Big Data Plant Phenomics

Emmanuel Gonzalez and Travis Simmons

Emerging technologies produce a lot of data

Robots

(Wall Street Journal, LemnaTec)

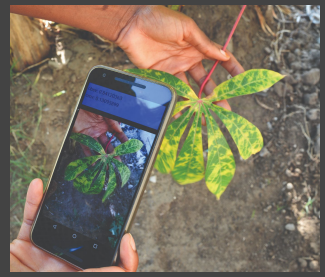
Carts

(USDA)

Drones

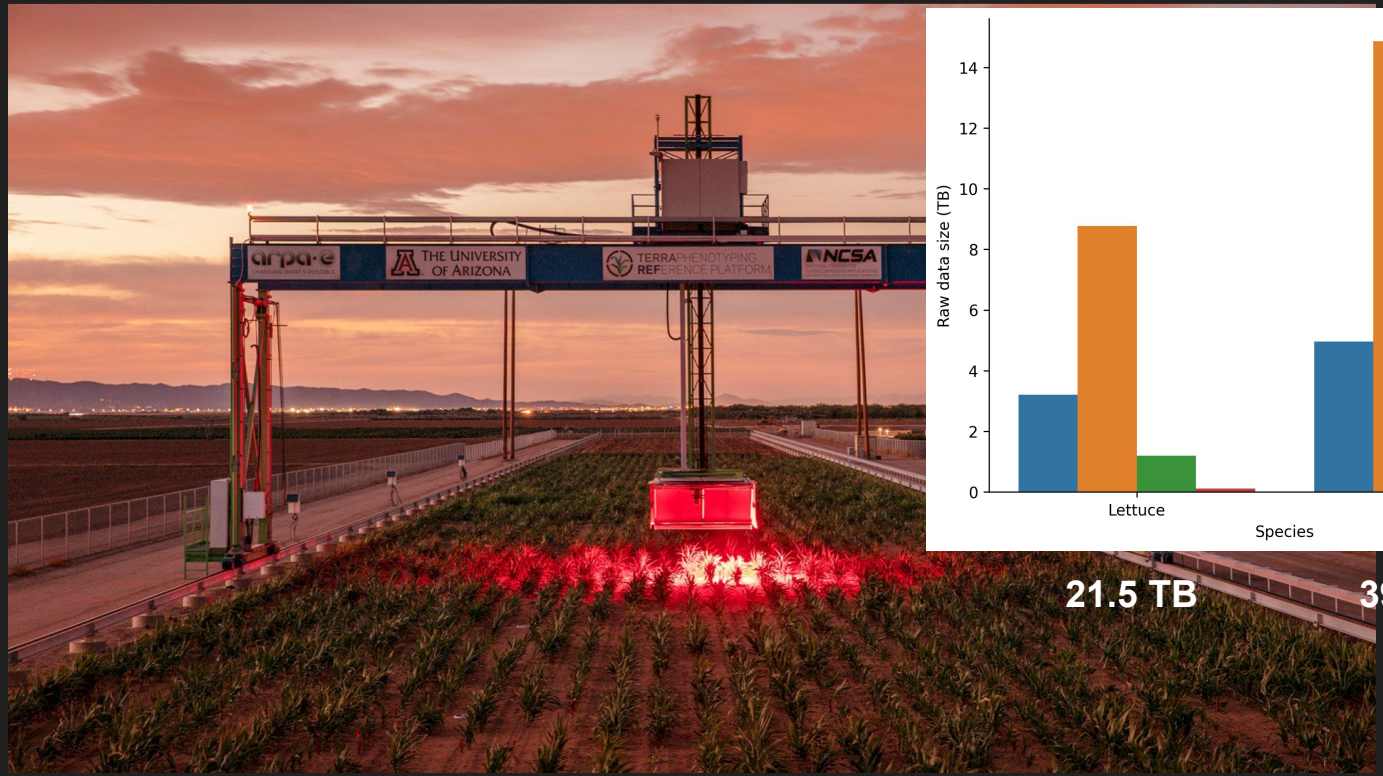
(DJI)

Phones



(IITA)

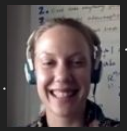
Increasing data volumes necessitate scalable frameworks

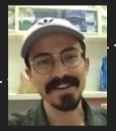


The UA team


Dr. Pauli

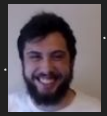

Dr. Hendler


Holly

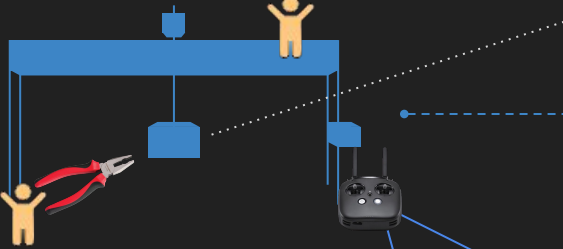

Emmanuel


Michele


Ariyan


Travis

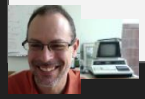
PhytoOracle



Cache Server


Sean

CYVERSE™ DataStore


Dr. Lyons


Bob


Jeffrey


Sebastian


Bruno

DONALD DANFORTH PLANT SCIENCE CENTER

SAINT LOUIS UNIVERSITY™

GW

General-use frameworks for generalizability

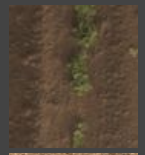
GeoTIFF image/s + GeoJSON/Shapefile + PyTorch Faster RCNN model

Input

Output

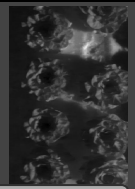
RGB

Faster R-CNN



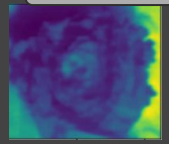
Bounding area

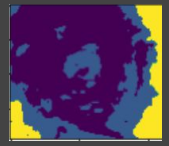
Thermal



Faster R-CNN

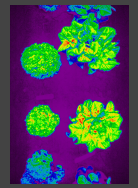
K-means



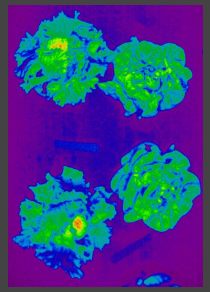


Canopy temperature

PSII



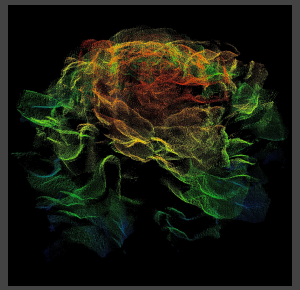
FLIP



F_v/F_m

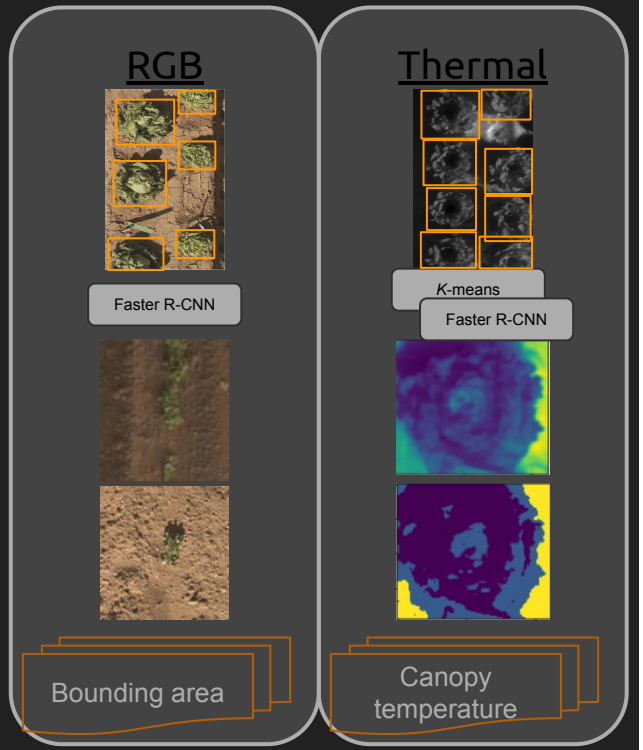
3D Laser

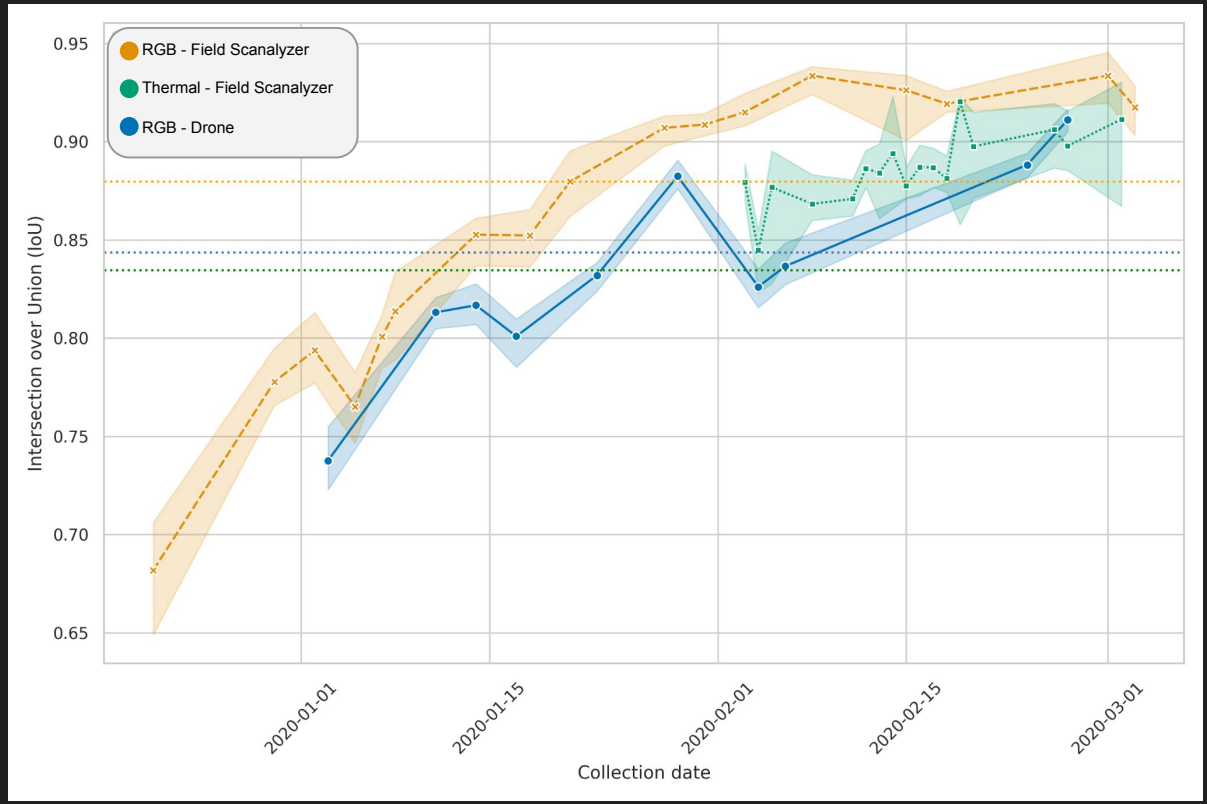
Semantic segmentation



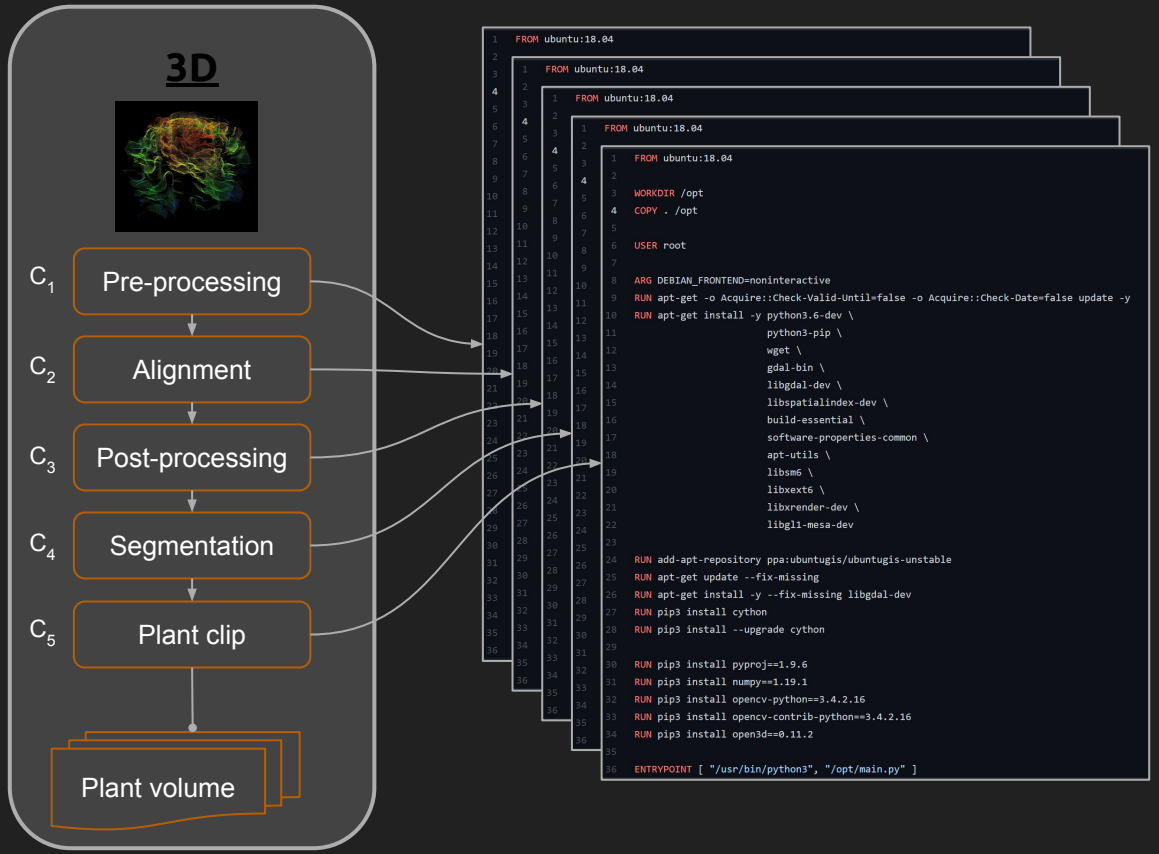
Plant volume

General-use frameworks for generalizability





Focus on reproducible science

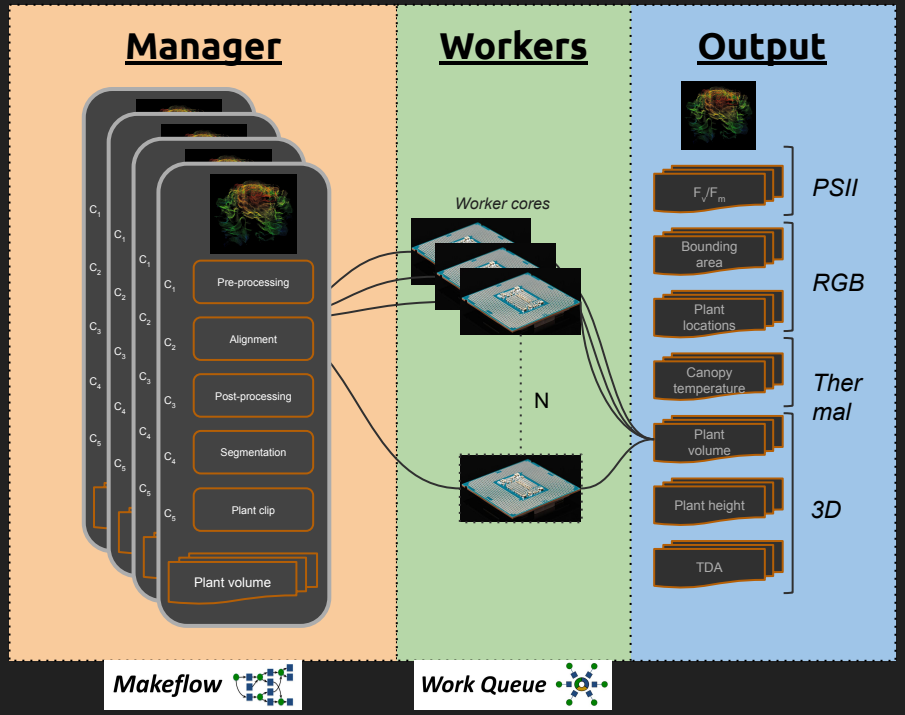


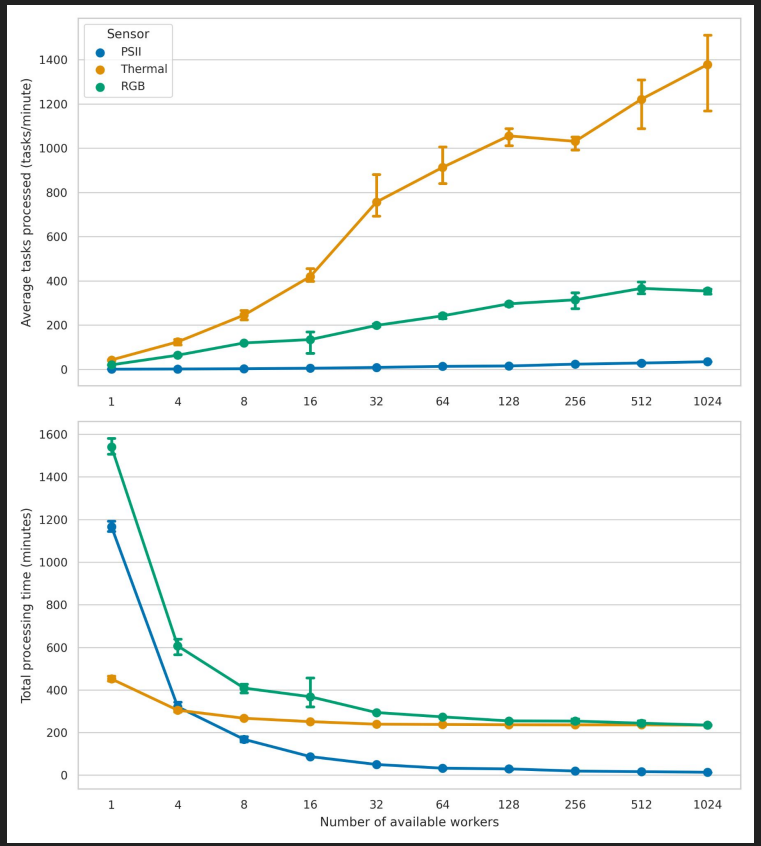
Computational tools

Docker containers provide **extensibility** and **reproducibility**

Singularity enables **distributed** processing on high performance computer (**HPC**) clusters

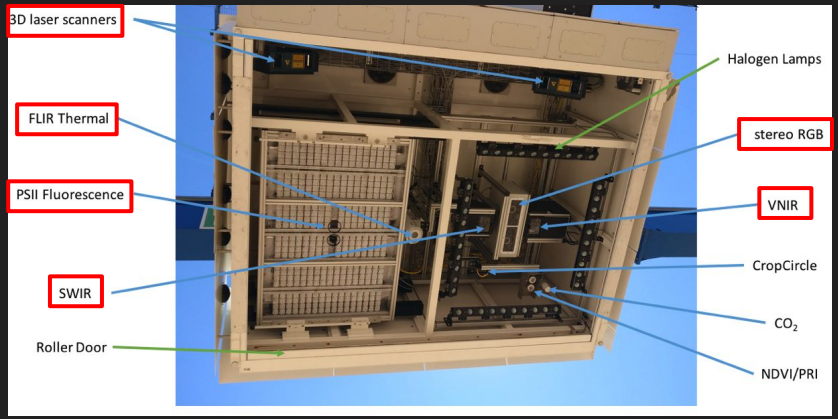
Leveraging distributed processing



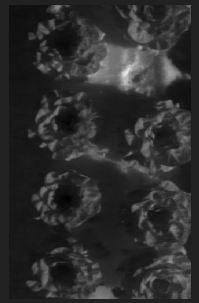


Processing a whole lot of data

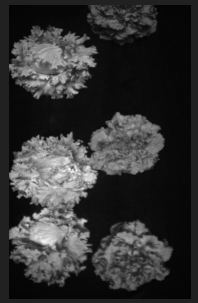
- The world's biggest scanalyzer
- Data volume
  - Max capacity of 10 TB/day
  - Typical performance of 1.5 TB/day



RGB
~550GB

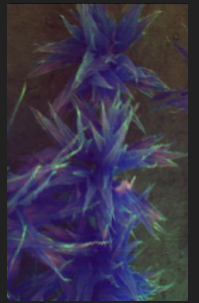


Thermal
~5.5GB



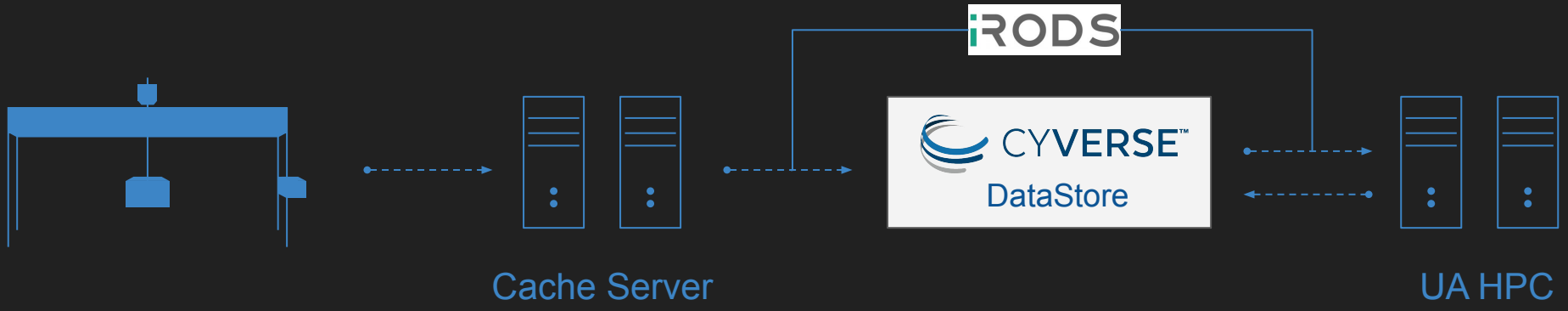
Fluorescence
~80GB

3D
~300GB



Hyperspectral
~600GB

Data transfer



- Collect data
- Compression
- Checksums
- Storage (raw + processed)
- Processing

How much time would it take to process* a single season worth of RGB data (50TB) on a 4-core, regular lab computer?

* From raw data to a quantifiable phenotype

How much time would it take to process* a single season worth of RGB data (50TB) on a 4-core, regular lab computer?

55 years!

* From raw data to quantifiable phenotypes

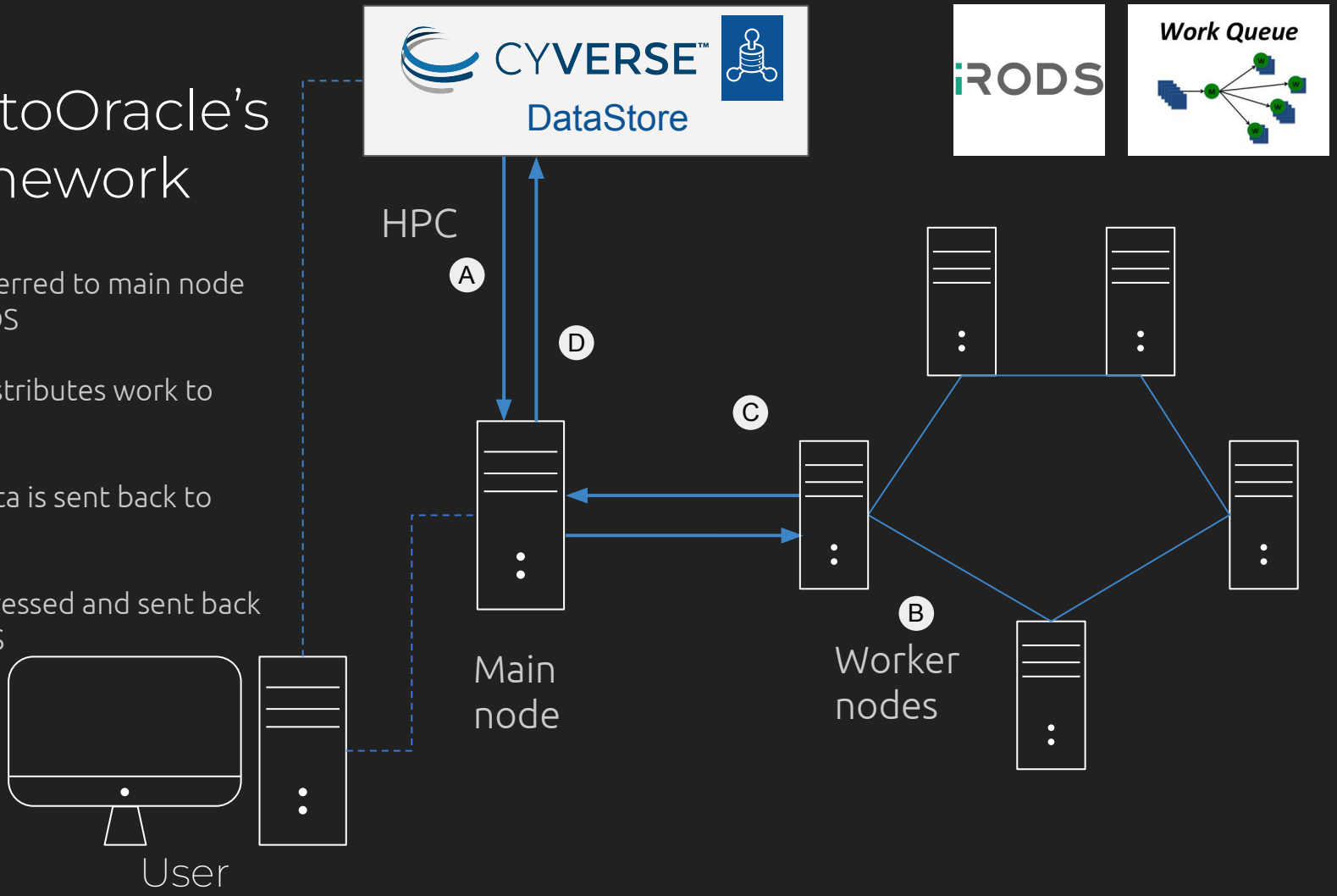
PhytoOracle's framework

A. Data is transferred to main node through iRODS

B. Main node distributes work to worker nodes

C. Processed data is sent back to main node

D. Data is compressed and sent back to CyVerse DS



How much time would it take PhytoOracle to process* a single season worth of RGB data (50TB)?

* From raw data to quantifiable phenotypes

How much time would it take PhytoOracle to process* a single season worth of RGB data (50TB)?

Only 6 days!

* From raw data to quantifiable phenotypes