Centaurus: Scalable Clustering for Agriculture Analytics

Nevena Golubovic, Chandra Krintz, Rich Wolski – University of California, Santa Barbara
Olivier Jerphagnon, Thomas Kuo, Kevin Langham – AgMonitor
Bo Liu – California State University, San Luis Obispo
Balaji Sethuramasamyraja – Fresno State University

UC SANTA BARBARA

AgMonitor

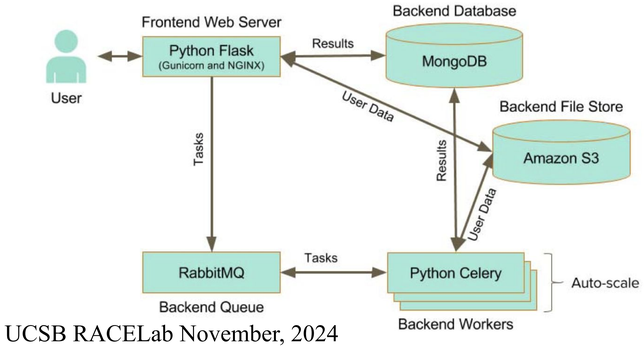
Motivation

- Cloud computing systems have evolved to be low cost, easy to use, and provide access to virtually unlimited compute power
- Recent advances in machine learning (ML) & data analytics have revolutionized e-commerce
- Unsupervised clustering is an effective form grouping similar observations with minimal human intervention and no model training or labeling
- Potential solutions to many agricultural problems are amenable to clustering
- The scale required to cluster accurately and the sheer number of clustering implementations make it difficult to use clustering correctly

Research Goals

- Develop an easy to use cloud service that implements clustering at scale
- Facilitate experimentation with different clustering implementations
- Develop a scoring metric that can be used to identify the "best" clustering implementation for a particular solution and dataset
- Develop new agricultural applications that leverage clustering to solve real problems that farming communities face

Centaurus Architecture



Centaurus Cloud Service

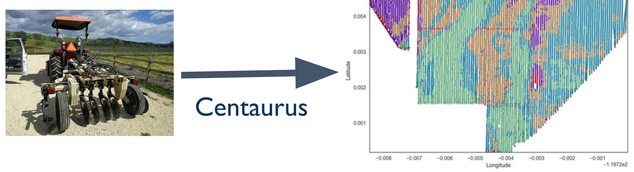
- Part of the UCSB SmartFarm effort
- Performs statistical clustering & model selection
- Upload data and select features to cluster
- Performs runs in parallel and across options
- Selects the best option using Bayesian information criterion (BIC), validated against groundtruth
- Visualizes the results for users

Key Findings

- Accurate validation against datasets with known clusters
- Best cluster is rare – requiring thousands of parallel runs. Most commonly found solution is not always the best
- Degenerate clusters must be removed

Source: UCSB. (NSF CCF-1539586)

Precision Agricultural Zone Management

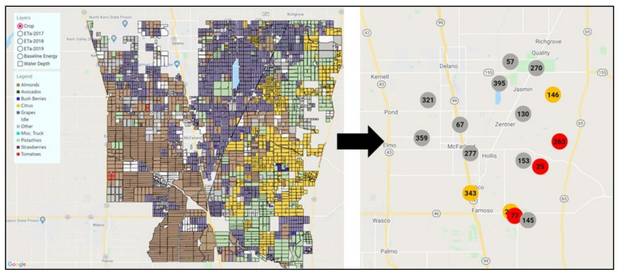


- Goal: Delineate sections of farmed blocks with similar soil characteristics which can vary widely across a field to enable growers to apply water & nutrients in similar areas more precisely
- Cluster based on electrical conductivity data that is fast and inexpensive to collect; use as an estimate for soil health
- Validate using soil core sample analysis
- Provide visualization tool for boundary identification of similar blocks and field regions. IT significantly outperforms commonly used tools such as Management Zone Analyst (MZA) tool from USDA

Comparison: Percent Error

	Dataset-1	*Dataset-2*	*Dataset-3*
Centaurus	0.0%	3.6%	0.1%
MZA	0.0%	13.8%	11.6%

Estimating Farmed Areas at Risk from Energy Use and Satellite Data

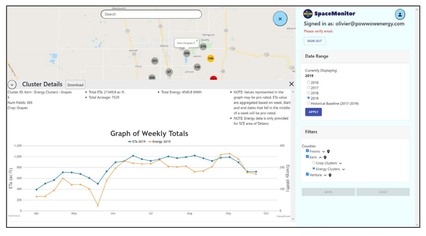


Results

- Alert level classes identify differences in irrigation practices for the farms studied
- Validation at pump level with growers (pilot sites) and basin level with districts

- Goal: Identify ranches facing similar water and energy effiency challenges and apply incentives for remediation (pump retrofits, crop change, groundwater recharge, etc.)
- Evapotranspriation (ETa) is an estimate of crop water consumption from Formation Environmental (private) or OpenET (public)
- Cluster ranches by crop type, water table level, and location for alerts & incentives
- Compute average energy use of clusters from "groundtruth" meters and estimate ETa using data and pump model
- Compare to validated model; find anomalies
- Share findings with growers and districts

SpaceMonitor



Findings

- Clustering across data layers can help identify risky areas
- Solar penetration can also be optimized (e.g., Westlands)

Source: AgMonitor (CEC EPC-16-051)