Deep learning **on microcontrollers**

Jan Jongboom

IETF 101, London

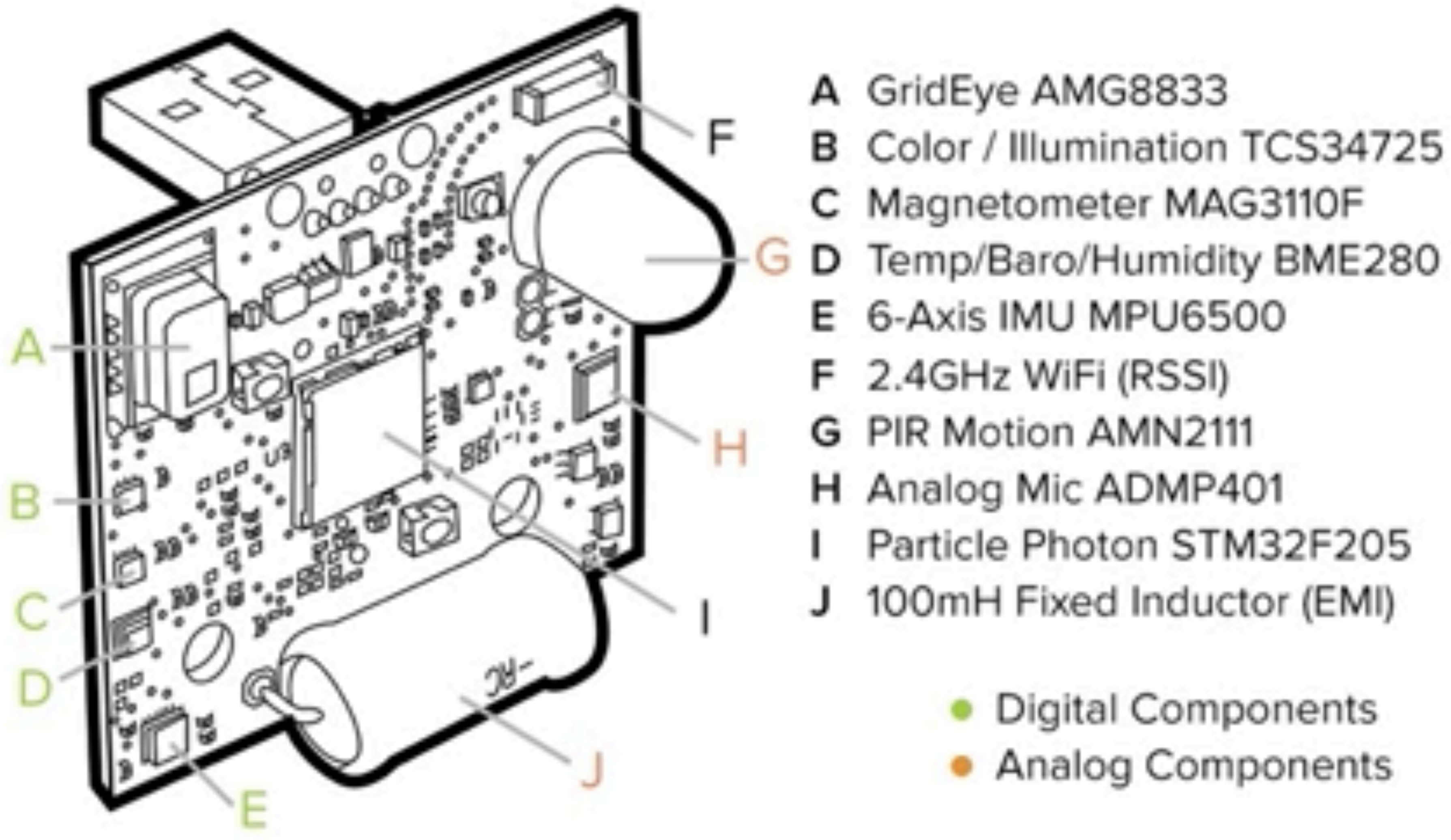
22 March 2018

https://twitter.com/billyd/status/701804370560729089


Data Science 2017 (DSA 2017) - Summer School & Workshop
17th - 21st July, 2017 at NM-AIST, Arusha
PULSE LAB KAMPALA
The University Of Sheffield.
ARMmbed
twas
MOUSER ELECTRONICS

Machine learning

Why machine learning on the edge?

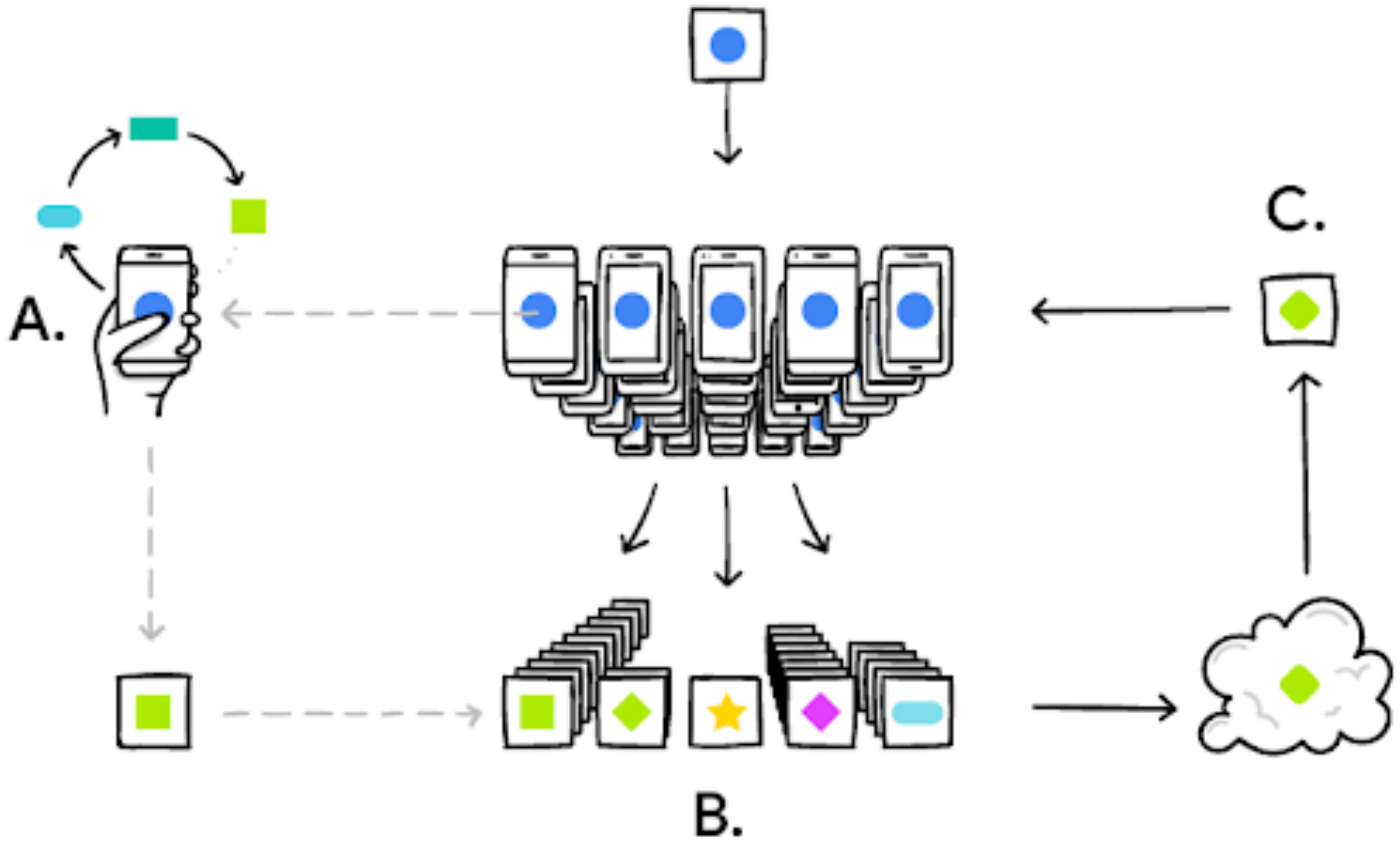


Sensor fusion

http://www.gierad.com/projects/supersensor/

arm

Why machine learning on the edge?



Federated learning

https://research.googleblog.com/2017/04/federated-learning-collaborative.html

Why machine learning on the edge?

LPWANs

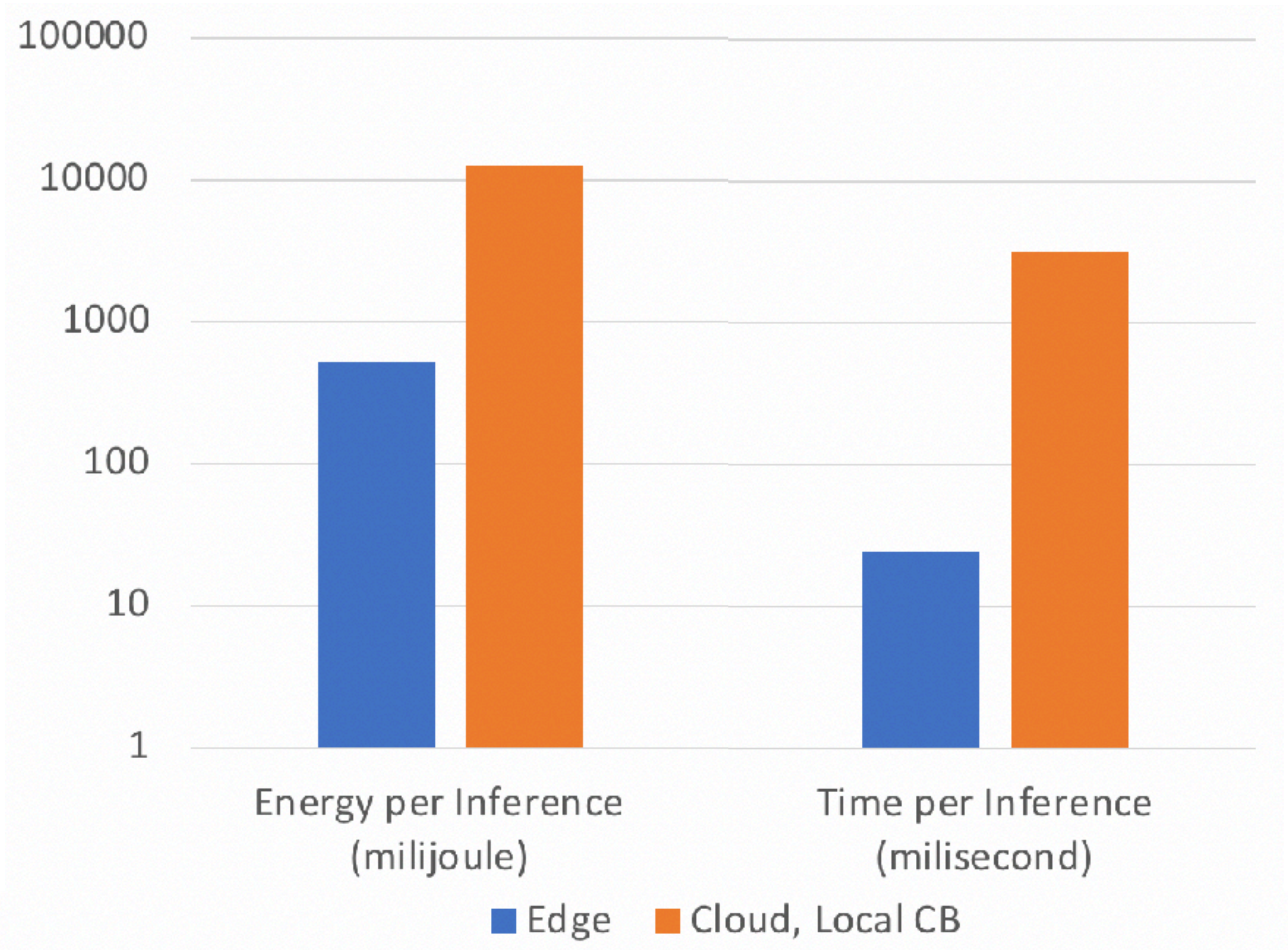
Why machine learning on the edge?

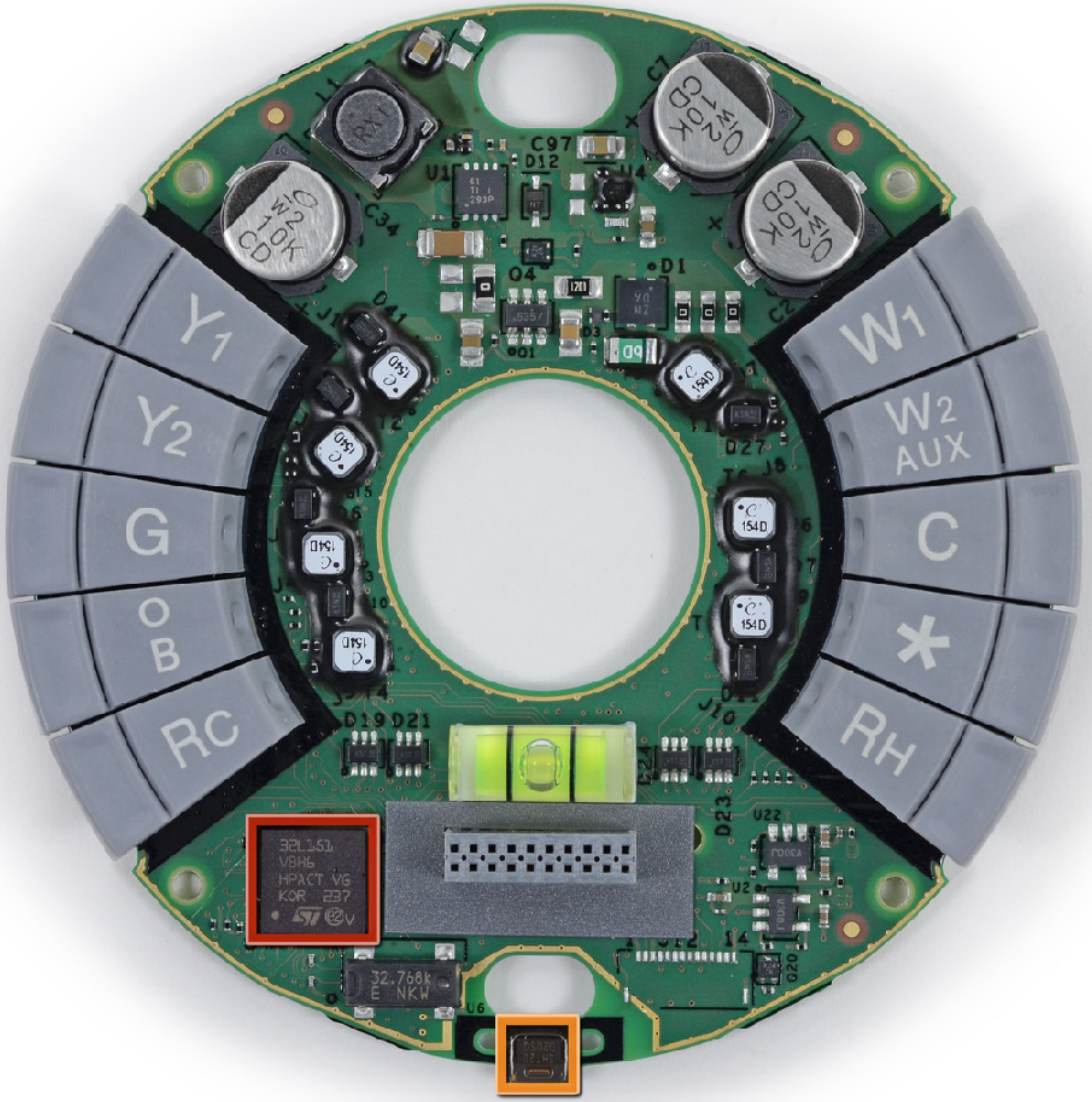
Offline self-contained systems

https://os.mbed.com/blog/entry/streaming-data-cows-dsa2017/

arm

Edge vs. Cloud





Microcontrollers

Small (1cm²)

Cheap (~1$)

Efficient (standby: 0.3 µA)

Downsides

Slow (max. 100 MHz)

Limited memory (max. 256K RAM)

uTensor

Machine learning for microcontrollers

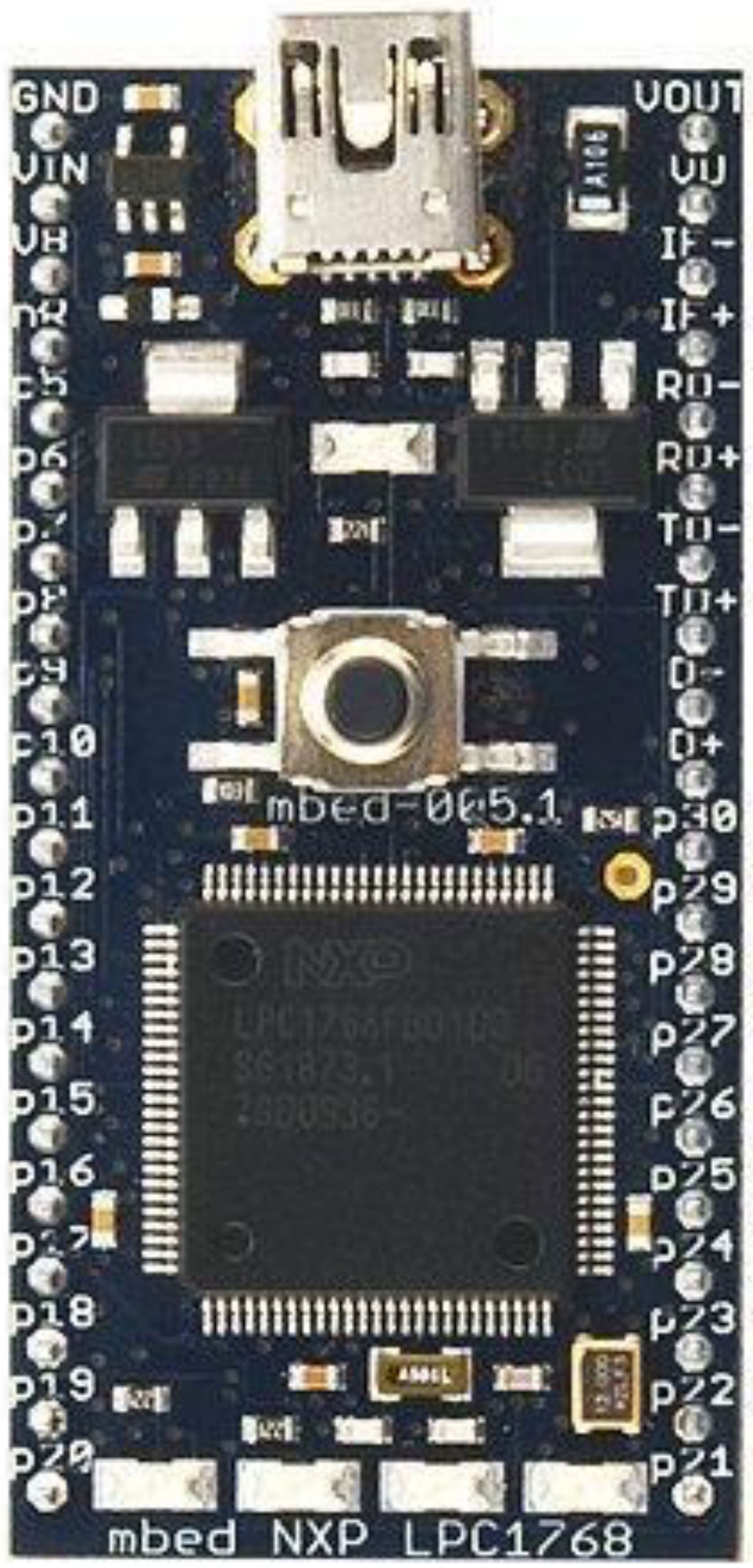
Runs in <256K RAM

TensorFlow compatible

Built on top of Mbed OS 5

(file systems, drivers, 150 boards compatible)

Open source, Apache 2.0 license



arm

uTensor Team

Neil Tan
Arm

Michael Bartling
Arm

Dboy Liao
Piniko

Kazami Hsieh
Academia Sinica


109 YEARS
MB1274 D-01

How?

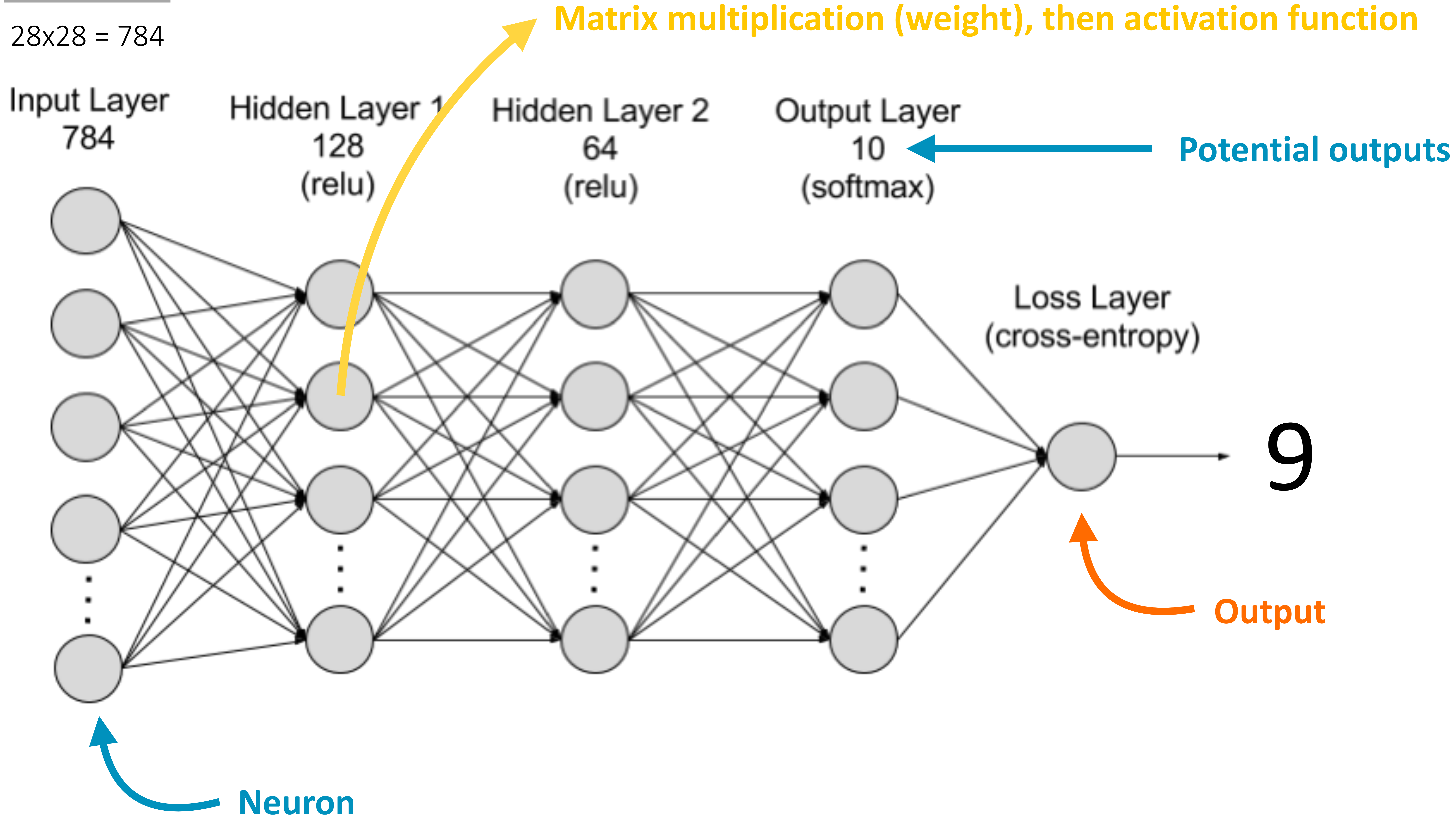
MNIST data set

Training set: 60,000 images

Every drawing is downsampled to 28x28 pixels

Supervised learning through backpropagation

Multi-layer perceptron (MLP) classification



arm

How?

Quantization

8-bit integers instead of 32-bit floats

Only during classification

79.9% accuracy vs. 80.3% accuracy (CIFAR-10)

TensorFlow requires floating-point de-quantization between layers

https://petewarden.com/2016/05/03/how-to-quantize-neural-networks-with-tensorflow/

Memory usage

Matrix multiplication in first hidden layer dominates RAM usage:

`Input elements:`	`784`
`Number of neurons (1st layer):`	`128`
`Number of weight (input to 1st layer):`	`128 * 784`
`Resulting values (Pre-activation function):`	`128`
`Data type:`	`8-bit integer (1 byte)`

`1 byte * (784 + (128 * 784) + 128) = 98.891 kB`

Other tricks

Paging of memory for larger models (sacrifices speed)

Graph in ROM (requires pre-processing) (MNIST: 26K)

Take advantage in sparsity of data, sacrifice accuracy *(TBD)*

Operators

Add, Subtract

Min, Max, ArgMax

ReLU, Matrix multiplication, Reshape, Quantization

Convolution *(WIP)*

Pooling *(WIP)*

Tensors

RAM tensor

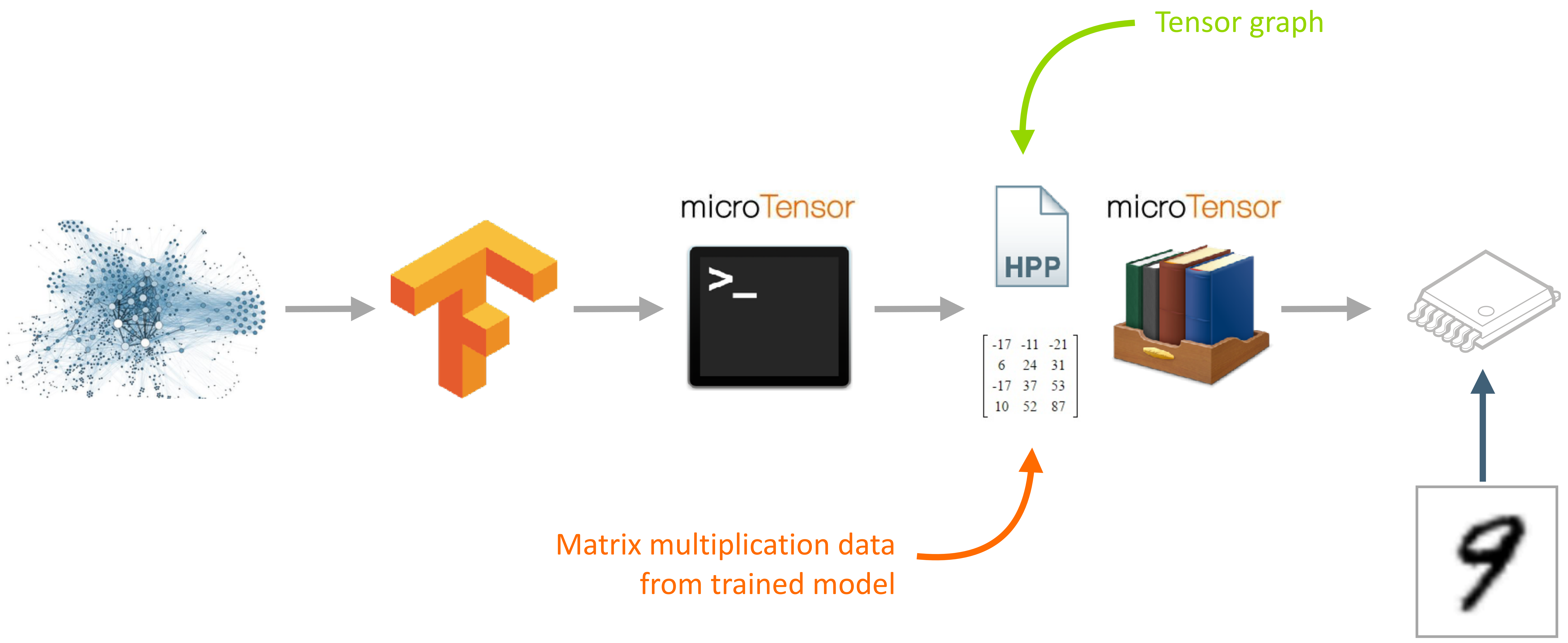
Flash tensor

Sparse tensor

Networked tensor

Tensors can be paged to fit larger networks

Workflow


Tensor graph
microTensor
HPP
microTensor
-17 -11 -21
6 24 31
-17 37 53
10 52 87
Matrix multiplication data
from trained model
9

Developing using the simulator

CMSIS-NN

New neural network kernel functions

Leverages the DSP/SIMD functions in silicon

See speedup of 4-5x

Hardware acceleration for convolution, pooling, etc.

uTensor will be built on top of CMSIS-NN

Recap

1. Buy a development board (http://os.mbed.com/platforms)
2. Clone uTensor (https://github.com/uTensor/uTensor)
3. ???
4. PROFIT!!!

Thank you!

https://labs.mbed.com

Jan Jongboom, Arm