MSDZ06 Common Tools for Data Science

Causal Inference

To build truly intelligent machines, teach them cause and effect. Judea Pearl

Min Ding

April 26, 2022

Correlation Does Not Imply Causation

- Francis Galton (1889): Height and forearm length are “co-related”. Interpreted both of them are caused by genetic inheritance.
- Karl Pearson (1892): Causation can never be proven–that mere data is all there is to science.

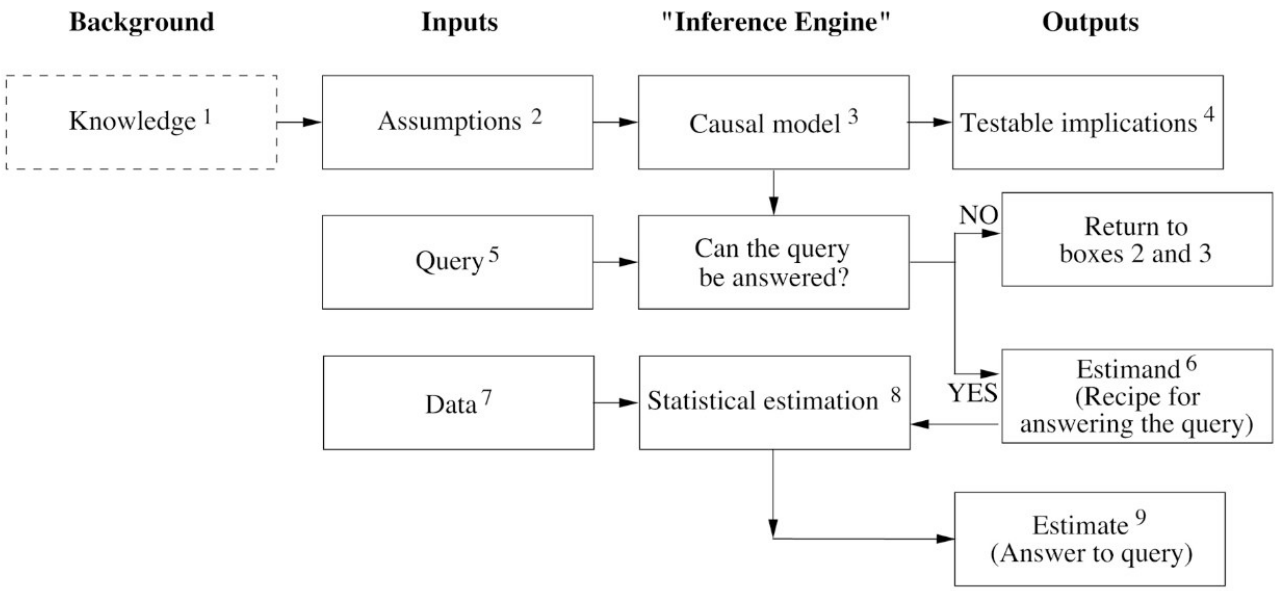
- But should (did) we stop here?
    - Most people did, some does not.

Causal Revolution

- Judea Pearl: “This idea must have seemed simple to Wright but turned out to be revolutionary because it was the first proof that the mantra ‘correlation does not imply causation’ should give way to ‘some correlations do imply causation.”

2012 Turning Award

Inference Engine



What is causality?
How may we quantify it?
Could you give some examples on causal questions?

Brief History - Koch's Postulates (1884)

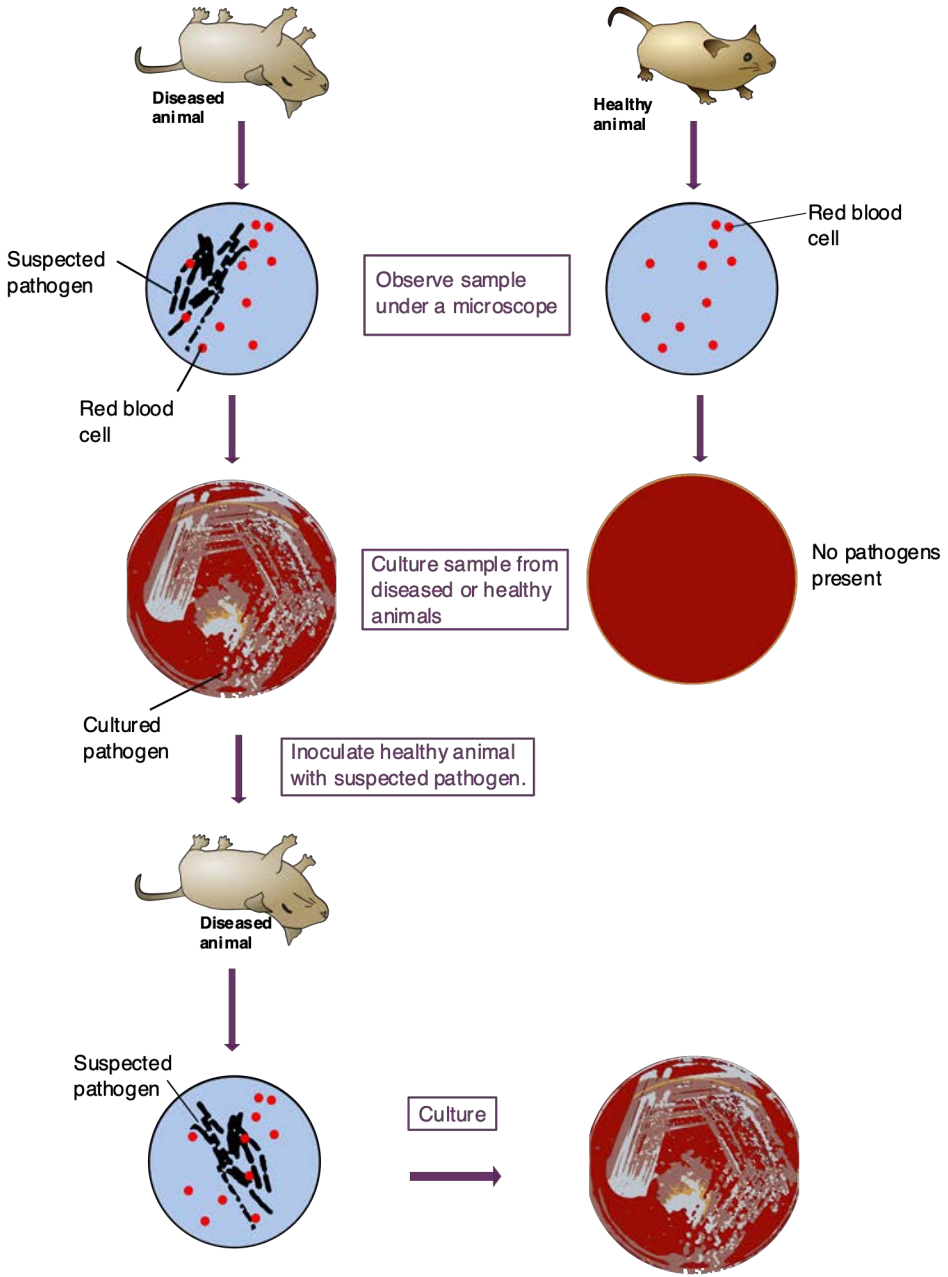
Koch's Postulates:

① The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms.

② The microorganism must be isolated from a diseased organism and grown in pure culture.

③ The cultured microorganism should cause disease when introduced into a healthy organism.

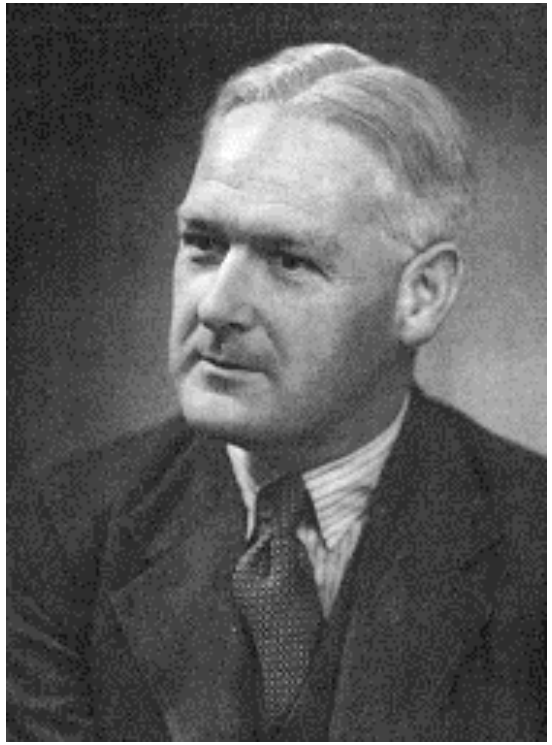
④ The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.



- Koch's postulates are four criteria designed to determine whether there is a causative relationship between a microbe and a disease.
- They have been used since the 19th century to decide if a microorganism was the cause of a disease.

Bradford Hill criteria for causation (1965)

In 1965, the English statistician Sir Austin Bradford Hill proposed a set of nine criteria to provide epidemiologic evidence of a causal relationship between a presumed cause and an observed effect:



	Criteria	Definition
1	Strength	The size of the risk as measured by appropriate tests.
2	Consistency	The association is consistent when results are replicated in studies in different settings using different methods.
3	Specificity	When a single putative cause produces a specific effect.
4	Temporal sequence	Exposure always precedes the outcome.
5	Dose response	An increasing level of exposure (in amount and/or time) increases the risk.
6	Experimental evidence	The condition can be altered (prevented or ameliorated) by an appropriate experimental regimen
7	Biologic plausibility	The association agrees with currently accepted understanding of pathobiological processes.
8	Coherence	The association should be compatible with existing theory and knowledge.
9	Analogy	A finding of analogous associations between similar factors and similar diseases.

Causal Graph - Sewall Wright 1921

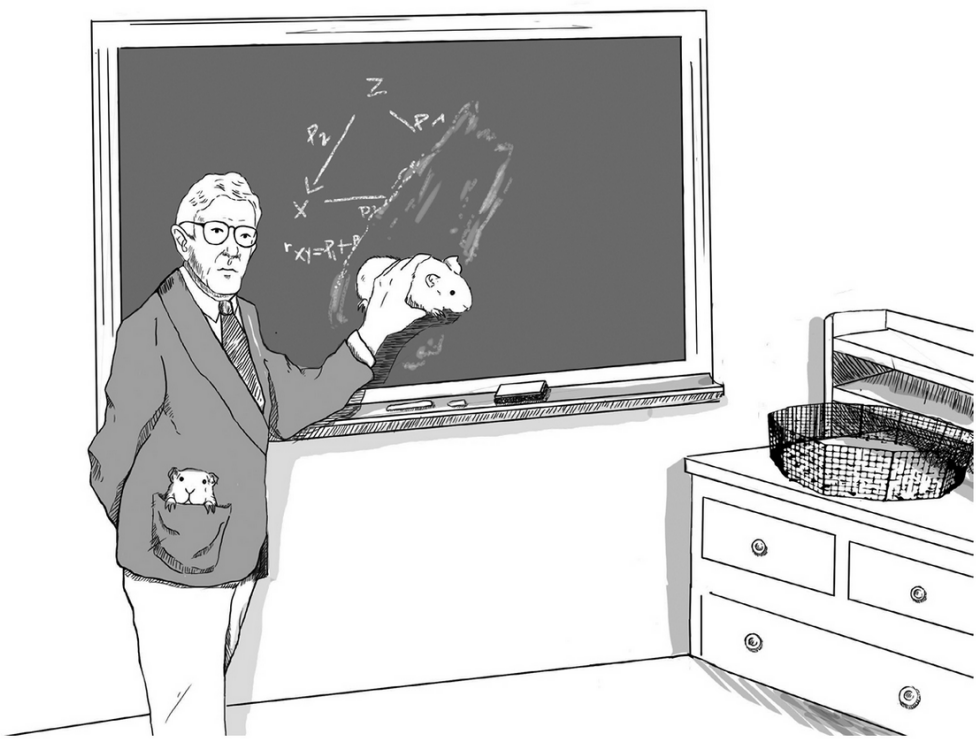


圖 2.6　西瓦爾．萊特首先提出依據資料解答因果問題的數學方法，稱為路徑圖。他對數學的喜愛僅次於對天竺鼠的熱情。來源：Dakota Harr 繪圖。

THE RELATIVE IMPORTANCE OF HEREDITY AND ENVIRONMENT IN DETERMINING THE PIEBALD PATTERN OF GUINEA-PIGS

By Sewall Wright

Bureau of Animal Industry, United States Department of Agriculture

Communicated by R. Pearl, March 17, 1920

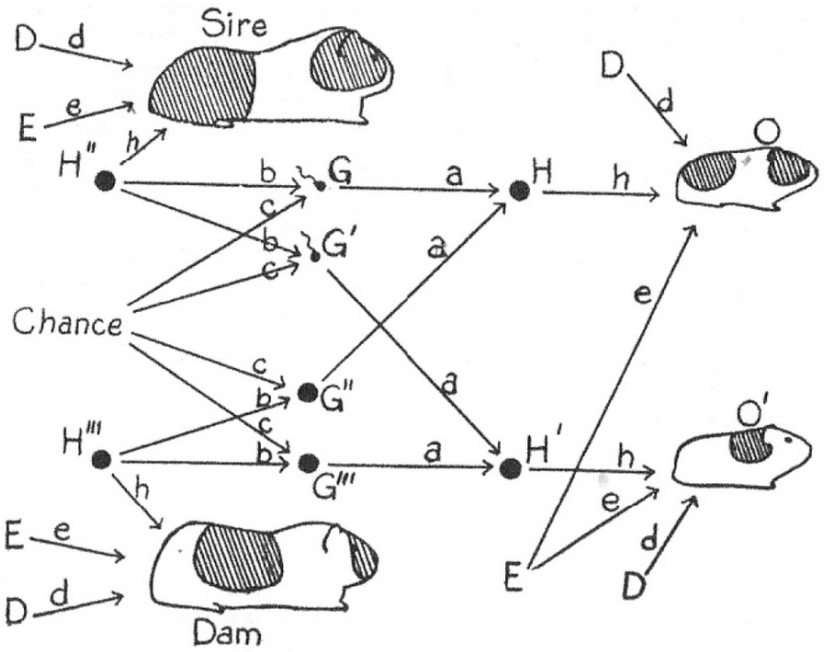


圖 2.7　西瓦爾．萊特的第一幅路徑圖說明決定天竺鼠毛色的因素。D = 發育因素（受孕後到出生前），E = 環境因素（出生後），G = 分別來自雙親的基因因素，H = 來自雙親的綜合遺傳因素，O, O′ = 後代。分析的目的是判定 D、E、H（圖中寫成 d、e、h）的影響強度。
來源：Sewall Wright, Proceedings of the National Academy of Sciences (1920, 320–332)。

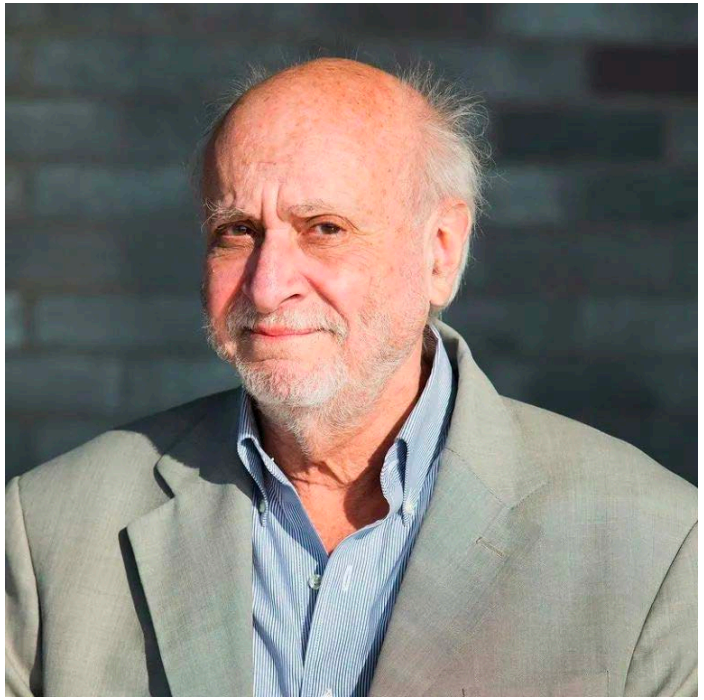
What is Causality?
In Potential Outcome Framework (1974) 潜在结果模型

- Use potential outcomes to quantify causal effects
- e.g., for a population, define Average Treatment Effect (ATE):

$$
\begin{aligned}
ATE &= E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)] = E_x[E[Y(1)|X] - E[Y(0)|X]] \\
&= E_x[E[Y(1)|T=1,x] - E[Y(0)|T=0,x]] \\
&= E_x[E[Y|T=1,x] - E[Y|T=0,x]]
\end{aligned}
$$

Type Causality (forward-looking)

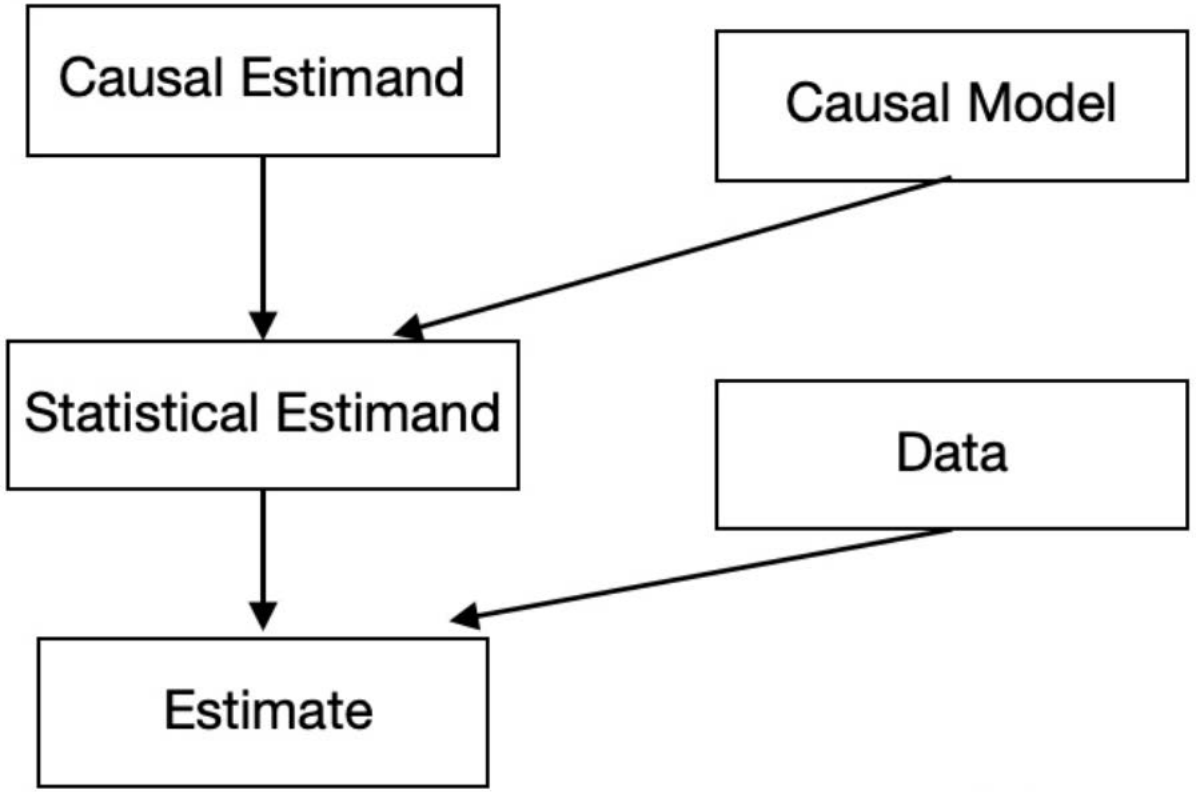
- 吸烟是否会导致肺癌？



Donald B. Rubin

What is Causal Inference?

- 根据已有或者假设的因果模型，从数据中提取因果关系 Use observed data to infer a causal relationship based on a known or hypothesized causal model.



The Fundamental Problem of Causal Inference

- It is not possible to observe both Y(1) and Y(0) for the same individual and so individual causal effects.
- Among Y(1) and Y(0), one is observed, the other is counterfactual (反事实)。

Example: HIV treatment and survival

- **Treatment label** (ART = antiretroviral therapy)

$$\begin{aligned} a &= 1 \text{ label for scenario that ART is received} \\ &= 0 \text{ label for scenario that ART is not received} \end{aligned}$$

- **Potential outcomes:** hypothetical outcomes *for a single individual* under the two possible scenarios.

$$\begin{aligned} Y_0 &= \text{survival time if } a = 0 \\ Y_1 &= \text{survival time if } a = 1 \end{aligned}$$

- **Causal effect of receiving ART:** the contrast of these potential outcomes *within the same individual*

$$Y_1 - Y_0$$

Slide: Joseph Hogan

	Potential Outcomes			Observables		
Person	Y_0	Y_1	Causal effect	Received ART?	Y	
1	300	500	200	yes	500	
2	200	900	700	yes	900	Ave = 600
3	500	600	100	yes	600	
4	100	400	300	yes	400	
5	800	1200	400	no	800	
6	800	1000	200	no	800	Ave = 867
7	1000	1300	300	no	1000	
Average	529	843	314			

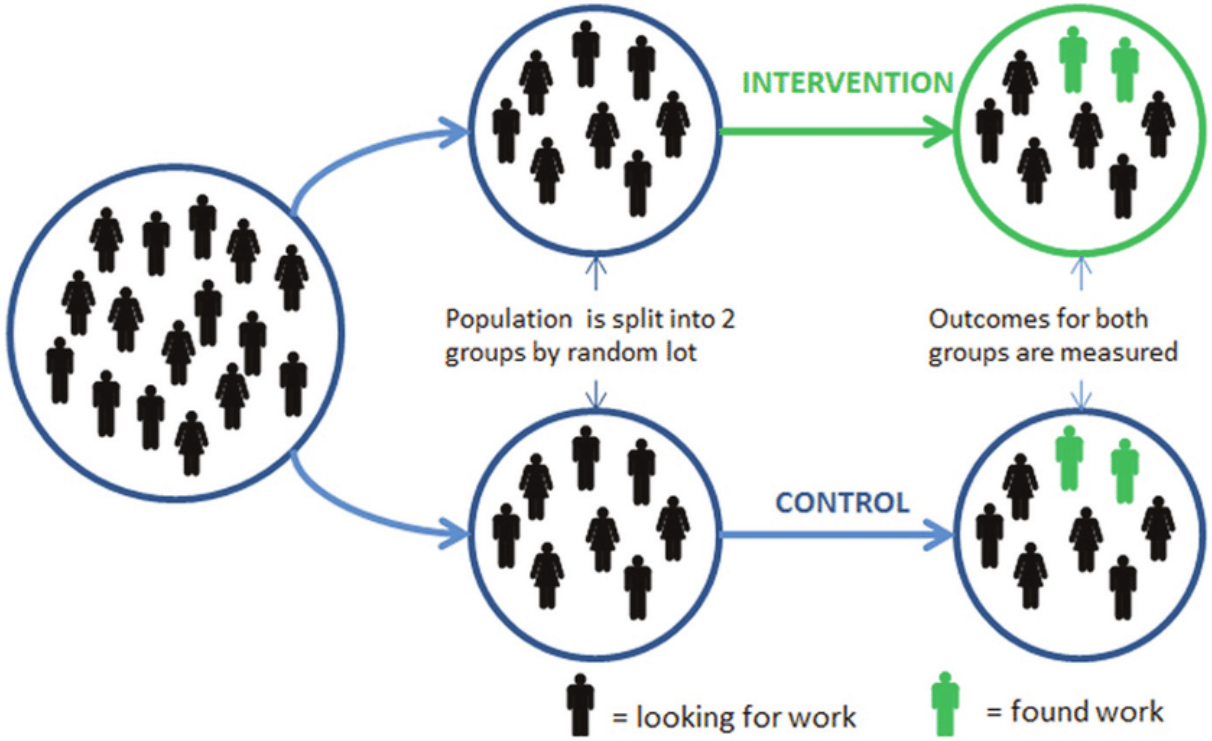
Predictive effect: $E(Y \mid A=1) - E(Y \mid A=0) = 600 - 867 = \boxed{-267}$

Causal effect: $E(Y_1) - E(Y_0) = 843 - 529 = \boxed{+314}$

- Need to correct for this bias (due to confounders)!

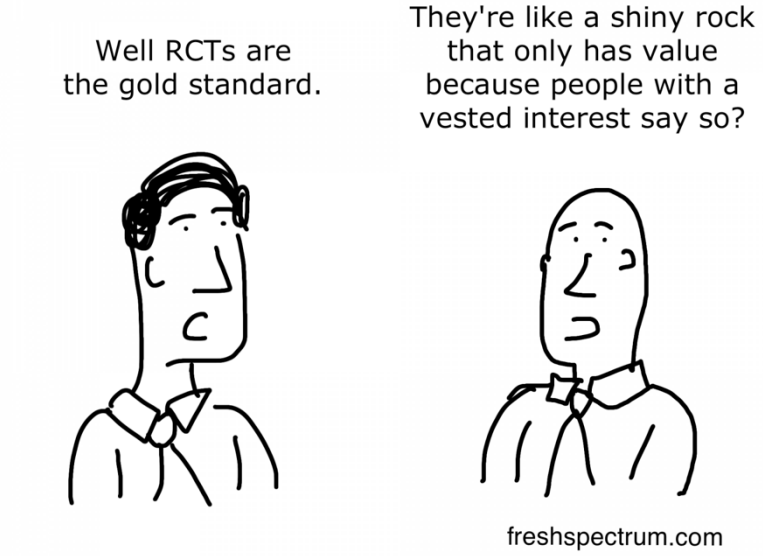
Solution: Golden standard - Randomized Control Trials (RTCs)

- Every participant in the trial has the same chance as any other of receiving the intervention and receiving the control.
- And thus there are no systematic differences between the treatment and control groups.
- So that the two groups are balanced. And effects of confounders can be de avoided.



Problems in RTCs?

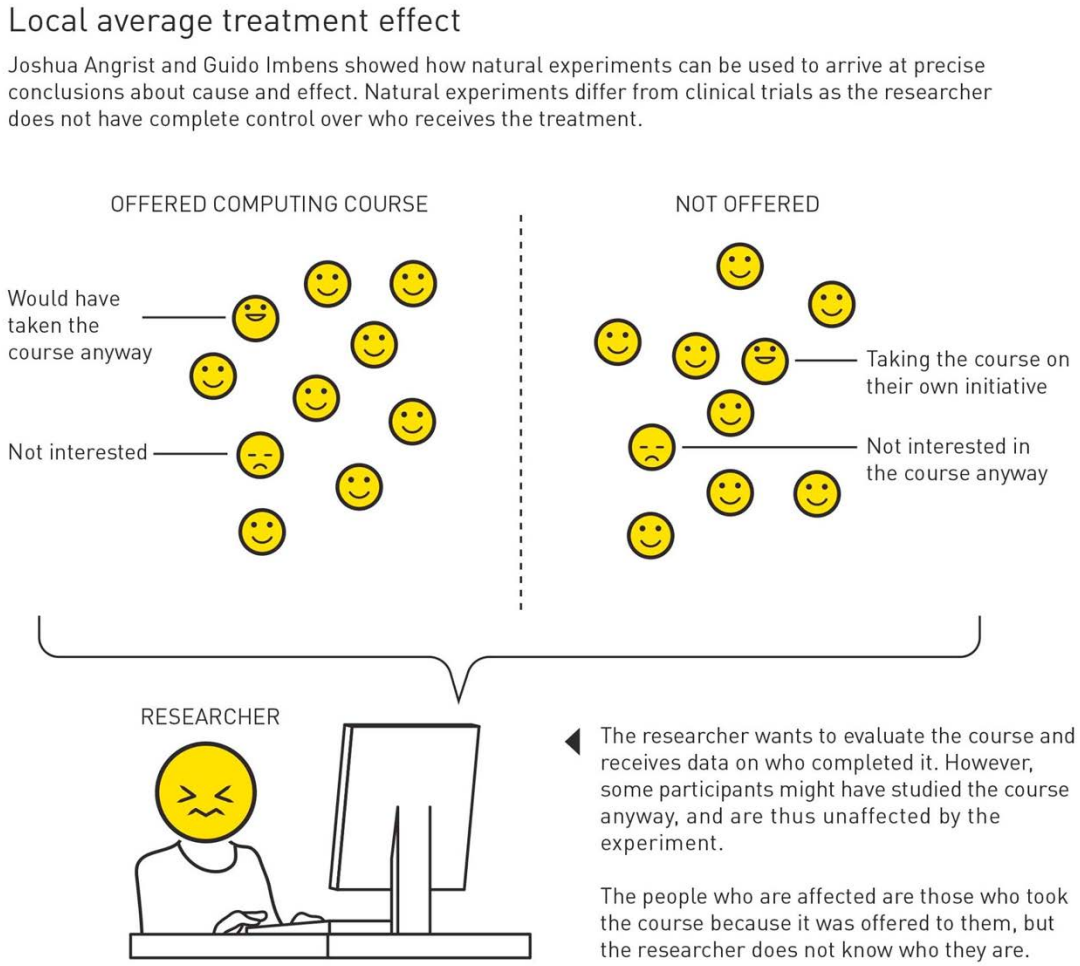
- RTCs can be impossible, unethical, unwarranted, and even misleading.
- Endogenic problems (in blackbox?):
  - Example (Incentive problem): advertisement campaign is championed by a senior manager, then analysts who evaluate the investment return may game the data to make it look good. (Holmstrom and Milgrom, 1991; Prendergast, 1993)
  - Too big to fail amongst financial institutions and firms.
  - Algorithms that tell us which topics are trending don't merely re trends; they can also help create them" by Russel Smith “知识陷阱”



Examples from Ye Luo

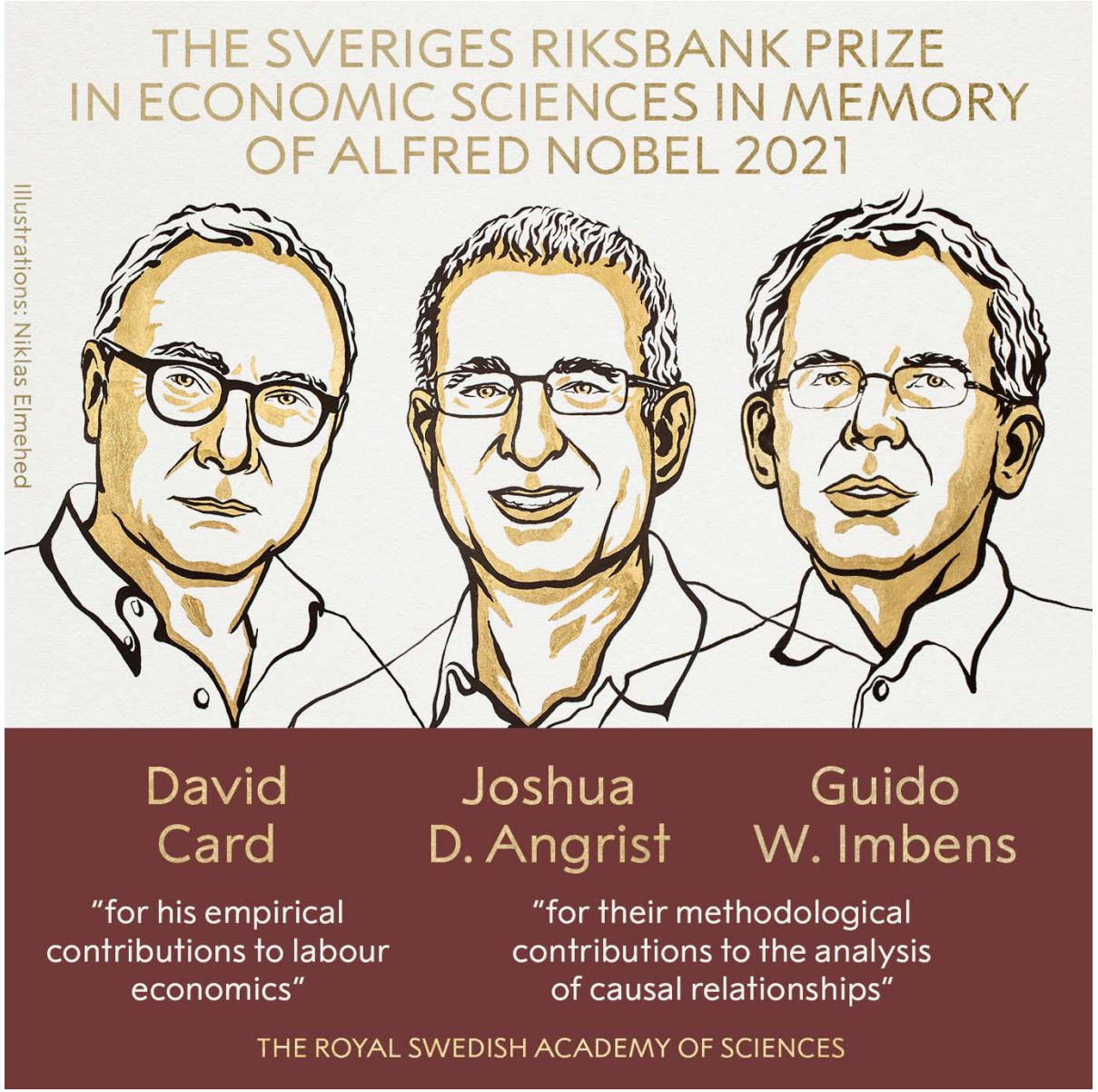
Natural Experiments

- But nature can be the laboratory.



©Johan Jarnestad/The Royal Swedish Academy of Sciences

2021 Nobel Prize in Economic Sciences

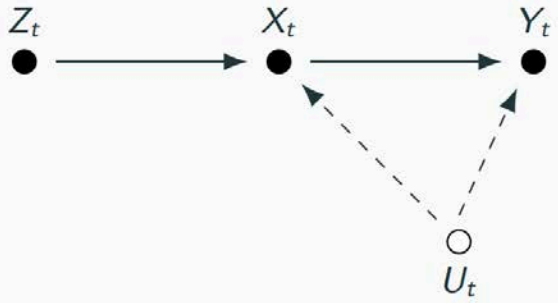


- Based on Rubin's Potential Outcome Model
- Apply to Natural Experiments
- Use Instrumental Variable (IV)

IV for Bias Correction

- Consider a linear model $Y = X\beta + \epsilon$.
- ϵ is a residual.
- Standard assumption for identification of the model: $\mathrm{corr}(X, \epsilon) = 0$.
- But some times this assumption can fail. Technically, when $\mathrm{corr}(X, \textit{epsilon}) \neq 0$, we call X endogenous variable.

- A "third" variable Ut (confounder or confounding variable) may influence both the independent and dependent variables, which can leads to a spurious correlation.



- An additional instrumental variable Z_t is needed for causal analysis of this model.
- Two requirements for IV in linear regression (and many other models):
  - Z_t correlated with X_t
  - Z_t and ϵ_t are conditionally independent given X_t.
  - Consequently we have that $\mathrm{corr}(Z_t, \epsilon_t) \neq 0$.
- IV helps identify β by "creating" exogenous shocks to X.

Slide: Ye Luo

IV Approach to Debias

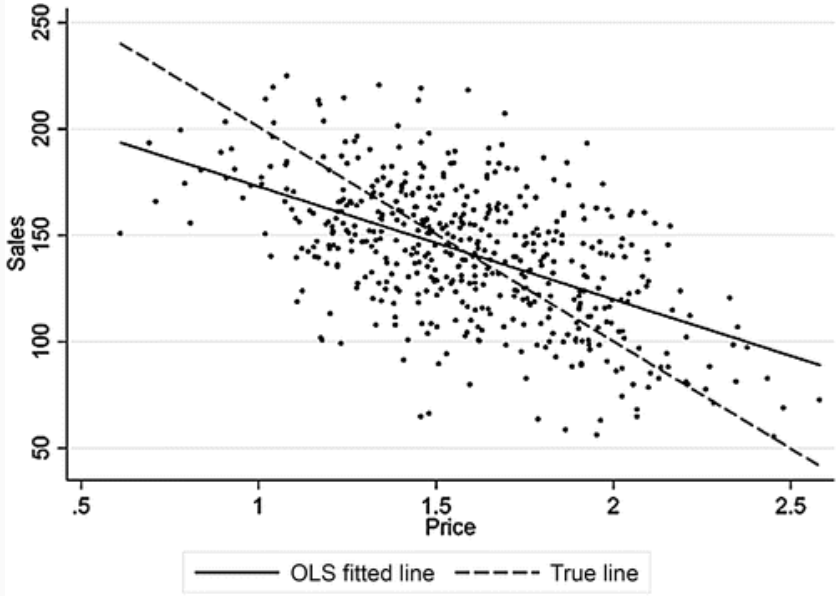
Example: Supply/Demand Model

- Supply and Demand equations: the price is determined by supply and demand simultaneously.

$$Q_S = A + B \cdot P + \epsilon_S.$$

$$Q_D = C - D \cdot P + \epsilon_D.$$

- Market Clearing: $Q_S = Q_D$.
- If we regress the market sales Q on market price P, what do we get: supply or demand? Neither.



- If we would like to learn the demand equation, we need to find variables that affects P, but not ϵ_D.
- Shocks in supply equation can serve this purpose: change of marginal cost (oil condition, weather condition, etc.)

Well-known IVs in Economics

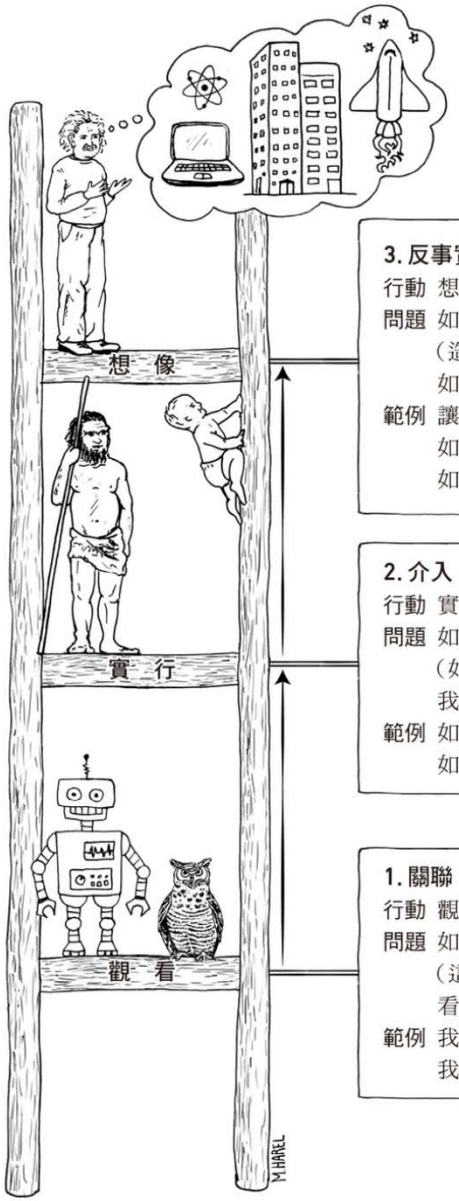
- Levitt (1997): Effect of police on crime.
  - Concern: high-crime areas require more police.
  - IV: Mayor electoral cycle – affect the number of police without affecting crime.
- Angrist and Krueger (1991): Effect of schooling on future income.
  - Concern: unobserved ability affects both schooling and income.
  - IV: quarter of birth – affect years of schooling without affecting ability.
- Angrist and Evans (1998): Impact of childbirth on women labor participation.
  - Concern: women with lower earning potential have more kids and do not work.
  - IV: gender mix of the first two kids – affect the number of children without affecting earning potential.

Other Statistical Techniques for Bias Correction

- Matching
- Weighting
- Extrapolation
- Propensity scores
- Others

Enough for causal inference?
Can we now answer all causal questions?

Pearl 因果推理三级结构



3. 反事實
行動 想像、回想、理解
問題 如果當時我做……會怎麼樣？為什麼？
（造成Y的原因是X嗎？如果X沒有發生會怎麼樣？
如果當時我的反應不同會怎麼樣？）
範例 讓我頭不痛的原因是阿斯匹靈嗎？
如果奧斯華沒有暗殺甘迺迪的話，甘迺迪還活著嗎？
如果我這兩年沒有抽菸的話會怎麼樣？

2. 介入
行動 實行、介入
問題 如果我做……會怎麼樣？怎麼做？
（如果我做X的話，Y會怎麼樣？
我要怎麼讓Y發生？）
範例 如果我服用阿斯匹靈，頭就不會痛了嗎？
如果禁止抽菸會怎麼樣？

1. 關聯
行動 觀看、觀察
問題 如果我看到……會怎麼樣？
（這些變項之間的關係是什麼？
看到X會不會改變我對Y的相信程度？）
範例 我們如何從症狀了解疾病？
我們如何從民調了解選舉結果？

3-LEVEL HIERARCHY

3. COUNTERFACTUALS
ACTIVITY: Imagining, Retrospection, Understanding
QUESTIONS: *What if I had done . . . ? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES: Was it the aspirin that stopped my headache?
Would Kennedy be alive if Oswald had not killed him? What if I had not smoked the last 2 years?

2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: *What if I do . . . ? How?*
(What would Y be if I do X?)
EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see . . . ?*
(How would seeing X change my belief in Y?)
EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Counterfactual Definition of Causal Relationship

- David Lewis (1973): An event E *causally depends* on C if, and only if, (i) if C had occurred, then E would have occurred, and (ii) if C had not occurred, then E would not have occurred.

Actual Causality (backward-looking)

- 恐龙灭亡是小行星撞击地球造成的吗？

Pearl – Structural Causal Model SCM

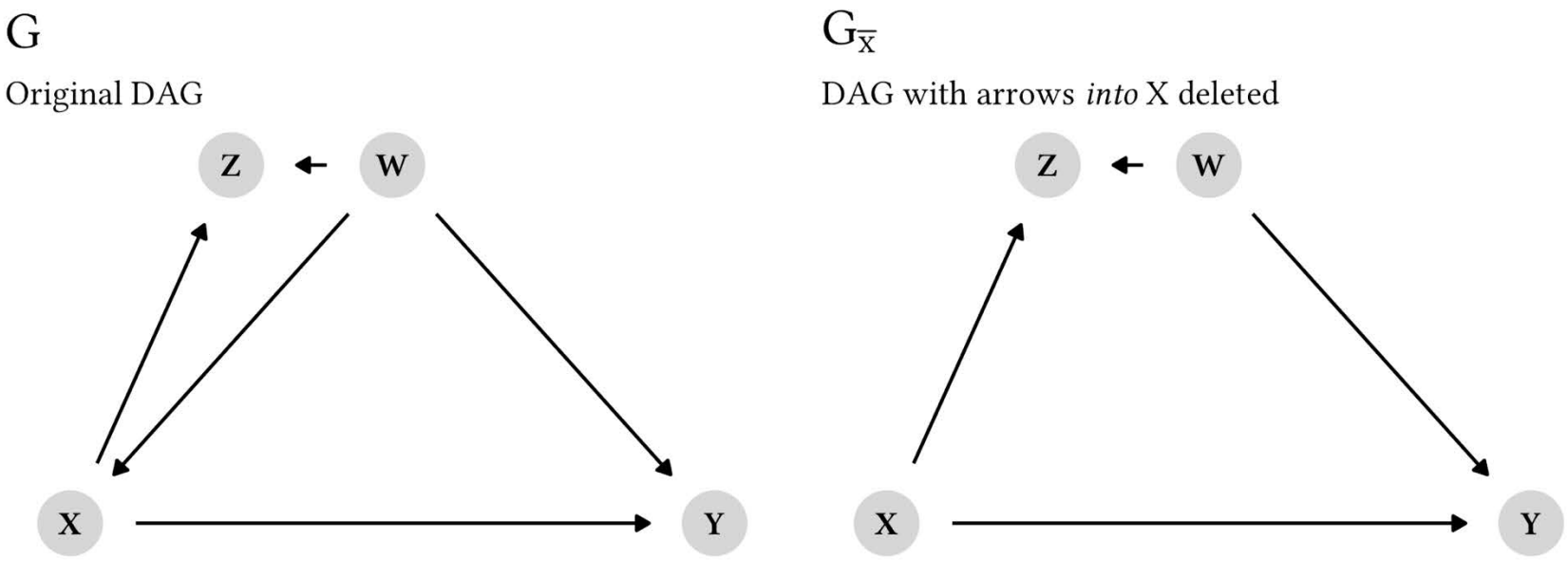
- A probabilistic Structural Causal Model (SCM) is defined as a tuple $M=\langle U,V,F,P(U)\rangle$ where U is a set of exogeneous variables, V a set of endogenous variables, F is a set of structural equations that determines the values of each endogenous variable and $P(U)$ a probability distribution over the domain of U.
- The effect of an intervention on a variable X by a submodel $Mx=\langle U,V,Fx,P(U)\rangle$ where Fx indicates that the structural equation for X is replaced by the new interventional equation. For example, the atomic intervention of setting the variable X to a specific value x --- usually denoted by $do(X=x)$ --- consists of replacing the equation for X with the equation $X=x$.

Model M		Modified M_x
$Z = U_z$	Intervention $\longrightarrow$	$Z = U_z$
$X = f(Z, U_x)$		$X = f(Z, U_x)$
$Y = g(X, Z, U_y)$		$Y = g(X, Z, U_y)$

do-calculus for Deriving P(Y | do(X))

- **Rule 1 (ignoring obs.):** We can ignore any observational node if it doesn't influence the outcome through any path, or if it is d-separated from the outcome.

$$P(y \mid z, \mathrm{do}(x), w) = P(y \mid \mathrm{do}(x), w) \qquad \text{if } (Y \perp Z \mid W, X)_{G_X}$$

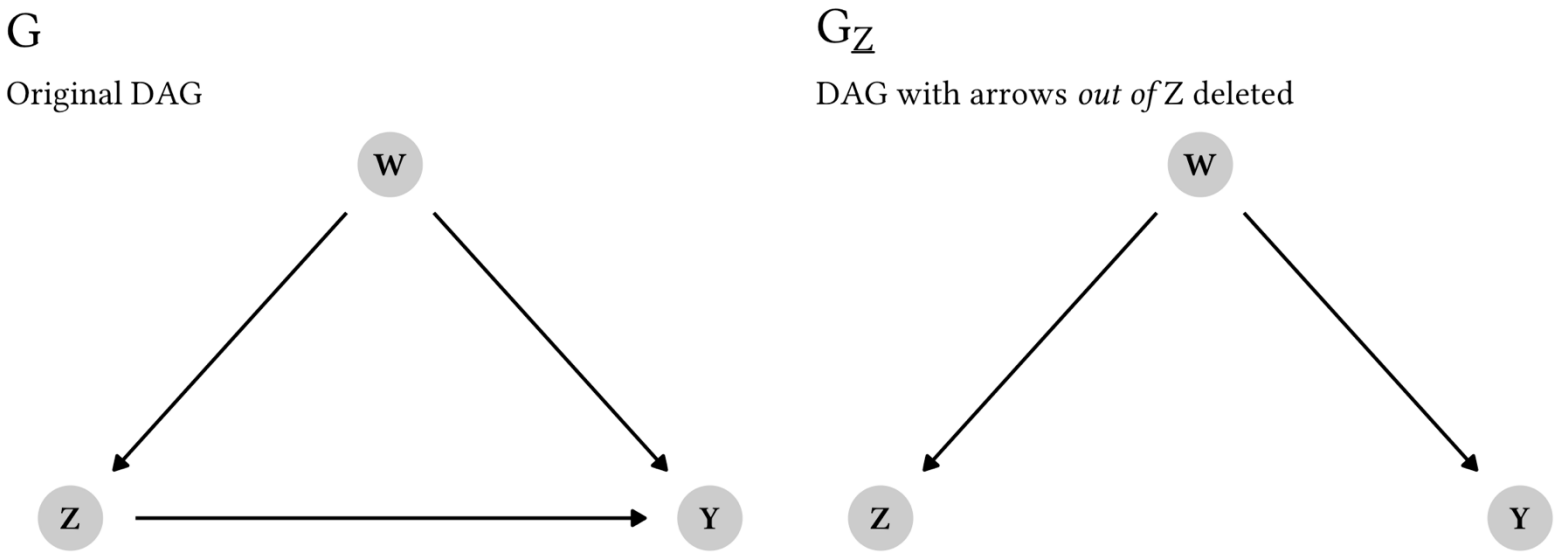


From https://www.andrewheiss.com/blog/2021/09/07/do-calculus-backdoors/

do-calculus for Deriving P(Y | do(X))

- **Rule 2 (inventions as obs.):** Interventions (or do(x)) can be treated as observations (or x) when the causal effect of a variable on the outcome (X→Y) only influences the outcome through directed paths.

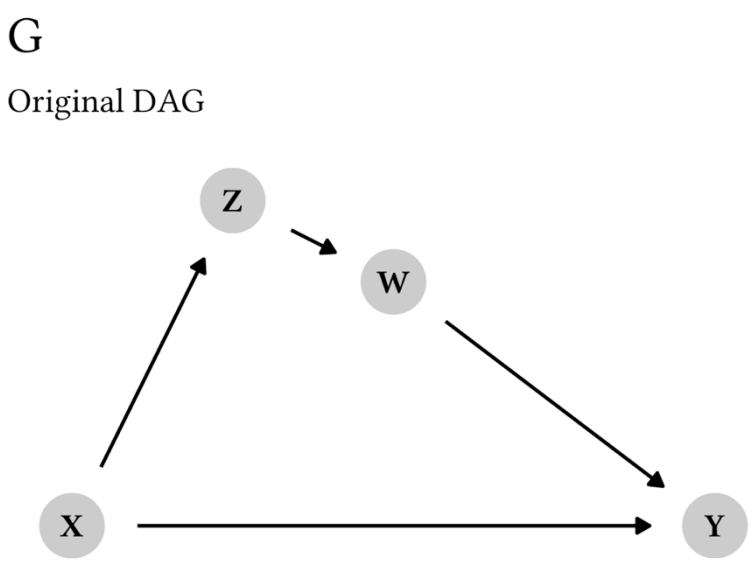
$$P(y \mid \mathrm{do}(z), \mathrm{do}(x), w) = P(y \mid z, \mathrm{do}(x), w) \qquad \text{if } (Y \perp Z \mid W, X)_{G_{X,Z}}$$

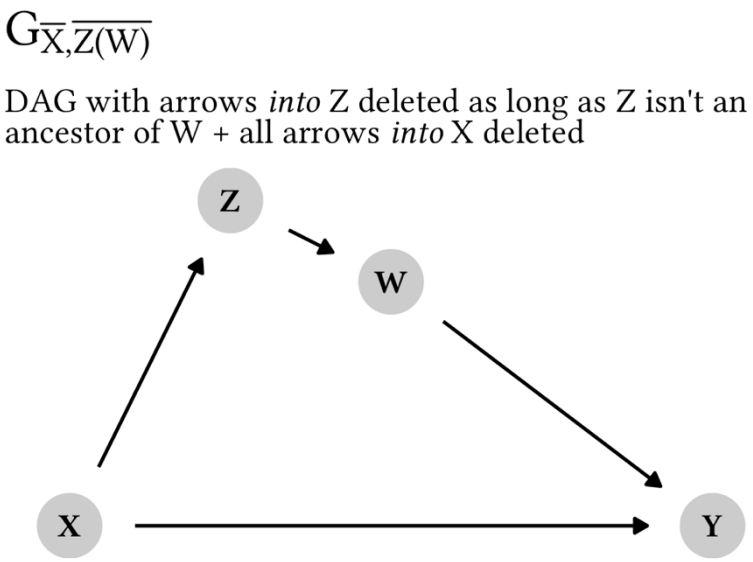


do-calculus for Deriving P(Y | do(X))

- Rule 3 (ignoring interventions): We can ignore an intervention (or a do-expression) if it doesn't influence the outcome through any uncontrolled path—we can remove do(z) if there is no causal association (or no unblocked causal paths) flowing from Z to Y.

$$P(y \mid \mathrm{do}(z), \mathrm{do}(x), w) = P(y \mid \mathrm{do}(x), w) \qquad \text{if } (Y \perp Z \mid W, X)_{G_{X,Z(W)}}$$





Back-Door Criterion

- Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.

Proof:

$$
\begin{aligned}
& [\text{Marginalization across } z + \text{chain rule for conditional probabilities}] \\
P(y \mid \mathrm{do}(x)) &= \sum_z P(y \mid \mathrm{do}(x), z) \times P(z \mid \mathrm{do}(x)) \\
& [\text{Use Rule 2 to treat } \mathrm{do}(x) \text{ as } x] \\
&= \sum_z P(y \mid x, z) \times P(z \mid \mathrm{do}(x)) \\
& [\text{Use Rule 3 to nuke } \mathrm{do}(x)] \\
&= \sum_z P(y \mid x, z) \times P(z \mid \text{nothing!}) \\
& [\text{Final backdoor adjustment formula!}] \\
&= \sum_z P(y \mid x, z) \times P(z)
\end{aligned}
$$

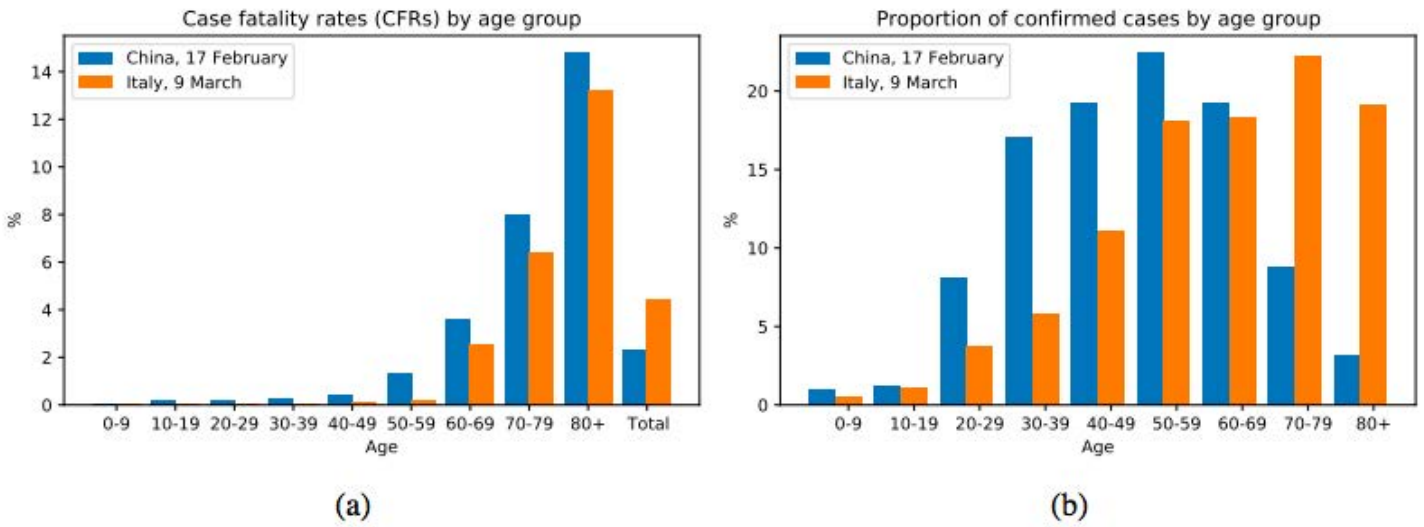
And front-door criterion

Comparison of Two Causal Models

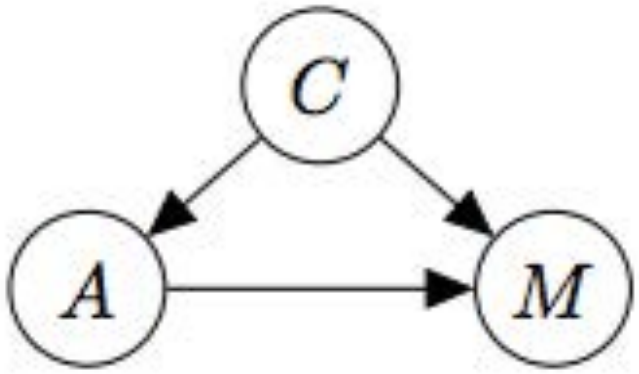
Model	Predict in IID setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

- 基于图的因果贝叶斯网络因果只有干预层的信息，所以只能回答干预和关联层的问题，而不能回答反事实问题。
- 基于结构的因果模型，它的能力最接近物理模型，用函数关系表示原因和结果之间的关系，三个层级的问题都能够回答。
- 因此可以认为，因果模型是从统计模型走向物理模型的过渡。

Example: 意大利的新冠病毒死亡率高于中国，那么是否能够说明意大利的医疗水平低于中国呢？



- 为了确定国家（C）对死亡(M) 的影响，需要消除年龄（A, confounder）的影响



国家-年龄-死亡因果回路示意图

From 因果科学：连接统计学、机器学习与自动推理的新兴交叉领域, 龚鹤扬、张妍, https://mp.weixin.qq.com/s/l-05jRYabGI-JoXedU-PLA

Do-Calculus，这里的本质就是干预，我们要求单纯国家这个因素对死亡率的影响，就要把年龄这个因素控制成一样(debias)，来求不同国家对死亡率的影响，具体做法如下：

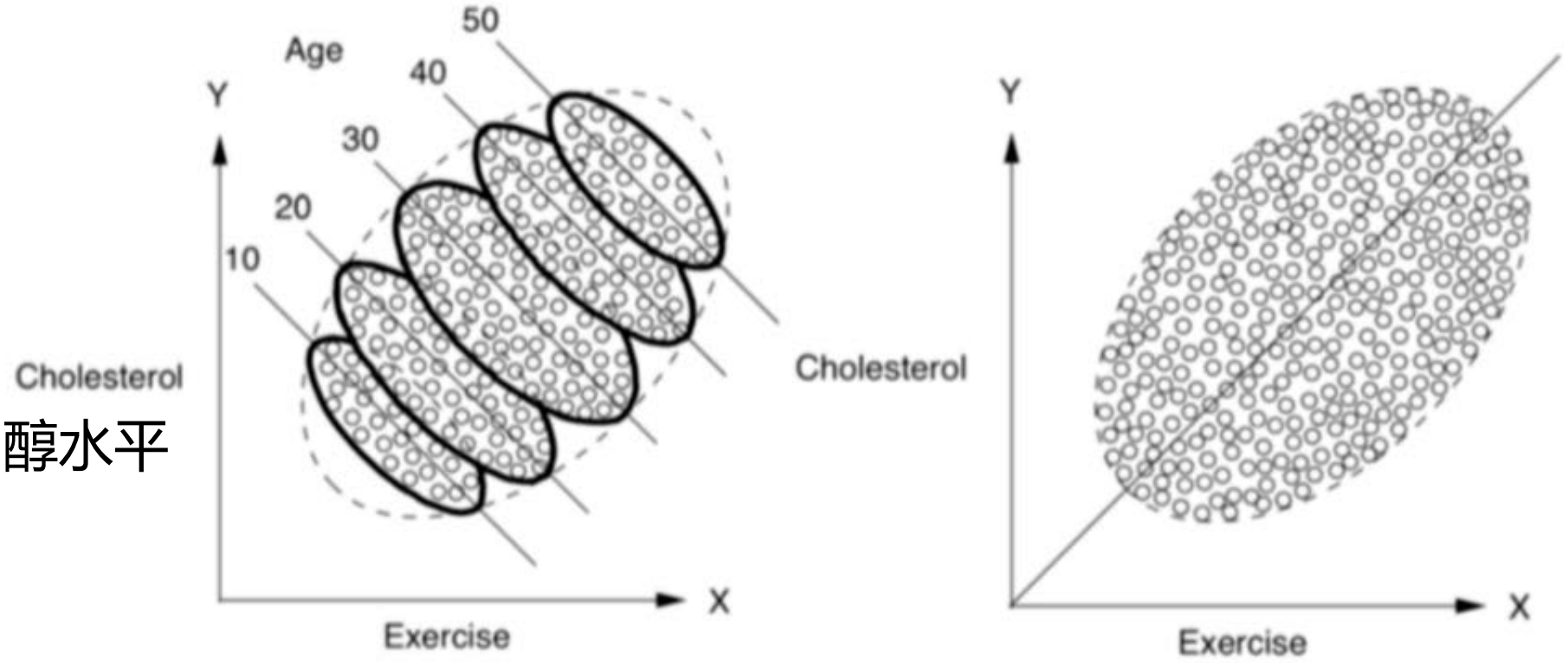
Applying our assumptions, in particular causal sufficiency, we can calculate the NDE to answer Q_{NDE} for our running example as follows,

$$
\begin{aligned}
\text{NDE}_{\text{China}\to\text{Italy}} &= \mathbb{E}[M_{A(\text{China})}|do(C=\text{Italy})] - \mathbb{E}[M_{A(\text{China})}|do(C=\text{China})] \\
&= \sum_a P_{A|do(C)}(a|do(\text{China}))\left[P_{M|do(A,C)}(1|do(a,\text{Italy})) - P_{M|do(A,C)}(1|do(a,\text{China}))\right] \\
&= \sum_a P_{A|C}(a|\text{China})\left[P_{M|A,C}(1|a,\text{Italy}) - P_{M|A,C}(1|a,\text{China})\right] \\
&= \mathbb{E}_{A|C=\text{China}}\left[\text{CDE}_{\text{China}\to\text{Italy}}(A)\right] \approx -0.8\%.
\end{aligned}
$$

结果得到：**国家对死亡率的直接因果效应反而是意大利比中国少 0.8%。**

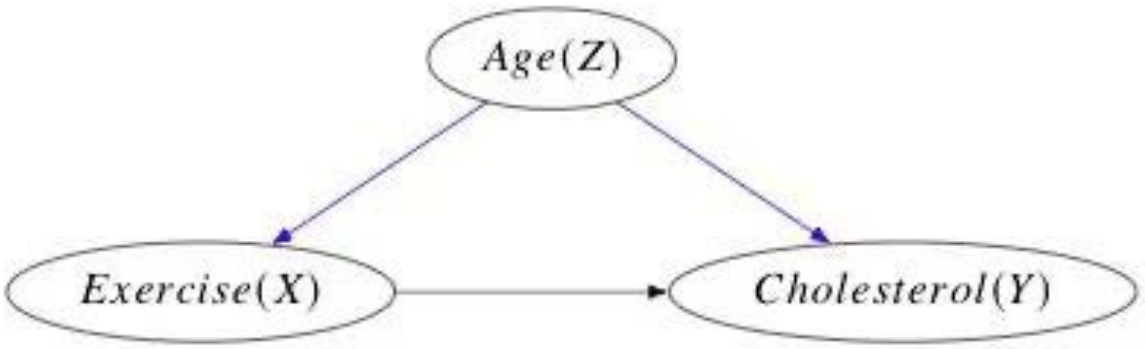
辛普森悖论

- 锻炼身体是否有利于健康？



Exercise is helpful in every age group but harmful for a typical person. **IS EXERCISE HELPFUL OR NOT?**

辛普森悖论



锻炼是否有益于健康？

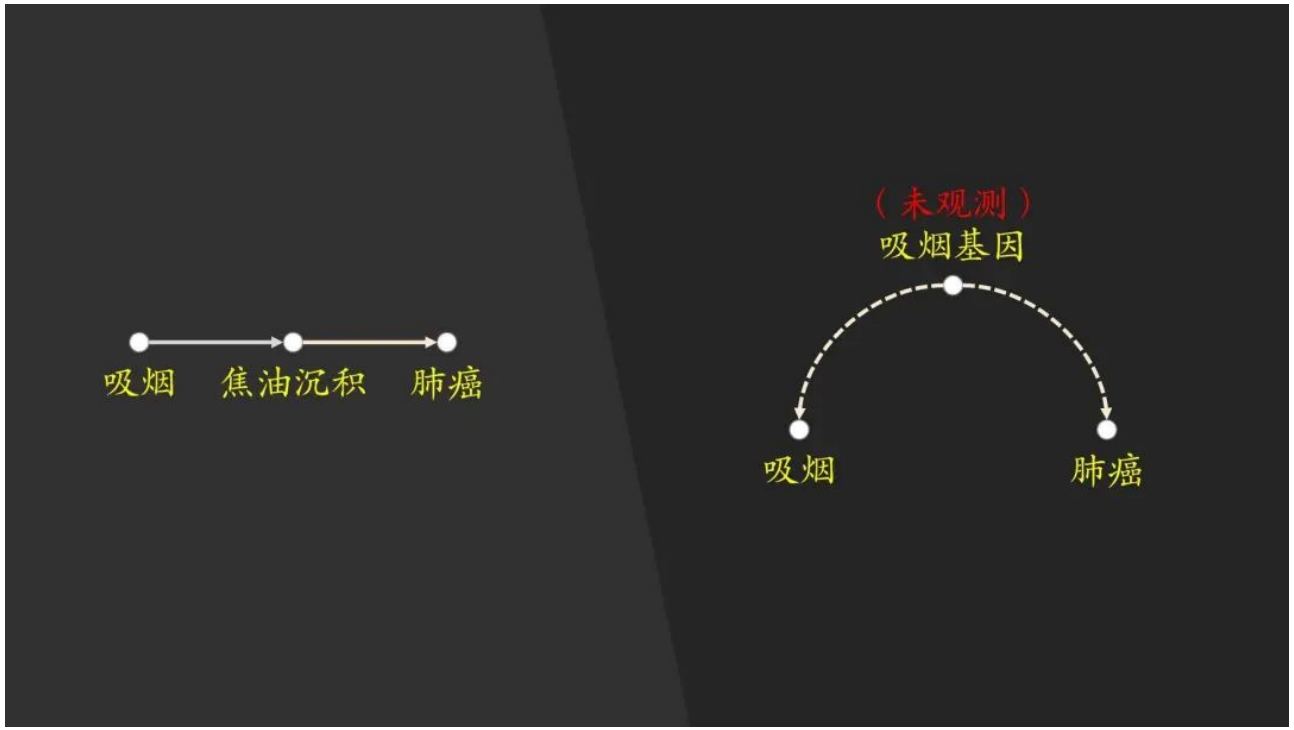
- 选择锻炼对健康的影响是：

$$E[Y|do(X=x)] = \sum_y yP(y|do(X=x)) = \sum_y \sum_z yP(y|x,z)P(z)$$

- 而不是：

$$E[Y|X=x] = \sum_y yP(y|x)$$

Example: 如何才能证明吸烟致癌？



- 一部分研究者认为，吸烟与肺癌的关系如左图所示，吸烟会直接导致肺部的焦油沉积，而焦油沉积会直接导致肺癌的产生。
- 而右侧的图是另一派研究者的观点，他们认为吸烟并不是肺癌产生的原因，而是有一个当时还没有观测到的吸烟基因，既会导致一个人容易尼古丁上瘾，又会导致肺癌的产生，因此吸烟和肺癌两者之间产生了相关性。
- 哪一种模型是对的？是否能够用数据出来呢？

The book of why, Judea Pearl, Chapter 5
https://mp.weixin.qq.com/s/SLBXgf8rkJaQwVzZhD_yAQ

幸存者偏差和因果倒置
Example: 费米悖论

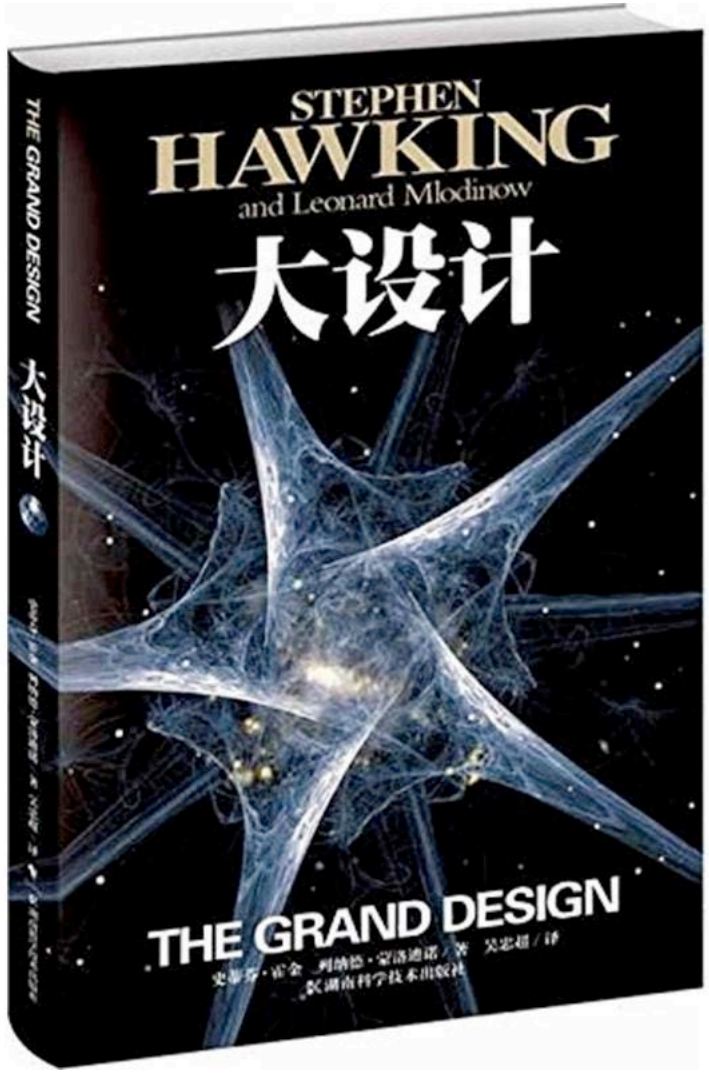
宇宙显著的尺度和年龄意味着高等地外文明应该存在。

The apparent size and age of the universe suggest that many technologically advanced extraterrestrial civilizations ought to exist.

但是，这个假设得不到充分的证据支持。

However, this hypothesis seems inconsistent with the lack of observational evidence to support it.

- 为什么没有地外文明？银河系大约有2500亿（2.5×10^{11}）颗恒星，可观测宇宙内则有700垓（7×10^{22}）颗。即使智慧生命以很小的概率出现在围绕这些恒星的行星中，那么仅仅在银河系内就应该有相当大数量的文明存在。
- 或者如果地外文明存在的话，为什么我们没有观测到智慧生命的迹象？无线电观测（射电望远镜）、遥感光谱分析均未能提供证据。



可能的一个解释是弱人择原理（幸存者偏差）：

- 看上去：地球的某些特质→地球产生了文明
- 实际上：没有上述因果关系，地球只是产生了文明，同时拥有一些特质。

- 在因果框架下，可以证明/证伪吗？

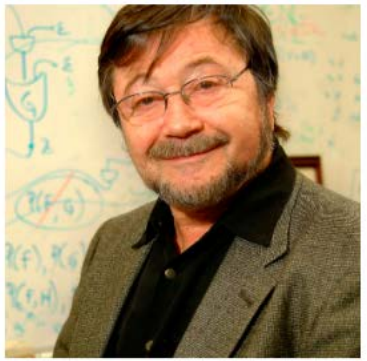
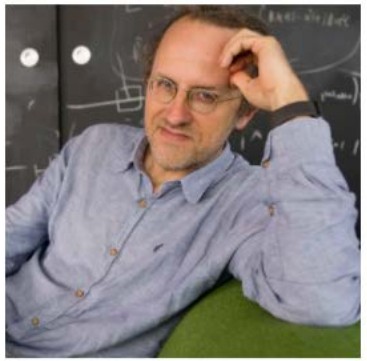
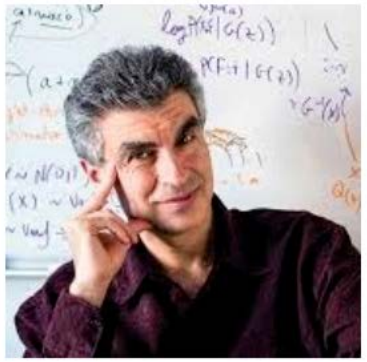
Is Time Important?

- Granger 因果检验，若X存在与否让下一时刻Y出现概率发生了变化，则X是Y的一个格兰杰因

$$\mathbb{P}[Y(t+1) \in A \mid \mathcal{I}(t)] \neq \mathbb{P}[Y(t+1) \in A \mid \mathcal{I}_{-X}(t)],$$

(There is a whole group of studies, but not the focus of this class.)

Causality for Machine Learning

Judea Pearl, Bernhard Schölkopf, Yoshua Bengio

1 简介 **3**

2 信息处理的机械化 **3**

3 从统计模型到因果模型 **4**
3.1 独立同分布数据驱动的方法 4
3.2 共同原因原则 5
3.3 结构因果模型(SCMs) 5

4 因果建模的层级 **6**

5 独立因果机制原理 **7**
5.1 机制相关的度量方法 8
5.2 算法独立性 9

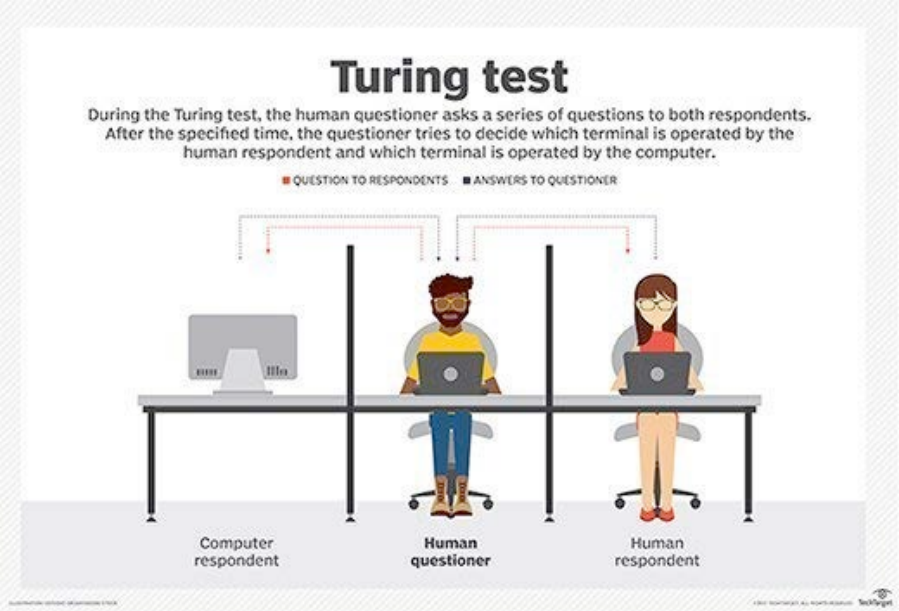
6 因果发现 **10**

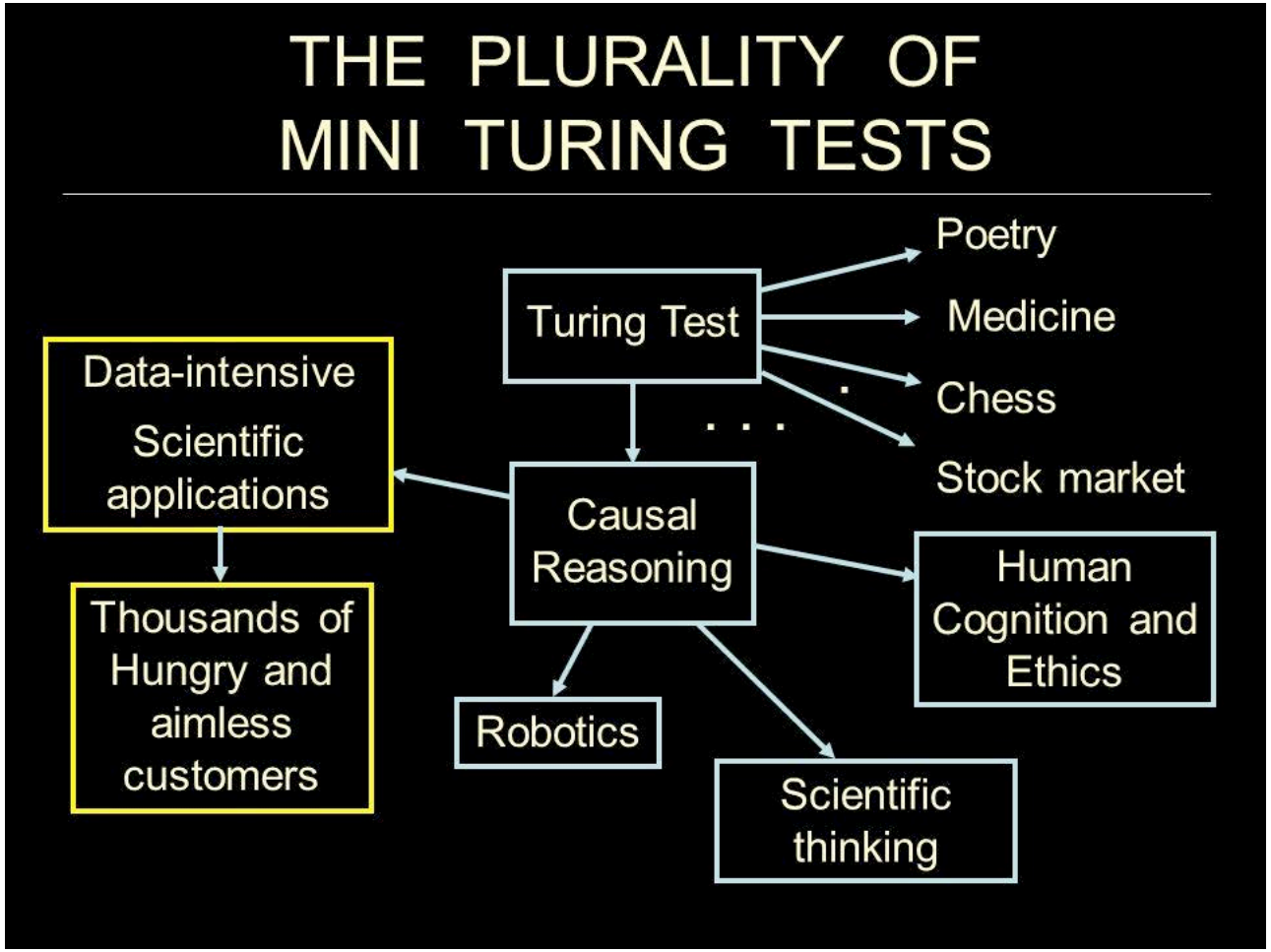
7 半兄弟回归和系外行星检测 **11**

8 不变性，鲁棒性和半监督学习 **11**
8.1 半监督学习 12
8.2 对抗脆弱性 12
8.3 多任务学习 13
8.4 强化学习 13

9 因果表示学习 **14**
9.1 学习可迁移机制 15
9.2 学习解纠缠表示 15
9.3 学习干预模型和推理 16

Pearl's Mini Turning Test





Potential Scientific Discovery

COGNITIVE SCIENCE 5, 31–54 (1981)

Data-Driven Discovery of Physical Laws

PAT LANGLEY

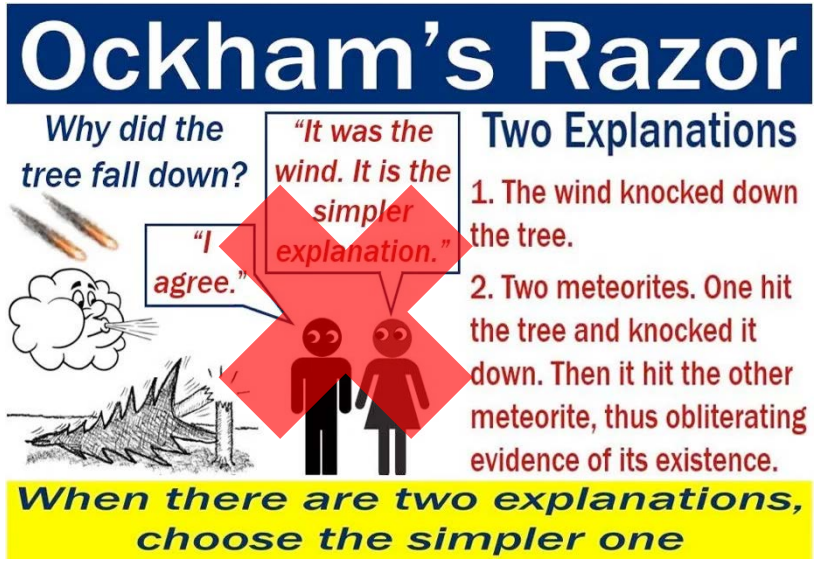
Department of Psychology
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

BACON.3 is a production system that discovers empirical laws. Although it does not attempt to model the human discovery process in detail, it incorporates some general heuristics that can lead to discovery in a number of domains. The main heuristics detect constancies and trends in data, and lead to the formulation of hypotheses and the definition of theoretical terms. Rather than making a hard distinction between data and hypotheses, the program represents information at varying levels of description. The lowest levels correspond to direct observations, while the highest correspond to hypotheses that explain everything so far observed. To take advantage of this representation, BACON.3 has the ability to carry out and relate multiple experiments, collapse hypotheses with identical conditions, ignore differences to let similar concepts be treated as equal, and to discover and ignore irrelevant variables. BACON.3 has shown its generality by rediscovering versions of the ideal gas law, Kepler's third law of planetary motion, Coulomb's law, Ohm's law, and Galileo's laws for the pendulum and constant acceleration.

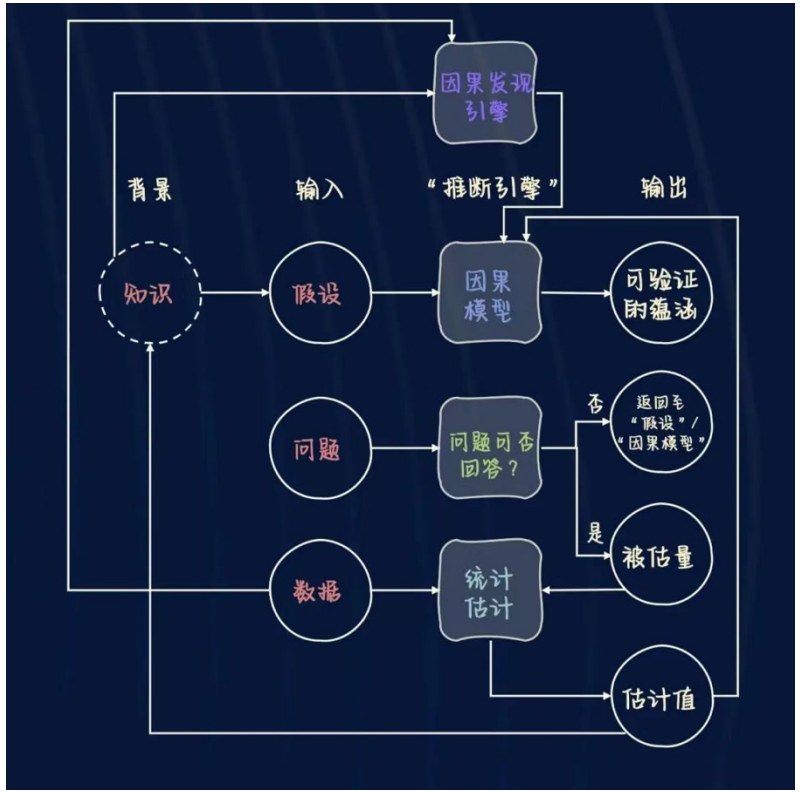
Summary

- Under the framework of causal inference, models must come first: Science first, data later.
- The question is: Can we use observables to identify the model and how?

Karl Pearson (1892): Causation can never be proven–that mere data is all there is to science.

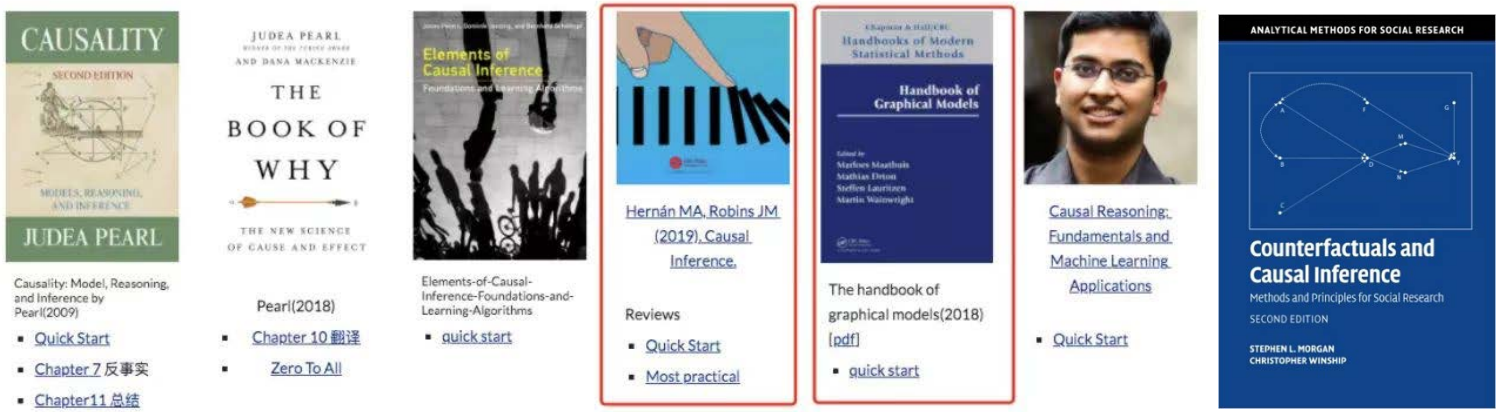


因果引擎

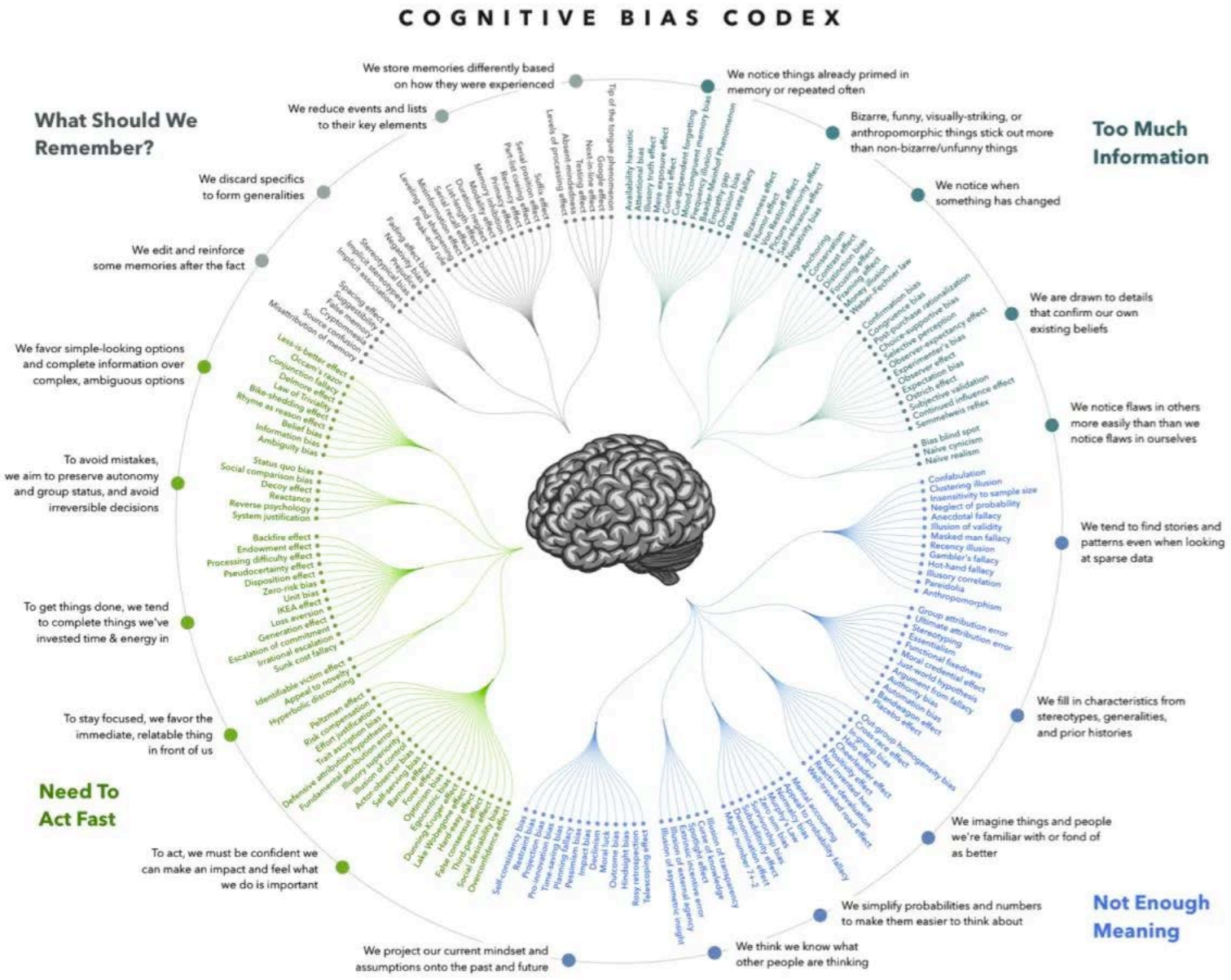


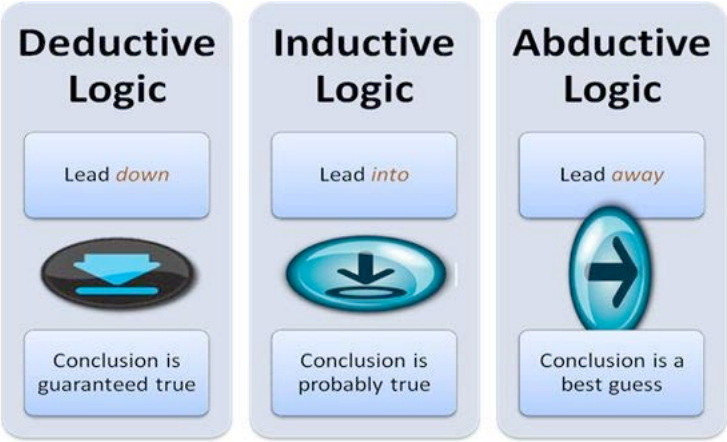
References

- 集智俱乐部公众号“因果推断”
- A Crash Course in Causality - https://www.coursera.org/learn/crash-course-in-causality#syllabus
- Causality by Jonas Peters https://www.youtube.com/watch?v=zvrcyqcN9Wo
- Causal Inference by Joseph Hogan https://www.youtube.com/watch?v=NzBTnzWiQno
- https://why19.causalai.net/



Logic


COGNITIVE BIAS CODEX
What Should We Remember?
We store memories differently based on how they were experienced
We reduce events and lists to their key elements
We discard specifics to form generalities
We edit and reinforce some memories after the fact
Too Much Information
We notice things already primed in memory or repeated often
Bizarre, funny, visually-striking, or anthropomorphic things stick out more than non-bizarre/unfunny things
We notice when something has changed
We are drawn to details that confirm our own existing beliefs
We notice flaws in others more easily than than we notice flaws in ourselves
We tend to find stories and patterns even when looking at sparse data
We fill in characteristics from stereotypes, generalities, and prior histories
We imagine things and people we're familiar with or fond of as better
Not Enough Meaning
We simplify probabilities and numbers to make them easier to think about
We think we know what other people are thinking
We project our current mindset and assumptions onto the past and future
Need To Act Fast
To act, we must be confident we can make an impact and feel what we do is important
To stay focused, we favor the immediate, relatable thing in front of us
To get things done, we tend to complete things we've invested time & energy in
To avoid mistakes, we aim to preserve autonomy and group status, and avoid irreversible decisions
We favor simple-looking options and complete information over complex, ambiguous options


Deductive Logic
Lead down
Conclusion is guaranteed true
Inductive Logic
Lead into
Conclusion is probably true
Abductive Logic
Lead away
Conclusion is a best guess