# Nathan Russell

Junior, Industrial Engineering
Mentored by Visual Analytics Group

I study Operations Research and I am the president and founder of the INFORMS chapter at UIUC. I am interested in applications of machine learning, optimization and signal processing to technical problems arising in science and engineering. I especially enjoy problems with sensing and vision. I will be working with NCSA during the Summer of 2014 on supervised feature ranking for P>>N mixed attribute data sets that arise in bioinformatics.

### Project Background

The modern farm is a large scale operation and rich with technologies. The goal is to leverage a combination of machinery sensors and environmental data to give decision makers insight.

### Project Objectives

- Develop prototype software to collect, clean and visualize data collected from John Deere machinery
- Develop a set of metrics and classifiers to aide in high level interpretation of machine utilization and field efficacy
  - Machine State Classifier
  - Field Irregularity metric

## Data Architecture

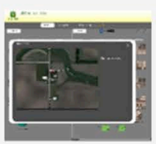
Hundreds of files per machine – inconsistent formatting/ naming

JD SQL DB
1 table per machine per year + summary tables

Public Weather data

Python Scripts

D3



Prototype GUI is built on Java's D3 library and combines real-time computation with stored data in an SQLITE DB. Users can view high level summary information in a calendar view and then pick timespans to drill down on.

## Knowledge Discovery & Machine State Classifier

Unfiltered sensory data is collected from John Deere vehicles during every second of operation. To make this data useful to farm managers, data is summarized in a set of high level metrics presented in the prototype visualization software. For Some metrics, It is necessary to know what task a vehicles was engaged in at a given time, otherwise known as the "Machine's State". The process of transforming Raw data into actionable information is described in the diagram below.

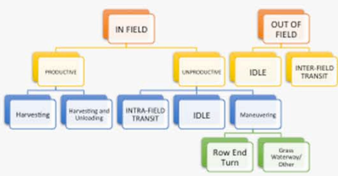
### Evaluation Criteria for Classification

Classifier Efficacy was evaluated through a combination of cluster seperability, distributional moments and feasibility review by subject matter experts.

**Existing Boolean logic constructed from subject matter experts.**
- ✓ Clear reason behind class
- ✓ Unclassified > 75%
- ✓ Poor classification on Atypical solution

**Temporal-Spatial Two Step Clustering on Mixed Attributes**
- ✓ Identifies many potential classes
- ✓ No Class Labels
- ✓ Dependent on clustering initial conditions

**Hierarchical classification based on weighed similarity to class centroid**
- ✓ 100% Classified
- ✓ Improved Class Efficacy for all Classes
- ✓ Parent Class Efficacy > Daughter Class Efficacy




### Data Refinement Process

1. Drop irrelevant sensor information from data set
2. Sensor output is smoothed (Kalman filtering)
3. Continuous and Boolean Features are generated to enrich data set
4. Data is abstracted for clustering and classification

**Raw Data:** 143 sensors per machine: unfiltered/ Mixed attribute/ temporal-Spatial

**Low Level Metric:** P<20 Filtered Features: GPS and Heading features, Engine Output, status of harvesting equipment and field boundary information

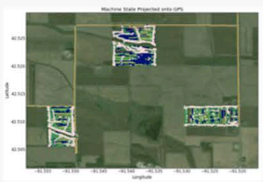
**Abstraction:** Similarity Transformation based on co-ocurance of sensor status and Minkowski norms

**High Level:** Projection of Estimated Machine operation State onto Satellite imagery of fields and roads

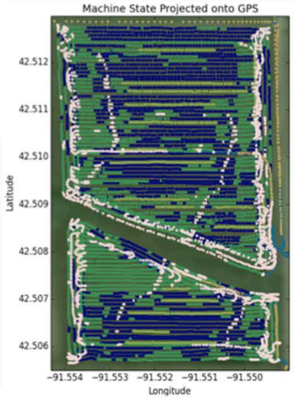
### Clustering Interpretation:

**Harvesting (higher Yield)**
**Intra Field Transport**
**Harvesting (lower Yield)**
**Harvesting and Unloading**
**Inter Field Transit**
**Maneuvering**

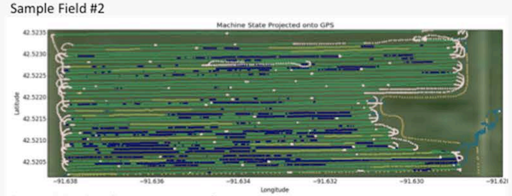
Fields and Roads



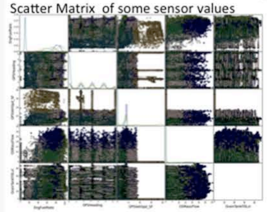

Sample Field #1




Sample Field #2




Scatter Matrix of some sensor values



### Next Steps

- Investigate class centroid sensitivity to different data sets (Other farms)
- Develop new features to better describe and distinguish machine states
- Form focus groups to discuss metric importance and machine state efficacy
- account for delay in Grain Mass flow Sensors

## Field Irregularity Metric

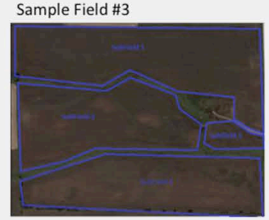
Frequently, fields can have irregular shapes and this makes it more difficult to plant, treat and harvest. A measure of field irregularity is therefore useful in quantifying the potential worth and productivity of a field.

**(Time Spent Maneuvering) / (Time Spent Harvesting)**
- ✓ Easy to compute given historical data for a field
- ✓ Metric skewed by efficiency of combine operator

**2-dimensional orthogonal packing of randomly sampled border polygons**
- ✓ Accounts for rectangular bias of crop rows
- ✓ Computationally Intensive
- ✓ Yields multiple metrics

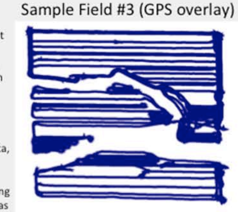
**Border Detection -> Fourier Border Irregularity + Hausdorff Distance between subfields**
- ✓ Accounts for inter & intra field complexity
- ✓ Computationally Intensive
- ✓ Arbitrary weighting of multiple metrics

Sample Field #3



Unfortunately, not all fields can be easily segmented into subfields with just the visible spectrum. However, when provided with historical GPS data, it is possible to circumvent this problem by treating GPS overlay data as a binary image.

Sample Field #3 (GPS overlay)



### Next Steps

The task of quantifying irregularity can be accomplished many ways and any single metric would be an arbitrarily weighted agglomeration of other metrics. Future work will involve comparisons of different methods for measuring irregularity to determine relative merits. Such work would have potential use in computer vision in general