

金融大数据 反诈技术白皮书



蚂蚁集团
ANT GROUP



清华大学
Tsinghua University

2022年12月

完成单位

Company

蚂蚁科技集团股份有限公司
清华大学

总策划

Producer

赵闻飙、徐恪、李俊奎

主要完成人

Major Character

王维强、李琦、金宏、张震、任炬、
高丽、肖凯、崔世文、洪丹、张天
翼、李健雄、张哲、马俊杰、张超、
郑亮、朱丛、赵亮

其他参与人

Character

郑霖、刘腾飞、兰钧、吴星、彭凤
超、都金涛、许小龙、王宝坤、田
胜、付大鹏、刘谦、傅欣艺、应缙
哲、谭潇、苗书宇、王宁涛、李志
峰、刘京、金小蓉、薛兰青、王可、
叶帆帆、徐峰、孟昌华、傅幸、祝慧
佳、陈帅、李哲、陈倩华、陈锣斌、
王晓东、郭振宇、朱传群、盛闯、尹
攀、黄海、孙博文、王兴驰、郭真
林、禹航、梁磊、李强、鲁玮

PREFACE

序言

当前以电信网络诈骗为首的各类互联网欺诈行为呈现出多发、高发、多元化的态势，严重威胁着人们的财产安全。与传统的电信网络诈骗相比，互联网时代的电信网络诈骗活动不断利用新理念、新技术来实现新的诈骗手法，并与网络赌博、薅羊毛、互联网洗钱等欺诈行为一起，形成了一条相互关联的黑色产业链，使得互联网欺诈行为的风险识别难度不断增加，风险对抗也日趋激烈。为了遏制和惩治电信网络诈骗活动，我国于2022年12月1日正式实施了《中华人民共和国反电信网络诈骗法》。然而，如何从技术层面有效打击和治理电信网络诈骗仍然面临着极大的挑战。

近年来，随着5G、物联网、大数据、云计算等技术的快速发展与应用，人工智能技术迎来了新一轮的蓬勃发展。在反电信网络诈骗领域，如何通过先进的人工智能技术针对性地识别全链路、团伙化的电信网络诈骗行为是人工智能应用的重中之重。目前，基于人工智能算法的异常检测、多模态融合学习、图神经网络、端云协同等技术已被广泛运用在了电信网络诈骗风险识别中，在事前风险感知、事中攻击检测和事后威胁处置上都表现出明显的性能优势，能够有效提升电信欺诈的检测率及打击力度。

然而，人工智能在反诈领域的广泛应用也

同时带来了新的挑战。人工智能的技术安全程度和应用可信赖程度正逐渐成为反诈场景中的焦点问题，大大加速了可信人工智能技术的研究进度。可信人工智能技术通过将可信算法的指导性原则与人工智能技术相融合，在数据收集和处理、算法设计和实现、运维等多个环节上提升了人工智能技术的可信赖程度，解决了人工智能技术在应用过程中面临的数据隐私保护、算法可解释性、算法鲁棒性和公平性等问题，规范了人工智能技术的应用。

目前，可信人工智能技术在电信网络诈骗领域已经进行了非常多的应用尝试。例如，通过应用隐私计算机制保证算法在应用过程中的数据安全、借助算法可解释性提升模型决策的透明度和公平性、利用对抗生成技术提升模型在风险攻防中的鲁棒性等，这些技术能够显著提升人工智能技术在反诈场景中应用的可信赖程度，极大发挥人工智能技术在电信网络诈骗风险识别中的性能优势，同时有效缓解人工智能技术的“黑箱”特性，对于网络反诈有着重要意义。

本白皮书针对网络反诈的技术挑战、技术架构、关键技术、应用实践等方面展开论述，希望能为学术界和工业界开展面向反诈技术的理论、实践和应用提供参考和指导。

徐恪、李琦、金宏

STENZON CONTENT

目录

01

金融大数据 反诈系统通用框架

- 01 背景
- 02 电信网络诈骗概述
- 04 反电信网络诈骗金融治理

04

事中攻击检测 与防护

- 37 事中风险防控概述
- 38 异常检测
- 42 实时风控
- 50 风险决策
- 54 风险阻断

02

反诈基础设施

- 12 反诈基础设施概述
- 13 多维异构超大规模交互图
- 17 可信技术设施和可信基础算法

03

事前风险感知与预测

- 29 事前风险感知概述
- 30 全网威胁探测处理
- 33 全网情报采集处理

05

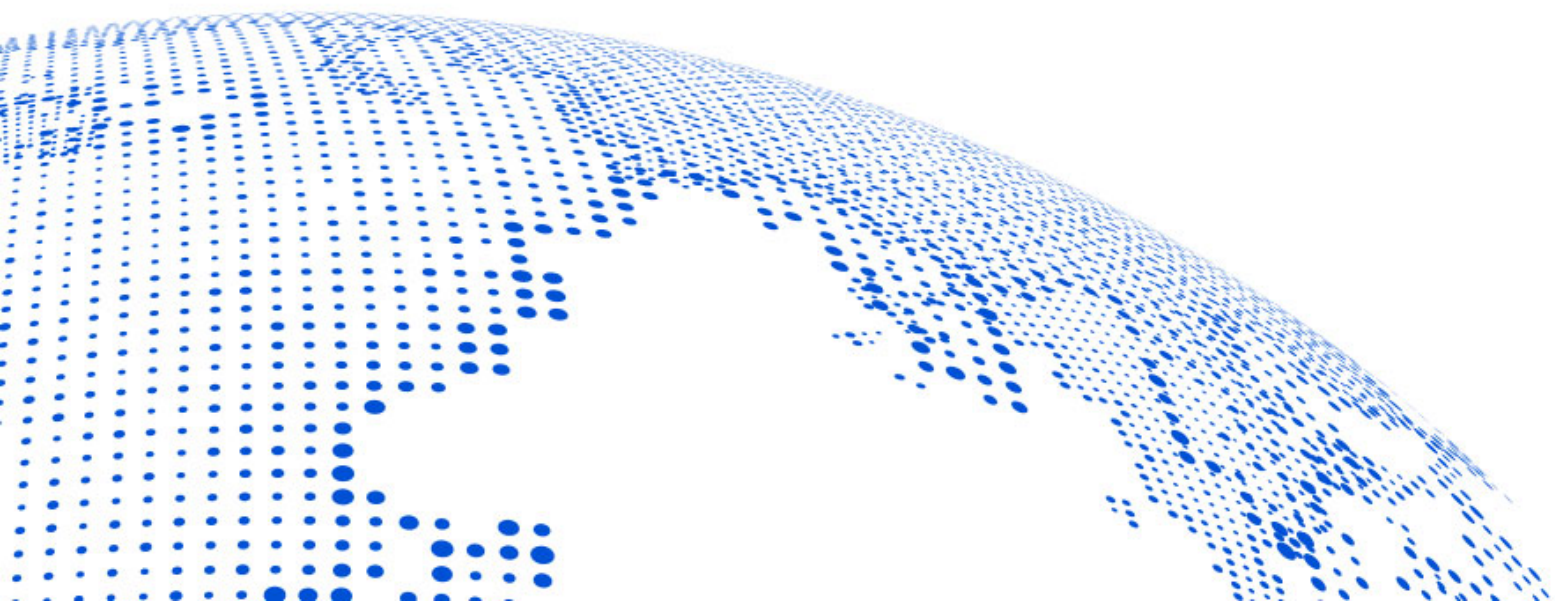
事后威胁反制与应对

- 57 事后威胁应对概述
- 58 智能审理方案
- 60 类案检索方案
- 62 威胁知识提炼

06

实践与案例分析

- 65 事前防控案例-公网反诈信息处置
- 67 事中防控案例-支付宝双十一反诈风控
- 69 事后防控案例-支付宝投诉体验提升案例





金融大数据 反诈系统通用框架



1.1 背景

当今，银行业金融机构与非银行支付机构正面临愈发严峻的电信网络诈骗风险。随着互联网技术的不断发展，恶意诈骗分子利用电信网络技术实施诈骗的手段不断升级，其基于新技术与新场景，不断寻找新方法，如软硬件攻击、社会工程学等，实施更具专业化、智能化的欺诈手段，形式上包括诱骗投资理财、杀猪盘、借贷等等。这使得需要防范的欺诈风险案例的数量快速增长，案例的复杂性也不断提升，传统的风控审查以人工的方式进行经验控制，难以应对当下反诈形势。

基于大数据风控的反欺诈技术与体系应运而生，运用大数据方法构建风控系统，对黑产风险行为进行风险预测、阻断、管理、溯源。依托于人工智能、云计算等技术，通过大量庞杂的数据建模的大数据风控能够更加科学有效的提高风控安全能力，同时降低人工审

查成本与交易处理成本，在保障安全的同时提升金融服务的效率，为金融行业保驾护航。

2022年12月《反电信网络诈骗法》正式颁布，国家明确了对电信治理、金融治理、互联网治理三个领域的治理要求。目前市场还未出现一份完整的金融领域反电信网络诈骗的框架搭建规范，本白皮书针对金融风控场景下反诈系统的基础设施建设，以及风险全生命周期反诈核心技术展开了阐述，期待为产业界和学术界开展面向金融反诈的研究、实践和应用提供有益的参考和指导。



预测



阻断



管理



溯源

1.2 电信网络诈骗概述

电信网络诈骗的概念

据《中华人民共和国电信网络诈骗法》^[1]，电信网络诈骗，是指以非法占有为目的，利用电信网络技术手段，通过远程、非接触等方式，诈骗公私财物的行为。

电信网络诈骗的常见手段

电信网络诈骗手段多样、隐蔽性高，主要包括以下几种：

兼职刷单类诈骗

骗子谎称可以提供网络兼职刷单工作，被害人误以为可以通过兼职刷单劳务获取报酬，在虚假/正规购物网站、社交软件上被骗子引导缴纳各种费用，最终既没有收到报酬，也没有返回本金，造成财产损失。

信贷类诈骗

骗子谎称可以提供贷款、套现等服务，被害人在虚假借贷网站/APP、社交软件等平台上进行相关操作时，骗子通过后台控制、骗术引导设置障碍，让被害人缴纳各种费用，造成财产损失。

仿冒类诈骗

骗子假冒客服、公检法、亲友等身份，虚构被害人涉及违法违规行为（比如校园贷、洗钱等）、中奖返利、理赔补偿（如已购产品不合格等）等事实，诱导被害人缴纳各种费用或转账，最终造成财产损失。

色情类诈骗

骗子谎称可以提供色情上门服务，诱导被害人缴纳定金、交通费、安全保证金等各种费用，最终对方失联或继续要求转钱，被害人意识到被骗，造成财产损失。

杀猪盘类诈骗

骗子以各种交友套路与被害人建立网络恋爱关系，取得被害人信任后，虚构在某平台赌博或投资可以赚大钱的事实，诱导被害人投入资金，最终在发现对方失联、资金无法提现、平台关停等情况时意识到被骗，造成财产损失。

游戏相关商品交易类诈骗

骗子虚构出售游戏账号、游戏装备、代练/陪玩服务等事实，被害人误以为付款后对方会发货或提供服务，付款后对方失联或不发货，造成财产损失。

网络交友类诈骗

骗子虚构美国大兵、卖茶女等身份，假装在网络上与被害人进行恋爱交友，以给被害人寄送高价值财物（美金、贵金属等），在海关被扣押需要缴纳关税等名义诱导被害人转钱，或谎称家族产业（茶叶、白酒等）被亲戚侵占，博取被害人同情，让被害人高价购买商品，最终不发货、货不对板、失联等，造成财产损失。

虚假购物消费类诈骗

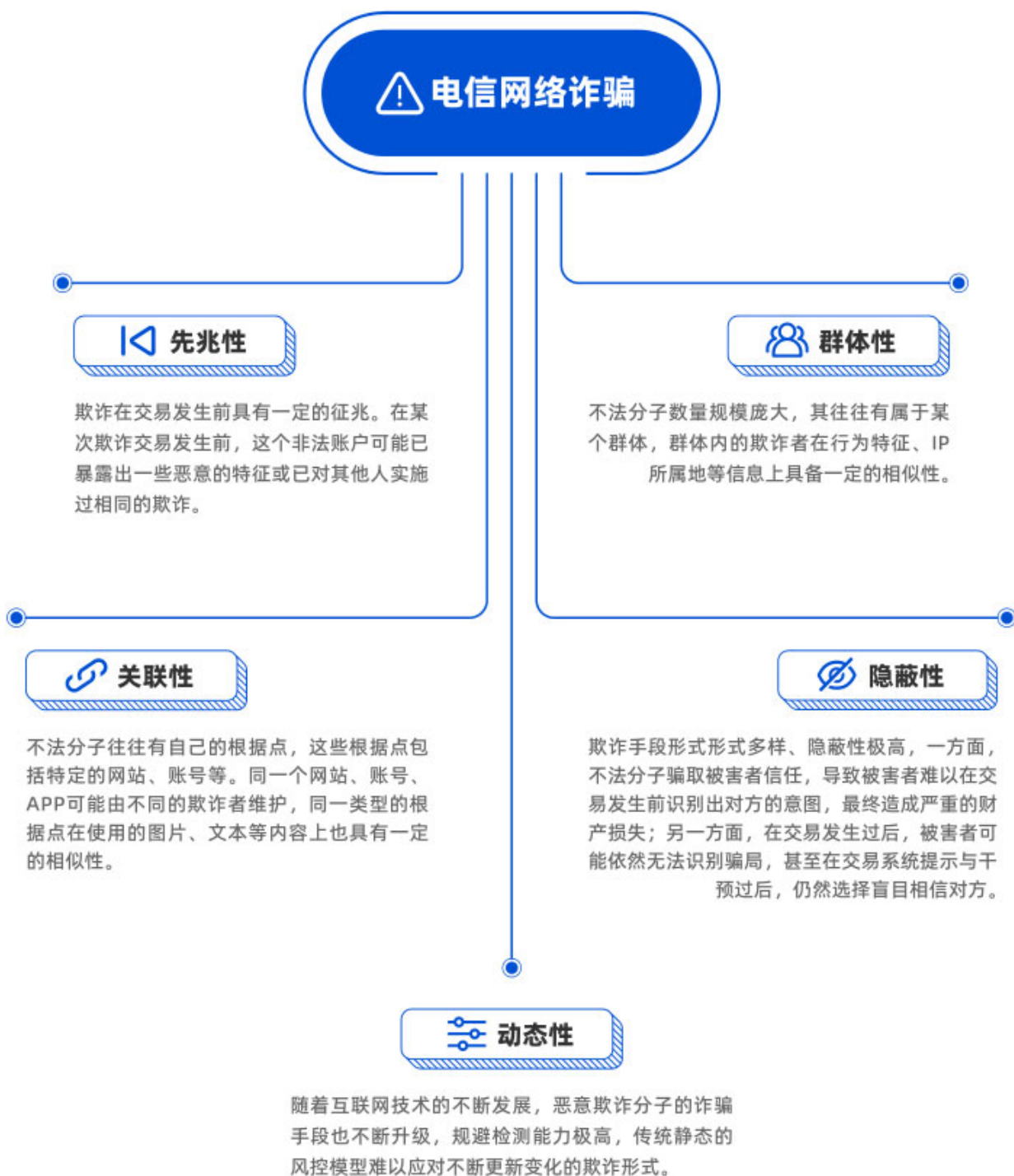
骗子虚构出售游戏相关以外的商品或服务事实，被害人误以为付款后对方会发货或提供服务，付款后对方失联或不发货，造成财产损失。

中奖返利类诈骗

骗子谎称被害人中奖，诱导被害人缴纳定金、税费、保证金等各种费用，最终失联活继续要求转钱，导致被害人财产损失。

电信网络诈骗的特点

电信网络诈骗行为具有先兆性、群体性、关联性、隐蔽性、动态性等特点。具体如下：



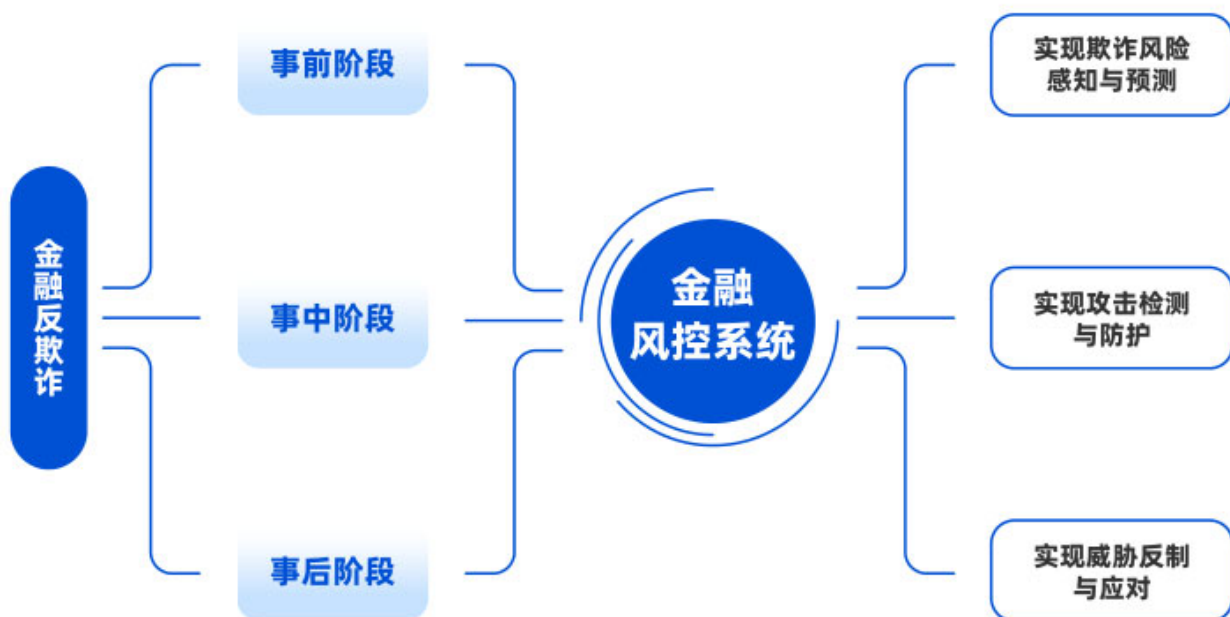
1.3 反电信网络诈骗金融治理

概述

金融机构通过搭建金融风控系统达到反欺诈的目的，保障用户的切身利益。金融反诈需要从问题出发，根据欺诈事件、欺诈分子的特点找寻可靠的反制应对方案。

● 一方面，根据欺诈的先兆性特点，可知欺诈发生前许多信息（例如舆情信息、站点风评等）对欺诈检测有辅助补充作用。而传统风控方案仅在**事中**（即交易发生中）进行检测与防护，往往会遗漏相当一部分事前信息，导致判断结果不准确；此外，由于金融领域交易数据量庞大，对实时防控需求极高，因此，需要风控系统具备一定的先验知识，或在交易前对交易双方已有基本的判断。为此，**事前**（即交易前）欺诈风险的感知与预测十分必要。

● 另一方面，欺诈的动态性与隐蔽性特征表明，尽管事前事中的检测部署层层严密，仍然存在一些欺诈事件被“漏检”和“逃逸”的现象。为此，在**事后**（即交易后）的威胁反制与应对急需重视。



分阶段检测是金融风控需要遵循的上层运作逻辑，而搭建一套完善的金融风控系统还需要解决数据处理与算法设计的问题。反欺诈的三个阶段对数据与算法的需求具有相似性，因此，从数据与算法层面抽象出一套通用的**底层基础设施**，是整套框架的基石。

不同阶段的需求与挑战

金融行业的业务逻辑庞杂错综，交易发生的各个阶段都存在诈骗风险隐患，因此，需要对基础设施与各个阶段可能存在的问题与挑战进行分析，总结归纳出每个阶段需要的风控业务和技术。

1. 数据层面

金融支付环节产生的数据形式多样关联性强且涉及用户隐私，如何对其进行快速存取和有效表征，需要考虑以下几点：

01

数据规模庞大

金融支付交易的日吞吐量巨大，庞大的数据量与用户对交易效率的要求成为了金融数据处理的一大问题。

02

多模态异质数据

金融风控安全场景的主体对象多种多样，包括有支付过程、用户、商户、站点、程序等等，不同对象的数据存在多种差异，例如，一个支付过程包括有聊天记录、收款时间、收款数额、参与用户等信息；构成一个站点、程序的画像信息包括访问历史、访问用户、关联IP、首页截图等等；一个用户对象包含了访问历史、聊天记录、访问内容、行为序列；一个商户对象有商户照片、营业执照图片、收款信息、地理位置等内容。这些数据形式多样，包括了图片、文字、数字、语音、序列等等，如何处理这些多模态异质数据是金融风控安全面临的一大难题。

03

数据强耦合、关联性大

不同主体之间的信息存在强关联、强耦合的特征，例如，一个用户访问某个站点，其访问记录既是用户画像的构成部分，也是评判站点是否可疑的主要特征。因此，如何将不同主体的不同类型的关联信息以最佳的形式保存、建模，从而保留主体之间的关系特征、并实现跨主体信息补充，也是金融风控需要解决的一大问题。

04

用户的敏感隐私问题

用户在使用互联网金融服务时，会产生大量隐私数据，如果这些数据被上传至云端，会造成不可估计的后果。因此，出于对用户隐私安全的保护，在搭建框架时，必须考虑敏感数据的存储与处理问题。

05

数据实时调用的需求

用户对交易效率的要求越来越高，因此数据的存取与调用必须快速。

2. 算法层面

人工智能模型是算法的核心，AI（Artificial Intelligence，人工智能）模型在风控场景落地应用时，涉及到AI安全可信的问题，算法层面包括鲁棒性、可解释性等，工程层面包括全链路部署的问题。



安全风险场景的业务处理链路复杂，需要从模型全生命周期安全出发，考虑可信技术设施问题。



金融安全风控是一个双向对抗的过程，黑产团伙、欺诈方会通过各种手段逃避检测，例如上传带有PS痕迹的商户图片、对宣传文字进行变种操作等等。因此，无论是模型学习过程，还是部署阶段，都存在逃逸检测的风险，因此，需要增加模型的**鲁棒性**；对支付的风险监测需要保证公平性、避免黑盒问题，为此，需要人能够“理解”算法；同时，为了让AI能够适应新的知识，又需要算法能够“理解”人，这就是**AI可解释技术**。

3. 事前阶段

事前阶段，需要对互联网站点提前预测风险、并收集外部情报以备事中检测。

● 一方面，恶意分子往往驻根于现有的互联网站点，包括APP、URL网站、小程序等实体，大部分欺诈案件的发生与恶意风险站点密不可分。金融反诈系统需要对现有互联网站点进行主动扫描与风险预判，以期防患于未然，这部分称作**全网威胁主动探测**。

● 另一方面，事中检测对实时性要求高，预先收集外部情报数据有助于提前阻断风险，并为事中检测欺诈补充有力依据，这部分称作**多源情报智能采集**。



4. 事中阶段

事中阶段，需要对可疑的风险交易进行识别、决策与阻断。

01

由于金融业务的特殊性，部分业务场景缺乏有效的风险标签，即冷启动问题，而**异常检测**可以通过无标签数据解决该场景下的攻击检测。

02

针对有风险标签样本的具象风险防控，则通过**实时风控**技术，综合端、云信息，及时识别可疑风险，解决数据传输负载高与用户隐私安全问题。

03

当感知到交易风险后，需要根据风险识别结果，结合用户的画像信息，评估综合风险水平，并给出合适的管控决策方案，即**风险决策**。

04

考虑到电信网络诈骗存在“隐蔽性”与“动态性”的特点，即部分欺诈交易过程中，欺诈分子对用户的欺骗极深，风控系统应当主动发起与用户的交互，包括了解更多的用户信息和情绪、提醒用户交易风险，从而做出更合理的判断，更有针对性的进行风险劝阻。这部分内容即**风险阻断**。



5. 事后阶段

在事后阶段，需要对用户的投诉进行审理、同时回看发掘潜在风险事件，并对事中模型进行更新与升级。

01

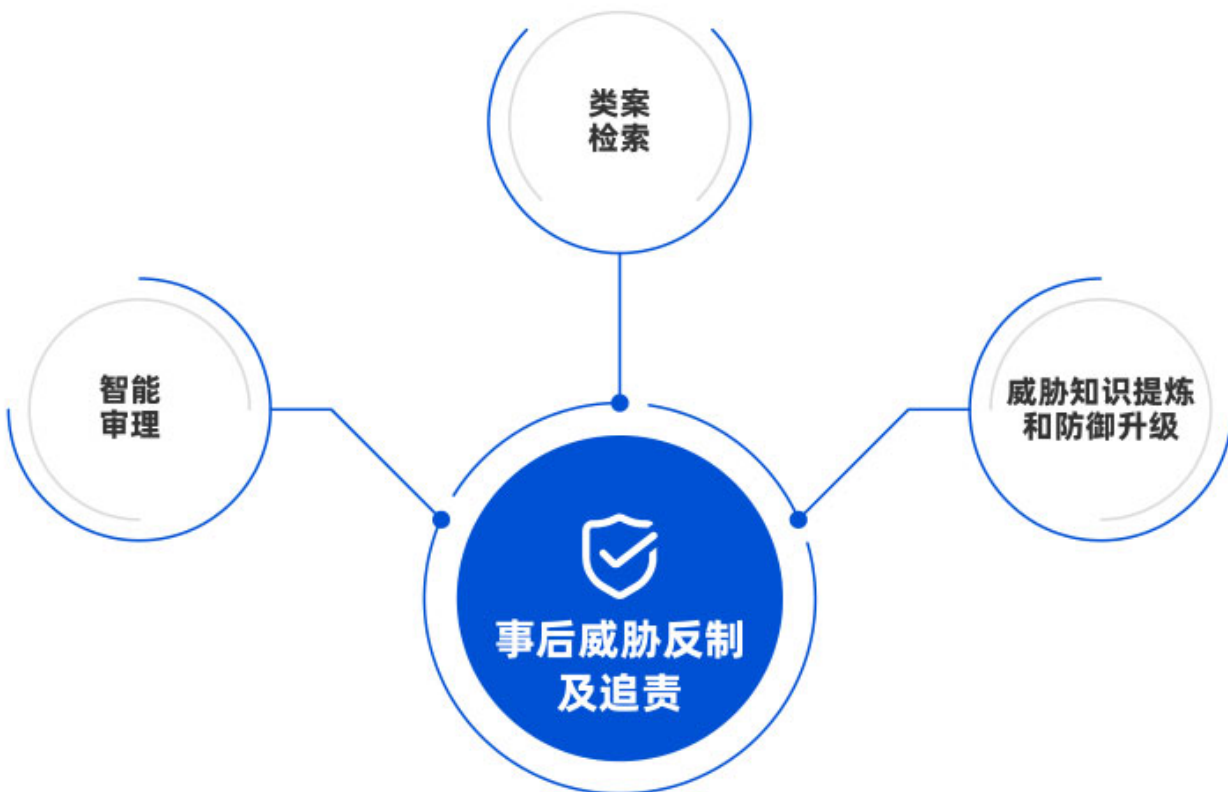
一些风险事件能够逃避事前风险感知与事中攻击检测（隐蔽性与动态性），而用户对这类风险事件的投诉信息是一份有力的指认证据，如何客观的利用这份信息进行判断，是**智能审理**需要解决的问题。

02

必须认识到，仍然存在许多隐蔽性极高的事件可以躲过用户的发现和检举，如何通过大数据的方式找寻这类未被投诉的潜在风险事件是事后阶段的一大重点，这部分称作**类案检索**。

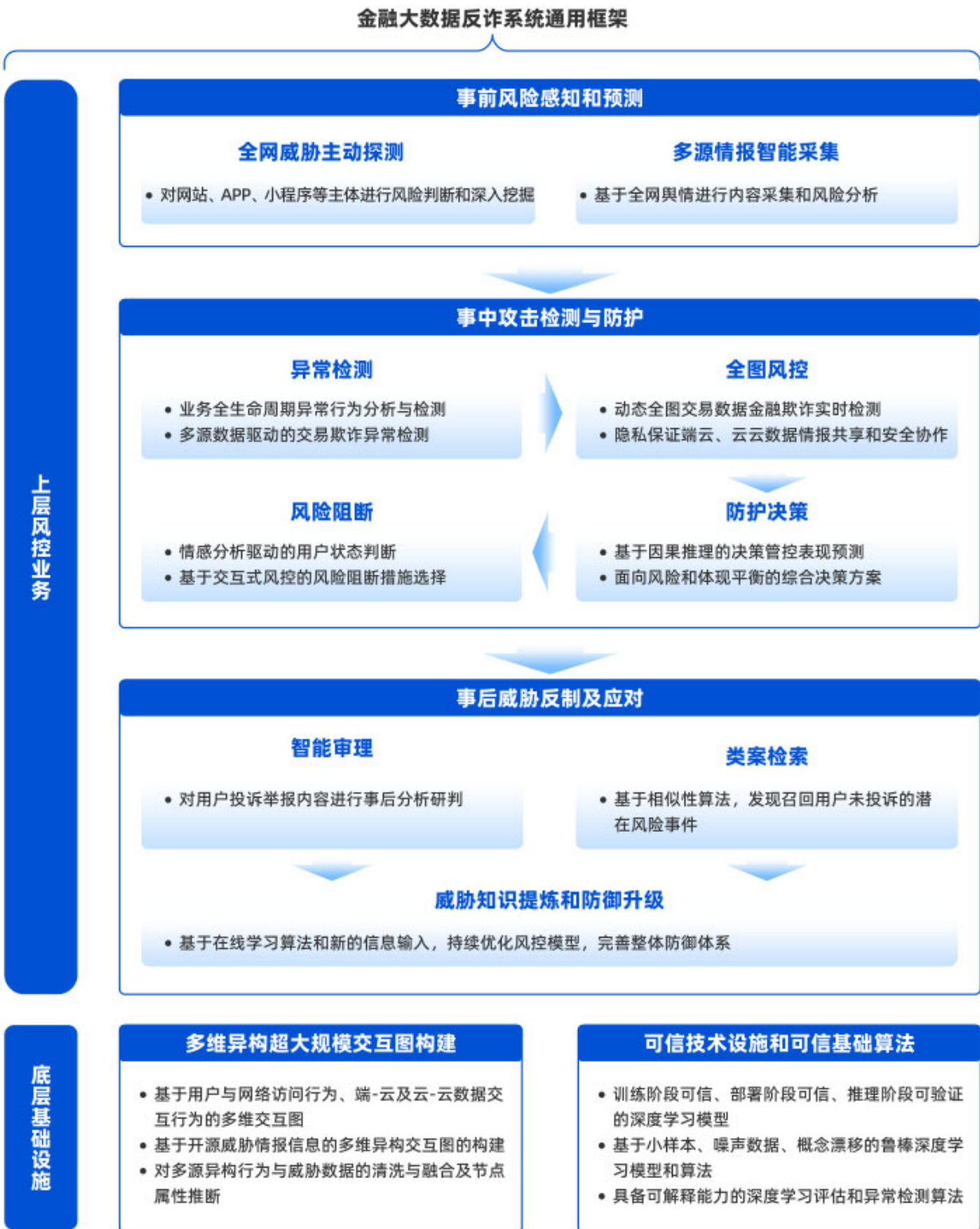
03

欺诈的动态性特征表明黑产团伙、恶意分子的作案手段日益更新，部署的模型需要动态更新以适应动态对抗场景，同时，由于交易风控系统要求模型具备稳定性，因此模型的动态更新与稳定性成为了一个两难问题，这部分称作**威胁知识提炼与防御升级**。



金融反诈系统通用框架

本书旨在面向银行业金融机构、非银行支付机构等承担支付风险防控责任的相关部门，提供一份完整的金融反诈系统通用框架的构建方案，以期帮助相关部门建立反电信网络诈骗的内部控制机制，防范业务中存在的涉诈风险。具体内容与各模块作用如下图所示：



1. 数据与算法基础设施



数据方面

构建多维异构超大规模交互图，实现大规模、强关联、异质、隐私数据的储存、调用与表征。



算法方面

在全链路，即模型训练、模型部署、模型运营三个阶段部署可信AI，以解决用户数据的隐私安全问题，提高模型的鲁棒性和可解释性，保障信息的真实性、决策的透明性和结果的公平性。

2. 事前阶段

全网威胁主动探测技术对所有可得的互联网站点进行风险判断与深入挖掘，从而将被动风控变为主动风控；智能情报引擎技术基于全网舆情，实现内容采集与风险分析，为事中攻击检测提供有力的数据支撑。

* 这部分内容将在第三章中详细介绍

3. 事中阶段

首先，无监督异常检测技术能够有效解决冷启动问题，实现业务全生命周期的异常行为分析与检测，其分别应用于结构化数据与图数据，实现多源数据驱动的交易欺诈检测；

针对有标签数据则采用有监督方法，利用端、云结合的实时风控技术保障用户隐私、共享情报与安全协作、解决实时的需求；

之后，基于因果推理的风险决策技术，预测决策管控表现，以平衡用户体验与管控强度；

最后，基于交互式风控的风险阻断技术，通过分析用户情感、合理推测用户状态，主动阻断可疑交易。

* 这部分内容将在第四章中详细介绍

4. 事后阶段

智能审理技术对用户投诉举报内容进行事后分析研判；

同时，利用类案检索技术，基于相似性算法，发现高隐蔽、未被投诉的潜在风险事件；

最后，搭建威胁知识提炼框架，基于在线学习算法与新信息输入，持续优化风控模型，完善整体防御体系。

* 这部分内容将在第五章中详细介绍

2.1 反诈基础设施概述

反诈基础设施是支撑整个反诈系统的重要基石。从要解决的问题来看又分为两类问题：数据和算法。



基于此类问题，传统的依靠堆表链接的方式并不能很好解决此类问题；“多维异构超大规模交互图”是行业有效可行的技术解决方案，下文将展开介绍如何用该方案解决数据问题。



2.2 多维异构超大规模交互图

■ 概述

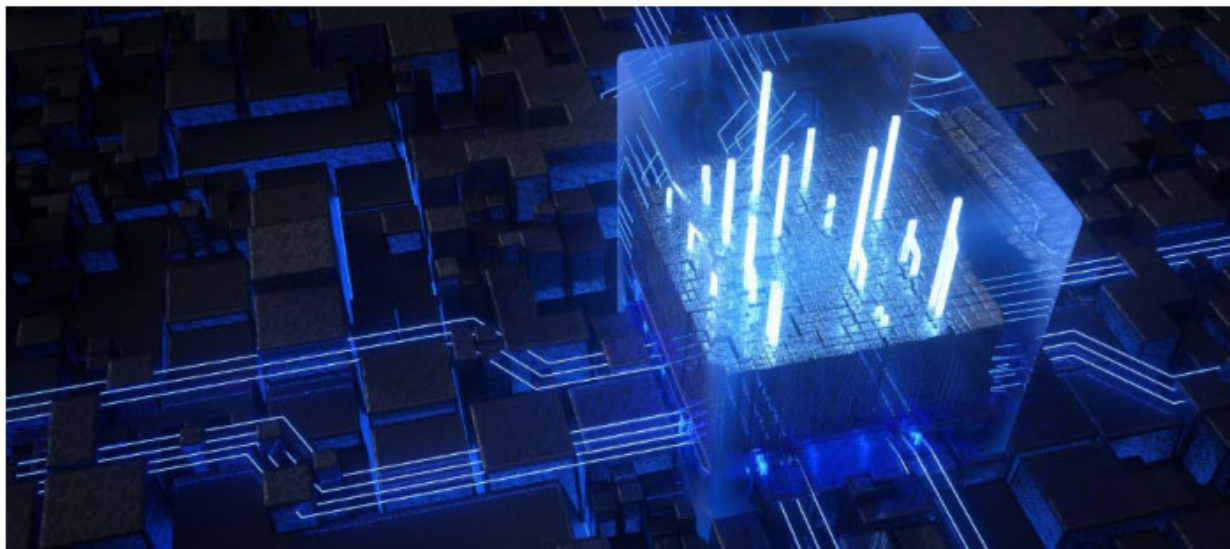
多维异构超大规模交互图要解决的主要问题是什么？金融反诈场景中存在大量的异质数据，如：图片、文字、行为序列等。这些异质数据以用户为中心，包含了访问历史、聊天记录、访问内容。如果用户为商户，信息则更加丰富，也会包含商户照片，营业执照照片，收款信息，地理位置等。要想高效合理的使用这个数据挑战大，多维异构超大规模交互图技术可以有效的将这些信息整合起来，同时提供高效强大的多模态数据分析能力。不仅可以解决多模态多源数据信息可视化能力，同时也可以将这些数据进行数据标准化和强大的建模的能力。传统同质图或者异质图，往往为静态图，同时节点和边关系也限于结构化数据。与传统同质图或者异质图不同的是，多维异构超大规模交互图是包含事件、行为序列、图片、文字的多模态异质动态事件图结构。这样就可以最大程度的保留原始丰富信息，保证下游任务可以更加全面地进行信息整合。

名词解释

名词	定义
结构化数据	以标准csv或excel文件存储的数据，数据内容为字符串、浮点数、整数等，存在可以直接用来做数据分析的特征
非结构化数据	常常以字符串形式为主，需要处理后可以入模进行训练。如自然语言中的段落文字等。
行为数据	带时间戳的数据，在金融风控场景中常按照用户操作顺序作为行为序列数据，行为序列的属性可以为结构化或非结构化
实体	指的是具有可区别性且独立存在的某种事物。如用户、银行卡、地理位置、设备等等，一个实体存储在图谱上就是图的一个节点。
概念	具有同种特性的实体构成的集合
属性	用于区分概念的特征，不同概念具有不同的属性。不同的属性值类型对应于不同类型属性的边。如果属性值对应的是概念或实体，则属性描述两个实体之间的关系，称为对象属性；如果属性值是具体的数值，则称为数据属性

■ 数据结构

整个技术方案将数据格式分成了三类：结构化数据、非结构化数据，行为数据。针对结构化数据，在导入过程中，相对比较规范，直接按照属性值进行导入即可。针对非结构化数据，可以用string格式、JSON格式进行储存。行为数据常常以[[time, action, propoties], ...]的格式存储。在所有数据进入图谱时，需要对数据中是否已存在相关实体及边关系进行判断。



■ 知识获取

知识获取的目的是根据非结构化文本构建知识图谱，补全已有的知识图谱以及发现和识别实体和关系。知识获取的任务主要包括实体识别、实体关系学习以及联合抽取等。

1. 实体识别

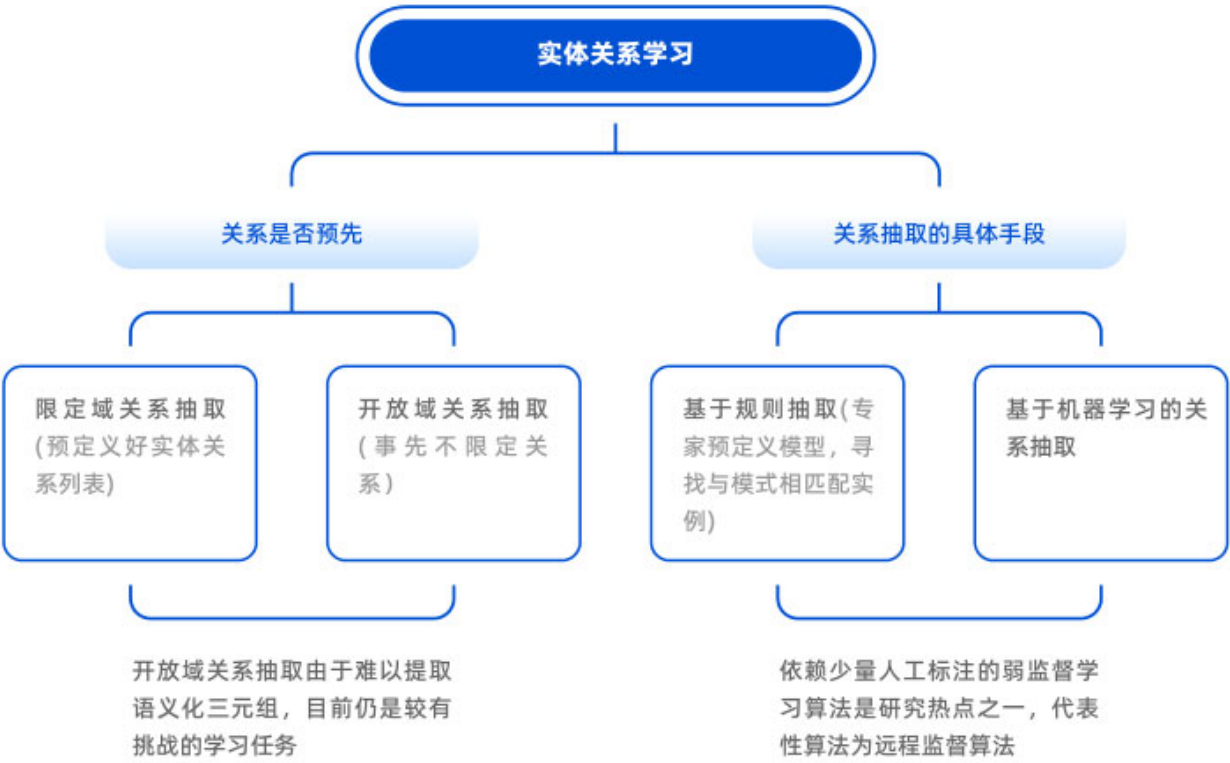
实体识别，即从待识别文本中指定如姓名、地理位置、商户名等类别实体的过程。

常见算法如：

传统统计模型如最大熵模型、SVM（Support Vector Machine，支持向量机）、HMM（Hidden Markov Model，隐马尔可夫模型）、CRF（conditional random field，条件随机场）等；深度学习类模型常见的如[NN]-CRF结构【其中[NN]部分可以为：BERT（Bidirectional Encoder Representation from Transformers，谷歌文本预训练模型）/CNN（Convolution Neural Network，卷积神经网络）/LSTM（long short-term memory，长短期记忆网络等】。

2. 实体关系学习

实体关系学习(又称关系抽取), 指自动从文本中检测和识别实体之间的某种语义关系, 其往往建立在已完成实体识别的基础上。



3. 联合抽取

不同于传统串行抽取(先做实体识别、再做关系抽取)步骤, 联合抽取算法将实体识别与关系抽取融合在端对端框架内统一实现。常常需要上述多个方案进行组合使用。

知识融合

知识融合的目的是为了能够将不同的知识图谱融合为一个统一、一致、简洁的形式, 为使用不同知识图谱的应用程序间的交互建立操作性。

数据校正

无论针对结构化数据与非结构化数据，均存在用户信息缺失或数据不准确的情况。针对数据缺失，采用知识图谱的数据补全方法进行补全。信息校正按照节点属性及节点关系进行分类。如果有场景业务规则即可完成信息校正，则可以纠正高质量具有可解释性的边关系及节点属性。如果规则无法进行校正，则需要借助算法能力进行。针对节点属性，常用图谱点分类算法进行校正。针对节点关系常常使用关系预测进行校正。

01 点属性校正

金融场景中，在数据导入之后，很多点属性并不是完全准确的，为了保证下游任务的稳定性及有效性。修正有误的点属性尤为重要。当前行业在点属性修正过程中，通常采用的方案为：基于规则和算法。基于规则在不同场景中，不同的业务差异较大，这些修正具有比较好的解释性并强依赖业务。基于算法的，往往使用深度学习算法，与图分类算法、图回归算法相似。

02 边关系校正

在金融场景中，因为用户隐私及数据源获取等原因，会导致信息不全的情况，如夫妻、同事、同学等关系。补全这些信息可以提升后续推理的性能，同时也便于进行金融知识管理、风险提前发现、提升用户体验等。行业内补全主要分为两种：基于专家规则、基于算法。基于专家规则的常常根据专家经验，利用节点属性及其边关系的阈值进行判断，从而实现自动推理关联等。基于算法的常常使用MLN，Trans系列，pLogicNet等算法。基于不同数据源获取知识图谱数据时，不仅会存在数据缺失的情况，也会存在节点关系误连的情况。针对这种情况，与关系补全类似。

知识推理

知识推理是知识图谱应用中的一块重要内容，其目标是通过图谱中已知的实体/关系/属性/模式等，推测出隐藏在图谱多度关系之后的隐含知识。

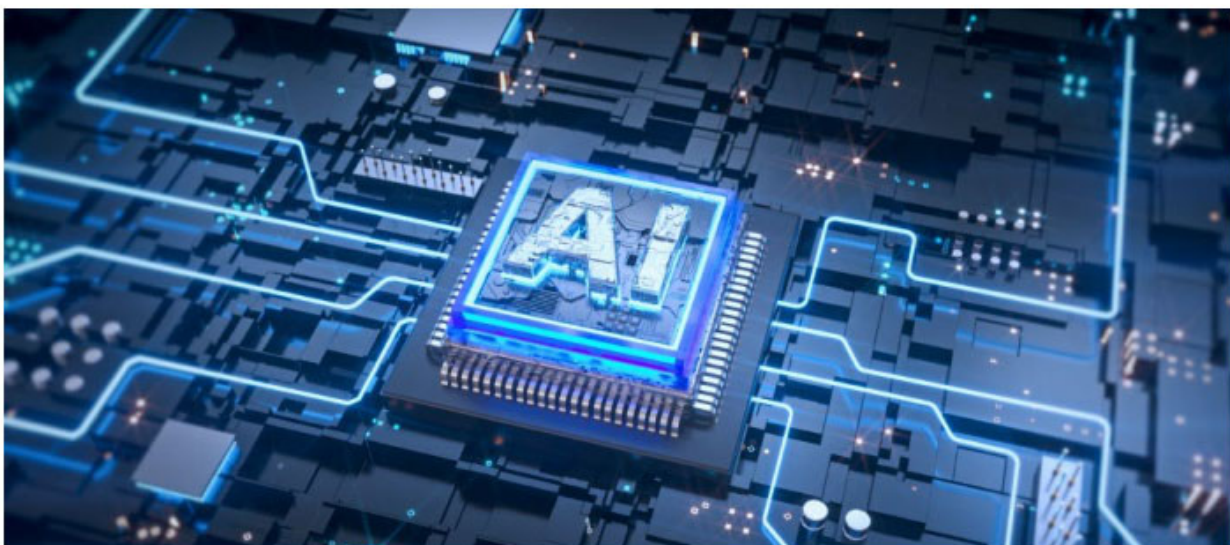
知识推理算法门类繁多，比较有代表性的如FOIL/PRA(基于逻辑规则)、TransX系列(基于距离表示)、DeepWalk/HeGNN(基于图神经网络)。由于知识图谱表征提取后，叠加一层或数层神经网络即可用于下游推理任务，故有时知识推理算法与知识表征算法并无明确界限。针对金融知识图谱场景，尤其大规模情况下，基于距离表示和逻辑规则的方式，常常会出现适应性差，针对新节点需要重新训练等不同的问题。行业内通常使用基于图神经网络的方式进行推理，往往可以达到高效准确的目标。不同的图神经网络模型也可以用于不同的下游任务，如：关系判断、节点分类、节点回归任务等。

2.3 可信技术设施和可信基础算法

如2.1节所述，反诈技术设施和基础算法在实际业务中会遇到算法和工程两方面问题：算法是人工智能模型的核心，所有算法应用都需要具备鲁棒性、可解释性等基本可信AI方面的能力；工程方面，AI模型在风控场景落地应用时，需要解决全链路技术设施可信的问题。以下从工程上的可信技术设施到可信算法展开介绍总体方案。

可信技术设施

随着这几十年AI的快速发展，人工智能模型也在安全风控中起到越来越重要的作用。当AI在安全风控场景应用时，面临AI的安全可信问题，如信息的真实性、决策的透明性、结果的公平性、隐私保护、鲁棒性、可解释等问题。这一些系列的问题不是单靠一个算法技术就能解决的，许多问题前后关联相互耦合，需要系统性的去考虑设计。



当前在提到可信AI时，更多的是从模型本身出发，讨论鲁棒性、可解释、隐私保护和公平性。但是真正要在工业界落地时，需要从模型全生命周期来考虑可信AI问题。并且可信AI的任何一个属性都需要在人工智能平台全链路的各个阶段进行努力。全部阶段整体上可以分为：模型训练，模型部署，模型运营三个阶段。

训练阶段

01

在训练阶段，需要保障训练全过程可靠和模型可靠，例如数据处理、稳健算法设计、对抗训练、中毒防护等等。在数据处理时，需要对敏感数据进行脱敏或加密存储，通过防篡改等技术防止数据被恶意修改，建立授权访问机制对数据进行分级授权访问，通过数据血缘追溯数据引用关系。对抗训练，通过在训练期间同时输入原始数据和对抗样本来防御相应的攻击方法，对抗训练不仅可以提升模型对对抗样本的防御能力，还能提升对原始样本的泛化能力。中毒防护，典型的中毒或后门攻击会污染训练数据，从而误导模型行为。

02

在模型上线前，会在测评中心进行模型质量的管控，对模型质量进行测评，达到上线标准才可以上线发布应用。对模型质量的各个维度进行评估，如模型的性能、模型是否安全可靠、模型的稳定性、线上推理时的RT（Response Time，响应时间）和QPS（Query Per Second，每秒查询率）等等。

03

在模型文件安全上，需要保障模型文件的完整性，对文件加密防止非法访问与泄漏，通过模型文件的水印技术，保护版本，通过模型混淆，使之不可反编译等等。

部署阶段

01

在部署发布阶段，需要确保部署的模型文件版本是否正确，验证部署链路是否可以正确调用，部署的服务测试的效果要和线下测试的一致，部署过程要可灰度，可监控，可应急来保障发布过程的安全可信，并能通过AB测试等方法验证模型效果后逐渐推全应用。

02

在推理阶段，系统应能对相关攻击进行防御，如对抗攻击、成员推理攻击、模型逆向攻击、模型后门攻击、对抗资源消耗攻击等等。另外，在推理时，也应能对推理结果给出相应的解释。

03

在平台安全上，需对推理链路中数据、中间结果、推理结果、模型进行保护，并保障AI系统完整性、机密性、可用性、可靠性。

运营阶段

01

在监控环节，需要对业务上的性能、模型性能、模型与特征的稳定性、模型调用量、推理耗时、推理异常等进行监控，并提供应急处置和持续更新手段。

02

在透明可审计可复现方面，需对全链路建模各环节过程和资产进行沉淀（建模pipeline、建模元数据等），方便审计和复现，并依据相关政策法规要求对模型的内容对公众和政府监控部门进行披露。

03

在运营安全方面，主要关注应急响应能力建设，提升平台的防危性。

04

建设可信AI可信度量化的能力。融合在可信AI各维度的能力，给出一个量化指标，帮助模型的开发者和使用者了解人工智能的可信程度，能对人工智能模型的可信程度有一个可量化可度量的认知。

可信基础算法-鲁棒性算法

1.1 对抗防御和攻击

(1) 概述

“对抗”是安全风险永恒的主题。黑产团伙在利益的驱使下，会不断尝试攻破平台的风控防御体系，例如：黑产团伙在应对内容安全环节时使用变种文字、图片逃避检测；欺诈交易方面会尝试变换交易金额账户等交易属性。AI模型的复杂性带来极高性能提升的同时，其本身也存在一些鲁棒性问题。这些鲁棒性问题在黑灰产的攻击下会更加明显。那么如何解决反欺诈风险防御中的动态对抗的问题是行业中一个重要的研究方向。

“对抗智能”融入了经济学“博弈理论”，能够借助“双手互搏”的演习机制，让AI模拟黑产发起对自身的攻击演练，提前发现算法存在漏洞，从而帮助AI在风险识别能力、抗打击能力上持续、自动进化升级。其技术本身由“攻击”和“防御”两个方向组成，“攻击”可以帮助我们提前发现模型存在的问题和漏洞，“防御”可以帮助我们提升模型的防御能力和鲁棒性。

(2) 基于对抗防御的模型训练

在面对分布漂移和黑产攻击时，业务中常用防御算法分为以下几种类型：

01 基于样本对抗 的防御算法

这一种最直观防御方法，即直接将生成的攻击加入到模型训练当中，例如：FGSM（Fast Gradient Sign Method）对抗训练和PGD（Projected Gradient Descent）对抗训练，都是将使用FGSM和PGD攻击算法产生的攻击样本加入到模型训练中。

此方法优点是使用简单且当测试集攻击样本和训练集攻击样本手法一致时效果明显，缺点是对非对抗测试集通常会带来一定性能损失。

02 基于参数对抗 的防御算法

和基于样本对抗的防御算法类似，一个是“参数”一个是“样本”，此方法是在深度学习训练过程中通过对模型权重进行梯度上升来获得扰动之后的攻击参数，在此基础上再进行梯度下降，从而获得更加鲁棒的模型，常见的基于参数对抗的防御方法有FGM（Fast Gradient Method）和AWP（Adversarial Weight Perturbation）。

03 基于损失函数 的防御算法

算法通过设置模型损失函数增加模型的鲁棒性，例如：Trades算法，模型训练时同时使用原始样本和对抗样本，并添加两个样本之间的交叉熵。

04 基于检测 的防御算法

这是一种相对简单有效的方法，它是指在样本进入模型之前再添加一道检测变种攻击样本的环节，来判断样本是否属于变种攻击样本。例如人脸识别模型中的活体检测，内容安全文本检测中的对抗文本检测等。

05 基于模型压缩 蒸馏的防御算法

实际业务场景中由于硬件设备、网络通信等原因需要对模型进行精简压缩，我们发现基于压缩蒸馏的蒸馏的模型在面对数据漂移和样本攻击时，往往会有更好的表现。

06 基于特征预处理 的防御算法

特征在进入模型之前经过一次降噪预处理，再进行模型预测，常用的预处理方法有：PCA（Principal Component Analysis）特征压缩、图片压缩、特征降噪等。特征预处理在一些较弱的攻击时会有一个好的表现。

(3) 基于对抗攻击的AI安全检测

AI对抗检测能力通过多种对抗算法生成对抗样本，用对抗样本对模型进行测评，从而生成一份“安全检测报告”，告知当前模型所存在的风险，给出合理的建议，从而有助于AI模型风险的提前发现和防范。整体评估思路和方案分为：对抗攻击的数据、对抗攻击的模型、对抗攻击的攻击算法、对抗攻击的测评任务、对抗攻击的评估报告。

对抗攻击的数据

不同的数据任务类型，使用的攻击算法测评指标，评估指标会有非常明显的差异，常见的数据类型如下：



表格

样本特征为无序固定长度的数字组合，常见场景有：账户审理，欺诈预测等；



图像

样本特征为图片格式，常见场景有：证件审核，ocr任务、人脸识别等；



文本

样本特征为文本，常见场景有：投诉文本审核，内容文本分类、文本匹配等；



音频

样本特征为音频数据，常见场景有：智能外呼，身份认证等；



序列

样本特征为序列，常见场景有：事件序列预测，行为操作序列预测等；



视频

样本特征为视频，是多张图片的有序组合，常见场景有：内容安全审核，身份认证等；



多模态

多种数据类型的组合，是多张图片的有序组合，常见于多种数据源，且信息有互补性的场景；

对抗攻击的模型

业务中被检测模型形态有两种（模型文件和API接口），第一种模型文件，包括树模型、神经网络、逻辑回归等，第二种API接口相对更加通用。

模型的对抗测评需要用到对抗测试集，对抗测试集通常是由对抗攻击算法，攻击原始测试集所产生的新的测试集，攻击算法根据攻击强弱和攻击的条件可以分为四个等级（Level），具体如下：



L1-随机攻击和L2-盲盒攻击，相对攻击强度较弱，但是却广泛发生在真实的业务中。L3黑盒攻击-对黑产攻击的成本较高，也需要一定的专业性，但攻击成功率一般较高，一旦攻击成功危害也较大。L4-白盒攻击需要攻击者拿到被攻击模型的详细信息，在实际业务不太可能发生，但是对于对抗防御和迁移攻击有较大意义。



对抗攻击的测评任务

对抗测评任务主要分为3个步骤：

1. 准备待测评的模型和测试集；
2. 选择合适的攻击算法，并产生对应的对抗测试集；
3. 使用对抗测试集对模型进行评估分析。

整体流程如下：



1.2 小样本学习与噪声学习

(1) 概述

在金融安全风控中，基于用户举报的案件标注（黑样本）相较于用户的正常交易（白样本）往往是比较稀少的，绝大多数高风险交易会在被风控体系拦截。与此同时，这部分因风控而中断的高风险交易（灰样本，也即噪声样本），因为缺少事后举报，无法进行有效的标注。这种标签数据的不平衡性，小样本，与噪声问题的叠加，给风险防控带来了巨大的挑战，也形成了完全不同于主流学术研究方向的算法问题。

(2) 小样本问题

学界关注的小样本问题，是所有标签的样本量都非常稀少的；而在金融安全风控中，往往只有黑样本稀少，相对应的白样本与带噪声的灰样本并不稀少。因此学界的大部分小样本方案在风控中并不适用。在反诈风控场景可以采取如下MixUp方案：与Borderline-SMOTE等业界传统方法不同，在MixUp的插值方案中，新的数据点样本只会在原数据的邻域中生成。因为SMOTE插值方案往往会在不合理的区域生成数据，使用MixUp来的插值方案便显得更为合理。此外，MixUp框架可以在黑白样本之间进行插值，这样能产生更丰富的黑样本，也打破了SMOTE只能在同标签内插值的桎梏。最后，MixUp可以与过采样结合，形成了一套带风控独有特点的全新解决方案。

(3) 噪声样本

上文小样本问题的另一个解决思路，便是从大量的灰样本中挖掘风险信号，并补充到黑样本中。因此，风控中噪声样本问题本身便是小样本问题的一个解决方案。

在业界传统的解决方案中，会根据人工专家经验，利用事后信息，把灰样本中的一部分标记为黑样本加入到模型训练中。这样可以显著提升模型的分类效果。但此方案也严重依赖于人工专家经验对事后信息的判断，无法推广到更多风险类型上。

此外，常用的解决不确定样本问题的Positive-Unlabeled Learning框架，并没有给风控中白样本留有可适配的空间，仅靠黑样本与灰样本是无法解决风控中的噪声问题。

结合学界最新的研究进展，当前行业里有两套相对成熟的解决方案

Partial Label Learning

在Positive-Unlabeled Learning框架的思路中，灰样本可以被表达成黑白标签按概率的混合。把这种思路与全新的，基于logit的自蒸馏框架相结合，便可以有效的解决噪声标签的问题。

自监督

另一个解决方案来借鉴于图像领域的自监督学习。在包含黑白灰的全量数据上，训练自监督masked autoencoder模型，然后在仅含黑白数据上做fine-tuning。这也可以有效的解决噪声样本问题。

可信基础算法-AI可解释

1. 概述

AI可解释是一种用算法解释AI模型，方便人工理解AI模型，进而运用这种理解迭代模型、处置业务问题的技术。具体来说，研究AI可解释具有三重意义：

01 避免黑盒

预防灾难或重大缺陷。比如AI医疗、自动驾驶，这些场景的决策失误容易导致灾难性后果，如果我们对AI决策的原理不够了解，就将大幅增加此类严重问题发生的几率。

02 新的知识

我们也希望在机器做得比人更好的一些场景，能够带来新的知识挖掘。当AlphaGo走出人类成百上千年的围棋历史中没有走出的招法时，我们希望能够更好地学习AI。

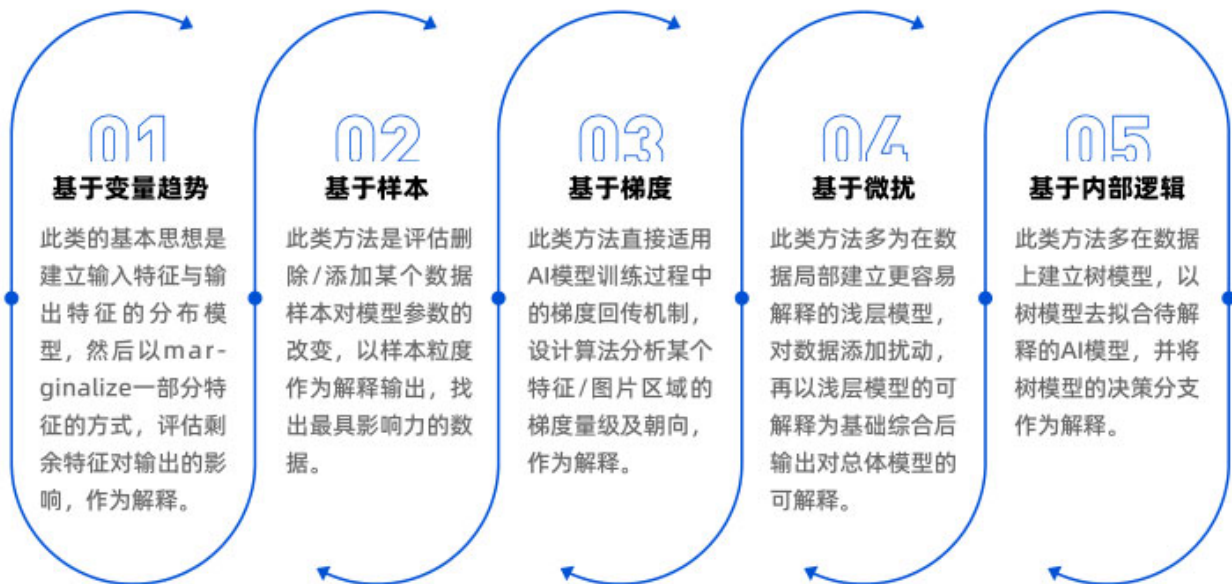
03 合法合规

无论是面对监管还是用户，可解释能够增强人们对AI的信任，对提高AI模型的隐私性、公平性具有重要意义。

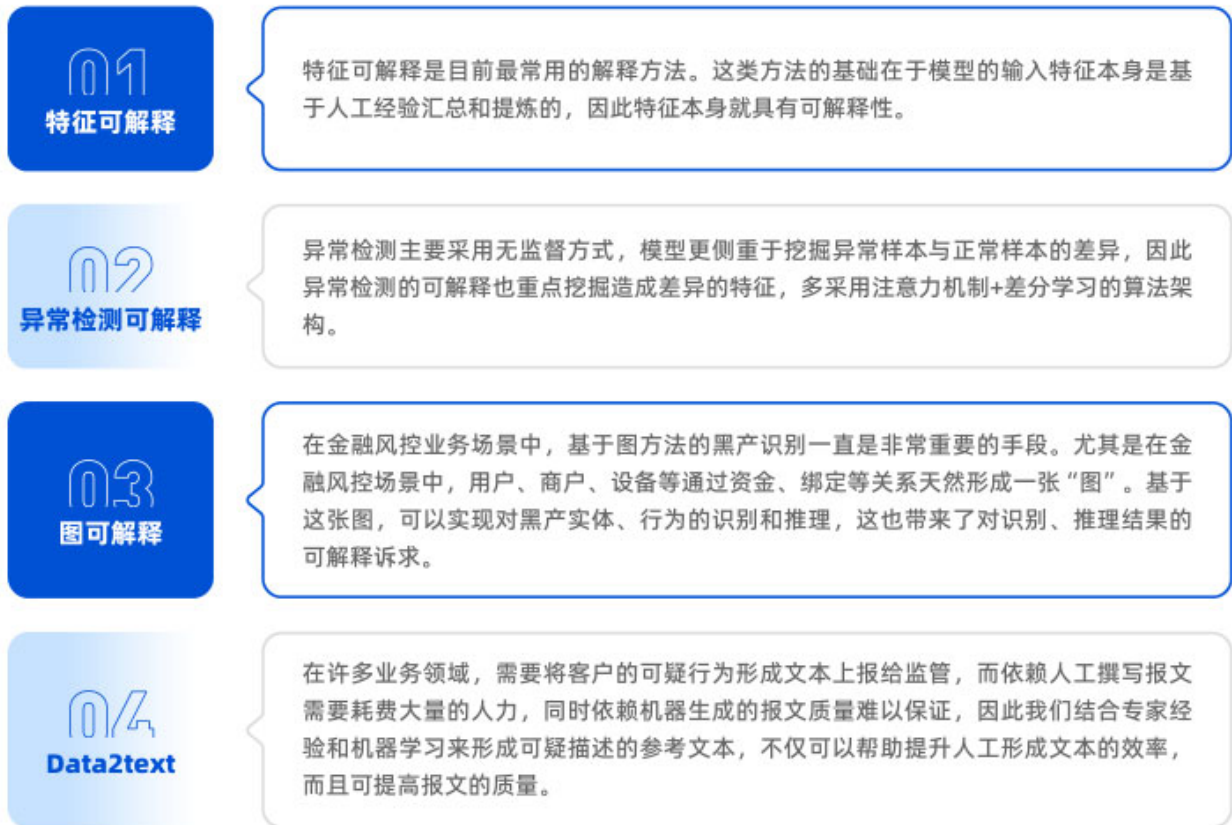
2. 可解释技术体系

归纳法	基于变量趋势	- Partial Dependence Plot (PDP) - Individual Conditional Expectation (ICE) - Accumulated Local Effect (ALE) Plot
	基于样本	- Influential Instances - Counterfactual Instance
演绎法	基于梯度	- Saliency Maps - Gradient*Input
	基于微扰	- LIME - SHAP
	基于内部逻辑	- Treeinterpreter - XNN / PLNN

AI可解释以方法大类来划分，如图所示，大致有五类：



AI可解释目前的应用方向有以下几个，上述5大类方法都可以按需应用于以下任何方向：



05 逻辑可解释

基于相似历史判例进行判别，有助于判别的稳定性与解释性。相比单一分数阈值的模型输出，提供更丰富的、带有聚类性质的历史相似案例知识。若需要历史判例的修正，可以体现在未来判别。

3. AI可解释核心技术

AI可解释用到的具体技术非常多，包括但不限于以下几类：

01 特征可解释

这类方法通过分析模型的输出与输入特征的关系，给出对输出影响较大的特征。常用特征可解释方法有SHAP、ELI5等，以及他们的衍生方法如Tree-SHAP。

02 差分学习

用异常样本和正常样本构建三元组，用Triplet Loss让模型学习三元组中数据的差异。基于该方法的可解释算法已经被证实是目前最适合用于异常检测的可解释算法。Integrated Gradient等梯度可解释算法也可以归为此类。

03 注意力机制

是用神经网络为某个中间层输出一个相同shape的权重层，与该中间层相乘，网络通过调节权重达到控制后续网络层对该层不同元素的注意力大小的目的。在可解释算法中，可以在输入特征层增加一个注意力层，该层的权重可以用来评估特征的贡献度，起到解释的作用。

04 深度表征学习

用神经网络将输入数据映射为一个表征向量的技术。在逻辑可解释中需要借助该技术对风控案件、事件进行向量表征标准化，借助使用快速向量检索算法、存储能力实现对历史案件、事件的快速匹配。

05 图可解释

图可解释问题可拆解为点可解释、边可解释、子图可解释。点可解释问题代表性的方法是PG-explainer；边级别可解释的代表方法有KPRN；子图可解释的常用方法为通过采样+图同构判定，实现风险子图挖掘，输出子图作为解释结果。

06 文本生成

传统文本生成一般着重在表面的客观描述，而现在研究趋势是如何生成内含一定逻辑推理的文本，这需要新的图谱构建方式，将专家知识以逻辑图谱的形式进行表达，以提高复杂逻辑文本生成的质量。

03

事前风险感知
与预测



3.1 事前风险感知概述

在金融反诈领域，传统的思路是在诈骗风险发生时进行防控。这种思路存在两个问题。一个问题是在风险发生时再进行防控，需要大量的实时计算，有可能出现计算资源不足，计算结果不可用的情况。另一个问题是风险发生时，我们能利用的数据信息不足，结果不可靠的情况。针对这两个问题，相应的解法就是将一些数据收集工作和相应的计算任务前置。孙子兵法《计篇》中说，夫未战而庙算胜者，得算多也，未战而庙算不胜者，得算少也。多算胜，少算不胜。孙子强调了战前庙算的重要性，而庙算的前提就是信息，获取更多的信息和数据，就能知己知彼，立于风控的不败之地。所以，智能金融反诈的重要一环，就是事前风险感知与预测。风险感知是在风险发生前获取相关数据，预测是对相关数据进行事先的计算。感知与预测是同一阶段的上下游环节，下文不做明显的划分，统一称之为风险感知。

就数据来源而言，我们将一个金融组织内部的业务系统产生的信息称为域内数据，将外部可以获取到的数据称为域外数据。所以事前风险感知，天然的分为两个大类，即域外风险感知和域内风险感知。域外风险感知和域内风险感知的区别是数据来源和计算方法上的。域外风险感知的重要目标就是捕获域外数据，以补充丰富域内数据。域内风险感知和事中防控的区别是时间上的，事中防控是在欺诈风险交易真实发生时刻，才进行风控决策和处置。而域内风险感知是在交易发生之前，例如用户在注册、认证、登录等环节，对相关数据进行汇聚和计算的工作。

本章侧重于对域外信息进行事前风险感知。域外数据来源庞杂，种类众多，可以粗分为两个大类，即实体信息和情报信息。实体信息是关于一个站点，一个APP有关的信息。情报信息是关于舆情，用户投诉等内容有关的信息。所以我们将事前风险感知分为全网威胁探测处理和全网情报采集处理。我们将对这两种事前风险感知进行详细介绍。



3.2 全网威胁探测处理

全网威胁探测处理的工作，主要处理来自于APP、URL网站、小程序等站点的数据和风险预判。和某域外公安数据源交叉对比发现，有70%以上的案件与APP或网站有关。如果我们能及时发现这些风险站点，就能为内部风险防控带来巨大增益。

对互联网站点进行风险定性，并深入站点进行信息挖掘的任务，我们称之为巡检。通过巡检将被动风控变为主动风控，是风控体系的一大突破。下面，对全网巡检的系统架构和算法体系分别介绍。

■ 全网威胁探测处理系统架构

全网巡检系统架构包括了四层结构：全网主动发现、定性引擎、巡检引擎和站点形式分析。



非法四方支付平台：早期黑产站点在收款时会用自己的账号，但这种方式很容易被稽核，于是发展出四方支付平台，他们整合第三方支付公司和银行的支付能力，为黑产提供支付结算服务，这种平台背后一般会有成千上万个账号，而这些账号很多都是租来的，会有正常用户在使用，在很多笔正常交易中会掺杂一笔黑产交易，所以很难发现。

■ 全网威胁探测处理算法体系

算法体系中输入数据是全网站点，包括APP、网站、小程序等，输出则是对这些站点的风险识别结果。算法首先会对站点进行特征提取，提取的特征包括首页截图、源码、图标、IP地址、OCR文本等。这些特征作为站点定性和站点图谱的输入。站点定性模块接受包含站点内容的特征（包括首页截图、站点源码）进行风险定性和行业定性，定性结果一方面输入到站点巡检对黑产站点进行态势感知，产出风险内容，包括账号、黑产四方等；另一方面定性结果会输入到基于自监督图表征的站点图谱，由站点图谱进行图算法分析挖掘，弥补站点定性内容缺失的问题，提升定性的覆盖量级。最后流量巡检对站点图谱、站点定性、站点巡检产出的非法四方或站点进行流量监控，产出黑产账号，该模块通常采用时序聚类算法。

该体系主要由定性算法和巡检算法两大核心算法组成，下面作具体介绍。

1. 定性算法

定性算法是为了判断某个站点是否和某种风险有关。其的整体流程包括内容页定位、新风险发现、内容定性、站点图谱四个部分。

01 内容页定位

根据站点的内容进行定性，其中主要的内容来自于站点的首页截图和其中的文字信息。通常打开站点首先显示的可能是广告页、加载页、注册页等，需要依赖多模态页面分类和模仿学习来识别和跳过这些页面才能获取有效用于定性的内容。

02 新风险发现

挖掘到内容后我们会进行新风险发现，外部的风险形式是不断变化的，经常会冒出一些新的风险，像最近比较火爆的虚拟数字藏品的欺诈，我们希望快速发现这些新的风险来不断完善我们的定性体系，提升风险类型的覆盖。我们通过异常样本识别和新样本聚类来发现新风险。

03 内容定性

我们主要的内容信息是首页的图文信息，但是有时候会出现首页打不开或渲染不成功的情况，我们也引入源码中的内容进行定性。我们也探索了多模态融合、模态缺失和带噪学习等技术来提升定性的准确率和覆盖率。

04 站点图谱

最后我们用站点图谱对识别出的风险站点进行黑种子传播。站点图谱两个比较重要的部分是表征提取和图自监督表征。人工对站点提取的很多特征通常无法直接使用，例如icon、logo、atg图等，我们需要通过自监督算法提取比较有意义的表征来进行下游的检索和聚类。同时自监督图算法可以更好地融合不同的表征，提升黑种子扩散的准确和覆盖。

2. 巡检算法

巡检是对目标站点进行深入下探，挖掘可疑内容（收款账号和非法四方支付平台等）。我们对整个巡检流程进行拆解，将其分为三个模块：巡检目标管理、注册登录、支付寻路。为提升巡检的灵活度，我们引入了页面分类、按键识别等算法串联整个流程。以下，具体介绍三个模块：

01 巡检目标管理

在巡检之前加入巡检目标管理模块，可以帮助我们快速发现主要问题，提高模型迭代效率，实现运营智能化。

02 注册登录

在注册登录模块我们引入表单识别算法和多种验证码破解算法，提升注册登陆成功率，实现注册登录智能化。

03 支付寻路

其目标是进入充值页面模拟**支付路径**，为此，我们引入智能寻路算法，提升寻路的成功率，实现支付寻路智能化。

支付路径：类似选择支付渠道、选择支付方式、输入充值金额、确认支付这样的支付流程。



3.3 全网情报采集处理

全网情报采集，是指从舆情、黑灰产情报渠道中，提前获取风险相关信息。目标是将外部信息，引入体系内进行风险布控、并成功检测出该类案例。然而，域外的情报数据是海量的，包括且不仅限于舆情信息（盗用、欺诈、非法融资）、黑灰产（TG群聊、白帽子）、客服来电、投诉等内容。情报智能引擎从海量情报中提取有用信息是一项艰巨的任务，需要通过多种技术的整合应用，例如语音识别、自然语言处理和计算机视觉技术，以进行内容理解，提高情报处理的速度。

一套情报智能引擎可以通过信息抽取技术，实现对海量的多源异构情报进行三个流程步骤处理，分别是情报分类、事件分组、要素提取。情报分类将情报分为不同的风险类型，例如，欺诈、盗用；事件分组将描述相同事件的不同表达形式的情报分至同一个组，剔除重复信息。在要素提取阶段，进行文本结构化和视觉文档结构化、提取事件要素，提高风险的感知和研判的速度。

经情报智能引擎处理后的的情报信息会被应用于各种业务场景，例如获取的欺诈、盗用事件涉及的app、商户会作为风控的名单；识别出来的风险身份证号会被旗下贷款服务机构做背调；识别出来有风险的公司，也会在商户准入、凭证审核时作为参考信息。



情报分类

情报中含有大量的无效内容，过滤无效内容可以大幅降低情报量级。由于场景多、风险类型多，造成建模需求多，可以采用常见的领域预训练+下游微调finetune的范式建设文本分类模型，并在技术实现上将领域预训练、层次分类、多标签分类算法进行组件化，提升开发效率。领域预训练阶段使用领域数据，进行全词隐蔽（Whole Word Mask）的预处理之后，通过MLM（Masked Language Model）任务构建基于Transformer的语言模型；下游适配阶段，对于短文本选择Roberta语言模型，对于多语言选择mBERT语言模型，对于长文本选择Bigbird语言模型。

事件分组

相同的事件可能存在不同表达形式的情报中，尽管其内容不完全相同，但表达的事件要素相同，为此，需要将情报进行分组。考虑到事件组的数量是不可枚举的，因此需要采用聚类相关算法。把重复事件分到同一个组里，可以快速压缩情报审核量级。情报是流式数据，需要进行实时聚类，因此，传统的DBSCAN、KMEANS方案不适用这个场景。

基于情报数据的特点，可以采用一个分层方案。

01 第一层

由于情报被转发后句子不变，通过句子指纹检索和精排得到组ID。

02 第二层

对于情报被改写的特点，拼接情报中的事件类型和元素得到组ID。

03 第三层

如果抽取不到合理的事件信息，采用无监督/有监督关键词抽取算法，例如textRank、LSTM+CRF等，并通过词组树得到组ID。

最后融入百度贴吧、微博盖楼的分组逻辑，将跟帖和主帖归并，最后使用simhash算法兜底。

要素提取

仅仅识别情报的风险类型是不够的，还需要通过更细粒度分析，将文本信息结构化，即识别情报中的事件要素，才能提高风险事件的研判处置速度。要素指的是欺诈受害人、欺诈地点、非法集资主体、欺诈黑中介、赌博APP等实体。



1. 文本结构化

对于纯文本情报，采用BERT-CRF基础模型进行实体抽取（NER，Named Entity Recognition）。但是由于实体打标非常耗时，所以有标签数据（Strong Label）是比较少，模型面临冷启动问题。可以利用好无标注数据（Unlabel Data）、实体词典以及启发式规则的角度出发设计方案。通过远程监督，对无标注数据打标得到带噪声的弱标签（Weak Label），也可以通过有监督模型在无标注数据上预测得到概率形式的软标签（Soft Label）。最后使用知识蒸馏的思路（Teacher-Student），将强标签、弱标签和软标签迁移到学生模型，达到扩展标签的效果。抽取到的实体可能并不是标准的名称，例如公司名称。还需要利用实体链指技术，依据实体所在上下文进行消歧得到标准的公司名称。

2. 视觉文档结构化

情报中还包含视觉文档型的数据，例如订单截屏、监管处罚公告、黑灰产证件等。由于视觉文档不像纯文本数据那样是序列数据，其中的实体类型和所处的位置有较强的关系。需要先进行布局分析，将文档拆解成不同的板块。对于不同的板块采用不同的解决方案。

例如，针对监管处罚公告的文书主体信息处理，先利用OCR（Optical Character Recognition）技术抽取情报图片中的文字及其位置信息，再利用卷积神经网络（CNN）进行文本编码，最后利用图卷积网络（GCN）得到文本编码和OCR位置特征的融合编码，通过融合编码增强文本表征。这样的方法称之为视觉NER。然而，正文部分属于自由文本，则使用BERT-CRF和规则配合通用NER进行辅助提取。



事中攻击检测 与防护



4.1 事中风险防控概述

事中风险防护是风险交易发生时进行识别并做出风险决策的过程。

事中防控的首要任务是风险识别，判断当前交易是否存在风险，存在什么样的风险。从技术上区分，风险识别可以分为有监督方法和无监督方法。对于新业务冷启动、无准确样本反馈等业务场景，最大问题是缺乏有效风险标签，主要依赖无监督异常检测发现风险，将展开介绍基于[结构化数据的异常检测](#)和基于[图的异常检测](#)这两种重要技术方案；针对有风险样本的具象风险防控，核心技术是有监督方法，在当前时代背景下，金融服务往往通过用户端侧提交服务请求，云侧完成服务确认的方式形成闭环，系统需要同时在端和云两侧同时做好[实时风控](#)。在云侧，[全图风控技术](#)能通过挖掘用户/黑产多维度的关系拓扑信息进而解决黑产隐蔽性高、信息孤岛等挑战，是当前事中防控前沿而有效的核心技术。在端侧，移动互联网时代用户在端上的行为蕴含巨大的信息量，技术核心挑战是如何在考虑用户隐私保护因素下进行[端风控](#)。

有效的风险决策是识别能否起作用的关键。决策首先要做的是平衡风险与体验，识别一定会存在准确率问题，在识别能力一定的前提下，如果选择更全的风险召回势必带来更多误打扰，如果选择降低打扰就会造成更多的风险漏过。所以决策水位如何制定要依赖业务考量动态调整，这套解决方案即[风险决策](#)。进一步，如果能够在决策环节和用户交互，获取到更多用户信息，判断当前被骗的阶段、情绪等状态，就可以做更好地[风险阻断](#)，这个方案就是[交互式风控](#)。对深度被诈骗的用户来说，传统的失败交易、出具核身校验等方式都不再有效，这些场景下交互式风控的风险阻断能力是非常显著的。



4.2 异常检测

异常检测的一般目标是在无可靠风险样本的情况下，用无监督方法发现业务中隐藏的异常，从而驱动风控策略的升级提升防控水位，这在业务冷启动或者针对营销作弊等无风险反馈的业务问题中是核心手段。

在反电诈风控业务上，我们一般面临着两种不同数据源，第一种是结构化数据，如一条交易信息里的数据可能由商品信息、对手方信息、介质环境信息等结构化的数据组成；第二种是图数据，如一个交易资金网络由交易方、设备、银行卡等实体，以及实体之间的连接关系构成，这种数据里蕴藏丰富的风险连接关系。风险异常检测的任务就是要从这两种数据中找出可能的异常信息，技术上来说不同数据源需要不同的处理方式，以下分别展开介绍。

■ 基于结构化数据的异常检测

1. 概述

黑样本标签的获取在很多场景中往往比较困难，例如欺诈场景需要用户主动投诉，赌博、洗钱场景隐蔽性很强，有时甚至完全无法获得。传统的有监督学习技术在这种情况下难以得到施展，这给欺诈风险防控带来了极大的挑战。在无标签场景下，系统如何从数据中快速发现异常情况、及时感知并定位潜在风险，是亟需解决的重要课题。针对这一情况，可以采用基于无监督异常检测技术的通用解决方案。

从异常检测的角度出发，可以风控场景中比较常见的各种“异常”按照异常模式划分为三大类：

01

单点异常

指单独看某条数据样本就可以发现它跟大多数的数据样本不一样，比如数据分布中比较孤立的样本点。

02

上下文异常

指要根据数据样本所在的具体“背景”或者“语境”才能判断的异常，比如时序数据中的异常样本点。

03

集合异常

是指单独看某条数据样本可能无法看出异常，但是把多条数据样本放在一起时就可能发现不正常，比如团伙类风险。

2. 无监督异常检测方案及主要技术

解决方案主要包括以下四个模块：

01 特征筛选

实际业务中经常出现数据缺失、填充默认值等低质量字段，或者重复建设的字段，特征筛选技术用以去除冗余、噪声特征以及相关性比较高的特征，避免这些特征对模型带来的负面影响。冗余特征、噪声特征或者相关性比较高的重复特征，对于异常检测算法的影响较大。常用的特征选择方法有CUFS(Coupled Unsupervised Feature Selection)、CORR方法等。CUFS是通过构图的方式综合分析特征冗余度和噪声，进行过滤。CORR则是通过简单的相关性检验过滤特征。

02 集成检测

通过集合主流的异常检测算法，基于集成算法的异常融合，即将多个单一的异常检测算法的预测结果融合为准确率更高的预测结果，并通过知识蒸馏的方式，将实际的推理模型压缩为一个轻量级、可部署在亿级规模数据集的小模型。核心技术为集成学习(Ensemble Learning)：通过设计集成算法，将多个算法的预测结果融合为一个结果。融合后的结果通常准确率要高于任何一个单一模型的结果。异常融合常用的方法有SelectH、Kemeny Young、Inverse Rank Aggregation等。单一异常检测方法，常见有聚类、重构、树模型、深度学习等方式，代表性方法有LOF、AE、iForest、OCNN等。知识蒸馏是一种将一个轻量级模型与大模型一起训练的方法，这样训练出的轻量级模型能够以较小的计算量获得与大模型接近的预测效果，可以用于模型压缩，适用于在大规模数据集上运行。

03 异常归因

通过树模型或NN模型可解释技术，输出异常样本同正常样本之间区别最为明显的特征信息，用以反映该样本被识别为异常的主要原因，对于帮助风控专家快速验证潜在欺诈风险案例非常有帮助。主要涉及技术为模型可解释技术，做法为训练一个解释模型，分析检测模型的输出与输入特征的关系，对检测模型的预测行为做出解释，让人理解一条数据被识别为异常的原因。这项技术对于特征筛选、模型迭代、模型结果触达给实际业务具有至关重要的作用。常用的两类一类是树模型归因，这种方法是训练一个树模型拟合检测模型输出，然后分析树模型的决策路径；另一类是NN归因，这类方法常用NN模型+注意力机制对检测模型做拟合，以注意力权重作为解释。

04 反馈学习

由于风控场景的特殊性，“异常”并不一定等于“风险”。为了让无监督集成式异常检测结果的指向性更明确，更偏向业务专家关心的风险模式，增加反馈学习模块。跟风控专家交互之后的反馈信息，借助半监督学习和弱监督学习技术，利用少量的专家输入信息，进一步提升异常检测的准确性。常用的反馈学习有半监督学习和弱监督学习，半监督学习可以兼具无监督的分析数据分布，与有监督的学习标签信息，达到仅使用少量标签就能得到远远优于无监督算法，接近有监督算法的效果。常用的半监督学习算法有DevNet、XGBOD等；弱监督学习解决标签不准确、不确切或者不完全的问题。在异常检测人机交互过程中，把不同异常检测模型的结果作为弱标签，同时配合利用专家打标的极少部分真实标签来迭代式训练得到更准确、稳定的异常检测模型。

基于图的异常检测

基于图的异常检测是风控中非常重要的一部分，利用图的结构可以挖掘业务中的潜在风险。在交易支付过程中，用户和用户间的资金关系、用户和设备间的使用关系等构成一个大的关系网络。图异常检测就是从这个大的关系网络中挖掘风险。图异常检测可以分为三个维度，分别是点异常检测、链路异常检测和社区异常检测。其中点异常检测可以使用图的方法，发现针对目标节点的异常；链路异常可以发现存在风险的关系路径，这个路径可以是一些绑定关系，或者是一些资金链上的上下游关系；社区异常主要来发现关系网络中存在的团伙，可以通过社区检测等方法发现潜在的风险。



1. 点异常检测

点异常检测是指确定某个目标节点是否存在异常，因此针对点异常检测问题，首先会从当前点出发，基于图关系进行拓展得到一个子图，然后判断当前节点所在的子图是否存在异常，如果存在异常，那么就会对当前的目标节点进行管控。在金融风控当中，根据不同的业务逻辑会有不同的构图方式，发现异常之后也会有不同的处置方式。比如在用户注册场景，如果发现新注册的用户跟历史的欺诈团伙有很多资金关系，那就可以认为当前注册的用户是有风险的，可以对注册进来的账户进行一些权益上的限制。

针对节点的异常检测，一般通过基于规则的方法和基于监督学习的方法训练得到。基于规则的方法会计算当前的目标节点是否与已知的异常存在显式的关系，并且在特征上存在一些共性；监督模型一般基于图神经网络GNN模型对目标节点进行打分，如果显著高分可以判定为异常。一般来讲基于规则的方法的解释性更强、GNN方法的覆盖比较高，但解释性可能稍弱。

2. 链路异常检测

链路异常检测可以定义为两个问题，第一个问题是已知特定的风险链路模式，在关系网络上进行搜索和匹配，另外一个问题是自动地发现关系网络中存在的一些异常链路或者拓扑结构。

问题一

以支付风控场景为例，在反洗钱风险中，经常存在一笔资金的快入快出，需要从某笔特定的交易出发找到快入快出的资金链路模式。除此之外还有一些其他类型的风险模式，比如资金闭环检测、菱形网络检测等，这些都是给定一个特定的模式或者拓扑结构，在关系网络中找到对应的账户。不同类型的拓扑匹配依赖不同的类型的算法，一般来讲资金闭环检测、菱形网络检测这种属于严格匹配，一般通过数据结构算法即可实现；对一些洗钱的链路模式，需要通过多目标优化算法来解决，要求金额相似、时间相近等多个目标，同时也需要增加一些约束条件。

问题二

存在风险的链路或者拓扑结构并不一定能够完全被业务专家总结出来，因此需要通过算法来找到关系网络中频繁出现的一些链路结构，因为整个关系网络存在时序信息，因此这个过程称为时序图上的风险motif挖掘，把风险motif挖掘出来之后，同样的需要在整个关系网络中进行匹配，找到同样有类似链路结构的账户。算法上一般通过频繁子图挖掘挖掘潜在的风险结构，然后在推理阶段使用子图匹配的算法来进行。业务对应的图规模可能非常大，因此频繁子图挖掘往往是非常耗时的，未来的一个研究重点是如何在大规模图上进行风险子图的挖掘。

3. 社区异常检测

社区异常检测，主要来发现关系网络中存在的异常社区，根据数据的结构，分为同质图上的异常社区检测、异质图上的异常社区检测、属性图上的异常社区检测。存在风险的社区一般称之为“团伙”，下面分别介绍下这三种数据形式下的社区异常检测。

同质图

针对同质图（节点和边的类型都只有一种）上的团伙挖掘，一般使用基于Louvain、LPA等社区划分方法，将关系网络中紧密的社区结构挖掘出来，一般来讲在挖掘的紧密社区中，如果一部分节点（账户）已经被证明是存在风险的，那么剩下的账户大概率也是有风险的。

异质图

与同质图不同，异质图是指图上节点的类型不止一种，面对异质图关系，可以对社区发现算法对应的公式进行修改，从而处理异质图上的社区发现，比如针对二部图，可以用二部图社区发现算法如Bi-louvain发现社区关系。总之，异质图上的社区检测主要是把不同的边类型、特有的拓扑结构考虑进去进行社区发现。

属性图

属性图是指图中的节点存在不同类型的属性，在进行社区划分的时候，不仅要求结构之间要紧密，而且要求同一个社区内的节点属性比较相似。在一些场景中，节点上的属性包括多个不同的部分，比如名称、地址、统计特征等，因此需要针对这种数据设计专门的属性图聚类方案。

4.3 实时风控

在移动互联网时代，实时风控需要我们从服务侧（云侧）和用户侧（端侧）一起做防控。云侧的优点是，可以获取到用户与用户、用户与介质之间动态的、全量的连接信息，我们可以利用全图风控技术进行高效能的风险识别；端侧的优点是，不需要将敏感信息传到云侧做计算，可以直接在端上处理的方式保护了用户隐私，而且技术上可以针对每个用户生成个性化的模型提升端侧防控性能。本节展开介绍这两项实时防控技术的核心功能和技术。

全图风控

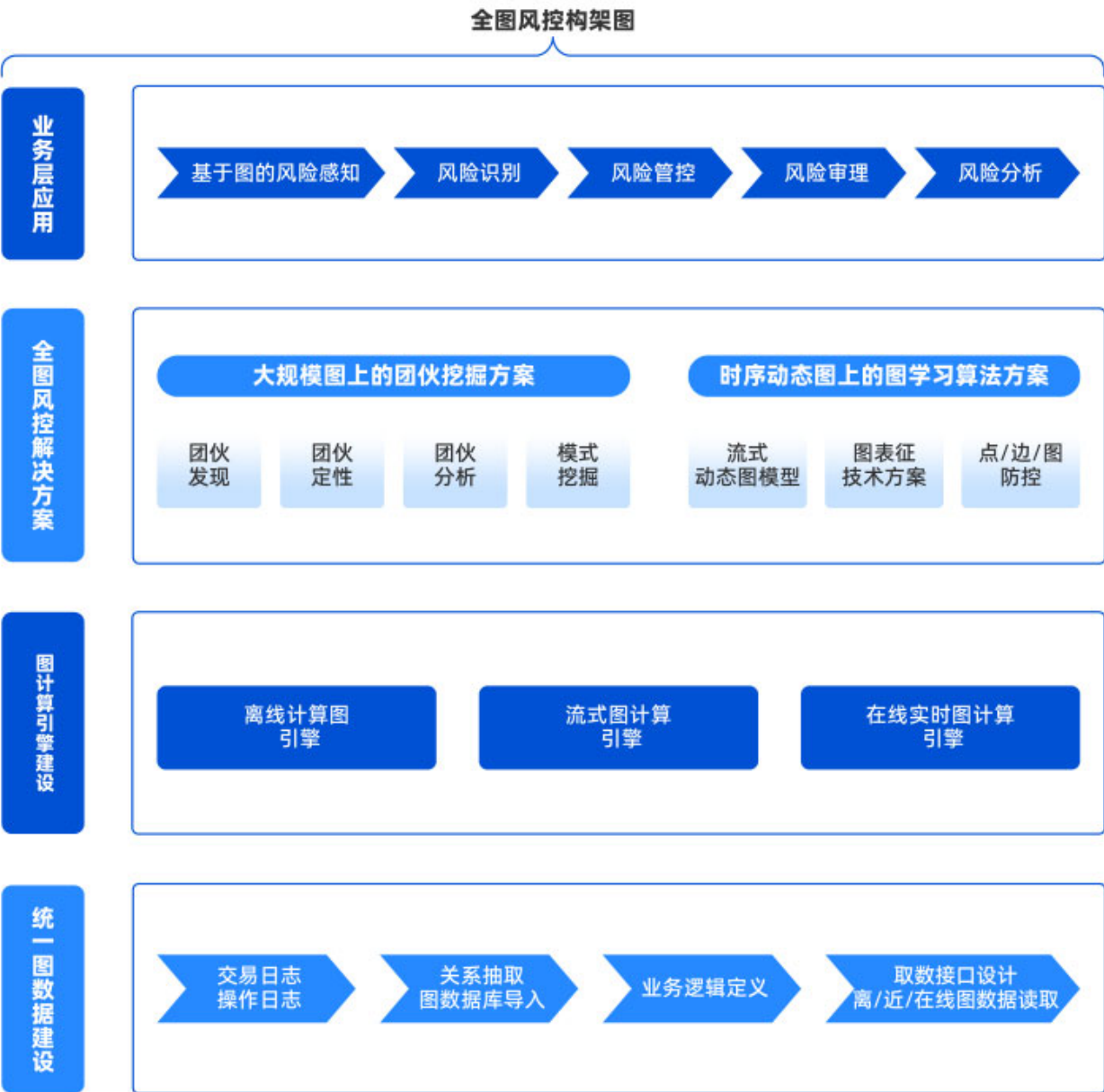
1. 概述

随着数字化的深入，安全风险形势也在发生深刻的变化，电信诈骗等风险手段不断升级且手法日趋隐蔽，风险识别难度越来越大，原来依靠手动设计交易统计特征的方式越来越难以捕捉黑产行为，需要对黑产信息刻画升维，而图是一种关键的信息载体。全图风控的目的就是在现有风控识别的基础上，挖掘用户多维度的关系拓扑信息，解决黑产隐蔽性高、信息孤岛的挑战，实现高性能的风险图识别体系。



2. 图风控功能架构

全图风控功能整体自底向上可以分为三大模块。第一部分是图计算的基础设施，其中包括统一图数据建设、图计算引擎建设、以及配套的风控引擎，能够支持千亿级别数据量的构图以及高性能的图数据更新与查询，提供多维度信息提升风控的识别性能。第二部分是全图风控解决方案，通过图算法加工图数据，产出团伙挖掘、图表征、图风险识别等上层数据资产助力风险发现，主要包括近线秒级团伙挖掘、图自表征学习技术与流式图实时风控决策模型，基于全场景多维度图数据，在风险交易发生前，对团伙及其他批量异常风险要素提前感知，在交易发生过程中进行精准防控。第三层是顶层的业务应用，包括基于图的风险感知、风险识别、风险管控、风险审理、风险分析等，全图风控的架构图如下：



3. 图计算基础设施建设

01 图数据库 建设

全图风控需要依赖业务系统中产生的海量图数据。一般来讲，业务系统中任意实体都可以当做全图中的“节点”，实体之间的交互行为都可以构成全图中的“边”。比如在交易系统中，账户、设备、银行卡等都可以作为一个节点，账户到账户之间的交易，账户和设备之间的绑定/登录等关系都可以作为一条边。这些在业务系统中产生的图数据是全图风控引擎的“燃料”，后续的风险识别、防控等都基于这些图数据关系。一般来讲，可以通过关系型数据库或者图数据库来进行存储，典型的存储方案有各种本地存储如ORACLE等，云存储如阿里云odps等，图数据库GeaBase等。为了有效的读取数据，需要设计相关的读取接口，读到业务需要的图数据。数据规模上一般支持在线/离线千亿、近线百亿级数据量的点边构图。

02 图计算 引擎建设

图计算引擎用来处理海量的图数据，根据图算法的复杂性和使用数据的不同，一般可以分为离线、流式、在线图分析引擎等。离线图计算引擎包括传统的基于python的networkx，基于spark的GraphX等，流式图引擎可以基于分布式计算引擎技术如Ray等开发，在线图查询引擎一般通过kv存储的实时关系查询，在高性能应用场景需要支持3度之内的毫秒级查询，5度之内的秒级查询。

4. 全图风控解决方案

01 图数据库 建设

团伙挖掘一般包括发现、定性、分析等步骤，也可以按照特定模式进行全图挖掘或者匹配。面对团伙作案时效性不断提高的挑战，在平衡性能与成本的基础上，可以基于近线图计算引擎实现端到端秒级团伙识别，在不同场景自动选择合适算法识别局部子图中紧密连接的团伙，并同时产出包含团伙规模，风险节点占比，未知风险节点占比等一系列通用团伙拓扑属性。这里的算法一般包括连通图、社区检测、基于业务经验的后处理等操作。

02 图计算 引擎建设

风控场景的图数据天然存在时序性，可以通过对时序性进行建模，从而捕捉其中存在的风险信息。动态图建模算法包括基于动态图的预训练技术、表征技术等，可以实现对节点、边、图等形式的防控。典型的算法包括TGAT、EvolveGCN、SDGNN、DDGCL、AMAP等。

5. 基于图方法的业务应用

基于上述的团伙挖掘、图学习等方案，运用到业务系统上，可以实现对风险的感知、识别、管控、审理、分析等操作。所有的风险防控都基于图数据产出，在图计算引擎执行，结合业务系统的风控引擎进行防控。

端风控

1. 概述

为了提供精准的风控服务，通常的做法是数十亿的智能终端向中心服务器发送大量原始数据，中心服务器根据收集到的数据，进行一系列的大数据分析与建模，最终将风控结果返回给终端，进行后续的业务流程。这样的大数据风控模式，存在以下几个挑战：

01

低成本与高负载

对客户端来说，实时发送大量原始数据，需要较大的带宽成本，对服务端来说，实时处理大量的原始数据，需要较大的计算成本。

02

个人信息保护

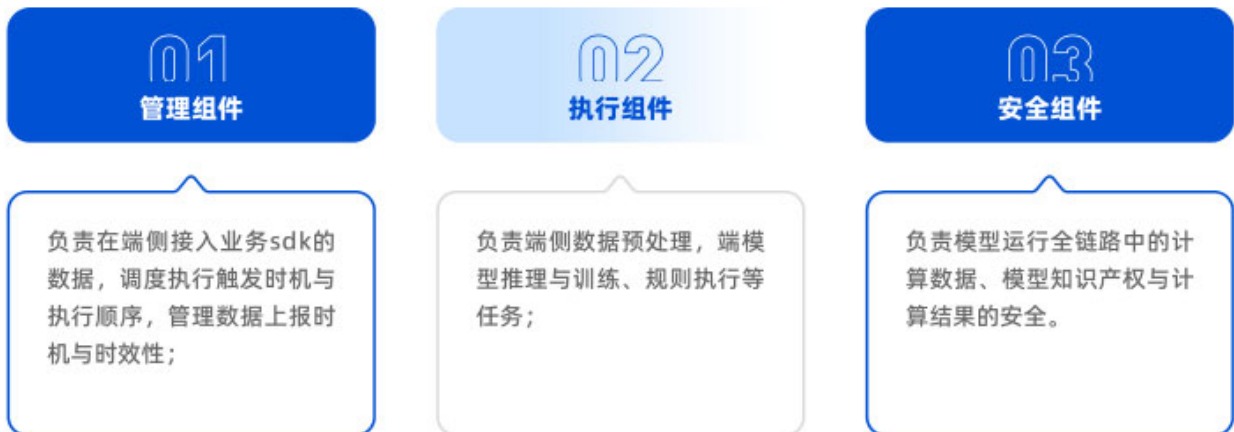
终端采集的数据，包含了用户与设备、与app、与环境交互的数据，这类数据有可能存在个人敏感信息，原始数据在服务端进行保存与处理，会带来数据安全风险。

随着终端硬件的升级，大部份的用户手机终端都拥有强大的算力，可以支撑深度学习模型的推理和训练。端智能技术是指在智能终端进行特征、规则和模型的计算，将处理后的结果上报到服务端，可以有效节省带宽成本与中心服务器负载压力，同时可以使得敏感数据留在用户本地终端，彻底避免个人敏感数据滥用与泄漏的风险。端云协同风控综合运用端智能与中心服务器大规模计算的能力，实现能力互补，共同提供体验最优、个人隐私保护的风控服务。

2. 技术架构



端云协同风控架构分成client侧与server侧两部分，互相传递加工后的结果数据与模型参数数据等。client侧可以分成三大部分，包括管理组件、执行组件与安全组件。



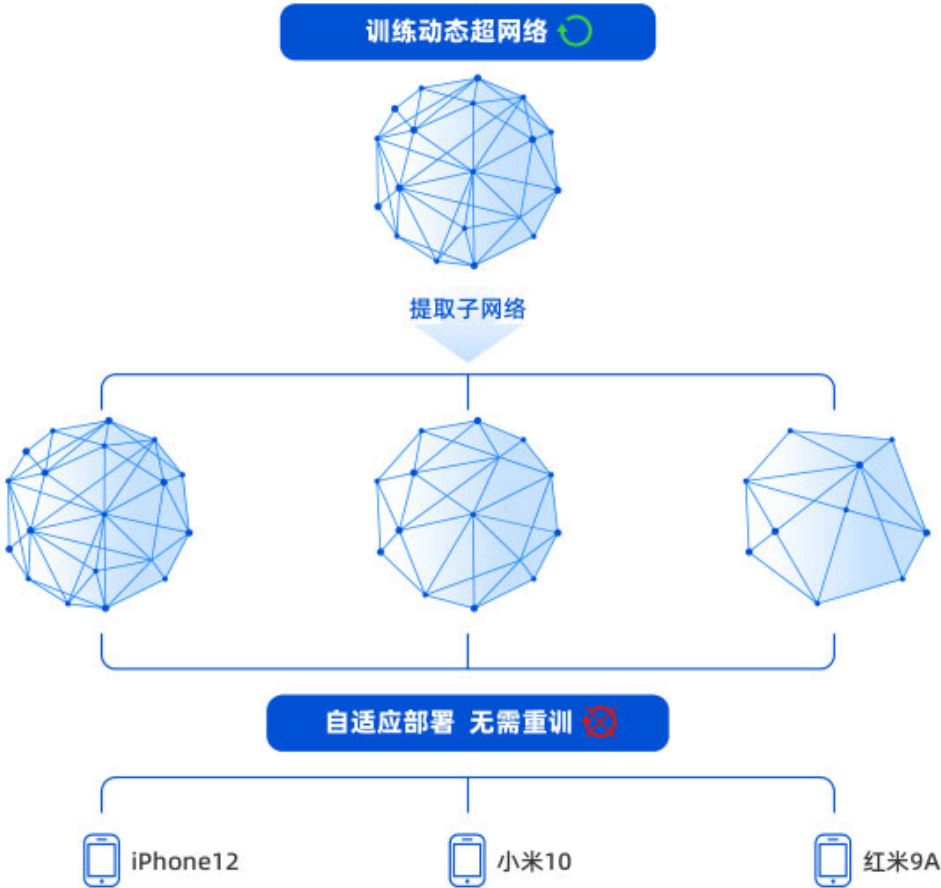
server侧包括端模型的研发工具，端模型客户端发布工具和紧急情况下动态下发的应急管理能力。server侧的数据组件负责接收端侧上报的数据进行存储与轻量的汇聚加工，由策略路由组件判断如何与服务端的风控引擎交互。

以资金风控应用举例，服务端风控引擎每天需要处理数亿笔交易，每一笔交易背后有上千个实时特征和上百个实时模型，风控耗时面临很大的压力，大促等高峰期还会有超时的风险，影响用户体验。在端侧利用app内交互访问数据进行交易综合风险的初步判断，模型结果和策略结果加密上报至服务端，无风险的流量直接放行，有风险的流量流入服务端风控引擎进行深度分析，提升用户体验。

3. 核心技术

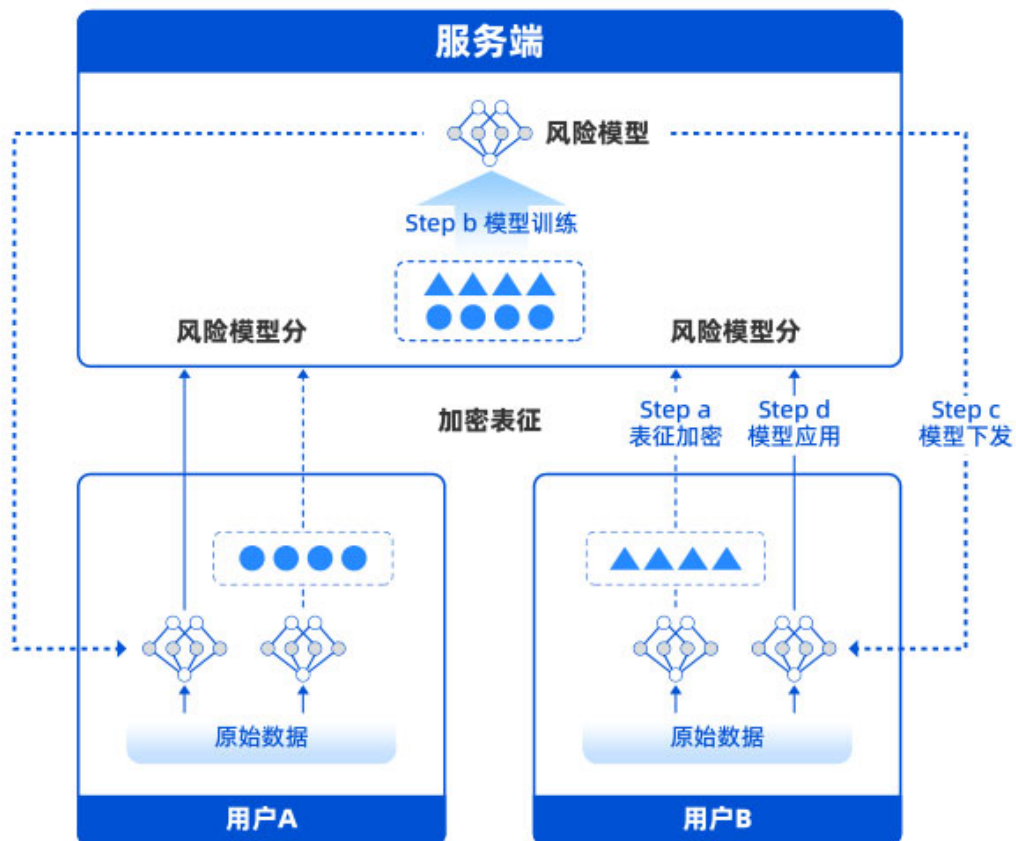
模型压缩技术

模型压缩是指为了让风控模型运行在资源相对受限的智能终端，通过一系列的手段减少模型的尺寸与推理耗时，包括模型结构搜索、模型剪枝、模型蒸馏、模型量化等。由于智能终端的性能差异较大，为了保障低端机用户的体验公平性，还需要千机千模的解决方案，即通过神经网络搜索技术，自动化地对每一种机型定制不同尺寸的模型，提升每一种机型的推理成功率。



首先需要严格按照个保法的要求，处理用户个人信息，并且站在用户视角，对部分用户无需授权、或者已经明确授权的信息，利用端智能的隐私计算技术进行数据最小化使用，进一步提升数据安全。具体步骤如下：第一步在非隐私数据集上训练表征模型，部署到用户终端，将原始数据转换为向量表征，然后对表征进行混淆加密，可使用的典型算法有InstaHide，将表征加密后上传至服务端；第二步服务端收集到加密的表征后，进行风险模型训练；第三步将训练好的模型下发部署到终端。这种方案，在模型训练阶段，用户上传的是加密后的表征，模型应用阶段上传的是风险模型分，都无需上传原始数据，可以降低或者避免隐私泄漏风险。

此外，方案中表征模型也可以通过联邦学习的方法训练产生，端到端训练得到的表征模型更符合实际数据分布。联邦学习是一种分布式联合建模技术，可以联合多个设备，在保证原始数据不离开设备本地的基础上，共同完成模型训练。具体做法是，客户端从服务端下载模型，根据本地的数据进行训练，然后将训练后的结果上传至服务端，进行聚合，完成一轮训练。相比于集中式建模需要上传原始数据，联邦学习中只需要上传模型梯度，可以保护数据隐私。联邦学习在落地应用中，端侧AI引擎需要具有训练能力、样本管理能力、训练编排能力，云侧AI引擎需要训练调度能力、模型聚合能力，并且需要引入TEE、多方安全计算等技术，增强训练过程中的安全性。联邦学习的训练效率比集中式建模低，且技术成本比较大，可以根据实际情况进行技术选型。



个性化建模技术

端侧每台设备的模型是用户专属的，使得极致个性化的模型部署成为可能。由于风控场景单用户样本较少，千人千模收益并不稳定，更适合的是对用户进行分群，为不同的群体部署个性化模型。在风控场景，少部分活跃的群体贡献了大部分的数据，对所有数据统一建模，会使得对小众群体拟合不够好，难以保证小众群体的公平性。可以通过元学习的方法，利用群体数据的同时，为每一个群体更新一个个性化模型，可采用的典型算法有MAML等。

模型安全技术

在端模型应用的全生命周期中，存在着输入输出篡改、模型窃取攻击等风险，模型窃取攻击除了影响模型知识产权，还会提升对抗攻击、隐私攻击的成功率。因此需要综合应用数据加密、安全权限控制等手段保障模型调用安全，利用模型计算图混淆、可信执行环境（TEE）等技术保护模型文件安全。



4.4 风险决策

■ 概述

风险用户往往带有多个风险域的识别结果，如欺诈、赌博、色情等。当感知到用户的风险之后，需要根据风险识别结果，结合用户的画像信息，评估综合风险水平，并根据决策方案采取合适的管控方式。无底线风险可以考虑放过，有风险就进行挑战或者处罚，直至风险可控。假如管控方式过于严格，用户会通过申诉的方式解除管控动作；如果风险防控不到位，则会导致风险进一步扩大。好的管控方案需要在做到好的风险覆盖的同时，平衡用户的被管控体验。

从最终目标上看，这是一个带约束的目标优化问题，典型的风险控制场景里，需要在用户打扰率（或其他体验指标）控制在一定量下，最小化风险漏过；有时候约束和目标可以倒过来，比如在体验保障场景里，需要在风险漏过控制在一定量前提下，最小化用户打扰率。这种情况下，就可以将问题纳入背包优化问题框架里解决，以风险控制场景为例，分两步走：

01

预测用户的管控效果：预测用户在每一个决策产品管控后的风险发生概率，以及管控后的客诉概率（或其他体验指标）。管控后满意度和风险率都是一个0-1之间的概率值，可以由一个分类模型结果做分值校准后得到。

02

以整体风险率量为优化目标，以整体客诉量为约束，以可用的决策产品为优化变量，进行0-1背包优化。

■ 决策效果预估

为了输出合理的管控手段，需要评估已经被识别出风险的用户的风险水位。这里一方面需要建模预测用户在管控后的客诉概率；另一方面也需要基于历史上管控后的业务反馈，建模预测管控后依然发生风险漏过可能造成的损失。由于线上对一个用户只能出一种决策产品（比如说人脸核验），因此实际上系统是不知道用户对其他产品的管控后效果的。所以这里训练数据存在两种偏差，分别是用户的选择偏差和策略管控体系导致的曝光偏差。

1. 标签噪音和选择偏差

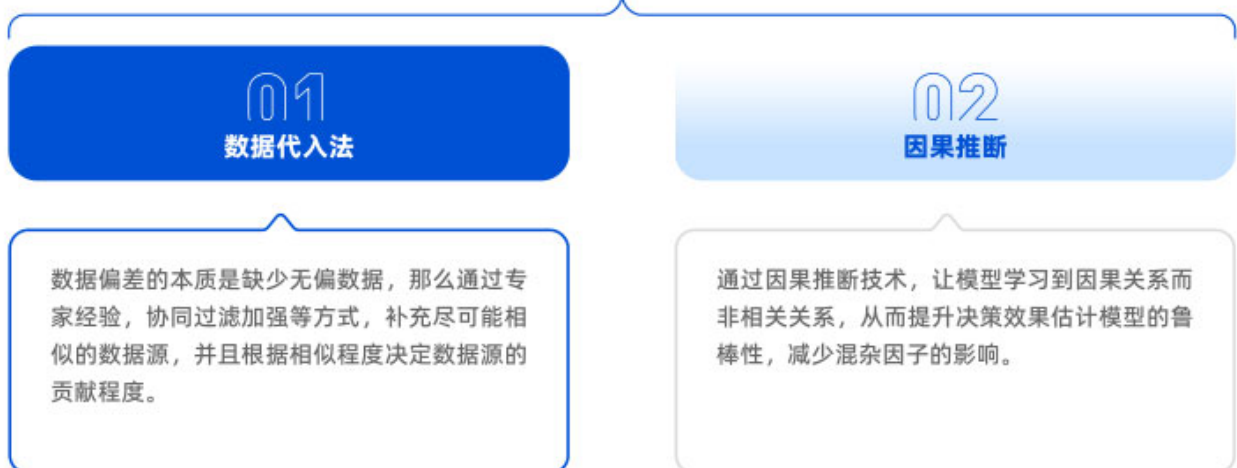
因为多数情况下风控系统并不会与用户进行可感知的交互，除非用户主动对“是否管控合理”这一情况进行反馈，此时几乎无法将用户未来的自然行为直接作为监督学习的目标。比如对于无风险/低风险的终端用户，只有当管控类型的严格程度或是管控的次数超过了用户的容忍度的时候商户才会进行投诉，并且不同商户对于打扰的容忍度也是不一样的；而对于高风险商户的资损确认，一般是通过用户给平台提供的举报信息。对于选择偏差，主流的处理方法包括：ESSM算法，这阿里巴巴算法团队提出的多任务训练方法减少选择偏差的影响。其在信息检索、推荐系统、在线广告投放系统的CTR、CVR预估中广泛使用。IPW（逆概率加权），主要是通过预估**倾向得分对样本重新加权**来减少选择偏差的影响，因果推断，这个方法的本质目标是消除观察性研究中实验组和对照组用户的不同质，进而得到因果效应的估算。尽管风险管控推荐与商品推荐中的“曝光-点击-转化”在因果链路上有一定的差异，但通过简单的网络结构修改，就能使其适应风控场景的去偏需求。

2. 曝光偏差

训练推荐模型最好的方法是通过将各类管控动作随机曝光，收集无曝光偏差的训练样本。但是在风险管控场景中，各类管控动作在输出的时候很难做到随机曝光，通常是由策略团队的基于业务经验输出管控动作。这就会导致不同的管控方式积累的管控样本，无论是样本的数量，还是样本的特征分布，都存在很大的差异，而这主要是由风险管控策略体系的曝光方式决定的。由于训练样本和打分样本的特征分布存在差异，且不存在无偏的验证集合，使用存在曝光偏差的数据训练得到的模型对全量人群的打分可信度不会很高。

这里对于风险管控决策模型中的曝光偏差，有两类主流解决方案，分别是

1.在模型训练环节引入模型假设修正曝光偏差



2. 在模型迭代环节引入探索机制获取无偏训练集

01 主动学习

主动学习基于已有的数据进行建模，并且主动发现最急需的数据，主动向外界发出询问，获得反馈，从而加速整个学习过程，生成更全面的模型。在目前的风控中，主动学习需要和风控策略的同学密切合作，以免在边界探索的过程中漏过底线风险。

02 探索与利用机制

不同于主动学习中基于模型认知程度主动发现急需的数据，探索与利用机制是在“探索新数据”和“利用旧数据”之间进行平衡，使系统即能利用旧数据进行推荐，又能高效地探索冷启动的管控类型是否是“优质”的，快速的收集冷启动数据。除了快速冷启动，“探索与利用”机制可以更好的挖掘风险潜在的管控方式，维持系统长期收益状态。在风控场景中，管控动作的层面上的探索是很难在业务上落地的，更合理的方式是设计出多套不同风格的管控方案，在管控方案的推荐上可以做充分的探索。

决策方案设计

将平衡风险管控和用户体验建模为背包优化问题进行求解，即在分发管控手段的过程中，在满足全部约束的同时，最大化核心指标。以风险最小化场景为例，

公式求解

$$\begin{aligned} \min & \left(\sum_{ij} risk_{ij} * x_{ij} \right) \\ s.t. & \sum_j x_{ij} = 1, x_{ij} \in (0, 1) \\ & \sum_{ij} exp_{ij} * x_{ij} \leq a \end{aligned}$$

在风险管控场景里， $risk_{ij}$ 为第*i*个用户在第*j*个决策产品后的风险发生概率， x_{ij} 表示是否对第*i*个用户使用第*j*个决策产品，约束1表示，每个用户只能输出一个决策产品，约束2表示总投诉率（体验指标）需要低于一定阈值。

用优化问题的方式来处理管控推荐，对比于传统的推荐问题，有以下几个优点：

01
方便权衡多目标

这个场景需要同时权衡风险管控和用户体验，而传统的推荐问题只有一个目标（例如点击率）。当然也可以把风险率和转好率捏合成一个目标函数，但这个函数的业务含义没有那么直观。优化的做法业务含义更清晰，即风险率有限的情况下最大化转好。

02
方便控制全局约束

用优化问题的方式，方便对管控手段的占比进行约束。而推荐的方式只考虑单个商户自身的情况，难以动态控制全局的管控分布。未来甚至可以加入更多业务需要的约束控制。

03
可随业务动态调整

随着风险变化或转好率、风险率模型衰退，在线优化会逐渐调整参数，以保证全局约束依然满足，而推荐模型缺乏阈值动态调整机制，效果可能逐渐衰退。



4.5 风险阻断

■ 概述

随着与黑灰产的对抗不断深入，黑灰产在进行非法活动的过程中，在一个平台上留下的信息越来越少，这对于风控来讲面临风险信息发现和挖掘的天花板，用户遇到风险的意图预测越来越难，为了进一步压缩风险水平，势必会付出更多的打扰成本，影响用户支付体验。另外，无法准确的预测用户的意图，就难以针对性的进行风险阻断，难以保护用户的资金安全。

风险阻断的核心技术是交互式主动风控，目的是在现有风控识别基础上，主动发起与用户的风险交互，获取用户面临的真实风险信息，通过风险模型进行风险识别后，针对性的进行风险决策处置。交互式主动风控能帮助系统更准确的去判断风险信息，更有针对性的去进行风险揭示和劝阻，提升用户资金保护效能，减少风险发生。

■ 交互式风控核心能力

交互式风控应该由三个功能模块构成：触达、交互和防控。首先需要触达到相关的用户；然后针对不同的用户选择相应的交互方式，交互中用户可以反馈相关的信息、风控系统也可以向用户传递安全信息；根据这些信息，风控系统最终需要决定采用哪些合适的防控方式。在实现这三个功能模块的同时，风控系统还应当以用户为中心，在保障用户安全的基础上，尽可能解决且及时解决用户诉求，优化用户的使用体验。因此，建成一个完整的交互式风控系统需要具备以下能力：



交互式风控核心技术

实现智能化的安全服务技术需要强化学习算法和知识驱动算法。强化学习又称再励学习、评价学习或增强学习，是机器学习的范式和方法论之一，用于描述和解决智能体在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。在交互式风控这个领域，需要解决用户与风控系统的交互过程中，通过怎样的交互流程（例如：每一轮交互中需要向用户收集什么信息、最终需要给用户怎样的安全提示等），来提升用户和风控系统的交互体验，从而引导用户积极配合风控系统的交互过程。其中，可以将用户的交互体验、配合程度或者交互信息收集率作为强化学习中的回报，用DDPG的算法，实现风控交互的强化学习。知识驱动算法指在缺少交互样本、或者没有交互样本的情况下，可以通过专家经验或者常识知识来帮助完成交互过程。



达成个性化的实时决策技术需要意图识别算法和情感识别算法。通过对用户当前意图和情感的识别，可以实时洞察用户在交互过程中的状态，从而更好的调整后续交互过程。用户在每个时刻可能呈现多种意图，因此，可以将用户意图识别抽象成大规模多标签分类问题。意图识别算法可以基于BERT模型设计，针对大规模多标签分类问题效果的提升，实践中在head设计上加入多重池化和归一化设计是有效的做法。情感识别则需要根据交互时用户的音频特征，分析用户情绪状态。



搭建精细化的管控运营技术需要谎言识别算法、可解释性算法和人机结合的能力。谎言识别算法通过一系列谎言指征（例如：言语闪烁、支支吾吾、喜用叠词等）来判断用户在交互过程中是否有隐瞒或者欺骗行为，从而采取不同的管控方式或者风控策略。可解释性算法针对用户的回答和模型的结果进行可解释识别，通过扩充停用词评价句子中所有字的重要度得分，大于停用词的即为重要词，从而方便了后续的安全服务效果评估。此外，通过设置转人工服务，可以对复杂用户进行交互，提升交互体验。

05

事后威胁反制 与应对

5.1 事后威胁应对概述

金融风险防控的主要挑战在于动态对抗。尽管事前感知和事中防控体系非常严密，但正所谓道高一尺，魔高一丈，诸如黑灰产团伙的诈骗分子仍会找到规避的手段或者系统漏洞，进而出现“漏网之鱼”，导致用户的权益受到损失。因此，风险防控体系中需要加入事后模块。

事后模块，具体称为事后威胁反制与应对，主要处理三类问题。

01

第一类问题

处理用户投诉

处理用户投诉，此时风险事件已经发生，利益受损的用户一般会在平台上进行投诉反馈问题，平台方需要对用户的投诉进行审理定性，基于事后信息，判断投诉相关的风险是否成立。这类问题的解决方案，称为**智能审理**。

02

第二类问题

识别漏网之鱼

用户不一定会百分之百的投诉，会有相当量的风险事件发生了，但没有被投诉审理流程暴露出来，此时，需要基于事后信息进行回溯反查。这类问题的解决方案，称为**类案检索**。

03

第三类问题

对风控体系进行事后回看

进行必要的复盘和调整。也就是说，智能审理和类案检索的基础上，我们需要对事前事中的防控体系做出什么样的调整，特征体系是不是要修订补充，模型参数是不是要更新，这种问题的解决方案，称为**威胁知识提炼**。

从模型角度来看，智能审理是给事中风险识别模型的打分进行修正，通过对投诉信息进行审理，发现是有效的风险事件后，打上标签，回头看事中模型是否识别正确。类案检索是对智能审理的补充，智能审理是基于可见的投诉事件进行判断，而冰山之下可能有更多的类似风险事件，可以基于关联性或相似性来召回更多的潜在案件，来补充标签。这两种标签都可以用来让事中风险模型进行更学习和再训练，调整其参数，提升其性能。

5.2 智能审理方案

智能审理架构

智能审理产品整体上可以分为三大功能模块。

01

用户投诉模块

该模块主要用于投诉信息的收集，主要关注如何提升收集的投诉质量和用户体验，保证后续审理定性环节有充分的信息依据，该功能模块类似于法院起诉案件的证据调查和收集环节。

02

审理定性模块

该模块主要用于对用户投诉信息进行审理定性，判断投诉是否成立，主要关注审理定性的准确性和时效性；审理定性可以同时应用多种技术方案，如抗辩审理和类案检索技术，当同时应用多种技术进行审理定性时，需要根据业务上要求的准确率来确定选择哪一种方案作为最终决策结果，准确率可以通过建模时调整风险阈值，以评估测试集效果的方式进行确定；该功能模块类似于法院起诉案件的庭审环节。

03

诉后服务模块

该模块主要根据审理定性结果，对投诉双方用户进行相应的诉后管控和服务，主要关注诉后服务用户的满意度和管控准确性，该功能模块类似于法院起诉案件的宣判环节。



智能审理核心技术

交互式投诉技术

当用户到平台系统（如举报中心）投诉时，用户投诉信息填写完后，系统会通过语义理解模型识别用户填写的信息是否满足举报信息的完整性。当识别举报信息完整时，用户即可以提交信息完成举报。当模型判断填写的信息不完整时，系统会针对缺少的举报要素信息，针对性地与用户进行交互，让用户补充缺少的信息。信息更新后，模型会进行再一次判断。若信息完整，用户即可提交完成举报，反之，系统会进行下一轮的交互信息补充。交互过程可以通过语音电话或者文本交互进行。交互式投诉主要涉及文本语义理解、语音识别和知识图谱等技术。

多模态学习

多模态机器学习（MultiModal Machine Learning, MMML），旨在通过机器学习的方法实现处理和理解多源模态信息的能力。审理场景天然存在多模态数据，如用户的交易行为结构化特征，投诉过程中产生的文本和图像等信息，有效融合利用这种多模态信息是提升审理定性效能的一大利器。



抗辩审理技术

抗辩理论（argumentation theory）是一种建模不完全、不一致、不确定信息的推理方式，属于逻辑学范畴。对于审理业务，在接到欺诈受害人的投诉之后，如果由法律专家来审理，其会依据刑法中的法条，欺诈罪有所谓的四要素：非法占有、实施诈术，陷入错误、资产转移等，而这些可以和具体的行为做到对应。抗辩审理技术旨在模拟法律专家的审理过程，其通过构建欺诈罪的论辩推理图谱，利用置信度传播的抗辩模型的论据汇总方式进行概率论辩推理，首先判断欺诈要素是否成立，进而推断欺诈是否成立。

5.3 类案检索方案

■ 类案检索架构

类案检索主要通过案件学习、案件检索和案件定性指导三部分组成。其目标是为了在历史案件中找到与当前案件的相似案件。

01 案件学习

首先需要将历史案件进行特征抽取，案件中各种要素信息结构繁杂，例如举证图片、投诉描述文本、交易行为等。要将这些信息通过特征抽取转化为结构化数值。例如针对举证图片，分别提取类似聊天记录的文字信息、类似网站、APP截图等图片的视觉信息；针对文本数据，除了投诉描述文本，还利用聊天记录包含的文字信息。

02 案件检索

案件检索需要同时兼顾底库内存、召回情况、搜索时间这三个方面。构建索引内存过大，对存储要求过高；召回案件的精度是精准召回还是模糊召回，以及案件搜索的时间成本是一个需要综合考虑的问题。类案检索为了实现实时案件指导，选择了相对折衷的方案，内存占用小、模糊召回、搜索时间快。通过级联架构的粗召回和精排的方式解决模糊召回的准确率问题。

03 隐案定性

检索出的多个类案可能风险类别不一致，面对不完全一致的类案，需要筛选并甄别出其中最恰当的类案并参照其中的案件结果对隐案的风险类别做出审理定性。可以利用投票、加权、最相似等方法确定隐案的风险类别。



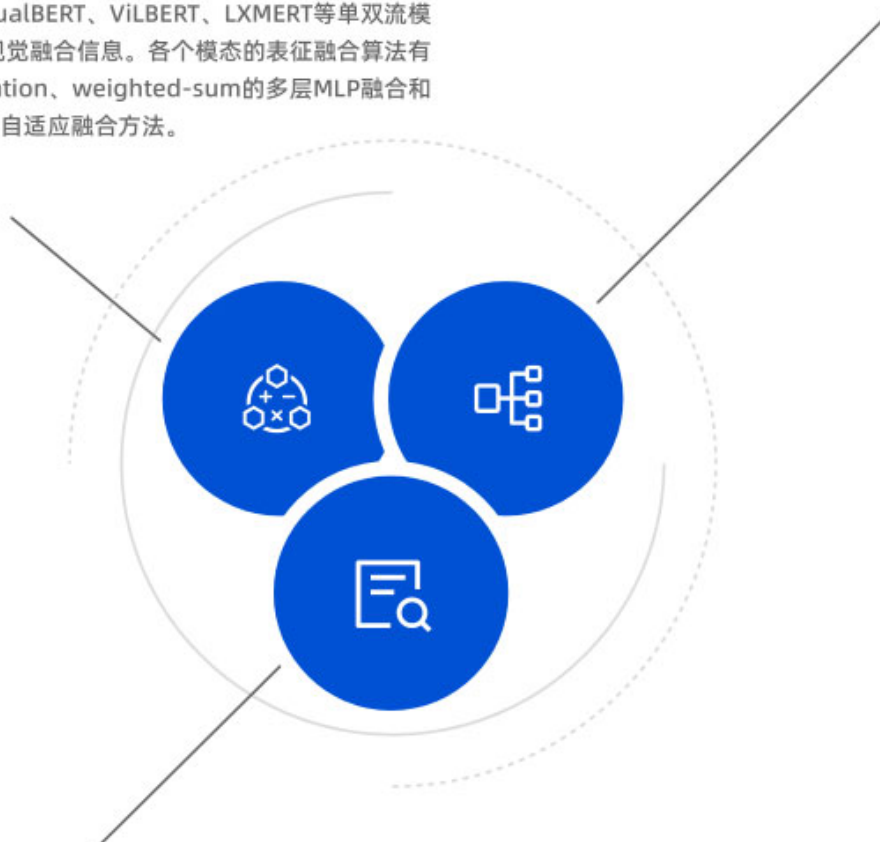
类案检索核心技术

多模态融合算法

针对图像、文本、结构化等不同模型的举证信息，分别对单模态数据进行表征学习，再通过多模态算法进行多模数据的融合。目前，文本数据的表征学习算法有Roberta、tinyBert、LongFormer等模型适应长短文本；结构化数据的表征学习算法有tabNet、xdeepfm、xgboost leaves等方法；图片数据的文字信息可以利用通用的ocr能力，提取后和举报文本结合学习，其视觉信息的表征算法目前有ResNet、layout-XL，也可以直接利用VisualBERT、ViLBERT、LXMERT等单双流模型直接学习文本和视觉融合信息。各个模态的表征融合算法有基于concat、attention、weighted-sum的多层MLP融合和基于transformer的自适应融合方法。

级联架构

首先通过多种召回算法召回相似案件，再根据具体策略进行排序，从而实现模糊召回到精准召回的转化。粗召回+精排的级联架构可以实现既要又要的要求，同时避免多任务的跷跷板缺陷。



检索库构建

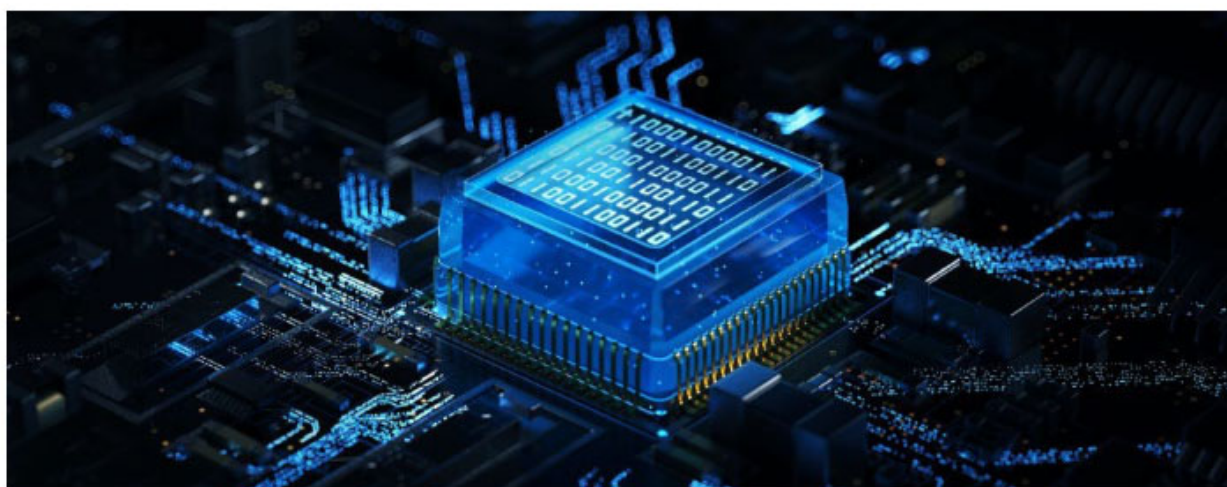
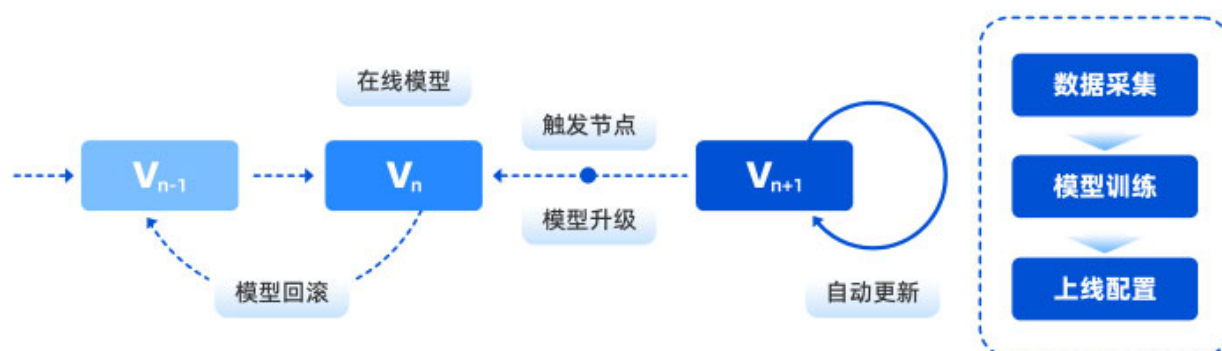
检索库的构建包括三个步骤：1) 计算得到案件向量库，2) 构建数据索引并将向量添加到索引中，3) 用索引检索。其主要差异在于构建索引和索引检索，这个决定了检索的精度召回和性能。索引的构建方法有倒排、树、哈希、图和矢量化等。

5.4 威胁知识提炼

威胁知识提炼整体框架

如前所述，金融风控体系的主要挑战在于动态对抗。在面对动态攻击时，传统风控体系的主要应对方式是人工调整策略。人工调整策略的可控性比较好，容易理解，调整也更加直接迅速。然而长远来看，风控策略会不断膨胀，一方面增加了风控系统的复杂性，应对措施是否合理也取决于运营人员的专家经验，另一方面策略运营人员也疲于应对，7×24小时的值班大大增加了人工运营成本。由此，能够动态的完成威胁知识提炼和防御提升是智能风控的重要模块。

为了更好的应对这种动态风险，需要搭建一套快速训练部署模型的链路，最理想的状态就是采用在线学习的方法，然而考虑到交易风控系统对与稳定性的要求，这套系统在传统的在线学习基础上新增了两个功能。一个是模型回滚，当系统遇到意外情况可以快速回切到原来的模型版本。一个是模型自更新，模型会在异步状态进行不断训练更新，是否上线取决触发节点，触发节点可以是人工触发，也可以是定制规则，具体更新的时间频率和触发规则可以根据具体情况设置。



威胁知识提炼核心技术

整个算法体系分为三大部分：基础模型、攻防模型、知识驱动，三类算法共同决策，但是所负责功能和决策权重会有区别。

基础模型

基础模型在这套系统中的主要作用是稳定，相对决策权重也最高，管控持续存在的常规风险。模型训练可以使用数据时间跨度久，标签丰富的数据集，算法上可以采用更加稳定鲁棒的模型。模型更新频率可以更加慎重低频。

攻防模型

攻防模型在这套系统中的主要作用是迅速，相对决策权重略低，管控难以防范的动态风险，模型训练可以使用数据时间跨度短，重点使用新风险手法标签的数据集，算法上可以采用迁移学习，增量学习，博弈对抗等对小样本较为有效的模型算法。模型更新评率也相对较高，可以根据风险变化来判断。



知识驱动

知识驱动模型这套系统中的主要作用是让模型充分使用专家经验来提高模型的鲁棒性。整体算法逻辑是将人工策略知识做为一种补充标签信息，输入到模型中，让模型去学习正常案件标签的同时，也学习策略结果。模型输出与专家知识的契合程度越大，说明模型的预测与专家经验契合程度越高。这样一来，可以有效提升模型风险识别能力。



实践与案例分析



6.1 事前防控案例-公网反诈信息处置

背景

支付宝提供了用户投诉举报的产品入口，当用户遭到电信网络诈骗出现资金损失时，可以通过举报入口来进行报案。但有很多用户并不清楚支付宝投诉举报的使用方法，会在公网相关平台下发声，描述案情和个人情绪。为了更好地服务用户和提早发现黑产异常信息，需要将这些反馈信息进行分析处理，并将此类信息引入到体系内进行处置，判断识别案件中的犯罪分子以及他们的作案手法，并帮助对用户进行反诈的宣传教育。

挑战

公网下的信息种类非常复杂，而且由于是文本形式的信息，很难直接输入算法进行处理。所以最重要的挑战就是对于文本信息的NLP处理，将其处理成结构化的信息。第二个挑战是如何将用户公网反馈信息和体系内信息相匹配，补充丰富体系内的信息，以更准确的处置案件，给用户更好的体验。

方案

基于公网信息搜集和分析系统，支付宝采用了五步工作流程来解决这个问题。

01 反馈信息分类

采集的反馈信息内容是极为复杂的，内容呈现出多种多样的表达方式。首先需要粗加工，将反馈信息进行初步的梳理抽取，整理归类到不同的风险类目中，例如电信网络杀猪盘诈骗就是一类典型手法。可以通过文本分类等NLP算法进行识别判断。

02 事件分组

同一组诈骗风险信息内容可能会被多个信息源用户所转发评论，因此需要对分类后的事件按内容再细分梳理。目标是一个组内所涉及到的都是讲的同一批反诈案件信息。支付宝平台采取聚类的算法思路，将讲述同一内容的相关文本归在同一组中，同时将该组的其它属性进行刻画，例如传播次数，转发条数等，以反映某组诈骗案件的重要程度。

03 要素提取

从大的诈骗信息池子中，通过分类和分组梳理出某个案件后，就需要进行更细致的分析。需要将其文本内容进行结构化处理，抽取出一些核心关键要素。例如风险类型，受害人，作案时间，作案地点，涉案平台，订单号等信息。这一步骤主要是基于NLP算法中的实体抽取。

04 事件研判

研判阶段是需要将第三步处理后的案件事件，和支付宝体系内信息进行打通匹配。还原其体系内对应的事件号，交易双方的支付宝账号，并对其账号进行图谱分析，找到关联风险账号。通过相应风险信息的累计，也可以以反哺帮助改进事中风险防控的策略规则。

05 案件处置

最后的处置动作涉及体系内的账户风控动作，例如关闭犯罪分子的交易权限等。也涉及到体系外的案件风控动作，例如协助公关进行用户安抚和信息宣导，提供信息协助警方追索赃款等操作。



6.2 事中防控案例-支付宝双十一反诈风控

背景

双十一购物狂欢节成为全民活动之后，也吸引了黑产的注意，基于种种原因，黑产会在双十一零点开启后的交易高峰期间，对支付宝支付系统发起攻击。传统的风控做法是，计算支付过程中的详细特征信息，对每笔交易风险进行各个维度上的准确计算，最终输出风险，如果有风险就拦截当笔交易。但是这种做法在双十一期间失效了，因为双十一峰值交易量会冲到平时两个数量级以上，平时系统可以做每笔交易的详细计算和管控，但这时如果仍然这么做，巨大的计算量将导致系统计算负荷超载。

本案例将介绍支付宝在事中如何设计实时识别和决策方案，从而同时满足在大促极端环境下和正常情况下的风控要求：如何在可负担的计算量下做高性能风险计算，如何在平衡打扰体验的情况下做风险档位切换管控。

方案

1. 识别分层

实时交易环节的计算资源是很珍贵的，实时防控的核心思路是将计算分流，有两个基本方法：将实时计算分流到提前一天的离线计算，或者将不同风险程度的交易分配不同的计算资源。

在这个思路指导下，依据使用资源的多少以及检测的精细程度，将实时风险检测划分成一个3层的漏斗体系：

01 第一层

只使用离线计算产生并存储的数据，结合简单的实时属性对最可信的请求快速放行，例如一个账户每天中午都会到一家饭馆吃饭，那这个账户-商家就可以作为一个可信放行对提前计算并保存好，实时交易时若查到这个交易对存在可行关系，就可以直接放行。

02 第二层

对于第1层无法判断的少量请求，结合离线数据与少量实时数据，通过轻量化算法进行再次判断，放过相对可信的请求。过程中决策阈值可以通过模型开发阶段的效能评估确定。

03 第三层

对于第2层无法判断的极少量请求，使用深度实时风险检测与对抗技术进行识别与响应。由于一般金融数据风控系统中，大部分请求都是无风险的，该技术方案可以从整体上保证大部分请求只消耗少量资源就可以低延时快速放行，同时对于风险请求则继续进行精准检测，兼具了绿色环保、降低成本、严格风险控制和良好用户体验等优势。

2. 动态管控

有别于经典的基于专家经验的风控策略，以及单一模式的核身推荐，动态管控方案通过半监督算法和进化算法实现了用户个性化的风险控制策略，不同用户的核身认证方式因场景、时间和地点的不同而不同，同时大大提升了风险控制精确性、实现风控运营自动化能力。

01 用户分群

用户分群是通过决策树算法+德尔斐法相结合而得，综合考虑了分群的稳定性、业务含义和风险概率，既从大数据角度出发科学分群，又包含了特定的业务含义。

02 多目标优化

风险决策策略推荐需要解决的问题是求满足多业务目标(打扰率和覆盖率等风险指标，失败率和限权率等体验指标)的最优解。这是一个典型多目标优化问题。而现实世界的多目标优化问题存在两个困难：相互制衡或冲突的目标和复杂的解空间。因此多目标问题不存在单一最优解，而是存在一组帕累托最优前沿(Pareto-optimal)，在缺乏主观偏好函数下无法进行解之间的权衡，使得解空间可能非常复杂和庞大，所以高效率而精确的求解极为困难。

在不同的阶段，对体验的要求是不一样的，双十一时允许的打扰率要远远低于平常时段。在双十一体验保障场景，设计了比平时更加严格的体验指标约束，在更少的打扰下最小化风险漏过。针对这种黑盒优化问题，采取了群体进化算法进行优化求解，其优点是无需先验知道优化目标函数的性质。

通过结合专家经验预设的方式，设定不同情况下的管控档位要求，算法输出相应的决策管控手段。这套动态决策能力实现了风险防控策略的自助运营，极大减少人工干预。历年双十一大促，系统根据交易流量和风险变化动态自动调整模型和策略的管控力度，实现了无人调配策略的可能，并经受住了大促时期黑产的攻击。



6.3 事后防控案例-支付宝投诉体验提升案例

背景

被诈骗用户在完成支付后，第一时间到支付宝完成投诉是非常重要的事情。对用户来说可以第一时间止损，对支付宝来说可以第一时间发现和定性黑产，阻止更多用户被骗。在这个过程中，用户投诉过程会受到很多原因中止掉，比如投诉产品交互链路复杂、信息录入体验差等。为了进一步提升用户安全体验，支付宝把对整个投诉产品进行系统梳理和升级，补全反诈全周期的关键一环能力。



挑战

用户在投诉流程中，最大的时间消耗是人工客服的处理流程上。因为投诉审核流程中，需要用户提供包括文字、图片等复杂形态的证据，来支持其诉求。同样也需要人工客服来阅读理解这些证据，甚至需要和用户交互以拿到进一步的证据，最后做出案件定性。所以需要有一个智能系统来自动化收集用户信息，理解用户数据并进行定性判断。此外，非正常用户的虚假投诉骗赔账户安全险行为也会影响投诉系统，骗赔识别也是投诉定性流程中的一个重要挑战。

方案

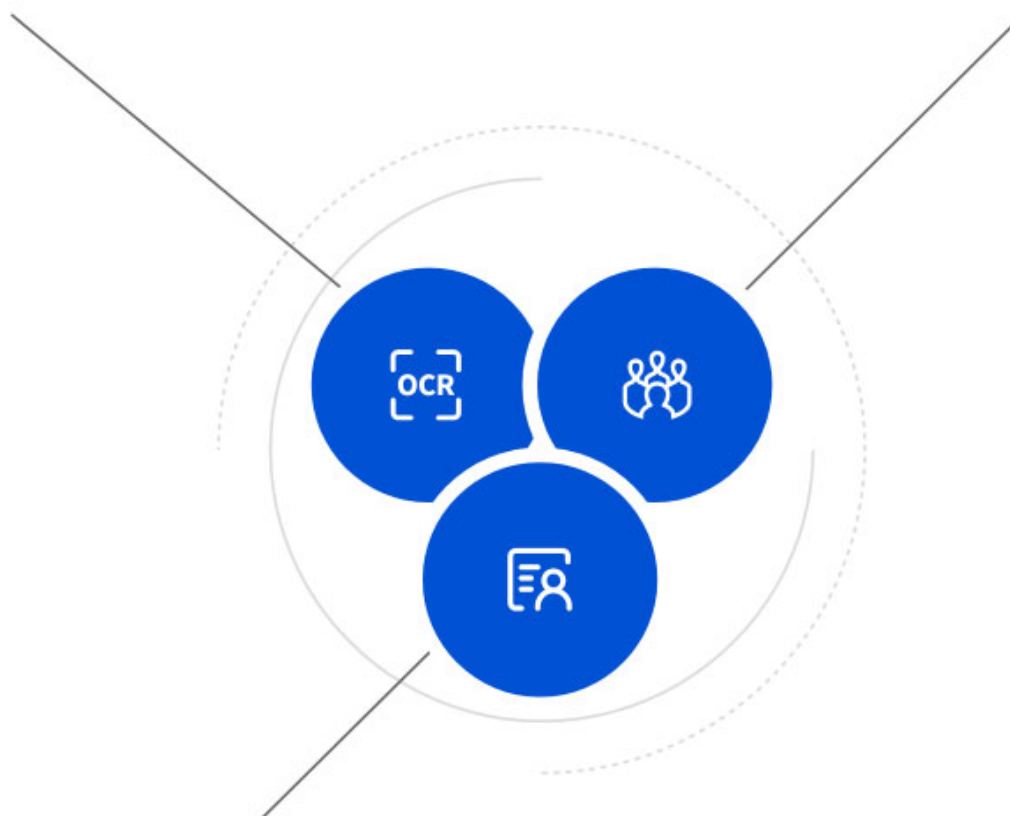
在投诉体验提升上，支付宝使用了三种底层算法能力来共同完成这项挑战。

智能凭证审核算法

对用户提供的材料进行OCR信息提取，通过对理赔场景图片去噪，旋转变换，透视变换等数据增强，优化OCR识别能力，达到身份证100%准确率。从报案立案材料的OCR解析结果，利用NER识别模型提取报案人、报案时间、公安机关、报案/立案编号、受案登记表文号，作为投诉定性的关键信息输入源。

骗赔团伙识别

在账户安全险骗赔识别上，将历史的多个团伙骗赔识别模型，通过t+1拓展黑节点，能够对账户进行实时的骗赔团伙风险识别。同时为了进一步挖掘骗赔案件，从设备、资金、手法等多个维度进行场景无监督/半监督的挖掘。此外，骗赔案例中有半数以上的案件属于单点骗赔行为，基于以往的策略及人工核赔标签，进行有监督的学习，从而有效提升单点本人操作骗赔识别的覆盖。



基于用户风险行为序列的案情还原

人工审理投诉案件时，常常需要根据用户的历史行为来进行案情还原，同时相较于传统人工衍生特征，直接对用户的原始行为数据进行挖掘可以补充更多未挖掘到的风险信息。因此引入用户的操作行为序列，来进一步提升场景识别效果。行为序列构建时，会考虑案件回溯到交易时刻，通过账户支出方与收款方的操作情况来判断。也会考虑用户在支出交易后到报案前这段时间内的行为。构建好的行为序列再输入到深度学习算法中进行学习训练，可以更加有效地还原用户投诉意图、做出更优的投诉诉求回应。