ML PIPELINE

Data in Academia

In Academia

- collected for research purposes
- focuses on understanding a specific phenomenon / hypothesis.
- usually collected in a controlled environment
- sample size is often small (relatively)
- often collected with a high level of accuracy and precision

In Production

- collected in a real-world setting
- the goal is making decisions or improving a particular process
- collected on a large scale
- come from a variety of sources, (sensors, customer interactions, ...)
- often collected with a focus on efficiency and cost-effectiveness, and may not be as accurate or precise

Deploying on Mobile

- With regards to mobile devices, there are 2 deployment methods:
  1. On a server: user access through a network
  2. On the user device

Deploying on Mobile (On user device)

Factors driving this Trend:

- More capable devices
- Cost-effectiveness: no need to pay server bills
- Privacy: data stays on user's device

Deploying on Mobile (Challenges)

On user device

- Less processing power
- Need smaller models
- Updates demand an app update

On server

- Cost optimization
- Need constant network connection
- Privacy concerns

Pipelines

- Using TensorFlow input pipeline for high performance data ingestion (the use of parallelization)
- High performance modeling: overcoming memory constraints for training large models

Model Monitoring

Why do we need to monitor models?

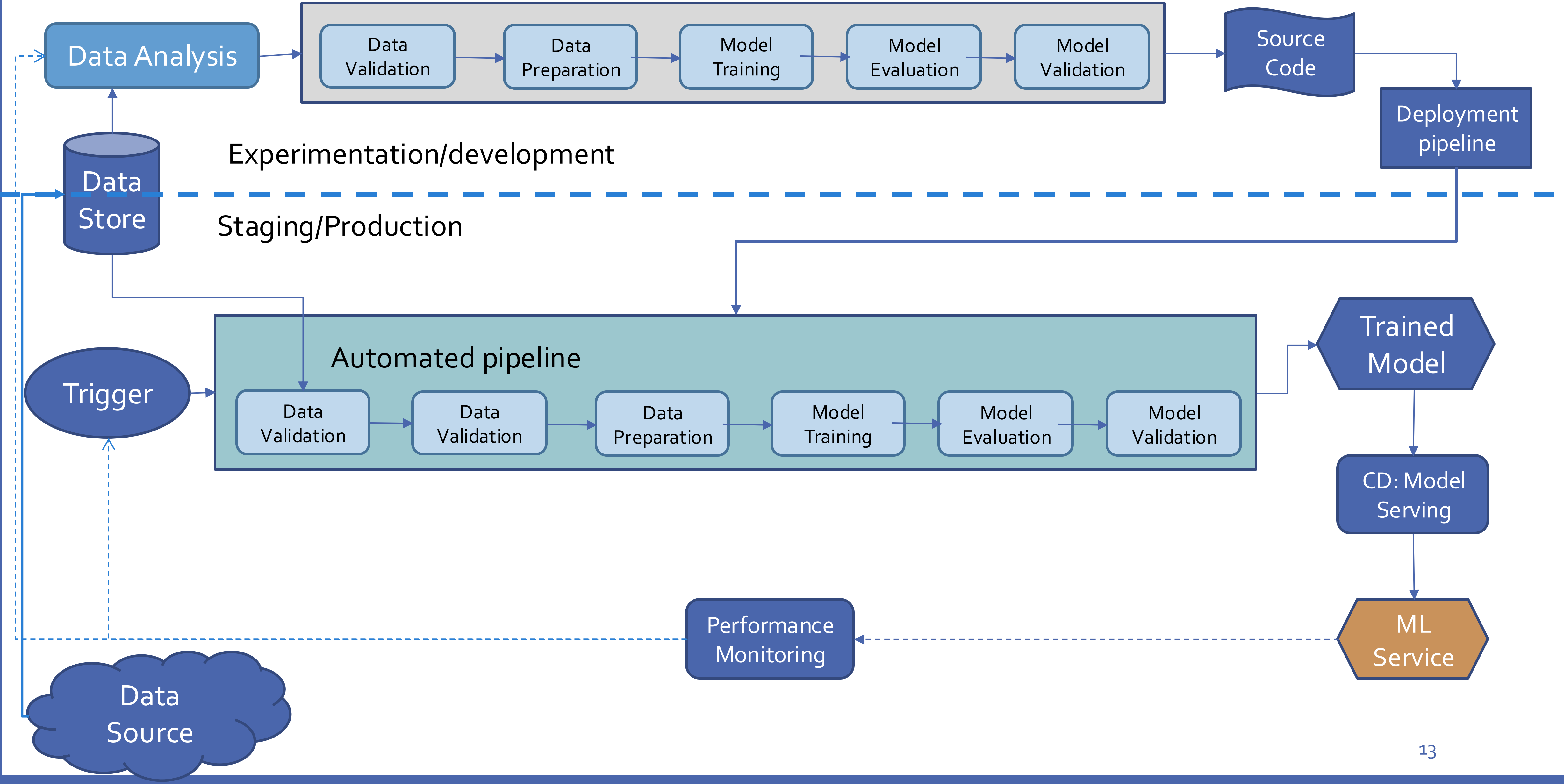
- Data changes over time, which means the model output at one point might not be correct anymore after some time (due to data drift and shift).
- We need early warnings that the model performance is changing.

Model Analysis

- After training and deployment, you might notice that your model performance is dropping
- Check the possibility for performance improvement
- Anticipate data changes

Model Analysis

- 2 main ways to analyze the performance of your model:
  - Black box evaluation: you are not interested in the internal structure of your model, you just look at the performance metrics (example tool is **TensorBoard**)
  - Model introspection: for inspecting more fine-tuned details of model performance and data flow within the model (**TensorFlow Model Analysis TFMA**).


Data Analysis
Data Validation
Data Preparation
Model Training
Model Evaluation
Model Validation
Source Code
Deployment pipeline
Data Store
Experimentation/development
Staging/Production
Trigger
Automated pipeline
Data Validation
Data Validation
Data Preparation
Model Training
Model Evaluation
Model Validation
Trained Model
CD: Model Serving
ML Service
Performance Monitoring
Data Source