个性化：大数据信息暗海的领航员

百分点科技

生活全面向互联网和移动互联网转移

- 可获取和面对的信息成指数式增长
- 用户全景数据的获取和处理成为可能

- 用户注意力严重碎片化：多任务、多渠道
- 人脑的处理能力并未增长

我们的目标

- 以大数据为基础，应用个性化技术，帮助用户从海量信息中筛取所需的信息
  - 数据：entity数据和用户在entity上的行为数据
  - 个性化：用户场景 = 用户意图 + 用户偏好

用户偏好

- Known Likes
- Unknown Likes
- Known Unlikes
- UnKnown Unlikes

- 时间：长期、短期
- 人群：个体、群体

问题定义

- 过滤：屏蔽Unlikes
- 发现：推荐Likes

- Known：根据历史行为提取
- 预测：如何从Known推出UnKnown

Known推出Unknown的基础

- 过去可以预测未来：偏好的可延续性
- 物以类聚、人以群分

用户意图

- 状态判定
- 状态迁移

用户意图提取的基础

- 行为建模
- 行业知识：零售学、传播学等

个性化：问题定义

假设U是用户集合， I是信息集合，个性化技术要解决：

- 令$R(u,\alpha)$是向用户$u \in U$推荐一集信息$\alpha \subset I$的收益，则对于给定的u_0，要求满足$\max R(u_0,\alpha)$的α

如果简化这个问题：

- 令$R(u,i)$是向用户$u \in U$推荐信息$i \in I$的收益，则对于给定的u_0，要求满足$\max R(u_0,i)$的i
- 此时$R(u_0,\alpha)$等价于求top k个i

个性化：收益函数$\boldsymbol{R(u,i)}$

- KPI为导向
- 根据业务需求定义
- 根据业务效果修正
- 连接现实业务和技术实现

个性化：基本技术

- Content Based
- Behavior Based
- Social Based
- Hybrid

面临的挑战

- 数据稀疏
- 冷启动
- 大数据处理与增量计算
- 多样性与精确性
- 用户行为模式的挖掘和利用
- 多维数据的交叉利用
- 效果评估

百分点实时个性化模型（RTPM）

- $U \subseteq S \times (0,1]$：用一系列的场景来代表用户
- $S = L \times T$：场景，对于$(l,t) \in S$：
  - l：代表了用户当前的意图
  - t：描述了用户的偏好目标
- L：用户状态集合
- T，I：类目和标签空间上的实向量集合
- 收益函数： $R(u,i) = \sum_{(s,p)\in u} Q(s,i)p$
- $Q(s,i)$：场景收益，根据业务需求定义

RTPM实例

在为媒体提供的个性化阅读推荐中，我们假设用户的阅读意图分为聚焦和发散两种，用0和1表示，并定义：

$$Q((l,t),i)=\begin{cases}\dfrac{t'i}{|t|\cdot|i|}, & l=0\\ 1-\dfrac{t'i}{|t|\cdot|i|}, & l=1\end{cases}$$

- 用户意图为聚焦时，推荐相关的信息
- 用户意图为发散时，推荐新奇的信息

RTPM关键

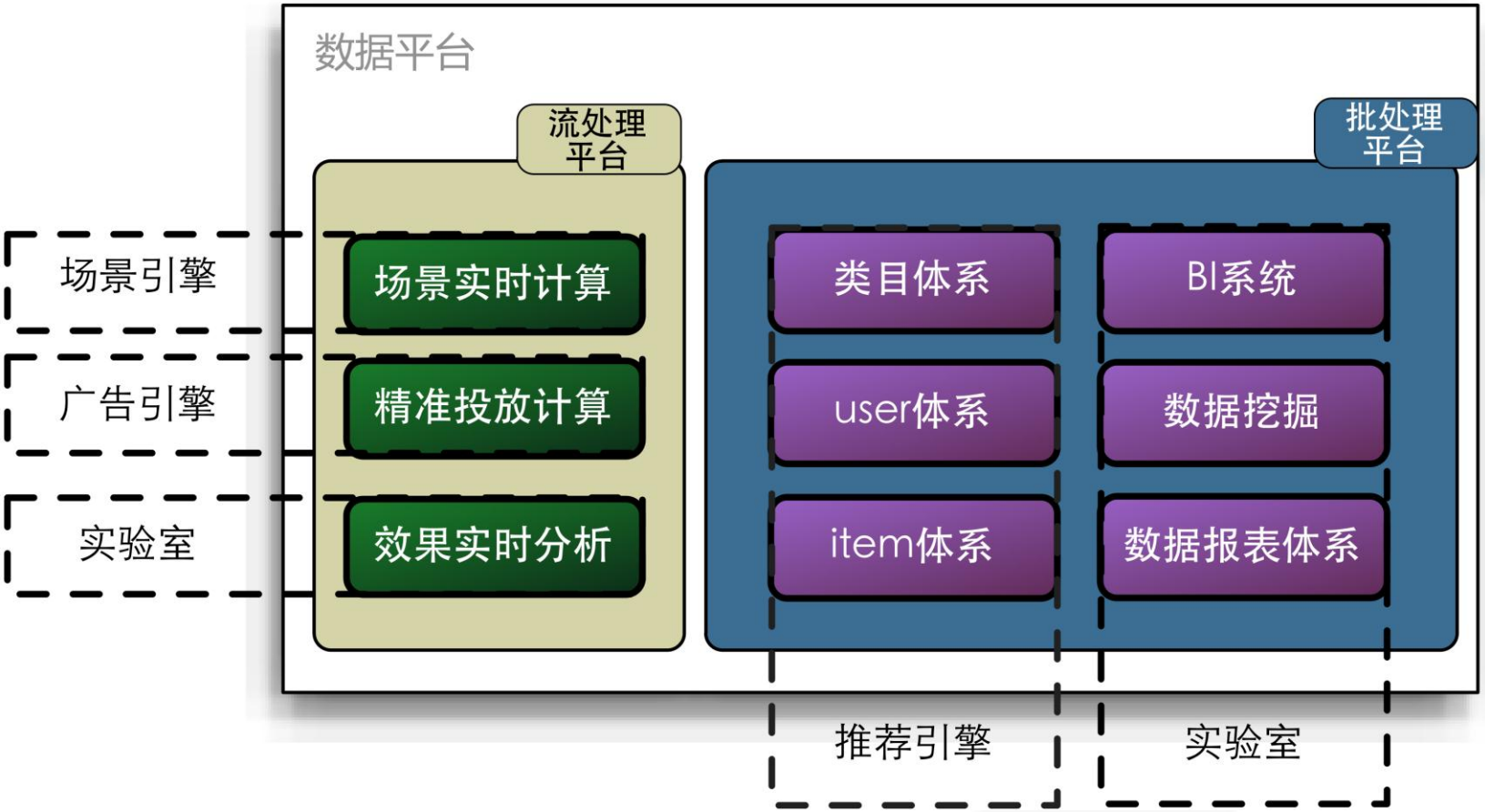
- T, I：
  - 标准化的类目和标签体系
  - 全网信息分类

- L, S：
  - 从业务中抽象出可观测可解释的用户状态
  - 用户行为的状态映射和状态迁移
  - 大数据

RTPM实现

- 标准类目和标签体系
  - 数据抓取
  - 大数据挖掘，自然语言处理
- 用户意图
  - 电商：零售学，购物心理学
  - 媒体：传播学，心理学
  - 从用户行为轨迹中抽象出合理的特征
- 用户偏好
  - 长期偏好+短期偏好
  - 实时+离线处理

个性化数据体系



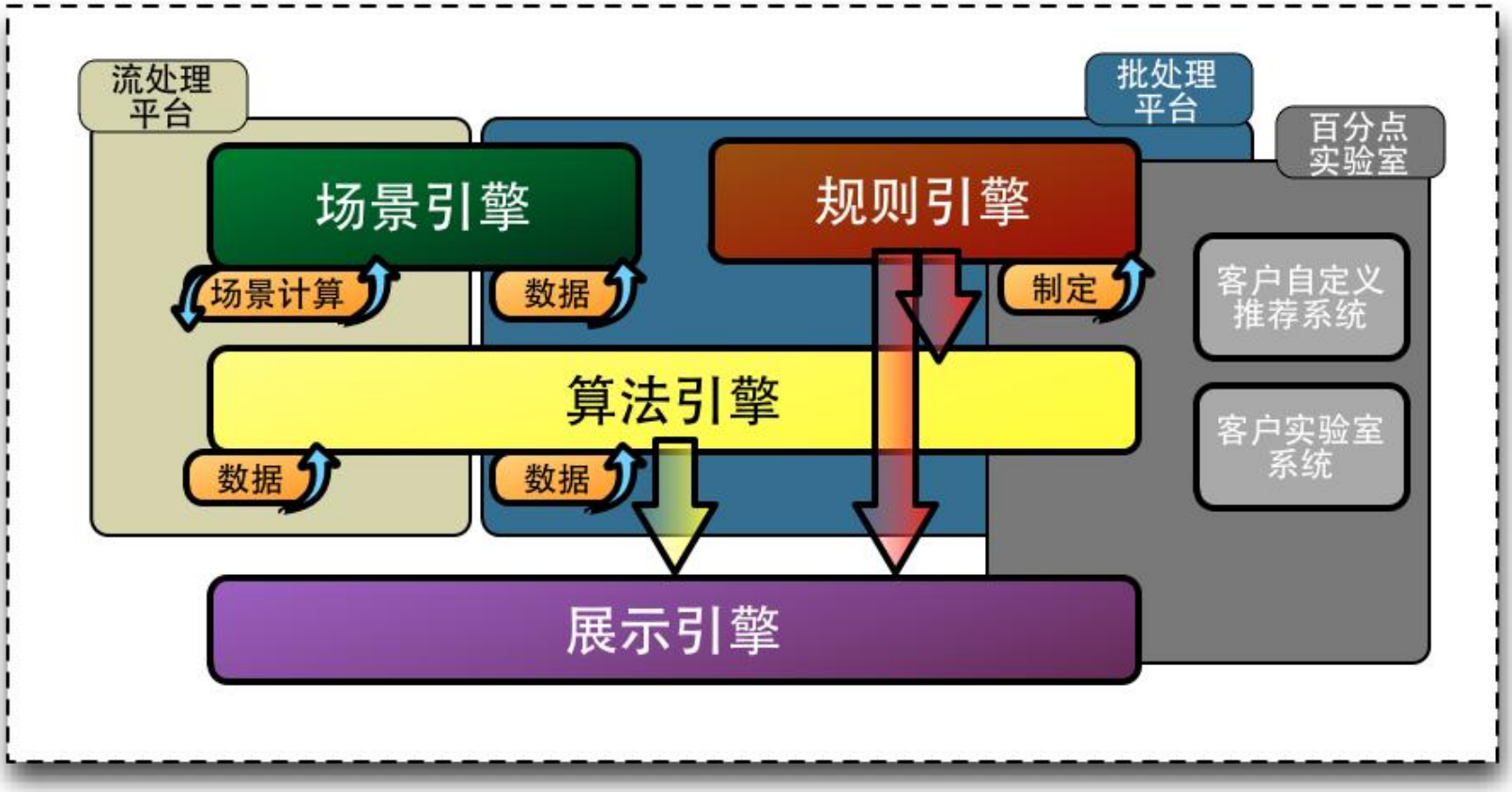
个性化数据体系

- 将网络上形形色色的数据信息，划分为三部分：行为主体“人”(User)、 行为对象“物”(Item)、所有信息粘附与其上的大一统的类目体系
- **类目体系:** 为不等深的树状结构，叶子节点的选择原则为在同一叶子节点下的各个User/Item在选择中“可以相互替代”，针对每级节点，都有特征的标签集和行业词库
- **UserProfile**：人口统计学信息、属性标签、行为方式特征、对Item的偏好
- **ItemProfile**：可抓取的基本信息、数据挖掘的扩展信息、适应的User群体信息

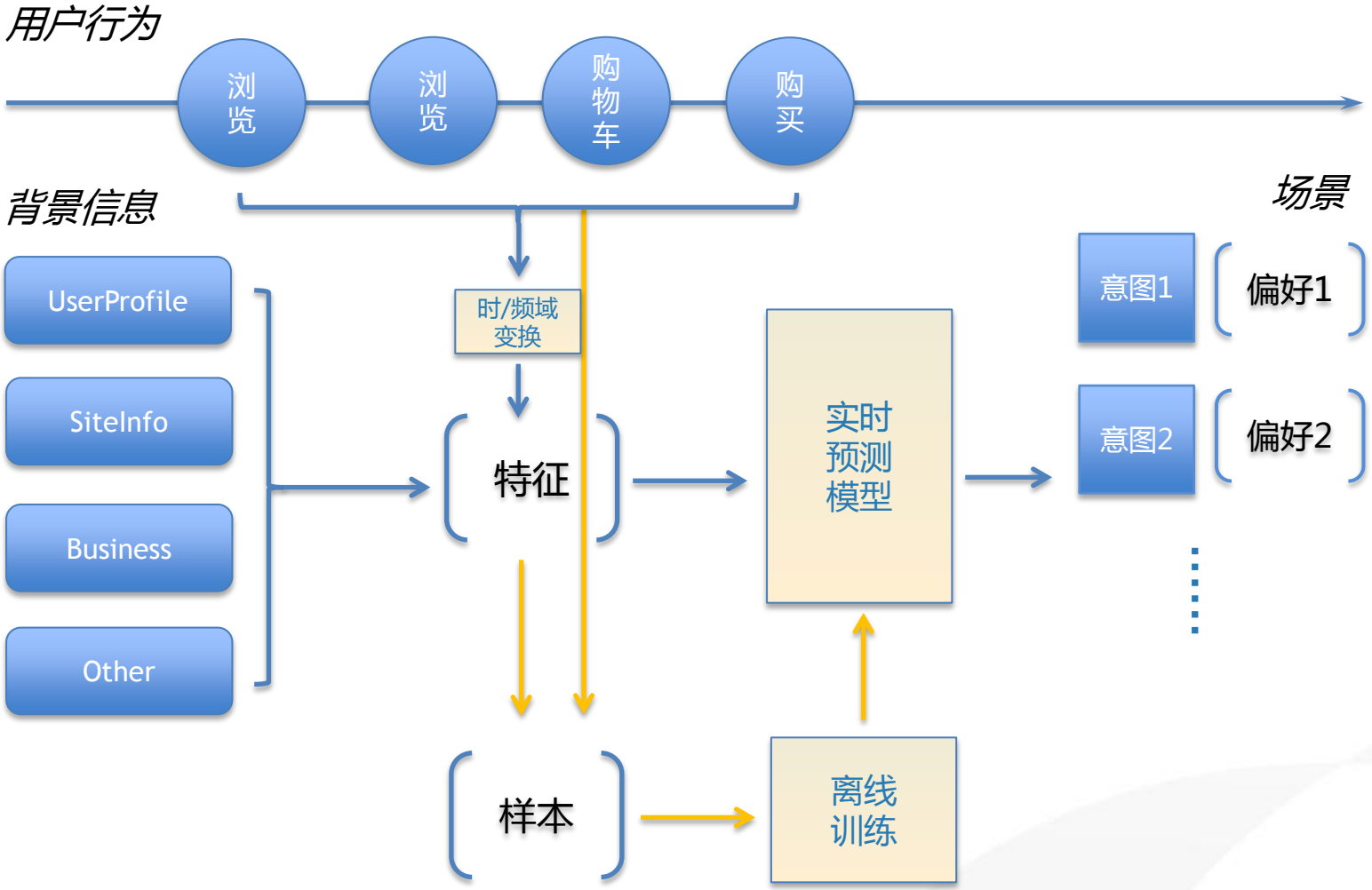
基于Hadoop的大数据挖掘

- 分类：支持向量机(SVM)、贝叶斯（Bayes），Hadoop分布式实现
- 标签发现、标签提取：条件随机场(CRF)模型，Hadoop分布式实现
- MapReduce海量历史数据合并问题：布隆过滤

个性化推荐引擎架构



场景引擎



场景引擎

- 大规模逻辑回归

- 在线
  - 特征处理，预测
  - 5000事件/(秒*服务器)

- 离线
  - 模型训练，评测
  - 3600亿事件 * 50次实验 / 周

规则引擎

- 行业经验，决策树
- 在线
  - 实时计算收益函数
  - 推荐规则语言，快速实现需求
  - 4000请求/(秒*服务器)
- 离线
  - Hadoop分布式训练
  - 决策树，评测
  - 500亿事件 * 10次实验/(周*行业)

算法引擎

- 在线
  - 局部计算、近似计算，精度换时间
  - 增量式CF/TagCF/Content Based/Diffusion
  - 4000事件/(秒*服务器)

- 离线
  - Hadoop批量计算
  - 全量TagCF，关联规则，LDA，统计模型
  - 分级计算
    - 70亿事件/周
    - 300亿事件/月
    - 900亿事件/季

RTPM优点

	业务模型	标准类目标签	场景	实时+离线
数据稀疏	数据多用	数据多用	数据多用	
冷启动	减少冷数据	减少冷信息	减少冷用户	
大数据处理与增量计算		降维	局部计算	分散计算压力
多样性和精确性	方向性指导		技术实现	
用户行为模式的挖掘和利用	方向性指导		技术实现	分散计算压力
多维数据的交叉利用		全网数据互通	全网数据互通	分散计算压力
效果评估	实践检验			

最古老的个性化案例

我问自己——

萨福

对于一个

拥有一切的人

像阿弗洛狄忒

你能给她什么呢？

于是我说——

我将焚烧

一只白母山羊的

肥大腿骨

在她的祭坛上

——（古希腊）萨福

630B.C－592 B.C