OUTLINE

1. The causal revolution: From Haavelmo to graphs and counterfactuals
2. The two fundamental laws of causal inference and their mathematical implications
3. Applications:
   • Policy evaluation,
   • Misspecification tests,
   • Mediation,
   • Generalizability (external validity),
   • Latent heterogeneity
   • Missing data
His example:

\[
\begin{align*}
y &= ax + 1 \\
x &= by + 2 \\
ax &= E(Y | X = x)
\end{align*}
\]

What information did the modeler intend \( a \) to carry in Eq. 1 and what information would \( a \) provide if we are able to estimate its value?

Perhaps \( a = g[P(x,y,z\ldots)] \)?

There is no such \( g[\ ] \).

"We have in mind some actual experiment, or some design of an experiment, which we could at least imagine arranging." (Haavelmo, 1944)
TRADITIONAL STATISTICAL INFERENCE PARADIGM

Data \rightarrow \text{Joint Distribution} \rightarrow Q(P) \text{ (Aspects of } P) \rightarrow \text{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 $P'(\text{cancer})$ if we ban smoking.
e.g., Estimate the probability that a customer who bought $A$ would buy $B$ if we were to double the price.
PLATO'S CAVE...
THE STRUCTURAL MODEL PARADIGM

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

“Think Nature, not experiment”
THE STRUCTURAL MODEL PARADIGM

Joint Distribution → Data Generating Model → $Q(M)$ (Aspects of $M$)

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

“A painful de-crowning of a beloved oracle!”
Assume that the Government decides, through public spending, taxation, etc., to keep income, \( r \), at a given level,... the only change in the system being that, instead of

\[
    r_i = u_i + v_i
\]

we now have

\[
    r_i = u_i + v_i + g_i
\]

where \( g_i \) is Government expenditure, so adjusted as to keep \( r \) constant, whatever be \( u \) and \( v \),..."

(Haavelmo, 1943, p. 12)
1. An economic model is a set of hypothetical experiments, qualitatively encoded in a system of equations.

2. An economic model is capable of answering policy intervention questions, with no further assistance from the modeller.

3. There is a formal way of taking an arbitrary model, combining it with data, and derive valid answers to policy questions.
THE EVOLUTION OF CAUSAL CALCULUS

- Haavelmo's surgery (1943)
  Add adjustable force
- Strotz and Wold surgery (1960). “Wipe out” the equation $r_i = u_i + v_i$, and replace it with $r_i = \text{constant}$
- Graphical surgery (Spirtes et al., 1993; Pearl, 1993).
  Wipe out incoming arrows to $r$

\[
P(u, v, r, y) = P(u)P(v)P(r | u, v)P(y | r)
\]

- \textit{do}-calculus (Pearl, 1994)
  \[
P(Y = y | do(r)) \quad \text{New operator} \quad \neq P(Y = y | r)
\]
- Structural counterfactuals (Balke and Pearl, 1995)
  \[
  Y_r(u) = Y(u) \quad \text{in the } r - \text{mutilated} \quad \text{model}
  \]
- Unification with Neyman-Rubin $Y_x(u)$ and Lewis (1973)
Definition: A structural causal model is a 4-tuple $\langle V, U, F, P(u) \rangle$, 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., $y = a + bx + u_Y$
- $P(u)$ is a distribution over $U$

$P(u)$ and $F$ induce a distribution $P(v)$ and a graph $G(v)$ over the observable variables
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:

\[
Y_x(u) = Y_{M_x}(u)
\]
WHAT KIND OF QUESTIONS CAN THE MODEL 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?) $P(y_{A\Box} \mid A)$

• Options:
  “With what probability?”

THE CAUSAL HIERARCHY - SYNTACTIC DISTINCTION
THE TWO FUNDAMENTAL LAWS OF CAUSAL INFERENCE

1. The Law of Counterfactuals

\[ Y_x(u) = Y_{M_x}(u) \]

\(M\) generates and evaluates all counterfactuals.

2. The Law of Conditional Independence \((d\text{-separation})\)

\((X \text{ sep } Y \mid Z)_{G(M)} \quad (X \quad Y \mid Z)_{P(v)}\)

(Separation in the model \(\Rightarrow\) independence in the distribution.)
Reading Counterfactuals from a Structural Model

Data shows: $X = 0.5$, $Z = 1.0$, $Y = 1.9$

A student named Joe, measured $X = 0.5$, $Z = 1.0$, $Y = 1.9$

Q$_1$: What would Joe’s score be had he doubled his study time?

Answer: $Y_{Z=2} = 0.7 \cdot 0.5 + 0.4 \cdot 2.0 + 3 = 2.30$
1. The Law of Counterfactuals

\[ Y_x(u) = Y_{Mx}(u) \]

\((M\) generates and evaluates all counterfactuals.\)

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

\[(X \text{ sep } Y \mid Z)_{G(M)} \quad (X \quad Y \mid Z)_{P(v)}\]

\((\text{Separation in the model } \Rightarrow \text{ independence in the distribution.})\)
THE LAW OF CONDITIONAL INDEPENDENCE

Each function summarizes millions of micro processes.
Each function summarizes millions of micro processes.

Still, if the $U$'s are independent, the observed distribution $P(C,R,S,W)$ must satisfy certain constraints that are:

1. independent of the $f$'s and of $P(U)$
2. can be read from the structure of the graph.
Every missing arrow advertises an independency, conditional on a separating set.

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

Applications:
1. Model testing
2. Structure learning
3. Reducing "what if I do" questions to symbolic calculus
4. Reducing scientific questions to symbolic calculus
THE FIVE NECESSARY STEPS FOR CAUSAL INFERENCE

Define: Express the target quantity $Q$ as a property of the model $M$.

\[ ATE = E(Y_1 - Y_0) = E(Y \mid do(X = 1)) - E(Y \mid do(X = 0)) \]

Assume: Express causal assumptions in structural or graphical form.

Identify: Determine if $Q$ is identifiable.

Estimate: Estimate $Q$ if it is identifiable; approximate it, if it is not.

Test: If $M$ has testable implications
THE FIVE NECESSARY STEPS FOR CAUSAL INFERENCE

Define: Express the target quantity $Q$ as a property of the model $M$.

$$ATE = E(Y_1 - EY_0) = E(Y_1 | do(X = 1)) - E(Y | do(X = 0))$$

Assume: Express causal assumptions in structural or graphical form.

Identify: Determine if $Q$ is identifiable.

Estimate: Estimate $Q$ if it is identifiable; approximate it, if it is not.

Test: If $M$ has testable implications
THE FIVE NECESSARY STEPS FOR CAUSAL INFERENCE

Define: Express the target quantity $Q$ as a property of the model $M$.

\[ ETN = E(Y_0 = Y_0 | X = 1, Y = 1) \]

Assume: Express causal assumptions in structural or graphical form.

Identify: Determine if $Q$ is identifiable.

Estimate: Estimate $Q$ if it is identifiable; approximate it, if it is not.

Test: If $M$ has testable implications
FORMULATING ASSUMPTIONS
THREE LANGUAGES

1. English: Smoking \((X)\), Cancer \((Y)\), Tar \((Z)\), Genotypes \((U)\)

\[
\begin{align*}
Z_x(u) &= Z_yx(u), \\
X_y(u) &= X_zy(u) = X_z(u) = X(u), \\
Y_z(u) &= Y_zx(u), \\
Z_x &\{Y_z, X\}
\end{align*}
\]

2. Counterfactuals:

\[
\begin{align*}
x &= f_1(u, 1) \\
y &= f_3(z, u, 3) \\
z &= f_2(x, 2) \quad 1 \quad 2 \quad 3
\end{align*}
\]

Not too friendly:
Consistent?, complete?, redundant?, plausible?, testable?
ILLUSTRATING THE WORKING OF CAUSAL CALCULUS

Model 2 (Linear version)

\[ Y = g(W_3, Z_3, W_2, u) + u \]

\[ W_3 = g_3(X_1, u_3') \]

\[ Z_3 = g_3(Z_1 + Z_2 Z_2') + u_3' \]

\[ W_2 = g_2(Z_2, u_2') \]

\[ X = g_1(W_1, Z_2, Z_3') + u'' \]

\[ W_1 = g_1(Z_1, u_1') \]

\[ Z_1 = f_1(u_1) \]

\[ Z_2 = f_2(u_2) \]

\( U' \) s are mutually independent
WHAT ARE THE TESTABLE IMPLICATIONS OF THE MODEL?

Model 1

\[ Y = f(W_3, Z_3, W_2, u) \]
\[ W_3 = g_3(X, u''_3) \]
\[ Z_3 = f_3(Z_1, Z_2, u_3) \]
\[ W_2 = g_2(Z_2, u'_2) \]
\[ X = g(W_1, Z_3, u'') \]
\[ W_1 = g_1(Z_1, u'_1) \]
\[ Z_1 = f_1(u_1) \]
\[ Z_2 = f_2(u_2) \]

Missing edges:
\[ Z_1 - Z_2, Z_1 - Y, Z_2 - X \ldots \]

Separating sets:
\{0\}, \{X, Z_2, Z_3\}, \{Z_1, Z_3\} \ldots

Testable implications (FINITE!):
\[ Z_1 \rightarrow Y \mid \{X_1, Z_2, Z_3\}, \]
\[ Z_2 \rightarrow X \mid \{Z_1, Z_3\}. \]

Zero regression coefficients:
\[ r_{\{Z_1 Z_2\} Z_2} = 0 \]
\[ r_{\{Y Z_1 \cdot X Z_2 Z_3\}} = 0 \]

These imply *all* misspecification tests
Question 4:
If we regress $Z_1$ on all other variables in the model, which regression coefficient must be zero?

Answer: All but these three

Question 5:
If we regress $Z_1$ on $Z_3$ and $W_1$, which regression coefficient might change if we add $Y$ as a regressor?

Answer:
The coefficient of $Z_3$ might change and the coefficient of $W_1$ must remain invariant. Non-invariance may not be misspecification.
Suppose we wish to estimate the average causal effect of $X$ on $Y$:

$$ACE = P(Y = y \mid do(X = 1)) - P(Y = y \mid do(X = 0)).$$

- Which subsets of variables need to be adjusted to obtain an unbiased estimate of $ACE$?
- Is there a single variable that, if measured, would allow an unbiased estimate of $ACE$?
EFFECT OF WARM-UP ON INJURY (After Shrier & Platt, 2008)

Watch out!

Frosty Door

Warm-up Exercises (X) → Intra-game proprioception → Injury (Y)

Team motivation, aggression

Pre-game Proprioception

Previous Injury

Coach

Fitness Level

Neuromuscular fatigue

Connective Tissue Disorder

Tissue Weakness

Genetics

Contact Sport

Injury (Y)

No, no!
Rule 1: Ignoring observations
\[ P(y \mid \text{do}\{x\}, z, w) = P(y \mid \text{do}\{x\}, w) \]
if \((Y \uparrow Z \mid X, W)_{G_X}\)

Rule 2: Action/observation exchange
\[ P(y \mid \text{do}\{x\}, \text{do}\{z\}, w) = P(y \mid \text{do}\{x\}, z, w) \]
if \((Y \uparrow Z \mid X, W)_{G_{XZ}}\)

Rule 3: Ignoring actions
\[ P(y \mid \text{do}\{x\}, \text{do}\{z\}, w) = P(y \mid \text{do}\{x\}, w) \]
if \((Y \uparrow Z \mid X, W)_{G_{XZ(W)}}\)

Sound and complete
**DERIVATION IN CAUSAL CALCULUS**

\[
P(c \mid do\{s\}) = tP(c \mid do\{s\}, t)P(t \mid do\{s\})
\]

= \[tP(c \mid do\{s\}, do\{t\})P(t \mid do\{s\})\] \hspace{1cm} \text{Rule 2}

= \[tP(c \mid do\{s\}, do\{t\})P(t \mid s)\] \hspace{1cm} \text{Rule 2}

= \[tP(c \mid do\{t\})P(t \mid s)\] \hspace{1cm} \text{Rule 3}

= \[s' \times tP(c \mid do\{t\}, s')P(s' \mid do\{t\})P(t \mid s)\] \hspace{1cm} \text{Probability Axioms}

= \[s' \times tP(c \mid t, s')P(s' \mid do\{t\})P(t \mid s)\] \hspace{1cm} \text{Rule 2}

= \[s' \times tP(c \mid t, s')P(s')P(t \mid s)\] \hspace{1cm} \text{Rule 3}
WHAT ELSE CAN CAUSAL CALCULUS DO FOR US?

- Equivalent models
- Identifying Counterfactual queries (ETT, PC)
- Causes of Effects
- Mediation (Victory II)
- External validity (Victory III)

Finding instruments
Is there an instrumental variable for the $Z_3 \rightarrow Y$ relationship?

Answer: No
Can we turn $Z_1$ into an IV?
Answer: Yes, condition on $W_1$. 
THE STRUCTURAL-COUNTERFACTUAL SYMBIOSIS

1. Express theoretical assumptions in structural language.
2. Express queries in counterfactual language.
3. Translate (1) into (2) for algebraic analysis, or (2) into (1) for graphical analysis.
4. Use either graphical or algebraic machinery to answer the query in (2).
MEDIATION: A SYMBIOTIC TRIUMPH

Why decompose effects?
1. To understand how Nature works
2. To comply with legal requirements
3. To predict the effects of new type of interventions:
   - Signal re-routing and mechanism deactivating,
   - rather than variable fixing
COUNTERFACTUAL DEFINITION OF INDIRECT EFFECTS

\[ z = f(x, u) \]
\[ y = g(x, z, u) \]

No Controlled Indirect Effect

Robins-Greenland (1992) and Pearl (2001)

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 attained had \( X \) changed to \( x_1 \).

\[ E[Y_{x_0}Z_{x_1} \mid Y_{x_0}] \]
What is the indirect effect of $X$ on $Y$?

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

Deactivating a link – a new type of intervention
THE MEDIATION FORMULAS IN UNCONFOUNDED MODELS

\[ DE = \left[ E(Y \mid x_1, z) - E(Y \mid x_0, z) \right] P(z \mid x_0) \]

\[ IE = \left[ E(Y \mid x_0, z) \right] \left[ P(z \mid x_1) - P(z \mid x_0) \right] \]

\[ TE = E(Y \mid x_1) - E(Y \mid x_0) \]

\[ z = f(x, u_1) \]

\[ y = g(x, z, u_2) \]

\[ u_1 \text{ independent of } u_2 \]

Fraction of responses explained by mediation (sufficient)

Fraction of responses owed to mediation (necessary)
THE MEDIATION FORMULAS IN UNCONFOUNDED MODELS

\[ D_E = [E(Y \mid x_1, z) - E(Y \mid x_0, z)]P(z \mid x_0) \]
\[ I_E = [E(Y \mid x_0, z)[P(z \mid x_1) - P(z \mid x_0)] \]
\[ T_E = E(Y \mid x_1) - E(Y \mid x_0) \]

\[ z = f(x, u_1) \]
\[ y = g(x, z, u_2) \]
\[ u_1 \text{ independent of } u_2 \]

\( DE + IE \) (sufficient)

Complete identification conditions for confounded models with multiple mediators.

\( T_E = \text{Fraction of responses owed to mediation} \) (necessary)
WHAT CAN MEDIATION FORMULA DO FOR PARAMETRIC ANALYSTS?

What combination of parameters gives the effect mediated by \( M \)?

\[
IE(M) = 1(1 + 2)
\]

What combination of parameters gives the effect owed to \( M \)?

\[
TE - DE(M) = (1 + 3)(1 + 2)
\]

\[
y = 1m + 2x + 3xm + 4w + u_1
\]

\[
m = 1x + 2w + u_2
\]

\[
w = x + u_3
\]
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?
Extrapolation across studies requires “some understanding of the reasons for the differences.” (Cox, 1958)

“`External validity’ asks the question of generalizability: To what population, settings, treatment variables, and measurement variables can this effect be generalized?” (Shadish, Cook and Campbell, 2002)

“An experiment is said to have “external validity” if the distribution of outcomes realized by a treatment group is the same as the distribution of outcome that would be realized in an actual program.” (Manski, 2007)

"A threat to external validity is an explanation of how you might be wrong in making a generalization." (Trochin, 2006)
MOVING FROM THE LAB TO THE REAL WORLD . . .

Real world

Everything is assumed to be the same, trivially transportable!

Lab

Everything is assumed to be the different, not transportable!
MOTIVATION
WHAT CAN EXPERIMENTS IN LA TELL ABOUT NYC?

Experimental study in LA
Measured:
\[ P(x, y, z) \]
\[ P(y \mid do(x), z) \]

Needed:
\[ P^*(y \mid do(x)) = \text{?} = P(y \mid do(x), z)P^*(z) \]

Transport Formula (calibration):
\[ F(P, P_{do}, P^*) \]

Observational study in NYC
Measured:
\[ P^*(x, y, z) \]
\[ P^*(z) \neq P(z) \]

(Intervention) (Outcome) (Observation) (Outcome)

X (Intervention) Y (Outcome) Z (Age)
X Y Z (Age)

\[ \sum_z P(y \mid do(x), z)P^*(z) \]
TRANSPORT FORMULAS DEPEND ON THE STORY

(a) \( Z \) represents age

\[
P^*(y \mid do(x)) = P(y \mid do(x), z)P^*(z)
\]

(b) \( Z \) represents language skill

\[
P^*(y \mid do(x)) = P(y \mid do(x))
\]

Factors producing differences
TRANSPORT FORMULAS DEPEND ON THE STORY

a) \( Z \) represents age
\[
P^* (y \mid do(x)) = P(y \mid do(x), z)P^*(z)
\]
b) \( Z \) represents language skill
\[
P^* (y \mid do(x)) = P(y \mid do(x))
\]
c) \( Z \) represents a bio-marker
\[
P^* (y \mid do(x)) = ? P(y \mid do(x), z)P^*(z \mid x)
\]
GOAL: ALGORITHM TO DETERMINE IF AN EFFECT IS TRANSPORTABLE

INPUT: Annotated Causal Graph

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 $\square$ to $\square^*$ if and only if it is reducible, using the rules of $do$-calculus, to an expression in which $S$ is separated from $do(\ )$.

$$R(\ *) = P^*(y \mid do(x)) = P(y \mid do(x), s)$$
$$ = P(y \mid do(x), s, w)P(w \mid do(x), s)$$
$$ = P(y \mid do(x), w)P(w \mid s)$$
$$ = P(y \mid do(x), w)P^*(w)$$
RESULT: ALGORITHM TO DETERMINE IF AN EFFECT IS TRANSPORTABLE

INPUT: Annotated Causal Graph

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)

\[ P^* (y | do(x)) = \]

\[
P(y | do(x), z) \quad P^* (z | w) \quad P(w | do(w), t) P^* (t)
\]
WHICH MODEL LICENSES THE TRANSPORT OF THE CAUSAL EFFECT $X \rightarrow Y$

**Diagram Description**

- **(a)** External factors creating disparities create disparities.
- **(b)** External factors creating disparities create disparities.
- **(c)** External factors creating disparities create disparities.
- **(d)** External factors creating disparities create disparities.
- **(e)** External factors creating disparities create disparities.
- **(f)** External factors creating disparities create disparities.

**External factors creating disparities**

- Yes
- No

**Diagram Notes**

- The arrows indicate the direction of transport and causality.
- The yellow square represents the external factor creating disparities.
Why should we transport statistical information?

i.e., Why not re-learn things from scratch?

1. **Measurements** are costly.
   Limit measurements to a subset $V^*$ of variables called “scope”.

2. **Samples** are scarce.
   Pooling samples from diverse populations will improve precision, if differences can be filtered out.
**STATISTICAL TRANSPORTABILITY**

**Definition: (Statistical Transportability)**
A statistical relation $R(P)$ is said to be *transportable* from $\Box$ to $\Box^*$ over $V^*$ if $R(P^*)$ is identified from $P, P^*(V^*)$, and $D$ where $P^*(V^*)$ is the marginal distribution of $P^*$ over a subset of variables $V^*$.

$R=P^*(y \mid x)$ is transportable over $V^* = \{X,Z\}$, i.e., $R$ is estimable without re-measuring $Y$

$$R = P^*(z \mid x)P(z \mid y)$$

**Transfer Learning**
If few samples ($N_2$) are available from $\Box^*$ and many samples ($N_1$) from $\Box$ then estimating $R = P^*(y \mid x)$ by

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

achieves a much higher precision.
META-ANALYSIS OR
MULTI-SOURCE LEARNING

Target population

\[ R = P^* (y \mid do(x)) \]
**R** = *P*(y | do(x))

Is R identifiable from (d) and (h)?

\[
R = \frac{P(h)(y | do(x), w)P(d)(w | do(x))}{w} \cdot \frac{P(h)(y | do(x), w)P(d)(w | x)}{w}
\]

*R(∏*) is identifiable from studies (d) and (h).

*R(∏*) is not identifiable from studies (d) and (i).
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.
Theorem
{\Pi_1, \Pi_2, \ldots, \Pi_K} – a set of studies.
{D_1, D_2, \ldots, D_K} – selection diagrams (relative to \textquotedblleft\textquotedblright). A relation \( R(\square^*) \) 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 \( D_k \).

Sufficient and necessary condition established (Bareinboim & Pearl, 2013).
Principle 1: Calibrate estimands before pooling (to minimize bias)

Principle 2: Decompose to sub-relations before calibrating (to improve precision)
BIAS VS. PRECISION IN META-SYNTHESIS

\[
P^*_{(i,d,k)}(y \mid do(x)) = \sum_{w} P^*_{(h)}(y \mid w, do(x))P^*_{(i,d)}(w \mid do(x))
\]
Latent heterogeneity:
Nameless groups that react differently to treatments or policies.

Health science:
Is the drug uniformly beneficial, or kills some and saves more?

Social science:
Do those who have access to a program benefits most from it?
ASSESSING POPULATION HETEROGENEITY

COVARIATE-INDUCED HETEROGENEITY

c-specific effect: $E(Y_1 - Y_2 \mid C = c)$

Effect difference

$$D(c_i, c_j) = \left| E(Y_1 - Y_0 \mid C = c_i) - E(Y_1 - Y_0 \mid C = c_j) \right|$$

Lower Bound on population heterogeneity

$$LB = \max_{(c_i, c_j)} D(c_i, c_j)$$
When is the $c$-specific effect identifiable? $E(Y_1 - Y_2 \mid C = c)$
ASSESSING POPULATION HETEROGENEITY

Latent Heterogeneity

No $c$-specific effect available
Assess $TT - TUT$

$TT = E(Y_1 - Y_0 | X = 1)$
$TUT = E(Y_1 - Y_0 | X = 0)$
When is $TT-TUT$ identifiable?

1. When $X$ is binary, and $(E(Y_1), E(Y_0))$ are identifiable by some method (e.g., randomized trials)

2. When $X$ is arbitrary, and $ATE$ is identifiable by adjustment for a sufficient set of covariates

3. When $ATE$ is identified through mediating instruments
2. When $X$ is arbitrary, and $ATE$ is identifiable by adjustment for a sufficient set $Z$ of covariates

$$E(Y_{x} \mid x') = \sum_{z} E(Y \mid x, z)P(z \mid x')$$
3. When $ATE$ is identified through mediating instruments

\[ E(Y_x | x') = E(Y | z, x')P(z | x) \]
CONDITIONS THAT UNRAVEL LATENT HETEROGENEITY (TT ≠ TUT)

1. RCT with binary treatment
   \[ p[E(Y \mid X = 1) \cdot E(Y_1)] - (1 - p)[E(Y \mid X = 0) \cdot E(Y_0)] \]

1. Observational studies with admissible covariates Z
   \[ TT - TUT = z\left[E(Y \mid X = x', z) - E(Y \mid X = x, z)\right] \]
   \[ z\left[P(z \mid X = x') - P(z \mid X = x)\right] = 0 \]

1. Observational studies with mediating instruments Z
   \[ TT - TUT = z\left[E(Y \mid X = x', z) - E(Y \mid X = x, z)\right] \]
   \[ z\left[P(z \mid X = x') - P(z \mid X = x)\right] = 0 \]
MISSING DATA: A SEEMINGLY STATISTICAL PROBLEM (Mohan & Pearl, 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.
- Consequence: Like Alchemy before Boyle and Dalton, the field is craving for (1) theoretical guidance and (2) performance guarantees.
**ESTIMATE** \( P(X,Y,Z) \)

<table>
<thead>
<tr>
<th>Sample #</th>
<th>Observations</th>
<th>Missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X* Y* Z*</td>
<td>Rx Ry Rz</td>
</tr>
<tr>
<td>1</td>
<td>1 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>1 0 1</td>
<td>0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>1 m m</td>
<td>0 1 1</td>
</tr>
<tr>
<td>4</td>
<td>0 1 m</td>
<td>0 0 1</td>
</tr>
<tr>
<td>5</td>
<td>m 1 m</td>
<td>1 0 1</td>
</tr>
<tr>
<td>6</td>
<td>m 0 1</td>
<td>1 0 0</td>
</tr>
<tr>
<td>7</td>
<td>m m 0</td>
<td>1 1 0</td>
</tr>
<tr>
<td>8</td>
<td>0 1 m</td>
<td>0 0 1</td>
</tr>
<tr>
<td>9</td>
<td>0 0 m</td>
<td>0 0 1</td>
</tr>
<tr>
<td>10</td>
<td>1 0 m</td>
<td>0 0 1</td>
</tr>
<tr>
<td>11</td>
<td>1 0 1</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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?

- None of these questions can be answered by statistical characterization of the problem.
- All can be answered using causal models.
MISSING DATA: TWO PERSPECTIVES

Causal inference is a missing data problem.  
(Rubin 2012)

Missing data is a causal inference problem.  
(Pearl 2012)

Why is missingness a causal problem?

• Which mechanism causes missingness makes a difference in whether / how we can recover information from the data.

• Mechanisms require causal language to be properly described – statistics is not sufficient.

• Different causal assumptions lead to different routines for recovering information from data, even when the assumptions are indistinguishable by any statistical means.
### ESTIMATE \( P(X,Y,Z) \)

<table>
<thead>
<tr>
<th>Sample #</th>
<th>Observations</th>
<th>Missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( X^* )</td>
<td>( R_x )</td>
</tr>
<tr>
<td>1</td>
<td>1 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>1 0 1 1</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>1 m m m</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>4</td>
<td>0 1 m m</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>5</td>
<td>m 1 m m</td>
<td>1 0 1 1</td>
</tr>
<tr>
<td>6</td>
<td>m 0 1 m</td>
<td>1 0 0 0</td>
</tr>
<tr>
<td>7</td>
<td>m m 0 m</td>
<td>1 1 0 0</td>
</tr>
<tr>
<td>8</td>
<td>0 1 m m</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>9</td>
<td>0 0 0 m m</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>10</td>
<td>1 0 m m</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>11</td>
<td>1 0 1 1</td>
<td>0 0 0 0</td>
</tr>
</tbody>
</table>

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

### Missingness graph

- **Graph Description**
  - Nodes: \( X \) and \( X^* \)
  - Edges: \( R_X \)
  - \( X \) and \( X^* \) are connected.

- **Equation**
  \[
  X^* = \begin{cases} 
  X & \text{if } R_X = 0 \\
  m & \text{if } R_X = 1
  \end{cases}
  \]
### NAIVE ESTIMATE OF $P(X,Y,Z)$

#### Complete Cases

<table>
<thead>
<tr>
<th>Row #</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>$R_x$</th>
<th>$R_y$</th>
<th>$R_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Line deletion estimate is generally biased.

\[
P(X,Y,Z) = P(X,Y,Z | R_x = 0, R_y = 0, R_z = 0)
\]

<table>
<thead>
<tr>
<th>Sample #</th>
<th>Observations</th>
<th>Missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X^*$</td>
<td>$Y^*$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>m</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>m</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>m</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- MCAR
SMART ESTIMATE OF \( P(X,Y,Z) \)

\[
P(X,Y,Z) = P(Z \mid X,Y)P(X \mid Y)P(Y)
\]

\[
P(Y) = P(Y \mid R_y = 0)
\]

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

\[
P(Z \mid X,Y) = P(Z \mid X,Y, R_x = 0, R_y = 0, R_z = 0)
\]
SMART ESTIMATE OF $P(X,Y,Z)$

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

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

$P(Y \mid R_y = 0)$

<table>
<thead>
<tr>
<th>Sample #</th>
<th>$X^*$</th>
<th>$Y^*$</th>
<th>$Z^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>$m$</td>
<td>$m$</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>$m$</td>
</tr>
<tr>
<td>5</td>
<td>$m$</td>
<td>1</td>
<td>$m$</td>
</tr>
<tr>
<td>6</td>
<td>$m$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>$m$</td>
<td>$m$</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>$m$</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>$m$</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>$m$</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
**SMART ESTIMATE OF $P(X,Y,Z)$**

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

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

\[ P(Y \mid R_y = 0) \]

<table>
<thead>
<tr>
<th>Sample #</th>
<th>$X^*$</th>
<th>$Y^*$</th>
<th>$Z^*$</th>
<th>Compute $P(Y \mid R_y = 0)$</th>
<th>Row #</th>
<th>$Y^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>m</td>
<td>m</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>m</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>m</td>
<td>1</td>
<td>m</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>m</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>m</td>
<td>m</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>m</td>
<td>8</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>m</td>
<td>9</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>m</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## 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 # | $X^*$ | $Y^*$ | $Z^*$ | Compute $P(Y \mid R_y = 0)$ | \multicolumn{2}{c|}{Row #} | \multicolumn{2}{c|}{$Y^*$} | Compute $P(X \mid Y, R_x = 0, R_y = 0)$ | \multicolumn{2}{c|}{Row #} | \multicolumn{2}{c|}{$X^*$} | \multicolumn{2}{c|}{$Y^*$} |
|----------|-------|-------|-------|-----------------------------|----------------|--------|----------------|-----------------------------|----------------|--------|-------|-------|--------|-------|
| 1        | 1     | 0     | 0     |                             | 1              | 0      |                |                             |                |        |        |        |        |        |
| 2        | 1     | 0     | 1     |                             | 2              | 0      |                |                             |                |        |        |        |        |        |
| 3        | 1     | m     | m     |                             | 4              | 1      |                |                             |                |        |        |        |        |        |
| 4        | 0     | 1     | m     |                             | 6              | 0      |                |                             |                |        |        |        |        |        |
| 5        | m     | 1     | m     |                             | 8              | 1      |                |                             |                |        |        |        |        |        |
| 6        | m     | 0     | 1     |                             | 10             | 0      |                |                             |                |        |        |        |        |        |
| 7        | m     | m     | 0     |                             | 11             | 0      |                |                             |                |        |        |        |        |        |
| 8        | 0     | 1     | m     |                             |                |        |                |                             |                |        |        |        |        |        |
| 9        | 0     | 0     | m     |                             |                |        |                |                             |                |        |        |        |        |        |
| 10       | 1     | 0     | m     |                             |                |        |                |                             |                |        |        |        |        |        |
| 11       | 1     | 0     | 1     |                             |                |        |                |                             |                |        |        |        |        |        |
| -        |       |       |       |                             |                |        |                |                             |                |        |        |        |        |        |
SMART ESTIMATE OF $P(X,Y,Z)$

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

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

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

<table>
<thead>
<tr>
<th>Sample #</th>
<th>$X^*$</th>
<th>$Y^*$</th>
<th>$Z^*$</th>
<th>Compute $P(Y \mid R_y = 0)$</th>
<th>Row #</th>
<th>$Y^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>m</td>
<td>m</td>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>m</td>
<td></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>m</td>
<td>1</td>
<td>m</td>
<td></td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>m</td>
<td>0</td>
<td>1</td>
<td></td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>m</td>
<td>m</td>
<td>0</td>
<td></td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>m</td>
<td></td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>m</td>
<td></td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>m</td>
<td></td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Compute $P(X \mid Y, R_x = 0, R_y = 0)$</th>
<th>Row #</th>
<th>$X^*$</th>
<th>$Y^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ P(X,Y,Z) = P(Z \mid X,Y)P(X \mid Y)P(Y) \]

\[ P(Y) = P(Y \mid R_y = 0) \]

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

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

\[ P(X,Y,Z) \text{ is not recoverable} \]
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.)

Theorem:
$Q$ is recoverable iff it is decomposable into terms of the form $Q_j = P(S_j \mid T_j)$, such that:
1. For each variable $V$ that is in $T_j$, $R_V$ is also in $T_j$.
2. For each variable $V$ that is in $S_j$, $R_V$ is either in $T_j$ or in $S_j$.

e.g.,

$$Q_j = P(Y, X, R_Y \mid R_X, Z, R_Z)$$
AN IMPOSSIBILITY THEOREM FOR MISSING DATA

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

In (a): $P(X) = P(X^*|Y)P(Y)$, while

in (b): $P(X) = P(X^*)$

No universal algorithm exists that produces an unbiased estimate whenever such exists.
CONCLUSIONS

- **Counterfactuals**, the building blocks of scientific thought, are encoded meaningfully and conveniently in structural models.
- **Identifiability** and **testability** of counterfactuals in recursive NP-models are **solved problems**.
- The counterfactual-graphical **symbiosis** has led to major advances in the empirical sciences, including policy evaluation, mediation analysis, generalizability, credit-blame determination, missing data, and heterogeneity.
- **THIS IS NOT AN “APPROACH” BUT AN INDISPENSIBLE SCIENTIFIC TOOL.**
Thank you
For $x = (0,1)$ and $x' = 1-x$, we have:

$$E(Y_x) = E(Y_x \mid X = x')P(X = x') + E(Y_x \mid X = x)(1 - P(X = x'))$$

$$= E(Y_x \mid X = x')p + E(Y \mid X = x)(1 - p)$$

Conclusion:

$E(Y_x \mid X=x')$ is identifiable from $\{E(Y_x), E(Y \mid X=x), p\}$

e.g.

$$E(Y_0 \mid X = 1) = E(Y_0) / p \quad E(Y \mid X = 0)(1 - p) / p$$

$$E(Y_1 \mid X = 0) = E(Y_0) / (1 - p) \quad E(Y \mid X = 1)p / (1 - p)$$

TT AND TUT ARE IDENTIFIABLE FROM RCT WITH BINARY TREATMENTS
NO ESTIMATION WITHOUT CAUSATION

Two ways of estimating $\text{cov}(X,Y)$

1. $\hat{\text{cov}}_1(X,Y) = \hat{\text{var}}(X) \times \hat{YX}$
   
   $= \hat{\text{var}}(X) \left[ \text{on observed } X \right] \times \hat{YX} \left[ \text{on observed } (X,Y) \right]$  

2. $\hat{\text{cov}}_2(X,Y) = \hat{\text{var}}(Y) \times \hat{XY}$

   $= \hat{\text{var}}(Y) \left[ \text{on observed } Y \right] \times \hat{XY} \left[ \text{on observed } (X,Y) \right]$  

- $\hat{\text{cov}}_1(X,Y) \rightarrow ? \hat{\text{cov}}_2(X,Y)$
- Generally NO! So which one is correct, if any?
NO ESTIMATION WITHOUT CAUSATION (cont)

- $\widehat{\text{cov}}_1(X,Y) \rightarrow ? \widehat{\text{cov}}_2(X,Y)$
- Generally NO! So, which one is correct, if any?

Answer: It depends on the causal model

(a) $X \rightarrow Y \rightarrow L \rightarrow R_Y \leftarrow R_X$
(b) $X \rightarrow R_X \rightarrow Y \rightarrow R_Y \leftarrow L

(a) and (b) are statistically indistinguishable, yet

$\widehat{\text{cov}}_1(X,Y)$ is unbiased in (b), and biased in (a)
$\widehat{\text{cov}}_2(X,Y)$ is unbiased in (a), biased in (b)