Causal Data Science for Business Decision Making

Graphical Causal Models I

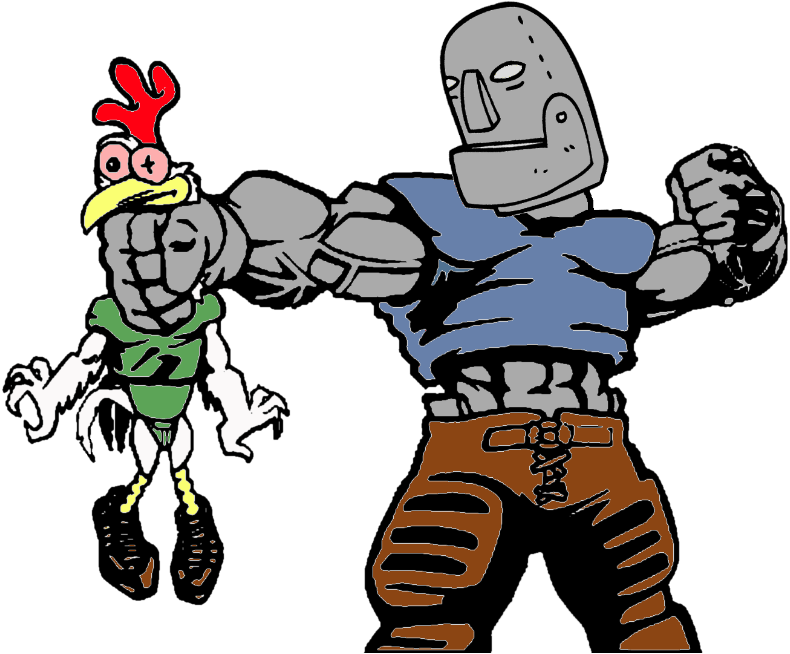
Paul Hünermund

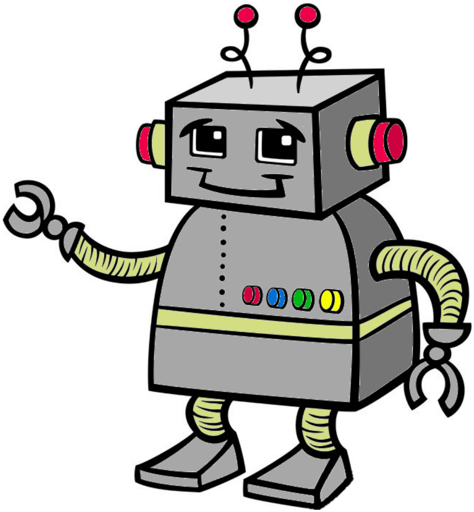
COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Historical Background

- Causal inference is arguably the most important goal in econometrics
  - Inform policy-makers, legislators, and managers about the likely impact of their actions by uncovering quantitative relationships in statistical data
- Since the end of the 1980s, an extensive literature on causal inference was developed in the computer science and artificial intelligence field
  - Builds on the graph-theoretic approach to causality developed by
  - Interest emerged from older AI techniques such as Markov random fields and Bayesian nets
  - Shares several mutual intellectual roots with econometrics
- Why do AI scholars care about causality?

How can we prevent a future robot from trying to make the rooster crow at 3am in order to make the sun come up?





“Beyond Curve Fitting” in Machine Learning and AI

“To Build Truly Intelligent Machines, Teach Them Cause and Effect”

— Judea Pearl

- The notion of causality is a fundamental concept in human thinking
- Current ML / AI techniques remain purely prediction-based
- In other words: machine learning is very good at high-dimensional pattern recognition (“is this a cat or dog?)
- But nothing in the theoretical basis of ML allows to capture the asymmetry inherent to causal relationships
- If we want machines to be able to interact meaningfully with us, they should be equipped with a notion of cause and effect

"If you can't explain it to a ~~six year old,~~ ROBOT

you don't understand it yourself."

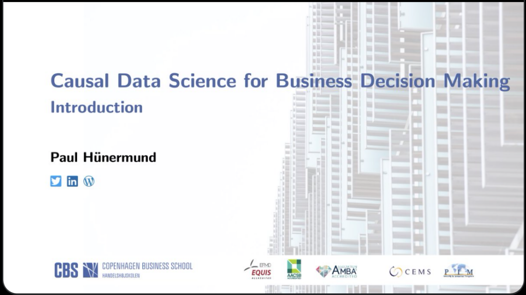
— attributed to *Albert Einstein*

Judea Pearl @yudapearl

If "Data Science" was truly data science (as defined eg here ucla.in/3iEDRVo), you wouldn't need to add "Causal" ahead of the title. But, given what it is today, this is an effective way of telling students: "This is not another function-fitting class."

Tweet übersetzen

> **Paul Hünermund** @PHuenermund · 14 Std.
>
> First session of my new elective in the @CBScph master program today. What a great feeling to be back in the class room and teach the stuff you're really excited about!



10:50 vorm. · 3. Sep. 2021 · Twitter Web App

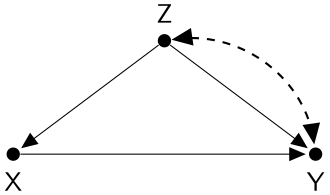
Structural Causal Models

$$
\begin{aligned}
z &\leftarrow f_Z(u_z) \\
x &\leftarrow f_X(z, u_x) \\
y &\leftarrow f_Y(x, z, u_Y)
\end{aligned}
$$

- The f_i's denote the causal mechanisms in the model
  - Are not restricted to be linear as in traditional structural equation models
- The u_i's refer to background factors that are determined outside of the model
- Assignment operator ($\leftarrow$) captures asymmetry of causal relationships
  - $x \leftarrow a \cdot z \;\neq\; z \leftarrow x/a$
- Similar to definition of "structure" according to Cowles foundation

Directed Acyclic Graphs

$$z \leftarrow f_Z(u_z)$$
$$x \leftarrow f_X(z, u_x)$$
$$y \leftarrow f_Y(x, z, u_Y)$$



- In a fully specified SCM, every counterfactual quantity is computable
- In most social science contexts it's hard to know the causal mechanisms f_i and distribution of background factors $P(U)$
- Therefore, restrict attention to qualitative causal information of the model, which can be encoded by a graph G
  - Nodes V: variables in the model
  - Directed edges E: causal relationships in the model

Directed Acyclic Graphs

- No functional form or distributional assumptions means that framework remains fully nonparametric
  - Particularly helpful in fields where theory is purely qualitative and no shape restrictions on can be derived
- $Z \dashleftarrow\dashrightarrow Y$ is a shortcut notation for unobserved common causes $Z \leftarrow U \rightarrow Y$
- Acyclicity
  - Directed cycles such as $A \rightarrow B \rightarrow C \rightarrow A$ are excluded
  - This means there are no feedback loops
  - Otherwise A could be a cause of itself
  - Gives rise to what economists call a *recursive* model
  - Extensions of the SCM framework to cyclic graphs exist

D-Separation

- DAGs are such a useful tool because they are able to efficiently encode conditional independence relationships:

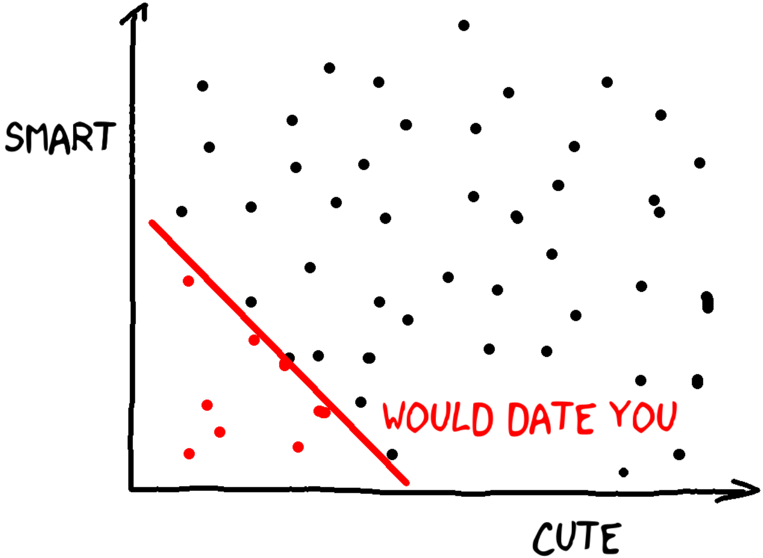
$$\begin{array}{llcl}
\underline{\text{Chain:}} & A \rightarrow B \rightarrow C & \Rightarrow & A \not\perp\!\!\!\perp C \ \text{ and } \ A \perp\!\!\!\perp C|B \\
\underline{\text{Fork:}} & A \leftarrow B \rightarrow C & \Rightarrow & A \not\perp\!\!\!\perp C \ \text{ and } \ A \perp\!\!\!\perp C|B \\
\underline{\text{Collider:}} & A \rightarrow B \leftarrow C & \Rightarrow & A \perp\!\!\!\perp C \ \text{ and } \ A \not\perp\!\!\!\perp C|B
\end{array}$$

- Independence: knowledge that B occurred gives no additional information about the probability of A

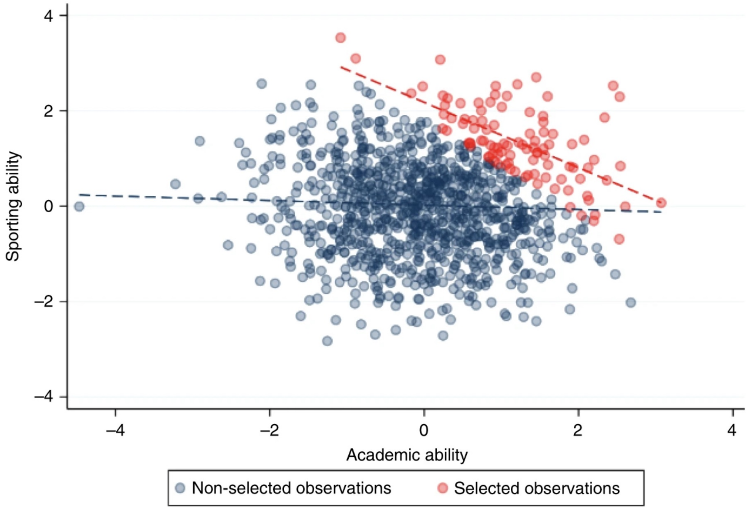
$$A \perp\!\!\!\perp B \Rightarrow P(A|B) = P(A)$$
$$A \perp\!\!\!\perp B|C \Rightarrow P(A|B, C) = P(A|C)$$

- The same holds for longer paths in the graph
  - Conditioning on a variable along a chain or fork blocks (*"d-separates"*) the path
  - Conditioning on a collider opens the path

Collider Bias Example



Collider Bias Example II



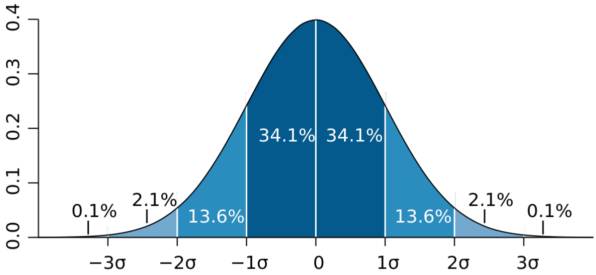
Source: "Collider bias undermines our understanding of COVID-19 disease risk and severity" (2020, Nature Comm.)

Refresher: Variance and Standard Deviation

- The variance of a variable X is a measure for how "spread out" values of X are

$$Var(X) = E((X - \mu)^2)$$

  where $\mu = E(X)$
- The standard deviation σ_X is defined as the square root of the variance



Refresher: Covariance and Correlation

- The covariance between two variables X and Y measures the degree to which they *covary*

$$\sigma_{XY} = E[(X - E(X)(Y - E(Y))]$$

- More specifically, it measures the degree to which they *linearly* covary
- The covariance depends on the scale of the variables, that's why we often normalize it by the standard deviation to arrive at the *correlation coefficient*

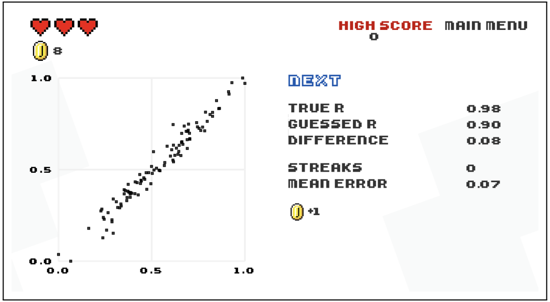
$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

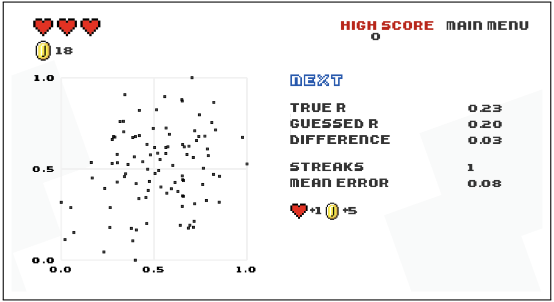
- It holds that

$$X \perp\!\!\!\perp Y \Rightarrow \sigma_{XY} = 0,$$

  but not the other way round

Guessthecorrelation.com



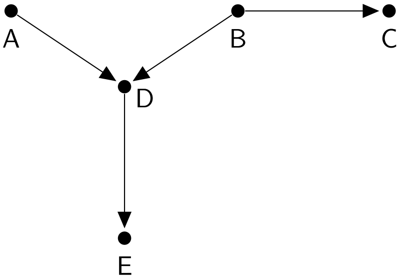


Collider Bias – R Example

```r
# Create two independent uniformly distributed variables
cute <- runif(1000)
smart <- runif(1000)

# Plot
plot(scute, smart, pch=20)

# Construct date equal to one if smart + cute < 0.5, and zero
    otherwise
date <- 1*(smart + cute > 1)

# By design, there is no correlation between the two variables
cor(smart, cute)

# But if we condition on date==1, we find a negative correlation
cor(smart[date==1], cute[date==1])
```

Testable Implications

- D-separation provides testable implications of a model



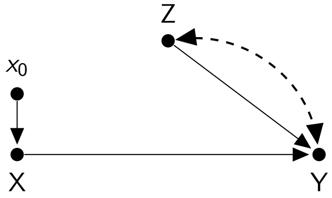
Testable implications:

$$A \perp\!\!\!\perp B \qquad A \perp\!\!\!\perp C$$
$$A \perp\!\!\!\perp E|D \qquad B \perp\!\!\!\perp E|D$$
$$C \perp\!\!\!\perp D|B \qquad C \perp\!\!\!\perp E|D$$
$$C \perp\!\!\!\perp E|B$$

- If one of these conditional independence relations do not hold in the data, the model can be rejected
- *"Causal discovery"*: try to learn compatible model from conditional independence relations found in the data
  - We will talk about that in week 38

Interventions in Structural Causal Models

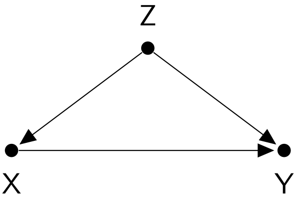
$$z \leftarrow f_Z(u_z)$$
$$x \leftarrow x_0$$
$$y \leftarrow f_Y(x, z, u_Y)$$



- Causal inference $\triangleq$ predict the effects of interventions (policy initiatives, social programs, management initiatives, etc.)
- Interventions in SCMs amount to *"wiping out"* of causal mechanisms, an idea that originally came from econometrics (Strotz and Wold 1960)
  - Delete naturally occurring causal mechanism $f_X(\cdot)$ from model and set X to constant value x_0
  - This operation is denoted by *do-operator*: $do(X = x_0)$
- Query of interest: post-interventional distribution $P(Y = y|do(X = x))$

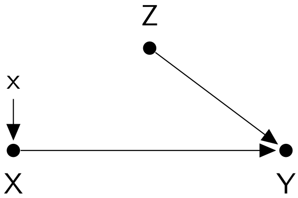
Pre- versus Post-intervention Distribution

<u>Pre-intervention</u>



$$Z = f_z(u_z)$$
$$X = f_x(Z, u_x)$$
$$Y = f_y(X, Z, u_y)$$

<u>Post-intervention</u>



$$Z = f_z(u_z)$$
$$X = x$$
$$Y = f_y(X, Z, u_y)$$

- The intervention changes the data-generating process; thus, $P(Y|X)$ (pre-intervention) is generally not equal to $P(Y|do(X))$ (post-intervention)

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

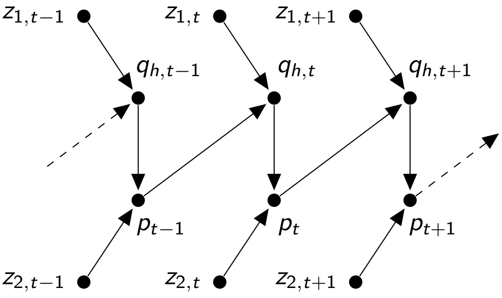
Email: phu.si@cbs.dk

Recursive versus Interdependent Systems

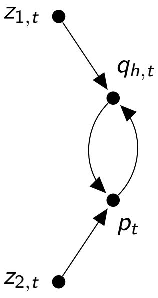
- DAGs represent recursive systems, but many standard models in economics are interdependent (Marshallian cross, game theory, etc.)
- This connects to an old debate within econometrics about the causal interpretation of recursive versus interdependent models that emerged in the aftermath of Haavelmo's celebrated 1943 paper
- One central argument (Strotz and Wold, 1960):
  - Individual equations in an interdependent model do not have a causal interpretation *in the sense of a stimulus-response relationship*
  - Interdependent systems with equilibrium conditions are regarded as a *shortcut* description of the underlying dynamic behavioral processes

Recursive versus Interdependent Systems

- In this context, Strotz and Wold (1960) discuss the example of the cobweb model:



$$p_{t-1} = p_t$$
$$\Rightarrow$$



$$\begin{aligned} q_{h,t} &\leftarrow \gamma + \delta p_{t-1} + \nu z_{1,t} + u_{1,t}, \\ p_t &\leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t}. \end{aligned}$$

$$\begin{aligned} q_{h,t} &\leftarrow \gamma + \delta p_t + \nu z_{1,t} + u_{1,t} \\ p_t &\leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t} \end{aligned}$$

Recursive versus Interdependent Systems

- However, equilibrium assumption $p_{t-1} = p_t$ carries no behavioral interpretation
- Individual equations in interdependent system do not represent autonomous causal relationships in the stimulus-response sense
  - Endogenous variables are determined jointly by all equations in the system
  - Not possible, e.g., to directly manipulate p_t to bring about a desired change in $q_{h,t}$
- Equilibrium models can of course still be useful for learning about causal parameters
- But, if individual mechanisms are supposed to be interpreted as stimulus-response relationships, cyclic patterns need to be excluded

Back