Data Science for Food, Energy and Water: A Workshop Report

Naoki Abe[1], Yiqun Xie[2], Shashi Shekhar[2], Chid Apte[1], Vipin Kumar[2], Mitch Tuinstra[3], Ranga Raju Vatsavai[4]
[1]IBM Research,[2]University of Minnesota,[3]Purdue University,[4]North Carolina State University
[1]{nabe,apte}@us.ibm.com, [2]{xiexx347,shekhar,kumar001}@umn.edu, [3]mtuinstr@purdue.edu, [4]rrvatsav@ncsu.edu

ABSTRACT

At the 22nd ACM SIGKDD conference on Knowledge and Data Discovery (KDD), a workshop on Data Science for Food, Energy and Water (DSFEW) was held to foster an interdisciplinary community intersecting data science and societally important domains of food, energy and water. The workshop included keynotes, panel discussion, presentations and posters, and introduced the emerging area of DSFEW to ACM SIGKDD audience, and triggered interdisciplinary idea-sharing in DSFEW research. The workshop website is sites.google.com/site/2016dsfew.

Keywords

food, energy and water nexus; data science

1. BACKGROUND

In the coming decades, the world population is projected to grow significantly (Fig. 1). Thus, securing the essential resources of food, energy and water (FEW), is one of the most pressing challenges the world faces today. The challenge is made harder due to climate change, rising economies and interactions among food, water and energy systems.

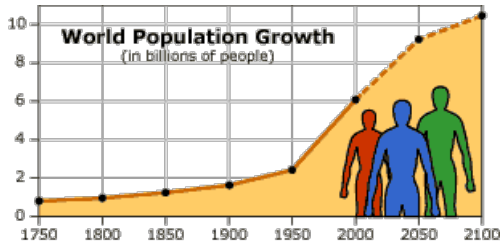

Figure 1: Projected world population growth [7].

It is difficult to consider food, water or energy security in isolation due to their complex interactions. For example, energy production needs water for cooling and may use bio-fuels. Conversely, food production requires energy and water as shown in Fig. 2. Trying to achieve energy security in isolation may lead to unanticipated surprises for food and water security [13]. For example, food prices rose in many parts of the world in 2008 coincident with increased subsidies

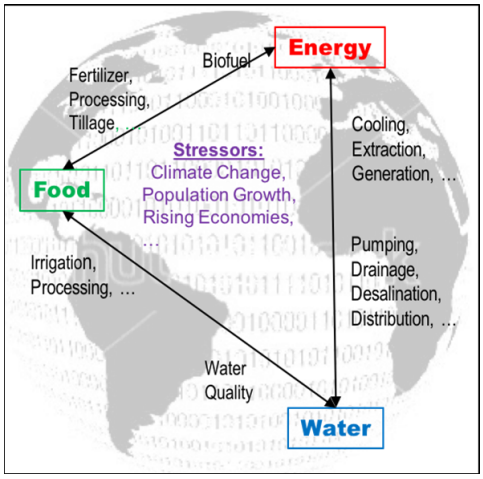

Figure 2: Interactions among Food, Energy, Water Systems (best in color) [2]. For example, food production not only needs water for irrigation and energy for fertilizer but also may degrade water quality due to run-offs.

for biofuels. Similarly, incentives for growing crops have depleted water resources (e.g., Aral Sea, Ogallala aquifer) and affected water quality (e.g., dead zone in Gulf of Mexico). To reduce such unanticipated consequences, the nexus approach jointly considers the interactions among food, energy and water systems [11].

Understanding the FEW nexus is among the highest priorities at the United Nations [8] as well as many countries. In 2011, Stockholm Environment Institute initiated a conference on "The Water, Energy and Food Security Nexus, Solutions for the Green Economy" to better understand the nexus [10]. In 2014, U.K. funded a set of research proposals on FEW (e.g., WEFWEBs at University of Glasgow) [1]. In U.S., a recent National Intelligence Council report identified it among the greatest challenges facing our world in the coming decades [12]. The US National Science Foundation (NSF) has also started a multi-year cross-directorate initiative titled Innovations at the Nexus of Food, Energy and Water Systems (INFEWS) [4]. More international research efforts are in need to address global FEW challenges (e.g., global FEW choke point in China, US and India [3]).

In 2015, US NSF sponsored a set of workshops to engage a diverse set of research communities to identify research chal-

lenges and opportunities. This ACM SIGKDD workshop on DSFEW is motivated by the NSF workshop, "A Workshop to Identify Interdisciplinary Data Science Approaches and Challenges to Enhance Understanding of Interactions of Food Systems with Energy and Water Systems" [2].

The US NSF INFEWS data science workshop [2] identified research needs in five key areas: 1) Integration of data sets and data-driven models of multiple types at many spatial and temporal scales, 2) Predictive and causal modeling of spatial and temporal data, with particular attention to auto-correlation, non-stationarity, and scale of FEW data, 3) Multi-stakeholder decision support, including methods for eliciting and sharing preferences, supporting negotiation and consensus building, 4) FEW nexus life cycle thinking, and 5) FEW data uncertainty, incompleteness, and bias. The workshop also underscored the need for community infrastructure, such as shared data sets, evaluation metrics, models, and tools, and the training of a new generation of scientists with the requisite training in the data sciences and the FEW sciences to facilitate progress at their interface.

The goals of this workshop were: (1) To introduce the emerging area of DSFEW to the KDD community; (2) To invite scientists and practitioners in FEW domains to the KDD community, and interest them in leveraging our technology and expertise; and (3) To innovate new technology, leveraging existing KDD technology where appropriate, to address the challenges in FEW, by bringing together a multidisciplinary audience and enticing them to synergize.

The workshop included two keynote presentations on predictive phenomics of plants and remote sensing in agricultural applications, followed by a set of presentations and posters on phenotyping, object recognition, change detection, prediction, regression and optimization. A panel discussion was held with leading experts in DSFEW on the motivation, problems, current stage, challenges and next steps in DSFEW research.

2. KEYNOTE PRESENTATIONS

Dr. Naoki Abe and Dr. Chid Apte opened the workshop and introduced the keynote speakers. Prof. Patrick Schnable, Director of Plant Sciences Institute and Center for Plant Genomics at Iowa State University, highlighted how field sensors and data science approaches are helping to build automated phenotyping systems and predict crop performance. Crop phenotypes (e.g., yield and drought tolerance) are determined by genotype, environment and their interaction (GxE). Genotyping data have been made accessible for all major crops. However, phenotypic data, a powerful source to statistically model GxE, remains with a limiting volume. Prof. Schnable described a framework to collect phenotypic data using high-throughput and high-resolution field sensors and robots. The dataset is analyzed using data science techniques (e.g., machine learning, correlation analysis) that help model relationships between phenotype properties and performance of the plants. Prof. Schnable has also explored interactions between neighboring plants. This analysis revealed that seed orientation influenced performance of adjacent plants.

The second keynote, Prof. Melba Crawford, associate dean for research in the College of Engineering at Purdue University, summarized the types of data collected through remote sensing and how they are applied to agriculture. With recent developments in remote sensing, data are generated via a variety of platforms (e.g., space-based, airborne, proximal sensing platforms). For space-based platforms, sensors are evolving from complex, multi-purpose to lower cost and measurement specific constellations of small satellites. Hyperspectral satellite images contain rich information for detailed spectral analysis of crops. Advanced data science techniques are necessary to explore the increased volume of data introduced by hyperspectrum, high resolution and high time frequency. For airborne and proximal sensing platforms (e.g., unmanned aerial vehicles, autonomous vehicles), new data sources, such as LiDAR point cloud, are becoming increasingly popular in agricultural applications. Topographical dataset (e.g., digital elevation models) can be derived from LiDAR point cloud and provide additional height information in analysis. These massive, multi-modality datasets from new sensors offer opportunities in agriculture applications from plant-level phenotyping to large-scale crop mapping and plantation monitoring. New algorithms in data science are needed to address challenges of multi-temporal, multi-scale and multi-sensor analysis.

3. PANEL DISCUSSION

Four major questions were posed in the panel discussion: (1) Context: How are food, energy and water communities exploiting data and data-science? (2) Gap Analysis: What are the key pain-points in leveraging data and data-science for food, energy and water? What are the data and data-science knowledge gaps in context of food, energy and water? (3) Nexus: Past approaches to solving Food problems (e.g., via fertilizers) had unanticipated negative impacts to water (e.g., water quality degradation). How may data and data-science help improve understanding of the interactions among food, energy, and water systems? How may they reduce unintended consequences ? (4) Engagement, Community Building: Why should data scientists engage with food, energy and water? How may we build and sustain a community of food, energy and water data science?

Dr. Sivan Aldor-Noiman, director of data science at The Climate Corporation, discussed the complex decision-making challenge faced by food growers and pointed out a major flaw in general machine learning methods when applied to FEW. Machine learning methods broadly have a well-known assumption on their variables that they are independent and identically distributed (i.i.d.). I.i.d assumption does not well-fit FEW datasets since the variables (e.g., soil property, water quality) are spatially autocorrelated under the first law of geography. Considering autocorrelation and brining it into machine learning models poses a great challenge in DSFEW research. Prof. Ronald Turco, the director of the Indiana Water Resources Research Center at Purdue University, discussed the importance of visiting farms to understand farmers' vocabularies and concerns. For example, farmers care about heavy rainfalls and droughts which are worse due to climate change. Dr. David Lapen, research scientist at Agriculture and Agri-Food Canana, shared his experience on water management in agricultural fields. Dr. Charlie Messina, senior scientist at Dupont Pioneer, described his perspectives on DSFEW-related research from his domain expertise. He emphasized the importance of incorporating expert knowledge and on-site experience into data science instead of finding solutions by modeling webpage click-statistics as is often done by some industry companies. Prof. Shashi Shekhar, McKnight Distinguished Pro-

fessor of Computer Science at University of Minnesota, reviewed how food-water-energy relate to the United Nations' 2030 goals for sustainable development. He then outlined the importance of spatial computing in these endeavors [9] and noted research challenges that lie ahead. He concluded by sharing successful stories in DSFEW (e.g., GeoGlam [6]).

In audience question section, Dr. Ramasamy Uthurusamy, a founder of ACM SIGKDD, challenged the audience to identify the moonshots of DSFEW if extensive funding (e.g., one US billion dollars) is available. Examples given by panel speakers include solving world hunger and clear drinking water to everyone. One interesting industry perspective is to crowd-source ideas and solutions from a series of hackathons without huge investment.

Dr. Ramasamy Uthurusamy also encouraged holistic thinking on the life cycle of DSFEW. The workshop focuses mostly on model selection in DSFEW. However, model selection accounts for only a limited portion (e.g., 10%) of the DSFEW life cycle. Data collection, data cleaning, business modeling, policy making, etc., all play critical roles in DSFEW life cycle. Research efforts on all stages of the life cycle require a commitment to community building and interdisciplinary collaboration.

Dr. Nikunj Oza, group lead of Data Sciences in Intelligent Systems Division at NASA, asked about the use of secondary information in FEW domains. Panelists mentioned value of such information (e.g., tweets) in detecting food-borne illnesses and food contamination.

Other discussions include applying advances in precision agriculture to small-farm owners, privacy and security protection with data sharing which are of farmers' concerns, and solutions to climate change under political challenges.

4. PRESENTATIONS AND POSTERS

The presentations and posters are grouped into three tiers: (1) **Domain application**: food, energy and water, as well as their nexus; (2) **Data analytics**: machine learning, data mining, regression and remote sensing techniques; and (3) **Infrastructure**: computation platform and system build. Majority of the work presented in the workshop belong to domain application. A summary of publications is shown in Table 1. In the table, data collection and data management belong to the category "infrastructure".

Food, energy and water: In food-related applications, interet was shown in automated phenotyping of crop plants and pest monitoring. Phenotyping, the process of characterizing properties and traits of plants, can be used to predict the status and performance of plants. Different types of image data are used to compute the phenotypes, including aerial imagery/LiDAR collected by UAV, ground images captured by field cameras and high contrast plant images taken in the lab. Unlike images captured in the field and labs, aerial imagery collected by wide-angle lens cameras on UAVs needs to be preprocessed into ortophotos before phenotyping. In the phenotyping phase, LiDAR point cloud can be used to generate digital elevation models to further incorporate height property of plants. Computer vision techniques (e.g., feature point extraction, pattern recognition) are applied to identify plant structures (e.g., leaves, stems, tassels), measure geometric properties and compute phenotypes. 3D images taken at different angles are also used to generate new phenotype characteristics by comparing morphological metrics among multiple angles. For pest monitoring, a convolution neural network approach is used to find soybean Cyst Nematode eggs with different rotations, shapes and scales. A network-based approach is applied to model the dynamics of invasive plant pests.

In water and energy, applications presented focus on water conservation and hydro-based power generation modeling. For water conservation, an estimation model was created to predict how much water can be saved through turf removal in California urban landscape to deal with the ongoing severe drought. Hydro-based power generation was modeled based on hydro-lake river inflows. A regularized linear regression Lasso method is applied on large scale oceanic-climatic predictors with high-dimensional data but small sample sizes. The forecasted stream flow information is used to estimate electricity production of hydro power stations in Waitaki catchment in New Zealand, which yields about 40% electricity of the country.

Data Analytics: Data analytics approaches were proposed to analyze remote sensing datasets to monitor land cover, identify land cover changes and optimize future spatial allocations of land covers. As a social concern, ethical issues in data science were also discussed.

To assist land cover monitoring, an automatic plantation mapping approach with ensemble learning and hidden Markov model was proposed to estimate palm oil cultivation in southeast Asia and enforce sustainability standard.

For land cover change detection, a support vector machine and convolutional neural network based method was constructed to classify land covers and locate urbanization (e.g., agriculture to residential) in West Bengal, India. Additionally, a multi-instance and multi-view learning framework was proposed to identify spatiotemporal change footprints where lake and river shrinkage happens.

Planning ahead with land cover allocation, a geodesign optimization tool was introduced to facilitate redesign of landscape in agricultural watersheds in the mid-western US. The nexus goal of geodesign is to improve water quality while still providing enough food under economic budget.

Infrastructure: Two infrastructure building efforts on data collection and management were presented. The first introduced an integrated knowledge graph for FEW, built on semantic web technologies and statistical relational learning. The goal is to harmonize diverse FEW data sources to perform ontology analysis. The second system, SmartFarm, combines sensor technologies and cloud computing platform to assist growers making decisions in precision farming with local farm statistics and a variety of external data inputs (e.g., weather predictions, satellite imagery).

5. CONCLUSION & NEXT STEPS

The ACM SIGKDD workshop on Data Science for Food, Energy and Water (DSFEW) introduced the emerging area of DSFEW to KDD data science community and inspired interdisciplinary idea-sharing [5]. Recent research results in FEW domain applications, data science approaches and system infrastructures were presented through keynotes, presentations and posters. Critical research questions in DSFEW were discussed in the panel discussion. The workshop participants are looking forward to growing the DSFEW community through publications (e.g., conference with special interest group, journal special issue) and competitions (e.g., KDD DSFEW challenge).

Table 1: Summary of workshop publications [5]

Title	Food	Energy	Water	Data collection	Data management	Data Analytics
A Knowledge Ecosystem for the Food, Energy, and Water System	✓	✓	✓		✓	
An end-to-end convolutional selective autoencoder approach to Soybean Cyst Nematode eggs detection	✓					✓
Automated Sorghum Phenotyping and Trait Development Platform	✓			✓		✓
Automated Vegetative Stage Phenotyping Analysis of Maize Plants using Visible Light Images	✓			✓		✓
Predictive Modeling of Sorghum Phenotypes with Airborne Image Features	✓					✓
Fast, automated identification of tassels: Bag-of-features, graph algorithms and high throughput computing	✓			✓		✓
Estimating Phenotypic Traits From UAV Based RGB Imagery	✓			✓		✓
What spins the turbine? Finding spatial climate precursors of hydro-lake inflows: Waitaki catchment, New Zealand		✓	✓			✓
How Much Water Does Turf Removal Save? Applying Bayesian Structural Time-Series to California Residential Water Demand		✓	✓			✓
Satellite Image Analytics, Land Change and Food Security						✓
A Bayesian Network approach to County-Level Corn Yield Prediction using historical data and expert-knowledge	✓					✓
Modeling the Food-Energy-Water Nexus in Critical Biodiverse Landscapes: A Case Study of Tonle Sap, Cambodia and Tulalip Tribe, USA	✓	✓	✓			✓
Plantation Mapping in Southeast Asia	✓					✓
SmartFarm: Improving Agriculture Sustainability Using Modern Information Technology	✓				✓	✓

6. ACKNOWLEDGEMENTS

The workshop was supported by NSF, the Computing Community Consortium and the Midwest Big Data Hub. We thank Jayant Gupta for his workshop notes.

7. REFERENCES

[1] Water energy food: WEFWEBs, research councils UK. http://gtr.rcuk.ac.uk/project/EDD08B5A-61DE-41CA-B967-3554BE85CCBA, 2014.

[2] Computing research association, NSF workshop to identify interdisciplinary data sci. approaches and challenges to enhance understanding of interactions of food sys. with energy and water sys. *Computing Research News*, 27(10), 2015.

[3] KSIConnect, Global choke point: Water-energy-food confrontations in china, us and india. https://www.youtube.com/watch?v=6F2tx903FMI, 2015.

[4] National Science Foundation, Innovations at the nexus of food, energy and water systems. www.nsf.gov/about/budget/fy2016/pdf/37_fy2016.pdf, 2015.

[5] ACM SIGKDD 2016 workshop on data science for food, energy and water. https://sites.google.com/site/2016dsfew, 2016.

[6] GeoGlam (global agricultural monitoring initiative). www.geoglam-crop-monitor.orgf, 2016.

[7] Museum of Paleontology, The ecology of human populations. http://evolution.berkeley.edu, 2016.

[8] United Nations, Transforming our world: the 2030 agenda for sustainable development. sustainabledevelopment.un.org/?menu=1300, 2016.

[9] E. Eftelioglu, Z. Jiang, R. Ali, and S. Shekhar. Spatial computing perspective on food energy and water nexus. *J. of Env. Studies and Sci.*, 6(1):62–76, 2016.

[10] M. Escobar. The nexus of water-energy-food: an introduction to the Inter-ADB sustainability report 2011. *Inter-American development bank sustainability report 2011*, pages 7–9, 2012.

[11] J. Liu, H. Mooney, et al. Systems integration for global sustainability. *Science*, 347(6225):1258832–1–9, 2015.

[12] National Intelligence Council. Global trends 2030: Alternative worlds. globaltrends2030.files.wordpress.com/2012/11/global-trends-2030-november2012.pdf, 2012.

[13] USDOE. The water energy nexus: Challenges and opportunities. http://www.energy.gov/sites/prod/files/2014/06/f16/Water%20Energy%20Nexus%20Report%20June%202014.pdf, 2014.