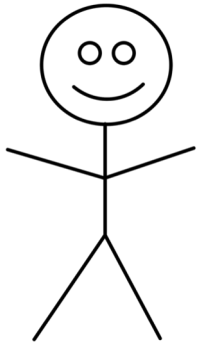
From Cloud to Edge: Advances in Mobile AI

Dr. Aakanksha Chowdhery
Software Engineer ML, Google Brain
chowdhery@google.com

machines in 2010

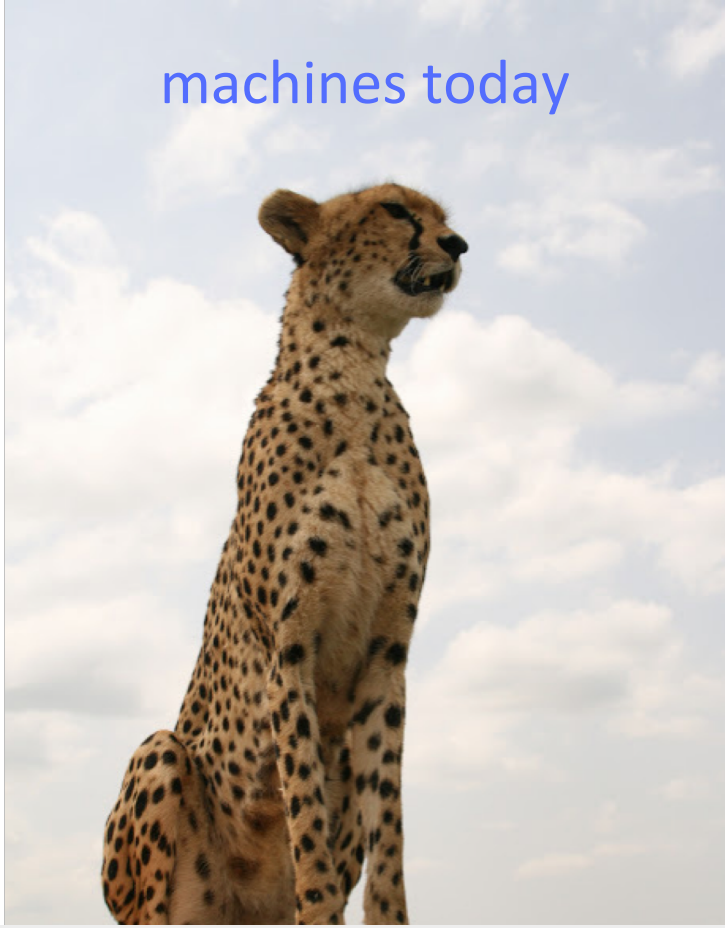
~25% errors

humans



5% errors

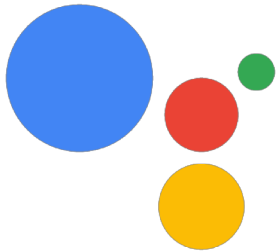
machines today



~3% errors

Deep Learning ML Use Cases

Computer Vision and Image Understanding



Speech and Language Understanding

Text Understanding

Gesture Recognition

Mobile AI is Important

Why ?

Sensitive Data

Latency

Bandwidth

Offline Performance

Cost

But hard ...

Tight memory constraints
Low energy usage to preserve batteries
Little compute power

Key Advances in Mobile AI

Lightweight ML platforms
Accelerators for ML computations
Model optimization
Privacy/Security

TensorFlow Lite

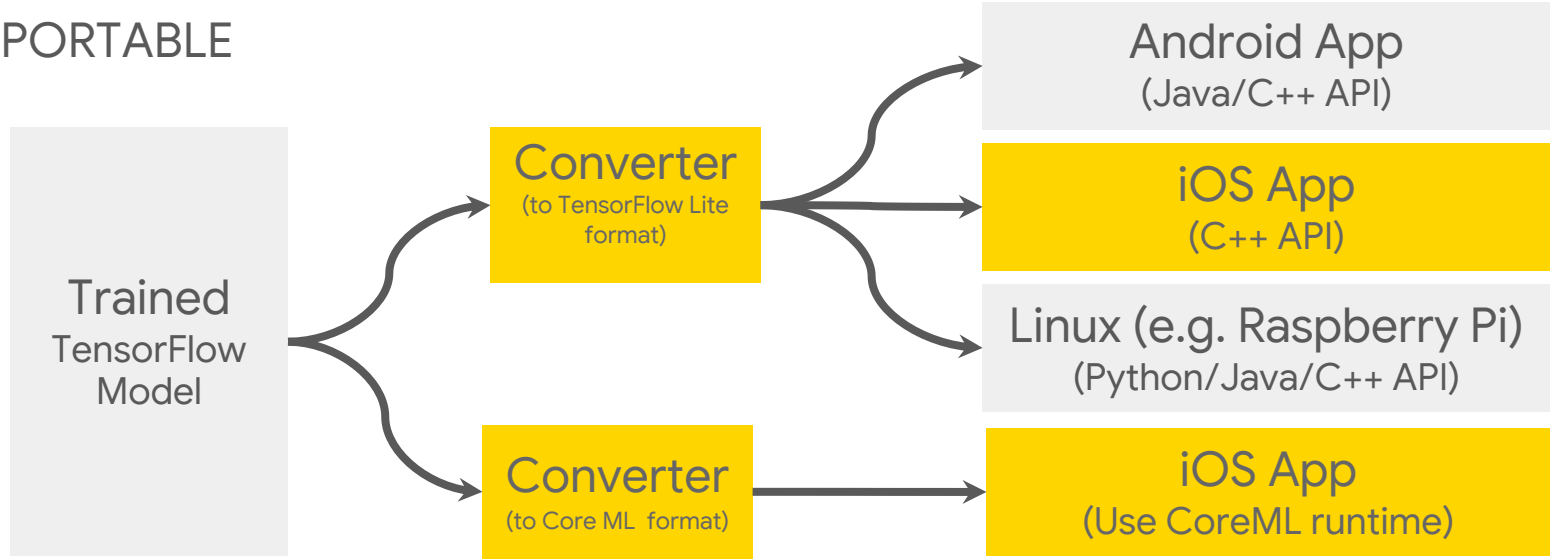
TensorFlow's solution for running **on-device machine learning** with **low latency** & a **small binary size** on **many platforms.**

TensorFlow Lite is **portable**

- iOS, Android
- Raspberry Pi or other Linux SOCs
- Micro-Controllers
- Works on PCs too

TensorFlow Lite

PORTABLE



TensorFlow Lite is **optimizable**

- Selective Registration
- CPU kernel fusion
- Optimized SIMD kernels

TensorFlow Lite

DESIGNED FOR SPEED

Converter
(to TensorFlow Lite format)

Interpreter
Core

Operation
Kernels

Hardware acceleration
delegates

TensorFlow
FALL SYMPOSIUM

TensorFlow Lite

DESIGNED FOR SPEED

Specially optimized kernels optimized for NEON on ARM

Converter
(to TensorFlow Lite format)

Interpreter
Core

Operation
Kernels

Hardware acceleration
delegates

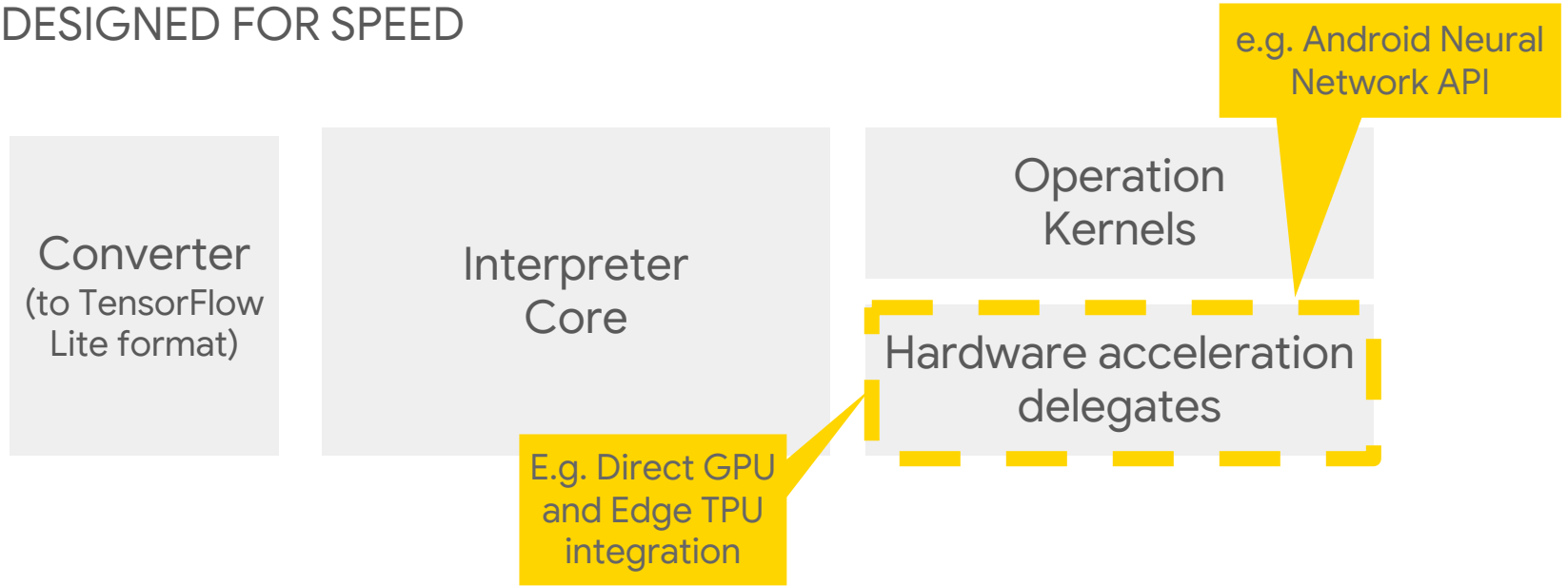
TensorFlow
FALL SYMPOSIUM

TensorFlow Lite connects with **accelerators**

- GPUs
- Edge-TPUs!
- NNAPI-supported accelerators

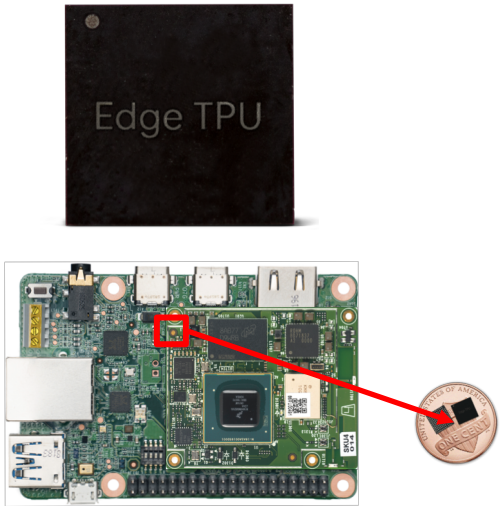
TensorFlow Lite

DESIGNED FOR SPEED



TensorFlow Lite Edge TPU™

DESIGNED FOR SPEED



- Google's AI ecosystem for the Edge
- High performance in small footprint
- High performance/Watt
- Enable broad deployment of high-quality AI
- Available in a Edge TPU Dev kit in Q4 2018

Model Optimization Toolkit

Model Optimization

SMALLER, FASTER MODELS

What it means?

- Alter the **original graph topology** into a more efficient one with reduced parameters and/or faster to execute (i.e. distillation)
  - Reduce parameter precision (e.g. network weights, pruning).
  - Execute operations between the static parameters and dynamic inputs/activations.

Why optimize models?

SMALLER, FASTER MODELS

BALANCING ACT

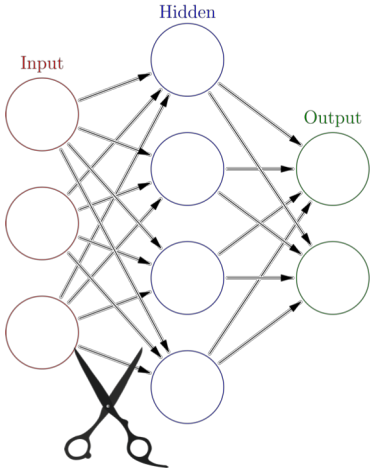
Memory vs Computation vs Accuracy

Model Optimization Toolkit

SMALLER, FASTER MODELS

Quantization

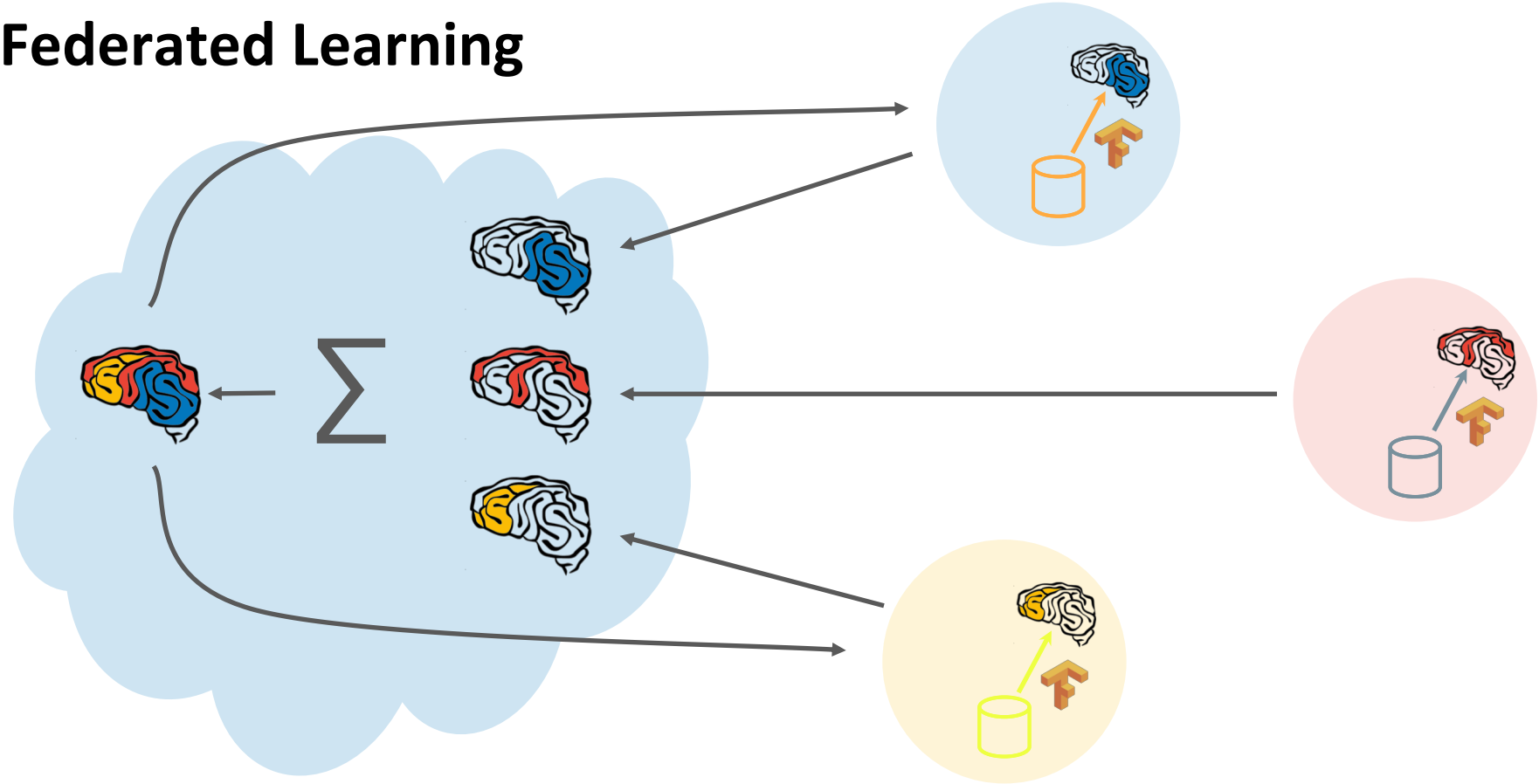
Pruning / sparsity



0	0	7	4
9	6	0	0
0	0	1	3
2	3	0	0

Privacy Challenges

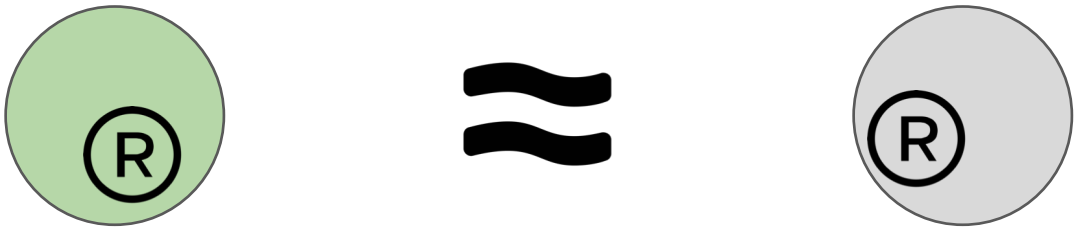
Training with Sensitive Data: Federated Learning



Security & IP Protection: On-device model use risks

1. Unauthorized model use and duplication
2. Reversing of model data

Model watermarking and fingerprinting



Original model with special pattern embedded

Stolen model with modifications

Conclusions

Key Advances in Mobile AI

Lightweight ML platforms
Accelerators for NN computations
Model optimization
Privacy/Security

Get it. Try it.

Code: github.com/tensorflow/tensorflow

Docs: tensorflow.org/lite/

Discuss: tflite@tensorflow.org mailing list