Project Summary

As our food sources, producers, and environment become increasingly diverse, and as demands on the food system grow, concerns of food safety, security, waste, and provenance are paramount. Beyond traditional concerns of contaminants and food-borne illnesses, factors relating to pesticide use, antibiotic resistance, and environmental factors are important. An essential component of a comprehensive food safety and security infrastructure is the ability to accurately track the source and movement of food products through the supply chain. We propose a fundamentally new framework for food provenance based on profiling of plant DNA to provide robust end-to-end tracking. Combined with existing instrumentation in food processing and supply chains, the proposed framework provides a powerful framework for food safety diagnoses, analyses, and prognostics. It also provides necessary basis for identifying origin of waste and inefficiencies in the supply chain, while enabling powerful studies correlating genotype and growth environment with phenotypic outcomes. The project aims to do for food sources, what digital watermarking and blockchains have done for digital content.

Intellectual Merit. The food industry is a critical socioeconomic driver – one that has been massively underserved by innovations in computing. The goal of this project is to bring disruptive innovation to food safety, quality, distribution, and production. It achieves these goals through: (i) novel computational genomics techniques for food provenance; (ii) use of large-scale instrumentation for supply chain modeling and optimization; (iii) incentive mechanisms for participation across the supply chain; (iv) integrated blockchains for an open marketplace for producers, processors, distributors, retailers, and consumers; and (v) use of provenance information combined with seed, field management, and weather data to comprehensively map quality and productivity.

Broader Impact. Our broader impact program scales from outreach to local producers, distributors, and retailers, to the nationwide community of stakeholders through the proposed CropHub, and to the international community, through targeted outreach to NGOs in India, with the goal of customizing and delivering context-specific solutions. Using the existing and highly successful Purdue Outreach program, we will offer a sequence of workshops and tutorials (both in-person and online) to practitioners and policymakers. A concrete broader impact deliverable is the development of CropHub, a web resource for integrating diverse data streams, from crop sources, soil and weather data, nutrient use data, and supply chain data. We will design the Hub to be fully integrated, customizable, and extensible. We will initiate the Hub with a rich set of data from field experiments at Purdue. Beyond research infrastructure, CropHub will also provide a powerful platform for education, outreach, and broadening participation.

Education Programs. The unique combination of areas constituting the project presents tremendous potential for innovative curricula and instruction. Our educational activities span all levels, from high-school students and undergraduates to graduate students and postdocs. Our overarching goal is to create and train the next generation of intellectual leaders in this new area.

Broadening Participation in Computing. The project includes a comprehensive plan for broadening participation to (i) first-time college attendees; (ii) students from rural backgrounds; and (iii) women and underrepresented minorities. Our proposed programs are supplemented by well-defined success metrics, reliance on best-practices, continual assessment, and program evaluation.

To achieve these goals, the project brings together a cohesive and committed group of investigators, comprised of faculty from Computer Science, Food Science, Agriculture, and Plant Biology. Beyond comprehensive intellectual coverage, the team also has significant experience in running large multidisciplinary programs, education and outreach efforts, and contributions to broadening participation to underrepresented groups. The project significantly leverages institutional infrastructure in all aspects of the project.

The food industry is a critical socioeconomic driver – one that has been massively underserved by innovations in computing. The goal of this project is to bring disruptive innovation to food safety, quality, distribution, and production. It achieves these goals through: (i) novel computational techniques for food provenance; (ii) use of large-scale instrumentation for supply chain modeling and efficiency; (iii) incentive techniques for participation across the supply chain; (iii) integrated blockchains for an open marketplace for producers, processors, distributors, retailers, and consumers; and (v) use of provenance information combined with seed, field management, and weather data to comprehensively map quality and productivity. The multidisciplinary project scope presents exciting new opportunities for education, broader impact, and broadening participation in computing. In each of these facets, the proposed project targets important underrepresented groups, including aspiring students from rural backgrounds, first-time college attendees, women and minorities, professionals in the food industry, and companies across the spectrum for technology transfer. The project also has international scope – reaching out to developing countries with the goal of customizing low cost technologies to specific markets and contexts.

1 MOTIVATION AND PROJECT OVERVIEW

The global food security challenge impacts nearly all aspects of economies and societies [30]. The aspirational vision of a 'a world free from hunger and malnutrition, where food and agriculture contribute to improving the living standards of all, especially the poorest, in an economically, socially, and environmentally sustainable manner' is absolutely critical [57, 178, 179]. These goals are not new, yet despite dedicated efforts, and estimates that enough food is currently produced to feed the global population, hunger persists in both developed and developing countries. One-sixth of Americans do not have enough food to eat [129]; worldwide over 795 million people suffer from hunger, and the prevalence of undernourishment has been rising [57]. Challenges facing the food and agriculture sector include: (i) population growth that is anticipated to result in doubling agricultural demand by 2050, with disproportionately higher consumption of meat, fruits, and vegetables relative to cereals; (ii) growth of crop yields is slowing despite innovations in productivity; (iii) food losses and waste claim a significant proportion of agricultural output, globally, from production through to the consumer; (iv) critical parts of food systems are becoming more capital intensive; and (v) food distribution networks are becoming more complex and geographically distributed, which increases the resource-, energy-, and emission- intensity of the global food system [57].

An estimated 20 billion pounds of produce is lost on farms in the US each year for a variety of reasons (lack of planning, overplanting, improper field management, disease, adverse weather, cosmetic imperfections, and market price fluctuations leading to decisions not to harvest when prices don't cover the cost of bringing product to market) [157]. Food loss then propagates through distribution and on to the consumer. Although cold chain and supply chain logistics efforts are reducing food loss and waste during transit between harvest and retail sale, estimates indicate that $7-15.4 billion worth of fresh produce spoils annually before reaching the consumer in the US. This corresponds to approximately 12.3% of fruits and 11.6% of vegetables [32, 57]. The consumer then accounts for an additional 19% of fruit and 22% of vegetable losses [32, 74]. In September 2015, the UN, USDA, and EPA set goals to cut food waste in half by 2030 [74]. To meet this goal, transformative change is needed at all levels of the food system.

The United States, arguably, has the safest food supply in the world. Yet even in the US, the annual burden of foodborne illness, as estimated by the Centers for Disease Control and Prevention (CDC), is 48 million cases, 128,000 hospitalizations, and 3,000 deaths, with a cost of $152 billion in medical expenses, lost productivity and business, lawsuits, and compromised branding [161, 194]. Globally, the annual numbers are higher: 600 million foodborne illnesses, 230,000 deaths from foodborne diarrheal disease agents, and over 33 million disability adjusted life years [185]. The World Health Organization has concluded that foodborne diseases are 'a significant impediment to socio- economic development worldwide'. Diagnosing the causes of foodborne illnesses and their mitigation are significantly hampered by the length of time and complexity of tracing back the illness to its source and implementing effective control and prevention measures [185].

Recent outbreaks associated with romaine lettuce offer insight into the challenges involved in identifying and resolving foodborne outbreaks. The number of cases involving leafy greens in the US has dramatically increased in the last two decades [33, 44, 50]. Fresh fruits and vegetables were associated with less than 1% of reported outbreaks in the 1970s but over 6% in the 1990s, with most involving E. coli O157:H7, Salmonella species, and Listeria monocytogenes [164]. Between 1998-2008, the CDC reported that leafy greens accounted for 23% of all cases of foodborne illness in the US [148]. Leafy greens are now recognized as a significant reservoir of human bacterial pathogens [26, 50]. Leafy greens grow low to the ground and therefore can easily be contaminated by bacteria that are naturally present in the soil or the surrounding environment [56]. Leafy greens are typically consumed fresh with no processing steps at the time of consumption that would reduce microbial loads [19]. Cut and prepackaged salad products, designed to meet consumer demands for convenience, create scenarios for cross-contamination. The increasing complexity of food distribution networks, and consumer demands for year-round access to seasonal products, create extended times between harvest and consumption, wherein bacteria are given longer times to proliferate. This is accompanied by an increasingly globalized food network, with vastly varying regulations and practices – most of the fresh fruit and almost a third of the fresh vegetables Americans buy come from other countries.

The consensus opinion is that transformative change in both agriculture and food systems is critically needed to address numerous challenges facing the food and agriculture sector and to minimize, if not eradicate, hunger and improve food safety and quality (FAO, 2017). A critical component of this change is the use of computing technologies in all components of the food systems – provenance for auditing food systems, optimization for minimizing wastage and increasing production, incentive mechanisms for best-practices, data and software resources for maximizing output and quality, and enabling open markets for food producers, processors, distributors, and retailers. Given the diversity of food systems, environmental considerations, and regulatory regimes, challenges in developing robust, reliable, integrated computing infrastructure are profound. These intellectual challenges and critical societal need motivate our proposed project.

Project Overview. We propose a comprehensive research, development, outreach, and education program, aimed at bringing disruptive computing innovation to current food systems. Our proposed effort comprises the following five themes: (A) end-to-end provenance tracking, using genomic watermarking, and dense supply chain and cold chain instrumentation; (B) optimizing all aspects of production, processing, and distribution using provenance data, demand tracking, forecasting, and context specific constraints and considerations; (C) incentive mechanisms for best-practices across the food system; (D) creating a blockchain enabled open marketplace for producers, processors, distributors, and retailers; (E) developing software for integrating provenance, field management, weather, and phenotype data to enable phenotypic characterization at massive scale. Beyond this, we plan (i) outreach to producers, food industry professionals, policy-makers, broader research community, and to agencies in developing countries; (ii) a novel education program at the intersection of computing, operations research, supply chain management, and food science; and (iii) a broadening participation in computing (BPC) program, that aims to bring significant new cohorts into computing nationwide.

Theme A: End to End Provenance Tracking (§2.1). We propose an ambitious new approach to food safety, through a-priori, robust and fine-grained genetic watermarking of crops. Our proposed approach uses existing simple sequence repeat (SSR) markers in cultivars to establish proof-of-concept DNA profiling for precise tracking of crops back through the distribution and processing systems to the farm or field where they were produced. SSRs are commonly used for seed genotyping and provide an excellent platform for demonstrating the feasibility of our computational framework. Building on this framework, we propose a number of novel and ambitious DNA profiling schemes that allow rapid and inexpensive detection, traceback, and auditing of food sources. Such provenance information has tremendous impact and value for food safety, quality, distribution efficiency, and phenotyping. The ability to track recombinations and hybridization also provides gold-standard tools for intellectual property protection. While the project primarily deals with crops, the technologies developed could also be used to tag and track animal products and feed.

Theme B: Optimization Techniques for Food Systems (§2.2). There is a critical need for an integrated computing environment that optimizes production and waste in the processing and distribution system. While large vendors optimize for their cost efficiencies, an open system that routes food and food products to minimize total waste and maximize nutrition and freshness, has tremendous potential to revolutionize the industry. To accomplish this, we propose an integrated distribution and processing framework that utilizes available provenance data, supply chain instrumentation, demand forecasting, and parameters associated with food products to formulate stochastic optimization problems and solution techniques. These formulations are designed to work with highly incomplete/ noisy parametrizations, their solutions guaranteed to yield optimal plans (for specified notions of optimality), and along with extensions for dealing with data privacy.

Theme C: Incentive Mechanisms for Participation (§2.3). Designing an effective provenance and supply chain optimization framework does not, in itself, guarantee participation from various entities. While regulation is an alternative, we propose to develop incentive mechanisms that motivate various entities to participate based on suitably designed incentives. The goal is to design algorithms for decentralized systems comprised of nodes (food system entities) operating based on their own incentives. Our provenance and tracking framework provides unique opportunities for formulating powerful new incentive models, with the goal of redesigning the food system to increase efficiency and reduce waste. To this end, we will investigate a number of important and challenging problems – we first study design of dynamic mechanisms, where strategic agents act in a decentralized manner and take (myopic) decisions repeatedly over time, based on partial view of the system. Second, we formulate the problem as one of multi-player learning, an area at the interface of games and learning that has gained significant recent attention. In this context, the design of a suitable information structure regarding supply chain traces is critical (what information to revel, and to whom). Finally, based on our studies, we will design, implement, and evaluate in testbeds, suitable incentive mechanisms for optimizing food system parameters under various real-world scenarios.

Theme D: Creating Open Markets Through Blockchains (§2.4). The ability to track produce through the supply chain has also been recently recognized in industry. As an example, Walmart has teamed up with IBM to explore complete blockchain solutions for auditing their supply chain. While provenance using blockchain has its benefits (and limitations), vendor-specific standalone solutions do not support market efficiency, and more importantly, are vulnerable to cross-vendor counterfeiting. To address this challenge, we aim to develop an open, privacy-preserving marketplace for producers, processors, distributors, and vendors, based on novel adaptations of blockchains. To accomplish this, we must address a number of technical challenges related to validation (defining the secure distributed architecture for validation nodes), access control (understanding privacy needs and achieving those using cryptographic techniques) and data standardization (defining universal ID formats and supply-chain operations) in a blockchain framework. The proposed infrastructure interfaces with users (producers, distributors, and retailers) at the front end, and with the proposed optimization and incentive frameworks at the back end.

Theme E: Scalable Phenotyping Through Large-Scale Data Integration (§2.5). Crop productivity and quality are critical phenotypes associated with agricultural produce. Vast amounts of research and investments in crop genetics and breeding are focused on collecting and analyzing phenotypic data for various plant genotypes, with the goal of identifying genomic regions associated with desirable phenotypic traits. Indeed, this mapping of genotype to phenotype has been recognized as one of the Big Ideas by NSF. The complexity of this problem is rooted in the high dimensionality of the phenotype prediction problem, with genomic variants, weather, field management (irrigation, fertilizer and pesticide use/ schedules), providing important inputs to the problem where only tiny fractions of this space has been explored due to the high cost of experimentation. Our proposed provenance framework provides the opportunity, for the first time, to explore the richness of this space, by overlaying phenotypic data with weather and field management data through our provenance framework. This enables turning virtually any batch of produce into a phenotypic trait experiment, significantly reducing costs and accelerating development through a data driven approach. To enable this framework, we propose a number of innovations – from construction of a scalable extensible

data resource integrating phenotypic, genomic, and field management data, to novel formulations, algorithms, and web-accessible tools for in-silico phenotyping experiments. Issues of noise, missing data, privacy considerations, and scale, present significant intellectual challenges towards this goal.

Education Programs (§3). The unique combination of areas constituting the project presents tremendous potential for innovative curricula and instruction. We will specifically focus on the following: (i) with a view to recruiting high-school students, we will create a weekly seminar program (online and local), exposing students to challenges in real-world food systems, computational modeling and optimization, and use of algorithmic and analytical tools; (ii) at the undergraduate level, we will create a number of Honors courses accessible to students across campus; (iii) at the graduate level, we will create a graduate specialization in Computational Food Sciences (along the lines of the top-ranked Computational Science and Engineering program at Purdue) at the intersection of computer science, operations research, supply chain management and food science; and (iii) at the postdoctoral level, we will create a fellows program, built on best practices gleaned from the NSF Science and Technology Center for Science of Information. Taken in totality, the project will define an entirely new focus area for education, at various levels. We will design a broad curriculum focused on next generation of food safety and distribution infrastructure, detailed instructional material, software infrastructure for enabling a wide variety of classroom (in-silico) experiments integrating diverse data sources and plant phenotype. All of this material will be made available for nation-wide adoption.

Broader Impact Summary (§4). To maximize broader impact, we have integrated into our research program, a broader impact plan that scales from outreach to local producers, distributors, and retailers, to the nationwide community of stakeholders through the proposed CropHub, and to the international community, through targeted outreach to NGOs in India, with the goal of customizing and delivering context-specific solutions (please see Letters of Collaboration). Using the existing and highly successful Purdue Outreach program, we will offer a sequence of workshops and tutorials (both in-person and online) to practitioners. We will invite policymakers to these meetings, with the goal of influencing regulation and law. All of our technical and educational material will be made available on the CropHub. CropHub, relying extensively on Purdue's HubZero technology, has four major platforms – (i) the research platform focusing on datasets (field management data, overlaid weather data, phenotype data), software tools, and research artifacts (publications, talks); (ii) the education platform, focusing on modules and instructional material; (iii) outreach platform, focusing on workshops, tutorials, and meetings aimed at professionals, policy-makers, and NGOs; and (iv) broadening participation platform, for recruitment, mentoring, and assessment of BPC programs. CropHub is built using Purdue's HubZero technology, that allows for rapid development of complex web-based portals.

Broadening Participation in Computing Summary (§4.1). The project includes an ambitional plan for broadening participation to, among others, a cohort that is not traditionally targeted by others. This group, comprised of first-time college attendees and students from rural backgrounds, often arrive on campus with untapped potential for contributions to computing and related disciplines. We build this program around the Purdue University Horizon's program, which is a federally funded TRIO program. As part of this project, we will create a nationwide cohort of eligible students, bring them to campus for summer programs, keep them engaged over the academic year through distance offerings, and provide mentoring on a continual basis. We will also reach out to more traditional cohorts – women and underrepresented groups. In each case, these efforts are aided by the multidisciplinary nature of the project (women are not underrepresented in food sciences) and geographic reach (UC-Riverside is a Hispanic serving institution). We emphasize that the goal is not to increase participation within our institutions, but rather, nationwide. To accomplish this, we will also engage students at conferences such as Tapia, Grace Hopper, and SACNAS. We will create a nationwide network of engaged students and institutions looking to recruit these students. These efforts are deeply rooted in our ongoing programs aimed at broadening participation at the Center for Science of Information. These programs are supplemented by well-defined success metrics, reliance on best-practices, continual assessment, and program evaluation.

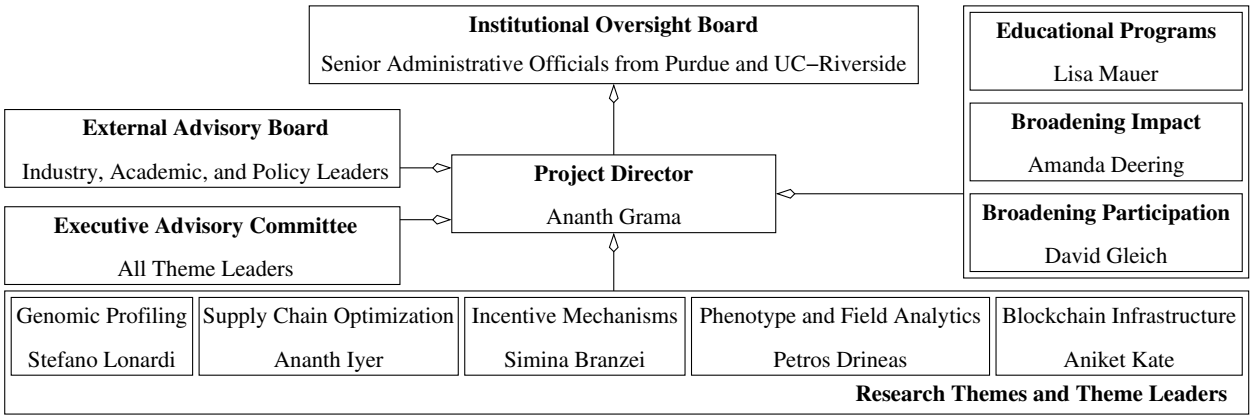

Figure 1: An overview of the project team and organization in terms of themes.

Project Team. To achieve these goals, the project brings together a cohesive and committed group of investigators that includes faculty from Computer Science, Food Science, Agriculture, and Plant Biology. See Figure 1. The team represents a mix of senior investigators with extensive experience in running large successful programs, as well as junior faculty. The project team is integrated through a collaboration plan that ensures timely accomplishment of all project goals, while maximizing real-world impact, worldwide. The project leverages significant existing resources at Purdue, including the Center for Science of Information (CSoI), the Purdue Horizons program for first time college attendees, HubZero program for building user-centric scientific hubs, and CSoIHub providing extensive education and broadening participation resources.

2 INTELLECTUAL MERIT

2.1 Theme A: End-to-End Provenance Through DNA Profiling

Theme Objective. *Our goal is to design and develop novel computational genomic technologies that allow robust and high resolution provenance for produce. They will be able to identify, down to the field, where produce was grown, by profiling its DNA.*

Theme Investigators. Stefano Lonardi (Lead), Wojtek Szpankowski, Katy Rainey.

Theme Motivation. DNA profiling provides a compelling technology for end-to-end provenance tracking. We first demonstrate how existing genomic markers such as Simple Sequence Repeats (SSRs) can be used to construct a provenance system. Having established the feasibility of the concept, we focus on inexpensive and fast technologies based on exogenous markers. This poses a number of interesting and hard computational challenges on design, incorporation, and detection of watermarks, along with tradeoffs between robustness and complexity of the watermark.

Background, current state of the art, and related work. Current approaches to traceability rely on analysis of protein and DNA sequences. While technologies for analyzing proteins are more mature (electrophoresis, chromatography), DNA sequencing provides a more robust and accurate method for tracking raw material, because DNA from plants in animal feed is present and detectable both in animal tissue and human blood [166]. DNA barcoding relies on intrinsic genomic markers (e.g., simple sequence repeats – SSRs or microsatellites, single nucleotide polymorphism, 16S ribosomal RNA in microbes) to identify individuals and/or species. For plants, it specifically uses rbcL and matK chloroplast genes in mitochondrial DNA (mtDNA) [73]. DNA fingerprinting [184, 95], relies on short tandem repeats (STRs) to identify individuals (and species). The recent discovery of horse meat in factory-produced burgers was verified through DNA fingerprinting [143]. Crucially, both barcoding and fingerprinting rely on intrinsic markers and we use the terms interchangeably, although there are domain specific uses this contradicts.

In contrast, DNA watermarking relies on extrinsic features introduced into the genome. It has been used to track growth and ecology of microbial species for bioremediation of groundwater contamination [21].

DNA watermarks are incorporated into the host DNA through biolistics (DNA tethered to a gold or titanium nanoparticles shot into plant tissue), use of agrobacterium (replacing bacterial T-DNA from the bacterial plasmid and with watermarked gene), electroporation (using electric fields to render the cell membrane permeable to plasmid DNA), and CRISPR [37, 64, 112, 188, 6] These techniques are relatively well developed, to the point where over 80% of US production of corn, soybean, sugar beet, and cotton have some form of genetic modifications. Watermarking techniques have been proposed for both coding and non-coding regions. For non-coding regions, comma codes [165], alternating codes [165], and DNA-Crypt [76] have been proposed. Recently, such watermarks have been incorporated into a variety of crop genomes, including rice, tomato, and barley. Their heritability to T1 and T2 generations, with no side effects has also been demonstrated [195, 11, 127, 4].

Research goal: Provenance Using Intrinsic Signatures (Combinatorial Mixing and Deconvolution). A significant fraction of eukaryotic genomes (conservatively, over 50%) is composed of repetitive sequences of varying complexity. Microsatellites (1-10 nucleotides) and minisatellites (more than 10 nucleotides) are subcategories of tandem repeats that, together with the predominant interspersed repeats (or remnants of transposable elements), make up the large majority of these repetitive regions. Microsatellites (also known as Simple Sequence Repeats or SSRs) are arguably most commonly used for defining cultivar DNA fingerprints, and for forensics and parentage analysis in plants [182, 98, 99]. They are informative, co-dominant, multi-allele genetic markers that are reproducible and transferable among related species [182].

As a first step, we propose to use SSRs to genotype specific cultivars. To establish provenance down to individual farms, one would require a large number of cultivars. This is economically infeasible. We rely, instead, on *combinatorial pooling* to achieve specificity. For a given plant for which cultivars are available, we associate a source with a particular mix of these cultivars. Assuming F farms, each farm with C cultivated plant varieties (cultivars), the question is one of choosing fractions of cultivars, and establishing limits on number of farms we can safely distinguish. To design reliable combinatorial pooling, we must separate different fractions of a given cultivar, and distinguish between farms; simple rules of thumb exist to accomplish this objective.

Research goal: Using Extrinsic Signatures (DNA Watermarking). While deconvolution of SSRs across cultivars can be used for provenance, this process is expensive and noisy. We propose to investigate extrinsic watermarks that: (i) induce silent mutations, i.e., no changes to the phenotype; (ii) minimize change to DNA, so that the scheme is practical and cost-effective; (iii) chance of a random occurrence of the watermark is low; and (iv) minimize the impact of recombination and/or cross-pollination. We leverage redundancy in the genetic code to design watermarks. In coding regions, the synonymous translation of multiple codons into the same amino acid provides opportunities for embedding additional bits of information. For instance, the amino acid Serine is coded by UCA (00), UCC (01), UCG (10), and UCU (11). This allows one to embed two bits of the watermark into the codons. Histidine is coded by CAC(0) and CAU(1). This allows one to embed one bit of the watermark in these codons. A two-amino acid group Ser-His can therefore code three bits. A watermark 000 would correspond to the sequence UCA CAC; 101 would correspond to UCG CAU. Such codes can be complemented with suitable error detection codes (checksums, cyclic redundancy codes) to protect the watermark. As a proof of concept, Haider *et al.* used this approach to embedded a 16-bit watermark into the Vam7 gene in *S. cerevisiae* with no functional modification. While in principle synonymous substitutions do not change the primary sequence of a protein, they can induce changes in transcription, splicing, mRNA transport, and translation. To minimize the chances of a phenotypic change, we propose to embed the watermark on plants' pseudogenes. Pseudogenes are non-functional genomic sequences with significant sequence similarity to functional RNA or protein-coding genes. Studies have shown that in plants few pseudogenes have expression evidence, and their expression levels tend to be lower compared with annotated genes [197].

Watermarks embedded into cultivars can be validated at various points through the supply chain. This is done through sparse sampling of the genome, typically through hybridization or shallow sequencing. A

number of computational questions arise in this context, namely: (i) how much data is needed to validate (with high probability) a watermark; (ii) what are the associated statistical models and algorithms for validating the watermark; (iii) in the event of DNA degradation during food processing, can we make these methods tolerant to models of degradation; (iv) can we use models of degradation to optimize the process of watermark design. As an example of the type of analytical questions raised, we focus on its uniqueness property. If we assume that the watermark is embedded as a subsequence in a particular genomic region, we want the probability of observing this watermark by chance to be very low. To avoid such undesirable occurrences we will develop new methods to analyze subsequence occurrences in a (random) string, using prior work that leverages multivariate generating functions over specially designed de Bruijn graphs [62, 93, 172].

Theme Outcomes. For intrinsic fingerprints, we will develop pooling techniques, conduct large-scale studies on feasibility, implement software for signature design and detection, and for SSR-based signature design, based on combinatorial mixing and detection. For extrinsic schemes, we will demonstrate feasibility, design watermarking techniques, and implement software for watermark design and detection.

Risks and Mitigation. The technical risk associated with use of intrinsic markers is relatively low. The question remains whether food producers can be convinced to mix seeds in specific proportions. We address this problem in the Theme on incentive mechanisms. Successful demonstration of our techniques may also influence policy-makers to explore suitable regulatory instruments. For extrinsic watermarks, the primary risk is associated with inducing silent mutations. This is mitigated by the observation that introducing a short DNA oligo at a random position in the genome is highly likely to not induce a phenotypic change.

2.2 Theme B: Supply Network Optimization of Tracking and Tracing Foodborne Contaminants

Theme Objective. *We develop models and methods for optimizing supply chain architecture, location of tracking as well as the optimal response to contamination events, given the network structure and tracking choices. We study overall supply chain response to contaminations and the impact of the availability of tracking and testing solutions on optimal deployment by industry.*

Theme Lead and Researchers. Ananth V Iyer (lead), Paul Preckel, Ken Foster

Theme Motivation. In a network model with nodes corresponding to producers, processors distributors, and retailers, and edges characterized by costs, capacities and flows, the availability of provenance/ traceback data enables construction of models closer to real-world deployments. Individual entities in this model (or groups thereof), have the incentive to ensure reliable food supply, while being cost competitive. Current fresh food supply chains are highly fragmented; consequently, traceback of recent episodes of food contamination is time and effort intensive. This motivates critical questions of: (i) supply chain designs that support rapid and accurate traceback; (ii) location of sensors in supply chain to tradeoff data collection and storage overheads with accuracy of traceback; (iii) uncertainty associated with partial data from supply chain nodes; and (iv) identification of spread of contamination from identified sources. Solutions to these problems would alleviate current state of the art, which involves destruction of all supply in the pipeline (e.g, romaine lettuce) and recommendations to stop consumption of specific foods.

Related Work. Our initial efforts focus on the impact of tracing regulations on flow volumes through various paths in the supply chain [75, 145]. Finding optimal locations of tracking devices, given heterogeneous choices by individual supply chain participants, and the need for the industry to provide guarantees that outbreaks can be detected quickly, can be formulated as a scenario based capacity planning program, and solved as a mathematical programming model [55, 49]. Food supply chains often have two stages, where product is sold by one stage to a successor stage that then divides up the product into marketable pieces [83, 54, 80]. This raises issues of impact of co-mingling of batches versus maintaining batch integrity on the efficiency of tracing. The decision regarding "what is a lot" is highlighted in our recent work [69]. The tradeoffs of the technical solutions to be investigated in these abbreviated supply chains will be between chain performance measures such as revenues and costs, contract efficiency, and efficacy of the tracing system. The goal is to characterize how the cost of tracing is affected by the extent of co-mingling and other channel

parameters [79, 175, 82]. The accuracy and cost of sensed data raises another set of tradeoffs between the size of the unit of observation (i.e., individual items versus batches), the cost of tracing, and the accuracy with which tracebacks can be performed. This strategy has limits, due to the economies of scale at various points along the supply chain. For example, it is unlikely to be cost-effective to keep products from individual farms separate while packing or to segregate packages from individual farms during transportation. The nature of these tradeoffs will also vary depending on the scale of operations of the entity within the supply chain and the sensing/ traceback technology [72, 68, 144, 145].

Proposed supply chain model. Our proposed model can be broadly formulated as:

$$\min \quad h(x,z) + \sum p_i f(x, y_i, z) \quad \text{such that} \quad g_j(x, y_i, z) \geq 0 \text{ for all } i, j$$

Here, vector x refers to the choice of supply chain network structure, z refers to the location of the tracking and testing devices for flows. The function $h(x,z)$ refers to the capacity costs (e.g., fixed amortized cost) associated with the choice of network structure and tracking locations. Index i refers to possible contamination scenarios, and index j refers to the network constraints. The vector y_i refers to the optimal flows under scenario i, given the choice of supply chain network x, and tracking and testing locations z. Note that the capacity choices are made ex–ante, before the contamination scenario unfolds. The probability values p_i refer to the probability associated with contamination scenario i. Note that a scenario can also refer to a change in network structure e.g., imports may be halted or impacted by a change in tariffs, thus changing the flows that are feasible and impacting the associated feasible choices of network flow y_i. Finally the function $f(x, y_i, z)$ refers to the system impact from choice of flows y_i under contamination scenario i, given a choice of capacity x and z. The specific functions chosen reflect: (i) the specific spatial locations of production for the produce; (ii) the nature of the processing steps; and (iii) downstream retail locations, and possible relevant scenarios and associated probabilities.

Finding the optimal location of tracking devices, given heterogeneous choices by individual supply chain participants, and the need for the system to provide a guarantee that outbreaks will be detected quickly is thus a scenario based capacity planning program, solved as a mathematical programming model [55]. The precise nature of the solution can also incorporate expected downside risk to reflect the extent of destruction of both product and brand (taken to mean consumer propensity to consume the specific fresh product), which is determined by the network structure (x), the testing locations (z), and the optimal recovery of flows y_i for scenario i. The specific solutions generated will be tested using our proposed testbed to reflect different product specific supply chain features. A key aspect of this design model is the incorporation of specific tracking and tracing technologies with associated costs and speed of response.

Research Goals. We will first model the optimal amount of product to be quarantined/ destroyed, given an outbreak. This problem is challenging in the presence of optimal choices of the flow during a period, and the need to model the extent of mixing upstream and downstream of these locations. We will create models of prototypical supply chains that are specific to different products. This is used to characterize the nature of the processing steps and the extent of mixing of different inputs. We will investigate two stage supply chains to assess the impact of co-mingling of batches versus maintaining batch integrity, on the efficiency of tracing. The tradeoffs of the technical solutions to be investigated in these abbreviated supply chains will be between chain performance measures such as revenues and costs, contract efficiency, and efficacy of the tracing system. Our modeling approach relies on single and multi-period stochastic optimization as the solution tools.

We will then study tradeoffs of simplicity and cost of instrumentation and accuracy of traceback. At one extreme, each quanta of produce would be traceable to its source distributer, processor, and producer, along with various cohorts encountered along the supply chain. This level of detail would allow very precise elimination of contaminated or infected products at the cost of having to maintain and track information on a very large number of items [80]. However, complications associated with blending, diminish the effectiveness

of such a strategy [142, 176]. At the other end of the spectrum, packages would be tracked in batches of a size that facilitates tracing through the supply chain. Upon the occurrence of an outbreak, all product from the batch could be recalled for testing and/or destruction. The benefit of such an approach could be a simpler tracing system, while the costs come from an increase in the quantity of product destroyed during an outbreak.

Theme Outcomes. We identify the following outcomes from the theme: *(i)* Developing and solving the scenario based planning model while incorporating prototypical supply chain structures and tracking and testing technologies by food type. We aim to develop a testbed that incorporates prototypical supply chain structures. *(ii)* Developing a model of the optimal choice across alternate technology choices at the farm level; i.e., cost of instrumentation vs. expected benefit (reduced waste). *(iii)* Modeling the inventory and pricing impact of better tracing on producer-retailer supply chains. *(iv)* Modeling and evaluating the impact of customer segments with different willingness to pay for tracing, on traceability adoption in supply chains. Specifically, rapid traceability that reduces supply chain disruption could result in more stable prices and increased consumer welfare.

A key challenge is that the results from using a single decision maker that generates *first best* solutions, does not incorporate the economic value generated across the supply chain under competitive conditions [146, 147, 176]. Theme C focuses on mechanisms for sharing the benefits of enforcement, certification, and supply chain management across owners of individual parts of the food supply chain. The goal is to develop implementable schemes that anticipate the market consequences.

Technical Risk and Mitigation. The research team has worked in the past on mathematical programming models as well as contexts within agriculture [22]. While we do anticipate that we can incorporate industry constraints in the testbed, we will use aggregate industry data to simulate alternate structures in the absence of concrete datasets. We also expect to match model inputs to industry margins, tracking, and testing costs. We are confident of the analytic results and required modeling and computational analysis required to ensure near optimal results.

2.3 Theme C: Modeling mechanisms to enable tracking in competitive fresh food supply chains

Theme Objective. *Food supply chains involve independent agents (producers, processors, distributors, retailers), who make independent decisions that endogenize their incentives. Developing suitable mechanisms for incentivizing agents to join the provenance/ traceback framework is essential to a scalable and sustainable food system. We will develop incentive mechanisms, study the information structure associated with these mechanisms, how they can be implemented in real-world supply chains, associated overheads and tradeoffs, and applications of multiplayer learning.*

Theme Investigators. Simina Brânzei (lead), Ananth Iyer, Ken Foster, Paul Preckel

Theme Motivation. Redesigning the food industry to increase efficiency and reduce waste is an important application of game theory in the "wild". It has the potential to boost US economic growth and decrease carbon emissions significantly. Historically, food safety has been treated as a public good in the U.S. For this reason, it has been regulated through policy instruments. The proposed effort takes an alternate view by casting food safety as an incentive for various agents in the food system. In formulating food safety as a mechanism design problem, the project investigates novel mechanisms with significant potential payoffs.

Background and Related Work. Problems in mechanism design for supply chains are commonly formulated as Stackelberg and Cournot games [59, 78, 58, 128]. It has been shown that partial information sharing is a competitive outcome of these games, in specific cases. In recent work [190], we show that there are two period games in which partial demand state information sharing is the equilibrium outcome when there is a learning effect to supply costs. Cournot games with supply traceability provide an effective modeling context for understanding traceability. However, since fresh food products are highly perishable, previous research on food industry consolidation and the economics of food safety suggests that a Bertrand (price competition) framework may be more appropriate for at least some segments of the supply chain [43, 163, 70].

The food traceability setting is a promising application of multi-player learning, an area at the interface of games and learning that has been emerging in recent years [122], with prominent real world scenarios where companies set up "race to the bottom" games that were in their best interest, but which had negative effects on others. Recent work [28] shows that a natural decentralized dynamic, where players update their bids on the goods proportionally to how useful the investments were in the past round, *leads to growth of the economy* in the long term (whenever growth is possible), but also creates growing inequality, i.e. *very rich and very poor players emerge over time*. Additionally, other research studies the cost of outbreaks, paying for traceability, mathematical models and techniques, and externalities [59, 58, 153, 58, 27, 49, 38, 81].

Proposed Research. Our proposed mechanism design is formulated as a combination of models of competition – Cournot competition (where participants choose capacity and prices unfold based on the aggregate capacity seen by the market), Bertrand competition (where prices are the choice variables and capacities unfold accordingly), and the impact of large decentralized systems enabled by digital platforms. The primary role of these investigations is the exploration of choices regarding incentives for adoption and their impact. Given the network structure and the associated tracking and testing locations our model consists of a simple mathematical model. This formulation is flexible enough to handle a variety of structures from: (i) Stackelberg models, where there is a leader (e.g., Walmart) that sets incentives to get followers (distributors or producers to comply); (ii) Cournot competition, where individual producers choose their flows, and prices are determined based on the volume entering the market and associated expected waste during a contamination; and (iii) Bertrand competition, where prices get announced with associated tracking, and volumes are the result of demand curve effects.

While the model above solves for the prices that are the equilibrium outcome, one can also imagine other rules that determine the shared benefit from tracking and tracing. These aspects of tracing systems will be investigated as part of the Theme. One of the issues is the composition of payments in a supply chain, which involves tracing and how it affects individual firm behavior. This investigation focuses on cost allocation (e.g., Shapley value) and uses cooperative gaming solutions to shed insights into the payment question.

When we apply ideas from multi-player learning, it may be optimal to reveal some of the information available but not all of it, such as narrowing the source of a contamination down to a small enough area, but not to the exact producer. In such games with incomplete information, with asymmetric situations where some players don't know what game they play [7], designing the rules so that not everyone knows everything may boost the social welfare in the long run. In recent work [29], we have started exploring the connections between games and learning in the framework of online learning from expert advice, where we provided an essentially optimal learning algorithm for the setting where a decision maker must choose among multiple available options repeatedly over time. We also establish lower bounds by exploiting a connection with zero-sum games.

We will develop a testbed to focus on the fresh food supply chain. The testbed will be structured so as to be extensible to other commodities. The stages of food production to be included in the modeling system begin with seeds, the farm and its management, processing and transportation, associated warehousing and delivery to the retailer. Producers will be permitted to endogenously choose between producing traceable or non-traceable lots with higher tracing lots garnering higher prices (according to market conditions) and non-traceable lots having lower costs (due to reduced needs for bookkeeping and instrumentation). At the end of the supply chain, consumers will endogenously determine the mix of traceable and non-traceable products. This selection will be based on the heterogeneity of consumer preferences for food that has different levels of traceability. We will use this testbed to benchmark our mechanisms. We will also make this testbed available on CropHub for broader community use and experimentation.

One possible challenge is that rewards may appear too far in the future is to be useful. In this case, a virtual currency system circumvents the problem. Here, each player has a budget, a part (but not all) of which could be cashed immediately. Market mechanisms, where agents use a point system to receive allocations have been shown to favor desirable market outcomes (see, e.g. [27]), such as achieving a balance between

fairness and efficiency by maximizing the Nash social welfare.

Theme Outcomes. The theme will result in the following outcomes: *(i)* Establishing the equilibrium level of tracing adoption in competitive supply chains; e.g., non-cooperative competition across supply chains may increase the adoption rate of tracing once the technology for full tracing in a supply chain is feasible. *(ii)* Development of an equilibrium model of benefits, and costs associated with tracing and calibrating using empirical data from food and transport supply chains. *(iii)* Development of solutions with large number of participants to incentivize adoption of tracing by introducing virtual currency payments. *(iv)* Development of rules for sharing under different degrees of asymmetric information that allow participants to understand that social welfare is boosted in the long run, thus incentivizing participation. *(v)* Development of the testbed that will be used to evaluate the costs and benefits of alternative policies.

Risk and Mitigation. While we understand the methodological details, the modeling of the costs of tracking and tracing, as accrued to participants in the food industry context, is a key issue. We will collaborate with industry partners (see letters of support) to gather and prototype this detail. In the absence of primary data, we will rely on estimates and industry experience to guide our models. The solution generation for large problems will also involve balancing the precision of solutions with computational costs. The testbed realism will require us to share details with industry observers to validate whether our representations reflect industry detail.

2.4 Theme D: An Open Privacy-preserving Blockchain Platform for Multi-vendor Provenance

Theme Objective. *By coupling DNA profiling and densely instrumented food supply chains with a distributed ledger (or blockchain) framework, this theme aims to enhance food traceability and to deter counterfeiting. Observing inherent limitations of the current commercial inclinations towards building blockchains inside and at the periphery of organizational boundaries, we aim to design and develop an open multi-vendor blockchain platform offering a powerful trade-off between privacy, traceability, and performance.*

Theme Investigators. Aniket Kate (Lead), Ananth Grama

Theme Motivation. Traditional food supply chains fail to make their distribution processes transparent to involved participants as well as to end consumers. Among other issues, this raises significant concerns relating to food safety and traceability. The provision of a distributed and append-only ledger jointly governed by farmers, processors, distributors, and vendors themselves, by means of a distributed consensus process, makes permissioned blockchains such as Hyperledger Fabric [5] a promising alternative for proactively mitigating counterfeiting and offering end-to-end traceability. This is increasingly recognized in the industry; notably, Walmart has teamed up with IBM to develop complete blockchain solutions for auditing their supply chain [63]; however, vendor-specific standalone solutions cannot account for cross-vendor transactions by players and remain vulnerable to cross-vendor counterfeiting: an insider may exchange a real item by counterfeit, if the identity is not robustly attached to it, e.g., through bar codes. Although approaches such as DNA profiling are not susceptible these attacks, they are expensive to measure extensively throughout the supply chain. This motivates us to conceptualize and develop an open multi-vendor blockchain platform that can integrate different vendors into an open market, while supporting extensive provenance.

Background and Related Work. At its core, a blockchain [139, 5, 189, 31] is a decentralized ledger in which records are append-only and cannot be altered. This allows participants to verify and audit transactions. Although it was initially designed for cryptocurrencies [139], with the appearance elaborate smart contracts on Ethereum [189] and Hyperledger Fabric [5], it has been used to enforce complex business logic (such as supply-chain transfers) in automated fashion without human interaction [63].

A typical supply chain is a series of bilateral contractual links that are put next to each other to form a chain. Traditionally, each of these supply chain links becomes an information flow bottleneck, and reduces the trust in the system as whole. Blockchains have the potential to mitigate these problems through the use of a permissioned ledger and a set of smart contracts, creating an ecosystem in which information flows

openly. Several academic as well as industrial efforts are underway along these directions for different supply-chains from food systems [9] to diamonds [110]. In other projects, the PIs are currently developing complete blockchain solutions for automotive supply-chains [119].

With a significant emphasis on transparency, a distributed ledger inherently negatively affects privacy. Although privacy vulnerabilities and defenses have been extensively explored for cryptocurrencies [20, 12, 158, 77, 109, 130, 126], privacy challenges in blockchain-enabled supply-chains remain unresolved.

Proposed Research. By coupling the densely instrumented supply chains and DNA profiling in the food supply chain with blockchains, this project aims to significantly enhance traceability. Contrary to popular belief, blockchains or distributed ledgers, do not entirely solve the problem of uniquely associating identities to assets, except if the identity is inherent to the products and not physically unclonable, or if the application scenario discourages exchange of real and counterfeit produce. As physical unclonability is not possible (or at least unknown) in some cases, and expensive in other cases, such as DNA profiling, broadening the scope from vendor-specific to multi-vendor systems can significantly enhance traceability and counterfeit resistance. The proposed research aims to develop an open, privacy-preserving platform for mutually distrusting producers, processors, distributors, and vendors. This involves addressing three key challenges. First, in multi-vendor environments, it may not be acceptable to supply-chain processes that their transaction validation logic and business processes are visible across the platform. Defining a secure distributed validation architecture that offers acceptable agreement between processes is a challenge. Second, since most multi-vendor distributed ledger information should not be ubiquitously accessible across the platform and should be protected from competitors, it is necessary to understand privacy needs, and to achieve these using cryptographic and distributed computing techniques. Finally, it is important to standardize identities and operations associated with produce such that identities are not conflicting and operations are not incompatible.

Proposed Platform. We aim to develop the proposed multi-vendor platform by extending Hyperledger Fabric [5], which is an open source permissioned distributed ledger platform designed for use in enterprise contexts. We use it as our starting point, since the Fabric has a highly modular and configurable architecture that supports Turing-complete smart contracts (referred to herein as chaincode), which are stored and executed by endorsing peers who maintain the ledger. The current Hyperledger Fabric fails to correctly address challenges associated with the multi-vendor environment, and we plan to significantly enhance the Fabric as part of this project. In the current Hyperledger Fabric, endorsing peers do not interact with each other directly. This simplistic message flow prevents us from allowing validation logic, requiring participation (in particular, multi-party computation) of two or more endorsing peers. As a first step, as shown in Figure 2, we aim to define and develop the Fabric validation process as a multi-party computation, so that it can execute diverse policies required for a multi-vendor environment. Although Hyperledger Fabric can encrypt transactions payload such that they are only visible on ledgers in a confidential manner, it does not offer any privacy for meta-data. Indeed, achieving strong privacy guarantees entails several challenges regarding data leaks, handling of meta-data, and the communication patterns themselves. Using off-the-shelf privacy techniques such as anonymous communication and oblivious computation can help build stronger privacy protection, at the cost of higher system complexity, and reduced efficiency. We will perform a careful analysis and design necessary building blocks to strike a balance between privacy, traceability, and performance.

The proposed platform development will require understanding and development of the identity format and transaction interfaces, derived from food supply-chain examples. An important technical challenge is the definition of secure identity formats for products having multiple sub-components that do not introduce any false positives or false negatives. Towards achieving our goal of extensive traceability, we will also associate DNA profiles into product descriptors, when available. Finally, towards demonstrating efficiency and scalability of the approach, we will perform empirical evaluations with real food supply-chain data sets.

Theme Outcomes. Our platform forms a distributed source of shared truth for food supply chains, which, along with smart contracts and cryptographic primitives, enables mutually distrusting sets of entities,

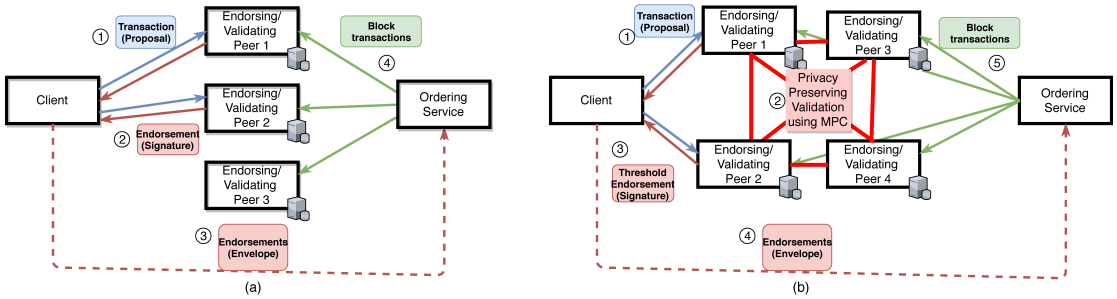

Figure 2: **Transaction Flow for Open Blockchain Platform.** Panel (a) shows message flow for the existing Hyperledger fabric architecture. Panel (b) shows message flow for our proposed open blockchain platform. The key change is here individual node-based validation is replaced by a privacy-preserving validation, implemented as a multi-party computation (MPC) between mutually distrusting endorsing/validating nodes.

with possibly adversarial interests, to collaborate with a secure set of rules. Although specialized for food, our primitives, platform, and findings are broadly applicable to different supply-chain scenarios.

Risk and Mitigation. We note an inherent three-way trade-off between privacy, traceability, and performance. A key risk is that simultaneous strong (meta-data) privacy and traceability demand will result in the need for highly interactive cryptographic protocols, which will result in significant communication overhead. We will investigate suitable techniques for trading off performance to provide tightest possible guarantees, while satisfying application latency requirements.

2.5 Theme E: A Fully Reified Analysis Pipeline

Theme Objective. *The forthcoming decade will see a tremendous growth in the amount of field management, genomic, and phenotypic data collected from a wide variety of sources. We will create a new generation of analytics tools, based on Randomized Numerical Linear Algebra (RandNLA) algorithms and higher-order analysis that will offer an unprecedented ability to understand what impacts the final consumer product (in terms of micronutrients, decomposition, and many more) as food traverses through the food system. This will be realized in a fully reified fashion, which identifies specific data that support the conclusions of the analytics as well as makes it easy to communicate.*

Theme Investigators. Petros Drineas (Lead) and David F. Gleich

Theme Motivation. It is now essentially possible to monitor and track all elements of the food system from seed to store or farm to fork [41] – and the other themes of our proposal fill in a number of missing pieces on how computational approaches are essential to make such as system robust, reliable, and traceable. One of the key benefits of a fully monitored system is that it enables unprecedented analytical insights. Indeed, many subtle data science insights rely on the collection and aggregation of a large volume of data. Long term, aspirational ideas we hope to enable through our analytics research and CropHub systems include critical problems such as studying the impact of pollution on the production of crops. This is important because pollution from China can be measured in much of the agricultural areas on the west coast [187, 115]. As a concrete example, radiation from the Fukishima Daishii nuclear disaster can be detected in Napa Valley wines grown after the incidence [1, 152]. Furthermore, areas in the west such as the Salt Lake Valley frequently experience high-pressure "inversions" that trap pollution and particulate matter for weeks. While the acute impact of pollution on crops is recognized (e.g. [71]), the chronic impact of low to moderate pollution, as experienced in the US, on micronutrient and other is expected to be subtle and hard to pin down without a wide-ranging analysis such as would be enabled by our overall project.

Background. The major challenge that will arise in the data analytic aspect of this proposal is interpretability and explainability of our findings. During the numerous PI meetings that led to the design and formation of this proposal, it became clear that a major bottleneck and disconnect between the communities

that design algorithms for Machine Learning and Data Analysis and practitioners using such algorithms is related to the two concepts of **interpretability** and **explainability**. Interpretability and explainability are broad terms that different communities perceive in different ways. Depending on the situation, they are connected to feature selection or coreset construction; in other applications, they could be related to whether or not the output of an algorithm can be explained in a court of law; and, in still other applications, they are connected to understanding what objectives are optimized by highly complicated ML/DA algorithms, such as deep neural networks. However, having principled data analysis methods that are readily-interpretable and/or explainable in terms of the processes in the domain from which the data are drawn is difficult to accomplish: on the one hand, internet and social media applications provide a major forcing function for much of the Big Data phenomenon, and thus methodologically-inclined students naturally gravitate toward those areas without paying particular attention to the development of interpretable or explainable methods, since they are not necessary in this domain. On the other hand, in many scientific applications (such as the ones that are of interest to this proposal), methodology is secondary, i.e., there is not a strong desire to develop methods per se, and so narrow methods get developed that are interpretable or explainable but are heavily-tailored to that domain and are not portable to or useful in other domains. Making our analytics explainable and reified will pose a number of challenges in the analytics pipelines that we will address via our research.

Understanding the genotype-phenotype relationships is one of the NSF Big Ideas proposed this year. It is also at the center of many analytics efforts by large seed companies such as Monsanto (which was recently acquired by Bayer). Commercial efforts focus on planting schedules and matching seeds to soil and expected weather. Our goal is to build open-tools that enable others to understand these situations retrospectively and build hypotheses for future studies rather than to replicate commercial offerings.

Proposed Research: Interpretable analytics via higher-order analysis and RandNLA We have a wide-ranging set of analytic methods to explore to tackles these issues ranging from simple to complex. On the simple side, we intend to build tools akin to "FarmGrep" that will enable quick lookup of specific field management information given both natural language criteria – as might be offered by a farmer – and quantitative information – as might be generated by an algorithm. On the complex side, we intend to develop completely novel methods to address key questions above.

In past efforts, we have found tremendous utility for higher-order methods in analyzing data from science and society (including social networks, food networks, connectomes, and protein-interaction networks) [17, 16, 15, 193, 191, 192, 53, 133, 137]. In the context of this proposal, we seek to use higher-order and parameterized dimension reduction techniques [40, 39, 167] for hypothesis generation [114, 97] from field management data and phenotypic characterizations of the plant. The goal is to use these advanced analytic techniques to identify new places in the parameter space that have not been explored through existing field management data and breeder experiments. In other past efforts, we have used randomized methods and randomized numerical linear algebra successfully to study population genetics data [23, 35, 8, 25, 24, 36, 151, 141, 51, 150, 177]. In this proposal, we will use randomized techniques to study the relationships between genotype and phenotype using existing field studies as a first data source. This will be used to identify relevant characteristics including weather, fertilizer, soil. Moreover, the goal is to use this type of analysis to understand places in the parameter space that are missing.

We will pay particular attention to the investigations of several complementary approaches to designing improved algorithms that, in addition to coming with strong algorithmic and statistical guarantees, are also are interpretable, in the sense that the output of the algorithms can be understood by domain scientists in terms of the processes generating the data. Our work will proceed along several complementary directions. **First,** we will investigate the concept of interpretability from a feature selection and coreset construction viewpoint. We will focus on understanding aspects of feature selection and coreset construction for classification algorithms (such as Support Vector Machines), as well as for the general problem of preserving distances to affine subspaces (a problem that is closely related to problems such as Principal Components Analysis, regression, k-means clustering, the non-negative matrix factorization, etc.). We will also introduce the concept of *cost* in

feature selection and coreset construction for the aforementioned problems: in our experience, not all features and not all samples have been borne equal and certain features (samples) are preferable to others from a practitioner's perspective. **Second,** we will evaluate the proposed approaches on the data applications that are the focus of this proposal. All research directions have significant intellectual merit and the end product of this project is to transform popular and highly applicable data analysis frameworks to become a global toolbox that can deal with data for which interpretability is an important concern.

Expected outcomes. *(i)* The above ideas will be implemented in the CropHub system to make it easy for others to use our analytics pipelines and to search for similar scenarios. *(ii)* Development of higher-order analysis techniques for field management data; *(iii)* Development of similarity search for field management experiments to enable searching in CropHub; *(iv)* Interpretable analytics for field management data to characterize the impact of various features; *(v)* RandNLA tools for field management data; *(vi)* New experiments to conduct for breeding that explore new aspects of the large multidimensional parameter space. These will all be evaluated in terms of the statistical guarantees, the algorithmic power, and the types of down-stream insights into the crop data they generate.

Risk and Mitigation. The commercial and startup landscape surrounding farm and food-supply analytics is extremely active at the moment. Our goal is to complement rather than to compete with industry, and having our analysis pipelines be open and interpretable is an important difference. Another key risk is access to data to support our analysis. We have secured access to the local Wabash Heartland data [3, 2] – see more in Facilities – which is where we will conduct the majority of our initial studies. As the project progresses, we will seek to address ongoing analytics challenges in the other themes as well.

3 EDUCATIONAL INITIATIVES

We propose a broad set of educational initiatives that target students at Purdue and UCR, initiate novel programs and summer schools for students nationwide, create a complete set of instructional material for broad dissemination, and more broadly, initiate a new multidisciplinary educational program in Computational Food Science, at the intersection of Food Science, Agricultural Economics, and Computer Science. We summarize these targets and outcomes in Table 1 and briefly discuss them here: **(i) Educational Activities for High School Students.** We will initiate two major initiatives in collaboration with area high schools. The first is a series of five lectures on food systems, computing challenges, and software platforms. These lectures will initiate students into the complexities of our current food systems, the challenges in safety, efficiency, and quality, the need for computational solutions, and real-world considerations of regulation and policy. Second, students will be given access to software simulation testbeds, accessible through powerful online portals, that enable them to manipulate elements of the food system and see the results. The goal of this effort is to excite students about everyday challenges, and to recruit them into broad computing disciplines. We will make all material available over the public domain, for other institutions to adopt and build on this program. **(ii) Honors Undergraduate Curriculum.** We will initiate a number of courses at the intersection of computer science, food science, and agricultural economics. In particular, we will create courses on Modeling and Optimization of Food Systems, Incentives and Mechanism Design for Food Systems, Blockchains for Supply Chains, and Introduction to Genomic Barcoding and Watermarking. All of these courses will be offered as honors courses (HONR designation at Purdue) and will be available to all undergraduate majors. The associated coursework will be designed to be accessible to broad audiences, and lab work will leverage the intended multidisciplinary student subscription. We will make all of the material (courseware, online exercises, datasets), available online, and will create an resource for continued development and exchange of instructional material. **(iii) Graduate Program in Computational Food Science.** We will create a new graduate specialization in Computational Food Science. Patterned after the highly successful Computational Science and Engineering and Computational Life Sciences programs at Purdue, these specializations supplement existing degrees. Students are required to take four courses from designated groups of courses, participate in a colloquium, and write a report. As part of the proposed effort,

Level	Goal	Proposed Programs	Assessment Mechanisms
High School	Introduce Concepts in Computational Food Science and motivate students to consider computational majors	Lecture series at local high schools, access to software tools and real datasets for students to experiment with.	Number of students recruited into majors directly related to the project.
Undergraduate	Introduce Core Concepts in Computational Food Science at the Undergraduate Level	Honors Courses on Modeling and Optimization of Food Systems, Incentives and Mechanism Design for Food Systems, Blockchains for Supply Chains, and Introduction to Genomic Barcoding and Watermarking; Honors Projects, Online Material, Technical Modules	Course Participation, Course Outcomes and Evaluations, Access statistics and evaluations of Online Material
Graduate	Create an Intellectual Core at Intersection of Supply Chains, Food Science, and Computer Science	New graduate specialization in Computational Food Systems, new courses, refocusing existing courses, research seminar series. An online space for sharing instructional material.	Course Participation, Broad Adoption of Material, Participation in the Graduate Specialization
Post-Doctoral	Create a unique multidisciplinary skill set in Computational Food Science	Postdoctoral fellowships across two or more disciplines. Joint faculty mentorship. Postdoctoral mentoring and professional development.	Professional success metrics for project postdoctoral fellows.
Professional	Dissemination of project findings and best practices to the broader community.	Releasing lectures, tutorials, software, datasets, and best practices on CropHub. Joint hot topics symposia held by the computational, ag-production, and food-science groups.	Subscription statistics and reviews for CropHub.

Table 1: Educational Program: Cohorts, Activities, and Assessment Mechanisms.

we will: (a) create new graduate courses that will serve as core courses for the specialization; (b) identify existing courses across campus that may serve as electives for the program; and (c) where needed, work with relevant faculty to refocus/ reorient existing courses to provide necessary background for the specialization. **(iv) Multidisciplinary Postdoctoral Fellow.** Building our our experiences from the Center for Science of Information, we will create a Fellows program that will recruit postdoctoral researchers to work jointly with two or more investigators from different disciplines. Our experience shows that this is (a) a particularly effective way of fostering collaboration across disciplines; (b) trains individuals in skills that are highly valued in the community (graduates from this program at CSoI have gone on to faculty positions at premier institutions such as MIT, Caltech, Berkeley, Illinois, CMU, among others); and (c) the strong multidisciplinary orientation of their training sets them up for highly productive research careers. We will broadly advertise and recruit for this program.

4 BROADER IMPACTS

Food systems are of paramount importance to our societal well-being. To this end, the impact of the proposed effort is broad and immense. We propose a number of activities aimed at further maximizing the impact of the project in terms of specific stakeholder constituencies (discussed below). These are evaluated and assessed as discussed below to ensure seamless interactions and to maximize effectiveness.

Project Stakeholders. There are a number of important stakeholders in the food system that are impacted by our project. We intend to involve such stakeholders in various phases of the execution of the project: **(i) Local farm communities.** Purdue has a large and successful outreach program to local farm communities to educate them on best practices in agriculture and handling produce. We will leverage these programs to first educate stakeholders regarding this project and then disseminate project resources (see CrobHub below) to the local farm communities. Members of the PI team have had involvement in the past with such outreach programs (e.g., PIs Deering, Mauer). We would like to emphasize that the interaction with the local farm communities will shape an appropriate subset of project resources to be shared with these

communities (software tools, newsletters, scientific findings, etc.). Given the status of Purdue's outreach programs in the local communities and the state of Indiana more generally, we anticipate that attracting local stakeholders will be a straight-forward task. **(ii) Broader farming, food processing, supply chain, and retail institutions.** We will reach out to professionals nationwide in broad areas of the food system. The goal is to disseminate project insights to the community through the most relevant communication channel – be it seminar, YouTube videos, or farm-based demonstration. Through these interactions, we also hope to better understand unexpected challenges in each aspect of the food supply chain. This part of our outreach activities will build upon our experience from interacting with local farm communities first. **(iii) Policy-makers and NGOs.** We will interface with policy makers (primarily for non-profits) with the goal of guiding policy and regulations. Members of our team have considerable past experience in this. **(iv) International partners.** We aim to reach out to partners in developing countries, in particular, India, with the goal of understanding the suitability of our tools to different environments, as well as to maximize impact of the project, worldwide. PI Deering has extensive experience with international engagement, particularly with developing countries. PI Iyer has ongoing collaborations with food supply chain vendors in India. **(v) Broader research and development community.** Finally, as we have done in prior projects of this scale (e.g., NSF-funded Center for Science of Information), we aim to reach out to the broader community of researchers and practitioners in diverse areas, to create an active community of contributors and beneficiaries.

CropHub. With the goal of reaching the broadest possible set of stakeholders, we will create a Hub for all computational and data resources associated with food systems. Built around Purdue's unique HubZero platform, CropHub is the repository for all technical and educational material, both developed in the project, and contributed by the broader community. CropHub, has four major platforms – *(i)* the research platform focusing on datasets (field management data, overlaid weather data, phenotype data), software tools, and research artifacts (publications, talks); *(ii)* the education platform, focusing on modules and instructional material; *(iii)* outreach platform, focusing on workshops, tutorials, and meetings aimed at professionals, policy-makers, and NGOs; and *(iv)* broadening participation platform, for recruitment, mentoring, and assessment of BPC programs. CropHub employs state of the art web technologies (virtualization, web services, workflows) to enable seamless online execution of complex toolchains, with powerful compute backends, to enable powerful analyses and simulations, for research, education, and development.

Proposed Broader Impact Activities For Engagement With Our Stakeholders. For each stakeholder group identified above, we propose activities to effectively engage with the group. **(i) Purdue Extension for outreach to local farmer/ food processing communities.** Purdue Extension is an organization at Purdue tasked specifically with establishing connections between industry and organizations in the State of Indiana and researchers at Purdue. It provides facilities for, and access to, a large cohort of local farmers. We will organize summer workshops and conferences to showcase computational tools and techniques in support of seed selection, field management, processing markets for produce, and optimal distribution venues. Tools supporting these activities will be made available on CropHub. The project, with computing resources from Purdue, will facilitate online access to these tools to the stakeholders. **(ii) Annual project showcase meeting.** With the goal of reaching out to the broader farming, processing, distribution, and retail communities, we will organize an annual showcase of project developments, available expertise, and key challenges. This will be done through presentations and demonstrations, where stakeholders will be able to engage with project personnel. We will create mechanisms for matching project expertise with interested external partners to foster a vibrant industry-university consortium, with the goal of seamless knowledge and technology transfer to and from the project. **(iii) Policy forum.** With the goal of reaching out to policy-makers and NGOs, we will organize a policy forum, in conjunction with the annual meeting of the Center. We will invite local government officials, officials from agencies such as USDA and FDA, along with other NGOs in Indiana and beyond. The proposed forum specifically focuses on policy implications, guiding regulation, and key identified bottlenecks impacting the efficiency and safety of the food system. **(iv) International scope and interaction.** We have established interactions with a farmers consortium in India. We will build a formal

collaboration with this group, initially through online access to tools and techniques developed as part of the project. We will then customize our tools to specific environmental constrains, in order to enhance our toolchains and to make substantive difference in developing food systems, worldwide.

Beyond these interactions, we have established collaborations with faculty at the University of Agricultural and Horticultural Sciences in Bangalore, India, and with the University of Herat in Herat Province, Afghanistan. As part of this collaboration, faculty were trained in the area of postharvest technology. PI Deering has visited many growing and processing operations throughout Bangalore and Mysore to better understand the challenges and opportunities that growers and processors face in the region. This exchange of information allows for better collaboration in training of GAPs in both the U.S. and India.

In 2018, Universidad Nacional de San Agusta (UNSA) in Arequipa, Peru and Purdue University partnered to create a new research, education, and innovation institute that focuses on key challenges to a sustainable future for the Arequipa region. The Arequipa Nexus Institute for Sustainable Food, Energy, Water and the Environment (The Nexus Institute) was started in March of 2018. PI Deering is part of the Nexus Institute and has been working on assessing water quality for drinking and agricultural activities and level of pesticide residue from locally grown produce, and to determine the prevalence of foodborne pathogens in high value food products. There are several training components in this work to improve the current growing practices in the Arequipa region and to bring awareness to food safety issues. Introducing new technologies that improve food safety and traceability for fresh fruits and vegetables to this region would be extremely valuable.

Assessment Plan. With each proposed activity, we will constitute a corresponding assessment plan. We briefly describe the backbone of such plans, depending on the particular target activity. Starting with workshops and conferences, in addition to attendance, we will track engagement metrics including collaborative projects, internships/ industry experiences for students, and funded research projects from external sources. For international partnerships, we will assess technology transfer and human resource development. This is quantified in terms of deployed technologies, international collaborative projects, and personnel exchanges. For CropHub, we will use access metrics (page views, video views, access to online material, amount of data archived, number of simulations executed). We have extensive experience with assessment of online resources as part of the Center for Science of information (specifically, the Science of Information Hub, soihub). In addition, we will also assess the impact of our online tools by tracking a equivalent H-Index of the Hub. This considers the number of papers that cite CropHub and the number of citations to these papers. These measures were developed by the HubZero team at Purdue and successfully used at Purdue's NanoHub.

4.1 Broadening Participation in Computing

The broadening participation in computing plan has been deeply integrated into our project efforts because we will work with and interact with a number of groups that are not traditionally exposed to computing and computer science. We note that while we have specific programs in order to target various cohorts (Table 2), these programs are not restrictive – our programs are designed to significantly enhance participation from a broad range of under-represented groups.

Broadening Participation: Target Cohorts

We identify the following target cohorts, and design specific programs aimed at recruitment, retention, mentoring, and graduation into professional careers: **(i) Students from Rural Backgrounds.** While a number of urban schools now offer AP Computer Science courses, corresponding numbers are much lower for rural schools. Specifically, in academic year 2017-18, 135,992 high school students took AP Computer Science exams. During the same time, the number of rural students taking AP Computer Science exams was 14,184 (under 11%). For reference, the total number of students in rural, suburban, and urban high schools are 24%, 34%, and 29%, respectively). These numbers clearly demonstrate the significant extent to which students from rural backgrounds are disadvantaged when it comes to selecting computer science or computation-heavy degrees as a major. **(ii) First Time College Attendees.** Challenges for first time college

Cohort	Goals	Activities	Assessment Mechanisms
Students from Rural Backgrounds	Increase number of students from rural backgrounds in computing and related disciplines.	Presentations to local area schools, creation of learning communities, mentoring, and assistance in placement.	Number of students from rural backgrounds at each stage in the pipeline. Impact of program nationwide.
First Time College Attendees	Initiate students to various aspects of computing with food science and agriculture as driving applications.	Recruitment and mentoring, working with the Horizon's program.	Number of students, student success metrics, scaling program nationwide.
Women in Computing	Create a pathway for women students in various project application domains into computing.	collaborations, Grace Hopper, mentoring, and graduation.	Number of students, student success metrics, scaling program nationwide.
Underrepresented Communities	Recruit, mentor, and graduate students from underrepresented groups to programs in computing, supply chains, and food science	Recruitment through Tapia, SACNAS, UCR programs, mentoring, success. Creating a nationwide network.	Number of Students, Student Success Metrics, Scaling Program Nationwide.

Table 2: Cohorts, programs, and assessment mechanisms for broadening participation.

attendees in families are increasingly recognized on college campuses and the lack of computing background further exacerbates them. **(iii) Women.** Women have been traditionally underrepresented in computing disciplines. Our past work at the Center for Science of Information has led to significant institutional infrastructure, understanding of best practices, and a history of success stories to build on including a pipeline of how to find talented female candidates and how to retain them. **(iv) Traditionally Underrepresented Groups.** We focus on Hispanic and African American students from high schools entering undergraduate programs, and from undergraduate programs to graduate school. Beyond this, we will make efforts at identifying and recruiting individuals from these groups as postdocs.

Broadening Participation: Specific Activities

(i) Recruitment Activities. We will develop a number of distinct pathways for attracting first-time college attendees. First, working with the Food Sciences and Agricultural Economics programs, we will recruit students to minor in computer science. Once we establish this channel, we will subsequently create dual majors, tapping into these other majors where students from rural backgrounds are better represented. With respect to first-time college attendees, we will create a special program in collaboration with Purdue's Horizons program at Purdue. As part of this outreach, we will create a weekly seminar series for all Horizon's students. The purpose of this seminar series is to expose students from other majors to exciting opportunities at the intersection of Computer Science and their respective majors. This seminar also helps the students develop support networks composed of students with similar backgrounds and facing similar challenges. Finally, we will work with large area high schools (Harrison, McCutcheon, Jefferson, Montgomery) to present recruitment seminars, and create summer research opportunities for students after their junior year. We have a number of established programs for recruiting women and underrepresented students into our program. These programs include summer schools (where we bring in 40 students each year for two weeks), outreach to meetings such as Grace Hopper, SACNAS, and Tapia, and our network of collaborators at institutions such as Howard and Bryn Mawr Universities. Our project partner, University of California, Riverside, is a Hispanic serving institution, which provides another avenue for recruitment. **(ii) Retention and Mentoring.** Engagement and mentoring are critical aspects of retention. To this end, we will create a learning community for our cohort. The learning community will facilitate interactions among students through research projects, coursework, and a monthly seminar. At the end of each academic year, we will organize a day-long retreat that may include a seminar and poster session, where we will invite external speakers as well as student presentations, or it may involve a field day to visit an interesting food supply location, or other enrichment and

cohort building activities. Each student will be assigned a faculty mentor. Students will be expected to meet with their mentors thrice each semester, and write a short report on their mentor meeting. **(iii) Graduation to Professional Careers.** We will provide extensive professional mentoring, exposing students to various career opportunities. We will invite industry and academic leaders to present to these cohorts. We will use the industry outreach as part of the project to secure internships for students. Professional mentoring will scale from low level tasks such as preparation and review of resumes, mock interviews, and expectations of site visits. With respect to graduate schools, we will assist students in selecting appropriate schools to apply to, securing suitable reference letters, reviewing statements of purpose, and interviews. **(iv) Graduate and Post-Doctoral Programs.** We will create programs that match students and post-doctoral research candidates with professors, reach out through our Center for Science of Information partners to recruit individuals from target groups, and to help these individuals into research positions upon successful completion of their programs. **(v) Scaling to a Nationwide Program.** Our goal is not simply to broaden participation at project sites, but rather, nationwide. To this end, we will create an online network of institutions willing to participate in such a program. This will enable us to scale both the recruitment, as well as placement of students.

Broadening Participation: Continual Assessment Mechanisms and Institutional Resources

For each aspect of the BPC, we define a clear set of assessment metrics (Table 2), and report on them every six months. Annual assessments are made available both to the external advisory committee, as well as on the annual report to NSF. We will engage in a strategic planning effort on initiation of the project. This planning effort establishes annual targets associated with each metric. Our recruitment metrics identify cohorts and programs, sets targets for each cohort, and monitors progress towards these targets. In the even of unsatisfactory progress towards targets, recommendations are made to the executive committee on action items for course correction. With respect to retention and mentoring, we will maintain both programmatic metrics (student engagement events, mentor engagement), as well as student satisfaction, through annual surveys. These surveys will be conducted as part of the annual meetings. We will also monitor professional placement and subsequent professional trajectories of our students. We will create a virtual space for program alumni to report their status, interact with other alumni, and to create an environment of shared success.

We understand and appreciate the need for professional conduct of assessments. To this end, we will rely on the Evaluation and Learning Center of the School of Education at Purdue University to design and conduct all of our assessments. The project leverages significant existing resources at Purdue and University of California, Riverside in support of our BPC program. In particular, we rely on significant infrastructure and cache of best practices from the Center for Science of Information (CSoI). CSoI is a Science and Technology Center of the National Science Foundation (the Center runs through 2020), and has built a network of personnel across the country to support its diversity and outreach mission. It also provides significant dedicated staff to run these programs. CSoI also provides the PIs with a committed set of collaborators at Bryn Mawr and Howard Universities to initiate new programs, before scaling them nationwide. Purdue has also initiated a new program called DataMind, a learning community for students majoring in disciplines related to Data Science. We will use this learning community to provide a fostering environment for our cohorts.

4.2 Intellectual property and ethics

All PIs are committed to respecting privacy, intellectual property, and the highest ethical standard. We will make publicly available all outcomes of the project, including data, software, publications, etc. (see our data management plan for more discussion). None of the project PIs have competing interests to declare. Our program of ethics will leverage current tools at Purdue and UCR, which necessitates that all personnel involved complete training modules on preserving data privacy, responsible code of conduct for research, ethics training, Title IX training, etc. We will follow all standard practices to ensure that all personnel involved in this project will maintain the highest standards when it comes to ethical and responsible research.

References

[1] Fukushima's nuclear signature found in california wine. https://www.technologyreview.com/s/611654/fukushimas-nuclear-signature-found-in-california-wine/, 2018.

[2] Iot for ag testbed at acre. http://whin.org/resources/Poster07_IoT_Ag_Testbed_at_ACRE.pdf, 2018.

[3] Wabash heartland innovation network. http://whin.org/, 2019.

[4] N. A. Abdallah, C. S. Prakash, and A. G. McHughen. Genome editing for crop improvement: Challenges and opportunities. *GM Crops & Food*, 6(4):183–205, 2015. PMID: 26930114.

[5] E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich, S. Muralidharan, C. Murthy, B. Nguyen, M. Sethi, G. Singh, K. Smith, A. Sorniotti, C. Stathakopoulou, M. Vukolić, S. W. Cocco, and J. Yellick. Hyperledger fabric: A distributed operating system for permissioned blockchains. In *EuroSys*, pages 30:1–30:15, 2018.

[6] M. Arita and Y. Ohashi. Secret signatures inside genomic DNA. *Biotechnol Prog*, 20:1605–1607, 2004.

[7] R. J. Aumann and M. Maschler. *Repeated Games with Incomplete Information*. MIT Press, 1995.

[8] H. Avron, C. Boutsidis, S. Toledo, and A. Zouzias. Efficient dimensionality reduction for canonical correlation analysis. *SIAM Journal on Scientific Computing*, 36(5):111–131, 2014.

[9] S. Banker. Blockchain gains traction in the food supply chain, July 2018.

[10] M. Bayati, D. F. Gleich, A. Saberi, and Y. Wang. Message-passing algorithms for sparse network alignment. *ACM Trans. Knowl. Discov. Data*, 7(1):3:1–3:31, Mar. 2013.

[11] M. Begemann, B. Gray, E. January, G. Gordon, Y. He, H. Liu, and M. Oufattole. Precise insertion and guided editing of higher plant genomes using cpf1 crispr nucleases. *Scientific Reports*, 7(11606), 2017.

[12] E. Ben-Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer, and M. Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *IEEE Symposium on Security and Privacy, SP*, pages 459–474, 2014.

[13] A. Benson, D. Gleich, and J. Demmel. Direct tall-and-skinny QR factorizations in MapReduce architectures. In *Big Data, 2013 IEEE International Conference on*, pages 264–272, Oct. 2013.

[14] A. Benson and D. F. Gleich. Computing tensor z-eigenvectors with dynamical systems. *arXiv*, math.NA:1805.00903, 2018.

[15] A. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[16] A. Benson, D. F. Gleich, and L.-H. Lim. The spacey random walk: a stochastic process for higher-order data. *SIAM Review*, 59(2):321–345, May 2017.

[17] A. R. Benson, D. F. Gleich, and J. Leskovec. Tensor spectral clustering for partitioning higher-order network structures. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 118–126, 2015.

[18] A. R. Benson, J. D. Lee, B. Rajwa, and D. F. Gleich. Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrices. In *Proceedings of Neural Information Processing Systems*, pages 945–953, 2014. Selected for Spotlight Presentation.

[19] C. Berger, S. Sodha, R. Shaw, P. Griffin, D. Pink, P. Hand, and G. Frankel. Fresh fruit and vegetables as vehicles for the transmission of human pathogens: Fresh produce as vehicles for transmission of human pathogens. *Environmental microbiology*, 12:2385–97, 09 2010.

[20] A. Biryukov and I. Pustogarov. Bitcoin over Tor isn't a good idea. In *Proceedings of IEEE S&P 2015*, pages 122–134, San Jose, CA, USA, May 2015.

[21] S. Block. DNA barcodes and watermarks. Technical Report JSR-03-305, The MITRE Corporation, June 2004.

[22] D. Boussios, P. V. Preckel, Y. A. Yigezu, P. N. Dixit, S. Akroush, H. C. M'hamed, M. Annabi, A. Aw-Hassan, Y. Shakatreh, O. A. Hadi, A. Al-Abdallat, J. A. E. Enein, and J. Ayad. Modeling producer responses with dynamic programming: a case for adaptive crop management. *Agricultural Economics*, 50(1):101–111, oct 2018.

[23] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-Optimal Column-Based Matrix Reconstruction. In *Proceedings of the 52nd IEEE Symposium on Foundations of Computer Science (FOCS)*, 2011.

[24] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-Optimal Column-Based Matrix Reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

[25] C. Boutsidis and A. Gittens. Improved Matrix Algorithms via the Subsampled Randomized Hadamard Transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3), 2013.

[26] M. Brandl. Fitness of human enteric pathogens on plants and implications for food safety. *Annual Review of Phytopathology*, 44:367–392, 2006.

[27] S. Brânzei, V. Gkatzelis, and R. Mehta. Nash social welfare approximation for strategic agents. *EC*, 2017.

[28] S. Brânzei, R. Mehta, and N. Nisan. Universal growth in production economies. *NIPS*, 2018.

[29] S. Branzei and Y. Peres. Online learning with an almost perfect expert. 2018.

[30] K. Breene. Food security and why it matters, 2016.

[31] R. G. Brown, J. Carlyle, I. Grigg, and M. Hearn. Corda: An introduction. *R3 CEV, August*, 2016.

[32] J. Buzby, H. Farah-Wells, and J. Hyman. The estimated amount, value, and calories of postharvest food losses at the retail and consumer levels in the united states. *SSRN Electronic Journal*, 01 2014.

[33] CDC. Reports of selected e. coli outbreat investigations, 2018.

[34] Y. Choi and W. Szpankowski. Compression of graphical structures: Fundamental limits, algorithms and experiments. *IEEE Trans. on Information Theory*, 58:620–638, 2012.

[35] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. The Fast Cauchy Transform and Faster Robust Linear Regression. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, volume 45, pages 466–477, 2013.

[36] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. The Fast Cauchy Transform and Faster Robust Linear Regression. *SIAM Journal on Computing*, 45(3):763–810, 2016.

[37] C. Clelland, V. Risca, and C. Bancroft. Hiding messages in DNA microdots. *Nature*, 399:533–534, 1999.

[38] R. Clemens. Meat Traceability and Consumer Assurance in Japan . *MATRIC Briefing Paper 03-MBP 5*, September 2003.

[39] P. G. Constantine and D. F. Gleich. Computing active subspaces. *arXiv*, math.NA:1408.0545, 2014.

[40] P. G. Constantine, D. F. Gleich, Y. Hou, and J. Templeton. Model reduction with MapReduce-enabled tall and skinny singular value decomposition. *SIAM J. Sci. Comput.*, 36(5):S166–S191, November 2014.

[41] M. Corporation. Farmbeats. `https://www.microsoft.com/en-us/research/project/farmbeats-iot-agriculture/`, 2019.

[42] M. Coskun, A. Grama, and M. Koyutürk. Indexed fast network proximity querying. *PVLDB*, 11(8):840–852, 2018.

[43] J. Crespi and S. Marette. How should food safety certification be financed? *AMERICAN JOURNAL OF AGRICULTURAL ECONOMICS*, 83(4):852–861, NOV 2001.

[44] F. Critzer and M. Doyle. Microbial ecology of foodborne pathogens associated with produce. *Current opinion in biotechnology*, 21:125–30, 02 2010.

[45] A. Daskin, A. Grama, and S. Kais. Multiple network alignment on quantum computers. *Quantum Information Processing*, 13, 2014.

[46] A. Daskin, A. Grama, and S. Kais. Quantum random state generation with predefined entanglement constraint. *International Journal of Quantum Information*, 2015. (to appear).

[47] A. Daskin, A. Grama, G. Kollias, and S. Kais. Universal programmable quantum circuit schemes to emulate an operator. *Journal of Chemical Physics*, 137, 234112, 2012.

[48] A. Daskin, S. Kais, and A. Grama. A universal quantum circuit scheme for finding complex eigenvalues. *Quantum Information Processing*, 13:333–353, 2014.

[49] D. Dickinson and D. Bailey. Meat traceability: Are U.S. Consumers Willing to Pay for it ? *Journal of Agricultural and Resource Economics*, 27(2):348–364, 2002.

[50] M. Doyle and M. Erickson. Summer meeting 2007 - the problems with fresh produce: an overview. *Journal of Applied Microbiology*, 105:317–330, 02 2008.

[51] P. Drineas, K. Fountoulakis, and A. Kundu. A Randomized Rounding Algorithm for Sparse PCA. pages 1–18, aug 2015.

[52] M. Drmota and W. Szpankowski. Precise minimax redundancy and regrets. *IEEE Trans. Information Theory*, 50:2686–2707, 2004.

[53] N. Eikmeier, A. S. Ramani, and D. F. Gleich. The hyperkron graph model for higher-order features. In *Proceedings of the International Conference on Data Mining (ICDM)*, 2018.

[54] G. Eppen and A. Iyer. Backup agreements in fashion buying - the value of upstream flexibility. *Management Science*, 43(11):1469–1484, November 1997.

[55] G. Eppen, R. Martin, and L. Schrage. A scenario based approach to capacity planning. *Operations Research*, 37(4):517–527, July-August 1989.

[56] M. Erickson, C. Webb, J. Diaz-Perez, S. Phatak, J. Silvoy, L. Davey, and M. Doyle. Infrequent internalization of escherichia coli o157:h7 into field-grown leafy greens. *Journal of Food Protection*, 73:500–506, 3 2010.

[57] FAO. *The future of food and agriculture - Trends and challenges*. Rome, 2007.

[58] M. Ferris, S. Dirkse, J.-H. Jagla, and A. Meeraus. An Extended Mathematical Programming Framework . *Computers and Chemical Engineering*, 33(12):1973–1982, 2009.

[59] M. Ferris and J. Pang. Engineering and economic applications of complementarity problems. *SIAM Review*, 39(4):669–713, 1997.

[60] J. Fill, H. Mahmoud, and W. Szpankowski. On the distribution for the duration of a randomized leader election algorithm. *Annals of Applied Probability*, 6:1260–1283, 1996.

[61] P. Flajolet and W. Szpankowski. Analytic variations on redundancy rates of renewal processes. *IEEE Trans. Information Theory*, 48:2911 – 2921, 2002.

[62] P. Flajolet, W. Szpankowski, and B. Vallee. Hidden word statistics. *Journal of the ACM*, 53:1–37, 2006.

[63] D. Galvin. IBM and Walmart: Blockchain for Food Safety, 2017.

[64] A. Gehani, T. LaBean, and J. Reif. DNA-based cryptography. *Dimacs Series In Discrete Mathematics and Theoretical Computer Science*, 54:233–249, 2000.

[65] A. Ghoshal, J. Zhang, M. A. Roth, K. M. Xia, A. Y. Grama, and S. Chaterji. A distributed classifier for microrna target prediction with validation through TCGA expression data. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 15(4):1037–1051, 2018.

[66] D. F. Gleich. PageRank beyond the web. *SIAM Review*, 57(3):321–363, August 2015.

[67] D. F. Gleich, L.-H. Lim, and Y. Yu. Multilinear PageRank. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1507–1541, 2015.

[68] E. Golan, S. Vogel, P. Frenzen, and K. Ralston. Tracing the costs and benefits of improvements in food sefety. *Agricultural Economic Report*, 791, October 2000.

[69] E. H. Golan, B. Krissoff, F. Kuchler, L. Calvin, K. Nelson, and G. Price. Traceability in the us food supply: economic theory and industry studies. Technical report, United States Department of Agriculture, Economic Research Service, 2004.

[70] J. H. Grant, D. M. Lambert, and K. A. Foster. A seasonal inverse almost ideal demand system for north american fresh tomatoes. *Operations Research Letters*, 58(2):215–234, JUN 2010.

[71] H. Griffiths. Air pollution on agricultural crops, order no. 85-002. Technical report, Ontario Ministry of Agriculture, Food, and Rural Affairs, 2003.

[72] A. C. Group. *Food Safety Management Systems - Costs, Benefits and Alternatives*. May 2002.

[73] C. P. W. Group. A DNA barcode for land plants. *PNAS*, 106(31):12794 – 12797, August 4, 2009.

[74] D. Gunders, J. Bloom, J. B. D. Hoover, A. Spacht, and M. Mourad. Wasted: How america is losing up to 40 percent of its food from farm to fork to landfill, 2017.

[75] M. Harris, P. Kaufman, S. Martinez, and P. C. The US Food Marketing System : Competition, Coordination and Technological Innovations Into the 21st Century. *Agricultural Economic Report*, 811, June 2002.

[76] D. Heider and A. Barnekow. DNA-based watermarks using the DNA-crypt algorithm. *BMC Bioinformatics*, 8(176), February 2007.

[77] E. Heilman, L. Alshenibr, F. Baldimtsi, A. Scafuro, and S. Goldberg. TumbleBit: An untrusted Bitcoin-compatible anonymous payment hub. In *Proceedings of NDSS 2017*, San Diego, CA, USA, February–March 2017.

[78] R. Ivanic, P. Preckel, and Z. Yu. Market Power and Welfare Effects in DC Power Flow Electricity Models with Thermal Line Losses . *Decision Support Systems*, 2003.

[79] A. Iyer and M. Bergen. Quick response in manufacturer-retailer channels. *Management Science*, 43(4):559–570, April 1997.

[80] A. Iyer, V. Deshpande, and Z. Wu. A capacity planning model with demand postponement. *Management Science*, 49(7), July 2003.

[81] A. Iyer and A. Huchzermeier. Smart forecasts for smart customers. *International Commerce Review - The ECR Journal*, 3(1), Spring 2003.

[82] A. Iyer, A. Huchzermeier, and J. Stolle. The supply chain impact of smart customers in a promotion environment. *M & SOM - Manufacturing and Service Operations Management*, 4(2), Fall 2002.

[83] A. Iyer and A. Jain. The logistics impact of a mixture of order streams in a manufacturer-retailer system. *Management Science*, 49(8), August 2003.

[84] P. Jacquet, D. Milioris, and W. Szpankowski. Classification of markov sources through joint string complexity: Theory and experiments. In *ISIT 2013*, pages 2289–2293, Istanbul, 2013.

[85] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden markov process. *Theoretical Computer Science*, 395:203–219, 2008.

[86] P. Jacquet and W. Szpankowski. Autocorrelation on words and its applications. analysis of suffix trees by string-ruler approach. *J. Combinatorial Theory*, 66:237–269, 1994.

[87] P. Jacquet and W. Szpankowski. Asymptotic behavior of the lempel-ziv parsing scheme and digital search trees. *Theoretical Computer Science*, 144:161–197, 1995.

[88] P. Jacquet and W. Szpankowski. Analytical depoissonization and its applications. *Theoretical Computer Science*, 201:1–62, 1998.

[89] P. Jacquet and W. Szpankowski. Markov types and minimax redundancy for markov sources. *IEEE Trans. Information Theory*, 50:1393–1402, 2004.

[90] P. Jacquet and W. Szpankowski. Noisy constrained capacity for bsc. *IEEE Trans. Information Theory*, 56:5412 – 5423, 2010.

[91] P. Jacquet and W. Szpankowski. Joint string complexity for markov sources. In *23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, pages 303–322, Montreal, 2012. AofA'12, DMTCS Proc.

[92] P. Jacquet and W. Szpankowski. On the limiting distribution of lempel ziv'78 redundancy for memoryless sources. *IEEE Trans. Information Theory*, 60:6917–6930, 2014.

[93] P. Jacquet and W. Szpankowski. *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, Cambridge, 2015.

[94] S. Janson and W. Szpankowski. Analysis of an asymmetric leader election algorithm. *Electronic J. of Combinatorics*, 4, 1997. R17.

[95] A. Jeffreys, V. Wilson, and S. Thein. Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314:67 – 73, 1984.

[96] B. Jiang, D. F. Gleich, and M. Gribskov. Differential flux balance analysis of quantitative proteomic data on protein interaction networks. In *Symposium on Signal Processing and Mathematical Modeling of Biological Processes with Applications to Cyber-Physical Systems for Precise Medicine*, GlobalSIP, pages 977–981. IEEE, 2015.

[97] B. Jiang, K. Kloster, D. F. Gleich, and M. Gribskov. AptRank: an adaptive PageRank model for protein function prediction on bi-relational graphs. *Bioinformatics*, 33(12):1829–1836, June 2017.

[98] P. Jonah, L. Bello, O. Lucky, A. Midau, and S. Moruppa. The importance of molecular markers in plant breeding programmes. *Global Journal of Science Frontier Research*, 11(5):4–12, 2011.

[99] R. K. Kalia, M. K. Rai, S. Kalia, R. Singh, and A. Dhawan. Microsatellite markers: an overview of the recent progress in plants. *Euphytica*, 177(3):309–334, 2011.

[100] K. Kambatla, G. Kollias, V. Kumar, and A. Grama. Trends in big data analytics. *J. Parallel Distrib. Comput.*, 74(7):2561–2573, 2014.

[101] K. Kambatla, V. Yarlagadda, I. Goiri, and A. Grama. UBIS: utilization-aware cluster scheduling. In *2018 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2018, Vancouver, BC, Canada, May 21-25, 2018*, pages 358–367, 2018.

[102] X. Kang, D. F. Gleich, A. H. Sameh, and A. Grama. Distributed fault tolerant linear system solvers based on erasure coding. In *37th IEEE International Conference on Distributed Computing Systems, ICDCS 2017, Atlanta, GA, USA, June 5-8, 2017*, pages 2478–2485, 2017.

[103] A. Khan, D. F. Gleich, M. Halappanavar, and A. Pothen. A multicore algorithm for network alignment via approximate matching. In *Proceedings of the 2012 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 64:1–64:11, Los Alamitos, CA, USA, Nov. 2012. IEEE Computer Society Press.

[104] S. G. Kim, N. Theera-Ampornpunt, C. Fang, M. Harwani, A. Grama, and S. Chaterji. Opening up the blackbox: an interpretable deep neural network-based classifier for cell-type specific enhancer predictions. *BMC Systems Biology*, 10(S-2):54, 2016.

[105] K. Kloster and D. F. Gleich. Heat kernel based community detection. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1386–1395, New York, NY, USA, 2014. ACM.

[106] G. Kollias, E. Gallopoulos, and A. Grama. Surfing the network for ranking by multidamping. *IEEE Trans. Knowl. Data Eng.*, 26(9):2323–2336, 2014.

[107] G. Kollias, M. Sathe, S. Mohammadi, and A. Grama. A fast approach to global alignment of protein-protein interaction networks. *BMC Research Notes*, 6:35, 2013.

[108] G. Kollias, M. Sathe, O. Schenk, and A. Grama. Fast parallel algorithms for graph similarity and matching. *Journal of Parallel and Distributed Computing*, 74(5): 2400 - 2410, 2014.

[109] P. Koshy, D. Koshy, and P. D. McDaniel. An analysis of anonymity in Bitcoin using P2P network traffic. In *Proceedings of FC 2014*, volume 8437 of *LNCS*, pages 469–485, Christ Church, Barbados, March 2014.

[110] A. Kovacevic. Blockchain and diamonds: The future of the diamond industry, Sept 2018.

[111] S. B. Kylasa, H. M. Aktulga, and A. Y. Grama. Reactive molecular dynamics on massively parallel heterogeneous architectures. *IEEE Trans. Parallel Distrib. Syst.*, 28(1):202–214, 2017.

[112] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe. Cryptography with DNA binary strands. *BioSystems*, 57:13–22, 2000.

[113] H. Lim. Coordinated capacity investment in clean engine technology under critical material r&d, August 2016. Purdue University.

[114] C.-H. Lin, D. M. Konecki, M. Liu, S. J. Wilson, H. Nassar, A. D. Wilkins, D. F. Gleich, and O. Lichtarge. Multimodal network diffusion predicts future disease–gene–chemical associations. *Bioinformatics*, page bty858, oct 2018.

[115] J. Lin, D. Pan, S. J. Davis, Q. Zhang, K. He, C. Wang, D. G. Streets, D. J. Wuebbles, and D. Guan. China's international trade and air pollution in the united states. *Proceedings of the National Academy of Sciences*, 111(5):1736–1741, jan 2014.

[116] G. Louchard and W. Szpankowski. Average profile and limiting distribution for a phrase size in the lempel-ziv parsing algorithm. *IEEE Trans. Information Theory*, 41:478–488, 1995.

[117] G. Louchard and W. Szpankowski. On the average redundancy rate of the lempel-ziv code. *IEEE Trans. Information Theory*, 43:2–8, 1997.

[118] G. Louchard, W. Szpankowski, and J. Tang. Average profile for the generalized digital search trees and the generalized lempel-ziv algorithms. *SIAM J. Computing*, 28:904–934, 1999.

[119] D. Lu, P. Moreno-Sanchez, A. Zeryihun, S. Bajpayi, S. Yin, K. Feldman, J. Kosofsky, P. Mitra, and A. Kate. Reducing automotive counterfeiting using blockchain: Benefits and challenges. *Under Submission*, 2018.

[120] L. Lu, P. V. Preckel, D. Gotham, and A. L. Liu. An assessment of alternative carbon mitigation policies for achieving the emissions reduction of the clean power plan: Case study for the state of indiana. *Energy Policy*, 96:661–672, sep 2016.

[121] T. Luczak and W. Szpankowski. A suboptimal lossy data compression based in approximate pattern matching. *IEEE Trans. Information Theory*, 43:1439–1451, 1997.

[122] G. Lugosi and A. Mehrabian. Multiplayer bandits without observing collision information. 2018. arXiv.

[123] A. Magner, J. Duda, W. Szpankowski, and A. Grama. Fundamental bounds for sequence reconstruction from nanopore sequencers. *T-MBMC*, 2(1):92–106, 2016.

[124] A. Mahgoub, S. Ganesh, F. Meyer, A. Grama, and S. Chaterji. Suitability of nosql systems - cassandra and scylladb - for iot workloads. In *9th International Conference on Communication Systems and Networks, COMSNETS 2017, Bengaluru, India, January 4-8, 2017*, pages 476–479, 2017.

[125] A. Mahgoub, P. Wood, S. Ganesh, S. Mitra, W. Gerlach, T. Harrison, F. Meyer, A. Grama, S. Bagchi, and S. Chaterji. Rafiki: a middleware for parameter tuning of nosql datastores for dynamic metagenomics workloads. In *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference, Las Vegas, NV, USA, December 11 - 15, 2017*, pages 28–40, 2017.

[126] G. Malavolta, P. Moreno-Sanchez, A. Kate, M. Maffei, and S. Ravi. Concurrency and privacy with payment-channel networks. In *Proceedings of CCS 2017*, pages 455–471, October–November 2017.

[127] A. Malzahn, L. Lowder, and Y. Qi. Plant genome editing with talen and crispr. *Cell & Bioscience*, 7(21), 2017.

[128] E. Maskin. Mechanism design: How to implement social goals. *American Economic Review*, 98(3):567–576, 2008.

[129] T. McMillan. The new face of hunger., 2014.

[130] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage. A fistful of Bitcoins: Characterizing payments among men with no names. *Communications of the ACM*, 59(4):86–93, 2016.

[131] N. Merhav and W. Szpankowski. Average redundancy of the shannon code for markov sources. *IEEE Trans. Information Theory*, 59, 2013.

[132] F. Meyer, S. Bagchi, S. Chaterji, W. Gerlach, A. Grama, T. Harrison, T. Paczian, W. Trimble, and A. Wilke. Mg-rast version 4 – lessons learned from a decade of low-budget ultra-high throughput metagenome analysis. *Briefings in Bioinformatics*, Sept. 2017.

[133] S. Mohammadi, D. F. Gleich, T. G. Kolda, and A. Grama. Triangular alignment (TAME): A tensor-based approach for higher-order network alignment. *Transactions on Computational Biology and Bioinformatics*, 14(6):1446–1458, November 2017. Published online (July 2016) ahead of print.

[134] S. Mohammadi and A. Grama. A convex optimization approach for identification of human tissue-specific interactomes. *Bioinformatics*, 32(12):243–252, 2016.

[135] S. Mohammadi and A. Grama. De novo identification of cell type hierarchy with application to compound marker detection. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2016, Seattle, WA, USA, October 2-5, 2016*, pages 251–260. ACM, 2016.

[136] S. Mohammadi, G. Kollias, and A. Grama. Synthetic genetic interactions and pathway crosstalk. In *Pacific Symposium on Biocomputing (PSB)*, 2012.

[137] S. Mohammadi, V. Ravindra, D. F. Gleich, and A. Grama. A geometric approach to characterize the functional identity of single cells. *Nature Communications*, 9(1):1516, April 2018.

[138] S. Mohammadi, N. S. Zuckerman, A. Goldsmith, and A. Grama. A critical survey of deconvolution methods for separating cell types in complex tissues. *Proceedings of the IEEE*, 105(2):340–366, 2017.

[139] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system," http://bitcoin.org/bitcoin.pdf, 2008.

[140] T. Narayanan, M. Gersten, S. Subramaniam, and A. Grama. Modularity detection in protein-protein interaction networks. *BMC Research Notes*, 4:569, 2011.

[141] N. H. Nguyen, P. Drineas, and T. D. Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference*, 4(3):195–229, sep 2015.

[142] T. Nilsson, P. V. Preckel, B. Ohlmer, and H. Andersson. Revisiting the role of common labeling in a context of asymmetric information: Critique and extensions. *European Regional Science Association 2003 Congress*, 2003.

[143] F. S. A. of Ireland, 2014.

[144] M. Ollinger and N. Ballenger. Weighing incentives for food safety in meat and poultry. *Amber Waves*, April 2003.

[145] M. Ollinger and V. Mueller. The economics of sanitation and process controls in meat and poultry plants. *Agricultural Economic Report*, 817, April 2003.

[146] L. Opara. Traceability in agricultural supply chains: concepts, technological implications and prospects. *International Journal of Food, Agriculture and Environment*, 1(1), 2002.

[147] L. Opara and F. Mazaud. Food traceability from field to plate. *Outlook on Agriculture*, 30(4):239–247, 2001.

[148] J. Painter, R. Hoekstra, T. Ayers, R. Tauxe, C. Braden, F. Angulo, and P. Griffin. Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, united states, 1998–2008. *Emerging Infectious Diseases*, 19:407–415, 3 2013.

[149] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profile of tries. *SIAM J. Comput.*, 38:1821–1880, 2009.

[150] P. Paschou, P. Drineas, E. Yannaki, A. Razou, K. Kanaki, F. Tsetsos, S. S. Padmanabhuni, M. Michalodimitrakis, M. C. Renda, S. Pavlovic, A. Anagnostopoulos, J. a. Stamatoyannopoulos, K. K. Kidd, and G. Stamatoyannopoulos. Maritime route of colonization of Europe. *Proceedings of the National Academy of Sciences of the United States of America*, 111(25):9211–6, jun 2014.

[151] S. Paul, M. Magdon-Ismail, and P. Drineas. Column Selection via Adaptive Sampling. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2015.

[152] M. S. Pravikoff and P. Hubert. Dating of wines with cesium-137: Fukushima's imprint. *https://arxiv.org/abs/1807.04340*, physics.pop-ph:1807.04340, 2018.

[153] P. Preckel. A alternative algorithms for computing economic equilibria. *Mathematical Programming Studies*, 23:163–172, 1985.

[154] N. Rapolu, S. Chakradhar, and A. Grama. Vayu: Accelerating stream processing applications through dynamic network-aware topology re-optimization. *Journal of Parallel Distributed Computing*, 111:13–23, 2018.

[155] N. Rapolu, S. Chakradhar, A. Hassan, and A. Grama. M-lock: Accelerating distributed transactions on key-value stores through dynamic lock localization. In *2013 IEEE Sixth International Conference on Cloud Computing, Santa Clara, CA, USA, June 28 - July 3, 2013*, pages 179–187, 2013. Best Student Paper.

[156] N. Rapolu, K. Kambatla, and S. J. A. Grama. TransMR: Data-centric programming beyond data parallelism. In *3rd Workshop on Hot Topics in Cloud Computing (HotCloud)*, 2011.

[157] ReFed. A roadmap to reduce us food waste by 20 percent, 2016.

[158] T. Ruffing, P. Moreno-Sanchez, and A. Kate. P2P mixing and unlinkable Bitcoin transactions. In *Proceedings of NDSS 2017*, San Diego, CA, USA, February–March 2017.

[159] D. Savel, T. Laframboise, A. Grama, and M. Koyuturk. Suffix-tree based error correction of ngs reads using multiple manifestations of an error. In *ACM International Conference on Bioinformatics, Computational Biology, and Biomedical Informatics (BCB)*, 2013.

[160] D. M. Savel, T. LaFramboise, A. Grama, and M. Koyutürk. Pluribus - exploring the limits of error correction using a suffix tree. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 14(6):1378–1388, 2017.

[161] E. Scallan, R. Hoekstra, F. Angulo, R. Tauxe, M. Widdowson, S. Roy, J. Jones, and P. Griffin. Foodborne illness acquired in the united states-major pathogens. *Emerging Infectious Diseases*, 17:7–15, 1 2011.

[162] G. Seroussi, W. Szpankowski, and M. Weinberger. Deinterleaving finite memory processes via penalized maximum likelihood. *IEEE Trans. Information Theory*, 58:7094 – 7109, 2012.

[163] R. Sexton. Industrialization and consolidation in the us food sector: Implications for competition and welfare. *AMERICAN JOURNAL OF AGRICULTURAL ECONOMICS*, 82(5):1087–1104, NOV 2000.

[164] S. Sivapalasingam, C. Friedman, L. Cohen, and R. Tauxe. Fresh produce: a growing cause of outbreaks of foodborne illness in the united states, 1973 through 1997. *Journal of Food Protection*, 67:2342–2353, 10 2004.

[165] G. C. Smith, C. C. Fiddes, J. P. Hawkins, and J. P. Cox. Some possible codes for encrypting data in DNA. *Biotech. Lett.*, 25(14):1125 – 1130, July 2003.

[166] S. Spisák, N. Solymosi, P. Ittzés, A. Bodor, D. Kondor, G. Vattay, B. K. Barták, F. Sipos, O. Galamb, Z. Tulassay, Z. Szállási, S. Rasmussen, T. Sicheritz-Ponten, S. Brunak, B. Molnár, and I. Csabai. Complete genes may pass from food to human blood. *PLOS ONE*, 8(7):1–11, 07 2013.

[167] K. Stinson, D. F. Gleich, and P. G. Constantine. A randomized algorithm for enumerating zonotope vertices. *arXiv*, math.NA:1602.06620, 2016.

[168] S. Stutzman, B. Weiland, P. Preckel, and M. Wetzstein. Optimal replacement policies for an uncertain rejuvenated asset. *International Journal of Production Economics*, 185:21–33, mar 2017.

[169] W. Szpankowski. Asymptotic properties of data compression and suffix trees. *IEEE Trans. Information Theory*, 39:1647–1659, 1993.

[170] W. Szpankowski. A generalized suffix tree and its (un)expected asymptotic behaviors. *SIAM J. Compt*, 22:1176–1198, 1993.

[171] W. Szpankowski. On asymptotics of certain sums arising in coding theory. *IEEE Trans. Information Theory*, 41:2087–2090, 1995.

[172] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, New York, 2001.

[173] W. Szpankowski and S. Verdu. Minimum expected length of fixed-to-variable lossless compression without prefix constraints. *IEEE Trans. Information Theory*, 57:4017 – 4025, 2011.

[174] W. Szpankowski and M. Weinberger. Minimax pointwise redundancy for memoryless models over large alphabets,. *IEEE Trans. Information Theory*, 58:4094–4104, 2012.

[175] S. Tayur, M. Magazine, and R. Ganesan. *Quantitative Methods in Supply Chain Management, Iyer, A.V., Modeling the Impact of Information on Inventories*. December 1998.

[176] M. Thakur and C. Hurburgh. Framework for Implementing Traceability System in Bulk Grain Supply Chain. *Journal of Food Engineering*, 95:617–626, 2009.

[177] F. Tsetsos, S. S. Padmanabhuni, J. Alexander, I. Karagiannidis, M. Tsifintaris, A. Topaloudi, D. Mantzaris, M. Georgitsi, P. Drineas, and P. Paschou. Meta-analysis of Tourette Syndrome and Attention Deficit Hyperactivity Disorder provides support for a shared genetic basis. *Frontiers in neuroscience Section Child and Adolescent Psychiatry*, 10:340, jul 2016.

[178] UN. Un treaty collection. chapter iv. human rights. 3. international covenant on economic, social and cultural rights. (no. 14531), 1976.

[179] UN. The sustainable development goals 2015-2030, 2015.

[180] A. Vedantam. Essays in sustainable operations, August 2015. Purdue University.

[181] A. Verstak, N. Ramakrishnan, L. Watson, J. He, C. Shaffer, and A. Grama. Using hierarchical data mining to characterize performance of wireless system configurations. *Advances in Engineering Software*, 65:66–77, 2013.

[182] M. L. C. Vieira, L. Santini, A. L. Diniz, and C. d. F. Munhoz. Microsatellite markers: what they mean and why they are so useful. *Genetics and molecular biology*, 39(3):312–328, Jul-Sep 2016.

[183] N. B. Villoria and P. V. Preckel. Gaussian Quadratures vs. Monte Carlo Experiments for Systematic Sensitivity Analysis of Computable General Equilibrium Model Results. *Economics Bulletin*, 37(1):480–487, 2017.

[184] J. Wambaugh. *The Blooding*. Perigord Press Book, New York, 1989.

[185] WHO. Who estimates of the global burden of foodborne diseases: Foodborne disease burden epidemiology reference group 2007-2015, 2015.

[186] A. Wilke, J. Bischof, W. Gerlach, E. Glass, T. Harrison, K. Keegan, T. Paczian, W. Trimble, S. Bagchi, A. Grama, S. Chaterji, , and F. Meyer. The mg-rast metagenomics database and portal in 2015. *Nucleic Acids Research*, 44: 590-594, 2016.

[187] E. Wong. China exports pollution to u.s., study finds. `https://www.nytimes.com/2014/01/21/world/asia/china-also-exports-pollution-to-western-us-study-finds.html`, 2014.

[188] P. Wong, K. Wong, and H. Foote. Organic data memory using the DNA approach. *Communications of the ACM*, 46, 2003.

[189] G. Wood. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*, 151:1–32, 2014.

[190] J. Wu, A. Iyer, and P. Preckel. Information visibility and its impact in a supply chain. *Operations Research Letters*, 44:74–79, July 2016.

[191] T. Wu, A. Benson, and D. F. Gleich. General tensor spectral co-clustering for higher-order data. In *Advances in Neural Information Processing Systems 29*, pages 2559–2567, 2016. http://arxiv.org/abs/1603.00395.

[192] T. Wu and D. F. Gleich. Retrospective higher-order markov processes for user trails. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1185–1194, New York, NY, USA, 2017. ACM.

[193] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 555–564, New York, NY, USA, 2017. ACM.

[194] D. Zhang, H. Zhang, L. Yang, J. Guo, X. Li, and Y. Feng. Simultaneous detection of lsiteria monocytogenes, staphylococcus aureus, salmonella enterica and escherichia coli)157:h7 in food samples using multiplex pcr method. *Journal of Food Safety*, 29(3):348–363, 2009.

[195] C. Zhu, L. Bortesi, C. Baysal, R. M. Twyman, R. Fischer, T. Capell, S. Schillberg, and P. Christou. Characteristics of genome editing mutations in cereal crops. *Trends in Plant Science*, 22(1):38 – 52, 2017.

[196] Y. Zhu, D. F. Gleich, and A. Grama. Erasure coding for fault-oblivious linear system solvers. *SIAM J. Scientific Computing*, 39(1), 2017.

[197] C. Zou, M. D. Lehti-Shiu, F. Thibaud-Nissen, T. Prakash, C. R. Buell, and S.-H. Shiu. Evolutionary and expression signatures of pseudogenes in arabidopsis and rice. *Plant Physiology*, 151(1):3–15, 2009.

Facilities and Equipment

A range of facilities are available to the PIs at Purdue University.

Research Space

The PIs have research space allocated in the Lawson Computer Science Building. The PIs' lab provides a diverse and stimulating research environment. In addition, the PI's have space available for long-term visitors and post-doctoral researchers at the Purdue Center for Science of Information.

Dedicated Computing Facilities

The PIs' lab currently supports several large compute platforms: (ii) a 48-core (4 X Xeon E7-8857 v2 Processor 3.0GHz, 30M Cache, 12 Core) server with 256 GB of RAM and 22 TB of disk, and (ii) thirty quad core Xeon servers, each with 16GB of RAM and 1 TB disk.

- A large shared memory machine with 6TB of RAM and 8 Skylake 24-core, 2.7GHz processors (8168 Xeon Platinum) and 8 NVidia Tesla P100 16GB HBM GPUs. This system also features 12TB of SSD flash that can be accessed at 8GB/second.
- A large disk array with 240TB of space. This is partitioned into 120TB that can be used for experimental purposes and 120TB that will be available as scratch computing space.
- a 160 core (16 X Intel(R) Xeon(R) CPU E7- 8870 Processor, 2.40GHz, 10 Core) with 256 GB of RAM and 100 TB of disk;
- A large shared memory machine with 2TB of RAM and 64 computing cores.
- a 48-core (4 X Xeon E7-8857 v2 Processor 3.0GHz, 30M Cache, 12 Core) server with 256 GB of RAM and 22 TB of disk
- Another large shared memory machine with 512GB of RAM and 32 computing cores.
- A bulk computing machine with 256GB of RAM and 24 computing cores.
- A fast SSD array with 24TB of storage space that can be accessed via 40 GbE by the other computers.
- thirty quad core Xeon servers, each with 16GB of RAM and 1 TB disk

In addition, the lab also has two dual GPU (P100 Nvidia) servers for code development. These provide dedicated development platforms for diverse large-scale applications. In addition, the PI's lab also supports older clusters for software development, comprising over 128 cores in various configurations. These facilities are exclusively available to the PIs and team.

General Computing Facilities at Purdue

The Department of Computer Science at Purdue University is dedicated to providing high-quality computing facilities for use by faculty, students, and administrative personnel. The facilities are operated by a technical staff that is not only responsible for the installation and maintenance of the systems, but also assists faculty and students in the development of software systems for research projects. General computing facilities are available for both administrative activities (such as the preparation of research reports and technical publications) and research needs that are not supported by other dedicated equipment.

I/O Equipment

The department operates both special-purpose I/O devices as well as general output equipment, including more than 100 laser printers, color printers, color scanners, video projectors, digital video editing capabilities, and video conferencing equipment.

Networking Services

The department is strongly committed to state-of-the-art networking technology to provide access to and communication among its systems, as well as to systems elsewhere on campus and throughout the world. A large number of Gigabit Ethernet switches installed in the Computer Science Building connect the workstations to the departmental computing facilities. Experimental wireless networks and production wireless networks are also available. The department supports high-throughput connectivity to other systems on campus, as well as to the Internet community via both "commodity" and Internet2/I-Light connections.

Information Technology at Purdue (ITaP)

In addition to the facilities described above, the PIs and their students have access to computing systems owned and operated by ITaP. General instructional facilities operated by ITaP include large servers and several Intel workstation laboratories. In addition, ITaP provides systems for use in courses taught by the department. Research projects can make use of other facilities provided by ITaP. These include large clusters (Halsted, Rice, Conte, and Coates) with over 10K cores and 256 P100 Nvidia GPUs, smaller SMPs, and commodity clusters that can be set up as per PIs' needs. Extensive visualization facilities are also available at the Envision Center for Data Perceptualization.

Wabash Heartland Project

We also have substantial interactions with the Wabash Heartland Project, see their letter, whose goal is to investigate IoT sensors and other advanced sensing equipment for farms and fields. Our initial analytics experiments will utilize their field management and sensing data. For more information, see [2].

Data Management Plan

The proposed effort critically relies on a range of data collected on field management (planting schedule, pesticide use, fertilizer use, irrigation), sensed data from fields, harvest data, genomic barcodes, transportation and processing data, warehouse, storage and retail data, and consumption data. In addition, the project imports data on weather from third party vendors. Putting all of this data together into a single consistent easy-to-use resource presents interesting challenges, which we will address in this project.

I. Types of Data

The project collects a diverse set of complex data, which must be geographically and temporally co-registered. This data includes:

- Geotagged data on field management and sensed data from fields on soil, moisture, and nutrient content.
- Geotagged data on genomic barcodes.
- Temporal and geotagged weather data at the field-level.
- Data through supply chain – tagged with time, location, and transportation parameters such as temperature and humidity.

In addition, we will publish models of existing and synthetic supply chains, with well defined parameters for optimization experiments.

II. Data and Metadata Standards

All of the aforementioned data will be stored in clearly described XML formats. We will use existing data formats to the extent possible. Where such formats do not exist, we will use clearly described conceptual, logical, and physical data models. We will provide web services interfaces to all of these data repositories for each online querying.

III. Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy

All of the data will be accessible through CropHub and through analysis software executed on CropHub's backend. There will be no charge for accessing any data; however, the PI's reserve rights to use the data before publishing it into the public domain.

There are no ethical issues associated with the data. All data on CropHub can be tagged during submission time as private, public, or anonymized. For private data, we will provide authentication and access controlled environments. For anonymized data, we will suitably de-identify data, or make data available in differentially private databases. We will not store any 'personal data'; the terms of the Data Protection Act 1998 (the DPA or equivalent HIPAA requirement) and IRB Protocols do not apply. The data is not copyrighted and no licenses pertain to it.

IV. Policies and Provisions for Re-use, Re-distribution

We express no restrictions on reuse and redistribution of data, provided the sources are acknowledged. The intended use of the data is by the broader scientific community to validate their own methods, or to use data to develop new methods and applications. We wholeheartedly welcome these efforts.

V. Plans for Archiving and Preservation of Access

All data, software, and research products will be preserved through regular backups at Purdue University. Our use of open formats for all data implies that there is no specific consideration needed for storing metadata.

VI. Software Releases and Maintenance

We will release all software with detailed use cases and documentation as open source over the public domain. Our team is very familiar with version control, release management, and providing helpdesk facilities for our software. Web services will be hosted at Purdue, with analyses software running at servers at Purdue.

Result Dissemination Plan

Disseminating research and educational artifacts through the CropHub is a major activity for the project. We will pay particular attention to the dissemination of the following major results from the project: *(i)* research results in the form of publications and presentations; *(ii)* models, methods, and software; *(iii)* datasets; *(iv)* educational materials; *(v)* outreach platform for external stakeholders; and *(vi)* a platform for broadening participation. All of the aforementioned artifacts will be made available, free of cost, on the CropHub. All software will be released in open source form under a GPL license. This allows third parties to download, modify, use, and (re)release the software (with original attributions). The project team has an extensive track record of such open source software development and release.

Dissemination of Software. All software will be disseminated in the following forms from CropHub.

- In source form: All development will be done through a git repository. All software will be released in source form, both through the git repository, as well as through suitable archives maintained at CropHub. All releases will be clearly indexed, and README files will catalog all release notes. Software will come packaged with clear instructions on installation and use, along with sample inputs and detailed examples.

- In containerized executable form: We will create containerized executables, accessible through CropHub. These containerized executables will enable execution at client sites with no compilation/ setup requirements.

- As web-accessible services: We will make major software packages available as web-accessible services, along with the ability to create customizable workflows, the ability to save intermediate datasets, and to make available complete workflows as third party services to the broader community.

Dissemination of Data. We will release all datasets, (pre)processing scripts, and analysis software for output data on CropHub. We will release all raw execution data from publications. Where possible, we will make our hardware infrastructure available for verification of benchmarks used in our publications (please also see Data Dissemination Plan).

Dissemination of Educational Material. We will fully develop the education portal at CropHub and use it to deliver a wide variety of educational material (from presentations and material targeted to high school students to research presentations). This educational material comprises of slides, presentations, software tools, datasests, and complete hands-on learning modules. We have extensive experience with scalable delivery of educational material to over 10,000 students at a time through the Science of Information Hub (soihub). We are keenly familiar with challenges associated with helpdesk services, scaling hardware backends, and issues of presentation quality and targeting. An important aspect of our CropHub education portal is that we will allow contributions from external stakeholders as well, with the goal of making CropHub the central resource for all such material.

Dissemination of BPC Services. The "Broadening Participation in Computing" portal of the CropHub will serve as a focal point for recruitment of underrepresented groups in computing – both into Center activities, as well as recruitment by external partners of project personnel. To this end, the BPC team will maintain material relating to opportunities within the Center, profiles of project personnel, external partner profiles, and opportunities for external stakeholders to get involved in the BPC effort. The BPC portal will serve as a national resource – as a focal point for broadening participation, nationwide.

Postdoctoral Researcher Mentoring Plan

The partner institutions have extensive infrastructure for successfully mentoring the budgeted postdoctoral researchers in the project. Past postdoctoral researchers in the group have gone on to build highly successful careers in academia as well as industry.

1. Expectations from the postdoctoral researchers will be clearly articulated to them. They will be counseled on various aspects of independent research, interactions within various research groups, productivity (both in terms of publications and software artifacts), and ethical conduct of research.
2. They will be provided career counseling, directed towards various options, building a research program, teaching, and interacting with students at all levels.
3. They will be provided opportunities for participating in grant proposal preparation, writing project reports, and project management.
4. They will be provided opportunities to attend technical meetings and conferences, presenting their results, and collaborating with other researchers as appropriate.
5. They will be given opportunities to teach and mentor students, develop group projects, and to guide student groups through these research projects.
6. They will be given opportunities relating to technology transfer, patenting and licensing, and interacting with industry, as appropriate.
7. Their progress and success will be assessed in terms of their accomplishment of their career goals, and accomplishing project goals in a timely manner.

Results from prior support – Ananth Grama

Ananth's research, over the past five years, has been supported by the following grants:

- Ananth Grama, Saurabh Bagchi, Somali Chaterji, and Folker Meyer, Continued Development and Maintenance of the MG-RAST Metagenomics Pipeline, NIH, $3.8M, 10/15 - 9/20.
- Douglas Lacount and Ananth Grama, Emerging virus-host cell protein interaction networks, NIH, $3.2M, 4/15 - 4/20.
- Mehmet Koyuturk, Ananth Grama, Wojciech Szpankowski, and Shankar Subramaniam, Theoretical Foundations and Software Infrastructure for Biological Network Databases, NIH, $960K, 6/15 - 6/19.
- David Gleich, Ahmed Sameh, and Ananth Grama, Fault Oblivious Computations, DOE, $750K, 8/15 - 8/18.
- Ananth Grama, XPS: EXPL: SDA: Scalable Concurrency Control Techniques for Distributed Systems, NSF, $300K, 9/15 - 9/18.
- Graham Cooks, Ananth Grama, David Thompson, Zoltan Nagi, and Eric Barker, Analytic-Directed Multi-scale Synthesis System, DARPA, $8.3M, 1/16 - 12/19.
- David Gleich, Ananth Grama, and JEnnifer Neville, BIGDATA: F: Models, Algorithms, and Software for Spatial-Relational Networks, NSF, 900K, 9/15 - 8/19.
- Ananth Grama, Software Infrastructure for Online Analytics, NSF, $469K, 8/14 - 7/17.
- Ananth Grama and Markus Lill, Network Analysis Aided Drug Re-purposing for Degenerative Diseases, OVPR Incentive Grant, Purdue University, $398K, 1/14 - 5/16.
- Wojciech Szpankowski, Ananth Grama (and others), Center for Science of Information, NSF Science and Technology Center, $25M, 9/15 - 8/21.
- Priya Vashistha, Rajiv Kalia, Aiichiro Nakano, and Ananth Grama, Probing Complex Dynamics of Small Interfering RNA (siRNA) Transfection by Petascale Simulations and Network Analysis, NSF, $1.9M, 9/11 - 8/16.
- Jan Vitek, Ananth Grama, Douglas Adams, and Suresh Jagannathan, Robust Distributed Wind Power Engineering, NSF, $1.6M, 9/11 - 8/16.
- Suresh Jagannathan and Ananth Grama, Profile-Guided Speculation for Multicore Architectures, Intel Corp., $80K, 9/06 - 12/35.
- Anthony Hosking, Suresh Jagannathan, Jan Vitek, and Ananth Grama, Language and Runtime Support for Safe and Scalable Programs, Microsoft, $329K, 08 -.
- Wojciech Szpankowski, Ananth Grama, and Daisuke Kihara, Information Transfer in Biological Systems, NSF, $480K, 7/08 - 6/15.
- Wojciech Szpankowski, Ananth Grama (and others), Center for Science of Information, NSF Science and Technology Center, $25M, 9/10 - 9/15.
- Eric Jacobsson, Shankar Subramaniam, and Ananth Grama, Hierarchical Modularity in Evolution and Function, NSF, $480K (Purdue's Budget), 10/08 - 9/14.
- Jayathi Murthi, Ananth Grama, Anil Bajaj, Weinong Chen, PRISM: NNSA Center for Prediction of Reliability, Integrity and Survivability of Microsystems, Department of Energy, $21M, 4/08 - 4/14.

These projects have resulted in the following major results:

Systems Infrastructure: Grama's work over the past five years has focused on: (i) development of suitable domain-specific abstractions, along with an API for specification of analytics dataflows; (ii) support for dynamic updates and user-interaction geared towards scalable analytics applications; (iii) development of a runtime system infrastructure for scheduling, resource management, performance, and fault tolerance; (iv) development of a kernel library of online versions of important analytics operations; and (v) validation through exemplar applications. These efforts have resulted in three core sets of results. The first set addresses a number

of challenges associated with resource management and scheduling. The VAYU system for lightweight traffic shaping in analytics applications, the UBIS resource manager/ scheduler, and the RAFIKI runtime configuration manager of cloud storage systems. The second set of results is aimed at algorithms and a kernel library for graph operations. The third set of results focuses on theoretical characterization and performance modeling of the software systems developed. The project has resulted in open source software, along with a number of publications, including [154, 132, 196, 186, 108, 100, 106, 42, 101, 125, 124, 102, 155].

Models and Methods for Analytics: Grama's work on data analytics has focused on models and methods for quantifying, extracting, and manipulating information contained in data, with considerations for space, time, resource constraints, semantics and other application-oriented factors. Specifically, his work has focused on models and methods for analysis of graph structured data. Specific aspects of the work included development of statistical methods, analysis of graph models (for quantification of statistical significance), design of efficient and accurate algorithms (including optimization techniques), implementation on large-scale computing platforms, and validation in the context of social network analysis and systems biology. The project has resulted in software, datasets, and a number of publications, including [46, 45, 108, 100, 106, 48, 181, 107, 47, 140, 159, 155, 136, 156].

Applications in Computational Sciences: Grama's work in Computational Sciences has focused on two major application areas – reactive atomistic modeling of systems, and systems modeling of cellular processes. Work on reactive atomistic modeling has resulted in extensive algorithmic and software development on the ReaxFF force field. The Purdue Reactive Molecular Dynamics (PuReMD) package is used aroudn the world for modeling systems ranging from PETN and RDX to oxidative stress on biomembranes. His work on systems modeling has resulted in models and methods for tissue-specific analysis of human interactomes, disease implicated networks, and single-cell transcriptome analyses. Representative publications from these efforts include [137, 65, 138, 160, 133, 111, 134, 104, 123, 135].

Broader Impact: These research results have significant broad impact on the community. This includes public domain software – the PuReMD molecular dynamics package, Action for analysis of transcriptomes, NSD and TAME for network alignment, UBIS for data center scheduling, VAYU for fine-grained topology reoptimization for stream processing, among others. He was responsible for (co)developing the Center for Science of Information Hub (SoIHub), which is a major community resource hosting research artifacts, educational material, and outreach efforts.

Education and Outreach: As part of these projects, PI Grama has developed extensive instructional material. As Director of the Computational Science and Engineering and Computational Life Sciences programs at Purdue, he significantly redesigned these graduate specializations. He has created a number of online courses and modules. He has organized/ participated in a number of summer schools at the Center for Science of Information, aimed at increasing participation from minority groups, which have yielded significant results at the Center. The material and summaries of outreach programs and results are available at the SoIHub.

Service to Community and Recognition: Recognition from these research results enabled PI Grama to serve on a number of important committees and panels. These include Pres. Obama's Precision Medicine Initiative, Chairing the Biodata Management and Analysis Study Section at NIH, and other ad-hoc sections, including MIDAS (Modeling Infectious Diseases). He was recognized as a Fellow of the American Association for Advancement of Sciences (AAAS) (2015), Distinguished Alumnus of the University of Minnesota (2015), and Outstanding Research Award from the School of Science at Purdue (2017).

Results from Prior Grants – Petros Drineas

Drineas' research, over the past five years, has been supported by the following grants:

- AF Small: Collaborative Research: Practice-friendly theory and algorithms for linear regression problems, NSF, 2018 - 2021, $499,982.
- NSF/DMS FRG: Collaborative Research: Randomization as a Resource for Rapid Prototyping, NSF, 2018-2021, $1,143,223.
- III: Small: Novel Statistical Data Analysis Approaches for Mining Human Genetics Datasets, NSF, 2017-2020, $499,100.
- NSF/BIGDATA Randomized Numerical Linear Algebra (RandNLA) for multi-linear and non-linear data, NSF, 2016 - 2019, $800,000.
- III: Small: Fast and Efficient Algorithms for Matrix Decompositions and Applications to Human Genetics, NSF, 2016-2019, $329,455.

These projects have resulted in a large number of publications on theoretical foundations of Randomized Numerical Linear Algebra (RandNLA), including numerous novel algorithms that leverage randomization and sampling in order to speed up matrix computations, as well as numerous publications on applications of such algorithms to data analysis with a particular focus on population genetics data. As highlights of such publications we mention [23, 35, 8, 25, 24, 36, 151, 141, 51, 150, 177]; a full list of publications with acknowledgements to the respective grants can be found at PI Drineas' web page. These research results have had significant broad impact on the community; we would like to highlight Prof. Drineas' numerous publications in the population genetics community, which has adopted and is using many RandNLA algorithm in data analysis. These projects have also enabled the PI to make significant educational contributions, such as the 2015 Gene Golub SIAM Summer School on RandNLA, which trained over 50 graduate and undergraduate students in RandNLA. Finally, the PI has also made significant contributions to broadening participation of underrepresented groups. These efforts include advising three female PhD students, as well as an African-American student, who has been awarded an NSF GRF.

Results from prior support – Aniket Kate

Aniket's research, over the past five years, has been supported by the following grants:

- SaTC-BSF: CORE: Small: Collaborative: Making Blockchains Scale Privately and Reliably, NSF, 2017-20, lead PI (jointly with Ryan Henry, IU), share $257,669 (of $512,302)
- Ford-Purdue Alliance: Supply Chain Transparency and Control using Blockchain Technology, Ford Motors, 2017-19, sole PI, $278,031
- DUST-BT: Detection of Unauthorized Supply Chain Tampering using Blockchain Technology, Northrop Grumman Cybersecurity Research Consortium (NGCRC), 2016-18, $226,579
- TWC: Small: Practical Assured Big Data Analysis in the Cloud, NSF, 2017-19, sole PI, $439,000

These projects have resulted in the following major results:

The first project, so far, has (i) laid the foundations for privacy in payment channel networks and their routing, by presenting a formal definition as well as practical and provably several secure solutions, (ii) initiated the discussion regarding blockchain access privacy in the form of a position paper, and (iii) investigated the fundamental constraints of anonymous communication necessary for accessing blockchains privately. The project so far has resulted in four top-tier conference papers and an IEEE security and privacy magazine article.

As part of our second and third projects, we have developed an in depth understanding of blockchain's applications to automotive and digital supply-chain respectively. This includes performance and efficiency considerations, privacy issues, and identity formats and requirements.

These research results have had significant broad impact on the community. Our payment channel networks work is not only foundational, but also actively getting considered for real-world use: the Lightning Network group has already incorporated our solutions, which is getting used by thousands of Bitcoin users daily. Our results are available public, actively discussed on blockchain mailing lists and dispersed through blockchain events (e.g., ScalingBitcoin, Stanford Blockchain Conference, and MIT Bitcoin Expo) offering free video access.

These projects have also enabled the PI to make significant educational contributions. In Spring 2019, Aniket is offering a three-credit full-fledged course on Blockchain and Cryptocurrencies, which is attended by 14 graduate and undergraduate students from different colleges at Purdue university. A graduate student funded through the project is joining TU Vienna as a junior faculty, while an undergraduate student funded as NSF REU is joining Cornell for his Ph.D.

Finally, the PI has also made significant contributions to broadening participation of underrepresented groups. These efforts include having a long-term female intern from Bogota, Columbia as the part of Purdue's Undergraduate Research Experience Purdue-Colombia (UREP-C), giving an invited talk on blockchains to Black Graduate Student Association at Purdue University, and delivering a Global Science Leadership Seminar for the College of Science, Purdue University.

Results from Prior Grants – Ananth Iyer

Ananth Iyer's research has been suported by the following grants:

1144843-DGE:IGERT: Global Traineeship in Sustainable Electronics Grant Amount: $ 3,199,979, Period: July 2012 to June 2018

The goal of this IGERT was to create a sustainable electronics program for doctoral students in materials science, mechanical engineering, management, political science, anthropology, industrial engineering and electrical engineering. The impact on their thesis will be to incorporate a model holistic global supply chain view of the impact of materials choices while designing electronics components.
Summary of Results: Doctoral students from Purdue and Tuskegee have been taught content by Purdue and Tuskegee faculty tfor the past six years. The students went on a global tour to India and were hosted by the Indian Institute of Management in Udaipur (IIMU). IIMU coordinated visits to mines, manufacturers, recyclers both nonprofit and forprofit and pollution regulators. Several students from the program have graduated with thesis content in the area of sustainable elctronics
Summary of Broader Impact: All course content was taught to doctoral students from Purdue, and, via webex to students at Tuskegee. Both Purdue and Tuskegee students went on the global trip and worked in teams together. Students thesis content included data from electronics recyclers, hard drive shredders etc.
Publications and other products: These results were presented in [180, 113].

PFI:BIC MAKERPAD: Making everyone a maker through intuitive Shape-Modeling and 3D Printing Service Platform Grant Amount: $1,000,000, Period: September 2016 to June 2019

Our goal is to develop a system that dynamically configures the mix of appropriate technologies to deliver the optimal solution mix consisting of intent and gesture sensing, onject tracking and shape fusion and cloud integration. The approach is aimed to enable seamless creation of 3D designs using hands and everyday objects.
Summary of Results:Advised regarding customized product service operations, traveled to manufacturers to determine capabilities, conducted workshop, brought in 3D printing future suppliers such as UPS, medium companies with 3D printing capabilities, and virtual reality and augmented reality companies to workshop.
Summary of Broader Impact: This is a project in progress. But early discussions with Jay–Randolph Dveelopment Services (JRDS), a nonprofit in Indiana, that works with adults with developmental disabilities, has resulted in initial adoption of tools to improve their business success.

SPR-4228: Developing a Business Ecosystem around Autonomous Vehicles in Indiana Grant Amount: $50,000, Period: October 2017 to April 2019

Our goal is to collect data from industry participants, survey current choices across states and explore ways that INDOT can plan to develop the economic impact of autonomous vehicles and other sensor based innovations. We have provided a first report to the state and are now completing a survey to generate industry perspectives.

SPR-4229: Cost Effectiveness of Constructing Minimal Shelter to Store INDOT Equipment Grant Amount: $50,000, Period: December 2017 to June 2019

We have collected data regarding the benefit of storing snow equipment and trucks in a shelter, developed a spreadsheet based model to estimate the benefit and built a simulation model to provide inputs to the spreadsheet. The benefit has been presented to INDOT and recommendations are being incorporated in a report.

SPR-4203: Synthesis Study: Facilities (Enterprise Development, Sponsorship and Privatization Grant Amount: $50,000, Period: December 2017 to June 2019

Analysis of private partnership initiatives by other state DOTs were collected and possible options presented. We are doing follow up work to respond to INDOT feedback and working on a final report.

SPR-4200: Ohio River Bridges East End Crossing, ORB, Project After Action Review of Procurement Models Grant Amount: $100,000, Period: JUne 2017 to April 2019

The project report summarized the financing and associated toll revenues and penalties for the Ohio River Bridges project which was constructed as a public–private partnership. Sensitivity analysis of the contract parameters and NPV calculations using observed too revenue as well as data gathered from direct interviews with decision makers are part of this report.

Results from Prior Research – W. Szpankowski

PI Szpankowski's research is supported by the following grants:

- **DMS-0800568** ($500K): "Information Flow in Biological Systems", 2008-2014;
- **CCF-0830140** ($300K): " Information Theory of Data Structures", 2008-2014;
- **CCF-0939370** ($48M): "Emerging Frontiers of Science of Information", 2010-2020;
- **CIF-1524312** ($500K): "Towards Structural Information", 2015-2108.

Intellectual merit and publications: Over the last two decades, we have worked on the analysis and design of algorithms and data structures on strings [60], asymptotic properties of data compression [34]–[173], information theory [52]–[171], big data [91, 84, 162] random structures, combinatorial optimization [117, 170], and the asymptotic behavior of large distributed systems [169]. Recently, we estimated the entropy of a hidden Markov model [85] (a long standing open problem), and in [90] we obtained, for the first time an asymptotic expansion of the *noisy constrained capacity* for small noise. Finally, in a series of papers we proposed novel tools for studying redundancy of Markov sources and Markov types (cf. [52, 89, 174]). For example, in [174] we studied the minimax redundancy of universal coding for large alphabets over memoryless sources. In [92] we presented a simplified derivation for the limiting distribution of the redundancy of the Lempel-Ziv'78 algorithm, a result that has been wanting since the algorithm's inception. These results are obtained by analytic techniques such as tree-like generating functions and the saddle point method. Our book [172] *Average Case Analysis of Algorithms on Sequences*, presents a variety of mathematical tools used in the analysis of algorithms and analytic information theory. In our recent new book *Analytic Pattern Matching: From DNA to Twitter* (with P. Jacquet) [93] we focus on analytic methods dealing with pattern matching and their applications to data compression, bioinformatics, and data mining. We also made progress in structural information. In the past, information theory has traditionally dealt with "conventional data," be it textual data, image, or video data. However, repositories of of "unconventional data", including biological, social, medical, web data, and topographical maps, are now commonplace. In compressing such data, one must consider two types of information: the information conveyed by the structure itself, and that conveyed by the data labels implanted in the structure. In [34], we address the former problem by studying information of graphical structures (i.e., unlabeled graphs). In [34], we derived an expression for the structural entropy, and then proposed a two-stage compression algorithm that, for Erdos-Renyi graphs, asymptotically achieves the structural entropy, up to the first two leading terms. This is the first provable (asymptotically) optimal graph compressor. We use combinatorial and analytic techniques such as generating functions, Mellin transform, and Poissonization to establish these findings.

Broader Impact . An essential component of the broad-based research effort of the PI has been the development of an active and thriving interdisciplinary community of students and researchers in broad areas of the science of information. The Center on Science of Information (CSoI) directed by the PI developed a multidisciplinary information sciences track at the undergraduate level and graduate level, developmed of comprehensive courseware, maintained the CSoI web portal `http://soihub.org` for dissemination of projects information, and instructional materials;supported over a 100 students and around 20 post-docs.

Results from Prior Grants: Simina Brânzei

PI Branzei started her first tenure track position at Purdue in 2018 and has not received prior grants.

Results from Prior NSF Support – Stefano Lonardi

Stefano's research, over the past five years, has been supported by the following grants:

1. Establishing the thermotolerant yeast *Kluyveromyces marxianus* as a host for biobased fuels and chemicals production, DOE, 2018–2021, $1,499,999
2. Improving *de novo* Genome Assembly using Optical Maps (IIS-1814359), NSF, 2018–2021, $499,978
3. Algorithms for Genome Assembly of Ultra-deep Sequencing Data (IIS-1526742), NSF, 2015–2019, $499,000
4. Advancing the Cowpea Genome for Food Security (IOS-1543963), NSF, 2015–2019, $1,587,345
5. Feed the Future Innovation Lab: Advanced Tools for Climate-Resilient Cowpeas, US-AID, 2013–2019, $4,972,542
6. Acquisition of a Big Data Compute Cluster for Interdisciplinary Research (DBI-1429826), NSF, 2014–2015, $548,476
7. Algorithms and Software Tools for Epigenetics Research (IIS-1302134), NSF, 2013–2017, $994,370
8. Acquisition of a Scalable Storage Cluster for Data Intensive NIH Research, NIH, 2014–2015, $592,816

These projects have resulted in the following major results:

Algorithms for Genome Assembly of Ultra-deep Sequencing Data. The goal of this project was to investigate the problem of *de novo* genome assembly under the assumption of ultra-deep (i.e., $>$1000x) sequencing data. We showed in Mirebrahim et al. (*Bioinformatics*, 2015) that when the depth of sequencing increases over a certain threshold, sequencing errors make the genome assembly problem harder and harder, and as a consequence the quality of the solution degrades with more and more data. In Mirebrahim et al. (*Bioinformatics/ISMB*, 2016) we presented a possible solution that uses a novel meta-assembler that partitions the ultra-deep input data into optimal-sized "slices" and uses a standard assembly tool (e.g., Velvet, SPAdes, IDBA, Ray) to assemble each slice individually. Assemblies produced for each slice are merged using a majority voting approach. Experimental results clearly demonstrate the efficacy of this divide-and-conquer approach. In Alhakami et al. (*Genome Biology*, 2017), we analyzed several assembly reconciliation tools proposed in the literature. The strengths and weaknesses of these tools have never been compared on a common dataset. We filled this need with our work, in which we reported on an extensive comparative evaluation. The surprising finding was that none of the tools we tested consistently improved the quality of the input assemblies.

Improving *de novo* Genome Assembly using Optical Maps. The objective of this project is to develop innovative algorithmic solutions for automatically and accurately improve de novo genome assembly. Specifically, we want to provide user-friendly software tools to enable users to enhance assembly contiguity and reduce assembly errors (e.g., mis-joins) using optical maps. We investigated how to take advantage of one or more optical maps i) to accurately detect and split chimeric contigs and chimeric molecules (see Pan et al., *Bioinformatics*, 2018), ii) to accurately create scaffolded genome assemblies (see Pan et al., *Proceedings of RECOMB*, 2019), iii) to accurately stitch multiple (redundant) genome assemblies (see Pan et al., *Bioinformatics/ISMB*, 2018.)

Barcoding-Free Multiplexing: Leveraging Combinatorial Pooling for High-Throughput Sequencing. In this project we investigated a new sequencing protocol for hierarchical (i.e., BAC-by-BAC) genome sequencing and assembly of large eukaryotic genomes. At the core of the protocol is the ability to solve a set of hard computational questions, which are the focus of the research plan. The sequencing protocol was published in Lonardi et al, (*PLoS Computational Biology*, 2013). The protocol was applied on a BAC library for barley (*Hordeum vulgare*) and cowpea (*Vigna unguiculata*). Our work on these barley BAC clones was included (among other resources) in the paper appeared in the journal *Nature* on the first draft of the barley genome (Stein et al., *Nature*, 2012).

These research projects have had significant broad impact on the community.

Release of the barley genome (Algorithms for Genome Assembly of Ultra-deep Sequencing Data). We have used the techniques developed this project to help with the BAC assembly of the barley genome and the cowpea genome (see below for cowpea). The map-based reference genome sequence of barley was constructed by the International Barley Genome Sequencing Consortium using hierarchical shotgun sequencing, as reported in Mascher et al. (*Nature*, 2017). In Beier et al. (*Scientific Data*, 2017), we reported the experimental and computational procedures used in Mascher et al. to i) sequence and assemble more than eighty thousand bacterial artificial chromosome (BAC) clones along the minimum tiling path of a genome-wide physical map, ii) find and validate overlaps between adjacent BACs, iii) construct non-redundant sequence scaffolds representing clusters of overlapping BACs, and iv) order and orient these BAC clusters along the seven barley chromosomes using positional information provided by dense genetic maps, an optical map and chromosome conformation capture sequencing (Hi-C).

Release of the cowpea genome (Advancing the Cowpea Genome for Food Security – Feed the Future Innovation Lab: Advanced Tools for Climate-Resilient Cowpeas). In Munoz-Amatriain et al., (*The Plant Journal*, 2016), we described a set of fundamental resources developed from a cowpea African cultivar that include i) bacterial artificial chromosome (BAC) libraries, ii) a BAC-based physical map, iii) assembled sequences from over four thousand BACs, and iv) whole-genome shotgun (WGS) assembly. In Huynh et al., (*The Plant Journal*, 2018) we developed a MAGIC population for cowpea from eight founder parents. These founders were genetically diverse and carried many abiotic and biotic stress resistance, seed quality and agronomic traits relevant to cowpea improvement in the United States and sub-Saharan Africa, where cowpea is vitally important in the human diet and local economies. The eight parents were inter-crossed using structured matings to ensure that the population would have balanced representation from each parent, followed by single-seed descent, resulting in 305 F8 recombinant inbred lines each carrying a mosaic of genome blocks contributed by all founders. We have recently completed the sequencing of the cowpea genome using PacBio and Bionano, and released the genome in the public domain at Phytozome (Lonardi et al., bioRxiv, 2019).

Training of graduate students (All projects). Funding for these project provided opportunities for training for the following PhD students: Hamid Mirebrahim (currently Principal Scientist at Roche Sequencing), Hind Alhakami (currently Bioinformatics Scientist at Dovetail Genomics), Rachid Ounit (currently CTO of Biotia a start-up which focuses on metagenomics analysis in clinical settings), Anton Polishko (currently CTO at Digibuild Software), Denise Duma (currently post-doc at Icahn School of Medicine at Mount Sinai, New York), Weihua Pan (4^{th} year PhD student), Abbas Ardakany (4^{th} year PhD student), Md Abid Hasan (4^{th} year PhD student), Qihua Liang (3^{th} year PhD student) Dipankar Ranjan Baisya (2^{nd} year PhD student). Undergrad students Alan Venegas (hispanic) and Matthew Goldberg (now PhD student at UMD) were also involved in the project.

Results from Prior Grants: Katherine Rainey

PI Rainey's research has been supported by the following grants over the past five years:

- Clarifying The Genetic Architecture Of Components Of Yield With Soynam, Smithbucklin Corp, 03/01/2012 - 02/28/2014, $272,680.
- Groundwork For Improving Nutritional Value Of Indiana Soybeans, Indiana Soybean Alliance, 05/01/2012 - 04/15/2015, $46,324.
- Program Support For Soybean Breeder And Geneticist, Indiana Soybean Alliance, 05/09/2012 - 12/31/2017, $200,000.
- Nested Association Mapping To Identify Yield QTL In Diverse High Yielding Elite Soybean Lines Continued Evaluation, Univ Of Illinois At Champaign-Urbana, 04/12/2012 - 03/31/2014, $84,000.
- Exploring Soybean Yield Potential Through Modification Of Plant Architecture, United Soybean Board, 03/01/2014 - 12/31/2015, $385,021.
- High-Impact Public Research For Modified Carbohydrate Composition In U.S. Soybeans, United Soybean Board, 10/01/2013 - 09/30/2016, $3,225,169.
- Characterization Of Soynam Population For Node And Pod Number, Dow Agrosciences, 01/01/2014 - 12/31/2015, $507,733.
- Characterization And Enhancement Of Soybean Genetic Resources For Soilborne Disease Resistance, University Of Minnesota, 05/01/2014 - 09/30/2015, $265,198.
- Deciphering The Molecular Basis Of Soybean Stem Growth Habit, National Inst Of Food & Agriculture, 01/01/2015 - 12/31/2018, $491,962.
- Acceleration Of Soybean Yield And Composition Improvement Through Genomic Selection, University Of Illinois, 05/01/2014 - 09/30/2015, $106,924.
- Initiation Of A Genomic Selection Pipeline For Public Soybean Breeders In The North Central Region, University Of Minnesota, 10/01/2015 - 03/29/2017, $62,160.
- Improving Efficiency Of Soybean Breeding With Drone-Based Canopy Measurements, Indiana Soybean Alliance, 04/01/2016 - 03/31/2017, $55,983.
- Acceleration Of Soybean Yield And Composition Improvement Through Genomic Selection, University Of Illinois, 10/01/2015 - 09/30/2016, $42,631.
- Increasing The Rate Of Genetic Gain For Yield In Soybean Breeding Programs, Ohio State University, 10/01/2016 - 11/30/2017, $402,472.
- A Public-Private Partnership To Use Drone-Acquired Metrics To Increase Accuracy Of Yield Estimation In Multi-Environment Yield Trials Of Soybeans, Indiana Soybean Alliance, 04/01/2017 - 06/28/2018, $66,423.
- Increasing The Rate Of Genetic Gain For Yield In Soybean Breeding Programs, Ohio State University, 10/01/2017 - 11/30/2018, $181,695.
- Characterization Of The High Stearic Acid Soybean Oil Trait, Agricultural Research Service, 01/01/2015 - 08/31/2019, $55,000.
- Modifying Soluble Carbohydrates In Soybean Seed For Enhanced Nutritional Energy Meal, United Soybean Board, 10/01/2016 - 09/30/2019, $2,439,127.
- Development Of Pipeline & Tools For Drone-Based Canopy Phenotyping In Crop Breeding, National Inst Of Food & Agriculture, 02/01/2017 - 01/31/2020, $703,000.
- A Public-Private Partnership To Use Drone-Acquired Metrics To Increase Accuracy Of Yield Estimation In Multi-Environment Yield Trials Of Soybeans -Yr2, Indiana Soybean Alliance, 04/01/2018 - 03/31/2019, $73,788.
- I-Corp Market Research For Uas Imaging Of Agricultural Fields, National Science Foundation,

04/01/2018 - 03/31/2019, $50,000.

- Increasing The Rate Of Genetic Gain For Yield In Soybean Breeding Programs, Ohio State University, 10/01/2018 - 08/31/2019, $169,597.
- Acquisition Of Goods And Services, Agricultural Research Service, 09/01/2018 - 08/31/2019, $10,000.
- Pilot Project With Nils United Soybean Board 10/01/2017 - 09/30/2018 $75,000
- Modifying Soluble Carbohydrates In Soybean Seed For Enhanced Nutritional Energy Meal - Year 2 United Soybean Board 10/01/2017 - 09/30/2018 $842,882
- Development And Implementation Of Genome-Wide Analysis Tools For Function-Valued Traits Describing Crop Growth National Inst Of Food & Agriculture 09/01/2019 - 08/31/2021 $333,665
- Exploiting Fall-Planted Barley To Increase Agricultural Productivity And Improve Resource Use Efficiency In Sustainable Cropping Systems University Of Minnesota 06/01/2019 - 05/31/2024 $951,463
- Enabling Quantification And Prediction Of Soybean Maturity And Adaptation To Production Systems Indiana Soybean Alliance 04/01/2019 - 03/31/2021 $150,969

PI Rainey's work as part of these projects focuses on genetic improvement, yield improvement, and quality improvement. Specifically, the projects have resulted in the following important results:

i) Dr. Rainey's work as part of these projects focuses on genetic improvement of soybeans for increased yield and better quality using multidisciplinary approaches. Her work integrating diverse sources of information to demonstrate new approaches to soybean breeding. Working with geneticists, agronomists, economists, engineers, and other soybean breeders in the public and private sectors, Dr. Rainey has been breeding soybeans for over 10 years and has released specialty cultivars. Dr. Rainey's cultivar Glenn had the highest yields in the 2008 Virginia variety trial, higher than cultivars from Monsanto and other private companies. In 2009 - 2010, certified seed of Glenn was produced on 1,100 acres in 5 states, which generated enough seed to plant 10,000 acres for commercial production.

ii) For yield improvement, her work focuses on dissection of yield into components traits in productive environments. This includes goals to describe new traits associated with yield that can measured using new technologies such as precision and high-throughput phenotyping. She has also explored how to predict yield and maximize gain from selection in soybeans using the components of phenotypic variance, high density markers, and mixed models for analyses.

iii) For quality improvement, Dr. Rainey has worked on modifying the commodity paradigm in soybean to create new markets and grow value across the entire value chain. An example of the application of this to her research is using poultry feeding studies to describe metabolizable energy of soybean meal as a new trait that can be phenotyped.

Education and Outreach Dr. Rainey has contributed to a review of the Agronomy Department Plant Genetics, Breeding, and Biotechnology curriculum. Dr. Rainey has developed new content for a number of courses as part of these efforts. Notable experiential, group, and flipped learning that Dr. Rainey has developed includes: (i) curricula to help students learn basic genetic terminology and techniques before class; (ii) assignments and rubrics for group discussion based on diverse reading; (iii) field-based lab agronomy classes where students collected real data illustrating the concept of yield component compensation; (iv) an assessment based on the Facebook game Spore Islands, where students developed and shared virtual environments and evolved creatures adapted to them to learn breeding principles; and (v) a greenhouse lab to observe examples of major crop plants and their wild relatives, such as maize and teosinte, illustrating the major morphological changes that can occur when one gene mutates.

By way of outreach, she has released cultivars for use by farmers, and she has released improved germplasm for use by other breeders, and she contributes to cooperative multi-environment yield trials for the development of improved germplasm and genetic analyses. Many of the multidisciplinary aspects of her research are focused on economic outcomes.

Results from prior support - David F. Gleich

Gleich's research, over the past five years, has been supported by the following grants.

- *ADMM for Power-Grid Optimization*
  Agency MISO, Amount $73,600, Dates 2016-07-01 – 2018-06-30
- *Sloan Research Fellowship Award*
  Agency Sloan Foundation, Amount $55,000, Dates 2016-09-01 – 2018-08-30
- *Swept time-space domain decomposition rule for breaking the latency barrier*
  Agency NASA, Amount $690,891 , Dates 2015-09-01 — 2018-08-31
- ***CCF-BIGDATA-Medium*** *Models, algorithms, and software for spatial-relational networks*
  Agency NSF, Amount $900,000, Dates 2015-09-01 — 2019-08-31
- *Erasure coded computations*
  Agency DOE, Amount $275,000 33% of total $750,000, Dates 2015-08-01 — 2018-07-31
- ***STC*** *Science of Information – Seed grant*
  Agency NSF, Amount Approx $200k, Dates 2015-09-15 — 2019-09-14 (estimated)
- ***SIMPLEX*** *Multimodal Networks*
  Agency DARPA, Amount $392,322.94, Dates 2015-05-22 — 05-31-2018
- ***IIS-III-Small*** *Spectral clustering with tensors*
  Agency NSF, Amount $339,546, Dates 2014-08-01 — 2019-08-01
- ***Career*** *Modern Numerical Matrix Methods for Network and Graph Computations*
  Agency NSF, Amount $499,586 , Dates 2012-05-01 — 2019-04-22
- *Engineering Data Science Algorithms*
  Agency Purdue Data Science Initiative, Amount $234k, Dates 2018-08 – 2020-05

These projects have resulted in the following major results.

Higher-order analysis. Higher-order and multiway correlations are necessary to identify important structures in complex data. We discuss how this arises in the problem of community detection [15] and how this enables a new class of methods with exceptional performance in the detailed descriptions below. Methods that utilize multilinear correlations and attributes of those data typically utilize three-way or higher relationships, which can be represented as a tensor or hypermatrix. A common problem with such methods is that they typically result in NP-complete problems. Gleich developed a new type of stochastic process that applies to data with higher-order information that, in some cases, is polynomial time computable [67, 16]. This stochastic process is a computationally tractable analogue of random walk methods that are at the heart of many existing tools for scientific data analysis. In fact, it enables an interesting class of generalizations of spectral clustering directly on these stochastic processes [17, 191]. In another line of work, we have new results when using higher-order versions of the network alignment problem to produce accurate alignments of protein-protein interaction networks; these produce the most accurate alignments using the simplest algorithms [133]. This work has opened a number of novel avenues to study higher-order data and tensors including novel methods to compute Z-eigenvectors [14].

Bioinformatics. In concert with a number of collaborators, we have developed a suite of routines that address key challenges in single-cell (scRNAseq) data analysis including novel characterizations of the functional identities of cells [137]. We also have state of the art hypothesis generation mechanisms that posit novel drug-repurposing efforts [114].

Erasure coded high-performance algorithms. In concert with Co-PI Grama, we have developed state of the art fault-tolerant computations for distributed systems and any systems where faults are likely [196, 102].

Broader impacts. The mathematical methods and software tools Gleich has written are commonly used to analyze data from the brain, posit new functions of genes in biology, minimize jet engine noise, and study communities or clusters in proprietary industry datasets and social networks [103, 13, 10, 18, 105, 96]. Developers at Twitter, Ebay, and Facebook have implemented Gleich's procedures inside their internal libraries. For instance, the Scalding library developed at Ebay and Twitter and the Spark Machine Learning library use Gleich's tall-and-skinny QR factorization routine.

Education and Outreach. Gleich has thousands of views of his slides on the website slideshare. He has organized multiple tutorial presentations at international conferences, including a recent tutorial on higher-order and tensor methods for the SIAM Applied Linear Algebra Meeting in Hong Kong.

Service and Recognition. Gleich served as the co-chair of the SIAM Annual Meeting in 2016, which was the largest SIAM Conference at the time with over 1600 attendees. He was also awarded a Sloan Research Fellowship. More recently, he received a SIAM Outstanding Paper Prize for his work describing PageRank [66].

Results from prior support – Lisa Mauer

Lisa Mauer's research, over the past five years, has been supported by the following grants:

- Advanced development of innovative technologies and systematic approaches to foodborne hazard detection and characterization for food safety. USDA-ARS. 4/2016-4/2021. $7,600,000.
- The effects of sugars and salts on starch retrogradation. USDA-NIFA. 5/2018-4/2020. $165,000.
- Improving thiamin (Vitamin B1) delivery in foods by understanding its physical and chemical stability in natural form and enriched products. USDA-AFRI. 1/2016-12/2019. $427,025
- Crystallization inhibition in food components. Nestle. 8/2015-7/2017. $209,103.
- Foods for health. USDA-AFRI Higher Education. 8/2012-7/2016. $229,500.
- Effects of formulation and unit operations on the ingredient architecture, finished product texture, and stability of cookies. PepsiCo. 8/2015-7/2016. $54,282.
- Effects of water and other ingredients on starch thermal properties in wheat flour. 1/2016-1/2016. $15,000.
- Improvement of potato French fry quality and reduction of added ingredients by using and manipulating endogenous pectin/starch. McCain Foods Limited. 1/2015-12/2016. $66,642.
- Fundamentals and consequences of water-solid interactions: Investigating vitamin C stability in crystalline-amorphous blends. Institute of Food Technologists Marcel Loncin Research Prize. 5/2014-4/2016. $50,000.
- The science behind molecular gastronomy techniques: Food chemistry fundamentals meet experimental cuisine. Purdue University Provost Instructional Equipment Program. 8/2013-7/2014. $16,028.
- Improved detection techniques for foodborne pathogens. USDA-ARS. 1/2011-12/2015. $7,000,000.
- Crystallization characterization for sugar replacement considerations in cookie and sugar syrup products. PepsiCo. 8/2015-7/2015. $26,027.
- Enhanced delivery of phytochemicals by nanodispersion in polysaccharide matrices. USDA-NIFA. 8/2009 – 7/2014. $445,000.
- Use of non-wheat cereal proteins as functional viscoelastic polymers. 8/2009 – 7/2014. USDA-NIFA. $370,890.

These projects have resulted in the following major results:

i) The ability to rapidly detect pathogens in complex food products is an ongoing industry and regulatory need, with different detection technologies often needed for different food products and different types of pathogenic organisms. Dr. Mauer's team has developed novel technologies for the detection and analysis of harmful foodborne bacteria, yeasts, and molds ranging from spark induced breakdown spectroscopy, lateral flow immunoassays with magnetic capture and net fishing enhancements, and bioluminescent bacteriophages, to elastic light scattering with image analysis and smartphone based assays. The technologies are being evaluated by the USDA agriculture research service (ARS) and food safety inspection service (FSIS) agencies.

ii) Thiamine (vitamin B1) deficiencies are problematic in both developed and developing countries. Dr. Mauer has developed seven new salt forms of the vitamin, with enhanced delivery and stability traits compared to the two commercially available ingredient salt forms (hydrochloride and mononitrate). Her work has also identified key driving factors in the vitamin phase transformations and stability in foods, enhancing the understanding of the degradation pathways and barriers for improving delivery of thiamine in foods and dietary supplements.

iii) In an international effort to reduce the amount of sugar added to low moisture baked products (such as cookies), high moisture baked products (such as cakes), and sweetener syrups/binding agents (such as those that hold 'granola' bar components together), Dr. Mauer has characterized the effects of numerous sweeteners and oligosaccharides on starch functionality (gelatinization, pasting, and retrogradation). This work led to her development of the governing theory by which sweeteners alter starch functionality (intermolecular hydrogen bonding).

iv) It is essential to control the physical state of sucrose in many food products. From the combined results of a series of studies that investigated the effects of a wide variety of molecular structures and properties on the physical state of sucrose, Dr. Mauer developed a hierarchical scheme for how the different ingredient structures/properties influence sucrose functionality and phase transformation.

Broader Impacts. These research results have had significant broad impact on the community. The USDA is using the pathogen detection technologies developed by Dr. Mauer's team to improve the inspection and safety of foods, and companies have licensed 4 of the technologies. Over the past 5 years we have had 10 material transfer agreements, 23 invention disclosures, 17 patent applications filed, and 4 commercial licenses executed.

The food industry is using the recommendations developed by Dr. Mauer to: 1) reduce the amount of added sugar to food products while maintaining gold standard texture; 2) improve the stability and delivery of thiamine in foods (a patent application has been filed); and 3) control the crystallization of sucrose (a patent has been granted).

Education. These projects have also enabled the PI to make significant educational contributions. Dr. Mauer is known for mentoring students in and out of the classroom and laboratory. She teaches numerous classes ranging from food chemistry, food packaging, and ingredient technology to the science of experimental cuisine, all of which incorporate experiential learning activities. She has included numerous students from diverse backgrounds in undergraduate and graduate research projects and product development competitions.

Broadening Participation. Finally, the PI has also made significant contributions to broadening participation of underrepresented groups. These efforts include mentoring women undergraduate and graduate students in research projects (17 of the 28 graduate students who have matriculated from Dr. Mauer's lab were female; and 61 of 70 undergraduate students who have completed research projects in Dr. Mauers' lab were female); completing the first African American female Ph.D. student in the food science discipline (in 2015), and numerous K-12 outreach activities related to food safety and food science such as 'professors in the classroom' around the state of Indiana and hosting booths at Springfest at Purdue University.

Results from prior support – Paul Preckel

Paul's research over the past five years has been supported by the following grants:

1. "Continuing the Work of the State Utility Forecasting Group III," Indiana Utility Regulatory Commission, July 2013-June 2015, $1,100,000

2. "Building a Bio-economic Farm Household Model for Simulating the Impacts of Socio-economic, Policy, Technological and Climate Changes in Vulnerable Areas of Jordan," International Center for Agricultural Research in Dry Areas (ICARDA), October 2013-September 2015, $162,000

3. "Independent Energy and Peak Demand Forecasts to the Midcontinent System Operator (MISO)," Midcontinent Independent Systems Operator, December 2014-November 2017, $1,527,055

4. "Continuing the Work of the State Utility Forecasting Group IV," Indiana Utility Regulatory Commission, July 2015-June 2017, $1,300,000

5. "The role of international trade in adapting US agriculture to increased global climate variability," USDA/NIFA October 2016-September 2018, $120,000

6. "Opportunities for Agriculture and Tourism in the Orinoquía Region of Colombia," Government of Colombia, November 2016-December 2017, $1,000,000

7. "Coal-power plants rejuvenated with biomass: An economic, social, and environmental sustainable transition to clean power," USDA/NIFA, October 2017-September 2019, $492,099

8. "Independent Energy and Peak Demand Forecasts Update to the Midcontinent System Operator (MISO)," Midcontinent Independent Systems Operator, December 2017-November 2018, $208,148

9. "Continuing the Work of the State Utility Forecasting Group V," Indiana Utility Regulatory Commission, July 2017-June 2019, $1,300,000

10. "Annual Independent Load Forecasts for the Midcontinent Independent System Operator (MISO)," Midcontinent Independent System Operator, December 2018-November 2022, $2,586,437

These projects have produced the following results:

Grants 1, 4 and 9 have provided essential input to the Indiana Utility Regulatory Commission in the form of a series of 20-year forecasts of electricity demand, pricing and generating resources needs. This information serves as input to decisions regarding approval of utility proposals for new generating capacity. Other studies provide information regarding energy policy. This work resulted in the publication of one refereed journal article, 7 technical reports, and one Ph.D. thesis [120].

Grants 3, 8 and 10 have provided essential input to the Midcontinent Independent System Operator regarding decisions related to electricity network transmission infrastructure investments in the form of long-term forecasts of electricity demand and pricing for a swath of states that cuts through the middle of the United States from Louisiana to Minnesota and further north into Manitoba. This work has resulted in one refereed journal article (accepted, in press), three technical reports and two Ph.D. theses.

Grant 7 examines alternatives for extending the life of existing coal-fired electricity generating plants while reducing greenhouse gas emissions by co-firing with wood pellets. The principle focus is on determining the optimal timing of investments in retro-fitting of these plants for co-firing and/or replacing them. One refereed journal article has been published from this research [168].

Grant 5 explores the role of international trade for reducing the price impacts of weather-based agricultural production shocks. One journal article has resulted from this research [183].

Grant 2 determines sustainable crop-livestock production strategies in arid dryland farming for the country of Jordan. One journal article and one Ph.D. thesis have resulted from this research [22].

These research projects have resulted in contributions to educational programs by providing timely, relevant examples of analyses that are incorporated in Paul's course offerings at the graduate level.

While not directly related to these projects, Paul is currently supervising the M.S. project of a female, African-American who came to Purdue after completing a B.S. degree at Alcorn State University.

Results from prior support – Kenneth Foster

Kenneth Foster's research, over the past five years, has been supported by the following grants:

- Water Supply for Developing Countries: Community Scale Water Treatment System for Las Canas, DR, Dr. Scholls Foundation, 2014-15, $2500

These projects have resulted in the following major results:

A community scale water treatment system was designed, constructed, and implemented at the local school in Las Canas, Dominican Republic. The system was designed by Purdue students, co-constructed by Purdue students and community members, and is operated by the community.

Results from prior support – Amanda Deering

Amanda Deering's research, over the past five years, has been supported by the following grants:

- Amanda Deering, Scott Monroe. State and Territory Cooperative Agreement to Enhance Produce Safety in Preparation of Implementation of FDA's Rule: Standards for the Growing, Harvesting, Packing, & Holding of Produce for Human Consumption, FDA, $3.6M, 9/16-8/21.
- Peter Hirst, Amanda Deering, Ariana Torres, Klein Ileleji. Appropriate Postharvest Handling, Processing, and Marketing of Dried Apricots in Tajikistan, USAID, $599,000 2/17-1/19.
- Amanda Deering, Scott Monroe. Development of a Food Safety Training Center in Collaboration with Vincennes University, FDA, $467,903, 4/18-4/19.
- Amanda Deering, Haley Oliver. Cultivating Shared Solutions for Locally and Responsibly Sourced Cereals in Nigeria, Cultivating New Frontiers in Agriculture, $108,000, 4/17-7/19.
- Michael Ladisch, Amanda Deering, Eduardo Ximenes. Physical method for concentrating Salmonella from tomato samples to detectable levels using automated microfiltration, FDA, $320,000, 8/15-7/16.
- Scott Monroe, Amanda Deering, Intification of Food Safety Best Practice for Indiana Cantaloupe Production, USDA Specialty Crop Block Grant, $105,000, 1/17-12/19.
- Klein Ileleji, Ariana Torres, Amanda Deering. A Collaborative Study on the Feasibility of Value-added Solar Drying of Specialty Crops for Small Growers in Georgia and Indiana, USDA-NIFA, $499,617, 2/17-1/2020.
- Amanda Deering, Robert Pruitt, Cathie Aime. Bacterial and Fungal Communities Associated with Fresh Produce, Center For Food Safety Engineering, $350,000, 4/16-4/19.
- James Krogmeier, Dennis Buckmaster, Aaron Ault, Amanda Deering. An Open Source Framework and Community for Sharing Data and Algorithms, Foundation for Food and Agricultural Research, $936,699, 1/18-1/21.
- Suranjan Panigrahi, Amanda Deering. A systems-based management practices for enhancing quality and safety of organic produce: Planning Grant, $50,000, USDA-NIFA, 8/15-7/16.
- Amanda Deering, Haley Oliver. Assessment of Water Quality for Drinking and Agricultural Activities and Level of Pesticide Residue from Locally Grown Produce, $1.5M, UNSA NEXUS – Peru, 1/19-12/21.

These projects have resulted in the following major results:

(i) Dr. Deering's research program addresses both basic and applied research questions. The basic research done in her lab focuses on understanding the movement and internalization of human pathogenic bacteria in plants. In addition, Dr. Deering works on testing and developing novel sanitizers for the fresh produce industry. Such work has led to Dr. Deering receiving over $270,000 in funds from industry to improve the safety and quality of fresh produce. In addition, Dr. Deering is a member of the Center for Food Science Engineering at Purdue University and works to improve the detection capabilities for a scatter pattern-based detection technology, as well as working to detect human pathogenic bacteria in plant tissue using immunohistochemical techniques. Dr. Deering works closely with industry to help them address food safety and quality issues in the products they produce. In addition, she works with growers to help address problems they are facing in regards to postharvest sanitizers and food safety issues. Dr. Deering conducted trials that she started in Herat, Afghanistan to test the use of hermetic storage as a possible postharvest storage method that would be viable and economically feasible in Afghanistan. These trials were started in August 2015 with Herat University undergraduates and in December 2015 with AAEP II staff members.

(ii) Dr. Deering is currently working with the Open Ag Technologies and Systems (OATS) group at Purdue university who work to use open source data to create technologies and tools to address specific agricultural problems. The group received funding for this work in 2018 from the Foundation for Food and Agricultural Research. Dr. Deering's role is to bring ideas on how food safety in agriculture can be improved and she is currently working with Dr. Amy Reibman (Computer and Electrical Engineering) on a

project where open source codes and video analytics will be used to evaluate good agricultural practices in a packinghouse. This work will provide the tools to develop similar evaluations that can be used in other areas of food safety.

These research results have had significant broad impact on the community and educational contributions.

Dr. Deering works with Stakeholders to improve food safety of fresh fruits and vegetables grown in Indiana. This includes teaching Good Agricultural Practices (GAPs) courses, working with growers on Good Manufacturing Practices (GMPs) in their packinghouses, as well as providing resources to growers regarding appropriate postharvest sanitizers. Dr. Deering works closely with small growers to help them develop the best postharvest treatments and provides mock audits for growers to help them prepare for 3rd party audits. The FDA Food Safety Modernization Act (FSMA) Standards for the Growing, Harvesting, Packaging, and Holding of Produce for Human Consumption (better known as the Produce Safety Rule finalized in January 2016) has changed the way fresh produce is being produced and helping growers understand these changes is an important aspect of Dr. Deering's Extension program. She works closely with state agencies such as the Indiana State Department of Health, the Indiana State Department of Agriculture, and the Office of the Indiana State Chemist to help align fresh produce food safety efforts in Indiana and provides Stakeholders with the most accurate information available. Dr. Deering also works with local and national companies for the development of novel sanitizers that can be used in the fresh produce industry, as well as to assist them with any food safety related issues. She also has conducted trainings in Afghanistan and India to help improve GAPs, food safety, and postharvest storage issues

Finally, the PI has also made significant contributions to broadening participation of underrepresented groups.

Dr. Deering has worked for the last five years to support capacity building efforts in Afghanistan. She provides trainings regarding food safety, health, and hygiene to Extension educators from the Directorates of Agriculture, Irrigation, and Livestock (DAIL). In collaboration with Dr. Haley Oliver, Dr. Deering worked with the Afghan Ministries of Agriculture, Irrigation, and Livestock (MAIL) and Public Health to increase food safety knowledge and laboratory capacity within the government. This has allowed for increased food safety standards in Afghanistan.