Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

面向数据权利、数据定价和隐私计算的数据驱动学习

徐基珉[a]，洪暖欣[a]，许哲宁[a]，赵洲[a]，吴超[a]，况琨[a,*]，王嘉平[b]，朱明杰[c]，周靖人[d]，任奎[a]，杨小虎[a]，卢策吾[e]，裴健[f]，沈向洋[b]

[a] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[b] International Digital Economy Academy, Shenzhen 518045, China
[c] Craiditx, Shanghai 200050, China
[d] Antgroup, Hangzhou 310023, China
[e] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[f] School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

ARTICLE INFO

Article history:
Received 11 January 2022
Revised 17 October 2022
Accepted 25 December 2022
Available online 9 February 2023

关键词
数据科学
人工智能
数据权利
数据定价
隐私计算

摘要

近年来，数据已成为数字经济中最重要的生产要素之一。与传统生产要素不同，数据的数字化性质使其难以签约和交易。因此，建立一个高效和标准的数据交易市场体系将有利于降低成本，提高行业各方的生产力。尽管许多研究致力于数据法规和其他数据交易问题，如隐私和定价，但对机器学习和数据科学领域的这些研究进行全面回顾的工作很少。为了提供对这个主题的完整和最新的理解，本文涵盖了数据交易过程中的三个关键问题：数据权利、数据定价和隐私计算。通过厘清这些主题之间的关系，本文提供了一个数据生态系统的全貌，其中数据由个人、研究机构和政府等数据主体生成，而数据处理者出于创新或运营目的获取数据，并通过适当的定价机制根据数据主体各自的所有权分配收益。为了使人工智能（AI）能够长期有益于人类社会的发展，人工智能算法需要通过数据保护法规（即隐私保护法规）进行评估，以帮助构建日常生活中值得信赖的人工智能系统。

1. 引言

近年来，互联网、大数据、云计算、人工智能（AI）等技术加速创新，并日益融入经济和社会发展各个领域的完整过程。数字经济发展速度之快、辐射范围之广、影响之深前所未有。作为数字经济中的一种新的生产要素，数据已经积累了巨大的数量，并包含了大量的经济价值。随之而来的是，机器学习等数据驱动方法已被广泛应用于许多领域，包括化学反应预测[1]、蛋白质结构预测[2]和科学计算[3]等。因此，建立一个高效、标准的数据交易市场体系将有利于挖掘数字经济中的新型生产要素所蕴含的价值。最近，Pei [4]发表了一篇从经济学原理出发，将数字产品定价和数据产品定价联系起来的综述，重点讨论了数据定价和数字产品定价的基本经济学和数学原理。Cong等[5]的另一篇综述侧重于介绍机器学习中的数据定价流程，并涵盖了数据标签定价的研究内容。与这些现有的综述不同，本文从数字经济中建立数据交易市场体系所面临的三个关键问题出发，对数据权利、数据定价和隐私

* Corresponding author.
E-mail address: kunkuang@zju.edu.cn (K. Kuang).

英文原文：Engineering 2023, 25(6): 66–76
引用本文：Jimin Xu, Nuanxin Hong, Zhening Xu, Zhou Zhao, Chao Wu, Kun Kuang, Jiaping Wang, Mingjie Zhu, Jingren Zhou, Kui Ren, Xiaohu Yang, Cewu Lu, Jian Pei, Harry Shum. Data-Driven Learning for Data Rights, Data Pricing, and Privacy Computing. *Engineering*, https://doi.org/10.1016/j.eng.2022.12.008

计算进行了展开讨论，并将其作为数据要素计算框架中的要素加以整合。

数据权利包括权利主体和权利内容，是数据交易的前提，由相关法律法规规定和保障。近来，越来越多的国家和地区开始关注大数据立法。例如，欧盟（EU）发布了《通用数据保护条例》（GDPR），中国发布了《个人信息保护法》（PIPL）。随之而来的，是对保障法律规定的数据权利所需的技术解决方案的迫切需求。此外，数据定价和隐私计算技术在数据交易过程中至关重要。与传统的商品交易不同，数据的特殊性要求制定相应的定价策略和保护数据隐私的技术解决方案。

数据定价和隐私计算在数据交易过程中相辅相成。本文，我们将通过三种典型的数据交易场景介绍数据定价和隐私计算的技术解决方案。第一种数据交易场景由单个数据所有者和多个数据购买者组成。在这种场景下，不同的客户通常从某个数据公司，如Twitter、Bloomberg或Pistachio等，购买数据集，以便访问所需的数据。在此数据交易场景中，根据不同的数据购买者的需求，需要客制化多种定价策略和隐私要求。第二种数据交易场景由多个数据所有者和一个数据购买者组成。在这种场景下，为了利用存储在不同数据所有者之间的数据，通常需要构建可信的隐私计算方法来实现模型的分布式训练，以及公平的数据定价方法来确保来自不同数据所有者的贡献的激励机制。第三种数据交易场景由多个数据所有者和多个数据购买者组成。在这种场景下，通常需要数据中介的参与，来为数据所有者设计合理和公平的补偿函数，为数据购买者设计无套利价格函数，以实现收入最大化的目标。在这种多方数据交易过程中，数据所有者、数据中介和数据购买者之间必须满足多种隐私要求。

在接下来的第2节中，我们将讨论随着数字经济活动的不断增加而出现的数据权利问题，如数据所有权和数据隐私权等。对数据权利问题的广泛关注最终迎来了GDPR等大数据法律法规的发布。目前，数据是否应该受到严格监管的问题仍在激烈争论中。符合现有大数据法律法规的新技术正成为行业的新焦点。在这一节中，我们将根据上述主题概述数据权利，并介绍这些领域的一些潜在解决方案。

在第3节中，我们将讨论最近提出的数据定价技术解决方案。随着移动终端设备的普及，越来越多的端到端个人信息或个人数据产生并被赋予某些属性。数据处理者可以使用这些数据来训练模型，并从中获得商业利益。作为数据资产的所有者，个人应因其数据被使用而获得补偿。在这一节中，我们将从三种典型数据交易场景出发，对数据定价的技术方案进行概述和介绍，包括三条技术路线：基于查询的定价、基于模型的Shapley值定价和基于数据市场的定价。

在第4节中，我们将讨论隐私计算技术，这是一系列加密计算技术的组合。当数据被访问时，数据处理者可能通过某些方法反向获取数据中的敏感信息，从而导致数据主体的敏感信息被披露和滥用。为了防止此问题的发生并确保数据被合法使用，必须采取某些技术措施来保护数据隐私和安全。隐私计算通过在数据交易过程中保护数据的敏感信息，在数据要素和数据价值之间架起了一座桥梁。在这一节中，我们将从三种典型的数据交易场景出发，对隐私计算的技术方案进行概述和介绍，包括三条技术路线：加密技术、可信执行环境（TEE）和协作学习。

在本文中，我们将数据权利、数据定价和隐私计算整合到数据要素计算框架中。如图1所示，数据权利、数据定价和隐私计算是数据要素计算框架中的相关技术。在不同的行业中，数据由不同参与者（如个人、商业平台、政府机构）的行为生成，数据要素计算的第一步需要确定数据相关权利，如数据使用权、数据所有权和数据隐私权；在确定数据权利后，需要对数据的价值进行评估，并根据数据财产所有权的归属分配收入；最后，有必要在数据交易过程中添加必要的隐私保护，以防止私人信息的泄露或恶意窃取。基于这个数据要素计算框架，我们将介绍三个主要的数据交易问题：数据权利、数据定价和隐私计算。我们还为未来可能引起关注的研究挑战提供了可行的建议和分析。

2. 数据权利

工业数据的产生和交换在如今的数字经济中发挥着关键作用；一些研究甚至表明，工业数据正在取代石油成为最有价值的资源[6]。国际商业机器公司（IBM）估计，每天有2.5万亿字节的数据被产生[7]。随着数据传输、分析等的新兴技术的发展，数据的产生和交换速度变得越来越快。数据的高交易量正引起人们对数据内在的高外部性成本的关注，因为对于数据而言，事前协商的交易方式是不可行的[8–9]。因此，数据权利作为产权的一种形式正在兴起，因为数据权利主体内部化数据的外部性比事前协商更有利可图[10]。数据公司提供基于收集到的用户数据的智能推荐服务，但某些推荐的内容可能存在偏见、误导性和有利于服务提供商的暗示内容。因此，数据权利主体容易受到所谓的数据资本主义的影响，并依赖于这种数据资本主义[11]。因此，学界正在寻求对数据的最终裁决，

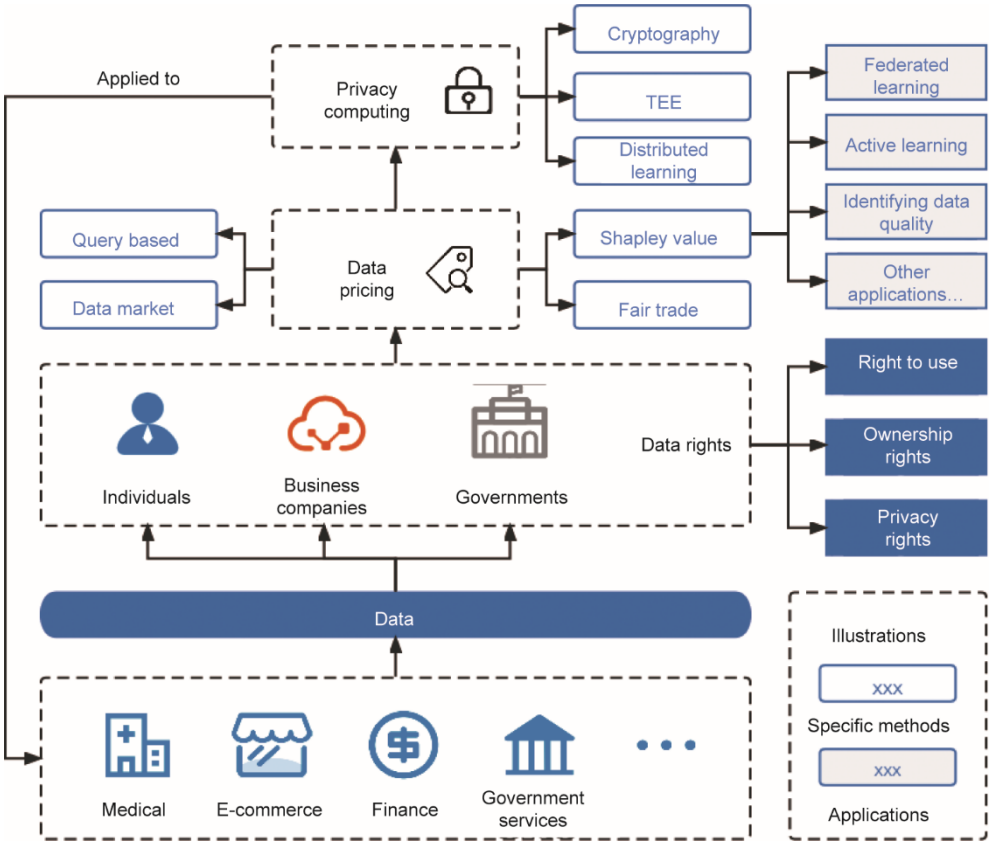

图1. 数据要素计算。

如数据所有权等相应权利[12]。

数据权利表示数据主体对所述类型信息的所有权和控制权，如图2所示。我们将数据权利分为三类：个人数据权利、商业数据权利和政府数据权利[13]。

GDPR于2018年5月在欧盟所有国家生效。很快，欧盟以外的许多国家都进行了类似的立法工作，如中国的PIPL。GDPR的目的是确保“合法、公平和透明的方式”处理个人数据，并确保数据主体获得知情权、访问权、纠正权、删除权、反对权和自动化个人决策权[14]。

为了遵守GDPR，研究人员提出了一些解决方案，其中之一是基于区块链的系统的实现。对于具有集中式中心化客户端-服务器体系结构的服务提供商来说，几乎不可能确保他们持续遵守GDPR。而区块链技术是一个完美的解决方案，因为它具有去中心化、难以篡改和易于访问的

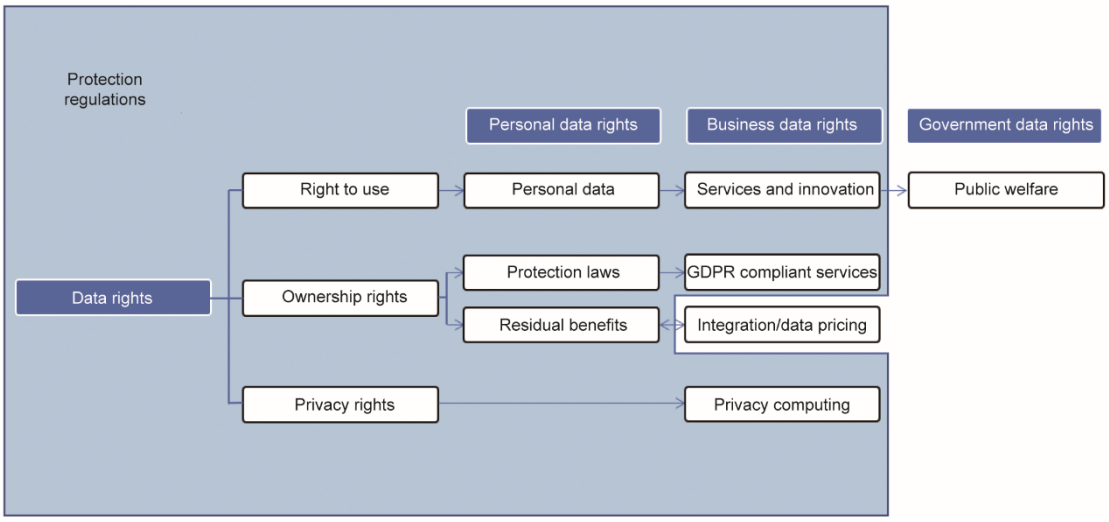

图2. 数据权利。

特性。Truong等[15]提供了一个基于Hyperledger Fabric许可的区块链框架构建的遵守GDPR的个人数据管理平台的示例。未来的主要挑战是实现相应的机制来解决缺乏可信的集中式资源服务器的问题，并在区块链网络上提供潜在的计算能力。

遵守GDPR的其他方法主要围绕满足GDPR要求所构建的详细框架[16–19]。除了对这些方法本身的研究外，迄今为止很少有关联数据权利这一主题的研究成果。联邦学习[20]是隐私限制下的机器学习中的一个热门话题[19]，但为了遵守GDPR所规定的数据权利，会面临一些新的挑战。例如，Ginart等[18]试图通过构建有效的数据删除模型解决其中的一个挑战，即数据被遗忘权。此外，大数据法律法规通常包含被称为“解释权”的个人权利。更具体地说，由于联邦学习的全局模型是本地模型的聚合，因此很难描述数据主体的数据贡献[19]。

随着GDPR和PIPL等数据法律法规的发布，许多规模较小的数据权利组织也开始采取行动。例如，英国特许营销协会（CIM）敦促各成员组织就客户数据的负责任管理问题采取行动。他们要求成员组织在管理客户数据时保持公开透明，向客户展示共享数据的好处，并尊重客户的数据。CIM声称，如果成员组织对其客户数据使用过程更加公开透明，67%的客户将愿意分享更多的个人信息[21]。开放的沟通和诚实的使用可以赢得客户的信任；对于其他情况，如谷歌（Google）和YouTube等大型数据公司向数十亿人提供廉价甚至免费的服务，以收集数据来改进算法。然而，如果对这种行为增加严格的监管，公司提供此类服务和创新的能力将大大降低。这种对数据开放访问的严重依赖促使人们更加关注数据访问权，而不是人们对其数据的专有产权[22]。

商业数据权利主要指知识产权和专利[13]。Atkinson[7]从商业角度讨论了与数据权利相关的许多挑战。他主张通过市场的力量来确保交易的公平性和健康的数据关系，只有当反竞争行为限制了创新或伤害了客户时，政府才应该采取行动。此外，他还主张各国政府应利用其易于获取整合数据的能力，并应将数据发布给公众，以便其他人将数据用于创新目的，最终提高公共福利总额。总之，Atkinson主张数据在默认情况下应保持开放，政府只应在必要时进行干预。

数字经济中存在的高外部性意味着整合数据权利是有益的[10,23]，通常由在合作中做出最大贡献的一方在整合中处于领导地位。基于这种数据权利的整合，可以通过事前协商的方式，为其他各方提供补偿。为了进一步降低协商成本，企业甚至可以选择用匹配博弈的相关方法（如最小核心或核仁）来近似收益分配[24]。

3. 数据定价

一个公平有效的数据交易市场可以引导数据要素的合理分配，从而促进各种资源要素的快速流动，加速各种市场主体的整合，帮助市场主体重构组织模式，实现跨界发展，打破时间和空间限制，延长产业链，畅通国家间经济循环。作为数据交易过程中的一个关键问题，数据定价将受到越来越多的关注。随着移动终端设备的普及，越来越多的端到端个人信息或个人数据被产生，并被赋予某些属性。数据处理者可以使用这些数据来训练模型，并从中获得商业利益。作为数据资产的所有者，个人应因其数据的使用而获得补偿。为了激励数据所有者提供高质量数据和数据处理者以挖掘更多信息，进而优化数字经济中数据要素的分配，各种数据产品的公平和有效定价策略至关重要。

目前，数据定价的研究和应用尚处于起步阶段；这里，我们回顾了基于三种典型数据交易场景的数据定价研究路线，如图3所示。

3.1. 单个数据所有者，多个数据购买者

在此场景中，一般是以公司作为数据交易中的数据所有者，收集数据并将其整理成数据库的形式进行数据交易。然后，多个客户直接从公司购买所需的数据。公司的数据定价策略必须满足客户的各种需求。在这种情况下，一般采用直接数据定价，即基于数据集本身的定价策略。这种定价策略通常由原始数据的固有因素决定，如数据质量、数据数量等。直接数据定价的一个典型技术路线是基于查询的数据定价，即基于所涉及的查询或数据项的数量的定价方式。直观地说，数据卖家可以将数据集的一个访问视角视为一个销售版本。不同访问视角的定价需满足无套利原则。Koutris等[25]将基于查询的定价问题转化为网络流问题，使得能从给定少量查询的价格计算出任何查询的价格。之后，他们采用了不同的方法，将基于查询的定价问题转化为优化问题中的整数线性规划，基于卖方指定的价格点对结构化查询语言（SQL）查询进行定价，并使用查询历史记录避免对重复数据的收费[26]。Deep等[27]提出了一种支持各种定价功能的实时定价系统，可以有效地计算大规模SQL查询的价格。

3.2. 多个数据所有者，单个数据购买者

在多个数据所有者和单个数据购买者的场景中，大量

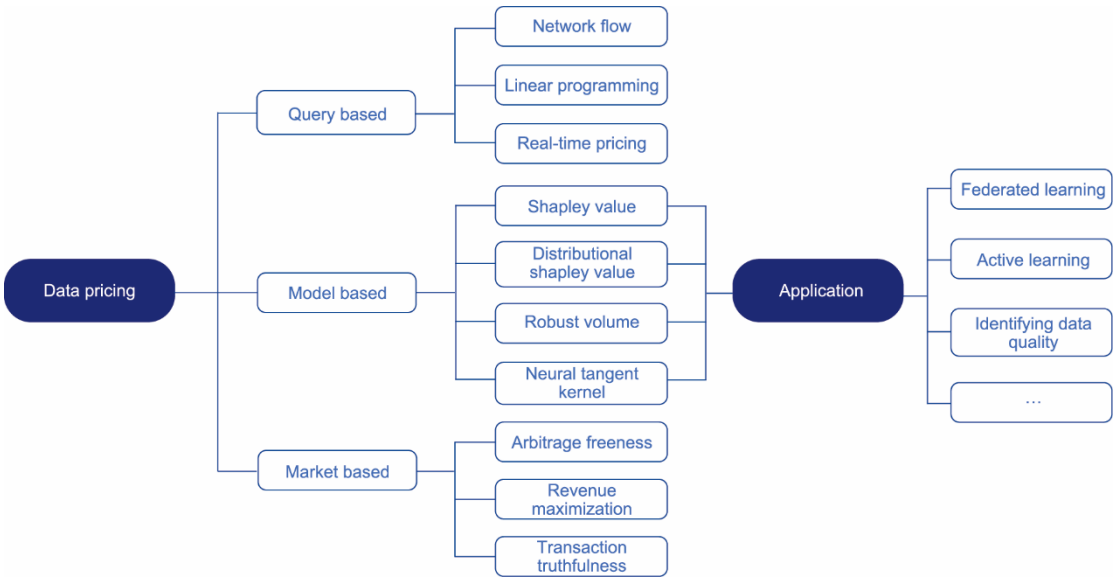

图3. 数据定价。

数据由个人生成并存储在其移动终端设备中。利用这些数据训练模型的数据处理者必须对数据所有者的数据使用进行补偿。数据处理者的定价策略应公平评估不同数据所有者对模型训练的贡献。在这种情况下，一般采用基于模型的数据定价的技术路线。基于模型的数据定价是一种基于通过数据集训练获得的模型的数据定价策略。这种定价策略通常由不同数据对模型训练的贡献决定。对于深度学习模型，单个数据通常对模型的训练没有直接的效用，因为深度学习模型必须从由大量数据组成的数据集中进行学习。也就是说，对于深度学习模型，通常很难直接衡量单条数据的贡献，单条数据的贡献只能与其他数据结合反映。因此，此类定价策略通常通过计算单条数据的边际贡献来确定该数据对模型训练的贡献。这种通过模型计算数据贡献的方法，在机器学习领域通常被称为“数据估值”。

数据估值可以通过多种技术来实现，如留一法[28]、杠杆或影响力分数[29]和强化学习[30]。Shapley值[31]是合作博弈论中的经典概念，得益于其深厚的理论背景，是一种比较典型的数据估值方法。在合作博弈论中，Shapley给出了公平收入分配的定义[31]。假设有k个代理一起合作参与奖金为v的游戏（其中，k代表参与游戏的代理数量，v代表游戏奖金）。我们将D记为k个代理组成的集合，将$V(S)$记为$S\subset D$的联盟收入（其中，S是一些代理组成的联盟），并将ϕ_i记为代理i的收入。为了将奖金公平分配给每个代理，应满足以下四个公理：

（1）**效率**。对于全集D，$\sum_{i\in D}\phi_i=V(D)$。换言之，所有代理的收入之和应等于游戏奖金。

（2）**对称**。对于任意子集$S\subset D-\{i,j\}$，如果对于代理i和j，存在关系$V(S\cup\{i\})=V(S\cup\{j\})$，那么$\phi_i=\phi_j$。换言之，如果代理$i$和$j$在与其他代理组成的每个联盟中总是提供相同的边际贡献，那么他们的收入应该相等。

（3）**零元**。对于任意子集$S\subset D-\{i\}$，如果$V(S\cup\{i\})=V(S)$，那么$\phi_i=0$。我们将这类代理称为“零元”。换言之，如果代理i没有在与其他代理组成的任何联盟中做出贡献，它就不应该得到任何收入。

（4）**可加**。对于由相同的代理集合D参与的两个不同的合作博弈游戏，这两个游戏的联盟收入分别记为V_1和V_2，那么对于代理i，存在关系$\phi_i(V_1+V_2)=\phi_i(V_1)+\phi_i(V_2)$。

Shapley值是唯一一个满足上述四个公理的奖金分配方法；它划分了全集D的总奖金v，并满足对称性、零元素和可加性公理。代理i的Shapley值由下式给出：

$$\phi_i(V)=\frac{1}{|D|}\sum_{S\subseteq D-\{i\}}\frac{V(S\cup\{i\})-V(S)}{\binom{|D|-1}{|S|}} \tag{1}$$

Shapley值在全集D的所有不同排列上取平均值，从而获取代理i的平均边际贡献。

然而，当直接按照公式给出的Shapley值计算机器学习模型中的数据贡献时，将存在许多问题。例如，由于需要获得每个数据子集的联盟收益，就需要在每个数据子集上都训练一个模型，并评估由此得到的模型，这种计算是随着数据数量增加呈指数复杂度增长的。研究机器学习领域贡献计算问题的现有方法可分为两类：侧重于优化公平收入分配算法的方法和侧重于设计联盟收入函数的方法。

早期的研究通常侧重于优化公平收入分配算法。为了解决计算Shapley值的指数复杂性问题，Ghorbani等[32]首次将Shapley值引入监督机器学习中用于公平数据估值，使用蒙特卡洛和基于梯度的方法来有效地估计数据的Shapley值。Jia等[33]介绍了一种快速计算Shapley值的方法，与通过定义计算精确Shapley值的指数复杂性相比，该方法允许在$O(k\log k)$时间内计算K近邻（KNN）模型上精确的Shapley值。为了解决数据Shapley值的稳定性问题，Amirata等[34]提出了分布Shapley，通过在基础数据分布的条件下定义点的值，以改进Shapley值的统计解释；分布Shapley可以评估不同分布的数据价值。Kwon等[35]进一步改进了这项工作，他们推导出了分布Shapley的解析表达式和可解释公式，以便在线性回归和二分类问题中有效地估计分布Shapley。

近来的研究已开始侧重于联盟收入函数的设计。早期的数据估值方法通常使用在特定数据集上训练的模型的分类精度作为该数据集的联盟收入函数。然而，这种联盟收入函数依赖于评估收敛模型在验证集上的性能，这对于大型复杂模型[如深度神经网络（DNN）]来说，由于其不可避免的长期模型训练，计算成本很高。此外，验证集在实际应用中可能不可获得，数据提供商可能难以就验证集的选择达成一致。近来的研究[36–37]提出了有效的技术来估计大型复杂模型的完全收敛性能，以此设计数据贡献计算中的联盟收入函数。更具体地说，基于稳健的体积Shapley值（RVSV），Xu等[36]采用了一种观点，即数据的价值由数据的内在性质决定，以此将数据集的体积设计为联盟收入函数。数据集的体积定义为其左Gram矩阵的行列式，如下所示：

$$\mathrm{Vol}(\boldsymbol{X}) := \sqrt{\left|\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)\right|} = \sqrt{|\boldsymbol{G}|} \tag{2}$$

式中，$\boldsymbol{X}$是数据矩阵；$\boldsymbol{G}$是$\boldsymbol{X}$的左Gram矩阵；Vol是$\boldsymbol{X}$的体积。

与使用验证集上的性能作为联盟收入函数相比，这种方法的计算复杂度更低，并且数据估值不受模型和任务的限制。此外，RVSV是一种基于健壮的体积度量的方法，理论上保证了复制的健壮性，即避免了通过直接数据复制造成的数据估值问题。健壮的体积度量通常将数据空间离散化为一组d立方体（其中，d是数据空间的维度），并合并同一组d立方体中的数据点作为其统计量（如平均向量），这样能使得复制的数据被合并在同一组d立方体中，从而确保方法对于直接数据复制的健壮性。基于RVSV，Xu等[36]从理论上证明了，对于线性模型和一维情况，体积和稳健体积作为联盟收入函数的适用性。然而，这种理论保证不适用于非线性模型或高维情况。此外，当应用于复杂的深度学习模型时，由于这些模型通常是非线性和高维的，缺乏理论保证仅仅依靠经验证明可能会导致问题。基于模型初始化时的数据估值（DAVINZ），Wu等[37]引入了统计学习理论（SLT）来估计DNN的完全收敛性能，以作为联盟收入函数，这完全避免了数据估值过程中模型训练的需要。更具体地说，DAVINZ通过在近来提出的神经正切核（NTK）理论中引入域差异，推导出了一个域感知的泛化界限。DNN模型$f(x,\theta)$在给定数据集上的NTK矩阵$\boldsymbol{\Theta} \in \mathbb{R}^{m\times m}$定义如下：

$$\boldsymbol{\Theta}\left(x,x';\theta\right) = \nabla_{\theta} f(x,\theta)^{\top} \nabla_{\theta} f\left(x',\theta\right) \tag{3}$$

其中，x和x'表示数据集中的数据点；θ是DNN模型的参数。近来对NTK理论的研究表明，基于初始化模型参数的NTK矩阵，可以通过理论推导得到DNN的泛化误差的上界。此外，NTK可以表征基于梯度下降的任何合理架构DNN的训练动态。DAVINZ基于NTK的这些性质，仅使用初始化的模型参数即可估计DNN的性能，并基于NTK推导出的泛化误差上界设计联盟收入函数，这一过程无需任何的模型训练。与RVSV相比，DAVINZ基于SLT，这在理论上对于深度学习模型更合理。另一方面，与精确估计相比，DAVINZ基于验证性能的上限设计联盟收入函数可能会导致更多误差。

以Shapley值为代表的贡献度计算方法在机器学习领域有各种应用。Shapley Q值[38]在多智能体强化学习中引入Shapley值来估计每个智能体对全局奖励的贡献。Wang等[39]提出了Shapley流，使用Shapley值来计算分配给因果图边缘的信度，以推断模型输入对其输出的影响。Ghorbani [40]使用Shapley值来标注无标签的数据，以提高批量主动学习的效率，同时保持性能有效性。Fan等[41]提出了用于联邦学习中公平数据估值的联邦Shapley值。Xu等[42]在联邦学习中设计了一种新的训练时梯度奖励机制，该机制根据每一轮中余弦梯度Shapley值（CGSV）计算的贡献，将不同质量的梯度分配给本地客户端。通过对原梯度向量的不同百分比掩模获得不同质量的梯度。此外，已经有一些工作将Shapley值用于真实场景下的数据估值。Tang等[43]使用Shapley值来计算大型胸部X射线数据集中训练数据的价值，这为使用Shapley值来进行大型数据集的数据估值提供了一个参考框架。

3.3. 多个数据所有者，多个数据购买者

在此场景中，多个数据所有者由各种不同的数据主体组成，从个人到数据公司和政府。数据交易涉及数据本身和数据产品，例如，从数据中训练得到的模型。数据中介

通常是各种数据所有者和数据购买者之间的复杂交易所必需的中间人。现有的关于这种场景的研究通常在其数据定价模型中考虑市场信息。我们将这些定价策略称为基于市场的定价。基于市场的数据定价是基于数据市场中的供需关系和其他信息的数据定价策略。这种定价策略的制定通常取决于数据市场中数据所有者、数据购买者和数据中介建立的三方博弈模型。这里，我们总结了数据所有者、数据购买者和数据中介在数据市场中的作用。

数据所有者是源数据的提供者；在一定程度上，它们承担着将源数据集成和处理为可在数据市场交易的数据产品的功能。数据所有者以不同的隐私保护要求向数据中介提供数据，并获得数据中介分配的相应数据使用补偿。

数据购买者是数据产品的最终购买者。数据产品不仅指数据本身，还指从数据挖掘中获得的信息或从数据中训练学习得到的模型。数据购买者根据自己的需求和预算购买不同质量的数据产品。通常可以通过向模型参数或训练数据添加不同级别的噪声来获得不同质量的数据产品。

数据中介为不同类别的数据产品提供定价模型和相应的技术支持。在做出市场决策时，数据中介必须为数据所有者设计合理和公平的补偿函数，为数据购买者设计无套利价格函数，以实现收入最大化的目标。

对于数据市场中由数据所有者、数据中介和数据购买者构建的多方博弈模型，数据中介应向数据所有者提供数据使用补偿，并制定价格函数以满足数据购买者的需求。为了设计这些函数，Niu等[44]从数据市场中的数据中介的角度出发，研究了含噪声的聚合统计交易，并提出了定价模型，该模型支持对私人相关数据聚合统计的定价，并考虑了数据所有者之间的依赖公平性。Chen等[45]首先提出了数据市场中基于模型定价的正式框架，重点是避免套利，并提供了数据中介如何将价格分配给模型以实现收入最大化的算法解决方案。更具体地说，对于具有严格凸损失函数的机器学习模型，研究人员将高斯噪声添加到模型参数中，以实现无套利定价。Liu等[46]和Lin等[47]也采用了基于模型的定价的观点，并提出了定价框架Dealer，该框架使用差分隐私（DP）来构建几个不同的模型版本，采用动态规划算法来制定定价策略以实现收入最大化，并将Shapley值应用于数据所有者的收入公平分配。Zheng等[48]通过考虑每个数据所有者的有界个性化DP提出了定价框架，并证明了无套利约束可以通过部分无套利在有界条件下合理放松。

为了设计定价策略，数据中介必须不可避免地事先从数据所有者那里访问数据，这对数据所有者是不公平的，因为数据中介可能会从访问数据中获取信息，而不会对数据所有者进行补偿。验证数据中介是否诚实收集和处理了数据非常重要。一种直接的解决方案是在建立数据市场时加密敏感信息，如数据市场中的诚实性和隐私保护（TP-DM）[49]。另一种解决方案是让中介在不通过隐私计算技术获取数据的情况下对数据进行定价。然而，该解决方案引入了一个公平交易问题：数据所有者可以在定价期间提供高质量数据，但在数据交易期间提供低质量数据。为了解决这个问题，Zhou等[50]提出了一个新概念，称为零知识附带模型支付（ZKCMP），该概念允许经过训练的机器学习模型和加密货币支付之间的公平交易。

4. 隐私计算

隐私计算是一系列加密计算技术的组合，如图4所示。它涉及高等数学、计算机科学、密码学、网络通信技术和其他学科（即安全多方计算、DP、同态加密、零知识证明、TEE）。它是数据要素和数据价值之间的桥梁，是数字经济和数据要素市场成熟的基础。通过利用隐私计算技术，数据变得可用而不可见。

世界各地都在发生数据隐私泄露事件。例如，2018年，剑桥分析公司（Cambridge Analytica）[51]涉嫌窃取Facebook用户的信息，以操纵美国大选和英国公投。各种隐私泄露问题表明，对数据隐私保护的研究对于充分利用数据的价值极为必要。近年来，无论是国内还是国外，与数据隐私相关的法律法规都日趋成熟和完善。例如，欧盟的GDPR和中国的《数据安全管理办法》都规定了保护个人信息隐私的责任和规范。总体而言，隐私计算是实现数据隐私保护和安全的关键。

在实际应用方面，每种隐私计算技术都有其自身的特点、优点和缺点。根据应用场景、安全要求和效率要求，有必要为每种应用场景选择最合适的隐私计算技术。在隐私计算中，关键问题如下：

（1）谁拥有数据？

（2）谁使用数据和数据衍生品？

显然，当数据由使用数据本身的一方拥有时，不需要隐私计算。因此，在本节中，我们对数据所有者和数据使用者是相互不信任的主体的场景感兴趣。

4.1. 单个数据所有者，多个数据购买者

在单个数据所有者和多个数据购买者的场景下，数据由单个数据所有者持有，该数据所有者希望委托给单个不受信任的计算节点，以便在联合数据库上进行计算。前面提到的同态加密技术也可以用于这种场景；然而，这样做

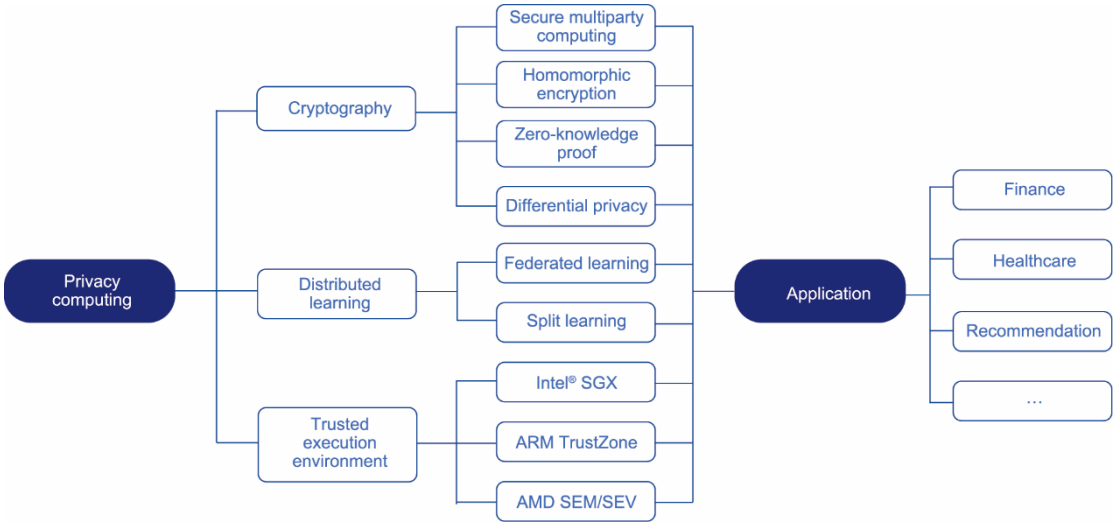

图4. 隐私计算。SGX：软件保护扩展；ARM：进阶精简指令集机器；AMD：超微半导体；SEM：安全加密存储器；SEV：安全加密虚拟化。

可能需要分布式密钥生成和复杂的密钥管理。对于同态加密来说，计算任务有时过于繁重。DP [52]是一种廉价的隐私增强技术，它基于密码学，建立在严格的数学定义之上，提供了一种定量评估方法。DP的主要思想是通过删除个人特征同时保留统计特征来保护用户隐私。如果算法在两个数据库上运行，而这两个数据库正好相差一个条目，并且产生的差异由ε限定，则该算法称为ε-差分隐私[53]。较小的ε表示该算法可以确保更强的隐私。换言之，算法处理两个相似数据集所获得的输出越接近，特定条目数据的隐私保护就越好。近来，研究提出了本地DP（LDP）。LDP机制不像DP机制那样向聚合结果添加噪声，而是在向中央服务器发送数据之前由每个用户添加噪声。因此，用户不依赖中央服务器的可信度。DP和LDP机制都可以与机器学习相结合。DP机制可以通过将随机噪声添加到目标函数、梯度和输出结果中来提供隐私保护，例如，通过添加拉普拉斯噪声或高斯噪声[53]。LDP机制可用于保护各种类型的训练数据集，如项目数据集[54]、项目集[55]和图[56]。

4.2. 多个数据所有者，单个数据购买者

多个数据所有者和单个数据购买者的场景中可以进一步分为几个子案例。当多个数据所有者也是计算节点时，多方计算（MPC）是一种理想的技术。安全多方计算[57]于1982年被提出，当时图灵奖得主Chi-Chi Yao提出了著名的百万富翁问题，这需要多方在不泄露私人数据的情况下，合作解决问题。自提出以来，安全多方计算一直受到广泛关注和研究，这一领域新的方法和工具在不断涌现。其中，用于安全双方计算的协议通常是混淆电路（GC）[57]与不经意传输（OT）[58]相结合，而用于安全MPC（即三方或更多方）的协议通常为与OT相结合的秘钥分享（SS）。前者（即GC+OT）的主要问题是计算开销可能更高，尽管通信轮数的需求更少。后者（即SS+OT）[59]通常需要OT的多次迭代和大量通信轮数，尽管其计算开销较小。

在模型训练方面，传统MPC通常需要非常大量的通信。在这种情况下，可以采用协作学习来提高效率。协作学习是一类MPC协议，旨在利用多方的数据训练数据模型，并保持多方的数据隐私。联邦学习和分割学习是协作学习中两个重要的框架。在联邦学习[60]中，中央服务器将当前模型分发给客户端。每个客户端使用自己的本地数据训练模型，然后将模型上传到服务器进行聚合。重复此过程，直到模型收敛。这一技术概念于2016年首次由谷歌引入，当时谷歌提出了移动终端的联邦学习。此后，WeBank为金融行业提出了第一个“联邦迁移学习”[61]解决方案，将迁移学习和联邦学习相结合。目前，各种开源联邦学习框架，如Federated AI Technology Enabler (FATE)和TensorFlow Federated，在人工智能领域不断涌现和成熟。

在实际应用场景中，假设N个客户端$\{U_1, ..., U_N\}$持有自己的数据集$\{D_1, ..., D_N\}$，其他客户端无法直接访问这些数据集。联邦学习通过从分布式设备收集训练信息来学习模型。它包括三个基本步骤：

（1）服务器向每个客户端发送初始模型。

（2）客户端U_i不需要共享自己的本地数据，只需要用本地数据D_i训练自己的模型W_i（W_i是U_i的本地模型）。

（3）服务器将收集的本地模型$\{W_1, ..., W_N\}$聚合为全局模型W'，并下传聚合后的全局模型给各个客户端以更

新其本地模型。

随着联邦学习的快速发展，联邦学习模型的效率和准确性越来越接近集中式训练得到的模型。基于数据的样本空间和特征向量空间的不同分布模式，联邦学习可以分为三类：横向联邦学习、纵向联邦学习和联邦迁移学习。横向联邦学习适用于数据集之间的用户特征向量大量重叠但用户很少重叠的情况。换句话说，不同的数据行具有相同的特征向量（在特征向量维度中对齐）。因此，横向联邦学习可以增加用户样本量。例如，Kim等[62]提出了一个称为BlockFL的横向联合联邦学习框架，其中每个移动设备使用区块链网络来更新本地模型，Smith等[63]提出了一种称为MOCHA的联邦学习方法，以解决多任务中的安全问题，该方法允许多个客户端共同完成任务并确保隐私和安全。多任务联邦学习还改善了原始分布式多任务学习的通信成本，并增强了容错能力。

纵向联邦学习适用于数据集之间的用户特征向量很少重叠，但用户大量重叠的情况。因此，纵向联邦学习可以增加训练数据的特征向量的维数。例如，Cheng等[64]提出了一种称为SecureBoost的纵向联邦学习系统，其中各方将用户特征向量组合在一起进行训练，以提高决策的准确性，这是一种无损训练方案。Hardy等[65]提出了一种基于纵向联邦学习的逻辑回归模型，能够保护数据隐私。该模型使用流水线实体分析和Paillier半同态加密进行分布式逻辑回归，可以有效保护隐私，并提高分类器的准确性。

联邦迁移学习适用于数据集之间的用户和用户特征向量没有太多重叠但迁移学习可以用来解决数据和标签不足的情况。迁移学习适用于试图优化任务的性能，但没有足够的相关数据用于训练。例如，医院放射科很难收集许多X射线扫描来建立一个有效的放射诊断系统。迁移学习可以通过与其他相关和不同任务（如图像识别任务）相结合，使建立一个有效的放射诊断系统成为可能。通过联邦迁移学习，我们不仅可以保护数据隐私，还可以将辅助任务的模型迁移到目标模型学习，从而解决数据量小的问题。

联邦学习强调数据层面的分离，而分割学习[66–67]的核心思想是分离网络结构。在最简单的分割学习示例中，网络结构被拆分为两部分，一部分存储在客户端，另一部分存储在服务器端。客户端无法访问服务器端模型，服务器端也无法访问客户端模型。与联邦学习相比，分割学习减少了客户端的计算量。

4.3. 多个数据所有者，多个数据购买者

在多个数据所有者和多个数据购买者的场景下，通常需要数据中介的参与，保护数据隐私是基本要求。同态加密[68]技术适用于这种情况。同态加密是一种允许用户在不解密加密数据的情况下对加密数据进行计算的加密形式。这些计算的结果以加密形式存储，并且在解密结果之后，输出与通过对未加密数据执行相同操作而获得的结果相同。常见的同态加密类型包括部分同态[69]、有点同态[70]、分级全同态[71]和全同态加密[68]。自从IBM科学家Gentry构建了第一个真正的全同态加密方法[69]以来，密码学在这一领域进行了深入的研究。已经创建了第二代[72–73]、第三代[74]和第四代[75]全同态加密系统。

TEE [76]也可用作此场景下的有效解决方案。TEE通过硬件技术隔离保护数据。在启用TEE的中央处理器（CPU）中，可以创建一个特定的隔离区域，作为敏感数据及其应用程序代码的安全内容容器，确保其机密性和完整性。即使攻击者控制了操作系统和其他特权级别的软件，也无法访问隔离区域（即无法修改或读取信息）。TEE上运行的应用程序称为可信应用程序；它们彼此隔离，未经授权无法读取和操作其他可信任应用程序的数据。显然，通过软件算法和硬件技术实现的隔离确保了可以安全地计算、存储、传输和删除私人信息。TEE技术通常取决于特定的技术平台和实施供应商；常见技术包括英特尔软件保护扩展（SGX）、进阶精简指令集机器（ARM）TrustZone和超微半导体（AMD）安全加密存储器（SEM）/安全加密虚拟化（SEV）。

此外，可以采用许多其他可验证的计算技术来确保计算完整性。零知识证明技术是可验证计算中广泛使用的解决方案。在这个证明系统中，证明者知道问题的答案，并且必须向验证者证明“他或她知道答案”，但验证者除了“他或她知道答案”这一事实之外，无法获得任何其他信息。零知识证明[77]最早由Shafi Goldwasser、Silvio Micali和Charles Rackoff在他们的论文“交互式证明系统的知识复杂性”中首次提出。随后，零知识证明技术继续发展，直到2013年，密码学家创建了第一个高效且可商用的通用简洁非交互式零知识证明协议：零知识简洁知识论证（zk-SNARKs）[78]。

5. 挑战和开放性问题

在本节中，我们将讨论未来可能工作中一些有趣的尚未探索的挑战。我们希望这些讨论将引起对这一快速增长领域的更广泛研究兴趣。

5.1. 保障数据权利的合适技术解决方案

近来，机器学习模型被广泛用于数据处理。虽然通过训练完成的这些模型可以独立于用于训练的数据，但它们仍然必须满足数据主体的要求。机器学习模型的黑箱特性给保障各种数据权利带来了挑战。例如，被遗忘权是GDPR规定的数据主体的一项权利。数据主体有权要求控制者删除其个人数据，不得无故拖延。与传统数据库不同，可以直接删除相应的数据。然而，让机器学习模型忘记学习到的数据是一个具有挑战性的任务。在访问权方面，数据主体有权从控制者处获得有关其个人数据是否正在处理的确认。为了防止互联网上共享的数据被非法抓取用于模型训练，需要相应的技术解决方案，以使数据成为可见但不可利用的不可学习样本。机器学习的复杂模型和数据依赖性是确保数据权利的主要挑战。

5.2. 数据定价和隐私计算的结合

数据定价为数据交易和流通过程中的所有权收益提供了一种技术解决方案，而隐私计算为保护数据交易和流通过程中的隐私提供了技术解决方案。数据定价和隐私计算在数据交易过程中相辅相成。最近，分布式场景下的机器学习模型训练已经成为一个研究热点。在分布式场景下，数据交易在机器学习模型的训练过程中频繁发生。这种情况需要设计实时高效的数据定价和隐私计算技术，以满足分布式场景中机器学习模型的训练。在本文中，我们概述了分布式场景中基于模型定价的数据定价技术和基于联邦学习的隐私计算技术。一个例子是Xu等[42]结合联邦学习的隐私计算技术提出的新定价机制，该技术通过联邦学习机制补偿数据所有者。我们认为，对于分布式场景，挑战来自数据定价和隐私计算技术的结合。应通过利用隐私计算技术，例如，通过联邦学习的相关机制，设计有效和公平的定价策略。

5.3. 符合数据交易市场实际情况的数据要素计算

实际的数据交易市场包含各种类型的数据交易，数据产品的形式从原始数据到通过数据训练获得的机器学习模型。数据要素计算应基于实际的交易类型，以帮助完成交易。数据交易市场是复杂的，随着供求信息的变化而变化。数据要素计算提供了对市场的解释，指导市场的每个主体做出判断，确保主体的权利，稳定市场价格，保护数据隐私，以完成数据交易。数据要素计算的研究不仅应建立一个基于数据科学的模型，还应包括对市场机制和用户行为等因素的综合考量，以符合数据市场的实际情况。数据要素计算应从跨学科的角度进行研究，包括数据科学、经济学和市场营销学等。

6. 结论

在大数据时代，大数据治理已成为社会各界普遍关注的问题，需要适当的算法方法来确保大数据的流通和交易。本文概述了数据交易市场体系中的数据要素计算，并回顾了数据交易过程中的三个主要问题：数据权利、数据定价和隐私计算。本文还讨论了未来可能开展的研究所面临的挑战，希望本文的讨论将引起对这一快速增长领域的更广泛的研究兴趣。

Compliance with ethics guidelines

Jimin Xu, Nuanxin Hong, Zhening Xu, Zhou Zhao, Chao Wu, Kun Kuang, Jiaping Wang, Mingjie Zhu, Jingren Zhou, Kui Ren, Xiaohu Yang, Cewu Lu, Jian Pei, and Harry Shum declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 2019;5(9):1572–83.

[2] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020; 577(7792):706–10.

[3] Lu L, Meng X, Mao Z, Karniadakis GE. DeepXDE: a deep learning library for solving differential equations. SIAM Rev 2021;63(1):208–28.

[4] Pei J. A survey on data pricing: from economics to data science. IEEE Trans Knowl Data Eng 2020;34(10):4586–608.

[5] Cong Z, Luo X, Jian P, Zhu F, Zhang Y. Data pricing in machine learning pipelines. Knowl Inf Syst 2021;64:1417–55.

[6] Parkins D. The world's most valuable resource is no longer oil, but data [Internet]. New York City: The Economist; 2017 May 6 [cited 2022 Dec 27]. Available from: https://www.economist.com/leaders/2017/05/06/the-worldsmost-valuable-resource-is-no-longer-oil-but-data.

[7] Atkinson RD. IP protection in the data economy: getting the balance right on 13 critical issues. Report. Washington, DC: Information Technology & Innovation Foundation; 2019 Jan 22.

[8] Klein B, Crawford RG, Alchian AA. Vertical integration, appropriable rents, and the competitive contracting process. J Law Econ 1978;21(2):297–326.

[9] Williamson OE. Transaction-cost economics: the governance of contractual relations. J Law Econ 1979;22(2):233–61.

[10] Demsetz H. Toward a theory of property rights. Am Econ Rev 1967; 57(2): 347–59.

[11] Balkin JM. The fiduciary model of privacy. Harv Law Rev Forum 2020; 134: 11–33.

[12] Ritter J, Mayer A. Regulating data as property: a new construct for moving forward. Duke Law Technol Rev 2018;16:220–77.

[13] Michael K, Kobran S, Abbas R, PrivacyHamdoun S., data rights and cybersecurity: technology for good in the achievement of sustainable development goals. In: Proceedings of 2019 IEEEInternational Symposium on Technology and Society ISTAS); 2019 Nov 15–16; Medford, MA, USA. New York City: IEEE; 2019. p. 1–13.

[14] Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR). Brussels: European Commission; 2017.
[15] Truong NB, Sun K, Lee GM, Guo Y. GDPR-compliant personal data management: a blockchain-based solution. IEEE Trans Inf Forensics Secur 2020;15:1746–61.
[16] Wingerath W, Gessert F, Witt E, Kuhlmann H, Bücklers F, Wollmer B, et al. Speed Kit: a polyglot & GDPR-compliant approach for caching personalized content. In: Proceedings of 2020 IEEE 36th International Conference on Data Engineering (ICDE); 2020 Apr 20 – 24; Dallas, TX, USA. New York City: IEEE; 2020. p. 1603–8.
[17] Agostinelli S, Maggi FM, Marrella A, Sapio F. Achieving GDPR compliance of BPMN process models. In: Cappiello C, Ruiz M, editors. Information systems engineering in responsible information systems. New York City: Springer; 2019.
[18] Ginart AA, Guan MY, Valiant G, Zou J. Making AI forget you: data deletion in machine learning. In: Proceedings of 33rd Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada; 2019.
[19] Li Q, Wen Z, Wu Z, Hu S, Wang N, Li Y, et al. A survey on federated learning systems: vision, hype and reality for data privacy and protection. IEEE Trans Knowl Data Eng 2023;35(4):3347–66.
[20] McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. Communicationefficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS); 2017 Apr 20–22; Lauderdale, FL, USA; 2017.
[21] The Chartered Institute of Marketing (CIM). Data right: best data practice [Internet]. Berkshire: CIM; c2018 [cited 2022 Dec 27]. Available from: https://www.cim.co.uk/more/data-right/.
[22] Kerber W. A new (intellectual) property right for non-personal data? An economic analysis. J Eur Int IP Law 2016;11:989–99.
[23] Grossman SJ, Hart OD. The costs and benefits of ownership: a theory of vertical and lateral integration. J Polit Econ 1986;94(4):691–719.
[24] Yan T, Procaccia AD. If you like Shapley then you’ll love the core. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2021 Feb 2–9; online. Palo Alto: AAAI Press; 2021. p. 5751–9.
[25] Koutris P, Upadhyaya P, Balazinska M, Howe B, Suciu D. Query-based data pricing. J ACM 2015;62(5):1–44.
[26] Koutris P, Upadhyaya P, Balazinska M, Howe B, Suciu D. Toward practical query pricing with QueryMarket. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data; 2013 Jun 22–27; New York City, NY, USA. New York City: Association for Computing Machinery; 2013. p. 613–24.
[27] Deep S, Koutris P. QIRANA: a framework for scalable query pricing. In: Proceedings of the 2017 ACM International Conference on Management of Data; 2017 May 14 – 19; Chicago, IL, USA. New York City: Association for Computing Machinery; 2017. p. 699–713.
[28] Cook RD. Detection of influential observation in linear regression. Technometrics 2000;42(1):65–8.
[29] Cook RD, Weisberg S. Residuals and influence in regression. New York City: Chapman and Hall; 1982.
[30] Yoon J, Arik S, Pfister T. Data valuation using reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Vienna, Austria; 2020.
[31] Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, editors. Contributions to the theory of games. Princeton: Princeton University Press; 2016.
[32] Ghorbani A, Zou J. Data Shapley: equitable valuation of data for machine learning. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA; 2019.
[33] Jia R, Dao D, Wang B, Hubis FA, Gurel NM, Li B, et al. Efficient task-specific data valuation for nearest neighbor algorithms. Proc VLDB Endow 2019; 12(11): 1610–23.
[34] Amirata G, Kim M, Zou J. A distributional framework for data valuation. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jun 12–18; Vienna, Austria. 2020. p. 3535–44.
[35] Kwon Y, Rivas MA, Zou J. Efficient computation and analysis of distributional Shapley values. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics; 2021 Apr 13–15; online. 2021. p. 793–801.
[36] Xu X, Wu Z, Foo CS, Low BKH. Validation free and replication robust volumebased data valuation. In: Proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS 2021); 2021 Dec 7–10; online. 2021. p. 10837–48.
[37] Wu Z, Shu Y, Low BKH. DAVINZ: data valuation using deep neural networks at initialization. In: Proceedings of International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MA, USA. 2022. p. 24150–76.
[38] Wang J, Zhang Y, Kim TK, Gu Y. Shapley Q-value: a local reward approach to solve global reward games. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence; 2020 Feb 7 – 12; New York City, NY, USA. Palo Alto: AAAI Press; 2020. p. 7285–92.
[39] Wang J, Wiens J, Lundberg S. Shapley flow: a graph-based approach to interpreting model predictions. In: Proceedings of 23rd International Conference on Artificial Intelligence and Statistics; 2020 Aug 26 – 28; online. New York City: Society for Artificial Intelligence and Statistics; 2021. p. 721–9.
[40] Ghorbani A, Zou J, Esteva A. Data Shapley valuation for efficient batch active learning. 2021. arXiv:2104.08312.
[41] Fan Z, Fang H, Zhou Z, Pei J, Friedlander MP, Liu C, et al. Improving fairness for data valuation in federated learning. 2021. arXiv:2109.09046.
[42] Xu X, Lyu L, Ma X, Miao CL, Foo CS, Low BKH. Gradient driven rewards to guarantee fairness in collaborative machine learning. In: Proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS 2021); 2021 Dec 7–10; online. 2021. p. 16104–17.
[43] Tang S, Ghorbani A, Yamashita R, Rehman S, Dunnmon JA, Zou J, et al. Data valuation for medical imaging using Shapley value and application to a largescale chest X-ray dataset. Sci Rep 2021;11:8366.
[44] Niu C, Zheng Z, Wu F, Tang SJ, Gao X, Chen G. Unlocking the value of privacy: trading aggregate statistics over private correlated data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 Aug 19 – 23; London, UK. New York City: Association for Computing Machinery (ACM); 2018. p. 2031–40.
[45] Chen L, Koutris P, Kumar A. Towards model-based pricing for machine learning in a data marketplace. In: Proceedings of the 2019 International Conference on Management of Data; 2019 Jun 30 – Jul 5; Amsterdam, the Netherlands. New York City: Association for Computing Machinery (ACM); 2019. p. 1535–52.
[46] Liu J, Lou J, Liu J, Xiong L, Pei J, Sun J. Dealer: an end-to-end model marketplace with differential privacy. Pro VLDB Endow 2021;14:957–69.
[47] Lin Q, Zhang J, Liu J, Ren K, Lou J, Jun L, et al. Demonstration of Dealer: an endto- end model marketplace with differential privacy. Pro VLDB Endow 2021;14 (12):2747–50.
[48] Zheng S, Cao Y, Yoshikawa M. Trading data with personalized differential privacy and partial arbitrage freeness. 2021. arXiv:2105.01651.
[49] Niu C, Zheng Z, Wu F, Gao X, Chen G. Trading data in good faith: integrating truthfulness and privacy preservation in data markets. In: Proceedings of 2017 IEEE 33rd International Conference on Data Engineering (ICDE); 2017 Apr 19–22; DiegoSan, CA, USA. New York City: IEEE; 2017. p. 223–6.
[50] Zhou Z, Cao X, Liu J, Zhang B, Ren K. Zero knowledge contingent payments for trained neural networks. In: Bertino E, Shulman H, Waidner M, editors. Computer security—ESORICS 2021. New York City: Springer; 2021. p. 628–48.
[51] Isaak J, Hanna MJ. User data privacy: Facebook, Cambridge Analytica, and privacy protection. Computer 2018;51(8):56–9.
[52] Dwork C. Differential privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I, editors. International colloquium on automata, languages, and programming. Berlin: Springer; 2006. p. 1–12.
[53] Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 2014;9(3–4):211–407.
[54] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: randomized aggregatable privacypreserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security; 2014 Nov 3–7; Scottsdale, AZ, USA. New York City: Association for Computing Machinery (ACM); 2014. p. 1054–67.
[55] Qin Z, Yang Y, Yu T, Khalil I, Xiao X, Ren K. Heavy hitter estimation over setvalued data with local differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016 Oct 24–28; Vienna, Austria. New York City: Association for Computing Machinery (ACM); 2016. p. 192–203.
[56] Qin Z, Yu T, Yang Y, Khalil I, Xiao X, Ren K. Generating synthetic decentralized social graphs with local differential privacy. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; 2017 Oct 30– Nov 3; Dallas, TX, USA. New York City: Association for Computing Machinery (ACM); 2017. p. 425–38.
[57] Yao AC. Protocols for secure computations. In: Proceedings of 23rd Annual Symposium On Foundations Of Computer Science (SFCS 1982); 1982 Nov 3–5; Chicago, IL, USA. New York City: IEEE; 1982. p. 160–4.
[58] Rabin MO. How to exchange secrets with oblivious transfer. 2005. IACR Cryptology ePrint Archive:187.

[59] Tassa T. Generalized oblivious transfer by secret sharing. Des Codes Cryptogr 2011;58(1):11–21.
[60] Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh TA, Bacon D. Federated learning: strategies for improving communication efficiency. 2016. arXiv: 1610.05492.
[61] Liu Y, Kang Y, Xing C, Chen T, Yang Q. A secure federated transfer learning framework. IEEE Intell Syst 2020;35(4):70–82.
[62] Kim H, Park J, Bennis M, Kim SL. Blockchained on-device federated learning. IEEE Commun Lett 2020;24(6):1279–83.
[63] Smith V, Chiang CK, Sanjabi M, Talwalkar A. Federated multi-task learning. In: Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4 – 9; Long Beach, CA, USA. Red Hook: Curran Associates Inc.; 2017. p. 30.
[64] Cheng K, Fan T, Jin Y, Liu Y, Chen T, Papadopoulos D, et al. Secureboost: a lossless federated learning framework. IEEE Intell Syst 2021;36(6):87–98.
[65] Hardy S, Henecka W, Ivey-Law H, Nock R, Patrini G, Smith G, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. 2017. arXiv:1711.10677.
[66] Zhao S, Zhou L, Wang W, Cai D, Kam TL, Xu Y, et al. Splitnet: divide and co-training. 2020. arXiv:2011.14660.
[67] Vepakomma P, Gupta O, Swedish T, Raskar R. Split learning for health: distributed deep learning without sharing raw patient data. 2018. arXiv: 1812.00564.
[68] Gentry C. Fully homomorphic encryption using ideal lattices. In: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing; 2009 May 31 – Jun 2; Bethesda, MD, USA. New York City: Association for Computing Machinery (ACM); 2009. p. 169–78.
[69] Shoukry Y, Gatsis K, Alanwar A, Pappas GJ, Seshia SA, Srivastava M, et al. Privacy-aware quadratic optimization using partially homomorphic encryption. In: Proceedings of 2016 IEEE 55th Conference on Decision and Control (CDC); 2016 Dec 12–14; VegasLas, NV, USA. New York City: IEEE; 2016. p. 5053–8.
[70] Damgård I, Pastro V, Smart N, Zakarias S. Multiparty computation from somewhat homomorphic encryption. In: Safavi-Naini R, Canetti R, editors. Advances in cryptology—CRYPTO 2012. Berlin: Springer; 2012. p. 43–62.
[71] Gorbunov S, Vaikuntanathan V, Wichs D. Leveled fully homomorphic signatures from standard lattices. In: Proceedings of the 57th Annual ACM Symposium on Theory of Computing; 2015 Jun 14 – 17; Portland, OR, USA. New York City: Association for Computing Machinery (ACM); 2015. p. 469–77.
[72] Brakerski Z, Vaikuntanathan V. Efficient fully homomorphic encryption from (standard) LWE. SIAM J Comput 2014;43(2):831–71.
[73] López-Alt A, Tromer E, Vaikuntanathan V. On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In: Proceedings of the 44th Annual ACM Symposium on Theory of Computing; 2012 May 19–22; New York City, NY, USA. New York City: Association for Computing Machinery; 2012. p. 1219–34.
[74] Chillotti I, Gama N, Georgieva M, Izabachène M. Faster fully homomorphic encryption: bootstrapping in less than 0.1 seconds. In: Proceedings of 22nd International Conference on the Theory and Application of cryptology and Information Security; 2016 Dec 4 – 8; Hanoi, Vietnam. Berlin: Springer; 2016. p. 3–33.
[75] Cheon JH, Kim A, Kim M, Song Y. Homomorphic encryption for arithmetic of approximate numbers. In: Takagi T, Peyrin T, editors. Advances in cryptology—ASIACRYPT 2017. Berlin: Springer; 2017. p. 409–37.
[76] Sabt M, Achemlal M, Bouabdallah A. Trusted execution environment: what it is, and what it is not. In: Proceedings of the 2015 IEEETrustcom/ BigDataSE/ ISPA; 2015 Aug 20 – 22; Helsinki, Finland. New York City: IEEE; 2015. p. 57–64.
[77] Goldwasser S, Micali S, Rackoff C. The knowledge complexity of interactive proof systems. SIAM J Comput 1989;18:186–208.
[78] Bitansky N, Canetti R, Chiesa A, Tromer E. 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 Jan 8 – 10; Cambridge, MA, USA. New York City: Association for Computing Machinery (ACM); 2012. p. 326–49.