竞赛经验分享
个性化推荐，搜索广告，RTB

@严强Justin
scmyyan@gmail.com

报告内容

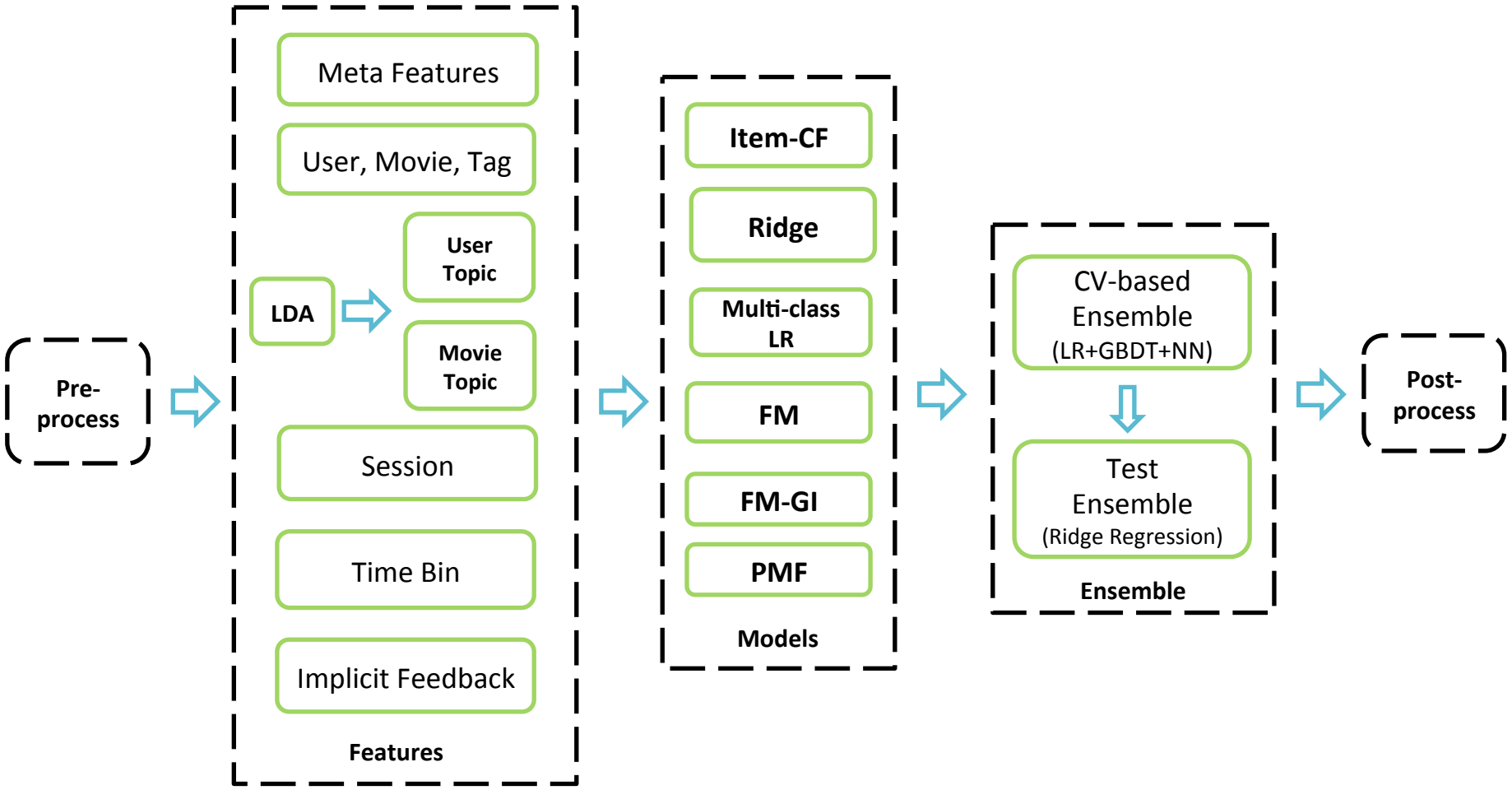
- 介绍
- 方法
  - Recommendation
  - Churn Prediction
  - Search Ads CTR Prediction
  - RTB
- 经验
  - Feature
  - Model
  - Ensemble
- 总结

比赛

时间	比赛	成绩	排名
2013.12	ICDM 2013 Personalize Expedia Hotel Searches Contest	NDCG@38: 0.53102	第五
2013.05	品友RTB算法大赛 Season 1 (DSP – CTR Prediction + Bidding +Pacing)	Score：1960	第一
2013.04.14	Data Science London Big Data Hackathon (Find Influencers in SNS)	AUC: 0.8782	第二
2013.03 -- 2013.05	百度电影推荐算法大赛 (Movie Rec – Rating Prediction)	RMSE：0.5920	第二
2012.11 -- 2012.12	WSDM Challenge 2013 (SE User Churn Prediction)	AUC： 0.8433	第三
2012.03 -- 2012.05	KDD CUP 2012 (Search Ads CTR Prediction)	AUC： 0.8030	第三
2011.03 – 2011.06	KDD CUP 2011 (Music Rec – Rating Prediction)	RMSE: 19.90	第五

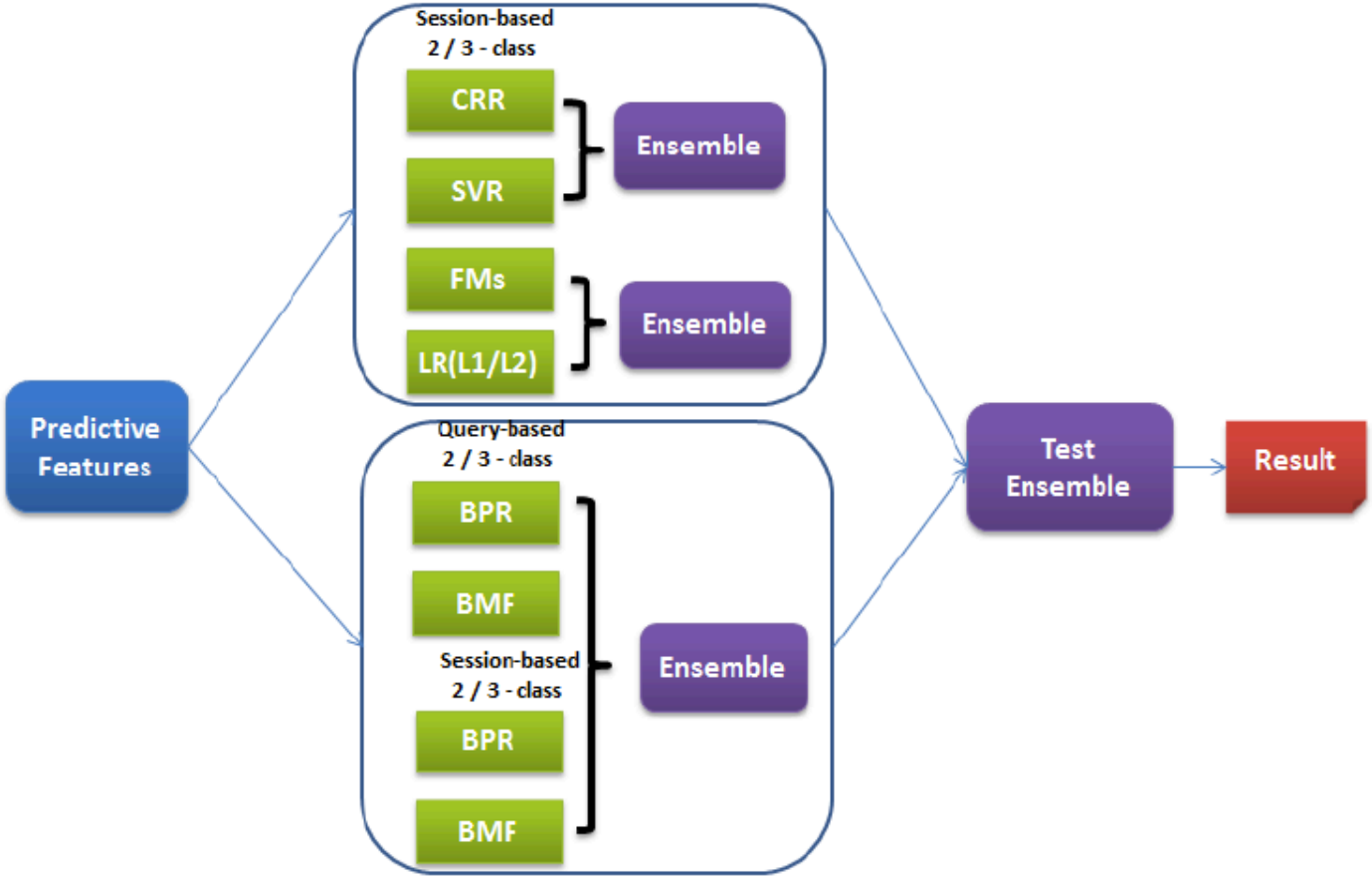
方法

- Recommendation – 百度电影推荐比赛



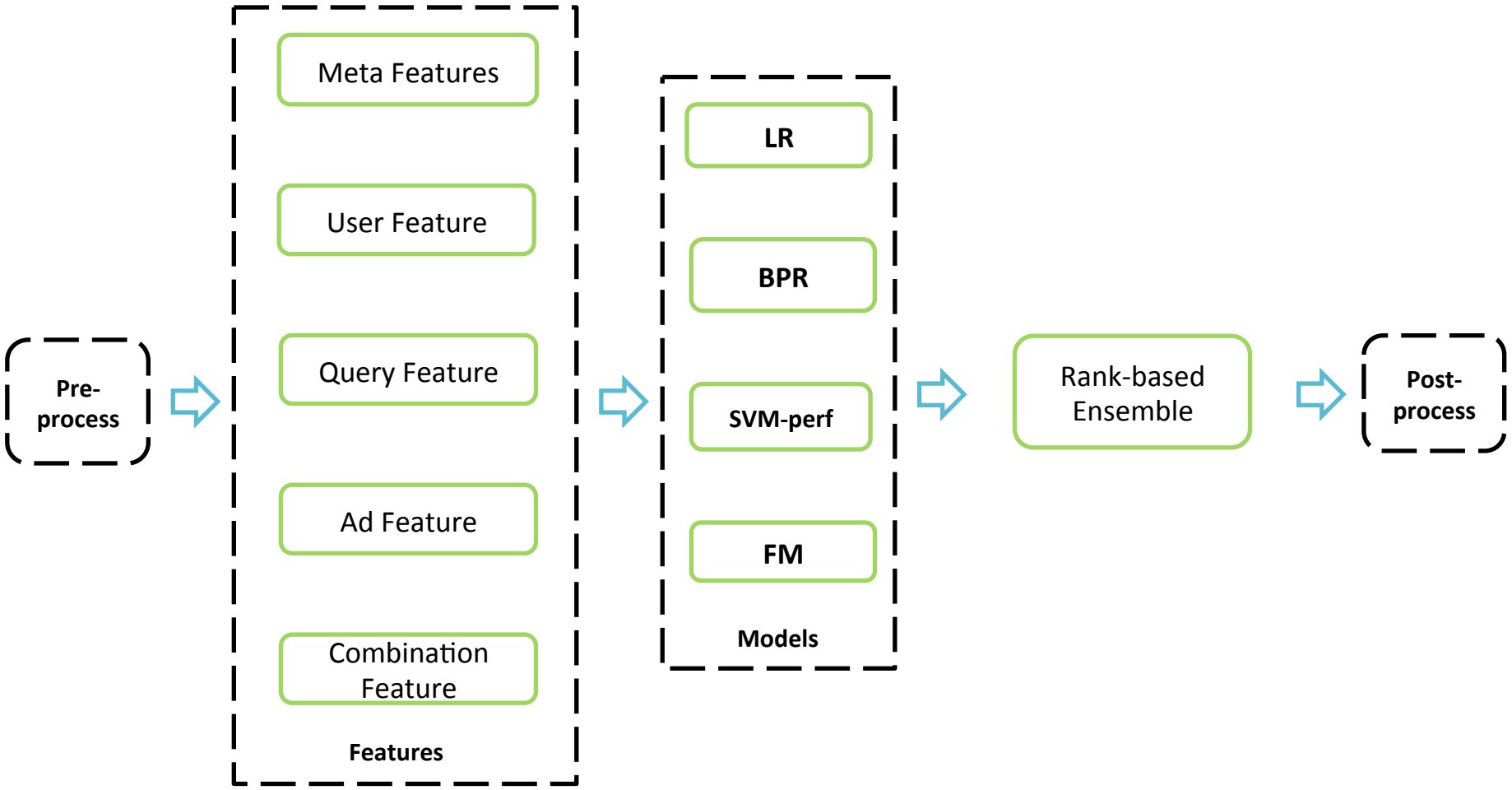
方法

- Prediction – Churn Prediction



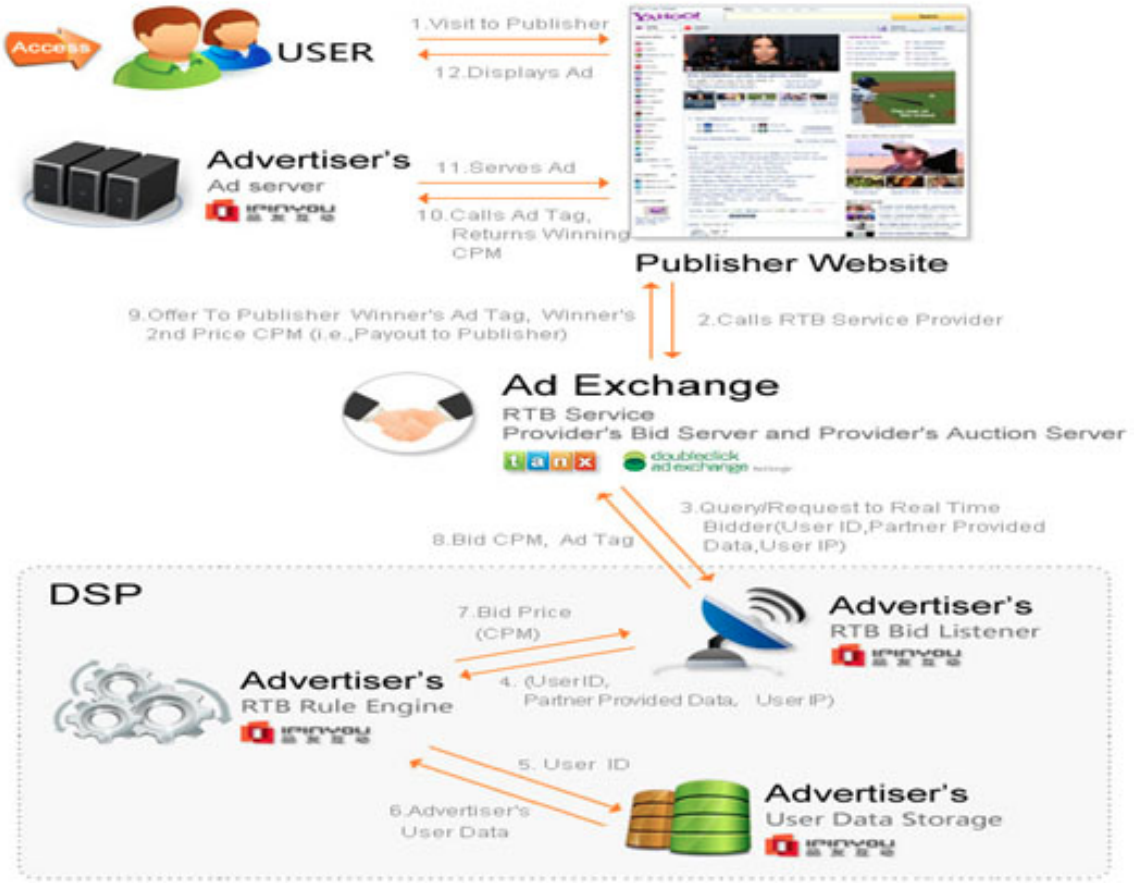
方法

- Prediction – Search Ads CTR Prediction



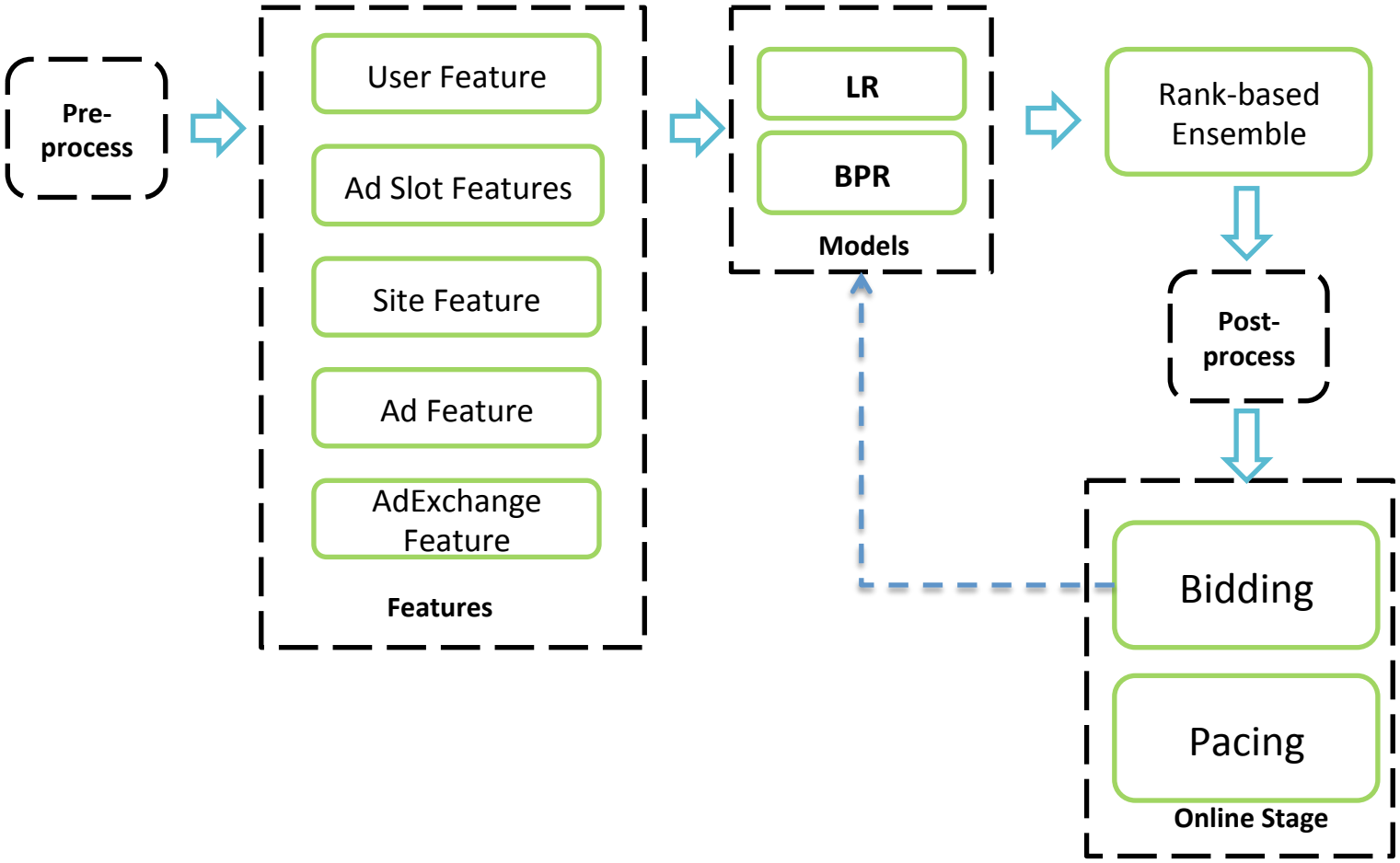
RTB

- DSP算法
  - CTR Prediction
  - Bidding
  - Pacing



方法

- RTB


Pre-process
User Feature
Ad Slot Features
Site Feature
Ad Feature
AdExchange Feature
Features
LR
BPR
Models
Rank-based Ensemble
Post-process
Bidding
Pacing
Online Stage

RTB

- CTR – Prediction
  - Model
    - LR-l1 (Sparse)
    - AdPredictor (Online Learning)
  - Feature
    - 用户: 区域、城市、User Agent、(User Tags)
    - 广告: 广告主ID、创意ID
    - 广告位: type、size、可见性、形式
    - Site: 域名
    - Ad Exchange: Adx/Tanx/Tencent
- 经验
  - CTR预估不是关键
  - 优化Conversion Rate很难

RTB

- Bidding

基于价值的出价(与M6D的算法类似)

展现的价值 = 点击概率*点击价值

出价模型：

$$bid = Base\,\mathrm{Pr}\,ice * (\frac{P(c \mid u,i,a)}{BaseCTR})^{\lambda}$$

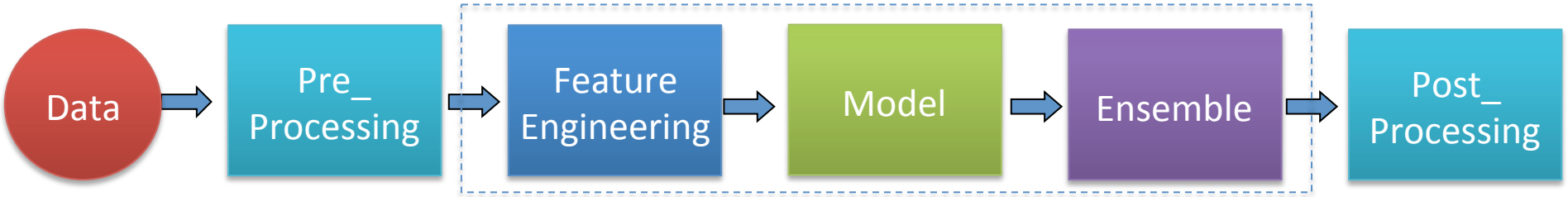
λ参数调节CTR对出价的影响程度。

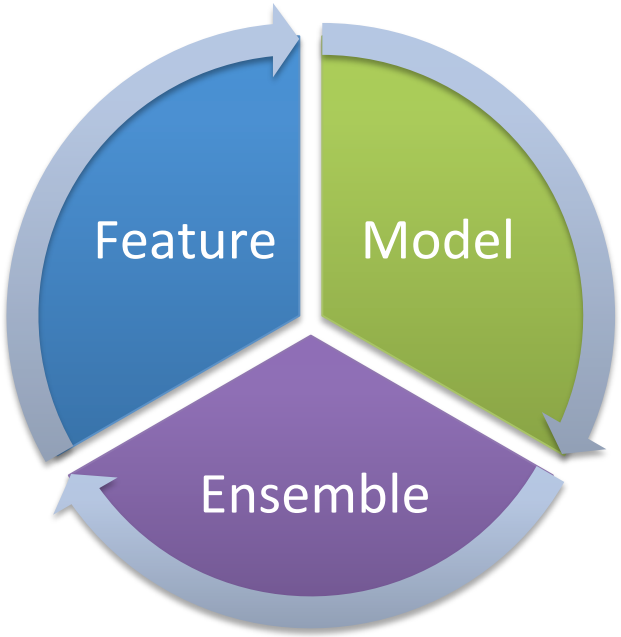
BasePrice和BaseCTR，λ三个参数，通过实验决定。调整原则是使得预算刚好在时间结束时用完。

RTB

- Pacing
  - 预估流量
  - 预算控制(预算、BaseCTR、BasePrice)
    - 分AdExchange
    - 分Campaign

Framework

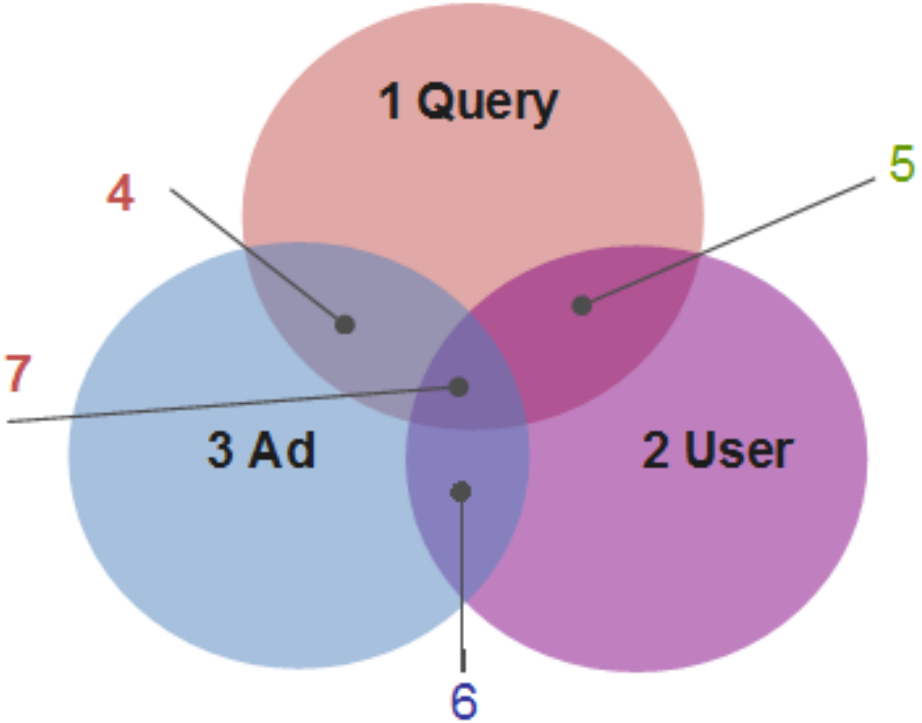

Data
Pre_ Processing
Feature Engineering
Model
Ensemble
Post_ Processing


Feature
Model
Ensemble

Feature

Feature

- 特征分类
  - Low-level vs. High-level
  - 简单特征 vs. 组合特征



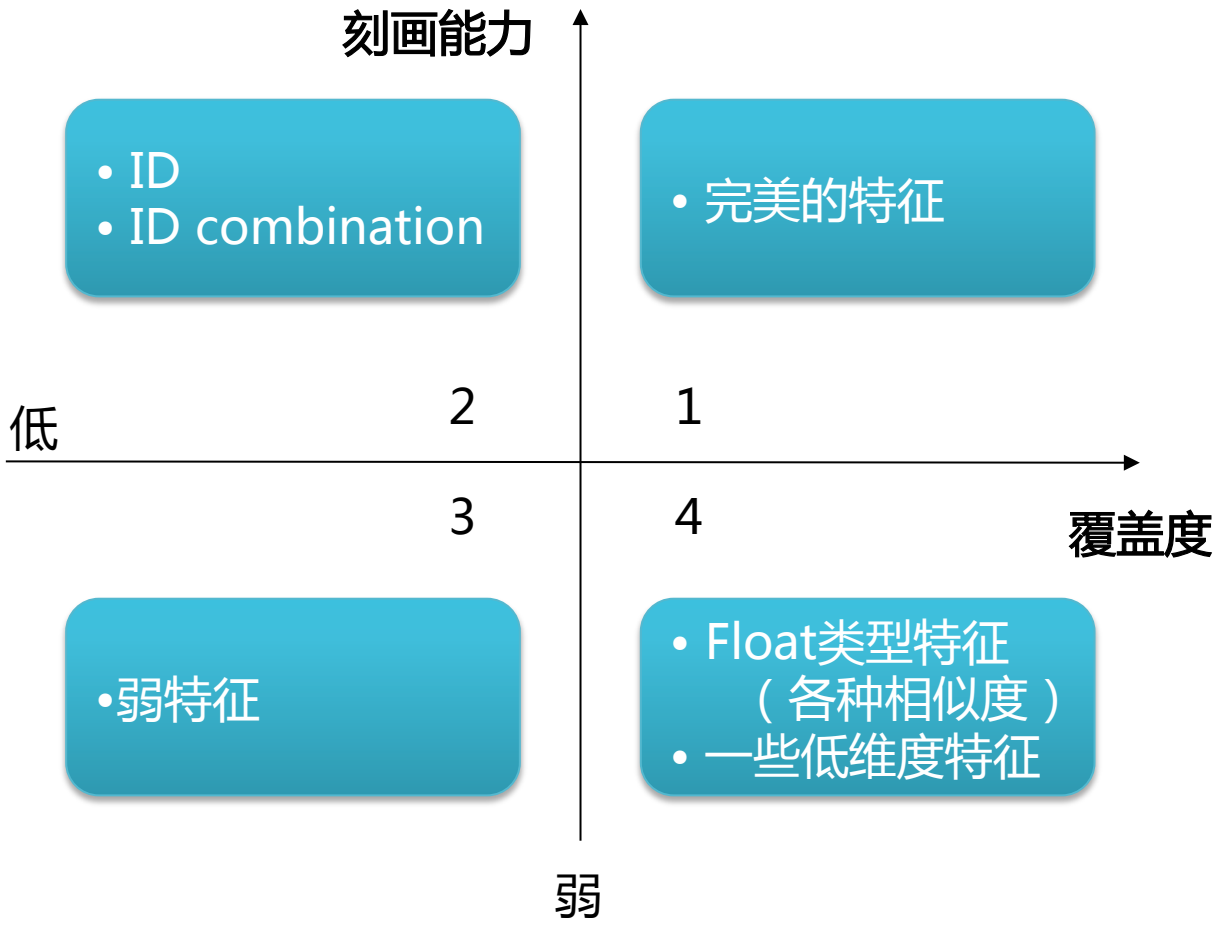
以CTR预估为例:
1，Query：长度、历史CTR；
2，User：年龄、性别、历史CTR；
3，Ad(Ader\BidWord\Title\Desc等)：各种长度、各种历史CTR；

4，Query与Ad的组合：
5，Ad与User的组合；
6，Query与User的组合：

7，Query、Ad、User的组合。

Feature

- 特征设计
  - 刻画能力
  - 覆盖度



Model

Model

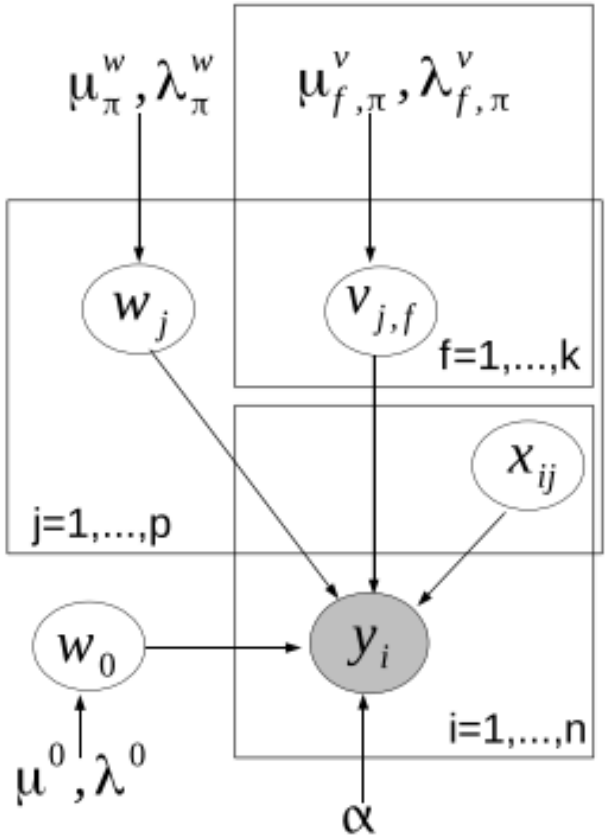
- 模型选择
  - 问题类型
    - 推荐问题(MF, **FM**)
    - 排序问题(**BPR**, Pair-wise, Rank LR)
    - 分类/回归
  - 数据规模(样本、特征)：
    - KDD CUP(十亿维特征，百万级样本）
      - Online Learning
      - L1 > L2
    - WSDM（百万维特征，百万级样本）
      - Rank SVM
      - L2 > L1
  - 评价指标
    - NDCG,AUC (Ranking, Classification)
    - RMSE (Regression)

FM: Factorization Machine

- Model:

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j>i}^{p} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

$$w_0 \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^p, \quad \mathbf{V} \in \mathbb{R}^{p \times k}$$



- 场景: 推荐、回归、分类
- 优化:
  - SGD、ALS、MCMC
- 优点
  - Generalized Model Framework
  - Automatic Feature Combination

FMGI: FM with Group-wise Interaction

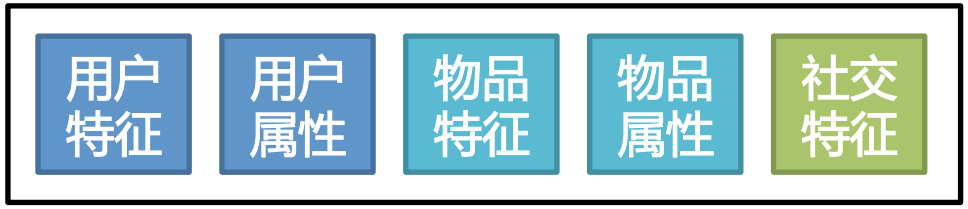
- FM存在的问题
  - 复杂度
  - 精度

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j>i}^{p} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

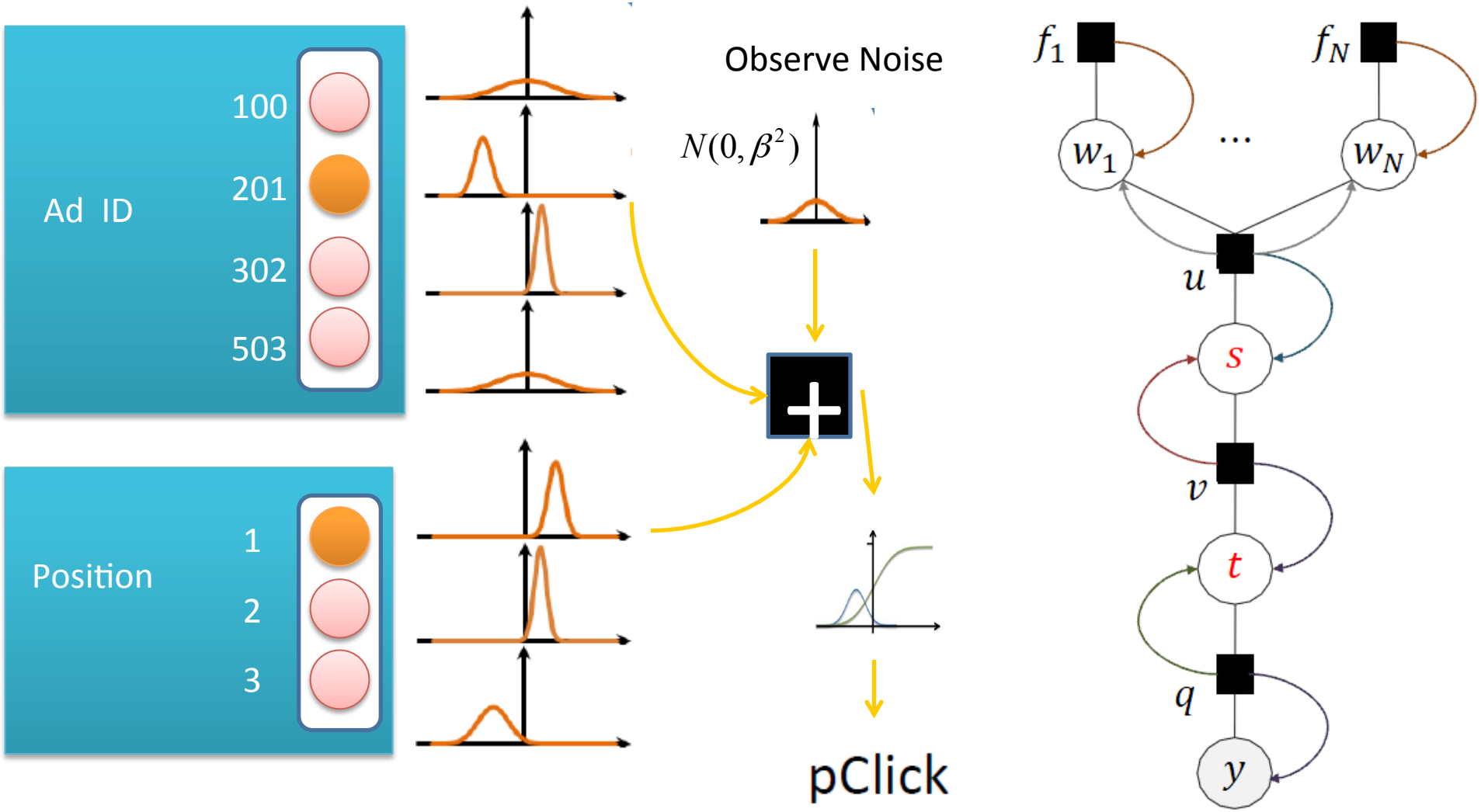
- 模型

$$\widehat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{F} w_i x_i + \sum_{e=1}^{M} \sum_{f=e+1}^{M} \left[I(e,f) \sum_{i \in G_e} \sum_{j \in G_f} x_i x_j \sum_{k=1}^{K} v_i v_j \right]$$

- 优化
  - SGD
  - MCMC



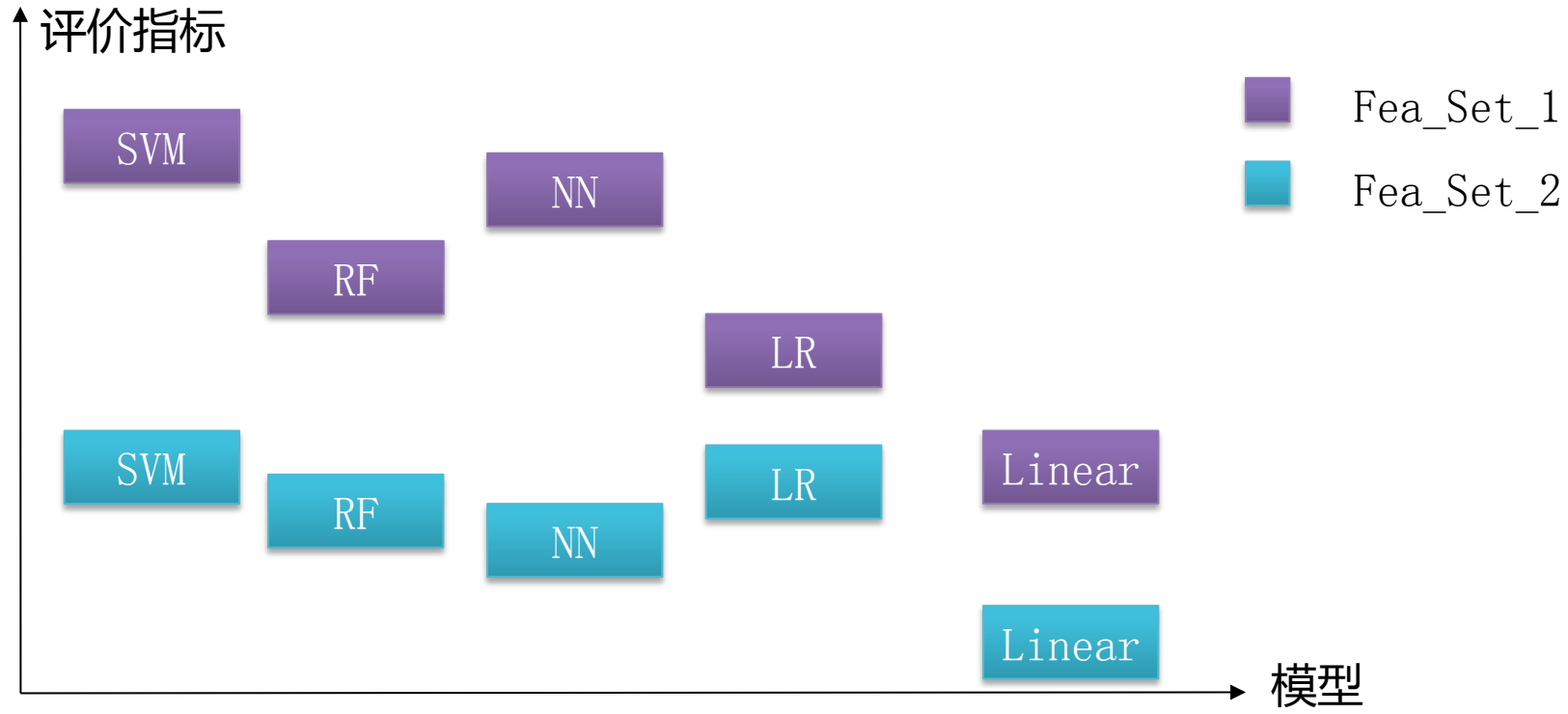
AdPredictor:
Online Bayesian Probit Regression



AdPredictor: Online Bayesian Probit Regression

- 应用场景: CTR/Churn Prediction (Search Ads, RTB)
- 优点
  - Bayesian Model (Easy to add domain knowledge)
  - Easy to parallelize
  - Fast: Online Learning
  - Less Parameters to tune
  - Model Uncertainty Explicitly
  - Natural Exploration
    - Provide a way to add randomness elegantly
- 缺点
  - L2-Norm, Not Sparse (vs. LR-L1)
    - Pruning
  - Poor performance when unbalanced/Rare data without sampling

Model vs. Feature

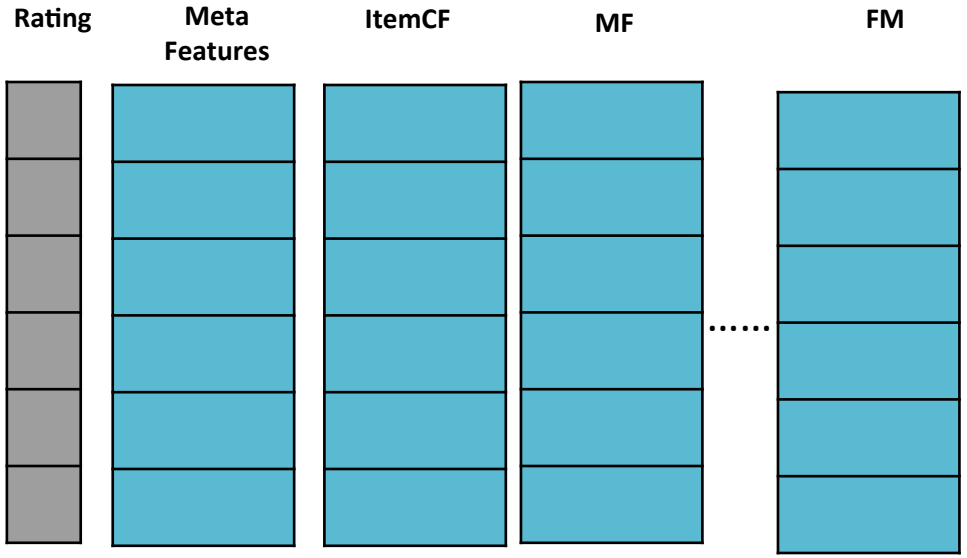


* Feature决定 UpperBound
* Model决定接近UpperBound的程度
* 不同问题下Model的表现是不一样的

Ensemble

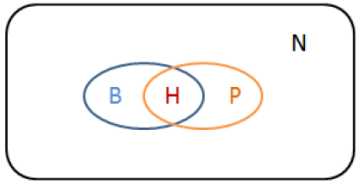
Ensemble

- 方法:
  - Validation Based
  - CV Based



Ensemble

- Diversity



Classification Regression Ranking

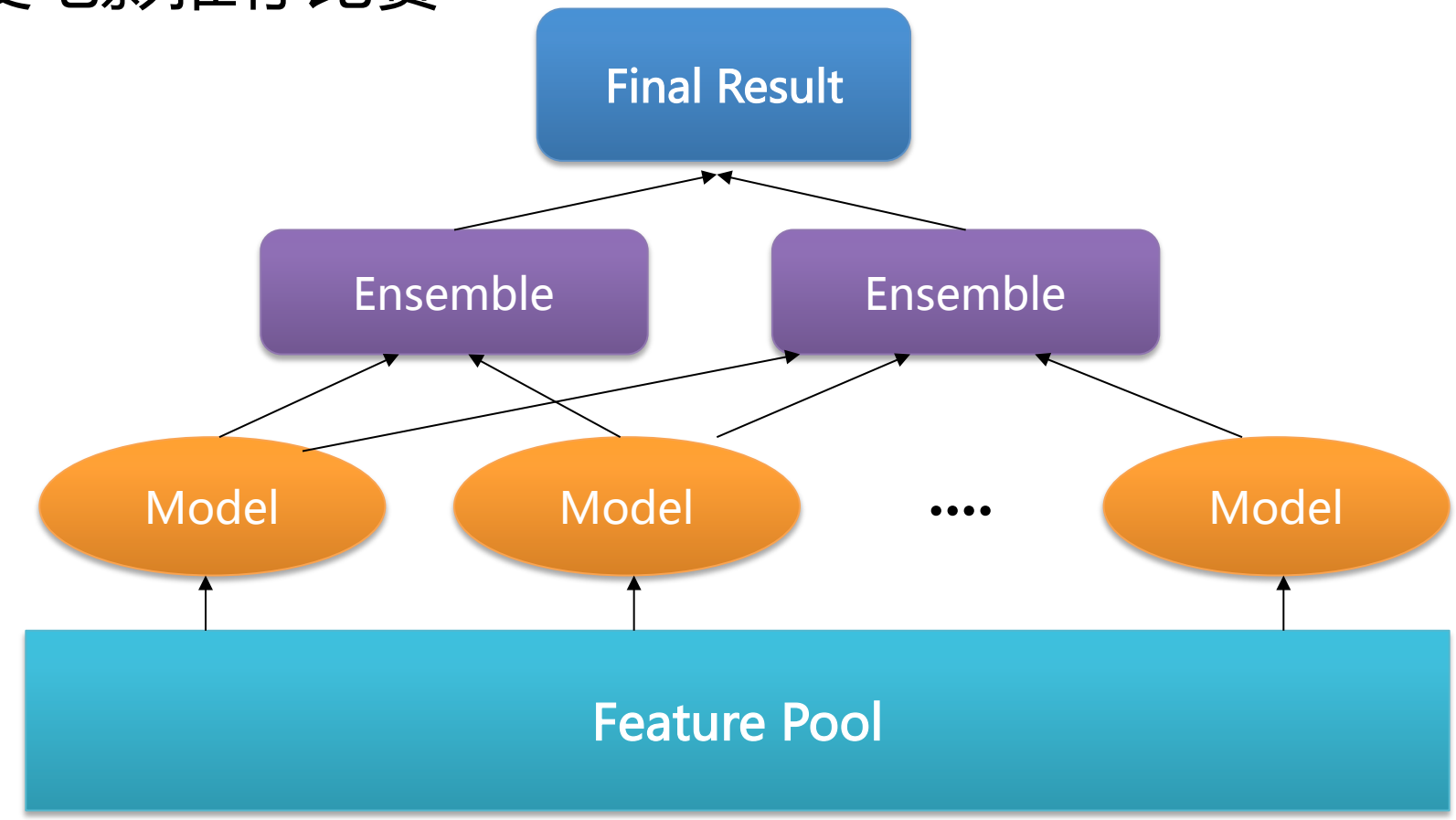
Feature Set Feature Set

Ensemble

- 方法
  - Search Based
    - 参数搜索
  - Learning Based
    - 线性融合
      - 感知机、LR
    - 非线性融合
      - NN, GBDT
    - 基于pair-wise
  - Multi-Stage Ensemble

Ensemble

- 示例
  - 百度电影推荐比赛



总结

- 竞赛 vs. 工业界

	竞赛	工业界
数据	固定，类干净的	流动，非常脏
关注点	特征、模型	数据
模型的重要程度	100%	<<100%
数据集大小	小	大
实时性要求（特征、模型）	基本无	强
评测指标	通常1个，且可以直接优化	通常多个,且不可直接优化

总结

- 竞赛的意义
  - 码农的运动会
  - 接触工业界问题，可以拿到实际数据
  - focus在模型、特征
  - 利于算法的创新、推广，技术的交流

Acknowledgement

- MLRush Team@CAS
- RP Team@baidu
- Liang Xiang@hulu
- Danny Bickson@CMU
- Quan Yuan@taobao

第三期个性化推荐技术周末实战班

- 2014年3月30日开课
  - 上午9点—12点
  - 下午1点--5点半
- 内容:
  - 推荐系统基础
  - 基于投票的推荐算法
  - 基于内容的推荐算法
  - 基于近邻模型的推荐算法设计
  - 基于矩阵分解及隐因子族模型的推荐算法
  - 企业级推荐系统设计和实践

CFP: ACM RecSys 2014 workshop on Large Scale Recommendation Systems (LSRS 2014)

- Tao Ye, tye@pandora.com, Pandora Inc.
- Danny Bickson, bickson@graphlab.com, GraphLab Inc.
- Qiang Yan, yanqiang.yq@taobao.com, Taobao Inc.

We are hiring!

一淘及搜索事业部
技术类 – 搜索与算法职位

描述:
在最具挑战的无线客户端中，从事大数据分析和机器学习、个性化推荐系统算法的研发。包括深度理解用户的Query语义、分析挖掘无线用户时空特征和兴趣偏好、融合PC和无线端数据预测用户行为等。

要求

1、扎实的编程功底，对C/C++/Java/Python等主流语言至少精通一门，熟悉2门；
2、在推荐系统、自然语言处理、搜索相关性、排序模型中的一方面有较深入的动手实践经验
3. 有责任心、对技术有热情、团队合作精神佳

简历发送到yanqiang.yq@taobao.com

Thanks

竞赛经验分享: 个性化推荐，搜索广告，RTB
@严强Justin