TETRA TECH

Enabling Crop Analytics At Scale

ENABLING CROP ANALYTICS AT SCALE (ECAAS)

Agricultural Ground Data (AGData) Acceleration Facility Innovation Agenda

Contents

Introduction

The Enabling Crop Analytics at Scale (ECAAS) initiative is a global effort supported by the Bill and Melinda Gates Foundation to improve the availability, sharing, and use of the ground truth training data needed to support the development of advanced crop analytics for smallholder farmers using artificial intelligence (AI) and Machine Learning (ML) techniques. ECAAS aims to unlock the tremendous potential of remote sensing and Earth observation in ways that could transform smallholder agriculture.

To support the achievement of ECAAS's goal, the AGData Acceleration Facility aims to catalyze improved collection, processing, sharing, and use of ground-truth data, targeting key bottlenecks and technical challenges that inhibit use of advanced crop analytics in support of smallholder farmers. The facility invests in innovations and partnerships focused on widely applicable use cases.

BOX 1:
AGData Acceleration Facility Priority Use Cases

The ECAAS Initiative mapped and prioritized use cases for its initial phase (2020–2022). Priority use cases include:

- Integrated digital farm management, including tailored extension services and advice;
- Improved smallholder farmer access to financial products, including traditional and non-traditional borrowing, investment, and insurance; and
- Food security monitoring and response, including in the context of food supply chains disrupted by the novel coronavirus outbreak.

By focusing on major use cases, the AGData Acceleration Facility invests in scalable technologies and opportunities that can have large scale positive impacts. During its first phase, ECAAS is prioritizing three use cases (*Box 1*) and focusing in three critical ground data parameters (*Box 2*).

BOX 2:
AGData Acceleration Facility Priority Parameters

The priority parameters required for analytic support to these uses cases include:

- Field boundary identification and area planted;
- Crop type classification, including for intercropped systems and varietal detection; and
- Yield estimation and forecasting.

The AGData Acceleration Facility offers exciting potential to leverage existing research and development (R&D) pipelines, which will in turn will inform the next generation of R&D. Through partnerships with leading R&D institutions, the Facility aims to reduce the time it takes to get technologies ready for product development into testing and commercial applications. The Facility will also support creative partnership models between commercial, civil society, and public organizations to reduce first mover obstacles in technology investments and data limitations, and enable pathways toward a sustainable data sharing ecosystem.

The AGData Acceleration Facility innovation agenda represents the result of multiple in-depth stakeholder consultations and prioritization exercises undertaken to identify areas for investment by the Facility over the next 2-3 years. The agenda is meant to evolve as the crop analytics landscape evolves. The agenda is also intended to support other innovation funders and investors to identify where they can have the greatest impact to catalyze large-scale and cost-effective collection, processing, sharing, and use of ground truth training data that can drive the expansion of use of advanced crop analytics in support of smallholder farmers.

AgData Acceleration Facility Innovation Agenda

Crop analytics that leverage remote sensing and other data using machine learning and artificial intelligence are rapidly advancing. To accelerate progress so that these tools are available to public, private, and research actors working to support smallholder famers to improve their productivity, resilience, and sustainability, the AgData Acceleration Facility will prioritize four areas for innovation across the crop analytics data chain (Figure 1):

1
Standardized data collection tools, quality standards, and alternative data sources

- Developing or deploying standardized mobile or other on-farm data collection tools to reduce the cost and time required for data collection and improve dataset interoperability
- Advancing adoption of guidelines and standards for collection processes
- Identifying optimal and user-centric sampling schemes
- Determining minimum volume and accuracy of training data for applications and accuracy plateaus
- Methods for generating synthetic training data
- Testing of transfer learning methods

2
Methods and systems for semi-automated ingestion and processing

- Improving linkages and interoperability between data collection mechanisms and data hosting
- Streamlining and automating pipelines connecting ground data to accessible hosting platforms
- Development of systems for semi-automated data processing;
- Advancing methods for semi-automated labeling
- Quality Assessment/Quality Control (QA/QC), and resolution enhancement
- Standardizing catalogs through STAC specifications or similar means.

3
Tools, standards and methods for data storage, exchange, and addressing privacy

- Advancing methods, technologies, and standards for data management and dissemination
- Assessing the existing landscape of regulations, laws, and emerging guidelines for data privacy
- Assessing methods for anonymization and impact on model performance
- Testing vulnerabilities in anonymized datasets through re-identification
- Development of an open source Application Programming Interfaces (API)
- Development data marketplaces and exchanges

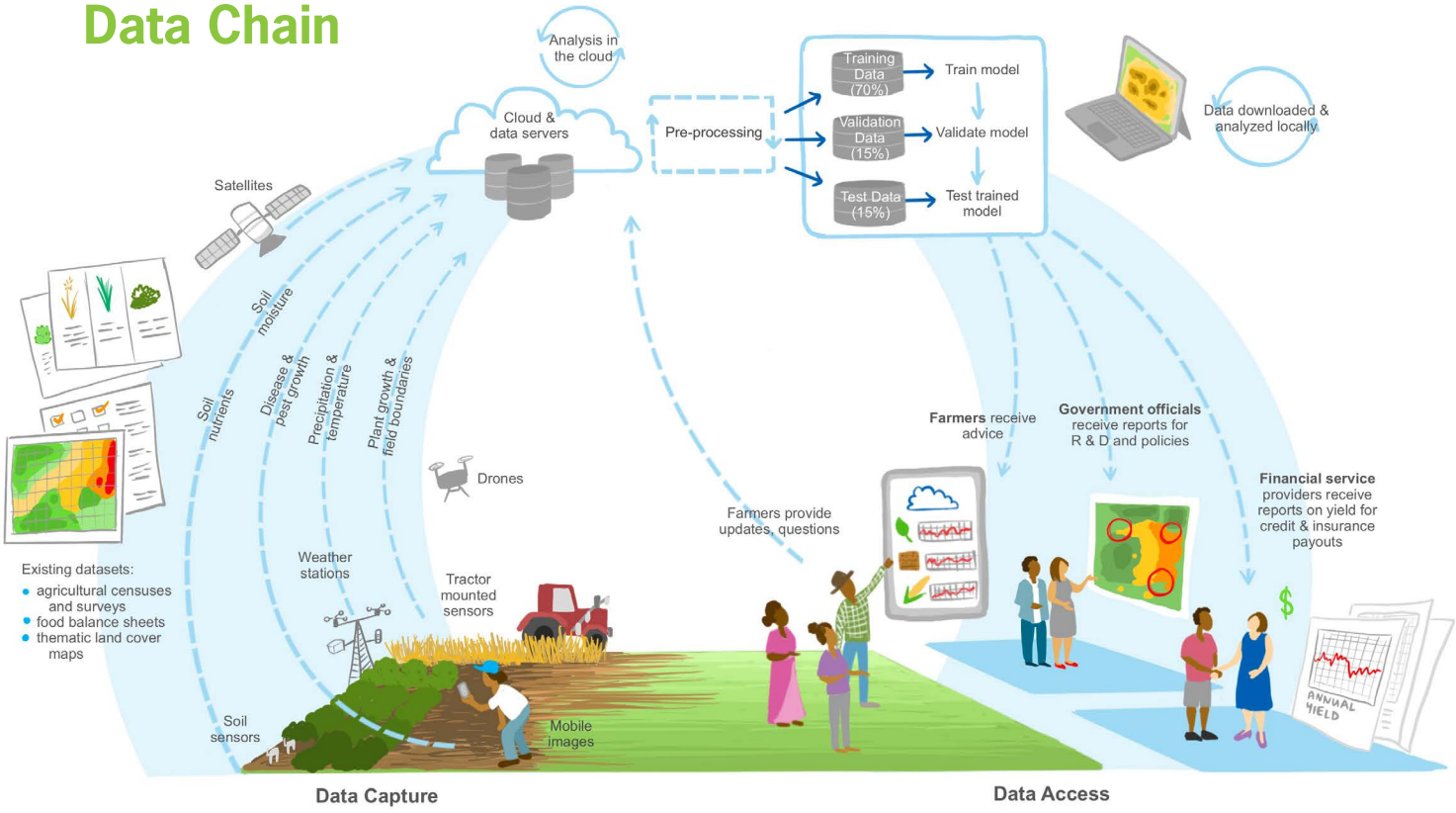
4
Model training, calibration, and validation

- Advancing understanding of machine learning modeling families and bridging silos to enable comparative assessment between models across different geographies or production systems
- Advancing research and application of unsupervised learning models for crop analytics
- Building of benchmark datasets, models, and comparisons
- Determining minimum model accuracy required for use cases
- Development of a black box optimized for model back-testing

Within each research area the analytics community has identified specific technologies, methodologies, or approaches which hold significant potential for catalyzing advances within this ecosystem.

These are detailed on the next pages.

Figure 1: Illustrative Crop Analytics Data Chain



1

Standardized data collection tools, quality standards, & alternative data sources

In order to catalyze a vibrant data ecosystem of ground truth training data for advanced crop analytics, cheaper, more effective, and scalable approaches to collect data and create data sets are needed. Innovation is needed to improve the apture of data, the standardization data collection, and the development of alternatives to field level data collection.

Collection Technologies

The paucity of high-quality georeferenced ground data that can train and calibrate machine learning/AI models for higher accuracy and generalizability is pronounced for smallholder agriculture in developing countries, where the scale, remoteness, complexity, and heterogeneity of farming practices intensify data and analytical requirements. Conducting traditional ground truthing campaigns with manual in-field collection by a trained enumerator is expensive and slow, and can introduce bias and error which limit dataset usability for ML applications. Organizations involved in conventional data capture use a disparate set of tools for the collection of ground data, leading to inconsistencies in the definition of data points and the formats that they are stored in, limiting the potential for dataset exchange and publication. Priorities include:

Testing and scaling cost-effective technologies for data capture through low-cost technologies for data capture including Internet of Things (IoT) and other sensors (e.g., LidAR), drones, and mobile phone image capture, etc. (*Box 3*).

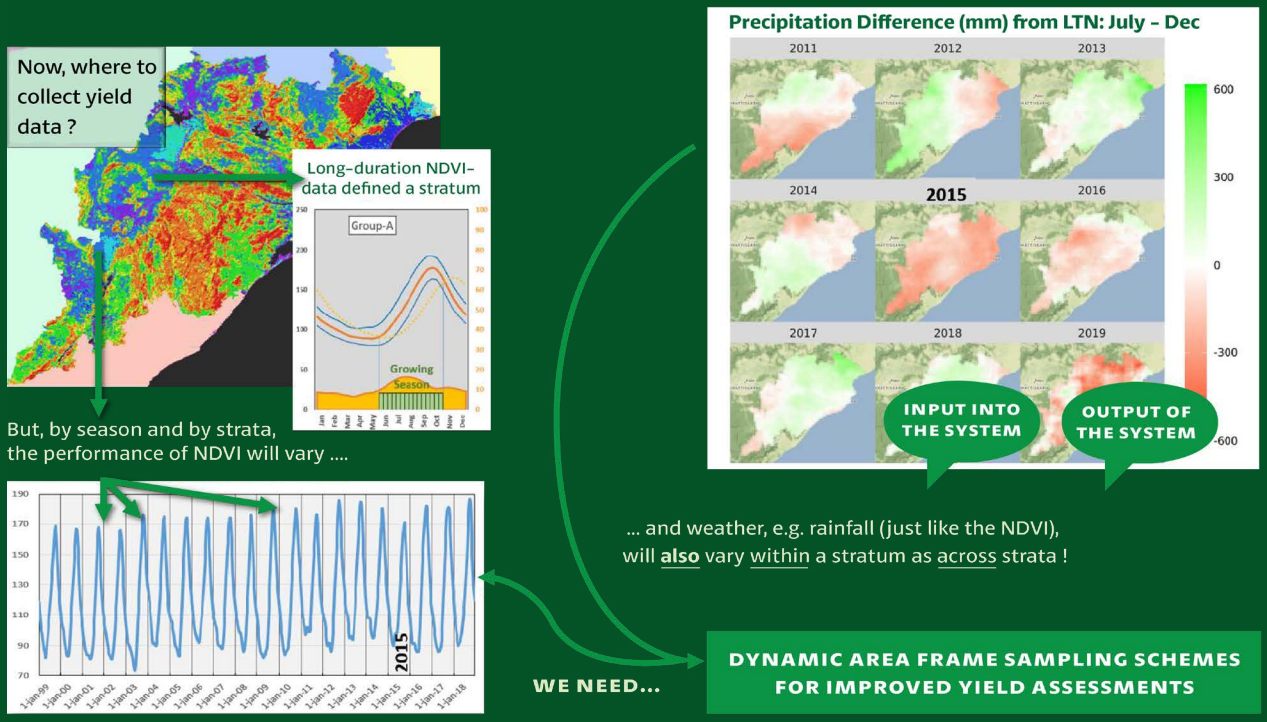
Digitizing data collection tools and systems to standardize and improve the value of traditional agricultural data capture in the public and private sectors. This includes the development of standardized digital toolkits for collection of data through mobile phones or tablets, or improving existing tools such as the Open Data Toolkit (ODK) (*Box 4*).

Usability testing and scaling of bi-directional data flows and user tasking/feedback loops for data collection applying intuitive feedback mechanisms. For example, developing a platform to collect near-real time feedback from a farmer or other end user about the ability to easily delineate or confirm field boundaries in a dataset, and adjust data collection parameters based on this input.

Testing new methods to pre-populate parameters at scale, e.g., where satellite imagery is used to automatically draw field boundaries allowing enumerators to confirm/deny data points.

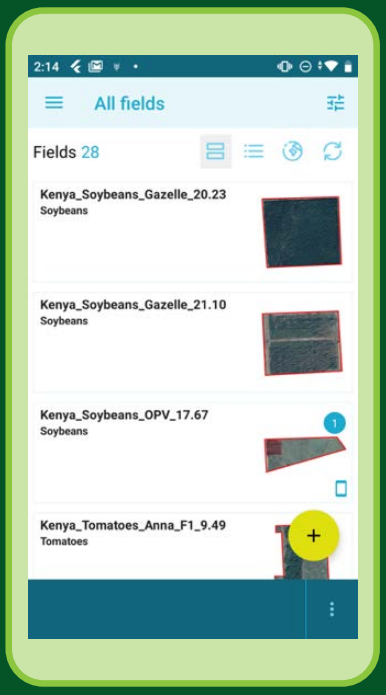
BOX 3:
Next-Generation Crop Production Analytics Using Smartphone 3D Imaging and Dynamic Area Sampling Frames

IFPRI, University of Twente, aWhere, and the Ethiopian Agricultural Transformation Agency leverage cutting-edge remote sensing and geo-statistical techniques to address the dual issues of inefficient ground-truth sampling design and inaccurate in-field crop yield measurement methods. Researchers will leverage spatially detailed weather information to cue field data collection that oversamples areas of high in-season variability to augment crop production zone maps derived from long-duration temporal NDVI profiles. Terrestrial LiDAR is used to collect gold-standard 3D information in-field. These data will be used along with photos of the crop to train a deep-learning model to estimate yield, which can then be bootstrapped for use in smartphones.



BOX 4:
Scalable Agriculture Data through Conversational Farm Record Keeping

6th Grain and farm.ink have developed a highly scalable ground data collection method through the combination of multiple proprietary mobile applications to gather high quality field data at a lower cost per datapoint than currently possible via household surveys. This will be achieved by integrating farm.ink's chatbot technologies within 6th Grain's FieldFocus mobile tool to train a large base of smartphone owning farmers to digitize their field boundaries, while also collecting crop type, varietal, farm management, and previous year production data. These data will be used to generate maize crop maps and, eventually, yield estimate maps. In exchange, farmers will receive a tailored farm record keeping and remote management platform.



Standards & Guidelines for Training Data Capture

Existing standards for agricultural data collection are technically diverse and adoption is inconsistent, leading to inefficient data collection, poor data quality, limited interoperability and increased effort required for data cleaning. Existing standards endorsed at the international level primarily cover traditional methods of data capture (e.g., crop cutting, survey enumeration) and not creation of ML-ready agricultural training data. Many datasets are not useful due to issues such as geolocations tagged to households rather than fields, or unclosed polygons created with active tracing on GPS units. Different types of crop analytic applications also have unique training data requirements such as spatial-temporal resolution, accuracy, sampling frames, or other parameters. Priorities include:

Determining the minimum required type and accuracy of data points and assessing the impact on model performance by training several models on a single dataset using different aspects of the dataset to determine the impact on classification.

Determining the minimum required quantity of training data points and assessing the impact of sample amount on model performance by training models with different sub-samples of a dataset to determine the improvement in performance achieved by collecting more data.

Developing new or promoting existing standards for collecting data for use in training ML models, taking into account specific needs required by specific use case applications.

Alternatives to Active Data Collection

Due to the time and cost-intensive nature of active data collection, alternatives such as use of synthetic datasets, ancillary datasets, or those collected in one geography and usable in another, are attractive areas for research. Innovation in this area would be especially useful as COVID-19 public health restrictions have limited the ability for active data capture campaigns. Even in normal conditions gathering ground data can be difficult due to lack of connectivity or access to fields, insecurity, data security regulations, etc. Priorities include:

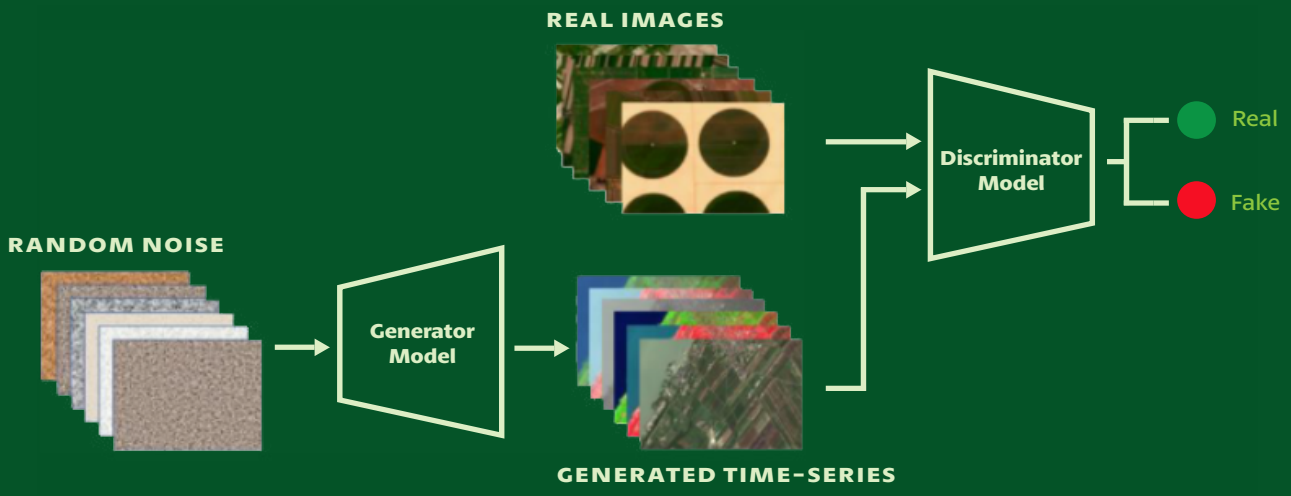
Testing methods for synthetic training data generation using techniques such as General Adversarial Networks (GAN; Box 5).

Testing of transfer learning methods to support analysis in geographically and agronomically similar but data poor areas using information from data rich areas.

Using existing non-traditional datasets which could be used to train models for crop analytics including non-agricultural datasets (such as nutrition or household survey data). Investments in this area should include a cost-benefit analysis to identify where data cleaning would be worth the investment vs. where existing data quality is such that active data collection is required.

BOX 5:
Generating Synthetic Agricultural Ground Reference Data from Satellite Observations

A research team led by Radiant Earth Foundation, Google AI Research Center, and Benson Hill have developed a novel approach to overcome model limitations due to lack of ground reference data using Generative Adversarial Networks (GANs). GANs have been used in medical applications, astronomy and to create photorealistic images, and hold promise for agricultural applications. The project will develop a framework to generate synthetic crop type labels to improve classification model performance. The framework will consist of two neural network models, a discriminator model applied to time series of multi-spectral satellite images, and a generator model using existing crop type and landcover data. The findings of this research will be used to assess the volume and geographic distribution of ground-truth data required to develop accurate crop classification models.



2

Methods & systems for semi-automated ingestion & processing

Once data are collected at the field level, they must be transferred from the point of collection to hosting platforms such as a desktop computer or to the cloud so that they can be used and shared. At present, this process is highly manual, non-standard, and inefficient. Hosting platforms currently perform significant data cleaning and quality control on data sets before posting them. Innovations in data ingestion and fusion as well as data processing would significantly improve the sharing of ground truth data.

Data Ingestion & Fusion

Ground-truth data currently reside in a variety of storage locations including local servers or other systems. While cloud storage is increasingly used, analysts, researchers, and others must often download datasets from several sources in time and computing intensive processes. Crop analytics actors universally agree that building pipelines for better dataset ingestion and cloud uploads, starting with publicly available data is a priority for innovation investment. In addition, in-situ sensors are important data sources for AI/ML application calibration and development. Analysts generally rely on multiple sensors or other sources of data in order to capture key variables but must compensate for the deficiencies associated with each capture mechanism. Sensor fusion techniques enhance ground truth accuracy and improve model performance. Centralization and improved interoperability can improve accessibility for non-experts or organizations without large storage capacity or compute power by eliminating the need to download and process a multitude of datasets locally. Priorities include:

Pipelines for automatically ingesting publicly available data using Extract, Transform, Load (ETL) and other technologies to enable the development of a centralized access points or hubs for public and private datasets, reducing the time, storage, and compute power currently required for dataset download.

Infrastructure such as the SpatioTemporal Asset Catalog (STAC) specifications or APIs, which enable dataset searchability in order to improve accessibility and prevent data siloes.

Techniques for multi-sensor fusion or similar functions to aggregate data from disparate sources in a single location in order to improve data accuracy and spatial and temporal resolution.

Data Processing

The Committee on Earth Observation Satellites estimates analysts invest 80% of time and effort in data processing and cleaning. Improved processing efficiency early within workflows would especially help those organizations without significant in-house analytics capabilities. It would also improve dataset quality and interoperability, and ultimately support further model testing and performance measurement. Many organizations are working on proprietary data processing methods and pipelines which automatically process data and contain embedded quality control, enhancement, and standardization. Identifying ways to expand use or efficiency of these models holds significant potential. Priorities include:

Building broad partnership models and infrastructure to automatically process data, enabling development of a set of Analysis Ready Data (ARD) which has been pre-processed and requires minimal user effort for further use.

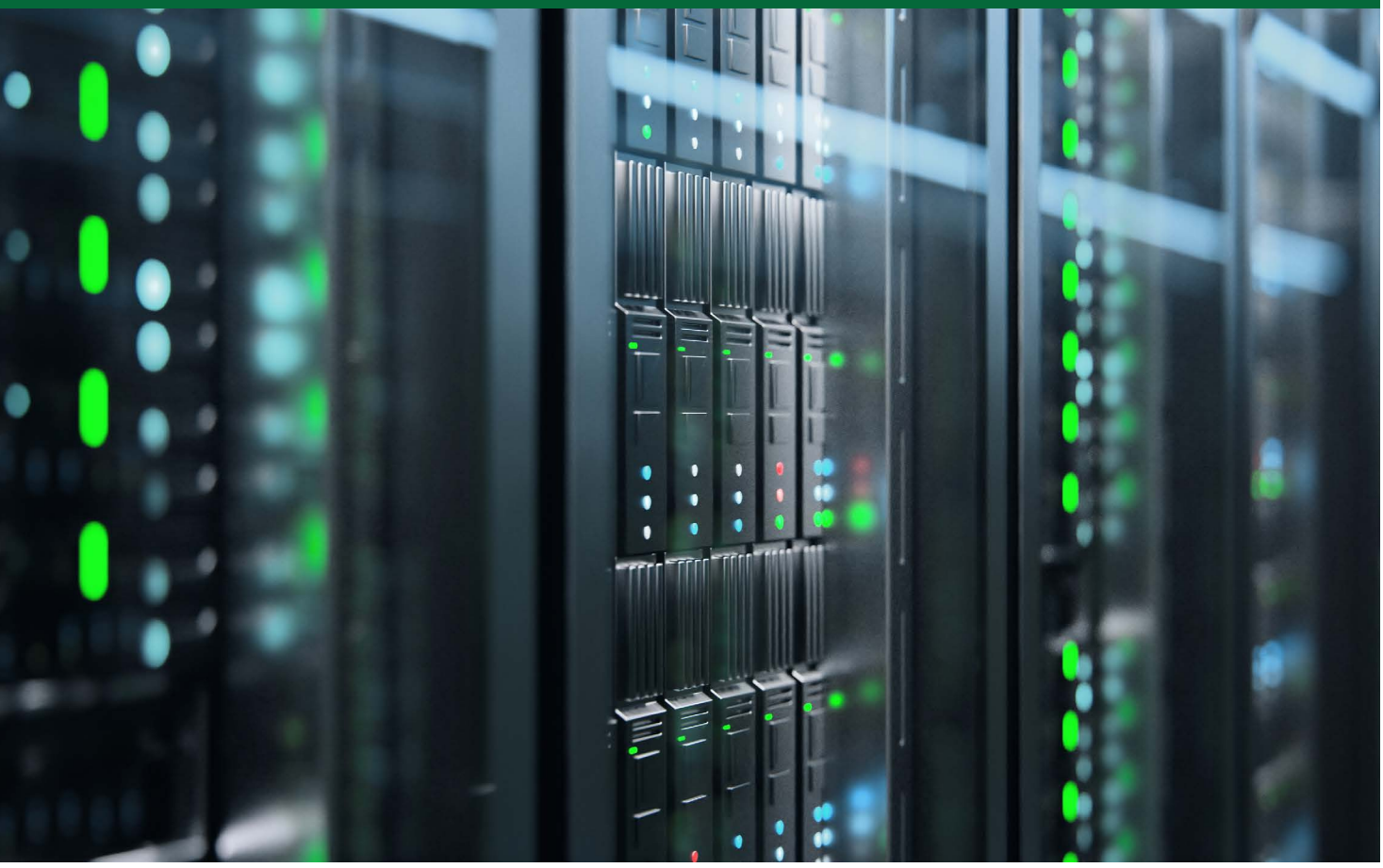
Testing new or disseminating existing methods for automated QA/QC to improve dataset accuracy and standardization while reducing the effort required to process data.

Testing new or disseminating existing methods for semi-automated data labeling to standardize the labeling process and reduce the need for expensive manual labelers.

3

Tools, standards, & methods for data storage, exchange, & addressing privacy

Ultimately, for the benefits of advanced crop analytics to be realized at scale, high quality ground truth data needs to be shared and exchanged while maintaining data privacy and meeting regulatory requirements. Innovation in the standards for data storage, data sharing platforms and marketplaces, and method for addressing ethical and privacy concerns are also needed.



Standards for Data Storage

The lack of clear standards for data storage (metadata, ontologies, semantics) is a consistent challenge faced in advanced crop analytics. There are few widely adopted standards and many organizations are not using external standards. Priorities include:

Testing established standards and guidelines for application across multiple use cases based on data storage requirements such as metadata, formats, etc.

Expanding adoption of existing standards through scaling mechanisms and partnerships, quantifying the speed and cost of various dissemination approaches.

Data Sharing Platforms & Marketplaces

The need to develop effective data sharing pipelines through repositories, effective APIs, or other exchange structures is a very high priority for the crop analytics community. Data sharing is hindered by organizations' inability to monetize proprietary data and derived products or publications. Within public and civil society sectors some hesitate to publish data which may be used for commercial purposes. Incentives for data sharing, such as access to processing, storage, and analytics services in exchange for data, "pay to play" access to non-public datasets, or payment for high-quality data can be tested to counteract these barriers. To identify and leverage these incentives innovative approaches are required which consider both the partnership and business models required to sustain these mechanisms and the technologies and data sharing agreements or structures which make them work. Priorities include:

Developing open source APIs to link data from disparate sources so that a user can search a geolocation and access all data (survey, imagery, field data) collected from that location.

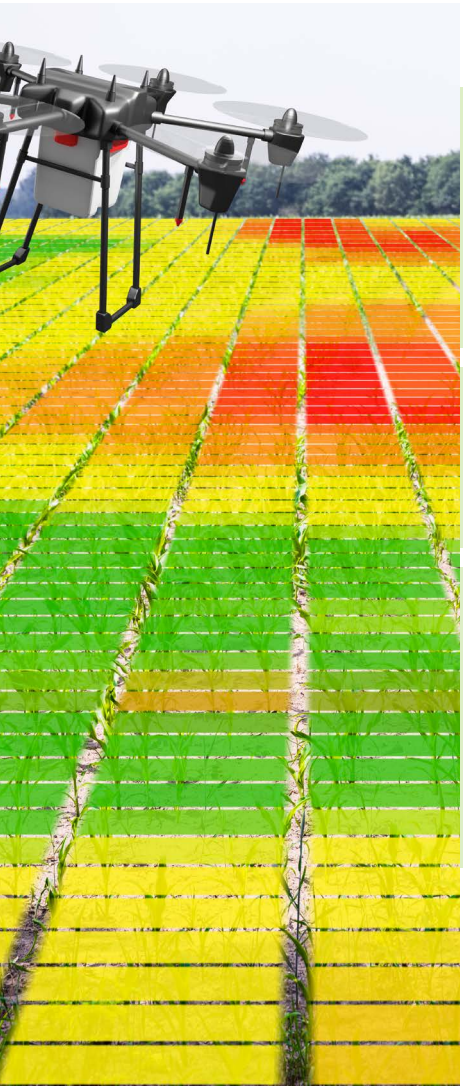
Developing and testing data marketplaces and exchanges for data sets using different pricing and exchange models that generate long term value for users including smallholder farmers contributing data. Testing both financial and non-financial incentives, such as access to tailored analytics and extension in exchange for data.

Encouraging data sharing through challenges/hackathons by partnering with organizations capturing large amounts of data with gaps in processing or analytics capabilities who are willing to offer data publicly and a small prize in exchange for crowdsourced labeling and model building.

Testing partnership and data sharing models which incentivize data sharing including creative licensing strategies, delayed publication of data to allow some level of competitive advantage, and partnerships to overcome first mover obstacles. One such example could be an agreement between multiple donor organizations to incentivize or require their implementing partners to collect ground data in accordance with ML-ready standards, thus beginning to populate a robust global dataset.

Privacy & Ethical Challenges

One of the largest barriers to achieving an open agricultural data ecosystem is the tension between the need for precise, georeferenced data to train accurate models and the regulatory, legal and ethical concerns associated with the use and sharing of data with personally identifiable information (PII). This barrier hinders the flow of datasets from organizations with open data mandates and geographically diverse operations. As these data hold significant promise for advancing crop analytics and associated smallholder services, it is crucial to identify pathways for addressing these concerns and develop a set of guidelines around the sharing of data which protects and benefits smallholder farmers. Farmer ownership of data is a critical issue to consider. Financial and non-financial incentives for sharing of data by the farmer need to be considered. Priorities include:



Assessing the impact of regulations, laws, and emerging guidelines for data privacy as they pertain to data capture and storage within applications supporting priority use cases. Quantifying gains in accuracy, deployment at scale, or other areas which could be used to influence policy or encourage data sharing arrangements. Policy toolkits that support national authorities and others to improve policy and regulation to better support agricultural data sharing and use are also needed.

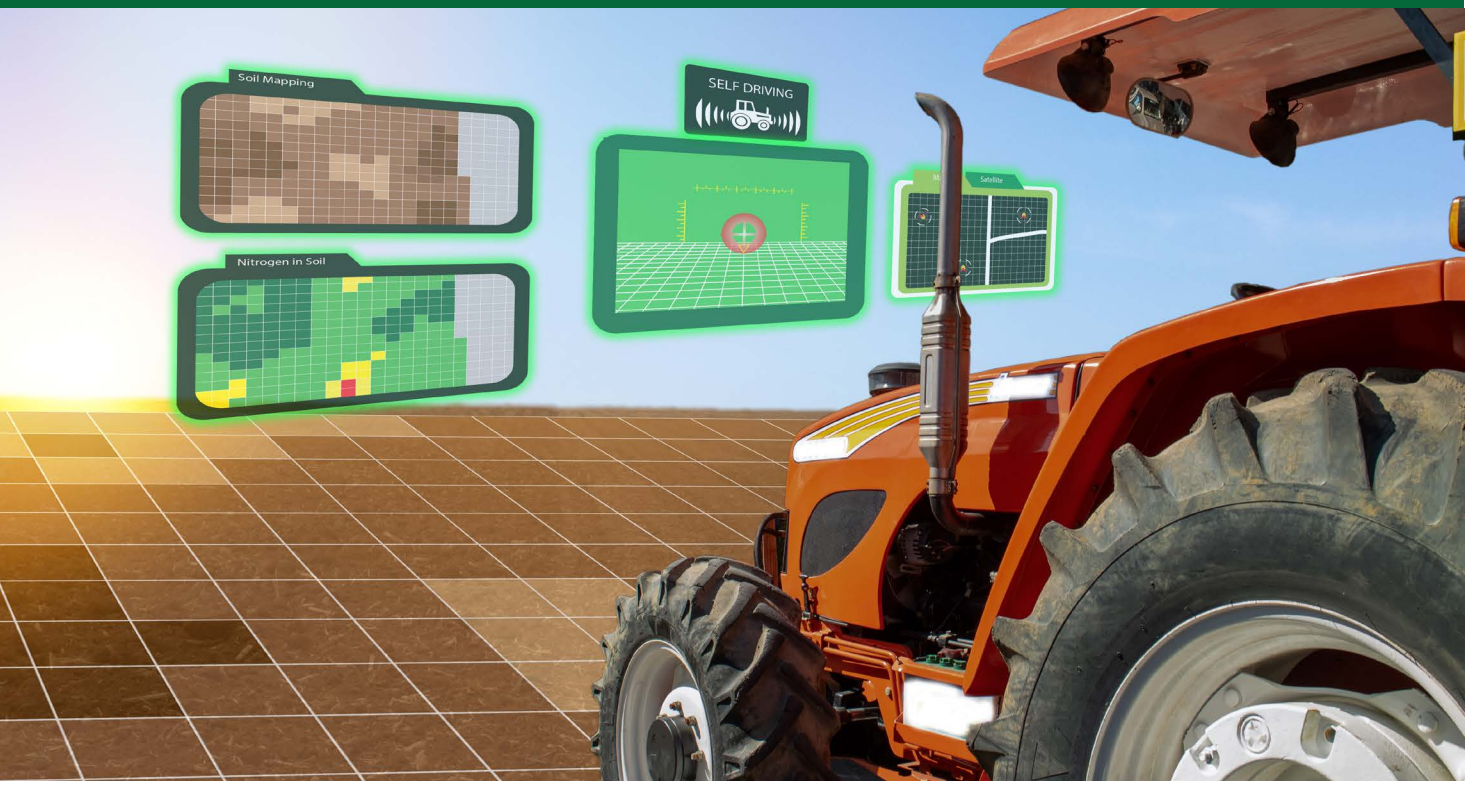
Testing new data sharing arrangements, potentially across geographies, developing and disseminating best practices from one location or user group to another. Developing guidance or materials to increase understanding of regulations and develop decision support tools to better share sensitive data.

Testing various methods for anonymizing data and implications on model accuracy. Research is needed to test whether methods such as geographic fuzzing are robust enough to avoid re-identification of anonymized datasets through other analytic means. In addition, the impact of using anonymized data on model accuracy across different production contexts needs to be better understood. Finally, black box testing could be used to address issues of data privacy, allowing for model testing without requiring access to potentially sensitive input data, as demonstrated in *Box 6* below.

4

Model selection, training, and validation

Ground truth data are collected in order to train and validate models to generate useful information around a number of parameters including estimating and predicting crop yield, assessing crop health and growth, and mapping field boundaries. This information is used to provide a range of services to actors throughout the supply chain, each of which has specific needs for accuracy, spatial resolution, periodicity, etc. Research and development in machine learning in agriculture is highly siloed and has not allowed for direct comparison of models or the development of benchmark models and datasets. Innovation is needed in model selection, exchange, calibration, and validation.



Model Selection & Exchange

Training and reference data requirements are highly specific to the algorithm and application they support. Several advancements in machine learning approaches have emerged as an area of interest for further investment due to their potential to greatly decrease the amount of training data needed to train traditional models. These include unsupervised learning methods and deep learning methods because of their ability to run without labeled training data. However, these models are generally opaquer and are more difficult to replicate unless publicly hosted. Priorities include:

BOX 6: Black Box Testing

Black box testing allows for testing of a model's functionality without knowledge of its internal structure. The tester knows what the application is supposed to do but has no knowledge of the code or input data.

Test cases can be derived from a simple description of the application's specifications and requirements, allowing for model testing without sharing potentially proprietary information (e.g. model code or structure) or sensitive data (e.g. georeferenced PII).

Source: Ehmer, Mohd, and Farmeena Khan. "A Comparative Study of White Box, Black Box and Grey Box Testing Techniques." International Journal of Advanced Computer Science and Applications, vol. 3, no. 6, 2012, doi:10.14569/ijacsa.2012.030603.

Advancing research and application of unsupervised learning methods with an emphasis on field boundary classification, crop type, and crop yield estimation and prediction.

Developing pipelines for sharing open models and code to allow for more rapid improvements in models and to enable direct comparison of models.

Model Calibration & Validation

Advancing domain-specific validation metrics is a key area for innovation identified by the crop analytics community. The emergence of easy-to-use analytics tools have resulted in a proliferation of crop models with variable levels of quality and accuracy. The lack of standards for validation is both a barrier to adoption of crop analytics tools and creates risks for users who take decisions based on poor information. More rapid and trusted use-case specific validation is necessary to ensure quality and build trust in derived products. Tools to support model validation, creative benchmarking as a service business models, and the pipelines for sharing open models discussed above need to be developed to enable standardized model validation. Priorities include:

Establishing benchmark data sets and common benchmarking metrics for specific crop analytics use cases that meet agreed up standards developed with input from analysts and users.

Assessing required model accuracy and determining the minimum level of precision in datasets required for model application under different use cases.

Developing black-box tools optimized for model back-testing to validate model performance helping to demonstrate accuracy and usability to stakeholders (see *Box 6*).

Enabling Crop
Analytics At Scale

Unlocking the Potential of Satellite-based Data and Analytics for Smallholder Farmers

The Enabling Satellite-based Crop Analytics at Scale (ECAAS) Initiative is a multi-phase project that aims to catalyze the development, availability, and uptake of agricultural remote-sensing data and subsequent applications in smallholder farming systems. The initiative is funded by The Bill & Melinda Gates Foundation and implemented by Tetra Tech.

info.ecaas@tetratech.com
cropanalytics.net