Agriculture Big Data (AgBD) Challenges and Opportunities From Farm To Table: A Midwest Big Data Hub Community† Whitepaper

Shashi Shekhar[1], Patrick Schnable[2], David LeBauer[3], Katherine Baylis[4] and Kim VanderWaal[5]

[1] Dept. of Computer Science & Engineering, University of Minnesota, Twin Cities
[2] Dept. of Agronomy, Dept. of Genetics, Development and Cell Biology, Iowa State University
[3] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign
[4] Dept. of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign
[5] Dept. of Veterinary Population Medicine, University of Minnesota, Twin Cities

Abstract: Big data is critical to help agriculture meet the challenges of growing world population, climate change and urbanization. Recent success stories include precision agriculture, phenotyping, and global agricultural monitoring. Many of these initiatives are made possible by novel data sources such as satellite imagery, instrumented tractors and initiatives such as the Global Open Data for Agriculture and Nutrition (GODAN). This whitepaper surveys agricultural big datasets, characterizes their limitations, lists transformative opportunities and suggests a plan to engage and nurture Agriculture Big Data (AgBD) research community. Public big data includes satellite imagery (e.g., Earth on Amazon Web Services, Google Earth Engine), surveys (e.g., National Agricultural Statistics Service), financial statistics (e.g., Economic Research Service), social media (e.g., Twitter), etc. Private datasets describe yield (e.g., precision agriculture, Farm Service Agency), farm loss (e.g., Risk Management Agency) and condemnation (Food Safety and Inspection Service), etc. Limitations include data and metadata gaps, insufficient data storage, preservation, and documentation, lack of scalable spatiotemporal big data analytics methods, and inadequate secure data-sharing mechanisms. Transformative opportunities include workforce development, Cyber-Infrastructure (e.g., long-term, curated data repository services), data norms and sharing models, metadata, big data aided mechanistic models, spatiotemporal big data analytics for data-driven hypothesis generation and testing, etc. These transformative opportunities cannot be realized without federal leadership. To make progress towards the transformative opportunities, the whitepaper also lists resources to engage researchers from agriculture and big data in collaborative efforts with federal support.

1. Big Data

In general, big data is defined by the 3Vs: volume, velocity, and variety. Volume refers to the exponential growth in the amount of data collected (e.g., high-resolution and high-frequency of satellite and aerial imagery). Velocity refers to the speed of data collection (e.g., real-time in-field cameras for monitoring plant growth). Variety refers to the large number of data sources and formats (e.g., traditional survey data vs. social media posts about food and food-borne illnesses).

In the agriculture community, big data is often viewed as a combination of technology and analytics that can collect and compile novel data, and process data in a more useful and timely way to assist decision-making (Stubbs, Big Data in U.S. Agriculture, Congressional Research Service, 2016). This view simultaneously considers both the data and the processing methods used to extract value from the data.

†This whitepaper paper was co-authored by a few volunteers following an April 2017 Midwest Big Data Hub workshop [57] titled "Machine Learning: Farm-to-Table," which brought together academic researchers from Agriculture, Food-Energy-Water, Engineering, Computer Science, and Agricultural Economics, with representatives from startups, small & medium size enterprises (SMEs), and large corporations, to discuss current applications of Machine Learning, big data and other computational approaches to understanding Ag-food-energy-water systems.

2. The Potential of Big Data for Agriculture

With the global population projected to exceed 9 billion by 2050 [2], it will be critical to optimize agricultural production and food supply chains to more efficiently produce and deliver food, fiber and fuel to meet growing demand [3] [4]. This goal is further complicated by climate change and urbanization. Agricultural Big Data (AgBD) will be an essential component of the second green revolution that will be required to meet these needs.

AgBD sets are already used by many countries and commodity markets for the early detection of disruptions in supply chains for commodity crops such as wheat, rice, corn, and soybean [5] [6] [7] [8] [9]. Precision agriculture has developed with advances in remote sensing data collection, including improved spatial and temporal resolution, spectral resolution, variety of sensor platforms (e.g., satellite, aerial, ground-based), etc. [10]. A recent congressional reception also reported that precision agriculture has shown promise in increasing on-farm yields [11]. In addition, a recent Fortune magzine [12] quoted the potential of increasing farm profits by almost $100 per acre via prescriptive farming that uses predictive modeling and AgBD to optimize farm management practices ranging from customized seed planting density to fertilizer application based on local soil characteristics and long-range weather forecasts. In animal agriculture, AgBD and predictive modeling are critical for surveillance and control of infectious diseases.

Beyond agricultural production, GPS-enabled sensors are being used to track food and generate AgBD of supply chains. Such technologies are estimated to help reduce food-borne illnesses by 76 million in the US every year [13]. AgBD can also be used to improve supply chain security. For example, spatial data mining techniques (e.g., hotspot detection) [14] [15] [16] [17] [63] [64] [65] can be used with AgBD to identify crops (e.g., California almonds [18], Cocoa [19]) produced in small geographic regions or a set of regions that are vulnerable to climate change and natural disasters. Their supply chain maps can then predict geographic chokepoints of these sensitive crops and animal-based commodities, informing industry and consumers of risks before they hit. Similarly, spatial data mining may also help select sustainable sources (e.g., avoid deforestation based palm oil) in a supply-chain [20]. In addition, detailed data on consumer and market behavior can be used to improve food access and nutritional outcomes, and geo-social media can be leveraged for timely detection of food contamination events and control related illnesses.

We envision that AgBD will assist decision-making in agriculture at four levels: [21] [22]:

Descriptive: For precision agriculture and high throughput phenotyping applications, the aim of AgBD collection is to characterize spatial and temporal variability in soil, land cover, crop and weather characteristics and identify stressors, traits, or infectious disease risk factors that need better management.

Prescriptive: Using data collected and associated maps of individual characteristics, traits, or exposures to infectious agents, a prescriptive analysis is conducted to determine necessary farm management interventions.

Predictive: A predictive analysis using historic datasets as well as integrated soil, crop, weather and market models may forecast outcomes such as crop yields and food insecurity. Predictive analytics can also be used to improve decision making to forecast spread and limit the impact of infectious agents on crops and livestock.

Proactive: A proactive level involves observations of crop development and stress on multiple farms over large regions and time scales. AgBD from these observations are pooled and mined to obtain relationships between site characteristics, weather and crop performance under a range of management conditions. These relationships can be used to customize management practices and seed selection to local conditions.

3. Current Agricultural Big Data

Public Agricultural Big Data: Various types of AgBD have been made publicly available by a number of providers as shown in Table 1. The U.S. Department of Agriculture remains the major provider of services for managing and sharing most types of agricultural data (e.g., survey, financial, scientific, etc.). In 2013, in its response to the G8 International Conference on Open Data for Agriculture [23], USDA also launched the

GODAN (Global Open Data for Agriculture and Nutrition) Initiative [24] to support the sharing of open data to help ensure world food security.

Besides USDA, the U.S. Bureau of Labor Statistics, the National Oceanic and Atmospheric Administration (NOAA), and the National Aeronautics and Space Administration (NASA) also continue to provide a huge volume of satellite imagery with increasingly high resolution and high frequency. Many open imagery datasets are also accessible through platforms such as Amazon Web Services S3 (AWS), Google Earth Engine and NASA Earth Exchange. Imagery datasets are critical AgBD resources in many agricultural applications, including precision agriculture and yield prediction. In addition, with the growth of popularity of online social media, more and more users are sharing agriculture-related information (e.g., food consumption, food-borne illness). For many social media platforms (e.g., Twitter, Google Search), this information can be retrieved and downloaded for analysis.

Table 1. Examples of Public Agricultural Big Data (*Adapted from* [1])

Type of Public AgBD	Provider
Satellite imagery and meteorological information	**Cloud-computer based** (e.g., Earth on Amazon Web Services [28], Google Earth Engine [29], and NASA Earth Exchange [30]) and others (e.g., National Oceanic and Atmospheric Administration (NOAA) [25], National Aeronautics and Space Administration (NASA) [26] and U.S. Bureau of Labor Statistics [27])
Survey data	National Agricultural Statistics Service (NASS)* [31]
Financial data	Economic Research Service (ERS)* [32] National Water Economy Database (NWED) [33]
Scientific data	Agricultural Research Service (ARS - U)* [34]
Soil, water, and geospatial data	Natural Resources Conservation Service (NRCS)* [35]
Price and sales data	Agricultural Marketing Service (AMS)* [36]
Commodity and market data	World Agricultural Outlook Board (WAOB)* [37]
Generic data	Global Open Data for Agriculture and Nutrition (GODAN)* [24] VegScape* [38]
Animal disease incidence data	World Animal Health Information System (WAHIS) [39] EMPRES Global Animal Disease Information System (Empres-i2) [40]
Citizen data	Social media platforms (e.g., Twitter posts about food)

* Service provided by U.S. Department of Agriculture (USDA)

Private Agricultural Big Data: One source of private data is generated through administrative processes in government agencies (e.g., Farm Service Agency) as shown in Table 2. These data may contain individual level or other private information and are not publicly available. Other key types of privately held data are collected by agricultural companies, financial institutions and individual farmers [1]. For example, sub-field level and plant-level data have been collected using drones, field-sensors and cameras. Private AgBD have benefited many groups of users, including farmers (e.g., improved production), ranchers, retailers, industry groups and environmental scientists (e.g., reduced usage of fertilizer or antibiotics). The private sector has also seen the introduction of a variety of new technologies. For example, positioning techniques are used on dairy farms to track the movements of herds. However, due to privacy or business concerns (e.g., competing in commodity markets), these data are usually kept for internal use only.

3.1. Limitations of Current Agricultural Big Data

Limited data storage and preservation: The increasing volume, variety and velocity of agricultural big data (AgBD) sets demands excessive computing power and computational resources to manage and analyze. High-resolution (e.g., sub-meter) datasets can be collected at high temporal resolution (e.g., daily or more frequent) via ground-based sensors, low-flying UAVs and remote sensing satellites. It has become increasingly difficult to store and

maintain AgBD without significant investment in big data platforms such as high-performance computing. More importantly, some types of data need to be preserved for long-term use and analysis, which requires a highly centralized and reliable platform managed by dedicated administrators. The agriculture research community is facing the dilemma that valuable AgBD is being collected at an unprecedented pace and scale, but the effort, expense and time required for data management discourage the efficient sharing and reuse of these datasets.

Table 2. Examples of Private Agricultural Big Data (*Adapted from* [1])

Type of Private AgBD	Owner
Yield and loss data	Risk Management Agency (RMA)* [41]
Farm record data of individual producers, federal payments, and loan information	Farm Service Agency (FSA)* [42] Farmers
Conservation plans, geospatial data, and conservation program activities and payments	Natural Resources Conservation Service (NRCS)* [35]
Generic research data	Agricultural companies (e.g., equipment manufacturers, chemical companies) Farmers and research groups (e.g., at universities)
Animal movement data	Farmers and companies
Herd production and health data	Industry groups; Information-sharing services; Farmers
Diagnostic animal health data	Universities; Commercial laboratories
Carcass condemnation data	Food Safety and Inspection Service (FSIS)* [43]
Genetic, metagenetic, microbiome data	Agricultural companies; Research groups
Citizen data (processed)	Processed social media data at university research labs (e.g., Food Protection and Defense Institute at the University of Minnesota [44])

* Service provided by U.S. Department of Agriculture (USDA)

Data sharing barriers: Data privacy is a major concern of AgBD sharing, and private AgBD owners may be reluctant to share the data. Even if private data (e.g., administrative data; field treatment data) is shared, it is not certain that one private dataset will be compatible with another private dataset or with other public datasets. For example, sampling by public agencies is usually conducted following statistical standards [45], whereas the same is not necessarily true of private data collection. Security is another concern: Many private agriculture datasets are stored locally rather than on cloud computing platforms due to security concerns, and these concerns need to be addressed before many data owners will be willing to share their data [1].

Insufficient data documentation: To support big data analytical methods in agriculture, such as data mining, it is increasingly common for satellite imagery to be supplemented with more and more field data. Although original owners may be fully aware of their data's properties (e.g., coordinate system, error range), it is often difficult for others to reuse the data or combine it with other data due to a lack of this sort of documentation. Incomplete metadata significantly limits the value of the laborious and time-consuming data collection tasks in many of the existing research projects.

Lack of connection between observation and theory: Empirical models (e.g., machine learning outputs) generated by AgBD can be difficult to generalize outside of their target geography or time frame. Current empirical crop models [5] [46] cannot be used to extrapolate beyond training data, making them inappropriate for non-stationarity such as with new climate and weather patterns, or new or improved crops [47]. Beyond empirical models trained solely based on observations, mechanistic models need to be developed to make simulations based on theory to address the non-stationarity in AgBD prediction problems.

Missing crucial data: While novel data collection and analytical methods have generated valuable insights for agricultural production, data availability remains sparse in domains related to natural resources (e.g., water use),

as well as other phases in the food life cycle (e.g., food movements, food consumption, household outcomes, etc.) as shown in Figure 1. For example, data on water use for agriculture are not collected for many locations in the U.S. Big data analytical techniques provide the ability to predict such outcomes using alternate data sources (e.g., remote sensing imagery, SMS-based surveys). These models, however, require substantial data for validation. For example, predicting food security crises would greatly improve the global ability to respond with food and funds to limit harm, but little data exist on global household food insecurity at the spatial and temporal scale needed for validation.

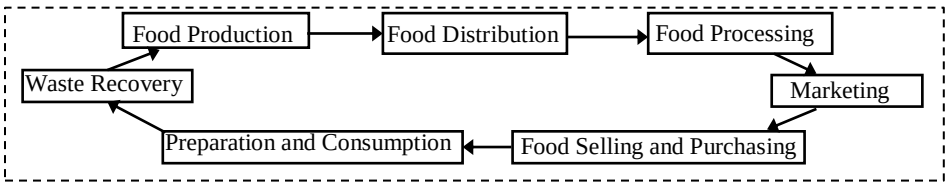

Figure 1. Food life cycle

4. The Case for Federal Actions on Agricultural Big Data

The case for federal actions are summarized in Table 3, and several of these are detailed in the subsections.

4.1. Workforce development

Specific coursework must be designed to develop an agricultural research workforce with big data skills to continue to grow the AgBD economy and improve food production. The courses will train scientists and others interested in data-driven agriculture (who lack a thorough computer science background) to understand, adapt and correctly apply big data approaches (e.g., machine learning, spatiotemporal data mining, high-performance computing). These courses could reach beyond the academy with modules designed with and for people in agribusiness and government. Federal (e.g., NIFA) funding can make a difference to help universities and other educational organizations develop such courses and programs and prepare skills for AgBD economy.

4.2. Cyber-Infrastructure: The rationale for centralized, long-term, curated data repository services

Success stories of centralized curated storage center in agriculture include seed-banks (e.g., USDA Plant introduction station, and Fort Collins, CO). Given the value of curated, long-term, geographically-diverse emerging big data (from precision agriculture and beyond), to support longitudinal studies (e.g., gene-environment interaction, impact of climate change), there is a ground-breaking opportunity to create NIFA-supported data centers that provide long-term storage (e.g., 10 years, permanent), and curation for sharing, discovery and dissemination.

We envision a centralized data center, where researchers and others can share their own AgBD and access others' AgBD to facilitate big data analytics without worrying about data storage, persistency, security and privacy. A critical step to realize this is to first build a catalog of current data repositories (e.g., data.gov portal) of value for modeling. Identify existing data formats, and facilitate community engagement around developing standards for data collection and documentation. The path towards long term, accessible, and harmonized (i.e. curated) data will begin with the development of APIs and associated software that can transform from current data structures into shared data models. With translation tools and APIs, data creators and maintainers can work together as a community to put similar information into shared formats.

4.3. Data norms and sharing models

We envision a set of NIFA-approved norms that define and implement conventions for agriculture big data (AgBD) research. The norms would cover: (1) experimental design and sampling methods to use for data collection in different scenarios; (2) conditions that must be considered in controlled experiments (case by case) and (3) minimum acceptable spatial and temporal precision and resolution of data (e.g., field sensors, UAVs).

The NIFA-AgBD norms can have multiple levels (similar to database norm) to allow flexibility in data selection for different research. These best practices also open up opportunities to develop tools to automate them in research (e.g., R package that uses power analysis to estimate required sample size). A recent work also suggested the use of standards across agricultural Information and Communications Technology (ICT) to facilitate the sharing and reuse of open agricultural data and models [48].

We also envision a NIFA sharing model that implements privacy protection rules (e.g., database access control) and techniques (e.g., data aggregation) to protect sensitive personal information and facilitate private data sharing. Sharing models can also be created for special scenarios. For example, in public safety, Enhanced-911 allows sharing of private information (e.g., location) during an emergency. Similarly, in agriculture, sharing private data in case of emergency (e.g., disease outbreak in cattle farms) may help control the situation and reduce loss to farmers.

Table 3. Agricultural Big Data Opportunities and Research Needs

Areas	AgBD Opportunities and Research Needs
Workforce Development	Training for farmers and AgBD companies with coursework on big data methods at land-grant universities and beyond in collaboration with department of education.
CyberInfrastructure	NIFA-supported storage‡ for valuable AgBD sets along the lines of NIFA support for genome databases (e.g., MaizeGDB) and seed banks (e.g., the USDA Plant Introduction System). Improve rural broadband infrastructure to support AgBD collection in rural areas.
Private data sharing and compilation	Models for sharing private AgBD. For example, administrative data may provide behavioral and societal information that are not well studied. Standards for sharing of private AgBD (e.g., data format, statistical guidance). Methods for compiling public and private AgBD.
Novel Data Collection	New data collection methods for model validation, combined with funding the development and testing of algorithms to fill in data gaps using predictions from existing information (e.g., remotely sensed data, market data). Public data on food movement and food consumption. New approaches to improve data transfer capacity between farms and data center (e.g., use of TV white space or other less frequently used channels).
Spatiotemporal Machine Learning	Leveraging of new high-resolution (e.g., daily, 1 meter) satellite data to monitor crops on a large scale. Spatiotemporal hotspot detection to identify risks in supply-chain (e.g., heavily localized plants that are subject to climate change). Spatial optimization for land-use and land-cover allocation, and identification of potential production improvements through changes in management. Disease risk forecasting for livestock, including environmental, epidemiological, and weather-related data.
Mechanistic Models	Combining empirical models and mechanistic models to link observation with theory.
Citizen Engagement	Social Media, Apps, Easy-to-use Decision Support for growers and ranchers. Downstream behavioral change through apps (e.g., reduce food waste). Cognitive and behavioral science applied to enhance feedback for technology improvement, scientific advancement and innovation.
Data analytics in animal agriculture	Application of data science methods to detect aberrations in AgBD data streams, which may indicate changing or emerging threats to animal health. New approaches to optimize animal health and production through linking processes occurring at multiple spatial and temporal scales. Development of data pipelines to promote analysis of data in near real-time.

‡ Other federal agencies have faced similar challenges and included long-term data storage and preservation as part of their core missions (e.g., cancer registries, Census Bureau, NASA Earth Explorer).

4.4. Metadata

Enriching AgBD sets with detailed metadata is important to enhance their value beyond their original purpose. Metadata contains necessary details describing a dataset, such as time, location and explanations of properties. Creating metadata and documentation for existing AgBD is a laborious task. NIFA funding could provide an important opportunity to complete the metadata and help repurpose shared datasets for new scientific research, thereby facilitating collaboration and opening many new avenues for scientific discovery. The increased amount of re-usable data will also provide more training data for computational approaches (e.g. supervised learning) to improve the accuracy and robustness of big data analytics for predicting agricultural outcomes.

4.5. Novel data collection and compilation

Where data do not exist, NIFA could support novel data collection and compilation through funding calls, and by facilitating interactions with other government agencies engaged in data provision. NIFA might target data for model development and validation, combined with funding the development and testing of algorithms to fill in data gaps using predictions from existing information (e.g., remotely sensed data, market data). Strategically targeting methods to fill data gaps, combined with systematically-collected sub-samples of field data could dramatically improve researchers' abilities to develop and test new models of environmental and socio-economic outcomes of agricultural processes.

4.6. Spatiotemporal big data analytics for data-driven hypothesis generation and testing

AgBD analytics provides great opportunities to generate new hypotheses from large datasets that are otherwise tedious and time-consuming for human researchers to inspect and identify. For example, in food supply chain analysis, it is important to identify high-risk nodes whose failure has severe implications (e.g., type of crops that are grown only in a few small geographic regions or that rely on specific ports threatened by sea-level rise). Analytic methods can also assist hypothesis testing in longitudinal studies that require coordinating experiments across multiple sites. For example, testing a relationship between plant growth rate and plant geometry must be done in field trials across multiple seasons and locations. AgBD analytics can assist in controlled experiment design (e.g., spatial optimization [49] [50]) to improve the efficiency of field trials.

However, agricultural data exhibit spatiotemporal auto-correlation and non-stationarity [63] [64] [65], which violate common assumptions (e.g., independence and identical distribution of learning samples) underlying common machine learning and big data analytics methods. In addition, current spatial statistical methods (e.g., spatial auto-regression) do not scale up to big datasets due to their computational complexity. Thus, there is a great need to develop computationally scalable methods to analyze spatiotemporal datasets in agriculture. NIFA's funding is important to support researchers to develop scalable spatiotemporal data analytics methods and explore AgBD-based hypothesis generation and experiment design.

4.7. Agriculture big data aided mechanistic models

Mechanistic models enable the use of scientific understanding to harmonize datasets from diverse experimental designs and scales by directly simulating the system states under the experimental and environmental conditions under which the data were generated. In the context of theory, it is possible to better interpret and utilize data that are available to identify knowledge gaps. Similarly, using mechanistic models to simulate dynamics at multiple scales simultaneously allows coherent evaluation and reanalysis of data collected at different scales and temporal frequencies [51]. This approach has been used for decades in the geosciences including weather forecasting and the development of historical climate data products for decades, and more recently has been applied to agricultural sciences.

Linking predictions to mechanistic behavioral models holds the potential to better evaluate how human behavior may affect agricultural outcomes. Agricultural production outside of test plots result from a combination of agroecological characteristics, weather and human behavior. Forecasting models may be unable

to capture the human responses to changing climates or market conditions, and require a coupling with behavioral models to capture outcomes. Further, to understand these outcomes and identify how best to improve agricultural productivity requires insights into causal mechanisms underlying these processes.

4.8. Data analytics in animal agriculture

While crop-based agriculture and human medicine have harnessed big data to optimize "precision" approaches to improve production and health outcomes [52], AgBD in animal agriculture has been mostly focused on spatial analyses and bioinformatics [53] [54]. However, the use of AgBD for animal disease surveillance is a small but rapidly growing field, with applications ranging from targeting specific populations to tracking or even anticipating spatial and temporal trends. The development and refinement of such capabilities in animal agriculture could significantly improve our ability to identify and respond to emerging animal health concerns, especially if collection and analysis of data occurs in near real-time rather than retrospectively [55].

Animal agriculture data that are or are becoming "big" include "-omics" data, geospatial data, publicly available data repositories such as World Animal Health Information System and EMPRES Global Animal Disease Information System (Empres-i2), clinical and diagnostic data for food animal diseases, and data on animal movement from local to international scales [56]. In addition, data associated with production constraints in food animal industries (such as infectious disease, mastitis, nutrition, physiological metrics, etc.) are often housed in industry-based data warehouses that have the participation of large proportions of the industry. The analysis of such data can be used to understand health risks and minimize the impact of adverse animal health issues by, for example, increasing the effectiveness of control and surveillance by identifying high-risk populations through the analysis of spatial and animal movement data; combining disparate data or processes acting at multiple scales through mechanistic modeling approaches; and harnessing high velocity data to monitor animal health trends and for early detection of emerging health threats [55].

5. Engaging Big Data Research Community in Agriculture

Future innovations in AgBD research will require engaging multi-sector communities across the nation. The Big Data Regional Innovation Hubs (BD Hubs) were launched by the National Science Foundation (NSF) to strengthen the data ecosystem, and develop effective academic-industry-government-NGO networks to address scientific and social issues of regional and national interest. The BD Hubs cultivate communities and resource networks, and build collaborations that reduce barriers to data sharing and access, and develop activities that build capacity in Data Science and Big Data applications.

To address the challenges and opportunities outlined above, interdisciplinary collaboration is needed between research communities in computer science and agriculture. We envision the collaboration happening in two directions, namely, from computer-science to agricultural-science (C2A) and from agricultural-science to computer-science (A2C).

In C2A, computer scientists introduce big data ideas to data-driven agricultural scientists through specifically designed workshops, seminars and courses. Successful AgBD applications (e.g. GEOGLAM) will be introduced to illustrate how AgBD can be used to achieve new goals in agricultural science. In A2C, agricultural scientists will thoroughly explain and list open agricultural problems to computer scientists so that computer scientists can know exactly which domain problems to work on and which goals to achieve. In addition, agricultural scientists will create benchmark datasets for the open problems so that computer scientists can evaluate the performance of new algorithms and avoid over-fitting. NIFA funding is critical for both the C2A workshops and seminars, as well as the careful defining of open problems in agriculture and the generating of A2C benchmark datasets.

The development of knowledge-exchange software will be another core component of this collaboration. Software provides a framework for formal collaboration and encoding of knowledge. In addition, software and data provide a formal representation of knowledge, and thus a 'truth' that cuts across the interdisciplinary barriers in language and understanding. Conversion of textbook knowledge into tools for interoperability and QAQC are excellent applied exercises for training both agricultural and computer scientists. The subsequent extension of such tools to incorporate more modern statistical tools and concepts in informatics will provide opportunities for graduate student and postdoc level work to build new research and interdisciplinary skills that are valuable for both academic research and industry.

6. Acknowledgement

We thankfully acknowledge the contributions of Yiqun Xie, Melissa Cragin and participants of the following workshops supported or led by the Midwest Bug Data Hub: Machine Learning: Farm To Table (MBDH) [57], Agricultural Data Integration: From Genomics to Unmanned Systems (Institute for Mathematics and its Applications and MBDH@University of North Dakota) [58], and Data Science in Agriculture Summit (National Institute of Food and Agriculture) [59], ACMSIG KDD Workshop on Data Science for Food, Energy and Water [60] [61]. This material is based upon work supported by the National Science Foundation under Grants No. 1550320, 1541876, 1029711, IIS-1320580, 0940818 and IIS-1218168, the USDOD under Grants No. HM1582-08-1-0017 and HM0210-13-1-0005, ARPA-E under Grant No. DE-AR0000795, USDA under Grant No. 2017-51181-27222, USDA NIFA under Grant No. 2016-38831-25874, the CRA Computing Community Consortium and the OVPR Infrastructure Investment Initiative and Minnesota Supercomputing Institute (MSI) at the University of Minnesota. We also thank Kim Koffolt for improving the readability of the article.

References

[1] M. Stubbs, "Big Data in U.S. Agriculture, Congressional Research Service," 2016. [Online]. Available: https://fas.org/sgp/crs/misc/R44331.pdf.

[2] "World Population Prospects, The 2015 Revision: Key Findings and Advance Tables. Department of Economic and Social Affairs, Population Division, United Nations," 2015. [Online]. Available: https://esa.un.org/unpd/wpp/publications/files/key_findings_wpp_2015.pdf.

[3] N. Abe and et al, "Data Science for Food, Energy and Water: A Workshop Report," ACM SIGKDD Explorations Newsletter, 2017.

[4] "NSF Workshop to Identify Interdisciplinary Data Science Approaches and Challenges to Enhance Understanding of Interactions of Food Systems with Energy and Water Systems," Computing Research News (ISSN 1069-384X), vol. 27, no. 10, 2015.

[5] "Global Crop Monitoring. Group on Earth Observations Global Agricultural Monitoring (GEOGLAM) Initiative," 2017. [Online]. Available: https://www.earthobservations.org/geoglam.php.

[6] "Remote Sensing Technology Trends and Agriculture, Digitalglobe," 2015. [Online]. Available: https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/31/DG-RemoteSensing-WP.pdf.

[7] C. Rosenzweig and et al, "The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies," Agricultural and Forest Meteorology, vol. 170, pp. 166-182, 2013.

[8] J. D. George Hanuschak, "Utilization of Remotely Sensed Data and Geographic Information Systems (GIS) for Agricultural Statistics in the United States and the European Union," Advances in planning, design and management of irrigation systems as related to sustainable land use, pp. 14-17, 1993.

[9] M. E. Bock, N. J. Kirkendall and et al., "Improving Crop Estimates by Integrating Multiple Data Sources," The National Academies Press, 2017.

[10] D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," Biosystems Engineering, vol. 114, no. 4, pp. 358-371, 2013.

[11] "Capitol Hill Presentation on Deconstructing Precision Agriculture," Computing Research News (ISSN 1069-384X), vol. 27, no. 4, 2015.

[12] K. Noyes, "Big data poised to change the face of agriculture, Fortune," 2014. [Online]. Available: http://fortune.com/2014/05/30/cropping-up-on-every-farm-big-data-technology/.

[13] V. Estes, "How Big Data is Disrupting Agriculture from Biological Discovery to Farming Practices, Ag Funder News," 2016. [Online]. Available: https://agfundernews.com/how-big-data-is-disrupting-agriculture-from-biological-discovery-to-farming-practices5973.html.

[14] S. Shekhar, S. Feiner and W. Aref, "Spatial computing," Commun. ACM, vol. 59, no. 1, pp. 72-81, 2015.

[15] R. R. Vatsavai and et al., "Spatiotemporal data mining in the era of big spatial data: algorithms and applications," Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, pp. 1-10, 2012.

[16] Z. Jiang and S. Shekhar, "Spatial and Spatiotemporal Big Data Science," in Spatial Big Data Science, Springer, 2017, pp. 15-44.

[17] "DARPA. Broad Agency Announcement, Geospatial Clould Analytics (HR001118S0004)," 2017. [Online]. Available: https://www.fbo.gov/index?s=opportunity&mode=form&id=30e9d3053a666eca911148b744ec9602&tab=core&_cview=1.

[18] K. Sofer, "The California Drought's Lessons for Food Security, Slate," 2016. [Online]. Available: http://www.slate.com/blogs/future_tense/2016/06/22/the_california_drought_s_lessons_for_food_security.html.

[19] P. Huttner, "Climate Cast: No choco-pocalypse yet but cocoa could become scarce, MPR News," 2017. [Online]. Available: https://www.mprnews.org/story/2017/05/04/cocoa-and-coffee-might-become-less-available-more-expensive.

[20] "Sustainable palm oil," 2014. [Online]. Available: https://www.cargill.com/sustainability/palm-oil/sustainable-palm-oil.

[21] S. Shekhar and et al., "Intelligent Infrastructure for Smart Agriculture: An Integrated Food, Energy and Water System, Computing Community Consortium," 2017. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1705/1705.01993.pdf.

[22] B. Mynatt and et al, "A National Research Agenda for Intelligent Infrastructure, CCC Led Whitepapers," [Online]. Available: http://cra.org/ccc/resources/ccc-led-whitepapers/.

[23] "G-8 International Conference on Open Data for Agriculture, " 2013. [Online]. Available: https://sites.google.com/site/g8opendataconference/home.

[24] "Global Open Data for Agriculture and Nutrition," 2017. [Online]. Available: http://www.godan.info/.

[25] "National Oceanic and Atmospheric Administration (NOAA)," [Online]. Available: http://www.noaa.gov/.

[26] "National Aeronautics and Space Administration (NASA)," [Online]. Available: https://www.nasa.gov/.

[27] "U.S. Bureau of Labor Statistics," [Online]. Available: https://www.bls.gov/.

[28] "Earth on AWS," [Online]. Available: https://aws.amazon.com/earth/. [Accessed 2017].

[29] "Google Earth Engine," [Online]. Available: https://earthengine.google.com/. [Accessed 2017].

[30] "NASA Earth Exchange," [Online]. Available: https://nex.nasa.gov/nex/. [Accessed 2017].

[31] "National Agricultural Statistics Service (NASS)," [Online]. Available: https://www.nass.usda.gov.

[32] "Economic Research Service (ERS)," [Online]. Available: https://www.ers.usda.gov/.

[33] R. R. Rushforth and B. L. Ruddell, "A Spatially Detailed and Economically Complete Blue Water Footprint of the United States," Hydrology and Earth System Sciences Discussions, pp. 1-54, 2017.

[34] "Agricultural Research Service (ARS-U)," [Online]. Available: https://www.ars.usda.gov/.

[35] "Natural Resources Conservation Service (NRCS)," [Online]. Available: https://www.nrcs.usda.gov/.

[36] "Agricultural Marketing Service (AMS)," [Online]. Available: https://www.ams.usda.gov/.

[37] "World Agricultural Outlook Board (WAOB)," [Online]. Available: https://www.usda.gov/oce/commodity/.

[38] "VegScape," [Online]. Available: https://nassgeodata.gmu.edu/VegScape/.

[39] "World Animal Health Information System (WAHIS)," [Online]. Available: www.oie.int/wahis/.

[40] "EMPRES Global Animal Disease Information System (Empres-i2)," [Online]. Available: empres-i.fao.org/.

[41] "Risk Management Agency (RMA)," [Online]. Available: https://www.rma.usda.gov/.

[42] "Farm Service Agency (FSA)," [Online]. Available: https://www.fsa.usda.gov/.

[43] "Food Safety and Inspection Service (FSIS)," [Online]. Available: https://www.fsis.usda.gov/.

[44] "Food Protection and Defense Institute, University of Minnesota," [Online]. Available: https://foodprotection.umn.edu/.

[45] "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies," [Online]. Available: https://www.whitehouse.gov/omb/fedreg_final_information_quality_guidelines.

[46] "USDA Agricultural Projections to 2026," 2017. [Online]. Available: https://www.usda.gov/oce/commodity/projections/USDA_Agricultural_Projections_to_2026.pdf.

[47] C. Rosensweig and et al, "Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison," PNAS, vol. 111, no. 9, 2014.

[48] S. J. Janssen, C. H. Porter, A. D. Moore, I. N. Athanasiadis, I. Foster, J. W. Jones and J. M. Antle, "Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology," Agricultural Systems, vol. 155, pp. 200-212.

[49] Y. Xie, B. Runck, S. Shekhar, L. Kne, D. Mulla, N. Jordan and P. Wiringa, "Collaborative Geodesign and Spatial Optimization for Fragmentation-Free Land Allocation," ISPRS International Journal of Geo-Information, vol. 6, no. 7, 2017.

[50] Y. Xie and S. Shekhar, "FF-SA: Fragmentation-Free Spatial Allocation," Spatial and Temporal Databases, 2017.

[51] M. Dietze, D. Lebauer and R. Kooper, "On improving the communication between models and data," Plant, Cell & Environment, vol. 36, no. 9, pp. 1575-1585, 2013.

[52] S. S Schneeweiss, "Learning from big health care data," New England Journal of Medicine, 2014.

[53] R. Kao, D. Haydon, S. Lycett and P. Murcia, "Supersize me: how whole-genome sequencing and big data are transforming epidemiology," Trends in Microbiology, 2014.

[54] D. Pfeiffer and K. Stevens, "Spatial and temporal epidemiological analysis in the big data era," Preventive Veterinary Medicine, 2015.

[55] K. VanderWaal and et al., "Translating Big Data into smart Data for veterinary epidemiology," Frontiers in Veterinary Science, vol. 4, no. 110, 2017.

[56] M. Gates, L. Holmstrom, K. Biggers and T. Beckham, "Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: challenges and opportunities," Frontiers in Public Health, 2015.

[57] "Workshop on Machine Learning: Farm To Table, Midwest Big Data Hub," 2017. [Online]. Available: https://publish.illinois.edu/machine-learning-farm-to-table-workshop/.

[58] "Workshop on Agricultural Data Integration: From Genomics to Unmanned Systems, Institue for Mathematics and its Applications," [Online]. Available: https://www.ima.umn.edu/2017-2018/SW10.26-28.17.

[59] "Data Science in Agriculture Summit, the National Institute of Food and Agriculture (NIFA)," 2017. [Online]. Available: https://nifa.usda.gov/data-science-agriculture-summit.

[60] "ACM SIGKDD Workshop on Data Science for Food, Energy and Water.," 2016. [Online]. Available: https://sites.google.com/site/2016dsfew/home.

[61] "ACMSIG KDD Workshop on Data Science for Intelligent Food, Energy, and Water (DSIFER)," 2017. [Online]. Available: http://ai4good.org/few17/.

[62] "Food and Agriculture Cyberinformatics and Tools (FACT) Initiative, the National Institute of Food and Agriculture (NIFA)," 2017. [Online]. Available: https://nifa.usda.gov/announcement/usda-announces-150-million-funding-through-agriculture-and-food-research-initiative.

[63] Y. Xie, E. Efteliogl, R.Y. Ali, X. Tang, Y. Li, R. Doshi, & S. Shekhar. Transdisciplinary Foundations of Geospatial Data Science. ISPRS International Journal of Geo-Information, 6(12), 2017.

[64] S. Shekhar; Z. Jiang; R.Y. Ali; E. Eftelioglu; X. Tang; V. Gunturi; X. Zhou. Spatiotemporal Data Mining: A Computational Perspective. ISPRS International Journal of Geo-Information, 4, 2306–2338, 2015.

[65] S. Shekhar, M.R. Evans, J.M. Kang, P. Mohan. Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery , 1(3), 193-214, 2011.