How might technology rise to the challenge of data sharing in agri-food?

Aiden Durrant[a,b,*], Milan Markovic[b,*], David Matthews[c], David May[d], Georgios Leontidis[b,**], Jessica Enright[a]

[a] *School of Computing Science, University of Glasgow, G12 8RZ, Glasgow, United Kingdom*
[b] *Department of Computing Science, University of Aberdeen, AB24 3UE, Aberdeen, United Kingdom*
[c] *Upton Beach Consulting Limited*
[d] *Lincoln Institute for Agri-food Technology, University of Lincoln, LN2 2LG, Lincoln, United Kingdom*

Abstract

Data sharing is often hindered by a number of real word challenges caused by a mixture of technological and social factors. To date, the agri-food sector significantly lags behind other sectors in overcoming these challenges. However, the benefits of data sharing are too great to be ignored as they have a potential to address many historical failings such as issues related to food safety, traceability and transparency, and must be carefully considered as the sector is undergoing a widespread digitalisation. In this article, we explore the potential of different technologies in addressing the challenges presented by data sharing in the agri-food sector, and how the use of these technologies in the narrative of a Data Trust may address many of these obstacles. We argue the importance of utilising semantic web technologies, distributed ledger technologies, machine learning, and privacy preserving technologies to enable future transformative data sharing infrastructures in the agri-food sector. The utilisation of holistic statistical analysis of the shared data is also discussed, vital in supporting many of the sectors optimisation and sustainability goals.

Keywords: Data Trusts, Data Sharing, AI Technologies, Agri-food Supply

[*] authors contributed equally to this work
[**] Corresponding author
Email address: `georgios.leontidis@abdn.ac.uk` (Georgios Leontidis)

Chains

1. Introduction

Data sharing poses many opportunities to address the historical failings impacting the agri-food supply chain, such as traceability and transparency in cases of food fraud [63, 56]. Many domains have seen great success from data sharing, notably genomic research, which for the last two decades have shared large quantities of data enabling extensive analysis of rare diseases, only possible through collaborative sharing [69, 13]. Other domains, in particular ones that have a more sensitive commercial component, such as drug discovery and pharmaceuticals, could also benefit from data sharing (e.g. in the need for rapid solutions to the Covid-19 pandemic); however, in such commercial settings, where intellectual property might be worth hundred of millions of pounds, sharing data might be seen as a less attractive aspect, resulting in reluctance to adopt data sharing principles.

In addition, data sharing mechanisms are easily impeded by technical issues related to e.g. data quality, transparency and privacy protection concerns, and interoperability issues, as well as a host of social considerations including inequality and complex power dynamics [66, 37, 72]. Data sharing therefore presents both a challenge and an opportunity to the agri-food sector. Opportunities include increased transparency and traceability (whether to address food safety concerns or to meet consumer demand) [29], the potential for increased overall system efficiency via data science insights [60], and a general increase in cooperation and decrease in administrative friction. Beyond the technical challenges, which are the focus of this article, there are other major blockers in understanding ownership of data and correctly incentivising sharing by providing significant benefits. For example, reluctant partners can easily stymie data sharing, particularly if they are powerful actors in the food system [38].

The primary gains possible from data sharing depend on the style and purpose of the sharing. One model to help support sharing is that of a Data Trust, in which independent, fiduciary stewardship of data is provided [39]. In a regulatory setting where data sharing is required for compliance purposes, Data Trusts could remove some data maintenance and storage burden from industrial actors. However, perhaps more importantly Data Trusts could increase the speed of traceability for food safety related incidents and product recalls. In a wider setting of sharing production or pre-competitive data, there is significant potential for improved efficiency and profit, both for the sector as a whole and for individual operators. To focus our discussion more concretely and meet our objectives, we consider two particular use-cases for data sharing within agri-food, aligning with these two example types of gains, and outline some associated challenges and opportunities. Finally, the objectives of this article are a) to explore the potential of different technologies – some established and others emerging – in addressing the challenges presented by data sharing in the agri-food sector; and b) to discuss how these technologies, within the context of a Data Trust, can be used in practice to overcome the obstacles associated with data sharing.

1.1. Food and Drink Sector

The UK Food and Drink sector is the largest manufacturing sector in the whole country, valued at about £100bn [62]; it is larger than automotive and aerospace combined. In 2018, total food and drink export figures were worth more than £23bn, with the top three export markets being Ireland, USA and France. The food supply chain employs about 4 million people and generated over £121bn of added value for the economy each year. It is largely a small and medium enterprises (SME) sector, as 97% of food and drink businesses are SMEs [62]. Although the effect of Brexit is still unclear, it is estimated that more than 100,000 new recruits will be needed by 2024 to feed the ever-growing population and meet market demand.

UK Food and Drink sector includes several sub-sectors, e.g. animal feeds, bakery, confectionary, dairy, fish, fruit and vegetables, meat, etc., with the top 10 export products being whisky, salmon, chocolate, wine, cheese, gin, beef, pork, breakfast cereals and beer. Such is the growth of the sector that several of these products have seen more than a 10% increase on export demand in the last couple of years [62].

1.2. Use Case 1: Traceability beyond one-up and one-down

Traceability of goods within the food system is desireable from a number of perspectives, including regulatory requirements (e.g. in the case of bovine livestock in Europe [19]), for food safety monitoring, and to satisfy consumer demand. Currently, apart from systems where central government reporting is required, a system of 'one-up-one-down' traceability is the norm [73]: here each actor in a supply chain will know the source and the destination of their products or goods, but will not know the next eventual destination or the original source. While this system does preserve the necessary information to reconstruct a product history through contacting all actors in the chain, it is cumbersome and may be too slow in a food safety emergency.

1.3. Use Case 2: Sharing data for production optimisation

Insights leading to improved production efficiency may be possible from data analytic approaches on production data either from learned insights from the overall data or by an actor comparing their efficiency directly to their peers in an effort to identify areas of possible improvement [79]. Both of these would naively seem to require direct pooling of data between competitors, which data holders may be uncomfortable with. However, as we expand on below, there are a number of technological approaches that may allow some of this advantage to be gained while protecting individual commercial sensitivities.

1.4. Challenges to data sharing

Several social challenges to data sharing are reported in an Open Data Institute report describing a pilot study of a Data Trust for reducing food waste [37]. In particular, they report that businesses feared that sharing data might reveal commercially sensitive information to competitors, or that it might result in bad publicity for the sharer, and hence pose a reputational risk. Overall, they noted that manufacturers and retailers saw little value in sharing data about food waste, feeling that the benefits were insufficient to justify the competitive and reputational risks and the effort required to prepare the data.

Trust and power between actors play a significant role in the uptake and ethical value of data sharing, even where legislation designed to support equitable sharing is in place [72]. Where legislation is limited or absent, the situation may be even more untenable - agricultural actors have sometimes been reluctant to share data over a lack of trust with those who are gathering, collating, and sharing the data, and uncertainty about how data will be used and shared eventually [78, 40]. It is clear that if actors in agri-food systems are to willingly participate in data sharing (via Data Trusts or otherwise) we must mitigate these concern, and the benefits of the sharing must be sufficient to overcome these social challenges.

In addition to critical social challenges, there are a variety of technological challenges to effective data sharing [29, 15, 33]. These span the entire data lifecycle and relate to the data format and encryption levels used by Data Trusts, security and transparency of data access, and ultimately the utilisation of data assets through analysis and as evidence. A pre-agreed data sharing system for traceability could significantly improve the efficiency and reliability of product tracing, regardless of whether the data required resided in a central location or locally with actors under careful federated agreements. However, widespread adoption of digital technologies including data sharing in sectors such as Agri-Food have yet to materialise; indeed almost 60% of UK farm data are paper-based [3].

In this paper, we expand on opportunities and challenges of Data Trusts, in particular discussing how Data Trusts and various technological approaches to data integration might address them.

2. Related work on data trusts and data sharing

Data sharing has been largely enabled through the advent of cloud-based technologies, with the competition between large players pushing the limits in infrastructural capabilities and costs. Across Microsoft Azure [1], Amazon Web Services [2] and Google Cloud Computing [3], one can find a plethora of cloud-based solutions spanning end-to-end pipelines, i.e. from data input and aggregation to model development and business intelligence. Such pipelines allow for a streamlined process of keeping data in remote locations - in a fault tolerant manner - enabling companies to rely less on local infrastructure.

Along the lines of automation and cloud computing, cloud manufacturing has become popular over the past few years, as a resource sharing paradigm [54]. Through cloud manufacturing, businesses can have remote access to a pool of manufacturing resources and capabilities, akin to other cloud-based resources. This new paradigm enables businesses to leverage the Industrial Internet of Things (IIoT) and its underlying infrastructure as a service including architecture models, and data and information exchange protocols. In this process, interoperability is a key component as it allows the implementation of vertically or horizontally integrated cyber-physical systems for production engineering [54].

In 2018, the EU launched the EU code of conduct for agricultural data sharing by contractual agreement, which encourages transparency about data use [72]. The scope of this EU code has been to enable trust through the establishment of contractual agreements between the parties concerned with the data exchange process. In such cases, fostering trust goes beyond contractual agreements, whereby power relationships might be influencing the direction these agreements might take, as argued in [72]. Therefore contractual agreements are only one small component of the processes needed to accommodate data sharing.

[1] https://azure.microsoft.com/en-gb/industries/discrete-manufacturing/

[2] https://aws.amazon.com/

[3] https://cloud.google.com/solutions

ODI [39] have made significant progress in standardising concepts and processes around Data Trusts through their engagement with various stakeholders and UK universities. For example, the UK Biobank [10] has collected and maintained data since 2006; stewardship is achieved via its status as a charitable company with a board of directors that "act as charity trustees under UK charity law and company directors under UK company law" [39]. Other similar examples provided by ODI revolve around data stewardship within the context of a Trust: an independent legal entity requiring a careful framework of agreements. For small enterprises in particular the complexity of the legal framework required can be daunting, and discourage involvement.

Within agri-food, many of the efforts in organised data sharing have included only an informal Data Trust, often within one organisation or company and focussing around one product or chain e.g. multi-IoT based sensor data aggregation as a systems approach [16, 35], with [28] providing a comprehensive overview of how modern IoT and data analytics approaches can enable smart agriculture towards boosting productivity and sustainability. Blockchain technologies have contributed to a number of agri-food data sharing systems [59], and have shown potential value when coupled with other technologies such as IoTs [71] and machine learning [44]. Machine learning approaches and ontologies have been primarily used in the context of agri-food data analytics for precision agriculture [53, 4, 18, 8].

In this diverse technological landscape, privacy preserving approaches have started gaining momentum as components of larger pipelines that might also include blockchain [48] or machine learning [7]. The appeal of privacy preserving technologies is clear: they enable computation despite the fact that raw data are obscured, hence any data exchange occurs with encrypted data. This could be invaluable in overcoming concerns about data privacy, competitiveness, and reputational risk.

Despite existing and emerging technologies, we are still far from converging on an established model of a Data Trust - either legally or technologically. Some reasons that might have led to this are touched upon in [21], where among others, the concept about “one size fits all” in data governance is considered as a hindering factor, going on to suggest that a “plurality of Trusts” could be a way forward. Regulatory concerns and how personal data can be protected within a Data Trust are essential to their adoption, where [66] identifies the necessity of legal and social foundations implementing a ’data protection by design’ philosophy into Data Trusts. This is also considered in [21], which builds upon the “Growing the Artificial Intelligence Industry in the UK” report published by the UK Government in 2017 [32].

Another direction to Data Trust in agri-food revolves around approaches that increase transparency and availability of information as a means to increase trust [61, 40]. In particular, the concept around transparency is considered in the context of sociotechnical factors and conditions that influence the development of smart farming [40], touching upon issues around use of data, sceptisism about the value of smart technologies and balancing expectations within the farming industry. It becomes evident that data sharing quickly becomes a very complex problem when considering all factors involved, let alone in the context of Data Trusts. Regarding adoption issues and lack of a universally accepted Data Trust system, we speculate that the complexity of data collected across sectors, interoperability issues, analytics as a knowledge extraction process, and business intelligence as a decision making mechanism for improving performance, could mean that multiple Data Trusts might need to be proposed to accommodate the specific needs found within individual sectors, along the lines discussed in [21]. As data control processes become more and more complex, Data Trust systems could be developed and tailored to a sub-sector level, adopting some components and/or implementing new ones.

3. Role of technology

A wide variety of technological approaches may be useful in overcoming the aforementioned challenges to data sharing previously outlined, thus allowing the agri-food sector to best take advantage of the opportunities available. Below we mention a number of technologies that we believe have potential to address a spectrum of technological and social complexities at various stages of the data lifecycle.

At the most basic level, data interoperability and reuse will require **agreement on data formats and reproducible data pipelines**. A suitable format for incoming data is crucial; while this point may seem very straightforward, it really is critical to enable value-adding data processing or inference to occur. Machine-readable data that can be accessed by a programmatic interface will allow best value: that is, a structured filetype (e.g. YAML, JSON, HDF5) with an agreed format will allow smoother updates and quicker use of new data than a non-readable or changing format (e.g. a scanned pdf or a spreadsheet with a non-agreed format).

Semantic web technologies can take the description of machine-readable data even further. Data is represented in a form of a graph described using the Resource Description Framework (RDF) [58] and ontology languages such as OWL [34] are used to produce formal models of semantic annotations for a specific domain. Applying such annotations to raw data will produce semantic metadata that can formally describe the types of individual data elements and their relationships to other data represented within the same or any other data set, thus forming a *knowledge graph*. Data described using standard semantic frameworks may be further processed using automated pipelines to infer new knowledge, to integrate with other datasets, or to validate for missing information. For example, SPARQL query language [36] can be used to query and transform data stored in multiple distributed repositories and recently developed standards such as SHACL [45] can be used to define data quality rules to support management of knowledge graphs.

Recently, **distributed ledger technologies (DLTs)** have been gaining traction in agrifood through both commercial platforms (e.g. Food Trust from IBM[4]) and proposed research prototypes and pilot systems [14, 68, 41]. Key

[4] https://www.ibm.com/uk-en/blockchain/solutions/food-trust

characteristics of DLT solutions include the immutability of stored records (i.e., records cannot in retrospect be falsified), transparency of data operations, and decentralised data access and storage. These characteristics make DLTs a strong candidate for applications where data availability is crucial and data is contributed by a range of third parties with varying levels of trust (e.g. within a traceability scenario where there is risk of retrospective record fraud). However, it is important to note that DLTs on their own are not a complete solution to data sharing challenges, which is further discussed in section 4.4.

In the presence of ingested and structured data, **machine learning** approaches can be very good candidates to perform analyses and extract knowledge. Such analyses have shown much success in agri-food, ranging from yield prediction [4] to disease detection [64]. Federated computing and federated machine learning address one of the resulting challenges from the holistic view of data analysis, seeing vast adoption in big data to maintain local ownership [11]. Under such federated computing models, standard data analysis methodologies can be undertaken, sharing only model updates of the training procedure rather than the raw data itself. Additionally, machine learning can assist in the earlier and fundamental stage of interoperability. Outlier detection [24] and imputation [47] are a few proven machine learning methodologies modelling high-dimensional data input to provide high quality and error free data for further analysis.

Privacy preserving technologies (PPT) are emerging technologies that have gained momentum recently and which have the ability to allow knowledge extraction and machine learning without compromising privacy. Various cryptographic schemes have been developed that can accommodate the analysis of data even without revealing information about the raw data at the individual level. Techniques such as fully homomorphic encryption (FHE) [30], secure multiparty computation (SMPC) [74], order preserving encryption (OPE) [2] and differential privacy [1] can be used as part of machine learning end-to-end pipelines, albeit carrying a large overhead. While many of these techniques are still in their early days of application and so the technological barriers could be significant, their potential to contribute to data sharing is very high: they provide a way in which any party in the trust or federation could share encrypted data and compute functions on aggregated data without anyone ever holding enough information to infer any other party's individual data.

4. A call for technologically-mediated data sharing in agri-food

We believe that the advantages of data sharing across a variety of use cases within agri-food justify pursuing Data Trusts for data sharing within the sector, and that technologies can help overcome the challenges associated with trust, data formatting and understanding, and data privacy. As examples of possible systems, we refer back to our two use-cases. Within our **traceability** use case, we can imagine the use of a number of different technologies depending on the regulatory framework and the motivation for tracing. Our **production optimisation** would benefit from a variety of technologies, with the focus depending on the preferences of data sharers and the sensitivity of the data. In the following sections we expand on the potential of several different approaches and technologies, making reference to our use cases throughout.

4.1. From Data Sharing to Model Sharing?

In the context of Data Trusts, the *data* shared typically takes the form of raw or pre-processed data (transformed, aggregated, cleansed, etc.), but not models derived from those data. If our goal is not the data sharing in itself but instead the use case of production optimisation (Section 1.3) via collective intelligence to improve productivity, performance, sustainability, etc. one wonders whether this can be achieved via means other than data sharing, e.g. sharing trained models instead. This process is akin to sharing physical models of a system that describes its behaviour and can be used as a data generation mechanism (simulation systems, see [5]).

Along these lines machine learning approaches and particularly a sub-domain called Deep Learning, have enabled the transfer of knowledge between and within domains through the exchange of trained models. As Deep Learning approaches are based on Artificial Neural Networks, which are high dimensional non-linear models, there has been an abundance of approaches concerning transfer learning and domain adaptation that allow the transfer of knowledge through sharing learned parameters [70, 77].

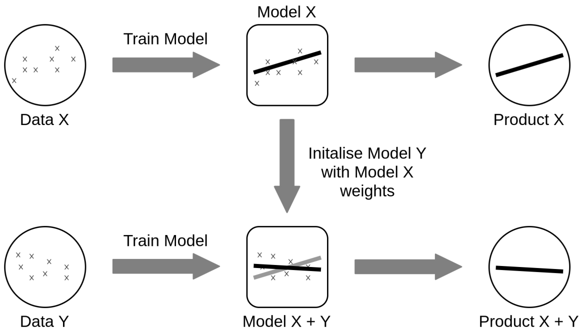


Figure 1: Visual depiction of the model sharing methodology. Model Y is trained with data Y, but its parameters are initialised with those of model X. Black lines represent the statistical model, black crosses represent local data.

In various real-life problems, there exist many high level concepts that are consistent across various domains, e.g. shapes, texture, spatial distribution, etc. Therefore, extracting representations for such concepts from domain X can be very relevant to a cognate domain Y. This line of thinking can be extrapolated to our specific use cases, investigating production optimisation via crop yield data. In this case, data from Farm X can be used to develop a machine learning-based yield forecasting model for Farm X. This high dimensional complex system can be used as a surrogate model, which upon fine-tuning it with data from Farm Y (hence incorporating knowledge from Farm X as well), Farm Y can gain an advantage over Farm X, which has only used its own data (Farm X's data). At no point in this process has Farm Y seen any of the raw data used by Farm X to develop the initial model, which was shared with Farm Y. This process is visually depicted in Figure 1. In addition, given the non-linearities involved along with extensive hyperparametrisation, it becomes a non-trivial problem to reconstruct the exact raw data given that the parameters one is sharing (model sharing) are highly sensitive to hyperparametrisation.

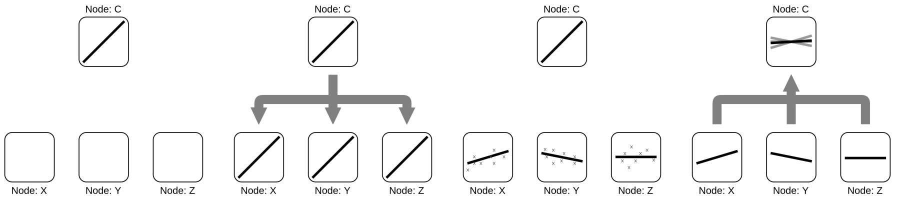
We believe model sharing to be another exciting possibility, that when thinking of Data Trusts as a more holistic approach, could actually be one direction that organisations can take to accommodate knowledge exchange rather than solely focusing on data sharing. That aligns with our earlier statement that "one size fits all" approach might not be the way forward and that agri-food organisations should be aware of the various ways available that can allow for a step-change to occur within this industry. Besides going from 0% to 100% (with 0% being no data/Knowledge sharing/exchange at all, and 100% being fully open data) is a continuum, in that step-wise improvement can take various shapes and make use of the whole spectrum of possible adoption levels that can suit the organisational needs.

4.2. Federated learning for decentralised analysis

Following from model sharing, yet maintaining the concept of our production optimization for actors across the supply chain via a holistic view of data, we return to the notion of more traditional and alternative view of data analytics from a collection of data. We therefore ponder the idea of performing machine learning training to produce one model on a number of independent and decentralised datasets simultaneously, without the exchange of raw data. On the contrary to model sharing, it is simpler to gain a holistic analysis of data when presented with all available data during training, so we ask, can we leverage multiple independent, distributed datasets in the training procedure to produce one model encapsulating data from multiple sources at scale?

This approach, commonly known as federated learning, have shown vast success in large scale industry applications, trained with highly sensitive data, most notably powering predictive texting in our phones [11]. Conceptually, rather than data being centralised, data analysis is performed across all decentralised data stores/nodes simultaneously via aggregation and dissemination of model updates. Such a technological implementation fits hand-in-hand with the proposition of decentralised models of Data Trusts, providing holistic analysis that leverages data across sources whilst maintaining privacy and eliminating raw data sharing.

The previous example of production optimisation via crop yield forecast (Section 1) can be also be applied to here, where Farm X and Farm Y each have their own data, we follow the regime presented in [80] and represented in Figure 2. To train the model, a central node which we will call C, controls the communication and computes the model updates. For each step in the training procedure, C will send a copy of the model parameters to each of the Farm nodes X and Y, each farm node will then perform a forward step on the model (each node has identical models) and the resulting variables are sent to back to C. The C node will aggregate these variables and compute a set of updates based on the data from X and Y, the updates are sent back to each node X and Y to update each model.



(a) Central node initialises the statistical model.

(b) Each node receives the initialised model from the central node.

(c) Each node trains the model locally on their own data.

(d) The individual models are aggregated by the central node to generate one model.

Figure 2: Depiction of centralised federated learning in which no access or sharing of data between nodes is undertaken. The temporal process moves left to right (a-d). Black lines represent the statistical model, and black crosses the local data on each node [76].

This entire procedure eliminates raw data sharing, and allows a single model to be developed leveraging data from all participating datastores. Privacy has been at the forefront of the design and implementation of federated learning, differential privacy and encryption has played a key role in securing individual's data [46], the former of which is elaborated on in Section 4.3. Additionally, the introduction of blockchain technologies have allowed for improved accountability reducing the effectiveness of malicious attacks [22]. Furthermore, scalability is a driving factor, where continuous and low compute updates to the model are trivial, essential in sectors that record frequently. Consequently, this opens the possibility of Internet of Things (IoT) devices (the collection, processing and analysis of data from interconnected devices within a system, for the provision of smart solutions in the argi-food supply chain[42]) being integrated into the federated network seamlessly.

Federated learning and federated computing in general could power the data analytics behind decentralised Data Trusts, operating across all participating organisations and even on IoT devices directly, potentially reducing management impacts on actors [12]. The scalability and adaptability of federated models lends itself to the advancements in technology, not only providing solutions to data sharing issues that have arisen now, but providing a proven platform for expansion as the adoption of data acquisition increases and subsequent technological challenges arise. Additionally, leading from the belief that the "one size fits all" approach may not be the ideal solution, federated learning can be implemented into the argi-food supply chain one step and one model at a time, expanding to new organisational challenges as trust in the Data Trust develops.

4.3. Privacy Preserving Technologies

Regardless of which technologies help achieve traceability and transparency, or improve product optimisation, all beneficiaries gain from increased privacy to help develop trust within data sharing. Simply ensuring that data privacy is maintained addresses the most apparent social implication of commercial sensitivity [37], both between actors within a data sharing environment, but also from malicious actors outside.

Previous technologies in Section 4.1 and 4.2 can be seen as privacy preserving technologies (PPT), following a more implicit approach decoupling data analysis from traditional data centralisation. Yet it is common for more explicit methodologies to preserve privacy to be blended alongside. Differential privacy operates under the notion that the addition of noise to a statistical model of data inhibits an adversary to exactly reverse-engineer sensitive data of an individual [27, 1]. This concept has seen a recent re-emergence in deep learning [6], while it has been commonplace and the forefront of development of federated learning algorithms [17, 75], where the process defining federated learning lends itself to the implementation of differential privacy [31]. As such, the implementation of differential privacy is not only achievable but practical in terms of computational overhead, yet the benefits to ensuring trust in knowledge sharing can be significant.

Extending the concept of privacy preserving statistical data analysis to the most complete case, Fully Homomorphic Encryption (FHE) aims to extract knowledge through statistical models like the former, yet doing so on fully encrypted data without decryption the data first. This maintains privacy throughout the entire process of knowledge extraction, data to result. Successful implementations have shown the potential applicability of FHE in industrial settings [52], with a recent toolkit by IBM providing a platform for development[5]. However, FHE for complex tasks comes with large computational overhead, and subsequent loss of performance due to the necessity of approximation [23], reducing the attractiveness of an approach in its current state. Although the applicability of FHE is non-trivial and still a major research topic, the fundamental principle is very attractive for our propositions, all whilst being easily explainable, assisting in knowledge dissemination and understanding, therefore eliciting trust between actors within a data sharing model.

PPT are not only essential in building trust for data sharing, but also a necessary precaution in a modern world of data sensitivity and regulatory obligations. As such, we believe that PPT lies at the heart of all data sharing, where privacy can be maintained between individuals data whilst holistic and novel knowledge extraction can be achieved assisting in our ongoing challenges related to traceability and product optimisation.

[5] https://fhe-website.mybluemix.net/

4.4. Leveraging Semantic Technologies and Blockchain to Facilitate Centralised and Federated Data Sharing

Whether it is raw data or data models that are to be shared, a suitable underlining data sharing infrastructure must exist. Following from our earlier discussions such infrastructure should support automation based on machine-readable data formats, and secure and trusted data assessment mechanisms.

Blockchain networks (a type of DLT) can be formed to hold data about arbitrary domains [81]. For example, a single transaction recorded on a blockchain can describe an exchange of money between two parties trading an item, but it can also record results of sensor readings such as location of an animal. Blockchain networks have gained popularity for their ability to deliver immutability of shared data, transparency of data transactions, and high data availability due to data duplication across multiple nodes. Certain blockchains such as Ethereum[6] and Hyperledger[7] also allow for deployment of decentralized applications (DApps) where data input and retrieval is managed via so called *smart contracts* [55], which is an implementation of an idea originally proposed by Nick Szabo in 1997[8]. Smart contracts are programs that can execute on the blockchain network and define, for example, the structure of transactions that participating entities can complete on the network (e.g. to record purchase and sale of assets as they are moving through a supply chain). Business blockchain networks also provide functionalities for permissioned data access where actors can be restricted from accessing different parts of the blockchain functionalities and the data it stores. Given our traceability use case scenario, blockchain technologies possess an obvious advantage over the existing systems. For example, in case of an outbreak of a food borne disease such as that caused by E. coli, the real-time availability of data across the whole supply chain is important in order to identify potentially contaminated products and issue recalls. Transparent

[6] https://ethereum.org/en/

[7] https://www.hyperledger.org

[8] https://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/idea.html

records of transactions between different food supply actors and the immutability of records may be also useful for regulators in aiding discovery of potential cases of food fraud.

However, several challenges relating to the use of blockchain have been highlighted to date, including scalability and negative environmental impact due to high carbon footprint [59, 81, 43, 67]. In addition, the blockchains technologies typically operate with only simple data types such as string, integer, date, etc. As such, it may be difficult to integrate data represented by different blockchain data models without further knowledge that is not captured in a machine-understandable form, hence hindering an effective automation of such processes. Given such limitations, it is therefore unfeasible to expect that, for example, a very complex description of a product and the full context in which such product was produced, transported, or stored would be stored on a blockchain. This may prevent traceabillity systems from realising their full potential by aiding investigations, for example, through discovery of common factors (e.g. use of river water) in processing workflows of the entities affected by an outbreak. Similar challenges will apply in case of model sharing and federated learning discussed in previous sections. The shared models and data may be too large to be shared via blockchain networks as well as additional metadata descriptions (e.g., model limitations, biases, algorithms used, etc.) which are required for effective reuse. However, such information may still exist within company records, possibly in a proprietary format and requiring a human expert to be retrieved.

On the other hand, semantic web technologies that exist today stemmed from the idea of the Semantic Web which envisioned a globally interconnected *web of data* where much of the meaning is machine-readable [9]. This resulted in creation of data sharing protocols for RDF graphs, distributed graph repositories, and standards for federated data querying [57], paving a way towards generic data sharing infrastructures based on web technologies. Such technology could then support the traceability scenario in both directions: as a centralised repository where all data is stored within a single knowledge graph; and also as a collection of distributed repositories, each under the control of individual data owners and publishing only a portion of a knowledge graph. Many semantic resources such as vocabularies, ontologies, and taxonomies already exist for different parts of the agri-food sector [26, 25, 51, 50]. However, these are currently mostly used by the research community while the industry actors are influenced by, for example, GS1 family of standards[9]. While data sharing systems based on distributed semantic repositories and federated queries for traceability in food supply chains have been proposed in the past [65], such approaches, on their own, do not address challenges related to trust, transparent data access, and performance including data availability from remote repositories.

We believe that Data Trust solutions for agri-food sector may benefit from both blockchain and semantic technologies. In fact, as we outlined in this sub-section, a certain level of symbiosis between these technologies will be required to deliver effective data sharing mechanisms in the agri-food domain. Immutability of blockchain records may support trust in data sharing mechanisms while semantic web technologies are used to describe the data in machine-understandable form. In order to reduce the amount of data stored on blockchain networks to address scaleability challenges, semantically annotated raw data may be automatically abstracted to form more concise reports [51, 49], or the data may include only digital signatures required to validate less critical data stored as external resources (e.g., large RDF datasets accessed from distributed repositories) [20]. For example, a traceability system may store the most critical information regarding the product on a blockchain to enable basic traceability functionality and also provide information to discover and validate more elaborate description of a production context stored in a third party semantic repository. Blockchains could also store information relating to data access for external semantic repositories and thus increasing transparency of data sharing in agri-food domain.

[9] https://www.gs1.org/standards/barcodes

5. Conclusion

Given the central importance of food systems to global heath, economy, and culture, it is no surprise that vast quantities of data are meticulously collected and collated by a wide variety of actors within these systems. While the volume of data collection within food systems is extensive, data sharing and integration has lagged behind, limiting the added-value that could be made available. Both social and technological challenges have contributed to this lag, and we have discussed the potential for technological approaches to overcome these obstacles.

In addition, given the inherent structure of food supply chains with multiple actors present, then the various challenges mean that it is unlikely that physical data will ever be conveniently co-located and that whatever the overarching legal structure, one would still need technologies to address sharing and interoperability considerations.

While the potential is clear, much work remains to be done to allow the agri-food sector to unlock the value of its collective data. This work will include further investigations into the drivers of data sharing and the key benefits within particular systems, as well as engineering work in building systems that trial data sharing via appropriate technologies. The most exciting future work may combine these two strands of research and involve both study of the actors in these systems and the regulatory frameworks in which they operate, alongside implemented prototypes of systems allowing these actors to understand the benefits of data sharing and contribute to design that fits their needs and priorities. As argued in this article, data sharing is a complex technical and social challenge and each of the technologies discussed in this article has a potential to contribute towards solving only part of this challenge. Therefore, it is very likely that any future data sharing solutions will utilise a combination of multiple technologies resulting in complex socio-technical systems requiring interdisciplinary expertise from various disciplines including computer science, social sciences, and business.

Acknowledgements

This work was supported by an award made by the UKRI/EPSRC funded Internet of Food Things Network+ grant EP/R045127/1. We would also like to thank Mr Steve Brewer and Professor Simon Pearson for supporting the work presented in this paper.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. Order preserving encryption for numeric data. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 563–574, 2004.

[3] Agrimetrics. Farm data is not digital data. `https://agrimetrics.co.uk/blog/whitepaper/barriers-to-agricultural-data-sharing/`. Accessed: 24-11-2020.

[4] Bashar Alhnaity, Simon Pearson, Georgios Leontidis, and Stefanos Kollias. Using deep learning to predict plant growth and yield in greenhouse environments. *arXiv preprint arXiv:1907.00624*, 2019.

[5] APSIM. Apsim: The leading software framework for agricultural systems modelling and simulation. `https://www.apsim.info/about-us/`. Accessed: 24-11-2020.

[6] Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 2019.

[7] Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. A trustworthy privacy preserving framework for machine learning in industrial iot systems. *IEEE Transactions on Industrial Informatics*, 16(9):6092–6102, 2020.

[8] Sahin Aydin and Mehmet Nafiz Aydin. Semantic and syntactic interoperability for agricultural open-data platforms in the context of iot using crop-specific trait ontologies. *Applied Sciences*, 10(13):4460, 2020.

[9] Tim Berners-Lee and Mark Fischetti. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. DIANE Publishing Company, 2001.

[10] UK Biobank. About uk biobank. *Available at h ttps://www. ukbiobank. ac. uk/a bout-biobank-uk*, 2014.

[11] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

[12] Christopher Brewster, Ioanna Roussaki, Nikos Kalatzis, Kevin Doolin, and Keith Ellis. Iot in agriculture: Designing a europe-wide large-scale pilot. *IEEE communications magazine*, 55(9):26–33, 2017.

[13] James Brian Byrd, Anna C Greene, Deepashree Venkatesh Prasad, Xiaoqian Jiang, and Casey S Greene. Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, pages 1–15, 2020.

[14] Miguel Pincheira Caro, Muhammad Salek Ali, Massimo Vecchio, and Raffaele Giaffreda. Blockchain-based traceability in agri-food supply chain management: A practical implementation. In *2018 IoT Vertical and Topical Summit on Agriculture-Tuscany (IOT Tuscany)*, pages 1–4. IEEE, 2018.

[15] Winner Dominic Chawinga and Sandy Zinn. Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research*, 41(2):109–122, 2019.

[16] Dong Chen, Baoguo Wu, Tian'en Chen, and Jing Dong. Development of distributed data sharing platform for multi-source iot sensor data of agriculture and forestry. *Transactions of the Chinese Society of Agricultural Engineering*, 33(1):300–307, 2017.

[17] Olivia Choudhury, Aris Gkoulalas-Divanis, Theodoros Salonidis, Issa Sylla, Yoonyoung Park, Grace Hsu, and Amar Das. Differential privacy-enabled federated learning for sensitive health data. *arXiv preprint arXiv:1910.02578*, 2019.

[18] Sai Sree Laya Chukkapalli, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam, Anupam Joshi, Ravi Sandhu, and Karuna Joshi. Ontologies and artificial intelligence systems for the cooperative smart farming ecosystem. *IEEE Access*, 8:164045–164064, 2020.

[19] Council of European Union. Directive 92/65/eec, 1992.
`https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31992L0065`.

[20] Hao Dai, H. Patrick Young, Thomas J. S. Durant, Guannan Gong, Mingming Kang, Harlan M. Krumholz, Wade L. Schulz, and Lixin Jiang. Trialchain: A blockchain-based platform to validate data integrity in large, biomedical research studies. *CoRR*, abs/1807.03662, 2018.

[21] Sylvie Delacroix and Neil D Lawrence. Bottom-up data trusts: disturbing the 'one size fits all'approach to data governance. *International Data Privacy Law*, 9(4):236–252, 2019.

[22] Harsh Bimal Desai, Mustafa Safa Ozdayi, and Murat Kantarcioglu. Blockfla: Accountable federated learning via hybrid blockchain architecture. *arXiv preprint arXiv:2010.07427*, 2020.

[23] Josep Domingo-Ferrer, Oriol Farràs, Jordi Ribes-González, and David Sánchez. Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges. *Computer Communications*, 140:38–60, 2019.

[24] Rémi Domingues, Maurizio Filippone, Pietro Michiardi, and Jihane Zouaoui. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74:406–421, 2018.

[25] Damion M Dooley, Emma J Griffiths, Gurinder S Gosal, Pier L Buttigieg, Robert Hoehndorf, Matthew C Lange, Lynn M Schriml, Fiona SL Brinkman, and William WL Hsiao. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2(1):1–10, 2018.

[26] Brett Drury, Robson Fernandes, Maria-Fernanda Moura, and Alneu de Andrade Lopes. A survey of semantic web technology for agriculture. *Information Processing in Agriculture*, 6(4):487–501, 2019.

[27] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[28] Olakunle Elijah, Tharek Abdul Rahman, Igbafe Orikumhi, Chee Yen Leow, and MHD Nour Hindia. An overview of internet of things (iot) and data analytics in agriculture: Benefits and challenges. *IEEE Internet of Things Journal*, 5(5):3758–3773, 2018.

[29] Huanhuan Feng, Xiang Wang, Yanqing Duan, Jian Zhang, and Xiaoshuan Zhang. Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits and challenges. *Journal of Cleaner Production*, page 121031, 2020.

[30] Craig Gentry and Shai Halevi. Implementing gentry's fully-homomorphic

encryption scheme. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 129–148. Springer, 2011.

[31] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[32] UK Gov. Growing the artificial intelligence industry in the uk. `https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk`, 2017.

[33] Sam Grabus and Jane Greenberg. The landscape of rights and licensing initiatives for data sharing. *Data Science Journal*, 18(1), 2019.

[34] W3C OWL Working Group. OWL 2 web ontology language document overview (second edition). W3C recommendation, W3C, December 2012. https://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[35] Angappa Gunasekaran, Nachiappan Subramanian, Manoj Kumar Tiwari, Bo Yan, Chang Yan, Chenxu Ke, and Xingchao Tan. Information sharing in supply chain of agricultural products based on the internet of things. *Industrial Management & Data Systems*, 2016.

[36] Steven Harris and Andy Seaborne. SPARQL 1.1 query language. W3C recommendation, W3C, March 2013. http://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[37] Open Data Insitute. Exploring the potential of data trusts in reducing food waste. `https://theodi.org/article/data-trusts-food-waste/`, 2019.

[38] Open Data Insitute. Data sharing in the private sector. `https://theodi.org/article/new-survey-finds-just-27-of-british-businesses-are-sharing-data/`, 2020.

[39] Open Data Insitute. Data trusts. `https://theodi.org/article/data-trusts-in-2020/`, 2020.

[40] Emma Jakku, Bruce Taylor, Aysha Fleming, Claire Mason, Simon Fielke, Chris Sounness, and Peter Thorburn. "if they don't tell us what they do with it, why would we trust them?" trust, transparency and benefit-sharing in smart farming. *NJAS-Wageningen Journal of Life Sciences*, 90:100285, 2019.

[41] Andreas Kamilaris, Agusti Fonts, and Francesc X Prenafeta-Boldύ. The rise of blockchain technology in agriculture and food supply chains. *Trends in Food Science & Technology*, 91:640–652, 2019.

[42] Andreas Kamilaris, Feng Gao, Francesc X Prenafeta-Boldu, and Muhammad Intizar Ali. Agri-iot: A semantic framework for internet of things-enabled smart farming applications. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pages 442–447. IEEE, 2016.

[43] Ghassan Karame. On the security and scalability of bitcoin's blockchain. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1861–1862, 2016.

[44] Prince Waqas Khan, Yung-Cheol Byun, and Namje Park. Iot-blockchain enabled optimized provenance system for food industry 4.0 using advanced deep learning. *Sensors*, 20(10):2990, 2020.

[45] Dimitris Kontokostas and Holger Knublauch. Shapes constraint language (SHACL). W3C recommendation, W3C, July 2017. https://www.w3.org/TR/2017/REC-shacl-20170720/.

[46] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 2020.

[47] Yuzhe Liu and Vanathi Gopalakrishnan. An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data*, 2(1):8, 2017.

[48] Imran Makhdoom, Ian Zhou, Mehran Abolhasan, Justin Lipman, and Wei Ni. Privysharing: A blockchain-based framework for privacy-preserving and secure data sharing in smart cities. *Computers & Security*, 88:101653, 2020.

[49] Milan Markovic and Peter Edwards. Semantic stream processing for iot devices in the food safety domain. *Proceedings of Semantics 2016*, 2016.

[50] Milan Markovic, Peter Edwards, Martin Kollingbaum, and Alan Rowe. Modelling provenance of sensor data for food safety compliance checking. In *International Provenance and Annotation Workshop (IPAW 2016)*, pages 134–145. Springer, 2016.

[51] Milan Markovic, Naomi Jacobs, Konrad Dryja, Pete Edwards, and Norval Strachan. Integrating internet of things, provenance and blockchain to enhance trust in last mile food deliveries. *Frontiers in Sustainable Food Systems*, 2020.

[52] Oliver Masters, Hamish Hunt, Enrico Steffinlongo, Jack Crawford, Flavio Bergamaschi, Maria Eugenia Dela Rosa, Caio Cesar Quini, Camila T Alves, Fernanda de Souza, and Deise Goncalves Ferreira. Towards a homomorphic machine learning big data pipeline for the financial services sector. *IACR Cryptol. ePrint Arch.*, 2019:1113, 2019.

[53] Gota Morota, Ricardo V Ventura, Fabyano F Silva, Masanori Koyama, and Samodha C Fernando. Big data analytics and precision animal agriculture symposium: machine learning and data mining advance predictive big data analysis in precision animal agriculture. *Journal of animal science*, 96(4):1540–1550, 2018.

[54] Mohamed H Mourad, Aydin Nassehi, Dirk Schaefer and Stephen T Newman. Assessment of interoperability in cloud manufacturing. In *Robotics and Computer-Integrated Manufacturing*, 61, 2020.

[55] Kristoffer Nærland, Christoph Müller-Bloch, Roman Beck, and Søren Palmund. Blockchain to rule the waves-nascent design principles for reducing risk and uncertainty in decentralized environments. In *International Conference on Information Systems (ICIS)*, 2017.

[56] PJ O'mahony. Finding horse meat in beef products—a global problem. *QJM: An International Journal of Medicine*, 106(6):595–597, 2013.

[57] Jeff Z Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu. *Exploiting linked data and knowledge graphs in large organisations*. Springer, 2017.

[58] Peter Patel-Schneider and Patrick Hayes. RDF 1.1 semantics. W3C recommendation, W3C, February 2014. http://www.w3.org/TR/2014/REC-rdf11-mt-20140225/.

[59] Simon Pearson, David May, Georgios Leontidis, Mark Swainson, Steve Brewer, Luc Bidaut, Jeremy G Frey, Gerard Parr, Roger Maull, and Andrea Zisman. Are distributed ledger technologies the panacea for food traceability? *Global Food Security*, 20:145–149, 2019.

[60] Xuan Pham and Martin Stack. How data analytics is transforming agriculture. *Business Horizons*, 61(1):125–133, 2018.

[61] Wolfgang Nikolaus Probst. How emerging data technologies can increase trust and transparency in fisheries. *ICES Journal of Marine Science*, 77(4):1286–1294, 2020.

[62] Food and Drink federation and Santader Food and Industry Report, 2020 *https://www.santander.com/en/press-room/news/opportunities-ahead-for-uk-food-and-beverage-sector*.

[63] Sam Sarpong. Traceability and supply chain complexity: confronting the issues and concerns. *European Business Review*, 2014.

[64] U Shruthi, V Nagaveni, and BK Raghavendra. A review on machine learning classification techniques for plant disease detection. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 281–284. IEEE, 2019.

[65] Monika Solanki and Christopher Brewster. Consuming linked data in supply chains: Enabling data visibility via linked pedigrees. In *COLD*, 2013.

[66] Sophie Stalla-Bourdillon, Gefion Thuermer, Johanna Walker, Laura Carmichael, and Elena Simperl. Data protection by design: Building the foundations of trustworthy data sharing. *Data &amp; Policy*, 2:e4, 2020.

[67] Christian Stoll, Lena Klaaßen, and Ulrich Gallersdörfer. The carbon footprint of bitcoin. *Joule*, 3(7):1647–1661, 2019.

[68] G Sylvester. E-agriculture in action: Blockchain for agriculture (opportunities and challenges). 2019.

[69] Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, and Casey S Greene. Multiplier: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell systems*, 8(5):380–394, 2019.

[70] Mamatha Thota, Stefanos Kollias, Mark Swainson, and Georgios Leontidis. Multi-source domain adaptation for quality control in retail food packaging. *Computers in Industry*, 123:103293, 2020.

[71] Mohamed Torky and Aboul Ella Hassanein. Integrating blockchain and the internet of things in precision agriculture: Analysis, opportunities, and challenges. *Computers and Electronics in Agriculture*, page 105476, 2020.

[72] Simone van der Burg, Leanne Wiseman, and Jovana Krkeljas. Trust in farm data sharing: reflections on the eu code of conduct for agricultural data sharing. *Ethics and Information Technology*, pages 1–14, 2020.

[73] Carol Anne Wallace and L Manning. Food provenance: Assuring product integrity and identity. *CAB Reviews*, 2020.

[74] Xiao Wang, Samuel Ranellucci, and Jonathan Katz. Global-scale secure multiparty computation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 39–56, 2017.

[75] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 2020.

[76] Wikimedia Commons, Jeromemetronome. Federated learning process in central orchestrator case, 2019.

[77] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[78] Leanne Wiseman, Jay Sanderson, Airong Zhang, and Emma Jakku. Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. *NJAS-Wageningen Journal of Life Sciences*, 90:100301, 2019.

[79] Sjaak Wolfert, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt. Big data in smart farming–a review. *Agricultural Systems*, 153:69–80, 2017.

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

[81] Zibin Zheng, Shaoan Xie, Hong-Ning Dai, Xiangping Chen, and Huaimin Wang. Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services*, 14(4):352–375, 2018.