

大数据机器学习应用架构实战

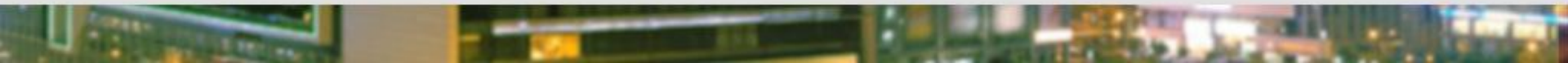
何锐邦 2016.04.23

1

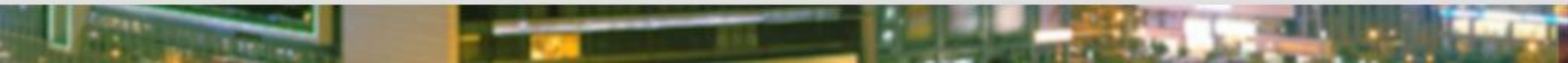
大数据机器学习应用架构

2

机器学习实践中的经验分享



- 大数据+机器学习=智能决策
 - 判断性别
 - 身高、体重、头发长度、三围、腿毛、.....
 - Adaboost
 - 啤酒与尿布
 - 关联规则



- 基本原理

- 每种事物都具有很多特征
 - 身高、体重、头发长度、三围、腿毛、.....

- 特征的分布决定事物的类别

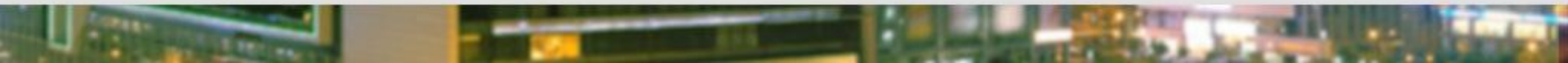
- $a_0x_{00} + a_1x_{01} + a_2x_{02} + \dots + a_nx_{0n} = 0$
- $a_0x_{10} + a_1x_{11} + a_2x_{12} + \dots + a_nx_{1n} = 1$
- $a_0x_{20} + a_1x_{21} + a_2x_{22} + \dots + a_nx_{2n} = 1$
-
- $a_0x_{k0} + a_1x_{k1} + a_2x_{k2} + \dots + a_nx_{kn} = 0$

- 方程规模

- n: 可达亿万级
- k: 可达千万级

- 机器学习目标: 求解特征权重

- 人工猜权重 vs 计算机求解超大规模数学方程

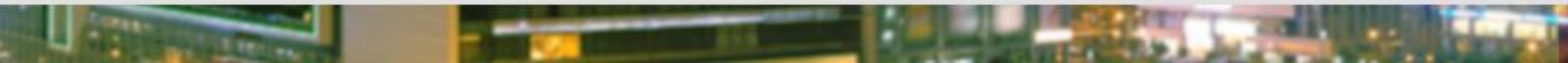


- why机器学习？

- 人工处理的数据量非常有限，不全面；机器可处理的数据量远大于人工，因此考虑更加全面
- 人工总结的规律多数凭感觉，不精确，往往局部最优也达不到；而机器学习算法从数学理论上保证至少能局部最优

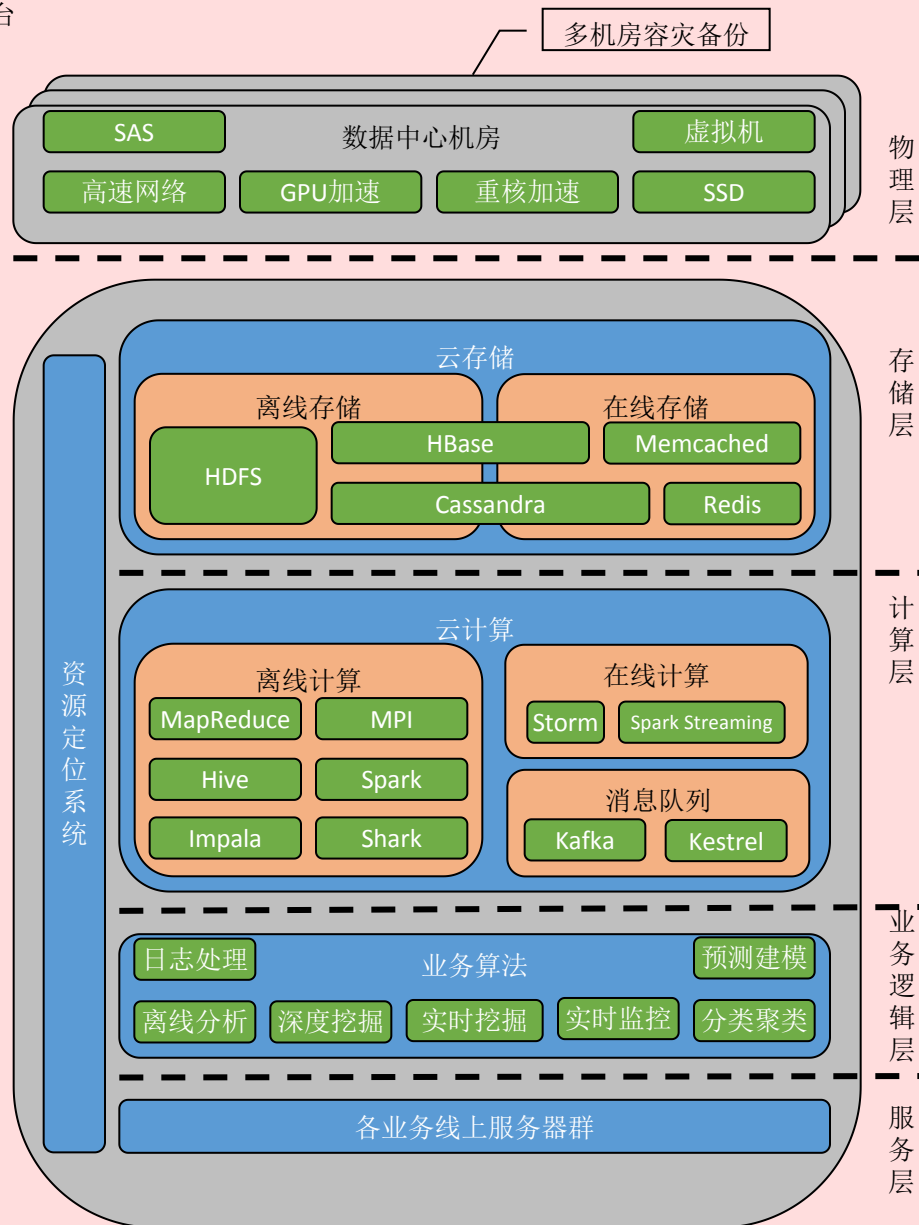
- 机器学习的应用场景

- 搜索
- 广告
- 个性化推荐
- 用户画像
- 安全
- 语音/人脸识别
- 互联网金融
- 下棋——AlphaGo
-



底层平台概览

大数据平台



运维监控

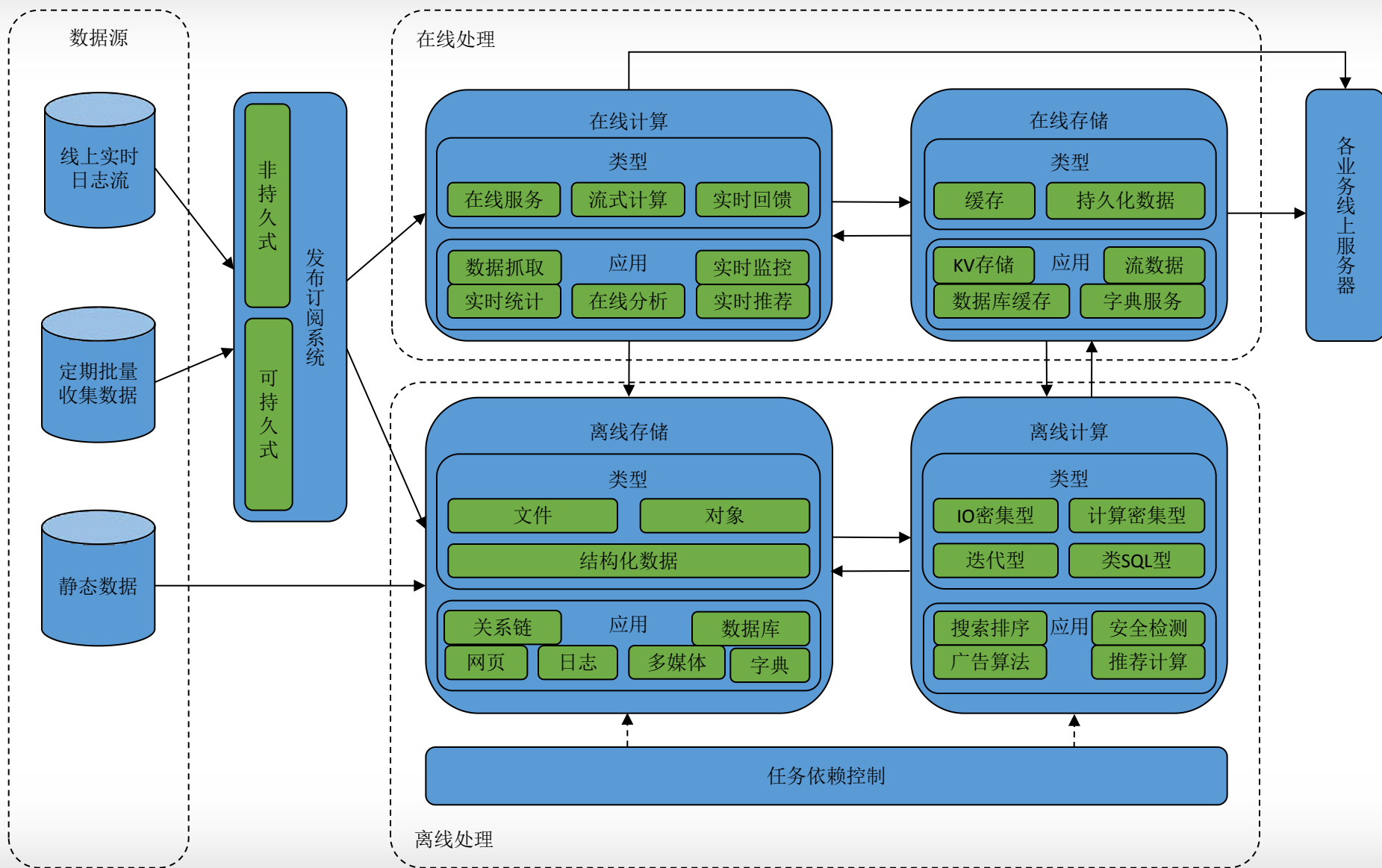
系统状态

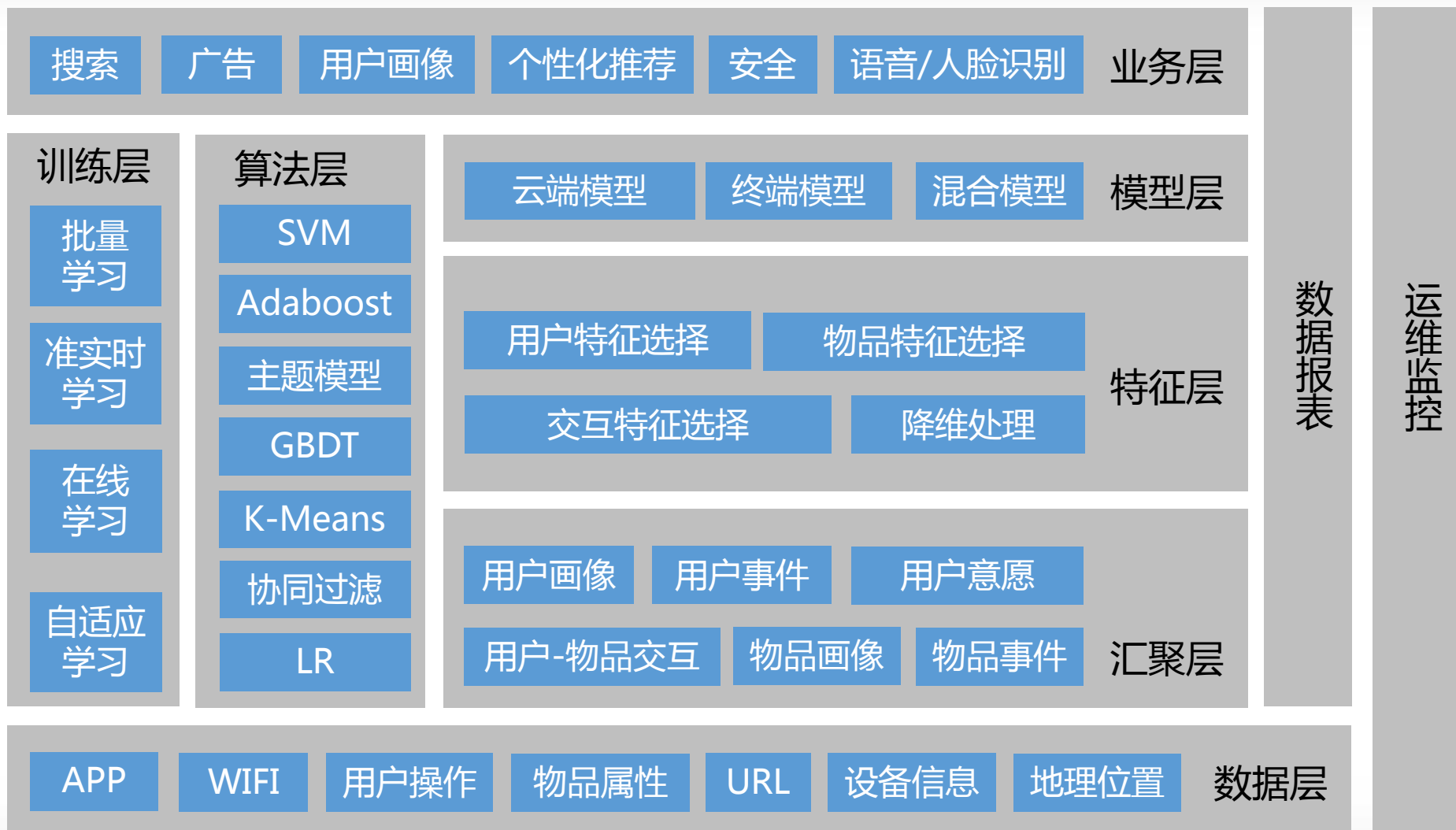
性能分析

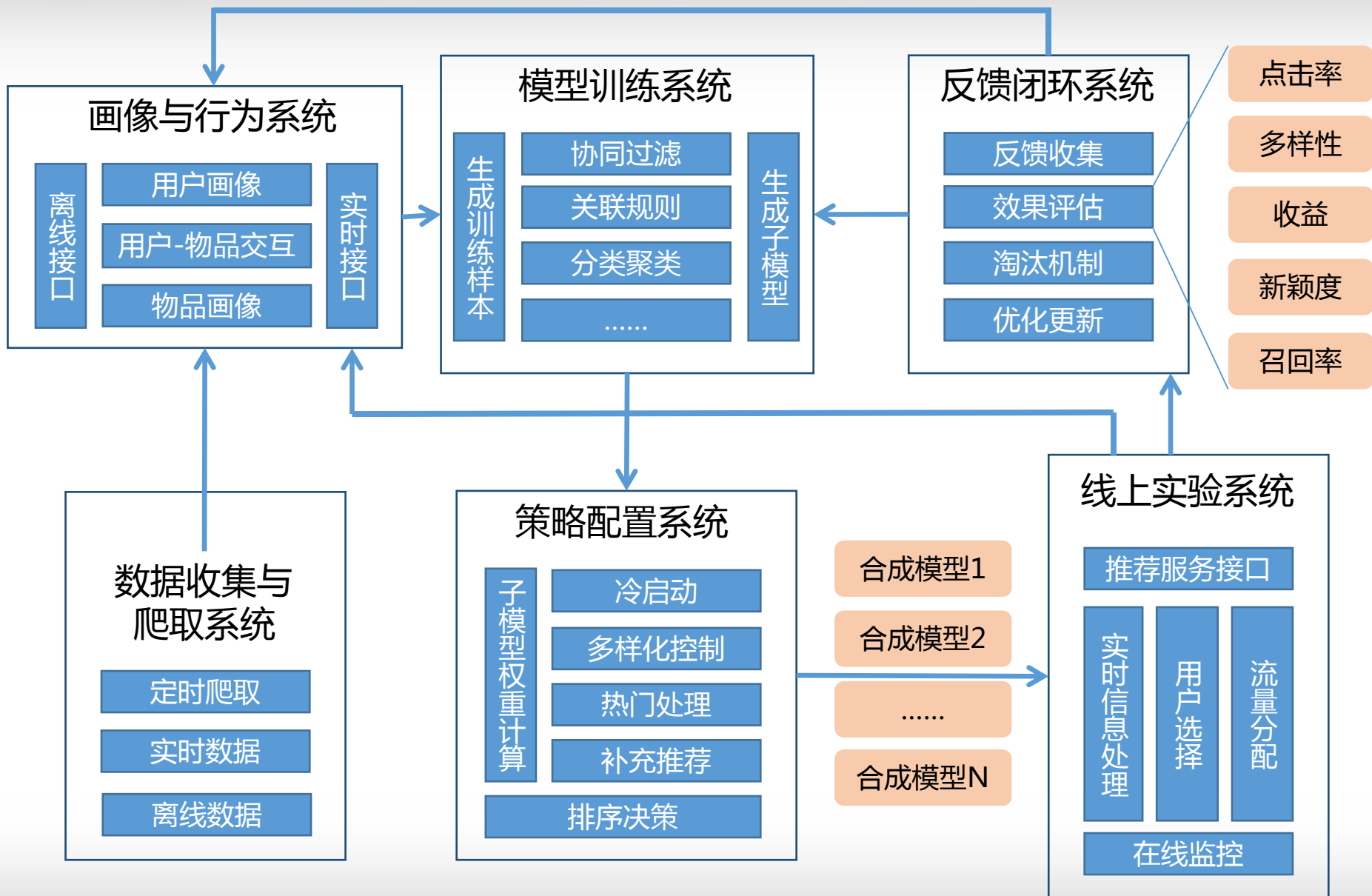
监控报警

自动发布与回滚

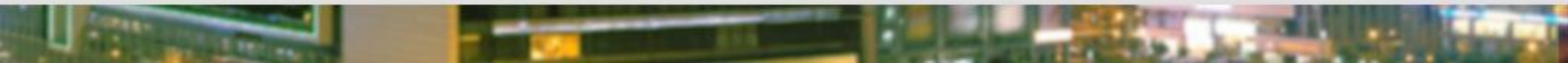
自动监控运维系统

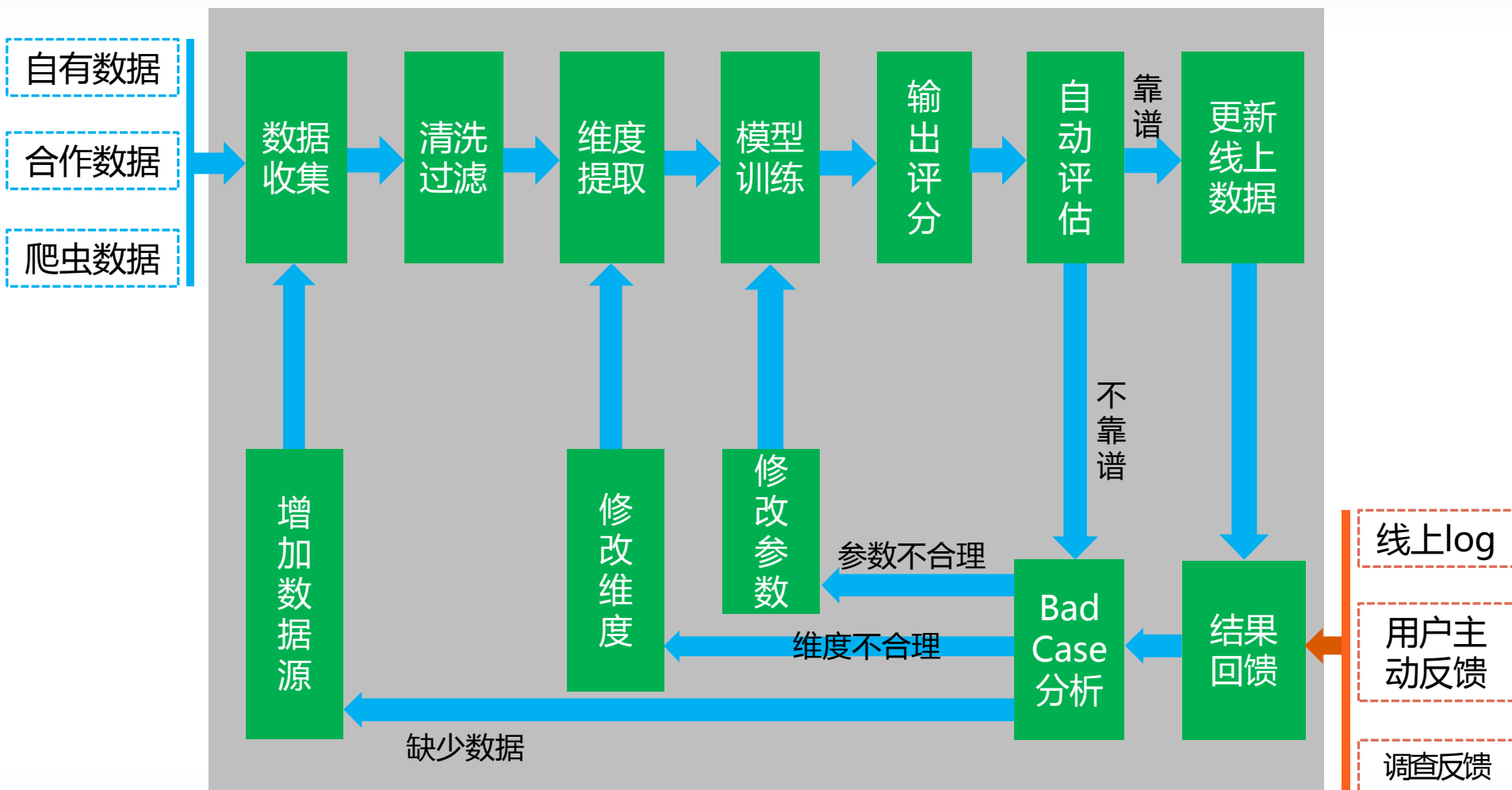




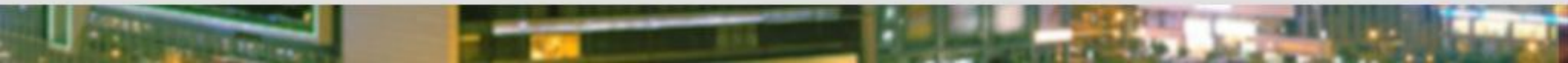


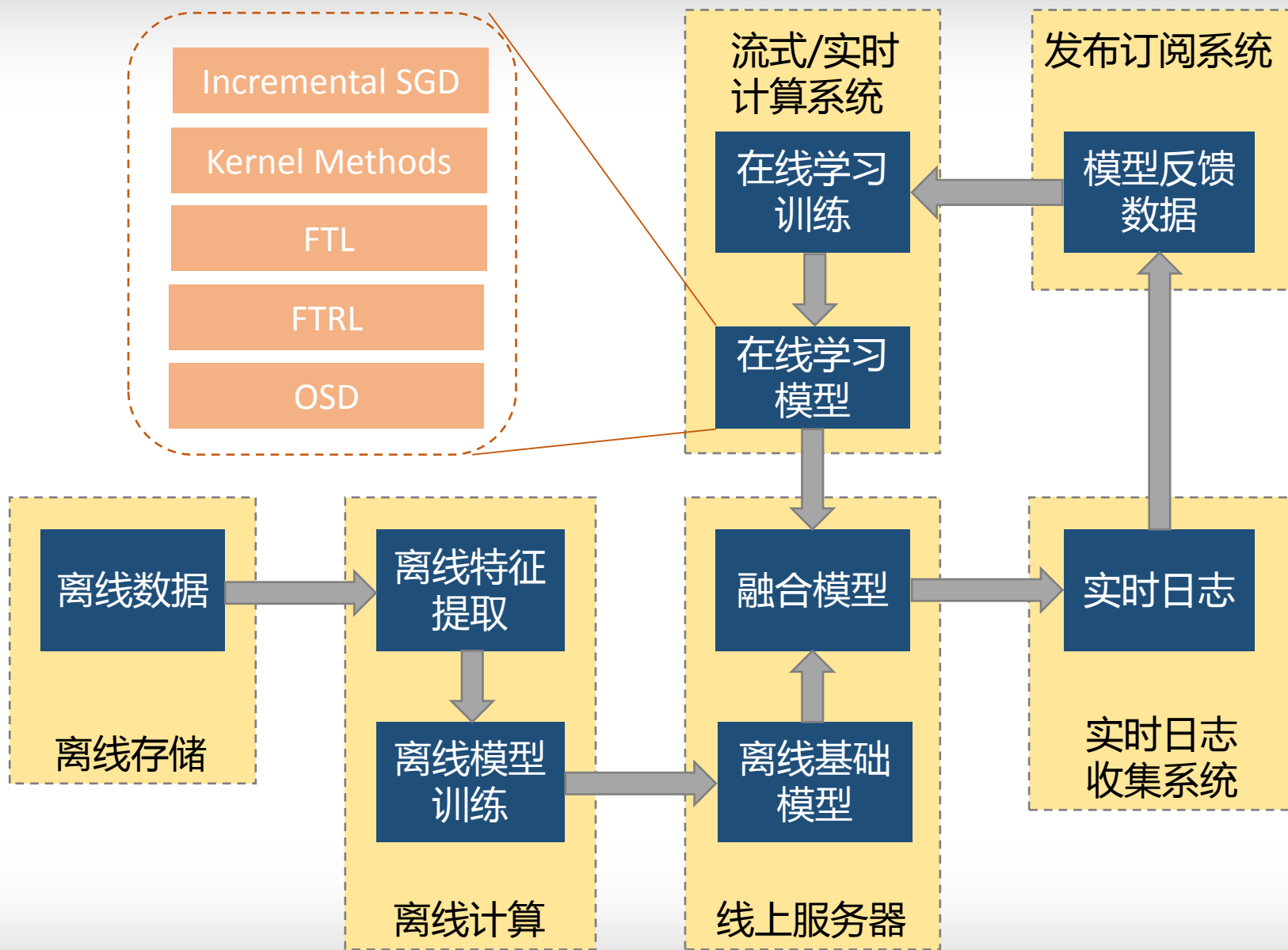
- 闭环回馈的作用
 - 验证模型正确性
 - 修正错误
 - 优化模型





- 批量学习
 - 优点
 - 基于大量数据，模型精确
 - 缺点
 - 滞后性，响应时间慢
- 在线学习
 - 优点
 - 快速响应，即时学习
 - 缺点
 - 有累积偏差



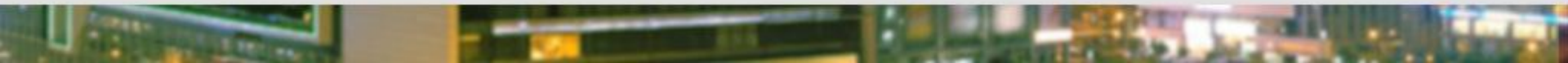


1

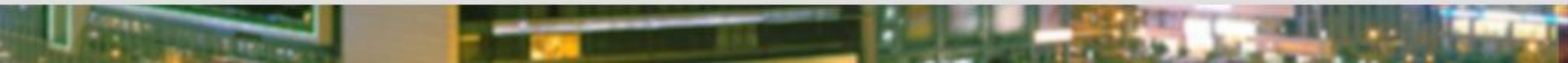
大数据机器学习应用架构

2

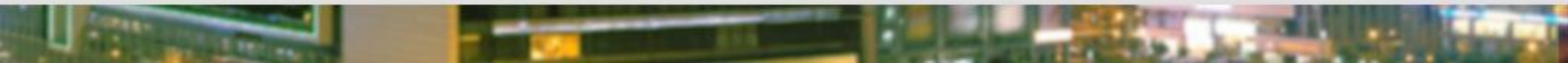
机器学习实践中的经验分享



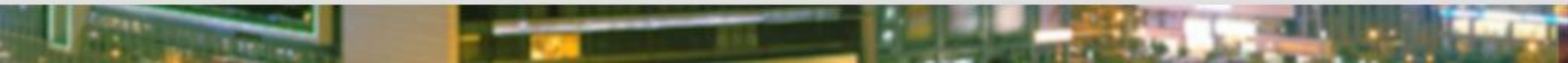
- 分类
 - SVM、Adaboost、Minhash、Simhash、GBDT、随机森林、朴素贝叶斯、KNN、最大熵分类
- 聚类
 - 凝聚层次聚类、K-means、Disjoint Set聚类、Query Clustering
- 自然语言处理
 - PLSA、LDA、N-gram、EBMT、SMT
- 排序
 - 逻辑回归、PageRank、BrowserRank、KNN
- 推荐
 - User-Based协同过滤、Item-Based协同过滤、主题模型、聚类、分类
- 社交网络
 - K-means、KNN、HAC
- deep learning
 - DNN、多层BP神经网络、贝叶斯神经网络



- 关键因素
 - 特征、样本、算法、平台、评估
- 难点1：数据量级
 - 没有现成的标注数据，怎么办？
 - 人工标注得到的数据量太少，怎么办？
 - 数据集终于很大，但还是不全面，怎么办？
 - 数据集较全面了，但计算时间很长，怎么办？

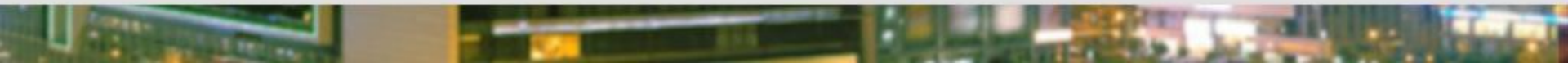


- 难点2：算法使用
 - 算法太多，不知道用哪个
 - 算法计算量大，速度太慢
 - 有些算法天生不能被并行
- 难点3：效果评估
 - 选取什么指标
 - 评估数据来源与覆盖面
 - 线上线下评估结果不一致

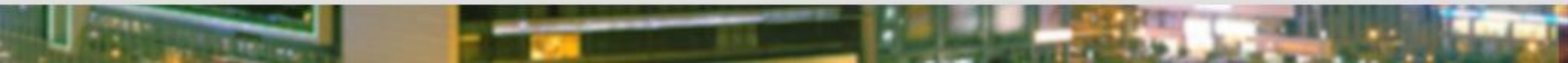


终于解决了 **数据**、**算法**、**评估** 的问题！

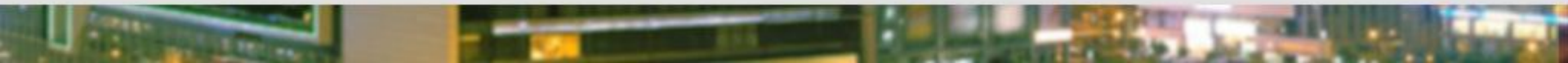
但 跑程序的可是  啊！有木有！



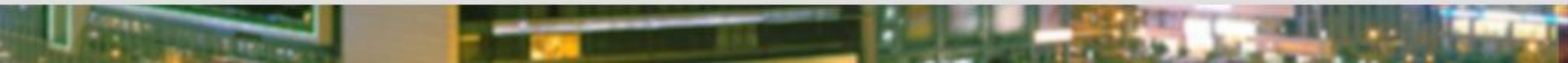
- 难点4：手机终端
 - 资源有限
 - 计算
 - 存储
 - 内存
 - 很多策略不能放在客户端
 - 规则库大小
 - 分词
 - 模型不能太复杂



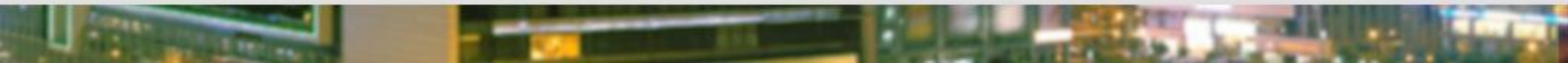
- 可解释性
- 虚假显著特征
- 特征反推
- 过拟合



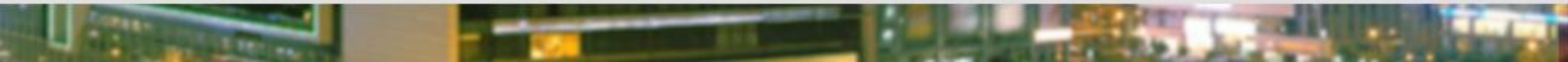
- 只有正样本
- 标注数据少
- 样本不均衡
- 不同数据源有冲突



- 小数据量 vs 大数据量
- 恶意识别 vs 排序
- 单算法 vs 多算法
- 长期爱好 vs 短期兴趣



- 常见问题
 - 多样性
 - 大众化
 - 计算量大
 - 争议性item
 - 推荐新item
 - 用户个人数据缺失
 - 评分数量很少
 - 评分距离不等
 - 时间影响
 - 周期性、季节性、长时期
 - 突发热点
 - 产品认可度改变
 - 用户taste改变
 - 用户情绪改变
 - 用户评分标准改变
 - 账号相同人不同
- 准确率评估



谢谢！
Q & A