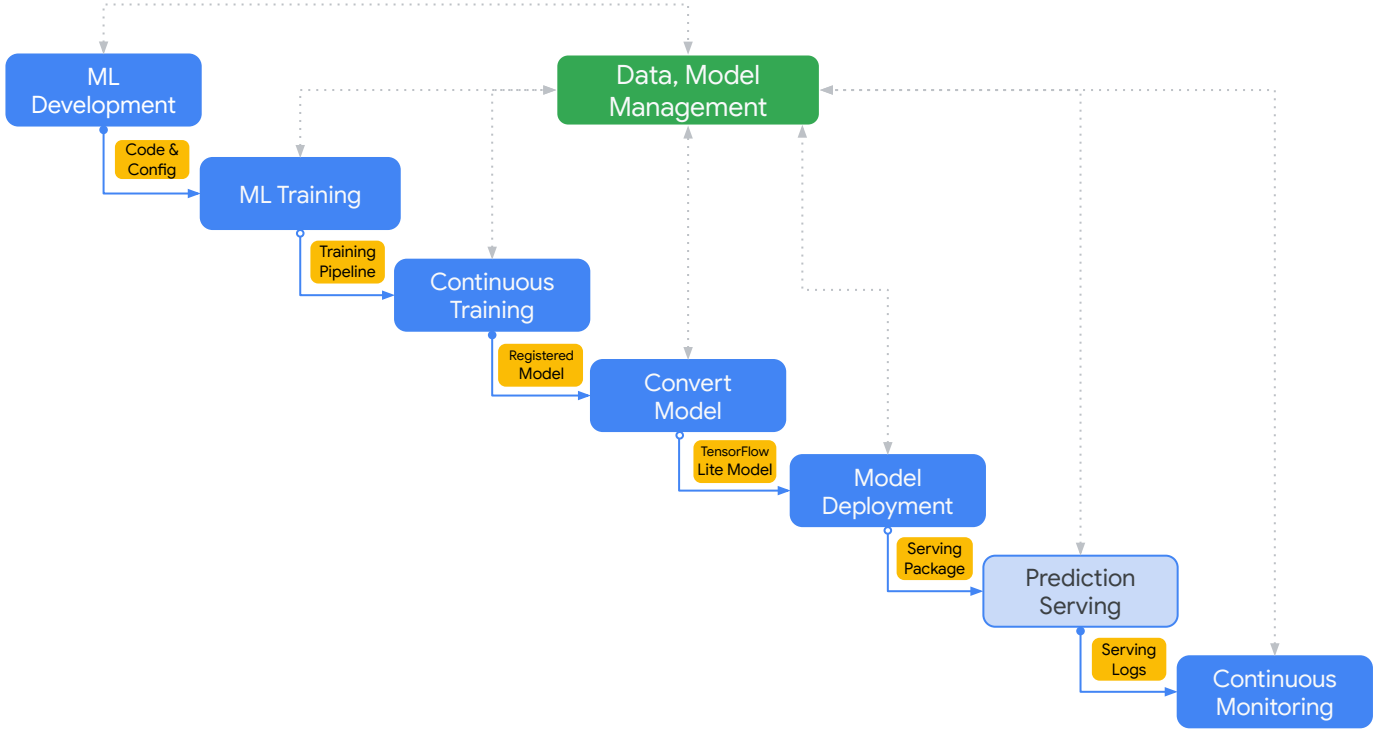
Prediction Serving

Provisioning online inference serving

Dr Lara Suzuki
Google

The MLOps Process for **TinyML**



MLOps: Prediction Serving

Concerns about **serving** the model that is deployed in production for inference. It includes tasks such as accepting prediction requests (serving data) and to serve responses.

Core MLOps Capabilities:

- Dataset & feature repository
- Model serving

MLOps: Prediction Serving