人工智能行业应用
—Fintech

天云大数据： 邢建伟

BeagleData
天云大数据

获取机器智能像读书一样简单

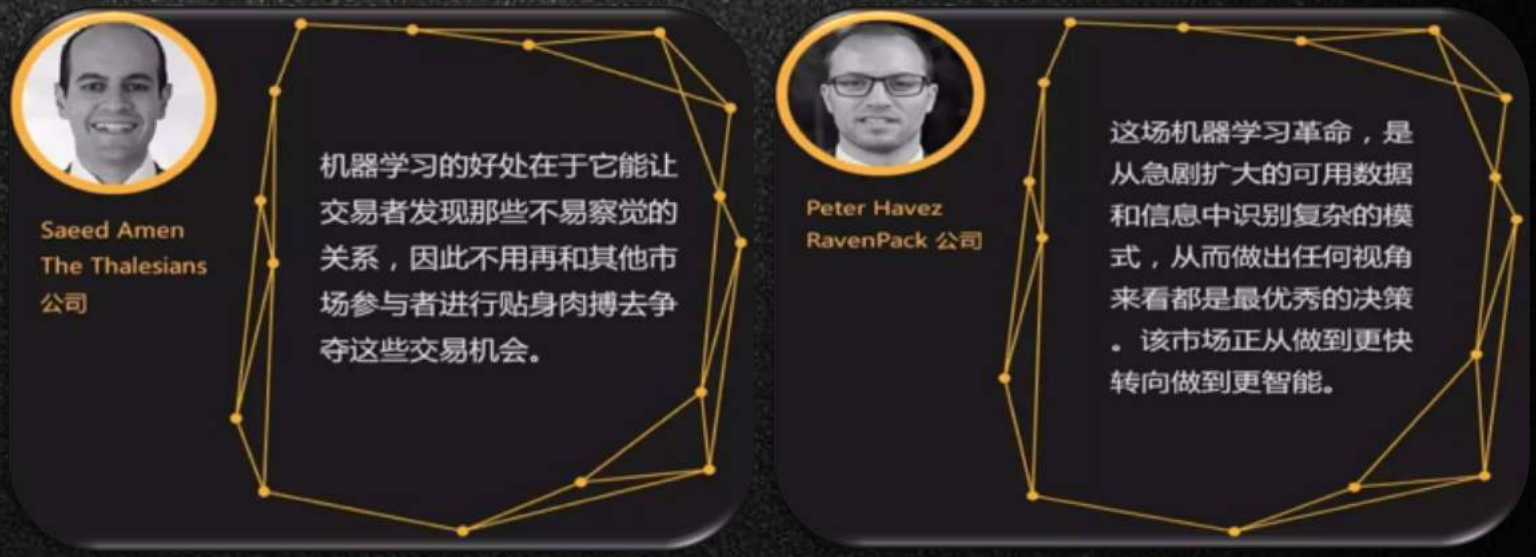
Get machine intelligence as simple as reading

FinTech

IT DT

规则流程驱动到智能数据驱动

机器的角色，从快速思维到智能思维



在过去的几十年里，计算机被广泛用于完成自动化任务，后者往往是被清晰的规则和算法描述的。如今，机器学习技术允许我们在难以精确描述规则的边界内完成同样的任务。

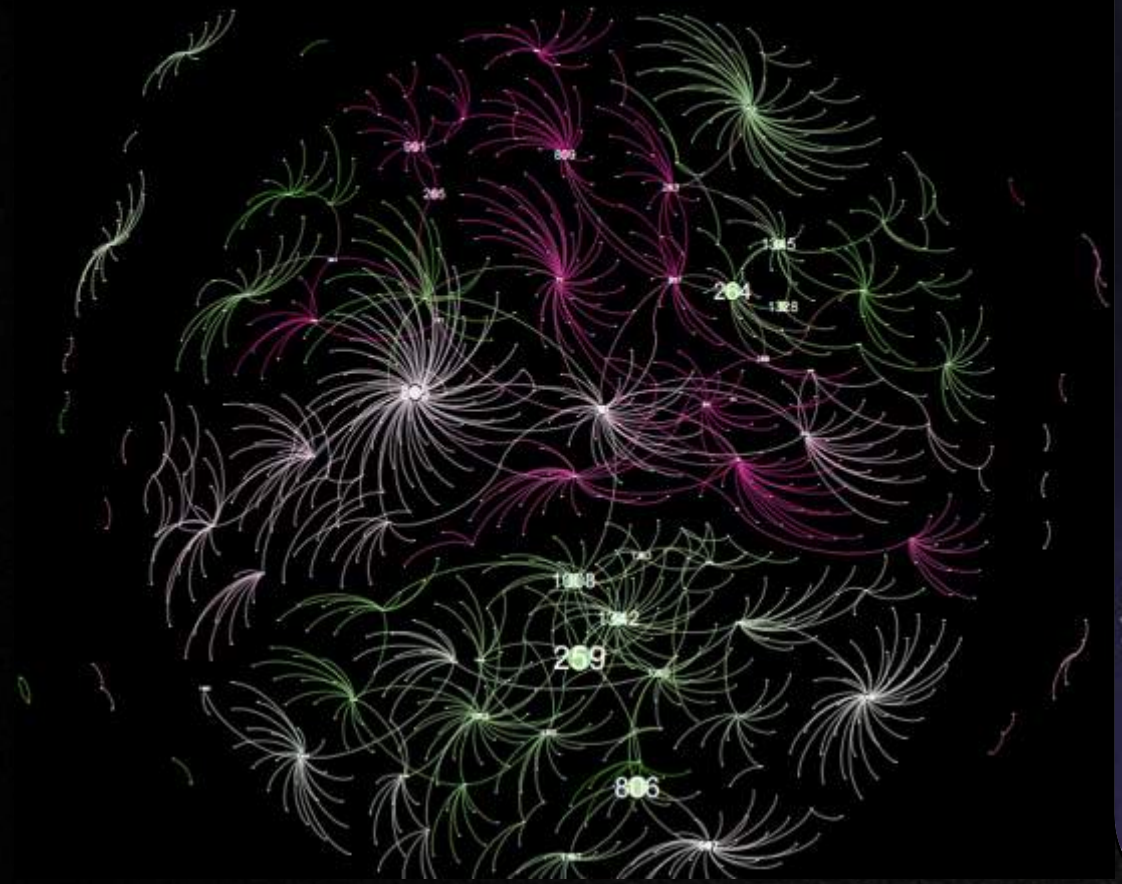
— 来源自亚马逊创始人杰夫·贝索斯（Jeff Bezos）2017年度致股东的公开信。

人类习惯抽象认知

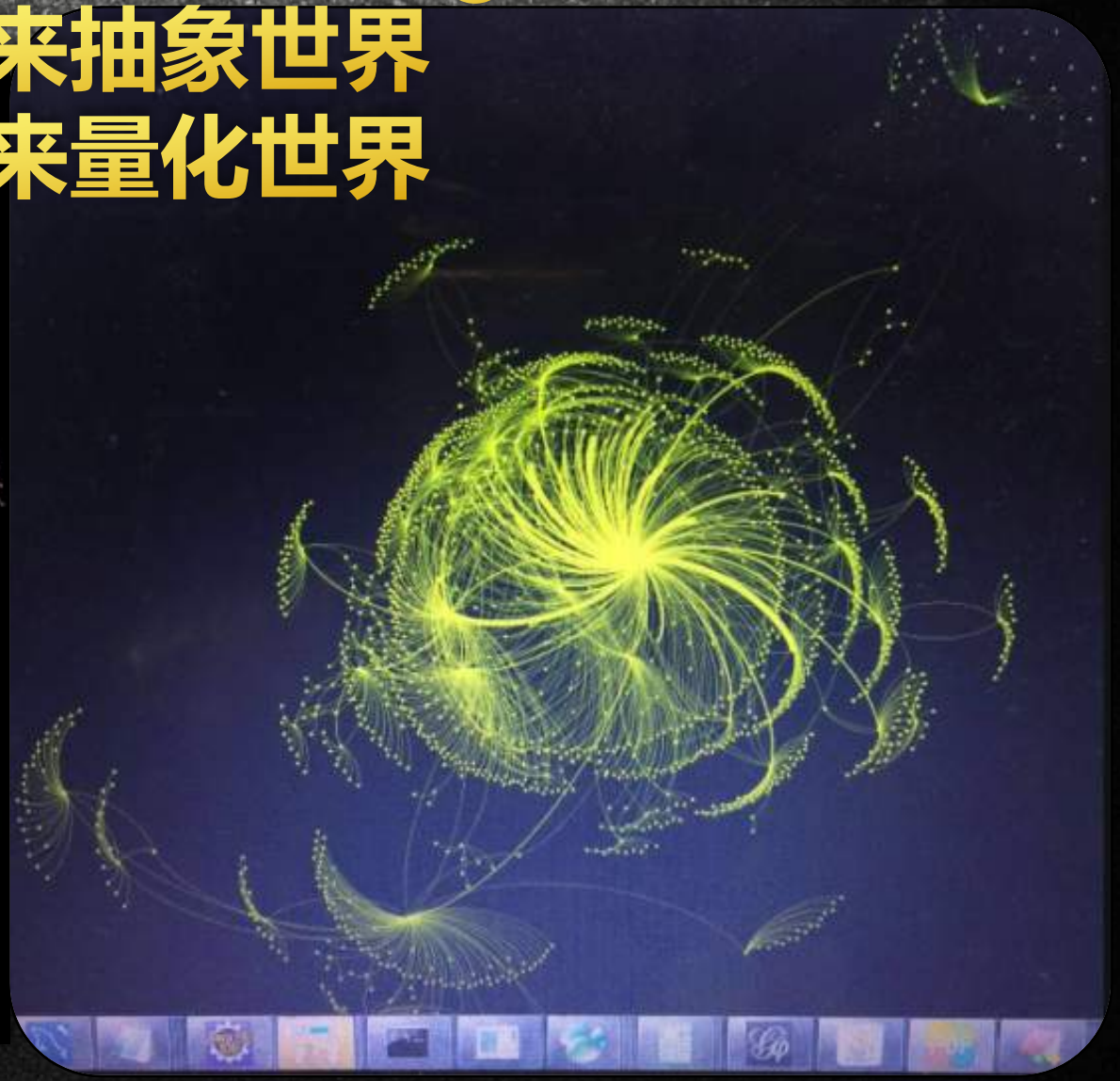
350年前，科学巨人牛顿用三个简约表达式，揭示了自然规律，客观抽象了传统参照系中的世界。

BeagleData
天云大数据

人类用简单来抽象世界
机器用复杂来量化世界



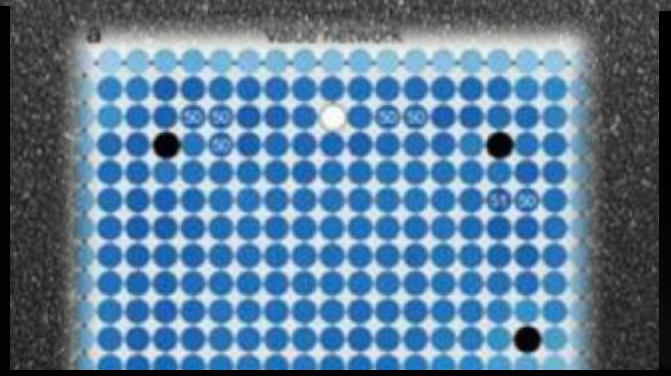
某互金理财产品的营销获客传播的复杂网络

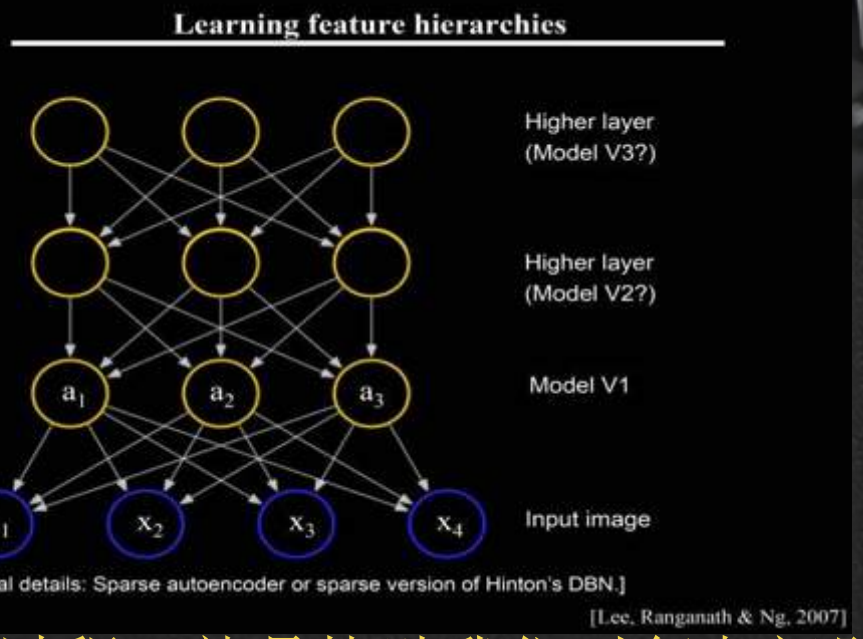


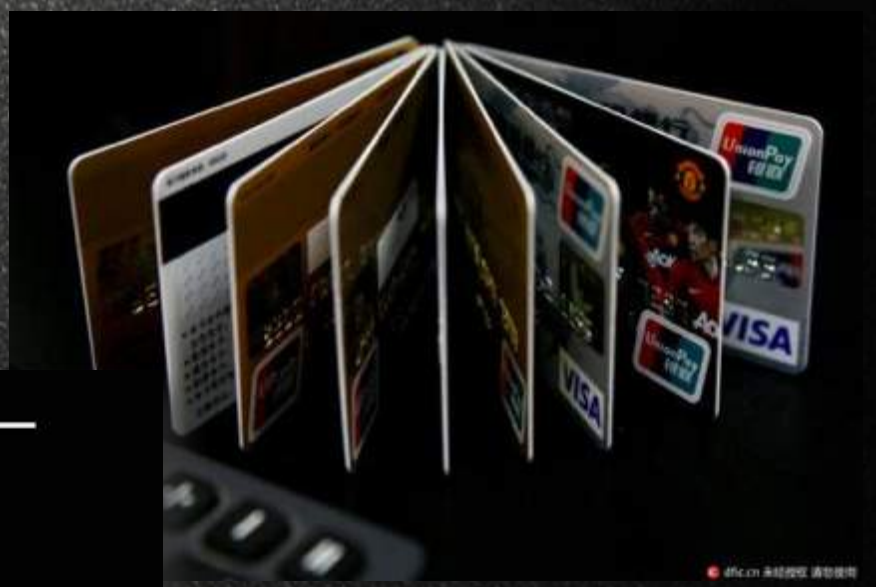
某保险公司的代理人成功销售的获客网络

人工智能依靠质朴的数学和超强的计算能力，还原了世界的复杂性。

如何用RGB像素色差等信号体系描述图片内容？







如何制订动态防范的欺诈规则？

深度学习的特征建立过程，就是协助我们对复杂问题描述的精确量化。

今天是笛卡尔的时代，等待着艾萨克牛顿的到来

笛卡尔为之后的牛顿准备了一个坐标系，使F=ma的推演成为经典。
今天我们可能不知道目的地像什么，但我们现在知道如何绘制一张地图。

Yann LeCun: 1960 - Current
Father of Convolutional Neural Networks

René Descartes: 1596 - 1650
Father of Cartesian Geometry

- 物理学上牛顿之前的日子。许多聪明的科学家能够使用数学来预测物体的运动，曾经聪明的笛卡尔教会我们如何将我们的物理思维考虑到坐标系统中。Yann LeCun（深度学习之父其中之一）就是一个现代的笛卡尔，开创性的工作是指日可待。ConvNets思想框架就像是一个必备的坐标系统。

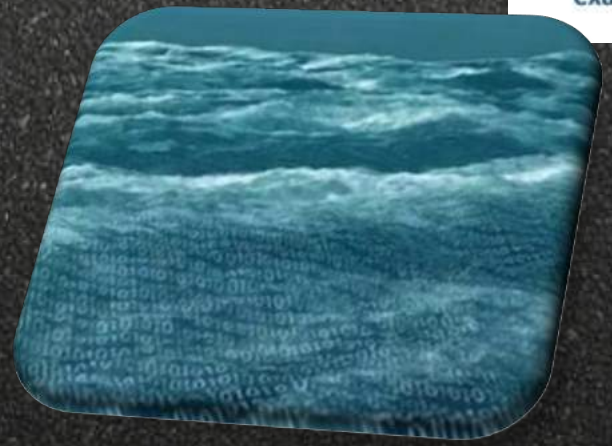
开放的 B 价值远大于性感的 A

年份	人工智能的突破	年份	数据集（首次可用）	年份	算法（首次提出）
1994	人类自发的语音识别	1991	华尔街日报文章与其他文本	1984	隐马尔可夫模型
1997	IBM深蓝打败Garry Kasparov	1991	70万国际象棋大师赛，又称“扩展的书”	1983	主变量搜索算法(Negascout算法)
2005	谷歌阿拉伯语-中文-英语翻译软件	2005	1.8万亿符号的搜索，基于谷歌网站与新闻网页	1988	统计机器翻译算法
2011	IBM Watson机器人成为世界“危险”节目冠军	2010	860万文件的上传，基于维基百科，维基词典引用以及古登堡计划	1991	集合训练算法
2014	GoogLeNet在ILSVRC达到人类同等水平	2010	150万标签的图片1000项目分类的视觉数据库	1989	卷积神经网络算法
2015	谷歌deepmind游戏水平达到与人类同等水平	2013	超过50种超高难度雅达利游戏的学习环境数据集	1992	Q-learning算法
平均突破年限：		3 年		18 年	

数据湖—数据融合方案

DATALAKE

> Think of a Data Mart as a store of bottled water—it's cleansed, packaged, and structured for easy consumption. The Data Lake, meanwhile, is a large body of water in a more natural state. The contents of the Data Lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in or take samples.
>
> James Dixon
> *Pentaho CTO and creator of the term Data Lake*



2010年，Pentaho 的 CTO James Dixon介绍了“数据湖Data Lake”的概念并且影响至今。Dixon认为数据湖是一种数据仓库的架构，他如此描述：

“如果你把数据市集看做一个贩卖瓶装水的商店——贩卖清洁的、被包装过的、标准化的可被轻易销售的产品—数据湖则是一个在更自然状态的大体量水体。数据的不同支流源源不断的丰富着湖泊，而各种各样的用户都可以到湖水中检验、潜水，或者抽取样品。”

数据湖

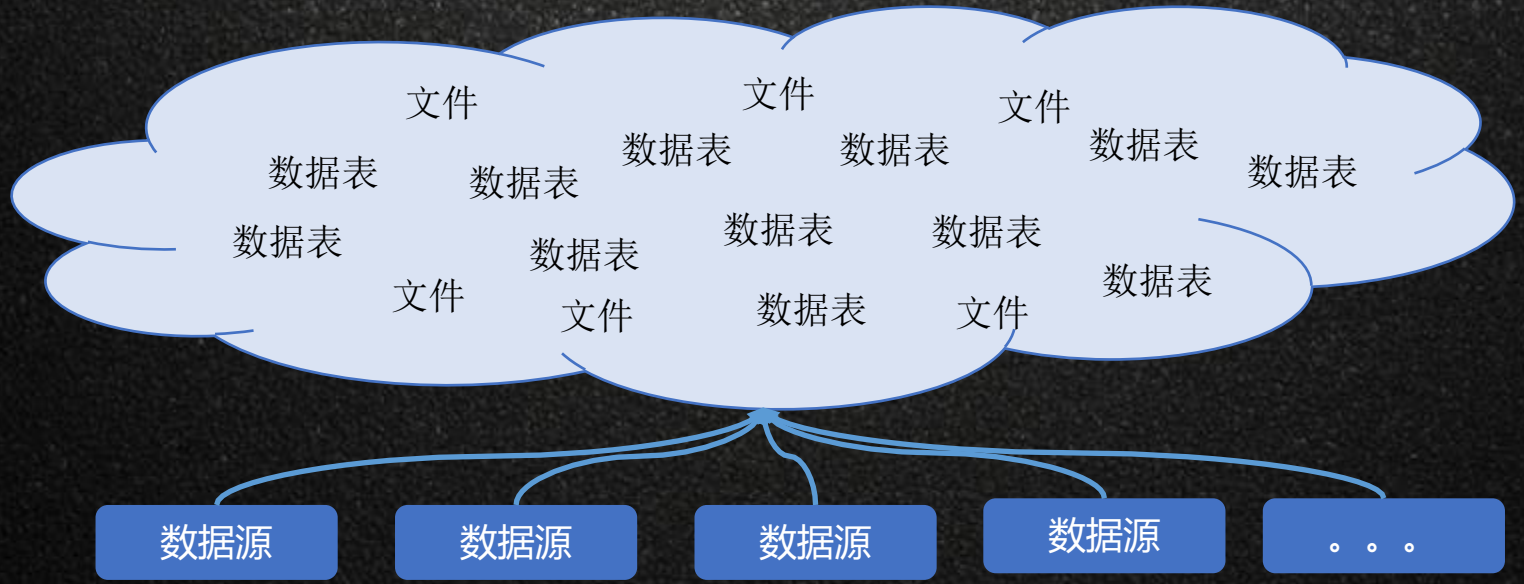
数据湖-元数据管理

- 我要的数据有哪些？
- 我要的数据在哪里？

数据库/表/字段
存放目录/文件名

如果要把所有的数据的元数据定义全部梳理出来，知道元数据的含义，将会面临巨大的工作量。

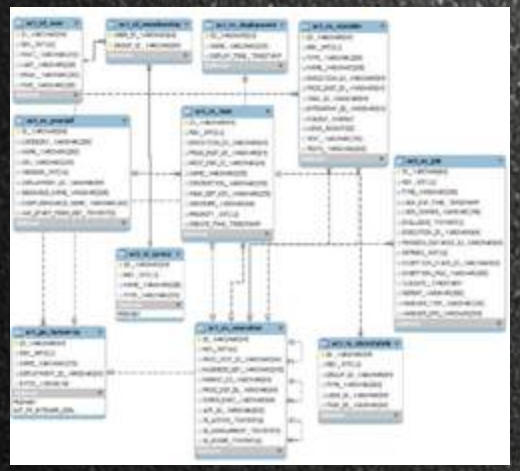
面临数千甚至数万的元数据信息的治理

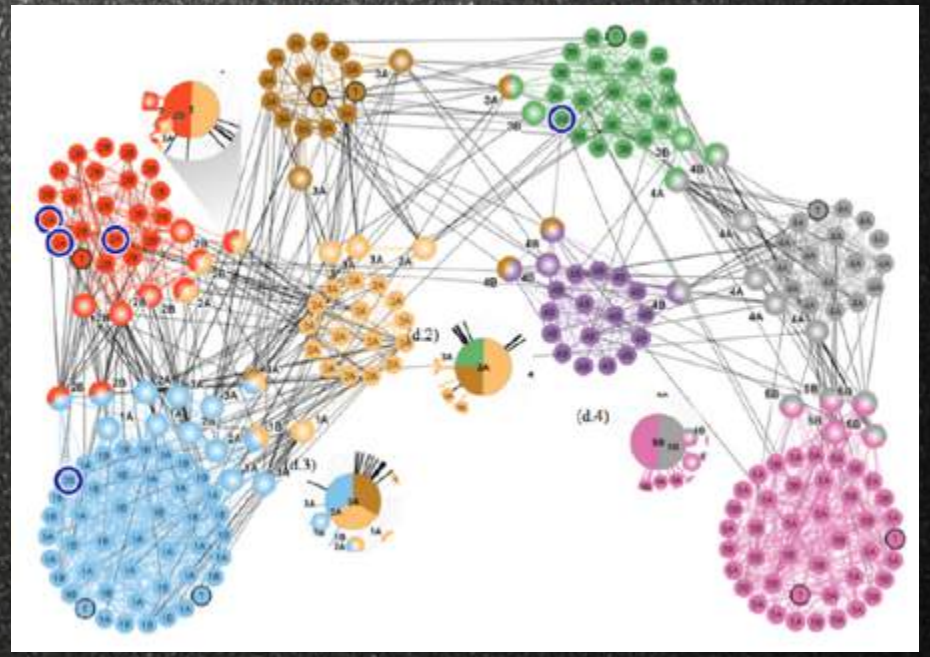


数据湖-标签索引

- 为每一张表和每一个字段赋予标签含义，通过标签索引相关字段，未来通过标签检索，敏捷定位到所需字段项以及字段数据。

数据湖--数据管理





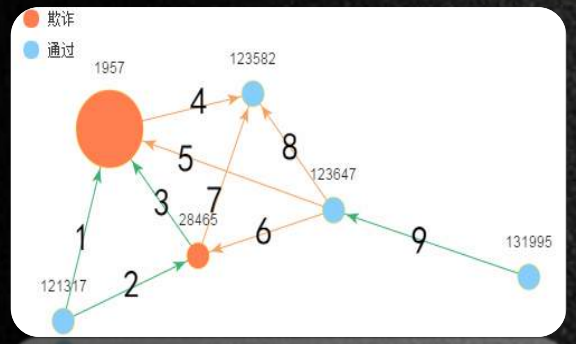
```
select *
from  bank_statements as a
here a.category in ('bop', '通付通')
nd exists(select 1
   from (
      select
         ba_merchant_id
         case category
            when 'bop' then ba_income_money
            when '通付通' then ismull( ba_income_money , ba_expenditure_money )
            end as amount
      from tb
      where category in ('bop', '通付通')
      ) as b
   where b.ba_merchant_id = a.ba_merchant_id
   and b.amount != case a.category
            when 'bop' then a. ba_income_money
            when '通付通' then ismull(a. ba_income_money , a. ba_expenditure_money )
            end
   )
```

将数千张表字段构建关联起来，形成网络关联图谱，探查元数据逻辑联系，实现对数据结构的透彻了解和灵活掌控。

贷前—信用申请反欺诈

通过分析银行信用卡的“通过信用卡”信息和“欺诈信用卡”信息，找到注册信息中包含的关系，同时对关系信息统计分析，计算相关指标，然后通过统计分析的结果构建社交网络，最终支撑欺诈用户发现。

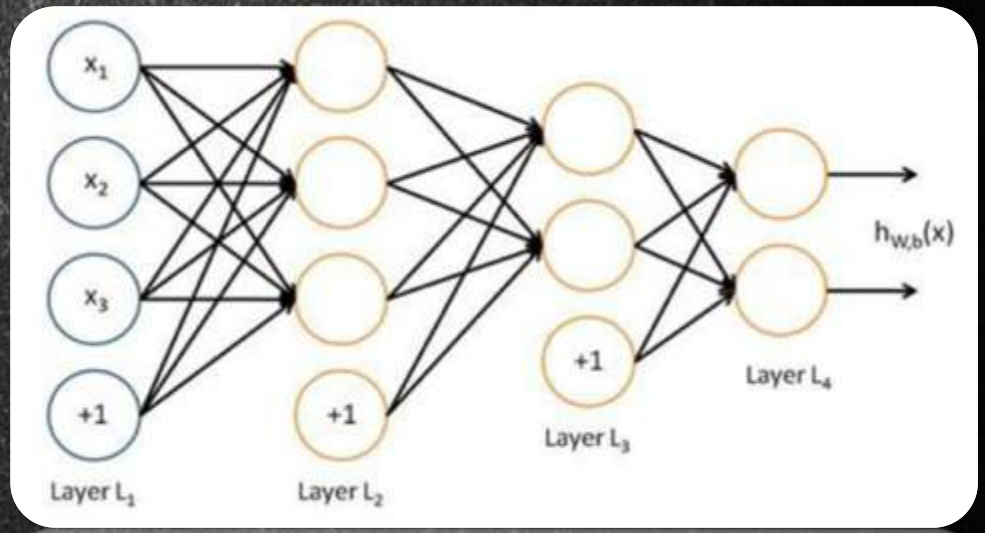
申请进件的关联特征



基础金融属性

年龄	
年收入	
学历	
职位	
区域	
职业	
第三方信用卡	
。。。	

深度学习网络



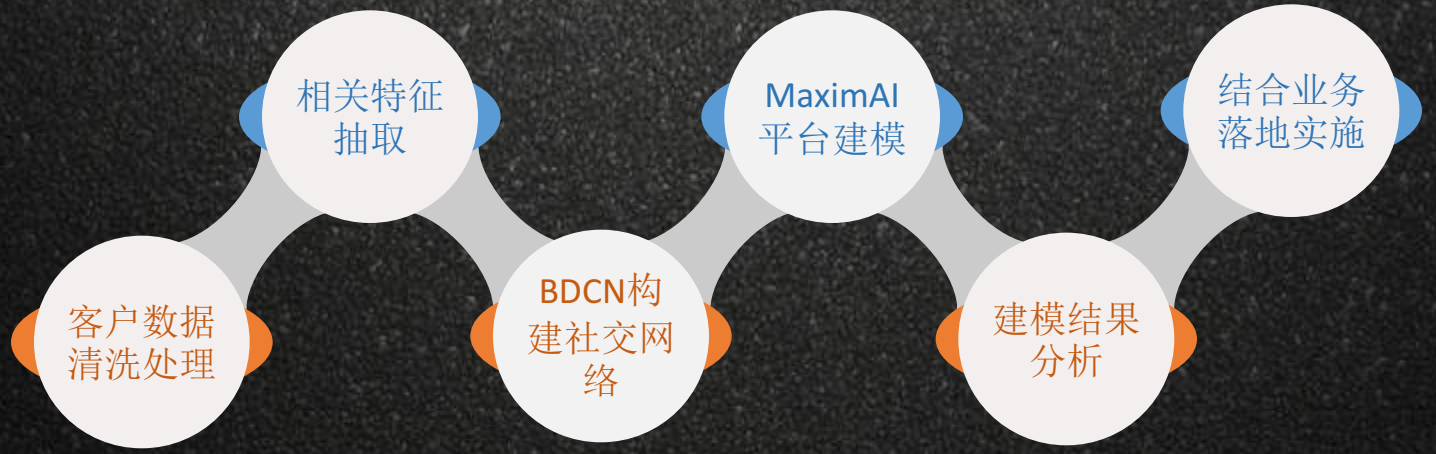
Combine	AUC	Accuracy	Precision	Recall	F1-measure	Lift
LR	0.85	0.86	0.90	0.90	0.90	1.25
DL	0.90	0.87	0.90	0.92	0.91	1.32
RF	0.93	0.88	0.93	0.91	0.92	1.35

欺诈识别具体实施步骤

数据转换
社交特征抽取

随机森林
梯度提升
深度学习

串联模式
结合原有审批流程



客户基本申请信息
客户社交属性信息

抽取一阶度、
一阶欺诈比等特征

AUC/KS等
评估模型准确性

复杂网络特征提取建模

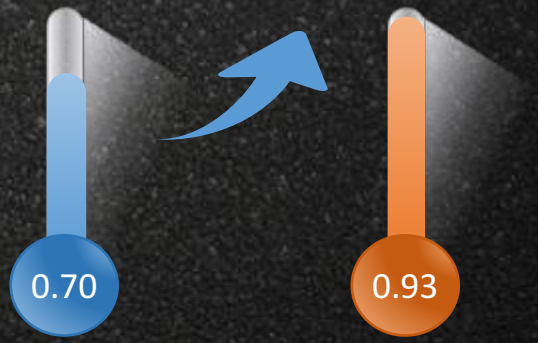
原始特征

- 年龄
- 年收入
- 学历（数值化）
- 职位（数值化）
- 手机号
- 单位电话
- 电子邮箱
- ……

建立复杂网络

网络特征

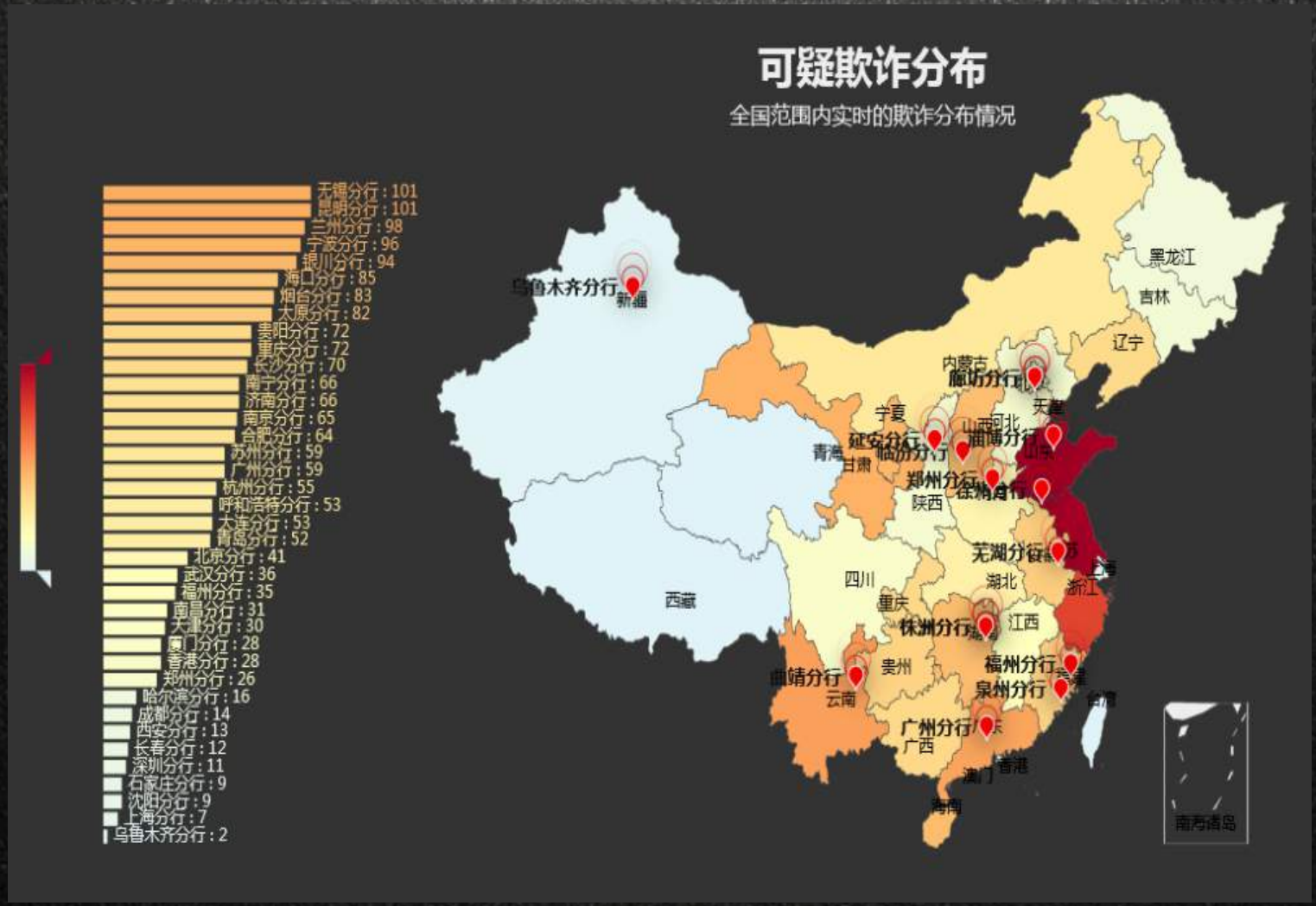
- 一阶度，一阶欺诈数，一阶欺诈比
- 二阶度，二阶欺诈数，二阶欺诈比
- 最短路径（距离欺诈节点）
- ……

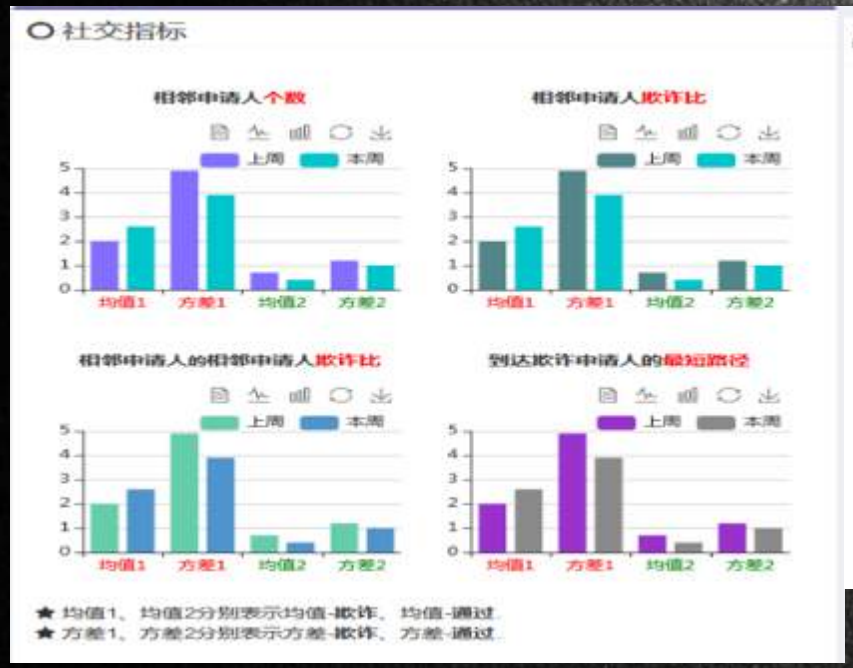


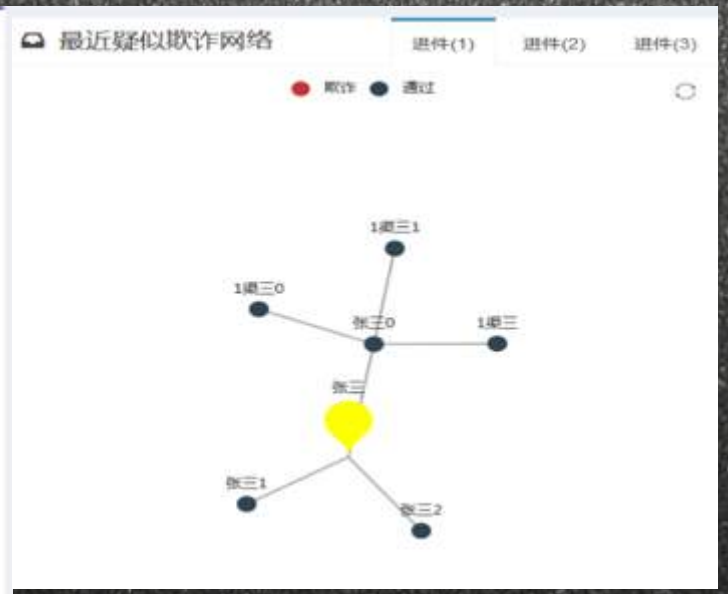
仅使用申请人原始基础信息的AUC值

加入网络特征属性后的AUC值

全国欺诈分布图


可疑欺诈分布
全国范围内实时的欺诈分布情况
无锡分行：101
昆明分行：101
兰州分行：98
宁波分行：96
银川分行：94
海口分行：85
烟台分行：83
太原分行：82
贵阳分行：72
重庆分行：72
长沙分行：70
南宁分行：66
济南分行：66
南京分行：65
合肥分行：64
苏州分行：59
广州分行：59
杭州分行：55
呼和浩特分行：53
大连分行：53
青岛分行：52
北京分行：41
武汉分行：36
福州分行：35
南昌分行：31
天津分行：30
厦门分行：28
香港分行：28
郑州分行：26
哈尔滨分行：16
成都分行：14
西安分行：13
长春分行：12
深圳分行：11
石家庄分行：9
沈阳分行：9
上海分行：7
乌鲁木齐分行：2
乌鲁木齐分行
新疆
黑龙江
吉林
辽宁
内蒙古
临沂分行
宁夏
天津
河北
延安分行
淄博分行
青海
甘肃
临沂分行
郑州分行
陕西
芜湖分行
四川
湖北
浙江
西藏
重庆
株洲分行
江西
福州分行
贵州
南靖分行
泉州分行
云南
广州分行
广西
香港
南海诸岛


社交指标
相邻申请人个数
相邻申请人欺诈比
相邻申请人的相邻申请人欺诈比
到达欺诈申请人的最短路径
上周
本周
均值1
方差1
均值2
方差2
★ 均值1、均值2分别表示均值-欺诈、均值-通过
★ 方差1、方差2分别表示方差-欺诈、方差-通过

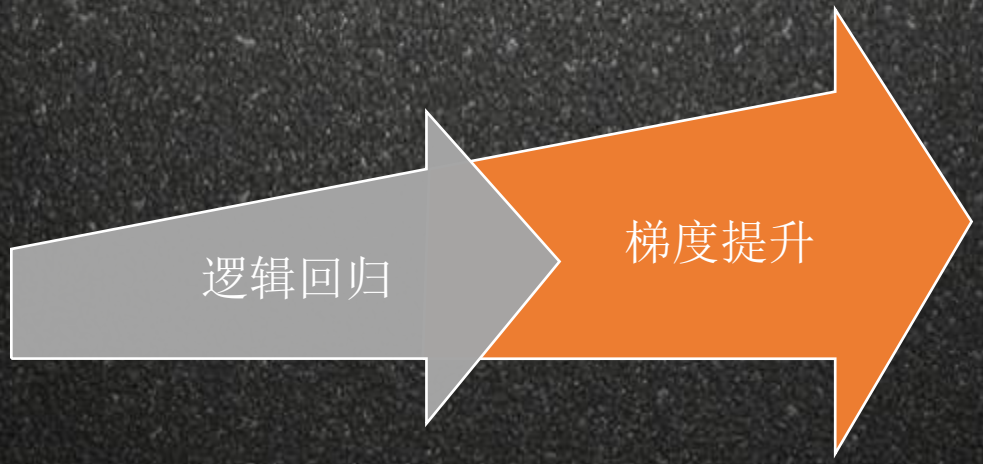

最近疑似欺诈网络
进件(1)
进件(2)
进件(3)
欺诈
通过
1逾三1
1逾三0
张三0
1逾三
张三
张三1
张三2


分析明细查询
社交网络查询
查询结果


分析明细查询
社交网络查询
查询结果
信用卡客户：王某某，男，23岁，销售经理，身份证号：110226201703241113，申请时间：

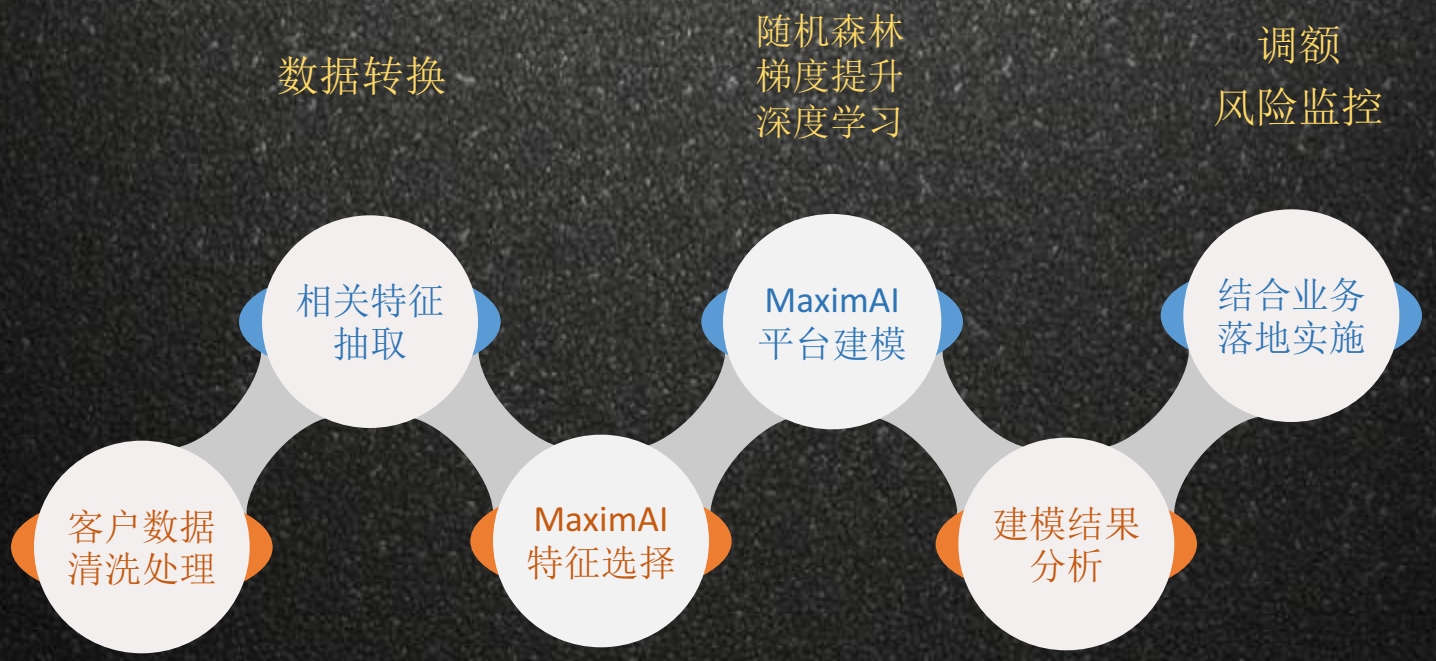

个人数据
风险概率
43%
53%
53%
53%
个人社交网络
基础指标信息
姓名 王某某
年龄 23
学历 本科
年收入 100000
身份证号 131123456789012345
手机号前三位 131
单位电话前三位 131
社交指标信息
一阶度总数 10
二阶度总数 20
一阶欺诈比 25%
二阶欺诈比 21%
最短路径 1

行为风险评分的新方法——通过梯度提升模型建模



- 传统的行为评分模型主要使用逻辑回归模型
- 使用梯度提升模型能提升模型的性能

某分期产品行为评分卡

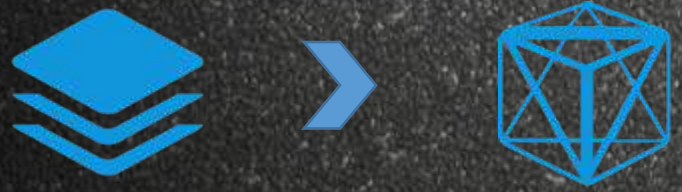


客户静态申请信息
客户动态行为信息

静态属性信息
动态属性信息

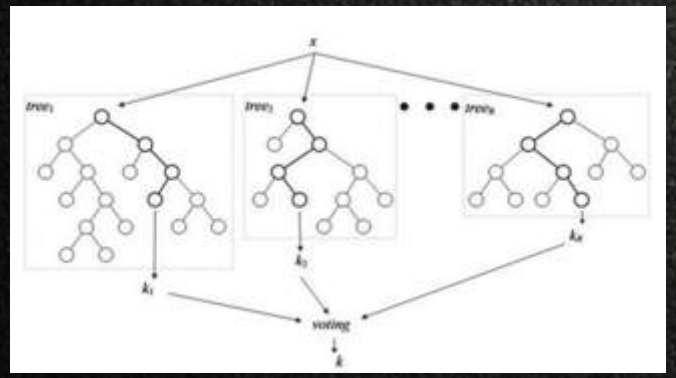
AUC/KS等
评估模型准确性

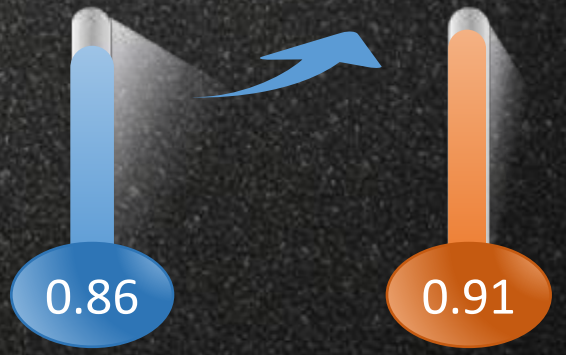
提升评分模型性能



原始无变换特征：
- 客户基本属性
- 原始行为数据
- 变换行为数据
- 人行信息数据

模型：
- **梯度提升**
- 随机森林
- 深度学习

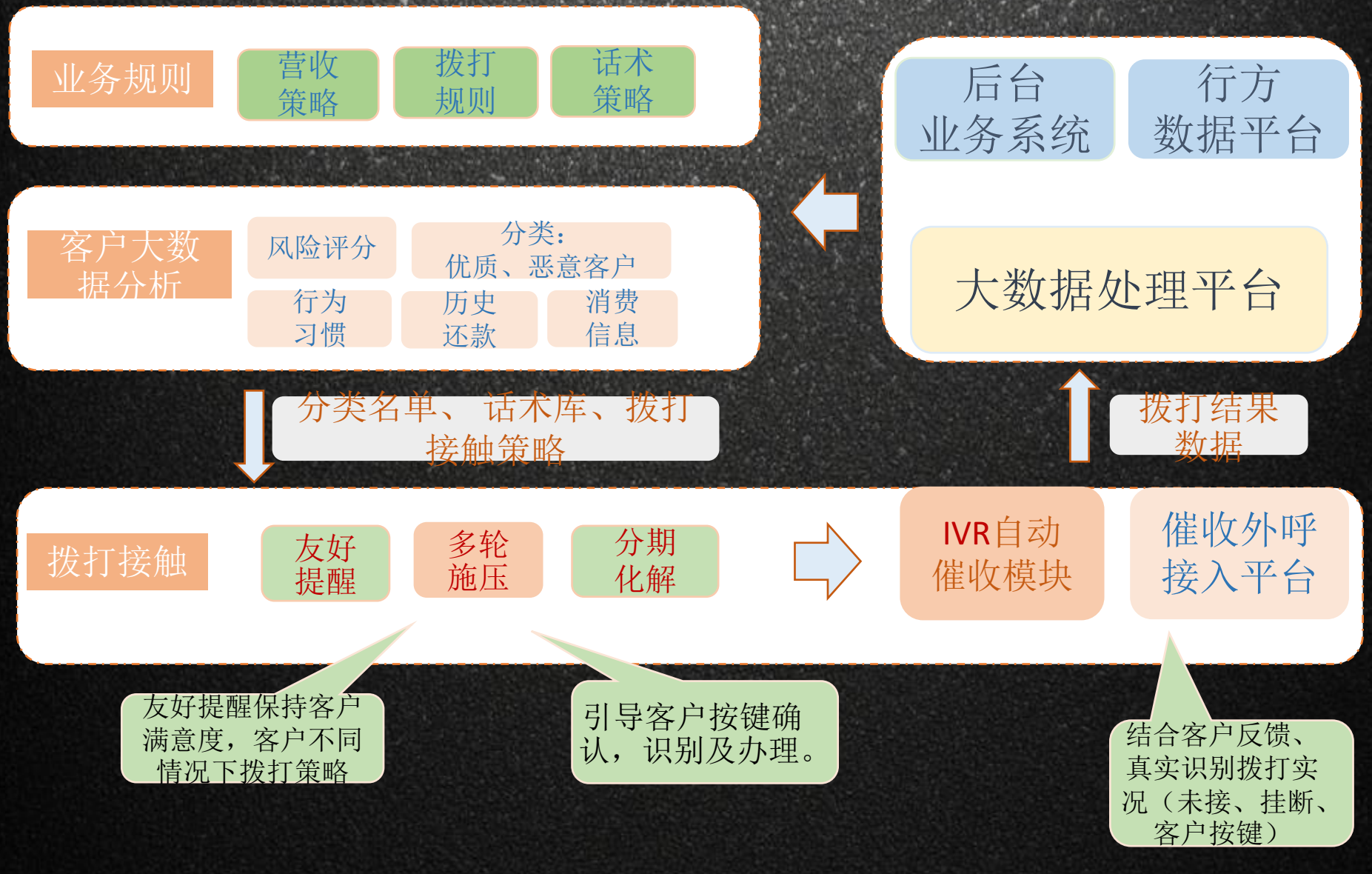




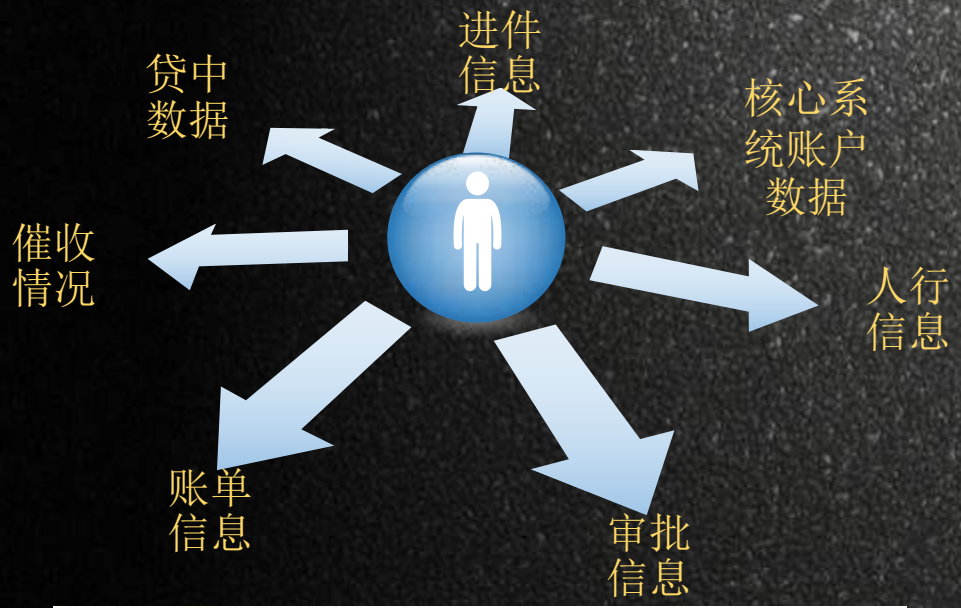
使用逻辑回归
模型的AUC值

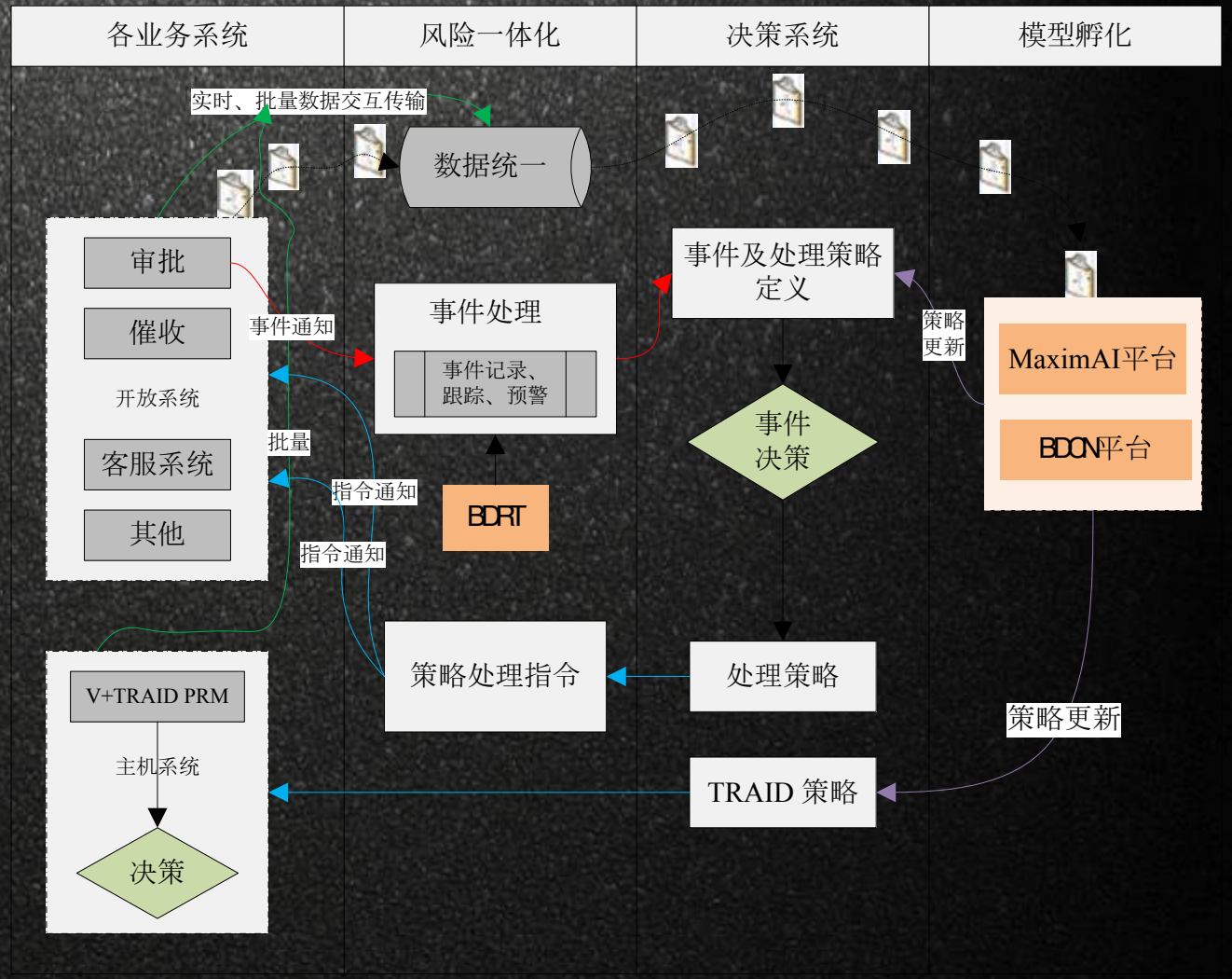
使用梯度提升
模型的AUC值

智能自动催收


业务规则
营收策略
拨打规则
话术策略
后台业务系统
行方数据平台
客户大数据分析
风险评分
分类：优质、恶意客户
行为习惯
历史还款
消费信息
大数据处理平台
分类名单、话术库、拨打接触策略
拨打结果数据
拨打接触
友好提醒
多轮施压
分期化解
IVR自动催收模块
催收外呼接入平台
友好提醒保持客户满意度，客户不同情况下拨打策略
引导客户按键确认，识别及办理。
结合客户反馈、真实识别拨打实况（未接、挂断、客户按键）

风控系统平台总体流程图





融合Algorithm Bigdata Cloud

MaximAI企业级人工智能平台产品

融合计算能力：

从并行计算到分布式计算的创新

Scala分布式程序的算法代码重构，充分发挥SPARC/Alluxia内存计算能力。

融合在线数据：

从流程驱动到数据驱动的创新

数据无需在生产系统和挖掘系统间抽取离线，实时的全量数据建模

融合业务价值：

从零到一的创新

从业务问题定义到前沿算法模型反复迭代，最终体现商业价值化的模型，可以在平台中发布、分享和继承。业务创新可以规模化复制。

MaximAI企业级人工智能平台产品（续）

FreeCoding

采用完全界面化的操作
用户无需任何编程背景，
也可轻松使用数据挖掘技术

Subscription

通过REST接口整合、订阅算法包和数据分析模版
面向高阶用户，自主编写Spark Scala,R,Python代码

向导服务

订阅服务

Scala

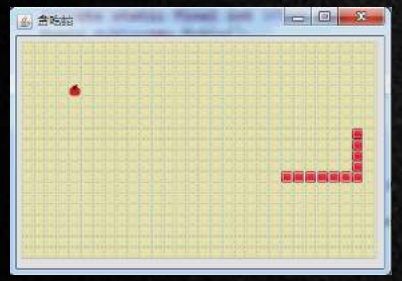
python

R

轻AI的前沿算法民主化

AI的平台化，催生Fintech

移动互联网前夜



开放平台应用开发

通讯科技巨头专利科技

Android/IOS平台化，屏蔽了技术复杂度，推动了移动互联网的繁荣。AI平台的兴起，将释放巨大AI潜能和催生更广阔人工智能市场。例：Google Tensorflow, Facebook FBLearner, MS CNTK,腾讯Angel, 天云MaximAI.

获取机器智能 像读书一样简单