因果关系与时空数据挖掘

张文涛

wechat : zwt532586242

大纲

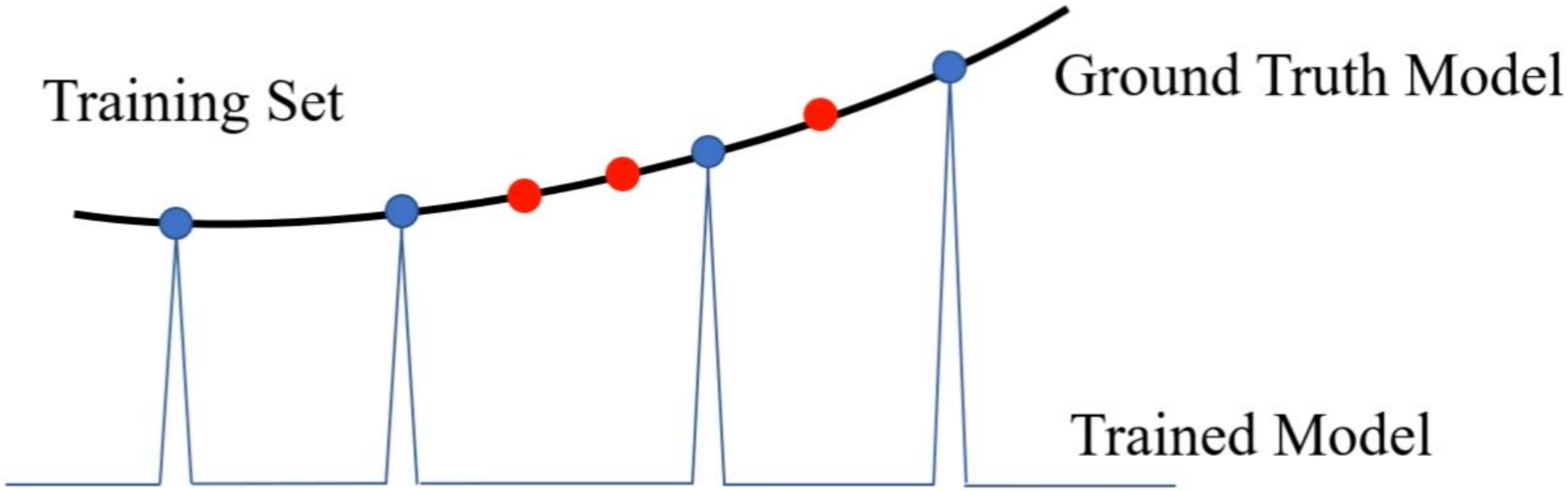
为什么要研究因果关系

当前深度学习的难题-外泛化问题

- 我们希望最终学得的模型可以在任意数据上给出正确的判断
- 然而，现在的学习范式只是在最小化训练损失（Empirical Risk Minimization，简称ERM）
- 当训练损失最小时，模型可以对任意数据给出正确的判断么？

泛化性能

图片来自网络



- 我们考虑上面这个极端例子，蓝色为训练样本，红色为测试样本。
- 很明显，Trained Model的train loss 为0，而test/population loss很大。
- 模型只学习到了训练集中的信息，但是不能泛化到其他样本上。

泛化性能

如何提高模型的泛化性能呢?

现有的一些研究方向:

1. 扩充数据集: NICO[1];
2. 无/自监督增广数据集: ST-SSL[2];
3. 改变学习范式: IRM[3]、HRM[4];
4. **引入因果关系:** CausalAtt[5]、Causal4Rec[6]、STNSCM[7];

Reference:
[1] NICO Challenge: Out-of-Distribution Generalization for Image Recognition Challenges.
[2] Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction.
[3] Invariant Risk Minimization.
[4] Heterogeneous Risk Minimization.
[5] Causal Attention for Interpretable and Generalizable Graph Classification.
[6] Causal Recommendation: Progresses and Future Directions.
[7] Spatio-temporal Neural Structural Causal Models for Bike Flow Prediction.

因果关系是什么

因果关系

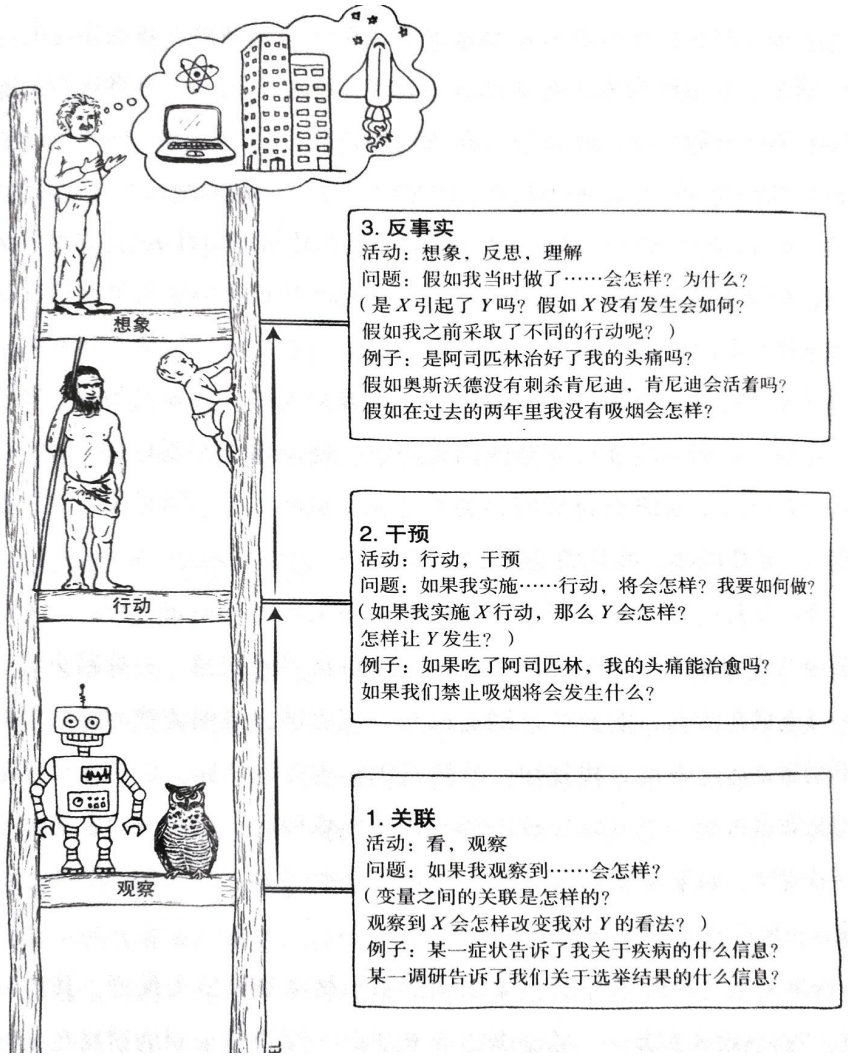
- 不论在中国的那里，或者你是否可以抬头看到太阳，我们都能知道，太阳在天上

  因：现在是白天 ⟶ 果：太阳一定在天上

- 因果关系最白话的解释：在任何场景下都成立的关系
- 因此，引入因果关系可以提高模型泛化性能

人工智能中的三层因果关系之梯

图片来自网络



图灵奖获得者Peral & Mackenzie（2019）《The Book of Why》：提出因果关系的三个层级：

③ 反事实：想象

张三没打疫苗患了新冠；

假若当初打疫苗，是否不患新冠

② 干预：决策

如果打疫苗，疫情会减轻么？

① 关联/相关：预测

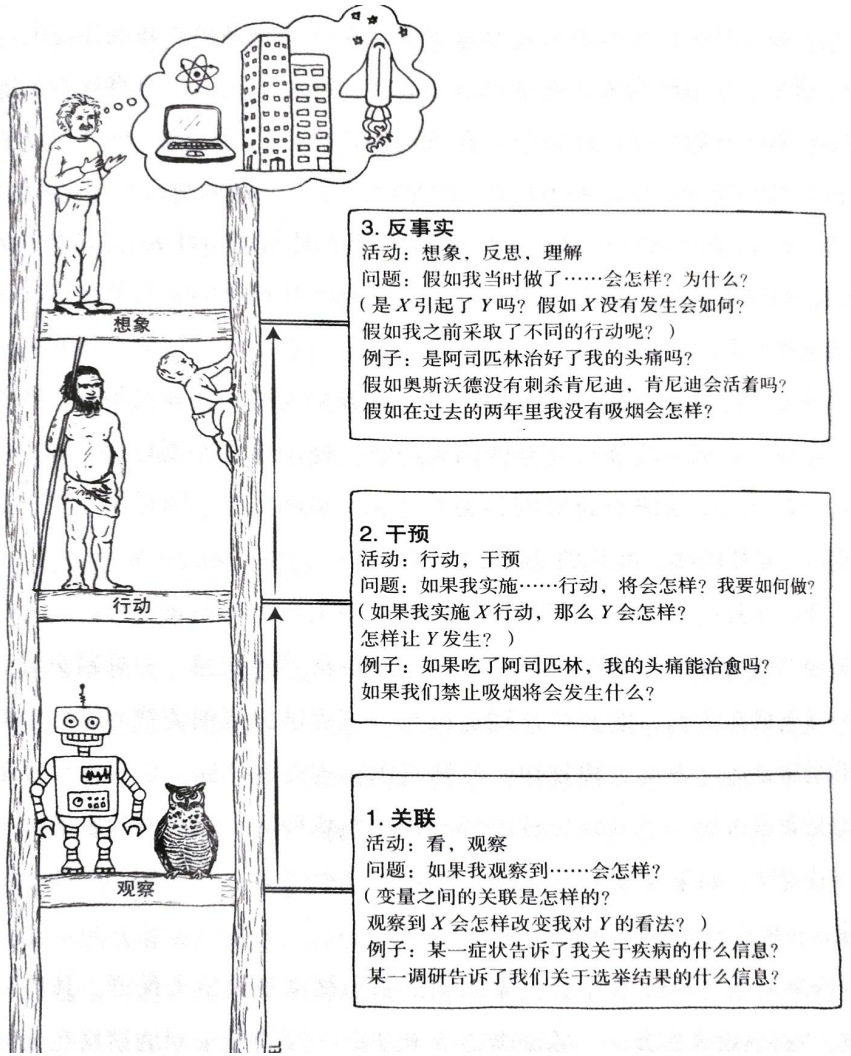
打疫苗越多的地方或时期，疫情越重

当今人工智能处于最低层级：相关，

无论数据多大或神经网络多深，都无法回答“干预”问题

人工智能中的三层因果关系之梯

图片来自网络



图灵奖获得者Peral & Mackenzie（2019）《The Book of Why》：提出因果关系的三个层级：

③ 反事实：想象

张三没打疫苗患了新冠；

假若当初打疫苗，是否不患新冠

② 干预：决策

如果打疫苗，疫情会减轻么？

① 关联/相关：预测

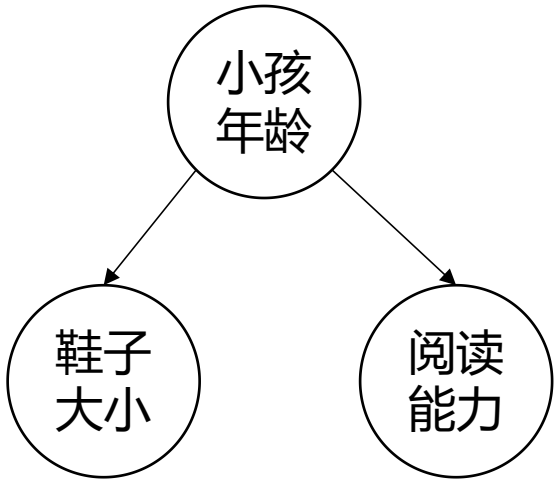
打疫苗越多的地方或时期，疫情越重

那是否不关注“干预”，相关也够用呢？

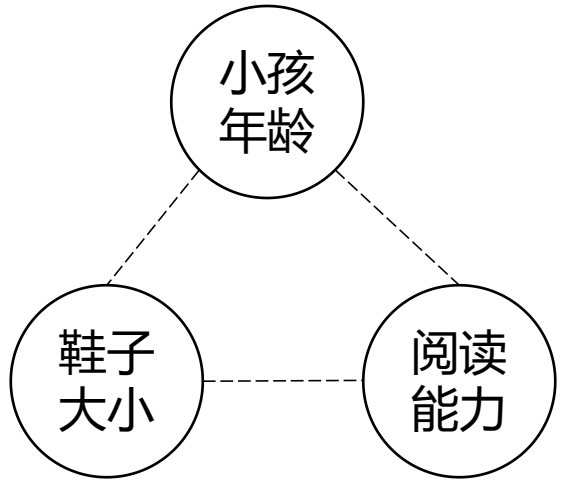
“相关关系”不同于“因果关系” (1)

无因果关系可能会表现出虚假相关关系：

- Freedman（1991）：小学生阅读能力与鞋尺寸有强相关；
- 根据小孩鞋尺寸能预测他的阅读能力！
- 然而人为地改变鞋的尺寸，不会提高他们的阅读能力。



因果关系

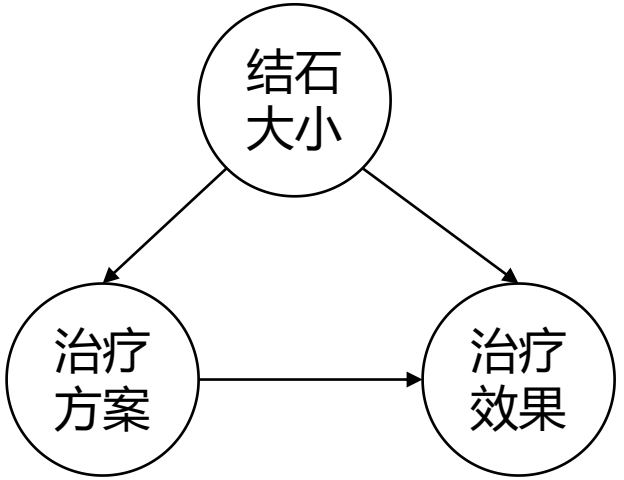


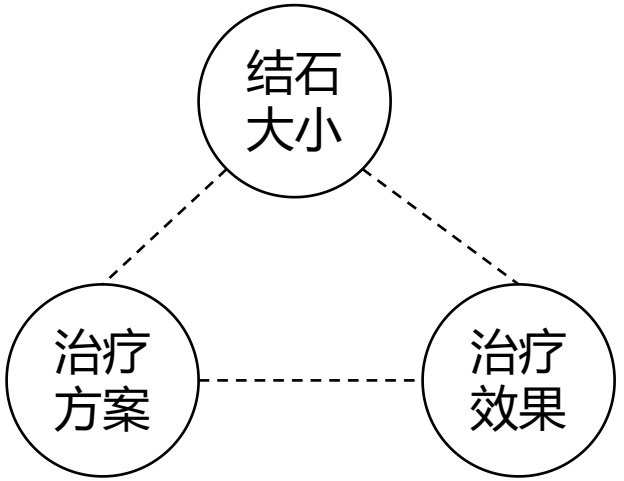
相关关系

相关关系导致的悖论

- 辛普森悖论

肾结石	A 疗法	B 疗法
小型	**81/87 = 93%**	234/270 = 87%
大型	**192/263 = 73%**	55/80 = 69%
总计	273/350 = 78%	**289/350 = 83%**

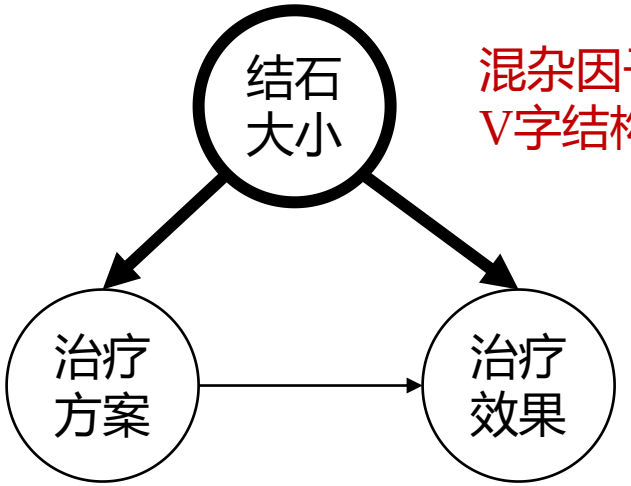




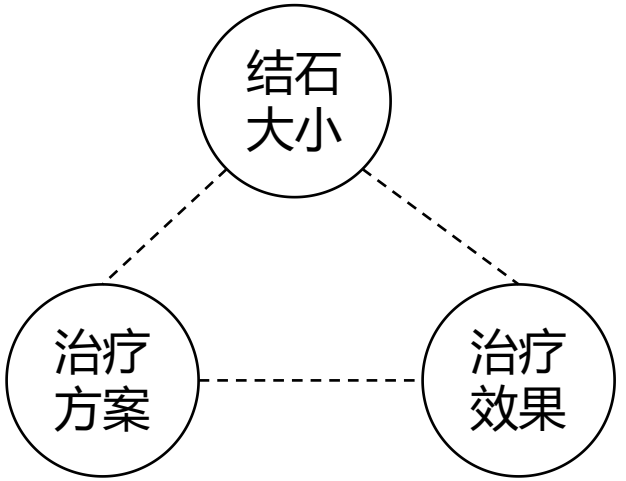
相关关系导致的悖论

- 辛普森悖论

肾结石	A 疗法	B 疗法
小型	**81/87 = 93%**	234/270 = 87%
大型	**192/263 = 73%**	55/80 = 69%
总计	273/350 = 78%	**289/350 = 83%**



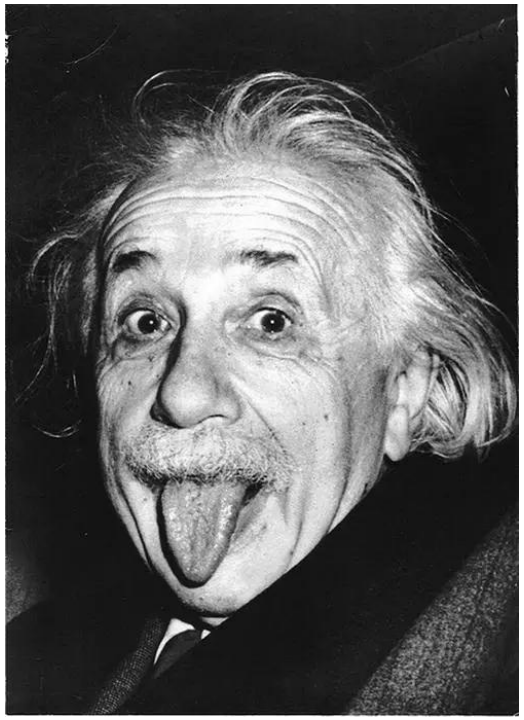
混杂因子
V字结构

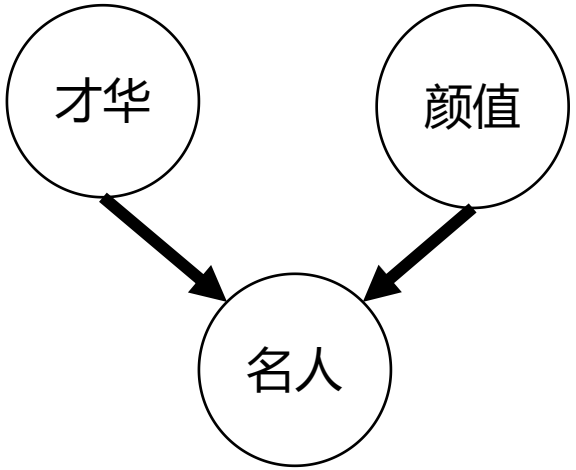


混杂因子的存在可能会导致虚假的相关性

“相关关系”不同于“因果关系” (2)

- 通常“才华”和“颜值”是两个独立的特质，但是若我们已知一个名人颜值不高，那么就可以推断他很有才华。
- 所以，“才华”和“颜值”并不独立而是成反比?
- 其实，是出现了对撞结构。当选择名人作为数据集时就会出现选择偏差。





如何研究因果关系

Potential outcome

VS

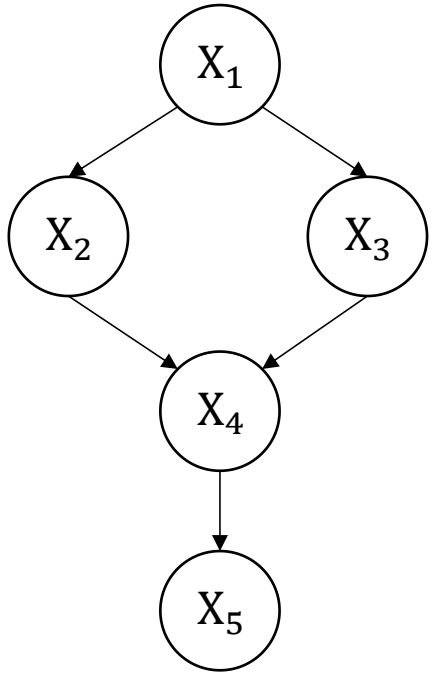
Structure causal model

Potential outcome [Rubin, 1991]

- Potential Outcome 主要关注于对干预影响的估计。例如，干预为是否读大学，A同学有两种潜在结果，大学毕业，工资8000；高中毕业，工资6000。则该干预的因果效应为+2000。
- 有了干预的因果效应（Treatment Effect），我们可以推测某一同学的潜在结果。
- 但需要满足以下约束：
  - 稳定性假设（SUTVA）：不同个体间的潜在结果不会互相影响；干预水平对所有个体相同。
  - 一致性：如果实际接受的治疗是w，治疗w的潜在结果等于观察到的结果。
  - 可忽略性：给定协方差，即协变量影响治疗，作为干预的治疗与潜在结果无关。(Unconfoundedness)
  - 可能行：每一种潜在结果必须是可能的。

Structure causal model [Pearl, 2000]

- 用贝叶斯网络（DAG）建模变量之间的因果关系



Causal Graph

- $X_1 = f_1(\epsilon_1)$
- $X_2 = f_2(X_1, \epsilon_2)$
- $X_3 = f_3(X_1, \epsilon_3)$
- $X_4 = f_4(X_2, X_3, \epsilon_4)$
- $X_5 = f_5(X_4, \epsilon_5)$

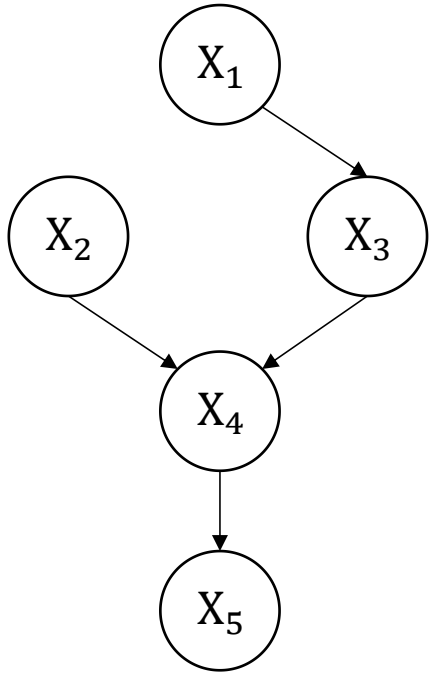
Function

联合分布：

$$P(X_1,\ X_2,\ X_3,\ X_4,\ X_5) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2,X_3)P(X_5|X_4)$$
$$= \prod P(X_i \mid pa(X_i))$$

$pa(X_i)$ 表示 X_i 的父节点集合。

干预的定义



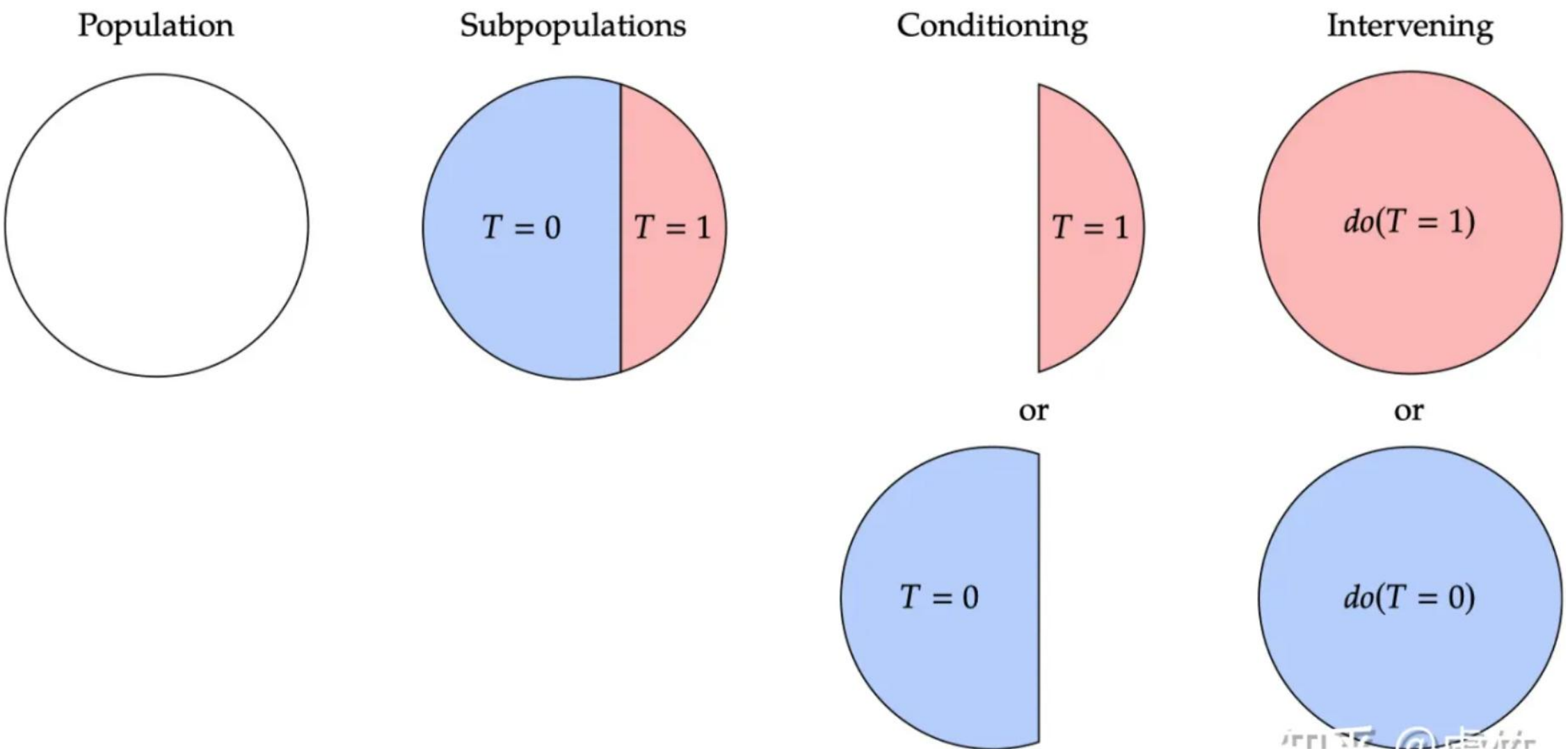
外部干预$X_2 = \mathrm{x}_2$

- $X_1 = f_1(\epsilon_1)$
- $X_2 = set(X_2 = x_2)$
- $X_3 = f_3(X_1, \epsilon_3)$
- $X_4 = f_4(X_2, X_3, \epsilon_4)$
- $X_5 = f_5(X_4, \epsilon_5)$

联合分布：

$$P(X_1,\ X_3,\ X_4,\ X_5|set(x_2)) = I(X_2 = x_2)\, P(X_1)P(X_3|X_1)P(X_4|X_2, X_3)P(X_5|X_4)$$

do 算子


Population
Subpopulations
Conditioning
Intervening
T = 0
T = 1
T = 1
or
T = 0
do(T = 1)
or
do(T = 0)

图片来自网络

Structure causal model VS Potential outcome

- Structure causal model
- 拥有一套完整的数学定义及运算符号。
- 对先验要求高，需要知道各变量之间的因果关系。

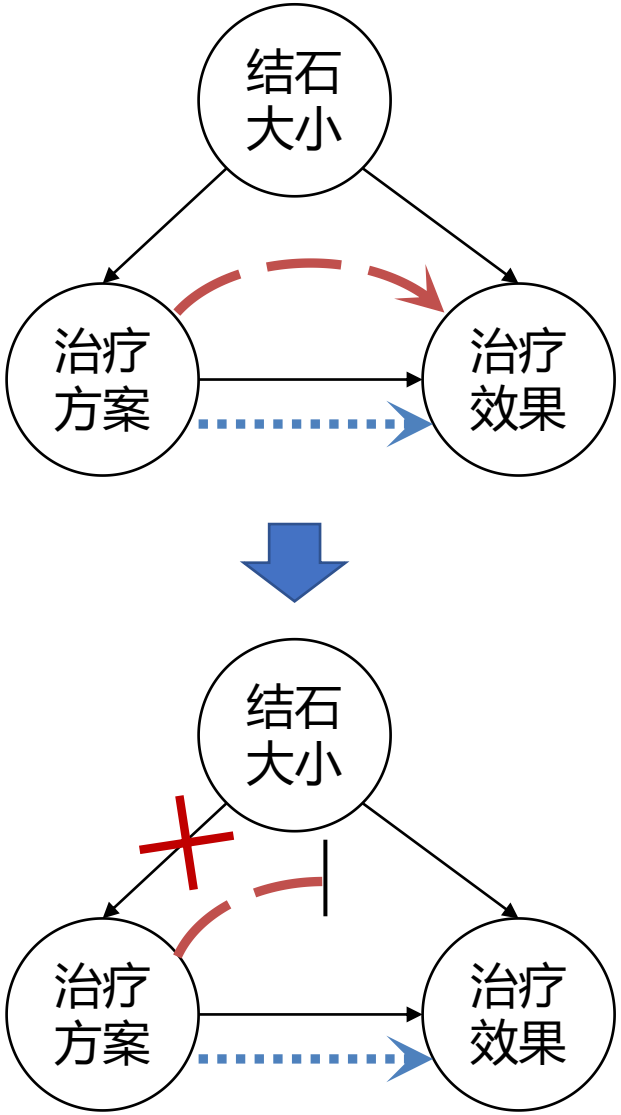
- Potential outcome
- 只关注变化带来的影响，即causal effect，对先验知识要求低。
- 假设条件强，在大部分领域都难以满足4项假设。

应用：反事实推断，因果学习

利用SCM来解决混杂偏倚

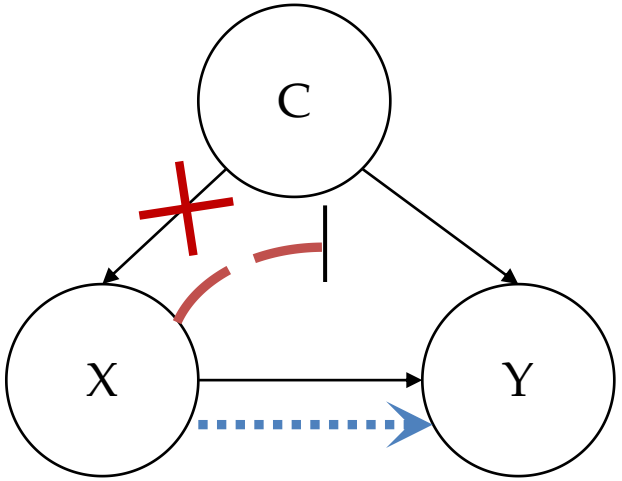
- 辛普森悖论

肾结石	A 疗法	B 疗法
小型	81/87 = 93%	234/270 = 87%
大型	192/263 = 73%	55/80 = 69%
总计	273/350 = 78%	289/350 = 83%



do算子

- 我们令C表示结石的大小，X表示治疗方案，Y表示治疗效果
- 当C可以阻断X和Y之间的所有后门路径时，可以通过do算子进行后门调整



$$P_\Theta(Y|do(X)) = \sum_k P_\Theta(Y|do(X), C_k)P(C_k|do(X))$$ 贝叶斯准则

$$= \sum_k P_\Theta(Y|X, C_k)P(C_k)$$ C_k与X独立

$$\boldsymbol{P_\Theta(Y|do(X)) = \sum_k P_\Theta(Y|X, C_k)P(C_k)}$$

利用SCM来解决混杂偏倚

$$P(C_{small}) = \frac{87 + 270}{700} = 0.51$$

$$P(C_{\text{large}}) = \frac{192 + 55}{700} = 0.49$$

$$P(\text{Y}|do(X_A)) = 0.51 * 93\% + 0.49 * 73\% = 83.2\%$$

$$P(\text{Y}|do(X_B)) = 0.51 * 87\% + 0.49 * 69\% = 78.18\%$$

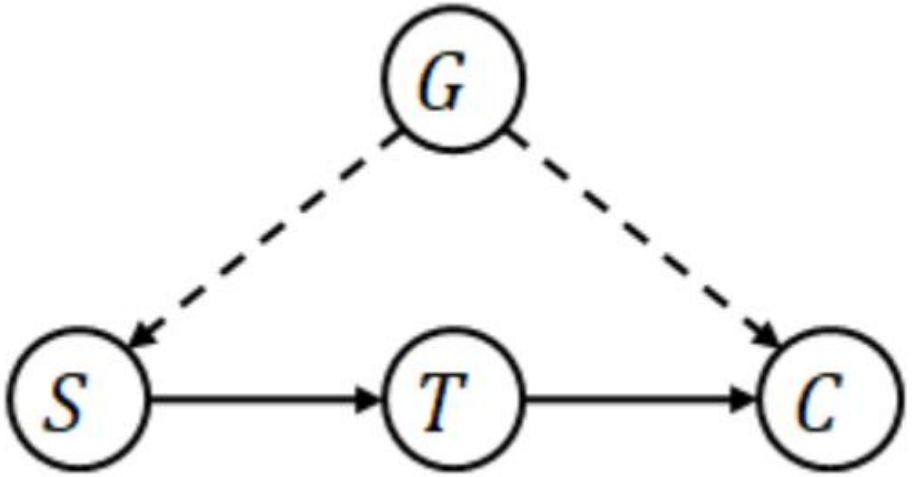
调整后

肾结石	A 疗法	B 疗法
小型	81/87 = 93%	234/270 = 87%
大型	192/263 = 73%	55/80 = 69%
总计	83.2%	78.18%

前门调整

图片来自网络

- 吸烟与肺癌。S=吸烟、T=焦油，C=肺癌，G=基因。
- 若我们只看非吸烟者，体内有焦油可以的患癌率从10%降到了5%；若们只看吸烟者，体内有焦油可以的患癌率从90%降到了85%，可见焦油有防护作用。
- 我们需要估计吸烟对肺癌的**因果效应**。



	组别	P(S,T) (每个组别所占百分比)	P(C=1|S,Z) (每组内患癌症的百分比)
S=0，T=0	非吸烟者，肺内无焦油	47.5	10
S=1，T=0	吸烟者，肺内无焦油	2.5	90
S=0，T=1	非吸烟者，肺内有焦油	2.5	5
S=1，T=1	吸烟者，肺内有焦油	47.5	85

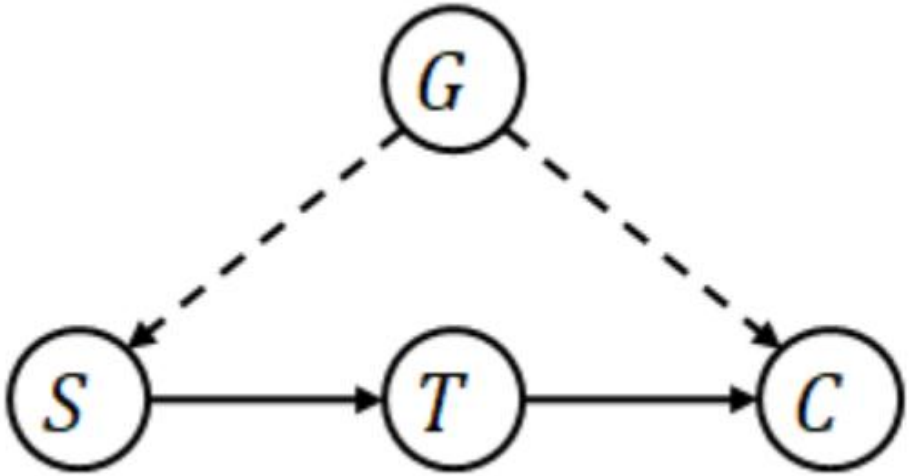
前门调整

后门调整

$$P_{\Theta}(C|do(S)) = \sum_{k} P_{\Theta}(C|S, G_k)P(G_k)$$

无从得知

图片来自网络



前门调整

$$P_{\Theta_1}(T|do(S)) = P_{\Theta_1}(T|S)$$ 无混杂

$$P_{\Theta_2}(C|do(T)) = \sum_k P_{\Theta_2}(C|T, S_k)P(S_k)$$ 后门调整

$$P_{\Theta}(C|do(S)) = \sum P_{\Theta_2}(C|do(T))\, P_{\Theta_1}(T|do(S))$$ 综上

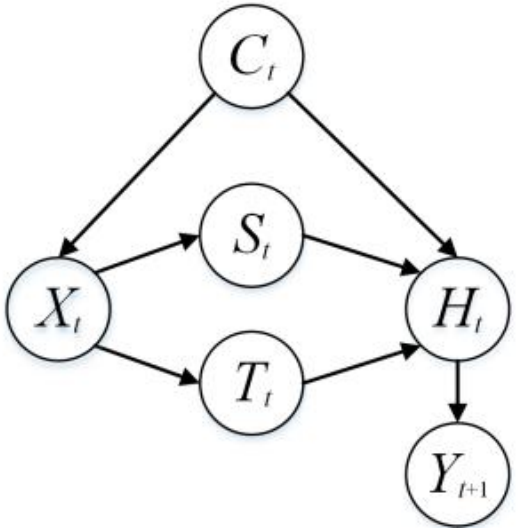
$$\boldsymbol{P_{\Theta}(C|do(S)) = \sum_{k_1} P_{\Theta_1}(T|S_{k_1}) \sum_{k_2} P_{\Theta_2}(C|T, S_{k_2})P(S_{k_2})}$$

大纲

研究动机

研究动机

- 时空数据同样存在辛普森悖论



C: Traffic context

X: Historical traffic flow data

Y: Ground truth of future traffic

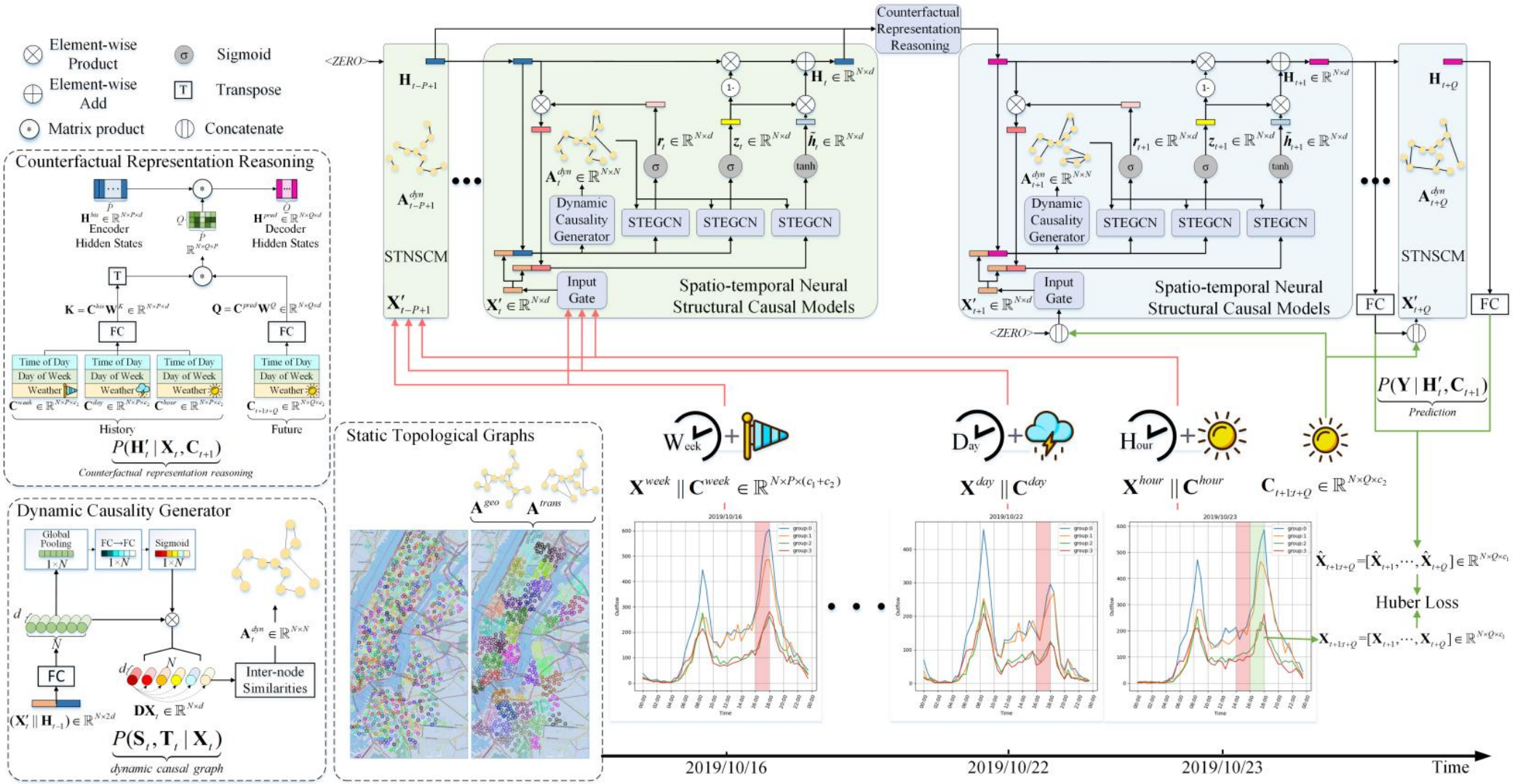
$$\boldsymbol{P(\mathbf{H}|\mathbf{X_t})}$$

$$P(\mathbf{H}_t|do(\mathbf{X}_t)) = \sum_{\mathbf{S}_t,\mathbf{T}_t} P(\mathbf{S}_t,\mathbf{T}_t|do(\mathbf{X}_t))P(\mathbf{H}_t|do(\mathbf{S}_t,\mathbf{T}_t))$$

$$= \sum_{\mathbf{S}_t,\mathbf{T}_t} P(\mathbf{S}_t,\mathbf{T}_t|\mathbf{X}_t) \sum_{\mathbf{X'}_t} P(\mathbf{H}_t|\mathbf{S}_t,\mathbf{T}_t,\mathbf{X'}_t)P(\mathbf{X'}_t) \quad (1)$$

Reference： Spatio-temporal Neural Structural Causal Models for Bike Flow Prediction.

方法介绍


Element-wise Product
Element-wise Add
Matrix product
Sigmoid
Transpose
Concatenate
Counterfactual Representation Reasoning
Encoder Hidden States
Decoder Hidden States
History
Future
Counterfactual representation reasoning
Dynamic Causality Generator
Global Pooling
FC→FC
Sigmoid
Inter-node Similarities
dynamic causal graph
Static Topological Graphs
STNSCM
Dynamic Causality Generator
STEGCN
Input Gate
Spatio-temporal Neural Structural Causal Models
Counterfactual Representation Reasoning
Prediction
Huber Loss
2019/10/16
2019/10/22
2019/10/23
Time

实验结果

dataset	Category	Models	30min			60min			Avg		
			MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
BJ-Bike	Deep Learning	LSTM	14.2031	28.2169	19.5308%	19.8906	41.4004	26.1706%	17.378	35.3137	22.9824%
		GRU	14.5063	30.5848	20.1777%	20.2897	42.8662	27.1207%	17.7293	37.0763	23.7914%
	Predefined Graph	STGCN	11.5225	32.3062	15.4271%	14.1583	36.1548	18.3616%	13.2019	33.4340	16.9863%
		STGODE	13.0722	25.5082	17.6931%	18.2863	38.2222	23.7129%	15.9745	32.0672	20.8288%
	Adaptive Graph	GWNet	11.3790	23.4735	15.7683%	14.4496	30.3216	20.0026%	13.0956	25.7935	17.9991%
		HGCN	12.3808	23.7278	16.5783%	14.4550	29.5130	19.1782%	13.6700	25.9349	18.0185%
		CCRNN	13.0028	40.7625	16.2663%	15.1600	43.8774	19.0467%	14.4291	40.7677	17.7535%
		DMSTGCN	11.3967	23.1091	15.6620%	14.1286	29.5651	18.9005%	12.9 921	25.6640	17.3526%
	Attention Graph	GMAN	14.2979	32.5408	19.4858%	19.5415	43.8546	25.7455%	17.3069	38.4888	22.7454%
		ASTGNN	13.0494	26.2419	17.4269%	17.8318	40.4308	23.1511%	15.8104	33.4645	20.4270%
	Dynamic Graph	DGCRN	11.4163	27.7059	16.0872%	14.0468	33.2321	19.2002%	12.9966	29.8171	17.7582%
		STNSCM	**11.2833**	**23.2897**	**15.2978%**	**13.3415**	**28.1254**	**17.4721%**	**12.5180**	**24.4383**	**16.4879%**
NYC-Bike	Deep Learning	LSTM	3.1259	6.0591	24.8426%	3.8342	7.9119	30.9824%	3.4809	6.9045	28.5997%
		GRU	3.1359	6.0629	24.8762%	3.8442	7.8694	30.9406%	3.4908	6.8827	28.6149%
	Predefined Graph	STGCN	2.6015	4.9071	20.5065%	2.9732	6.1072	23.3678%	2.7879	5.4049	22.4650%
		STGODE	2.7222	5.2398	21.4485%	3.2079	6.7118	25.4689%	2.9649	5.8408	23.8943%
	Adaptive Graph	GWNet	2.5686	4.8377	20.5743%	2.9589	6.1023	23.7937%	2.7644	5.3748	22.6851%
		HGCN	2.5865	4.9755	20.2884%	2.9728	6.4410	23.4383%	2.7803	5.5742	22.3060%
		CCRNN	2.5945	4.8986	20.4141%	2.9670	6.2435	23.4624%	2.7814	5.4886	22.4958%
		DMSTGCN	2.5539	4.7400	20.1860%	2.9159	6.0370	23.2154%	2.7395	5.3201	22.1471%
	Attention Graph	GMAN	3.1153	6.2649	23.6753%	3.1816	6.4593	24.7447%	3.1469	6.1648	24.7131%
		ASTGNN	2.9774	5.6568	23.3848%	3.3349	6.8282	26.2812%	3.1576	6.0232	25.3607%
	Dynamic Graph	DGCRN	2.6175	4.9426	20.5837%	2.9653	6.1419	23.3420%	2.7918	5.4210	22.4856%
		STNSCM	**2.5289**	**4.6835**	**19.8475%**	**2.8099**	**5.6129**	**22.4450%**	**2.6701**	**5.0397**	**21.5300%**

Thanks!

张文涛

wechat : zwt532586242