TinyMLOps:
Scaling Deployments

Colby Banbury
PhD Student
Edge Computing Lab, Harvard University
“Hidden Technical Debt in ML Systems”

Source: D. Scully, 2015
“Hidden Technical Debt in ML Systems”
Close The Loop

- Collect Data
- Preprocess Data
- Design a Model
- Train a Model
- Evaluate Optimize
- Convert Model
- Deploy Model
- Make Inferences
MLOps = ML Workflow + Automation
MLOps means...

5 Focus Areas

Running end-to-end
Managing Complexity
Evaluating Results
Improving Models
Tracking deployments
Running System

Traditional System Testing

ML-based System Testing

- Code
- Running System
- Unit Tests
- Integration Tests
- System Monitoring

ML Infrastructure Tests
Model Tests

Data Tests
Skew Tests
Data Monitoring

Training code
Running System
Running System Tests

- Unit Tests
- Integration Tests
- System Monitoring

Traditional System Testing

Training code

Running System

- Data Tests
- Skew Tests
- Data Monitoring

ML Infrastructure Tests

Model Tests

ML-based System Testing
MLOps Tools

- Custom Scripts
- Frameworks
- End-to-end Platforms

- Python
- TensorFlow Extended (TFX)
- EDGE IMPULSE
Edge Impulse Workflow

Data Collection
- Data Collection
- Data Analysis

Development & Deployment
- DSP Pipeline
- ML Design & Training
- Estimation & Evaluation
- Optimization & Compression
- Conversion & Compilation

Monitoring
- IoT Device Management
- Production Monitoring

Edge Impulse Workflow

- Raw Data Store
- Metadata & Labels
- Feature Store
- Model Registry
- Firmware Versions

Active Learning

AutoML
Edge Impulse Workflow

**Challenges Addressed**
- Data Collection
- Development & Deployment
- Monitoring
- Co-Optimization & Cross-Stack Collaboration
- Software Fragmentation
- Hardware Heterogeneity

**Edge Impulse Workflow**
- Raw Data Store
- Metadata & Labels
- Feature Store
- Model Registry
- Firmware Versions

**Active Learning**
**AutoML**
Basic features of MLOps tool

Way to manage and process data

Way to train models

Way to version models

Way to evaluate and compare models

Way to deploy models
The MLOps Process

- ML Development
  - Code & Config
- ML Training
  - Training Pipeline
- Continuous Training
  - Registered Model
- Model Deployment
  - Serving Package
- Prediction Serving
  - Serving Logs
- Continuous Monitoring
Continuous Monitoring Example
Concept Drift

Decision boundary fit to training data
Concept Drift

Decision boundary fit to training data

Concept drift, detrimental performance impact
Li-ion Battery Example
Acceptable Threshold for Performance

Model Drift
Trigger Retraining

Class Frequency Shift?
Poor performance on test data?
Connectivity Trade-Off
Connectivity Trade-Off
Connectivity Trade-Off
You’re **always** doing MLOps, so make sure you do it well.
MLOps for Scaling TinyML

This course introduces learners to Machine Learning Operations (MLOps) through the lens of TinyML (Tiny Machine Learning). Learners explore best practices to deploy, monitor, and maintain (tiny) Machine Learning models in production at scale.

Choose your session:
After a course session ends, it will be archived.

- **Starts Apr 18**
- **Starts Apr 26**

Enroll

Free
Optional upgrade available