人工智能困境

况琨
浙江大学
https://kunkuang.github.io/

研究背景: 人工智能面临不可解释的重大挑战

- 深度学习面临不可解释、不稳定等挑战

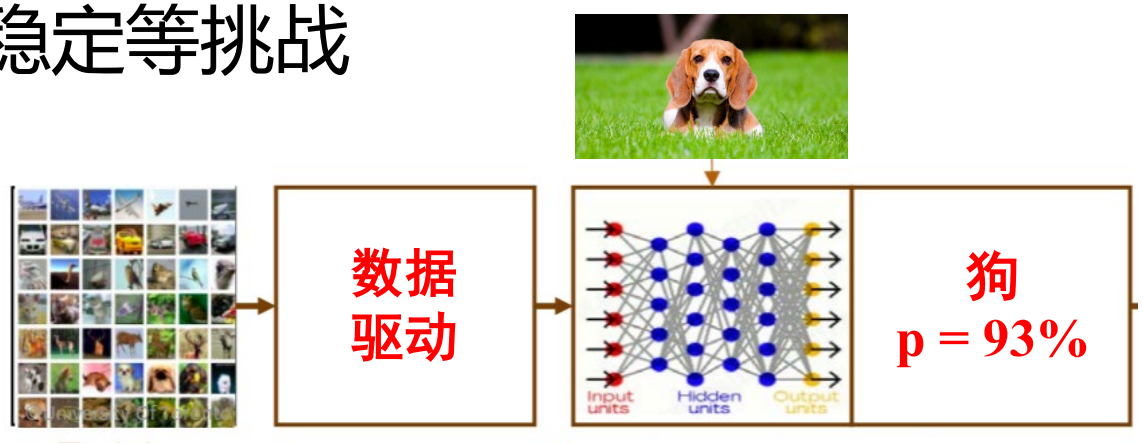
Geoffrey E. Hinton

Yoshua Bengio

Yann LeCun

2018图灵奖获得者
表彰其在深度学习领域的
巨大贡献

深度学习特点

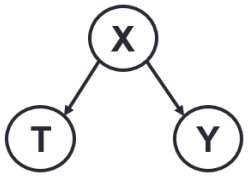
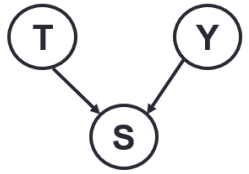


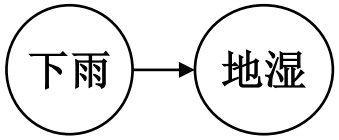
关联学习　概率输出

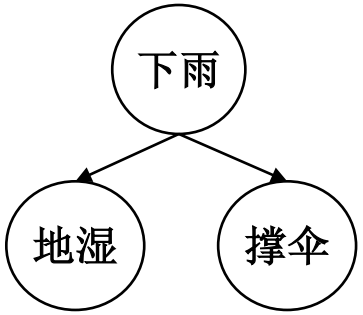
模型难以解释且预测不稳定：为什么图像会被识别为“狗”？
为什么会用“草地”预测狗？为什么不同测试结果差异大？

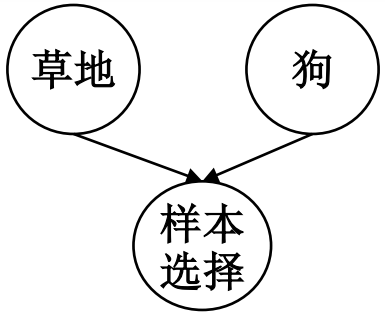
关联 vs. 因果

- 关联的三种来源

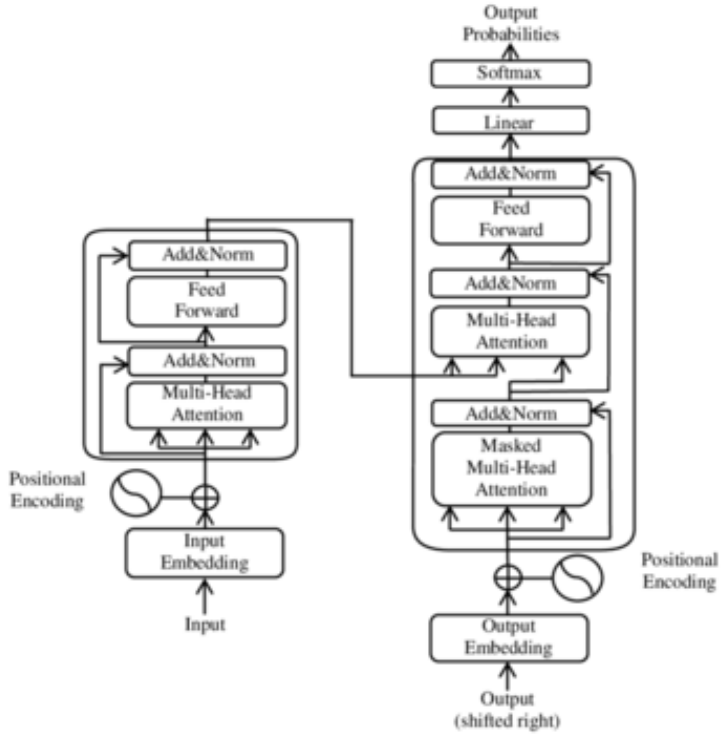
因果	混淆偏差	选择偏差
		
可解释 稳定/鲁棒 可决策	虚假关联: 当忽略 X 时，T 和 Y 相关	虚假关联: 当给定 S 时，T 和 Y 相关







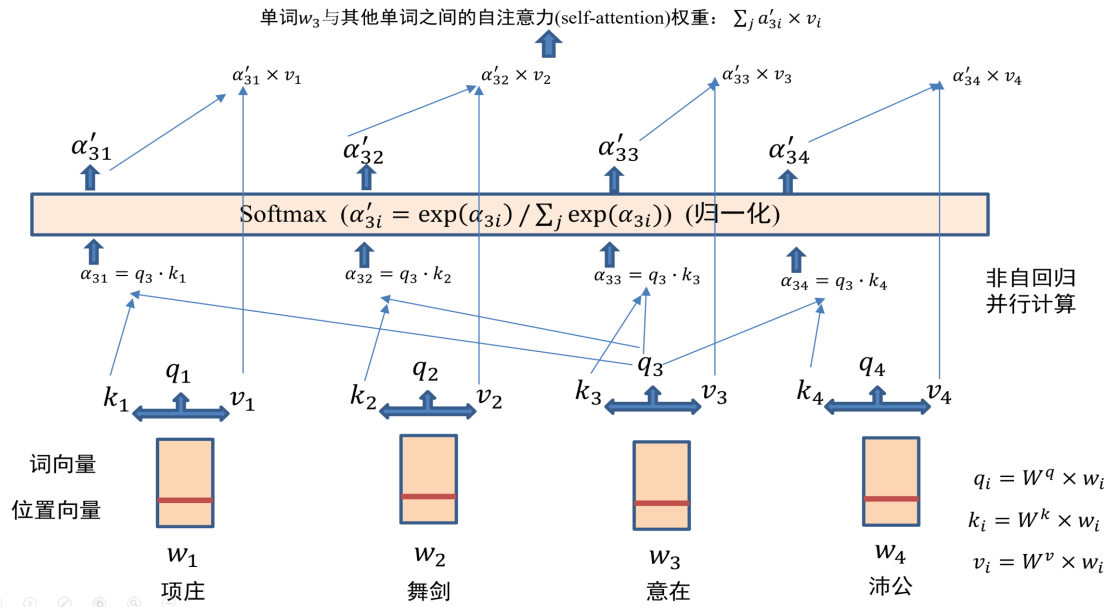
大模型可解释可泛化问题：由关联到因果



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

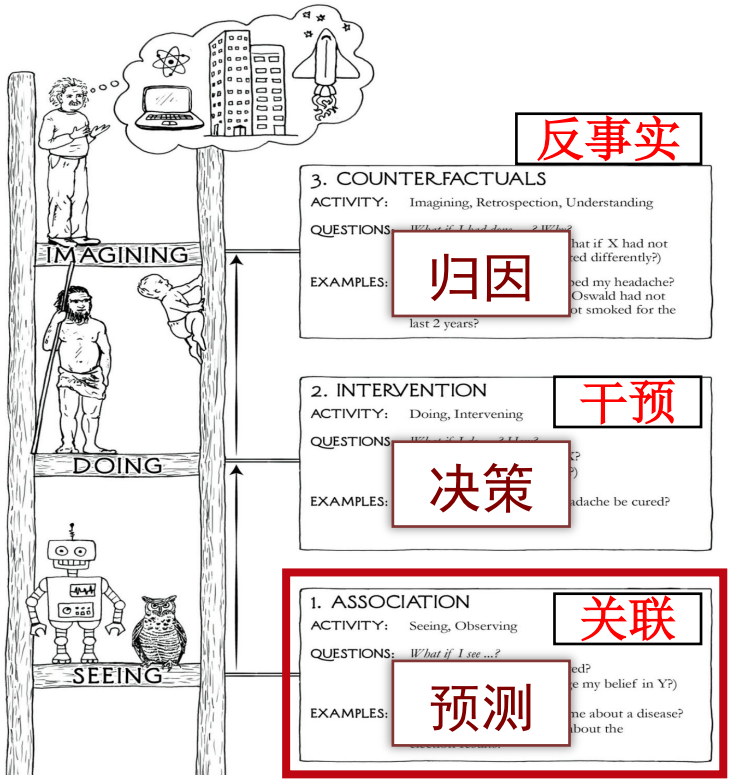
学习单词和单词之间**关联关系**（in-context meaning）

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

研究背景: 人工智能面临不可解释的重大挑战

- 深度学习面临不可解释、不稳定等挑战

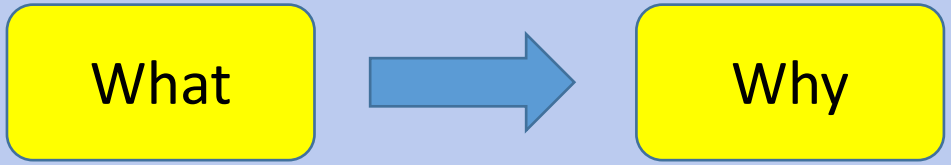
Geoffrey E. Hinton

Yoshua Bengio

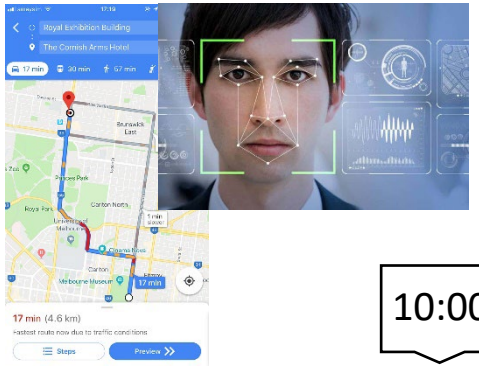
Yann LeCun

2018图灵奖获得者
表彰其在深度学习领域的
巨大贡献

深度学习需要从关联分析跨越到因果推理
“知其然，并知其所以然”

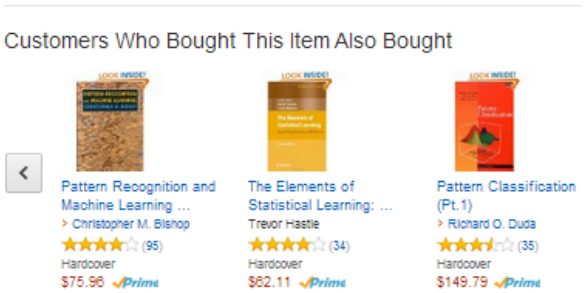


人工智能算法在我们生活中无处不在



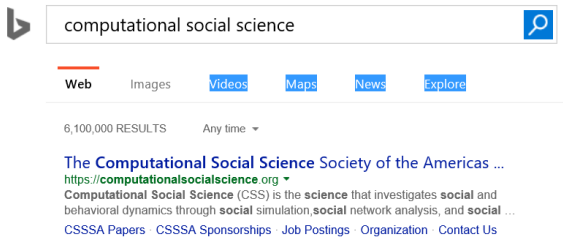
8:00 am

10:00 am

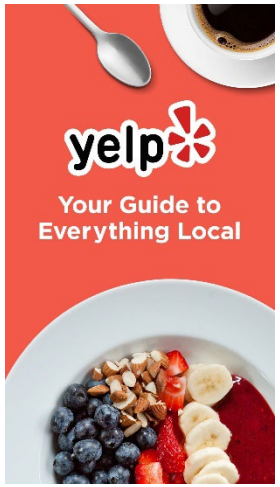


6:00 pm

8:30 am



4:00 pm



8:00 pm

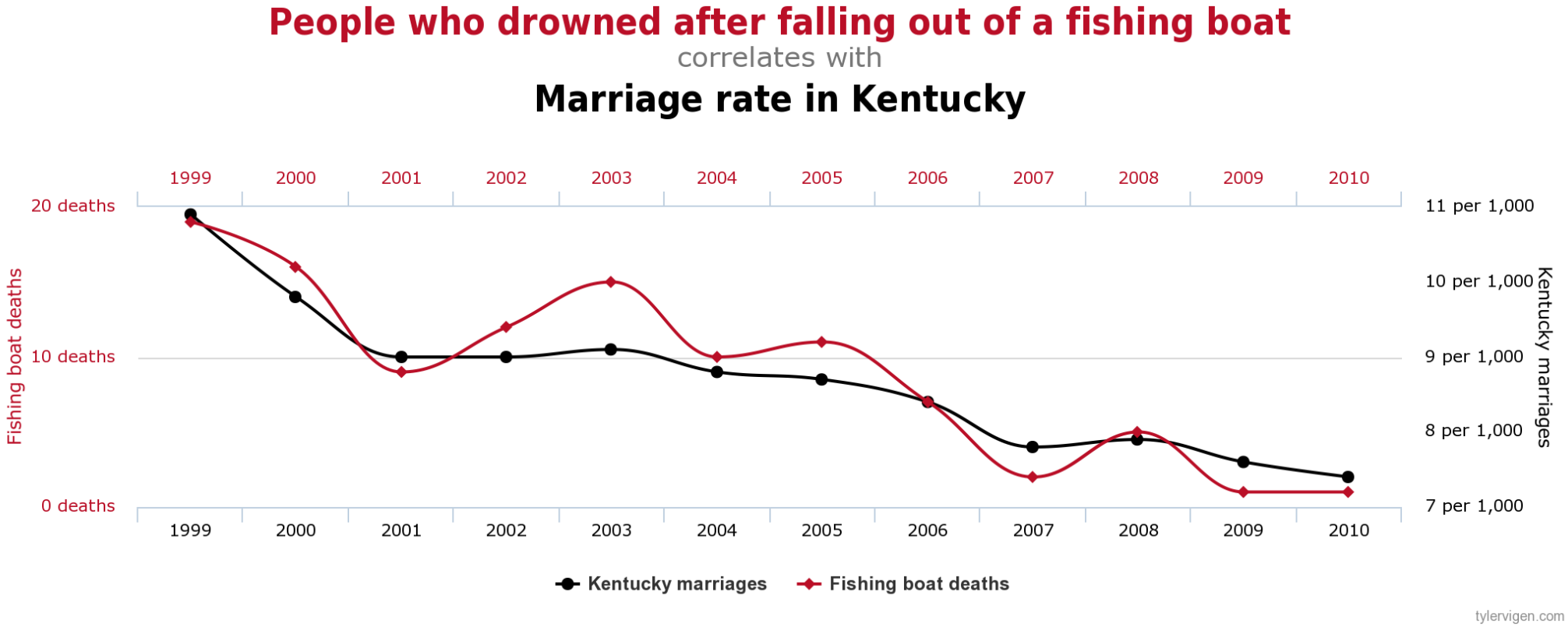
人工智能算法在我们生活中无处不在

- 以人为中心的风险敏感领域

模型优化应从**性能第一**转到**风险控制**

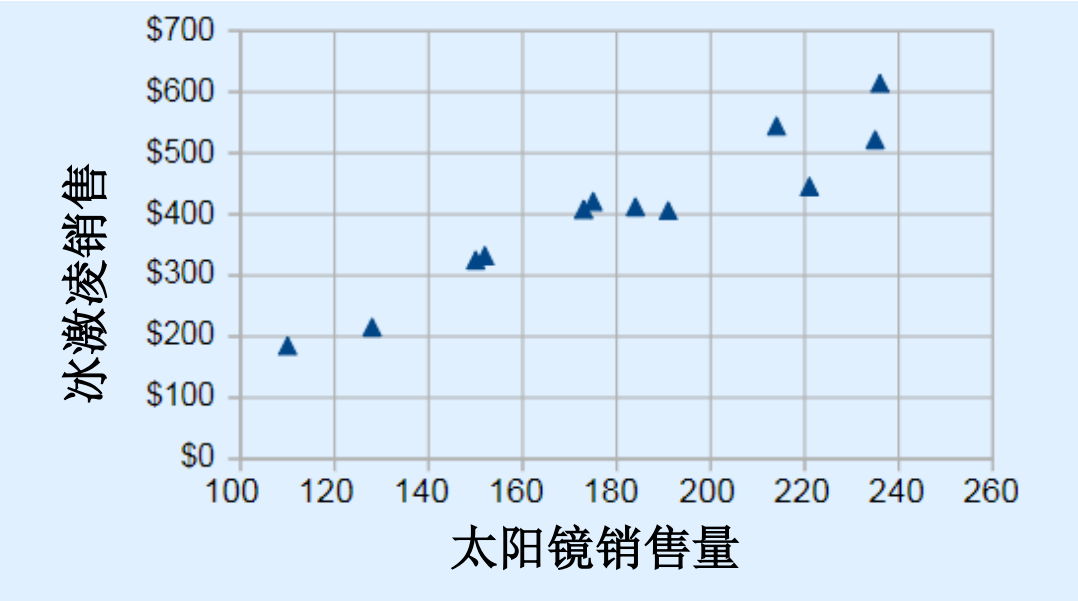
为什么需要从数据关联到因果推理

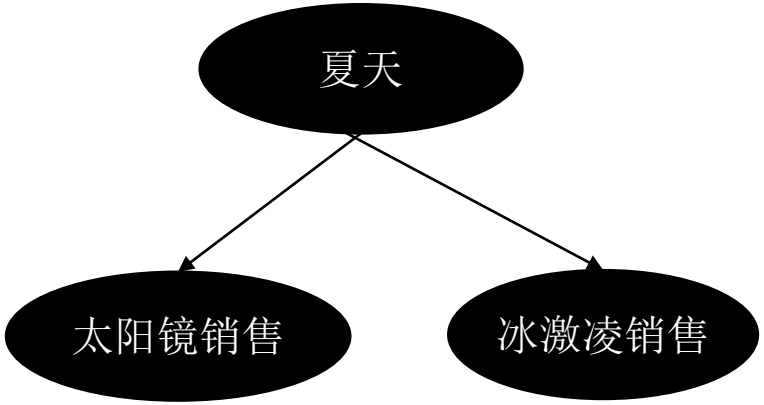
- 数据关联不可解释



为什么需要从数据关联到因果推理

- 数据关联不可解释

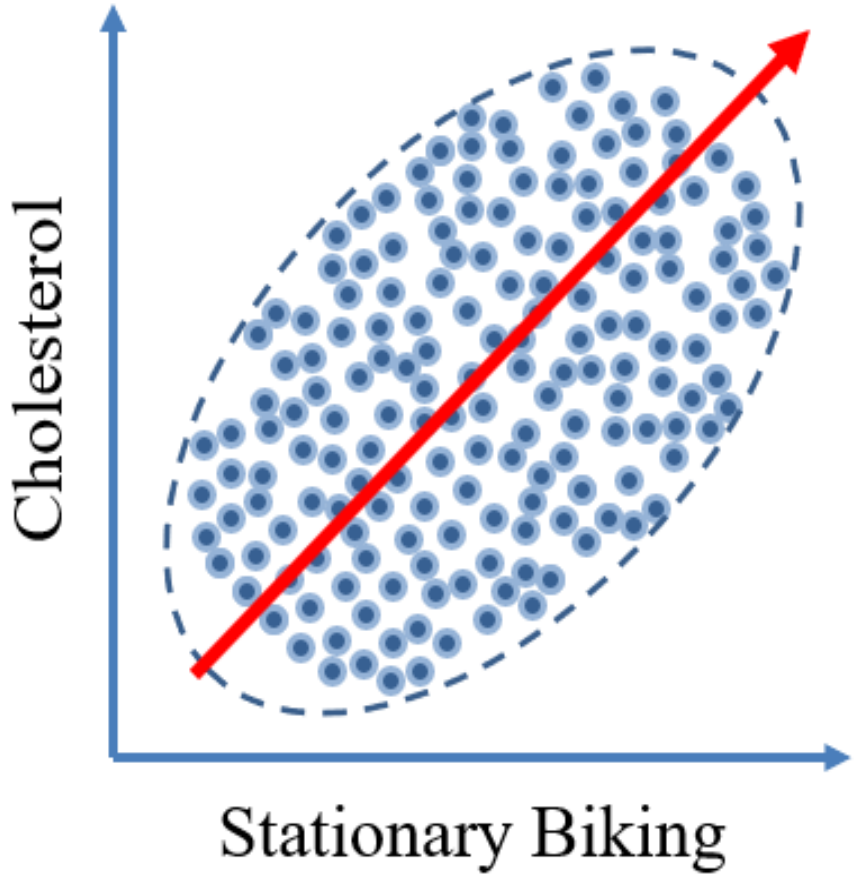


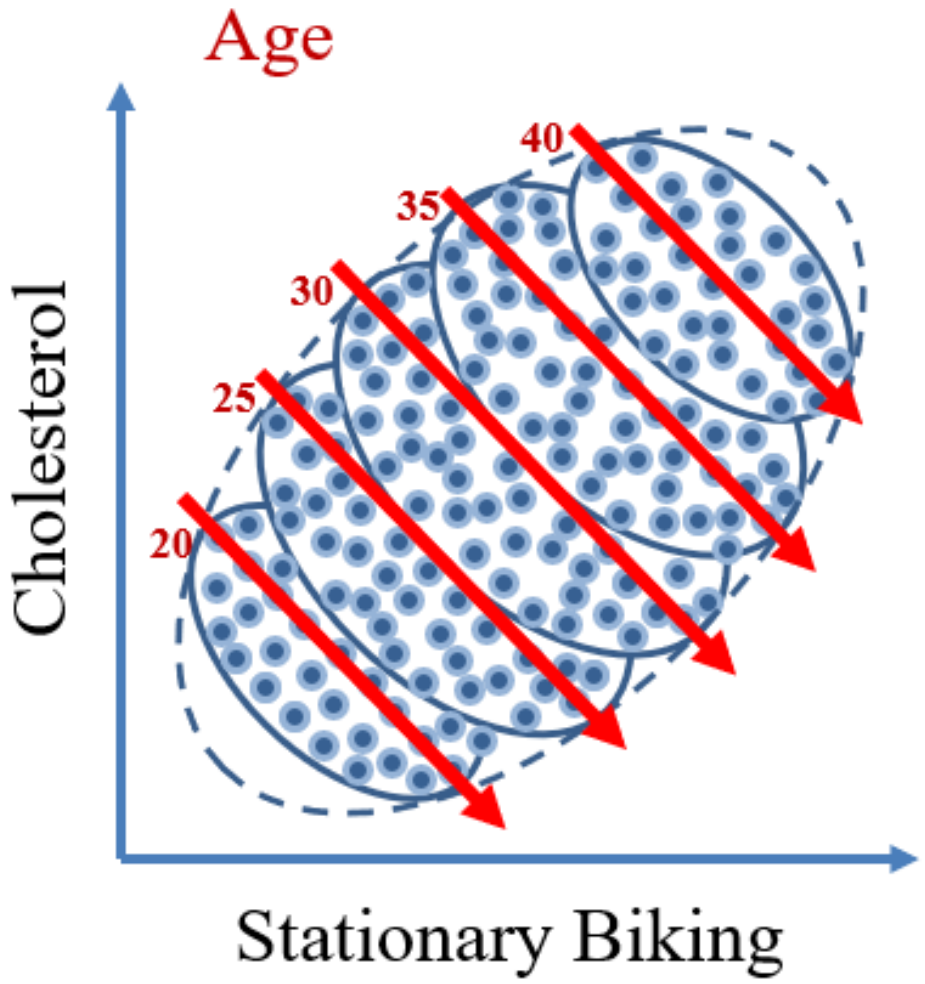


虚假相关

为什么需要从数据关联到因果推理

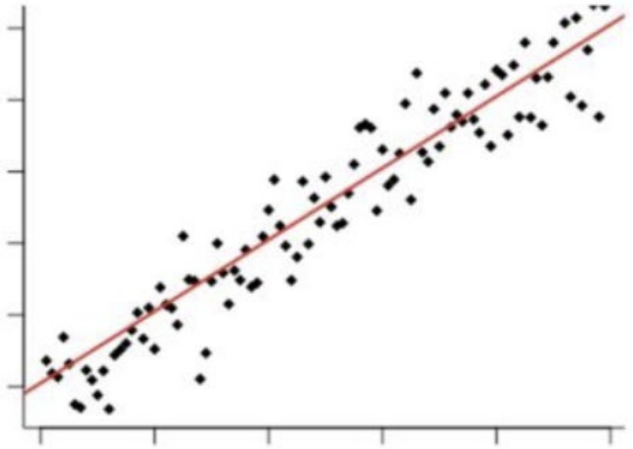
- 数据关联不可用于支撑决策





为什么需要从数据关联到因果推理

- 数据关联不可用于支撑决策



- 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

- 数据关联不可用于支撑决策
- 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

- 数据关联不可用于支撑决策
- 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？

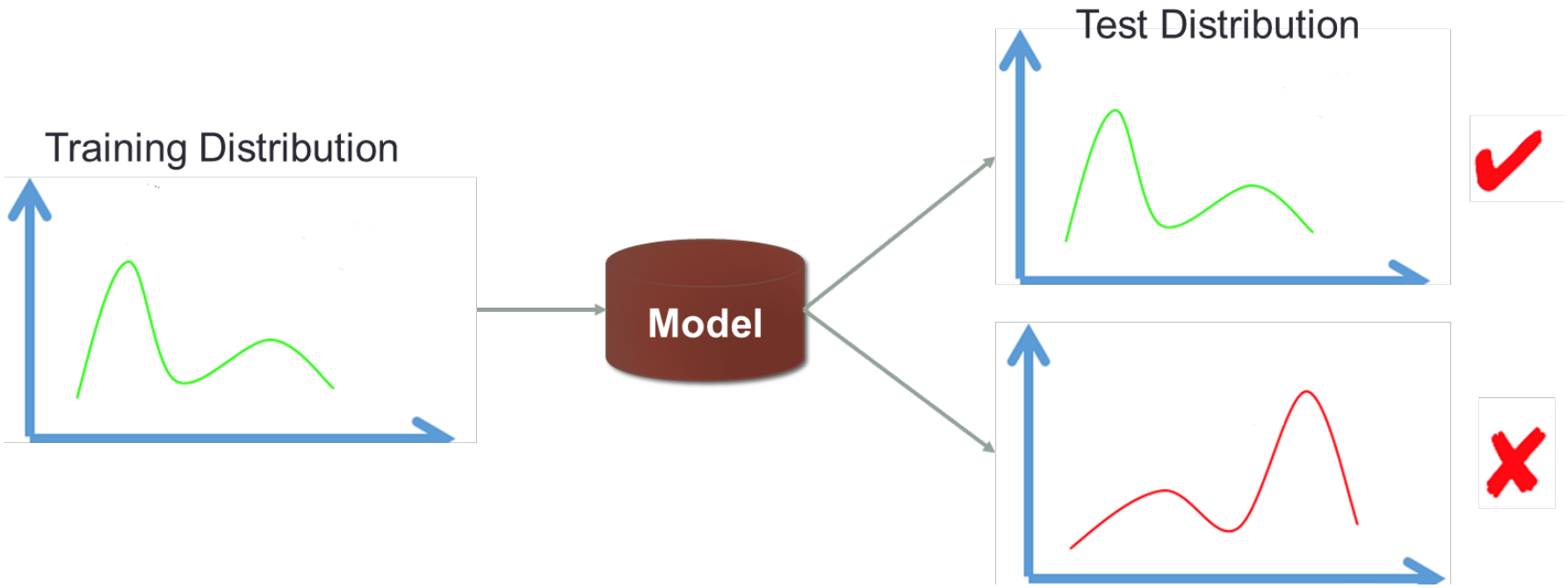
	算法A	算法B
低收入用户	10/400 **(2.5%)**	4/200 **(2%)**
高收入用户	40/600 **(6.6%)**	50/800 **(6.2%)**
整体	50/1000 **(5%)**	54/1000 **(5.4%)**

到底哪个算法更好?

为什么需要从数据关联到因果推理

- 数据关联不稳定，会随着时间、数据、环境等变化而变化

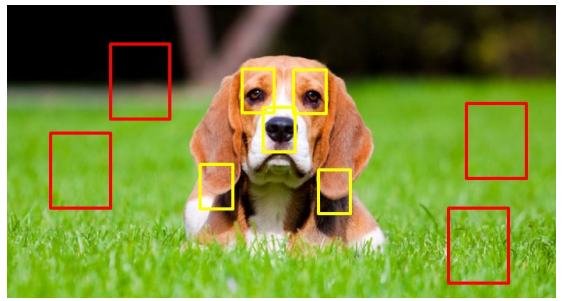
绝大多数机器学习方法需要独立同分布假设

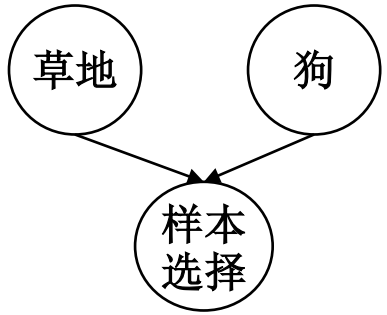


为什么需要从数据关联到因果推理

- 数据关联不稳定，会随着时间、数据、环境等变化而变化

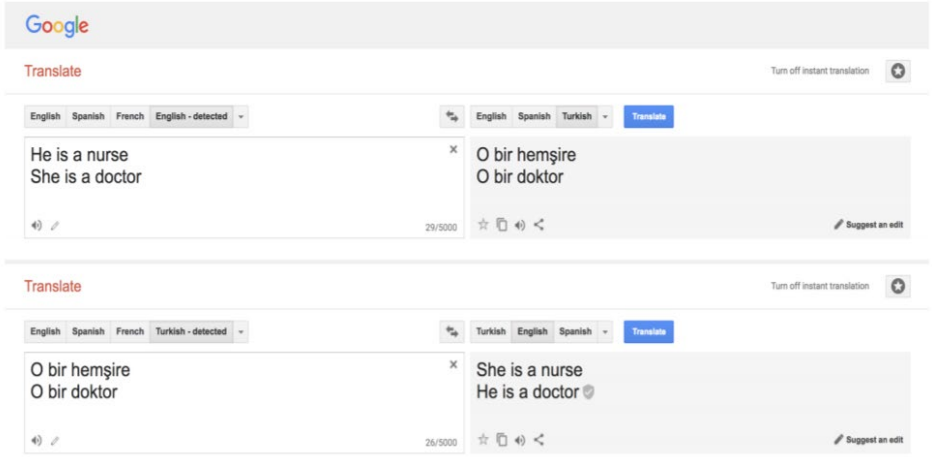
绝大多数机器学习方法需要独立同分布假设





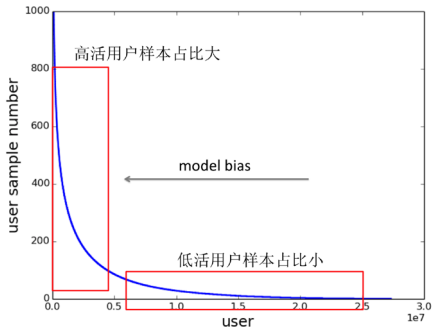
为什么需要从数据关联到因果推理

- 数据关联可能会带来不公平性问题



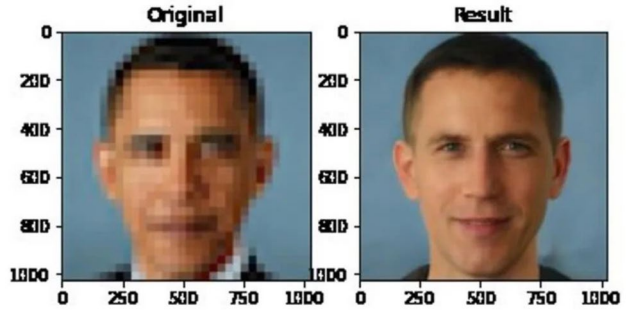
自然语言翻译

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

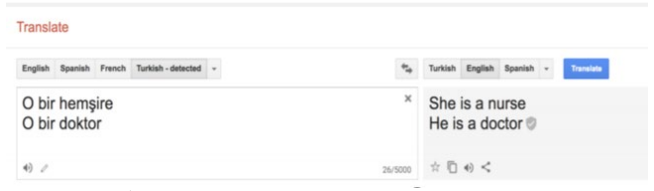


计算机视觉和检索

为什么需要从数据关联到因果推理

- 数据关联可能会带来不公平性问题

Ranking Bias
Popularity Bias
…



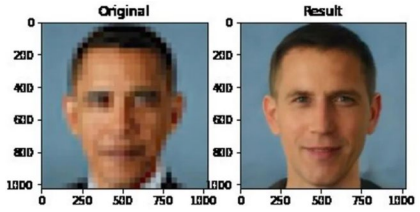
Algorithm

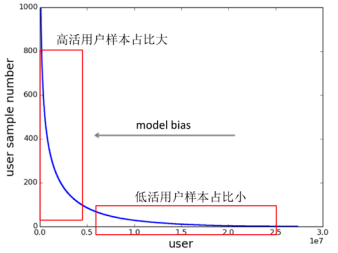
User

Data

Sample Selection Bias
Omitted Variable Bias
…

User Behaviors Bias
Content Production Bias
…



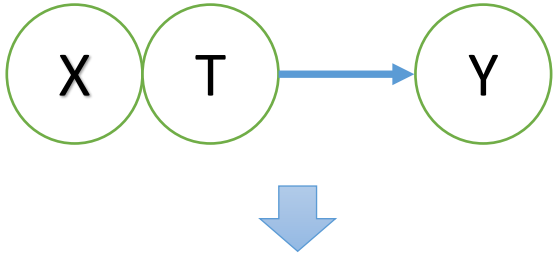


Micro AUC提升，表面收益增加，但可能Macro AUC下降，用户平均满意度下降（模型牺牲低活用户体验，换取更多收益）。

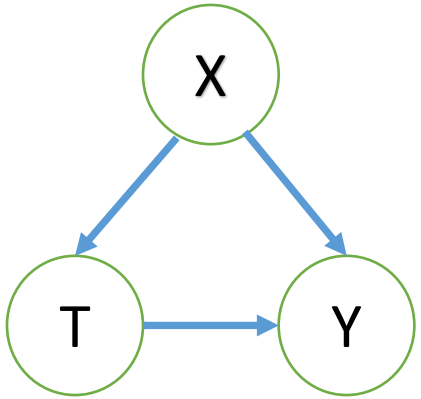
为什么需要从数据关联到因果推理

- 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

为什么需要从数据关联到因果推理

- 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

可信人工智能

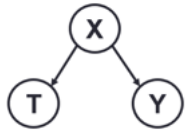
为什么需要从数据关联到因果推理

■ 关联的三种来源

因果

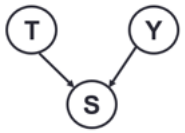
可解释
稳定/鲁棒
可决策

混淆偏差



虚假关联: T is correlated with Y ignoring X

选择偏差

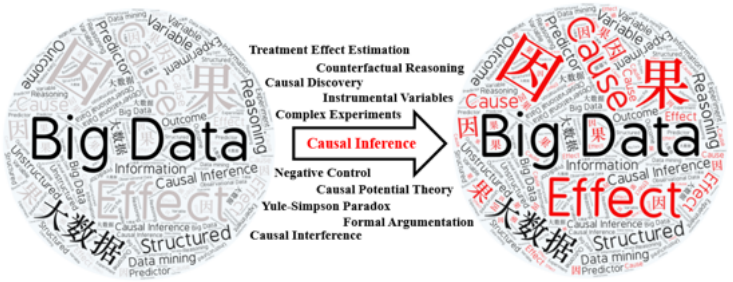


虚假相关: T is correlated with Y given S

机器学习赋能
因果推理

因果启发机器学习

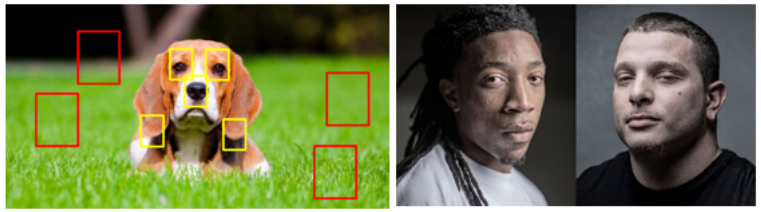
■ 大数据驱动因果推理



■ 因果表征学习

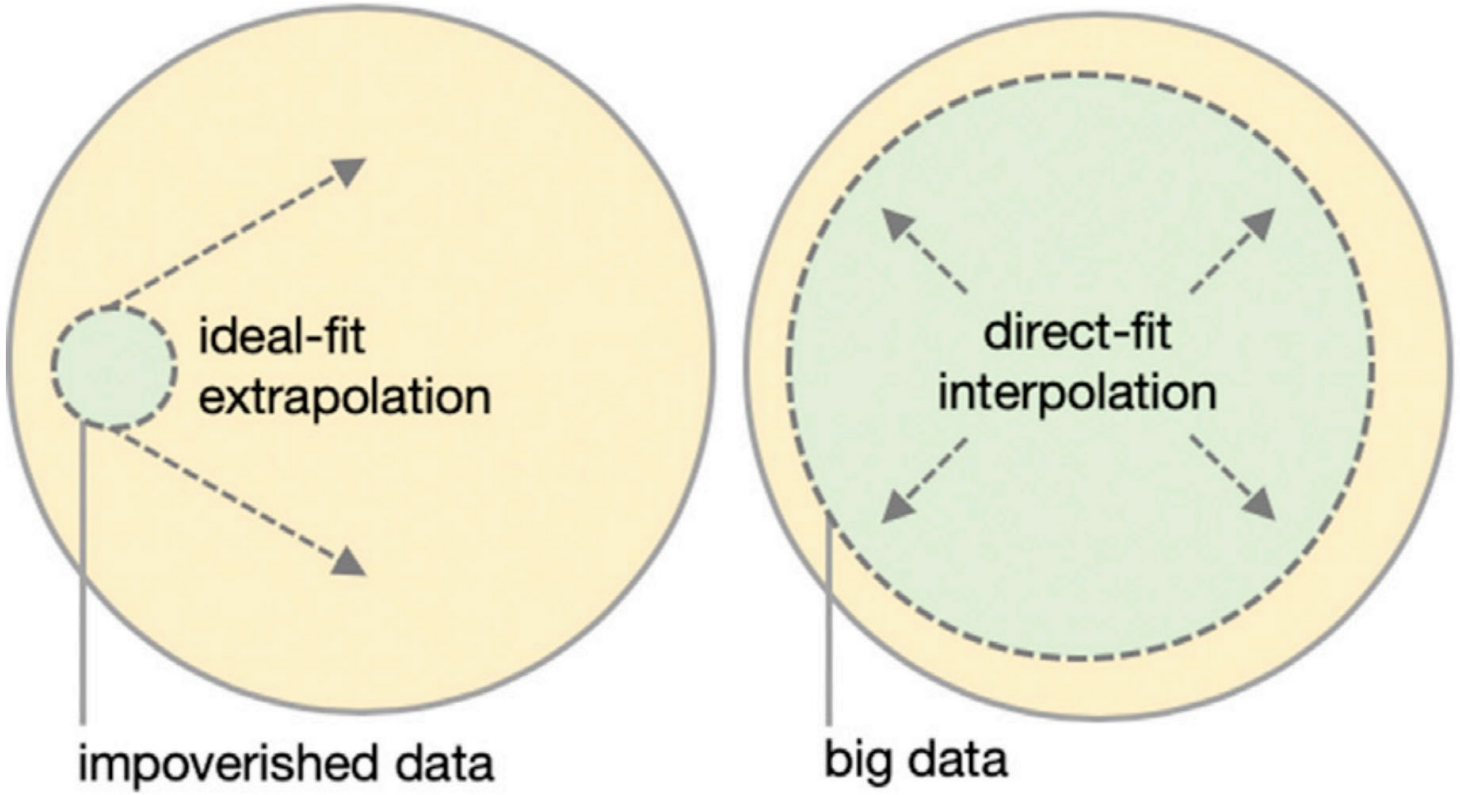
可解释性、稳定性

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

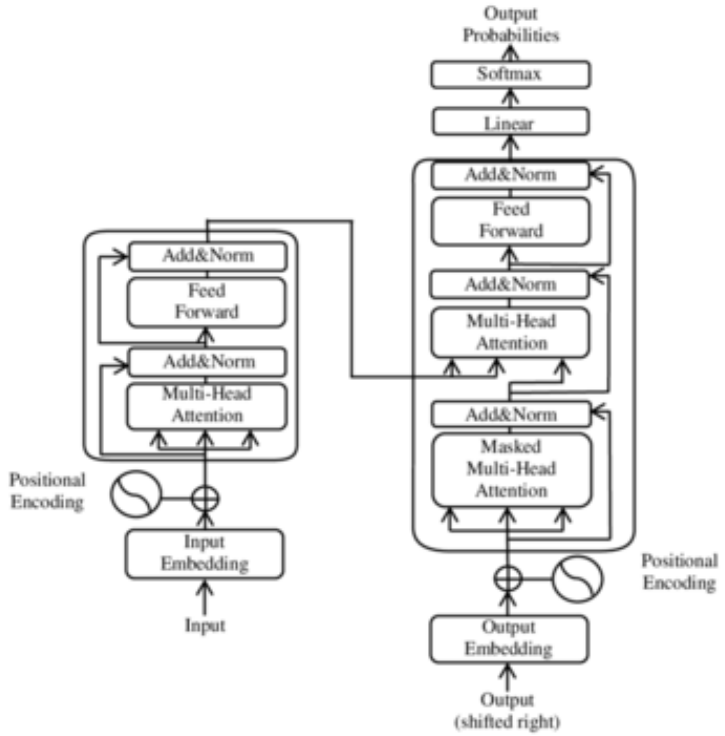
大模型泛化问题：由关联到因果



Wisdom v.s. Knowledge

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

大模型泛化问题：由关联到因果



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

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

Human: Saratoga is a city in California. Saratoga has many parks.

In which city was San Jose's mayor Sam Liccardo born?

text-davinci-003: Sam Liccardo was born in Saratoga, California.

GPT-3.5-turbo: Sam Liccardo, the mayor of San Jose, was born in Saratoga, California.

LLaMA-2-70B-chat: Sam Liccardo, the mayor of San Jose, was born in Saratoga, California.

利用指令调整 LLM，通过移除不相关的文本，来重写上下文

1）给定上下文 x，S2A 首先重新生成上下文 x'，从而删除会对输出产生不利影响的上下文的不相关部分。

2）给定 x′，然后使用重新生成的上下文而不是原始上下文生成 LLM 的最终响应：y ~ LLM (x′)。

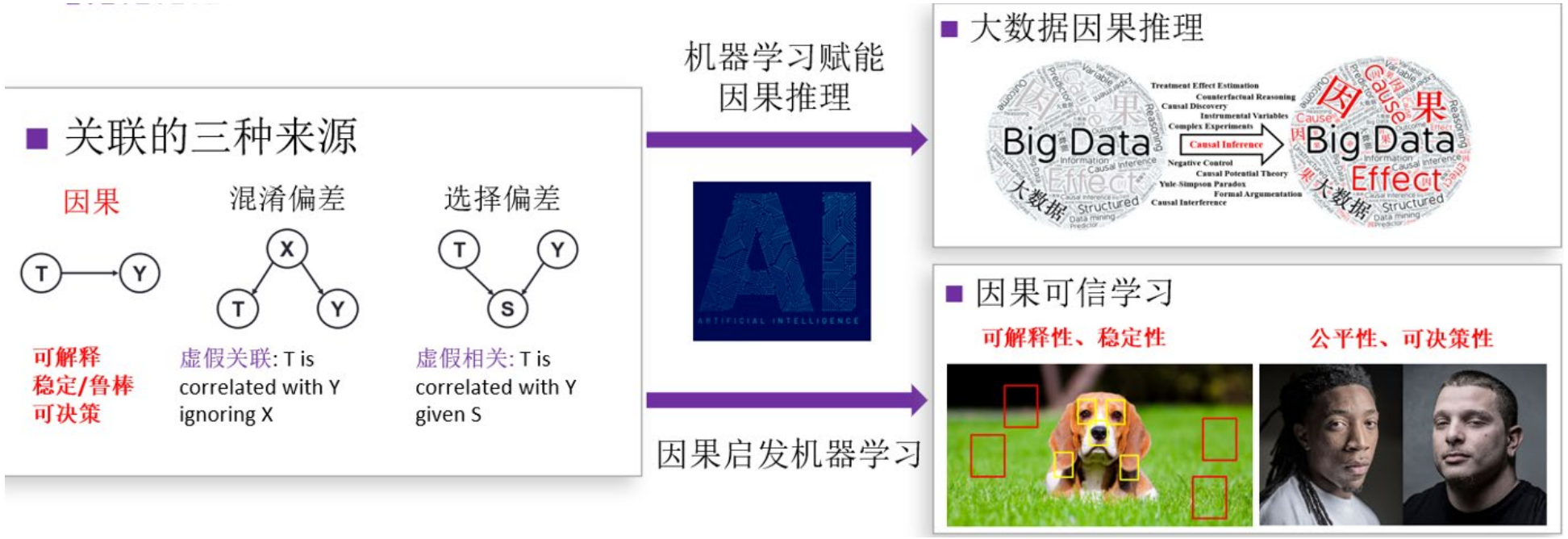
System 2 Attention (is something you might need too), Meta 2023

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

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

Causal Transformer：引用因果，计算单词之间的因果关联

Causal Transformers: Improving the Robustness on Spurious Correlations



Thank You!

kunkuang@zju.edu.cn
https://kunkuang.github.io/