

AI科技金融的实践

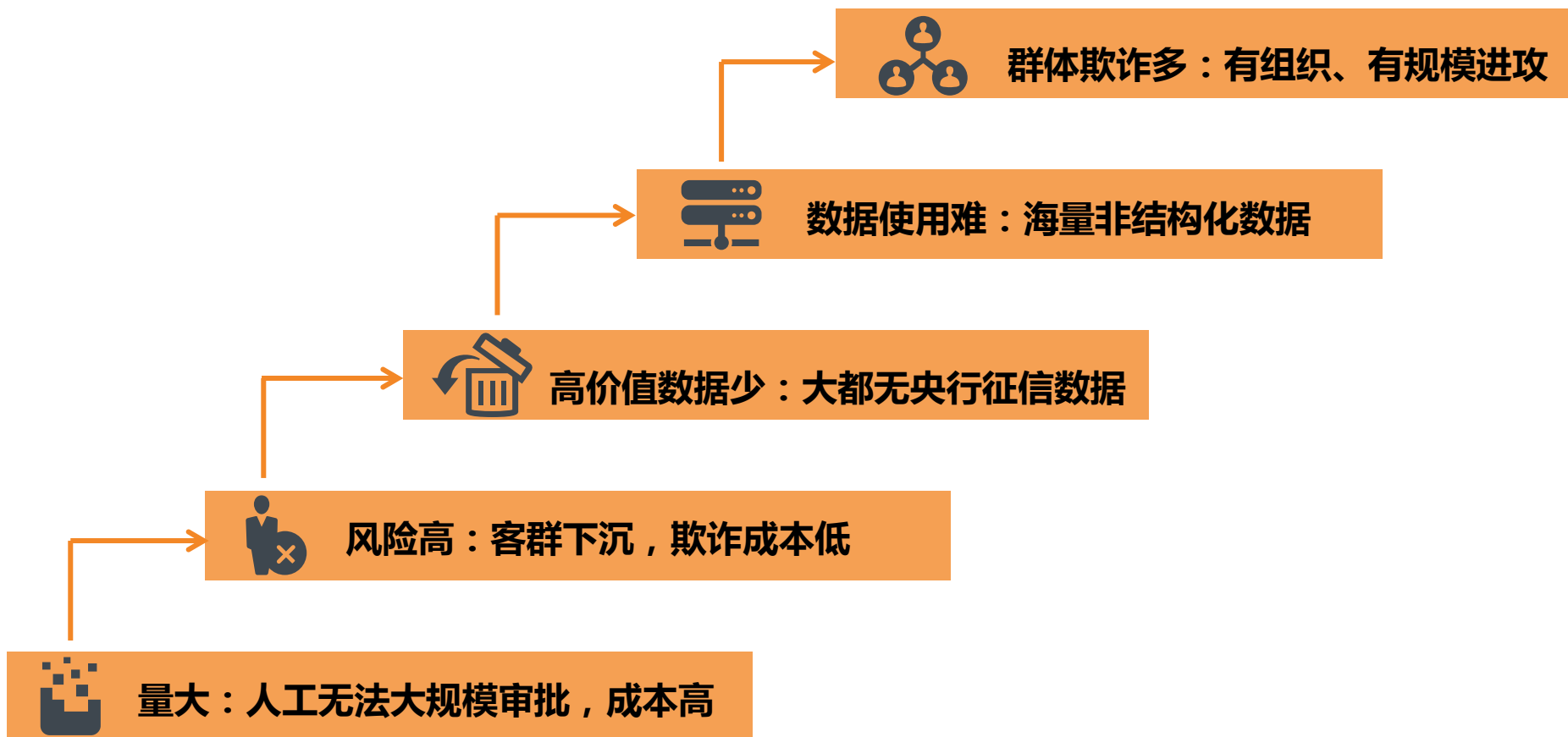


CreditX 氮信

AI应用三要素



新金融业务面临的风控挑战



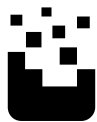
AI解决风控挑战



对AI技术的需求



数据太少



数据太多



半监督学习解决冷启动



深度学习解决特征提取



复杂集成模型解决维度灾难



LIME模型解释器

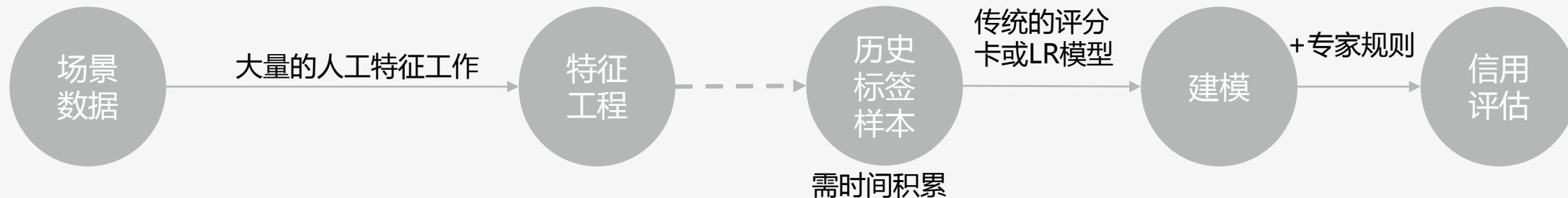


SNA社交网络分析

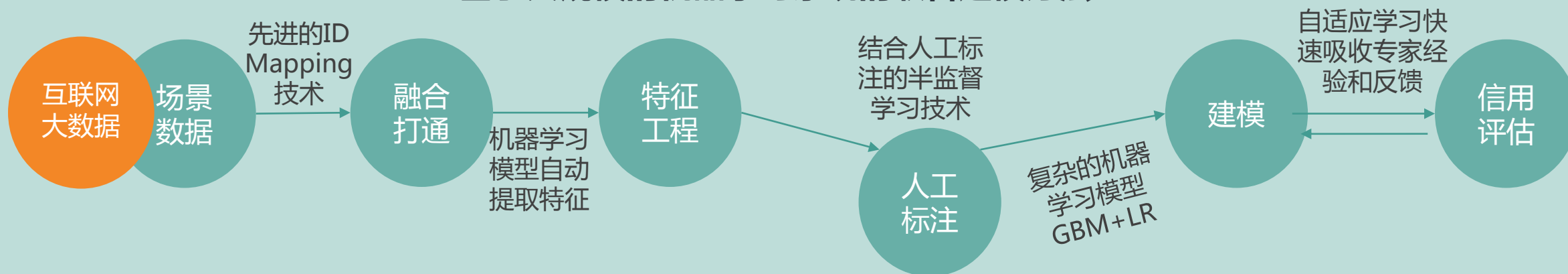
...

互联网级别的机器学习

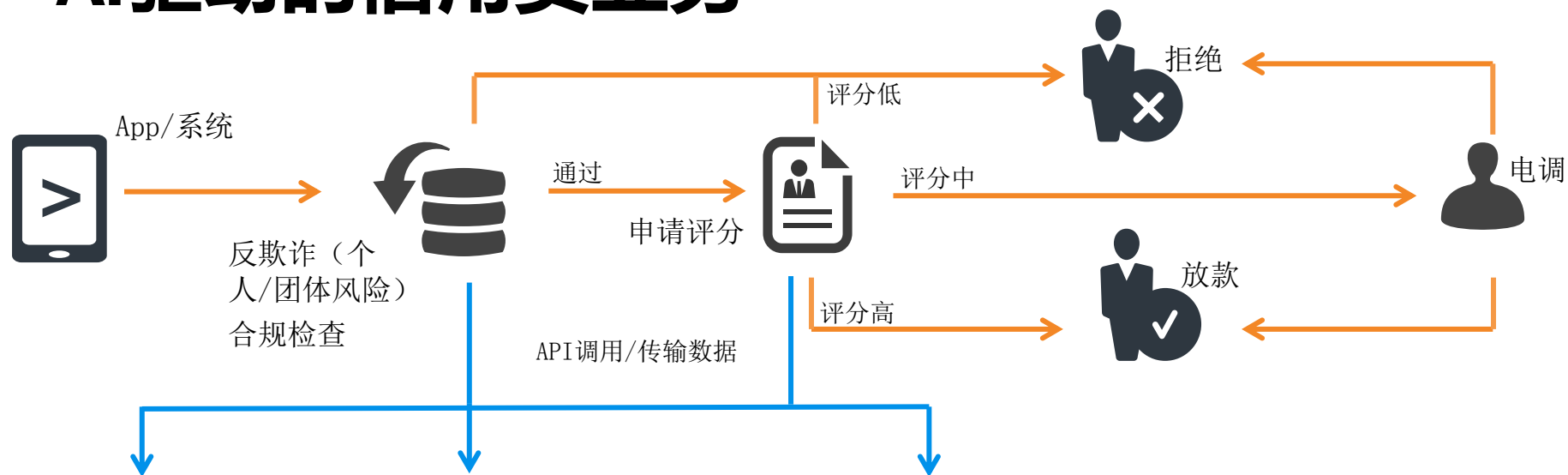
传统信用评估模型方法



基于大规模的机器学习系统的联合建模方法



AI驱动的信信贷业务



机器学习模型与算法



基于网络的反欺诈模型



规则模型



申请评分
违约概率模型



模型快速
自动迭代/
优化

公安户籍数据

第三方学历数据

App相关数据

手机设备相关数据

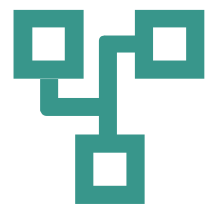
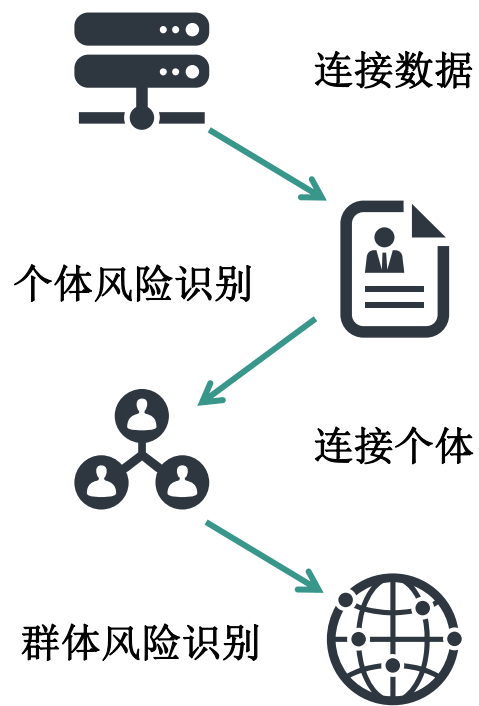
运营商数据

互联网行为数据

申请材料图片数据

个体风险到网络全局风险的识别

基于社会网络分析 (SNA) 算法



分析方法的^{不同}

- > SNA抛开独立性假设，通过图论方法，用中心度、介数、聚集系数等指标在欺诈专有拓扑网络上识别欺诈团伙

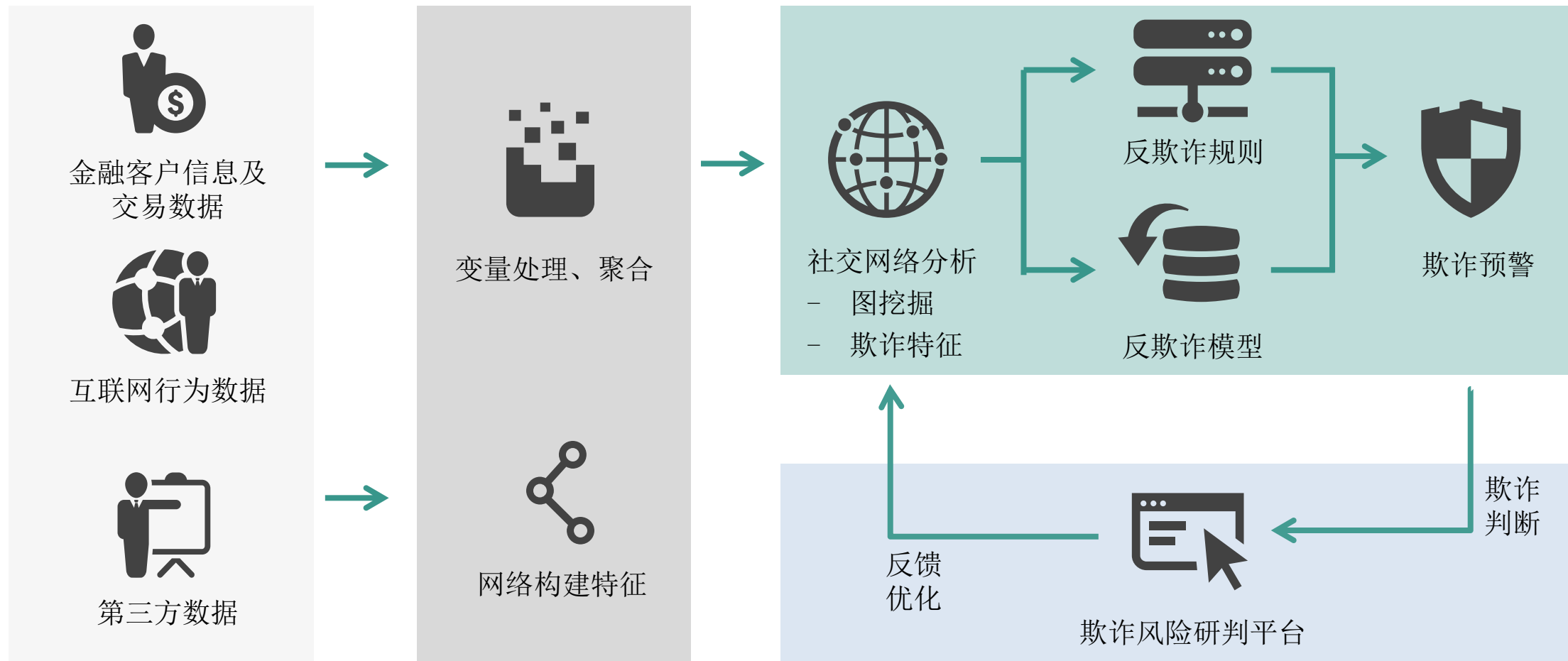


识别模式的^{不同}

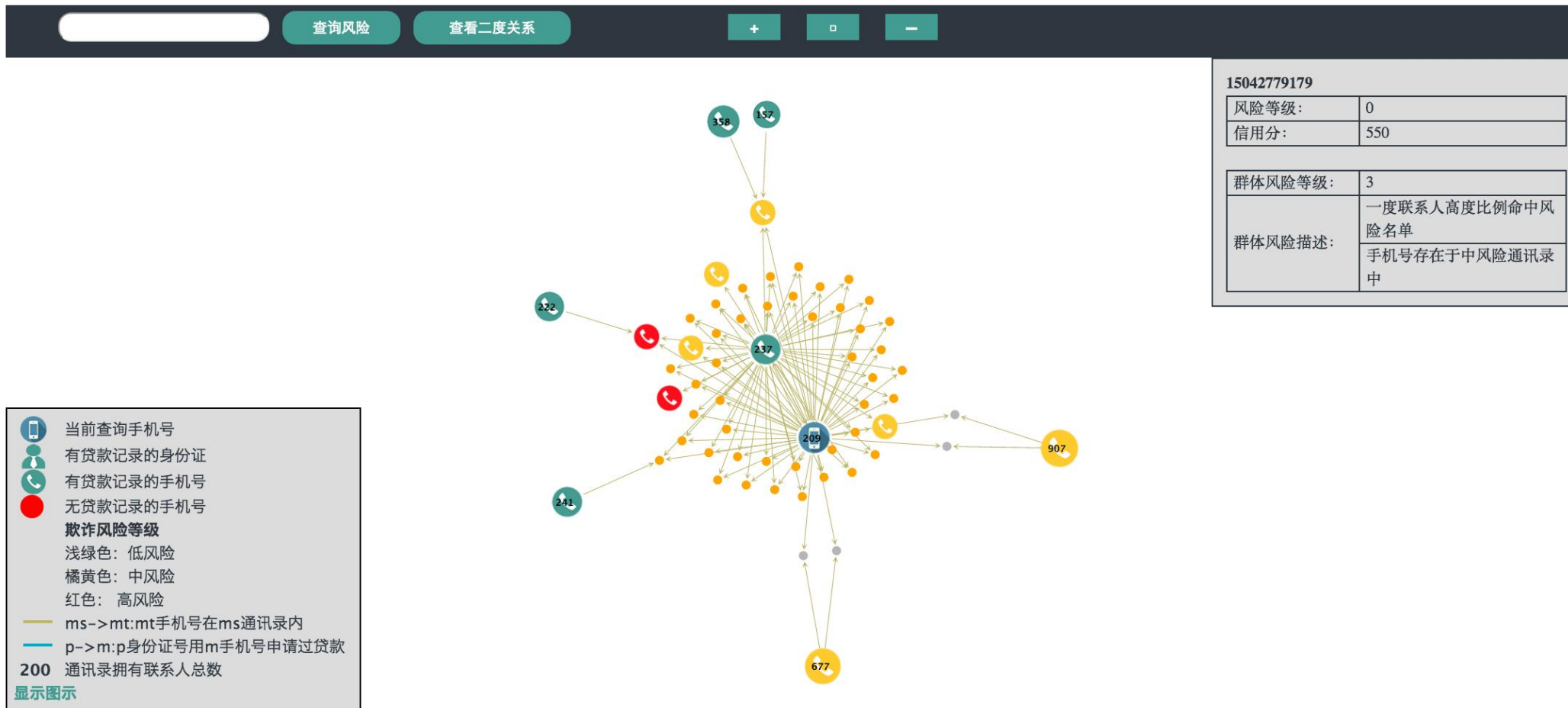
- > 欺诈网络分析最重要的是分析其与邻近节点的关系，从关系的角度把具有相同行为的节点聚合在一起

技术实现方案

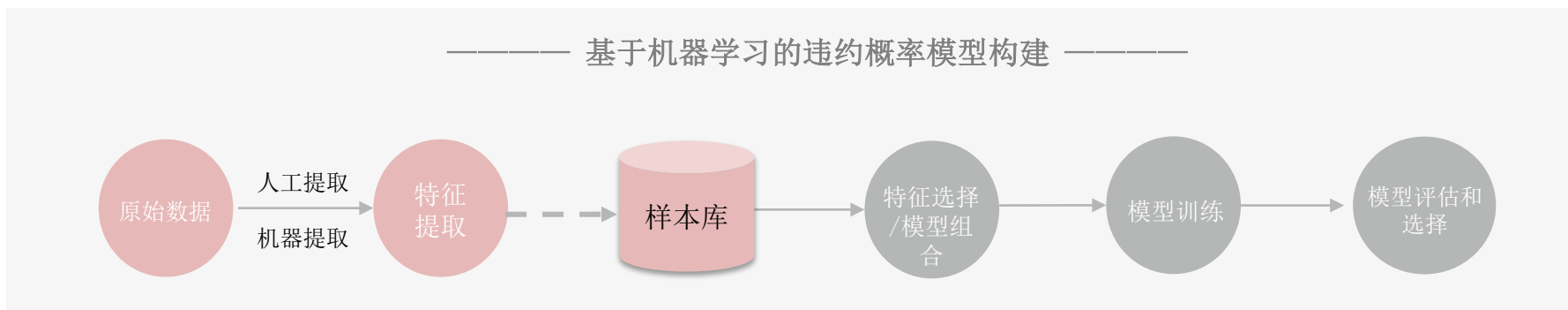
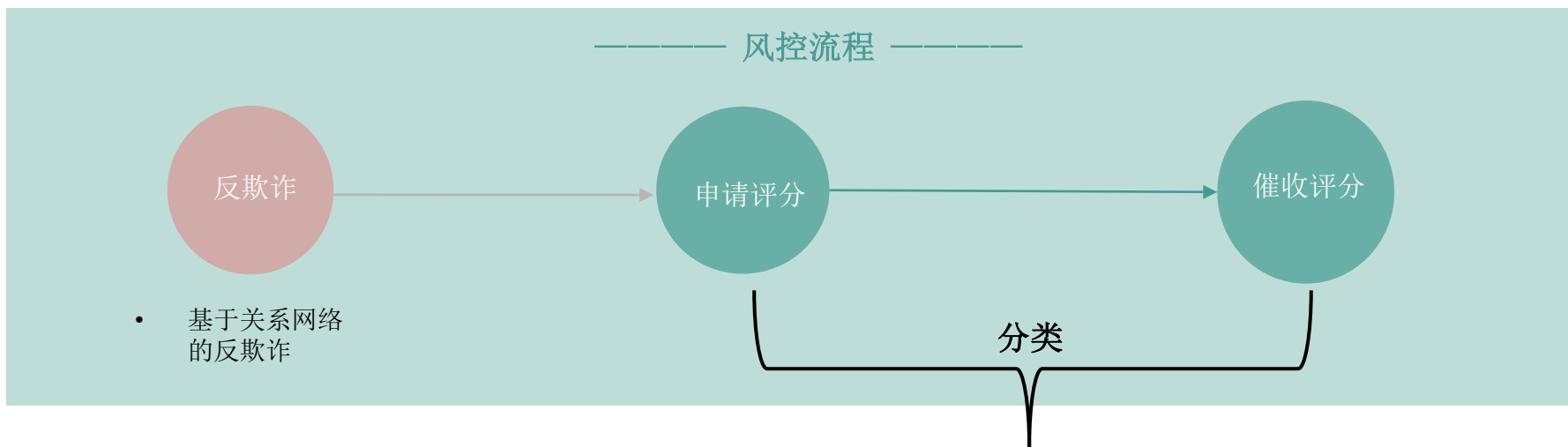
基于社会网络分析的反欺诈风险识别



基于申请网络识别欺诈风险



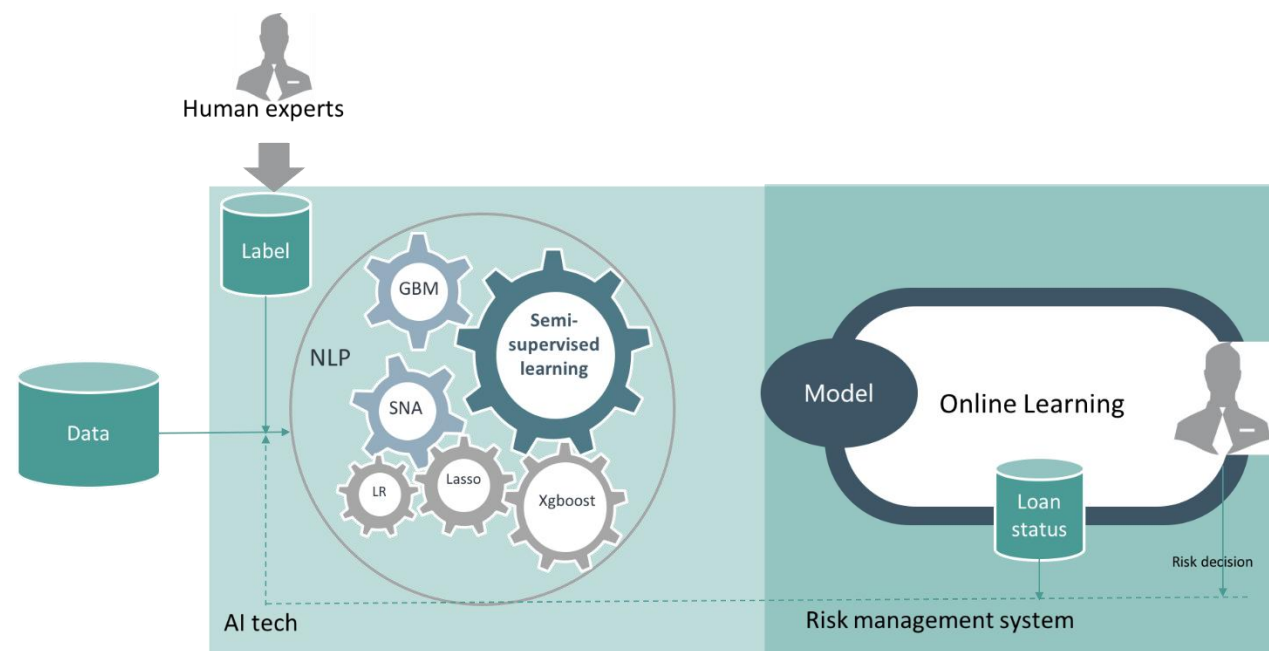
核心授信决策- 机器学习模型



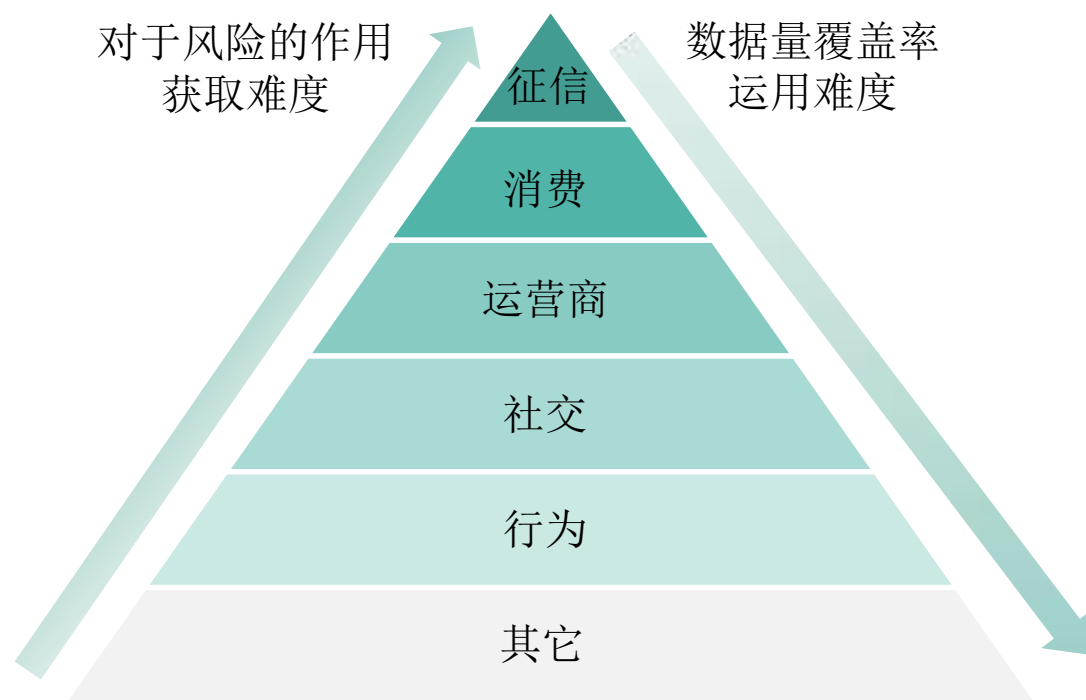
- 样本少：“冷启动”
- 数据多：非结构化数据特征提取
- 数据使用难：高维、稀疏、低饱和

样本少-“冷启动”

- 主动式学习：
 - 专家提供样本的标记
- 半监督学习：
 - 结合使用有标记和无标记样本来
 - 提高单一使用有标记样本性能
- Online-learning：
 - 线上迭代



数据多：所有数据都是风险数据



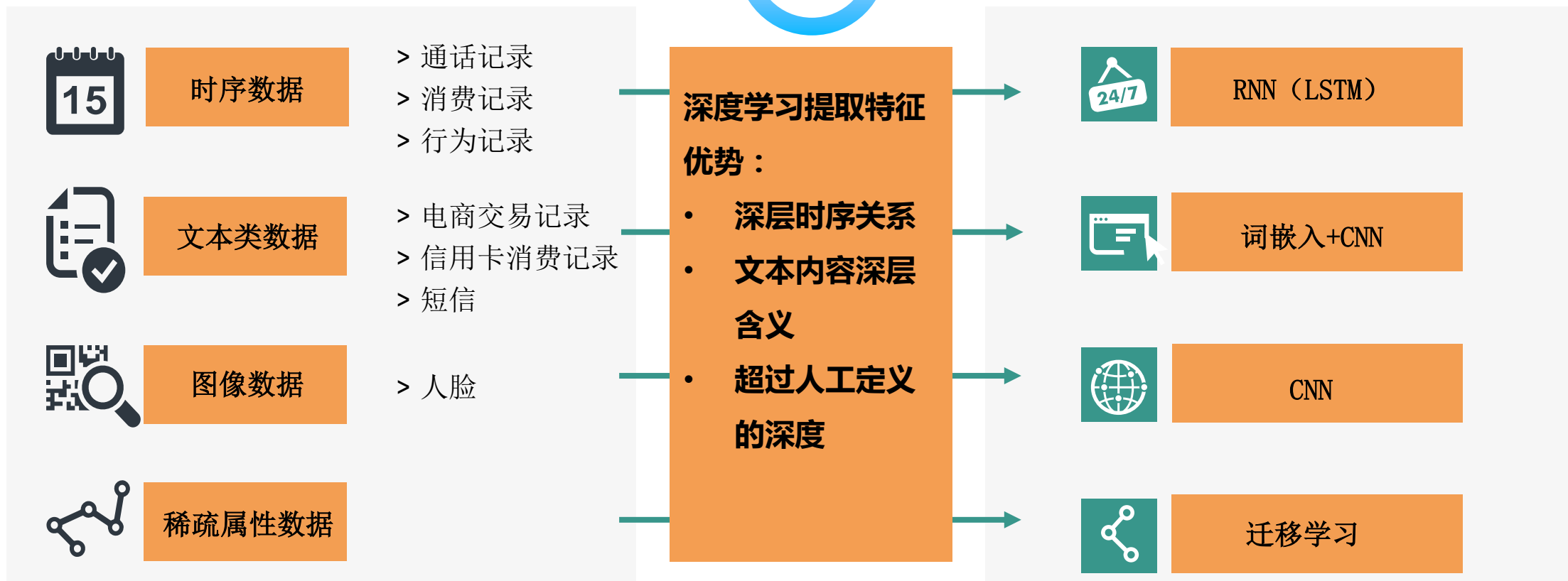
- > 短信
- > 通话记录

- > 消费记录
- > 消费地点
- > 消费金额

- > App使用数据
- > 设备信息

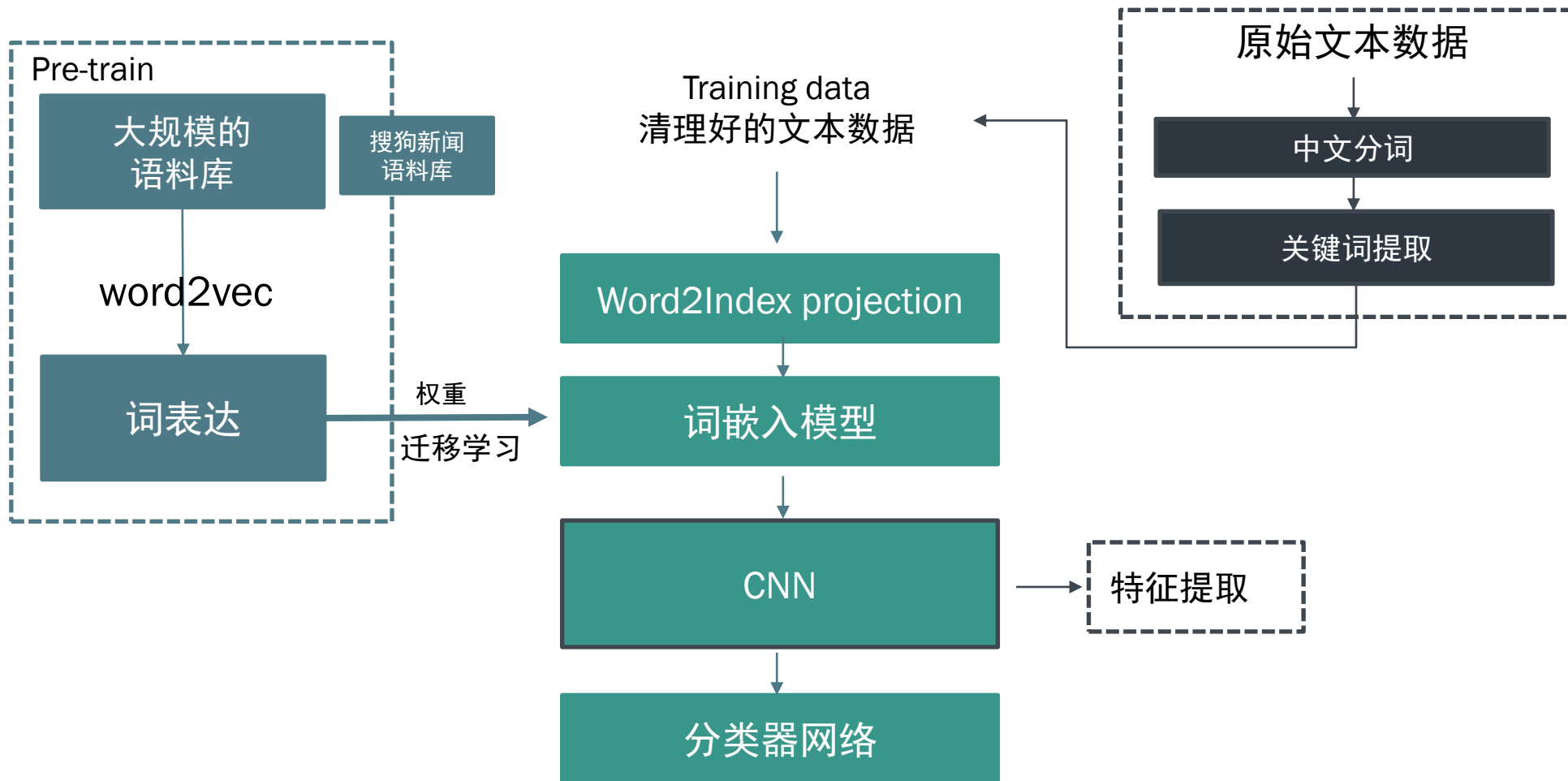
机器提取特征

深度学习

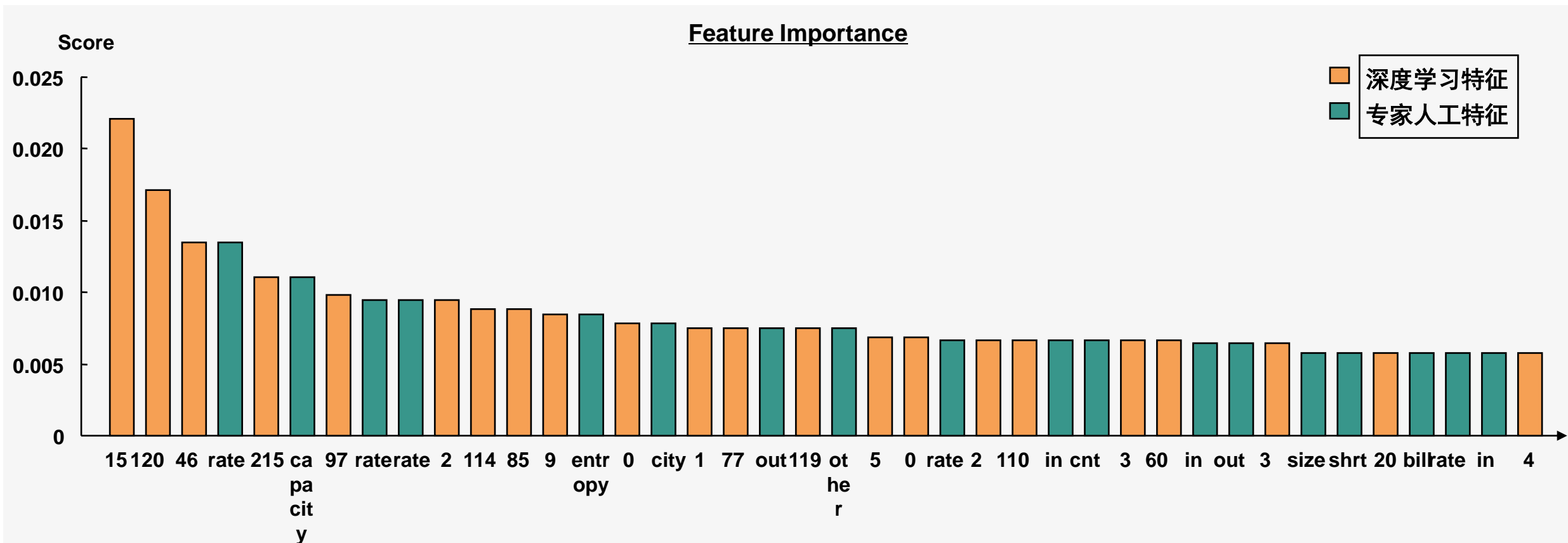


文本特征提取网络

word2vector , domain adaption



机器生成特征对模型效果的明显提升 与专家人工特征融合



数据使用难

先进建模技术



> 数据维度多

> 上千位特征变量

> 弱变量

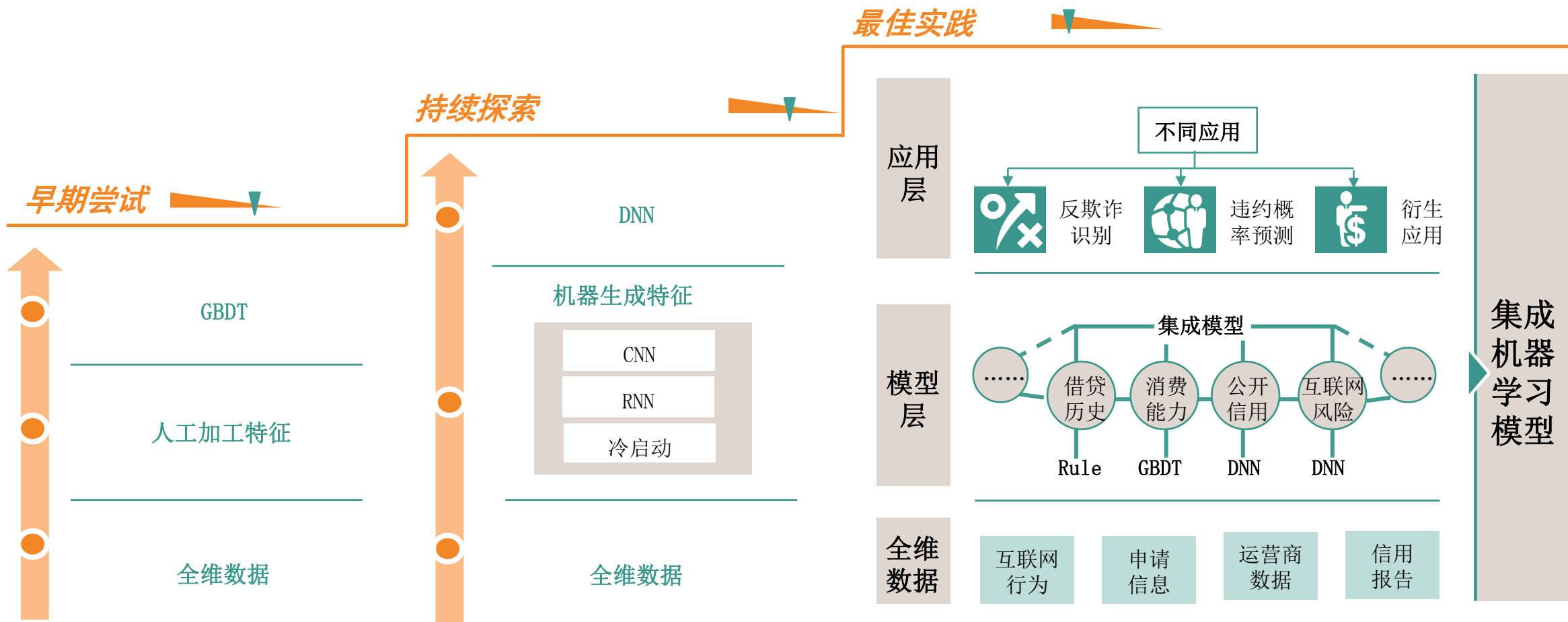
> 大量特征是弱变量，无法直接使用线性模型

> 数据稀疏、低饱和

> 存在大量缺失值问题

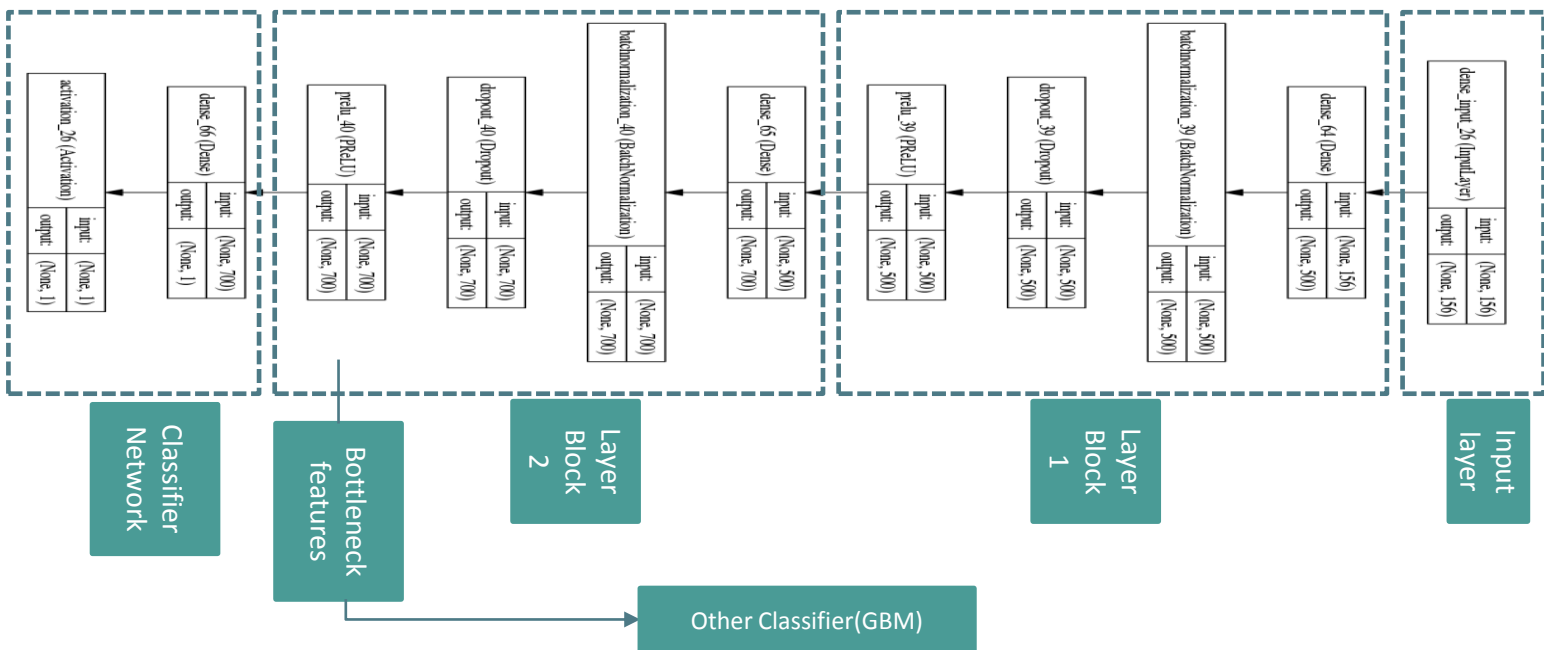
机器学习建模

机器提取特征+专家人工特征



模型效果对比

信贷贷场景



模型	准确率	查对率	最大KS值
DNN	0.41	0.70	0.32
Bottleneck features +GBM	0.42	0.65	0.28
Logistic regression	0.37	0.53	0.19

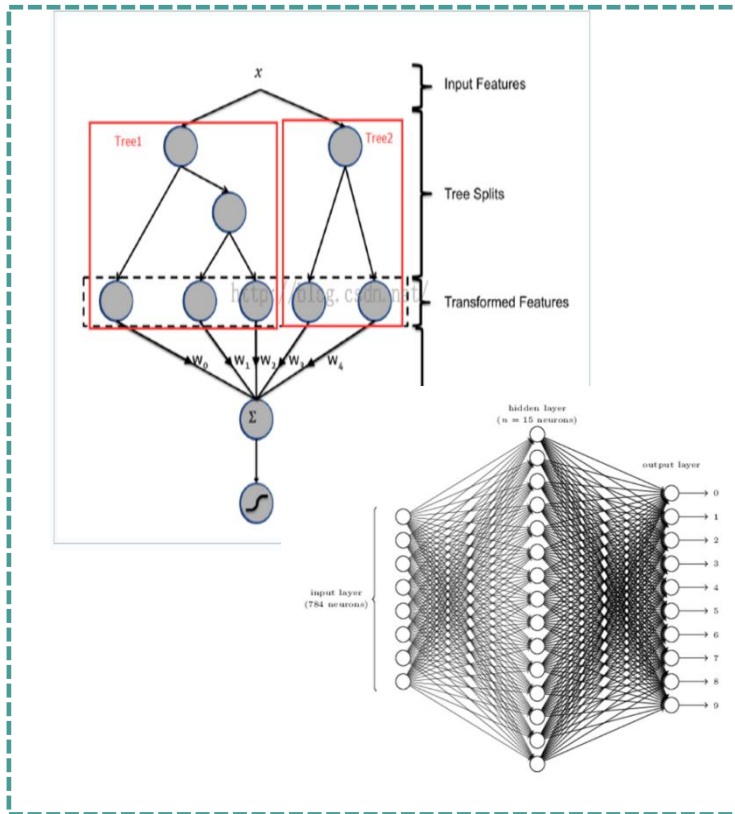
提升14%

提升45%

模型可解释

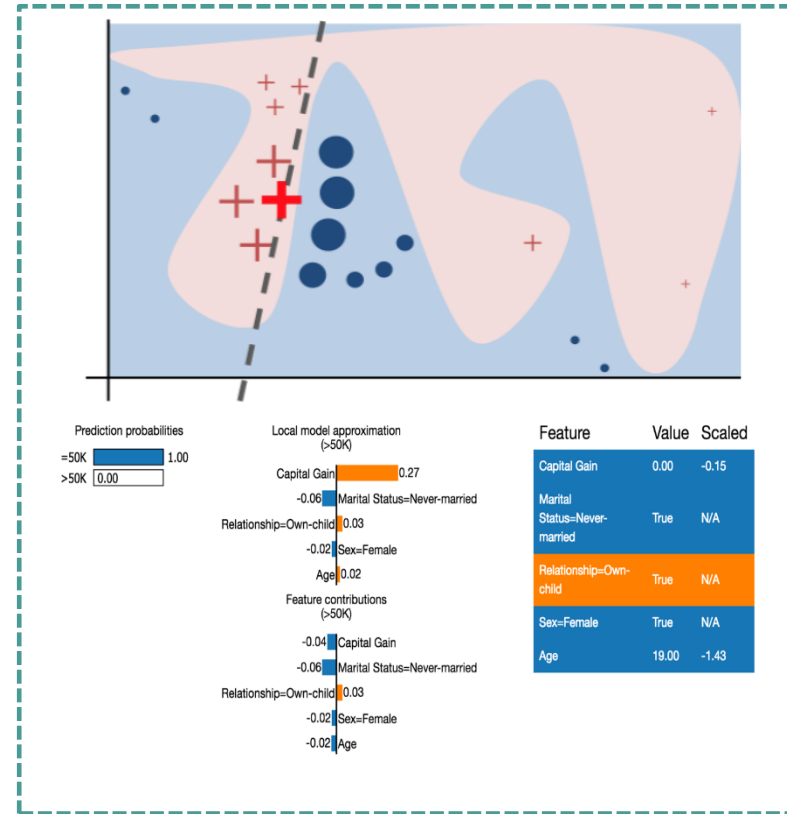
捕获影响结果的关键变量

Ensemble Black-box Models

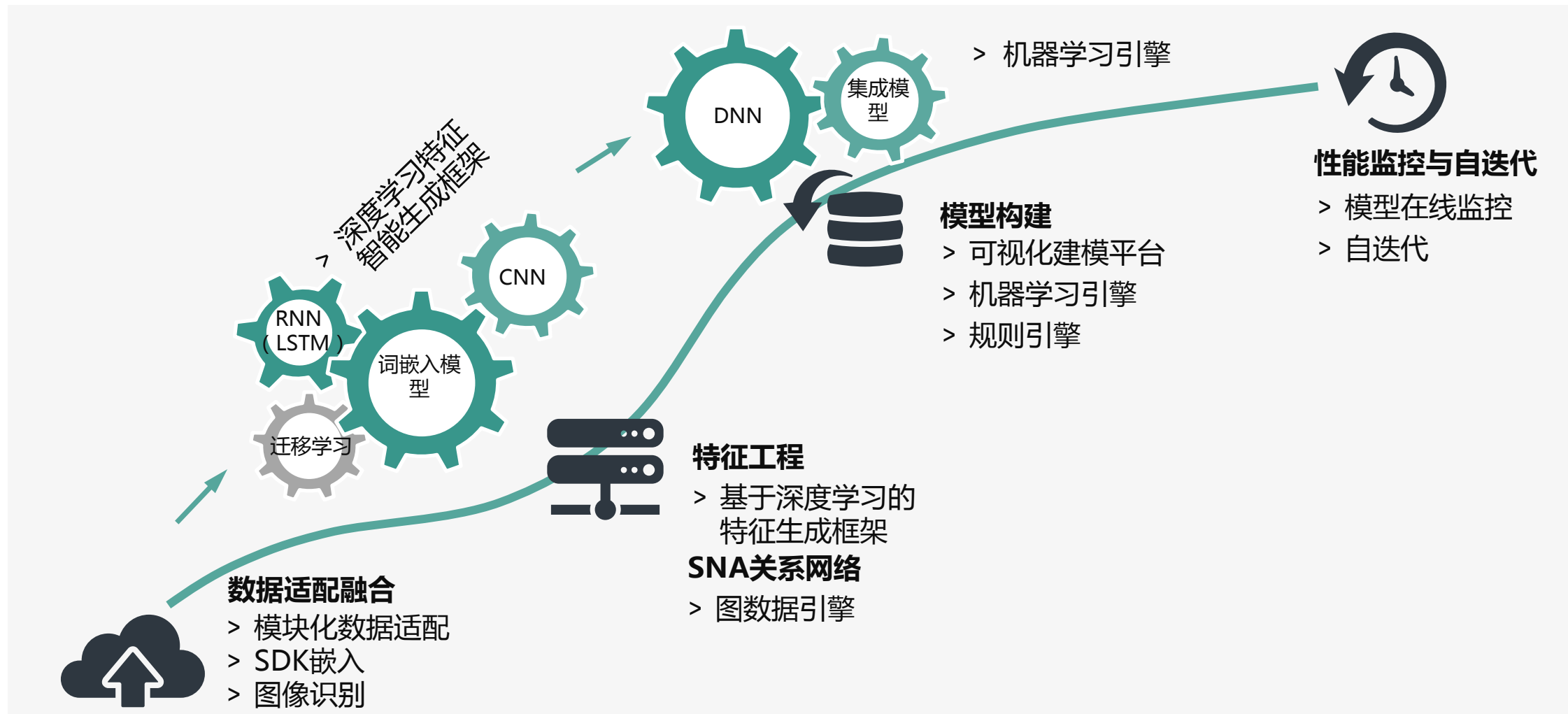


LIME

Local Interpretable Model-agnostic Explainer



AI-DRIVEN 大数据风控



合作伙伴与客户



Thanks

