INTERPRETABILITY AND EXPLAINABILITY FROM A CAUSAL LENS

Judea Pearl
IPAM Workshop
October 16, 2019
Twitter: @yudapearl

OUTLINE

- What is a causal lens?
- Why causal understanding needs a new logic, and a new inference engine
- The two fundamental laws ("double-helix") of causal inference
- The Seven Pillars (Tools) of Causal Wisdom
  - o how they are revolutionizing science,
  - o how they clarify social, legal, and ethical questions

WHAT IS A CAUSAL LENS?

- There exists an unknown but true Data Generating Process (DGP) that explains the world.
- The DGP comes as a set of CAUSAL equations
- Task: Infer properties of the DGP using data and assumptions about other properties of the DGP.
- Central: Consequences of pending policies on various populations or subpopulations.
- Central: Qualitative understanding of the DGP structure (in graphical form).

WHAT IS CAUSAL INFERENCE?

- A method of taking three inputs and producing answers to two types of causal questions.

Inputs: (1) What we wish to know
(2) What we do already know
(3) Available data

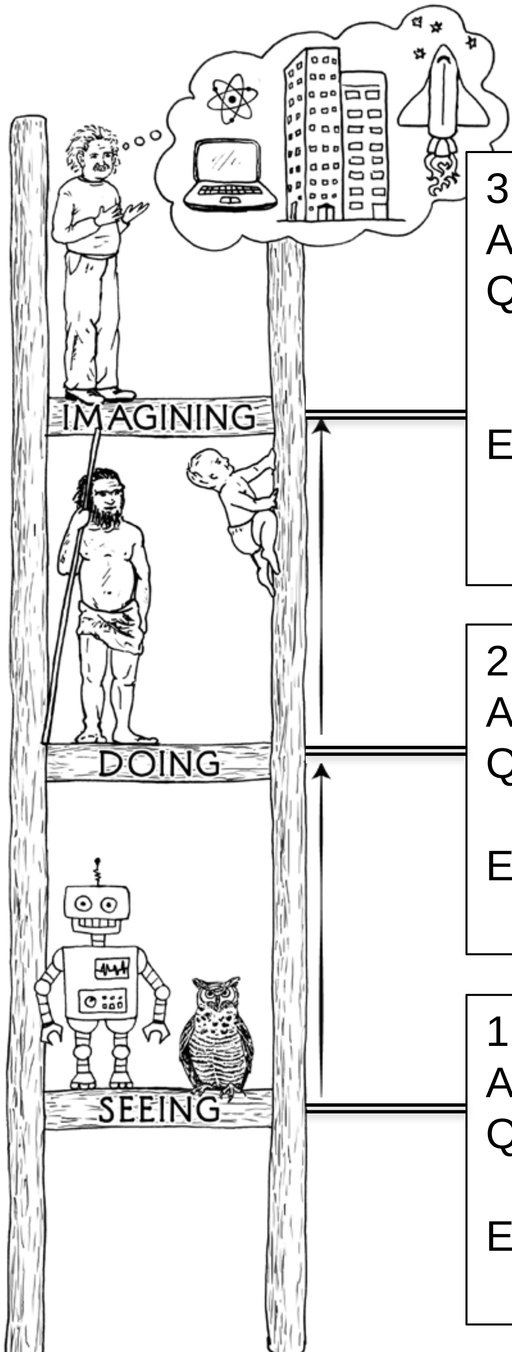
Outputs: (1a) effects of pending interventions
(1b) effects of undoing past events

TYPICAL CAUSAL QUESTIONS

1. How effective is a given treatment in preventing a disease?
2. Was it the new tax break that caused our sales to go up? Or our marketing campaign?
3. What is the annual health-care costs attributed to obesity?
4. Can hiring records prove an employer guilty of sex discrimination?
5. I am about to quit my job, will I regret it?
   - Unarticulatable in the standard grammar of science.

$$Y = aX \quad \text{vs.} \quad Y \leftarrow aX$$

3-LEVEL HIERARCHY



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done . . . ? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache?
Would Kennedy be alive if Oswald had not killed him? What if I had not smoked the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do . . . ? How?*
(What would Y be if I do X?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

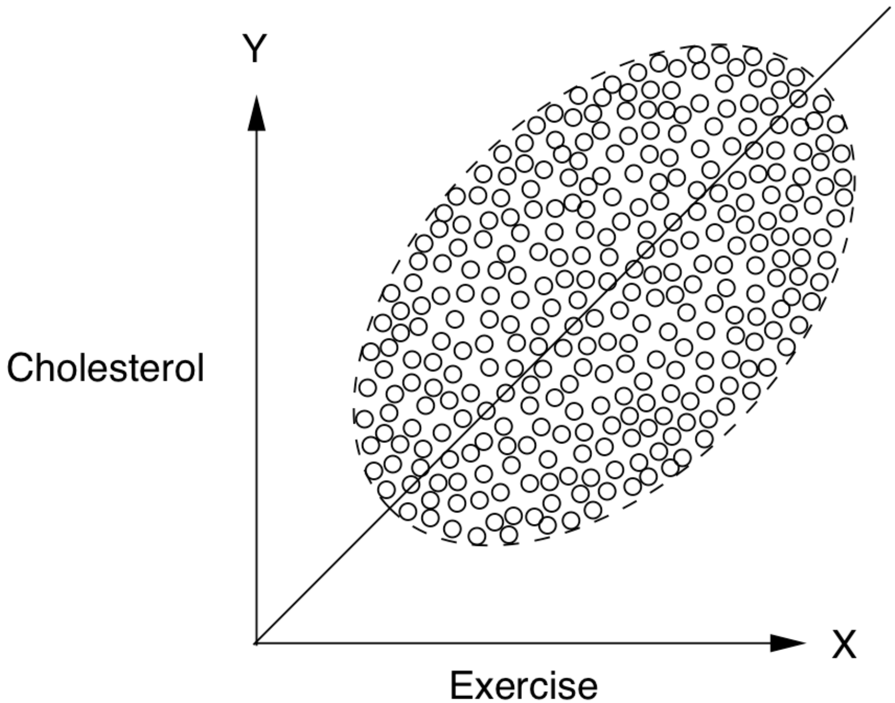
1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see . . . ?*
(How would seeing X change my belief in Y?)

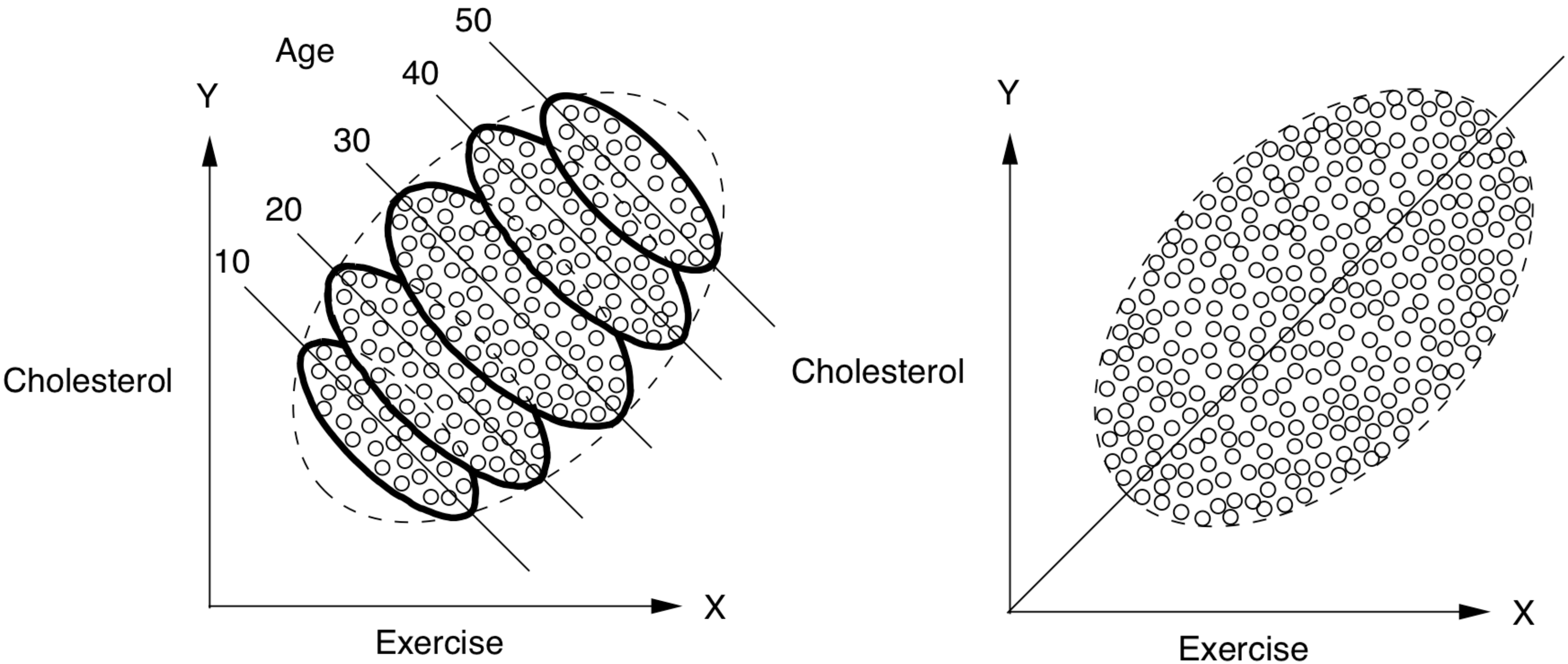
EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

WHY DATA CAN BE DUMB



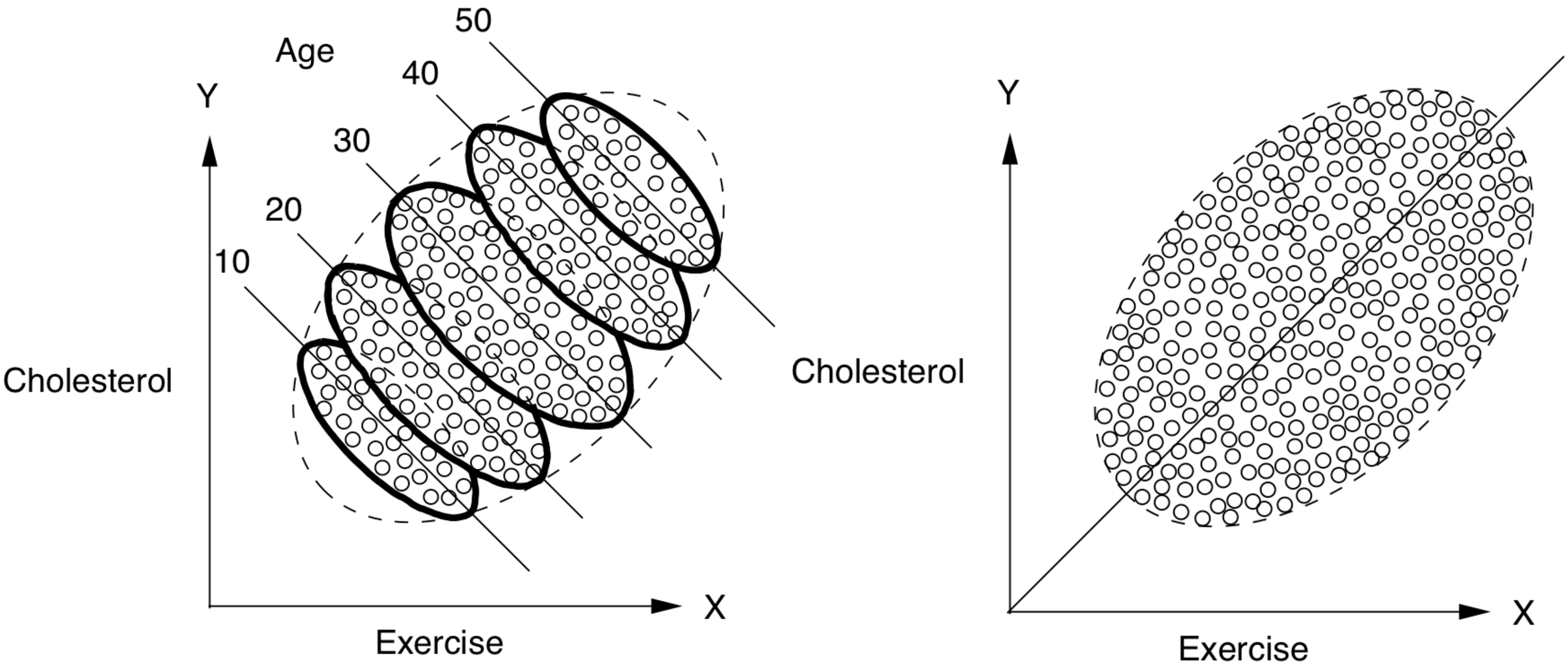
Exercise seems to increase cholesterol level in this population.

WHY DATA CAN BE DUMB



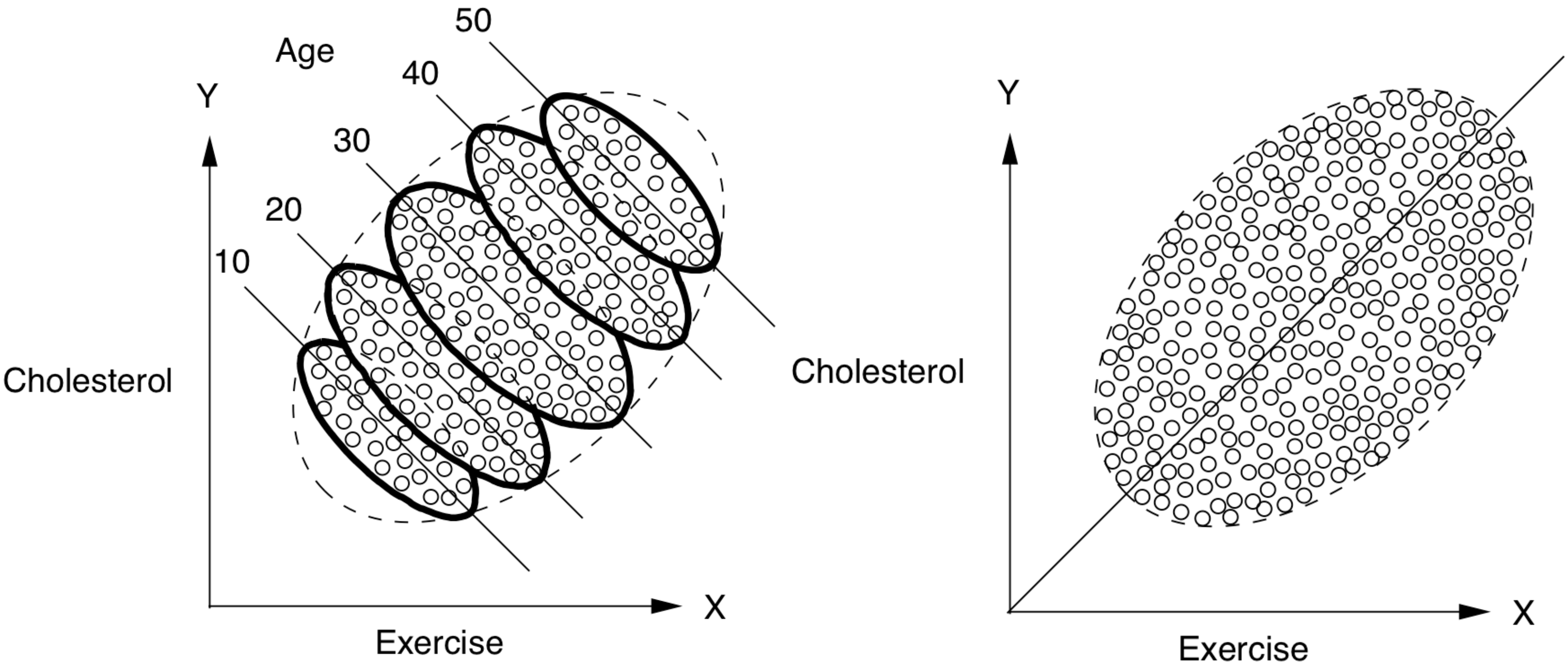
Exercise is helpful in every age group but harmful for a typical person. Why not?

WHY DATA CAN BE DUMB



Exercise is helpful in every age group but harmful for a typical person. Is exercise helpful or not?

WHY DATA CAN BE DUMB



Exercise is helpful in every age group but harmful for a typical person. Is exercise helpful or not? More specific? What about seatbelt usage?

EXPLAINABILITY
DEEP-LEARNING STYLE

Q. Why was my loan denied?
A. Because you are a female.

Q. What if I were a male?
A. It would be denied too.

Q. So who gets a loan?
A. Those who do not divulge their gender.

Q. But this does not make sense.
A. It explains WHY I made the decision.

ALGORITHMIC FAIRNESS DEEP-LEARNING STYLE

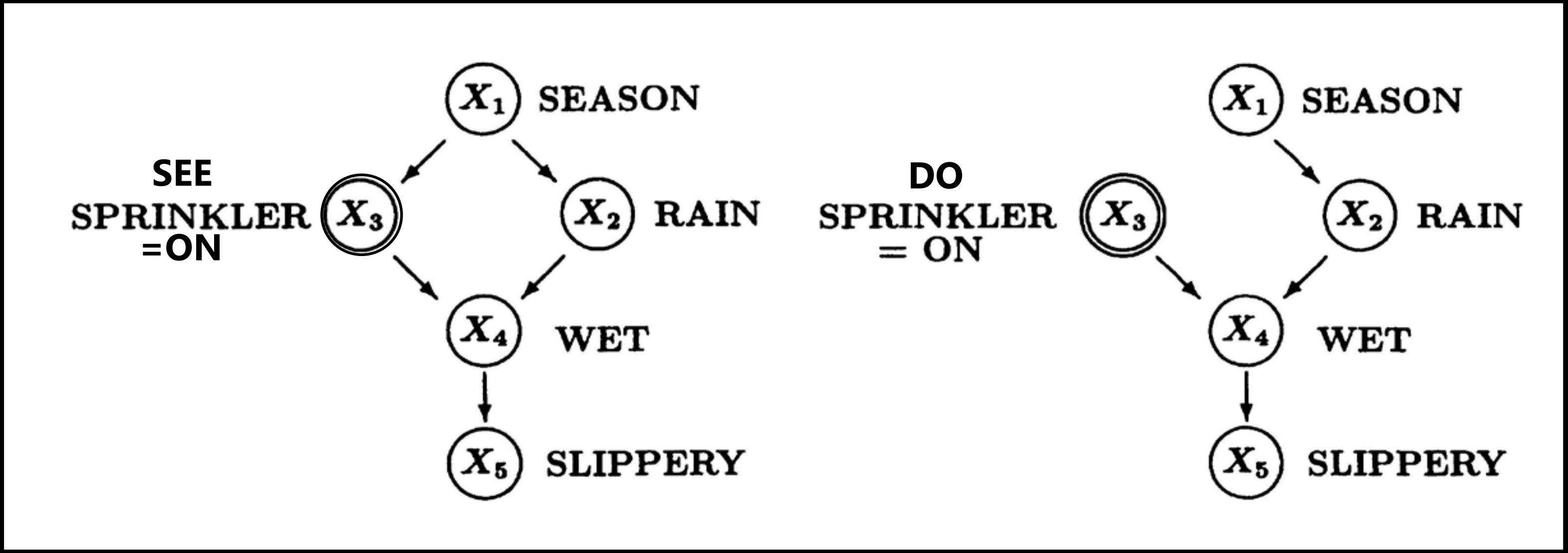
Q. Why was my insurance cancelled?
A. Because you had a traffic violation.

Q. What if I had no traffic violation?
A. It would have been cancelled too.

Q. So who gets insurance?
A. New drivers, with no record.

Q. This does not help safe driving.
A. It is at least "fair."

THE SECRET TO CAUSAL REASONING DISTINGUISH SEEING FROM DOING



What if we see the Sprinkler ON?

What if we turn the Sprinkler ON?

What if the Sprinkler were ON?

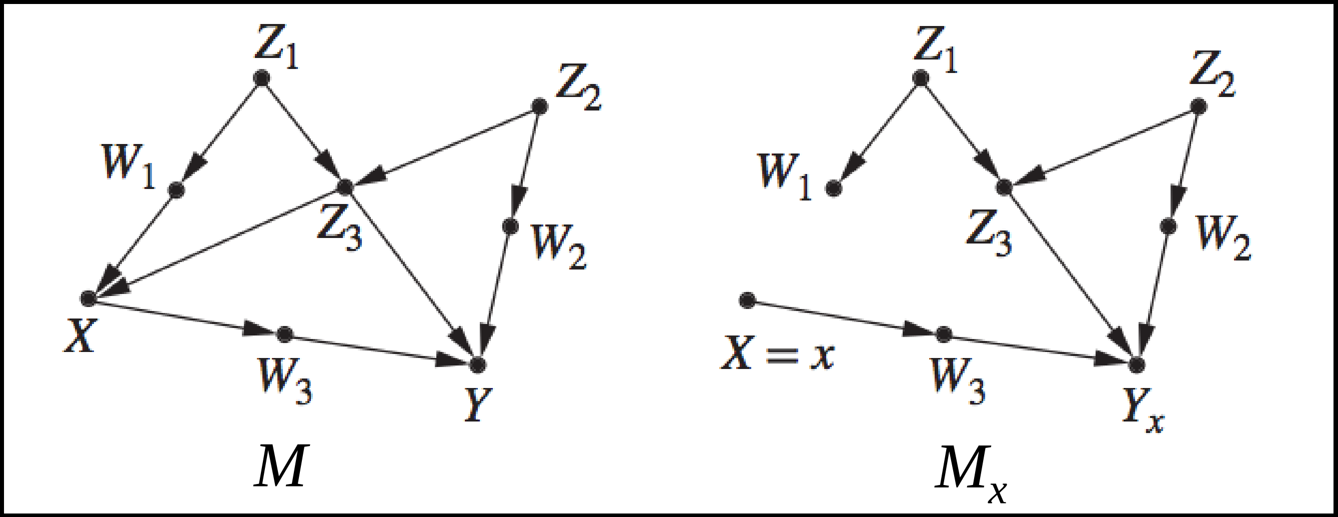
3 steps to counterfactuals

THE TWO FUNDAMENTAL LAWS OF CAUSAL INFERENCE

1. The Law of Counterfactuals (and Interventions)

$$\boxed{Y_x(u) = Y_{M_x}(u)}$$

(Y_x is equal to Y in a mutilated model M_x)



THE TWO FUNDAMENTAL LAWS OF CAUSAL INFERENCE

1. The Law of Counterfactuals (and Interventions)

$$\boxed{Y_x(u) = Y_{M_x}(u)}$$

(Y_x is equal to Y in a mutilated model M_x.)

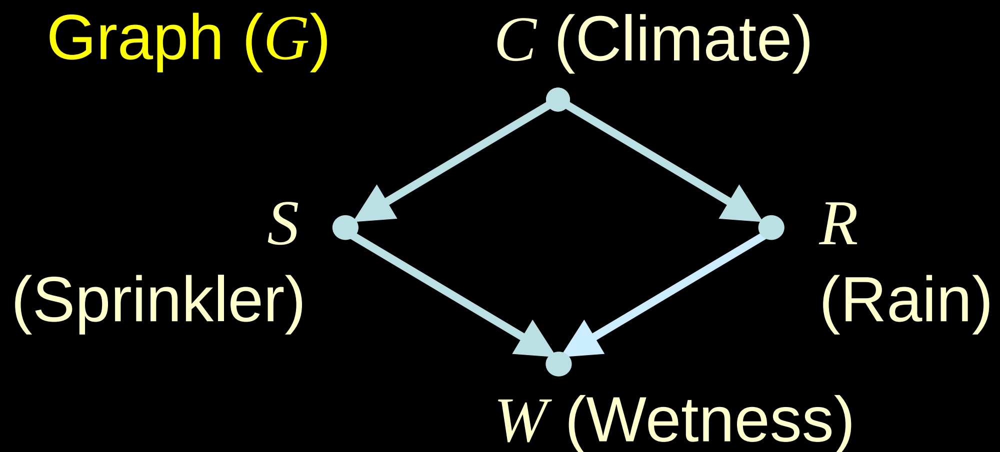
2. The Law of Conditional Independence (d-separation)

$$\boxed{(X \text{ sep } Y \mid Z)_{G(M)} \Rightarrow (X \perp\!\!\!\perp Y \mid Z)_{P(v)}}$$

(Separation in the model ⇒ independence in the distribution.)

READING INDEPENDENCIES

Graph (G)



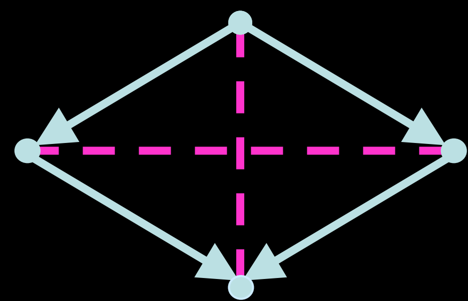
Model (M)

$$C = f_C(U_C)$$
$$S = f_S(C, U_S)$$
$$R = f_R(C, U_R)$$
$$W = f_W(S, R, U_W)$$

Every missing arrow advertises an independency, conditional on a separating set.

READING INDEPENDENCIES

Graph (G)



Model (M)

$$C = f_C(U_C)$$
$$S = f_S(C, U_S)$$
$$R = f_R(C, U_R)$$
$$W = f_W(S, R, U_W)$$

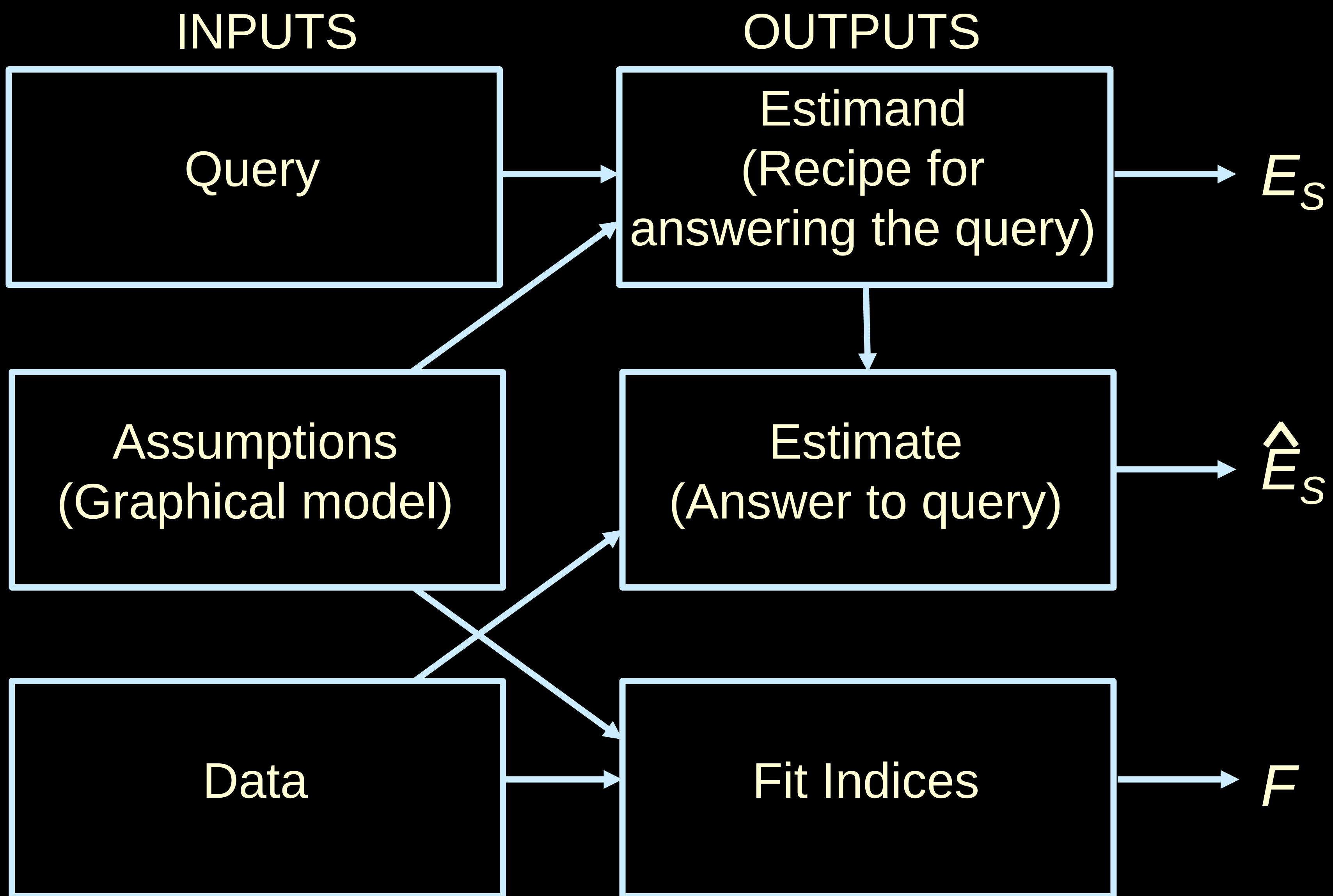
Every missing arrow advertises an independency, conditional on a separating set.

e.g., $C \perp\!\!\!\perp W \mid (S, R)$ $\qquad$ $S \perp\!\!\!\perp R \mid C$

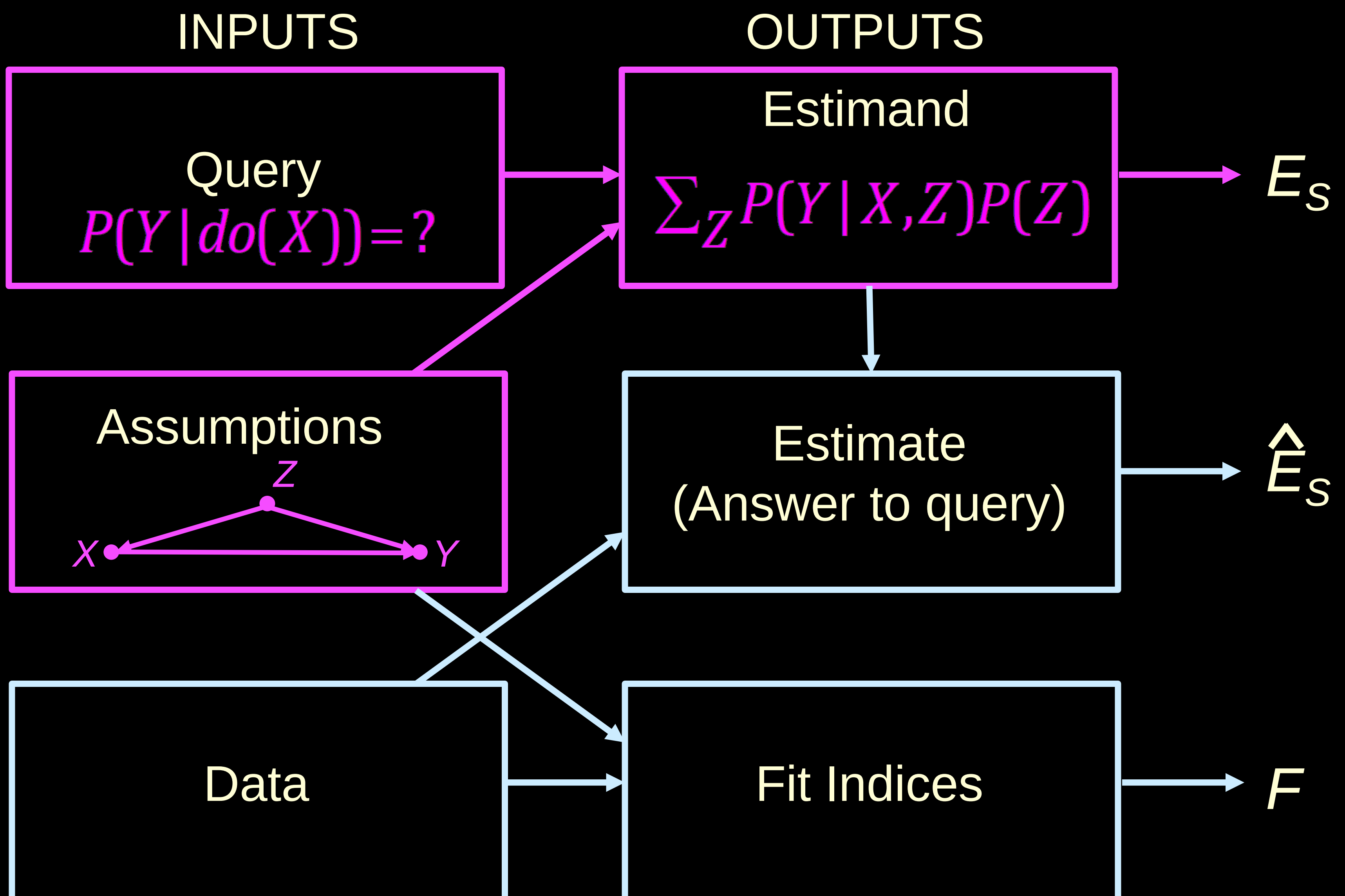
Applications:

1. Model testing
2. Structure learning
3. Reducing scientific questions to symbolic calculus

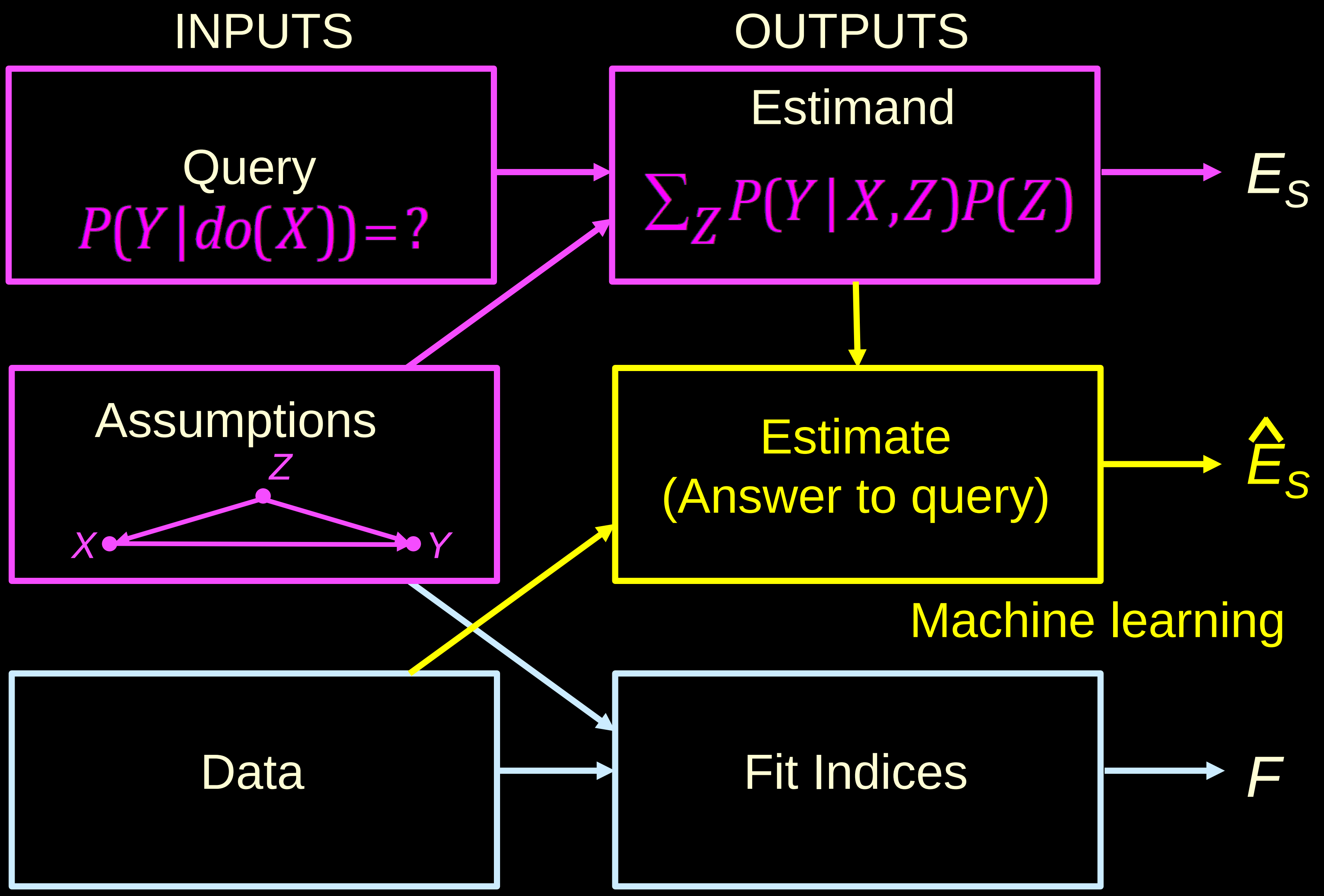
THE STRUCTURAL CAUSAL MODEL (SCM) INFERENCE ENGINE



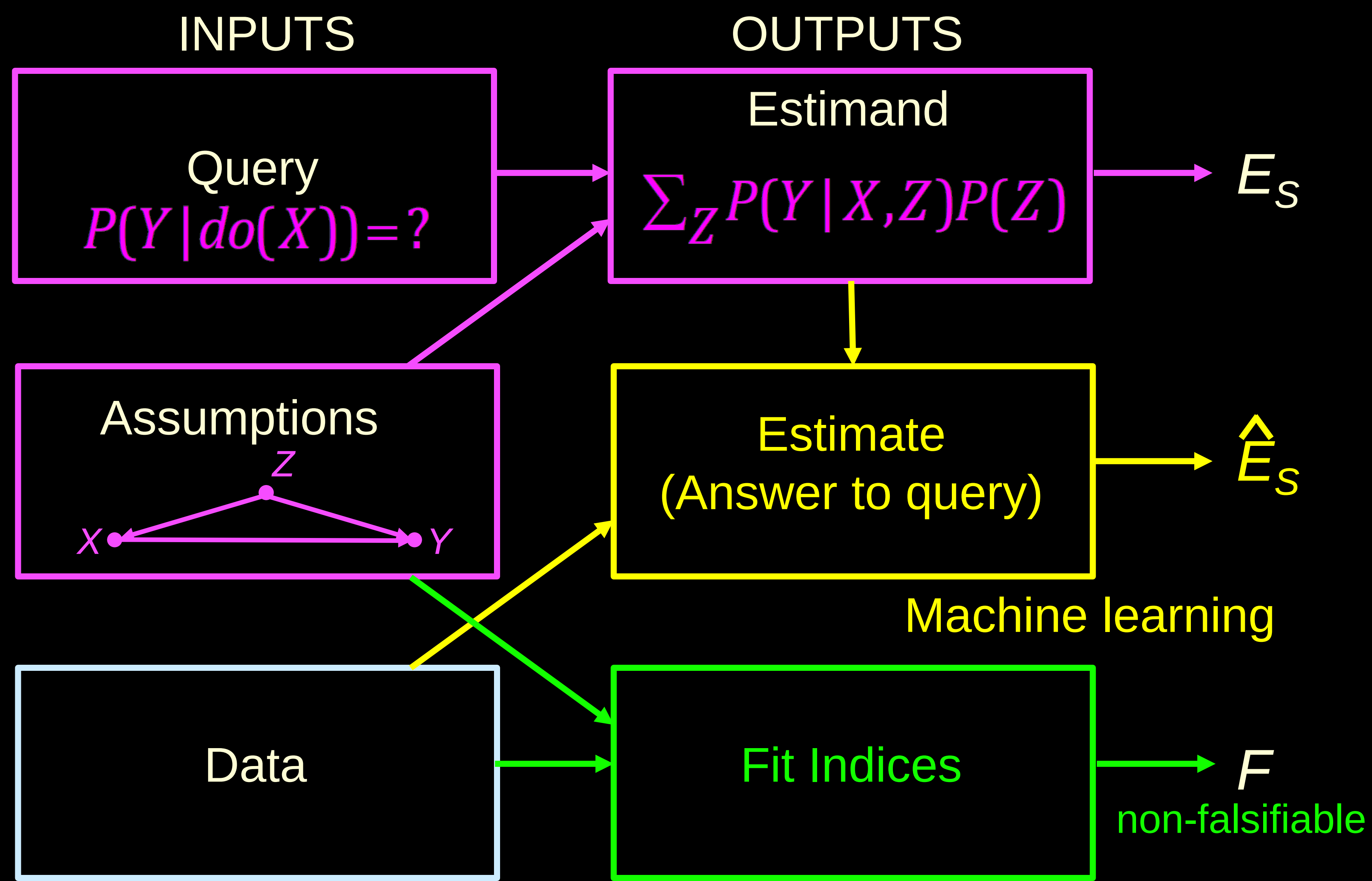
THE INFERENCE ENGINE IN ACTION



THE INFERENCE ENGINE IN ACTION



THE INFERENCE ENGINE IN ACTION



THE SEVEN PILLARS

Pillar 1: Transparency and Testability of Causal Assumptions
Pillar 2: Effect of Policies - Estimability
Pillar 3: Counterfactuals Algorithmitized (attribution, explanation, susceptibility)
Pillar 4: Direct and Indirect Effects (discrimination and inequities)
Pillar 5: External Validity and Sample Selection Bias
Pillar 6: Missing Data
Pillar 7: Causal Discovery

PILLAR 1:
MEANINGFUL COMPACT REPRESENTATION FOR CAUSAL ASSUMPTIONS

Task: Represent causal knowledge in compact, transparent, and testable way.

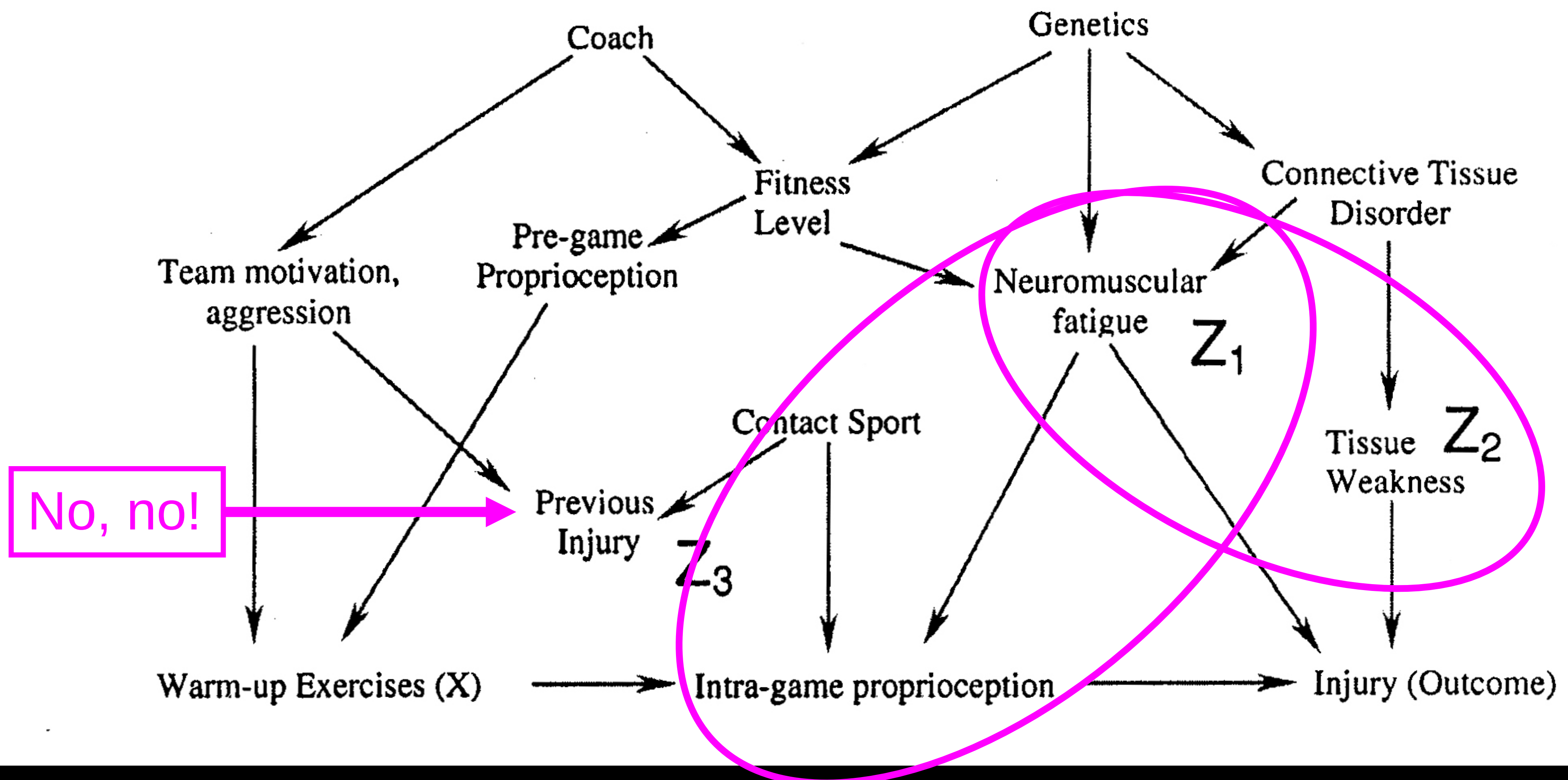
PILLAR 1:
MEANINGFUL COMPACT REPRESENTATION FOR CAUSAL ASSUMPTIONS

Task: Represent causal knowledge in compact, transparent, and testable way.

Result: Graphical models

- Graphs permit plausability checks over scientific knowledge.
- Graphical criteria tell us, for any pattern of paths, what pattern of dependencies hold in the data.
- Graphs compute for us the logical implications of our scientific assumptions.

EFFECT OF WARM-UP ON INJURY (After Shrier & Platt, 2008)



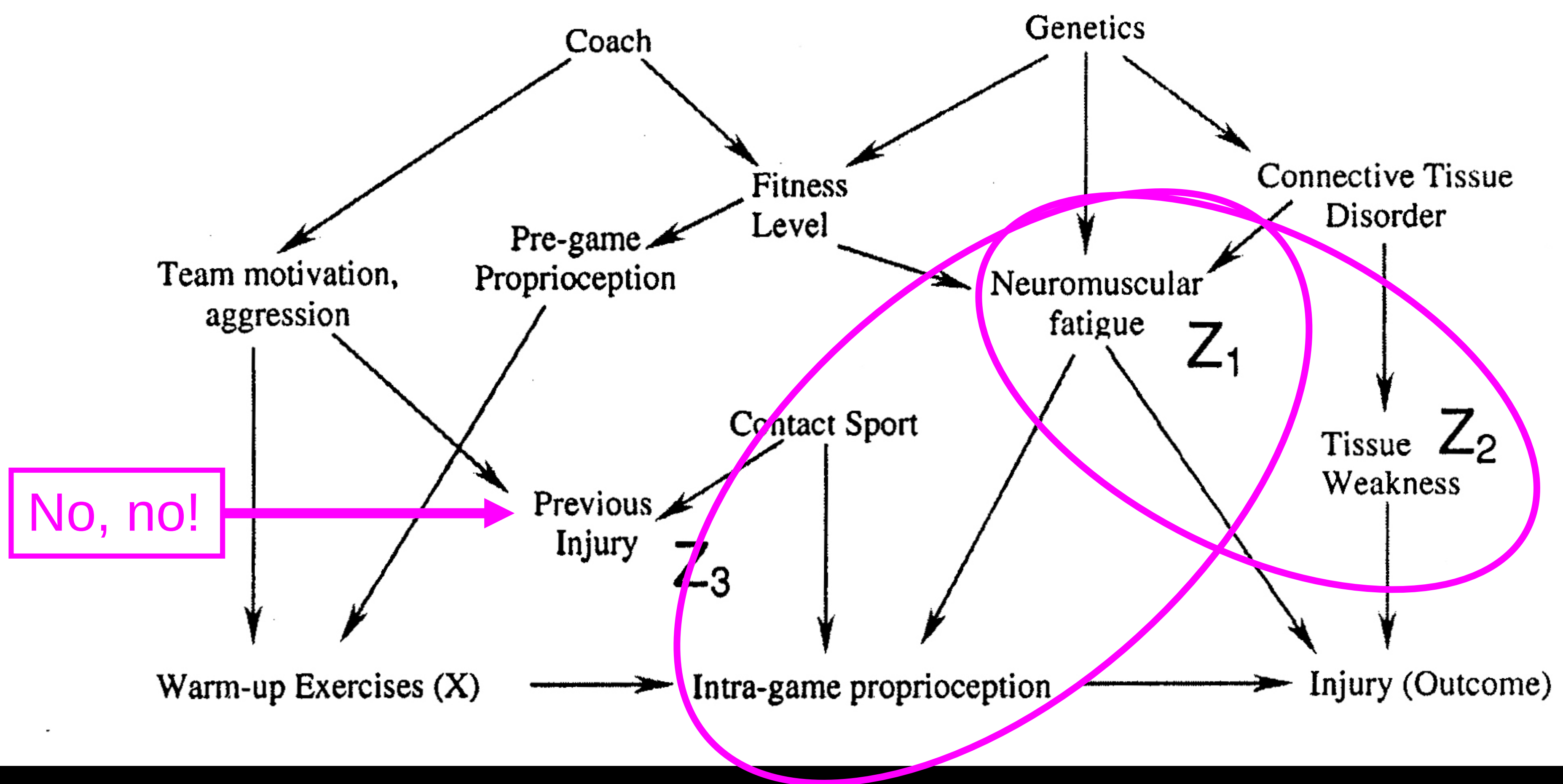
PILLAR 2:
EVALUATING EFFECTS OF NEW POLICIES

Problem: Determine if a *do*-expression can be estimated from data and how.

Solution: Reduced to a game-like calculus

- "back-door" – adjustment for covariates
- "front door" – extends it beyond adjustment
- *do-calculus* – predicts the effect of policy interventions whenever feasible

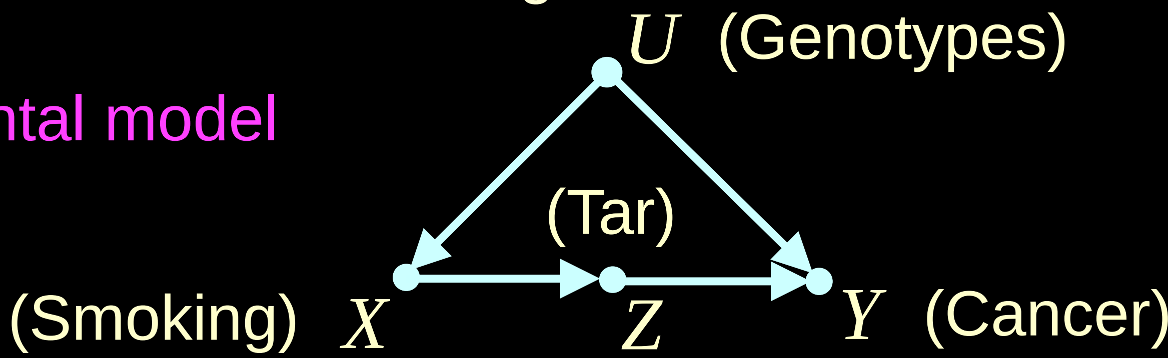
EFFECT OF WARM-UP ON INJURY (After Shrier & Platt, 2008)



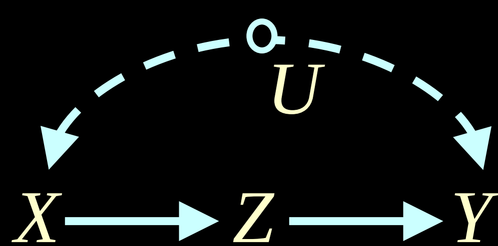
FORMULATING A PROBLEM IN THREE LANGUAGES

1. English: Given samples from $P(x, y, z)$
   Find: Effect of Smoking on Cancer

Mental model



2. Structural:
   Find: $P(Y = y \mid do(X = x))$



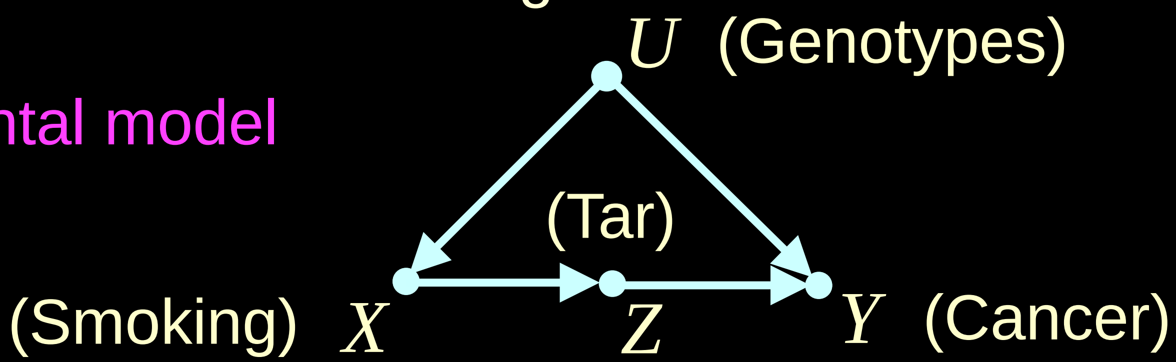
$$x = f_x(u, \varepsilon_x) \qquad y = f_y(z, u, \varepsilon_y)$$

$$z = f_z(x, \varepsilon_z) \qquad \varepsilon_z \perp\!\!\!\perp u, \varepsilon_x, \varepsilon_y$$

FORMULATING A PROBLEM IN THREE LANGUAGES

1. English: Given samples from $P(x, y, z)$
   Find: Effect of Smoking on Cancer

Mental model



3. Potential Outcome:
   Find: $P(Y_x = y)$

$$Z_x(u) = Z_{yx}(u),$$

$$X_y(u) = X_{zy}(u) = X_z(u) = X(u),$$

$$Y_z(u) = Y_{zx}(u), \quad Z_x \perp\!\!\!\perp \{Y_z, X\}$$

Not too friendly:

Consistent?, complete?, redundant?, plausible?, testable?

PILLAR 3:
THE ALGORITHMIZATION OF COUNTERFACTUALS

Task: Given {Model + Data}, determine what Joe's salary would be, had he had one more year of education.

Solution: The probability of every counterfactual can be computed or bounded using the "surgery" procedure.

Corollary: "Causes of effects" and "Attribution" formalized.

ATTRIBUTION

- Your Honor! My client (Mr. A) died BECAUSE he used this drug.

-

ATTRIBUTION

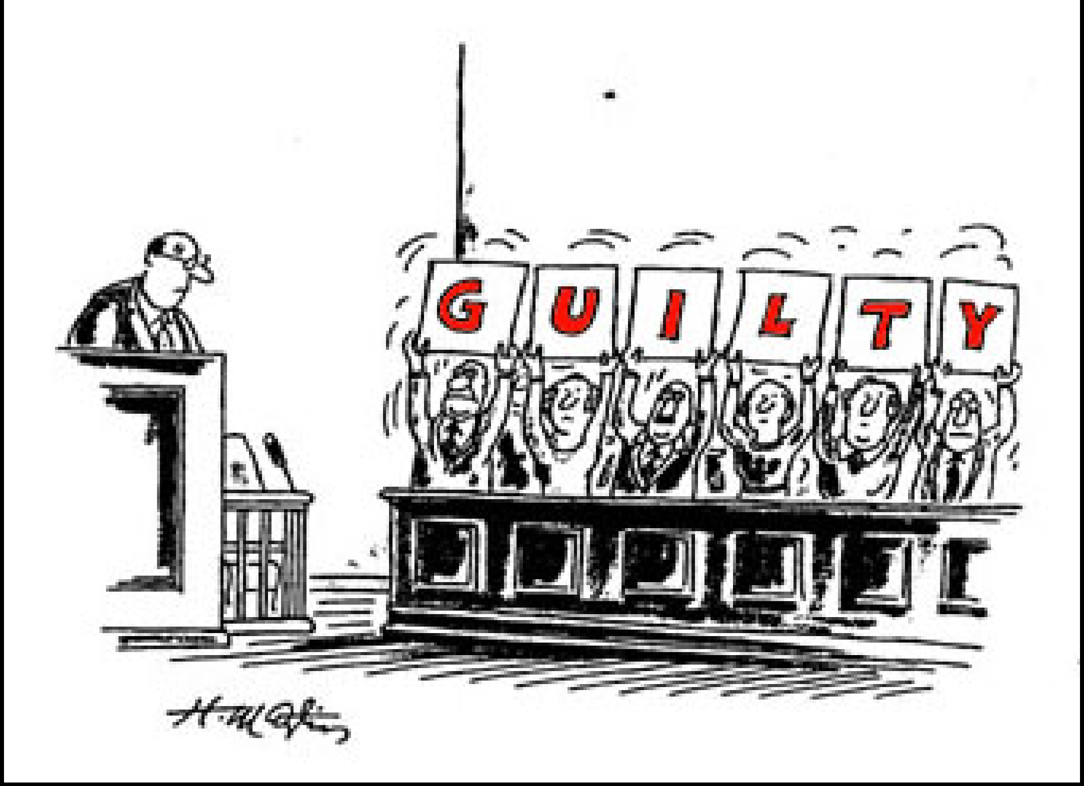
- Your Honor! My client (Mr. A) died BECAUSE he used this drug.

- Court to decide if it is MORE PROBABLE THAN NOT that Mr. A would be alive BUT FOR the drug!
- $PN = P(alive_{\{no\ drugs\}} \mid dead, drug) \geq 0.50$

CAN FREQUENCY DATA DETERMINE LIABILITY?

Sometimes:

When *PN* is bounded above 0.50.



- WITH PROBABILITY ONE $1 \leq PN \leq 1$
- Combined data tell more that each study alone

IDENTIFYING "SWING VOTERS"

Wikipedia: Voters that are uncommitted.
Counterfactual: Voters susceptible to persuasion.

PNS = Probability that a voter with characteristics c will vote yes IF AND ONLY IF enticed.

$$P(Y(1) = 1, Y(0) = 0 | C = c)$$

Derived (or bounded) from experimental and observational studies.

Only the gullible will be targeted.

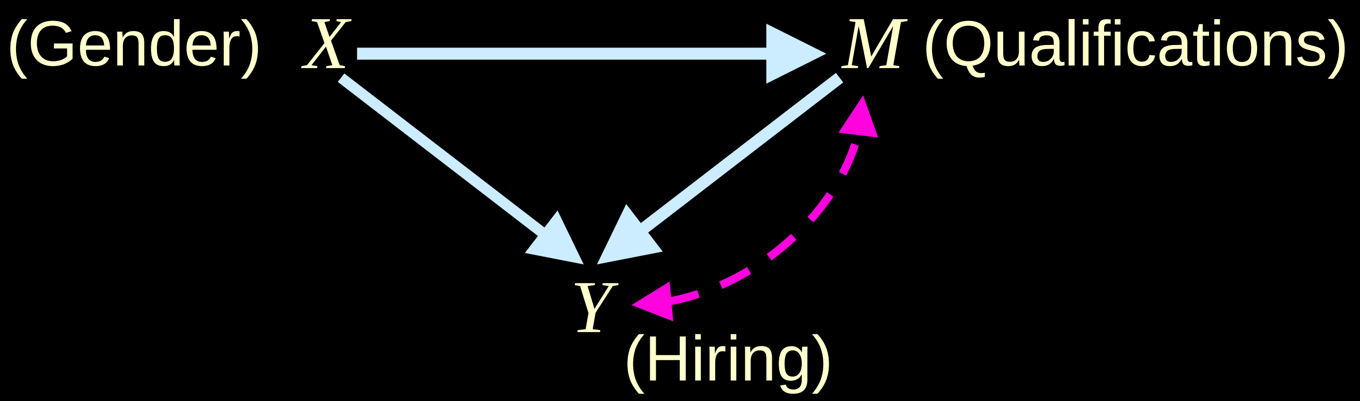
PILLAR 4:
MEDIATION ANALYSIS – DIRECT AND INDIRECT EFFECTS

Task: Given {Data + Model}, unveil and quantify the mechanisms that transmit changes from a cause to its effects.

Result: The graphical representation of counterfactuals tells us when direct and indirect effects are estimable from data, and, if so, how necessary (or sufficient) mediation is for the effect.

LEGAL IMPLICATIONS OF DIRECT EFFECT

Can data prove an employer guilty of hiring discrimination?



What is the direct effect of X on Y?

$$CDE = E(Y|do(x_1), do(m)) - E(Y|do(x_0), do(m))$$

(m-dependent) Adjust for M? No! No!

CDE Identification is completely solved

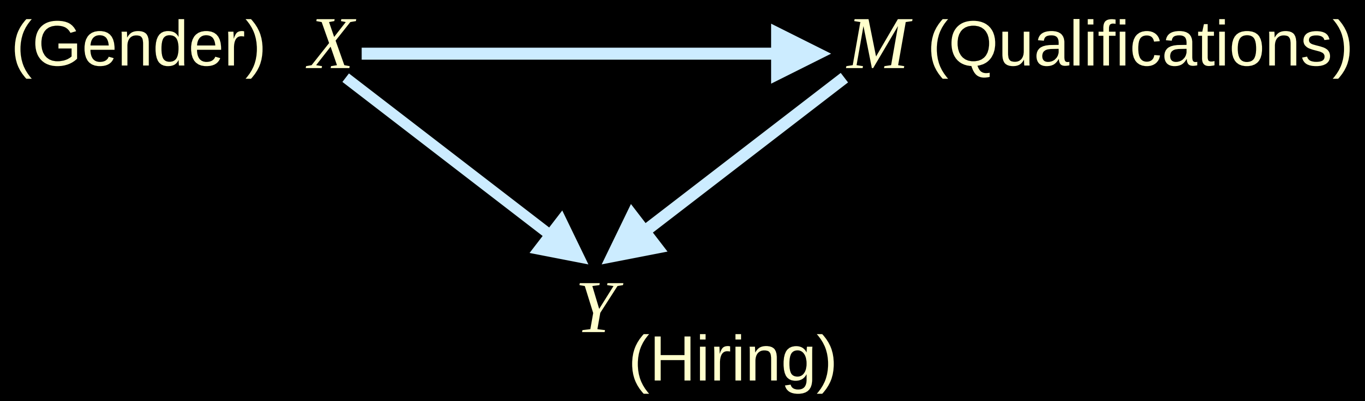
COUNTERFACTUAL DEFINITION OF DESCRIMINATION

"The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin, etc.) and everything else had been the same."

(In Carson vs Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996).)

LEGAL DEFINITION OF DISCRIMINATION

Can data prove an employer guilty of hiring discrimination?

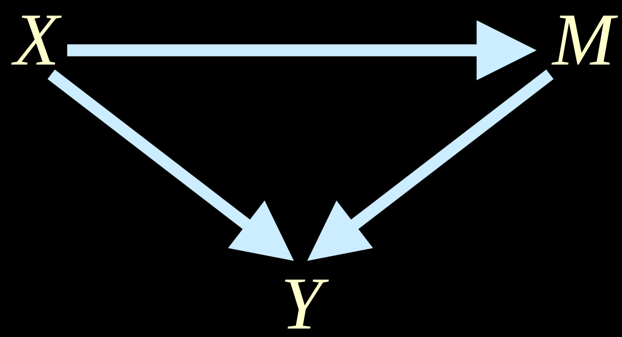


The Legal Definition:

Find the probability that "the employer would have acted differently had the employee been of different sex and qualification had been the same."

NATURAL INTERPRETATION OF AVERAGE DIRECT EFFECTS

Robins and Greenland (1992), Pearl (2001)



$$m = f(x, u)$$
$$y = g(x, m, u)$$

Natural Direct Effect of X on Y: $DE(x_0, x_1; Y)$

The expected change in Y, when we change X from x_0 to x_1 and, for each u, we keep M constant at whatever value it attained before the change.

$$E[Y_{x_1 M_{x_0}} - Y_{x_0}]$$

Note the nested counterfactuals

PILLAR 5: GENERALIZABILITY AND DATA FUSION

The problem

- How to combine results of several experimental and observational studies, each conducted on a different population and under a different set of conditions,
- so as to construct a valid estimate of effect size in yet a new population, unmatched by any of those studied.

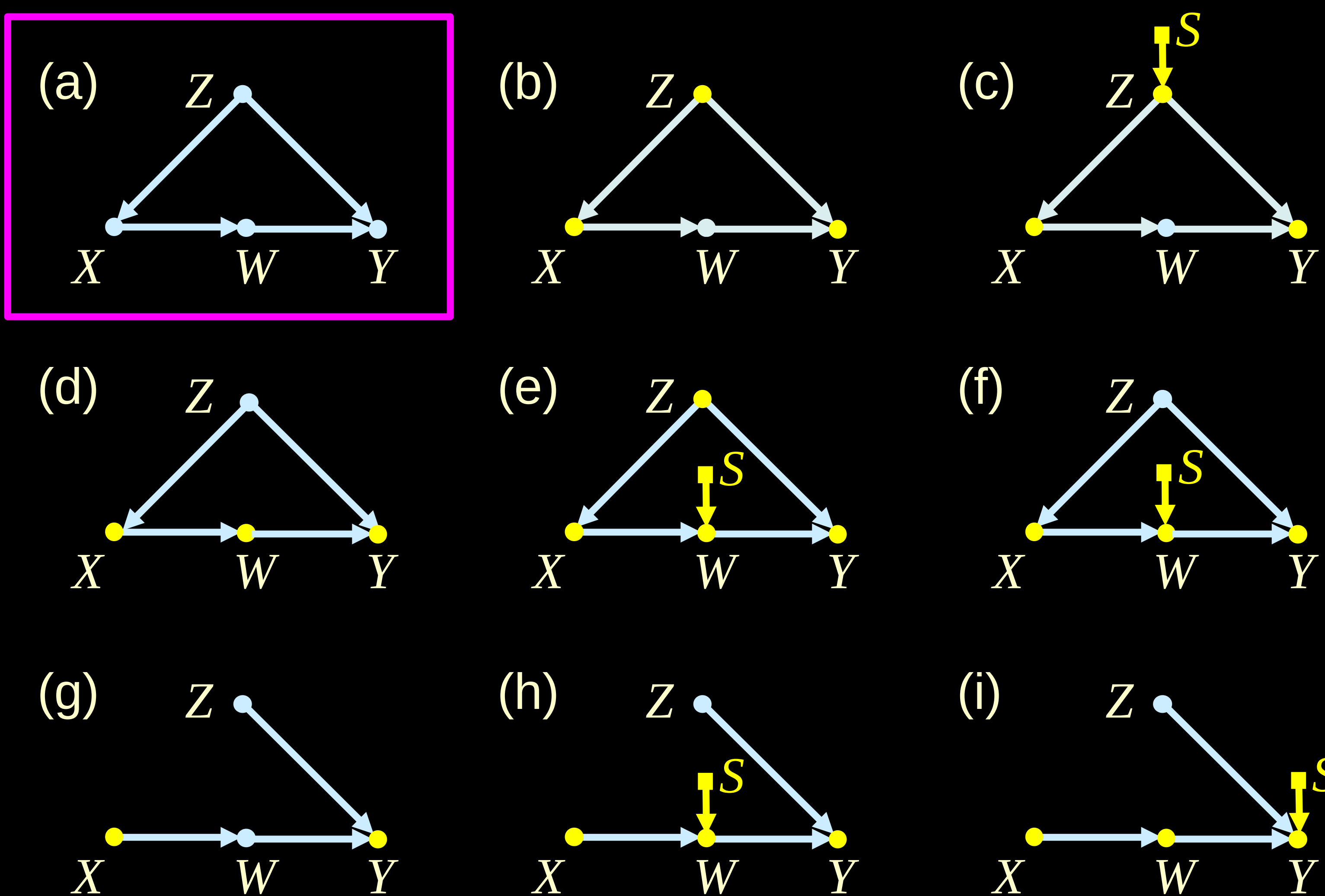
THE PROBLEM IN REAL LIFE

Target population Π^* Query of interest: $Q = P^*(y \mid do(x))$

(a) Arkansas Survey data available	(b) New York Survey data Resembling target	(c) Los Angeles Survey data Younger population
(d) Boston Age not recorded Mostly successful lawyers	(e) San Francisco High post-treatment blood pressure	(f) Texas Mostly Spanish subjects High attrition
(g) Toronto Randomized trial College students	(h) Utah RCT, paid volunteers, unemployed	(i) Wyoming RCT, young athletes

THE PROBLEM IN MATHEMATICS

Target population Π^* Query of interest: $Q = P^*(y \mid do(x))$



PILLAR 6: MISSING DATA (Mohan, 2017)

Problem: Given data corrupted by missing values and a model of what causes missingness. Determine when relations of interest can be estimated consistently "as if no data were missing."

Results: Graphical criteria unveil when estimability is possible, when it is not, and how.

Missing Data is a causal problem.

PILLAR 7:
CAUSAL DISCOVERY

Task: Search for a set of models (graphs) that are compatible with the data, and represent them compactly.

Results: In certain circumstances, and under weak assumptions, causal queries can be estimated directly from this compatibility set.

(Spirtes, Glymour and Scheines (2000); Jonas Peters etal (2018))

CONCLUSIONS

"More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history."

(Gary King, Harvard, 2014)

The peak of this revolution is still ahead of us (social intelligence, free-will, compassion).

UCLA has all the credentials to be its epi-center.

Paper available: http://ftp.cs.ucla.edu/pub/stat_ser/r475.pdf
Refs: http://bayes.cs.ucla.edu/jp_home.html

THANK YOU

Joint work with:
Elias Bareinboim
Karthika Mohan
Ilya Shpitser
Jin Tian
Many more . . .

Time for a short commercial

For a trailer, click WHY on my home page.