

## THE MATHEMATICS OF CAUSE AND EFFECT

Judea Pearl  
University of California  
Los Angeles  
([www.cs.ucla.edu/~judea](http://www.cs.ucla.edu/~judea))

## OUTLINE

---

- Causality: Antiquity to robotics
- Modeling: Statistical vs. Causal
- Causal Models and Identifiability
- Inference to three types of claims:
  1. Effects of potential interventions
  2. Claims about attribution (responsibility)
  3. Claims about direct and indirect effects

## ANTIQUITY TO ROBOTICS

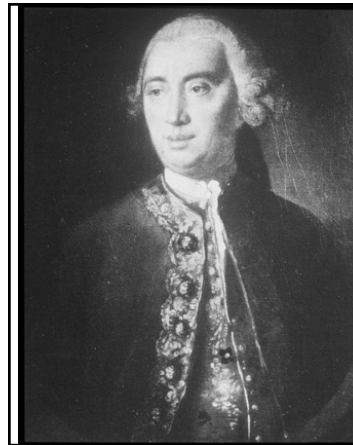
---

“I would rather discover one causal relation than be King of Persia”

Democritus (430-380 BC)

Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).

A. Einstein, April 23, 1953



*David Hume*  
(1711–1776)

## HUME'S LEGACY

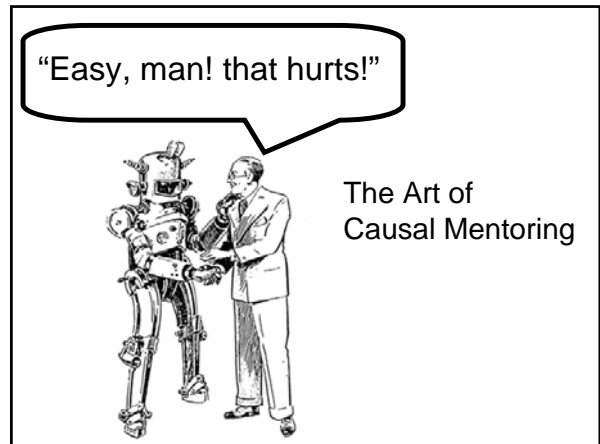
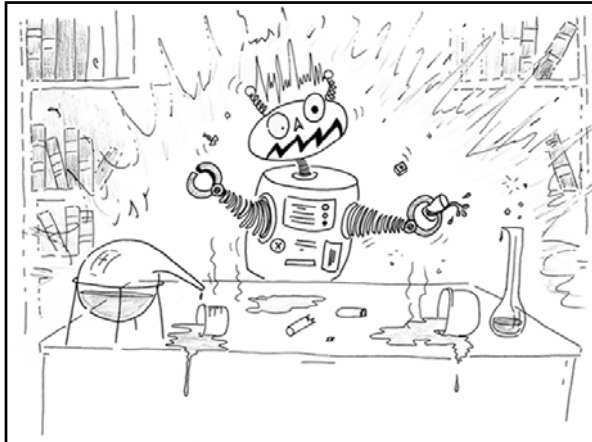
---

1. Analytical vs. empirical claims
2. Causal claims are empirical
3. All empirical claims originate from experience.

## THE TWO RIDDLES OF CAUSATION

---

- What empirical evidence legitimizes a cause-effect connection?
- What inferences can be drawn from causal information? and how?



## OLD RIDDLES IN NEW DRESS

---

1. How should a robot acquire causal information from the environment?
2. How should a robot process causal information received from its creator-programmer?

## CAUSATION AS A PROGRAMMER'S NIGHTMARE

---

Input:

1. "If the grass is wet, then it rained"
2. "if we break this bottle, the grass will get wet"

Output:

"If we break this bottle, then it rained"

## CAUSATION AS A PROGRAMMER'S NIGHTMARE (Cont.) ( Lin, 1995)

---

Input:

1. A suitcase will open iff both locks are open.
2. The right lock is open

Query:

What if we open the left lock?

Output:

The right lock might get closed.

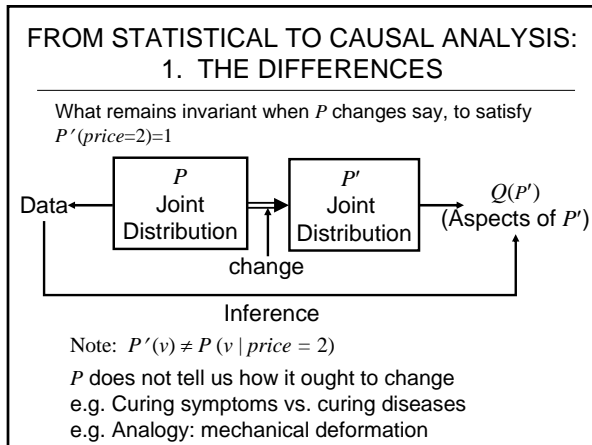
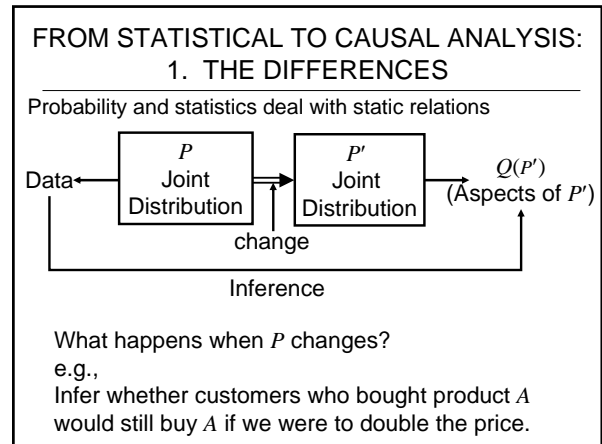
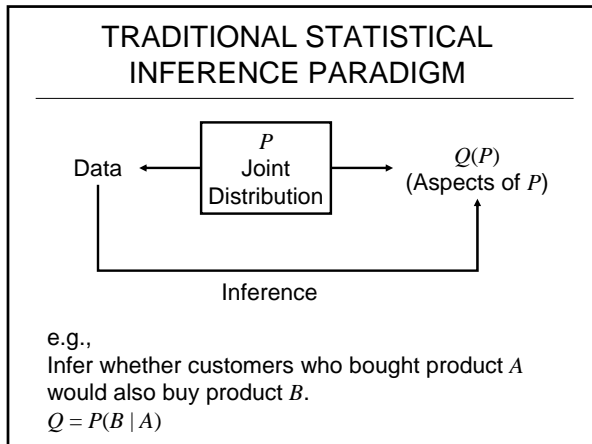
## THE BASIC PRINCIPLES

---

Causation = encoding of behavior under interventions

Interventions = surgeries on mechanisms

Mechanisms = stable functional relationships  
= equations + graphs



- ### FROM STATISTICAL TO CAUSAL ANALYSIS: 1. THE DIFFERENCES (CONT)
- 
- Causal and statistical concepts do not mix.
 

<b>CAUSAL</b> Spurious correlation Randomization Confounding / Effect Instrument Holding constant Explanatory variables	<b>STATISTICAL</b> Regression Association / Independence "Controlling for" / Conditioning Odd and risk ratios Collapsibility Propensity score
---	---
  - 
  - 
  -

- ### FROM STATISTICAL TO CAUSAL ANALYSIS: 1. THE DIFFERENCES (CONT)
- 
- Causal and statistical concepts do not mix.
 

<b>CAUSAL</b> Spurious correlation Randomization Confounding / Effect Instrument Holding constant Explanatory variables	<b>STATISTICAL</b> Regression Association / Independence "Controlling for" / Conditioning Odd and risk ratios Collapsibility Propensity score
---	---
  - No causes in – no causes out (Cartwright, 1989)  
 $\left. \begin{array}{l} \text{statistical assumptions + data} \\ \text{causal assumptions} \end{array} \right\} \Rightarrow \text{causal conclusions}$
  - Causal assumptions cannot be expressed in the mathematical language of standard statistics.
  -

- ### FROM STATISTICAL TO CAUSAL ANALYSIS: 1. THE DIFFERENCES (CONT)
- 
- Causal and statistical concepts do not mix.
 

<b>CAUSAL</b> Spurious correlation Randomization Confounding / Effect Instrument Holding constant Explanatory variables	<b>STATISTICAL</b> Regression Association / Independence "Controlling for" / Conditioning Odd and risk ratios Collapsibility Propensity score
---	---
  - No causes in – no causes out (Cartwright, 1989)  
 $\left. \begin{array}{l} \text{statistical assumptions + data} \\ \text{causal assumptions} \end{array} \right\} \Rightarrow \text{causal conclusions}$
  - Causal assumptions cannot be expressed in the mathematical language of standard statistics.
  - Non-standard mathematics:
    - Structural equation models (Wright, 1920; Simon, 1960)
    - Counterfactuals (Neyman-Rubin ( $Y_x$ ), Lewis ( $x \boxrightarrow Y$ ))

**FROM STATISTICAL TO CAUSAL ANALYSIS:  
2. THE MENTAL BARRIERS**

1. Every exercise of causal analysis must rest on untested, judgmental causal assumptions.
2. Every exercise of causal analysis must invoke non-standard mathematical notation.

**WHY CAUSALITY NEEDS  
SPECIAL MATHEMATICS**

SEM Equations are Non-algebraic:

$$\begin{array}{ll} Y = 2X & X = 1 \\ X = 1 & Y = 2 \end{array}$$

Process information                      Static information

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

**TWO PARADIGMS FOR  
CAUSAL INFERENCE**

Observed:  $P(X, Y, Z, \dots)$   
Conclusions needed:  $P(Y_x = y), P(X_y = x | Z = z) \dots$

How do we connect observables,  $X, Y, Z, \dots$   
to counterfactuals  $Y_x, X_y, Z_z, \dots$  ?

<p><u>N-R model</u> Counterfactuals are primitives, new variables Super-distribution <math>P^*(X, Y, \dots, Y_x, X_y, \dots)</math> <math>X, Y, Z</math> constrain <math>Y_x, Z_y, \dots</math></p>	<p><u>Structural model</u> Counterfactuals are derived quantities Subscripts modify a data-generating model</p>
---	---

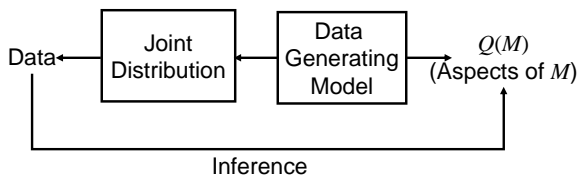
**"SUPER" DISTRIBUTION  
IN N-R MODEL**

X	Y	Z	$Y_{x=0}$	$Y_{x=1}$	$X_{z=0}$	$X_{z=1}$	$X_{y=0} \dots$	U
0	0	0	0	1	0	0	0...	$u_1$
0	1	1	1	0	1	0	1...	$u_2$
0	0	0	1	0	0	1	1...	$u_3$
1	0	0	0	0	0	1	0...	$u_4$

inconsistency:  $x = 0 \Rightarrow Y_{x=0} = Y$        $Y = xY_1 + (1-x)Y_0$

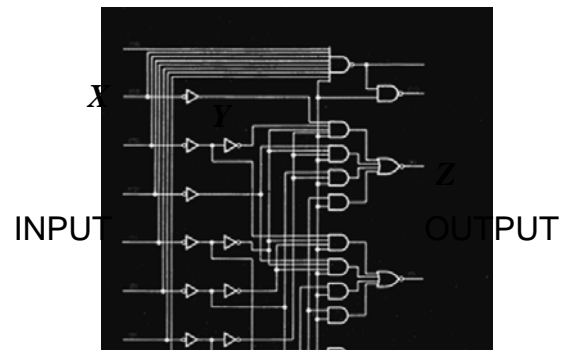
Defines:  $P^*(X, Y, Z, \dots, Y_x, Z_y, \dots, Y_{xz}, Z_{xy}, \dots \dots)$   
 $P^*(Y_x = y | Z, X_z)$   
 $Y_x \perp\!\!\!\perp X | Z_y$

**THE STRUCTURAL MODEL  
PARADIGM**



M – Oracle for computing answers to Q's.  
e.g.,  
Infer whether customers who bought product A would still buy A if we were to double the price.

**FAMILIAR CAUSAL MODEL  
ORACLE FOR MANIPULATION**

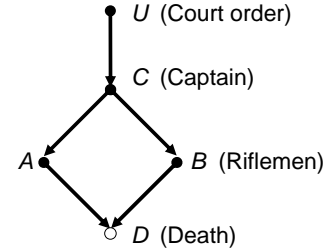


## STRUCTURAL CAUSAL MODELS

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

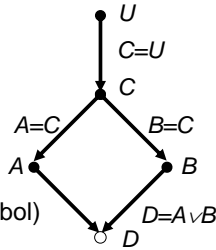
- $V = \{V_1, \dots, V_n\}$  are observable variables
  - $U = \{U_1, \dots, U_m\}$  are background variables
  - $F = \{f_1, \dots, f_n\}$  are functions determining  $V$ ,  
 $v_i = f_i(v, u)$
  - $P(u)$  is a distribution over  $U$
- $P(u)$  and  $F$  induce a distribution  $P(v)$  over observable variables

## CAUSAL MODELS AT WORK (The impatient firing-squad)



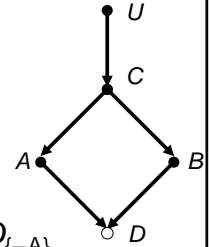
## CAUSAL MODELS AT WORK (Glossary)

- $U$ : Court orders the execution
- $C$ : Captain gives a signal
- $A$ : Rifleman-A shoots
- $B$ : Rifleman-B shoots
- $D$ : Prisoner dies
- $=$ : Functional Equality (new symbol)



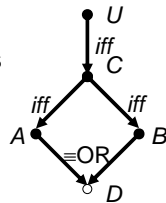
## SENTENCES TO BE EVALUATED

- S1. prediction:  $\neg A \Rightarrow \neg D$
- S2. abduction:  $\neg D \Rightarrow \neg C$
- S3. transduction:  $A \Rightarrow B$
- S4. action:  $\neg C \Rightarrow D_A$
- S5. counterfactual:  $D \Rightarrow D_{\{\neg A\}}$
- S6. explanation: Caused( $A, D$ )



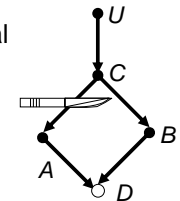
## STANDARD MODEL FOR STANDARD QUERIES

- S1. (prediction): If rifleman-A shot, the prisoner is dead,  
 $A \Rightarrow D$
- S2. (abduction): If the prisoner is alive, then the Captain did not signal,  
 $\neg D \Rightarrow \neg C$
- S3. (transduction): If rifleman-A shot, then  $B$  shot as well,  
 $A \Rightarrow B$



## WHY CAUSAL MODELS? GUIDE FOR SURGERY

- S4. (action):  
 If the captain gave no signal and Mr. A decides to shoot, the prisoner will die:  
 $\neg C \Rightarrow D_A$ ,  
 and  $B$  will not shoot:  
 $\neg C \Rightarrow \neg B_A$



## WHY CAUSAL MODELS? GUIDE FOR SURGERY

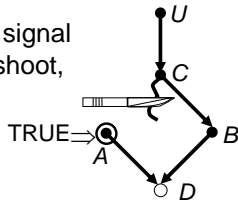
S4. (action):

If the captain gave no signal  
and Mr. A decides to shoot,  
the prisoner will die:

$$\neg C \Rightarrow D_A,$$

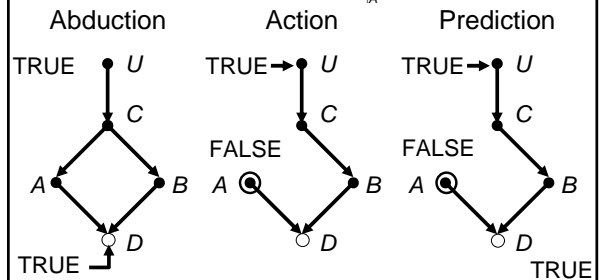
and B will not shoot:

$$\neg C \Rightarrow \neg B_A$$



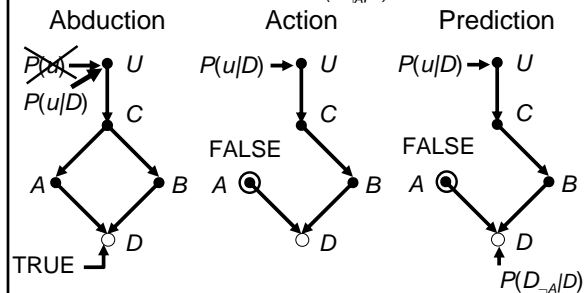
## 3-STEPS TO COMPUTING COUNTERFACTUALS

S5. If the prisoner is dead, he would still be dead  
if A were not to have shot.  $D \Rightarrow D_{\neg A}$



## COMPUTING PROBABILITIES OF COUNTERFACTUALS

P(S5). The prisoner is dead. How likely is it that he would be dead  
if A were not to have shot.  $P(D_{\neg A}|D) = ?$



## CAUSAL MODEL (FORMAL)

$$M = \langle U, V, F \rangle \text{ or } \langle U, V, F, P(u) \rangle$$

$U$  - Background variables

$V$  - Endogenous variables

$F$  - Set of functions  $\{U \times V \rightarrow V_i\}$

$$v_i = f_i(p_{a_i}, u_i)$$

Submodel:  $M_x = \langle U, V, F_x \rangle$ , representing  $do(x)$

$F_x$  - Replaces equation for  $X$  with  $X=x$

Actions and Counterfactuals:

$$Y_x(u) = \text{Solution of } Y \text{ in } M_x$$

$$P(y | do(x)) \triangleq P(Y_x=y)$$

## WHY COUNTERFACTUALS?

Action queries are triggered by (modifiable) observations,  
demanding abductive step, i.e., counterfactual processing.

E.g., Troubleshooting

Observation: The output is low

Action query: Will the output get higher –  
if we replace the transistor?

Counterfactual query: Would the output be higher –  
had the transistor been replaced?

## APPLICATIONS

1. Predicting effects of actions and policies
2. Learning causal relationships from assumptions and data
3. Troubleshooting physical systems and plans
4. Finding explanations for reported events
5. Generating verbal explanations
6. Understanding causal talk
7. Formulating theories of causal thinking

## CAUSAL MODELS AND COUNTERFACTUALS

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.

Joint probabilities of counterfactuals:

$$P(Y_x = y, Z_w = z) = \sum_{u: Y_x(u)=y, Z_w(u)=z} P(u)$$

The super-distribution  $P^*$  is derived from  $M$ . Parsimonious, consistent, and transparent

## AXIOMS OF CAUSAL COUNTERFACTUALS

Y would be y, had X been x (in state  $U = u$ )

1. Definiteness  
 $\exists x \in X \text{ s.t. } X_y(u) = x$
2. Uniqueness  
 $(X_y(u) = x) \& (X_{y'}(u) = x') \Rightarrow x = x'$
3. Effectiveness  
 $X_{xw}(u) = x$
4. Composition  
 $W_x(u) = w \Rightarrow Y_{xw}(u) = Y_x(u)$
5. Reversibility  
 $(Y_{xw}(u) = y \& (W_{xy}(u) = w) \Rightarrow Y_x(u) = y$


## GRAPHICAL – COUNTERFACTUALS SYMBIOSIS

Every causal graph expresses counterfactuals assumptions,

e.g.,

$$X \rightarrow Y \rightarrow Z$$

1. Missing arrows  $X \rightarrow Z$   $Z_{x,y}(u) = Z_y(u)$

2. Missing arcs   $Y_x \perp\!\!\!\perp Z_y$

Assumptions are guaranteed consistency.  
Assumptions are readable from the graph.

## RULES OF CAUSAL CALCULUS

Rule 1: Ignoring observations

$$P(y | do(x), z, w) = P(y | do(x), w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}}}$$

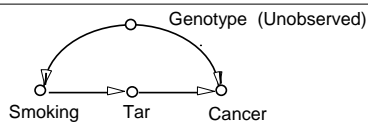
Rule 2: Action/observation exchange

$$P(y | do(x), do(z), w) = P(y | do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}Z}}$$

Rule 3: Ignoring actions

$$P(y | do(x), do(z), w) = P(y | do(x), w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}\bar{Z}\bar{W}}}$$

## DERIVATION IN CAUSAL CALCULUS

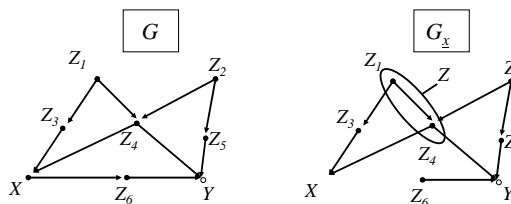


$$\begin{aligned}
 P(c | do(s)) &= \sum_t P(c | do(s), t) P(t | do(s)) && \text{Probability Axioms} \\
 &= \sum_t P(c | do(s), do(t)) P(t | do(s)) && \text{Rule 2} \\
 &= \sum_t P(c | do(s), do(t)) P(t | s) && \text{Rule 2} \\
 &= \sum_t P(c | do(t)) P(t | s) && \text{Rule 3} \\
 &= \sum_{s'} \sum_t P(c | do(t), s') P(s' | do(t)) P(t | s) && \text{Probability Axioms} \\
 &= \sum_{s'} \sum_t P(c | t, s') P(s' | do(t)) P(t | s) && \text{Rule 2} \\
 &= \sum_{s'} \sum_t P(c | t, s') P(s') P(t | s) && \text{Rule 3}
 \end{aligned}$$

## THE BACK-DOOR CRITERION

Graphical test of identification

$P(y | do(x))$  is identifiable in  $G$  if there is a set  $Z$  of variables such that  $Z$  d-separates  $X$  from  $Y$  in  $G_{\bar{X}}$ .



Moreover,  $P(y | do(x)) = \sum_z P(y | x, z) P(z)$   
("adjusting" for  $Z$ )

## RECENT RESULTS ON IDENTIFICATION

---

- *do*-calculus is complete
- Complete graphical criterion for identifying causal effects (Shpitser and Pearl, 2006).
- Complete graphical criterion for empirical testability of counterfactuals (Shpitser and Pearl, 2007).

## DETERMINING THE CAUSES OF EFFECTS (The Attribution Problem)

---

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



- 

## DETERMINING THE CAUSES OF EFFECTS (The Attribution Problem)

---

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



- Court to decide if it is MORE PROBABLE THAN NOT that *A* would be alive BUT FOR the drug!  
 $PN = P(? | A \text{ is dead, took the drug}) \geq 0.50$

## THE PROBLEM

---

Semantical Problem:

1. What is the meaning of  $PN(x,y)$ :  
"Probability that event *y* would not have occurred if it were not for event *x*, given that *x* and *y* did in fact occur."

- 

## THE PROBLEM

---

Semantical Problem:

1. What is the meaning of  $PN(x,y)$ :  
"Probability that event *y* would not have occurred if it were not for event *x*, given that *x* and *y* did in fact occur."

Answer:

$$PN(x, y) = P(Y_{x'} = y' | x, y)$$

Computable from *M*

## THE PROBLEM

---

Semantical Problem:

1. What is the meaning of  $PN(x,y)$ :  
"Probability that event *y* would not have occurred if it were not for event *x*, given that *x* and *y* did in fact occur."

Analytical Problem:

2. Under what condition can  $PN(x,y)$  be learned from statistical data, i.e., observational, experimental and combined.

## TYPICAL THEOREMS (Tian and Pearl, 2000)

- Bounds given combined nonexperimental and experimental data

$$\max \left\{ \frac{0}{P(x,y)} \right\} \leq PN \leq \min \left\{ \frac{1}{P(x,y)} \right\}$$

- Identifiability under monotonicity (Combined data)

$$PN = \frac{P(y/x) - P(y/x')}{P(y/x)} + \frac{P(y/x') - P(y/x)}{P(x,y)}$$

corrected Excess-Risk-Ratio

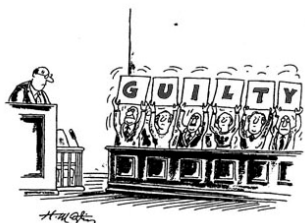
## CAN FREQUENCY DATA DECIDE LEGAL RESPONSIBILITY?

	Experimental		Nonexperimental	
	$do(x)$	$do(x')$	$x$	$x'$
Deaths (y)	16	14	2	28
Survivals (y')	984	986	998	972
	1,000	1,000	1,000	1,000

- Nonexperimental data: drug usage predicts longer life
- Experimental data: drug has negligible effect on survival
- Plaintiff: Mr. A is special.
  1. He actually died
  2. He used the drug by choice
- Court to decide (given both data):  
Is it more probable than not that A would be alive but for the drug?

$$PN \triangleq P(Y_{x'} = y' | x, y) > 0.50$$

## SOLUTION TO THE ATTRIBUTION PROBLEM



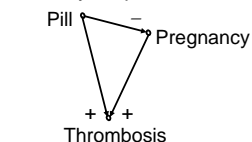
- WITH PROBABILITY ONE  $1 \leq P(y_{x'} | x, y) \leq 1$
- Combined data tell more than each study alone

## EFFECT DECOMPOSITION

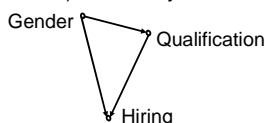
- What is the semantics of direct and indirect effects?
- What are their policy-making implications?
- Can we estimate them from data? Experimental data?

## WHY DECOMPOSE EFFECTS?

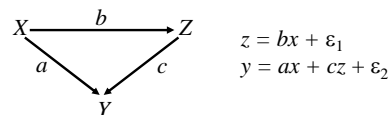
1. Direct (or indirect) effect may be more transportable.
2. Indirect effects may be prevented or controlled.



3. Direct (or indirect) effect may be forbidden



## TOTAL, DIRECT, AND INDIRECT EFFECTS HAVE SIMPLE SEMANTICS IN LINEAR MODELS



$$z = bx + \varepsilon_1$$

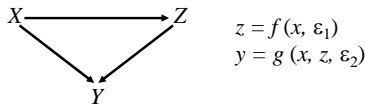
$$y = ax + cz + \varepsilon_2$$

$$TE \triangleq \frac{\partial}{\partial x} E(Y | do(x)) = a + bc$$

$$DE \triangleq \frac{\partial}{\partial x} E(Y | do(x), do(z)) = a \quad Z \text{ - independent}$$

$$IE \triangleq TE - DE = bc$$

**SEMANTICS BECOMES NONTRIVIAL  
IN NONLINEAR MODELS**  
(even when the model is completely specified)

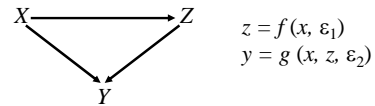


$TE \triangleq \frac{\partial}{\partial x} E(Y | do(x))$

$DE \triangleq \frac{\partial}{\partial x} E(Y | do(x), do(z))$     Dependent on  $z$ ?

$IE \triangleq \text{????}$     Void of operational meaning?

**THE OPERATIONAL MEANING OF  
DIRECT EFFECTS**



"Natural" Direct Effect of X on Y:  
The expected change in Y per unit change of X, when we keep Z constant at whatever value it attains before the change.

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

In linear models,  $NDE = \text{Controlled Direct Effect}$

**LEGAL DEFINITIONS TAKE THE  
NATURAL CONCEPTION**  
(FORMALIZING DISCRIMINATION)

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

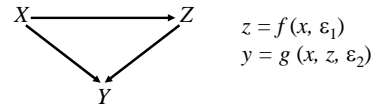
[Carson versus Bethlehem Steel Corp. (70 FEP Cases 921, 7<sup>th</sup> Cir. (1996))]

$x = \text{male}, x' = \text{female}$   
 $y = \text{hire}, y' = \text{not hire}$   
 $z = \text{applicant's qualifications}$

**NO DIRECT EFFECT**

$Y_{x'Z_x} = Y_x, \quad Y_{xZ_{x'}} = Y_{x'}$

**THE OPERATIONAL MEANING OF  
INDIRECT EFFECTS**



"Natural" Indirect Effect of X on Y:  
The expected change in Y when we keep X constant, say at  $x_0$ , and let Z change to whatever value it would have under a unit change in X.

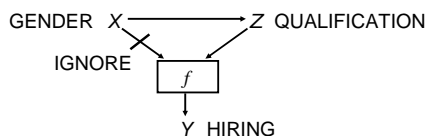
$E[Y_{x_0 Z_{x_1}} - Y_{x_0}]$

In linear models,  $NIE = TE - DE$

**POLICY IMPLICATIONS  
OF INDIRECT EFFECTS**

indirect  
What is the ~~direct~~ effect of X on Y?

The effect of Gender on Hiring if sex discrimination is eliminated.



**SEMANTICS AND IDENTIFICATION  
OF NESTED COUNTERFACTUALS**

Consider the quantity

$Q \triangleq E_u [Y_{xZ_{x^*}}(u)]$

Given  $\langle M, P(u) \rangle$ ,  $Q$  is well defined

Given  $u$ ,  $Z_{x^*}(u)$  is the solution for Z in  $M_{x^*}$ , call it  $z$

$Y_{xZ_{x^*}}(u)$  is the solution for Y in  $M_{xz}$

Can  $Q$  be estimated from  $\left\{ \begin{array}{l} \text{experimental} \\ \text{nonexperimental} \end{array} \right\}$  data?

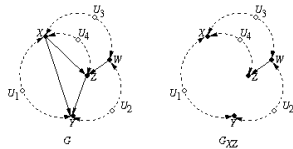
## GRAPHICAL CONDITION FOR EXPERIMENTAL IDENTIFICATION OF AVERAGE NATURAL DIRECT EFFECTS

Theorem: If there exists a set  $W$  such that

$$(Y \perp\!\!\!\perp Z | W)_{G_{XZ}} \text{ and } W \subseteq ND(X \cup Z)$$

$$NDE(x, x^*; Y) = \sum_{w, z} [E(Y_{xz} | w) - E(Y_{x^*z} | w)] P(Z_{x^*} = z | w) P(w)$$

Example:

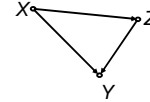


## IDENTIFICATION IN MARKOVIAN MODELS

Corollary 3:

The average natural direct effect in Markovian models is identifiable from nonexperimental data, and it is given by

$$NDE(x, x^*; Y) = \sum_z [E(Y | x, z) - E(Y | x^*, z)] P(Z_{x^*} = z)$$



$$NDE(x, x^*; Y) = \sum_z [E(Y | x, z) - E(Y | x^*, z)] P(z | x^*)$$

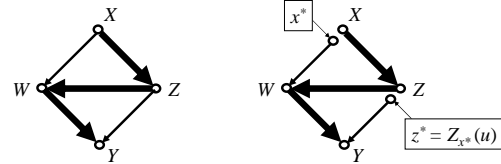
## RELATIONS BETWEEN TOTAL, DIRECT, AND INDIRECT EFFECTS

Theorem 5: The total, direct and indirect effects obey the following equality

$$TE(x, x^*; Y) = NDE(x, x^*; Y) - NIE(x^*, x; Y)$$

In words, the total effect (on  $Y$ ) associated with the transition from  $x^*$  to  $x$  is equal to the difference between the direct effect associated with this transition and the indirect effect associated with the reverse transition, from  $x$  to  $x^*$ .

## GENERAL PATH-SPECIFIC EFFECTS (Def.)



Form a new model,  $M_g$ , specific to active subgraph  $g$

$$f_i^*(pa_i, u; g) = f_i(pa_i(g), pa_i^*(\bar{g}), u)$$

Definition:  $g$ -specific effect

$$E_g(x, x^*; Y)_M = TE(x, x^*; Y)_{M_g^*}$$

Nonidentifiable even in Markovian models

## EFFECT DECOMPOSITION SUMMARY

- Graphical conditions for estimability from experimental / nonexperimental data.
- Graphical conditions hold in Markovian models
- Useful in answering new type of policy questions involving mechanism blocking instead of variable fixing.

## CONCLUSIONS

Structural-model semantics, enriched with logic and graphs, provides:

- Complete formal basis for causal reasoning
- Powerful and friendly causal calculus
- Lays the foundations for asking more difficult questions: What is an action? What is free will? Should robots be programmed to have this illusion?