Embracing Opportunities of Livestock Big Data Integration with Privacy Constraints

Franz Papst[12], Olga Saukh[12], Kay Römer[1], Florian Grandl[4], Igor Jakovljevic[5], Franz Steininger[3], Martin Mayerhofer[3], Jürgen Duda[4], Christa Egger-Danner[3]
[1]Institute of Technical Informatics, Graz University of Technology, Austria [2]Complexity Science Hub Vienna, Austria
[3]ZuchtData EDV-Dienstleistungen GmbH, Austria, [4]LKV Bayern, Germany, [5]smaXtec, Austria
{papst, saukh, roemer}@tugraz.at, egger-danner@zuchtdata.at

Abstract
Today's herd management undergoes a major transformation triggered by the penetration of cheap sensor solutions into cattle farms, and the promise of predictive analytics to detect animal health issues and product-related problems before they occur. The latter is particularly important to prevent disease spread, ensure animal health, animal welfare and product quality. Sensor businesses entering the market tend to build their solutions as end-to-end pipelines spanning sensors, proprietary algorithms, cloud services, and mobile apps. Since data privacy is an important issue in this industry, as a result, disconnected data silos, heterogeneity of APIs, and lack of common standards limit the value the sensor technologies could provide for herd management. In the last few years, researchers and communities proposed a number of data integration architectures to enable exchange between streams of sensor data. This paper surveys the existing efforts and outlines the opportunities they fail to address by treating sensor data as a black box. We discuss alternative solutions to the problem based on privacy-preserving collaborative learning, and provide a set of scenarios to show their benefits for both farmers and businesses.

CCS Concepts
• **Security and privacy → Domain-specific security and privacy architectures**; • **Applied computing → Agriculture**.

Keywords
data privacy, privacy-preserving data analysis, agriculture

ACM Reference Format:
Franz Papst[12], Olga Saukh[12], Kay Römer[1], Florian Grandl[4], Igor Jakovljevic[5], Franz Steininger[3], Martin Mayerhofer[3], Jürgen Duda[4], Christa Egger-Danner[3]. 2019. Embracing Opportunities of Livestock Big Data Integration with Privacy Constraints. In *9th International Conference on the Internet of Things (IoT 2019), October 22–25, 2019, Bilbao, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3365871.3365900

1 Introduction
In the past few years, a wide range of sensors for monitoring cattle appeared on the market. Those sensors usually focus on monitoring animals' behavioral and well-being parameters, *e.g.*, body temperature, feed intake and milk yield. Although a farmer can enjoy diverse real-time information about the status of his own facilities, receive alarms and early warnings, the insights he gets are as good as the data gathered on his own farm or on a set of farms that use the technology from the same sensor manufacturer. However, the farmers would benefit significantly from putting these data into the following two related contexts. First, a farmer needs to have an *integrated view of all the data generated on his farm*, *e.g.*, animal parameters, milk recordings, weather and farm information, independently of the APIs, business models and specifics of sensor service providers. Second, no farm should be considered in isolation and, therefore, a farmer should benefit from *the insights obtained from other farms* to improve event detection and insight quality gained from his own data. Even though there are different standards for exchanging agricultural data like AgroXML [16] or ISOBUS [18], none of these neither has gained a large market share nor managed to incorporate all sub-fields of agriculture [13].

With the advent of cloud computing, IoT solutions and the pervasiveness of machine learning models, the disparity between the benefits of integrating data from different sources and today's state of affairs in herd management gets more apparent than ever. The main factor for the lack of data sharing are privacy concerns. Information about the herd and its state are not personal data which is affected by the privacy regulations like GDPR. Farm data has not been legally classified, but it is argued, that farm data should be viewed as the *farmers' trade secret* [15][1]. This implies that it is up to the farmers to decide whom to share this information with. The drawback of data privacy is that it complicates data sharing and processing of joint data. At the same time, the farmers are generally willing to share their data for better insights and service quality in return, but only if they are assured that their data is protected and secure. The data exchange solutions offering transparency with respect to how the data is shared and who is in control of this data increase trust in these systems and therefore their adoption.

Another obstacle to data sharing are sensor manufacturing companies that host and analyze their customer's data. Sensor companies largely implement their own *proprietary algorithms* to gain insights from the data. The value created usually constitutes the companies' *intellectual property* and is in their control. These companies are interested in utilizing these data for product development. The gained insights are often linked to the companies' unique value proposition and give them an advantage on the market. This may pose a barrier to sharing the data even in exchange for a financial compensation.

IoT 2019, October 22–25, 2019, Bilbao, Spain

ACM ISBN 978-1-4503-7207-7/19/10…$15.00
https://doi.org/10.1145/3365871.3365900

[1]In this paper, we extend the notion of privacy to also include trade secrets.

An additional problem arises from the *lack of standards* concerning the following three aspects of the data: i) what type of sensor data needs to be measured and at what frequency, ii) how accurate and precise the measurements should be, and iii) the common APIs to access the data. For all the above reasons, the valuable cattle data remains in data silos.

Existing systems are either *farm-centric solutions* [2] that connect the data generated by sensor systems on a farm into a single information system to provide the farmer with an integrated view of his data and derived insights, or *data exchange services* [3, 5] between sensor solution providers that maintain high data privacy standards and treat the data they share as *black boxes*. The former systems treat every farm in isolation whereas the latter fail to make sense of the data they share due to a strong privacy focus.

Recent advances in privacy-preserving collaborative machine learning [15, 24] make it possible to think beyond data exchange and provide *data integration solutions that run data analytics on joint data while preserving privacy constraints of all parties involved*. This paper makes the following contributions: First, we survey the state-of-the-art data exchange systems for agricultural data focusing on cattle and dairying data. We discuss the opportunities with respect to data integration they fail to address in Sec. 2. Second, Sec. 3 introduces different approaches based on privacy-preserving data processing and discusses their value in the context of herd management. We provide a set of use cases that can benefit from large-scale data integration and analytics, and discuss the privacy trade-offs explored and exploited by these solutions.

2 Existing Systems

This section surveys the state-of-the-art agricultural data exchange systems, also summarized in Table 1. All systems emerged in the past few years and are promoted by farmer communities, governments and standardization agencies. The systems strongly focus on data privacy protection. We classify the systems as *centralized* and *decentralized* with respect to the concentration of the ownership rights on the exchange system itself.

2.1 Centralized Solutions

The systems below present the data exchange solutions and farm management platforms where the data flows through a system are managed by a single authority. Centralized architectures generally allow deploying an analytical model on top of the architecture to derive valuable insights from the data shared through the system.

Nordic Cattle Data Exchange (NCDX) [14] is a standardized interface for the exchange of cattle data developed for the Northern European countries. It supports all local farm management software providers resulting in an easier and more automated data exchange between different information subsystems such as farm management software and milk recording databases. The design of the interface is strongly influenced by the guidelines from ICAR[2], an international non-governmental organization, aiming to standardize data recording for livestock data. NCDX does not store data, but provides a transport layer encryption for data transmission.

Open Farm Information System (OFIS) [13] is a farm information management system proposed by researchers from Seoul National University. It is a cloud-based platform for data collection and data sharing. The system was developed for arable farmers, therefore it only collects and stores basic farm information, weather data, sensor observations and historical data. OFIS provides a standardized API for data sharing, privacy issues are not taken into account.

JoinData [5] is a Dutch non-profit platform for sharing farming data. It connects the data from different sensor companies to give a farmer an integrated view of the processes happening on the farm. JoinData initially focused on the dairying sector, but extended it to all areas of agriculture, especially to those prevalent in the Netherlands. The farmers can decide whom they share the data with. JoinData puts a focus on transparency, privacy and usability.

365FarmNet [1] is a farm management software developed and maintained by Claas, a German manufacturer of farm machinery and one of the largest companies in the field. It is a cloud-based platform covering different aspects of farming, *e.g.*, recording field usage, tracking the use of fertilizers and herd management. 365FarmNet provides data access through a unified interface and therewith makes data access more convenient for the farmer. The platform is modular and supports data import from other sources.

Barto [4] builds on 365FarmNet and provides an all-in-one solution for data-driven precision agriculture for the Swiss market. The platform offers interfaces to exchange data with federal agencies or consumers, the farmer decides whom to share the data with. It integrates with the federal animal tracing database[3] and the federal database for balanced use of fertilizers[4].

The primary focus of the above systems is on providing farmers with the standard means to record, access and exchange their data. Integration with federal databases and other chain parts adds significant value to these systems and fosters their adoption. The systems such as JoinData and Barto make a step forward towards providing an *integrated view of the data to individual farmers* by deploying analytics and models that run within the *local scope* of farmer's own and shared sensor data.

2.2 Decentralized Solutions

Decentralized solutions are particularly popular as they provide an elegant way to exchange data without a third party being involved.

Agrar-Daten-Austausch (ADA) [3] is a data bus and provides a standardized interface for data exchange. It is developed by a non-profit organization with the goal to help farmers increase the market value of their products, reduce costs, improve efficiency. ADA utilizes free and open standards and software. The system allows individual farmers to decide how their data is used. ADA addresses two main challenges: data redundancy and data communication, *i.e.*, how many copies of data are in circulation, and whom the data is shared with. To tackle these challenges, ADA utilizes a distributed ledger based on Hyperledger Fabric [12]. This allows the users of ADA to trace their data sharing and usage.

Offene Software-Plattform für Dienstleistungsinnovationen in einem Wertschöpfungsnetz in der Landwirtschaft (ODiL) [17] is an open software platform for the agricultural sector. It is developed by the German Research Center for Artificial Intelligence in cooperation with German universities and companies. The system provides a directory where the users can find each other and exchange data in a peer-to-peer fashion, without the need for a third

[2] ICAR: https://www.icar.org/

[3] Swiss Animal Tracing Database, https://bit.ly/2wCJcpz

[4] Suisse-Bilanz, https://bit.ly/2WRwIcn

System	Go live	Target markets	Type of data to exchange	Data storage	Analytics	Privacy	Architecture
NCDX [14]	2015	Finland, Sweden, Norway, Denmark, Iceland	Basic animal information, milk recordings, related events	✗	✗	✓	centralized
OFIS [13]	2015★	Republic of Korea†	Farm and field information, weather data, related sensor data	✓	✗	✗	centralized
JoinData [5]	2018	Netherlands	Any type of agricultural data	✓	✗‡	✓	centralized
365FarmNet [1]	2013	Germany, Austria, France, Poland, Bulgaria	Herd and field related data, weather data, fertilizer data, related sensor data	✓	✗	✓	centralized
Barto [4]	2018	Switzerland		✓	✗‡	✓	centralized
ADA [3]	2018		Any type of agricultural data	✗	✗	✓	decentralized
ODiL [17]	not yet	Germany		✗	✗	✓	decentralized
HARA [25]	2019	Indonesia, Vietnam, Thailand, Kenya, Uganda, Mexico, Peru	Any type of data, but the primary focus is on farmers' data, *e.g.*, location and boundaries of their land and property	✓*	✗*	✓	decentralized

★ – Paper publication. † – OFIS is not available to the public and does not have a target market. * – Off-chain data storage. Users can buy data to do analytics. ‡ – Integrated view.

Table 1: A survey of the state-of-the-art agricultural data exchange systems. All systems strongly focus on data privacy protection.

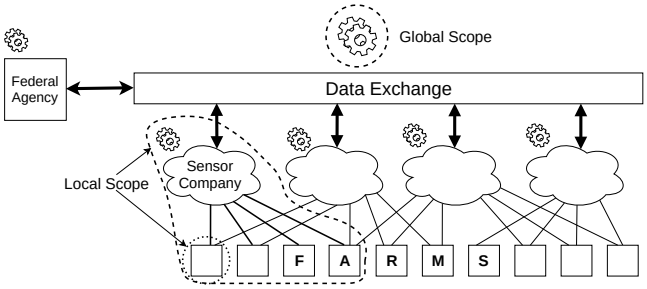

Figure 1: Network architecture supporting privacy-preserving data analytics in the global scope.

party. Thus, farmers are in full control over the usage of their data, yet have to host the data on their own infrastructure.

HARA [25] is an Indonesian start-up building a data exchange platform based on blockchain technology. The first use case for HARA is the exchange of agricultural data with an emphasis on the farmers' land location and ownership, cultivation and ecological data. The data comes from farmers, NGOs or IoT devices. Verification of the data is crowdsourced. The users are incentivized to use HARA by receiving rewards for the data they provide. The data can be acquired by interested entities like banks, insurance companies or government agencies. The platform ensures data privacy by default by storing the data encrypted off-chain, yet the data providers have full control over it. HARA is targeting markets in developing countries near equator, where agriculture plays a vital role in the economic development of these countries.

There are a number of general-purpose data exchange architectures [6, 11] which we do not cover in the paper, since they focus on standardizing data access and data exchange, yet fail to address the specific needs of herd management. All systems summarized above (see Table 1) strongly focus on protecting data privacy. A few systems step beyond data exchange and deploy data analysis models. Those end up with a *limited local scope* of their solutions due to the strong privacy constraints they face. The high-level architecture is sketched in Fig. 1. The insights these systems gain originate from the data limited to isolated farms or to sets of farms that implement the same sensor technologies. Yet local scope yields biased results. It is well known that the quality of today's machine learning models depends on the amount and quality of data used to train them. In the next section, we discuss a set of *privacy-preserving collaborative machine learning* approaches which can improve the quality of analytical models by training these in *global scope* without violating privacy constraints of the parties involved.

3 Beyond Data Exchange

Privacy is an important feature but is also a hurdle of data exchange platforms when it comes to data analytics. In this section, we look into privacy-preserving analytics on top of data exchange infrastructure and showcase its benefits for several scenarios specific to precision agriculture and herd management.

3.1 Privacy-preserving Data Analytics

Extracting insights from data under privacy constraints is a challenging task, especially for machine-learning techniques which often require substantial amounts of data to train a reasonably accurate model. Below we summarize privacy-preserving data analytics approaches that can be valuable for the analysis of cattle data.

Data Anonymization and Obfuscation. Traditional methods modify the data to limit the impact of individuals whose information is used in data analysis. *k-anonymity* [23] removes unique identifiers, *e.g.*, names, and groups quasi-identifiers, *e.g.*, age = 34 gets replaced by age $\in [30, 35]$. The parameter k indicates that an individual is indistinguishable from $k-1$ other individuals within the same group. *l-diversity* [10] builds on top of k-anonymity and further improves data privacy by enforcing diversity in sensitive attributes. *Differential privacy* [9] adds noise to the data to prevent that it can be traced back to an individual, but keeps the overall statistical properties of the original data set. It has been successfully used with a number of machine learning methods including boosting, PCA, linear and logistic regression and SVMs [24].

Secure Multi-party Computation [21] (SMC) helps to protect intermediate computation steps when multiple parties perform collaborative machine learning on their proprietary inputs. The input data is encrypted and computations are performed on this encrypted data. SMC has been used for learning decision trees, linear regression, Naive Bayes classifiers, and k-means clustering [24]. SMC techniques generally impose high performance overheads.

Federated Learning [24] is a novel approach for training machine learning models when training data are privately managed by a collection of nodes. Each of these nodes can train a local model from its data and share it with the others. The input data is never shared. A central entity is only required to collect the updated gradients learned by the local models. These gradient updates are aggregated and used to update the global model, which is shared with all the nodes. Each node then uses the global model and its local model to compute a highly accurate *personalized model*. Federated learning is particularly interesting in the context of precision agriculture and herd management, since the global model can generalize over sensor measurements across many farms, yet adapt to the specifics of a

local farm, *e.g.*, herd size, barn conditions and the quality of the green fodder.

Synthetic Data Generation [15]. The goal is to generate synthetic data which share statistical properties with the real data set. This approach is used to share data, which is similar to the original, but does not contain sensitive information.

3.2 Applications and Benefits for the Farmers

In the following subsection we give examples of application scenarios from precision agriculture and herd management and discuss the methods applicable to every scenario.

Predicting Missing and Delayed Data. The data fed into a data exchange system often originates from sensors that operate under challenging conditions, *e.g.*, on a collar or in the rumen of a cow. Thus, sensor failures, communication delays and transmission failures are common. Data analytics can enhance data quality by interpolating missing values to improve robustness of the applications that use this data. *Differential privacy* can be used to address privacy issues when training prediction models.

Anomaly and Event Detection. Farmers are interested in detecting unusual patterns in their cattle's data to be able to intervene. For example, a sensor in the collar of a cow can mistakenly be exchanged with another cow leading to wrong data interpretation by information systems. Automatic detection of such anomalies is necessary to ensure data consistency and reduce the number of false alarms. Detecting anomalies is, however, difficult because their occurrence is rare and gathering sufficient training data to train a model with a limited number of farms may be problematic. In this case, the models deployed in the global scope are able to learn from a higher number of events and thus outperform local models. Anomaly and event detection systems often employ RNNs [22] or LSTMs [20] to learn recurrent event patterns or use autoencoders [19] to reduce dimensionality and identify patterns outside of the obtained latent space representation. *SMC* and *federated learning* provide a good basis to protect data privacy when building such models for herd management.

Autocompletion and Verification of Manual Data Input. Farmers and veterinarians have to manually add multiple values into their information systems. Depending on the size of the farm, this can be a laborious and time-consuming task. Machine learning models such as keyboard key prediction,—also known as autocomplete,—can save time and reduce input errors. For example, the system may automatically suggest if a cow is ready to inseminate based on the related data already known to the system. Verifying suggested inputs by an expert further decreases the likelihood of wrong values being entered. Autocompletion is a primary use-case for *federated learning*, which learns the most probable autocompletion patterns without sharing person's input.

Monitoring Cattle Health and Welfare. Federal agencies, sensor solution providers, veterinaries and data analytics businesses in herd management work on gaining valuable insights from the data they own. Privacy-preserving analytics offers a way to train truly large-scale models and customize these to every farm and animal. This approach enables *precision farming at unprecedented scale*. Running health-related analytics at global scope is particularly important to increase animal welfare and let veterinaries and farmers make faster and more targeted decisions, *e.g.*, to prevent the spread of a disease. k-anonymity, l-diversity and differential privacy are valuable in this use-case and have already proven to be useful in processing human health data while respecting privacy [8].

4 Discussion

Privacy is not binary, but rather a spectrum spanning *full access* to the data and treating data as a *black box* as two extremes. All privacy-preserving data processing approaches discussed above fall in-between. They may disclose statistical features about the data distribution but hide the contribution of each individual to the data set [7]. Except for federated learning, privacy-preserving data processing alters input data. This may negatively impact the quality of the trained models. k-anonymity, l-diversity and differential privacy trade data quality for privacy protection. Synthetic data circumvents these issues, yet the quality of the generated data depends on the quality of the original data.

Acknowledgments. This work was conducted within the COMET-Project D4Dairy that is supported by BMVIT, BMDW and the provinces of Lower Austria and Vienna.

References

[1] 2013. 365FarmNet. (2013). https://www.365farmnet.com/
[2] 2017. Farmit. (2017). Retrieved 2019-06-11 from https://farmit.at/
[3] 2018. Agrar Datenaustausch / Des échanges de données agricoles. (2018). Retrieved 2019-06-01 from https://www.ada-eda.org/
[4] 2018. Barto. (2018). Retrieved 2019-06-01 from https://www.barto.ch
[5] 2018. Join Data. (2018). Retrieved 2019-06-01 from https://www.join-data.nl
[6] K. Aberer, M. Hauswirth, and A. Salehi. 2006. Global Sensor Networks. (2006).
[7] R. Agrawal and R. Srikant. 2000. Privacy-preserving Data Mining. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*.
[8] F. K. Dankar and K. El Emam. 2012. The Application of Differential Privacy to Health Data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*.
[9] C. Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming (Lecture Notes in Computer Science)*, M. Bugliesi *et al.* (Ed.).
[10] A. Machanavajjhala *et al.*. 2006. l-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*.
[11] B. Otto *et al.*. 2019. International Data Spaces Association — Reference Architecture Model. (2019).
[12] E. Androulaki *et al.*. 2018. Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains. In *Proc. of EuroSys Conf. (EuroSys '18)*.
[13] J.Y. Kim *et al.*. 2015. Open Farm Information System data-exchange platform for interaction with agricultural information systems. *Agricultural Engineering International: CIGR Journal* 17, 2 (2015).
[14] J. Kyntäjä *et al.*. 2018. Nordic Cattle Data eXchange – a shared standard for data transfer. In *Cooperation, Networking and Global Interactions in the Animal Production Sector. Proc. of ICAR Conf.*
[15] J. W. Anderson *et al.*. 2014. Synthetic data generation for the internet of things. In *2014 IEEE International Conference on Big Data (Big Data)*.
[16] M. Schmitz *et al.*. 2009. agroXML Enabling Standardized, Platform-Independent Internet Data Exchange in Farm Management Information Systems. In *Metadata and Semantics*, M.A. Sicilia and M.D. Lytras (Eds.).
[17] S. Stiene *et al.*. 2017. Architektur einer offenen Software-Plattform für landwirtschaftliche Dienstleistungen. (2017).
[18] T. Oksanen *et al.*. 2005. ISO 11783 – Standard and its Implementation. *IFAC Proceedings Volumes* 38, 1 (2005).
[19] G. Hinton and R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. 313, 5786 (2006).
[20] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* (1997).
[21] Y. Lindell. 2005. Secure Multiparty Computation for Privacy Preserving Data Mining. *Encyclopedia of Data Warehousing and Mining* (2005).
[22] D. P. Mandic and J. Chambers. 2001. *Recurrent Neural Networks for Prediction: Learning Algorithms,Architectures and Stability*. John Wiley & Sons, Inc.
[23] P. Samarati and L. Sweeney. 1998. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. (1998).
[24] R. Shokri and V. Shmatikov. 2015. Privacy-Preserving Deep Learning. In *Prof. of SIGSAC Conf. on Computer and Communications Security (CCS '15)*.
[25] R. Wahyu, I. Zuhri, and A. Jatra. 2019. *HARA Token Whitepaper*. Technical Report. https://haratoken.io/doc/HARA_Token_White_Paper_v20190325.pdf