Towards a Decentralized Infrastructure for Data Marketplaces: Narrowing the Gap between Academia and Industry

Lodovico Giaretta*
KTH Royal Institute of Technology
Stockholm, Sweden
lodovico@kth.se

Thomas Marchioro*
Evangelos Markatos
Foundation for Research and Technology Hellas
Heraklion, Greece
{marchiorot,markatos}@ics.forth.gr

Šarūnas Girdzijauskas
KTH Royal Institute of Technology
Stockholm, Sweden
sarunasg@kth.se

ABSTRACT

One big challenge for Industry 4.0 is leveraging the large amount of data that remain unused after collection. A variety of commercial data marketplaces have emerged in recent years to tackle this task. Despite their different business models and target markets, such marketplaces share a number of common issues that slow the growth of the industry, including data discovery, transparency, data privacy and data valuation. Many academic designs have been proposed to address these issues, yet most of them remain unimplemented, due to complexity or inefficiency.

We argue that these issues can be addressed with a combination of blockchain-based infrastructure, privacy-preserving computing and machine learning-based valuation metrics. Furthermore, we discuss key enabling technologies in each of these areas that are feasible to deploy at scale and could thus be implemented in real-world marketplaces in the near future. We select such technologies based on their current maturity and their industrial prominence.

CCS CONCEPTS

• **Security and privacy** → *Cryptography*; *Access control*; **Privacy-preserving protocols**; *Distributed systems security*; **Privacy protections**; • **Information systems** → **Information integration**; • **Applied computing** → **Electronic data interchange**; • **Networks** → **Network privacy and anonymity**; • **Computing methodologies** → **Distributed computing methodologies**; *Machine learning*; • **Computer systems organization** → **Distributed architectures**;

ACM Reference Format:
Lodovico Giaretta, Thomas Marchioro, Evangelos Markatos, and Šarūnas Girdzijauskas. 2022. Towards a Decentralized Infrastructure for Data Marketplaces: Narrowing the Gap between Academia and Industry. In *Data Economy (DE '22), December 9, 2022, Roma, Italy.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3565011.3569060

*Both authors contributed equally to this work.

DE '22, December 9, 2022, Roma, Italy

ACM ISBN 978-1-4503-9923-4/22/12.
https://doi.org/10.1145/3565011.3569060

1 INTRODUCTION

Companies and organizations of all kinds are collecting more data each year. Yet, over 68% of these data remains unused, as shown by a Seagate report in 2020 [63]. At the same time, other organizations struggle to find enough high-quality data to develop data-driven applications based on machine learning and artificial intelligence. A new *data economy* is thus growing, centered around the concept of *data marketplaces*, where companies can trade data and unlock previously unimaginable business opportunities.

A multitude of commercial data marketplaces have emerged, encompassing different domains and business models. However, these implementations share a number of weaknesses, which have been extensively addressed by the academic community with several marketplace designs. Nevertheless, most academic solutions have not taken hold in the commercial space, due to their complexity or inefficiency, leading to a widening gap between academia and industry.

In this work, we analyze the following key issues: data discovery, transparency, data privacy and data valuation. For each of them, we indicate high-level guidelines and examine enabling technologies, with a focus on practical feasibility. Finally, we discuss which among these technologies are more likely to bridge the gap between academic and industrial designs in the near future.

On the infrastructure side, we note that the current fragmentation of existing marketplaces hampers **data discovery** efforts. A unified infrastructure is required to realize the data market potential. However, the creation of monopolies as result of a consolidation process would harm openness and **transparency**, key requirements for a thriving market. We identify blockchain technology as a promising solution to prevent such an outcome and discuss the benefits and drawbacks of permissionless and permissioned approaches. Finally, we argue that a permissioned design is the most likely to be accepted by the industry in the near term.

In regards to **data privacy**, we argue that the direct sale of *personal* data is incompatible with high privacy standards, and legally challenging given current and future regulations. We therefore suggest a shift towards designs that allow consumers to submit data processing jobs to the marketplace and receive the final outputs without obtaining direct access to the data. We compare different enabling technologies, such as homomorphic encryption, secure multiparty computation, trusted execution environments, differential privacy and federated learning. Finally, we suggest that a combination of trusted execution environments, differential privacy

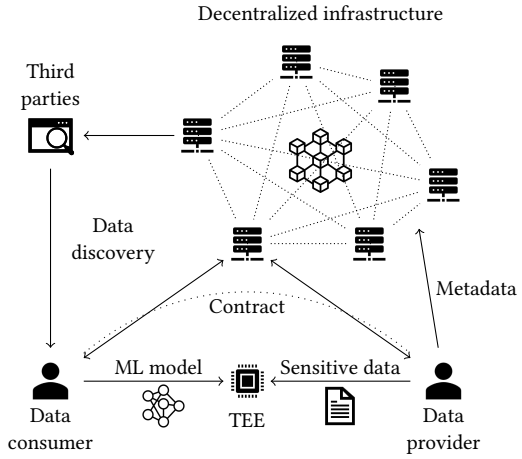

Figure 1: Vision for a decentralized, transparent, privacy-preserving marketplace.

and federated learning provides strong privacy guarantees while at the same time being feasible and scalable.

On the **data valuation** side, we note the challenge, for both providers and consumers, to put a price tag on a dataset, as its value depends on how it is used and combined with other data sources. We therefore argue that data prices should be agreed upon on a case-by-case basis, based on the actual value generated. We analyze different techniques to estimate this value and apportion it across multiple dataset. Finally, we argue that a combination of model-independent and model-specific metrics can be employed to provide accurate valuations while still being computationally feasible.

We believe that our selected techniques are sufficiently mature for practical applications and can be easily combined. They can thus be employed to build a decentralized, transparent and privacy-preserving marketplace, as shown in fig. 1, which could feasibly bring some of the recent academic advancements to the real data economy.

2 OPENNESS AND TRANSPARENCY

Data can present themselves in different forms: continuous streams or fixed datasets, raw or curated, bulk-downloadable or accessible via APIs. Data from different domains may have different characteristics and requirements. Additionally, as the data marketplaces industry is still in its early steps, companies are experimenting with a variety of business models, interfaces and domains. All of this leads to substantial fragmentation in the landscape of data marketplaces, which limits the benefits for both data providers and consumers. Data consumers lack a simple way to identify all potentially useful datasets, as they are scattered across a large number of isolated marketplaces, some of which they may not know. And when useful datasets are identified across multiple marketplaces, dealing with the different business models and access patterns may represent a significant barrier, especially for smaller consumers. These barriers also affect providers, who may not be able to reach all their potential customers without a substantial investment to support multiple marketplaces concurrently.

Permissionless	Permissioned
Open membership with self-enrollment	Closed membership, approval required
Anonymous transactions	Verifiable identities
Limited throughput, somewhat high latency *	High throughput and low latency out of the box
High computational and energy costs *	Relatively efficient and cost effective

Table 1: Comparison of permissionless and permissioned blockchains. * = advances in permissionless blockchains may deliver improvements (see section 2.1)

As the data marketplaces industry matures, a consolidation process inevitably takes hold, eventually resulting in a monopoly or an oligopoly, where few large entities control a majority of the market. This centralization solves the challenges caused by fragmentation, but also exacerbates a different class of issues, centered around the concept of *transparency*. In this context, transparency is the ability for all stakeholders in the marketplace to have full knowledge of all providers, consumers and transactions, and to be able to audit all processes taking place. When the entire marketplace is controlled by a single organization, this is typically not the case. A lack of transparency leads to issues of *fairness*. For example, it is not possible to verify whether certain sellers or buyers are being given preferential treatment, or whether information about certain datasets is being withheld to influence the market dynamics.

2.1 Blockchain-Based Infrastructure

To prevent the issues above, a marketplace infrastructure must provide all consumers with full access to all providers, while ensuring full transparency and accountability. Blockchain technology [9] has emerged as the most promising tool to achieve these goals. As all datasets and transactions are recorded on a decentralized, tamper-proof ledger, buyers and sellers are guaranteed access to all relevant information when choosing datasets to buy and when negotiating prices, thus ensuring a fair environment. Furthermore, the public availability of this information empowers third parties to build services on top of the marketplace, such as data discovery or brokerage, boosting the ecosystem. Finally, *smart contracts* allow different entities to set out and enforce unambiguous requirements and processes to be followed in each transaction, enabling trustless interactions.

Many different blockchains have been proposed over the years, each presenting different benefits and drawbacks. It is therefore necessary to examine the following question:

Q1. What type of blockchain can *feasibly* provide openness and transparency in industrial marketplaces?

Permissionless Blockchains. In permissionless blockchains, anyone can participate in the ledgering activities in a fully anonymous fashion. A large number of marketplace designs have been proposed based on such blockchains, mostly focused on IoT or personal data gathered by a large number of sources [29, 31, 41, 60, 64, 68, 78],

as the lack of identification requirements allows frictionless self-enrollment of users at large scale . The most commonly-used permissionless blockchain is Ethereum [10], due to its stability, powerful smart contracts and rich ecosystem. However, Ethereum does not support confidentiality, with any data handled by smart contracts required to be public. If a marketplace requires the smart contracts to handle private information, a different blockchain must be used. That is why Sterling [35] employs Oasis, a variant of the Ekiden blockchain, which provides confidential smart-contracts [13, 57]. One of the main limitation of permissionless blockchains is that their trust-less architecture requires complex protocols to prevent anonymous attackers from corrupting the ledger. This often leads to low throughput and high latency in the processing of transactions, limiting the scalability of any marketplaces design [17]. Furthermore, these protocols consume huge amounts of energy to operate and produce high levels of carbon emissions [42, 48].

Attempts are being made to tackle these issues. By allowing independent transactions to be published concurrently, and by using lighter cryptographic protocols, IOTA [65] promises high throughput, low latency, and energy efficiency. Furthermore, Ethereum is undergoing a slow transition [22] that should increase its throughput and reduce its computational requirements. However, it will take time before these approaches can achieve widespread adoptions and their real-world advantages have not been proven yet [30].

Permissioned Blockchains. On the other hand, permissioned blockchains require all entities to be authorized before they can participate in the protocol. In exchange, verifiable identities are associated to each activity, increasing accountability. The improved trust leads to lighter ledgering protocols that provide better scalability. Thus, permissioned blockchains are more often found in business-centric architectures, where the marketplace is designed for a relatively small number of large data-handling organizations. In this context, the additional friction that comes with the approval of each new participant is acceptable. One of the most well-known permissioned blockchains is Hyperledger Fabric [5], on top of which several data marketplaces have been proposed [18, 70]. In addition to the aforementioned benefits, Fabric also provides a high degree of customizability and can thus be easily tailored to the requirements of specific applications.

Market Directions. Based on the analysis above, we argue that permissioned blockchains are the most likely to be deployed in a real-world setting in the near term, thanks to their many advantages. In particular, verifiable identities may increase trust in the system and deter (or help to identify) potential high-level manipulations. This direction towards permissioned blockchains is also backed by previous studies in various commercial domains [25, 47, 55]. However, a permissioned approach also brings downsides that may grow more severe as the market matures. First, the high bar to participate in the marketplace means that single persons or small groups may not be able to individually sell their data, thus requiring brokers to aggregate and sell data on their behalf. This can add value to the data (in the form of aggregation and curation), simplify bulk purchase by consumers and increase the chances of monetization. But it can also reduce the users' control over their data. Second, if a set of established large companies dominates the system, they may be unwilling to let new players in, as these may undermine the incumbents' position. Thus, in the absence of strong regulatory supervision, a closed oligopoly might form. The third issue is that when real, verifiable identities are associated to each transactions, it is possible for an organization to see which datasets have been purchased by its competitors, thus leaking potentially sensitive business directions.

In the longer term, a shift towards more open, permissionless blockchains may be possible, especially if ongoing attempts towards higher throughput, lower latency and better efficiency bear fruit and reach stability and wide adoption.

2.2 Data Discovery and Standardization

As already mentioned, data discovery is one of the most important issues that affect the data marketplaces landscape. Effective data discovery solutions are necessary to maximize the utility of data marketplaces for both consumers and providers, and thus support the growth of the sector [59]. However, it is also one of the most challenging issues, and one that is not often addressed in marketplace architectures proposed by academia.

The use of blockchain as a fully transparent ledger where all datasets are recorded can help tackling this issue. Data discovery can be decoupled from the underlying infrastructure, enabling third parties to provide it as a specialized service, by scraping and indexing all datasets metadata available on the blockchain. Recent works have even shown the possibility of building a decentralized keyword index to enable efficient multi-keyword lookups [77].

However, automated semantic reasoning and query systems will be necessary to build a truly large-scale ecosystem. Substantial research has been devoted to the use of *ontologies* [33] to build semantic, machine-friendly resource indices, in particular regarding large-scale IoT data sources [43, 44]. These would require the entire industry to collaborate and build a hierarchy of interoperable ontologies [23] to unambiguously describe all datasets. Additionally, the means of formatting and storing both data and metadata would have to be standardized.

3 PERSONAL DATA AND PRIVACY

Another key issue that commercial data marketplaces need to address is personal data management. In the European Union and United Kingdom, personal data are subject to the General Data Protection Regulation (GDPR) [1]. GDPR guarantees to individuals the "right to be forgotten", meaning that one may request to erase their personal information. Even in the event that a person explicitly consents to sell her or his data, this permission can be revoked at any point in time.

Thus, buying and selling personal data is not practical in countries that are subject to GDPR. Furthermore, the increase of privacy awareness is likely to impede personal data sales also outside Europe [15]. However, personal information is arguably the most valuable in many applications such as medical research, financial studies, and recommendation systems. Therefore, just forfeiting the possibility to leverage such information would severely hinder the economy of data marketplaces.

Anonymization and de-identification. Issues related to personal data are often addressed via *anonymization* or *de-identification*.

GDPR defines anonymized information as "information which does not relate to an identified or identifiable natural person" [1]. Similar notions data are expressed also in other laws, e.g., the California Consumer Privacy Act (CCPA) [2]. While the laws are rigorous in requiring resilience to any kind of linkage, popular anonymization techniques are more lax. The best established, k-anonymity [67], relies on the assumption that the dataset curator is able to determine a subset of attributes that may contribute to de-anonymization. While there are cases where this assumption applies, often identifying information is entangled with useful data [69]. It has been shown, in fact, that individuals can be identified by unexpected information such as a list of movies they have seen [53], connections on social media [54], or daily activity [37].

A natural conclusion is that data should not be accessed at all, at least at a low aggregation level. In the remainder of this section, we will address the following question:

Q2. Is it possible to extract value from personal data without seeing them?

While several techniques have been proposed to process private data "blindly", our discussion focuses on those which can realistically be employed by data consumers. Our discussion will mostly focus on training machine learning models, which is the most prominent use case for said techniques. More specifically, we review blinding methods that have been thoroughly studied in recent years, namely *homomorphic encryption* (HE) [76], *secure multiparty computation* (SMC) [45], and *trusted execution environments* (TEE) [36]. We evaluate existing solutions based on the following criteria:

Q2.1 *Do they prevent privacy leaks from the dataset?* Any solution that requires to release personal data may expose private information.
Q2.2 *Do they keep the model private?* The specifics of a model and the processing details are intellectual property of the data consumer, and as such must be protected.
Q2.3 *Do they prevent privacy leaks from the model?* Models may retain personal information during training [12].
Q2.4 *Are they computationally efficient?* As ML training often demands many matrix operations, it is preferable to run models on GPUs and minimize the overhead.
Q2.5 *Are they scalable for a large number of data sources?* Data will often come from more than one source, requiring to handle interactions with multiple providers.
Q2.6 *Are they platform-independent?* Methods that rely on specific hardware or software might have limited compatibility, and may constitute a cost for organizations.

Homomorphic encryption. Homomorphic encryption schemes allow to perform some operations on encrypted data that get reflected on the clear data after decryption. Albeit powerful in many applications [4], in the field of machine learning HE is hampered by two drawbacks. The first is its computational overhead, due in part to encryption and decryption, but mainly to the more complex steps required to process encrypted data. The second drawback is that only linear and piece-wise linear operations can be performed exactly. This implies that some non-linear functions used in ML (e.g., sigmoid, softmax), can only be approximated, and ordinal operations (e.g., max, median) cannot be computed [46, 52]. Overall, these shortcomings do not appear to excessively impact the model accuracy. However, training deep models requires significantly more time compared to other solutions [32].

Secure multiparty computation. SMC is a broad class of cryptographic protocols, which can be based on homomorphic encryption [16], garbled circuits [34], secret sharing, or a combination of them [38]. SMC solutions are often distinguished from techniques purely based on homomorphic encryption, as the former usually require many exchanges between the data storage and the model [71]. The implication is that training must happen through online communications between the data provider and the consumer. Under low latency, SMC offers an efficient alternative to homomorphic encryption [32]. However, issues may arise when multiple providers are involved in the training, as all the parties need to participate in online updates at the same time.

Trusted Execution Environments. TEE are isolated portions of hardware, located inside the CPU, that allow to run trusted applications without the possibility of interference by any external entity, even the machine's OS. Although many TEE have emerged, the most successful are undoubtedly the enclaves provided by Intel SGX [14], which comes embedded in most Intel processors. In order to train a machine learning model on personal data within SGX, a secure channel must be established between the enclave and the data providers [58], so that the data can be sent to the secure portion of memory. Once inside the enclave, data can be processed in clear, and the training can happen as in standard non-private applications. A main drawback of utilizing TEE is that they are hardware-specific, and tools need to be built on top of their drivers. Nevertheless, this is the case also for many non-private ML applications (e.g., relying on CUDA drivers to perform operations on the GPU). With recent attempts of making confidential code execution compatible with GPU [21, 51, 56], TEE are probably the most competitive candidate solution for private machine learning.

3.1 Differential Privacy and Federated Learning

Above-mentioned techniques may be used alone or combined with other privacy-enhancing paradigms such as *differential privacy* (DP) [20] and *federated learning* (FL) [74].

Differential privacy. DP is a property that can be enforced on aggregation algorithms through randomization [19] to prevent sensitive inferences. In the case of neural networks, DP can prevent unintended memorization of private data in the final model. DP is achieved by "randomizing" the gradient with additive Gaussian noise [3] during training. The magnitude of the noise depends on a parameters called *privacy budget*. Small values of the privacy budget provide stronger guarantees against memorization, but also require to inject more noise. Experimental results have shown that even small amounts of noise, despite not achieving strong theoretical guarantees, yields high resilience to privacy leaks [49]. DP can be easily made compatible with other privacy preserving solutions based on HE, SMC and TEE.

Federated learning. FL is a learning paradigm that allows to train machine learning models in a distributed fashion across multiple data sources. Data providers compute model updates based on their

	ML	HE	SMC	TEE
Prevents privacy leaks from the dataset	○	●	●	●
Protects intellectual property on the model	○	◐	◐	●
Prevents privacy leaks from the model	○	○	○	○
Preserves model accuracy	●	●	●	●
Is computationally efficient	●	○	◐	●
Is scalable for a large number of data sources	◐	◐	○	◐
Is hardware-independent	●	●	●	○

	ML+DP	HE+DP	SMC+DP	TEE+DP
Prevents privacy leaks from the model	●	●	●	●
Preserves model accuracy	◐	◐	◐	◐

	FL	HE+FL	SMC+FL	TEE+FL
Is scalable for a large number of data sources	●	●	◐	●
Preserves model accuracy	◐	◐	◐	◐

Table 2: The top table compares different solutions for machine learning on private data sources. ML = machine learning on clear data, HE = homomorphic encryption, SMC = secure multiparty computation, TEE = trusted execution environments. The bottom two tables highlight only the differences when the solutions are combined differential privacy (DP) and federated learning (FL).

local data, and exchange such updates with a central aggregator[1] that merges them [40]. The principal advantage of FL is that data consumers do not need to see the data and gather them in a central storage. Nonetheless, although the model updates are arguably less sensitive than clear data, they still might leak private information [26]. Recent works [66] have adopted HE and SMC to prevent data consumers from observing them. TEE can also be employed to prevent access to the gradients, with the additional advantage of protecting model architectures, and thus the data consumers' intellectual property [50].

4 DATA VALUATION

Current industrial marketplaces adopt different business models and pricing strategies, including fixed prices, volume-based payments, revenue sharing, and bids [8]. Dataset prices are quite variable, depending on many aspects such as domain, volume and update frequency [6]. However, these properties are not sufficient to determine the value of a dataset, which is a challenging question that neither providers nor consumers are typically able to answer with confidence [24]. Solutions for data valuation are thus necessary for vendors to set prices and for consumers to decide whether they are willing to make a purchase.

For consumers, that value is determined by the insights that they can extract from the data. This, in turn, depends not only on the features of the dataset itself, but also on the way it is processed and on other data sources that are combined with it. Therefore, it can be simpler to assign a value to the outputs of a data processing flow, as those outputs are then used by the consumers to produce value for their business, and thus key performance indicators (KPIs) are typically available to evaluate them [11]. Focusing on marketplaces for ML applications, consumers can assign value to accuracy, error rate or other metrics that they are able to easily interpret from a business perspective.

Two main approaches exist when acquiring data for ML training. Either the datasets are purchased in advance, or a contract is signed with the data providers that allows the payment to happen after training, possibly based on the KPIs achieved by the trained model. In the former case, the key question is how a buyer can estimate how using a dataset will affect KPIs. In the latter case, the core problem is how to apportion the total payment to different datasets, based on their contribution in achieving the final result. This leads to the following research question:

Q3. Which methods can be practically used to valuate individual datasets in a ML-based marketplace?

Predicting dataset value. Volume-based valuation [73] assesses a dataset value based on the volume of "diverse" information contained in it. The underlying assumption of this approach is that valuable information is encoded in the variability among the data points, and that an excessively homogeneous dataset provides no useful insights. This solution may not be accurate in cases where similar entries actually provide value to a model, but can be used to give a prior estimate that is model-independent.

A better tailored value assessment can be obtained by taking into account the task at hand. The "try before you buy" approach [7] evaluates the target model on metrics of interest once per each dataset individually. Based on these metrics, a subset of datasets is selected to maximizes the total value while minimizing the cost. This can be seen as an inexpensive approximation of the Shapley value, presented below.

Assessing dataset contribution. Shapley value [61] measures the relative contribution of different data sources to the total value of a metric. The Shapley value of a specific dataset is defined as the average marginal improvement provided by its addition to each of the possible combinations of other datasets. However, evaluating the Shapley value of n candidate datasets requires repeating the training process $O(2^n)$ times. Many works devised faster approximations [27, 62], e.g., via Monte Carlo simulations or gradient-based methods. Albeit less computationally demanding, such approximations still require to train the target ML model multiple times on different dataset combinations. At the end of the process, a single model is obtained by the consumer, wasting any resources spent on

[1]There are alternatives, such as gossip learning [28], where the data providers exchange updates with each other, but for the sake of discussion we use "federated learning" as an umbrella term to include them.

training the others, making solutions based on the Shapley value mostly impractical.

Relative usefulness may also be gauged *online* during a single training process. A recent study [75] suggests using trainable estimators to assess the potential value of each datapoint. The estimator is trained online, using reinforcement learning, based on the real-time improvements in the target loss function. However, the approach is application-specific, and a different estimator needs to be trained for each task.

A simpler yet effective solution consists in using influence functions [39]. This approach requires to train an initial model on all candidate datasets. Taylor approximations are then used to estimate the performance loss caused by the removal of individual datasets. The final model can then be re-trained using only the best performing datasets.

Recent works have shown that even more efficient and practical solutions are possible, completely forgoing the need to train any additional component or model. DAVINZ [72] allows to compute domain-specific generalization bounds on the error rate of a model without requiring any training, by exploiting the mathematical behaviour of untrained models. Experiments on known classification models have shown that such bounds allow to give reasonable estimates of the marginal contributions of individual datasets.

Practical Directions. Overall, despite substantial research efforts, approaches that directly mimic or approximate Shapley values are unlikely to be feasible on a large-scale in the near term. Instead, we argue for the use of a combination of indirect measures. These should include both task-agnostic and task specific metrics. The former, such as volume-based valuation, can be computed a priori for each dataset and used for a first filtering of potential sources. The latter, such as DAVINZ, can then be used to fine-tune the dataset selection.

5 CONCLUSIONS

In this work, we analyzed the key issues affecting industrial data marketplaces: data discovery, transparency, data privacy and data valuation. We discussed how a combination of blockchain-based infrastructure, privacy-preserving machine learning and data valuation techniques can be employed to overcome these issues. Furthermore, we identified key technologies in each of these areas that are ready for real-world applications and can therefore be implemented across existing industrial marketplaces in the near future. This will hopefully help bridge the gap between academia and industry and foster further growth in the sector.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813162. The content of this paper reflects the views only of their author (s). The European Commission/ Research Executive Agency are not responsible for any use that may be made of the information it contains.

REFERENCES

[1] 2016. General Data Protection Regulation. https://gdpr-info.eu. Accessed: 2022-09-21.
[2] 2018. California Consumer Privacy Act. https://ccpa-info.com/home/1798-140-definitions. Accessed: 2022-09-21.
[3] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.
[4] Mohamed Alloghani, Mohammed M Alani, Dhiya Al-Jumeily, Thar Baker, Jamila Mustafina, Abir Hussain, and Ahmed J Aljaaf. 2019. A systematic review on the status and progress of homomorphic encryption technologies. *Journal of Information Security and Applications* 48 (2019), 102362.
[5] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, et al. 2018. Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth EuroSys conference*. 1–15.
[6] Santiago Andrés Azcoitia, Costas Iordanu, and Nikolaos Laoutaris. 2021. What is the price of data? A measurement study of commercial data marketplaces. *arXiv preprint arXiv:2111.04427* (2021).
[7] Santiago Andrés Azcoitia and Nikolaos Laoutaris. 2020. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. *arXiv preprint arXiv:2012.08874* (2020).
[8] Santiago Andrés Azcoitia and Nikolaos Laoutaris. 2022. A survey of data marketplaces and their business models. *arXiv preprint arXiv:2201.04561* (2022).
[9] Marianna Belotti, Nikola Božić, Guy Pujolle, and Stefano Secci. 2019. A Vademecum on Blockchain Technologies: When, Which, and How. *IEEE Communications Surveys & Tutorials* 21, 4 (2019), 3796–3838. https://doi.org/10.1109/COMST.2019.2928178
[10] Vitalik Buterin et al. 2014. A next-generation smart contract and decentralized application platform. *white paper* 3, 37 (2014), 2–1.
[11] Jian Cai, Xiangdong Liu, Zhihui Xiao, and Jin Liu. 2009. Improving supply chain performance management: A systematic approach to analyzing iterative KPI accomplishment. *Decision support systems* 46, 2 (2009), 512–521.
[12] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*. 267–284.
[13] Raymond Cheng, Fan Zhang, Jernej Kos, Warren He, Nicholas Hynes, Noah Johnson, Ari Juels, Andrew Miller, and Dawn Song. 2019. Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contracts. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 185–200.
[14] Victor Costan and Srinivas Devadas. 2016. Intel SGX explained. *Cryptology ePrint Archive* (2016).
[15] Munther Dahleh. 2018. Why the Data Marketplaces of the Future Will Sell Insights, Not Data. https://sloanreview.mit.edu/article/why-the-data-marketplaces-of-the-future-will-sell-insights-not-data/. Accessed: 2022-09-21.
[16] Ivan Damgård, Valerio Pastro, Nigel Smart, and Sarah Zakarias. 2012. Multiparty computation from somewhat homomorphic encryption. In *Annual Cryptology Conference*. Springer, 643–662.
[17] Richard Dennis and Jules Pagna Disso. 2019. An Analysis into the Scalability of Bitcoin and Ethereum. In *Third International Congress on Information and Communication Technology*, Xin-She Yang, Simon Sherratt, Nilanjan Dey, and Amit Joshi (Eds.). Springer Singapore, Singapore, 619–627.
[18] Akanksha Dixit, Arjun Singh, Yogachandran Rahulamathavan, and Muttukrishnan Rajarajan. 2021. FAST DATA: A Fair, Secure and Trusted Decentralized IIoT Data Marketplace enabled by Blockchain. *IEEE Internet of Things Journal* (2021), 1–1. https://doi.org/10.1109/JIOT.2021.3120640
[19] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.
[20] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
[21] Anne C Elster and Tor A Haugdahl. 2022. Nvidia Hopper GPU and Grace CPU Highlights. *Computing in Science & Engineering* 24, 2 (2022), 95–100.
[22] Ethereum Foundation. 2022. *Ethereum Vision.* Retrieved 2022-09-22 from https://ethereum.org/en/upgrades/vision/
[23] Jérôme Euzenat, Pavel Shvaiko, et al. 2007. *Ontology matching*. Vol. 18. Springer.
[24] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. 2020. Data market platforms: Trading data assets to solve data problems. *arXiv preprint arXiv:2002.01047* (2020).
[25] Rosa M Garcia-Teruel. 2020. Legal challenges and opportunities of blockchain technology in the real estate sector. *Journal of Property, Planning and Environmental Law* (2020).
[26] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems* 33 (2020), 16937–16947.
[27] Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*. PMLR, 2242–2251.

[28] Lodovico Giaretta and Šarūnas Girdzijauskas. 2019. Gossip learning: Off the beaten path. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 1117–1124.

[29] Lodovico Giaretta, Ioannis Savvidis, Thomas Marchioro, Šarūnas Girdzijauskas, George Pallis, Marios D Dikaiakos, and Evangelos Markatos. 2021. PDS 2: A user-centered decentralized marketplace for privacy preserving data processing. In *2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, 92–99.

[30] Fengyang Guo, Xun Xiao, Artur Hecker, and Schahram Dustdar. 2020. Characterizing IOTA Tangle with Empirical Data. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. 1–6. https://doi.org/10.1109/GLOBECOM42002.2020.9322220

[31] Pooja Gupta, Volkan Dedeoglu, Salil S. Kanhere, and Raja Jurdak. 2021. Towards a blockchain powered IoT data marketplace. In *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. 366–368. https://doi.org/10.1109/COMSNETS51098.2021.9352865

[32] Veneta Haralampieva, Daniel Rueckert, and Jonathan Passerat-Palmbach. 2020. A systematic comparison of encrypted machine learning solutions for image classification. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*. 55–59.

[33] Ian Horrocks. 2008. Ontologies and the semantic web. *Commun. ACM* 51, 12 (2008), 58–67.

[34] Yan Huang, David Evans, Jonathan Katz, and Lior Malka. 2011. Faster Secure {Two-Party} Computation Using Garbled Circuits. In *20th USENIX Security Symposium (USENIX Security 11)*.

[35] Nick Hynes, David Dao, David Yan, Raymond Cheng, and Dawn Song. 2018. A demonstration of sterling: a privacy-preserving data marketplace. *Proceedings of the VLDB Endowment* 11, 12 (2018), 2086–2089.

[36] Patrick Jauernig, Ahmad-Reza Sadeghi, and Emmanuel Stapf. 2020. Trusted execution environments: properties, applications, and challenges. *IEEE Security & Privacy* 18, 2 (2020), 56–60.

[37] Andrei Kazlouski, Thomas Marchioro, and Evangelos P. Markatos. 2022. What your Fitbit Says about You: De-anonymizing Users in Lifelogging Datasets. In *SECRYPT*.

[38] Marcel Keller. 2020. MP-SPDZ: A versatile framework for multi-party computation. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 1575–1590.

[39] Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. 2019. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems* 32 (2019).

[40] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527* (2016).

[41] Vlasis Koutsos, Dimitrios Papadopoulos, Dimitris Chatzopoulos, Sasu Tarkoma, and Pan Hui. 2021. Agora: A Privacy-Aware Data Marketplace. *IEEE Transactions on Dependable and Secure Computing* (2021), 1–1. https://doi.org/10.1109/TDSC.2021.3105099

[42] Max J Krause and Thabet Tolaymat. 2018. Quantification of energy and carbon costs for mining cryptocurrencies. *Nature Sustainability* 1, 11 (2018), 711–718.

[43] Danh Le-Phuoc, Hoan Nguyen Mau Quoc, Josiane Xavier Parreira, and Manfred Hauswirth. 2011. The linked sensor middleware–connecting the real world and the semantic web. *Proceedings of the Semantic Web Challenge* 152 (2011), 22–23.

[44] Danh Le-Phuoc, Hoan Nguyen Mau Quoc, Hung Ngo Quoc, Tuan Tran Nhat, and Manfred Hauswirth. 2016. The graph of things: A step towards the live knowledge graph of connected things. *Journal of Web Semantics* 37 (2016), 25–35.

[45] Yehuda Lindell. 2020. Secure multiparty computation. *Commun. ACM* 64, 1 (2020), 86–96.

[46] Jian Liu, Mika Juuti, Yao Lu, and Nadarajah Asokan. 2017. Oblivious neural network predictions via minionn transformations. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 619–631.

[47] Sin Kuang Lo, Yue Liu, Su Yen Chia, Xiwei Xu, Qinghua Lu, Liming Zhu, and Huansheng Ning. 2019. Analysis of blockchain solutions for IoT: A systematic literature review. *IEEE Access* 7 (2019), 58822–58835.

[48] Kyle McDonald. 2021. Ethereum Emissions: A Bottom-up Estimate. *arXiv preprint arXiv:2112.01238* (2021).

[49] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963* (2017).

[50] Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. 2021. PPFL: privacy-preserving federated learning with trusted execution environments. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 94–108.

[51] Fan Mo, Zahra Tarkhani, and Hamed Haddadi. 2022. SoK: Machine Learning with Confidential Computing. *arXiv preprint arXiv:2208.10134* (2022).

[52] Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 19–38.

[53] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.

[54] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*. IEEE, 173–187.

[55] Nawari O Nawari and Shriraam Ravindran. 2019. Blockchain and the built environment: Potentials and limitations. *Journal of Building Engineering* 25 (2019), 100832.

[56] Lucien KL Ng, Sherman SM Chow, Anna PY Woo, Donald PH Wong, and Yongjun Zhao. 2021. Goten: Gpu-outsourcing trusted execution of neural network training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14876–14883.

[57] Oasis Protocol Project. 2020. *The Oasis Blockchain Platform*. Technical Report.

[58] Olga Ohrimenko, Felix Schuster, Cédric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. 2016. Oblivious {Multi-Party} machine learning on trusted processors. In *25th USENIX Security Symposium (USENIX Security 16)*. 619–636.

[59] Evangelos Psomakelis, Anastasios Nikolakopoulos, Achilleas Marinakis, Alexandros Psychas, Vrettos Moulos, Theodora Varvarigou, and Andreas Christou. 2020. A scalable and semantic data as a service marketplace for enhancing cloud-based applications. *Future Internet* 12, 5 (2020), 77.

[60] Gowri Sankar Ramachandran, Rahul Radhakrishnan, and Bhaskar Krishnamachari. 2018. Towards a Decentralized Data Marketplace for Smart Cities. In *2018 IEEE International Smart Cities Conference (ISC2)*. 1–8. https://doi.org/10.1109/ISC2.2018.8656952

[61] Alvin E Roth. 1988. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

[62] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The Shapley Value in Machine Learning. *arXiv preprint arXiv:2202.05594* (2022).

[63] Seagate. 2020. Rethink Data. https://www.seagate.com/files/www-content/our-story/rethink-data/files/Rethink_Data_Report_2020.pdf.

[64] Nicolás Serrano and Fredy Cuenca. 2021. A Peer-to-Peer Ownership-Preserving Data Marketplace. In *2021 IEEE International Conference on Blockchain (Blockchain)*. 394–400. https://doi.org/10.1109/Blockchain53845.2021.00062

[65] Wellington Fernandes Silvano and Roderval Marcelino. 2020. Iota Tangle: A cryptocurrency to communicate Internet-of-Things data. *Future Generation Computer Systems* 112 (2020), 307–319. https://doi.org/10.1016/j.future.2020.05.047

[66] Oana Stan, Vincent Thouvenot, Aymen Boudguiga, Katarzyna Kapusta, Martin Zuber, and Renaud Sirdey. 2022. A Secure Federated Learning: Analysis of Different Cryptographic Tools. In *SECRYPT*.

[67] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems* 10, 05 (2002), 557–570.

[68] Matias Travizano, Carlos Sarraute, Mateusz Dolata, Aaron M. French, and Horst Treiblmaier. 2020. *Wibson: A Case Study of a Decentralized, Privacy-Preserving Data Marketplace*. Springer International Publishing, Cham, 149–170. https://doi.org/10.1007/978-3-030-44337-5_8

[69] Ardhendu Tripathy, Ye Wang, and Prakash Ishwar. 2019. Privacy-preserving adversarial networks. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 495–505.

[70] Hien Thi Thu Truong, Miguel Almeida, Ghassan Karame, and Claudio Soriente. 2019. Towards Secure and Decentralized Sharing of IoT Data. In *2019 IEEE International Conference on Blockchain (Blockchain)*. 176–183. https://doi.org/10.1109/Blockchain.2019.00031

[71] Sameer Wagh, Shruti Tople, Fabrice Benhamouda, Eyal Kushilevitz, Prateek Mittal, and Tal Rabin. 2020. Falcon: Honest-majority maliciously secure framework for private deep learning. *arXiv preprint arXiv:2004.02229* (2020).

[72] Zhaoxuan Wu, Yao Shu, and Bryan Kian Hsiang Low. 2022. DAVINZ: Data Valuation using Deep Neural Networks at Initialization. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 24150–24176. https://proceedings.mlr.press/v162/wu22j.html

[73] Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. 2021. Validation free and replication robust volume-based data valuation. *Advances in Neural Information Processing Systems* 34 (2021), 10837–10848.

[74] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.

[75] Jinsung Yoon, Sercan Arik, and Tomas Pfister. 2020. Data Valuation using Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 10842–10851. https://proceedings.mlr.press/v119/yoon20a.html

[76] Hana Yousuf, Michael Lahzi, Said A Salloum, and Khaled Shaalan. 2021. Systematic review on fully homomorphic encryption scheme and its application. *Recent*

Advances in Intelligent Systems and Smart Applications (2021), 537–551.

[77] Mirko Zichichi, Luca Serena, Stefano Ferretti, and Gabriele D'Angelo. 2021. Towards Decentralized Complex Queries over Distributed Ledgers: a Data Marketplace Use-case. In *2021 International Conference on Computer Communications and Networks (ICCCN)*. 1–6. https://doi.org/10.1109/ICCCN52240.2021.9522165

[78] Kazim Rifat Özyilmaz, Mehmet Doğan, and Arda Yurdakul. 2018. IDMoB: IoT Data Marketplace on Blockchain. In *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)*. 11–19. https://doi.org/10.1109/CVCBT.2018.00007