Theme E: Actionable Analytics & Phenotyping

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

Challenges.

- Field management data is a sparse, high dimensional space 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 largely incomplete due to legal and privacy issues.
- The community wants *data and information to support hypotheses generation,* and not black box *solutions*.
- The data itself is tremendously heterogeneous and multimodal, and we need to analyze mixtures and more complex large-scale data integrations.

Technical approach for large scale integrated data analytics

- Our goal is to formulate hypothesis generation frameworks. These are techniques that identify useful pieces of raw data and potential relationships
- We infer the relevance of information to an extremely general notion of a query.
- These approaches are commonly used for protein function discovery in bioinformatics, where they are sometimes called “guilt by association”
- As CS problems, they involve solving large systems of linear equations, eigenvectors, and increasingly for higher-order scenarios, nonlinear systems of equations and tensor eigenvectors.

Motiving example based on current state of the art. What are similar agricultural regions to West Lafayette?

MAPSPAM is a spatial production allocation model based on a fusion of satellite data and simple model.

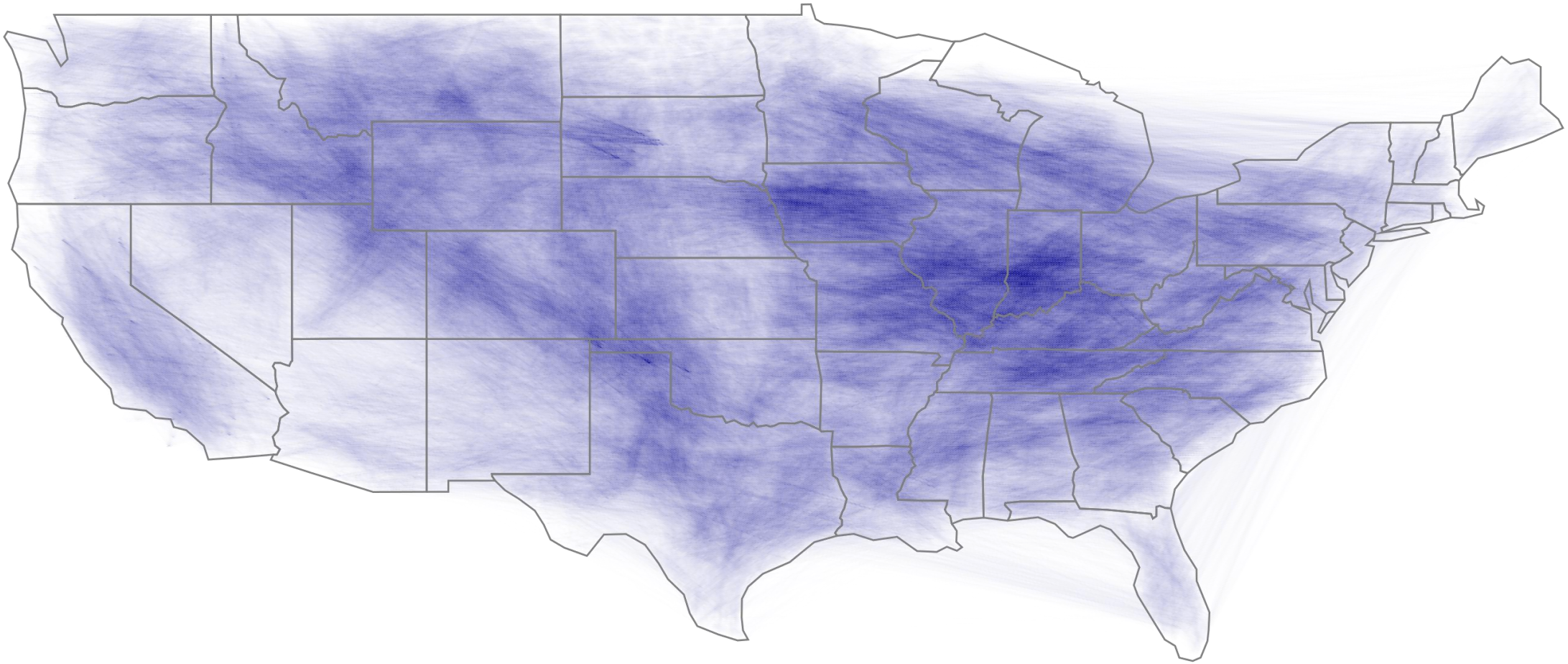
For each of 800k “pixels” on a world-grid, we get various quantities on 48 standard crops.

- Two regions are similar if they produce similar crops.
- Applying a k-NN threshold gives us a simple undirected graph.

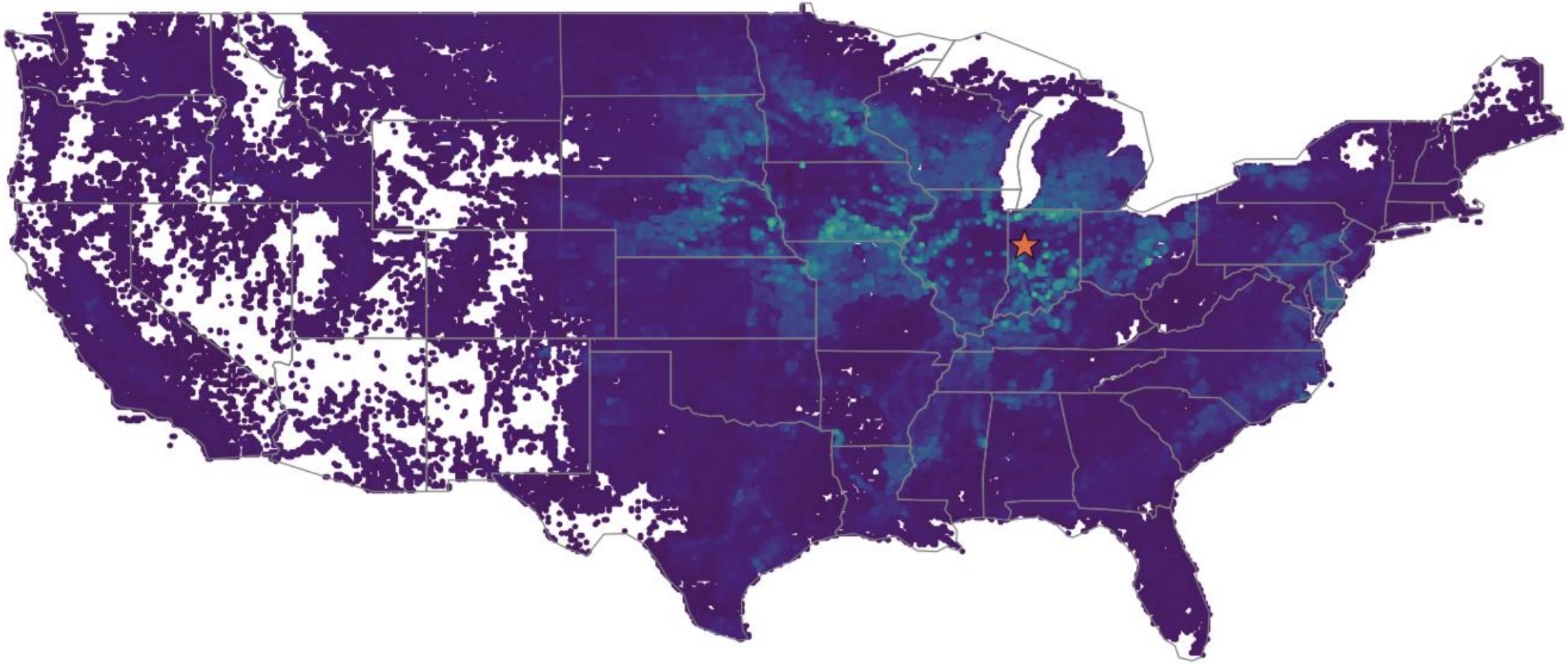
We focus on the US-48

Where there are many bananas grown.

Edges of an 80k-by-80k graph that show similar regions based on harvested area.



We can use a diffusion and PageRank-style similarity to find other regions similar based on harvested crops.



This motivates our approach in this section for integrating large scale data for finding hypotheses

A hypothesis is a new relationship that is not explicit in the raw data.

Graphs and local search on graphs, mixtures of graphs, higher-order relationships on networks, hypergraphs, and nonlinear processes on graphs are extremely useful ways to identify latent, or hypothetical, relationships.

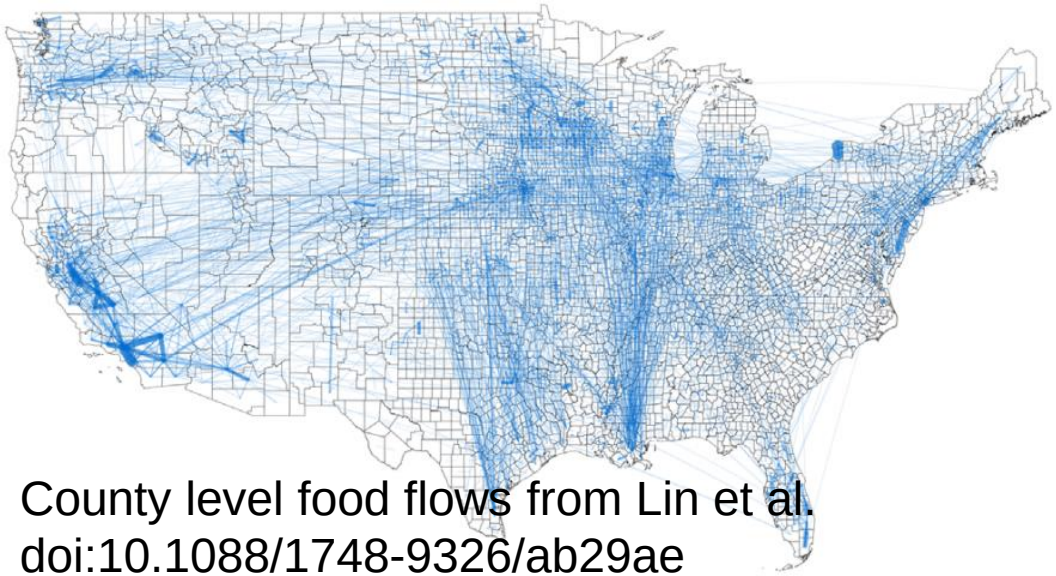
Classical CS uses in link prediction, ranking, semi-supervised learning, image segmentation, etc.

The starting point of our technical approach is to use these to design a large scale integrated analytics framework.

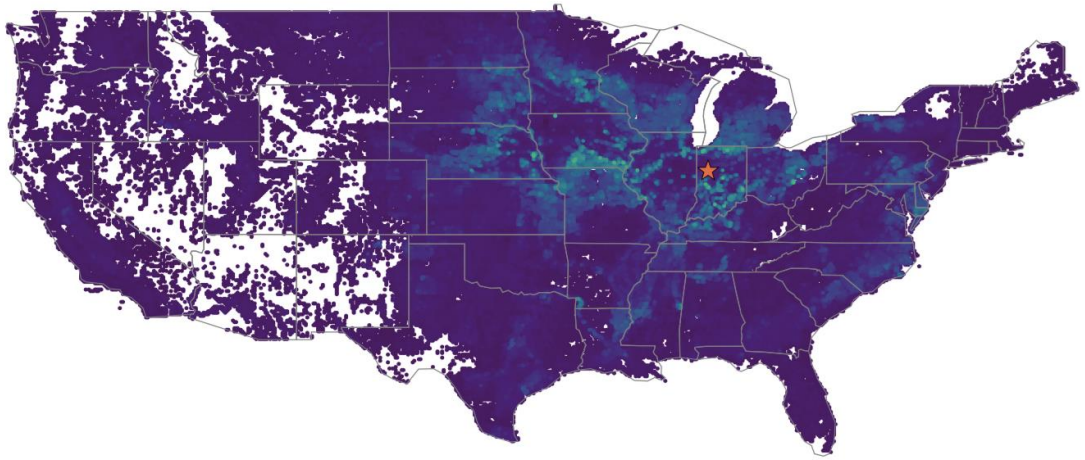
A related class of techniques seeks explainable decompositions of low-rank matrices (PCA-like) and tensors, which gives an alternate strategy for multiple types of data.

State of the art

- There are no large scale integrated data analyses of seed-to-store data.
- The current state of the art for localized graph search methods all require explicit neighborhood access, which will not be possible in many of our scenarios.
- Mixtures of diffusions have traditionally been done either on a multimodal graph representation or on some type of combined graph; we view these nonlinearly and wish to study processes that diffuse multimodally.



County level food flows from Lin et al. doi:10.1088/1748-9326/ab29ae



Specific instances of challenges we anticipate with this type of approach

We wish to compute localized-PageRank-like eigenvectors but on mixtures of data where direct neighbor access is not available.

- e.g. the entries of the matrix are given by a mixture of low-rank and sparse

We wish to find vectors that are simultaneously near eigenvectors of multiple sets of data.

We can relate PageRank to optimization problems

$$\text{maximize} \quad \mathbf{x}^T \boldsymbol{A} \mathbf{x}$$

$$\text{maximize} \quad \min(\mathbf{x}^T \boldsymbol{A} \mathbf{x}, \mathbf{x}^T \boldsymbol{B} \mathbf{x})$$

For multiple datasets, we want the best worst-case

This ties into Laplacian solvers and a host of other active areas.

Specific Data Analysis Problems

FarmGrep. A 21st century farmer's almanac.

Field management data will become as ubituiqous as sequencing and expression data are in bioinformatics.

Key CS Challenge: Making vast databases of raw information useful.

- "Google-like" input to identify field management data relevant to a query.
- Tell me what happened in a situation like …
- Spatio-temporal queries and sparse, high dimensional, structured search
- Surface information useful to make decisions, the data sources, etc.

For many of these, we can encode them as vectors over our database and use local search to find answers.

Initial Data Sources for Analytics

- USAID sources make their data available to the public
- Partnership with WHIN will make sensor & field data available
- Co-PI Rainey performs extensive experiments and user-surveys, and will provide data

Differences from “Explainable AI”

The goal is not AI or decision making. We want explicit algorithms to reveal and refine useful relationships, akin to how GIS tools are used today.

The goal in FarmGrep is surfacing raw information to a decision maker.

The goal of our advanced analytic studies is to probe specific phenotype questions.

- Does pollution impact micronutrient content? This requires a more complex setting where we view the result as a computation as an simulation or experiment.
- A phenotype is any observable feature, so this framework is general

Our view is that these analytics will generate hypotheses that would need to be tested with either designed experiments or a natural experiment.