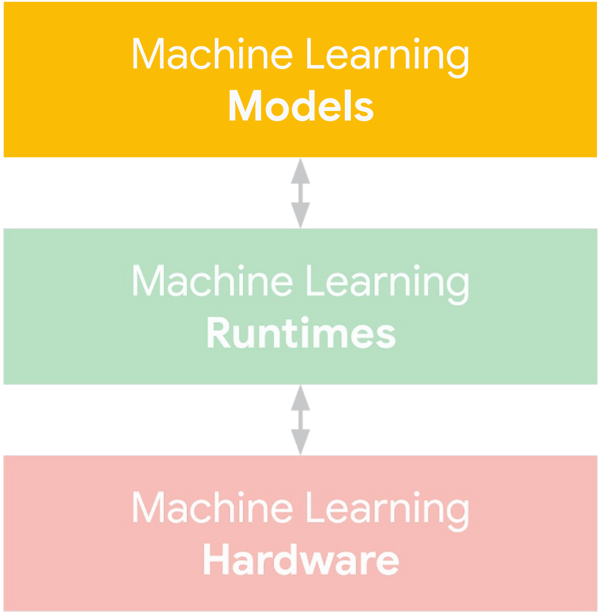
What are the Challenges for TinyML?

Part D



Model Compression Techniques

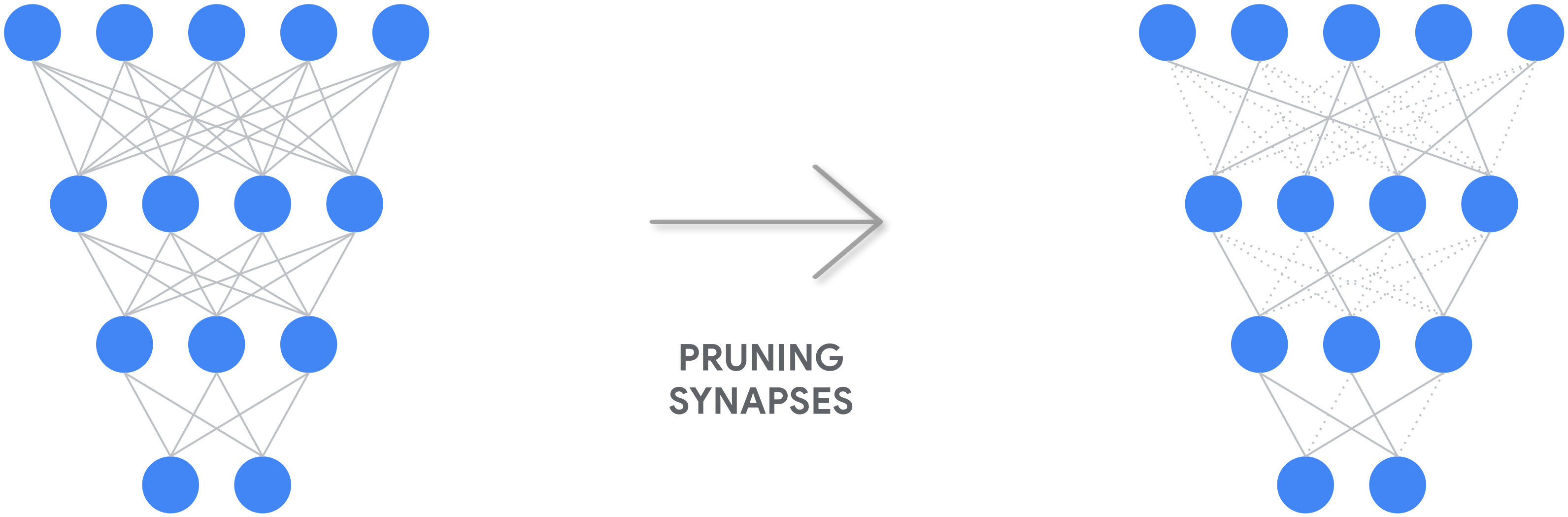
Pruning

Quantization

Knowledge Distillation

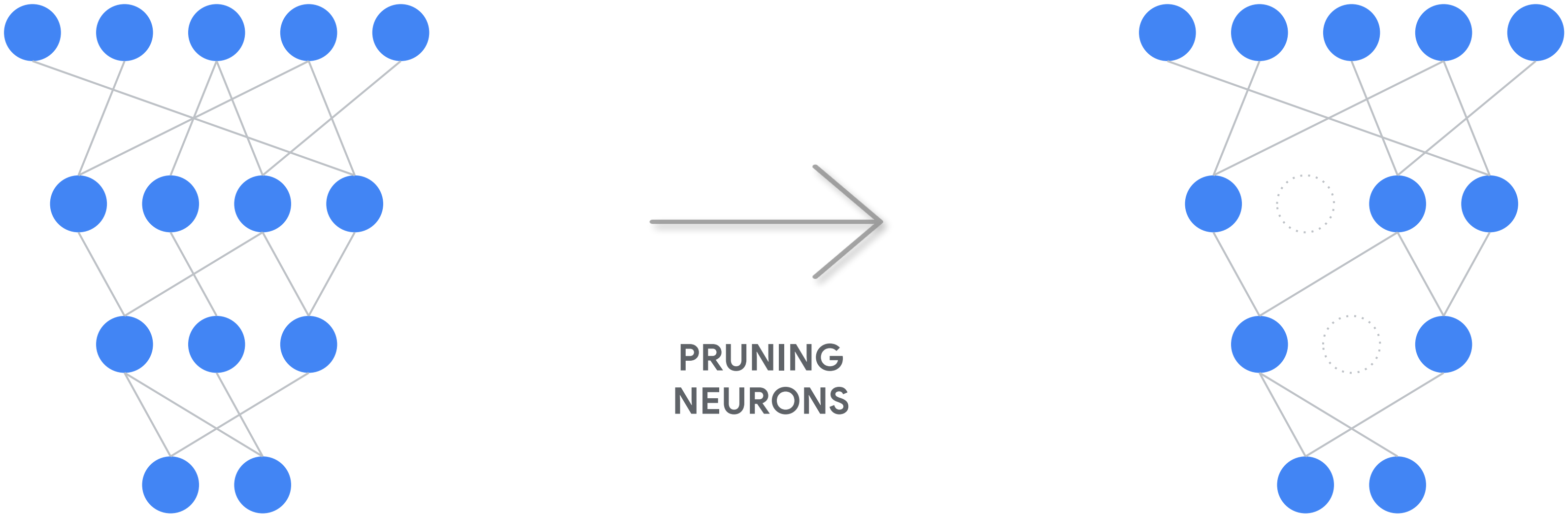
...

Pruning

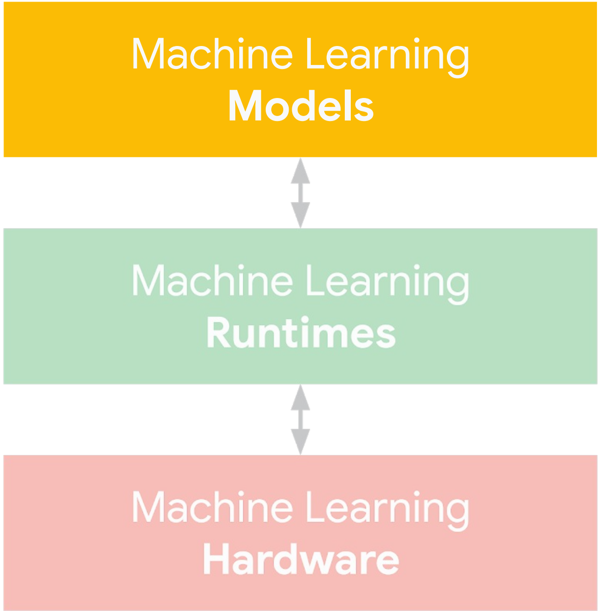


PRUNING SYNAPSES

Pruning



PRUNING NEURONS



Machine Learning **Models**

Machine Learning **Runtimes**

Machine Learning **Hardware**

Model Compression Techniques

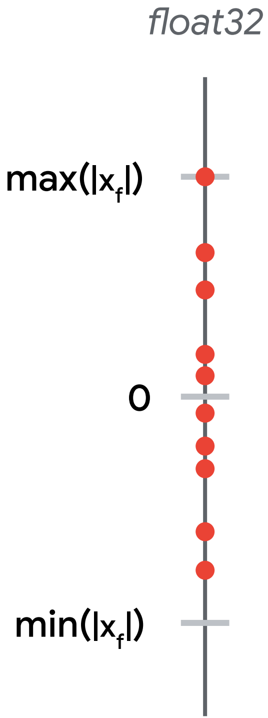
Pruning

Quantization

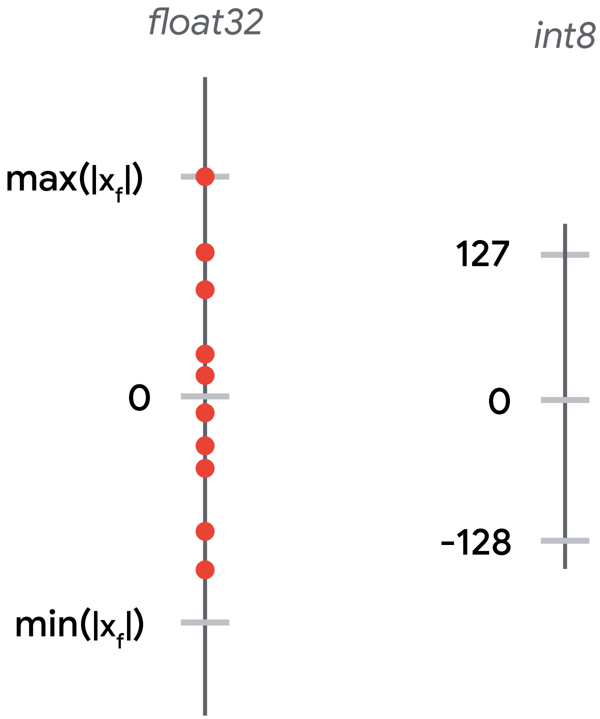
Knowledge Distillation

...

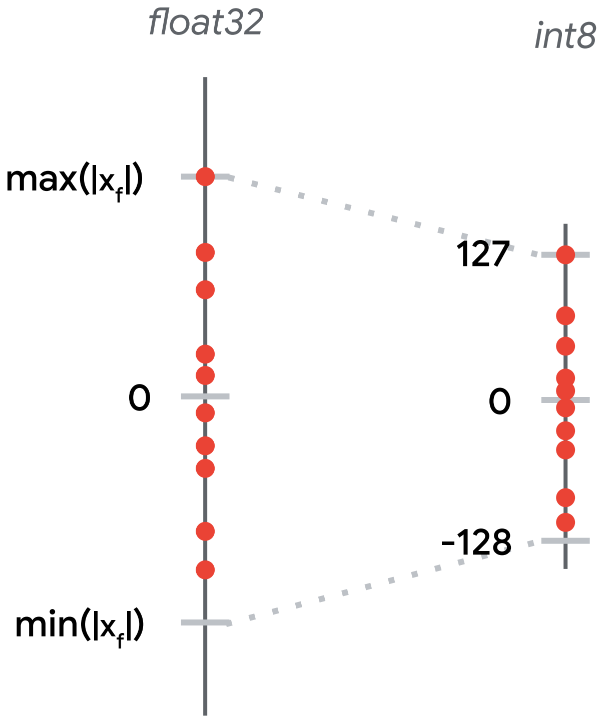
Quantization

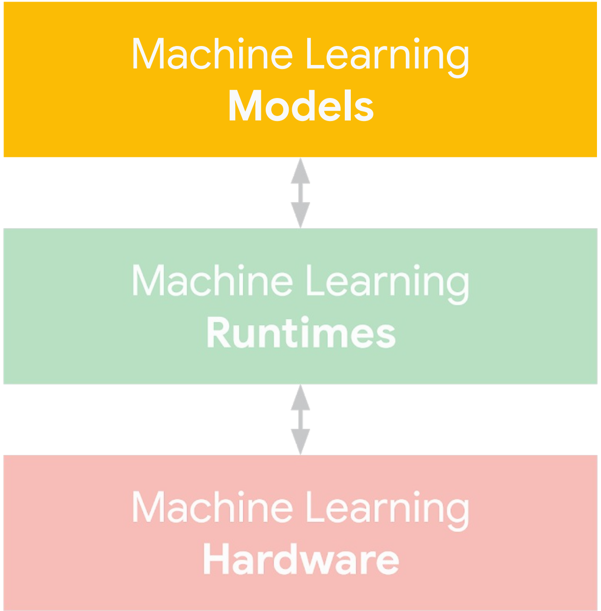


Quantization



Quantization





Machine Learning **Models**

Machine Learning **Runtimes**

Machine Learning **Hardware**

Model Compression Techniques

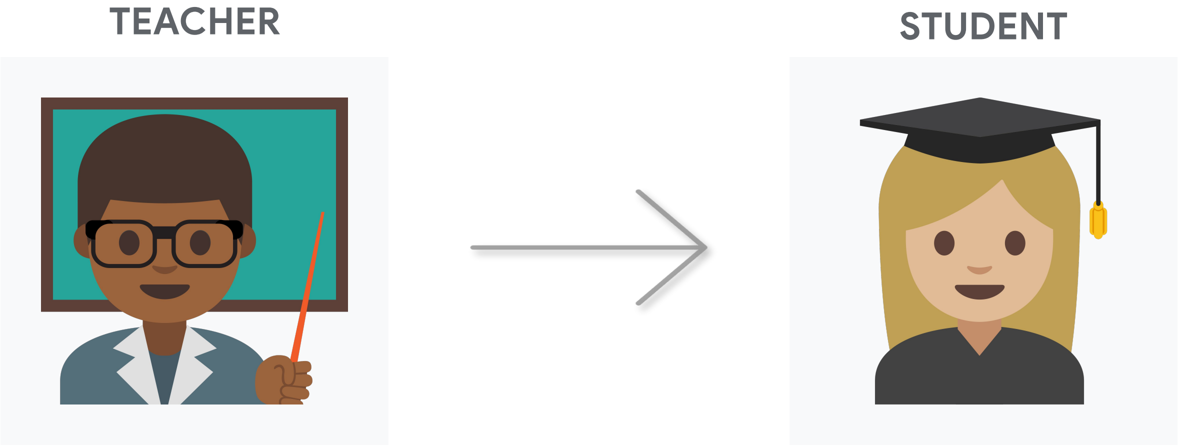
Pruning

Quantization

Knowledge Distillation

...

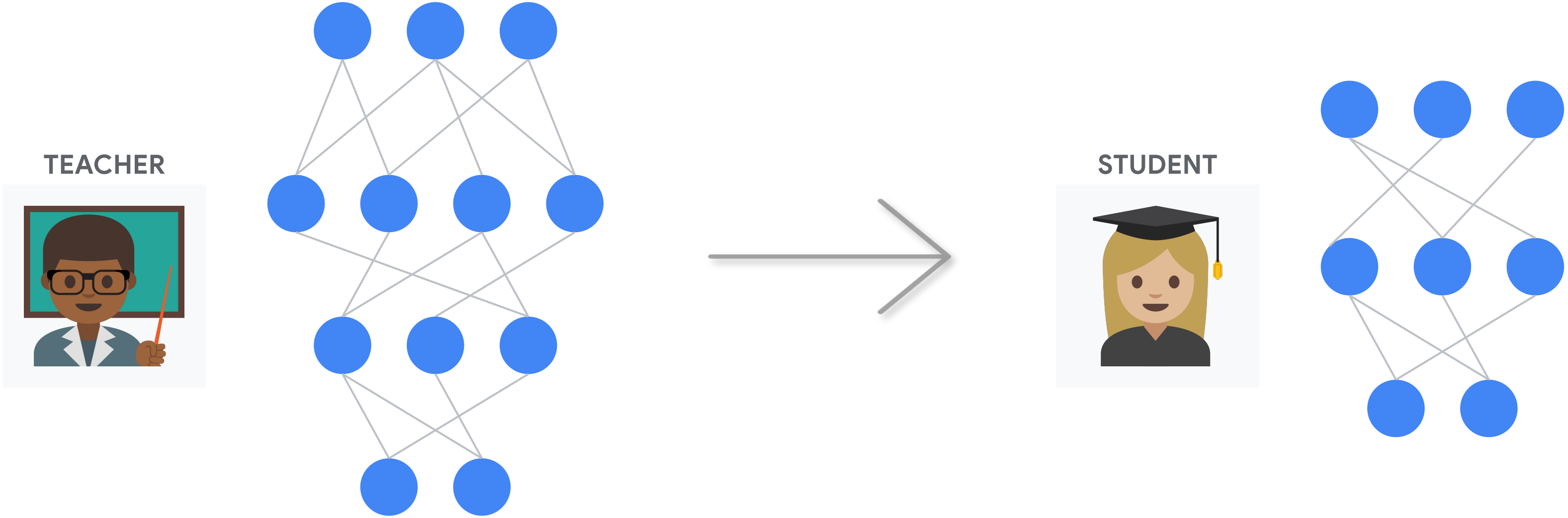
Knowledge Distillation



TEACHER

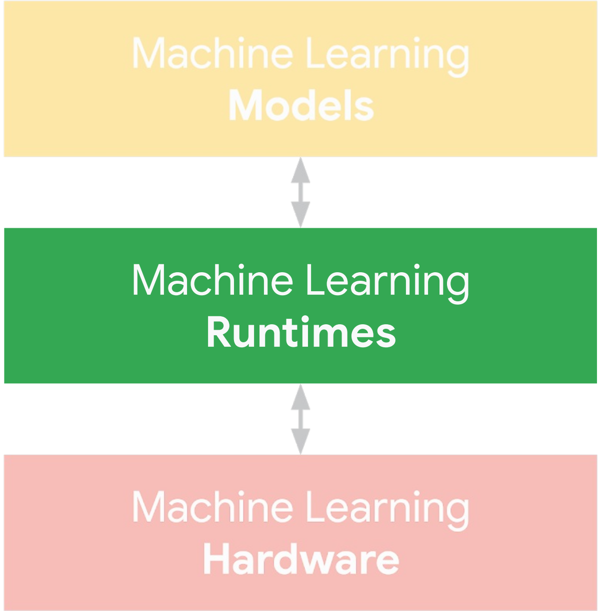
STUDENT

Knowledge Distillation



TEACHER

STUDENT



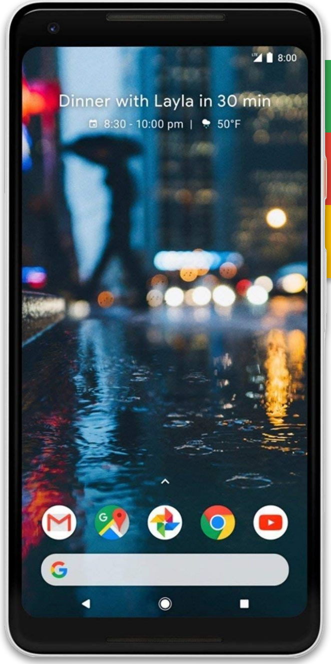
Machine Learning
Models

Machine Learning
Runtimes

Machine Learning
Hardware

TensorFlow

TensorFlow



Less memory

Less compute power

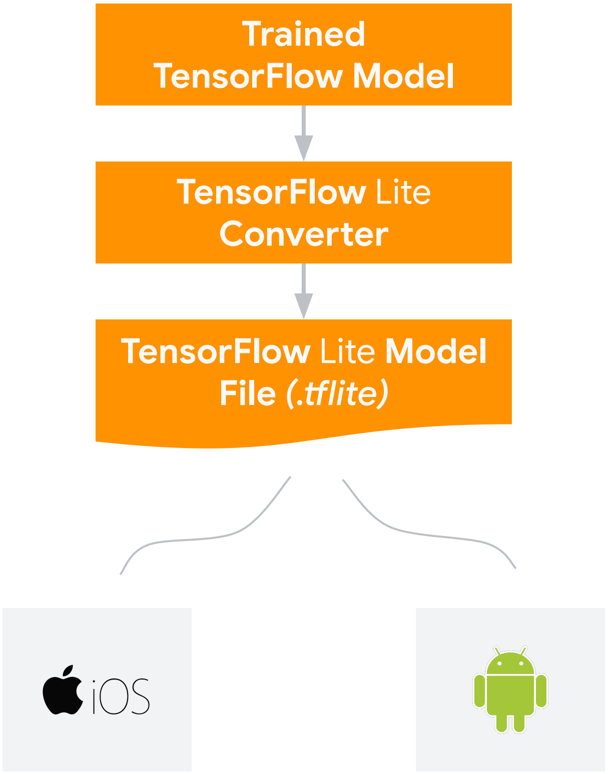
Only focused on *inference*

TensorFlow Lite

Key Differences

	TensorFlow	TensorFlow Lite
Topology	Variable	Fixed
Weights	Variable	Fixed
Binary Size	Unimportant	High Priority
Distributed Compute	Needed	Not Needed
Developer Background	ML Researcher	Application Developer

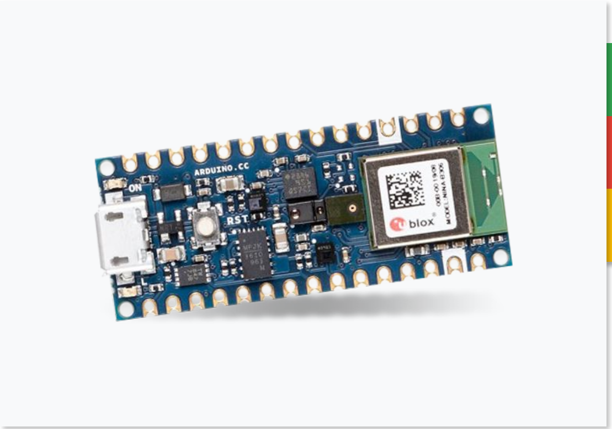
Architecture



Trained TensorFlow Model

TensorFlow Lite Converter

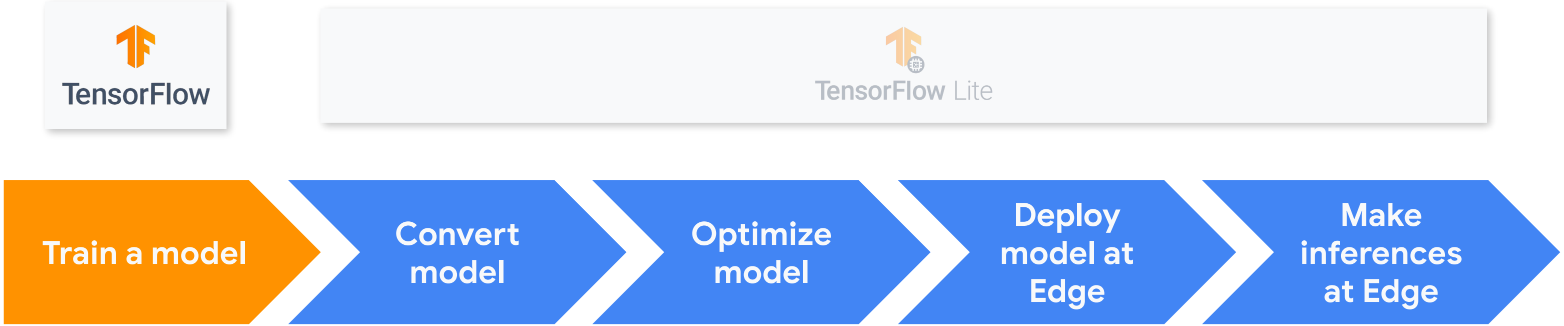
TensorFlow Lite Model File (*.tflite)*



Even less memory

Even less compute power

Also, only focused on *inference*



TensorFlow

TensorFlow Lite

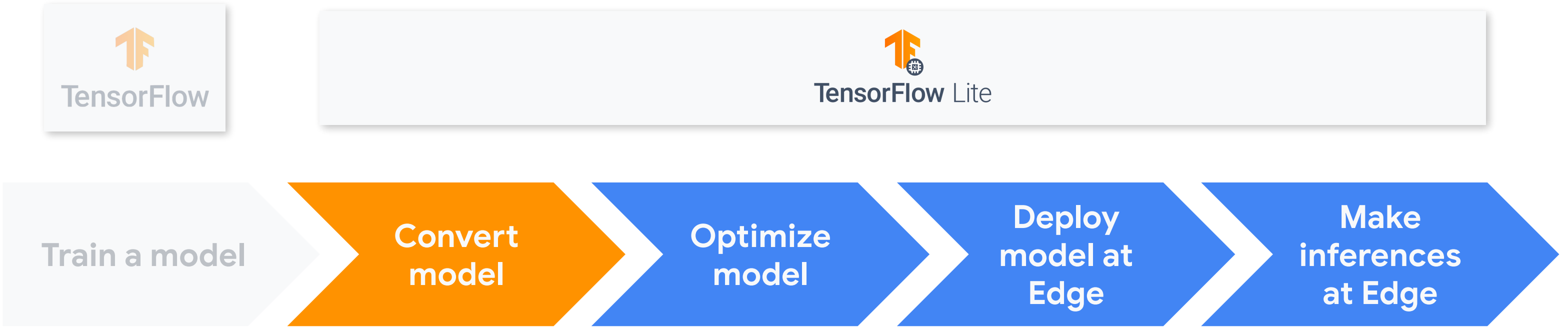
Train a model

Convert model

Optimize model

Deploy model at Edge

Make inferences at Edge



TensorFlow

TensorFlow Lite

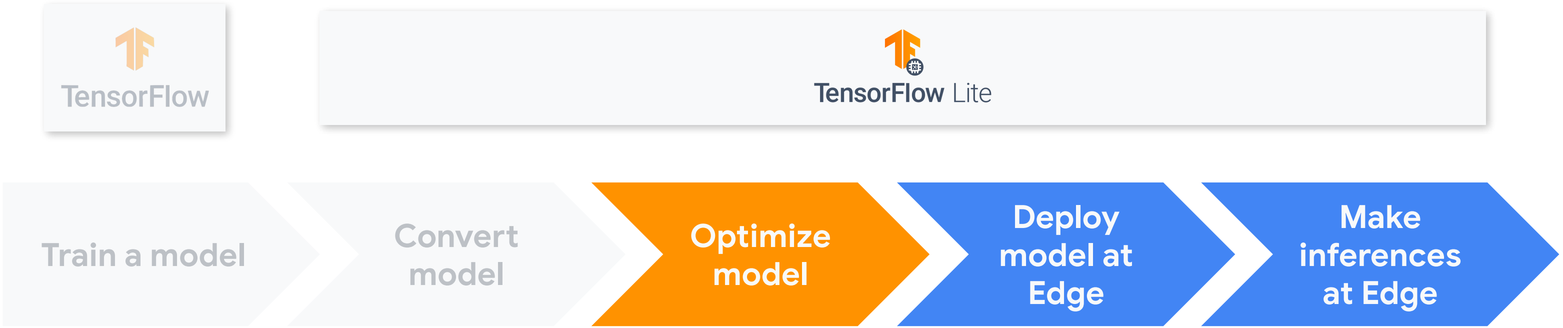
Train a model

Convert model

Optimize model

Deploy model at Edge

Make inferences at Edge



TensorFlow

TensorFlow Lite

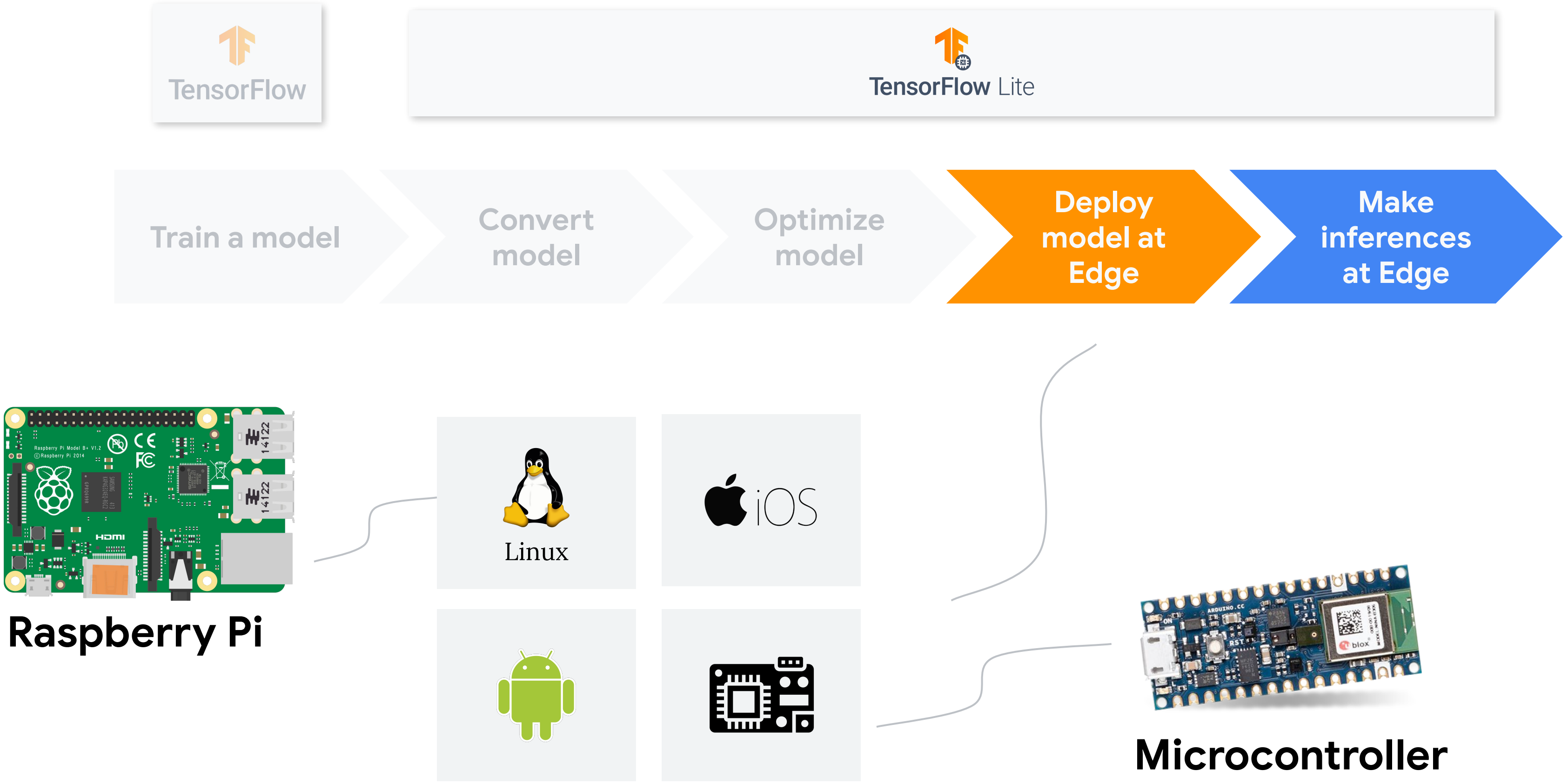
Train a model

Convert model

Optimize model

Deploy model at Edge

Make inferences at Edge



TensorFlow

TensorFlow Lite

Train a model

Convert model

Optimize model

Deploy model at Edge

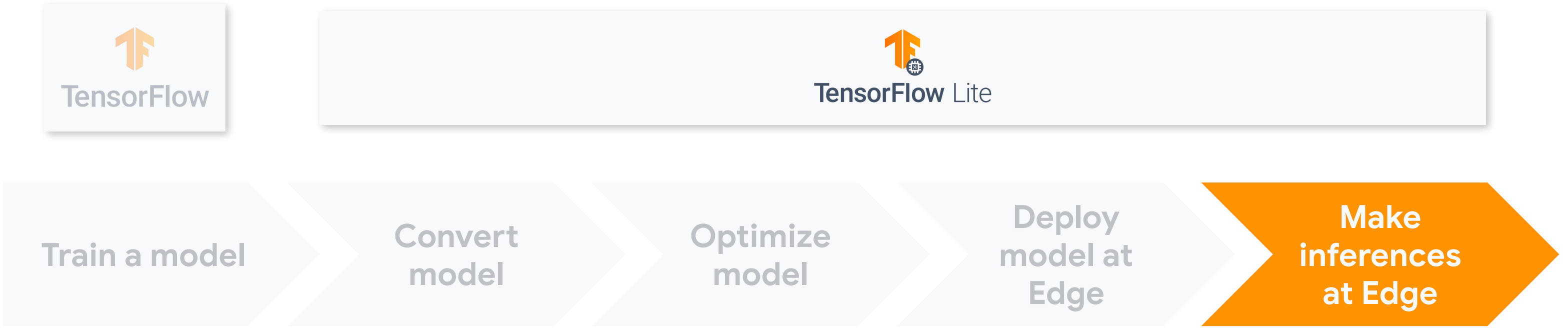
Make inferences at Edge

Raspberry Pi

Linux

iOS

Microcontroller



TensorFlow

TensorFlow Lite

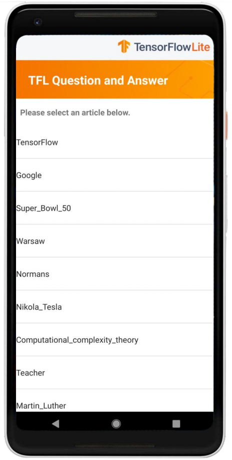
Train a model

Convert model

Optimize model

Deploy model at Edge

Make inferences at Edge



TensorFlowLite

TFL Question and Answer

Please select an article below.

TensorFlow

Google

Super_Bowl_50

Warsaw

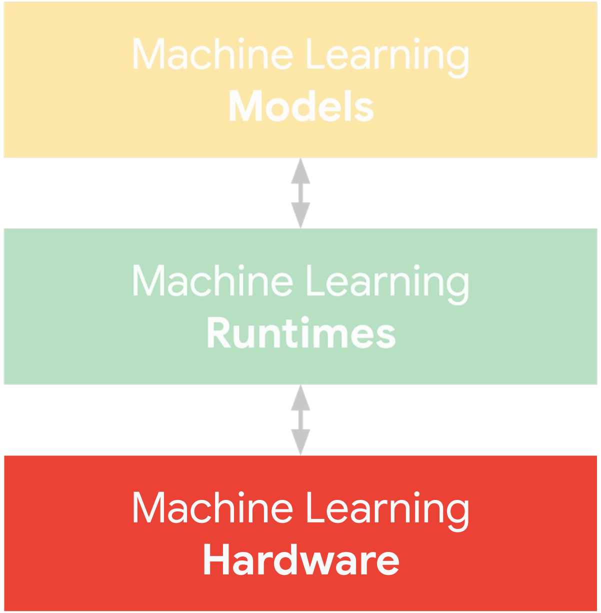
Normans

Nikola_Tesla

Computational_complexity_theory

Teacher

Martin_Luther



Machine Learning
Models

Machine Learning
Runtimes

Machine Learning
Hardware