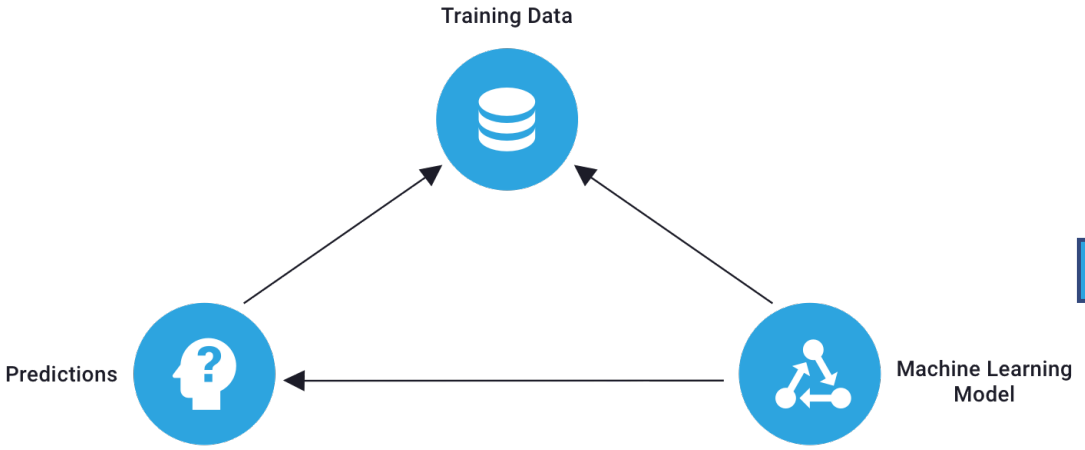
MLOPS INTRODUCTION

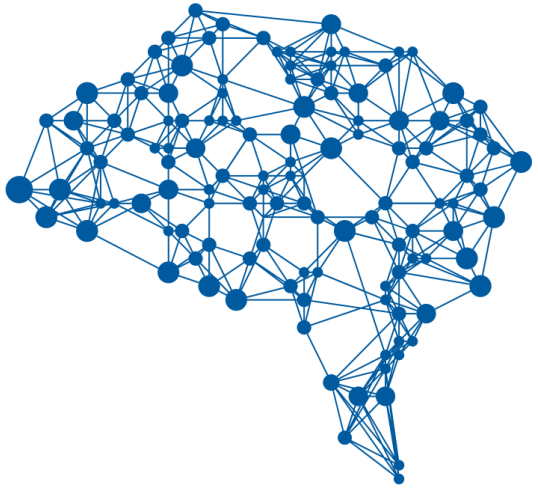
What is MLOps

Definition

- Wikipedia: MLOps or ML Ops is a set of practices that aims to deploy and maintain machine learning models in production reliably and efficiently
- Microsoft: MLOps is based on DevOps principles and practices that increase the efficiency of workflows. Examples include continuous integration, delivery, and deployment.
- Amazon: MLOps refers to a methodology that is built on applying DevOps practices to machine learning workloads.

Traditional ML Model Creation





Production ML is Much More

ML Workflow:

- Configuration
- Data Collection
- Feature extraction
- Data Verification
- ML Model (aprox. Just 5% of the code required to put an ml application into production)
- Machine Resource Management
- Analysis Tools
- Process Management Tools
- Serving Infrastructure
- Monitoring

ML Modeling VS Production

Academic ML	Production ML
static data	dynamic-shifting data
priority is highest accuracy	priority is fast inference / good results / low cost
model training is based on a singular result	model is continuously re-trained and assessed
fairness (model doesn't descriminate) is important	fairness (model doesn't descriminate) is crucial
main challenge is high accuracy	main challenge is the entire system

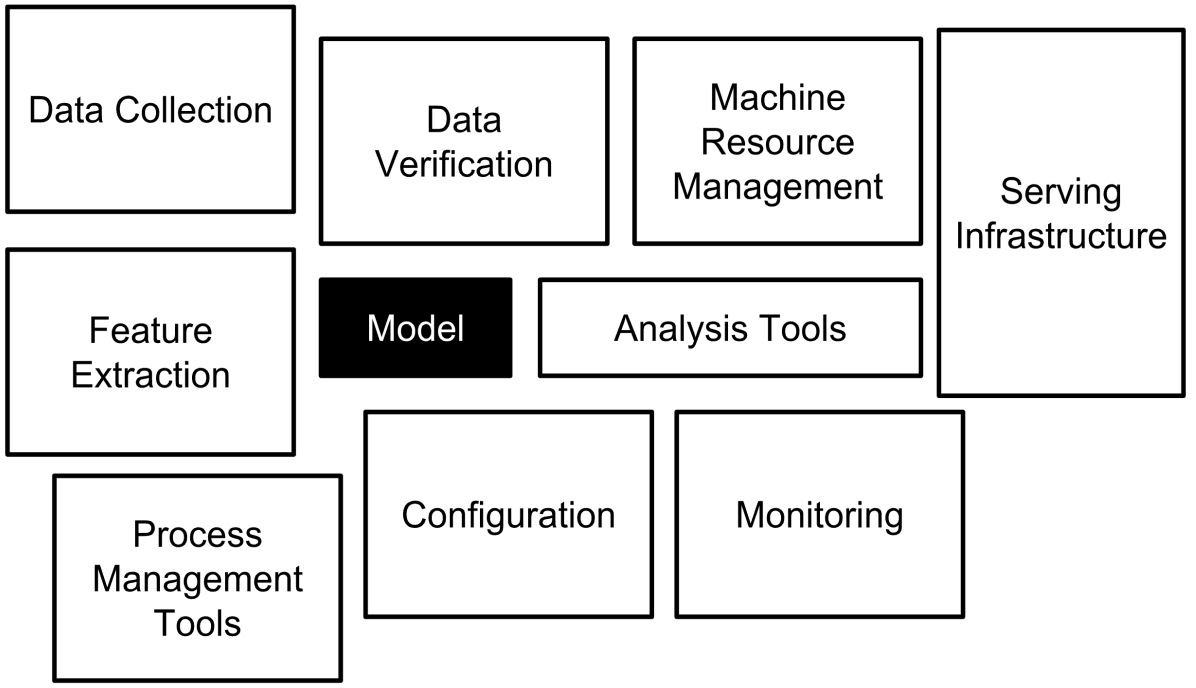
Learning From Software Engineering

Developing a machine-learning application has similar problems as modern software development:

- Scalability
- Extensibility
- Clear configuration
- Consistent
- Security
- Modularity
- Testability
- Monitoring

Learning From Software Engineering

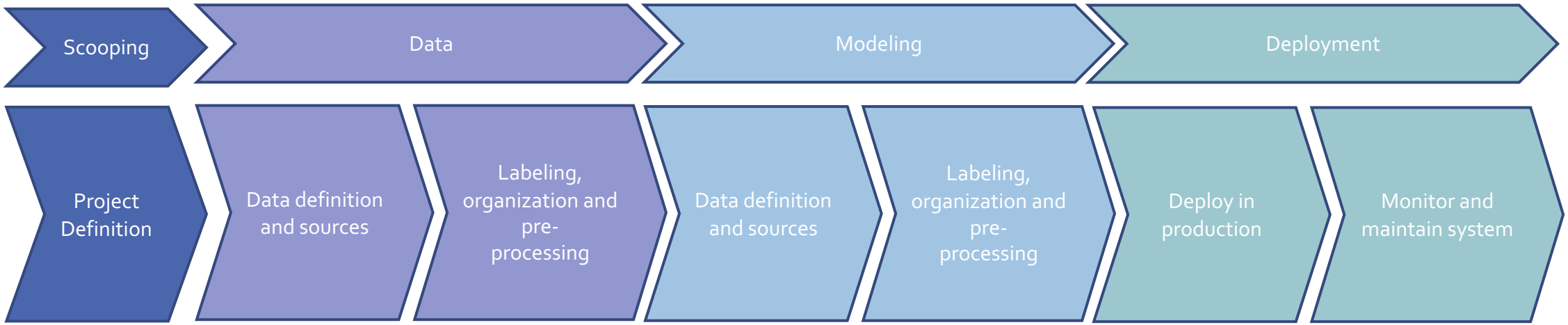
- When deploying an ML application, it goes through a repetitive cycle of an ML workflow.
- This is made possible through an ML pipeline.
- Orchestrators (like Tensorflow Extended, or TFX) can help build such pipelines.



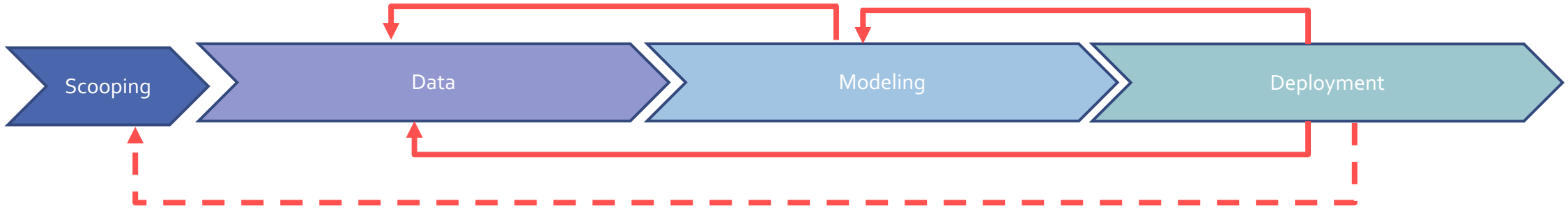
A schematic of a typical machine learning pipeline.

reference: https://developers.google.com/machine-learning/testing-debugging/pipeline/overview

Production ML System



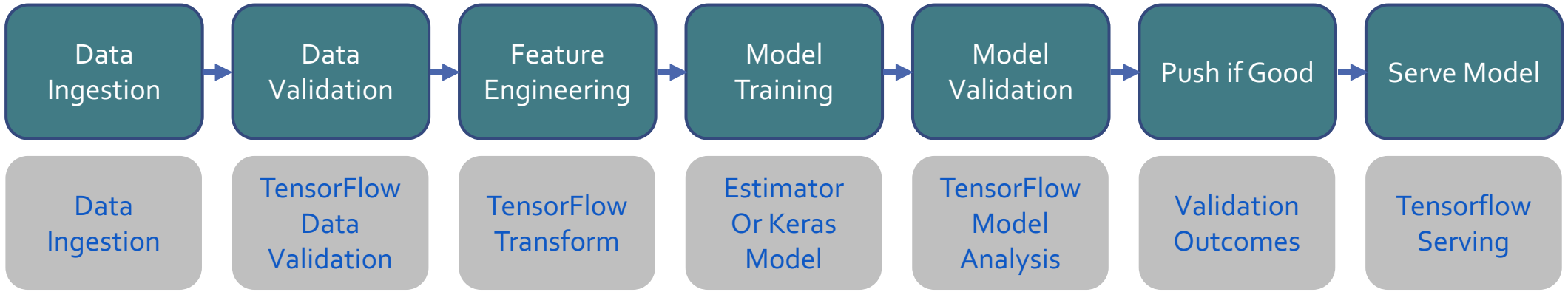
Production ML System



TensorFlow Extended (TFX)

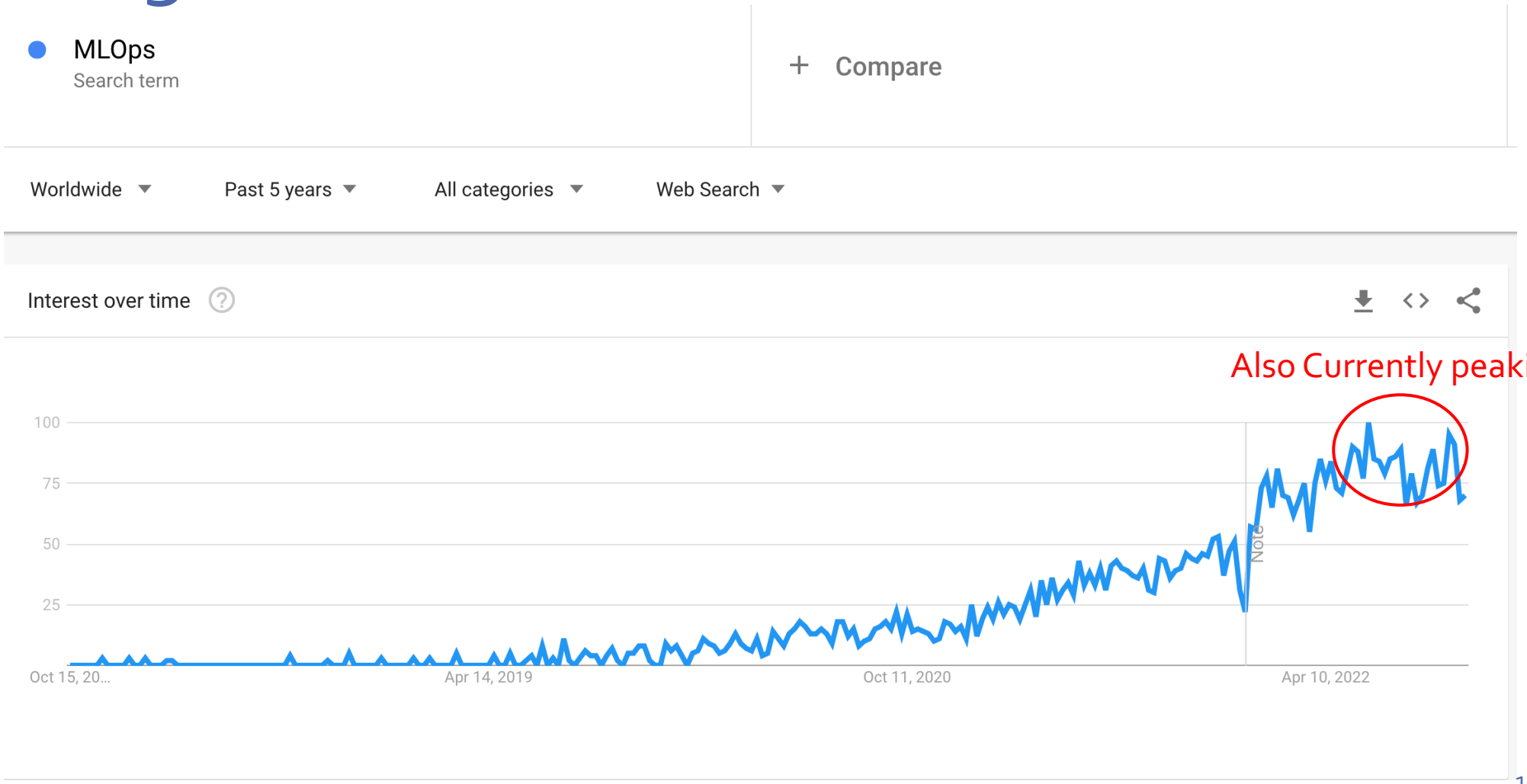
- Is an end-to-end platform for deploying production ML pipelines
- TFX pipeline: a sequence of scalable components that can handle large volumes of data.

TFX Production Components

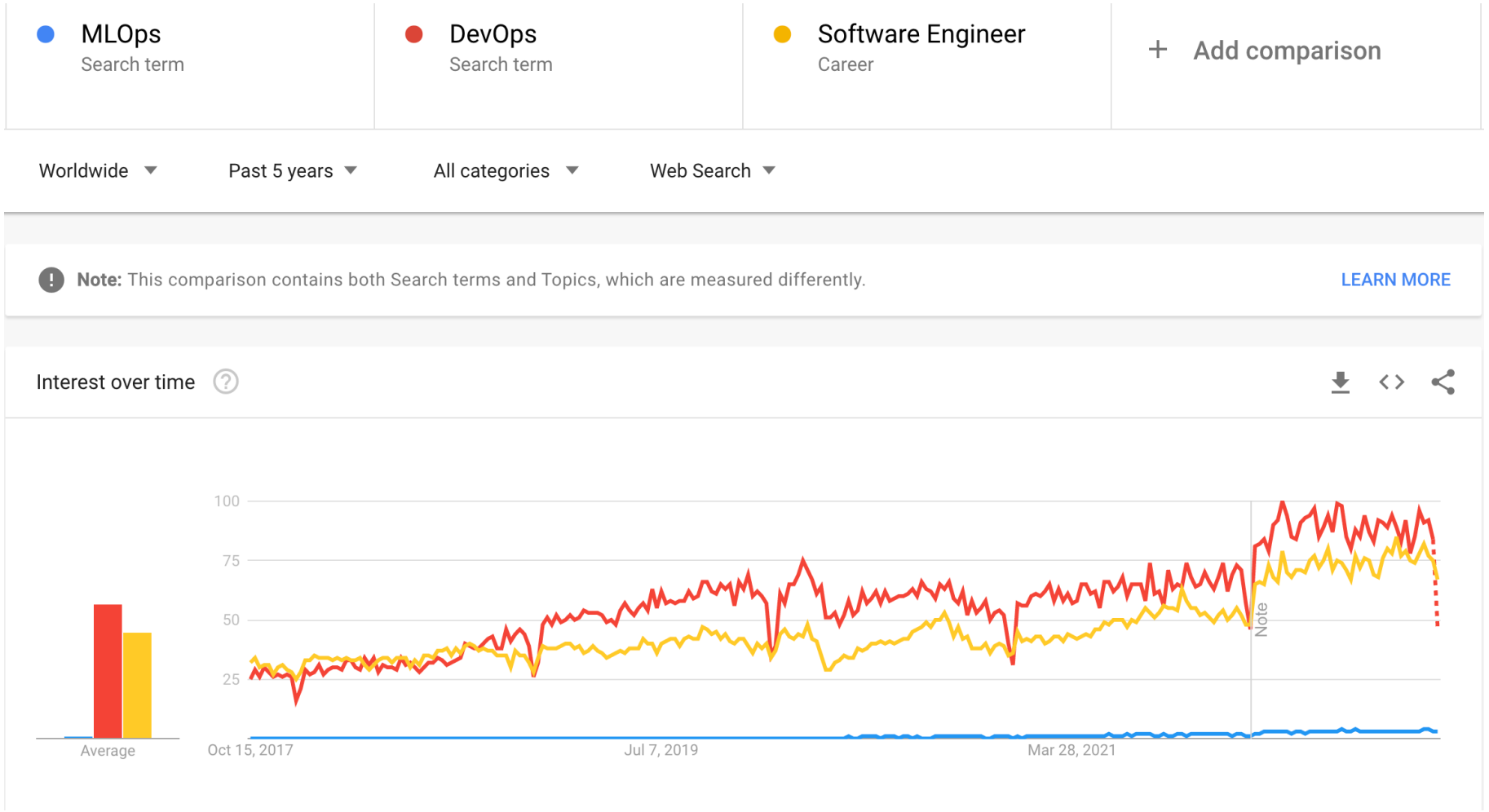


Motivation

Google Trends (World Wide)



Google Trends (World Wide)



Model Deployment

- ML Kit (Firebase): targets mobile platforms and uses TensorFlow lite.
- Core ML (Apple): train your own ML model and deploy on apple devices
- TensorFlow Lite (Google): optimized for on-device deployment, and supports IoT devices