Data Collaboratives with the Use of Decentralised Learning

Maciej Krzysztof Zuziak*
Knowledge Discovery and Data Mining Laboratory (KDD Lab) National Research Council of Italy (CNR)
maciejkrzysztof.zuziak@isti.cnr.it

Onntje Hinrichs†
Research Group on Law, Science, Technology and Society (LSTS) Vrije Universiteit Brussel (VUB)
onntje.marten.hinrichs@vub.be

Aizhan Abdrassulova
Jagiellonian University
aizhan.abdrassulova@uj.edu.pl

Salvatore Rinzivillo
Knowledge Discovery and Data Mining Laboratory (KDD Lab) National Research Council of Italy (CNR)
rinzivillo@isti.cnr.it

ABSTRACT

The endeavor to find appropriate data governance frameworks capable of reconciling conflicting interests in data has dramatically gained importance across disciplines and has been discussed among legal scholars, computer scientists as well as policy-makers alike. The predominant part of the current discussion is centered around the challenging task of creating a data governance framework where data is 'as open as possible and as closed as necessary'. In this article, we elaborate on modern approaches to data governance and their limitations. It analyses how propositions evolved from property rights in data towards the creation of data access and data sharing obligations and how the corresponding debates reflect the difficulty of developing approaches that reconcile seemingly opposite objectives – such as giving individuals and businesses more control over 'their' data while at the same time ensuring its availability to different stakeholders. Furthermore, we propose a wider acknowledgement of data collaboratives powered by decentralised learning techniques as a possible remedy to the shortcomings of current data governance schemes. Hence, we propose a mild formalization of the set of existing technological solutions that could inform existing approaches to data governance issues. Our proposition is based on an abstractive notion of collaborative computation as well as on several principles that are essential for our definition of data collaboratives. By adopting an interdisciplinary perspective on data governance, this article highlights how innovative technological solutions can enhance control over data while at the same time ensuring its availability to other stakeholders and thereby contributing to the achievement of the policy goals of the European Strategy for Data.

CCS CONCEPTS

• **Collaborative and social computing theory, concepts and paradigms**; • **Collaborative and social computing systems and tools**; • **Privacy policies**;

KEYWORDS

Data Governance, Decentralised Learning, Data Access, Data Sharing, European Strategy for Data

ACM Reference Format:
Maciej Krzysztof Zuziak, Onntje Hinrichs, Aizhan Abdrassulova, and Salvatore Rinzivillo. 2023. Data Collaboratives with the Use of Decentralised Learning. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3593013.3594029

1 INTRODUCTION

The interest in data governance has spiked over the last years, gathering the worldwide attention of policy-makers and various research communities. The machine learning community is most concerned with gaining access to an undisrupted source of high-quality data. From the engineering side of view, the problem concerns mostly building trust between data providers and deploying an infrastructure that may support data access. The European Commission, taking into account the current market disadvantage of smaller entities, wants to guarantee such access, albeit still ensuring a high degree of protection for fundamental rights. From the European policy perspective, access to data should therefore be "as open as possible, as closed as necessary", which generally implies finding a point of equilibrium between conflicting interests in data. Consequently, the creation of data governance frameworks that enable the emergence of a data-agile economy [18] where tensions, resulting from the seemingly opposed objectives of enabling control of data and guarantee its availability to other stakeholders at the same time, exceeds the boundaries of a single discipline.

This paper provides an interdisciplinary perspective on how to solve tensions resulting from diverging interests in data by providing insights in both, the regulatory as well as technological layer in data governance. While the provided solutions are mainly discussed in the European context - they may not necessarily be limited to such. The main contribution of this paper consists of the following:

First, it analyses how the creation of property rights in data was envisioned by scholars and policy-makers as a possible solution to reconcile conflicting interests: enhancing control over data while

*Both authors contributed equally to this paper
†Both authors contributed equally to this paper

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594029

at the same time ensuring its availability to stakeholders. Furthermore, it explains the shortcomings of this idea and why it was subsequently abandoned and replaced by a new approach in EU policy that should achieve similar objectives but by different means: the creation of mandatory data access and sharing rights.

Second, the analysis then shifts towards the technological layer in data governance and how distributed computation methods might promote data access and data sharing. Building upon the concept of data collaboratives, it shows how technological solutions through distributed and collaborative processes might be capable of achieving the seemingly opposite objectives of data control and availability and could thereby contribute to reduce existing tensions in data economies. This framework should provide constraints on the access and control of the shared individual data and give back to each participant an acknowledgement of the contribution by means of a repertoire of metrics.

The paper is structured in the following way: Section 2 outlines how data governance has turned into a key policy concern for the European legislator. Section 3 is dedicated to the discussion of past and current regulatory approaches to data governance in EU policy as well as their shortcomings. Section 4 introduces the concept of data collaboratives, defined through the set of four principles, as a possible tool shifting data governance in the technological layer towards mutually beneficial cooperation. Section 5 presents conclusions of this work.

2 DATA GOVERNANCE AS A KEY POLICY CONCERN

Whereas the European Commission does not offer a comprehensive definition of its "European way of data governance" [23], data governance has in general been defined by scholars as the exercise of authority and control over the management of data, whilst its purpose would be to increase the value of data and minimize data-related costs and risks [1]. Von Grafenstein differentiates between three analytical layers that are equally mentioned by the legislators with each having their own challenges: (i) the regulatory layer (ii) the organisational layer (iii) the technological layer [32].

Since the Juncker Commission identified the creation of a Digital Single Market as a critical policy objective in 2014, the EU has been extremely active in proposing regulation for a digital economy that ensures a high level of data protection while, at the same time, boosting competitiveness and innovation capacities. With the recognition of data as an "essential resource for economic growth, job creation, and societal progress" [18], the interest in data governance has spiked over the last years, gathering worldwide attention of policy-makers and research communities. Whereas the regulation of data flows originated in attempts to protect the fundamental rights to privacy and data protection [62], regulatory perspectives evolved with the growing understanding of benefits that result from data-driven innovation that relies on data as a resource [59]. Since data would not have any intrinsic value but its value would rather depend on the context of use as well as the extent to which it can be reused [60] ensuring availability and access to quality data has become a key concern.

Whilst the European Data Protection Framework has been conceived in fundamental rights terms, it has a dual purpose to also ensure the free flow of personal data [48]. EU policy of the past years therefore equally aimed at adding a fifth to the four existing freedoms of the European Single Market (free flow of goods, capital, services, and labour): the free flow of data.[12, 79] However, in addition to growing concerns about the internal consistency of this patchwork of rules and regulations [13] the question comes up if the current path to data governance taken by the EU in its Data Strategy is capable of achieving its objectives. How a framework should look like that ensures both the availability of personal and nonpersonal data to a wide variety of actors as well as the protection of fundamental rights of different stakeholders is far from clear. Whilst substantial gains are expected from exploiting the non-rivalrous character of data by increasing its availability, neither companies nor individuals would necessarily want their data widely available due to potential threats to privacy, intellectual property rights, or trade secrets if data is disclosed [11], [12]. Resolving these tensions lies at the heart of discussions on how to design a framework for the data-agile economy. It, therefore, comes as no surprise that commentators described the regulation of data markets as "shaping up to be one of the major challenges of the twenty-first century" [76].

3 FROM PROPERTY RIGHTS TO DATA ACCESS AND SHARING OBLIGATIONS

Current approaches and propositions with regard to appropriate frameworks for data governance, however, do not come out of the void. Instead, they build upon previous discussions and proposals that equally attempted to find solutions on how conflicting interests in data could be resolved. Therefore, before turning to our model of data collaboratives based on decentralised learning techniques, this section first retraces how property rights in data have been considered as a means to create a flourishing data economy. It thereby sheds light on objectives which regulatory frameworks should achieve. Second, it analyses how a property approach was abandoned and replaced by a regulatory approach that focuses on creating data access and sharing obligations instead. In line with the distinction of three analytical layers in data governance [32], this section therefore focuses on the first, i. e. the regulatory layer and different approaches that have been discussed by scholars and policy-makers.

3.1 Data Property

A frequently used example for the explanation of the necessity of property rights in society constitutes the "tragedy of the commons" elaborated by Hardin [53]. Scarce resources that are accessible to all members of society would ultimately end up being overexploited and depleted – freedom in a commons would bring ruin to all [35]. This tragedy could only be averted either with the introduction of a system of private property rights that should enable efficient resource allocation according to market mechanisms, or via government regulation that define rules for the use of that resource. Whilst the characteristics of data as an intangible, non-rivalrous resource that does not get depleted upon consumption make it a difficult fit for traditional economic and legal categories [39], debates emerged how property rights in data could contribute to the preservation of the 'privacy commons' which might otherwise degrade due to

concerns about an (over)use of personal information [69]. As described by Thouvenin et al., data propertization gave "hope to those wishing to unlock the potential of the data economy and to those trying to re-empower individuals that have lost control over their data" [77].

Whereas a certain degree of legal protection can already be offered by non-property regimes such as trade secrets or contracts, and de facto ownership positions via technical protection measures, none of these regimes create rights erga omnes – i.e. enforceable rights against the world [36]. It is property law that confers exclusive rights that can be invoked against more than a number of specific persons and it is precisely this universality that has been described as the touchstone of property rights [8]. The existing European intellectual property rights regimes, however, do not provide for a comprehensive property right in data as such. Instead, they offer different degrees of protection based on criteria such as creativity, investment, or secrecy. Some commentators have therefore rightfully used the metaphor of a patchwork to describe the protection offered by existing intellectual property regimes in data [33, 86]. Furthermore, neither the Data Protection Directive nor its successor the General Data Protection Regulation, grounded in the unalienable fundamental right to data protection, created any property rights for data subjects over their data.

Whether or not the introduction of a new comprehensive unitary property rights regime might be the appropriate approach to solve diverging interests in data has been subject to a long-lasting scholarly debate. Early proponents in the US argued that property rights in data would empower individuals to regain control over their data and to value their privacy according to their preferences [45]. The protection of data would thereby be linked with the incentives of the market by using the laws of property as a control mechanism [45]. Others, however, feared that a property rights approach would ultimately erode existing levels of privacy if individuals simply traded away their personal data [10]. Market solutions based on a property rights model might thus not cure any of the problems related to control, but only legitimize them [46].

Particularly in Europe, where information privacy is considered as a fundamental right and where personal data should consequently not be considered as a commodity that could be bought and sold on a market, a property rights approach appeared problematic. At the same time, critics observed that maintaining the argument that personal data should be nobody's property would be illusional in a data-driven economy [65]. Instead, the status quo devoid of well-defined property rights in data would result in a de facto assignment of (economic) property rights to the information industry, eroding data subject's autonomy, privacy and informational self-determination [65].

Other scholars discussed the introduction of property rights in data not from the angle of the empowerment of the individual but the creation of data markets [85]. Property rights in non-personal data might create incentives to generate and disclose data – thereby enhancing the general availability of data for other market actors. Whilst the proposition was taken up in the proposal by the European Commission on the possible introduction of a data producer's right [17], criticism prevailed. The reliance of Big Data on permanent, dynamic access to real-time data sources could hardly be aligned with individual ownership [15]. Instead, creating an additional layer of rights in data might cause disruptive overlaps with existing copyright and the sui generis database right [36]. Consequently, competing claims of ownership would risk resulting in a "tragedy of the anticommons", i.e. an underuse of data due to complex interrelations of overlapping ownership claims [74].

With the European Strategy for Data, a turning point in the EU's approach to the regulation of the data economy has been reached. After decades of discussions, the creation of property rights in data seems to have been discarded as a suitable mechanism that is capable of resolving conflicting interests in data. However, the rationale behind its proposition, i. e. empowering individuals and companies with regard to 'their' data as well as enabling the emergence of a data-agile economy, is still existent. Before turning to our model of data collaboratives and how it might contribute to solving those tensions, the following section elaborates on the shift away from property rights towards data access and sharing in EU policy.

3.2 The EU's Shift towards Data Access and Data Sharing

As the academic debate shifted away from conceiving property rights in data as a potential legal solution to resolve conflicting interests in data, a similar tendency could be observed in EU policy. Whilst the goal of the European Strategy for Data remains to realize a 'genuine single market for data', any reference to (legal) data ownership rights disappeared from the European Commission's communications and proposals. Commentators hence described the strategy as a 'paradigm shift' in EU policy from data ownership towards facilitating data access and sharing [56].

Since data would constitute an essential resource for companies to develop new products, services or train AI systems, a regulatory environment should enable stakeholders to have easy access to an almost infinite amount of data. Regulation should enable data to flow easily while at the same time ensure that European rules and values are fully respected. Key pillars of the European Strategy for Data thereby attempt to enable precisely the creation of such a regulatory environment that ensures 'better access to data and its responsible usage'[18]. Regulating data access instead of creating property rights in data have thus become the new instrument that should solve conflicting interests in data.

First, the Data Governance Act (DGA) [23] intends to facilitate access and encourage use of data held by public sector bodies. Whilst the Open Data Directive already established minimum rules governing the re-use for data that is not subject to, for instance, intellectual property rights or commercial confidentiality, the DGA harmonizes access conditions for reuse of protected categories of data. Moreover, the DGA puts in place a framework for 'data intermediation services' and 'data altruism organization' which should play a key role in the data economy by supporting, facilitating, and promoting (voluntary) data sharing by bringing demand and supply of data together. A notification scheme for data sharing services should thereby increase trust in these data intermediaries (and hence in data sharing) by companies and data subjects.

Second, the European Commission intends to create Common European Data Spaces [20] in strategic sectoral fields [21] that bring

together relevant data infrastructures and governance frameworks to facilitate data pooling and sharing. Key features of this data spaces should be that they provide a secure infrastructure to pool, access, share, and use data while ensuring the protection of European rules. The structure should thereby provide access to and use of data in a fair, transparent, and non-discriminatory manner via trustworthy data governance mechanisms. The DGA is thus supposed to play a fundamental role for the creation of data spaces as it should create the foundation for the establishment of trustworthy, neutral data intermediaries.

Third, the Data Act proposal [22] constitutes together with the DGA the second horizontal legislative instrument of the European Data Strategy. Whilst the DGA aims at opening up data held by public sector bodies, the Data Act proposal creates mandatory access rules to data held by private actors in B2C, B2B as well as B2G relations. The data access rights are thereby supposed to fulfil the twofold purpose of empowering individuals and companies over 'their' data while at the same time ensuring that data is available for other stakeholders. It is still unclear, however, how mandatory data access rights would have to be designed in practice to comply with, for instance, the requirements of the current Data Act proposal. Whether they would have to be designed as ex-situ rights that enable the porting of data directly to another platform or as in-situ rights that would require external algorithms to be transferred to the data location to perform data analysis [41].

Since the Data Act proposal merely demands that data should be 'made available' or to implement the principles of data minimization and data protection by design by using technology that 'permits algorithms to be brought to the data', it seems likely that in many instances it might favor in-situ access rights that do not require the transfer of data.

Finally, this paradigm shift towards fostering data access is not only visible in the European Strategy for Data but has also been observed in other new European digital regulations [59]. Provisions in the Digital Markets Act (DMA) [24], the Digital Service Act (DSA) [25] and the proposal for an AI Act [19] would also contain provisions that reflect the EU's attempt to open up data while at the same time finding a balance between new access and transparency obligations and competing interests in data. New access and/or portability obligations in the DMA, DSA, and the AI Act would therefore equally constitute part of the latest regulatory effort of the EU to end data enclosure while departing from a "single-minded rights-based and data ownership approach" [58].

At the same time, however, authors have been critical of approaches based on access rights alone. Focusing exclusively on the problem of whether a company should get access to data would be insufficient. The "data as a resource" framing of the European Strategy for Data that regards 'data abundance' as desirable would necessarily continue to conflict with the purpose limitation and data minimization principles in European Data Protection law [76]. This tension would not be automatically resolved in a regulatory governance model that merely proscribes data access and sharing obligations. Viljoen argues that a common flaw in existing data governance frameworks based on propertarian or fundamental rights based approaches would be its consistent focus on attempting to reassert individual instead of more collective forms of control over ones datafication. Since data processing practices by major companies would primarily aim at deriving population-level insights, data governance approaches would have to move beyond individualist claims and develop institutional responses to represent the relevant interests at stake on a more collective level [82]. Other scholars insist that a broader analytical approach would be necessary, for instance, by the use of data trustees or technological solutions that impact the allocation of de facto control of data [40].

The following section thus presents a model of data collaboratives based on decentralised learning techniques that might remedy some of the existing shortcomings by moving beyond individual control of data. Instead, it employs a more collective approach that enables participants in the collaborative to have access to different datasets and to collectively train one model without the need for transferring any raw data. It exemplifies how a solution on the technological layer could contribute to reduce tensions which the regulatory layer alone could not solve, i.e. enabling control over data while at the same time making it available to other stakeholders. Furthermore, the following proposition mirrors equally the demand of the Data Act to conceive data access rights in a way they implement the data protection by design and data minimization principle

4 DATA COLLABORATIVES AND DATA ACCESS

4.1 Data Governance and the Commons Management Problem

When Hardin saw either a market-based approach via property rights or government intervention as the only possible solutions to prevent the tragedy of the commons to occur, scholarship on the commons [61] argued that an alternative approach was possible. This approach would rely on collective resource management, i.e. the institutionalized sharing of resources among members of a community [49]. Commons thereby consist of three elements: (i) a resource (ii) a community that has access to and takes care of the resource (iii) collective action of creating, maintaining, and governing in common [51]. Since they are based on the idea of access rather than exclusion, they would challenge the dominance of private property [51] or even constitute the opposite of property [54]. The main function of commons would therefore be to institutionalize freedom to operate within symmetric constraints, free of the particular risk that any other can deny use of that resource set [6]. The commons-framework therefore does not imply unmanaged access to a particular resource, but requires that groups engage in managed resource sharing [48].

Ostrom famously defined eight design principles [61] that enable successful collective resource management, such as collective-choice arrangements that allow groups to adapt governance conditions to their needs and local circumstances. Whilst initially conceived for the management of natural resources whose characteristics are evidently very distinct from data, they would still be highly relevant with regard to data governance as they could provide insights for assessing the types of institutions needed to govern access and use of data [49]. How data could be subject to regulation via commons governance institutions has been subsequently further analysed within research on governing knowledge commons that

have been defined as "institutionalized community governance of the sharing and, in some cases, creation, of information, science, knowledge, data, and other types of intellectual and cultural resources" [29]. Its key insight would be how information resources are governed as a shared resource via a collective and not in markets via intellectual property rights regimes or state intervention [48]. New technologies thereby would have played a key role for the possible development of knowledge commons as they facilitated the ability to capture the previously uncapturable and to draw value from it by means of advanced data analytics [66].

Studies therefore attempted to adapt Ostrom's design principles for collective self-managing institutions and translate them to and extract useful insights for data governance [2, 11, 63]. This further resulted in discussions and analyses, [2, 84] on various models of data stewardships, such as data trusts, data foundations, data cooperatives or data collaboratives [84]. Data collaboratives thereby have been described as new emerging forms of partnerships where privately held data is made accessible for analysis. The collaboration between participants that is facilitated within these structures aims to result in new insights and innovation and to unlock the public good potential of previously siloed data [31]. Our approach to data collaboratives and collaborative data sharing is based on many advancements in the technological field that were presented in the span of the last 20 years. We draw from such solutions as personal data stores [58, 72], distributed learning [42, 81, 82] or even data licensing [4]. What we propose is a mild formalization of the set of existing technological solutions adjusted to mitigate the issues arising from the current data governance methods. For the sake of better generalization, we limit ourselves to the method-agnostic notation, not to exclude any new or existing technological solutions that may be adopted to serve as a building block of new data collaboratives. Our method of defining the boundaries and goals of data collaboratives is based on the identification of common principles and basic computational operations that can be performed by the participants. In the following, we introduce the goal of data collaboratives, together with the four essential principles that define it. Next, we elaborate on the notation of collaborative computing that is a baseline for the presented structure. Lastly, we introduce current limitations of proposed approach.

4.2 Goal of the Data Collaborative

The proposed approach is based on an application of decentralised learning for better data governance and could provide a substantial leap towards independent and self-sovereign data management inside trusted communities.

The main goal of data collaboratives is to allow trusted parties to establish a private or hybrid (public-private) collaboration and train one shared model without transferring the raw data beyond the local storage. In this sense, the data collaborative is an applied case of data access and data sharing - because each party is allowing only access to derivative aggregates of its resources – and does not consent to any transfer of raw data. It therefore reflects demands set by the Data Act proposal with regard to access rights that should implement the data protection by design and the data minimization principle (in-situ instead of ex-situ access right). The trained model is then shared between all parties. We deliberately do not define here the notion of a party or a participant to the collaborative. We assume that it may be a physical person or an organization, and that the shared data may or may not constitute personal data. All these scenarios will require different legal approaches, as different types of rules may or may not apply, depending on a specific use-case. However, any approach to the data governance in the technological layer requires a degree of abstraction that would allow for a fine generalization of the presented method, and any synthetic definition of a participant may be harmful to this objective.

The main objective of the data collaborative is to always perform at least one collaborative computation (which is abstractly defined in the next section of this paper). At the same time, the collaborative may be capable of tracking the individual contribution of all participants with regard to the final output. Although this is a distinct problem on its own, that is overviewed as an open challenge, a data collaborative might therefore strengthen control over data in several ways. First, by not requiring participants to transfer raw data which is directly embodied in the four essential principles that define the data collaborative. Data access within our data collaborative would therefore be less invasive with regard to potential privacy, intellectual property, or commercial interests the stakeholders might have with regard to their data. Second, if we assume that the data collaborative is able to track and evaluate the contribution of each participant, this could enhance trust in the collaboratives as it empowers individuals to understand their contribution and thereby promotes the participation and collaboration to the collective model. Furthermore, one might envision the possible implementation of rules that enable proprietary or contractual claims with regard to the trained model that are made dependent on the proportion of the registered contribution by participants. At the same time, the collaborative could succeed in breaking up previously siloed data and thus ensuring its availability to other stakeholders. Where previous regulatory approaches failed, data collaboratives might contribute to reconcile conflicting interests in data.

As briefly mentioned in the previous paragraph, we define our commons management system through the set of four essential principles that are used classify it as a data collaborative and which are loosely connected to previous attempts to transpose Ostrom's principles on managing the commons to data governance frameworks. Firstly, the data collaboratives should provide an accessible infrastructure for performing various analytical tasks without the necessity to transfer raw data beyond the participants' devices. Consequently, the proposed architecture strengthens control of participants over their local datasets. [Decentralised Data Storage]. Once the model is trained (or once an analytical task is accomplished) it should be governed by all the members (in proportion to their marginal contribution). Shared governance is a key guarantee that all the members will benefit from joint participation in the analytical tasks. Collective-choice arrangements could be realized by allowing participants to the collaborative to create their own rules and governance conditions. What's more, it can be envisaged that during the establishment of the collaborative, parties establish operations and actions that require a certain number of votes and can only be initialized if a certain quorum has been reached. [Shared Model Governance]. The formal structure of a data collaborative should be mostly implementation-agnostic. This is because any

structure or implementation that satisfies the baseline definition and four essential principles can be treated as a data collaborative - irrespective of the implementation details [Universality]. Finally, data collaboratives can be established and executed in many different ways, combining the available technology with local needs. However, each data collaborative should be able to perform at least one analytical operation in a distributed environment [Minimal Utility – Collaborative Computation].

These principles can be fulfilled by the utilization of many already existing technologies. The last decade has seen the development of a large-batch synchronous distributed learning [14], federated learning [3, 5, 9, 42, 43, 50, 55, 67, 78, 83, 87, 88] and other" no-peek" approaches such as split neural network [80] that allowed us to develop a completely different mindset regarding machine learning, escaping the dogma of collecting the whole data in one centralized location in order to develop a model based on the chosen objective. Although many challenges regarding those methods are still open [38], decentralised learning has seen a number of use-cases that may be seen as a showcase of the potential hidden in that approach [9, 34, 67]. Most importantly, decentralised learning may serve as the backbone of a different approach to data governance, where the data "owners" are able to establish closed communities of high and medium trust and then collaboratively participate to the training of ML models.

4.3 Problem Statement and Basic Definitions for Collaborative Computations

The data collaborative is established between three or more parties that share an interest in performing one or more collaborative computation and retaining controls of outputs of those computations throughout the existence of the partnership. The principle of Minimal Utility requires the parties to be interested in performing at least one collaborative computation throughout the existence of the collective partnership. Therefore, we need to define – at least briefly – the notion of this term. For this, we use mostly the notation used in the theory of decentralised machine learning [70, 81, 83]. We assume to have a set of available clients (that are synonymous to the notion of parties of participants) $C = C_1, C_2, C_3, \ldots, C_n$ and each client C_i stores a set of data $X_i = \{(\bar{x}_1, y_2), (\bar{x}_2, y_2), (\bar{x}_3, y_4), \ldots, (\bar{x}_p, y_p)\}$ where each $(\bar{x}_i, y_i)$ is a one sample, $\bar{x}_i$ is a vector n-dimensional vector containing all features and y_i is a label that is attributed to the vector $\bar{x}_p$. The objective is to estimate a surrogate for a hidden function $h(\bar{x}) \rightarrow y$, that assign a label y to the observation $\bar{x}$.

Machine Learning training method recreates a hypothesis function $\tilde{h}(\bar{x}) \rightarrow y$ as a proxy of $h(\bar{x})$ such that $\tilde{h}(\bar{x}) \sim h(\bar{x})$. We want this approximation to be the closest to the unknown function $h(\bar{x})$ as possible. We can generally express that idea as:

$$min\, \mathbb{P}\left[\tilde{h}\left(\bar{x}\right) \neq h\left(\bar{x}\right)\right] \leq \varepsilon \tag{1}$$

where ϵ is some previously defined margin of error that we can tolerate given the projected task. The generalised error on distribution D and in relation to some underlying function $h(\bar{x})$ can be defined as generalization error:

$$\mathcal{L}\left[\tilde{h}\left(\bar{x}\right)\right] \stackrel{\text{def}}{=} \mathbb{P}\left[\left(\bar{x}\right) \neq h\left(\bar{x}\right)\right. \tag{2}$$

Equation 2) describes the generalization error that our hypothesis function is making on the given population. During training, we don't have access to the whole population, but only to the training sample that we have gathered for that particular purpose. In a centralized scenario, the dataset will be defined as a collection of data gathered from all available clients, i.e. $X = s \cup X_2 \cup X_3 \cup \ldots \cup X_i$where X_i is data stored at the client C_i. In such an environment, we can directly train one hypothesis function $\tilde{h}(\bar{x}) \rightarrow y$, and then evaluate it on the possessed data:

$$\mathcal{L}\left(\tilde{h}\left(\bar{x}_i\right)\right)_{ELF} = \frac{1}{n}\sum_{i=1}^{n}[\tilde{h}\left(\bar{x}_i\right) \neq y_i] \tag{3}$$

Equation 3) describes the empirical loss function, where we sum over all the n samples from a dataset, comparing the output of the function to the original label (independent) variable attached to the feature vector $\bar{x}_i$. Equation 3) also assumes that we can gather the whole dataset X in one (or few) centralised locations, irrespective of the number of clients that have generated said data. Although this is a common for a centralised scenario, such an action devoid independent actors from any control over their own data.

Therefore, we assume that clients do not want to consent to direct data transfer and are only considering engaging in a collaboration which do not require them to take such steps. Given that the participants have limited knowledge about the information coming from other distributions than their own, they still may be interested in information that could help them better understand the whole population. In such cases, all the actors may engage in a collaborative activity of collaborative computation (e. g. distributed analytics or distributed learning): in the first case (distributed analytics) they are interested in a single statistic regarding samples from other distributions; in the second (distributed learning) they participate to learn a hypothesis function $\tilde{h}(\bar{x})$ that minimizes the local loss functions on the datasets distributed differently than their own.

Both distributed computations, as well as the notion of **collaborative computation** in general, can be formalized as the composite function:

$$F\left(\bar{X}\right) = F\left(\bar{X}_1\right) \oplus F\left(\bar{X}_2\right) \oplus \ldots \oplus F\left(\bar{X}_i\right) \tag{4}$$

where: $F(\bar{X}_1), F(\bar{X}_2), \ldots, F(\bar{X}_i)$ are functions computed locally at each client, such that $F(\bar{X}) : \mathbb{R}^d \rightarrow \mathbb{R}$, and equation 4) denotes the composition of all local contributions. Providing an example let us assume we have p clients each one possessing a dataset $X_p = (\bar{x}_1, y_2), (\bar{x}_2, y_2), (\bar{x}_3, y_4), \ldots, (\bar{x}_p, y_p)$ that may be distributed differently, and the maximum length of the dataset $|X_p|$ may differ from client to client. Let us also assume that each vector x_i contains n different features. In such case, one may be interested in statistics regarding the selected features across all the population. Function $F(\bar{X}_i)$ may in that case locally compute mean of the sample

$$\overline{X_i} = \frac{1}{|\boldsymbol{X_i}|}\sum_{\boldsymbol{i}=1}^{\left|\overline{X_i}\right|}\boldsymbol{x_i} \tag{5}$$

where for each dimension n we obtain one scalar describing the mean of the local samples. Such results can be then aggregated and

once again averaged before returning the final vector:

$$\bar{X} = \frac{1}{|\bar{X}|} \sum_{i=1}^{|\bar{X}|} \bar{X}_i \tag{6}$$

is calculated and returned to all participants of the collaborative.

Another example of distributed computation is a collaborative approach to establishing the common hypothesis function $\tilde{h}(\bar{x}) \rightarrow y$ that is indirectly trained on data from all the clients. Given that the random variables are not independent and identically distributed , each local client can compute $F(\bar{X}_i) : \tilde{h}_i(\bar{X}_i) \rightarrow y$ and send back only the aggregate representing the final hypothesis function $\tilde{h}_i$ to the other members of the collaborative. Compound function from Equation 4) will be then a global hypothesis function that is made from composition of the local hypotheses functions. One simple algorithm that can be utilized to perform such aggregation is FedAvg – a vanilla baseline for Federated Learning [42].

The following definitions are independent of any particular implementation design. It may be possible that all the local aggregates are send back to the Central Orchestration Node (CON) that is deployed by a third-party as a service [44]. It is also possible that the following schema is implemented in a fully decentralized manner with peer-to-peer connections [5, 25, 51]. The universal notation of compound operations denoted as $\oplus$ was employed specifically, not to limit the domain of possible implementation designs. Hence, collaborative computation can be defined as any function $F(\bar{X})$ that accepts aggregates of the local functions $(\bar{X}) = F(\bar{X}_1) \oplus F(\bar{X}_2) \oplus \ldots \oplus F(\bar{X}_i)$ and broadcast the output to all the members who provided some aggregates of their local data.

Considering the multitude of circumstances under which the aforementioned methodology of data collaboratives can be used, it is advisable to define this concept as a set of principles that are fundamental to its existence, while the particularities of implementation may vary from case to case. Given an example, the personal data stores [57] may be seen as data collaborative, although their real implementation is quite different from what was presented here. On the other hand, existing solutions based on federated learning can be turned into data collaboratives. This flexibility is something that the concept of data property was devoid of. Hence, we believe, that the concept of data collaboratives is something that may fill up a gap in the current system of data governance.

4.4 Open Challenges Regarding Data Collaboratives and Directional Proposals

We can distinguish two major issues concerning the implementation of such an architecture, introduced here by a name of data collaborative, namely: metrics for individual contributions and the trade-off between local and global utility. The metrics for individual contributions could add an additional layer of utility to the functionality of data collaborative, as they would allow participants not only to perform collaborative computations, but also to track their individual contributions, thus enabling them to reward those who contribute more to the global data exploration. The trade-off between local and global utility is characteristic for methods of decentralized learning, although in the case of Data Collaborative, it is gaining even more importance, as generally such a solution should seek some kind of balance between interests of a different parties by its very definition. Therefore, we want to elaborate on these issues a little bit more, posing them as a challenge for further research on data collaboratives. Firstly, it is not a trivial task to establish common metrics for individual contribution, especially taking into account privacy constraints. If the information about contributions is disclosed to all members of the collective, each such disclosure may add an additional risk of infringing privacy or business interests in the data - and the magnitude is defined by the amount and type of information being disclosed. It is also a non-trivial question how we can measure individual contributions. The simpler the measure, the less informative it is regarding the real value of the contribution. The more complicated, the more risk there is that the measure will indeed cause a privacy infringement or will give raise to unjust and arbitrary model allocation.

Three approaches to the problem of contribution evaluation have been addressed in the literature so far, but the work on this issue should still deemed to be at an early stage of development. The simplest and most intuitive approach to this problem is to take into account the number of times that the particular client has been selected in respect to the total number of selections during the computational task. Let's assume that a computational task T requires at least k rounds, where in each round T_k only a subset of clients $C_S \subset C$ of cardinality $|C_S|$ is selected. Each round T_k we have a possibility to form $bincoef(C, |C_S|)$ different subsets. Depending on the relation of $|C_S|$ to $|C|$, a client C_i will have a different probability of being chosen multiple times. We can simply register the number of times each client C_i is chosen, and then divide it by the number of clients we choose each round, i.e. contribution of the client i can be defined as:

$$Con\,(C_i) \stackrel{\text{def}}{=} \frac{selection\,(C_i)}{k * |C_s|} \tag{7}$$

Approach presented in equation 7) requires a few additional assumptions that may not necessarily hold in our application scenario. Firstly, we must treat all the clients' participation equally, irrespective of the result. In many cases, this will not hold, because some clients' updates may be less valuable that others (we can just imagine client holding in its local dataset n samples, and a client possessing more than 10^n samples). In such a scenario, the disproportion of the contribution is striking. Secondly, we don't take into consideration the real computational resources needed to prepare the aggregates before sending it to the other members of the collective. In reality, such a simple metric will not hold in majority of cases.

The most common method for data evaluation, presumably started by [90] in a context of Federated Learning is the deletion diagnostic used to evaluate the difference between the selected metric of the model trained with and without the client i, i.e.

$$Con\,(C_i) = V\,(S \cup \{C_i\}) - V\,(S) \tag{8}$$

where V is a selected value function (e. g. the model accuracy on a centralized test set). This way we can measure the difference that agent i is making on the aggregated (global) model. The [73] has extended this measure further by employing Data Shapley, an equation based on the Shapley Value [72]. This method was further used elaborated on in [91-93]. If we use Shapley Value for

calculating the client contribution, then for each agent

$$Con\,(C_i) = \frac{1}{n} \sum_{C_s \subseteq C \setminus \{i\}} \binom{|C| - 1}{k}^{-1} \left[V\,(S \cup \{C_i\}) - V\,(S)\right] \quad (9)$$

Formula (9) can be interpreted as simply calculating marginal contribution of client i in every possible combination of coalitions. Some authors have also moved beyond the Shapley value, proposing other methods of establishing individual contributions [3, 27, 47, 89]. The deletion diagnostic is a static method of contribution evaluation. Therefore, it requires computing a model a number of times, which is far from a feasibility criterion posted in the real-life implementations. Additionally, the usage of Shapley value itself to calculate individual contributions may pose additional challenges, as the axiomatization of Shapley value presented in [72] may not be necessarily compatible with machine learning applications. One possible remedy that comes along consists of either reconstructing the models belonging to particular coalitions from gradients [90] or performing Monte Carlo sampling to reduce the computational complexity [91, 92]. Both methods are explored in the literature and deliver promising results. Resolution of the time complexity problem does not change the fact that the metric itself will only reflect the contribution expressed as the marginal change of the metric quality on the tested against preselected dataset. Thus, deletion diagnostic does not take into consideration the computational resources that are consumed during the collaborative computations nor the usefulness of the global model for the local collaborator.

Finally, we can take into consideration only the computational resources that are consumed during the training phase. The [83] presented some early and general considerations on the quantification of resource allocation in the Federated scenario [84], this works is also reflected by many other authors [27, 78, 87, 88]. It may be up to the debate how this can be reflected in measuring the individual contributions.

Another open issue is connected to the trade-off between individual utility and global utility that was briefly mentioned in the preceding paragraph. As mentioned at the beginning of this subsection, in machine learning we distinguish between the generalised error and the empirical error. In ideal circumstances, we would like the global predictor to have smaller empirical loss on each of the local datasets than the locally trained predictors, i.e.

$$\mathcal{L}_{ELF}(\tilde{h}\left(\overline{X_i}\right) \leq \mathcal{L}_{\mathrm{ELF}}(\tilde{h}_i\left(\bar{X}_{\mathrm{i}}\right) \ \forall\, \overline{X_i} \in C \quad (10)$$

On the other hand, such situation may not occur, as empirical loss of the global hypothesis function may be greater than the local one on a number of local clients, i.e.

$$\exists \bar{X}_{\mathrm{i}} \in C : \mathcal{L}_{\mathrm{ELF}}\tilde{\mathrm{h}}_i\left(\overline{X_i}\right) \geq \mathcal{L}_{ELF}\tilde{\mathrm{h}}_i\left(\bar{X}_i\right) \quad (11)$$

In such cases, the general performance of the global model (hypothesis function) on the local client may seem lower than the performance of the local model. On the other hand, such model may generalise better even if the empirical loss (tested against the local dataset) is greater than the loss of the model trained on local data. Such model may still be more profitable to the individual agent, even if the empirical loss function would suggest otherwise. It is important to acknowledge, that the main goal of the hypothesis function $\tilde{h}(\bar{x})$ is always to predict new samples, which generally will be sampled from the distribution of the population, not the distribution of the client. Even if the empirical loss function seems to be higher for some particular local dataset, as the agent will encounter new samples sampled from the space of the population, the general loss function will behave better than the local one.

Taking into consideration all presented above we can formulate three additional directional proposals regarding data collaboratives. Primarily, each collaboration should be reflected in the fair division of shares between the collaborators. While those shares do not entitle to monetary compensation per se, they may be used – for example – as a quantification of the voting power of a particular participant. This way, while still formally being collaborative, we implement a self-sovereignty mechanism for controlling commons. Secondly, the calculation of shares could not only reflect the contribution brought into the model (defined as a marginal increase in a model's performance) but also consumed resources or other utilities spent during the training phase. Only hypothetically, one could open collaboratives for participants, who bring contributions not in terms of data, but in terms of raw processing power. This could also lead to distinguishing different roles inside the collaborative itself. Whether such a solution should also be considered under the umbrella term introduced in this article is an open matter for debate. Although we want to signalise that the 'collaboration' may go beyond the action of sharing personal data, we want to limit the scope of cooperation only to the process of sharing the latter for the sake of this article. Lastly, collaboratives should measure not only the objective contribution of each member but also the local utility that is obtained by each of its members. The local utility may be measured, for example, in the distance between the performance of the global model on the other models' data sets and the local data set. The local utility could be further offset by the consumed resources – if the local utility is excessively below the consumed resources, each participant should have the opportunity to opt-out from such a collaboration.

It can be noticed that all of the directional proposals are derived directly from the four essential principles presented before. Those essential principles are – in turn – derived from Ostrom's design principles for collective self-managing institutions. In this way, the proposed concept is not an arbitrary proposal, but a set of implementable solutions that are based on a well-established set of guidelines, while the implementation details are not constrained in any way by the imposed set of definitions. The transition between those three layers is even more manageable given the fact that Ostrom's principles were already explained in the context of data governance [2, 11, 64].

5 CONCLUSION

This paper took an interdisciplinary perspective on how to solve tensions resulting from diverging interests in data by providing insights in both, the regulatory as well as technological layer in data governance. First, it analysed the regulatory layer and the debates among scholars and policy-makers on how the seemingly opposed objectives of enhancing control over data while at the same time ensuring its availability to different stakeholders could be reconciled via a legal framework. Whilst property rights were discussed for several decades as a potential mechanism to empower

individuals and companies with regard to 'their' data and to thereby enable the emergence of a data-agile economy, this approach has been abandoned by now. Instead, a 'paradigm shift' in EU policy has taken place from (legal) data ownership towards facilitating data access and data sharing.

Turning to the technological layer, we have framed data collaboratives as a commons management problem, therefore seeking a solution that may enable stakeholders to retain control over their local data while also allowing them to participate in the exchange of utilities (computational power) and collaboration (collaborative computations). We have presented our own novel approach to the problem of defining data collaboratives, introducing the notion of collaborative computation and four essential principles as a cornerstone of our definition. This approach exemplifies how technological solutions contribute to reconcile the seemingly opposed objectives of simultaneously achieving data control and data availability. The main difference between the standard use case of federated learning (or any other decentralised learning paradigm) and the solution proposed here is that in the data collaborative, there is no central entity that profits from the training and deployment of the model - every collaborative entity has its own interest in the model sharing and without the collaborative effort - the model would probably never be created. It is the sum of common interests, combined computational power and granted access to the data that has given rise to the model creation (shared model governance). The four essential principles encapsulate the idea of shared data governance through collaborative partnership.

Whilst it was argued that data collaboratives can contribute to reduce tensions that result from the objectives of data control and availability, several open questions remain. Even though raw data does not have to be transferred among participants in the collaborative, it might remain possible to make inferences about the individual contributions of participants. This might consequently negatively impact their respective privacy or intellectual property interests. Furthermore, this article did not address the question on how to incentivize stakeholders to participate in the collaborative. More research is thus necessary to further elucidate how data collaboratives can be implemented in practice to resolve tensions in the data economy – requiring an interdisciplinary perspective from law and data science.

We hope to draw more attention towards this topic, as it may prove to become a practical solution to problems that are unsolvable by classical regulatory instruments. Moreover, the extensive work on the contribution index measurements and data quality evaluation that was carried out by the community in the last few years (including the work currently carried out by some of the authors of this article) may result in wider adoption of the collaborative data sharing and model training. While the concept is still at a very early stage of development, we further encourage the community to explore it and assess whether it could be suitable for their use cases and scenario. The success of data collaboratives certainly does not depend as much on the regulatory and technological layers (as those already provide suitable solutions), as on the wider acceptance of the alternative data governance method in the community.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Arianna Rossi from University of Luxembourg and Prof. Dr. Giovanni Comande' from Scuola Superiore Sant'Anna for their helpful feedback. The research is part of the Legality Attentive Data Scientists project. The Project has received funding from the European Union's Horizon Marie Skłodowska-Curie Actions (MSCA) 2020 Innovative Training Networks (ITN). Grant Agreement ID: 956562

REFERENCES

[1] Rene Abraham, Johannes Schneider, and Jan vom Brocke. 2019. Data governance: A conceptual framework, structured review, and research agenda. Int. J. Inf. Manag. 49, (December 2019), 424–438. DOI:https://doi.org/10.1016/j.ijinfomgt.2019.07.008

[2] Ada Lovelace Institute. 2021. Exploring principles for data stewardship — a case study analysis. Retrieved from https://docs.google.com/spreadsheets/d/1hAN8xMJuxobjARAWprZjtcZgq1lwOiFT7hf2UsiRBYU/edit#gid$=$432908716

[3] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, and Antonio Ferrara. 2019. Towards Effective Device-Aware Federated Learning. Retrieved November 28, 2022 from http://arxiv.org/abs/1908.07420

[4] Kalinda Basho. 2000. The Licensing of Our Personal Information: Is It a Solution to Internet Privacy? Calif. Law Rev. 88, 5 (2000), 1507–1545. DOI:https://doi.org/10.2307/3481264

[5] Monik Raj Behera, Sudhir Upadhyay, Suresh Shetty, and Robert Otter. 2021. Federated Learning using Peer-to-peer Network for Decentralized Orchestration of Model Weights. DOI:https://doi.org/10.36227/techrxiv.14267468.v1

[6] Yochai Benkler. 2014. Between Spanish Huertas and the Open Road: A Tale of Two Commons? In Governing Knowledge Commons. Oxford University Press, 69–98.

[7] Andreas Boerding, Nicolai Culik, Christian Doepke, Thomas Hoeren, Tim Juelicher, Charlotte Roettgen, and Max Schoenfeld. 2018. Data Ownership—A Property Rights Approach from a European Perspective. J. Civ. Law Stud. 11, 2 (2018), 323–369.

[8] Bridge, Michael. 2015. Personal Property Law (Fourth Edition ed.). Oxford University Press, Oxford, New York.

[9] Mingqing Chen, Rajiv Mathews, Tom Ouyang, and Françoise Beaufays. 2019. Federated Learning Of Out-Of-Vocabulary Words. DOI:https://doi.org/10.48550/arXiv.1903.10635

[10] Julie E Cohen. 2000. Examined Lives: Informational Privacy and the Subject as Object. Stanford Law Rev. 52, (2000), 1373–1438.

[11] Diane Coyle, Stephanie Diepeveen, Julia Wdowin, Lawrence Kay, and Jeni Tennison. 2020. The Value of Data - Policy Implications. Bennett Institute for Public Policy, Cambridge. Retrieved from https://www.bennettinstitute.cam.ac.uk/wpcontent/uploads/2020/12/Value_of_data_Policy_Implications_Report_26_Feb_ok4noWn.pdf

[12] Bart Custers and Gianclaudio Malgieri. 2022. Priceless Data: Why the EU Fundamental Right to Data Protection is at Odds with Trade in Personal Data. Comput. Law Secur. Rev. 45, (July 2022), 105683. DOI:https://doi.org/10.1016/j.clsr.2022.105683

[13] Paul De Hert. 2022. 'Post-GDPR lawmaking in the Digital Data Society: mimesis without integration. Topological understandings of twisted boundary setting in EU data protection law'. In Data at the Boundaries of European Law, Deirdre Curtin and Mariavittoria Catanzariti (eds.). Oxford University Press, Oxford.

[14] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc' aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Ng. 2012. Large Scale Distributed Deep Networks. In Advances in Neural Information Processing Systems, Curran Associates, Inc. Retrieved August 31, 2022 from https://papers.nips.cc/paper/2012/hash/6aca97005c68f1206823815f66102863-Abstract.html

[15] Josef Drexl. 2017. Designing Competitive Markets for Industrial Data - Between Propertisation and Access. JIPITEC 8, 4 (2017), 257–292. DOI:https://doi.org/10.2139/ssrn.2862975

[16] Josef Drexl, Reto Hilty, Luc Desaunettes, Franziska Greiner, Daria Kim, Heiko Richter, Gintare Surblyte, and Klaus Wiedemann. 2016. Data Ownership and Access to Data - Position Statement of the Max Planck Institute for Innovation and Competition of 16 August 2016 on the Current European Debate. Max Planck Inst. Innov. Compet. Res. Pap. 16, 10 (2016). DOI:https://doi.org/10.2139/ssrn.2833165

[17] European Commission. 2017. Building a European Data Ecoomy.

[17] European Commission. 2017. Building a European Data Economy.

[18] European Commission. 2020. A European Strategy for Data.

[19] European Commission. 2021. Proposal for a Regulation laying down harmonised rules on artificial intelligence (AI Act).

[20] European Commission. 2022. Common European Data Spaces.

[21] European Commission. 2022. Proposal for a Regulation on European Health Data Space.

[22] European Commission. 2022. Proposal for a Regulation on harmonised rules on fair access to and use of data (Data Act) COM/2022/68.

[23] European Union. 2022. Regulation (EU) 2022/868 of the European Parliament and of the Council on European data governance and amending Regulation (EU) 2018/1724.

[24] European Union. 2022. Regulation (EU) 2022/1925 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

[25] European Union. 2022. Regulation (EU) 2022/2065 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[26] Amin Fadaeddini, Babak Majidi, and Mohammad Eshghi. 2020. Secure decentralized peer-to-peer training of deep neural networks based on distributed ledger technology. J. Supercomput. 76, 12 (December 2020), 10354–10368. DOI:https://doi.org/10.1007/s11227-020-03251-9

[27] Zhenan Fan, Huang Fang, Zirui Zhou, Jian Pei, Michael P. Friedlander, and Yong Zhang. 2022. Fair and efficient contribution valuation for vertical federated learning. Retrieved January 27, 2023 from http://arxiv.org/abs/2201.02658

[28] Eros Fani'. 2021. On the Challenges of Class Imbalance in Federated Learning for Semantic Segmentation. laurea. Politecnico di Torino. Retrieved February 3, 2023 from https://webthesis.biblio.polito.it/20566/

[29] Brett Frischmann, Michael Madison, and Katherine Strandburg. 2014. Governing Knowledge Commons. Oxford University Press.

[30] Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. DOI:https://doi.org/10.48550/arXiv.1904.02868

[31] GovLab. 2020. Wanted: Data Stewards - (Re-) Defining the Roles and Responsibilities of Data Stewards for an Age of Data Collaboration.

[32] Max von Grafenstein. 2022. Reconciling Conflicting Interests in Data through Data Governance. An Analytical Framework (and a Brief Discussion of the Data Governance Act Draft, the Data Act Draft, the AI Regulation Draft, as well as the GDPR). HIIG Discussion Paper Series 2022-2. DOI:https://doi.org/10.5281/zenodo.6457735

[33] Malte Grützmacher. 2016. Dateneigentum - ein Flickenteppich. Comput. Recht 32, 8 (2016). DOI:https://doi.org/10.9785/cr-2016-0803

[34] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2019. Federated Learning for Mobile Keyboard Prediction. ArXiv181103604 Cs (February 2019). Retrieved February 10, 2022 from http://arxiv.org/abs/1811.03604

[35] Garrett Hardin. 1968. The Tragedy of the Commons. Science 162, 3859 (1968).

[36] P. Bernt Hugenholtz. 2017. Data Property in the System of Intellectual Property Law: Welcome Guest or Misfit? Nomos, 73–100. DOI:https://doi.org/10.5771/9783845288185-73

[37] P. Bernt Hugenholtz. 2018. Against 'Data Property.' In Kritika: Essays on Intellectual Property. Edward Elgar Publishing, 48–71. DOI:https://doi.org/10.4337/9781788971164.00010

[38] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and Open Problems in Federated Learning. ArXiv191204977 Cs Stat (March 2021). Retrieved February 11, 2022 from http://arxiv.org/abs/1912.04977

[39] Wolfgang Kerber. 2017. Rights on Data: The EU Communication "Building a European Data Economy" from an Economic Perspective. In Trading Data in the Digital Economy: Legal Concepts and Tools. Nomos, 109–134.

[40] Wolfgang Kerber. 2021. From (Horizontal and Sectoral) Data Access Solutions Towards Data Governance Systems. In Data Access, Consumer Interests and Public Welfare. Nomos, 441–476.

[41] Wolfgang Kerber. 2022. Governance of IoT Data: Why the EU Data Act will not Fulfill Its Objectives. GRUR Int. (2022), 1–16. DOI:https://doi.org/10.1093/grurint/ikac107

[42] Jakub Konečný, Brendan McMahan, and Daniel Ramage. 2015. Federated Optimization:Distributed Optimization Beyond the Datacenter. ArXiv151103575 Cs Math (November 2015). Retrieved February 11, 2022 from http://arxiv.org/abs/1511.03575

[43] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2017. Federated Learning: Strategies for Improving Communication Efficiency. DOI:https://doi.org/10.48550/arXiv.1610.05492

[44] N. Kourtellis, Kleomenis Katevas, and Diego Perino. 2020. FLaaS: Federated Learning as a Service. (2020). DOI:https://doi.org/10.1145/3426745.3431337

[45] Lawrence Lessig. 1999. The Architecture of Privacy: Remaking Privacy in Cyberspace. Vanderbilt J. Entertain. Technol. Law 1, 1 (1999), 56–65.

[46] Jessica Litman. 2000. Information Privacy/Information Property. Stanford Law Rev. 52, (2000), 1283–1313.

[47] Hongtao Lv, Zhenzhe Zheng, Tie Luo, Fan Wu, Shaojie Tang, Lifeng Hua, Rongfei Jia, and Chengfei Lv. 2021. Data-Free Evaluation of User Contributions in Federated Learning. Retrieved November 28, 2022 from http://arxiv.org/abs/2108.10623 [48] Orla Lynskey. 2015. The Foundations of EU Data Protection Law. Oxford University Press, Oxford.

[48] Michael Madison. 2020. Tools for Data Governance. Technol. Regul. (2020), 29–43. DOI:https://doi.org/10.26116/TECHREG.2020.004

[49] Michael J Madison, Brett M Frischmann, and Katherine J Strandburg. 2010. Reply: The Complexity of Commons. Cornell Law Rev. 95, 4 (2010), 839–550.

[50] Dylan Mäenpää. Towards Peer-to-Peer Federated Learning: Algorithms and Comparisons to Centralized Federated Learning.

[51] Maria Rosaria Marella. 2017. The Commons as a Legal Concept. Law Crit. 28, 1 (2017), 61–86.

[52] Bertin Martens. 2021. Data Access, Consumer Interests and Social Welfare - An Economic Perspecctive on Data. In Data Access, Consumer Interets and Public Welfare. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden, 69–102.

[53] Ugo Mattei. 2000. Basic Principles of Property Law: A Comparative Legal and Economic Introduction. Greenwood Publishing Group.

[54] Ugo Mattei and Alessandra Quarta. 2017. Théorie du Droit. Principles of Legal Commoning. Rev. Jurid. L'Environnement 42, 1 (2017), 67–81.

[55] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. ArXiv160205629 Cs (February 2017). Retrieved February 10, 2022 from http://arxiv.org/abs/1602.05629

[56] Maria Lillà Montagnani and Antonia von APPEN. 2021. IP and data (ownership) in the new European strategy on data. Eur. Intellect. Prop. Rev. 43, 3 (2021).

[57] Yves-Alexandre de Montjoye, Erez Shmueli, Samuel S. Wang, and Alex Sandy Pentland. 2014. openPDS: Protecting the Privacy of Metadata through SafeAnswers. PLOS ONE 9, 7 (July 2014), e98790. DOI:https://doi.org/10.1371/journal.pone.0098790

[58] Guido Noto La Diega. 2023. Ending Smart Data Enclosures: The European Approach to the Regulation of the Internet of Things between Access and Intellectual Property. In The Cambridge Handbook on Emerging Issues at the Intersection of Commercial Law and Technology. Cambridge University Press.

[59] OECD. 2015. Data-Driven Innovation: Big Data for Growth and Well-Being. OECD Publishing, Paris.

[60] OECD. 2016. Maximising the Economic and Social Value of Data: Understanding the Benefits and Challenges of Enhanced Data Access. OECD. DOI:https://doi.org/10.1787/9789264229358-en

[61] Elinor Ostrom. 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.

[62] Przemysław Pałka. 2020. Data Management Law for the 2020s: The Lost Origins and the New Needs. Buffalo Law Rev. 68, 2 (2020), 559–640. DOI:https://doi.org/10.2139/ssrn.3435608

[63] Barbara Prainsack. 2019. Logged out: Ownership, exclusion and public value in the digital data and information commons. Big Data Soc. 6, 1 (2019). DOI:https://doi.org/10.1177/2053951719829773

[64] Nadezda Purtova. 2010. Property in Personal Data: A European Perspective on the Instrumentalist Theory of Propertisation. Eur. J. Leg. Stud. 2, (2010), 193.

[65] Nadezhda Purtova. 2015. The Illusion of Personal Data as no One's Property. Law Innov. Technol. 7, 1 (2015), 83–111. DOI:https://doi.org/10.1080/17579961.2015.1052646

[66] Nadezhda Purtova. 2017. Health Data for Common Good: Defining the Boundaries and Social Dilemmas of Data Commons. In The Interplay between eHealth and Surveillance. Springer, 177–210.

[67] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. 2019. Federated Learning for Emoji Prediction in a Mobile Keyboard. DOI:https://doi.org/10.48550/arXiv.1906.04329

[68] Christian Reimsbach-Kounatze. 2021. Enhancing Access to and Sharing of Data: Striking the Balance between Openness and Control over Data. In Data Access, Consumer Interests and Public Welfare. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden, 27–68.

[69] Paul M. Schwartz. 2004. Property, Privacy, and Personal Data. Harv. Law Rev. 117, 7 (2004), 2056–2128. DOI:https://doi.org/10.2307/4093335

[70] Shai Shalev-Shwartz and Shai Ben-David. 2014. Understanding Machine Learning: From Theory to Algorithms (1st ed.). Cambridge University Press. DOI:https://doi.org/10.1017/CBO9781107298019

[71] Yashothara Shanmugarasa, Hye-Young Paik, Salil S. Kanhere, and Liming Zhu. 2021. Towards Automated Data Sharing in Personal Data Stores. In 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 328–331. DOI:https://doi.org/10.1109/PerComWorkshops51409.2021.9431001

[72] Lloyd S. Shapley. 1952. A Value for N-Person Games. RAND Corporation. Retrieved November 25, 2022 from https://www.rand.org/pubs/papers/P295.html
[73] Tianshu Song, Yongxin Tong, and Shuyue Wei. 2019. Profit Allocation for Federated Learning. In 2019 IEEE International Conference on Big Data (Big Data), 2577–2586. DOI:https://doi.org/10.1109/BigData47090.2019.9006327
[74] Ivan Stepanov. 2020. Introducing a property right over data in the EU: the data producer's right – an evaluation. Int. Rev. Law Comput. Technol. 34, 1 (January 2020), 65–86. DOI:https://doi.org/10.1080/13600869.2019.1631621
[75] Thomas Streinz. 2021. The Evolution of Europen Data Law. In The Evolution of EU Law, Paul Craig and Gráinne De Búrca (eds.). Oxford University Press, 902–936.
[76] Linnet Taylor, Hellen Mukiri-Smith, Tjaša Petročnik, Laura Savolainen, and Aaron Martin. 2022. (Re)making data markets: an exploration of the regulatory challenges. Law Innov. Technol. 0, 0 (August 2022), 1–40. DOI:https://doi.org/10.1080/17579961.2022.2113671
[77] Florent Thouvenin, Rolf H. Weber, and Alfred Früh. 2017. Data Ownership: Taking Stock and Mapping the Issues. In Frontiers in Data Science (1st ed.), Matthias Dehmer and Frank Emmert-Streib (eds.). CRC Press, 111–145. DOI:https://doi.org/10.1201/9781315156408-4
[78] Silvana Trindade, Luiz F. Bittencourt, and Nelson L. S. da Fonseca. 2022. Resource management at the network edge for federated learning. Digit. Commun Netw. (October 2022). DOI:https://doi.org/10.1016/j.dcan.2022.10.015
[79] Sjef Van Erp. 2019. Comparative Property Law. In The Oxford Handbook of Comparative Law (2nd edn). Oxford University Press, Oxford.
[80] Praneeth Vepakomma, Tristan Swedish, Ramesh Raskar, Otkrist Gupta, and Abhimanyu Dubey. 2018. No Peek: A Survey of private distributed deep learning. Retrieved August 10, 2022 from http://arxiv.org/abs/1812.03288
[81] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S. Rellermeyer. 2020. A Survey on Distributed Machine Learning. ACM Comput. Surv. 53, 2 (March 2020), 30:1-30:33. DOI:https://doi.org/10.1145/3377454
[82] Salomé Viljoen. 2021. A Relational Theory of Data Governance. Yale Law J. 131, 2 (2021), 573–654. DOI:https://dx.doi.org/10.2139/ssrn.3727562
[83] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. 2021. A Field Guide to Federated Optimization. ArXiv210706917 Cs (July 2021). Retrieved May 12, 2022 from http://arxiv.org/abs/2107.06917
[84] Janis Wong, Tristan Henderson, and Kirstie Ball. 2022. Data Protection for the Common Good: Developing a Framework for a Data Protection-Focused Data Commons. Data Policy 4, (2022). DOI:https://doi.org/doi:10.1017/dap.2021.40
[85] Herbert Zech. 2016. Data as a Tradeable Commodity. In European Contract Law and the Digital Single Market. Intersentia, Cambridge, 51–79.
[86] Herbert Zech. 2021. Exclusivity in data: How to best combine the patchwork of applicable European legal instruments. In Research Handbook on Information Law and Governance. Edward Elgar Publishing, 69–76.
[87] Yufeng Zhan, Peng Li, and Song Guo. 2020. Experience-Driven Computational Resource Allocation of Federated Learning by Deep Reinforcement Learning. In 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 234–243. DOI:https://doi.org/10.1109/IPDPS47924.2020.00033
[88] Hao Zhang, Tingting Wu, Siyao Cheng, and Jie Liu. 2022. CCFL: Computationally Customized Federated Learning. DOI:https://doi.org/10.48550/arXiv.2212.13679
[89] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. 2020. Hierarchically Fair Federated Learning. Retrieved November 23, 2022 from http://arxiv.org/abs/2004.10386
[90] Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. 2019. Measure Contribution of Participants in Federated Learning. Retrieved November 24, 2022 from http://arxiv.org/abs/1909.08525"
[91] Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. 2022. GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning. ACM Trans. Intell. Syst. Technol. 13, 4 (May 2022), 60:1-60:21. DOI:https://doi.org/10.1145/3501811
[92] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. 2020. A Principled Approach to Data Valuation for Federated Learning. In Federated Learning: Privacy and Incentive, Qiang Yang, Lixin Fan and Han Yu (eds.). Springer International Publishing, Cham, 153–167. DOI:https://doi.org/10.1007/978-3-030-63076-8_11
[93] Shuyuan Zheng, Yang Cao, and Masatoshi Yoshikawa. 2023. Secure Shapley Value for Cross-Silo Federated Learning. DOI:https://doi.org/10.14778/3587136.3587141