µTensor
AI on Microcontrollers

ARM MBED DEVX

NEIL TAN

github.com/uTensor

WHO

CONTRIBUTORS

NEIL TAN (ARM)

Physicist, Electrical Engineer

Mbed Mad Scientist

KAZAMI HSIEH

ML Researcher, Academia Sinica, Taipei

Startup CTO, Open Source Contributor, Mbed Enthusiast

MICHAEL BARTLING (ARM)

Data Scientist, Mathematician

Mbed Speed Coder

DBOY LIAO

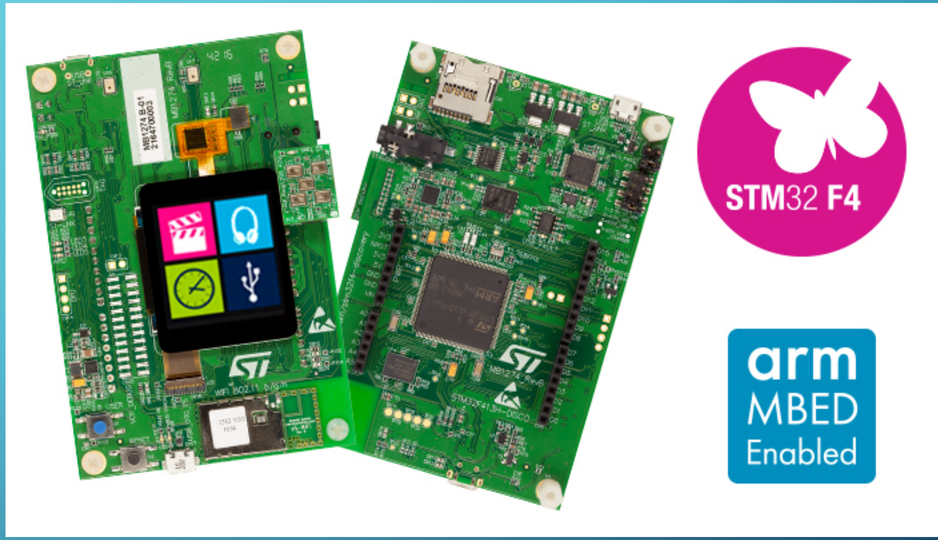
Data Scientist, Piniko, Taipei

Mathematician, Economist, Software Developer

WHAT

MCU EXAMPLE

- 100 MHz
- 320 KB SRAM
- 1.5 MB Flash
- 112 µA/MHz (1.7 V to 3.6 V)

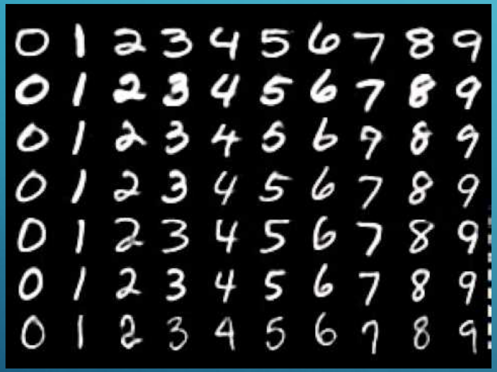


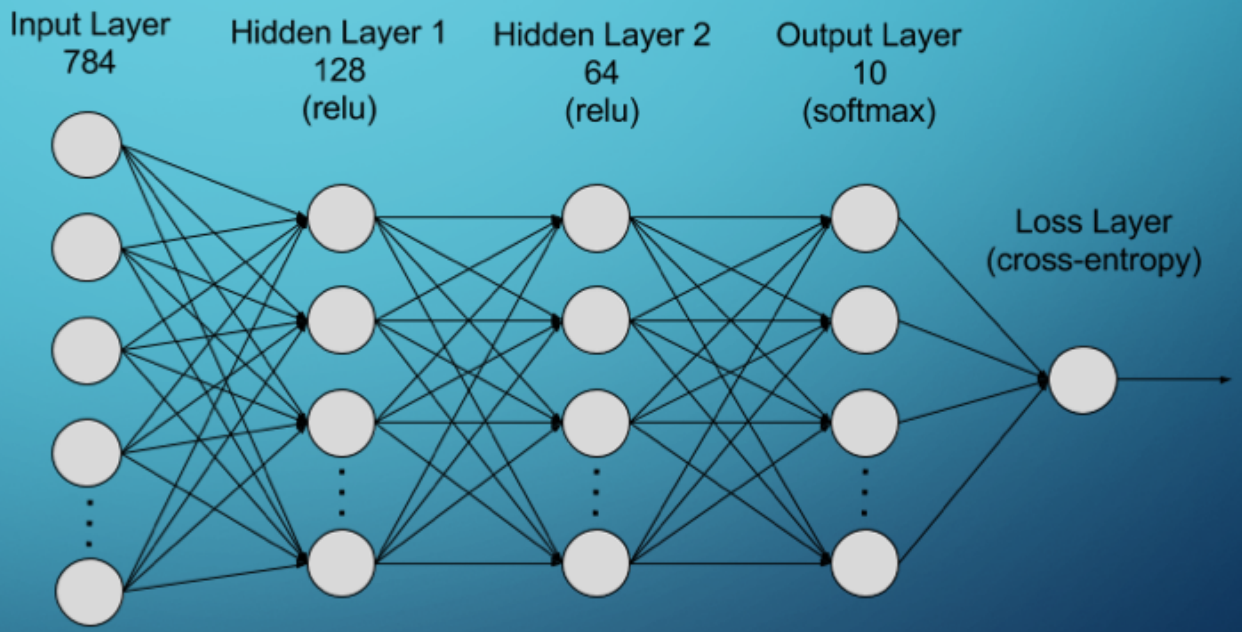
DEMO

goo.gl/uhqgUf

HELLO WORLD

- MNIST Dataset
- 3-layer MLP
- Less than 256Kb of RAM



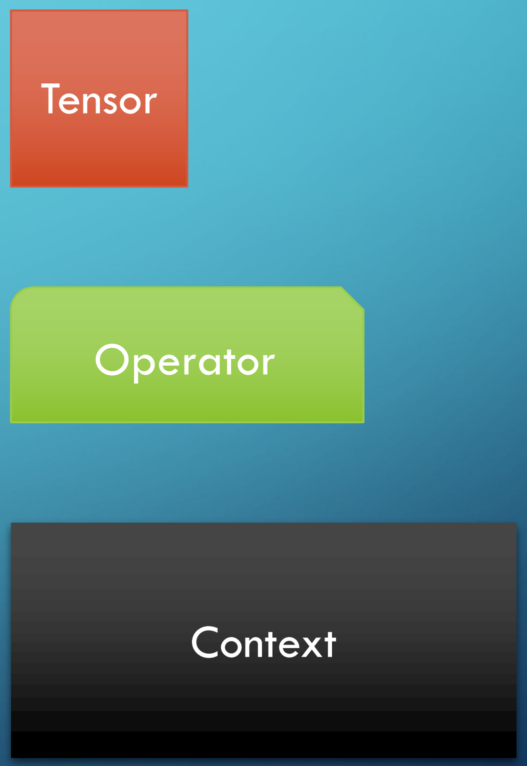


https://mxnet.incubator.apache.org/tutorials/python/mnist.html

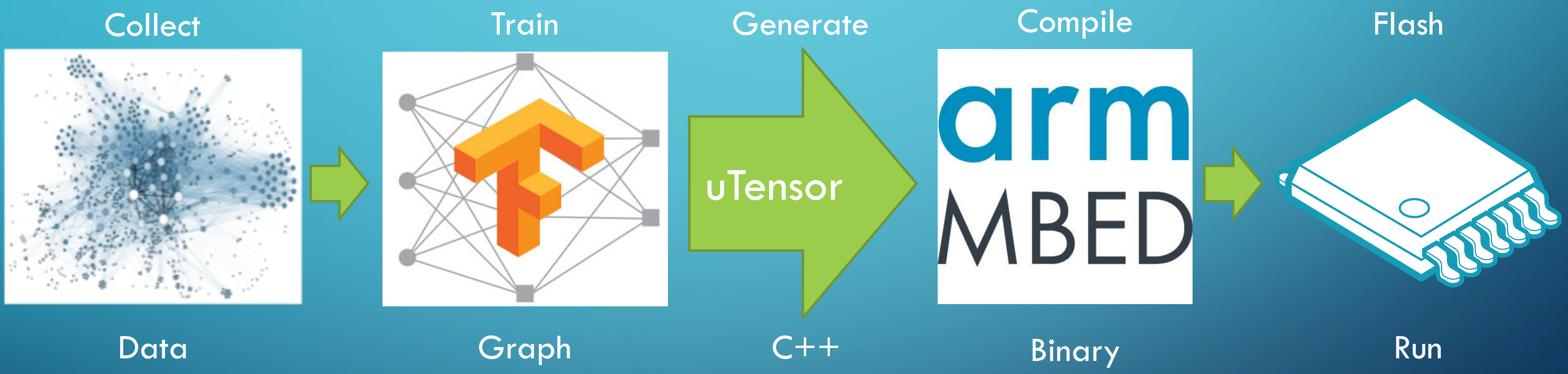
HOW

DESIGN

- Quantization
  - Float to 8-bit
  - 75% Memory reduction
  - See Pete Warden's Blog
- Dynamic Tensor Allocation
- Graph to C++ Code Generation
- Optimized Memory Usage
- Mbed for rapid development



WHAT – UTENSOR WORKFLOW



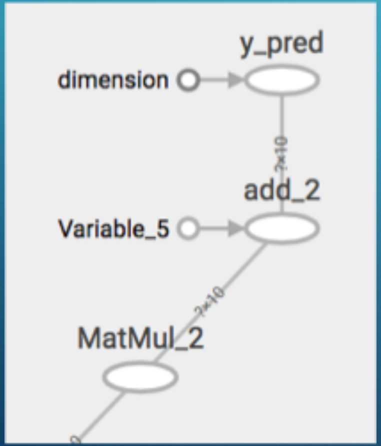
CODE GENERATOR

```
(ut) $ utensor-cli tests/deep_mlp/quant_mnist.pb
  ...
  saving constants/quant_mnist/OuputLayer_Variable_1_0.idx
  saving constants/quant_mnist/y_pred_dimension_0.idx
  Generate header file: models/quant_mnist.hpp
  Generate source file: models/quant_mnist.cpp
```

Graph

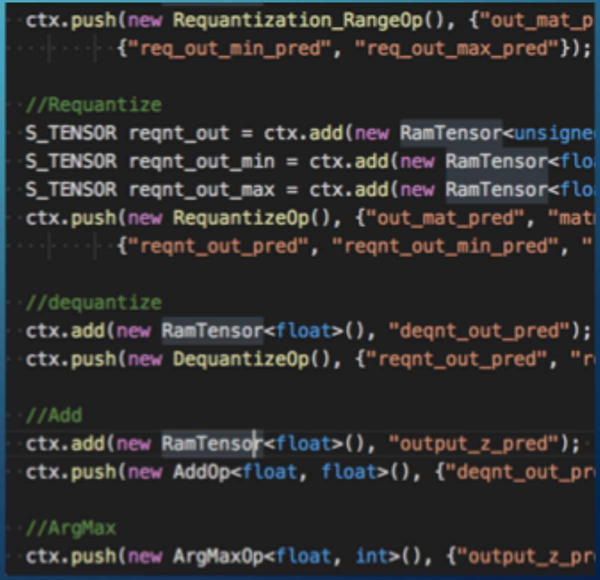
Parameters

CPP files



Graph

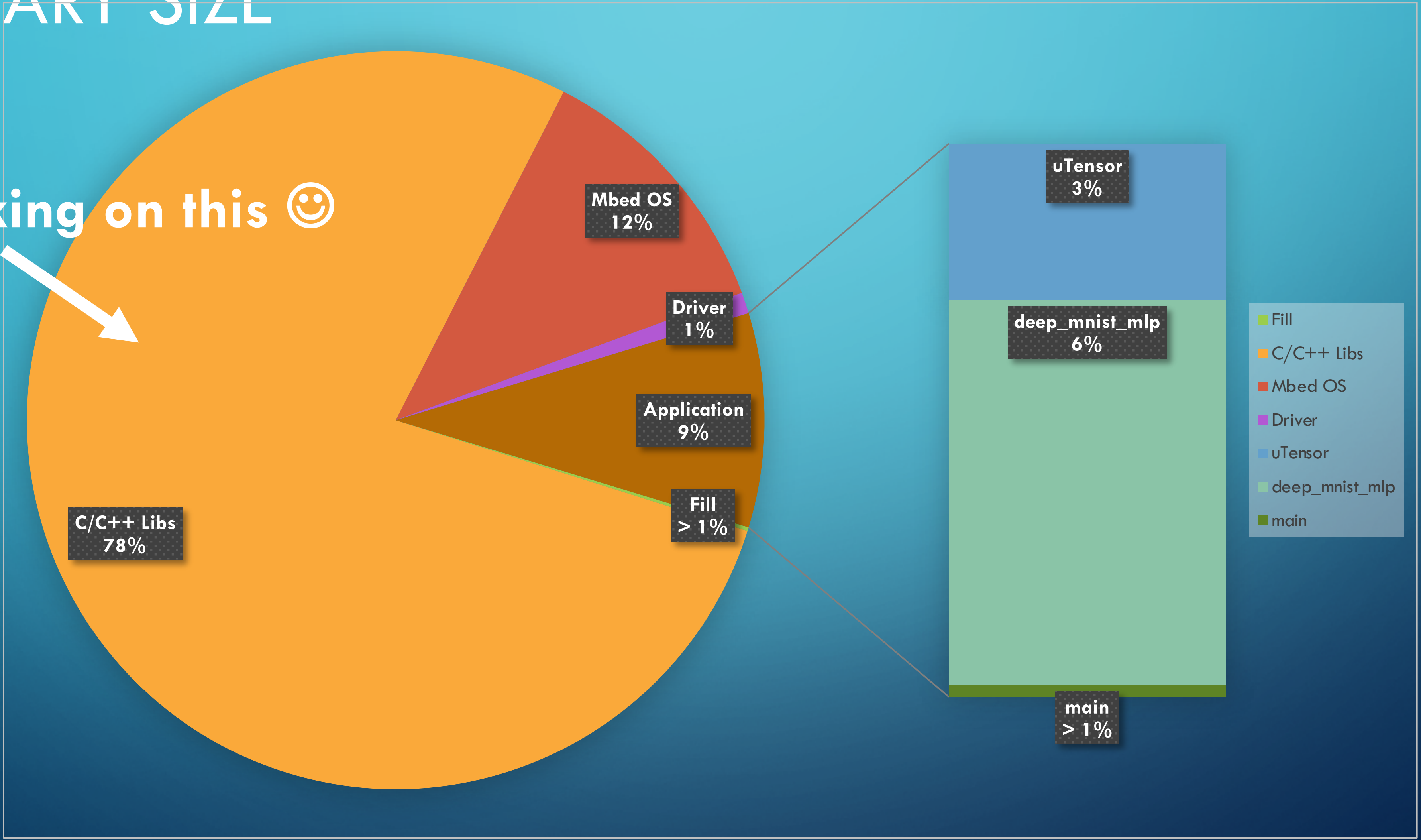
C++



```
ctx.push(new Requantization_RangeOp(), {"out_mat_p
        {"req_out_min_pred", "req_out_max_pred"});

//Requantize
S_TENSOR reqnt_out = ctx.add(new RamTensor<unsigne
S_TENSOR reqnt_out_min = ctx.add(new RamTensor<flo
S_TENSOR reqnt_out_max = ctx.add(new RamTensor<flo
ctx.push(new RequantizeOp(), {"out_mat_pred", "mat
        {"reqnt_out_pred", "reqnt_out_min_pred", "

//dequantize
ctx.add(new RamTensor<float>(), "deqnt_out_pred");
ctx.push(new DequantizeOp(), {"reqnt_out_pred", "r

//Add
ctx.add(new RamTensor<float>(), "output_z_pred");
ctx.push(new AddOp<float, float>(), {"deqnt_out_pr

//ArgMax
ctx.push(new ArgMaxOp<float, int>(), {"output_z_pr
```

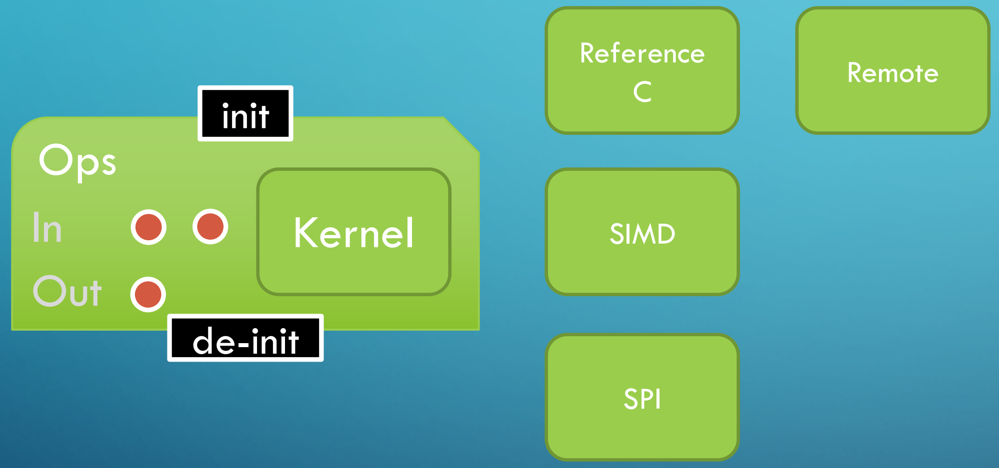
BINARY SIZE

We're working on this ☺

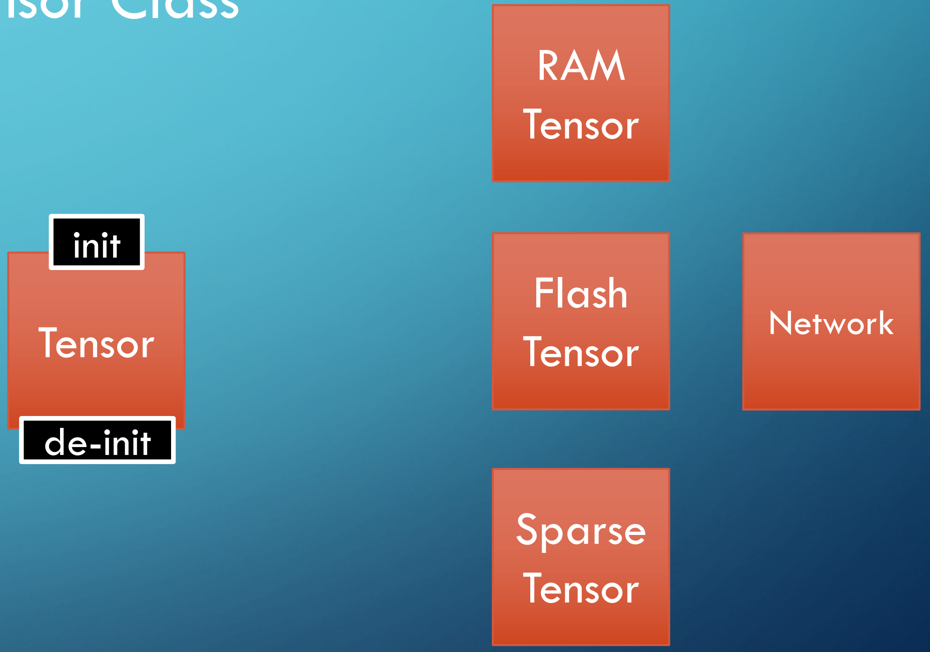


FLEXIBLE CLASSES

Operator Class

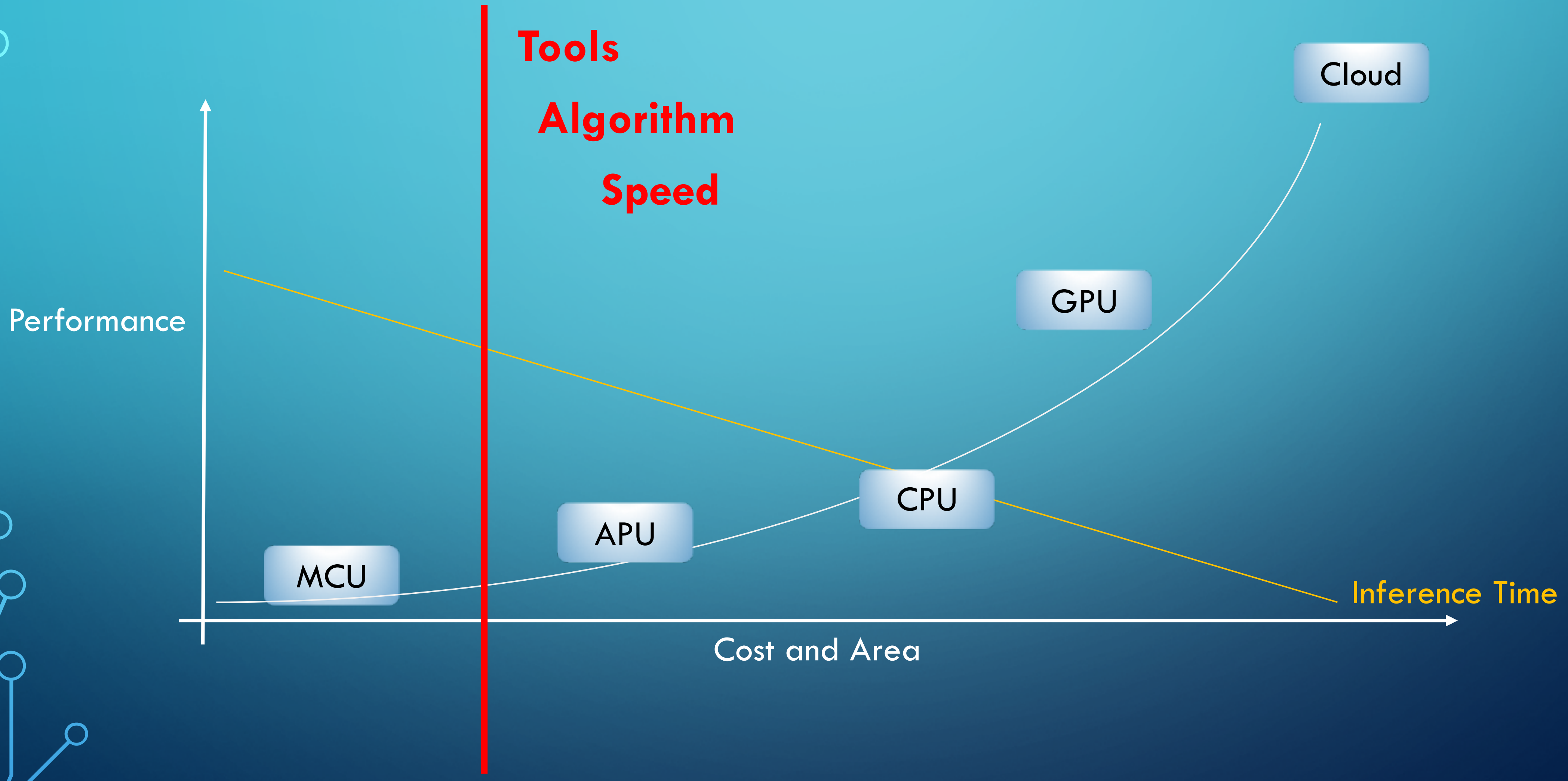


Tensor Class



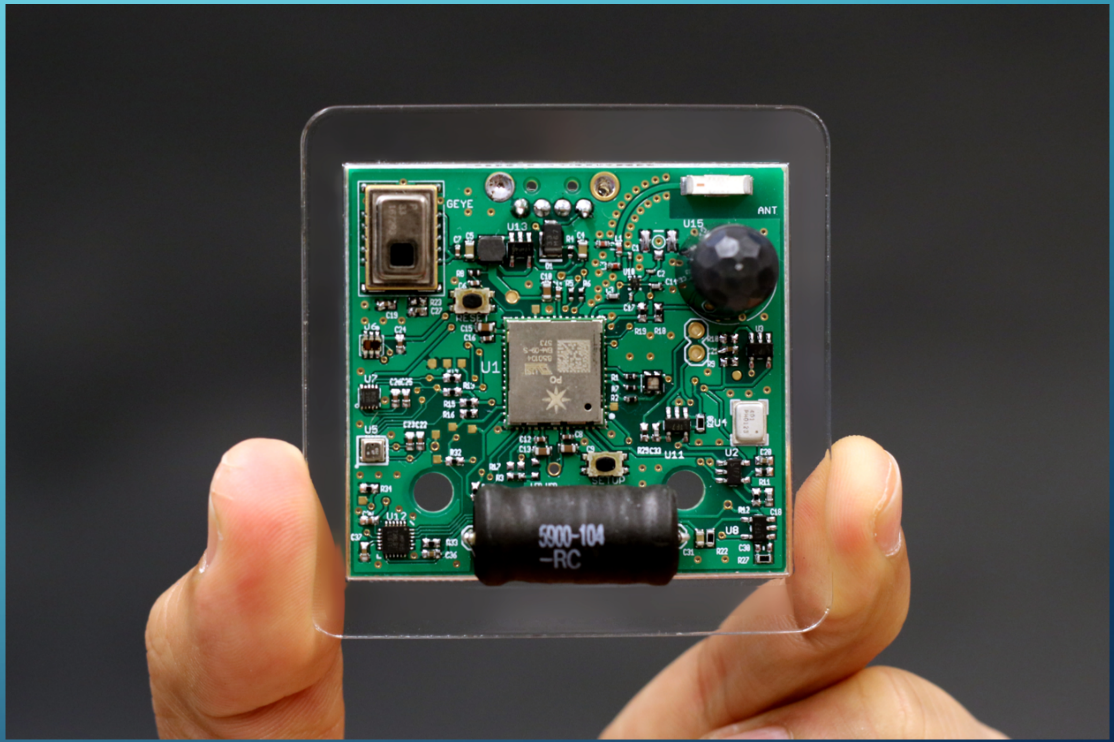
WHY

COMPUTATIONAL DEVICES



WHY DEVICE AI

- Sensor Fusion
- Lora and NB-IoT
- Long Stand-By Time
- Low Cost
- Novel Use-Cases



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

WHERE

THANK YOU!

github.com/uTensor

- PR Welcome!
- Embedded Developers
- Data Scientists