金融知识图谱建设及其在银行业务中的应用

梁栋
• 北京松鼠山科技有限公司
• 哈尔滨工业大学智能金融联合实验室

关于松鼠山科技

松鼠山科技，是国际领先的金融科技与人工智能领域开拓者， 运用大数据及人工智能相关技术服务于**以银行业、保险业与基金业为主的金融行业，致力于推动各金融垂直场景的赋能**。松鼠山科技的核心竞争力来源于前沿扎实的科研能力和传统金融的深厚经验。

其专注于金融科技领域赋能建设，其中的**业务核心团队具有15年中国金融行业信息化建设经验。**

“松鼠山”（Squirrel Hill）位于美国**卡耐基梅隆大学**校区一侧，我们的创始人以此命名公司。

专注的业务范围

经过四年多的高速发展，松鼠山科技逐渐形成了以Beeler（结构化大数据分析）为基础， Fusion 为核心，结合海量企业工商数据、专利信息、投融资信息、舆情信息、监管数据、司法数据，打通整合金融机构第一方自有数据，凭借松鼠山科技一流的工程实现能力、算法设计能力和模型建设能力，以多年金融行业的从业经验为指导，覆盖“风险控制”，“精准营销”等一系列金融关键环节的应用服务体系。

松鼠山科技 SQH Tech

CONTENTS

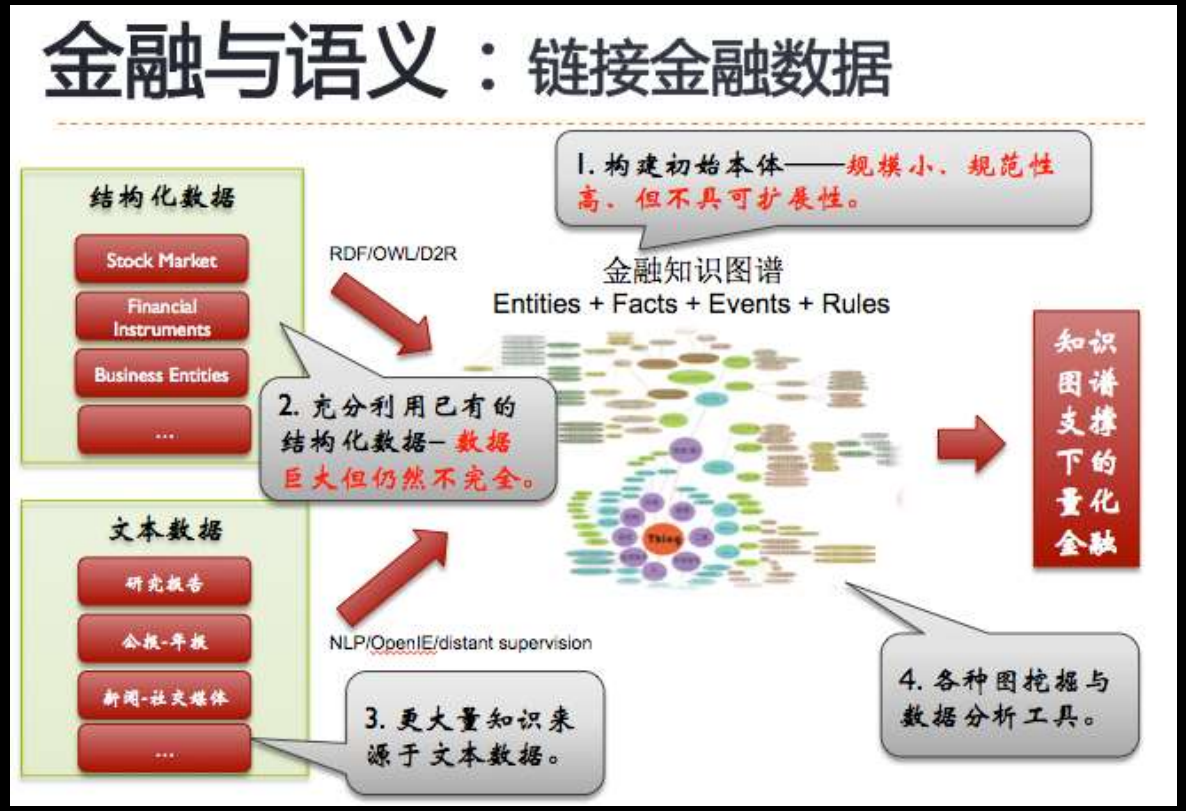
知识图谱的建设背景

– **客户认知**的问题：在外界市场纷繁的变化中，哪些因素发生了变化，对我行的客户造成了影响、价值（风险）有多大？

知识图谱：提供了从企业群体关系角度分析问题的能力。进一步可以拓展企业和行业、企业和客户、企业和市场价格等要素的关联传导。

大数据：完成从原始信息到企业关系、风险特征、风险事件的提取和整合。

机器学习：建立预测模型，实现风险识别、传导和预测的能力。



知识图谱的建设背景

松鼠山科技 SQH Tech

大数据感知

结构化数据
- 银行机构
- 监管/政府

非结构化数据
- 互联网媒体
- 垂直行业

流数据
- 交易流水
- 订单采购

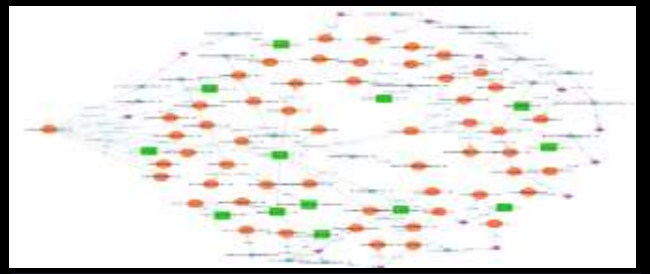
知识提取


投资关系


交易关系


贸易关系

知识存储



知识表现

规则 + 推理

知识检索

机器学习

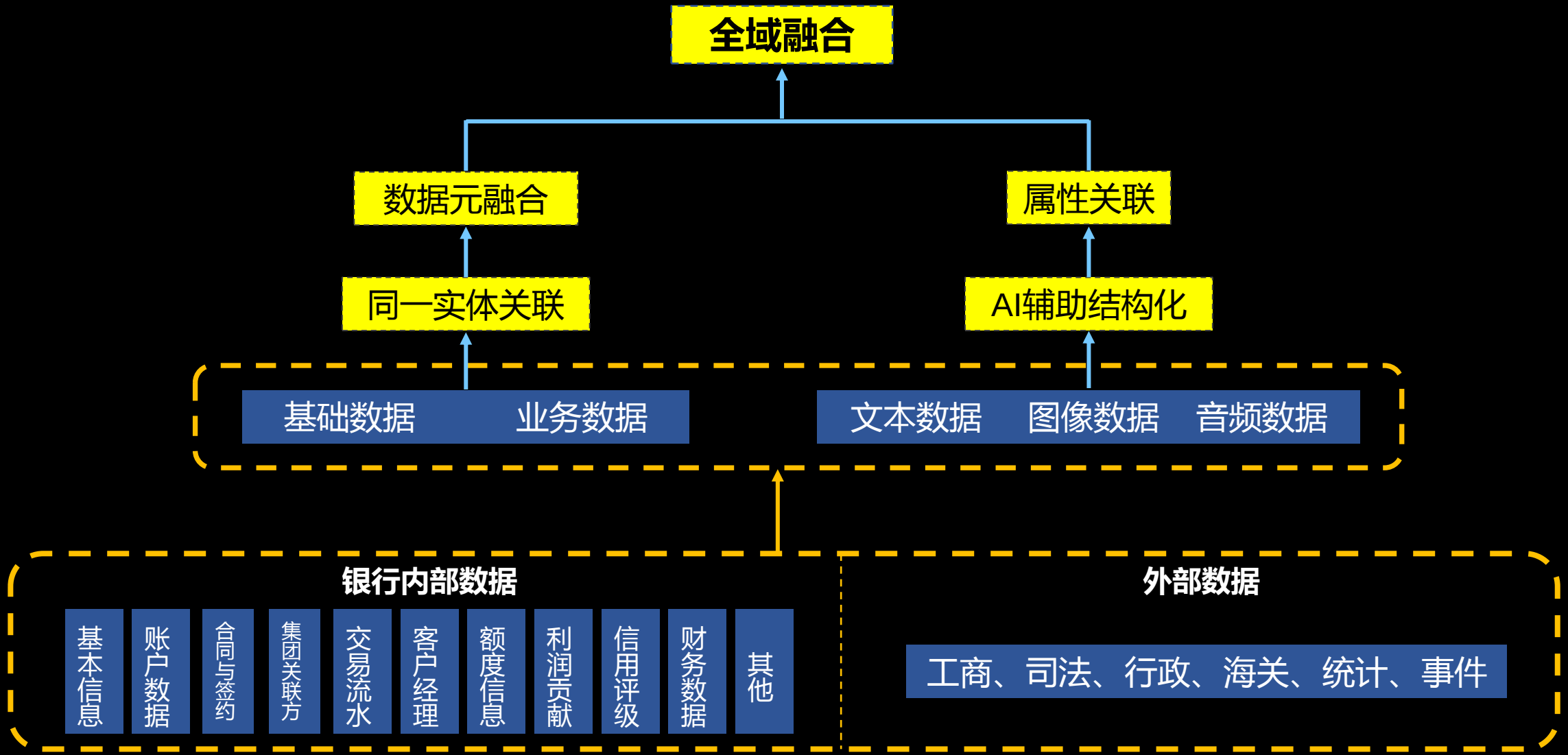
逻辑回归 + 决策树 + 神经网络 + 复杂网络 + 集成算法

松鼠山科技 SQH Tech

CONTENTS

关键路径：数据引入（策略）

松鼠山科技 SQH Tech



关键路径：数据引入（范围）

松鼠山科技 SQH Tech

结构化数据

内部数据

外部数据

非结构化数据

- CRM：对公客户基本信息、
- 信贷：对公客户贷款信息、授信信息、担保信息、关键关联人信息
- 评级器：对公客户评级信息、财务信息
- 中间业务：用水用电信息
- ……

- 银监会数据：银监大额不良、银监大额预警、银监零售违约、银监限制行业等
- 人民银行：征信报告
- 工商局数据：客户集团关系、客户股权关系
- 税务局数据：税收情况、开票信息（验证客户收入情况）
- 海关数据：贸易往来关系、 进出口贸易金额、频次
- 法院数据：法院诉讼文件、法院判决文件

- 授信报告：客户核心竞争力、重大事项等信息
- 评审报告：审议决定内容中的密切需要关注内容、督促内容等信息
- 融资规划报告：重点任务、资金供需分析结论等信息
- ……

- 社交媒体：新浪微博（需购买数据），微信公众号，QQ
- 新闻门户：新浪、网易、搜狐、腾讯、凤凰，财新网，证券日报、京华时报，人民网、新华网、百度新闻等
- 公司资讯：同花顺、大智慧
- 行业论坛：行业报告、行业趋势分析、行业专家看点
- 其他数据源：天涯社区、时事点评、行政法规

构建知识图谱-单一客户谱系（数据与算法）

松鼠山科技 SQH Tech

单一客户谱系——基于大数据的精准场景服务

非结构化数据
处理加工（NLP)

结构化数据
清洗整合

企业关联
知识库

4800余万
商事主体

1.5亿
企业关联自然人

2.9亿
投资、任职关系链

1100万
疑似关系链

单一客户谱系算法模型

- “权益法”算法
- 递归穿透优化算法
- “企业关联方”规则模型
- 实际控制人模型
- 受益所有人规则模型
- “一致行动人”规则模型
- 最短关联路径算法
- 最强全局路径算法
- 节点关联度模型

- 企业关联谱系
- 企业股权架构
- 关联探查
- 企业实际控制人
- 企业关联方
- 受益所有人

客户尽职调查

贸易背景真实性调查

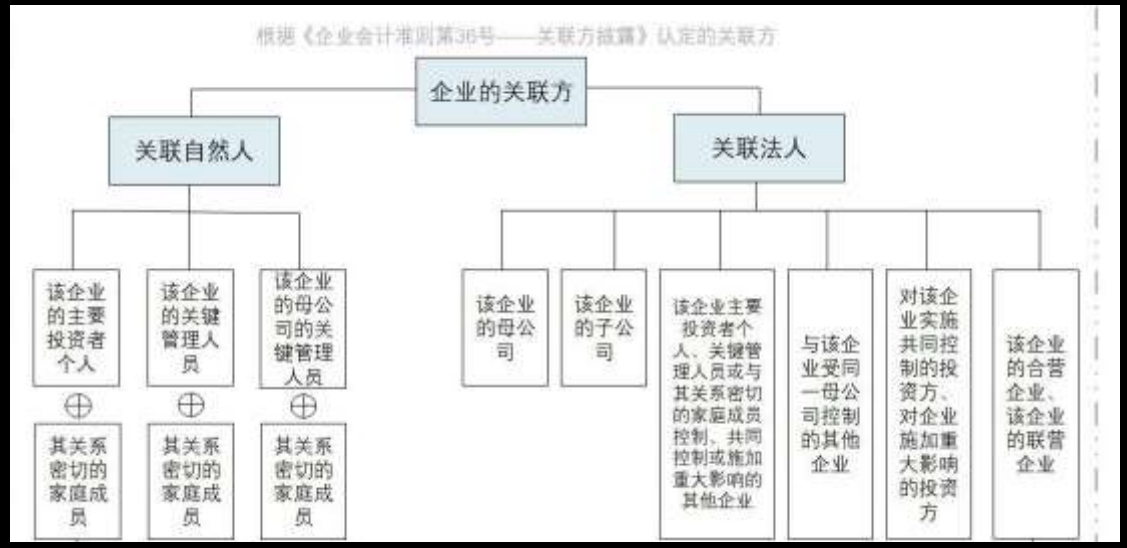
财务会计审计

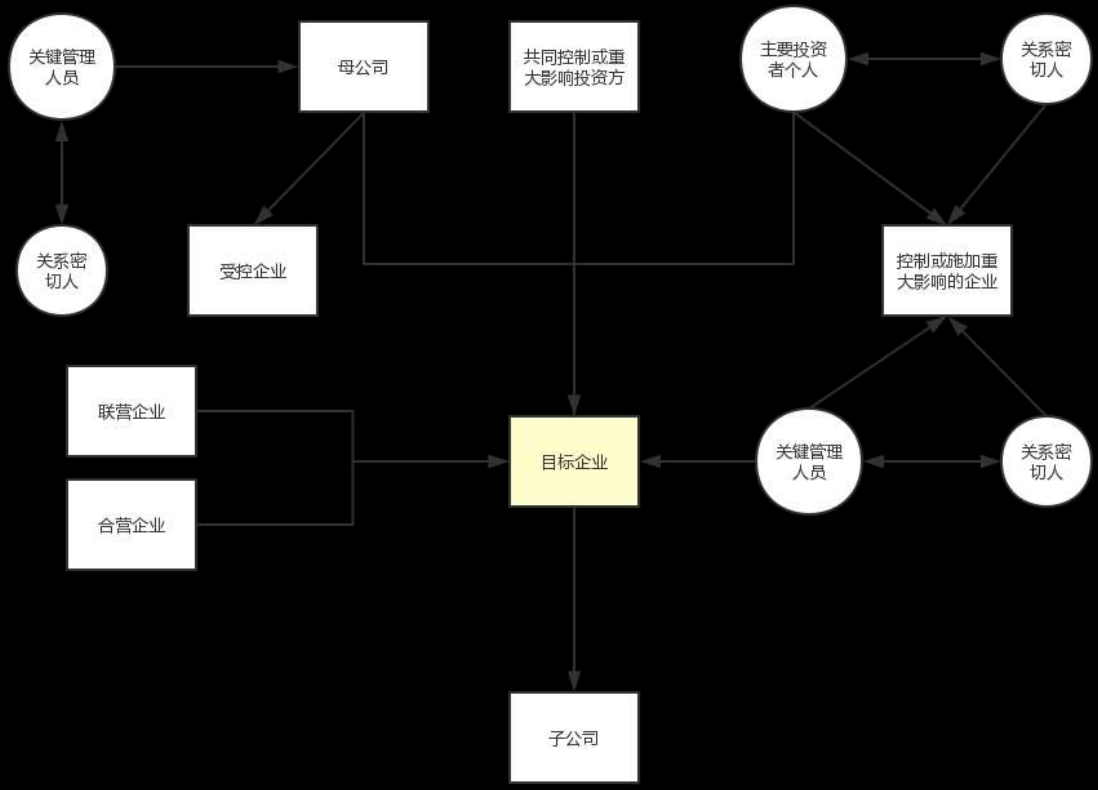
风险预警监控

信贷准入

构建知识图谱-企业关联

财政部《企业会计准则第36号——关联方披露》中明确描述了非上市公司的关联方定义标准，因此，银行在进行尽职调查时，需明确排查企业关联方交易、担保等情况。我们的业务经验认为，通过知识图谱的搭建与分布式图计算技术，将企业外部数据（股权、任职）与行内数据（自然人亲属关系）进行融合，可以对定性规则加以量化，形成一套基于知识图谱的挖掘算法，满足“穿透制”与“实质重于形式”原则，对企业的关联方关系进行快速认定，减少银企之间的信息不对称。

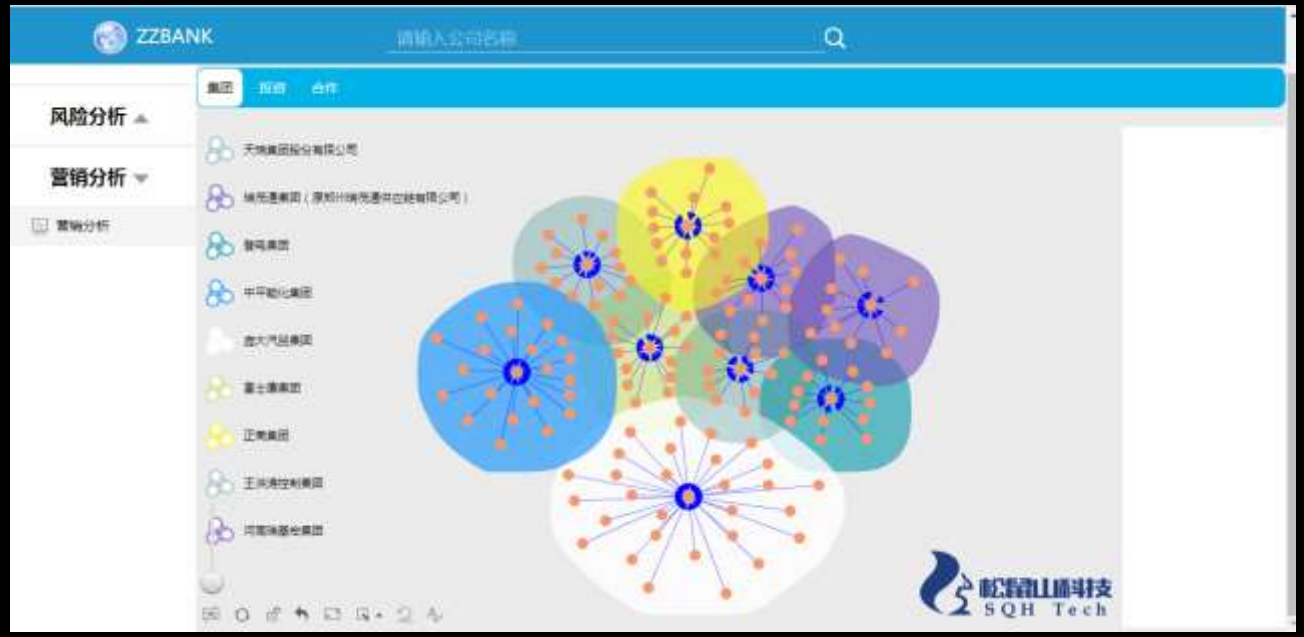




构建知识图谱-图计算模型

-基于图连通性的关联模型

- **单源最短路径模型**
  在复杂关系图中，找到两个不同节点最近的关联路径
- **基于权重的最强关联搜索**
  在有权重的关系图中，找到两个不同节点多条路径中，权重和最大的一条路径（最重要）
- **多种社团发现模型**
  在复杂关系图中，找到关联关系联系紧密的一个节点群，主要用于 欺诈团伙发现，利益共同体发现，企业集团/系发现等场景
- **桥发现模型**
  在复杂关系图中，找到多个关联团中的关键联系节点
- **传导计算模型**
  在复杂关系图中，某一个节点发生状态改变，根据关联关系，寻找状态改变影响的范围和关联节点影响的程度。
- **自主设计的带参数的局部结构检索模型**
  比较难于理解，举例说明：查找销售额大于10（数值用户定义）亿的企业的客户，并且必须是高科技企业。
- **模式匹配模型**
  在图中查找某种固定的拓扑结构，场景：循环担保发现等。
- **紧密中心度模型**
  发现局部网络中，发现最中心的节点（最容易到达其他所有节点）



构建知识图谱-企业集团谱系（数据方法体系3）

松鼠山科技 SQH Tech

对于集团客户的核定标准银监会《银行集团客户授信业务风险管理办法》给出了可参考借鉴的定义。

我们认为集团客户的形式复杂多样，在界定其具体范围时，应自始至终遵循实质重于形式和“穿透制”的原则，从控制与施加重大影响的视角出发，重点分析隐匿在复杂关系链路中的施控与被控关系。

《中国银行业监督管理委员会关于修改（商业银行集团客户授信业务风险管理指引）的决定》——2010年第4号

第一章 总则

第三条 本指引所称集团客户是指具有以下特征的商业银行的企事业法人授信对象：

（一）在股权上或者经营决策上直接或间接控制其他企事业法人或被其他企事业法人控制的；

（二）共同被第三方企事业法人所控制的；

（三）主要投资者个人、关键管理人员或与其近亲属（包括三代以内直系亲属关系和二代以内旁系亲属关系）共同直接控制或间接控制的；

（四）存在其他关联关系，可能不按公允价格原则转移资产和利润，商业银行认为应当视同集团客户进行授信管理的；

前款所指企事业法人包括除商业银行外的其他金融机构。

商业银行应当根据上述四个特征结合本行授信业务风险管理的实际需要确定单一集团客户的范围。

指引解读

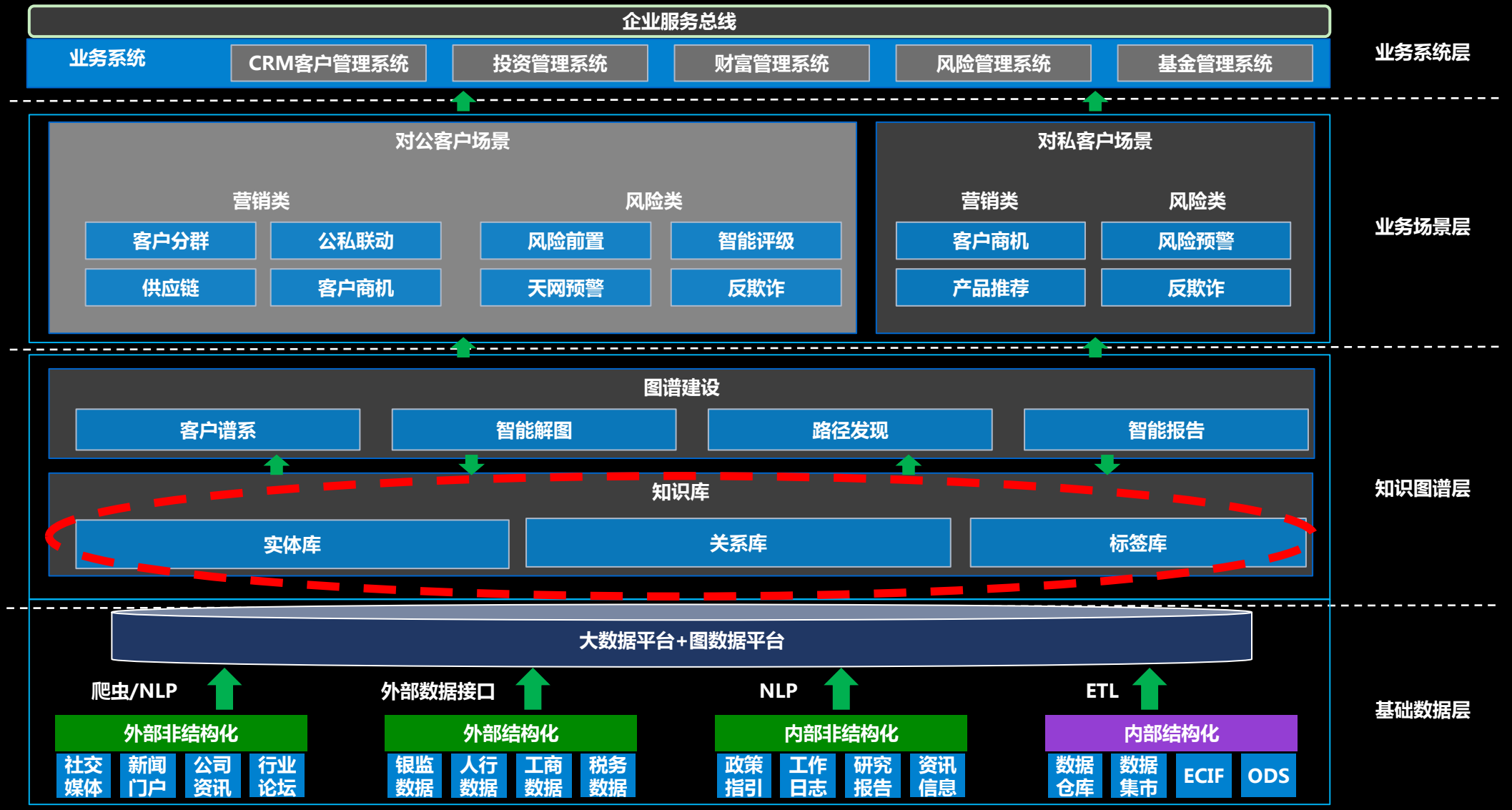
- 控股母公司
- 控股子公司
- 共同被第三方控制
- 关键管理人员或与其近亲属
- 施加重大影响

遵循实质重于形式的原则

松鼠山科技 SQH Tech

CONTENTS

知识图谱的关键路径（概览）



知识库建设-标签库体系（1）

松鼠山科技 SQH Tech

企业标签库体系（200+）

基础信息

生命周期

生存周期
年限分层
发展阶段

价值周期
客户分层
资产分层
负债分层
交易分层
核心客户
潜力升优质
普通升潜力
优质向下迁徙

交往周期
客户发展状态
客户关系时间
存款持有周期
贷款持有周期
理财持有周期

公司轮廓

企业性质
国有企业

企业规模
大型企业

治理结构
双层结构

集团标识
所属集团

关联企业
控股比例

上市企业
是否上市企业

产品持有

产品偏好
产品持有偏好
产品风险偏好
产品流动偏好
产品持有组合
持有产品数量
持有产品额度
首次购买产品

业务客群
主账户客群
存款客群
融资客群
中间业务客群
小微客群

资金流量
收款金额分层
下月到期资产分层
下月到期存款分层

交易行为

周期频率
客户购买频率分层

交易场景
交易行为偏好
转账行为偏好
交易时间偏好

渠道偏好
渠道使用频率

转账汇款
频繁大额转出
定期全额转出

交易趋势
金额变化趋势

价值贡献

直接贡献
存款贡献度
贷款账户贡献度
债券承销贡献度
票据承销贡献度
远期结售汇贡献度
资产托管贡献度

潜在贡献
资产收入价值分层
抵押物价值
担保合同金额
贷款授信额度

经营特征

贸易数据
进口总额分层
出口总额分层
主要贸易对象国
进出口主要商品

经营活动
纳税总额分层用电量分层
用水量分层

经营活动事件
重大战略发布重大投资项目
经营业绩预警
上市公司违规

往来关系

资金往来特征
往来户标识
是否我行客户
资金方向
结算工具
结算金额分层
前N大上下游企业

分布特征
往来户行业分布
往来户地域分布
与我行核心企业往来特征
与集团内关联企业往来特征
上下游往来户变化特征

风险特征

风险评级
客户评级
债项评级

违约历史
逾期次数
违约次数

特殊名单
银监大额预警
银监大额不良

外部事件
高管潜逃

关联关系
供应链圈子
担保圈子

知识库建设-事件库体系

客户风险事件预警中事件分类与分级是重心，不同类别事件对客户影响不同，同一类型事件，不同的严重程度对客户的影响不同。

风险预警事件分类

外部事件

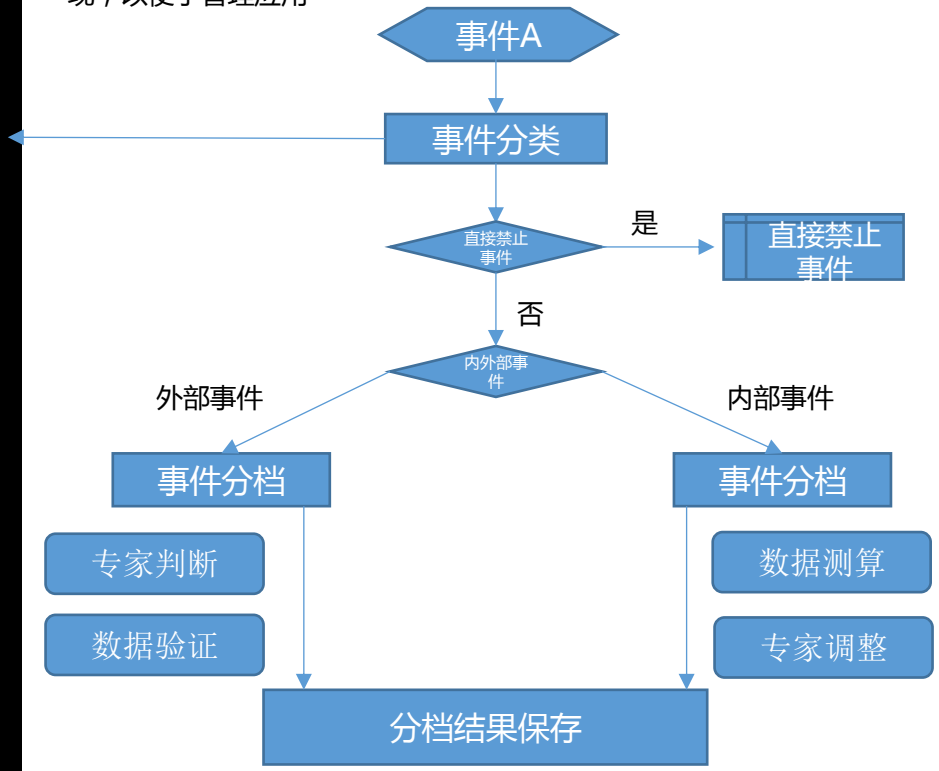
企业经营相关事件		外部环境相关事件	
公司事件	违法事件	宏观环境	行政法规
公司上市 ……	高管潜逃 ……	GDP下降 ……	支持新能源 ……
管理事件	意外事件	行业情况	物理环境
员工留存率过低 ……	工厂起火 ……	行业增速下滑 ……	地震 ……
经营事件	其它事件	外部评价	其它事件
客户大幅度流失 ……	客户举报 ……	客户负面评价 ……	美国竞选 ……

内部事件

企业信用变化事件		企业交易行为事件	
财务事件	担保事件	授信事件	产品事件
资产负债率骤增 ……	抵押物贬值 ……	挪用信贷资金 ……	工程进度不匹配 ……
	其它事件	账户事件	内部操作
	列入银监不良 ……	大额资金转出 ……	违规放贷 ……

风险预警事件分级

预警事件严重程度是判断外部事件是否触发风险预警的重要输入指标；预警事件严重程度需要基于历史数据及专家经验的反复调整过程，并最终及分档的形式体现，以便于管理应用



CONTENTS

金融风控模型

松鼠山科技 SQH Tech

输入

根据单一关联生成企业图谱

根据各种关联关系生成企业图谱

关联关系

股权关系

交易关系

担保关系

行业关系

……

传导

在关联关系图谱中，部分信用主体发生变化，该类变化沿着图谱网络以一定效能进行传导

传导前

A

B*3

C

D

E

F*4

输出

根据传导结果，以资产规模，违约逾期，营业收入等维度反映并输出企业传导后风险情况

传导后：信用主体风险情况

A*1

B*2

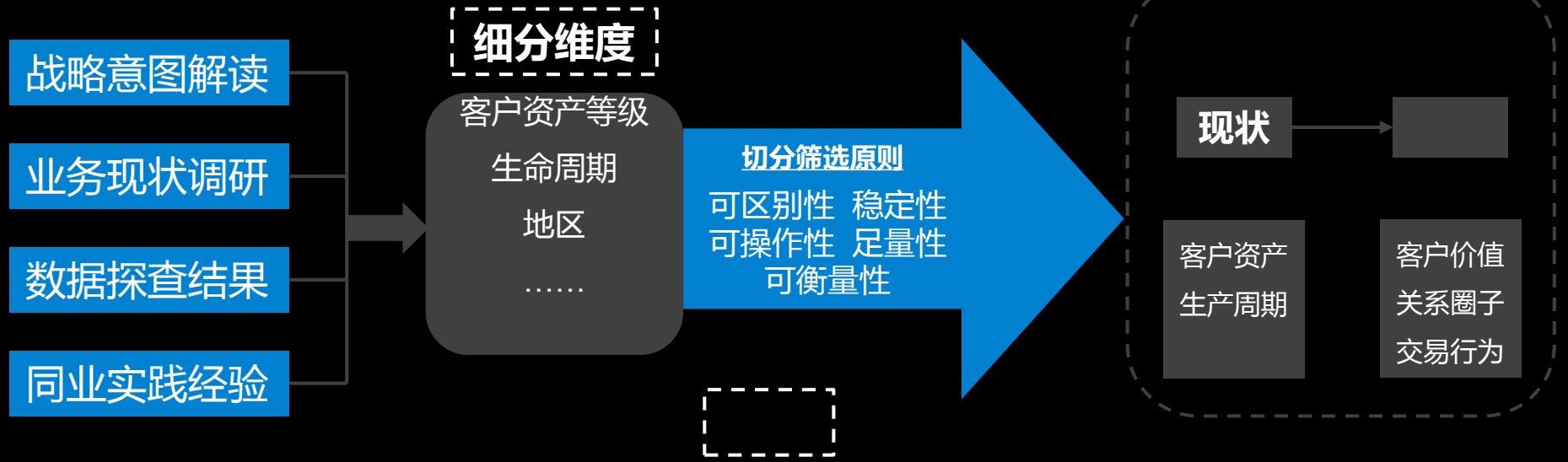
C*1

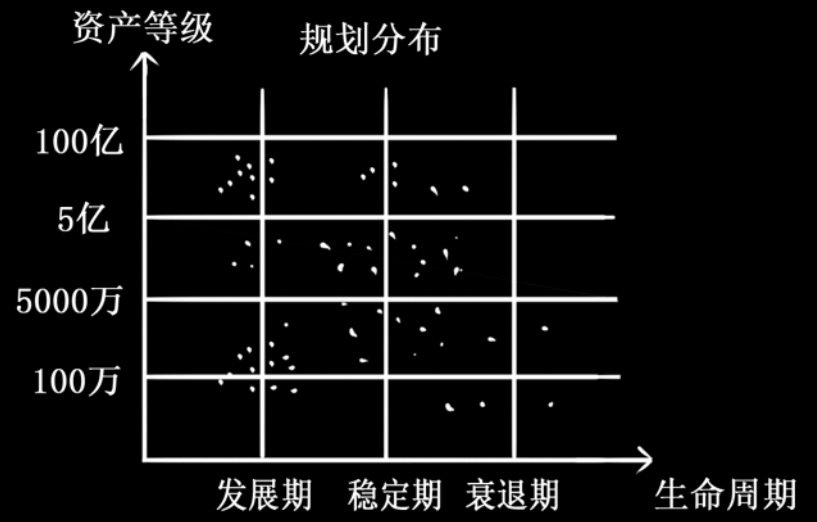
D*2

E*5

F*6

金融营销模型


战略意图解读
业务现状调研
数据探查结果
同业实践经验
细分维度
客户资产等级
生命周期
地区
……
切分筛选原则
可区别性 稳定性
可操作性 足量性
可衡量性
现状
客户资产
生产周期
客户价值
关系圈子
交易行为


资产等级
规划分布
100亿
5亿
5000万
100万
发展期
稳定期
衰退期
生命周期

我们的合作伙伴

Neo4j亚太区合作伙伴

中国金融行业总代理商

- 结合世界上最领先最流行的图形数据库产品，打造银行、PE/VC、保险、财富管理、证券等细分金融行业的专业知识图谱，建立起各实体关联图，并将其直接关联、间接关联的各种实体、概念相联系，帮助各参与者洞察市场脉搏。

- 目前为止，中国金融行业企业版用户（中国平安集团）的唯一服务商。

后端存储

松鼠山科技 SQH Tech

关系数据库
AgensGraph(PostgreSQL)

原生单机存储

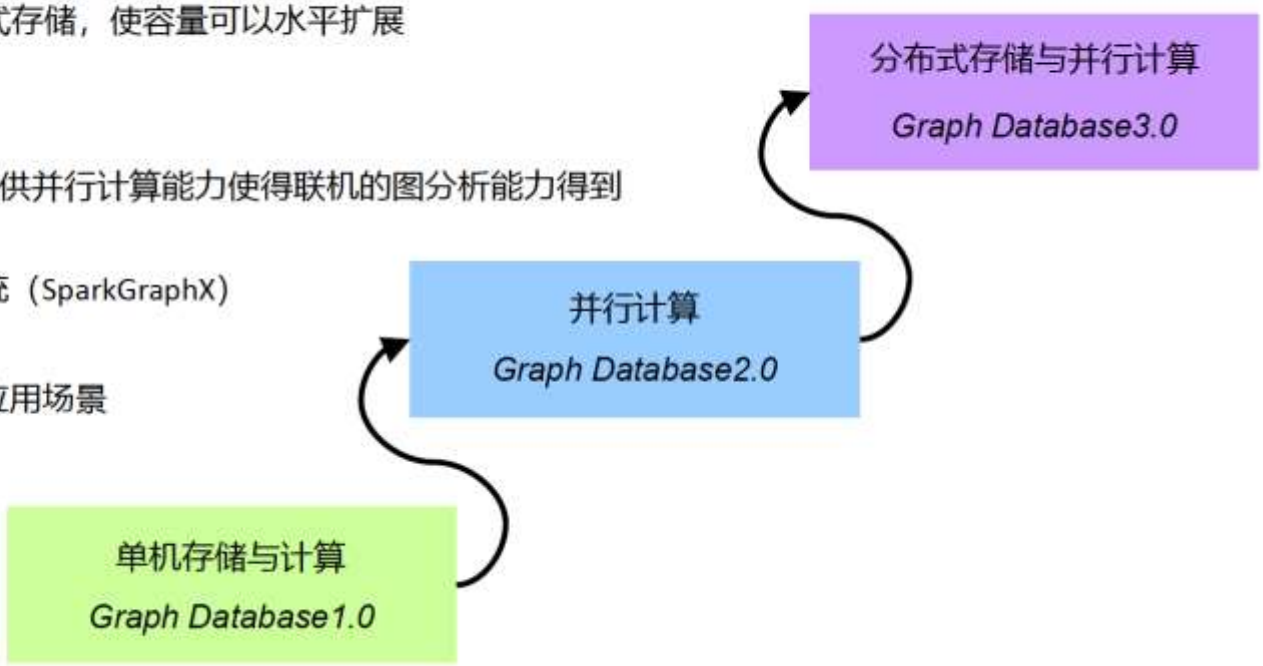
Neo4j

分布式存储
DSE(Cassandra)
JanusGraph/Titan

原生分布式存储
ArangoDB(Raft + RcoksDB)
Dgraph(Raft + Badger)
TigerGraph

发展趋势

- 分布式存储
  - Neo4j 存储受限于单机容量
  - 基于HBase/Cassandra或原生分布式存储，使容量可以水平扩展
- 计算并行化
  - JanusGraph 单点计算
  - ArangoDB / Dgraph / TigerGraph ,提供并行计算能力使得联机的图分析能力得到大幅度提升。
  - 并行计算能力不再局限于离线系统（SparkGraphX）
- 数据一致性（Raft）
  - 提供了更好的数据一致性，拓宽应用场景
- 底层存储引擎优化
  - HBase -> RocksDB -> Badger(SSD)
  - 大幅提升数据更新及查询能力




松鼠山科技
SQH Tech
感 谢 您 的 聆 听