A Survey on Data Pricing: From Economics to Data Science

Jian Pei, *Fellow, IEEE*

Abstract—Data are invaluable. How can we assess the value of data objectively, systematically and quantitatively? Pricing data, or information goods in general, has been studied and practiced in dispersed areas and principles, such as economics, marketing, electronic commerce, data management, data mining and machine learning. In this article, we present a unified, interdisciplinary and comprehensive overview of this important direction. We examine various motivations behind data pricing, understand the economics of data pricing and review the development and evolution of pricing models according to a series of fundamental principles. We discuss both digital products and data products. We also consider a series of challenges and directions for future work.

Index Terms—Data pricing, data economics, data science, privacy, digital product, data product, data market.

1 Introduction

In this digital economics era, data are well recognized as an essential resource for work and life. Many products and services are delivered purely in digital forms. Many big data applications are built on the second use or reuse of data [1], that is, the same data are customized and reused by many applications for different purposes. The extensive sharing and reusing data has profound implications to economy. For example, digital maps are often produced for traffic and directions as the immediate usage. However, Nagaraj [2] finds that mining activities were strongly benefited by open maps or maps sponsored by governments, particularly for smaller firms with less resources. Universal availability of data often helps minority parties and emerging initiatives.

In economic activities where data are shared, exchanged and reused, it is essential to measure the value of data properly. While there exist many possible ways to appreciate and represent the value of data, a general approach that can be scalable for massive applications and acceptable to many parties is to set a price at which data can be sold or purchased, that is, data pricing. The importance of pricing in business is well recognized in financial modeling [3], as price being one of the four Ps of the marketing mix.[1]

Pricing data is far from trivial. Data have many different aspects. Consequently, the term "price of data" may carry different meanings and refer to different properties of data. To illustrate the complexity, let us quickly consider three scenarios involving price information related to data.

- *Data transmission.* Imagine the scenario where a mobile service provider offers a smart phone user the price of its data package. Here, the price is quoted for the data transmission service and is decided by several factors, such as the amount of data the user wants to transmit in a month time, the location (roaming or not, for example), and the transmission speed. The price does not include and is independent from the content, that is, what the data are about, such as data quality, and how the data are collected, stored or processed.
- *Digital products.* Imagine that a person wants to watch a movie at home. This is a purchase of data, since the movie is sent to the customer's home as a stream of bits. The price here typically is related to the content, but is independent from the data transmission service, that is, how the data are transmitted to the user's home.
- *Data products.* Many logistics companies want to pay for weather information to support their business operations. While historical data are relevant, more often than not those companies want to subscribe to weather forecasting information instead. Some companies may want weather predictions at a higher granularity while some may want detailed predictions at specific locations. Moreover, some may want long term predictions while some others may want short term projections. Here, prediction services are sold as data products.

The above three cases just elaborate some representative scenarios where data prices are used, and are by no means exhaustive. To appreciate data pricing, including ideas, principles and methods, we have to take an interdisciplinary approach from multiple fields, economics and data science being the two most prominent. Indeed, the studies and practice of data pricing started as early as the dawn of digital economics, and are highly diversified and rich in innovative thinking.

In this article, we try to present a comprehensive survey on data pricing, an emerging research and practice area that plays a more and more important role in the current big

1. The four Ps are product, price, place and promotion [3].

- *The author is with the School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. E-mail: jpei@cs.sfu.ca.*

Manuscript received 10 Sept. 2020; revised 26 Nov. 2020; accepted 15 Dec. 2020. Date of publication 21 Dec. 2020; date of current version 12 Sept. 2022. (Corresponding author: Jian Pei.)
Recommended for acceptance by L. Chen.
Digital Object Identifier no. 10.1109/TKDE.2020.3045927

data and AI economics era. Our survey is highly related to the current strong rising of data science. To a large extent, data pricing is an overdue pillar in data science research and practice.

Data and information as goods discussed in this article are those that are distributed purely in digital form. We focus on two categories of the most interest: pricing digital products and pricing data products, demonstrated by the last two aforementioned scenarios, respectively. In this article, *digital products* refer to those intangible goods but can be consumed through electronics, such as e-books, downloadable musics, online ads, and internet coupons. Many digital products have physical correspondences in one way or another, though not absolutely necessary. *Data products* refer to data sets as products and information services derived from data sets. We build the linkage between these two categories by pointing out many ideas and methods on pricing digital products can be generalized and applied to pricing data products. In some scenarios, the boundary between digital products and data products is also blurry. Hereafter, we use the term *information goods* to refer to both digital products and data products.

1.1 Related Surveys

The research into data pricing happens simultaneously in multiple domains, including but not limited to economics, marketing, e-commerce, databases and data management, operational research, management science, machine learning and AI. However, to the best of our knowledge, there exists very limited effort to provide an interdisciplinary survey of the related work. This article presents our endeavor to produce a comprehensive picture.

There are some previous surveys related to data pricing. For example, Liang *et al.* [4] survey the life cycle of big data, and reviews 11 data pricing models. They also discuss data trading and protection. Fricker and Maksimov [5] report a literature survey over 18 research articles regarding several research questions, including maturity of the pricing models. Very recently, Zhang and Beltrán [6] review the state-of-the-art data pricing methods. They categorize data pricing methods according to two important data properties, granularity and privacy, This article covers a substantially broader scope than those [4], [5], [6]. We connect economics, digital product pricing and data product pricing. We also discuss a series of desirable properties in data pricing, including arbitrage-freeness, revenue maximization, fairness, truthfulness, and privacy preservation, and review the techniques achieving those properties.

Data pricing is related to cloud pricing, since a lot of data for pricing and trading are hosted on cloud. Wu *et al.* [7] present a comprehensive survey on cloud pricing models. Data pricing is selling data, while cloud pricing is selling cloud resources (e.g., storage and computation), including physical resources, virtual resources and stateless resources. In addition, Sen *et al.* [8] survey the major broad-band pricing proposals, including the realizations in various consumer data plans around the world. Murthy *et al.* [9] list different pricing models and pricing schemes used by some popular IaaS (infrastructure-as-a-service) providers. Wu *et al.* [10] propose pricing as a service, which is essentially a personalized pricing service for IaaS. Aazam and Huh [11] propose broker as a service, which matches cloud services among cloud service providers and users. The key idea is to predict resource demands and thus derive prices.

As data are often hosted online, one interesting question is the fair sharing of the cost among data owners, data users and brokers. This is related to data pricing, because the costs of data hosting and processing have to be recovered from data pricing. Kantere *et al.* [12] study the fair allocation of costs in query services. They develop a stochastic model, which predicts the extent of cost amortization in time and number of services based on query traffic statistics. The model can be implemented on top of a cloud DBMS. Al-Kiswany *et al.* [13] provide a cost assessment tool to evaluate the cost of a desired data sharing. One useful feature of the tool is that a user can explore the cost space of alternative configurations using various factors, such as quality, staleness, and accuracy. The technique is based on what-if analysis.

1.2 Structure of This Survey

We take a multi-disciplinary approach in this survey. The rest of the article is organized as follows.

In Section 2, we start from economics and focus on two aspects. First, we discuss cost reduction in information goods that contributes to their prices and has impact on economics. Then, we discuss the differences between digital products and data products.

In Section 3, we discuss the fundamental principles of data pricing. We first present versioning as a general framework for pricing information goods. Then, we identify several desirable properties in data pricing, including truthfulness, fairness, revenue-maximization, arbitrage-freeness, privacy preservation and computational efficiency.

In Section 4, we discuss pricing digital products. We first review the three major streams of revenues for digital products. Then, we revisit the bundling and subscription planning pricing models. Last, we consider auctions, which are widely used in pricing digital products.

In Section 5, we discuss pricing data products. We first overview the structures, players, and ways to produce data products in data marketplaces. Then, we examine several important areas in pricing data products, including arbitrage-free pricing, revenue maximization pricing, fair and truthful pricing, privacy preservation in pricing. We also discuss dynamic data pricing, online pricing, and pricing in federated and collaborative learning.

Last, in Section 6, we discuss challenges and future directions.

2 Economics of Data Pricing

In general, *pricing* is the practice that a business sets a price at which a product or a service can be sold. Pricing is often part of the marketing plan of a business. To set prices, a business often considers a series of objectives, such as profitability, fitness in marketplace, market positioning, price consistency across categories and products, and meeting or preventing competition. Some major pricing strategies in literature [14], [15], [16], [17], [18] include operation-oriented pricing, revenue-oriented pricing, customer-oriented pricing, value-oriented pricing, and relationship-oriented pricing. There is a

rich body of studies in economics and marketing research on pricing tactics, which are far beyond the scope and capacity of this survey.

In this section, to understand the economic factors specific to data pricing, we examine the cost reduction in information goods. Then, we inspect the differences between digital products and data as products.

2.1 Cost Reduction in Information Goods

"Technology changes. Economic laws do not." [19] The production, distribution, and consumption of information goods, comparing to those of physical products in the long history of human economies, are distinguished by significant cost reductions on five aspects, namely search costs, production costs, replication costs, transportation costs, and tracking and verification costs. Essentially, digital and data economics investigates how standard economic models adjust when those costs are reduced dramatically. Goldfarb and Tucker [20] present a thorough discussion, whose framework is largely followed here.

2.1.1 Search Costs

"*Search costs* are the costs of looking for information" [19], which are incurred in any information collection activities. Information goods allow more effective and efficient online search. The consequent low search costs facilitate users' discovering digital products and data sets, as well as comparing prices of similar products and services. For example, Brynjolfsson and Smith [21] show that online prices of books and CDs are clearly lower than offline, though the price dispersion, however, does not shrink accordingly.

Low search costs facilitate the sales of rare and long tail products [22], [23]. Thus, more variety is often observed in information goods and services. The degree of variety may be heavily impacted by recommender systems. Specific to consumption of media, one of the major categories of digital products, Gentzkow and Shapiro [24] show that online media consumption is more diverse than offline. At the same time, customers may tend to consume more that aligns more or less with their viewpoints, which is called the "echo chamber" effect [25].

Low search costs give strong rise to the prevalent platform businesses, which provide extensive matching services to customers and improve trade efficiency [26]. Interoperability, compatibility and standards are strategic tools for both building platforms and running platform businesses [27].

2.1.2 Production Costs

Producing digital products, such as online courses, eBooks, software, graphics and digital arts, and photography, is very different from manufacturing physical products, like bread, shoes, and jackets. Moreover, collecting and processing massive data so that parts of data can be sold and can meet customers' needs is also different from traditional production. A wide spectrum of production costs in traditional products are substantially reduced in information goods.

First, some essential major costs in traditional production, such as materials, semi-finished products and their transportation, are dramatically reduced in producing information goods. In many cases, the costs of obtaining, producing and transporting raw materials and physical semi-finished products can be reduced to very low or can even approach zero in making information goods. Second, a substantial cost of a traditional physical product often belongs to the product itself and cannot be further reduced through sharing. The unit costs of information goods can approach zero through sharing as long as there are sufficient reuses and sales volume. Last, smart manufacturing and customer-to-manufacturing can reduce the supply chain costs in traditional physical production [28], [29]. Information goods often can reduce the costs of customization to extreme.

The substantial reduction in production cost in materials, semi-finished products, customization and sharing gives rise to a series of innovative business models, such as economics of sharing, pay-as-you-go and query-based data consumption. This also encourages innovation and long tail products that address diverse and smaller groups of potential customers.

2.1.3 Replication Costs

One distinct feature of information goods versus traditional products is that information goods are non-rival. That is, one customer consuming an information good does not reduce the amount or quality of the product available to other customers. The zero marginal costs and the non-rival property of information goods empower innovative opportunities and bring in new challenges.

In order to structure pricing of a large variety of non-rival information goods with zero marginal costs, bundling is often used [19], that is, multiple products are sold together at a single price. Since a large number of information goods can be bundled together without a substantial increase in cost, economically it may be optimal to bundle thousands of digital products together to meet diverse and independent customer preferences [30], [31], [32].

Due to the zero marginal costs and the non-rivalrous property, many information goods are made publicly available, such as Wikipedia[2] and open source software [33]. People contribute to open source or publicly available digital products and data to demonstrate their professional skills to potential employers. Companies support those products to complement their sales on other products.

The zero marginal costs and non-rivalrous property post challenges to copyright policies and enforcement. Waldfogel [34] shows that low replication costs, though may reduce revenue, help supplies and demands, and thus boost quality. Williams [35] shows that the protection of intellectual properties indeed has negative impact on follow-on innovation in gene sequencing.

At the same time, there are evidences showing that governments mandate "open data" may lead to data leakages and privacy breaches that affect citizens' offline welfare [36]. On the negative side, the zero marginal costs or non-rivalrous nature also ease the way for spamming [37] and online crime [38].

2.1.4 Transportation Costs

Thanks to the Internet, the costs of transporting information goods approach zero. This may imply, in many scenarios,

2. https://www.wikipedia.org

that local communities may not affect adoptions and consumptions of information goods, often known as the effect of flat world [39]. Interestingly, this is not true all the time, as some studies demonstrate that tastes may still be local in music [40] and content consumption [41].

While the physical transportation may approach zero, regulation may put sophisticated constraints on locations. For example, when Wekipedia was blocked in China in October 2005, more contributors from outside China were motivated to contribute [42]. Copyright policies may also affect the availability and consumption of information goods in different regions, such as news media [43], and thus may be reflected by price.

2.1.5 Tracking and Verification Costs

The capability of tracking users with relatively low costs is an important feature of information goods [19]. The low tracking costs give the rise to extensive personalized markets and possible price discrimination [44], [45]. Behavioral price discrimination is an immediate type, which sets prices according to customers' previous behavior. Correspondingly, if customers are well aware of the benefits of tracking information to a monopoly, they may likely choose to be privacy sensitive and hold the information [46]. Another type of price discrimination is versioning [47], which sells information at different prices to different customers using different versions. Versioning is discussed in detail in Section 3.1.

The advantage of low tracking costs also leads to the blooming businesses of personalized advertising [48]. A challenge for a company, however, is how to set prices for many advertisements that may be shown to massive customers? The same advertisement may have different prices for different customers. Auctions are often used to address the challenge [49], and can even be used to discover prices for information goods [50]. At the same time, auctions may be less useful when online marketplaces become mature [51].

The low tracking costs and the consequences, such as price discrimination, lead to serious concerns on privacy [52]. As to be discussed later in this article, whether privacy should be treated as goods and how privacy is priced are investigated [53], [54]. Moreover, privacy regulation and the impact on welfare are important topics, though they are far beyond the scope of this survey.

As a byproduct of low tracking costs, the costs of verifying identity and reputation of producers and users of information goods are dramatically lower than those in traditional scenarios. The low verification costs facilitate online transactions extensively and lower the costs of trust dramatically.

2.2 Digital Products and Data Products: Differences

While digital products and data products share a series of common ideas and methods in pricing, they are also essentially different from each other on at least four aspects.

First, the units of digital products are often well defined and fixed. For example, individual movies and musics are often priced and sold in whole. The consumption of a digital product is often independent from each other. For example, it would be rare that two digital books have to be read at the same time. In contrast, although the basic unit in a data set can be at a very small granularity, such as a record in a relational table, the units for pricing and consumption often vary from one customer to another. For example, a customer may be interested in the sales data of female customers in a province, while another customer may be interested in the sales data on electronics during the Christmas season. Correspondingly, one individual unit of data at the lowest granularity may not be valuable as a data product. For example, one customer purchase record, after proper anonymization, may not be useful for a retailer. Instead, more often than not, many basic units of data are combined, aggregated and consumed together.

Second, different from digital products, data sets as data products have very strong and flexible aggregateability. Customers often aggregate data using various dimensions. The aggregateability, on the one hand, enables many opportunities for innovations in data business, and, on the other hand, posts many technical and business challenges, such as ensuring arbitrage-freeness as to be discussed later in this article. In many business scenarios, digital products like movies and musics are bundled. However, bundles are not aggregates. Customers still get digital products and consume them individually. Bundling is to take the advantage of low replication costs of digital products to boost sales and meet customers' diverse demands [30], [31], [32].

Third, the means of consuming digital products and data products are also very different. Typically digital products are consumed directly by people, such as movies watched by people and musics enjoyed by fans. Data sets are more often than not consumed by computers. They are, for example, analyzed, summarized or used to train machine learning models. The outputs of models are used to automate operations or support human decision making.

Last, digital products and data products are dramatically different in ways to be reused and resold. Digital products are easy to be consumed by others, that is, to be reused, or even to be resold to others in whole. Data sets, to the contrary, can be reused by others in different ways, such as aggregation in different dimensions and analysis for different purposes. Moreover, data can be easily processed and transformed so that they can be resold in a hard-to-detect manner.

The above differences between digital products and data products lead to different considerations in pricing principles and methods, which are discussed later. Before we leave this topic, we want to point out that it is possible that the same information can be regarded as digital products in some situations and as data products in some other situations. For example, social media like tweets and customer reviews can be regarded as digital products when a customer reads them online. At the same time, they can be collected and processed in batch by analytic tools to detect events, discover customer profiles and feed recommender systems. In this situation, a systematic collection of social media can be priced and sold as a data product.

2.3 Summary

Information goods, including digital products and data products, distinguish themselves from the traditional physical products in significant cost reductions, particularly in search costs, production costs, replication costs, transportation costs, and tracking and verification costs. The significant reduction

of costs has profound impact on pricing information goods, which is discussed in the later sections. There are several major differences between digital products and data products, including consumption units, aggregatebility, means of consumption, and reusing and reselling.

3 Fundamental Principles of Data Pricing

In this section, we first review the idea of versioning [19], [47], which is a fundamental framework of designing information goods and pricing them. Then, we review several important properties in cost models of digital and data products.

3.1 Versioning

As the replication costs of information goods are very low, even approaching zero in many cases, the price of an information good tends to be very low in marketplaces, too. The potential of very low prices of information goods, on the one hand, makes information goods economically appealing, and, on the other hand, may also make information goods economically dangerous, as the competitors may easily enter the market [19], [47]. This dilemma keeps many traditional pricing strategies far away from being effective for information goods.

To tackle the dilemma, the core idea is "linking price to value", that is, setting the price reflecting the value that a customer places on the information. Specifically, the *versioning strategy* [47] makes different versions to appeal to different types of customers. For example, for a piece of software, different versions have different subsets of features. Essentially, versioning divides customers into subgroups so that each subgroup may regard some features highly valuable and some other features of little value.

There are many different ways to produce different versions of information goods. For example, as information is often time sensitive, delay is often a good basis. In stock market information services, an expensive version may deliver real time quotes while a basic version delivers the same information 20 minutes later. In addition, versions may be defined by convenience (e.g., data can be accessed only by PDF file or by downloadable spreadsheet), comprehensiveness (e.g., the length of historical data available), manipulation (e.g., whether users can store, duplicate, print the information), community (e.g., availability of posting and reading discussion boards), annoyance (e.g., the option of no advertisements), the means of customer support (e.g., by website only or by talking to experts), and many other factors. Most versions of information goods are created by subtracting value from the most technologically advanced and complete version.

In many situations where customers may not realize the value of an information good unless they try it, even the free versions may be provided. The rationale is that the free versions can provide opportunities to potential customers to test out. The objectives of offering free versions include building awareness, gaining follow-on sales, creating a customer network, attracting attentions, and gaining competitive advantages.

The number of versions of an information good may be decided by two major considerations. First, the characteristics of the information to be sold is important. An information good that can be used in many different ways opens the door to many different versions. The second important factor is the value that different customers may place on it. The larger the variance, the more versions may be needed.

The versioning strategy has been investigated in pricing data products, for example, relational data sets and query results [55], [56]. Relational views provide a natural and flexible technical mean to produce versions of an information source. A series of technical challenges are identified, such as arbitrage in pricing, fine-grained data pricing, pricing updates, integrated data and competing data sources, which are reviewed further in this article.

3.2 Important Desiderata in Data Pricing

There are many different ways to design and implement pricing models for information goods. There are a small number of desiderata pursued by most models. How to implement those desiderata in pricing models is discussed in the later sections.

3.2.1 Truthfulness

To make a market efficient, the market is preferred to be truthful. A market is truthful if every buyer is selfish and only offers the price that maximizes the buyer's true utility value. In other words, in a truthful market, no buyer pays more than sufficient to purchase a product. Here, different buyers may have different utility values on the same product. Truthfulness can facilitate a wide spectrum of pricing mechanisms, such as many kinds of auctions [57].

3.2.2 Revenue Maximization

Pricing models can optimize different objectives, such as lowest cost, highest profit, and largest sales. The objective of maximizing revenue is often of special interest in designing pricing strategies. The rationale is that, for a business to be successful long term, a more immediate and important requirement is to win over as many customers as possible.

For traditional physical products, it is often assumed that the marginal cost goes up after a certain number of units are manufactured, and thus the profit can be maximized if the output level is set so that the marginal revenue is equal to the marginal cost, and the revenue can be maximized if the marginal revenue becomes zero. However, given that the replication costs of information goods are very low, revenue maximization and profit maximization for information products become quite different from those for physical products [57], [58].

3.2.3 Fairness

Essentially, a market is fair if each seller gets the fair share of the revenue in coalition. In his seminal article [59], Shapley lays out the fundamental requirements of fairness in markets. Suppose there are k sellers cooperatively participate in a transaction that leads to a payment v. There are four basic requirements for being fair.

- *Balance*: the sum of the payment to each seller should be equal to v. That is, the payment is fully distributed to all sellers.

- *Symmetry*: for a set of sellers S and two additional sellers s and s' who are not in S, that is, $s, s' \notin S$, if $S \cup \{s\}$ and $S \cup \{s'\}$ produce the same payment, then s and s' should receive the same payment. That is, the same contribution to utility should be paid the same.
- *Zero element*: for a set of sellers S and an additional seller $s \notin S$, if $S \cup \{s\}$ and S produce the same payment, then s should receive a payment of 0. That is, no contribution, no payment.
- *Additivity*: If the goods can be used for two tasks T_1 and T_2 with payment v_1 and v_2, respectively, then the payment to complete both tasks $T_1 + T_2$ is $v_1 + v_2$.

In the above well celebrated Shapley fairness, the *Shapley value* is the unique allocation of payment that satisfies all the requirements.

$$\psi(s) = \frac{1}{n} \sum_{S \subseteq D \setminus \{s\}} \frac{\mathcal{U}(S \cup (s)) - \mathcal{U}(S)}{\binom{n-1}{|S|}}, \tag{1}$$

where $\mathcal{U}()$ is the utility function, D is the complete set of sellers, $S \subseteq D$ is a set of sellers, and s is a seller.

Equivalently, Equation (1) can also be written as

$$\psi(s) = \frac{1}{N!} \sum_{\pi \in \Pi(D)} (\mathcal{U}(P_s^\pi \cup \{s\}) - \mathcal{U}(P_s^\pi)), \tag{2}$$

where $\pi \in \Pi(D)$ is a permutation of all sellers, and P_i^π is the set of sellers preceding s in π.

Agarwal *et al.* [57] observe that, as the replication costs of information goods are very low, the marginal costs of production are close to zero, a seller can produce more units of the same information good to obtain a larger Shapley value and thus a larger portion of the payment unjustified in business. This is a challenge in designing fair marketplace for information goods.

3.2.4 Arbitrage-Free Pricing

Arbitrage is the activities that take advantage of price differences between two or more markets or channels. For example, consider a scenario where a user wants to purchase the access to an article, whose listed price is \$35. Suppose that the journal publishing the article has a monthly subscription rate of \$25. Then, the user can conduct arbitrage to subscribe to the journal for only one month and obtain the article at a price cheaper than the listed price.

Arbitrage is often undesirable in pricing models. At least it should be able to check whether a pricing model is arbitrage-free. However, arbitrage can sneak in pricing models that are not thoroughly designed. For example, suppose a data service provider sells query results with prices based on variance [60], a variance of 10 for \$5 each query result and a variance of 1 for \$100 each query result. Each answer is perturbed independently. A customer who wants to obtain an answer of variance of 1 can purchase the query 10 times and compute their average. Due to the independent noise in perturbation, the aggregated average has variance 1, and thus the customer saves \$50 by arbitrage.

3.2.5 Privacy-Preservation

Privacy is becoming a more and more serious concern about information goods. In general, privacy is the ability of an individual or a group to keep themselves or the information about themselves hidden from being identified or approached by other people. Privacy is highly related to information and information exchange, which are what information goods about.

As explained in Section 2.1.5, due to the low tracking costs of information goods, it is easier to collect data about user privacy [52]. Whether privacy should be treated as goods and how privacy is priced are investigated [53], [54].

It is highly desirable to preserve privacy in marketplaces of information goods. In general, transactions in a marketplace may disclose privacy of various parties in many different ways. First, privacy of buyers is highly vulnerable. Their identities, the location and time of purchases, specific products purchased, the purchase prices and total amount may reflect their privacy. It has been reported from time to time that e-commerce providers leak customer information by mistakes, such as an accident reported recently.[3] Second, privacy of information good providers may also be disclosed. For example, medical treatment information in hospitals is highly valuable for many business companies, such as pharmacy and medical equipment companies. Imagine that hospitals can collect and anonymize medical treatment data properly and provide the corresponding data products in marketplaces so that individual patients cannot be re-identified. Buyers, however, may be able to infer from the data the successful rates of a specific treatment in a hospital, which may be regarded as the privacy of the hospital. Last, transactions in marketplaces may also disclose privacy of a third party involved. For example, an AI technology company may provide machine learning model building services to data product buyers. However, machine learning models may be stolen [61], which are regarded privacy of the AI technology company.

To protect privacy in marketplaces of information goods, various directions are being explored, such as hiding the information about what, when and how much a buyer purchases [62], building decentralized and trustworthy privacy preservation data marketplace [63], [64], investigating the tradeoff between payments and accuracy when privacy presents [65], and aggregating non-verifiable information from a privacy-sensitive population [66]. There are many studies on preserving privacy in information goods. We refer interested readers to consult the rich body of surveys [67], [68], [69], [70], [71], [72], [73], [74] and others. We do not discuss further details about general privacy preservation techniques in this article, since privacy preservation techniques are far beyond the scope and capacity of this survey.

3.2.6 Computational Efficiency

As many information goods may be sold to a huge number of potential buyers, a pricing model has to match goods/sellers and buyers with an appropriate price. Computing prices efficiently with respect to a large number of goods and buyers presents technical challenges [55].

3. https://www.telegraph.co.uk/technology/2020/03/10/leak-millions-amazon-ebay-transactions-exposes-customer-addresses/

For example, one reasonable expectation is that a marketplace is polynomial, that is, the complexity of computing prices has to be polynomial with respect to the number of sellers, and cannot grow with respect to the number of goods/buyers when prices are updated [57]. When auctions are used in determining prices, auction efficiency [75] is required to be fast, which is the time needed to process bids.

3.3 Summary

Versioning is a common mechanism in designing and pricing information goods, so that prices of different versions can be linked to values placed by various customer groups. There are a series of important requirements on pricing information goods, including truthfulness, revenue maximization, fairness, arbitrage-free pricing, privacy preservation, and computational efficiency. Those requirements post technical challenges to pricing models.

4 Pricing Digital Products

Although the focus of this article is about pricing data products, we provide a brief review on pricing digital products here, since some general ideas in pricing digital products can be borrowed and extended to data products. In some cases, the boundary between digital products and data products is even blurry.

We first discuss the three major streams of revenues for digital products. Then, we look at two major types of pricing models. The first is bundling and subscription, and the second is auctions. These pricing models are popularly adopted by digital product marketplaces.

4.1 Streams of Revenues

As discussed in Section 3.2.2, revenue maximization often serves as the basic objective in pricing mechanisms, including pricing digital products. Therefore, the understanding of pricing digital products can naturally start with an analysis of possible ways where revenues of digital products may come from. Lambrecht *et al.* [76] summarize that there are three streams of revenues for digital products that are delivered online.

- *Money*. A provider can sell to customers content or, more broadly, services, such as movies and e-books.
- *Information/privacy*. Instead of charging customers directly, a provider can collect customer information by tracking (e.g., using cookies) and sell the information about customers to generate revenues.
- *Time/attention*. A provider can sell space in their digital products to advertisers to produce revenue.

Often, a firm has to design a revenue model for its digital products that combine more than one revenue stream. The three streams are not independent. Instead, they compete with each other, and thus a good tradeoff has to be settled [77]. On the one hand, in some situations, revenues from money stream may be increased at the cost of those from time/attention stream. For example, customers may pay for the content and avoid ads [78], [79], or convert from free versions to premium versions with fitting functions [80]. On the other hand, customers may be highly price sensitive in some digital products, and thus growth in time/attention stream may be easier. For example, an online news site experiences a dramatic loss of customer visits after introducing a paywall [81]. Free samples may stimulate long-term sales [82]. A possible tradeoff between money and time/attention has to be carefully designed.

Typical approaches in revenue models of content and services [83] include rigid pricing (e.g., each movie is priced at a fixed price), designing pricing tiers (e.g., basic versus premium versions), setting up duration of subscription plans (e.g., 6 months of promotion period with very low subscription price) and designing freemium models. One important and unique feature in digital product consumption is micropayments, which means a customer can pay a very small amount that is typically impractical in traditional transactions using standard credit cards due to network service fees. Micropayments and subscriptions have different effects on consumer behavior [84].

As a concrete example of revenue models, consider pricing software products [85]. The major parameters of pricing models include formation of price, structure of payment flow, assessment base, price discrimination, price building and dynamic strategies. The formation of price considers price determination, that is, cost-based, value-based or competition-oriented, as well as degree of interaction, unilateral versus interactive. In terms of payment flow, it may be by single payment, recurring payments or combination. The assessment base of pricing may be usage-dependent (e.g., by transaction or time) or usage-independent (e.g., server types and GPU).

As the tracking costs of digital products are low, a firm can collect customer personal data and sell such data for revenue, that is, generating revenues from information/privacy stream. Typically, personal data may include customers' identities, behavior patterns, preferences and needs. There are various ways to sell customer data, which are also discussed in Section 5 when data products and their marketplaces are discussed. For example [86], [87], a website can provide direct marketing companies user activity information. Moreover, websites can also collaborate with data management platforms (DMP, for advertising) [88] and produce revenues by facilitating businesses to identify audience segments. For example, the information about how customers are connected in social networks can be used to design customized discounts in marketing campaigns [89]. Bergemann and Bonatti [90] develop a model of pricing customer-level information such that the data about each customer are sold individually and individual queries to the database are priced linearly. As new technologies of customer tracking become available, more pricing models may emerge.

We want to point out that selling customer data, though serves the purpose of selling digital products, crosses the boundary between selling digital products and data products. We review some studies on setting prices for customer data and privacy information in the next section.

To produce revenues from time/attention stream, many digital product producers and service providers embed advertisements in their products in one way or the other, and obtain remarkable or even dominant advertising income. However, as John Wanamaker (1838-1922) wisely said, "Half the money I spend on advertising is wasted; the trouble is I don't know which half." It is well recognized

that it is hard to accurately measure advertising effects [91], [92]. Advertisers customize ads for online display [93], [94].

One feasible way to improve advertising effectiveness is to combine user information and advertising opportunities. Retargeted advertising [95] is such an approach, which combines customer online and offline behavior data and makes firms focus on customers showing prior interest in the related products. For example, Athey *et al.* [96] consider customers with multiple homes and investigate the advertising strategies and effectiveness.

In summary, digital product and service suppliers produce revenues through three major streams, money, information/privacy and time/attention. Orthogonally, a firm can bundle its digital products and also design subscription plans that provide products and services in a specific period for a price, which is discussed next.

4.2 Bundling and Subscription Planning

Product bundling organizes products or services into bundles, such that a bundle of products or services are for sale as one combined product or service package. Product bundling is a common marketing practice, particularly in the traditional industry like telecommunication services, financial services, healthcare, and consumer electronics.

As discussed in Section 2.1.3, the low replication costs of information goods allow prevalent adoption of bundling in pricing digital products [19]. Designing product bundles essentially is a combinatorial optimization problem. The basic and static setting is that a customer wants to buy either one or multiple products at a time, which is investigated well before digital products are available [97]. A series of studies [98], [99], [100] develop pricing strategies with two products under different types of bundling. They share the basic assumption that demand for a bundle is elastic comparing to demand for individual products. For example, Armstrong [100] studies the scenarios where products may be substituted or provided by separate sellers.

Bundling multiple products is analyzed, often under the independent value distribution framework [101]. Consider the situation where there are n heterogeneous products for one buyer, and the objective is to maximize expected revenue. Assume that the value distributions on products are independent. That is, for each product x_i, the price that a buyer would like to pay for is an arbitrary distribution D_i in range $[a_i, b_i]$, where $0 \leq a_i \leq b_i < \infty$, and those distributions $D_1, \ldots, D_i$ are independent from each other. Further assume that the buyer is additive, that is, the buyer's value for a set of products is the sum of the buyer's values of those individual products in the set. Babaioff *et al.* [102] show that either selling each item separately or selling all items together as a grand bundle produces at least a constant fraction of the optimal revenue. This interesting and important result allows a simple yet effective bundling strategy: either pricing each product individually or pricing the grand bundle in the expected price. In practice, many platforms, such as Hulu and Amazon Prime Video, offer grand bundle subscription for their products.

More recently, Haghpanah and Hartline [103], [104] show that grand bundle is optimal if more price-sensitive buyers consider the products more complementary. When multiple buyers are considered, whose preferences are unknown, Balcan *et al.* [105] give a simple pricing model that achieves a surprisingly strong guarantee: in the case of unlimited supplies, a random single price achieves expected revenue within a logarithmic factor for customers with general valuation functions. This result allows great convenience in practice, that is, setting a uniform price for all products. It is easier to price a bundle of a larger number of products, since the law of large numbers allows to predict customers' valuations more accurately for a larger bundle of products [106].

Orthogonal to bundling, subscription is to price the interactions between customers and a platform over a period of time. Subscribing customers are in general heterogeneous in both usage rate and value of products. On the one hand, customers with higher usage rates may prefer subscribing to larger subscription sets. On the other hand, in order to maximize revenue, the platform wants customers with lower usage rates to subscribe, and customers with higher usage rates to rent. Moreover, different users may have different values for a product. Many platforms offer subscription and renting at the same time. For a platform, the *subscription model* is to select a subscription fee and the period for each set of products and also set the rental price for each product [107].

Alaei *et al.* [107] follow the model of grand bundle and consider grand subscription, a single rental price for the set that includes all products. They establish the sufficient and necessary condition for the optimality of grand subscription. They also show that subscription fees can be set proportional to the cardinality of a set of products and can achieve $\frac{1}{4\log 2m + \log n}$ of the optimal revenue for n types of customers and m types of products. This approximation is tight in the sense that it cannot be improved more than $\Omega(\frac{1}{\log n})$ in polynomial time.

After all, modeling bundling and subscriptions is computationally challenging due to the combinatorial nature. Dynamic pricing bundles and subscriptions, such as promotions and coupons, have rarely been touched yet.

4.3 Auctions

Auctions have a long history back to the Babylonian and Roman empires [108]. There are many excellent surveys on auctions (e.g., [109], [110], [111], [112]). A comprehensive review on auctions is far beyond the scope and capacity of this article. Here we instead only focus on the important role of auctions as a pricing mechanism for digital products.

4.3.1 Basics About Auctions

There are four basic types of auctions widely used.

- In the *ascending-bid auction* (also known as English auction), the price is raised successively until only one bidder remains, who wins the object at the final price.
- The *descending auction* (also known as the Dutch auction) works the other way by starting at a very high price and lowering the price continuously, until the first bidder calls out and accepts the current price.
- In the *first-price sealed-bid auction*, every bidder submits a bid without knowing the others' bids. The one making the highest bid wins and pays at the named price.

- The *second-price sealed-bid auction* (also known as the Vickrey auction [113]) works in the same way as the first-price sealed-bid auction does, except that the winner pays only the second highest bid.

There are two basic models of the value information in auctions. The *private-value model* assumes that every bidder has an independent value on the object for sale. The value is also private to the bidder only. The *pure common-value model* assumes that the actual value of the object for sale is the same for all bidders, but bidders have different private information about that actual value. Every bidder adjusts her/his estimate of the actual value by learning other bidders' signals. There are also models considering both values private to individual bidders and common to all bidders.

One fundamental principle in auction theory is the *revenue equivalence theorem* [101], [113], [114], [115], which essentially states that, for a set of risk-neutral bidders with independent private valuation of an object drawn from a common cumulative distribution that is strictly increasing and atomless on $[v_{min}, v_{max}]$, any auction mechanism yields the same expected revenue and thus any bidder with valuation v makes the same expected payment if (1) the object is allocated to the bidder with the highest valuation; and (2) any bidder with valuation v_{min} has an expected utility of 0. Based on the revenue equivalence theorem, the four basic types of auctions lead to the same payment by the winner and the same revenue.

While most studies in auction theory make some simple assumptions about independence of customer valuations, empirical studies [116] demonstrate that, in practice, the wrong assumption of valuation independence causes inefficient auctions in e-commerce.

4.3.2 Sponsored Search Auctions

Online ad and sponsored search auctions [117], [118], [119] are one important application of auctions in pricing digital products. Sponsored search [120] is the business model where content providers pay search engines for traffic to their websites. In sponsored search, advertisers and, more generally, content providers bid for keywords in search engines, and search engines decide which ad to display in which position to answer a query from a user. GoTo.com created the first sponsored search auction [120].

Different pricing models can be used in sponsored search auctions, such as pay-per mille[4]/pay-per impression (PPM), pay-per-click (PPC), and pay-per-action (PPA). In the early days of sponsored search, a generalized first price auction is used. Each advertiser bids on multiple keywords, and can set a bidding price for each keyword. When a user query is answered, which is a keyword, the top k bids on the keyword in price are displayed. If an ad is clicked by the user, the corresponding advertiser pays the bidding price. The first price auction mechanism is unstable, costs advertisers time and reduces search engine profits [121]. Later, Google generalizes the second price auction mechanism [122], and enhances the ranking of bids by additional information, such as the ad's click-through-rate (CTR), keyword relevance, and ad's landing-page/site quality.

There are many in depth analyses about sponsored search auction mechanisms (e.g., [119]). For example, some studies analyze auction mechanisms based on assumptions about rationality, budget constraints and CTR distributions. Some other studies look at practical sponsored search systems and discuss auction mechanisms when the standard assumptions do not hold. Another group of studies, such as [123], [124], [125], [126], conduct empirical studies to understand bidding behavior and statics. Last and latest, deep learning approaches are used to develop auction strategies in sponsored search [127], [128].

4.3.3 Auctions on Digital Products With Unlimited Supplies

One unique feature of digital products is that the replication costs are very low and thus may have almost unlimited supply. Products of unlimited supplies lead to new challenges and opportunities to auction mechanism design. For example, the second price auction can be straightforwardly generalized for k identical products – the top k highest bidders win and each pays the $(k+1)$-th bidding price. However, when there are unlimited identical products, the $(k+1)$-th bidding price approaches 0. The lack of competition due to obsessive supplies prevents bidders from offering any high prices. In other words, the challenge is how to ensure the bids are truthful, that is, reflecting the bidders' true valuation of the digital products.

Denote by B the set of bidders, and by $b_1, b_2, \ldots$ the bidding prices in descending order, that is, $b_i \geq b_{i+1} \geq 0$ for any $i > 0$. Suppose the generalized second price auction mechanism is used. That is, if k bids are taken, those winning bidders each pays the cost b_{k+1}. The auction objective is to maximize $k \cdot b_{k+1}$. An auction is *competitive* if it yields revenue within a constant factor of the optimal fixed pricing. It is tricky that, when there is unlimited supply, the Vickrey auction is not competitive if the seller chooses the number of products to sell before knowing the bids, and is not truthful if the seller chooses after knowing the bids [75].

Goldberg *et al.* [75] propose the first competitive auction for digital goods with unlimited supplies. The major idea is the smart framework of *random sampling auction*. An auction is *bid-independent* if bidder i's bid value should only determine whether the bidder wins the auction, but not the price. We select a sample B' of B at random, independent from the bid values. We use the bids in B' to compute the optimal bid threshold $f_{B'}$ that maximizes the revenue in B', and every bidder in $B - B'$ whose bid value is over $f_{B'}$ wins. Symmetrically, we use the bids in $B - B'$ to compute the optimal bid threshold $f_{B-B'}$ that maximizes the revenue in $B - B'$, and every bidder in B' whose bid value is higher than $f_{B-B'}$ wins. In general, $f_{B'} = f_{B-B'}$ does not necessarily hold. Random sampling auctions are competitive, no matter the single-price version or the multi-price version. Indeed, random sampling auctions are 15-competitive in the worst case [129] and 4-competitive for a large class of instances where there are at least 6 bids that are as good as the optimal sale price [130]. There are a series of improvements on random sampling auctions. For example, Hartline and McGrew [131] further improve the competitiveness.

4. That is, the cost of 1,000 advertisement impressions.

Goldberg and Hartline [132] extend the scope from single digital product with unlimited supply to multiple products with unlimited supplies. Given a set of bids, they show that the bidder-optimal product assignment given the bids and the optimal sale prices can be determined by solving an integer programming problem.

Then, we can solve the optimal pricing problem in the following random sampling auction. Let B be the set of bidders. First, we obtain a sample B' of bidders. Second, we compute the optimal sale prices for B'. Last, we run the fixed-price auction on $B - B'$ using the sale prices computed in the integer programming problem. All bidders in B' lose the auction. The random sampling auction is shown truthful and competitive [132].

Most of the proposed auctions for digital goods with unlimited supply are randomized auctions. Goldberg *et al.* [75] show that no deterministic auction can be competitive. Aggarwal *et al.* [133] later point out that the result does not hold for asymmetric auctions [134]. In a symmetric *ex ante* auction, buyers' preference parameters are drawn from a symmetric probability distribution, and thus there exists a symmetric equilibrium if an equilibrium exists at all. In an asymmetric auction, each buyer has the same information about the product but a different opportunity cost of obtaining the product, that is, bidders' valuations are drawn from different distributions. Aggarwal *et al.* [133] give an asymmetric deterministic auction that can approximate the revenue of any optimal single-price sale in the worst case. Indeed, they develop a general derandomization technique to transform any randomized auction into an asymmetric deterministic auction with approximately the same revenue. The general idea follows the deterministic maximum flow solution to the well-known hat problem [135].

4.3.4 Envy-Free Auctions

One drawback in random sampling auctions is that some bidders may lose even they make bids higher than some winning bidders do, since the bidders in B' and $B - B'$ use different thresholds (i.e., $f_{B-B'}$ and $f_{B'}$, respectively) in the one product version and all bidders in B' lose in the multiple product version.

Goldberg and Hartline [136] establish a fundamental result: an auction cannot be truthful, competitive and envy-free at the same time. They also explore possible tradeoffs between truthfulness and envy-freeness based on the consensus revenue estimate (CORE) technique [137]. Specifically, using a similar idea in combinatorial auctions with single parameter agents [138], we can relax the truthfulness requirement by requiring being truthful with probability $(1 - \epsilon)$, and always guarantee envy-free. The auction is highly truthful when ϵ approaches 0 and the number of winners in the auction approaches infinity. The other type of auctions relaxes the envy-free requirement to being envy-free with probability $(1 - \epsilon)$, and guarantees truthfulness. Both auctions are competitive and the probability is over random coin tosses made by the randomized auction mechanism and not the input.

4.3.5 Online Auctions

In addition to potentially unlimited supply, another important feature of digital goods is that a digital good may be sold repetitively, such as a movie and a song. Therefore, auctions on digital goods may run continuously instead of only one round. Moreover, customers may want to have prompt answers to their bids.

Online auctions [139] are designed to address the setting where different customers bid at different times. The auction mechanism has to make decision about each bid as it arrives. An (online) auction is *incentive compatible* if the bidders are rationally motivated to reveal their true valuations of the object. Lavi and Nisan [139] show that an online auction is incentive compatible if and only if it is based on supply curves under the assumption of limited supply, that is, before it receives the ith bid $b_i(q)$, it fixes the supply curve $p_i(q)$ based on the previous bids, and (1) the quantity q_i sold to customer i is the quantity q that maximizes the sum $\sum_{j=1}^{q}(b_i(j) - p_i(j))$; and (2) the price paid by i is $\sum_{j=1}^{q} p_i(j)$.

To tackle the challenges when there is unlimited supply, Bar-Yossef *et al.* [140] point out that supply curves are not available anymore. Instead, they propose an extremely simple incentive-compatible randomized online auction. Each bidder i picks a random number $t \in \{0, \ldots, \lfloor \log h \rfloor\}$ and sets the price threshold to $s_i = 2^t$, where h is the ratio of the highest valuation against the lowest valuation among all bidders. This auction is $O(\log h)$-competitive.

The auction mechanism can be further improved to achieve even better incentive-compatibility. Specifically, we can divide a sequence of bids $b_1, b_2, \ldots$ into $l = (\lfloor \log h \rfloor + 1)$ buckets, such that bucket B_j contains the bids with indexes in range $[2^j, 2^{j+1})$. The weight of bucket B_j is the sum of bids within B_j, that is, $w_j = \sum_{i \in B_j} i$. A new bidder can choose one of the buckets at random with the probability proportional to the bucket weight, and pays the price of the lowest bid of the bucket. The price s_i that bidder i pays follows the probability distribution $Pr[s_i = 2^j] = \left(\frac{w_j}{\sum_{r=0}^{l-1} w_r}\right)^d$, where d is a parameter. The auction is shown $O(3^d (\log h)^{\frac{1}{d+1}})$-competitive. By setting $d = \sqrt{\log \log h}$, the auction is $O(\exp(\sqrt{\log \log h}))$-competitive.

4.4 Summary

As revenue maximization plays a fundamental role in pricing digital products, we review the three major streams of revenues for digital products, namely money, information/privacy, and time/attention. Then, we revisit bundling and subscription planning for digital products, which echoes the opportunities and challenges due to low replication costs of information goods. Auctions are widely used in pricing digital products. We review some basic types of auctions and their applications in digital products, including sponsored search auctions, auctions with unlimited supplies, envy-free auctions and online auctions. Some ideas employed by pricing digital products are also used in pricing data products, as to be discussed in the next section.

5 Pricing Data Products

In this section, we discuss pricing in marketplaces of data. We first obtain an overall understanding about data markets and the major players in such markets. Then, we look into several most studied technical problems in data product pricing, including arbitrage-free pricing, revenue maximization pricing, fair and truthful pricing and privacy

preservation in data marketplaces. Last, we discuss pricing in novel application scenarios, including dynamic data pricing, online pricing and federated learning pricing.

5.1 Data Markets and Pricing, What Are They?

Marketplaces for data have been actively developed for over a decade. An early survey [141] identifies different categories and dimensions of data marketplaces and data vendors in 2012. There are many studies on various issues about data markets and pricing strategies. Before we discuss any specifics in detail, it is important to obtain an overall understanding about data markets, such as what are sold and for what purposes, who are the sellers, who are the buyers, and what are the basic pricing models.

Pantelis and Aija [142] present a brief economic analysis of data taxonomy as a market mechanism. Data and databases are legally protected by either copyright or database right. Copyright protects expression and significant creative effort that creates and organizes data. Database right protects a whole database. One challenge is that both copyright and database right are hard to enforce due to the non-rivalrous nature of data.

In general, data may be owned by governments, private parties or individuals. Consequently, data can be categorized into three types: open, public, and private data [142]. Open data are common pool resources [143], such as the data made available by the open data initiatives. Public data, such as the data collected by the government in the United States, are valuable resources subject to the "tragedy of the commons" [144]. Public data are often produced by individuals or organizations for research and used by governments and local authorities, but may also be employed by commercial parties to enhance their proprietary resources or services. Private data are generated by private applications or services.

To understand what are sold in data markets and for what purposes, Muschalle *et al.* [145] consider the common queries and demands on data markets, as well as the pricing strategies. They observe two major types of queries. The first type is to estimate the value of a "thing" or compare the values of "things", where examples of the "things" are like webpages for advertisements, starlets, politicians and products. The second type is to show all about a "thing". Those queries are raised by seven categories of beneficiaries, namely analysts, application vendors, data processing algorithm developers, data providers, consultants, licensing and certification entities, and data market owners. The authors also identify three types of market structures. First, in a monopoly, a supplier is powerful enough to set prices to maximize profits. Second, an oligopoly is dominated by a small number of strong competitors. Last, in strong competition markets, prices may align with marginal costs.

A series of pricing strategies and models may be considered in data markets [145]. First, free data may be obtained from public authorities, may help to attract customers and suppliers of commercial data, and may be integrated into private and not-free data products. Second, prices can be based on usages, such as charging customers per hour of data usage. Third, package pricing allows a customer to obtain a certain amount of data or API calls for a fixed fee. A few studies [12], [146] try to optimize package pricing models. Fourth, in the flat fee tariff model, a data product or service is offered at a flat rate, regardless of usage. It is simple, easy to use. The drawback is the lack of flexibility, particularly for buyers. Fifth, combining package pricing and flat fee tariff results in two-part tariff, that is, a fixed basic fee plus additional fee per unit consumed. This model is popular in data services. Specifically, Wu and Banker [147] show that, under zero marginal costs and monitoring costs, flat fee and two-part tariff pricing are on par, and two-part tariff is the most profitable strategy. Last, in the freemium model, users can use basic products or services for free and pay for premium functions or services.

Recently, machine learning, particularly deep learning [148], becomes disruptive in many applications, such as computer vision [149], [150] and natural language processing [151]. In most situations, powerful deep models heavily rely on large amounts of training data [152]. Monetization of data and machine learning models built on data through markets gains stronger and stronger interests from industry. Specific to data as an economic good and data pricing as a monetization mechanism in this context, a series of studies focus on data utility for model building and the associated pricing, particularly considering privacy.

Some data owners may have detailed knowledge of specific machine learning tasks and thus dedicate corresponding effort to collect high quality data for building better models. Babaioff *et al.* [153] study the design of optimal mechanisms for a monopoly data provider to sell her/his data. Specifically, they show that it is feasible to achieve optimal revenue by a simple one-round protocol, that is, a protocol where a buyer and a seller each sends a single message, and there is a single money transfer. The optimal mechanism can be computed in polynomial time. For a buyer who may abort the interaction with a seller prematurely, multiple rounds of partial information disclosure interleaved by payments may be needed to ensure optimal revenue. Cummings *et al.* [154] study the optimal design for data buyers to purchase data estimators with different variances and combine the estimators to meet a required quality guarantee on variance with the lowest total cost.

The role of privacy in data collection and machine learning model building is investigated. For example, Ghosh and Roth [155] develop auctions that are truthful and approximately optimal for data buyers to obtain accurate estimates on data from owners who are compensated for privacy loss. They show that the classic Vickrey auction [113] can minimize the buyer's total payment and meet the accuracy requirement. They also develop a mechanism that can maximize the accuracy given a budget.

In general, modeling data owners' costs of privacy loss is very difficult, since the costs may be correlated with private data arbitrarily. It is impossible to design a direct revelation mechanism that can provide a non-trivial guarantee on accuracy and, at the same time, is rational for individual data owners. To tackle the issue, Ligett and Roth [156] design a take-it-or-leave-it mechanism, which randomly approaches individuals from a population and makes offers. This mechanism can be used for some data collection scenarios, such as surveys.

Versioning is an important strategy in data pricing. A data seller can customize data into different versions

according to buyers' needs. Bergemann *et al.* [157] develop the optimal menu of information products that a monopoly data supplier can offer to a data buyer, so that one product can fit the buyer's willingness to buy the information at the offered price, and the revenue is maximized. One important finding is that information products indeed allow larger scopes of price discrimination. There are at least two dimensions that sellers can explore to derive various subsets of a data set, namely data quality and data position.

When data are used to build machine learning models, it is important to assess the value of each data record within a data set. There exist various methods for assessment, such as leave-one-out [158], leverage or influence score [159].

Ghorbani and Zou [160] propose to apply the Shapley fairness on the data used to train a machine learning model, and thus define the data Shapley value for a record i in a training data set D as

$$\psi_i = C \sum_{S \subseteq D-\{i\}} \frac{\mathcal{U}(S \cup \{i\}) - \mathcal{U}(S)}{\binom{n-1}{|S|}},$$

where C is an arbitrary (positive) constant, and $\mathcal{U}(S)$ is the performance score of the model trained on data $S \subseteq D$. One challenge is that computing the exact data Shapley values on large data sets for sophisticated models, such as deep neural networks, is computational prohibitive. Ghorbani and Zou [160] also develop Monte Carlo and gradient-based methods for estimation.

If a data point p appears in two samples D_1 and D_2 from the same data distribution, intuitively the Shapley value of p in D_1 and D_2 should be similar. Mathematically, the intrinsic Shapley value of p in a distribution should be the expectation of the Shapley value of p in the distribution. Based on this intuition, Ghorbani *et al.* [161] propose the notion of distributional Shapley. Let $\mathcal{Z}$ be a universe in question. Let $\mathcal{D}$ be a data distribution in $\mathcal{Z}$. Assuming a potential function or a performance metric $U : \mathcal{Z}^* \rightarrow [0, 1]$ and a sample size $m > 0$, the distributional Shapley value of a point $z \in \mathcal{Z}$ is the expected Shapley value over data sets of size m containing z, that is, $\nu(z; U, \mathcal{D}, m) = \mathbb{E}_{S \sim \mathcal{D}^{m-1}}[\psi(z; U, S \cup \{z\}]$, where $S \sim \mathcal{D}^{m-1}$ is a set of m points sampled i.i.d. from $\mathcal{D}$. They show that distribution Shapley values are stable. Kwon *et al.* [162] further derive the computationally tractable expressions for distributional Shapley for a series of models, including linear regression, binary classification and non-parametric density estimation.

In machine learning, influence functions [163], [164] approximate leave-one-out to assess the value of a data item. Cai *et al.* [165] propose strategy-proof mechanisms for data elicitation and trade off between model accuracy and reward. Richardson *et al.* [166] focus on the case of linear regression. Recently, Yoon *et al.* [167] propose data valuation using reinforcement learning. They use a data value estimator to learn how much a data item as an element in the training data contributes to improving model performance. One distinct advantage is that the model being trained and the data value estimator can improve each other's performance.

Data quality is an important issue [168]. There are many studies on assessment of data quality [168], [169], [170]. Some studies specifically focus on pricing based on data quality and the impact on data markets. Heckman *et al.* [170] propose a simple linear model, $Value\ of\ data = fixed\ cost + \sum_i w_i \cdot factor_i$, where the factors include but are not limited to age of data, periodicity of data, volume of data, and accuracy of data, and w_i is the associated weight. One practical difficulty in using the model is that the parameters in the model are hard to estimate. Another difficulty is that many data sets do not have public prices associated. Yu and Zhang [171] consider pricing multiple versions formed by multiple factors of data quality and build a two-level model. The first level is the data platform where a single owner is assumed, who designs the number of versions. The second level is the customers who want to maximize the data utility. Each level is modeled as a maximization problem and thus the whole model is a bi-level programming problem, which is NP-hard.

Another way to form multiple versions of data products is to charge by queries [172], [173], [174], [175]. Intuitively, a data seller may treat a view of a data set as a version. However, setting the price for every possible view is not only tedious but also tricky. If prices on views are not set properly, arbitrages or less than highest prices may happen. Koutris *et al.* [173], [174] propose a framework of query and view based data pricing. The major idea is that a seller only needs to specify the prices on a few views, and then the prices of other views can be decided algorithmically. Thy advocate two desiderata, arbitrage-freeness and discount-freeness. Theoretically, they show the existence and uniqueness of pricing functions satisfying the requirements. They also show the complexity of computing the pricing functions. Unfortunately, only selection views and conjunctive queries without self-joins are tractable. They present polynomial time algorithms for chain queries and cyclic queries.

Technically, the core idea in the view and query based pricing framework is query determinacy [176], [177], [178]. A query Q is said to be determined by a set of views V if the answer to Q can be completely derived from the views. Query determinacy enables the feasibility of arbitrage detection. If V determines Q, then arbitrage happens if and only if the price of V is cheaper than that of Q.

Koutris *et al.* [172], [175] further explore the technical challenges in practical implementation of view and query based data pricing. Specifically, they develop an integer linear programming formulation for the pricing problem with a large number of queries. Considering the scenario where a user may purchase multiple queries over time or the database is updated, such that information in multiple queries and updates may have overlaps, they also leverage query history to avoid double charging. To handle the situation where there are multiple sellers, they define the share of a seller as the maximum revenue that the seller can get among all minimum-cost solutions, and accordingly define a fair revenue distribution policy.

Tang *et al.* [179] follow the view and query based pricing framework and consider each tuple as the the minimum granularity of data. Their model assigns to each tuple a price and prices queries based on minimal provenances. Tang *et al.* [180] extend view and query based pricing to XML documents and prices of samples of query results.

5.2 Arbitrage-Free Pricing

As introduced in Section 3.2.4, arbitrage is undesirable in many pricing models. Unfortunately, arbitrage may sneak in

pricing models without rigorous design. For example, Balazinska *et al.* [55] analyze that subscription based pricing possibly with a query limit allows arbitrage. Muschalle *et al.* [145] point out that a pricing model charging users a certain amount of API calls for a fixed rate may potentially allow arbitrage, depending on the package size.

Arbitrage-freeness is one of the fundamental properties of pricing models in query and view based pricing [172], [173], [174], [175]. Li and Miklau [181] and Li *et al.* [60] develop frameworks of pricing linear aggregate queries. Specifically, Li *et al.* [60] consider linear queries. Given a data set of n tuples $x_1, \ldots, x_n$, a linear query $\mathbf{q} = (q_1, \ldots, q_N)$ is a real-valued vector, and the answer $\mathbf{q}(\mathbf{x}) = \sum_{i=1}^{n} q_i x_i$. For a multiset of queries $\mathbf{S} = \{\mathbf{Q_1}, \ldots, \mathbf{Q_k}\}$ and query $\mathbf{Q}$, if the answer to $\mathbf{Q}$ can be linearly derived from the answers to the queries in $\mathbf{S}$, then $\mathbf{Q}$ is said to be determined by $\mathbf{S}$, denoted by $\mathbf{S} \rightarrow \mathbf{Q}$. A pricing function $\pi(\mathbf{Q})$ is arbitrage-free if for any multiset $\mathbf{S}$ and query $\mathbf{Q}$ such that $\mathbf{S} \rightarrow \mathbf{Q}$, $\pi(\mathbf{Q}) \leq \sum_{i=1}^{k} \pi(\mathbf{Q}_i)$.

Under the general intuition of arbitrage-freeness, Li *et al.* [60] consider a specific form of queries, linear queries with variance $(\mathbf{q}, v)$, that is, the estimation of the answer to query $\mathbf{q}$ should have a variance no larger than v. Using different values of v, different versions are formed. A pricing model not carefully designed may allow arbitrage. Li *et al.* [60] first observe $\pi(\mathbf{q}, v) = \Omega(\frac{1}{v})$. Then, they synthesize pricing function $\pi(\mathbf{q}, v) = \frac{f^2(\mathbf{q})}{v}$, which is arbitrage-free if f is positive and semi-norm.[5] For any arbitrage-free pricing functions $\pi_1, \ldots, \pi_k$, $f(\pi_1(\mathbf{q}), \ldots, \pi_k(\mathbf{q}))$ is also arbitrage-free if f is subadditive[6] and nondecreasing.

As Roth [182] summarizes, the framework by Li *et al.* [60] still faces three important challenges. First, arbitrage is still possible to derive answers to a bundle of queries from another bundle of queries and their answers. Second, arbitrage is still possible on biased estimators for statistical queries. Last, it is unclear whether we can obtain arbitrage-free pricing maximizing profit given the distribution of buyer demands. Later, Deep and Koutris [183] provide some interesting insights into arbitrage-free pricing for bundles.

Lin and Kifer [184] investigate arbitrage-free pricing for general data queries. They consider three types of pricing models for query bundles, where a query bundle is a set of queries posted simultaneously as a batch. First, an instance-independent pricing function depends on the query bundle but not the database instance. Second, an up-front dependent pricing function depends on both the query bundle and the database instance. A customer knows an un-front dependent pricing function, and decides whether to purchase or not the query answers. Last, a delayed pricing function depends on both the query bundle and the answers computed by the query bundle on the current database instance. The customer knows the pricing function, but do not know the exact price. Once agreeing, the customer is charged when the answers are computed. They also summarize five different types of arbitrage situations. First, if prices are quoted by queries, in order to avoid price-based arbitrage, answers to queries should not be deduced from prices along. Second, a buyer may use multiple accounts to derive answers to a query bundle. To avoid separate account arbitrage, the price of a query bundle $[q_1, q_2]$ should be at most the sum of the prices of q_1 and q_2. Third, if the answers to a query bundle q' can always be deduced from answers to another query bundle q, to prevent post-processing arbitrage from happening, the price of q should be no cheaper than that of q'. Fourth, although the answers to a query bundle q may not be always derivable from the answers to another query bundle q' on all database instances, still for a specific database instance $\mathcal{I}$, the answers to q may be derived from the answers to q'. If so, a serendipitous arbitrage happens. Last, if two queries behave almost identical but their prices are dramatically different, almost-certain arbitrage happens. Based on the above categorization, they discuss conditions that can prevent various types of arbitrage situations.

Pricing many queries in real time with formal guarantees on arbitrage freeness is challenging. Many theoretical methods are not scalable in practice. Qirana [185], [186] is a system for query-based pricing, which allows data sellers to choose from a set of pricing functions that are information arbitrage-free, which covers both post-processing arbitrage-freeness and serendipitous arbitrage-freeness in Lin and Kifer's taxonomy [184]. Qirana also supports history-aware pricing. Qirana has been shown highly efficient and scalable on TPC-H[7] and SSB[8] benchmark datasets as demonstration. The key idea is that it regards a query as an uncertainty reduction mechanism. Initially, a buyer faces a set of possible databases $\mathcal{I}$ defined by a database schema, primary keys and predefined constraints. Once a buyer obtains the answer E to a query Q, all possible databases D such that $E \neq Q(D)$ are eliminated. The price assigned to Q should be a function of how much the set of possible databases shrinks. Let $\mathcal{S}$ be the set of possible databases before the query Q is answered. $\mathcal{S}$ is called the support set. Then, a weighted coverage function assigns a weight w_i to every $D_i \in \mathcal{S}$, and computes the price to a query by $p^{wc}(Q, D) = \sum_{Q(D_i) \neq Q(D)} w_i$. Alternatively, consider the equivalence relation in $\mathcal{S}$: $D_i \sim D_j$ if and only if $Q(D_i) = Q(D_j)$. Assign to each possible database $D_i \in \mathcal{S}$ a weight w_i such that $\sum_{D_i \in \mathcal{S}} w_i = 1$. Let $\mathcal{P}_Q$ be the set of equivalence classes. For each class $B \in \mathcal{P}_Q$, denote by $w_B = \sum_{D_i \in B} w_i$. The Shannon entropy function is used to compute the price of query Q as the entropy of the query output $P^H(Q, D) = -\sum_{B \in \mathcal{P}_Q} w_B \log w_B$. The q-entropy function (also known as Tsallis entropy) for $q = 2$ is used to assign to Q the price $P^T(Q, D) = \sum_{B \in \mathcal{P}_Q} w_B(1 - w_B)$. Using the complete set of possible databases as the support set leads to a $\#P$-hard problem. To make the price calculation computationally feasible, Qirana uses uniform random sample and random neighboors as the support sets. Deep and Koutris [183] show that the weighted coverage function, the Shannon entropy function and the 2-entropy function are all arbitrage-free.

In targeted advertising markets, user data, such as opt-in email addresses, and user impressions are sold as data products. How to price users[9] properly to avoid arbitrage is

5. A function $f : \mathcal{R}^n \rightarrow \mathcal{R}$ is semi-norm if for any $c \in \mathcal{R}$ and any query $\mathbf{Q} \in \mathcal{R}^n$, $f(c\mathbf{q}) = |c| f(\mathbf{q})$; and for any $\mathbf{q}_1, \mathbf{q}_2 \in \mathcal{R}^n$, $f(\mathbf{q}_1 + \mathbf{q}_2) \leq f(\mathbf{q}_1) + f(\mathbf{q}_2)$.

6. A function f is subadditive if for any $x_1, \ldots, x_k$, $f(\sum_{i=1}^{k} x_i) \leq \sum_{i=1}^{k} f(x_i)$.

7. http://www.tpc.org/tpch.

8. http://www.cs.umb.edu/?poneil/StarSchemaB.PDF

9. Here, "buying a user" is short for purchasing the impression of a user in online advertising and a user email in targeted email advertising, for example.

important. Xia and Muthukrishnan [187] consider the following problem. Denote by q_i a selection query over user attributes, by U_i the set of all users satisfying q_i, and by p_i the price of each user in U_i. If a buyer purchases n users ($1 \leq n \leq |U_i|$) in U_i, she/he has to pay $n \cdot p_i$. If prices of different queries are not well coordinated, version-arbitrage may arise. If two queries q_i and q_j return similar user sets but q_i is dramatically more expensive than q_j, then a user who wants q_i may purchase q_j instead. Xia and Muthukrishnan [187] point out that uniform pricing, that is, every query has the same price, is arbitrage-free, but is a logarithmic approximation to the maximum revenue arbitrage-free pricing solution. Then, they present a greedy non-uniform pricing design. The design starts with the optimal uniform pricing that is arbitrage-free, and then iteratively updates the pricing function. If the price of a query can be updated to increase the revenue, it is increased so that the arbitrage-free property is retained. This greedy algorithm is still a logarithmic approximation to the maximum revenue arbitrage-free pricing solution.

Chen *et al.* [188] develop an arbitrage-free pricing design for multiple versions of a machine learning model. They assume that a broker trains the optimal model on the complete raw data. Then, random Gaussian noises are added to the optimal model to produce different versions for different buyers. The assumption is that the error of a machine learning model instance is monotonic with respect to the variance of the noise injected into the model. In this setting, a pricing function is arbitrage-free if and only if the price of a randomized model instance is monotonically increasing and subadditive with respect to the inverse of the variance.

5.3 Revenue Maximization Pricing

Revenue maximization pricing for data products is a relatively less explored area. A possible reason is that, comparing with pricing digital products, some other factors in pricing data products need more urgent accommodation, such as arbitrage.

Xia and Muthukrishnan [187] also consider the situations where both the maximum number (i.e., maximum demand) and the minimum number (i.e., minimum demand) of users that a buyer purchases are specified, and provide an $O(D)$ approximation algorithm to maximize revenue, where D is the largest minimal demand among all buyers.

Chawla *et al.* [58] consider query and view based pricing for arbitrage-free revenue maximization under the assumption that all buyers are single-minded and the supply is unlimited. A buyer is single-minded if the buyer wants to purchase the answer to a single set of queries. They consider three types of pricing functions. Uniform bundle pricing sets the price of every bundle identical. Additive or item pricing prices each item and charges a bundle the sum of prices for the items in the bundle. Fractionally subadditive pricing or XOS sets k weights $w_j^1, \ldots, w_j^k$ for each item j, and for a bundle e, the price is set to $\max_{i=1}^{k} \sum_{j \in e} w_j^i$. Building on the extensive studies on revenue maximization with single-minded buyers and unlimited supply [189], [190], [191], they develop new heuristics.

It is well known that there exists uniform bundle pricing that is $O(\log m)$ approximation of revenue maximization, where m is the number of bundles. Swamy and Cheung [192] show that item pricing can achieve an $O(\log B)$ approximation of maximum revenue, where B is the maximum number of bundles an item can involve. Chawla *et al.* [58] show some new lower bounds, that is, uniform bundle pricing, item pricing and XOS pricing combining a constant number of item pricing functions are still $\Omega(\log m)$ away from maximum revenue. They also present approximation algorithms.

To maximize revenue in machine learning models, Chen *et al.* [188] show that the optimization problem is coNP-hard. Thus, they relax the subadditive constraint $p(x+y) \leq p(x) + p(y)$ by $\frac{q(x)}{x} \geq \frac{q(y)}{y}$ for every $0 < x \leq y$, and turn to finding a pricing function $q()$ such that $\frac{q(x)}{x}$ is decreasing with respect to x. They show that, for every well standing pricing function $p()$, there exists a pricing function $q()$ with the relaxed subadditive constraint such that $\frac{p(x)}{2} \leq q(x) \leq p(x)$, and $q(x)$ can be computed using dynamic programming in $O(n^2)$ time, where n is the number of interpolated price points.

5.4 Fair and Truthful Pricing

Fairness and truthfulness are important for data product markets. Recall that fairness refers to that the revenue generated by a sale transaction in the data market is distributed among sellers in an unprejudiced manner so that they are paid for their marginal contributions. Truthfulness means a market where buyers are well motivated to report their internal valuations of data products unwarily.

Agarwal *et al.* [57] propose a mathematical model of data marketplaces that are fair, truthful, revenue maximizing, and scalable. They assume each seller j supplies a data stream X_j and each buyer n conducts a prediction task Y_n, where $X_j, Y_n \in \mathcal{R}^T$. For example, X_j may be a stream of customers' interest on different products, and Y_n is a task predicting a new customer's interest. Taking a prediction task Y_n and an estimate $\hat{Y_n}$, a prediction gain function $\mathcal{G}_n : \mathcal{R}^{2T} \rightarrow [0,1]$ measures the quality of the prediction. The value that buyer n gets from estimate $\hat{Y_n}$ is $\mu_n \cdot \mathcal{G}(Y_n, \hat{Y_n})$, where μ_n is the price rate that the buyer is willing to pay for a unit increase in $\mathcal{G}$. A machine learning model $\mathcal{M} : \mathcal{R}^{MT} \rightarrow \mathcal{R}^T$ uses data from M sellers to produce an estimate $\mathcal{Y}_n$ for buyer n's prediction task Y_n. Let p_n and b_n be the price and the bid, respectively. Then, allocation function $\mathcal{AF} : (p_n, b_n; X_M) \rightarrow \widetilde{X}_M$ measures the quality at which buyer n obtains that is allocated to the sellers on sale X_M, where $\widetilde{X}_M \in \mathcal{R}^M$. Revenue function $\mathcal{RF} : (p_n, b_n, Y_n; \mathcal{M}, \mathcal{G}, X_M) \rightarrow r_n$ calculates how much revenue $r_n \in \mathcal{R}^+$ to extract from the buyer. The utility that buyer n receives by bidding n_n for Y_n is $\mathcal{U}(b_n, Y_n) = \mu_n \cdot \mathcal{G}(Y_n, \hat{Y_n}) - \mathcal{RF}(p_n, b_n, Y_n)$, where $\hat{Y_n} = \mathcal{M}(Y_n, \widetilde{X}_M)$ and $\widetilde{X}_M = \mathcal{AF}(p_n, b_n; X_M)$. A market is truthful if for all prediction tasks Y_n, $\mu_n = \arg\max_{z \in \mathcal{R}^+} \mathcal{U}(z, Y_n)$. They adopt the notion of fairness following the famous Shapley fairness [59].

Shapley fairness [59] is popularly adopted as the foundation of fairness in data markets. However, computing Shapley value is exponential [193]. Maleki *et al.* [194] present a permutation sampling method that approximates Shapley value for any bounded utility functions. The basic idea is to use Equation 2 and tackle $\psi(s) = E[\mathcal{U}(P_s^\pi \cup \{s\}) - \mathcal{U}(P_i^\pi)]$ by sample mean. Following Hoeffding's inequality [195], to achieve an (ϵ, δ)-approximation, that is, $P(|\hat{s} - s|_p \leq \epsilon) \geq$

$1-\delta$, where $\hat{s}$ is the estimate, we need $\frac{2r^2N}{\epsilon^2}\log\frac{2N}{\delta}$ samples and evaluate the utility function $O(N^2\log N)$ times, where r is the range of the utility function $\mathcal{U}$.

Jia *et al*. [196] present approximation algorithms for Shapley values that can substantially reduce the number of times that the utility function is evaluated. First, they apply the idea of feature selection using group testing [197], [198]. For user s, let β_s be the random variable that s appears in a random sample of sellers. Then, for sellers s_i and s_j, the difference in Shapley values between s_i and s_j is

$$\begin{aligned}\psi(s_i)-\psi(s_j) &= \frac{1}{N-1}\sum_{S\in D\setminus\{s_i,s_j\}}\frac{\mathcal{U}(S\cup\{s_i\})-\mathcal{U}(S\cup\{s_j\})}{\binom{N-2}{|S|}}\\ &= E[(\beta_{s_i}-\beta_{s_j})\mathcal{U}(\beta_{s_1},\ldots,\beta_{s_j})],\end{aligned}$$

where $\mathcal{U}(\beta_{s_1},\ldots,\beta_{s_j})$ is the utility computed using the sellers appearing in the random sample. They can use group testing to first estimate the Shapley differences and then derive the Shapley value from the differences by solving a feasibility problem. They show that this algorithm is an (ϵ,δ)-approximation that evaluates the utility function at most $O(\sqrt{N}(\log N)^2)$ times. They further observe that most of the Shapley values are around the mean. Exploiting this approximate sparsity, they give an (ϵ,δ)-approximation algorithm that evaluates the utility function only $O(N(\log N)\log(\log N)$ times.

Ghorbani and Zou [160] propose a principled framework of fair data evaluation in supervised learning, and Monte-Carlo and gradient-based approximation methods. Their Monte-Carlo method follows a general idea similar to that in Jia *et al*. [196]. They generate Monte-Carlo estimates until the average empirically converges. They also argue that, in practice, it is sufficient to estimate Shapley values up to the intrinsic noise in the predictive performance on the test data set. Adding one tuple as a training data point does not significantly affect the performance of a model trained using a large training data set. Therefore, truncation can be used in practice based on the bootstrap variation on the test set. In their gradient Shapley method, they train a model using one "epoch" of the training data, and then update the model by gradient descent on one data point at a time, where the marginal contribution is the change in the performance of the model.

In general, computing Shapley values requires an exponential number of model evaluations. However, for some specific model, the computation may be reduced dramatically. For example, Jia *et al*. [199] show that for unweighted kNN classifiers, the exact computation needs only $O(N\log N)$ time and an (ϵ,δ)-approximation can be achieved in $O(N^{h(\epsilon,k)}\log N)$ time when ϵ is not too small and k is not too large. They also propose a Monte-Carlo approximation of $O(\frac{N(\log N)^2}{(\log k)^2})$ for weighted kNN classifiers. A key enabler of the progress is the specific utility function of a kNN classifier

$$\mathcal{U}_{kNN}(S)=\frac{1}{k}\sum_{i=1}^{\min\{k,|S|\}}\mathbb{1}[y_{\alpha_i(S)}=y_{\text{test}}],$$

where $\alpha_i(S)$ is the index of the training feature that is the kth closest to x_{test} among the training examples in S. Moreover, the sublinear approximation for unweighted kNN classifiers is facilitated by locality sensitive hashing [200].

Recently, Jia *et al*. [201] leverage the efficient computation of Shapley values in kNN [199] to tackle general classification problems. They propose to first train a target model, such as a deep neural network, and identify the features. Then, they conduct a model distillation to kNN by training a kNN classifier using the features to mimic the performance of the original model and tune parameter k, the number of nearest neighbors considered. Last, they apply the Shapley value estimation method in kNN [199] to approach the Shapley values in the target model.

Many classic rewarding methods, such as Shapley values, may be vulnerable to data-replication attacks. One data provider may replicate its data and act as an additional provider to obtain extra unconscionable rewards. To prevent data-replication attacks from happening, replication-robust payoff mechanisms are proposed. Han *et al*. [202] propose a fix to Shapley value based payoff mechanisms. The idea is to down-weigh the Shapley value – a data provider gets a less reward if there are multiple copies of its data in the coalitions.

Related to fairness and truthfulness in a market, cooperation among different agents in a market may happen. Building trust in a sub-community within a data marketplace becomes an interesting subject. Armstrong and Durfee [203] analyze factors that may influence the efficiency of building trust and conducting cooperation in a data market. For each agent in a market, the other agents can be divided into two categories, namely those remembered agents and those strange or forgotten agents. They have a few interesting findings. Cooperations arising from iterated interactions is inversely proportional to the rate of system mixing, the number of initially misbehaving agents, and the rate at which agents explore alternative strategies. Cooperation is also initially inversely proportional to population size. At the same time, cooperation is proportional to average member size and better estimation of the likelihood of strange agents to misbehave.

5.5 Privacy Preserving Marketplaces of Data

Privacy is a serious concern and also a critical tipping point in designing marketplaces of data. When a user shares her/his data with some others, the user may disclose her/his privacy to some extent. Therefore, it is important to explore how to protect or minimize the privacy leakage. At the same time, it is also important to understand how a seller's privacy disclosure may be properly compensated through data pricing.

Ghosh and Roth [155] design truthful marketplaces where data buyers want to purchase data to estimate statistics and sellers want compensation for their privacy loss. In the design, there is only one query and the individual evaluations of their data are private. Data owners are asked to report the costs for the use of their data. Under the assumption of differential privacy [70], [204], they transform the problem into variants of multi-unit procurement auction. They show that, when a buyer holds an accuracy goal, the classic Vickrey auction can minimize the buyer's total cost and guarantee the accuracy. When the buyer has a budget, they give an approximation algorithm to maximize the accuracy under the budget constraint.

The method by Ghosh and Roth [155] may not work well when the costs and the data are correlated. For example, a store with more customer traffic may request a higher cost in using the data. Correspondingly, reporting the cost may reveal the privacy of the store. Fleischer and Lyu [53] tackle the scenario where costs are correlated with data and propose a posted-price-like mechanism. Given a set of data sellers categorized into different types and the associated distributions of costs, the mechanism offers each user a contract with the expected payment corresponding to the type. If a seller takes the offer, the payment is determined by the seller's verifiable type and the associated payment in the contract. All sellers have the same probability to take or reject their contracts independently. The sellers are truthful, that is, a user takes the offer if the payment is larger than or equal to the privacy loss. This posted-price-like mechanism is Bayesian incentive compatible (i.e., every seller's strategy is Bayesian-Nash equilibrium), ex-interim individually rational (i.e., the expected utility is non-negative for every seller when the seller decides truthfully), $O(\epsilon^{-1})$-accurate, perfectly data private (i.e., whenever the mechanism's posterior belief about a seller's data differs from its prior belief, the mechanism pays the seller) and ϵ-differentially private.

Li *et al.* [60] tackle the same problem as Ghosh and Roth [155] do, but assume that individual valuations are public and focus on returning unbiased estimations and pricing multiple queries consistently. To address the concerns on privacy loss, they develop a theoretical framework to divide the price among data owners who contribute to the aggregate computation and thus have loss of privacy. Their framework extends several principles from both differential privacy and query pricing in data markets.

The fairness mechanism considered by Li *et al.* [60] only compensates a seller whose data are used. Niu *et al.* [54] further consider the scenario where multiple sellers' data are correlated and extend to dependent fairness. In dependent fairness, a seller s is still compensated if the data of another seller s' are used that are correlated with the data of s. They propose two approaches to privacy compensation. In the bottom-up approach, the broker first satisfies each individual seller's privacy compensation and then decides the price for the statistic selling to a buyer. In the top-down design, the broker decides the total price of a data aggregate product sold to a buyer, and then spares a fraction of the total price for privacy compensation. The privacy compensation is divided and assigned to individual data sellers by solving a budget allocation problem. Each seller receives a compensation roughly proportional to the privacy loss due to the data sharing. Niu *et al.* [205] further extend to time series data that may have temporal correlations. They adopt Pufferfish privacy [206] to measure privacy losses under temporal correlations.

While various efforts have been made to address the challenges of privacy loss compensation when user data are correlated in one way or another, as Ghosh and Roth [155] point out, in general, it is impossible for any mechanism to compensate individuals for privacy loss properly if correlations between their private data and their cost functions are unknown beforehand.

In the classical setting of physical goods [207], using contract theory [208] with hidden information, that is, unobservable types of buyers, a seller can design a set of contracts with different consumption levels to maximize revenue from buyers. Naghizadeh and Sinha [209] extend the contract design model to price a bundle of queries at different privacy levels to maximize revenue. They also consider adversarial users. Their work also adopts differential privacy [70], [204]. For a query bundle $\{Q_1, \ldots, Q_k\}$, a contract is a tuple (p, ϵ, s), where $p > 0$ is the price paid by a buyer, ϵ is the privacy budget, such that a buyer can get an answer to query Q_i $(1 \leq i \leq k)$ with ϵ_i-differential privacy guarantee, and $\epsilon \geq \sum_{i=1}^{k} \epsilon_i$, and p is the post-hoc fine to be paid if the buyer is found misusing the query answers. It is assumed that an adversarial buyer derives a benefit $C(\epsilon)$, which is monotonically increasing and convex, $C(0) = 0$. One interesting finding is that, in the traditional contract theory, if there are n types of honest buyers and one type of adversarial buyers, the seller should design up to $n + 1$ contracts. In the data marketplace situation, they show that up to n contracts are sufficient. In other words, a data seller should not design a contract for the adversary. Instead, the seller should adjust the contracts' pricing to account for the risks from adversarial users. They also design post-hoc fines in pricing query bundles that can help to reduce loss due to privacy leakage by adversarial buyers. They provide a fast approximation algorithm to compute the contracts.

A data owner has to decide a tradeoff between privacy and data utility. Li and Raghunathan [210] design an economics-based incentive-compatible mechanism for a data owner to price and disseminate private data. Specifically, let two-part tariff pricing function $R(s, x) = \alpha_s + \beta_s x$ be the price for x amount of data at sensitivity level s, where α_s and β_s are the fixed and variable price factors, respectively. Assuming two types of data users, one type for aggregate information and patterns in data and the other type for individual identity and personal information, the proposed mechanism works in four stages. First, the data owner selects a variety of sensitivity types to offer. Second, the data owner offers different prices for data with different sensitivity types. Third, a data user selects a certain sensitivity type with corresponding price, and thus reveals the user type. Last, the data user selects the optimal amount of data with the chosen sensitivity type. The core idea is that the data owner can identify the sensitive attributes in the data, such as the identifying attributes, which are not useful for aggregate analysis but necessary at individual communication. A data owner can offer a lower price for data without sensitive attributes, and charge for a higher price for data with sensitive attributes. This approach provides an orthogonal idea to the popular ways of tuning the parameter in differential privacy.

Due to the privacy concerns, when a company may have opportunities to collect data about its customers, should it do it (i.e., collecting and revealing the data) or not (i.e., a blanket policy of never collecting)? Jaisingh *et al.* [211] find that the company should not collect customer data if the total gains from trading the data cannot cover the privacy loss. In practice, there is an increasing tendency for consumers to overestimate their loss of privacy, particularly when the use of the private data is uncertain. In other cases, the company should offer two contracts on their services and products. One contract collects the customer data at a certain price, and the other contract does not collect any customer data at a different price.

While most of the studies on privacy preserving data marketplaces focus on the privacy of data owners, transactions may also disclose privacy of data buyers, such as what, when and how much they buy. For example, a retail company purchasing query results may consider what queries (e.g., the products or customer groups involved in the queries), when (e.g., the periods where the queries are concerned), and how much data it purchases as privacy, and may want to keep the information confidential from any others, including the data sellers and the broker. Aiello *et al.* [62] design a mechanism such that after making an initial deposit and maintaining a sufficient balance, a buyer can engage in an unlimited number of price-oblivious transfer protocols where the sellers and the broker cannot know anything other than the amount of interaction and the initial deposit amount. The broker even cannot know the buyer's current balance and when the buyer's balance runs out. This is achieved by adapting conditional disclosure [212] to the two-party setting.

Distribution and use of private data are another important step where privacy may leak. Hynes *et al.* [63] demonstrate Sterling, a decentralized marketplace for private data, which supports privacy-preserving distribution and use of data. The central technical idea comes from privacy-preserving smart contracts on a permissionless blockchain. To provide strong security and privacy guarantees, they combine blockchain smart contracts, trusted execution environments and differential privacy. Particularly, smart contracts allow enforcement of constraints on data usage and enables payments and rewards.

5.6 Data Pricing in Novel Applications: Dynamic Data Pricing, Online Pricing and Federated Learning Pricing

The demand of data pricing arises in many novel application scenarios. In this subsection, we particularly discuss three emerging situations: dynamic data pricing, online pricing and pricing in federated learning.

Many applications are built on dynamic and online data. How to price temporal views on data streams properly is an important issue for practical data markets. One central task is to estimate and optimize the operational costs, which are the costs to evaluate queries of different users on the fly. The pricing decisions involve not only data sellers but also data buyers. For example, suppose two data buyers b_1 and b_2 purchase two queries q_1 and q_2, such that q_2 can be written as a further selection on top of q_1 (e.g., q_1 is about all customers in North America, while q_2 keeps all the same as q_1 but focuses on only customers in Canada). The optimal pricing of q_1 and q_2 should take the advantage of the overlap between the two queries so that the sharing can save the operational costs, and, at the same time, be fair to b_1 and b_2.

Al-Kiswany *et al.* [13] propose a greedy method that enumerates all possible sharing plans and selects the one with the minimum additional cost. It does not come with any quality guarantee. Liu and Hacigümüş [213] propose an improved method that takes some risk in sharing plan. If the costs of the previous sharings are already cumulated to a high level, and the additional cost of a new sharing (i.e., the risk) is moderate and can be amortized well by the previous sharings, then the new sharing may be taken. They also give five rules to ensure fair pricing. Let $AC(S)$ be the cost attributed to a sharing S. First, for two identical sharings $S_1 = S_2$, $AC(S_1) = AC(S_2)$ should hold. Second, for any sharing S, $AC(S)$ should be no higher than the lowest cost of S if no other sharing exists. Third, for two sharings S_1 and S_2, if the query of S_1 is contained by the query of S_2, that is, the result of S_1 is a subset of the result of S_2, and the lowest cost of S_1 is smaller than the lowest cost of S_2 if no other sharing exists, then $AC(S_1) \leq AC(S_2)$. Fourth, a sharing plan with common subexpressions with other sharings should be compensated. Last, the cost of the global plan should be equal to the sum of costs attributed to all sharings.

In order to purchase dynamic data, a buyer may have to call a seller's API repeatedly. A buyer may have to pay for the same data multiple times. Upadhyaya *et al.* [214] explore how to modify APIs to achieve optimal history-aware pricing, that is, buyers are charged only once for data purchased and not updated. The central idea is the introduction of the notion of refund – a user can ask for refunds of data that she/he has bought before. For each query, the seller issues a coupon in addition to the query result, where the coupon records the identity information of the data in the query result. Specifically, a coupon $c = ((id, uid, v), \tau, \mathcal{H}(id \oplus \tau \oplus \kappa))$, where id is a tuple identifier, uid is a user-id, v is a version-id that is monotonically increasing, τ is a query identifier that is also monotonically increasing, $\mathcal{H}$ is a cryptographic hash function [215], such as SHA-1, SHA-256 and SHA-3, and κ is a secret key only known to the seller. If a buyer gets two coupons c_1 and c_2 in two different purchases such that $c_1[(tid, uid, v)] = c_2[(tid, uid, v)]$, then the buyer can ask the seller for a refund by showing the two coupons. As pointed out by Deep and Koutris [185], the refund mechanism does not provide any arbitrage-free guarantee.

Qirana [185], [186] can support history-aware pricing. To incorporate a query history, suppose a buyer already purchases queries $\mathbf{Q} = Q_1, \ldots, Q_k$ and pays for a total of $p(\mathbf{Q}, D)$ so far. When a new query Q_{k+1} comes, let the support set $\mathcal{S}_{k+1} = \{D_i \in \mathcal{S} \mid \mathbf{Q}(D_i) = \mathbf{Q}(D), Q_{k+1}(D_i) \neq Q_{k+1}(D)\}$. Then, the new total price $p((Q_1, \ldots, Q_k, Q_{k+1}), D) = p(\mathbf{Q}, D) + \sum_{D_i \in \mathcal{S}_{k+1}} w_i$. This history-aware pricing function is shown arbitrage-free.

Zheng *et al.* [216] consider online pricing for mobile crowd-sensing data markets. Different from most of the work on data markets, they assume that data providers are distributed in space and there are three types of spatial queries from buyers, namely single-data query (e.g., inquiring the value at a specific location), multi-data query (e.g., inquiring the mean in a region) and range query (e.g., inquiring the probability that the data at a region falls in a given range). The vendor uses raw data from data providers and produces a statistical model through Gaussian process to answer queries. To form different versions of data products, the vendor generates different conditional Gaussian distribution with respect to locations and uses the conditional entropy to quantify the quality of the versions. They propose a randomized online pricing strategy so that the price can be adaptive from the historical queries. They show that the pricing mechanism is arbitrage-free and is a constant factor approximation of revenue maximization.

Niu *et al.* [217] consider online data market where a query may be sold to different buyers at different time and the broker can adjust prices over time. The objective is to maximize the broker's cumulative revenue by posting reasonable prices for sequential queries. They design a contextual dynamic pricing mechanism with the reserve price constraint. The central idea is to use the properties of ellipsoid for efficient online optimization. Their method can support both linear and non-linear market value models with uncertainty.

Federated learning [218], [219] trains a machine learning model across multiple decentralized parties, where each party holds local data without any peer-wise data exchanging. The parties and their data sets are often ordered in a federated learning process. To accommodate the participation order and value data in federated learning, Wang *et al.* [220] develop federated Shapley value. Let I be the set of participants and $\mathcal{U}$ be the utility function, where $\mathcal{U}(A+B)$ is the utility of training first on A and then on B. For participant i at round t in a federated learning process, the federated Shapley value is

$$\psi_t(i) = \frac{1}{|I_t|} \sum_{S \subseteq I_t \setminus \{i\}} \frac{1}{\binom{|I_t|-1}{|S|}} [\mathcal{U}(I_{1:t-1} + S \cup \{i\})) - \mathcal{U}(I_{1:t-1} + S)],$$

if $i \in I_t$ and $\psi_t(i) = 0$ otherwise. The federated Shapley value of a party is the sum of the values of all rounds, that is, $\psi(i) = \sum_{t=1}^{T} \psi_t(i)$. Wang *et al.* [220] show that the federated Shapley values have instantaneous group rationality, that is, $\sum_{i \in I_t} \psi_t(i) = \mathcal{U}(I_{1:t}) - \mathcal{U}(I_{1:t-1})$. The fairness is guaranteed at each round. That is, for any two parties i and j, $\psi_t(i) = \psi_t(j)$ at round t if $\forall S \subseteq I_t \setminus \{i, j\}$, $\mathcal{U}(I_{1:t-1} + (S \cup \{i\})) = \mathcal{U}(I_{1:t-1} + (S \cup \{j\}))$. Moreover, for any party i at round t, $\psi_t(i) = 0$ if $\forall S \subseteq I_t \setminus \{i\}$, $\psi(I_{1:t-1} + (S \cup \{i\})) = \mathcal{U}(I_{1:t-1} + S)$. They also extend the previous Shapley value approximation techniques to compute federated Shapley values.

Sim *et al.* [221] consider the more general situation of collaborative machine learning and advocate using information gain as the utility function. For a model θ trained on data D, the information gain $\mathbb{I}(\theta; D) = \mathbb{H}(\theta) - \mathbb{H}(\theta|D)$, which is the reduction in uncertainty. They generalize to ρ-Shapley fairness by assigning a reward $r_i = k\psi_i^{\rho}$ to a party i. By tuning parameter ρ, they can trade off among Shapley fairness, individual rationality, stability of the grand coalition and group welfare.

Hu and Gong [222] consider privacy leaking in federated learning and design an incentive mechanism to compensate the costs of privacy leakage of the users that are most likely to provide reliable data. Their problem is formulated in a two-stage Stackelberg game [223]. Richardson *et al.* [166] use influence functions to reward data contributions to linear regression in the federated learning setting.

5.7 Summary

In this section, we review the topic of pricing data products. We first analyze the structures, players, and ways to produce data products in data marketplaces. Then, we examine several important areas in pricing data products, including arbitrage-free pricing, revenue maximization pricing, fair and truthful pricing and privacy preserving pricing. We also discuss how to price dynamic data and online pricing. When pricing data products in a data marketplace, those several considerations are typically incorporated and integrated in one way or another.

6 Discussion and Open Challenges

Data pricing comes from practical demands and has been tackled in multiple disciplines. Although there is a rich body of literature addressing a series of issues in data pricing, there are still many questions remained unexplored. In this section, we discuss some interesting challenges for possible future work. By no means our list is exhaustive. Instead, we hope our discussion can intrigue more extensive interest and research effort into this fast growing area.

6.1 Data Supply Chain: A Grand Challenge

At the macro level, although many studies focus on different steps in data marketplaces, we clearly observe a lack of systematic investigation on data supply chains and development of end-to-end solutions. As data products are abundant and diversified, to develop ecologically sustainable marketplaces, supply chains of data products have to be built. Here, we introduce and advocate the notion of *data supply chains*, which connect all parties involved in data production and consumption, including data providers, data processors, data analysts, data product and services consumers and other possible roles. Each party in a data supply chain connects its upstream providers and its downstream consumers, provides its value-added contributions and obtains rewards. Feedback mechanisms through pricing and marketing have to be created in a data supply chain so that supply and consumption can be matched, coordinated and balanced. Most of those problems are not thoroughly thought about.

Although the notion of data supply chain is not mentioned in literature, some specific trends and challenges are discussed sporadically. For example, Muschalle *et al.* [145] identify some trends and challenges in data consumption and marketplaces. First, they assert that many essential data processing tasks are essential for data markets, such as labeling, annotating and aggregating data. Second, data markets will be integrated with numerous application domains. To enable domain data markets, it is important to customize general data processing technologies for niche domains. Third, customers want to have data faster. Thus, it is important to create online data query services and develop corresponding pricing models. Fourth, as there are more data, more data providers and more analysts, a data product may be substituted by others. To hatch a healthy ecological data marketplace, it is important to establish standard data processing mashups to facilitate data product substitution. Fifth, to maintain a fair data market overall, it is important to provide price transparency so that data product providers have to optimize their data and data processing/analysis services. Last, customer preferences and experience are critical for data markets.

Recently, Acemoglu *et al.* [224] present an insightful study on the ecological effect of data markets. They demonstrate that a user's sharing of data may likely reveal some other users' privacy and depress the price of other users' data. The depressed prices lead to excessive data sharing

and thus further reduce welfare. Their study suggests the need of mediation in data sharing in data markets.

Most recently, Fernandez *et al.* [225] analyze the challenges and propose a research agenda around constructing a data market platform to address the sharing, discovery and integration of data among many parties. Their big picture covers both market design and system development. The focus is to create the incentives and mechanisms to connect data supply and demand. As the middlemen, arbiters build data mashups to match data supply and demand. The market platforms advocated by the authors can be regarded as the data exchange mechanisms in data supply chain.

One challenge associated with the macro view of data supply chain is the interdisciplinary nature of data pricing research. As can be observed in this article, data pricing is studied in many different disciplines, such as economics, marketing, electronic commerce, data management, data mining and machine learning. The communication and dialog among different areas have to be strengthened.

6.2 Some Technical Challenges at the Micro Level

At the micro level, there are many research problems remained open. We name a few examples of fundamental problems.

First, most of the studies suggest relative prices of data products. Very few studies connect theoretical models with data pricing practice and investigate absolute prices of data products and their marketing effect. As data pricing is a market mechanism and user behavior in practice is hard to modeled completely, experimental studies of data pricing models are essential and should be connected to theoretical investigations.

Second, pricing is based on valuation and equilibrium among multiple parties. Different parties may have different valuation on data, data products and data services. It is important to systematically establish the principles of value assessment for various parties in data marketplaces, such as data providers, data owners, data users, and data brokers. Moreover, it is important to understand what messages are passed to different parties in data marketplaces through data pricing actions, and how. So far, value assessment of data and negotiations among different parties in data marketplaces are largely not analyzed in detail.

Third, many pricing models are proposed in literature. It is important to understand how data pricing models and their assumptions can be implemented and enforced in practice. Specifically, accounting and auditing in data marketplaces are critical to achieve transparency in data pricing and efficiency in data marketplaces. Accounting and auditing in data marketplaces, however, are interesting problems that have not been investigated in depth yet. We need principles, quality guarantees and designs of operational procedures for accounting and auditing in data pricing, transactions and adversary detection.

Fourth, most of the studies on data pricing develop general models. At the same time, as data science transforms many application domains, data pricing has to deal with specific applications. Mechanisms, regulations and constraints in a specific domain may facilitate data pricing in some aspects, and post challenges in some other aspects. For example, Jia *et al.* [199] show that, although fair pricing in general is exponential in computation time but can be achieved polynomially in kNN models (Section 5.4). It is interesting and highly desirable to explore fairness, truthfulness, and privacy preservation of data pricing in specific applications.

Last but not least, almost all applications are dynamic in nature. The values of data, data products and data services may also evolve over time. The changes may be caused by the updates in demands and supplies. It is important to develop mechanisms to capture and monitor changes in demand and supply of data, data products and data services, and explore corresponding dynamic pricing.

Acknowledgments

This research is supported in part by the NSERC Discovery Grant program. All opinions, findings, conclusions and recommendations in this paper are those of the author and do not necessarily reflect the views of the funding agencies.

References

[1] S. van de Sandt, S. Dallmeier-Tiessen, A. Lavasa, and V. Petras, "The definition of reuse," *Data Sci. J.*, vol. 18, no. 1, 2019, Art. no. 22.
[2] A. Nagaraj, "The private impact of public information: Landsat satellite maps and gold exploration," 2016.
[3] P. Kotler, *Marketing Management: The Millennium Edition*. Boston, MA, USA: Pearson, 2000.
[4] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, "A survey on big data market: Pricing, trading and protection," *IEEE Access*, vol. 6, pp. 15132–15154, 2018.
[5] S. A. Fricker and Y. V. Maksimov, "Pricing of data products in data marketplaces," in *Software Business*, A. Ojala, H. Holmström Olsson, and K. Werder, Eds. Cham, Switzerland: Springer, 2017, pp. 49–66.
[6] M. Zhang and F. Beltran, "A survey of data pricing methods," *SSRN*, Apr. 2020. [Online]. Available: https://ssrn.com/abstract=3609120
[7] C. Wu, R. Buyya, and K. Ramamohanarao, "Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges," *ACM Comput. Surv.*, vol. 52, no. 6, Oct. 2019, Art. no. 108.
[8] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "A survey of smart data pricing: Past proposals, current plans, and future trends," *ACM Comput. Surv.*, vol. 46, no. 2, Nov. 2013, Art. no. 15.
[9] M. K. M. Murthy, H. A. Sanjay, and J. P. Ashwini, "Pricing models and pricing schemes of IaaS providers: A comparison study," in *Proc. Int. Conf. Advances Comput. Commun. Inform.*, 2012, pp. 143–147.
[10] X. Wu, W. Zhang, and W. Dou, "Pricing as a service: Personalized pricing strategy in cloud computing," in *Proc. IEEE 12th Int. Conf. Comput. Inf. Technol.*, 2012, pp. 1119–1124.
[11] M. Aazam and E. Huh, "Broker as a service (BaaS) pricing and resource estimation model," in *Proc. IEEE 6th Int. Conf. Cloud Comput. Technol. Sci.*, 2014, pp. 463–468.
[12] V. V. Kantere, D. Dash, G. Gratsias, and A. Ailamaki, "Predicting cost amortization for query services," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2011, pp. 325–336.
[13] S. Al-Kiswany, H. Hacigümüş, Z. Liu, and J. Sankaranarayanan, "Cost exploration of data sharings in the cloud," in *Proc. 16th Int. Conf. Extending Database Technol.*, 2013, pp. 601–612.
[14] S. Dibb, L. Simkin, W. M. Pride, and O. Ferrell, *Marketing: Concepts and Strategies*. 5th ed., Abingdon, U.K.: Houghton Mifflin, April 2005.
[15] T. T. Nagle and J. Hogan, *The Strategy and Tactics of Pricing: A Guide to Growing More Profitably*. Englewood Cliffs, NJ, USA: Prentice Hall, 2010.
[16] R. Brennan, L. Canning, and R. Mcdowell, *Business-to-Business Marketing*. Newbury Park, CA, USA: Sage Publications, 2013.
[17] M. Neumeier, *The Brand Flip: Why Customers Now Run Companies–and How to Profit From It*. San Francisco, CA, USA: New Riders, 2015.
[18] G. Irvin, *Modern Cost-Benefit Methods*. London, U.K.: Macmillan, 1978.
[19] C. Shapiro, S. Carl, H. Varian, and H. B. Press, *Information Rules: A Strategic Guide to the Network Economy*. Brighton, MA, USA: Harvard Business School Press, 1998.
[20] A. Goldfarb and C. Tucker, "Digital economics," *J. Econ. Literature*, vol. 57, no. 1, pp. 3–43, Mar. 2019.
[21] E. Brynjolfsson and M. D. Smith, "Frictionless commerce? a comparison of internet and conventional retailers," *Manage. Sci.*, vol. 46, no. 4, pp. 563–585, 2000.

[22] H. Yang, "Targeted search and the long tail effect," *RAND J. Econ.*, vol. 44, no. 4, pp. 733–756, Dec. 2013.

[23] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*. New York, NY, USA: Hyperion, 2006.

[24] M. Gentzkow and J. M. Shapiro, "Ideological segregation online and offline," *Quart. J. Econ.*, vol. 126, no. 4, pp. 1799–1839, 11 2011.

[25] C. Sunstein, *Echo Chambers: Bush V. Gore, Impeachment, and Beyond*. Princeton, NJ, USA: Princeton Univ. Press, 2001.

[26] B. Jullien, "Two-sided b to b platforms," in *The Oxford Handbook of the Digital Economy*, M. Peitz and J. Waldfogel, Eds. London, U.K: Oxford Univ. Press, 2012.

[27] H. Halaburda and Y. Yehezkel, "Platform competition under asymmetric information," *Amer. Econ. J., Microeconomics*, vol. 5, no. 3, pp. 22–68, 2013.

[28] A. Gilchrist, *Industry 4.0: The Industrial Internet of Things*, 1st ed. New York, NY, USA: Apress, 2016.

[29] B. Squire, S. Brown, J. Readman, and J. Bessant, "The impact of mass customisation on manufacturing trade-offs," *Prod. Operations Manage.*, vol. 15, pp. 10–21, 2009.

[30] Y. Bakos and E. Brynjolfsson, "Bundling information goods: Pricing, profits, and efficiency," *Manage. Sci.*, vol. 45, no. 12, pp. 1613–1630, Dec. 1999.

[31] Y. Bakos and E. Brynjolfsson, "Bundling and competition on the internet," *Marketing Sci.*, vol. 19, no. 1, pp. 63–82, Feb. 2000.

[32] L. Aguiar and J. Waldfogel, "As streaming reaches flood stage, does it stimulate or depress music sales?" *Int. J. Ind. Org.*, vol. 57, no. C, pp. 278–307, 2018.

[33] J. Lerner, P. A. Pathak, and J. Tirole, "The dynamics of open-source contributors," *Amer. Econ. Rev.*, vol. 96, no. 2, pp. 114–118, May 2006.

[34] J. Waldfogel, "Copyright research in the digital age: Moving from piracy to the supply of new products," *Amer. Econ. Rev.*, vol. 102, no. 3, pp. 337–42, May 2012.

[35] H. L. Williams, "Intellectual property rights and innovation: Evidence from the human genome," *J. Political Economy*, vol. 121, no. 1, pp. 1–27, 2013.

[36] A. Acquisti and C. Tucker, "Guns, privacy, and crime," Working paper, 2011. [Online]. Available: https://www.heinz.cmu.edu/~acquisti/papers.htm

[37] J. M. Rao and D. H. Reiley, "The economics of spam," *J. Econ. Perspectives*, vol. 26, no. 3, pp. 87–110, Sep. 2012.

[38] T. Moore, R. Clayton, and R. Anderson, "The economics of online crime," *J. Econ. Perspectives*, vol. 23, no. 3, pp. 3–20, Sep. 2009.

[39] T. L. Friedman, *The World is Flat: A Brief History of the Twenty-First Century | Thomas L. Friedman.*, 1st ed. New York, NY, USA: Farrar, Straus and Giroux, 200.

[40] F. Ferreira and J. Waldfogel, "Pop internationalism: Has half a century of world music trade displaced local culture?" *Econ. J.*, vol. 123, no. 569, pp. 634–664, Jun. 2013.

[41] N. Gandal, "Native language and internet usage," *Int. J. Sociology Lang.*, vol. 2006, no. 182, pp. 25–40, 2006.

[42] X. M. Zhang and F. Zhu, "Group size and incentives to contribute: A natural experiment at chinese wikipedia," *Amer. Econ. Rev.*, vol. 101, no. 4, pp. 1601–15, Jun. 2011.

[43] L. Chiou and C. Tucker, "Content aggregation by platforms: The case of the news media," *J. Econ. Manage. Strategy*, vol. 26, no. 4, pp. 782–805, 2017.

[44] A. Odlyzko, "Privacy, economics, and price discrimination on the internet," in *Proc. 5th Int. Conf. Electron. Commerce*, 2003, pp. 355–366.

[45] D. Fudenberg and J. M. Villas-Boas, "Price discrimination in the digital economy," in *The Oxford Handbook of the Digital Economy*, M. Peitz and J. Waldfogel, Eds. London, U.K.: Oxford Univ. Press, 2012.

[46] C. R. Taylor, "Consumer privacy and the market for customer information," *RAND J. Economics*, vol. 35, no. 4, pp. 631–650, 2004.

[47] C. Shapiro and H. R. Varian, "Versioning: The smart way to sell information," *Harvard Business Rev.*, vol. 76, no. 6, pp. 106–114, Nov./Dec. 1998.

[48] D. S. Evans, "The online advertising industry: Economics, evolution, and privacy," *J. Econ. Perspectives*, vol. 23, no. 3, pp. 37–60, Sep. 2009.

[49] N. Arnosti, M. Beck, and P. Milgrom, "Adverse selection and auction design for internet display advertising," in *Proc. 16th ACM Conf. Economics Comput.*, 2015, Art. no. 167.

[50] A. Ockenfels, D. Reiley, and A. Sadrieh, "Online auctions," National Bureau of Economic Research, Working Paper 12785, Dec. 2006.

[51] L. Einav, C. Farronato, J. Levin, and N. Sundaresan, "Auctions versus posted prices in online markets," *J. Political Economy*, vol. 126, no. 1, pp. 178–215, 2018.

[52] A. Acquisti, C. Taylor, and L. Wagman, "The economics of privacy," *J. Econ. Literature*, vol. 54, no. 2, pp. 442–92, Jun. 2016.

[53] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proc. 13th ACM Conf. Electron. Commerce*, 2012, pp. 568–585.

[54] C. Niu, Z. Zheng, F. Wu, S. Tang, X. Gao, and G. Chen, "Unlocking the value of privacy: Trading aggregate statistics over private correlated data," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 2031–2040.

[55] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," *Proc. VLDB Endowment*, vol. 4, no. 12, pp. 1482–1485, 2011.

[56] M. Balazinska, B. Howe, P. Koutris, D. Suciu, and P. Upadhyaya, "A discussion on pricing relational data," in *Search of Elegance in the Theory and Practice of Computation: Essays Dedicated to Peter Buneman*, V. Tannen, L. Wong, L. Libkin, W. Fan, W.-C. Tan, and M. Fourman, Eds. Berlin, Germany: Springer, 2013, pp. 167–173.

[57] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *Proc. ACM Conf. Economics Comput.*, 2019, pp. 701–726.

[58] S. Chawla, S. Deep, P. Koutrisw, and Y. Teng, "Revenue maximization for query pricing," *Proc. VLDB Endowment*, vol. 13, no. 1, pp. 1–14, Sep. 2019.

[59] L. S. Shapley, "A value for n-Person games," RAND Corporation, Santa Monica, CA, Tech. Rep. P-295, 1952.

[60] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," *ACM Trans. Database Syst.*, vol. 39, no. 4, Dec. 2015, Art. no. 34.

[61] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Conf. Secur. Symp.*, 2016, pp. 601–618.

[62] W. Aiello, Y. Ishai, and O. Reingold, "Priced oblivious transfer: How to sell digital goods," in *Proc. Int. Conf. Theory Appl. Cryptogr. Techn.*, 2001, pp. 119–135.

[63] N. Hynes, D. Dao, D. Yan, R. Cheng, and D. Song, "A demonstration of sterling: A privacy-preserving data marketplace," *Proc. VLDB Endowment*, vol. 11, no. 12, pp. 2086–2089, Aug. 2018.

[64] D. Dao, D. Alistarh, C. Musat, and C. Zhang, "Databright: Towards a global exchange for decentralized data ownership and trusted computation," *CoRR*, vol. abs/1802.04780, 2018.

[65] K. Nissim, S. Vadhan, and D. Xiao, "Redrawing the boundaries on purchasing data from privacy-sensitive individuals," in *Proc. 5th Conf. Innovations Theor. Comput. Sci.*, 2014, pp. 411–422.

[66] A. Ghosh, K. Ligett, A. Roth, and G. Schoenebeck, "Buying private data without verification," in *Proc. 15th ACM Conf. Economics Comput.*, 2014, pp. 931–948.

[67] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, Jun. 2010, Art. no. 14.

[68] C. C. Aggarwal and P. S. Yu, "Privacy-preserving data mining: A survey," in *Handbook of Database Security: Applications and Trends*, M. Gertz and S. Jajodia, Eds. Boston, MA, USA: Springer, 2008, pp. 431–460.

[69] E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacy preserving data mining algorithms," in *Privacy-Preserving Data Mining: Models and Algorithms*, C. C. Aggarwal and P. S. Yu, Eds. Boston, MA, USA: Springer, 2008, pp. 183–205.

[70] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Germany: Springer, 2008, pp. 1–19.

[71] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *SIGKDD Explor. Newsl.*, vol. 10, no. 2, pp. 12–22, Dec. 2008.

[72] M. A. Ferrag, L. Maglaras, and A. Ahmim, "Privacy-preserving schemes for ad hoc social networks: A survey," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 4, pp. 3015–3045, Fourth Quarter 2017.

[73] X. Wu, X. Ying, K. Liu, and L. Chen, "A survey of privacy-preservation of graphs and social networks," in *Managing and Mining Graph Data*, C. C. Aggarwal and H. Wang, Eds. Boston, MA, USA: Springer, 2010, pp. 421–453.

[74] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng, "Differential privacy and its applications in social network analysis: A survey," *ArXiv*, vol. abs/2010.02973, 2020.

[75] A. V. Goldberg, J. D. Hartline, and A. Wright, "Competitive auctions and digital goods," in *Proc. 12th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2001, pp. 735–744.

[76] A. Lambrecht *et al.*, "How do firms make money selling digital goods online?" *Marketing Lett.*, vol. 25, pp. 331–341, 2014.

[77] J. M. Gallaugher, P. Auger, and A. BarNir, "Revenue streams and digital content providers: an empirical investigation," *Inf. Manage.*, vol. 38, no. 7, pp. 473–485, 2001.

[78] A. Prasad, V. Mahajan, and B. Bronnenberg, "Advertising versus pay-per-view in electronic media," *Int. J. Res. Marketing*, vol. 20, no. 1, pp. 13–30, 2003.

[79] K. Pauwels and A. Weiss, "Moving from free to fee: How online firms market to change their business model successfully," *J. Marketing*, vol. 72, no. 3, pp. 14–31, 2008.

[80] T. Wagner, A. Benlian, and T. Hess, "Converting freemium customers from free to premium–the role of the perceived premium fit in the case of music as a service," *Electronic Markets*, vol. 24, pp. 259–268, 2014.

[81] L. Chiou and C. Tucker, "Paywalls and the demand for news," *Inf. Economics Policy*, vol. 25, no. 2, pp. 61–69, 2013.

[82] A. Boom, "download for free": When do providers of digital goods offer free samples?," School of Bus. & Econ., Free University Berlin, Discussion Papers 2004/28, 2004. [Online]. Available: https://EconPapers.repec.org/RePEc:zbw:fubsbe:200428

[83] A. Rao, "Online Content Pricing: Purchase and Rental Markets," *Marketing Sci.*, vol. 34, no. 3, pp. 430–451, May 2015.

[84] S. Athey, E. Calvano, and J. Gans, "The impact of the internet on advertising markets for news media," National Bureau of Economic Research, Working Paper 19419, Sep. 2013. [Online]. Available: http://www.nber.org/papers/w19419

[85] S. Lehmann and P. Buxmann, "Pricing strategies of software vendors," *Bus. Inf. Syst. Eng.*, vol. 1, pp. 452–462, 2009.

[86] J. Benfield and W. Szlemko, "Internet-based data collection: Promises and realities," *J. Res. Practice*, vol. 2, no. 2, 2006, Art. D1.

[87] S. J. Best and B. S. Krueger, *Internet Data Collection*. Thousand Oaks, CA, USA: SAGE, 2004.

[88] H. Elmeleegy *et al.*, "Overview of turn data management platform for digital advertising," *Proc. VLDB Endowment*, vol. 6, no. 11, pp. 1138–1149, Aug. 2013.

[89] Y. Yang, X. Mao, J. Pei, and X. He, "Continuous influence maximization: What discounts should we offer to social network users?" in *Proc. Int. Conf. Manage. Data*, 2016, pp. 727–741.

[90] D. Bergemann and A. Bonatti, "Selling cookies," *Amer. Econ. J.: Microeconomics*, vol. 7, no. 3, pp. 259–94, Aug. 2015.

[91] R. Lewis and J. Rao, "On the near impossibility of measuring the returns to advertising," *SSRN Electron. J.*, 2013, 10.2139/ssrn.2367103.

[92] B. R. Gordon, F. Zettelmeyer, N. Bhargava, and D. Chapsky, "A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook," *Marketing Sci.*, vol. 38, no. 2, pp. 193–225, 2019.

[93] Y. Yang, Q. S. Lu, G. Tang, and J. Pei, "The impact of market competition on search advertising," *J. Interactive Marketing*, vol. 30, no. C, pp. 46–55, 2015.

[94] G. Tang, Y. Yang, and J. Pei, "Price information patterns in web search advertising: An empirical case study on accommodation industry," in *Proc. IEEE Int. Conf. Data Mining*, 2013, pp. 737–746.

[95] A. Lambrecht and C. Tucker, "When does retargeting work? information specificity in online advertising," *J. Marketing Res.*, vol. 50, no. 5, pp. 561–576, 2013.

[96] S. Athey, E. Calvano, and J. S. Gans, "The impact of consumer multi-homing on advertising markets and media competition," *Manage. Sci.*, vol. 64, pp. 1574–1590, 2018.

[97] W. J. Adams and J. L. Yellen, "Commodity bundling and the burden of monopoly," *Quart. J. Economics*, vol. 90, no. 3, pp. 475–498, 1976.

[98] D. Menicucci, S. Hurkens, and D.-S. Jeon, "On the optimality of pure bundling for a monopolist," *J. Math. Econ.*, vol. 60, pp. 33–42, 2015.

[99] A. Pavan, I. Segal, and J. Toikka, "Dynamic mechanism design: A myersonian approach," *Econometrica*, vol. 82, no. 2, pp. 601–653, 2014.

[100] M. Armstrong, "A more general theory of commodity bundling," *J. Econ. Theory*, vol. 148, no. 2, pp. 448–472, 2013.

[101] R. B. Myerson, "Optimal auction design," *Math. Operations Res.*, vol. 6, no. 1, pp. 58–73, Feb. 1981.

[102] C. Daskalakis, A. Deckelbaum, and C. Tzamos, "Strong duality for a multiple-good monopolist," *Econometrica*, vol. 85, no. 3, pp. 735–767, 2017.

[103] N. Haghpanah and J. Hartline, "When is pure bundling optimal?," PA State Univ., Working Paper, April 2020.

[104] N. Haghpanah and J. Hartline, "Reverse mechanism design," in *Proc. 16th ACM Conf. Econ. Comput.*, 2015, pp. 757–758.

[105] M.-F. Balcan, A. Blum, and Y. Mansour, "Item pricing for revenue maximization," in *Proc. 9th ACM Conf. Electron. Commerce*, 2008, pp. 50–59.

[106] T. Abdallah, "On the benefit (or cost) of large-scale bundling," *Prod. Operations Manage.*, vol. 28, no. 4, pp. 955–969, 2019.

[107] S. Alaei, A. Makhdoumi, and A. Malekian, "Optimal subscription planning for digital goods," *SSRN Electron. J.*, 2019. [Online]. Available: https://ssrn.com/abstract=3476296

[108] M. Shubik, "Auctions, bidding, and markets: An historical sketch," in *Auctions, Bidding, and Contracting*, M. Shubik and J. Stark, Eds. New York, NY, USA: New York Univ. Press, 1983, pp. 33–52.

[109] P. Klemperer, "Auction theory: A guide to the literature," *J. Econom. Surv.*, vol. 13, no. 3, pp. 227–286, 1999.

[110] R. Engelbrecht-Wiggans, "Auctions and bidding models: A survey," *Manage. Sci.*, vol. 26, no. 2, pp. 119–142, 1980.

[111] P. Bajari and A. Hortacsu, "Economic insights from internet auctions," *J. Econ. Literature*, vol. 42, no. 2, pp. 457–486, Jun. 2004.

[112] R. P. McAfee and J. McMillan, "Auctions and Bidding," *J. Econ. Literature*, vol. 25, no. 2, pp. 699–738, Jun. 1987.

[113] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *J. Finance*, vol. 16, no. 1, pp. 8–37, 1961.

[114] J. Riley and W. F. Samuelson, "Optimal auctions," *Amer. Econ. Rev.*, vol. 71, no. 3, pp. 381–392, 1981.

[115] W. Vickrey, "Auctions and bidding games," in *Recent Advances in Game Theory*. Princeton, New Jersey. USA: Princeton Univ. Conf., 1962, pp. 15–27.

[116] W. Hu and A. Bolivar, "Online auctions efficiency: A survey of ebay auctions," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 925–934.

[117] S. Lahaie, D. M. Pennock, A. Saberi, and R. V. Vohra, "Sponsored search auctions," in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2007, pp. 699–716.

[118] H. R. Varian, "Online ad auctions," *Amer. Econ. Revi.*, vol. 99, no. 2, pp. 430–34, May 2009.

[119] T. Qin, W. Chen, and T.-Y. Liu, "Sponsored search auctions: Recent advances and future directions," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 4, Jan. 2015, Art. no. 60.

[120] J. Jansen and T. Mullen, "Sponsored search: An overview of the concept, history, and technology," *Int. J. Electron. Bus.*, vol. 6, pp. 114–131, 2008.

[121] B. Edelman and M. Ostrovsky, "Strategic bidder behavior in sponsored search auctions," *Decis. Support Syst.*, vol. 43, no. 1, pp. 192–198, Feb. 2007.

[122] B. Edelman, M. Ostrovsky, and M. Schwarz, "Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords," *Amer. Econ. Rev.*, vol. 97, no. 1, pp. 242–259, Mar. 2007.

[123] Y.-K. Che, S. Choi, and J. Kim, "An experimental study of sponsored-search auctions," *Games Econ. Behav.*, vol. 102, pp. 20–43, 2017.

[124] K. Ganchev, A. Kulesza, J. Tan, R. Gabbard, Q. Liu, and M. Kearns, "Empirical price modeling for sponsored search," in *Internet and Network Economics*, X. Deng and F. C. Graham, Eds. Berlin, Germany: Springer, 2007, pp. 541–548.

[125] D. Davydov, S. Izmalkov, and A. Smirnov, "Sponsored-Search Auctions: Empirical and Experimental Works," *J. New Econ. Assoc.*, vol. 28, no. 4, pp. 56–73, 2015.

[126] J. Auerbach, J. Galenson, and M. Sundararajan, "An empirical analysis of return on investment maximization in sponsored search auctions," in *Proc. 2nd Int. Workshop Data Mining Audience Intell. Advertising*, 2008, pp. 1–9.

[127] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, and Y. Yu, "Deep landscape forecasting for real-time bidding advertising," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 363–372.

[128] J. Zhao, G. Qiu, Z. Guan, W. Zhao, and X. He, "Deep reinforcement learning for sponsored search real-time bidding," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1021–1030.

[129] U. Feige, A. Flaxman, J. D. Hartline, and R. Kleinberg, "On the competitive ratio of the random sampling auction," in *Proc. 1st Int. Conf. Internet Netw. Econ.*, 2005, pp. 878–886.

[130] S. Alaei, A. Malekian, and A. Srinivasan,"On random sampling auctions for digital goods," in *Proc. 10th ACM Conf. Electron. Commerce*, 2009, pp. 187–196.
[131] J. D. Hartline and R. McGrew, "From optimal limited to unlimited supply auctions," in *Proc. 6th ACM Conf. Electron. Commerce*, 2005, pp. 175–182.
[132] A. V. Goldberg and J. D. Hartline, "Competitive auctions for multiple digital goods," in *Proc. 9th Annu. Eur. Symp. Algorithms*, 2001, pp. 416–427.
[133] G. Aggarwal, A. Fiat, A. V. Goldberg, J. D. Hartline, N. Immorlica, and M. Sudan, "Derandomization of auctions," in *Proc. 37th Annu. ACM Symp. Theory Comput.*, 2005, pp. 619–625.
[134] E. Maskin and J. Riley, "Asymmetric Auctions," *Rev. Econ. Stud.*, vol. 67, no. 3, pp. 413–438, 2000.
[135] T. Ebert, "Applications of recursive operators to randomness and complexity," Ph.D. dissertation, Dep. Math. Univ. California, Santa Barbara, CA, 1998.
[136] A. V. Goldberg and J. D. Hartline, "Envy-free auctions for digital goods," in *Proc. 4th ACM Conf. Electron. Commerce*, 2003, pp. 29–35.
[137] A. V. Goldberg and J. D. Hartline, "Competitiveness via consensus," in *Proc. 14th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2003, pp. 215–222.
[138] A. Archer, C. Papadimitriou, K. Talwar, and E. Tardos, "An approximate truthful mechanism for combinatorial auctions with single parameter agents," in *Proc. 14th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2003, pp. 205–214.
[139] R. Lavi and N. Nisan, "Competitive analysis of incentive compatible on-line auctions," in *Proc. 2nd ACM Conf. Electron. Commerce*, 2000, pp. 233–241.
[140] Z. Bar-Yossef, K. Hildrum, and F. Wu, "Incentive-compatible online auctions for digital goods," in *Proc. 13th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2002, pp. 964–970.
[141] F. Schomm, F. Stahl, and G. Vossen, "Marketplaces for data: An initial survey," *SIGMOD Record*, vol. 42, no. 1, pp. 15–26, May 2013.
[142] K. Pantelis and L. Aija, "Understanding the value of (big) data," in *Proc. IEEE Int. Conf. Big Data*, 2013, pp. 38–42.
[143] E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[144] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968.
[145] A. Muschalle, F. Stahl, A. Löser, and G. Vossen, "Pricing approaches for data markets," in *Enabling Real-Time Business Intelligence*, M. Castellanos, U. Dayal, and E. A. Rundensteiner, Eds. Berlin, Germany: Springer, 2013, pp. 129–144.
[146] S. Wu and P. Pavlou, "On the optimal fixed-up-to pricing for information services," *J. Assoc. Inf. Syst.*, vol. 20, no. 10, pp. 1447–1474, Jan. 2019.
[147] S. Wu and R. Banker, "Best pricing strategy for information services," *J. Assoc. Inf. Syst.*, vol. 11, no. 6, pp. 339–366, Jan. 2010.
[148] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[149] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349.
[150] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
[151] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [review article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.
[152] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in Big Data analytics," *J. Big Data*, vol. 2, no. 1, Feb. 2015, Art. no. 1.
[153] M. Babaioff, R. Kleinberg, and R. Paes Leme, "Optimal mechanisms for selling information," in *Proc. 13th ACM Conf. Electron. Commerce*, 2012, pp. 92–109.
[154] R. Cummings, K. Ligett, A. Roth, Z. S. Wu, and J. Ziani, "Accuracy for sale: Aggregating data with a variance constraint," in *Proc. Conf. Innovations Theor. Comput. Sci.*, 2015, pp. 317–324.
[155] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. 12th ACM Conf. Electron. Commerce*, 2011, pp. 199–208.
[156] K. Ligett and A. Roth, "Take it or leave it: Running a survey when privacy comes at a cost," in *Proc. 8th Int. Workshop Internet Netw. Economics*, 2012, pp. 378–391.
[157] D. Bergemann, A. Bonatti, and A. Smolin, "The design and price of information," *Amer. Econ. Rev.*, vol. 108, no. 1, pp. 1–48, Jan. 2018.
[158] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, Feb. 1977.
[159] R. Cook and S. Weisberg, *Residuals and Influence in Regression*, New York, NY, USA: Chapman & Hall, 1982.
[160] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 2242–2251.
[161] A. Ghorbani, M. P. Kim, and J. Zou, "A distributional framework for data valuation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3535–3544.
[162] Y. Kwon, M. A. Rivas, and J. Zou, "Efficient computation and analysis of distributional shapley values," *ArXiv*, vol. abs/2007.01357, 2020. [Online]. Available: https://arxiv.org/abs/2007.01357
[163] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1885–1894.
[164] Z. Wang, H. Zhu, Z. Dong, X. He, and S. Huang, "Less is better: Unweighted data subsampling via influence function," *CoRR*, vol. abs/1912.01321, 2019. [Online]. Available: https://arxiv.org/abs/1912.01321
[165] Y. Cai, C. Daskalakis, and C. Papadimitriou, "Optimum statistical estimation with strategic data sources," in *Proc. 28th Conf. Learn. Theory*, 2015, pp. 280–296.
[166] A. Richardson, A. Filos-Ratsikas, and B. Faltings, "Rewarding high-quality data via influence functions," *CoRR*, vol. abs/1908.11598, 2019. [Online]. Available: https://arxiv.org/abs/1908.11598
[167] J. Yoon, S. Arik, and T. Pfister, "Data valuation using reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 10842–10851.
[168] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, Apr. 2002.
[169] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.
[170] J. Heckman, E. Peters, N. G. Kurup, E. Boehmer, and M. Davaloo, "A pricing model for data markets," in *Proc. iConf.*, 2015.
[171] H. Yu and M. Zhang, "Data pricing strategy based on data quality," *Comput. Ind. Eng.*, vol. 112, pp. 1–10, 2017.
[172] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with querymarket," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 613–624.
[173] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proc. 31st ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, 2012, pp. 167–178.
[174] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," *J. ACM*, vol. 62, no. 5, Nov. 2015.
[175] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Querymarket demonstration: Pricing for online data markets," *Proc. VLDB Endowment*, vol. 5, no. 12, pp. 1962–1965, Aug. 2012.
[176] A. Nash, L. Segoufin, and V. Vianu, "Views and queries: Determinacy and rewriting," *ACM Trans. Database Syst.*, vol. 35, no. 3, Jul. 2010, Art. no. 21.
[177] A. Nash, L. Segoufin, and V. Vianu, "Determinacy and rewriting of conjunctive queries using views: A progress report," in *Proc. 11th Int. Conf. Database Theory*, 2007, pp. 59–73.
[178] L. Segoufin and V. Vianu, "Views and queries: Determinacy and rewriting," in *Proc. 14th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, 2005, pp. 49–60.
[179] R. Tang, H. Wu, Z. Bao, S. Bressan, and P. Valduriez, "The price is right," in *Database and Expert Systems Applications*, H. Decker, L. Lhotská, S. Link, J. Basl, and A. M. Tjoa, Eds. Berlin, Germany: Springer, 2013, pp. 380–394.
[180] R. Tang, A. Amarilli, P. Senellart, and S. Bressan, "Get a sample for a discount," in *Database and Expert Systems Applications*, H. Decker, L. Lhotská, S. Link, M. Spies, and R. R. Wagner, Eds. Cham, Switzerland: Springer, 2014, pp. 20–34.
[181] C. Li and G. Miklau, "Pricing aggregate queries in a data marketplace," in *Proc. 15th Int. Workshop Web Databases*, 2012, pp. 19–24.
[182] A. Roth, "Technical perspective: Pricing information (and its implications)," *Commun. ACM*, vol. 60, no. 12, Nov. 2017, Art. no. 78.
[183] D. D. Freydenberger, "A Logic for Document Spanners," in *Proc. 20th Int. Conf. Database Theory, {ICDT}*2017, Venice, Italy, pp. 13:1–13:18, 2016. [Online]. Available: https://doi.org/10.4230/LIPIcs.ICDT.2017.13
[184] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *Proc. VLDB Endowment*, vol. 7, no. 9, pp. 757–768, May 2014.

[185] S. Deep and P. Koutris, "Qirana: A framework for scalable query pricing," in *Proc. ACM Int. Conf. Manage. Data*, 2017, pp. 699–713.
[186] S. Deep, P. Koutris, and Y. Bidasaria, "Qirana demonstration: Real time scalable query pricing," *Proc. VLDB Endowment*, vol. 10, no. 12, pp. 1949–1952, Aug. 2017.
[187] C. Xia and S. Muthukrishnan, "Arbitrage-free pricing in user-based markets," in *Proc. 17th Int. Conf. Auton. Agents Multi Agent Syst.*, 2018, pp. 327–335.
[188] L. Chen, P. Koutris, and A. Kumar, "Towards model-based pricing for machine learning in a data marketplace," in *Proc. Int. Conf. Manage. Data*, 2019, pp. 1535–1552.
[189] M.-F. Balcan and A. Blum, "Approximation algorithms and online mechanisms for item pricing," in *Proc. 7th ACM Conf. Electron. Commerce*, 2006, pp. 29–35.
[190] P. Briest and P. Krysta, "Single-minded unlimited supply pricing on sparse instances," in *Proc. 17th Annu. ACM-SIAM Symp. Discrete Algorithm*, 2006, pp. 1093–1102.
[191] V. Guruswami, J. D. Hartline, A. R. Karlin, D. Kempe, C. Kenyon, and F. McSherry, "On profit-maximizing envy-free pricing," in *Proc. 16th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2005, pp. 1164–1173.
[192] C. Swamy and M. Cheung, "Approximation algorithms for single-minded envy-free profit-maximization problems with limited supply," in *Proc. IEEE 49th Annu. Symp. Foundations Comput. Sci.*, 2008, pp. 35–44.
[193] X. Deng and C. H. Papadimitriou, "On the complexity of cooperative solution concepts," *Math. Operations Res.*, vol. 19, no. 2, pp. 257–266, 1994.
[194] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, and A. Rogers, "Bounding the estimation error of sampling-based shapley value approximation with/without stratifying," *CoRR*, vol. abs/1306.4265, 2013. [Online]. Available: https://arxiv.org/abs/1306.4265
[195] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.
[196] R. Jia *et al.*, "Towards efficient data valuation based on the shapley value," in *Proc. Mach. Learn. Res.*, 2019, pp. 1167–1176.
[197] Y. Zhou *et al.*, "Parallel feature selection inspired by group testing," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3554–3562.
[198] D.-Z. Du and F. K. Hwang, *Combinatorial Group Testing and Its Applications*, 2nd ed. Singapore: World Scientific, 1999.
[199] R. Jia *et al.*, "Efficient task-specific data valuation for nearest neighbor algorithms," *Proc. VLDB Endowment*, vol. 12, no. 11, pp. 1610–1623, Jul. 2019.
[200] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on P-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry*, 2004, pp. 253–262.
[201] R. Jia, X. Sun, J. Xu, C. Zhang, B. Li, and D. Song, "An empirical and comparative analysis of data valuation with scalable algorithms," *CoRR*, vol. abs/1911.07128, 2019. [Online]. Available: https://arxiv.org/abs/1911.07128
[202] D. Han, S. Tople, A. Rogers, M. Wooldridge, O. Ohrimenko, and S. Tschiatschek, "Replication-robust payoff-allocation with applications in machine learning marketplaces," *ArXiv*, vol. abs/2006.14583, 2020 [Online]. Available: https://arxiv.org/abs/2006.14583
[203] A. A. Armstrong and E. H. Durfee, "Mixing and memory: Emergent cooperation in an information marketplace," in *Proc. 3rd Int. Conf. Multi Agent Syst.*, 1998, Art. no. 34.
[204] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Germany: Springer, 2006, pp. 265–284.
[205] C. Niu, Z. Zheng, S. Tang, X. Gao, and F. Wu, "Making big money from small sensors: Trading time-series data under pufferfish privacy," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 568–576.
[206] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *Proc. 31st ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, 2012, pp. 77–88.
[207] E. Maskin and J. Riley, "Monopoly with incomplete information," *RAND J. Economics*, vol. 15, no. 2, pp. 171–196, 1984.
[208] A. Mas-Colell, M. Whinston, and J. Green, *Microeconomic Theory*. London, U.K.: Oxford Univ. Press, 1995.
[209] P. Naghizadeh and A. Sinha, "Adversarial contract design for private data commercialization," in *Proc. ACM Conf. Economics Comput.*, 2019, pp. 681–699.
[210] X.-B. Li and S. Raghunathan, "Pricing and disseminating customer data with privacy awareness," *Decis. Support Syst.*, vol. 59, pp. 63–73, 2014.
[211] J. Jaisingh, J. Barron, S. Mehta, and A. Chaturvedi, "Privacy and pricing personal information," *Eur. J. Oper. Res.*, vol. 187, no. 3, pp. 857–870, 2008.
[212] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," *J. Comput. Syst. Sci.*, vol. 60, no. 3, pp. 592–629, Jun. 2000.
[213] Z. Liu and H. Hacigümüş, "Online optimization and fair costing for dynamic data sharing in a cloud data market," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 1359–1370.
[214] P. Upadhyaya, M. Balazinska, and D. Suciu, "Price-optimal querying with data apis," *Proc. VLDB Endowment*, vol. 9, no. 14, pp. 1695–1706, Oct. 2016.
[215] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Trans. Inf. Theory*, vol. 22, no. 6, pp. 644–654, Nov. 1976.
[216] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "An online pricing mechanism for mobile crowdsensing data markets," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2017, Art. no. 26.
[217] C. Niu, Z. Zheng, F. Wu, S. Tang, and G. Chen, "Online pricing with reserve price constraint for personal data markets," *CoRR*, vol. abs/1911.12598, 2019. [Online]. Available: https://arxiv.org/abs/1911.12598
[218] B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," *Google AI Blog*, Apr. 2017. [Online]. Available: https://ai.googleblog.com/2017/04/federatedlearning-collaborative.html
[219] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[220] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. Song, "A principled approach to data valuation for federated learning," *ArXiv*, vol. abs/2009.06192, 2020 [Online]. Available: https://arxiv.org/abs/2009.06192
[221] R. H. L. Sim, Y. Zhang, M. C. Chan, and B. K. H. Low, "Collaborative machine learning with incentive-aware model rewards," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8927–8936.
[222] R. Hu and Y. Gong, "Trading data for learning: Incentive mechanism for on-device federated learning," *ArXiv*, vol. abs/2009.05604, 2020. [Online]. Available: https://arxiv.org/abs/2009.05604
[223] H. von Stackelberg, *Market Structure and Equilibrium*. Berlin, Germany: Springer, 1934.
[224] D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar, "Too much data: Prices and inefficiencies in data markets," National Bureau of Economic Research, Inc, NBER Working Papers 26296, Sep. 2019. [Online]. Available: http://www.nber.org/papers/w26296
[225] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, "Data market platforms: Trading data assets to solve data problems," *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 1933–1947, Jul. 2020.

Jian Pei (Fellow, IEEE) is currently working to facilitate efficient, fair, and sustainable usage of data and data analytics for social, economical and ecological good. Through inventing, implementing and deploying a series of data mining principles and methods, he produced remarkable values to academia and industry. His algorithms have been adopted by industry, open source toolkits and textbooks. His publications have been cited more than 100,000 times. He is also an active and productive volunteer for professional community services, such as chairing ACM SIGKDD, running many premier academic conferences in his areas, and being editor-in-chief or associate editor for the flagship journals in his fields. He is recognized as a fellow of the Royal Society of Canada (i.e., the National Academy of Canada), the Canadian Academy of Engineering, ACM and IEEE. He received a series of prestigious awards, such as the ACM SIGKDD Innovation Award, the ACM SIGKDD Service Award, and the IEEE ICDM Research Award.