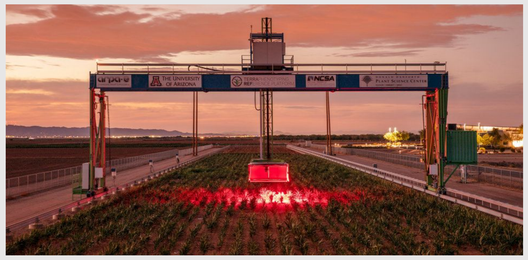
PhytoOracle: Modular, Scalable Phenomics Data Processing Pipelines

Emmanuel Gonzalez*[1] (emmanuelgonzalez@email.arizona.edu), Ariyan Zarei,[2] Nathanial Hendler,[1] Michele Cosi,[1] Jeffrey Demieville,[1] Travis Simmons,[1,3] Holly Ellingson,[4] Nirav Merchant,[4] Eric Lyons,[1,4] Duke Pauli,[1,4] and Andrea Eveland[5]

[1]School of Plant Sciences, University of Arizona, Tucson, AZ; [2]Department of Computer Sciences, University of Arizona, Tucson, AZ; [3]College of Coastal Georgia, Brunswick, GA; [4]Data Science Institute, University of Arizona, Tucson, USA; and [5]Donald Danforth Plant Science Center, St. Louis, MO

Introduction

Phenomics generates large, high-dimensional datasets, making data processing computationally intensive and time-consuming. PhytoOracle addresses these bottlenecks by reducing processing time and generating reproducible results by leveraging current computational tools, such as distributed workflows, container technology and high performance computing (HPC).



(The Wall Street Journal)

PhytoOracle was developed for processing data collected by the world's largest plant phenotyping platform, the Field Scanalyzer, which is located at the University of Arizona's Maricopa Agricultural Center (pictured above). It is equipped with a diverse set of sensors able to capture a variety of phenotypic data with a maximum generation of 10 terabytes a day. PhytoOracle efficiently quantifies phenotypic information for subsequent studies aimed at identifying genetic components associated with abiotic stress tolerance. Data extraction is achieved using multiple machine learning algorithms that detect and track individual plants throughout the growing season, allowing for the precise scoring of stress phenotypes such as photosynthetic efficiency and growth rate. A typical experiment consists of multiple scans over a growing season, capturing 30,000 plants per scan. This level of data capture requires distributed computing to extract phenotypic trait data efficiently.

Efficiently processing multimodal data

Data processing steps:

- Pre-processing
- Orthomosaic using MegaStitch
- Clip images to agricultural plots
- Plant detection and phenotype quantification
- Clustering of time-series data

Pipeline outputs:

- Multimodal time-series phenomic datasets
  - Laser 3D, RGB, thermal, and PSII chlorophyll fluorescence
- Large processed image datasets for various ML applications

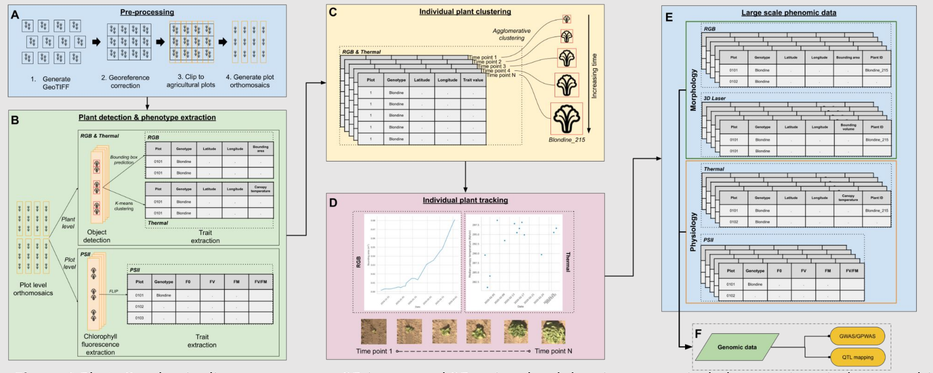


Figure 1 PhytoOracle pipelines process raw 2D image and 3D point cloud data into numerical phenotypic trait data, resulting in large, time-series phenomic data.

PhytoOracle leverages CCTools' Makeflow and Workqueue frameworks to distribute processing tasks across thousands of workers on high performance computer (HPC), Cloud, and/or local compute resources. Tasks are executed in parallel, allowing efficient processing of various data types across thousands of compute nodes. PhytoOracle processing times at a maximum of 1,024 workers (Fig. 2):

- 235 minutes for 9,270 RGB images (140.7 GB)
- 235 minutes for 9,270 thermal images (5.4 GB)
- 13 minutes for 39,678 PSII images (86.2 GB)

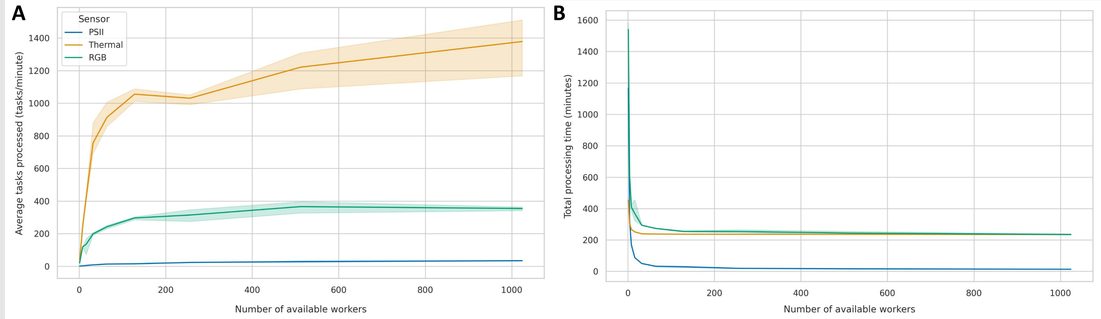
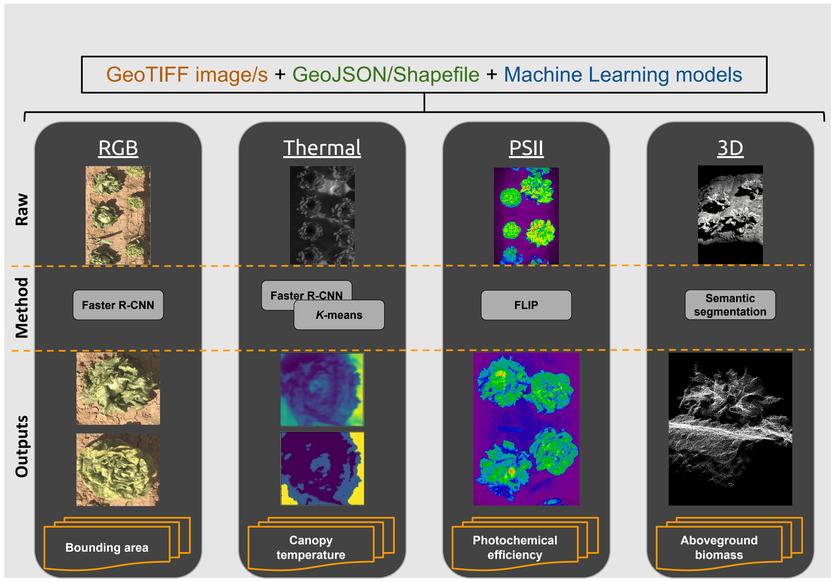


Figure 2 Processing times for PSII, thermal, and RGB data. (A) Average tasks processed with increasing workers. (B) Total processing time in minutes with increasing workers. Benchmark datasets for each sensor were processed over the following range of available workers: 1, 4, 8, 16, 32, 64, 128, 256, 512, and 1024. Each configuration was run three times, for a total of 30 benchmark observations.

Pipeline Overview



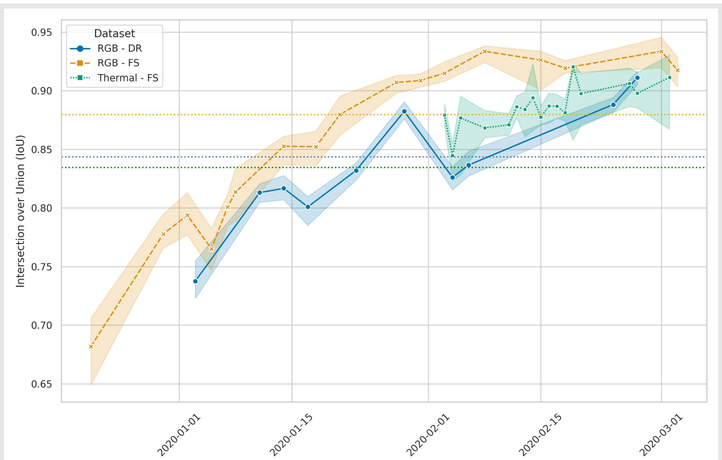
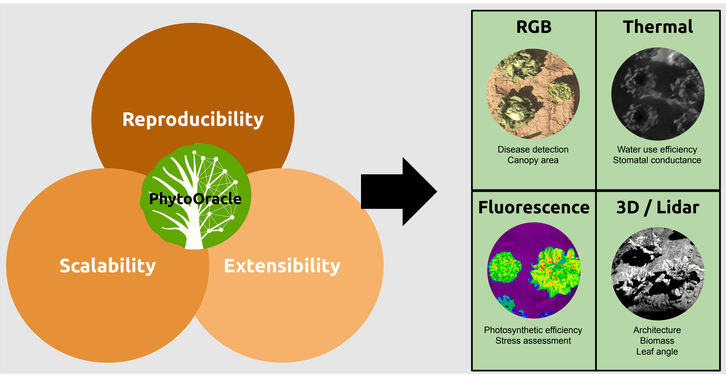


Figure 3 Change in median IoU across the collection dates represented in our test sets. RGB collections, both Field Scanalyzer scans (RGB-FS) and drone flights (RGB-DR), began earlier than thermal, allowing us to capture the temporal effect of collection date, a proxy to plant size, on the median IoU. Error bands represent a 95% CI around the median.

Summary & Documentation



Scan the QR code to access our GitHub repository and documentation

Acknowledgements

U.S. Department of Energy | Office of Science
College of Agriculture & Life Sciences
University of Notre Dame
arpa·e
Donald Danforth Plant Science Center
CyVerse