

大数据产业和产品链条

南京大学 软件学院 刘嘉

PART-1

大数据是否能改造你的业务

大数据和数据分析数据挖掘的差别

大数据、人工智能和互联网产品

人工智能是大脑

人工
智能



大数
据

数据是所有工作的基础、资产

产品

商业化产品离不开业务场景

深度学习是什么？

基于规则的系统



人工设计程序



传统机器学习



人工设计特征

特征映射结果



深度学习



原始特征

额外的层和抽象特征

特征映射结果



表示学习

深度学习和数据的关系



浅层模型



深度模型

浅层模型与深度模型



旧的优化方法



新的优化方法



优化方法是关键



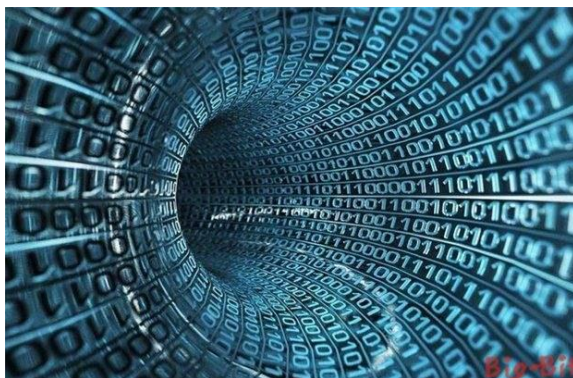
方法

数据

模型

数据的作用

大数据进入行业的三要素



全量加工



行为数据



自动化应用

交易数据

- 业务流程必须要记录的数据
 - 电信：通话记录、话费
 - 银行：存取款、利息
 - 医疗：病历
- 特点
 - 数据规模中等
 - 一致性要求极高

行为数据

- 业务流程中非必须记录的数据
 - 互联网：后台日志
 - 电信：通信内容、上网记录
 - 医疗：日常健康指标
- 特点
 - 数据规模巨大
 - 一致性要求相对较低

采样分析

- 通过小规模数据取得相当准确的结果
 - 用户教育程度分布统计
 - 人口普查
 - 百度迁徙地图
- 特点
 - 实际上不需要大规模计算

全量加工

- 必须分析全量数据才能得到问题的结果
 - 个性化推荐
 - 计算广告
 - 个人征信
- 特点
 - 大规模计算实际上无法避免

洞察应用

- 全局或局部性的统计信息获取
 - 企业财务报表
 - 日常运营报表
- 特点
 - 主要用于宏观决策支持
 - 面向管理者和运营人员

自动化应用

- 个体的行为和兴趣特征捕获
 - 定向广告
 - 客户关系维护
- 特点
 - 主要用户微观业务实施
 - 面向机器和市场销售人员

广告行业




行为
数据

用户行为日志、包括搜索、
交易、网页浏览、分享等等。



全量
加工

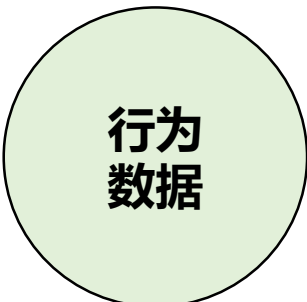
所有用户都需要分析用户画
像，无法通过采样计算。



自动化
应用

根据用户行为得到用户画
像，
再将用户画像自动匹配广
告。

保险行业（展望）



行为
数据


“上一年未出险人群”

“癌症发病率是平均三倍”



全量
加工

出险率预估+个性化定价



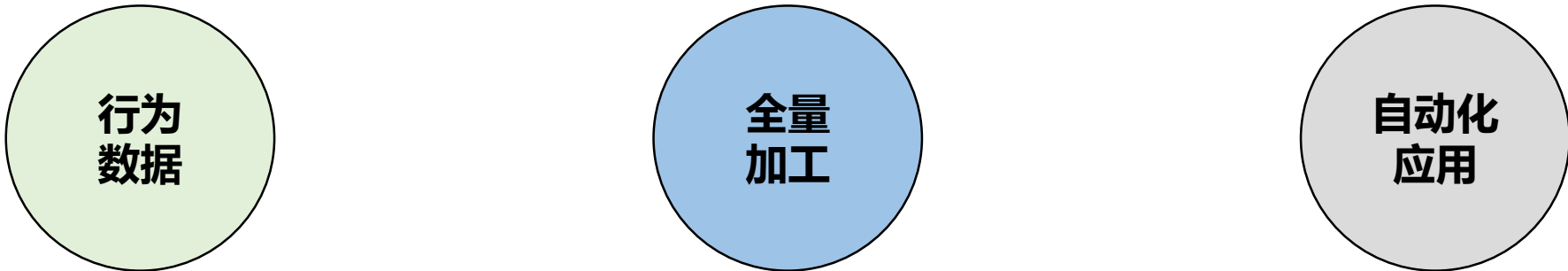
自动化
应用

非理财险的利润率来源于信息不对称

利润=保费-平均赔偿***出险率**

通过对出险率的准确预测，可以极大地提升保险产品利润

医疗行业（展望）



行为
数据

可穿戴设备和云存储的普及，
个人健康数据规模将爆发增长

全量
加工

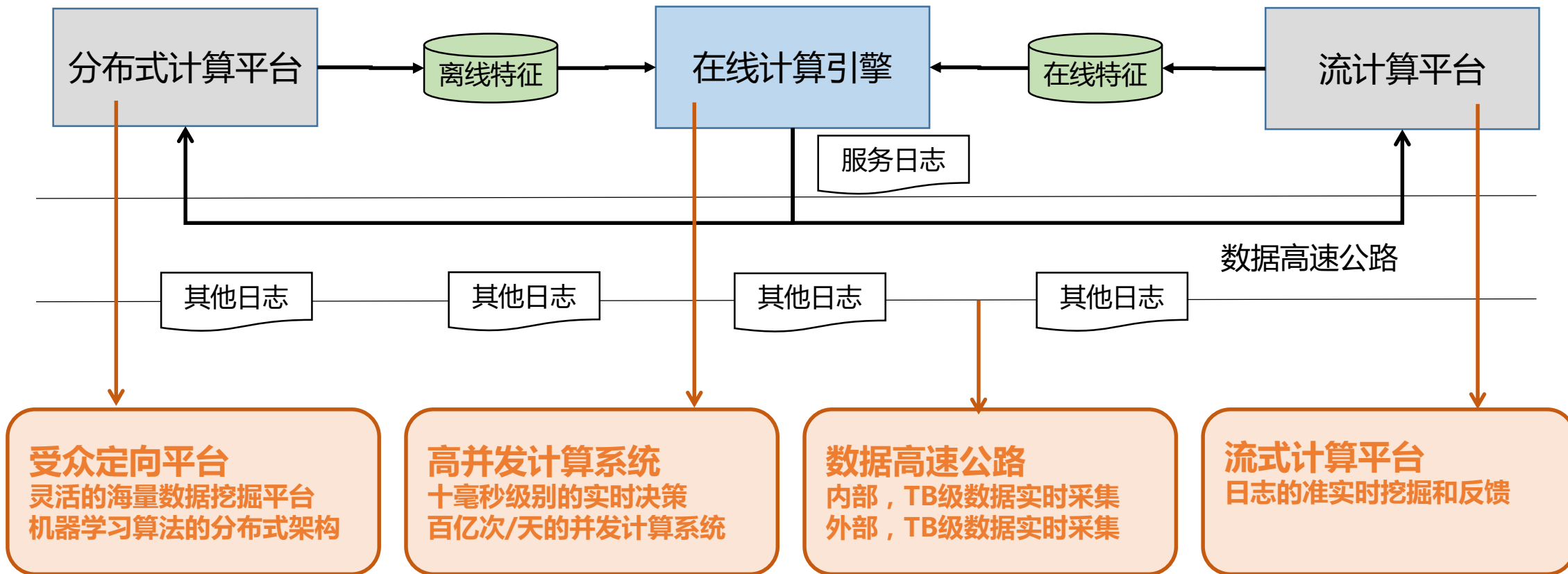
个人健康建模+疾病的管理预防

自动化
应用

基于个人健康数据，实现个性化医疗和点对点医疗新模式

洞察应用：海量健康数据将催生数据驱动的医学研究方法

大数据自动化系统一般框架



开源软件的使用

- 优势
 - 大量细分使用场景都有开源方案
 - 大型互联网公司的开源产品经过了充分的测试
- 顾虑
 - 需要专业能力去鉴别好的和不太好的开源项目
 - 在遇到深层次的bug时无能为力
- 核心业务逻辑不应该选择开源软件

数据作为资产

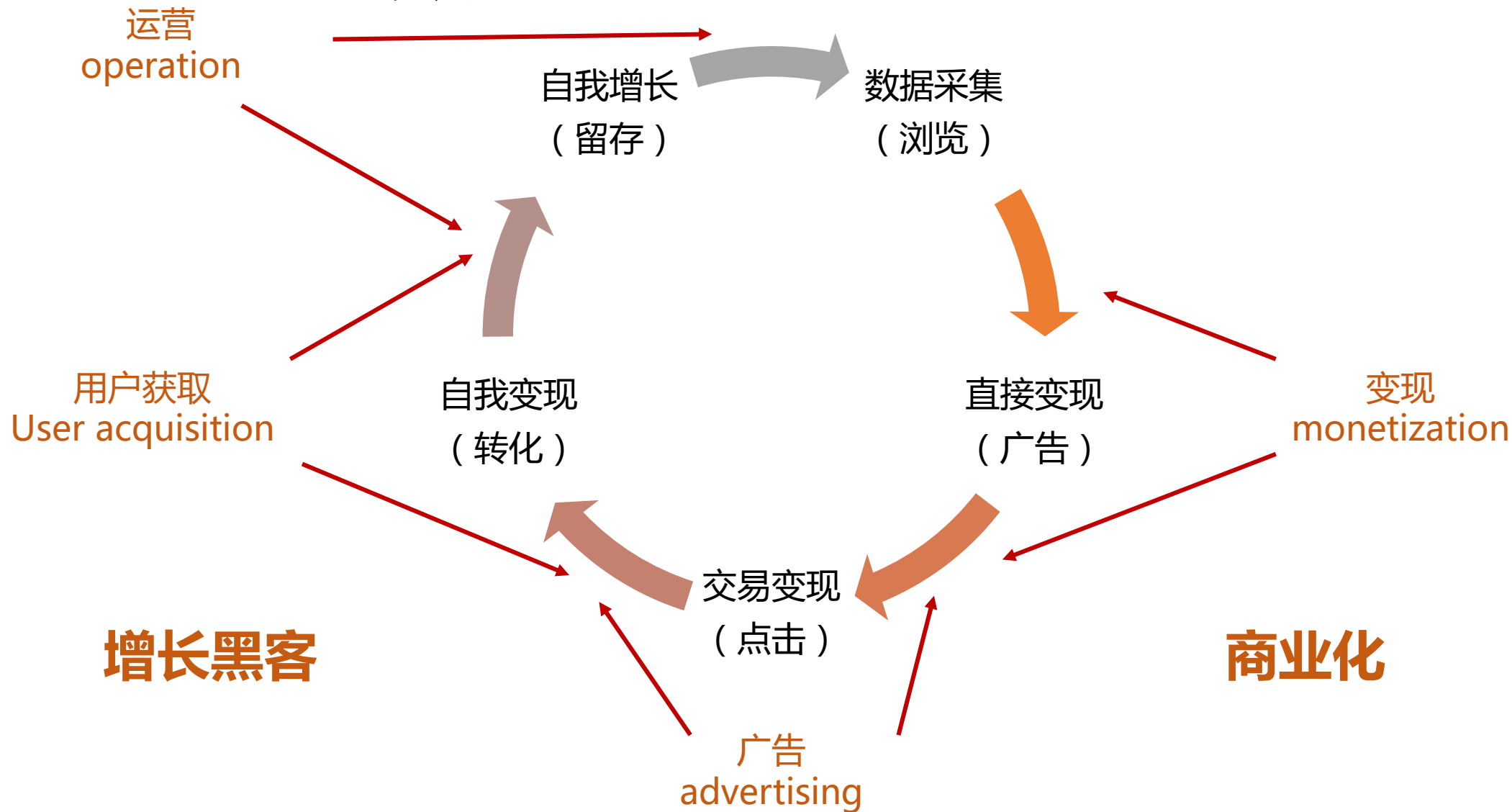
- 如何让数据帮你赚钱
- 如何用数据洞察你的用户
- 如何让数据进行交易

数据用于交易

- 如何让数据组织你的运维

数据用于自我提升

数据商业化体系闭环



PART-2

数据作为资产的变现

免费模式

- 免费模式的本质
 - 能够个性化传播信息的产品，售价都会趋向其边际成本
- 免费模式的举例
 - 网站、App应用：边际成本 ≈ 0
 - 手机、智能电视：边际成本 \approx 硬件成本
- 免费模式的目的
 - 获得其他资产，通过后向渠道变现

广告是最直接的数据赚钱模式

公司	Alphabet	Facebook	腾讯	阿里巴巴	百度
总收入 (亿美元)	817.62	179.28	158.41	122.93	102.23
广告营销 (亿美元)	732.23	170.83	26.90+87.14	77.04	100.78
广告占比	89.6%	95.3%	17.0+55.0%	62.7%	98.6%

线上可变现资产



流量变现到数据变现



10000



6000

流量变现



6000

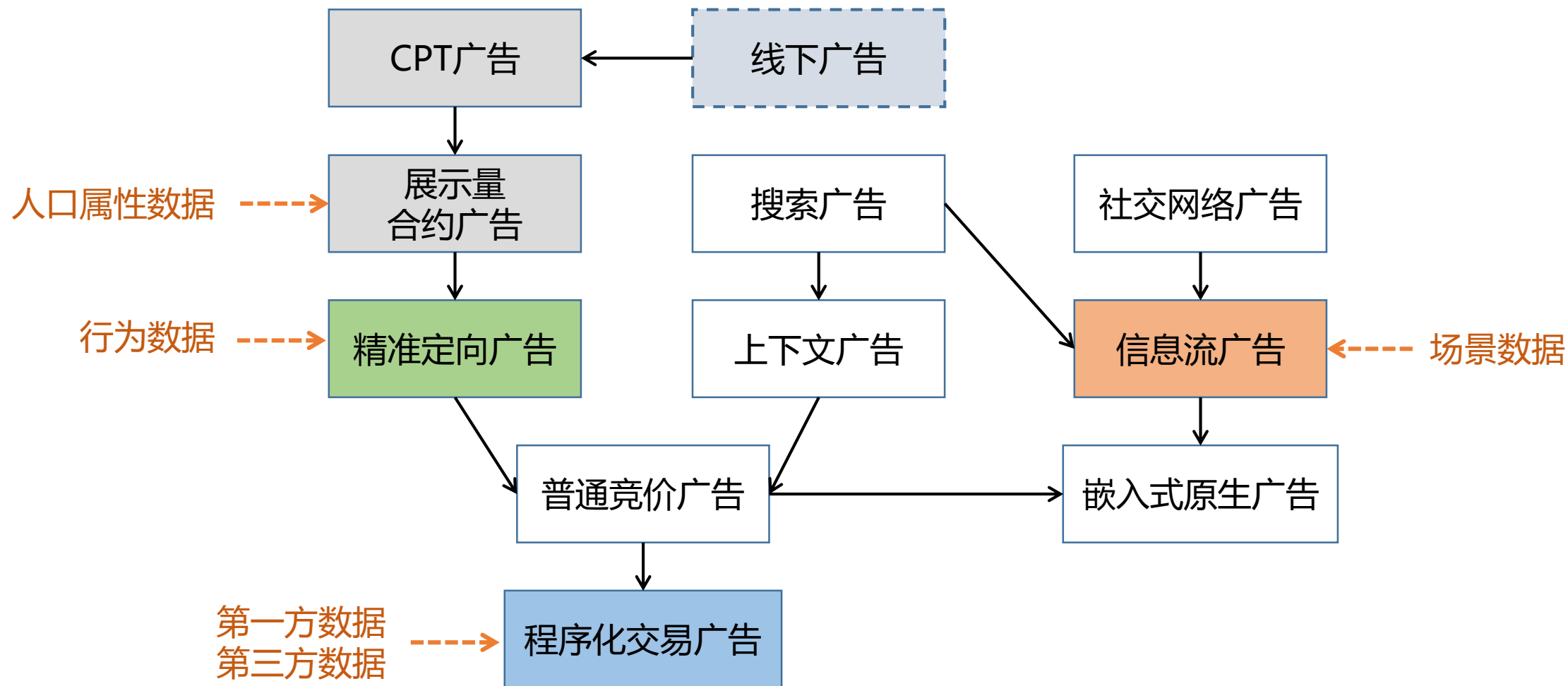
数据变现

2000

数据价值

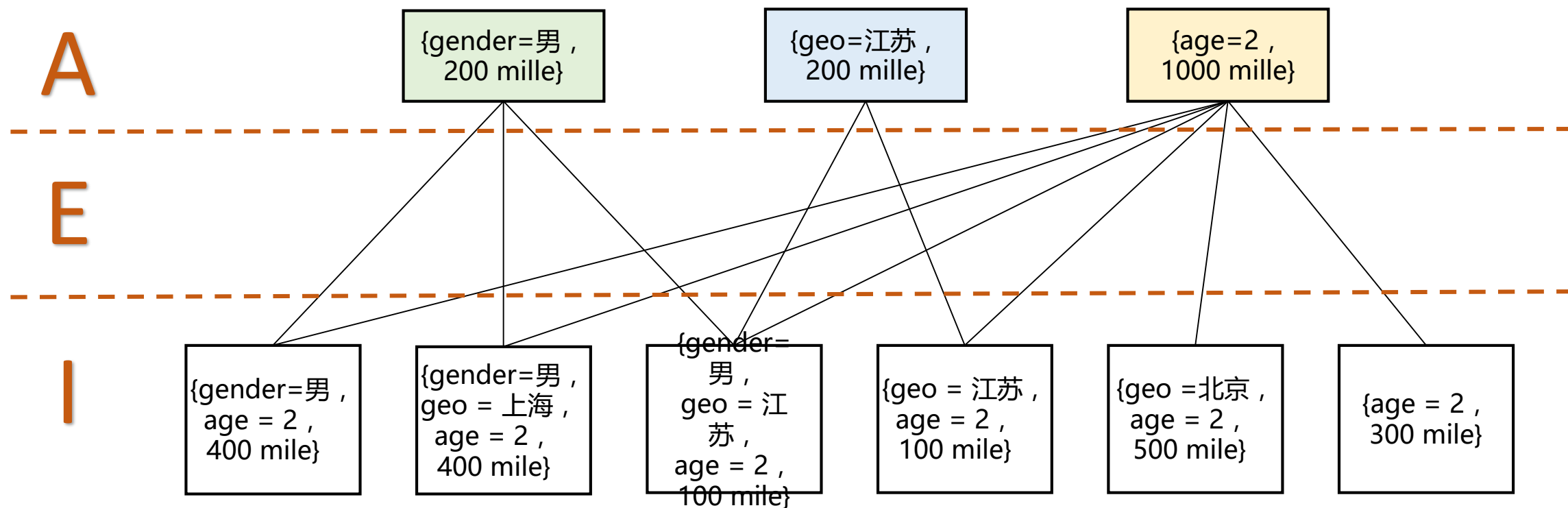
12000

数据驱动的商业产品发展历程



合约分配的交易模式

需求节点 (Demand Nodes , 订单要求的定向标签组合)



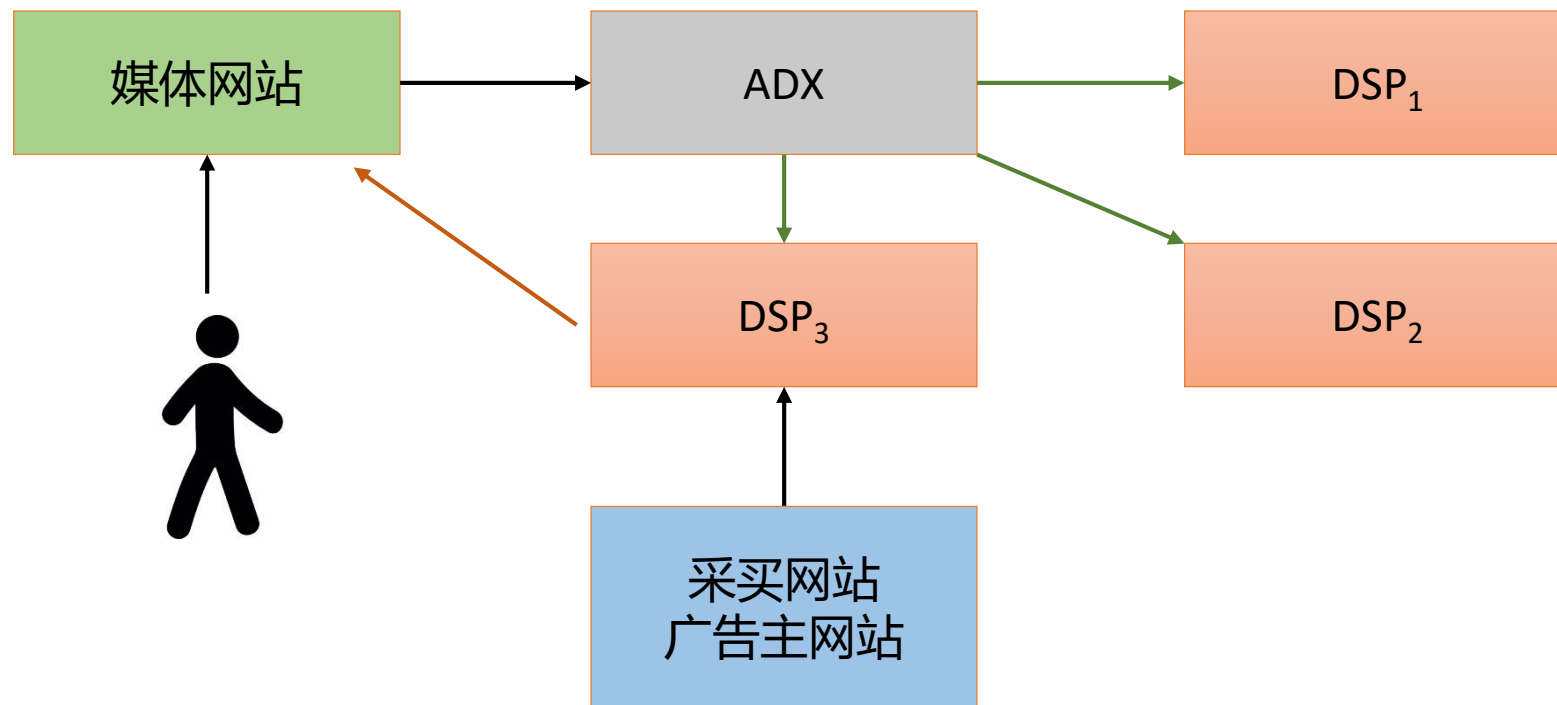
供给节点 (Supply Nodes , 定向标签的最细组合)

假设：节点内部的流量差异可以忽略

竞价的交易模式

- 将对象 $a=\{1,2,...A\}$ 排放到位置 $s=\{1,2,...S\}$
- 对象 a 的出价bid为 b_a ，而其对位置 s 的计价为 $r_{as}=u_s v_a$ ，
($u_1 > u_2 > u_3 \dots > u_s$)
- V_a 为点击价值， u_s 为点击率
- 对称纳什均衡 (Symmetric Nash Equilibrium)
 - ($v_s - p_s$) $x_s \geq (v_s - p_s) x_t$, 其中 $p_t = b_s + 1$
 - 寻找收入最大化且稳定的纳什均衡状态是竞价系统设计的关键

程序化交易模式



重定向的数据变现模式

- 网站重定向 (Site retargeting)
 - 根据用户在广告主网站上的行为进行重定向
- 搜索重定向 (Search Retargeting)
 - 根据用户与广告主相关的搜索行为进行重定向
- 个性化重定向 (Personalized Retargeting)
 - 根据用户再广告主网站上关注的具体产品和购买阶段，推送商品粒度的广告，
可以视为一个站外推荐引擎

合作的数据变现—Look-alike

- 问题：
 - 对于中小电商，仅对老用户定向营销远远不够
 - 对于某些类型的广告商，大多数用户无法通过重定向渠道捕捉，例如银行
- 新客推荐
 - 由广告商提供一部分种子用户，DSP通过网络行为的相似性为其找到潜在用户
 - 是一种广告商自定义标签，可视为扩展重定向
 - 在同样reach水平下，效果应好于通用标签

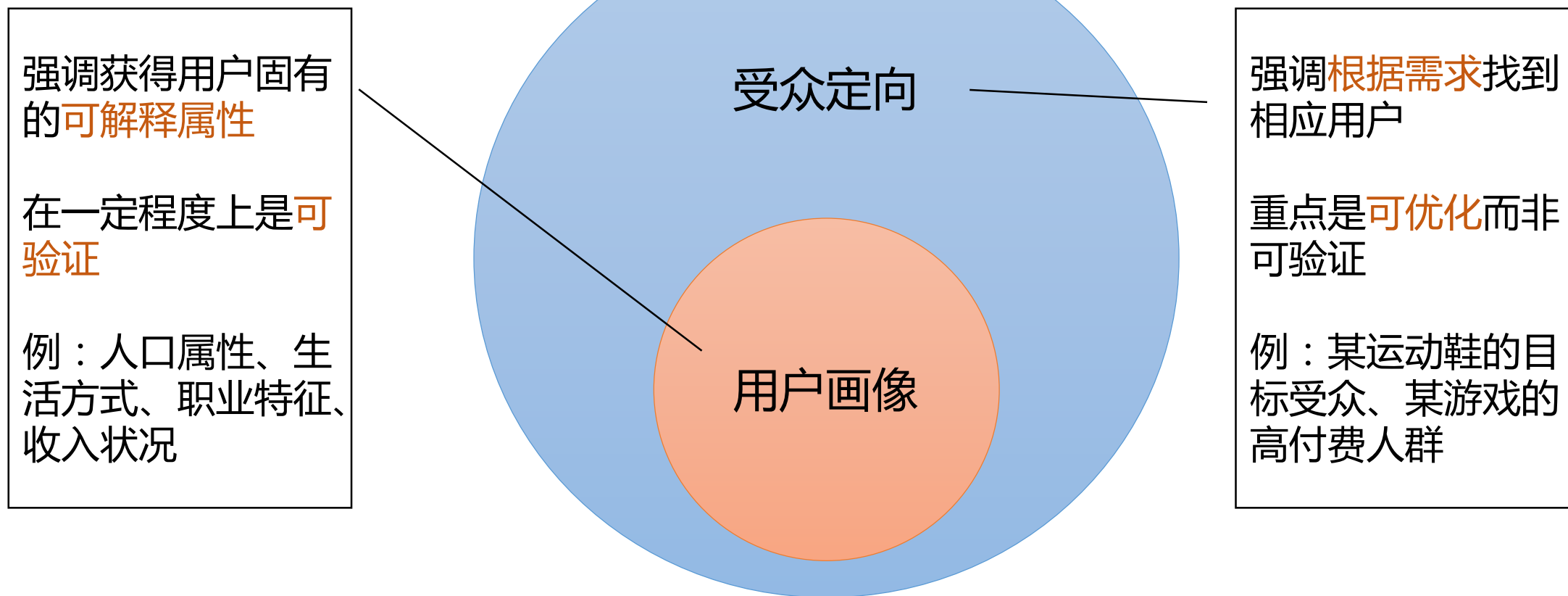
数据变现的逻辑

- 数据变现，数据交易都依附于流量
- 数据变现
 - 人口属性这样粗粒度的分类：合约
 - 高价值的小数据，长尾效益：竞价
 - 广告主拿自己的数据变现：程序化交易
 - 移动的模式：场景数据
- 数据不仅能变现，还能是一个巨大的市场，2200亿超过传统媒体

PART-3

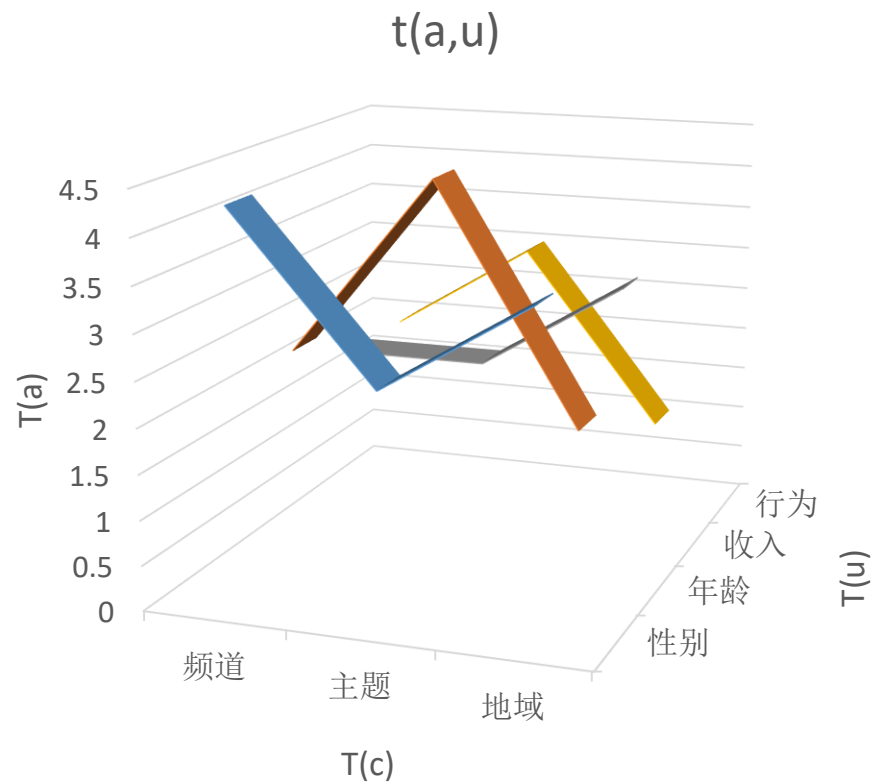
数据作为资产的预先处理

受众定向与用户画像

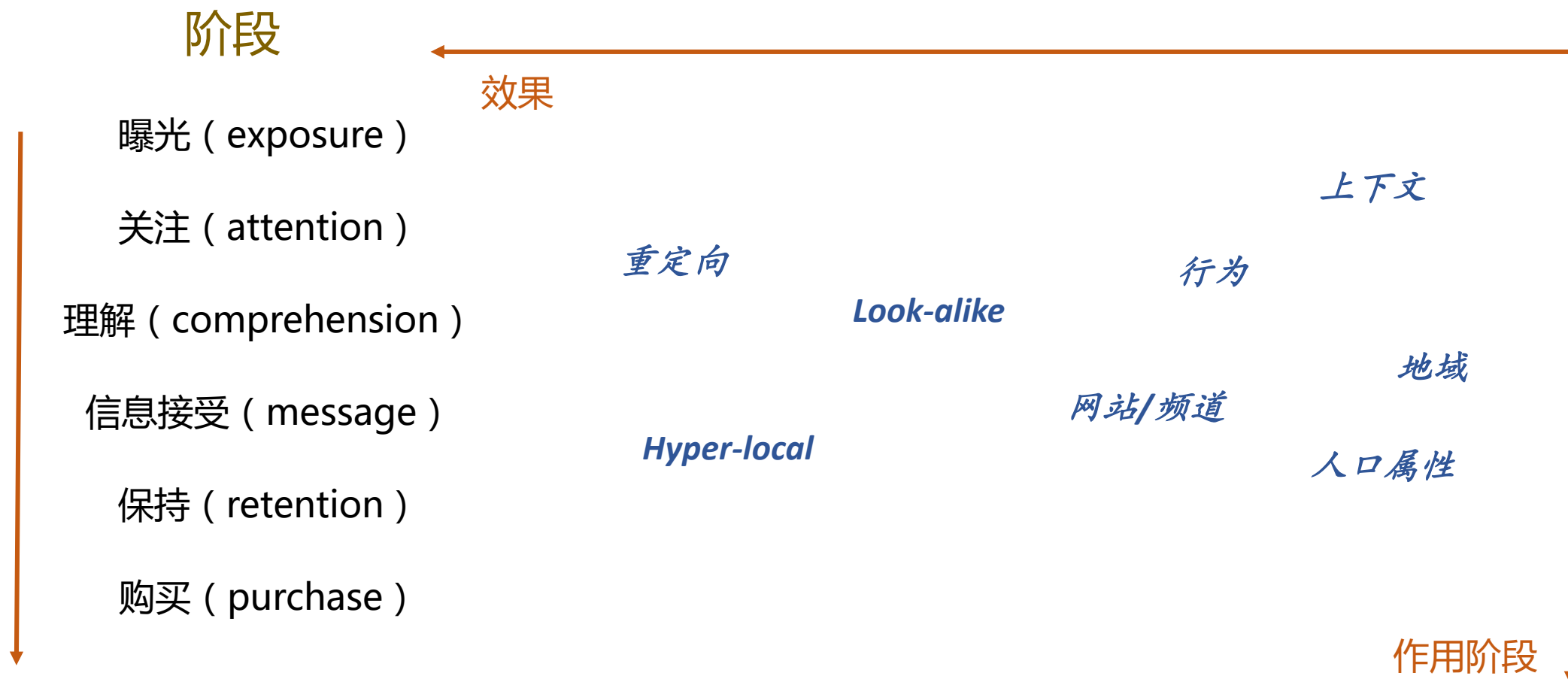


受众定向方法分类

- 受众定向即为 (a, u, c) 打标签的过程
 - 上下文标签可以认为是即时受众标签
- 标签的两大主要作用
 - 建立面向直接变现 (广告主) 的流量售卖体系
 - 为数据业务估计模块 (如CTR预测) 提供特征



常见受众定向方式



受众定向标签体系

- 结构化标签体系
 - 按照某分类法（Taxonomy）制定一个层次标签体系，父节点与子节点在人群覆盖上是包含关系
 - 主要用于面向品牌广告的受众定向（GD系统），固定领域知识下的标签系统
- 非结构化标签体系
 - 根据某类定向需求设置标签，标签并不能为同一个分类体系所描述
 - 适用于多种目标、特别是效果目标并存的精准选择要求
- 关键字
 - 按照搜索或浏览内容的关键词划分人群
 - 非结构化，容易理解，但操作和优化不容易

受众定向标签



一级标签	二级标签
Finance	Bank Accounts, Credit Cards, Investment, Insurance, Loans, Real Estate, ...
Service	Local, Wireless, Gas & Electric, ...
Travel	Europe, Americas, Air, Lodging, Rail, ...
Tech	Hardware, Software, Consumer, Mobile, ...
Entertainment	Games, Movies, Television, Gambling, ...
Autos	Econ/Mid/Luxury, Salon/Coupe/SUV, ...
FMCG	Personal care, ...
Retail	Apparel, Gifts, Home, ...
Other	Health, Parenting, Moving, ...



类别	描述	数据来源	用户规模
Intent	最近输入词表现出某种产品或服务需求的用户	Bluekai Intent	160+MM
B2B	职业上接近某种需求的用户	Bizo	90MM
Past Purchase	根据以往消费习惯判断可能购买某产品的用户	Addthis, Alliant	65+MM
Geo/Demo	地理上或人口属性上接近某标签的用户	Bizo, Datalogix, Expedia	
Interest/LifeStyle	可能喜欢某种商品，或某种生活风格的用户	Forbes, i360, IXI, ...	103+MM
Qualified Demo	多数据源上达成共识验证一致的人口属性	多数据源	90+MM
Estimated Financial	根据对用户财务状况的估计做的分类	V12	

行为定向 (Behavioral Targeting)

- 根据用户历史上网记录和其他数据计算出用户兴趣



张亮田亮讨论佳偶天橙 《结婚吧》穿帮镜头 · 红配绿竟也能搭出时尚气质/图 · 中国10大最美空姐你遇到过吗

坚持1个月 听懂CNN 每天只要 5分钟, 坚持30天, 效果真的不一样! **立即行动**

汽车 降价 二手车 优惠 | 论坛

房产 家居 二手房 旅居 | 租房

汽车

- 奥迪Q3降6.28万/翼虎降1万 豪华B级车
- 8万大空间狂野SUV 标致2008预10-14万
- 新款卡宴曝光动力变化 8万“国产揽胜”
- 大众捷达碰撞后什么样 13款最热门SUV

本田廉价SUV将8万

• 哈弗H8为何敢卖23万 6大豪华品牌2.0T发动机 最便宜车!

• 2014款劳斯莱斯魅影 日系车漂移之王 极速豪华警车曝光

• 2013北京人最爱车型 逛店记:姚家园 豪车北京降70.8万!

• 与马有关的汽车! 奥迪A6L优惠9.7万 大众途锐降14.65万

• 论坛| 大美女开宝马X5 提日系红色奇骏! 车友评高尔夫7

• 热点| 开这些车的男人易出轨 搜狐汽车车商宝等你加入

选品牌 选车型 查询 5万 8万 12万 SUV 查违章 找底价

房产

- 1月11看房团8线报名中 60万百平三居
- 堪比自住商品房3项目盘点 1月开盘汇
- 环京区域房价炒翻天 房山低价别墅汇
- 马年楼市5大特征 北京二手房价格下跌

80后小夫妻理想家

• 导购| 东南5环地铁综合体23000 第2空港联排送250平

• 资讯| 京今年3成住宅地建低价房 李嘉诚抛资产逃离内地

• 市场| 贫困县新政府楼造价过亿 京供暖费拖欠至少数亿

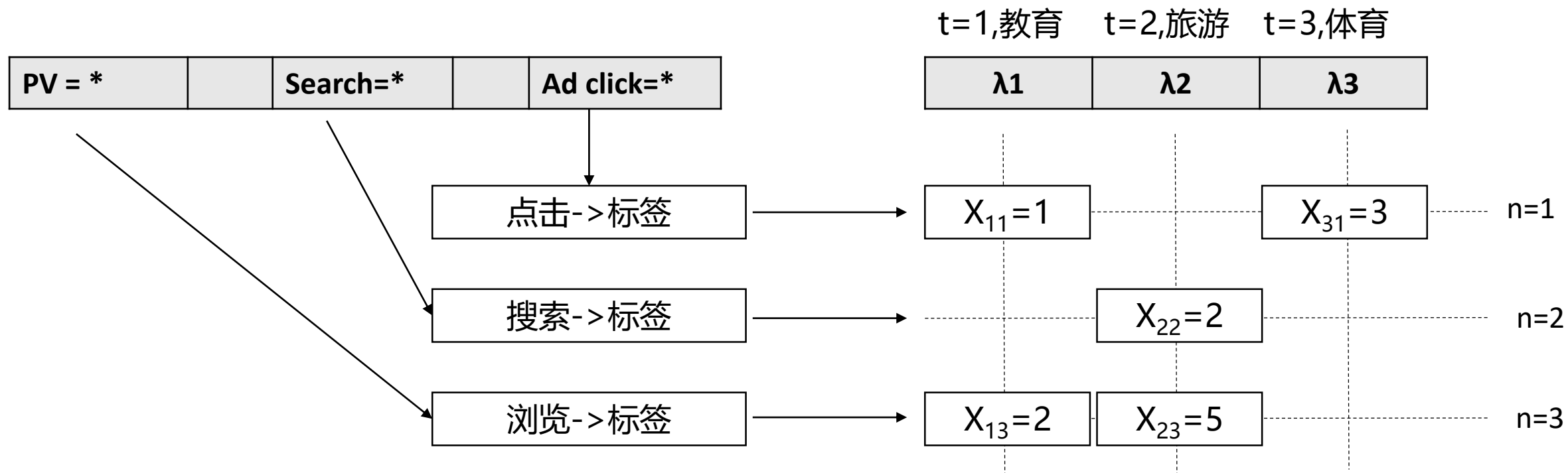
• 购房| 100万-500万住房山低价别墅 5环优质学区房推荐

• 优惠| 给钱就拿钥匙两居现房推荐 学区现房减八万96折

• 热点| 京现房146盘直降50万 小户型最低不到25万

北京 查询 新房 二手房 租房 商铺 产业园

行为定向特征选择过程



行为定向建模

泊松分布 \longrightarrow

定向标签 \longrightarrow

该标签使用的归一化点击数 \longrightarrow

$$p_t(h) = \frac{\lambda_t^h \exp(-\lambda_t)}{h!}$$

频繁性参数 \longrightarrow

$\lambda_t = \sum_{n=1}^N w_{tn} x_{tn}(b)$

N个特征选择函数 \longrightarrow

待优化参数 \longrightarrow

原始行为 \longrightarrow

The diagram illustrates the behavioral targeting model. It starts with the Poisson distribution $p_t(h)$, which is defined by the frequency parameter λ_t and the count h . The frequency parameter λ_t is then defined as a weighted sum of N features, where w_{tn} are the parameters to be optimized and $x_{tn}(b)$ are the original behaviors. Red arrows indicate the flow of information and the mapping of terms to their definitions.

行为定向数据组织

- Session log

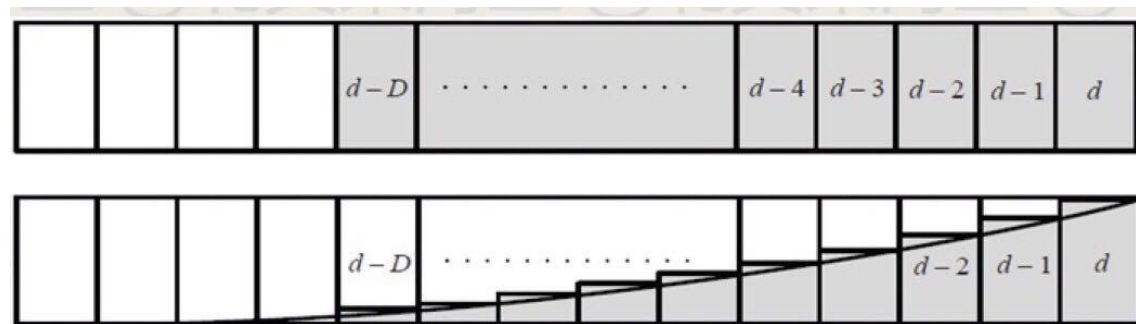
- 将各种行为日志整理成以用户ID为key的形式，作为各数据处理模块的输入源，可以将targeting变成局部计算

- 行为定向两种长期特征累积方式

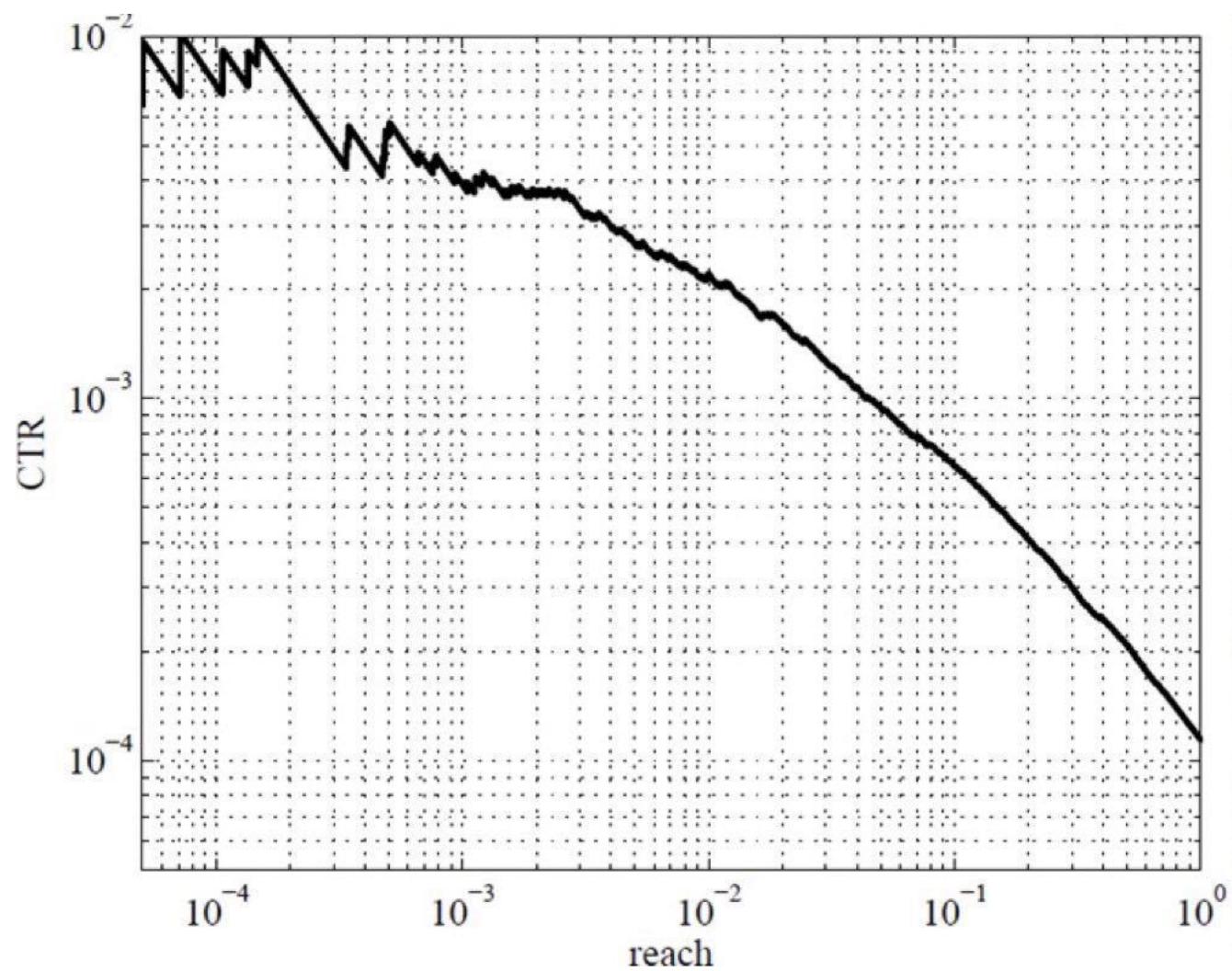
- 滑动窗方式 $\tilde{x}(d) = \sum_{\delta=0}^D x(d - \delta)$

- 时间衰减方式

$$\tilde{x}(d) = \alpha \tilde{x}(d - 1) + x(d)$$



受众定向评测-Reach/CTR曲线



场景数据

- 用户当前所处的场合和状态
 - 例如：地铁上、上厕所、开会中、运动中.....
- 丰富的场景是移动设备所特有的
 - 台式机：位置固定，只有极简单的场景
 - 笔记本：可以移动，但只有工作和娱乐类场景
 - 移动设备：人体器官，具有人们所有可能的场景
- 场景不是上下文
 - 场景是用户的状态，而非媒体的特征

如何检测场景

- 根据移动多种信息来源和传感器进行检测
- 例：检测用户是否处于工作状态
 - 每天上午10点，对用户地理位置采样，如果大多数采样在同一个位置，则该位置为用户上班地点
 - 如果采样没有明显位置规律，则用户为销售等无固定地点工作者
 - 检测到用户处于上班地点，则认为用户处于工作状态
- 注意，场景检测不需要逻辑上完全正确

人口属性定向

- 人口属性
 - 由于监测的原因，实践中主要使用的是性别、年龄
 - 在传统广告中为人群选择的主要语言
- 人口属性定向
 - 以性别定向为例，为二分类问题

$$g = \arg \max_{g \in \{M, F\}} p(g | \mathbf{b})$$

- 需要有一定数量标注样本，特征来自于用户行为

PART-4

数据作为资产的交易

有价值的行为数据来源（一）

- 决策行为
 - 转化（Conversion）、预转化（Pre-conversion）
 - 对应着非常明确的用户兴趣，价值最高
- 主动行为
 - 搜索（Search）、广告点击（Ad click）、搜索点击（Search click）
 - 在明确意图支配下主动产生的行为，价值也很高
- 半主动行为
 - 分享（Share）、网页浏览（Page View）
 - 量最大，用户意图较弱，但也有一定价值

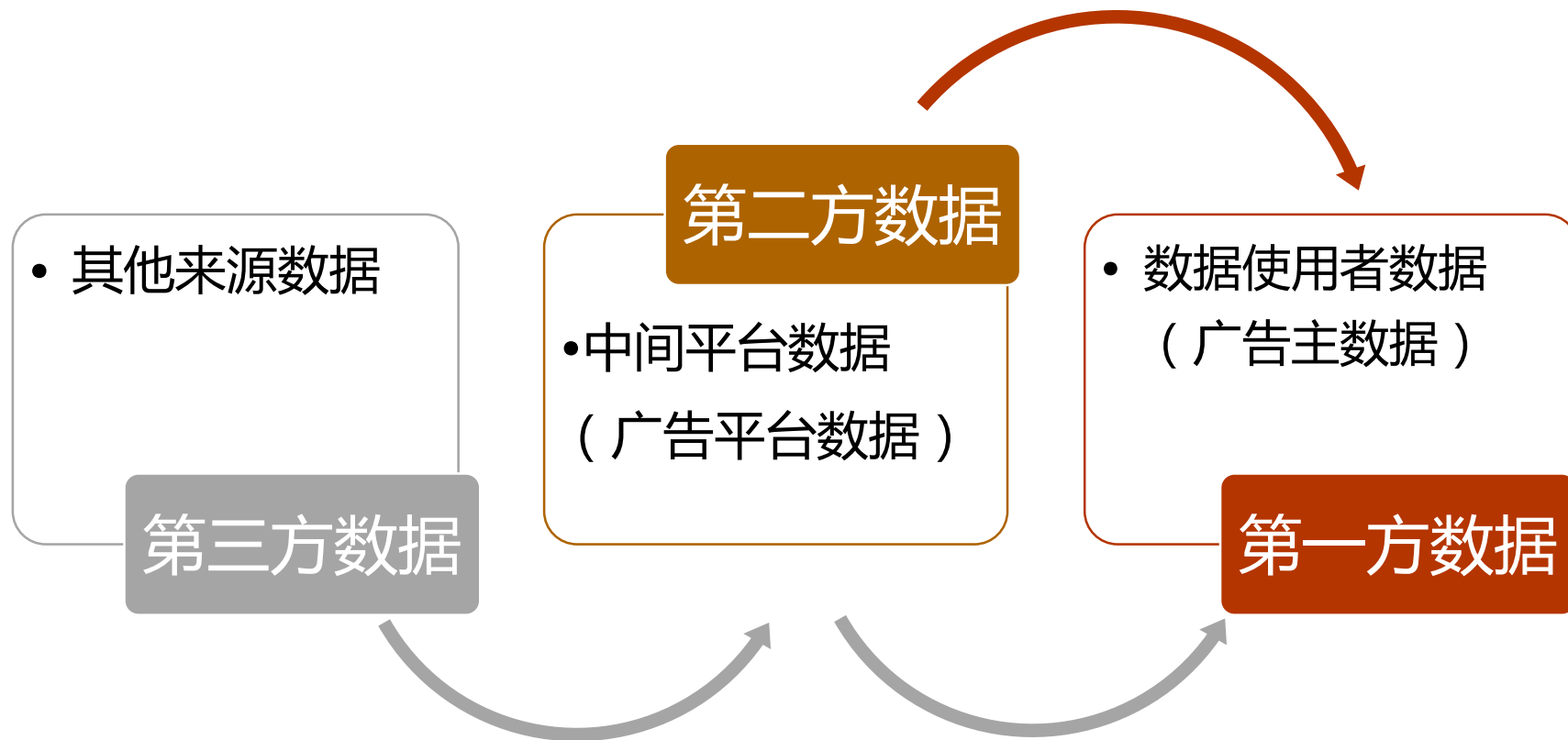
有价值的行为数据来源（二）

- 被动行为
 - 广告浏览（Ad View）
 - 负面的加权因素
- 用户ID
 - 最重要的数据，一串0前面的那个1
 - 稳定、精确的用户ID能大幅提高行为数据使用效率
- 社交关系
 - 可以用与用户兴趣的平滑：当某个人的行为不足，无法进行精准的行为定向时，可以考虑借鉴其社交网络朋友的行为和兴趣。

如何标识一个用户？

- Web/WAP环境
 - **Cookie**：存续性差，跨域时需要映射
- iOS应用
 - **IDFA**：存续性好于cookie，但iOS10有更严格的政策
- Android应用
 - **Android ID**：存续性好于IDFA；IMEI：在中国部分使用
- 无以上ID场景
 - **FingerPrint (IP+User Agent)**；存在http头中，可作缺省标识

三方数据的概念



第一方DMP

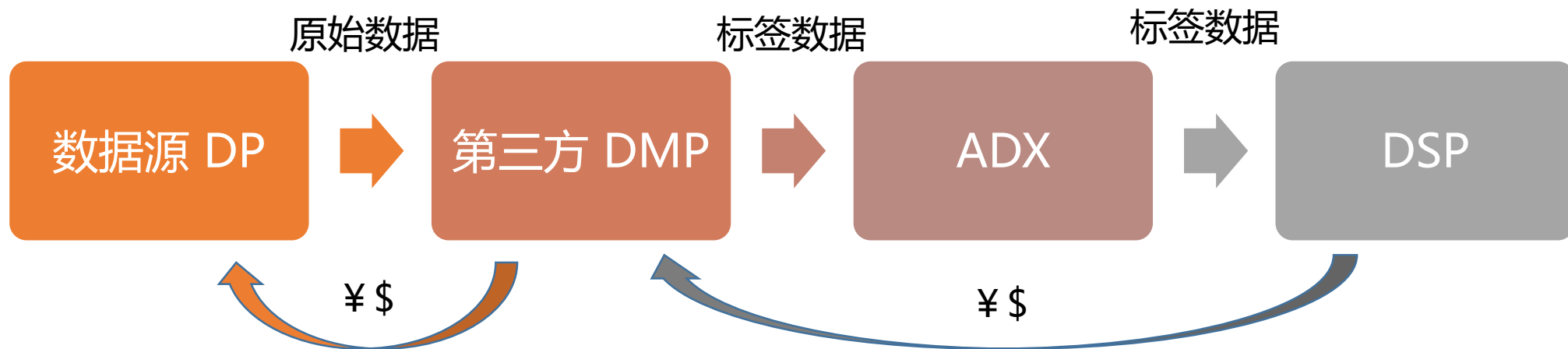
- 目的
 - 为应用/网站提供第一方数据加工和应用能力
 - 结合公开市场第三方数据，加工跨媒体用户标签，支持应用/网站业务运营和可能的广告投放
- 主要特征
 - 第一方用户定制化划分能力
 - 统一的对外数据接口

第一方DMP商业模式



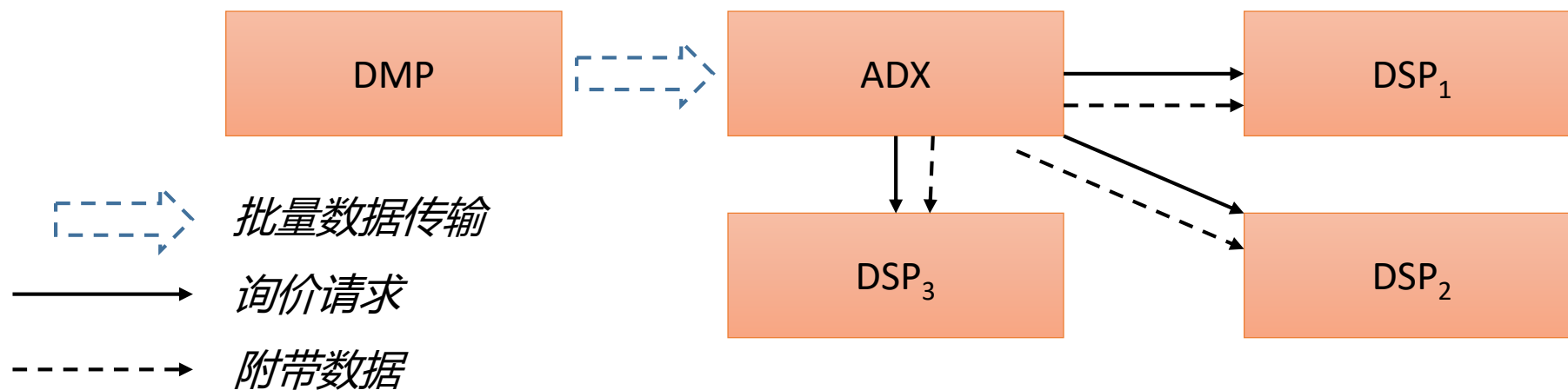
- DMP应数据源（Data Provider，DP）的要求，收集第一方数据，并加工成第一方需要的用户标签
- DP可以根据这些用户标签进行站内的运营，也可以用来指导DSP进行广告投放
- DMP会向DP收取费用，但是绝对不会把数据二次变现

第三方DMP商业模式



- DMP从多个DP那里收集原始数据，按照自己的逻辑加工成用户标签，并向DSP出售标签数据收入
- DMP获得的收入再按照一定的比例分成个DP

数据交易该怎么做？（广告售卖模式）



- 数据传输附着在实时竞价过程中，无额外开销
- 需求方可以自由地选择需要的部分人群数据，并且按照实际的广告展示/数据交易付费

为什么数据不能共享

- 疑问：数据交换似乎在发生啊？
 - 那往往是因为有更高层次的交换，即投资关系
- 为什么大公司不把数据共享出来
 - 你见过大公司把钱共享出来么？
 - 短时间的贴补性共享是可行的
- 政府数据是可以共享的，这本质上是转移支付

如何给数据定价？

- 市场化的定价方式是唯一的选择
- 目前数据的价值是被低估的
 - 上页的交易方式并未限制数据供给次数
 - 这间接地抬高了流量价格，而低估了数据价格
- 能否采用竞价交易方式？
 - 不限量供应的商品，是无法竞价的
 - 数据的限量供应怎么做？

数据隐私的初步认识

- 隐私安全基本原则
 - A29：欧盟负责隐私保护条例制定的委员会
 - A29原则
 - Personal Identifiable Information (PII) 不能使用
 - 用户可以要求系统停止记录和使用自己的行为数据
 - 不能长期保存和使用用户的行为数据
- Quasi-identifier与K-anonymity
 - Quasi-identifier：鼓楼区，36岁，在苏宁易购上班
 - K-anonymity：南京市，30-40岁，互联网行业

互联网行为数据隐私问题

- 稀疏行为数据的新挑战
 - 从一个人观影或购物记录，能否反推他是谁？
 - 实际案例：Netflix推荐大赛，有人从数据集发现同事是同性恋
 - 理论研究：Robust De-anonymization of Large Sparse Datasets
- 深度个性化系统也有隐私安全风险
 - 相关研究课题是差分隐私 (Differential Privacy)
 - 最大化个性化推荐准确率和最小化隐私泄露风险
 - 原始数据的随机化处理
- 隐私是大数据头上的达摩克里斯之剑

PART-5

数据作为资产的内部使用

增长黑客（数据化运营）

- 建立用户转化漏斗
 - 总体的数据发生变化，到底是哪个环节变化了？
 - 每个产品/运营岗位都对漏斗上的某个环节负责
- 用多维报表找到问题
 - 某个环节的数据变化了，原因到底是什么？
- 建立灵活的实验框架
 - 除了被动地发现问题，更要主动地探索新方案
- 数据对于产品的改进，作用是有限的

用户转化漏斗示例

移动用户获取



关注比例而非数量

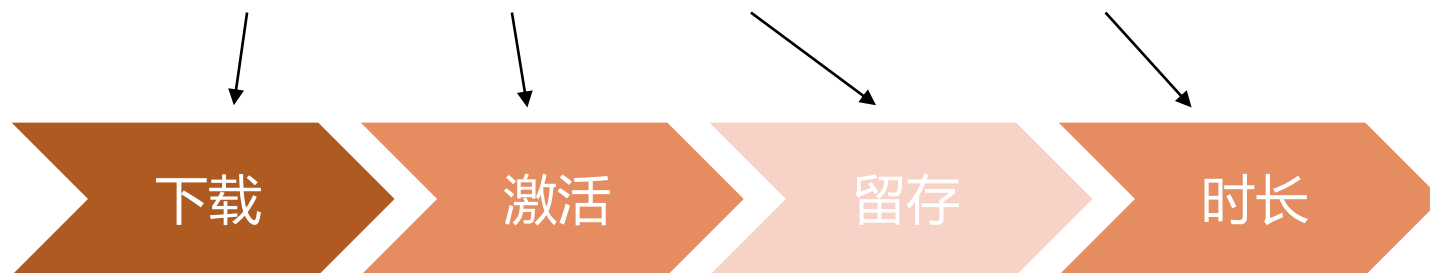
电商用户转化



漏斗设计的原则与作用

- 原则：整个漏斗过程用于优化一个唯一的目标
- 作用：将该目标分解为若干比率的乘积，便于发现问题并优化
- 示例

$$\text{总用户时长} = \text{下载量} \times \text{激活率} \times \text{留存率} \times \text{平均用户时长}$$



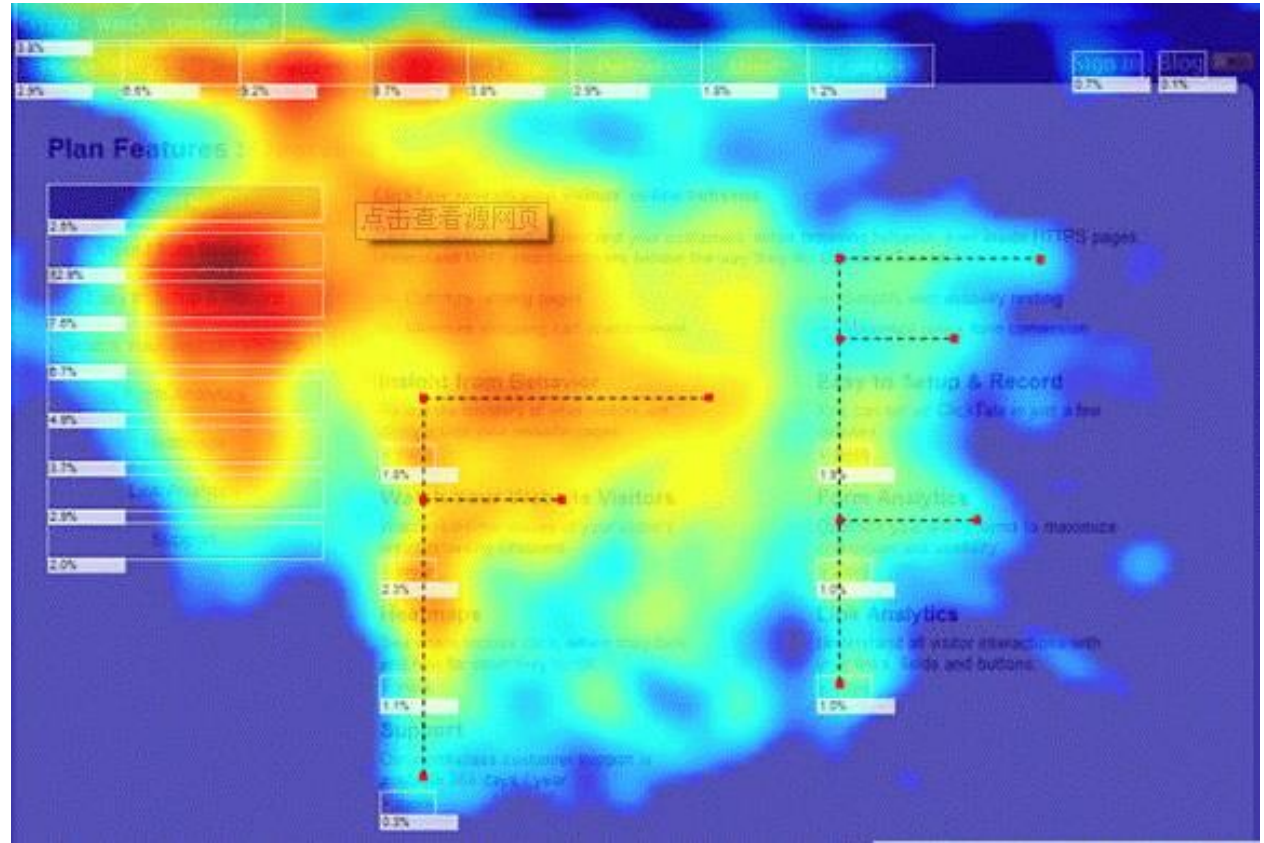
应用分析常见度量

- 转化率
 - 激活数与点击数的比
- {次日/七日/月}留存率
 - 某日激活的用户中，{次日/七日/月}后活跃的用户占比
- {日/月}活跃用户 (DAU、MAU)
 - 每{日/月}活跃的独立用户数
- 用户时长
 - 每个活跃用户平均消耗的时间

所有指标都是可量化可度量

网站分析常见度量

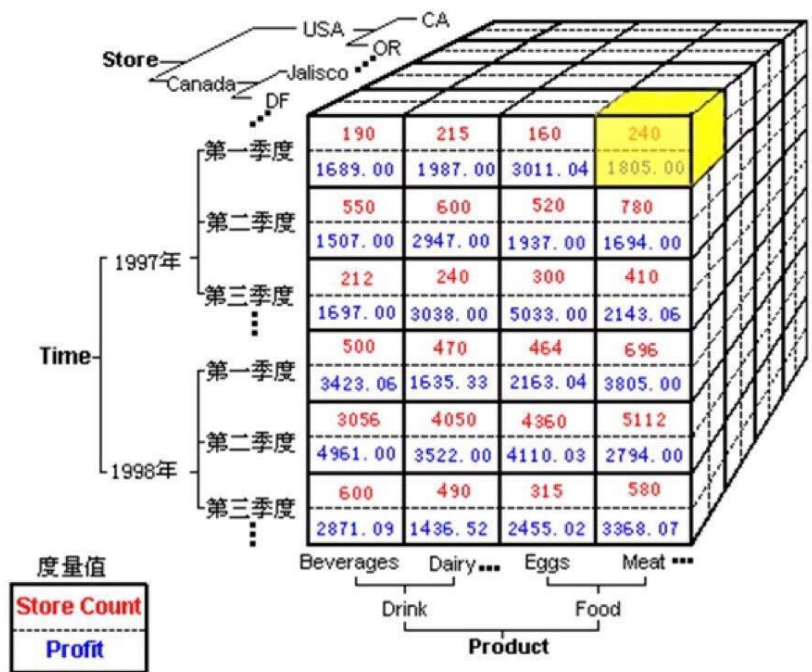
- 访客数
 - UV (Unique Visitor)
- 浏览量
 - PV (Page View)
- 页面停留时长
- 跳出率
 - Bounce Rate
- 网站热力图



网站/应用分析工具

- 网站分析工具
 - Google Analytics、百度统计、CNZZ
 - 无埋点：Heap Analytics
- 应用分析工具
 - TalkingData、友盟+
 - Flurry、Google Analytics
- 应用归因工具
 - Appsflyer、Tune、Adjust、TalkingData

数据魔方 (Data Cube)



- 什么是数据魔方？
 - 用户可以较灵活选择维度组合，得到定制化报表
 - 为人工决策提供便利
- 技术方案
 - OLAP数据库
- 开源方案
 - Saiku+MySQL

为什么需要A/B测试？

- 多维情况下，魔方里大部分区域数据非常稀疏
 - 极端情况：对于新Feature，需要主动分配测试流量
- 某维度上的两个选项（例如两个不同的模型），数据并不是完全可比
- 因此，我们需要一个主动的A/B测试框架，以便
 - 主动分配流量给新的产品特征
 - 保证对比实验的各组在数据上完全可比
 - 尽可能在同样的流量规模上容纳更多的实验

分层实验框架



A/B测试并不是万能的

- 用户产品过于依赖数据会丧失对**关键创新**的把握
 - 汽车无法从“跑得更快的马”进化而来（例：Zynga）
- 多数情形下，需要测试的**可行组合太多**，必须先经过人的筛选，或更复杂的E&E策略
 - 例：每天数十万的新闻，那些有可能最受用户欢迎？
- **博弈性场景**无法通过A/B测试获得可靠结论
- A/B测试最适合的场景
 - **理性产品**、**被动反应**场景

PART-6

数据领域的职业生涯和技能树

应该有怎样的大数据行业视野



本质

什么是大数据？如何利用大数据？

产品

大数据都能做什么？市场上是怎么做的？

技能

我应该准备好哪些能力

大数据职业发展方向

系统工程师
系统架构师

- 数据处理系统
- 高并发服务系统搭建

算法工程师
数据科学家

- 大规模数据集
 - 统计
 - 建模
 - 优化

数据产品经理
产品架构师

- 数据使用逻辑
- 功能设计与优化

大数据职业发展阶段

掌握基础技能和思想

参与完整的业务闭环

独立负责项目或产品

数据产品经理基本素养

基础工具：EXCEL、SQL、Hive/Pig

典型问题：机制设计、冷启动、标签体系

思维模式：依赖数据做决策，建立产品闭环

职能分工：功能产品，策略产品

系统工程师基本素养

编程语言：C/C++、Go、Java

开源工具：No-SQL、Nginx、Spark、Kafka

主要问题：大数据存储与计算、高并发服务

算法工程师基础素养

编程语言：Java、Python、C/C++

知识准备：ML/DL、最优化、分布式计算

主要问题：报表/BI、用户画像、预测模型

参与一个业务闭环过程

- 什么是一个业务闭环—OPEN?
 - 建立明确的优化目标 (Objective)
 - 打通数据记录和分析流程 (Process)
 - 建立A/B测试优化的框架 (Experiment)
 - 将目标按转化网络分解 (Net)
- 通过闭环的优化过程，抚摸和感知数据

示例：在线广告—Objective

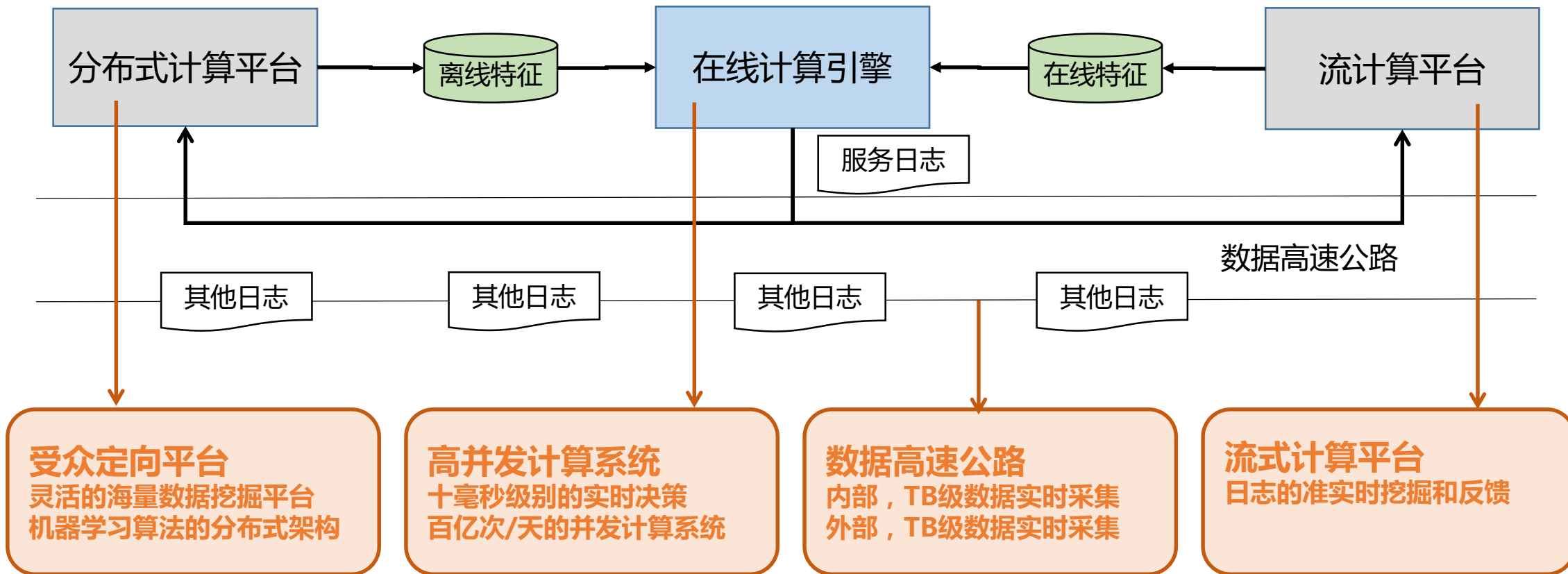
- 计算广告的核心问题，是为一系列用户与环境的组合，找到最合适的广告投放策略以优化整体广告活动的利润。
- 优化问题描述：

$$\max_{a_1, \dots, T} \sum_{i=1}^T \{r(a_i, u_i, c_i) - q(a_i, u_i, c_i)\}$$

广告 用户 上下文

决策对象：一组广告展示 收入 (eCPM) 成本

示例：在线广告—Process



示例：在线广告—Net

展示页



广告页

落地页



广告主网站

转化页



$$eCPM = r(a,u,c) = \mu(a,u,c) - v(a,u)$$

点击率 点击价值

示例：搜索

- Objective :
 - nDCG/Bad case/用户反馈
- Process :
 - 爬虫、倒排索引、相关性排序、用户数据增强
- Experiment
- Net :
 - 搜索->结果展示->点击->翻页/新搜索

主持一项数据产品

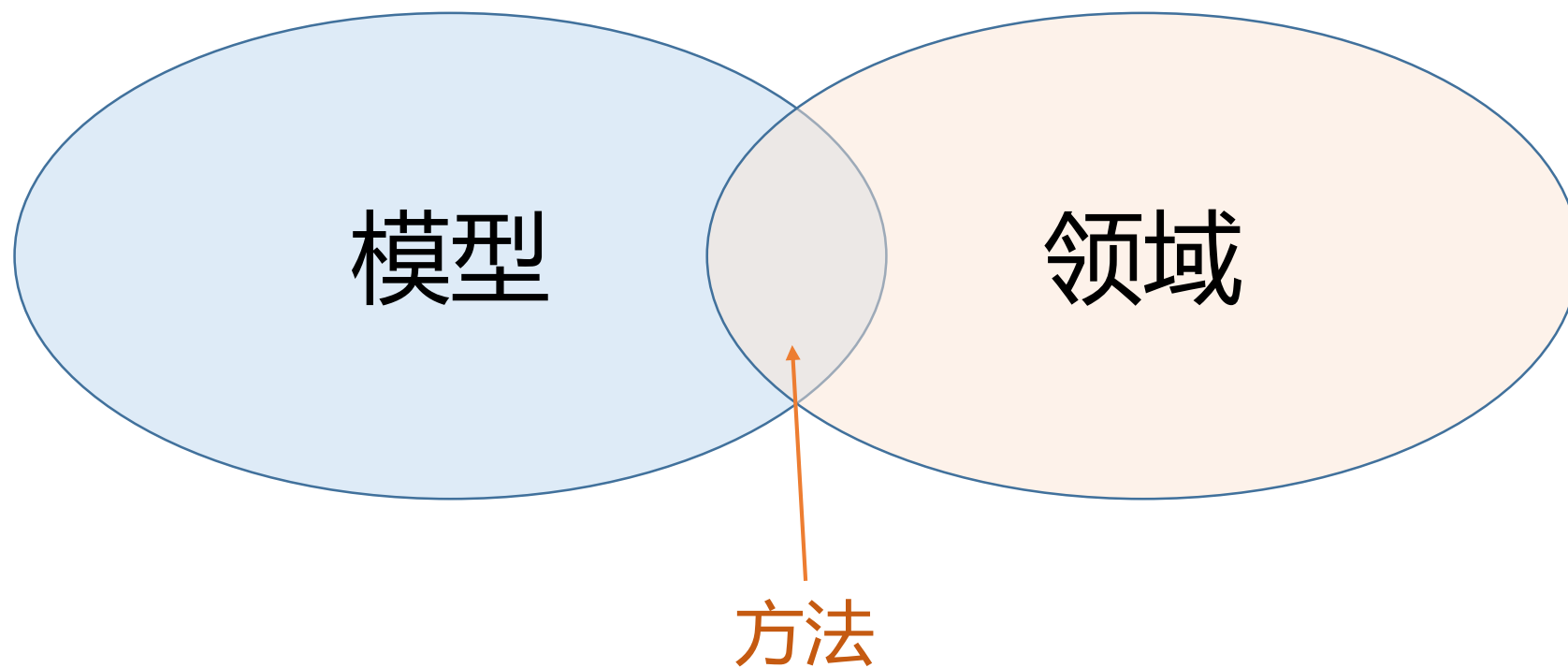
O
定目标

P
搭流程

E
建实验环境

N
做分解

误区一：数据科学的关键在于模型



两个重要的定理

没有免费午餐定理 (No Free Lunch , NFL)

- 没有任何算法在所有数据情形下有天然的优势，哪怕跟随机猜测相比。

丑小鸭定理 (The Ugly Duckling)

- “丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别一样大”

误区二：我需要成为全栈工程师

- 全栈工程师：
 - 全栈工程师，指同时具备前端和后台能力，并能利用多种技能独立完成产品的人。什么是产品？
 - 杂则必然不精
- 不同的技术岗位，需要不同的素养
 - 系统工程师：良好的代码习惯，严谨的测试流程
 - 算法工程师：很多算法逻辑是不容易测试的

误区三：我不是技术出身所以不懂产品

- 什么是技术？
 - 实现产品逻辑的代码、模型和架构
 - 例：深度学习、爬虫、Nginx服务器
- 什么是产品？
 - 定义问题、解决问题的逻辑
 - 例：用户标签体系，冷启动策略，语义检索
- 所有的岗位，都必须深入理解产品

误区四：不断切换从事的业务领域

- 业务领域先验知识的积累，是成功进行数据建模的基础
- 业务领域的商业逻辑需要花时间搞清楚，这是产品决策的基础
- 同一个领域的不断努力可以形成个人口碑

Q&A