# Expeditions: Collaborative Research: A Computational Approach to Provenance, Efficiency, Effectiveness, and Quality of Food Systems

## Response to Reviews

We thank the reviewers for their many insightful comments and questions. We first address general questions relating to the overall research, development, education, and broader impact plan. We then respond to specific questions.

### Innovations in Computer Science

One of the panel questions relates to contributions of the project to computer science. The project aims to re-engineer the food system – the world's most important socio-economic process – into a computing-driven ecosystem. This change is brought about by critical advances in computing, targeted to problems and solutions in the food system. We liken this to the development of the web in the 1990s, where similar questions were raised about innovation in web protocols, servers, and browsers, particularly given existing Gopher infrastructure. What was often missed in these arguments is the fact that the innovative underpinnings of the web provided the foundation for much of today's information and business infrastructure. Along the same lines, we aim to develop core computing substrates that enable a similar revolution in food safety, quality, productivity, and system efficiency. These substrates provide critical data infrastructure, guiding theory, and enabling technology for new enterprises throughout food systems. We note that there are virtually no integrated computational efforts focused on this important problem, further highlighting the novelty and innovation in the overall formulation. Our proposed computing efforts are organized into five inter-related thrusts that feature fundamental questions in computing.

**Theme A.** The problem of DNA profiling poses new questions on design of optimal barcodes that allow for rapid and inexpensive detection, traceback, and auditing of food sources. Minimizing barcode cost corresponds to minimizing perturbations to non-coding regions of DNA; robust coding corresponds to distribution of codes over disparate parts of the DNA; inexpensive detection corresponds to shallow reads and mapping, which in turn relates to the repeat complexity of selected regions of the genome; and accurate traceback corresponds to barcodes that are maximally distant from each other. Optimizing for these factors poses significant challenges in string editing and analyses over specified distributions. These problems are entirely novel in computer science, with broad applications beyond plant genome barcoding to tagging and tracking biological samples and processes.

**Theme B.** Multiobjective optimization problems on food networks involve a number of features that, considered together, represent fundamental advances in computing. These include large-scale data and noisy, incomplete model parameters that pose significant challenges for modeling and defining measures of optimality. These models warrant methods (such as stochastic optimization, sketching-based approaches), analyses (optimality in weakly sampled regimes, real-time decision

processes, and privacy preserving parameterizations) and implementations (distributed environments with varying communication and computation capabilities and constraints) that do not currently exist.

**Theme C.** The unique context of the food supply chain (perishability, critical importance of biological contamination) combined with differential market power with chain agents closer to consumers having greater power, yet the consumer having little individual power, gives rise to novel frameworks for optimization-based agent behavior and alternative game theoretic structures. Our approach is unique in that it jointly optimizes information-based transparency, supply chain efficiency, and mechanism design to create optimal market structures with high rates of participation. Ensuring that individual participants in a supply chain follow recommended protocols to update data in an accurate and timely manner, in the presence of agency effects (i.e., inability to observe actions), requires protocols to be incentive compatible, i.e., be the optimal choice across alternatives. In this context, the development of inventive compatible protocols is an open problem in computer science. Our issues of interest are somewhat different from conventional decision support or economic incentive models funded by other NSF programs, primarily because of the close connection to specific computing elements such as protocols, information breadth, and information value. Our proposed solutions enrich core computing principles, while developing new methodologies to solve optimization models with sensor inputs.

**Theme D.** Architecting a privacy preserving open blockchain (as opposed to current vendor specific solutions), poses significant challenges and presents exciting opportunities. In a closed system, it is possible for a malicious entity to substitute an authentic item with a counterfeit item in the absence of intrinsic watermarks (DNA watermarks), without being identified through traceback; this is not possible in our proposed open system. A key challenge in an open framework is the need for privacy across the supply chain. We aim to develop privacy preserving solutions, and to comprehensively study tradeoffs between cross-(supply)chain traceability and privacy. In consortium scenarios, we do not need permissionless systems (e.g., Bitcoin and Ethereum) that allow anybody to join the market. We propose a consortium blockchain, where the number of validating nodes is small – on the order of the number of participating organizations. Our proposed secure multiparty computation (MPC)-based approach allows us to further reduce the number of validating nodes. In an open system many of the parties may not have the capacity to add individual blockchain nodes. Current consortium blockchain solutions expect large surrogate entities (such as IBM) to function trustfully on their behalf. These solutions have significant privacy issues, which are particularly problematic in the context of data protection laws (such as CCPA and GDPR). We develop MPC-based protocols that allow low capacity nodes to instead rely on a small set of MPC nodes such that their security and privacy are preserved even when one MPC node is honest. Since different supply-chain contexts involve different threat models, the project also focuses on developing tailored MPC protocols. These formulations, and associated solutions, represent significant advances to the state of the art, and have broad applicability beyond our domain.

**Theme E.** Mapping genotype to phenotype has been recognized as one of the Big Ideas by NSF, and computing is an essential component of the solution. The complexity of this problem is rooted in the high dimensionality of the phenotype prediction problem, with genomic variants, weather, field management (irrigation, fertilizer, pesticide use, and associated schedules), providing inputs to the problem. Only a tiny fraction of the input space has been explored due to the high

cost of experimentation. Constructing interpretable and actionable maps from inputs to phenotypic outputs poses significant challenges in data analysis and machine learning due to sparsity of samples and missing data, heterogeneity of input parameters, biological constraints on models, identification of causal effects, and need for refined statistical models. These problems, and proposed solutions represent significant advances in computing.

**Interconnections among themes.** The five themes are strongly interconnected; modeling interactions between these themes is an important and innovative aspect of the project. For instance, DNA watermarking, traceback, and phenotyping are closely related, since the accuracy of DNA watermarking and pooling determine accuracy of traceback, which in turn influences supply chain instrumentation to accurately retrace contamination events. Similarly, for in-silico phenotyping experiments, assessing the impact of genotype and environmental factors depends on the accuracy of traceback. Incentive mechanisms, cost models, and supply chain instrumentation are closely linked, since incentive mechanisms must balance instrumentation and data costs with benefits from instrumentation in an environment with partial participation and regulatory constraints. Supply chain instrumentation and multiobjective food network optimization techniques are linked through instrumentation cost, data quality, and savings from optimization. Likewise, in-silico phenotyping at consumer end relies on supply chain data (transit time, refrigeration, processing), in addition to properties at source. Characterizing in-transit loss as a function of different transit modes is an essential input to the optimization procedure. These are only some of the examples of strong links between all of the thrusts.

**Feasibility of DNA Barcoding and Industry Adoption.**

A second question raised by the reviewer(s) relates to the feasibility of the concept of DNA Barcoding, its adoption by seed manufacturers, and acceptance by the broader community. We submit the following observations in response:

- **Genomic Manipulations are Commonplace in Industry.** A number of methods for modifying genomes have been developed over the past three decades. Genes are transfected into cells using conventional techniques such as Biolistics, Electrophoration, PEG-Mediated delivery, or Agrobacterium-Mediated delivery. More recently, a number of nanoparticle based delivery techniques have also been developed. Gene editing techniques based on CRISPR/Cas9 and, recently, more precise DNA endonuclease enzymes have been developed. These techniques are commonly used in labs, as well as industry, and can be used to insert precise barcodes that can be read back exactly (with vanishingly small error probabilities) relatively cheaply (about $100 at current state of technology).

- **Barcodes Can be Easily Constructed Entirely Using Current Genomic Manipulation Techniques.** 98% of a given crop genome does not directly encode proteins; rather it encodes (non-coding) RNAs or is involved in regulation and structure of the genome and gene expression. Our proposed barcoding technique uses synonymous substitutions in codons, therefore the associated protein sequence and structure are unchanged. In an abundance of caution, we propose these changes to non-coding regions, resulting in *no* change to the protein expression or structure. Our genetic modifications are completely transparent from a protein profile perspective.

- **Regulatory Processes for Such Seeds are Rapidly Maturing.** Scientifically, genetically modified crops are considered completely safe. A recent au-

thoritative commentary in Science (https://www.sciencemag.org/news/2016/05/once-again-us-expert-panel-says-genetically-engineered-crops-are-safe-eat), the accompanying 2016 report by the National Academy of Science and the National Academy of Engineering, all say that genetically engineered crops are safe. The Science article cites two cases in which USDA deemed two CRISPR-edited crops to be exempt from its review process because neither contained genetic material from species considered to be "plant pests". In March 2018, U.S. Secretary of Agriculture announced that the USDA would not regulate new plant varieties developed with new technologies like genome editing that would yield plants indistinguishable from those developed through traditional breeding methods (https://www.usda.gov/media/press-releases/2018/03/28/secretary-perdue-issues-usda-statement-plant-breeding-innovation).

- **Societal Perceptions of Genetic Modifications have Changed.** In view of the increasingly conclusive scientific evidence of safety of genetic modifications, societal acceptance of these plants has also rapidly changed. A recent survey of consumers (https://doi.org/10.1016/j.gfs.2018.10.005) indicates that a majority of US consumers (56%) would consume genetically modified and CRISPR edited plants. This number is rapidly growing, and we do not believe this is an impediment to the project, given widespread industrial use and regulatory acceptance.

- **The Project Team has Ongoing Collaborations with Seed Manufacturers on these Topics.** CoI Rainey has extensive funded efforts with seed producers and related agencies, including Dow Agro, Smithbucklin, Indiana Soybean Alliance, United Soybean Board, and the National Institute of Food and Agriculture. CoIs Grama and Gleich have initiated a recent collaboration with Bayer on problems closely related to the proposed project (please see letter of support from Bayer in the Proposal). These collaborations ensure that the proposed concepts can be fully realized in practice.

## Broader Impact, Outreach, and Education

**Engagement with External Stakeholders.** The proposed project builds on extensive existing outreach programs from the PIs' groups to various stakeholders. CoI Deering directs Purdue's extension efforts in the area, working with a large set of stakeholders on compliance with the FDA Food Safety Modernization Act (FSMA) Standards for the growing, harvesting, packaging, and holding of produce for human consumption. She works closely with state agencies, including Indiana State Department of Health, Indiana State Department of Agriculture, and Office of the Indiana State Chemist, to help align fresh produce food safety efforts in Indiana. These outreach efforts include 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 post-harvest sanitizers. Beyond the US, she has conducted training courses in Afghanistan, India, and Tajikistan to help improve GAPs, food safety, and postharvest storage. CoI Mauer has extensive collaborations with food processors and producers, including PepsiCo, McCain Foods, and Nestle, in addition to a number of projects with USDA. CoIs Iyer, Preckel, and Foster have collaborations with supply chain vendors, including those in India and Colombia. These active collaborations with stakeholders across the spectrum, from seed producers to consumers and policy-makers will enable rapid deployment and testing of all research artifacts and technologies to real-world scenarios, instead of a "build-it-and-they-shall-come" model of engagement.

**Cyberinfrastructure Ecosystem.** The proposed CropHub is a key element of outreach to the broader community. CropHub hosts educational material, research products in the form of software, services, publications, and media, outreach to external stakeholders in the form of training material, external/ cooperative research and projects, as well as outreach to underrepresented groups in the form of workshops, summer schools, and avenues for sustained engagement. CropHub significantly leverages Hub technologies pioneered by Purdue and used in a number of critical scientific resources, such as the nanoHub and SoIHub. CropHub builds on resources from `http://www.SafeProduceIN.com`, to seamlessly provide services to a broad class of stakeholders (`http://www.SafeProduceIN.com` is run by CoI Deering and has a large userbase of external stakeholders). For educational initiatives, CropHub links to the Purdue Education Store (`https://edustore.purdue.edu`), which provides online services for instruction, and to the Science of Information Hub (`http://www.soihub.org`) for a broader CS engagement. Finally, for software and services, CropHub follows best practices gleaned from HubZero, which powers over 60 science gateways worldwide and a total of over 2M users.

**Management Structure and Team.** We are strongly committed to building a cohesive and collaborative group, where the whole far exceeds the sum of the parts. To this end, we adopt a set of best-practices from our Center for Science of Information on fostering collaboration. These mechanisms include shared students (graduate students are required to have advisors from multiple thrusts), project fellows (postdoctoral researchers working with more than one CoI), student led projects (students are funded for small multidisciplinary projects that they design), undergraduate project teams (groups of undergraduates mentored by graduate students on interdisciplinary projects), and weekly project seminars. We also benefit from largely being on the same campus, facilitating frequent face-to-face interactions.

With respect to role of various committees, the external advisory committee is comprised of industry as well as academic leaders. Their role is to assess project progress, to make recommendations for streamlining processes, and to provide guidance on emerging trends in industry. The executive committee of the project is charged with the operation of the project, setting project priorities, allocation of funds, assessment of broader impact, education, and outreach efforts, as well as performance of individual investigators in terms of their research alignment with project goals and their collaborative efforts.

The project benefits greatly from Purdue's top-rated Agriculture, Food Sciences, and Operations Research programs. The Computer Science investigators on the project have well-established credentials in various aspects of the project. We believe that we are fortunate to have a comprehensive, highly-qualified team capable of successfully executing this challenging project largely collocated at Purdue.