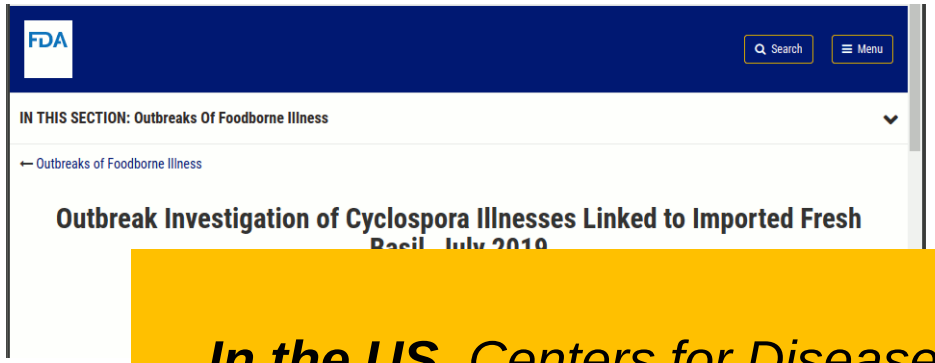
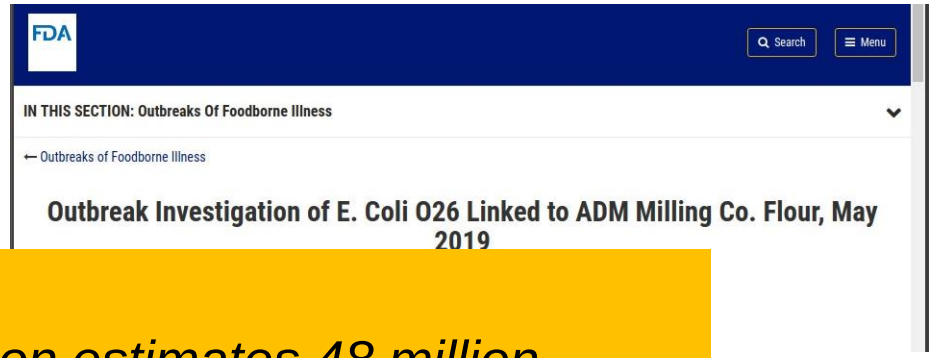
A Computational Approach to Provenance, Efficiency, Effectiveness, and Quality of Food Systems

Overview

1. Project Motivation
2. Computational Challenges and Project Themes
3. Education, Broadening Participation and Broader Impacts
4. Management Structure

Motivation: Safety of Food Systems is a Massive Problem





In the US, *Centers for Disease Control and Prevention estimates 48 million cases of foodborne illnesses, 128,000 hospitalizations, and 3,000 deaths, with a cost of $152 billion in medical expenses, lost productivity and business, lawsuits, and compromised branding.*

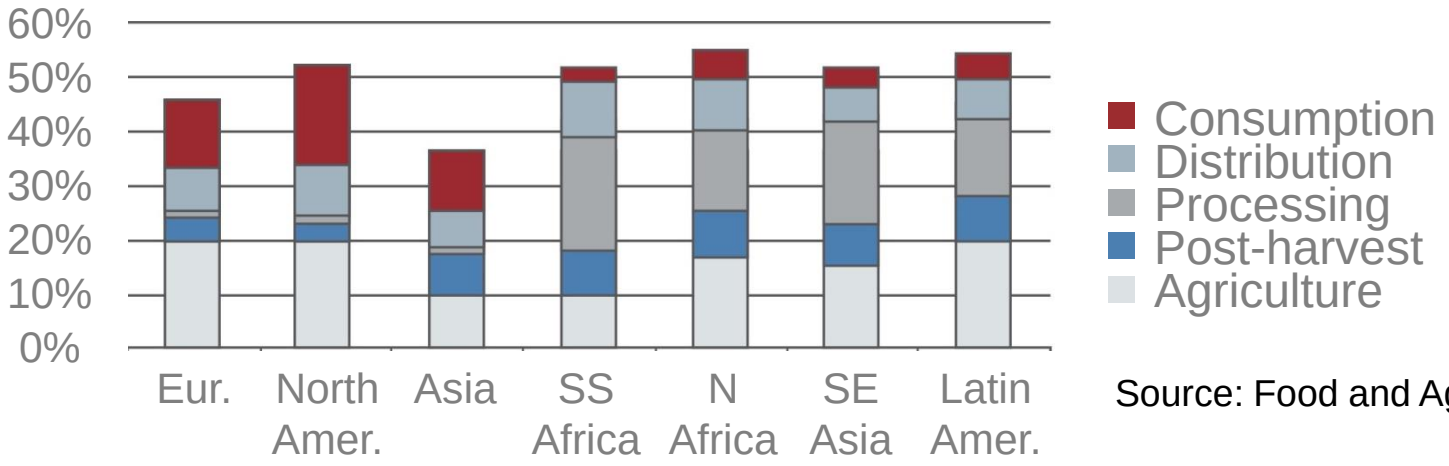
Globally*, the annual numbers are 600 million cases, 230,000 deaths from foodborne diarrheal disease, and over 33 million disability adjusted life years!*

Motivation: Waste in Food Networks is a Challenge

An estimated *20 billion pounds of produce is lost* on US farms each year.

Estimates indicate that *$7-15.4 billion worth of fresh produce spoils* in the supply chain annually before reaching the consumer in the US. This corresponds to approximately 12.3% of fruits and 11.6% of vegetables.

The consumer then accounts for an additional 19% of fruit and 22% of vegetable losses.



Source: Food and Agriculture Organization

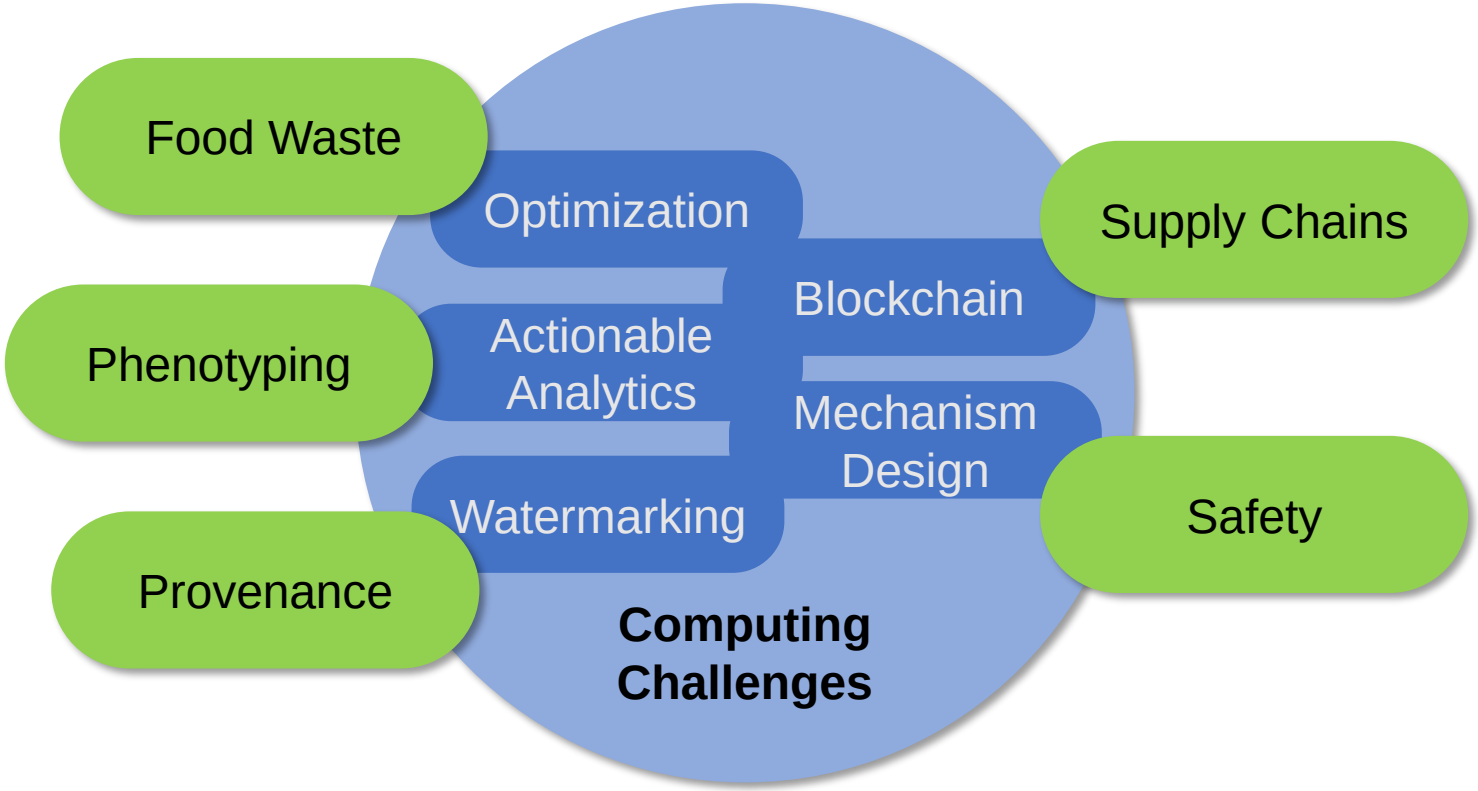
All this, while *one-sixth of Americans do not have enough food to eat*; worldwide over 795 million people suffer from hunger, and the prevalence of undernourishment has been rising.

Motivation: Food Quality and Nutrition is a Challenge

Fresh produce, on average, spends up to 50% of its shelf-life (typically 21 days from harvest to consumption) in the supply chain! This has significant impact on quality and nutritional value. Supply chain conditions (e.g., cold chains) strongly determine quality.

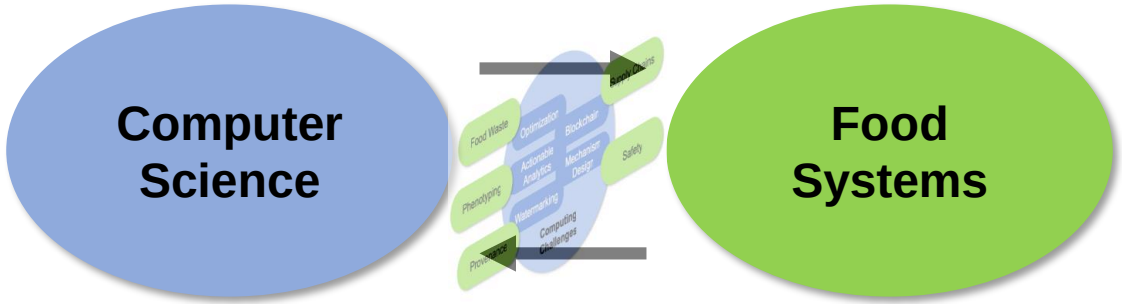
Supply chains are highly fragmented with little (highly asymmetric) information flow across various entities (the EU recently passed guidelines for price transparency across supply chains).

The goal of this project is to bring disruptive innovation through computing to food safety, quality, distribution, and production.



The goal of this project is to bring disruptive innovation through computing to food safety, quality, distribution, and production.

Our Computer Science Lens will enable us to address these issues as well as bring improved methodology and insights to computing



1. Novel computational techniques for food provenance
2. Use of large-scale instrumentation for supply chain modeling and optimization
3. Incentive mechanisms for participation across the supply chain
4. Integrated blockchains for an open marketplace for producers, processors, distributors, retailers, and consumers
5. Use of provenance information combined with seed, field management, and weather data to comprehensively map quality and productivity.

Theme A: Computational Approaches to End-to-End Food Provenance

Goal

- To design and develop novel computational models and methods that allow robust and high resolution provenance for produce

Challenges

- Design of DNA barcodes that are minimal, robust to perturbations (both local and global), can be read back easily, and are constrained to specific regions of the DNA
- Use of intrinsic sequence features (simple sequence repeats) as barcodes
- Design of low-cost shallow sequencing, deconvolution, and mapping algorithms

Connections

- Theme B. Optimization of traceback accuracy
- Theme D. Mapping product to entities in the blockchain ledger
- Theme E. Virtual phenotyping

Theme A: Computational Approaches to End-to-End Food Provenance

CS Challenges. Sequence Modeling, Barcode Design, Readback Algorithms

- Design of optimal barcodes for rapid and inexpensive detection, traceback, and audit
- Analytical proofs of uniqueness of barcodes for real models of DNA sequences
- Ensuring robustness through (constrained) distribution of barcodes over the sequence in the presence of point mutations, hybridization, and other common perturbations
- Establishing minimality of barcodes with constraints on uniqueness and robustness
- Maximizing traceback accuracy through maximally distant barcodes
- Establishing feasibility of use of intrinsic sequence features as barcodes
- Analyzing sequence depth in the context of repeat complexity of the DNA sequence for inexpensive readback

These challenges pose significant problems in sequence modeling, analyses, and design of associated methods.

Theme B: Food Supply Chain Modeling and Optimization

Goal

- Develop mathematical programming models to: (i) optimize supply chain architecture; (ii) determine location of tracking devices; and (iii) determine optimal responses to contamination events.

Challenges

- The mathematical programming models are a natural approach to model such systems, but might only be able to represent a *partial, distorted view* of the system.
- Understanding whether approximately optimal solutions can still be reached in such settings is a major open problem.
- Understanding how to approximate the optimal solution in a distributed setting, where each actor optimizes over his own set of constraints is also a major challenge.

Connections

- Theme C. Design mechanisms to incentivize the actors of the supply chain to reveal information about the system.

Theme B: Food Supply Chain Modeling and Optimization

CS Challenges. Solve optimization problems with missing or noisy variables or constraints or objective function in a distributed manner.

- Let G be the underlying (unknown) optimization problem and G' be the observed, severely perturbed, one. Characterize (i) whether the feasible sets of G and G' significantly overlap, and (ii) whether the optimal solutions of G and G' are close.
- Difficult even for Linear Programs; related work goes under the name sensitivity analysis. Very open for convex optimization.
- **Related problem.** In the presence of noise, early stopping of iterative optimization algorithms might lead to better generalization via implicit regularization.
- **Related problem.** In the distributed version of the problem, each actor might be able to locally solve an optimization problem with access to all his constraints/variables. What is the minimal information that each actor should globally reveal in order to (at least approximately) solve the overall optimization problem?

Theme C: Incentive Mechanisms for Participation

Goal

- Design algorithms (e.g. auctions, multi-round games), where the inputs are provided by correlated strategic agents (e.g. by a network of supply chains), and the result is truthful, max. welfare, etc.
- Design mechanisms for dynamic environments, where the actors are myopic and have asymmetric, limited information, and are subject to external forces (weather, pricing, trade)

Challenges

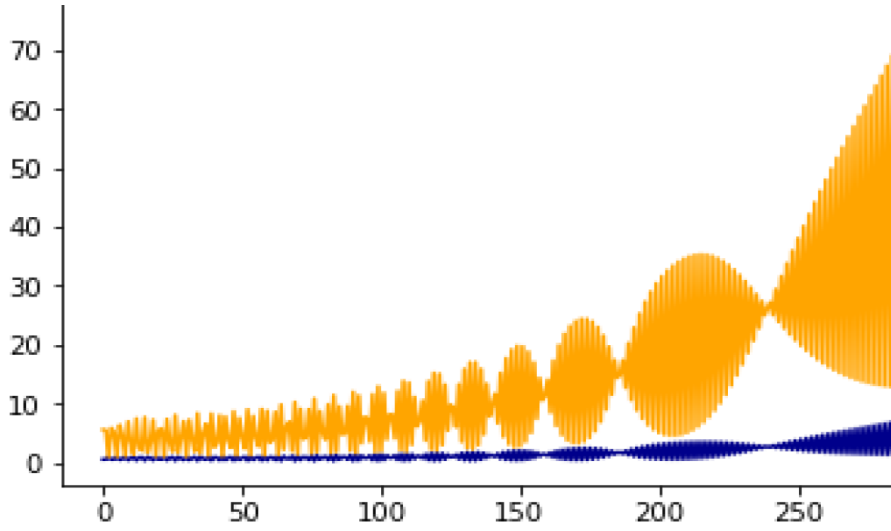
- In general, incentive compatible mechanisms do not compose.
- Myopic actors and limited, asymmetric information environments complicate mechanisms to achieve a desirable global solution.

Connections

- Theme B. Supply chain models and sensing mechanisms will influence modeling assumptions in this theme.
- Theme D. Blockchain models can be used to ensure privacy.

Theme C: Incentive Mechanisms for Participation

CS Challenges. Dynamics on networks of interacting agents, where agents continually adapt to the observed state of the world



- Myopic actions – e.g. cutting production costs in places difficult to observe – can lead to suboptimal products or the market unraveling completely over time.
- Results are highly dependent on information assimilation methods including regret minimization, multiplicative weight updates, proportional updates.
- How to measure properties of these networks when we assume entities are strategic.

Oceanic games model scenarios where there are big and small players.

- This results in complex market games and learning problems where players are not independent of each other (e.g. have correlated utilities, are big vs small).

Theme D: Systems Infrastructure for Open Markets

Goal

- To design and develop an open multi-vendor blockchain platform offering a powerful trade-off between privacy, traceability, and performance.

Challenges

- Current vertically integrated provenance solutions cannot account for cross-vendor transactions by players without a trusted facilitator (such as IBM) and remain vulnerable to cross-vendor counterfeiting.
- Multi-vendor blockchain information should not be ubiquitously accessible across the platforms and should be protected from competitors.
- Incentive mechanism for supply-chain players to join and participate in the distributed ledger needs to be combined with the provenance ledger.

Connections

- Theme A. DNA profiles provide strong identities.
- Theme C. Incentive system for fair participation.

Theme D: Systems Infrastructure for Open Markets

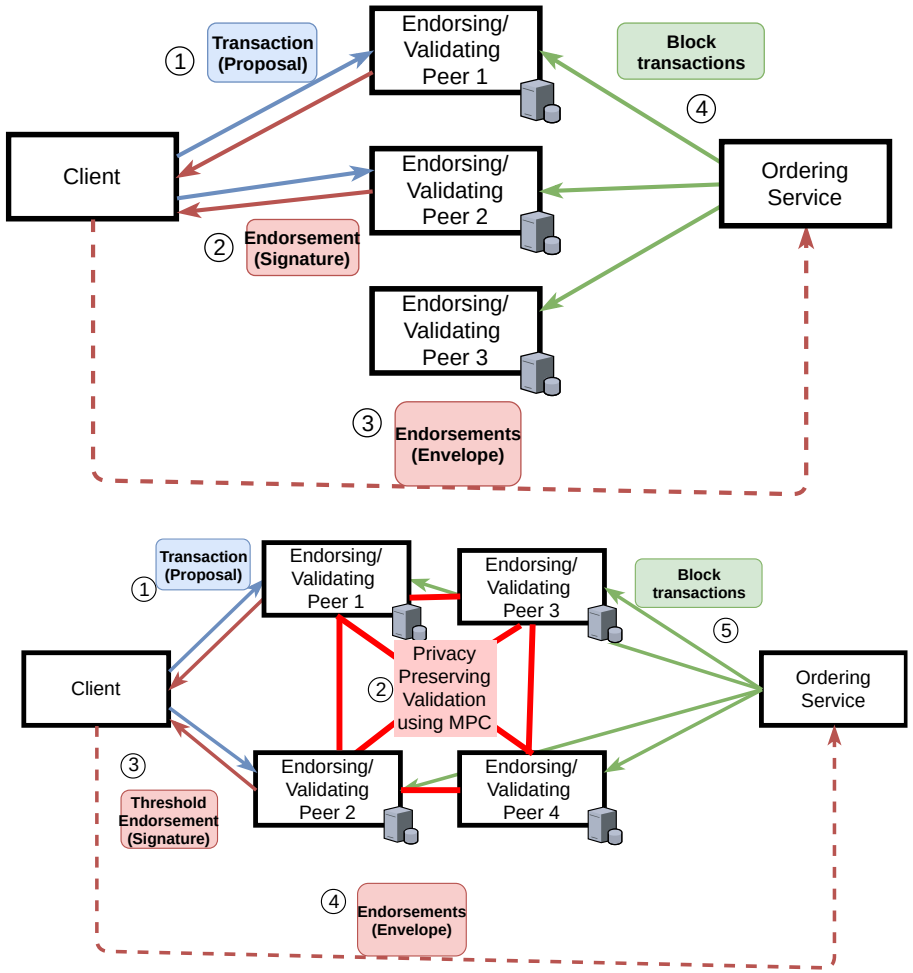
CS Challenges.

In multi-vendor environments, it may not be acceptable to supply-chain processes that their business processes are visible across the platform.

- This suggests a secure distributed validation architecture that offers acceptable agreement between processes using execute-order-validate (EOV) processing of Hyperledger.

Most multi-vendor distributed ledger information should not be ubiquitously accessible across the platform and should be protected from competitors

- This type of privacy is novel and has not been formalized.
- Beyond EOV, we will achieve it using innovative solutions based combinations of secure multi-party computation (MPC) and non-interactive zero-knowledge proofs (NIZK).



Theme E: Scalable Phenotyping Through Large-Scale Data Integration

Goal

- To build scalable data analysis pipelines to address simple and complex questions about what impacts end-product phenotypes.

Challenges

- Field management data is sparse & high dimensional with strong spatiotemporal effects (weather).
- Food quality phenotypes – taste and nutrition – are results of extremely complex systems spanning seed-to-store.
- Data on food safety and microbial contamination is typically very incomplete due to legal and privacy issues.
- The community needs *data and information to support possible hypotheses* and not black box *solutions*.

Connections

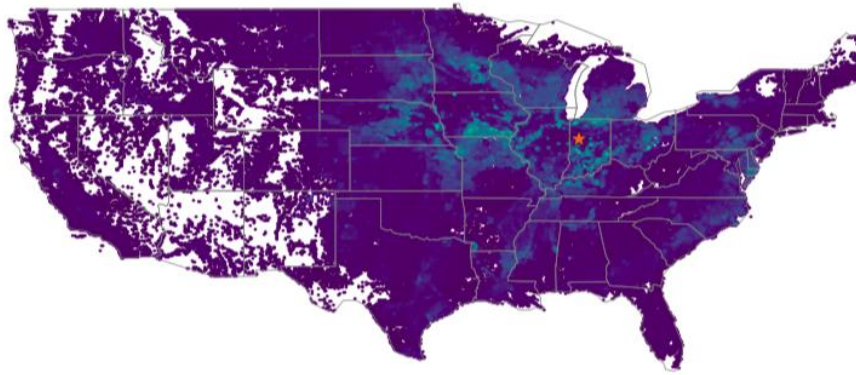
- Theme A. Provenance data provided by watermarking
- Theme B. Supply chain flows influence quality

Theme E: Scalable Phenotyping Through Large-Scale Data Integration

CS Challenges. Hypothesis generation

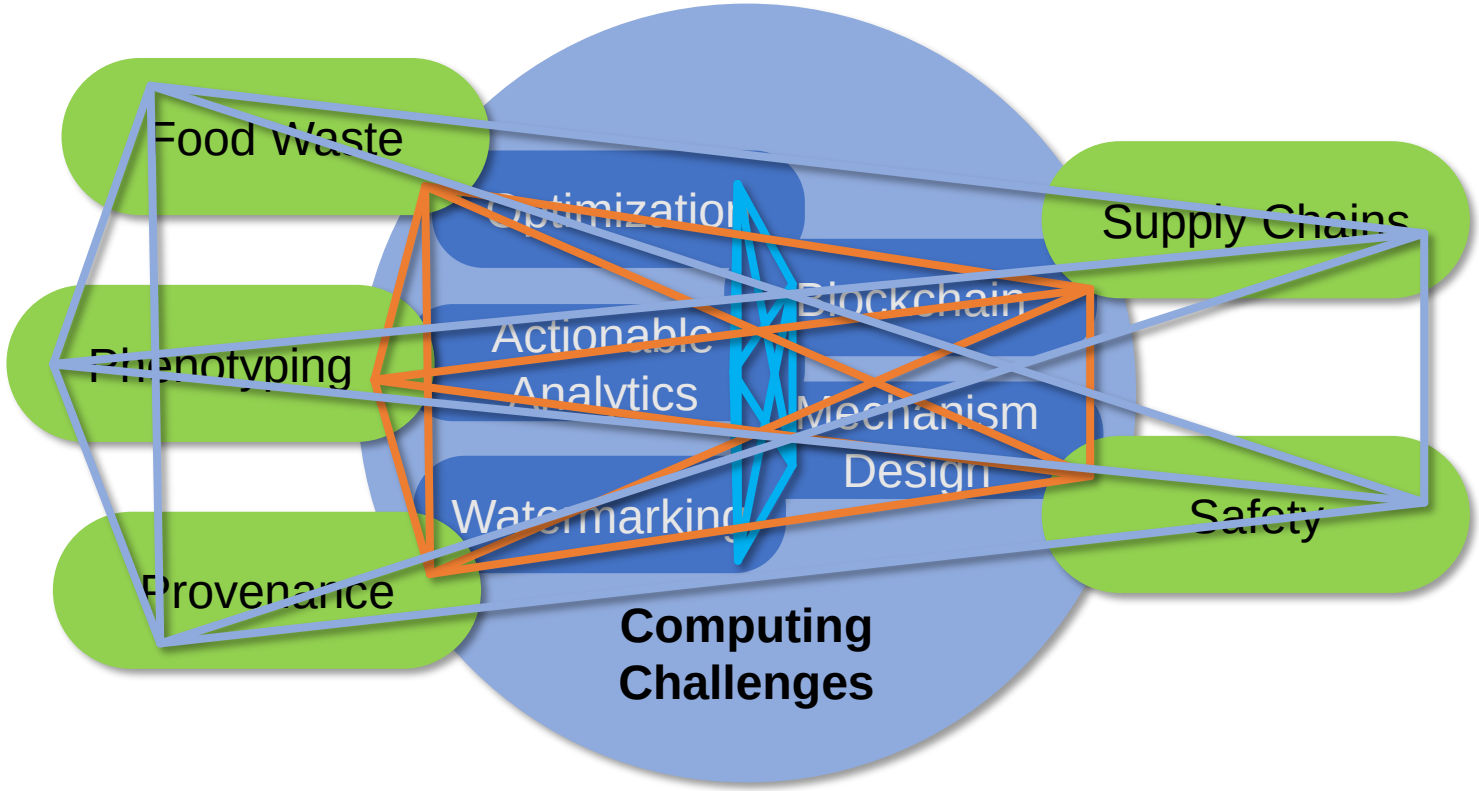
- Graph and tensor-based local search (random-walk / PageRank / eigenvector) approaches will be used to generate non-trivial hypotheses for future experiments.
- There is no state of the art on using these with the types of complex mixtures (low-rank + sparse) models of data we need.
- Higher-order expansions of data are extremely useful to find subtle and tunable insights into the data and in terms of new hypotheses
- Matrix completion and imputation are needed to missing data and outliers
- Our goal is to build algorithms that work in order of the size of the input dataset, not the intermediate (expanded) graph.
- RandNLA and tensors will be an enabling technique.

All of the analysis and results must be explainable in terms of raw data (fully reified).



Graph-based local search for agricultural regions similar to West Lafayette

Our research themes are decoupled in failure and intertwined in success and help make the whole greater than the sum of the parts.



- Important new intersections of areas within computing
- New connections at the interface between computing and food systems
- New possibilities for how computation can be a subtrate to improve our food systems

Educational Programs

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.

Broadening Participation

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 prograr 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.

Broader Impacts of the Proposed Work

Local Farm Communities

- Education programs, training programs, and computational resources

Farming, Food Processing, Supply Chain, and Retail Communities

- Exchange of data, models, tools, and optimization techniques

Policymakers and NGOs

- Guiding policy and informed regulation

International Partners

- Collaborations with stakeholders in India, Kenya, Bangladesh, Cambodia on best practices in data collection and use

Broader Scientific Community

- Computer Science communities to motivate new modeling paradigms, problems, and solutions.
- Food science communities to motivate new problems and solutions.

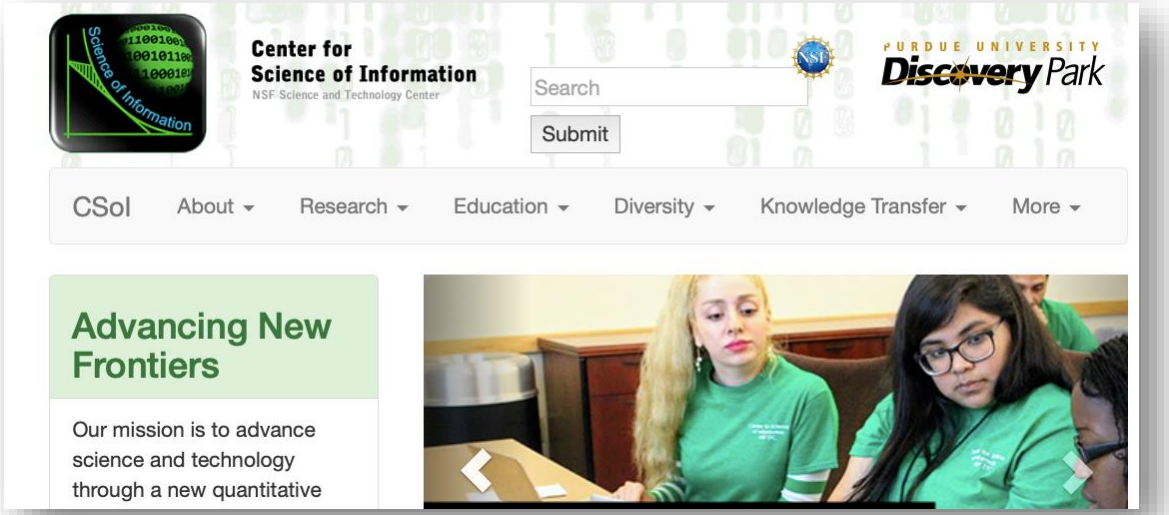
Broader Impacts of the Proposed Work

CropHub. Repository for all technical and educational material.

- The research platform focusing on datasets, software tools, and research artifacts
- The education platform, focusing on modules and instructional material;
- Outreach platform, focusing on workshops, tutorials, and meetings aimed at professionals, policy-makers, and NGOs; and
- Broadening participation platform, for recruitment, mentoring, and assessment of BPC programs.

Built using Purdue's unique HubZero platform.

- https://soihub.org
- https://nanohub.org
- https://hubzero.org



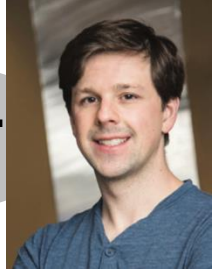
Institutional Oversight Board
Senior Officials from Purdue & UCR

External Advisory Board

Executive Committee

Project Director - Ananth Grama

Broadening Part.
David Gleich



Education
Lisa Mauer

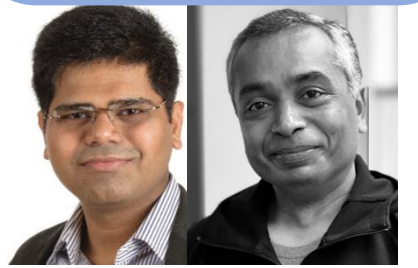
Broader Impacts
Amanda Deering

Genomic Profiling
Stefano Lonardi

Supply Chain Opt.
Ananth Iyer

Incentive Mech.
Simina Brânzei

Blockchain
Aniket Kate



Phenotyping
Petros Drineas

Project Management Structure

Project Contingencies

Personnel contingencies

- The project is led by Ananth Grama.
- David Gleich and Petros Drineas serve as Associate Directors of the project.
- All other project teams have sufficient extra personnel to handle unforeseen contingencies.

Technical contingencies

- Each of the thrusts has formulated technical contingency plans (see proposal).

Institutional contingencies

- The project is primarily housed at Purdue, thus largely minimizing contingencies due to partnerships and collaborations across institutions.

A Computational Approach to Provenance, Efficiency, Effectiveness, and Quality of Food Systems