DEECAMP

机器学习实践

-- 通用方法泛讲

燕鹏

2018.07

美团业务

跨平台终端

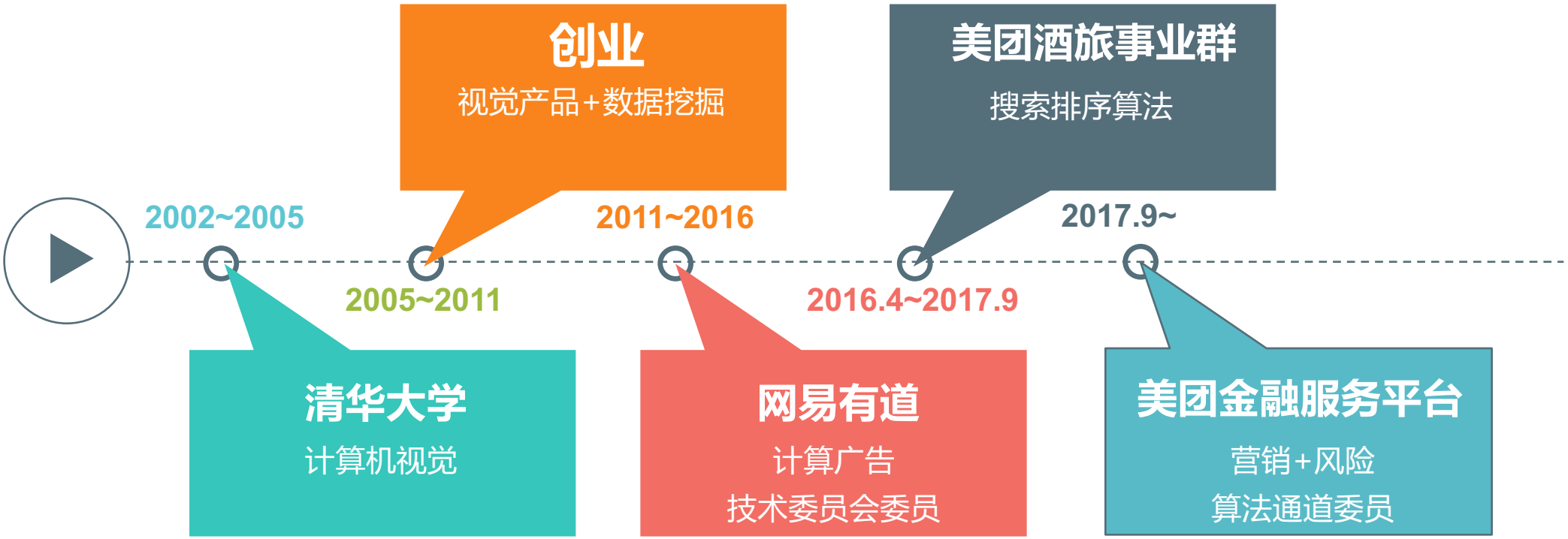
全场景联通

到家 | 到店 | 出行 | 旅行

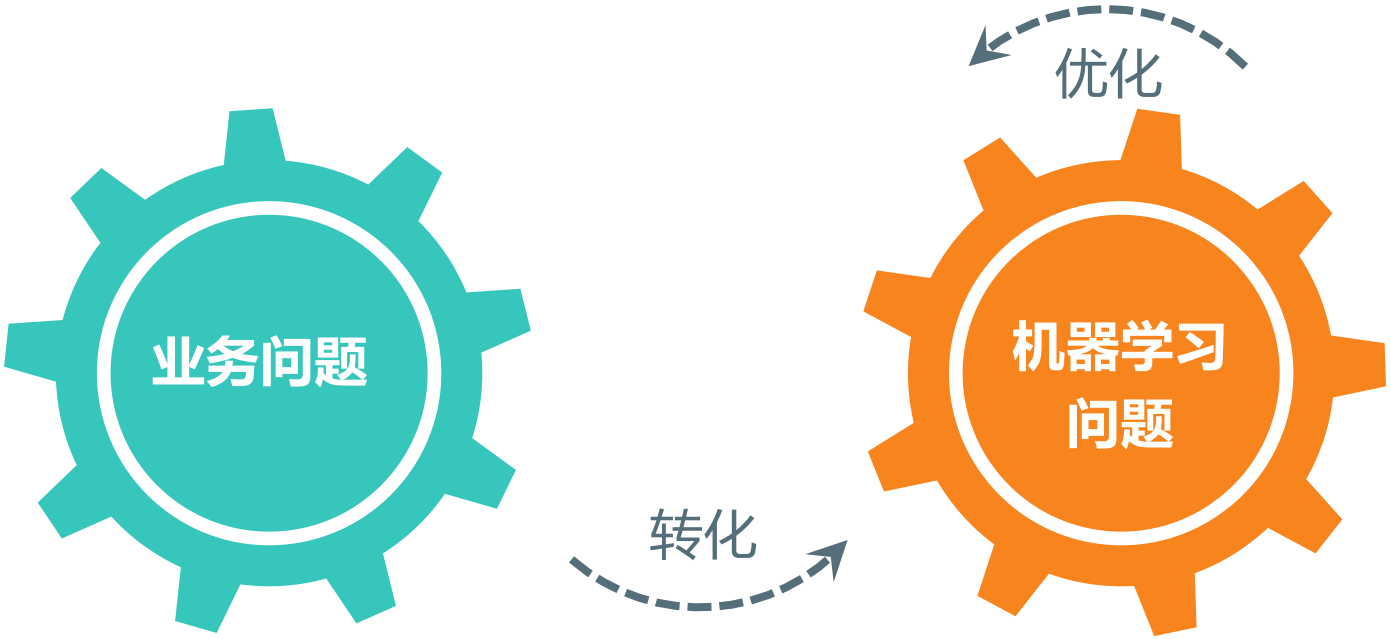
多业务引擎

餐饮 酒旅 电影 休闲娱乐 打车 美容美发.....

个人介绍


创业
视觉产品+数据挖掘
美团酒旅事业群
搜索排序算法
2002~2005
2005~2011
2011~2016
2016.4~2017.9
2017.9~
清华大学
计算机视觉
网易有道
计算广告
技术委员会委员
美团金融服务平台
营销+风险
算法通道委员

两个步骤



四种问题

数据挖掘

应用：排序，经营数据，ETA
数据：结构化数据
信息：不完备

语音

应用：语音识别、语音合成
数据：非结构化
信息：完备

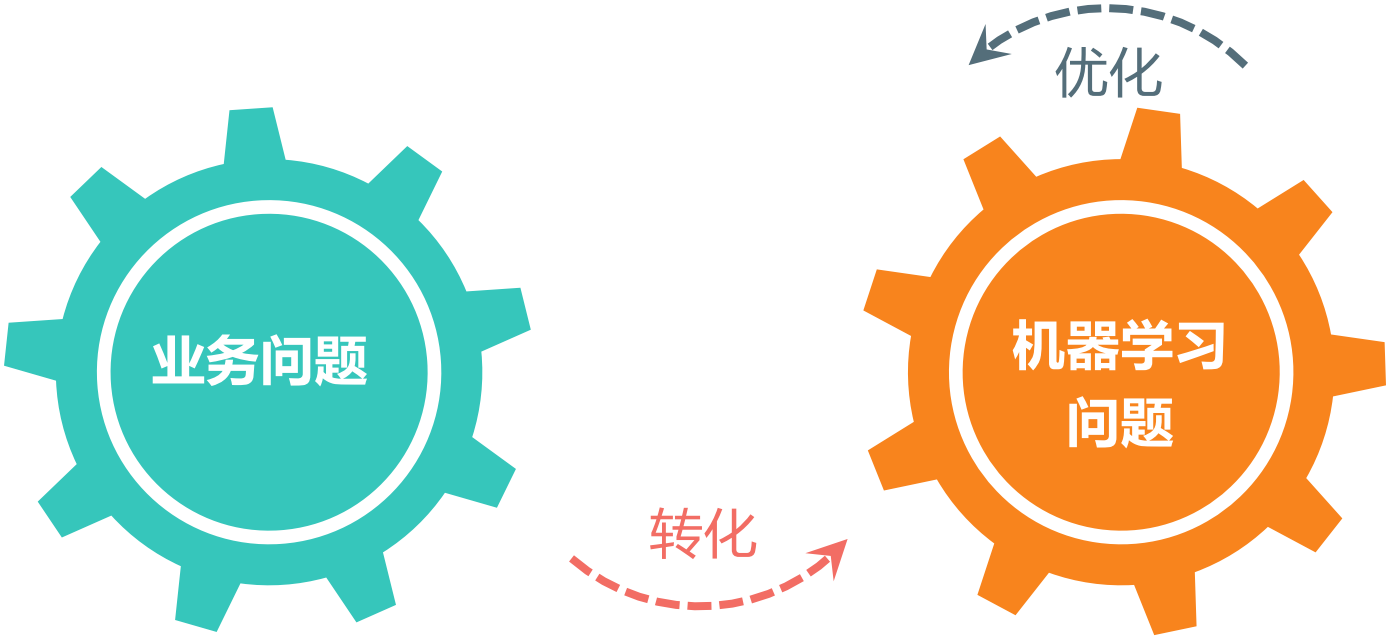
图像

应用：图像分类、分割
数据：非结构化
信息：完备

自然语言

应用：机器翻译、文本分类
数据：非结构化
信息：完备

两个步骤



问题转化


业务分析
业务现状
期望收益
方案选择
评价指标
线上评估
线下评估
数据准备
信息完备
样本分布
标注
模型需求
更新方式
线上性能
模型大小

业务分析

01 **业务现状**

新业务 VS 老业务

02 **业务目标**

开源（用户，收入，利润），节流（人力，财力）

03 **方案评估**

机器学习不是唯一方法！

数据准备

01

样本

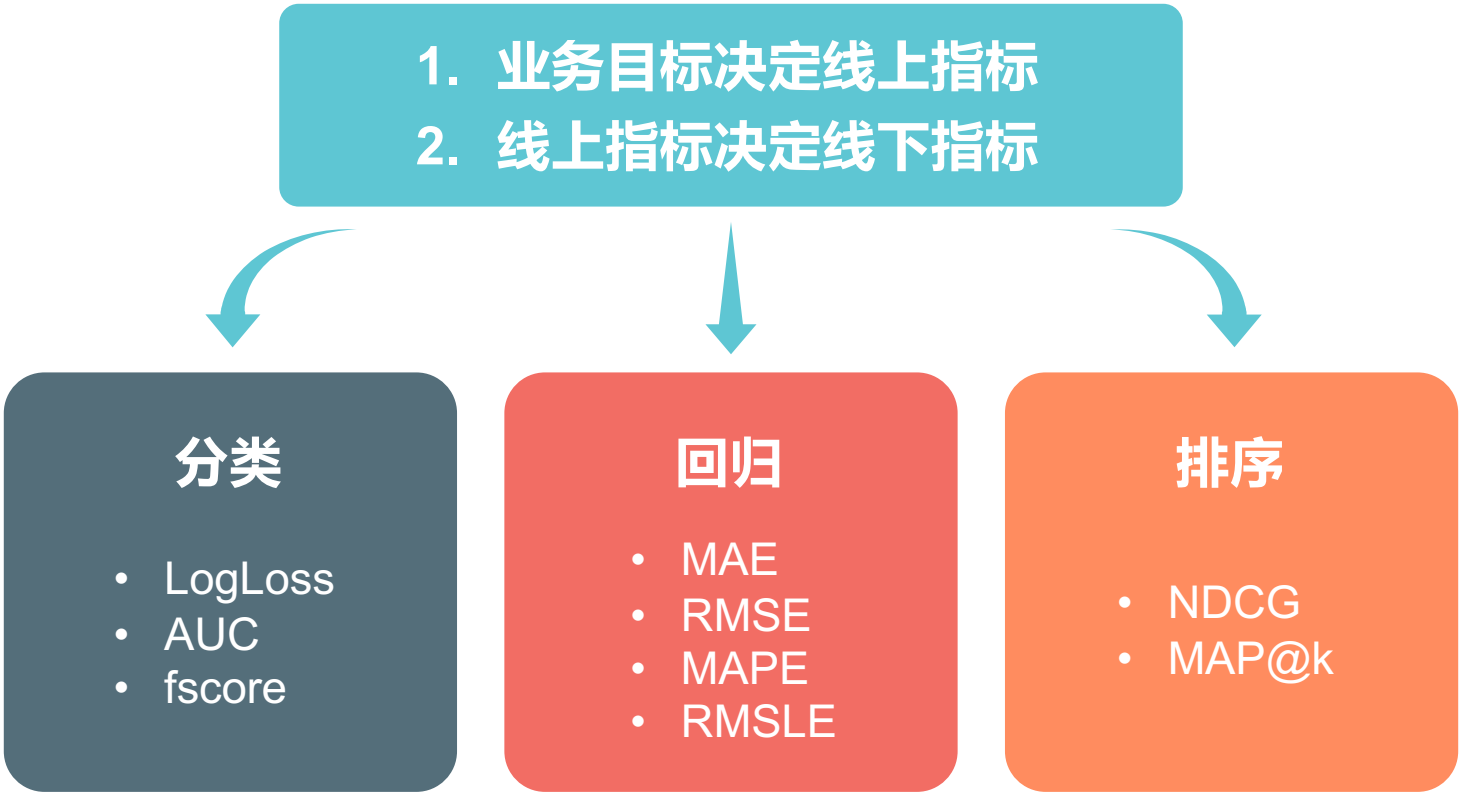
- 时间累积：电商双十一，新业务
- 外部数据：金融信用评级

02

标注

- 自动标注：广告点击率，转化率
- 人工标注：搜索排序，人脸特征点
- 被动标注：金融反欺诈

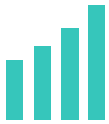
评价指标


1. 业务目标决定线上指标
2. 线上指标决定线下指标
分类
• LogLoss
• AUC
• fscore
回归
• MAE
• RMSE
• MAPE
• RMSLE
排序
• NDCG
• MAP@k

模型需求

更新方式

- 更新方式
- 更新频率



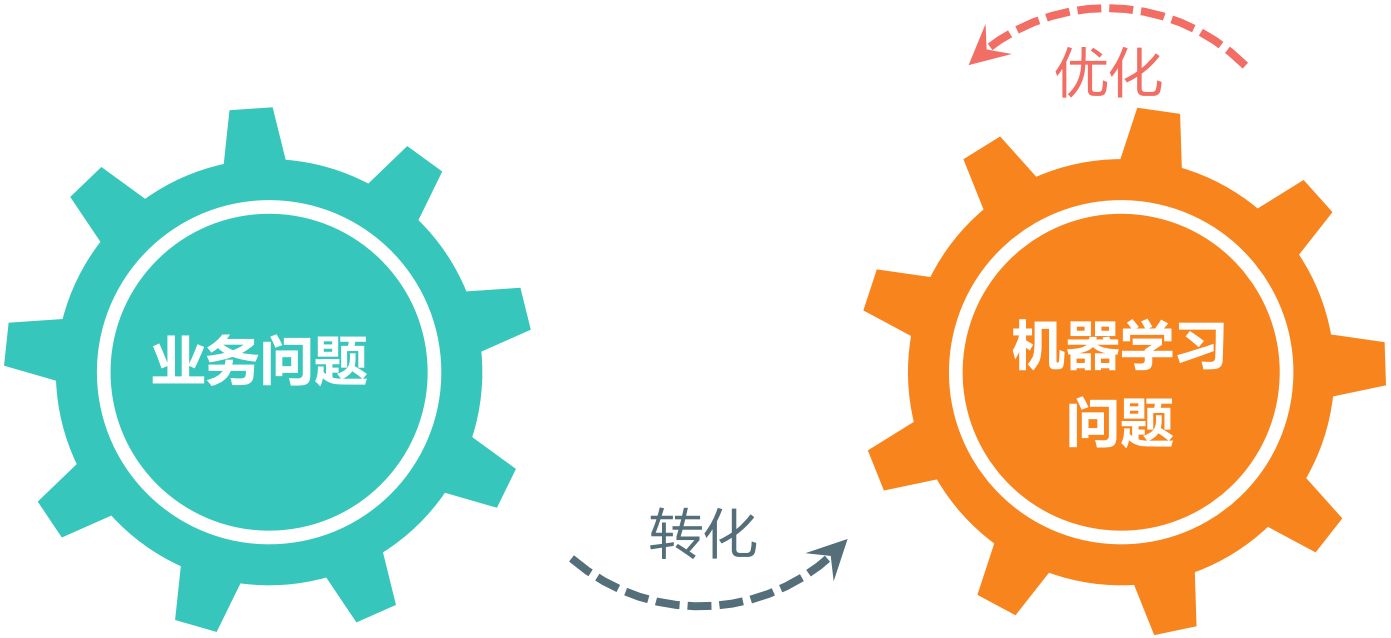
性能要求

- 响应时间
- 开发时间
- 可解释性

模型大小

- 移动端
- 云端
- 芯片

两个步骤



模型优化

01 样本选择

02 特征工程

03 模型选择

04 模型融合

05 模型评估

样本选择

01

样本选择

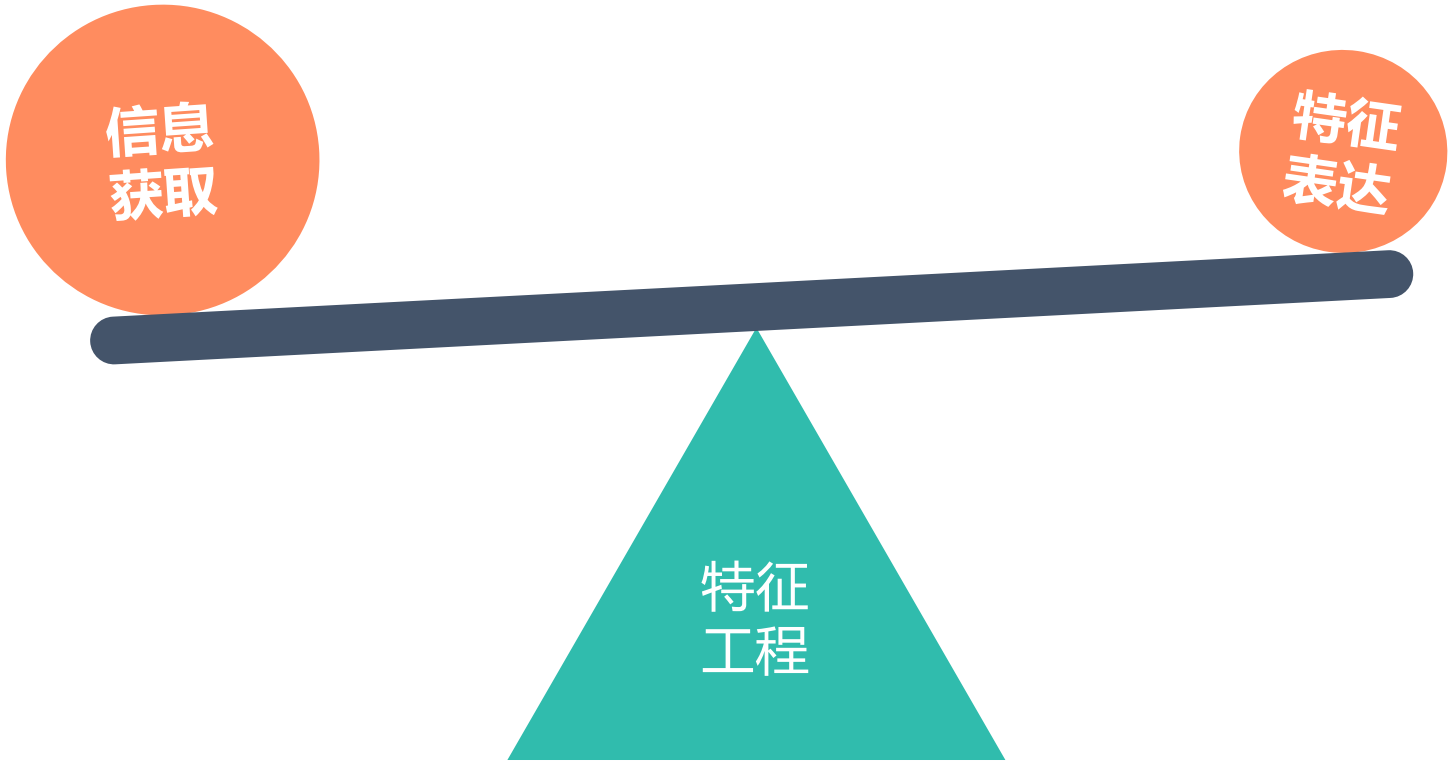
- 去噪声
- 采样：时间窗 vs 样本密度
- 分模型预测：样本分布 vs 样本量

02

样本划分

- 本地验证集+本地测试集
- 时间序列相关：按时间划分
- 时间序列无关：cross validation

特征工程



特征工程

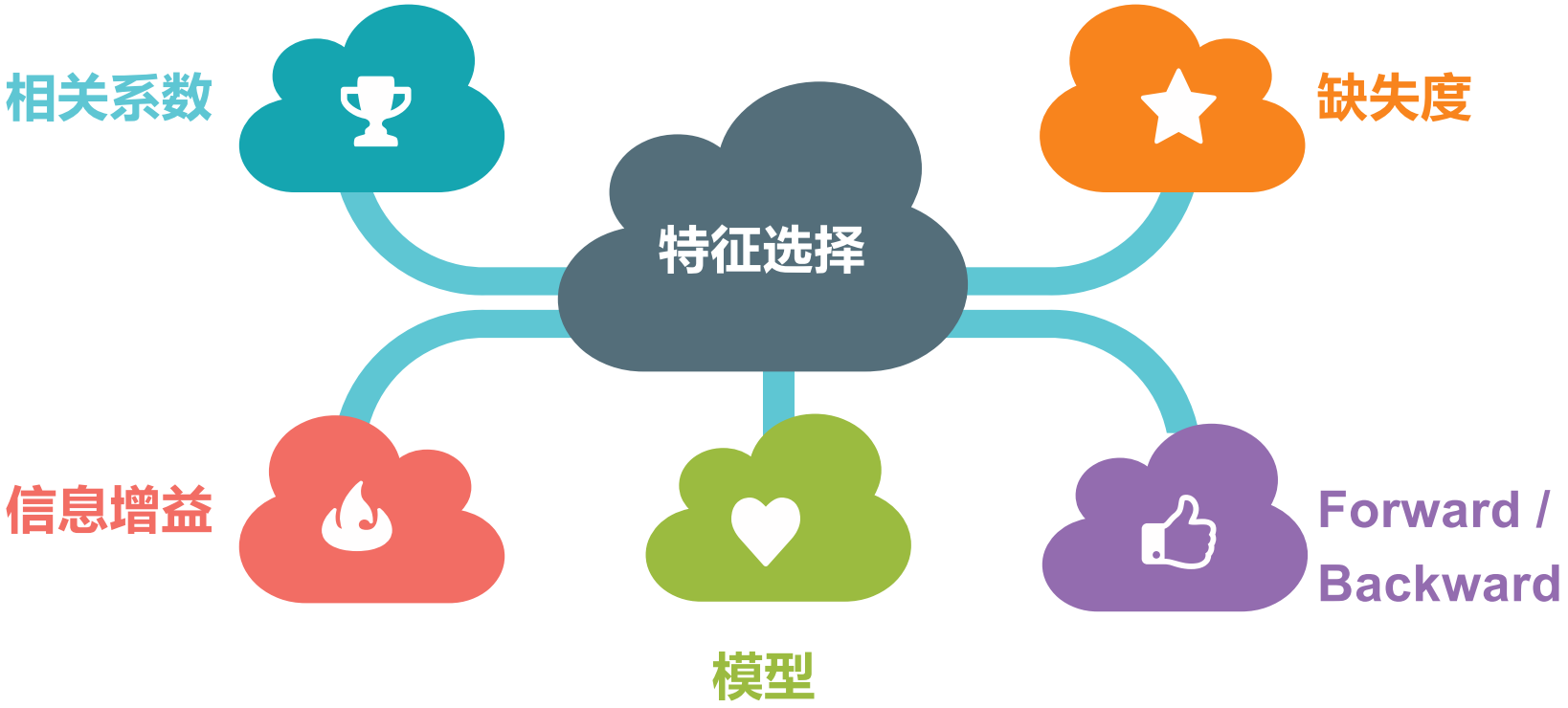
信息获取

- **系统数据**：用户决策所有可能用到的信息
- **外部数据**：覆盖度，互补性
- **原有规则**：规则输出，规则输入

特征表达

- **类别特征**：count，target encoding，one-hot，embedding...
- **连续特征**：min-max，quantile，standard...
- **组合特征**：类别+类别=更细类别，类别+连续=原类别，连续+连续=新连续
- **时间序列**：时间窗 + 统计(min,max,mean,median,std...)
- **其他表达**：聚类、降维...

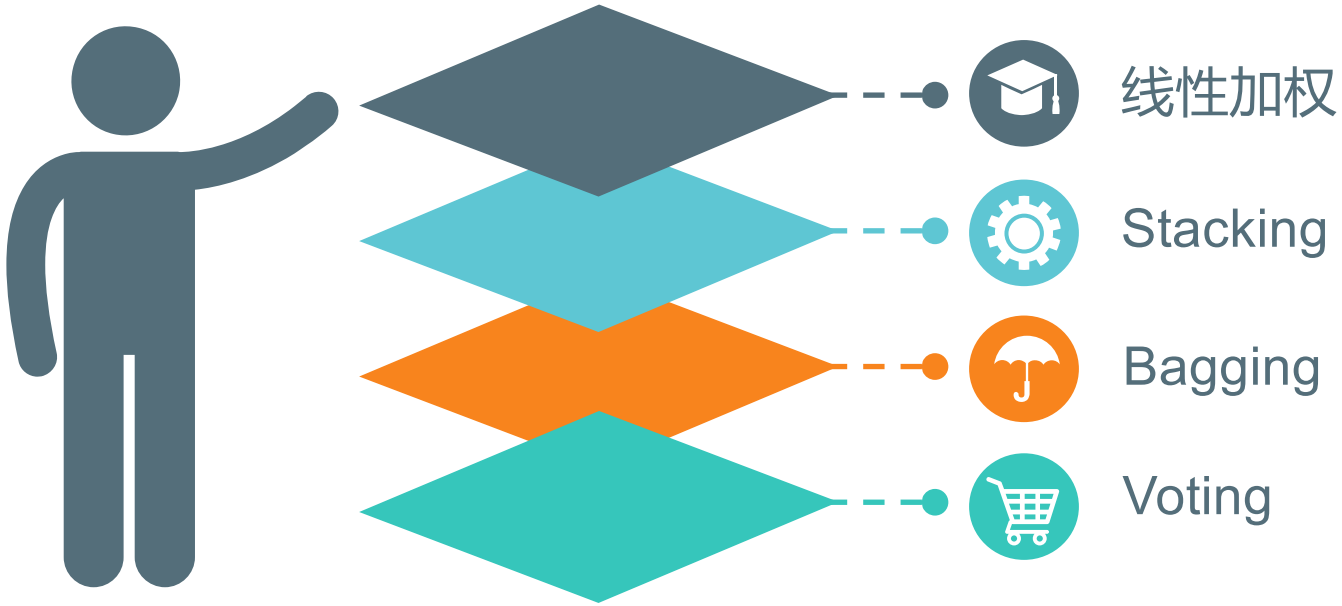
特征选择



模型选择

LR (FTRL)	Liblinear, Vowpal Wabbit
FM	libfm
FFM	libffm
GBDT	xgboost, lightgbm, catboost
NN	tensorflow, mxnet, caffe, kaldi
Others	scikit-learn

模型融合



模型评估

1. 模型改动尽量小
2. 模型稳定性
3. 线下与线上一致性

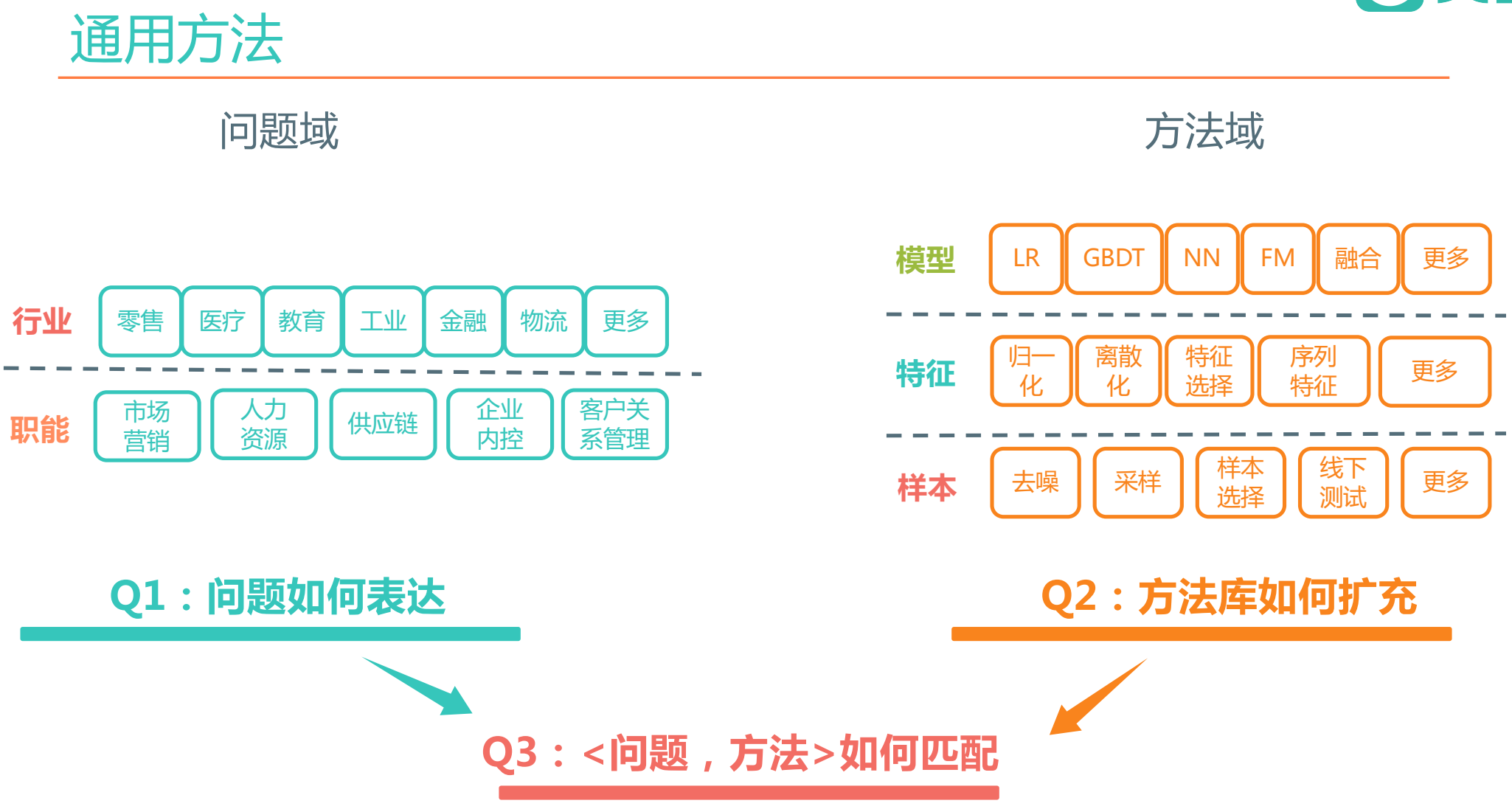
线下异常好

- 特征穿越
- 验证集划分
- 特征引入标注

线上线下不一致

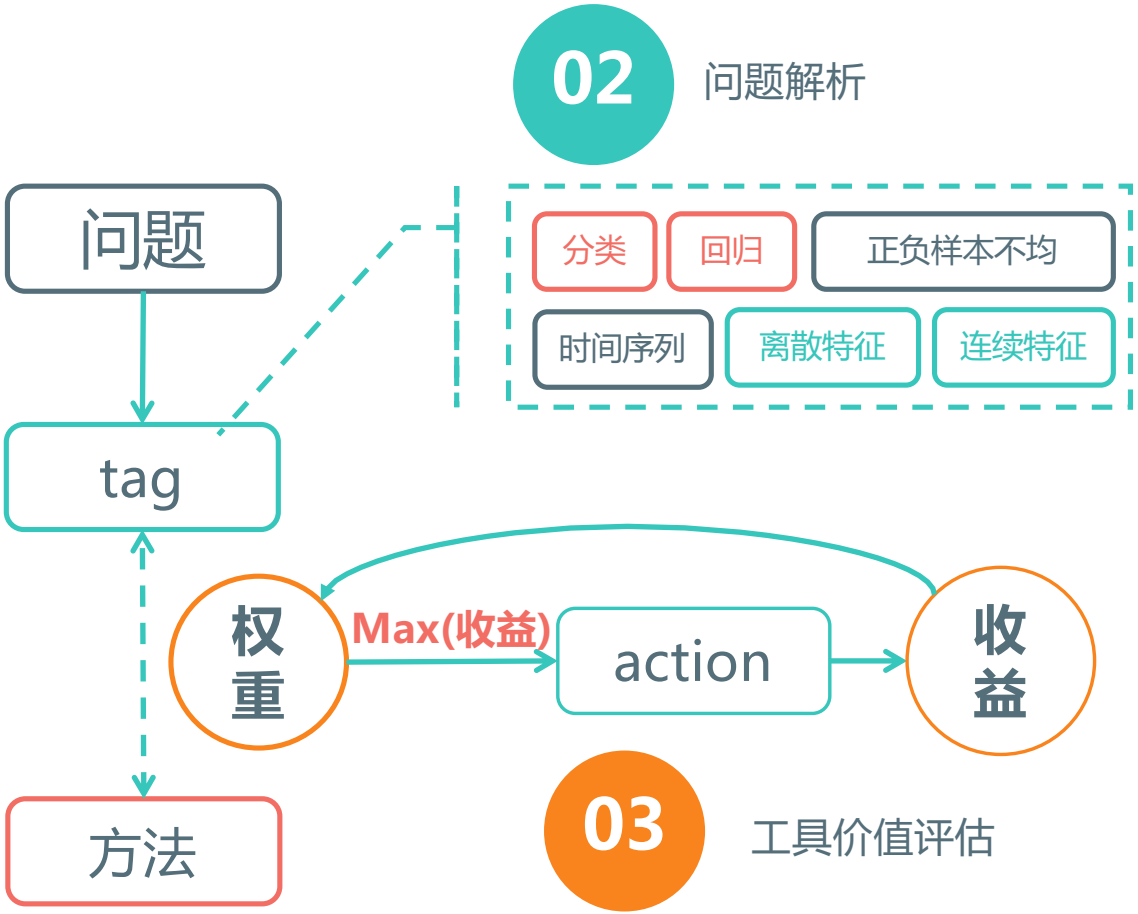
- 特征生产
- 预测模块
- 线上特征变化
- 线下评估指标

通用方法


问题域
方法域
行业
零售
医疗
教育
工业
金融
物流
更多
职能
市场营销
人力资源
供应链
企业内控
客户关系管理
模型
LR
GBDT
NN
FM
融合
更多
特征
归一化
离散化
特征选择
序列特征
更多
样本
去噪
采样
样本选择
线下测试
更多
Q1：问题如何表达
Q2：方法库如何扩充
Q3：<问题，方法>如何匹配

通用方法


01
获取方法
方法
论文
工作
比赛


02
问题解析
问题
分类
回归
正负样本不均
时间序列
离散特征
连续特征
tag
权重
Max(收益)
action
收益
方法
03
工具价值评估

DEECAMP

Thanks!