IEEE Open Journal of the Computer Society

Received 27 April 2021; accepted 27 May 2021. Date of publication 1 June 2021;
date of current version 29 June 2021. The review of this article was arranged by Associate Editor Xiang Sun.

Digital Object Identifier 10.1109/OJCS.2021.3085846

LATTICE: A Vision for Machine Learning, Data Engineering, and Policy Considerations for Digital Agriculture at Scale

SOMALI CHATERJI [1], NATHAN DELAY[1], JOHN EVANS[1], NATHAN MOSIER[1], BERNARD ENGEL[1], DENNIS BUCKMASTER[1], MICHAEL R. LADISCH[1], AND RANVEER CHANDRA[2] (Fellow, IEEE)

[1] Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, IN 47907 USA
[2] Microsoft Research, Microsoft Azure, Redmond, WA 98052 USA

CORRESPONDING AUTHOR: SOMALI CHATERJI (e-mail: schaterji@purdue.edu)

This work was supported in part by the Purdue University (150 Giant Leaps Selected Project), in part by the WHIN-Purdue Funding from Lilly Endowment Inc (WHIN is Wabash Heartland Innovation Network), and in part by the Microsoft Research.

ABSTRACT Digital agriculture, with the incorporation of Internet-of-Things (IoT)-based technologies, presents the ability to control a system at multiple levels (individual, local, regional, and global) and generates tools that allow for improved decision making and higher productivity. Recent advances in IoT hardware, e.g., networks of heterogeneous embedded devices, and software, e.g., lightweight computer vision algorithms and cloud optimization solutions, make it possible to efficiently process data from diverse sources in a connected (smart) farm. By interconnecting these IoT devices, often across large geographical distances, it is possible to collect data at different time scales, including in near real-time (i.e., with delays of only a few tens of seconds). This data can then be used for actionable insights, e.g., precise applications of soil supplements and reduced environmental footprint. Through LATTICE, we present an integrated vision for IoT solutions, data processing, and actionable analytics for digital agriculture. We couple this with discussion of economics and policy considerations that will underlie adoption of such IoT and ML technologies. Our paper starts off with the types of datasets in typical field operations, followed by the lifecycle for the data and storage, cloud and edge analytics, and fast information-retrieval solutions. We discuss what algorithms are proving to be most impactful in this space, e.g., approximate data analytics and on-device/in-network processing. We conclude by discussing analytics for alternative agriculture for generation of biofuels and policy challenges in the implementation of digital agriculture in the wild.

INDEX TERMS Data integration, data analysis, internet of things, Sensor systems, cloud computing.

I. INTRODUCTION

BY 2050, the world's population is projected to increase to nine billion, which will intensify the food-water-energy nexus challenges. Demand will also rise because of increase in people's wealth resulting in higher meat consumption plus the increasing use of cropland for biofuels. Site-specific farm management (precision farming) has the potential to nourish the world while increasing farm profitability under constrained resource conditions. Despite advancements in field sensors, the global positioning system (GPS), and grid soil sampling, adoption of technology by farm operators has fallen short of expectations. Moreover, it is unclear how profitable the adoption of such technologies will be. Use of variable rate technology (VRT), for example, has lagged that of yield monitors and automated guidance systems. A thorough study [1], including rigorous analysis, has shown how the lack of widespread adoption of VRT can be attributed to the paucity of site-specific data. Specifically in this study, the authors attributed the scant adoption of variable rate nitrogen application to the lack of site-specific yield data. The generalizable insight from this is that site-specific detailed data about the effects of digital agriculture interventions are important to drive their adoption. **Identified gaps motivating LATTICE**

Some of the gaps on the agricultural side include:

1) Mapping seed variety to performance with better seeds or engineered varieties (*e.g.*, via genome editing [2]).
2) Mapping soil supplementation needs to regional/temporal conditions using interpretable data science [3].
3) Defining functional properties of derived bioproducts (*e.g.*, ethanol from corn), distinct from traditional products (*e.g.*, heavier kernels using genome editing).

Conversely, some of the gaps in the data engineering and machine learning areas for digital agriculture include:

1) Approximate data analytics for processing data from multiple inexpensive sensors deployed on connected farms. This is especially important for compute-intensive workloads, *e.g.*, vision workloads from drone imaging for object classification/detection, *e.g.*, in our recent approximate object classification and detection work [4], [5].
2) Tradeoff between privacy and utility when analyzing data from multiple farms, similar to federated computing, used profitably in other fields, *e.g.*, genomics [6].
3) Network management for sparsely connected farms using newer networking solutions that are bandwidth-aware and do not require cell towers [7], [8].
4) Drones and tractors for data ferrying when needed especially under sparse network connectivity [9], [10].
5) Effective sharing of data processing and analytics load between sensors, edge devices [11], [12], and the cloud for maximizing throughput or latency, based on the client (farmer/farm manager) preferences [13], [14].
6) Optimized cloud computation for beefier machine learning workloads using vision APIs, *e.g.*, Azure Vision or Amazon Rekognition [15] or processing streaming workloads using on-premise database optimization or using optimized clustered cloud instances [16]. Examples of such optimization can be found in our recent work for on-premise database optimization [17] for streaming workloads or cloud/serverless optimization [18], [19] for computationally-heavier vision [5] or lighter-weight, but latency-sensitive, IoT workloads [15].
7) Data ethics when sharing farm data with agricultural companies or insurance providers.

On the economics side, VRT for fertilizer treatment, as an example, depends on accurate intra-field soil data, which is expensive. Unless the economic returns to site-specific management cover both the up-front investment and the cost of collecting quality data, adoption will be low. Guidelines for VRT use for fertilizer application illustrate the need for:

- Marking management zones for the VRT system.
- Identifying whether the system will be guided by map-based inputs or finer-granularity sensor-based inputs. While the sensor-based inputs are more sophisticated because they reflect the changing conditions in the farm, they are also logistically and computationally more expensive because they are battery-powered and will need to continuously or intermittently be guided by anomaly or bottleneck detection [20].
- Identifying the kind of data that will be used for mapping or the kind of data for the actuation of VRT dispensers.

For farmers to adopt these technologies (VRT provided as an example and others, *e.g.*, edge-cloud data partitioning, discussed in this article), concrete savings on resources (*e.g.*, supplements or fertilizers) need to be demonstrated with potential yield increase and environmental protection from decreased farm effluents from nutrient pollution and reducing farm runoff and eutrophication (hypertrophication), such as from high levels of nitrogen and phosphorous in fresh water. In the case of livestock farmers, this translates to the decreased use of hormones, supplements, or antibiotics for the livestock, resulting in ecological gains [21], [22].

Relevance of our team's ongoing efforts. Digital agriculture—encompassing precision agriculture, data analytics and edge-cloud computing, and data privacy and ownership—has the promise to transform agricultural throughput. It can do this by applying data science for mapping input factors to crop throughput and that too in a region-specific and crop-specific manner, while bounding the available resources, both tangible farm-specific resources such as seeds, nutritional supplements, and farm machinery, and computational resources (*e.g.*, cloud credits or CPU/GPU cycles) or networking expenses (*e.g.*, LoRA or NBIoT towers). In addition, as the volumes and varieties of data increase with the increase in sensor deployment in agricultural fields, data engineering techniques will also be instrumental in collection of distributed data as well as distributed processing of the data. These have to be done such that the latency requirements of the end users and applications are satisfied.

At the same time, Microsoft has developed and is looking to spread the reach of the FarmBeats program [23], which has the vision of empowering farmers with low-cost digital agriculture solutions using low-cost sensors, drones, and computer vision and machine learning (ML) algorithms. Understanding how farm technology and big data can improve farm productivity can significantly increase the world's food production by 2050 in the face of constrained arable land and with the water levels receding. While much has been written about digital agriculture's potential, little is known about the economic costs and benefits of these emergent systems.

There are important questions to be answered before data analytics for agriculture, questions related to technical viability, economic feasibility, sustainability, and data protection and ownership are implemented. These questions cannot be looked at in isolation—for example, if some algorithm needs data from multiple data owners to be pooled together, that raises the question of data ownership and data privacy.

In summary, the paper reviews the current state of the following questions and presents a look forward at the challenges that need to be resolved, namely: *data lifecycle* for digital agriculture data; *applied ML* techniques for the domain; *low-power communication protocols* for the domain; *economics, policy, and decision making* driving adoption. LATTICE is the

first to bring together these questions under one roof, discussing the goals, path forward, and challenges of digital agriculture to surmount the challenges of the food-water-energy nexus. We present an integrated effort LATTICE,[1] which will culminate in the incorporation of foundational AI advances , driven by the domain constraints and requirements of digital agriculture systems, resulting in efficient distributed computational infrastructure to execute the algorithms.

II. DATA GENERATION FROM SENSORS

This section will cover the modalities of data gathering and controlled dissemination, the volume of data generated, and the quality of the data, collected and processed from ubiquitous sensors on farms and includes the following.

Soil sampling: Soil samples are extracted from field (may or may not be georeferenced) and sent to a lab for analysis. Lab results come back in a report detailing fertility levels. In most cases, one must manually assign geo-referenced points to report values. **Fertilizer application**: Based on soil sampling, a fertilizer recommendation is generated. If geo-referenced points are used, a variable rate prescription can be generated. If fertility results are aggregated across the field, a flat rate is applied (proprietary or shapefile). **Planting**: Generated as applied maps with information on population, singulation, misses (proprietary or shapefile).

Scouting: Conducted as frequently as once a week during growing season. Traditionally data comes in the form of a report that details presence of disease, insect, and weed pressure. Currently much research is being conducted on drone-based scouting using multi-spectral imagery to determine nutrient and water deficiencies as well as detect disease, insect, and weed pressure. Data format is large image files that need post processing. **Spraying**: Based on the results of the scouting reports spraying operations are conducted. Most modern sprayers can generate as applied maps. Files are saved in a proprietary format based on the sprayer manufacturer. **Harvesting**: Yield maps are generated by harvesters and saved in a proprietary, manufacturer-specific format. In irrigated fields, soil moisture sensors are often used to determine irrigation intervals. Despite the wealth of available data in most operations, very little is actually analyzed and used to inform future decisions. The most commonly used data sets are scouting data (used to make chemical application decisions) and soil sampling data (used to make fertilizer recommendations). More progressive producers use yield data to determine variable rate fertilizer application and to compare seed varieties. Much of the data collected only when the richer context of the data is known. This can be known through processing such as aggregation with additional local or regional data and attaching the correct metadata. **Biorefining**: Different bioprocess operational modes for converting components of a crop (*e.g.*, corn) are recorded during enzyme catalyzed conversion, fermentation, biocatalysis, and separations to achieve the desired purity of the target. These data can be used to identify inherent barriers, which can be overcome by modifying feedstock.

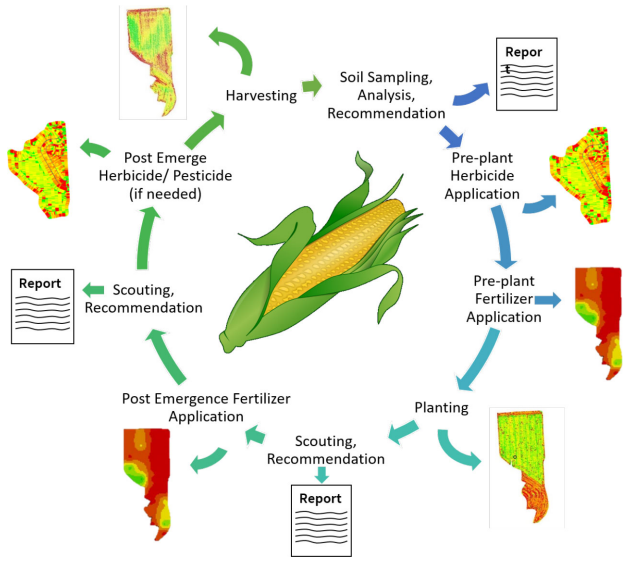


FIGURE 1. Example of no-till corn operation, data generation, and lifecycle. In modern agriculture, data is generated from almost every operation. This data may be site-specific or more broadly defined but often there is no clear way to aggregate data layers.

For VRT, with advanced electronic controls and improved communication, applications include: fertilizer/nutrient applications, manure, seed applications, tillage as a function of soil compaction, and irrigation. Thus, if the VRT leverages sensors rather than static maps, farm processes can benefit.

One primary challenge facing digital agriculture is the lack of data sharing, which happens due to technological as well as human reasons. The technological impediment centers around the lack of interoperability of data collection, processing, and visualization tools. Producers are reluctant to share data due to fears of regulatory issues and the lack of perceived value.

III. DATA LIFECYCLE

This section will cover the various phases in the lifecycle of agricultural data—sanitization, loading, processing, storing, summarization, and analysis; an example is shown in **Fig. 1**. This will go into some of the general-purpose approaches (*e.g.*, data deduplication, calibration using sensor metadata) as well as agriculture-specific approaches (*e.g.*, known variations in hyperspectral maps from ground sensing and aerial image data and effective fusion among sensor arrays).

Data generation sources: Data generation is the first stage of a data lifecycle. There are many ways in which data can be generated. The sources of data generation can be broadly classified into two types.

Localized data or private data: This is the data that is generated on the farm such as soil nutrient composition, water, and fertilizer usage. This type of data is generated from sensors that are present on the farm.

[1]Our name derives from Distributed *L*earning for Agriculture Systems through *Art*ificial in*t*elligen*ce*. This is an inspiration from arrayed lattices that integrate into innovative structures for cohesion and creativity — a reflection on how the correctly structured and integrated data pipeline elements can lead to leaps in crop productivity.

Public data: Data such as historic weather conditions and market prices fall under this category. Imported data is often generated at outside sources and shared with the farmers to use in precision agriculture. Such data is not farm specific. An example of data that is at the crux of localized and public data is topography and soil type, which may be somewhat localized but follow a trend for farms in geographical proximity.

Data warehousing: Data generated then needs to be stored in repositories called data warehouses. Data warehousing allows integration of different data from multiple sources and helps restructure the data for better performance. One recent example of data warehousing is an initiative taken from the government of India [24], titled INARIS (*Integrated National Agricultural Resources Information System*). Woodard [25] discusses Ag-Analytics—a platform that provides data warehousing in the field of precision agriculture. Although there are readily available platforms for data warehousing, there are several constraints when using these platforms directly in precision agriculture. Some of these constraints are discussed in [26].

Metadata annotation: Metadata annotation can be done manually or automated. Since the data that we are considering is prone to be complex, automated metadata annotation is preferred. Roy, Sarkar, and Ghose [27] provides a comparative study of the different learning techniques used for metadata annotation. Fiehn, Wohlgemuth, and Scholz [28] provides an algorithm for metadata annotation. Haug and Ostermann [29] uses a human expert to first mark crops from raw images. Masks were then derived from using these markings. These masks are then used to acquire the metadata. Similar techniques can be used in different aspects of precision agriculture.

Data annotation and cleaning: Due to the large size of data, it is important to perform annotation and cleaning before data analysis. Data annotation is subjective and depends on the particular use case of precision agriculture. The choice of the data annotation technique is dictated by the size of the data set, cost of annotation per sample among many other guidelines. Schoofs, Guerrieri, Delaney, O'Hare, and Ruzzelli [30] proposes a data annotation technique for electricity data in wireless sensor networks (WSNs). Similar annotation techniques could be developed in precision agriculture that can be performed in a WSN infrastructure. Data cleaning removes or corrects errors that are present in the data. There are several existing works that propose data cleaning techniques in precision agriculture. Simbahan, Dobermann, and Ping [31] proposed a screening algorithm for cleaning yield data that provided an increase in map precision. Sun, Whelan, McBratney, and Minasny [32] proposes an integrated framework for software that increases mean yield through data cleaning.

Data processing: Steven [33] provides a good overview of the constraints faced in data processing in precision agriculture. In [33], the authors consider the case of using satellite images for remote sensing in precision algorithms. In such use cases, one of the important aspects of data processing must be to make the data more readable. One such scenario where these images can not be directly used is in the case of cloud cover, where the images need to be processed before utilizing the data. This can be extended to data acquired through other means as well. Data acquired from soil sensors may need to be processed in order to make it more utilizable. Honkavaara *et al.* [34] proposes a processing chain that uses data collected from unmanned airborne vehicles to generate meaningful results. Loreto and Morgan [35] proposes an automated system that performs both the data acquisition and data processing of soil nitrate measurements. Murakami, Saraiva, Junior, Cugnasca, Hirakawa, and Correa [36] proposes a data processing algorithm that processes yield data on a distributed framework.

IV. DATA ANALYTICS FOR DIGITAL AGRICULTURE

Here, we will cover approximate processing for in-sensor analytics, advanced processing for backend analytics on the edge or cloud platforms, and interpretable data analytics. Analytics workloads are often quite demanding, and do not fit, out-of-the-box, into the embedded devices, deployed in agriculture. Advanced processing for backend analytics may leverage edge platforms, *e.g.*, Azure IoT Edge device [23], [37], and there may be sensitivity of farmers to upload personal data to the cloud. Interpretable analytics are important because the farmers will require insights into the results of the algorithm, at their level of understanding, to potentially take action. In Fig. 2, we show a high-level architecture of the nodes deployed in different locations, illustrating the execution of analytics routines on the heterogeneous nodes.

Data analytics plays an important role in precision agriculture. It can help farmers decide what crop to grow when, monitor the crop growth, and decide on the logistics of farm management. But agricultural data is often large and noisy and needs careful processing to distill insights from them. The following subsections elaborate on the advanced ML capabilities that can be used for such analysis.

Another relevant technology in this context is the use of scalable databases to house and process these data sets for downstream processing and retrieval. With this in mind, we also include some innovations in NoSQL database technologies to assist in high-throughput information retrieval from the evolving agricultural data. Traditionally, this domain did not have requirement for low latency—the decisions were more mid-term tactical or strategic and could be made in the time frame of hours or days. However, with sophistication in farm machinery and automated nutrient application, some use cases of low-latency data analytics have evolved. One example is farm machinery moving over a region of the farm, queries ground sensors for soil quality or cameras on-board probe for pest infestation on the plants. It then analyzes the sensor feed and decides on the appropriate application of nutrients/pesticides, actuating the on-board applicator on the dispenser, real-time. In Fig. 3, we illustrate data flow in FarmBeats as an example.

Approximate processing for in-sensor analytics: A sensor network acquires real-world measurements at discrete

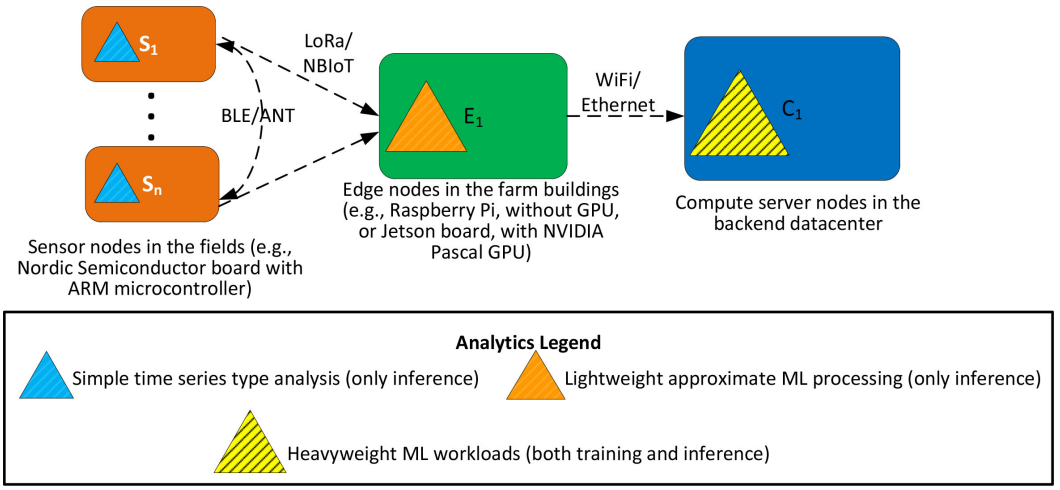


FIGURE 2. System architecture of the distributed analytics relevant to digital agriculture.

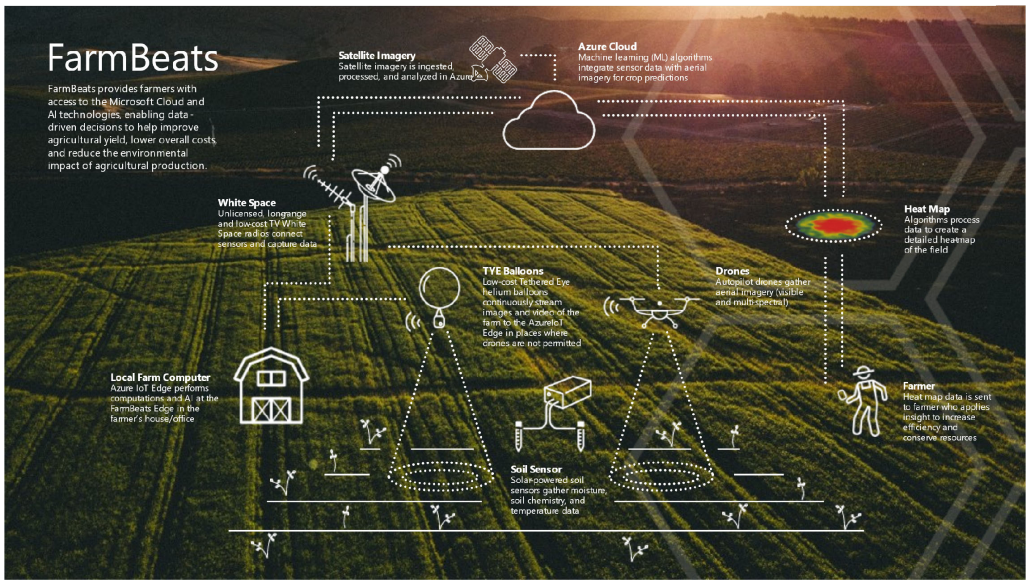


FIGURE 3. *FarmBeats*, an example of a data-driven agricultural platform from Microsoft, captures data from sensors and drones, and sends it to the Cloud for processing with other data streams, such as satellite data and weather stations.

points, where each measurement is a snapshot in time and space. In most scenarios in which sensor networks are deployed, the sensors are frequently queried resulting in continuously monitoring alongside high energy costs. Thus, one of the chief problems faced by WSN, often deployed in farm settings, is the constrained availability of resources to these devices.

The sensors used in such networks are low-power embedded devices that are expected to last for long periods of time (order of months) on standard batteries [38].[2] This issue can be mitigated by leveraging a more distributed architecture and using more energy-efficient algorithms. Further, these devices generate large volumes of data, which ideally will be processed in real time in a streaming manner for usable insights.

As a balance between computational load and accuracy, approximate computing tools and techniques have become popular in several domains, *e.g.*, computer vision [5], [39] and scientific computing [40]. The idea is to perform approximate computation over carefully chosen subsets of the entire input dataset. In digital agriculture, some degree of approximation or error in the output of the algorithm is tolerable, either because humans cannot perceive these differences or downstream algorithms are not affected by such approximations.

Relevant to our discussion, an example is the approximate computation can relay whether a particular soil nutrient concentration is above or below a threshold, rather than the exact value of it. Also, it may compute this over a uniform random subsample of say one in every 10 samples. Alternately, we can also use information theory principles, such as the Nyquist-Shannon sampling theorem to decide on the spacing

[2]One class of sensors that does not have this energy constraint is those sensors that are mounted on farm machinery that are connected to power outlets. Such sensors can also be powered by AC line power.

of the sensors in the WSN [41] and also the actual redundancy needed for robust sampling from sensors [42].

An important requirement of our target applications is low latency. This can be achieved by using multiple nodes to parallelize the work. However, this has to be done carefully so that the load is approximately balanced and there is not much overhead of energy to perform the distribution. For example, ApproxIOT [43] proposes an algorithm based on Apache Kafka [44] that uses IoT devices to generate data and forward it to edge computers, managed by service providers. As shown in Fig. 2, wireless communication, like LoRa/NBIoT, is used to forward data, compressed or otherwise, to the edge devices. These data streams are then sampled and forwarded to a central location (compute server C_1 in the figure), where user-specific queries can be made. The sampling in such systems are based on two techniques: *Stratified Sampling:* The streams of data are categorized based on their distribution. A random sampling is done on these distributions, with prior knowledge of the data required for this kind of sampling. *Reservoir Sampling:* A reservoir size R is maintained and at most R items are uniformly sampled from the data set. Here, prior knowledge of the data is not required. ApproxIOT extends both these techniques to a weighted hierarchical sampling. The nodes conduct sampling over data generated and compute statistics, with a 1.3X – 10X speedup.

In-sensor analytics is relevant to our domain because it means that the data being sensed will be (partially) analyzed locally at the sensor itself. The value proposition is that raw data will not have to be sent over the wireless network, thus saving on wireless bandwidth and energy, and thus potentially yielding low-latency decisions. In-sensor analytics algorithms can either be **value tolerant**, where the approximation of values can be made and the resulting algorithms are lightweight or they can be **delay tolerant**, where the resulting algorithms can be made more accurate at the expense of latency.

SERENE [45] is a framework that selects *representative* nodes, among clusters of correlated sensors. Here, the framework also takes into account the dynamic changes in the network topology and outliers. Since sensor data acquisition and communication are energy intensive, and sensors are typically battery-powered, SERENE uses clustering algorithms to spatially and temporally aggregate the data. For example, it uses a density-based clustering algorithm, DBSCAN [46] for robustness against outliers and noise. This algorithm can cluster based on any shape, as the sensor readings may be correlated. Based on the cluster shape, availability of battery power, and distance of the target node from other nodes, the representative nodes —*M-sensors*— are queried.

Another lightweight approach for in-sensor analytics is **Snapshot Queries** [47]. Here the representative nodes are elected through a localized process. Each node maintains a data distribution model of its neighboring nodes. Based on this model, the algorithm predicts the values of the neighboring nodes. If the error between the predicted and actual value is less than a threshold, the predicting node can represent the neighboring node in question. The data model maintained is based on the previous correlation between the values of the node and its neighbors. Here, the model is updated frequently, making it robust to dynamic network changes. Another work that leverages the correlation between sensor values is Kartakis *et al.* [48] where they use this property to detect anomalies with a Kalman filter to reduce false positives.

Distributed neural network inferencing: The concept of lightweight algorithms can also be extended to deep neural networks. Distributed Deep Neural Networks (DDNNs) provide better scalability and fault tolerance than DNNs. The data generated by sensor nodes is processed locally at the edge. DDNNs are used to utilize the advantages of distributed computing hierarchy in DNNs [49]. Further, simple ANNs have been used in tandem with Bayesian loops to reduce the number of hyperparamters in models and thus improve the interpretability of models and decrease the need for additional hyperparameter tuning. Having a multitude of hyperparameters to tune is often an overkill for simpler processing needs with energy considerations in mind, as is often the case for lightweight edge processing for sensor node analytics [50]. This and other approaches, such as Bayesian neural networks (BNNs), sped up using hardware accelerators, can be a panacea when data volume is limited (as in cases where sensor nodes have been initialized on a farm) and to prevent overfitting plus allow for limited memory footprint [51]. Limited memory space may be the case for microcontroller-class devices or lower-resourced edge-class devices used in sensor nodes or gateway nodes, where memory is of the order of a few GBs rather than 100 s of GBs at server-class machines.

Another aspect to reduce the energy consumption of sensor nodes is by *reducing the duty cycles (sleep-wake cycling) of sensors*. By activating the sensors only when required, the energy consumed by the sensors will be reduced. In our target domain, this is highly feasible since the sensing frequency can be kept low (of the order of a few minutes) due to the nature of the underlying events being sensed.

Database management and backend analytics for large-scale agricultural sensor data: Precision agriculture allows for site-specific crop management to increase throughput and achieve more sustainable farming by applying data science to agriculture practices, learning from local data trends. More and more agricultural sensing data is being live-streamed from farms, whether it be through on-board cameras, on manned or unmanned aerial vehicles, or through ground sensors in the farms. There is thus a need for centralized databases to store and process these data sets, often in real-time, to get actionable insights for farmers. Plus, there may be some degree of federation in storage and compute resources that may be needed as the computing needs of this domain increases, as has been seen in the genomics domain [6].

Further, the sensing data is multi-dimensional and noisy, coming from ground sensors deployed in farms to measure an array of soil characteristics, such as, moisture, nutrient levels, temperature profiles, soil acidity, etc. These live-streamed data sets need to be stored in a fail-safe repository of nodes, such

as Redis installations in Amazon Web Services (AWS) Elastic Compute Cloud (EC2) [52]. Redis is a popular NoSQL datastore that is being currently used by Twitter, Instagram, Microsoft, Groupon, etc. It is an in-memory key-value store that supports various abstract data structures, such as strings, lists, maps, etc. By storing the data in the main memory, Redis serves as both a datastore and a cache. This allows for a very fast and flexible storage model, which is essential for real-time processing pipelines, such as in digitized agriculture pipelines, *e.g.*, FarmBeats [53]. Moreover, Redis supports durability by periodically saving a snapshot of its main memory to the permanent storage. This snapshot can be loaded back into main memory in the case of a failure.

The aggregate live-streamed updates and queries represent a unique workload for such NoSQL datastores, as these queries vary in rate and type over time. In such cases, a workload-aware tuning system is needed to reconfigure the NoSQL cluster, whether locally or on the cloud, to provide high performance. This is becoming more important as the data from small-scale farms across the country is burgeoning both in size and diversity, slowly replacing the previously used manual data-collection processes. Maximizing the throughput of these processing pipelines of digital farm data will enable actionable insights from agricultural sensing data. Also, given that these pipelines are hosted on the cloud and this domain is very cost conscious, we want to maximize the performance within a user-defined cost bound.

Our recent work on cost-aware optimization of NoSQL database has shown promise [17], [18]. A pipeline for analytical workloads (OLAP, online analytical processing) consists of two main parts: a storage cluster (*e.g.*, Redis or Cassandra) and a computing cluster (*e.g.*, Spark), the latter operates on the top of the storage cluster to execute queries and thus perform analysis on the stored data. The task of the optimization is to find the best combination of configurations maximizing the objective metric, modeled as:

$$P^* = \arg\max_{Confs} f(Confs(t), WL(t)) \qquad (1)$$

Where P^* is the optimal performance that is achieved by the best combination of configurations $Confs$. The search space of $Confs$ is large, *e.g.*, NoSQL databases, *e.g.*, Cassandra and Redis, have 50+ and 40+ performance-sensitive parameters, respectively. Hence, an exhaustive search through all possible configurations is impractical. Therefore, evolutionary search techniques are preferred in this case due to their ability to find close-to-optimal solutions in practical time, *e.g.*, **1**.

Experimental Evaluation: We perform our experiments on multi-modal sensor data, mimicked on real data collected from our experimental digital agriculture farms. First, we identified the most impactful NoSQL parameters for our agricultural processing pipeline. We use D-optimal design to specify the data points to collect to reveal these impactful parameters. We collect 128 data points. These data points represent different combinations of configuration parameters and their corresponding performance (in terms of throughput, *i.e.*, Operations/s or Ops/s). Second, we use the collected data points to train and test our performance prediction model. We estimate the importance of each feature and select the most impactful parameters (Fig. 4). We then use these top-k impactful parameters in training our model.

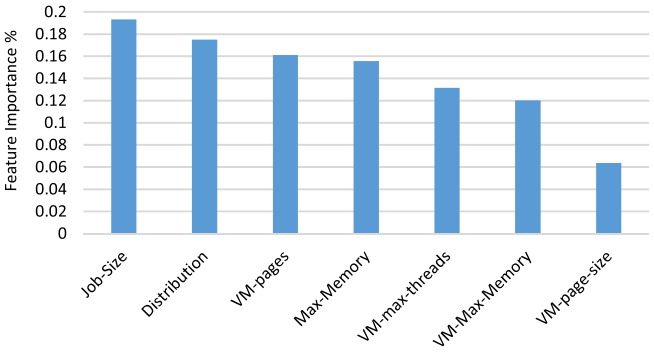


FIGURE 4. Feature importance for accurate performance prediction.

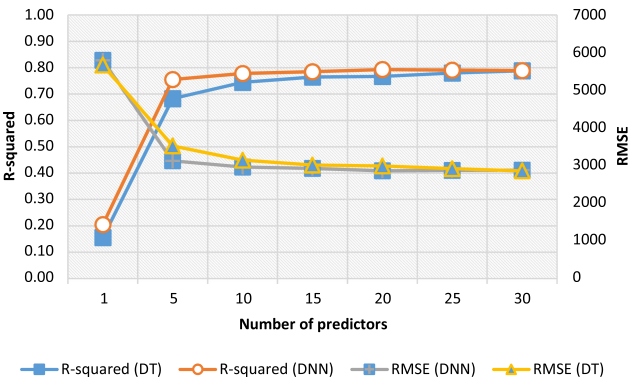


FIGURE 5. Improved performance with increasing ensemble size.

Performance prediction model training: In this experiment, we study the impact of using an ensemble of predictors on the prediction accuracy. Ensemble of predictors can achieve better bias–variance tradeoff than a single predictor by combining the predictions of many predictors. Each predictor in the ensemble is trained with 75% random sampling of the training data. This sampling step is repeated for every model, which increases the diversity of the models and hence improve the prediction accuracy of their combined results. As shown in Fig. 5, a single DNN shows a very poor prediction performance with an R-squared value of 0.16 and an RMSE of 5671. However, a significant improvement is observed when we combine the predictions (by taking the average) of 5 or more models. We notice that there is a diminishing improvement in the performance prediction (for both metrics) after increasing the size of the ensemble beyond N = 15.

We evaluate three possible single server prediction models for evaluating the prediction of throughput for NoSQL databases such as Redis and Cassandra. The prediction is done for different configurations of the NoSQL DB as well as the cloud VM on which the VM is hosted. In each case, we use a Random Forest using 75%:25% for training and prediction.

TABLE 1 Comparison of Different Single-Server Prediction Techniques. OptimusCloud, Which is Our Cloud-Based Optimization Technique Achieves Better Performance in Terms of R^2 and *RMSE* Over All Baselines

Workload	MG-RAST		BUS		HPC	
Metric	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
N-Solitary Models	0.2	3401.4	0.127	109.9	0.04	2778
Selecta	-0.14	4149.3	0.66	110.6	0.932	2451
OptimusCloud -Categorical	0.41	1334.2	0.986	21.87	0.983	1172.9
OptimusCloud -Numerical	**0.89**	**1260.9**	**0.988**	**19.77**	**0.986**	**1076.2**

1) *N-Solitary-Models*: This builds a separate prediction model per VM instance type (referred to as "architecture"). It predicts the performance of a given architecture/configuration combination using previously collected data points from the same architecture. Thus, there is no scope for transfer learning here.

2) *Combined-Categorical*: This builds a combined model using all points from all architectures, while it represents the architecture as a categorical parameter (with integral values). Thus, knowledge transfer is limited across architectures, *e.g.*, C4.large to C4.xlarge, for AWS VM instances.

3) *Combined-Numerical*: This also builds a combined model for all architectures. However, it describes the architecture in terms of its resources *e.g.*, C4.large is represented as vCPU 8, RAM 3.75 GB, Network-Bandwidth 0.62 Gbits/s. This model allows extrapolation across all VM architectures, even across cloud vendors. We test the accuracy of each predictor using the same number of data points (100 points per architecture) and show the result in terms of R^2 (Table 1). We see that using a separate model per architecture (N-Solitary Models) gives poor performance due to the lack of knowledge transfer between architectures. Further, the numerical representation shows a significant improvement in prediction performance over categorical from better knowledge transfer across architectures.

Interpretable and interoperable data analytics: Although there are several predictive models, it is important that the analysis be interpretable. Molnar *et al.* [54] explains the importance of interpretable ML. This could help clients better understand why a model is predicting a certain intervention for crop growth. Vellido, Martín-Guerrero, and Lisboa [55] provides an overview of several works that address the issue of making ML techniques more interpretable. Dimensionality reduction is a popular choice here. A comparison of such existing techniques is provided in [56]. After dimensionality reduction, it becomes tractable to rank order the different features by their importance, providing insights [57]. Another aspect of interpretability is to resolve post-hoc "sanity check" questions [58], such as, after a rain event, does the model predict the moisture content in the soil is higher. If the model fails to answer correctly a sanity-check question, then further examination of the model is done. An important aspect of interpretability is to allow domain experts to parametrize the models, *e.g.*, by feeding the appropriate parameter values for different regions of operation. This can be fed in as manual input by experts or by blending simulation models (typically built by experts) into the data analytics models. Another important design consideration is **interoperability** of the software systems used to acquire and analyze data. Too often software vendors provide silo-ed software packages either due to lack of design consideration or due to conscious decision to force vendor lock-in. In all such cases, it becomes a burden, often insurmountable, to make them interoperate [59].

V. EDGE COMPUTING AND LOW-POWER COMMUNICATION TECHNOLOGIES

Here we discuss recent developments in edge computing and low-power communication technologies that aid our use domain. Edge computing is rapidly evolving, so it is timely to consider the lessons that can be learned and further customized for digital agriculture. The use of low-power communication is a common requirement in several domains. However, our target domain places some distinctive requirements and opportunities that we discuss next.

Microservices and edge-cloud partitioning for low-latency communications: The world of connected devices has fueled the IoT era, where applications rely on a multitude of devices aggregating and processing data sets across highly heterogeneous networks. In this context, distributed deployment alongside containerization of the different information channels will shield the systems from isolated failures, conferring resiliency. The other important aspect is the partitioning of the data stream for computing at different degrees of latency—computing nodes at the edge are used for user-facing applications (face recognition, reconnaissance from a video stream, etc.). Owners will react negatively if the computers become unusable due to intermittent edge analytics. Therefore, the prioritization of the processes needs to change dynamically. In contrast, the application itself needs to be designed in a way that it is insensitive to such dynamic, and unpredictable, changes to the priority level, *e.g.*, it will not time out if there are client-server interactions. Another aspect of the prioritization is that the different analytics results are needed with vastly differing timing requirements. Such high-level, user-expressed requirements will be used to dynamically prioritize in the face of unpredictable arrivals of the events (*e.g.*, a flash flood event, or onset of a locust infestation). Thus, overall the partitioning needs to happen in a top-down or bottom-up manner. Top-down means that we take the high-level user requirements on latency and accuracy and define the partitioning based on that. Bottom-up means that depending on the available resources on each platform, the resource handler decides where to run the application component. Top-down requirements naturally have a higher priority. This will leverage the significant amount of work that has been done in automatic partitioning of applications to run on mobile devices and the cloud [60]–[64].

In-network processing: Sensors used in precision agriculture are *Internet of Things* or IoT devices and like all IoT

devices have constraints on power and connectivity. It is necessary to have the first step of data processing that takes place at the sensor end to be energy efficient. Thus, anomalous data can be suppressed and need not be communicated. Alternately, in some scenarios, the exact opposite is desired—when an anomalous event is detected, that event needs to be communicated to the gateway promptly.

AutoRegressive Integrated Moving Average (ARIMA) is used for fitting time-series data in order to predict or estimate the trend. This can then be used for suppressing data communication — if the sensor node and the receiving gateway node use the same model to predict values and the predicted value is close to the sensed value, the sensor node can suppress the communication [65], [66]. While ARIMA models work on a linear process, more sophisticated ML algorithms can model non-linear processes. Examples of traditional ML techniques such as *k-Nearest Neighbors* (KNNs) and *Support Vector Regression* (SVR) as well as more complex *Recurring Neural Networks* (RNNs) [67] and *Long Short-Term Memory Neural Networks* (LSTMs) [68] have been used for prediction of sensor network values [69], [70]. The predicted values can again be used for suppression of redundant data communication. The tunable configuration allows us to navigate the tradeoff space between accuracy and data communication.

While RNNs and LSTMs give more accurate prediction over statistical methods like ARIMA and ETS, they are more complex in nature. For simpler user queries and in-sensor analytics, ARIMA and ETS models can be used as a lightweight alternative. A more detailed analysis can be done on the cloud backend with some context achieved through *attention pooling* for example.

Low-power communication technologies Low-power communication for wireless IoT communication is key for digital agriculture and falls in three broad categories:

1) Low-power wide area networks (LPWAN), with a greater than 1 km range, essentially low-power versions of cellular networks, with each "node" covering thousands of end devices. Examples include LoRaWAN, Sigfox, DASH7, and weightless.
2) Wireless personal area networks (WPAN), typically ranging from 10 to a few 100 meters. Examples include Bluetooth and Bluetooth Low Energy (BLE), ANT, and ZigBee, which are applicable directly in short-range personal area networks or if organized as mesh networks and with higher transmit power, larger coverage areas.
3) Cellular solution of IoT, including any protocol that are reliant on the cellular connection

Some of the bottlenecks in wireless transmission in farm settings include the harsh physical conditions in farm settings and the proliferation of inexpensive and less reliable sensors coupled with the intrinsic challenges of the LoRaWAN and cellular network technologies, such as 5 G. The core problem in farm settings is to acquire and transfer disparate data sources to support the various demands of a wide range of computation tasks while meeting the stringent constraints of heterogeneous wireless connectivity available in agricultural domain (especially for the livestock application). Data traffic from different sources and for distinct computation tasks raise diverse requirements to wireless connectivity in terms of its availability, bandwidth, responsiveness, resilience and energy efficiency. For example, sparse data relating behavior changes needs highly responsive, highly reliable communication; streaming videos captured by the drones[3] or surveillance cameras expects high bandwidth but is elastic to varying-bandwidth with appropriate rate adaptation; A large amount of livestock herd-level information can be delay-tolerant but requires energy-efficient connectivity. Further, compared to wireless networks in the urban and civil uses, wireless networking for the livestock experiences more practical challenges regarding coverage holes, fast-fading channels, high interference and time-varying performance. Wireless connectivity is not always available in the wild rural areas: WiFi or other local-area communication (white space, ZigBee, LoRA) to the edge is unavailable when the cattle is far away from the farm facility, and cellular connectivity is missing at places given the poor coverage in the countryside.

In most operational large-scale farms in the US, WiFi connectivity is not present. While cellular providers in the US do provide cellular plans for IoT devices (prominently NB-IoT), the cost of such plans makes them infeasible for farms, considering that the cost is per device and we anticipate many IoT devices in a large-scale farm. Hence, the use of LoRa, and its variants LoRaMesh or LoRaWAN, is popular. The IoT orchestrators that are provided by Microsoft, Amazon, and Google, *i.e.*, Microsoft Azure IoT, AWS IoT Greengrass Greengrass, and Google IoT Core, can also leverage any of these networking modalities. However, Azure IoT uses a hybrid of per-device and data-based pricing system and AWS Greengrass uses a per-device pricing system, while Google's IoT Core uses a pricing system based on data volume alone. Thus, all existing commercial edge-based orchestrators need additional operational costs to function. So using a non-metered network like LoRa and using homegrown orchestration software will avoid these additional charges. A suite of techniques for data acquisition and wireless networking must work in concert, and at a price-point, to meet the diverse demands in digital agriculture.

Agricultural data types relevant to algorithmic design:

It is important to set up the data architecture prior to ingestion of the data. A broad and yet useful principle relates to the FAIR data principles, which refer to the Findability, Accessibility, Interoperability, and Reuse of digital assets. This is particularly important as in this domain there is close interaction between man and machine throughout the data

[3]The use of drones is gaining favor in large farm settings because they are often not set up for WiFi or other wireless coverage. Therefore drones can act as data ferries for visiting the sensor nodes and acquiring the data, when the data is not latency sensitive. For example, time-lapse video showing the growing characteristics of plants can be brought back by the drone to the backend for heavy compute processing. The use of drone preserves the low power operation of the sensor nodes as the expensive long-range wireless communication is avoided and the low duty cycle of the nodes is maintained (the nodes are woken up by the drones on an on-demand basis).

pipeline. Now let us consider the primary types of data as they relate to our theme here. **Batch Data** mainly includes information collected from sensing devices, such as information of cropping, feeding, waste, livestock, etc. As the data is often sparse in time and space, it has a loose synchronization requirement and does not require a continuous wireless connectivity. Therefore, batch data can be periodically recorded and asynchronously updated. The main limitation for sensing devices is that they are usually battery powered and have limited operational lifetime. An optimized device management and transmission protocol is quite essential for energy efficient data acquisition and transmission, device lifetime maximization and sustainable development of the livestock system. **Streaming Data** is usually dynamic video/audio recoded by surveillance cameras or drones for real-time monitoring and early anomaly detection. Such data, especially for real-time monitoring, analytics and diagnosis is time sensitive and demands a continuous high-volume bandwidth for intensive content storage and delivery. A joint procedure of data acquisition and transmission is thus required to adapt the high data volume, handle varying wireless conditions, and support a reliable information flow.

Multi-class data acquisition, storage, and processing: We consider two main classes: delay-sensitive and delay-tolerant. Delay-sensitive data is used for real-time monitoring, early anomaly detection and other highly-responsive tasks. It is further divided into multiple sub-classes depending on their bandwidth requirement: continuous high bandwidth (e.g., streaming videos recorded by the cameras or drones), continuous low bandwidth (e.g, streaming audio/voice, GPS or other in-cattle sensing data), or instantaneously-available but short connectivity (e.g., information like critical alarms).

Delay-tolerant data is used for strategic tasks and includes sensing data collected for cropping, feeding, waste, livestock, etc. For different classes of data, we propose to develop a suite of techniques: 1) task-aware sampling schemes to reduce data volumes but retain data samples essential to the computation tasks, 2) content-aware data aggregation and compression to eliminate statistical redundancy and adjust data volume to fit the dynamic channel capacity and varying wireless condition (*e.g.*, content-aware video frame compression).

Integrated network over heterogeneous wireless communication: We have designed an integrated network to enable both delay-sensitive communication and delay-tolerant communication (in an opportunistic manner) over heterogeneous wireless connectivity [8]. We consider both infrastructure-based and ad hoc modes. Hybrid mode is also designed for whereby data resides at a location till infrastructure becomes available such as a drone as a data ferry. Network adaptation needs to be performed at multiple levels spanning from wireless technologies (e.g., cellular, cellular-IoT, WiFi, ZigBee, LoRA etc), transport layer (MPTCP, delay-driven optimization), lower-layer techniques like resource allocation, scheduling and rate adaption, to name a few. Many techniques can be modified at runtime to further optimize data acquisition and transmission under time-varying conditions. These include predictive optimization based on time varying wireless connectivity and data requirements and joint optimization of local storage, local processing, and wirelesss transmissions.

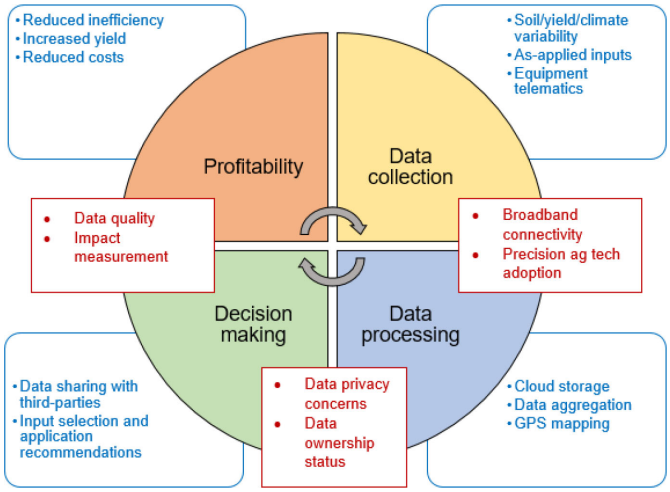


FIGURE 6. Farm economic decision making pipeline and barriers to profitability in digital agriculture.

VI. ECONOMICS, POLICY, AND DECISION MAKING

The most relevant policy questions in digital agriculture and related business data and are summarized in Fig. 6. Though farm data enjoy some of the IP protections afforded to trade secrets, its legal ownership remains ambiguous [71]. The decision to subscribe to a data service provider stem from fears of personally identifiable information (PII) being misappropriated. Yet, farm data generates positive network externalities when aggregated across a large number of operations.

Profitability and on-farm decision making: Site-specific farm management has the potential to enhance farm profitability while conserving resources. But despite advancements in field sensors, GPS guidance, and grid soil sampling, adoption by farmer operators has fallen short of expectations. Operator demographics, operation size, and perceived benefits influence the decision to invest in site-specific management practices. Subsequent economic returns to adoption depend on the nature of the adopted technology and its interface with on-farm decision making. Miller *et al.* [72] identify two types of precision agriculture technologies: embodied knowledge—tools that generate value in isolation, *e.g.*, GPS guidance systems or automatic section control—and information intensive—tools that produce data for use in future decision making such as yield monitors, grid soil sampling, or electrochemical sensors. Embodied knowledge technologies create immediate convenience while the benefits of information-intensive technologies are revealed over a longer time horizon and depend on their role in the on-farm decision making process. Differences in the immediacy and measurability of realized gains may explain differences in adoption rates across technology types. Use of VRT and GPS soil mapping, for example, has consistently lagged that of

GPS guidance systems. Rather than assessing technologies in isolation, agricultural economists are increasingly interested in how producers bundle complementary tools to create an overall precision technology strategy. Profitable use of "hard technologies" such as variable rate planters and fertilizer spreaders depends crucially on the availability of accurate intra-field soil data, or "soft technology" inputs [1]. These data sources however, are themselves costly to obtain. Moreover, the optimal data collection frequency or sampling density is not obvious and likely varies by field. Unless the economic returns to site-specific management cover both the up-front investment and the cost of collecting and using actionable information, adoption will be low.

Policy: The most relevant policy questions in digital agriculture regard the value and legal status of farm data. Agricultural data generates benefits when aggregated across a large number of farm operations. These benefits—referred to as network effects or network externalities—grow with the number of participants. Business models such as Farmers Business Network, Inc. (FBN) have demonstrated the value of data sharing through its crowd-sourced database of input costs and performance benchmarking. But farm data often remains siloed within the farm gate. An individual farm's data leads to more reliable recommendations when pooled with comparable operations using similar practices and inputs. To overcome this "small data" problem, the perceived benefits of joining a big data community must exceed the perceived costs—*e.g.*, privacy concerns over data misappropriation. To better understand the privacy concerns of producers, the nature and legal status of agricultural data must be considered. Miller *et al.* [71] discuss farm data's place on the private-public good spectrum. For a good to be a "private good," *i.e.*, its benefits and costs are fully realized by the owner, it must be both rivalrous—consumption by one party prohibits the consumption of another—and excludable—access to the good can be restricted. A private good allows for the highest possible degree of privacy protections for the owner. Copies of farm data can be shared without inhibiting its use by the original owner. In this way, farm data is clearly non-rivalrous.

The ability of a farm operator to exclude others from using their data depends on their relationships with data service providers and the data sharing agreements that govern those relationships. For example, equipment manufacturers collect telematics data on new products they sell for improving performance and service. The equipment owner has no reasonable expectation of excludability and may not even be aware that they opted into such an agreement. Farms that subscribe to a data service provider to manage and analyze their data, *e.g.*, Climate FieldView, are similarly forfeiting excludability. However, data may be partially excludable if access is limited to the network, or "club" of subscribers. Farm data most closely satisfies the definition of a "club good", meaning the degree to which a farmer's privacy is at risk depends on how extensive and excludable their data network is [71].

A farmer's data is not legally protected from disclosure in the way medical records are protected by the Health Insurance Portability and Accountability Act (HIPAA) or education information is protected by the Family Educational Rights and Privacy Act (FERPA). Without overarching legal safeguards for farm data, individual sharing agreements dictate the terms of access and use. Though farm data enjoy some of the intellectual property protections afforded to trade secrets, its legal ownership structure remains ambiguous. There may be a role for liability insurance that addresses unintended consequences of information sharing. Over large regions, such data might be reported on a non-attributional basis, thus shielding specific individuals. The flip side of this is that this reduces the precision of mitigation actions, *e.g.*, for a specific geographical region. Possible solutions to the conundrum might consider cryptography coupled to access keys, facilitating access to large databases, without revealing specific information to those who do not hold keys. Keys could be made available based on a regional user group, and to those who contributed to the database. For further sophisticated use cases, secure multi-party computation can be used, allowing a group of members of a minimum size coming together to access the information, but individuals or smaller-sized groups cannot. For example, homomorphic encryption in federated computing serves this purpose of balancing privacy and data utility.

VII. LOOKING AHEAD

Big data and precision agriculture will likely be a disruptive force in the farm economy over the medium to long-term range. Digital agriculture, with the incorporation of Internet-of-Things (IoT)-based technologies, presents the ability to evaluate a system at multiple levels (individual, local, regional, and global) and generate tools that allow for improved decision making in every sub-process related to digital agriculture. In this article, we have reviewed the different types of datasets and relevant data science processing algorithms in typical field operations together with the typical lifecycle for the data to be contributory to the digital farm economy. We have then discussed the optimized NoSQL-based data storage solutions. Then, we have developed the idea of Machine Learning being adapted for use in digital agriculture, which means putting domain-specific requirements regarding interpretability, distribution, ability to handle intermittent wireless connectivity, and low cost. Finally, we have reported on results from real-world data from real-world testbeds with respect to which features are important in the analytics and the performance of analytics-based prediction. We conclude by discussing the policy challenges, the farm economic decision making pipeline, and barriers to profitability in digital agriculture.

REFERENCES

[1] D. S. Bullock, M. L. Ruffo, D. G. Bullock, and G. A. Bollero, "The value of variable rate technology: An information-theoretic approach," *Amer. J. Agricultural Econ.*, vol. 91, no. 1, pp. 209–223, 2009.

[2] S. Chaterji, E. H. Ahn, and D.-H. Kim, "CRISPR genome engineering for human pluripotent stem cell research," *Theranostics*, vol. 7, no. 18, 2017, Art. no. 4445.

[3] R. Kumar, A. Y. Mahgoub, D. Buckmaster, and S. Chaterji, "Yield trends of corn and soybean in the midwest," in *Proc. ASABE Annu. Int. Meeting*, Amer. Soc. Agricultural Biol. Engineers, 2019, pp. 1–5.

[4] R. Xu *et al.*, "ApproxNet: Content and contention-aware video object classification system for embedded clients," *ACM Trans. Sensor Netw. (TOSN)*, pp. 1–27, 2021.

[5] R. Xu *et al.*, "Approxdet: Content and contention-aware approximate object detection for mobiles," in *Proc. 18th Conf. Embedded Networked Sensor Syst.*, 2020, pp. 449–462.

[6] S. Chaterji, J. Koo, N. Li, F. Meyer, A. Grama, and S. Bagchi, "Federation in genomics pipelines: Techniques and challenges," *Brief. Bioinf.*, vol. 20, no. 1, pp. 235–244, 2019.

[7] S. Chaterji *et al.*, "Resilient cyberphysical systems and their application drivers: A technology roadmap," 2019, *arXiv:2001.00090*.

[8] X. Jiang *et al.*, "Hybrid low-power wide-area mesh network for IoT applications," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 901–915, Jan. 2021.

[9] D. Kim, L. Xue, D. Li, Y. Zhu, W. Wang, and A. O. Tokuta, "On theoretical trajectory planning of multiple drones to minimize latency in search-and-reconnaissance operations," *IEEE Trans. Mobile Comput.*, vol. 16, no. 11, pp. 3156–3166, 2017.

[10] A. W. Malik, A. U. Rahman, T. Qayyum, and S. D. Ravana, "Leveraging fog computing for sustainable smart farming using distributed simulation," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3300–3309, Apr. 2020.

[11] E. Cuervo *et al.*, "Maui: Making smartphones last longer with code offload," in *Proc. 8th Int. Conf. Mobile Syst., Appl., Serv.*, 2010, pp. 49–62.

[12] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[13] R. Newton, S. Toledo, L. Girod, H. Balakrishnan, and S. Madden, "Wishbone: Profile-based partitioning for sensornet applications," in *Proc. NSDI*, vol. 9, 2009, pp. 395–408.

[14] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. Liu, "Big data cleaning based on mobile edge computing in industrial sensor-cloud," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1321–1329, Feb. 2020.

[15] K. Shankar, P. Wang, R. Xu, A. Mahgoub, and S. Chaterji, "Janus: Benchmarking commercial and open-source cloud and edge platforms for object and anomaly detection workloads," in *Proc. IEEE 13th Int. Conf. Cloud Comput.*, 2020, pp. 590–599.

[16] A. Y. Mahgoub, R. Kumar, and S. Chaterji, "Iris: Tuning the configuration parameters of NoSQL databases for high-throughput digital agricultural processing pipelines," in *Proc. ASABE Annu. Int. Meeting*, 2019, pp. 1–9.

[17] A. Mahgoub *et al.*, "SOPHIA: Online reconfiguration of clustered nosql databases for time-varying workloads," in *Proc. USENIX Annu. Tech. Conf.*, Renton, WA, USA, 2019, pp. 223–240.

[18] A. Mahgoub *et al.*, "OPTIMUSCLOUD: Heterogeneous configuration optimization for distributed databases in the cloud," in *Proc. USENIX Annu. Tech. Conf.*, 2020, pp. 189–203.

[19] A. Mahgoub, K. Shankar, S. Mitra, A. Klimovic, S. Chaterji, and S. Bagchi, "SONIC: Application-aware data passing for chained serverless applications," in *Proc. USENIX Annu. Tech. Conf., (Virtual), USENIX Assoc.*, 2021, pp. 1–15.

[20] T. E. Thomas, J. Koo, S. Chaterji, and S. Bagchi, "Minerva: A reinforcement learning-based technique for optimal scheduling and bottleneck detection in distributed factory operations," in *Proc. 10th Int. Conf. Commun. Syst. Netw.*, 2018, pp. 129–136.

[21] Y. He *et al.*, "Antibiotic resistance genes from livestock waste: Occurrence, dissemination, and treatment," *npj Clean Water*, vol. 3, no. 1, pp. 1–11, 2020.

[22] N. Udikovic-Kolic, F. Wichmann, N. A. Broderick, and J. Handelsman, "Bloom of resident antibiotic-resistant bacteria in soil following manure fertilization," *Proc. Nat. Acad. Sci.*, vol. 111, no. 42, pp. 15202–15207, 2014.

[23] D. Vasisht *et al.*, "Farmbeats: An IoT platform for data-driven agriculture," in *Proc. 14th USENIX Symp. Networked Syst. Des. Implementation*, 2017, pp. 515–529.

[24] S. Sharma, R. Singh, and A. Rai, "Integrated national agricultural resources information system (inaris)," *Indian Agricultural Statist. Res. Inst.*, New Delhi, 2000.

[25] J. Woodard, "Big data and ag-analytics: An open source, open data platform for agricultural & environmental finance, insurance, and risk," *Agricultural Finance Rev.*, vol. 76, no. 1, pp. 15–26, 2016.

[26] S. Nilakanta, K. Scheibe, and A. Rai, "Dimensional issues in agricultural data warehouse designs," *Comput. Electron. Agriculture*, vol. 60, no. 2, pp. 263–278, 2008.

[27] D. Roy, S. Sarkar, and S. Ghose, "A comparative study of learning object metadata, learning material repositories, metadata annotation & an automatic metadata annotation tool," *Adv. Semantic Comput.*, vol. 2, no. 2010, pp. 103–126, 2010.

[28] O. Fiehn, G. Wohlgemuth, and M. Scholz, "Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata," in *Proc. Int. Workshop Data Integration Life Sci.*, Berlin, Heidelberg, Germany: Springer, 2005, pp. 224–239.

[29] S. Haug and J. Ostermann, "A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks," in *Proc. Eur. Conf. Comput. Vis.*, Cham: Springer, 2014, pp. 105–116.

[30] A. Schoofs, A. Guerrieri, D. T. Delaney, G. M. O'Hare, and A. G. Ruzzelli, "Annot: Automated electricity data annotation using wireless sensor networks," in *Proc. 7th Annu. IEEE Commun. Soc. Conf. Sensor, Mesh Ad Hoc Commun. Netw.*, 2010, pp. 1–9.

[31] G. Simbahan, A. Dobermann, and J. Ping, "Screening yield monitor data improves grain yield maps," *Agronomy J.*, vol. 96, no. 4, pp. 1091–1102, 2004.

[32] W. Sun, B. Whelan, A. B. McBratney, and B. Minasny, "An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management," *Precis. Agriculture*, vol. 14, no. 4, pp. 376–391, 2013.

[33] M. D. Steven, "Satellite remote sensing for agricultural management: Opportunities and logistic constraints," *ISPRS J. Photogrammetry Remote Sens.*, vol. 48, no. 4, pp. 29–34, 1993.

[34] E. Honkavaara *et al.*, "Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight UAV spectral camera for precision agriculture," *Remote Sens.*, vol. 5, no. 10, pp. 5006–5039, 2013.

[35] A. Loreto and M. Morgan, "Development of an automated system for field measurement of soil nitrate," *Trans. ASAE*, 1996.

[36] E. Murakami, A. M. Saraiva, L. C. R. Junior, C. E. Cugnasca, A. R. Hirakawa, and P. L. Correa, "An infrastructure for the development of distributed service-oriented information systems for precision agriculture," *Comput. Electron. Agriculture*, vol. 58, no. 1, pp. 37–48, 2007.

[37] K. Shankar, P. Wang, R. Xu, A. Mahgoub, and S. Chaterji, "JANUS: Benchmarking commercial and open-source cloud and edge platforms for object and anomaly detection workloads," in *Proc. IEEE 13th Int. Conf. Cloud Comput.*, 2020, pp. 590–599.

[38] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low power wide area networks: An overview," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 855–873, Apr.–Jun. 2017.

[39] S. Mitra, M. K. Gupta, S. Misailovic, and S. Bagchi, "Phase-aware optimization in approximate computing," in *Proc. IEEE/ACM Int. Symp. Code Gener. Optim.*, 2017, pp. 185–196.

[40] P. V. Kotipalli, R. Singh, P. Wood, I. Laguna, and S. Bagchi, "Ampt-ga: Automatic mixed precision floating point tuning for gpu applications," in *Proc. ACM Int. Conf. Supercomputing*, 2019, pp. 160–170.

[41] Z. Qin, Y. Gao, M. D. Plumbley, and C. G. Parini, "Wideband spectrum sensing on real-time signals at sub-Nyquist sampling rates in single and cooperative multiple nodes," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3106–3117, Jun. 2016.

[42] D.-H. Shin and S. Bagchi, "An optimization framework for monitoring multi-channel multi-radio wireless mesh networks," *Ad Hoc Netw.*, vol. 11, no. 3, pp. 926–943, 2013.

[43] Z. Wen *et al.*, "ApproxIoT: Approximate analytics for edge computing," in *Proc. IEEE 38th Int. Conf. Distrib. Comput. Syst.*, 2018, pp. 411–421.

[44] K. Thein, "Apache kafka: Next generation distributed messaging system," *Int. J. Sci. Eng. Technol. Res.*, vol. 3, no. 47, pp. 9478–9483, 2014.

[45] E. Baralis and T. Cerquitelli, "Selecting representatives in a sensor network.," in *Proc. SEBD*, 2006, pp. 351–360.

[46] M. Ester *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Kdd*, vol. 96, pp. 226–231, 1996.

[47] Y. Kotidis, "Snapshot queries: Towards data-centric sensor networks," in *Proc. 21st Int. Conf. Data Eng.*, 2005, pp. 131–142.

[48] S. Kartakis and J. A. McCann, "Real-time edge analytics for cyber-physical systems using compression rates," in *Proc. 11th Int. Conf. Autonomic Comput.*, 2014, pp. 153–159.

[49] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst.*, 2017, pp. 328–339.

[50] J. Koo, J. Zhang, and S. Chaterji, "Tiresias: Context-sensitive approach to decipher the presence and strength of microrna regulatory interactions," *Theranostics*, vol. 8, no. 1, pp. 277–291, 2018.
[51] R. Cai *et al.*, "VIBNN: Hardware acceleration of Bayesian neural networks," *ACM SIGPLAN Notices*, vol. 53, no. 2, pp. 476–488, 2018.
[52] J. L. Carlson, *Redis in Action*. Manning Publications Co., CT, United States, 2013. [Online]. Available: https://www.manning.com/books/redis-in-action
[53] D. Vasisht *et al.*, "Farmbeats: An IoT platform for data-driven agriculture," in *Proc. 14th USENIX Symp. Networked Syst. Des. Implementation*, 2017, pp. 515–529.
[54] C. Molnar, Interpretable Machine Learning. United States: Lulu.com, 2020. [Online]. Available: https://www.google.com/books/edition/Interpretable_Machine_Learning/jBm3DwAAQBAJ?hl=en&gbpv=0
[55] A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable," in *Proc. ESANN*, vol. 12, 2012, pp. 163–172.
[56] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative," *J. Mach. Learn. Res.*, vol. 10, pp. 66–71, 2009.
[57] S. G. Kim, N. Theera-Ampornpunt, C.-H. Fang, M. Harwani, A. Grama, and S. Chaterji, "Opening up the blackbox: An interpretable deep neural network-based classifier for cell-type specific enhancer predictions," *BMC Syst. Biol.*, vol. 10, no. 2, pp. 243–258, 2016.
[58] D. Inouye, L. Leqi, J. S. Kim, B. Aragam, and P. Ravikumar, "Automated dependence plots," in *Proc. Conf. Uncertainty Artif. Intell.*, 2020, pp. 1238–1247.
[59] *Federal Communication Commission (FCC) Precision Ag Connectivity Task Force*, "Encouraging adoption of precision agriculture and availability of high-quality jobs on connected farms," [Online]. Available: https://www.fcc.gov/sites/default/files/precision-ag-adoption-jobs-wg-report-10282020.pdf, 2020.
[60] D. Kovachev, T. Yu, and R. Klamma, "Adaptive computation offloading from mobile devices into the cloud," in *Proc. IEEE 10th Int. Symp. Parallel Distrib. Process. Appl.*, 2012, pp. 784–791.
[61] H. Mora, J. F. Colom, D. Gil, and A. Jimeno-Morenilla, "Distributed computational model for shared processing on cyber-physical system environments," *Comput. Commun.*, vol. 111, pp. 68–83, 2017.
[62] G. Li, L. Liu, X. Wang, X. Dong, P. Zhao, and X. Feng, "Auto-tuning neural network quantization framework for collaborative inference between the cloud and edge," in *Proc. Int. Conf. Artif. Neural Netw.*, Cham: Springer, 2018, pp. 402–411.
[63] S. Dey, A. Mukherjee, A. Pal, and P. Balamuralidhar, "Partitioning of cnn models for execution on fog devices," in *Proc. 1st ACM Int. Workshop Smart Cities Fog Comput.*, 2018, pp. 19–24.
[64] J. Wang, J. Pan, F. Esposito, P. Calyam, Z. Yang, and P. Mohapatra, "Edge cloud offloading algorithms: Issues, methods, and perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–23, 2019.
[65] Y. Sangar and B. Krishnaswamy, "Wichronos: Energy-efficient modulation for long-range, large-scale wireless networks," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, 2020, pp. 1–14.
[66] Y. Zhu and R. Sivakumar, "Challenges: Communication through silence in wireless sensor networks," in *Proc. 11th Annu. Int. Conf. Mobile Comput. Netw.*, 2005, pp. 140–147.
[67] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.
[68] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[69] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep lstm-rnn," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 2727–2740, 2019.
[70] Z. Lv, J. Xu, K. Zheng, H. Yin, P. Zhao, and X. Zhou, "LC-RNN: A deep learning model for traffic speed prediction," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 3470–3476.
[71] N. J. Miller, T. Griffin, P. Goeringer, A. Ellixson, and A. Shanoyan, "Estimating value, damages, & remedies when farm data are misappropriated," Choices. Quart. 4. Available online: http://www.choicesmagazine.org/choices-magazine/submitted-articles/estimating-value-damages-and-remedies-when-farm-data-are-misappropriated
[72] N. J. Miller, T. W. Griffin, I. A. Ciampitti, and A. Sharda, "Farm adoption of embodied knowledge and information intensive precision agriculture technology bundles," *Precis. Agriculture*, vol. 20, no. 2, pp. 348–361, 2019.

SOMALI CHATERJI received the Ph.D. degree in biomedical engineering from Purdue University, West Lafayette, IN, USA. She is currently an Assistant Professor with the Department of Agricultural and Biological Engineering, Purdue University, where she leads the Innovatory for Cells and Neural Machines (ICAN). ICAN specializes in developing algorithms and statistical machine learning models for IoT and cloud/edge computing on the one hand and genome engineering on the other. She followed this up with two post-doctoral stints, in Biomedical Engineering from UT Austin and in Computer Science from Purdue University. Dr. Chaterji is a part of three centers at Purdue: CRISP (Center for Resilient Infrastructures, Systems, and Processes), OATS (The Open Ag Technology and Systems Center), and WHIN (Wabash Heartland Innovation Network) — a digital agriculture and digital manufacturing consortium harnessing the power of Internet-enabled sensors to develop the 10-county region in north-central Indiana into a global epicenter of digital agriculture and next-generation manufacturing. She was the recipient of the Chorafas Foundation Fellowship.

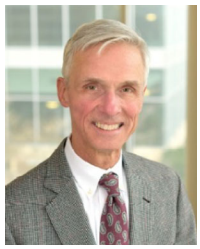
NATHAN DELAY received the B.S. degree in business administration from Rocky Mountain College, Billings, MT, USA, in 2009, the M.A. degree in economics from the University of Colorado Denver, Denver, CO, USA, in 2013, and the Ph.D. degree in economics from Washington State University, Pullman, WA, USA, in 2018. In 2018, he joined the Department of Agricultural Economics, Purdue University, West Lafayette, IN, USA, as an Assistant Professor. His research encompasses the economics of digital agriculture, production economics, and farm policy. His current research projects include estimating the value of farm data and its role in farmland markets.

JOHN EVANS received the B.S. and M.S. degrees in biosystems engineering from the University of Kentucky, Lexington, KY, USA, and the Ph.D. degree in biological engineering form the University of Nebraska, Lincoln, NE, USA. He is currently an Assistant Professor with the Department of Agricultural and Biological Engineering at Purdue. He is passionate about all things agriculture and how technology can be harnessed to help farmers feed a growing population. His research efforts focus on digital agriculture, machine logistics, and automation.

NATHAN MOSIER received the Ph.D. degree in agricultural and biological engineering from Purdue University. He is the Indiana Soybean Alliance New Uses Professor and the Head of Agricultural and Biological Engineering, Purdue University, West Lafayette, IN, USA. Dr. Mosier's research addresses fundamental topics in processing of agricultural materials such as grain, straw, and stover into biofuels, chemicals, and consumer products. Students from his research group have gone on to work in biofuels, biocatalysis, pharmaceutical, and food industries applying engineering principles to innovatively solve challenges in sustainable manufacturing, health, and food. Agricultural and Biological Engineering (ABE) faculty at Purdue has educational, research, and extension programs that span the agricultural value chain from producers to ag machinery to water resources, to food and pharmaceuticals. ABE faculty focus on solving the engineering challenges that enable new technologies that make agriculture more productive, more sustainable, and more profitable.

BERNARD ENGEL received the B.S. and M.S. degrees in agricultural engineering from the University of Illinois at Urbana Champaign, Urbana-Champaign, IL, USA, and the Ph.D. degree in agricultural engineering from Purdue University, West Lafayette, IN, USA. He is an Associate Dean for Agricultural Research and Graduate Education and a Professor of Agricultural and Biological Engineering, Purdue University. His research interests include geospatial sciences applications in agricultural, natural resources, and environmental sciences with a focus in hydrologic or water quality modeling.

DENNIS BUCKMASTER received the B.S. degree in agricultural engineering from Purdue University, West Lafayette, IN, USA, and the M.S. and Ph.D. degrees in agricultural engineering from Michigan State University, East Lansing, MI, USA. He was a faculty member with Penn State University in Agricultural and Biological Engineering with focus in forage systems engineering, machinery design, and farm systems modeling. He was an Assistant Dean and Associate Director for Academic Programs in the Purdue College of Agriculture. He is currently a Professor of agricultural and biological engineering and Dean's Fellow for Digital Agriculture. He co-coordinates and teaches in the Purdue Agricultural Systems Management program. As Co-Founder of the Open Ag Technology and Systems Center, much of his work revolves around the entire data pipeline in agricultural systems using and contributing to open source technologies. He was the recipient of the Purdue Agriculture Kohls Outstanding Teacher Award and is passionate about instruction in digital agriculture and data science for agriculture.



MICHAEL R. LADISCH received the B.S. degree from Drexel University, Philadelphia, PA, USA, in 1973, and the M.S. and Ph.D. degrees in chemical engineering from Purdue University, West Lafayette, IN, USA, in 1974 and 1977, respectively. He is the Director of the Laboratory of Renewable Resources Engineering (LORRE), and a Distinguished Professor of agricultural and biological engineering with a joint appointment in the Weldon School of Biomedical Engineering at Purdue University. He was CTO at Mascoma Corporation from 2007 to 2013, serves on the scientific advisory board of Agrivida, and is on the board for the Foundation for Food and Agricultural Research. Ladisch's research addresses transformation of renewable resources into biofuels and bioproducts, bioseparations, injectable biologics, and pathogen detection.

RANVEER CHANDRA (Fellow, IEEE) received the undergraduation degree from the Indian Institute of Technology Kharagpur, Kharagpur, India and the Ph.D. degree in computer science from Cornell University, Ithaca, NY, USA. He is the CTO of Agri-Food, Managing Director of Research for Industry, Microsoft, and the Head of Networking, Microsoft Research, Redmond, WA, USA. His research has shipped in multiple Microsoft products, including XBOX, Azure, Windows, and Visual Studio. He started the FarmBeats project in 2015, which shipped as a Microsoft product in 2019. Bill Gates featured his work on FarmBeats on GatesNotes, and he has been invited to present to the Secretary of Agriculture, and on TV White Spaces to the FCC Chairman. He has authored or coauthored more than 90 research papers and has more than 100 patents granted by the USPTO. He was the recipient of several awards, including the best paper awards in computer science conferences and the MIT Technology Review's Top Innovators Under 35 award in 2010.