

互联网金融风控中的数据科学

王婷 数据科学家 宜人贷

宜人贷

www.yirendai.com

技术架构未来

thegiac.com

个人介绍

- 计算机专业PH.D.
- 近5年从事数据挖掘、大规模社交网络分析、社会计算、知识图谱等机器学习算法实践工作
- 现任宜人贷数据科学家，从事反欺诈建模和创新技术自动化风控系统





背景

- 互联网金融行业在解决什么问题？ 面临怎样的风险？ 用什么样的技术手段来应对？
 - 传统金融 V.S. 互联网金融
 - 信用风险 V.S. 欺诈风险
 - 传统风控 V.S. 数据科学技术风控



互联网金融风控中的数据科学

- ① 互联网金融服务面临的风险
- ② 知识图谱在金融风控中的应用场景
- ③ 大规模金融服务中的实时架构



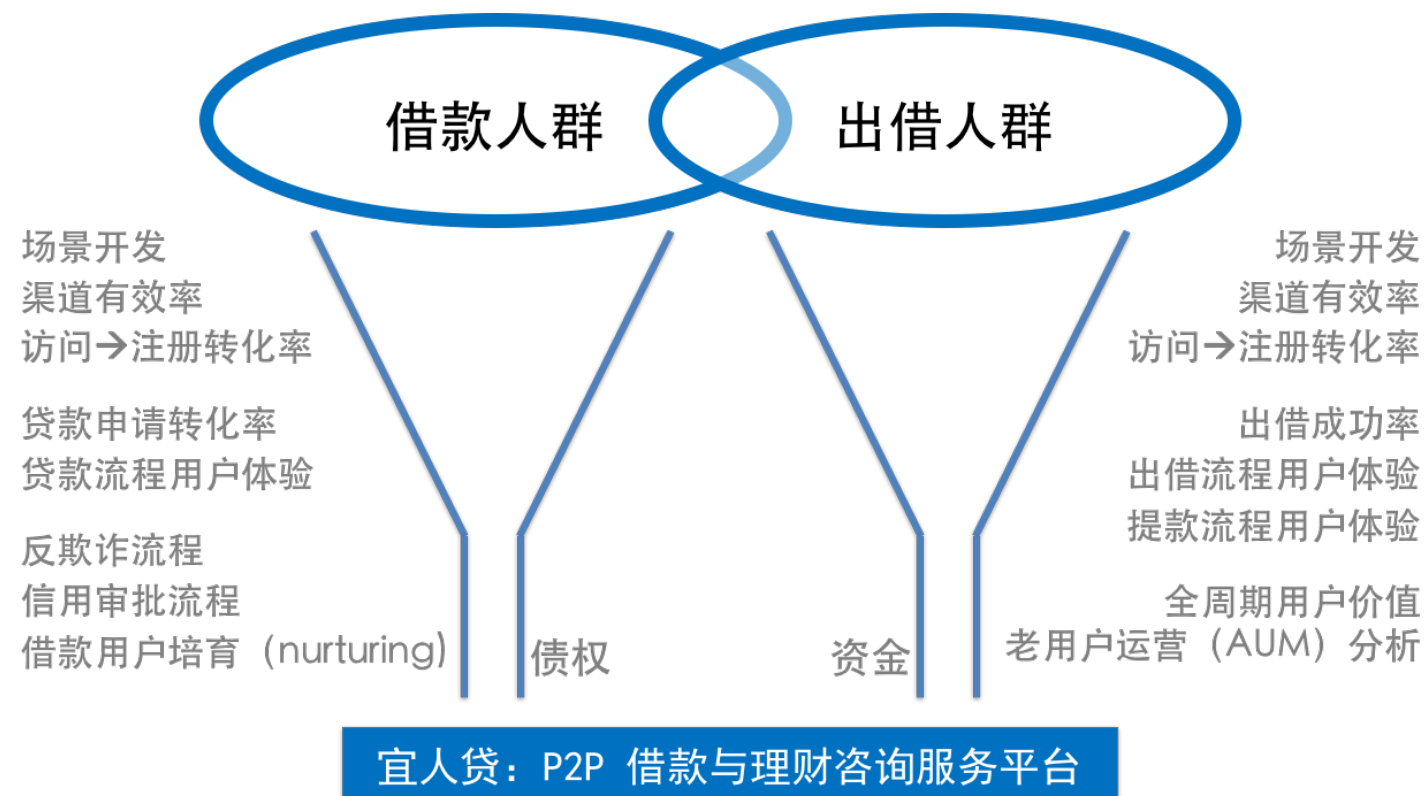
互联网金融风控中的数据科学

① 互联网金融服务面临的风险

② 知识图谱在金融风控中的应用场景

③ 大规模金融服务中的实时架构

互联网金融-个人对个人的信用贷款



宜人贷借款APP



宜人理财APP



宜人贷官方网站

宜人贷：极速信任 - 自动化信用评估

全流程线上借款与理财咨询服务

客户获取

信用评估

交易促成

客户服务



场景不同
人群不同
数据获取方式不同
数据维度不同
数据深度不同
信用评估机制不同

风险控制
是互联网金融
线上信贷工厂模式
最大的挑战

互联网金融行业中的风险

风险	遇到的问题	业界通常解决方法	业界的方法为什么无效
信用风险	还款能力	收集收入水平、消费水平、负债情况等对用户进行风险评分	无权威数据、数据收集难度大、传统评分卡有效特征挖掘难度大
欺诈风险	伪冒申请和欺诈交易	人工审查、信用黑名单、基于规则	人工效率低、无权威黑名单、无法自动发现异常、欺诈手段更新快

人群团体化



地区集中化



方式多样化



工具智能化





互联网金融风控中的数据科学

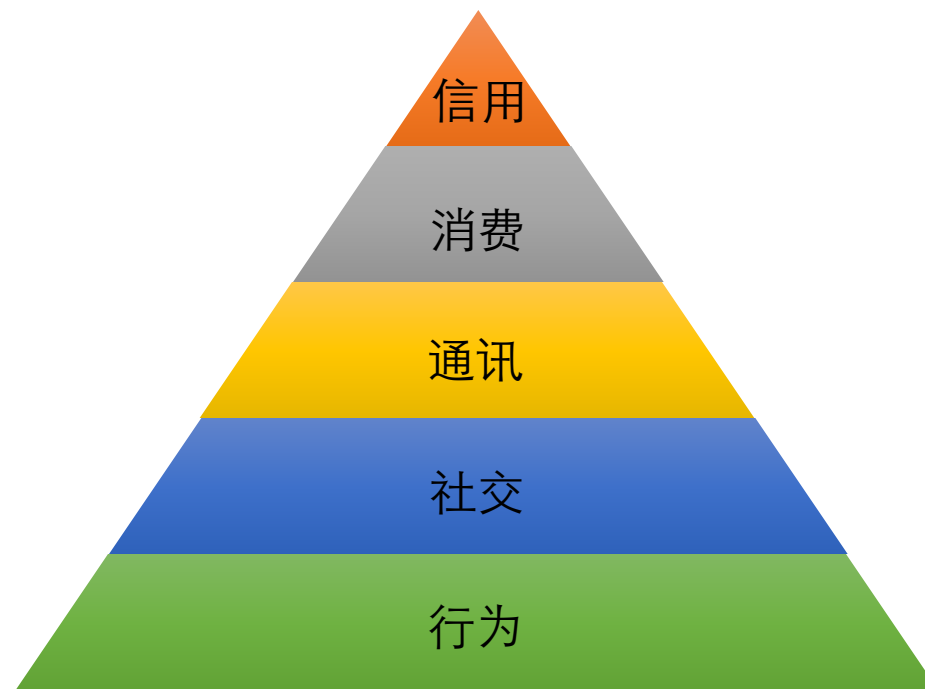
① 互联网金融服务面临的风险

② 知识图谱在金融风控中的应用场景

③ 大规模金融服务中的实时架构

互联网金融中的风控是一种机器学习过程

- Y目标： Benchmark选取
 - 好、坏用户定义
 - 训练、测试和跨时间验证样本
- X变量： 特征工程
 - 人工特征工程
 - 图谱特征挖掘技术
 - ✓ 知识图谱技术
 - ✓ 图挖掘技术



风险控制数据金字塔

知识图谱在金融风控中的优势

传统风控：

- 客户是否触黑
- 客户的消费记录是否异常

点



利用知识图谱风控：

- 客户一度、二度关系是否触黑
- 客户消费关联商家是否异常
- 一机多人
- 识别组团欺诈

面

构建金融知识图谱：FinGraph

FinGraph 平台系统

- 10种实体

- 电话、身份证、银行卡、信用卡、IP、设备号、地理位置等

- 约1.3亿节点

- 约10亿边关系

- 预计到2017年增长20倍

应用场景层面

智能搜索、反欺诈、贷后管理、营销分析、运营支撑 等

系统支持层面

特征工程、模型开发、异常监控、推荐系统 等
Spark+GraphX+Mllib+Streaming+TensorFlow

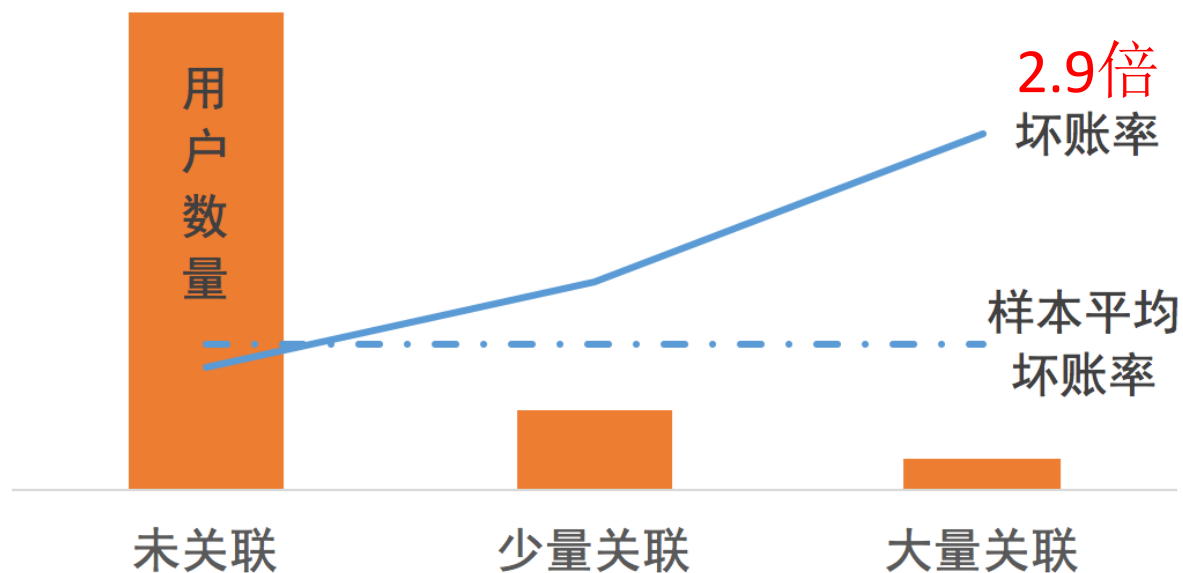
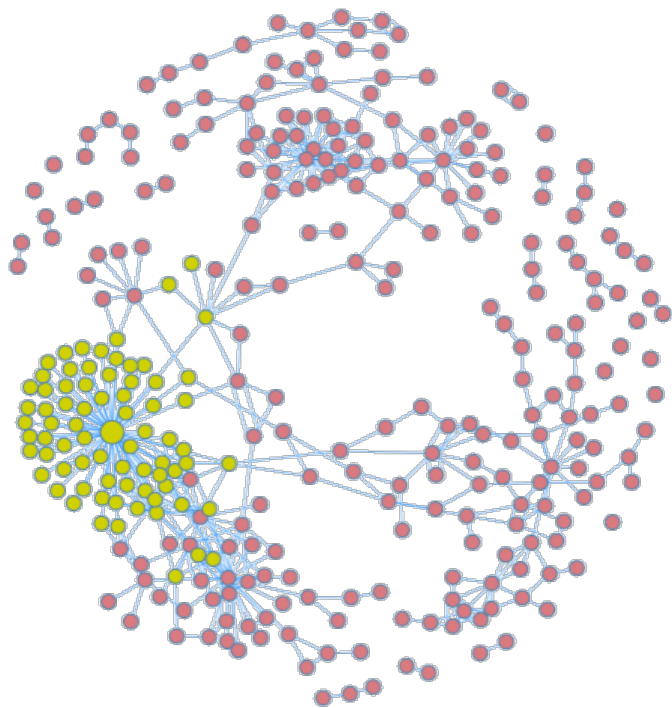
数据整合层面

信用数据、金融消费数据、行为数据、社交数据、
网络安全、第三方数据 等
图数据库neo4j

风险分析实践：人以群分(1)

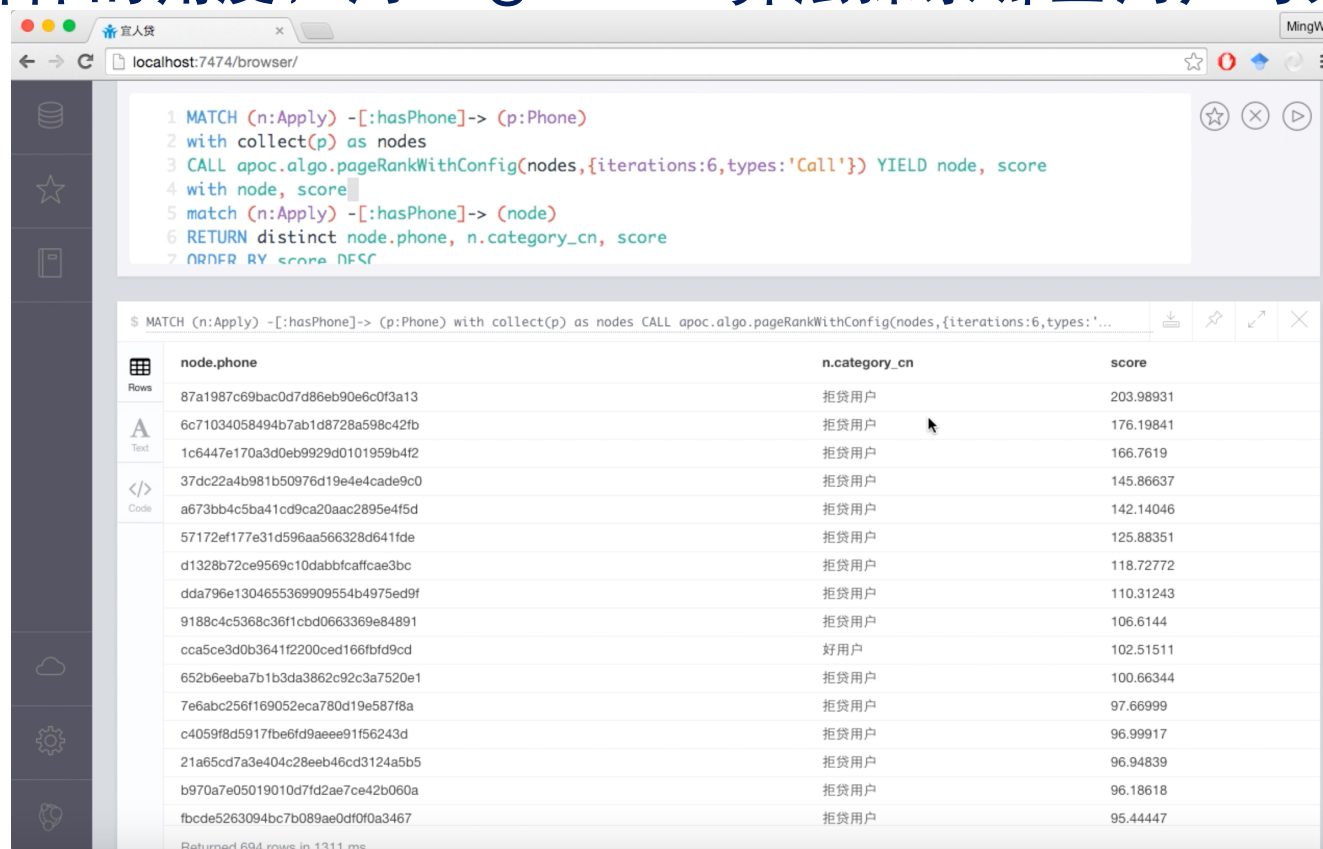
■ 借款用户通信社交网络与欺诈风险

- 结论：与坏用户有大量关联的借款用户的坏账率是未关联用户的2.9倍



风险分析实践：人以群分(2)

- 从整体借款群体的角度，用PageRank算法探索哪些用户与大量借款用户有关联关系



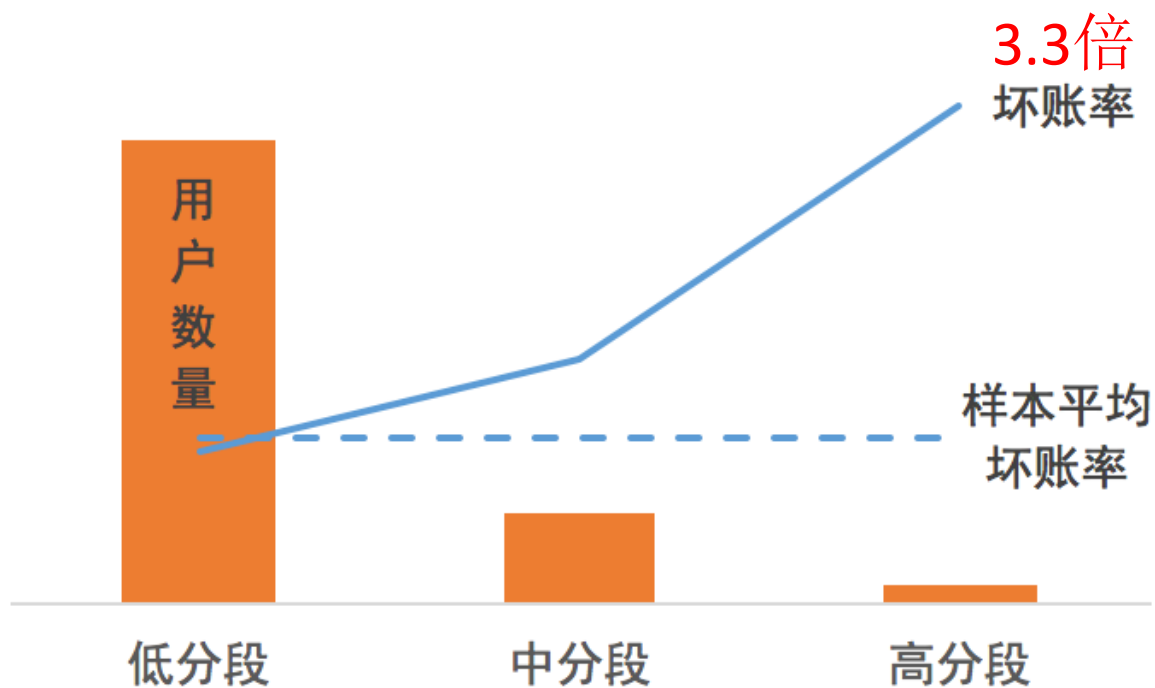
```
1 MATCH (n:Apply) -[:hasPhone]-> (p:Phone)
2 with collect(p) as nodes
3 CALL apoc.algo.pageRankWithConfig(nodes,{iterations:6,types:'Call'}) YIELD node, score
4 with node, score
5 match (n:Apply) -[:hasPhone]-> (node)
6 RETURN distinct node.phone, n.category_cn, score
7 ORDER BY score DESC
```

node.phone	n.category_cn	score
87a1987c69bac0d7d86eb90e6c0f3a13	拒贷用户	203.98931
6c71034058494b7ab1d8728a598c42fb	拒贷用户	176.19841
1c6447e170a3d0eb9929d0101959b4f2	拒贷用户	166.7619
37dc22a4b981b50976d19e4c4ade9c0	拒贷用户	145.86637
a673bb4c5ba41cd9ca20aac2895e4f5d	拒贷用户	142.14046
57172ef177e31d596aa566328d641fde	拒贷用户	125.88351
d1328b72ce9569c10dabbfcaffcae3bc	拒贷用户	118.72772
dda796e1304655369909554b4975ed9f	拒贷用户	110.31243
9188c4c5368c36f1cbd0663369e84891	拒贷用户	106.6144
cca5ce3d0b3641f2200ced166bfbd9cd	好用户	102.51511
652b6eeba7b1b3da3862c92c3a7520e1	拒贷用户	100.66344
7e6abc256f169052eca780d19e587f8a	拒贷用户	97.66999
c4059f8d5917f6e6fd9aeee91f56243d	拒贷用户	96.99917
21a65cd7a3e404c28eeb46cd3124a5b5	拒贷用户	96.94839
b970a7e05019010d7fd2ae7ce42b060a	拒贷用户	96.18618
fbcd5e263094bc7b089ae0df0f0a3467	拒贷用户	95.44447

Returned: 694 rows in 1311 ms.

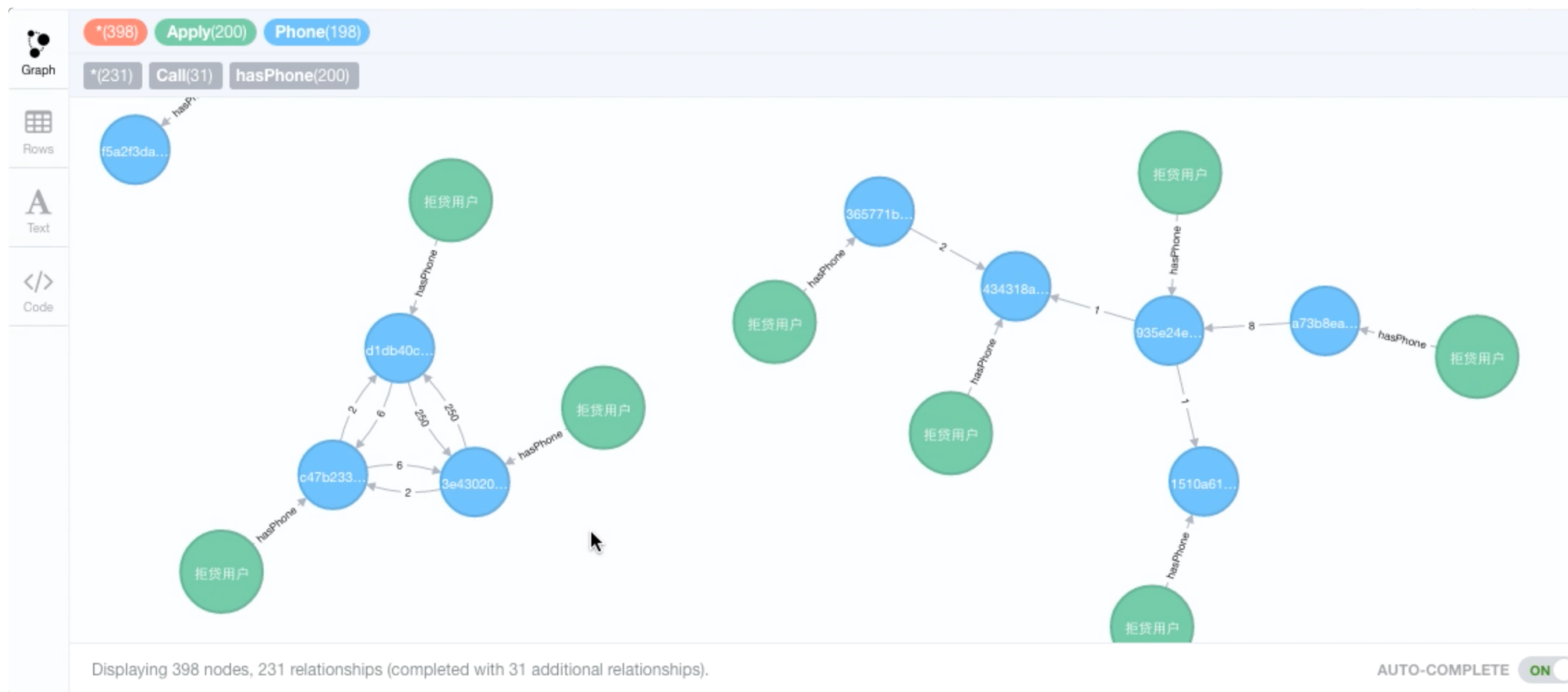
风险分析实践：人以群分(2)

- 借款用户通信社交网络与欺诈风险
 - 结论：PageRank高分段用户的坏账率是低分段用户的3.3倍



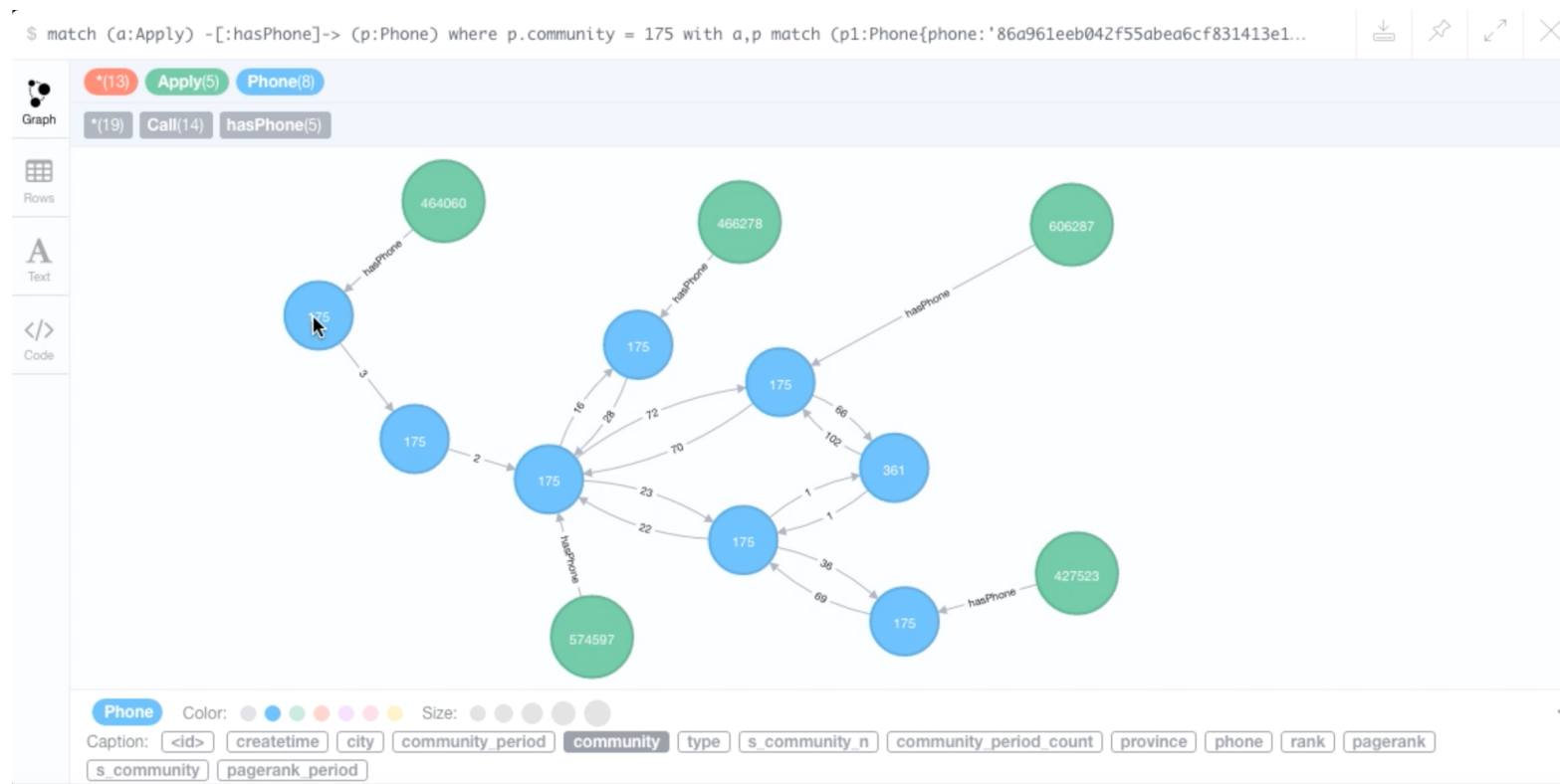
风险分析实践：识别组团欺诈风险

- 通过社区发现算法来实时评估每个用户的组团欺诈风险



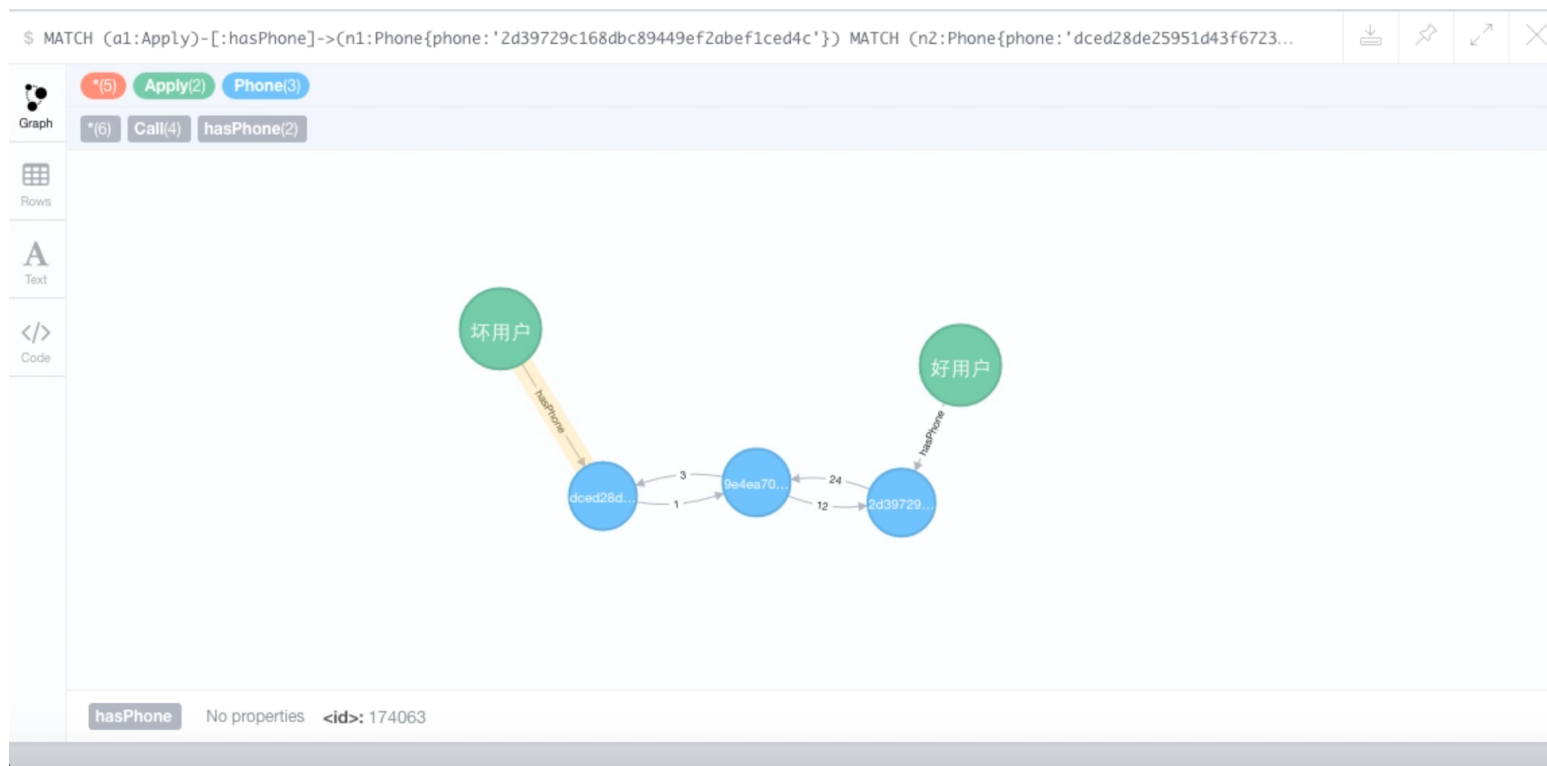
欺诈案例调查的挑战：失联用户找寻

- 通过社区发现算法挖掘失联用户的通信社交网络

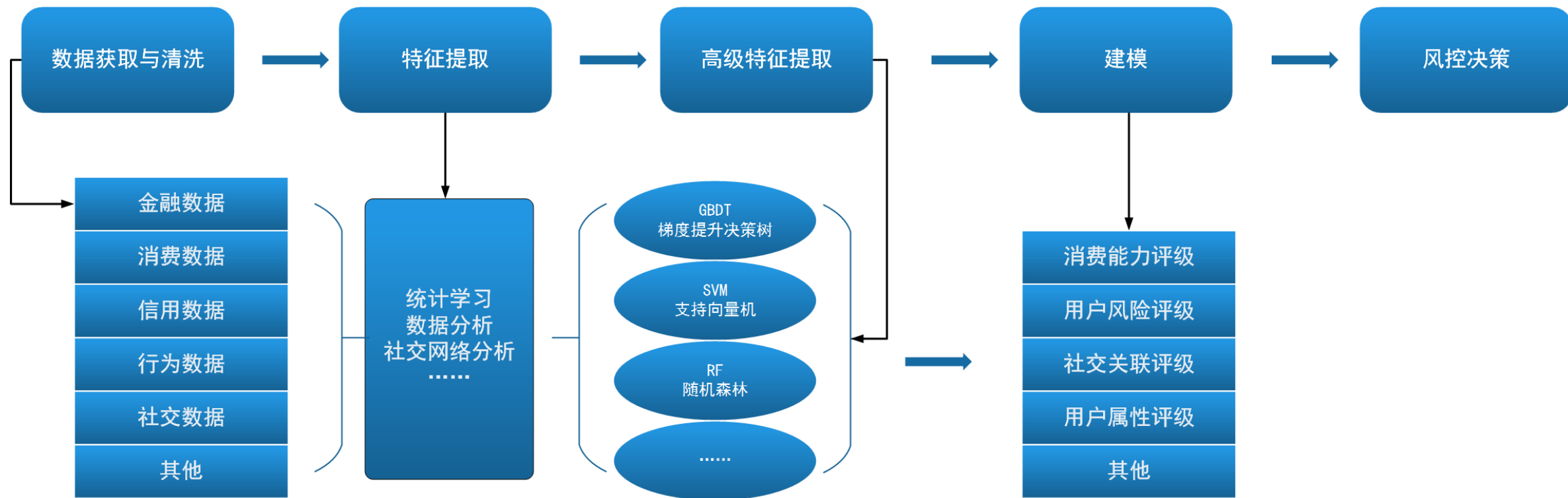


欺诈案例调查的挑战：失联用户找寻

- 在社区内使用最短路径算法来发现失联用户与一个正常还款用户的关系链



风控建模中的数据科学



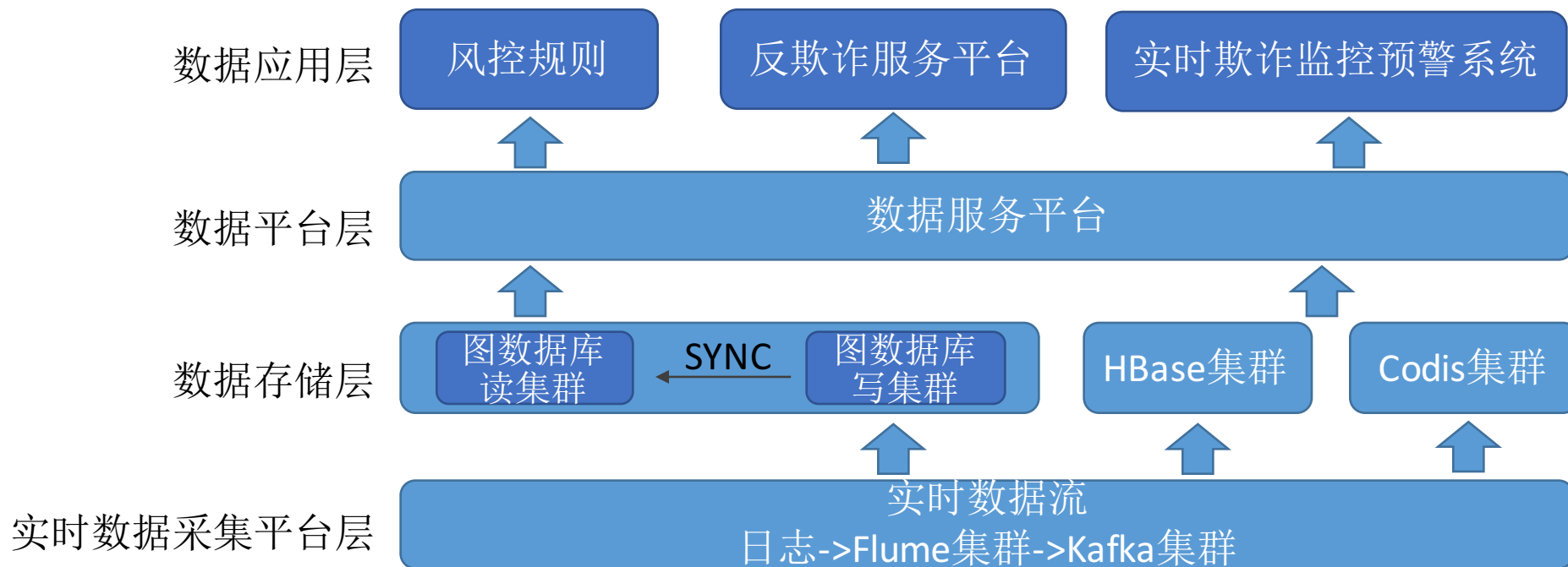


互联网金融风控中的数据科学

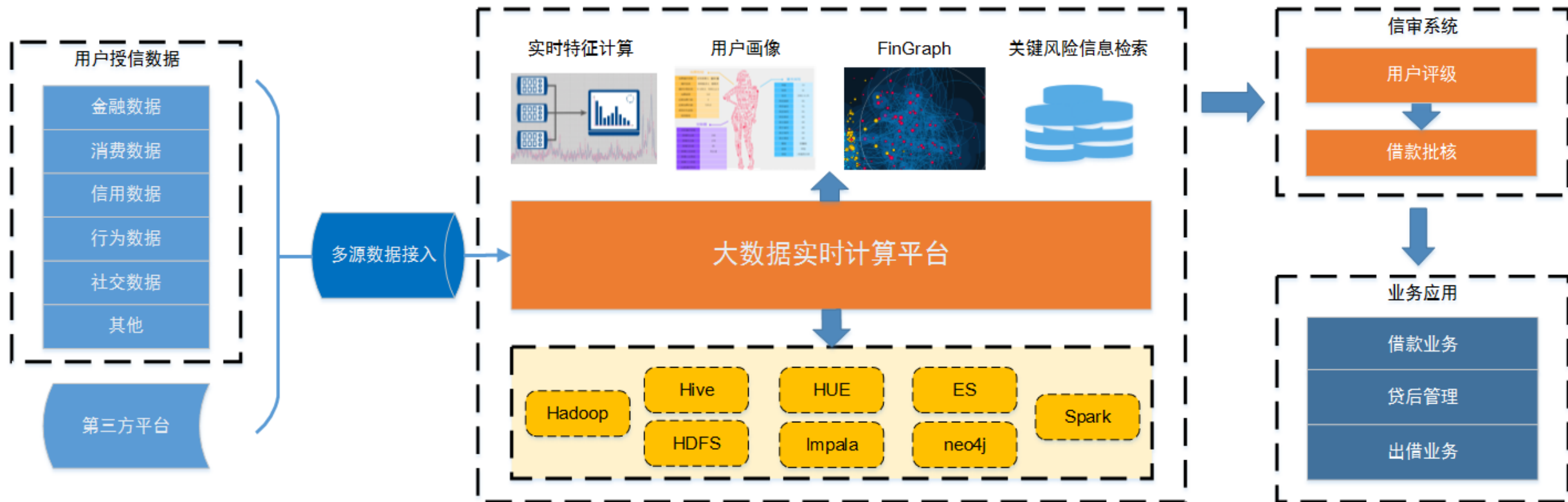
- ① 互联网金融服务面临的风险
- ② 知识图谱在金融风控中的应用场景
- ③ 大规模金融服务中的实时架构**

FinGraph是实时风险监控的重要支撑

- 基于实时数据采集平台和图数据库，可实时捕捉风险特征，控制风险
 - 数据在采集、传输、存储时达到99.999%的可靠



FinGraph是线上风控系统中的关键一环



总结：数据科学在互联网金融风控中发扬

互联网金融
风控工作
是一种机器学习过程

实时风险控制
在线上P2P借款与理财
咨询服务行业中，是全
自动信用评估的关键

图挖掘技术
可以把风控工
作，从局部考量
提升到全局考量



数据科学
已逐渐渗透到
离线建模流程
中的多个环节

FinGraph 知识图谱平台

- 结合了上亿节点和十亿关联关系提供全面风控能力
- 用户社交网络关系特征可以分辨出3倍欺诈风险的人群
- 用户的组团欺诈风险可以通过社区发现算法实时评估
- 失联用户也可以通过图挖掘算法来找寻



FinGraph
已逐渐成为线
上风控系统中的
关键一环

GIAC | BEIJING
Dec.12.16-17

技术架构未来

架構
ARCHNOTES
高 可 用 架 构

Thank You !

