

Causal inference and causal discovery

Shikui Tu

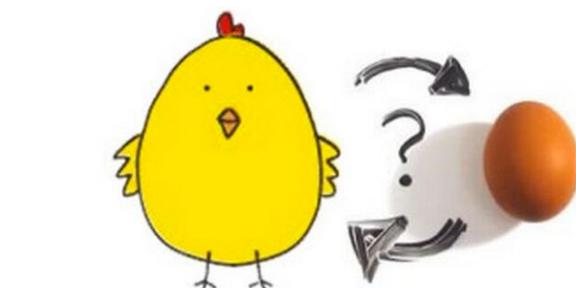
Department of Computer Science and
Engineering, Shanghai Jiao Tong University

2021-06-01

什么叫因果？

- 如下说法是因果吗？

- 善有善报，恶有恶报（佛法基本定律）
- 人在做，天在看，举头三尺有神明，多行不义必自毙
- 天网恢恢疏而不漏，不是不报时候未到
- 满屋老鼠跑，必定有窟窿
- 有风方起浪，无潮水自平
- 不听老人言，吃亏在眼前
- 种瓜得瓜，种豆得豆
- 前因后果
- 凡事预则立，不预则废



什么叫因果？

- **因果关系（西方哲学）**

- 亚里士多德的四因说：质料因、形式因、动力因、目的因
- 托马斯·阿奎那排序：目的因>动力因>质料因>形式因。第一动力因是上帝。
- 休谟：因果关系是两个事件先后的恒常联结。
- 马克思哲学原理：因果联系普遍存在，对立统一

- **因果关系（物理）**

- 原因先于效果（作为限制），可以预计信息不能比光速更快，否则，就可能在某个参照系（使用狭义相对论的洛伦兹变换）中观察者可以看到结果先于原因（即违反因果律的假定）。
- 热力学第二定律确定了时间之箭

- **因果系统（工程）**

- 是指该系统的输出和内部状态取决于当前和以前的输入值。

- **因果关系（法律）**

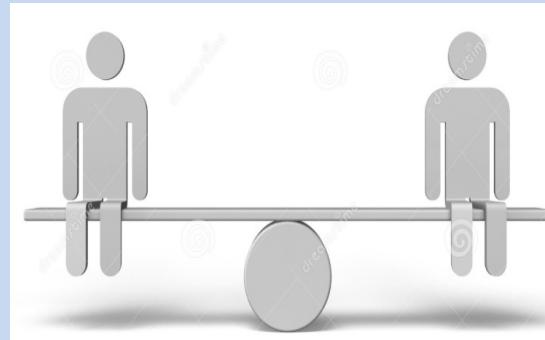
- 根据法学理论，要认定被告对犯罪或侵权行为承担责任，必须证明存在法律上的因果关系。

有势有导的因果势能理论 [Xu,2018]

Causal Potential Theory (CPT)

物理的视角看什么是因果**形成机制** (Xu 2018)

中国文化中的成语：
势必如此、
势比人强、
因势利导、
势不可挡、
孙子的势论
等等



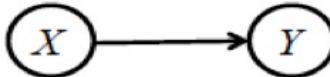
势能高的平衡被破坏后、变化趋势是向势能低处再平衡

因果是这种势必的、**动力学**的一种特殊变化趋势

- Lei Xu (2019), “An overview and perspectives on bidirectional intelligence: Lmser duality, double IA harmony, and causal computation,” IEEE/CAA J. Automatic Sinica, vol. 6, no. 4, pp. 865–893, Jul. 2019.
- Lei Xu (2018), Machine learning and causal analyses for modelling financial and economic data, Springer Nature OA Journal, Applied Informatics, 2018, 5:11, 42 pages, <https://doi.org/10.1186/s40535-018-0058-5>
- K. Kuang, L. Li, Z. Geng, Lei. Xu, K. Zhang, B. Liao, H. Huang, P. Ding, W. Miao, Z.Jiang, (2020), Causal Inference, Engineering, 2020 20(1):1-5.

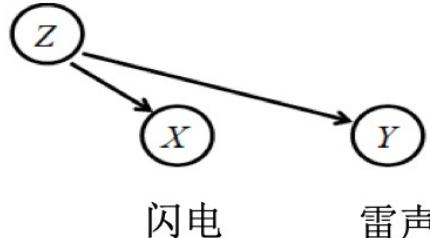
研究因果关系的关键挑战

即使前者引起后者变化



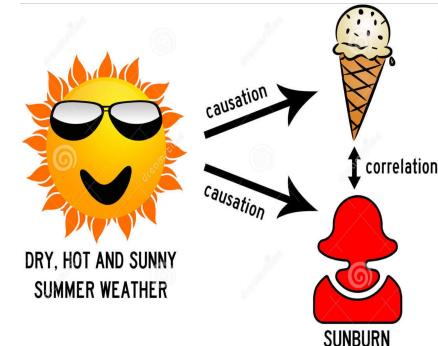
也不一定是前因后果

要排除



闪电

雷声



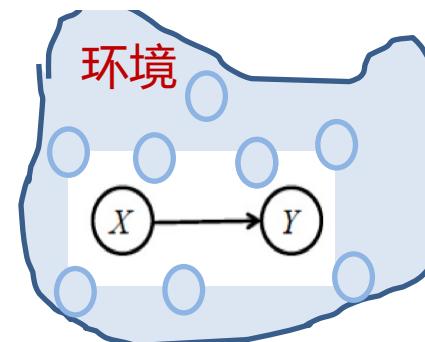
即要确认这二元关系不是环境中某些共同因素导致：

因果是全局关系在两变量上的投射

关键挑战：

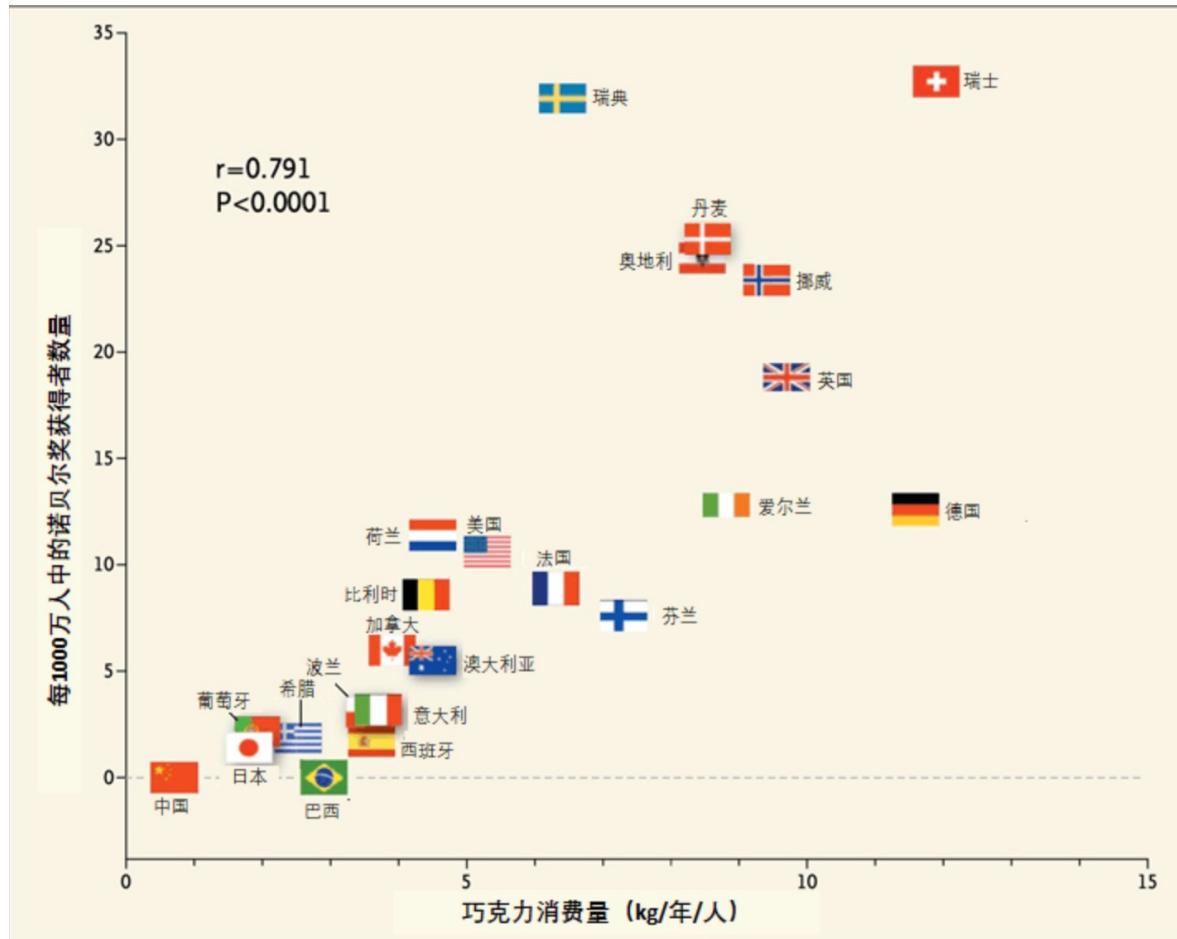
确认那些环境因素？

逻辑上不可能穷举！

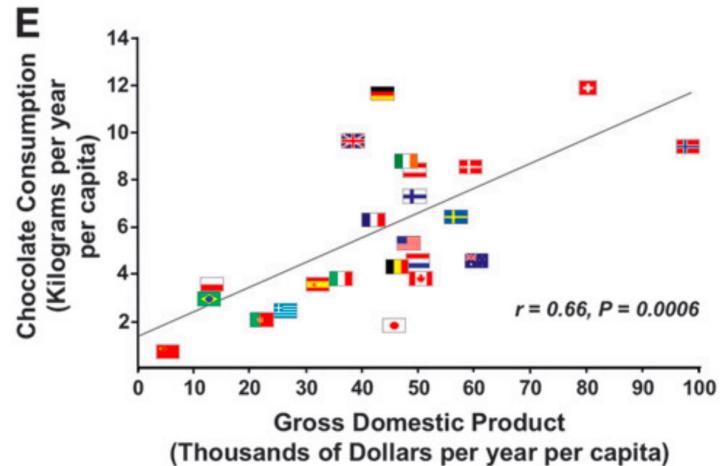
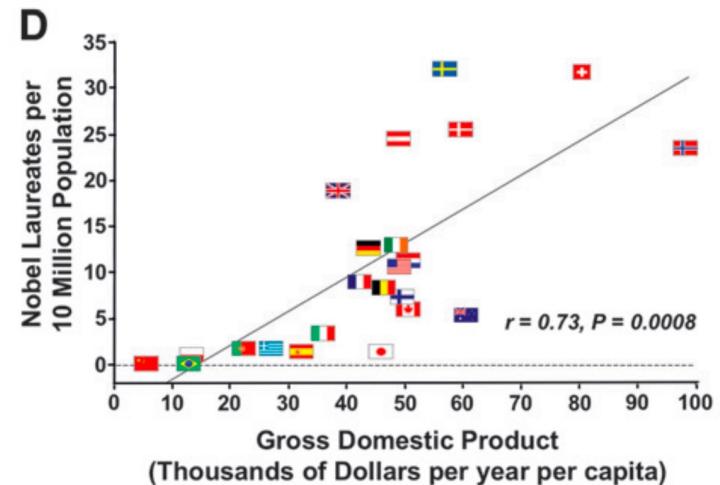
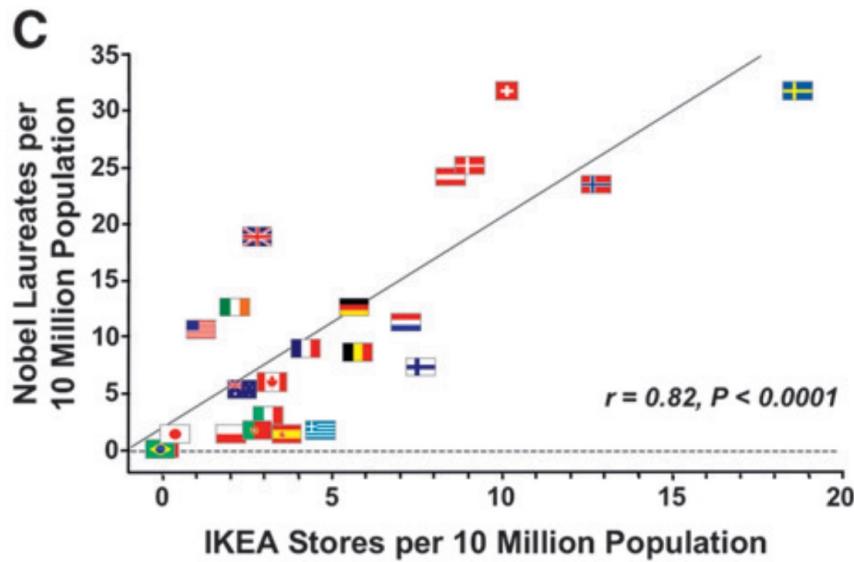


必须就具体情况对环境做假设或建模

巧克力消耗量 vs 诺贝尔奖获得者数量



诺贝尔奖获得者数量的其他相关性



Towards Causal Representation Learning

Bernhard Schölkopf [†], Francesco Locatello [†], Stefan Bauer ^{*}, Nan Rosemary Ke ^{*}, Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Abstract—The two fields of machine learning and graphical causality arose and developed separately. However, there is now cross-pollination and increasing interest in both fields to benefit from the advances of the other. In the present paper, we review fundamental concepts of causal inference and relate them to crucial open problems of machine learning, including transfer and generalization, thereby assaying how causality can contribute to modern machine learning research. This also applies in the opposite direction: we note that most work in causality starts from the premise that the causal variables are given. A central problem for AI and causality is, thus, causal representation learning, the discovery of high-level causal variables from low-level observations. Finally, we delineate some implications of causality for machine learning and propose key research areas at the intersection of both communities.

Mon, 22 Feb 2021

<https://arxiv.org/abs/2102.11107>

In the paper, Turing Award honoree Yoshua Bengio and colleagues make an effort to unite **causality** and **machine learning (ML)** research approaches, delineate some implications of causality for ML, and propose critical areas for future research.

An AI Pioneer Wants His Algorithms to Understand the 'Why'

Deep learning is good at finding patterns in reams of data, but can't explain how they're connected. Turing Award winner Yoshua Bengio wants to change that.



"It's a big thing to integrate [causality] into AI," says University of Montreal researcher Yoshua Bengio. PHOTOGRAPH: RENAUD PHILIPPE/REDUX

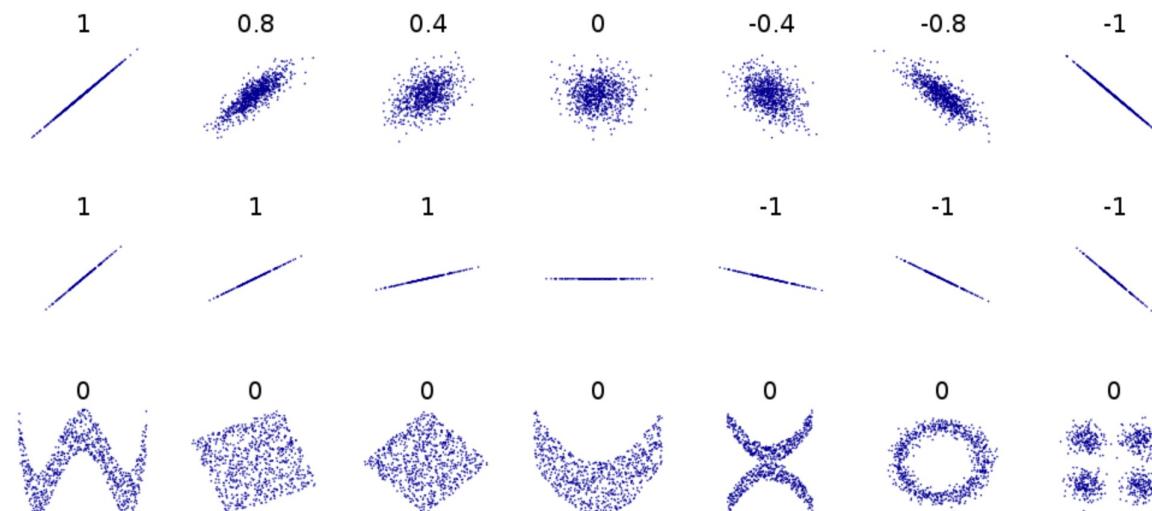
Outline

- **Association vs. Causation (Intervention)**
- Structural Causal Model
- Introduce a linear non-Gaussian model for causal discovery

What Is Association?

wiki

- In statistics, an **association** is **any** relationship between two measured quantities that renders them **statistically dependent**.
- The term "**association**" refers **broadly** to any such relationship, whereas the narrower term "**correlation**" refers to a **linear** relationship between two quantities.

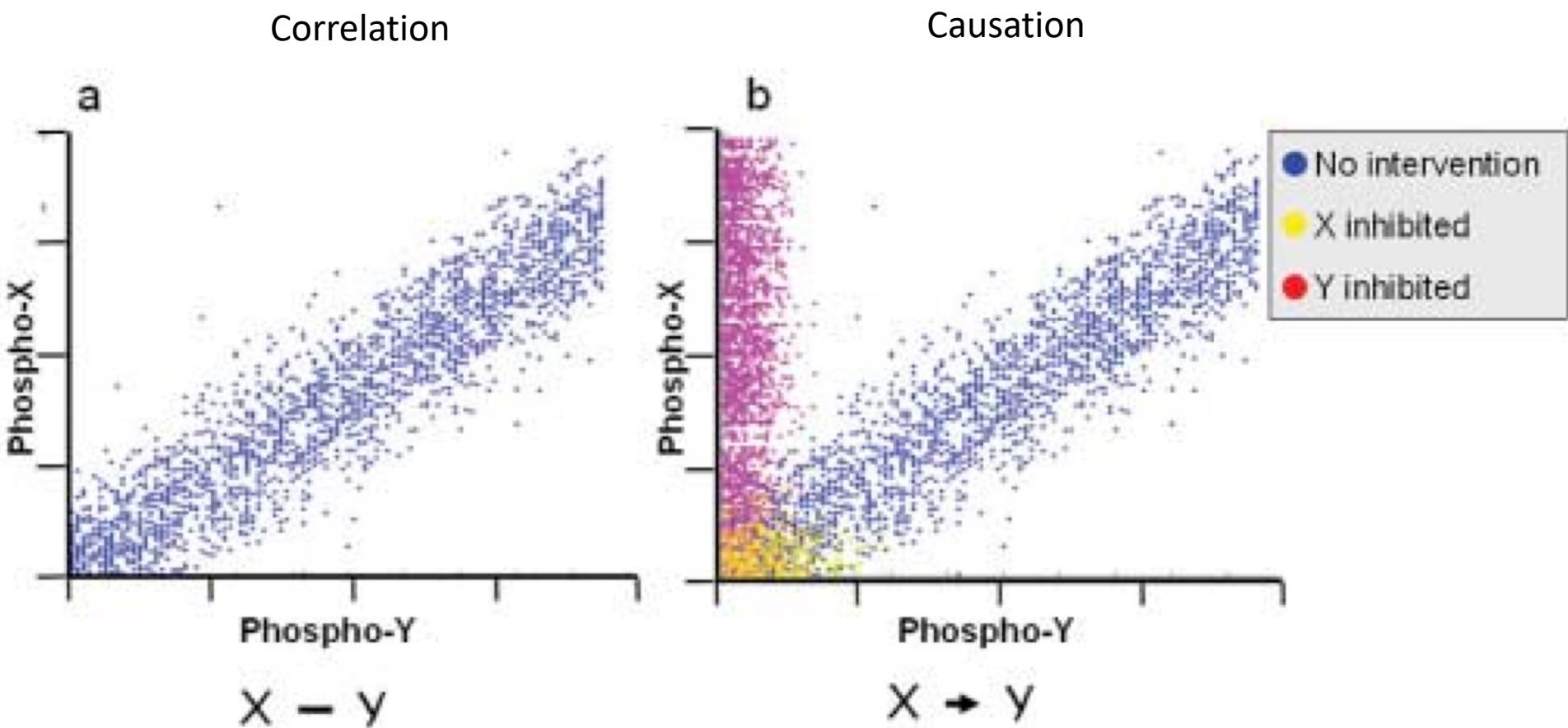


Distinction between Associational and Causal Concepts

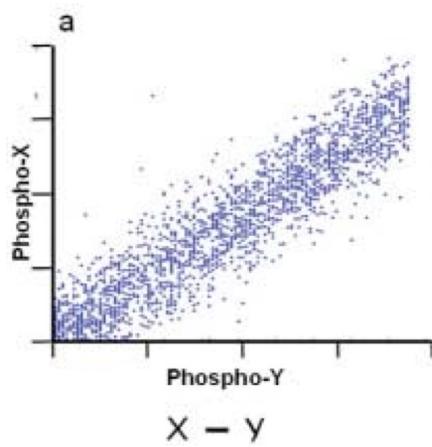
- Pearl ([2003], page 285) Pearl, J. (2003). Statistics and causal inference: A review. *Test* 12 281–318.
 - “An **associational** concept is any relationship that can be *defined in terms of a joint distribution of observed variables*, and a **causal** concept is any relationship that *cannot be defined from the distribution alone*. . . .
 - Every claim invoking **causal** concepts *must be traced to some premises* that invoke such concepts; it *cannot be inferred or derived from statistical associations alone*.”
- About **Judea Pearl**
 - the development of Bayesian networks
 - the 2011 winner of the **ACM Turing Award**, "... for probabilistic and causal reasoning"



A Two-Variable Example



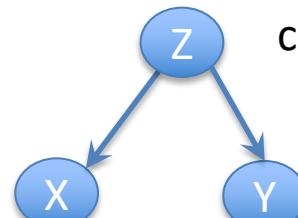
Correlation Does Not Imply Causation



X and Y are correlated.



Possible patterns



confounder



No connection

Association vs. Causation

Within a probability distribution

- Correlation
- Regression
- Dependence
- Conditional independence
- Likelihood
- Collapsibility
- Propensity score
- Risk ratio
- Odds ratio
- Marginalization
- Conditionalization
- “Controlling for”
- ...

Not distribution alone

- Randomization
- Influence
- Effect
- Confounding
- “holding constant”
- Disturbance
- Error terms
- Structural coefficients
- Spurious correlation
- Faithfulness/stability
- Instrumental variables
- Intervention
- Explanation, attribution
- ...

Yule-Simpson's Paradox

-- Can statistics be used to study causality?

[Pearl 2000]

Table 1: Yule-Simpson's Paradox

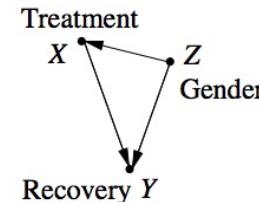
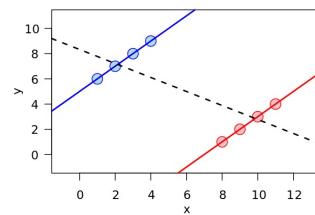
<u>Population</u>	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
<u>Male</u>			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
<u>Female</u>			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%

50% > 40%, positive

60% < 70%, negative

20% < 30%, negative

$$\frac{a}{b} < \frac{c}{d}, \frac{a'}{b'} < \frac{c'}{d'}, \frac{a+a'}{b+b'} > \frac{c+c'}{d+d'}$$



Paradox!

Paradox from Real Data

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naïve aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12,763$, $\chi^2 = 110.8$, d.f. = 1, $P = 0$ (18).

Applicants	Outcome				Difference	
	Observed		Expected		Admit	Deny
	Admit	Deny	Admit	Deny		
Men	3738	4704	3460.7	4981.3	277.3	- 277.3
Women	1494	2827	1771.3	2549.7	- 277.3	277.3

8442 male, 44% admitted
4321 female, 35% admitted

Science, New Series, Vol. 187,
No. 4175 (Feb. 7, 1975), pp. 398-
404

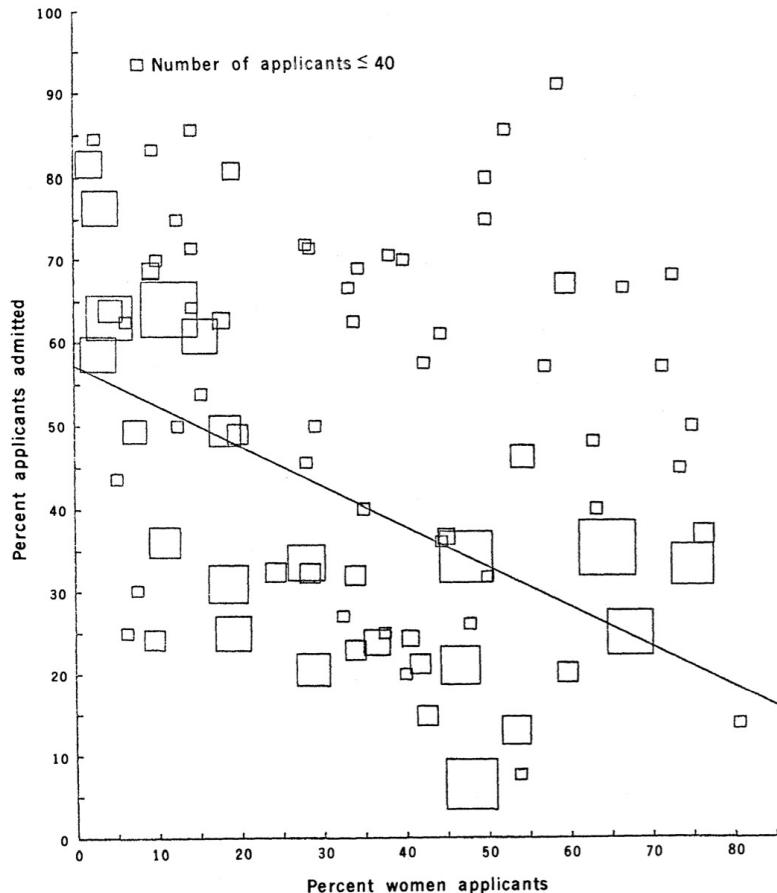


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

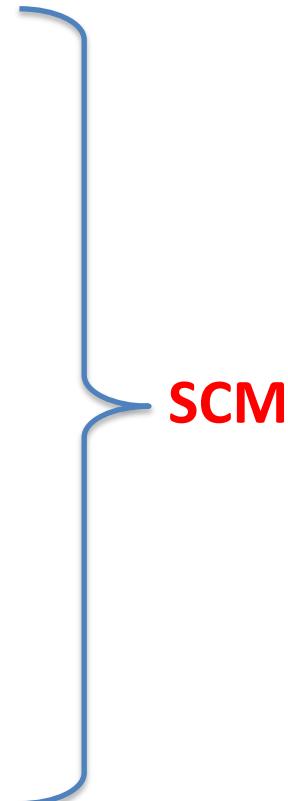
Outline

- Association vs. Causation (Intervention)
- **Structural Causal Model**
- Introduce a linear non-Gaussian model for causal discovery

Structural Causal Model (SCM)

a comprehensive **theory** of causation [Pearl 1995]

- **Potential-outcome framework** of Neyman (1923) and Rubin (1974)
- **Structural equation models (SEM)** used in economics and social science (Goldberger 1973, Duncan 1975)
- **Graphical models** developed for probabilistic reasoning and causal analysis (Pearl 1988, Lauritzen, 1996, Spirtes, Glymour, and Scheines, 2000, Pearl, 2000a).



POTENTIAL-OUTCOME FRAMEWORK

Neyman (1923) and Rubin (1974)

The Counterfactual (Potential Outcomes/Neyman-Rubin) Framework

Neyman (1923) and Rubin (1974)

- Accomplishments:
 - A precise definition of causal effects
 - A formal model of causality against which we can assess the adequacy of various estimators
- Approach:
 - Causal questions are “what if” questions.
 - Extend the logic of randomized experiments to observational data.

Example

- What is the causal effect of *attending catholic school* vs. *public school* on **high school graduation**?

Treatment $T = \{0, 1\}$

{
1: *attending catholic school*
0: *attending public school*

Each individual has two potential outcomes:

{
 Y_1 : *potential outcome if attending catholic school*
 Y_0 : *potential outcome if attending public school*

$$\text{Individual causal effect (ICE)} = Y_1 - Y_0$$

Outcome	Explanation
$Y_0 = Y_1 = 1 \rightarrow \text{ICE} = 0$	Kid would graduate from both catholic and public school. No effect.
$Y_0 = Y_1 = 0 \rightarrow \text{ICE} = 0$	Kid would neither graduate from catholic nor public school. No effect.
$Y_0 = 1, Y_1 = 0 \rightarrow \text{ICE} = -1$	Kid would not graduate from catholic but would graduate from public school. Negative effect.
$Y_0 = 0, Y_1 = 1 \rightarrow \text{ICE} = 1$	Kid would graduate from catholic but would not graduate from public school. Positive effect.

- Generally, either Y_1 or Y_0 can be observed, **not both**.
- The unobserved outcome is called the “counterfactual” outcome.

Fundamental Problem of Causal Inference

[Holland1986]

- One can never *directly observe* causal effects, because we can never observe both potential outcomes for any individual.

		Outcome	
		Treatment	Control
Assignment	Treatment	$(Y T=1) = (Y_1 T=1)$???
	Control	???	$(Y T=0) = (Y_0 T=0)$

Missing data

Average causal effect (ACE) = $E[Y_1] - E[Y_0]$ where $E[.]$ ranges over the entire population
population average of the individual level causal effects

Standard estimator $S^* = E[Y_1 | T=1] - E[Y_0 | T=0]$

where $E[.]$ ranges over the domain of the treatment and control groups, respectively

ACE measures causation \neq S^* measures association

Need assumptions (how the data were generated and collected) to fill in the missing values.

Solve the Problem by Randomized Experiments

$$\text{Average causal effect (ACE)} = E[Y_1] - E[Y_0]$$

$$\text{Standard estimator } S^* = E[Y_1 | T=1] - E[Y_0 | T=0]$$

if $E[Y_1 | T=1] = E[Y_1 | T=0] = E[Y_1]$
 $E[Y_0 | T=1] = E[Y_0 | T=0] = E[Y_0]$  ACE = S*

This condition can be achieved by *random assignment* of individuals to the treatment and control group (in a randomized experiment).

What if random assignment is not possible ?

“Ignorability” $(Y_0, Y_1) \perp T$  S^* is unbiased and consistent for ACE

“Conditional Ignorability” $(Y_0, Y_1) \perp T | X$

But how would we know whether ignorability holds?

Propensity score (倾向性得分)

- Rosenbaum & Rubin [1983] 提出一种对混杂因子降维的方法：倾向性得分

$$e(X) = P(T = 1|X)$$

Rubin认为，如果在给定 X 时， Z 对 Y 的因果效应消除了混杂因子的影响，那么，在给定一维的倾向性得分 $e(X)$ 时， Z 对 Y 的因果效应也是没有混杂因子的影响的，即：

$$Z_i \perp \{Y_i(1), Y_i(0)\} \mid e(X_i)$$

Then, ACE is calculated as:

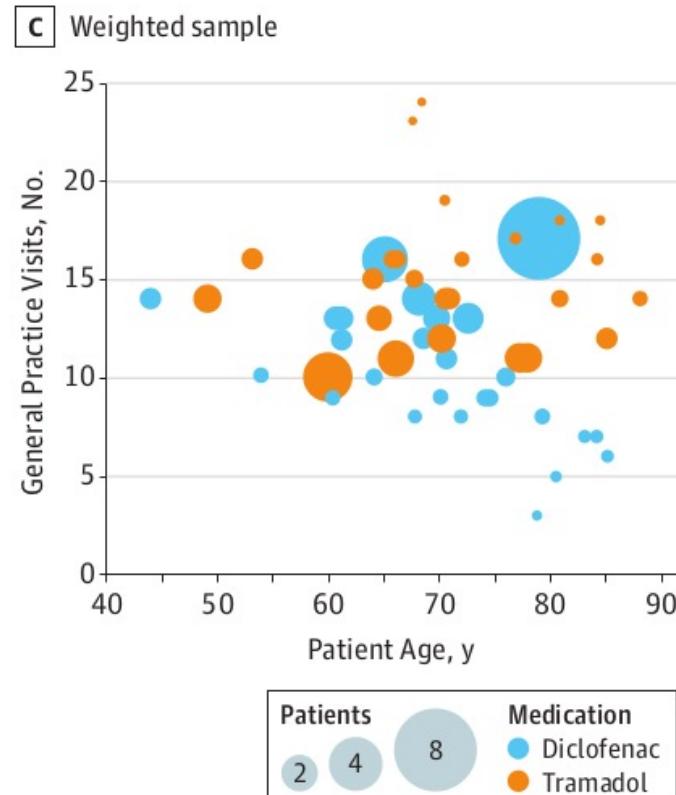
$$ACE = E[E(Y_i|T_i = 1, e(X_i))] - E[E(Y_i|T_i = 0, e(X_i))]$$

基于倾向性得分的逆概率加权

倾向得分是一维的，我们可以根据它：

- 分层（Rosenbaum 和 Rubin 建议分成 5 层），得到平均因果作用(ACE)的估计。
- 连续版本的分层，逆概率加权：

$$ACE = \frac{1}{n} \sum_i \frac{T_i Y_i}{e(X_i)} - \frac{(1 - T_i)Y_i}{1 - e(X_i)}$$



Rosenbaum, Paul R.; Rubin, Donald B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*. 70 (1): 41–55.

STRUCTURAL EQUATION MODELS

by the geneticist Sewall Wright (1921)

Wright's Structural Equation Models

How can one express mathematically the common understanding that symptoms do not cause diseases?

1921

Linear equation:

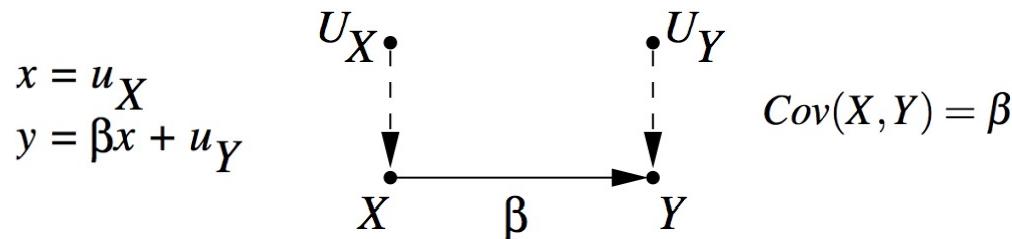
$$y = \beta x + u_Y$$

But if we rewrite it as:

$$x = (y - u_Y)/\beta$$

Then symptom influences disease!

Path Diagram:



- Encodes the possible existence of (direct) causal influence of X on Y
- Encode the **absence** of causal influence of Y on X

It is important to note that, in path diagrams, causal assumptions are encoded not in the links but, rather, in the missing links.

X : disease

Y : symptom

u_Y : all factors, other than the disease, that could possibly affect Y when X is held constant

β : path coefficient

U_X, U_Y : “exogenous”, represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence but are not influenced by the other variables (called “endogenous”) in the model.

DIRECTED ACYCLIC GRAPH

Judea Pearl and colleagues

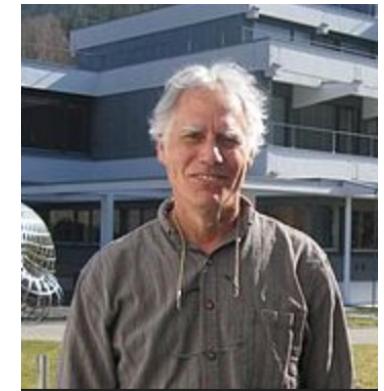
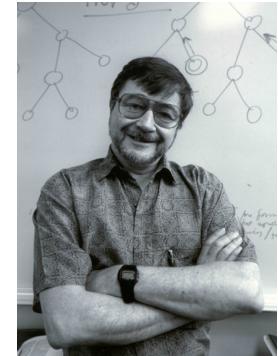
Directed Acyclic Graph (DAG)

- DAGs originate from
 - Structural equation models (1930s+)
 - Social science path models (1960s+)
 - Bayesian networks (1980s+)
- Judea Pearl and colleagues synthesized and generalized these approaches to develop a powerful graphical syntax for causal inference.
- Compatible with potential-outcome (Neyman-Rubin) framework
 - Same concepts, same theorems, different notations

Elwert, Felix. 2013. “Graphical Causal Models.” pp. 245-273 in S.L. Morgan (ed.), Handbook of Causal Analysis for Social Research. New York: Sage Publications.

Some key names in the field

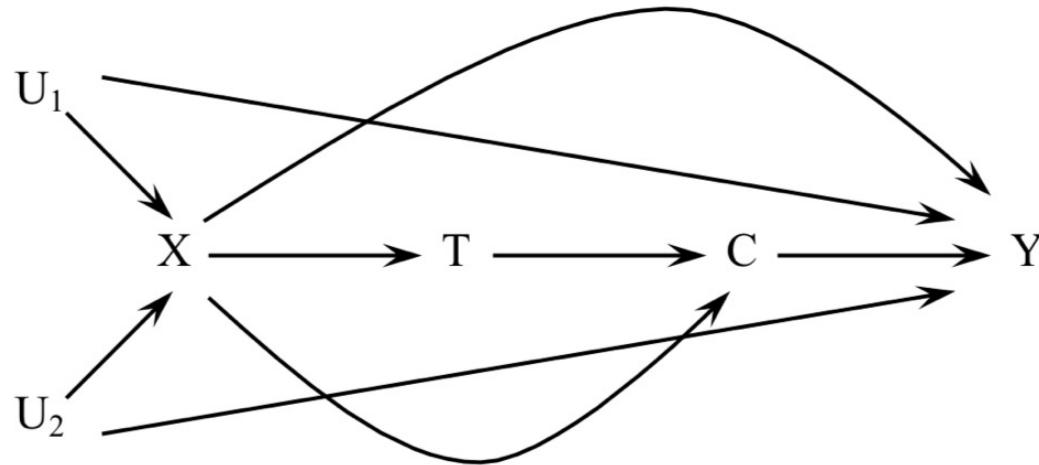
- Computer Science:
Judea Pearl, Jin Tian, Thomas Verma
- Philosophy:
Peter Spirtes, Clark Glymour, Richard Scheines
- Biostatistics:
James Robins, Sander Greenland, Tyler VanderWeele, Miguel Hernan



Identification: separating causal from non-causal associations

- Conditional ignorability
 - The total causal effect of T on Y is identifiable if
$$(Y_0, Y_1) \perp\!\!\!\perp T | X$$
- “What control variables, X, must be included—
and which variables mustn’t be included—
in the analysis to achieve identification? ”

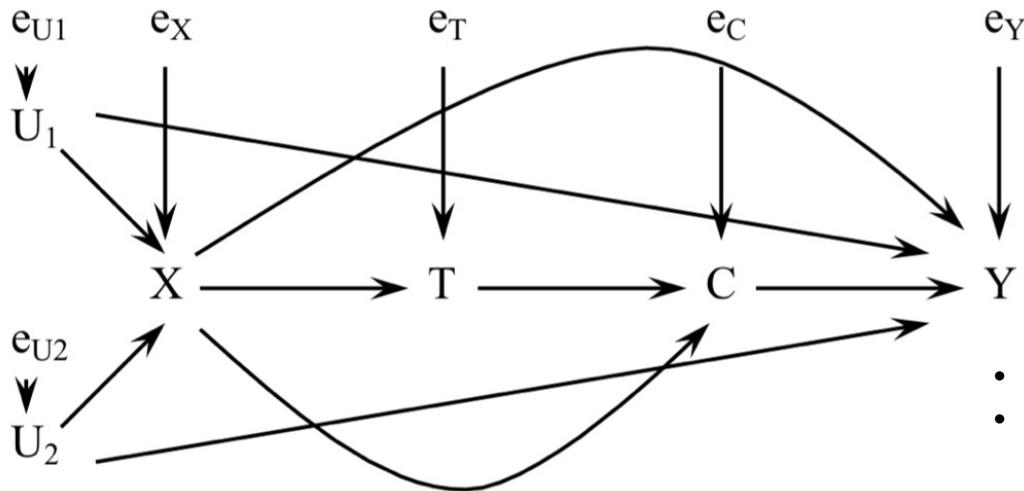
DAGs Encode Causal Knowledge



Arcs	Represent
Directed (presence)	Possible direct causal effects e.g., C may or may not directly cause Y
Missing (absence)	Sharp nulls of no-effect e.g., U_2 does not directly cause T

Only missing arcs encode causal assumptions, whereas directed arcs represent ignorance!

DAGs as SEM



usually omit the set of independent “error term” {e}

- Qualitative
- Population-level representation of data

Each variable, V, is generated by $V = f_V(\text{pa}(V), e_V)$

f_v : some arbitrary deterministic function;
 e_V : stochastic (independent) error

$$U_1 = f_{U_1}(e_{U_1})$$

$$U_2 = f_{U_2}(e_{U_2})$$

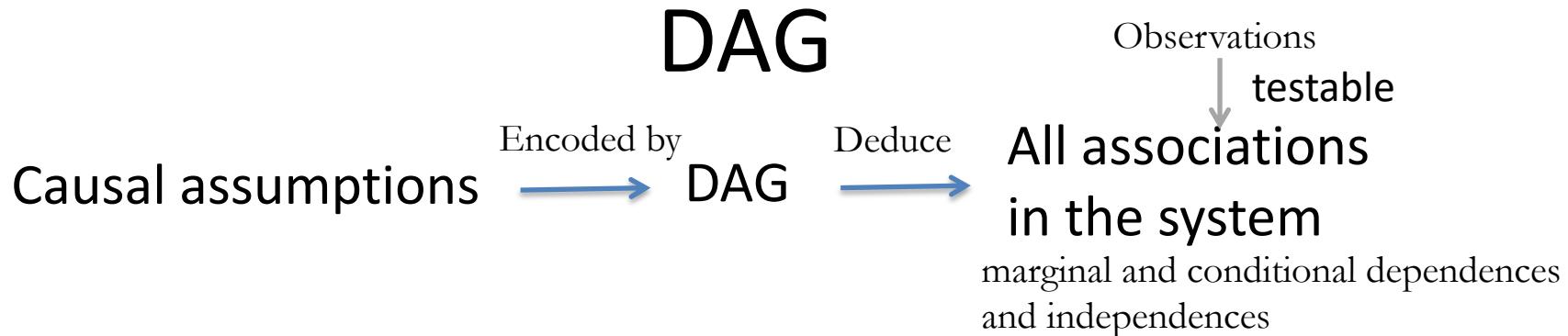
$$X = f_X(U_1, U_2, e_X)$$

$$T = f_T(X, e_T)$$

$$C = f_C(X, T, e_C)$$

$$Y = f_Y(U_1, U_2, X, C, e_Y)$$

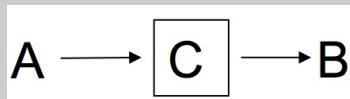
Deriving Testable Implications of A



Bias in estimating the causal effect of A on B

Overcontrol:
(conditional on C) intercepting the causal pathway

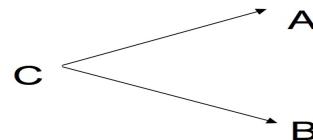
Three elements



Sources of associations between A and B

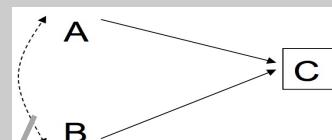
Direct and Indirect causation:
 $A \rightarrow B$ and $A \perp B | C$

Confounding bias:
failure to condition on a common cause



Common cause confounding:
 $A \rightarrow B$ and $A \perp B | C$

Endogenous selection bias:
mistaken conditioning on a common effect



Conditioning on a common effect (“collider”): Selection
 $A \perp B$ and $A \perp B | C$

Non-causal (spurious) association

[Elwert 2013]

d-separation

[Pearl 1988]

- A set S of nodes is said to block a path p if either
 - 1) p contains at least one arrow-emitting node that is in S , or
 - 2) p contains at least one collision node that is outside S and has no descendant in S .
- If S blocks all paths from X to Y , it is said to “d-separate X and Y ,” and then, X and Y are independent given S , written $X \perp Y | S$.

Active Trails

$d\text{-sep}_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \iff$ No active trails in G between \mathbf{X} and \mathbf{Y} given \mathbf{Z}

- $X \rightarrow Y$ blocking and the “flow” of association

- $X \leftarrow Y$

$W \notin Z$

- $X \rightarrow W \rightarrow Y$



- $X \leftarrow W \leftarrow Y$



- $X \leftarrow W \rightarrow Y$



- $X \rightarrow W \leftarrow Y$



v-structure (W is a collider), e.g.,

$W \in Z$

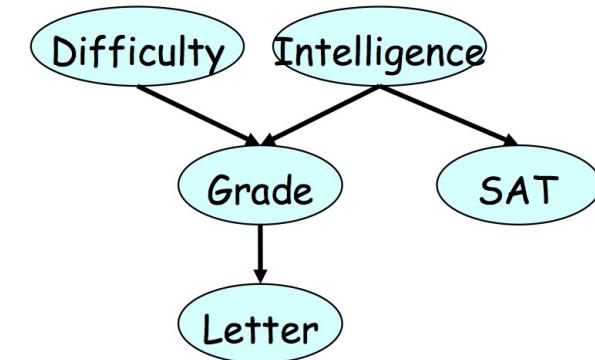


dead battery --> car won't start <-- no gas

[Pearl (1988)]

Berkson's paradox (1946):

Given two independent events, if you only consider outcomes where at least one occurs, then they become negatively dependent.



Daphne Koller

Thank you!