THE MATHEMATICS OF CAUSAL INFERENCE

With reflections on machine learning and the logic of science

Judea Pearl
UCLA

OUTLINE

1. The causal revolution – from statistics to counterfactuals – from Babylon to Athens
2. The fundamental laws of causal inference
3. From counterfactuals to problem solving

Old
- a) policy evaluation (ATE, ETT, …)
- b) attribution
- c) mediation

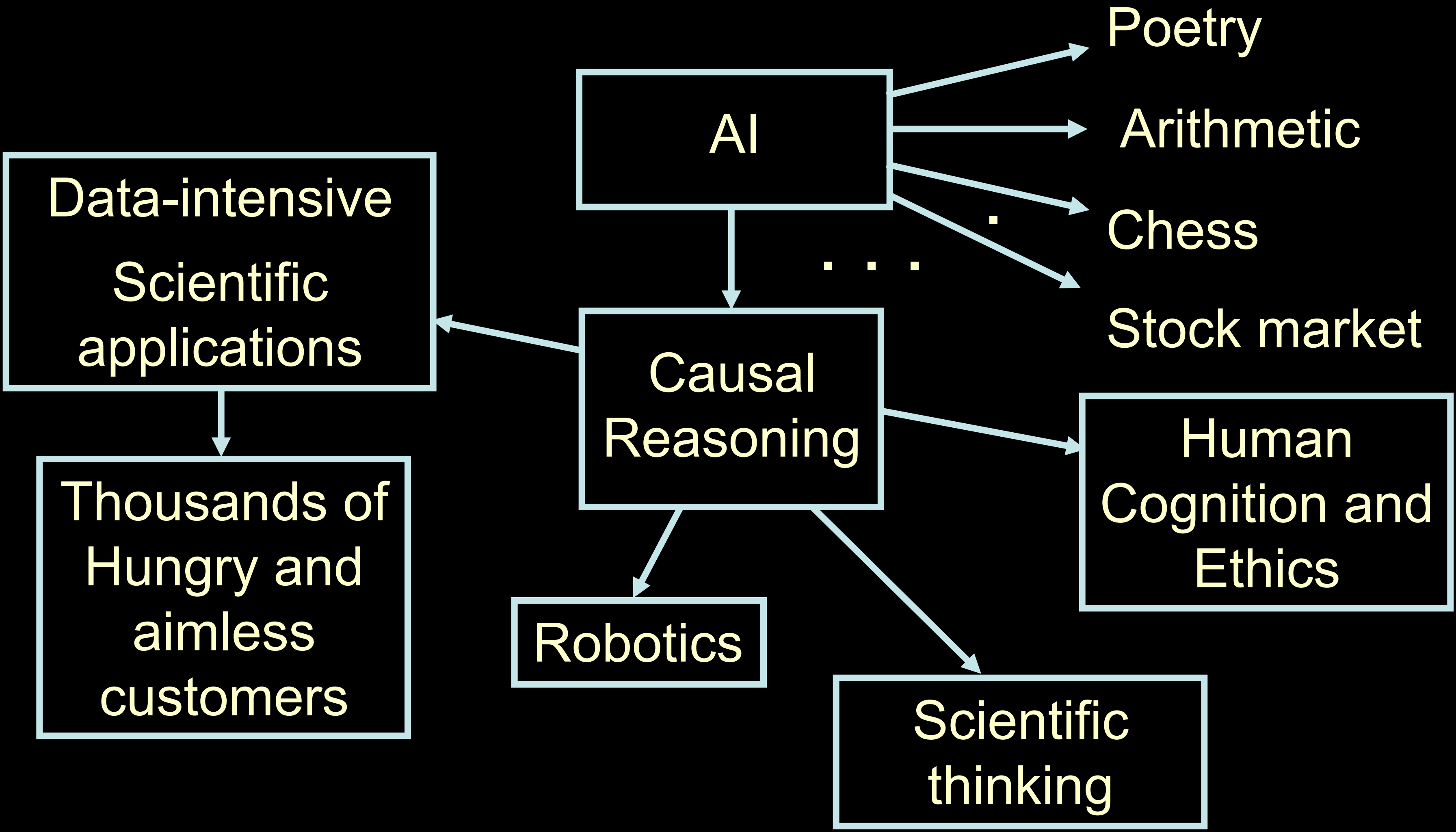
New
- d) generalizability – external validity
- e) latent heterogeneity
- f) missing data

TURING ON MACHINE LEARNING AND EVOLUTION

- The survival of the fittest is a slow method for measuring advantages.
- The experimenter, by exercise of intelligence, should be able to speed it up.
- If he can trace a cause for some weakness he can probably think of the kind of mutation which will improve it.

(A.M. Turing, 1950)

THE UBIQUITY OF CAUSAL REASONING



AI

Poetry

Arithmetic

Chess

Stock market

. . . .

Causal Reasoning

Data-intensive Scientific applications

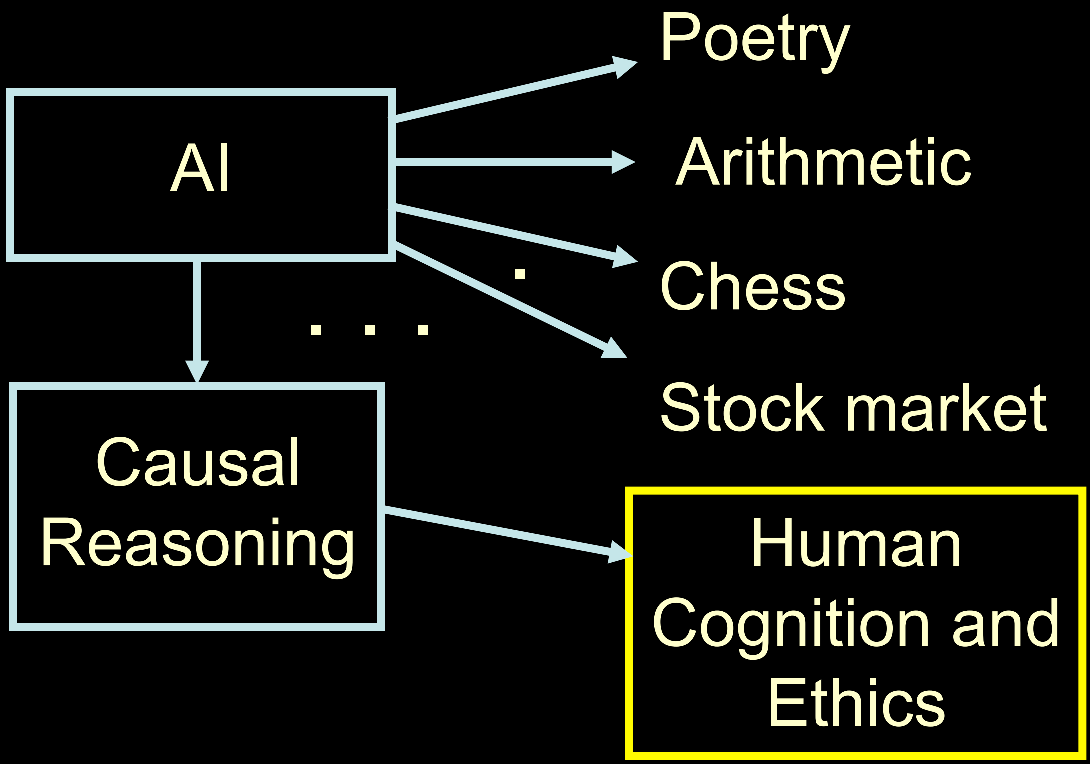
Thousands of Hungry and aimless customers

Robotics

Scientific thinking

Human Cognition and Ethics

THE UBIQUITY OF CAUSAL REASONING



AI

Poetry

Arithmetic

Chess

. . . .

Stock market

Causal Reasoning

Human Cognition and Ethics

Causal Explanation

"She handed me the fruit and I ate"

"The serpent deceived me, and I ate"

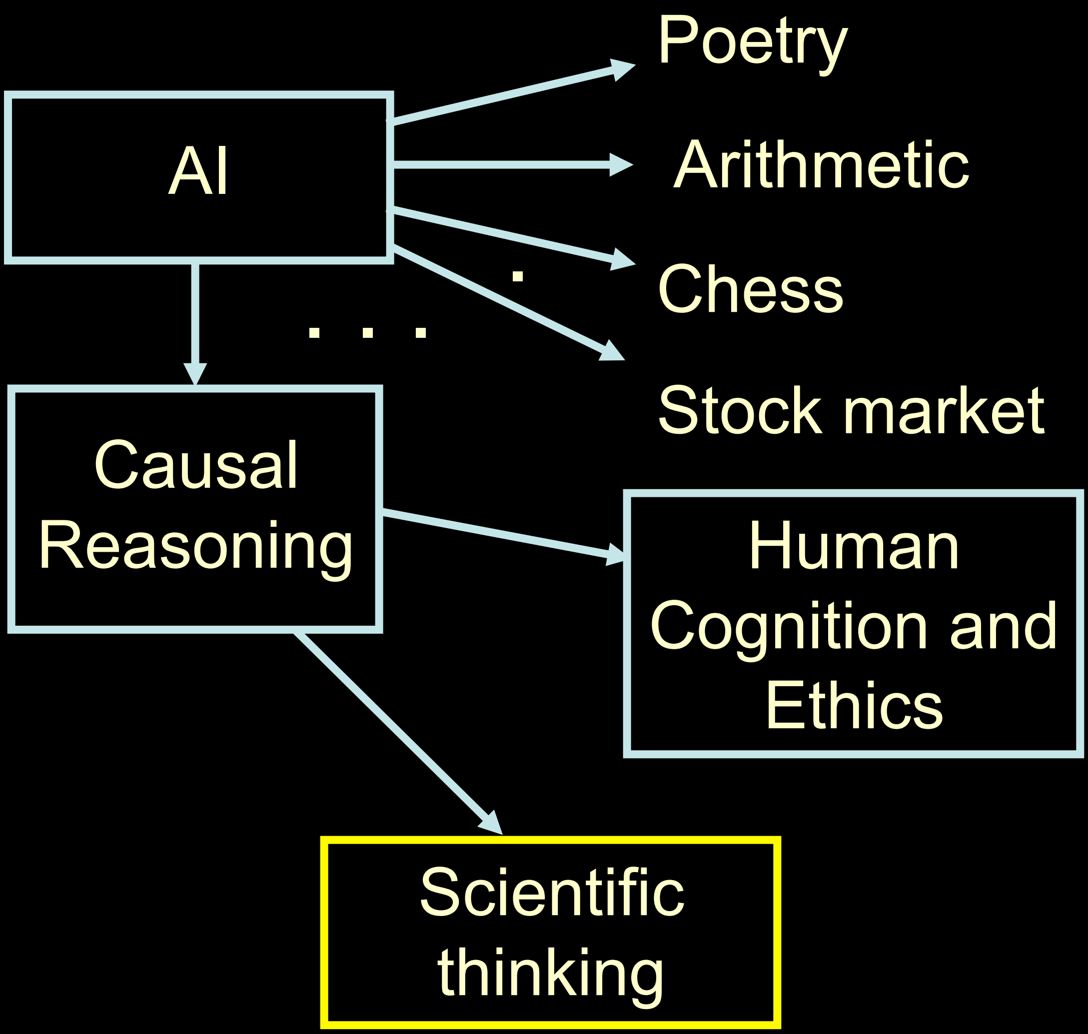
COUNTERFACTUALS AND OUR SENSE OF JUSTICE

Abraham:
Are you about to smite the righteous with the wicked?

What if there were fifty righteous men in the city?

And the Lord said,
"If I find in the city of Sodom fifty good men, I will pardon the whole place for their sake."
Genesis 18:26

THE UBIQUITY OF CAUSAL REASONING



AI

Poetry

Arithmetic

Chess

.

. . .

Stock market

Causal Reasoning

Human Cognition and Ethics

Scientific thinking

WHY PHYSICS IS COUNTERFACTUAL

Scientific Equations (e.g., Hooke's Law) are non-algebraic

e.g., Length (Y) equals a constant (2) times the weight (X)

Correct notation:

$Y = 2X$ $Y := 2X$
$X = 3$ $\cancel{X = 1}$
Process information

$X = 1$
$Y = 2$
The solution

$\cancel{X = \frac{1}{2} Y}$ $X = 3$
$Y = X+1$
Alternative

Had X been 3, Y would be 6.
If we raise X to 3, Y would be 6.
Must "wipe out" $X = 1$.

WHY PHYSICS IS COUNTERFACTUAL

Scientific Equations (e.g., Hooke's Law) are non-algebraic

e.g., Length (Y) equals a constant (2) times the weight (X)

Correct notation:

(or)

$Y \leftarrow 2X$

$X = 3$ $\cancel{X = 1}$

Process information

$X = 1$

$Y = 2$

The solution

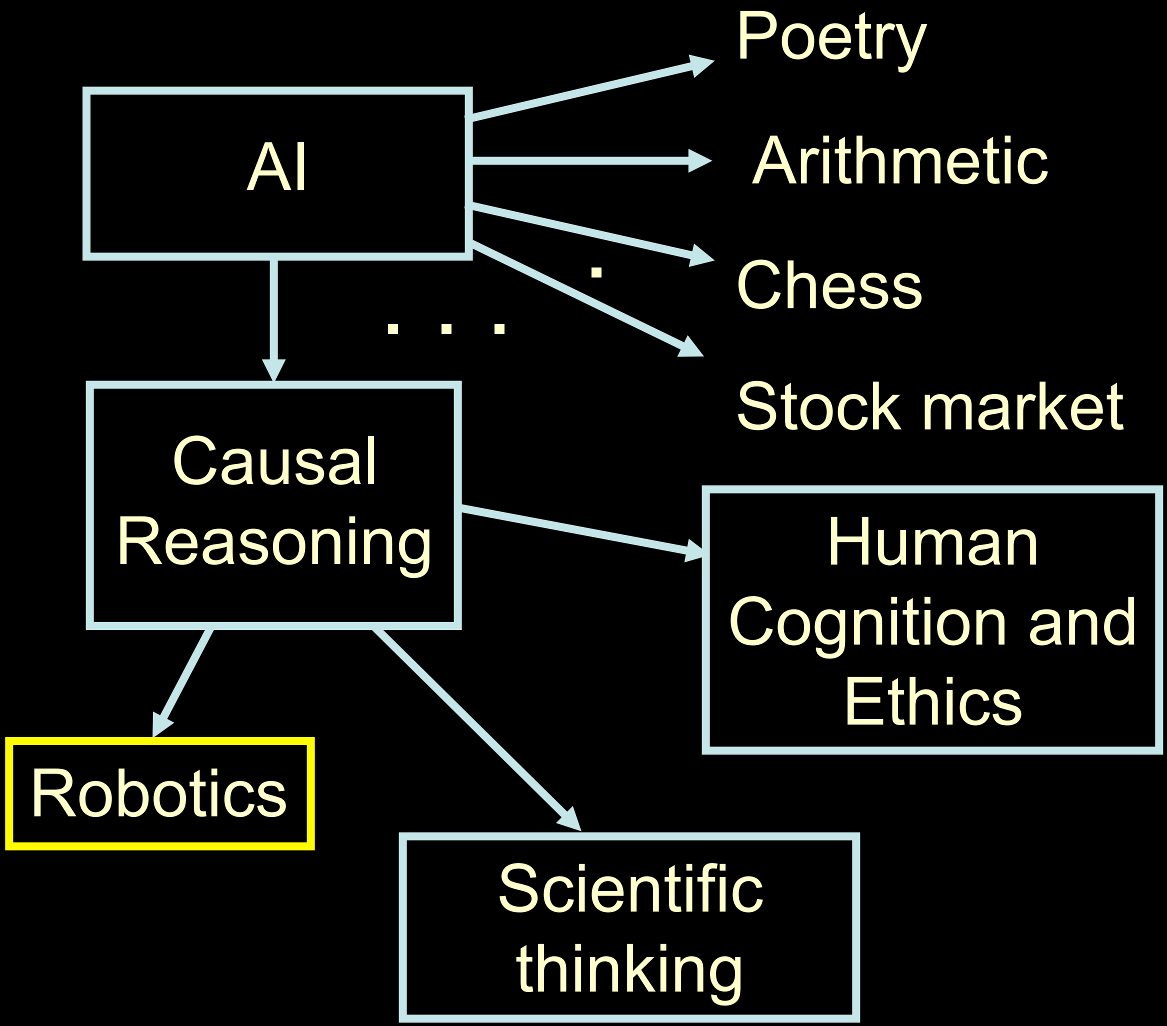
$\cancel{X = \frac{1}{2} Y}$ $X = 3$

$Y = X+1$

Alternative

Had X been 3, Y would be 6.
If we raise X to 3, Y would be 6.
Must "wipe out" $X = 1$.

THE UBIQUITY OF CAUSAL REASONING



AI

Poetry

Arithmetic

Chess

. . . .

Stock market

Causal Reasoning

Human Cognition and Ethics

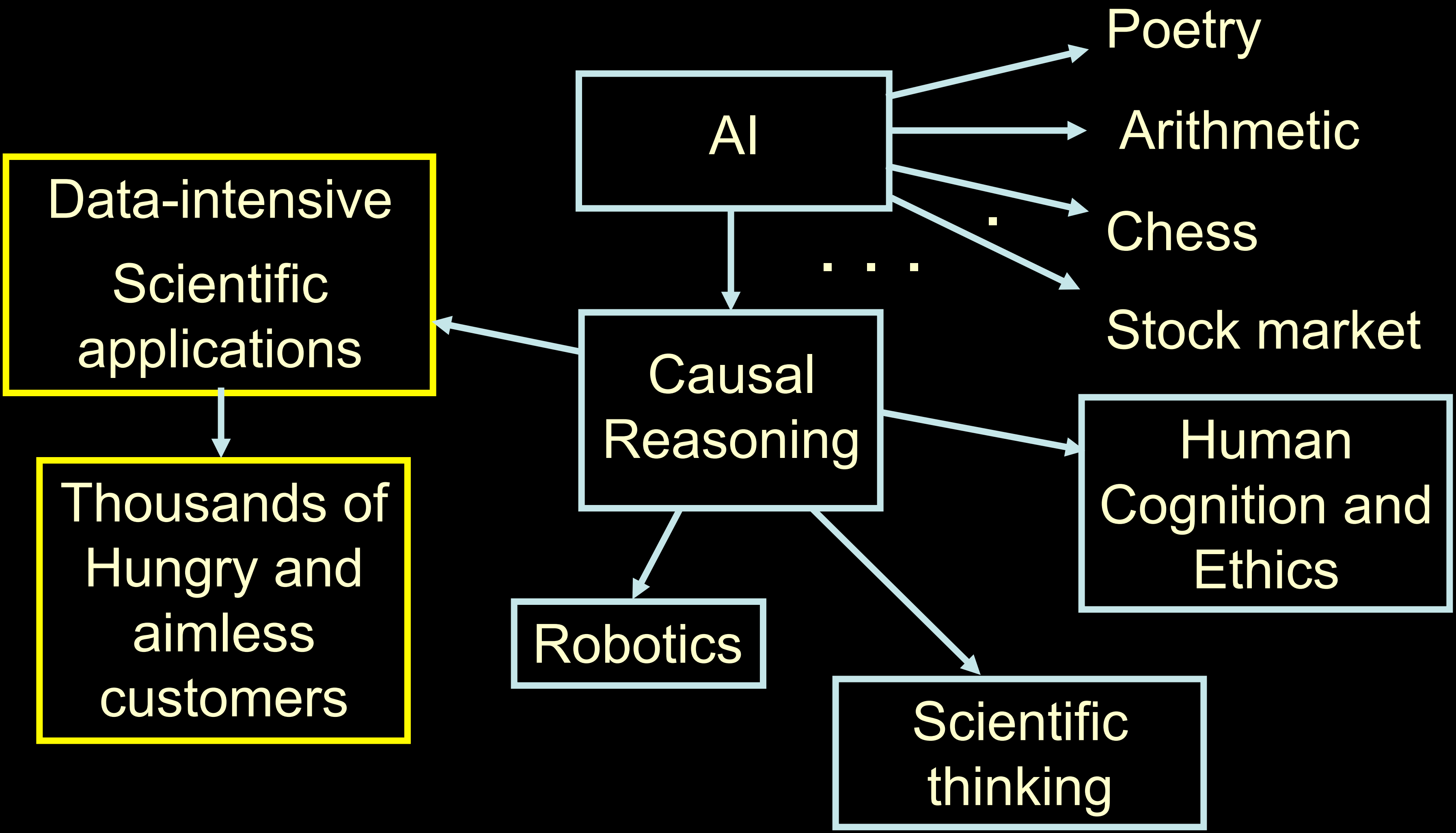
Robotics

Scientific thinking

WHAT KIND OF QUESTIONS SHOULD THE ROBOT ANSWER?

- Observational Questions:
  “What if I see A”
- Action Questions:
  “What if I do A?”
- Counterfactuals Questions:
  “What if I did things differently?”
- Options:
  “With what probability?”

THE UBIQUITY OF CAUSAL REASONING



AI

Poetry

Arithmetic

Chess

Stock market

. . . .

Causal Reasoning

Data-intensive Scientific applications

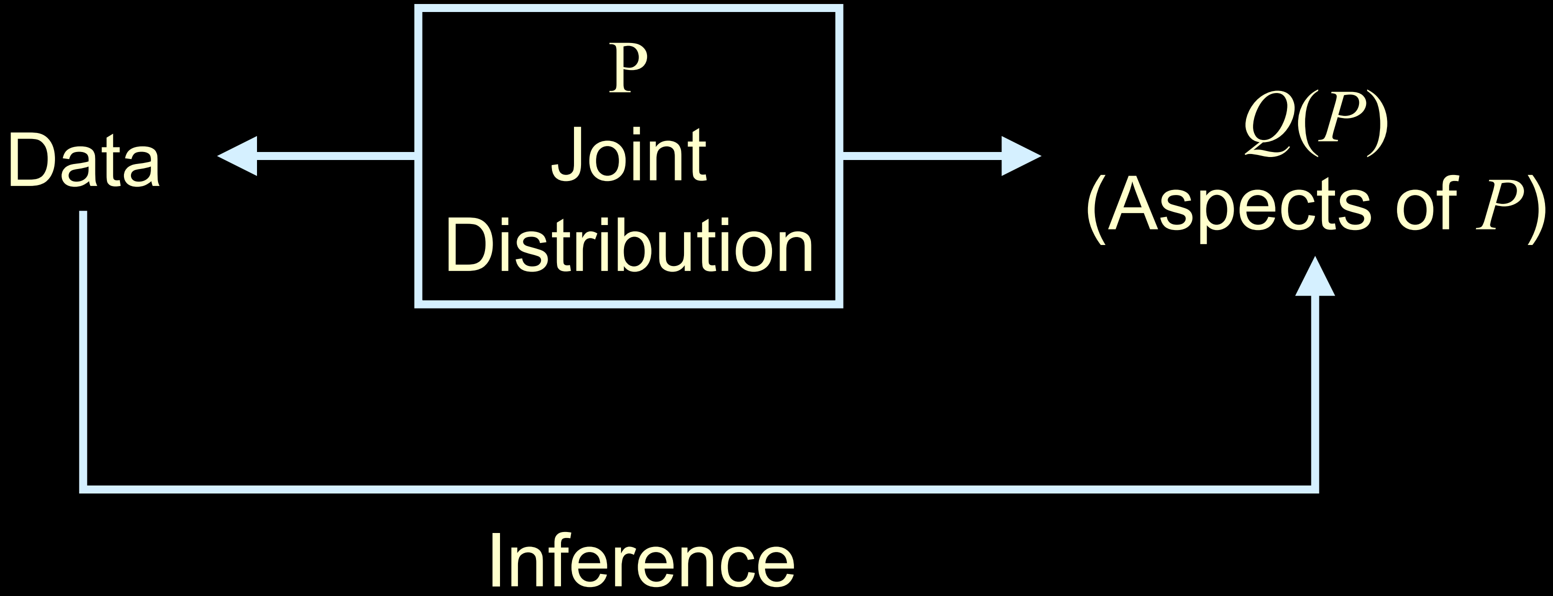
Thousands of Hungry and aimless customers

Robotics

Scientific thinking

Human Cognition and Ethics

TRADITIONAL STATISTICAL INFERENCE PARADIGM



Data

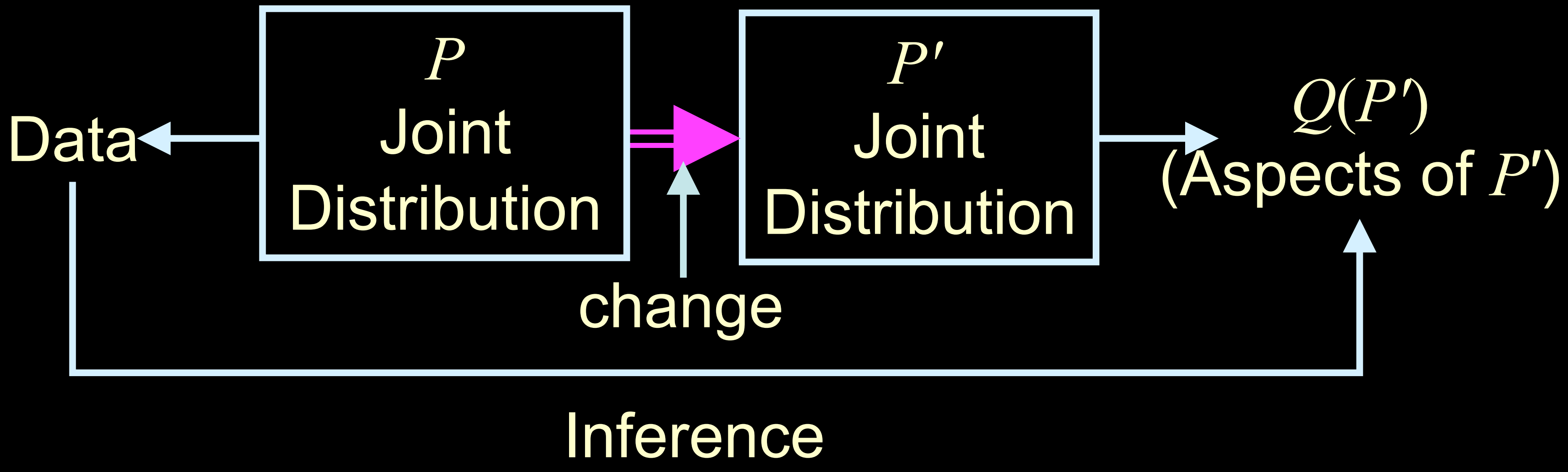
P
Joint
Distribution

$Q(P)$
(Aspects of P)

Inference

e.g.,
Infer whether customers who bought product A would also buy product B.
$Q = P(B \mid A)$

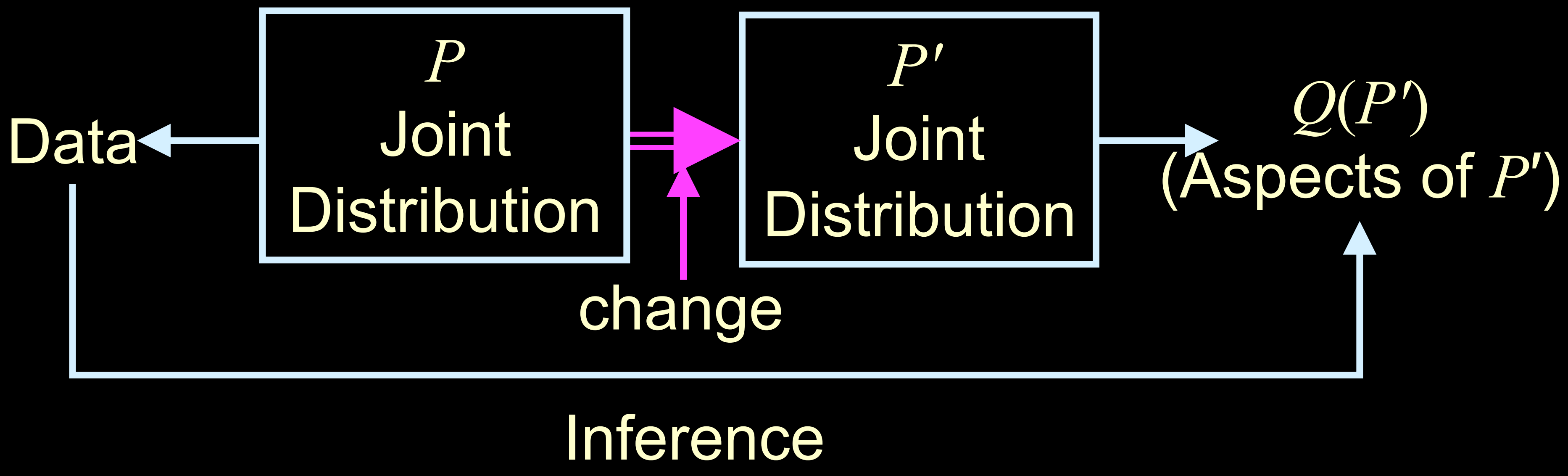
FROM STATISTICAL TO CAUSAL ANALYSIS:
1. THE DIFFERENCES



How does P change to P'? New oracle

e.g., Estimate P'(cancer) if we ban smoking.

FROM STATISTICAL TO CAUSAL ANALYSIS:
1. THE DIFFERENCES



Data

P

Joint Distribution

P'

Joint Distribution

$Q(P')$

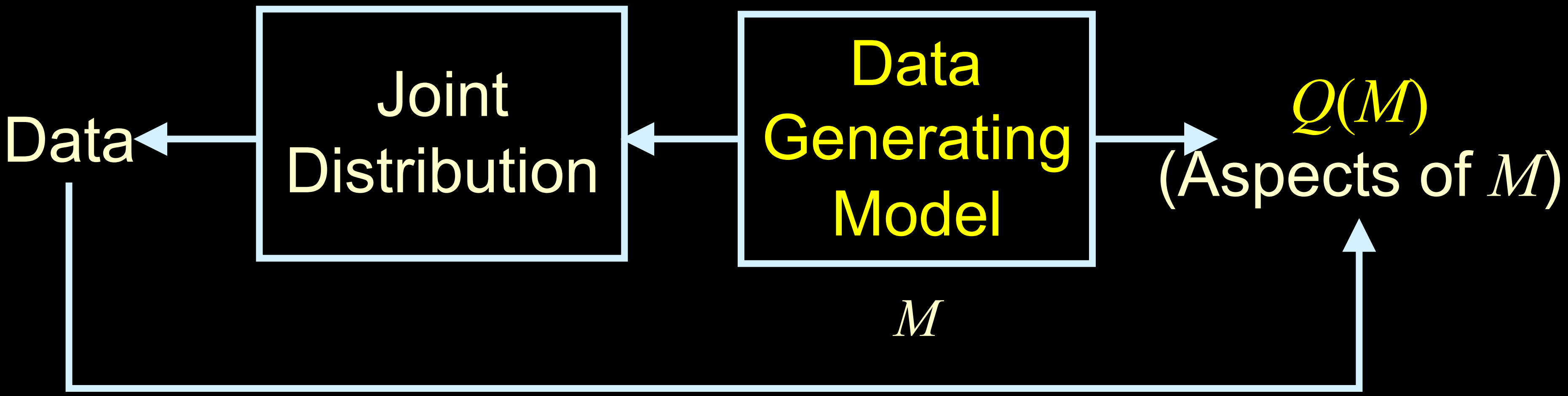
(Aspects of P')

change

Inference

e.g., Estimate the probability that a customer who bought A would buy B if we were to double the price.

THE STRUCTURAL MODEL PARADIGM



M – Invariant strategy (mechanism, recipe, law, protocol) by which Nature assigns values to variables in the analysis.

"A painful de-crowning of a beloved oracle!"

WHAT KIND OF QUESTIONS SHOULD THE ORACLE ANSWER?

- Observational Questions:
  "What if we see A" (What is?) $P(y \mid A)$
- Action Questions:
  "What if we do A?" (What if?) $P(y \mid do(A))$
- Counterfactuals Questions:
  "What if we did things differently?" (Why?) $P(y_{A'} \mid A)$
- Options:
  "With what probability?"

THE CAUSAL HIERARCHY - SYNTACTIC DISTINCTION

The fire
shadows cast on wall
Prisoners
Roadway where puppeteers perform

PLATO'S CAVE...

STRUCTURAL CAUSAL MODELS: THE WORLD AS A COLLECTION OF SPRINGS

Definition: A structural causal model is a 4-tuple $<V, U, F, P(u)>$, where

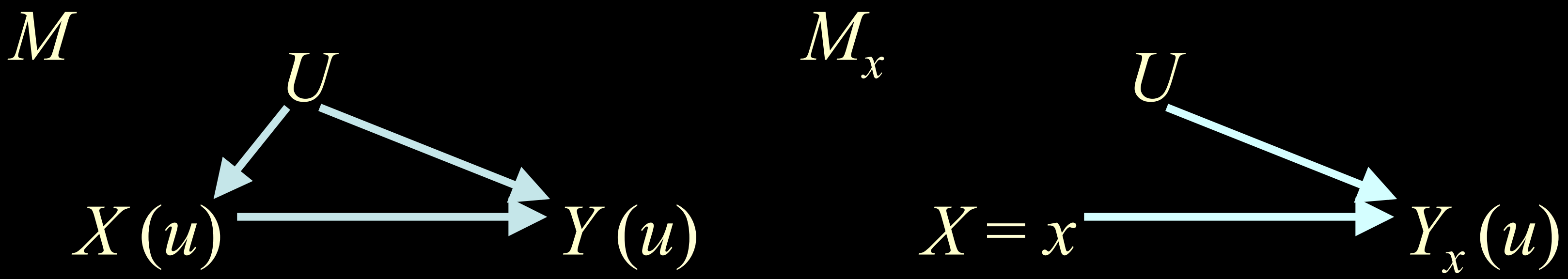
- $V = \{V_1, \ldots, V_n\}$ are endogenous variables
- $U = \{U_1, \ldots, U_m\}$ are background variables
- $F = \{f_1, \ldots, f_n\}$ are functions determining V,
  $v_i = f_i(v, u)$ e.g., $y = \alpha + \beta x + u_Y$ Not regression!!!!
- $P(u)$ is a distribution over U

$P(u)$ and F induce a distribution $P(v)$ over observable variables

COUNTERFACTUALS ARE EMBARRASSINGLY SIMPLE

Definition:

The sentence: "Y would be y (in situation u), had X been x," denoted $Y_x(u) = y$, means:
The solution for Y in a mutilated model M_x, (i.e., the equations for X replaced by $X = x$) with input $U=u$, is equal to y.



The Fundamental Equation of Counterfactuals:

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

THE TWO FUNDAMENTAL LAWS OF CAUSAL INFERENCE

1. The Law of Counterfactuals

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

(*M* generates and evaluates all counterfactuals.)

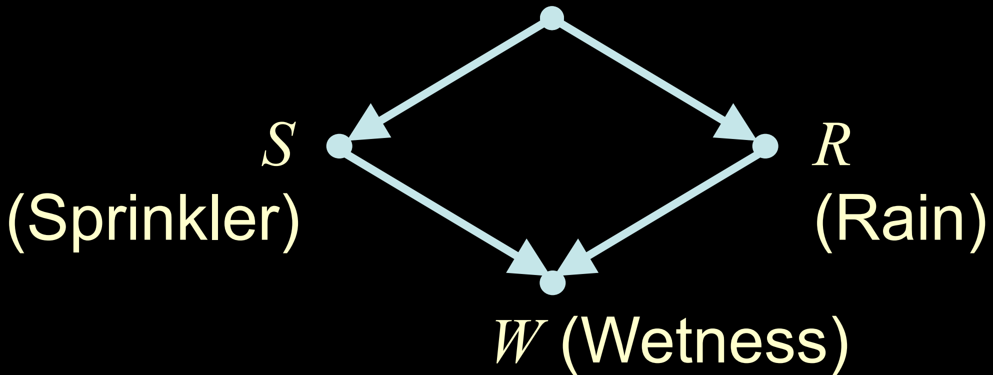
2. The Law of Conditional Independence (*d*-separation)

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

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

THE LAW OF CONDITIONAL INDEPENDENCE

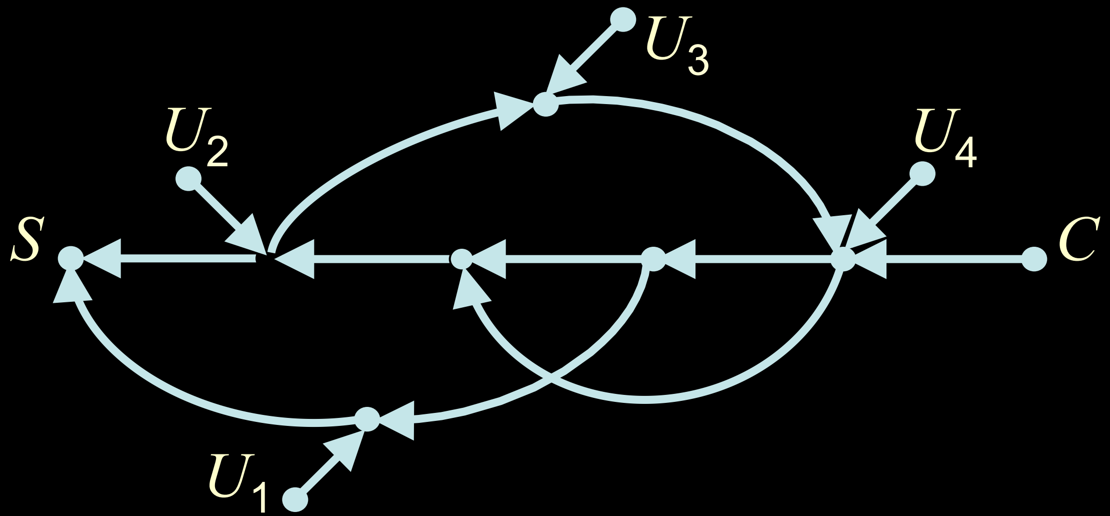
Graph (G)



Model (M)

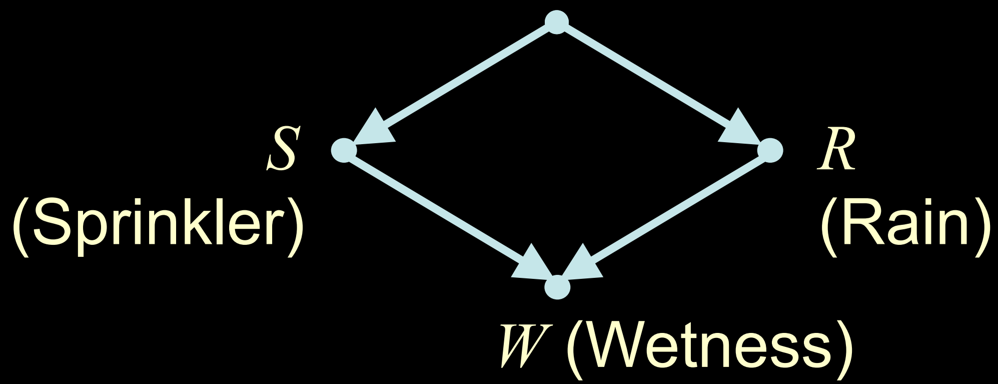
$$C = f_C(U_C)$$
$$S = f_S(C, U_S)$$
$$R = f_R(C, U_R)$$
$$W = f_W(S, R, U_W)$$

Each function summarizes millions of micro processes.



THE LAW OF CONDITIONAL INDEPENDENCE

Graph (G)



Model (M)

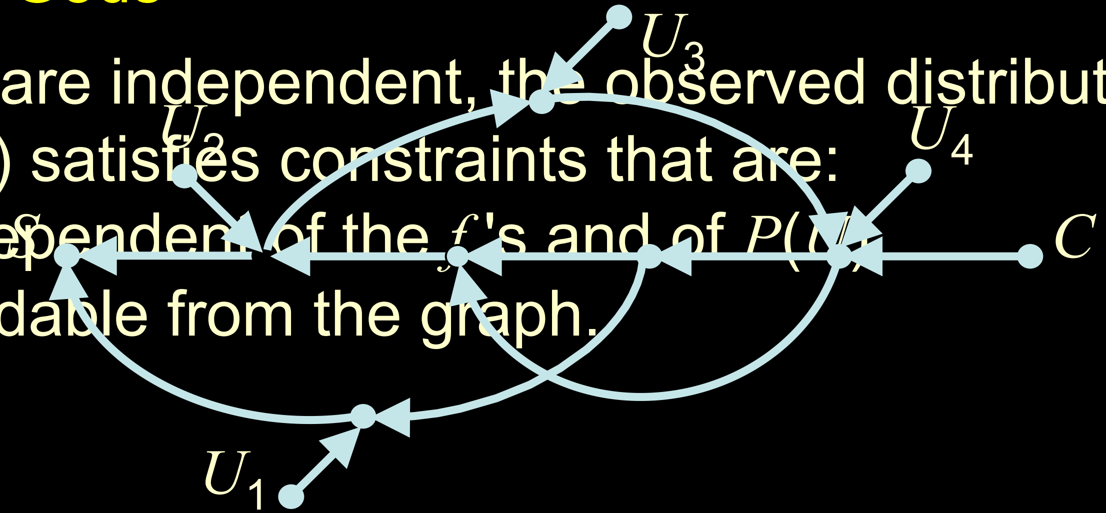
$$C = f_C(U_C)$$
$$S = f_S(C, U_S)$$
$$R = f_R(C, U_R)$$
$$W = f_W(S, R, U_W)$$

Gift of the Gods

If the U's are independent, the observed distribution $P(C,R,S,W)$ satisfies constraints that are:

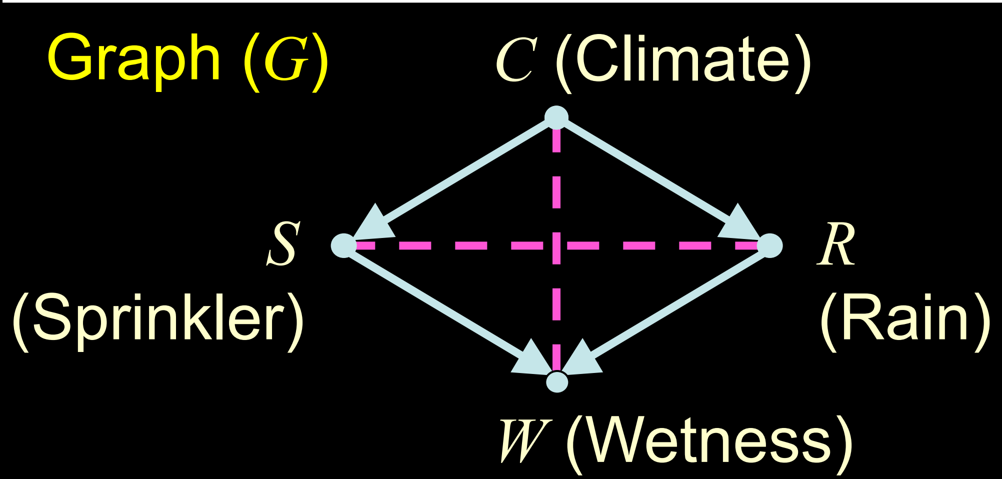
(1) independent of the f's and of $P(U)$,

(2) readable from the graph.



D-SEPARATION: NATURE'S LANGUAGE FOR COMMUNICATING ITS STRUCTURE

Graph (G)



Model (M)

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

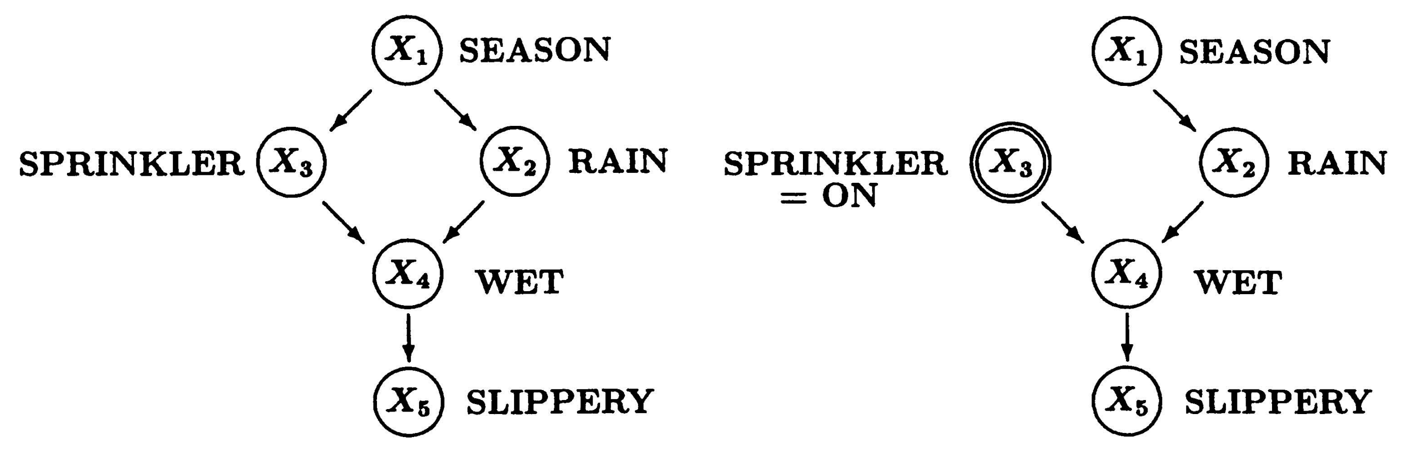
Every missing arrow advertises an independency, conditional on a separating set.

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

Applications:

1. Model testing
2. Structure learning
3. Reducing "what if I do" questions to symbolic calculus
4. Reducing scientific questions to symbolic calculus

SEEING VS. DOING



$$P(x_1,...,x_n) = \prod_i P(x_i \mid pa_i)$$

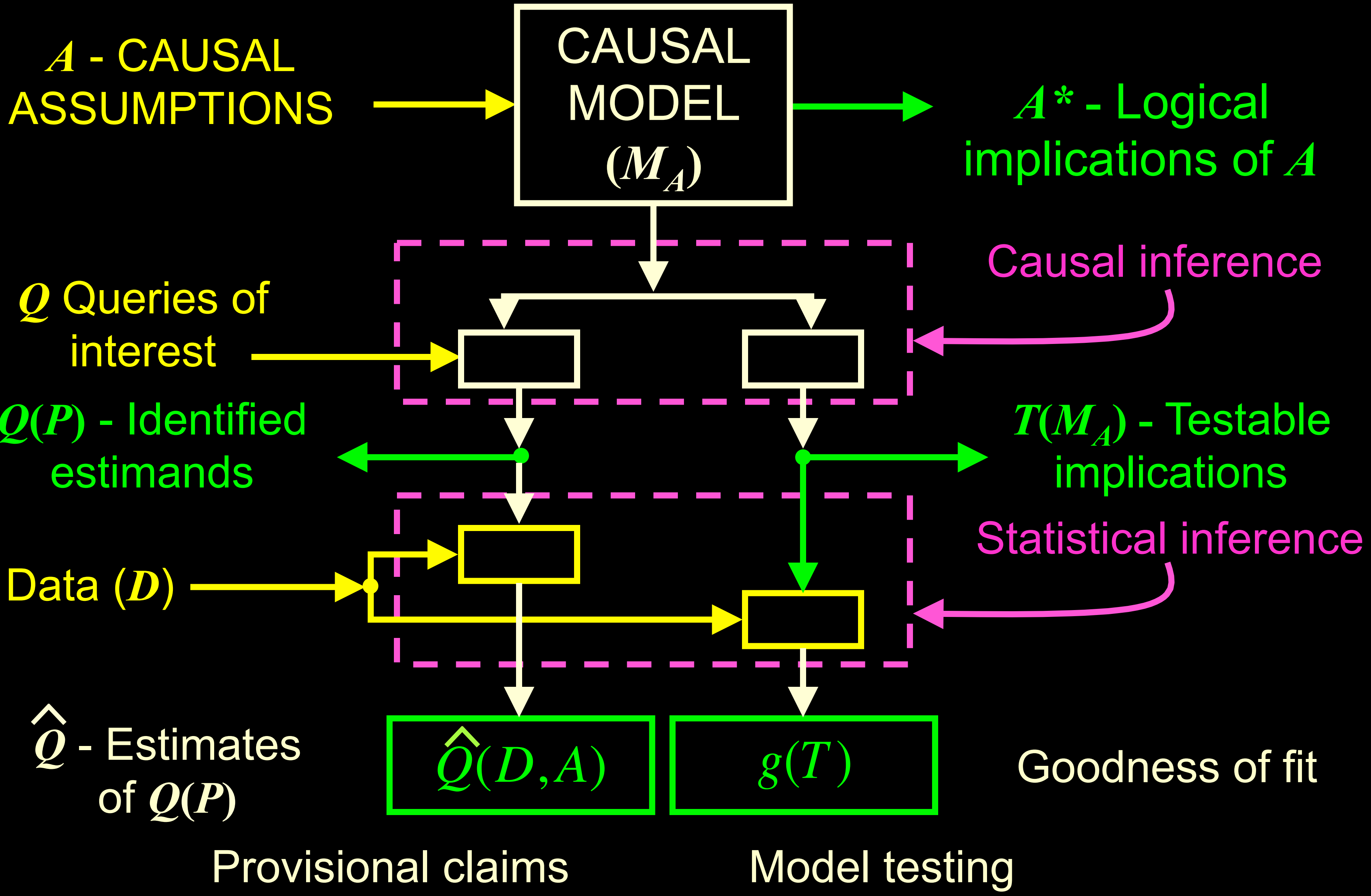
$$P(x_1,x_2,x_3,x_4,x_5) = P(x_1)P(x_2 \mid x_1)\xcancel{P(x_3 \mid x_1)}P(x_4 \mid x_2,x_3)P(x_5 \mid x_4)$$

Effect of turning the sprinkler ON (Truncated product)

$$P_{X_3=\text{ON}}(x_1,x_2,x_4,x_5) = P(x_1)P(x_2 \mid x_1)P(x_4 \mid x_2,X_3=\text{ON})P(x_5 \mid x_4)$$

$$\neq P(x_1,x_2,x_4,X_5 \mid X_3=\text{ON})$$

THE LOGIC OF CAUSAL ANALYSIS



A - CAUSAL ASSUMPTIONS

CAUSAL MODEL (M_A)

*A** - Logical implications of *A*

Causal inference

Q Queries of interest

Q(*P*) - Identified estimands

T(M_A) - Testable implications

Statistical inference

Data (*D*)

$\hat{Q}$ - Estimates of *Q*(*P*)

$\hat{Q}(D,A)$

$g(T)$

Goodness of fit

Provisional claims

Model testing

THE MACHINERY OF CAUSAL CALCULUS

Rule 1: Ignoring observations

P(y | do{x}, z, w) = P(y | do{x}, w)

if $(Y \perp\!\!\!\perp Z / X,W)_{G_{\overline{X}}}$

Rule 2: Action/observation exchange

P(y | do{x}, do{z}, w) = P(y | do{x},z,w)

if $(Y \perp\!\!\!\perp Z / X,W)_{G_{\overline{X}\underline{Z}}}$

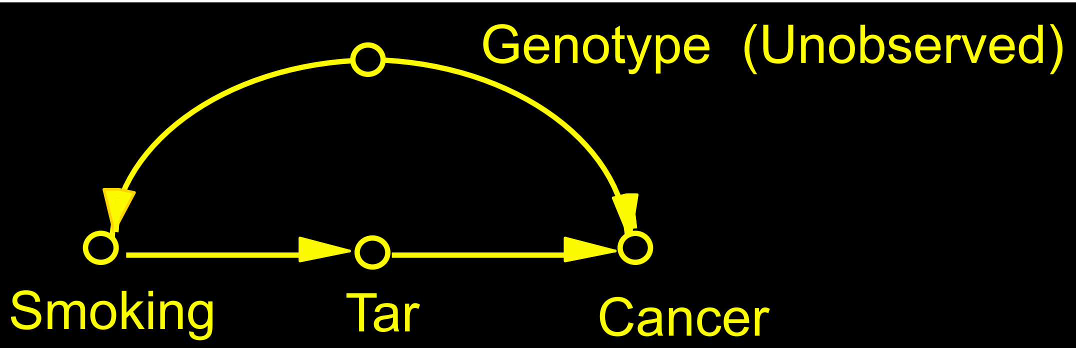
Rule 3: Ignoring actions

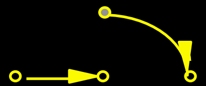
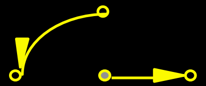
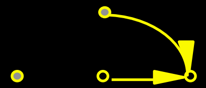
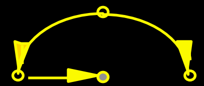
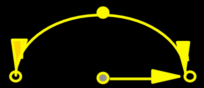
P(y | do{x}, do{z}, w) = P(y | do{x}, w)

if $(Y \perp\!\!\!\perp Z / X,W)_{G_{\overline{X}\,\overline{Z(W)}}}$

Completeness Theorem (Shpitser, 2006)

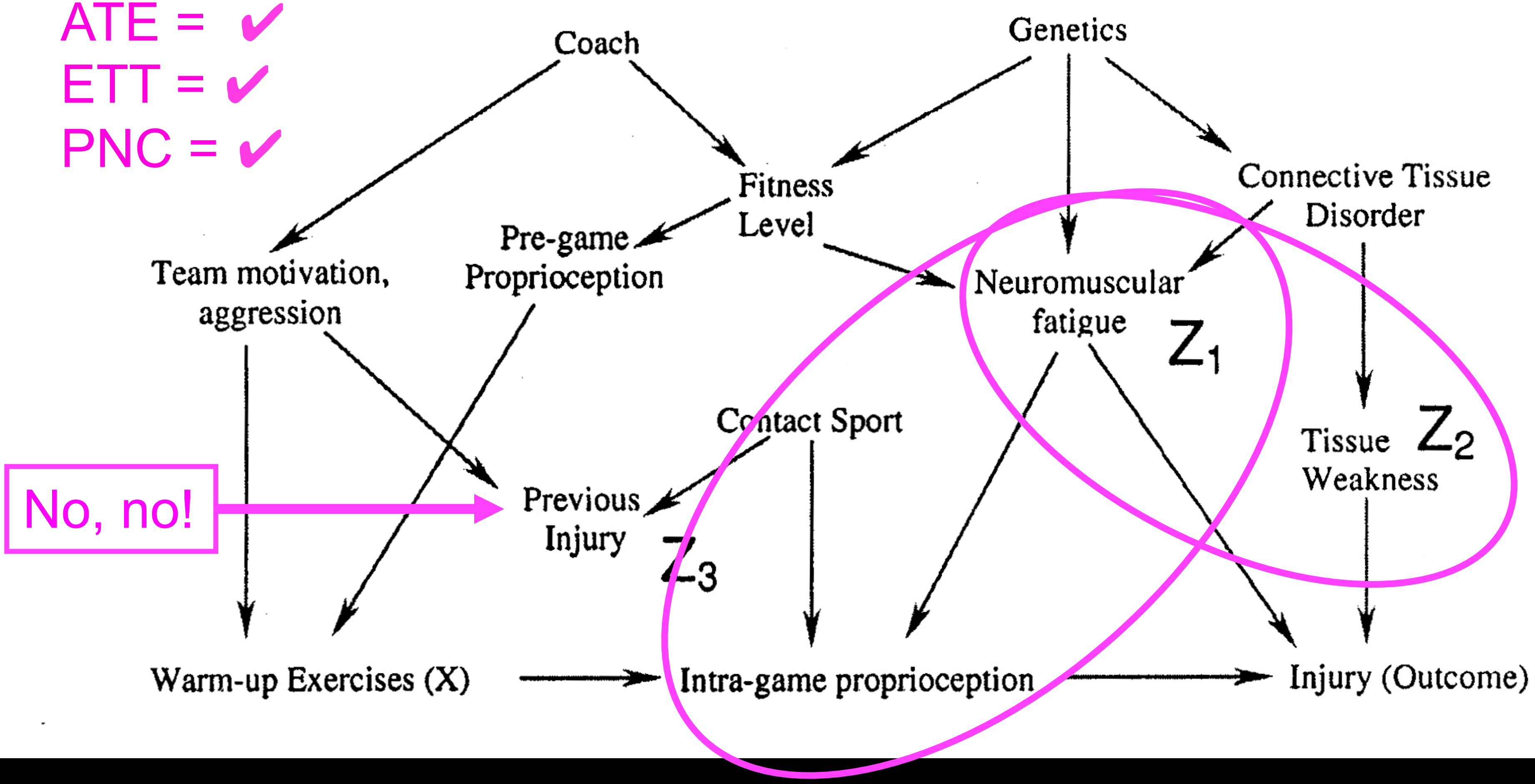
DERIVATION IN CAUSAL CALCULUS



$P(c \mid do\{s\}) = \sum_t P(c \mid do\{s\}, t) P(t \mid do\{s\})$	Probability Axioms	
$= \sum_t P(c \mid do\{s\}, do\{t\}) P(t \mid do\{s\})$	Rule 2	
$= \sum_t P(c \mid do\{s\}, do\{t\}) P(t \mid s)$	Rule 2	
$= \sum_t P(c \mid do\{t\} P(t \mid s)$	Rule 3	
$= \sum_{s'} \sum_t P(c \mid do\{t\}, s') P(s' \mid do\{t\}) P(t \mid s)$	Probability Axioms	
$= \sum_{s'} \sum_t P(c \mid t, s') P(s' \mid do\{t\}) P(t \mid s)$	Rule 2	
$= \sum_{s'} \sum_t P(c \mid t, s') P(s') P(t \mid s)$	Rule 3	

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

ATE = ✓
ETT = ✓
PNC = ✓



MATHEMATICALLY SOLVED PROBLEMS

1. Policy evaluation (ATE, ETT,…)
2. Attribution
3. Mediation (direct and indirect effects)
4. Selection Bias
5. Latent Heterogeneity
6. Transportability
7. Missing Data

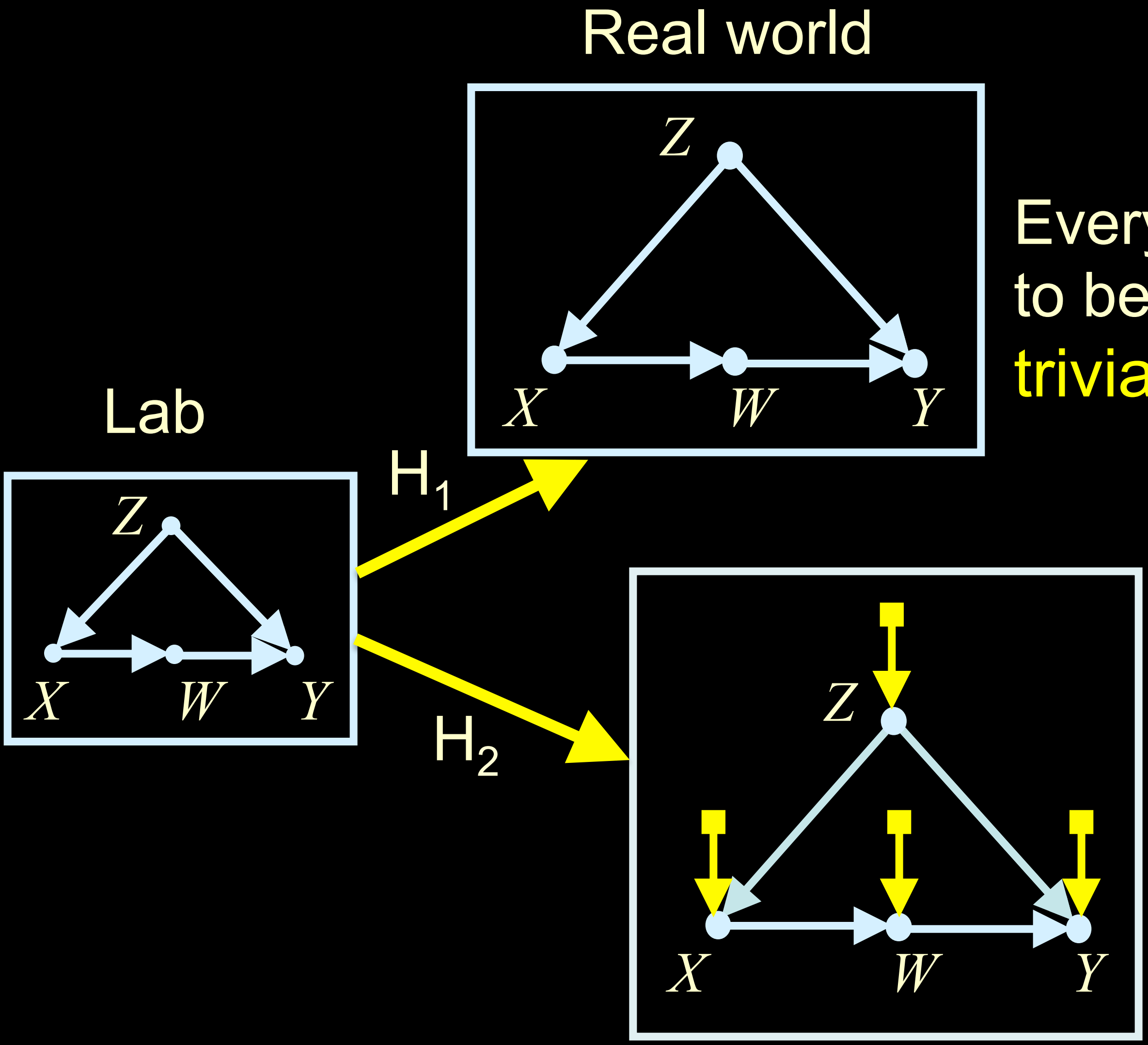
TRANSPORTABILITY OF KNOWLEDGE ACROSS DOMAINS
(with E. Bareinboim)

1. A Theory of causal transportability
   When can causal relations learned from experiments be transferred to a different environment in which no experiment can be conducted?

2. A Theory of statistical transportability
   When can statistical information learned in one domain be transferred to a different domain in which

   a. only a subset of variables can be observed? Or,
   b. only a few samples are available?

MOVING FROM THE LAB TO THE REAL WORLD . . .



Everything is assumed to be the same, trivially transportable!

Everything is assumed to be the different, not transportable!

MOTIVATION

WHAT CAN EXPERIMENTS IN LA TELL ABOUT NYC?

Experimental study in LA

Measured: $P(x,y,z)$

$P(y \mid do(x),z)$

Observational study in NYC

Measured:

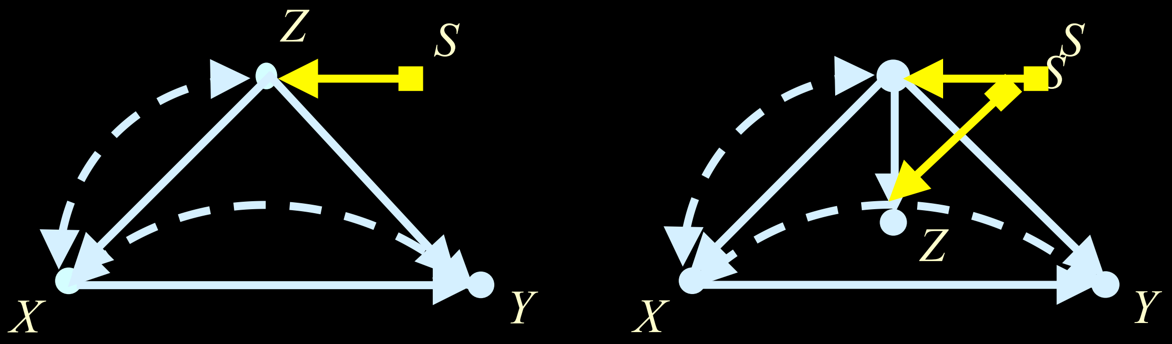
$P^*(x,y,z)$

$P^*(z) \neq P(z)$

Needed: $$P^*(y \mid do(x)) = \quad ? \quad = \sum_z P(y \mid do(x),z) P^*(z)$$

Transport Formula (calibration): $F(P, P_{do}, P^*)$

TRANSPORT FORMULAS DEPEND ON THE STORY



(a) (b)

S ■→ Factors producing differences

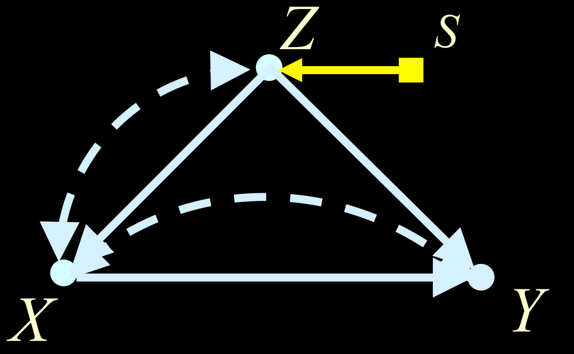
a) ***Z*** represents age

$$P^*(y \mid do(x)) = \sum_z P(y \mid do(x), z) P^*(z)$$

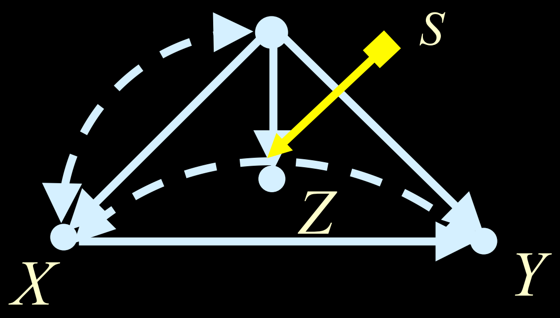
b) ***Z*** represents language skill

$$P^*(y \mid do(x)) \; = \; ? P(y \mid do(x))$$

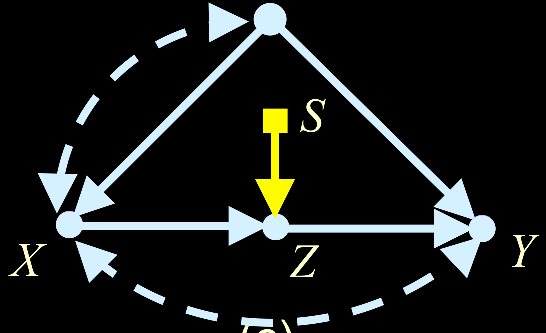
TRANSPORT FORMULAS DEPEND ON THE STORY



(a)



(b)



(c)

a) ***Z*** represents age

$$P^*(y \mid do(x)) = \sum_z P(y \mid do(x), z) P^*(z)$$

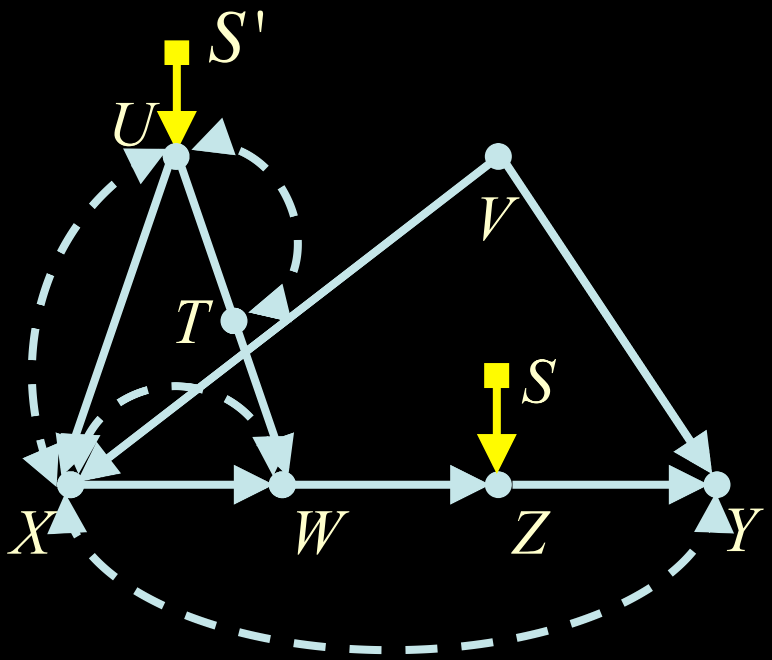
b) ***Z*** represents language skill

$$P^*(y \mid do(x)) = P(y \mid do(x))$$

c) ***Z*** represents a bio-marker

$$P^*(y \mid do(x)) = ?\sum_z P(y \mid do(x), z) P^*(z \mid x)$$

GOAL: ALGORITHM TO DETERMINE IF AN EFFECT IS TRANSPORTABLE



INPUT: Annotated Causal Graph

S → Factors creating differences

OUTPUT:

1. Transportable or not?
2. Measurements to be taken in the experimental study
3. Measurements to be taken in the target population
4. A transport formula

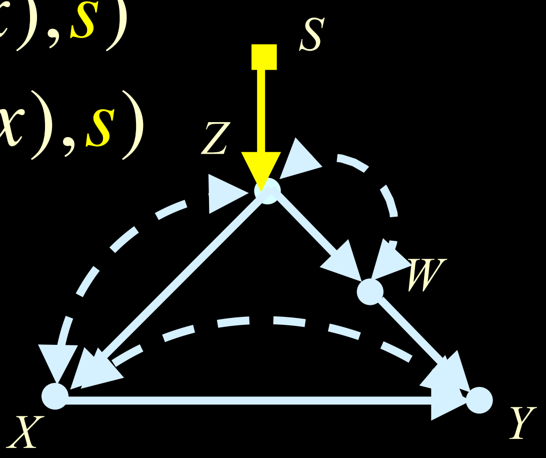
$$P^*(y \mid do(x)) =$$
$$f[P(y,v,z,w,t,u \mid do(x)); P^*(y,v,z,w,t,u)]$$

TRANSPORTABILITY REDUCED TO CALCULUS

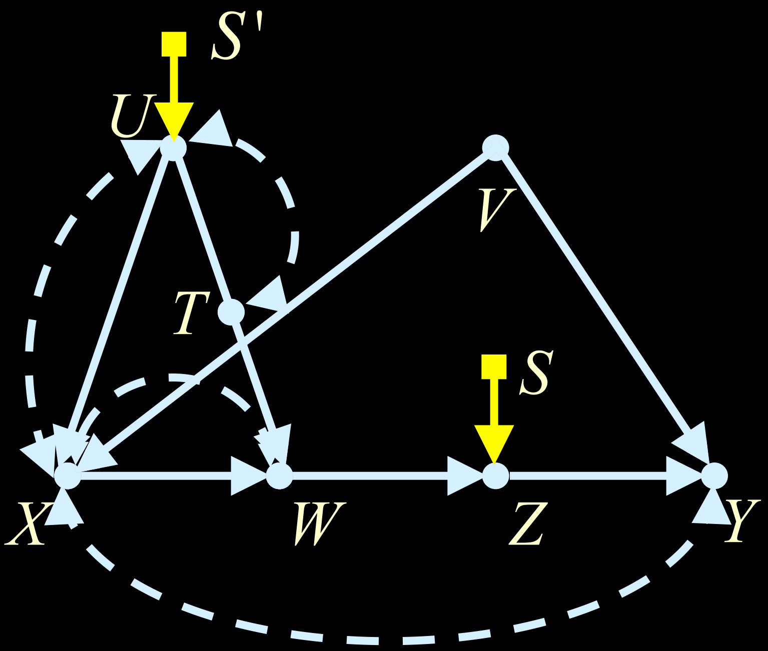
Theorem

A causal relation ***R*** is transportable from ∏ to ∏* if and only if it is reducible, using the rules of ***do***-calculus, to an expression in which ***S*** is separated from ***do***( ).

$$
\begin{aligned}
R\left(\prod *\right) = P^*(y \mid do(x)) &= P(y \mid do(x), s) \\
&= \sum_w P(y \mid do(x), s, w) P(w \mid do(x), s) \\
&= \sum_w P(y \mid do(x), w) P(w \mid s) \\
&= \sum_w P(y \mid do(x), w) P^*(w)
\end{aligned}
$$



RESULT: ALGORITHM TO DETERMINE IF AN EFFECT IS TRANSPORTABLE



INPUT: Annotated Causal Graph

S → Factors creating differences

OUTPUT:

1. Transportable or not?
2. Measurements to be taken in the experimental study
3. Measurements to be taken in the target population
4. A transport formula
5. Completeness (Bareinboim, 2012)

$$P^*(y \mid do(x)) =$$

$$\sum_z P(y \mid do(x), z) \sum_w P^*(z \mid w) \sum_t P(w \mid do(w), t) P^*(t)$$

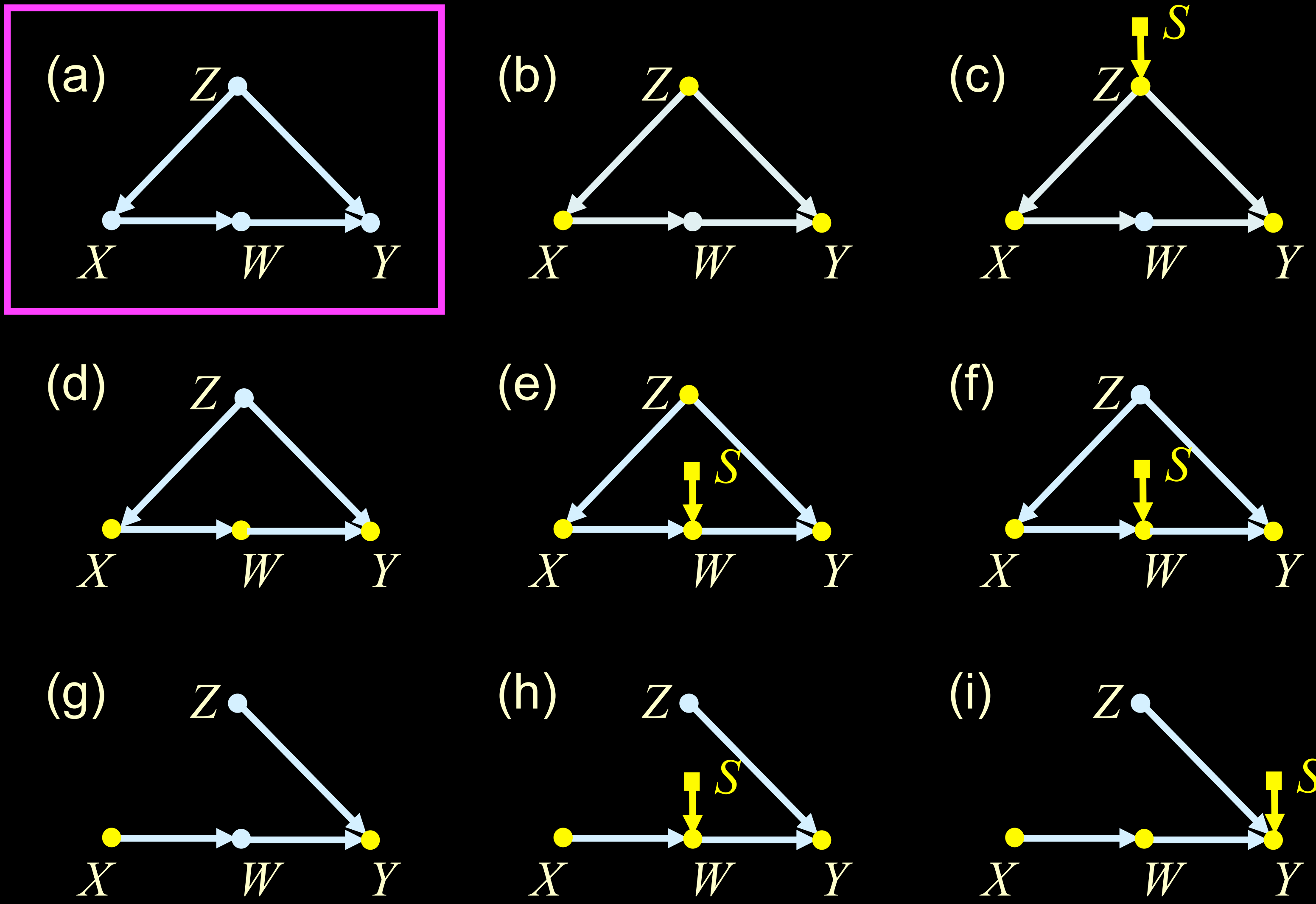
FROM META-ANALYSIS TO META-SYNTHESIS

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 an aggregate measure of effect size that is "better" than any one study in isolation.

META-SYNTHESIS AT WORK

Target population Π^* $R = P^*(y \mid do(x))$



META-SYNTHESIS REDUCED TO CALCULUS

Theorem

$\{\prod_1, \prod_2, \ldots, \prod_K\}$ – a set of studies.

$\{D_1, D_2, \ldots, D_K\}$ – selection diagrams (relative to $\prod^*$).

A relation $R(\prod^*)$ is "meta estimable" if it can be decomposed into terms of the form:

$$Q_k = P(V_k \mid do(W_k), Z_k)$$

such that each Q_k is transportable from D_k.

MISSING DATA:
A SEEMINGLY STATISTICAL PROBLEM
(Mohan, Pearl & Tian 2012)

- Pervasive in every experimental science.
- Huge literature, powerful software industry, deeply entrenched culture.
- Current practices are based on statistical characterization (Rubin, 1976) of a problem that is inherently causal.
- Needed: (1) theoretical guidance, (2) performance guarantees, and (3) tests of assumptions.

WHAT CAN CAUSAL THEORY DO FOR MISSING DATA?

Q-1. What should the world be like, for a given statistical procedure to produce the expected result?

Q-2. Can we tell from the postulated world whether any method can produce a bias-free result? How?

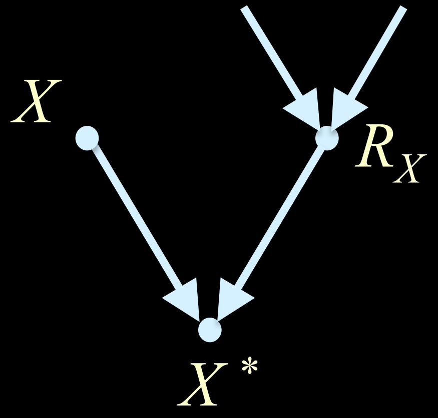
Q-3. Can we tell from data if the world does not work as postulated?

- To answer these questions, we need models of the world, i.e., process models.
- Statistical characterization of the problem is too crude, e.g., MCAR, MAR, MNAR.

GOAL: ESTIMATE $P(X,Y,Z)$

Sample #	Observations			Missingness		
	X*	Y*	Z*	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
3	1	m	m	0	1	1
4	0	1	m	0	0	1
5	m	1	m	1	0	1
6	m	0	1	1	0	0
7	m	m	0	1	1	0
8	0	1	m	0	0	1
9	0	0	m	0	0	1
10	1	0	m	0	0	1
11	1	0	1	0	0	0
-						

Missingness graph



$$X^* = \begin{cases} X \text{ if } R_X = 0 \\ m \text{ if } R_X = 1 \end{cases}$$

NAIVE ESTIMATE OF $P(X,Y,Z)$

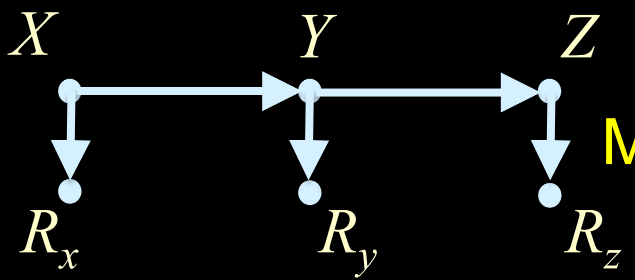
Sam-	Observations			Missingness		
ple #	X*	Y*	Z*	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
3	1	m	m	0	1	1
4	0	1	m	0	0	1
5	m	1	m	1	0	1
6	m	0	1	1	0	0
7	m	m	0	1	1	0
8	0	1	m	0	0	1
9	0	0	m	0	0	1
10	1	0	m	0	0	1
11	1	0	1	0	0	0
-						

Complete Cases

Row #	X	Y	Z	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
11	1	0	1	0	0	0
-						

- Line deletion estimate is generally biased.

$$P(X,Y,Z)$$
$$\neq P(X,Y,Z \mid R_x = 0, R_y = 0, R_z = 0)$$



MCAR

NAIVE ESTIMATE OF $P(X,Y,Z)$

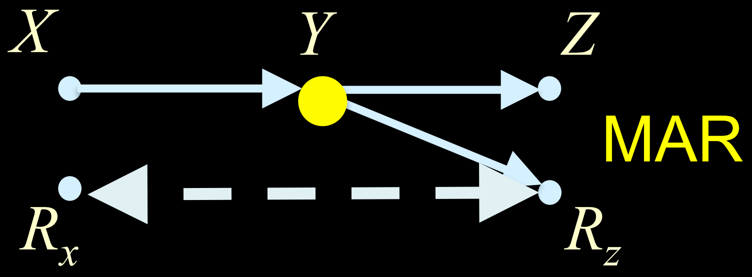
Sample #	Observations			Missingness		
	X^*	Y^*	Z^*	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
3	1	m	m	0	1	1
4	0	1	m	0	0	1
5	m	1	m	1	0	1
6	m	0	1	1	0	0
7	m	m	0	1	1	0
8	0	1	m	0	0	1
9	0	0	m	0	0	1
10	1	0	m	0	0	1
11	1	0	1	0	0	0
-						

Complete Cases

Row #	X	Y	Z	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
11	1	0	1	0	0	0
-						

- Line deletion estimate is generally biased.

$$P(X,Y,Z) \neq P(X,Y,Z \mid R_x=0, R_y=0, R_z=0)$$



NAIVE ESTIMATE OF $P(X,Y,Z)$

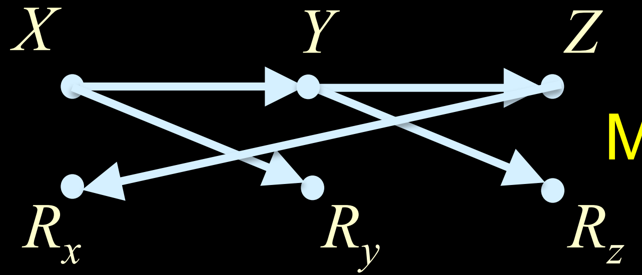
Sample #	Observations			Missingness		
	X*	Y*	Z*	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
3	1	m	m	0	1	1
4	0	1	m	0	0	1
5	m	1	m	1	0	1
6	m	0	1	1	0	0
7	m	m	0	1	1	0
8	0	1	m	0	0	1
9	0	0	m	0	0	1
10	1	0	m	0	0	1
11	1	0	1	0	0	0
-						

Complete Cases

Row #	X	Y	Z	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
11	1	0	1	0	0	0
-						

- Line deletion estimate is generally biased.

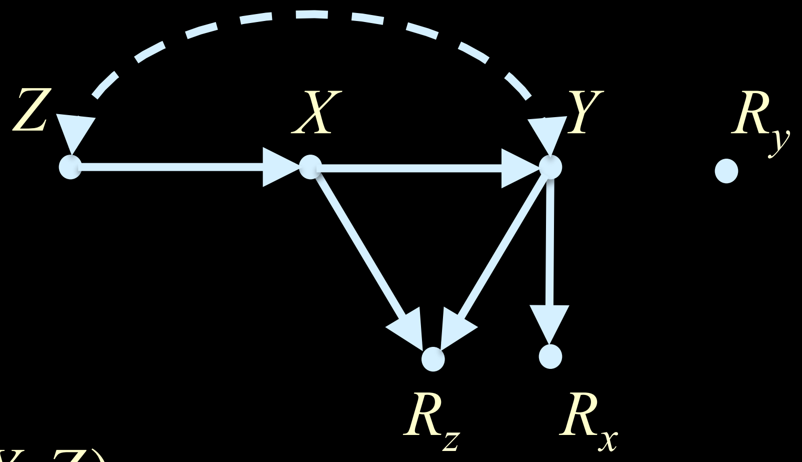
$$P(X,Y,Z) \neq P(X,Y,Z \mid R_x = 0, R_y = 0, R_z = 0)$$



MNAR

SMART ESTIMATE OF $P(X,Y,Z)$

Sample #	Observations			Missingness		
	X*	Y*	Z*	R_x	R_y	R_z
1	1	0	0	0	0	0
2	1	0	1	0	0	0
3	1	m	m	0	1	1
4	0	1	m	0	0	1
5	m	1	m	1	0	1
6	m	0	1	1	0	0
7	m	m	0	1	1	0
8	0	1	m	0	0	1
9	0	0	m	0	0	1
10	1	0	m	0	0	1
11	1	0	1	0	0	0
-						



$$P(X,Y,Z) \neq P(X,Y,Z \mid R_x=0, R_y=0, R_z=0)$$

$$P(X,Y,Z) = P(Z \mid X,Y)P(X \mid Y)P(Y)$$

$$P(Y) = P(Y \mid R_y = 0)$$

$$P(X \mid Y) = P(X \mid Y, R_x = 0, R_y = 0)$$

$$P(Z \mid X,Y) = P(Z \mid X,Y, R_x = 0, R_y = 0, R_z = 0)$$

SMART ESTIMATE OF $P(X,Y,Z)$

$$P(X,Y,Z) = P(Z \mid X,Y,R_x=0,R_y=0,R_z=0)$$

$$P(X \mid Y,R_x=0,R_y=0)$$

$$P(Y \mid R_y=0)$$

Sam-ple #	X*	Y*	Z*
1	1	0	0
2	1	0	1
3	1	m	m
4	0	1	m
5	m	1	m
6	m	0	1
7	m	m	0
8	0	1	m
9	0	0	m
10	1	0	m
11	1	0	1
-			

SMART ESTIMATE OF $P(X,Y,Z)$

$$P(X,Y,Z) = P(Z \mid X,Y,R_x=0,R_y=0,R_z=0)$$

$$P(X \mid Y,R_x=0,R_y=0)$$

$$P(Y \mid R_y=0)$$

Sample #	X*	Y*	Z*
1	1	0	0
2	1	0	1
3	1	m	m
4	0	1	m
5	m	1	m
6	m	0	1
7	m	m	0
8	0	1	m
9	0	0	m
10	1	0	m
11	1	0	1
-			

Compute $P(Y|R_y=0)$

Row #	Y*
1	0
2	0
4	1
5	1
6	0
8	1
9	0
10	0
11	0
-	

SMART ESTIMATE OF $P(X,Y,Z)$

$$P(X,Y,Z) = P(Z \mid X,Y,R_x = 0,R_y = 0,R_z = 0)$$

$$P(X \mid Y,R_x = 0,R_y = 0)$$

$$P(Y \mid R_y = 0)$$

Sam-ple #	X*	Y*	Z*
1	1	0	0
2	1	0	1
3	1	m	m
4	0	1	m
5	m	1	m
6	m	0	1
7	m	m	0
8	0	1	m
9	0	0	m
10	1	0	m
11	1	0	1
-			

Compute $P(Y|R_y=0)$

Row #	Y*
1	0
2	0
4	1
5	1
6	0
8	1
9	0
10	0
11	0
-	

Compute $P(X|Y,R_x=0,R_y=0)$

Row #	X*	Y*
1	1	0
2	1	0
4	0	1
8	0	1
9	0	0
10	1	0
11	1	0
-		

SMART ESTIMATE OF $P(X,Y,Z)$

$$P(X,Y,Z) = P(Z \mid X,Y,R_x=0,R_y=0,R_z=0)$$

$$P(X \mid Y,R_x=0,R_y=0)$$

$$P(Y \mid R_y=0)$$

Sample #	X*	Y*	Z*
1	1	0	0
2	1	0	1
3	1	m	m
4	0	1	m
5	m	1	m
6	m	0	1
7	m	m	0
8	0	1	m
9	0	0	m
10	1	0	m
11	1	0	1
-			

Compute $P(Y|R_y=0)$

Row #	Y*
1	0
2	0
4	1
5	1
6	0
8	1
9	0
10	0
11	0
-	

Compute $P(X|Y,R_x=0,R_y=0)$

Row #	X*	Y*
1	1	0
2	1	0
4	0	1
8	0	1
9	0	0
10	1	0
11	1	0
-		

Compute $P(Z|X,Y,R_x=0,R_y=0,R_z=0)$

Row #	X*	Y*	Z*
1	1	0	0
2	1	0	1
11	1	0	1
-			

RECOVERABILITY FROM MISSING DATA

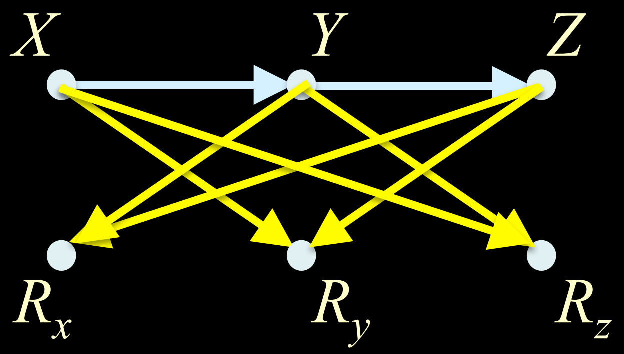
Definition:
Given a missingness model M, a probabilistic quantity Q is said to be recoverable if there exists an algorithm that produces a consistent estimate of Q for every dataset generated by M.

That is, in the limit of large sample, Q is estimable as if no data were missing.

RECOVERABILITY IN MARKOVIAN MODELS

Theorem:
If the missingness-graph is Markovian (i.e., no latent variables) then a necessary and sufficient condition for recoverability of $P(V)$ is that no variable X be adjacent to its missingness mechanism R_x.

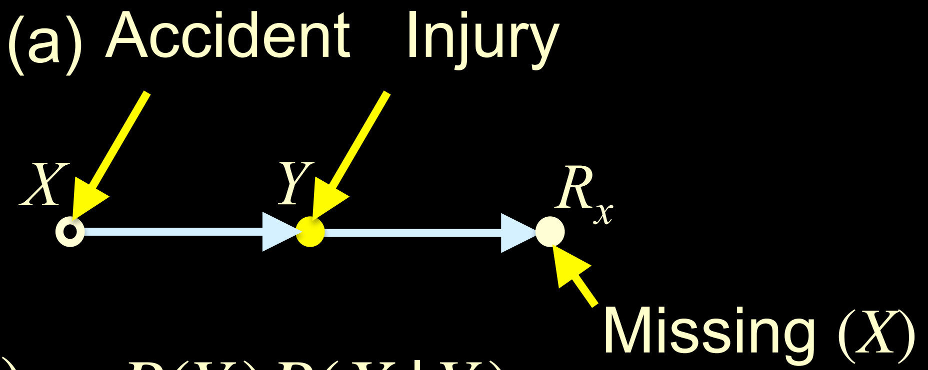
e.g.,



DECIDING RECOVERABILITY

Theorem:
Q is recoverable iff it is decomposable into terms of the form $Q_j = P(S_j \mid T_j)$ such that T_j contains the missingness mechanism R_v of every partially observed variable V that appears in Q.

e.g.,



$$Q_1 = P(X,Y) = P(Y)P(X \mid Y)$$

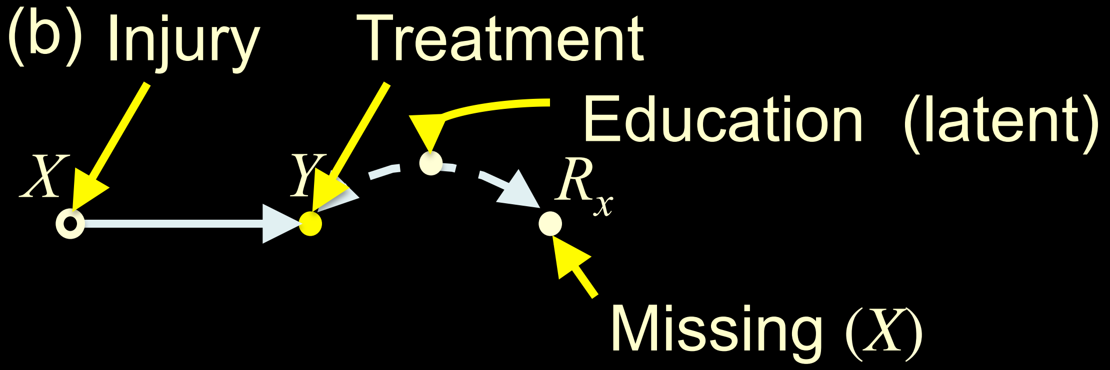
$$= P(Y)P(X \mid Y, R_x) \quad \text{recoverable}$$

$$Q_2 = P(X) = \sum_y P(X,Y) \quad \text{recoverable}$$

DECIDING RECOVERABILITY

Theorem:
Q is recoverable iff it is decomposable into terms of the form $Q_j = P(S_j \mid T_j)$ such that T_j contains the missingness mechanism R_v of every partially observed variable V that appears in Q.

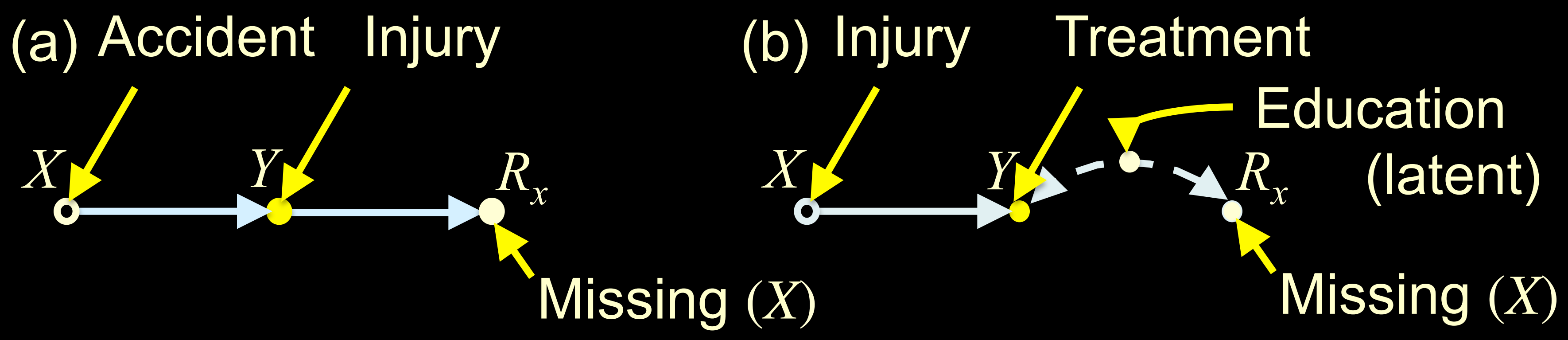
e.g.,



$$Q_1 = P(X,Y) \neq P(Y)P(X \mid Y, R_x)$$ nonrecoverable

$$Q_2 = P(X) = P(X \mid R_x)$$ recoverable

AN IMPOSSIBILITY THEOREM FOR MISSING DATA



(a) Accident Injury

X Y R_x

Missing (X)

(b) Injury Treatment

Education (latent)

X Y R_x

Missing (X)

- Two statistically indistinguishable models, yet $P(X,Y)$ is recoverable in (a) and not in (b).
- No universal algorithm exists that decides recoverability (or guarantees unbiased results) without looking at the model.

A STRONGER IMPOSSIBILITY THEOREM

- Two statistically indistinguishable models, $P(X)$ is recoverable in both, but through two different methods:

  In (a): $P(X) = \sum_y P(Y)P(X \mid Y, R_x = 0)$, while

  in (b): $P(X) = P(X \mid R_x = 0)$
- No universal algorithm exists that produces an unbiased estimate whenever such exists.

CONCLUSIONS

Deduction is indispensible in causal inference, as it is in science and machine learning.

1. Think nature, not data, not even experiment.
2. Counterfactuals, the building blocks of scientific and moral thinking can be algorithmitized.
3. Identifiability, testability, recoverability and transportability are computational tasks with formal solutions.
4. Think Nature, not data.

Thank you