cannot go back in time and expose it to the other treatment
- Fundamental problem of causal inference

Units	Covariates	Treatment assignment	Potential Outcome: Treatment	Potential Outcome: Control	Unit-level causal effects	Summary of causal effects
1	X_1	T_1	Y_{11}	Y_{01}	Y_{11} - Y_{01}	$E[Y_1 – Y_0]$
2	X_2	T_2	Y_{12}	Y_{02}	$Y_{12} – Y_{02}$	
…						
N	X_n	T_N	Y_{1N}	Y_{0N}	$Y_{1N} – Y_{0N}$	

Average Treatment Effect (ATE)

- Defined for a collection of units
- e.g.,
  - the mean (or expected) unit-level causal effect -- standard
  - the median unit-level causal effect for all males
  - the difference between the median Y_{1i} and Y_{0i} for all females
- Also, Average Treatment Effect for the Treated (ATT): $E[Y_1 - Y_0 \mid T = 1]$
- Conditional Average Treatment Effect (CATE): $E[Y_1 - Y_0 \mid Z = z]$

Units	Covariates	Treatment assignment	Potential Outcome: Treatment	Potential Outcome: Control	Unit-level causal effects	Summary of causal effects (ATE)
1	X_1	T_1	Y_{11}	Y_{01}	$Y_{11} - Y_{01}$	$E[Y_1 - Y_0]$
2	X_2	T_2	Y_{12}	Y_{02}	$Y_{12} - Y_{02}$	
…						
N	X_n	T_N	Y_{1N}	Y_{0N}	$Y_{1N} - Y_{0N}$	

Recall: Randomized Controlled Experiments

[Rubin, ‘74 ‘05]

Population $T \perp Y(1), Y(0)$

Treatment (vaccine)

Control (placebo)

Can be estimated from experimental observed data

Average Treatment Effect (ATE) = **E**[Y(1) - Y(0)]

$= E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0]$

- The assigned treatment is statistically independent of any (measured or unmeasured) covariate in the population before the experiment has been started
- The distribution of any covariate is the same in the treatment and control groups
- Any difference in outcomes is due to the treatment and not any other pre-existing differences
- The average of control/treatment group outcomes is an unbiased estimate of average outcome under control/treatment for whole population

SUTVA

Stable Unit Treatment Value Assumptions – needed even for randomized experiments

- Cox 1958, Rubin 1978

1. No “interference” or “spill-over effect” among units
   - For unit i, Y_{1i} and Y_{10} are NOT affected by what action any other unit j received
2. Unique Treatment Level or “Dose”
   - There are no hidden versions of treatments
   - No matter how (mechanism) unit i received treatment 1, the outcome that would be observed would be Y_{1i} -- similarly for treatment 0

Violations of SUTVA

1. (Violation of) No interference
    - (wiki) Two units Joe and Mary for effect of a drug for high blood pressure
    - They share the same household
    - Mary cooks
    - Mary got drug (treatment) – her pressure reduces – cooks salty food
        - In practice, Mary may not know if she got the drug or placebo
    - Joe’s pressure increases
2. (Violation of) Unique Treatment Level or “Dose”
    - Different doses of the medicine for drug pressure

More assumptions

- Compliance issue
  - People assigned to treatment may refuse it
  - People assigned to control may try to get treatment
    - [Barnard, Frangakis, Hill, and Rubin 2003]
  - People started taking a medicine, then stopped in the middle because it made them too sick to work

Observational Study (Rubin's model)

- Alternative to true randomized experiments
  - Tries to simulate the ideal situation

- Create treatment and control groups that appear to be random
  - at least on observed/measured variables by choosing individuals with similar covariate values
  - do not use the outcome while selecting the groups

Observational Study

Covariates (X)

T	Y1	Y0	Age (X_1)	Race (X_2)	Gender (X_3)	State (X_4)	Edu (X_5)
1	130	?	20s	W	M	NC	College
0	?	125	20s	W	M	NC	College
1	127	?	30s	B	F	MA	PhD
0	?	130	30s	L	F	CA	PhD

Average Treatment Effect (ATE) = $\mathbf{E}[Y(1) - Y(0)]$

$\cancel{T \perp Y1, Y0} \mid X$

(strong ignorability)

[Rosenbaum-Rubin, '83]

$\cancel{= E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0]}$

$= E_X[E[Y1 \mid T = 1, X] - E[Y0 \mid T = 0, X]]$

Can be (again) estimated from observed data

Strong Ignorability or Unconfoundedness

- Treatment assignment is

"strongly ignorable given a vector of covariates X" [Rosenbaum-Rubin 1983]

if

1. $(Y1, Y0) \perp T \mid X$
2. $0 < \Pr[T = 1 \mid X] < 1$

- E.g., (assume)

Conditioned on age, gender, ethnicity, socio-economic status (X)

whether someone smokes (T) and whether they have a lung disease (Y) are independent

- Unfortunately, untestable in most observational studies.

Methods to identify causal effects under unconfoundedness

1. “Matching”

Covariates (X)

T	Y1	Y0	Age (X_1)	Race (X_2)	Gender (X_3)	State (X_4)	Edu (X_5)
1	130	?	20s	W	M	NC	College
0	?	125	20s	W	M	NC	College
1	127	?	30s	B	F	MA	PhD
0	?	130	30s	L	F	CA	PhD

Average Treatment Effect (ATE) = $\mathbf{E}[Y(1) - Y(0)]$

Valid group

$= E_X[E[Y1 \mid T = 1, X] - E[Y0 \mid T = 0, X]]$

Each valid matched group must have

- at least one treated unit
- at least one control unit

2. Propensity Score Matching

- Propensity score (Rosenbaum and Rubin, 1983): The conditional probability of receiving a treatment T given pre-treatment covariates X:

$$e(X) = \Pr(T = 1 | X)$$

- The propensity score balances the observed covariates, but does not generally balance unobserved covariates
- In most observational studies, the propensity score e(X) is unknown and thus needs to be estimated

- Stage 1: Estimate the propensity score:
  - by a logistic regression or machine learning methods
- Stage 2: Given the estimated propensity score, estimate the causal effects through matching or Regression
- Limitation: May not be interpretable and depends on the model

Other methods

- Prognostic score matching
- Regression modeling potential outcome
- Doubly robust machine learning
- Causal BART

Not covered here

Comparing Rubin’s and Pearl’s Models

Neyman-Rubin vs. Pearl's Model

Disclaimer: only some excerpts, not exhaustive views and not the most recent ones..

Some authors (e.g., Greenland, Pearl, and Robins 1999; Dawid 2000) call the potential outcomes "counterfactuals," borrowing the term from philosophy (e.g., Lewis 1973). I much prefer Neyman's implied term "potential outcomes," because these values are not counterfactual until after treatments are assigned, and calling all potential outcomes "counterfactuals" certainly confuses quantities that can never be observed (e.g., your height at age 3 if you were born yesterday in the Arctic) and so are truly a priori counterfactual, with unobserved potential outcomes that are not a priori counterfactual (see Frangakis and Rubin 2002; Rubin 2004; and the discussion and reply for more on this point).

Despite other approaches advocated by people whom I greatly respect (e.g., Dawid 2000; Lauritzen 2004; Pearl 2000), the potential outcomes formulation of causal effects, whether in randomized experiments or in observational studies, has achieved widespread acceptance. The potential outcomes, together with

(Rubin, JASA, 2005, p325 & p329)

"Formally, the two frameworks are logically equivalent; a theorem in one is a theorem in the other, and every assumption in one can be translated into an equivalent assumption in the other. Therefore, the two frameworks can be used interchangeably and symbiotically, as it is done in the advanced literature in the health and social sciences….In summary, the PO framework offers a useful analytical tool (i.e.. an algebra of counterfactuals) when used in the context of a symbiotic SCM analysis. It may be harmful however when used as an exclusive and restrictive subculture that discourages the use of process-based tools and insights."

(Pearl 2012)

Read http://causality.cs.ucla.edu/blog/index.php/2012/12/03/judea-pearl-on-potential-outcomes/ for a detailed version

Neyman-Rubin vs. Pearl's Model

- Potential Outcome (Neyman-Rubin) = Do Operator/counterfactual (Pearl)
- Treatment (Neyman-Rubin) ≊ intervention (Pearl)
- Structural causal graph on variables assumed by Pearl
  - Causal inference is on (variable-value) pairs
- No causal structure assumed in Neyman-Rubin's model
  - Infers causal relationships by experiments or from evidence
- Pearl's method gives a systematic way to find the covariates to adjust for --- but you may not have a reliable causal DAG available.. In practice the directions might not be known
- Mathematically the two frameworks are connected, but each has different established goals, tools and applicable areas (Richardson and Robins, 2013)

Outline

Sudeepa

- Introduction
- Pearl's Graphical Causal Model
- Rubin's Potential Outcome Framework
- **Briefly: some recent research on causal inference techniques (scalability & relational)**

Babak

- Causal Fairness
- Causal Explainability

Scalable Matching Algorithms

Recall "Matching"

Covariates (X)

T	Y1	Y0	Age (X_1)	Race (X_2)	Gender (X_3)	State (X_4)	Edu (X_5)
1	130	?	20s	W	M	NC	College
0	?	125	20s	W	M	NC	College
1	127	?	30s	B	F	MA	PhD
0	?	130	30s	L	F	CA	PhD

Average Treatment Effect (ATE) = $\mathbf{E}[Y(1) - Y(0)]$

Valid group

$= E_X[E[Y1 \mid T = 1, X] - E[Y0 \mid T = 0, X]]$

Each valid matched group must have

- at least one treated unit
- at least one control unit

Exact Matching = Interpretability

Rosenbaum-Rubin'83

- "Match" on propensity score $e(X) = \Pr(T = 1 \mid X)$: need a model, hard to interpret

Go model free - Exact matching to the rescue!

- Highlights overlap between treatment and control populations
- Helps us to find uncertainty and determine what type of additional data must be collected
- Interpret causal estimates within matched populations as "conditional average treatment effects (CATE)" in addition to ATE

“Exact Matching” in Observational Data

Ideally…

(1) Find “units” (e.g. patients) with same/similar “confounding covariates”
- e.g., of same age, gender, height, ethnicity, …

(2) Make sure all groups have both treated and control units

(3) Estimate the causal effect within each group and take average

Exact Matching: Good but challenging

Rosenbaum-Rubin'83

"As a method of multivariate adjustment, subclassification has the advantage that it involves direct comparisons of ostensibly comparable groups of units within each subclass and therefore can be both understandable and persuasive to an audience with limited statistical training... "

- Subclassification = exact matching
- Direct comparisons = individualized effects
- Persuasive = intuitive, uncomplicated, reproducible

"A major problem with subclassification is that as the number of confounding variables increases, the number of sublcasses grows dramatically, so that even with only two categories per variable, yielding 2^P classes for P variables, most subclasses will not contain both treated and control units."

- Confounders = variables of potential interest
- Number of subclasses = types of individualized effects
- Empty subclasses = impossible to draw causal conclusions

How do we get more matched units?
Solution: "Almost" Matching Exactly!

FLAME: Fast Large-Scale Almost Matching Exactly

[Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky, JMLR'21]

Important Covariates **Unimportant Covariates**

covariates: age, gender, heart conditions, blood pressure, toenail length, eyeball width, etc.

treated patient Marietta	[50	F	1	0	1	1	68	1.5cm	2cm	1	0	3	0	.....]
control patient Lee Ann	[50	F	1	0	1	1	68	14cm	1cm	4	1	5	6	.....]

Use ML

- Match treatment and control units using as many important covariates as possible
- Handle large datasets

Using techniques from data management

Optimization Problem for FLAME

[Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky, JMLR'21]

Variable Selector Indicator: $\boldsymbol{\theta} \in \{0, 1\}^p$

Matched Group for i on variables :: $\boldsymbol{\theta}$

$$\mathcal{MG}_i(\boldsymbol{\theta}, \mathcal{S}) = \{i' \in \mathcal{S} : \mathbf{x}_{i'} \circ \boldsymbol{\theta} = \mathbf{x}_i \circ \boldsymbol{\theta}\}$$

Prediction Error on training set

$$\hat{\mathrm{PE}}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}, \mathcal{S}) = \min_{f^{(1)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \frac{1}{|\mathcal{S}_1|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_1} (f^{(1)}(\mathbf{x}_i \circ \boldsymbol{\theta}) - y_i)^2 + \min_{f^{(0)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \frac{1}{|\mathcal{S}_0|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_0} (f^{(0)}(\mathbf{x}_i \circ \boldsymbol{\theta}) - y_i)^2.$$

Objective:

$$\boldsymbol{\theta}^*_{i,\mathcal{S}} \in \arg\min_{\boldsymbol{\theta}} \hat{\mathrm{PE}}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}, \mathcal{S}) \text{ s.t. } \exists \ell \in \mathcal{MG}_i(\boldsymbol{\theta}, \mathcal{S}) \text{ s.t. } t_\ell = 0$$

For every treatment unit, find
The best possible match with at
least one control unit

"Best" = Low predictive error on
a holdout set

Find as many matched group as
possible

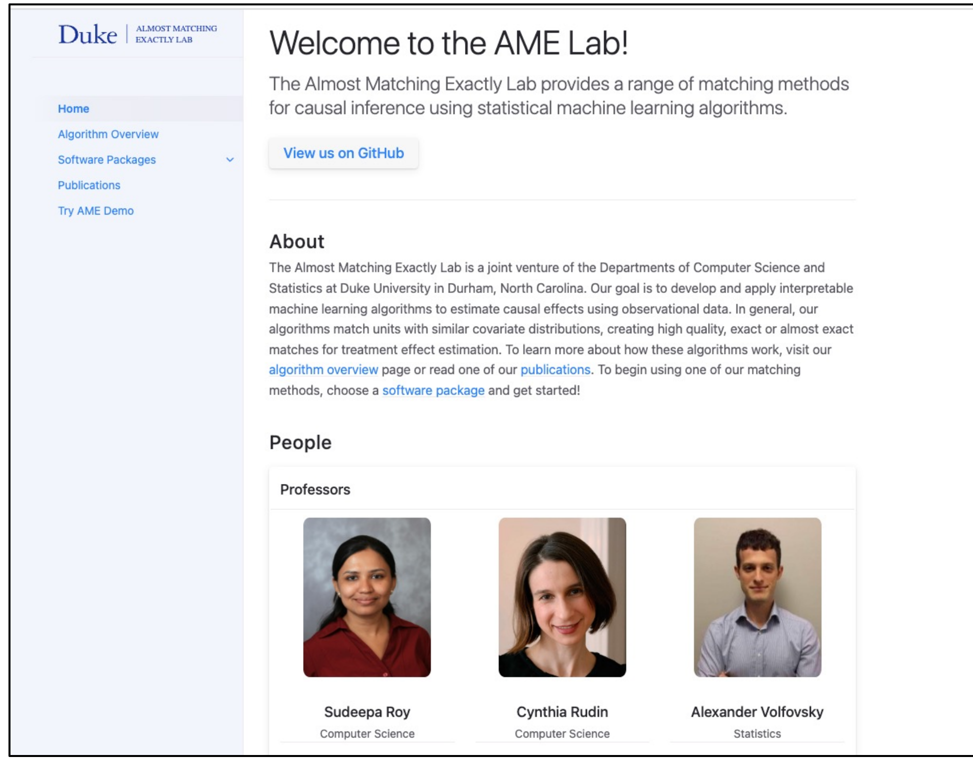
Drop least useful covariate and
REPEAT (greedy backward selection)

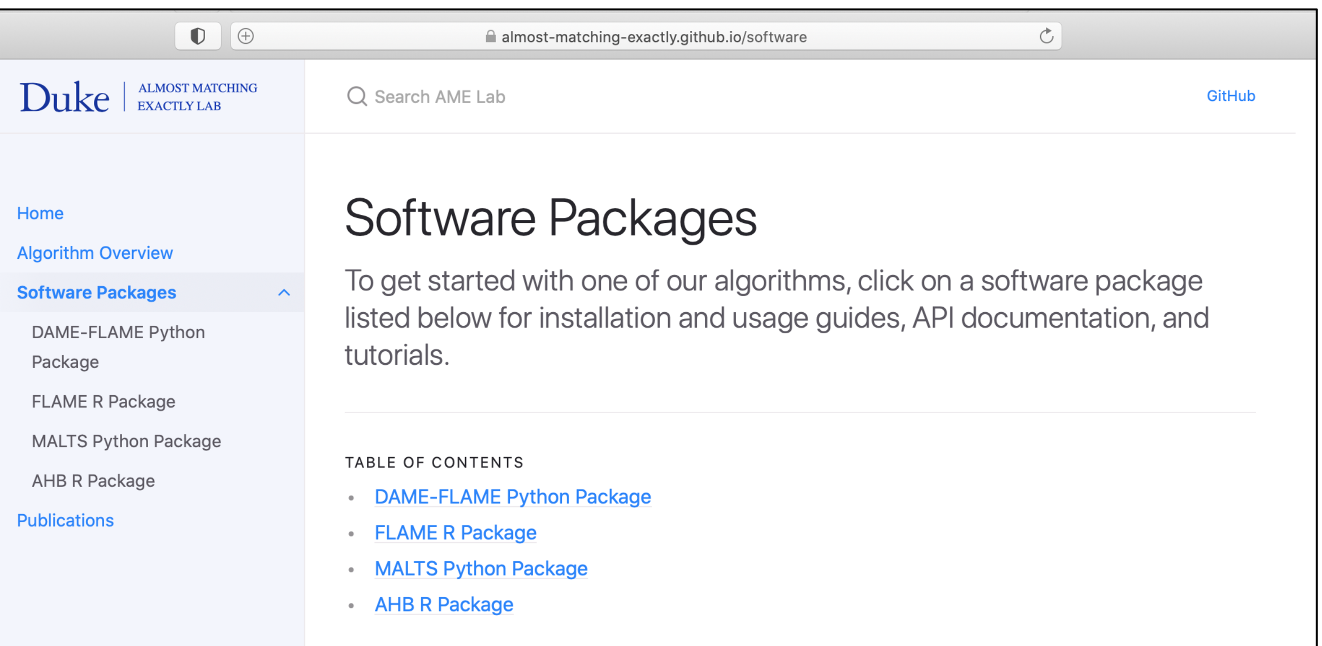
DAME: Dynamic Almost Matching Exactly
Go over all possible subsets: more accurate, more time to run
Can have FLAME-DAME hybrid too: first FLAME, then DAME
[Dieng-Liu-Roy-Rudin-Volfovsky, AISTATS'19]

Code, tutorial, examples on AME lab webpage

Duke UNIVERSITY

https://almost-matching-exactly.github.io

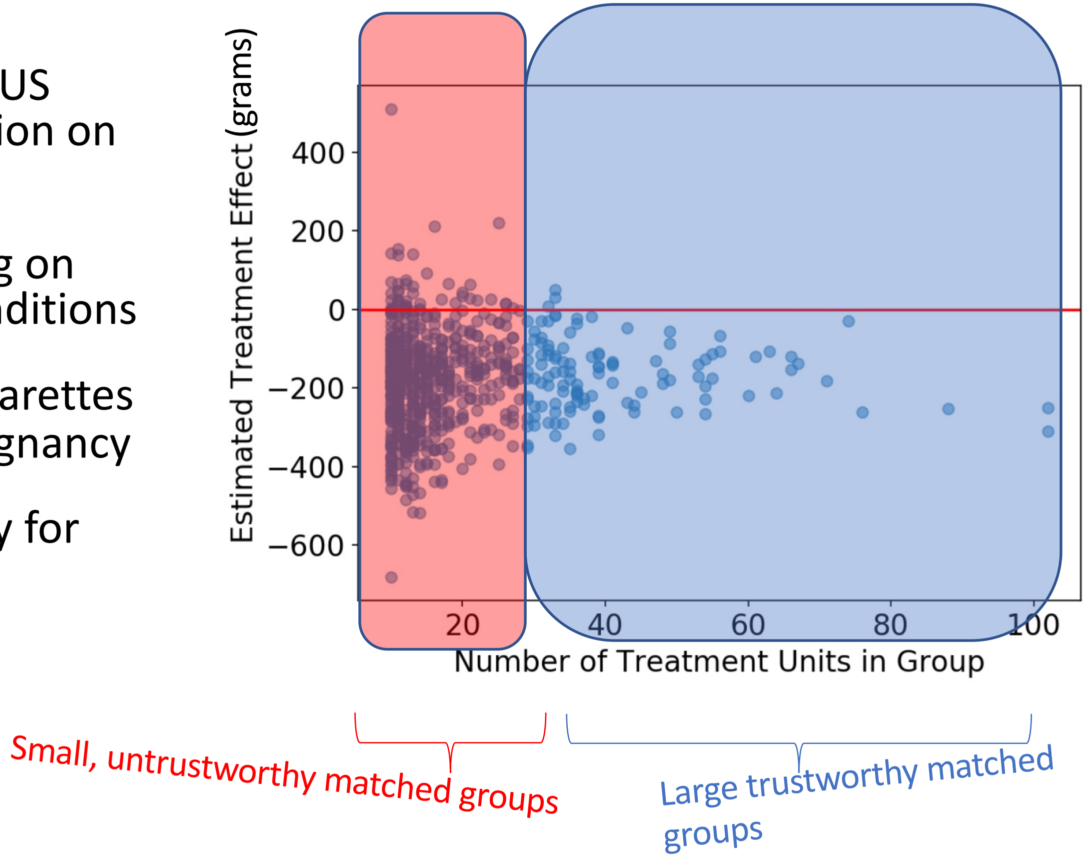




Advantage of Interpretable Matching Algorithms: Results on Natality Data

[Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky, JMLR'21]

- Natality Dataset, 2010: set of all US births, including health information on pregnant women and newborns
- Estimate causal effect of smoking on risk of child abnormal health conditions
- Treatment: Smoke at least 10 cigarettes a day for the duration of the pregnancy
- Control: Smoke 0 cigarettes a day for the duration of the pregnancy
- Outcome: birth weight
- ~2.1M units, 75K treated



Role of **Data Management** for Observational Causal Inference

			Covariates (X)				
T	**Y1**	**Y0**	**Age (X_1)**	**Race (X_2)**	**Gender (X_3)**	**State (X_4)**	**Edu (X_5)**
1	130	?	20s	W	M	NC	College
0	?	125	20s	W	M	NC	College
1	127	?	30s	B	F	MA	PhD
0	?	130	30s	L	F	CA	PhD

```
SELECT Age, Race, Gender, State, Education,
    ((SUM(T*Y)/SUM(T)) – (SUM(1-T)*Y)/(COUNT(*)-SUM(T))) AS ATE
FROM  Population
GROUP BY Age, Race, Gender, State, Education
HAVING SUM(T)>= 1 AND SUM(T) <= COUNT(*) - 1
```

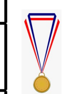
+ Robust
+ A few lines of code (declarative)
+ Scalable
(only the db-based method succeeded on > 1 million tuples)

[Wang, Morucci, Awan, Liu, Roy, Rudin, Volfovsky, JMLR'21]

Almost Matching Exactly (AME): Scalability with Database Queries

[Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky, JMLR'21]

Method	Time (hours)
FLAME-bit	Crashed
FLAME-db	1.33
Causal Forest	Crashed
1-PSNNM & GenMatch	> 10
Mahalanobis	> 10
Cardinality Match	> 10



US Census 1990 dataset

Units = 1.2 million

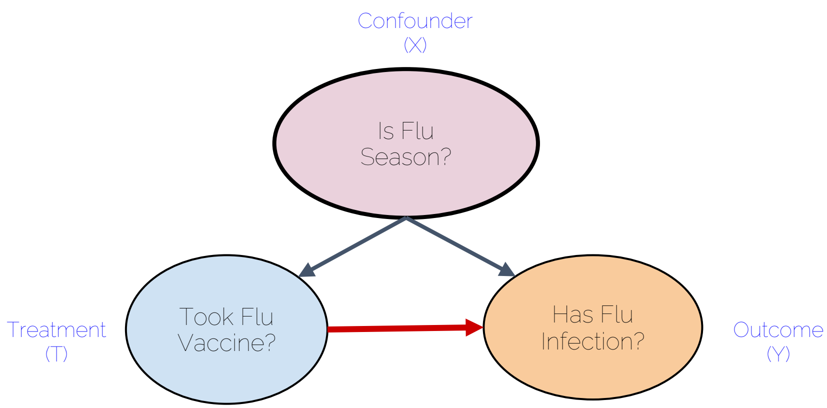
Covariates = 59

All these on a single “table”
with “Independent Units”

Causality for Network/Relational Data

.. When SUTVA or “no interference” is violated

Existing Causality Frameworks

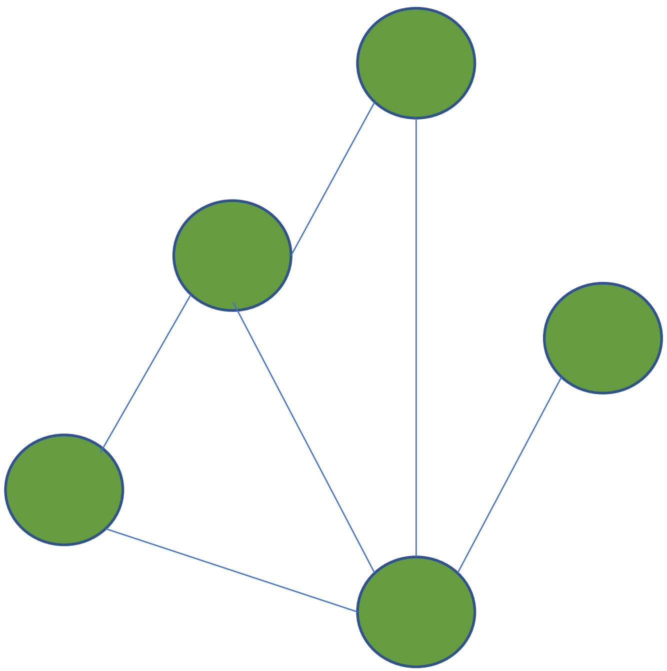


Key Assumptions –

- **Conditional Ignorability :**
  - $T \perp Y(1), Y(0) \mid X$
- **"SUTVA" :**
  - $Y\_a \perp T\_b$
  - Only one kind of treatment
- **Unit Homogeneity :**
  - T and Y on the same "entities"

Classical Causal frameworks require homogeneous units in a Single Flat Table.

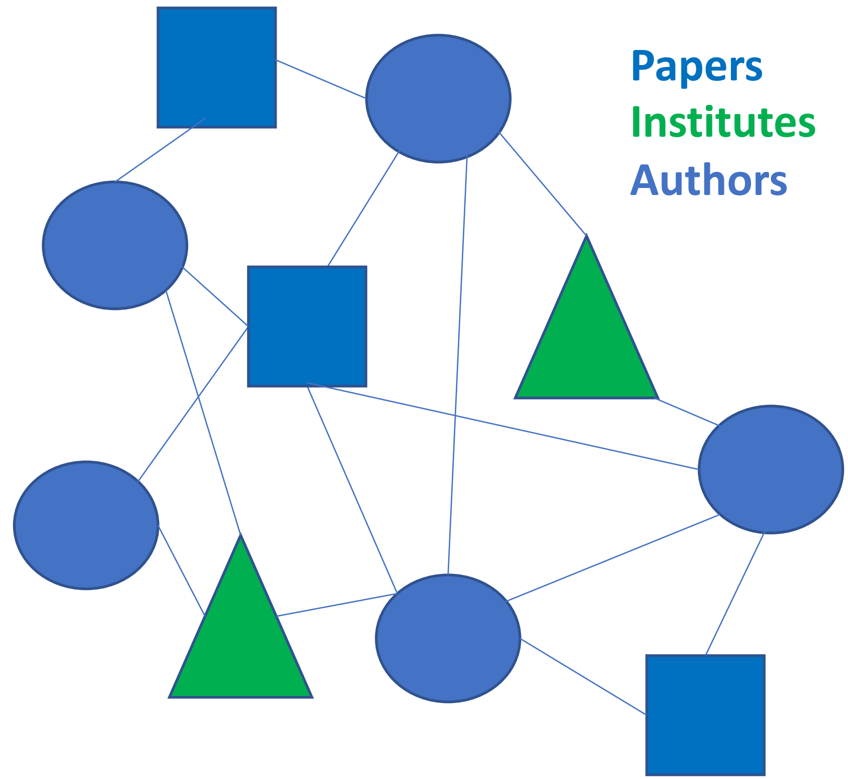
Units with Interference



Student sharing rooms in college dorms

"homogenous units"

Network data

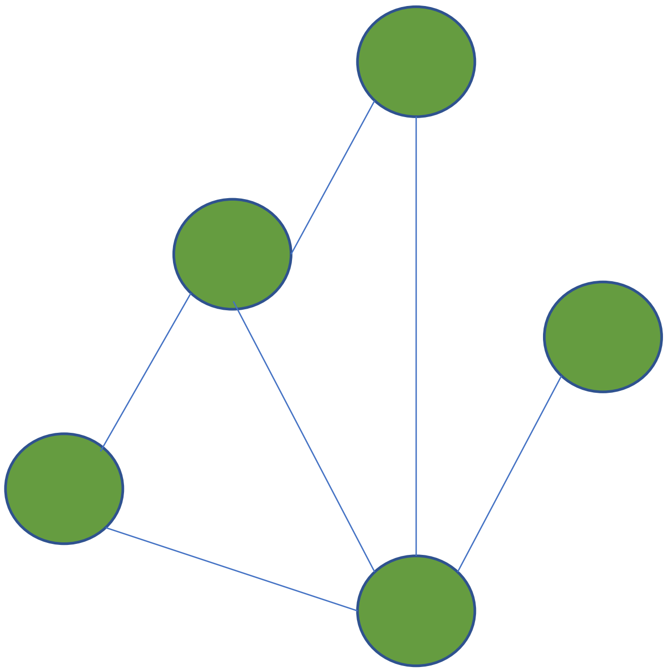


"heterogenous units"

Relational data

Homogenous units on a network

[Sherman-Shpitser, UAI'19]
[Bhattacharya-Malinsky-Shpitser, UAI'19]
[Morucci-Awan-Orlandi-Roy-Rudin-Volfovsky UAI '19]



Student sharing rooms in college dorms

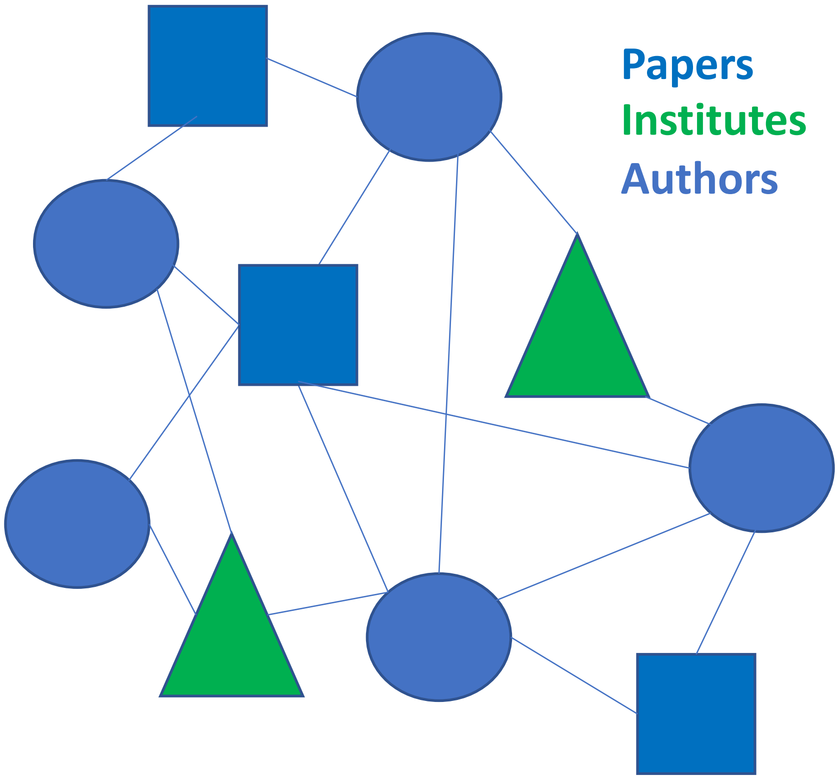
"homogenous units"

Basic assumptions like SUTVA do not hold

For two neighbors 1 and 2:
Interference T1 affects Y2
Contagion Y1 affects Y2
Entanglement T1 = T2

Ideas: can match on neighborhood structures
Or covariates of neighbors

Heterogenous "relational" data



"heterogenous units"

From two tables

Multiple tables:
Papers(pid, venue, year, title, …) T
Institute(iid, city, country, rank)
Authors(aid, name, position)
Affiliation(aid, iid)
Wrote(aid, pid) Y
Review(pid, rid, is-single-blind, **score**)

Does institutional rank (prestige) causally affect
Scores received by papers in reviews?

- For single-blind reviews?
- For double-blind reviews?

[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

Causality for Large Complex Data

Authors

person	prestige	qualification (h-index)
Bob	1	50
Carlos	0	20
Eva	1	2

Submissions

sub	score
s1	0.75
s2	0.4
s3	0.1

Authorship

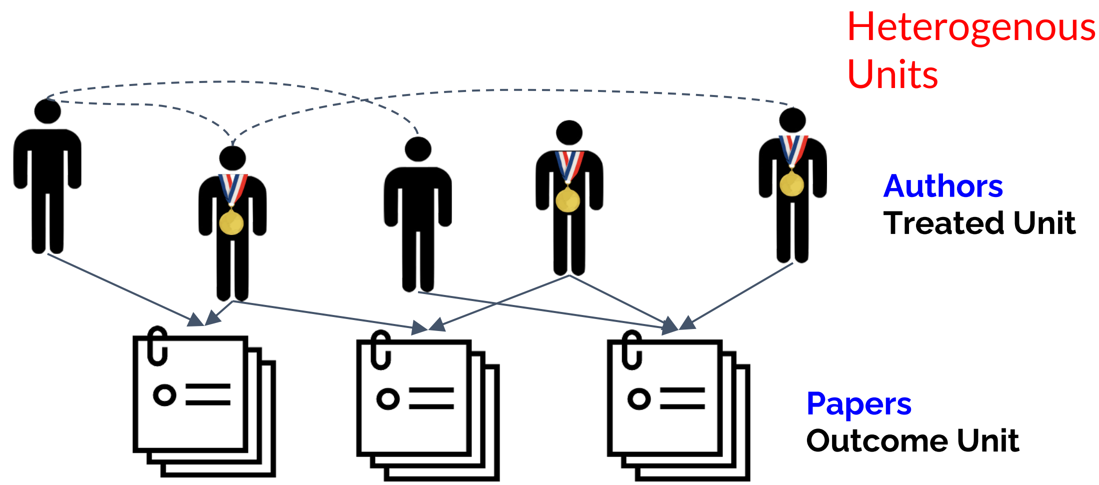
person	sub
Bob	s1
Eva	s1
Eva	s2
Eva	s3
Carlos	s3

Submitted

sub	conf
s1	ConfDB
s2	ConfAI
s3	ConfAI

Conferences

conf	blind
ConfDB	Single
ConfAI	Double



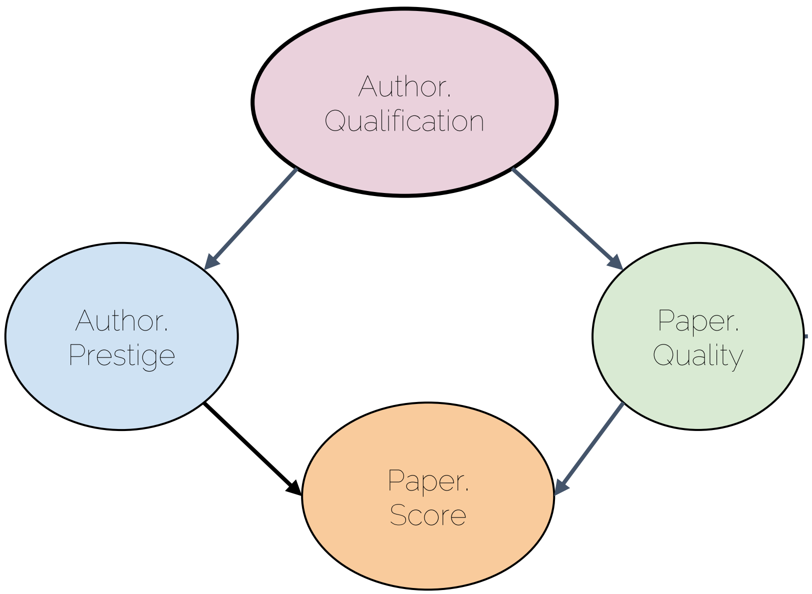
Relational DB has

- Multiple Tables with heterogeneous entities
- Many-Many Relationships and non-uniform treatment

[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

Background Knowledge

Potential Causal Links



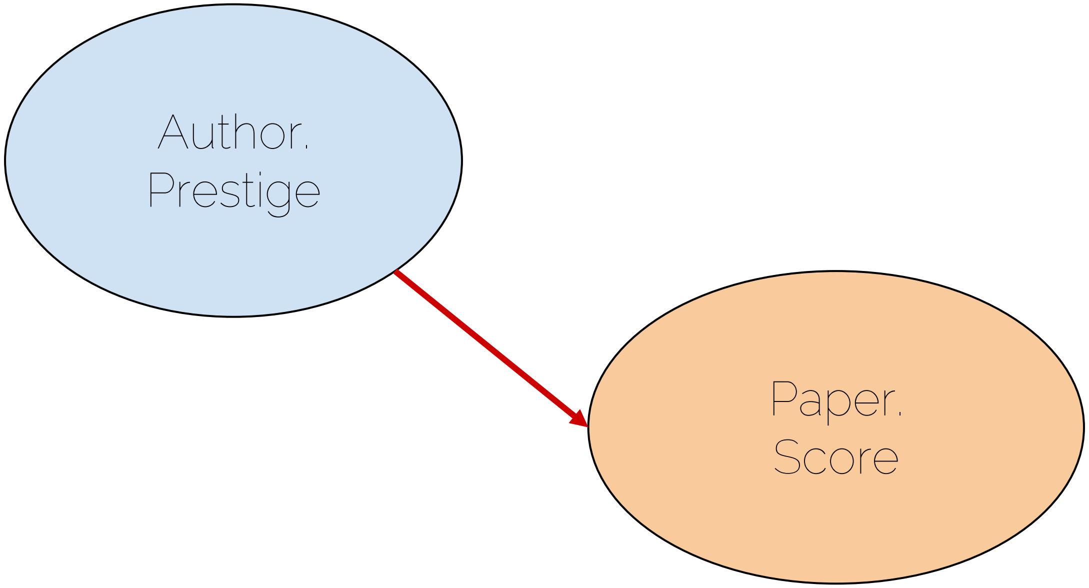

```
Prestige[A] <= Qualification[A]
            WHERE Author(A)

Quality[S] <= Qualification[A]
            WHERE Authorship(A, S)

Score[S] <= Prestige[A]
            WHERE Authorship(A, S)

Score[S] <= Quality[S]
            WHERE Paper(S)
```

Similar to Graphical Causal Model but Parameterized

Causal Query

[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

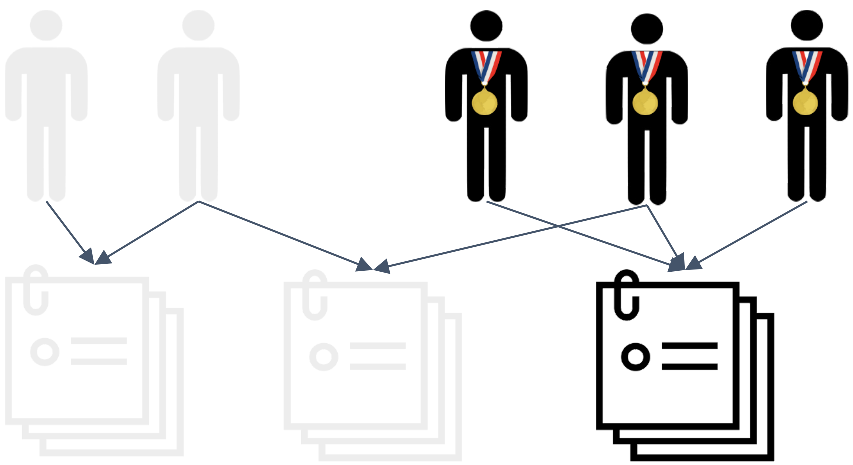
The Question of Interest



[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

Causal Query - Syntax

The Question of Interest



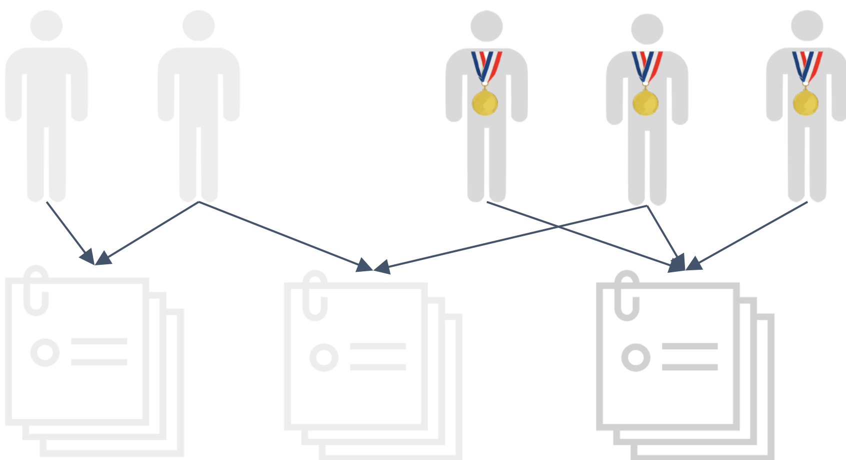
`Score [`S` ] ⇐ Prestige [`A`] ?`

`WHEN ALL AUTHOR TREATED`

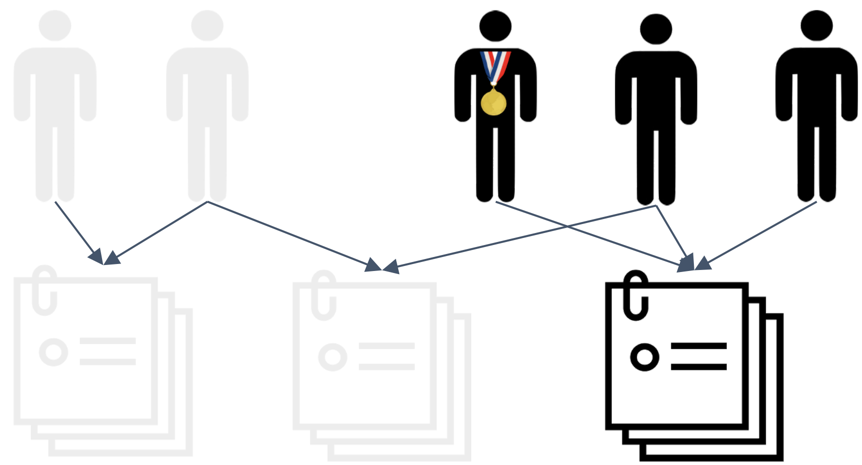
Causal Query - Syntax

[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

The Question of Interest



Score [S] ⇐ Prestige [A] ?
WHEN ALL AUTHOR TREATED



Score [S] ⇐ Prestige [A] ?
WHEN AT LEAST 1 AUTHOR TREATED

CARL Steps

[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

Translates multi-tables to the standard one-table form
Then uses off-the-shelf causal inference methods

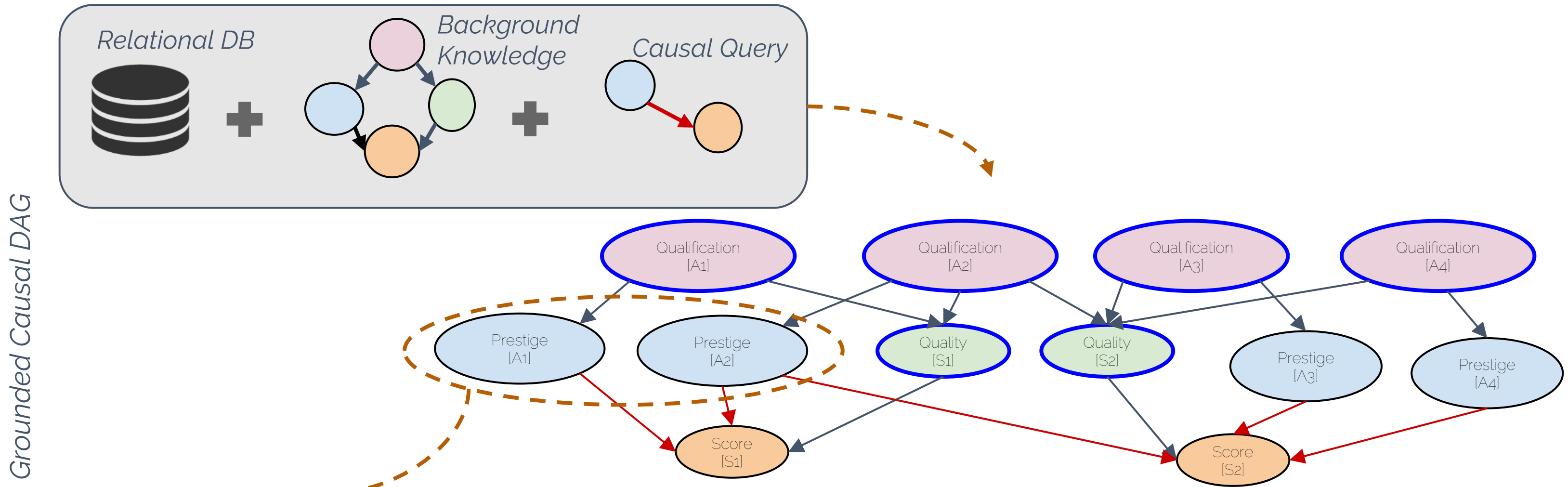
Skeleton Traversal

Grounding

Confounder Identification

Summary Functions

Causal Inference



Flat Table

Index	Summary-Prestige	Summary-Qualification	Score	Quality
S1	$g_{prestige}$(Prestige[A_1,A_2])	g_{qual}(Qualification[A_1,A_2])	Score[S1]	Quality[S1]
S2	$g_{prestige}$(Prestige[A_1,A_2,A_3])	g_{qual}(Qualification[A_1,A_2,A_3])	Score[S2]	Quality[S2]

\+ correctness proofs

Treatment Effect

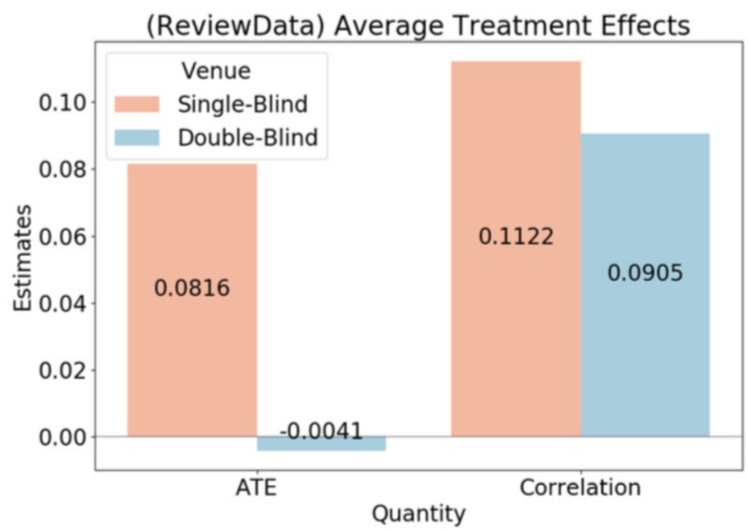
[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

Sample Results

OpenReview.net

Are reviewers influenced by authors' prestige?

≥ ⅓rd Authors Prestigious → Reviewer Score



(a)

Causation vs. Correlation

High correlation in both single and double blind
High causation only in single blind

Other interesting results on hospital And insurance data In the paper

Research Directions

Research Directions

- Explore the synergy in causal inference in Statistics and AI, and database research
- Improve scalability, e.g., matching on large and high-dimensional data
- Understand variance and confidence in matching
- Explore further causal analysis on complex relational and network data
- Work with domain experts on real applications

- After 30 mins, Causal Fairness & Explainability!

Outline

Introduction

Pearl's Graphical Causal Model

Rubin's Potential Outcome Framework

Briefly: some recent research on causal inference techniques (scalability & relational)

Causal Fairness

Causal Explainability

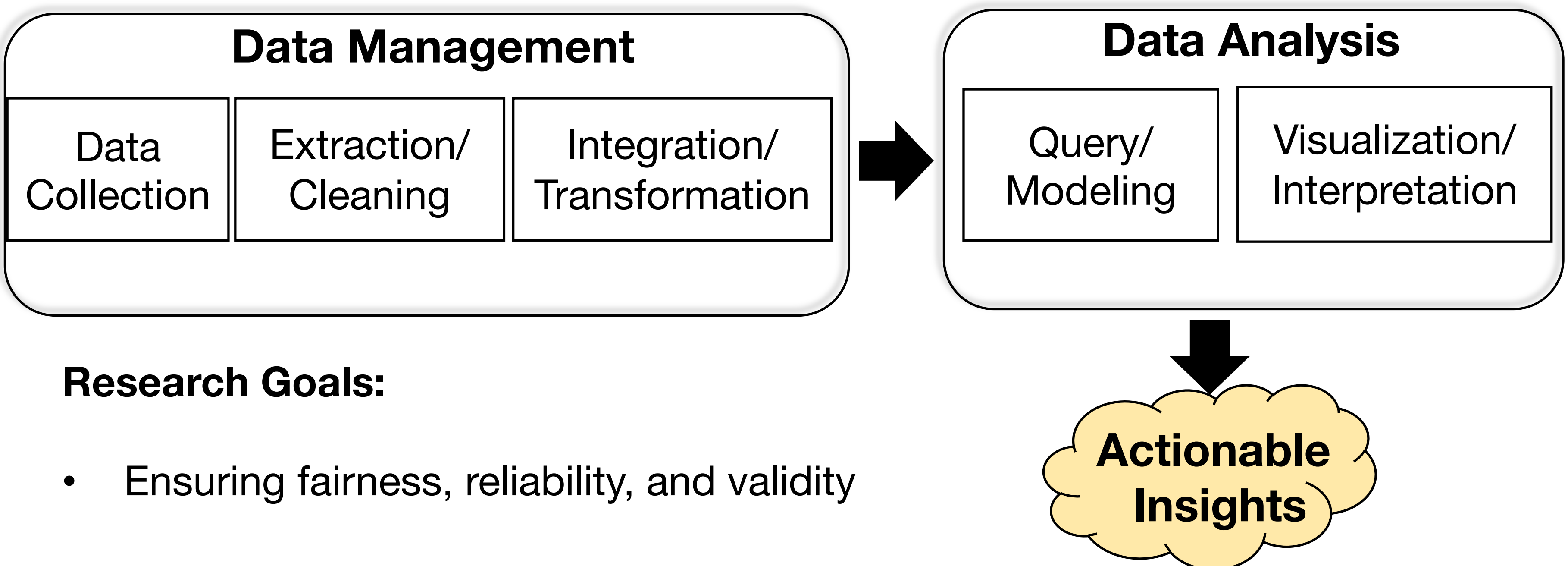
Sudeepa

Babak

Causal Inference and Fairness

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Responsible Data Management
and Analysis Lab
https://bsalimi.github.io/



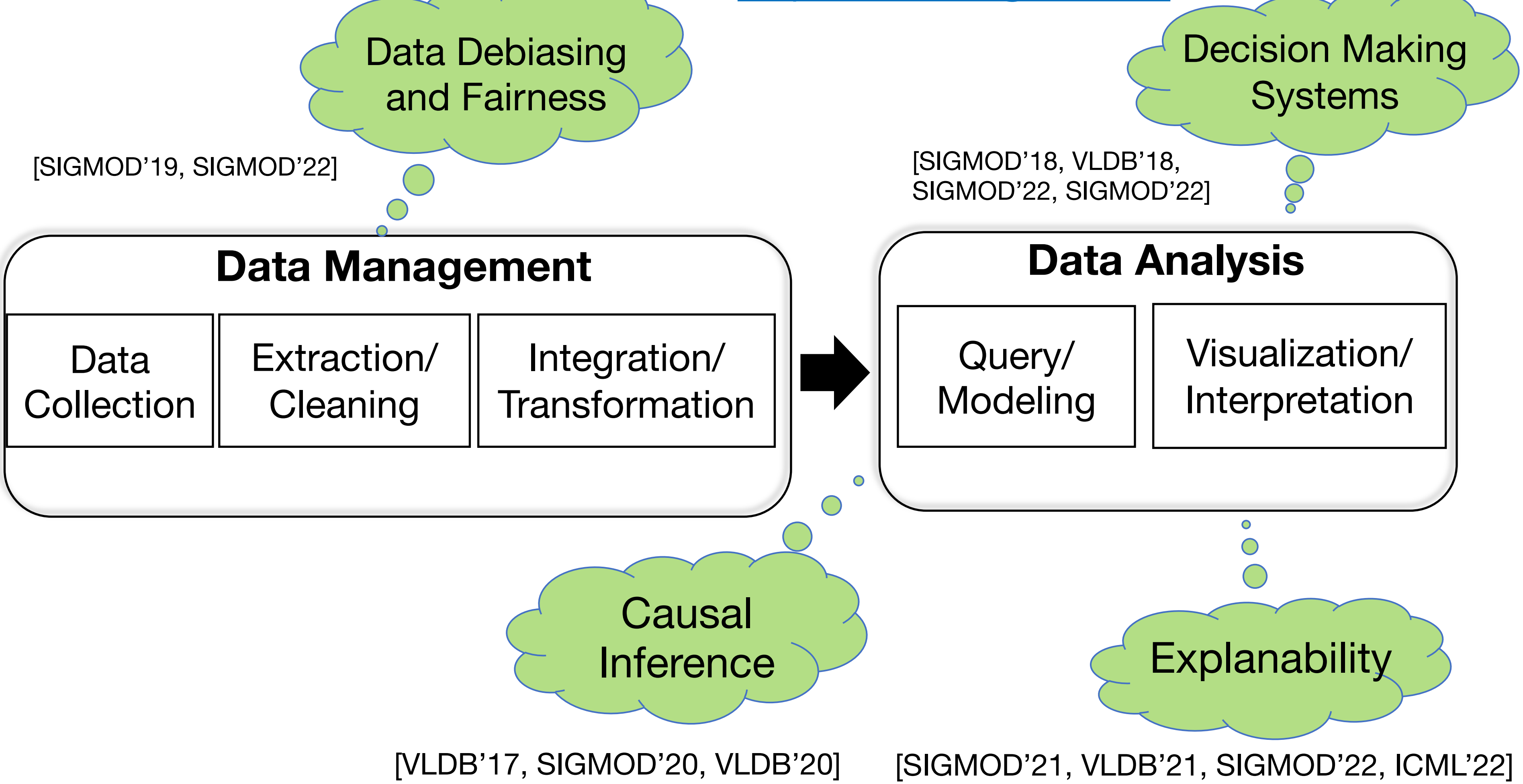
Research Goals:

- Ensuring fairness, reliability, and validity
- Operationalizing transparency and explainability
- Protecting data from bias in its lifecycle

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Responsible Data Management and Analysis Lab

https://bsalimi.github.io/



Algorithmic Fairness

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

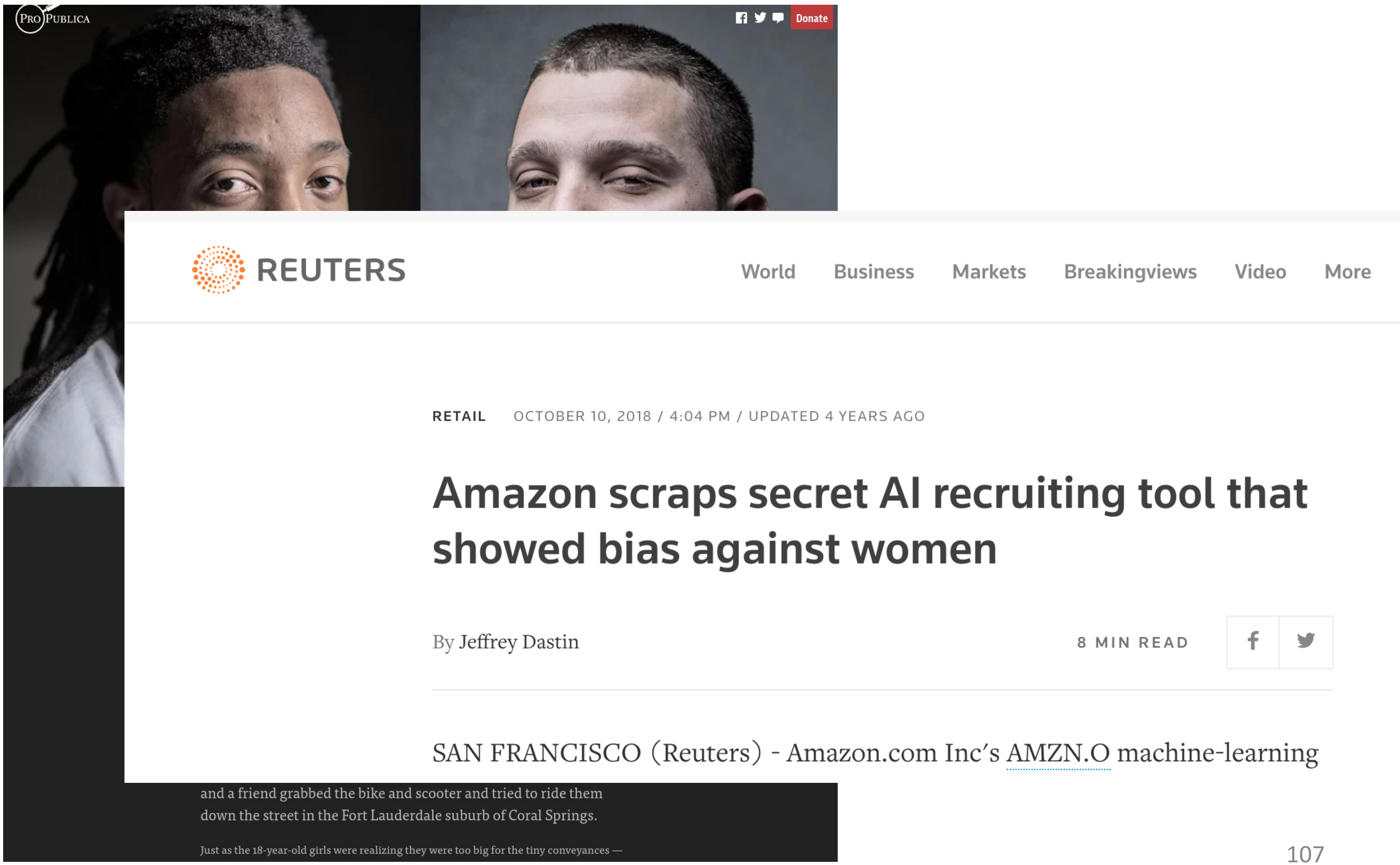
Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

source: https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70

Algorithmic Fairness



ProPublica

Donate

REUTERS

World Business Markets Breakingviews Video More

RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning

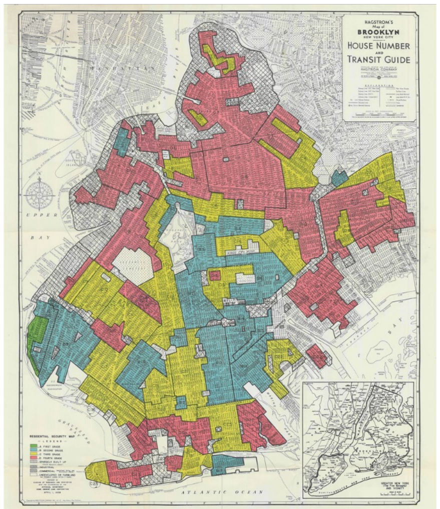
and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances —

What is algorithmic bias?

- Algorithm bias is the lack of fairness that emerges from the output of a computer system
- Fairness is typically defined in terms of invariance of algorithmic decisions to variables that considered as sensitive
- Examples of sensitive variables: gender, ethnicity, sexual orientation, disability, etc.

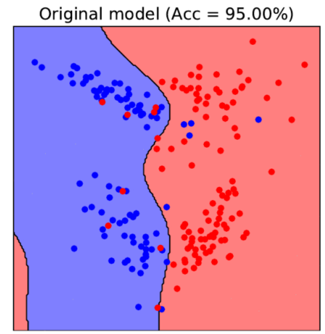
What are the sources of bias ?



src: NYTimes

Historical bias in training data

src: openai.com



src: https://labs.f-secure.com

Adversarial data attacks

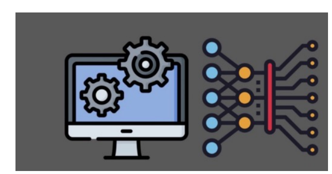
src: nagwa.com

Data integration

src: nagwa.com

src: datacubed.com

Selection bias



Model design choices

Hooker, Sara. "Moving beyond “algorithmic bias is a data problem”." *Patterns* 2.4 (2021): 100241.

Caveat: I am not lawyer, this is not a legal advice

Fair Classification

$X \rightarrow f \rightarrow O$

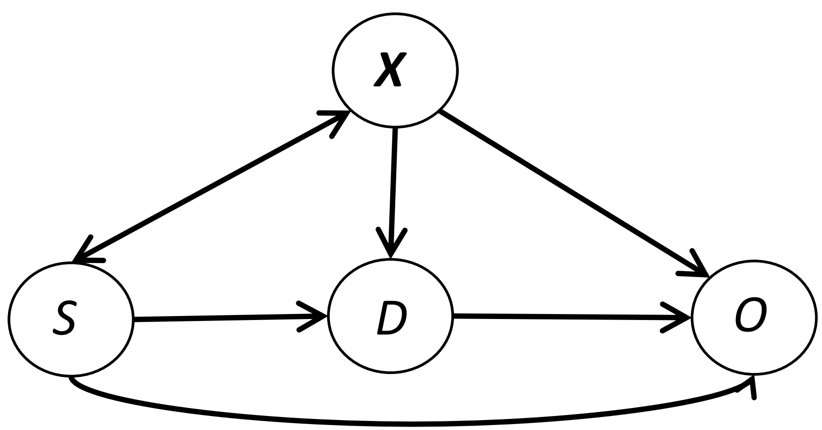
Inputs — Decision Procedure — Outcome

$O = 1$ Positive
$O = 0$ Negative

Sensitive attribute S :
$S = 1$ protected
$S = 0$ privileged

X: Features and qualifications: age, hobbies, test scores, grades, etc.

S: Gender



O: Admission Decisions

D: choice of dept.

Associational Fairness

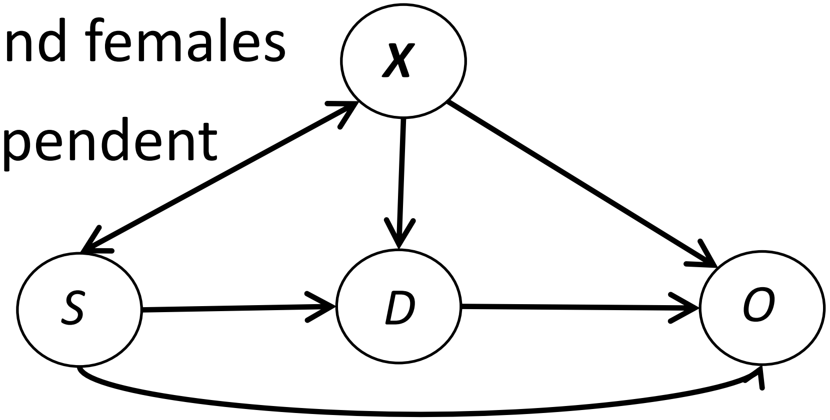
Demographic Parity
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$



Associational Fairness

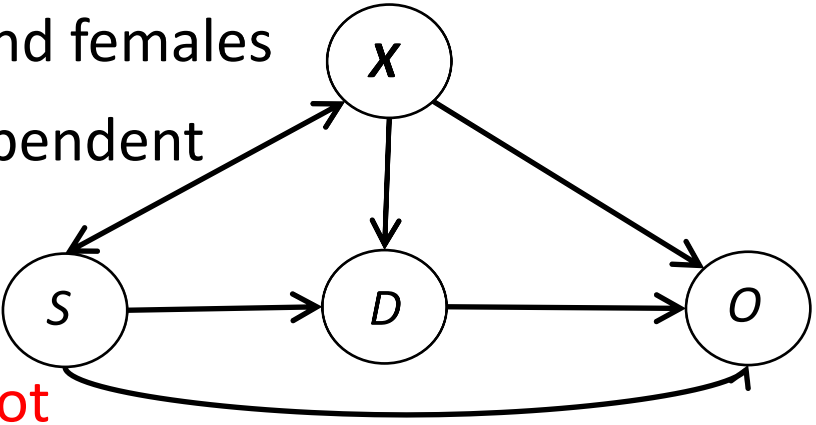
Demographic Parity
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$



Can it be ensured if decision are not based on S? (Fairness through Blindness)

Associational Fairness

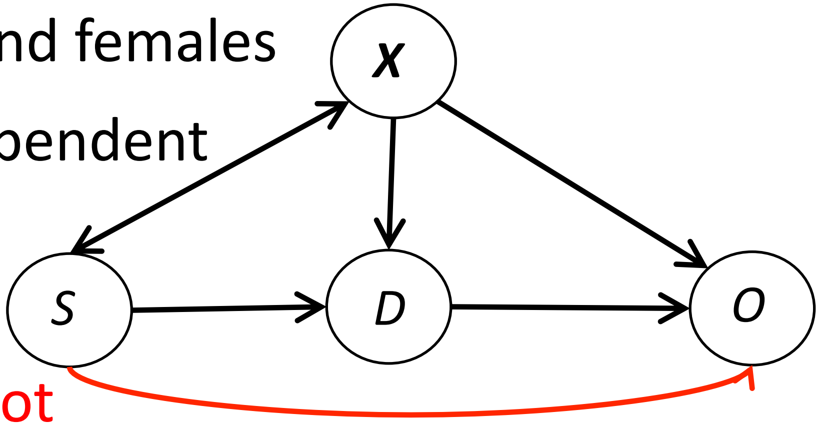
Demographic Parity
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$



Can it be ensured if decision are not based on S? (Fairness through Blindness)

Associational Fairness

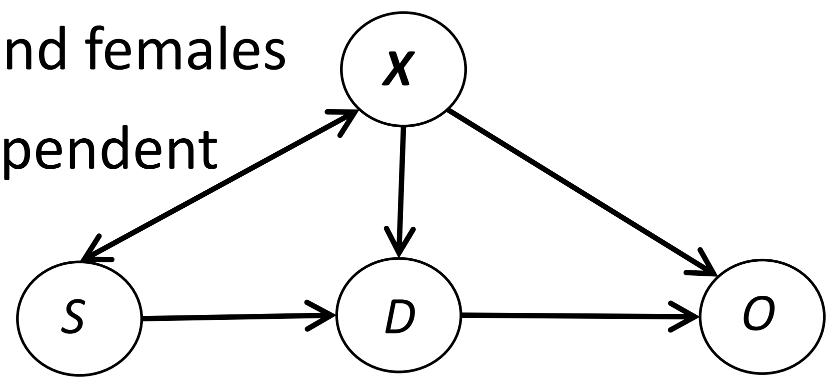
Demographic Parity
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$



Can it be ensured if decision are not based on S? (Fairness through Blindness)

Associational Fairness

Conditional Statistical Parity

Admissible attributes

For any **A=a**

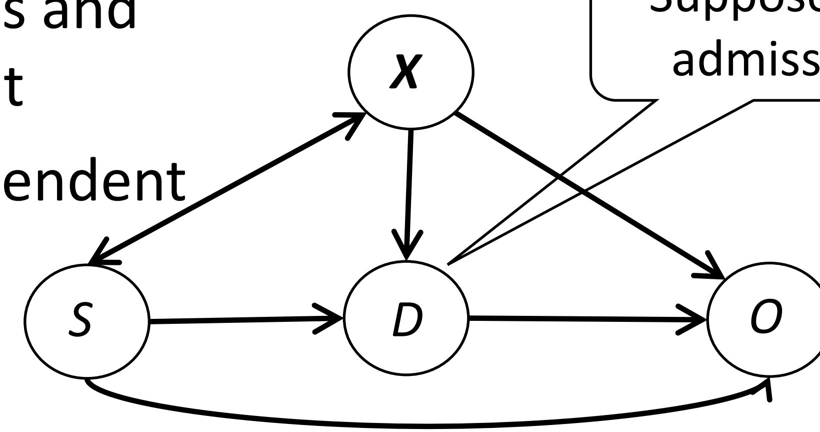
$$\mathbb{P}\{O=1|S=1, A=a\}=\mathbb{P}\{O=1|S=0, A=a\}$$

Same fraction of admitted males and females in each department

S and O should be marginally independent conditioned on D

$$O \perp\!\!\!\perp S \mid D$$

Suppose D is admissible



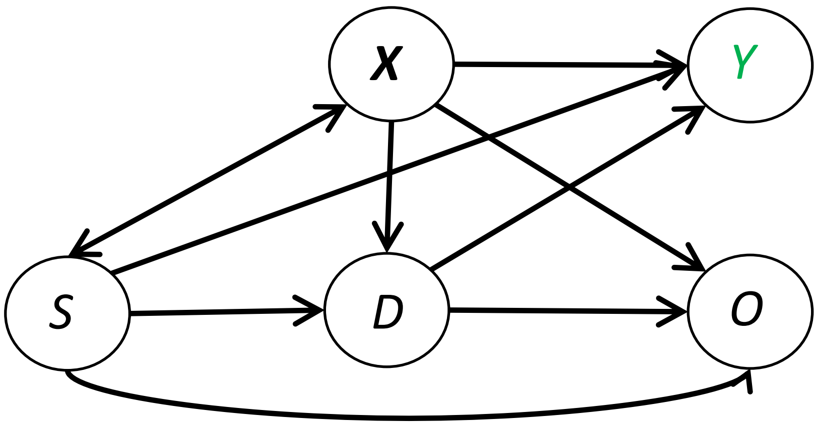
Associational Fairness

Equalized odds, conditional procedure accuracy equality and disparate mistreatment,

$$\mathbb{P}\{O=1|S=1,Y=1\}=\mathbb{P}\{O=1|S=0,Y=1\}$$
$$\mathbb{P}\{O=1|S=1,Y=0\}=\mathbb{P}\{O=1|S=0,Y=0\}$$

$$O \perp\!\!\!\perp S \mid Y$$

Among those applicant who (do not) graduate the rate of admitted students should be independent of applicants' gender.



Y be a binary variable that indicates degree attainment

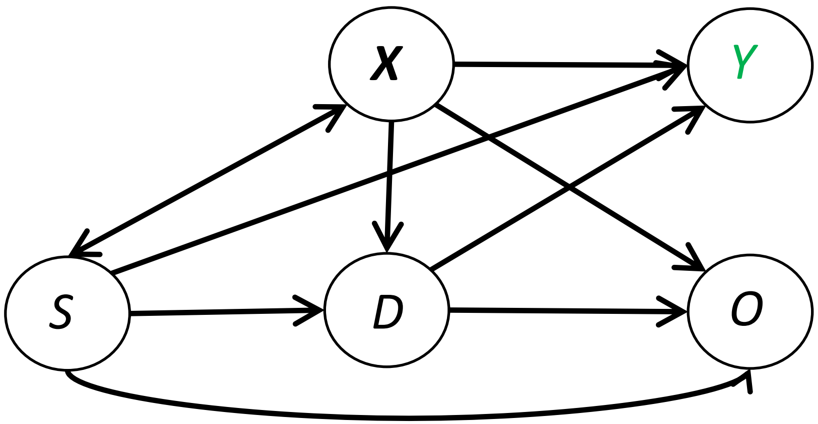
Associational Fairness

Predictive Parity, Outcome Test or Test-fairness or Calibration

$$\mathbb{P}\{Y=1|S=1,O=1\}=\mathbb{P}\{Y=1|S=0,O=1\}$$
$$\mathbb{P}\{Y=1|S=1,O=0\}=\mathbb{P}\{Y=1|S=0,O=0\}$$

$$Y \perp\!\!\!\perp S \mid O$$

Among those applicant that are admitted, the rate of those who attain colleague degree should be the same for males and females



Y be a binary variable that indicates degree attainment

An Associational Debate

FP rate for African-Americans (44.9%)
FP rate for white people (23.5%)
FN rate for whites (47.7%)
FN rate for African-Americans (28.0%)

The likelihood of recidivism among high-risk offenders is the same regardless of race

The COMPAS risk tool is *unfair it violates equalized odds*

ProPublica

The COMPAS risk tool is *fair*. It satisfies predictive parity.

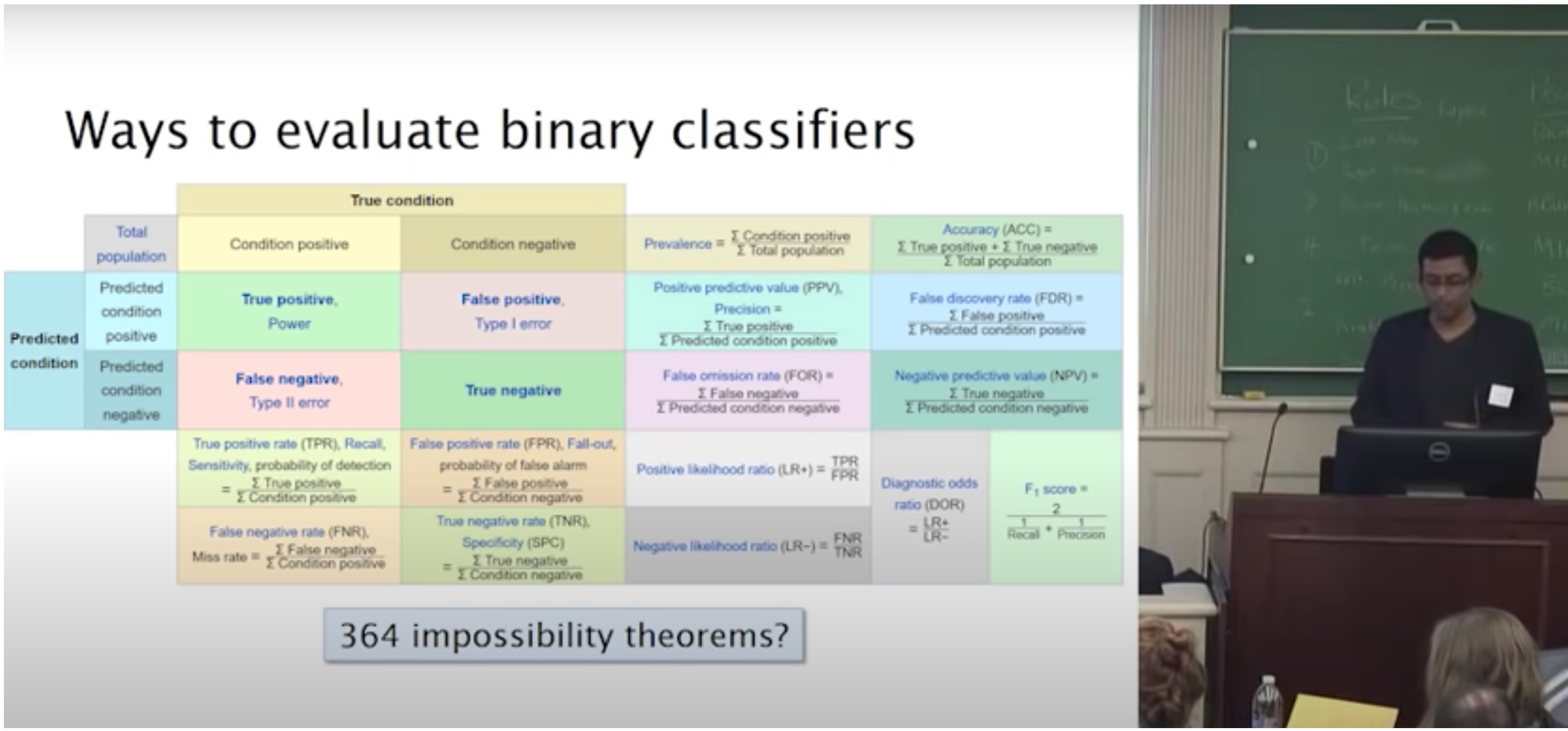
Northpointe

An Associational Debate

[Chouldechova 16], [Kleinberg, Mullainathan, Raghavan 16]:

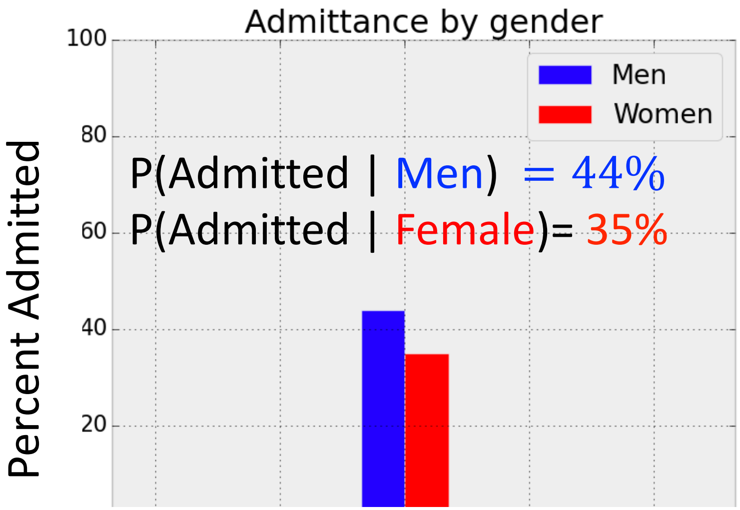
"If the base rates differ between two populations, then no non-trivial classifier can simultaneously equalized odds and predictive parity unless it is perfect".

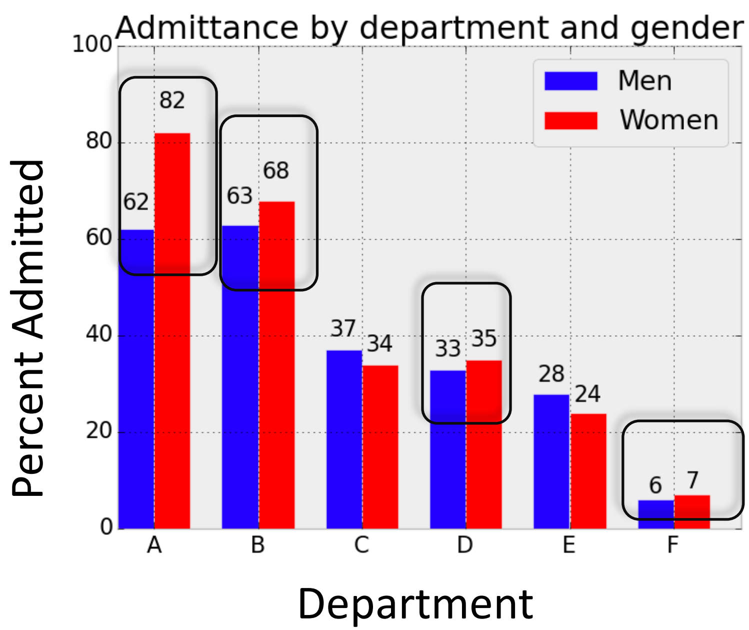
An Associational Debate



Tutorial: 21 fairness definitions and their politics
Arvind Narayanan
https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

An Associational Debate





[UC Berkeley 1973 grad school admissions]

Gender is correlated with Admitted

Disparity against females!

Disparity against males!

Discrimination is a causal concept

- Associational notions of fairness are inconsistent and could be misleading
- To prove discrimination, one must show sensitive attribute causes the decisions
- This conception can be traced back to legal systems and literature (The but-for test)
- The but for test broadly asks: **“But for the actions of the defendant (X), would the harm (Y) have occurred?”**

Discrimination in legal system

JUSTIA US Supreme Court

private developers leeway to state and explain the valid interest their policies serve, an analysis that is analogous to Title VII's business necessity standard. It would be paradoxical to construe the FHA to impose onerous costs on actors who encourage revitalizing dilapidated housing in the Nation's cities merely because some other priority might seem preferable. A disparate-impact claim relying on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity. A robust causality requirement is important in ensuring that defendants do not resort to the use of racial quotas. Courts must therefore examine with care whether a plaintiff has made out a prima facie showing of

source: https://supreme.justia.com/cases/federal/us/576/13-1371

or renovate housing units. And as Judge Jones observed below, if the [plaintiff] cannot show a causal connection between the Department's policy and a disparate impact—for instance, because federal law substantially limits the Department's discretion—that should result in dismissal of this case." *Id.* at 20-21.

source: https://www.jdsupra.com/legalnews/supreme-court-allows-disparate-impact-47404/

Causal Fairness

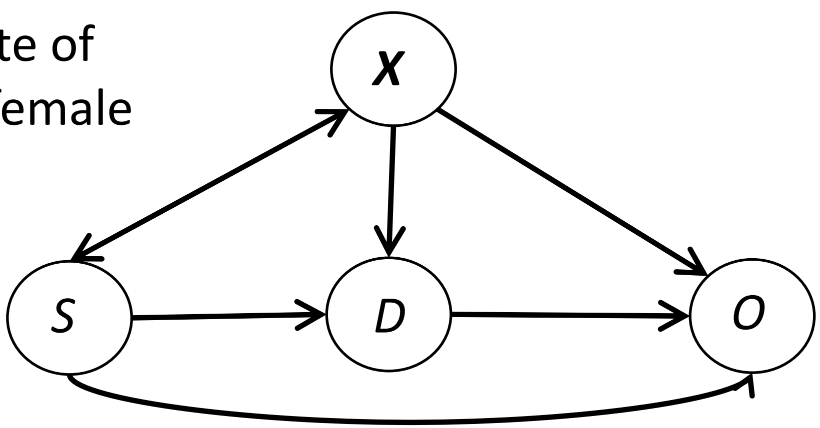
Total Causal Effect Fairness

$$\mathbb{P}(O=1 \mid Do(S=1)) = \mathbb{P}(O=1 \mid Do(S=0))$$

$$\mathbb{P}(O_{S\leftarrow 1}=1) = \mathbb{P}(O_{S\leftarrow 0}=1)$$

The rate of admitted students has all students were female should be equal to the rate of admitted student had all students were female

Sufficient Condition:
No causal path from S to O



Causal Fairness

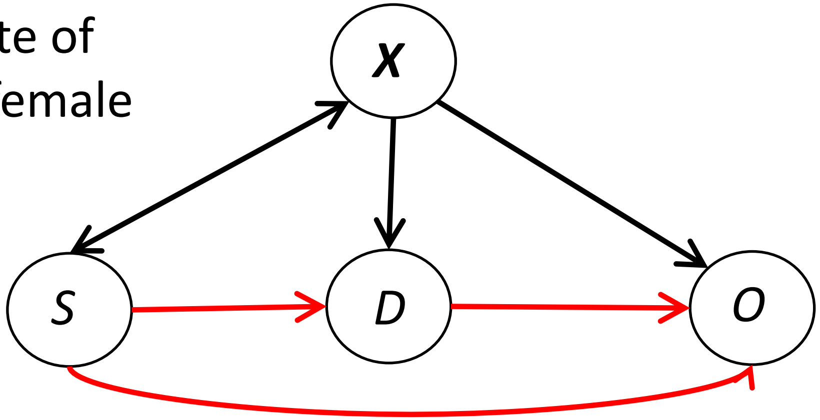
Total Causal Effect Fairness

$$\mathbb{P}(O=1 | Do(S=1)) = \mathbb{P}(O=1 | Do(S=0))$$

$$\mathbb{P}(O_{S\leftarrow 1} = 1) = \mathbb{P}(O_{S\leftarrow 0} = 1)$$

The rate of admitted students has all students were female should be equal to the rate of admitted student had all students were female

Sufficient Condition:
No causal path from S to O



Causal Fairness

Total Causal Effect Fairness

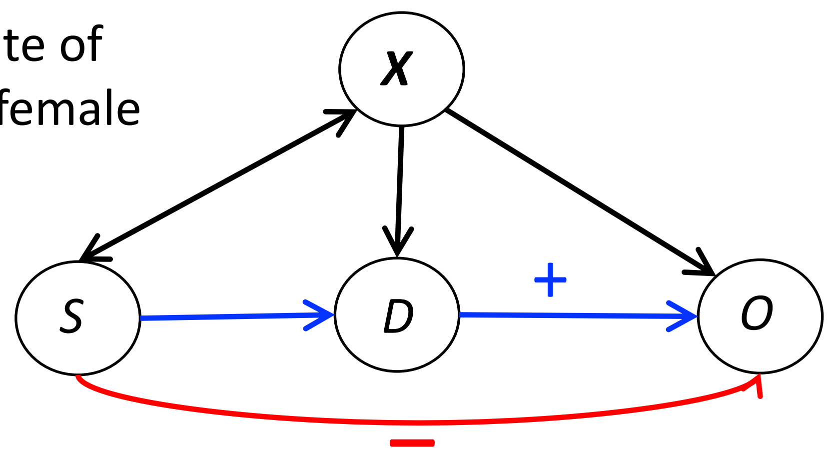
$$\mathbb{P}(O=1 \mid Do(S=1)) = \mathbb{P}(O=1 \mid Do(S=0))$$

$$\mathbb{P}(O_{S\leftarrow 1} = 1) = \mathbb{P}(O_{S\leftarrow 0} = 1)$$

The rate of admitted students has all students were female should be equal to the rate of admitted student had all students were female

Sufficient Condition:
No causal path from S to O

Dependence between S and O
= Spurious correlation + Causal effect



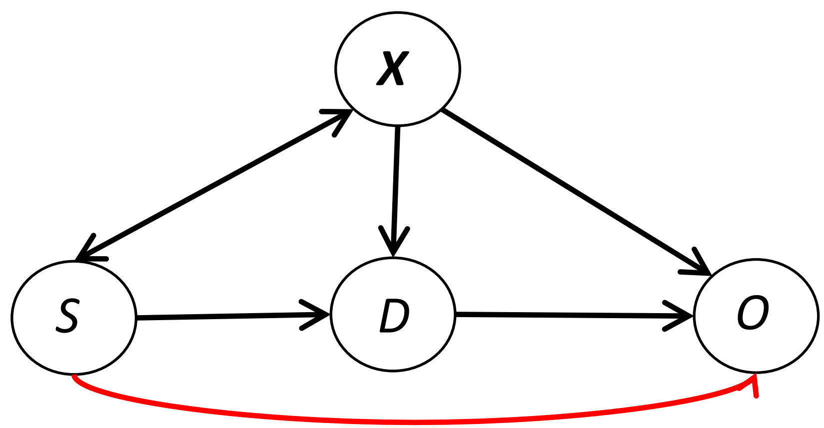
Causal Fairness

Direct Causal Effect Fairness

Total effect = Natural Direct Effect + Natural Indirect Effect

Forbids the natural direct causal effect of S on O

Dependence between S and O
= Spurious correlation + Direct causal effect + Indirect causal effect



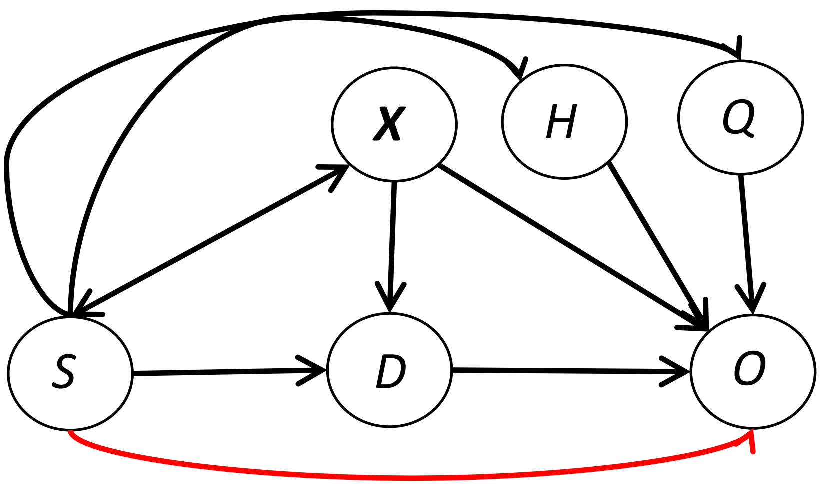
Causal Fairness

Direct Causal Effect Fairness

Total effect = Natural Direct Effect + Natural Indirect Effect

Forbids the natural direct causal effect of S on O

Dependence between S and O
= Spurious correlation + Direct causal effect + Indirect causal effect

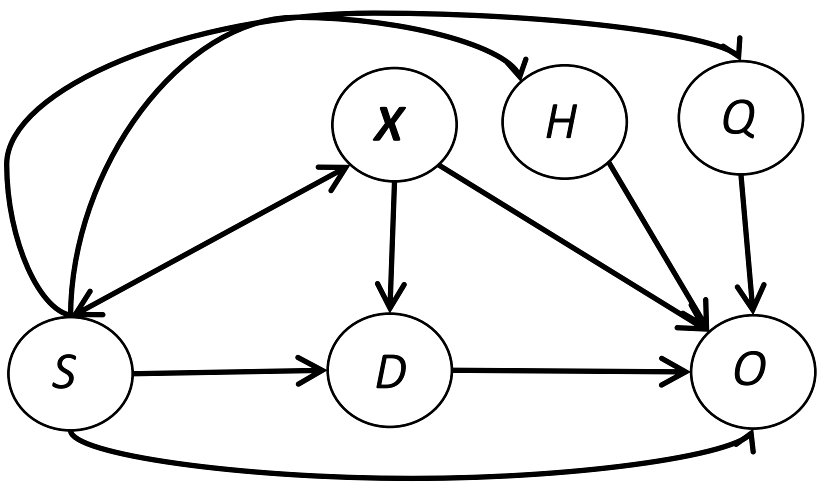


All indirect influences of S on O are allowed!

Causal Fairness

Path-Specific Fairness

Partition causal paths from S to O: fair/ discriminatory



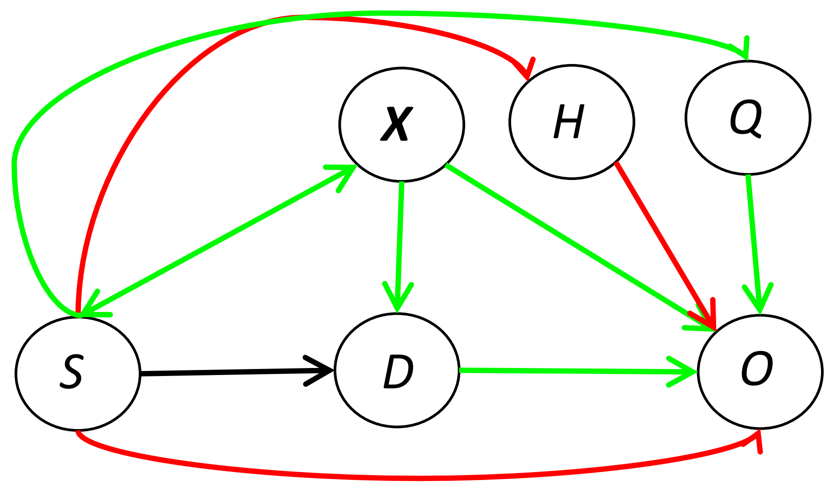
Causal Fairness

Path-Specific Fairness

Partition causal paths from S to O: fair/ discriminatory

S can influence O ONLY through fair causal paths

Red paths are discriminatory



Caveat: It is notoriously difficult to compute path specific effects

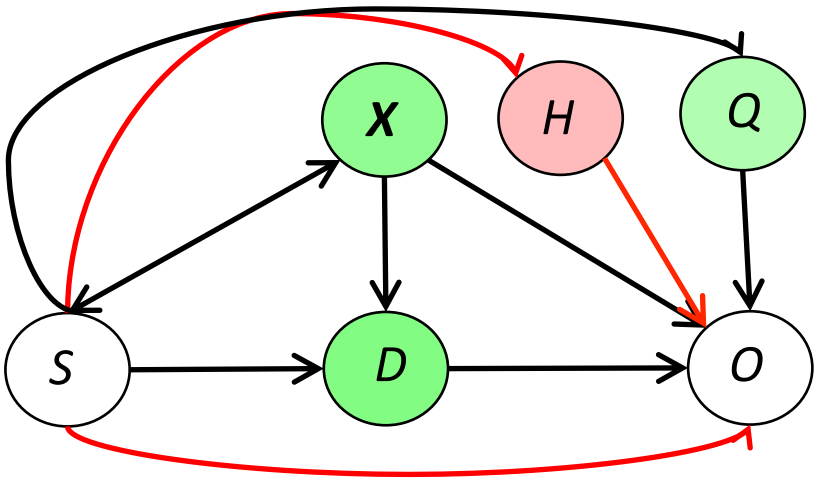
Causal Fairness

Interventional Fairness [SIGMOD'19]

Partition variables into: Admissible/ Inadmissible

For any $\mathbf{k} \in \mathrm{Dom}(\mathbf{K})$ and $\mathbf{K} \supseteq \{D, Q, X\}$

$$\mathbb{P}(O=1 \mid do(S=0), do(\mathbf{K}=\mathbf{k})) = \mathbb{P}(O=1 \mid do(S=1), do(\mathbf{K}=\mathbf{k}))$$



It is less expressive than path-specific fairness by easier to compute and enforce

Causal Fairness

Counterfactual Fairness

Total Causal Effect Fairness: $\mathbb{P}(O_{S\leftarrow 1} = 1) = \mathbb{P}(O_{S\leftarrow 0} = 1)$

$\mathbb{P}(O_{S\leftarrow 1} = 1) = \sum_u \mathbb{P}(O_{S\leftarrow 0}(u) = 1)\, \mathbb{P}(u)$

Exogenous variables

$$\mathbb{P}(O_{S\leftarrow 1} = 1 \mid X{=}x, S{=}1) = \mathbb{P}(O_{S\leftarrow 0} = 1 \mid X{=}x, S{=}1)$$

$$\mathbb{P}(O_{S\leftarrow 1} = 1 \mid X{=}x, S{=}0) = \mathbb{P}(O_{S\leftarrow 0} = 1 \mid X{=}x, S{=}0)$$

Can not be captured using the do-operator

$\mathbb{P}(O = 1 \mid X{=}x, do(S{=}0), S{=}1)$

Causal Fairness

Counterfactual Fairness

Total Causal Effect Fairness: $\mathbb{P}(O_{S\leftarrow 1}=1)=\mathbb{P}(O_{S\leftarrow 0}=1)$

$\mathbb{P}(O_{S\leftarrow 1}=1)= \sum_u \mathbb{P}(O_{S\leftarrow 0}(u)=1)\,\mathbb{P}(u)$

Exogenous variables

$\mathbb{P}(O_{S\leftarrow 1}=1 \mid X=x, S=1)=\mathbb{P}(O_{S\leftarrow 0}=1 \mid X=x, S=1)$

$\mathbb{P}(O_{S\leftarrow 1}=1 \mid X=x, S=0)=\mathbb{P}(O_{S\leftarrow 0}=1 \mid X=x, S=0)$

Can not be captured using the do-operator

$\cancel{\mathbb{P}(O=1 \mid X=x, do(S=0), S=1)}$

Causal Fairness

Equalized Counterfactual Odds

Equalized Odds:

$$\mathbb{P}\{O=1|S=1,Y=1\}=\mathbb{P}\{O=1|S=0,Y=1\}$$

$$\mathbb{P}\{O=1|S=1,Y=0\}=\mathbb{P}\{O=1|S=0,Y=0\}$$

Before intervention

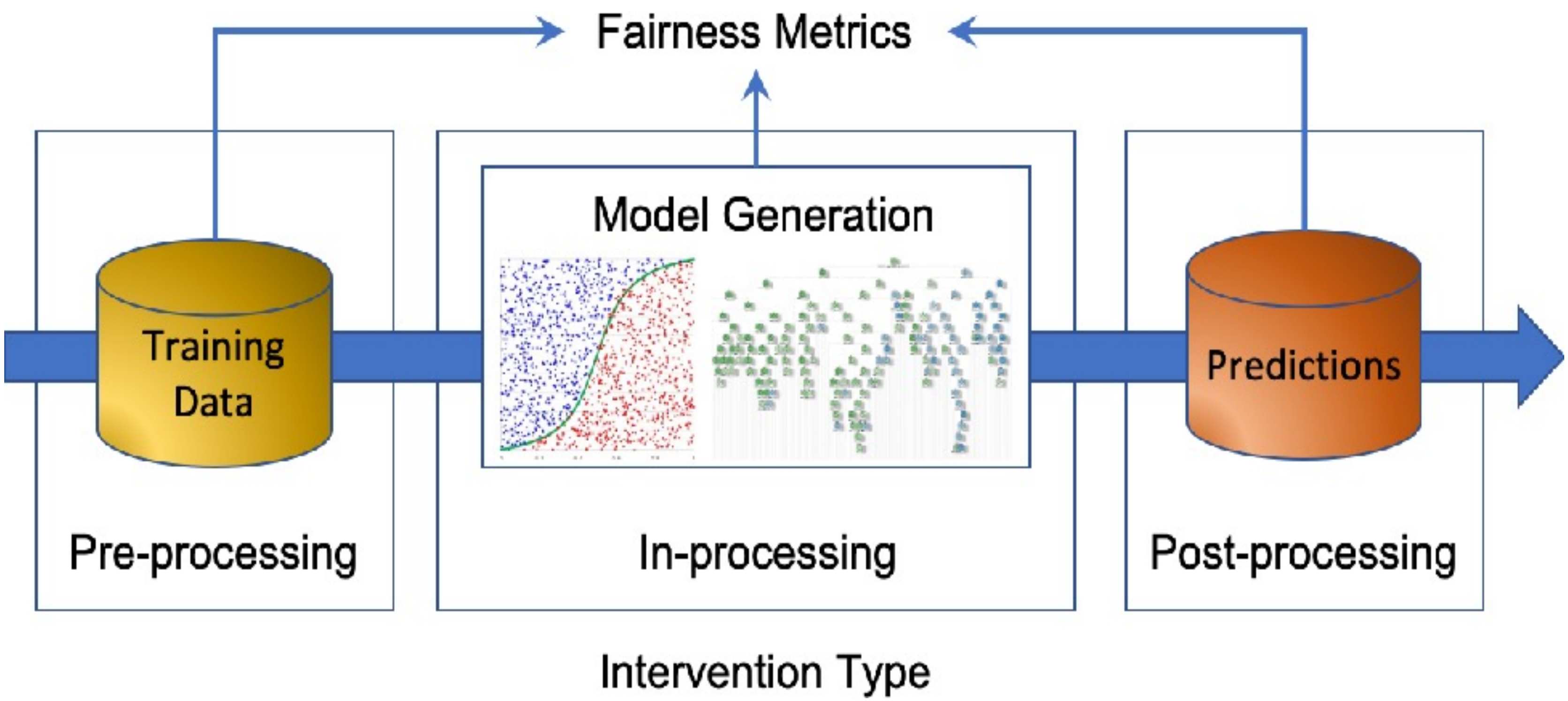
$$\mathbb{P}(O_{S\leftarrow 1}=1 \mid S=1, Y=0)=\mathbb{P}(O_{S\leftarrow 1}=1 \mid S=0, Y=0)$$

After intervention

$$\mathbb{P}(O_{S\leftarrow 1}=1 \mid S=1, Y_{S\leftarrow 1}=0)=\mathbb{P}(O_{S\leftarrow 1}=1 \mid S=0, Y_{S\leftarrow 1}=0)$$

$$\mathbb{P}(O_{S\leftarrow 1}=1 \mid S=1, X=x, Y_{S\leftarrow 1}=0)=\mathbb{P}(O_{S\leftarrow 0}=1 \mid S=0, X=x, Y_{S\leftarrow 1}=0)$$

Building Fair Models



Source: Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." *arXiv preprint arXiv:2010.04053* (2020)

Take Aways

☞ Fairness is causal concept

☞ One can define a causal counterpart for any existing associational notions of fairness

☞ Causal reasoning enable disentangling the observed statistical dependence between sensitive attribute and outcome into fine grained causal quantities

☞ Proving discrimination is as difficult as establishing causation

Outline

Sudeepa

- Introduction
- Pearl's Graphical Causal Model
- Rubin's Potential Outcome Framework
- Briefly: some recent research on causal inference techniques (scalability & relational)

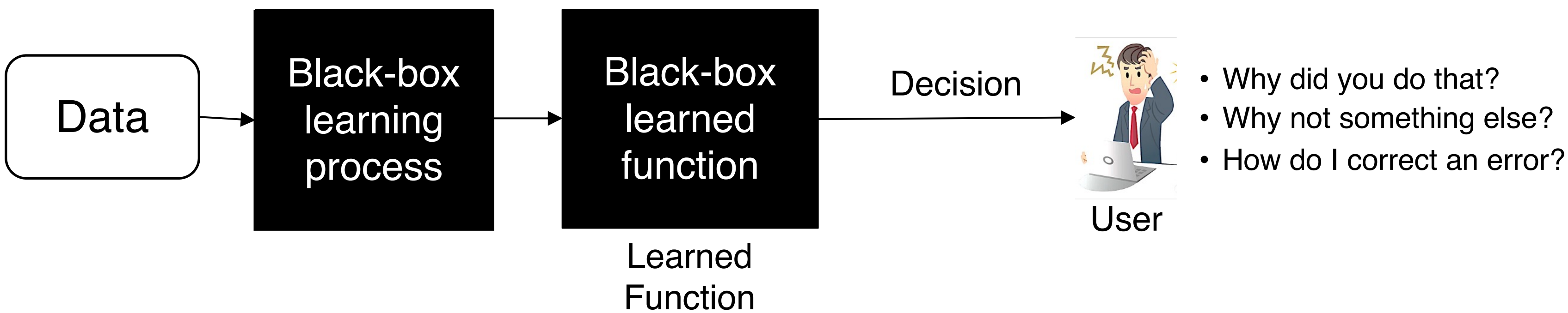
Babak

- Causal Fairness
- **Causal Explainability**

Causal Inference and Explainable AI

We interact with algorithmic decision-making on a daily basis

Sophisticated ML models shown to be highly accurate for many applications



- Complex models → difficult to trace decisions back to questions about why and how they were made
- ML systems are often opaque– complexity, proprietary
- What is XAI?

What is explainable AI?

- Human-understandable explanations of outcomes of algorithmic decision-making systems
- A powerful tool for answering How? and Why? questions about algorithmic systems

Not a new topic!

A Theory of Diagnosis from First Principles

Raymond Reiter

Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4; The Canadian Institute for Advanced Research

Recommended by Johan de Kleer and Daniel G. Bobrow

ABSTRACT

Suppose one is given a description of a system, together with an observation of the system's behaviour which conflicts with the way the system is meant to behave. The diagnostic problem is to determine those components of the system which, when assumed to be functioning abnormally, will explain the discrepancy between the observed and correct system behaviour.

We propose a general theory for this problem. The theory requires only that the system be described in a suitable logic. Moreover, there are many such suitable logics, e.g. first-order, temporal, dynamic, etc. As a result, the theory accommodates diagnostic reasoning in a wide variety of practical settings, including digital and analogue circuits, medicine, and database updates. The theory leads to an algorithm for computing all diagnoses, and to various results concerning principles of measurement for discriminating among competing diagnoses. Finally, the theory reveals close connections between diagnostic reasoning and nonmonotonic reasoning.

as a yardstick for measuring progress in the field. The paper describes

ons of XAI

Prevent bias and discrimination

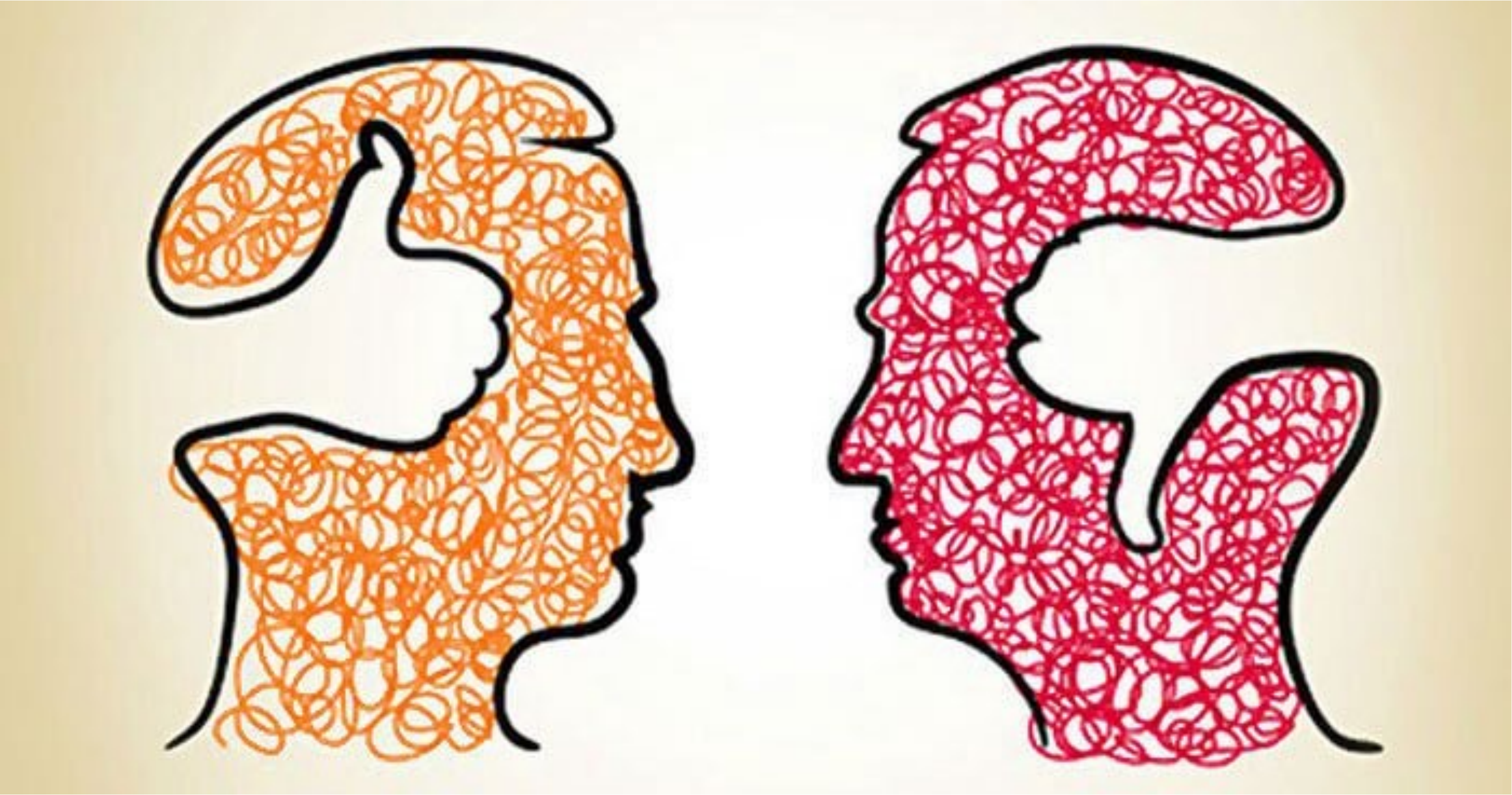
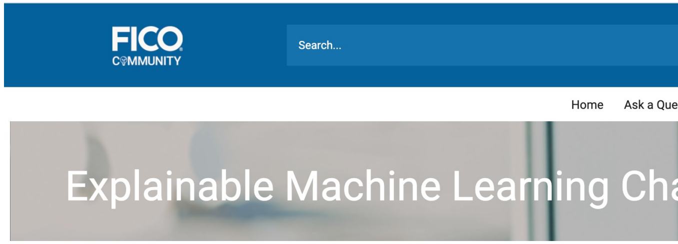
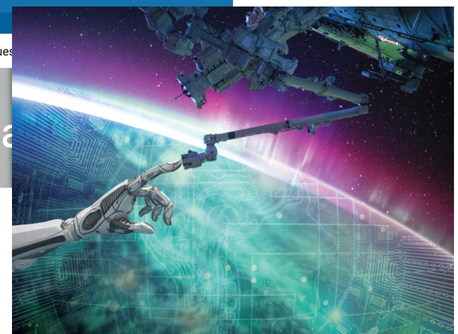


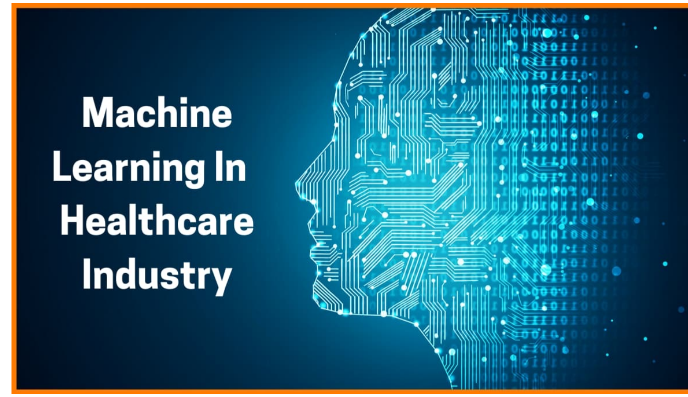
Image source: https://peasoup.deptcpanel.princeton.edu/2020/07/rima-basu-the-specter-of-normative-conflict-does-fairness-conflict-with-accuracy/

ons of XAI

Accuracy, trust, recourse and compliance with the law







"

Companies should commit to ensuring systems that could fall under GDPR, including AI, will be compliant. The threat of **sizeable fines of €20 million or 4% of global turnover** provides a sharp incentive.

Article 22 of GDPR empowers individuals with the **right to demand an explanation of how an AI system made a decision that affects them.**

"

European Commission

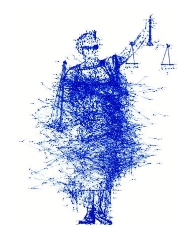
"The data subject shall have the **right not to be subject to a decision based solely on automated processing**…"

GDPR.EU

"… **monitor city use of algorithmic decision-making** and provide recommendations…"

NYC
Automated Decision Systems Task Force

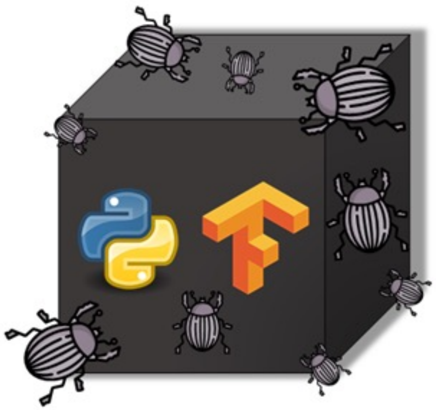
required companies to **study algorithms they use, identify bias in these systems and fix any discrimination** or bias they find

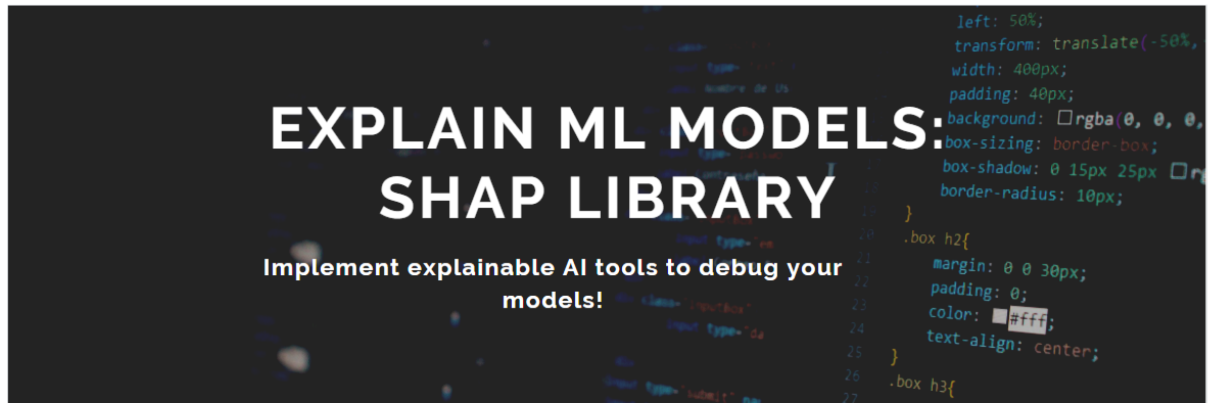


Algorithmic Accountability Act 2019

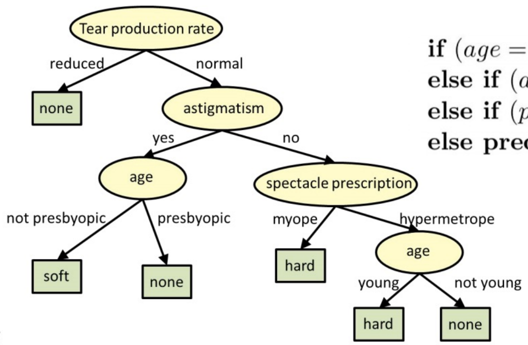
ons of XAI

Debugging ML model



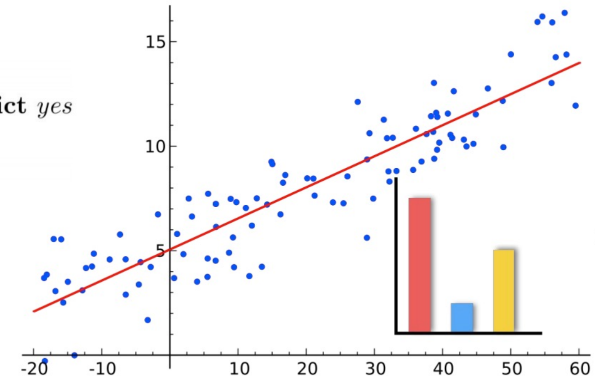


How are such decisions explained?



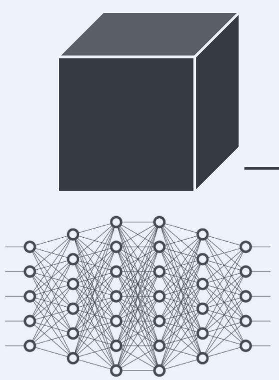
if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

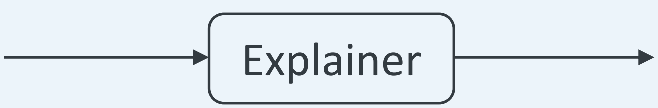
Intrinsic

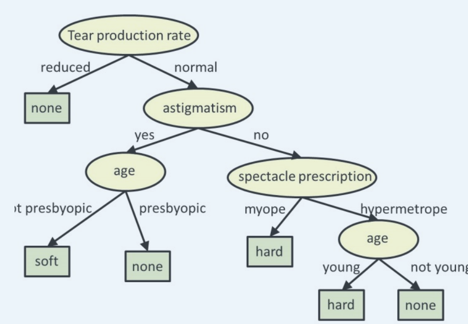


Post hoc

if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no





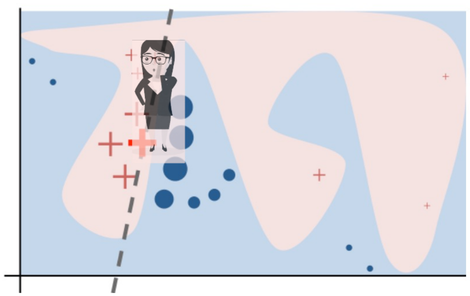


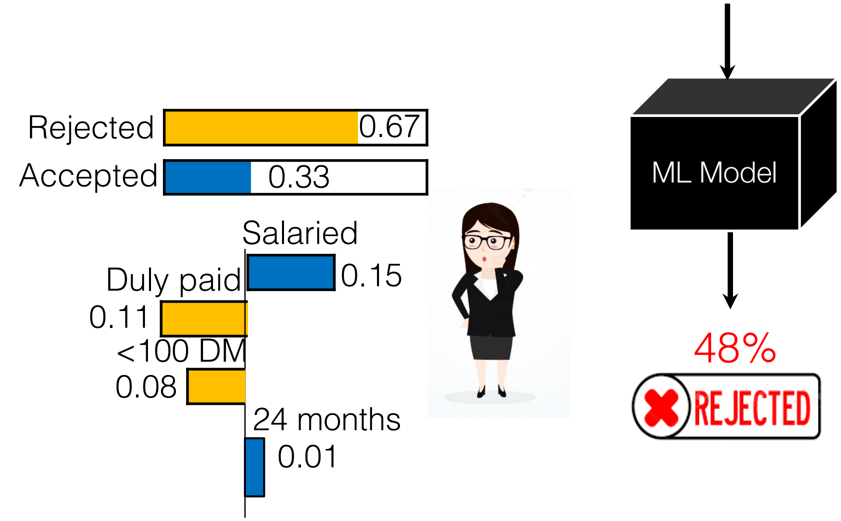
Focus of this Talk

Feature Attribution

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	24	…	Paid duly

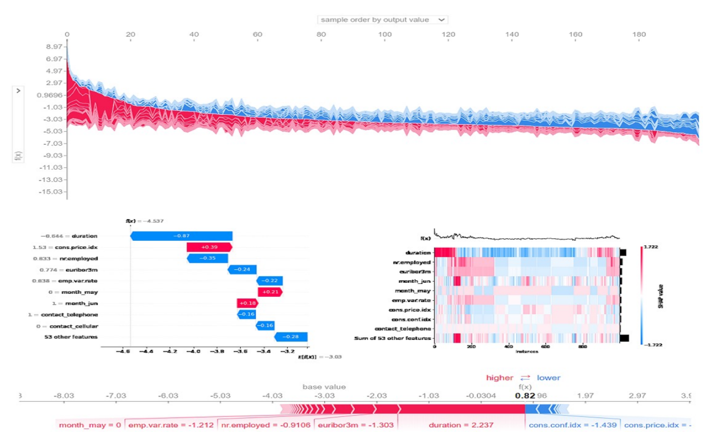
Feature Importance:
Identify important attributes and present their relative importance





LIME. Identify important attributes to explain outcome for a single instance

SHAP. Marginal contribution of each feature toward the prediction, averaged over all possible coalitions

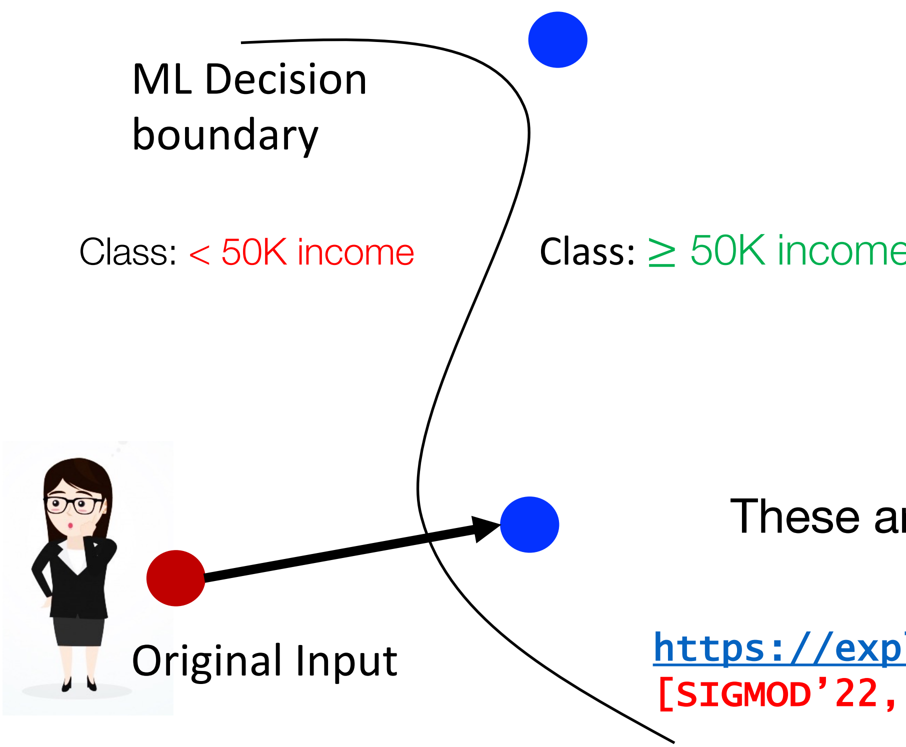


Counterfactual explanations

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	24	…	Paid duly

Counterfactual Explanations:

how the world would have (had) to be different for a desirable outcome to occur”



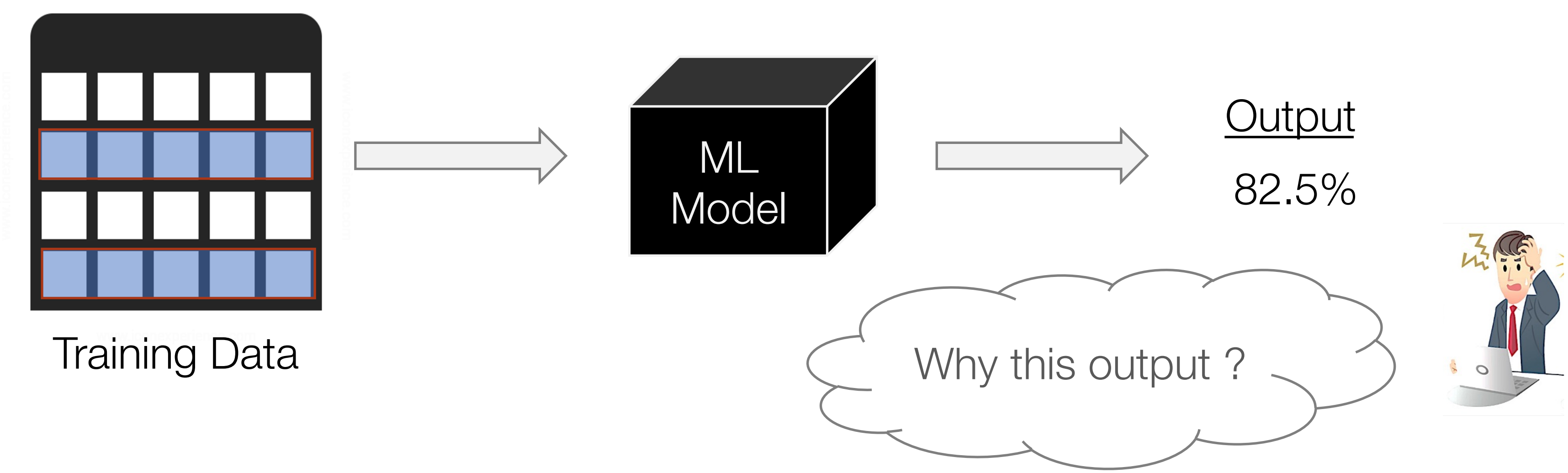
These are contrastive explanations!

https://explainable-ai-tutorial.github.io/

[SIGMOD’22, ICDE’22]

Limitations

- Focus on correlation between input/output
- Fail to account for the causal interaction between attributes
- Perturbations not translatable to real-world interventions
- Fall short in generating diagnostic explanations



What kind of explanations do humans seek?

CONTRASTIVE

"Explanatory relevant information is information that is potentially **relevant to manipulation and control**"
—James Woodward, Philosopher

The key insight is to recognize that one does not explain events per se, but that one explains **why the puzzling event occurred in the target cases but not in some counterfactual contrast case**.
—Denis Hilton, Psychologist

"An explanation is an **assignment of causal responsibility**"
—Josephson and Josephson, Computer Scientists

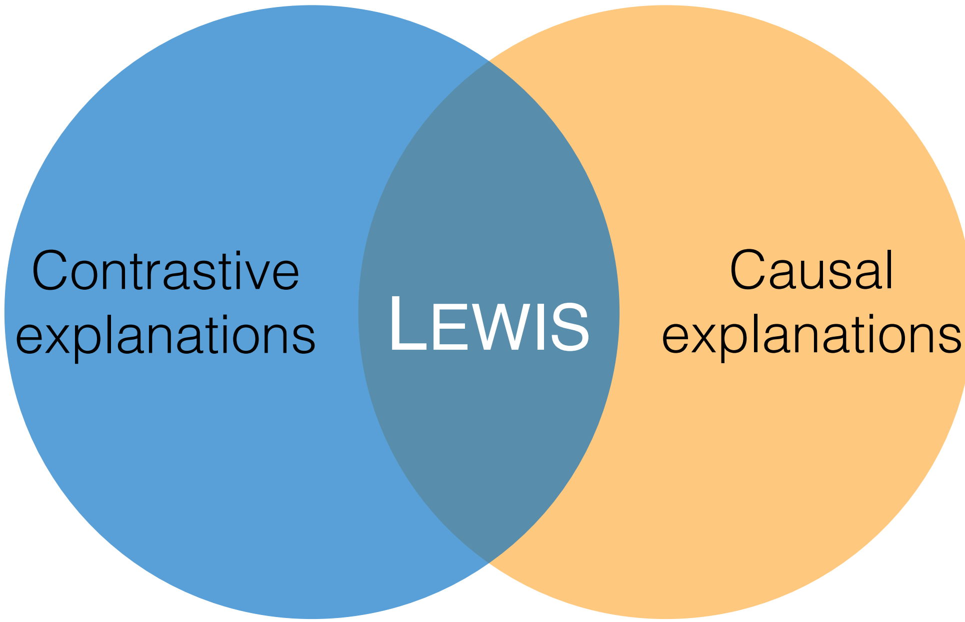
"To explain an event is to provide some **information about its causal history**."
"… think of a cause as something that makes a difference… **Had it been absent, its effects** – some of them, at least, and usually all – **would have been absent as well**"
—David Lewis, Philosopher

CAUSAL

"…the only way to **ensure that the recommended change is even possible**…"
"…and to **account for dependencies between features** is to model the outcome of interest using features that directly figure into the causal mechanism"
—Barocas et al. FAT 2020

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

Unifying Contrastive and feature attribution methods



[SIGMOD'21]
[VLDB'21]

Based on probabilistic contrastive counterfactuals

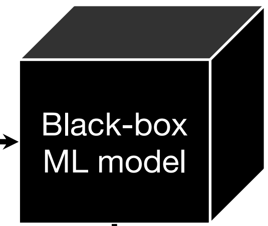
"For individual(s) with attribute(s) <actual-value> for whom an algorithm made the decision <actual-decision>, the decision would have been <foil-outcome> with probability <score> had the attribute been <counterfactual-value>."

To what extent is an attribute sufficient?

Maeve,
a bank customer

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	36	…	Existing paid duly

Maeve's data



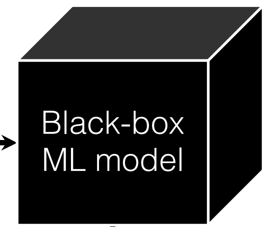
Maeve's credibility 87%

What is the probability that loan would be approved if month had been 36 months instead of 24 months?

To what extent is an attribute necessary?

Name	Age	Saving	Month	…	Credit History
Irrfan	>25	>500 DM	24	…	Existing paid duly

Irrfan's data



Irrfan,
a bank customer

Irrfan's credibility 48%

What is the probability that loan would be rejected if month had been 24 months instead of 36 months?

Our explanation scores

attribute X has value x
positive model outcome
instances in context

Necessity score

$$\textsc{Nec}_{x}^{x'}(\mathbf{k}) = \Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k})$$

Probability that outcome would have been negative had X been x’

Probability that
Loan=Rejected if Month were 24
for instances in context for whom
Loan=Approved when Month=36

Our explanation scores

$$\textsc{Nec}_x^{x'}(\mathbf{k}) = \Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k})$$

attribute X has value x'

negative model outcome

instances in context

Sufficiency score

$$\boxed{\textsc{Suf}_x^{x'}(\mathbf{k}) = \Pr(o_{X \leftarrow x} \mid x', o', \mathbf{k})}$$

Probability that outcome would have been positive had X been x

Probability that
Loan=Approved if Month were 36
for instances in context for whom
Loan=Rejected when Month=24

Our explanation scores

$$\textsc{Nec}_{x}^{x'}(\mathbf{k}) = \Pr(o'_{X\leftarrow x'} \mid x, o, \mathbf{k})$$

$$\textsc{Suf}_{x}^{x'}(\mathbf{k}) = \Pr(o_{X\leftarrow x} \mid x', o', \mathbf{k})$$

Probability that outcome would be positive
had X been x and negative had X been x’

Necessity & Sufficiency score

$$\boxed{\textsc{NeSuf}_{x}^{x'}(\mathbf{k}) = \Pr(o_{X\leftarrow x}, o'_{X\leftarrow x'} \mid \mathbf{k})}$$

Probability that
Loan=Approved were Month=36 and
Loan=Rejected were Month=24

Our explanation scores

$$\textsc{Nec}_x^{x'}(\mathbf{k}) = \Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k})$$

$$\textsc{Suf}_x^{x'}(\mathbf{k}) = \Pr(o_{X \leftarrow x} \mid x', o', \mathbf{k})$$

$$\textsc{NeSuf}_x^{x'}(\mathbf{k}) = \Pr(o_{X \leftarrow x}, o'_{X \leftarrow x'} \mid \mathbf{k}),$$

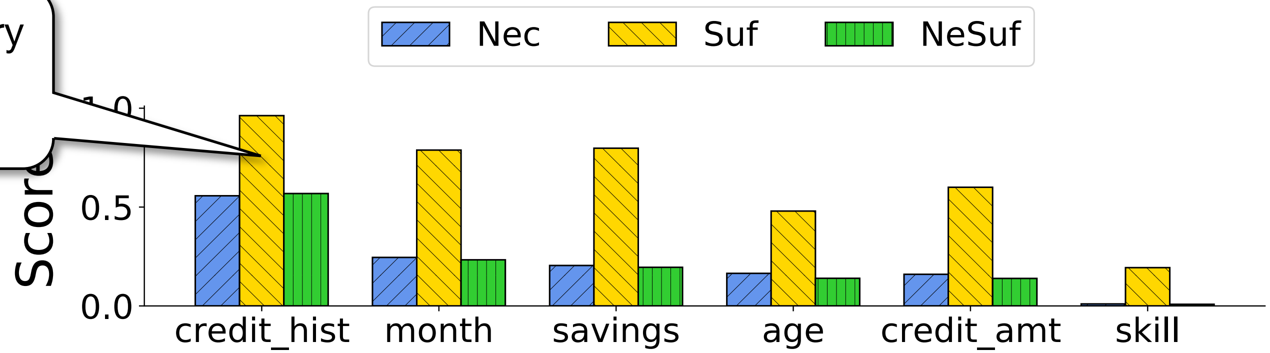
$$\begin{aligned}\textsc{NeSuf}_x^{x'}(\mathbf{k}) \leq\ & \Pr(o, \mathbf{x} \mid \mathbf{k}) \boxed{\textit{Nec}_{\mathbf{x}}(\mathbf{k})} \\ & + \Pr(o', \mathbf{x}' \mid \mathbf{k}) \boxed{\textit{Suf}_{\mathbf{x}}(\mathbf{k})} \\ & + 1 - \Pr(\mathbf{x} \mid \mathbf{k}) - \Pr(\mathbf{x}' \mid \mathbf{k})\end{aligned}$$

From scores to explanations

$$\textsc{Suf}_x^{x'}(\mathbf{k})$$

Global explanations

An improvement in credit history is most likely to change a negative decision into positive.



$$\mathrm{S{\small UF}}_x^{x'}(\mathbf{k} = \phi)$$

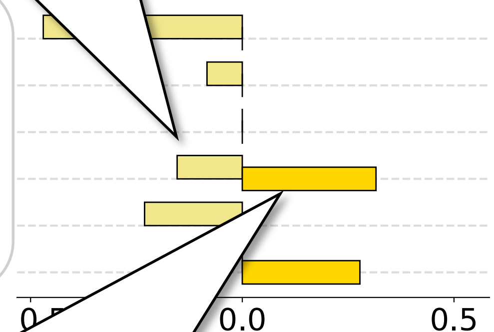
- Scores express the global influence of attributes on algorithm's decision
- Maximum score over all pairs of an attribute's values

Local explanations

$$\textsc{Suf}_x^{x'}(\mathbf{k} = \text{Maeve's data})$$

Age	< 25 years
Credit amount	1,275 DM
Month	10 months
Credit history	Existing paid duly
Savings	< 100 DM
Skill level	Skilled

Changing from current credit history value to a better value would have resulted in positive outcome with 15% probability.



- Contribution of attri oward
- Positive and negative contribution of an attribute value e.g., credit history

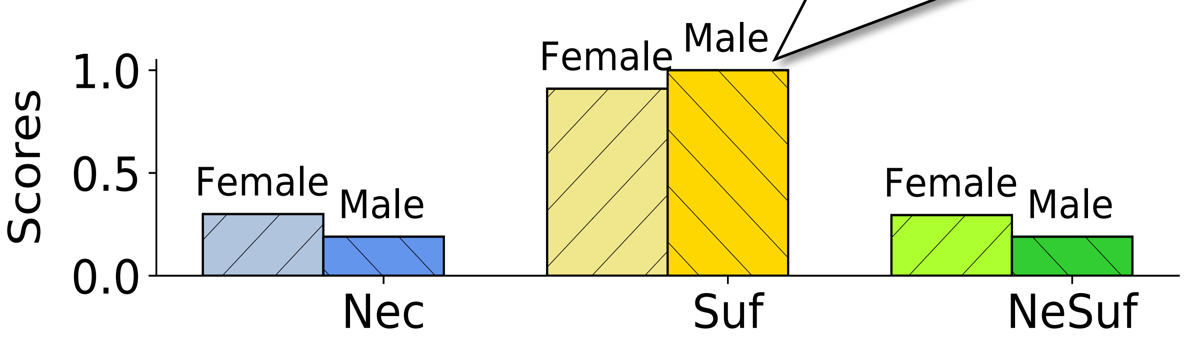
Staying at current credit history value instead of a worse value has positive contribution toward outcome.

Contextual explanations

$$\textsc{Suf}_{x}^{x'}(\mathbf{k} = \text{Sub-population})$$

Improving status of checking account is more likely to flip a negative decision for Sex=Male than for Sex=Female.

Effect of changing Status of checking account



How can I improve my chances of getting a loan?

User provides a set of actionable attributes A

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	24	…	Existing paid duly

Counterfactual recourse as interventions over actionable attributes

$$\underset{\mathbf{a}\in \mathrm{Dom}(\mathbf{A})}{\operatorname{argmin}} \quad \mathrm{Cost}\left(\mathbf{a}, \hat{\mathbf{a}}\right)$$
$$\text{s.t.} \quad \textsc{Suf}_{\hat{\mathbf{a}}}(\text{Maeve's data}) \geq 85\%$$

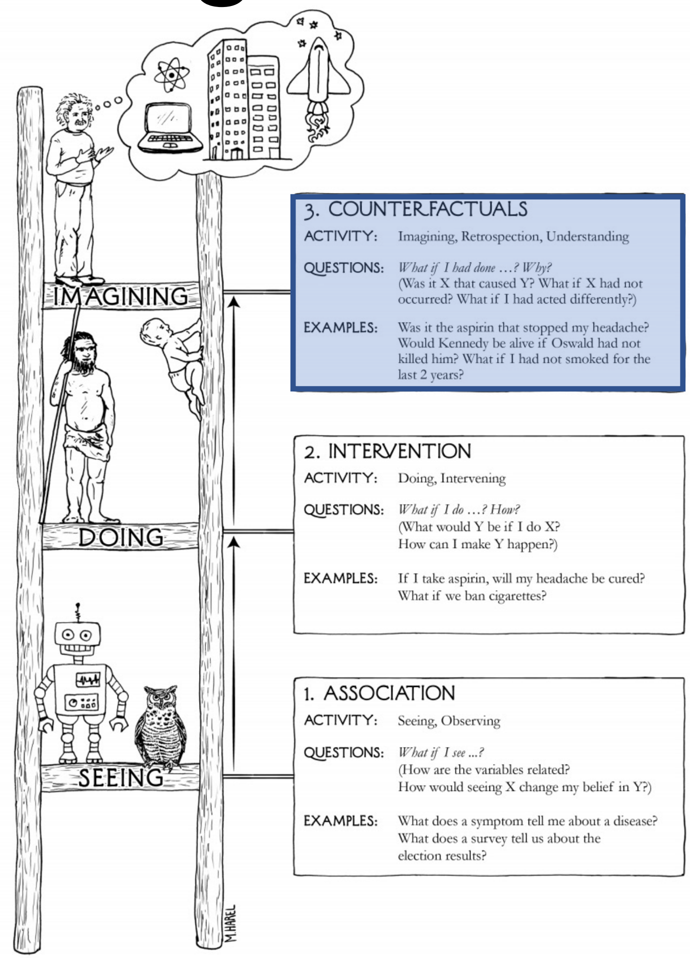
Recommended Recourse:

Actionable Attributes	Current Value	Required Value
Credit amount	1,275 DM	3,000 – 5,000 DM
Savings	< 100 DM	500 – 1,000 DM

This recourse will lead to a positive decision

Computational Challenges

$\textsc{Nec}_x^{x'}(\mathbf{k})$	$\Pr(o'_{X\leftarrow x'} \mid x, o, \mathbf{k})$
$\textsc{Suf}_x^{x'}(\mathbf{k})$	$\Pr(o_{X\leftarrow x} \mid x', o', \mathbf{k})$
$\textsc{NeSuf}_x^{x'}(\mathbf{k})$	$\Pr(o_{X\leftarrow x}, o'_{X\leftarrow x'} \mid \mathbf{k})$



Requires the full structural causal model

Age = U_{Age}

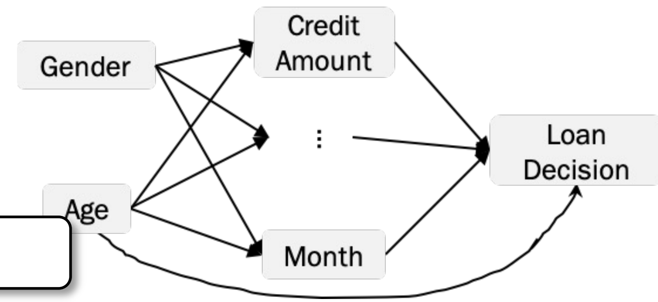
Gender = U_{Gender}

Month = 0.4(Age) + 0.8(Gender) + U_{month} …

Decision = Month + 0.5(Age) + … + $U_{Decision}$

Computing scores from data

- Scores can be **bounded** with background knowledge on causal graph (**confounders**)



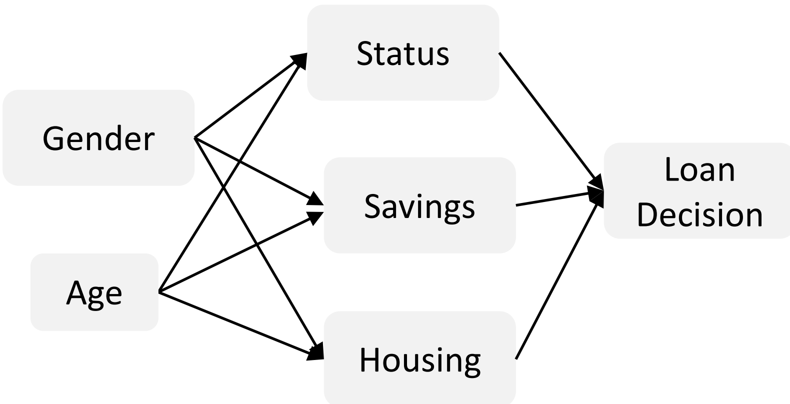
do-operator

$$\max\left(0, \frac{\Pr(o', \mathbf{x} \mid \mathbf{k}) + \Pr(o', \mathbf{x}' \mid \mathbf{k}) - \Pr(o' \mid \mathrm{do}(\mathbf{x}), \mathbf{k})}{\Pr(o', x' \mid \mathbf{k})}\right) \leq \textsc{Suf}_{\mathbf{x}}(\mathbf{k}) \leq \min\left(\frac{\Pr(o \mid \mathrm{do}(\mathbf{x}), \mathbf{k}) - \Pr(o, x \mid \mathbf{k})}{\Pr(o', x' \mid \mathbf{k})}, 1\right)$$

- Assuming **monotonicity** of the algorithm relative to attribute values, we can compute the scores from historical data

$$\textsc{Suf}_{\mathbf{x}}(\mathbf{k}) = \frac{\left(\sum_{\mathbf{c} \in Dom(\mathbf{C})} \Pr(o \mid \mathbf{c}, \mathbf{x}, \mathbf{k}) \Pr(\mathbf{c} \mid \mathbf{x}', \mathbf{k})\right) - \Pr(o \mid \mathbf{x}', \mathbf{k})}{\Pr(o' \mid \mathbf{x}', \mathbf{k})}$$

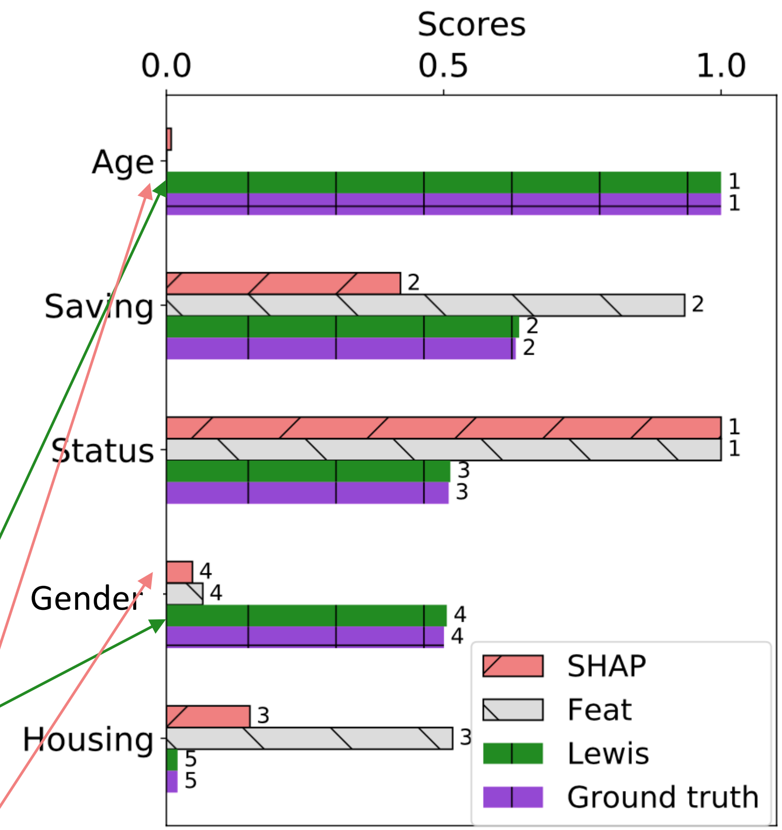
Experiments



Suppose the algorithm does not use gender and age for prediction

Age and gender still have a causal influence on the decision

SHAP picks up correlation between attributes– ranks age and gender the lowest



Synthetic dataset

Take Aways

A causal approach to generate feature-based explanations

- ☞ Scores based on probabilistic contrastive counterfactuals
- ☞ Captures causal dependencies between attributes
- ☞ Computes provably effective explanations at the global, local and contextual levels and generates recourse options
- ☞ Works with varying levels of background knowledge of causal model

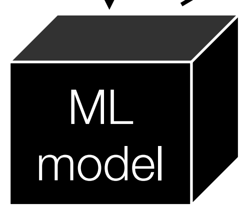
Experimental result: contrary to state-of-the-art approaches in XAI (e.g., LIME, SHAP), our framework provides causality-based explanations

Beyond feature-based explanations

Rosa

Name	Age	Education	Marital	…	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	female	40

Rosa’s data

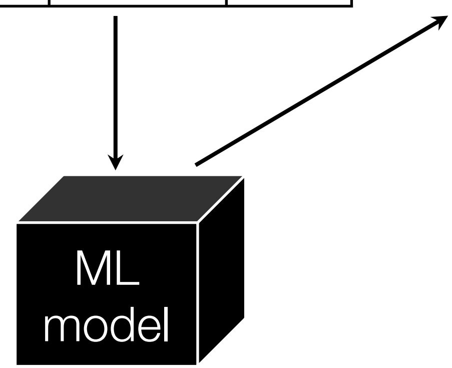


<50K

Rosa’s
Income

Example: Detecting bias

Name	Age	Education	Marital	…	Gender	Hours	
Rosa	34	Bachelors	Unmarried	…	female	40	<50K
Seok	29	Bachelors	Unmarried	…	male	25	>50K
Aisha	30	Masters	Married	…	female	35	<50K
Deka	55	Masters	Married	…	nonbinary	40	>50K



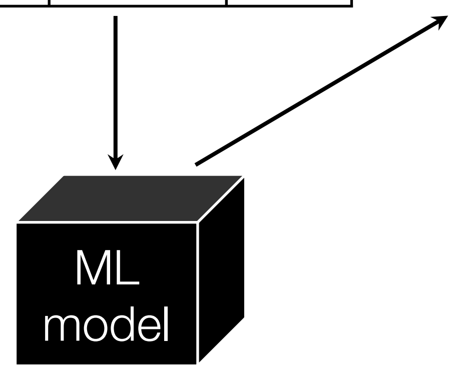
Example: Detecting bias

Name	Age	Education	Marital	...	Gender	Hours
Rosa	34	Bachelors	Unmarried	...	female	40
Seok	29	Bachelors	Unmarried	...	male	25
Aisha	30	Masters	Married	...	female	35
Deka	55	Masters	Married	...	nonbinary	40

<50K
>50K
<50K
>50K

20%

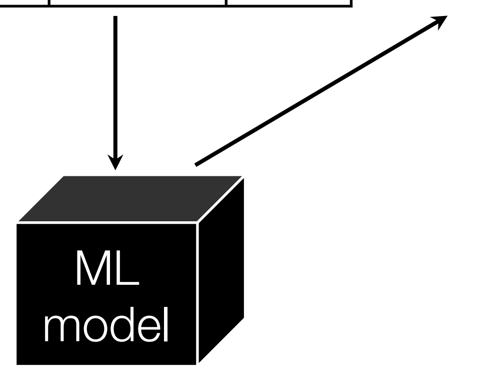
Fairness metric



Example: Detecting bias

Bias! Female persons are less likely to receive high salary predictions then male persons

Name	Age	Education	Marital	…	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	female	40
Seok	29	Bachelors	Unmarried	…	male	25
Aisha	30	Masters	Married	…	female	35
Deka	55	Masters	Married	…	nonbinary	40



<50K
>50K
<50K
>50K

20%

Fairness metric

Example: Detecting bias

Bias! Female persons are less likely to receive high salary predictions then male persons

Name	Age	Education	Marital	...	Gender	Hours
Rosa	34	Bachelors	Unmarried	...	female	40
Seok	29	Bachelors	Unmarried	...	male	25
Aisha	30	Masters	Married	...	female	35
Deka	55	Masters	Married	...	nonbinary	40

<50K
>50K
<50K
>50K

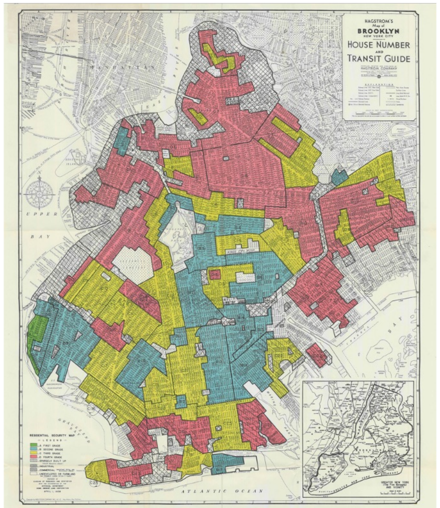
20%

Fairness metric

Why does the model exhibit this unexpected or discriminatory behavior?

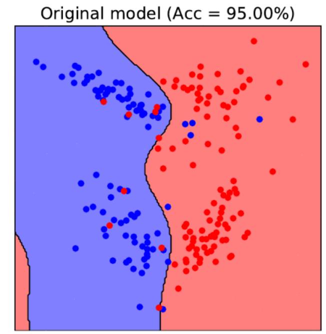
Model bias attributed to training data bias

src: NYTimes



Historical bias in training data

src: openai.com



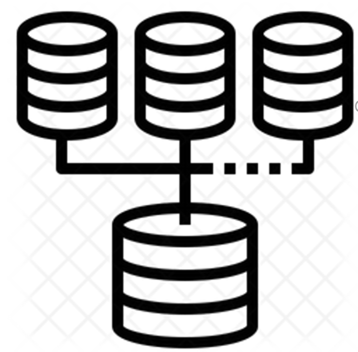
src: https://labs.f-secure.com

Adversarial data attacks

src: datacubed.com

Selection bias

src: nagwa.com

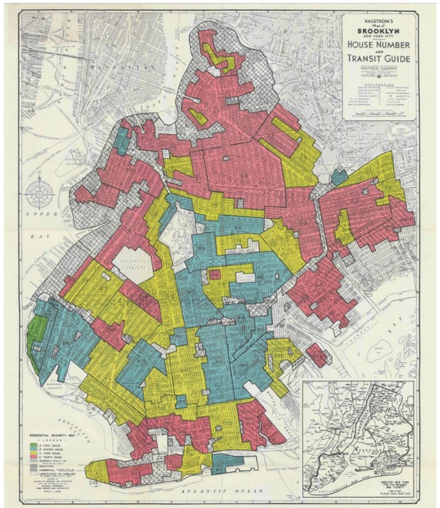


src: nagwa.com

Data integration

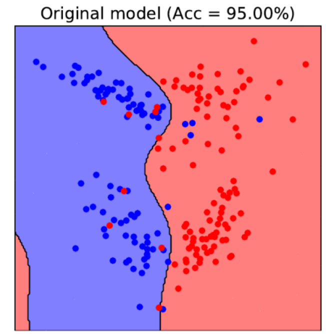
Model bias attributed to training data bias

src: NYTimes



Historical bias in training data

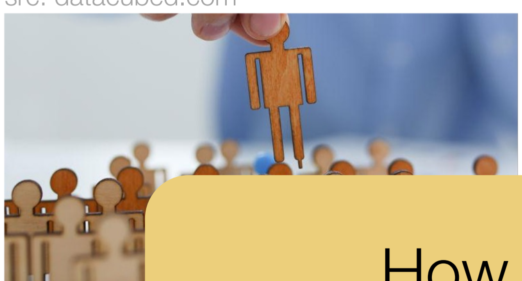
src: openai.com



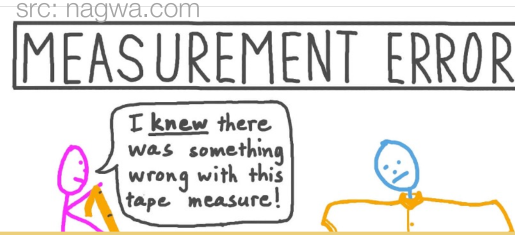
src: https://labs.f-secure.com

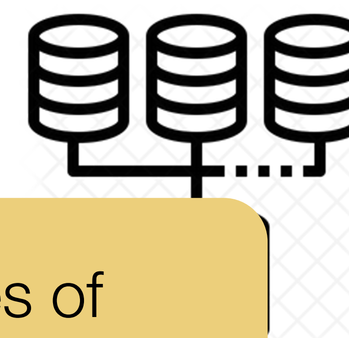
Adversarial data attacks

src: datacubed.com



src: nagwa.com





src: nagwa.com

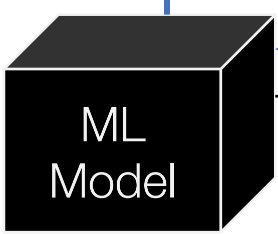
tion

How can we debug such sources of model bias?

Post hoc data-based explanations

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
Matt	26	Bachelors	Married	…	male	40	≥50K
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Training data



Gopher
[SIGMOD'22]

Name	Age	Education	Marital	…	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	female	40
Seok	29	Bachelors	Unmarried	…	male	25
Aisha	30	Masters	Married	…	female	35
Deka	55	Masters	Married	…	nonbinary	40

Test data

<50K
>50K
<50K
>50K

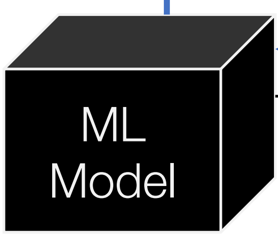
20%

Fairness metric

Post hoc data-based explanations

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
Matt	26	Bachelors	Married	…	male	40	≥50K
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Training data



- Root causes of bias
- Causal responsibility

Name	Age	Education	Marital	…	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	female	40
Seok	29	Bachelors	Unmarried	…	male	25
Aisha	30	Masters	Married	…	female	35
Deka	55	Masters	Married	…	nonbinary	40

Test data

<50K
>50K
<50K
>50K

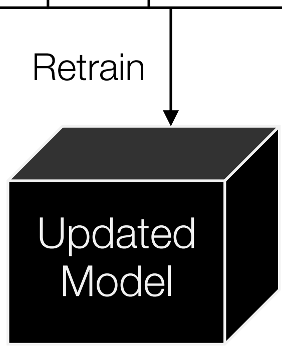
20%

Fairness metric

Causal responsibility

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
~~Matt~~	~~26~~	~~Bachelors~~	~~Married~~	~~…~~	~~male~~	~~40~~	~~≥50K~~
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Updated training data



Intervention
- Remove or upweight training data of interest
- Retrain the model
- Measure change in fairness

Causal responsibility

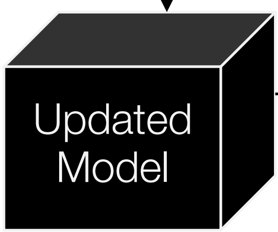
Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
~~Matt~~	~~26~~	~~Bachelors~~	~~Married~~	~~…~~	~~male~~	~~40~~	~~≥50K~~
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Updated training data

Intervention

- Remove or upweight training data of interest
- Retrain the model
- Measure change in fairness

Retrain



Name	Age	Education	Marital	…	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	female	40
Seok	29	Bachelors	Unmarried	…	male	25
Aisha	30	Masters	Married	…	female	35
Deka	55	Masters	Married	…	nonbinary	40

Test data

<50K
>50K
<50K
>50K

20%

18%

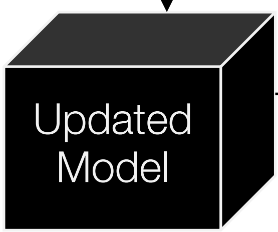
Fairness metric

Approximate causal responsibility using influence functions

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
~~Matt~~	~~26~~	~~Bachelors~~	~~Married~~	~~…~~	~~male~~	~~40~~	~~≥50K~~
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Updated training data

Retrain



Name	Age	Education	Marital	…	Gender	Hours
Rosa						
Seok						
Aisha						
Deka						

We want to compute the change in fairness metric without retraining the model

Influence functions

Cook, R. D., Weisberg, S. "*Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression*". Technometrics 1980

Koh, P. W., Liang, P. "*Understanding Black-box Predictions via Influence Functions*". ICML 2017

How to summarize data-based explanations?

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
Matt	26	Bachelors	Married	…	male	40	≥50K
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Training data

These instances are the most responsible for model unfairness

Gopher’s data-based explanations

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
Matt	26	Bachelors	Married	…	male	40	≥50K
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Training data

Most responsible for bias: Gender=‘male’

Gopher’s data-based explanations

Name	Age	Education	Marital	...	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	...	female	40	<50K
Matt	26	Bachelors	Married	...	male	40	≥50K
Yeji	50	Masters	Married	...	male	16	≥50K
Neel	45	Masters	Unmarried	...	male	28	<50K
...	...	...	...	...	...	...	...

Training data

Most responsible for bias: Gender=‘male’

To generate interpretable explanations, we need to know how much a **coherent** subset contributes to bias

Basu, S., You, X., Feizi, S. "*On Second-Order Group Influence Functions for Black-Box Predictions*". ICML 2020

Computational challenges

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
Matt	26	Bachelors	Married	…	male	40	≥50K
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Training data

Compute bias:
Gender='male'

- Compute causal responsibility of all subsets

Computational challenges

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
Matt	26	Bachelors	Married	…	male	40	≥50K
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Training data

Compute bias:
Gender='male' and
Marital='Married'

- Compute causal responsibility of all subsets

Computational challenges

Name	Age	Education	Marital	...	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	...	female	40	<50K
Matt	26	Bachelors	Married	...	male	40	≥50K
Yeji	50	Masters	Married	...	male	16	≥50K
Neel	45	Masters	Unmarried	...	male	28	<50K
...	...	...	...	...	...	...	...

Training data

Compute bias:
Education='Bachelors'

- Compute causal responsibility of all subsets

Computational challenges

Name	Age	Education	Marital	…	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	female	40	<50K
Matt	26	Bachelors	Married	…	male	40	≥50K
Yeji	50	Masters	Married	…	male	16	≥50K
Neel	45	Masters	Unmarried	…	male	28	<50K
…	…	…	…	…	…	…	…

Training data

Compute bias:
Marital='Unmarried'

- Compute causal responsibility of all subsets (large number of them)
- Sort in decreasing order of causal responsibility and select the top-k subsets with the highest responsibility
- A lattice-based search algorithm for pattern extraction
- Gopher also generates update-based explanations

Take Aways

A causal approach to generate data-based explanations

- ☞ Generate compact, interpretable, and causal explanations
- ☞ Identify coherent subsets of training data that are root causes for this bias
- ☞ Design efficient algorithms for generating top-k patterns that explain model bias
- ☞ Design efficient algorithms for generating top-k patterns that explain model bias
- ☞ Casting the problem of update-based explanation as as a constraint optimization problem

Thank you!

Questions?