软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

数据要素流通全流程隐私关键技术：现状、挑战与展望*

刘立伟[1], 傅超豪[1], 孙泽堃[1], 周耘[1], 阮娜[1], 蒋昌俊[2]

[1](上海交通大学 计算机学院,上海 200240)

[2](同济大学 计算机科学与技术系,上海 200092)

通讯作者: 阮娜, E-mail: naruan@sjtu.edu.cn

摘 要: 近年来以大语言模型为代表的一系列数据驱动型 AIGC 应用深刻地改变了人们的生活范式，引起国家对数据流通、数据隐私等问题的高度重视.健全数据市场规范，完善数据要素流通机制成为数字经济时代下又一重大研究课题.但是现有数据隐私研究普遍聚焦于数据流通中的单一环节，并未展现数据流通的全貌，技术研究相对孤立，存在不兼容性等问题.因此数据服务提供商在实际生产活动中往往需要投入额外人力成本以进行全方位的数据隐私保护.本文聚焦数据流通问题，依据数据生命周期将流通全过程划分为三个阶段，对各阶段的隐私关键技术建立系统的分类体系，并对各领域的最新进展、未来挑战等问题进行深入剖析.本工作以数据流通为载体，隐私技术为目标，涵盖数据流通全过程，有助于研究者快速建立对数据流通全流程隐私技术的系统认识，为后续研究建立完备的全流程数据流通隐私保护范式奠定基础.

关键词: 数据流通;数字水印;联邦学习;区块链;忘却学习

中图法分类号: TP181

中文引用格式: 刘立伟,傅超豪,孙泽堃等. 数据要素流通全流程隐私关键技术：现状、挑战与展望.软件学报,2025.

英文引用格式: Liu LW, Fu CH, Sun ZK, et al. Privacy Technologies in the Whole Stages of Data Circulation: State-of-the-art, Challenges, and Prospects. Ruan Jian Xue Bao/Journal of Software, 2025 (in Chinese).

Privacy Technologies in the Whole Stages of Data Circulation: State-of-the-art, Challenges, and Prospects

LIU Li-Wei[1], FU Chao-Hao[1], SUN Ze-Kun[1] ,ZHOU Yun[1],RUAN Na[1] and JIANG Chang-Jun[2]

[1](Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

[2](Department of Computer Science and Technology, Tongji University, Shanghai 200092, China)

Abstract: In recent years, a series of data-driven AIGC applications, represented by large language models, have profoundly changed people's lifestyles, drawing significant attention from nations to issues such as data circulation and data privacy. Establishing robust data market regulations and improving data element circulation mechanisms have become hot research topics. However, existing researchs on data privacy generally focus on individual aspects of data circulation and don't provide a comprehensive view of the entire process. Technical research tends to be isolated, leading to issues such as incompatibility. As a result, data service providers often need to invest additional manpower in practical production activities to implement comprehensive data privacy protection. Our paper focuses on data circulation, dividing the entire circulation process into three stages, and establishes a systematic classification framework for privacy technologies at each stage. It also provides an in-depth analysis of the latest developments and future challenges in various fields. To our knowledge, our work is the first review paper that that uses data circulation as a carrier and privacy technology as the subject, comprehensively covering the entire process of data circulation. It helps researchers quickly build a systematic understanding of privacy technologies throughout the data circulation process and lays a foundation for establishing a comprehensive full-process data circulation protection paradigm for future research.

* 基金项目: 国家自然科学基金(62472276); 国家重点研发计划(2023YFB2704700); 上海市科学技术委员会(23511101000, 24BC3200400)和国家电网公司科技项目(5700-202321603A-3-2-ZN)

收稿时间:2025-01-16; 修改时间: 2025-03-25; 采用时间: 2025-05-29

Key words: Data Circulation; Digital Watermarking; Federated Learning; Blockchain; Machine Unlearning

数据要素，作为数字经济时代的第五大生产要素，近年来已然成为推动社会进步和经济增长的关键资产.培育数据要素市场，加快数据要素流通交易，赋能数字经济发展成为各国共识.过去几年里以大语言模型为代表的数据驱动型 AIGC 技术的兴起，更是引起了全世界研究者的广泛讨论.AIGC 技术的应用横跨文本、图像到音频等多个领域，深刻改变了现代社会的内容生产模式，“数据即价值”的观念深入人心.但是层出不穷的“数据犯罪”，让用户数据隐私安全面临严峻挑战.

事实上，为加强数据犯罪打击力度，保障用户数据隐私，各国政府早已从立法层面采取相关措施.例如欧盟理事会于 2016 年 4 月通过，并在 2018 年 5 月开始强制实施的《通用数据保护条例》(General Data Protection Regulation, GDPR)，首次从法律层面对数据流通全流程中所应遵守的行业规范进行了严格规定.在该条例生效后，欧盟向谷歌和脸书两大知名互联网公司分别发起 39 亿欧元和 37 亿欧元的罚款诉讼, 在数据安全领域产生深远影响.在这之后，美国的《统一个人数据保护法》(UPDPA)、我国的《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》等多部数据保护法规相继出台，一定程度上遏制了数据犯罪的发生.

在技术层面上，隐私计算（Privacy Computing）技术在过去的几十年里飞速发展，不仅全同态加密、安全多方计算等密码学技术在可用性等方面取得可观进展，近年来逐渐兴起的联邦学习、数字水印、忘却学习等多个领域研究也取得显著成果.隐私计算这一概念最早由 Li[1]的团队提出，将其定义为“面向隐私信息全生命周期保护的计算理论和方法”.隐私计算研究发展至今，已经形成了横跨密码学、大数据、机器学习、计算机体系结构、联邦学习等多个领域的综合性研究体系，为数据安全流通提供可靠保障.此外，隐私计算中“数据可用不可见”，“不共享数据，而是共享数据价值”的理念对数据安全领域研究产生深远影响.

虽然隐私计算的定义中强调了其范围涵盖隐私信息全生命周期，但是现有的隐私计算相关综述在对相关技术进行总结时，或是因为提出时间过早而缺少对新兴技术如数据最小化原则的介绍[1]，缺少实时性;又或是围绕隐私计算中子领域进行归纳整理[2-4], 并未涵盖数据全生命周期，技术介绍相对孤立.从隐私计算发展角度而言，既需要研究者在隐私计算子领域深入挖掘，培养专业性人才，也需要研究者从数据流通的角度，总揽全局，为建立完备的全流程数据流通隐私保护范式而努力.后者在目前的研究中显得尤为稀缺.

数据流通一般是指数据在不同主体之间进行共享、传递和交换的过程.其实现了不同系统之间的数据互通，促进信息共享和协作，是跨系统、跨组织的数据应用和分析的重要基础.现有数据隐私研究的研究对象普遍关注的是数据流通中的信息共享，对于数据流通前的数据收集或是数据流通后的数据治理等问题关注较少.但是根据工业实践经验，上述所提到三个阶段在实践层面普遍存在且互不矛盾，在隐私性能层面互为补充，如联邦学习中同态加密技术常用于梯度信息传递过程中.因此，建立全流程的数据流通隐私保护范式有其独特的研究价值.

本文依据数据全生命周期以及不同阶段间的技术兼容性，将数据流通全流程划分为前、中、后三个阶段，并针对每个阶段所涵盖的隐私关键技术最新研究进展进行系统的分类整理，对不同领域的研究热点与未来挑战进行深入剖析.数据流通全流程隐私关键技术框架如图 1 所示.

数据流通前这一阶段主要包含数据生命周期中的数据收集和预处理环节，是数据流通不可缺少的环节.该领域在隐私计算研究初期相对其他阶段易被忽视，但是伴随数据最小化原则，数据溯源、数据所有权、模型所有权等诸多问题的发现，越来越多的研究者将目光转向该阶段.本文将从数据最小化原则、数字水印技术两方面对该阶段进行系统分类与整理.数据最小化原则要求数据采集者仅收集、处理为实现特定目的所必需的数据量，从数据采集阶段避免了数据滥用的可能性，但是如何实现其从法律定义到技术定义的转变需要更为细致的研究.数字水印技术则是在通过在数据中嵌入不可见的标识信息，防止未经授权的复制和分发，一方面解决了数据溯源问题，另一方面验证了数据的完整性和真实性.数据流通阶段涵盖数据分析、应用环节，本文依据技术类型进行分类，从基于密码学技术、基于合作学习技术、区块链技术三个方向进行归类整理.具体来说，同态加密、安全多方计算、零知识证明是密码学技术在数据隐私领域中的代表性成果；而近年来兴起的联邦

学习研究则是合作学习技术的代表，契合数据流通中普遍存在多个主体的分布式情形；区块链是数字经济的重要基础设施，其数据不可篡改性在金融交易、合同管理和知识产权保护等领域具有重要价值.数据流通后主要包括数据治理环节，研究数据使用后的安全性以及被遗忘权等问题，目前的代表研究是忘却学习，旨在使模型能够有选择地“忘记”某些特定的数据或知识.在某些情况下，用户可能要求删除其个人数据，忘却学习技术可以确保模型在删除这些数据后，不再保留或利用这些数据的任何信息.

据我们所知，本工作是首个以数据流通为载体，隐私技术为对象的综述类论文，将数据最小化原则、联邦学习、区块链等新兴技术纳入研究范畴，全面涵盖数据流通全过程，有利于研究者快速建立对数据流通全流程隐私技术的系统认识，为后续研究建立完备的全流程数据流通隐私保护范式奠定基础.

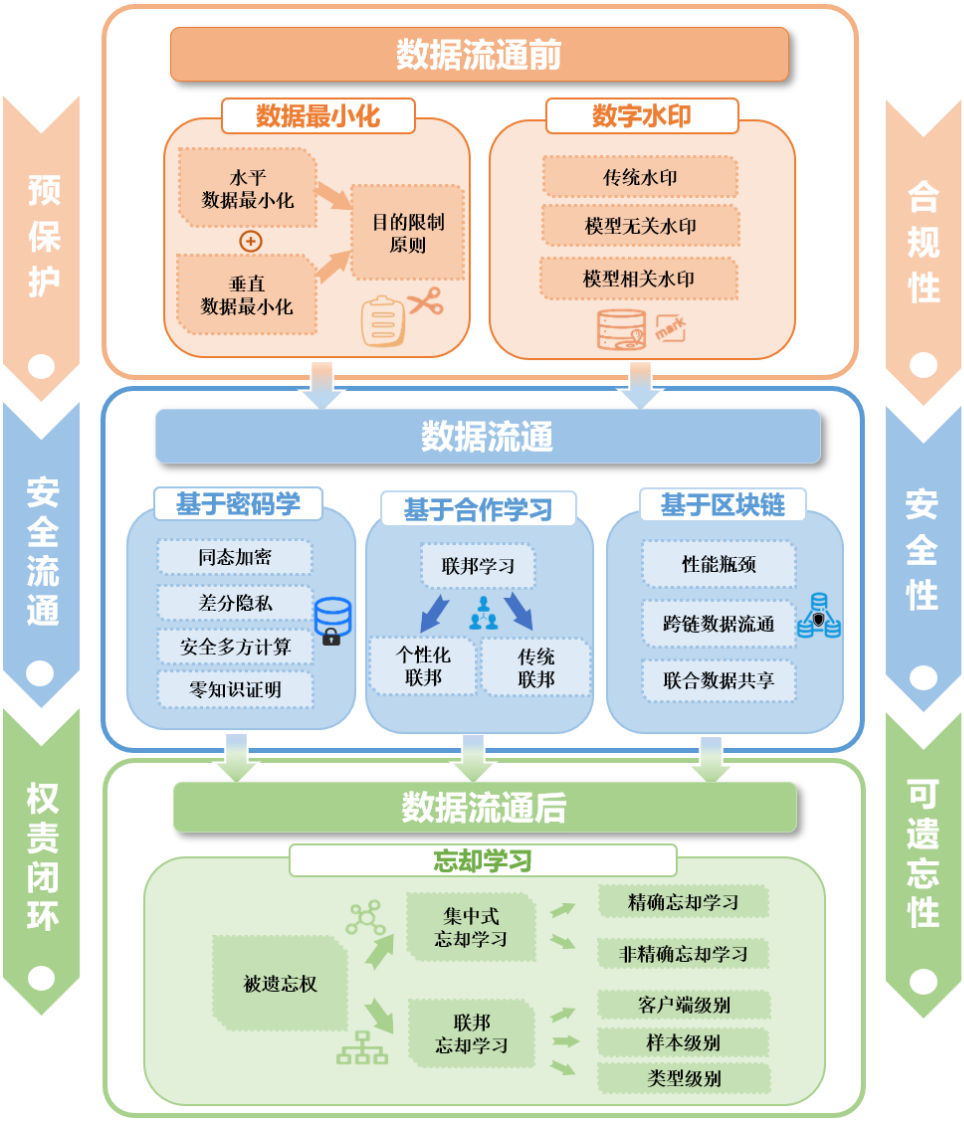


图 1　数据流通全流程隐私关键技术一览

本文第 1 节介绍数据流通三个阶段的隐私风险和相关前沿技术.第 2 节介绍数据流通前的隐私技术分类体系，包括数据最小化原则和数字水印技术.第 3 节依据技术基础从三个方向介绍数据流通阶段的隐私研究进展

与挑战，范围涵盖密码学、合作学习与区块链.第 4 节介绍数据流通后以忘却学习为代表的隐私保护技术.第 5 节介绍人工智能时代下数据流通的全新隐私挑战.最后总结全文.

1 数据流通隐私风险

在大数据时代，数据要素已成为驱动科学技术创新、经济发展增速和法制社会治理的核心资源，其流通共享可释放人工智能、智慧城市等领域的潜在价值.但是，伴随着数据要素市场化进程的加速，数据要素在多主体、跨领域的交互中面临严峻隐私风险.例如人脸、医疗记录等用户敏感信息因技术管理缺陷如数据过度采集或第三方滥用而泄露，引发身份盗用甚至公共信任危机.因此，现代社会亟需在数据价值挖掘与数据流通隐私安全之间构建动态平衡机制.结合数据流通的三个阶段，本节将对当前数据流通全流程面临的主要隐私风险进行分阶段总结.

数据流通前的隐私风险主要潜藏在数据收集、预处理与存储环节.其一是数据过度采集造成的敏感信息泄露问题.数据过度采集现象在当前社会广泛存在,企业为挖掘潜在价值往往倾向于超额收集数据,如健康类 APP 收集用户地理位置、通讯录等隐私信息，而手机号等个人隐私数据的泄露（如第三方非法交易）可能诱发精准诈骗等违法犯罪行为，给个人造成巨大经济损失.因此数据收集环节的合规性在近十年来得到全球政府的高度重视，其从源头上杜绝无关敏感信息的泄露风险.数据最小化原则是对全球数据保护法规中数据采集规范的抽象概括，要求企业采集数据前根据使用目的确定所必需的数据量，避免数据过度采集造成隐私泄露.但是如何建立法律条文到数学语言的映射，跨越形式与实践的鸿沟是目前数据最小化相关研究的重要一环.其二是数据脱敏不足造成的身份信息泄露.数据脱敏是预处理环节的标准步骤，但是传统的脱敏方法如删除直接标识符、k-匿名等方法已经被证明存在重识别攻击的风险.差分隐私能够实现更强的隐私性能，但是需要在隐私和效用之间实现权衡.其三是数据存储不当引发的数据泄露，例如：金融机构客户数据以明文形式存储于不安全服务器，引发大规模数据泄露、财产损失及金融诈骗.通过加密存储或弱加密存储，可以有效避免攻击者获取数据中的敏感信息.

数据流通中的隐私风险指数据要素在多个主体间的流转过程中由于通信协议设计缺陷、恶意中间主体、主体行为不可控等因素引发的隐私信息泄露.目前数据流通中的隐私风险主要包括：一是多方协作协议的缺陷.在涉及多方协作如协同计算、合作学习的过程中，不同主体间的原始数据如果进行直接传播，极易引发法律纠纷；因此，研究者尝试通过加密、传递中间信息等方式避免原始数据的直接传播.但是，在后续的研究中发现诸如合谋攻击、差分攻击、成员推断攻击等攻击方式，协议缺陷的存在引发了数据的直接或间接泄露.为解决多方协作过程中的隐私泄露问题，多方安全计算、联邦学习等研究领域应运而生.二是数据污染.数据流通过程中，恶意中间节点可能通过数据篡改、注入恶意数据等方式实现攻击目的，如在模型的训练集中添加少量精心设计的数据，为模型植入后门.针对数据篡改等问题，近年来兴起的区块链研究提供了一种可行的解决思路.作为一种基于密码学的分布式账本技术，区块链在不依赖可信第三方的前提下提供可信数据.共识协议的存在有效防止数据被恶意篡改.同时，区块链被认为是未来新一代的基础设施，在数据存证、可信溯源领域极具研究价值.

数据流通后的隐私风险主要依赖于数据生命末期对残留数据清理的及时性和完整性.其主要风险包括：一是第三方数据滥用.若数据流通后期缺乏数据审计机制，未在达成使用目的后及时清除，数据易被第三方滥用，如广告商利用用户行为数据定向投放骚扰广告；二是残留数据泄露问题.常规存储数据可以直接删除，主要防范的是硬件层面的数据恢复技术.但是在人工智能时代，大量的用户数据被用于模型的训练中，用户信息隐式地存储在模型的参数当中.当用户在退出 AI 服务的同时有权要求模型遗忘自己的隐私数据，即被遗忘权.如何让模型遗忘用户信息是忘却学习的研究领域.三是数据存证与可信溯源.数据的可信溯源对于加强数据监管具有重要意义，如打击虚拟货币非法洗钱.区块链技术的存在为数据存证和可信溯源提供了可靠的技术支撑.

在上文中我们对数据流通三阶段的主要隐私风险进行分类阐述并初步给出目前主流的解决策略，在后三节中将对各阶段的隐私关键技术进行系统总结.需要强调的是，部分技术事实上在多个阶段中均存在应用，如

区块链在数据流通中的应用防范数据篡改，同时在数据流通后可以实现数据的可信溯源.为了保证文章的逻辑连贯性，该类技术整合在单个阶段进行阐述.

2 数据流通前

数据收集与预处理是数据全生命周期的开始，也是数据正式进入流通前的关键环节.伴随国家对个人信息保护相关法律法规的完善，如何在合法前提下安全高效地采集并管理数据成为数据科学领域研究者关注的热点问题.本节内容将聚焦数据流通前的隐私关键技术，首先介绍数据最小化原则（Data Minimization, DM），随后介绍数字水印技术(Digital Watermarking).

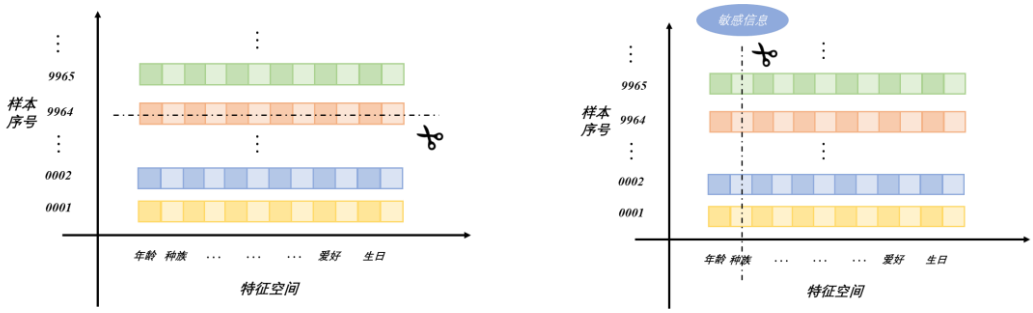
2.1 数据最小化原则

2.1.1 法律溯源

欧盟《通用数据保护条例》（GDPR）第 5 条中首次提出数据最小化原则，要求企业在采集个人数据的过程中应遵循适当，相关的原则，并要求采集数据受限于处理目的所需.简而言之，数据最小化原则旨在确保数据采集者仅收集、处理和存储为实现特定目的所必需的数据量，以最大限度地减少数据滥用、未经授权访问或数据泄露的可能性，契合隐私保护的基本原则，近些年来已得到包括《加州隐私权法案》（CPRA）、《中华人民共和国个人信息保护法》等多部数据保护法规的认可.

2.1.2 研究分类

实现数据最小化的过程本质上是对数据进行筛选、裁剪的过程，根据数据裁剪维度的不同，通常可将数据最小化研究领域细分为水平数据最小化（hDM）和垂直数据最小化（vDM）.



(a) 水平数据最小化示意图 (b) 垂直数据最小化示意图

图 2 数据最小化分类系统与工作机理

水平数据最小化(hDM) 水平数据最小化主要研究的是在数据层面通过减少数据点的数量以达成最小化的目的，在部分研究中也常将其称为深度型数据最小化.将数据集在二维平面上展开，以数据标识符作为纵坐标，数据字段或特征作为横坐标，数据点可以解释为数据集中某一条完整的数据或是用户数据集中用户的单条数据，减少数据点的过程即水平擦除的过程.这种方式直观且高效，在研究初期得到众多研究者的青睐.其核心思路是通过使用数据修剪等方法来减少唯一数据点的数量. Mansheej[5]等人为在训练早期阶段寻找数据中对于模型泛化性能影响较大的示例，提出两种分数——梯度归一化（GraNd）和误差 L2-范数（EL2N）分数，在不牺牲测试准确性的前提下在各种数据集上成功剪枝大量训练数据.Shanmugam[6]等人提出了一个用于限制数据过度收集的框架 FIDO，FIDO 利用分段幂律技术迭代更新性能曲线的估计值来提供数据收集停止准则.但是，在现实场景中，hDM 并不是解决数据最小化背后许多隐私问题的合适方案，因为它并不为那些数据已经被收集的个体客户提供隐私保护.

垂直数据最小化(vDM) 垂直数据最小化主要通过在特征层面对数据进行删减以减少数据集的大小，在部分论文中也称其为广度型数据最小化.类似前文将数据集在二维平面上展开，垂直数据最小化的过程即是减少数据字段或数据特征的过程.举例而言，一条用户数据中可能包含性别、年龄、种族、信仰等敏感信息，部分信息对于任务目标无效，根据数据最小化原则应进行删减.相比于水平数据最小化，垂直数据最小化更加贴合

现实场景，在近几年得到越来越多的关注.Ganesh[7]等人首次对垂直数据最小化中涉及的个性化特点（对某个人无关紧要的信息可能对另一个人至关重要）进行理论与实验分析，对机器学习任务的数据最小化进行了形式化； Staab[8]等人则提出了一种基于数据泛化的新型垂直数据最小化工作流程，通过引入对抗场景量化隐私风险.此外，他们同时提出一系列基准 vDM 算法，以及一种高效的 vDM 算法 Privacy-aware Tree (PAT).

2.1.3 热点问题

可行的数据最小化定义 虽然数据最小化原则早在 2018 年欧盟《通用数据保护条例》出台时便被提出，但是此举仅从法律层面对数据采集规范进行约束，并未从技术层面提供明确可行的应用与评判标准，在生产实践中缺少可操作性.数据最小化原则要求在数据收集中只获取为实现明确目的所需的最低限度的数据，但是根据过往实验规律，数据处理目的的实现往往与数据的质量和数量紧密相关，如何判断处理目的是否得到良好实现，所使用的数据是否满足数据最小化原则和如何选取数据最小化策略等问题都存在着进一步讨论的空间.举例而言，在机器学习领域，训练模型的质量与参与模型训练的数据量存在紧密联系，为判断是否实现模型的训练目的，研究者通常会使用一种或多种模型评价指标如准确率进行评估，要求性能指标在不低于一定阈值的前提下进行数据的筛选.在上述过程中，与性能相关的评估指标成为评价数据处理目的是否实现的标准，研究者选取合适的数据最小化策略对数据进行筛选，从而保证模型训练的过程满足数据最小化原则.类似地，在个性化推荐领域，Biega[9]等人首次注意到个性化领域缺乏对数据最小化原则的统一解释，根据选取性能评估指标的不同（个人性能指标和全局性能指标），创造性地提出两种基于性能评估的数据最小化定义，并通过充分实验对两种定义的可行性、不同最小化策略对推荐性能的影响等问题进行了深入讨论.实验结果表明数据最小化带来的全局性能下降可能并不显著，但会不同程度地影响不同用户，即在个人性能指标上出现较大差异.此外，数据最小化可能会损害边缘化群体，特别是如果这些群体在数据中占少数的话——多数人群的成员只需少数特征就能得到良好的服务，而少数人群需要提供更多特征才能获得同等质量的服务.

数据最小化审计 伴随着国家加大数据隐私监管力度，数据最小化审计得到广大研究者的重视.Clavell[10]等人为此在其开发的个性化健康推荐应用 REM!X 量身设计审计算法，在避免收集性别、年龄、种族、宗教或其他受保护属性的前提下，通过建立焦点小组、使用间接证据等方式尝试分析是否满足数据最小化等数据伦理原则.此外，针对黑匣子预测模型，Rastegarpanah[11] 等人为判断模型是否满足数据最小化原则，提出了基于模型不稳定性的 DM 合规性黑盒审核算法，其核心想法是通过为每个预测模型的输入特征分配一些固定值（插补），检查每个特征是否必要，并衡量模型结果的变化程度.

隐私度量 数据最小化原则的实施一定程度地降低潜在的滥用、未经授权访问或数据泄露的风险，但是如何在使用数据最小化策略后在隐私层面进行科学有效地评估同样是值得探讨的问题，这有助于研究者从除性能表现以外的层面判断一个数据最小化策略的好坏.为了更加全面地评估最小化策略的表现，Biega[9]等人创新性地提出可辨识性（identifiablity）作为个性化推荐领域的隐私性度量指标，表征用户数据对抗推断攻击的能力.该度量指标表示用户物品集中唯一确定该用户所需的最小数目，原理与海明距离有异曲同工之妙.相似地，Ganesh[7]等人在对机器学习领域的数据最小化进行形式化后，在实验部分通过对最小化数据集进行重新识别和重构攻击进行综合评估，分析了当前数据最小化法规要求在满足隐私期望方面存在的不足之处.值得一提的是，该工作首次对最小化的个性化特点进行分析，从个性化角度审视最小化与用户隐私的潜在联系.

2.1.4 未来挑战

泛用的数据最小化定义 在现有的数据最小化研究中，主要给出了机器学习领域、个性化推荐领域中可行的数据最小化定义[9,12]，但是这些研究中给出的最小化定义与该领域中数据的评估指标、数据要素的组成形式等因素密切相关，难以直接迁移到其他诸多领域.此外，现有研究中给出的数据最小化定义普遍依赖于性能评估指标.但是在实际生产实践中，数据使用者在大规模数据集上往往无法直接获取性能评估指标，难以根据该指标选取合适的数据最小化策略以满足不同用户的需求.因此，寻求基于性能评估指标之外的可行数据最小化定义成为后续研究的一大难点.

公平性问题 技术发展过程中不可避免会遭遇各种技术伦理问题,而已有众多研究表明数据最小化过程中

可能给公平性带来极大挑战.Rastegarpanah[12] 等人曾提出了一个基于分类准确性的公式,研究了数据最小化与满足其他公平性属性之间的联系,发现二者之间存在着明显的权衡关系.此外, Biega[9]等人在观察不同最小化策略的实验结果时发现, 数据最小化可能会损害边缘化群体,少数人群需要提供更多特征才能获得同等质量的服务.虽然这些研究一定程度上揭示了公平性和数据最小化原则间可能存在的权衡关系,但是如何平衡好二者仍然是颇具挑战的问题,亟待后来者的探索.

性能-隐私权衡 数据最小化原则要求只获取为实现明确目的所需的最低限度的数据,但是减少数据量并不等同于隐私保障.通过建立合适的隐私度量指标,有助于完善数据最小化策略评估体系.虽然在多个特定领域中,已有较为有效的数据最小化算法,但是最优解是否存在, 这些算法是否得到最优解仍然缺少理论层面的证明.

2.2 数字水印

伴随着以大语言模型为代表的人工智能生成内容（Artificial Intelligence Generated Content，AIGC）技术逐渐深入人们日常生活，数据要素的重要性不断凸显.在这个过程中，高价值的数据内容也诱发了一系列针对私人用户数据的违法犯罪行为.为了进一步加强对国内数据市场的监管，以数字水印技术为代表的数据溯源技术得到国家机关的高度关注.数字水印技术旨在通过在数字内容中嵌入不可见的标识信息，有效防止未经授权的复制和分发.这种技术不仅能追踪数据泄露源头，还能验证数据的完整性和真实性，从而确保隐私信息在数据流通过程中的安全性.本节将对传统水印和基于深度学习的水印技术进行归纳整理，其中依据模型与数据的作用模式，基于深度学习的水印可以进一步分为模型无关水印和模型相关水印.

2.2.1 传统水印

传统水印的生成一般依赖数据载体的人工特征如像素值、频率等，因此特征的选取对水印生成结果影响较大.针对图像数据，依据水印嵌入空间的不同，可以将其进一步分为空间域水印、变换域水印和混合域水印.空间域水印直接嵌入目标图像的像素值中，易受图像变换（压缩、裁剪、旋转）影响；变换域水印通过对信号进行频域变换（如离散傅里叶变换），在频域隐蔽地嵌入水印信息，难以被肉眼直接观察，但是具有较高的计算复杂度；混合域算法则综合前两者的优势，将水印部分嵌入到空间域中，部分嵌入到变换域中，可以根据不同需求调整水印在空间域和频域中的嵌入比例，实现更灵活的水印方案.类似地，面向视频数据、文本数据等同样可以进行适应性地迁移.这类传统水印可解释性强，但是过分依赖人工设计特征，导致单一方法的泛化性差，难以应对复杂场景.

2.2.2 模型无关水印

模型无关水印，与数据的使用方式相对独立，主要研究点是如何实现对现成的数据直接添加水印，关注数据本身.因为其直接作用于数据本身，并不依赖数据流通过程中的特定行为，模型无关水印通常与其他阶段的隐私保护技术具有较好的兼容性.进一步地，模型无关水印可以将水印生成方法细分为基于文本规则的水印生成和基于生成内容的水印生成.

基于文本规则 基于文本规则的水印生成方法，顾名思义，该类研究工作从文本固有规则作为切入点，包括文本格式、词汇语法、语义、句式等.其基本思路是通过检测添加水印数据中特定的文本特征进行判断.一种基于词汇替换的方法是 Sato[13]等人提出 EasyMark 的水印生成技术，其核心思路是对文本数据编码进行字符替换，通过检测水印文本数据中存在的替换字符进行水印检测.但是这种方法鲁棒性较差，攻击者在了解攻击机制的前提下仅需重新处理文本即可消除水印.类似地，Munyer 和 Zhong[14]通过使用预训练的 Word2Vec 模型进行语义建模，将选定的词汇转换为向量，并识别n个最近的向量作为替换候选，最后使用了一个包含预训练 BERT 模型和 转换器（transformer）模块的二分类器进行水印检测，但是该方法往往忽略目标词的上下文，可能会损害句子语义信息和文本质量.除此之外，Atallah 等[15] 引入了三种典型的句法转换进行水印嵌入——附加语移动、分裂句和被动化.每种转换类型被分配了一个唯一的信息位比如被动化对应数字 2，在水印检测过程中，原始文本和添加水印的文本都被转换为句法树，通过比较句法结构来提取水印信息.

基于生成内容 上述基于文本规则的水印生成方法研究虽然在过去几年例取得了显著进展，但是由于其普遍依赖特定的文本规则，易导致文本质量的下降.基于生成内容的水印生成方法，则是实现了利用原始文本和

水印信息直接生成水印文本.Abdelnabi[16]等人提出的一种端到端的水印生成方法 AWT.该方案在水印嵌入过程中，首先使用 transformer 的编码器对句子进行编码，并将句子和预先准备的水印信息后合并后输入 transformer 下游的解码器，最后得到水印文本.水印检测的时候，将水印文本输入 transformer 的编码器，反向解出其中的水印信息.伴随着近年来大语言模型的兴起，其生成高质量文本的能力引起水印研究人员的高度关注.大语言模型强大的释义能力能够增强生成文本的自然性，帮助产生更流畅和更自然的水印内容.Kang 和 Lau[17] 等人设计了一种基于大语言模型的水印生成方案 WATERFALL，该方法使用大语言模型对原始文本进行释义，并创新性地提出一种正交水印扰动方法，提高了该方法的性能和鲁棒性.

模型无关水印关注数据本身的特性使其在工程实现上具有相当的便利性，例如 EasyMark [13]在设计中提到该水印生成代码可以直接整合到大语言模型的输出中，这个过程无关模型的具体实现方式，相当于手动为模型输出添加水印.但是，模型无关水印方案的鲁棒性仍有待优化，如果方案设计过于简单且攻击者获得了水印方案相关的知识，水印信息易被移除.

2.2.3 模型相关水印

不同于模型无关水印，模型相关水印的生成方案在水印嵌入过程中往往需要利用模型进行特定操作，可能变更模型的参数、或是修改模型的输出机制.根据水印生成过程中模型作用的不同，模型相关水印生成方案可以细分为训练时水印和推断时水印.

训练时水印 训练时水印的特点是水印生成过程需要改变模型的参数，其中一种思路是通过在数据集中加入一些触发器（trigger）样本，在模型训练时植入后门来实现，因此也常被称作后门水印.当这些触发器出现在输入中时，模型会表现出特定的行为（如特定的格式或输出），因此这种水印可以由数据集提供者添加以保护数据集版权，或者由模型提供者添加以保护模型版权.需要注意的是，由于大模型训练开销大，重训练困难，因此训练时水印中涉及的模型普遍采用一些简单分类模型进行探讨.文献[18]中提出一种简单有效的基于后门的水印算法，通过给触发器样本指定一个错误的标签，在训练中为神经网络植入后门.但是，这种错误标签可能导致模型出现明显的性能下降，因此，文献[19]提出先通过对抗攻击生成不可察觉的扰动替换错误标签样本，掩盖样本本身有用的特征，确保水印样本与原始标签保持一致.然后植入后门触发器样本，让模型建立触发器到标签的映射.

推断时水印 推断时水印技术强调在不更改模型的参数的前提下完成水印的嵌入.其基本思路是通过改变模型在输出单词（token）时的操作机制把水印嵌入到生成的文本中.Kirchenbauer [20] 等人尝试通过修改模型输出 logits 进行水印嵌入.其方案实现基于一个“红绿表”的策略，通过更改 logits 的概率，使得偏差能被水印检测器检测到.生成红绿表的过程中利用私钥与哈希函数以保证随机性，其基础原理是随机筛除一半的候选词，则模型的输出必不含这些词，而人工生成的句子将以 50%的概率包含筛除列表中的词汇，通过检测概率分布的偏差实现水印检测. 此外，部分工作思路则是从 logits 到 token 间的采样过程入手，基于私钥和哈希函数等生成伪随机数，并将其转化为伪随机向量，应用指数最小采样来选择相应单词.水印检测则是通过评估文本与伪随机向量序列之间的对齐情况是否超过阈值.

2.2.4 未来挑战

鲁棒性问题 在介绍模型无关水印的过程中我们初步讨论了基于文本规则的水印生成算法的鲁棒性.事实上，在具有严格的句法或格式要求的场景下，水印文本空间的上限较低，水印生成更加困难，目前仍然缺乏合适的解决方案;另一方面，[3]中提到的公开可验证性场景也给水印生成带来极大挑战.在这种场景中，水印检测器对用户公开，用户可能利用检测器伪造水印.检测器公开给攻击者在设计目标攻击算法时带来较大的便利性，对水印算法的鲁棒性造成严峻挑战，例如使用检测器通过逆向工程获取生成器实现水印伪造[21].未来工作应针对上述所说的实际场景探索更加鲁棒的水印生成算法，抵御潜在的水印伪造攻击.

数据集版权 虽然前文中我们也讨论了相关的数据集版权保护工作[18-19], 但是这类工作在设计工程中普遍借鉴了后门攻击的思路，即在大量数据中植入部分触发器样本.在这种场景下，修改数据集的其它样本，可能对水印检测结果无明显影响，即忽视了整个数据集的版权保护.在现实场景下，恶意的数据篡改可能会对数

据的原始所有者造成重大影响，引发道德和法律争议.

3 数据流通

数据流通是指数据在不同主体（系统、组织或平台）之间进行共享、传递和交换的过程.数据流通实现了不同主体之间的数据互通，促进信息共享和协作，提高数据的可访问性和可用性，是跨主体数据应用和分析的基础，也是隐私计算（Privacy Computing）的主要研究对象.依据数据流通中所涉及的技术类型，数据流通过程中所涉及的隐私关键技术可分为基于密码学技术、基于合作学习技术、区块链技术三大类型，本节将依次对相关领域现状与挑战进行归纳整理，并初步探讨可能的多技术融合场景.

3.1 基于密码学技术

3.1.1 同态加密

同态加密(Homomorphic Encryption, HE) 指满足密文同态运算性质的加密算法.数据在经过同态加密之后，在密文上进行特定计算，用户可以利用密文的计算结果进行相应的同态解密，解密后的明文等同于对明文数据直接进行相同计算操作.根据支持运算操作的不同，可以分为全同态加密(Fully Homomorphic Encryption, FHE)和部分同态加密（Partially Homomorphic Encryption）.自第一代全同态加密方案问世以来，十余年间四代全同态加密在自举效率、明文打包技术等多个方面取得了长足进步，逐步成为安全多方计算、区块链、联邦学习等隐私保护领域的重要工具.

第一代全同态加密方案基于理想格上的数学困难问题如最短向量问题(Shortest Vector Problem, SVP)或其推广形式最近向量问题(Closest Vector Problem, CVP), 首次提出自举（Bootstrapping）技术，通过同态解密刷新密文噪声, 其核心思想是通过类同态方案的解密电路压缩和循环安全假设实现全同态计算，但存在自举效率低下、参数生成复杂、误差积累等短板；第二代全同态加密方案基于错误学习问题(LWE)和环上错误学习问题(RLWE)问题，利用密钥切换(KeySwitch)控制密文乘法的维数扩展,并基于模交换（ModSwitch）实现噪声线性增长.其引入 SIMD 技术，支持多明文打包和并行计算，目前 HElib 等开源库实现高效自举，在 128 位安全参数下单比特自举速度可达 1.3ms，但是仍需要加速自举中的比特提取步骤. 第三代方案则首次将矩阵作为一种密文结构引入，利用近似特征向量技术构造无需公钥就可以进行同态运算的同态加密方案，支持可编程自举和非线性运算，但是不支持打包技术，空间利用率低下.第四代方案采用近似计算策略，通过正则嵌入编码和噪声融合技术，牺牲精确性以适配机器学习等场景，精度与噪声管理为其核心挑战.四代全同态加密方案的优缺点如表 1 所示.需要注意的是,虽然学术界一般将同态加密算法分为四代，但是四代加密算法尤其是后三者并非是简单的优化替代关系，而是齐头并进，共同发展.

目前全同态加密方案研究存在三大共性挑战：一是打破效率瓶颈，提高算法实用性.较高的计算复杂度和较大的密文规模，导致全同态加密应用实际落地难度颇大，其在隐私安全领域巨大的应用潜力仍未被激活.如何提高全同态加密算法的计算效率使之真正落地仍是未来几年间该领域的重要课题.如 2018 年 Myers[29] 等人就针对该问题提出了一种打包密文的自举算法，在 FHEW 的基础上使用递归的计算方法，成功将计算复杂度从$O(3^{\rho}\lambda^{1/\rho})$降低到$O(1)$.二是安全性挑战.目前循环安全假设仍缺乏严格的数学证明，在后续方案研究中应尝试突破循环安全假设依赖.此外，提升方案的 IND-CPA/CCA2 安全性值得深入研究.Li 和 Micciancio[30]于 2021 年指出虽然 CKKS 方案满足 IND-CPA 安全，但是在面对更强的 IND-CPA^{D} 敌手时无法抵御恢复密钥攻击.因此，Li 进一步提出噪声泛洪（Noise Flooding）技术,但是导致解密明文精度出现较大幅度的下滑.Dai [31]等人则指出基于身份或属性的类 GSW 加密方案研究不满足 CCA2 安全.值得强调的是，目前对同态加密领域的后量子安全性证明的探索仍相对空白，是一个极具挑战和研究价值的领域. 三是应用拓展.同态加密技术作为隐私计算领域的重要基石，在安全多方计算、联邦学习、区块链等领域都具有极大应用潜力，但是如何降低引入同态加密带来的额外开销需要定制化的方案设计.

表 1 四代全同态加密方案对比

类型	代表工作	关键突破	优势	劣势
1st	Gentry [22], van Dijk[23]	自举技术	Gentry 的工作第一次实现了全同态加密，其创新性地提出自举技术为后续研究奠定基础	自举效率低、参数生成复杂、误差的积累会导致后续出现自举困难问题
2nd	BGV[24], BFV[25]	明文打包技术，模交换技术	采用了高效的明文打包技术,显著提高了计算效率	自举速度较慢,不适用于非线性计算
3rd	GSW[26], TFHE[27]	近似特征向量技术	显著提高了自举速度,能够高效地完成逻辑门形式的密文运算	不支持打包技术,计算效率受限；空间利用率低
4th	CKKS[28]	近似计算策略	支持高效地打包和计算,在机器学习领域等近似计算场景具有明显优势	自举速度较慢,存在精度损失,同样不支持非线性计算

3.1.2 差分隐私

差分隐私(Differential Privacy, DP), 通过对数据添加适当的噪声,使敌手难以分辨两个相邻数据分布之间的统计差别，进而达到数据隐私与可用性之间的平衡.根据模型架构设计，可以细分为中心化差分隐私(Central Differential Privacy, CDP)[32-33]、本地化差分隐私(Local Differential Privacy, LDP)[34]和差分隐私洗牌模型(Shuffle Model Differential Privacy)[35].中心化差分隐私即最早由 Dwork 提出的差分隐私模型，由中央服务器收集分布于若干个用户端的数据以进行统计分析,为数据库查询提供用户数据集的均值、方差等统计信息，其依赖于一个可信的第三方中央服务器的假设，在系统鲁棒性上存在明显不足，易出现单点故障问题.本地化差分隐私则是在中心化差分隐私的基础上进行了分布式改进，不再依赖可信的中央服务器.用户数据在本地进行随机化处理，服务器端仅负责进行聚合，使用与随机化处理对应的修正算法可以得到所需的无偏估计量.需要注意的是，中心化差分隐私的噪声添加对象是数据集的统计输出结果,而本地化差分隐私的噪声添加对象是单一用户的原始数据. 差分隐私洗牌模型则是在本地差分隐私模型的基础上进行了创新性地重构，提出 ESA 架构.三种差分隐私模型比较结果如表 2 所示.

表 2 差分隐私模型对比(●代表该项指标表现优异，○代表该项指标存在明显劣势
例如：●表示 CDP 的计算复杂度在三者间最低，通信开销最低)

评价指标	CDP	LDP	Shuffle Model DP
鲁棒性	○	●	◎
复杂度	●	◎	○
通信开销	●	○	◎
隐私性	○	●	◎
模型性能	●	○	◎

差分隐私技术的重要价值在于其提供严格的数学定义平衡数据的隐私性和可用性，其基本定义如定义 1 所示.

定义 1 $(\varepsilon-\delta)DP$ 假设有随机的机制 M，若 M 满足 $(\varepsilon-\delta)DP$，那么对于任意两个相邻数据集 D_1 和 D_2 且两个数据集有且只有一条记录不同，对于任意可能的输出满足 $O \subseteq Range(M)$，那么这两个相邻集合的概率分布满足如下约束条件：

$$\Pr(M(D_1)\in O)\le e^{\varepsilon}\Pr(M(D_2)\in O)+\delta \tag{1}$$

上述定义有时也被称作松弛型差分隐私. 其中 ε 被称为隐私预算，δ 为偏移. 隐私预算的大小直接影响着两个数据分布间的统计差距. 对攻击者而言，隐私预算越小，越难以通过差分攻击获取有效信息，隐私保护能力越强；对数据使用方而言，隐私预算越大，统计数据的精度就越高，数据可用性越强. 一般地，使用者通过注入噪声以满足 $(\varepsilon-\delta)$ 差分隐私，噪声从特定分布如高斯分布中采样获取.

差分隐私的本质在于保证相邻数据集对应输出的概率分布不可分，上述定义通过 $(\varepsilon-\delta)$ 量化数据分布的差异. 事实上，在数据科学中存在多种量化数据分布差异的指标. 其中瑞丽散度(Renyi Divergence)[36-38]在最新的差分隐私研究中被证明可以提供更加严格的隐私边界，为差分隐私证明尤其是机器学习领域提供新的视角. 值得一提的是, 瑞丽差分隐私(Renyi Differential Privacy, RDP)可推导得出对应的 $(\varepsilon-\delta)$ 差分隐私.

当前差分隐私研究的主要挑战包括：一是效用与隐私的权衡.高维数据中添加噪声易导致信息的显著损失，设计更精细的扰动策略尤为关键；二是动态场景下的噪声添加.在现实场景中用户的隐私需求存在动态变化，现有差分隐私机制的隐私预算分配仍依赖启发式方法，缺乏理论最优解，噪声无法根据隐私需求进行实时调整，造成不必要的效用损失.三是隐私攻防迭代.差分隐私通过噪声扰动实现对成员推断攻击的防御，但是伴随攻击方式的不断演变（如后门攻击、数据重构攻击），差分隐私的隐私保护性能被逐渐瓦解.例如 Tan[39]等人指出由于差分隐私的核心目标与数据重构攻击重构整个数据集分布的动机不契合，相比于成员推断攻击，差分隐私面对数据重构攻击的防御能力较为薄弱.

3.1.3 安全多方计算

安全多方计算 (Secure Multi-Party Computation, MPC)[40] 在无可信第三方的情况下允许多个参与方在不泄露各自私有数据的情况下，协同计算任意私有数据相关函数. 假设有 n 个参与方 $U_1,U_2,\cdots,U_n$,彼此协作计算函数 $f(x_1,x_2,\cdots,x_n)$，其中 x_i 为参与方 U_i 基于私人数据的输入变量，$f(\cdot)$ 是预先协商的函数.根据安全多方计算的定义，一方面要保证隐私性，即敌手或不诚实参与方无法在计算过程中获取到参与方 U_i 的输入信息 x_i；另一方面要求保证计算的正确性，即最终结果满足 $g(y_1,y_2,\cdots,y_n)=f(x_1,x_2,\cdots,x_n)$ 而非其他函数, 其中 y_i 为参与方 U_i 的输出，$g(\cdot)$ 是根据 MPC 协议设计的函数.目前，其设计思路主要有基于秘密共享(secret sharing)的方法[41-42],基于混淆电路(garbled circuit)的方法[43-45]和基于同态加密的方法[46].

基于秘密共享 其基本思想是通过数据分片，使得各参与方无法知悉完整数据，通过多方协作执行计算协议，最终仅恢复计算结果，而各方的输入数据全程保持隐私.进一步地，秘密分享可以划分为严格秘密分享和阈值秘密分享，严格秘密分享要求所有人参与分享过程，阈值秘密分享则只需要满足一定人数，就可以进行分享.阈值秘密分享的代表是 Shamir 秘密分享.其结合线性方程组的特性，通过拉格朗日插值构造相应多项式方程，确保少于阈值数量的参与方无法恢复原始数据.该类方案支持任意计算逻辑，从信息论角度安全性高，适用于大规模多方场景.但是其乘法协议需预先计算开销，恶意模型下通信复杂度较高.

基于混淆电路 该类方案将计算逻辑转换为布尔电路，通过加密和置换混淆电路中的逻辑门，使参与方在不解密输入的情况下合作完成计算.1984 年 Yao[40] 首先提出一种基于混淆电路和不经意传输的 S2PC 协议.生成方（Generator）加密电路，执行方（Evaluator）通过不经意传输（OT）获取输入标签，逐门解密输出.由于每个电路门只涉及恒定数量的对称的密码操作，电路极为高效.但是，Yao 协议的安全性建立在半诚实敌手模型下，无法应对恶意敌手，因此 Lindell[44]等人引入 Cut-and-Choose 技术，约束潜在的恶意敌手生成器以半诚实的方式行动，验证多数电路正确性，牺牲部分效率换取强安全性.基于混淆电路的方案主要适用于两方安全计算，在恶意敌手模型下实现强安全性.其挑战在于电路规模随计算复杂度指数增长，高延迟问题突出.

基于同态加密 参与方使用同态加密算法加密各自输入，在密文上直接执行计算，最终解密获取结果，过程中明文数据始终保密.基于部分同态加密的方案如 Paillier(加法同态)、ElGamal(乘法同态)，适用于特定计算场景, 该类方案实现难度小，计算开销小，已在实践中使用.基于全同态加密的方案适用范围广，支持任意次加法和乘法操作(如 BGV、CKKS 方案)，需要通过自举控制噪声，且全同态加密方案开销巨大，离实际应用还

有一定距离.在实际使用中需要权衡安全强度与性能选择合适方案.一般地，基于同态加密的方案需要结合其他协议进行混合协议设计，如结合阈值解密协议，避免单一密钥持有者成为信任瓶颈.为保证密文没有泄漏同态操作的秘密信息，同态计算后的密文需要满足 “电路隐私” (Circuit Privacy) 性质.

3.1.4 零知识证明

零知识证明(Zero-Knowledge Proof, ZKP), 是一种重要的密码学协议，它允许证明者（Prover）向验证者（Verifier）在不泄露任何相关信息的额外细节的前提下，证明某个论断是正确的.零知识证明通过密码学方法实现了“知识可验证而不可见”的范式突破，是隐私计算与可信验证的核心技术之一.

应用场景 目前，零知识证明协议作为数据隐私保护的工具被广泛使用. 以下介绍零知识证明的几类典型应用. 一是区块链与加密货币. 零知识证明被广泛运用于区块链中，在证明交易合法的同时将交易双方和金额作为证据隐藏起来，保障交易者的隐私. 如 Zcash、门罗币使用 zk-SNARKs 隐藏交易金额与地址. 此外，以太坊中在链下应用零知识证明 zk-Rollup，验证链下打包交易的有效性，节省链上空间. 二是身份认证. 用户可能不愿意通过密码等方式进行繁琐的身份认证，而利用零知识证明，用户无需透露密码或敏感信息即可证明其身份. 三是机器学习中的数据共享. 零知识证明可被用于验证 AI 模型推理过程正确性(如 zkML)或是与联邦学习结合保障数据隐私.

目前学界一般根据交互模式将零知识证明划分为交互式零知识证明(Interactive Zero-Knowledge Proofs)和非交互式零知识证明(Non-Interactive Zero-Knowledge Proofs, NIZK).交互式零知识证明通过多轮交互完成验证. 每一轮中，验证者发送随机挑战，证明者根据挑战生成响应.通过反复交互，验证者逐步确信陈述的真实性，同时不获取额外信息.交互式零知识证明适合小规模实时场景，但通信开销高.非交互式零知识证明仅生成一次证明，验证者无需额外交互即可独立验证.其通过公共参考字符串(Common Reference String, CRS)或随机谕示器（Random Oracle）实现，适用于大规模隐私计算场景，但需权衡信任假设与性能.特别地，Fiat 和 Shamir[47] 在研究基于二次剩余根的身份识别协议过程中发现可将交互式证明转换为非交互式证明，因此零知识证明协议的构造可从交互式证明入手.

目前零知识证明仍然存在如下挑战：一是性能瓶颈问题.零知识证明的生成和验证计算量大，如何提高效率并降低计算存储成本一直是一个重要挑战.尽管近年来出现了 zk-SNARKs[48] (Zero-Knowledge Succinct Non-Interactive Arguments of Knowledge)这样的高效零知识证明协议. 但是其证明生成时间仍较长(分钟级)，需在硬件加速与算法优化上开展进一步研究. 现有大部分协议还是需求对数或对数多项式级验证时间和证明长度,能否将三者降低到线性复杂度仍有待研究.二是安全与信任假设.zk-SNARKs[48]和 zk-STARKS 是近年来出现的两种高效零知识证明协议.其中 zk-SNARKs 基于椭圆曲线双线性配对和多项式承诺（如 KZG 承诺）,依赖公共参考字符串的可信设置, 一定程度上降低了协议的安全性和实用性.STARK 协议则基于哈希函数和默克尔树使用可公开验证的随机性解决交互问题，缺点是证明空间需求大，导致 STARK 证明在以太坊上的验证成本更高，难以适配低带宽场景.设计高效且不依赖可信设置的系统和基于 SNARK 或 STARK 协议的优化问题将是该领域的重要研究方向；三是标准化与用户友好性.现有的零知识证明技术框架繁琐，多种零知识证明框架(Circom、Noir、Leo)并存，缺乏统一标准.如何将其应用于实际系统并使其对终端用户友好是一个充满挑战性的课题.

3.2 基于合作学习技术

3.2.1 联邦学习

联邦学习[49-51]，由谷歌工程师于 2016 年首先提出.联邦学习实现了在数据不离开本地客户端的前提下进行多方数据的联合学习，一定程度上解决了“数据孤岛”问题,避免了隐私数据的直接泄露.论文中提出的 FedAvg[52] 算法建立了基于梯度聚合的服务器-客户端联邦学习框架，这种基于服务器-客户端的联邦学习框架结构被大多数的联邦学习算法所采用.虽然联邦学习的出发点契合当今时代对数据隐私的要求，但是后续的研究表明：联邦学习算法并未真正解决分布式数据的隐私安全问题，利用梯度代替客户端数据递交给服务器可能间接导致数据隐私的泄露[53-54].为进一步提高联邦学习的隐私性，隐私联邦研究者一方面围绕算法本身进行

改变，如 FedProto[55]中就提出使用原型代替梯度作为信息传递的媒介，利用原型计算的不可逆性防止数据信息的泄露；另一方面，联邦学习在交叉领域潜力极大，与其他隐私计算技术具有极佳的兼容性，代表性成果包括结合差分隐私技术[56]，根据隐私性需求对数据添加不同强度的噪声；结合同态加密技术[57]，对传递的梯度进行加密，从而在密文上实现梯度的聚合.

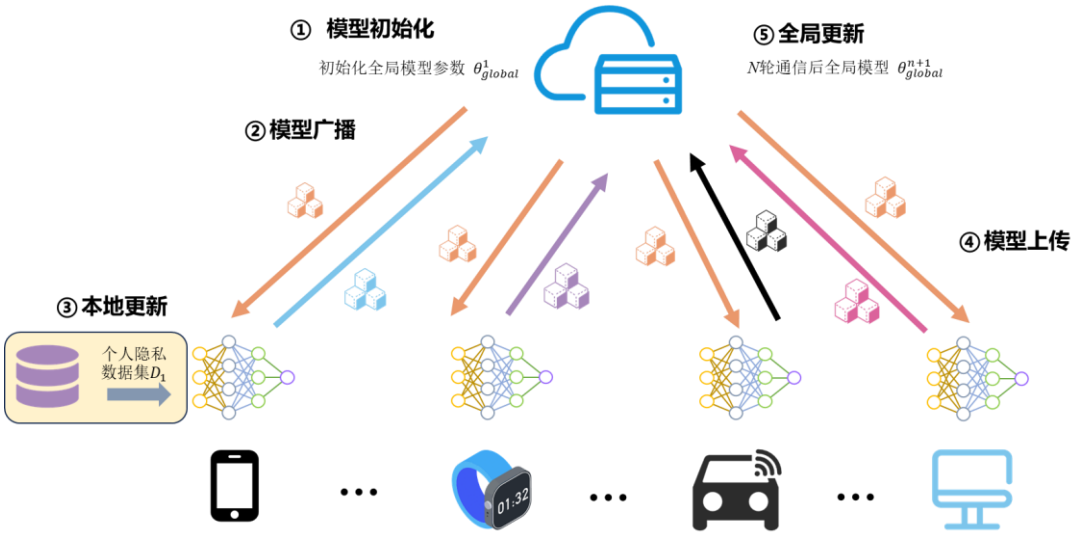


图 3 传统联邦学习框架示意图

3.2.2 传统联邦学习

传统联邦学习普遍采取的方式是使用梯度信息代替原始数据，用以进行联合训练.这类基于梯度聚合的联邦学习算法，本文将其归纳为传统联邦学习算法.一般地，每一通信轮次传统联邦学习算法的分为如下五个步骤：

1) 模型初始化.中央服务器挑选参与训练的客户端，并对全局模型的参数进行初始化设置.
2) 模型广播.传统联邦学习中服务器和客户端一般采用相同结构，因此，本步骤在操作中一般直接传递模型参数，以实现模型下发的目的.
3) 本地更新.客户端下载全局模型参数以更新本地模型，并使用本地数据集进行若干轮次的本地训练，准备将训练所得到的梯度或模型参数信息回传中央服务器.
4) 模型上传.本地客户端上传本地模型（梯度或其他参数），服务器接收来自各个客户端的回传信息，根据联邦学习算法进行模型聚合.
5) 全局更新.服务器根据聚合的结果对全局模型参数进行更新，用于下一轮次训练的初始化.

基于传统联邦学习的流程框架，假设每个客户端拥有的数据集为 D_i，数据集的采集来源遵从概率分布 $P_i(x,y)$，其中 x 和 y 分别代表数据集中样本的输入特征和样本标签.全局模型记为 $F(w;x)$，其中 $F(\cdot)$ 代表完成超参数设定的模型架构，代表模型参数，x 代表输入样本的特征向量.那么传统联邦学习的优化目标可以形式化如下：

$$\arg\min_{w}\sum_{i=1}^{n}\frac{\|D_i\|}{N}L_s(F(w;x),y) \tag{2}$$

上述流程可以发现传统的联邦学习算法如 FedAvg[52]等算法普遍存在两个共性问题：一是在理论和实验上普遍依赖于数据独立同分布的统计假设，与数据非独立同分布的现实场景相差较远,在实际优化过程中往往会出现“客户端漂移”现象[58],模型性能出现较大幅度下降；二是传统联邦学习研究中各客户端普遍采用本地模

型-全局模型的架构设计，即客户端和服务器采用一个相同的模型架构.事实上，这不仅牺牲了客户的个性化需求也忽视了训练中可能存在的公平性问题.

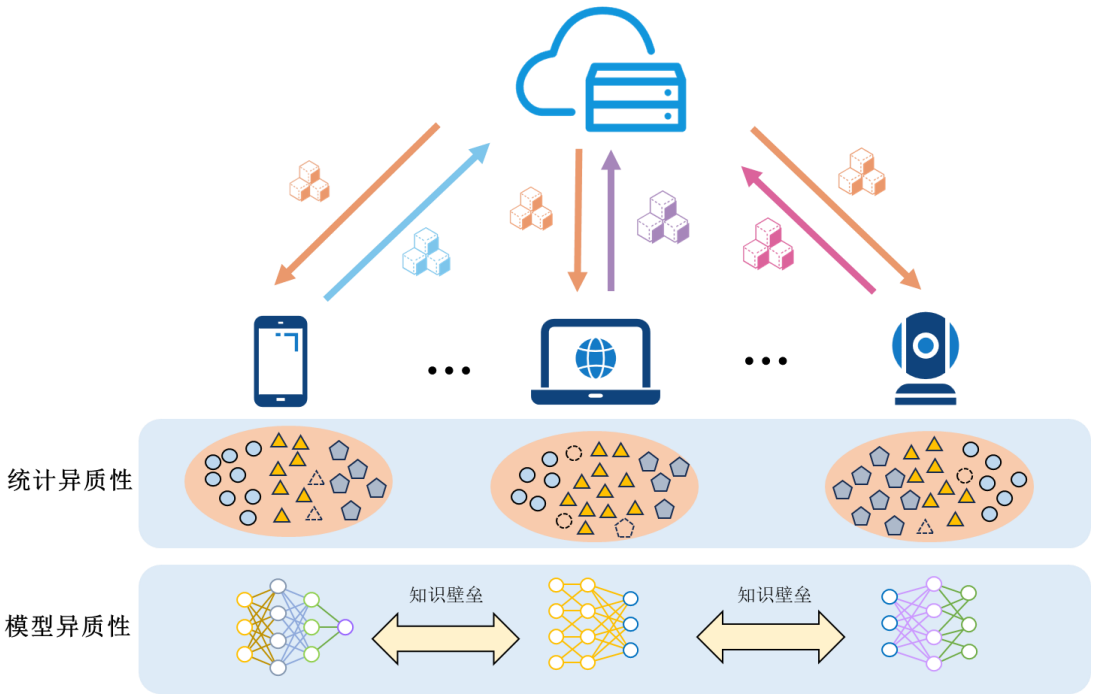


图 4 个性化联邦学习框架示意图

3.2.3 个性化联邦学习

个性化联邦学习，是一种旨在向用户提供个性化定制服务的联合学习范式.个性化联邦主要在两个方面对传统联邦学习算法提出挑战，一是要求当用户间的数据分布并不一致时，使得所有用户在联合训练后均取得理想的模型表现，即统计异质性（Statistical Heterogeneity）问题；二是要求模型能够按照用户需求执行不同的任务，如个性化推荐、关键词搜索等，即模型异质性（Model Heterogeneity）问题.

上述两种异质性问题是个性化联邦学习算法的主要研究对象，因此就研究范畴而言，个性化联邦与异质联邦有相当大的重叠.简单来说，两者在技术层面均是针对现实场景中联邦学习中可能遇到的各种异质性问题进行解决，但是个性化联邦从用户需求的出发，而异质联邦则是从异质性问题这一统计现象出发.除了统计异质性问题和模型异质性问题，Ye[59]等人对相关研究梳理后另外提出通信异质性和设备异质性作为异质性问题的有效补充.

个性化联邦场景下要求各个用户持有不同的模型，假设第 i 个用户拥有的数据集为 D_i，数据集遵从概率分布 $P_i(x,y)$，所持模型为 $F_i(w_i;x)$，w_i 代表模型参数，x 代表输入样本的特征向量.那么个性化联邦的优化目标可以形式化如下：

$$\arg\min_{w}\sum_{i=1}^{n}\frac{\|D_i\|}{N}L_s(F_i(w_i;x),y) \tag{3}$$

在个性化联邦领域，由于各个客户端持有不同的模型，在模型优化过程中，假设第 i 个客户端的模型参数最优解为 w_i^*,由于不同用户的模型参数最优解不同，传统联邦学习中训练单个全局模型难以兼顾用户需求.但是在个性化场景下，由于将本地模型参数 w_i 从全局参数 w 分离，在经过模型训练后，各个客户端的模型参数 w_i 均存在取得最优解 w_i^* 的可能，而不必采取平均最优解等折中方案.

针对统计异质性问题，近几年涌现出一批卓有成效的工作.具体来说，Liam[60]等人提出 FedRep 算法，创新地将交替最小化方法融入了联邦学习的模型训练流程，实现了模型的分层训练，在统计异质场景取得较好表现；Liang[61]提出 LG-FedAvg 算法,不同于共同训练上游特征提取器的传统分层思路,LG-FedAvg 算法通过共同训练下游的分类器解决数据异构问题.针对模型异质性问题，Wu[62]等人利用知识蒸馏，利用教师模型指导学生模型训练，实现不同模型间的知识传递，一定程度解决模型异质问题；Ghuhan[63]等人第一次提出个性化层的思想以实现模型的个性化架构，对于后续基于模型分割的个性化联邦学习算法具有重大启发意义. 部分工作在两个问题上都取得较好地结果,如 Tan[55]等人提出利用原型聚合代替传统梯度聚合方法的原型联邦算法，在极大提高模型在异质数据上性能的同时在模型异质问题上取得较优成果.虽然现有个性化联邦学习算法在模型性能上取得长足进步，但是其中伴随的隐私问题并没有得到相应的重视.Ye 等人[59]在对异质联邦工作梳理中发现，异质性问题加剧了隐私暴露的风险；原型联邦代表工作 FedProto[55]中在对框架隐私性讨论时仅从原型的维度和平均操作的不可逆上主观判断原型联邦具有较强的隐私保护能力，并未开展隐私性验证实验，因此原型联邦框架的隐私性仍有待探索.

3.2.4 未来挑战

联邦学习领域目前的挑战主要包括：一是联邦学习中隐私和性能间的权衡问题.联邦学习传递梯度信息的方式被后续研究者发现是不安全的，造成隐私数据的间接泄露.因此联邦学习常与其他隐私计算技术相结合.但是差分隐私带来的噪声对模型性能的影响较大，在高隐私性要求环境，目前模型性能的下降难以被接受.二是个性化联邦.个性化联邦更加贴近用户、贴合现实环境的特点使得近几年愈来愈多的研究者进入该领域.但是正如前文中的分析，同时较好地解决两种异质性问题的工作如原型联邦仍然较少，但是原型联邦中隐私性分析有所缺失，需要后来者的深入挖掘.三是联邦学习中的隐私攻防.针对联邦学习的代表性攻击目前主要包括后门攻击[64]、推断攻击[65]、样本重构攻击[54]、模型投毒攻击[66]、间谍攻击[67]等；联邦学习的防御性工作的主要方法包括差分隐私[56]、同态加密[57]、可信执行环境（TEE）[68], 这类防御性工作或是计算开销过大，或是影响模型整体表现，且普遍无法应对多种攻击方式，如何建立更加鲁棒、完备的隐私联邦学习范式在未来仍是颇具挑战性的工作.

3.3 区块链

3.3.1 技术特性

区块链，以密码学技术为基石，创新性地融合分布式数据库、博弈论、点对点网络、智能合约等多项技术，在不依赖中心机构的前提下提供可信数据.其主要包含以下特性：

1) 去中心化.相较于传统的中心化数据库，区块链采取多个节点共同管理账本的范式，通过共识算法保证节点间的一致性，消除了单点故障的风险，提高了系统的鲁棒性.
2) 不可篡改性.对于已有区块记录，区块链通过哈希函数建立链式结构，每个区块头部包含前一个区块的哈希值，从而形成一个连贯的链条，任何对数据的篡改都会破坏整条链的完整性.
3) 透明性.区块链采用分布式账本技术，所有参与节点都持有相同的数据副本，且所有交易和数据都可以被公开验证和追踪.这种透明性对防止欺诈和腐败具有重要意义，但也加剧了区块链系统可能遭遇的隐私挑战.
4) 可追溯性.区块链基于哈希值关联形成按时间顺序组织的链式结构，因此，在进行数据溯源时，我们可以通过区块时间戳开始快速追溯所有相关的历史信息.
5) 自动化.智能合约是部署在区块链上的自动执行代码，能够在满足特定条件时自动执行合约条款.智能合约的使用减少了人为干预和中介，确保数据流通过程中的自动化和安全性.
6) 隐私性.尽管区块链提供了透明性，但通过使用零知识证明（ZKP）等技术，区块链可以在保护隐私的同时验证交易的有效性.这种技术允许在不暴露具体数据的情况下验证数据的真实性，平衡了透明性和隐私性.

3.3.2 区块链数据流通安全

伴随数据信息的爆炸式增长，海量数据流通场景成为常态，区块链系统需要支持数据在多参与方之间的高效与安全流通.更大规模的数据和更复杂的参与方给区块链中多节点网络架构与机制的设计带来了新的挑

战，包括区块链本身的数据流通性能瓶颈问题，复杂参与方带来的数据跨链跨域需求等.本节将从区块链数据流通性能瓶颈，跨链数据流通，联合学习数据共享三个方面来总结研究现状.

数据流通性能瓶颈 目前区块链系统中随着链上节点数量的上升、数据规模的增长以及隐私保护方案的引入，链上数据流通会出现明显的性能下降，极大影响了其应用场景.为提高数据流通效率，目前的研究工作主要从以下两个角度切入：

1) 交易并行化.交易并行化方向目前的主流研究包括：排序-并行执行方案(order-parallel execute，OXII)和执行-排序-验证方案(execute-order-validate，XOV)两类方案，其代表成果分别是 ParBlockchain[69]和 Hyperledger Fabric[70].两种方法根据交易之间争用和冲突的程度和频率，在进行性能权衡后实现并行执行交易，大幅提升了系统性能.然而后者在支持冲突交易方面存在缺陷而前者在结果不一致时终止交易成本较高.
2) 分片技术.分片账本技术将完整的账本划分为多个分片，这些分片由不同的参与方子集维护.采用分片技术的区块链系统需要处理两种类型的交易：分片内和跨分片交易.分片技术大幅提高区块链系统的扩展性，从而更好地应对大规模的用户节点和数据.例如 AHL[71] 采用了可信执行环境技术来限制分片内节点的恶意行为.SharPer[72]是另一个分片联盟链系统，该系统由一组容错集群组成，每个集群维护区块链账本的一个分片.如何处理跨分片交易是该类研究的核心难点.

跨链数据流通 在区块链技术中，跨链数据流通是一个重要的研究和应用领域.它旨在解决不同区块链网络之间数据和价值传递的问题.现有关于跨链协议的研究工作主要集中在以下三个方面：基于公证人机制、基于侧链机制、基于密码学技术.

1) 基于公证人机制.在公证人机制中，受信任的第三方或一组各方负责见证一条链中的事件，并声明该事件对另一条链有效.通过引入第三方来验证交易信息是否一致、合法,无需对交易的细节进行验证.其优点是实现原理简单，无需复杂工作量证明,缺点是存在中心化风险，并且在联盟存在恶意节点的场景中无法保证跨链操作的安全性.
2) 基于侧链机制.侧链，也被称为锚定侧链，通过双向挂钩(two-way peg)实现不同区块链之间的资产或数据信息转移.在这一过程中，用户通过侧链将主链上的资产锁定，并在侧链上生成相应的原生资产或其表示.目前，实现双向挂钩的机制主要包括三种：中心化双向挂钩[73]、联合双向挂钩[74]、以及简化支付验证(SPV).这些机制分别采用单一公证人、公证人联盟与多重签名、以及轻客户端的形式作为资产的托管方式.侧链机制的优点在于其卓越的可扩展性和安全性，适用范围广泛.然而，其缺点是实现难度较高，并且在面对联盟中存在恶意节点或成员时，难以保证安全性.
3) 基于密码学技术.该方向主要研究通过零知识证明来实现跨链数据传输中的隐私保护.零知识证明可以在分布式账本上转移数字资产，而无需透露有关发送者、接收者或资产数量的任何信息.但是，使用密码学工具尤其是零知识证明来加密隐私数据需要消耗大量算力，将对系统性能造成不可忽视的负面影响.

联合学习数据共享 多参与方共同分享数据的联合学习范式在数据流通中应用广泛，而前文中的联邦学习正是该领域中最具代表性的技术，但是现有的联邦学习方案普遍依赖一个完全诚实的中央服务器进行模型参数聚合与分发，而在现实场景中往往难以找到被所有联合学习参与方共同认可的可信第三方.结合区块链的去中心化特性，部分学者尝试探索基于区块链的联合学习方案.基于区块链的联合学习方案的核心思路是通过共识协议随机选取某一参与方进行模型聚合，将参数聚合结果打包上链作为区块发布，其他参与方负责验证聚合结果.在上述流程中，分布式共识机制能够避免单一中央服务器可能带来的单点故障问题，而区块链上数据的透明性与不可篡改性也可以杜绝第三方中央服务器恶意错误聚合的可能性.但是，由于区块链的透明性，各参与方提交的梯度数据将会被链上所有参与方可见，一定程度上增大参与方隐私泄露的可能.部分研究尝试通过密码学方案如差分隐私[75]进行保护，但是对模型性能和运行效率上存在较大缺陷.此外，如何建立公平透明的参与激励机制，排除恶意的参与方也是一个重要问题.Kim 等[76] 提出了基于工作量证明共识协议的联邦学

习框架，通过出块奖励的方式将收益分配给矿工，而矿工再根据块内打包数据的提供方声明的样本数量与计算时间将出块奖励进行进一步分配.

3.3.3 关键挑战

跨链数据隐私 跨链技术使得不同的区块链网络能够进行数据与价值的传递，促进数据的交换与整合，但这同时带来了如何在不侵犯用户隐私的前提下进行数据跨链传输的挑战.为解决这一问题，传统的密码学技术如同态加密、零知识证明提供了有效思路，在不泄露具体交易信息的前提下保证了交易数据的真实性，但是如何缓解引入密码学技术带来的性能损失，寻求数据跨链效率和数据隐私的平衡点有待后续研究.

海量数据上链效率瓶颈 在区块链技术的快速发展过程中，海量数据上链交易的效率瓶颈已经成为制约区块链发展的重要问题，特别是在金融、医疗健康记录等关键领域，庞大的交易数量对区块链网络的处理速度和吞吐量提出了更高的要求.此外，链上数据量的急剧膨胀进一步加剧了处理和存储的负担.如何在不牺牲区块链去中心化、不可篡改等特性的同时，有效提高海量数据上链交易的处理效率，成为了区块链技术发展亟需解决的重要问题.

4 数据流通后

数据流通后期阶段主要关注数据生命末期的数据治理问题.当用户因故不再需要数据服务，该用户有权要求服务提供商删除用户个人数据，如删除搜索结果中与个人相关的链接，即“被遗忘权”.本节将围绕数据“被遗忘权”介绍机器学习领域中的忘却学习最新研究，建立系统分类体系，总结忘却学习领域的研究难点和未来挑战.

4.1 忘却学习

数据“被遗忘权”(Right to be Forgotten)近年来出现在世界各国的数据保护法案中，包括欧盟《通用数据保护条例》(GDPR)和美国的《加州消费者隐私法案》(CCPA)等.被遗忘权指个体有权要求服务提供商删除用户个人信息.但是在机器学习领域，用户个人数据的删除并非只是从公司数据库将其移除，也应包括消除用户个人数据样本对模型的贡献，但是直接进行重训练在复杂任务场景又将带来昂贵的训练成本.忘却学习这一研究领域专注于从已训练模型中移除指定样本的影响，近年来得到众多研究者的高度关注.

直接进行重新训练是实现忘却学习的有效方法，然而因为它需要存储整个原始数据集并从头开始重新训练模型，这在复杂的深度学习场景中会消耗大量的存储和计算资源，且耗时过长，效率低下.因此，研究人员尝试设计出高效的忘却学习机制.目前研究者发现在忘却学习设计过程中主要存在以下三大难点[77]：训练过程随机性、训练过程的递进性和灾难性性能下降.

根据忘却学习中数据参与训练方式，忘却学习可分为集中式忘却学习(Centralized Unlearning)和联邦忘却学习(Federated Unlearning).

4.2 集中式忘却学习

集中式忘却学习主要研究在中心化场景下消除原始数据对机器学习模型的影响.根据消除影响的精确度，集中式忘却学习研究可以分为精确忘却学习(Exact Unlearning)和非精确忘却学习(Inexact Unlearning).

精确忘却学习 精确忘却学习的基本思想是通过朴素的重新训练新模型的方式，精准移除用户数据对模型的潜在影响.但是前文中分析提到重训练的方式效率低下，对存储资源和计算资源要求高，因此该方向的研究围绕如何提升重训练的效率展开.目前精确忘却学习的普遍操作是通过数据集和模型的划分对原始机器学习流程进行转换，即将数据集划分为多个子集，分别用于多个子模型的训练，最后集成子模型的知识得到最终模型[78-79]，忘却学习过程中只需要重新训练包含被删除数据的子模型即可.这种分片和转换的思想极大降低了重训练的计算和存储成本，但是在大量数据被删除时，仍然不可避免的要重新训练多个子模型，因为被删除数据并不总是在同一个数据子集中.目前精确忘却学习的代表成果是 SISA[78], 通过对完整数据集的分片操作，并利用分片数据对子模型进行训练，进行数据擦除时仅需要重新训练包含被删除数据的分片.

非精确忘却学习 非精确忘却学习则是希望通过对已训练模型的直接操作移除数据样本的影响.其基本思想是通过其他技术手段如修改损失函数等拉近直接消除后的模型与重新训练模型在概率分布上的差距.不同于精确忘却学习基于定义的等价性，非精确忘却学习需要通过精准评估数据对模型的贡献才能判断采取何种方式进行相应地消除.非精确忘却学习的方法可以进一步分为基于理论路线和基于实验路线.基于理论路线的研究尝试通过严格的统计定义保证消除模型和重训练模型间的近似关系，其常见的方式有影响函数（Influence Function）和认证移除（Certified Removal）. 前者的思路是利用影响函数可以估计某个数据点对模型参数的影响量，根据计算结果对模型参数进行调整，以消除特定数据点的影响；后者则是一种类似于差分隐私的忘却学习技术[80-81]，例如 Guo[80]等人提出的$\varepsilon-$非精确忘却学习.其参考差分隐私中对概率分布不可区分的定义，提出忘却学习应确保消除前后的模型必须是$\varepsilon-$不可区分的.基于经验路线的方法则是结合辅助技术的直观经验，常见的策略包括知识蒸馏[82]、反向训练和随机训练的忘却学习.以知识蒸馏为例，知识蒸馏技术已有大量研究表明其可以实现教师模型和子模型间的知识传递.类似地，在忘却学习中可以尝试利用知识蒸馏的方式高效地训练消除后的模型.基于经验路线方法简单直观且有效，但是普遍需要进一步的恢复模型性能操作.

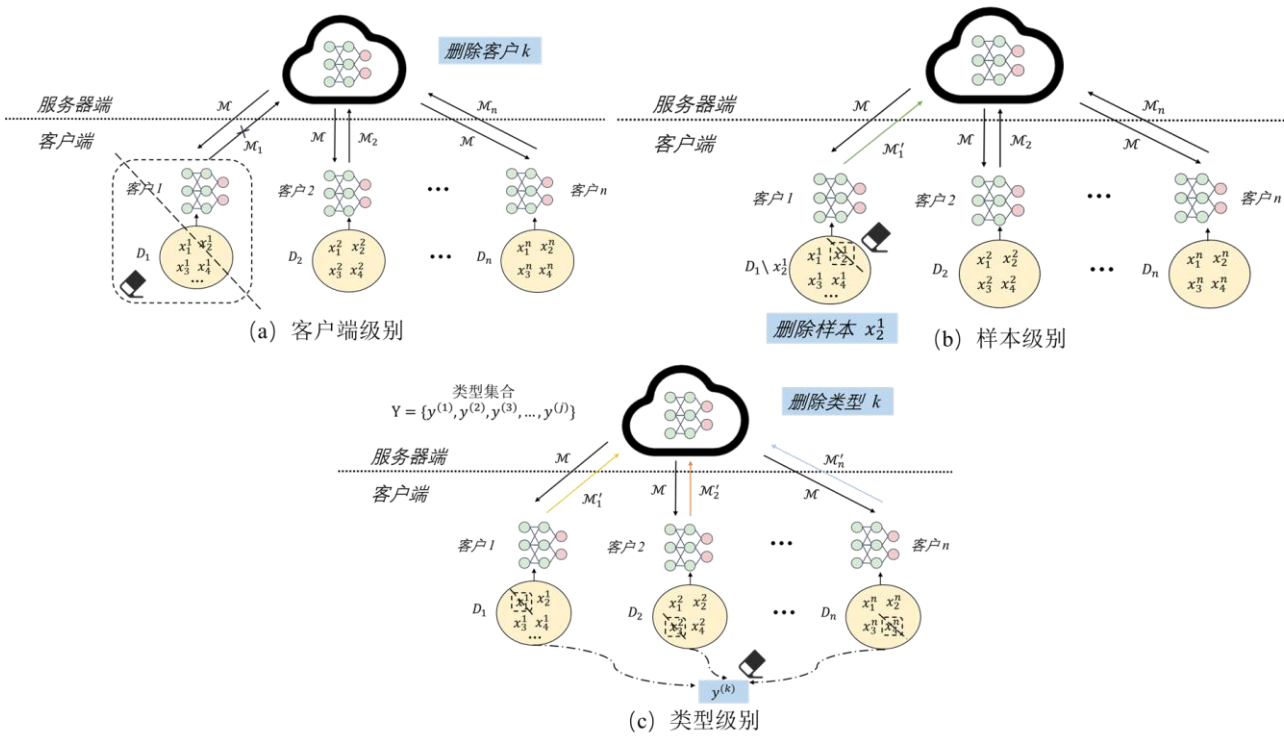


图 5 联邦忘却学习框架示意图

4.3 联邦忘却学习

数据流通过程中不同主体之间的数据传递不可避免，研究分布式环境下的忘却学习有其独特的学术与应用价值，在近年来愈来愈多的研究者将目光投向该领域.联邦忘却学习关注分布式环境即联邦学习设置下的忘却学习技术.在前文中的联邦学习介绍中可以得知，联邦学习中的每个客户端通过传递参数而非原始数据的方式实现模型的联合训练，以防止用户隐私的直接泄露.因此，相较于集中式的忘却学习，联邦忘却学习的数据集访问权限受限，服务器无法接触到客户端的原始数据.为解决这一问题，目前联邦忘却学习的主要方法是通过评估上传参数而非数据的贡献来消除特定客户端对训练模型的影响.根据数据擦除粒度的不同，

联邦忘却学习可以细分为客户端级别（Client-Level）、样本级别(Sample-Level)和类型级别(Class-Level).客户端级别指在学习过程中删除一个客户端的所有数据，代表工作是 FedEraser[80].其在联邦学习的过程中建立每一轮迭代的更新索引，从而在重新训练的过程中进行快速校准.样本级别则是关注消除客户端中单个样本或部分数据集样本，代表工作是 Liu[83]等利用一阶泰勒近似提出的忘却学习算法，利用费舍尔信息矩阵（Fisher Information Matrix, FIM）以低成本近似海森矩阵（Hessian Matrix），忘却学习过程中需要梯度减去该近似矩阵.类型级别则是研究如何选择性地删除某个类别的数据.其代表工作是文献[84] ,其发现卷积神经网络每个通道对不同类别的影响不同，并用一种称为 TF-IDF 的度量指标来量化通道的类别区分度，通过剪枝具有高 TF-IDF 分数的通道实现消除某一类别数据的目的.

4.4 未来挑战

忘却学习领域目前仍然存在以下挑战：一是忘却学习中可能带来的性能损失问题，执行数据擦除后模型性能往往会出现一定程度的下滑，如何在消除数据影响的同时保证模型性能不会出现大幅下降仍然需要不断的优化工作；二是联邦忘却学习中的性能退化问题.由于联邦学习中无法直接访问数据，只能使用非精确的方式，因此更容易出现性能退化问题；三是忘却学习中的隐私问题.部分文献提出忘却学习的更新操作可能泄露隐私信息，并提出了相应的攻击框架[85-86].如何提高忘却学习的隐私性能也将成为未来该领域研究的重要课题.

5 人工智能时代下的新挑战

从 2022 年底 ChatGPT 问世掀起大模型研究热潮，到 2025 年初国产大语言模型 DeepSeek-R1 发布引发全球关注，其在包括编写代码、回答专业问题和翻译在内的下游任务中展示了与人类水平相当甚至超越的准确性和熟练度.大模型的应用正在逐步改变着当今社会的内容生产范式，深层次重构人类与 AI 智能体间的关联，一个崭新时代——人工智能时代已经来临.在大模型研究中，尺度定律(Scaling Law)的存在将模型性能和数据量链接，数据资产的重要性不言而喻，人工智能时代下如何有效保护数据隐私成为重大课题.

人工智能时代下，数据流通中的隐私保护面临着全新挑战.首先，大模型的强大生成能力在给人类创作带来极大便利的同时，其生成内容若被恶意利用可能给社会带来巨大安全隐患，如深度伪造技术(Deepfake)[87-88]的滥用，通过伪造公众人物发言的图片、音频造成恶劣的舆论影响.此外，隐私数据的多元化给数据隐私保护拓宽全新维度.除传统的用户隐私数据外，模型数据如模型参数、模型架构信息等多元化的数据形式，在人工智能时代被赋予全新价值，其重要性有时甚至超过传统数据.最后，数据动态性给隐私机制设计带来极大挑战.现有 AI 应用普遍会通过在线学习过程中不断吸纳新的用户数据，新数据的引入可能使得旧的隐私防护机制失效.此外，大模型的记忆性使得大模型更加难以遗忘已学习的数据，现有忘却学习技术难以以较低成本实现理想的遗忘效果，例如某用户要求社交媒体删除历史记录，但 AI 推荐系统仍可能保留其行为特征.

生成式 AI 安全 伴随着人工智能研究的不断进步，生成式 AI 的生成能力在近年来飞速发展，多模态数据均可以通过 AI 生成理想成果，如文本生成模型 GPT 系列模型、图像生成模型 Midjourney 与 Stable Diffusion、视频生成模型 Sora.但是生成式 AI 的强大能力也可能成为恶意攻击者手中的武器.例如部分不法分子通过深度伪造技术，伪造人脸冒充公众人物发言，实现对舆论的操控，破坏社会信任和公共安全.深度伪造检测(Deepfake Detection)等技术的出现一定程度上缓解了 AI 换脸的危害.

除深度伪造领域之外，近年来伴随着大模型技术的发展，图像生成领域中的内容安全问题得到广泛关注，例如扩散模型微调攻击(Diffusion Finetuning Attack).恶意攻击者可以利用高效的算法微调预训练的扩散模型，使其能够从社交网络中非法提取信息.基于窃取的信息，攻击者可以合成未经授权的图像，例如伪造指定人物制作色情图像，典型微调算法包括 LoRA[89],DreamBooth[90] 等.为防止图像生成的滥用，研究者尝试通过在图像分享之前添加不可察觉的保护噪声，使得在这些图像上微调 DFA 难以有效地模仿它们，代表工作有 Glaze[91]、AdvDM[92]和 Pretender[93].

大模型数据隐私 大模型技术是人工智能领域的重要研究分支，其强大的能力为构建通用人工智能体奠定基础.但是，大模型的技术特性也引发了全新的隐私挑战.

一是大模型记忆性加剧了数据泄露风险.大模型的训练过程中包含大量的用户隐私数据，恶意攻击者可以通过模型的输出或中间信息开展逆向工程，还原原始文本，并从中提取敏感信息.大模型的记忆性导致数据隐式地嵌入模型之中，难以通过传统方法进行遗忘，加剧了隐私泄露风险.针对大语言模型的隐私泄露攻击主要包括嵌入向量反转攻击(Embedding Inversion Attack)[94-95]，成员推理攻击(Membership Inference Attack)[96]和属性推理攻击(Attribute Inference Attack)[97-98]. Liu[99]等人指出大语言模型中的嵌入向量(Embedding)存储了丰富的文本数据表示，捕捉了语义和句法属性,需要妥善保护以防止反转攻击.

二是数据多元化对隐私防护机制提出新挑战.出于商业价值和数据隐私的考量，大部分领先的大模型仅提供 API 接口，并不公布模型具体架构和参数.因此，模型架构和模型数据可被认为是一种新形式的商业数据资产.现有研究表明，仅通过 API 接口，攻击者可以通过查询系统并从查询结果中学习，将部署的机密模型的参数或功能复制到提取的替代模型中，即模型提取攻击(Model Extraction Attack)[100-101].

三是数据动态性引发的长期风险.持续学习是人工智能中智能实现的关键技能之一.大模型在持续学习的过程中吸纳大量数据，模型能力也在发生动态变化.已有研究表明，通过精心设计的数据或提示词，可以诱使模型绕过安全对齐机制，输出训练数据中的敏感内容如用户的身份信息、联系方式，典型攻击方式包括越狱攻击[102-103]、提示词注入攻击[104-105].如何平衡大模型的记忆与遗忘，不仅是实现大模型训练降本增效的需求，同时也是缓解大模型数据泄露风险的重要保证.

表 3 数据要素流通全阶段隐私关键技术对比（对比范围涵盖适用场景、优势、劣势以及当前工作面临的主要挑战）

流通阶段	技术类型		适用场景	优势	劣势	主要挑战
数据流通前	数据最小化		数据收集阶段，如涉及敏感数据、合规性要求高的医疗、金融领域	有效降低数据泄露风险，减少数据存储成本，符合 GDPR 等隐私保护法规的数据采集规范	限制数据后续分析价值，依赖对数据需求的精准预判；法规到技术实践存在一定鸿沟	缺乏统一的数据最小化形式定义；难以平衡数据效用与最小化，动态业务中需求预测困难
	数字水印		版权追踪、数据可信溯源	不破坏数据可用性，隐蔽性强，支持事后追责	水印可能被去除或破坏，鲁棒性高度依赖算法设计	鲁棒的水印算法设计，避免水印嵌入引入额外的信息泄露风险；欠缺对数据集版权的关注
数据流通	基于密码学	同态加密	数据流通中的计算阶段，如云端数据分析	支持在密文上进行数学计算，实现端到端隐私保护	计算开销大，仅支持有限运算，密钥管理复杂	性能优化与算法可扩展性
		差分隐私	数据发布与共享，如银行统计报表生成	提供可证明的隐私保证，抵御背景知识攻击	噪声添加降低数据精度，影响模型性能，需合理规划隐私预算.	更好地权衡隐私预算与数据效用；差分隐私对其他攻击方式的防御缺乏理论与实验验证
		安全多方计算	多方数据联合计算，如联合风控、精准营销	数据无需出域，支持隐私保护下的协同计算	通信开销高，计算效率低，协议设计复杂	抵御合谋攻击，优化大规模场景下的协议性能
		零知识证明	数据验证场景，如身份认证、交易合规性证明	验证过程不泄露原始信息，高效轻量	证明生成过程的计算复杂，适用场景有限	设计更加通用的零知识证明协议
	基于合作	传统联邦	分布式模型训练，如跨医院医疗模型训练	数据不出本地，降低集中泄露风险	通信成本高，模型易受推理攻击，异构数据兼容性差	防御模型反演、成员推断攻击，优化跨域数据对齐

	学习	个性化联邦	用户需求差异大的分布式训练场景，如个性化推荐	兼顾全局模型与本地特征，提升模型在异质数据上的性能，更加贴合现实场景	端云协调复杂度更高，隐私泄露风险分散化	防止本地个性化参数泄露用户数据分布；防止恶意用户在个性化训练阶段植入后门
	区块链		数据存证与溯源，如供应链审计	数据不可篡改，去中心化透明监管	存储成本高，吞吐量低，链上数据公开导致隐私风险	在不侵犯用户隐私的前提下进行数据跨链传输；解决海量数据上链交易的效率瓶颈
数据流通后	忘却学习	集中式忘却学习	中心化数据中心的数据删除需求	可以直接操作模型参数，删除效率高	残留信息难以彻底清除，可能影响模型性能；模型重训练可能带来高昂代价	量化数据擦除效果；减少忘却学习带来的性能损失；防御基于残留信息的重建攻击
		联邦忘却学习	分布式环境下的数据删除，如多机构协作场景	无需集中数据，符合数据本地化要求	协调成本高，删除一致性难保证	解决联邦忘却学习中的性能退化问题

6 总结

伴随着 AIGC 研究浪潮的兴起，用户数据被大规模收集、存储和处理，数据市场日新月异.如何在数据要素流通过程中妥善保护用户数据隐私成为监管者和数据服务提供商密切关心的问题.不同于传统隐私计算类综述，聚焦于数据流通过程单个环节的隐私问题，本文依据数据全生命周期和技术兼容性，将数据流通全过程分为三个阶段，内容全面涵盖各阶段的传统与新兴隐私研究领域，总结各领域的最新研究进展与未来挑战.表 3 中对三个阶段中的隐私关键技术的适用场景、优缺点以及主要挑战进行全面总结与横向对比.

本文首先对数据流通全流程中存在的隐私风险进行分阶段阐述，初步介绍应对各类隐私风险的前沿技术研究.数据流通前主要包含数据收集和预处理的过程.在这一部分，本文一方面对数据最小化原则这一新兴研究领域进行了系统分析，并对这一法规在具体实施过程的难点进行了深入分析，总结目前研究热点与进展；另一方面，针对数据水印技术，本文依据水印生成算法的技术基础，将其进一步细分为传统水印，模型无关水印和模型相关水印，并主要对近年来基于深度学习的水印研究成果进行梳理总结.数据流通环节，本文从密码学、合作学习、区块链等视角出发，对各视角下的研究子领域进行梳理，涵盖个性化联邦、区块链数据流通安全等多项热点话题.数据流通后这一环节最易被各方忽视，本文以忘却学习为切入点，对传统的集中式场景和分布式情形进行单独剖析，建立全新的分类体系，总结研究痛点，梳理最新成果.最后，结合当今人工智能技术的飞速发展，本文从数据流通的角度剖析人工智能时代下全新的数据隐私挑战,包括生成式 AI 安全与大模型数据隐私.

现有的相关综述对于数据流通过程中的隐私技术介绍相对孤立，易导致多种技术间的不兼容，造成不必要的人力物力浪费.从隐私计算发展角度而言，既需要研究者在隐私计算子领域深入挖掘，培养专业性人才，也需要研究者从数据流通的角度，总揽全局.据我们所知，本工作是首个以数据流通为载体，隐私技术为对象的综述类论文，内容涵盖数据最小化原则、个性化联邦学习、忘却学习等多项新兴研究领域.通过本文，研究者可以更加全面地认识数据流通全流程，建立系统的技术分类框架，从而为后续建立完备的全流程数据流通隐私保护范式奠定基础.

References:

[1] Li F H, Li H, Jia Y, et al. Privacy computing: concept, connotation and its research trend. Journal on Communications, 2016, 37(4): 1-11. [doi:10.11959/j.issn.1000-436x.2016078.]

[2] Guo Z J，Li M L，Zhou Y M, et al. Survey on digital watermarking technology for artificial intelligence generated content models. Journal of Cybersecurity, 2024, 2(1): 13-39.[doi: 10.20172/j.issn.2097-3136.240102]

[3] Liu A, Pan L, Lu Y, et al. A survey of text watermarking in the era of large language models. ACM Computing Surveys,2024, 57(2): 1-36. [doi: 10.1145/3691626]

[4] Huo W, Yu Y, Yang K, et al. Privacy-preserving cryptographic algorithms and protocols: a survey on designs and applications. Sci Sin Inform, 2023, 53: 1688–1733. [doi: 10.1360/SSI-2022-0434]

[5] Mansheej P, Ganguli S, Dziugaite G K. Deep learning on a data diet: Finding important examples early in training. Advances in neural information processing systems, 2021, 34: 20596-20607.

[6] Shanmugam D, Diaz F, Shabanian S, et al. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization.In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022,839–849.

[7] Ganesh P, Tran C, Shokri R, et al. The data minimization principle in machine learning . 2024. arXiv:2405.19471.

[8] Staab R, Jovanovic N, Balunovic M, et al. From principle to practice: Vertical data minimization for machine learning. In :Proceedings of the 2024 IEEE Symposium on Security and Privacy(SP), 2024, 4733-4752.

[9] Biega A J, Potash P, Daumé H, et al. Operationalizing the legal principle of data minimization for personalization.In :Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. 2020: 399-408.

[10] Galdon Clavell G, Martín Zamorano M, Castillo C, et al. Auditing algorithms: On lessons learned and the risks of data minimization. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020: 265-271.

[11] Rastegarpanah B, Gummadi K, Crovella M. Auditing black-box prediction models for data minimization compliance. Advances in Neural Information Processing Systems, 2021,34:20621–20632.

[12] Rastegarpanah B, Crovella M, and Gummadi K P. Fair inputs and fair outputs: The incompatibility of fairness in privacy and accuracy. In: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, 260–267.

[13] Sato R, Takezawa Y, Bao H, et al. Embarrassingly Simple Text Watermarks.2023. arXiv:2310.08920.

[14] Munyer T and Zhong X. Deeptextmark: Deep learning based text watermarking for detection of large language model generated text. 2023. arXiv:2305.05773.

[15] Atallah M, Raskin V, Crogan M, et al. Natural language watermarking: Design, analysis, and a proof-of-concept implementation.In: Information Hiding: 4th International Workshop, 2001, Springer,185–200.

[16] Abdelnabi S and Fritz M. Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding.In: Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP), 2021, 121-140.

[17] Lau G, Niu X Y, Dao H, et al.Waterfall: Framework for Robust and Scalable Text Watermarking. 2024. arXiv:2407.04411.

[18] Adi Y , Baum C , Cisse M, et al. Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring. In 27th USENIX Security Symposium,2018,1615-1631.

[19] Tang R, Feng Q, Liu N, et al. Did you train on my dataset? towards public dataset protection with cleanlabel backdoor watermarking. ACM SIGKDD Explorations Newsletter, 2023, 25(1): 43-53.

[20] Kirchenbauer J, Geiping J, Wen Y, et al. A Watermark for Large Language Models. In: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023,202: 17061-17084.

[21] Liu A W, Pan L Y, Hu X M, et al. An Unforgeable Publicly Verifiable Watermark for Large Language Models. 2023. arXiv:2307.16230.

[22] Gentry C. Fully homomorphic encryption using ideal lattices. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, 2009. 169–178.

[23] van D M, Gentry C, Halevi S, et al. Fully homomorphic encryption over the integers. In: Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2010. 24–43.

[24] Brakerski Z, Vaikuntanathan V. Efficient fully homomorphic encryption from (standard) LWE. SIAM Journal on computing, 2014, 43(2): 831-871.

[25] Fan J F, Vercauteren F. Somewhat practical fully homomorphic encryption.2012. https://eprint. iacr.org/ 2012/144.pdf.

[26] Gentry C, Sahai A, Waters B. Homomorphic encryption from learning with errors: conceptually-simpler, asymptotically-faster, attribute-based.In: Proceedings of Annual Cryptology Conference, 2013. 75–92.

[27] Chillotti I, Gama N, Georgieva M, et al. TFHE: fast fully homomorphic encryption over the torus. Journal of Cryptology, 2020, 33(1): 34-91.

[28] Cheon J H, Kim A, Kim M, et al. Homomorphic encryption for arithmetic of approximate numbers. In: Proceedings of International Conference on Advances in Cryptology, 2017. 409–437.

[29] Myers S and Shull A. Practical revocation and key rotation. The Cryptographers' Track at the RSA Conference, 2018.

[30] LI B Y and Micciancio D. On the security of homomorphic encryption on approximate numbers. The 40th Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2021.

[31] Dai Y R, Zhang J, Xiang B, et al. Overview on the Research Status and Development Route of Fully Homomorphic Encryption Technology. Journal of Electronics & Information Technology, 2024, 46:1775-1783.

[32] Dwork C, Kenthapadi K, McSherry F, et al. Our data, ourselves: Privacy via distributed noise generation.24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, 2006: 486-503.

[33] Dwork C. Differential privacy.International colloquium on automata, languages, and programming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 1-12.

[34] Kasiviswanathan S P, Lee H K, Nissim K, et al. What can we learn privately? SIAM Journal on Computing,2011,40(3):793-825.

[35] Bittau A, Erlingsson Ú, Maniatis P, et al. Prochlo: Strong Privacy for Analytics in the Crowd. In: Proceedings of the 26th Symposium on Operating Systems Principle. NewYork: ACM, 2017; 441-458.

[36] Mironov I. Renyi differential privacy. In : Proceedings of the IEEE 30th Computer Security Foundations Symposium, 2017:263–275.

[37] Jiang Y, Luo X, Wu Y, et al. Protecting label distribution in cross-silo federated learning. In: Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). IEEE, 2024: 4828-4847.[doi: 10.1109/SP54263.2024.00113]

[38] Renyi A. On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1961,547–562.

[39] Tan Q, Li Q, Zhao Y, et al. Defending against data reconstruction attacks in federated learning: An information theory approach. In 33rd USENIX Security Symposium. 2024: 325-342.

[40] Yao A. Protocols for secure computations.In: Proceedings of the 23rd Annual Symposium on Foundations of Computer Science, 1982; 160–164.

[41] Ben-Or M, Goldwasser S, Wigderson A. Completeness theorems for non-cryptographic fault-tolerant distributed computation.In: Proceedings of the 20th Annual ACM Symposium on Theory of Computing, 1988; 1–10.

[42] Goldreich O, Micali S, Wigderson A. How to play any mental game or a completeness theorem for protocols with honest majority.In: Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987; 218–229.

[43] Yao A. How to generate and exchange secrets.In: Proceedings of 27th Annual Symposium on Foundations of Computer Science, 1986;162–167.

[44] Lindell Y. Fast cut-and-choose based protocols for malicious and covert adversaries. In: Proceedings of the Advances in Cryptology-CRYPTO, Springer, 2013,1–17.

[45] Beaver D, Micali S, Rogaway P. The round complexity of secure protocols.In: Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, 1990; 503–513.

[46] Keller M, Pastro V, Rotaru D. Overdrive: making SPDZ great again.In: Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2018; 158–189.

[47] Fiat A, Shamir A. How to prove yourself: practical solutions to identification and signature problems.In: Proceedings of Annual Cryptology Conference, 1987. 186–194.

[48] Bitansky N, Canetti R, Chiesa A, et al. From extractable collision resistance to succinct non-interactive arguments of knowledge, and back again.In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012.

[49] Konecný J, McMahan H B, Ramage D, et al. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. 2016. arXiv:1610.02527.

[50] Konecný J, McMahan H, Felix X Y, et al. Federated Learning: Strategies for Improving Communication Efficiency.2016.arxiv: 1610.05492.

[51] McMahan H B, Moore E, Ramage D, et al. Federated learning of deep networks using model averaging. 2016, arXiv:1602.05629.

[52] McMahan H B, Moore E, Ramage D, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data.In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017.1-5.

[53] Zhu L, Liu Z, Han S. Deep leakage from gradients. Advances in neural information processing systems, 2019, 32.

[54] Geiping J, Bauermeister H, Dröge H, et al. Inverting gradients-how easy is it to break privacy in federated learning?. Advances in neural information processing systems, 2020, 33: 16937-16947.

[55] Tan Y, Long G, Liu L, et al. Fedproto: Federated prototype learning across heterogeneous clients. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022;1-13.

[56] Hu R, Guo YX, Li HN, et al. Personalized Federated Learning With Differential Privacy.IEEE Internet of Things Journal, 2020, 7: 9530-9539.

[57] Zhang C, Li S, Xia J, et al. BatchCrypt: Efficient homomorphic encryption for Cross-Silo federated learning.2020 USENIX annual technical conference (USENIX ATC 20). 2020: 493-506.

[58] Tan A Z, Yu H, Cui L, et al. Towards Personalized Federated Learning. IEEE Transactions on Neural Networks and Learning Systems, 2023,34: 9587-9603.

[59] Ye M, Fang X, Du B, et al. Heterogeneous federated learning: State-of-the-art and research challenges. ACM Computing Surveys, 2023, 56(3): 1-44.

[60] Collins L, Hassani H, Mokhtari A, et al. Exploiting shared representations for personalized federated learning.International Conference on Machine Learning,2021.

[61] Liang PP, Liu T, Liu Z Y, et al. Think locally, act globally: Federated learning with local and global representations.2020. arXiv:2001.01523.

[62] Wu C, Wu F, Lyu L, et al. Communication -efficient federated learning via knowledge distillation. Nature communications, 2022, 13(1): 2032.

[63] Arivazhagan MG, Aggarwal V, Singh AK, et al. Federated learning with personalization layers. 2019. arXiv:1912.00818.

[64] Bagdasaryan E, Veit A, Hua YQ, et al. How to backdoor federated learning. In the International conference on artificial intelligence and statistics. PMLR, 2020: 2938-2948.

[65] Fu C, Zhang XH, Ji SL, et al. Label inference attacks against vertical federated learning. In the 31st USENIX security symposium, 2022: 1397-1414.

[66] Arazzi M, Conti M, Nocera A, et al. Turning Privacy-preserving Mechanisms against Federated Learning.In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23). 2023. 1482–1495.

[67] Fu CH, Chen HB, Ruan N. Privacy for Free: Spy Attack in Vertical Federated Learning by Both Active and Passive Parties. IEEE Transactions on Information Forensics & Security, 2025.

[68] Mo F, Haddadi H, Katevas K, et al. PPFL: Privacy-preserving federated learning with trusted execution environments.In: Proceedings of the 19th annual international conference on mobile systems, applications, and services. 2021,94-108.

[69] Amiri M J, Agrawal D, El Abbadi A. Parblockchain: Leveraging transaction parallelism in permissioned blockchain systems.IEEE 39th International Conference on Distributed Computing Systems. 2019: 1337-1347.

[70] Androulaki E, Barger A, Bortnikov V, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains.13th EuroSys conference. 2018: 1-15.

[71] Dang H, Dinh TTA, Loghin D, et al. Towards Scaling Blockchain Systems via Sharding.the 2019 international conference on management of data. 2019: 123-140.

[72] Amiri MJ, Agrawal D, El Abbadi A. Sharper: Sharding permissioned blockchains over network clusters.the 2021 international conference on management of data. 2021: 76-88.

[73] Singh A, Click K, Parizi RM, et al. Sidechain technologies in blockchain networks: An examination and state-of-the-art review. Journal of Network and Computer Applications. 2020, 149: 102471.

[74] Dilley J, Poelstra A, Wilkins J, et al. Strong federations: An interoperable blockchain solution to centralized third-party risks. 2016. arXiv:1612. 05491.

[75] Adnan M, Kalra S, Cresswell J C, et al. Federated learning and differential privacy for medical image analysis. Scientific reports, 2022, 12(1): 1953.

[76] Kim H, Park J, Bennis M, et al. Blockchained on-device federated learning. IEEE Communications Letters, 2019,24(6): 1279-1283.

[77] Wang WQ, Tian ZY, Zhang CH. Machine Unlearning: A Comprehensive Survey. 2024. arXiv: 2405.07406.

[78] Bourtoule L, Chandrasekaran V, Christopher A, et al. Machine unlearning. In 2021 IEEE Symposium on Security and Privacy(SP), 2021;141–159.

[79] Cao Y Z and Yang J F. Towards making systems forget with machine unlearning. In 2015 IEEE Symposium on Security and Privacy(SP), 2015;463–480.

[80] Guo C A, Goldstein T, Hannun A, et al. Certified data removal from machine learning models. In: Proceedings of the 37th International Conference on Machine Learning, ICML, 2020;119:3832–3842.

[81] Sekhari A, Acharya J, Kamath G, et al. Remember what you want to forget: Algorithms for machine unlearning. Advances in Neural Information Processing Systems, 2021, 34: 18075-18086.

[82] Wu C, Zhu S C, Mitra P. Federated unlearning with knowledge distillation. 2022. arXiv:2201.09441.

[83] Liu Y, Xu L, Yuan X L, et al. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, 2022.1749–1758.

[84] Wang J X, Guo S, Xie X, et al. Federated unlearning via class-discriminative pruning. In: Proceedings of the ACM Web Conference, 2022;622–632.

[85] Chen M, Zhang Z K, Wang T H, et al. When machine unlearning jeopardizes privacy. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security(CCS), 2021; 896–911.

[86] Zanella-Béguelin S, Wutschitz L, Tople S, et al. Analyzing information leakage of updates to natural language models.In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020.363–375.

[87] Sun Z K, Ruan N, Li J H. DDL: Effective and Comprehensible Interpretation Framework for Diverse Deepfake Detectors, IEEE Transactions on Information Forensics & Security, 2025.

[88] Yisroel M, Wenke L. The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys, 2021, 54(1): 1-41.

[89] Hu EJ, Shen YL, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models.2021. arXiv:2106.09685.

[90] Ruiz N, Li YZ, Jampani V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject driven generation. In IEEE/CVF Conference on Com puter Vision and Pattern Recognition, 2023，22500-22510.

[91] Shan S, Cryan J, Wenger E, et al. Glaze: Protecting artists from style mimicry by text-to-image models. USENIX Security Symposium, 2023.

[92] Liang C M, Wu X Y, Hua Y, et al. Adversarial example does good: Pre venting painting imitation from diffusion models via adversarial examples. In International Conference on Machine Learning, 2023.

[93] Sun Z K, Liu Z J, Ji S L, et al. Pretender: Universal Active Defense against Diffusion Finetuning Attacks, The 34th USENIX Security Symposium, 2025.

[94] Hayet I, Yao ZJ, and Luo B. Inver-net: An inversion attack framework to infer fine-tuning datasets through word embeddings. In Findings of the Association for Computational Linguistics: EMNLP 2022. 5009–5018.

[95] Li HR, Xu MS, and Song YQ. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL),2023,14022–14040.

[96] Shokri R, Stronati M, Song CZ, et al. Membership inference attacks against machine learning models. In IEEE symposium on security and privacy (SP), 2017.3-18.

[97] Melis L, Song C Z, Cristofaro ED, et al. Exploiting unintended feature leakage in collaborative learning. In IEEE symposium on security and privacy (SP), 2019. 691–706.

[98] Wang Y, Zhao YY, Dong Y S, et al. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data.

[99] Liu T, Yao HW, Wu T, et al. Mitigating Privacy Risks in LLM Embeddings from Embedding Inversion, 2024.arXiv:2411.05034.
[100] Jagielski M, Carlini N, Berthelot D, et al. High accuracy and high fidelity extraction of neural networks. In 29th USENIX security symposium. 2020.
[101] Tang MX, Dai A N, Louis D, et al. MODELGUARD: Information-Theoretic Defense Against Model Extraction Attacks. In 33th USENIX security symposium. 2024.
[102] Yong Z X, Menghini C, and Bach S H. Low-resource languages jailbreak gpt-4, in NeurIPS Workshop, 2023.
[103] Wei A, Haghtalab N, and Steinhardt J, Jailbroken: How does llm safety training fail? NeurIPS, 2024.
[104] Perez F and Ribeiro I. Ignore previous prompt: Attack techniques for language models, in NeurIPS Workshop, 2022.
[105] Liu Y, Jia Y, Geng R, et al. Formalizing and benchmarking prompt injection attacks and defenses.In 33th USENIX Security, 2024.1831–1847.

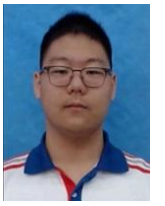
附中文参考文献:

[1] 李凤华,李晖, 贾焰等. 隐私计算研究范畴及发展趋势.通信学报. 2016, 37(4): 1-11. doi:10.11959/j.issn.1000-436x.2016078.
[4] 陈翔,霍炜,郁昱等.隐私保护计算密码技术研究进展与应用.中国科学:信息科学,2023,53:1688–1733,doi: 10.1360/SSI-2022-0434.
[31] 戴怡然, 张江, 向斌武, 邓燚. 全同态加密技术的研究现状及发展路线综述. 电子与信息学报, 2024, 46(5): 1774-1789. doi: 10.11999/JEIT230703.

刘立伟(2002－),男,硕士生,主要研究领域为数据隐私，个性化联邦学习，AI 安全.

傅超豪(1996－),男,博士生,主要研究领域为忘却学习、联邦学习、AI 安全.

孙泽堃(1997－),男,博士生,主要研究领域为计算机视觉、AI 安全、数据隐私.



周耘(2003－),男,科研助理,主要研究领域为区块链、AI 安全.

阮娜(1986－),女，博士生导师，副教授，计算机学会（CCF）杰出会员,主要研究领域为数据隐私、区块链、AI 安全.

蒋昌俊(1962－),男，中国工程院院士,讲席教授，计算机学会（CCF）会士,主要研究领域为网络计算技术、网络交易风控.