PhytoOracle

THE UNIVERSITY OF ARIZONA

THE UNIVERSITY OF ARIZONA
RESEARCH, DISCOVERY & INNOVATION
Data Science
Institute

Managing the Machine Learning Lifecycle with MLflow:

A Tech Preview Using PhytoOracle (and chest x-ray)

Artin Majdi UArizona, ECE Dept., Data Science Institute (Data7)

Ariyan Zarei UArizona, CS Dept., PhytoOracle

April 16, 2021

Personal experience

- Organization/Tracking
- Platform dependence
- Accessing old simulations
- Deployment

What We'll Cover Today

- Challenges in ML development
- How MLflow can help
- What is MLflow?
- Tech Preview with Case Studies

Typical ML Project Requirements (MLOps)

Data

Ethical fairness
Pre and post processing

Accessibility

Development

Design

Agnosticism & Reproducibility

Versioning & Tracking
(experiment, code, dependencies)

Deployment

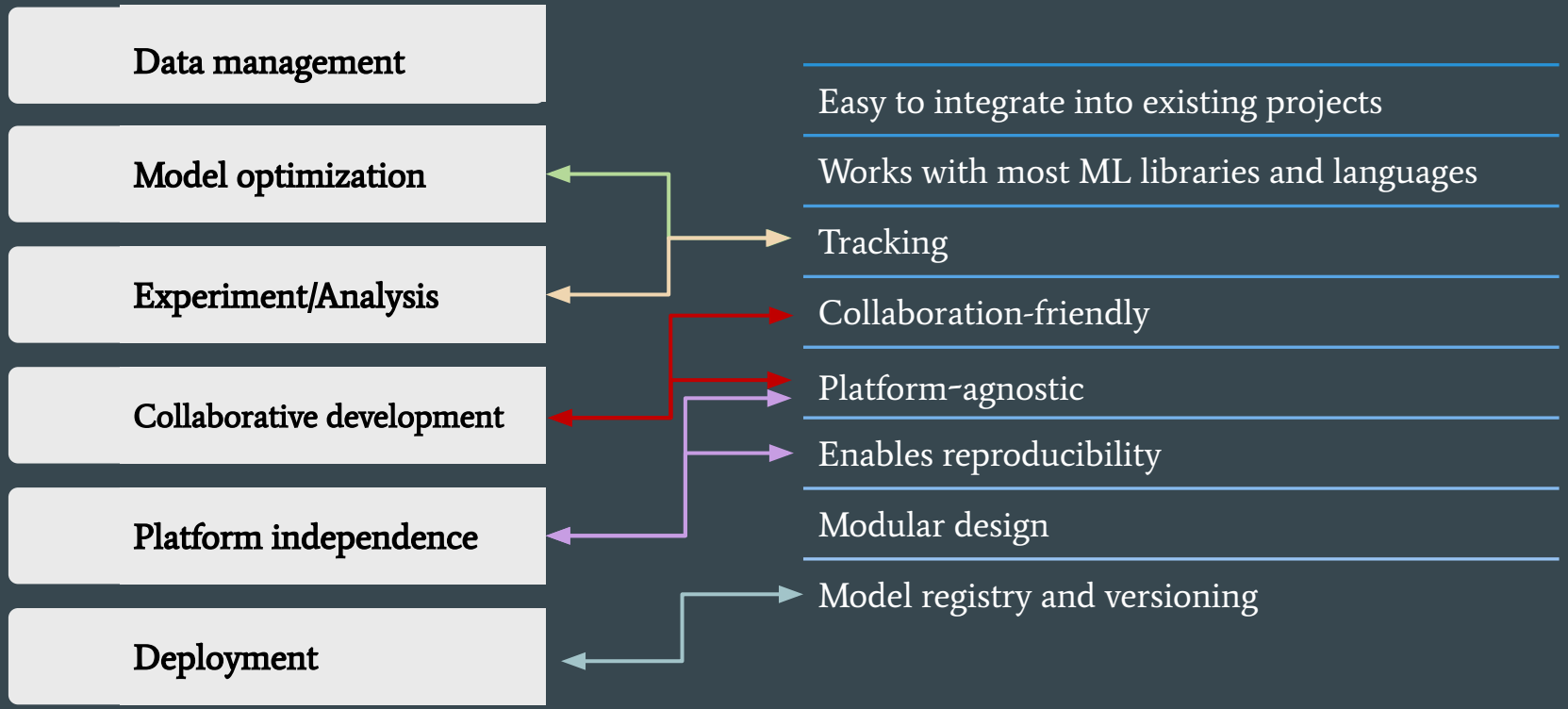
Continuous Monitoring

Multiple access mode
Visualization

Challenges in Development

Data management	Pre-processing Accessibility
Model optimization	Architecture design Hyper parameter tuning
Experiment/Analysis	Convergence Hyper parameter tuning
Collaborative development	Data/model accessibility Permission assignments
Platform dependence	OS Versioning conflicts

How MLflow Can Help with These Challenges



MLflow Components

Tracking

Record experiments config, results and sources code

Models

Standardized format for saving models

Projects

Reproducible packaging

Model registry

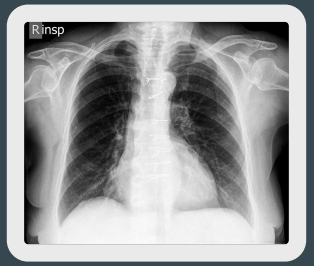
Centralized model management review & sharing

Plugins

Framework agnostic tool for ML

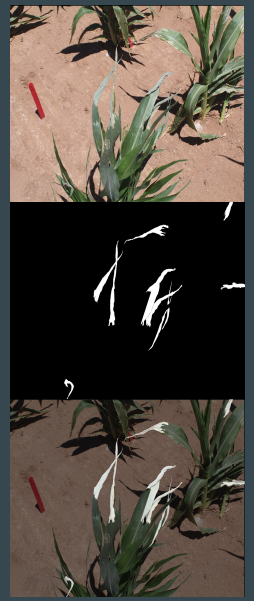
ML Case Studies

Chest X-ray

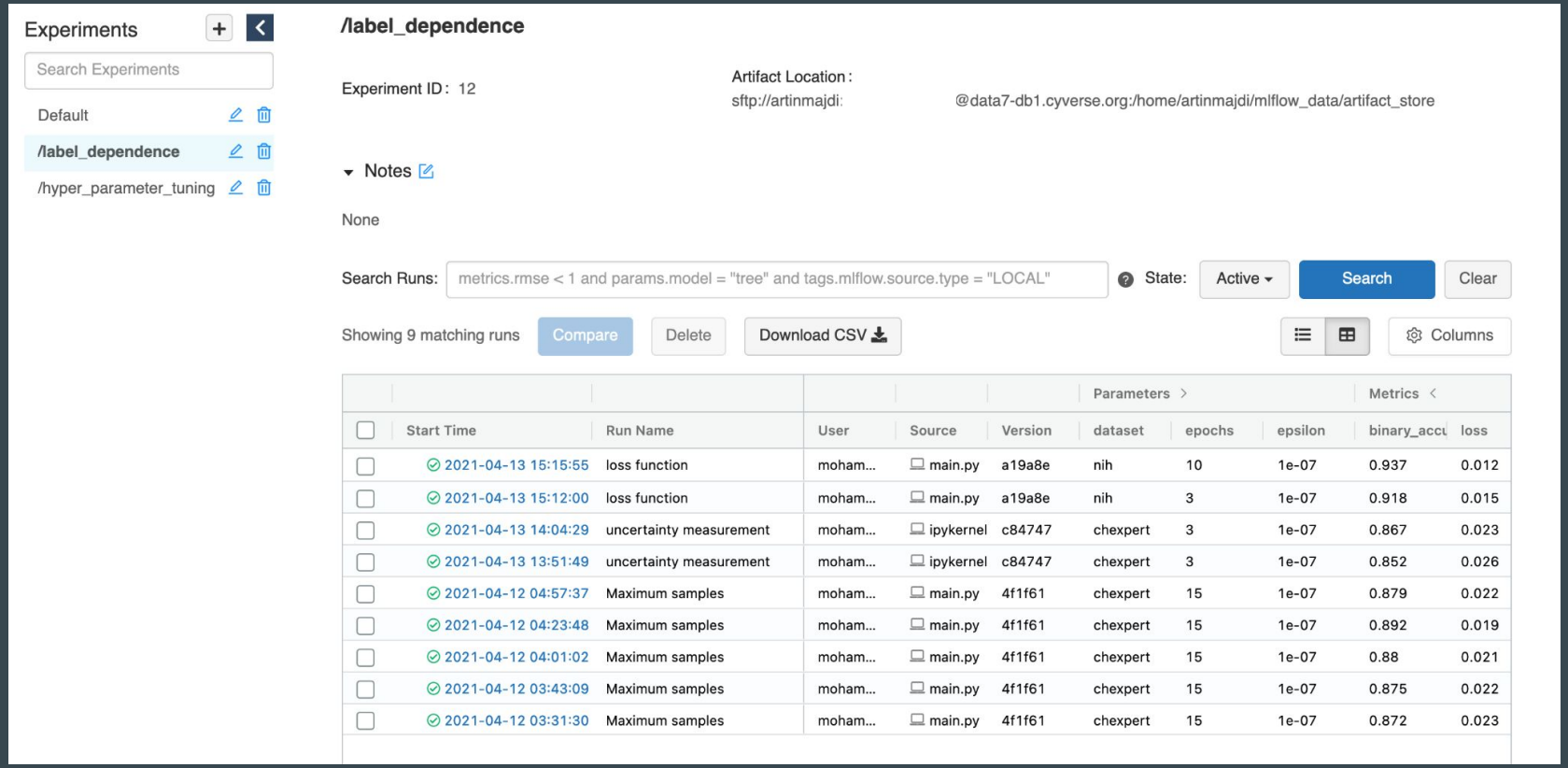


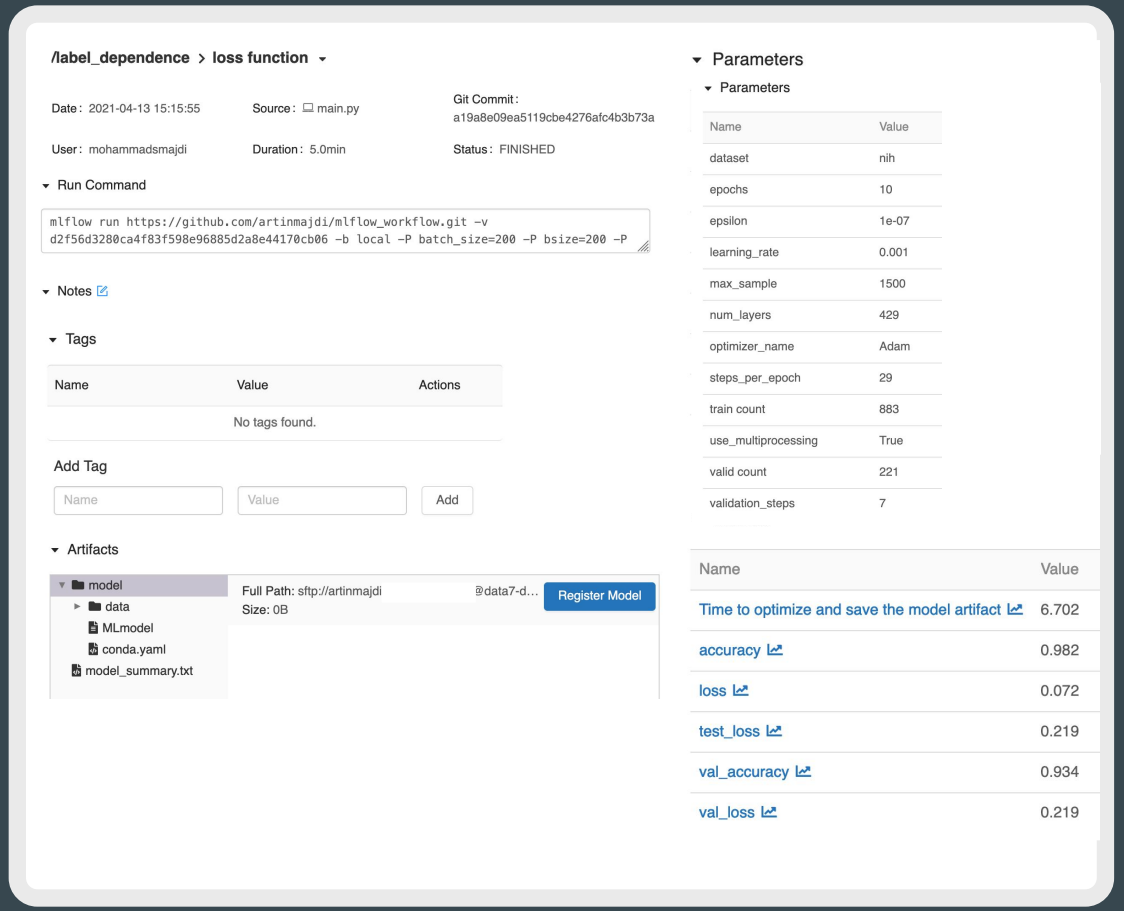
Pathology classification	Goal	Segmentation of plant disease
NIH and CheXpert	Data	Manually Collected/Labeled
X-ray	Data type	RGB Images
Python, Keras (TensorFlow)	Language	Python, Keras (TensorFlow)
Pandas, scikit-image,...	Dependencies	OpenCV, PIL, ...
Local (CPU) & HPC (GPU)	Processing	HPC (GPU)

PhytoOracle



MLflow User Interface for Chest Classification





Tracking Key Features

Parameters

Metrics

Tags and notes

Artifacts

Source code

How to Log Parameters/Metrics with MLflow

Automatic tracking of endless text/csv/pickle output files!

Logging parameters/metrics/artifacts

```
Run Cell | Run Above | Debug Cell
# %% ------------------------------------------------
""" Saving MLflow parameters & metrics """
mlflow.log_param("epochs", epochs)
mlflow.log_param("batch_size", batch_size)
mlflow.log_metric("accuracy", test_acc)
mlflow.log_metric("test_loss", test_loss)

mlflow.keras.log_model(model, "my_model_log")
mlflow.keras.save_model(model, 'my_model')

with open('predictions.txt', 'w') as f:
    f.write("predicted_classes")

mlflow.log_artifact('predictions.txt')
```

Using mlflow built-in automatic logging

```
""" Logging the parameters automatically """
mlflow.keras.autolog()
```

MLflow Project Structure

Environment

Entry point

```
name: My Project

conda_env: my_env.yaml                    1

docker_env:
   image:  mlflow-docker-example          2

entry_points:

  main:
    parameters:
      data_file: path
      regularization: {type: float, default: 0.1}

    command: "python train.py -r {regularization} {data_file}"

  validate:
    parameters:
      data_file: path

    command: "python validate.py {data_file}"
```

Model Registry

Audience

mlflow Experiments Models GitHub Docs

Registered Models

search model name

Name	Latest Version	Staging	Production	Last Modified
classification-mnist	Version 3	Version 2	Version 1	2021-02-04 13:40:08
classifier	Version 1	–	–	2021-02-14 21:58:05
model_A	Version 5	–	Version 4	2021-02-01 18:54:33

Page 1

MLflow on CyVerse

Tracking Server

- CyVerse Cloud Native Service

server = f'{dialect-driver}://{username}:{password}@{host}:{port}/{database-name}'

```
mlflow.set_tracking_uri(server)

mlflow.set_registry_uri(server)
```

- MySQL
- SQLite
- PostgreSQL

Artifact Storage

- CyVerse Data Store

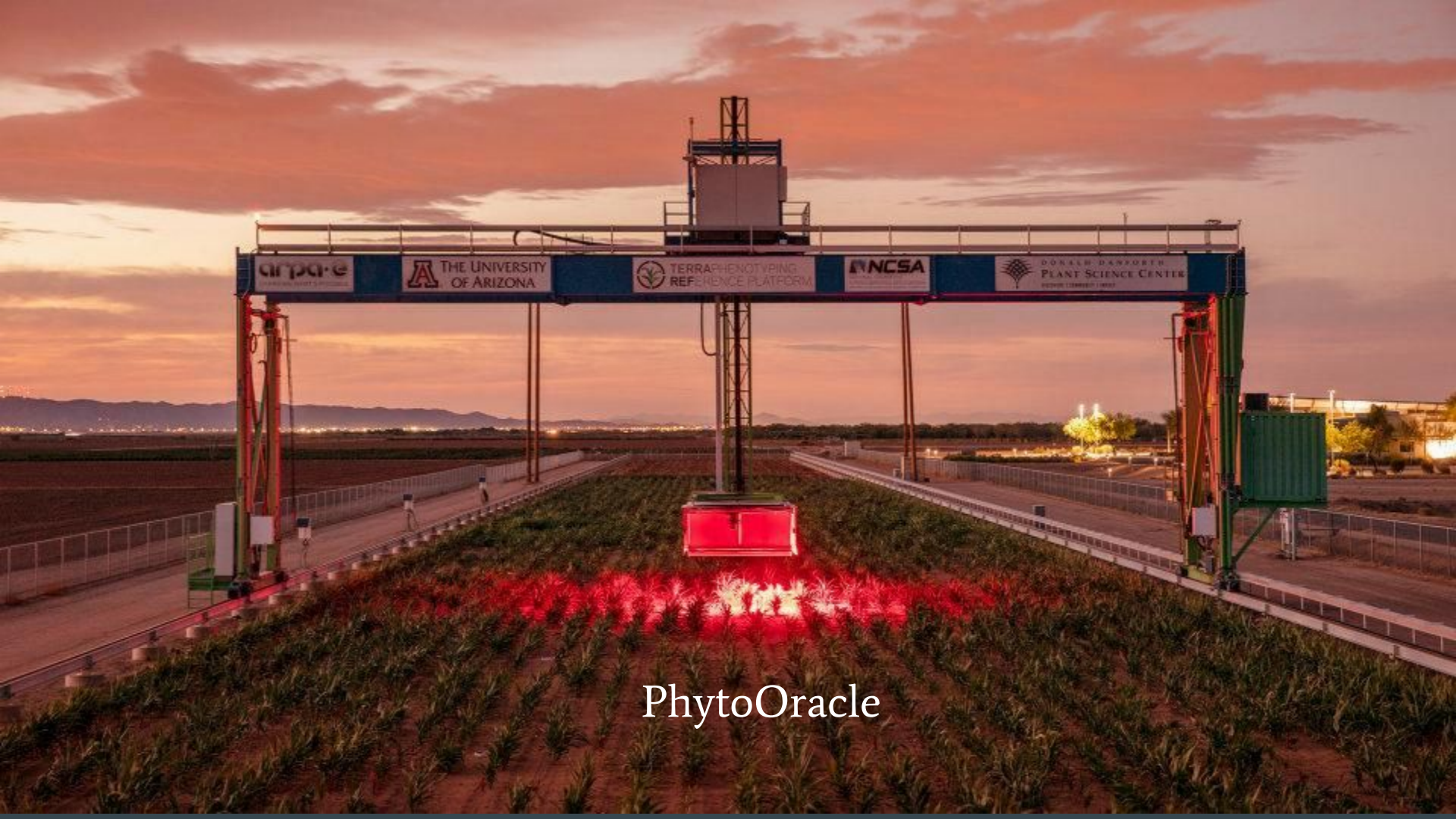
```
mlflow.create_experiment(name=experiment_name, artifact_location=artifact)

mlflow.set_experiment(experiment_name=experiment_name)
```

- Amazon S3 and S3-compatible storage
- Azure Blob Storage
- Google Cloud Storage
- FTP server
- SFTP Server
- NFS
- HDFS

Additional Tooling

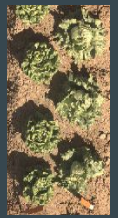
- DE-VICE (Flask, Python-Dask, R-shiny,...)

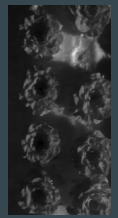


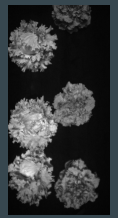
PhytoOracle

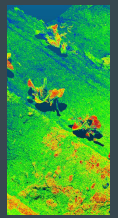
PhytoOracle

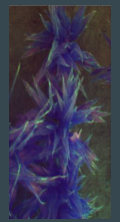
- Joint project
  - Danforth Center
  - School of Plant Science
  - Data Science Institute
  - CyVerse
- Funded by DOE
- Analyze plants in drought stress conditions
  - Genomics ⟷ Phenomics
  - Genomics ⟷ Disease
  - Disease detection
  - Predictive plant modeling
- 5+ cameras and sensors
- Previous CyVerse Webinar
  - https://cyverse.org/webinar-PhytoOracle


RGB


Thermal


Fluorescence


3D

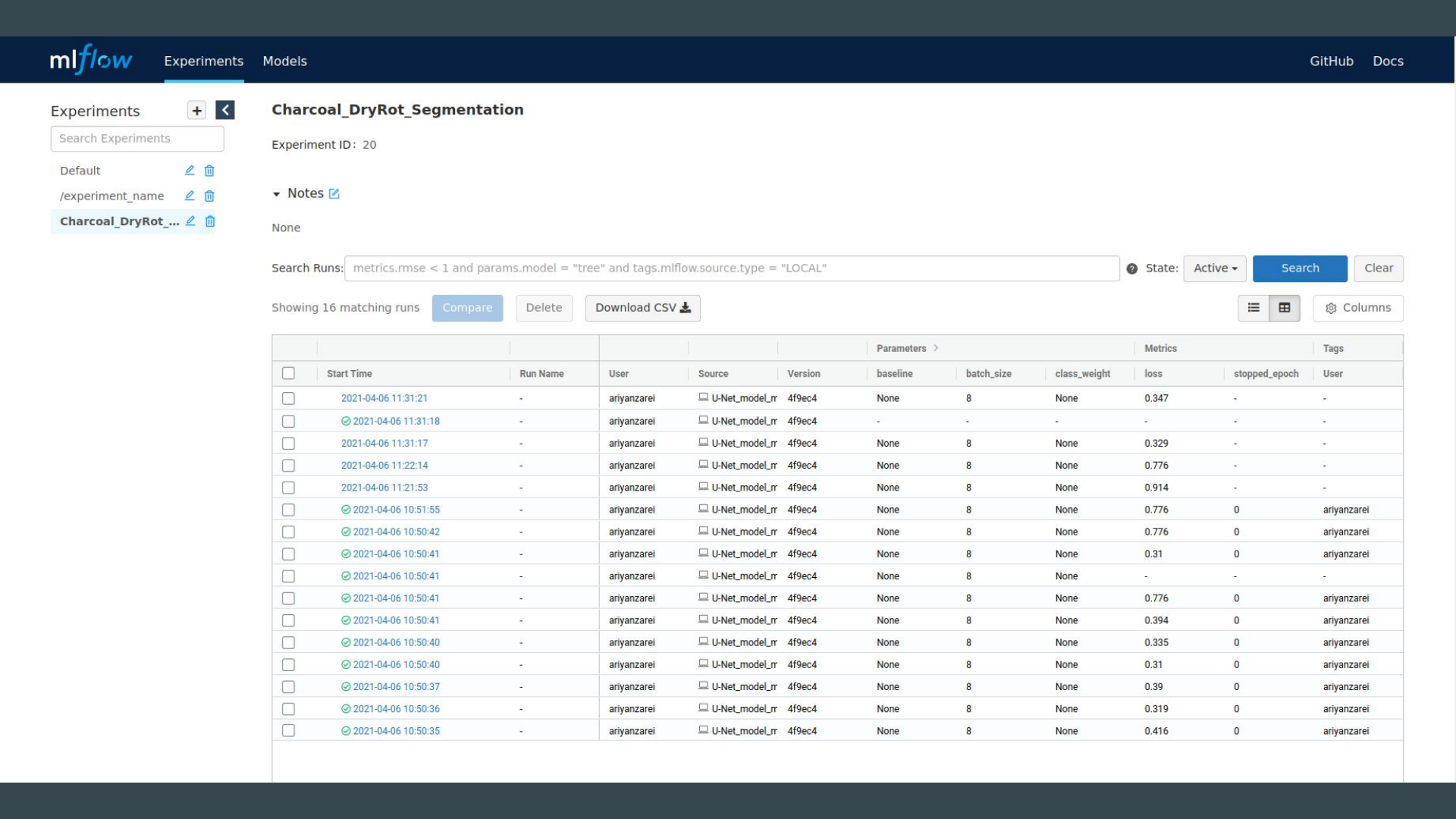

Hyperspectral

Charcoal Dry Rot

- A fungal disease in water-stressed sorghum plants
  - Caused by *Macrophomina phaseolina*
  - Dead tissue
  - Light gray - yellow
  - Starts from tips of the leaves
- Ultimate Goal
  - Disease detection
  - Locate affected regions using drones
  - Apply fungicides
- Train Neural Networks
  - Semantic segmentation
- Labeled 1400 Images
  - http://www.labelbox.com

Why Use MLflow?

- Cottage Industry → Collaborative/Distributed project
- MLflow helps with
  - Collaboration
  - Keeping track of experiments
  - Comparing the results
  - Designing new experiments
  - Storing the models
  - Deploying the models
  - Reproducibility and reusability


mlflow
Experiments
Models
GitHub
Docs
Experiments
Search Experiments
Default
/experiment_name
Charcoal_DryRot_...
Charcoal_DryRot_Segmentation
Experiment ID: 20
Notes
None
Search Runs: metrics.rmse < 1 and params.model = "tree" and tags.mlflow.source.type = "LOCAL"
State: Active
Search
Clear
Showing 16 matching runs
Compare
Delete
Download CSV
Columns
Parameters
Metrics
Tags
Start Time | Run Name | User | Source | Version | baseline | batch_size | class_weight | loss | stopped_epoch | User
2021-04-06 11:31:21 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.347 | - | -
2021-04-06 11:31:18 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | - | - | - | - | - | -
2021-04-06 11:31:17 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.329 | - | -
2021-04-06 11:22:14 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.776 | - | -
2021-04-06 11:21:53 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.914 | - | -
2021-04-06 10:51:55 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.776 | 0 | ariyanzarei
2021-04-06 10:50:42 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.776 | 0 | ariyanzarei
2021-04-06 10:50:41 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.31 | 0 | ariyanzarei
2021-04-06 10:50:41 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | - | - | -
2021-04-06 10:50:41 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.776 | 0 | ariyanzarei
2021-04-06 10:50:41 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.394 | 0 | ariyanzarei
2021-04-06 10:50:40 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.335 | 0 | ariyanzarei
2021-04-06 10:50:40 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.31 | 0 | ariyanzarei
2021-04-06 10:50:37 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.39 | 0 | ariyanzarei
2021-04-06 10:50:36 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.319 | 0 | ariyanzarei
2021-04-06 10:50:35 | - | ariyanzarei | U-Net_model_m | 4f9ec4 | None | 8 | None | 0.416 | 0 | ariyanzarei

mlflow Experiments Models GitHub Docs

Charcoal_DryRot_Segmentation > Run a822dbb50dcf47cb84518e3a1eb9ce18

Date: 2021-04-06 10:50:41 Source: U-Net_model_mlflow.py Git Commit: 4f9ec434311270d62893231c746b5fc02fedeb09

User: ariyanzarei Duration: 40.2min Status: FINISHED

Notes

None

Parameters

Name	Value
baseline	None
batch_size	8
class_weight	None
epochs	20
general_batch_size	8
general_epochs	20
general_loss	weighted_dice_coef
general_optimizer	Adam
general_patience	30
global_momentum	0.5
initial_epoch	0
learning_rate	0.0005
max_pooling_size	2
max_queue_size	10
min_delta	0

▾ Tags

Name	Value	Actions
User	ariyanzarei	

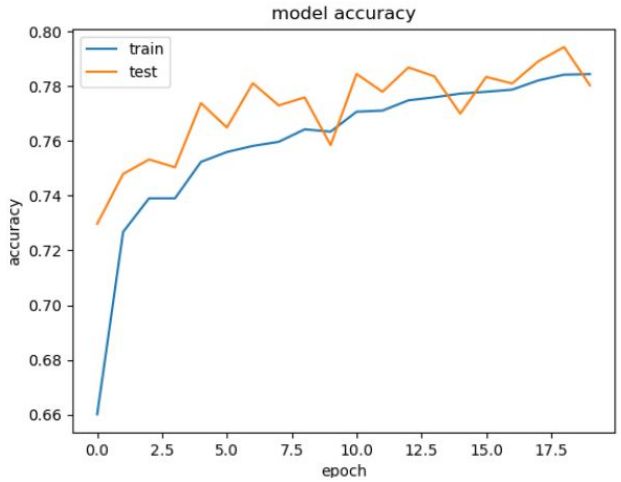
Add Tag

Name | Value | Add

▾ Artifacts

- ▸ model
- ▸ tensorboard_logs
- UNET_13_4_acc_plt.png
- UNET_13_4_val_plt.png
- model_summary.txt
- resfull_contour_UNET_13_4.png
- resfull_mask_UNET_13_4.png
- resfull_masked_UNET_13_4.png

Full Path: sftp://artinmajdi: @data7-db1.cyverse.org:/home/artinmajdi/mlflow_data/artifact_store/a822dbb50dcf47c...
Size: 32.01KB



Artifacts

- model
- tensorboard_logs
- UNET_13_4_acc_plt.png
- UNET_13_4_val_plt.png
- model_summary.txt
- resfull_contour_UNET_13_4.png
- resfull_mask_UNET_13_4.png
- resfull_masked_UNET_13_4.png

Full Path: sftp://artinmajdi:temp2_data7_b@data7-db1.cyverse.org:/home/artinmajdi/mlflow_data/artifact_store/a822dbb50dcf47c...
Size: 20.98MB

Charcoal_DryRot_Segmentation > Comparing 2 Runs

Run ID:	a822dbb50dcf47cb84518e3a1eb9ce18	61cbd68bba8e467583cfad28e1c1144a
Run Name:		
Start Time:	2021-04-06 10:50:41	2021-04-06 10:50:40

Parameters

baseline	None	None
batch_size	8	8
class_weight	None	None
epochs	20	20
general_batch_size	8	8
general_epochs	20	20
general_loss	weighted_dice_coef	weighted_dice_coef
general_optimizer	Adam	Adam
general_patience	30	30
global_momentum	0.5	0.5
initial_epoch	0	0
learning_rate	0.0005	5e-05
max_pooling_size	2	2
max_queue_size	10	10
min_delta	0	0
monitor	val_mean_io_u	val_mean_io_u
opt_amsgrad	False	False
opt_beta_1	0.9	0.9

Lessons Learned Using MLflow

01	Setting up the server and database
02	Comparing artifacts not possible
03	Autolog not working all the time

01	Prerequisites: Familiarity with Conda/Docker, SQL flavors, ssh-tunneling, . . .
02	Dependency version mismatch
03	Doesn’t support singularity

Getting Started with MLflow on CyVerse

Via External Collaborative Partnerships (ECP), researchers are paired with an expert to address their project's specific computational needs and more (postgreSQL . .)

➔ Request an ECP: https://cyverse.org/ecp

Helpful Resources

Links

- MLflow
  https://mlflow.org
- CyVerse
  www.cyverse.org
- PhytoOracle Docs
  https://tinyurl.com/phytooracle-rtd
- CyVerse Webinar on PhytoOracle
  https://cyverse.org/webinar-PhytoOracle
- MLflow use case with MNIST classification
  https://github.com/artinmajdi/mlflow_workflow
- MLflow use case with Chest X-Rays
  https://github.com/artinmajdi/chest-x-ray-classification

THE UNIVERSITY OF ARIZONA
RESEARCH, DISCOVERY & INNOVATION
Data Science
Institute

PhytoOracle