CHINA DATA ANALYST SUMMIT

CDA 数据分析师 www.cda.cn

BI to AI 演进路径

规则流程驱动到智能数据驱动

演讲人：天云大数据 雷涛

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

跨界互联 数聚未来

第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT

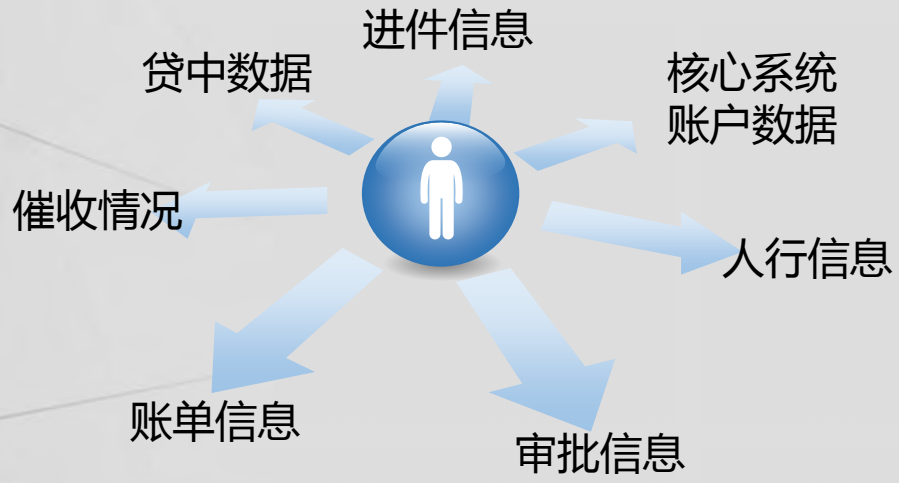
北京 中国大饭店 2017.07

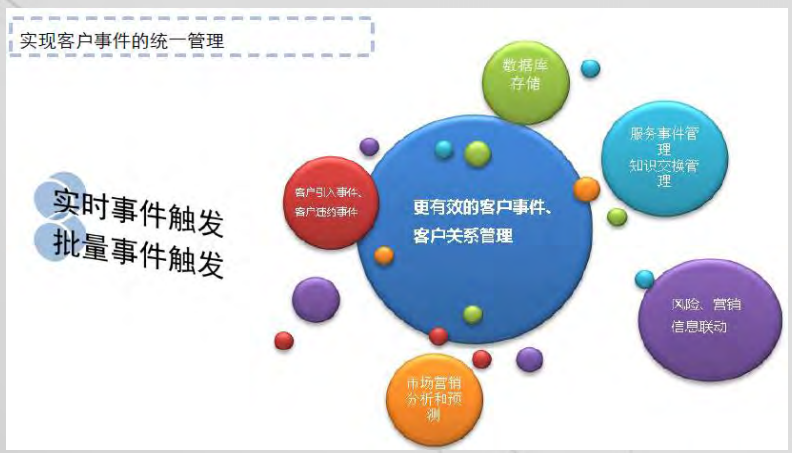
在过去的几十年里，计算机被广泛用于完成自动化任务，后者往往是被清晰的规则和算法描述的。如今，机器学习技术允许我们在难以精确描述规则的边界内完成同样的任务。

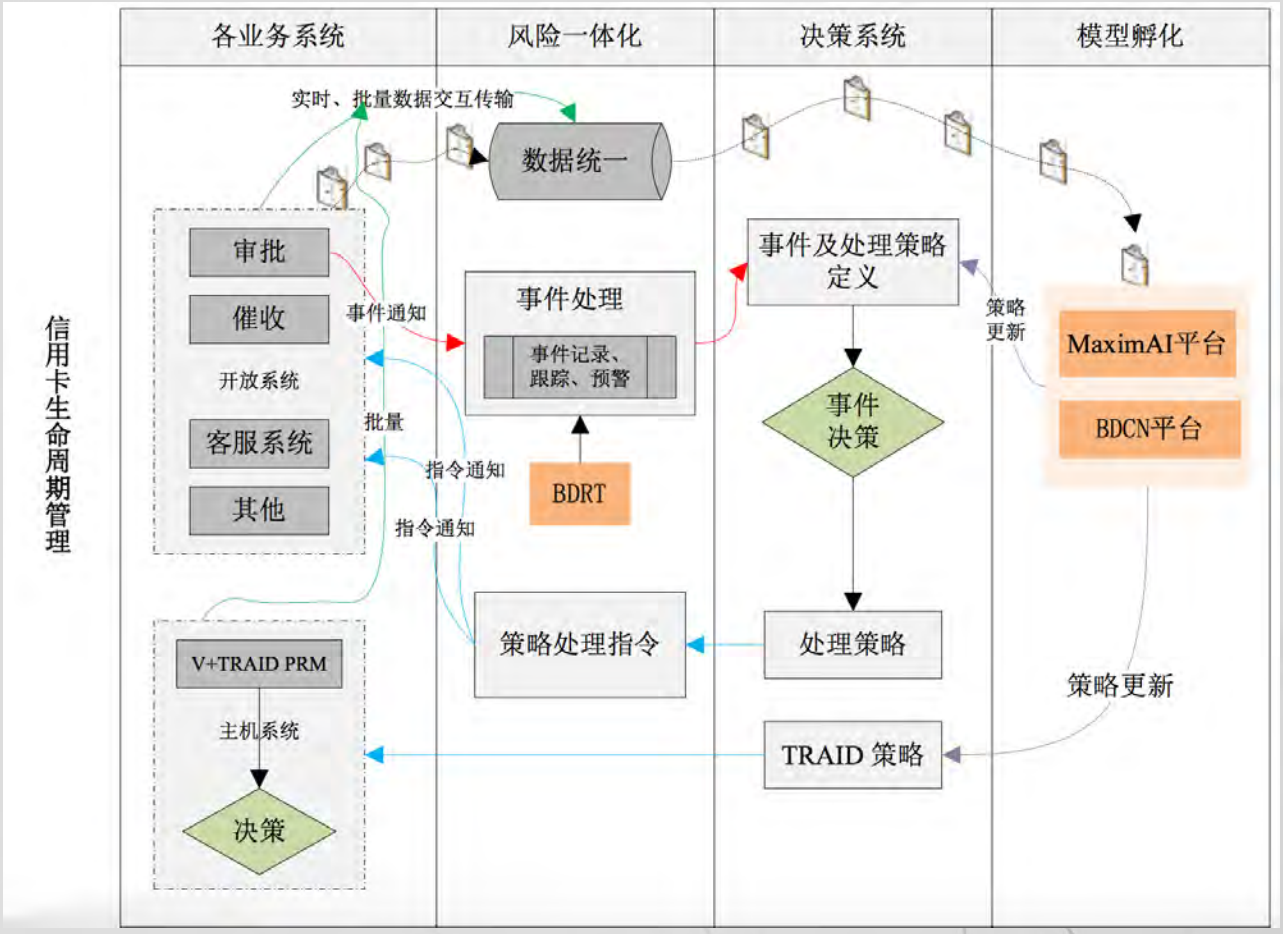
— 来源自亚马逊创始人杰夫·贝索斯（Jeff Bezos）2017年度致股东的公开信。

离线抽样 to 在线全量

风险一体化系统——实现客户数据与事件的统一管理



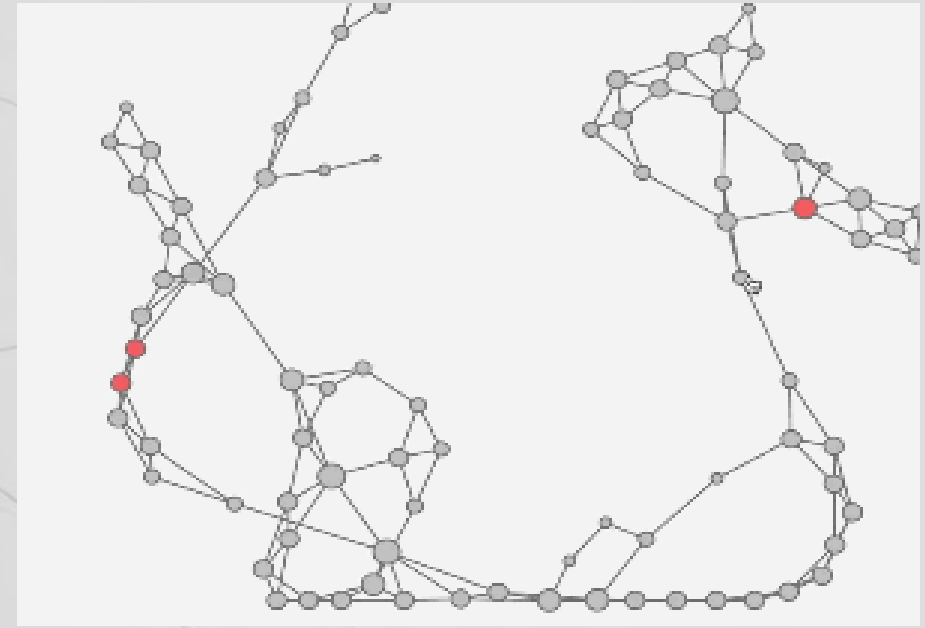


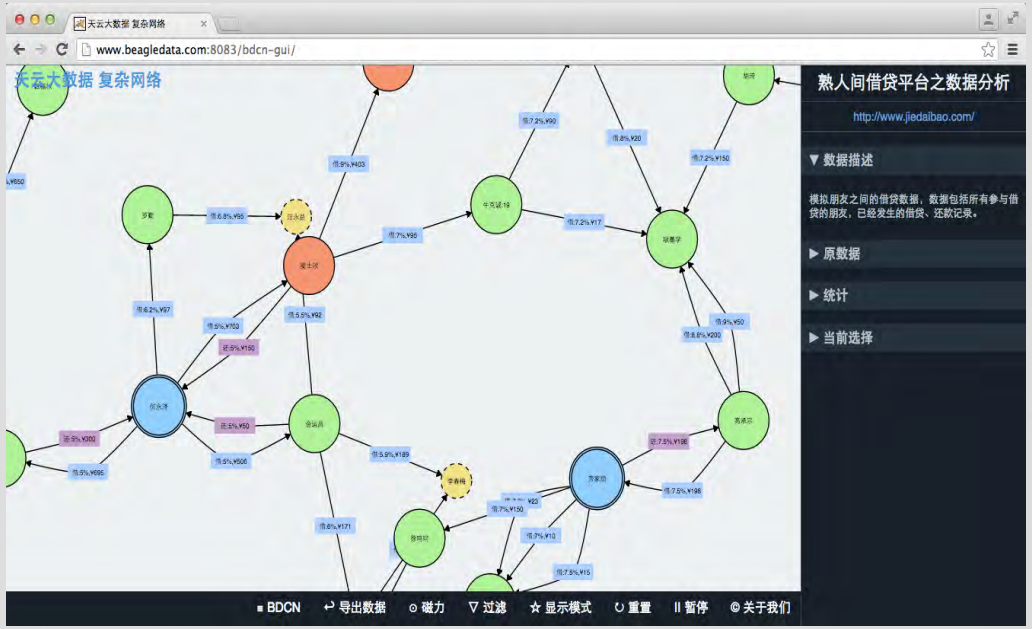


静态个体 to 动态关联

复杂网络风险量化模型

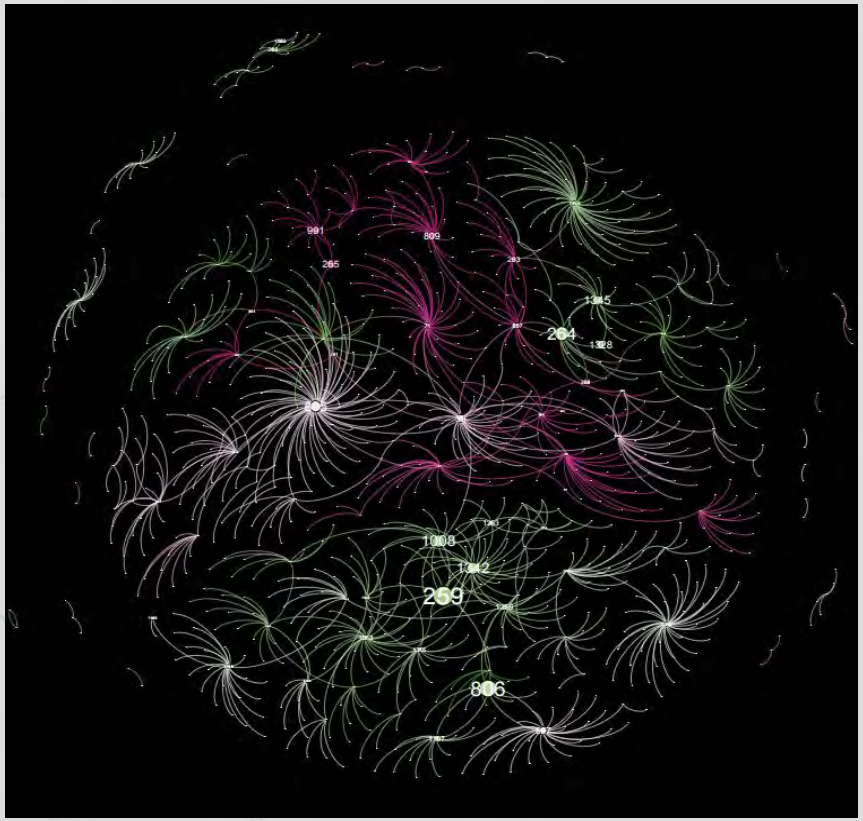
强调事物'联系'与全局量化的数据结构，主要用它来解决企业大量沉淀的结构化数据，使用向量矩阵替代数据库表结构计算。



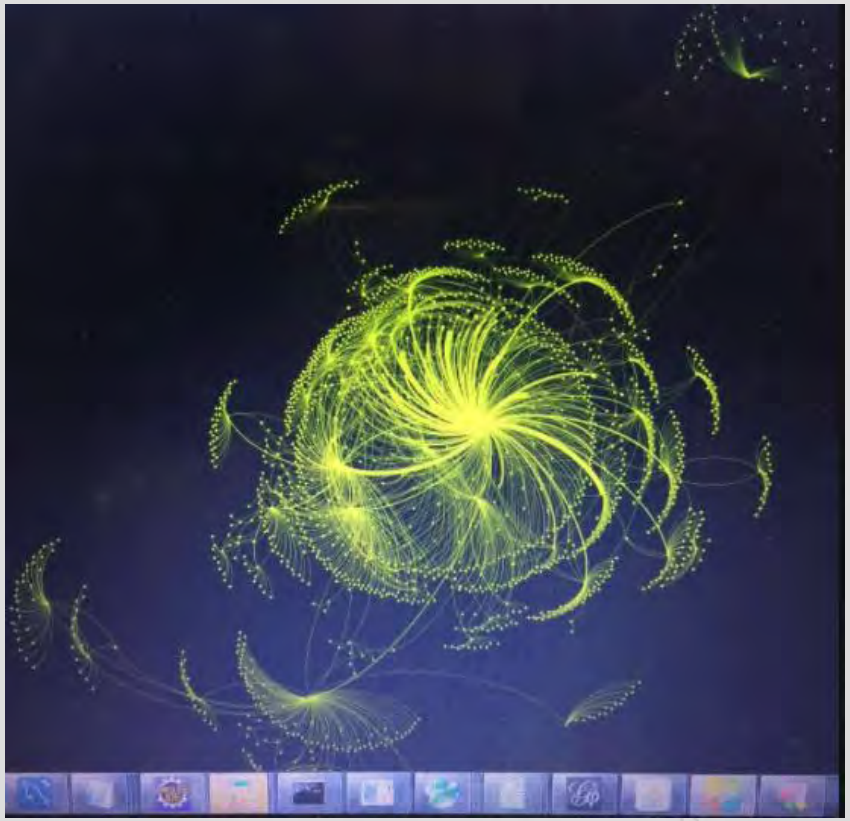


- 风险共同体：投资一致行动人，循环担保，重复抵押
- 社交P2P借贷
- 流动资本计量
- 风险种子揭示

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

统计评分 to 机器学习

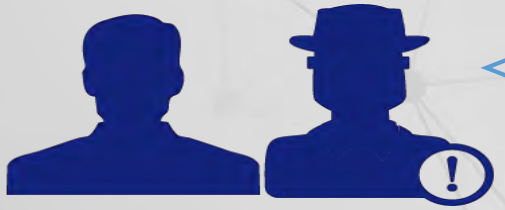
传统申请欺诈风险评分模型

验证如下信息：

- 填写地址与个人征信信息地址不符
- 填写地址在个人征信信息里第一次存档时间小于90天
- 填写地址在个人征信信息里仅为新信用账户所用
- 个人征信信息地址为高风险地址
- 个人征信信息地址为非住宅地址
- 个人征信信息地址曾有欺诈活动的记录
- ……

欺诈分子盗取、仿冒他人身份信息进行申请

人行征信系统难以覆盖



黑名单客户占比2%

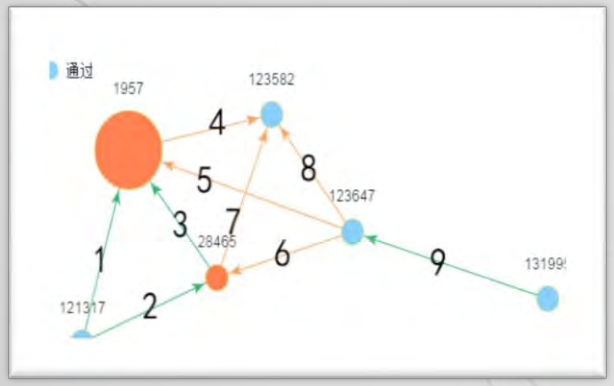
弱特征
信息量不足

效率较低

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。

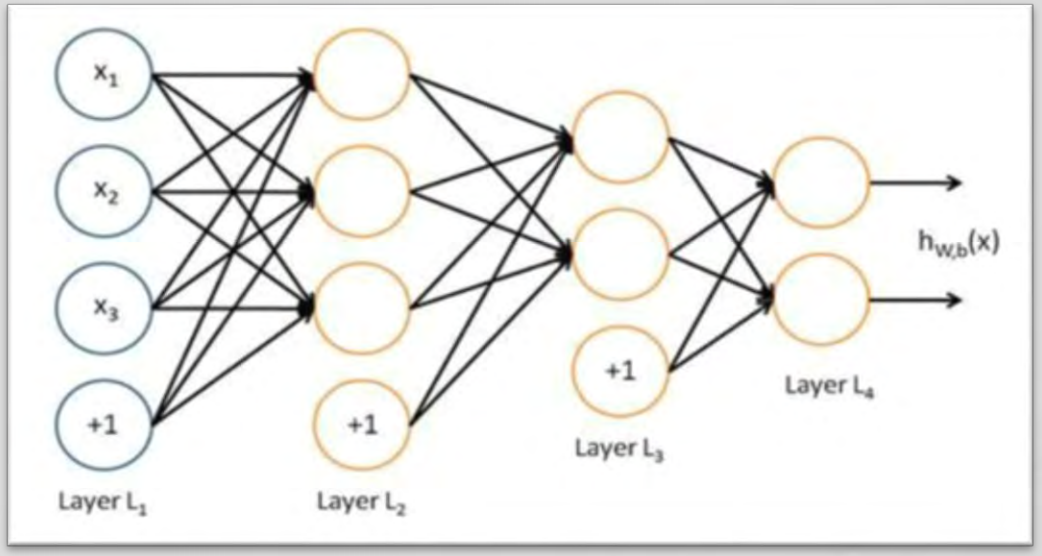
申请进件的关联特征



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	
。。。	

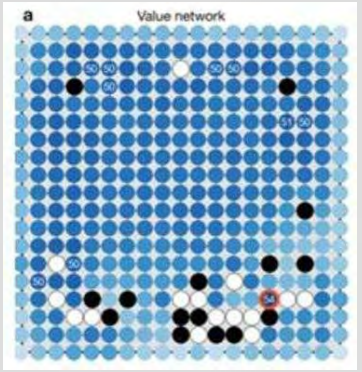
深度学习网络



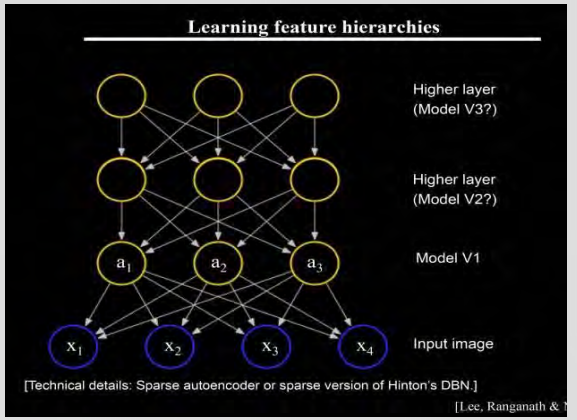
Combine	AUC	Accuracy	Precision	Recall	F1-measure	Lift
LR	0.85	0.86	0.90	0.90	0.90	1.25
DL	0.90	0.87	0.90	0.92	0.91	1.32
RF	0.93	0.88	0.93	0.91	0.92	1.35

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？



如何描述棋风，大局观？



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。

如何制订动态防范的欺诈规则？

传统行为风险评分模型

- 观察期
- 表现期

数据时间段划分

- 预测：严重性、近期性、频率性、货币价值性、组合性
- ……
- 表现：在表现期末为呆账、破产、3期以上拖欠的账户为『坏』、在表现期末无拖欠或仅为1期拖欠的账户为『好』……

预测变量/表现变量

概率 > 分值

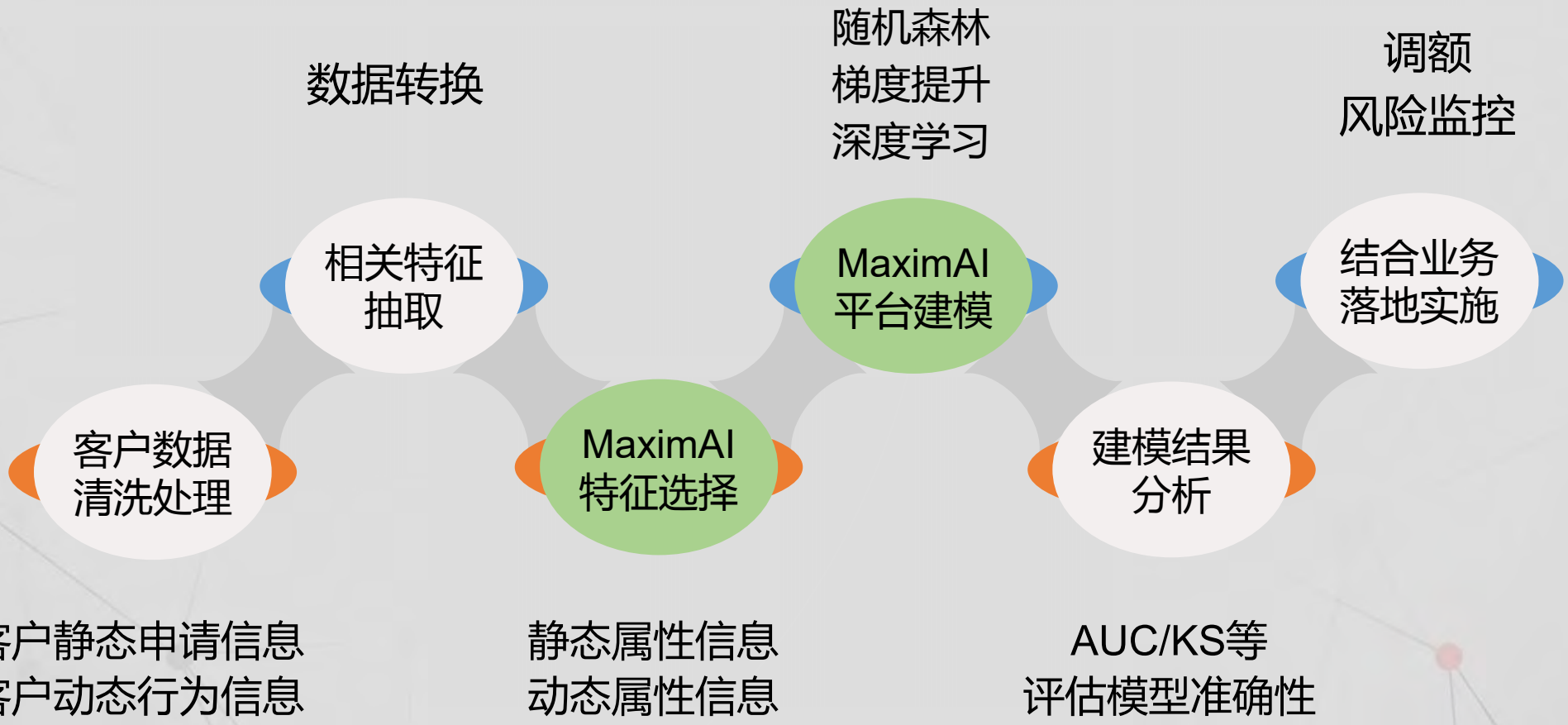
分值转换

排除

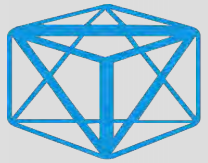
- 已经严重拖欠、呆账、破产的账户
- 开户时间少于6个月的账户
- 已经关闭的账户
- 处于欺诈、争议状态或账户持有人已故的账户
- ……

逻辑回归模型

某分期产品行为评分卡

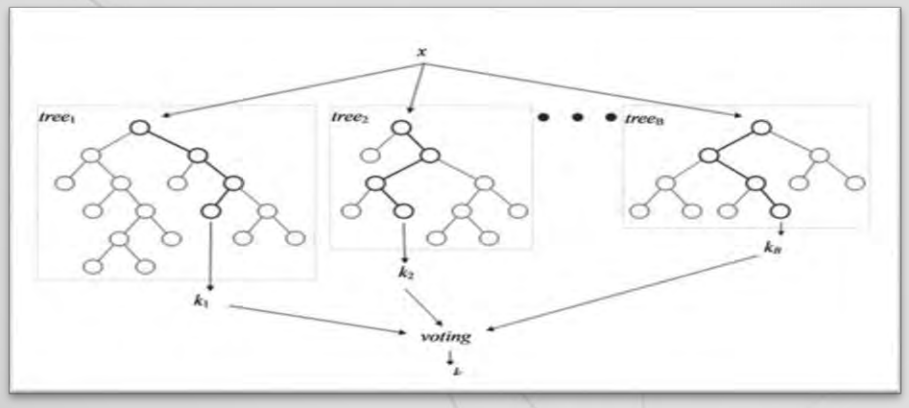

数据转换
随机森林
梯度提升
深度学习
调额
风险监控
相关特征
抽取
MaximAI
平台建模
结合业务
落地实施
客户数据
清洗处理
MaximAI
特征选择
建模结果
分析
客户静态申请信息
客户动态行为信息
静态属性信息
动态属性信息
AUC/KS等
评估模型准确性

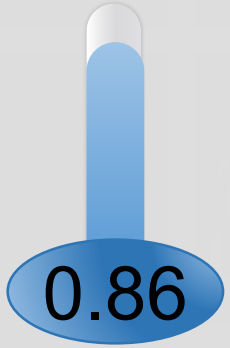
数据选择模型的规模化机器学习能力



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：
- 随机森林
- 梯度提升
- 深度学习

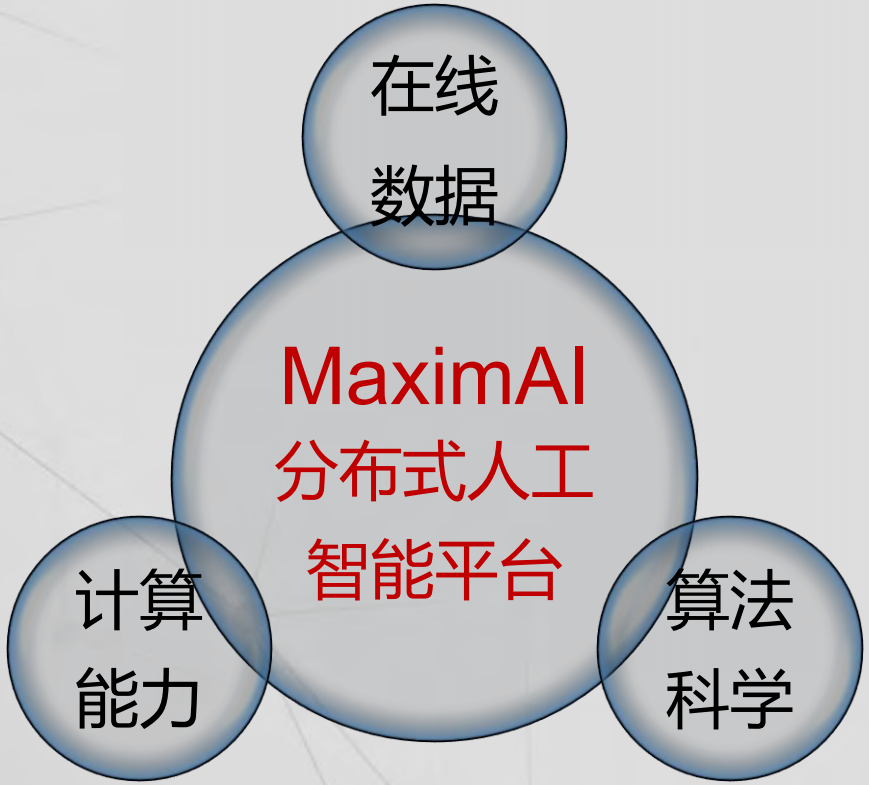




使用逻辑回归
模型的AUC值

使用梯度提升
模型的AUC值

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

向导服务

Scala
python
R

订阅服务

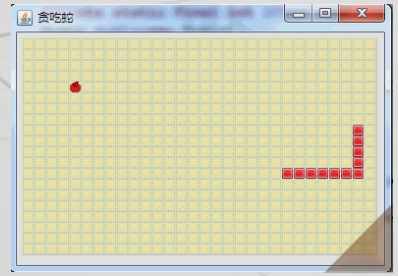
Subscription

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

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.

CHINA DATA ANALYST SUMMIT

CDA 数据分析师 www.cda.cn

THANKS

跨界互联 数聚未来

第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT