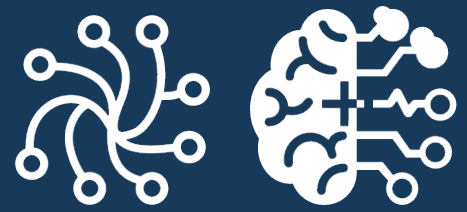
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



类脑认知-因果推理

叶南阳

第一节　因果推理的背景介绍

什么是因果推理?

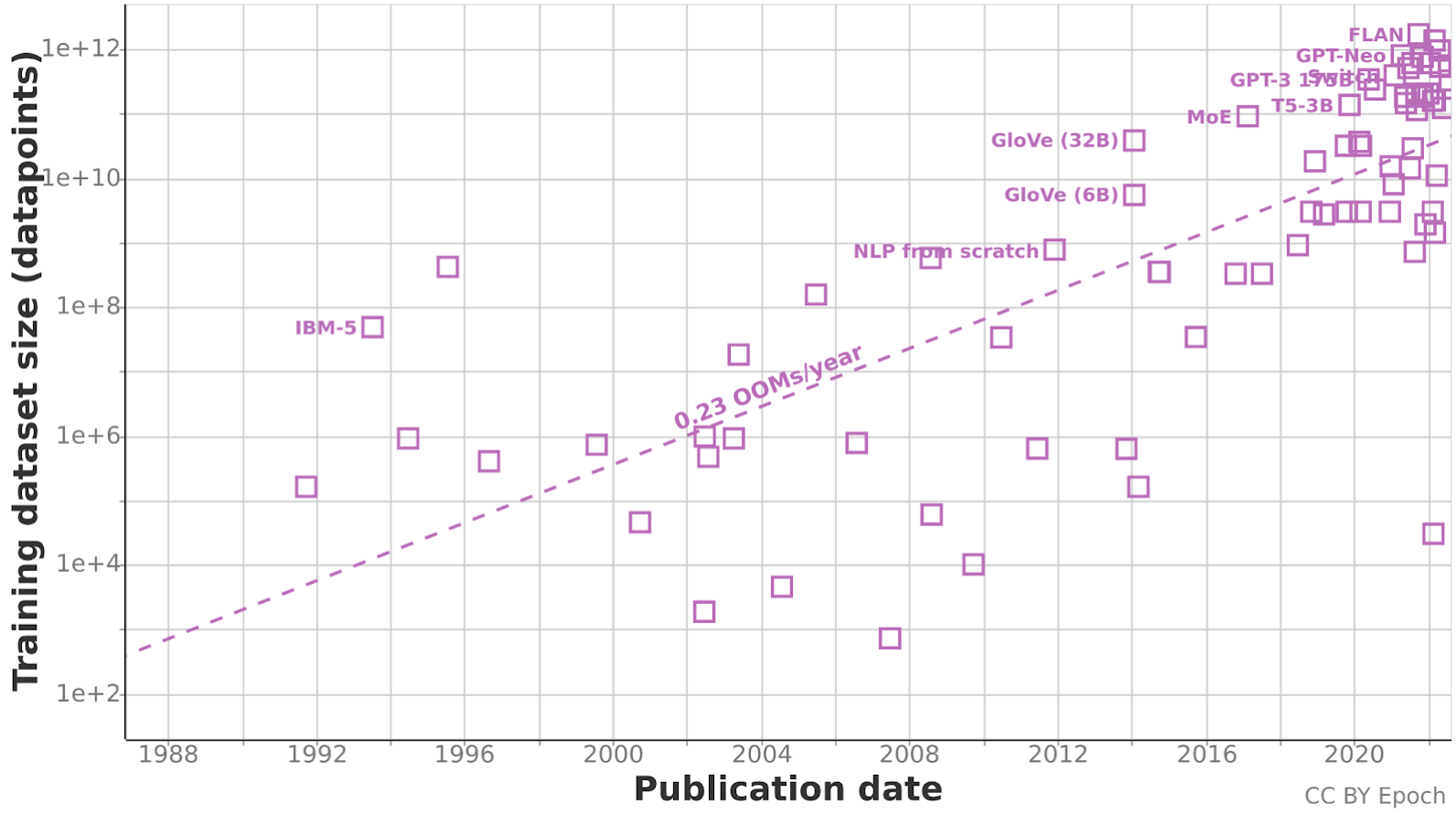
谷歌给出的解释：Causal inference refers to an intellectual discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data.

因果推理是一种考虑假设、研究设计和策略评估的思维方式，是**类脑智能**的一个重要部分。

维基百科给出的解释：Causal inference is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system.

因果推理是确定作为更大系统组成部分的某一特定现象怎样施加独立、实际影响的过程。

为什么我们需要因果推理？ I.I.D 假设数据中的每个样本来自于对同一分布的随机采样的结果。测试集的分布和训练集分布一样。导致提高机器学习系统泛化性能依赖大量数据。



随着时间的推移，论文中使用的数据集数据量越来越多。

为什么我们需要因果推理？ 传统的机器学习存在许多局限性。

机器学习严重依赖于大量预先设置好的数据集来提升准确率，在训练过程中难以学习到因果性。机器学习往往漠视许多人类和动物视觉常用的信息，例如：

(1)环境干扰（intervention）

(2)域偏移（domain shift）

(3)时序结构（temporal structure）

训练集:白天

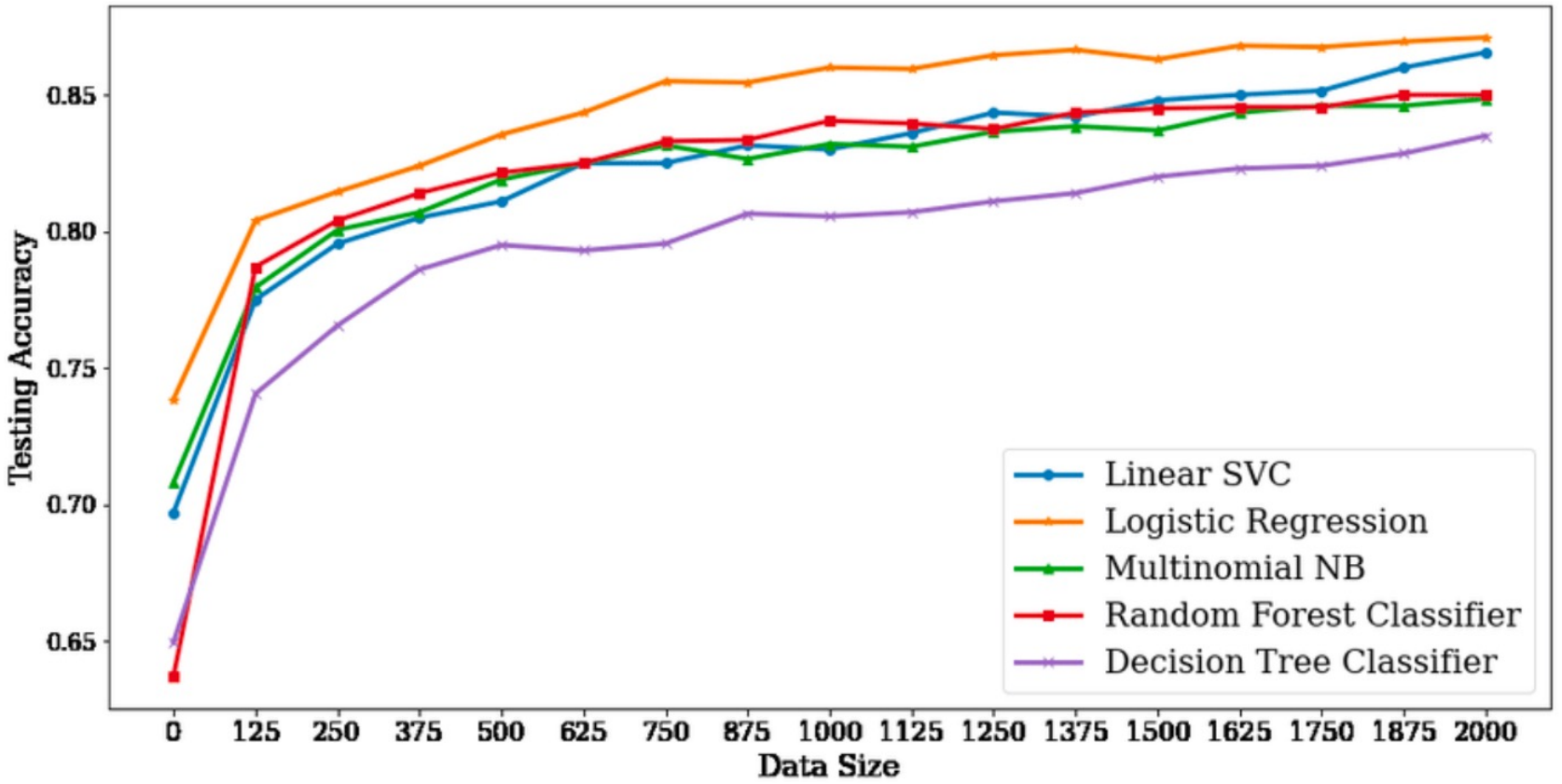
Generalization

测试集:黑夜

在机器学习中，我们当然希望能排除这些因素的影响，而当前机器学习的重点聚焦在基于独立同分布(I.I.D)数据的大规模模式识别[1]。

[1] Towards Causal Representation Learning, *Proceedings of IEEE.*

为什么我们需要因果推理？ 然而单纯通过数据集的大小增长，我们能解决机器学习系统的**泛化问题**么？



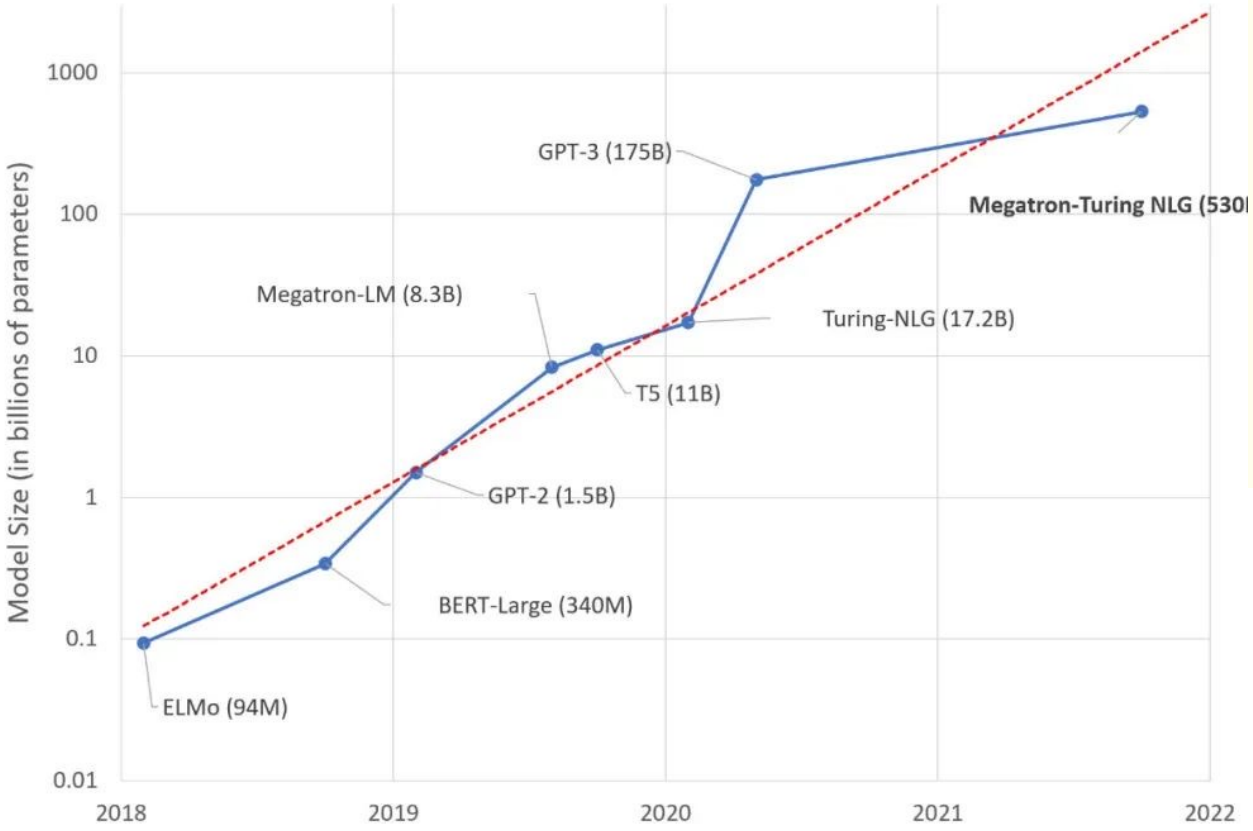
随着数据集，论文中使用的数据集数据量越来越多，精度却逐渐开始饱和。

为什么我们需要因果推理？ 然而单纯通过数据集的大小增长，我们能解决机器学习系统的**泛化问题**么？

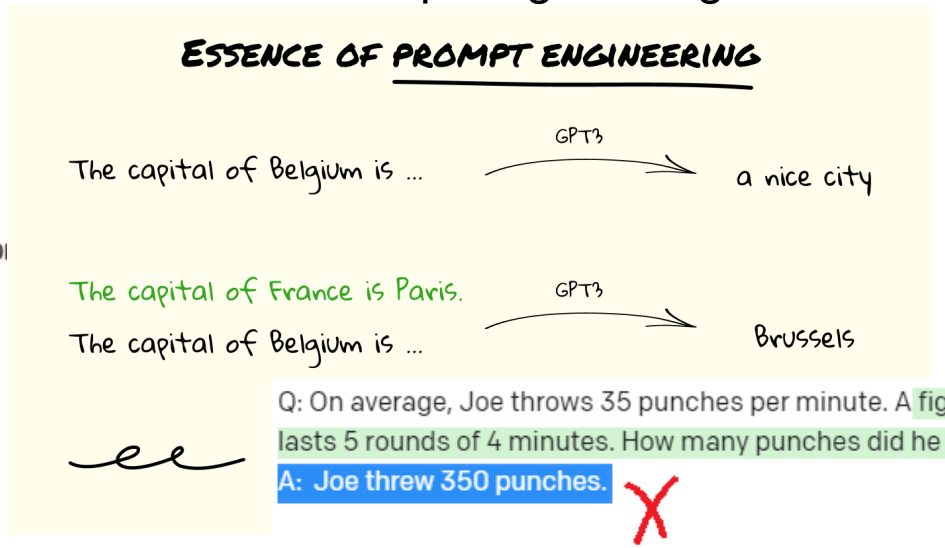
白色卡车被误识别为天空背景。

自动驾驶行业花费了巨大的成本来搜集数据，然而仍然不能完全避免事故，
而在这些情况中，清醒的人可以完全避免。

为什么我们需要因果推理？ 通过提升模型复杂度，我们能解决机器学习系统的**泛化问题**么？



Prompt engineering



Q: On average, Joe throws 35 punches per minute. A fight lasts 5 rounds of 4 minutes. How many punches did he throw?

A: Joe threw 350 punches. ✗

Q: On average, Joe throws 35 punches per minute. A fight lasts 5 rounds of 4 minutes. How many punches did he throw?

A: Let's think step by step.

In one minute, Joe throws 35 punches.

In four minutes, Joe throws 4 times as many punches as in one minute. So he throws 4 * 35 = 140 punches in four minutes. ✓

In five rounds, Joe throws 5 times as many punches as in one round. So he throws 5 * 140 = 700 punches in five rounds.

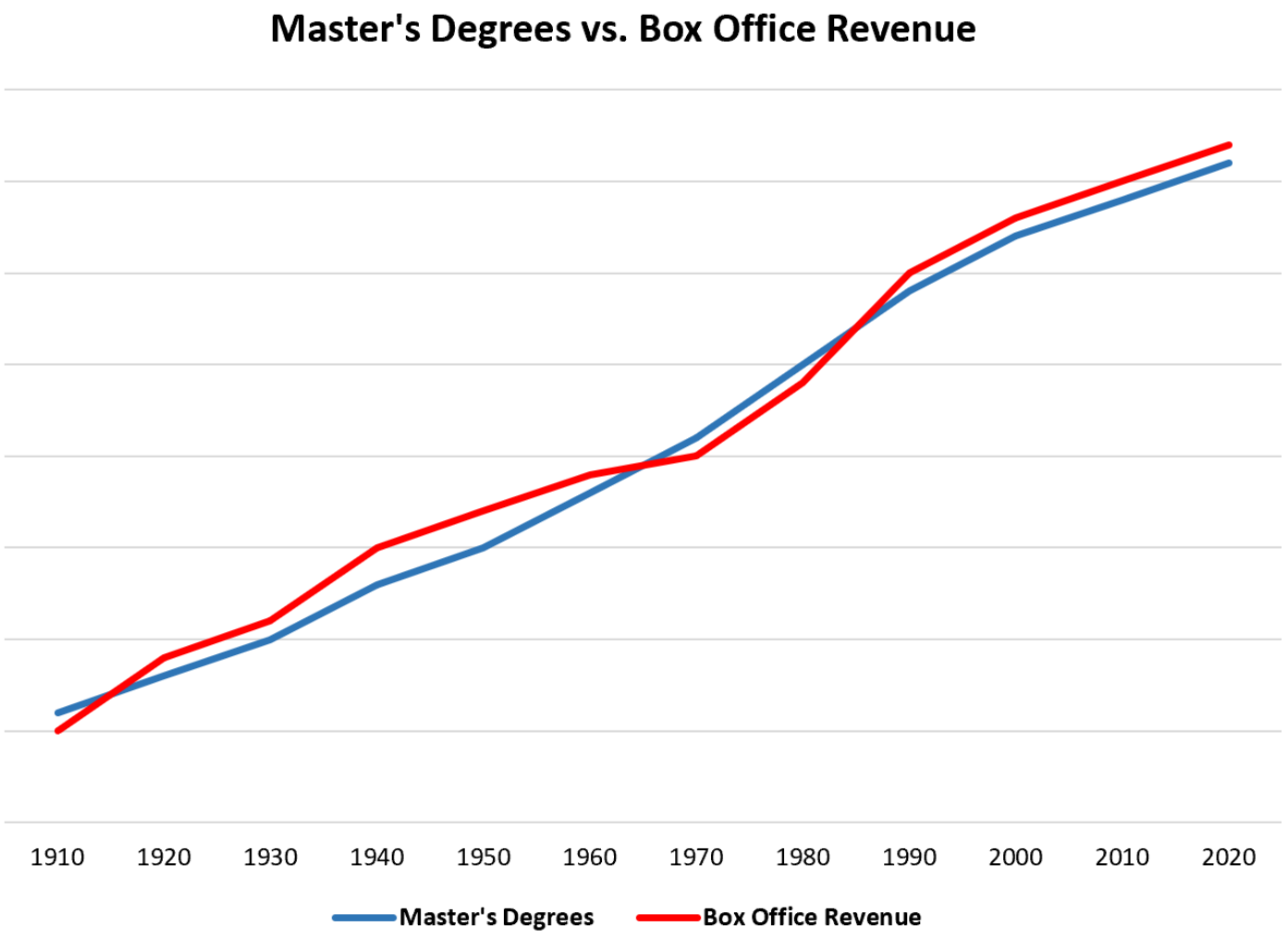
（语言）大模型复杂度近年来成指数级上升，然而很难讲大模型真正理解了人类语言。

为什么需要因果推理?

我们需要重新思考机器学习的范式

第二节 一切从一个“简单”问题开始

Correlation does not imply causality



硕士学位的增加

导致电影票房上升?

1.Simpson’s paradox 辛普森悖论

一种新药700例临床试验的对比，350例服药，350例不服药

患者	患者服药情况		患者未服药情况	
	痊愈患者数	痊愈率%	痊愈患者数	痊愈率%
男性患者	81/87	93	234/270	87
女性患者	192/263	73	55/80	69
合计	273/350	78	289/350	83

基于这个结果，可以得出这种药物对患者有效的结论么？

1.Simpson’s paradox 辛普森悖论

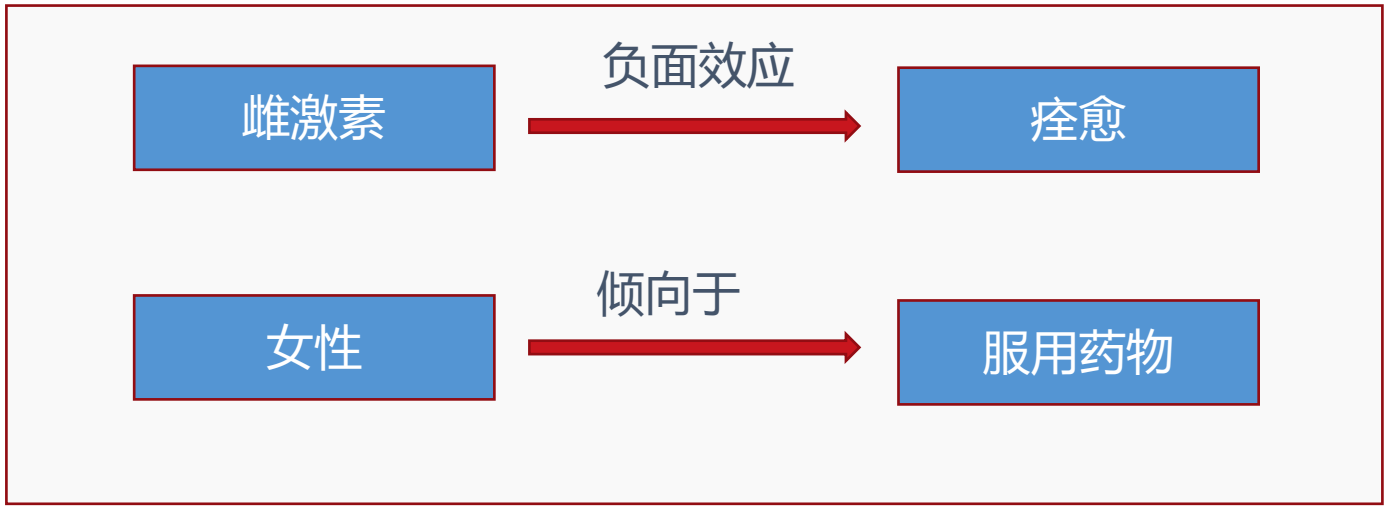
一种新药700例临床试验的对比，350例服药，350例不服药

患者	患者服药情况		患者未服药情况	
	痊愈患者数	痊愈率%	痊愈患者数	痊愈率%
男性患者	81/87	93	234/270	87
女性患者	192/263	73	55/80	69
合计	273/350	78	289/350	83

从表中可以看出，*如果已知性别，则可以使用药物，否则不能使用?*

1. Simpson’s paradox 辛普森悖论

数据背后的因果机制：



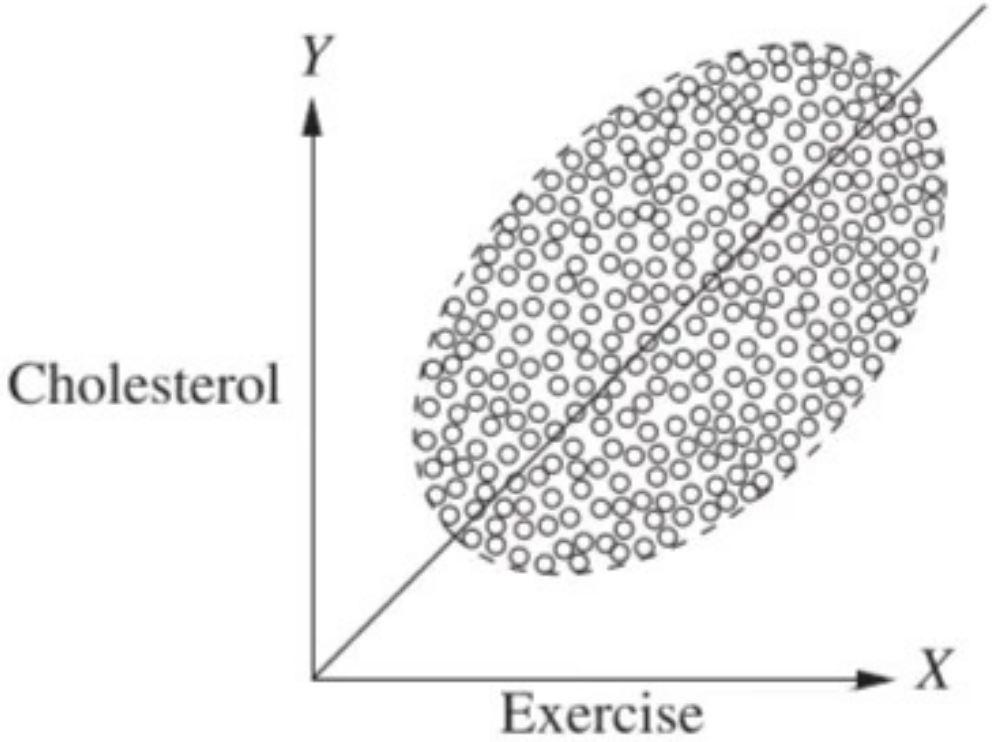
根据因果机制的分析方法：

比较同一性别的受试者，确保痊愈率差异并不归因于雌激素

结论：药物有效！

1. Simpson’s paradox 辛普森悖论

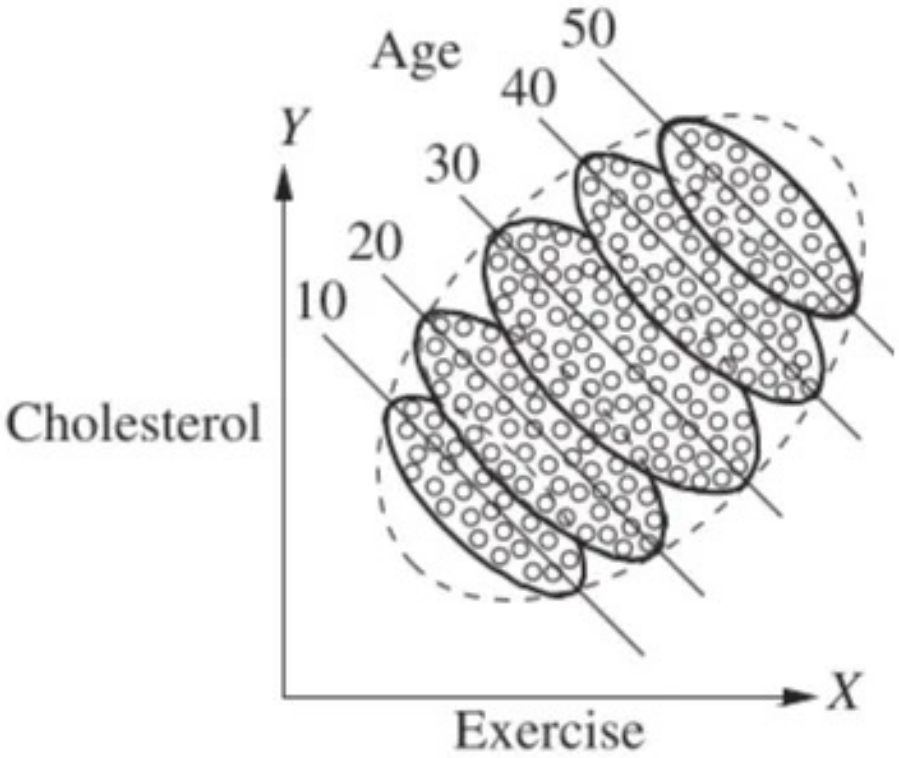
锻炼量与胆固醇关系图



越锻炼胆固醇越高?

1. Simpson’s paradox 辛普森悖论

锻炼量与胆固醇关系图



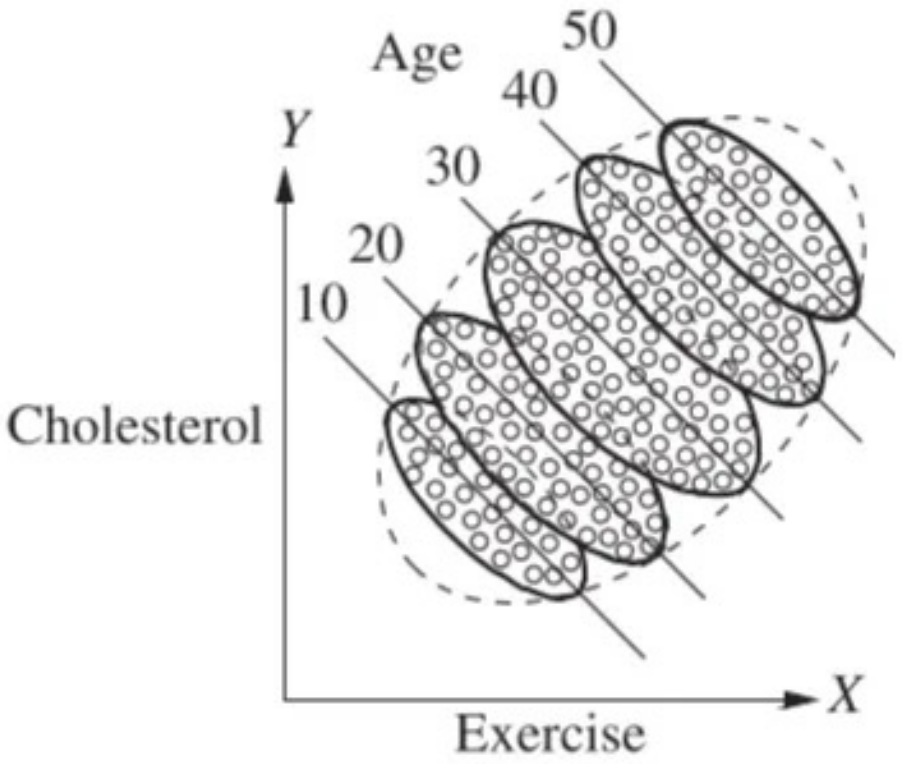
If we take a closer look

按照年龄划分，对每个年龄段来看，

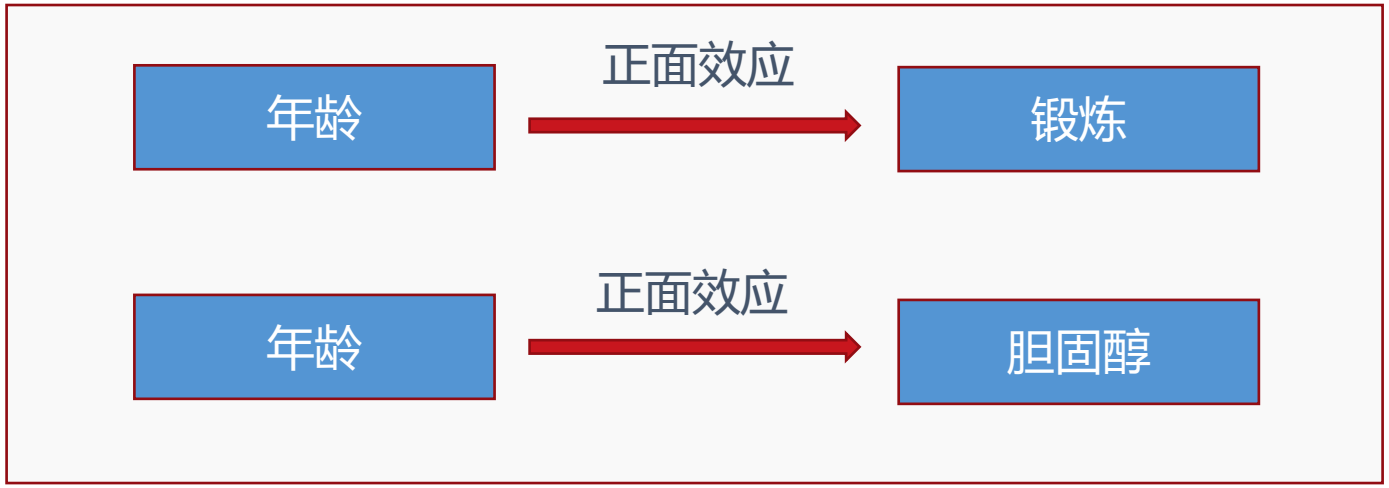
锻炼都可有效降低胆固醇

1. Simpson's paradox 辛普森悖论

锻炼量与胆固醇关系图



数据背后的因果机制：



根据因果机制的分析方法：

比较同一年龄段的受试者，确保胆固醇差异并不归因于年龄

结论：锻炼有效！

1. Simpson’s paradox 辛普森悖论

我们通过数据分类，解决了问题，然而数据分类是否总是有效?

患者	患者服药情况		患者未服药情况	
	痊愈患者数	痊愈率%	痊愈患者数	痊愈率%
患者(血压低)	81/87	93	234/270	87
患者(血压高)	192/263	73	55/80	69
合计	273/350	78	289/350	83

如果降低血压是药物的效果之一，则基于血压分类没有意义。但血压对治疗有影响则不同。

2. 相关性不等于因果性

- 例子：穿鞋睡觉的人往往醒来会头痛。
- 可能原因：睡前饮酒
- 1.穿鞋睡觉和不穿鞋睡觉的人群有别的关键不同（睡前饮酒），于是穿鞋就成了干扰因素

- 我们最终观察到的相关性是因果性和干扰因素的叠加

2. 相关性不等于因果性

思考题：下面的命题有什么错误?

- (a) 数据表明，随着火灾数量的增加，消防员的数量也会增加，因此，为了减少火灾，应该减少消防员的数量。
- (b) 数据表明，匆忙赶赴会议的人，通常会迟到。因此不要着急，否则你会迟到。

2. 相关性不等于因果性

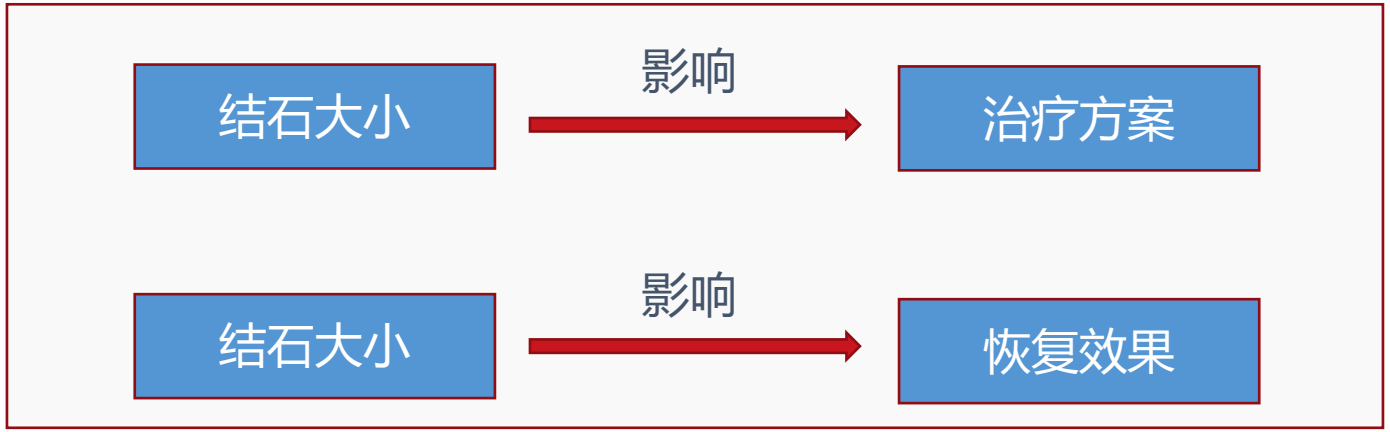
例题：对于下面的每一个因果问题，判断应该使用分类数据还是总体数据来获得正确的结论。

(a)肾结石的治疗有两种方法：方案A和方案B。医生对大的结石（因此病情更严重）更倾向于采用方案A，对小的结石倾向于方案B。如果一个患者不知道其体内结石的大小，为确定哪种治疗方案更有效，应该检索总体人群的数据还是结石大小不同亚群的数据？

2. 相关性不等于因果性

例题：对于下面的每一个因果问题，判断应该使用分类数据还是总体数据来获得正确的结论。

(a) 数据背后的因果机制：



根据因果机制的分析方法：

比较同一结石大小的患者，确保治疗效果差异并不归因于大小

2. 相关性不等于因果性

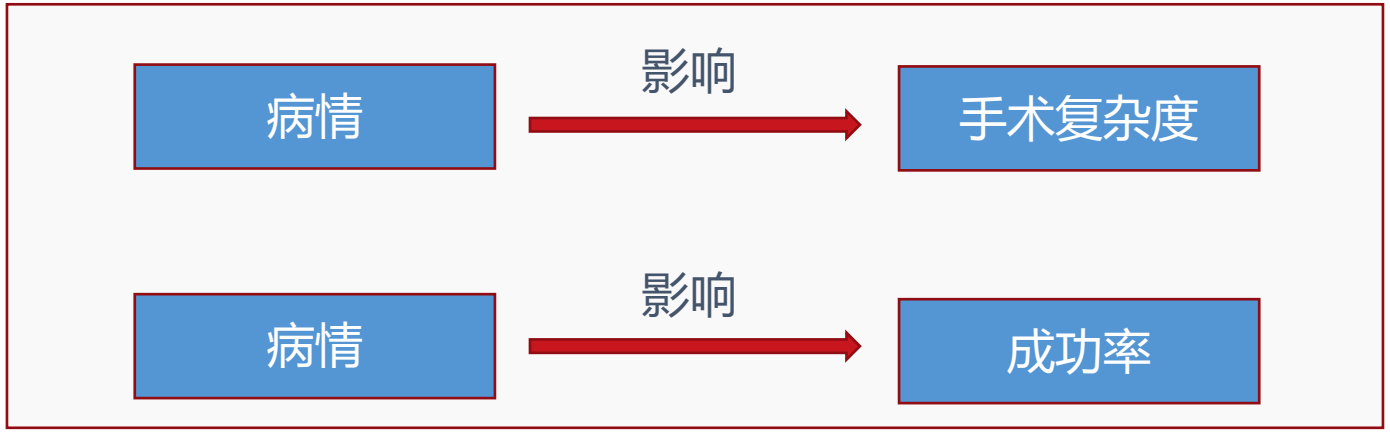
例题：对于下面的每一个因果问题，判断应该使用分类数据还是总体数据来获得正确的结论。

(b) 在一个小镇上有两位医生，二人在其职业生涯中做过100例手术。手术分为两类：一类非常复杂，一类非常简单。第一位医生经常做简单的手术，第二位医生经常做复杂的手术。假如你需要做手术，但不知道自己的情况是属于简单情况还是复杂情况。为尽可能提高手术成功概率，你应该查阅每位医生的总成功率，还是分别查阅他们各自简单和复杂手术的成功率?

2. 相关性不等于因果性

例题：对于下面的每一个因果问题，判断应该使用分类数据还是总体数据来获得正确的结论。

(a) 数据背后的因果机制：



根据因果机制的分析方法：

比较同一病情的患者，确保治疗效果差异并不归因于病情情况

2. 相关性不等于因果性

为评估一种新药的疗效，进行了一项随机试验。总体来说，50%的患者被分配服用新药，50%的患者接受安慰剂。在试验的前一天，某位护士给一些表现抑郁的患者分发棒棒糖，他们中的大多数是被指定第二天接受治疗的（也就是说，护士这轮查房正巧经过诊疗病房）。但是，实验数据显示了一个辛普森悖论：尽管证明药物对总体受试者是有益的，但在得到棒棒糖的亚群中和没有棒棒糖的亚群中，服用药物的患者比不服用药物的患者更不容易痊愈。假设吃棒棒糖本身对痊愈没有任何影响，回答以下问题。

(a)*药物对总体受试者是有益的还是有害的?*

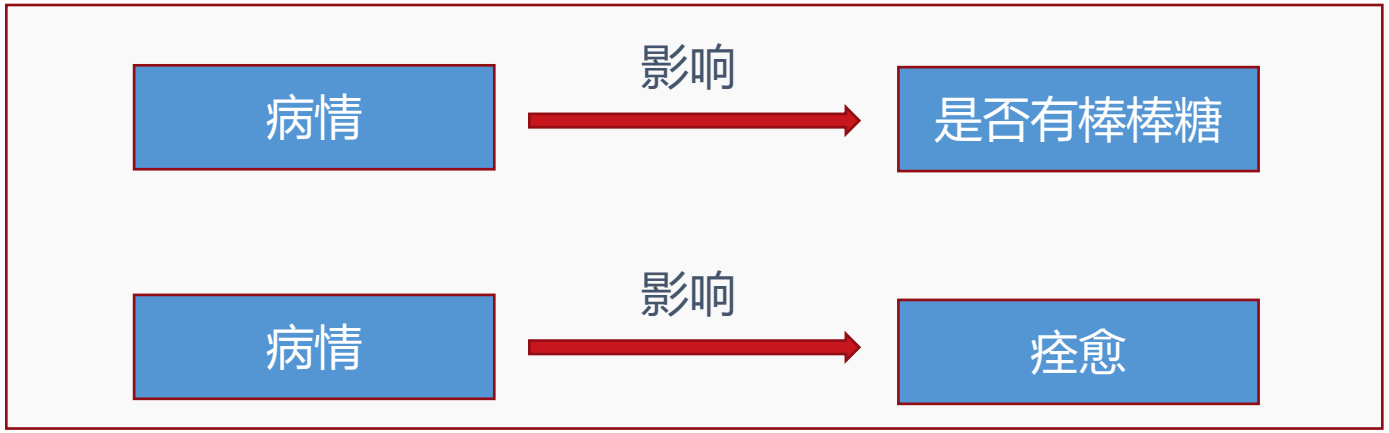
(b)*这个例子和此前按照性别分类的例子不一样么？*

(c)*如果棒棒糖在试验后的第二天分发（按同样的准则），答案会不同么？*

【提示：基于这样的事实，即接受棒棒糖表明患者更可能被安排药物治疗以及表现抑郁，而抑郁是降低痊愈可能性的风险因素】

2. 相关性不等于因果性

数据背后的因果机制：



根据因果机制的分析方法：

比较同一病情的患者，确保治疗效果差异并不归因于棒棒糖情况

2. 相关性不等于因果性

(a)药物对总体受试者是有益的还是有害的?

有益

(b)这个例子和此前按照性别分类的例子不一样么?

不一样

(c)如果棒棒糖在试验后的第二天分发（按同样的准则），答案会不同么?

不一定，取决于分发的标准。

2. 相关性不等于因果性

Training dataset bias in autonomous driving

在训练集中，天空大量是白色的。导致神经网络根据大片白色来判断是否是天空

然而，白色不是天空的本质特征

第三节 Potential Outcome

Potential Outcome Framework

Individual treatment effect (ITE)：对整体当中每一个个体 *i*，treatment的效果为

$$\tau_i \triangleq Y_i(1) - Y_i(0)$$

对于整体而言，Y 可视为随机变量，对单个个体*i*，Y_i 视为确定值。

Average treatment effect (ATE)：对整体而言，treatment的效果平均为

$$\tau \triangleq \mathrm{E}[Y_i(1) - Y_i(0)] = \mathrm{E}[Y(1) - Y(0)]$$

符号定义

T	施加的treatment (0或者1），例如是否服药
Y	感兴趣的结果变量，例如是否治愈
X	其他的影响变量，例如性别
i	指代的实验对象编号
$Y_i(1)$	施加的treatment为1的潜在结果
$Y_i(0)$	施加的treatment为0的潜在结果

ATE如何计算?

Recall1：我们知道统计相关性计算方法：

$$E[Y|T=1]-E[Y|T=0]$$

Recall2：我们知道Expectation运算的线性特性

$$\tau \triangleq \mathrm{E}[Y(1)-Y(0)]=\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)]$$

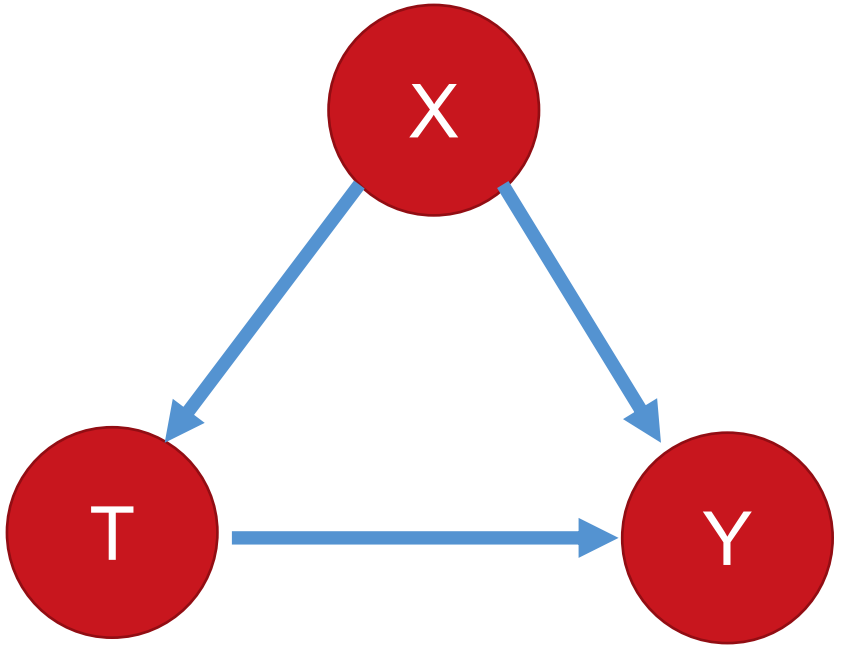
Thus：我们可以认为

$$E[Y|T=1]=E[Y(1)]?$$

$$E[Y|T=0]=E[Y(0)]?$$

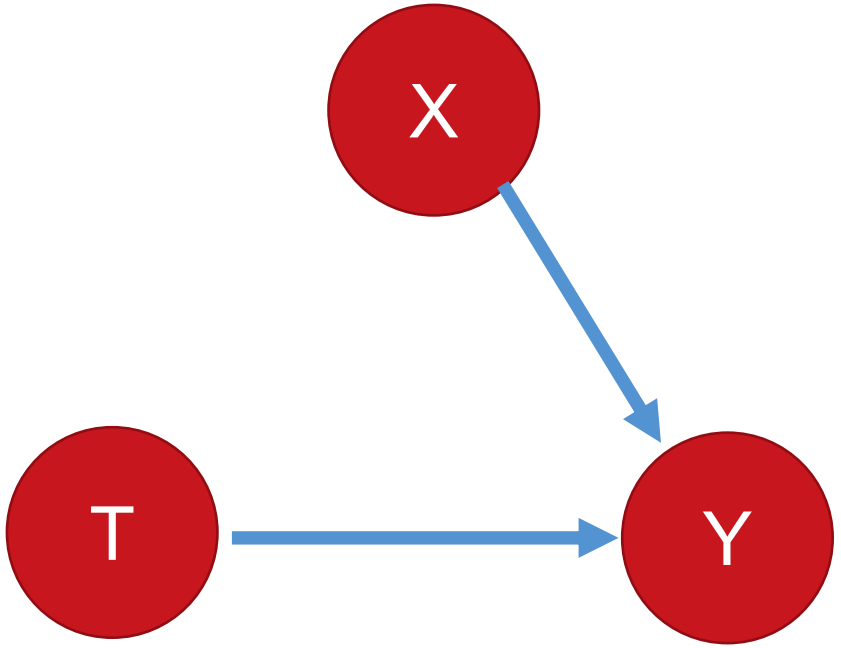
ATE如何计算?

如果X同时影响T和Y(Confounding)，则不成立



ATE如何计算?

如果treatment为随机指定，和X无关，则成立



Assumption (Ignorability/Exchangeability)

$$\big(Y(1), Y(0)\big) \perp T$$

即是否施加treatment和T完全独立

$$E[Y(1)] - E[Y(0)] = E[Y(1)|T = 1] - E[Y(0)|T = 0]$$

$$= E[Y|T = 1] - E[Y|T = 0]$$

Exchangeability:施加新的treatment会和原来的treatment观测到一样的结果。

Definition (Identifiability) 如果一个因果量 $E[Y(t)]$ 是identifiable的，即我们可以从纯统计量来计算 $E[Y|t]$

Note:1．Ignorability 假设对identifiability非常重要，但是现实数据中经常存在confounding。

Note:2．通过随机控制变量实验的方法。可以排除confounding。

Note:3．如果没有随机控制变量实验，是否还有其他方法。

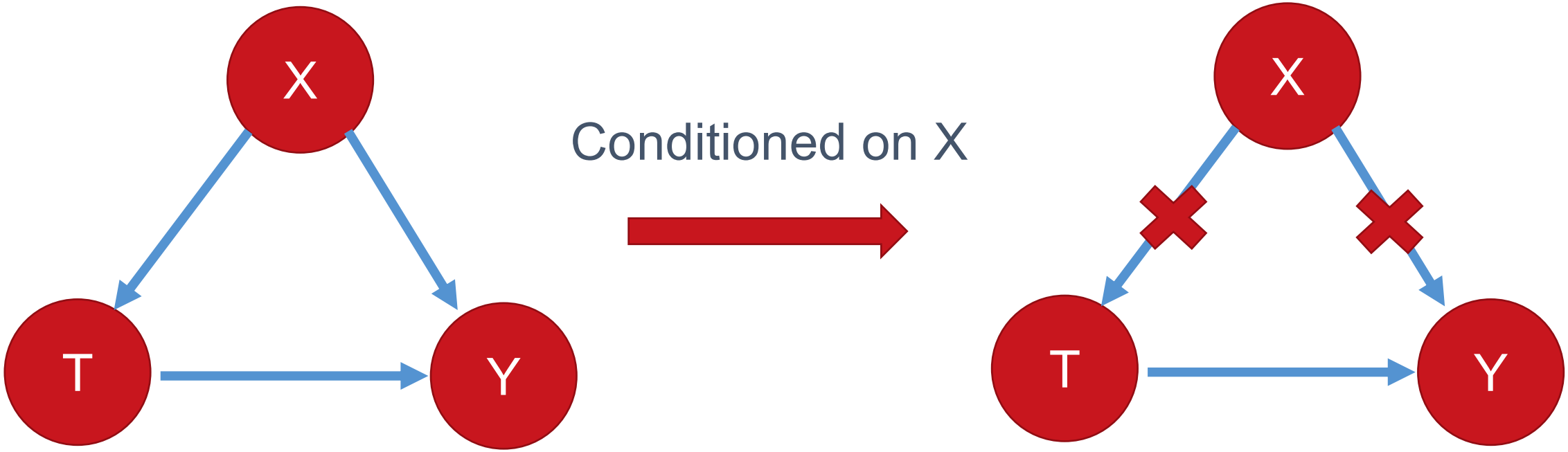
Assumption (Conditional Exchangeability)

$$\big(Y(1), Y(0)\big) \perp T|X$$

即假设所有变量都包含在 X 里。给定 X 了后满足Exchangeability条件。

这样，即使treatment和outcome可能相关, 在给定X后，它们变得无关，从而可以变成identifiable的。

Conditional Exchangeability



Assumption (Conditional Exchangeability)

$$\big(Y(1), Y(0)\big) \perp T|X$$

ATE计算 $$\mathrm{E}[(Y(1) - Y(0))|X] = \mathrm{E}[Y(1)|X] - \mathrm{E}[Y(0)|X]$$

$$= \mathrm{E}[Y(1)|T = 1, X] - \mathrm{E}[Y(0)|T = 0, X]$$

$$= \mathrm{E}[Y|T = 1, X] - \mathrm{E}[Y|T = 0, X]$$

Marginalization: $$\mathrm{E}[(Y(1) - Y(0))] = \mathrm{E}_X[\mathrm{E}[Y|T = 1, X] - \mathrm{E}[Y|T = 0, X]]$$

Theorem (Adjustment Formula) 给定Unconfoundedness, positivity, consistency, 和no interference假设， ATE可以用下面公式来identify：

$$\mathrm{E}[(Y(1) - Y(0))] = \mathrm{E}_X[\mathrm{E}[Y|T = 1, X] - \mathrm{E}[Y|T = 0, X]]$$

这里涉及到的假设是怎么来的呢?

Assumption (Positivity/Overlap/Common Support) 对所有的covariates x

只要 $P(X=x)>0$，就必须有 $0<P(T=1|X=x)<1$

$$\mathrm{E}[(Y(1)-Y(0))]=\mathrm{E}_X[\mathrm{E}[Y|T=1,X]-\mathrm{E}[Y|T=0,X]]$$

$$=\sum_x P(X=x)(\sum_y yP(Y=y|T=1,X=x)-\sum_y yP(Y=y|T=0,X=x))$$

$$=\sum_x P(X=x)(\sum_y y\frac{P(Y=y,T=1,X=x)}{P(T=1|X=x)P(X=x)}-\sum_y y\frac{P(Y=y,T=0,X=x)}{P(T=0|X=x)P(X=x)})$$

Assumption (No Interference)

$$Y_i(t_1, \dots, t_{i-1}, t_i, t_{i+1}, \dots, t_n) = Y_i(t_i)$$

Assumption (Consistency) 如果treatment是T，

$$T = t \quad \Rightarrow \quad Y = Y(t) \;\; or \; Equivalently \; Y = Y(T)$$

下面，我们基于这些假设，重新完整证明定理。

Theorem (Adjustment Formula) 给定Unconfoundedness, positivity, consistency, 和no interference假设， ATE可以用下面公式来identify：

$$\mathrm{E}[(Y(1)-Y(0))]=\mathrm{E}_X[\mathrm{E}[Y|T=1,X]-\mathrm{E}[Y|T=0,X]]$$

$$\mathrm{E}\big[\big(Y(1)-Y(0)\big)\big]=\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)]=E_X[E[Y(1)|X]-E[Y(0)|X]]$$

$$=E_X[E[Y(1)|T=1,X]-E[Y(0)|T=0,X]]$$

(Unconfoundedness and positivity assumption)

$$=E_X[E[Y|T=1,X]-E[Y|T=0,X]]$$

(Consistency assumption)

因果推理的一般步骤

综上所述，因果推理具有以下一般步骤：