An open platform for the machine learning lifecycle

Abdulrahman Alfozan

PyData Riyadh, August 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
4. Demo

ML Pipeline Lifecycle

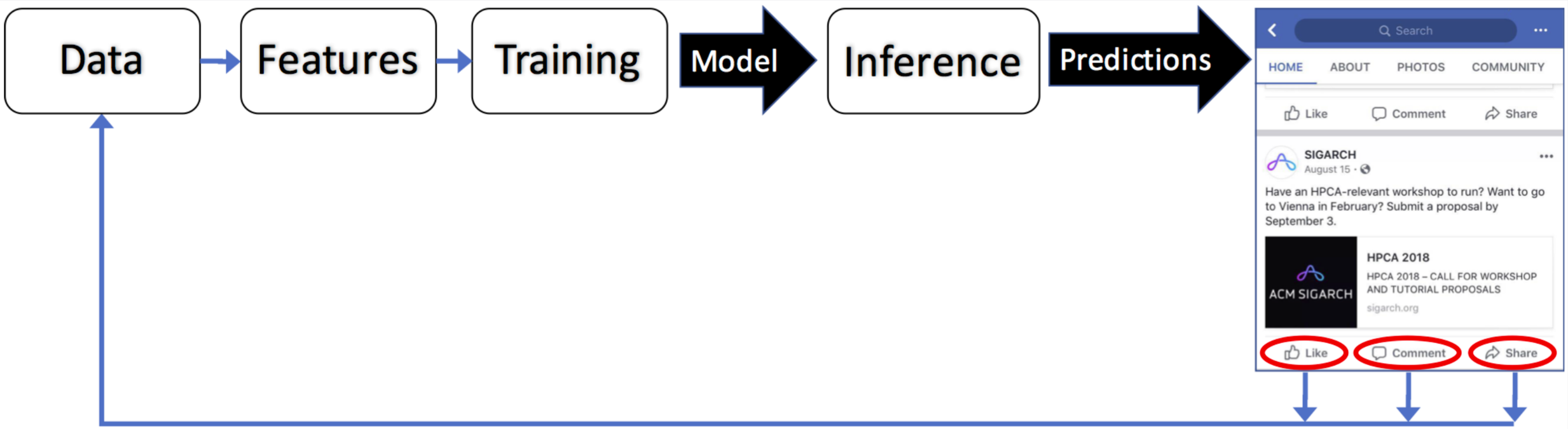


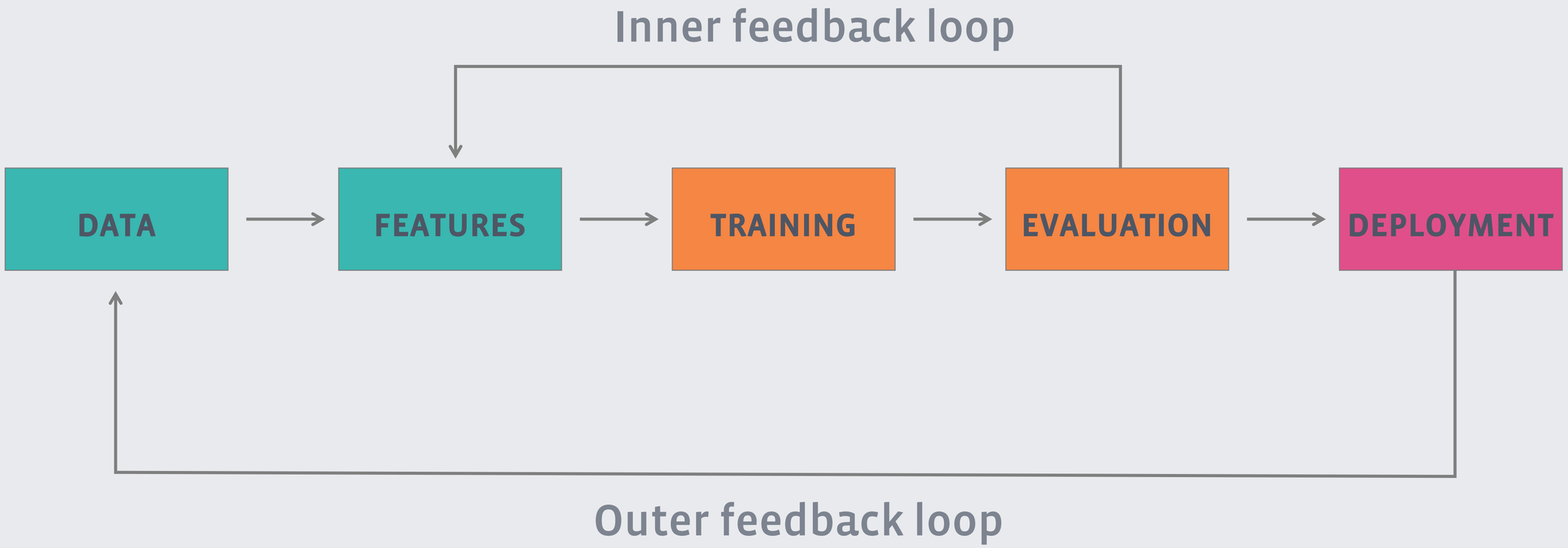
Fig. 1. Example of Facebook's Machine Learning Flow and Infrastructure.

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
4. Demo

“Hidden Technical Debt in Machine Learning Systems”

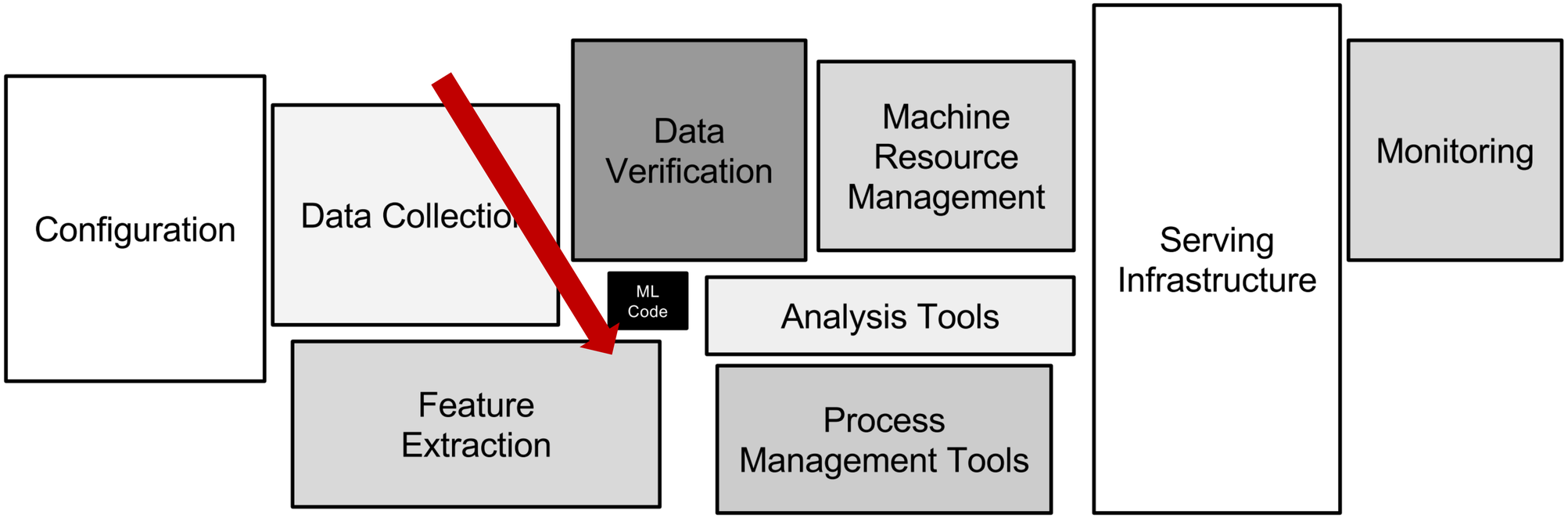


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

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

ML development challenges

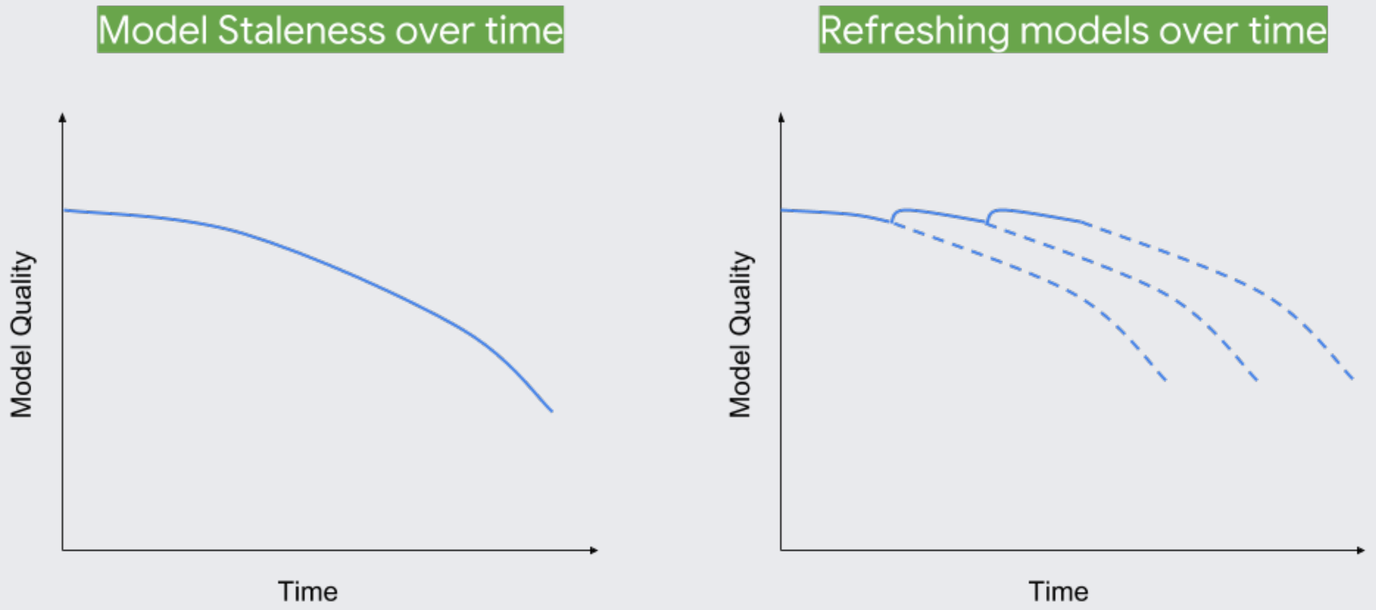
1) Experiment Standardization and Tracking

- Reproducibility
- Parameters Tunning

ML development challenges

2) Model Staleness

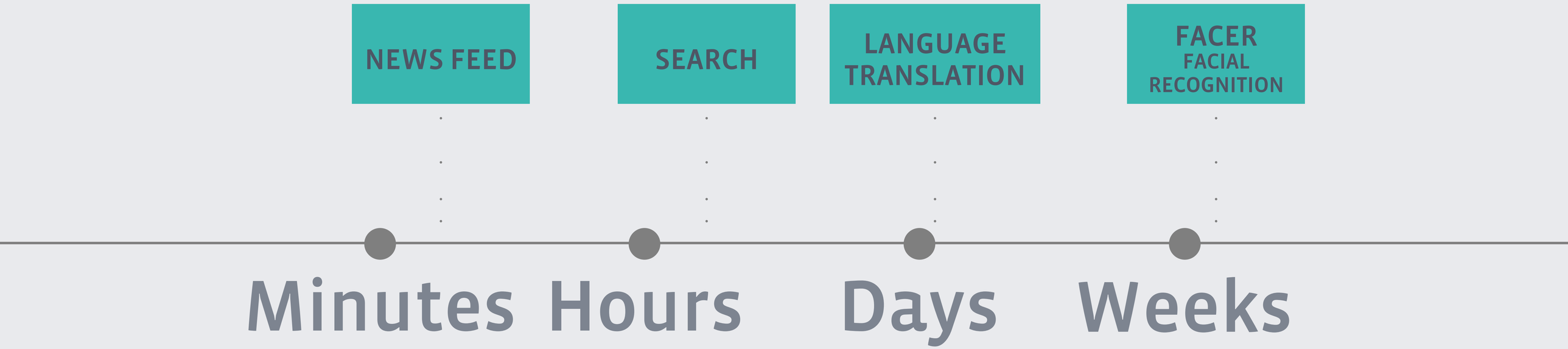
ML pipelines need to be constantly run with new data to avoid model staleness and close feedback loop



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

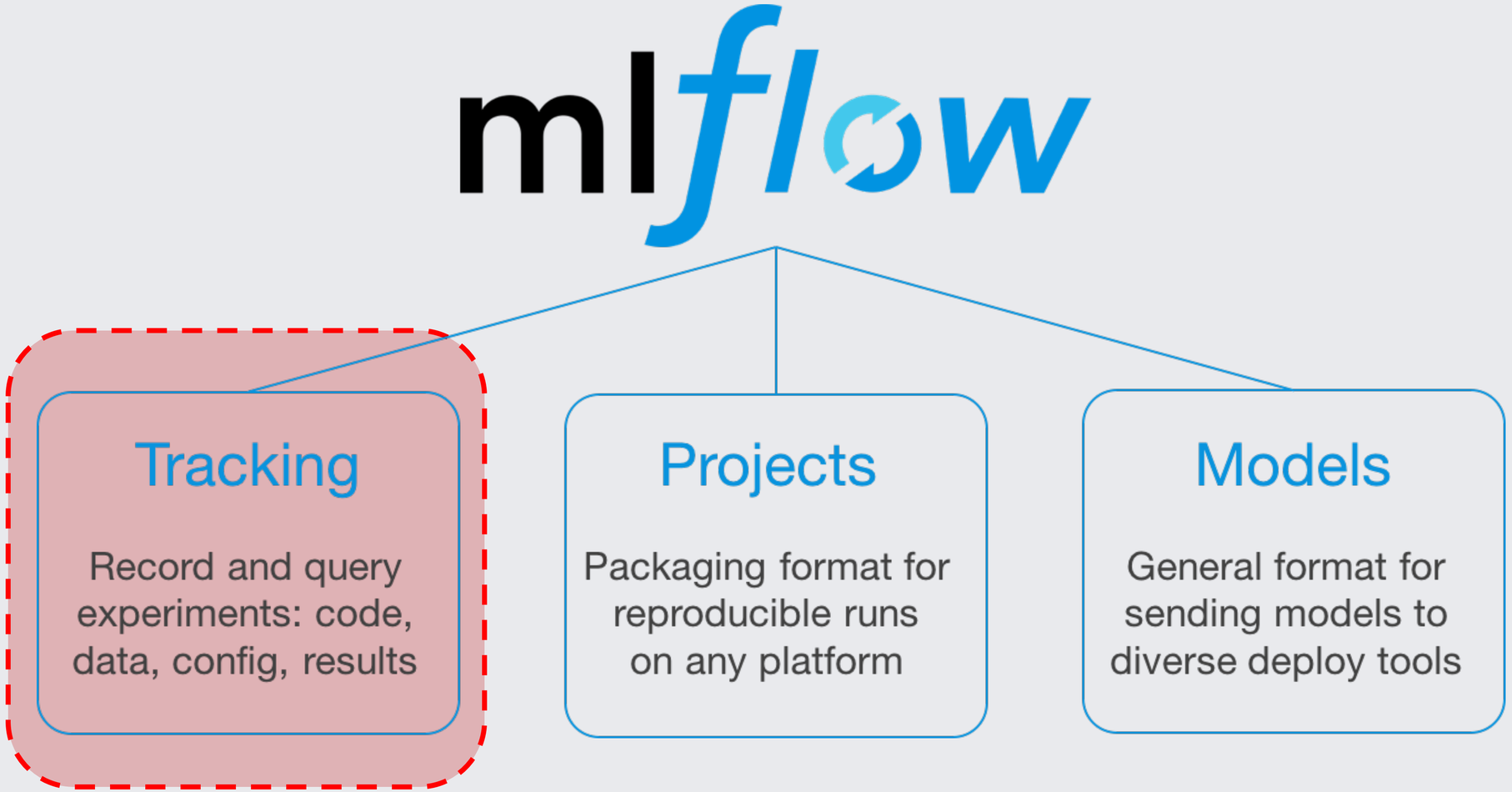
1. ML Pipeline Lifecycle
2. ML development challenges
3. MLflow platform
4. Demo

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

MLflow Components



Source: https://mlflow.org/

MLflow Components

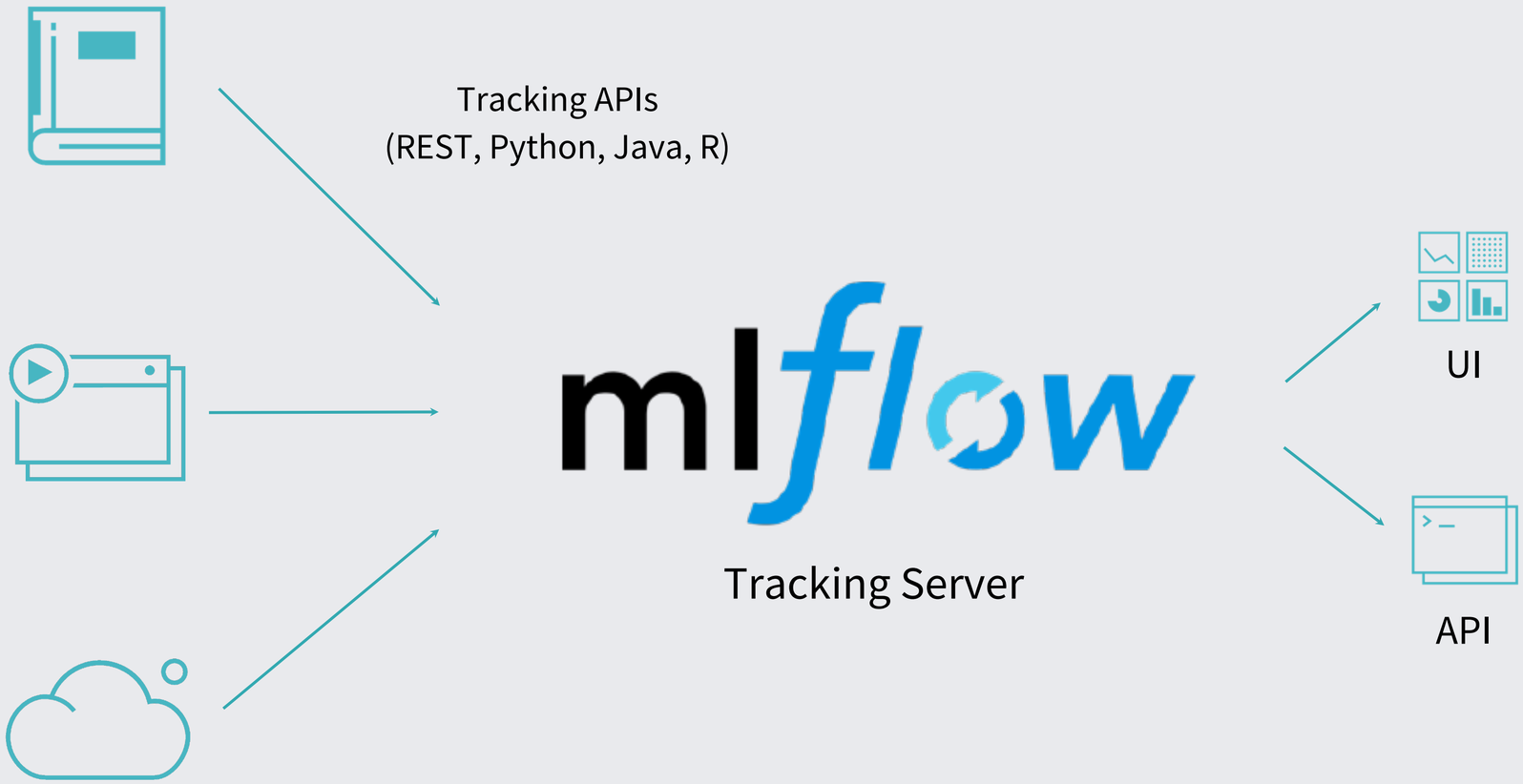
Experiment tracking

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

any additional information

MLflow Components

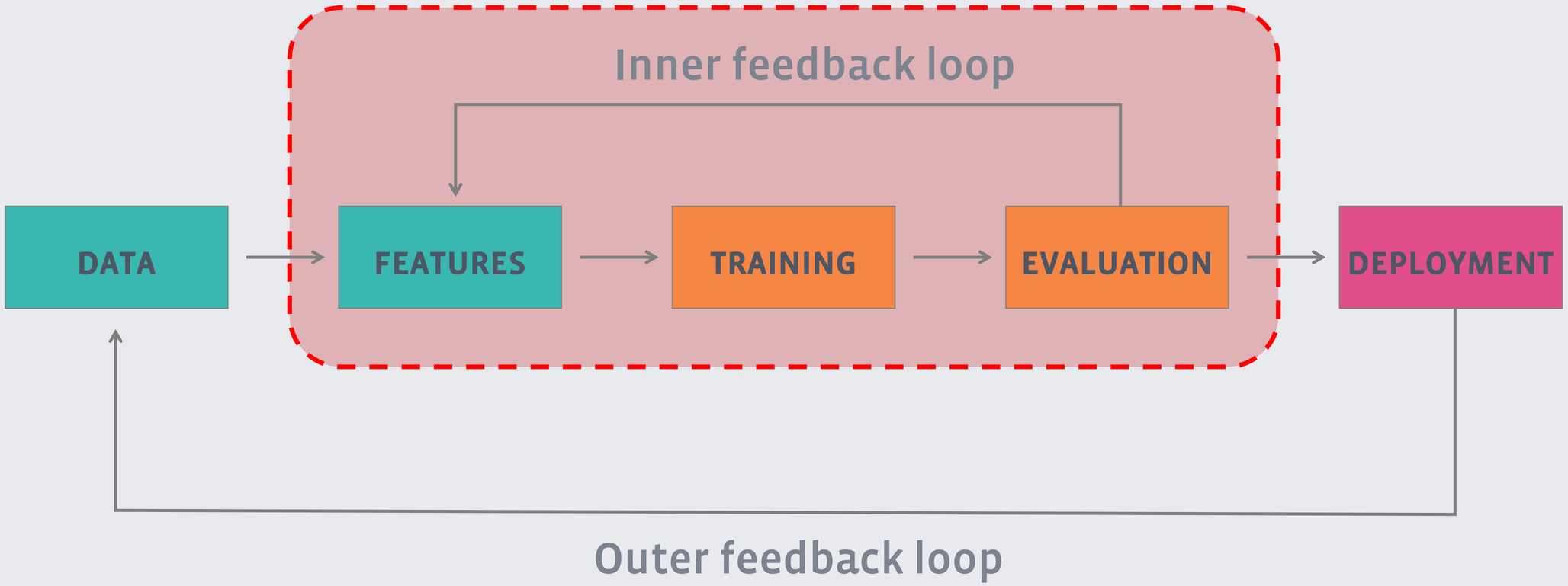
Experiment tracking



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

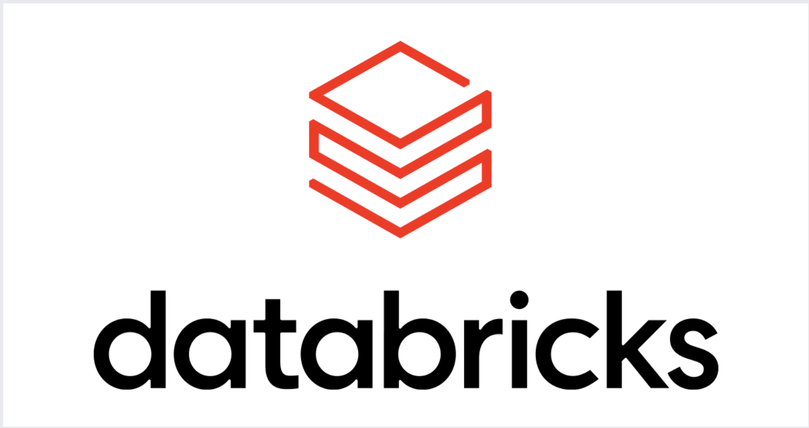
MLflow Components

Experiment tracking



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

Demo

https://github.com/alfozan/MLflow-GBRT-demo/blob/master/MLflow-GBRT-demo.ipynb

Questions

Thank you