The background of the slide features a photograph of a vast agricultural field. The field is organized into long, narrow plots, each containing a dense crop, likely corn or similar grain. In the distance, several tall, light-colored industrial towers or storage tanks are visible against a clear sky. A blue metal walkway or bridge structure crosses the field in the middle ground.

Big Data Plant Phenomics

Emmanuel Gonzalez and Travis Simmons

Emerging technologies produce a lot of data

Robots



Carts



Drones



Phones



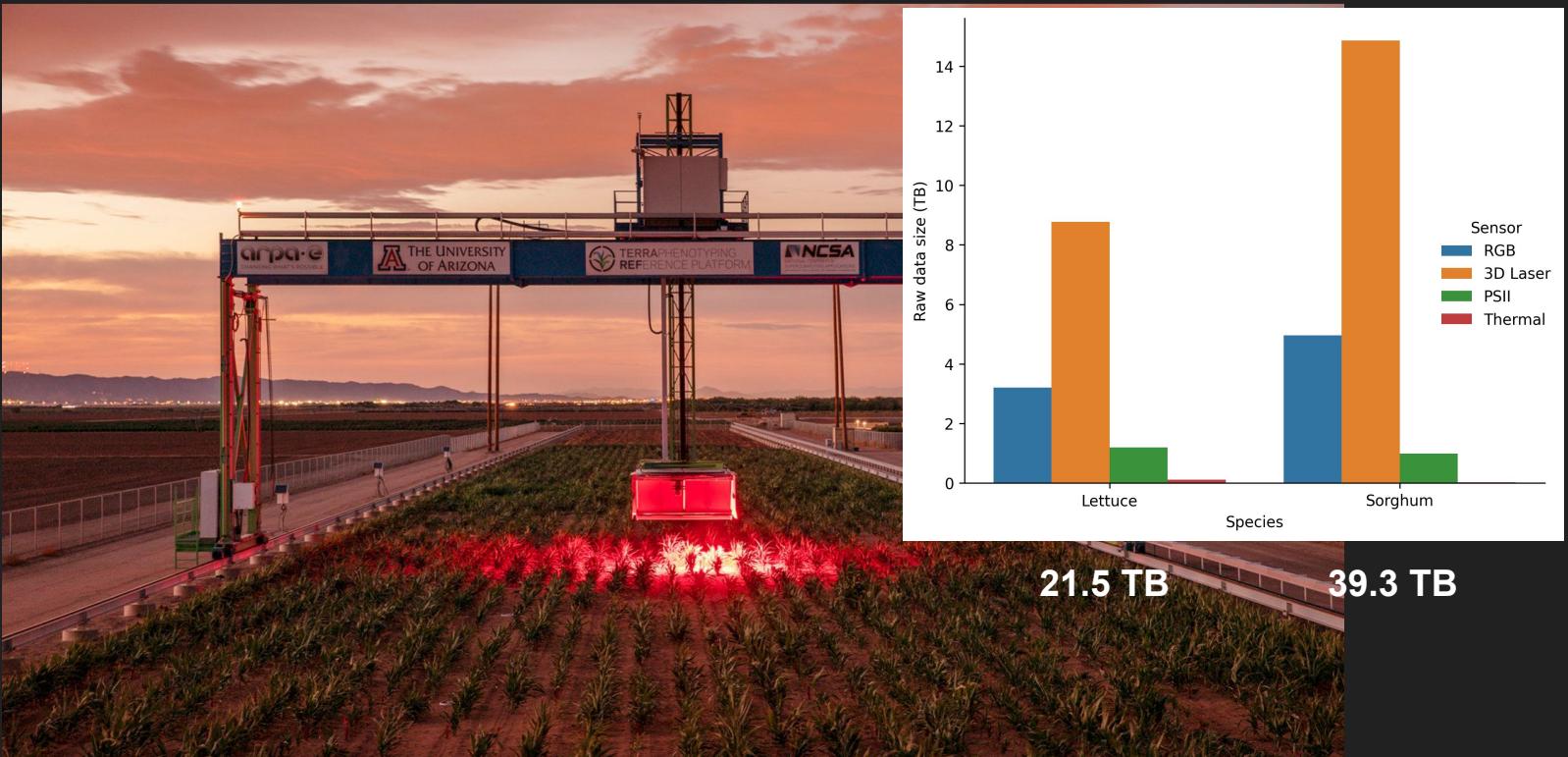
(Wall Street Journal, LemnaTec)

(USDA)

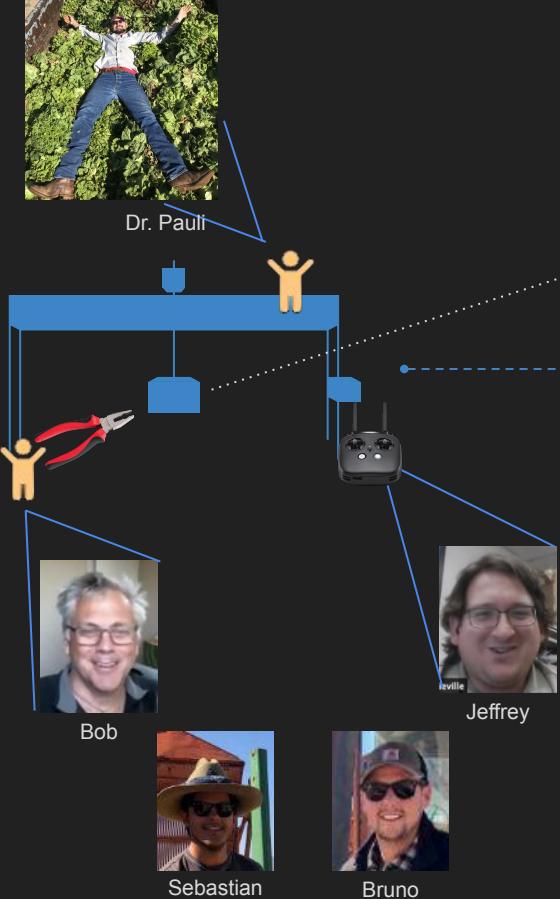
(DJI)

(IITA)

Increasing data volumes necessitate scalable frameworks

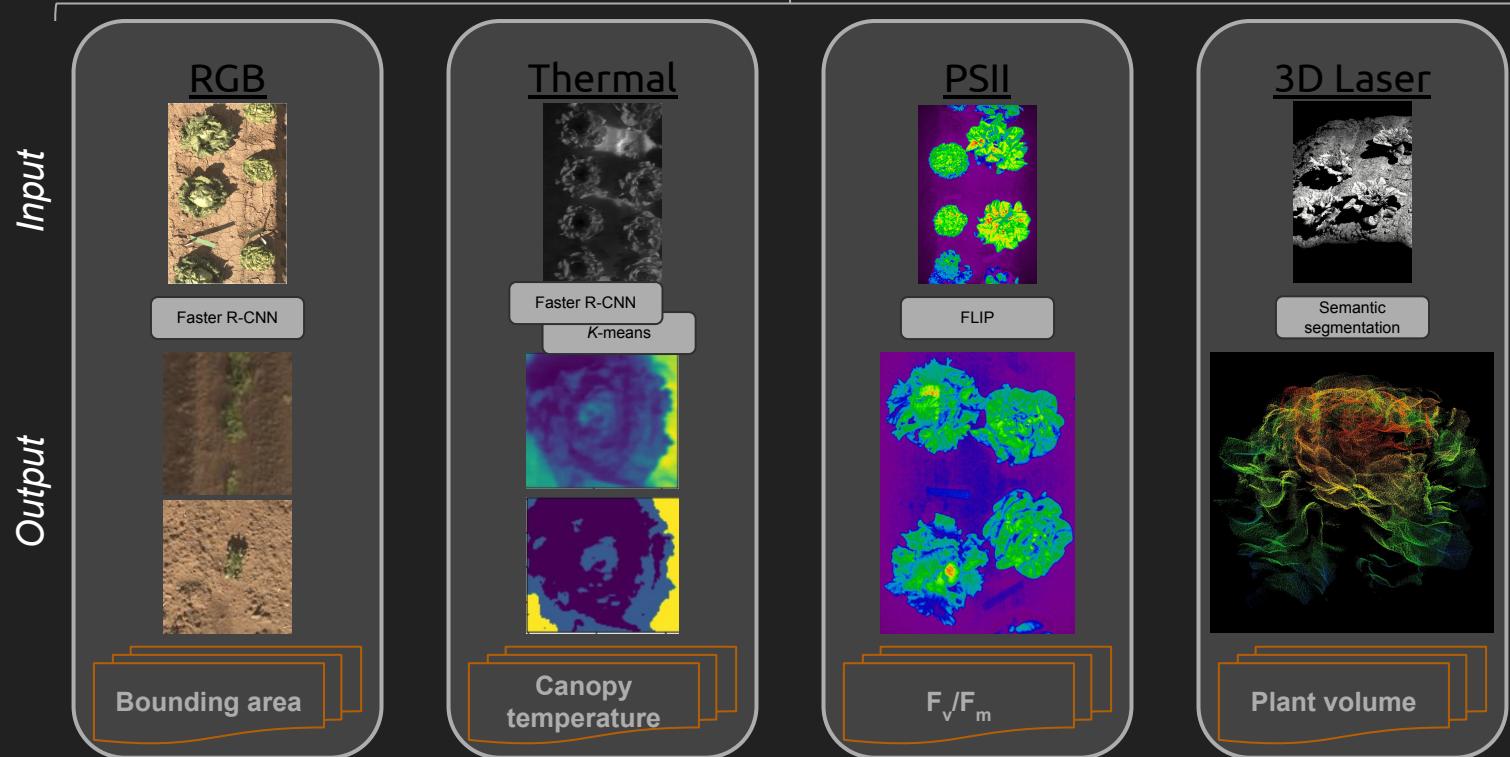


The UA team

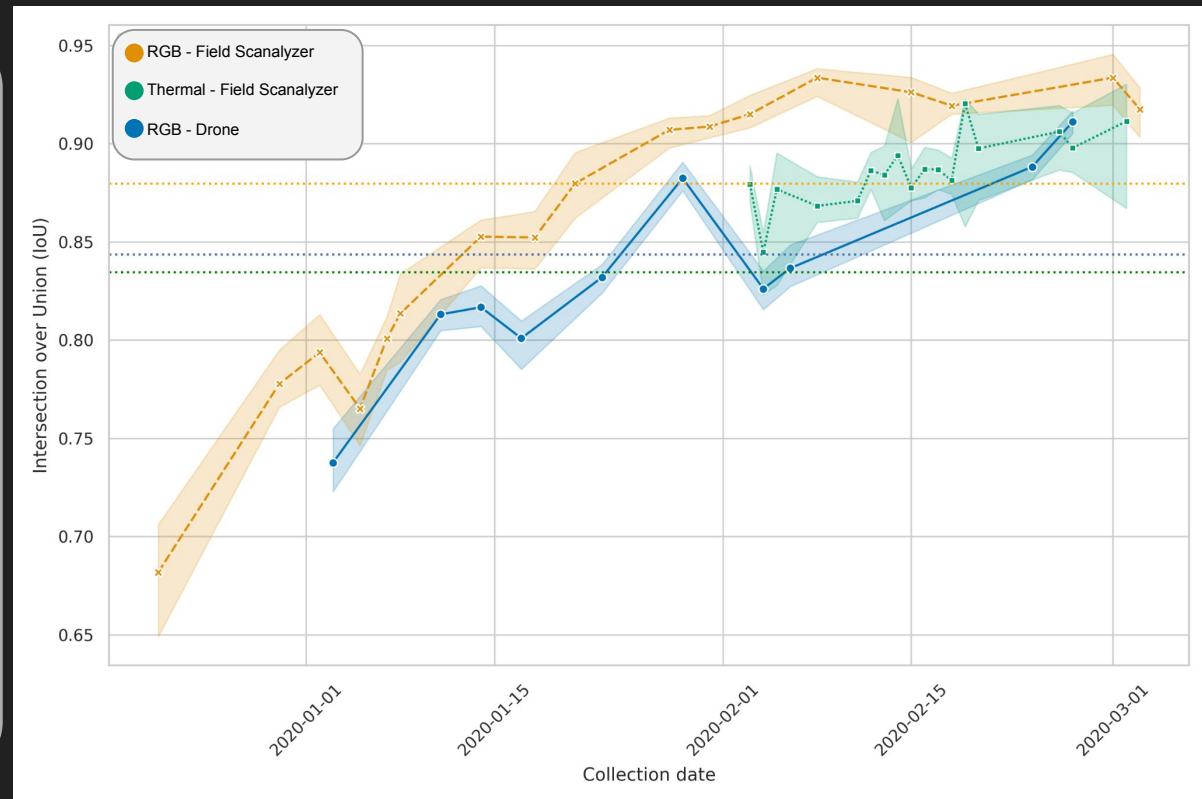
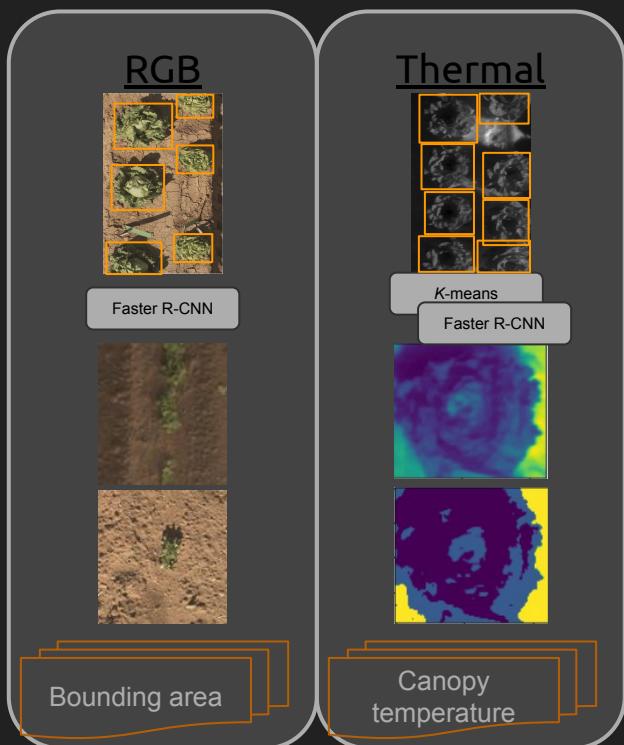


General-use frameworks for generalizability

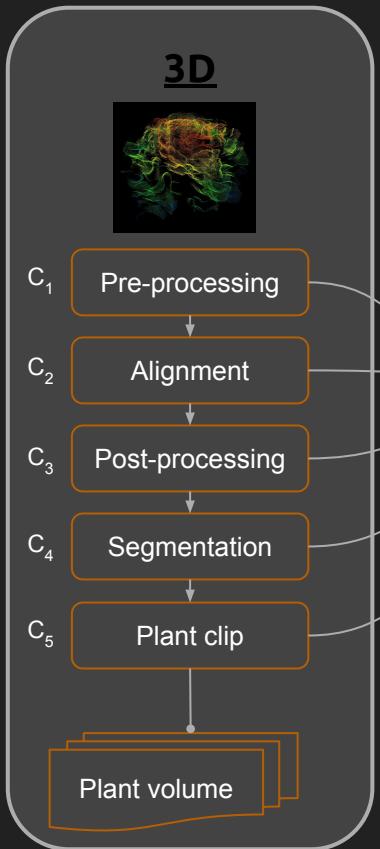
GeoTIFF image/s + GeoJSON/Shapefile + PyTorch Faster RCNN model



General-use frameworks for generalizability



Focus on reproducible science



```
1  FROM ubuntu:18.04
2
3  1  FROM ubuntu:18.04
4
5  2  FROM ubuntu:18.04
6
7  3  1  FROM ubuntu:18.04
8
9  4  2  WORKDIR /opt
10 5  3  COPY . /opt
11 6  4  USER root
12
13 7  5  ARG DEBIAN_FRONTEND=noninteractive
14 8  6  RUN apt-get -o Acquire:::Check-Valid-Until=false -o Acquire:::Check-Date=false update -y
15 9  7  RUN apt-get install -y python3.6-dev \
16 10   python3-pip \
17 11   wget \
18 12   gdal-bin \
19 13   libgdal-dev \
20 21   libspatialindex-dev \
21 22   build-essential \
22 23   software-properties-common \
23 24   apt-utils \
24 25   libsm6 \
25 26   libxext6 \
26 27   libxrender-dev \
27 28   libgl1-mesa-dev
28
29 29  RUN add-apt-repository ppa:ubuntugis/ubuntugis-unstable
30 30  RUN apt-get update -fix-missing
31 31  RUN apt-get install -y --fix-missing libgdal-dev
32 32  RUN pip3 install cython
33 33  RUN pip3 install --upgrade cython
34 34  RUN pip3 install pyproj==1.9.6
35 35  RUN pip3 install numpy==1.19.1
36 36  RUN pip3 install opencv-python==3.4.2.16
37 37  RUN pip3 install opencv-contrib-python==3.4.2.16
38 38  RUN pip3 install open3d==0.11.2
39
40 39  ENTRYPOINT [ "/usr/bin/python3", "/opt/main.py" ]
```

Computational tools

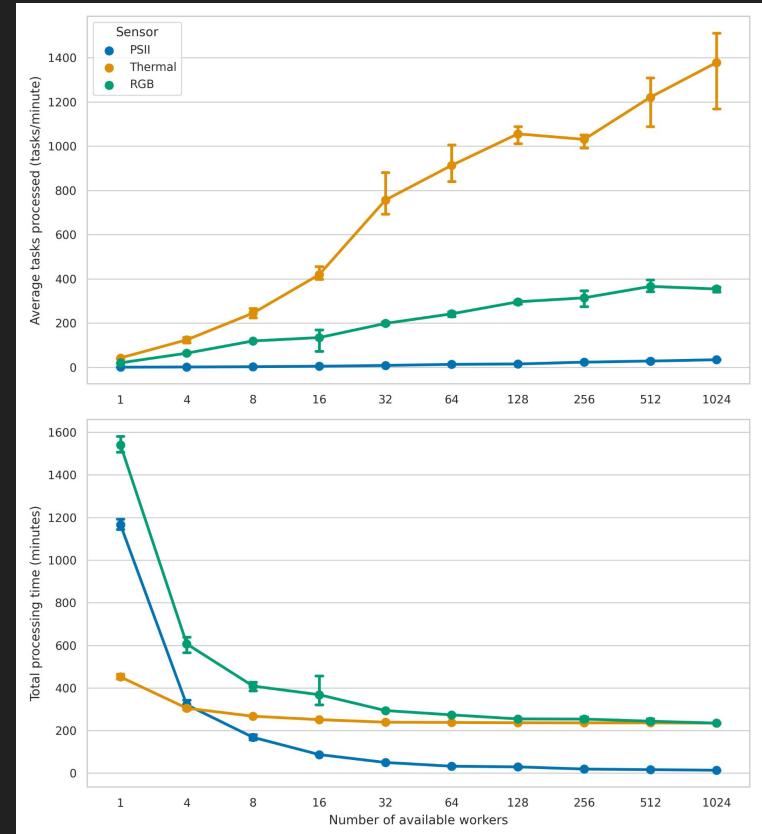
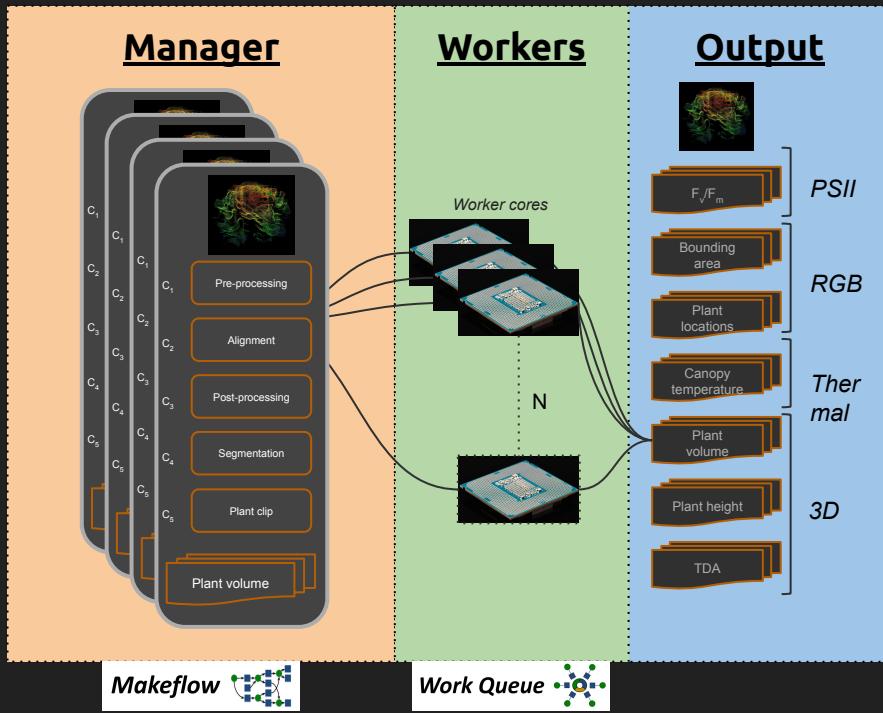


Docker containers provide **extensibility** and **reproducibility**



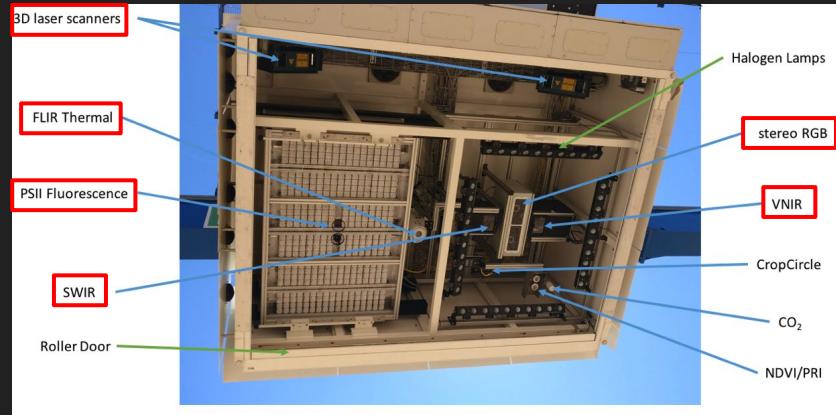
Singularity enables **distributed** processing on high performance computer (**HPC**) clusters

Leveraging distributed processing

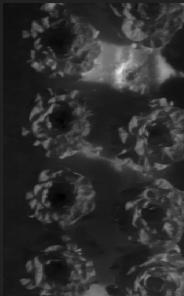


Processing a whole lot of data

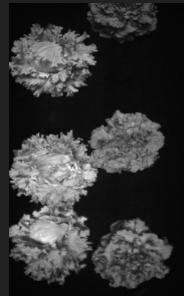
- The world's biggest scanalyzer
- Data volume
 - Max capacity of 10 TB/day
 - Typical performance of 1.5 TB/day



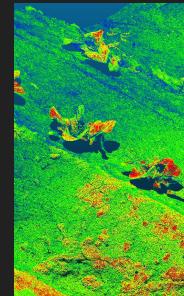
RGB
~550GB



Thermal
~5.5GB



Fluorescence
~80GB

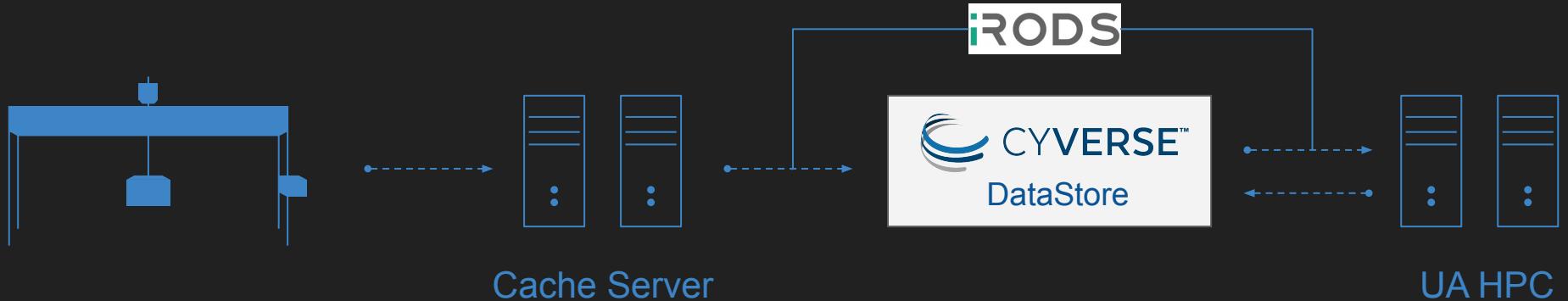


3D
~300GB



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?

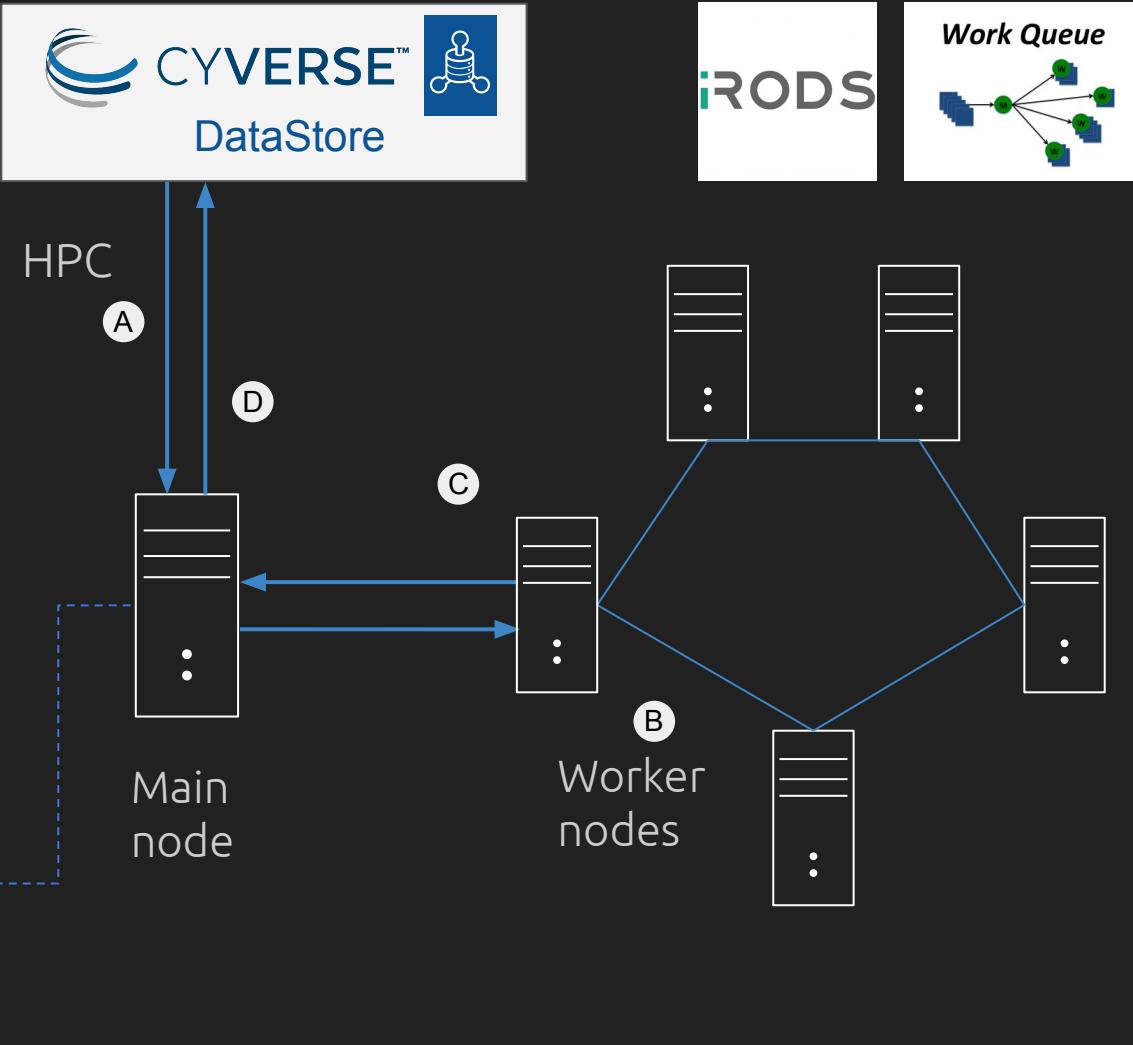
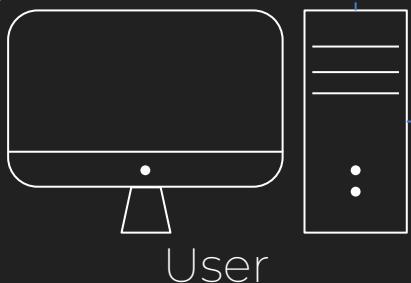
55 years!

* 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