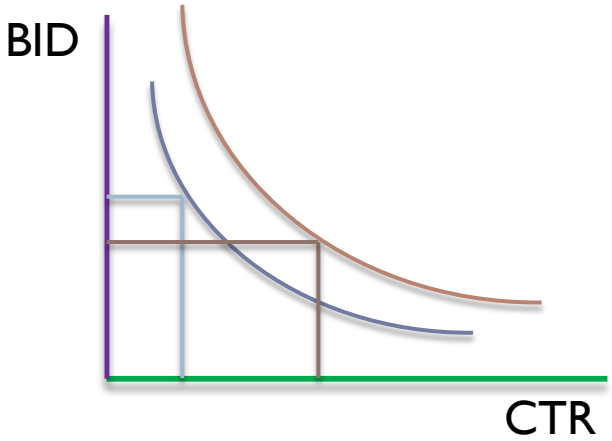
扶翼ECPM提升

回顾与计划

提升eCPM：问题描述

$$ad = Max_i \{ctr_i bid_i\}$$

- eCPM：收益的数学期望
- “提升CTR” => 预估更准的CTR
  - 排序
  - 绝对值 vs 相对值
- “提升bid” => 充分竞价



提升eCPM：技术工作

- 预估更准的ctr
  - 引入更多有效特征
  - 采用更全面的数据
  - 应用更复杂的模型
- 转化效果为目标
  - 引入转化数据
    - Transfer learning
- 充分竞价
  - 预算配速
    - pacing策略

预估更准的ctr

$$p(click \mid user, ad, page)$$

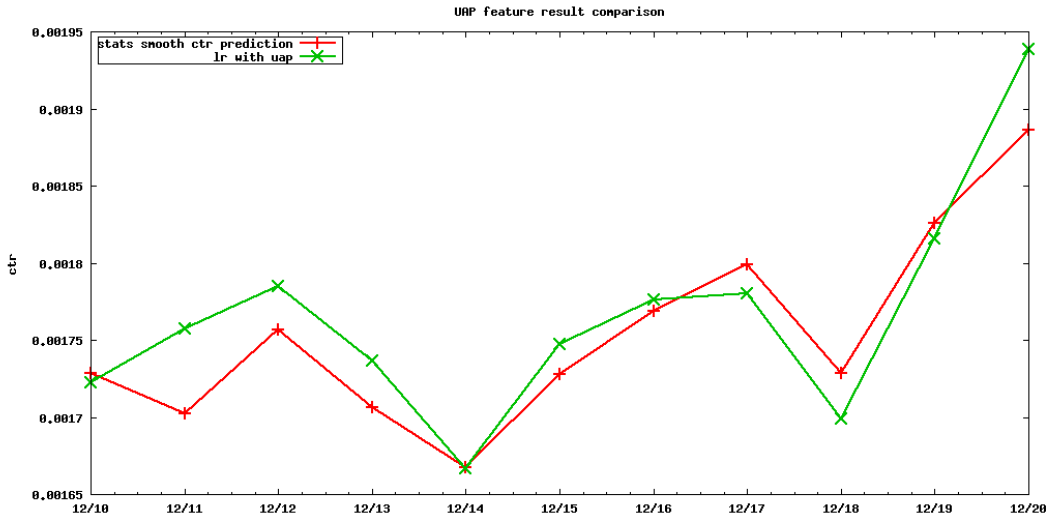
- 模型误差的来源
  - Bias：模型建模能力不足，underfitting
  - Variance：数据集不足以描述实际情况，overfitting
- 针对误差来源解决
  - 降低bias：提升模型复杂度，提高特征描述能力
  - 降低variance：应用更多的训练样本，增加数据的全面性

2015.12-2016.2 有效提升ctr/ecpm的工作

- 引入有效特征
  - User agent，pagechannel
- 增加数据的全面性
  - 单广告位 => 多广告位数据融合
- 增强模型建模能力
  - LR线性 => FM非线性
  - 引入特征层次关系
    - 行业-广告主-广告组-投放单-创意，解决新广告问题

引入有效特征

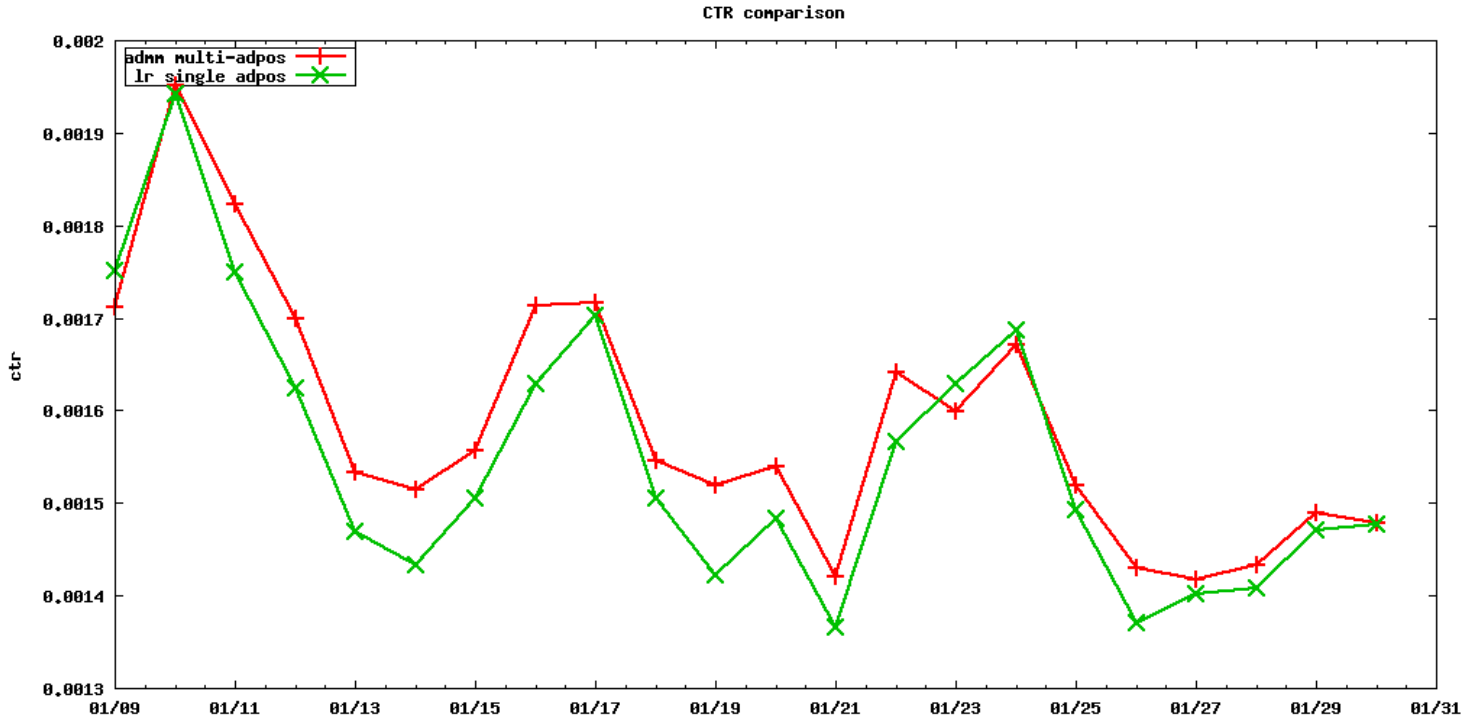
- 通过重构ETL
  - User agent, user profile, pagechannel, …



增加数据的全面性

- 原有不足
  - 单广告位模型，数据不全面，稀疏性
- 解决
  - 分解模型参数 $\vec{w}_t = \vec{w}_0 + \vec{v}_t$
    - $\vec{w}_0$:多广告位的整体因素，$\vec{v}_t$: 当前广告位的影响
  - ADMM算法框架解决的多任务学习

增加数据的全面性


CTR comparison
admm multi-adpos
lr single adpos
ctr
0.002
0.0019
0.0018
0.0017
0.0016
0.0015
0.0014
0.0013
01/09
01/11
01/13
01/15
01/17
01/19
01/21
01/23
01/25
01/27
01/29
01/31

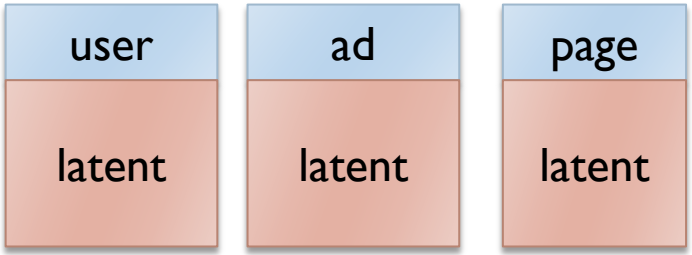
增强模型建模能力

- 原始特征数据

- LR：线性模型

$n = n_u + n_a + n_p + n_u n_a + n_p n_a + n_u n_p n_a$　　额外的特征工程与特征选择

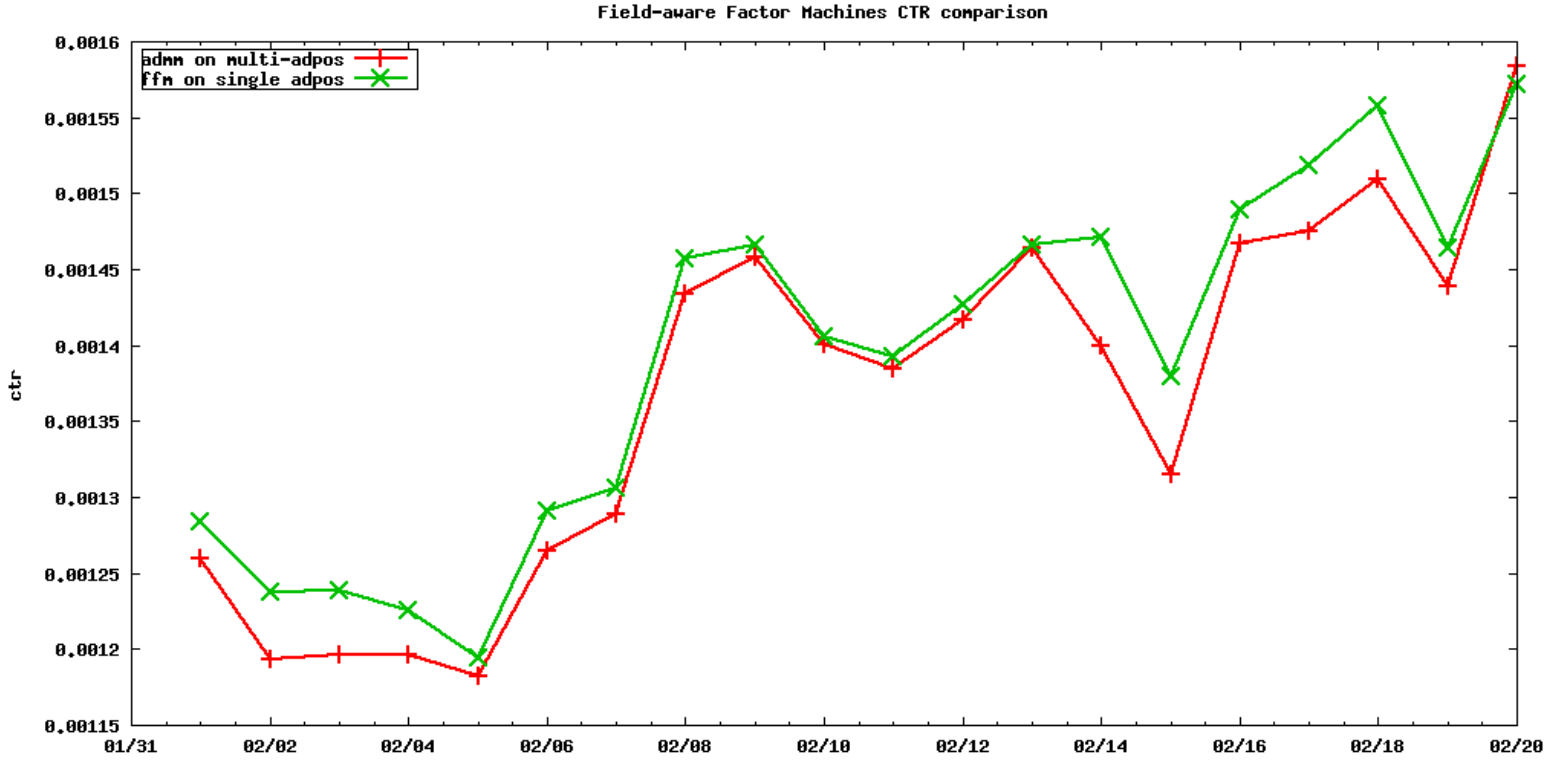
- FM：非线性模型



$m = (n_u + n_a + n_p)(1 + k)$

模型自身的高次项完成特征组合

增强模型建模能力

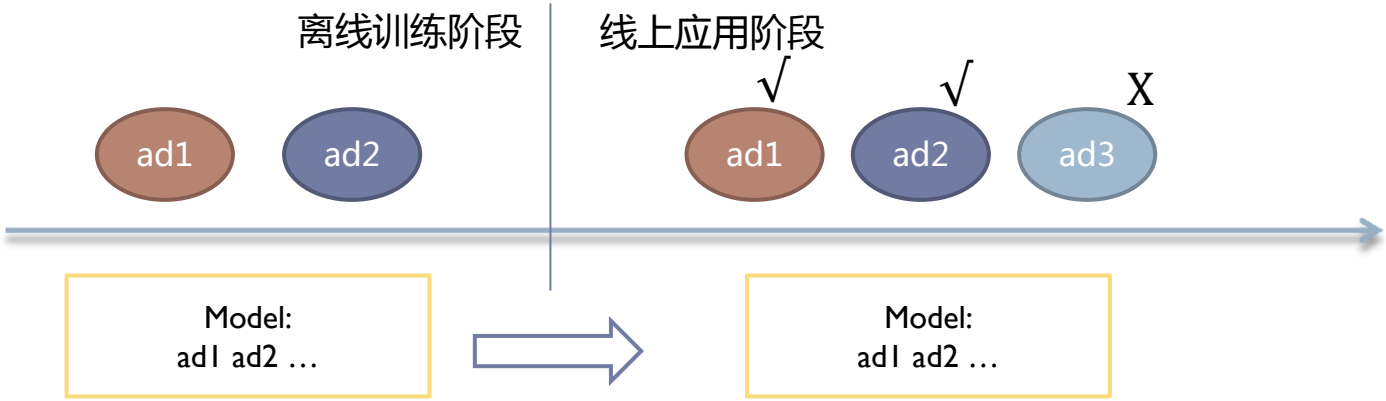

Field-aware Factor Machines CTR comparison
admm on multi-adpos
ffm on single adpos
ctr
0.0016
0.00155
0.0015
0.00145
0.0014
0.00135
0.0013
0.00125
0.0012
0.00115
01/31
02/02
02/04
02/06
02/08
02/10
02/12
02/14
02/16
02/18
02/20

预估更准的ctr：2016H1工作

- 特征引入
  - 标签化的广告特征
  - 创意特征挖掘
- 模型
  - 增强 数据全面性+模型建模能力
  - 建模目标：单纯点击率 => 点击+排序
  - 支持转化数据的引入

2016H1工作与计划：引入有效特征

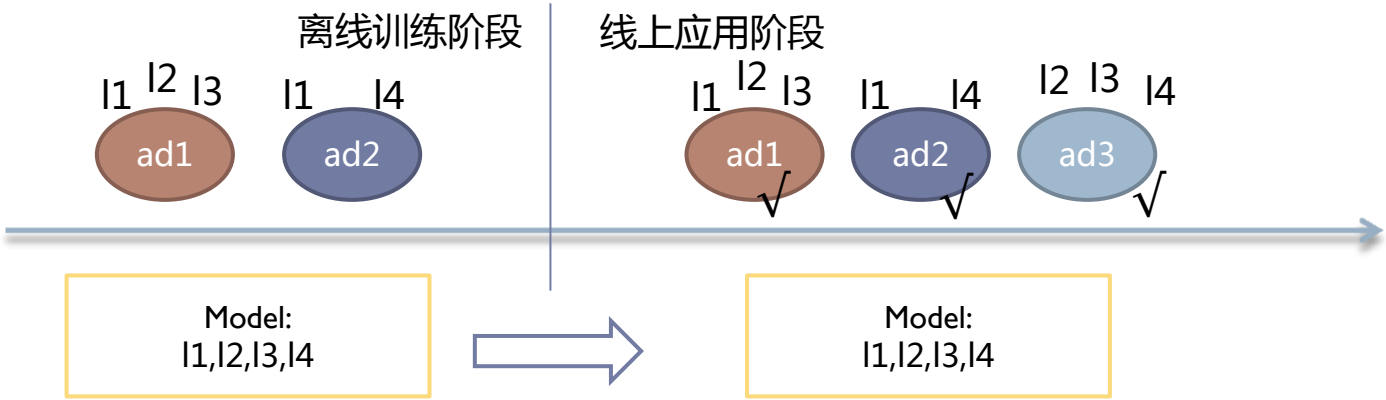
- 新广告冷启动问题



- 长期方案：实时在线模型训练
- 短期方案：标签化广告/广告主特征; 引入层次关系

2016H1工作与计划：引入有效特征

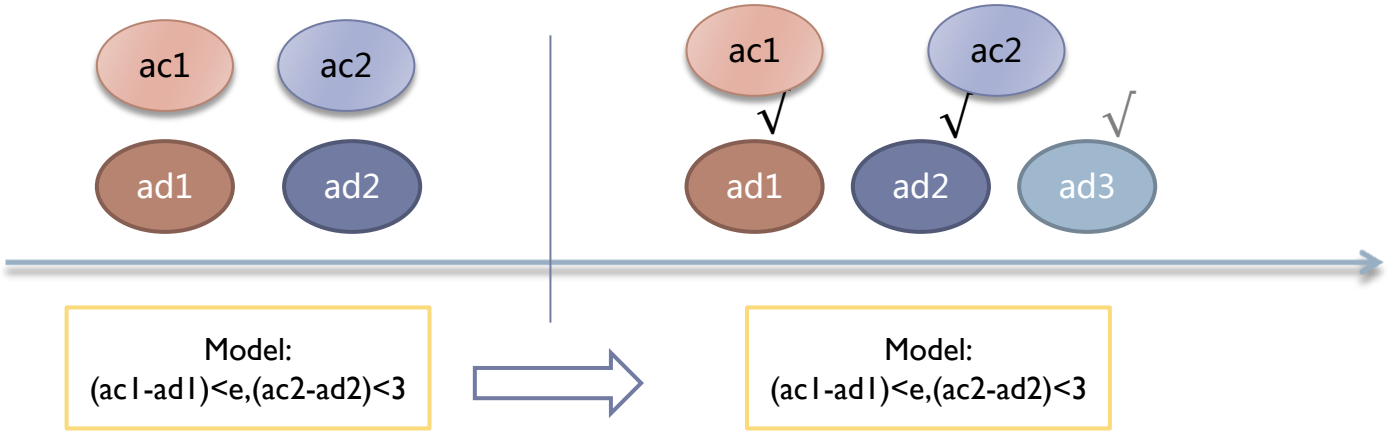
- 新广告冷启动问题



- 长期方案：实时在线模型训练
- 短期方案：标签化广告/广告主特征; 引入层次关系

2016H1工作与计划：引入有效特征

- 新广告冷启动问题



  - 长期方案：实时在线模型训练
  - 短期方案：标签化广告/广告主特征; 引入层次关系

2016H1工作与计划：引入有效特征

- 创意级别特征的挖掘
  - 标题文字
  - 图片
  - 落地页内容
  - 第三方数据
    - 百度搜索结果数等

2016H1工作与计划：数据+模型提升（1）

- ADMM多广告位+FFM模型训练
  - 已有的两项工作的合并
  - 数据全面性+模型建模能力
  - 目标eCPM提升1.5%

2016H1工作与计划：数据+模型提升（2）

建模目标的调整 目标ecpm提升 2.5%

- 点击概率的预估

  $p(click\ |user, ad, page)$

  - 行业惯例，标准方案
    - 要求概率模型足够精准

- 广告间的pk

  $p(ad_1 \succ ad_2\ |user, page, ctx)$

  - 与线上竞价一致
    - 增加对排序准确性的建模

$$w = argmin_w\ \eta Loss(w, D) + (1-\eta) Loss(w, P)$$

2016H1工作与计划:支持转化数据的引入

转化数据
- 严重稀疏
- 高度相关

- 模型上支持引入
  - 作为特征：过于稀疏
  - 作为样本：分布与点击不同
  - Transfer learning方案

提升eCPM：技术工作

- 预估更准的ctr
  - 引入更多有效特征
  - 采用更全面的数据
  - 应用更复杂的模型
- 转化效果为目标
  - 引入转化数据
    - Transfer learning
- 充分竞价
  - 预算配速
    - pacing策略

保证充分竞价：问题与目标

- 问题
  - 优质的广告会快速用光预算，并提前退出
    - 广告的流量倾斜，客户体验不佳
    - 竞价程度降低，GSP竞价下扣费会降低
    - 预算余额到投放控制之间有延迟，造成0扣费点击
- 优化目标
  - 平滑预算+提升广告效果，“预算配速”
  - 避免调整bid值

预算配速

$\boldsymbol{B} = \{B^{(1)}, \dots B^{(T)}\}$T个时段的预算分配, $\boldsymbol{C} = \{C^{(1)}, \dots C^{(T)}\}$T个时段的实际消耗;

给定广告投放单，对于一次广告展示机会i:

设定的参与竞价的概率r_i

$s_i \sim Bern(r_i)$ 参与竞价; w_i竞价胜出; c_i cost; p_i点击率; $q_i \sim Bern(p_i)$;

$$C = \sum_i s_i \cdot w_i \cdot c_i \ ; \ P = C / \sum_i s_i \cdot w_i \cdot q_i$$

预算控制+收益最大化：

$$\text{minimize} \quad P$$

$$subject\ to \quad C = B, \sqrt{\frac{1}{T}\sum_{t=1}^{T}(C^{(t)} - B^{(t)})^2} \leq \varepsilon$$

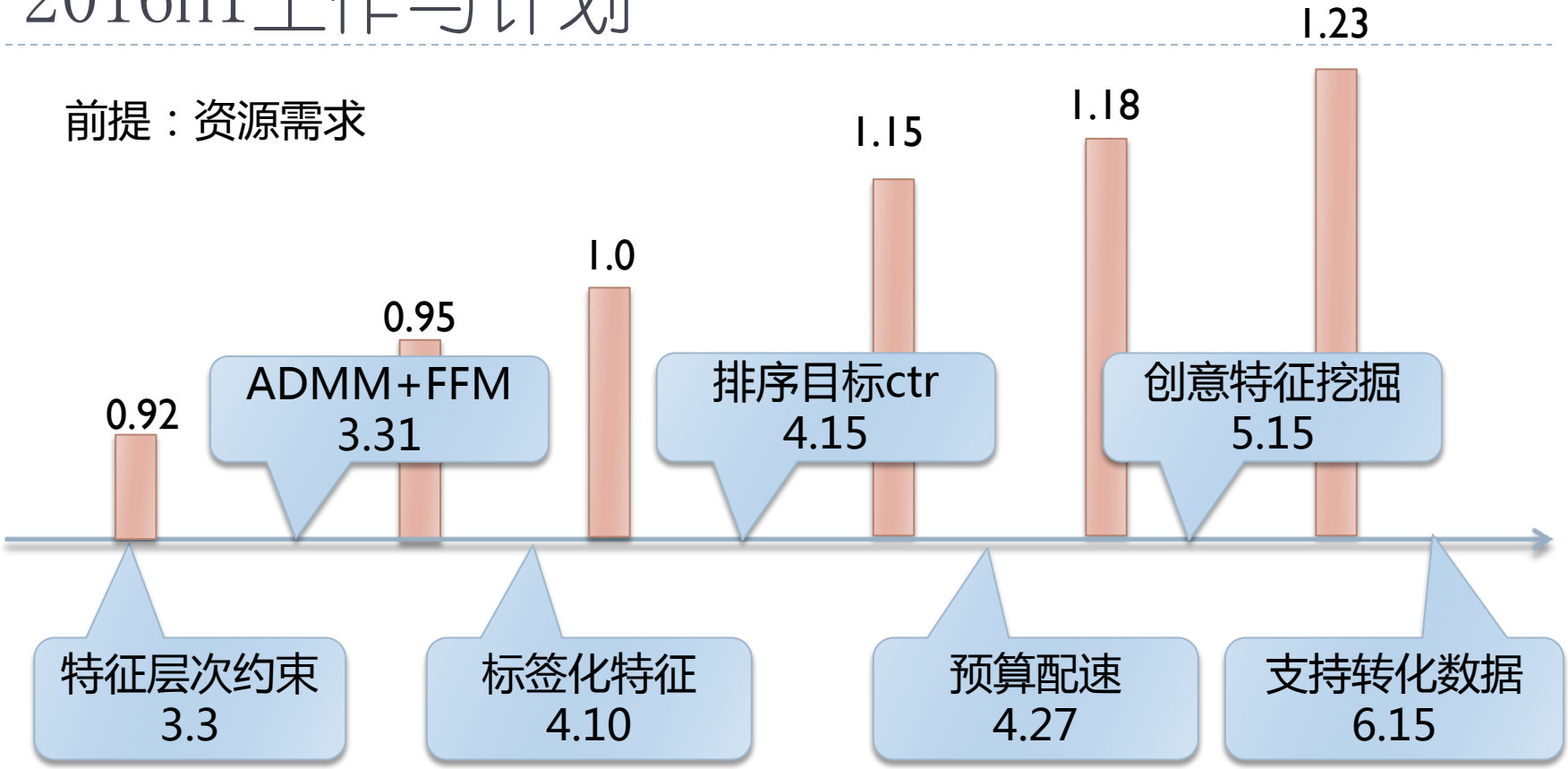
线上应用：实时根据$\{B^{(1)}, \dots B^{(t)}\}$和$\boldsymbol{C} = \{C^{(1)}, \dots C^{(t)}\}$来**调整$\boldsymbol{r_i}$**，

相当于online的求解这一最优化问题

预算配速

- 依赖
  - **数据实时流**统计扣费信息，更新pacing factor
  - **引擎**读取更新的pacing factor r向量
  - 算法逻辑中应用pacing factor

2016H1工作与计划



2016H1工作与资源需求

- 人
  - 2~3人，能力≈现有人员，4月底前
- 机器
  - 数据中心计算结点扩容
  - 自有parameter server：现有4台，需要~10台
- 配合
  - 引擎，数据实时流，日志，产品推动
    - 前端曝光日志数据
  - 算法服务与算法插件