Data Assets: Tokenization and Valuation

Responsible Use, Valuation and Monetization

Hirsh Pithadia
hirsh@valyu.network
hirsh.pithadia.12@ucl.ac.uk

Enzo Fenoglio
elf@valyu.network
e.fenoglio@ucl.ac.uk

Bogdan Batrinca
bogdan.batrinca@valyu.network
bogdan.batrinca.09@ucl.ac.uk

Philip Treleaven
p.treleaven@ucl.ac.uk

Andrei Bubutanu
andrei.bubutanu.19@ucl.ac.uk

Radu Echim
radu.echim.19@ucl.ac.uk

Charles Kerrigan
charles.kerrigan@cms-cmno.com

The version(s) and contents of this paper are managed by the ERC721 token contract: 0x3cf9c821b1c0bc350b5e8e8f72296e94ea128c9

***Abstract*—Your Data (new gold, new oil) is hugely valuable (est. $13T globally) but not a 'balance-sheet' asset. Tokenization—used by banks for payments and settlement—lets you manage, value, and monetize your data. Data is the ultimate *commodity* industry. This position paper outlines our vision and a general framework for tokenizing data and managing data assets and data liquidity to allow individuals and organizations in the public and private sectors to gain the economic value of data while facilitating its responsible and ethical use. We will examine the challenges associated with developing and securing a data economy, as well as the potential applications and opportunities of the decentralized data-tokenized economy. We will also discuss the ethical considerations to promote the responsible exchange and use of data to fuel innovation and progress.**

Keywords— data, tokenization, data economy, data liquidity, data governance, responsible exchange, dataDAOs, data trusts.

I. Introduction

Data is referred to as the *new gold* and *new oil*, highlighting its value as a resource in the digital age. However, unlike these traditional commodities, data cannot be considered a *balance sheet* asset. Instead, it is a unique asset that requires careful consideration of its value, ownership, and control. This can be addressed by tokenization, an infrastructure often used by investment banks for payments, settlement, and trade finance, by attaching value to data and making it more liquid. The concept of data as an asset unlocks many novel opportunities as data has the potential to become the ultimate commodity industry. In this paper, we explore the challenges of making data an asset and examine the potential implications of tokenization for data liquidity, responsible use, and monetization.

In economic terms, an *asset* is anything that has economic value and is owned or controlled by an individual or organization [1]. Data is increasingly recognized as a valuable resource for generating future business & individual incomes. Forthcoming regulations in the EU [2] [3], and the US [4] recognize the need for that. However, according to the current International Financial Reporting Standards (IFRS), the US Generally Accepted Accounting Principles (GAAP) and the Chinese Accounting Standards (CAS), data is not yet recognized as an asset in the statement of financial positions [5]. This does not mean that data has no value. Definitively, the value of data cannot be decided by a single authority, as its worth is determined by the dynamics of the market and the actions of its participants. In this paper, we will show that data is an asset because it possesses the characteristics of an asset, such as ownership, control, and value, which are attributes typically associated with traditional tangible and intangible assets.

Under English law, there is a long tradition of court decisions that, in broad terms, make the point that data is not property because sharing data and information is not analogous to a transfer of property—after they are shared— both parties have it rather than it moving from a transfer or to a transferee. Therefore, data must have legally recognized characteristics, and creating an asset out of data takes this into account —Data ownership can be established through legal rights, agreements, and regulations [6]. Control over data is essential for the effective management and utilization of data. Assigning value to data is more elusive and can be complex and subjective. To this end, we describe a data tokenization framework designed to create and monetize data assets, including data, models, and their derivatives. During this process, we were inspired by the maritime shipping containers industry, which facilitated the global trade of physical assets. However, we adapted and extended the framework

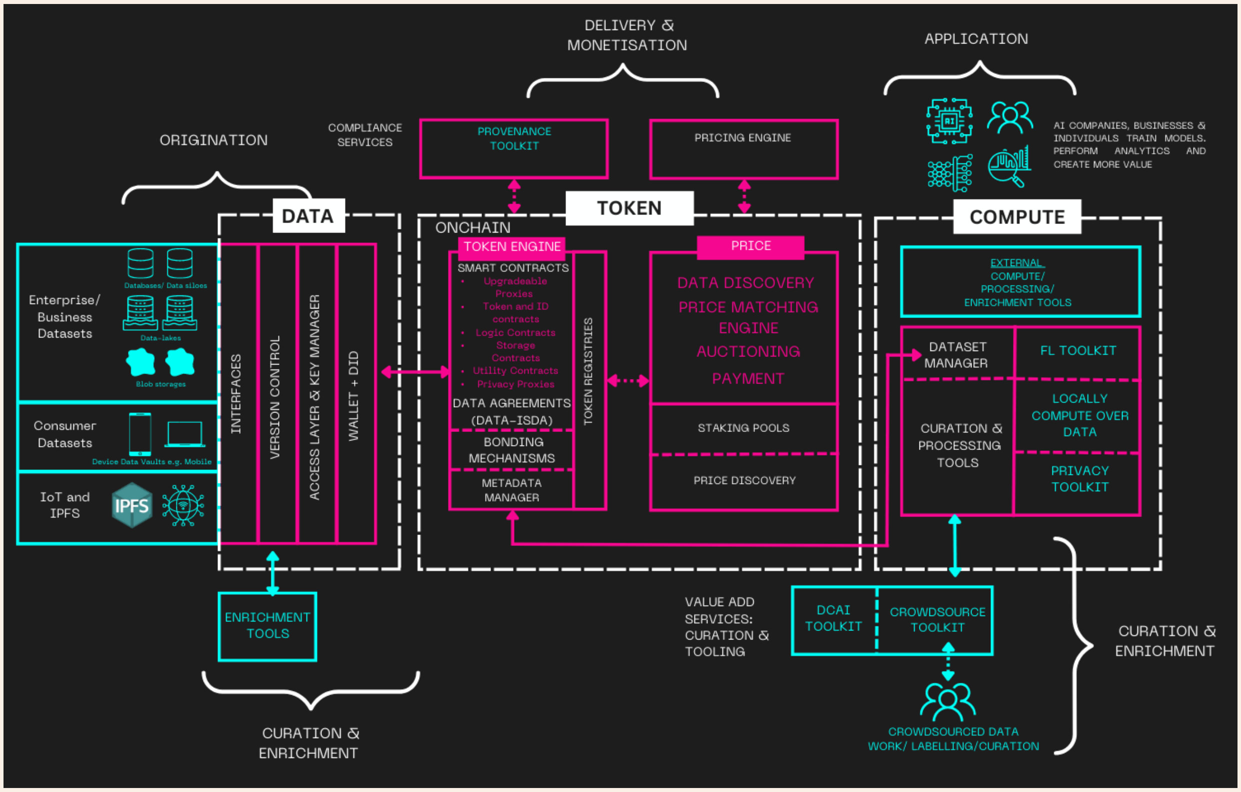

Fig. 1. The Data Tokenization and Valuation Framework

to meet the challenges of the new decentralized data economy. Ultimately, this paper presents a comprehensive framework for data tokenization and valuation. We outline the necessary steps and considerations to develop a rationalized set of requirements for data tokenization, focusing on its relevance for organizations operating within the data ecosystem and the digital economy [7]. We aim to provide a more thorough understanding of the possibilities and implications of data tokenization for managing data assets in the context of both data mesh and data fabric frameworks [8].

Our vision for Web 3.0 is one of a global, distributed marketplace of Digital Assets to ensure that any participant can earn from and trade data fairly and safely without sacrificing privacy or security. In our design, the Data Tokenization and Valuation Framework leverages foundational technologies such as blockchain, automated peer-to-peer services, decentralized storage, etc., to name a few, for creating new data-sharing and exchange platforms that naturally lead to enabling more secure, transparent, and equitable data and value exchange. Specifically, we envision a marketplace wherein data and their derivatives can be created, consumed, and exchanged for monetary and non-monetary value for stakeholders of all kinds, whether they are data subjects, data creators, data controllers, ML practitioners, algorithm/model creators, producers of compute capabilities or any other digital asset—in the pursuit of building a sustainable and equitable Decentralized Data Economy (DeData). By promoting transparency, ethics, and best practices, we can unlock the full potential of data and maximize its value for all stakeholders.

The process for a decentralized data economy can be decomposed into three main components: *Data, Token*, and *Compute*, as presented in Fig.1. In the following sections, we will defend the argument that only when data is tokenized, regulated, and governed we foster data security, equitable circulation, creation, and liquidity at scale:

- By tokenizing data, individuals and organizations can retain ownership and control over their data and choose to share for value —*rights-preserving* data access and monetization.

- By implementing regulations [7] that protect data privacy and security and enable decentralized governance of data-sharing networks, organizations can create a new economy in which data is shared and exchanged transparently, securely, and equitably.

- By implementing governance tools [9] [10], we ensure the quality, integrity, security, and usability of data collected and the development of clear and transparent policies and procedures for managing data and resolving disputes.

The ongoing work by the Law Commission of England and Wales [11] set out the basis on which digital and tokenized assets can be recognized under English private law and also proposes a law reform that would support this. This line of work is significant to the ideas in this paper because legal recognition of assets and relationships means that they can be enforced through the courts in the same way as other legally recognized items

Crafting regulations, establishing governance systems, and standardizing tokenization—complex and critical components of the decentralized data economy— presents several challenges which necessitate active management and planning for the successful implementation of a system with these characteristics.

The rest of the paper is organized as follows: in section (Sec. III), we provide background information on the evolution of data management and the challenges associated with traditional data architectures. We introduce the concept of data ownership and discuss the limitations of current models. Section (Sec. IV) introduces data tokenization using a metaphor drawn from the maritime shipping container industry (Sec. IV-A). Then we discuss the technical underpinnings of data tokenization, including blockchain technology and smart contracts. In section (Sec. IV-D), we explore various methods for data valuation, including traditional and data-centric valuation methods, the challenges associated, and the potential for data tokenization to enable more accurate and efficient methods. In section (Sec. V), we introduce the data value lifecycle and how data tokenization can be integrated into each stage. We also describe the data tokenization taxonomy components as Data (Sec. V-A), Compute (Sec. V-B), and Token (Sec. V-C). In section (Sec. VI), we discuss the advantages of data tokenization over existing data architectures. We explore the current (Sec. VII-A) and future (Sec. VII-B) applications of data tokenization, including its potential to enable decentralized data management and analysis, data mesh and data fabric approaches, trusted research environments, data ecosystems, data institutions, data trusts, funds, cooperatives, and DAOs. Section (Sec. VIII) presents a case study of the **Valyu Framework**, a data tokenization and valuation framework that is technically sound and commercially viable. We conclude in section (Sec. IX) emphasizing the potential for data tokenization to enable a decentralized data marketplace with improved regulation and governance, increased data liquidity, and interoperability.

II. Related Work

A growing number of papers investigate the multifaceted aspects of data. The existing literature is overwhelming in terms of the number of publications and variety of arguments. Nevertheless, we provide a selection of works that investigate some of the data aspects described in this paper but without our multidisciplinary focus. We hope this may help the reader to follow our vision better. In particular,

- *Data Liquidity.* A short introduction to data assets as a pathway to building high data liquidity is in [12]. How a liquid healthcare data system can facilitate managing large volumes of data from disparate sources is discussed in [13].
- *Data Tokenization.* In [14], data security and storage optimization of data, including tokenization, are discussed. The authors propose a view to removing privacy issues of IoT sensors. Other aspects of tokenization but for different data-based activities are presented in [15] [16] for data trusts, [17] for data dignity, [18] [19] for data-centric AI, or in [20] data provenance,
- *Data Value.* Big Data and Business Value are discussed in [21], which examines the role of big data in creating business value. Challenges and opportunities in the data-driven economy for the role played by data governance are discussed in [22]
- *Data Decentralization.* In [23], the authors present a system for healthcare data based on Ethereum private blockchain and IPFS for data sharing and uploading. A conceptual blockchain-based compromised firmware detection and self-healing approach for sharing IoT datasets are described in [24].
- *Data Capital.* Data as Capital is discussed in [25]. The author argues how understanding data as a form of capital can better analyze the meaning, practices, and implications of datafication as a new political and economic regime.

III. Motivation

A. The Data Paradox and Value Gaps

1) An Unrealized Value Gap

Data is created at an unprecedented scale (44 zettabytes as of 2022 and still growing). Nevertheless, innovation/research/businesses cannot find it when needed — then the data paradox. This imbalance, alone worth £80 billion in the UK [26], has led to an immense value gap between the two, as shown in Fig.2. Additionally, the demand for data (and the gap) is only growing due to developments in novel data-hungry architectures, such as transformer architectures for large language models (e.g., OpenAI's GPT-3.5). Few have derived this immense value for themselves because (1) they do not have the means and AI expertise, and (2) they do not own the data or spend millions of dollars curating the relevant datasets. This paradox is due

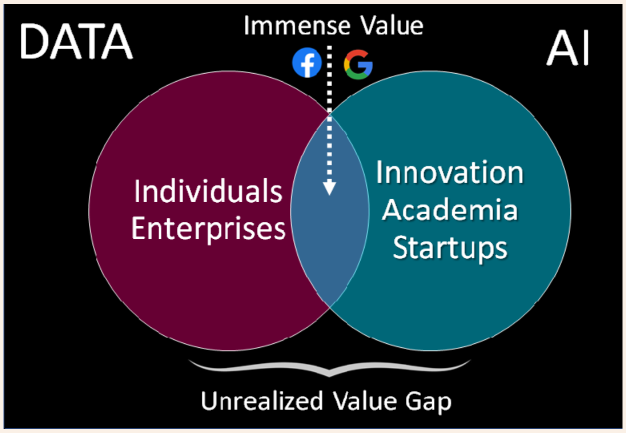

Fig. 2. The unrealized value gap

to a disconnect between those who create/own data and those who want to use it.

We argue that value-based incentives, governance, trustworthiness & transparency, and standardization are fundamental to reducing this gap by enabling the flow of data across the gap —Data Liquidity. However, they must be the focal point of the entire Data Value Cycle, from the creation/origination of data to its enrichment, delivery, and ultimate application/use (Fig.3). Developments in Smart Contracts, Blockchains & Distributed Ledgers, and Peer-to-Peer Networks can implement these three key ingredients. The idea (much like a shipping container) is simple; create an abstract representation of data through a token. This token acts as a *trustworthy* [27] instrument for managing Compute on the data, orchestrating governance (e.g., access controls), managing trust, and implementing incentives over the entire Value Cycle, as seen in Fig.4, which presents the top-level view of Fig. 1 divided into the three components: *Data, Token,* and *Compute* (Sec. V). Another benefit of this approach is that it treats data as a first-class *asset*, enabling it to be priced or traded easily, which is much needed for Data Liquidity. We remind the reader that an *asset* is a resource with economic value that an individual, corporation, or country owns or controls, with the expectation of providing a future benefit—Ownership, control, and value make an object an asset [1].

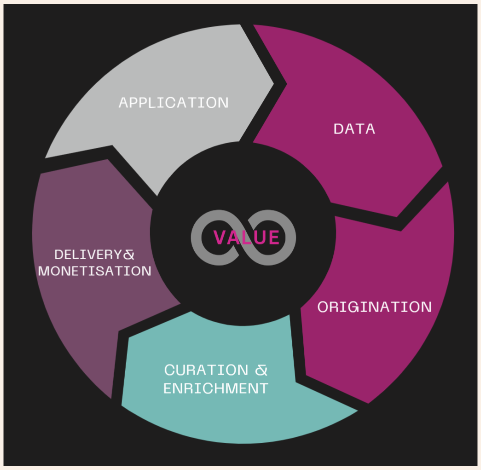

Fig. 3. The Data Value Cycle

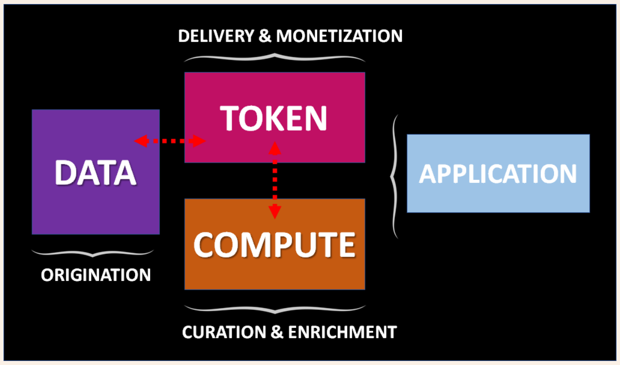

Fig. 4. The Data Value Cycle and Data Tokenization

2) Why does this gap exist?

The potential of the data economy remains locked due to several existing challenges:

- *Unavailability of data*: data is a ubiquitous resource. Data availability and accessibility remain a challenge. Aside from the government, much of the data generated today remains siloed and unavailable for use by the private sector (businesses and individuals).
- *Low quality of available data*: Even where datasets may be available, if they are incomplete, mislabeled, or in an unstructured format (e.g., Syslog from highly distributed networks), significant effort is required to clean, scrub or digitize data to derive potential or intended value.
- *Lack of interoperability of datasets*: Where data from different sources must be shared/analyzed/integrated, lack of uniform standards and protocols is a barrier to making sense of data. For example, in healthcare, the fact that different health service providers may be recording health records of a patient in different formats makes data portability arduous. Other than technical interoperability, jurisdictional interoperability, often overlooked, is a major challenge. Emerging and constantly evolving data sovereignty legislation is becoming significant to the flow/use of data across borders and data subjects.
- *Discoverability*: Where data is available, it cannot be easily found or understood. Metadata, cataloging, and standardization can address some of these challenges, but the more significant challenge is discovering the

value of data. Data is considered an *experienced good*[1] hence discovering their value or potential is difficult.

- *Lack of Trust*: Trustworthy systems and processes must be in place to ensure the responsible, ethical, and fair use of data because entities across the value gap may not trust each other.
- *Regulation*: data regulation is lagging behind exponentially growing data generation trends; lack of uncertainty, clarity, and fear have pushed data owners to enforce irrational protectionism. Some impactful regulation has been released long after e-commerce and social media has become ubiquitous, such as GDPR, the Digital Markets Act, Digital Services Act, etc. However, most of these acts apply to regional jurisdictions. They require specialized legal advice, which may hinder innovation across start-ups that may be wary of IP rights yet have limited funds.
- *Insufficient mechanisms to distribute risks and rewards*: Asymmetric risk and reward allocation across entities. Data owners often bear reputational, commercial, regulatory, and security risks when making data accessible. Whereas the data users often own most of the rewards.

The Data Economy is about translating this Data Value Gap into a Data Value Cycle: a positive sum value cycle. The Decentralized Data Economy enables everyone to achieve this.

B. Why we need a Decentralized Data Economy

Nowadays, the amount of data available is produced at an unprecedented scale: 44 zettabytes of data worldwide as of 2020, and it is expected to quadruple to 175 zettabytes by 2025 [28]. Sensors, wearable devices, and IoT devices continue translating physical movements into data points. When browsing the internet or using social media, tools process every click generating massive amounts of click-stream data. These instances of large-volume data producers sit on top of more traditional data generation mechanisms, ranging from financial time series, natural language processing (including human-labeled annotations) and various text corpora, image data sets, retail data (order/customer), individually-generated data (e.g., browsing history), etc. Enabling organizations to gather, store, manage, and manipulate vast amounts of data at the right speed is the scope of Big Data. In the last few years, big data has received much attention in academia and industry [29], [30]. The leading characteristics of Big data are traditionally described by the **Five V's**: data *Volume, Veracity, Velocity, Variety,* and *Value*

1) **Volume**: the massive amount of data generated, gathered, and processed. The generation of continuous data at high frequencies and volumes poses severe problems in terms of bandwidth and storage requirements.
2) **Veracity**: the quality of data, such as correctness, consistency, trust, security, and reliability.
3) **Velocity**: the speed at which data is generated, processed, and moved between different systems and devices.
4) **Variety**: the different types of data used to achieve the desired information or results.
5) **Value**: the different benefits of processing and analyzing data. However, although data is available, it can still be easily consumable and create immediate value. Due to its importance and impact, *value* represents the ethos of our framework.

The *Five V's* provides a valuable framework for understanding the challenges and opportunities posed by big data and helps organizations to plan and design their big data initiatives. However, we note that the *Five V's* oversimplify some of the complex and multifaceted nature of Big Data, overlooking other important factors worth considering beyond these five, such as data liquidity. In addition, the relative importance of each of the five characteristics will also vary depending on the specific context and use case and may evolve as new technologies and techniques emerge. Nevertheless, they can provide initial guidance while adopting a more dynamic and data-centric view, improving their organizational role. That requires focusing on data quality, governance, management, and a proactive approach to data analysis and decision-making. By taking a more active and data-centric approach, organizations can derive the maximum value from the data collected and achieve better outcomes from their big data initiatives.

In light of the limitations of the *Five V's*, it is also necessary to re-evaluate the traditional concept of *raw data* [31] and *data mining. Data mining* is a misleading term and a weak description of the data analysis process [32]. It implies that data is objectively and neutrally mined without the influence of human interpretation or prior knowledge. Data analysis is far more complex and subjective, as data scientists often have expectations and prior knowledge that influences their data collection, processing, and analysis. Furthermore, data is never neutral and can be subject to misuse, inaccuracy, falsification, and other forms of abuse. Therefore, it is more accurate to consider that *raw data* does not exist and that potential uses, expectations, and context influence the most basic perceptions and theoretical constructs. Data exists only as a solution to a practical problem and brings social, political, moral, and ethical connotations that determine what to collect and how to collect. Knowledge must exist before information can be formulated and data can be measured to form information [33]. Data can emerge if a meaningful structure, or semantics, is fixed and used to represent information—after knowledge and information are available.

[1] In Economics, an experienced good is one whose value cannot be determined without first consuming it.

The interpretation of the results of data analysis is also shaped by data scientists' understanding and prior knowledge of the subject of interest. So while the term *data mining* is useful as a shorthand to describe certain aspects of the data analysis process, it is important to consider that data is never given but taken [34]. Thus, the process of consuming data is much more complex and involves a significant amount of human interpretation and understanding. A better way to describe the data pipeline is *data processing pipeline* or *data manufacturing* [25]. This term captures the idea that consuming data involves a series of steps that involve collecting, cleaning, transforming, and analyzing data, as well as interpreting and communicating the analysis results. Nobody is *mining data* but instead *consuming data* or *capturing data* for knowledge discovery, which involves a deliberate and active effort to extract insights and knowledge from data and format rather than a neutral and objective process of extracting raw material.

Moreover, the *raw data* view is passive, emphasizing that data is a neutral starting point. In contrast, the *data processing pipeline* view for collecting, cleaning, and transforming data is an active view that highlights the effort and interpretation involved in the data analysis process. The active view acknowledges that data analysis is a complex and multi-step process that requires a range of skills and knowledge and that the analysis results are shaped by the data processing and cleaning that precedes it.

This paper proposes a data-centric approach that conceptualizes the product as a *whole product*. By understanding the needs of customers—the people who will consume data, we attempt to answer the question, *"What are the business insights and the economic value that data can offer?"*. We argue that the whole product concept [35] applied to data is an apt fit in our framework since it emphasizes looking at data as a holistic experience rather than individual characteristics or features associated with the data product. It is not just a matter of more sophisticated tooling for treating data volume, data speed, data discoverability, data liquidity, or other particular characteristics separately, but how to identify the value of data as soon as it is created and attach a price to a data product in the data processing pipeline [36]. In short, we must go beyond the data management and delivery activities of a data fabric and explore how a decentralized data economy (DeData) may provide better insights and contribute to the global economy (Sec. IV-A).

C. What is the value of my data?

Determining the value of data is challenging for several reasons: 1) data is an experienced good, 2) its value is not static and changes across the Value Cycle, 3) the value of data is often confused with the value of the information derived from the consumption of the data.

A complete valuation taxonomy and mechanisms are beyond the scope of this paper, but we briefly describe our approach. We view the value of data to consist of both *apparent value* and *latent value*:

- *Apparent Value*- is the observable value determined from datasets' inherent value drivers. Value drivers influence the readiness and usefulness of data for monetization and use.
- *Latent Value*- is the potential or inferred value a dataset may have when consumed/exposed to an application or market.

Apparent Value

Apparent Value is based on the intrinsic properties that characterize the usefulness of a dataset. Examples include:

- Consistency
- Completeness
- Accuracy
- Timeliness
- Usage (e.g., permissions/ restrictions)
- Interoperability

Latent Valuation

The latent valuation of data is a complex process. There is no one-size-fits-all model because 1) the actual and anticipated use of data may be different 2) the utility that a dataset offers can only be quantified once consumed. However, several approaches are used to determine this, depending on the specific context and intended use of data:

- *Cost approach*- estimates the value of the dataset based on the cost of replacing or replicating it. These costs are further broken down into the costs of collecting, curation, processing, and storing data.
- *Income approach* estimates the value of data based on the current and future income that data can generate. This can involve analyzing the revenue or cost savings the data can generate for the buyer and estimating the net present value of those benefits over time.
- *Market approach*- estimates the value of data based on comparable transactions in the market. This can involve researching the prices of similar data sets sold in the past and using those prices as a benchmark for valuing the data in question.
- *Real options approach*- estimates the value of data based on option valuation techniques [37]. It involves considering data as an intangible asset, much like a patent, and applying options pricing to determine the value [38].
- *ML approach*- uses machine learning algorithms based on statistical techniques that can learn from data to make decisions or predictions. These algorithms rely upon input features or variables to generate their output. Some of these variables are more informative than others and more influential to the model's prediction

process. It can involve training models to make predictions based on the data and using the accuracy of those predictions as a proxy for the data's value. It can also involve identifying the feature importance to determine compound classification and activity prediction using an ML model—similar to the use of model-agnostic models such as **SHAP** [39] in AI explainability

- *Network effect approach*- estimates the value of data based on the size and quality of the network that uses the data. The key idea is that data becomes more valuable as more people use them, so the value of the data is estimated based on the size and activity of the network

D. Data scarcity vs. Data abundance

Regardless of the relative abundance or scarcity of data, adopting a data-centric perspective and leveraging the data to produce enhanced outcomes is imperative. This necessitates an emphasis on data quality, governance, and management, as well as an active stance on data analysis and decision-making. In the dynamic data-centric view, the scarcity vs. abundance of data is an important consideration —the focus shifts from simply collecting and storing to using data to drive the best outcomes. When data is scarce, organizations may need to be more selective about what they collect and store. They may need to put more effort into verifying the quality and accuracy accordingly. In these cases, a clear understanding of the data value and how it will be used can help organizations to make more informed decisions and prioritize their data initiatives. When data is abundant, organizations may face different challenges, such as managing and storing the sheer volume of data and ensuring that data is of sufficient quality to support effective decision-making. In these cases, organizations may need to focus on implementing effective data management and governance processes and developing the necessary infrastructure to handle the volume collected and stored.

IV. Data Tokenization

A. Globalization and The Shipping Container Metaphor

This section briefly describes the decentralized data economy (DeData). We use the metaphor of container economies generated by the maritime shipping container industry [40] to introduce the concept of *tokenized data economies*.

Our analysis starts by observing that there are several terms in the shipping and global trade industry (containerized shipping industry, shipping container, containerization, maritime shipping, cargo, sea, etc.) that loosely correspond to terms of the data economy (decentralized data economy, tokenized data, tokenization, data transfer, tokenized data, blockchain, etc.). Indeed, this correspondence can serve as a metaphor and inspiration to foster intuition and describe the impact that DeData can make on the global economy.

The maritime shipment container is a technological innovation credited to Malcolm McLean[2], who realized that by encapsulating goods inside standardized containers, the loading and unloading of ships and trains could be at least partially mechanized. This made the logistics—the transfer from one mode of transportation to another—seamless. The main advantage was that goods/products could conveniently and securely remain in their containers from the point of manufacture to delivery, resulting in reduced costs and risks in terms of labor and potential damage. It also promoted the growth in trade and the standardization of trade-related administration and governance; the humble shipping container catalyzed globalization. Maritime shipping has been much more than containerized shipping, *"Maritime Transportation is not simply an enabling adjunct of trade, but is central to the very fabric of global capitalism"* [41]. Building on the shipping containers analogy, we investigate the potential impact of tokenization and the resulting data economy by drawing parallels to the transformation brought about by containerization in the shipping industry. The transfer of data through tokenization has the potential to create a profound transformation in the global economy, comparable to the impact of containerization on the shipping industry. Utilizing the shipping container industry as a metaphor, we explore how insights gained from the successful transition to containerization can inform organizations of various aspects of decentralized data tokenization in the data economy sector. We believe it could potentially play a similar role in the data economy as containerization did in the context of globalization. Specifically, it could be a standardized tool for trading, commercializing, managing logistics, and governing data, providing data liquidity at scale—data liquidity is expected to be a multi-billion dollar industry [7]. Thus, data tokenization could make data that are typically at rest or siloed available for increased profitability, whether the purpose is providing better outcomes for public health, agriculture, and scientific research, as well as providing access to data from smart devices at the edge.

B. What is Data Tokenization?

In technical terms, *tokenization* is the process of issuing tokens representing an asset where the rules and behaviors governing the asset are encoded in smart contracts (automatable agreements that may be enforced via tamper-resistant execution and even through courts if they have certain characteristics [42]), providing a powerful abstraction to manage and distribute value. *Tokenized economies*

[2] Apparatus for Shipping Freight US2853968A

refer to networks of digital assets and services powered by tokens used to access and pay for goods and services. Tokenized economies enable users to securely and quickly transact with each other without the need for a trusted third party. They also allow the creation of new digital assets, such as tokens and cryptocurrencies, that can be used to power new business models and services.

Data Tokenization is the process of issuing a token where the underlying *asset* is data, and the rules and behaviors pertinent to the asset are encoded as smart contracts. This creates *tokenized data*[3]. Similarly, *tokenized data economies* are digital economies where data is securely stored and tokenized, allowing it to be exchanged and tracked on distributed ledgers and across blockchains. This economy allows businesses/individuals to store, protect, and monetize their data, creating a transparent, fair, and secure data marketplace for data[4] to let humans and computers exchange freely and quickly. For example, data tokenization can be used for many data-based activities, including data trusts [15] [16], data dignity [17], data-centric AI [43], data provenance [20], data-driven investments, data-based payments, and the like.

From a purely economic perspective, *tokenized data economy* refers to how capital is constituted, moved, and destroyed through particular circulation and value accumulation methods during data transit and transfer of physical or non-physical assets. That is made possible and shaped by tokenization which is central to what we call the *decentralized data economy revolution*, by which we mean a much broader set of transformational activities that go beyond technology since it also embraces social, political, ethical, and environmental aspects at the intersection of different sciences: technological, business, social, and humanistic sciences.

C. A holistic approach to a multifactor problem

Data Liquidity is a multifactor problem involving technical, social, legal, and economic factors, among others. The economic perspective in data tokenization is important but is not the only one. The problem becomes even more complex by adding regulation and governance.

- From a technical perspective, data liquidity requires the development of robust and secure mechanisms that can handle large amounts of data and ensure that the data is exchanged in a compliant and secure way.
- From a social perspective, data liquidity requires the creation of a culture of trustworthiness and transparency among data providers and consumers. This is achieved by clearly explaining how data is used, who is accessing it, and for what purposes.

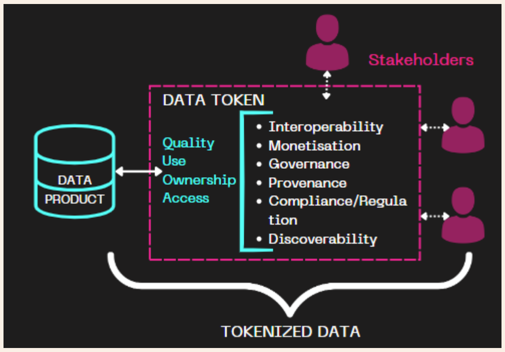

Fig. 5. Data Token

- From a legal perspective, data liquidity requires compliance with data protection laws, such as GDPR, and ensuring that data is shared in a way that respects people's privacy rights.
- From an economic perspective, data liquidity requires creating a viable business model that will allow data providers to monetize their data and data consumers to access the data they need at a reasonable cost.
- From a governance perspective, data liquidity requires the development of clear and transparent policies and procedures for managing data and resolving disputes.

All these factors need to be considered to create a sustainable ecosystem for data liquidity that respects the rights of data providers and consumers. To add to the complexity, each of these factors influences the other[5]. This affects data liquidity and manifests as the value gap discussed in Sec. III-A. Another significant challenge is that each factor has been addressed individually or in part but not together. Data tokenization can help address all these factors collectively.

What we need is a single *trustworthy container*, an instrument (i.e., an abstraction) of the underlying data to address all these factors holistically, as seen in Fig.5. This abstraction also decouples the data source from the purpose or intended use [7]. Therefore, we tokenize the entire value cycle and encode each factor and its underlying processes as part of the instrument. Above all, the leading aspect of **DeData** enabled by data tokenization is the standardization system rather than the token itself. That, by the way, perfectly corresponds with the maritime shipping containers metaphor (Sec. IV-A), where it was the standardization of cargo containers in the containerization that promoted the global trade revolution [44]

D. Capturing Value in Motion

Micro

From a micro perspective, we introduce the Data Value

[3] We will use the terms Data Token, Tokenized Data, and Data-Backed Token interchangeably

[4] Often referred to as Data Exchanges [7]

[5] For example, lack of technical & jurisdictional interoperability and discoverability can influence how a dataset is monetized

Cycle, Fig.3. It represents the data lifecycle from *production* to *consumption* across different value-add progressions. It starts with data, often siloed or inaccessible, and raw[6]. The first step is usually making this data into a *data product*. A data product is a self-contained data *container* that directly solves a business problem—through a product or service— or is monetized—as datasets or APIs. Raw data is productized through origination, where raw data is sliced, diced, and processed to be made available for use upstream. For the purposes of this discussion, datasets will be taken as the primary data product for tokenization, with all future references to data tokenization implying this premise.

Subsequent steps such as cleansing, enrichment, and delivery often play an intermediary and complementary role in enabling the production of **Data Quality**. Data delivery is a key yet often overlooked step that, if facilitated, may lead to financial reward via markets or other mechanisms—how/when data should be made available and what rights, purposes, and prohibitions are attached to their use. Ultimately, data used for analytics, model training, and services can create more data to be fed into the data tokenization and decentralized data economy, creating a cyclical process of value generation.

Each stage often has different stakeholders[7]. Stakeholders could include the data owners, data custodians, government and regulators, innovators/enterprises/researchers, and other ML and data practitioners. Most notably, each section of the cycle (and, therefore, the underlying data) is influenced by or influences the data value. Interoperability, discoverability, provenance, governance, monetization, compliance, and risk allocation become crucial to the flow of value.

Additionally, each stage can be a positive sum value gain function giving data the characteristic of capital—Data Capital, the idea that data is a factor of production for the data economy [45]. Most data scientists and data/ML practitioners are familiar with a *rolled out* version of this cycle, often referred to as data pipelines (Fig.6). Implementing this process on a large scale while maintaining proper controls is the practice of DataOps [46][8].

Macro

At the center of tokenized data economies, there is the need to keep *value in motion*—data liquidity at all costs.

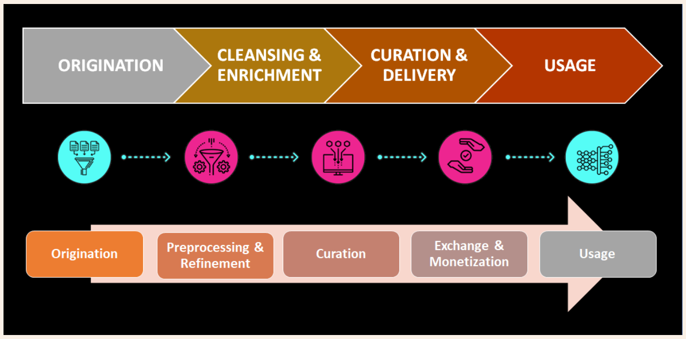

Fig. 6. Data Supply Chain: **Origination**- Datasets, including siloed or walled datasets, made available as *data products* and tokenized. **Preprocessing and Refinement**- These *data products* can be further processed or refined to improve quality. **Curation**- The tokenized data can be further curated by the owner for a particular use case by combining it with other tokenized datasets or improvements through crowdsourcing. **Exchange and Monetization**- The tokenized data product can be monetized or shared/given access to by the owner to counterparties/stakeholders within the value chain. **Usage**- The datasets can be used to train new models or for analytics. This usage can be explicit–sharing or passive– by providing compute access.

Value must be kept in motion to improve liquidity because liquidity is the ability to convert assets quickly and easily into cash. When the value is kept in motion, assets are actively managed and used to generate cash flow or revenue. That allows an organization to have a steady flow of cash on hand, which can be used to pay bills, invest in new projects, or return to shareholders. For example, a business that keeps inventory in motion by regularly selling products or services will likely have better liquidity than a competitor that holds onto inventory for long periods without selling it.

Similarly, an investment portfolio that is regularly bought and sold, or a real estate portfolio that is actively managed to generate rental income, is likely to have better liquidity than one that is stagnant. Additionally, keeping value in motion is vital for organizations because it allows them to adapt to changing market conditions, such as shifts in consumer demand or changes in the economy. By actively managing their assets and generating cash flow, organizations can respond to these changes and maintain their financial stability.

The imperative to capture all data from all sources by any means influences many key decisions about business models, political governance, and technological development [25]. Creating value and capturing value is not necessarily related [47]. It is possible to create value without capturing it. Farmers create value through the cultivation and growth of crops, but this value will not be captured

[6] Describing data as *raw* is a misnomer; what might be considered raw for one process might be the result of a previous business/compute process (Sec. III-B)

[7] For example, the data owners may be different from the stakeholders that use it to train the model

[8] DataOps partially addresses technical governance, interoperability, discoverability, etc.

unless the crops are harvested and sold. Capturing without creating value is illegal or impossible because it will be considered taking value that already exists, such as theft or fraud. However, creating less value does not necessarily mean capturing less value. A company can still capture value by being efficient and innovative in its business practices, even if producing less. Additionally, a company may capture more value by creating a unique or desirable product or service, even if the value created is less than that of its competitors.

In our analysis, we examine the circulation of data and capital in the tokenized data economy, drawing upon the theories of several authors such as K. Marx [48], K. Polanyi [49], M. Castells [50], A. Negri & M. Hardt [51]. We look at Marx's theory of production and circulation of physical commodities as a historical foundation and extend it to the digital age. In particular, we consider the two-fold problem represented by data circulation: the investment opportunities created by the production and distribution of digital data and the implications of the movement and exchange of these data within the marketplace. The circulation of data, like commodities, plays an important role in creating value. It is a significant source of profit and growth in a process called *datafication* [52]. The integration of datafication into capitalist co-processes reveals the potential of data, devices, and platforms to reconfigure the spatial organization of production and social reproduction to direct the circulation of money in favor of the corporate entities overseeing digital conditions of everyday life [53]. In this context, it is natural and also consequential from the existing literature on the social, political, and economic dimensions to take data as commodities and extend the argument of circulation of physical commodities to data for *keeping the flow of capital in motion* [54]. In this sense, it becomes apparent how tokenized data circulation will directly contribute to the growth of the global data economy, where capital accumulation in decentralized data economies is produced by adding, moving, and destroying value through the data supply chain, paralleling the process of capital accumulation in container economies through the maritime supply chain for physical goods. This is great as it suggests tokenized data circulation holds the same potential to become an integral factor in the global data economy as containerized goods circulation [40]. However, there are remarkable differences between the circulation of data and physical goods:

- In modern economic research, data is conceptualized as a form of capital rather than a commodity. This distinction is necessary to analyze digital capitalism and its dynamics of perpetual capital accumulation and circulation [25]. This paper follows the same approach and will not consider data a commodity. Data is neither a commodity nor a uniform, generic, static raw material like oil or gold, as it may appear in the first place. Instead, data is a product of several decisions on aggregation, filtering, deletion, and recording, which are usually irreversible. This variability makes it challenging to assign consistent values to data [55]. Nevertheless, we can consider tokenized data a commodity exchangeable, irrespective of the supplier [56]. Data-backed tokens can be assumed to be a commodity because they are a digital representation of a valuable asset stored on a distributed ledger. They can represent other assets, such as commodities or stocks, where the underlying asset's value determines the token's value and where the token can be used to facilitate the trading of assets in the DeFi economy.
- Another difference is liquidity. In the case of container economies, transnational trade and just-in-time logistics are not frictionless despite the normative efforts of multinational institutions and the unrealistic expectations [57] of investors. On the contrary, tokenized data can realistically be expected to be frictionless in such a way as to make the decentralized data economy even more impactful to the global economy than what the maritime shipping containers already did.
- A final difference is about the role of intellectual labor [58] in the production and circulation of capital. For the shipping industry, the circulation of capital depends on the labor of making cargo physically move. In digital data economies, it includes the work of data analysts, data scientists, and other personas who contribute to creating and distributing digital goods and services on their platforms. Here we can consider the value created by labor and the implications for distributing those profits among stakeholders in the global economy, which may bring forth societal and ethical implications in acquiring personal information from users of other products/services [59] and to data dignity (Sec. VII-B2). This is particularly true for the degradation of work conditions that may accompany the rise of competing digital platforms incorporated into wider capitalization processes of platforms capitalism [60].

We have presented some of the existing similarities between the circulation of data and goods, but also their differences. We have just introduced some of the approaches and theories designed to extend the traditional theory of production and circulation of physical commodities to the digital age [48] since the question of value creation in the digital economy is still a matter of debate that goes beyond the scope of this paper [61] [62] [22].

V. How can you Tokenize data?

This section introduces the components for data tokenization, which are the *Data*, *Compute*, and the *Token*, as depicted in Fig.1 and Fig.4. The token is an intermediate and abstract representation of the data. It manages all data-related aspects, its ownership, control, usage (Compute), and value. For clarity and ease of understanding,

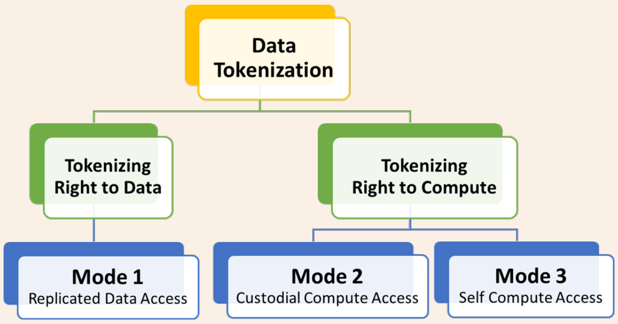

Fig. 7. Data Tokenization Taxonomy

this section will utilize the tokenization of datasets as a running example to explain the usage of data tokenization for Machine Learning. Using the complete set of features of each component is unnecessary, as their implementation largely depends on the specific use case. The goal is to provide a minimalistic framework for data tokenization based on these three components, each with several features that enable both form and function.

In this preamble, we will analyze the data tokenization taxonomy presented in Fig.7. Here are two key approaches to tokenizing data. The critical distinction is in the rights. The figure illustrates the different modes of tokenizing data that determine who owns/controls which component (i.e., Data, Token, and Compute). The first approach pertains to tokenizing the rights to the data, while the second approach consists of tokenizing the rights to compute over data. Each strategy has its advantages and drawbacks, but typically, the approach will focus on privacy and minimizing the risk of replicability of the data. (i.e., the *copy-paste* risk):

- **Mode 1:** The token provides the right to download and use a copy of the data, and additional usage restrictions, such as licensing arrangements, may be implemented. Replicability risk can only be mitigated using licensing agreements rather than systems-level approaches. This approach is the most widely used form of data-sharing rights. Whether no data enrichment enhancement or augmentation is necessary (e.g., for privacy reasons), then this mode requires only the *Data* and *Token* components (as shown in Fig. 4).
- **Mode 2:** The token provides rights to compute over the data, while other usage restrictions can be implemented using smart contracts via the *Token* component. Unlike the owner of the data and the user, a trusted third party is entrusted with the task of running computations. This mode requires the *Data*, *Token*, and *Compute* components (Fig.4). The *Compute* component is owned by a custodian. This form of data tokenization is particularly compelling when *Compute* is conducted simultaneously on multiple tokenized datasets, enabling the extraction of the combined value of data.
- **Mode 3:** The token grants rights for computing over data and may involve additional usage restrictions. The dataset owner provides a *Compute* environment that stays under their control to execute any required tasks. Critically data never leave their environment. This approach is necessary for highly sensitive data that cannot be exposed due to the *copy-paste* risk. The data proprietor must also provide a *Compute* environment that can be priced accordingly. This mode also considers regulations such as GDPR and Federated Learning (Sec. VII-A2). All the components required for this mode, such as *Data*, *Token*, and *Compute* (Fig.4) are owned by the data proprietor. We anticipate that the true potential of **Mode 3** can be achieved through use cases that require a *Compute* environment close to the location of the data, often under the control of the data proprietor, as seen with edge devices such as IoT, automobiles, and mobile phones.

Given the right level of transparency and oversight, all three modes can ensure that data is used responsibly, fairly, and ethically. For more sensitive data, **Mode 2** and **Mode 3** should be the preferred choice.

A. Data

The starting point is the underlying data collection to be represented and controlled by the token. The Data component has the following faculties:

1) Source and Interfaces

The data collection's source, format, and interfacing are established. This happens within the Origination stage of the Value Cycle. These are the minimum aspects needed to create a *data product* [63]. Data exist in structured (e.g., SQL/relational data), semi-structured(e.g., HTML, JSON, XML, NoSQL/document data), or unstructured storages (e.g., social media posts, presentations, chats, IoT sensor data) across the cloud-edge/ on-prem-edge spectrum. For businesses, data typically reside in databases, data lakes, objects storage like S3, or decentralized data stores such as the InterPlanetary File System (IPFS) for storing and accessing files, websites, applications, and data. Individuals and devices also possess data, typically stored on edge devices such as phones, laptops, personal clouds, and IoT devices. Commonly used formats for datasets include *CSV*, *JSON*, *Parquet* [64], and others such as *Avro* and *ORC*. Each format has advantages and disadvantages, and the choice often depends on the specific use case and the tools used to process the data. For example, *CSV* is a simple and widely supported format, while Parquet is optimized for performance and is often used in big data systems. *JSON* is popular for its human-readable format and support for nested data structures. *Avro* and *ORC* are optimized for storage efficiency and are commonly used in Hadoop environments.

Interfaces are the mechanisms required to transfer, consume or read the data. They play a crucial role in enabling data flow between different systems and can help ensure the data is accurate, consistent, and up-to-date. Interfaces entail different topologies. In particular,

- *File-based/ Format-based interfaces*: These interfaces transfer data by reading and writing files, which are format-dependent. File-based interfaces are common with datasets and in ML/DL applications as they provide a simple mechanism to work with and transfer to files.
- *Application Programming Interfaces*: (APIs) or Remote Procedure Calls (RPCs) allow different systems to communicate with each other. For example, Google RPC (gRPC) connects services in and across data centers with pluggable support for load balancing, tracing, health checking, and authentication.
- *Data virtualization tools*: These interfaces create a virtual representation of data from multiple data sources, making it easier to access and use.
- *Streaming*: distributed streaming platform can handle real-time data streams. For example, Apache Kafka provides a streaming interface that allows real-time data streams to be published and consumed through topics, with durable storage, providing low latency and high throughput, scalable and fault-tolerant platform.

2) Versioning

The data collection version is also critical. Data can constantly change as the collection gets modified knowingly (e.g., updates) or unknowingly (e.g., corruption during transmission). Versioning is necessary to keep track of any changes. It also helps preserve the veracity of data as the publisher/maintainer of the data intended. Like software engineering, Git-like version control tools can be used. Examples include *GitLFS* [65], *DVC* [66] [67], or *Oxen* [68].

3) Unique Identifier

We also require a globally unique identifier attached to the collection and its version. This allows the collection to be referred to unambiguously—An important requirement for attaching rights, permissions, value, and ownership. We propose the open W3C Decentralized Identifiers (DID) standard [69] as it is decentralized, globally unique, resolvable, highly available, and cryptographically verifiable. DIDs can also be associated with cryptographic material such as public keys and are complementary with W3C's Verifiable Credentials [70] that are used to ascertain claims about the data collection—necessary for assurance and provenance.

4) Metadata

The characteristics of the data collection need to be specified as metadata. Versioning, DIDs, and interfaces are also metadata. This provides useful information about the data collection, e.g., its quality, statistical distributions, known limitations, biases, etc., used for aspects such as provenance and apparent valuation. Google's Data Cards [71] and Hugging Face Data Cards [72] are some possible frameworks that can be used for metadata specification. We propose, as a minimum, a limited subset of attributes from the Google Data Card, including Dataset Overview (includes descriptive statistics), as an example of Data Points, Lineage, Annotations and Labelling, Validation Types, and Sensitive Attributes.

Metadata can also be perceived as the documentation for the underlying data collection- it can help researchers and practitioners understand the suitability of the data set for a particular task, such as training a machine learning model. Additionally, data cards can help researchers and practitioners share and discover data sets as they provide a standardized way to describe them and make them more easily searchable—e.g., discoverability.

5) Curation

The usefulness of data is characterized by aspects such as completeness, consistency, timeliness, and accuracy. These drive up the quality (value) of a dataset and data management throughout its lifecycle. Therefore, curation is necessary as it *refines* a dataset further. Paradigms such as *Data Centric AI* are based on data quality to avoid *garbage-in, garbage-out* issues during model training. Many data curation tools already exist [73] to help decide which data is relevant and which is not. Integrating these tools with tokenization can potentially help accelerate their adoption even further—especially for tools related to crowd-working.

Crowd Work

Human/Expert input (e.g., labeling and annotation) and validation are often necessary to improve data quality and accuracy (notable examples include Amazon Mechanical Turk). Leveraging tokenomics and the Web3 ecosystem (Sec. V-C8) can provide additional monetary and non-monetary payments for crowd-related data work. It also helps directly determine the value-add of a crowd-sourced task w.r.t. the dataset[9].

B. Compute

Compute is a key component for **Mode 2** and **Mode 3** data tokenization (Fig.7) and any additional enrichment, augmentation, or curation data needs. It relies on open tooling to work. The main distinction between the two is where Compute happens—either in the data owners' environment or a trusted third-party environment.. The Compute component has the following faculties:

[9] A means to determine the ROI on labeling and potential quality/value gain

1) Usage

Typically, an abstraction such as a containerized Docker Image [74] or *WASM* & *WASI* [75] [76] is needed to run Compute over data. This image/binary is typically provided by the entity wanting to compute over data. The image/binary must use the data interfacing discussed in the subsection above. The execution environment of this container can either be in the environment of the data owner or a custodial Compute provider. We even envisage complementary algorithm/ image audit services for additional assurance for all the stakeholders [77]—the data owner, Compute custodian, and algorithm/image developer.

This Compute can be mutable or immutable based on the restrictions imposed (Sec. V-C2). Race conditions resulting from mutability and multiple Compute components running concurrently on the same data must be handled accordingly. Similarly, the necessary security mechanisms (e.g., the ephemerality of the execution environment) also need to be considered.

Custodial Compute

The Compute custodian is responsible for providing a secure environment for the execution of computational tasks on behalf of the data owner and those requesting the computations. This may be achieved through cloud computing providers (e.g., AWS) or web3 compute protocols such as *Gensyn* [78]. This party must act neutrally and maintain proper security protocols to protect both parties from unauthorized data access or malicious activities—reducing the *copy-paste risk* of data sharing. The Compute environment can be run with Confidential Compute (Sec. V-B2) for added security.

Self Compute

This is similar to custodial Compute, but the distinction is that the data owner provides the Compute environment. The true potential of this is where the Compute environment is close to where the data resides. Edge devices, IoT devices, automobiles, and mobile devices are typical examples— *WebAssembly* (WASM) with the WebAssembly System Interface (WASI) can provide the right environment for such devices.

2) Privacy

Privacy is a challenge, especially when it comes to data liquidity. Privacy Enhancing Technologies (PETs) developments can help address some of these. The exact choice of PET depends on the security model and use cases for tokenizing data. We expect privacy during Compute to become particularly significant for custodial Compute-based data tokenization. Exploring PETs for tokenization is beyond the scope of this paper. However, for the benefit of the readers, we briefly introduce herein some of its key aspects[10]:

- *Differential Privacy*- captures the increased risk to one's privacy incurred by participating in a database [79]. A commonly employed technique to ensure differential privacy is the addition of random perturbations to the original data before its release. It can be effective in datasets that contain identifiable information about individuals. A distinguishing feature of this PET is that it has the notion of privacy budgeting and privacy loss- a means to keep track of the amount of privacy eroded during a computation. Accounting for these budgets with smart contracts (because of data tokenization) can provide a more holistic view of the privacy eroded over the life of the dataset.
- *Synthetic Data* involves artificial data generation by sampling from a given dataset. Rather than using the original dataset that may have sensitive attributes, the synthesized data is used in place—It is a widely used technique in privacy for machine learning.
- *Confidential Compute*- Confidential Computing protects data in use by performing compute in a hardware-based, attested Trusted Execution Environment (TEE). A TEE (e.g., Intel SGX) uses hardware-backed techniques to provide increased security guarantees for the execution of code and protection of data within that environment. This assurance is often missing in approaches that do not use a hardware-based TEE [80].
- *Secure Multiparty Computation*- SMPC involves jointly computing an algorithm/function from the private input (i.e., data) by each party without revealing these data to other parties[11]
- *Homomorphic Encryption*- HE allows computing and algorithm/function directly on ciphertext (i.e., encrypted data). Limitations include the inability to compute low-level operations and scale with large data.
  SMPC and HE are often used together; examples include SPDZ [81] and variants.

C. Token

In Sec. IV-C, we discussed the idea of a single trustworthy abstraction, a container, that decouples the source of data from purpose or intended use. We will now discuss how this abstraction can be the instrument for implementing technical, economic, legal, and governance perspectives that data is *subjected-by* and *subjective-to*. The nature of the Token is significant, not just in technical and economic terms but also in terms of law and regulation. Under English law and many other systems, depending on the characteristics of a Token, it may be subject to supervision by and registration with regulators.

[10] Significant developments are made in the field due to the growth in data regulation and blockchain industries

[11] in ML terms, this "function" could be a model's loss function during training or the model

The Token component is this abstraction, implemented with smart contracts—the token is a highly programmable instrument, and all these perspectives are directly embedded into the token. The Token component has the following faculties:

1) *Networks & Decentralization*

Networks are a key part of how tokens operate. If tokens are like shipping containers, networks are the ships and shipping infrastructure (Sec. IV-A). Tokens run on networks, the runtime environment for the token, and enforceability happens through them.

Depending on the level of centralization, trustworthiness, and privacy desired, data can be tokenized within various public, private-permissioned networks (consortium networks) [82] etc. This decision directly impacts the data liquidity that is achieved. Moreover, due to sector-specific requirements, there may be a need to tokenize certain information across different kinds of networks depending upon the level of privacy and macro-level governance needed. For example, for healthcare data, it may be suitable to tokenize within a private-permissioned network where network members have been thoroughly vetted to ensure data security and privacy[12].

Cross-chain protocols facilitate the tokenization of the same dataset across multiple blockchains, which affects the token's value or allow it to *move* from one ledger to another. The type of network used for the token will affect the enforceability and operational aspects discussed in this section.

2) *Governance*

Data governance is becoming increasingly important in the contemporary digital landscape. As the use of data becomes essential and commonplace, organizations must develop data governance models to ensure their data is used responsibly and well-managed. It is defined as the process of managing and overseeing the availability, usability, integrity, and security of data [84]. This process involves the **Five W's** (who, what, why, when, how) and provides an operational structure that defines the purpose, means, and conditions related to the data. In particular, the operational & often enforceable structure in relation to the purpose (why), means, and conditions/terms/duties (how/where/when) between parties (who) that need to work with or are responsible for a given dataset (what).

Tokenization offers the ability to program these policies, procedures, standards, and agreements directly into the token; the token becomes the means through which governance is administered. For the sake of intuition, the *Token* can be considered a passport for the underlying dataset. This programmable governance can be granular or high-level, depending upon the implementation semantics. We introduce three governance primitives: Ownership, Controls & Usage for implementation. They are the building blocks from which policies, procedures, and agreements are programmatically created, implemented, and enforced, depending upon the intended uses of the data (for example, the governance for a Data Trust[13] will be very different from that of Multiparty Data Access[14]).

We briefly discuss the objectives of each primitive [15]:

- *Ownership*: It manages all aspects related to the ownership of the underlying data, such as the ownership rights, prohibitions, permissions, and obligations. This primitive can also create radically new data ownership structures, such as fractional ownership of data assets [86].
- *Usage*: It specifies how data is used, the modes of data tokenization, and their respective requirements. For example, if **Mode 2** or **Mode 3** are applied, who and where is the Compute component, and what algorithms can be run—It works very closely with the Control primitive.
- *Control*: It defines access control to the usage and monetization of the data. This primitive can also be used to create a Data Custodian (Sec. VII-B1), an entity that controls the data on behalf of the owner

Control and Usage work closely together to specify rights, prohibitions, permissions, and obligations for monetizing and using the data. Usage determines the *what*, *where*, and *how*, whereas controls determine the *when* and *who* can use the data.

Augmentation: Legal Aspects & Voting

Smart Contracts offer the ability to augment each primitive with features such as voting and legal isomorphism. For example, voting can be used with the controls primitive to vote on who and how the data get to be used—if the tokenized data is under fractional ownership. Voting mechanisms such as quadratic voting [87] can potentially offer novel ways to govern resources (data) that affect many people.

It is also possible to augment these primitives with *Smart Legal Contracts* [42]. Smart Legal Contracts allow the legal narrative to be embedded with data agreements and policies. We believe that a large part of the issues regarding data liquidity and exchange is partially a result of the lack of legal standardization, composability, and automation regarding legal agreements for data use, exchange, and access. Exemplars and parallels can be borrowed from the

[12] Other attributes may include support for sector-specific technical standards, e.g., FHIR [83] or regulatory standards

[13] Sec. VII-B1

[14] Sec. VII-A1

[15] These primitives are expressed with deontic logic [85]

over-the-counter (OTC) derivatives market and their use of the International Swaps and Derivatives Association (ISDA) agreement [88] that standardized and automated the OTC derivatives market. As a result of this initiative, the relatively illiquid OTC derivatives market grew significantly, rising from approximately USD 100 billion in the early 1980s to a hyper-liquid value of USD 630 trillion by the end of 2022, illustrating a drastic improvement in liquidity [89].

3) Provenance

Data provenance is the information that describes the record of the origin, history, and description of a data set or data element, which includes information about the transformation and subsequent use of the data. It is the *biography* of the dataset and can be used to verify that data is trustworthy and is used for decision-making, compliance, and regulatory purposes. Data provenance is important in AI systems as it helps to ensure data quality and trustworthiness by allowing us to understand the origin of data and any potential issues or biases that may have been introduced during the data collection, processing, or storage. It can also help establish transparency and trust in models. Provenance is almost a *side-effect* of tokenization, with metadata, immutability, timestamp, digital signatures, and smart contracts used to keep track of the underlying datasets' key aspects. Data provenance documents the *who, what, and when* associated with a data element and helps ensure the accuracy, trustworthiness, and traceability of data.

4) Lineage

Data lineage is a vital component of data governance, allowing organizations to trace and document the origins and history of data from their sources to their current form. This process enables organizations to understand how and where data is used and stored, ensuring compliance with necessary regulations. Furthermore, data lineage is essential for data quality assurance, verifying that data entering the system is validated, cleansed, and transformed correctly—data lineage can inform business intelligence by providing insight into data relationships and dependencies.

These benefits are especially pertinent in industries such as healthcare, finance, and government, where compliance and accuracy of data are of utmost importance. Organizations can ensure that data provenance is maintained through data lineage, enabling transparency and compliance.

5) Discoverability

Data discoverability refers to the ease with which data is found, accessed, used, and understood. It is critical to data liquidity. It can enable efficient data reuse and facilitate the development and performance of new/existing models. Tokenization can improve discoverability as a result of the following:

- *Metadata*- Creating and maintaining detailed metadata for each data set, including information about the data source, format, quality, and other relevant characteristics that might affect its use.
- *Data catalog*- Creating data token catalogs can allow users to search for and discover data sets based on specific criteria, such as data type, source, or application area.
- *Interoperability*- Using common data formats, protocols, and standards to ensure that data is easily understood and used by different systems
- *Data annotation*- Annotating data (crowd work) with information can help potential consumers to understand the context of the data and how they are potentially used
- *Data provenance*- The provenance (history, origin, and processing) of data can improve the understanding of data quality and data lineage for decision making

A key aspect often overlooked is *price-discoverability*. Namely, how valuable is a dataset, and what is its price? Data buyers often want to know the cost of the data they are looking to buy, and sellers want to know the price they can sell a dataset—exemplars in fair asset valuation[16] and pricing can be borrowed from financial services to price this value. It is also possible to create marketplaces from data catalogs where buyers and sellers can access a transparent, automated, and liquid marketplace for efficient and effective discovery of data prices. Offering market-driven price discoverability can also be made possible with data exchanges.

6) Interoperability

Data interoperability refers to the ability of different systems, platforms, or applications to exchange and use data meaningfully. Interoperability enables data to be exchanged and reused across different systems without requiring manual intervention or special effort. Most literature considers interoperability as just technical interoperability. We additionally introduce the notion of jurisdictional interoperability and technical interoperability. Due to their abstract and instrumental nature, tokens are an appliance for interoperability.

Technical interoperability with tokenization is achieved by utilizing common data formats and standards in the data component. This helps to ensure that data is easily exchanged and understood by different systems. Furthermore, a shared data model and ontology enable easy linkage and integration of data throughout various systems. [90].

Jurisdictional interoperability concerns the use of data

[16] Sec. V-C7 for Data as an Asset.

across multiple jurisdictions. With the evolution of disciplines such as Data Sovereignty [91] and Data Regulation, jurisdictional interoperability of data is becoming critical. With governance (Sec. V-C2), sovereignty restrictions and obligations are maintained and enforced, facilitating jurisdictional interoperability.

Blockchains also enable protocol interoperability—the ability for tokens and protocols to *communicate* with each other. Tokenization standards such as the ERC1155 [92] enable tokens to be used with other protocols, tools, and blockchains. Combining these protocols can create a class of powerful instruments and derivatives. For example, it is possible to collateralize data-tokens using **DeFi** protocols. Protocol interoperability can also be used to facilitate incentivization.

Interoperability is important for data liquidity, as it allows data to be shared and reused across different systems, platforms, or applications. In ML, this can help to improve the performance of AI models and make it easier to upgrade and deploy new models in the field.

7) Data Assets & Monetization

Tokenization intrinsically links value to data and presents data as a first-class asset[17]. This abstraction offers the ability to create an asset class backed by data [93] that can be priced and monetized using concepts from the financial services domain, such as those included in the International Swaps and Derivatives Association (ISDA) agreement (Sec. V-C2). Furthermore, other implications include the ability to consider and account for data as an intangible asset within balance sheets.

Better interoperability, discoverability, provenance, and governance reduce the barriers and challenges, making it more readily monetizable than current methods. It also makes trading data more convenient as the actual data is not traded but rather is an instrument.

8) Incentives Alignment

When combined with blockchain technology, tokenization offers the potential for cryptoeconomic *Mechanism Design*[18] for managing the data value cycle (Sec. IV-D) [95]. This phenomenon, which we refer to as *Data Tokenomics*, can use tokens to combine usage and ownership with rewards to increase data quality and liquidity. Some of these include:

- *Token rewards*- token-based rewards to incentivize crowd work such as the labeling or curation of data, e.g., through staking [96]
- *Social tokens*- social tokens represent a person or brand's identity or reputation within a community. In Data Tokenization, they are used to represent a curator's reputation such that they may earn social tokens for contributing datasets or participating in discussions, which can help to build their reputation and influence
- *Tokenized access*- tokens to grant access to specific resources, e.g., users holding several tokens to access certain datasets
- *Tokenized governance*- tokens are used to give users a say in governance. For example, users may be able to vote on proposals or changes to the curated dataset(s), with their voting power proportional to the number of tokens they hold, if in a Data DAO or Data Fund (Sec. VII-B1)
- *Bounty program*- programs that offer users a number of tokens to complete specific tasks, e.g., finding new datasets or annotating new datasets

Secondary Markets [97] can also be built on top of the token component to provide additional liquidity.

In section III, we also discussed insufficient mechanisms and asymmetries in risk/reward allocation across the value cycle, as data owners/curators often bear the most risk in making data available as opposed to consumers who accrue most of the rewards. Integration of the token component with traditional services or DeFi protocols can provide additional mechanisms to allocate these risks and rewards better. For example, an escrow service is created to mitigate the risk of data misused[19] or creating insurance services for the use of data.

The network effects of the inherent P2P infrastructure and protocol interoperability can also amplify these incentives and risk allocation mechanisms.

9) Regulation & Compliance

Data regulation, such as CCPA[20] and GDPR, involves a set of laws, rules, and guidelines that aim to govern the collection, usage, storage, and sharing of data. English data protection law still largely follows European GDPR. While there is no political consensus, currently there are policy debates around the question of whether post-Brexit deviation from European rules can benefit deep tech firms in the UK while also preserving customer rights to privacy in an appropriate way. Data regulation applies to sector-specific regulations, such as those related to healthcare and financial data, as well as general-purpose data protection laws.

Tokenization presents an alternative approach in which

[17] Recall the definition of asset [1].

[18] *Mechanism Design* is a field in economics & game theory that takes an objectives-first approach to design economic mechanisms or incentives, toward desired objectives (data liquidity in our case), in strategic settings where players act rationally [94]. It is often termed *the reverse game theory*

[19] More relevant in the case of **Mode 1** and **Mode 2** tokenization

[20] California Consumer Privacy Act

the regulatory emphasis is shifted from the data to the instrument—and, therefore, the associated data. It has the potential to better scale and enforce regulation and compliance across multiple purposes and sectors due to the use of *contextual regulation*, where different regulations are applied to a single instrument depending on the context.

Demonstrating compliance is a key part of regulation. It concerns the possession, organization, storage, and management of data to prevent it from loss, theft, misuse, or compromise. Tokenization can be used to facilitate this assumption [98]. Compliance is a function of Provenance (Sec. V-C3) and Governance (Sec. V-C2). Governance stipulates the regulations and standards determining what data must be protected and the most suitable processes[21]. Provenance can keep track of how, when and to whom these measures were carried out.

Additionally, using token instruments allows for sector-specific regulations to be implemented within sector-specific networks (such as consortium networks/ private permissioned networks), as discussed in section V-C1, which outlines an example of such a network regarding healthcare data.

VI. Benefits

Data Tokenization has the potential to produce advantages over current data architectures, relationships, and uses, such as:

- *Permissioning-* Granular mechanisms provide individuals and businesses with clear control over how their data is used and by whom. Furthermore, the underlying infrastructure and provenance provide additional assurance that the data is only used in accordance with the owners' or controllers' agreements. Attaching permissions to multiple purposes further solidifies the trustworthiness of such transactions.
- *Reuse-* Tokenization allows data to be reused across various purposes and treat data differently depending on its actual and anticipated use, creating the commercial imperative for sharing data.
- *Protection-* Competing interests associated with data consumption are often overlooked with the *privacy by design* paradigm [99]. These imbalances adversely impact privacy, as data consumption (utility maximization) and privacy are antithetical. Tokenization facilitates the creation and embedding of incentives within the instrument that can reduce these asymmetries. Additionally, cryptography, identity, and provenance are core building blocks of tokenization and can promote better data security and transparency.
- *Adaptable Data Processing-* With **Mode 3**, tokenization can stay at the moment and point of collection or within the original collection system; the analysis of data and their use in the processing are performed in a privacy-preserving manner.
- *Provenance-* The entire dataset's biography is captured with the token instrument and its underlying immutable smart contract infrastructure enabling transparency, trust, and verifiability.
- *Portability-* A single tokenized dataset can be used for multiple purposes as tokenization decouples source from purpose.
- *Governance-* data is optimized over time for the different use cases, and governance rules between different data relationships are implemented and enforced through smart contracts, providing the adaptability required to regulate data utilization across numerous cases effectively.
- *Economic Value-* The ability to attach economic value directly to data [22] can provide better mechanisms to allocate risks and rewards across the use of data. For example, it can be possible to underwrite insurance for the use of data.
- *Data as a CapEx-* The concept of treating data as a capital expenditure is becoming increasingly relevant in industries such as finance and technology, where data is often a critical asset for business success. The data tokenization and valuation framework assign economic value to data and makes it possible to consider data a form of capital expenditure (CapEx) for businesses. For example, a company may invest in acquiring or generating data expected to provide future economic benefits. This investment can be recorded on the balance sheet as an intangible asset (Sec. V-C7) that could be treated as a form of CapEx.
- *Competition-* Tokenization can enable data liquidity: collaborative data access, data marketplaces, Trusted Research Environments, and Data Institutions can all enable data-driven innovation.

VII. Applications

In the previous sections, we described the challenges associated with current data infrastructures and how they inhibit data liquidity. This section introduces some applications that can be built through data tokenization. We have seen how data tokenization is a foundational technology for the data economy; it can advance current data infrastructures/relationships or build radically new and different data infrastructures that are potentially more equitable, open, trusted, and fair. We also stress that the potential applications of data tokenization and the decentralized economy are vast, spanning across industries such as finance, healthcare, government, supply chain, and data-intensive sectors in the context of both data

[21] Encoded with smart contracts in the case of tokenized data

[22] recall that tokenization creates a first-class asset out of data section III

mesh and data fabric paradigms. These organizations want to improve data privacy, security, and ownership while striving for ethical profitability and sustainability. They are exploring the potential benefits of using blockchain and other distributed ledger technologies to facilitate secure and efficient data management and analysis in a decentralized environment, which can lead to improved trust, transparency, and accountability.

To facilitate our readers' understanding, we will briefly describe each application and explain how data tokenization can enable them. Our goal is to make the reader realize that each of these applications requires governance, value-based incentives, standardization, transparency, or a combination and that a permutation of the three basic components (Data, Token, and Compute)[23] make them possible, as presented in Fig. 8. Similarly, we class potential applications based on what the Token component is mainly used for 1) governance-led, 2) incentives-led, 3) transparency-led 4) composites. Trust is induced through these and the underlying tokenization infrastructure.

A key aspect is that one data token could be used across several of these application topologies simultaneously. We believe the true potential of data tokenization is in the synergy across all these topologies. For example, the use of Crowd Work (Sec. VII-A5), with Data Cooperatives (Sec. VII-B1), or for creating DataDAOs.

We only scratch the surface; full technical & operational details and their wider implications are outside the scope of this paper. Nevertheless, we strongly encourage our readers to explore the potential of data tokenization and its applications in the subsequent parts discussed in this section. We begin by discussing some governance-led applications [24] followed by incentives-led [25] and finally composites [26].

A. *Today*

1) *Multiparty Data Access*

Multiparty data access (MPDA) refers to the ability of multiple parties or stakeholders to access and use data for a common purpose. This can include situations where different organizations or departments must collaborate and share data to achieve a common goal.

Multiparty data access is particularly valuable when organizations/ stakeholders across the value cycle have complementary data assets or expertise. Still, they may not achieve their goals without access to each other's data. Organizations can gain new insights, develop more accurate models, and make better decisions by collaborating and sharing data.

However, there are significant challenges associated with multiparty data access, including data privacy, security, legal agreements, and ownership. To facilitate responsible multiparty data access, clear policies and governance frameworks that protect the interests of all parties involved and their technological implementations are paramount. Tokenization can address these points.

With tokenization, access control is done with tokens and the underlying smart contract infrastructure, as discussed in Sec. V-C2, which could be further augmented by directly implementing tokens as smart legal data access agreements (also further discussed in Sec. V-C2). **Mode 2** and **Mode 3** tokenization (Sec. V) is the recommended mechanism. Transparency arising from Provenance (Sec. V-C3) also promotes trustworthiness in the infrastructure and across counterparties.

With respect to Fig. 8.1, the following variants of MPDA can be implemented with tokenization:

- *Multi-Owner*- involves one Data Accessor being able to access multiple different datasets that could be from multiple different data owners. This could be with a Compute custodian (**Mode 2** tokenization) or self-compute (**Mode 3** tokenization)—Diagram (8.1.I) illustrates **Mode 2** multi-owner access.
- *Multi-Accessor*- involves one dataset being accessed by multiple accessors. This could be with a Compute custodian (**Mode 2** tokenization) or self-compute (**Mode 3** tokenization)—Diagram (8.1.II) illustrates **Mode 3** multi-access.
- *Federated-Access*- involves multiple accessors accessing datasets owned by multiple owners as a combination of the previous two—Diagram (8.1.III) illustrates Federated access.

As evidenced by proposals such as the EU Data Governance Act [2] and the EU Data Act [3], we anticipate that regulation and legislation will be the primary drivers for MPDA systems.

2) *Federated Learning*

Federated Learning (FL) involves training a model across distributed data sets to prevent data leakage [100]. Instead of centralizing the data on a single server or entity, the data remains distributed across multiple edge devices and is processed locally [101]. The local updates are then communicated to a coordinating entity. Still, the approach is to keep the data decentralized and only share aggregated model updates, avoiding the centralization of raw data. Tokenization can be used to establish governance and orchestration of Compute and training (and retraining

[23] Presented in Fig. 4 and further explained in Sec. V

[24] Multi-Party Data Access, Federated Learning, and Data Mesh

[25] Crowd Work

[26] Data Exchanges, Trusted Data Infrastructures, Data Cooperatives, Data Trusts, DataDAOs, and Data Dignity

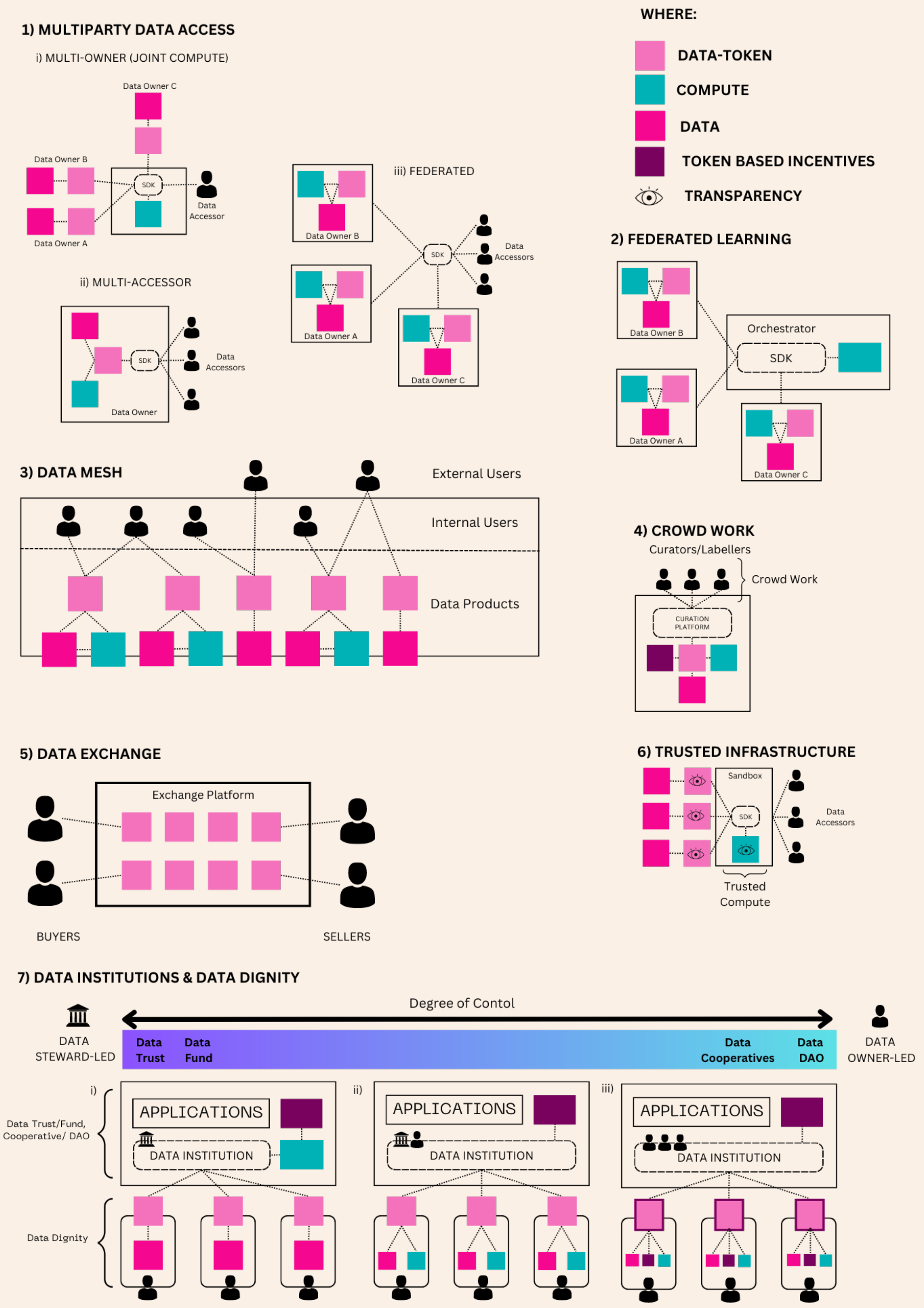

Fig. 8. Application Topologies: (1) Multiparty Data Access, (2) Federated Learning, (3) Data Mesh, (4) Crowd Work, (5) Data Exchange, (6) Trusted Infrastructure, (7) Data Institutions & Data Dignity

for continuous model updates). FL is an application of Federated Multi-Party Data Access, where the Compute is for training models locally.

There are two variants of aggregation: centralized and decentralized. Centralized aggregation requires a centralized entity to perform the aggregation of the weights, and decentralized aggregation makes use of mechanisms such as round-robin for aggregation.

Tokenization can also provide incentives which, combined with mechanism design, can also help mitigate risks such as Byzantine failures & malicious participants and contribute to data [102].

3) Data Mesh Architectures

Data Mesh is a relatively new architectural paradigm for managing data in large, complex organizations. It involves breaking down a monolithic data architecture into smaller, decentralized, domain-specific data platforms, each responsible for a particular set of data products [63]. In doing so, the goal of the approach is to promote data ownership, autonomy, and decentralization. This has the potential to enable teams to manage and scale their data products more effectively.

Data Mesh is a decentralized data management approach that emphasizes *domain-driven design*, *data-as-a-product*, *self-service data platforms*, *federated data governance*, and a culture of data collaboration. The main goal of Data Mesh is to enable organizations to elevate data to a first-class artifact promoting data quality, availability, and accessibility.

However, implementing and managing governance is challenging due to the decentralized nature of the architecture [103]. Robust governance mechanisms are required to facilitate collaboration and data management internally and externally[27]. Adopting these principles requires significant investment in infrastructure, tools, and processes aligned with business objectives to ensure maximum ROI. Overall, Data Mesh provides benefits that can be realized with the right governance mechanisms in place.

By tokenizing each of these data products, governance is managed through the token, as seen in Fig. 8.3. This is just a higher-level application using federated multi-party data access discussed in section VII-A1.

4) Data Fabric Architectures

Data Fabric is a data integration approach that enables organizations to connect and combine data from various sources and formats to provide a unified data view. Data fabric can be deemed as the technology part of data mesh. It emphasizes data virtualization, metadata management, and API-based integration to create a unified data fabric that stakeholders can access and consume [104].

The critical difference between Data Mesh and Data Fabric is that Data Mesh focuses on managing data as a valuable asset through decentralized data ownership, while Data Fabric emphasizes data integration and virtualization. Data Mesh promotes a culture of data collaboration across organizational silos, while Data Fabric seeks to create a unified fabric of data that can be accessed and consumed by stakeholders through a unified data layer. Both Data Mesh and Data Fabric aim to improve data quality, accessibility, and availability but use different approaches to achieve this goal [105].

A hybrid environment that combines the strengths of Data Mesh and Data Fabric could provide significant benefits for organizations [106]. For example, Data Mesh's decentralized architecture can make governance and data management challenging, but Data Fabric's data integration capabilities can help create a unified view of data across domains. Similarly, Data Fabric can sometimes struggle with data ownership and regulatory compliance, but Data Mesh's federated data governance can help ensure compliance and ownership.

Ultimately, the success of a hybrid environment depends on how well the organization can manage and govern data effectively across domains and the centralized layer, as well as how well stakeholders can access and consume data through a unified data layer. Even though technology plays a crucial role in data fabric, effective governance and management are critical to realizing the benefits of a hybrid environment that combines the strengths of both approaches.

5) Crowd Work

Crowd work is central in providing auxiliary services such as annotation and labeling datasets. The goal is to improve data quality through crowd work before its use for training. Incentive structures created with tokenization can be used to facilitate this. In recent times, *ghost work* has become a concern [107]. Tokenization and its use with the Web3 ecosystem can also potentially provide fairer and more equitable remuneration models to address this issue.

Data-Centric AI

The data-centric AI approach [43] is based on the idea that ML systems are built around the data used to make predictions and decisions more than around a model. Typically, ML systems rely on data to train models, and the quality and diversity of data directly impact the accuracy and performance of the system deployed. By focusing on the

[27] For example, a data product may need to be used internally across two different departments within an organization or externally between an organization and a client or across borders.

careful curation (e.g., through labeling and annotation) of data that relies on crowd work, it is possible to create better-performing models. Network effects, incentives (Sec. V-C8), and crowd work (Sec. VII-A5) can reinforce this paradigm.

Curation of Foundational Datasets

In the context of AI, foundational datasets and corpora to train machine learning models are typically large and of high quality (ground truth). These datasets may include thousands or millions of labeled data points to train models to perform specific tasks, such as image recognition, natural language processing, or predictive analytics.

Foundational datasets/Corpora in AI are essential for creating accurate and effective Machine Learning (ML) and Deep learning (DL) models. DL architectures have been developed to process raw structured data and facilitate rapid analyses of structured inputs, such as sequences, images, and videos, to predict complex outcomes with unprecedented accuracy and to generalize to new unseen data. ML/DL models may not be accurate or perform as well as expected for real-world data without high-quality foundational datasets.

Some examples of foundational datasets in AI include: *LAION-5B* [108] used to train Stable-Diffusion [109]; *ImageNet* [110], a dataset of millions of labeled images commonly used for image recognition tasks and the Common Crawl [111], a large dataset of web pages used for training natural language processing models.

The development of AI relies heavily upon the creation and curation of foundational datasets, which can be both time-consuming and resource-intensive. In response, many organizations, including those belonging to the academic, charitable, and corporate domains, are investing in producing high-quality foundational datasets employed across various applications. However, recent advancements in Large Language Models (LLMs), alongside other generative models which consume large amounts of data, have raised significant ethical considerations regarding proper attribution, usage, and remuneration. Notable cases include *Chat-GPT's* purported use of News Corp data [112], as well as Getty's role in Stable-Diffusion [109], [113].

Tokenization can offer a more transparent and equitable means to address some of these pressing challenges and foster the creation of even more foundational datasets. In particular,

- *fractional ownership*- fractionalized ownership with tokenization (Secs. V-C2, V-C7, V-C8) can facilitate the governance and equitable revenue share from the use of a dataset [28]
- *tokenized curation*- incentives (Sec. V-C8) with tokenization are used to curate more and better quality datasets, examples include Genomes IO [114]

6) Data Exchanges

Data markets and exchanges are platforms that allow individuals and organizations to buy and sell data [115]. They provide a way for data producers, such as companies and individuals, to monetize their data by selling to data consumers, such as researchers, businesses, and government agencies. There are multiple kinds of data exchanges [116]–[118]. Our critical assessment reveals that for an exchange to function optimally, three essential components are necessary[29]:

– Discoverability
– Liquidity
– Fulfillment

Much like financial markets, trading is carried out through the token instrument with tokenization. Marketplaces/ exchanges built for tokenized trading trade the instrument (i.e., the ownership and usage rights for the data) rather than the data itself, as illustrated in Fig.8. Liquidity, Discoverability, and Fulfillment then become a function of this instrument.

Section V-C5 describes how discoverability is achieved with tokenization; the exchange lists tokens rather than data. Liquidity can be induced through pricing, network effects, and aspects such as Crowd Work associated with tokenization. The decoupling of the source from the purpose also has implications for Liquidity. Similarly, fulfillment refers to how the data is consumed/used post-trade; the three modes of data tokenization (Fig. 7) facilitate a seamless achievement responsibly. Custodial Compute (Sec. V-B) services could be perceived as a *data clearing houses* [119] for the data exchanges. It is also possible to create derivatives and other exchange-traded instruments and secondary markets based on tokenized data—these have implications on liquidity and de-risking/ risk allocation in the use of data.

Token-based provenance (Sec. V-C3) and compliance (Sec. V-C9) can also be used for addressing regulation. In fact, with proposals such as the EU Data Governance Act [2] and the EU Data Act [3], we anticipate regulation and legislation as the key driver for the proliferation of data exchanges and marketplaces.

Data markets and Exchanges can be centralized (a single company, intermediary, or organization operates the

[28] See also Data Trusts, DAOs, and Dignity; Sec. VII-B1, VII-B2

[29] Once again, this is a brief explanation; a more in detail discussion will be provided as a supplement to this paper

platform) or decentralized (a network of users operates the platform.), such as a data exchange operated by its owners, a data cooperative, or DAO.

Open Data Exchanges

Open Data is data that anyone can access, use or share [120]. A growing number of proposals and legislation advocating Open Data Platforms (of public-sector data) [121]. In essence, these are the same as Tokenized Data Exchanges described above, with the element of monetization removed. **Mode 2** and **Mode 3** tokenization can also facilitate the use of private data.

7) Trusted Data Infrastructures

Trusted data infrastructures are an architecture or ecosystem of technologies that foster trustworthiness and responsible use of data through policies and practices. This ensures that data is reliable and secure and that its use/access is responsibly managed. Trusted Data Infrastructures/Ecosystems are typically composed of a collection of data management platforms, data integration tools, and data governance frameworks, as well as policies and procedures that govern how data is collected, stored, shared, and used. Similar to the use of tokenization for MPDA and Data Mesh, tokenization can be used to implement trusted data infrastructures (Fig. 8.6). A common kind of trusted data infrastructure is the Trusted Research Environment (TRE)[30].

Trusted Research Environments

A Trusted Research Environment (TRE) is a collection of datasets attached to a compute environment[31] that can be accessed securely and remotely by approved researchers [122]. TREs have enhanced security measures to ensure only accredited researchers can gain access, oversight measures to track research activities and purposes, and measures that ensure the data cannot be exported from the environment.

Governance, Transparency, and Privacy are key faculties of an TRE. Tokenization enables TREs as seen in Fig.8.6. **Mode 2** tokenization with Compute aspects such as Confidential Compute (Sec. V-B2) and PETs can enhance privacy. Provenance (Sec.V-C3) enhances transparency together with the underlying blockchain infrastructure.

TREs are used in various fields, including healthcare [123], social sciences, and market research, and enable researchers to analyze and derive insights from sensitive data without compromising the privacy or confidentiality of the individuals or organizations from whom the data was collected.

B. In the near future: Data Ecosystems

Data Ecosystems are sub-elements of a more comprehensive ownership-led data economy. They comprise people, communities, and organizations that create, curate, steward, and monetize their data. Data Tokenization is foundational for these data ecosystems [124]. These data ecosystems, which span public and private sectors and encompass the end-to-end view of data value cycles, are expected to play a pivotal role in the accelerated evolution of a data economy [26]. Two key tenets of Data Ecosystems are *Data Institutions* and *Data Dignity*, which this paper will not discuss. Instead, a brief overview and discussion of how Data Tokenization enables them are provided.

1) Data Institutions

A *Data Institution* is a broad term used to refer to a technical, legal, and financial structure designed to manage data for or on behalf of its data subjects (the owners of the data) to achieve specific financial, social, or public benefits. Examples of data subjects include individuals, IoT devices, or organizations.

The primary purpose of a Data Institution is to be a vehicle for managing and monetizing data in a way that is transparent, accountable, and often fair for the data subjects [26] while balancing the conflicting interests of data privacy and data utility. For example, a Data Institution can provide access to sensitive personal data for research purposes while ensuring that the data is protected and that data subjects' rights are respected—*rights-preserving data access*.

The utilization of Data Governance (i.e., *Ownership*, *Usage* and *Control*) is essential for Data Institutions to ensure robust and transparent stewardship of the usage and monetization of data. This is particularly beneficial in the case of data tokenization, as illustrated in Fig. 8.7. By tokenizing data subjects' data, Data Institutions can use tokens as an instrument for stewardship and control of data usage and monetization (further described in Sec. V-C2). Provenance (Sec. V-C3), as a result of tokenization, can also offer transparency to data subjects about how their data has been used by the institution.

There can be multiple kinds of institutions based on how and to what degree the data subjects' data is governed: Steward-Led and Owner-Led (Fig. 9). Owner-Led Institutions are more independent in that the data subjects actively decide how, where, and when their data can be used/monetized. In contrast, Steward-Led Institutions give the steward more control.

Fig. 9 has been overlaid on top of Fig. 8.7 to show how tokenization can enable these different kinds of Data Institutions across the spectrum. Examples of Data Insti-

[30] They are also referred to as *Data Safe Havens* (DSH) or *Secure Data Environments* (SDE)

[31] With tooling such as IDEs

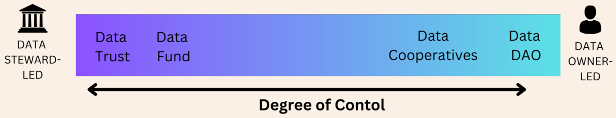

Fig. 9. Kinds of Data Institutions

tutions include:

1) *Data Trusts*- a *Data Trust* is a steward-led legal and technical structure designed to manage data for a specific group or community, aiming to achieve specific social or public benefits [15]. Data trusts are independent third-party organizations governed by a board of trustees that have the *fiduciary responsibility*[32] for the management and use of their data subjects' data for social good or profit. With tokenization, the trust can more effectively steward its subjects' data through the token as presented in Fig. 8.7.i. For example, individuals could *donate* for data altruism their healthcare data to a data trust whose objective could be to use those data only for research purposes.
2) *Data Funds*- a *Data Fund* is similar to a Data Trust. Still, its key objective (much like its investment fund counterpart) is profit for its data subjects (i.e., the *data investors*. For example, individuals could *invest* their retail data in a data fund. The data fund's objective could be to sell insights or ML models derived from this data and the resulting profit to be shared across the pool of data investors in the form of a *Data Dividend*.
3) *Data Cooperatives*- a *Data Cooperative* is similar in objectives to a Data Trust or Data Fund but is critically owner-led. It is a legal, technical, and financial structure designed to manage data for and by its members (i.e., the data subjects). Much like their agricultural or industrial counterparts, members of the cooperative collectively steward the data for their members' collective benefit. Tokenization can enable these as well through aspects such as voting—described in Sec. V-C2[33].
4) *Data DAOs*- *Data Decentralized Autonomous Organizations* are Data Institutions native to Tokenization. They are similar to cooperatives, but the entire socio-legal and financial structure is implemented with tokens and tokenization; members vote/ manage governance with their tokens and earn dividends in tokens. Value-based incentives (Sec. V-C8), implemented with tokens, become critical in enabling them—Fig. 8.7.iii presents the topology for DataDAOs.

2) *Data Dignity*

Data Dignity refers to the principle that individuals should have control over their personal data and that this data should be treated with respect and used in ways consistent with their values and interests [17]. Data Dignity naturally lends to data tokenization as with the token. Individuals can create a *data passport* of their personal data to control how, why, where, and for what purpose their data is used with the appropriate data security and transparency. For example, individuals could pool their banking, retail, healthcare, etc. data onto their devices, tokenize it and then appropriately monetize it—Fig. 8.7 illustrates tokenization and its use for Data Dignity with Data Institutions.

The foundation of data dignity is based on the idea that personal data is not just a commodity to be bought and sold but rather an extension of individuals' identities and personal experiences [125]. Data dignity is closely related to data sovereignty, which is the idea that individuals can control how their personal data is collected, used, and shared. It is also closely linked to data privacy principles (the right to control access to personal data) and data security (the protection of personal data from unauthorized access or misuse).

More recently, with the advent of generative AI, data dignity has become a major concern as large amounts of data (including personal, copyrighted, or licensed) are collected and used to train and operate AI models. Ensuring data dignity in this context requires mechanisms to ensure that individuals know how their data is being used & monetized and that they can control access to their data, of which tokenization is a natural enabler.

Data dignity is also closely tied to data ethics, which studies the moral and ethical implications of data collection, usage, and sharing. This includes ensuring that data is used in a way that is fair, equitable, and non-discriminatory and that it respects individuals' rights and dignity.

VIII. Case Study: The Valyu.network

Data tokenization is a complex but promising area of research and development that has the potential to revolutionize how organizations and individuals responsibly monetize, exchange, curate, and use data. The authors of this paper have developed the **Valyu Framework** to challenge the traditional approach to create and monetize data-centric assets such as data, models, and their derivatives for Web 3.0, emphasizing security and equity. The **Valyu** framework is an instance of the *Data Tokenization and Valuation Framework* discussed in this paper. It aims to prove that the ideas and concepts related to data tokenization are technically sound and commercially

[32] A fiduciary duty is considerably more onerous and requires considerations in *duties of care, loyalty, good faith, confidentiality, prudence, etc.*.

[33] Quadratic Voting (Sec. V-C2) can provide a more equitable voting mechanism for members of the cooperative

viable. This endeavor has been actively pursued through **Valyu.Network**[34], through which several PoCs and applications incorporating data tokenization have been developed:

- *Labelling Application*- **Valyu** has built a data labeling (crowd-work) application, currently being trialed in Kenya, to demonstrate the use of value-based incentives and tokenization for data quality
- *Data Exchange*- a data exchange to buy and sell datasets.
- *Data Fund*- a proof of concept Data Fund that allows users to monetize their healthcare data for research purposes. Users receive a Data Dividend for making their data available.
- *Data Bounty Platform*- A data curation platform that pays users to curate quality datasets

Additionally, **Valyu** has also developed the following research outputs and tooling:

- *A Formal Specification*- different applications will require implementations in private or permissioned blockchains. A formal specification of the *Data Tokenization and Valuation* framework described in this paper shall facilitate this objective. It will be released as an addendum to this position paper.
- *Reference Implementations*- reference implementations on the *Polygon*[35] and *Hedera*[36] networks
- *SDK and Tooling*- for building applications. **Valyu** currently has Crowd Work, Monetization, and Governance-related components.
- Pricing/Valuation Engine- for pricing data and facilitating price discovery

The decentralization of cryptocurrency markets and the transformation of foreign exchange (FX) markets from a traditional centralized structure have been major developments in the recent Fintech sector. In this context, the **Valyu** framework does not concern itself with currencies but instead focuses on a secure, decentralized data ecosystem.

IX. Conclusions And Future Work

The concept of data tokenization to create a decentralized marketplace for data is relatively new. We have described in this paper a general framework to address some of the multifactor aspects of data tokenization (technological, legal, social, economic) exploiting blockchain/distributed ledger technology (DLT) to create a decentralized marketplace for data while providing regulation and governance. We have described our decentralized data economy vision based on tokenization, which can disrupt the traditional data market and create new opportunities with several benefits. Among them, we see the creation of new business models to allow data providers to monetize data and data consumers to access data more efficiently at a lower cost. It can lead to new revenue flows for companies and individuals and drive innovation by making data more readily available to researchers and developers. Another benefit is the improvement of data liquidity and interoperability through which data is easily exchanged and shared among different parties for more efficient use of data resources and better decision-making. Data tokenization and smart contracts can improve data governance and control data circulation. Overall, we aim to foster a culture of data collaboration that enhances sustainability, equity, and profitability for organizations. Collaborative analysis of sustainability data can identify opportunities to reduce waste, improve energy efficiency, and mitigate environmental risks. Encouraging knowledge sharing among stakeholders can unlock the full potential of organizational data assets, leading to better decision-making, increased innovation, and more efficient operations. By fostering a culture of data collaboration, organizations can address environmental, social, and governance (ESG) challenges more effectively and improve overall performance, contributing to a more sustainable future.

As the technology and regulations surrounding data continue to evolve, these boundaries are liable to change, but at the time of writing, they are conventional and supported by regulatory opinions in England, and it is possible to predict the regulatory treatment of tokens. We nevertheless take account of this in the development of the framework. We are excited by the immense possibilities that this breakthrough offers and are confident that data tokenization will become a widely-used mechanism for the usage, valuation, and management of data assets, granting components of the decentralized data economy increased transparency, safety, and fairness. As part of our ongoing **Valyu** roadmap (Sec. VIII), we are committed to investigating the potential of data tokenization and exploring new ideas and opportunities to ensure that ethical considerations remain at the forefront of our research while utilizing the full potential of the decentralized data economy.

Acknowledgements

We would like to thank Adriano Koshiyama, Emre Kazim, Umar Mohammed, Holistic AI, Tu Tuan Ngo, Nigel Kingsman, Christopher Huntington and Ronnie Stafford for their constructive feedback and support.

∇ 🛠️ WE BUILD | JOIN US

108

References

[1] I.-A. Barone, “What is an asset? definition, types, and examples.” [Online]. Available: https://www.investopedia.com/terms/a/asset.asp

[34] The leading authors are part of the **Valyu** Network.

[35] polygon.technology

[36] hedera.com

[2] European Union, “Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act).” [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52020PC0767

[3] ——, “Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act).” [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN

[4] Congress.gov, “H.r.7120 - 117th congress (2021-2022): Dashboard act of 2022 | congress.gov | library of congress.” [Online]. Available: https://www.congress.gov/bill/117th-congress/house-bill/7120?s=1&r=23

[5] F. Xiong, M. Xie, L. Zhao, C. Li, and X. Fan, “Recognition and evaluation of data as intangible assets,” *SAGE Open*, vol. 12, no. 2, p. 21582440221094600, 2022. [Online]. Available: https://doi.org/10.1177/21582440221094600

[6] P. Voigt and A. Von dem Bussche, “The EU general data protection regulation (GDPR),” *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[7] World Economic Forum, “Data-driven Economies: Foundations for Our Common Future,” 2021. [Online]. Available: https://www3.weforum.org/docs/WEF_WP_DCPI_2021.pdf

[8] S.-J. Moon, S.-B. Kang, and B.-J. Park, “A Study on a Distributed Data Fabric-based Platform in a Multi-Cloud Environment,” *International Journal of Advanced Culture Technology (IJACT)*, vol. 9, no. 3, pp. 321–326, 2021.

[9] E. Eryurek, U. Gilad, V. Lakshmanan, A. Kibunguchy-Grant, and J. Ashdown, *Data Governance: The Definitive Guide*. O`Reilly Media, Inc., Mar 2021.

[10] N. Bhansali, *Data Governance: Creating Value from Information Assets*. CRC Press, Jun 2013.

[11] Law Commission of England & Wales, “Digital assets - law commission,” (Accessed on 04/15/2023). [Online]. Available: https://www.lawcom.gov.uk/project/digital-assets/

[12] J. Rodriguez and B. Wixom, “Increase data liquidity by building digital data assets,” *MIT Center For Information System Research*, Nov 2021.

[13] A. P. Abernethy, J. L. Wheeler, P. K. Courtney, and F. J. Keefe, “Supporting implementation of evidence-based behavioral interventions: the role of data liquidity in facilitating translational behavioral medicine,” *Translational behavioral medicine*, vol. 1, no. 1, Mar 2011.

[14] S. Roy, A. R. Shovon, and M. Whaiduzzaman, “Combined approach of tokenization and mining to secure and optimize big data in cloud storage,” in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 83–88.

[15] S. Mills, “Who owns the future? data trusts, data commons, and the future of data ownership,” *Data Trusts, Data Commons, and the Future of Data Ownership (September 24, 2019)*, 2019.

[16] J. Hardinges, “Data trusts in 2020.” [Online]. Available: https://theodi.org/article/data-trusts-in-2020

[17] J. Lanier, “A blueprint for a better digital society.” [Online]. Available: https://hbr.org/2018/09/a-blueprint-for-a-better-digital-society

[18] “Data-centric AI - what is Data-Centric AI & why does it matter?” [Online]. Available: https://landing.ai/data-centric-ai/

[19] M. Motamedi, N. Sakharnykh, and T. Kaldewey, “A data-centric approach for training deep neural networks with less data,” 2021. [Online]. Available: https://arxiv.org/abs/2110.03613

[20] K. Werder, B. Ramesh, and R. Zhang, “Establishing data provenance for responsible artificial intelligence systems,” *ACM Transactions on Management Information Systems (TMIS)*, vol. 13, no. 2, pp. 1–23, 2022.

[21] E. Raguseo, P. F., and C. Vitari, “Streams of digital data and competitive advantage: The mediation effects of process efficiency and product effectiveness,” *Information & Management*, vol. 58, no. 4, p. 103451, 2021. [Online]. Available: https://hal.science/hal-03323663

[22] N. Srnicek, *Platform capitalism*. Polity, Jan 2017.

[23] F. Tariq, Z. Khan, T. Sultana, M. Rehman, Q. Shahzad, and N. Javaid, *Leveraging Fine-Grained Access Control in Blockchain-Based Healthcare System*. Springer, 03 2020, pp. 106–115.

[24] M. Banerjee, J. Lee, and K.-K. R. Choo, “A blockchain future for internet of things security: a position paper,” *Digital Communications and Networks*, vol. 4, no. 3, pp. 149–160, 2018.

[25] J. Sadowski, “When data is capital: Datafication, accumulation, and extraction,” *Big Data & Society*, vol. 6, 2019. [Online]. Available: https://doi.org/10.1177/2053951718820549

[26] C. for Digital Ethics and Innovation, “Unlocking the value of data: Exploring the role of data intermediaries.” [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1004925/Data_intermediaries_-_accessible_version.pdf

[27] O. O’Neill and J. Bardrick, “Trust, trustworthiness and transparency,” *Brussels: European Foundation Centre*, 2015.

[28] B. Vuleta, “How much data is created every day? +27 staggering stats,” Oct 2021. [Online]. Available: https://seedscientific.com/how-much-data-is-created-every-day/

[29] M. Younas, “Research Challenges of Big Data,” *Service Oriented Computing and Applications*, vol. 13, no. 2, pp. 105 – 107, Jun 2019.

[30] J. Hiba, H. Hadi, A. Hameed Shnain, S. Hadishaheed, and A. Haji, “Big Data and Five V’s Characteristics,” *International Journal of Advances in Electronics and Computer Science*, pp. 2393–2835, Jan 2015.

[31] L. Gitelman, Ed., *Raw data is an oxymoron*. Cambridge, Massachusetts: The MIT Press, 2013. [Online]. Available: https://search.library.wisc.edu/catalog/9910134570702121

[32] N. Barrowman, “Why data is never raw,” *The New Atlantis*, Dec 2018. [Online]. Available: https://www.thenewatlantis.com/publications/why-data-is-never-raw

[33] I. Tuomi, “Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge

management and organizational memory," *Journal of Management Information Systems*, vol. 16, no. 3, pp. 103–117, 1999. [Online]. Available: https://doi.org/10.1080/07421222.1999.11518258

[34] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* SAGE, Aug 2014.

[35] T. Levitt, *The marketing imagination / Theodore Levitt.* Free Press ; Collier Macmillan New York : London, 1983.

[36] A. J. Berre, A. Tsalgatidou, C. Francalanci, T. Ivanov, T. Pariente-Lobo, R. Ruiz-Saiz, I. Novalija, and M. Grobelnik, *Big Data and AI Pipeline Framework: Technology Analysis from a Benchmarking Perspective*, E. Curry, S. Auer, A. J. Berre, A. Metzger, M. S. Perez, and S. Zillner, Eds. Springer International Publishing, 2022. [Online]. Available: https://doi.org/10.1007/978-3-030-78307-5{_}4

[37] J. C. Hull, *Options futures and other derivatives.* Pearson Education India, 2003.

[38] A. Damodaran, "Dealing with intangibles: Valuing brand names, flexibility and patents," *Flexibility and Patents (April 7, 2007)*, 2007.

[39] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions," *Journal of Computer-Aided Molecular Design*, vol. 34, 10 2020.

[40] H. H. Leivestad and J. Markkula, "Inside container economies," *Focaal*, vol. 2021, no. 89, pp. 1 – 11, 2021.

[41] L. Khalili, *Sinews of war and trade: Shipping and capitalism in the Arabian Peninsula.* Verso, 2020.

[42] C. D. Clack, V. A. Bakshi, and L. Braine, "Smart contract templates: foundations, design landscape and research directions," *arXiv*, pp. 1–15, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1608.00771

[43] A. Ng, "Data-centric AI Resource Hub," 2022. [Online]. Available: https://datacentricai.org/

[44] L. Singh, "The metal box that transformed global trade: The innovative vision of Malcom McLean behind the container revolution," *Legacy*, vol. 19, 2019. [Online]. Available: https://opensiuc.lib.siu.edu/legacy/vol19/iss1/4

[45] C. Tang, *Data Capital.* Springer Cham, 2021. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-60192-8

[46] H. Atwal, "DataOps Technology," in *Practical DataOps.* Apress, Dec. 2019, pp. 215–247. [Online]. Available: https://doi.org/10.1007/978-1-4842-5104-1_9

[47] P. Verdin and K. Tackx, "Are you creating or capturing value? a dynamic framework for sustainable strategy," *Harvard Kennedy School*, Jan 2015. [Online]. Available: https://www.hks.harvard.edu/centers/mrcbg/publications/awp/awp36

[48] C. Fuchs and V. Mosco, *Marx in the Age of Digital Capitalism.* Leiden, The Netherlands: Brill, 2015. [Online]. Available: https://brill.com/view/title/31597

[49] G. Dale, C. Holmes, and M. Markantonatou, *Karl Polanyi's Political and Economic Thought.* Agenda Publishing, Jul 2019. [Online]. Available: http://dx.doi.org/10.2307/j.ctvnjbfgk

[50] M. Wilenius, "Review Essay : A New Globe in the Making: Manuel Castells on the Information Age," *Acta Sociologica*, vol. 41, no. 2-3, pp. 269–276, 1998. [Online]. Available: https://doi.org/10.1177/000169939804100215

[51] M. Bost and M. S. May, "Michael Hardt and Antonio Negri in Communication Studies," *Oxford Research Encyclopedia of Communication*, 02 2017. [Online]. Available: https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-80

[52] S. Flensburg and S. Lomborg, "Datafication research: Mapping the field for a future agenda," *New Media & Society*, Sep 2021. [Online]. Available: https://doi.org/10.1177/14614448211046616

[53] J. Wagner, "Circulating value: convergences of datafication, financialization, and urbanization," *Urban Transformations*, vol. 3, 06 2021.

[54] D. Harvey, *Marx, capital and the madness of economic reason.* Profile Books, 2019.

[55] S. Olhede and R. Rodrigues, "Why data is not a commodity," *Significance*, vol. 14, pp. 10–11, 10 2017.

[56] R. Heines, C. Dick, C. Pohle, and R. Jung, "The tokenization of everything: Towards a framework for understanding the potentials of tokenized assets," *PACIS 2021 Proceedings*, Jul 2021. [Online]. Available: https://aisel.aisnet.org/pacis2021/40

[57] J. Schubert, "The work-intensive fiction of frictionless trade in the Angolan port of Lobito," *Focaal*, vol. 2021, pp. 64–78, 03 2021.

[58] E. Kazim, E. Fenoglio, A. Hilliard, A. Koshiyama, C. Mulligan, M. Trengove, A. Gilbert, A. Gwagwa, D. Almeida, P. Godsiff, and K. Porayska-Pomsta, "On the sui generis value capture of new digital technologies: The case of AI," *Patterns*, vol. 3, no. 7, p. 100526, 2022.

[59] N. Economides and I. Lianos, "Restrictions On Privacy and Exploitation In The Digital Economy: A Market Failure Perspective," *Journal of Competition Law & Economics*, vol. 17, no. 4, pp. 765–847, 04 2021. [Online]. Available: https://doi.org/10.1093/joclec/nhab007

[60] P. Langley and A. Leyshon, "Platform capitalism: The intermediation and capitalization of digital economic circulation," *Finance and Society*, vol. 3, no. 1, p. 11–31, Oct 2017. [Online]. Available: http://financeandsociety.ed.ac.uk/article/view/1936/0

[61] S. Marginson, "Value creation in the production of services: a note on Marx," *Cambridge Journal of Economics*, vol. 22, no. 5, pp. 573–585, 1998. [Online]. Available: http://www.jstor.org/stable/23600454

[62] C. Fuchs, *Social Media: a Critical Introduction.* SAGE Publication, 2013.

[63] Z. Dehghani, "How to move beyond a monolithic data lake to a distributed data mesh," *Martinfowler.com*, May 2019. [Online]. Available: https://martinfowler.com/articles/data-monolith-to-mesh.html

[64] Parquet, "Apache Parquet | Apache Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval." [Online]. Available: https://parquet.apache.org/

[65] "Git large file storage." [Online]. Available: https://git-lfs.com/

[66] "Versioning data and models | data version control DVC." [Online]. Available: https://dvc.org/doc/use-cases/versioning-data-and-models

[67] A. Barrak, E. E. Eghan, and B. Adams, "On the co-evolution of ML pipelines and source code - empirical

study of DVC projects," in *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2021, pp. 422–433.

[68] G. Schoeniger, A. Villanova del Moral, and J. Chaumond, "Oxen-ai, optimised vc for structured and unstructured data." [Online]. Available: https://github.com/Oxen-AI/oxen-release

[69] M. Sporny, D. Longley, M. Sabadello, D. Reed, O. Steele, and C. Allen, "Decentralized identifiers (DIDs) v1.0," *W3C Recommendation*, Jul 2022. [Online]. Available: https://www.w3.org/TR/did-core/

[70] M. Sporny, D. Longley, and D. Chadwick, "Verifiable credentials data model v1.1," *W3C Recommendation*, Mar 2022. [Online]. Available: https://www.w3.org/TR/vc-data-model/

[71] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards | Purposeful and transparent dataset documentation for responsible AI," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1776–1826.

[72] H. Face, "Create a dataset card." [Online]. Available: https://huggingface.co/docs/datasets/dataset_card

[73] "The best data curation tools for computer vision in 2022," 2022. [Online]. Available: https://www.lightly.ai/post/data-curation-tools-2022

[74] Docker, "Docker | Accelerated, Containerized Application Development." [Online]. Available: https://www.docker.com/

[75] W. WASM, "WASM WebAssembly." [Online]. Available: https://webassembly.org/

[76] WASI, "WASI | The WebAssembly System Interface." [Online]. Available: https://wasi.dev/

[77] A. Koshiyama, E. Kazim, P. Treleaven, P. Rai, L. Szpruch, G. Pavey, G. Ahamat, F. Leutner, R. Goebel, A. Knight *et al.*, "Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms," *Available at SSRN 3778998*, 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778998

[78] GensynAI, "Gensyn AI | decentralised compute to push the boundaries of machine learning." [Online]. Available: https://www.gensyn.ai/

[79] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

[80] Confidential Computing Consortium, "CCC outreach whitepaper (updated November 2022)," Nov 2022. [Online]. Available: https://confidentialcomputing.io/wp-content/uploads/sites/85/2023/01/CCC_outreach_whitepaper_updated_November_2022.pdf

[81] I. Damgård, V. Pastro, N. Smart, and S. Zakarias, "Multiparty computation from somewhat homomorphic encryption," in *Advances in Cryptology–CRYPTO 2012: 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012. Proceedings.* Springer, 2012, pp. 643–662.

[82] Cointelegraph, "A beginner's guide to the different types of blockchain networks."

[83] FHIR, "FHIR v4.3.0 | Fast Healthcare Interoperability Resources Specification." [Online]. Available: https://www.hl7.org/fhir/overview.html

[84] R. Mahanti, *Data Governance and Data Management.* Springer, 2021. [Online]. Available: https://link.springer.com/book/10.1007/978-981-16-3583-0

[85] Stanford University, "Deontic logic (stanford encyclopedia of philosophy)." [Online]. Available: https://plato.stanford.edu/entries/logic-deontic

[86] J. Chen and R. Charlene, "Fractional ownership: Definition, purpose, examples." [Online]. Available: https://www.investopedia.com/terms/f/fractionalownership.asp

[87] S. P. Lalley and E. G. Weyl, "Quadratic voting: How mechanism design can radicalize democracy," in *AEA Papers and Proceedings*, vol. 108, 2018, pp. 33–37.

[88] C. D. Clack and C. McGonagle, "Smart Derivatives Contracts: the ISDA Master Agreement and the automation of payments and deliveries," 2019. [Online]. Available: https://arxiv.org/abs/1904.01461

[89] BIS, "OTC derivatives statistics at end-june 2022." [Online]. Available: https://www.bis.org/publ/otc_hy2211.htm

[90] J. Goguen, "Data, Schema, Ontology and Logic Integration," *Logic Journal of the IGPL*, vol. 13, no. 6, pp. 685–715, 2005.

[91] P. Hummel, M. Braun, M. Tretter, and P. Dabrock, "Data sovereignty: A review," *Big Data & Society*, vol. 8, no. 1, p. 2053951720982012, 2021.

[92] W. Radomski, A. Cooke, and C. Philippe, "ERC-1155: Multi Token Standard." [Online]. Available: https://eips.ethereum.org/EIPS/eip-1155

[93] J. Klement, "What Is an Asset Class? | CFA Institute Enterprising Investor." [Online]. Available: https://blogs.cfainstitute.org/investor/2019/12/30/what-is-an-asset-class/

[94] T. Börgers and D. Krahmer, *An introduction to the theory of mechanism design.* Oxford University Press, USA, 2015.

[95] L. Zhang, T. Wu, S. Lahrichi, C.-G. Salas-Flores, and J. Li, "A data science pipeline for algorithmic trading: A comparative study of applications for finance and cryptoeconomics," in *2022 IEEE International Conference on Blockchain (Blockchain).* IEEE, 2022, pp. 298–303.

[96] L. W. Cong, Z. He, and K. Tang, "Staking, token pricing, and crypto carry," *Available at SSRN 4059460*, 2022.

[97] W. Kenton and S. Anderson, "What is the secondary market? how it works and pricing." [Online]. Available: https://www.investopedia.com/terms/s/secondarymarket.asp

[98] TIBCO, "What is Data Compliance? | TIBCO Software." [Online]. Available: https://www.tibco.com/reference-center/what-is-data-compliance

[99] S. Spiekermann, "The challenges of privacy by design," *Communications of the ACM*, vol. 55, no. 7, pp. 38–40, 2012.

[100] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1902.04885

[101] P. Treleaven, M. Smietanka, and H. Pithadia, "Federated learning: the pioneering distributed machine learning and privacy-preserving data technology," *IEEE Computer*, vol. 55, no. 4, pp. 20–29, 2022.

[102] Q. Yang, L. Fan, and H. Yu, *Federated Learning: Privacy and Incentive.* Springer Nature, 2020, vol. 12500.

[103] J. Bode, N. Kühl, D. Kreuzberger, and S. Hirschl, "Data Mesh: Motivational Factors, Challenges, and Best

Practices," 2023. [Online]. Available: https://arxiv.org/abs/2302.01713

[104] A. Ghiran and R. A. Buchmann, "The model-driven enterprise data fabric: A proposal based on conceptual modelling and knowledge graphs," in *Knowledge Science, Engineering and Management - 12th International Conference, KSEM 2019, Athens, Greece, August 28-30, 2019, Proceedings, Part I.* Springer, 2019, pp. 572–583. [Online]. Available: https://doi.org/10.1007/978-3-030-29551-6_51

[105] J. Bode, N. Kühl, D. Kreuzberger, and S. Hirschl, "Data Mesh: Motivational Factors, Challenges, and Best Practices," *CoRR*, vol. abs/2302.01713, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.01713

[106] K. Gulati, *Latest Data and Analytics Technology Trends That Will Change Business Perspectives.* CRC Press, 07 2020, pp. 153–184.

[107] M. L. Gray and S. Suri, *Ghost work: How to stop Silicon Valley from building a new global underclass.* Eamon Dolan Books, 2019.

[108] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5b: An open large-scale dataset for training next generation image-text models," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: https://openreview.net/forum?id=M3Y74vmsMcY

[109] J. Vincent, "AI art tools stable diffusion and midjourney targeted with copyright lawsuit - the verge." [Online]. Available: https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

[110] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2014. [Online]. Available: https://arxiv.org/abs/1409.0575

[111] C. Crawl, "Want to use our data? – common crawl." [Online]. Available: https://commoncrawl.org/the-data/

[112] G. Smith, "OpenAI's ChatGPT Criticized by News Media for Using Articles to Train Bot - Bloomberg." [Online]. Available: https://www.bloomberg.com/news/articles/2023-02-17/openai-is-faulted-by-media-for-using-articles-to-train-chatgpt

[113] J. Vincent, "Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement - The Verge." [Online]. Available: https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion

[114] "Genomes IO | DNA Bank." [Online]. Available: https://genomes.io/

[115] M. Zichichi, S. Ferretti, and V. Rodriguez-Doncel, "Decentralized personal data marketplaces: How participation in a DAO can support the production of citizen-generated data," *Sensors*, vol. 22, no. 16, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/16/6260

[116] R. Eichler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang, "Data shopping—how an enterprise data marketplace supports data democratization in companies," in *Intelligent Information Systems: CAiSE Forum 2022, Leuven, Belgium, June 6–10, 2022, Proceedings.* Springer, 2022, pp. 19–26.

[117] G. S. Ramachandran, R. Radhakrishnan, and B. Krishnamachari, "Towards a decentralized data marketplace for smart cities," in *2018 IEEE International Smart Cities Conference (ISC2).* IEEE, 2018, pp. 1–8.

[118] Q. Song, J. Cao, K. Sun, Q. Li, and K. Xu, "Try before you buy: Privacy-preserving data evaluation on cloud-based machine learning data marketplace," in *Annual Computer Security Applications Conference*, 2021, pp. 260–272.

[119] A. Ghanti and J. Mansa, "Clearinghouse: An Essential Intermediary in the Financial Markets." [Online]. Available: https://www.investopedia.com/terms/c/clearinghouse.asp

[120] Open Data Institute, "What is open data? – the odi." [Online]. Available: https://www.theodi.org/article/what-is-open-data/

[121] European Union, "Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast)." [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024

[122] Research Data Scotland, "What are Trusted Research Environments? | Research Data Scotland." [Online]. Available: https://www.researchdata.scot/what-are-trusted-research-environments

[123] NHS, "Trusted Research Environment service for England - NHS Digital." [Online]. Available: https://digital.nhs.uk/coronavirus/coronavirus-data-services-updates/trusted-research-environment-service-for-england

[124] Open Data Institute, "Data Ecosystem Mapping tool – The ODI." [Online]. Available: https://www.theodi.org/article/data-ecosystem-mapping-tool/

[125] J. Lanier, *Who Owns the Future?* Simon & Schuster, 2013.