浙江大学
ZHEJIANG UNIVERSITY

因果推理与大模型
理论、方法与应用

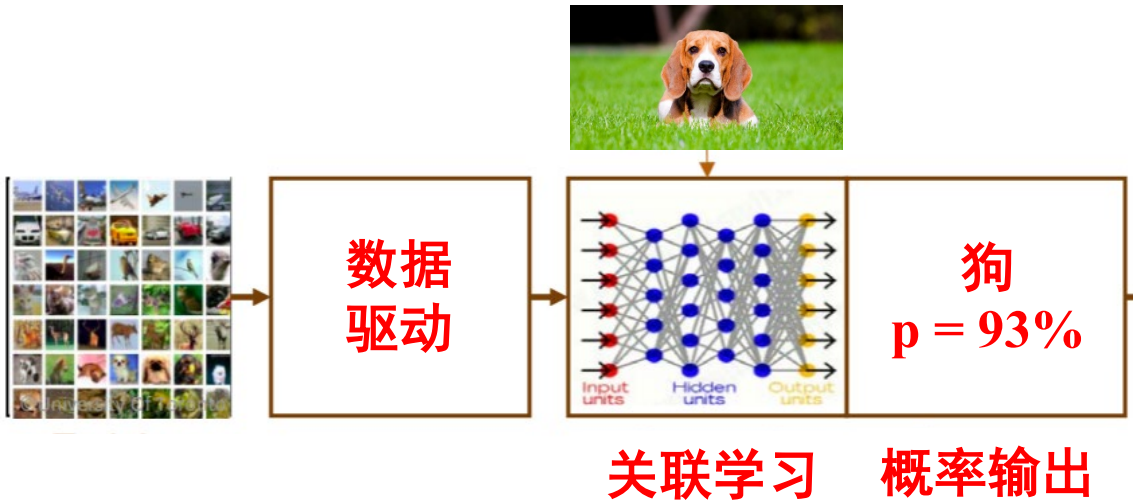
况琨（浙江大学）、吴飞（浙江大学）
张兴璇（清华大学）
王浩天（国防科技大学）

Tutorial安排

议程	嘉宾
因果与关联：从统计学习到因果范式	况琨
复杂环境下因果推断	
因果启发的稳定可泛化学习	
因果赋能大语言模型探索与思考	
茶歇	
因果赋能结构大数据模型：通用数据大模型引领结构化数据智能新范式	张兴璇
因果赋能物理模型与具身智能探索与思考	王浩天

人工智能学习特点

- 深度学习等人工智能学习特点



为什么图像会被识别为“狗”？为什么会用“草地”预测狗？为什么不同测试结果差异大？

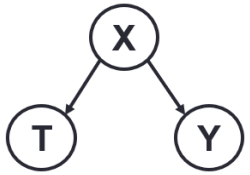
模型存在不可解释，不可泛化等问题

关联的三种来源

因果

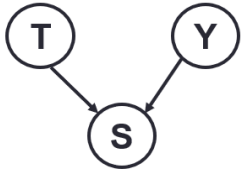
可解释
稳定/鲁棒
可决策

混淆偏差

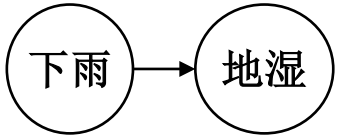


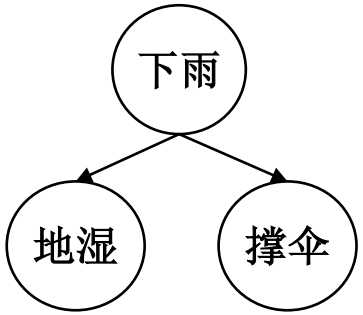
虚假关联: 当忽略 X 时，T 和 Y 相关

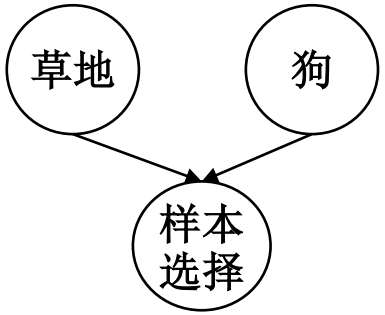
选择偏差



虚假关联: 当给定 S 时，T 和 Y 相关







大语言模型的学习特点也是数据驱动，关联学习

对齐（Alignment）

与数据空间对齐
与人类指令对齐
与人类反馈对齐

predict the next token

完形填空形式下文字接龙（自监督学习）

- 原话：一辆 列车 缓慢 行驶 在 崎岖 的 山路上
- 移除单词：一辆 列车 ___ 行驶 在 崎岖 的 山路上
- 预测填空：一辆 列车 缓慢 行驶 在 崎岖 的 山路上

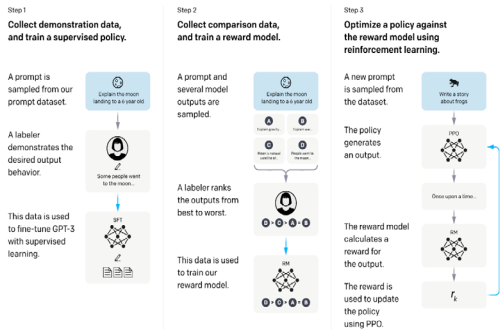
supervised fine-tuning

提示学习与指令微调（人教机学）

第一轮：{"instruction":本院查明,被告人酒后...请分析案情。
"output":根据上述内容，可以认定本案的核心要素包括醉酒驾驶、致人受伤、酒后逃逸...},
第二轮：{"instruction":根据上述分析，请预测罪名。
"output":本案预测的罪名是危险驾驶罪}
第三轮：{"instruction":请给出处罚意见。
"output":结合嫌疑人逃逸的情节，建议考虑拘役三个月，并罚款6000元}
...

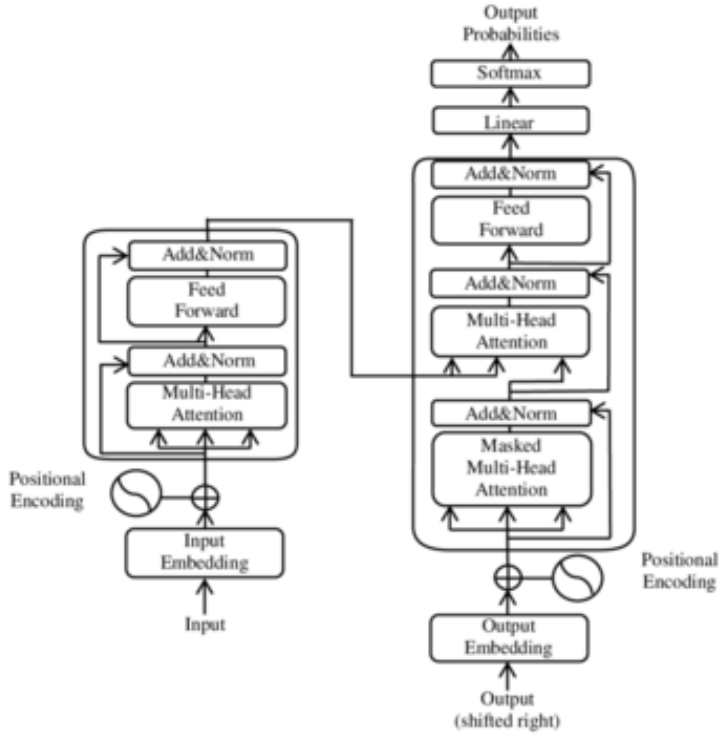
Reinforcement Learning from Human Feedback

人类反馈下强化学习（尝试与探索）



自然语言合成的核心神经网络是Transformer 模型

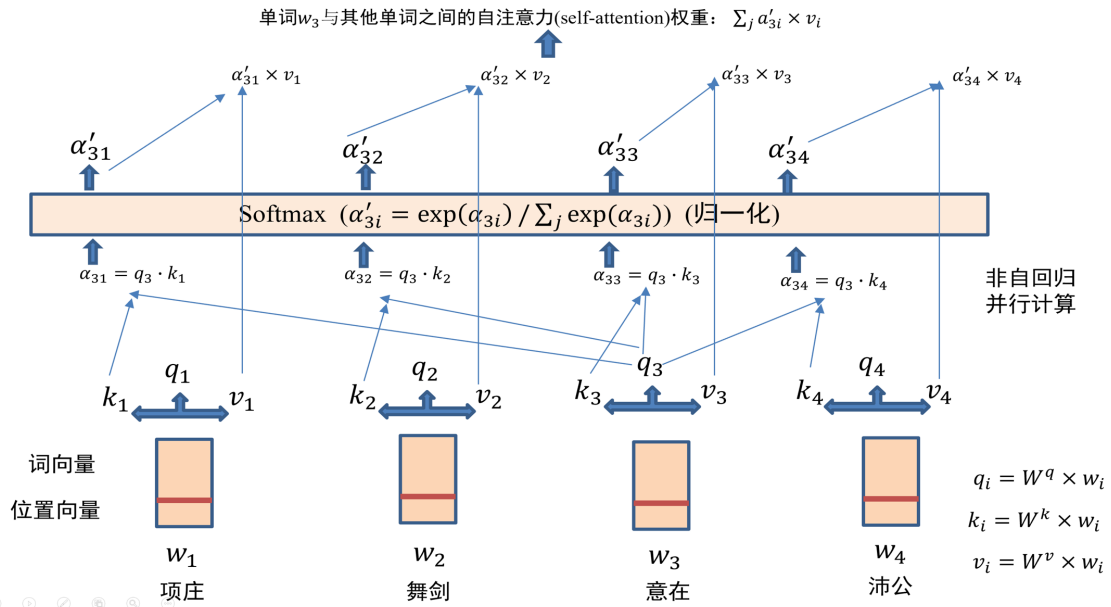
大语言模型的学习特点也是数据驱动，关联学习



消除反馈(recurrent)机制

Google (2017): Attention is all you need

attention：单词共生概率



项庄 舞剑 意在 沛公

学习单词和单词之间**关联关系** （in-context meaning）

大语言模型的学习特点也是数据驱动，关联学习

- 提示词：东京赏樱
- Sora可以认为是背诵、默写与组合出来“现实世界”

Sora “超级涌现力”将把AI引向何方 （文汇报：2024年2月24日）

从Chat到Sora: 对合成内容中最小单元进行有意义的关联组合，犹如昨日重现


文匯報

《习近平谈治国理政》第四卷
5种民族文字版出版发行

习近平主持召开中央财经委员会第四次会议强调
推动新一轮大规模设备更新和消费品以旧换新
有效降低全社会物流成本增强产业核心竞争力

为上海现代化建设提供坚强组织保障

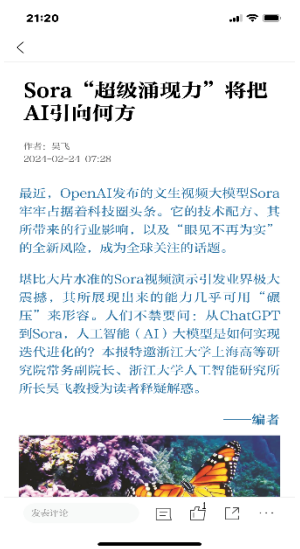
把群众诉求解决好情绪疏导好力量凝聚好

心系民生算大账 盘活农村边角料

申城龙腾狮跃闹元宵

沪郊村镇欢度佳节

Sora“超级涌现力”将把AI引向何方


21:20

Sora“超级涌现力”将把AI引向何方

作者：吴飞
2024-02-24 07:28

最近，OpenAI发布的文生视频大模型Sora牢牢占据着科技圈头条。它的技术配方、其所带来的行业影响，以及“眼见不再为实”的全新风险，成为全球关注的话题。

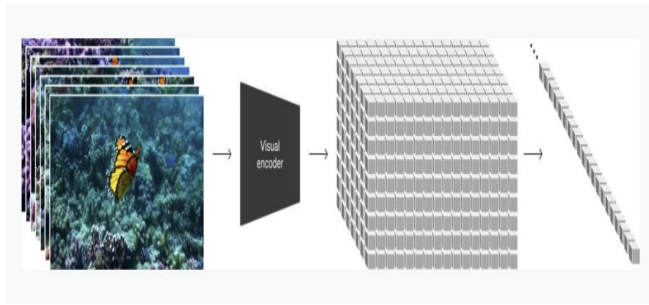
堪比大片水准的Sora视频演示引发业界极大震撼，其所展现出来的能力几乎可用“碾压”来形容。人们不禁要问：从ChatGPT到Sora，人工智能（AI）大模型是如何实现迭代进化的？本报特邀浙江大学上海高等研究院常务副院长、浙江大学人工智能研究所所长吴飞教授为读者释疑解惑。

——编者

发表评论

单词有意义的线性组合

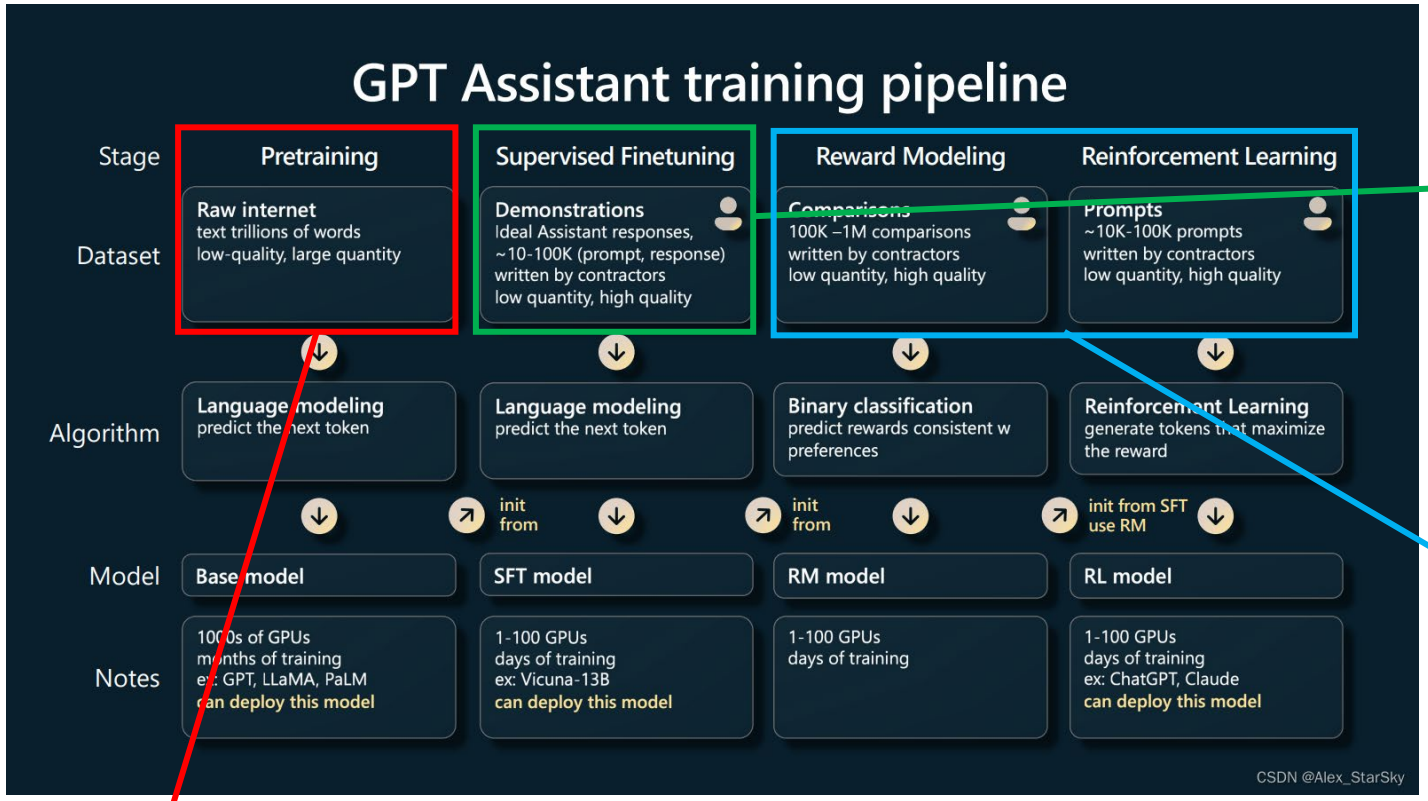
像素点有意义的空间组合



时空子块有意义的时序组合

在保持连贯的上下文语境中，对若干个单词进行有意义线性组合，从而连缀成一个会意句子；在保持合理的空间布局下，对众多图像小块进行有意义结构组合，拼合为一幅精彩图像；在保持一致的连续时空内，对一系列时空子块进行有意义时空组合，从而拼接成一段动感视频。

因果如何赋能大语言模型



- 因果去除数据偏置
  - 虚假相关问题
  - 灾难遗忘问题
- 因果支撑决策
  - 用户偏好对齐
  - 因果强化学习
- 因果赋能Transformer
  - 由关联自回归到因果回归机制
  - 因果Transformer架构
  - 基于因果知识增强的Transformer架构

Survey： Causality for LLMs

Causality for Large Language Models

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, Kun Zhang

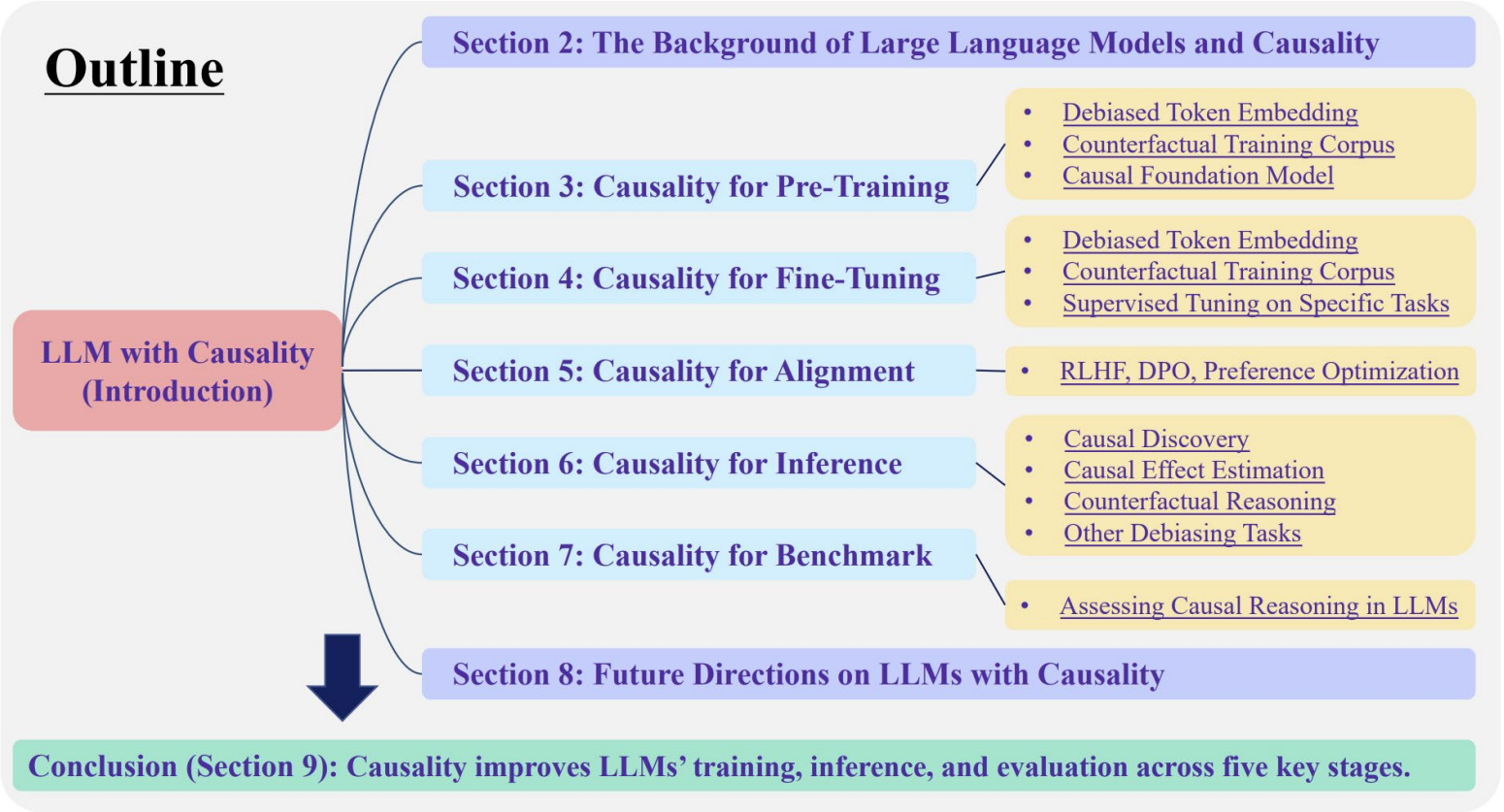
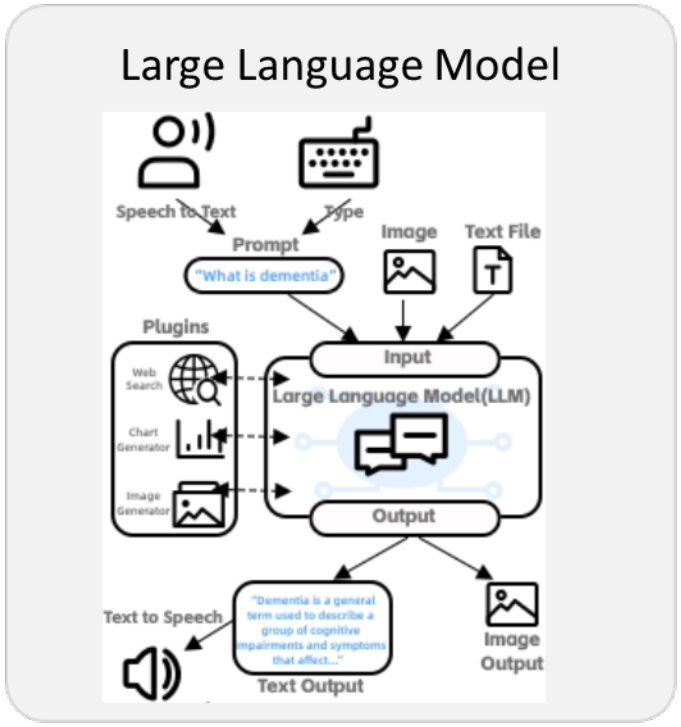
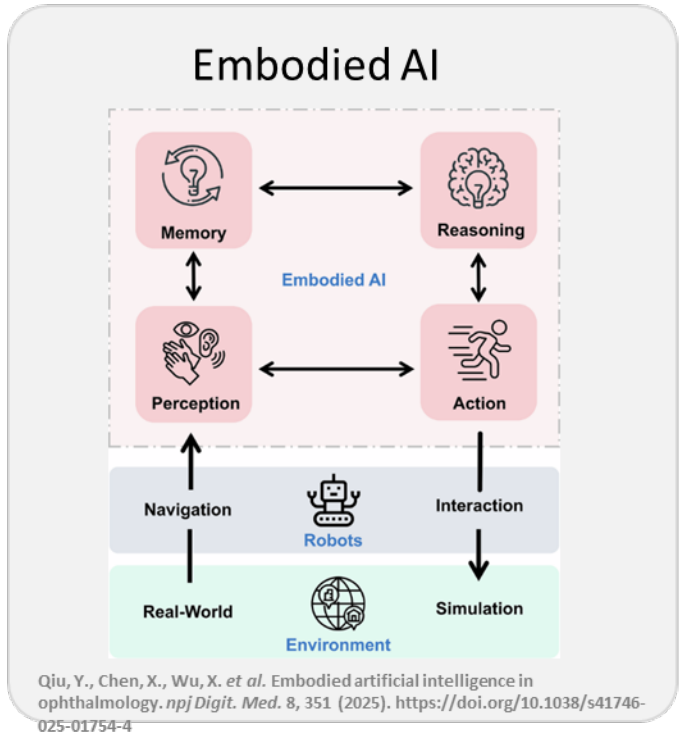


Figure 1: The Role of Causality in Enhancing LLMs: A Comprehensive Framework Across Development Stages

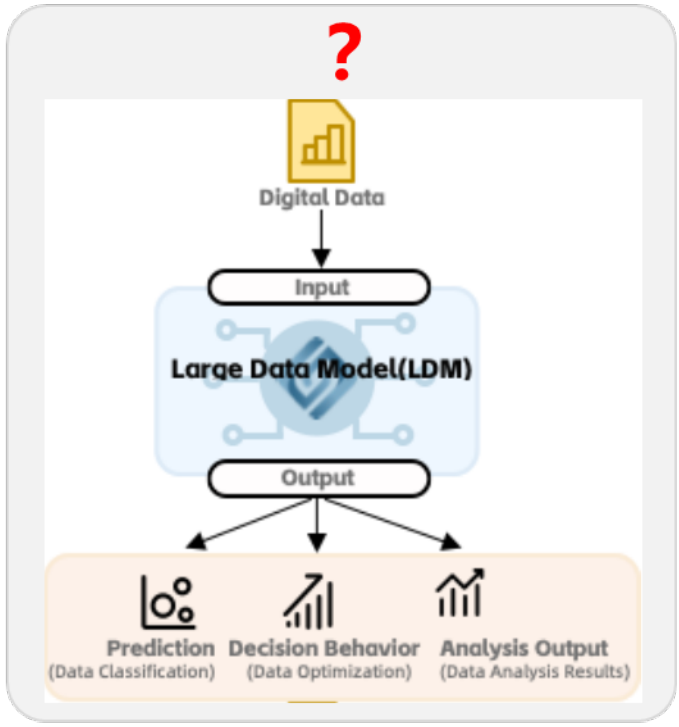
https://arxiv.org/abs/2410.15319

AGI需要三个通用世界模型





Qiu, Y., Chen, X., Wu, X. *et al.* Embodied artificial intelligence in ophthalmology. *npj Digit. Med.* 8, 351 (2025). https://doi.org/10.1038/s41746-025-01754-4



语义空间　　物理空间　　数据空间

Artificial general intelligence (AGI)—sometimes called **human-level intelligence AI**—is a type of artificial intelligence that would match or surpass human capabilities across virtually all cognitive tasks. <**Wikipedia**>

Tutorial安排

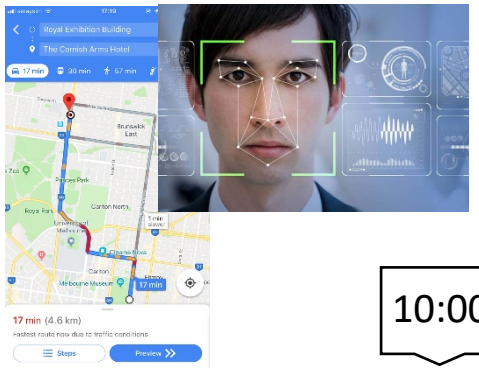
议程	嘉宾
因果与关联：从统计学习到因果范式	况琨
复杂环境下因果推断	
因果启发的稳定可泛化学习	
因果赋能大语言模型探索与思考	
茶歇	
因果赋能结构大数据模型：通用数据大模型引领结构化数据智能新范式	张兴璇
因果赋能物理模型与具身智能探索与思考	王浩天

浙江大学
ZHEJIANG UNIVERSITY

因果与关联：从统计学习到因果范式

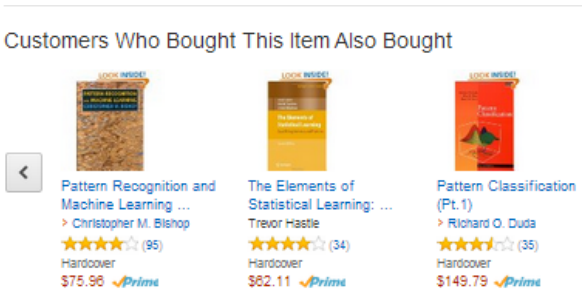
况琨
浙江大学计算机学院

人工智能算法在我们生活中无处不在



8:00 am

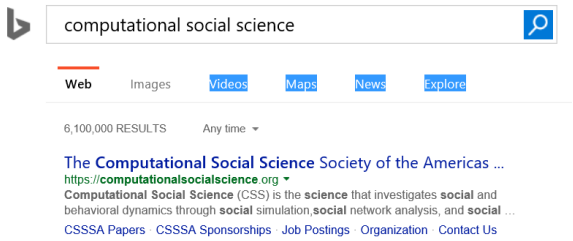
10:00 am



6:00 pm

8:00 pm

8:30 am

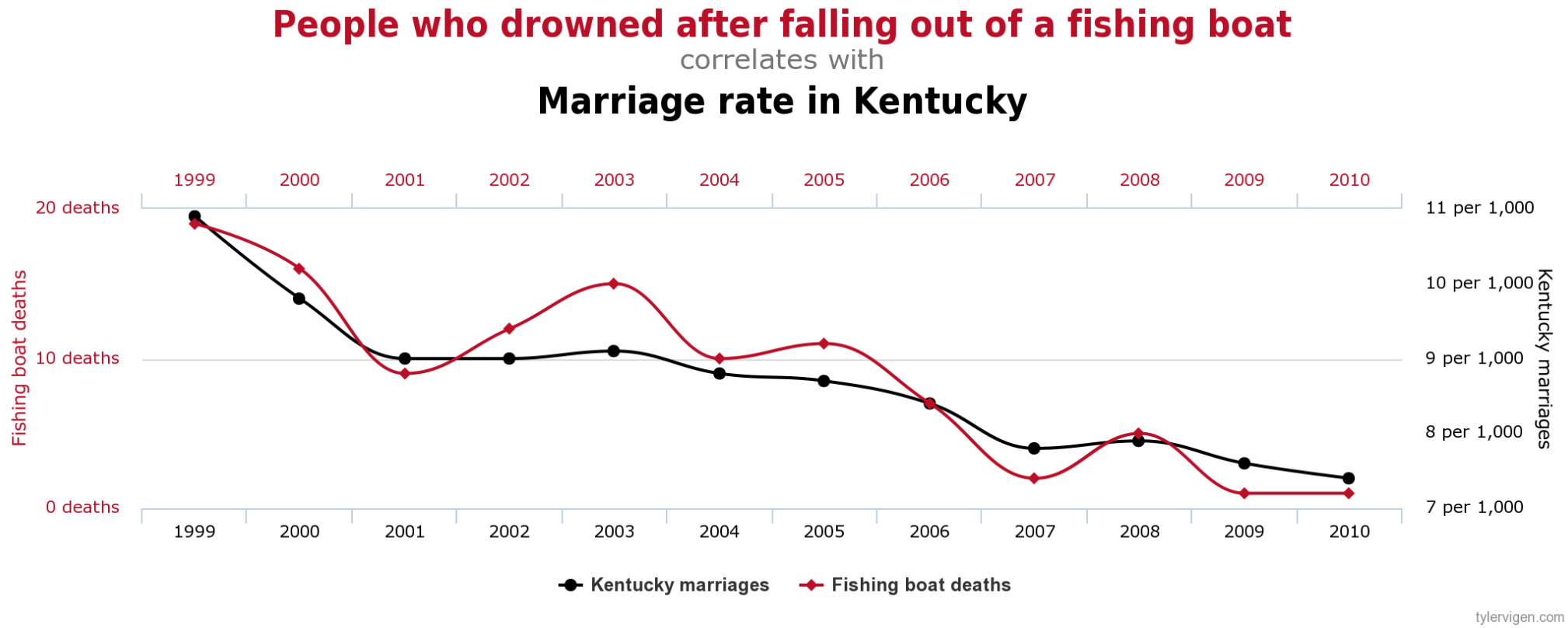


4:00 pm

人工智能学习特点：数据驱动，关联学习

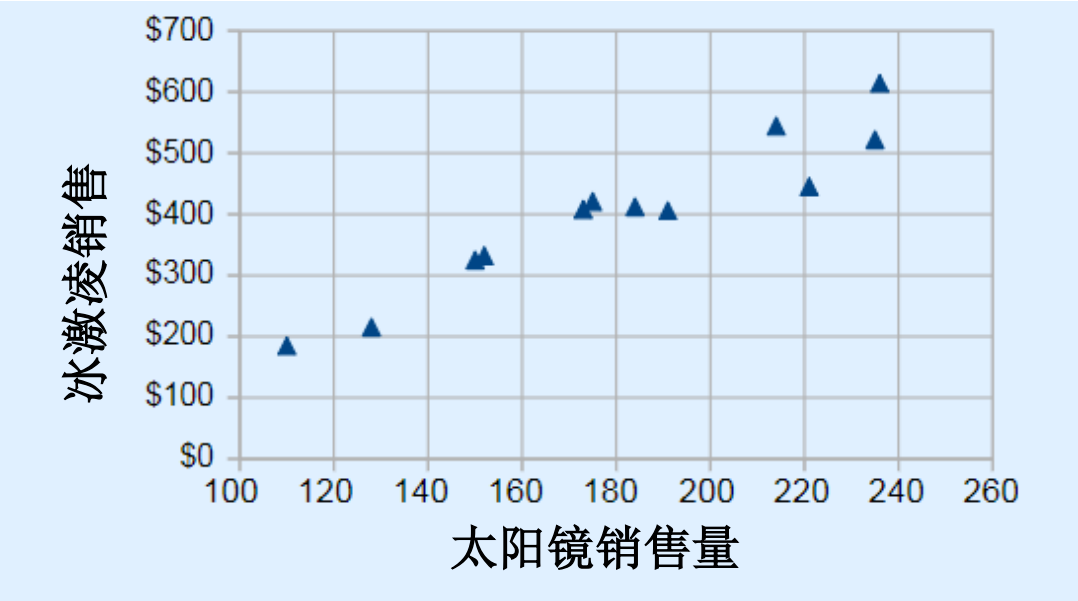
为什么需要从数据关联到因果推理

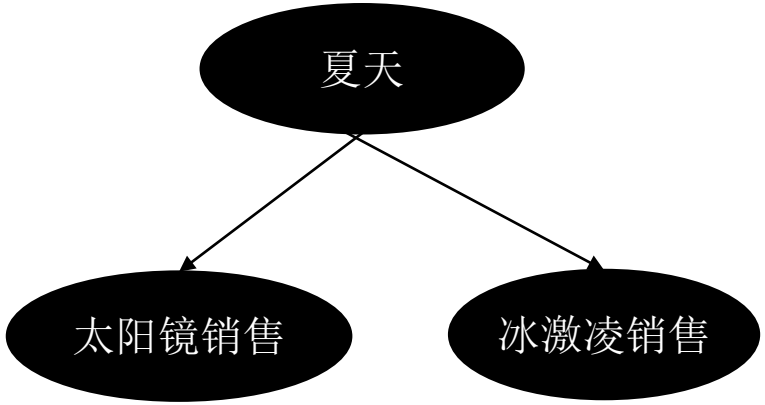
- 关联的局限性1：不可解释



为什么需要从数据关联到因果推理

- 关联的局限性1：不可解释

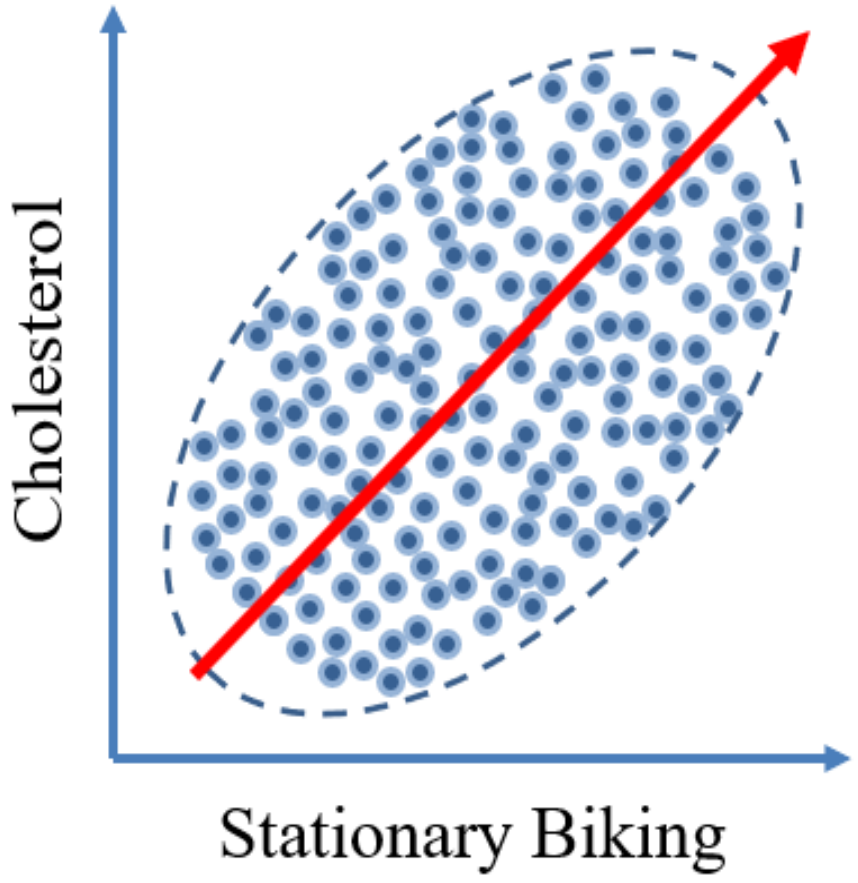


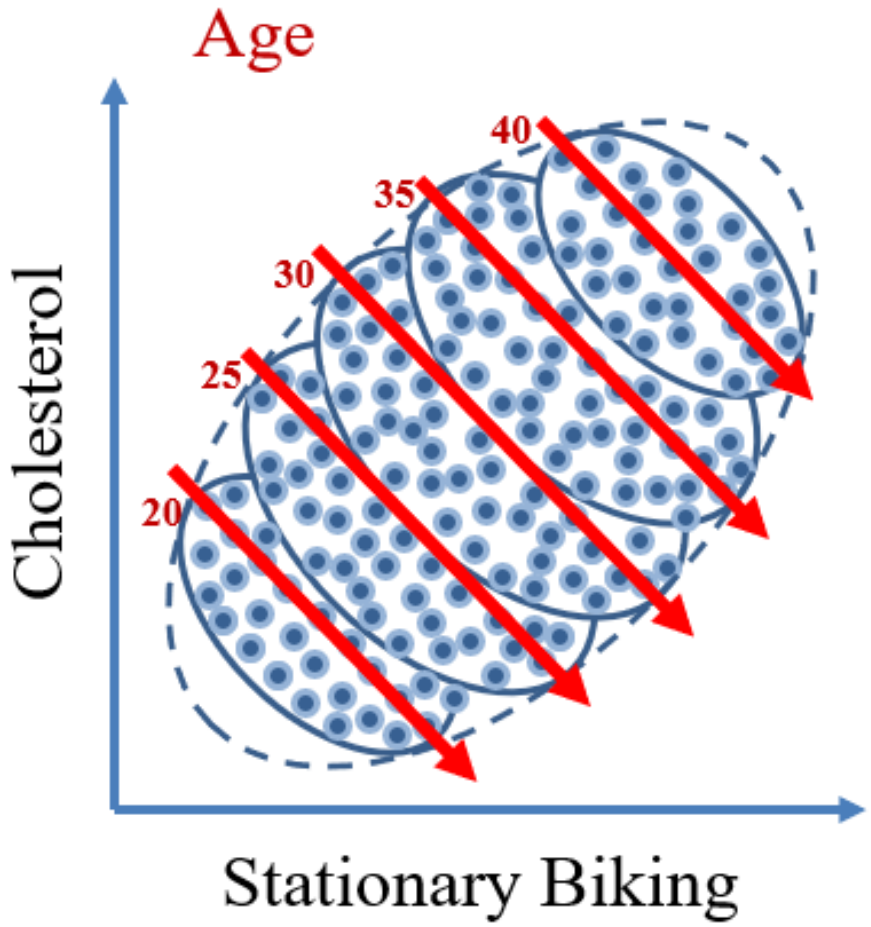


虚假相关

为什么需要从数据关联到因果推理

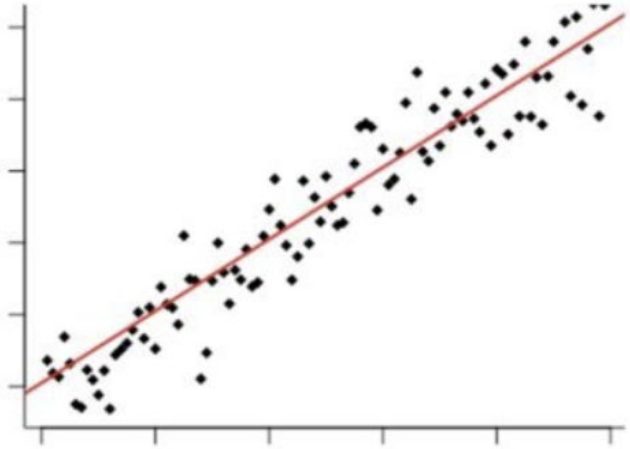
- 关联的局限性2：不可用于支撑决策





为什么需要从数据关联到因果推理

- 关联的局限性2：不可用于支撑决策



- 小孩子的阅读能力与鞋尺寸有强的正相关。
- 根据小孩鞋尺寸能预测他的阅读能力！
- 但是人为地改变鞋的尺寸，并不会提高他们的阅读能力。

为什么需要从数据关联到因果推理

- 关联的局限性2：不可用于支撑决策
- 预测模型能否指导我们决策？
- 举例：推荐算法A和B，推荐打折链接给用户
- 假设推荐系统需要更换算法，是否要将原来算法A调整到新算法B
- 是否新算法B的效果会更好一些？

算法A

算法B

为什么需要从数据关联到因果推理

- 关联的局限性2：不可用于支撑决策
- 测量两个算法的成功率

算法A	算法B
50/1000 **(5%)**	54/1000 **(5.4%)**

新算法B提升了推荐成功率，那么算法B就一定比算法A要好么？

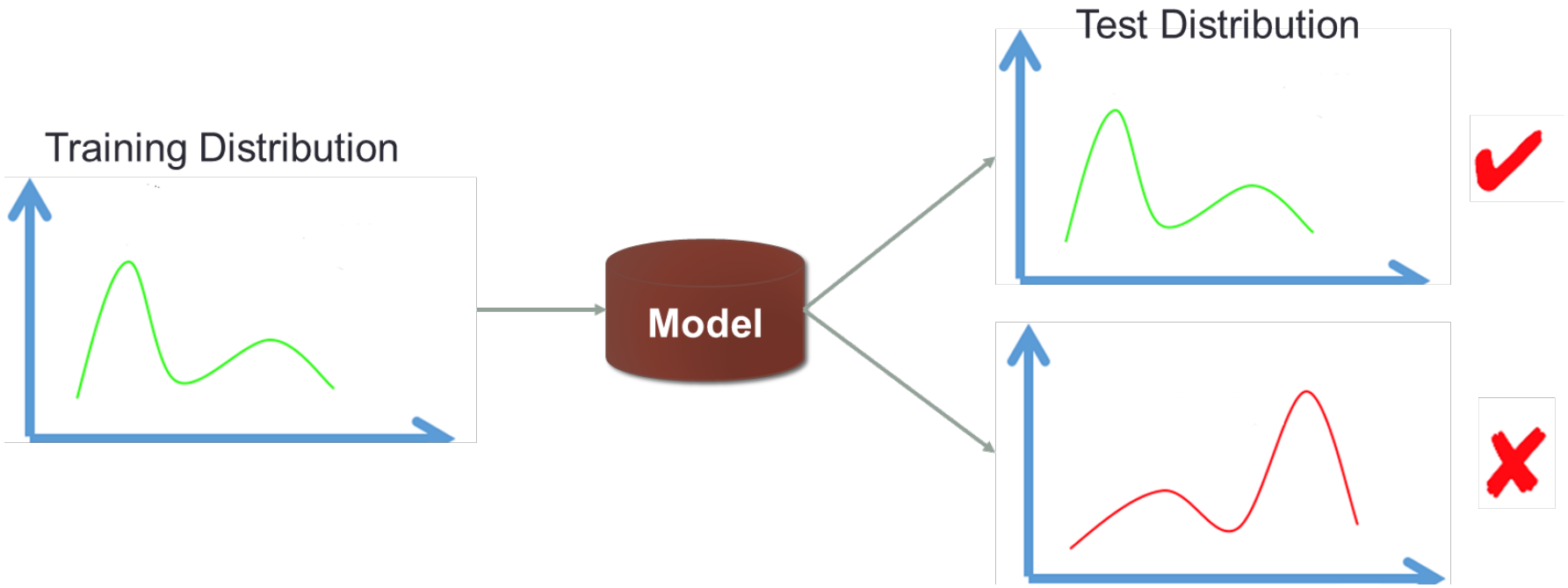
	算法A	算法B
低收入用户	10/400 **(2.5%)**	4/200 **(2%)**
高收入用户	40/600 **(6.6%)**	50/800 **(6.2%)**
整体	50/1000 **(5%)**	54/1000 **(5.4%)**

到底哪个算法更好?

为什么需要从数据关联到因果推理

- 关联的局限性3：不稳定，会随着时间、数据、环境等变化而变化

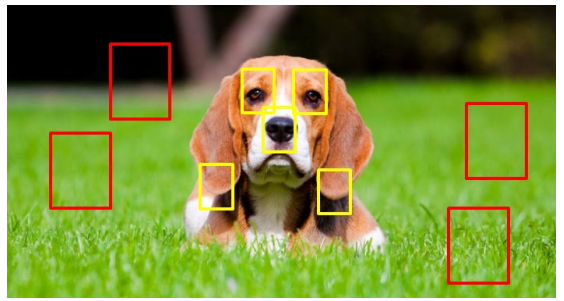
绝大多数机器学习方法需要独立同分布假设

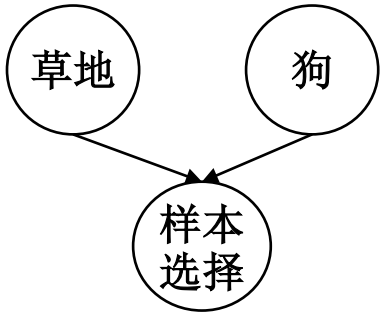


为什么需要从数据关联到因果推理

- 关联的局限性3：不稳定，会随着时间、数据、环境等变化而变化

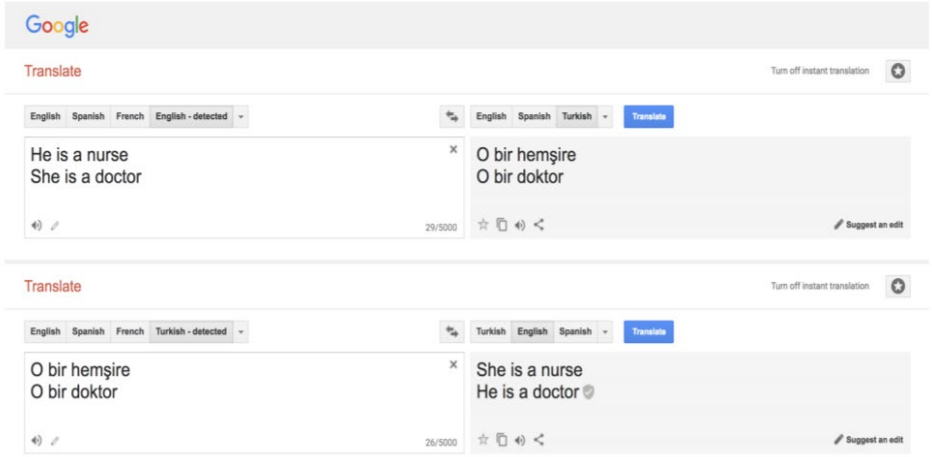
绝大多数机器学习方法需要独立同分布假设





为什么需要从数据关联到因果推理

- 关联的局限性4：数据关联可能会带来不公平性问题



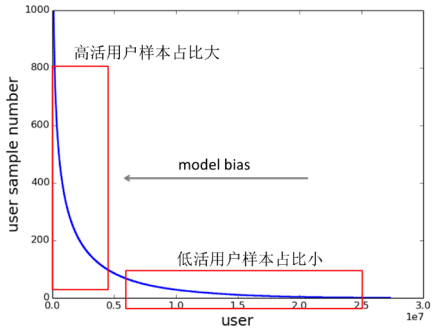
自然语言翻译

The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

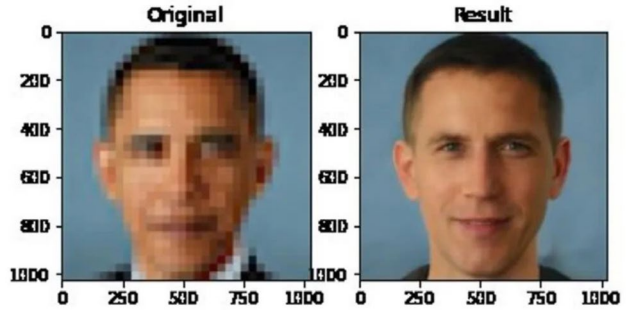
Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

人脸识别



Micro AUC提升，表面收益增加，但可能Macro AUC下降，用户平均满意度下降（模型牺牲低活用户体验，换取更多收益）。

推荐和预测

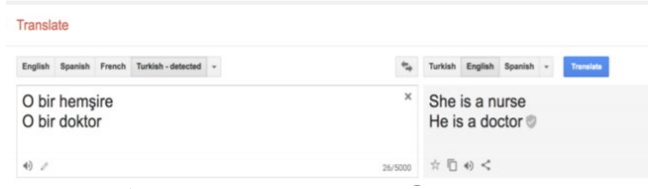


计算机视觉和检索

为什么需要从数据关联到因果推理

- 关联的局限性4：数据关联可能会带来不公平性问题

偏序偏差
流行度偏差
…



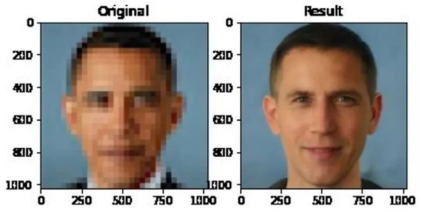
算法

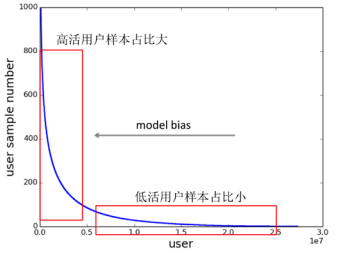
用户

数据

样本选择偏差
遗漏变量偏差
…

用户行为偏差
内容生产偏差
…





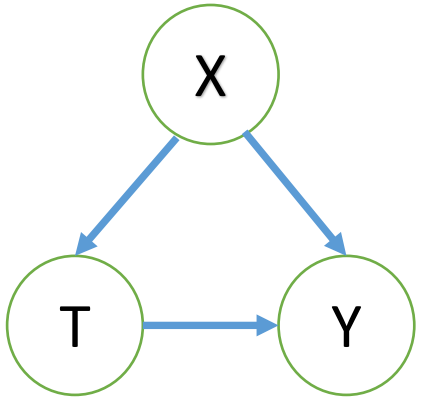
Micro AUC提升，表面收益增加，但可能Macro AUC下降，用户平均满意度下降（模型牺牲低活用户体验，换取更多收益）。

为什么需要从数据关联到因果推理

- 关联的局限性4：数据关联可能会带来不公平性问题

关联分析框架

因果推理框架



T：肤色
X：收入
Y：犯罪率

收入—犯罪率: 强相关

肤色—犯罪率: 强相关

收入—犯罪率: 强因果

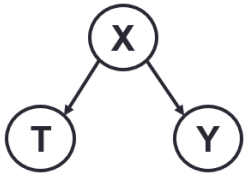
肤色—犯罪率: 弱因果

关联学习导致人工智能的不能

因果

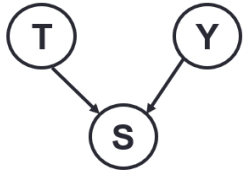
可解释
稳定/鲁棒
可决策

混淆偏差



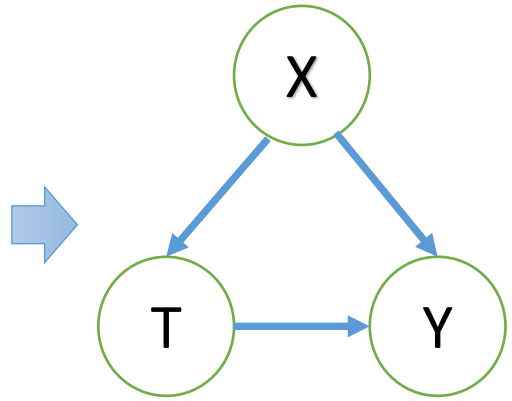
虚假关联: 当忽略 X 时，T 和 Y 相关

选择偏差



虚假关联: 当给定 S 时，T 和 Y 相关

关联分析框架



因果推理框架

- 关联不可解释，因果提升模型可解释性
- 关联不可决策，因果助力模型决策能力
- 虚假关联不稳定，因果关联具有不变性
- 虚假关联不公平，因果关联确保公平性

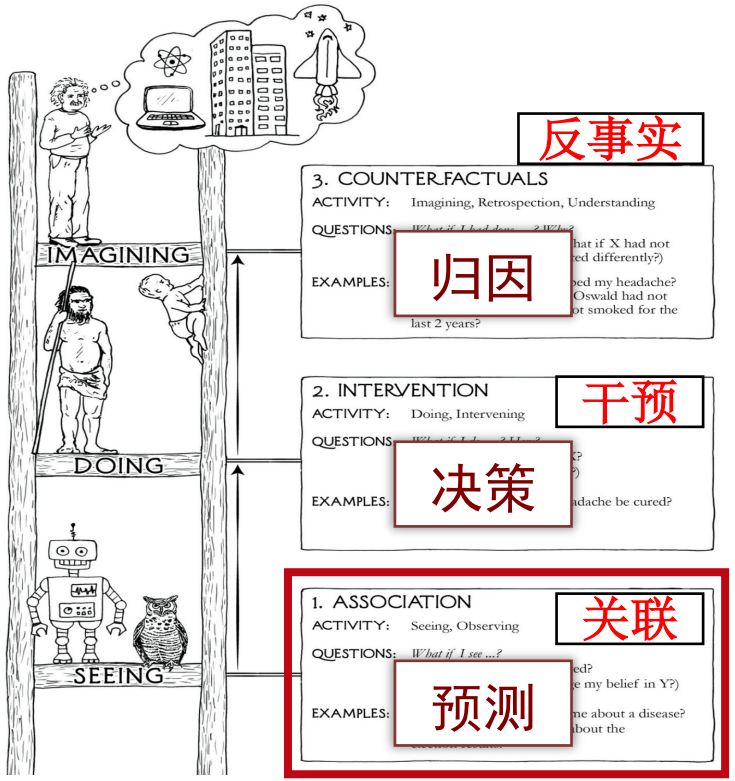
因果让不能变成可能，实现因果可信人工智能

问题的主要根源：因果机制尚未融入机器学习

- 问题主要根源：因果机制尚未融入机器学习

Judea Pearl

2011年图灵奖获得者
提出因果关系的
三个层级



公鸡打鸣是太阳升起的原因吗？
张三没打疫苗得病了；
假如当初打疫苗，是否还会得病？

如果不让公鸡打鸣，太阳还会升起吗？
如果打疫苗，疫情会减轻吗？

公鸡打鸣与太阳升起
打疫苗越多的地方或时期，疫情越重

当今人工智能处于最低层级：关联

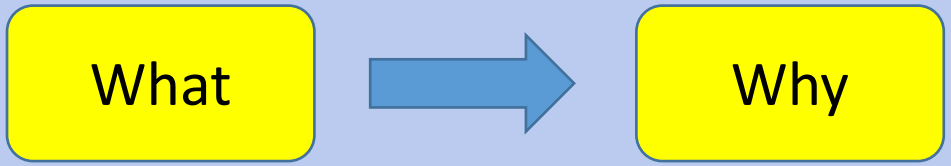
无论数据多大或神经网络多深，无法回答“干预”和反事实问题

将因果引入机器学习，提升模型稳定性、可解释性和决策能力

由关联到因果跨越

- 由关联学习到因果启发

人工智能需要从关联分析跨越到因果推理
“知其然，并知其所以然”



小结

- 人工智能的学习特点：数据驱动、关联学习、概率输出
- 关联有三种来源：因果关系、混淆偏差和选择偏差，后两者产生的关联称之为虚假关联
- 人工智能方法在关联学习过程中未能区分因果关联和虚假关联，会导致不可解释、不可决策、不稳定等不能

甄别因果关联，由关联到因果的跨越，
实现“知其然，并知其所以然”的人工智能

浙江大学
ZHEJIANG UNIVERSITY

复杂环境下因果推断

况琨
浙江大学计算机学院

什么是因？什么是果？

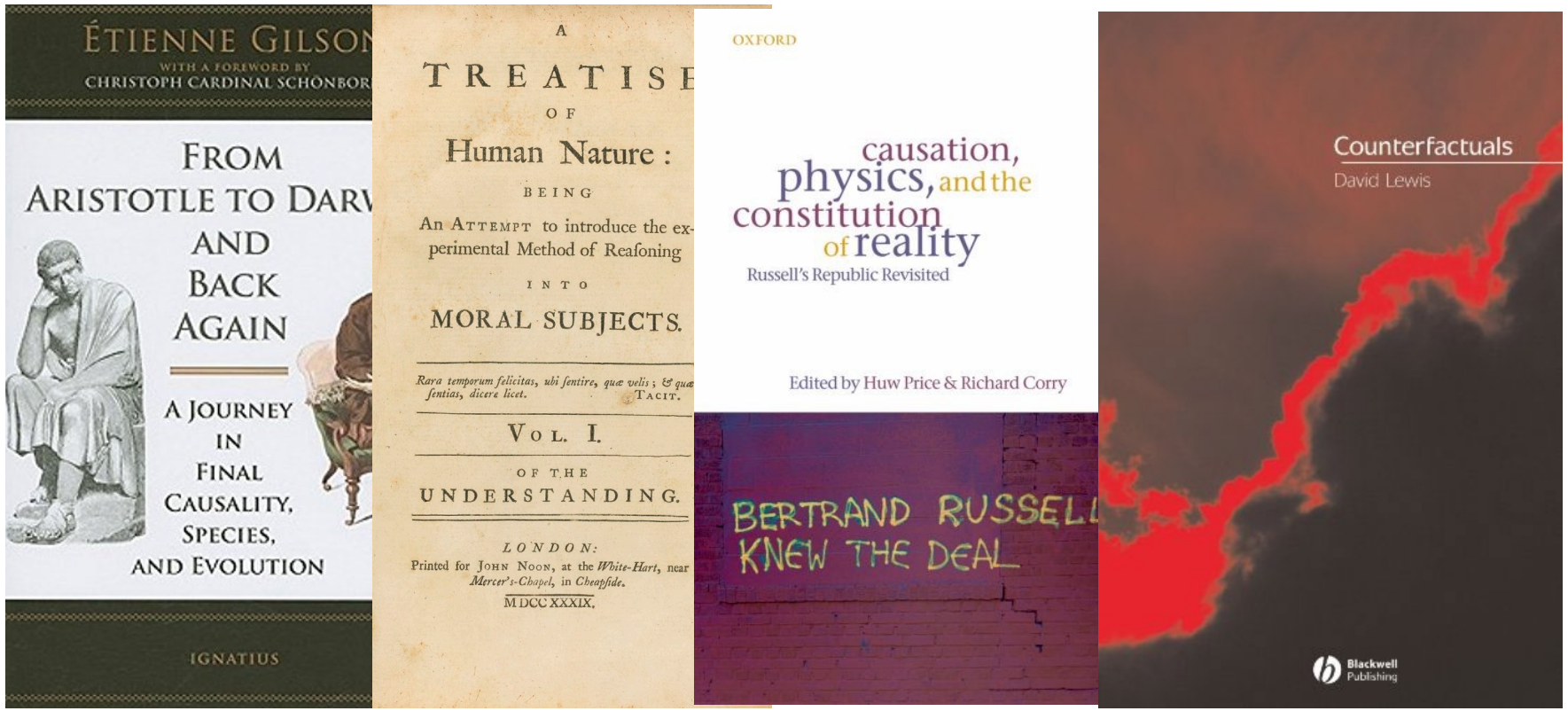
- 哲学上把现象和现象之间那种“引起和被引起”的关系，叫做因果关系，其中引起某种现象产生的现象叫做原因，被某种现象引起的现象叫做结果。

- 学科中因和果的问题：
  - 医学：新药的因果效应（疫苗是否能控制疫情）
  - 社会科学：政策的因果效应（禁烟的效应）
  - 广告学：营销策略的因果效应（投放广告的效果）
  - …

- **什么是因果？**

什么是因？什么是果？

- A big scholarly debate, from Aristotle to Russell

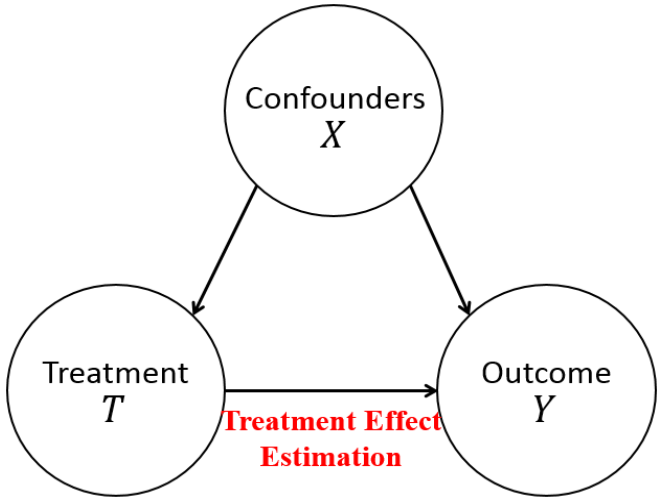


什么是因果？一个实用的定义

定义： T 导致了 Y 当且仅当

保持其他所有变量不变的情况下，

改变 T 发现 Y 也发生了变化。



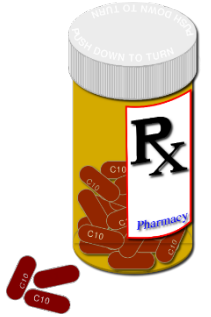
因果效应： T在改变一个单位时，Y的变化量。

两个关键条件: 保持其他所有变量不变，改变T

Interventionist definition [http://plato.stanford.edu/entries/causation-mani/]

因果效应估计

- 干预变量： $T = 1$ 或 $T = 0$
- 潜在结果变量： $Y(T = 1)$ 和 $Y(T = 0)$
- 个体因果效应（Individual Treatment Effect, ITE）：



$$ITE(i) = Y_i(T_i = 1) - Y_i(T_i = 0)$$

- 平均因果效应（Average Treatment Effect, ATE）：

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

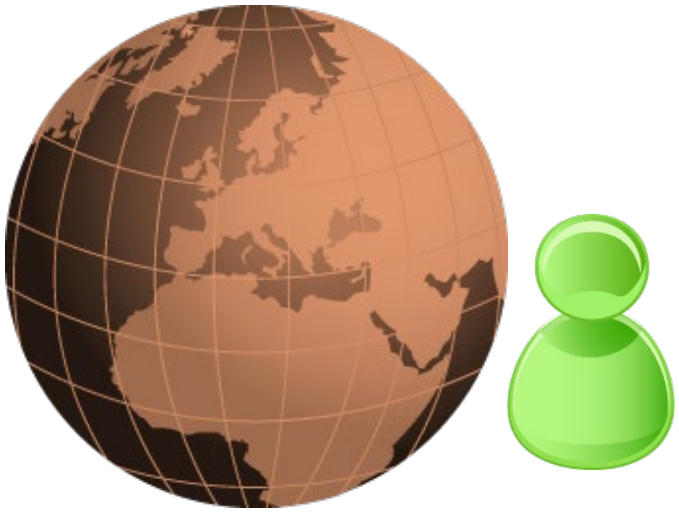
- 反事实问题：$Y(T = 1)$ 或 $Y(T = 0)$

两个关键条件: 保持其他所有变量不变，改变T

理想方案：存在平行世界

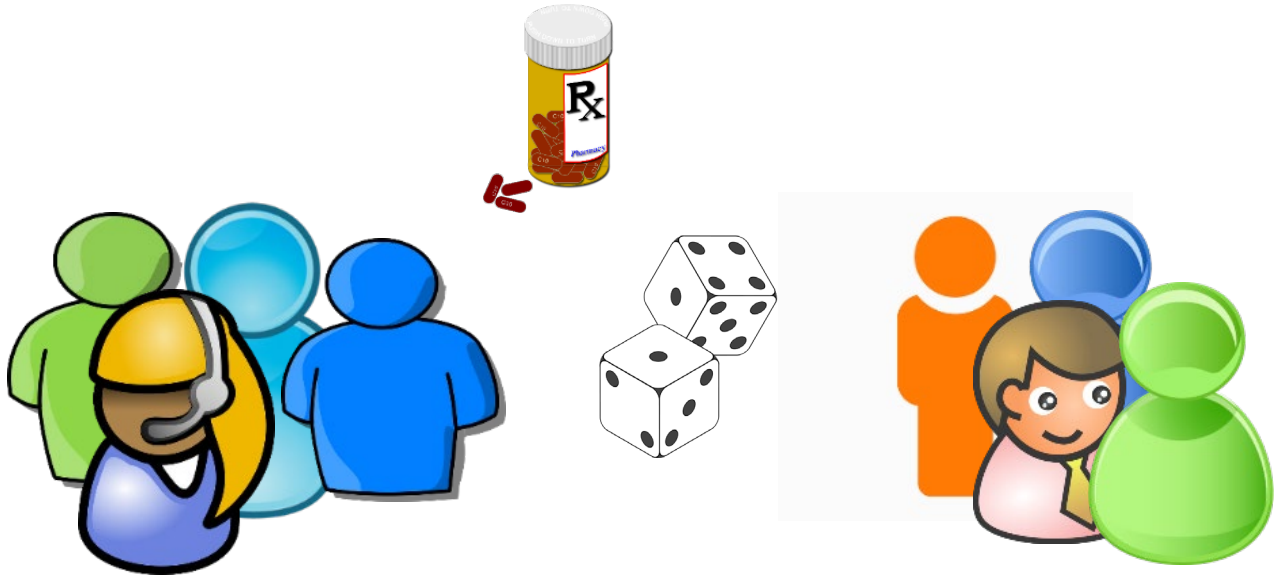
- 假设存在平行世界（真实世界和平行世界）
- 在真实世界和平行世界上，所有的变量都是一样的，但干预变量 T 不一样

$Y(T=1)$



$Y(T=0)$

金标准---随机实验



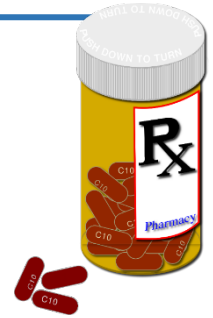
- 随机实验的缺点：
  - 耗钱，耗时，耗力
  - 很多时候可能存在伦理问题

金标准---随机实验

如果没有随机试验数据，我们该如何评估因果效应？
历史数据/观测学习！

- 随机实验
  - 耗钱，耗
  - 很多时候
    理问题

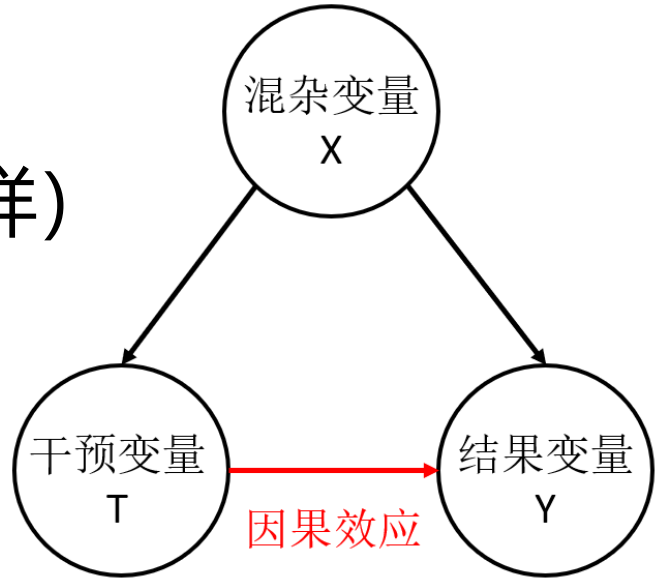
基于观测数据进行因果推断



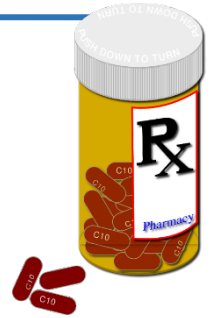
- 因果推断中的反事实问题：

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

- 能否通过计算T=1和T=0两个群体的均值，并相减来计算平均因果效应ATE？
  - 能，如果数据来自于随机实验（X是一样的）
  - 不能，如果是观测数据/历史数据（X可能不一样）
- 两个关键条件：
  - 保持其他所有变量不变
  - 改变T



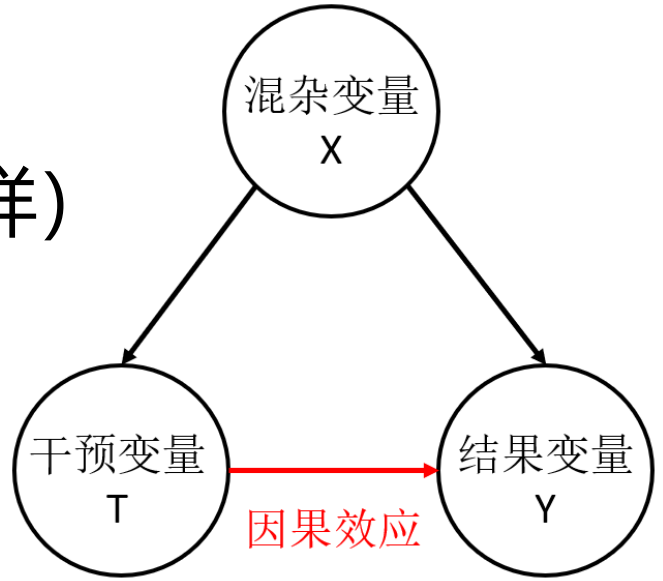
基于观测数据进行因果推断



- 因果推断中的反事实问题：

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

- 能否通过计算T=1和T=0两个群体的均值，并相减来计算平均因果效应ATE？
  - 能，如果数据来自于随机实验（X是一样的）
  - 不能，如果是观测数据/历史数据（X可能不一样）
- 两个关键条件：
  - 保持其他所有变量不变

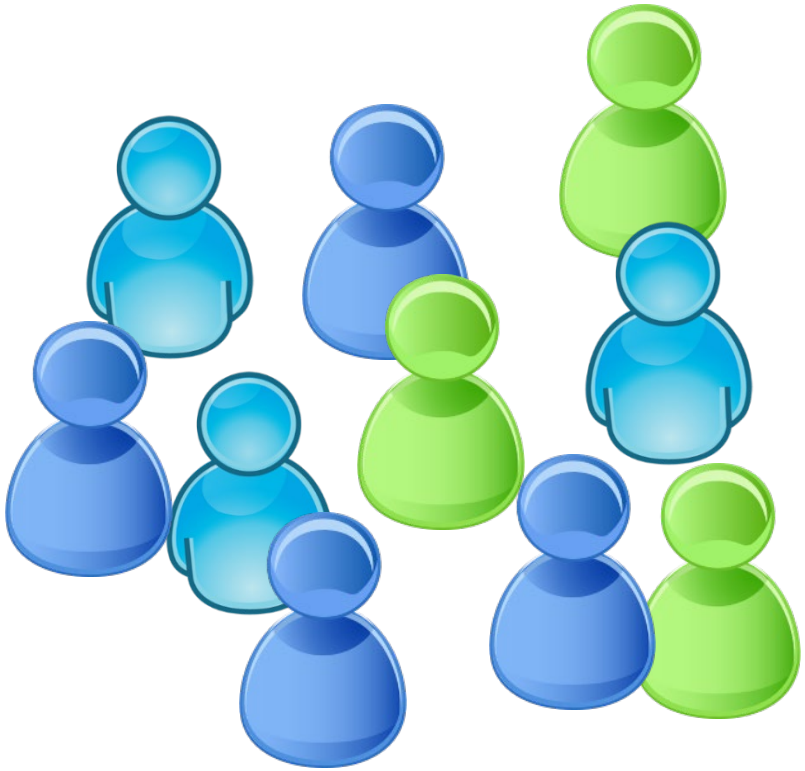


平衡混杂变量的分布（T=1 / T=0）

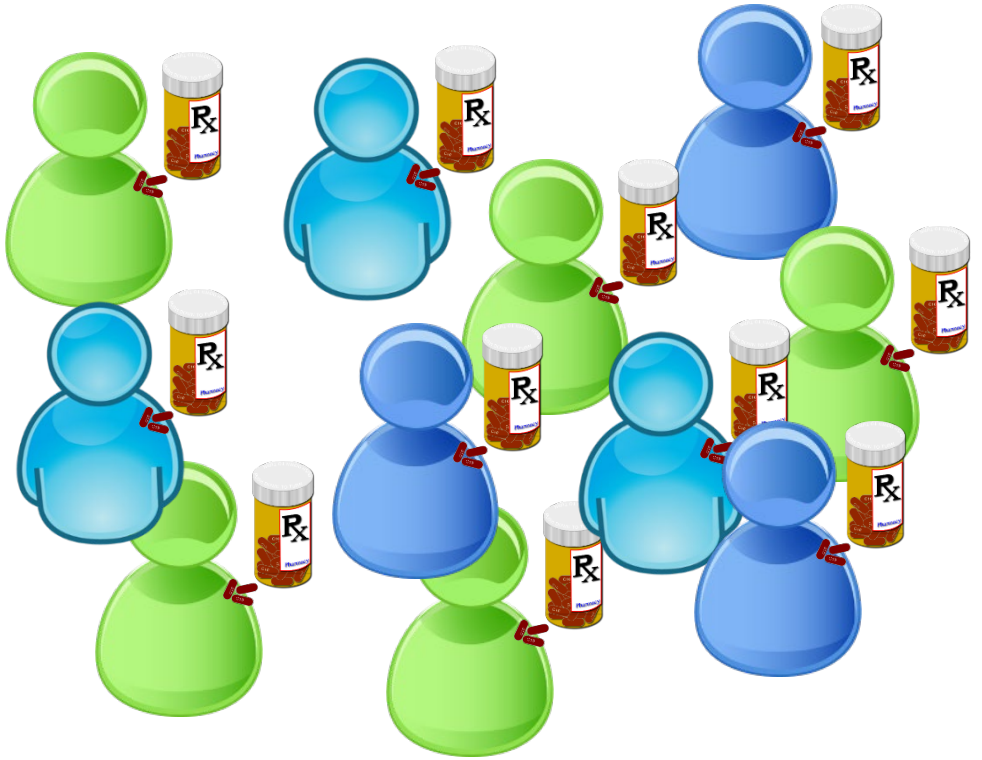
因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

匹配法

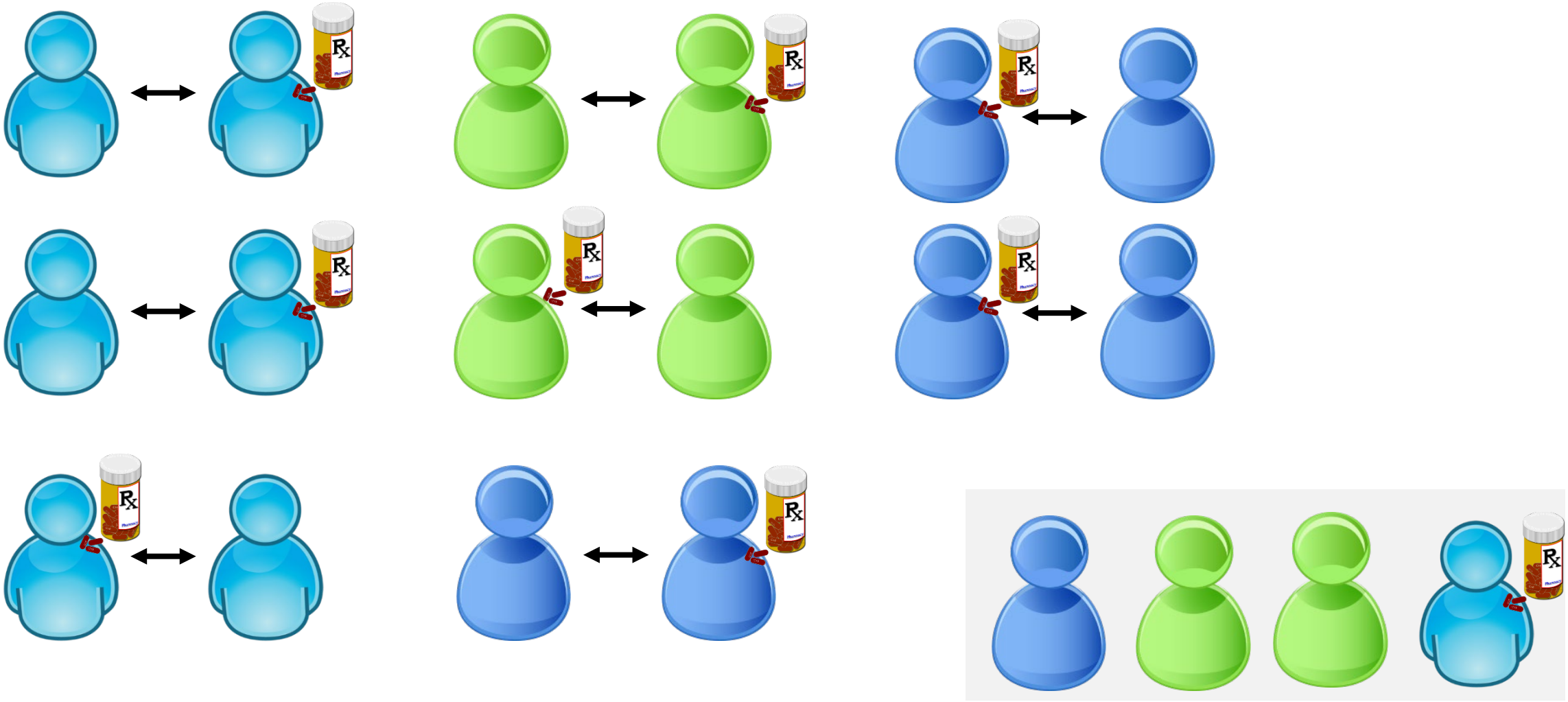


$T = 0$



$T = 1$

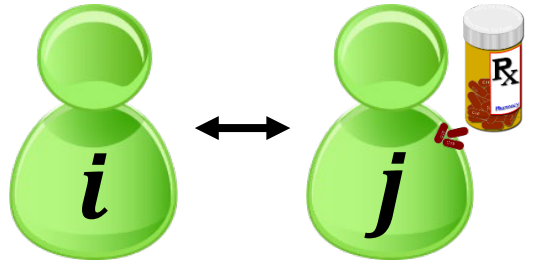
匹配法



匹配法

- 通过约束混杂变量X相似或一样，来匹配T=1和T=0的样本：

$$Distance(X_i, X_j) \leq \epsilon$$

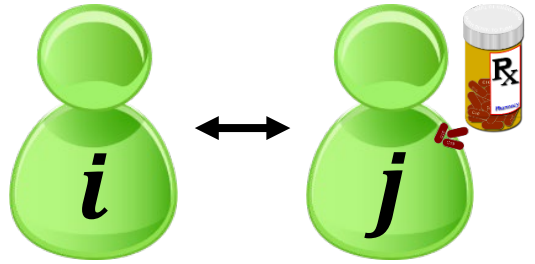


- 匹配的样本对能够保证T=1和T=0两组样本的混杂变量近似或一样。
- 平均因果效应可以通过计算T=1和T=0两组样本结果变量平均值的差值来计算。
- 参数ϵ越小，因果评估的偏差越小，但是方差大。

匹配法

- 精准匹配法：

$$Distance(X_i, X_j) = \begin{cases} 0, & X_i = X_j \\ \infty, & X_i \neq X_j \end{cases}$$



- 匹配法可用于混杂变量维度较低的场景

$$Distance(X_i, X_j) \leq \epsilon$$

- 但是当混杂变量是高维时，我们无法找到混杂变量完全一样的样本，匹配法就失效了

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

基于倾向得分的方法

- 倾向得分 $e(X)$ 定义为样本接受干预（T=1）的概率：

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

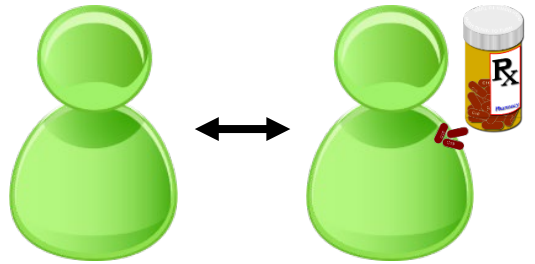
- 理论上， Donald Rubin证明了倾向得分足以总结混杂变量的信息，用于保证因果推断的无偏性。

$$T \perp\!\!\!\perp X \mid e(X) \quad \Rightarrow \quad T \perp\!\!\!\perp (Y(1), Y(0)) \mid e(X)$$

- 但是，倾向得分无法被观测到，需要进行评估

倾向得分匹配法

- 评估倾向得分： $\hat{e}(X) = P(T = 1|X)$
  - **监督学习**：基于样本的观测变量/混杂变量X来预测其干预变量T。
  - 二分类问题，逻辑斯蒂回归或深度神经网络



- 通过约束样本倾向得分相似或一样来匹配T=1和 T=0的样本：

$$Distance(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$$

$$Distance(X_i, X_j) \leq \epsilon$$

- 混杂变量高维的挑战：
  - 从匹配阶段转移到了倾向得分评估阶段

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

倾向得分逆加权法

- 为什么利用倾向得分的逆对样本加权有用?
  - 倾向得分导致了混杂变量X在T=1和T=0两组样本之间的分布不一致

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

样本	$e(X)$	$1 - e(X)$	样本数量	T=1组样本数量	T=0组样本数量
A	0.7	0.3	10	7	3
B	0.6	0.4	50	30	20
C	0.2	0.8	40	8	32

分布不一致

样本数量	T=1组样本数量	T=0组样本数量
A	10	10
B	50	50
C	40	40

混杂变量的分布一致了！

样本通过倾向得分的逆来加权：

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

倾向得分逆加权法

- 倾向得分逆加权法算法（Inverse of Propensity Weighting, IPW）：

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}$$

$$w_i = \frac{T_i}{e_i} + \frac{1-T_i}{1-e_i}$$

  - IPW算法通过利用倾向得分的逆来对样本进行重采样/加权，使得T=1和T=0两组样本的混杂变量X分布一致。

- 为什么IPW方法有效？考虑如下式子：

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)}$$

倾向得分逆加权法

- 如果评估的倾向得分等于真实的倾向得分 $\hat{e}(X) = e(X)$

$$
\begin{aligned}
E\left\{\frac{TY}{e(X)}\right\} &= E\left\{\frac{TY_1}{e(X)}\right\} = E\left[E\left\{\frac{TY_1}{e(X)}|Y_1, X\right\}\right] \\
&= E\left\{\frac{Y_1}{e(X)}E(T|Y_1, X)\right\} = E\left\{\frac{Y_1}{e(X)}E(T|X)\right\} \\
&= E\left\{\frac{Y_1}{e(X)}e(X)\right\} = E\left(Y_1\right)
\end{aligned}
$$

(1) $\boldsymbol{Y = T * Y_1 + (1 - T) * Y_0}$

(2) $\boldsymbol{T \perp (Y_1, Y_0) \mid X}$

(3) $\boldsymbol{e(X) = E(T|X)}$

- 同理，可得：

$$E\left\{\frac{(1-T)Y}{1-e(X)}\right\} = E(Y_0)$$

$$ATE = E[Y(1) - Y(0)]$$

倾向得分逆加权法

- 如果评估的倾向得分等于真实的倾向得分 $\hat{e}(X) = e(X)$，那么IPW方法是无偏的（unbiased）

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}(X_i)} = E(Y_1 - Y_0)$$

- IPW方法被广泛应用于各种实际场景中（经济学、社会学等）

- 但IPW方法对因果评估的无偏性依赖于倾向得分模型的准确性
- 且当倾向得分 e 接近于 0 或 1 时，因果评估的方差会很大

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

双稳健算法

- 平均因果效应的定义： $ATE = E[Y(T=1) - Y(T=0)]$
- 简单的结果变量回归方法：

$$m_1 = E(Y|T=1, X) \quad 以及 \quad m_0 = E(Y|T=0, X)$$

  - 如果结果变量回归模型是正确的，那么因果效应评估是无偏的
- 倾向得分逆加权方法：
  - 如果倾向得分回归模型是正确的，那么因果效应评估是无偏的

- 双稳健算法：将两者结合起来

双稳健算法

$$m_0 = E(Y|T=0, X)$$
$$m_1 = E(Y|T=1, X)$$

- 双稳健算法：

$$ATE_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{\{T_i - \hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{m}_1(X_i)\right]$$
$$- \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)\,Y_i}{1-\hat{e}(X_i)} + \frac{\{T_i - \hat{e}(X_i)\}}{1-\hat{e}(X_i)}\hat{m}_0(X_i)\right]$$

- 如果结果变量回归模型或者倾向得分回归模型是正确的，那么因果效应评估是无偏的
- 这种属性称之为双稳健

双稳健算法

- 理论证明：

$$
\begin{aligned}
& E\left[\frac{TY}{\hat{e}(X_i)} - \frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{m}_1(X_i)\right] \\
= \quad & E\left[\frac{TY_1}{\hat{e}(X_i)} - \frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{m}_1(X_i)\right] \\
= \quad & E\left[Y_1 + \frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\{Y_1 - \hat{m}_1(X_i)\}\right] \\
= \quad & E(Y_1) + E\left[\frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\{Y_1 - \hat{m}_1(X_i)\}\right]
\end{aligned}
$$

双稳健算法

$$m_0 = E(Y|T=0, X)$$
$$m_1 = E(Y|T=1, X)$$

- 双稳健算法：

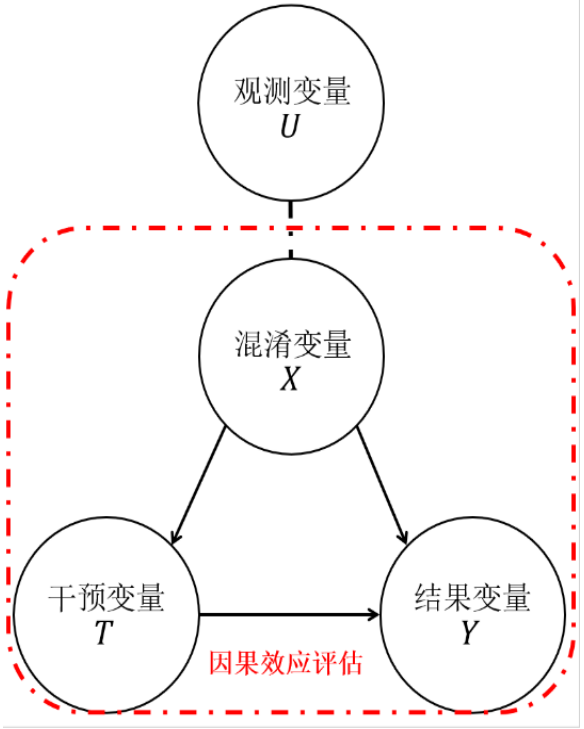
$$ATE_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{\{T_i - \hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{m}_1(X_i)\right]$$
$$- \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)\,Y_i}{1-\hat{e}(X_i)} + \frac{\{T_i - \hat{e}(X_i)\}}{1-\hat{e}(X_i)}\hat{m}_0(X_i)\right]$$

  - 如果结果变量回归模型或者倾向得分回归模型是正确的，那么因果效应评估是无偏的

- 但是如果两个模型都不对，会增加因果效应评估的偏差

基于倾向得分的方法

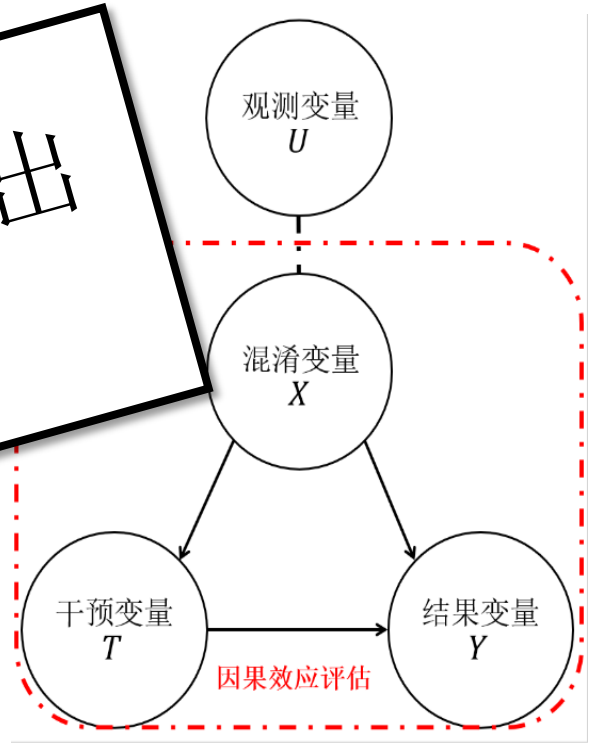
- 回顾：
  - 倾向得分匹配法
  - 倾向得分逆加权法
  - 双稳健算法
- 需要对倾向得分进行估计
  - 将所有的观测变量都视为混杂变量
  - 在大数据时代，数据往往高维
  - 但，并不是所有的变量都是混杂变量



基于倾向得分的方法

- 回顾：
  - 倾向得分匹配法
  - 倾向得分逆加权法
  - 双稳健算法
- 需要对倾向
  - 将所有
  - 在大数
  - 但，并

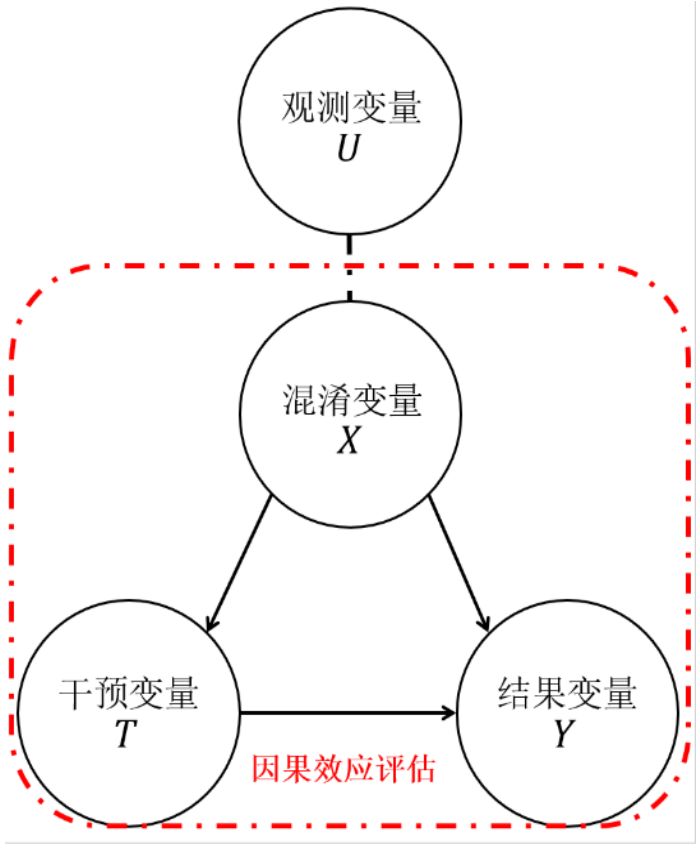
如何从高维变量中自动分离出混杂变量？



因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

倾向得分逆加权方法



之前的因果推断框架

- 将所有的观测变量 **U** 都视为混杂变量 **X**
- 倾向得分评估：

$$e(\mathbf{U}) = p(T=1|\mathbf{U}) = p(T=1|\mathbf{X}) = e(\mathbf{X})$$

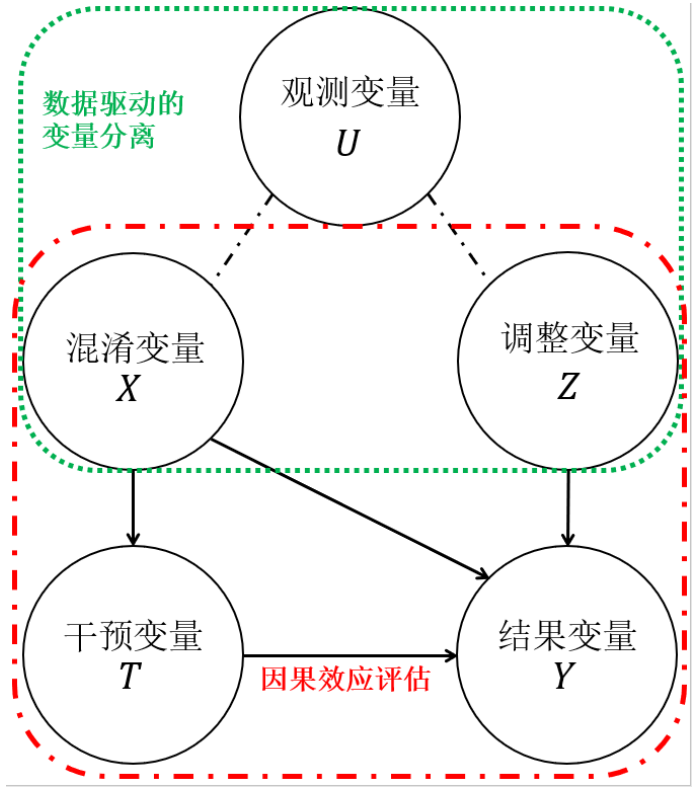
- 基于倾向得分逆加权调整结果变量：

$$Y^{\star} = Y^{obs} \cdot \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} = Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

- 基于倾向得分逆加权（IPW）计算平均因果效应：

$$\widehat{ATE}_{IPW} = \hat{E}(Y^{\star})$$

数据驱动的变量分解算法（D²VD）



我们提出的因果推断框架

- 可分离假设：
  - 所有的观测变量 U 可分解为3个部分：混杂变量 **X**, 调整变量 **Z**， 以及 无关变量 **I**（图中省略）。
- 倾向得分估计：

$$e(\mathbf{X}) = p(T = 1|\mathbf{X})$$

- 基于倾向得分调整结果变量：

$$Y^{+} = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

- 计算平均因果效应：

$$\widehat{ATE}_{D^2VD} = \hat{E}(Y^{+})$$

数据驱动的变量分解算法（D²VD）

$$minimize \quad \|Y^+ - h(\mathbf{U})\|^2 \qquad 其中，\quad Y^+ = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

$$e(\mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}\beta)} \qquad \phi(\mathbf{Z}) = \mathbf{Z}\alpha,$$

用 U 替换 X, Z

$$h(\mathbf{U}) = \mathbf{U}\gamma,$$

$$minimize \quad \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2, \qquad 其中，\quad W(\beta) := \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))}$$

$$s.t. \quad \sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)) < \tau,$$

$$\|\alpha\|_1 \le \lambda, \ \|\beta\|_1 \le \delta, \ \|\gamma\|_1 \le \eta, \ \|\alpha \odot \beta\|_2^2 = 0.$$

α, β, γ

- 调整变量： $\mathbf{Z} = \{\mathbf{U}_i : \hat{\alpha}_i \neq 0\}$
- 混杂变量： $\mathbf{X} = \{\mathbf{U}_i : \hat{\beta}_i \neq 0\}$
- 因果效应： $\widehat{ATE}_{D^2VD} = E(\mathbf{U}\hat{\gamma})$

数据驱动的变量分解算法（D²VD）

偏差分析：

我们数据驱动的变量分解算法对于因果效应评估是无偏的

THEOREM 1. *Under assumptions 1-4, we have*

$$E(Y^{+}|\boldsymbol{X},\boldsymbol{Z}) = E(Y(1) - Y(0)|\boldsymbol{X},\boldsymbol{Z}).$$

方差分析：

对于因果估计的渐进方差，数据驱动的变量分解算法要小于倾向得分逆加权法

THEOREM 2. *The asymptotic variance of our adjusted estimator* $\widehat{ATE}_{adj}$ *is no greater than IPW estimator* $\widehat{ATE}_{IPW}$*:*

$$\sigma^2_{adj} \leq \sigma^2_{IPW}.$$

真实数据上实验分析

- 数据描述：
  - 在线广告营销（LONGCHAMP）
  - 用户反馈：14,891 点赞； 93,108 不喜欢
  - 每个用户收集了56维特征：
    - 年龄，性别，朋友数量，设备型号，用户设定等

2015

- 实验设定：
  - 结果变量 Y：用户反馈
  - 干预变量 T：用户的某一维特征
  - 观测变量 U：用户其他特征

Y = 1，如果点赞
Y = 0，如果不喜欢

真实数据上实验分析

- 平均因果效应估计结果：

No.	Features	$\widehat{ATE}_{D^2VD}$ (SD)	$\widehat{ATE}_{IPW}$ (SD)	$\widehat{ATE}_{DR}$ (SD)	$ATE_{matching}$
1	No. friends (> 166)	0.295 (0.018)	0.240 (0.026)	0.297(0.021)	0.276
2	Age (> 33)	-0.284 (0.014)	-0.235 (0.029)	-0.302(0.068)	-0.263
3	Share Album to Strangers	0.229 (0.030)	0.236 (0.030)	-0.034(0.021)	n/a
4	With Online Payment	0.226 (0.019)	0.260 (0.029)	0.244(0.028)	n/a
5	With High-Definition Head Portrait	0.218 (0.028)	0.203 (0.032)	0.237(0.046)	n/a
6	With WeChat Album	0.191 (0.014)	0.237 (0.021)	0.097(0.050)	n/a
7	With Delicacy Plugin	0.124 (0.038)	-0.253 (0.037)	0.067(0.051)	0.099
8	Device (iOS)	0.100 (0.024)	0.206 (0.012)	0.060(0.021)	0.085
9	Add friends by Drift Bottle	-0.098 (0.012)	0.016 (0.019)	-0.115(0.015)	-0.032
10	Gender (Male)	-0.073 (0.017)	-0.240 (0.029)	0.065(0.055)	-0.097

- ✓ 数据驱动的变量分解算法（ D²VD ）评估平均因果效应**更准确**。
- ✓ 数据驱动的变量分解算法（ D²VD ） 能**降低因果效应估计的方差**。
- ✓ **年轻的女士**有更高的概率是 LONGCHAMP 的潜在用户（更喜欢 LONGCHAMP 的广告）。

真实数据上实验分析

- 变量分解实验结果：

Table 4: Confounders and adjusted variables when we set feature "Add friends by Shake" as treatment.

Confounders	Adjustment Variables
Add friends by Drift Bottle	No. friends
Add friends by People Nearby	Age
Add friends by QQ Contacts	With WeChat Album
Without Friends Confirmation Plugin	Device

✓ 分离出来的**混杂变量**是微信上其他加好友的方式。
✓ 分离出来的**调整变量**对结果变量有很强的效应。
✓ 数据驱动的变量分解算法（ D^2VD ） 能够**准确的**分离混杂变量和调整变量。

总结：基于倾向得分的方法

- 倾向得分匹配法（PSM）：
  - 基于倾向得分来匹配样本
- 倾向得分逆加权法（IPW）：
  - 基于倾向得分的逆来对样本加权
- 双稳健算法（DR）：
  - 结合倾向得分逆加权法和结果回归模型

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

将所有的观测变量都视为混杂变量，忽略了非混杂变量，如调整变量。

- **数据驱动的变量分解算法（D^2VD）：**
  - 自动分离混杂变量和调整变量
  - 混杂变量：用于准确评估倾向得分，用于样本逆加权
  - 调整变量：通过结果变量回归，降低因果效应评估的方差
  - 提升因果推断准确率的同时降低评估方差
- 但，这些方法都需要对倾向得分进行估计

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

混杂变量直接平衡法

- 回顾：基于倾向得分的方法
  - 通过样本加权来实现混杂变量平衡
  - 但，需要准确估计倾向得分
  - 权重会过大如果倾向得分趋近于 0 或 1

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

- 我们是否可以通过机器学习方式来学习样本权重，直接平衡混杂变量的分布？

当然可以！

混杂变量直接平衡法

- **动机**：变量的分布由变量的各阶矩（moments）唯一决定
- **方法**：学习样本使得混杂变量的矩在干预组（T=1）和对照组（T=0）一致，即实现了混杂变量直接平衡

$$\min_{W} \left\| \overline{\mathbf{X}}_t - \mathbf{X}_c^T W \right\|_2^2$$

干预组中
混杂变量 X 的一阶矩

对照组中
混杂变量 X 的一阶矩

基于矩能唯一决定变量分布，我们可以直接学习样本权重来实现混杂变量平衡，而无需任何模型假设

混杂变量直接平衡法

- **动机**：变量的分布由变量的各阶矩（moments）唯一决定
- **方法**：学习样本使得混杂变量的矩在干预组（T=1）和对照组（T=0）一致，即实现了混杂变量直接平衡

$$\min_{W} \left\| \overline{\mathbf{X}}_t - \mathbf{X}_c^T W \right\|_2^2$$

干预组中
混杂变量 X 的一阶矩

对照组中
混杂变量 X 的一阶矩

- 平均因果效应估计：$\widehat{ATT} = \sum_{i:T_i=1} \frac{1}{n_t} Y(1) - \sum_{j:T_j=0} W_j Y(0)$

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

熵平衡法

$$\begin{aligned} \min_{W} \quad & W\log(W) \\ s.t. \quad & \|\overline{\mathbf{X}}_t - \mathbf{X}_c^T W\|_2^2 = 0 \\ & \sum_{i=1}^{n} W_i = 1, W \succeq 0 \end{aligned}$$

- 学习样本权重 w 直接平衡混杂变量的分布（一阶矩）
- 样本权重可能不唯一，因此，最大化样本权重 w 的熵
- 但，方法将所有的变量都视为混杂变量，且在平衡过程中同等对待

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

近似残差平衡法

- 1. 计算样本权重 W，直接平衡混杂变量（一阶矩），如下：

$$W = \text{argmin}_W \left\{ (1-\zeta)\|W\|_2^2 + \zeta \left\| \overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^{\top} W \right\|_{\infty}^2 \text{ s.t. } \sum_{\{i:T_i=0\}} W_i = 1 \text{ and } W_i \geq 0 \right\}$$

- 2. 基于线性模型（参数为 β_c ）拟合结果变量（基于对照组数据）

$$\hat{\beta}_{\text{c}} = \text{argmin}_{\beta} \left\{ \sum_{\{i:W_i=0\}} \left(Y_i^{\text{obs}} - X_i \cdot \beta \right)^2 + \lambda \left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right) \right\}$$

- 3. 估计因果效应，如下：

$$\widehat{ATT} = \overline{Y}_t - \left(\overline{X}_t \cdot \hat{\beta}_c + \sum_{\{i:T_i=0\}} W_i \left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_{\text{c}} \right) \right)$$

- 双稳健性： 样本权重 W 或 回归模型参数 β_c 一个学好即可
- 但，方法将所有的变量都视为混杂变量，且在平衡过程中同等对待

混杂变量直接平衡法

- 回顾：
  - *熵平衡法，近似残差平衡法等*
  - 变量各阶矩（moments）唯一决定其分布
  - 学习样本权重 W 来直接平衡混杂变量的各阶矩：

$$\min_{W} \left\| \overline{\mathbf{X}}_t - \mathbf{X}_c^T W \right\|_2^2$$

干预组中
混杂变量 X 的一阶矩

对照组中
混杂变量 X 的一阶矩

  - 但，这些方法将所有变量都视为混杂变量，且平衡过程中同等对待
  - 不是所有变量都是混杂变量，且不同混杂变量会带来的偏差不一样

混杂变量直接平衡法

- 回顾：
  - *熵平衡法，近似残差平衡法等*
  - 变量各阶矩（moments）唯一决定其分布
  - 学习样本权重 W 来直接平衡混杂变量的

如何区分混杂变量以及它们带来的偏差？

照组中
混杂变量 X 的一阶矩

- 但，这 所有变量都视为混杂变量，且平衡过程中同等对待
- 不是所有变量都是混杂变量，且不同混杂变量会带来的偏差不一样

因果推断的经典方法

- **匹配法**
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法（Inverse of Propensity Weighting, IPW）
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

混杂变量区分平衡法

- **想法：** 同时学习*混杂权重* $\boldsymbol{\beta}$ 和*样本权重* $\boldsymbol{W}$。

- *混杂权重*决定哪些变量是混杂变量及其带来的混杂偏差的强度。
- *样本权重*用于平衡混杂变量的分布

$$\min \quad \left(\beta^T \cdot (\overline{\mathbf{X}}_t - \mathbf{X}_c^T W)\right)^2$$

如何学习混杂权重和样本权重？

混杂权重学习

- 考虑 X， T，和 Y 三者之间的一般关系：

$$Y = f(\mathbf{X}) + T \cdot g(\mathbf{X}) + \epsilon \quad \rightarrow \quad \begin{aligned} ATT &= E(g(\mathbf{X}_t)) \\ Y(0) &= f(\mathbf{X}) + \epsilon \end{aligned}$$

$$\begin{aligned} f(\mathbf{X}) &= \mathbf{a}_1\mathbf{X} + \sum_{ij} a_{ij}X_iX_j + \sum_{ijk} a_{ijk}X_iX_jX_k + \cdots + R_n(\mathbf{X}) \\ &= \alpha\mathbf{M}. \qquad\qquad \mathbf{M} = (\mathbf{X}, X_iX_j, X_iX_jX_k, \cdots). \end{aligned}$$

混杂权重　　混杂偏差

$$\widehat{ATT} = ATT + \sum_{k=1}^{p} \alpha_k \left(\sum_{i:T_i=1} \frac{1}{n_t} M_{i,k} - \sum_{j:T_j=0} W_j M_{j,k} \right) + \phi(\epsilon).$$

如果 $\alpha_k = 0$，则 M_k 就不是混杂便，没必要平衡。
不同的混杂变量具有不同的混杂权重。

混杂权重学习

推论：

- 在观测学习中，并不是所有的观测变量都是混杂变量，且不同的混杂变量对因果推断带来的混杂偏差是不同的，其偏差程度可通过混杂权重来表示。
- 混杂变量的混杂权重可通过潜在结果变量 $Y(0)$ 对增广的观测变量 M 的回归学习得到。

$$\mathbf{M} = (\mathbf{X}, X_i X_j, X_i X_j X_k, \cdots).$$

样本权重学习

$$\mathbf{M} = (\mathbf{X}, X_i X_j, X_i X_j X_k, \cdots).$$

- 任何变量的分布都由其各阶矩（moments）唯一决定
- 样本权重 W 可通过直接约束混杂变量各阶矩平衡来学习：

$$\min \quad \left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2$$

干预组中
混杂变量的各阶矩

控制组中
混杂变量的各阶矩

可以直接学习样本权重来实现混杂变量平衡，而无需任何模型假设

混杂变量区分平衡法

- 目标函数

$$\min \quad \left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2 + \lambda \sum_{j:T_j=0} (1+W_j) \cdot (Y_j - M_j \cdot \beta)^2,$$

$$s.t. \quad \|W\|_2^2 \le \delta, \ \|\beta\|_2^2 \le \mu, \ \|\beta\|_1 \le \nu, \mathbf{1}^T W = 1 \ \ and \ \ W \succeq 0$$

如果将上述混杂权重设为单位向量（全1向量），则熵平衡法和近似残差平衡法是我们混杂变量平衡法的特例。

实验结果及分析

- 对比方法：
  - **直接估计法（dir）：** 直接计算T=1和T=0两组样本结果期望的差值
  - **倾向得分逆加权法（IPW）：**基于倾向得分的逆加权样本
  - **双稳健算法（DR）：**结合倾向得分逆加权模型和结果变量回归模型
  - **熵平衡法（ENT）：** 学习样本权重直接平衡混杂变量分布
  - **近似残差平衡法（ARB）：** 结合混杂变量平衡模型和结果变量回归模型

- 评估指标：

$$
\begin{aligned}
Bias &= \left|\frac{1}{K}\sum_{k=1}^{K}\widehat{ATT}_k - ATT\right| \\
SD &= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATT}_k - \frac{1}{K}\sum_{k=1}^{K}\widehat{ATT}_k)^2} \\
MAE &= \frac{1}{K}\sum_{k=1}^{K}\left|\widehat{ATT}_k - ATT\right| \\
RMSE &= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATT}_k - ATT)^2}
\end{aligned}
$$

实验结果及分析

- LaLonde 数据集：评估参加职业培训项目对个人实际收入的影响
  - **随机实验数据：**提供了参加职业培训项目对个人实际收入的因果效应值
  - **观测数据：**可用于检测各个因果推断算法对因果效应评估的能力

- 实验设定：
  - **V-RAW：**仅仅利用原有的10维变量作为混杂变量进行平衡，变量包括年龄，教育程度，婚姻状态，种族信仰，当前实际收入等。
  - **V-INTERACTION：**基于10维变量、它们之间的一阶交互项以及平方项来进行混杂变量平衡。

实验结果及分析

因果推断的实验结果

Variables Set	V-RAW		V-INTERACTION	
Estimator	$\widehat{ATT}$	*Bias* (SD)	$\widehat{ATT}$	*Bias* (SD)
$\widehat{ATT}_{dir}$	-8471	10265 (374)	-8471	10265 (374)
$\widehat{ATT}_{IPW}$	-4481	6275 (971)	-4365	6159 (1024)
$\widehat{ATT}_{DR}$	1154	639 (491)	1590	204 (812)
$\widehat{ATT}_{ENT}$	1535	259 (995)	1405	388 (787)
$\widehat{ATT}_{ARB}$	1537	257 (996)	1627	167 (957)
$\widehat{ATT}_{DCB}$	1958	**164** (728)	1836	**43** (716)

混杂变量区分平衡法（DCB）比其他方法更精准

在V-INTERACTION设定下，混杂变量区分平衡法（DCB）效果更好，及考虑变量之间的交互有利于提升混杂变量平衡

总结：混杂变量直接平衡法

- **动机：**变量的各阶矩（moments）能唯一决定其分布
- **熵平衡法**：
  - 学习样本权重直接平衡混杂变量分布，同时约束权重熵最大
- **近似残差平衡法**：
  - 结合混杂变量平衡模型和结果变量回归模型，实现双稳健
- 将所有变量都视为混杂变量，且平衡过程中同等对待
- 但，并非所有变量都是混杂，且不同混杂变量会带来的偏差不一样
- **混杂变量区分平衡法（DCB）**
  - 理论上证明了混杂变量区分的必要性
  - 实验上验证了区分性平衡混杂变量能提升因果推断的准确性

总结：因果推断的经典方法

- **匹配法** 限定于低维数据场景
- **基于倾向得分的方法**
  - 倾向得分匹配法
  - 倾向得分逆加权法
  - 双稳健算法
  - 数据驱动的变量分解算法
- **混杂变量直接平衡法**
  - 熵平衡法
  - 近似残差平衡法
  - 混杂变量区分平衡法

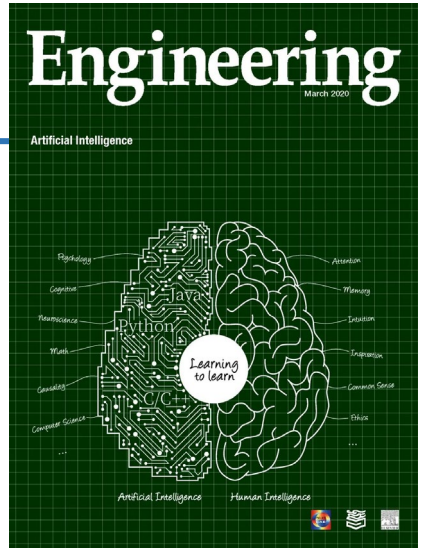
将所有观测变量都视为混杂变量

并非所有变量都是混杂变量

平衡变量过程中，所有混杂变量平等对待

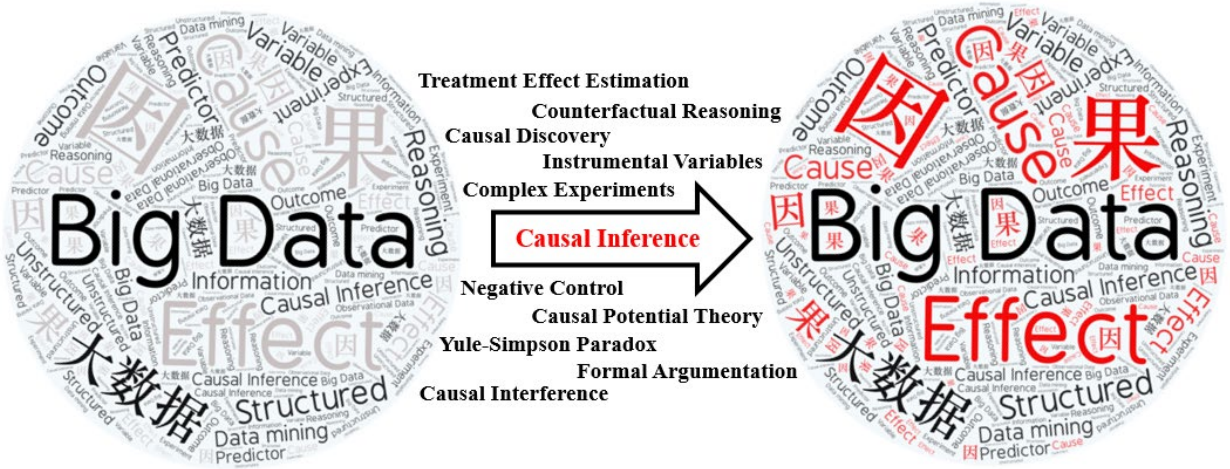
不同混杂变量的偏差不一样

Causal Inference – 因果推理



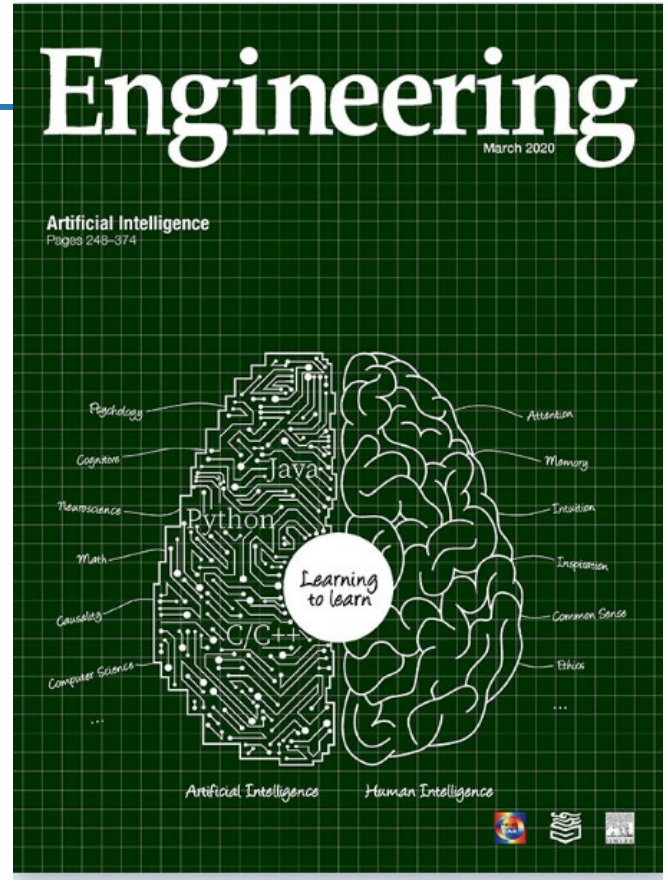
The official journal of the Chinese Academy of Engineering

Survey Paper：
Causal Inference（因果推理）



Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., Jiang, Z. (2020). Causal Inference. *Engineering*. http://www.engineering.org.cn/ch/10.1016/j.eng.2019.08.016

Causal Inference – 因果推理

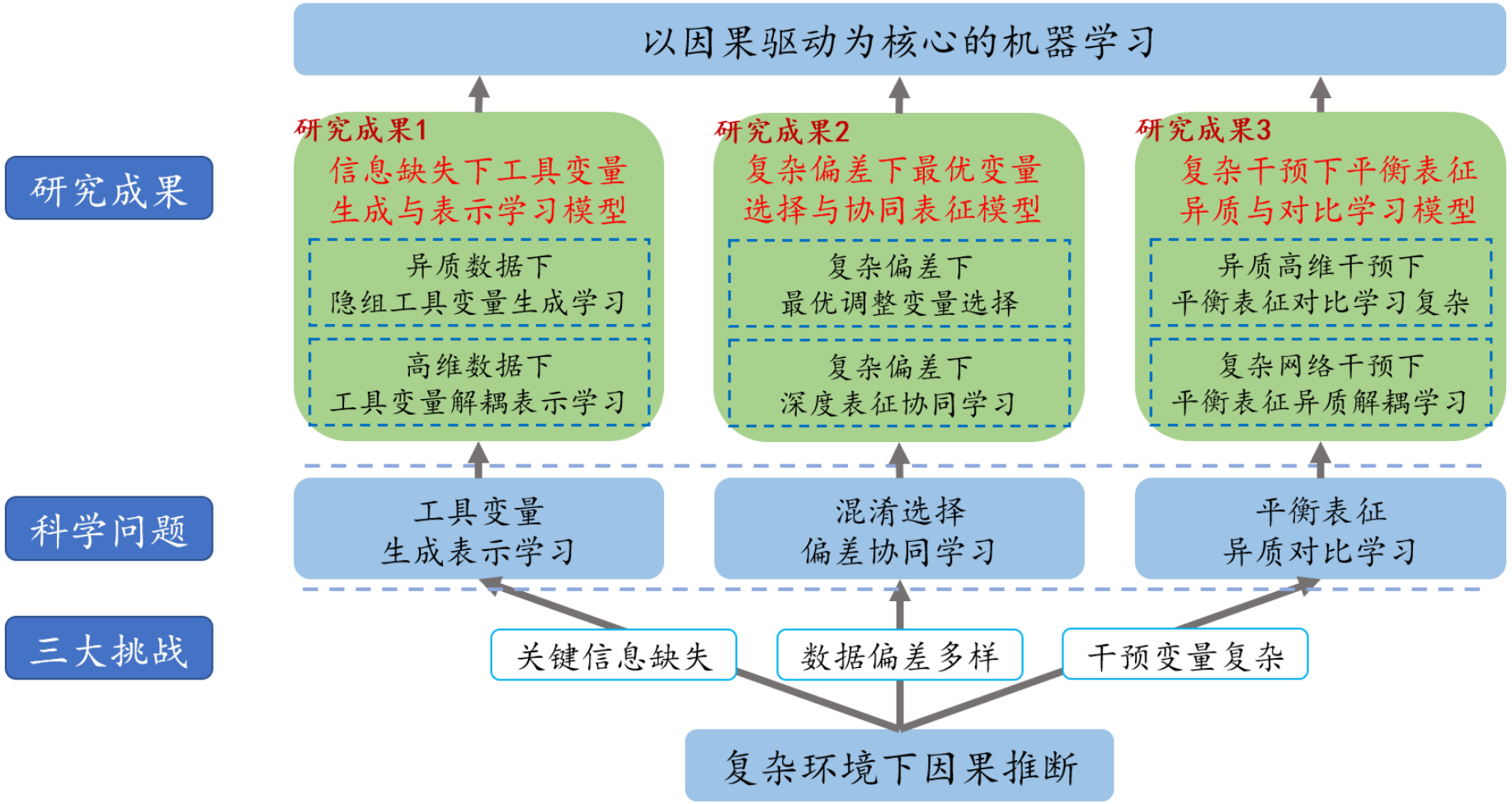


- **Engineering 综述论文**
- 况琨：平均因果效应评估-简要回顾与展望
- 李廉：反事实推理的归因问题
- 耿直：辛普森悖论和替代指标悖论
- 徐雷：因果发现CPT（因果势理论）方法
- 张坤：从观测数据中发现因果关系
- 廖备水，黄华新：形式论辩在因果推理和解释中的作用
- 丁鹏：复杂实验中的因果推断
- 苗旺：观察性研究中的工具变量和阴性对照方法
- 蒋智超：有干扰下的因果推断

http://www.engineering.org.cn/ch/10.1016/j.eng.2019.08.016

基础理论——复杂环境下因果推断

- 复杂环境下因果推断的三大挑战
  - 关键信息缺失、数据偏差多样、干预变量复杂



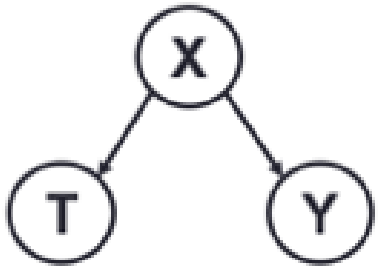
复杂偏差下因果推断

2021诺贝尔经济学奖：
基于工具变量的因果推断

因果

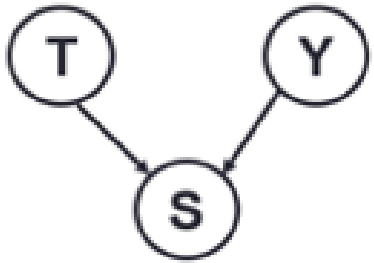
可解释
稳定/鲁棒
可决策

混淆偏差



虚假关联: T is correlated with Y ignoring X

选择偏差



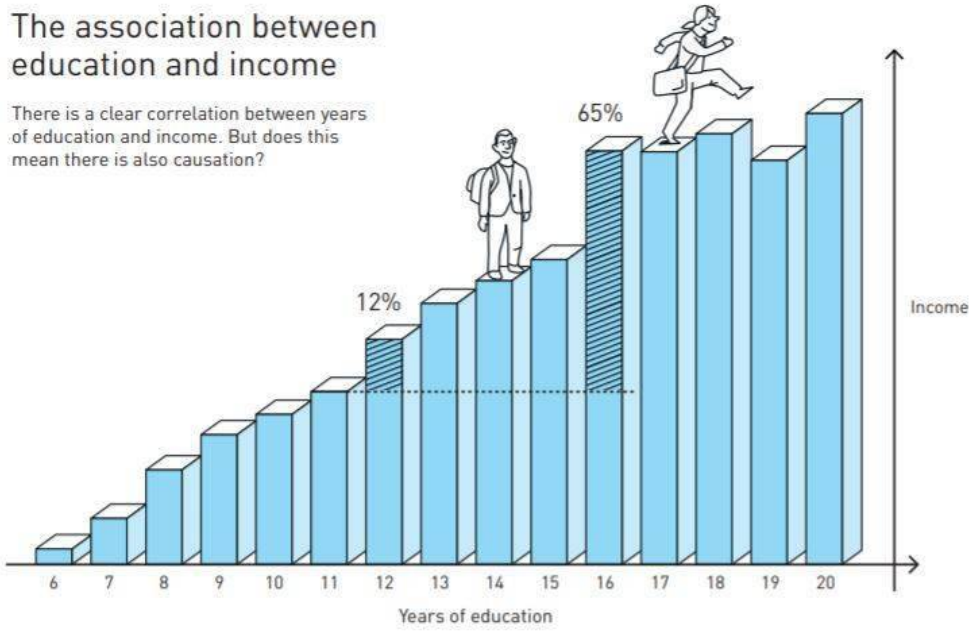
虚假相关: T is correlated with Y given S

混淆偏差下因果推断

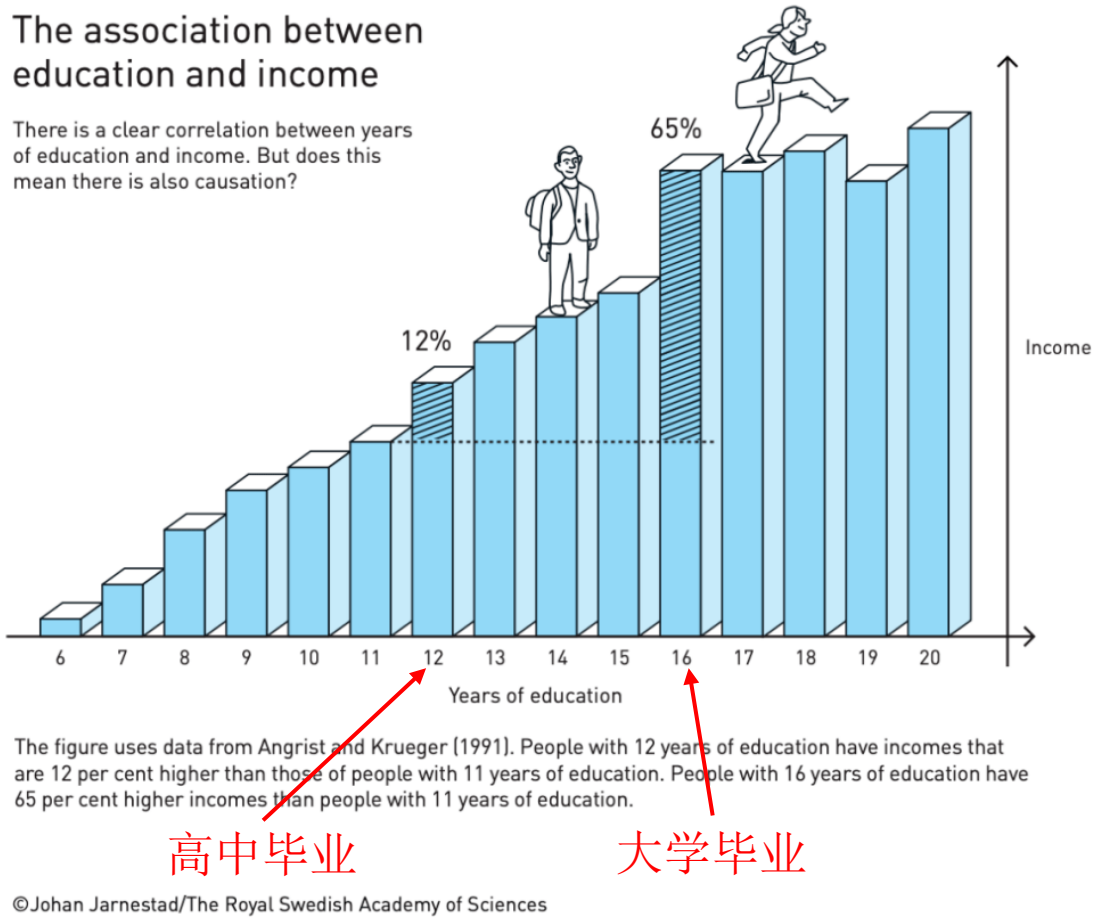
在自然实验中甄别混淆变量：知其然、且知其所以然

2021年度诺贝尔经济学奖一半奖项授予加州大学伯克利分校戴维·卡德（David Card），以表彰“他对劳动经济学的实证研究性的贡献”；另一半授予麻省理工学院的乔舒亚·安格里斯特（Joshua D. Angrist）和斯坦福大学的吉多·因本斯（Guido W. Imbens），以表彰“他们在因果关系分析方面的方法论贡献”。

“一个人的出生季节或月份”是一个神奇的**工具变量（Instrumental Variable）**，与其他（混淆）因素区分开，干净识别出了“多读一年书对未来收入造成的影响”。

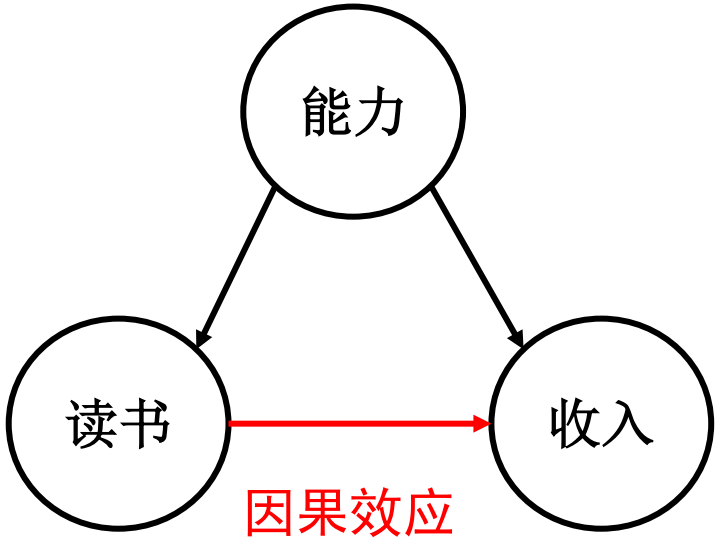


读书是否有用?

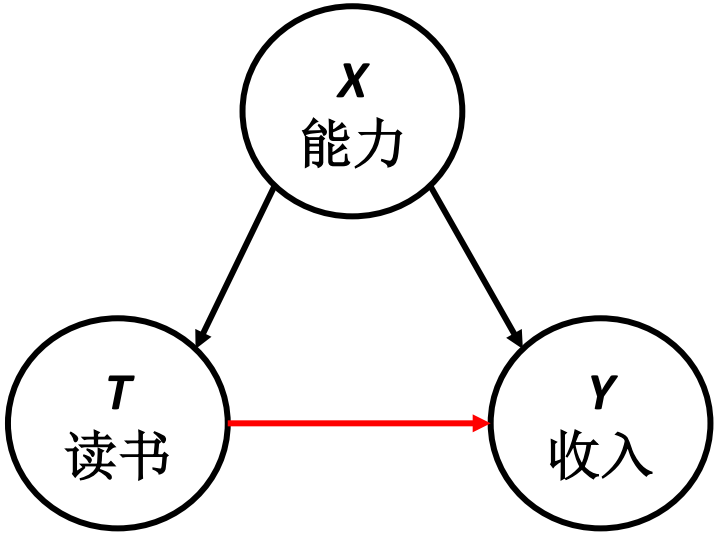


- 高中毕业 v.s. 高中肄业：12%
- 大学毕业 v.s. 高中毕业：53%

- 高出来的收入是否是“因为”多读了几年书?



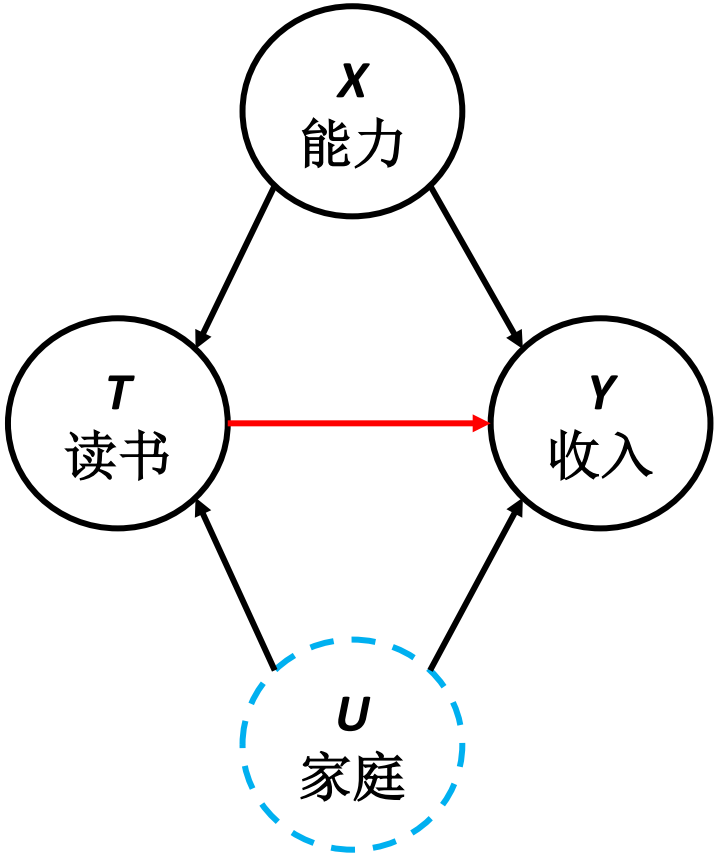
评估教育对收入的影响



- 混杂变量 X：能力
- 干预变量 T=1/0：是否高中毕业
- 结果变量 Y：收入

- 分析读书与收入之间的因果效应
  - 控制X（混杂变量）的分布
  - 方法：匹配法，基于倾向得分的方法等
  - 假设：所有的混杂变量都能被观测

评估教育对收入的影响

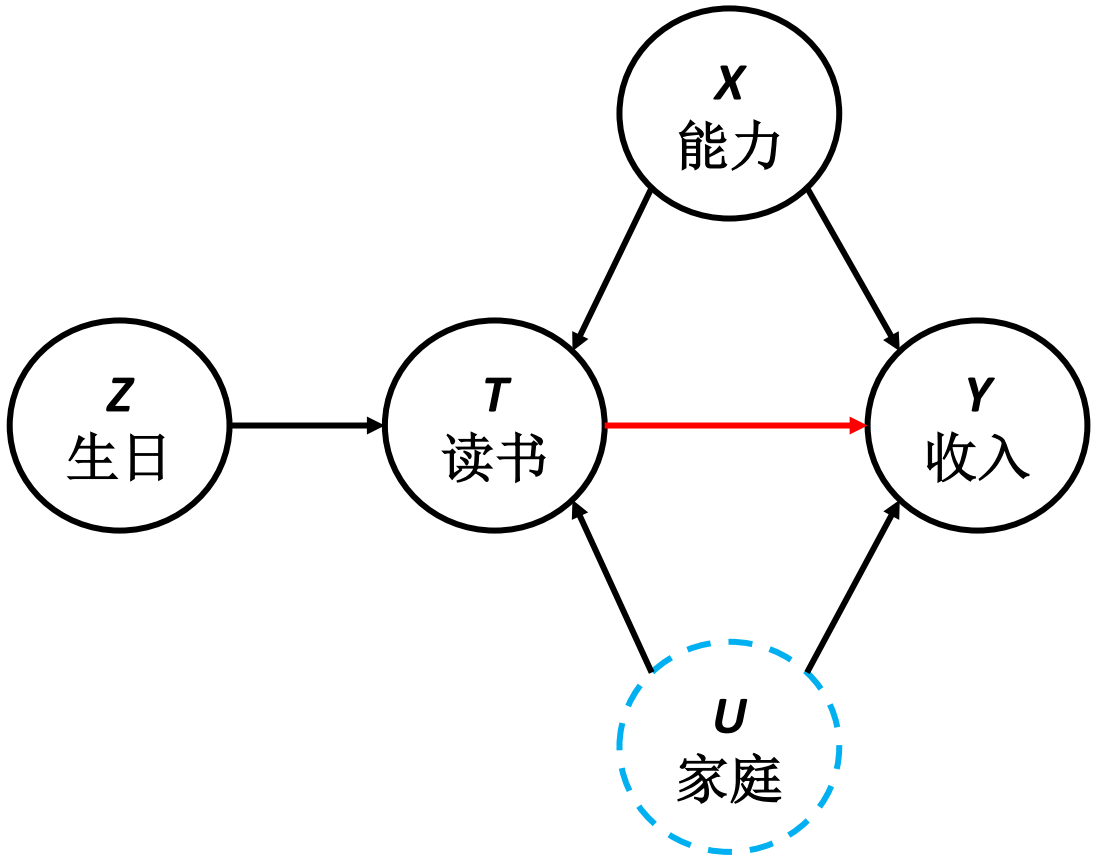


- 混杂变量 X：能力
- 干预变量 T=1/0：是否高中毕业
- 结果变量 Y：收入

- 分析读书与收入之间的因果效应
  - 控制X（混杂变量）的分布
  - 方法：匹配法，基于倾向得分的方法等
  - 假设：所有的混杂变量都能被观测

- 实际应用中：很多混杂变量无法观测
  - 如家庭情况（U）无法被测量

评估教育对收入的影响---工具变量



- 工具变量（Z）的条件
- 相关性：$P(T|Z) \neq P(T)$
- 排他性：$P(Y|Z,T,U) \neq P(Y|T,U)$
- 无混杂性：$Z \perp U$

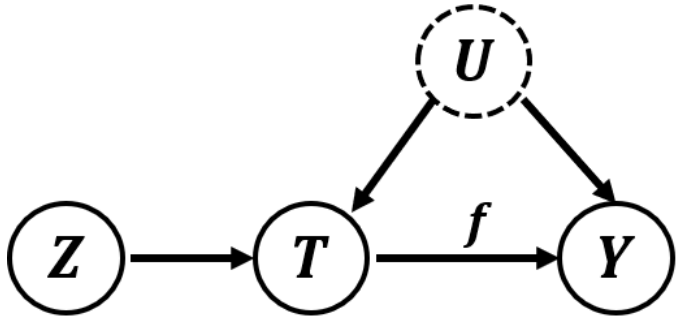
- 分析读书与收入之间的因果效应
- 生日可以作为工具变量

✓生日影响读书

✓生日不影响能力和家庭情况

✓生日不直接影响收入

评估教育对收入的影响---工具变量



- Z：生日，春季出生/冬季出生
- T：是否高中毕业
- Y：收入
- U：能力，家庭情况等

两阶段回归

第一阶段：用Z去回归T　$\hat{T} = \hat{g}(Z)$

第二阶段：用$\hat{T}$去回归Y　$\hat{Y} = \hat{f}(\hat{T})$

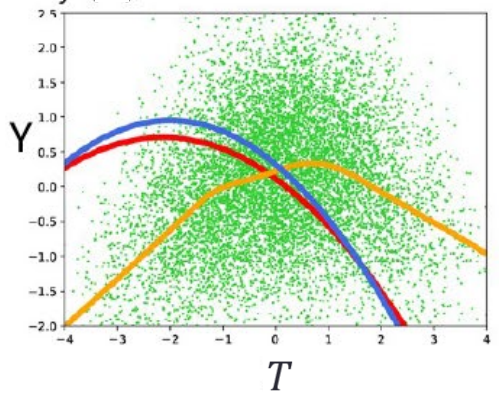
$Z \sim \mathcal{N}(0,1)$

$U \sim \mathcal{N}(0,1)$

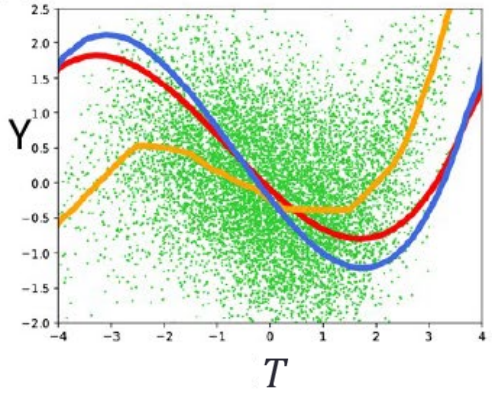
$T = Z + U$

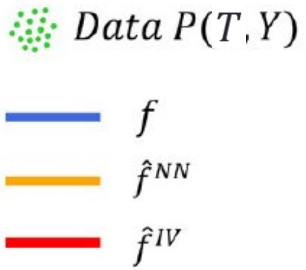
$Y = f(T) + U$

$f(T) = -0.1 \cdot T^2 - 0.4 \cdot T$



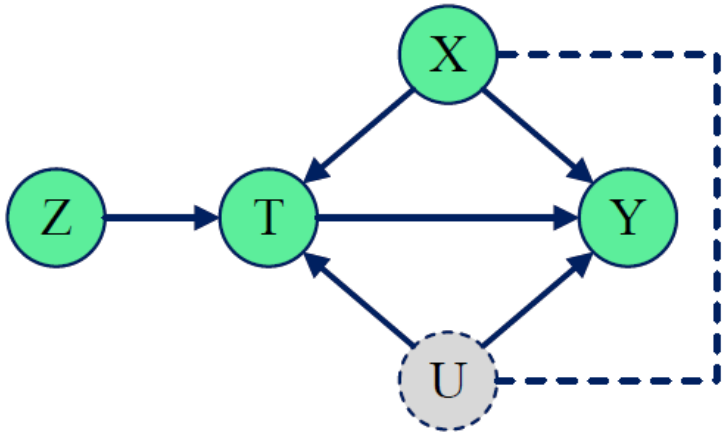
$f(T) = 0.05 \cdot T^3 + 0.1 \cdot T^2 - 0.8 \cdot T$

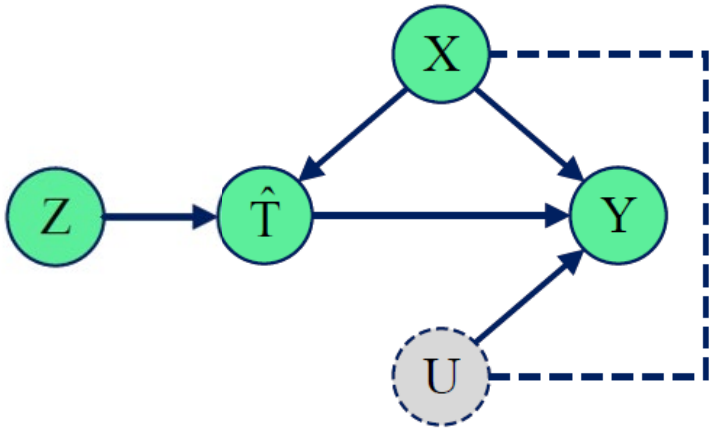




局限于线性假设
需要提前定义工具变量

非线性工具变量回归





非线性工具变量回归 (DeepIV, KernelIV et.al)

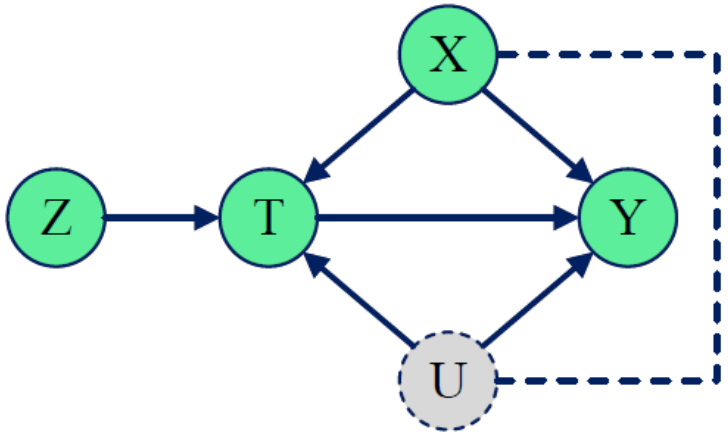
第一阶段：用 Z 和 X 去回归 T $\hat{T} = \hat{g}(Z, X)$

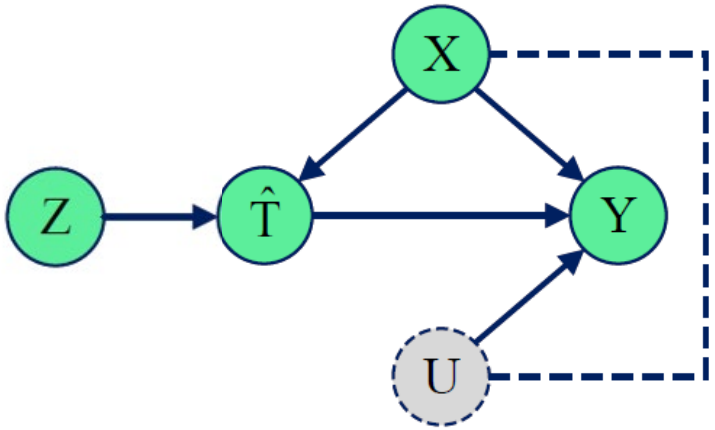
第二阶段：用 $\hat{T}$ 和 X 去回归 Y $\hat{Y} = \hat{f}(\hat{T}, X)$

第一阶段回归会给第二阶段回归引入混杂偏差（X）

混杂变量平衡 + 工具变量回归

混杂平衡的工具变量回归（CB-IV）





CB-IV (Confounder Balanced IV regression):

第一阶段（干预回归）：用 Z 和 X 去回归 T $\hat{T} = \hat{g}(Z, X)$

混杂平衡：学习混杂平衡的表征 $\phi(X)$，使其满足 $\hat{T} \perp \phi(X)$

第二阶段（结果回归）：用 $\hat{T}$ 和$\phi(X)$去回归 Y $\hat{Y} = \hat{f}(\hat{T}, \phi(X))$

Wu A, Kuang K, Li B, et al. Instrumental Variable Regression with Confounder Balancing, ICML 2022

实验分析和结果

Table 2: The bias (mean ± std) of ATE estimation on real-world data (Data-m_Z-m_X-m_U)

Within-Sample				
Method	**IHDP-2-6-0**	**IHDP-2-4-2**	**Twins-5-8-0**	**Twins-5-5-3**
DeepIV-LOG	2.8736 ± 0.0577	2.6227 ± 0.0651	0.0135 ± 0.0215	0.0237 ± 0.0111
DeepIV-GMM	3.7760 ± 0.0316	3.7396 ± 0.0402	0.0194 ± 0.0047	0.0221 ± 0.0041
OneSIV	1.7249 ± 0.3752	1.7411 ± 0.3422	0.0083 ± 0.0191	0.0080 ± 0.0167
DFIV	3.5543 ± 0.0891	3.6218 ± 0.1038	0.0268 ± 0.0005	0.0265 ± 0.0003
DFL	3.2018 ± 0.0496	3.1991 ± 0.0374	0.0624 ± 0.0586	0.0847 ± 0.0049
DirectRep	0.0675 ± 0.0562	0.4600 ± 0.0711	0.0167 ± 0.0171	0.0193 ± 0.0251
CFR	0.0854 ± 0.0579	0.4826 ± 0.0642	0.0115 ± 0.0167	0.0223 ± 0.0176
DRCFR	0.0553 ± 0.0644	0.4336 ± 0.0692	0.0114 ± 0.0221	0.0118 ± 0.0174
CB-IV	**0.0117 ± 0.3882**	**0.1601 ± 0.2499**	**0.0067 ± 0.0271**	**0.0014 ± 0.0249**
Out-of-Sample				
Method	**IHDP-2-6-0**	**IHDP-2-4-2**	**Twins-5-8-0**	**Twins-5-5-3**
DeepIV-LOG	2.8760 ± 0.0553	2.6226 ± 0.0692	0.0140 ± 0.0208	0.0238 ± 0.0111
DeepIV-GMM	3.7768 ± 0.0350	3.7388 ± 0.0416	0.0193 ± 0.0047	0.0221 ± 0.0040
OneSIV	1.7287 ± 0.3725	1.7351 ± 0.3430	0.0082 ± 0.0191	0.0081 ± 0.0168
DFIV	3.5538 ± 0.0904	3.6225 ± 0.1061	0.0268 ± 0.0005	0.0265 ± 0.0003
DFL	3.2038 ± 0.0496	3.1994 ± 0.0376	0.0624 ± 0.0584	0.0846 ± 0.0046
DirectRep	0.0608 ± 0.0817	0.4571 ± 0.0759	0.0162 ± 0.0175	0.0194 ± 0.0253
CFR	0.0785 ± 0.0810	0.4804 ± 0.0687	0.0110 ± 0.0163	0.0225 ± 0.0180
DRCFR	0.0450 ± 0.0953	0.4321 ± 0.0673	0.0113 ± 0.0219	0.0118 ± 0.0174
CB-IV	**0.0150 ± 0.3927**	**0.1578 ± 0.2540**	**0.0065 ± 0.0270**	**0.0015 ± 0.0247**

工具变量回归算法

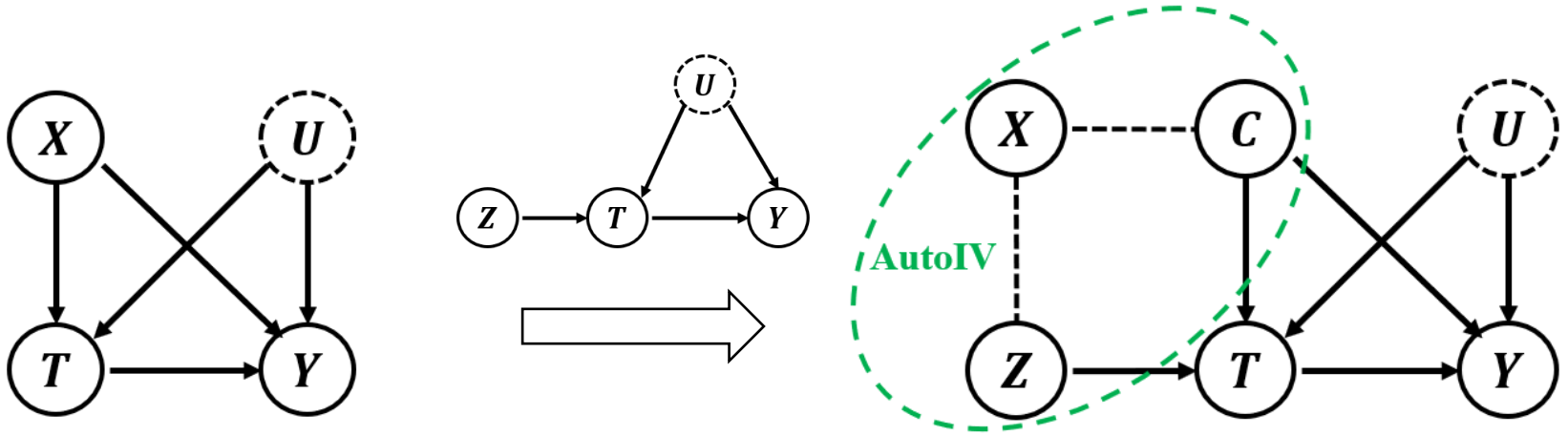
混杂变量平衡方法

混杂平衡+工具变量回归

需要提前定义工具变量

Wu A, Kuang K, Li B, et al. Instrumental Variable Regression with Confounder Balancing, ICML 2022

工具变量自动生成（AutoIV）



工具变量（Z）的条件

相关性：$P(T|Z) \neq P(T)$

排他性：$P(Y|Z,T,U) \neq P(Y|T,U)$

无混杂性：$Z \perp U$

⇦ 互信息约束
解耦表征学习

但是排他性可能不满足

Yuan J, Wu A, Kuang K, et al. Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition[J]. TKDD, 2022.

工具变量自动生成（AutoIV）

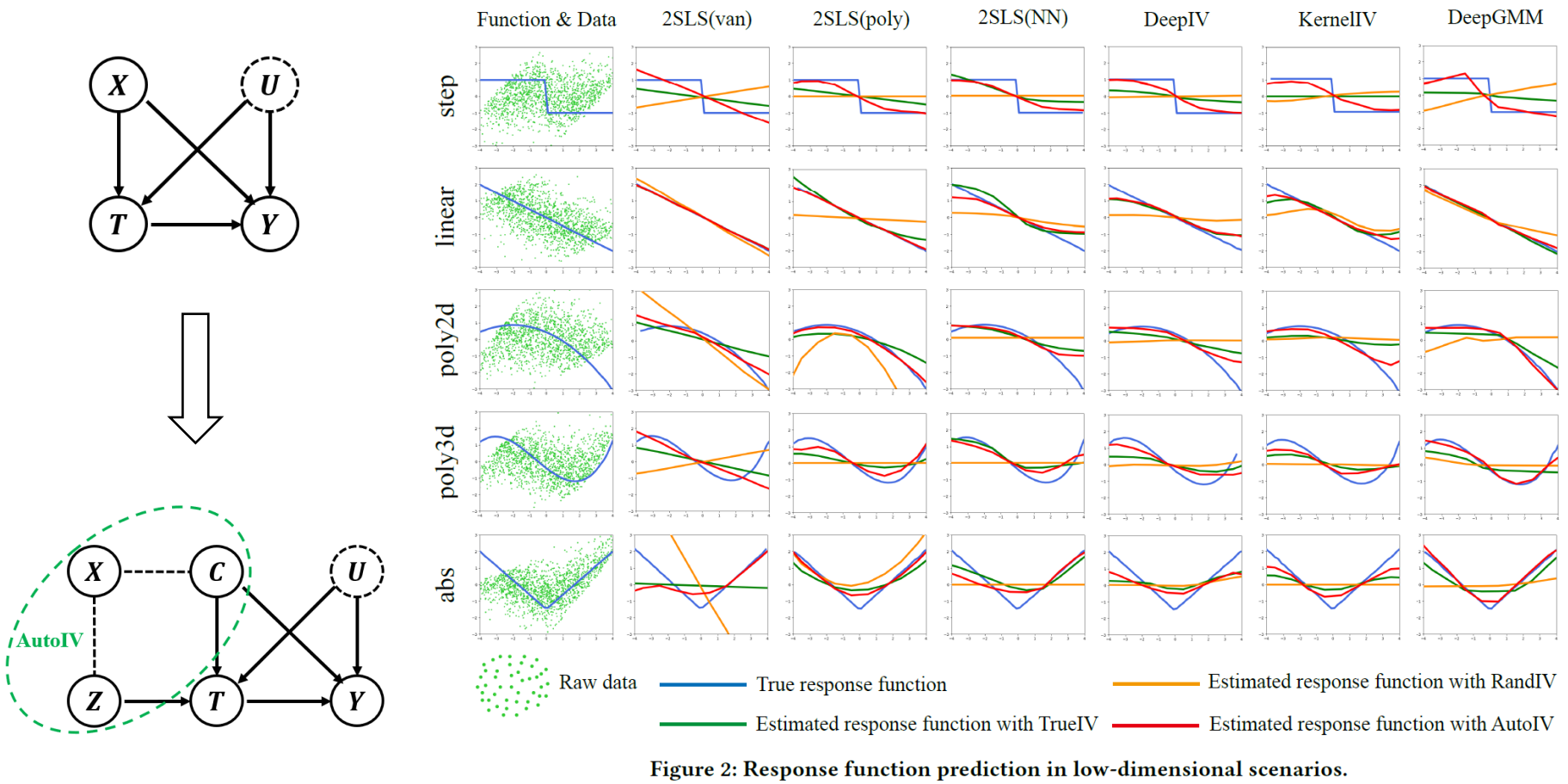


Figure 2: Response function prediction in low-dimensional scenarios.

Yuan J, Wu A, Kuang K, et al. Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition[J]. TKDD, 2022.

综述：因果推理和机器学习中的工具变量方法

Instrumental Variables in Causal Inference and Machine Learning: A Survey

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Fei Wu, *Senior Member, IEEE*

Abstract—Causal inference is the process of using assumptions, study designs, and estimation strategies to draw conclusions about the causal relationships between variables based on data. This allows researchers to better understand the underlying mechanisms at work in complex systems and make more informed decisions. In many settings, we may not fully observe all the confounders that affect both the treatment and outcome variables, complicating the estimation of causal effects. To address this problem, a growing literature in both causal inference and machine learning proposes to use Instrumental Variables (IV). This paper serves as the first effort to systematically and comprehensively introduce and discuss the IV methods and their applications in both causal inference and machine learning. First, we provide the formal definition of IVs and discuss the identification problem of IV regression methods under different assumptions. Second, we categorize the existing work on IV methods into three streams according to the focus on the proposed methods, including two-stage least squares with IVs, control function with IVs, and evaluation of IVs. For each stream, we present both the classical causal inference methods, and recent developments in the machine learning literature. Then, we introduce a variety of applications of IV methods in real-world scenarios and provide a summary of the available datasets and algorithms. Finally, we summarize the literature, discuss the open problems and suggest promising future research directions for IV methods and their applications. We also develop a toolkit of IVs methods reviewed in this survey at https://github.com/causal-machine-learning-lab/mliv.

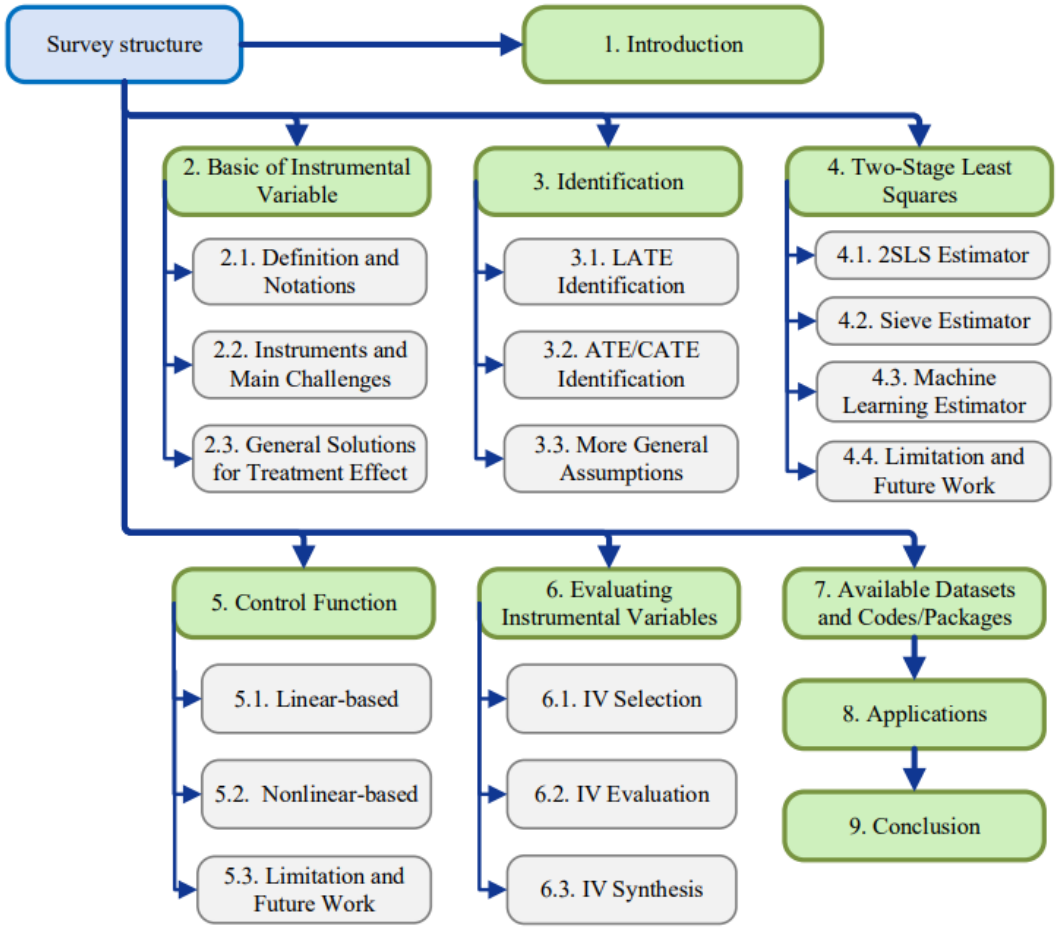


Fig. 2: Outline of the Survey.

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Fei Wu, Instrumental Variables in Causal Inference and Machine Learning: A Survey[J]. ACM Computing Surveys, 2025.

综述：因果推理和机器学习中的工具变量方法

mliv

```
from mliv.dataset.demand import gen_data
from mliv.utils import CausalDataset
gen_data()
data = CausalDataset('./Data/Demand/0.5_1.0_0.0_10000/1/')

from mliv.inference import Vanilla2SLS
from mliv.inference import Poly2SLS
from mliv.inference import NN2SLS
from mliv.inference import OneSIV
from mliv.inference import KernelIV
from mliv.inference import DualIV
from mliv.inference import DFL
from mliv.inference import AGMM
from mliv.inference import DeepGMM
from mliv.inference import DFIV
from mliv.inference import DeepIV           # Tensorflow & keras

for mod in [OneSIV,KernelIV,DualIV,DFL,AGMM,DeepGMM,DFIV,Vanilla2SLS,Poly2SLS,NN2SLS]:
    model = mod()
    model.config['num'] = 100
    model.config['epochs'] = 10
    model.fit(data)

    print(mod)
```

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Fei Wu, Instrumental Variables in Causal Inference and Machine Learning: A Survey[J]. ACM Computing Surveys, 2025.

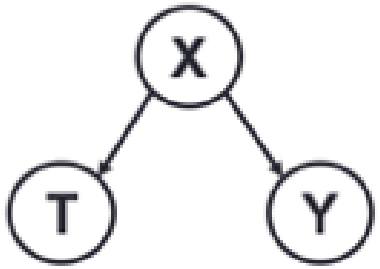
复杂偏差下因果推断

2000诺贝尔经济学奖：
面向选择偏差的因果推断

因果

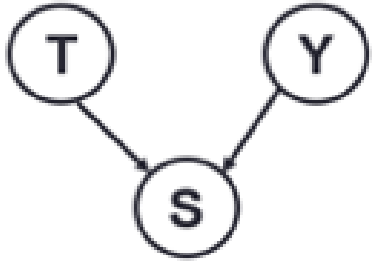
可解释
稳定/鲁棒
可决策

混淆偏差



虚假关联: T is correlated with Y ignoring X

选择偏差



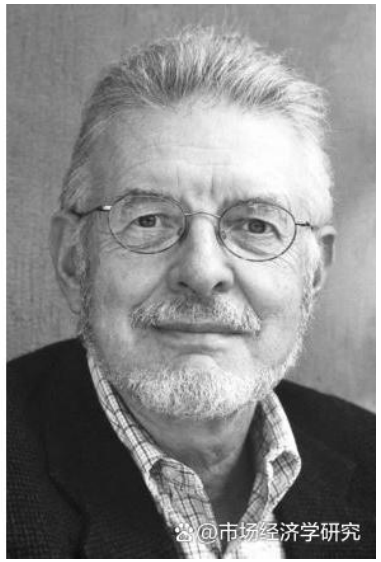
虚假相关: T is correlated with Y given S

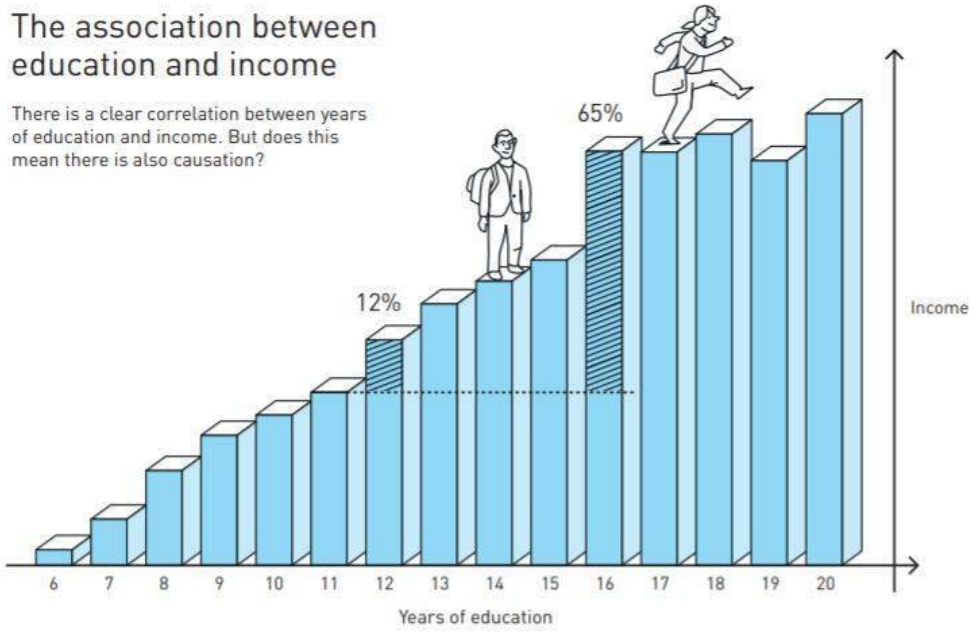
选择偏差下因果推断

选择性样本的理论与方法

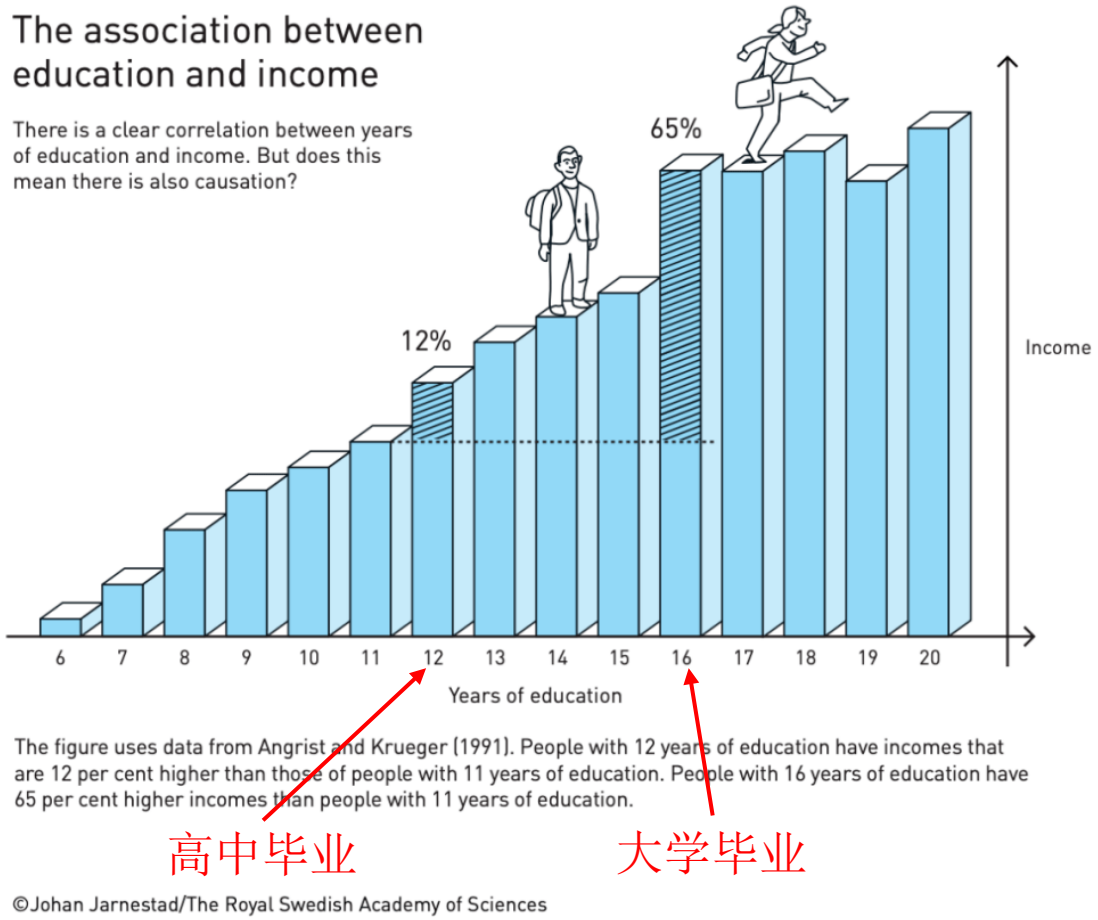
2000年诺贝尔经济学奖授予美国经济学家詹姆斯·赫克曼（James Heckman，生于1944年）和丹尼尔·麦克法登（Daniel McFadden，生于1937年），以表彰前者“发展了分析选择性样本的理论和方法”和后者“发展了分析离散选择的理论和方法”。

抽样问题是经济计量学中的基本问题。如果一个样本不能随机地代表其总体，则抽样偏差就可能发生。可获得的微观数据是选择性样本，而不是随机样本。例如，对工资数据进行统计分析时，如果选择工作或接受教育的个人具有某种研究者未观察到的特征，而这种特征在统计时未予以考虑，统计评估就可能产生偏差。



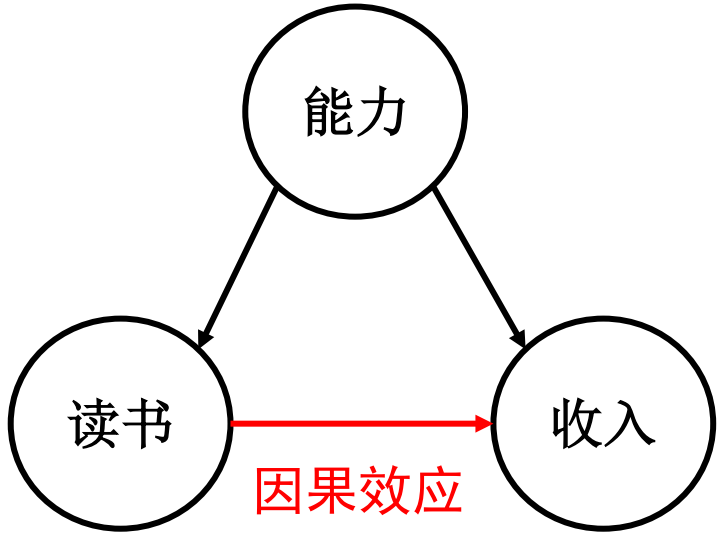


读书是否有用?



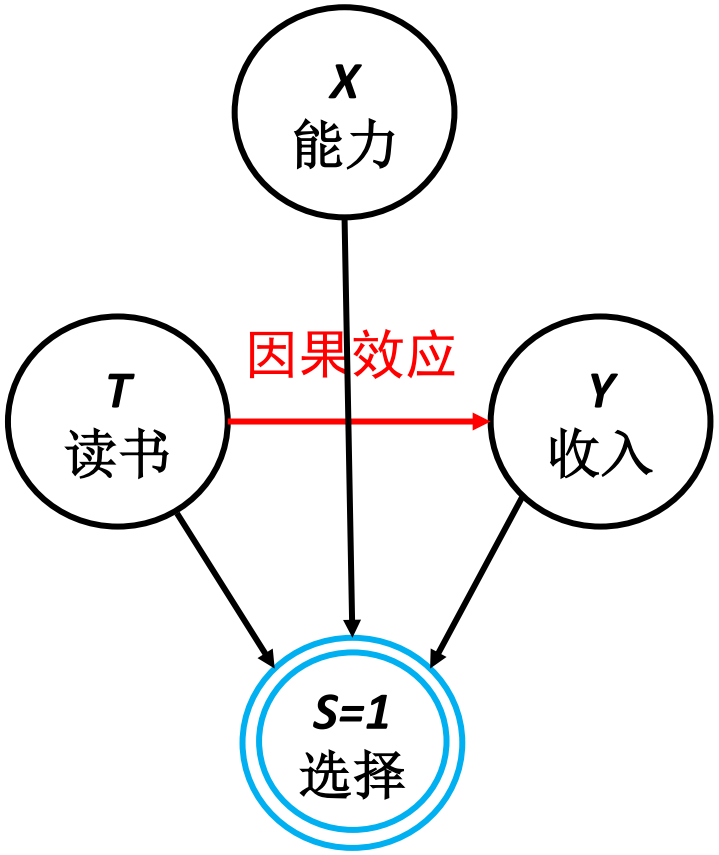
- 高中毕业 v.s. 高中肄业： 12%
- 大学毕业 v.s. 高中毕业： 53%

- 高出来的收入是否是“因为”多读了几年书?



直接做随机对照实验

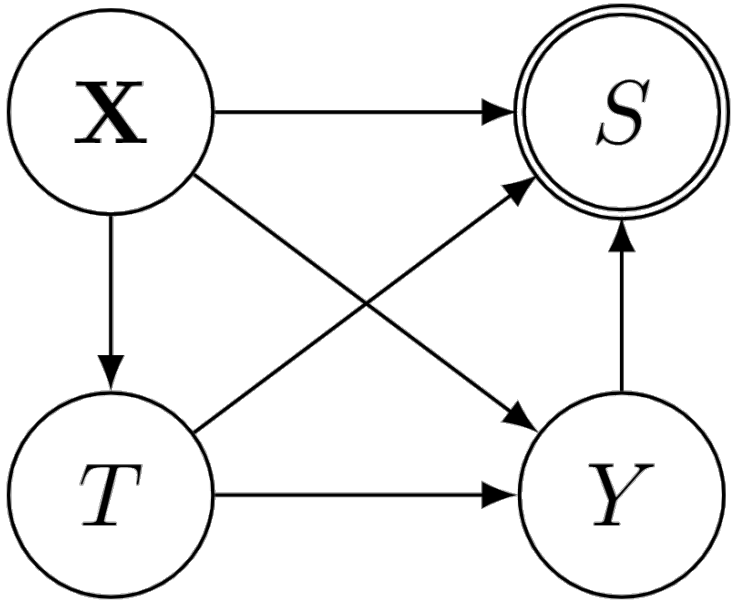
评估教育对收入的影响



- 分析读书与收入之间的因果效应
  - 随机对照实验：做实验、发问卷
  - 干预变量 T=1/0：是否高中毕业
  - 结果变量 Y：收入

- 实际应用中，RCT也可能存在选择偏差
  - 如高中毕业但收入低的受访者，不愿意回答
  - 而高中未毕业但收入高的，非常愿意回答
  - 选择变量 S，由 T 和 Y （甚至 X）同时影响
    - S=1：X，T，Y
    - S=0：X，T

碰撞偏差（Collider Bias）



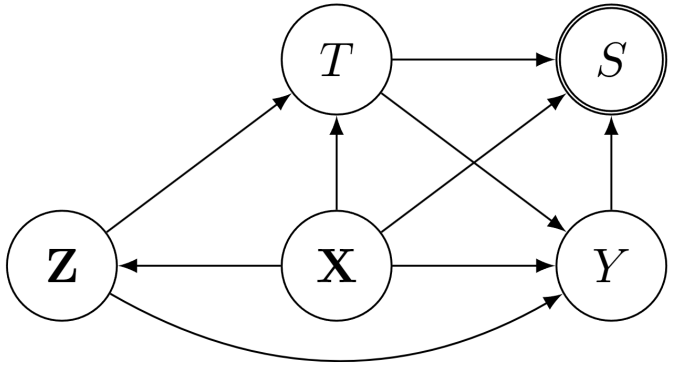
- 问题：读书与收入之间的因果效应
  - 混杂变量 X：能力
  - 干预变量 T=1/0：是否高中毕业
  - 结果变量 Y：收入
  - 选择变量 S：由 T, Y, X同时影响
    - S=1：X，T，Y
    - S=0：X，T

- 挑战：

✓分布偏移：$P(\mathbf{X}, T, Y, S = 1) \neq P(\mathbf{X}, T, Y)$

✓不可识别：存在碰撞偏差，因果效应不可识别

影子变量（Shadow Variable）



- 影子变量（$\boldsymbol{Z}$）的条件
- $\boldsymbol{Z} \perp\!\!\!\perp S|\mathbf{X}, Y, T$
- $\boldsymbol{Z} \not\perp\!\!\!\perp Y|\mathbf{X}, T, S = 1$

- 给定影子变量（$\mathbf{Z}$），则 $f(Y|\mathbf{X}, \mathbf{Z}, T, S = 0)$ 可识别

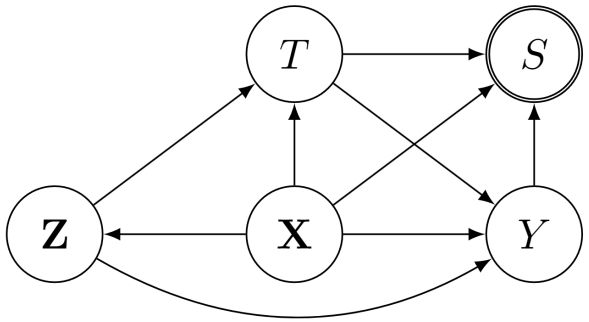
$$f(Y|\mathbf{X}, \mathbf{Z}, T, S = 0) = OR(\mathbf{X}, T, Y) \cdot f(Y|\mathbf{X}, \mathbf{Z}, T, S = 1)/E[OR(\mathbf{X}, T, Y)|\mathbf{X}, \mathbf{Z}, T, S = 1]$$

- 其中 $OR(\mathbf{X}, T, Y) := \frac{f(S=0|\mathbf{X},T,Y)\cdot f(S=1|\mathbf{X},T,Y=0)}{f(S=0|\mathbf{X},T,Y=0)\cdot f(S=1|\mathbf{X},T,Y)}$，可通过以下过程计算
  - $OR(\mathbf{X}, T, Y) = \widetilde{OR}(\mathbf{X}, T, Y)/\widetilde{OR}(\mathbf{X}, T, Y = 0)$
  - $E\left[\widetilde{OR}(\mathbf{X}, T, Y)\middle|\mathbf{X}, \mathbf{Z}, T, S = 1\right] = f(\mathbf{Z}|\mathbf{X}, T, S = 1)/f(\mathbf{Z}|\mathbf{X}, T, S = 0)$
  - 当且仅当$\mathbf{Z}$比Y有更大的 support set 时$\widetilde{OR}(\mathbf{X}, T, Y)$有唯一解

需要提前定义好影子变量

Miao W, Tchetgen Tchetgen E J. On varieties of doubly robust estimators under missingness not at random with a shadow variable[J]. Biometrika, 2016, 103(2): 475-482.

Learning Shadow Variable Representation



- 影子变量（$\boldsymbol{Z}$）的条件
- $\boldsymbol{Z} \perp\!\!\!\perp S|\mathbf{X}, Y, T$
- $\boldsymbol{Z} \not\perp\!\!\!\perp Y|\mathbf{X}, T, S = 1$

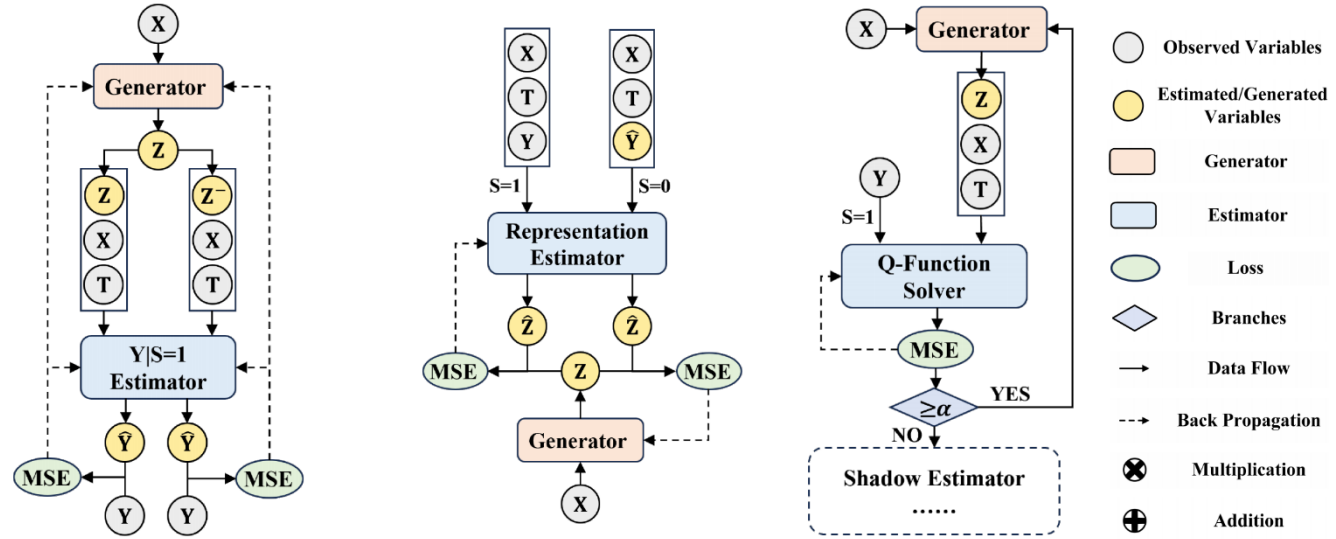
- 从混杂变量 $\mathbf{X}$ 中学习影子变量 $\mathbf{Z}$ 的表征，约束其满足以下条件：
  - $\boldsymbol{Z} \not\perp\!\!\!\perp Y|\mathbf{X}, T, S = 1$：只用观测变量即可显式约束
  - $\mathbf{Z} \perp\!\!\!\perp S|\mathbf{X}, Y, T$：$Y$ 在 $S = 0$ 时缺失，无法显式约束
- 假设 $\mathbf{Z} \perp\!\!\!\perp S|\mathbf{X}, Y, T$ 成立
  - 当且仅当 $E[S/Q(\mathbf{X}, T, Y)\,|\mathbf{X}, \mathbf{Z}, T] = 1$ 不存在 $(0,1]$ 之间的解时，可拒绝该假设
  - 其中 $Q(\mathbf{X}, T, Y)$ 是一个未知函数

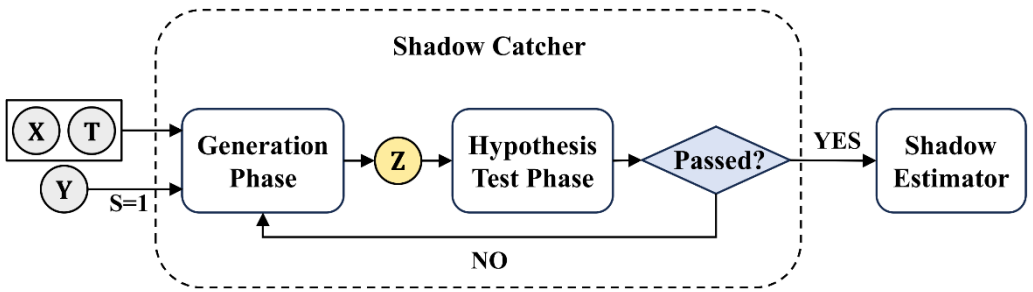
学习影子变量表征 → 假设检验确保条件满足 → 基于影子变量的因果推断

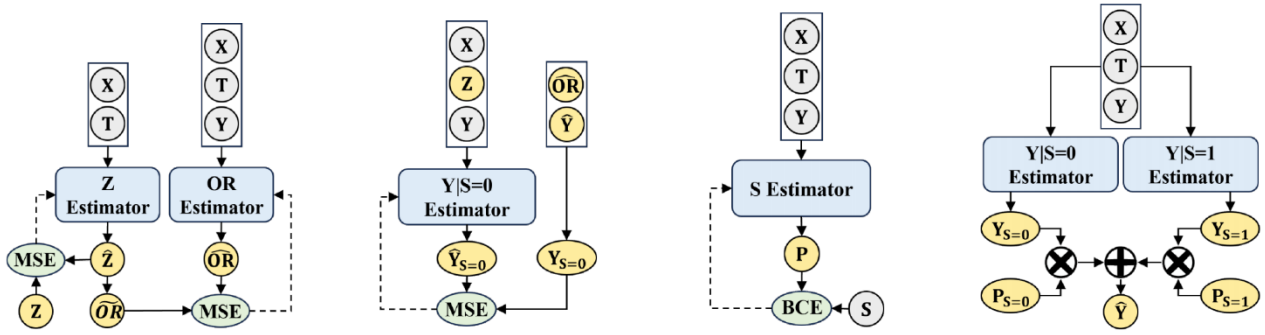
Baohong Li, Haoxuan Li, Ruoxuan Xiong, Anpeng Wu, Fei Wu, **Kun Kuang***. Learning Shadow Variable Representation for Treatment Effect Estimation under Collider Bias, ICML, 2024.

ShadowCatcher & ShadowEstimator

- ShadowCatcher:
  - 表征学习阶段
  - 假设检验阶段
- ShadowEstimator
  - 估计$OR(\mathbf{X}, T, Y)$
  - 估计$f(Y|\mathbf{X}, \mathbf{Z}, T, S=0)$







Baohong Li, Haoxuan Li, Ruoxuan Xiong, Anpeng Wu, Fei Wu, **Kun Kuang***. Learning Shadow Variable Representation for Treatment Effect Estimation under Collider Bias, ICML, 2024.

实验分析和结果

	$\beta = 1$		$\beta = 3$		$\beta = 5$	
ESTIMATOR	SELECTED DATA	UNSELECTED DATA	SELECTED DATA	UNSELECTED DATA	SELECTED DATA	UNSELECTED DATA
HECKIT	0.323±0.065	0.330±0.046	0.340±0.055	0.352±0.042	0.349±0.069	0.413±0.048
DR	0.298±0.032	0.316±0.042	0.331±0.048	0.357±0.053	0.367±0.033	0.448±0.017
IPSW	0.328±0.048	0.348±0.049	0.328±0.031	0.353±0.034	0.465±0.011	0.545±0.014
BNN	0.290±0.011	0.306±0.012	0.329±0.048	0.354±0.033	0.359±0.011	0.439±0.015
TARNET	0.295±0.012	0.312±0.011	0.329±0.030	0.357±0.053	0.366±0.071	0.436±0.087
CFR	0.290±0.009	0.307±0.008	0.324±0.009	0.350±0.013	0.359±0.008	0.436±0.030
CFOREST	0.310±0.030	0.331±0.038	0.338±0.019	0.368±0.022	0.373±0.026	0.453±0.043
DR-CFR	0.284±0.038	0.307±0.040	0.340±0.055	0.355±0.064	0.366±0.051	0.435±0.060
TEDVAE	0.281±0.056	0.419±0.070	0.378±0.063	0.420±0.059	0.394±0.054	0.431±0.067
DER-CFR	0.291±0.010	0.309±0.014	0.323±0.015	0.348±0.017	0.358±0.011	0.439±0.013
DESCN	0.295±0.002	0.312±0.002	0.326±0.003	0.357±0.004	0.365±0.003	0.449±0.011
ES-CFR	0.289±0.003	0.305±0.004	0.331±0.003	0.359±0.003	0.369±0.003	0.448±0.005
OURS	**0.227±0.001**	**0.229±0.001**	**0.249±0.013**	**0.255±0.021**	**0.299±0.008**	**0.300±0.008**

	IHDP ($\sqrt{\text{PEHE}}$)		**ACIC 2016** ($\sqrt{\text{PEHE}}$)		**JOBS** ($\hat{R}_{\text{Pol}}$)	
ESTIMATOR	**WITHIN-SAMPLE**	**OUT-OF-SAMPLE**	**WITHIN-SAMPLE**	**OUT-OF-SAMPLE**	**WITHIN-SAMPLE**	**OUT-OF-SAMPLE**
HECKIT	1.587±0.065	1.621±0.041	3.106±0.444	3.340±0.111	0.328±0.050	0.331±0.052
DR	1.355±0.123	1.572±0.205	2.346±0.129	2.653±0.222	0.316±0.007	0.317±0.036
IPSW	2.118±0.344	2.129±0.295	4.244±0.145	5.411±0.073	0.284±0.051	0.289±0.063
BNN	1.308±0.298	1.457±0.339	2.173±0.150	2.586±0.486	0.303±0.025	0.304±0.041
TARNET	1.240±0.158	1.416±0.154	2.275±0.756	2.805±0.766	0.315±0.012	0.316±0.050
CFR	1.283±0.186	1.401±0.238	2.107±0.297	2.361±0.587	0.313±0.018	0.314±0.072
CFOREST	1.702±0.292	1.948±0.429	4.137±0.295	4.605±0.137	0.326±0.012	0.326±0.059
DR-CFR	1.299±0.087	1.399±0.171	2.240±0.691	2.340±0.663	0.322±0.022	0.323±0.099
TEDVAE	4.246±0.394	4.347±0.563	3.501±0.708	4.468±0.813	0.296±0.046	0.300±0.031
DER-CFR	1.446±0.345	1.571±0.371	2.214±0.204	2.246±0.598	0.309±0.023	0.311±0.029
DESCN	1.193±0.057	1.665±0.246	2.185±0.150	2.306±0.236	0.331±0.010	0.331±0.051
ES-CFR	1.499±0.096	1.436±0.095	3.875±0.224	4.494±0.214	0.290±0.045	0.293±0.046
OURS	**0.703±0.106**	**0.723±0.102**	**1.911±0.126**	**2.047±0.351**	**0.279±0.017**	**0.280±0.018**

复杂偏差下因果推断

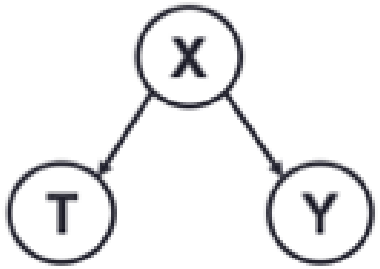
2021诺贝尔经济学奖：
基于工具变量的因果推断

2000诺贝尔经济学奖：
面向选择偏差的因果推断

因果

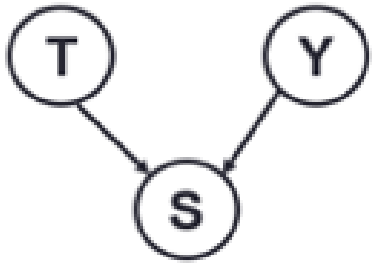
可解释
稳定/鲁棒
可决策

混淆偏差



虚假关联: T is correlated with Y ignoring X

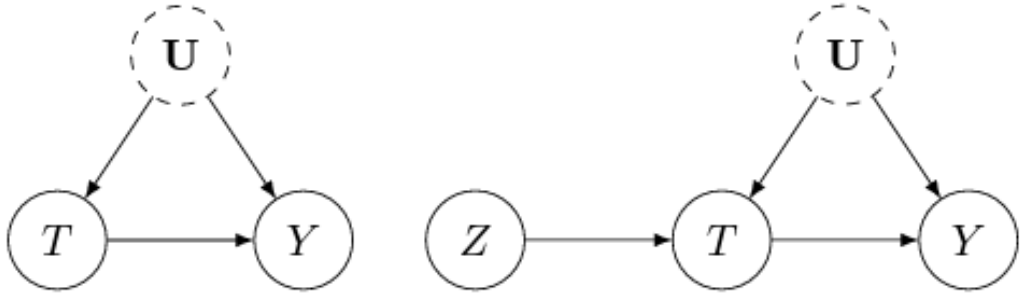
选择偏差



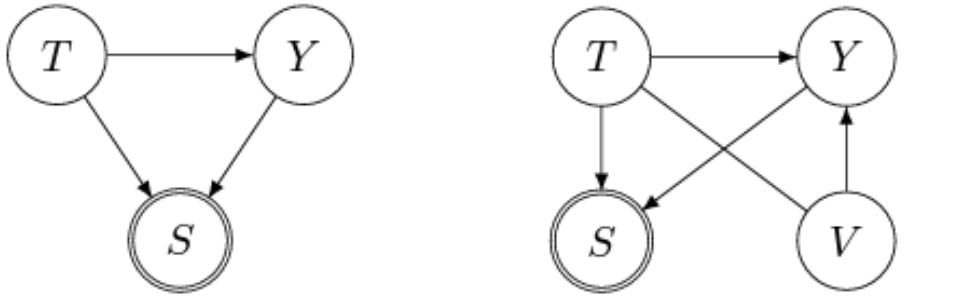
虚假相关: T is correlated with Y given S

同时存在混淆偏差和选择偏差，如何推断？

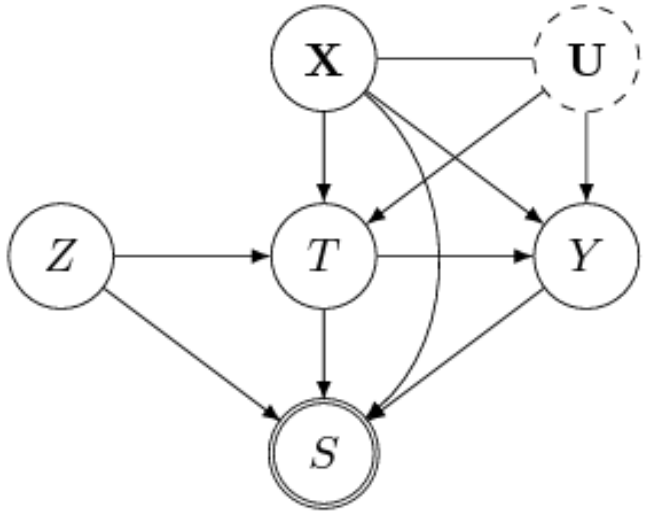
复杂偏差下因果推断



(a) Latent confounding. (b) Instrumental variable approach.



(a) Collider bias. (b) Shadow variable approach.

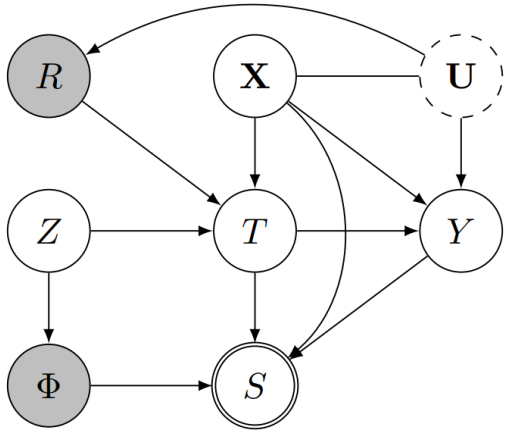


- 如何同时解决两种偏差：
  - IV is all you need

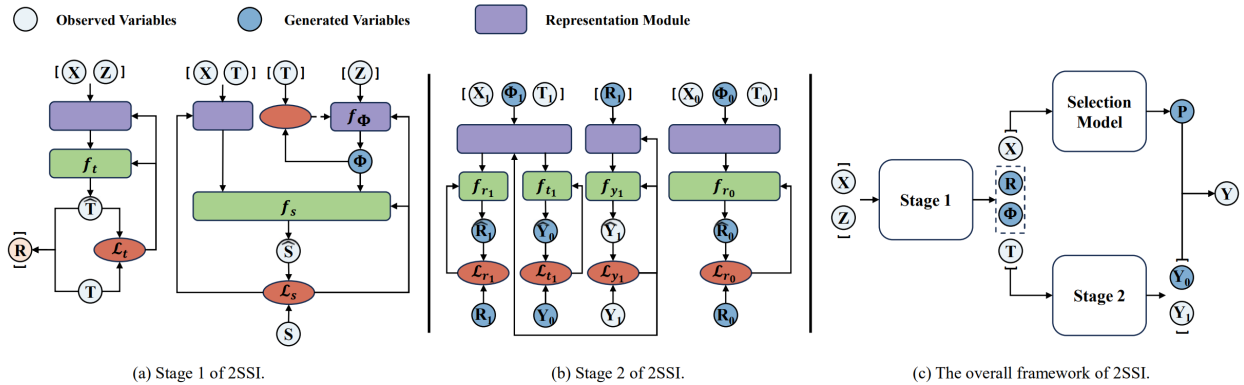
Baohong Li, Anpeng Wu, Ruoxuan Xiong, **Kun Kuang***. Two-Stage Shadow Inclusion Estimation: An IV Approach for Causal Inference under Latent Confounding and Collider Bias, ICML, 2024.

2SSI：两阶段Shadow Variable回归

- 第一阶段
  - 用S=1和S=0的X和Z回归T，根据T和$\hat{T}$求残差R
  - 用S=1和S=0的X, T和Z回归S，同时学习解耦表征φ
    - 用一个表征网络学习$\varphi = f(Z)$预测S，并约束其与T独立
      - 二值T时采用积分概率度量IPM，连续T时用互信息



- 第二阶段
  - 引入φ，将R视为φ的条件下的Shadow Variable
    - R在X,T,Y和φ的条件下与S独立，在X,T,S=1和φ的条件下与Y相关



- 影子变量（$\boldsymbol{Z}$）的条件
- $\boldsymbol{Z} \perp\!\!\!\perp S|\mathbf{X}, Y, T$
- $\boldsymbol{Z} \not\perp\!\!\!\perp Y|\mathbf{X}, T, S = 1$

Baohong Li, Anpeng Wu, Ruoxuan Xiong, **Kun Kuang***. Two-Stage Shadow Inclusion Estimation: An IV Approach for Causal Inference under Latent Confounding and Collider Bias, ICML, 2024.

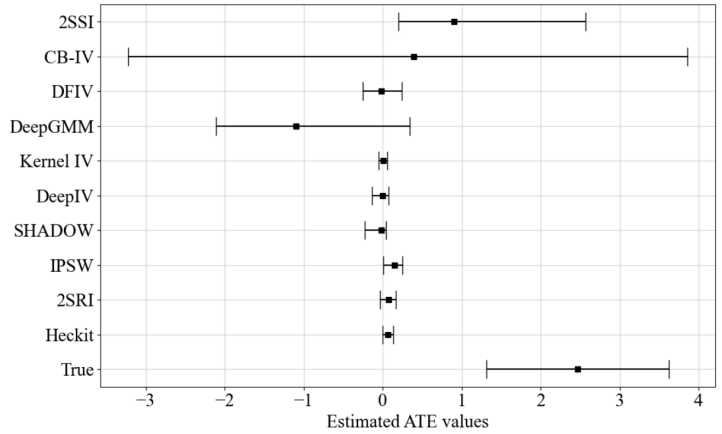
实验结果

- 在合成数据集上的实验结果

Estimator	$\alpha = 5$		$\alpha = 10$		$\alpha = 15$	
	$S = 1$ data	$S = 0$ data	$S = 1$ data	$S = 0$ data	$S = 1$ data	$S = 0$ data
Heckit	0.648±0.038	5.685±0.324	0.724±0.060	6.741±0.359	0.714±0.057	6.763±0.507
2SRI	0.998±0.070	6.076±0.262	1.052±0.097	6.673±0.219	1.016±0.063	6.799±0.363
IPSW	1.075±0.082	12.72±1.070	1.063±0.093	12.39±0.656	1.056±0.085	12.45±0.442
SHADOW	0.852±0.083	4.619±0.269	0.956±0.116	5.233±0.289	0.957±0.088	5.258±0.449
DeepIV	0.630±0.048	12.06±0.832	0.633±0.051	12.46±0.424	0.641±0.053	12.57±1.007
Kernel IV	0.297±0.097	6.450±0.367	0.400±0.175	6.995±0.526	0.428±0.109	7.220±0.903
DeepGMM	0.655±0.105	8.138±1.268	0.688±0.108	8.399±1.555	0.699±0.104	9.163±1.419
DFIV	0.572±0.090	12.72±0.686	0.580±0.106	12.97±1.667	0.620±0.052	13.60±1.181
CB-IV	1.202±0.228	7.910±0.482	1.270±0.225	8.442±0.574	1.343±0.153	8.771±0.712
2SSI	**0.147±0.020**	**1.320±0.335**	**0.154±0.016**	**1.278±0.155**	**0.155±0.024**	**1.275±0.227**

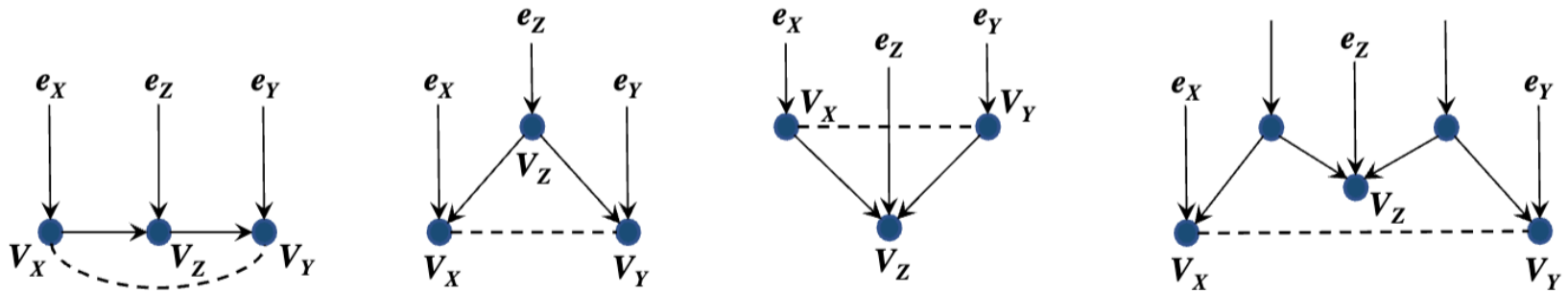
Estimator	$\beta = 5$		$\beta = 10$		$\beta = 15$	
	$S = 1$ data	$S = 0$ data	$S = 1$ data	$S = 0$ data	$S = 1$ data	$S = 0$ data
Heckit	0.089±0.007	6.008±0.230	0.724±0.060	6.741±0.359	1.819±0.122	9.490±0.778
2SRI	0.114±0.011	3.317±0.155	1.052±0.097	6.673±0.219	2.918±0.251	10.88±0.452
IPSW	0.136±0.035	5.588±0.777	1.063±0.093	12.39±0.656	3.150±0.396	24.65±2.918
SHADOW	0.099±0.009	3.220±0.149	0.956±0.116	5.233±0.289	2.512±0.260	7.350±0.520
DeepIV	0.065±0.007	4.697±0.278	0.633±0.051	12.46±0.424	1.945±0.117	24.97±1.596
Kernel IV	0.052±0.015	3.734±0.182	0.400±0.175	6.995±0.526	1.759±0.747	12.07±0.611
DeepGMM	0.040±0.009	3.893±0.443	0.688±0.108	8.399±1.555	2.445±0.396	13.59±2.546
DFIV	0.072±0.008	5.056±0.435	0.580±0.106	12.97±1.667	1.809±0.302	28.27±1.978
CB-IV	0.099±0.021	3.737±0.276	1.270±0.225	8.442±0.574	4.031±0.412	14.46±0.615
2SSI	**0.038±0.020**	**1.995±0.088**	**0.154±0.016**	**1.278±0.155**	**0.522±0.411**	**2.655±0.630**

- 在真实数据集上的ATE估计结果



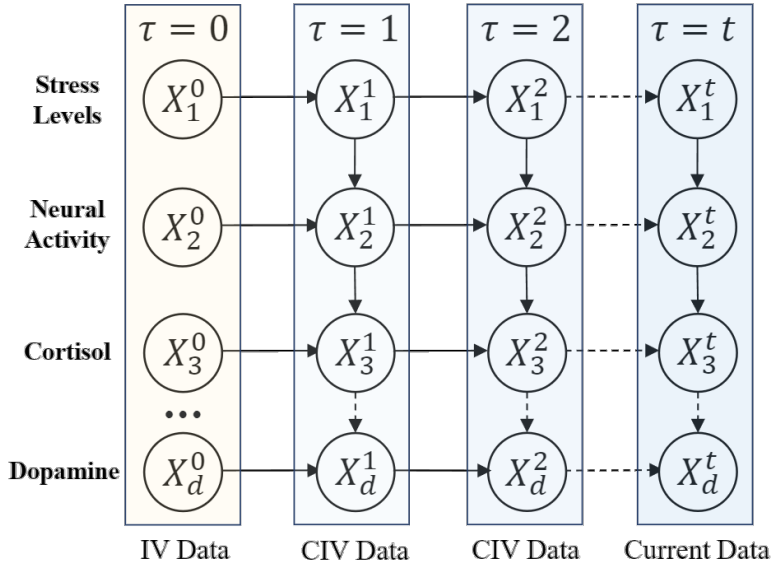
Baohong Li, Anpeng Wu, Ruoxuan Xiong, **Kun Kuang***. Two-Stage Shadow Inclusion Estimation: An IV Approach for Causal Inference under Latent Confounding and Collider Bias, ICML, 2024.

复杂偏差下因果推断与因果图学习



(a) Chain Structure. (b) Fork Structure. (c) Collider Structure. (d) M Structure.

Figure 1: Underlying bias present in unknown causal relation.

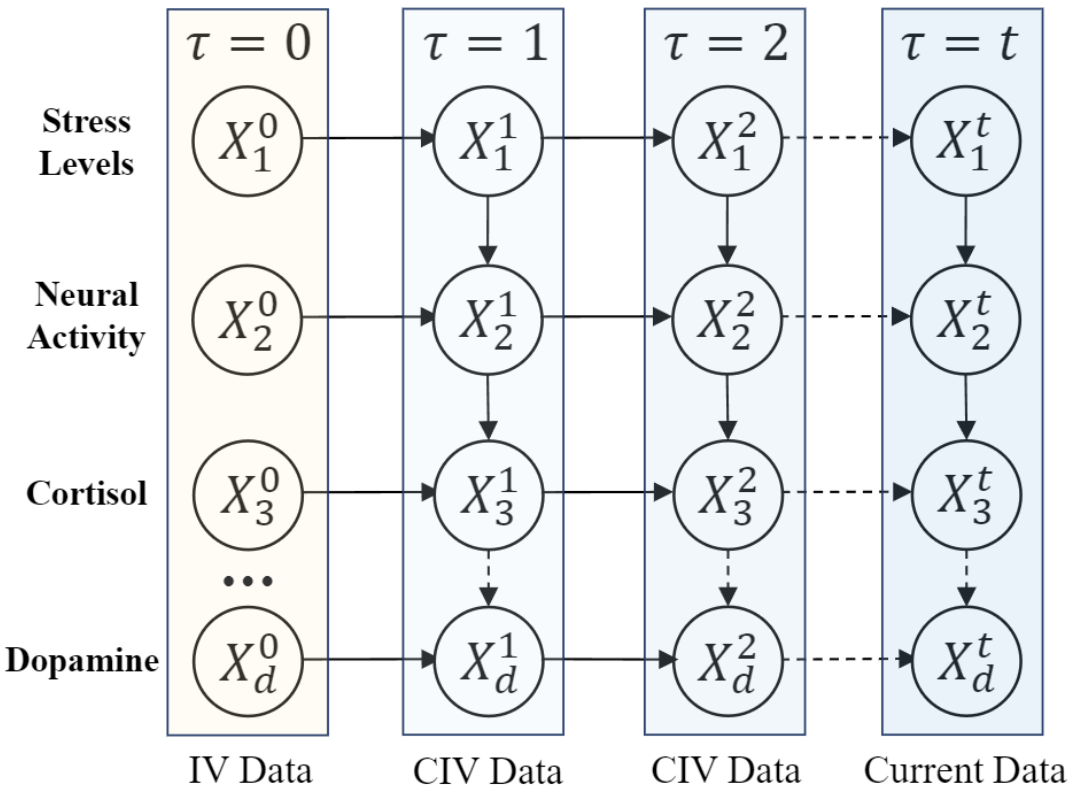


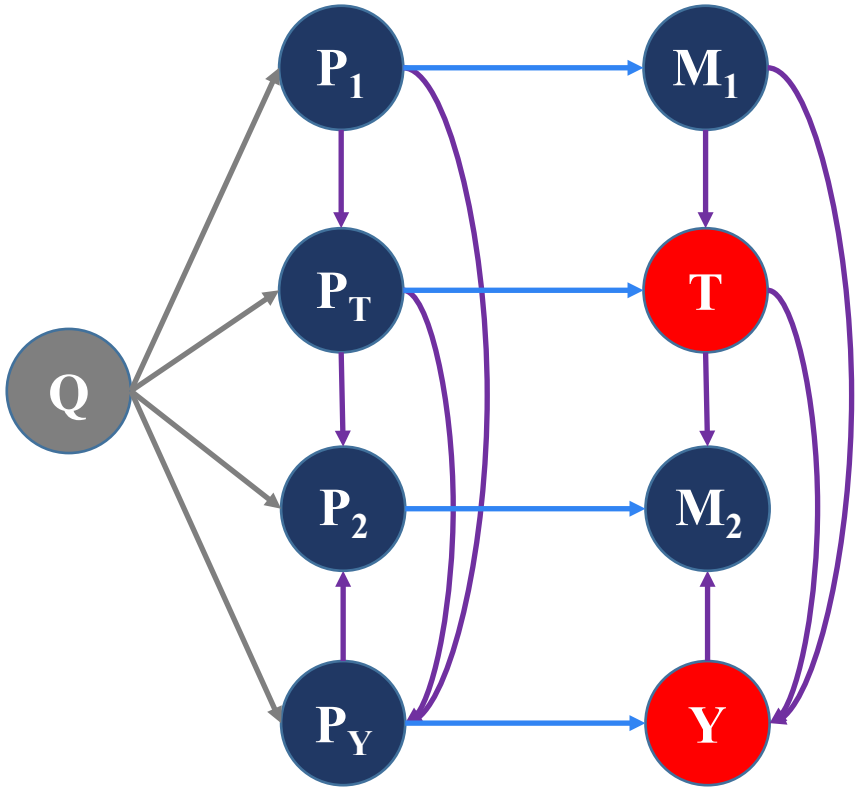
Assumption 1. SUTVA.
Assumption 2. Unconfoundedness.
Assumption 3. Positive Assumption.
Assumption 4. Markov Assumption.
Assumption 5. Temporal Dependency.

Assumption 6. Invariant Causal Mechanism. At different timestamps, the causal relationship between variables does not change over time, and the parent node of a variable in the previous timestamp remains its parent node in the next timestamp.

Anpeng Wu, Haoxuan Li, Kun Kuang*, Zhang Keli, Fei Wu. Learning Causal Relations from Subsampled Time Series with Two Time-Slices, ICML, 2024.

复杂偏差下因果推断与因果图学习





M_1 导致
混淆偏差

M_2 导致
内生选择偏差

基于前序时间变量P_1, P_2, P_Y，　P_T为T和Y的条件工具变量

Anpeng Wu, Haoxuan Li, Kun Kuang*, Zhang Keli, Fei Wu. Learning Causal Relations from Subsampled Time Series with Two Time-Slices, ICML, 2024.

复杂偏差下因果推断

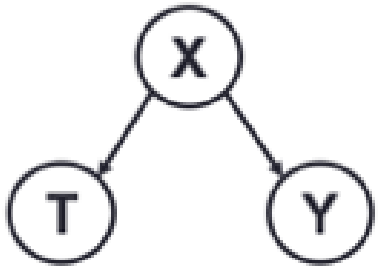
2021诺贝尔经济学奖：
基于工具变量的因果推断

2000诺贝尔经济学奖：
面向选择偏差的因果推断

因果

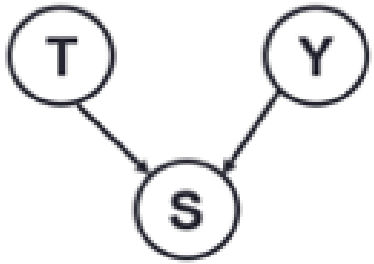
可解释
稳定/鲁棒
可决策

混淆偏差



虚假关联: T is correlated with Y ignoring X

选择偏差



虚假相关: T is correlated with Y given S

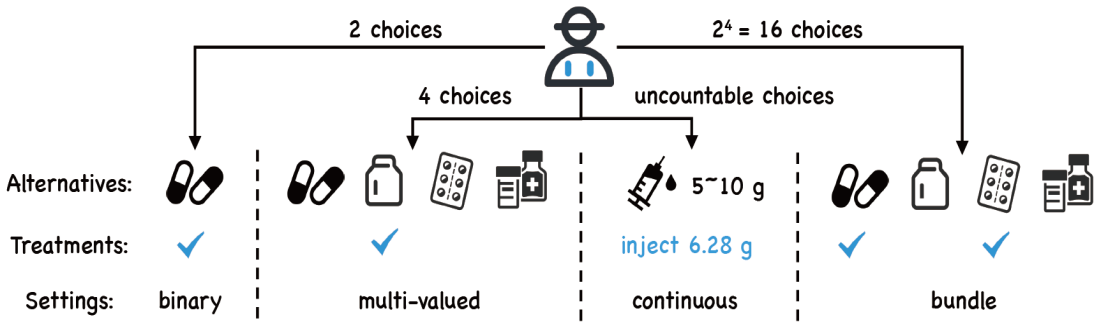
同时存在混淆偏差和选择偏差，（条件）工具变量

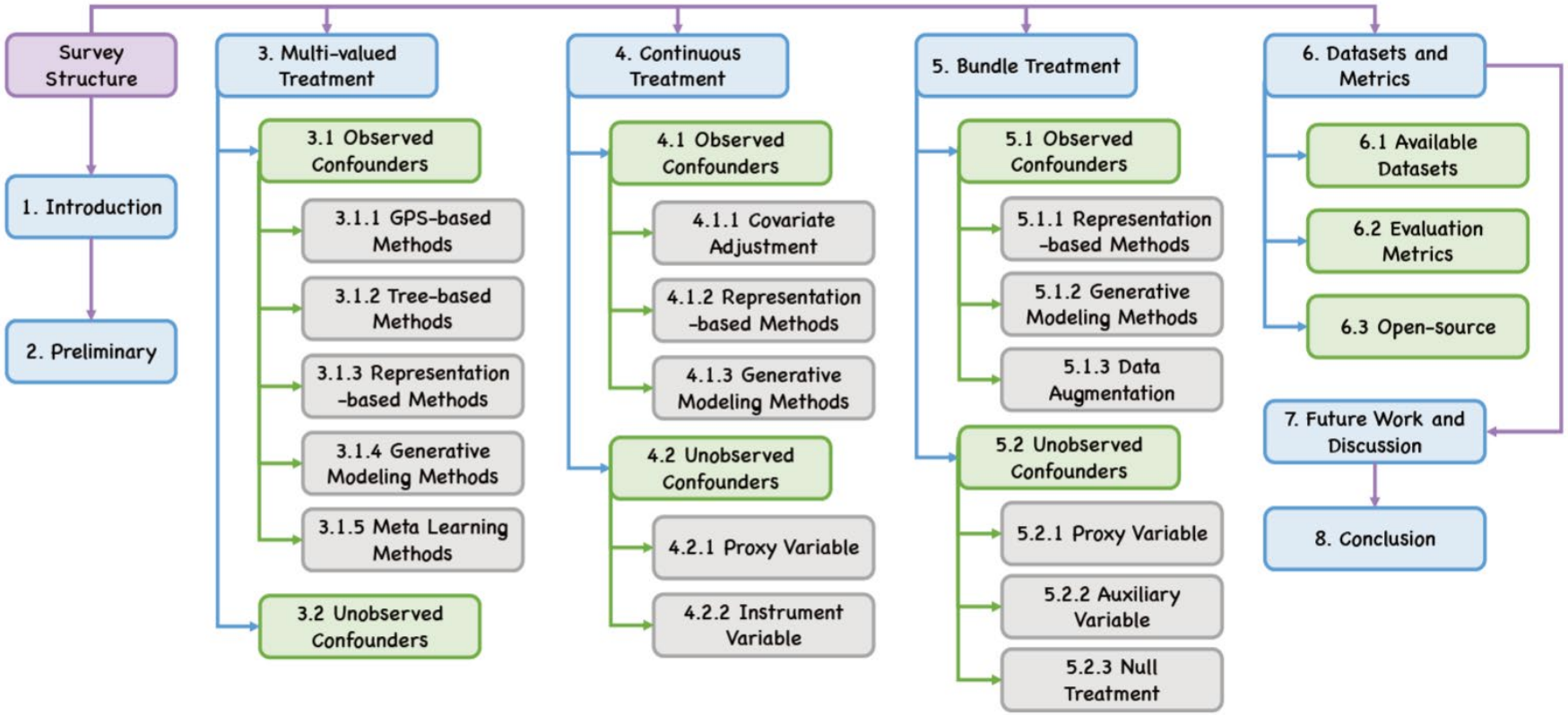
Continuous/Complex treatment effect estimation

Causal Inference with Complex Treatments: A Survey

YINGRONG WANG, Zhejiang University, China
HAOXUAN LI, Peking University, China
KUN KUANG, Zhejiang University, China

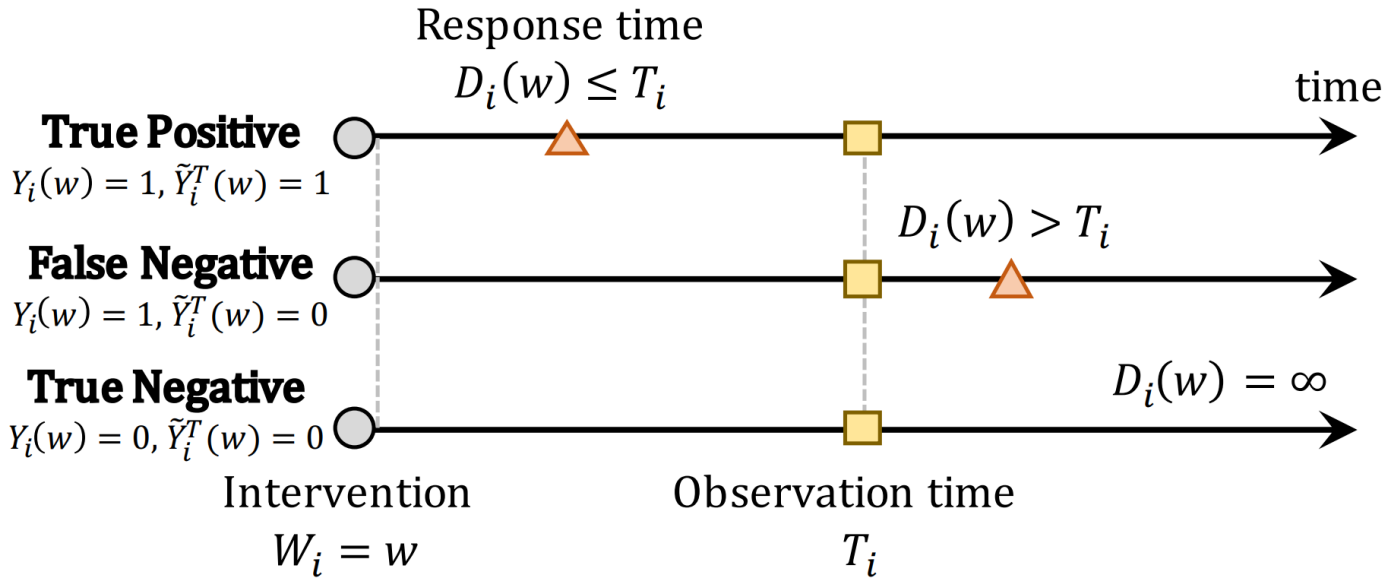
Causal inference plays an important role in ex
of a treatment. It has extensive applications of
marketing, health care, and education for a lo
of binary treatment that there is only one tre
can be much more complex in practice, such
attempt to introduce the causal inference metho
comprehensive manner. At first, we formally lo
and their possible variations on special condit
under the three treatment settings. In each situat
methods conforming to the unconfoundedness
is conducted according to their detailed implem
codes, together with evaluation metrics that are
provide a brief summary about these works and





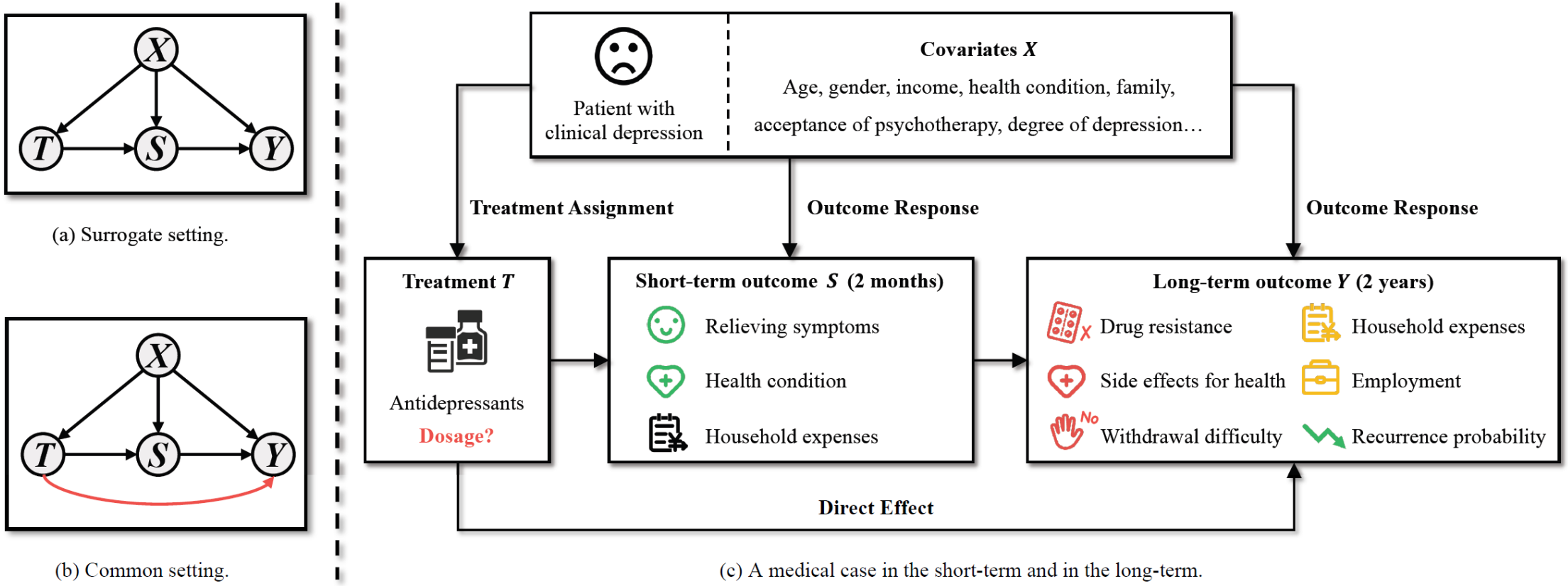
Some working papers

- Causal Inference with Delayed Response



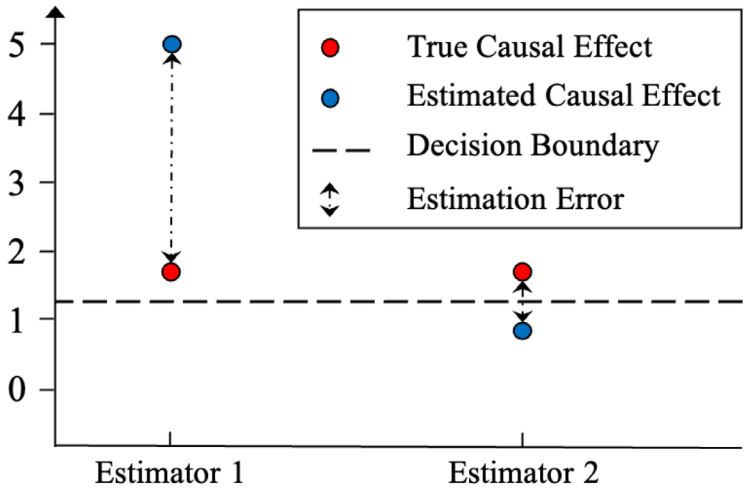
Some working papers

- Short-term and Long-term Treatment Effects



Some working papers

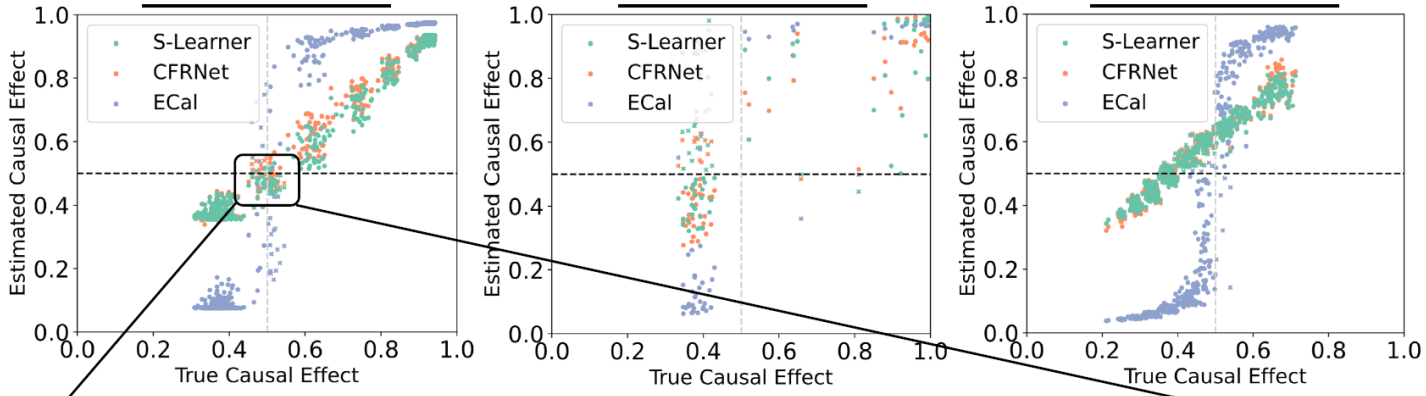
- Precise causal effect == accurate decision?



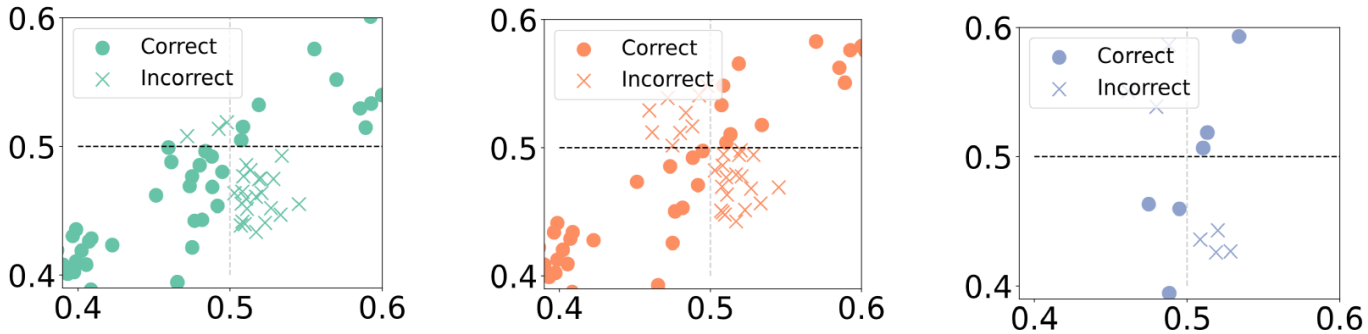
Method	Error Rate
S-Learner	13.33%
CFRNet	12.33%
ECal	**4.500%**

Method	Error Rate
S-Learner	48.00%
CFRNet	37.33%
ECal	**20.00%**

Method	Error Rate
S-Learner	28.50%
CFRNet	29.13%
ECal	**2.180%**



(a) Synthetic Data (b) IHDP Data (c) Jobs Data



(d) S-Learner (e) CFRNet (f) ECal

浙江大学
ZHEJIANG UNIVERSITY

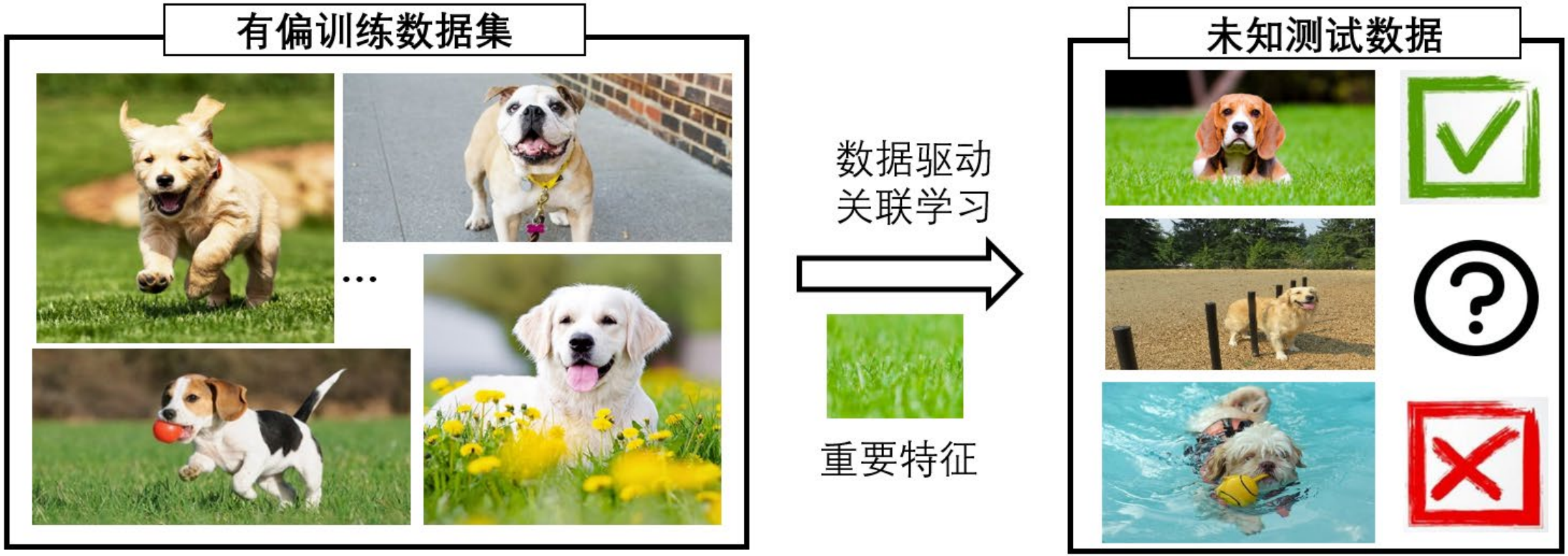
因果启发的稳定可泛化学习

况琨

浙江大学计算机学院

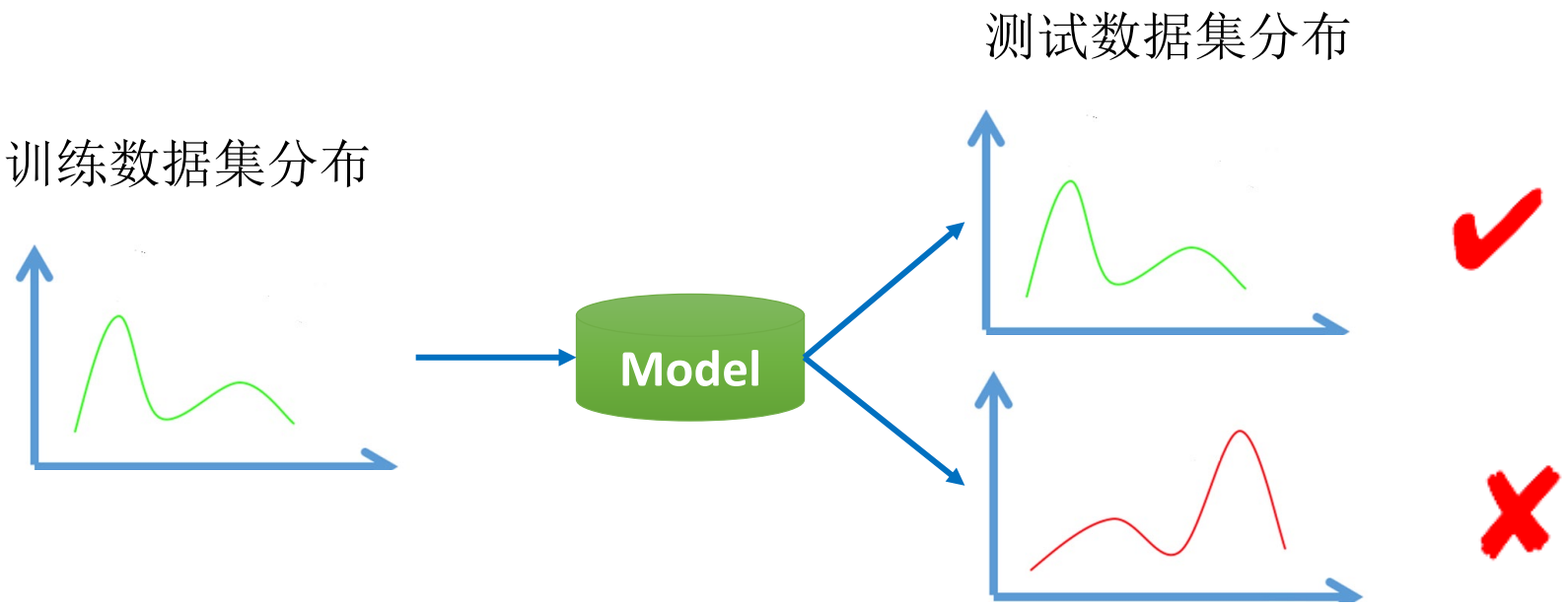
因果赋能机器学习—稳定预测

- 预测模型的能和不能



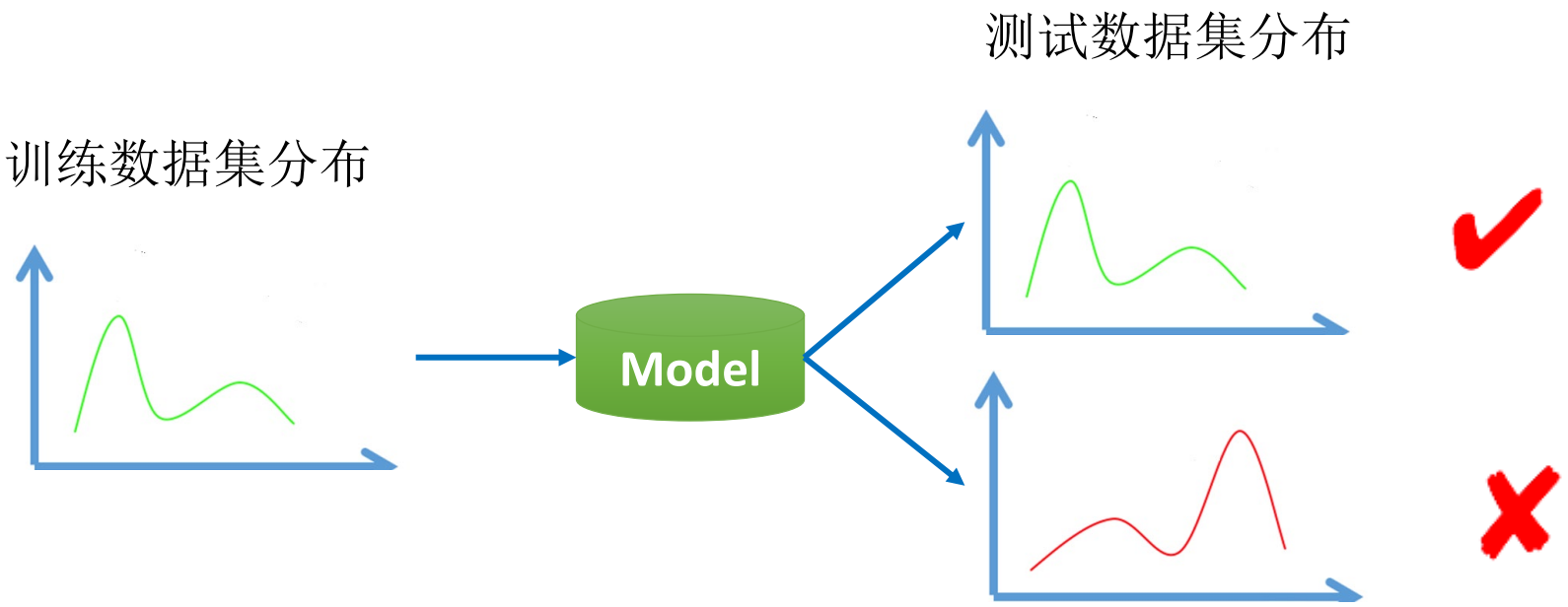
为什么会失败?

- 数据的问题:
  - 独立同分布假设（但实际应用中常常无法满足）
  - 样本选择偏差导致数据分布发生偏移（训练数据和测试数据分布不一致）
  - 小样本场景下，该问题更严重
  - 但，我们无法控制测试数据的产生



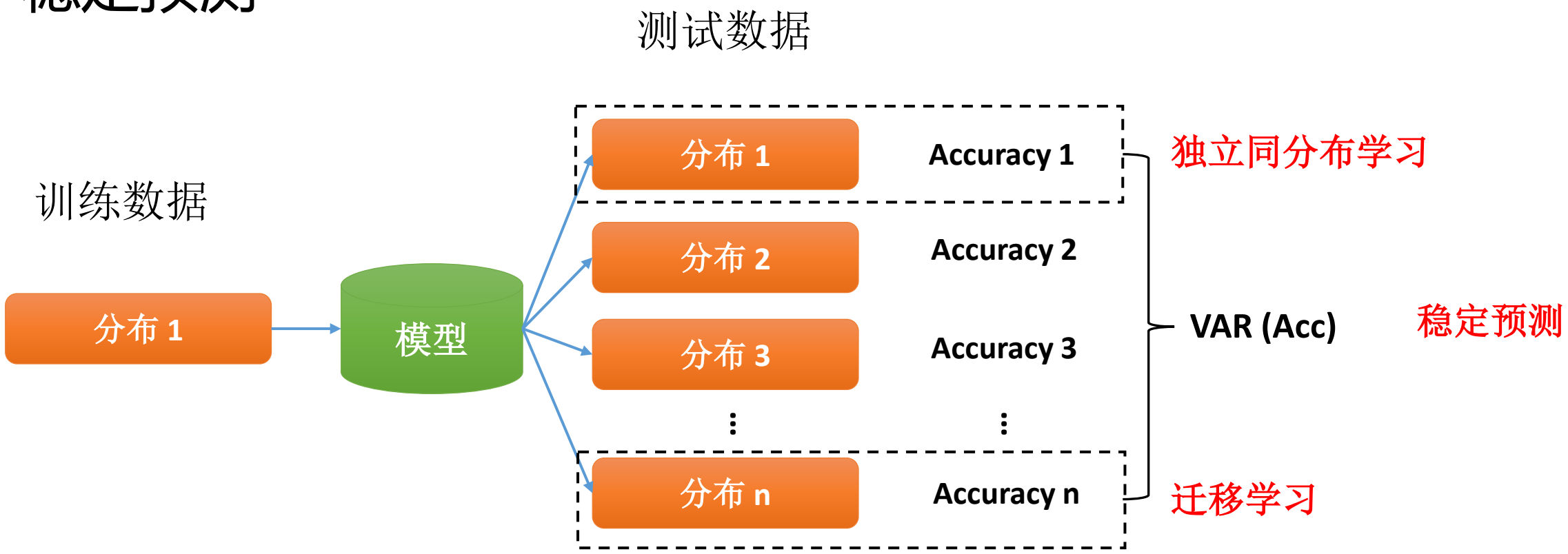
为什么会失败?

- 模型的问题:
  - 关联驱动学习
  - 关联的三种来源：因果，混杂偏差，选择偏差（因果关联和虚假关联）
  - 想法：从复杂关联中甄别因果关联，因果约束实现稳定预测



稳定预测

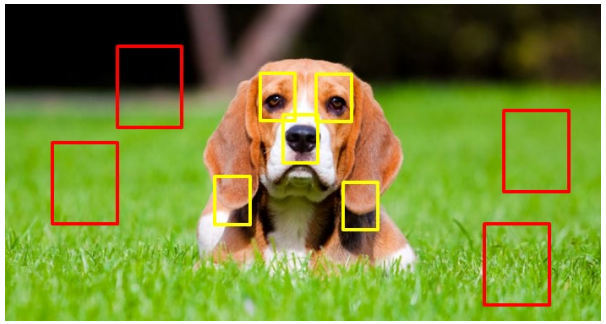
- 稳定预测



面向未知测试数据分布的稳定预测

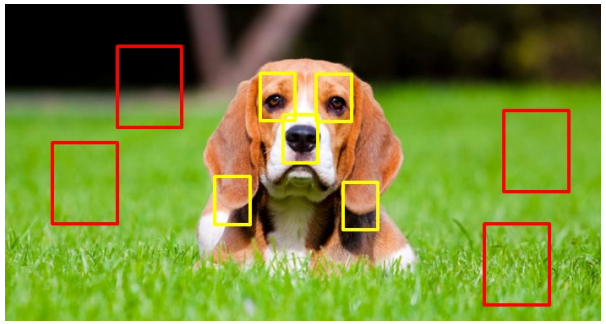
预测模型为什么不稳定

- 预测 / 分类
  - X: 特征向量； $Y = \{0,1\}$
  - 环境：X 和 Y 的联合分布，表示为 $P(XY)$
- 假设 $X = \{S, V\}$，以及 $Y = f(S) + \varepsilon$
  - S：因果特征集合（如狗的眼睛，耳朵等）
  - V：非因果特征集合（如草地等背景）
  - $P(Y|S)$ 是稳定不变的，但 $P(Y|V)$ 会随着环境的变化为变化
- V 和 Y 并不独立（由于样本选择偏差，存在虚假相关）
  - 关联学习驱动的方法
  - 一些非因果特征 $\boldsymbol{v} \subseteq \boldsymbol{V}$ 会被学习成很重要的预测特征



因果如何实现稳定预测

- 结果变量产生机制：
  - $Y = f(S) + \varepsilon,\ X = \{S, V\}$
- 因果变量 S 和非因果变量 V 的区别：
  - S 对 Y 有因果效应，
  - 但，V 对 Y 没有因果效应。

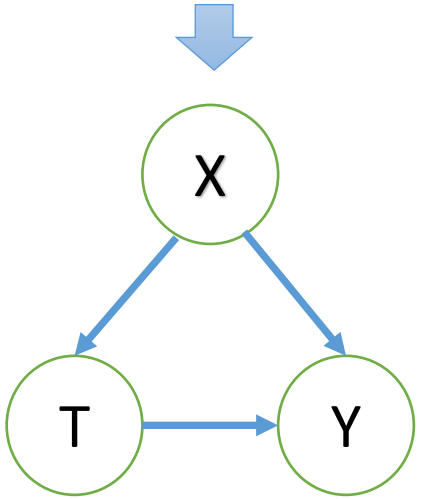


- **想法：** 恢复 X 和 Y 之间的因果关联，因为只有 S 和 Y 是因果关联的，这样就可以使得 $V \perp Y$。

因果如何实现稳定预测

- **丢弃虚假关联，拥抱因果关联。**

典型的关联分析框架



典型的因果推断框架

评估变量 *T* 和结果 *Y* 之间的关联关系和效应时，并没有考虑 *X* 和 *T* 之间的关系

评估变量 *T* 和结果 *Y* 之间的因果关系和效应时，需要平衡 *X* 的分布 (A/B 测试)

因果约束项及理论保证

- **因果约束项（近似全局平衡法）**
  - 学习全局的样本权重 W，使得变量集合 X 中的任意两个变量都相互独立

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{:,-j}^{T} \cdot (W \odot \mathbf{X}_{:,j})}{W^{T} \cdot \mathbf{X}_{:,j}} - \frac{\mathbf{X}_{:,-j}^{T} \cdot (W \odot (1-\mathbf{X}_{:,j}))}{W^{T} \cdot (1-\mathbf{X}_{:,j})} \right\|_{2}^{2}, \quad (4)$$

↓

0

PROPOSITION 3.3. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

因果约束的逻辑斯蒂回归

- 全局平衡回归算法（GBR）

$$\min \quad \sum_{i=1}^{n} W_i \cdot \log(1+\exp((1-2Y_i)\cdot(\mathbf{X}_i\beta))), \tag{5}$$

$$s.t. \quad \sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^T \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2 \leq \lambda_1, \quad W \geq 0,$$

$$\|W\|_2^2 \leq \lambda_2, \quad \|\beta\|_2^2 \leq \lambda_3, \quad \|\beta\|_1 \leq \lambda_4, \quad (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_5$$

样本加权的逻辑斯蒂损失函数

因果约束项

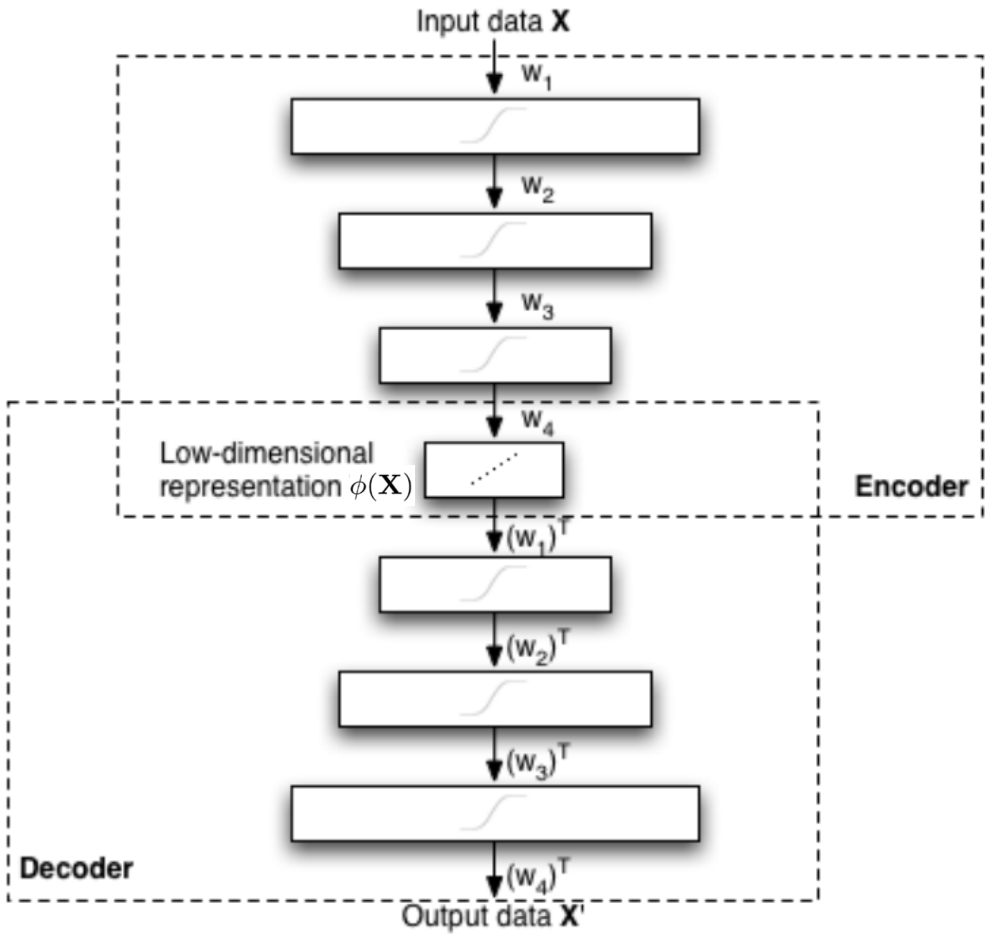
因果回归系数

- 因果回归系数：可解释且稳定
- 但，局限于线性模型

Shen Z, Cui P, Kuang K, et al. Causally regularized learning with agnostic data selection bias[C]//MM. 2018: 411-419.
Kuang K, Cui P, Athey S, et al. Stable prediction across unknown environments[C]//KDD. 2018: 1617-1626.

大数据时代的新挑战

- 数据变量维度高
  - 成千上百的变量
  - 降维
- 非线性
  - 预测变量与结果变量通常是非线性关系
  - 非线性方程拟合
- 深度自编码器



因果约束的深度学习网络

- 深度全局平衡算法（DGBR）

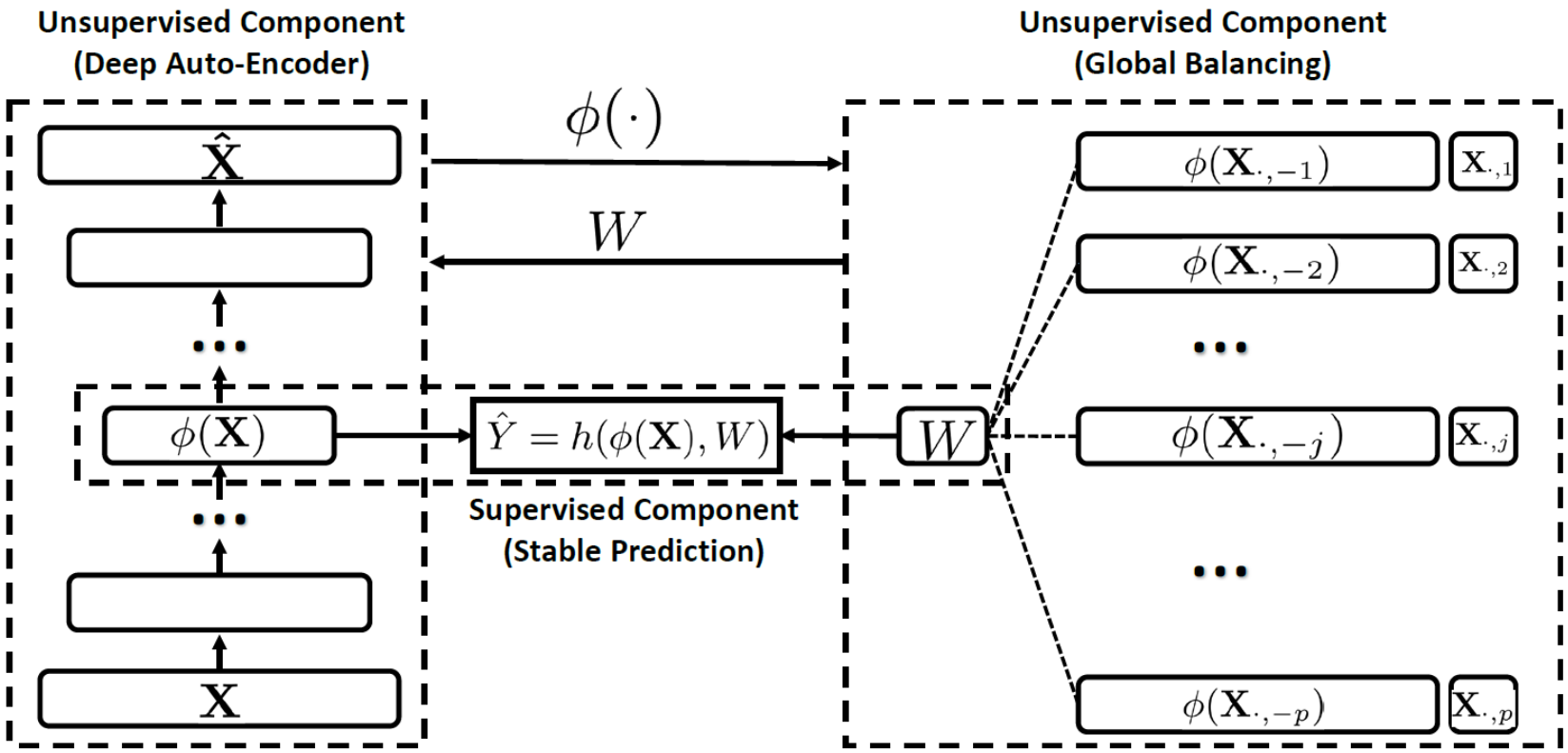
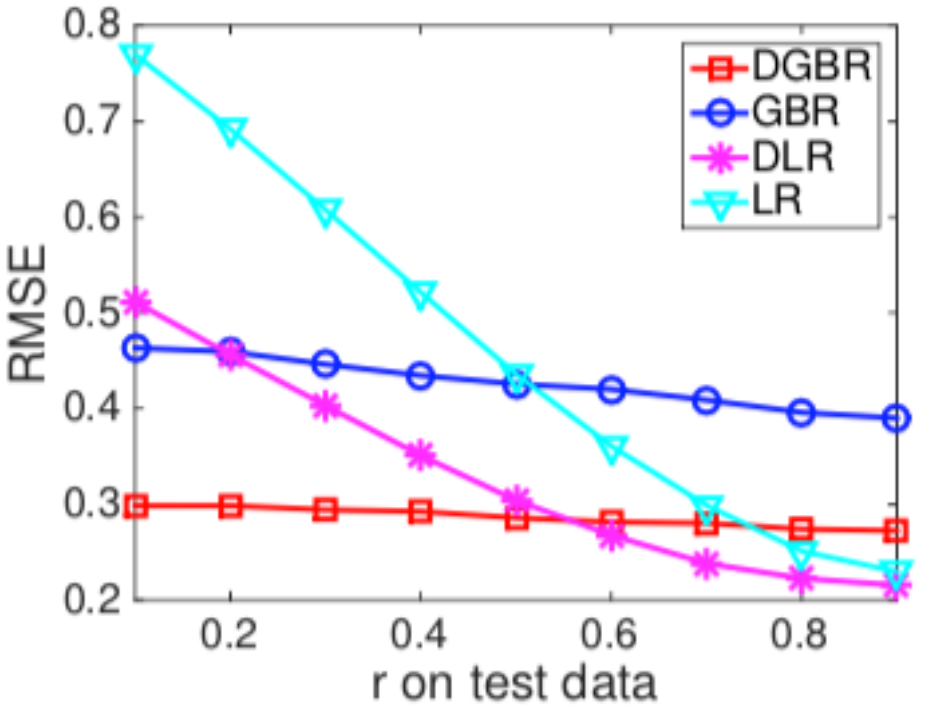


Figure 2: The framework of our proposed DGBR model.

实验结果及分析

- 仿真实验数据设定：
- 不同的 r 表示不同的数据分布
  - 训练数据集： $r = 0.85$
  - 测试数据集： $r = \{0.1,\ldots,0.9\}$

- LR： 逻辑斯蒂回归
- DLR： 深度逻辑斯蒂回归
- DBR： 全局平衡算法（我们）
- DGBR： 深度全局平衡算法（我们）



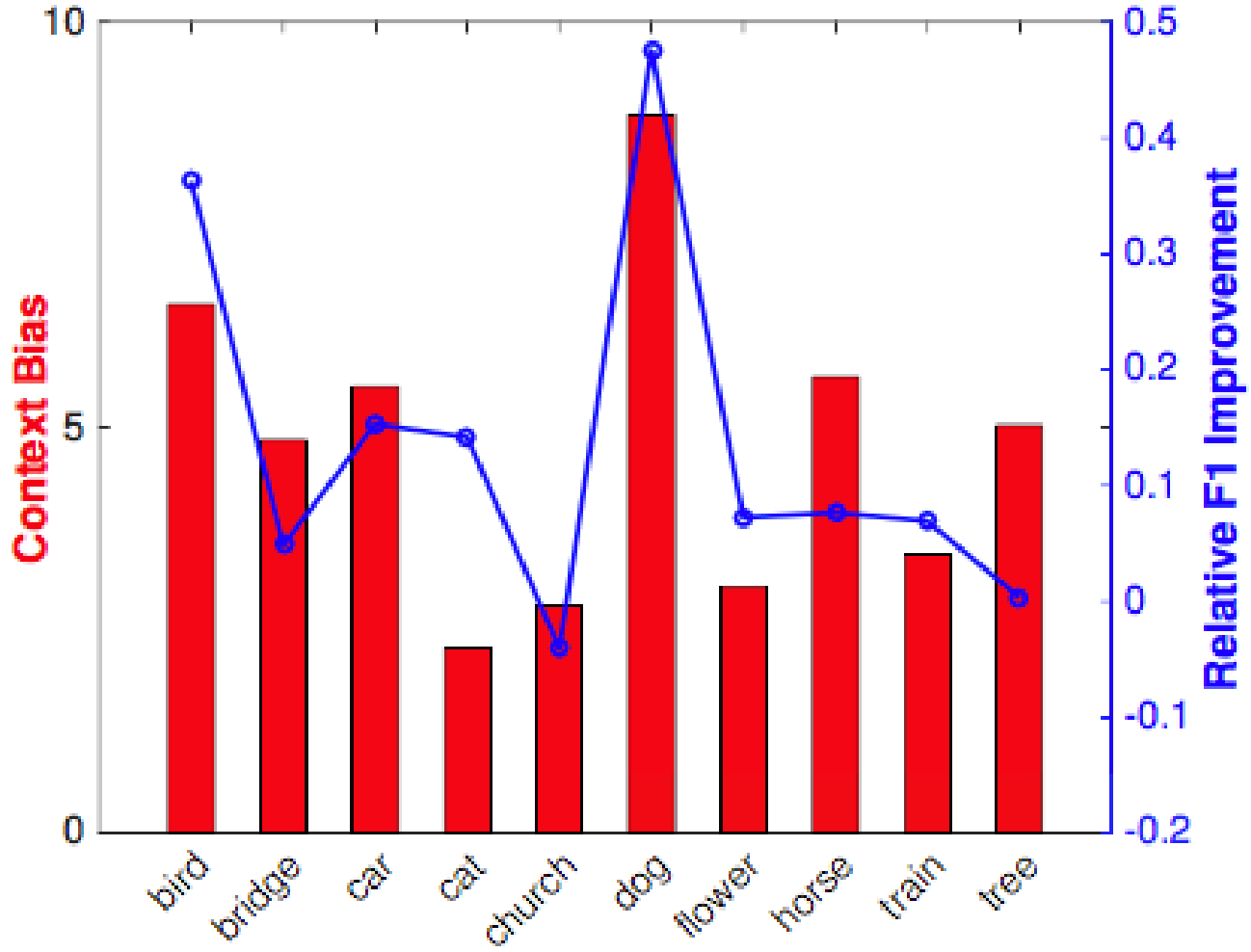
(f) Trained on $n = 2000$, $p = 20$, $r = 0.85$

Shen Z, Cui P, Kuang K, et al. Causally regularized learning with agnostic data selection bias[C]//MM. 2018: 411-419.
Kuang K, Cui P, Athey S, et al. Stable prediction across unknown environments[C]//KDD. 2018: 1617-1626.

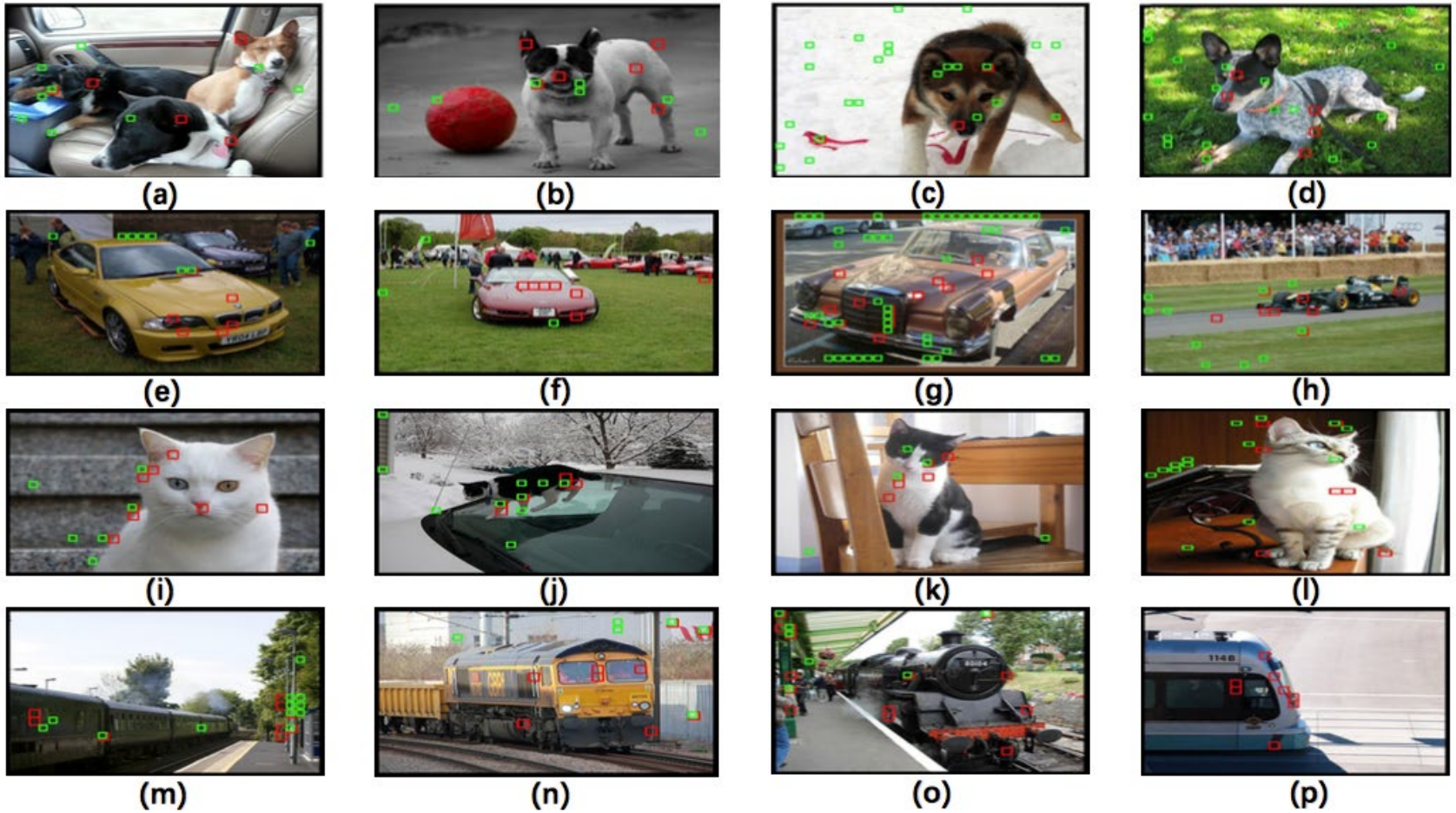
实验结果及分析

- 数据来源： *YFCC100M*
- 规模：10个类别，每个类别的数据约1000图片
- 数据构建：针对每个类别（如“狗”的图片），结合背景构造5个背景标签（如草地，水，车，地板、雪地

实验结果及分析


Context Bias
Relative F1 Improvement
10
5
0
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
bird
bridge
car
cat
church
dog
flower
horse
train
tree

实验结果及分析



Balance-Subsampled Stable Prediction

- Causal Regularizer (Approximate global balancing)
  - Making each variable in X become independent with others by learning a global sample weights W:

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{:,-j}^{T} \cdot (W \odot \mathbf{X}_{:,j})}{W^{T} \cdot \mathbf{X}_{:,j}} - \frac{\mathbf{X}_{:,-j}^{T} \cdot (W \odot (1 - \mathbf{X}_{:,j}))}{W^{T} \cdot (1 - \mathbf{X}_{:,j})} \right\|_{2}^{2}, \quad (4)$$

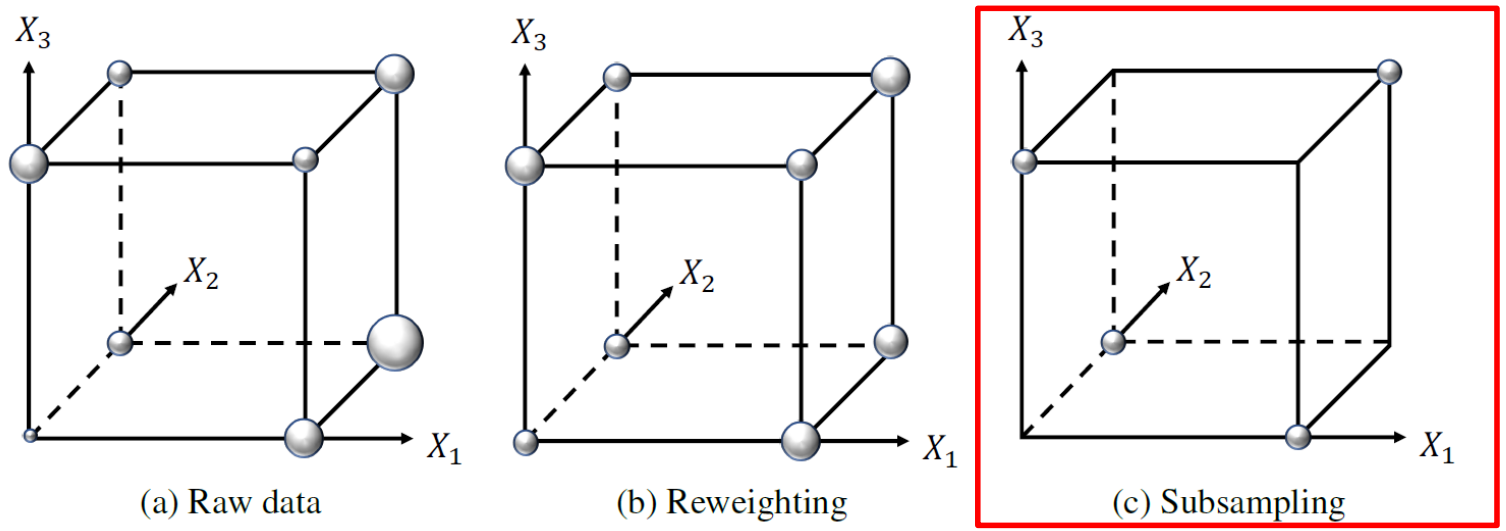


Figure 1: A toy example to illustrate the main idea of each deconfounding method.

Kuang K, Zhang H, Wu F, et al. Balance-Subsampled Stable Prediction[J]. TKDD 2021.

Variable Decorrelation for Stable Prediction

- Causal Regularizer (Approximate global balancing)
  - Making each variable in X become independent with others by learning a global sample weights W:

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^T \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2, \quad (4)$$

Variable Decorrelation by Sample Reweighting:

$$\min_{W} \sum_{j=1}^{p} \left\| \mathbb{E}[\mathbf{X}_{\cdot,j}^T \mathbf{\Sigma}_W \mathbf{X}_{\cdot,-j}] - \mathbb{E}[\mathbf{X}_{\cdot,j}^T W] \mathbb{E}[\mathbf{X}_{\cdot,-j}^T W] \right\|_2^2$$

Decorrelated Weighted Regression:

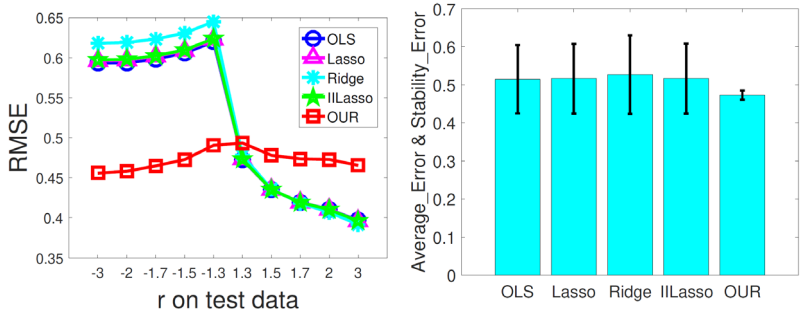
$$\min_{W,\beta} \sum_{i=1}^{n} W_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2 \quad (12)$$

$$s.t \quad \sum_{j=1}^{p} \left\| \mathbf{X}_{\cdot,j}^T \mathbf{\Sigma}_W \mathbf{X}_{\cdot,-j}/n - \mathbf{X}_{\cdot,j}^T W/n \cdot \mathbf{X}_{\cdot,-j}^T W/n \right\|_2^2 < \lambda_2$$

$$|\beta|_1 < \lambda_1, \quad \frac{1}{n}\sum_{i=1}^{n} W_i^2 < \lambda_3,$$

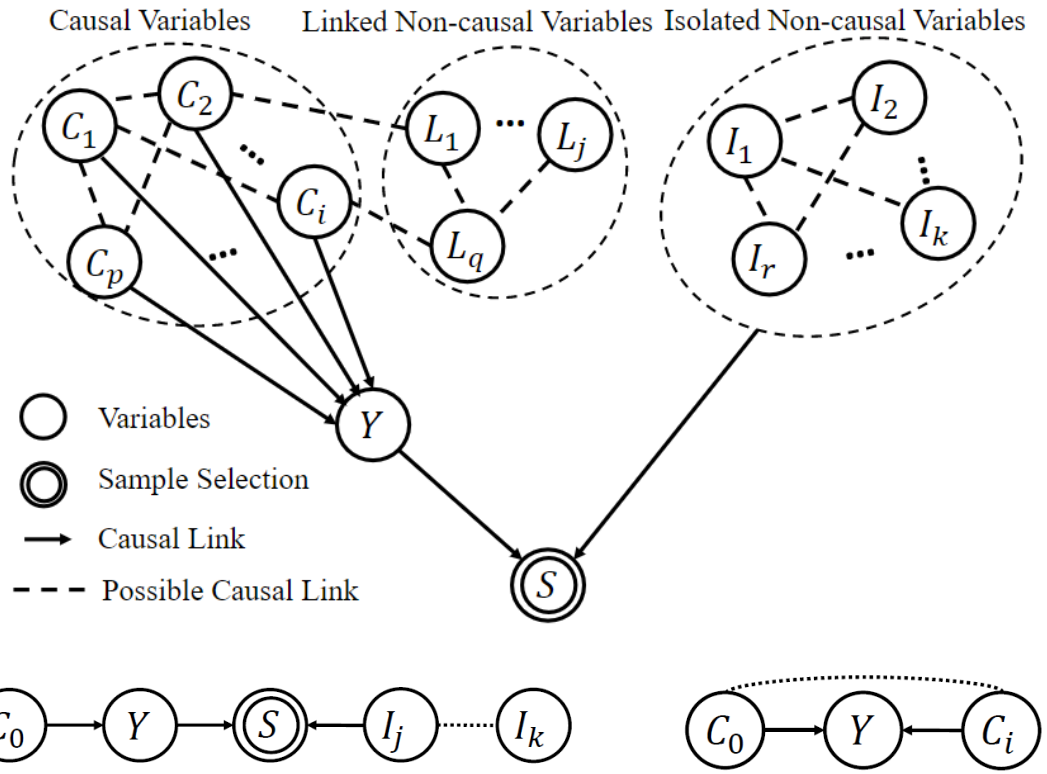
$$(\frac{1}{n}\sum_{i=1}^{n} W_i - 1)^2 < \lambda_4, \quad W \succeq 0,$$

Stable Prediction:



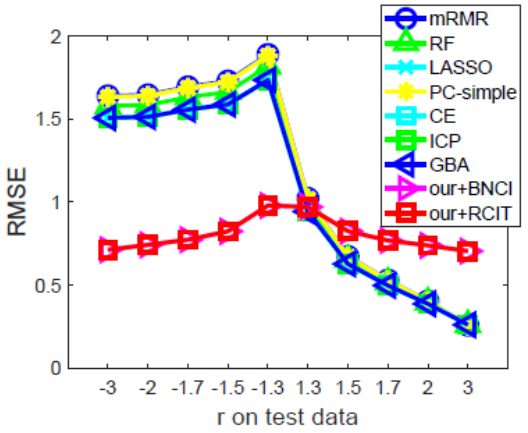
Kuang K, Xiong R, Cui P, et al. Stable Prediction with Model Misspecification and Agnostic Distribution Shift[C]//AAAI. 2020: 4485-4492.

Stable Prediction via Causal Features Selection

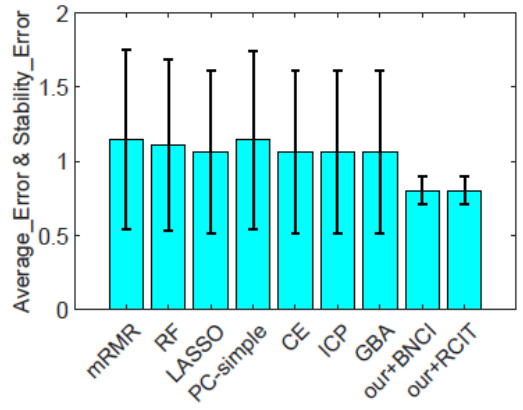


(a) Path between $\mathbf{C}_0$ and $\mathbf{I}_k$ (b) Path between $\mathbf{C}_0$ and $\mathbf{C}_i$

Theorem 1. *Given a causal variable* $\mathbf{C}_0$*, observed variables* $\mathbf{X}$ *and response variable* Y*, and assuming 1&2, then, for each causal variable* $\mathbf{C}_i \in \mathbf{C}$*, we have* $\mathbf{C}_i \not\perp\!\!\!\perp \mathbf{C}_0 \mid Y$*; and for each isolated non-causal variable* $\mathbf{I}_k \in \mathbf{I}$*, we have* $\mathbf{I}_k \perp\!\!\!\perp \mathbf{C}_0 \mid Y$.



(a) RMSE across testing environments



(b) Average Error (green bar) & Stability Error (black line)

Kuang K, Li B, Cui P, et al. Stable Prediction via Leveraging Seed Variable[J]. TKDE, 2022.

因果启发的可信机器学习（因果赋能机器学习）

发现了通过全局样本赋权实现因果关联挖掘的机制

PROPOSITION 3.3. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

$$\sum_{j=1}^{p}\left\|\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot\mathbf{X}_{\cdot,j})}{W^{T}\cdot\mathbf{X}_{\cdot,j}}-\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot(1-\mathbf{X}_{\cdot,j}))}{W^{T}\cdot(1-\mathbf{X}_{\cdot,j})}\right\|_2^2$$

↓

0

存在一组样本权重，使得任意输入变量与其他变量之间相互独立

PROOF. Since $\|\cdot\| \geq 0$, Eq. (8) can be simplified to $\forall j, \forall k \neq j$

$$\lim_{n\to\infty}\left(\frac{\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=1}} W_i}{\sum_i t\mathbf{X}_{i,j=1} W_i} - \frac{\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=0}} W_i}{\sum_i t\mathbf{X}_{i,j=0} W_i}\right) = 0$$

with probability 1. For W^*, from Lemma 3.1, $0 < P(\mathbf{X}_i = x) < 1$, $\forall x, \forall i, t = 1$ or 0,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,j=t} W_i^* &= \lim_{n\to\infty}\frac{1}{n}\sum_x x_{\cdot j=t}\sum_i t\mathbf{X}_{i=x} W_i^* \\ &= \lim_{n\to\infty}\sum_x x_{\cdot j=t}\frac{1}{n}\sum_i t\mathbf{X}_{i=x}\frac{1}{P(\mathbf{X}_i=x)} \\ &= \lim_{n\to\infty}\sum_x x_{\cdot j=t} P(\mathbf{X}_i = x)\cdot\frac{1}{P(\mathbf{X}_i=x)} = 2^{p-1}\end{aligned}$$

with probability 1 (Law of Large Number). Since features are binary,

$$\lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=1}} W_i^* = 2^{p-2}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,j=0} W_i^* = 2^{p-1},\quad \lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=0}} W_i^* = 2^{p-2}$$

and therefore, we have following equation with probability 1:

$$\lim_{n\to\infty}\left(\frac{\mathbf{X}_{\cdot,k}^{T}(W^*\odot\mathbf{X}_{\cdot,j})}{W^{*T}\mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,k}^{T}(W^*\odot(1-\mathbf{X}_{\cdot,j}))}{W^{*T}(1-\mathbf{X}_{\cdot,j})}\right) = \frac{2^{p-2}}{2^{p-1}} - \frac{2^{p-2}}{2^{p-1}} = 0.$$

□

样本重采样因果恢复模型

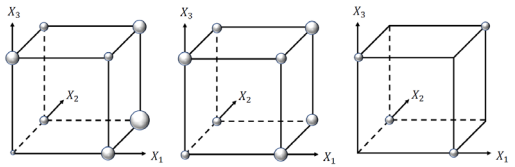
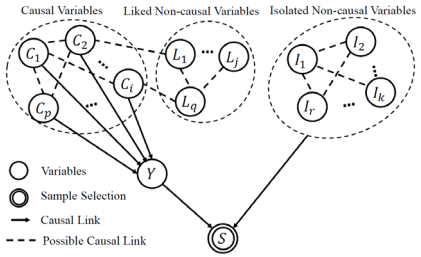
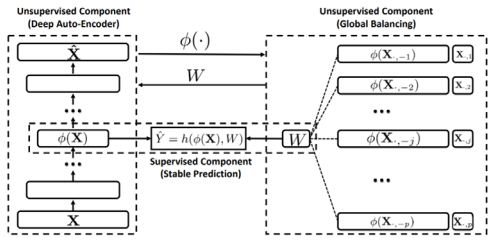


Figure 1: A toy example to illustrate the main idea of each deconfounding method.

因果特征选择解耦模型



因果约束的深度学习模型



有效提升机器学习模型的可解释性和稳定性

Haotian Wang, Kun Kuang, Long Lan, Wanrong Huang, Fei Wu, Wenjing Yang. Out-of-distribution Generalization with Causal Feature Separation, TKDE, 2023.
Kuang K, Li B, et al. Stable Prediction via Leveraging Seed Variable[J]. TKDE 2022.
Kuang K, Xiong R, et al. Stable prediction with model misspecification and agnostic distribution shift[C]//AAAI, 2020
Kuang K, Cui P, et al. Stable Prediction across Unknown Environments. KDD, 2018.

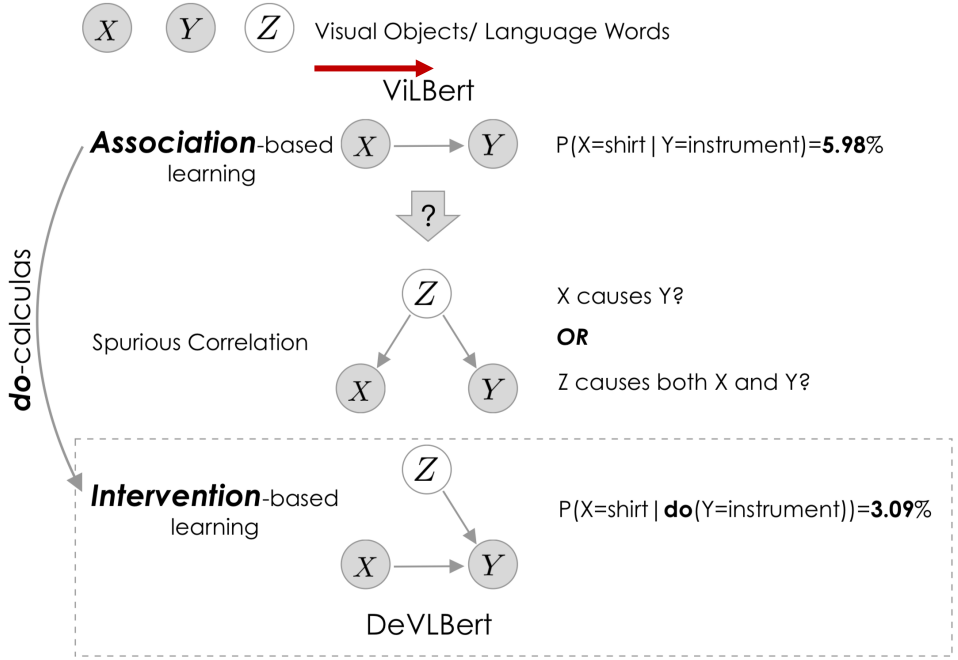
因果推理应用1

- 因果干预模型赋能多模态数据表征学习，应用于电商短视频推荐
- 问题：多模态数据表征学习
- 挑战：数据偏差导致文本特征与视觉特征虚假相关
- 方法：因果去偏差指导表征学习

因果干预
（do算子）

P(衬衫 | 乐器) = 5.98%

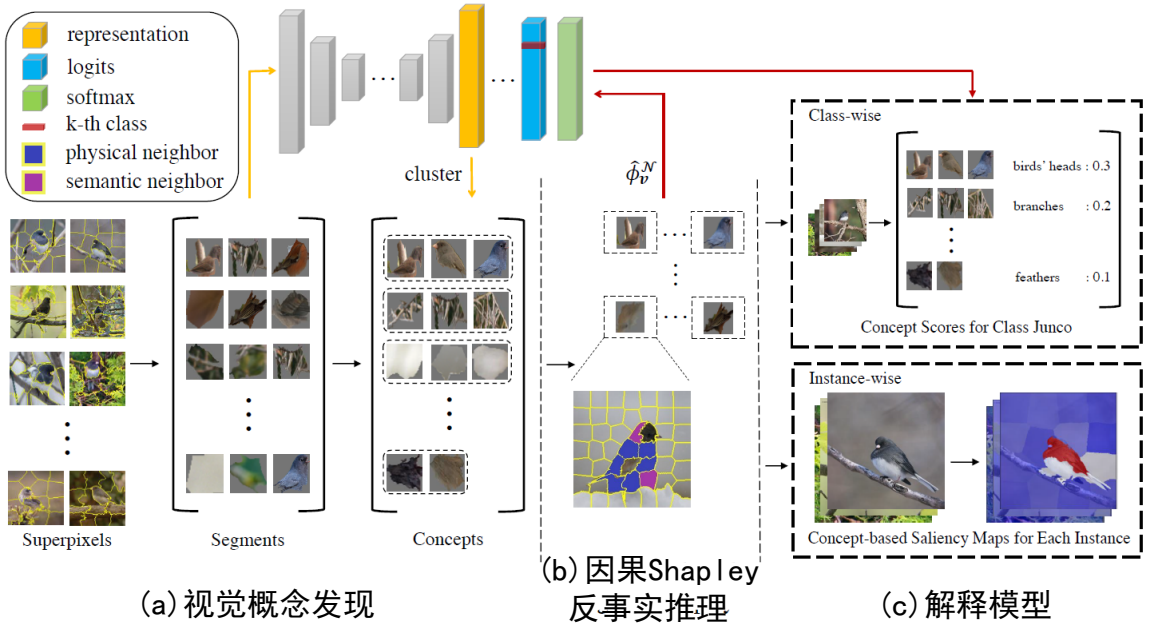
P(衬衫 | **do**(乐器)) = 3.10%



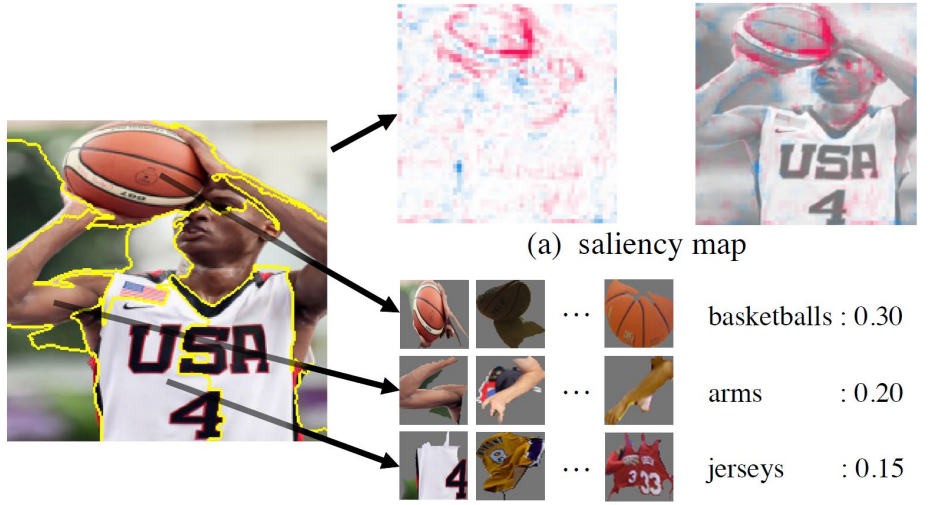
Zhang S, Jiang T, Wang T, et al. DeVLBert: Learning Deconfounded Visio-Linguistic Representations[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 4373-4382.

因果推理应用2

- 因果反事实+视觉知识赋能机器学习可解释
- 问题：机器学习中模型可解释性，为什么图中场景是在“打篮球”
- 挑战：虚假相关、如何从像素级别解释到视觉概念级别解释
- 方法：视觉知识+因果反事实推理



(a) 视觉概念发现　(b) 因果Shapley反事实推理　(c) 解释模型



(a) saliency map

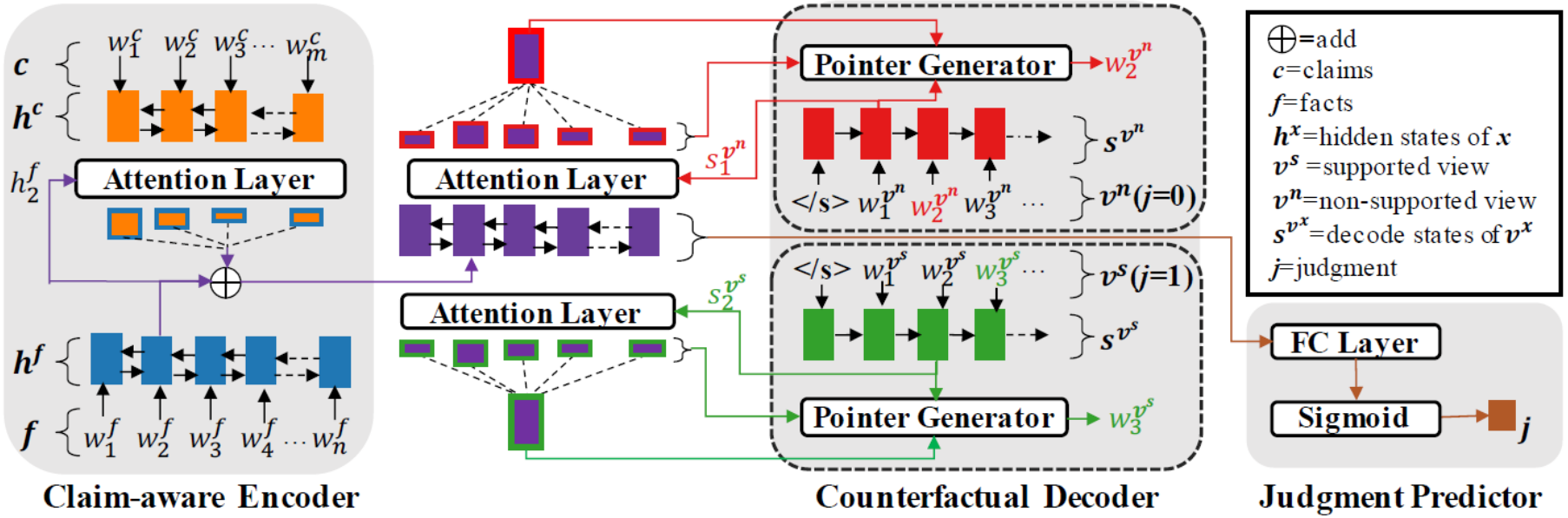
(b) important concepts and corresponding scores

Li J, Kuang K, Xiao J, et al. Instance-wise or Class-wise? A Tale of Neighbor Shapley for Concept-based Explanation, MM 2021

Li J, Kuang K, Xiao J, et al. Shapley Counterfactual Credits for Multi-Agent Reinforcement Learning, KDD 2021

因果推理应用3

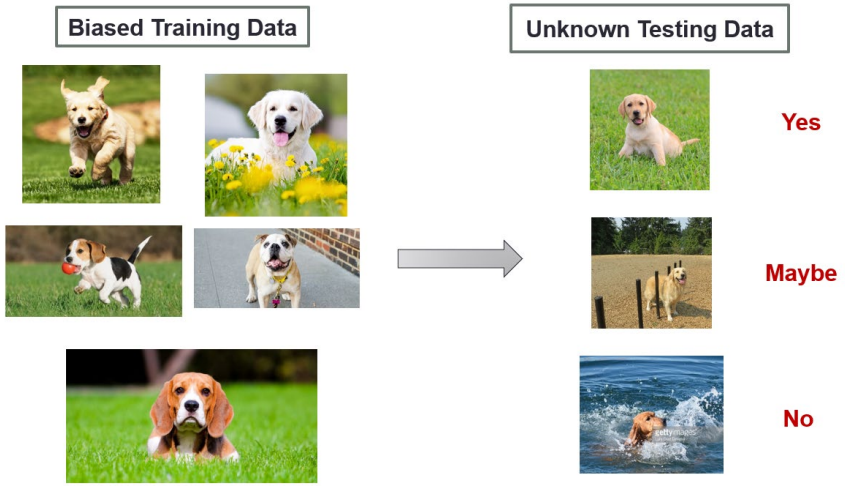
- 因果反事实赋能智能司法，实现法律文本生成
- 问题：法律裁判文书中“本院认为”生成
- 挑战：数据偏差/虚假关联（胜率高的案件才会起诉）
- 方法：因果反事实推理赋能文本生成

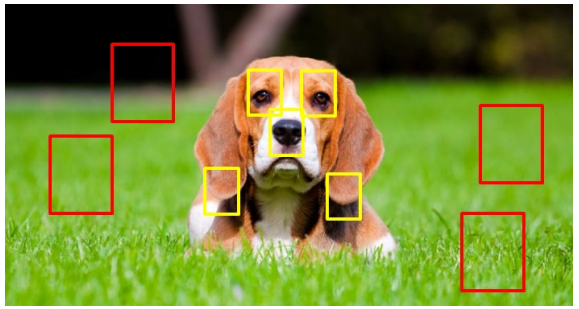


Wu Y, Kuang K, Zhang Y, et al. De-biased Court's View Generation with Causality[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 763-780.

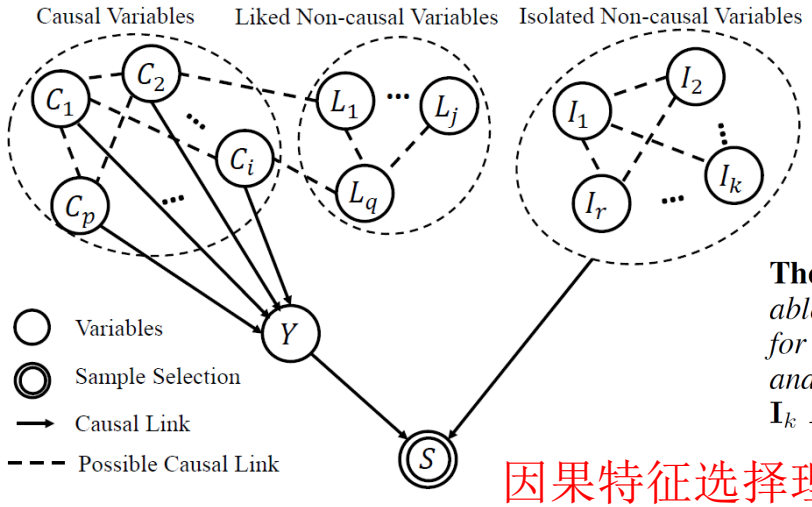
因果推理应用4

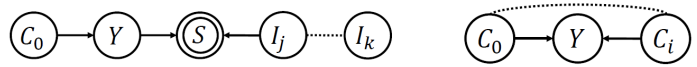
- 因果特征选择，实现面向未知环境的稳定预测
- 背景：传统方法是关联驱动，存在虚假关联
  - 训练图片中多数狗在草地上，草地与狗虚假相关
- 草地预测狗：不可解释，不稳定等
- 方法：区分因果特征和非因果特征





因果特征：狗眼睛、耳朵等
非因果特征：草地、背景等





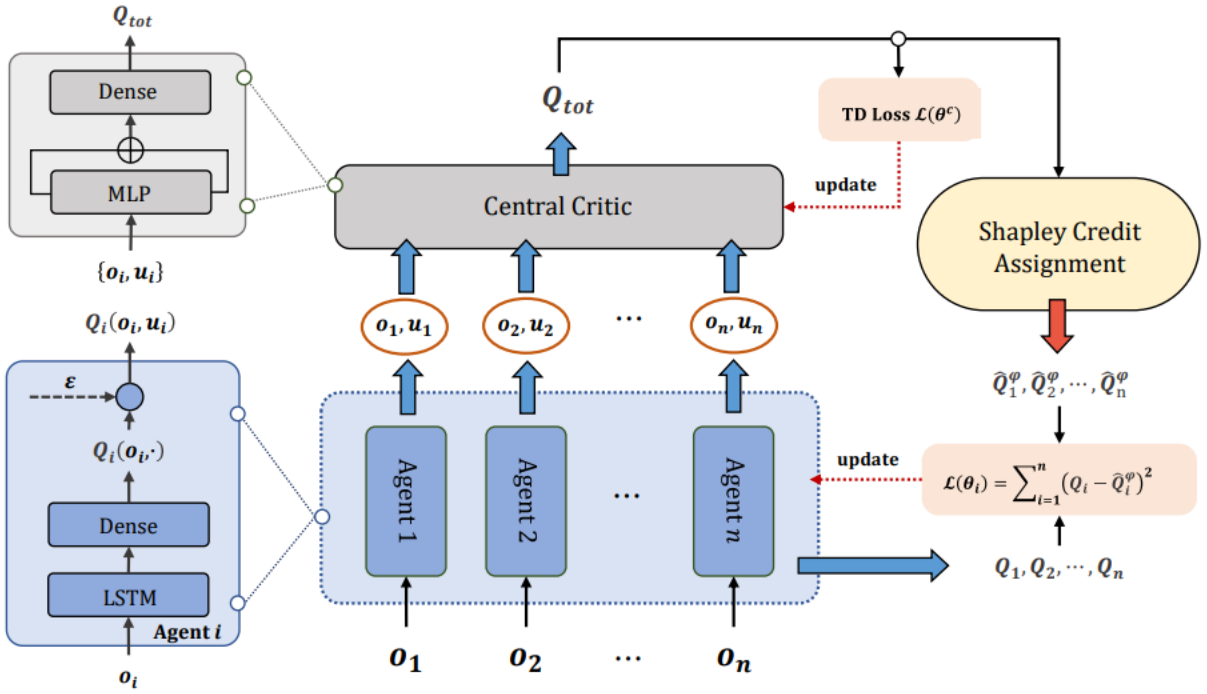
Theorem 1. *Given a causal variable* $\mathbf{C}_0$*, observed variables* $\mathbf{X}$ *and response variable* Y*, and assuming 1&2, then, for each causal variable* $\mathbf{C}_i \in \mathbf{C}$*, we have* $\mathbf{C}_i \not\perp\!\!\!\perp \mathbf{C}_0 \mid Y$*; and for each isolated non-causal variable* $\mathbf{I}_k \in \mathbf{I}$*, we have* $\mathbf{I}_k \perp\!\!\!\perp \mathbf{C}_0 \mid Y$.

因果特征选择理论保证

Kuang K, Li B, et al. Stable Prediction via Leveraging Seed Variable[J]. arXiv, 2020.
Liu J, Shen Z, et al. Invariant Adversarial Learning for Distributional Robustness[C]. AAAI, 2021.
Kuang K, Xiong R, et al. Stable prediction with model misspecification and agnostic distribution shift[C]//AAAI, 2020.

因果推理应用5

- 因果反事实推理赋能多智能体强化学习信度分配
- 问题：多智能体强化学习中，如何计算每个智能体对团体的贡献?
- 方法：提出因果沙普利值，通过因果反事实推理计算智能体贡献
  - 当某个智能体不参与团队合作，团队收益的减少量作为其对团队的贡献

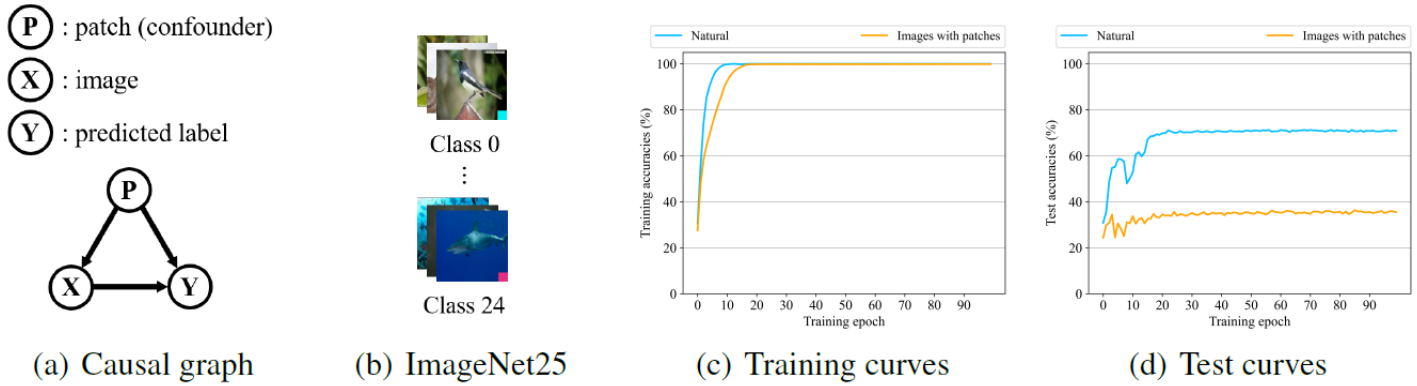


Map	Methods													
	VDN		QMIX		QTRAN		COMA		QPD		SQDDPG		OURS	
	$\tilde{m}$	avg	$\tilde{m}$	avg	$\tilde{m}$	avg	$\tilde{m}$	avg	$\tilde{m}$	avg	$\tilde{m}$	avg	$\tilde{m}$	avg
3m	**100**	**100**	**100**	**100**	**100**	**100**	95	96	99	99	64	65	99	99
8m	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	95	95	92	90	98	97
2s3z	**100**	**100**	**100**	**100**	92	91	45	45	99	98	60	55	**100**	**100**
1c3s5z	88	85	**95**	**90**	40	41	15	15	77	72	2	2	61	60
3s5z	80	69	80	67	12	13	5	3	79	80	1	1	**92**	**90**
3s5z_vs_3s6z	0	0	0	0	0	0	0	0	3	5	0	0	**20**	**20**

Li J, Kuang K, Xiao J, et al. Shapley Counterfactual Credits for Multi-Agent Reinforcement Learning, KDD 2021

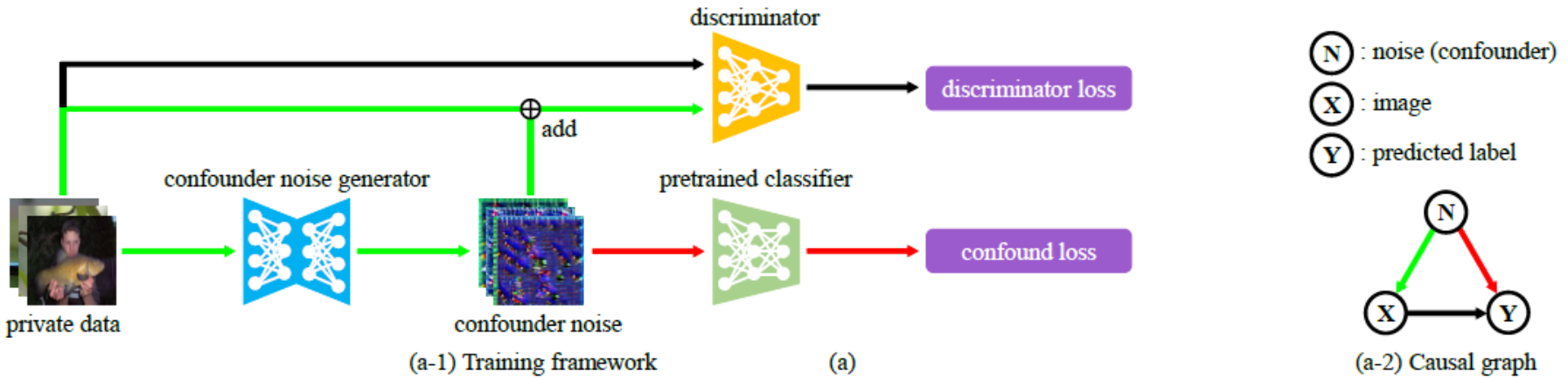
因果机器学习应用6

- 因果混淆机制赋能数据隐私保护



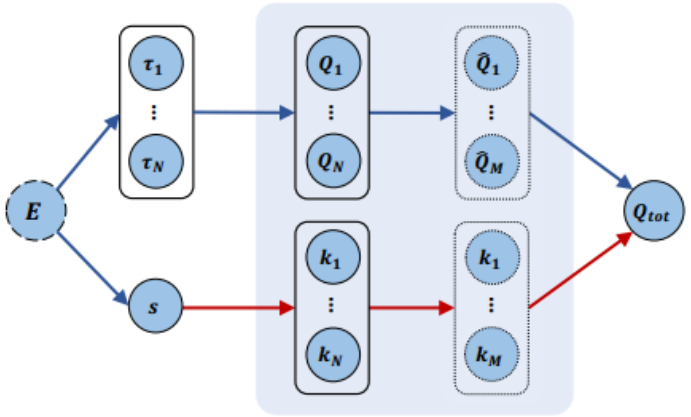
(a) Causal graph (b) ImageNet25 (c) Training curves (d) Test curves

Figure 2: Introducing confounders in ImageNet25 by class-wise patches.

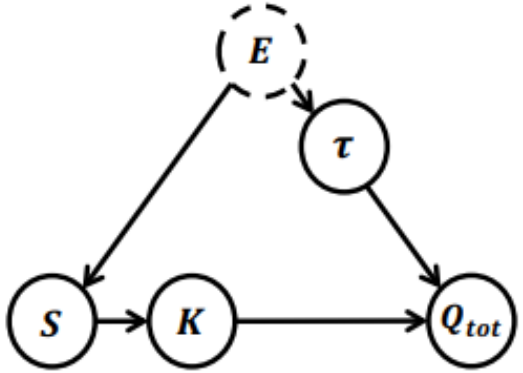


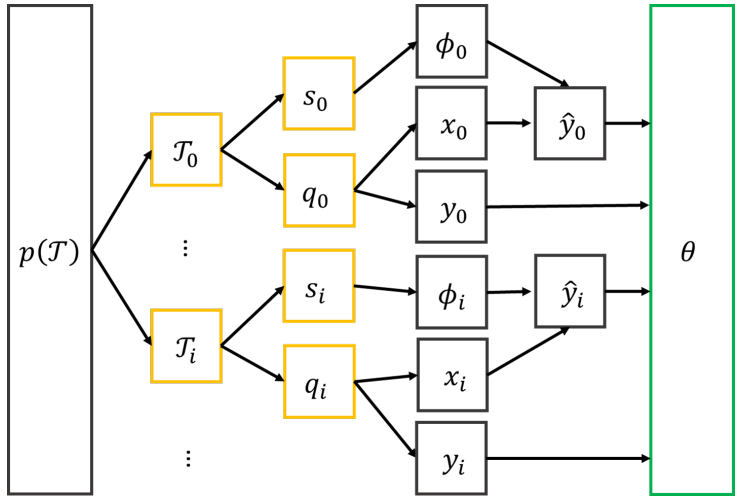
Qi Tian, Kelu Jiang, Kun Kuang*, Furui Liu, Zhihua Wang, Fei Wu. ConfounderGAN: Protecting Image Data Privacy with Causal Confounder, NeurIPS, 2022.

因果机器学习应用7

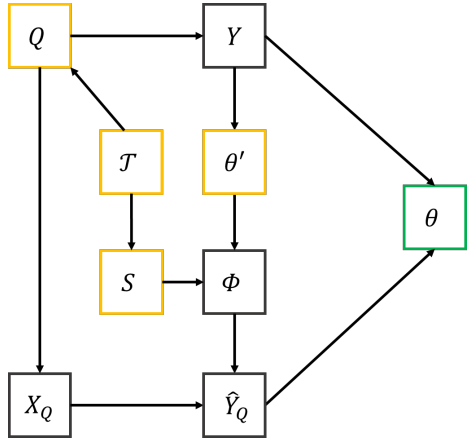


因果+多智能体强化学习





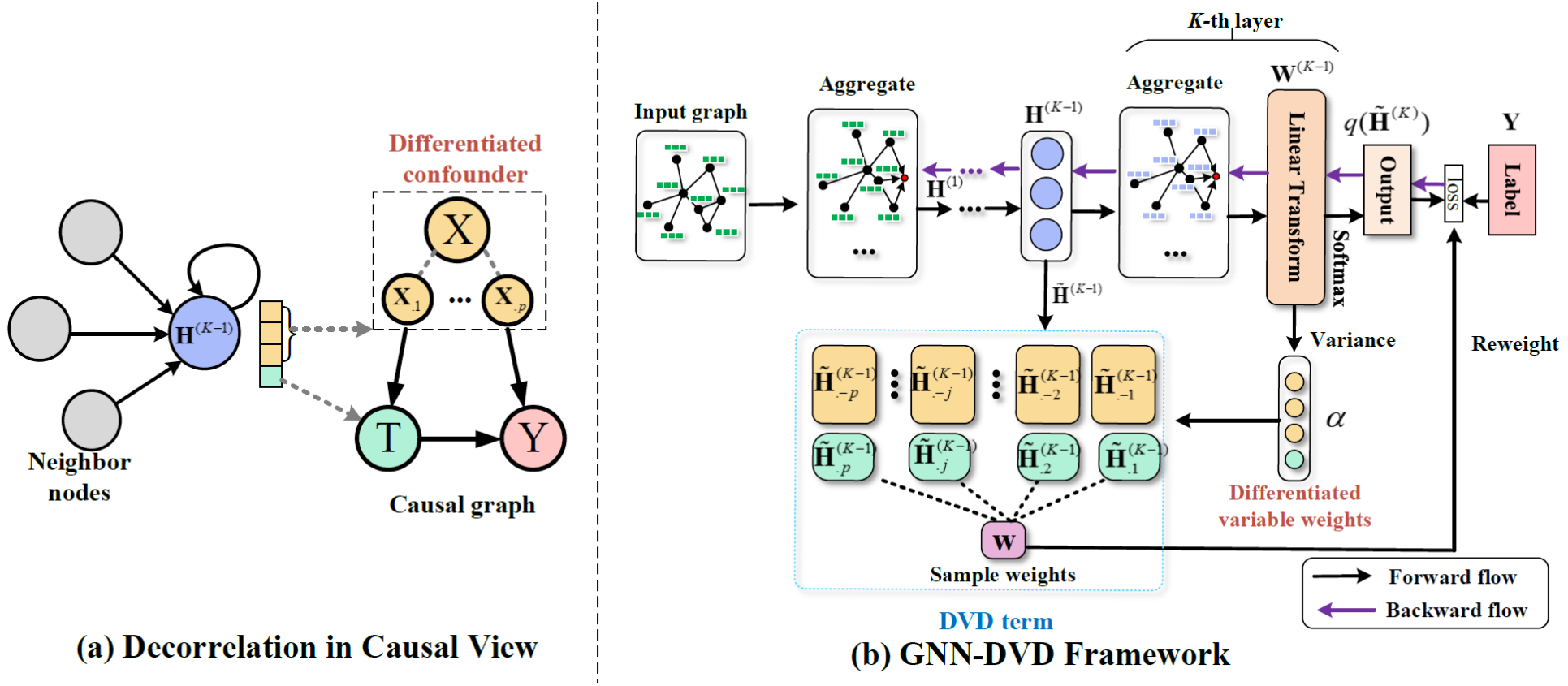
因果+元学习



Jiahui Li, Kun Kuang, et al. Deconfounded Value Decomposition for Multi-Agent Reinforcement Learning, ICML 2022

Yinjie Jiang, Zhengyu Chen, Kun Kuang et al. The Role of Deconfounding in Meta-learning, ICML 2022

因果机器学习应用8（因果图神经网络）

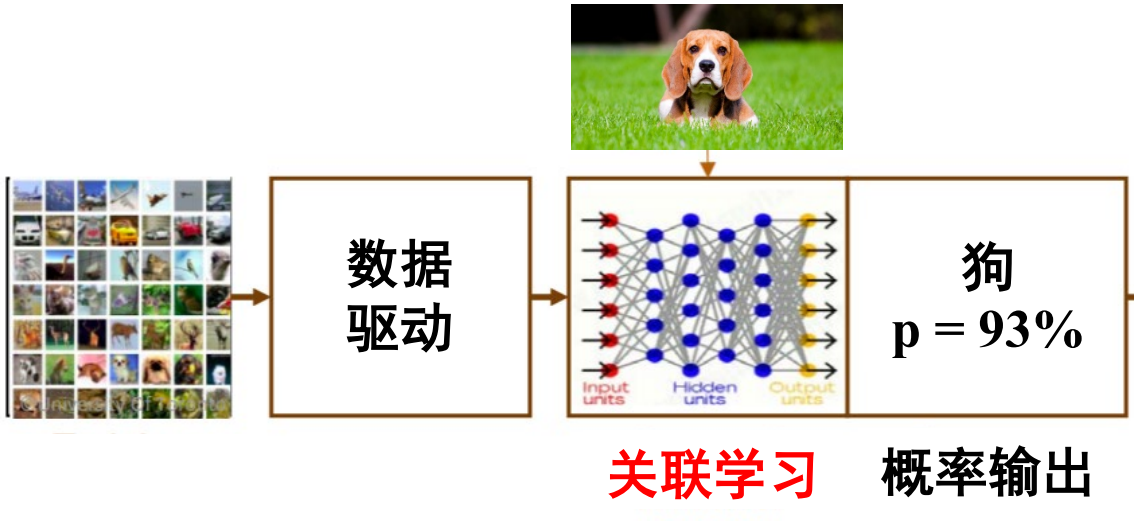


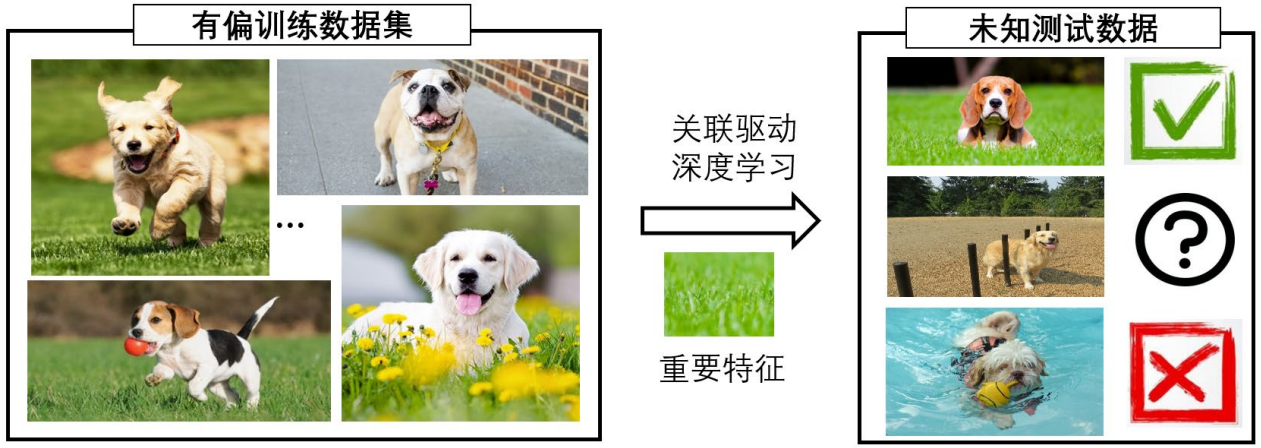
Zhu D, Wang D, Kuang K*, et al. Graph Neural Network with Two Uplift Estimators for Label-Scarcity Individual Uplift Modeling, The WebConf, 2023.

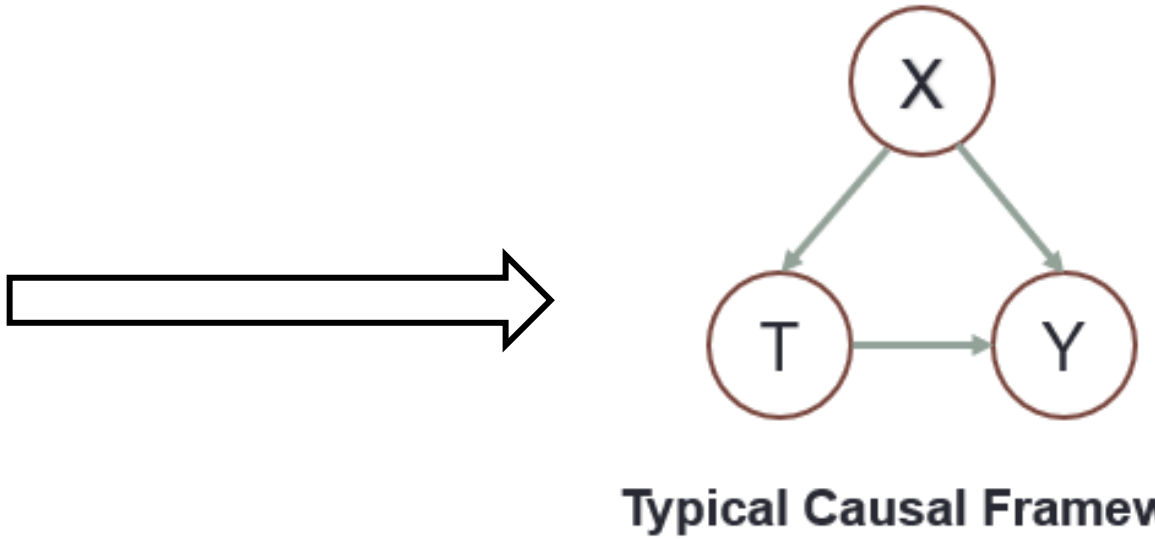
Shaohua Fan, Xiao Wang, Chuan Shi, **Kun Kuang**, et.al. Debiased Graph Neural Networks with Agnostic Label Selection Bias. TNNLS 2022.

Zhengyu Cheng, Teng Xiao, Kun Kuang*. BA-GNN: On Learning Bias-Aware Graph Neural Network, ICDE, 2022.

因果启发的可信机器学习（因果赋能机器学习）







浙江大学
ZHEJIANG UNIVERSITY

因果赋能大语言模型探索与思考

况琨
浙江大学计算机学院

大模型的学习特点也是数据驱动，关联学习

对齐（Alignment）

与数据空间对齐
与人类指令对齐
与人类反馈对齐

predict the next token

完形填空形式下文字接龙（自监督学习）

- 原话：一辆 列车 缓慢 行驶 在 崎岖 的 山路上
- 移除单词：一辆 列车 ___ 行驶 在 崎岖 的 山路上
- 预测填空：一辆 列车 缓慢 行驶 在 崎岖 的 山路上

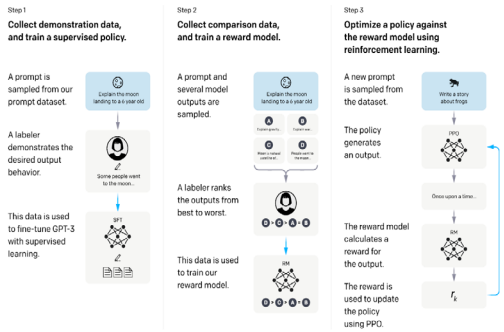
supervised fine-tuning

提示学习与指令微调（人教机学）

第一轮：{"instruction":本院查明,被告人酒后...请分析案情。
"output":根据上述内容，可以认定本案的核心要素包括醉酒驾驶、致人受伤、酒后逃逸...},
第二轮：{"instruction":根据上述分析，请预测罪名。
"output":本案预测的罪名是危险驾驶罪}
第三轮：{"instruction":请给出处罚意见。
"output":结合嫌疑人逃逸的情节，建议考虑拘役三个月，并罚款6000元}
...

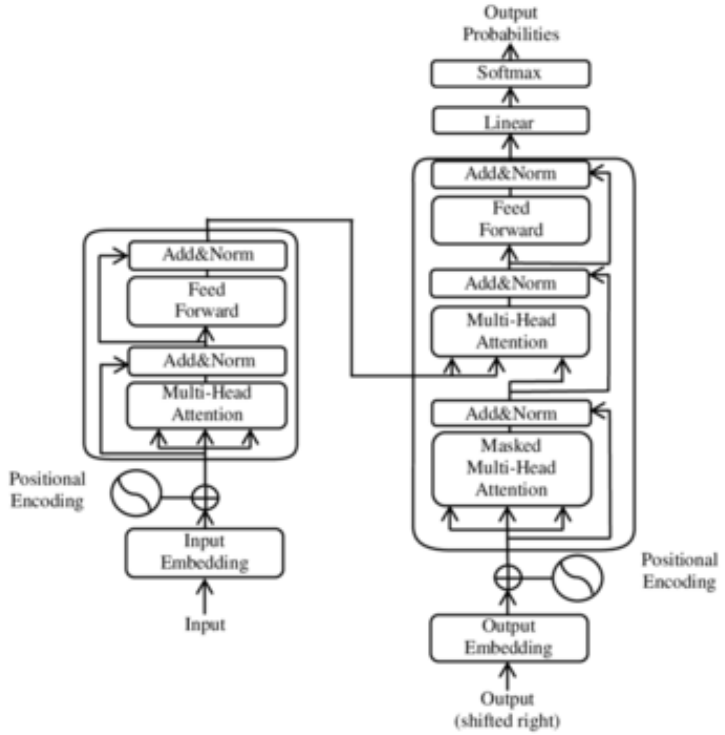
Reinforcement Learning from Human Feedback

人类反馈下强化学习（尝试与探索）



自然语言合成的核心神经网络是Transformer 模型

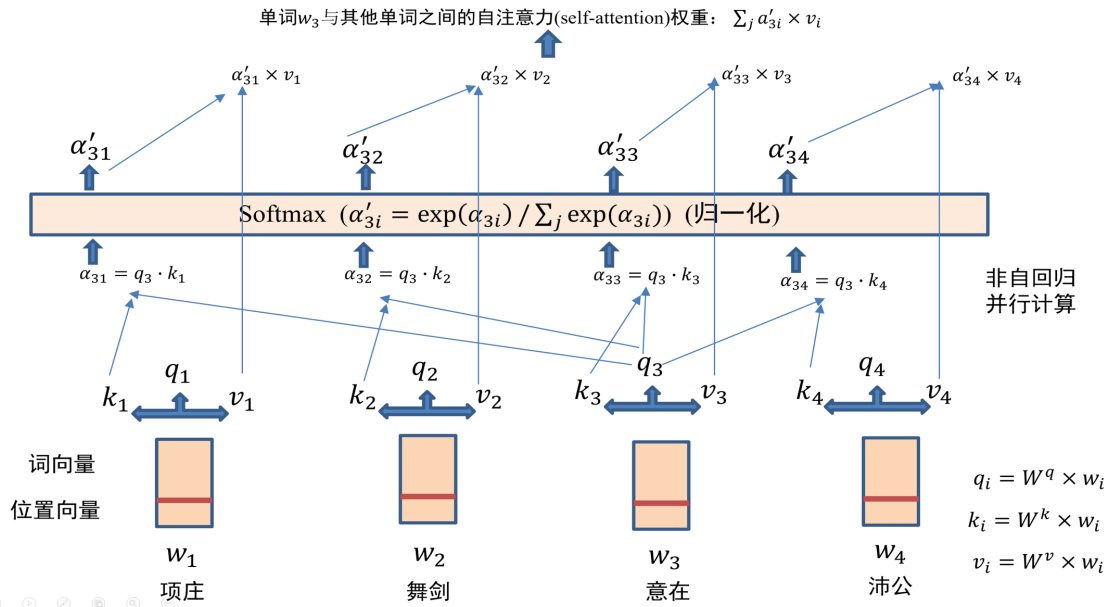
思考：大模型泛化问题---由关联到因果



消除反馈(recurrent)机制

Google (2017): Attention is all you need

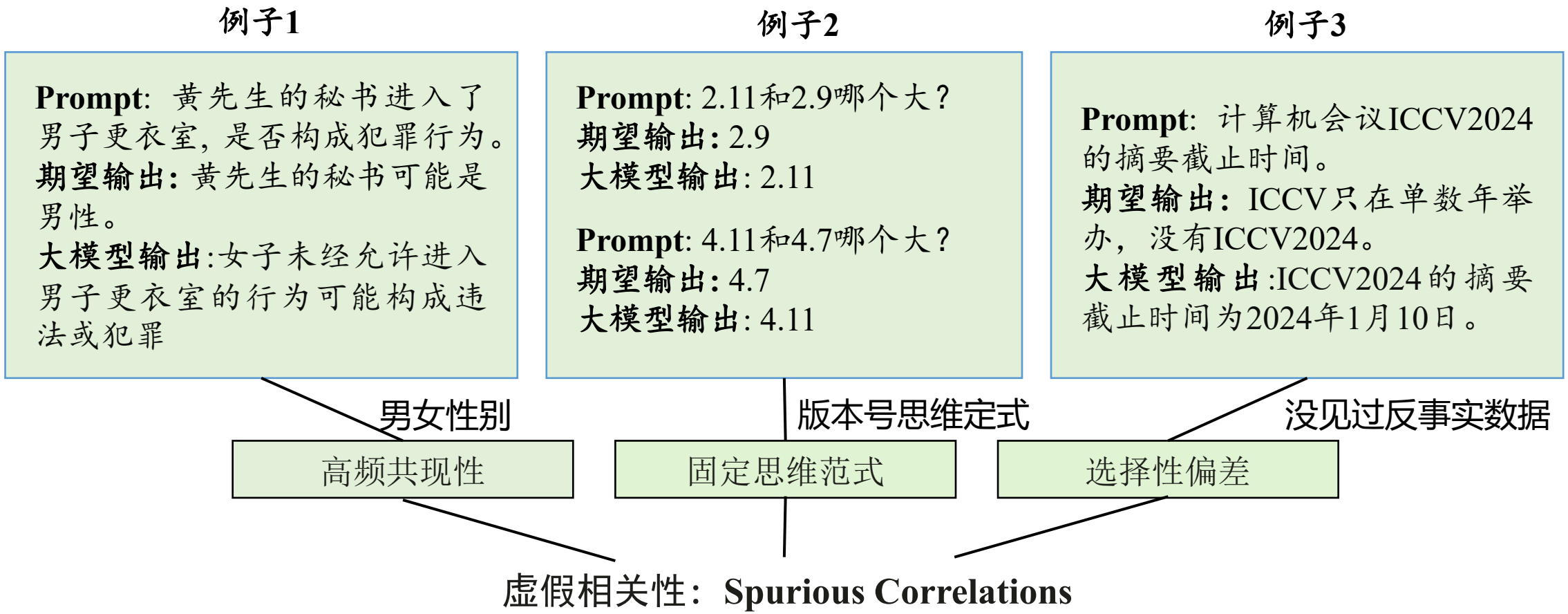
attention：单词共生概率



项庄 舞剑 意在 沛公

学习单词和单词之间**关联关系** （in-context meaning）

思考：大模型泛化问题---由关联到因果



LLM依赖于概率建模，这种建模通常捕捉到的是植根于语言模态和社会刻板印象的虚假相关性，而不是实体与事件之间的真正因果关系，导致推理出错。

Wu A, Kuang K, Zhu M, et al. Causality for large language models[J]. arXiv preprint arXiv:2410.15319, 2024.

大模型可解释可泛化问题：由关联到因果

从Chat到Sora：对合成内容中最小单元进行有意义的关联组合，犹如昨日重现

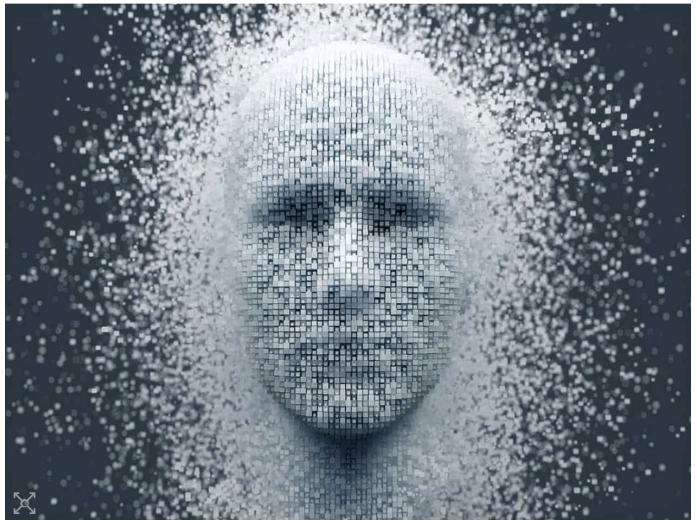
I am four years old.

There are five people in my family.

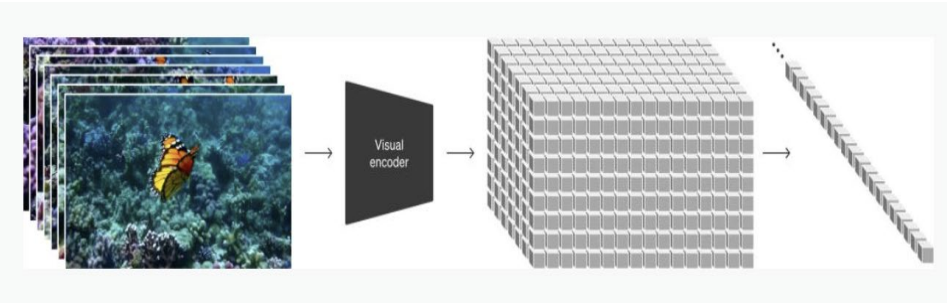
I am young, short and thin.

My dad is forty years old.

单词有意义的线性组合



像素点有意义的空间组合



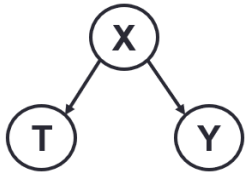
时空子块有意义的时序组合

关联的三种来源

因果

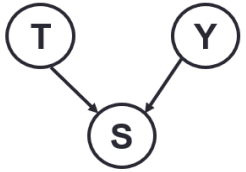
可解释
稳定/鲁棒
可决策

混淆偏差

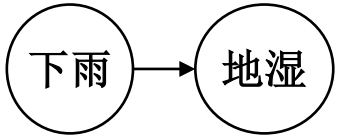


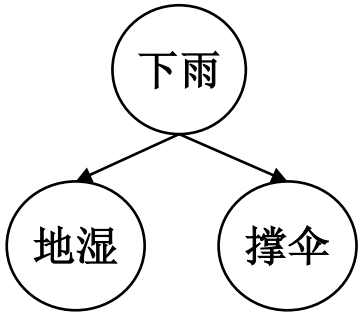
虚假关联: 当忽略 X 时，T 和 Y 相关

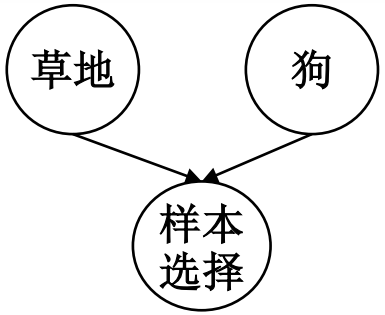
选择偏差



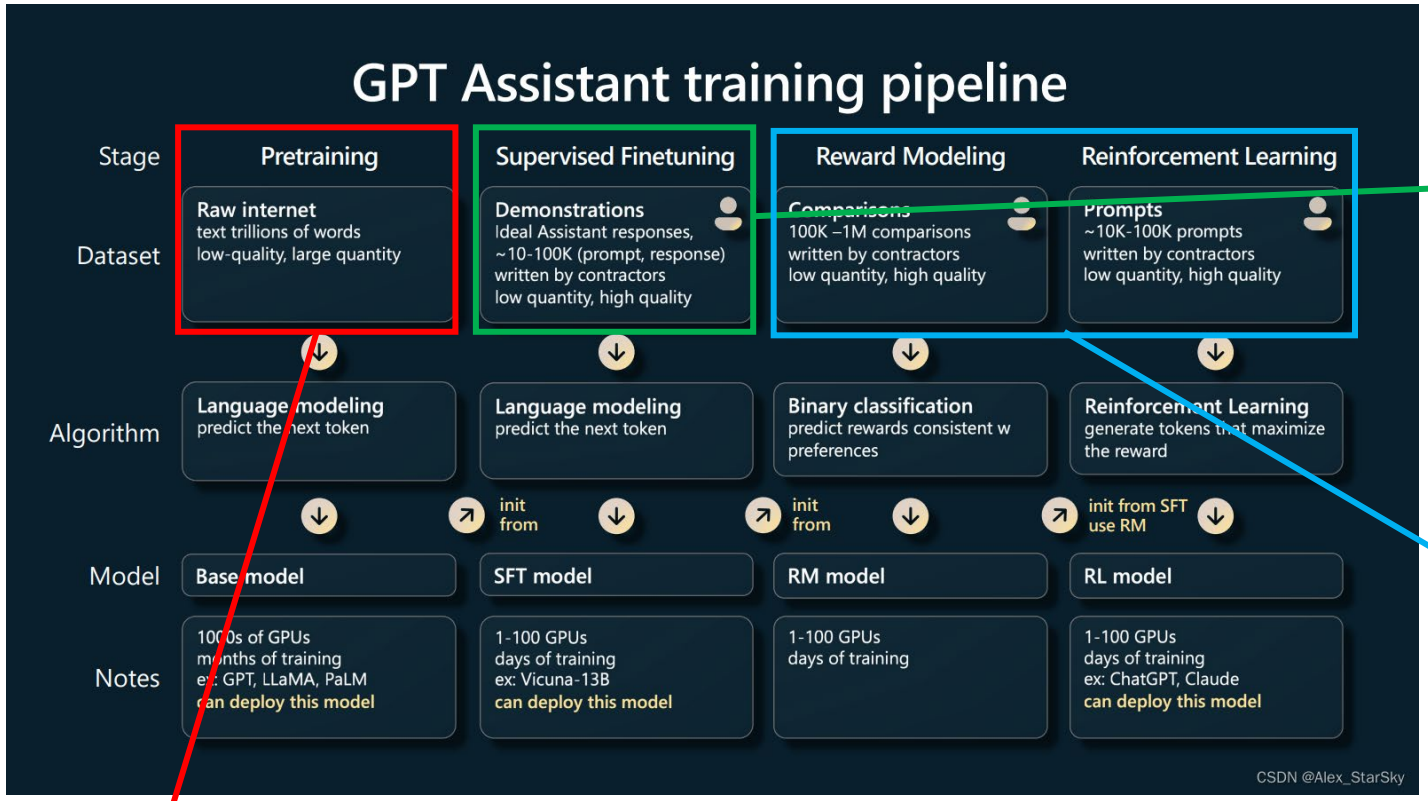
虚假关联: 当给定 S 时，T 和 Y 相关







因果如何赋能大模型



- 因果去除数据偏置
  - 虚假相关问题
  - 灾难遗忘问题
- 因果支撑决策
  - 用户偏好对齐
  - 因果强化学习
- 因果赋能Transformer
  - 由关联自回归到因果回归机制
  - 因果Transformer架构
  - 基于因果知识增强的Transformer架构

因果赋能Transformer

① 由关联自回归到因果回归机制

② 因果Transformer架构

③ 基于因果知识增强的Transformer架构

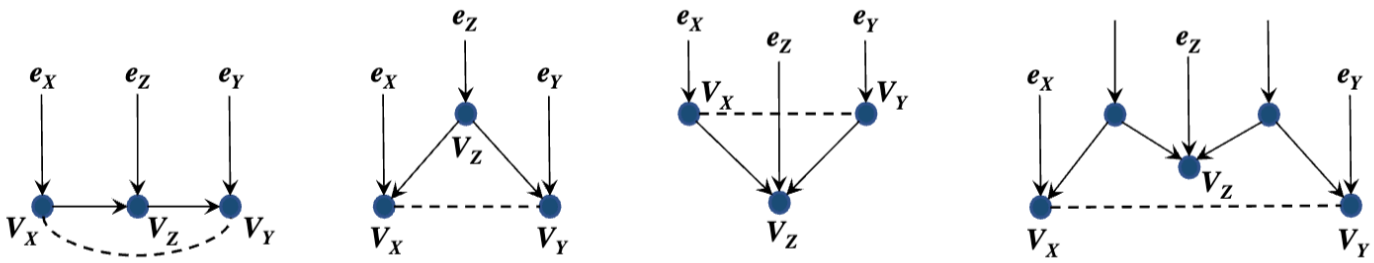
1. 由关联自回归到因果回归机制

- 关联回归学习：变量关联关系复杂，包含因果关联和虚假关联

OLS回归（最小二乘法）：

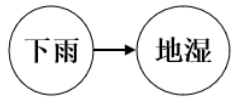
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

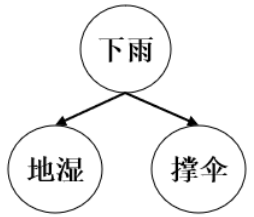
$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki})^2$$

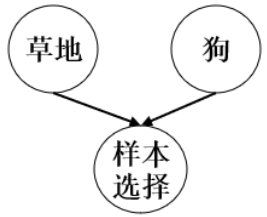


(a) Chain Structure. (b) Fork Structure. (c) Collider Structure. (d) M Structure.

Figure 1: Underlying bias present in unknown causal relation.



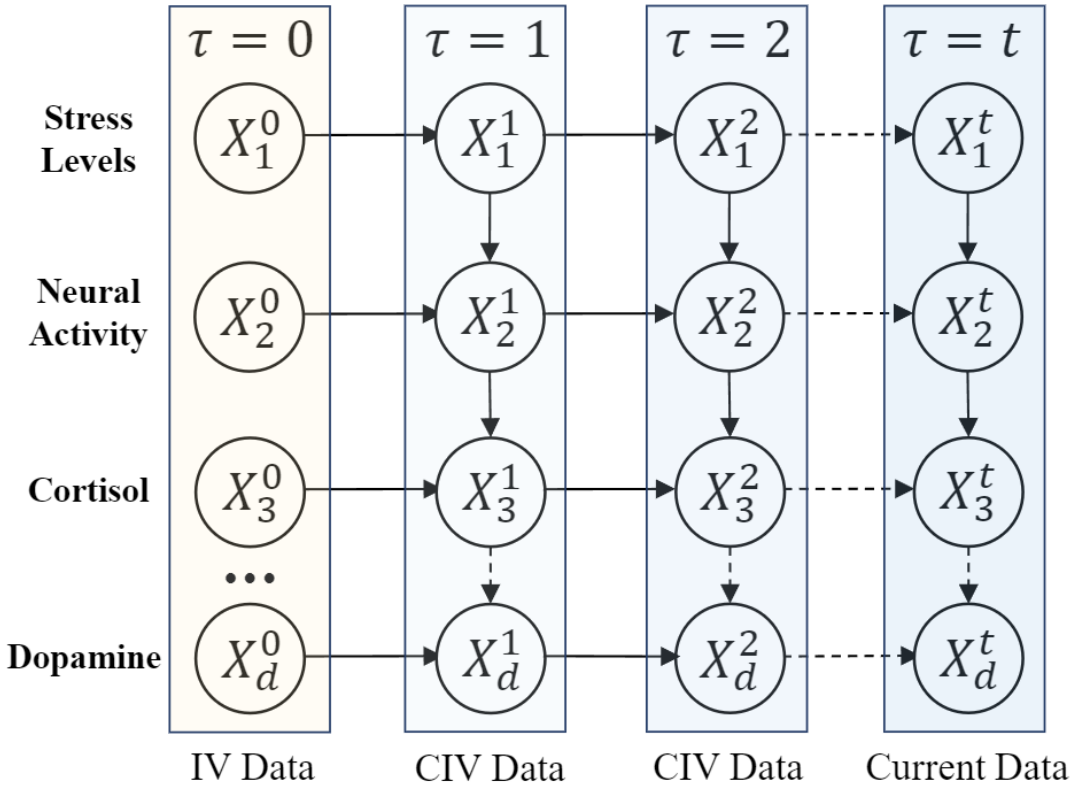


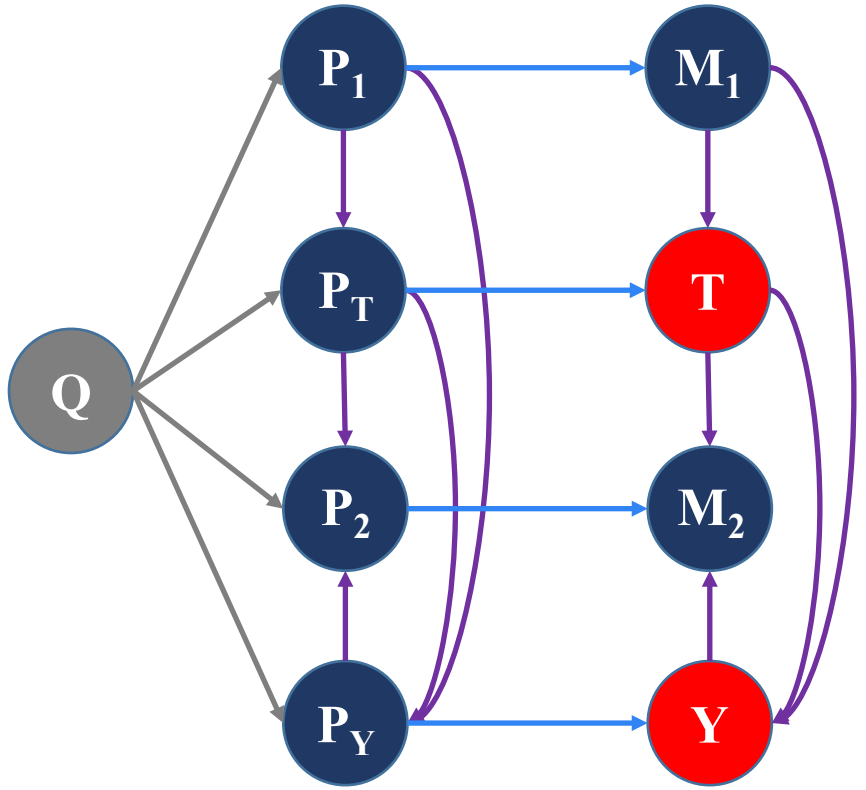


关联 v.s. 因果

1. 由关联自回归到因果回归机制

- 因果回归机制：基于时序数据，计算任意两个变量之间因果回归系数





M_1 brings Confounding Bias

M_2 brings Collider Bias

基于前序时间变量P_1, P_2, P_Y，P_T为T和Y的条件工具变量，可计算T和Y的因果系数

Anpeng Wu, Haoxuan Li, Kun Kuang*, Zhang Keli, Fei Wu. Learning Causal Relations from Subsampled Time Series with Two Time-Slices, ICML, 2024.

1. 由关联自回归到因果回归机制

- 由关联自回归到因果回归

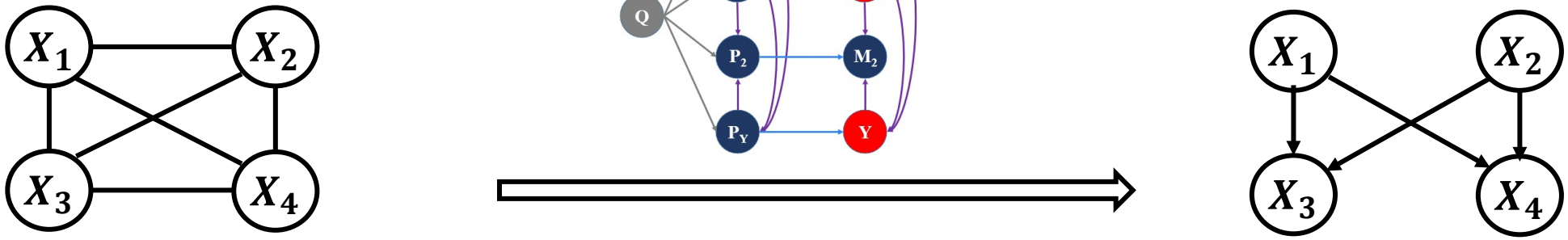
OLS回归（最小二乘法）：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki})^2$$

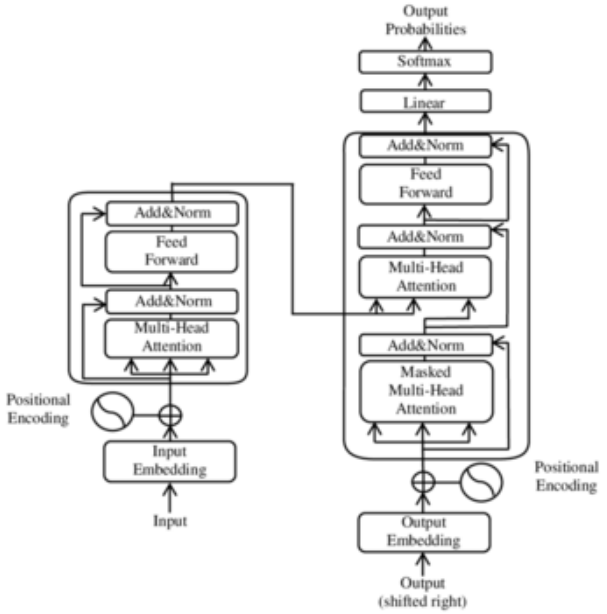
因果回归：

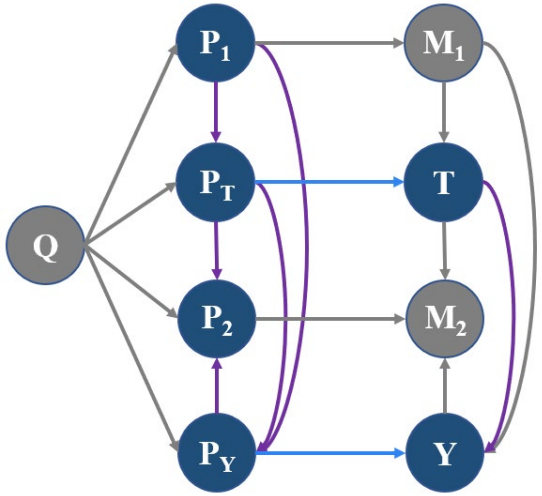
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_t X_{ti} + \hat{\beta}_k X_{ki}$$



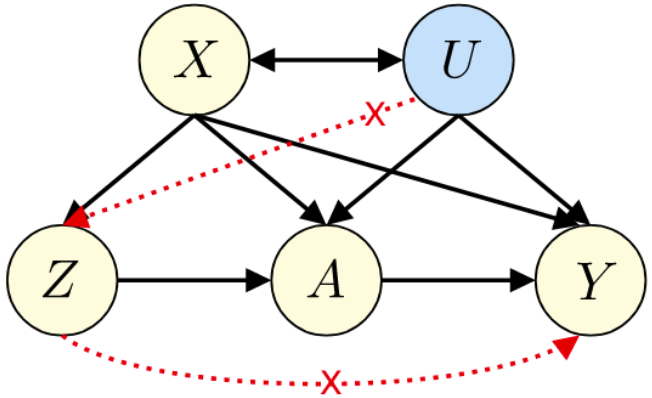
2.基于因果回归/因果发现的因果Transformer

- 由基于自回归的Transformer到基于因果回归/发现的因果Transformer





因果回归方法



因果发现方法

$$\text{Attention}(\boldsymbol{X}) = \text{softmax}\left(\frac{\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_k\boldsymbol{X}^T}{\sqrt{d}}\right)\boldsymbol{X}\boldsymbol{W}_v$$

消除反馈(recurrent)机制

Google (2017): Attention is all you need

attention：QK^T计算单词之间的相关性

因果回归

因果发现

$$\textbf{Causal Self-Attention}(\boldsymbol{X}) = \underbrace{\sigma\left(\mathbf{A}\sigma\left(\mathbf{A}\boldsymbol{X}\boldsymbol{W}_v^{(0)}\right)\boldsymbol{W}_v^{(1)}\right)}_{\text{2-layer GCN}}$$

$$\mathbf{A} = \text{softmax}\left(\frac{\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_k\boldsymbol{X}}{\sqrt{d_k}}\right)$$

Causal Transformers: Improving the Robustness on Spurious Correlations

2. 因果Transformer架构

- 由基于自回归的Transformer到基于因果回归/发现的因果Transformer

Causal Interpretation of Self-Attention:

- Self-attention is a mechanism that estimates a **linear structural equation model** in the deepest layer for each input sequence independently, which represents a **causal structure over the symbols** in the input sequence.
- **An equivalence class of the causal structure** over the input can be learned solely from the Transformer's estimated attention matrix.
- This enables learning the **causal structure over a single input sequence**, using existing constraint-based algorithms.

Causal: $X = (\mathbf{I} - \mathbf{G})^{-1}\Lambda U. \quad X = \mathbf{G}X + \Lambda U$

Attention: $\mathbf{A}\mathbf{A}^{\top} = ((\mathbf{I} - \mathbf{G})^{-1}\Lambda)\mathbf{C}_U((\mathbf{I} - \mathbf{G})^{-1}\Lambda)^{\top}.$

$\tilde{\mathbf{B}} = \mathbf{B}\mathbf{P}\Lambda,$ ⇒ Attention Matrix = Causal Effect * $\mathbf{P}\Lambda$

$\mathbf{A}\mathbf{A}^{\mathrm{T}} = \left((\mathbf{I} - \mathbf{G})^{-1}\Lambda\right)\mathbf{C}_U((\mathbf{I} - \mathbf{G})^{-1}\Lambda)^{\mathrm{T}}$

$\mathbf{A}\mathbf{A}^{\mathrm{T}} = \left((\mathbf{I} - \mathbf{G})^{-1}\Lambda\right) \cdot \mathbf{I} \cdot ((\mathbf{I} - \mathbf{G})^{-1}\Lambda)^{\mathrm{T}}$ The covariance are absorbed in Λ.

$\mathbf{A}\mathbf{A}^{\mathrm{T}} = \left((\mathbf{I} - \mathbf{G})^{-1}\right) \cdot \mathbf{C}_U \cdot ((\mathbf{I} - \mathbf{G})^{-1})^{\mathrm{T}}$ The scaling are absorbed in $\mathbf{C}_U$.

$\mathbf{B} = (\mathbf{I} - \mathbf{A})^{-1}$ ⇒ Causal Effect Matrix.

- Rohekar, Raanan Y., Yaniv Gurwicz, and Shami Nisimov. "Causal Interpretation of Self-Attention in Pre-Trained Transformers." Advances in Neural Information Processing Systems 36 (2024).

2. 因果Transformer架构

- 由基于自回归的Transformer到基于因果回归/发现的因果Transformer

原始数据公式转换:

$$\boldsymbol{X} = \mathbf{G}\boldsymbol{X} + \Lambda\boldsymbol{U}$$

$$\boldsymbol{X} = \mathbf{B}\Lambda\boldsymbol{U} = (\mathbf{I} - \mathbf{G})^{-1}\Lambda\boldsymbol{U}$$

$\mathbf{B} = (\mathbf{I} - \mathbf{G})^{-1}$

求协方差

$$\begin{aligned}\mathbf{C}_{\boldsymbol{X}} &= \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top] = \\ &= \mathbb{E}[(\mathbf{I} - \mathbf{G})^{-1}(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})^\top((\mathbf{I} - \mathbf{G})^{-1})^\top] \\ &= (\mathbf{I} - \mathbf{G})^{-1}\mathbb{E}[(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})^\top]((\mathbf{I} - \mathbf{G})^{-1})^\top \\ &= (\mathbf{I} - \mathbf{G})^{-1}\mathbf{C}_{\boldsymbol{U}}((\mathbf{I} - \mathbf{G})^{-1})^\top = \mathbf{B}\mathbf{C}_{\boldsymbol{U}}\mathbf{B}^\top,\end{aligned}$$

Attention公式转换:

$$\mathbf{Z}(\cdot, j) = \mathbf{Q}(\cdot, j)\mathbf{K}(\cdot, j)\mathbf{V}(\cdot, j) = \mathbf{A}\mathbf{V}(\cdot, j)$$

求协方差

$$\begin{aligned}\mathbf{C}_{\boldsymbol{Z}_j} &= \mathbb{E}[(\boldsymbol{Z}_j - \boldsymbol{\mu}_{\boldsymbol{Z}_j})(\boldsymbol{Z}_j - \boldsymbol{\mu}_{\boldsymbol{Z}_j})^\top] \\ &= \mathbf{A}\mathbb{E}[(\boldsymbol{V}_j - \boldsymbol{\mu}_{\boldsymbol{V}_j})(\boldsymbol{V}_j - \boldsymbol{\mu}_{\boldsymbol{V}_j})^\top]\mathbf{A}^\top = \\ &= \mathbf{A}\mathbf{C}_{\boldsymbol{V}_j}\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top,\end{aligned}$$

Causal & Attention: $\boldsymbol{X} = (\mathbf{I} - \mathbf{G})^{-1}\Lambda\boldsymbol{U}.$

$$\mathbf{A}\mathbf{A}^\top = ((\mathbf{I} - \mathbf{G})^{-1}\Lambda)\mathbf{C}_{\boldsymbol{U}}((\mathbf{I} - \mathbf{G})^{-1}\Lambda)^\top$$

令$\mathbf{C}_X$=$\mathbf{C}_Z$ **(Self-Attention)**

$\mathbf{A}=\mathbf{B}\Lambda\mathbf{P}$ → **Attention Matrix = Causal Effect * $\Lambda\mathbf{P}$**

$$\mathbf{A}\mathbf{A}^\mathrm{T} = \left((\mathbf{I} - \mathbf{G})^{-1}\Lambda\right)\mathbf{C}_{\boldsymbol{U}}((\mathbf{I} - \mathbf{G})^{-1}\Lambda)^\mathrm{T}$$

$$\mathbf{A}\mathbf{A}^\mathrm{T} = \left((\mathbf{I} - \mathbf{G})^{-1}\Lambda\right) \cdot \mathbf{I} \cdot ((\mathbf{I} - \mathbf{G})^{-1}\Lambda)^\mathrm{T}$$

$$\mathbf{A}\mathbf{A}^\mathrm{T} = \mathbf{B}\Lambda\mathbf{P}\mathbf{P}^\mathrm{T}(\mathbf{B}\Lambda)^\mathrm{T}$$

$$\mathbf{A}\mathbf{A}^\mathrm{T} = \left((\mathbf{I} - \mathbf{G})^{-1}\right) \cdot \mathbf{C}_{\boldsymbol{U}} \cdot ((\mathbf{I} - \mathbf{G})^{-1})^\mathrm{T}$$

The scaling are absorbed in $\mathbf{C}_{\boldsymbol{U}}$.

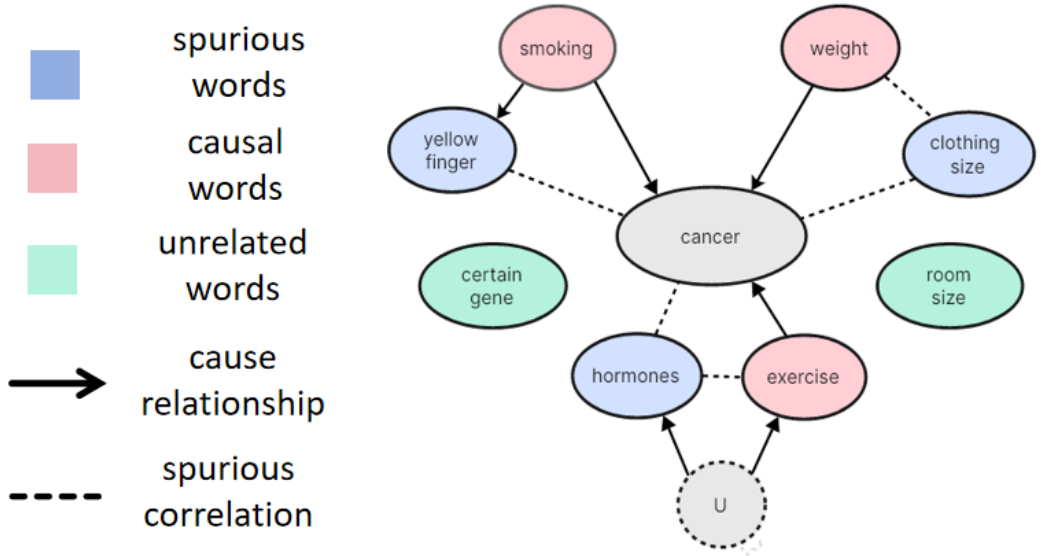
$\mathbf{B} = (\mathbf{I} - \mathbf{G})^{-1}$ → **Causal Effect Matrix.**

- Rohekar, Raanan Y., Yaniv Gurwicz, and Shami Nisimov. "Causal Interpretation of Self-Attention in Pre-Trained Transformers." Advances in Neural Information Processing Systems 36 (2024).

3. 基于因果知识增强的Transformer架构

问题: 为了探索虚假相关性对LLM推理的影响，我们首先通过仿真实验，在合成数据集上进行实验，提出Spurious Token Game（STG）基准。

Dataset: STG task | Answer: High Risk			Model Predicted
Setting: IID	Vanilla	Here is the statistical data for a person. Please predict the risk of cancer. Weight: 10, Certain gene: 1, Room size: 7, Yellow fingers: 13, Smoking: 9, Cloth size: 10, Hormones: 6, Exercise: 3.	High Risk ✓
	CAT	Here is the statistical data for a person. Please predict the risk of cancer. Weight: 10, Certain gene: 1, Room size: 7, Yellow fingers: 13, Smoking: 9, Cloth size: 10, Hormones: 6, Exercise: 3.	High Risk ✓
Setting: OOD	Vanilla	Here is the statistical data for a person. Please predict the risk of cancer. Weight: 10, Certain gene: 1, Room size: 7, Yellow fingers: 13, Smoking: 9, Cloth size: 2, Hormones: 6, Exercise: 3.	Low Risk ✗
	CAT	Here is the statistical data for a person. Please predict the risk of cancer. Weight: 10, Certain gene: 1, Room size: 7, Yellow fingers: 13, Smoking: 9, Cloth size: 2, Hormones: 6, Exercise: 3.	High Risk ✓



以吸烟肺癌黄手指的经典因果图，通过验证LLM能否自发利用因果知识，有效的预测给定属性和数值后的癌症风险

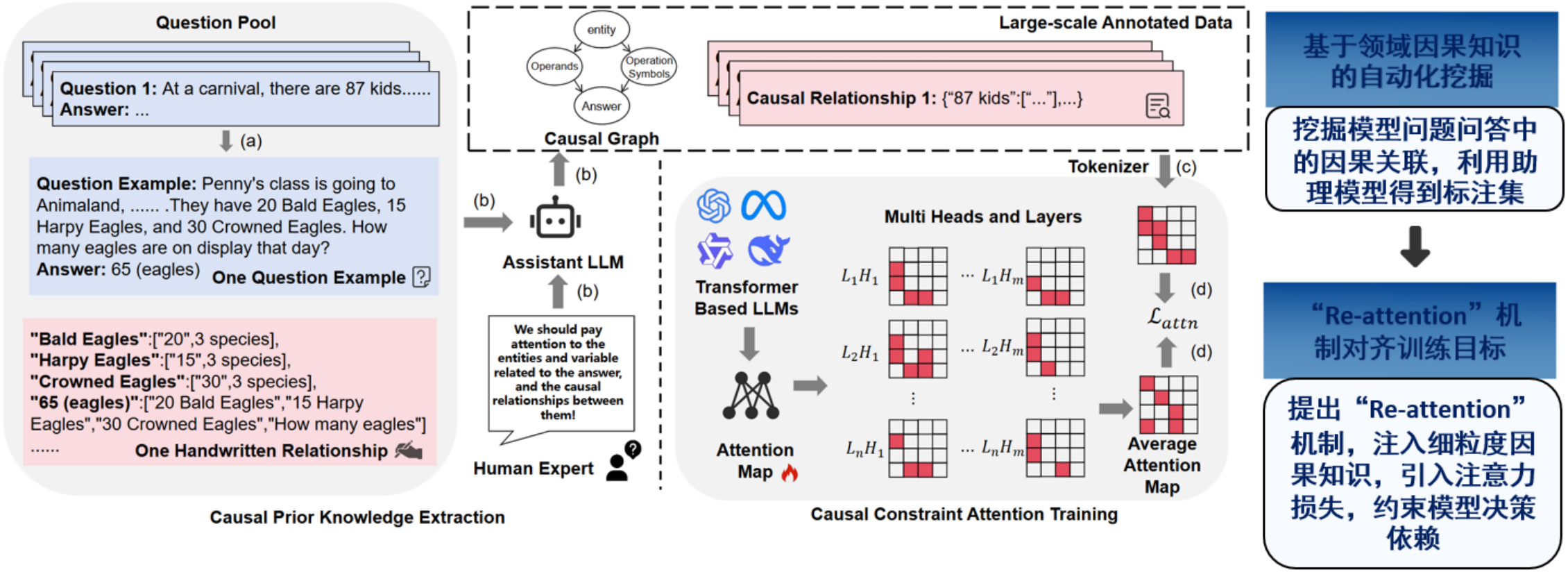
结论：LLM无法有效利用因果知识，错误的关注了虚假关联词。（红色越深表示attention分数越高），当Cloth size（虚假关联词）改变时，尽管理论上不会影响患癌风险，但LLM错误的改变了输出

Han, K., Zhao, W., Zhao, Z., Ye, J., Pan, L., & Kuang, K. (2025). CAT: Causal Attention Tuning For Injecting Fine-grained Causal Knowledge into Large Language Models. EMNLP2025.

3. 基于因果知识增强的Transformer架构

挑战：如何建模LLM token级别的因果关联，如何将因果知识融入Transformer架构

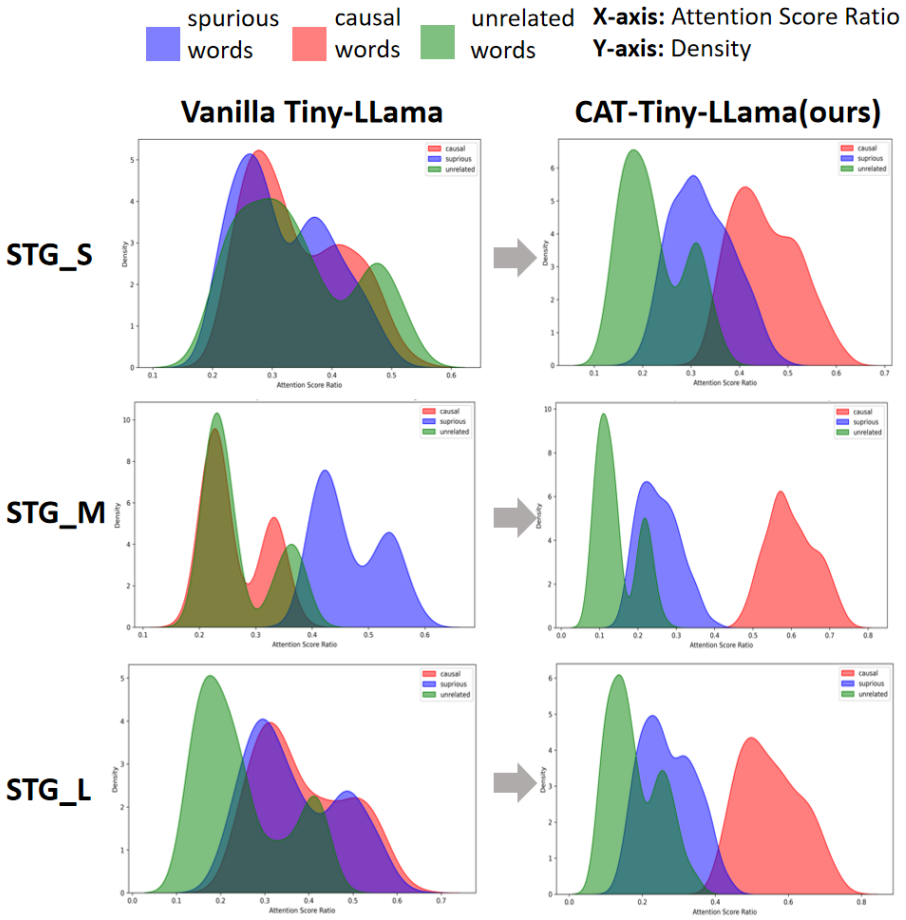
想法：（1）LLM赋能因果，发现复杂数据中token级别因果关系；（2）基于细粒度因果知识的注意力微调技术，提出因果知识约束的Re-attention机制。



Han, K., Zhao, W., Zhao, Z., Ye, J., Pan, L., & Kuang, K. (2025). CAT: Causal Attention Tuning For Injecting Fine-grained Causal Knowledge into Large Language Models. EMNLP2025.

3. 基于因果知识增强的Transformer架构

实验结果：STG数据集上，我们的方法有效关注到了因果token，实现模型的性能提升



Model	Setting	Task	Method	STG_S	STG_M	STG_L
TinyLlama-1.1B	Full	IID	Vanilla	62.25%	89.50%	95.50%
			CAT	**76.00%**	**94.00%**	**96.75%**
		OOD	Vanilla	53.50%	60.75%	65.25%
			CAT	**64.25%**	**73.00%**	**77.00%**
	LoRA	IID	Vanilla	62.75%	83.50%	96.00%
			CAT	**81.50%**	**89.75%**	**97.25%**
		OOD	Vanilla	59.25%	56.75%	61.50%
			CAT	**65.50%**	**63.50%**	**69.50%**
Qwen-1.5B	Full	IID	Vanilla	55.00%	94.50%	95.50%
			CAT	**74.00%**	**94.50%**	**96.00%**
		OOD	Vanilla	53.50%	79.00%	79.75%
			CAT	**64.75%**	**79.00%**	**83.25%**
	LoRA	IID	Vanilla	81.50%	93.25%	95.75%
			CAT	**82.00%**	**93.75%**	**95.75%**
		OOD	Vanilla	78.50%	82.00%	82.00%
			CAT	**78.50%**	**88.00%**	**90.50%**
LLaMA-7B	LoRA	IID	Vanilla	78.00%	95.25%	97.25%
			CAT	**86.00%**	**95.75%**	**97.50%**
		OOD	Vanilla	68.00%	90.25%	87.75%
			CAT	**69.00%**	**90.25%**	**94.25%**

Han, K., Zhao, W., Zhao, Z., Ye, J., Pan, L., & Kuang, K. (2025). CAT: Causal Attention Tuning For Injecting Fine-grained Causal Knowledge into Large Language Models. EMNLP2025.

3. 基于因果知识增强的Transformer架构

实验结果：六个通用任务进行测试：GSM8K，ARC-E，ASDiv，SVAMP，MAWPS, Date.

Model	Setting	Method	Date	MAWPS	ASDiv	SVAMP	GSM8K	ARC-E	Avg.
TinyLlama-1.1B	Full	Vanilla	20.55%	42.37%	25.51%	23.00%	10.08%	29.00%	25.09%
		CAT	**30.14%**	**46.25%**	**26.64%**	**23.50%**	**12.05%**	**30.05%**	**28.11%**
		impv.	+9.59%	+3.88%	+1.13%	+0.50%	+1.97%	+1.05%	+3.02%
	LoRA	Vanilla	19.18%	43.10%	21.22%	6.00%	9.40%	24.07%	20.50%
		CAT	**27.40%**	**45.04%**	**23.70%**	**7.00%**	**10.16%**	**24.12%**	**22.90%**
		impv.	+8.22%	+1.94%	+2.48%	+1.00%	+0.76%	+0.05%	+2.40%
Qwen-1.5B	Full	Vanilla	28.77%	69.98%	58.24%	62.50%	47.84%	89.73%	59.51%
		CAT	**28.77%**	**73.85%**	**59.14%**	**63.00%**	**48.45%**	**89.90%**	**60.52%**
		impv.	+0.00%	+3.87%	+0.90%	+0.50%	+0.61%	+0.17%	+1.01%
	LoRA	Vanilla	32.88%	74.33%	58.92%	50.50%	47.23%	89.14%	58.83%
		CAT	**36.99%**	**75.30%**	**58.92%**	**54.00%**	**47.54%**	**89.44%**	**60.37%**
		impv.	+4.11%	+0.97%	+0.00%	+3.50%	+0.31%	+0.30%	+1.54%
Llama-7B	LoRA	Vanilla	46.58%	64.16%	50.11%	55.50%	37.83%	80.81%	55.83%
		CAT	**52.05%**	**65.38%**	**52.14%**	**58.00%**	**39.42%**	**81.52%**	**58.09%**
		impv.	+5.47%	+1.22%	+2.03%	+2.50%	+1.59%	+0.71%	+2.26%

1. 提出Spurious Token Game（STG）基准，探究虚假关联对LLM的影响
2. 提出CAT注意力微调方法和重注意机制
3. CAT在多个模型性能上得到一致性提升，验证了方法有效性。

Han, K., Zhao, W., Zhao, Z., Ye, J., Pan, L., & Kuang, K. (2025). CAT: Causal Attention Tuning For Injecting Fine-grained Causal Knowledge into Large Language Models. EMNLP2025.

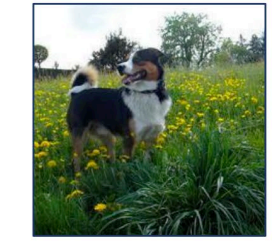
因果去除数据偏置

① 虚假关联问题

② 灾难遗忘问题

1. 多模态大模型中虚假关联问题

- 多模态大模型大多使用CLIP-based model作为Image Encoder，由于CLIP Image encoder在预训练中仅使用实例的对比损失，进而带来幻觉等问题。



Q: Is the dog facing left or right from the camera's perspective?

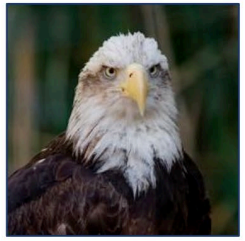
The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.

Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.

Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.

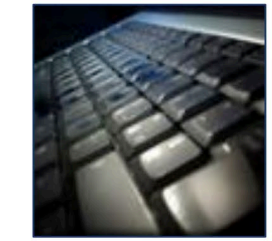


Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.



Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.

Q: From the camera's perspective, is the piano's back panel on the left or right side?

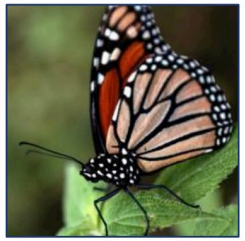
From the camera's perspective, the piano's back panel is **on the left side**. The keys visible in the foreground suggest that we looking at the piano from front-right side, thus placing the back panel on the left from this angle.

Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.

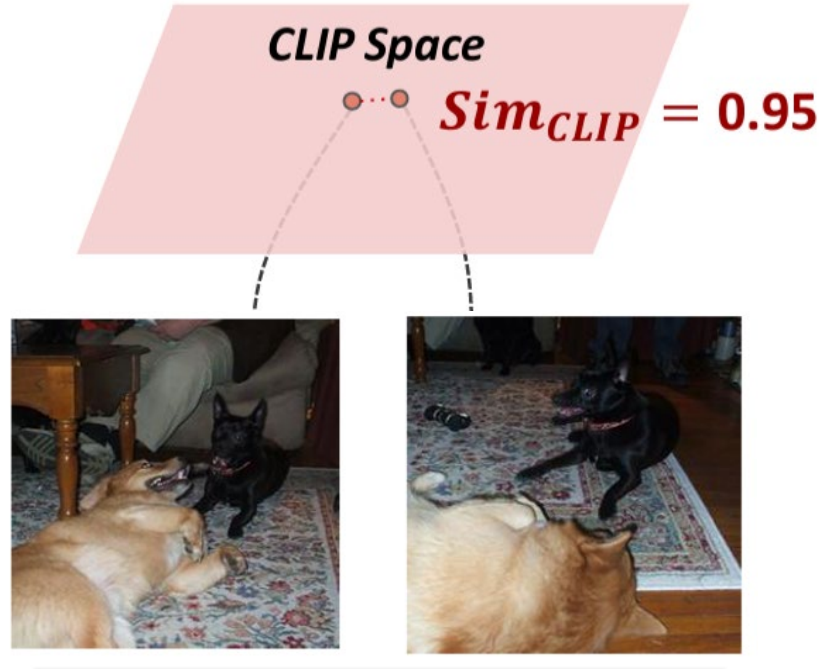
Q: Is the school bus parked facing the camera or away from the camera?

The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.



Q: Are the butterfly's feet visible?

The butterfly's feet are **not visible** in this image. The wings and part of the body of the butterfly are in view, but the feet are either hidden beneath the body or out of the frame of the picture. Butterflies have delicate, slender feet which are often not the most noticeable feature when they are perched, especially from the angle shown in this image.



Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, CVPR 2024.

1. 多模态大模型中虚假关联问题

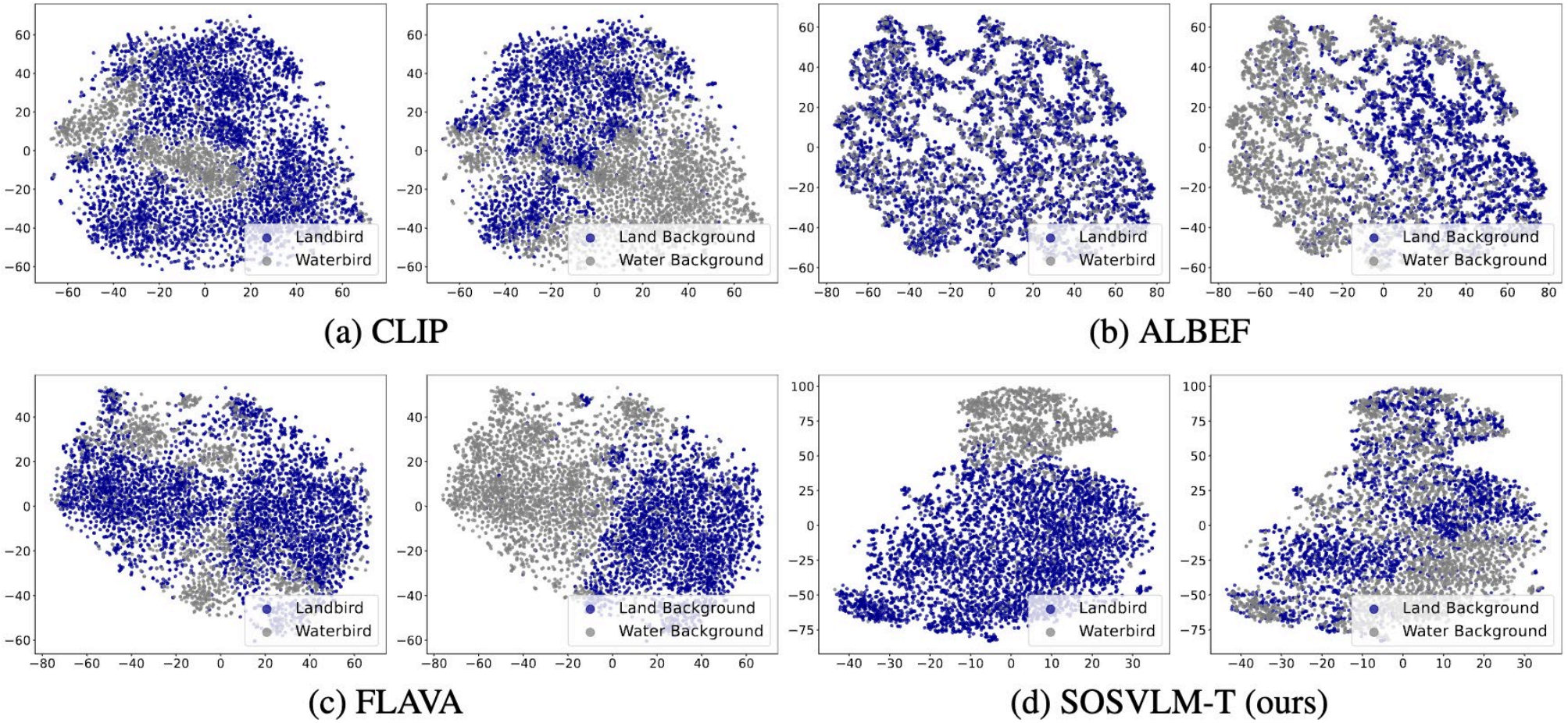
- 所有的视觉语言预训练VLMs方法，在可泛化问题上均出现不同程度的表现下降。尤其是CLIP模型在虚假问题上，下降接近60个点。

Method	Backbone	Waterbirds			CelebA		
		WG (↑)	AVG (↑)	**Gap (↓)**	WG (↑)	AVG (↑)	**Gap (↓)**
CLIP-0.4B [26]	ViT-B/16	22.74	79.24	56.50	68.89	88.36	19.47
CoCa-2.1B [42]	ViT-B/32	27.36	50.79	23.43	27.77	92.68	64.91
ALBEF-0.2B [18]	ViT-B/16	11.71	48.71	37.00	41.67	59.49	17.82
FLAVA-0.3B [31]	ViT-B/32	26.7	56.58	29.88	0.55	86.32	85.77
BLIP-0.2B [17]	ViT-B/16	29.53	57.21	27.68	40.56	85.38	44.82
ALIGN-0.8B [10]	EfficientNet-B7	47.35	69.83	22.48	31.66	60.69	29.03
AltCLIP-0.8B [3]	ViT-L/14	34.89	81.68	46.79	24.19	88.53	64.34
BEiT-3-0.7B [36]	ViT-Base	40.57	65.08	24.51	64.52	75.52	11.00

SOSVLM: Exploring and Mitigating Spurious Correlations in Vision-Language Model. Working paper.

1. 多模态大模型中虚假关联问题

- 当前主流多模态大模型视觉表征往往学到的是虚假关联（如背景特征等）



SOSVLM: Exploring and Mitigating Spurious Correlations in Vision-Language Model. Working paper.

1. 多模态大模型中虚假关联问题

- 由于CLIP Image encoder在预训练中仅使用实例的对比损失，进而导致虚假关联。
- 思路：结合前景和背景进行数据增强，基于 Image Encoder 参数微调

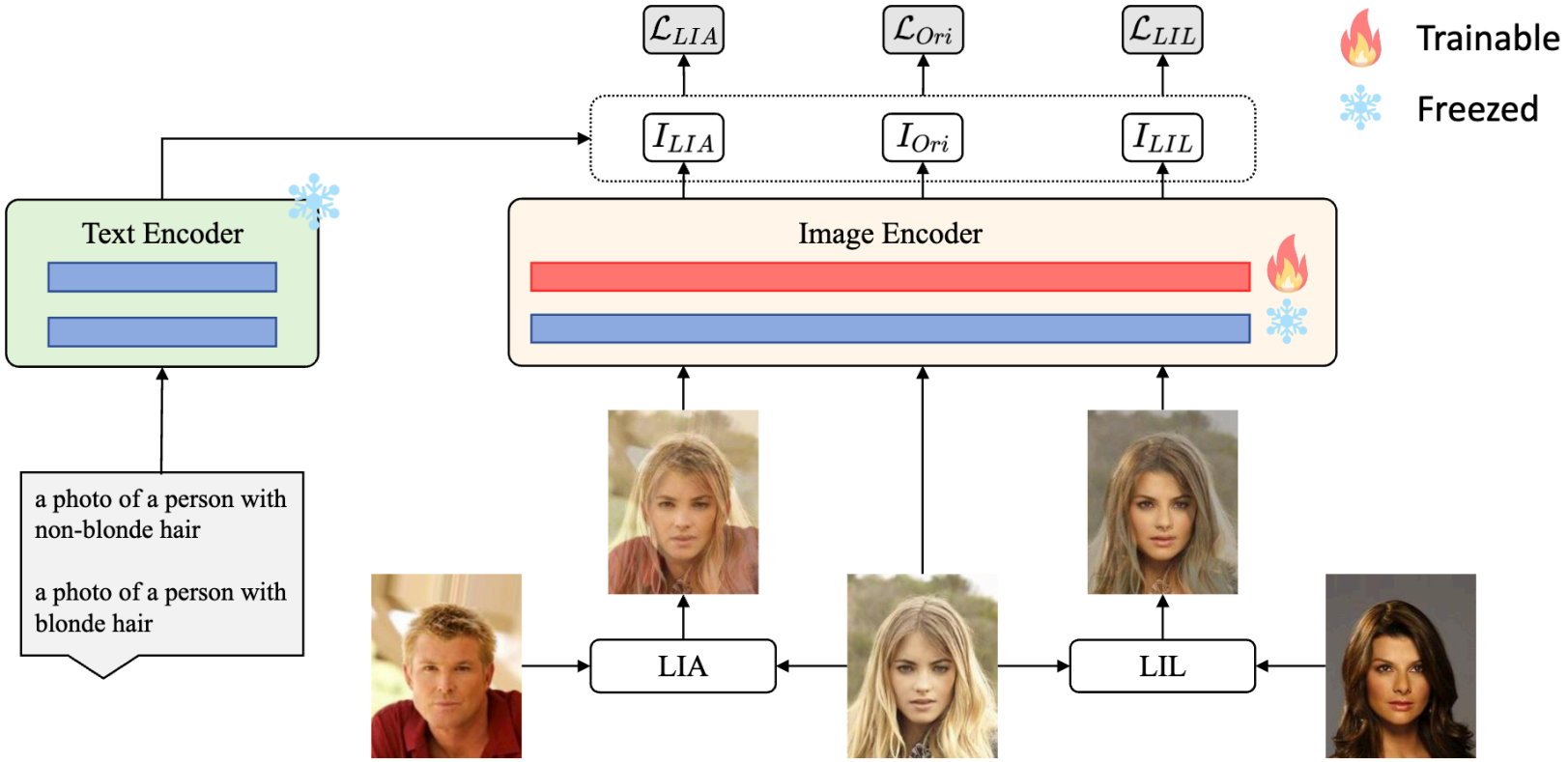


Figure 2: Overall framework.

SOSVLM: Exploring and Mitigating Spurious Correlations in Vision-Language Model. Working paper.

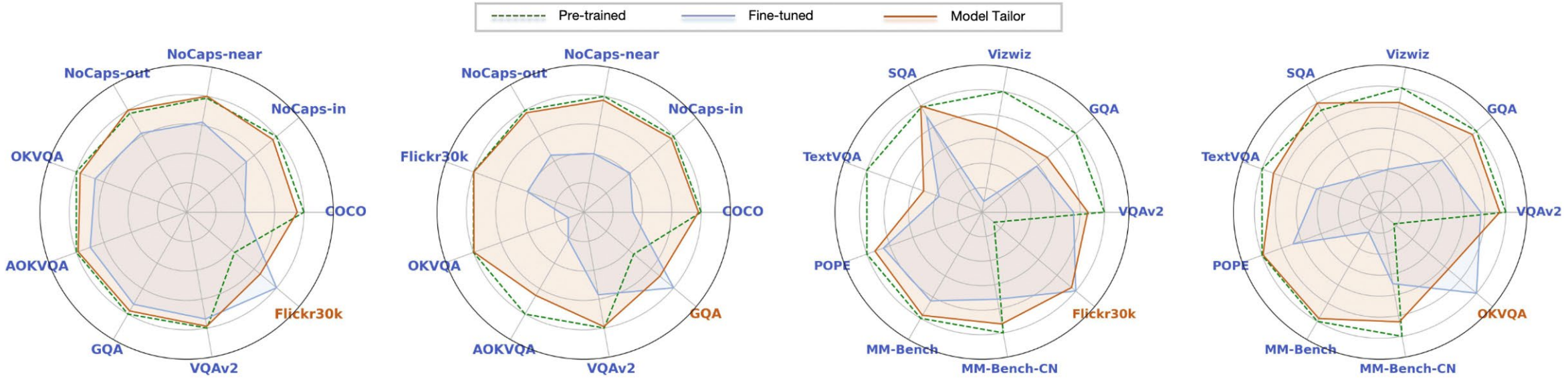
1. 多模态大模型中虚假关联问题

- 相较于SOTA的方法，可以有效缓解VLMs模型的虚假相关问题。

	Waterbirds			CelebA		
	WG (↑)	AVG (↑)	Gap (↓)	WG (↑)	AVG (↑)	Gap (↓)
ERM Linear Probe [15]	65.4	97.7	32.3	30.4	94.6	64.2
ERM Adapter [8]	76.1	97.8	21.7	40.0	94.3	54.3
DFR (Subsample) [13]	58.8	95.9	37.1	78.7	91.8	13.1
DFR (Upsample) [13]	66.5	96.4	29.9	83.9	91.2	7.3
WiSE-FT [37]	65.9	97.6	31.7	80.0	87.4	7.4
Contrastive Adapter [44]	86.9	96.2	9.3	84.6	90.4	5.8
FairerCLIP [6]	86.0	92.2	6.2	85.2	87.8	2.6
DPS+RNS [41]	88.2	96.8	8.6	84.8	87.8	3.0
SOSVLM-T+Tuning last layers	**90.6**	92.6	2.0	**89.3**	91.1	**1.8**
SOSVLM-T+CLIP Adapter	<u>89.0</u>	89.5	**0.5**	<u>87.4</u>	89.8	<u>2.4</u>
SOSVLM-T+Prompt tuning	87.9	89.6	<u>1.7</u>	87.0	90.1	3.1

SOSVLM: Exploring and Mitigating Spurious Correlations in Vision-Language Model. Working paper.

2. 多模态大语言模型中灾难遗忘问题



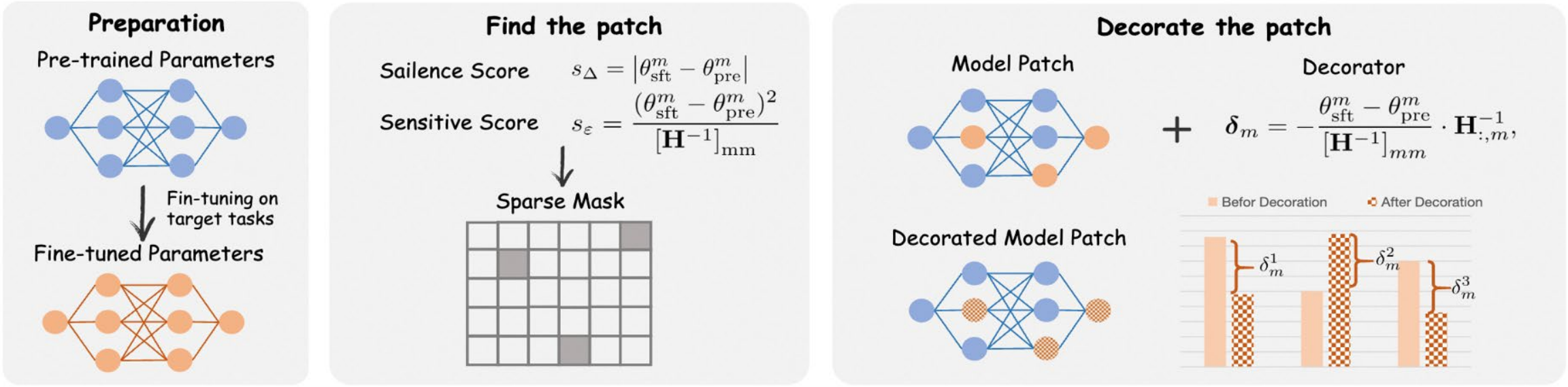
(a) InstructBLIP on Flickr30k (b) InstructBLIP on GQA (c) LLaVA on Flickr30k (d) LLaVA on OKVQA

多模态大语言模型中的灾难性遗忘现象

- 如图中pretrain阶段（绿色线段）到finetune阶段（蓝色线段）所示，在InstructBLIP和LLaVa1.5两个经典多模态大语言模型上，对未见过的目标任务(蓝色方框)进行微调后，会对原始的若干任务造成灾难性遗忘现象。
- 如图中finetune阶段（蓝色线段）到Model Tailor（橙色阶段）所示，我们的方法致力于提升MLLM在目标任务上的性能的同时，保证其在原始任务上的性能。

Didi Zhu, Zhongyisun Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, Kun Kuang*. Model Tailor: Mitigating Catastrophic Forgetting in Multi-modal Large Language Models, ICML, 2024.

2. 多模态大语言模型中灾难遗忘问题



Model Tailor总览：

给定预训练模型参数和微调模型参数，执行以下两个步骤

步骤1：识别稳定的模型补丁。识别微调模型参数中对目标任务性能稳定关键的参数子集，该部分参数子集被成为“模型补丁”。

步骤2：对模型补丁进行装饰。在确定了补丁后，对选定的子集进行有针对性的补偿。旨在缓解因去除其他未选中的微调参数而导致目标任务上损失增加的问题。这一补偿步骤被称为“补丁装饰器”。

上述两个步骤可以被建模为：$\boldsymbol{\Theta}_{\text{fusion}} = \mathcal{F}(\boldsymbol{\Theta}_{\text{sft}}, \boldsymbol{\Theta}_{\text{pre}}) = \boldsymbol{M} \odot (\boldsymbol{\Theta}_{\text{sft}} + \boldsymbol{C}) + (\boldsymbol{I} - \boldsymbol{M}) \odot \boldsymbol{\Theta}_{\text{pre}}$,

Didi Zhu, Zhongyisun Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, Kun Kuang*. Model Tailor: Mitigating Catastrophic Forgetting in Multi-modal Large Language Models, ICML, 2024.

因果支撑决策

① 用户偏好对齐

② 因果强化学习

Preference Optimization（RLHF）主流方法：PPO vs DPO

大模型训练的三个阶段：

1. **Pretraining**：利用海量无标注文本，拟合人类世界文本的分布、学习知识。
2. **SFT**：利用prompt-response的有监督文本，提高模型响应指令的能力。
3. **Preference Optimization**：利用prompt, preferred response (y_w)和dis-preferred response (y_l)，实现模型的对齐。

基于强化学习（Proximal Policy Optimization）

第一步：构造reward function $r_\phi(x, y)$，对y_w优于y_l的概率进行建模，利用NLL Loss优化r_ϕ最大化$P_{BT}(y_w \succ y_l|x)$.

$$P_{BT}(y_w \succ y_l|x) = \frac{\exp r_\phi(x, y_w)}{\exp r_\phi(x, y_w) + \exp r_\phi(x, y_l)}$$

第二步：利用强化学习方法（PPO），在π_{ref}的约束下（通常是SFT后的LLM）最优化LLM π_θ:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(\cdot|x)}\left[r_\phi(x, y)\right] - \beta \cdot \text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})$$

DPO (Direct Preference Optimization)

第一步：在PPO的优化目标下，最优解π_θ^*与reward function r ——对应（$Z(x)$用于配平概率分布）：

$$\pi_\theta^*(y|x) = \frac{1}{Z(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp\frac{1}{\beta} \cdot r(x, y)$$

$$r(x, y) = \beta \cdot \log\frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \cdot \log Z(x)$$

第二步：利用反推出的$r(x, y)$，我们可以通过对$P_{BT}(y_w \succ y_l|x)$进行最大似然估计来最优化LLM π_θ.

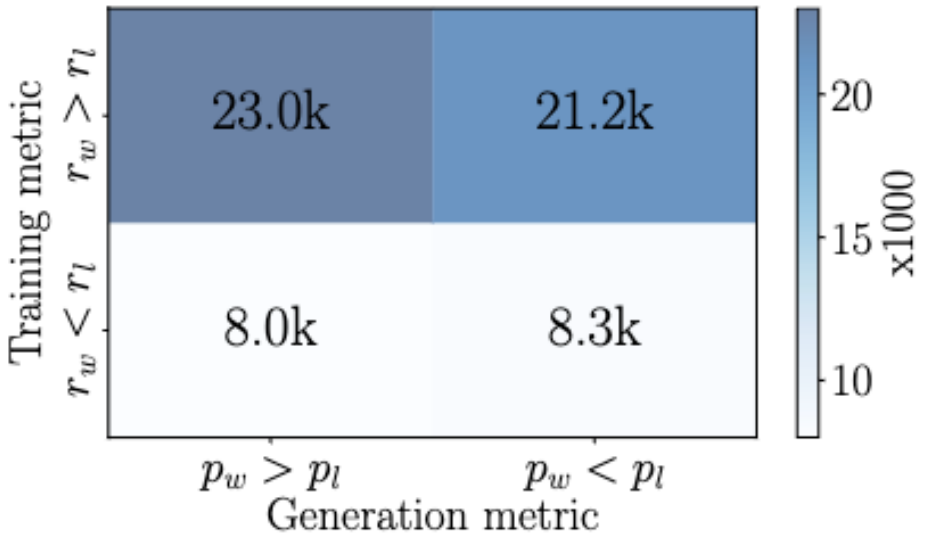
Preference Optimization（RLHF）主流方法：PPO vs DPO

$$P_{BT}(y_w \succ y_l|x) = \frac{\exp r_\phi(x, y_w)}{\exp r_\phi(x, y_w) + \exp r_\phi(x, y_l)}$$

优化目标和 Preference Alignment 并不一致：

$\boldsymbol{p_w > p_l}$ (概率上好) $\neq \boldsymbol{r_w > r_l}$（评分上好）

事实上，文本的好坏我们有大量的Direct Outcome Dataset进行标注 （如Imdb, Toxic等），我们可以利用这些数据集来实现Preference Optimization！



(SimPO, 2024, Danqi Chen)

因果赋能大语言模型用户偏好对齐

- 假设对于任意一个文本x，我们都能知道所有人G对这个文本的打分$Y(x)$（是否符合ta的取向），我们把这个分数的数学期望定义为$g(x) \coloneqq \mathbb{E}_{Y(\cdot)\sim G}[Y(x)]$，那么我们可以利用$g(x)$精准地将LLM f与人类的取向对齐：

$$\arg\max_f V(f) \equiv \arg\max_f \mathbb{E}_{X\sim Pf}[g(x)]$$

- 一般来说，所有人的打分很难得到，所以我们在一般的Direct Outcome Dataset D_O上优化会因为selection bias或unobserved confounder而有偏。

$$V(f) \neq \mathbb{E}_{X\sim Pf}[\mathbb{E}_{D_O}[Y|X]]$$

- 假如对Direct Outcome Dataset D_O作出一些限制，能否得到 $V(f)$ 的一个无偏估计？

将大语言模型和人类的取向对齐实际上是一个Causal Inference问题

Lin V, Ben-Michael E, Morency L P. Optimizing Language Models for Human Preferences is a Causal Inference Problem[J]. arXiv preprint arXiv:2402.14979, 2024.

因果赋能大语言模型用户偏好对齐

CPO (Causal Preference Optimization)

若数据集D^R是一个随机数据集（完全随机或收集来源足够充分，满足$\widehat{P^R}(X) = P^R(X)$）那么$\widehat{V_{IPW}}$是一个对$V_f$的无偏估计：

$$V_{IPW}(f) = \mathbb{E}_{X\sim P^R}\left[\mathbb{E}_{Y\sim P_y^R}\left[\frac{P^f(X)}{P^R(X)}\cdot Y\right]\right]$$

$$\widehat{V_{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{P^f(X_i)}{P^R(X_i)}\cdot Y_i\right]$$

IPW方法已经可以很好地对V_f进行估计，但是IPW存在方差较高的问题，我们可以通过对$g(x)$建模，以**Doubly Robust Estimation**的方法缓解此问题。

DR-CPO (Doubly Robust CPO)

将V_{IPW}分成两项得到V_{DR}：

$$V_{DR}(f) = \mathbb{E}_{X\sim P^R}\left[\mathbb{E}_{Y\sim P_y^R}\left[\frac{P^f(X)}{P^R(X)}\cdot(Y-g(X))\right]\right] + \mathbb{E}_{X\sim P^f}[g(X)]$$

第一项按CPO优化。

第二项难以优化，通过IPW固定f为f^0，再用蒙特卡洛采样文本$\widetilde{X_1},\cdots,\widetilde{X_m}\sim P^{f^0}$，计算$\widehat{V_{out}}$：

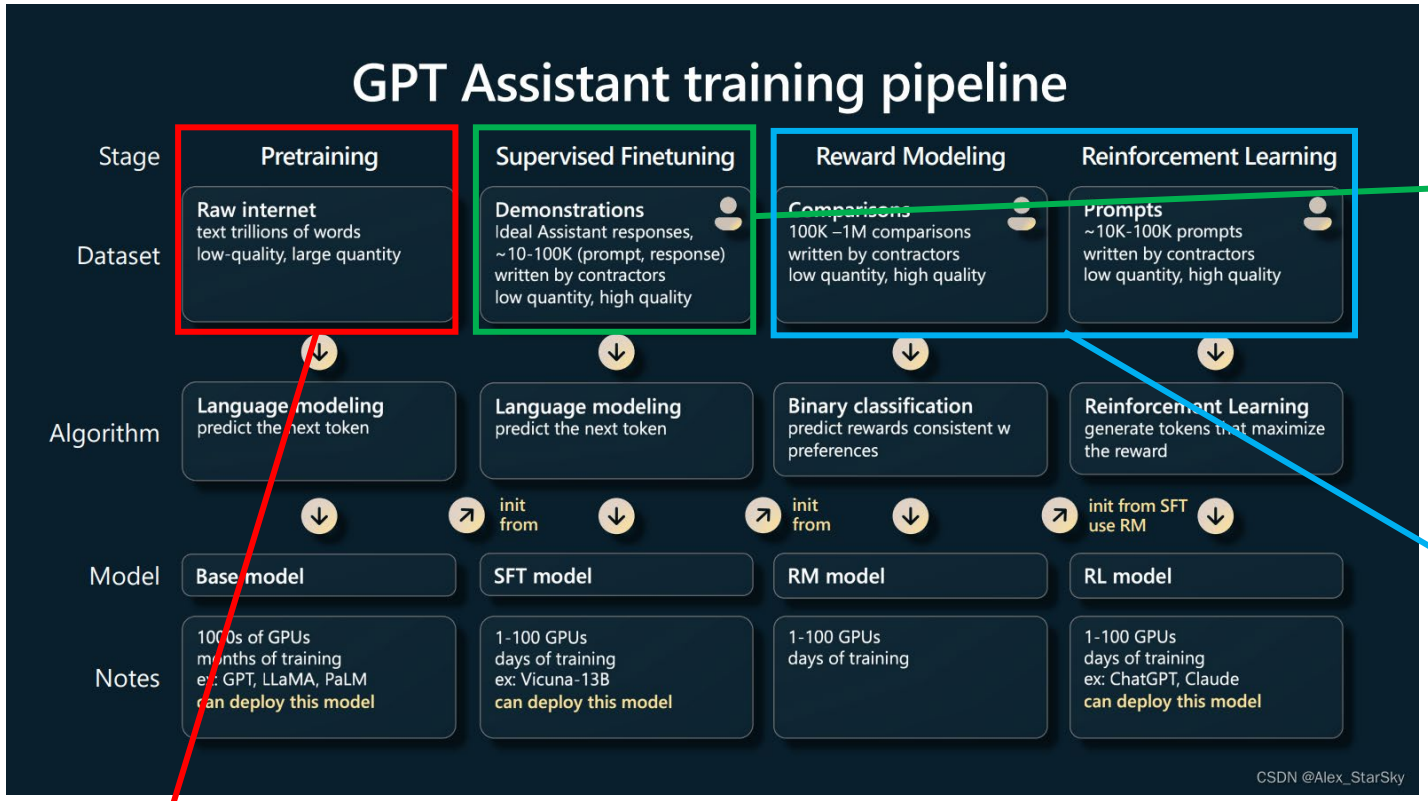
$$V_{out}(f) = \mathbb{E}_{X\sim P^f}[g(X)] = \mathbb{E}_{X\sim P^{f^0}}\left[\frac{P^f(X)}{P^{f^0}(X)}\cdot g(X)\right]$$

$$\widehat{V_{out}}(f) = \frac{1}{m}\sum_{j=1}^{m}\frac{P^f(\widetilde{X_j})}{P^{f^0}(\widetilde{X_j})}\cdot\hat{g}(\widetilde{X_j})$$

$$\widehat{V_{DR}}(f) = \frac{1}{n}\sum_{i=1}^{m}\frac{P^f(X_i)}{\widehat{P^R}(X_i)}\cdot(Y_i-\hat{g}(X_i)) + \widehat{V_{out}}(f)$$

若$\widehat{P^R}(X) = P^R(X)$或$\hat{g}(X) = g(X)$ **任意一项**成立，则$\widehat{V_{DR}}$是无偏的，可以证明$\widehat{V_{DR}}$比$\widehat{V_{IPW}}$的方差更小。

因果如何赋能大模型



- 因果去除数据偏置
  - 虚假相关问题
  - 灾难遗忘问题
- 因果支撑决策
  - 用户偏好对齐
  - 因果强化学习
- 因果赋能Transformer
  - 由关联自回归到因果回归机制
  - 因果Transformer架构
  - 基于因果知识增强的Transformer架构

Survey： Causality for LLMs

Causality for Large Language Models

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, Kun Zhang

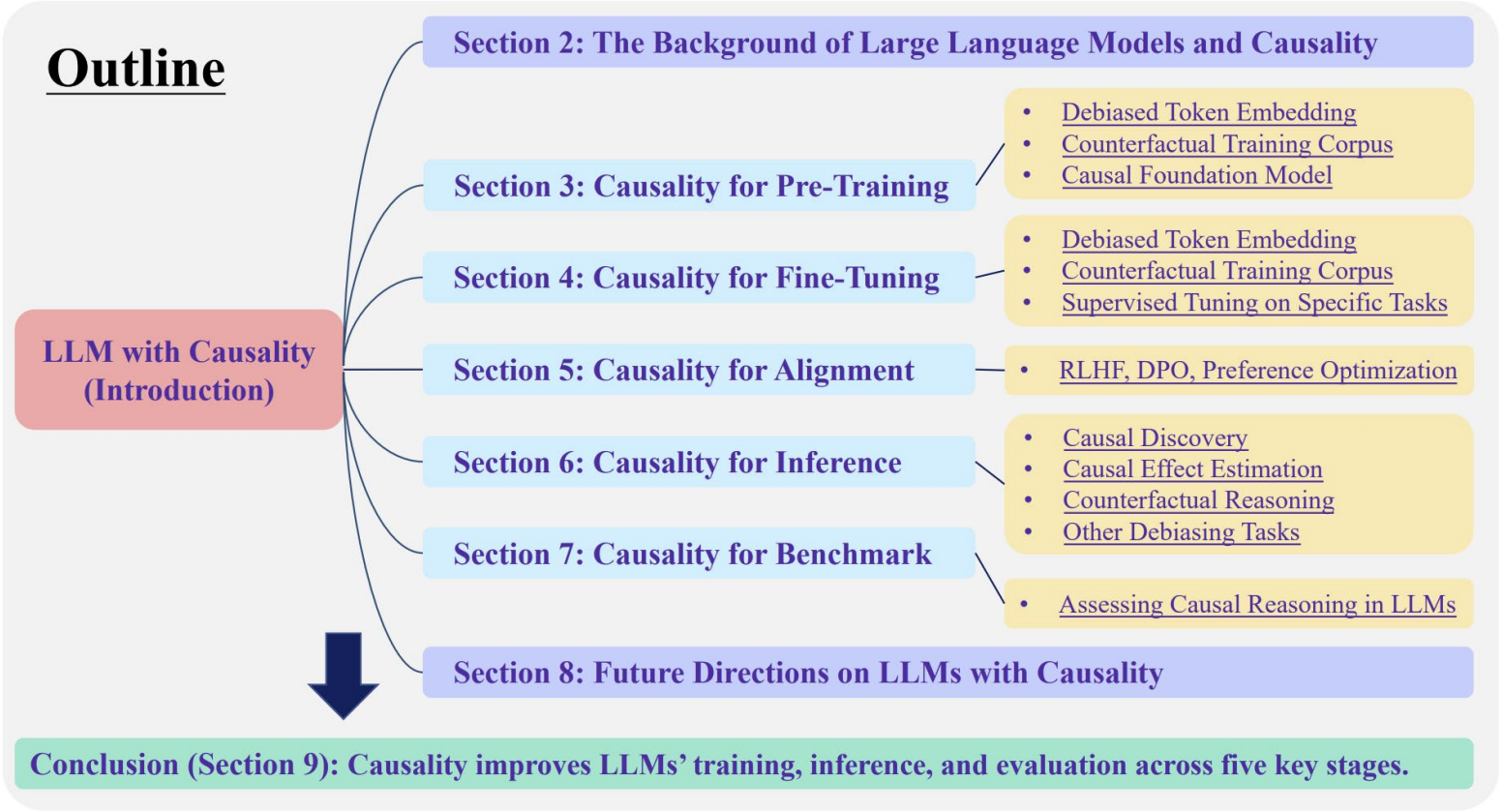
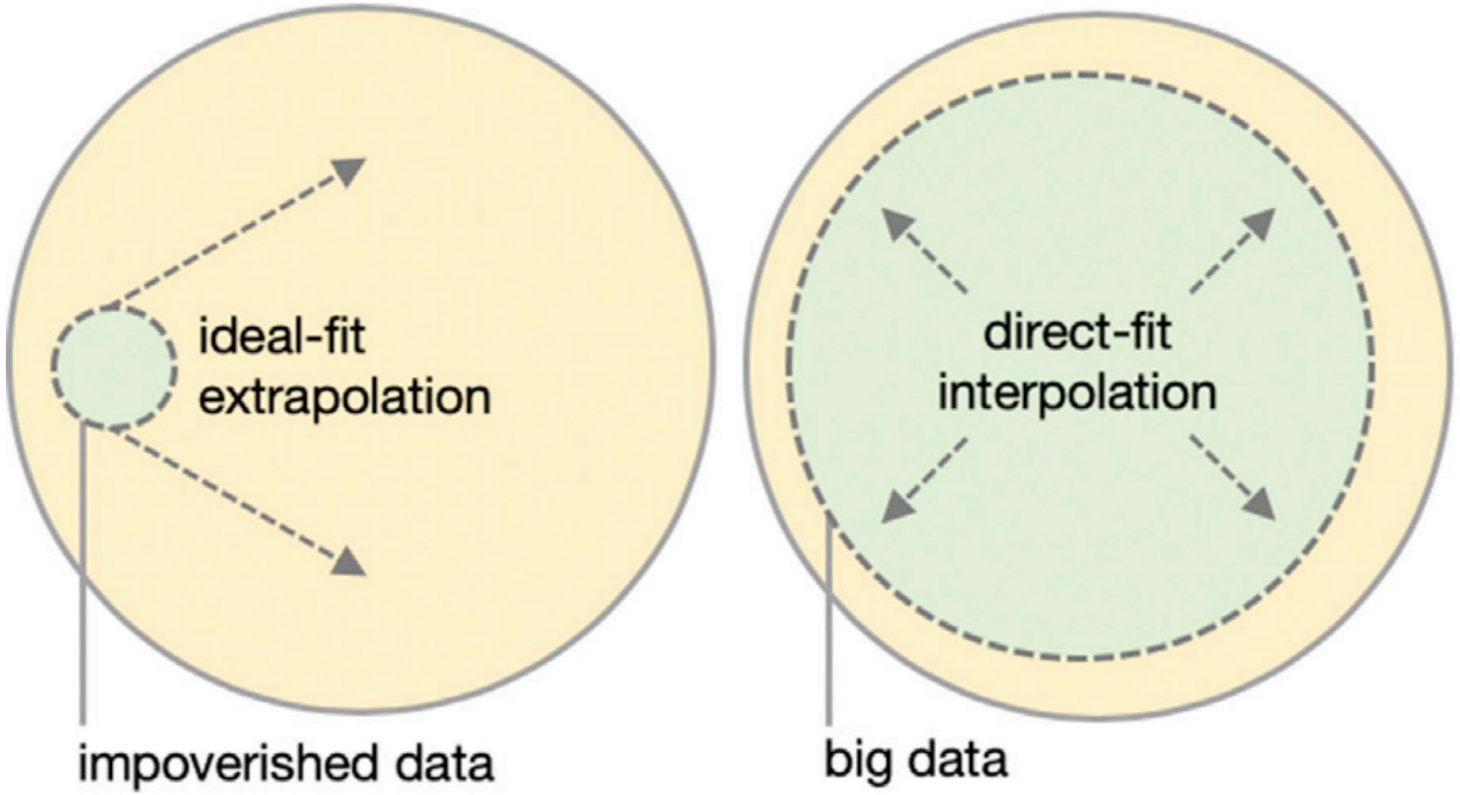


Figure 1: The Role of Causality in Enhancing LLMs: A Comprehensive Framework Across Development Stages

https://arxiv.org/abs/2410.15319

因果与大模型

因果：
以不变应万变



大模型：
见多识广

Uri Hasson et al. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron (2020)

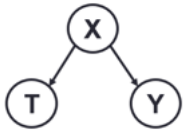
关联驱动 vs. 因果启发

■ 关联的三种来源

因果

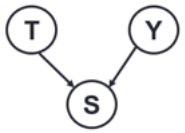
可解释
稳定/鲁棒
可决策

混淆偏差



虚假关联: T is correlated with Y ignoring X

选择偏差

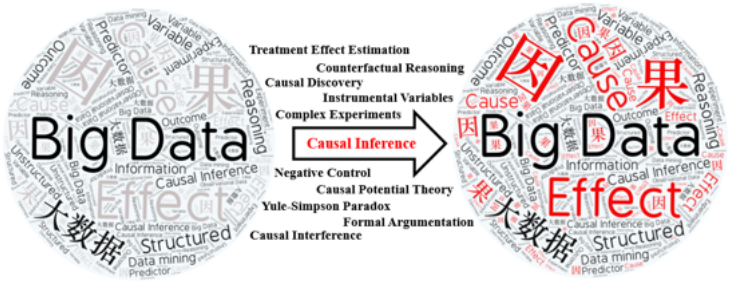


虚假相关: T is correlated with Y given S

机器学习赋能
因果推理

因果启发机器学习

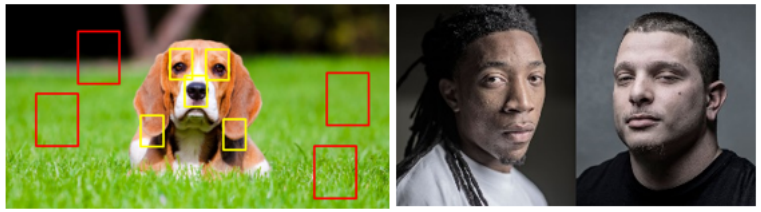
■ 大数据驱动因果推理



■ 因果表征学习

可解释性、稳定性

公平性、可决策性



机器学习赋能因果推理：从复杂数据关联中恢复因果关联
因果推理赋能人工智能：从数据关联驱动迈向因果启发学习

复杂偏差下因果推断（机器学习赋能因果推理）

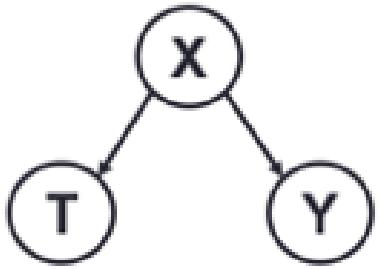
2021诺贝尔经济学奖：
基于工具变量的因果推断

2000诺贝尔经济学奖：
面向选择偏差的因果推断

因果

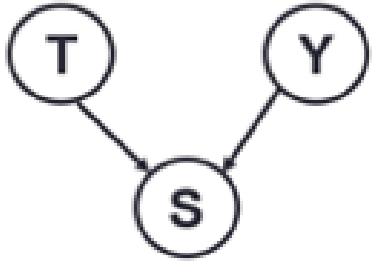
可解释
稳定/鲁棒
可决策

混淆偏差



虚假关联: T is correlated with Y ignoring X

选择偏差



虚假相关: T is correlated with Y given S

同时存在混淆偏差和选择偏差，（条件）工具变量

因果启发的可信机器学习（因果赋能机器学习）

发现了通过全局样本赋权实现因果关联挖掘的机制

PROPOSITION 3.3. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

$$\sum_{j=1}^{p}\left\|\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot\mathbf{X}_{\cdot,j})}{W^{T}\cdot\mathbf{X}_{\cdot,j}}-\frac{\mathbf{X}_{\cdot,-j}^{T}\cdot(W\odot(1-\mathbf{X}_{\cdot,j}))}{W^{T}\cdot(1-\mathbf{X}_{\cdot,j})}\right\|_2^2$$

↓

0

存在一组样本权重，使得任意输入变量与其他变量之间相互独立

PROOF. Since $\|\cdot\| \geq 0$, Eq. (8) can be simplified to $\forall j, \forall k \neq j$

$$\lim_{n\to\infty}\left(\frac{\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=1}} W_i}{\sum_i t\mathbf{X}_{i,j=1} W_i} - \frac{\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=0}} W_i}{\sum_i t\mathbf{X}_{i,j=0} W_i}\right) = 0$$

with probability 1. For W^*, from Lemma 3.1, $0 < P(\mathbf{X}_i = x) < 1$, $\forall x, \forall i, t = 1$ or 0,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,j=t} W_i^* &= \lim_{n\to\infty}\frac{1}{n}\sum_x x_{\cdot j=t}\sum_i t\mathbf{X}_{i=x} W_i^* \\ &= \lim_{n\to\infty}\sum_x x_{\cdot j=t}\frac{1}{n}\sum_i t\mathbf{X}_{i=x}\frac{1}{P(\mathbf{X}_i=x)} \\ &= \lim_{n\to\infty}\sum_x x_{\cdot j=t} P(\mathbf{X}_i = x)\cdot\frac{1}{P(\mathbf{X}_i=x)} = 2^{p-1}\end{aligned}$$

with probability 1 (Law of Large Number). Since features are binary,

$$\lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=1}} W_i^* = 2^{p-2}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,j=0} W_i^* = 2^{p-1},\quad \lim_{n\to\infty}\frac{1}{n}\sum_i t\mathbf{X}_{i,k=1,\mathbf{X}_{i,j=0}} W_i^* = 2^{p-2}$$

and therefore, we have following equation with probability 1:

$$\lim_{n\to\infty}\left(\frac{\mathbf{X}_{\cdot,k}^{T}(W^*\odot\mathbf{X}_{\cdot,j})}{W^{*T}\mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,k}^{T}(W^*\odot(1-\mathbf{X}_{\cdot,j}))}{W^{*T}(1-\mathbf{X}_{\cdot,j})}\right) = \frac{2^{p-2}}{2^{p-1}} - \frac{2^{p-2}}{2^{p-1}} = 0.$$

□

样本重采样因果恢复模型

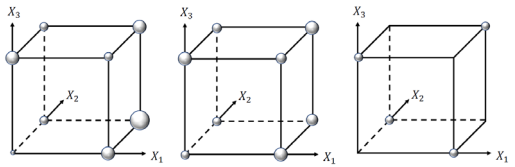
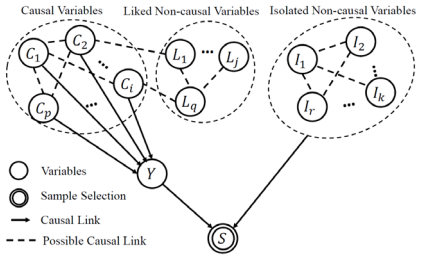
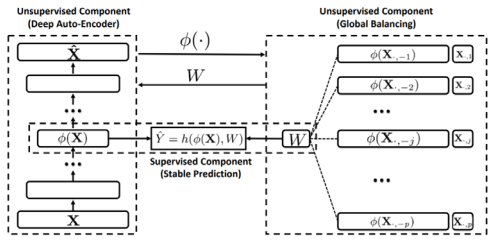


Figure 1: A toy example to illustrate the main idea of each deconfounding method.

因果特征选择解耦模型



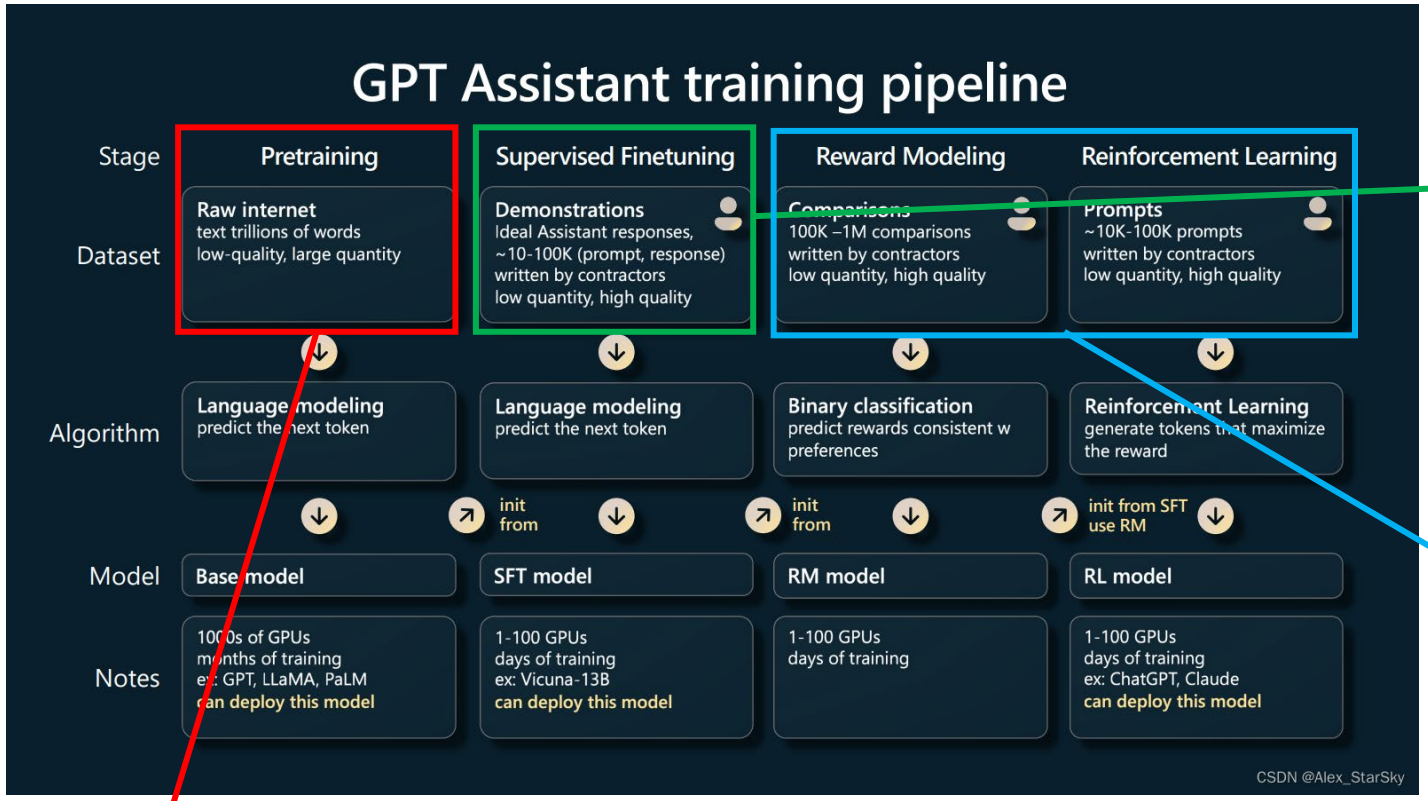
因果约束的深度学习模型



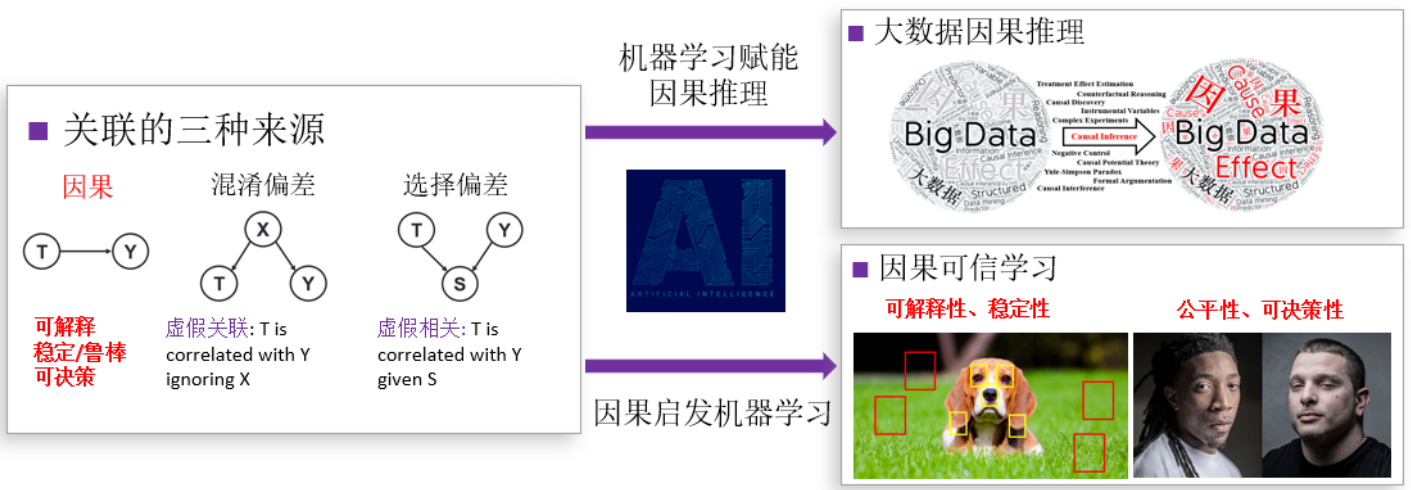
有效提升机器学习模型的可解释性和稳定性

Haotian Wang, Kun Kuang, Long Lan, Wanrong Huang, Fei Wu, Wenjing Yang. Out-of-distribution Generalization with Causal Feature Separation, TKDE, 2023.
Kuang K, Li B, et al. Stable Prediction via Leveraging Seed Variable[J]. TKDE 2022.
Kuang K, Xiong R, et al. Stable prediction with model misspecification and agnostic distribution shift[C]//AAAI, 2020
Kuang K, Cui P, et al. Stable Prediction across Unknown Environments. KDD, 2018.

因果如何赋能大模型



- 因果去除数据偏置
  - 虚假相关问题
  - 灾难遗忘问题
- 因果支撑决策
  - 用户偏好对齐
  - 因果强化学习
- 因果赋能Transformer
  - 由关联自回归到因果回归机制
  - 因果Transformer架构
  - 基于因果知识增强的Transformer架构



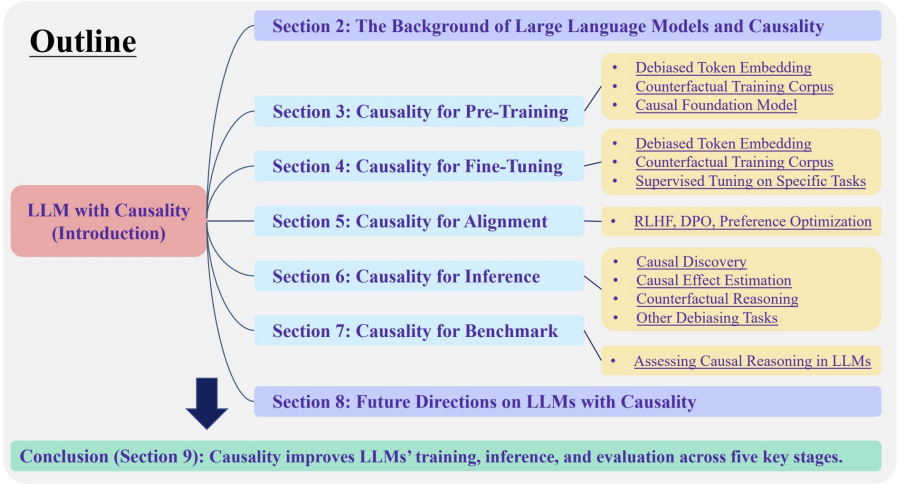


Figure 1: The Role of Causality in Enhancing LLMs: A Comprehensive Framework Across Development Stages

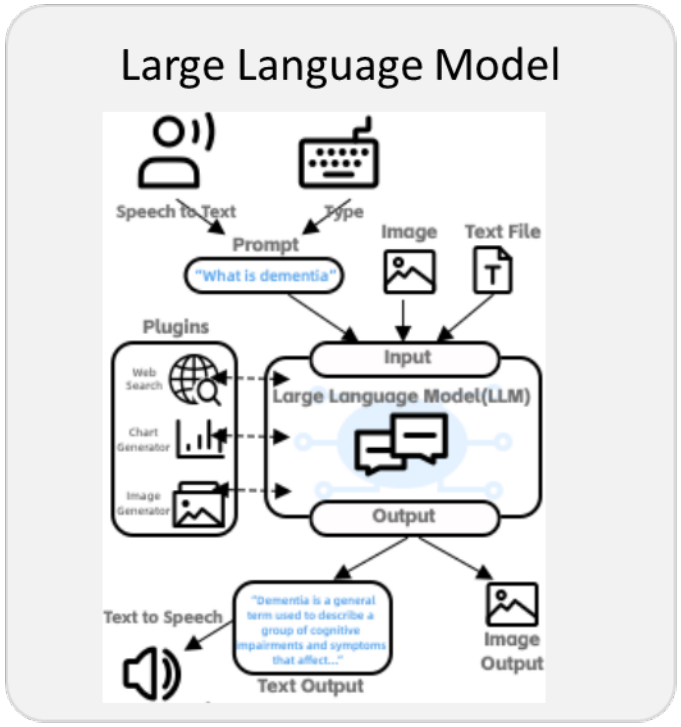
谢谢！

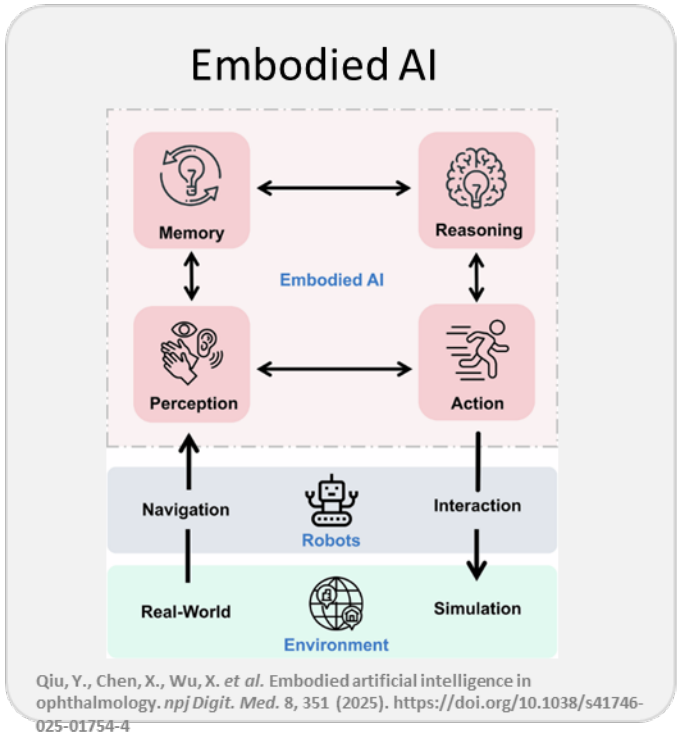
况琨

浙江大学计算机学院

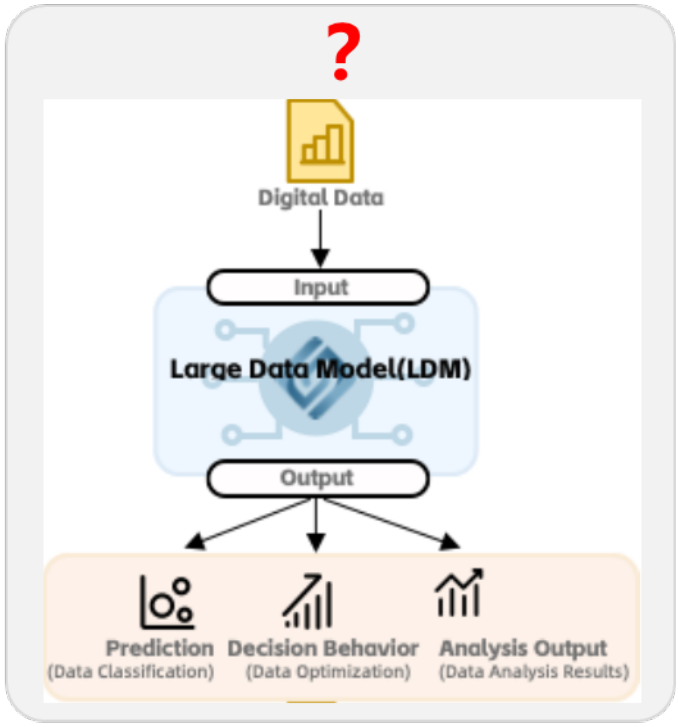
kunkuang@cs.zju.edu.cn

AGI需要三个通用世界模型





Qiu, Y., Chen, X., Wu, X. *et al.* Embodied artificial intelligence in ophthalmology. *npj Digit. Med.* 8, 351 (2025). https://doi.org/10.1038/s41746-025-01754-4



语义空间　　物理空间　　数据空间

Artificial general intelligence (AGI)—sometimes called **human-level intelligence AI**—is a type of artificial intelligence that would match or surpass human capabilities across virtually all cognitive tasks. <**Wikipedia**>

Tutorial安排

议程	嘉宾
因果与关联：从统计学习到因果范式	况琨
复杂环境下因果推断	
因果启发的稳定可泛化学习	
因果赋能大语言模型探索与思考	
茶歇	
因果赋能结构大数据模型：通用数据大模型引领结构化数据智能新范式	张兴璇
因果赋能物理模型与具身智能探索与思考	王浩天