Experts

Edge Computing using TensorFlow Lite

Bhavesh Bhatt

@_bhaveshbhatt

Google Developers

echo $(whoami)

Experts

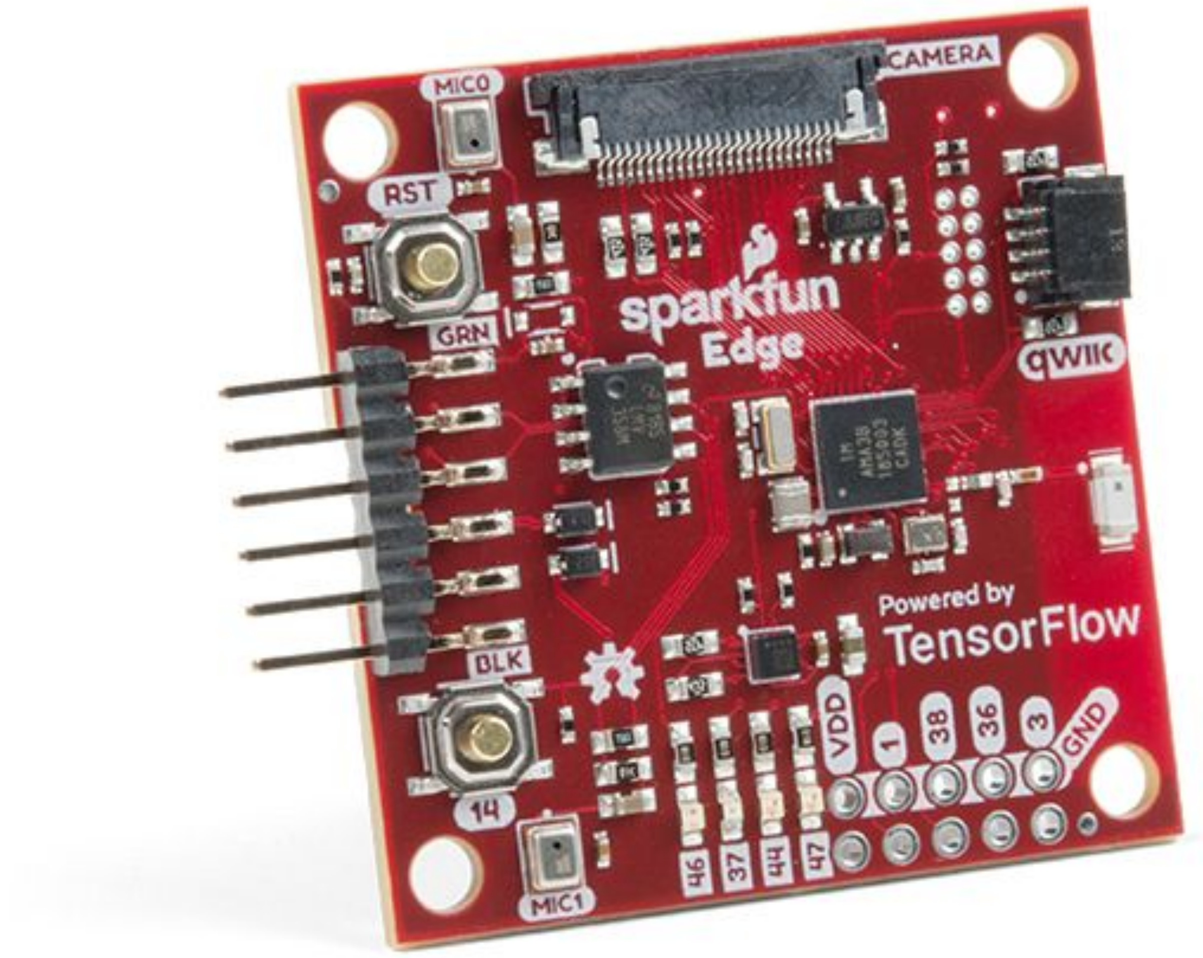
AIM

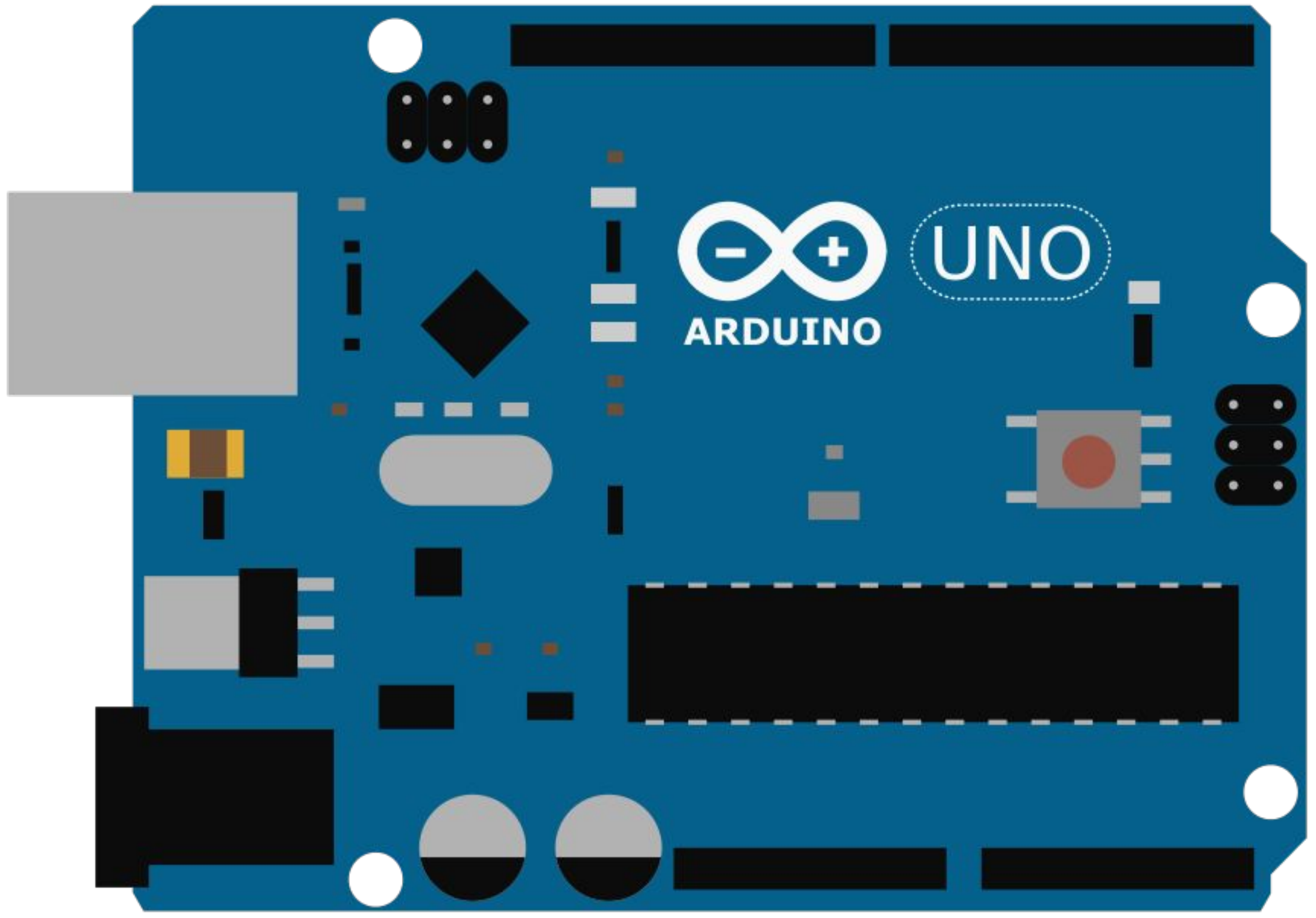
40 under forty

DATA SCIENTISTS

2020

7:00
October 4
TUESDAY, 2016


MIC0
CAMERA
RST
sparkfun
Edge
GRN
qwiic
Powered by
TensorFlow
BLK
VDD
1
38
36
3
GND
14
MIC1
46
37
44
47


UNO
ARDUINO

Experts

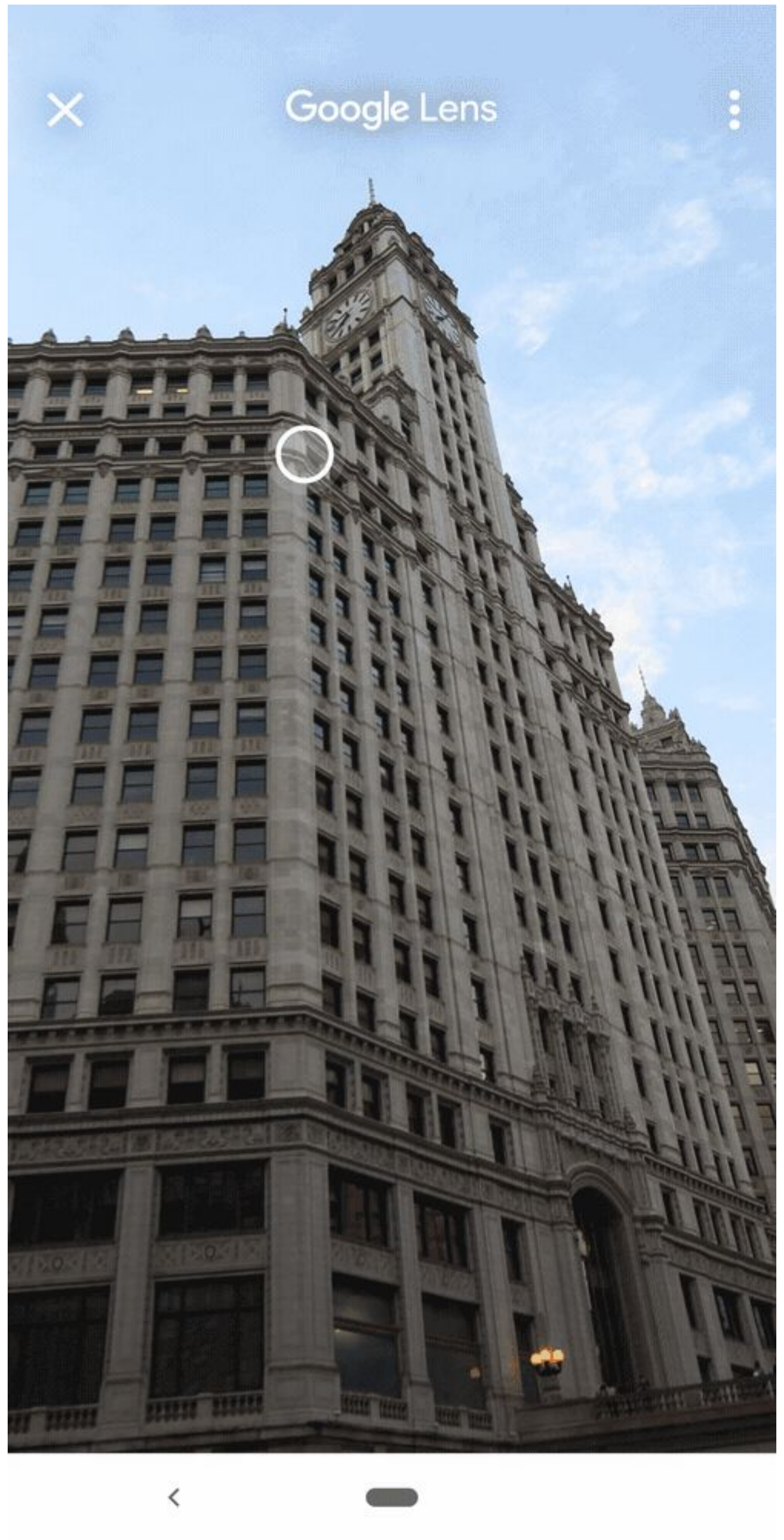


Image Courtesy : https://lens.google.com/

Edge ML Explosion

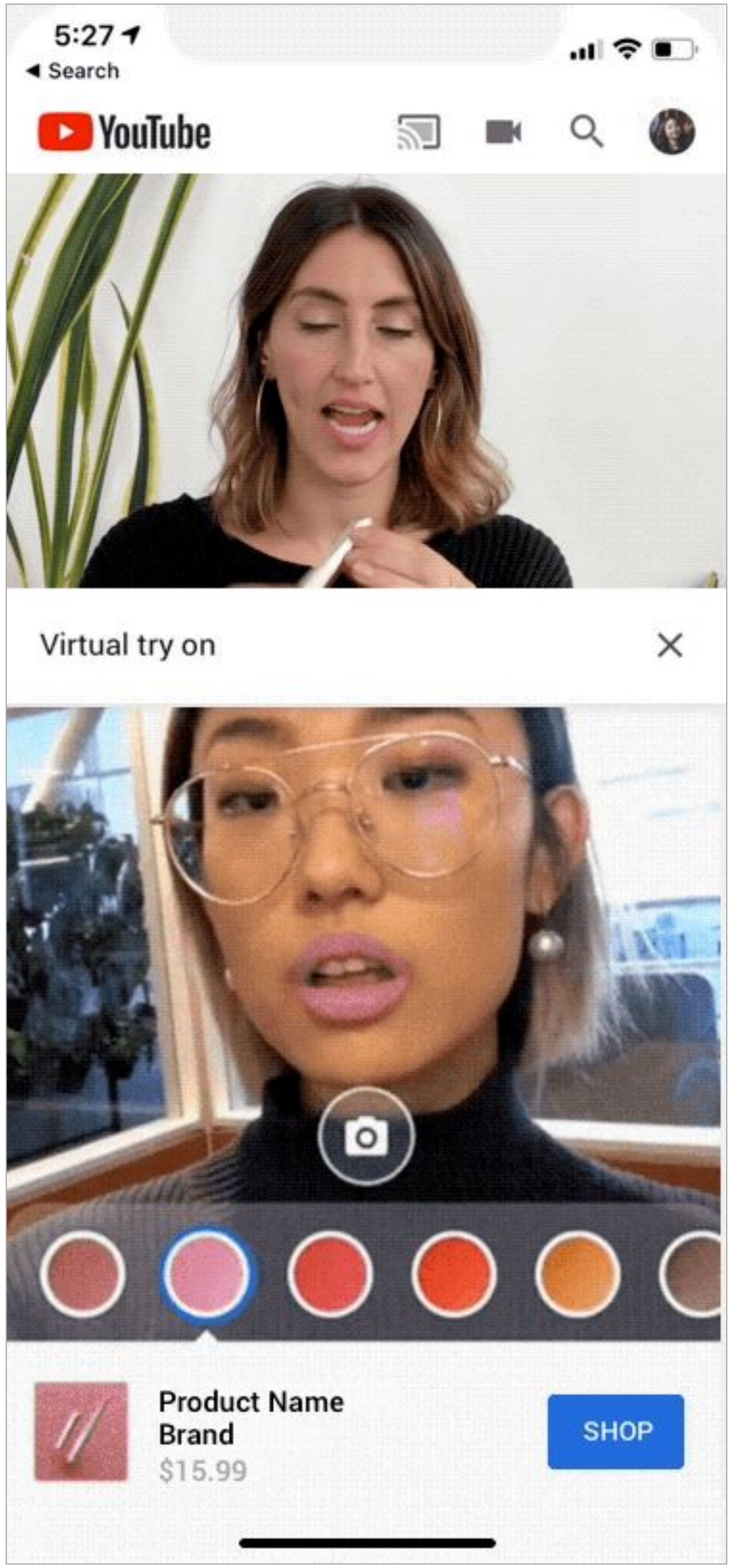
- Lower latency & close knit interactions

Edge ML Explosion

- Lower latency & close knit interactions
- Network Connectivity

Edge ML Explosion

- Lower latency & close knit interactions
- Network Connectivity
- Privacy preserving



Product Name
Brand
$15.99

牛奶

Challenges

- Limited compute power
- Limited memory
- Battery consumption
- App size

TensorFlow Lite is a production ready, cross-platform framework for deploying ML on mobile devices and embedded systems

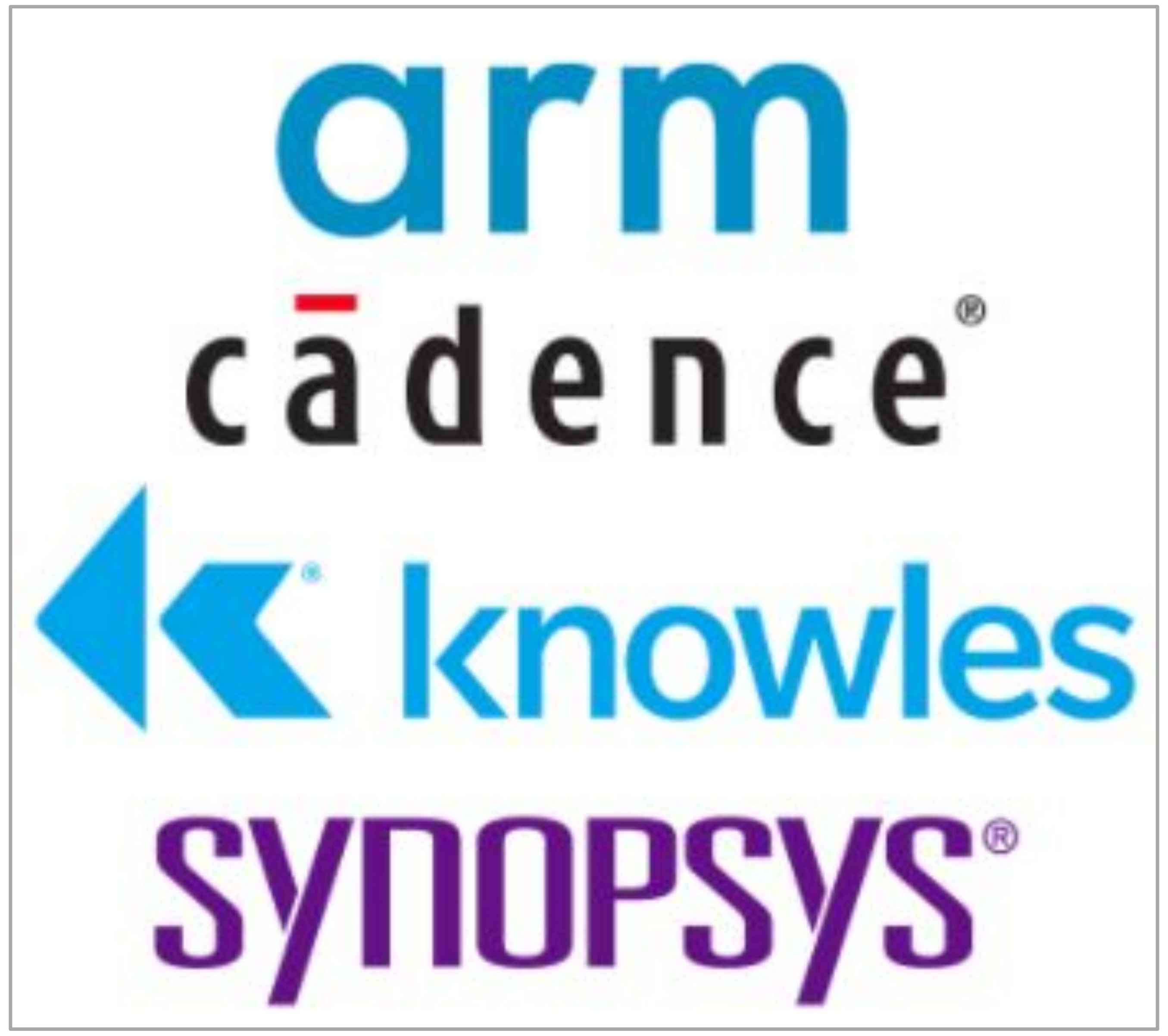
android

iOS

arm

Folks, hacking with embedded stuff & microcontroller stuff

- Build TensorFlow Lite for ARM64 boards
- Build TensorFlow Lite for Raspberry Pi



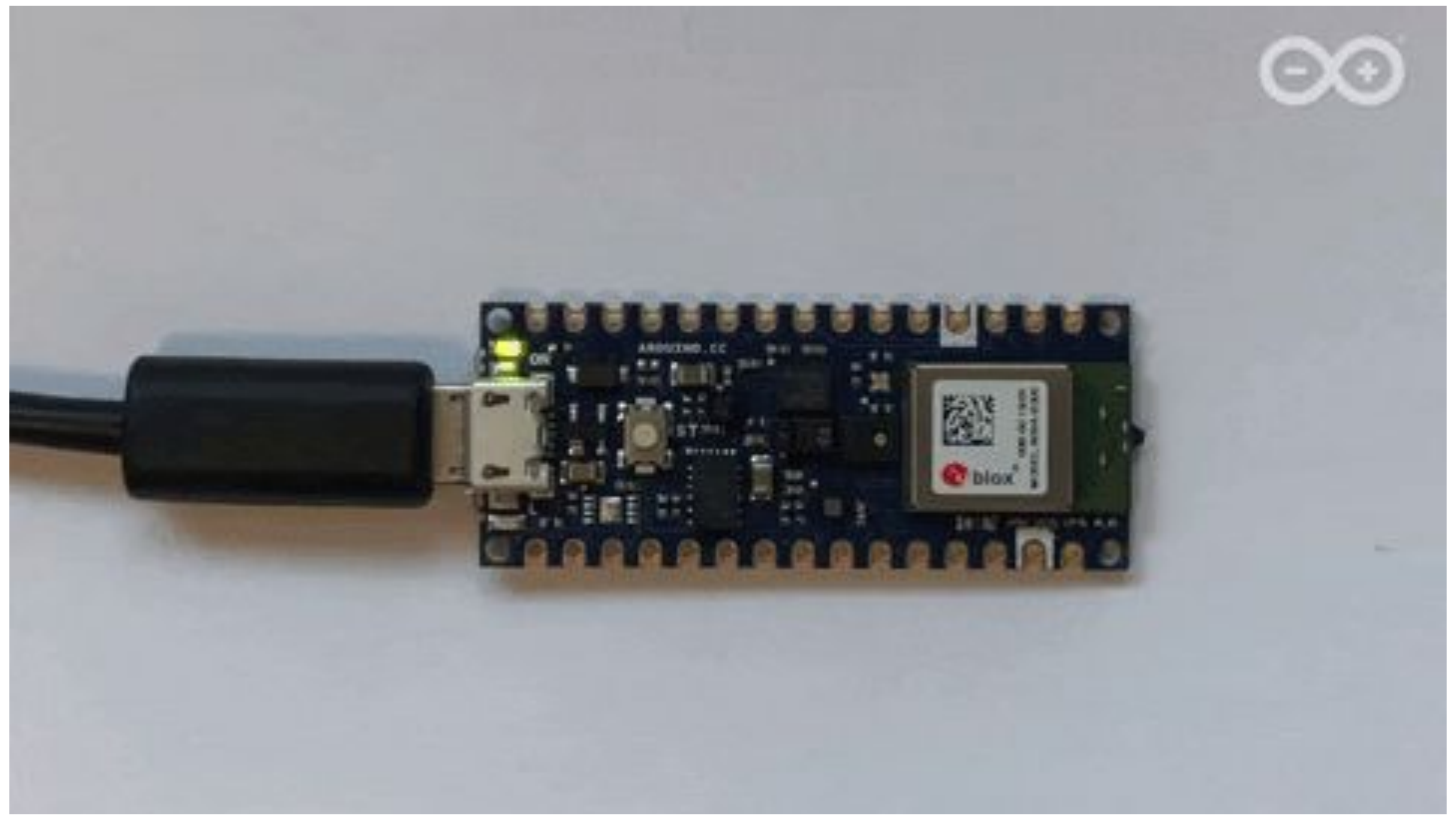
Folks, hacking with embedded stuff & microcontroller stuff

- Tremendous speed up with Edge TPU compatible TF Lite models

Check out here: Edge TPU performance benchmarks | Coral

Folks, hacking with embedded stuff & microcontroller stuff

- **Launch of official Arduino library** - run example code directly from desktop and web IDEs onto Arduino hardware
- **Speech detection in 5 minutes** - open source models available to get started quickly on Arduino

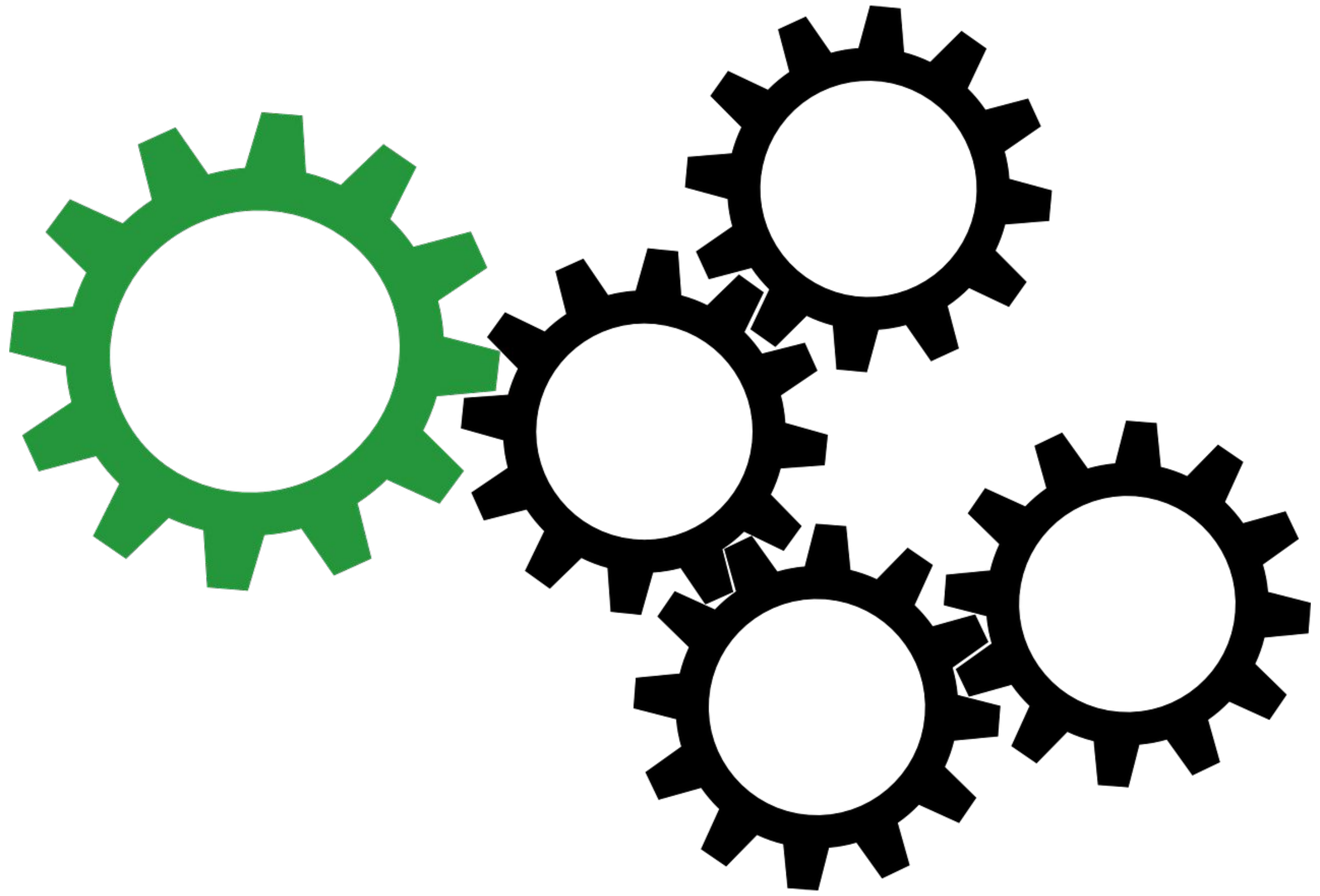


ARDUINO®

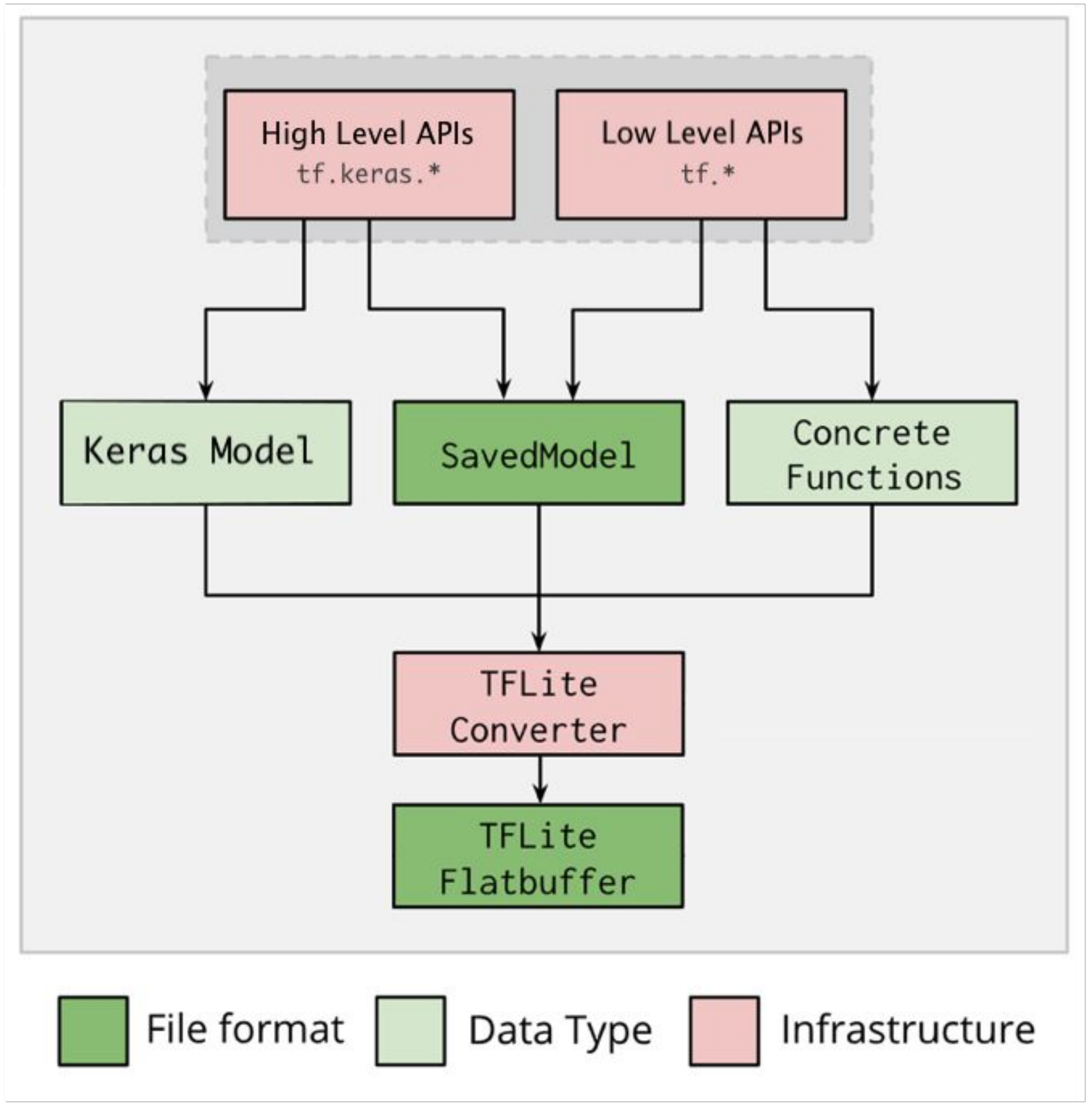
https://www.tensorflow.org/lite/microcontrollers

Workflow

1. Train a TensorFlow model
2. Convert to TensorFlow Lite format
3. Deploy and run on edge device



Workflow



High Level APIs
tf.keras.*

Low Level APIs
tf.*

Keras Model

SavedModel

Concrete Functions

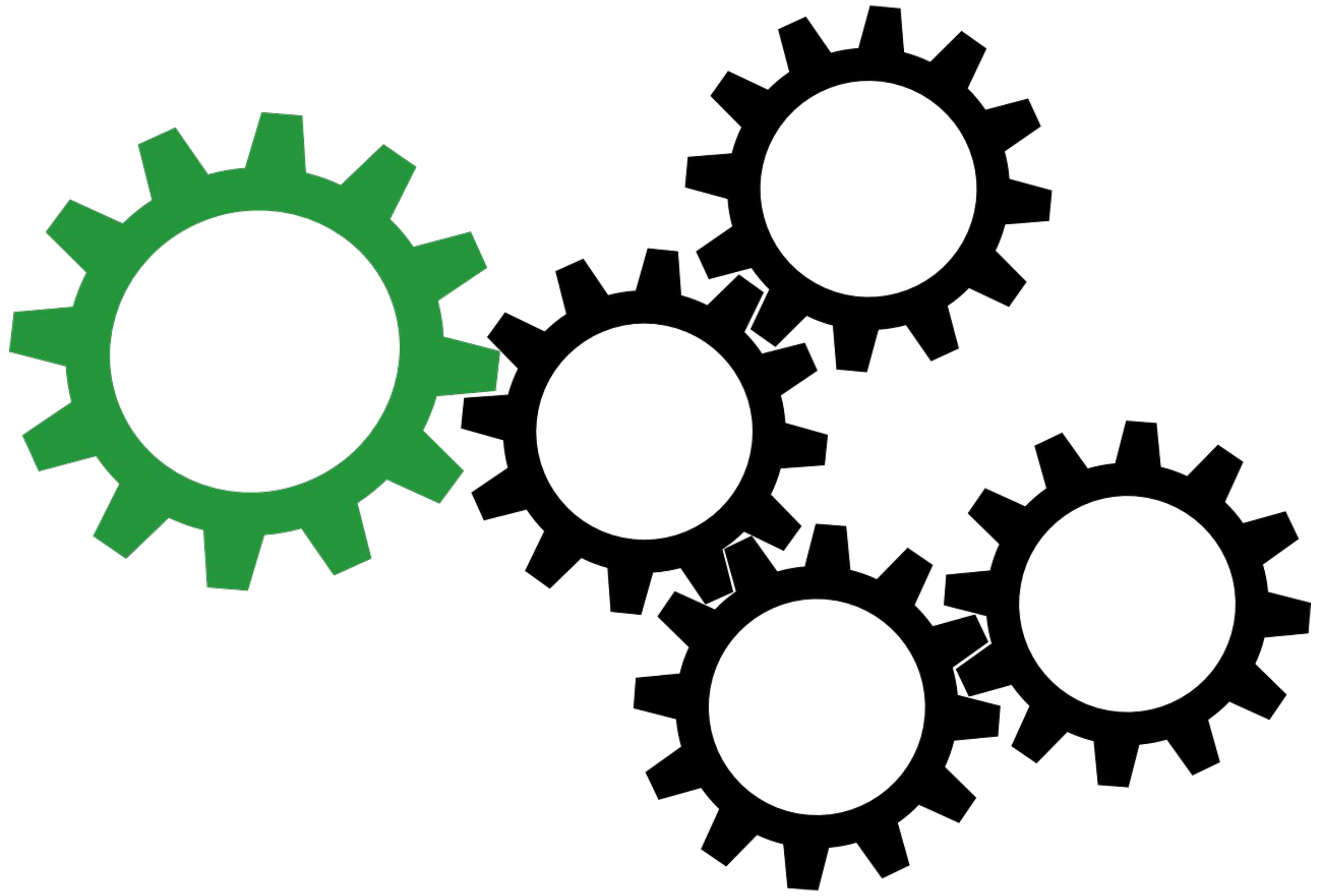
TFLite Converter

TFLite Flatbuffer

🟩 File format □ Data Type 🟥 Infrastructure

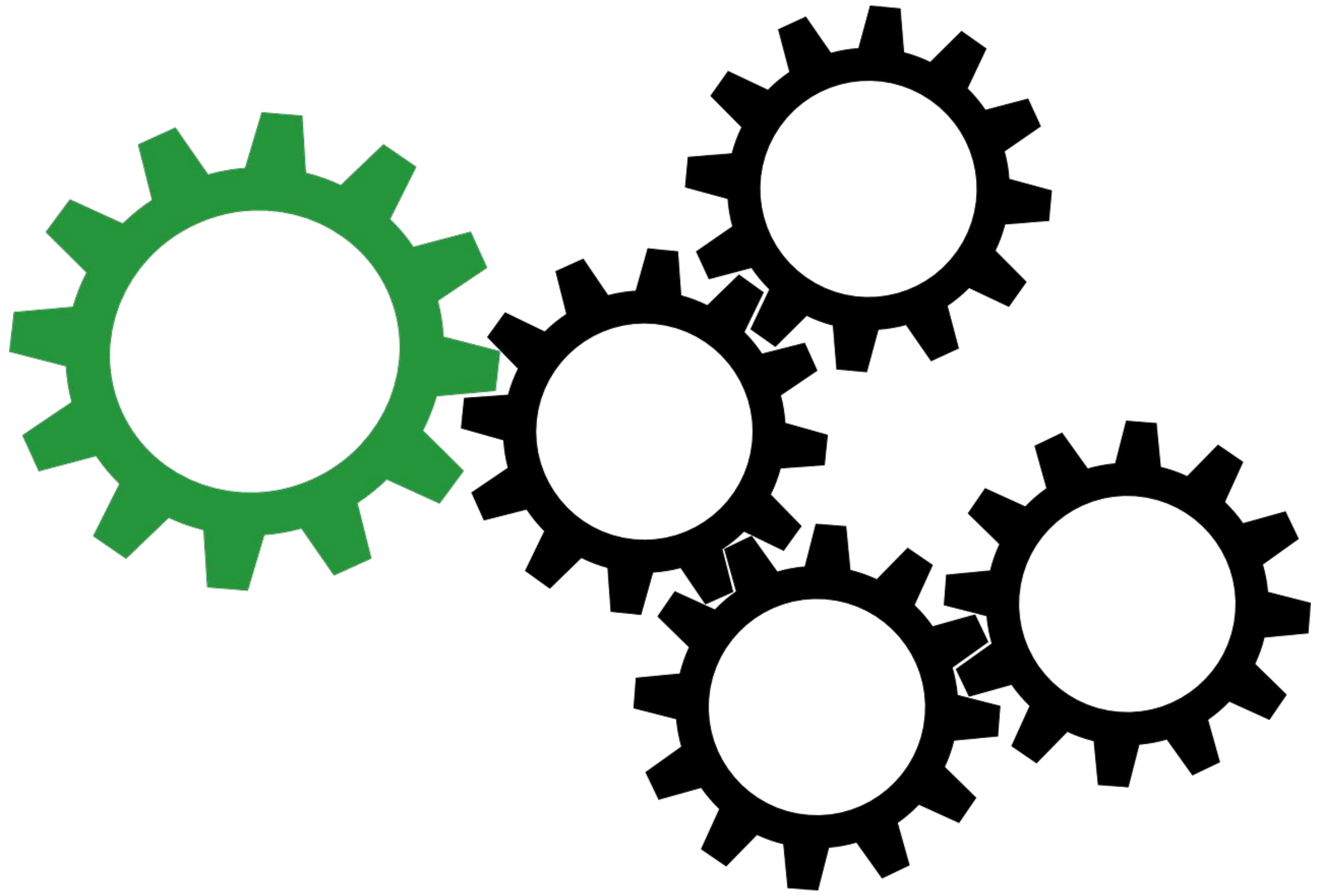
Workflow

FlatBuffers is an efficient cross platform serialization library for C++, C#, C, Go, Java, Kotlin, JavaScript, Lobster, Lua, TypeScript, PHP, Python, Rust and Swift. It was originally created at Google for game development and other performance-critical applications.



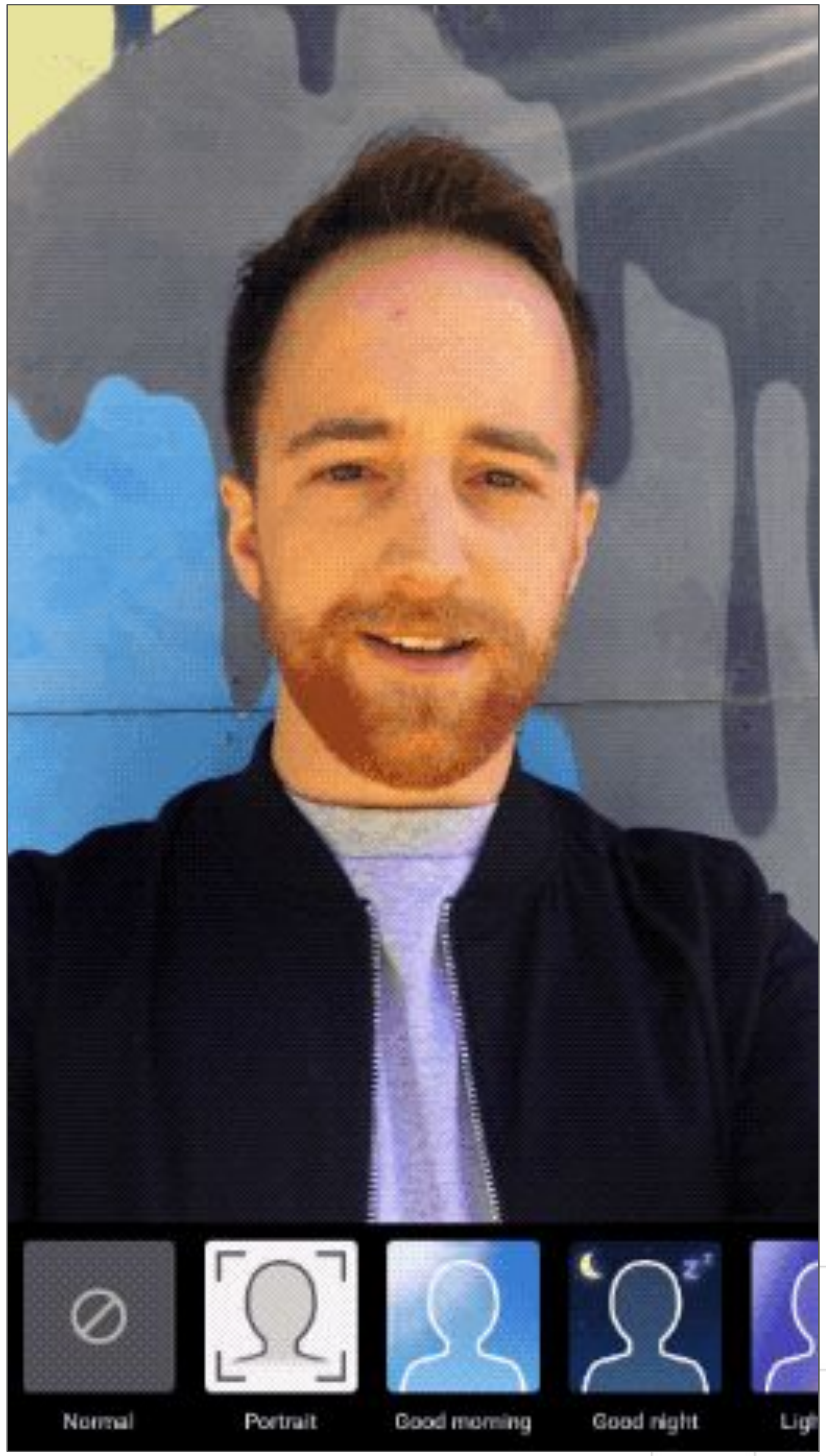
Why not use Protocol Buffers?

Protocol Buffers is indeed relatively similar to FlatBuffers, with the primary difference being that FlatBuffers does not need a parsing/ unpacking step to a secondary representation before you can access data, often coupled with per-object memory allocation. The code is an order of magnitude bigger, too.

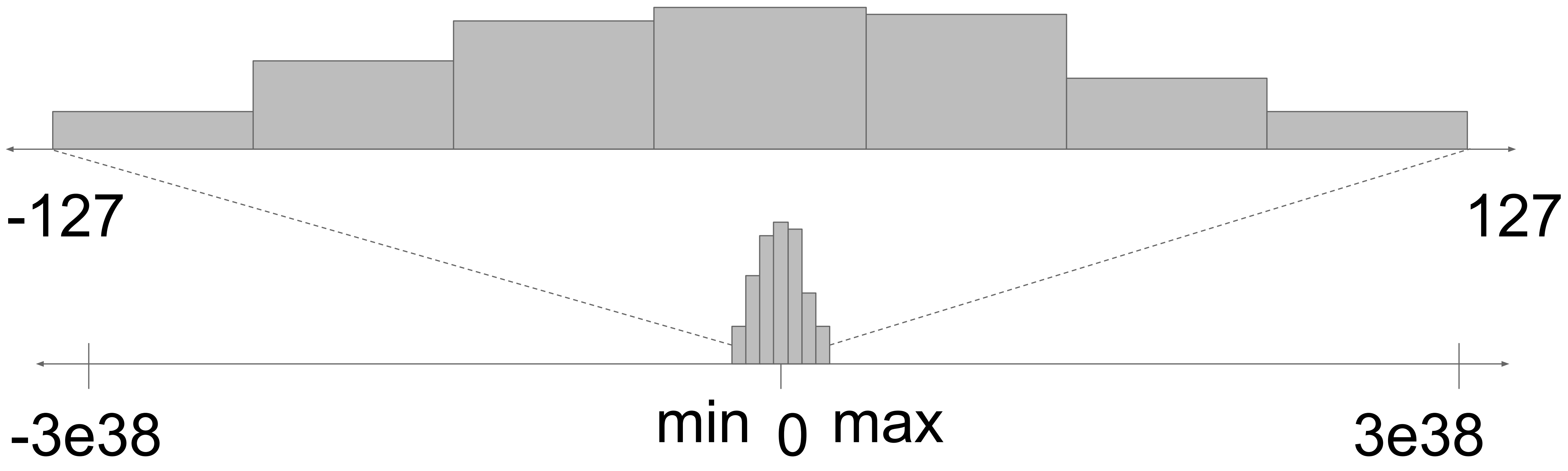


TensorFlow Lite

Lets Go Forward & witness the magic of TFLite

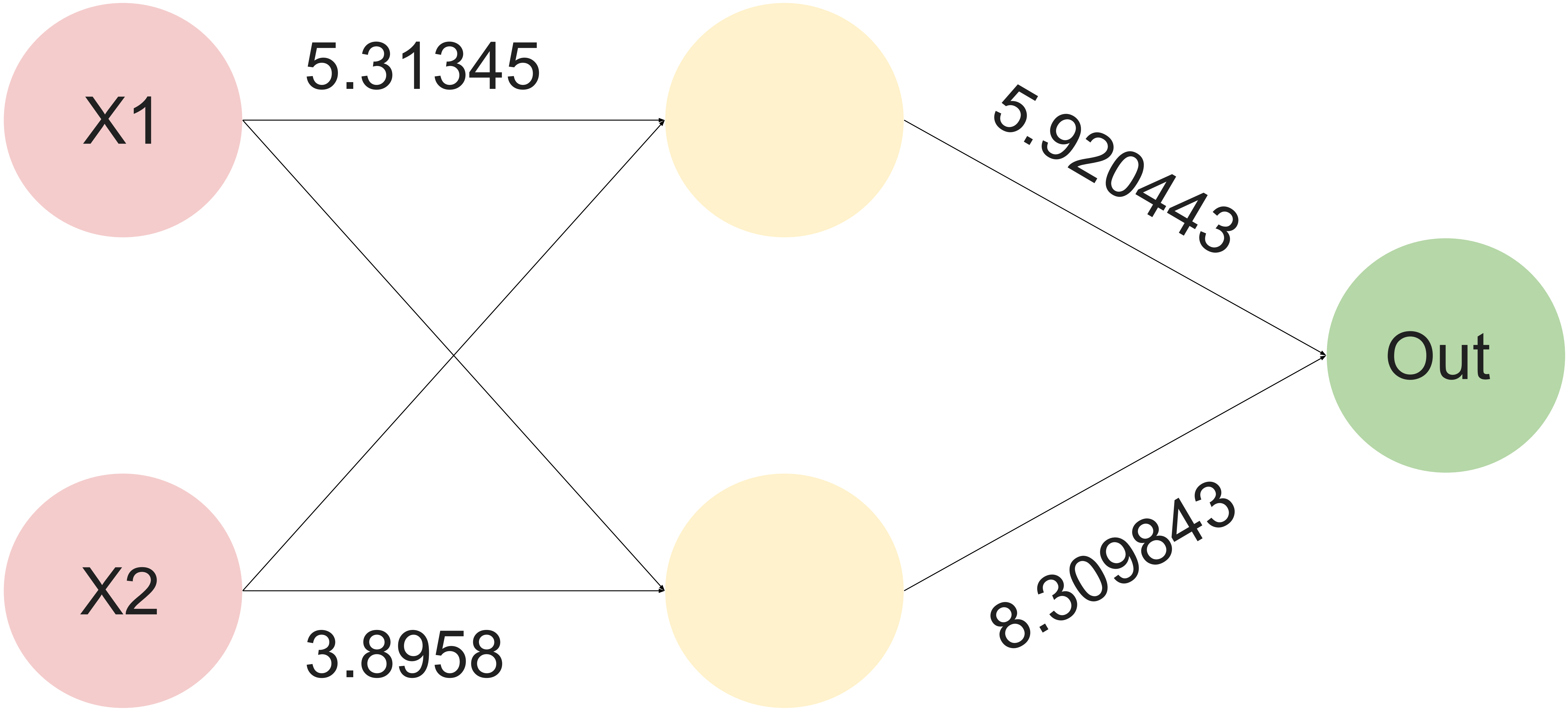


Quantization

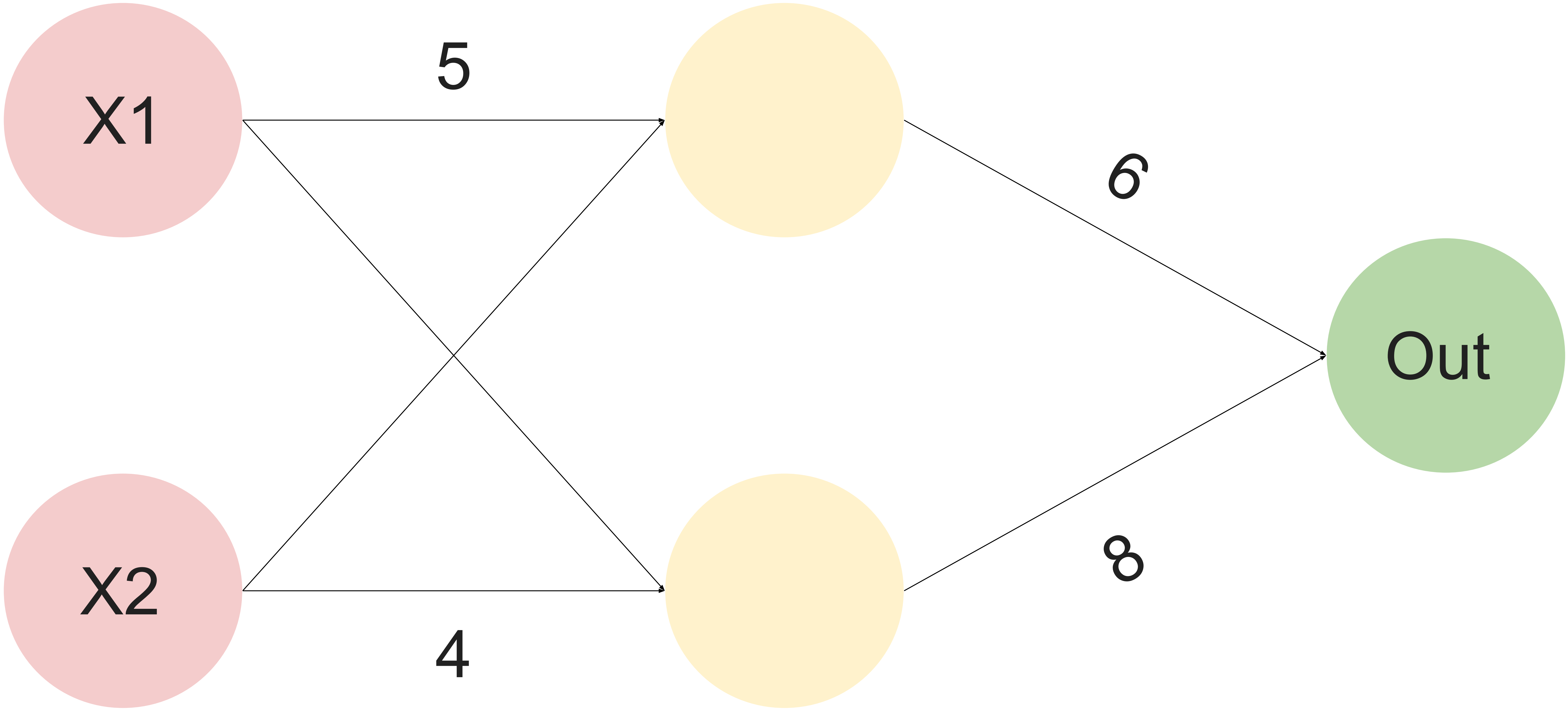


Reduce number of bits for model weights and activations

Quantization



Quantization



Q&A

THANK YOU FOR ATTENTION
Q&A SESSION
Spongebob Squarepants vector trace
memegenerator.net

Scan to access Slides & Code

Experts

Thank You!

Bhavesh Bhatt

@_bhaveshbhatt