MLOPS PILOT INTEGRATION WITH MLFLOW

Office of Compensation & Working Conditions

Sponsor: David Oh – Data Scientist

REMY STEWART
Cornell University
PhD Candidate in Sociology & Information Sciences

BACKGROUND

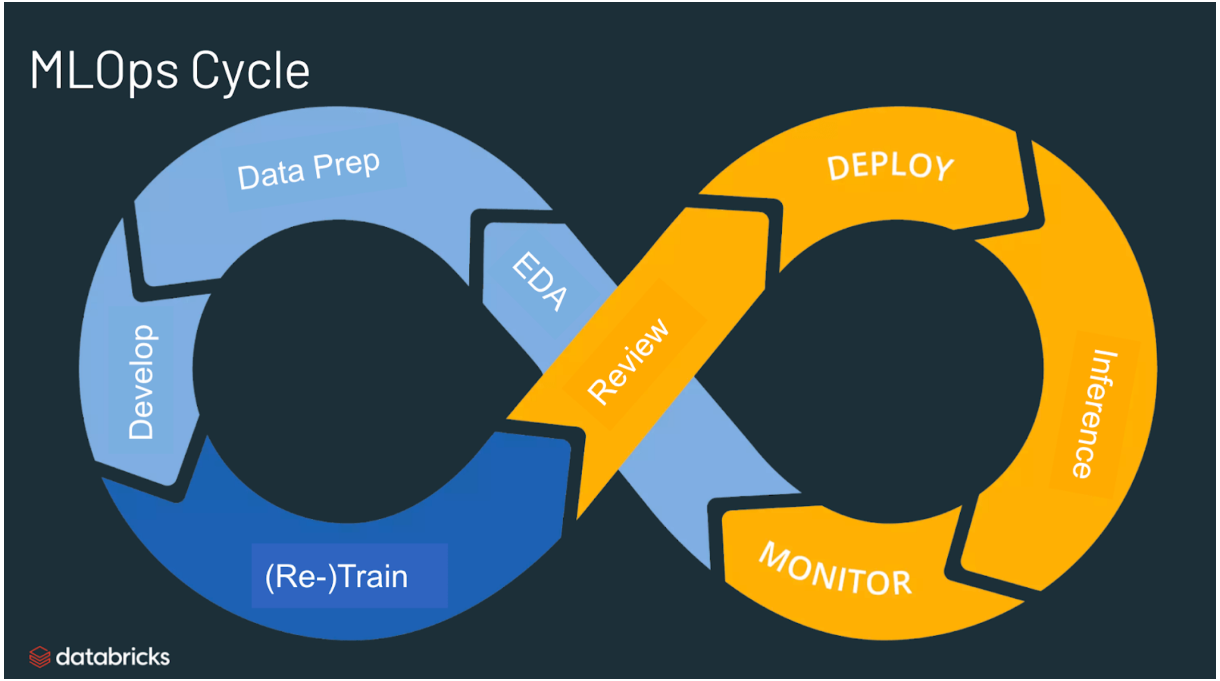
From starting a machine learning (ML) project to a model in production ...

- Acquiring data
- Transforming data into proper formats
- Running & tracking experiments
- Comparing experiment performance
- Packaging final models for deployment
- Successfully deploying models
- Monitoring live performance
- Retraining & redeploying over time

How can we streamline these steps to support reproducibility, transparency, and team collaboration?

CIVIC DIGITAL FELLOWSHIP

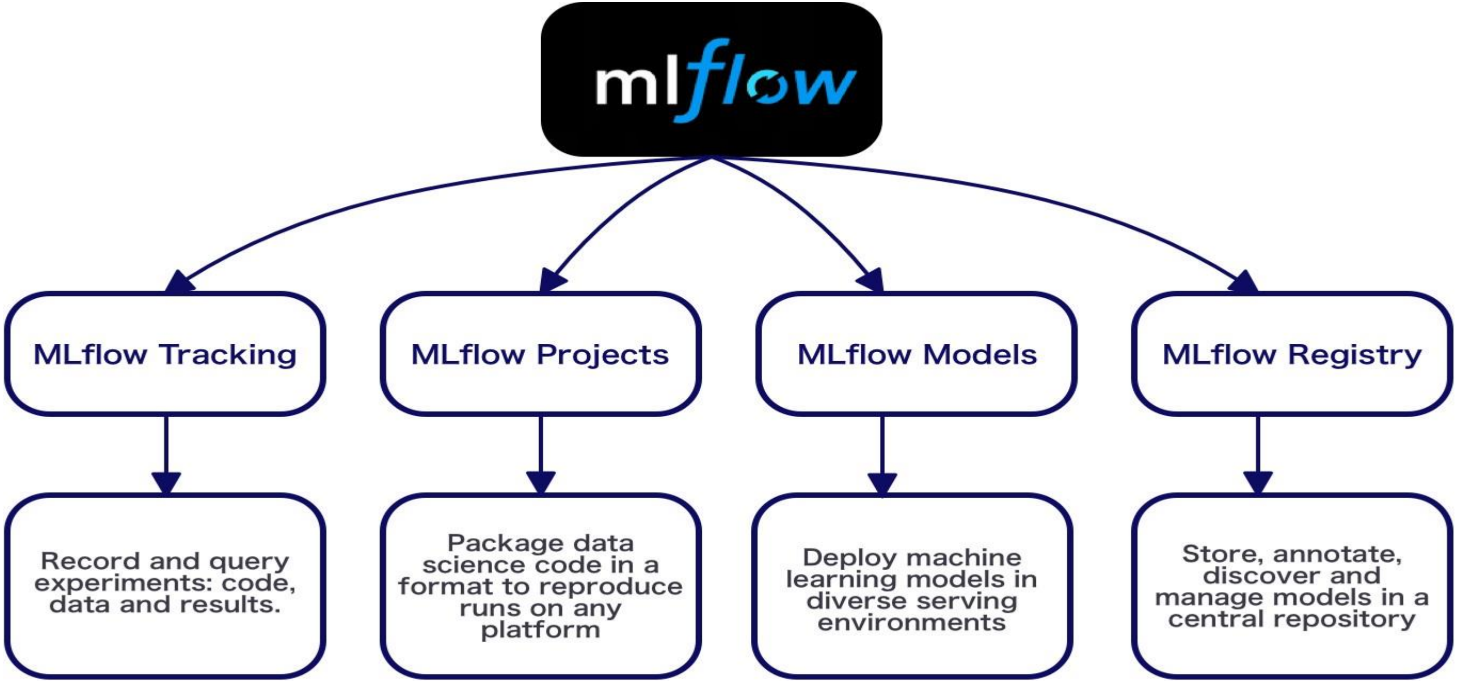
MACHINE LEARNING OPERATIONS (MLOPS)



CIVIC DIGITAL FELLOWSHIP

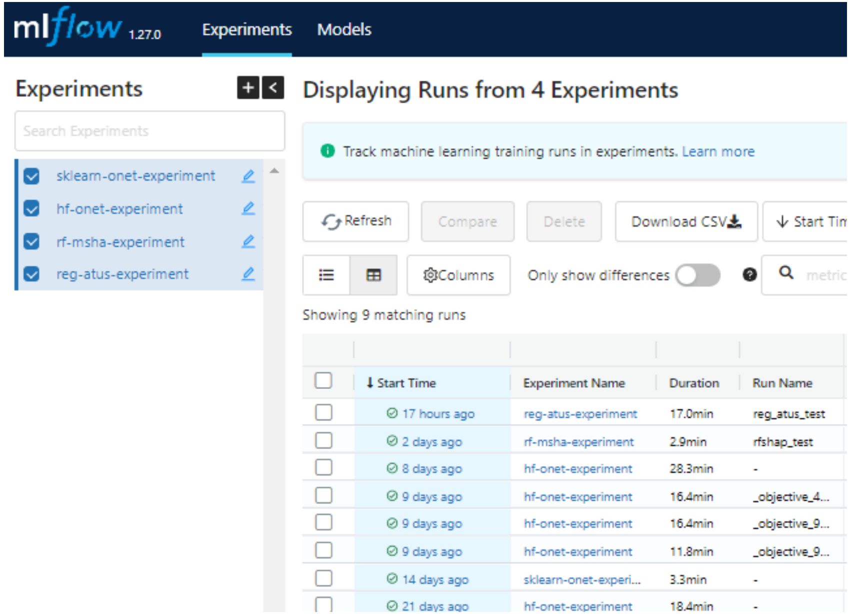
MLOps encompasses a range of best practices designed to support all stages of the ML project lifecycle.

MLFLOW


mlflow
MLflow Tracking
MLflow Projects
MLflow Models
MLflow Registry
Record and query experiments: code, data and results.
Package data science code in a format to reproduce runs on any platform
Deploy machine learning models in diverse serving environments
Store, annotate, discover and manage models in a central repository

TWO SIDES OF MLFLOW

User Interface (UI)



Python Application Programming Interface (API)

```
mlflow.log_params(best_cv_params)
predicted_prob_df = helpers.force_prediction(pd.DataFrame
signature = mlflow.models.infer_signature(task_train, pre
mlflow.sklearn.log_model(cross_val, "logreg_model", conda
                         registered_model_name="sklearn_o

# Logging all associated metrics into Mlflow- note the ar
mlflow.log_metric("Accuracy", accuracy_score)
mlflow.log_metrics(prf_micro)
mlflow.log_metrics(prf_sample)
mlflow.log_metric("Hamming Score", hamming_score_result)
mlflow.log_metric("Hamming Loss", hamming_loss)

# Logging classification label counts figure, current not
row_sums = predicted_prob_df.sum(axis=1)
row_sum_plot = row_sums.value_counts().plot(kind='bar')
row_sum_plot = row_sum_plot.get_figure()
mlflow.log_figure(row_sum_plot, "row_sum_plot.png")
mlflow.log_artifact("sklearn_mlflow_example.ipynb")
mlflow.log_artifact("./pyfiles/helpers.py")
```

PROJECT OBJECTIVES

1. Establish a central MLflow server supporting remote access

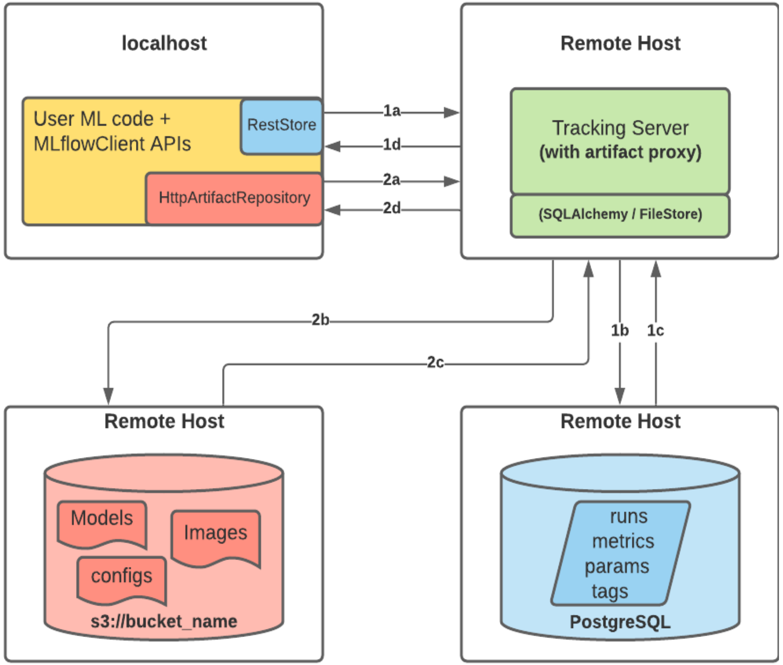
2. Test & document MLflow's features

3. Develop example Python scripts demonstrating MLflow code integration

4. Provide guidance on MLflow's ability to support BLS data science use cases

CIVIC DIGITAL FELLOWSHIP

ESTABLISHING THE REMOTE SERVER



* Not using S3 for BLS remote server

Central location to track projects & experiments

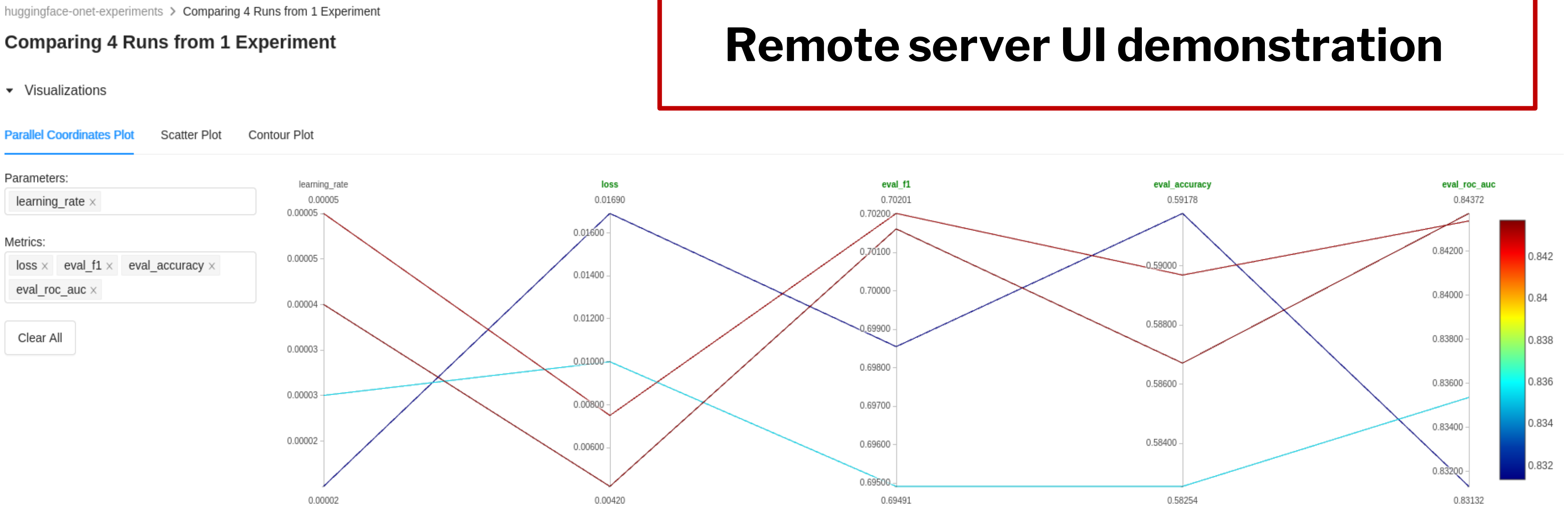
Accessible from individual work computers

Stores a range of ML project "artifacts"

- Models themselves
- Data visualizations
- Code files
- Chosen model parameters
- Performance metrics
- Project documentation

CIVIC DIGITAL FELLOWSHIP

TESTING MLFLOW FEATURES



Remote server UI demonstration

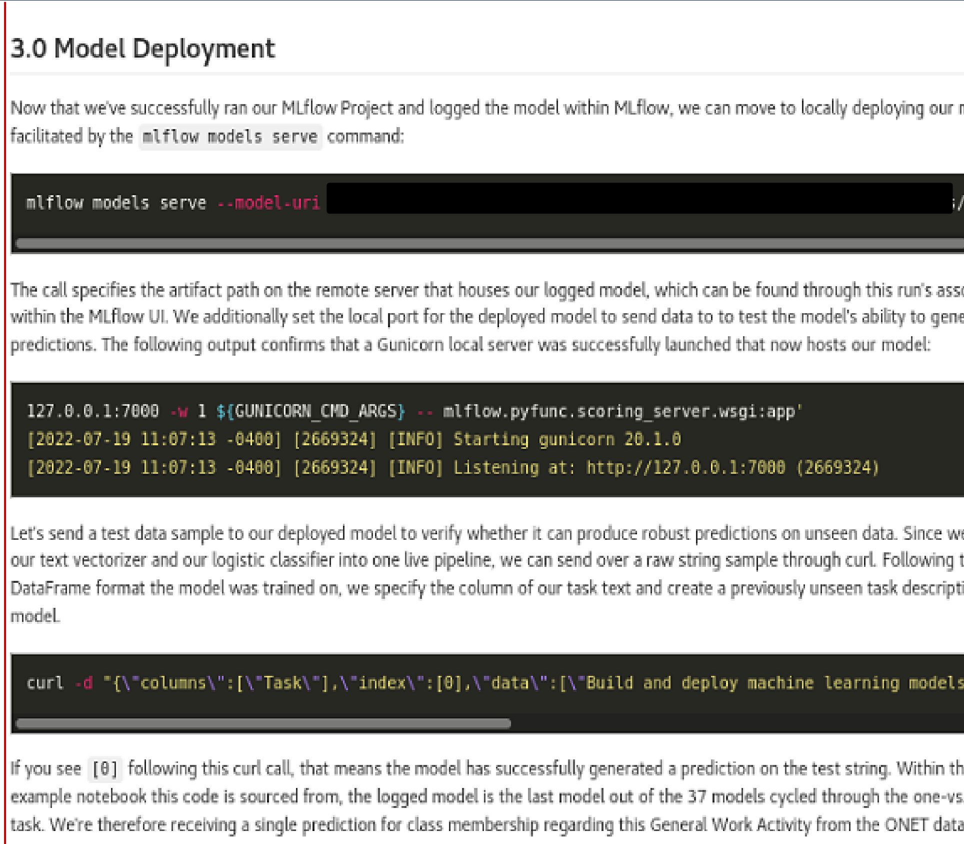
CIVIC DIGITAL FELLOWSHIP

MLFLOW EXAMPLE CODE

6 Example Code Walkthroughs

- 5 ML libraries
- 4 unique algorithms
- 3 public data sets

Delineating MLflow integration from initial data preparation to live model deployment



3.0 Model Deployment

Now that we've successfully ran our MLflow Project and logged the model within MLflow, we can move to locally deploying our r
facilitated by the `mlflow models serve` command:

```
mlflow models serve --model-uri [redacted]:/
```

The call specifies the artifact path on the remote server that houses our logged model, which can be found through this run's ass
within the MLflow UI. We additionally set the local port for the deployed model to send data to to test the model's ability to gene
predictions. The following output confirms that a Gunicorn local server was successfully launched that now hosts our model:

```
127.0.0.1:7000 -w 1 ${GUNICORN_CMD_ARGS} -- mlflow.pyfunc.scoring_server.wsgi:app'
[2022-07-19 11:07:13 -0400] [2669324] [INFO] Starting gunicorn 20.1.0
[2022-07-19 11:07:13 -0400] [2669324] [INFO] Listening at: http://127.0.0.1:7000 (2669324)
```

Let's send a test data sample to our deployed model to verify whether it can produce robust predictions on unseen data. Since we
our text vectorizer and our logistic classifier into one live pipeline, we can send over a raw string sample through curl. Following t
DataFrame format the model was trained on, we specify the column of our task text and create a previously unseen task descripti
model.

```
curl -d "{\"columns\":[\"Task\"],\"index\":[0],\"data\":[\"Build and deploy machine learning models
```

If you see `[0]` following this curl call, that means the model has successfully generated a prediction on the test string. Within th
example notebook this code is sourced from, the logged model is the last model out of the 37 models cycled through the one-vs
task. We're therefore receiving a single prediction for class membership regarding this General Work Activity from the ONET data

CIVIC DIGITAL FELLOWSHIP

MLFLOW AT BLS

BLS MLFlow Pilot Integration Documentation

Remy Stewart, BLS Civic Digital Fellow, Summer 2022

With the increase of machine learning applications across BLS offices, there is a growin
collaboration across teams, and assist with transitioning models from experimentation t
location to facilitate the practices of machine learning operations within organizations-
machine learning model lifecycles.

This documentation serves as an introduction to MLflow overall and towards its pilot ap
needs for BLS data scientists, and tips for how to best incorporate the platform's comp
Jupyter Notebooks within this repository's examples folder that offer code walkthrough
transformer models processing three different public BLS-affiliated data sets. The mlpro
scripts into production environments.

- Documenting how MLflow can best support BLS data science
- How to reproduce the remote server configuration
- Opportunities & challenges regarding MLflow for BLS use
- Future work towards scalability & comprehensive implementation

THANK YOU FOR A FANTASTIC SUMMER!

Special thanks to David Oh, Brandon Kopp, and Lowell Mason for their support of this project.

CIVIC DIGITAL FELLOWSHIP

U.S. BUREAU OF LABOR STATISTICS

REMY STEWART

Cornell University
PhD Candidate in Sociology & Information Sciences