基于因果推断的广告投后归因

飞猪算法平台：章凡

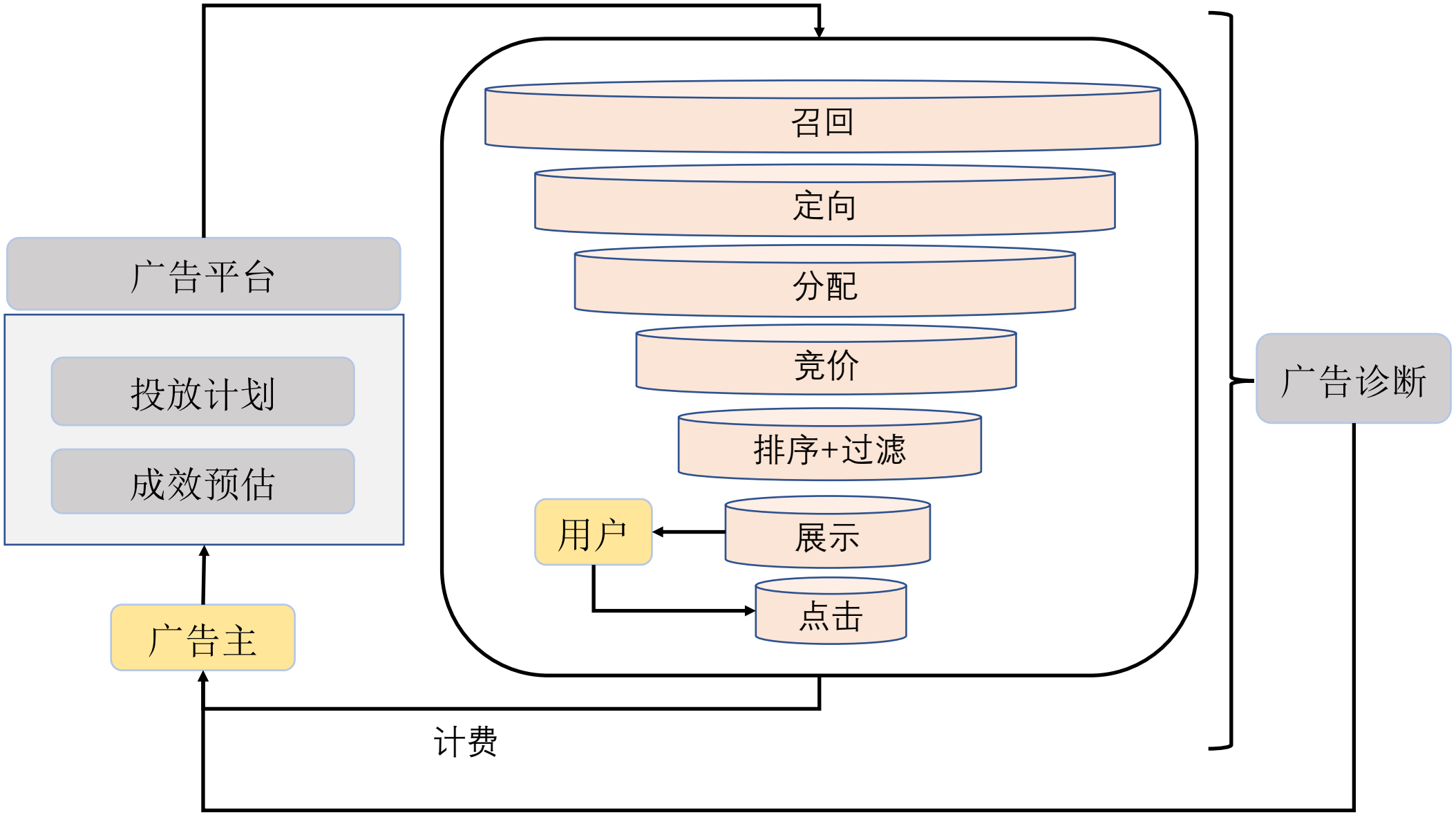
飞猪广告诊断系统简介

飞猪商业化广告主要为商家提供广告投放服务，酒店搜索广告诊断系统主要为广告主和广告运营提供服务，帮助广告主更及时、更准确的发现广告投放过程中存在的问题，并指导广告主对广告投放进行优化。。

通常在出现投放异常时，往往需要值班人员去检查整个投放漏斗，得出结论的过程需要丰富的经验。但随着广告酒店数量增多，快速诊断每个计划的投放情况变得非常复杂且耗时。因此我们需要一个能够快速定位问题的诊断系统。

随着因果推断技术的快速发展，我们开发了一套基于因果推断技术的归因方法。

飞猪广告诊断系统简介



基于因果发现的归因方法

贝叶斯网络应用于许多领域：

- 生物遗传学
- 机器学习
- 因果推断

构建有向无环图（directed acyclic graphs - DAG）也就是贝叶斯网络（BN）是一个NP-hard的问题，因为DAG的搜索空间是组合的，并且与节点数成超指数比例。

构建贝叶斯网络的方法：

- 基于贝叶斯统计的评分： BDeu
- 基于信息理论评分：MDL
- 搜索算法：HillClimbing算法、SGS 算法、PC 算法

$$\begin{aligned} \min_{\mathsf{G}} \quad & Q(\mathsf{G}) \\ \text{subject to} \quad & \mathsf{G} \in \mathbb{D} \end{aligned}$$

现有方法的一些弊端：

- 强依赖一些先验知识的假设
- 随着节点数量的增加，想要加速搜索方法需要很多技巧
- 概念复杂

基于因果发现的归因方法

主要参考：*DAGs with NO TEARS: Continuous Optimization for Structure Learning*

$$\begin{aligned} \min_{W \in \mathbb{R}^{d\times d}} \quad & F(W) \\ \text{subject to} \quad & \mathsf{G}(W) \in \mathsf{DAGs} \end{aligned} \iff \begin{aligned} \min_{W \in \mathbb{R}^{d\times d}} \quad & F(W) \\ \text{subject to} \quad & h(W) = 0, \end{aligned}$$

- 显式的构造了一个平滑可导的方程来刻画无环约束
- 开发了一个等式约束程序，用于从可能的高维数据中同时估计稀疏DAG的结构和参数，并展示了如何使用标准数值解算器来寻找平稳点。

Theorem 1. *A matrix $W \in \mathbb{R}^{d\times d}$ is a DAG if and only if*

$$h(W) = \operatorname{tr}\left(e^{W\circ W}\right) - d = 0,$$

where $\circ$ is the Hadamard product and e^A is the matrix exponential of A. Moreover, $h(W)$ has a simple gradient

$$\nabla h(W) = \left(e^{W\circ W}\right)^T \circ 2W,$$

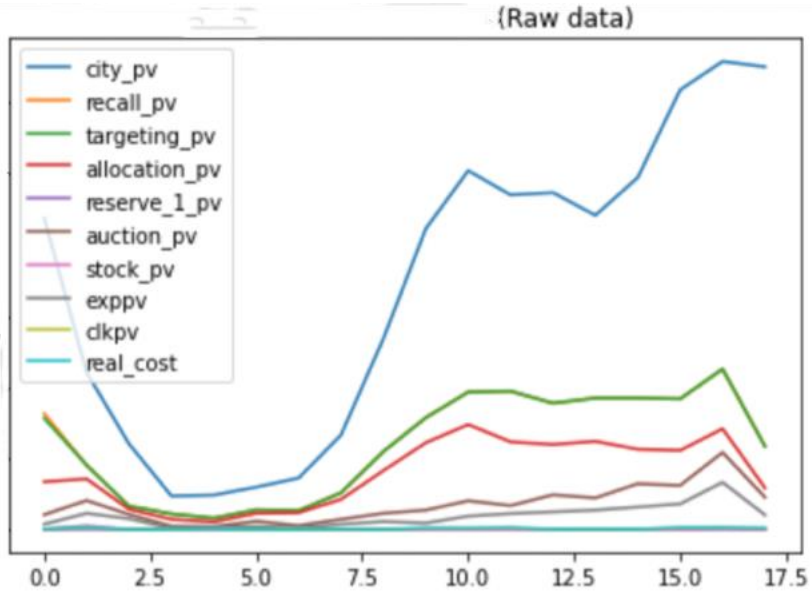
and satisfies all of the desiderata (a)-(d).

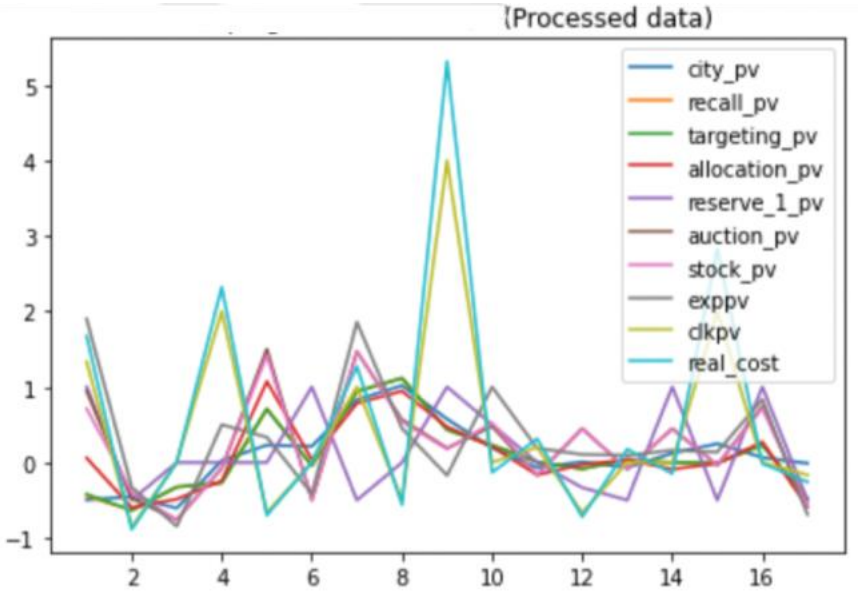
基于因果发现的归因方法

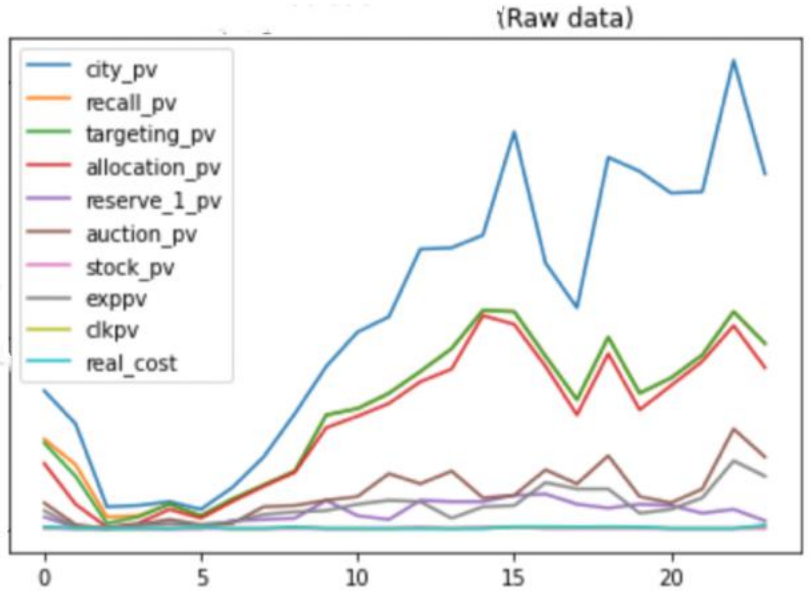
数据处理：

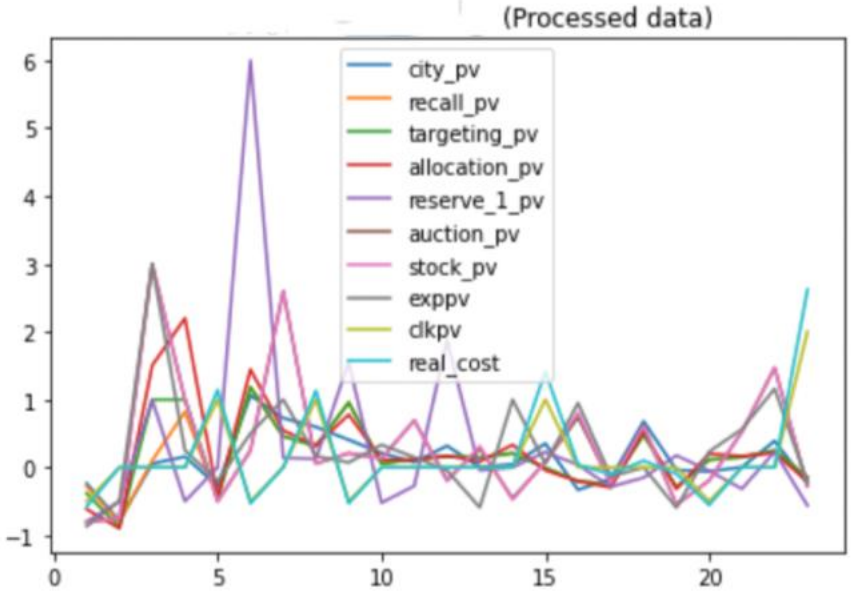
在前期实验过程中，使用了各种包括归一化、标准化、离散化、计算通过率等方法处理数据均没有效果。因为No Tears算法只能处理变量的变化值。

因此采集每个关注变量的绝对值，计算其相对t-1时刻的变化量。







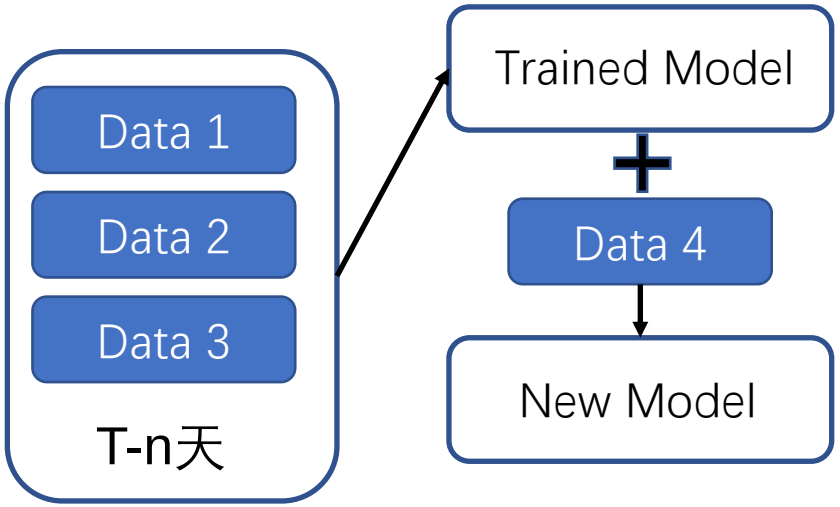


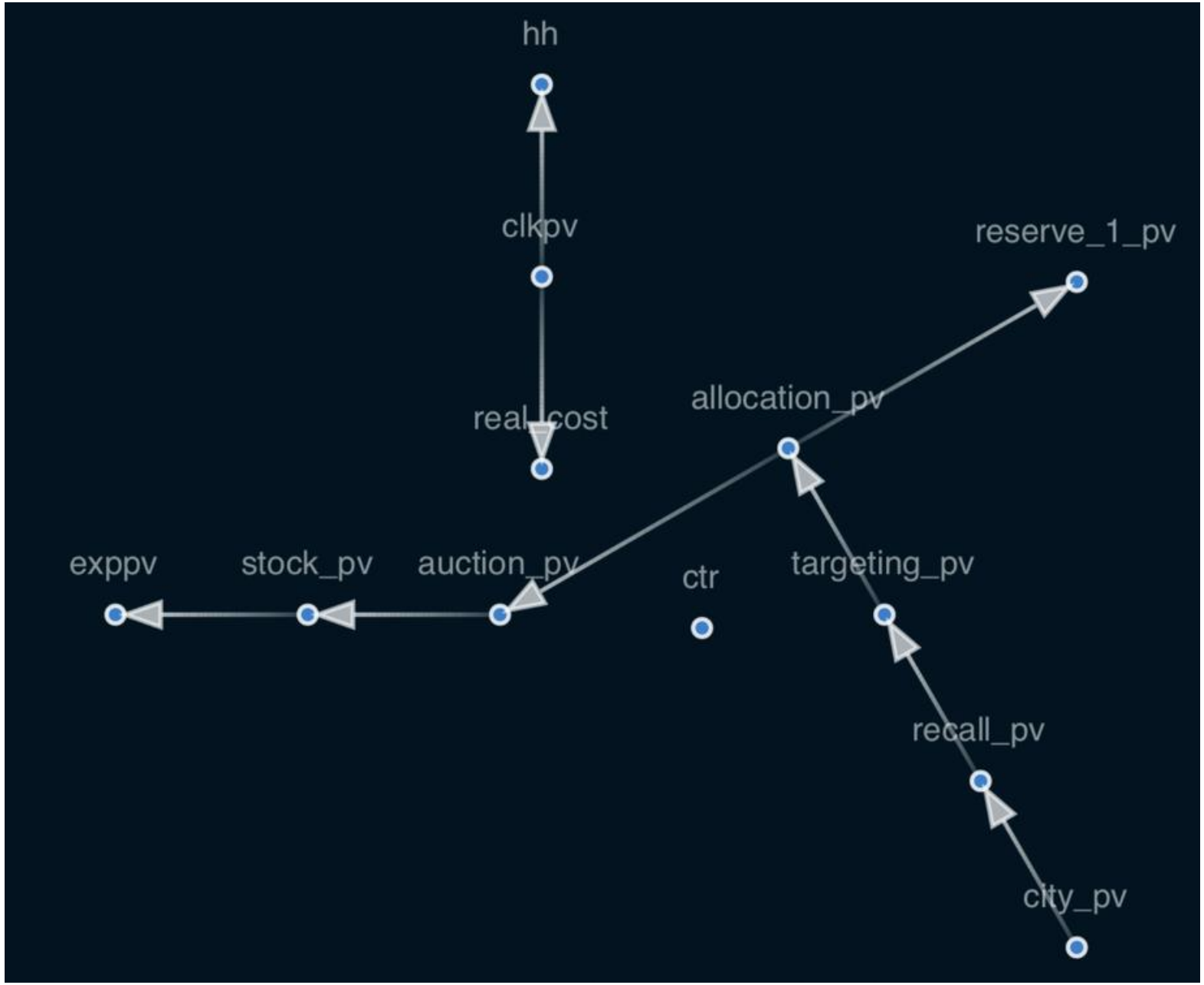
基于因果发现的归因方法

归因方法：

采集并处理t-n天的数据训练模型，使用当天归因的数据在训练好的模型上进行增量训练，计算邻接矩阵权值的变化量作为归因权重，变化幅度越大，则其归因权重越大。

将多个因子的归因权重进行排序，结果即为归因结果。





基于因果发现的归因方法

基于No Tears的归因准确率为80%。No Tears的因果发现归因缺点：

1. 计算速度较慢
2. 结果较为黑盒，不好干预
3. 输入变量需要较精细化处理

阿里达摩院提出更快速的因果发现方法：

Efficient and Scalable Structure Learning for Bayesian Networks: Algorithms and Applications

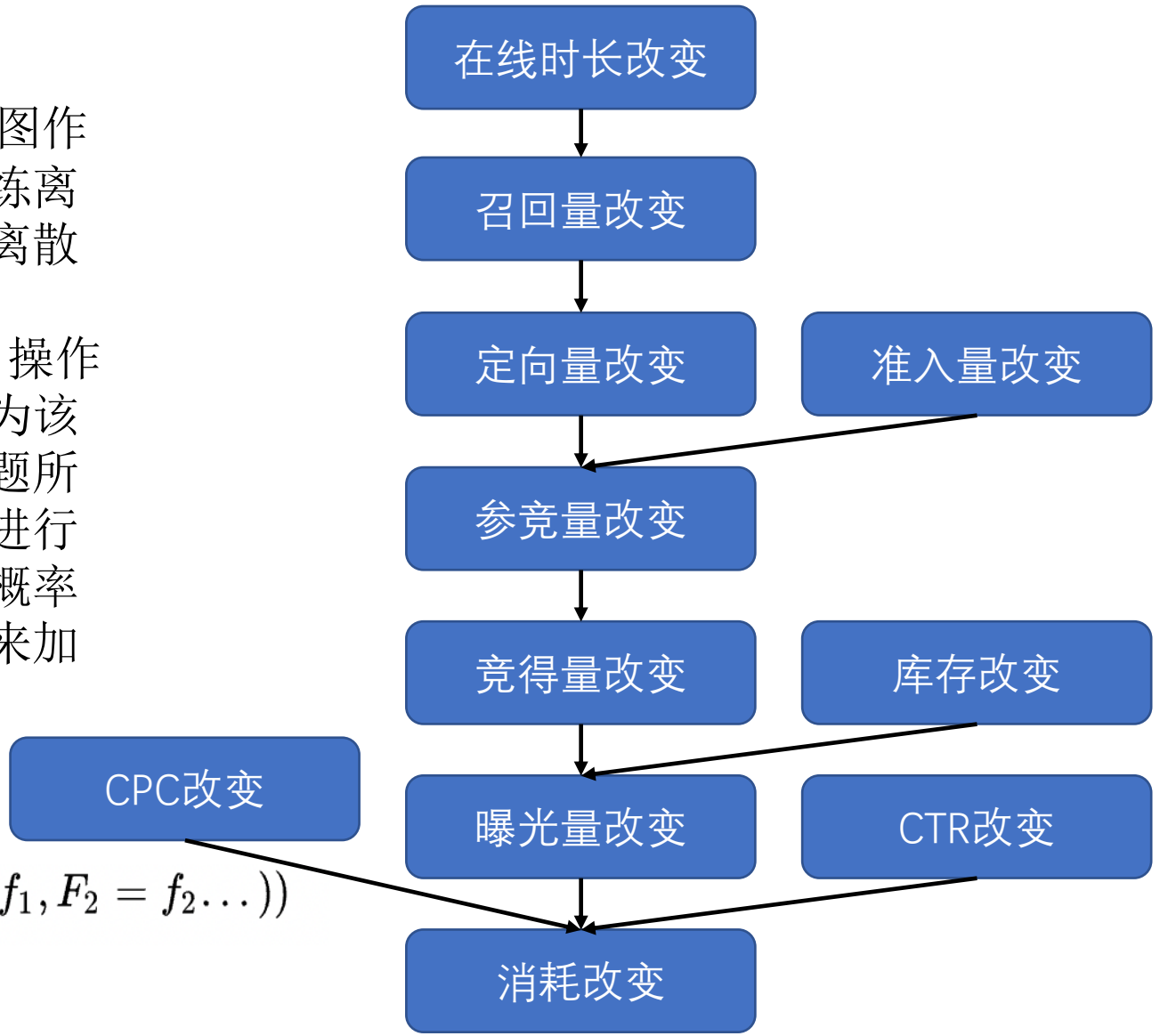
	hh	city_pv	recall_pv	targeting_pv	allocation_pv	reserve_1_pv	auction_pv	stock_pv	exppv	clkpv	real_cost	ctr
hh	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0
city_pv	0.000000	0.0	0.521339	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0
recall_pv	0.000000	0.0	0.000000	1.109799	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0
targeting_pv	0.000000	0.0	0.000000	0.000000	0.798122	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0
allocation_pv	0.000000	0.0	0.000000	0.000000	0.000000	0.674038	0.830487	0.00000	0.000000	0.0	0.000000	0.0
reserve_1_pv	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0
auction_pv	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.95802	0.000000	0.0	0.000000	0.0
stock_pv	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.911321	0.0	0.000000	0.0
exppv	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0
clkpv	0.600378	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	2.220125	0.0
real_cost	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0
ctr	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.0

基于贝叶斯网络的归因方法

根据No Tears方法构建好因果图之后，以因果图作为贝叶斯网络的结构。将数据离散化，使用数据训练离散贝叶斯网络获得条件概率表。其中消耗变化量被离散为5个值。

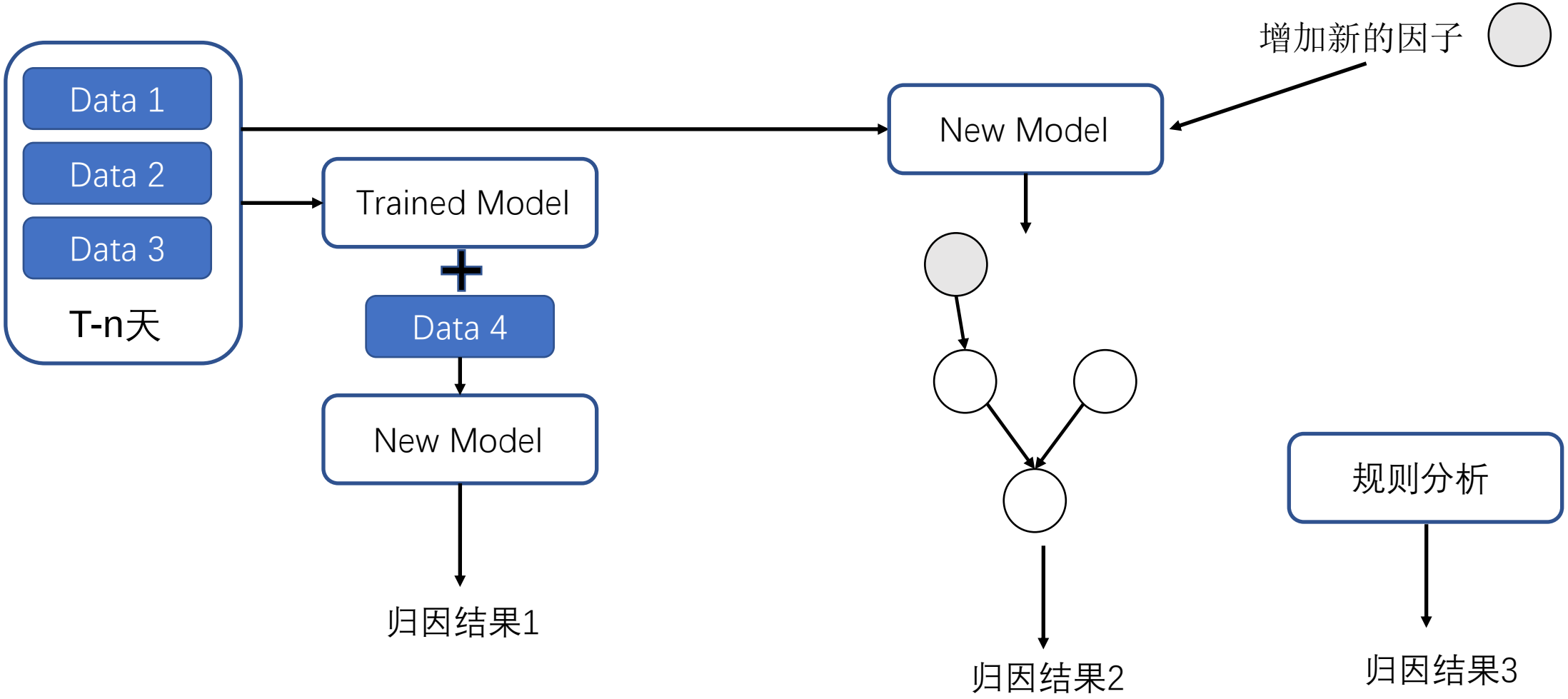
对于每个计划的消耗波动，使用*Do-Calculus* 操作计算“介入”每个因子对目标结果发生的概率，即为该因子的归因权重。除此之外，为了更好体现漏斗问题所在，我们设计对归因结果使用当前因子的后验概率进行微调来体现因子的波动是否受上游影响。通过查询概率表，计算当前因子的父节点对当前因子的后验概率来加权归因权重，最终的归因权重为：

$$weight_X = P(cost = c|Do(X) = x) * (1 - P(X = x|F_1 = f_1, F_2 = f_2 \ldots))$$



最终离散贝叶斯网络的归因准确率为85%。

基于贝叶斯网络的归因方法



感谢观看

欢迎与我交流&加入我们
邮箱：buyi.zf@alibaba-inc.com