Engineering 25 (2023) 66–76

Contents lists available at ScienceDirect

Engineering

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

Research
Artificial Intelligence—Review

Data-Driven Learning for Data Rights, Data Pricing, and Privacy Computing

Jimin Xu [a], Nuanxin Hong [a], Zhening Xu [a], Zhou Zhao [a], Chao Wu [a], Kun Kuang [a,*], Jiaping Wang [b], Mingjie Zhu [c], Jingren Zhou [d], Kui Ren [a], Xiaohu Yang [a], Cewu Lu [e], Jian Pei [f], Harry Shum [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

Keywords:
Data science
Artificial intelligence
Data rights
Data pricing
Privacy computing

ABSTRACT

In recent years, data has become one of the most important resources in the digital economy. Unlike traditional resources, the digital nature of data makes it difficult to value and contract. Therefore, establishing an efficient and standard data-transaction market system would be beneficial for lowering cost and improving productivity among the parties in this industry. Although numerous studies have been dedicated to the issue of complying with data regulations and other data-transaction issues such as privacy and pricing, little work has been done to provide a comprehensive review of these studies in the fields of machine learning and data science. To provide a complete and up-to-date understanding of this topic, this review covers the three key issues of data transaction: data rights, data pricing, and privacy computing. By connecting these topics, this paper provides a big picture of a data ecosystem in which data is generated by data subjects such as individuals, research agencies, and governments, while data processors acquire data for innovational or operational purposes, and benefits are allocated according to the data's respective ownership via an appropriate price. With the long-term goal of making artificial intelligence (AI) beneficial to human society, AI algorithms will then be assessed by data protection regulations (i.e., privacy protection regulations) to help build trustworthy AI systems for daily life.

1. Introduction

In recent years, the Internet, big data, cloud computing, artificial intelligence (AI), and other technologies have accelerated in innovation and have been increasingly integrated into the complete processes of various fields of economic and social development. The rapid development speed, wide radiation range, and deep influence of the digital economy are unprecedented. As a new factor of production in a digital economy, data has been generated in enormous quantities and contains a great deal of economic value. As a result, data-driven methods such as machine learning have been widely used in many areas, including chemical reaction prediction [1], protein structure prediction [2], and scientific computation [3], among others. Therefore, establishing an efficient and standard data-transaction market system would be beneficial for utilizing the value of the production factor of the digital economy. Very recently, Pei [4] presented a review connecting economics, digital product pricing, and data product pricing, which focused on economics and the fundamental mathematical principles of data pricing and digital product pricing. Another review by Cong et al. [5] focused on machine learning pipelines and covered studies on pricing data labels. Unlike these existing reviews, the current review discusses three key issues of the digital economy for establishing a data-transaction market system—namely, data rights, data pricing, and privacy computing—and consolidates them as data factors in a computing framework.

Data rights, including rights subjects and rights contents, are the premise of data transactions, which are identified and protected by laws and regulations. Recently, an increasing number

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

https://doi.org/10.1016/j.eng.2022.12.008

of countries have begun to pay attention to the legislation of big data. For example, the European Union (EU) has released the General Data Protection Regulation (GDPR), and China has released the Personal Information Protection Law (PIPL). Technical solutions are urgently needed to guarantee the data rights stipulated by laws. Data pricing and privacy computing are essential in the process of data transaction. Unlike traditional commodity transaction, the particularity of data requires technical solutions for the formulation of pricing strategies and the protection of data privacy.

Data pricing and privacy computing complement each other in the process of data transaction. Here, we will introduce technical solutions for data pricing and privacy computing using three typical data-transaction scenarios. The first data-transaction scenario consists of a single data owner and multiple data buyers. In this scenario, customers typically purchase datasets from companies—such as Twitter, Bloomberg, or Pistachio—in order to access the data. Multiple pricing strategies and privacy demands are needed in this data-transaction scenario. The second data-transaction scenario consists of multiple data owners and a single data buyer. In this case, to utilize the data stored among various data owners, it is typically necessary to build a trusted privacy computing method to realize the distributed training of the model, as well as a fair data-pricing method to ensure an incentive mechanism for contributions from various data owners. The third data-transaction scenario consists of multiple data owners and multiple data buyers. Here, data brokers are typically involved in designing reasonable and fair compensation functions for data owners and arbitrage-free price functions for data buyers, in order to achieve the objective of revenue maximization. In this multi-party data-transaction process, multiple privacy demands must be met among data owners, data brokers, and data buyers.

In the first section that follows, we discuss data-rights issues such as data ownership and privacy protection that have arisen with the ever-increasing activity in the digital economy. Extensive concern regarding data-rights issues has eventually led to legislations such as the GDPR. At present, the question of whether or not data should be under heavy regulation is still being hotly debated. New technologies that comply with the existing regulations are becoming the new focus of the industry. In this section, we provide an overview of data rights in accordance with the above topic and introduce a few potential solutions in these areas.

In the second section, we discuss technical solutions for data pricing that have recently been proposed. With the popularity of mobile terminal devices, an increasing amount of end-to-end personal information or personal data is being produced and endowed with certain property attributes. Data processors can use this data to train models and obtain commercial benefits from it. As the owner of data assets, individuals should be compensated for their data being used. In this section, we provide an overview and comprehensive review of data-pricing research based on three typical data-transaction scenarios: query-based pricing, Shapley value-for-model-based pricing, and data-market-based pricing.

In the third section, we discuss privacy computing, which is a combination of a range of cryptographic computing technologies. Sensitive information in data may be reverse acquired by data processors through certain methods when the data is accessed, resulting in the disclosure and abuse of sensitive information on the subjects of the data. Certain technical measures are necessary to achieve data privacy and security in order to prevent this problem and ensure that data is being lawfully used. Privacy computing forms a bridge between data factors and data value, by protecting sensitive information of the data during the data-transaction process. In this section, we introduce privacy computing from three typical data-transaction scenarios covering three technical threads: cryptography technology, trusted execution environment (TEE), and collaborative learning.

In this paper, we consolidate data rights, data pricing, and privacy computing into a data-factor computing framework. As shown in Fig. 1, data rights, data pricing, and privacy computing are the relevant technologies in a data-factor computing system. In different industries with data being generated by different participants (e.g., individuals, business platforms, and government agencies), it is first necessary to determine data-related rights, such as the right to use data, data ownership, and data privacy; next, it is necessary to assess the value of the data and to allocate revenue according to the attribution of data property ownership; finally, it is necessary to add the necessary privacy protection in the process of data utilization to prevent the leakage or malicious theft of private information. Using this data-factor computing framework, we review three main data-transaction issues: data rights, data pricing, and privacy computing. We also provide a perspective on interesting challenges for possible future work.

2. Data rights

The generation and exchange of industrial data play critical roles in the modern digital market; some have even claimed that industrial data is taking the place of oil as the most valuable resource [6]. International Business Machines (IBM) Corporation estimates that 2.5 quintillion bytes of data are being created daily [7]. With emerging technology that facilitates data transportation, analysis, and so forth, data can be created and collected even faster. The high volume of data trading is bringing attention to data's intrinsically high externality cost, in that it is *ex ante* noncontractible [8,9] due to its intrinsic nature of being different from any other data. As a result, data rights are rising as a form of property rights, as it is more profitable for entities to internalize externalities than to negotiate *ex ante* [10]. Digital companies provide the service of smart recommendations based on user data collection, but there is concern that some content is biased, misleading, and suggestive in favor of the service provider. As a result, data subjects are vulnerable to and dependent on so-called data capitalism [11]. Therefore, the data economy community is seeking definitive ruling on data, such as on data ownership and corresponding rights [12].

Data rights represent the entity's ownership and control over said types of information, as shown in Fig. 2. We separate data rights into three categories: personal data rights, industrial data rights, and government data rights [13].

The GDPR became enforceable in May 2018 in all EU countries. In quick succession, many nations outside the EU adapted similar legislations, such as China's PIPL. The aim of the GDPR is to ensure that personal data is processed "lawfully, fairly, and in a transparent manner," and to ensure that data subjects are granted the rights of transparency, modality, access, rectification, erasure, the ability to object, and automated individual decision-making [14].

To comply with the GDPR, researchers have come up with a few solutions, one of which is the implementation of a blockchain-based system. It is almost impossible for service providers with a traditional centralized client–server architecture to ensure that they follow the GDPR guidelines on a continuous basis. Here, blockchain technology is a perfect solution due to its decentralized, hard-to-tamper-with, and easy-access nature. Truong et al. [15] offered an example of a GDPR-compliant personal data-management platform built on top of the Hyperledger Fabric permissioned blockchain framework. The main challenge ahead is to implement mechanisms to resolve the lack of trusted centralized resource servers and to potentially provide computational capability on the blockchain network.

Other methods of GDPR compliance mainly revolve around an elaborate framework satisfying the GDPR requirements [16–19].

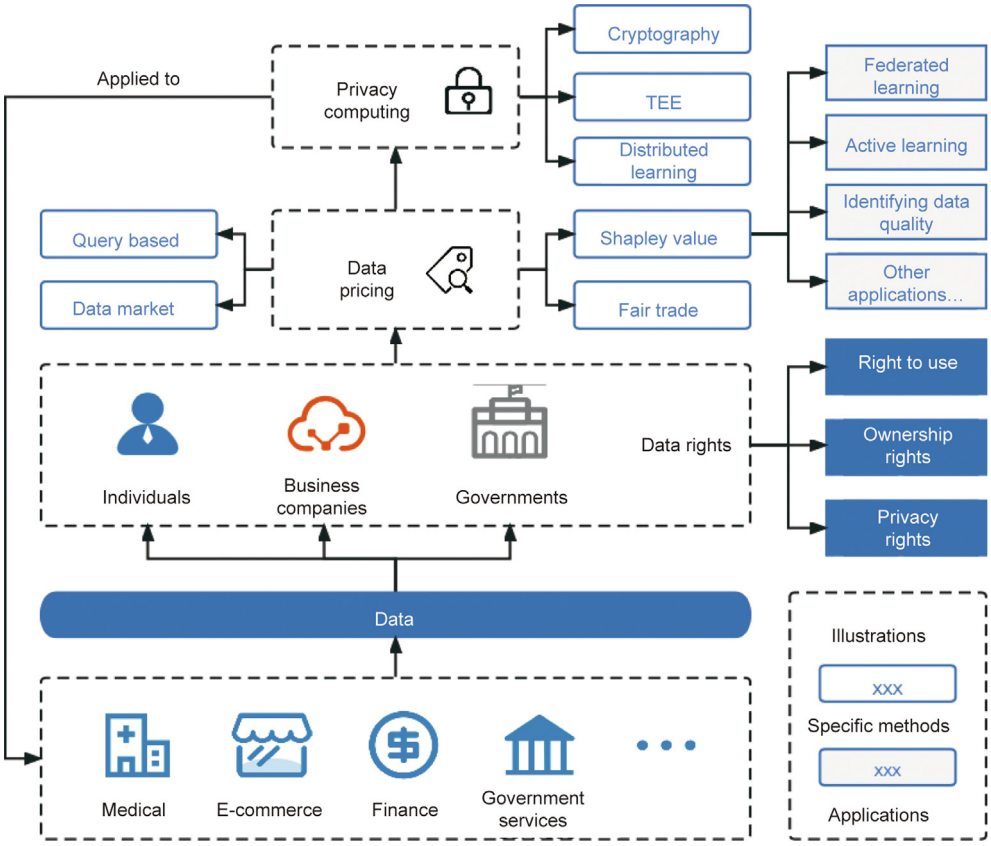

Fig. 1. Data-factor computing.

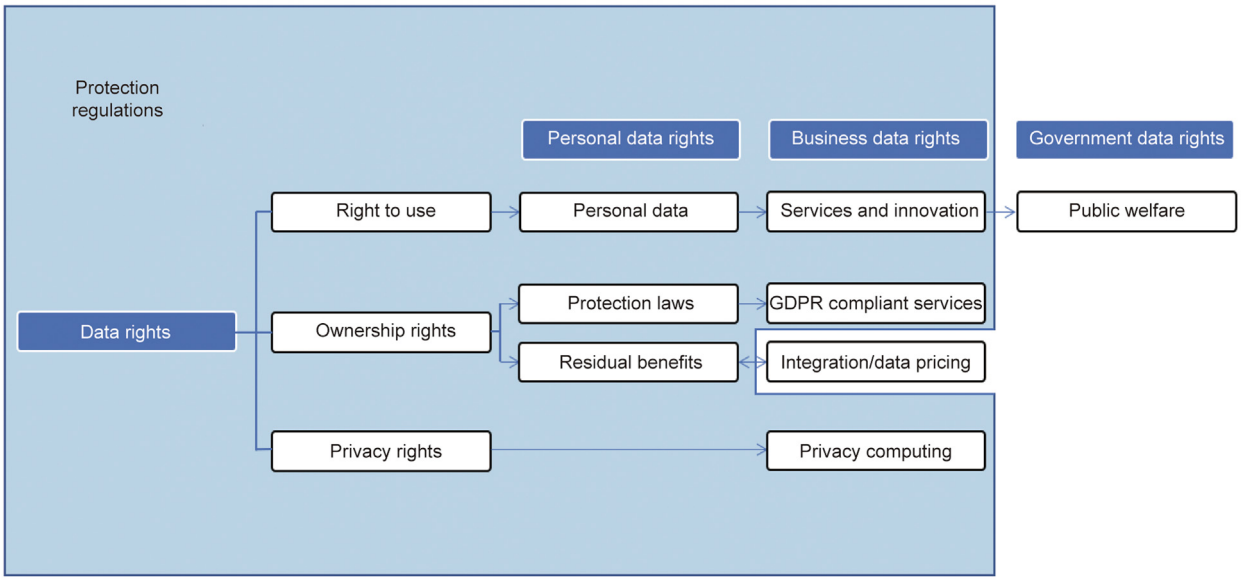

Fig. 2. Data rights.

Aside from studies on such methods, there are few publications on this topic so far. Federated learning [20] is a hot topic in machine learning under privacy restrictions [19], but it presents a few challenges regarding the GDPR. For example, Ginart et al. [18] attempted to address one of these challenges—namely, the right to be forgotten—with an efficient data-deletion model. Most data-regulation acts contain an individual right that is referred as the "right to explanation." More specifically, because federated learning's global model is an averaging of the local models, it becomes difficult to describe the data subject's data contribution [19].

In accordance with high-level initiatives such as the GDPR and the PIPL, many smaller scale data-rights organizations have also started to take action. For example, the UK's Chartered Institute of Marketing (CIM) has urged organizations to take action on the issue of responsible management of customer data. They ask their members to be clear when managing customer data, to show customers the benefits of sharing their data, and to show respect for customers'

data. The CIM claims that 67% of consumers would be willing to share more personal information if organizations were more open about their data usage [21]. Open communication and honesty can earn customers' trust; in other cases, large data companies such as Google and YouTube provide cheap or even free services to billions of people in order to gather data in exchange to improve their algorithms. However, if heavy regulation were to be added to such relationships, the companies' ability to provide such services and to innovate would be drastically reduced. Such heavy dependency on open access to data promotes a stronger focus on the right to data access than on people's exclusive property rights to their data [22].

Business data rights mostly refer to intellectual property and patents [13]. Atkinson [7] discusses many challenges related to data rights from a business viewpoint. He claims that the market force can ensure fairness of trade and a healthy data relationship, and that governments should only take action when anticompetitive behaviors limit innovation and harm customers. In addition, governments should take advantage of their ability to easily obtain aggregated data and should release the data for public access, so that others may use the data for innovative purposes, ultimately increasing the total public welfare. In conclusion, Atkinson is a strong believer that data should remain open by default, and that governments should only interfere while necessary.

The presence of high externality in the digital market implies that it is beneficial to integrate data rights [10,23], with whomever owns the greatest contribution in the collaboration coming out in the lead in the integration. With such an integration, *ex ante* negotiations can also be made to provide compensation for other parties. In order to further reduce negotiation costs, firms may even choose to approximate the payoff allocation with a matching game such as a least core or a nucleolus [24].

3. Data pricing

A fair and effective data-transaction market can guide the rational distribution of data factors, so as to promote the rapid flow of various resource elements, accelerate the integration of various market entities, help market entities restructure their organizational models, achieve cross-border development, break time and space constraints, extend the industrial chain, and smooth the economic cycle among countries. As a key issue in the process of data transaction, data pricing will receive an increasing amount of attention. With the popularity of mobile terminal devices, an ever-increasing amount of end-to-end personal information or personal data is being produced, endowed with certain property attributes. Data processors can use these data to train models and obtain commercial benefits from them. As the owner of data assets, individuals should be compensated for the use of their data. Fair and effective pricing strategies for various data products are essential in order to motivate data owners to provide high-quality data and data processors to mine more information, and then to optimize the allocation of data factors in the digital economy.

At present, research on and applications of data pricing are still in their infancy; here, we provide a review of data-pricing research based on three typical data-transaction scenarios, as shown in Fig. 3.

3.1. Single data owner, multiple data buyers

In this scenario, a company collects data and organizes it into databases as the data owner in data transactions. Multiple customers then directly purchase certain data from the company. The pricing strategies of the company must meet the various demands from the customers. For this scenario, direct data pricing is adopted, which refers to a pricing strategy that is based on the dataset itself. This pricing strategy is generally determined by the inherent factors of the original data, such as data quality, data quantity, and so forth. A typical example of direct data pricing is query-based pricing, which is based on the number of queries or data items involved. Intuitively, a data seller may treat a view of a dataset as a version. Prices on views should be appropriately set to avoid arbitrages or less than highest prices. Koutris et al. [25] transformed the query-based pricing problem into a network flow problem and automatically derived the price of any query from the price of a given set of views. After that, they adopted a different perspective, which transformed the query-based pricing problem into optimized integer linear programming, priced the structured query language (SQL) query based on the price point specified by the seller, and used the query history to avoid the repeated charging of overlapping information [26]. Deep and Koutris [27] proposed a real-time pricing system supporting various pricing functions, which can effectively calculate the price of large-scale SQL queries.

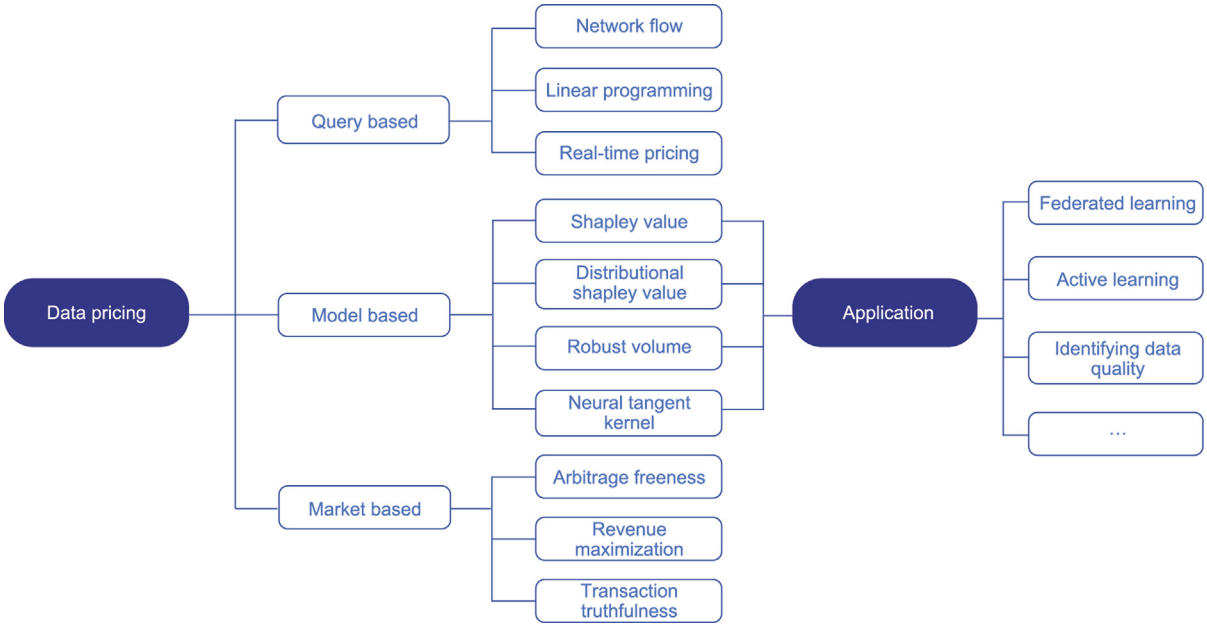

Fig. 3. Data pricing.

3.2. Multiple data owners, single data buyer

In the case of multiple data owners and a single data buyer, a large amount of data is produced by individuals and stored in their mobile terminal devices. A data processor that utilizes this data to train models must compensate the data owners for their data usage. The pricing strategy for a data processor should fairly evaluate the contribution made by various data owners to the model training. For this scenario, model-based pricing is adopted. Model-based data pricing is a data-pricing strategy based on the model obtained through dataset training. This pricing strategy is generally determined by the contribution of different data to the model training. For deep learning models, a single piece of data often does not directly assist in the training of models, as a deep learning model must learn from a dataset composed of a large amount of data. That is, for a deep learning model, it is often difficult to directly measure the contribution of a single piece of data, and the contribution of data can only be reflected in combination with other data. Therefore, such pricing strategies often determine the contribution of data to model training by calculating the marginal contribution of data. Such methods, which calculate data contribution through models, are often referred to as "data valuation" in the field of machine learning.

Data valuation can be implemented through a variety of techniques, such as leave-one-out [28], leverage or influence score [29], and reinforcement learning [30]. The Shapley value [31], a classic notion in cooperative game theory, is a well-known data valuation approach that benefits from its profound theoretical background. In cooperative game theory, Shapley gave a definition of fair revenue distribution [31]. Suppose that there are k agents cooperatively participating in a game that leads to a payment v (where k represents the number of agents and v represents game payment). We denote D as the complete set of k agents, $V(S)$ as the coalition revenue for $S \subset D$ (where S is the coalition of some agents), and ϕ_i as the valuation for agent i. To obtain fair allocation for each agent, the following four axioms should be satisfied:

(1) **Efficiency.** For the complete set D, $\sum_{i\in D}\phi_i = V(D)$. In other words, the sum of the payment to each agent should be equal to the full payoff.

(2) **Symmetry.** For any subset $S \subset D - \{i, j\}$, if for agent i and j, we have $V(S\cup\{i\}) = V(S\cup\{j\})$, then $\phi_i = \phi_j$ (where ϕ_j represents the valuation for agent j). In other words, if agent i and j always contribute the same amount to every coalition with the other agents, they should receive the same payments.

(3) **Zero element.** For any subset $S \subset D - \{i\}$, if $V(S\cup\{i\}) = V(S)$, then $\phi_i = 0$. We call this kind of agent the "zero element." In other words, if agent i does not contribute to any coalition with the other agents, it should not receive any payment.

(4) **Additivity.** For any two different coalitional games involving the same set of agents D but defined by two different coalition revenue functions V_1 and V_2, for any agent i, we have $\phi_i(V_1 + V_2) = \phi_i(V_1) + \phi_i(V_2)$.

The Shapley value is a unique payoff division that satisfies efficiency; it divides the full payoff v of the complete set D and the symmetry, zero element, and additivity axioms. The Shapley value of agent i is given by the following:

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

The Shapley value averages over all the different arrangements of the complete set D, thereby capturing the average marginal contribution of agent i.

However, when calculating the data contribution in a machine learning model using an exact Shapley value, there will be many problems. For example, such a calculation requires exponential model evaluations with regard to data quantity, as it is necessary to obtain the coalition revenue for every data subset. The existing approaches, which address problems of contribution calculation in the field of machine learning, can be divided into two categories: approaches that focus on the optimization of fair revenue distribution and approaches that focus on the design of coalition revenue functions.

Earlier approaches typically focus on the optimization of fair revenue distribution. To address the exponential complexity of computing the Shapley value, Ghorbani and Zou [32]—who first introduced the Shapley value to equitable data valuation in supervised machine learning—used Monte Carlo and gradient-based methods to efficiently estimate the Shapley values of data. Jia et al. [33] introduced a method that allows the computation of exact Shapley values on a K-nearest neighbors (KNN) model in $O(k\log k)$ time, compared with the exponential complexity of computing the exact Shapley value by definition. To address the stability problem of the data Shapley value, which provides no guarantee of consistency between the data Shapely value and the value of the data computed using a different dataset, Amirata et al. [34] proposed the distributional Shapley, where the value of a point is defined in the context of an underlying data distribution to improve the statistical interpretation of the Shapley value; this can evaluate the data value of different distributions. This work was further improved by Kwon et al. [35], who derived analytic expressions for a distributional Shapley and interpretable formulas in order to efficiently estimate the distributional Shapley in linear regression and binary classification problems.

Recent approaches have begun to focus on the design of coalition revenue functions. Earlier approaches typically use the classification accuracy of a model trained on a certain dataset as the coalition revenue function of that dataset. However, this coalition revenue function relies on obtaining validation performances of converged models, which is computationally costly for large complex models such as deep neural networks (DNNs), due to their inevitable long-term model training. In addition, a validation set may not be available in practice, and it can be challenging for data providers to reach an agreement on the choice of the validation set. Recent approaches [36,37] use efficient techniques to estimate the fully converged performances of large complex models as coalition revenue functions in data contribution computation. More specifically, with robust volume Shapley value (RVSV), Xu et al. [36] adopts a perspective in which the value of data is determined by the intrinsic nature of the data, and the volume of a dataset is proposed as a coalition revenue function. The volume of a dataset is defined as the determinant of its left Gram matrix, as follows:

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

where $\boldsymbol{X}$ is data matrix, $\boldsymbol{G}$ is the left Gram matrix of $\boldsymbol{X}$, and Vol is the volume of $\boldsymbol{X}$.

Compared with using validation performances as a coalition revenue function, the computation complexity is lower, and the data valuation is not limited by models and tasks. Furthermore, RVSV uses a robust volume measure that theoretically ensures the replication robustness via direct data copying. The robust volume typically discretizes the data space into a set of d-cubes (where d is the dimension of data space) and merges data points in the same d-cube as their statistic (e.g., mean vector) to ensure the robustness via direct data copying, as copied data is merged in the same d-cube. With RVSV, Xu et al. [36] theoretically proved the suitability of volume and robust volume as coalition revenue functions for linear models and one-dimensional cases. However, this theoretical guarantee cannot be generalized to nonlinear models or high-dimensional cases. Moreover, an empirical

demonstration may cause a problem when applied to complex deep learning models, which are typically nonlinear and high dimensional. With data valuation at initialization (DAVINZ), Wu et al. [37] introduced statistical learning theory (SLT) to estimate the fully converged performances of DNNs as coalition revenue functions, which completely avoids the need for model training. More specifically, DAVINZ derives a domain-aware generalization bound by introducing a domain discrepancy into the recent neural tangent kernel (NTK) theory. The NTK matrix $\boldsymbol{\Theta} \in \mathbb{R}^{m\times m}$ of a DNN model $f(x,\theta)$ on the dataset of size m is defined as follows:

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

where x and x' denote data points in dataset. θ is the parameters of the DNN model. Recent studies on NTK theory have shown that the generalization errors of DNNs can be theoretically bounded using the NTK matrix with initialized model parameters. In addition, NTK can characterize the training dynamics of any reasonable architecture DNNs with gradient descent. DAVINZ utilizes the properties of NTK, which can estimate the performances of DNNs using only the initialized model parameters, and uses a generalization bound derived by NTK as coalition revenue functions, without any model training. Compared with RVSV, DAVINZ utilizes SLT, which is theoretically and typically more reasonable for deep learning models. On the other hand, DAVINZ's utilization of the upper bound of the validation performance as the coalition revenue functions may cause more errors, compared with an exact estimation.

Contribution calculation methods represented by the Shapley value have various applications in the field of machine learning. The Shapley Q-value [38] utilizes the Shapley value in multi-agent reinforcement learning to estimate the contribution of each agent to a global reward. Wang et al. [39] proposed the Shapley flow, which uses the Shapley value to calculate the credit assigned to the edges of a causal graph in order to reason about the impact of a model's input on its output. Ghorbani et al. [40] used the Shapley value to annotate unlabeled data in order to increase the efficiency of batch active learning while preserving performance effectiveness. Fan et al. [41] proposed a completed federated Shapley value for fair data valuation in federated learning. Xu et al. [42] designed a novel training-time gradient reward mechanism in federated learning that distributes gradients of different quality to local clients according to the contribution calculated by the cosine gradient Shapley value (CGSV) in each round. Gradients of different quality are obtained by different percentage masks to parameters. Furthermore, there is already some work using the Shapley value for real scene data valuation. Tang et al. [43] used a Shapley value to calculate the value of training data in a large chest X-ray dataset, which provides a framework for using a Shapley value to estimate data valuation for large-scale datasets.

3.3. Multiple data owners, multiple data buyers

In this scenario, the multiple data owners comprise various data subjects, from individuals to data companies and governments. The data transactions involve data itself and data products such as models trained from data. Data brokers are usually necessary for the complex transactions between various data owners and data buyers. Existing studies on this scenario usually consider market information in their data-pricing models. We denote these pricing strategies as market-based pricing. Market-based data pricing is a data-pricing strategy based on the supply and demand relationship and other information in the data market. The formulation of this pricing strategy typically depends on a tripartite game model established by data owners, data buyers, and data brokers in the data market. Here, we summarize the functions of data owners, data buyers, and data brokers in the data market.

Data owners are the providers of the source data; to a certain extent, they undertake the function of integrating and processing the source data into data products that can be traded in the data market. Data owners provide data to the data brokers, with different requisites on privacy preservation, and receive corresponding compensation for their data usage, allocated by the brokers.

Data buyers are the final purchasers of data products. Data products refer not only to data itself but also to information obtained from data mining or models learned from data training. Data buyers purchase data products of different quality according to their own needs and budgets. Data products of different quality can typically be obtained by adding different levels of noise to model parameters or training data.

Data brokers provide pricing models and corresponding technical support for different categories of data products. When making market decisions, data brokers must design reasonable and fair compensation functions for data owners and arbitrage-free price functions for data buyers to achieve the objective of revenue maximization.

For a multi-party game model constructed by data owners, data brokers, and data buyers in the data market, data brokers should provide compensation for data usage to data owners and formulate price functions to meet the needs of data buyers. To formulate these functions, Niu et al. [44] studied noisy aggregate statistics trading from the perspective of a data broker in a data market and proposed the pricing model, which enables aggregate statistics pricing over private correlated data and considers dependency fairness among data owners. Chen et al. [45] first proposed a formal framework of model-based pricing in a data market by focusing on avoiding arbitrage and provided algorithmic solutions on how the data brokers can assign prices to models to achieve the objective of revenue maximization. More specifically, for a machine learning model with a strictly convex loss function, the researchers added Gaussian noise to the model parameters to realize arbitrage-free pricing. Liu et al. [46] and Lin et al. [47] also adopted the perspective of model-based pricing and proposed the framework Dealer, which uses differential privacy (DP) to build several different model versions, adopts dynamic programming algorithms to formulate pricing strategies in order to achieve revenue maximization, and applies a Shapley value for the fair distribution of revenue to data owners. Zheng et al. [48] proposed a pricing framework by considering the bounded personalized DP of each data owner and demonstrated that the arbitrage-freeness constraint can be reasonably relaxed under bounded utilities by partial arbitrage freeness.

To design a pricing strategy, data brokers must inevitably access the data from data owners *ex ante*, which is unfair to data owners, as data brokers may obtain information from accessing the data without compensating the data owners. It is important to verify whether the data brokers have truthfully collected and processed data. One direct solution is to encrypt sensitive information when setting up a data marketplace, such as truthfulness and privacy preservation in data markets (TPDM) [49]. Another solution is to make the brokers price the data without obtaining the data through privacy computing technology. However, this solution introduces a fair transaction problem: The data owner can provide high-quality data during pricing but provide low-quality data during data transaction. To address this problem, Zhou et al. [50] proposed a new notion named zero-knowledge contingent model payment (ZKCMP), which allows the fair exchange of a trained machine learning model and a cryptocurrency payment.

4. Privacy computing

Privacy computing is a combination of a series of cryptographic computing techniques, as shown in Fig. 4. It involves advanced

mathematics, computer science, cryptography, network communication technology, and other disciplines (i.e., secure multi-party computation, DP, homomorphic encryption, zero-knowledge proof, TEE). It is the bridge between data factors and data value, and the basis of maturity for the digital economy and the data-factor market. By leveraging privacy computing technologies, data becomes available yet invisible.

Data privacy breaches are happening all over the world. For example, in 2018, Cambridge Analytica [51] allegedly stole information from Facebook users to manipulate the US election and the UK referendum on the EU. The various privacy breaches show that research on data privacy protection is extremely necessary to fully exploit the value of data. That said, laws and regulations related to data privacy have become increasingly mature and complete in recent years, both domestically and globally. For example, both the EU's GDPR and China's Data Security Management Measures set out responsibilities and norms regarding the protection of personal information privacy. Overall, privacy computing is the key to achieving data privacy and security.

In terms of practical deployment, each privacy computing technique has its own features, advantages, and disadvantages. According to the deployment scenarios, security requirements, and efficiency requirements, it is necessary to choose the most suitable privacy computing techniques for each application. In privacy computing, the key questions are as follows:

(1) Who owns the data?

(2) Who consumes the data and data derivatives?

Clearly, when the data is owned by a single party that uses the data itself, privacy computing is not needed. Therefore, in this section, we are interested in scenarios in which the data owner(s) and data consumer(s) are mutually untrusted entities.

4.1. Single data owner, multiple data buyers

In the setting of a single data owner and multiple data buyers, the data is positioned with a single data owner who wants to delegate to a single untrusted computing node in order to compute on a joint database. The homomorphic encryption techniques mentioned before can also be used in this scenario; however, doing so might require distributed key generation and complicated key management. The computation task is sometimes too heavy for homomorphic encryption. DP [52] is a cheap privacy-enhancement technique that is rooted in cryptography and built on rigorous mathematical definitions, providing a quantitative evaluation method. The main idea of DP is to protect user privacy by removing individual features while preserving statistical features. An algorithm is called ε-differentially private [53] if the algorithm is run on two databases that differ by exactly-one entry and the resulting difference is bounded by ε. A smaller ε indicates that the algorithm can ensure stronger privacy. Informally, the closer the output obtained by an algorithm processing two similar datasets is, the better the privacy protection is for a specific piece of data. Recently, local DP (LDP) has been proposed. Instead of adding noise to the aggregated result as in DP mechanisms, LDP mechanisms add noise by each user before sending the data to the central server. Thus, users do not rely on the trustworthiness of the central server. Both DP and LDP mechanisms can be combined with machine learning. DP mechanisms can provide privacy protection by adding random noise to the objective function, gradient, and output results, such as by adding Laplace noise or Gaussian noise [53]. LDP mechanisms can be applied to protect various types of training datasets, such as item datasets [54], itemset [55], and graph [56].

4.2. Multiple data owners, single data buyer

The setting of multiple data owners and a single data buyer can be further divided into several sub-cases. When the multiple data owners are also the computing nodes, multi-party computation (MPC) is an ideal technique. Secure multi-party computation [57] is a research field that was created in 1982 when Turing Award winner Chi-Chi Yao proposed the famous millionaire's problem, which requires multiple parties to collaborate on solving a problem without revealing private data. Secure multi-party computing has received continuous attention and research investment since its inception, and new methods and tools are rapidly emerging in this area. Among them, the protocol used for secure two-party computation is generally garbled circuit (GC) [57] combined with oblivious transfer (OT) [58], while the protocol used for secure MPC (i.e., three or more parties) is generally secret-sharing (SS) combined with OT. The main problem with the former (i.e., GC + OT) is that the computational overhead can be higher, although fewer communication rounds are required. The latter (i.e., SS + OT) [59] usually requires multiple iterations of OT and a large number of communication rounds, although its computational overhead is smaller.

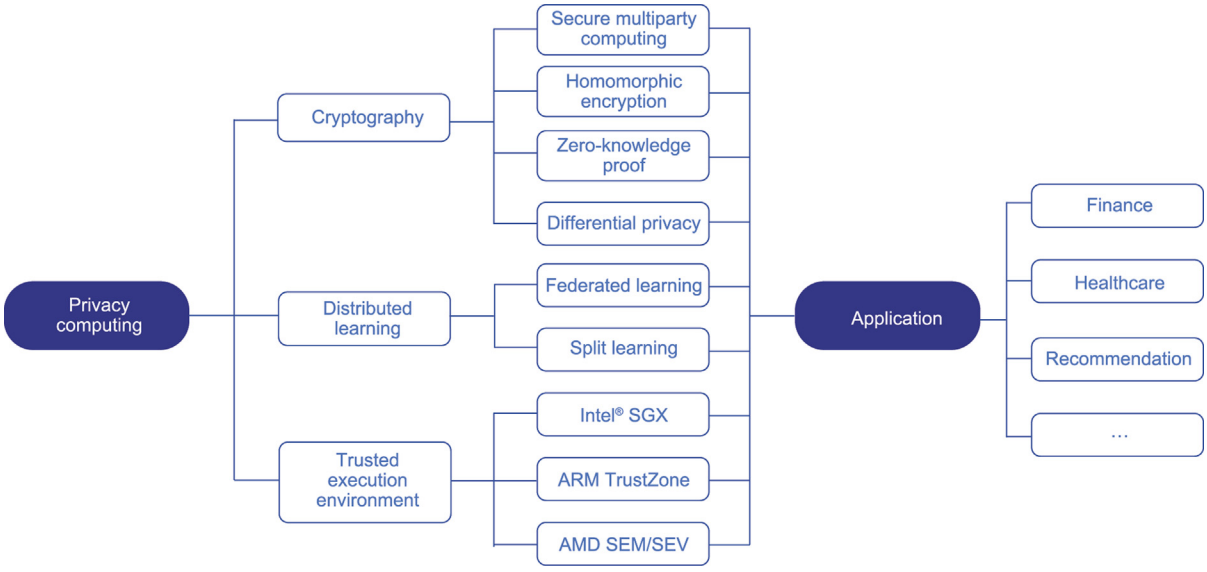

Fig. 4. Privacy computing. SGX: software guard extensions; ARM: Advanced Reduced Instruction Set Computer Machine; AMD: Advanced Micro Devices; SEM: secure encrypted memory; SEV: secure encrypted virtualization.

In terms of model training, conventional MPC typically requires extremely large amounts of communication. Instead, collaborative learning can be adopted for better efficiency. Collaborative learning is a class of MPC protocols that aims to train a model on data owned by multiple parties who want to keep their data private. Federated learning and split learning are two of the more important frameworks in collaborative learning. In federated learning [60], the central server distributes the current model to the participants. Each participant trains the model using its own local data and then uploads the model to the server for aggregation. This process is repeated until the model converges. This technology concept was first introduced by Google in 2016, when it proposed federated learning for mobile terminals. Since then, WeBank has proposed the first "federated transfer learning" [61] solution for the financial industry, combining transfer learning and federation learning. At present, various open-source federated learning frameworks such as Federated AI Technology Enabler (FATE) and TensorFlow Federated are continuing to emerge and mature in the field of AI.

In a practical application scenario, assume that N users $\{U_1, \ldots, U_N\}$ hold their own datasets $\{D_1, \ldots, D_N\}$, which are not directly accessible to other users. Federated learning is used to learn a model by collecting training information from distributed devices. It consists of three basic steps:

(1) The server sends the initial model to each device.

(2) Device U_i does not need to share its own resource data, but can federally train its own model W_i (W_i is the local model of U_i) with local data D_i.

(3) The server takes the individual local models collected, $\{W_1, \ldots, W_N\}$, aggregated into a global model W', and downlinks the global model to update the local model for each user.

With the rapid development of federation learning, the efficiency and accuracy of federation learning models are getting closer to those of centrally trained models. Based on the different distribution patterns of the sample space and feature vector space of the data, federal learning can be divided into three categories: horizontal federated learning, vertical federated learning, and federated transfer learning. Horizontal federation learning is suitable for scenarios in which the user feature vectors of two datasets overlap a great deal, but the users rarely overlap. In other words, different rows of data have the same feature vector (aligned in the feature vector dimension). Therefore, horizontal federation learning can increase the user sample size. For example, Kim et al. [62] proposed a horizontal federated learning framework called BlockFL, in which each mobile device uses a blockchain network to update the local model, and Smith et al. [63] proposed a federated learning approach called MOCHA to address security issues in multi-tasking, which allows multiple clients to work together to complete tasks and ensure privacy and security. Multi-task federated learning also improves the communication cost of the original distributed multi-task learning and enhances the fault tolerance.

Vertical federation learning is suitable for scenarios in which the user feature vectors of two datasets rarely overlap, but the users overlap a great deal. Therefore, vertical federation learning can increase the dimensionality of the feature vectors of the training data. For example, Cheng et al. [64] proposed a vertical federated learning system called SecureBoost, in which the parties combine user feature vectors to train together in order to improve the accuracy of decision-making, in what is a lossless training scheme. Hardy et al. [65] proposed a vertical federated learning-based logistic regression model with privacy protection. The model uses pipelined entity analysis with Paillier semi-homomorphic encryption for distributed logistic regression, which can effectively protect privacy and improve the accuracy of the classifier.

Federal transfer learning is applicable to this scenario: The users and user feature vectors of both datasets do not overlap much, but transfer learning can be used to overcome the lack of data and labels. The most appropriate situation for migration learning is when you try to optimize the performance of a task but do not have enough relevant data to put into training. For example, it is difficult for a hospital radiology department to collect many X-ray scans to build a good diagnostic radiology system. Transfer learning can make it possible to learn a diagnostic radiology system in combination with other related and different tasks (e.g., image-recognition tasks). Through federal migration learning, we can not only protect data privacy but also migrate the model of the auxiliary task to the target model learning, thereby solving the problem of small data volume.

While federal learning emphasizes splitting at the data level, the core idea of split learning [66,67] is to split the network structure. In the simplest example of split learning, the network structure is split into two parts—one stored on the client side and the other on the server side. The client has no access to the server-side model and vice versa. Compared with federated learning, split learning reduces the amount of computation on the client side.

4.3. Multiple data owners, multiple data buyers

In this setting, the data is owned by multiple entities, and there will be more than one data consumer. Such scenarios often require the involvement of a data broker(s), and data privacy is the fundamental requirement. Homomorphic encryption [68] is an ideal technique in this setting. Homomorphic encryption is a form of encryption that allows users to perform computations on their encrypted data without decrypting them. The results of these computations are stored in encrypted form and, after decrypting the results, the output is identical to the results obtained by performing the same operation on unencrypted data. Common types of homomorphic encryption include partially homomorphic [69], somewhat homomorphic [70], leveled fully homomorphic [71], and fully homomorphic encryption [68]. Since IBM scientist Gentry constructed the first true fully homomorphic cryptography method [68], the cryptographic community has conducted intensive research in this area. The second [72,73], third [74], and fourth [75] generations of fully homomorphic cryptosystems have been created.

A TEE [76] can also be used as an efficient solution in this setting. TEE protects data in isolation through hardware technology. In TEE-enabled central processing units (CPUs), a specific enclave can be created that acts as a secure content container for sensitive data and the code for its application, ensuring their confidentiality and integrity. Even if an attacker takes control of the operating system and other privileged-level software, the enclave cannot be accessed (i.e., the information cannot be modified nor read). Applications running on the TEE are called trusted applications; they are isolated from each other and cannot read and manipulate the data of other trusted applications without authorization. Clearly, isolation achieved through software algorithms and hardware technologies ensures that private information can be securely computed, stored, transmitted, and deleted. TEE technologies are often dependent on the specific technology platform and implementation vendor; common technologies include Intel® software guard extensions (SGX), Advanced Reduced lnstruction Set Computer Machine (ARM) TrustZone, and Advanced Micro Devices (AMD) secure encrypted memory (SEM)/secure encrypted virtualization (SEV).

In addition, many other verifiable computing techniques can be adopted to ensure computation integrity. The zero-knowledge proof technique is a widely used solution for verifiable computing. In this proof system, the prover knows the answer to a question

and must prove to the verifier that "he or she knows the answer," but the verifier cannot obtain any other information besides the fact that "he or she knows the answer." Zero-knowledge proofs [77] were first conceived in 1985 by Shafi Goldwasser, Silvio Micali, and Charles Rackoff in their paper "The knowledge complexity of interactive proof systems" [78]. Subsequently, zero-knowledge proof technology continued to evolve until 2013, when cryptographers created the first efficient and commercially available general-purpose succinct non-interactive zero-knowledge proof protocols: zero knowledge succinct arguments of knowledge (zk-SNARKs).

5. Challenges and open questions

In this section, we discuss some interesting unexplored challenges for possible future work. We hope this discussion will invite more extensive interest and research efforts into this fast-growing area.

5.1. Appropriate technical solutions to ensure data rights

Recently, machine learning models have been widely used for data processing. Although such models, which are completed by training, can work independently from the data used for training, they must still meet the requirements of the data subjects. The black-box characteristic of machine learning models makes it challenging to ensure various data rights. For example, the right to be forgotten is a right of a data subject that has been identified by GDPR. The data subject has the right to obtain from the controller the erasure of his or her personal data without undue delay. Unlike traditional databases, the corresponding data can be deleted directly. However, making a machine learning model forget the learned data is a nontrivial challenge. In right of access, data subjects have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her is being processed. In order to prevent the data shared on the Internet from being illegally crawled for model training, corresponding technical solutions are needed to make the data become unlearnable examples that are visible but unexploitable. The complicated model and data dependence of machine learning present a major challenge in ensuring data rights.

5.2. The combination of data pricing and privacy computing

Data pricing provides a technical solution for ownership benefits in the process of data transaction and circulation, while privacy computing provides a technical solution to protect privacy in the process of data transaction and circulation. Data pricing and privacy calculation complement each other in the process of data transaction. Recently, the training of machine learning models in a distributed scenario has become a research highlight. In this scenario, data transactions continue to occur in the training process of the machine learning model. This situation requires the design of real-time and efficient data-pricing and privacy-computing technologies that conform to the training of machine learning models in a distributed scenario. In this paper, we review data-pricing technology based on model pricing and privacy-computing technology based on federated learning in a distributed scenario. An example is the new pricing mechanism proposed by Xu et al. [42] along with the privacy-computing technology of federated learning, which compensates data owners through the mechanism of federated learning. We believe that, for a distributed scenario, challenges come from the combination of data-pricing and privacy-computing technologies. Efficient and fair pricing strategies should be designed by utilizing privacy-computing technologies, such as through the relevant mechanism of federated learning.

5.3. Data-factor computing that conforms to the practical situation of the data-transaction market

A practical data-transaction market contains diverse types of data transactions, and the form of the data products ranges from direct data to machine learning models obtained through data training. Data-factor computing should be based on practical transaction types in order to enable transaction completion. The data-transaction market is complex and changes with supply and demand information. Data-factor computing provides an explanation of the market, guides each subject of the market in making judgments, ensures the rights of the subject, stabilizes market prices, protects data privacy, and enales data transactions. Research on data-factor computing should not only build a model based on data science but also include a combination of market mechanisms and user behaviors in order to conform to the practical situation of the data market. It is essential to conduct data-factor computing from an interdisciplinary perspective, covering data science, economics, and marketing.

6. Conclusions

In the era of big data, big data governance has become a widespread concern in all sectors of society, and appropriate algorithmic methods are needed to ensure the circulation and transaction of big data. This paper provided an overview of data-factor computing in the data-transaction market system and reviewed the three main issues of data transaction: data rights, data pricing, and privacy computing. We also discussed interesting challenges for possible future work, in the hope that our discussion will invite more extensive interest and research efforts into this fast-growing area.

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-worlds-most-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, Hamdoun S. Privacy, data rights and cybersecurity: technology for good in the achievement of sustainable development goals. In: Proceedings of 2019 IEEE International 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. Communication-efficient 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 volume-based 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 large-scale 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 end-to-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; San Diego, 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 privacy-preserving 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 set-valued 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; Las Vegas, 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 IEEE Trustcom/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.