An open platform for the machine learning lifecycle

https://github.com/alfozan/mlflow-example

Abdulrahman Alfozan

PyData Riyadh, Oct 2020

Intro

Systems Engineer @ FB

- Apache Spark, Distributed Systems
- Data Engineering
- Applied Machine Learning

- Open-source distributed cluster-computing framework.
- **Unified analytics engine** for big data and machine learning.

Agenda

1. ML Pipeline Lifecycle
2. ML development challenges
3. MLflow platform
    a) Tracking
    b) Projects
    c) Models
4. Demo

ML Pipeline Lifecycle

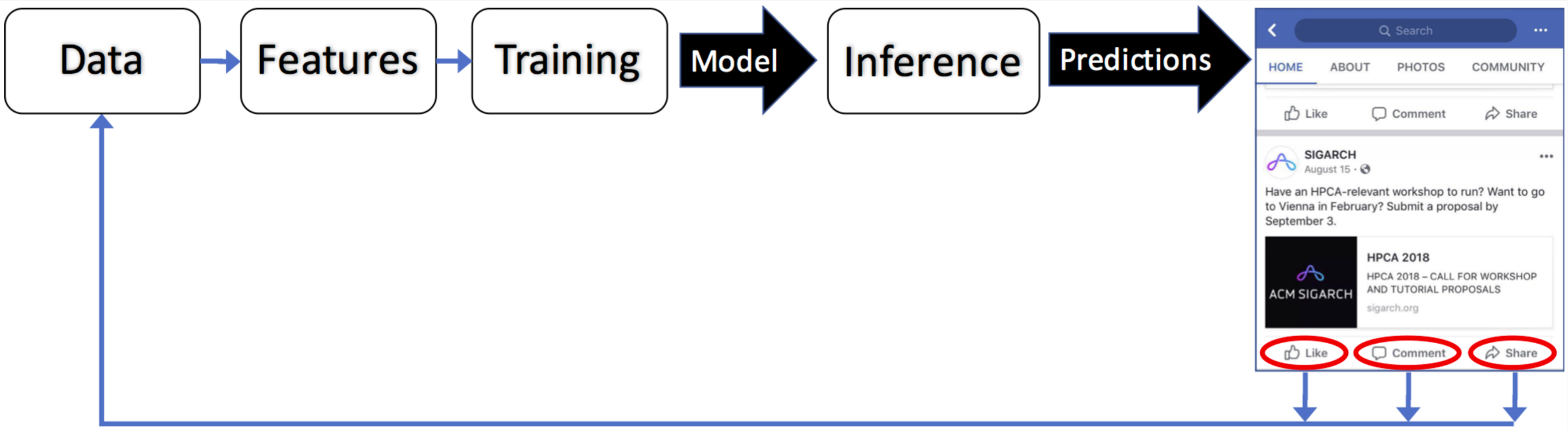


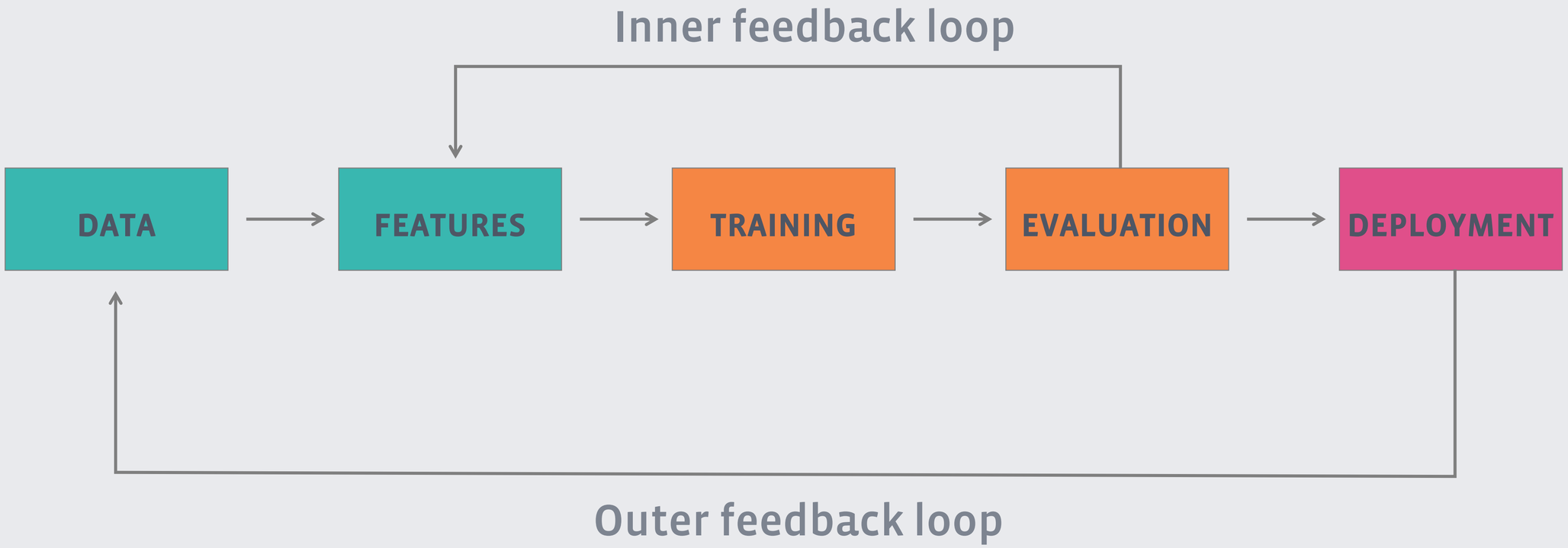
Fig. 1. Example of Facebook's Machine Learning Flow and Infrastructure.

offline

online

Source: https://research.fb.com/publications/applied-machine-learning-at-facebook-a-datacenter-infrastructure-perspective/

ML Pipeline Lifecycle



Source: https://atscaleconference.com/videos/scale-2018-applied-machine-learning-at-facebook-an-infrastructure-perspective/

Agenda

1. ML Pipeline Lifecycle
2. ML development challenges
3. MLflow platform
    a) Tracking
    b) Projects
    c) Models
4. Demo

“Hidden Technical Debt in Machine Learning Systems”

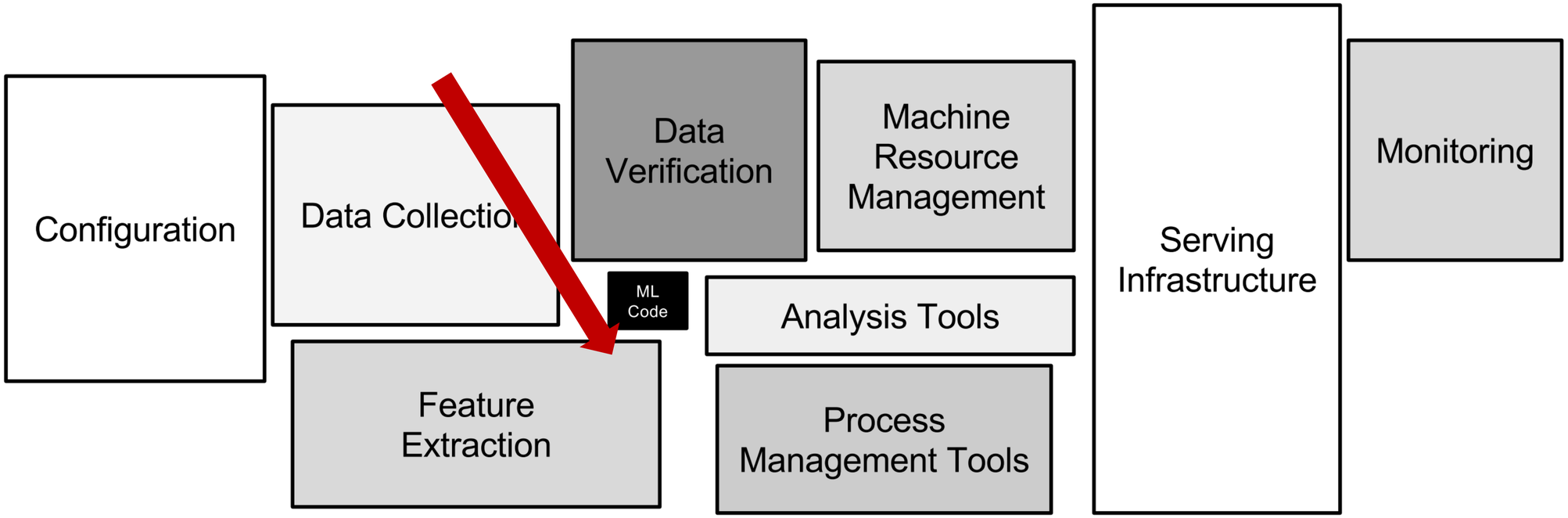


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

ML development challenges

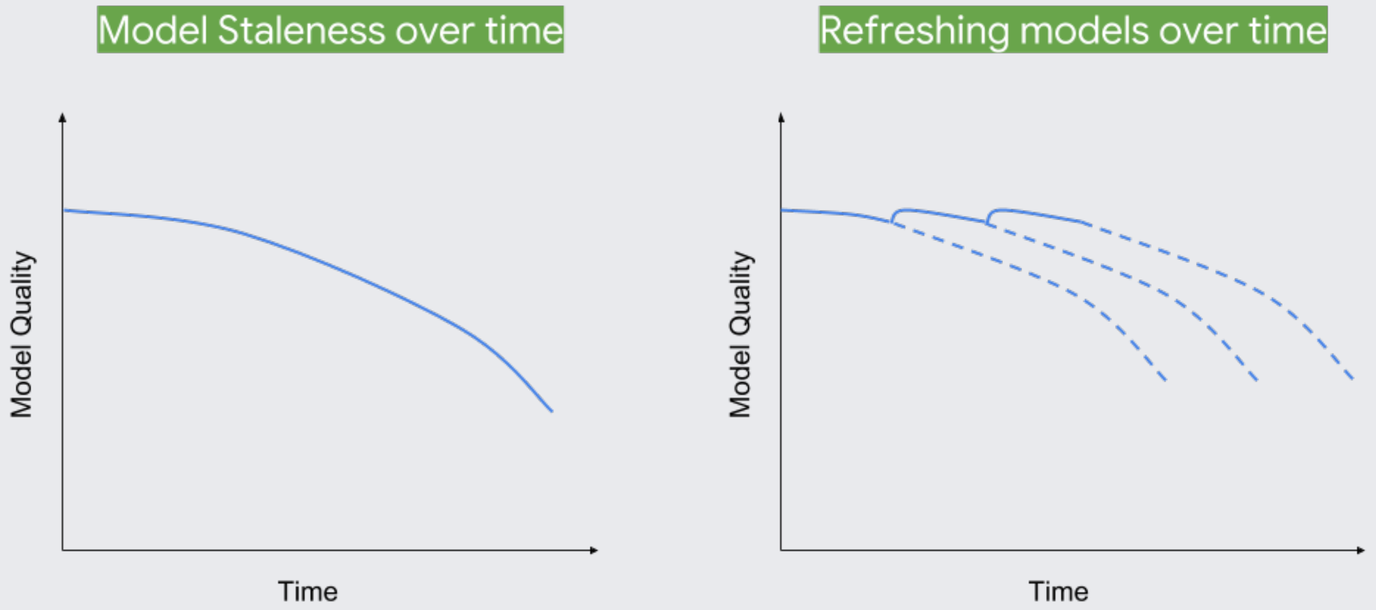
1) Experiment Standardization and Tracking

- Reproducibility
- Parameters Tunning

ML development challenges

2) Model Staleness

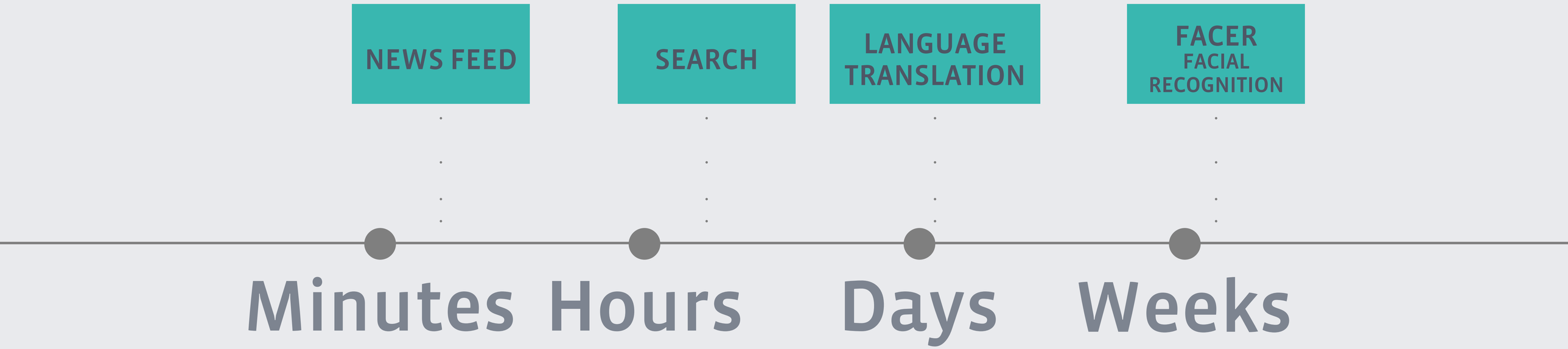
ML pipelines need to be constantly run with new data to avoid model staleness and close feedback loop



Source: https://medium.com/thelaunchpad/how-to-protect-your-machine-learning-product-from-time-adversaries-and-itself-ff07727d6712

Model Staleness

Training frequency for some model types at FB



Source: https://atscaleconference.com/videos/scale-2018-applied-machine-learning-at-facebook-an-infrastructure-perspective/

ML development challenges

3) Productionization

- ML code is difficult to productionize
- Development environment <> production environment

Agenda

1. ML Pipeline Lifecycle
2. ML development challenges
3. MLflow platform
    a) Tracking
    b) Projects
    c) Models
4. Demo

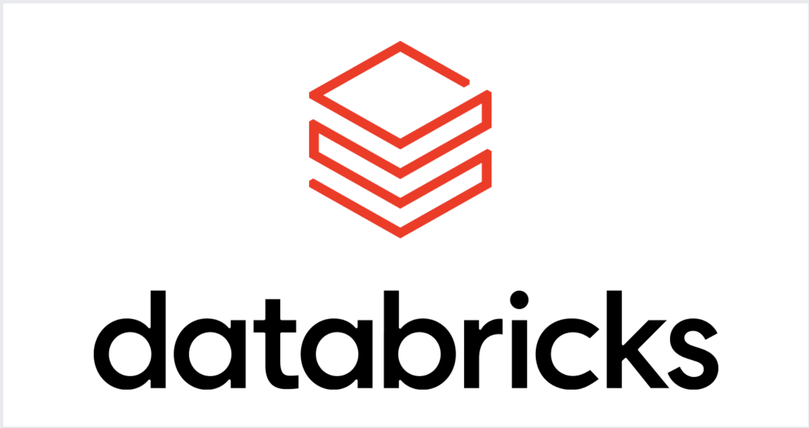
mlflow

- An open platform for the machine learning lifecycle
- Python Library; runs locally and on the cloud
- Built-in UI for experiment visualization
- Logging integrations for major frameworks: scikit-learn, PyTorch, TF,..

https://github.com/mlflow

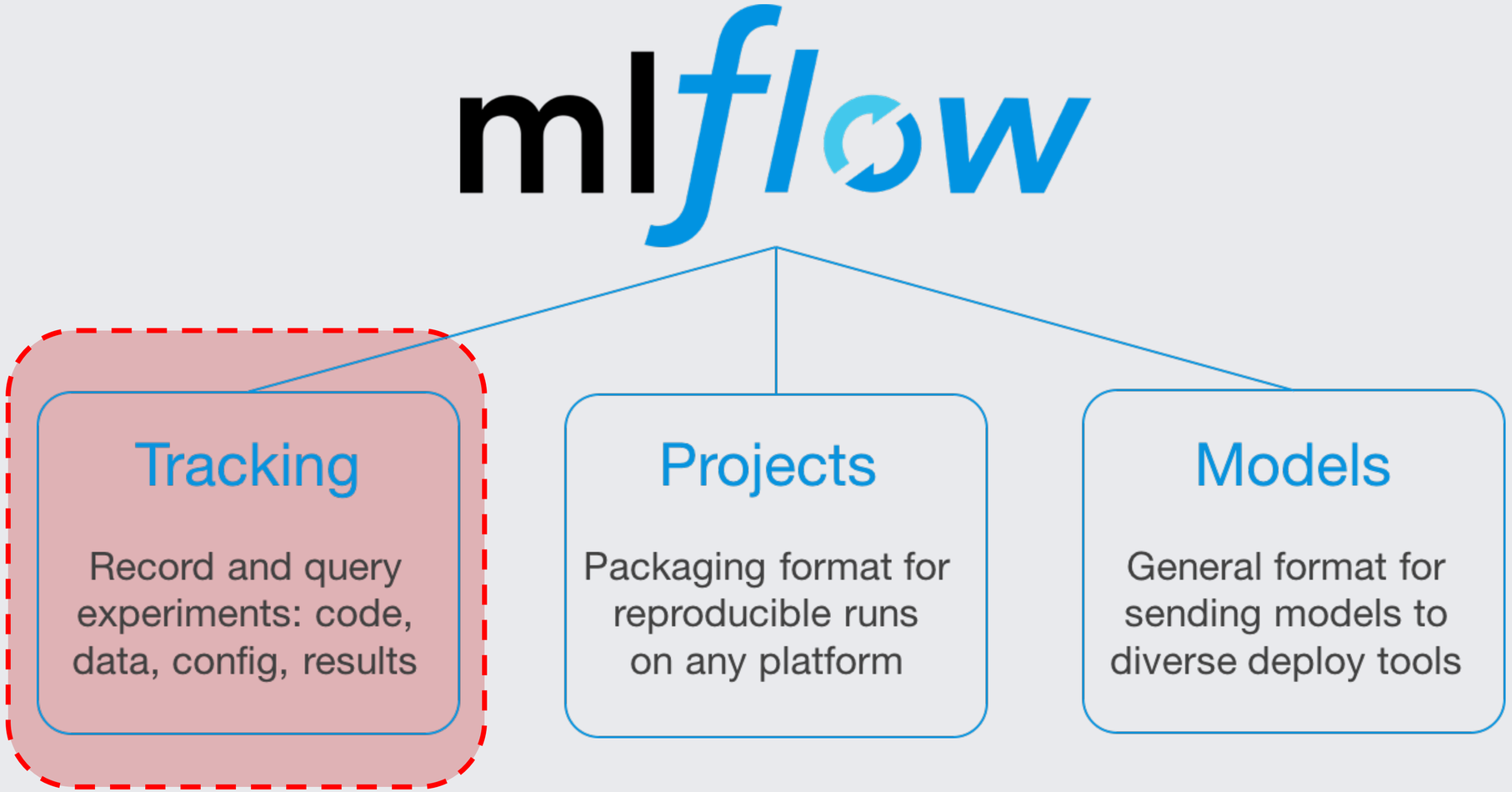
Getting Started with MLflow

- Install with `pip install mlflow`
- Find detailed tutorials at mlflow.org



- Repo: https://github.com/mlflow
- Main contributor and maintainer:
- Managed MLflow services offered by Databricks and Azure ML

MLflow Components



Source: https://www.mlflow.org/docs/latest/tracking.html

MLflow Components

Experiment tracking

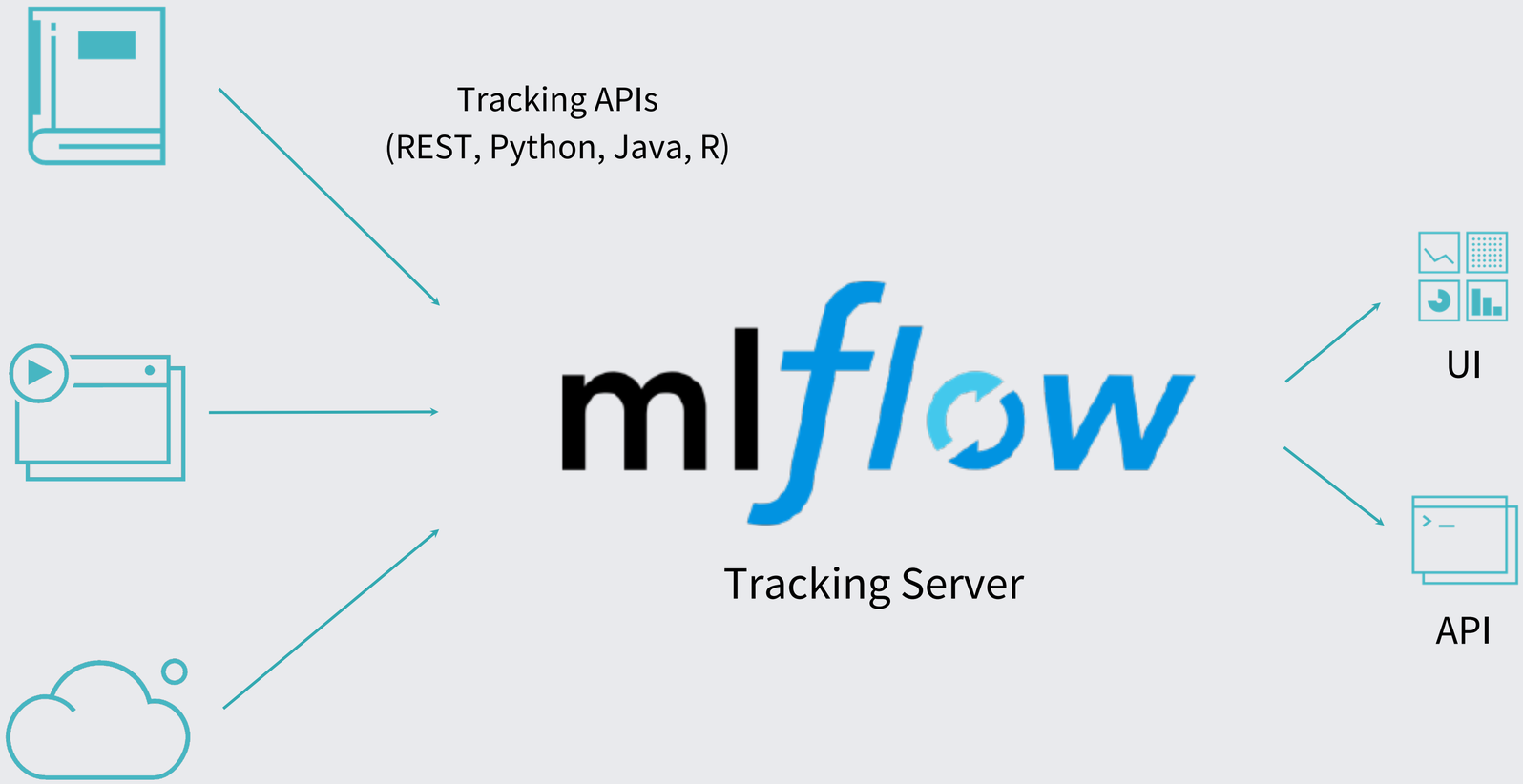
- **Hyper Parameters**: key-value inputs
- **Metrics**: numeric values (i.e. perf metrics)
- **Artifacts**: files, including data and models
- **Source**: training code

any additional information

Source: https://www.mlflow.org/docs/latest/tracking.html

MLflow Components

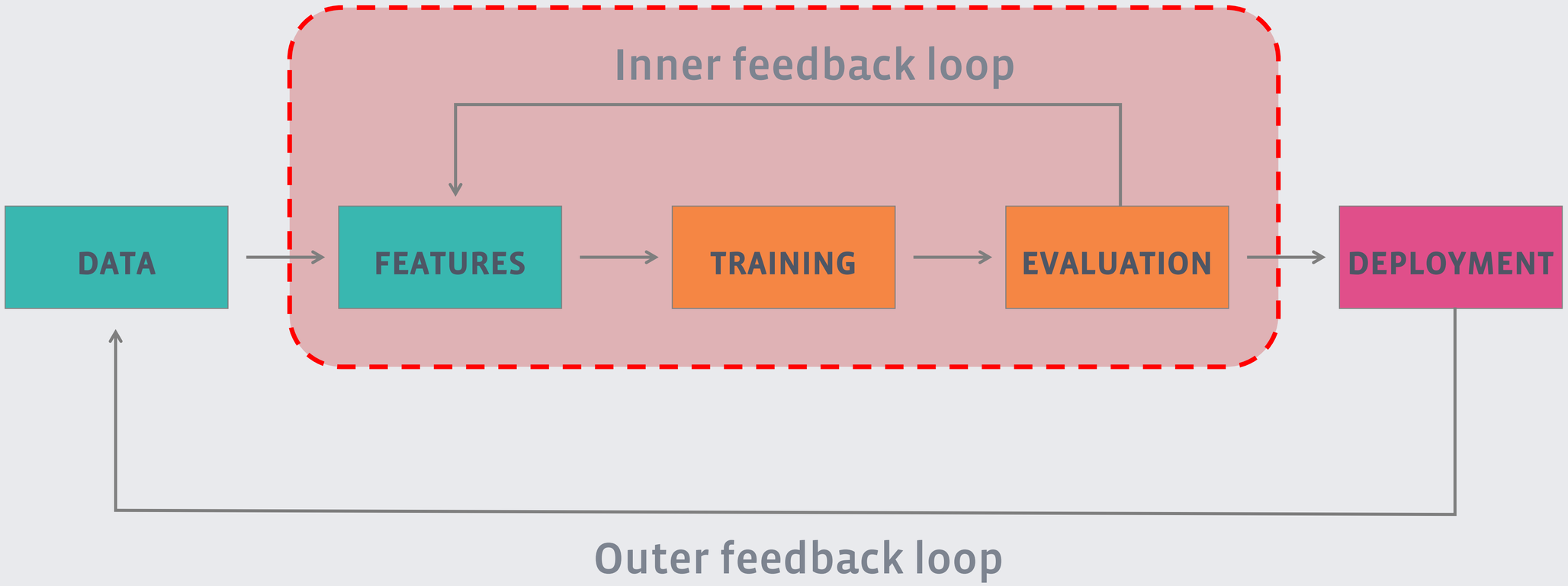
Experiment tracking



Source: https://databricks.com/blog/2019/09/24/a-guide-to-mlflow-talks-at-spark-ai-summit-2019-europe.html

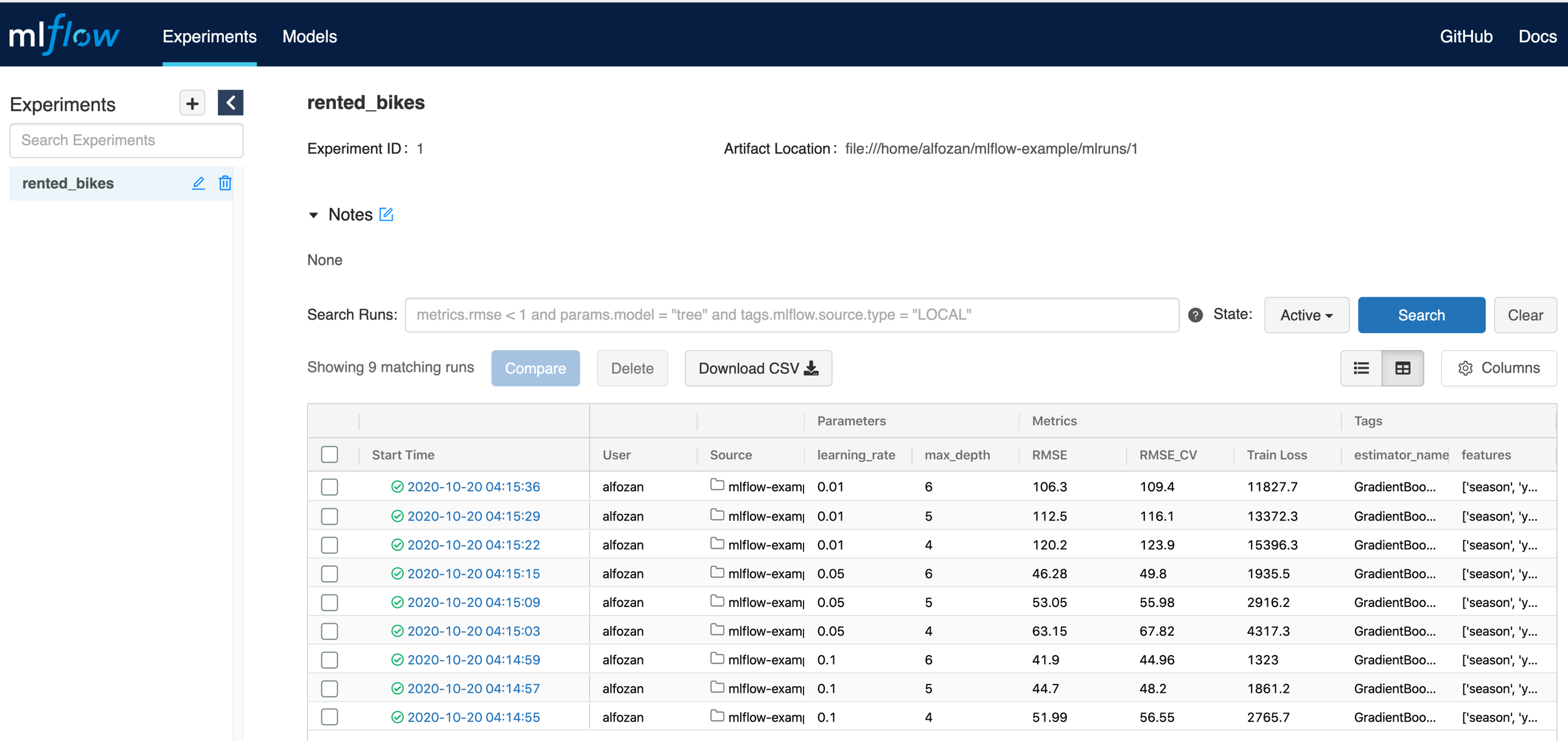
MLflow Components

Experiment tracking



Source: https://research.fb.com/publications/applied-machine-learning-at-facebook-a-datacenter-infrastructure-perspective/

MLflow tracking UI



				Parameters		Metrics			Tags	
	Start Time	User	Source	learning_rate	max_depth	RMSE	RMSE_CV	Train Loss	estimator_name	features
	2020-10-20 04:15:36	alfozan	mlflow-exam	0.01	6	106.3	109.4	11827.7	GradientBoo...	['season', 'y...
	2020-10-20 04:15:29	alfozan	mlflow-exam	0.01	5	112.5	116.1	13372.3	GradientBoo...	['season', 'y...
	2020-10-20 04:15:22	alfozan	mlflow-exam	0.01	4	120.2	123.9	15396.3	GradientBoo...	['season', 'y...
	2020-10-20 04:15:15	alfozan	mlflow-exam	0.05	6	46.28	49.8	1935.5	GradientBoo...	['season', 'y...
	2020-10-20 04:15:09	alfozan	mlflow-exam	0.05	5	53.05	55.98	2916.2	GradientBoo...	['season', 'y...
	2020-10-20 04:15:03	alfozan	mlflow-exam	0.05	4	63.15	67.82	4317.3	GradientBoo...	['season', 'y...
	2020-10-20 04:14:59	alfozan	mlflow-exam	0.1	6	41.9	44.96	1323	GradientBoo...	['season', 'y...
	2020-10-20 04:14:57	alfozan	mlflow-exam	0.1	5	44.7	48.2	1861.2	GradientBoo...	['season', 'y...
	2020-10-20 04:14:55	alfozan	mlflow-exam	0.1	4	51.99	56.55	2765.7	GradientBoo...	['season', 'y...

MLflow tracking UI

mlflow Experiments Models GitHub Docs

rented_bikes > Run a1cadce429dd46f29b64e58446ea5f9d

Date: 2020-10-20 04:15:36 Source: mlflow-example Entry Point: main

User: alfozan Duration: 1.7min Status: FINISHED

▸ Run Command

▸ Notes

▾ Parameters

Name	Value
learning_rate	0.01
max_depth	6

▾ Metrics

Name	Value
RMSE	106.3
RMSE_CV	109.4
Train Loss	11827.7

▾ Tags

Name	Value	Actions
estimator_name	GradientBoostingRegressor	
features	['season', 'year', 'month', 'hour_of_day', 'is_holiday', 'weekday', 'is_workingday', 'weather_situation', 'temperature', 'feels_like_temperature', 'humidity', 'windspeed']	

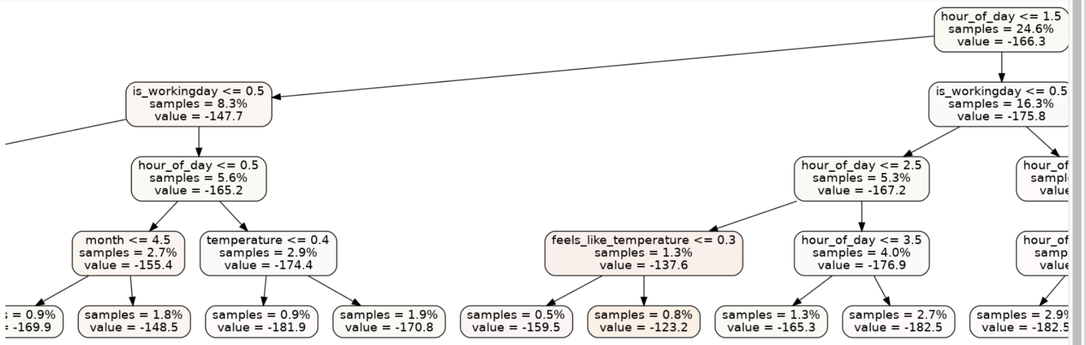
Add Tag

Name Value Add

▾ Artifacts

- model
  - MLmodel
  - conda.yaml
  - model.pkl
- Decision_Tree_Visualization.png
- feature_importance.png
- permutation_importance.png

Full Path: file:///home/alfozan/mlflow-example/mlruns/1/a1cadce429dd46f29b64e58446ea5f9d/artifacts/Decision_Tree_Visualization.png
Size: 804.31KB



MLflow tracking UI

Artifacts

- MLmodel
- conda.yaml
- model.pkl
- Decision_Tree_Visualization.png
- feature_importance.png
- permutation_importance.png

Size: 918B

```
artifact_path: model
flavors:
  python_function:
    env: conda.yaml
    loader_module: mlflow.sklearn
    model_path: model.pkl
    python_version: 3.7.9
  sklearn:
    pickled_model: model.pkl
    serialization_format: cloudpickle
    sklearn_version: 0.23.2
run_id: a1cadce429dd46f29b64e58446ea5f9d
signature:
  inputs: '[{"name": "season", "type": "long"}, {"name": "year", "type": "long"},
    {"name": "month", "type": "long"}, {"name": "hour_of_day", "type": "long"}, {"name":
    "is_holiday", "type": "long"}, {"name": "weekday", "type": "long"}, {"name": "is_workingday",
    "type": "long"}, {"name": "weather_situation", "type": "long"}, {"name": "temperature",
    "type": "double"}, {"name": "feels_like_temperature", "type": "double"}, {"name":
    "humidity", "type": "double"}, {"name": "windspeed", "type": "double"}]'
  outputs: '[{"type": "double"}]'
utc_time_created: '2020-10-20 11:17:19.015073'
```

mlflow Experiments Models GitHub Docs

rented_bikes > Run a1cadce429dd46f29b64e58446ea5f9d > Train Loss

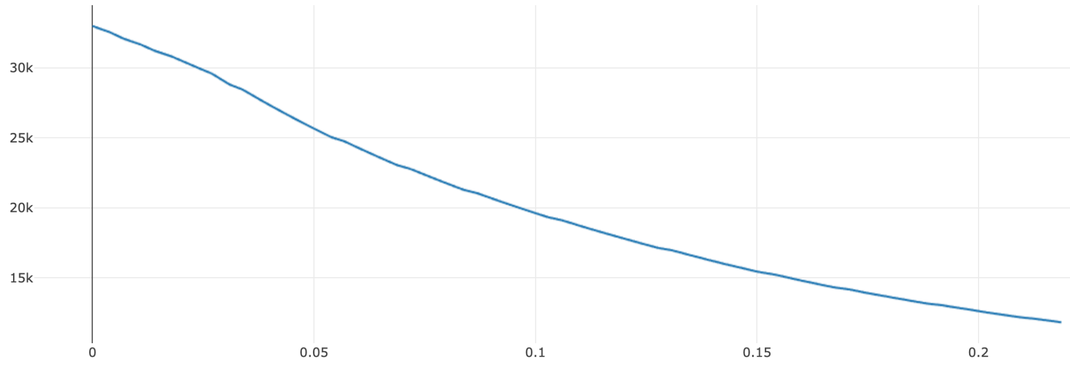
Points: Off

Line Smoothness 1

X-axis:
- Step
- Time (Wall)
- Time (Relative)

Y-axis: Train Loss ×

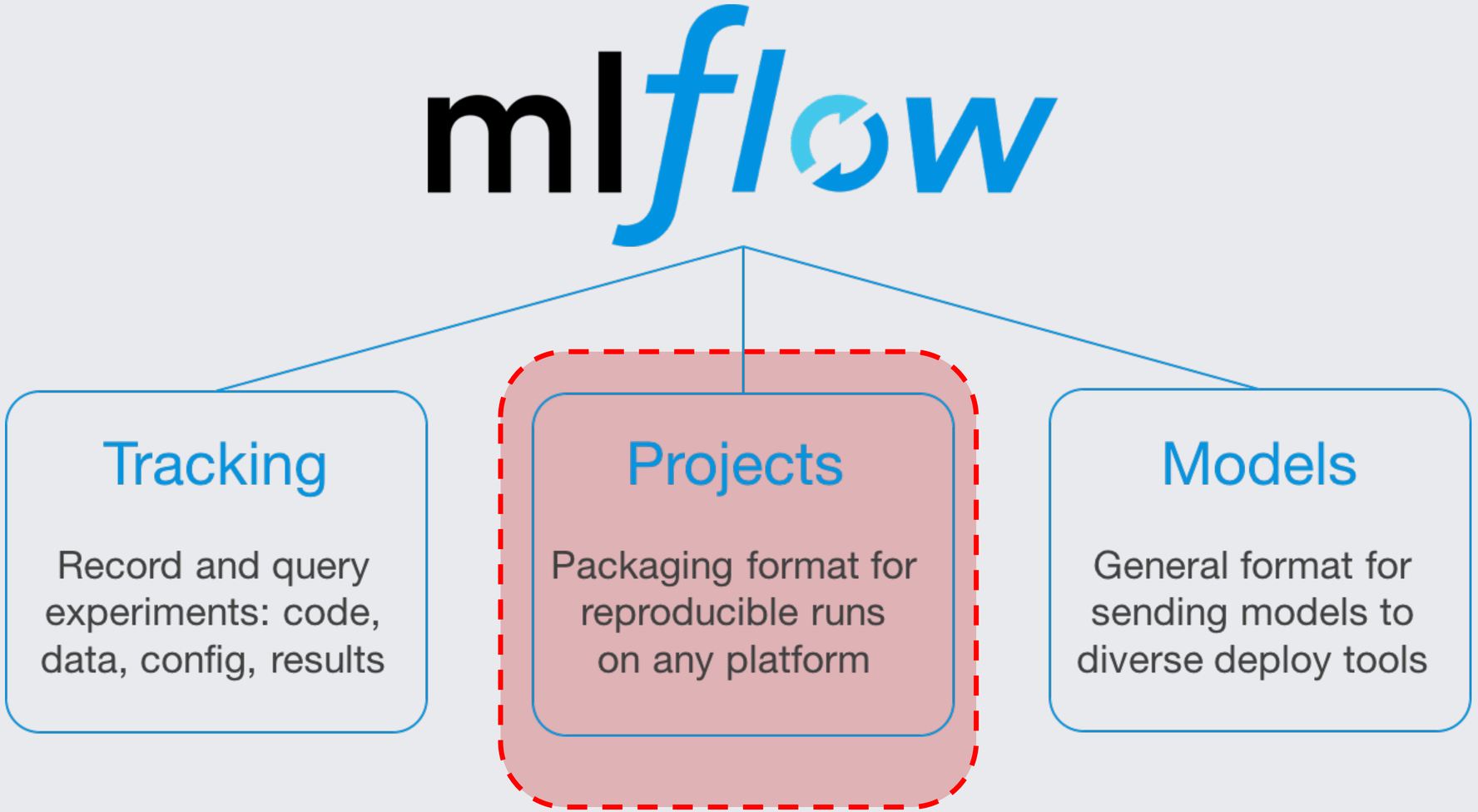
Y-axis Log Scale: Off



Demo

https://github.com/alfozan/mlflow-example

MLflow Components



Source: https://www.mlflow.org/docs/latest/projects.html

MLflow Projects

- a format for packaging data science code in a reusable and reproducible way, based primarily on conventions.
- In addition, the Projects component includes an API and command-line tools for running projects, making it possible to chain together projects into workflows.

Source: https://www.mlflow.org/docs/latest/projects.html

MLflow Projects

```
my_project/
├── MLproject
│
│
│
│
│
│
│
├── conda.yaml
├── main.py
└── model.py
    ...
```

```yaml
conda_env: conda.yaml

entry_points:
  main:
    parameters:
      training_data: path
      lambda: {type: float, default: 0.1}
    command: python main.py {training_data}
                    {lambda}
```

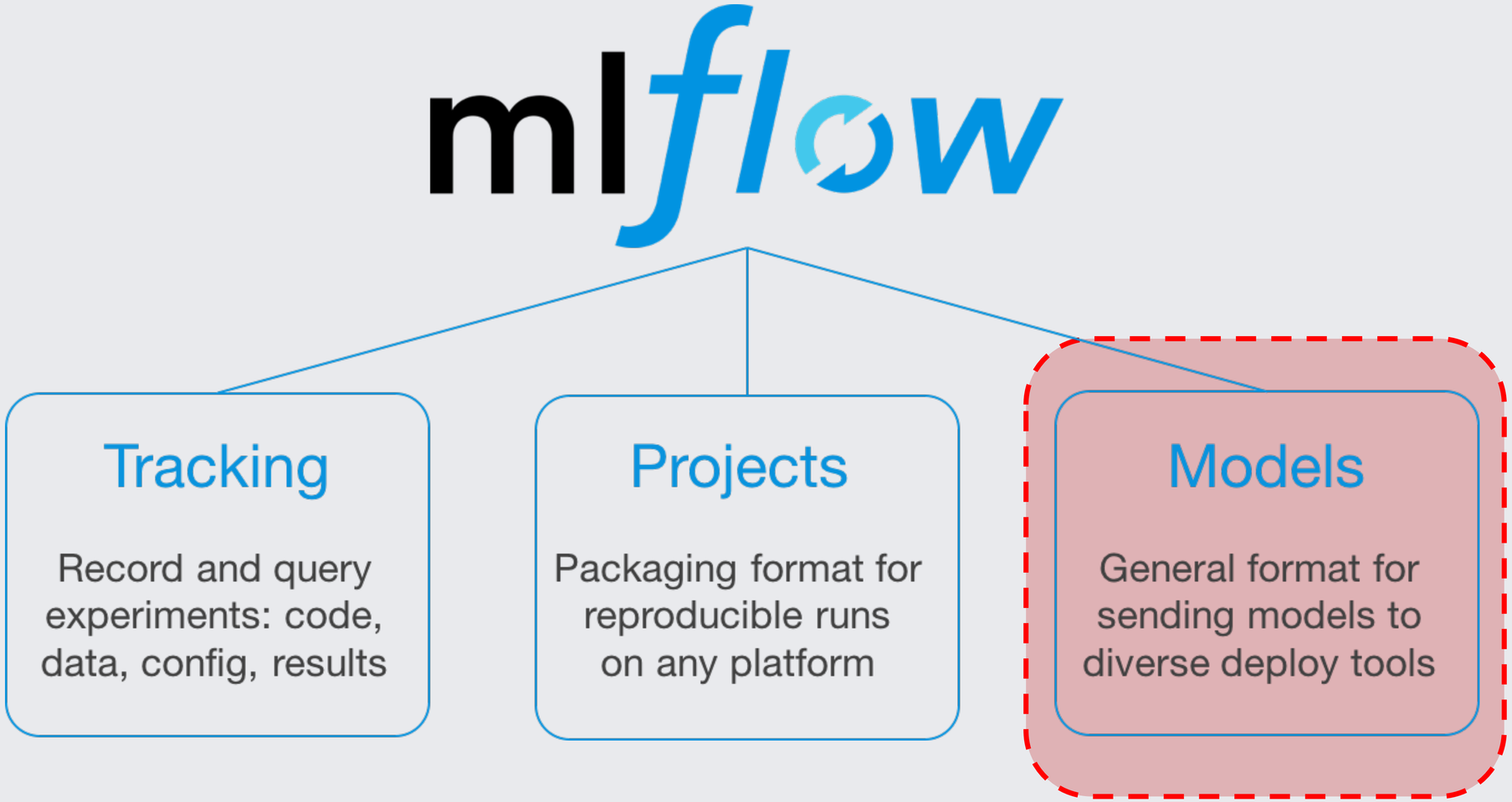
```
$ mlflow run git://<my_project>
```

Source: https://databricks.com/blog/2019/09/24/a-guide-to-mlflow-talks-at-spark-ai-summit-2019-europe.html

Demo

https://github.com/alfozan/mlflow-example

MLflow Components



Source: https://www.mlflow.org/docs/latest/models.html

MLflow Models

- a standard format for packaging machine learning models that can be used in a variety of downstream tools—for example, real-time serving through a REST API or batch inference.

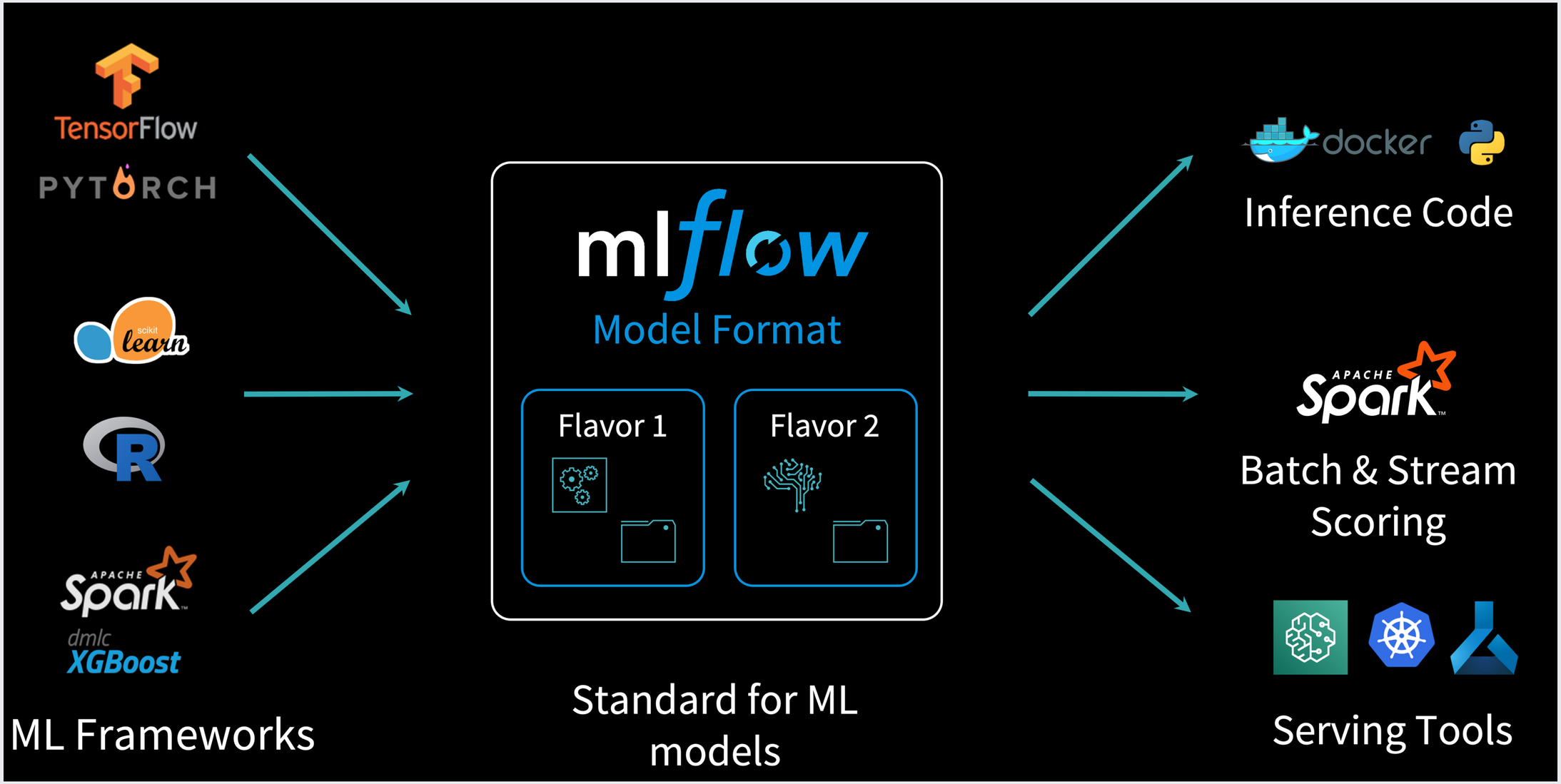
Source: https://www.mlflow.org/docs/latest/models.html

MLflow Models

- Packaging format for ML Models: MLmodel file
- Defines dependencies for reproducibility
- Model creation utilities
  - Save models from any framework in MLflow format
- Deployment APIs:
  - CLI / Python / R / Java

Source: https://www.mlflow.org/docs/latest/models.html

MLflow Models



Source: https://databricks.com/blog/2019/09/24/a-guide-to-mlflow-talks-at-spark-ai-summit-2019-europe.html

Demo

https://github.com/alfozan/mlflow-example

Questions