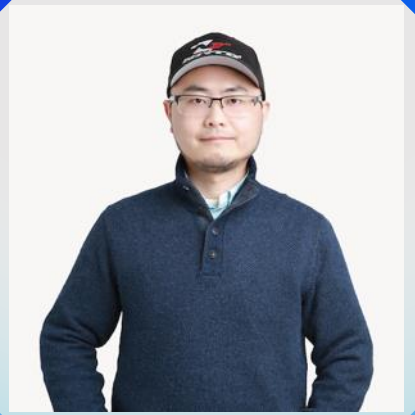
滴滴数据科学实践及展望

谢梁
滴滴首席数据科学家



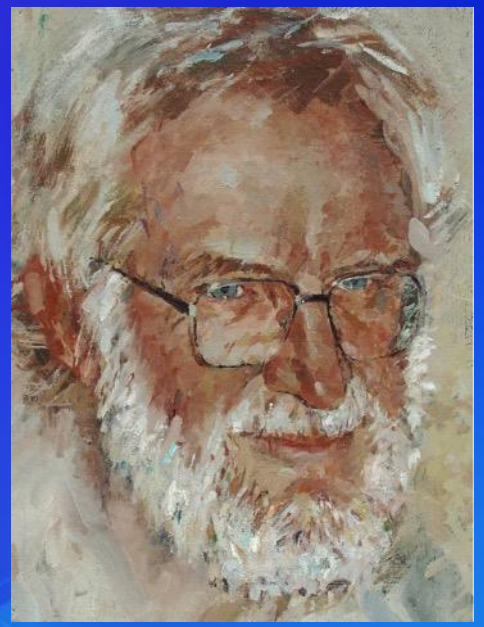
谢梁

滴滴首席数据科学家，纽约州立大学计量经济学博士，在滴滴主持运用机器学习和人工智能方法分析优化大规模交易市场效率和系统行为模式。 具有十余年ML/AI应用经验，熟悉各种业务场景下的应用，行业跨度包含金融，能源和高科技。加入滴滴之前担任微软总部云计算核心存储部首席数据科学家。

目录

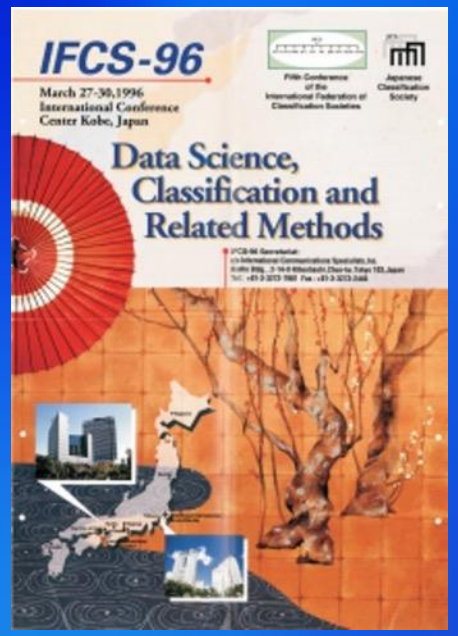
数据科学（DS）的历史

数据科学的历史



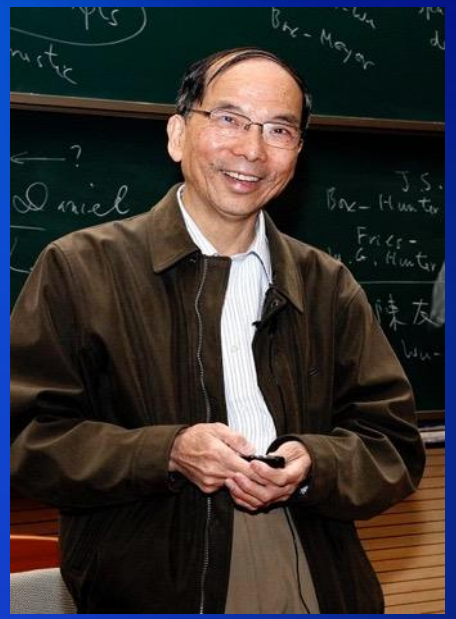
1974

第一次完整提出“Data Science”概念



1996

3月，International Federation of Classification Society



1997

提出Statistics = DS

2001

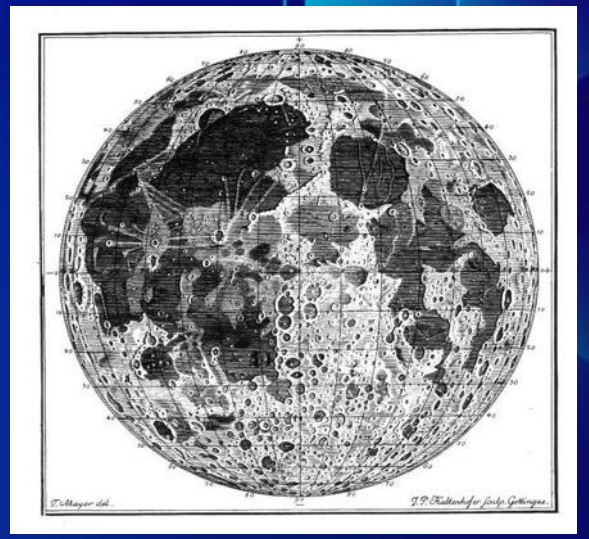
DS学科发展计划：提出6个支柱

数据科学的历史

- Tobias Mayer：月球的运动轨迹是怎样的?

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \cos\phi\sin\phi & 0 \\ -\sin\phi\cos\phi & 0 \\ 0 \quad 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\boldsymbol{\omega} = \begin{bmatrix} -\sin\phi\sin\theta \\ -\cos\phi\sin\theta \\ \cos\theta \end{bmatrix};$$



最小二乘估计

- 潘石屹：海南的房价合理吗?
  - 人均居住面积 = （现有楼盘+待开发楼盘面积） / 海南人口
  - 海南省人均居住面积 >>全国平均

滴滴数据科学前传

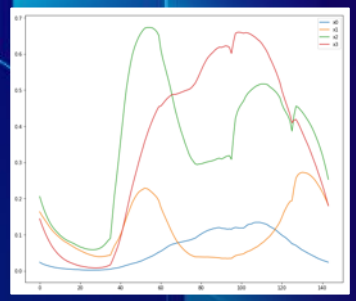
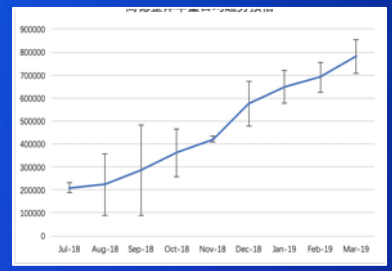
滴滴DS的前传

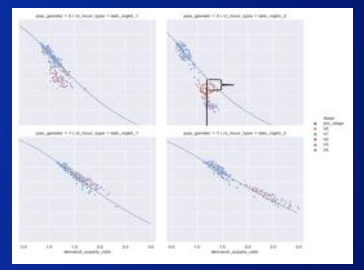
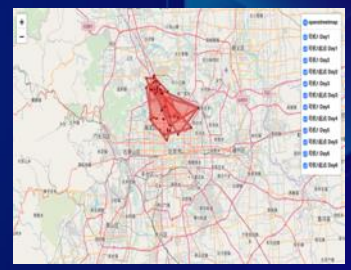
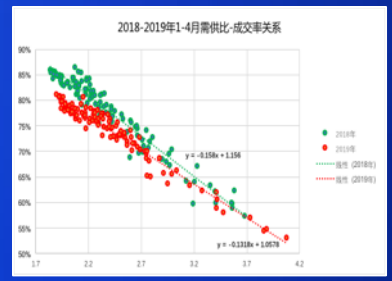
- 滴滴数据科学于2017年10月成立
- 行业传统BI以报表制作为主
  - 业务方给定数据需求
  - 数据分析师执行该具体需求
  - 以证明业务方想法正确为主线

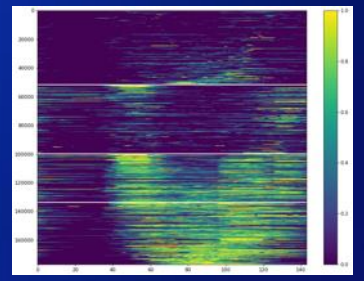
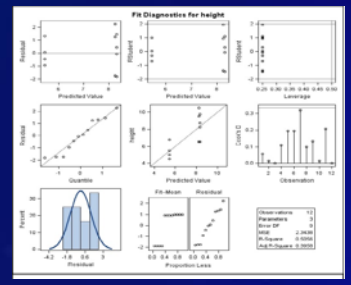
* 图片来自于网络

DS在滴滴带来的改变

- 使命：数据驱动，让决策更科学
- 愿景：用科学的理论和扎实的数据洞察影响决策：
  - 描述现状
  - 寻找规律
  - 推动改进

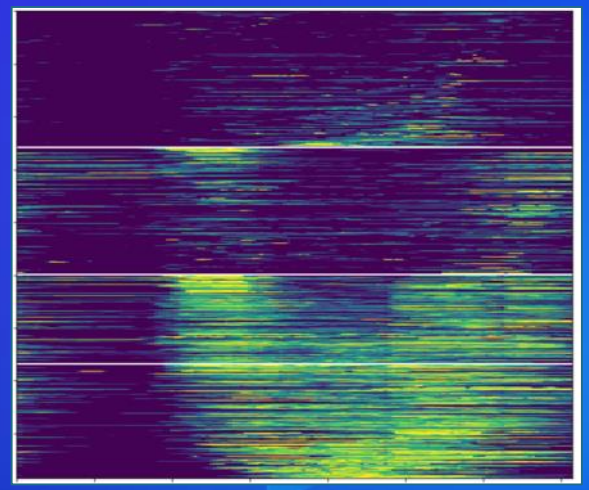



观测分析：ML + 多学科综合

观测分析：ML + 多学科综合

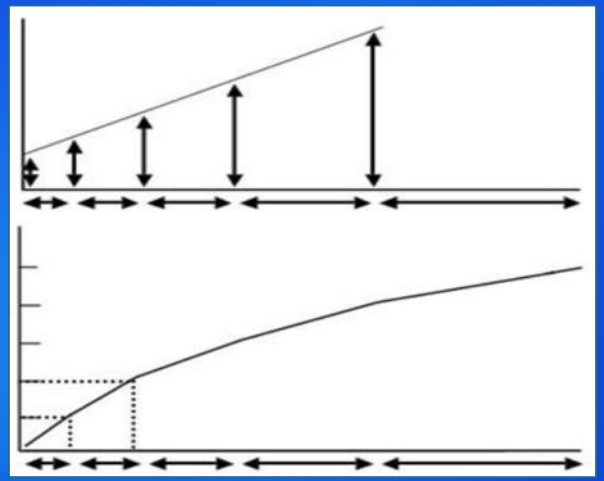
矩阵分解理论的应用：

基于多元奇异谱分析（MSSA）的运力分析



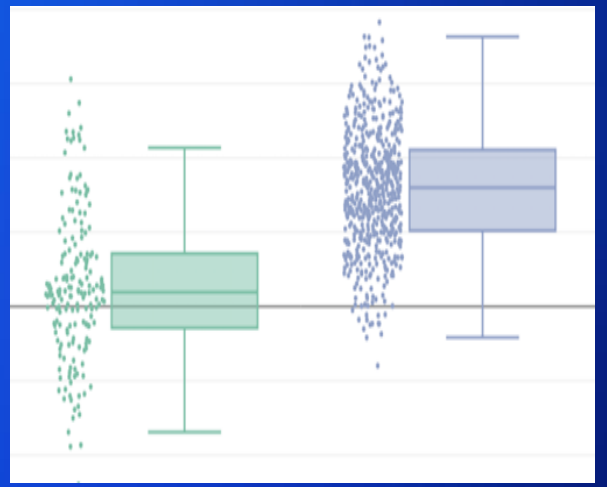
心理学理论的应用：

基于费希纳定律的用户行为分析



行为经济学理论的应用：

基于前景理论的策略效果评判

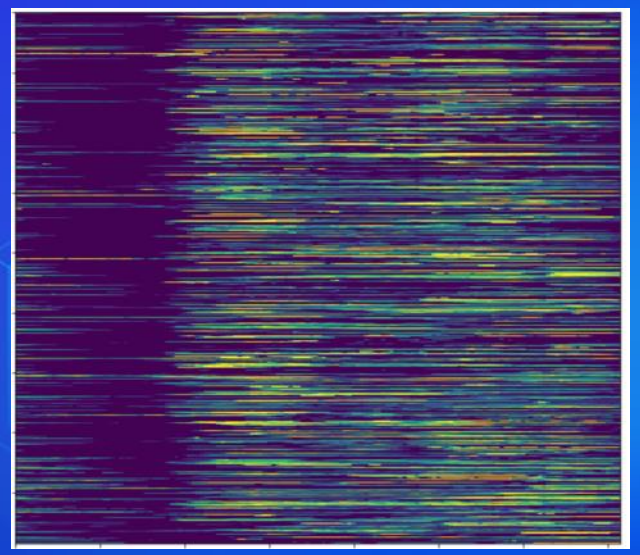


计量经济学的应用：

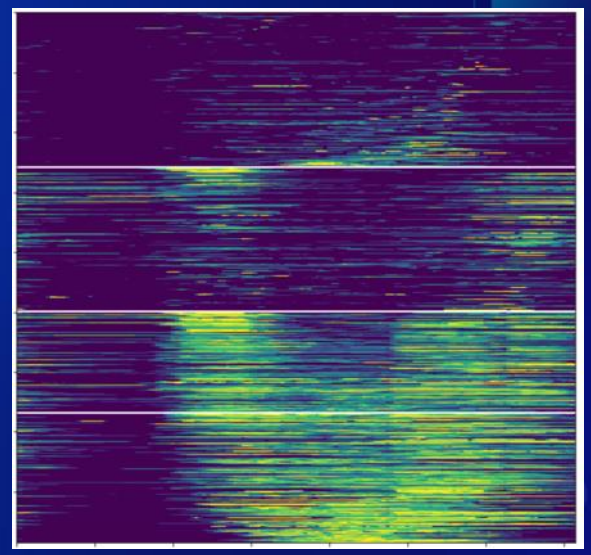
使用复杂离散选择模型刻画用户选择偏好

基于MSSA算法的运力分布分析

- 自动化挖掘每个城市的运力结构
- 自动定义运力分层特征



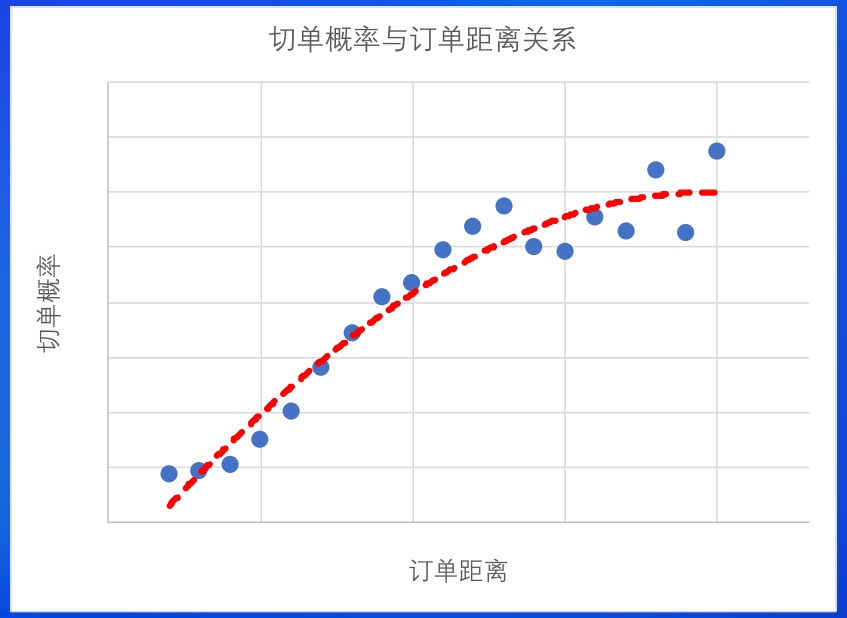
$$\begin{pmatrix} \mathbf{C}_{1,1} & \mathbf{C}_{1,2} & \cdots & \mathbf{C}_{1,D} \\ \mathbf{C}_{2,1} & \mathbf{C}_{2,2} & \cdots & \mathbf{C}_{2,D} \\ \vdots & \vdots & \mathbf{C}_{d,d'} & \vdots \\ \mathbf{C}_{D,1} & \mathbf{C}_{D,2} & \cdots & \mathbf{C}_{D,D} \end{pmatrix}$$

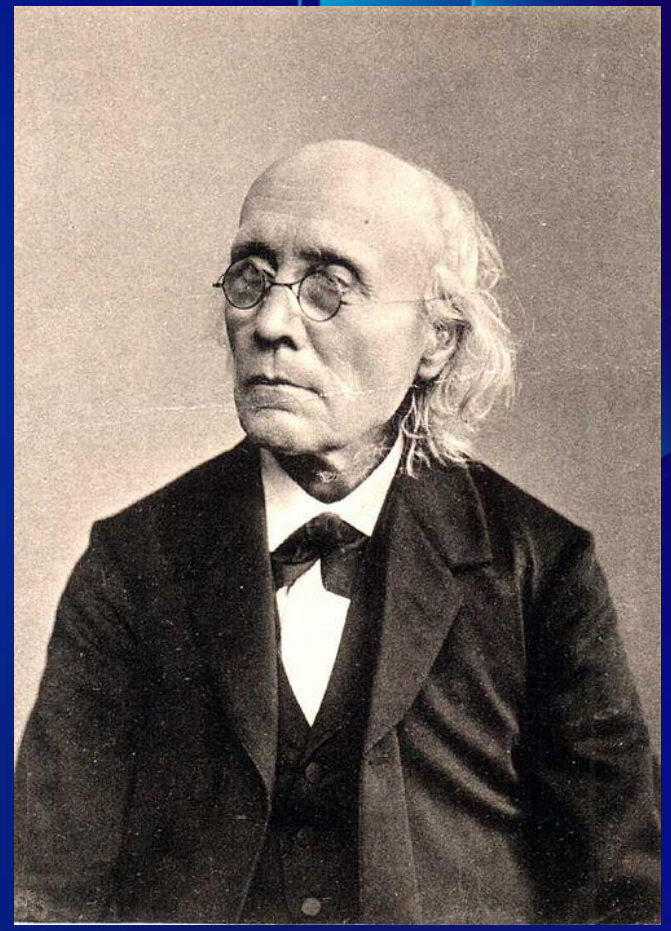


基于费希纳定律的用户分析和策略生成

怎么设计司机端奖励才能在降低线下交易的同时提高ROI?

$$\delta S = \beta \frac{\delta P}{P} => S = \text{Constant} + \beta \log(P)$$



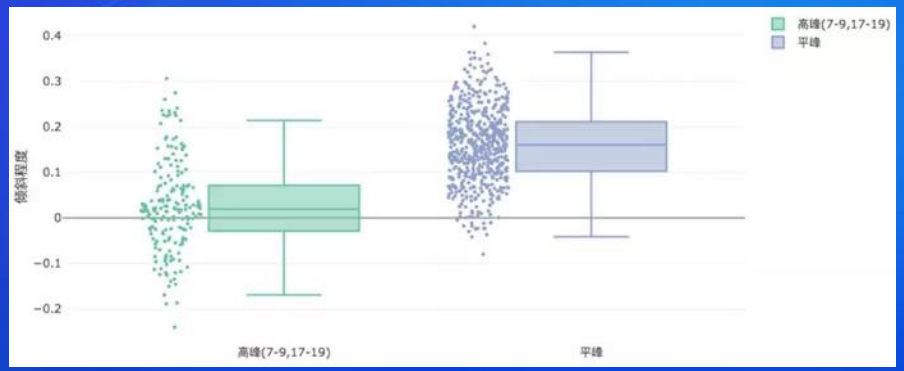


基于前景理论的策略效果评判

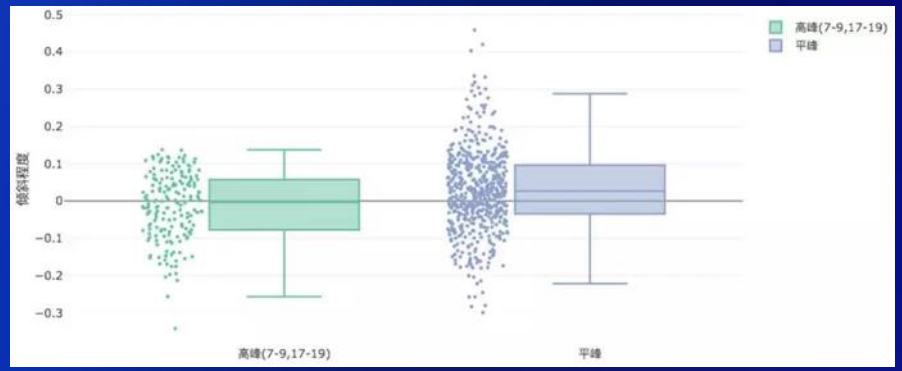
1. 一种新交易策略相对于原有策略可以明显改善乘客体验，但是在一些城市这个策略并不受高服务分司机的欢迎。为什么呢?

2. 我们发现新策略下订单分布在平峰期高度有偏，应用前景理论的损失厌恶概念，我们发现这种偏离原有策略订单分布的情况造成很强的负向心理感知，从而造成司机的抵触。也为改进提供了指导原则。

原有策略下司机感知的倾斜力度

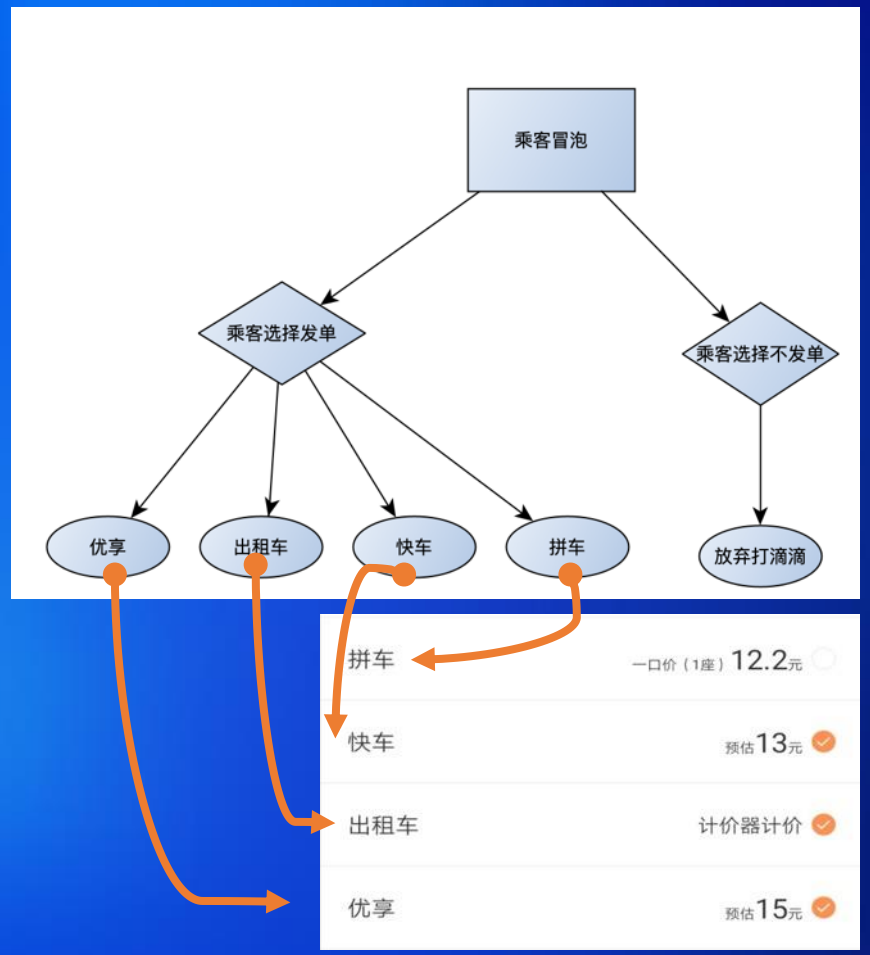


TopSpeed策略下司机感知的倾斜力度



使用离散选择模型刻画用户选择偏好

用户在端上的选择行为可以使用计量经济学中的嵌套Logit模型进行刻画和分析，从而辅助补贴方案的开发、体验产品的设计以及策略产品的迭代。

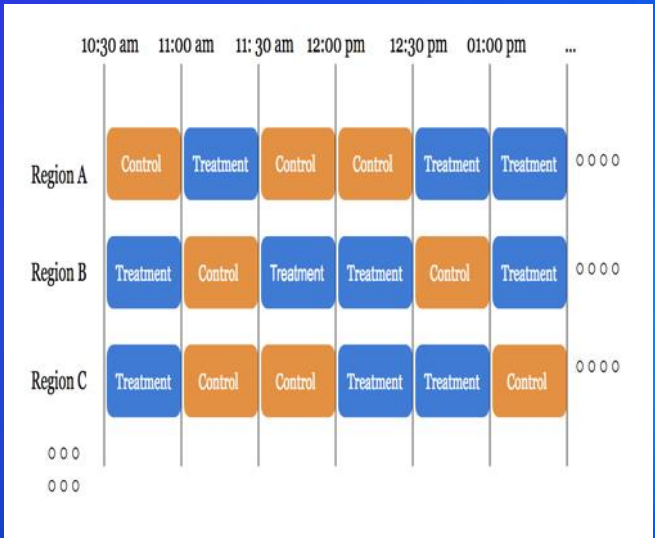


主动式探索：科学运营实验

主动式探索：科学运营实验

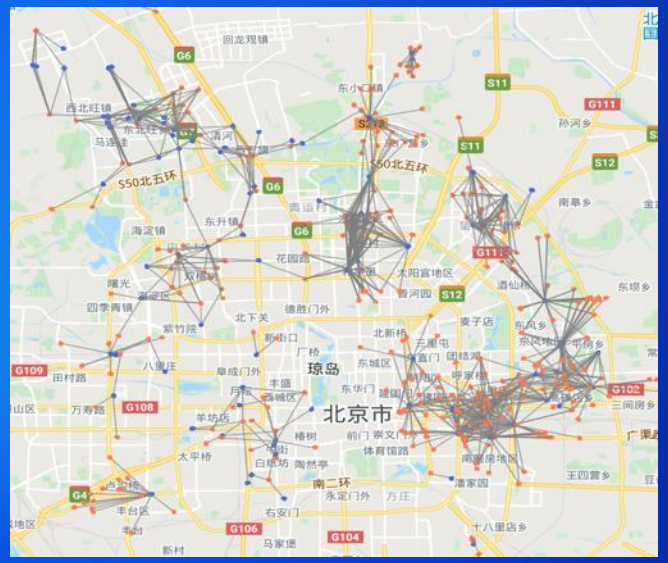
轮转实验：

不能在同一时空下对不同用户采用不同策略怎么办



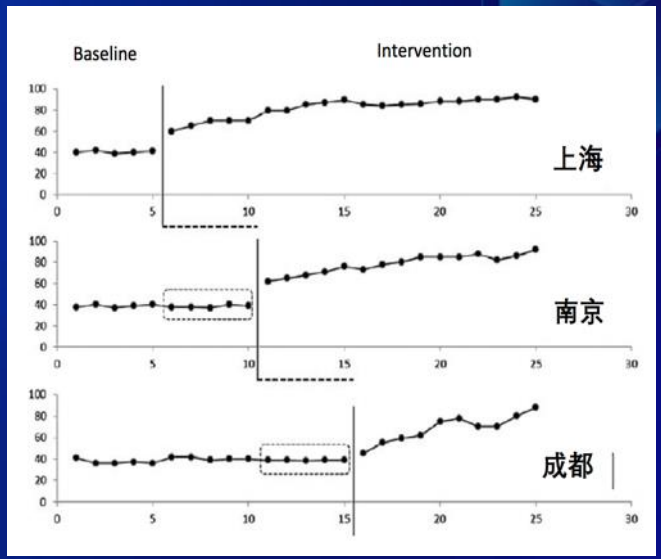
随机饱和实验：

在有溢出效应场景下的实验设计和操作



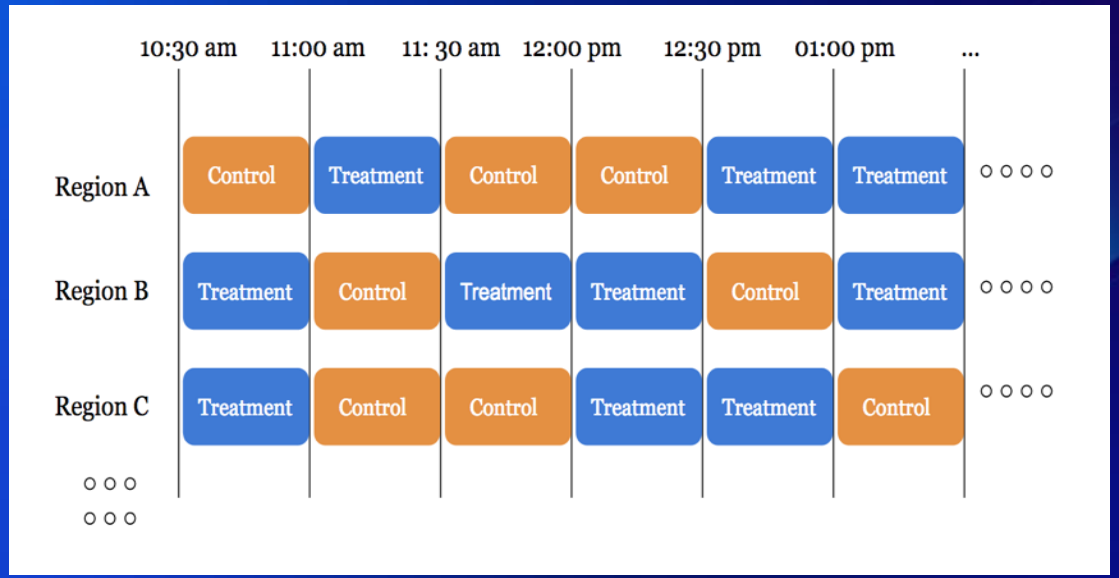
多基线实验：

全局性策略如何进行实验？



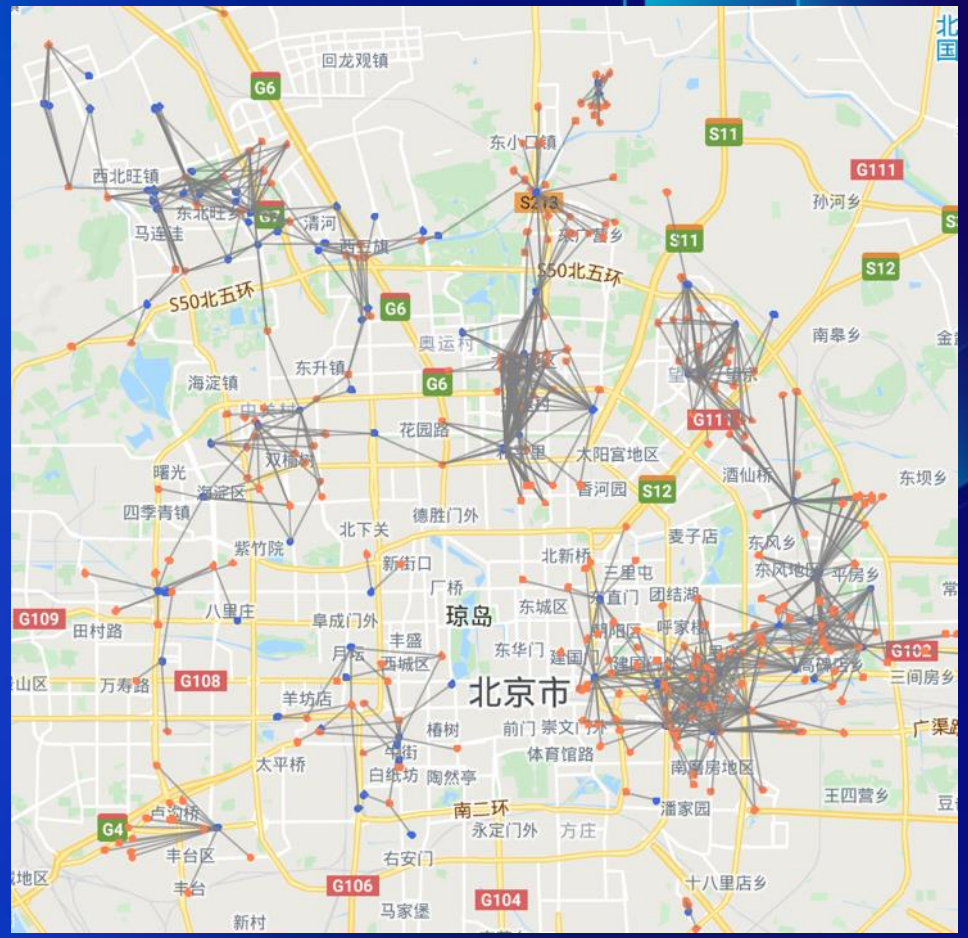
轮转实验

- 策略实验需要谨慎，尤其对于网约车这种涉及消费的场景，一定要杜绝大数据杀熟的问题
- 通过轮转实验，在保证了同一时空下的用户适用同样的策略情况下，也能有效获取统计显著的实验结果



带溢出效应场景下的实验设计

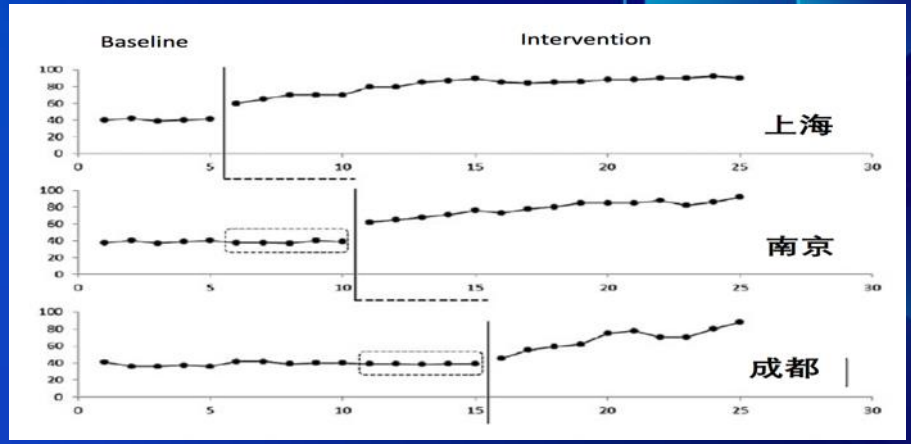
- 双边市场的统计实验有溢出效应，简单的AB分流实验会造成平均实验效果（ATE）误读
- 通过基于Geo-Spatial的聚类实现特殊的实验设计（Randomized Saturated Design)
- 特殊的设计和分析模型能尽量降低溢出效应的影响，看清纯粹实验因子带来的效果

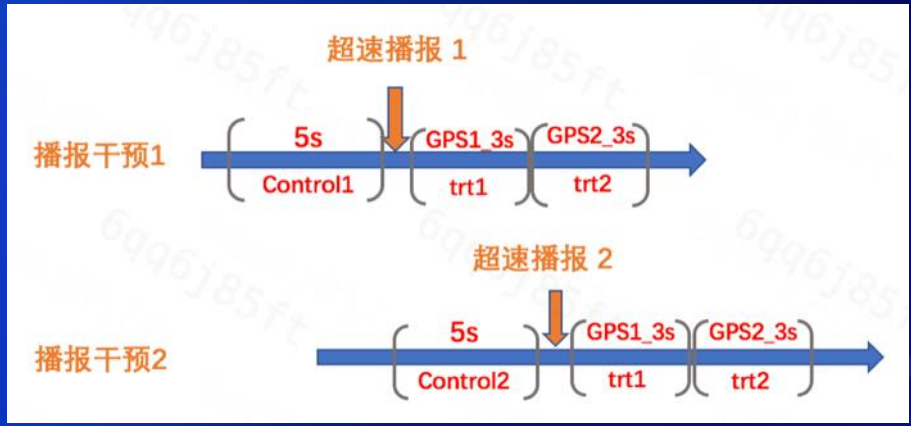


多基线实验

- 有些实验有很强的约束条件：
  - 策略上了以后不能反转
  - 保证同一时空的用户都被同样对待
  - 实验要有“温度”

- 全局性策略是该场景下实验的重要应用。采用多基线（MBL）设计，能在满足上述约束条件下获得一定统计显著性的结果，持续迭代





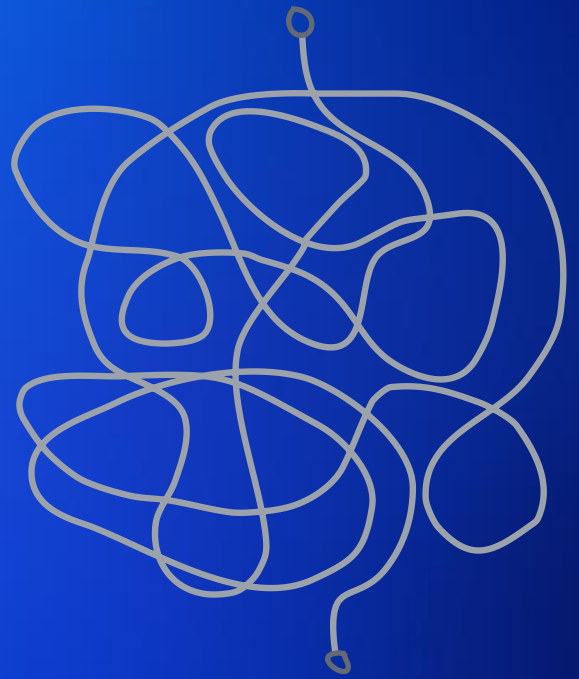
基于SQLFLow的自助式DS及展望

什么是SQLFlow

SQLFlow
=SQL + AI
= 人工智能大众化、普及化

- 只要懂商业逻辑就能用上人工智能，
- 让最懂业务的人也能够自由地使用人工智能

传统人工智能解决方案



VS

SQLFLow

SQLFlow的定位

普通BI分析师**即取即用**的数据和**AI赋能**的工具

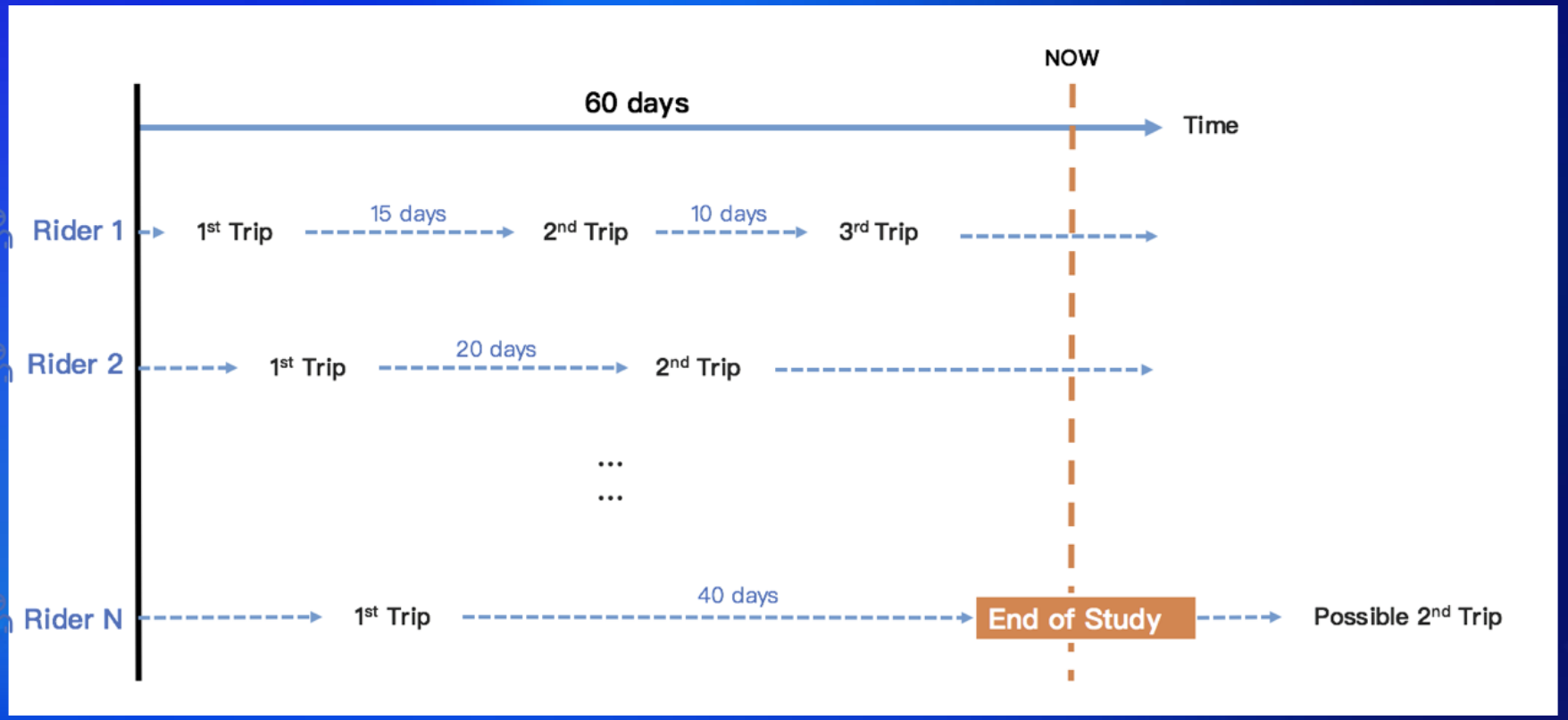
SQLFlow的应用

SQLFLow目前已经在蚂蚁金服和滴滴的实际业务中开始应用，受到DS、运营分析、PM的喜爱

	业务场景	数据库系统	AI 训练和预测	数学模型	模型解释
滴滴	商业智能（BI）	Apache Hive	xgboost	树模型	Shap
			Keras	聚类分析	
蚂蚁金服	精准营销	Alibaba MaxCompute	TensorFlow	深度神经元网络	

SQLFlow的应用：乘客活跃度因子分析

如何快速理解乘客活跃度及其影响因子，并应用于运营?



SQLFlow的应用：乘客活跃度因子分析

属性

Character

用户生命周期

注册天数

用户等级

价格

Price

平均折扣力度

预估价格

已付金额总量

体验

Experience

用户需求次数

接驾距离

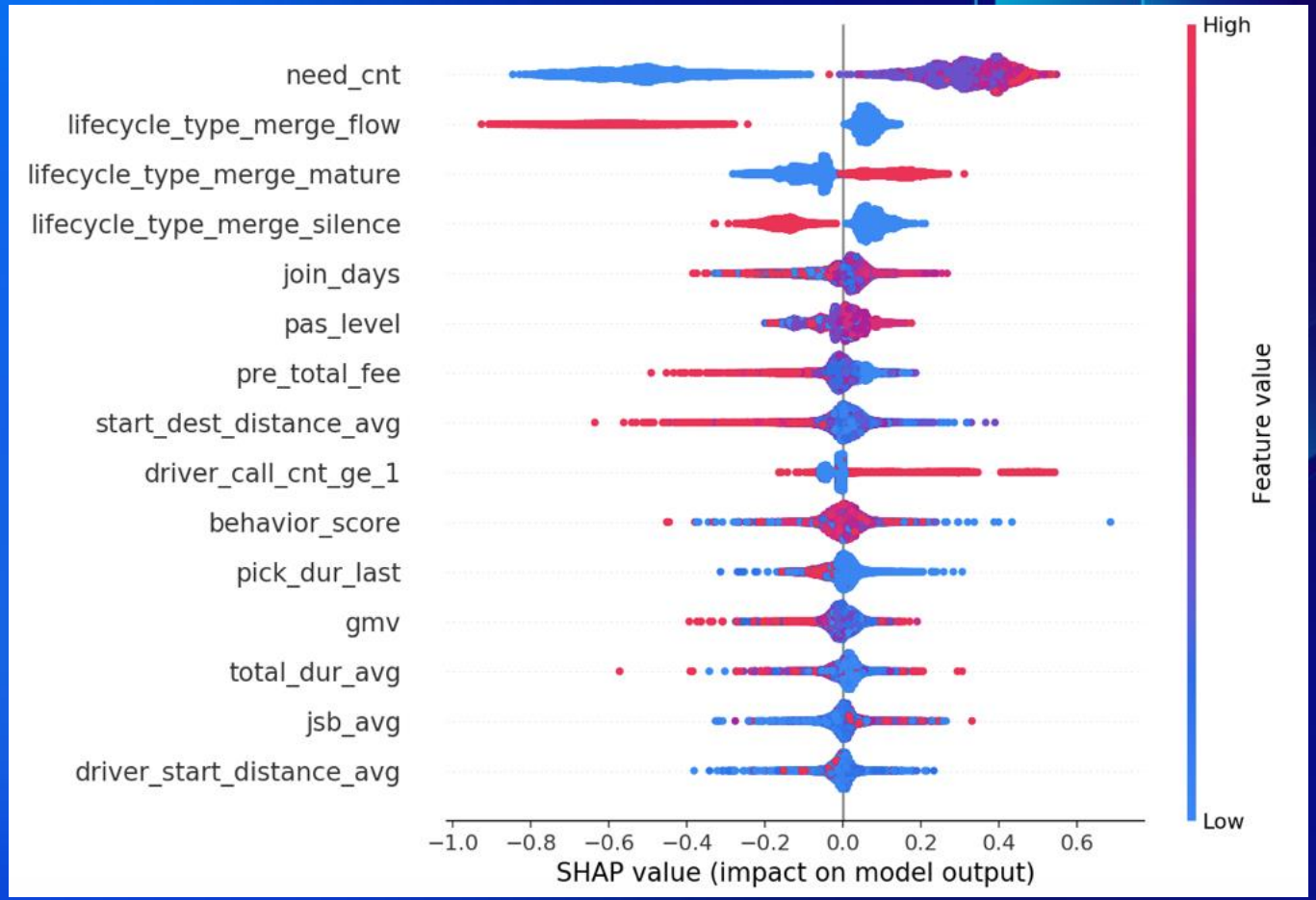
预估订单距离

预估订单距离

预估订单距离

SQLFlow的应用：乘客活跃度因子分析

- 使用Xgboost模型和Shapley值对潜在影响因子进行挖掘
- 一个模型可以被各城市运营直接在SQL中调用，并得到属于本城市的因子分析
- DS和城市BI也可以开发并发布新模型进行迭代升级



SQLFlow加持下未来DS的展望

让人工智能成为一种货架产品

业务需求和AI能力交易的市场

总结

1. 数据科学（DS）的历史
2. 滴滴DS前传
3. 观测分析：机器学习 + 多学科综合
4. 主动式探索：科学运营实验
5. 基于SQLFLow的自助式DS及展望

Thanks For Watching

本PPT来自2019携程技术峰会

更多信息请关注“携程技术中心”微信公众号~