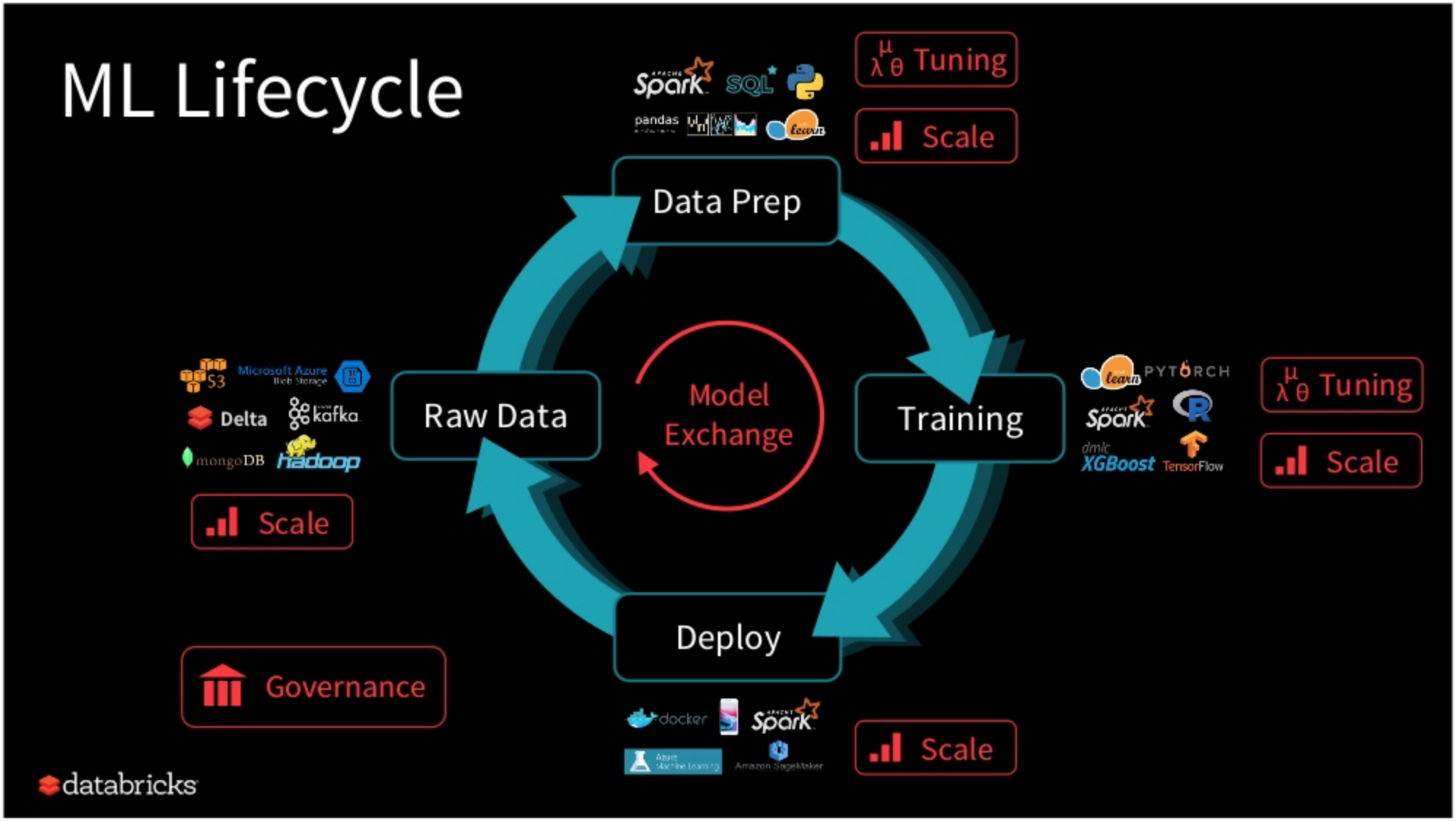
mlflow: Accelerating Production Machine Learning
Matei Zaharia
@matei_zaharia
databricks

Machine Learning Development is Complex


ML Lifecycle
Data Prep
Model Exchange
Training
Deploy
Raw Data
Tuning
Scale
Tuning
Scale
Scale
Scale
Governance
Delta
kafka
mongoDB
hadoop
Microsoft Azure
Blob Storage
PYTORCH
XGBoost
TensorFlow
docker
databricks

Custom ML Platforms

Facebook FBLearner, Uber Michelangelo, Google TFX

- + **Standardize the data prep / training / deploy cycle: if you work within the platform, you get these!**
- – **Limited to a few algorithms or frameworks**
- – **Tied to one company's infrastructure**

Can we provide similar benefits in an open manner?

Introducing mlflow

Open source machine learning platform

- Works with any ML library & language
- Runs the same way everywhere & cross-cloud
- Scales to big data with Apache Spark

Launched this June

- Already 48 contributors and many new features!

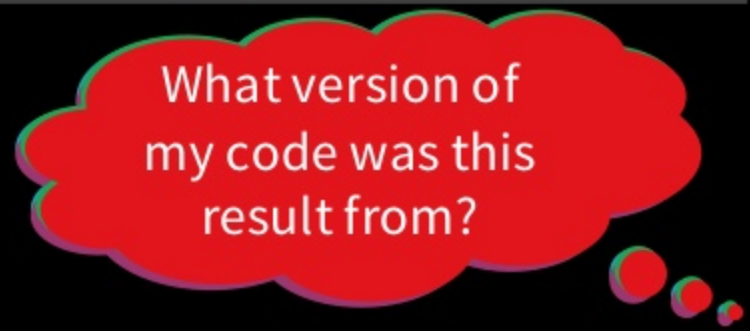
This Talk: MLflow Overview + New Announcements

How Does MLflow Work?

Model Development without MLflow

```
data   = load_text(file)
ngrams = extract_ngrams(data, N=n)
model  = train_model(ngrams,
             learning_rate=lr)
score  = compute_accuracy(model)

print("For n=%d, lr=%f: accuracy=%f"
      % (n, lr, score))

pickle.dump(model, open("model.pkl"))
```

```
For n=2, lr=0.1: accuracy=0.71
For n=2, lr=0.2: accuracy=0.79
For n=2, lr=0.5: accuracy=0.83
For n=2, lr=0.9: accuracy=0.79
For n=3, lr=0.1: accuracy=0.83
For n=3, lr=0.2: accuracy=0.82
For n=4, lr=0.5: accuracy=0.75
...
```



Model Deployment without MLflow

DATA SCIENTIST

Please deploy this ArXiv paper!

Code & Models

…

PRODUCTION ENGINEER

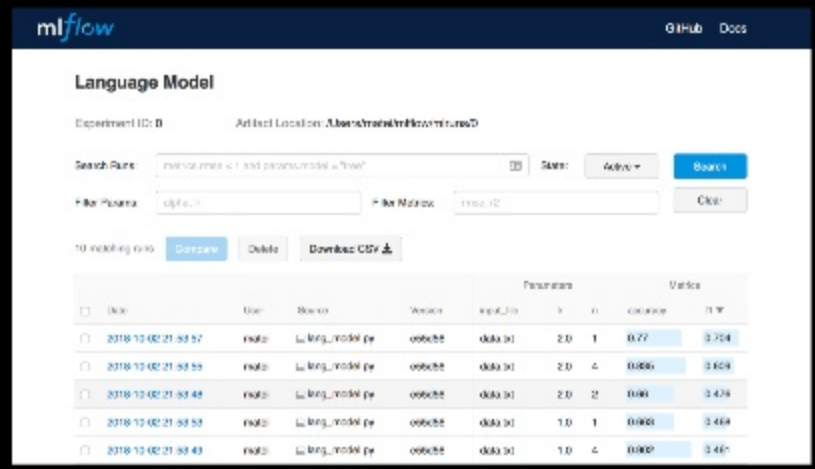
databricks

Model Development *with* MLflow

```
data   = load_text(file)
ngrams = extract_ngrams(data, N=n)
model  = train_model(ngrams,
             learning_rate=lr)
score  = compute_accuracy(model)

print("For n=%d, lr=%f: accuracy=%f"
      % (n, lr, score))

pickle.dump(model, open("model.pkl"))
```

Model Development *with* MLflow

```
data   = load_text(file)
ngrams = extract_ngrams(data, N=n)
model  = train_model(ngrams,
                     learning_rate=lr)
score  = compute_accuracy(model)

mlflow.log_param("data_file", file)
mlflow.log_param("n", n)
mlflow.log_param("learning_rate", lr)
mlflow.log_metric("score", score)

mlflow.sklearn.log_model(model)
```

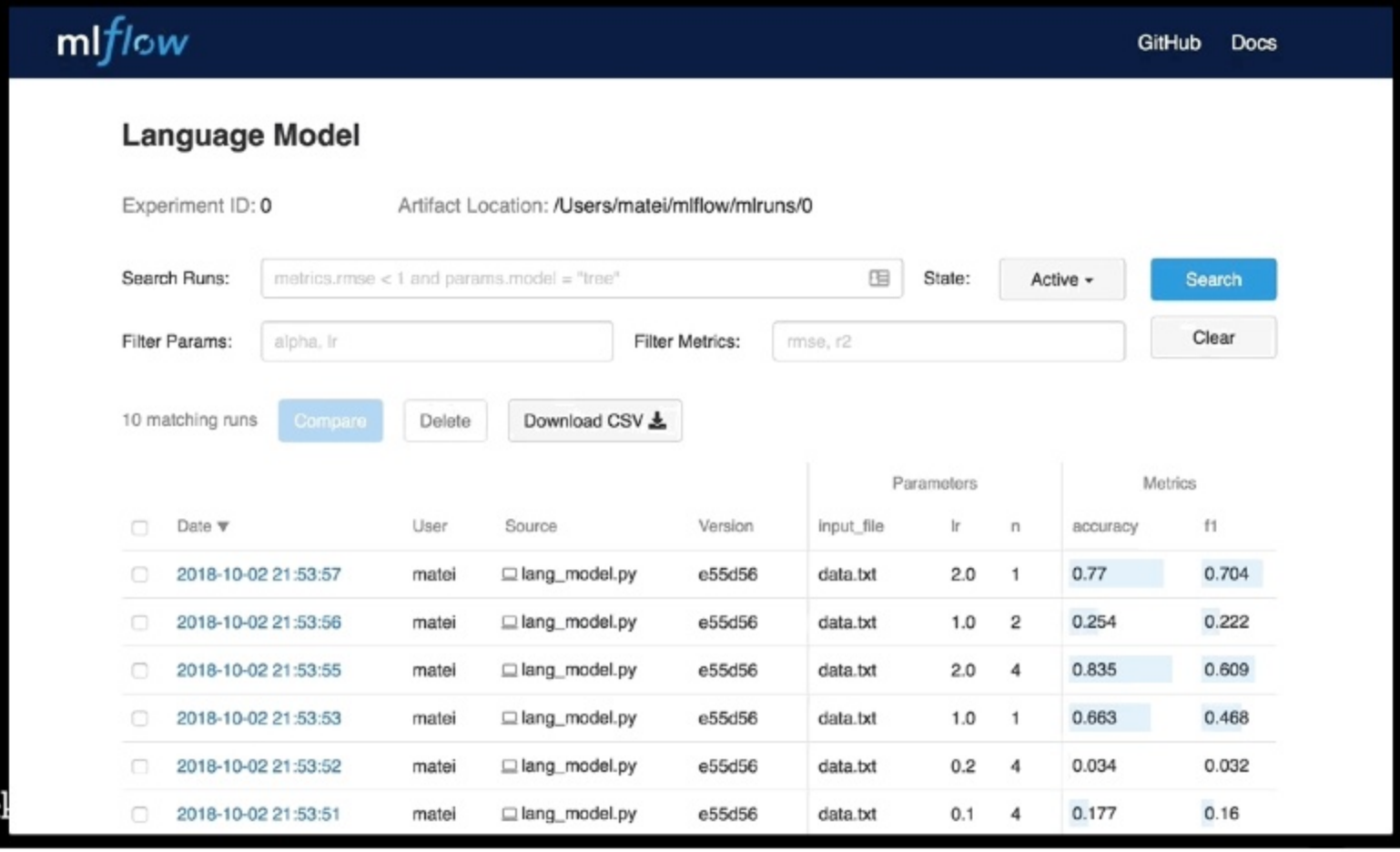


Track parameters, metrics, output files & code version

Search using UI or API

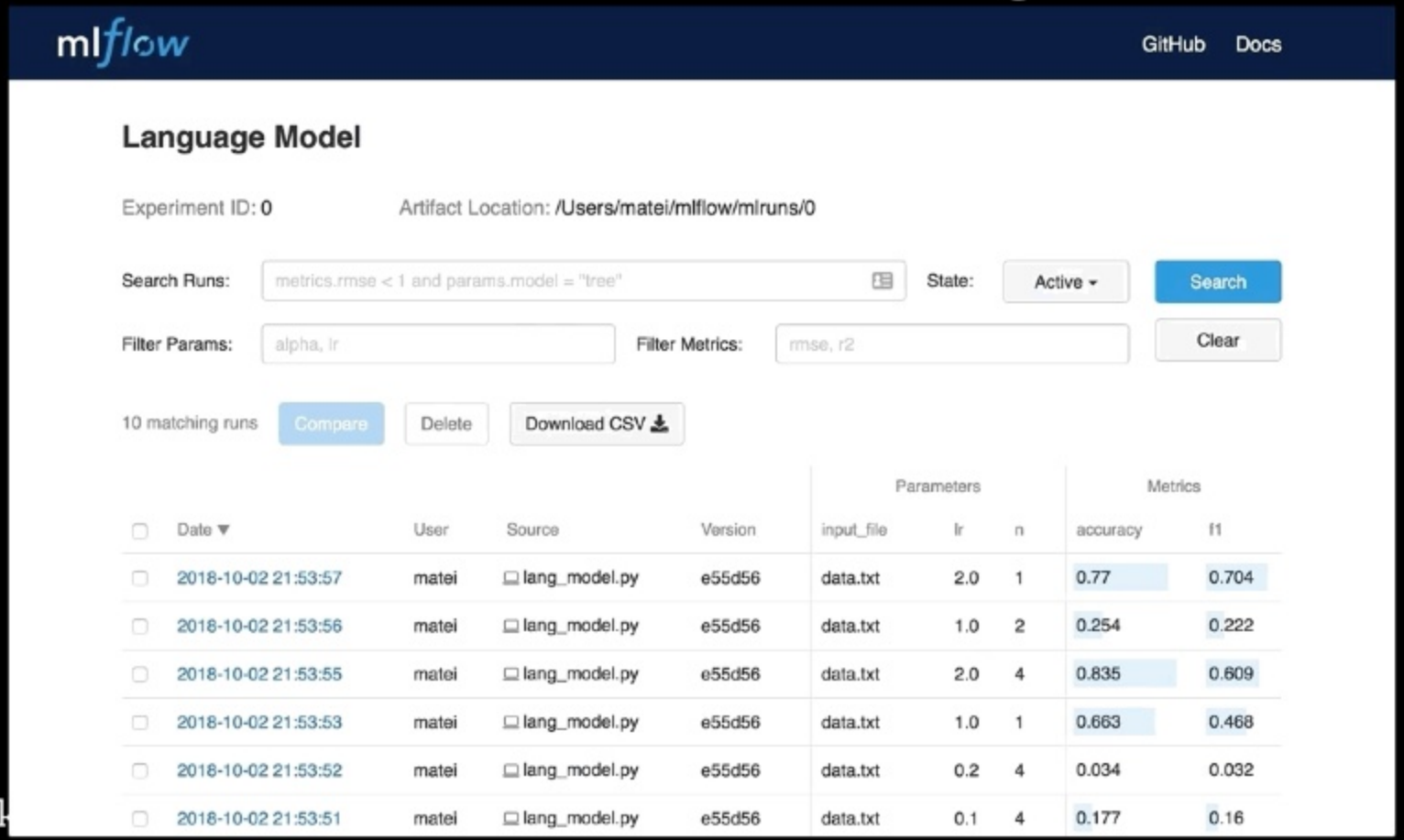
databricks

MLflow UI: Inspecting Runs



databrick

MLflow UI: Comparing Runs



Packaging Code: MLflow Projects

mlflow

Project Spec

Code Deps Config

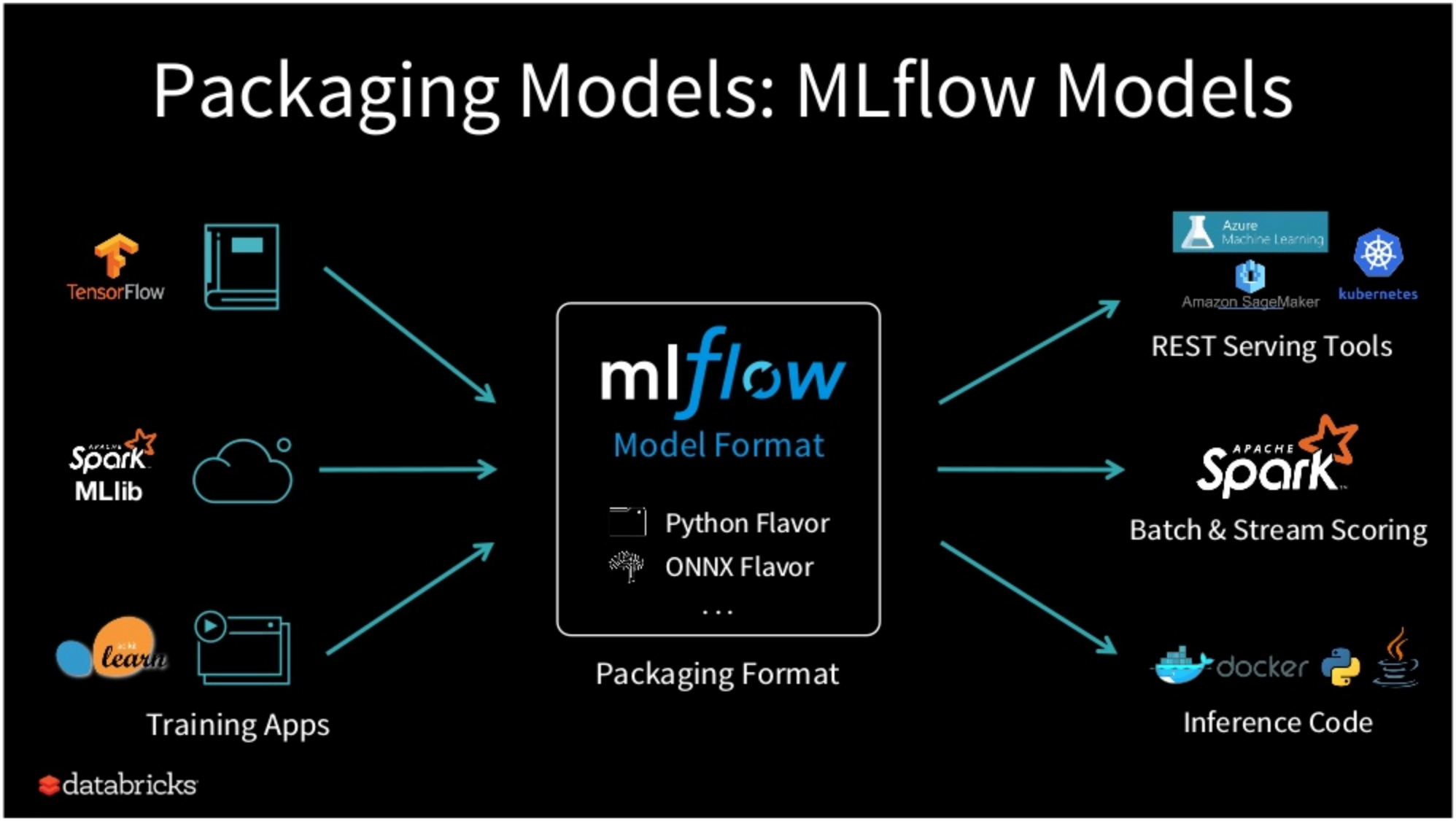
Local Execution

Remote Cluster

```
$ mlflow run git://...
```

databricks

Packaging Models: MLflow Models



Model Deployment *with* MLflow

DATA SCIENTIST

Please run this MLflow Project nightly for updates!

PRODUCTION ENGINEER

Don't even tell me what ArXiv paper that's from...

databricks

MLflow Development Status

Many new features since our release in June

- Model packaging for MLlib, H2O, TensorFlow, PyTorch, Keras
- Storage on Azure, AWS, Google, SFTP
- Java and Scala API
- New examples and UI features

Just released MLflow 0.7.0 today!

Major Announcement in MLflow 0.7.0

RStudio partnered with Databricks to add an MLflow R API

See Kevin Kuo's talk on this at 14:40 today!

Conclusion

Workflow platforms can greatly simplify ML development

- Improve usability for both data scientists and engineers

Get started at mlflow.org, and come see our other talks:

- MLflow R API: 14:40 today
- ML Factories: 11:40 Thursday
- 1h Deep Dive: 14:00 Thursday