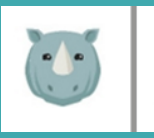


IT大咖说 知识共享平台

大规模机器学习在金融行业的应用

翟英博
第四范式咨询顾问

CONTENTS

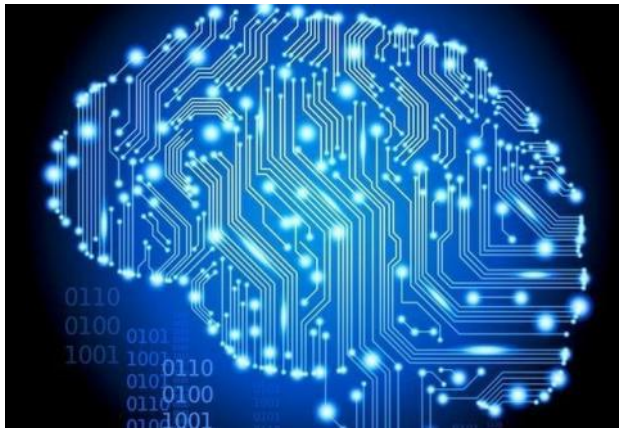
目录

背景

- 现代机器学习领域，高VC维模型已成为发展方向
- 分布式优化需要解决海量数据和海量特征参数高效训练的问题
- 通常的分布式计算方法如Spark，在数据传输、容错设计、同步机制等，效率无法满足需求

机器学习计算的特点

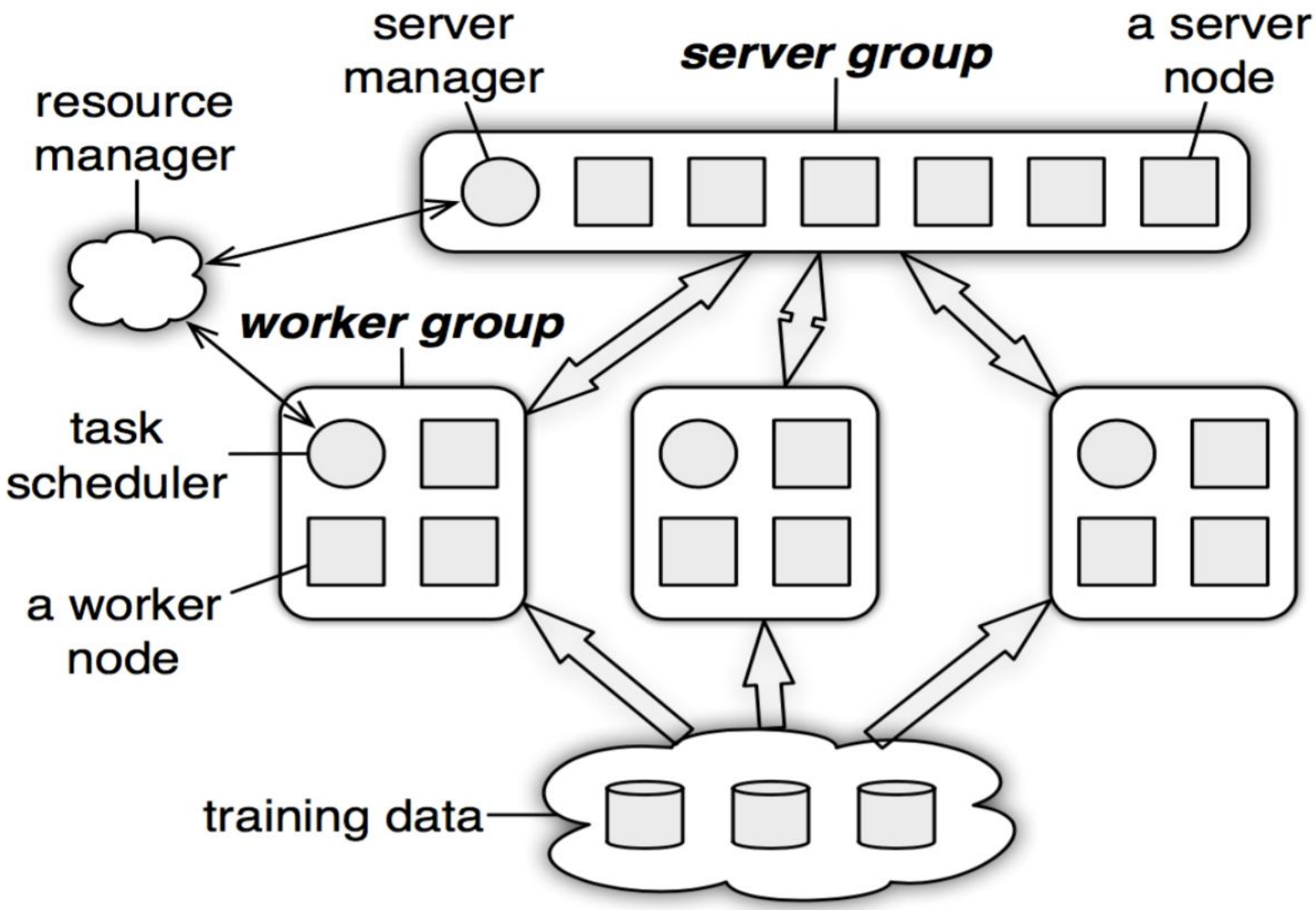
- 迭代性，模型的更新并非一次完成，需要循环迭代多次
- 容错性，即使在每个循环中产生一些错误，模型最终的收敛不受影响
- 参数收敛的非均匀性，模型中有些参数经过几个循环便不再改变，其他参数需要很长时间收敛



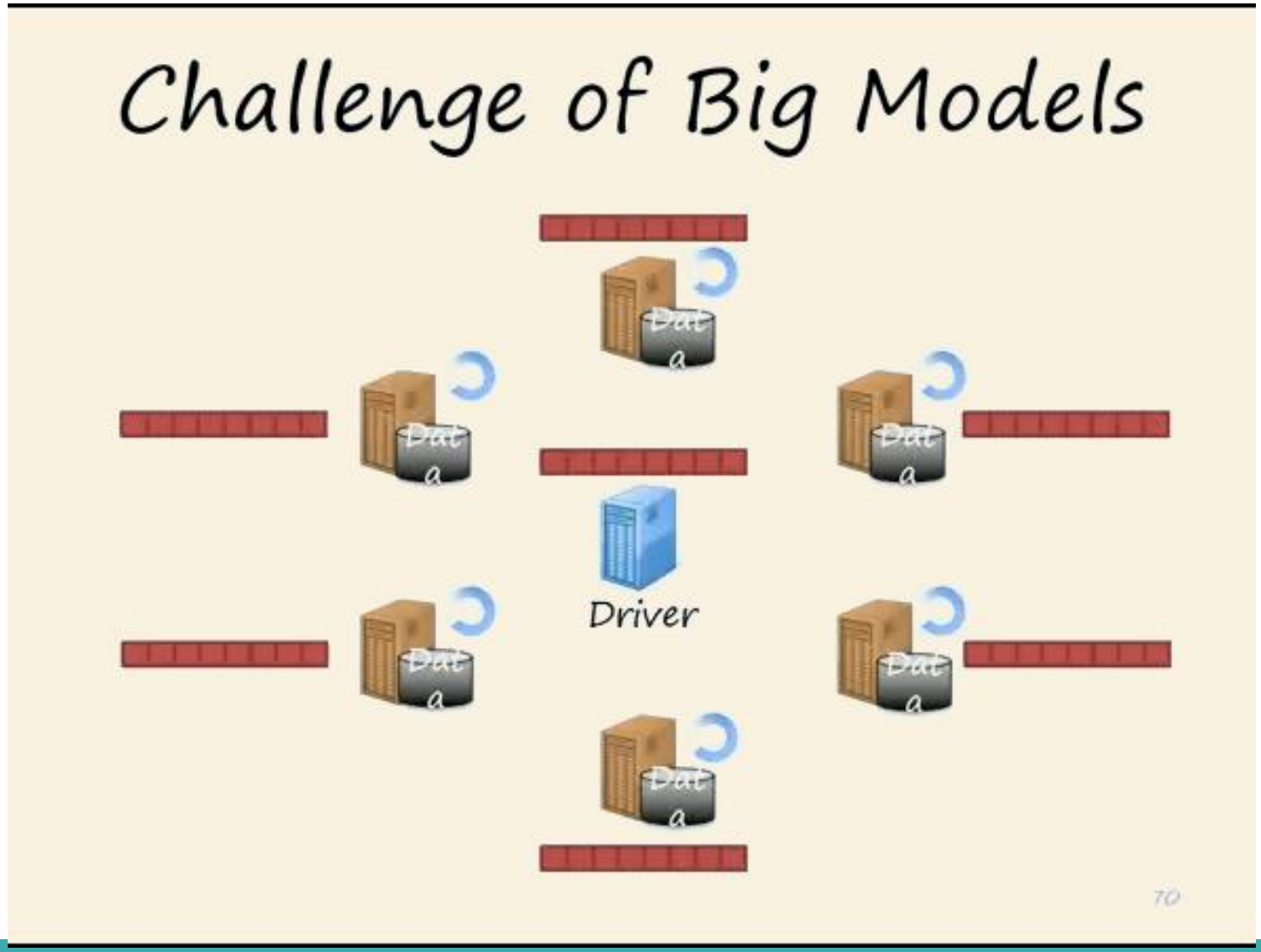
参数服务器的设计目标

- 高效的通信：异步通信不会拖慢计算
- 弹性一致：将模型一致这个条件放宽松，允许在算法收敛速度和系统性能之间做平衡
- 扩展性强：增加节点无需重启网络
- 错误容忍：机器错误恢复时间短，Vector Clock容许网络错误
- 易用性： 全局共享的参数使用向量和矩阵表示，而这些又可以用高性能多线程库进行优化

Parameter-Server设计架构

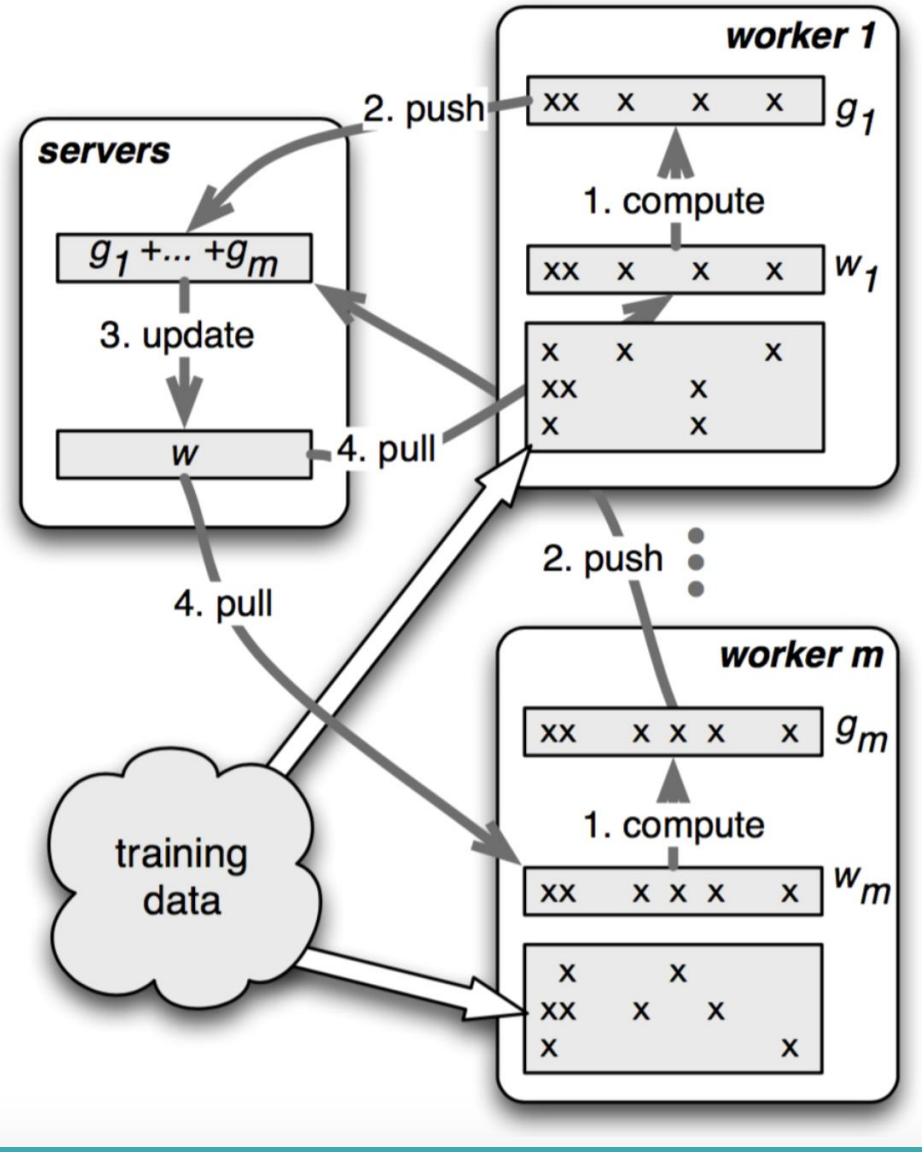


参数服务器的通信过程



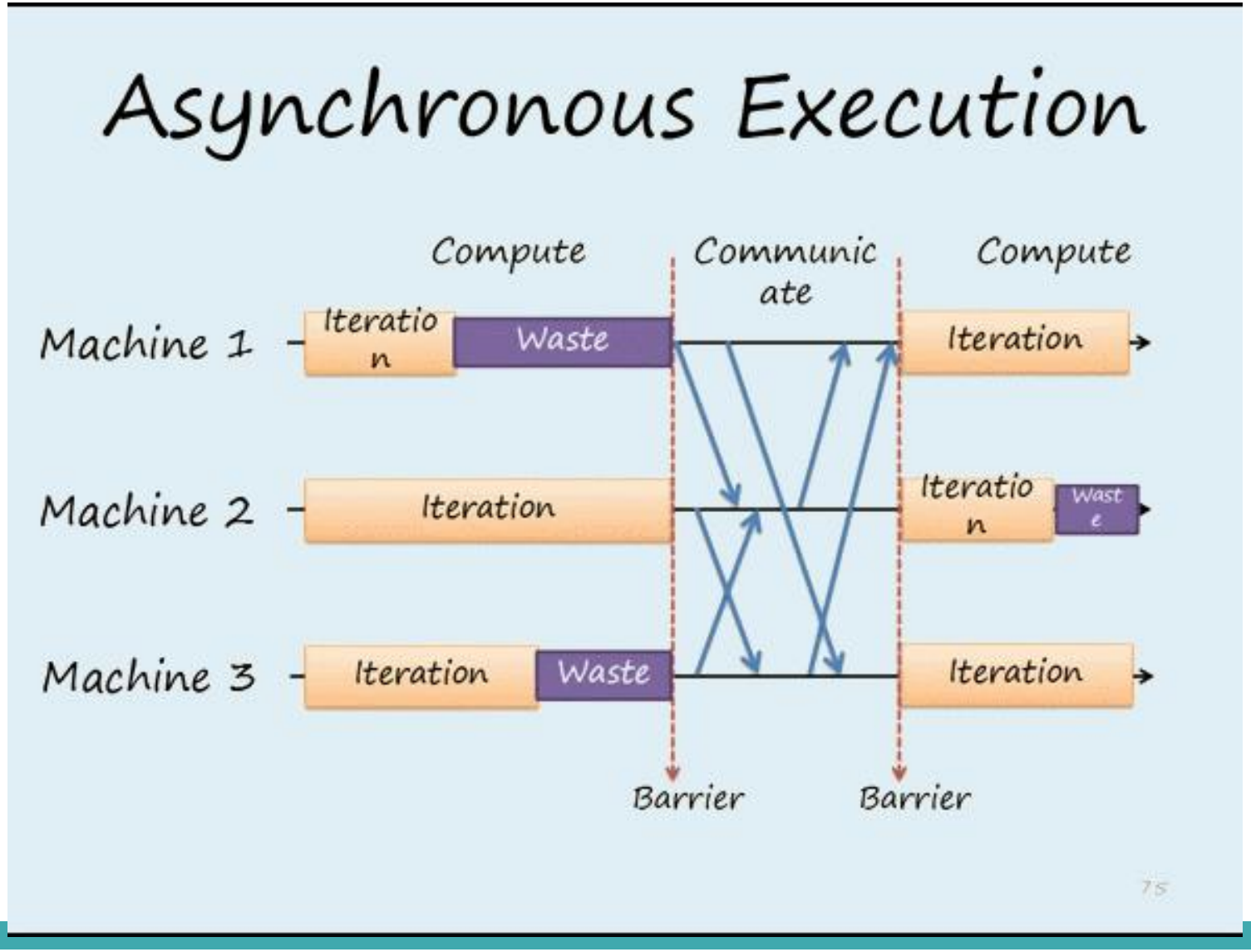
key-value vectors

- parameter server 中，参数都是可以被表示成(key, value)的集合
- workers 跟 servers 之间通过 push 跟 pull 来通信，通过Range来提高性能



异步计算提高效率——同步迭代计算模型

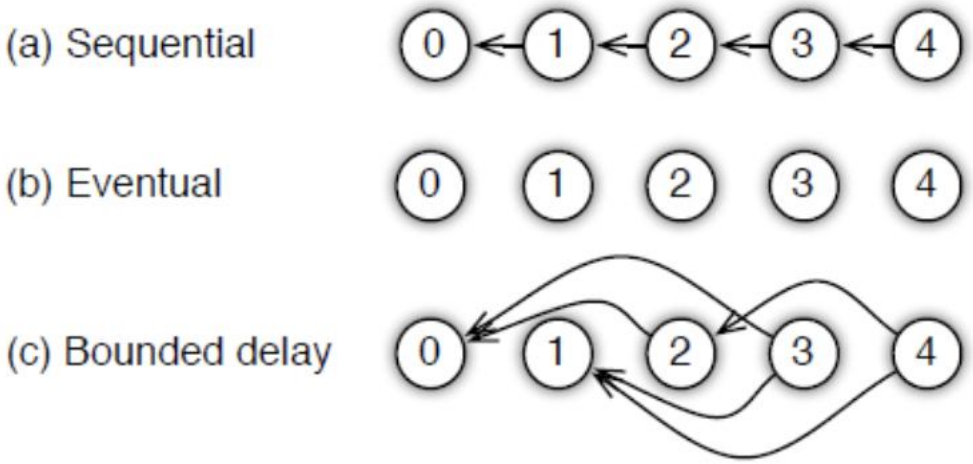
异步计算提高效率——异步迭代计算模型



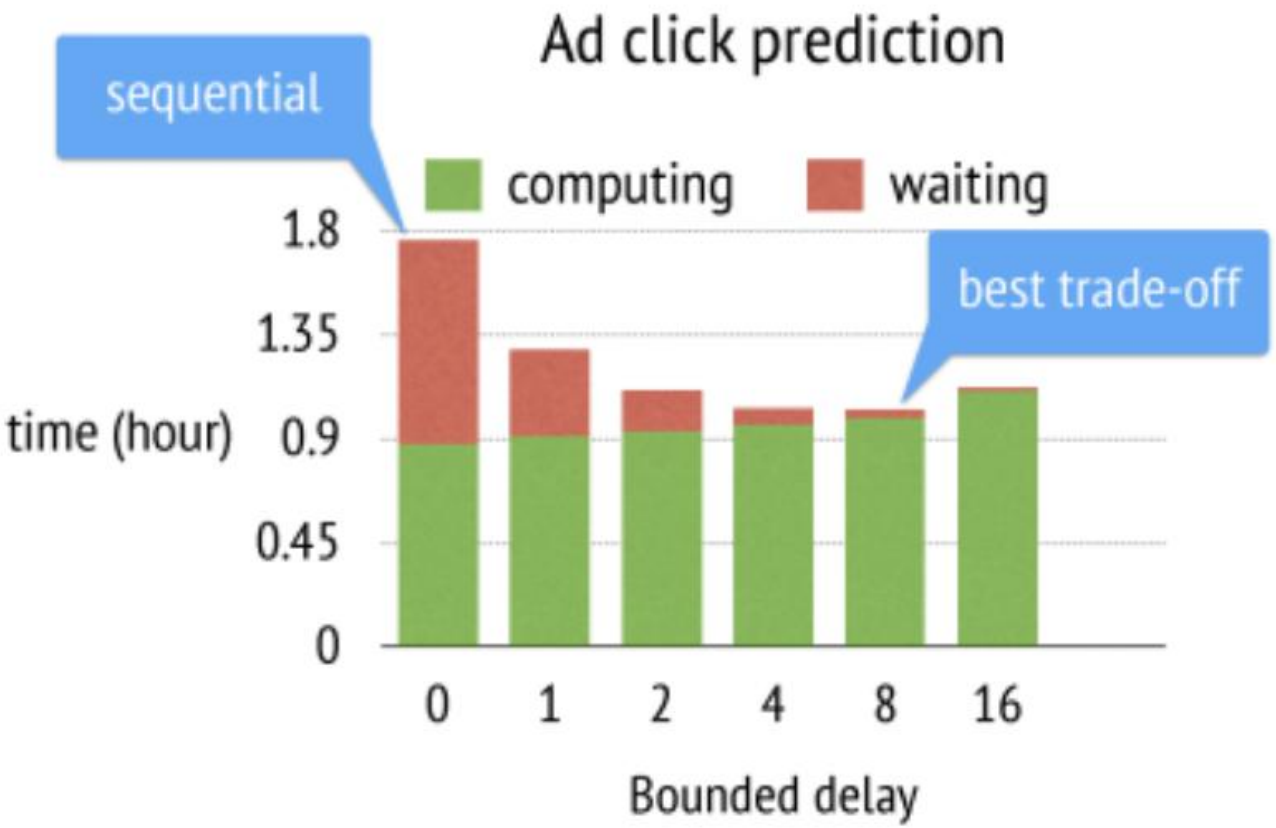
模型一致性设计

- parameter server 异步通信减少了通信时间、等待时间，牺牲了模型一致性，降低了算法的收敛速度

- 一致性选择
  - Sequential： 这里其实是 synchronous task，任务之间是有顺序的，只有上一个任务完成，才能开始下一个任务；
  - Eventual： 跟 sequential 相反，所有任务之间没有顺序，各自独立完成自己的任务
  - Bounded Delay： 这是sequential 跟 eventual 之间的trade-off，可以设置一个 τ τ 作为最大的延时时间。也就是说，只有 $>\tau>\tau$ 之前的任务都被完成了，才能开始一个新的任务；极端的情况：
    $\tau=0$ $\tau=0$，情况就是 Sequential；
    $\tau=\infty$ $\tau=\infty$，情况就是 Eventual

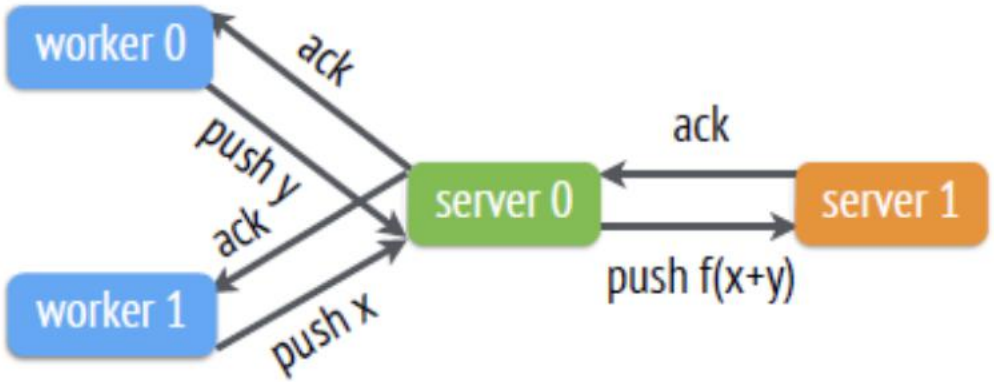


算法性能与算法收敛效率的 Trade-off



其他功能设计

- Vector Clock，基于Range-based记录参数更新时间戳，每次更新最多分裂3份
- Message（KV对）传输：K首次传输后cache，Value用 Snappy压缩
- 采用一致性Hash保证参数存储一致，Hash Ring基于server manager，采用链式存储
- 动态的节点增加与删除



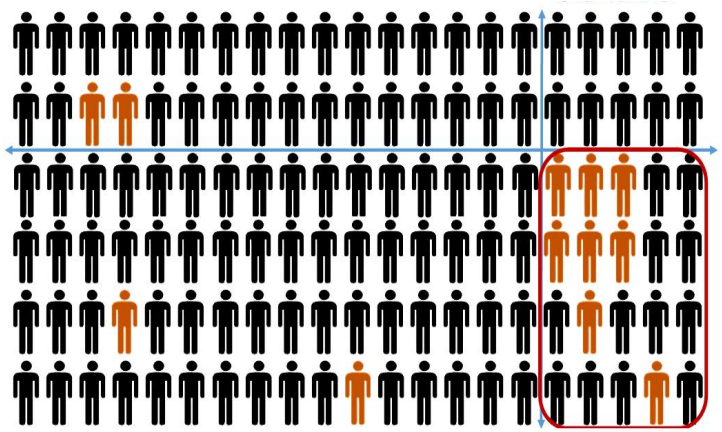
CONTENTS

目录

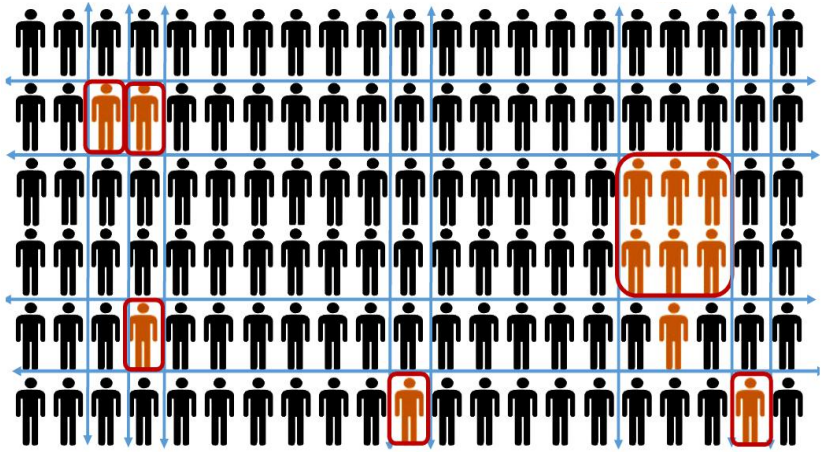
传统方式 VS 大规模机器学习方法，微观业务场景的分析和预测能力对比

传统客户触达用少量特征将用户较为粗放的划分到少量类别中，每个类别中的用户被认为有相似的属性和相同的意愿，丢失了对每个用户的个性化描绘，准确性有限。同时也无法覆盖到部分客群中的个性化用户。

传统客户触达



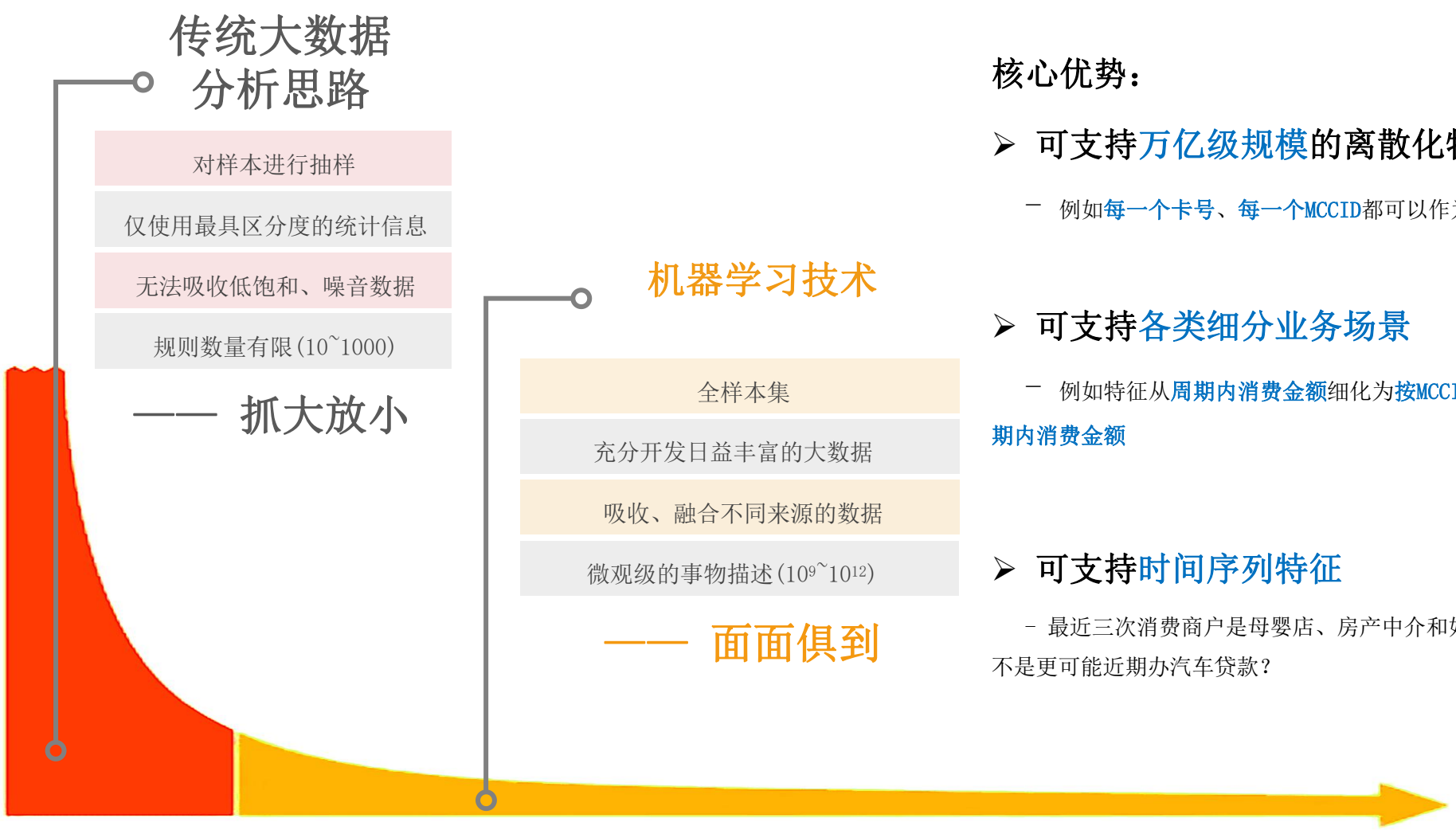
人工智能客户触达



大数据机器学习模型，基于日益丰富的海量数据样本，和千万以上量级数据特征，将用户细分到微观粒度，对每个用户做精细的个性化描述，直接定位到每个有意愿的用户，更精准，更全面。

有意愿的用户　　无意愿的用户

大规模机器学习技术：面面俱到



核心优势：

- **可支持万亿级规模的离散化特征量**
  - 例如每一个卡号、每一个MCCID都可以作为特征
- **可支持各类细分业务场景**
  - 例如特征从周期内消费金额细化为按MCCID及各属性统计周期内消费金额
- **可支持时间序列特征**
  - 最近三次消费商户是母婴店、房产中介和妇产科医院的人是不是更可能近期办汽车贷款？

信用卡交易分期营销—— 提升效果

信用卡中心，对每天数十万的信用卡交易，实时向有需求的客户发送交易分期营销短信，客户回复短信即刻办理

优化目标：提升短信营销响应率和手续费收入

效果：手续费+61%

响应率+68%

注：同等短信发送比例下

数据驱动的过程

1 目标拆解　　交易分期收入 = 短信发送量 * 短信响应率 * 分期费率

2 优化目标　　交易分期收入 = 短信发送量 * 短信响应率 * 分期费率

3 数据建模　　SMS 营销历史 → 短信响应预估模型

4 模型应用

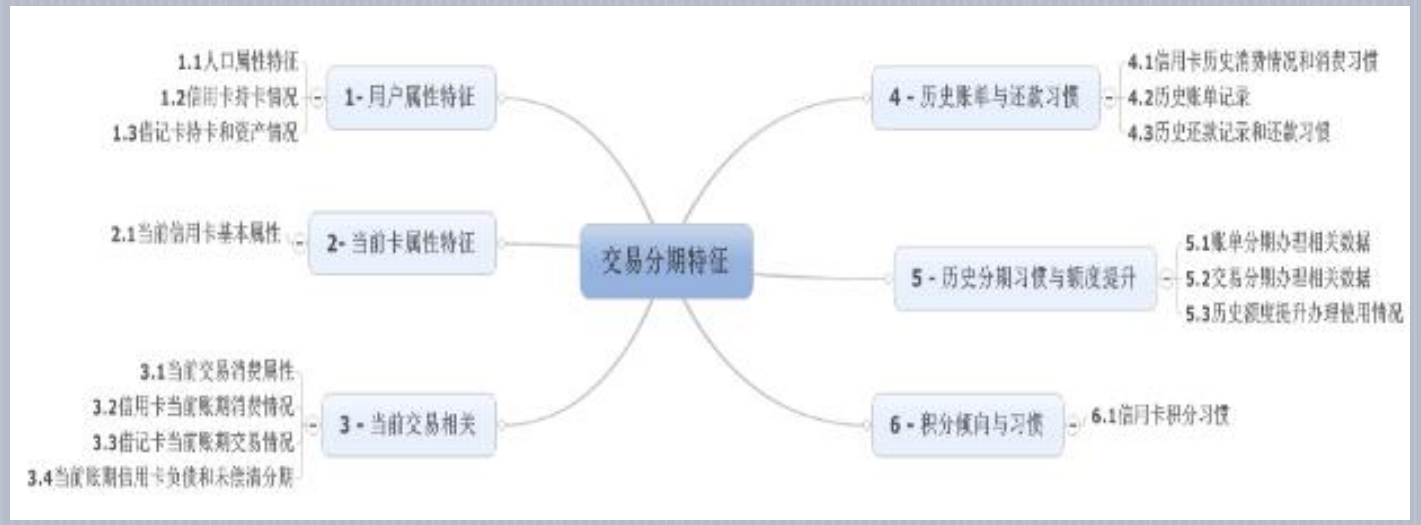
- 优化接受率：“预估响应率” > 阈值，则发送短信
- 优化收入：“预估响应率” * 分期费率 > 阈值，则发送短信

信用卡交易分期营销—— 建模过程

数据
（40,000,000+，2TB）

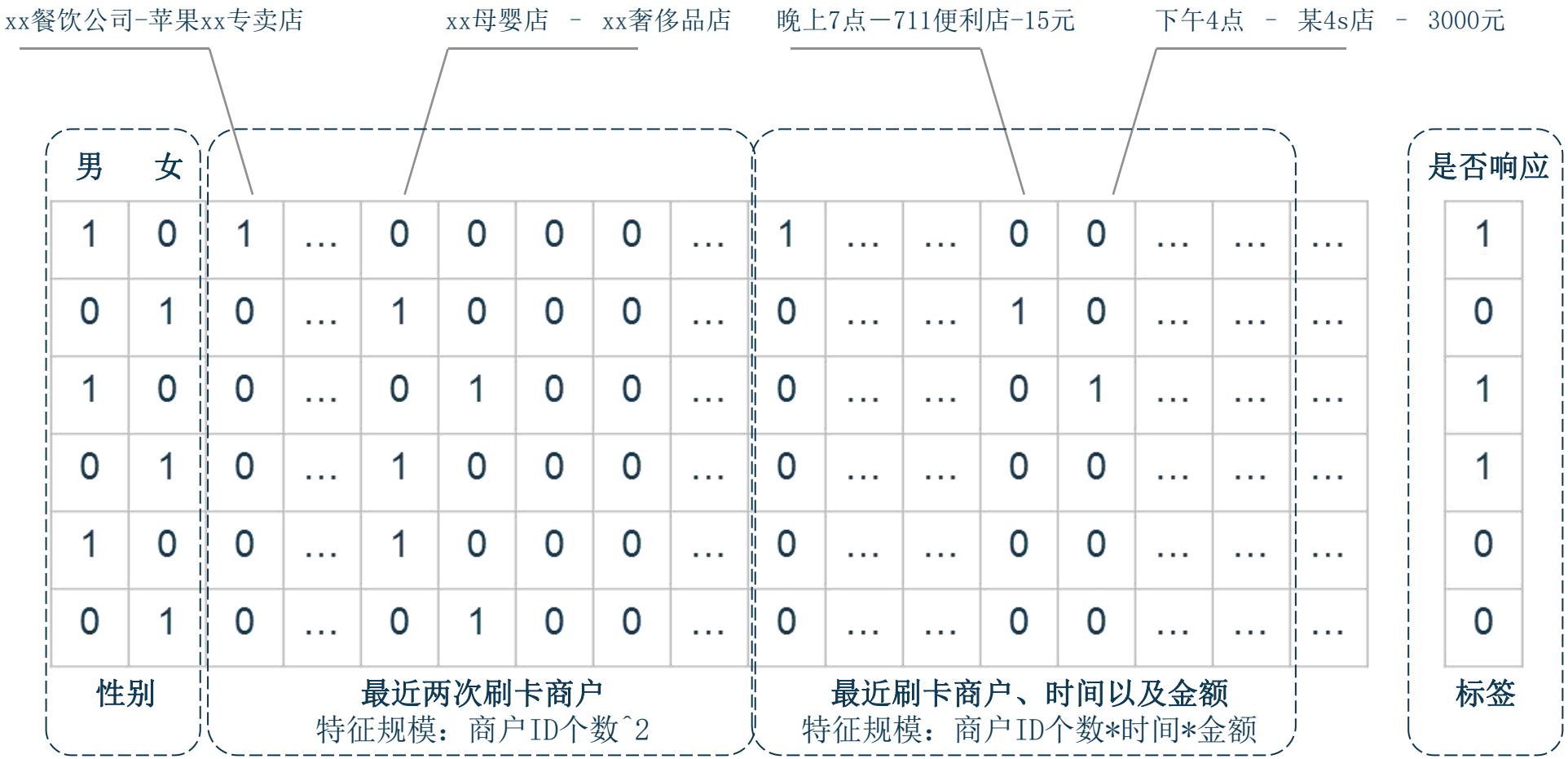
- SMS营销记录
- 账户信息
- 持卡人信息
- 信用卡信息
- 交易信息

维度
（50,000,000+）

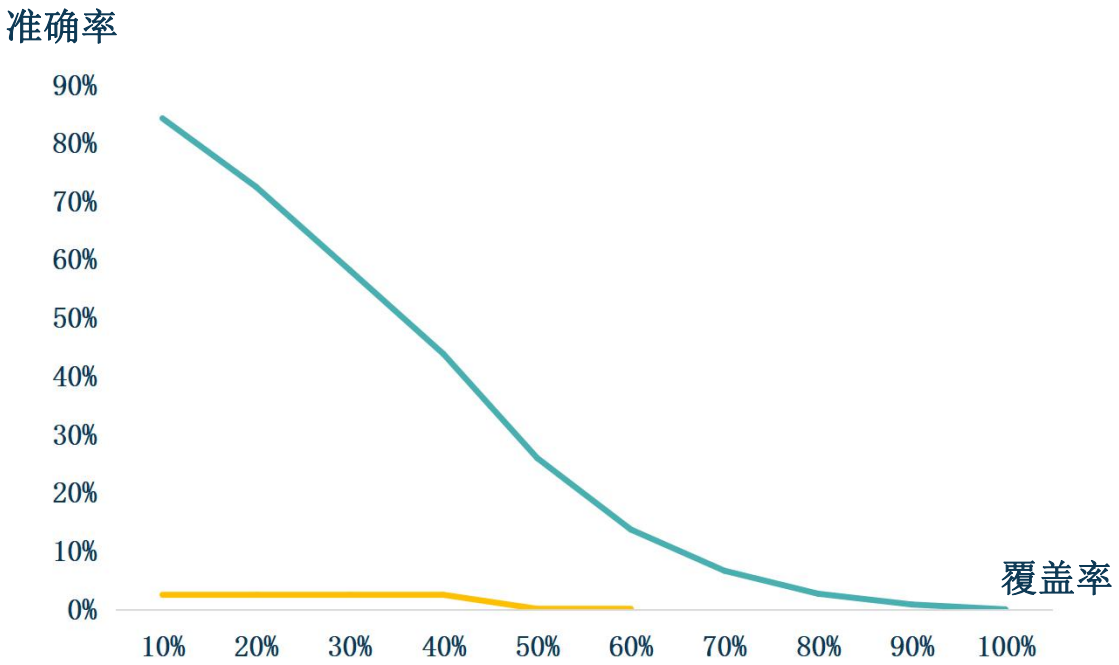


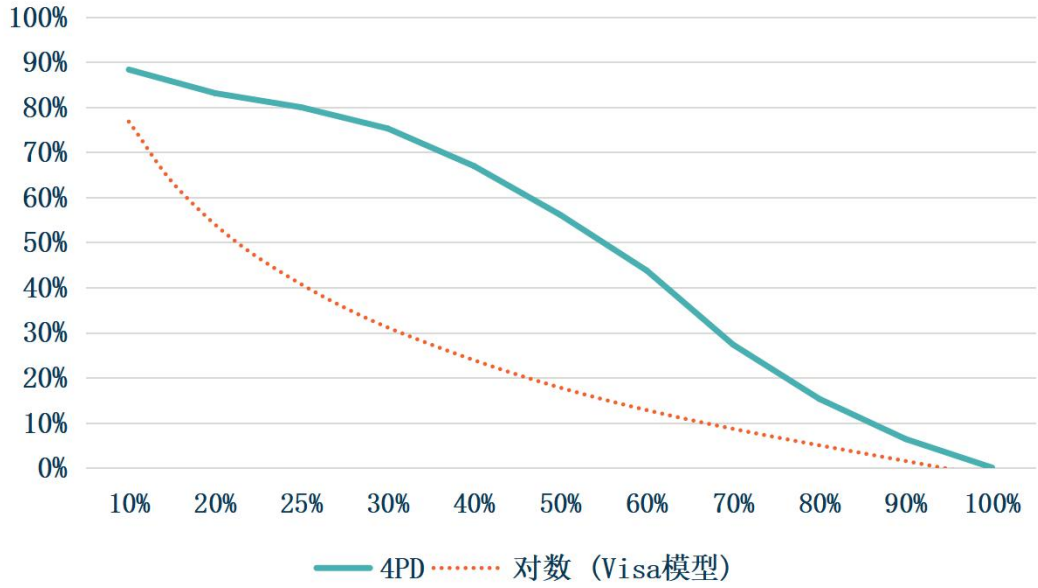
千万级维度特征设计

4000万行样本 X 5000万列特征稀疏矩阵



案例：某股份制商业银行信用卡线下交易反欺诈项目





效率提升：为现有专家规则提升40~260倍

- **覆盖率：**
  - 机器学习模型在超过**2.68%的准确率**下覆盖**80%**欺诈交易
  - 专家规则在低于1%的准确率的情况下覆盖**60%**欺诈交易
- **准确率：**
  - **Top 10%**：机器学习为现有专家规则的40倍
  - **Top 50%**：机器学习为现有专家规则的260倍

Visa渠道：与VAA模型（2016.11）

- 覆盖25%的交易，准确率提升**接近90%**
- 覆盖90%的交易，准确率**提升5倍**

亮点一、完整的特征体系 - 以欺诈防控场景为例

完整的特征体系

实现

充分 涵盖业务要素，完整 刻画行为偏好

为了精准刻画欺诈场景的特征，我们充分了解业务流程，结合业务经验，**识别出影响/判断目标（交易是否欺诈）的主要因素**，并形成特征体系的主要构成。

◘ 原始数据充分涵盖业务要素

我们尽可能地充分获取与欺诈场景相关的原始数据，**以期望获得完整的欺诈交易场景全貌：**

- 交易数据
- 客户数据
- 客户关系数据
- 账户信息
- 卡片信息
- 协议信息
- 黑交易记录
- 特约商户
- IP地点字典

◘ 历史 / 长期 / 短期特征结合完整刻画行为偏好

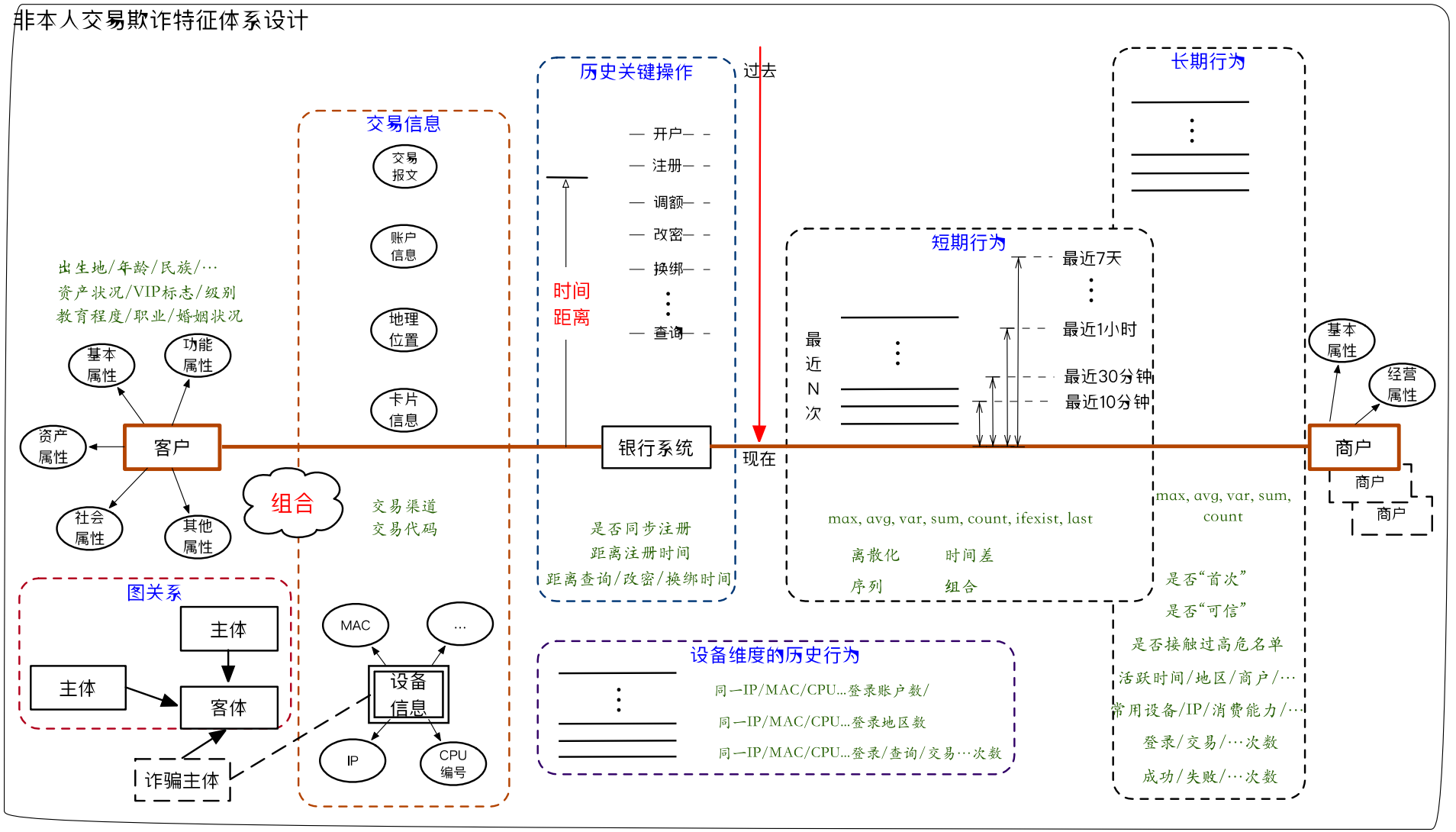
基于原始特征数据对交易场景中涉及的**持卡人、商户的时序特征进行了关联：**

- 短期行为方面，尝试更多种时间窗口，学习海量的短期行为序列；
- 长期行为方面，分析是否“首次”，“可信”等；
- 关键历史操作方面，分析开户、注册、调额、绑定等关键操作的时间距离。

亮点一、完整的特征体系 - 非本人交易欺诈特征体系设计

特征体系构成：

- *客户维度*
- *商户维度*
- *交易信息维度*
- *历史关键操作维度*
- *短期行为*
- *长期行为*
- *欺诈人维度*
- *图关系特征*



亮点二、高维离散特征 - 以欺诈防控场景为例

高维离散特征 实现 **精细 微观** 的特征工程

使用的原始特征为50个字段，通过时序、组合操作生成了约500小类特征，对应 **25亿** 维特征空间，模型保留有效特征约 **620万**。

特征空间的超高维度

我们尽可能地把所有相关的因素都设计为特征，并对**重点业务字段进行时序拼接和特征组合处理**，后续交给机器去学习。

特征离散化处理

我们对所有特征采取离散化处理，将连续值特征转化、切分为类别特征。特征离散化可以**精细地分析影响优化目标的因素**，例如对每一个年龄，或者每一个交易额都进行独立地刻画（即，年龄和金额的每一个取值都视为一个单独的特征）。

亮点二、高维离散特征 - 超高维特征空间示例

信用卡线下交易反欺诈场景为例

传统的低维模型 / 专家规则（维度<10000）	
同一张卡短时间内在两个城市交易	不可能事件
同一张卡短时间内在多个商户交易	不可能事件
同一张卡发生余额不足交易次数异常	试错
同一张卡发生密码错误次数异常	试错
同一张卡发送多笔错误交易且交易面额递减	试错
先小额试刷（小于50元）后，三十分钟内大额交易	试刷
在高危时间段交易	高危时间
商户一个小时内连续发生多张连续卡号的交易	信息泄漏
首次消费金额接近信用卡额度	首次行为
首次取现密码错误	首次行为

从低维到超高维特征工程方法

离散化

二阶特征组合

高阶特征组合

超高维特征结果

一阶特征		
持卡人属性（例如年龄）	细分成“一岁一档”	每一个年纪都有不同的权重分数
交易信息（例如金额）	可以按照交易金额分成非常细的金额段	每个金额段都有不同的权重分数
商户信息（例如MCC）	每个MCC都是一个特征	每一个MCC都有不同的权重分数

二阶特征		
持卡人属性组合商户属性	年龄28岁-某奢侈品店	概率偏低
交易金额与持卡人历史消费的关系	某次交易金额为历史中位数的1000倍，且金额大于10000	可疑，需进一步组合其他特征
交易地点与客户账单地点的距离	交易地点距离账单地点5000公里	可疑，需进一步组合其他特征

高阶特征（接近25亿特征）		
最近5次交易的金额与返回码的组合序列	例如：最近五次的交易金额和返回码分别为（由近及远，下同）： 4998.0，7800.0，8.9，298.0,105.8, 成功，额度不足，成功，成功，成功 对应的特征输出为（每行为一组特征，下同）： 4998.0，成功-7800.0，额度不足-8.9，成功 - 298.0，成功 - 105.8，成功 4998.0，成功-7800.0，额度不足-8.9，成功 - 298.0，成功 4998.0，成功-7800.0，额度不足-8.9，成功 4998.0，成功-7800.0，额度不足	前两次（按时间先后）为正常交易，第三次为盗卡的小额试探交易，试探成功，第四次为大额盗刷，超额度，第5次为降低额度盗刷
客户性别-学历-账单地址-商户MCC-交易金额-交易时段的组合	男-本科生-北京昌平-王府井某奢侈品店-18000.0-工作日晚11点	概率小

亮点三、基于各类算法的多模型对比建模 - 以智能营销场景为例

基于各类算法的多模型对比建模

实现

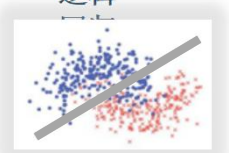
对比 建模，择优 选定算法

我们通过先知平台中集成的各类主流开源算法： LR / GBDT / SVM / FM等，对营销场景进行同步建模，对比分析。

➢ **先知机器学习平台高效支持对比建模：**

- 高性能的机器学习引擎，支持基于各算法同步建模
- 共享数据清洗、预处理阶段成果
- 横向比较分析，选取最优模型

逻辑

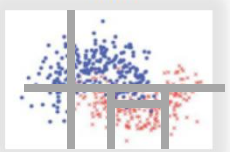


LR的适用场景：

- 离散特征较多
- 对数据和业务充分理解
- 加入足够多的组合特征
- 模型优化重点-特征

采用逻辑回归算法的场景：
融e联、存管通、理财、节节高

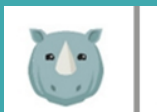
GBDT



GBDT的适用场景：

- 连续特征较多
- 不需要太多的特征调整
- 模型优化重点 - 参数

采用GBDT算法的场景：
基金、融e行、保险



Thanks!

Q & A