

# Embedded ML Software

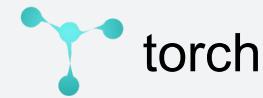
---

 PyTorch  
TensorFlow

Caffe



theano



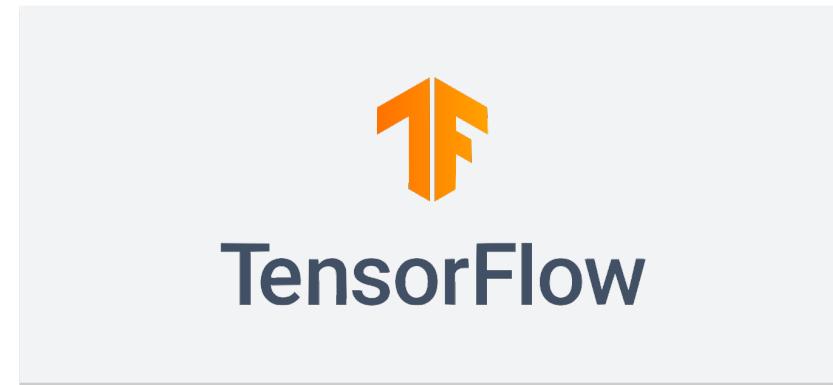
**Canned Estimators**

**Estimators**

**Keras Model**

**Layers**

**Datasets**



**Python Frontend**

**C++**

**Java**

**Go**

**...**

**TensorFlow Distributed Execution Engine**

**CPU**

**GPU**

**OS**

**XLA (Accelerated LinAlg Compiler)**

**CPU**

**GPU**

**TPU**

**...**



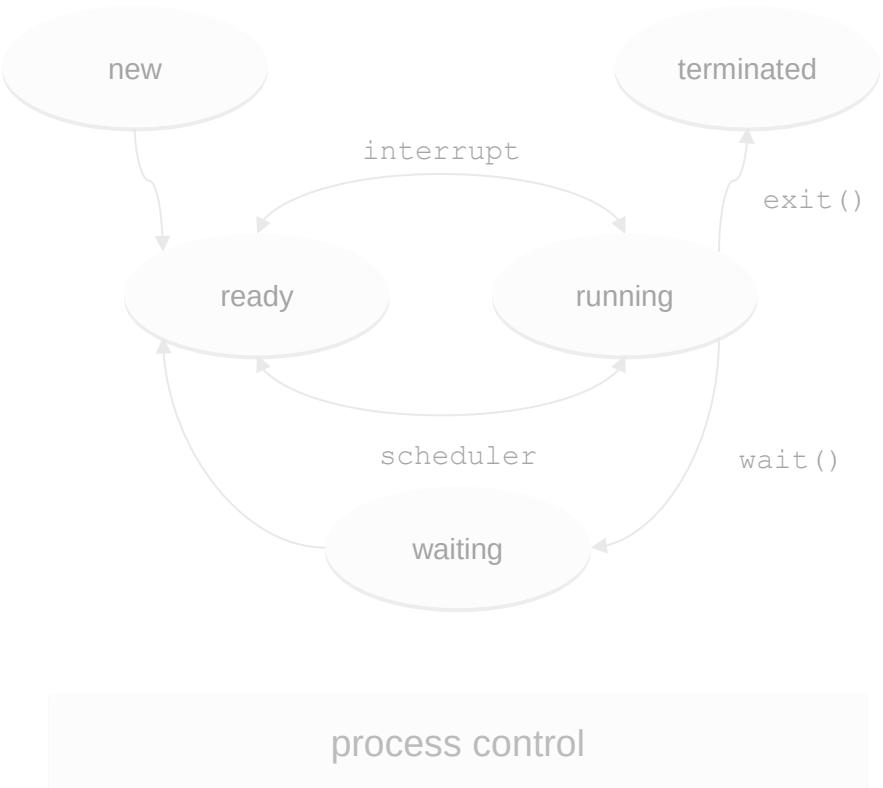
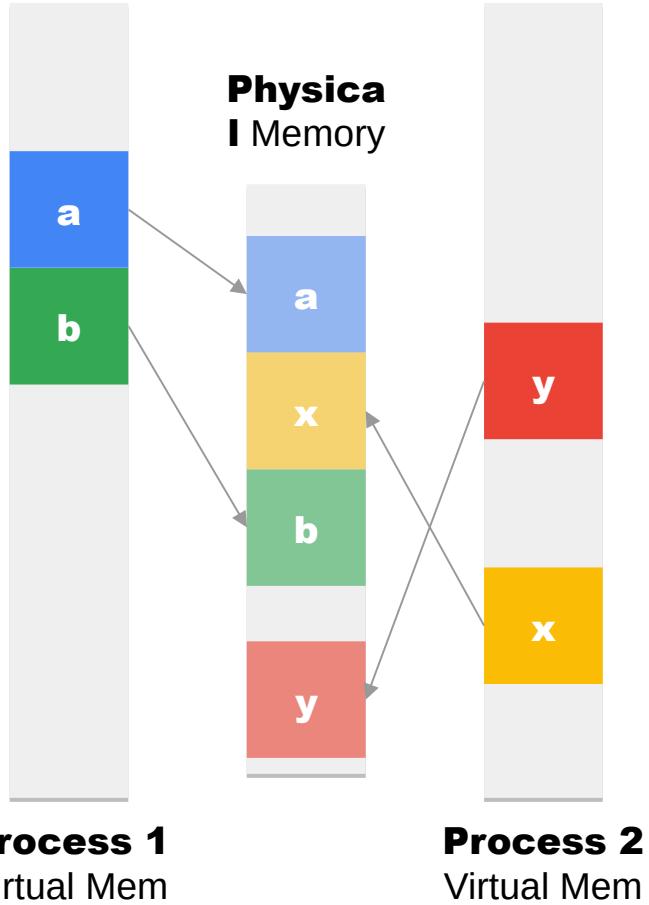
mac  
OS

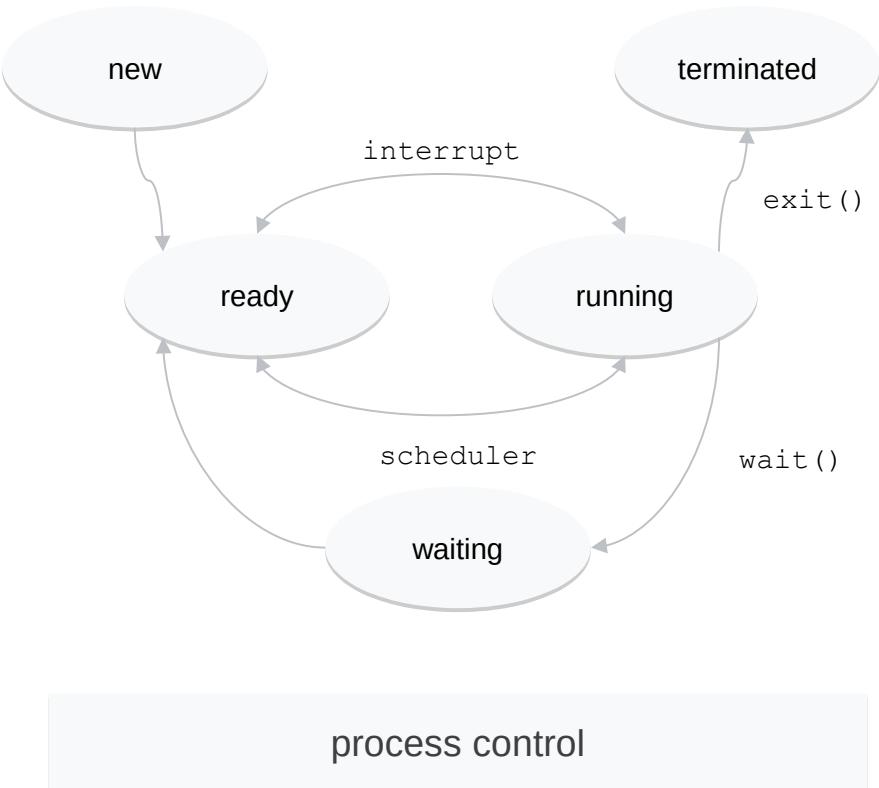
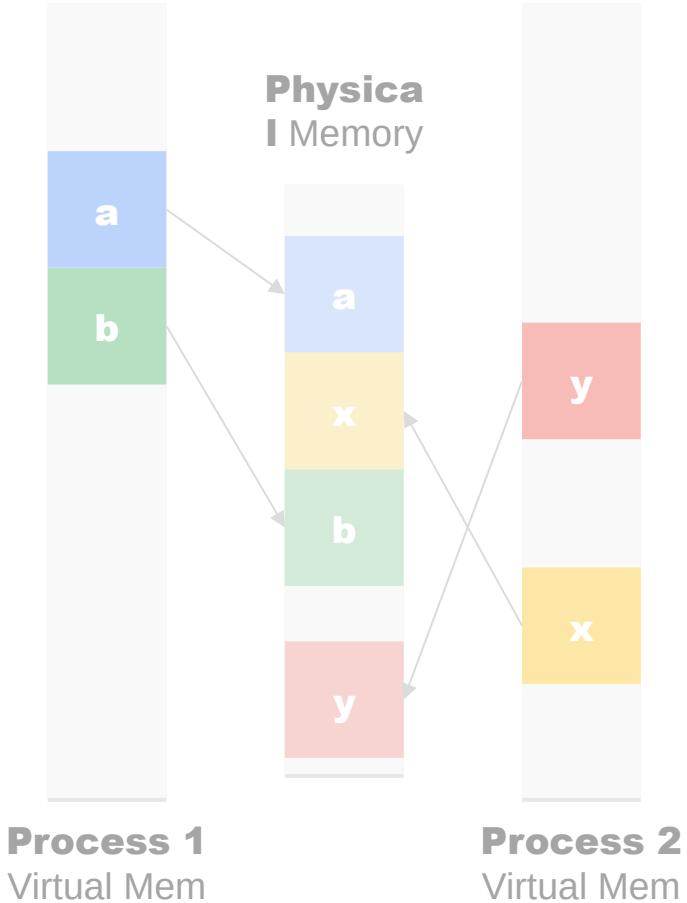


Linux



Mobile OS





# Embedded Systems

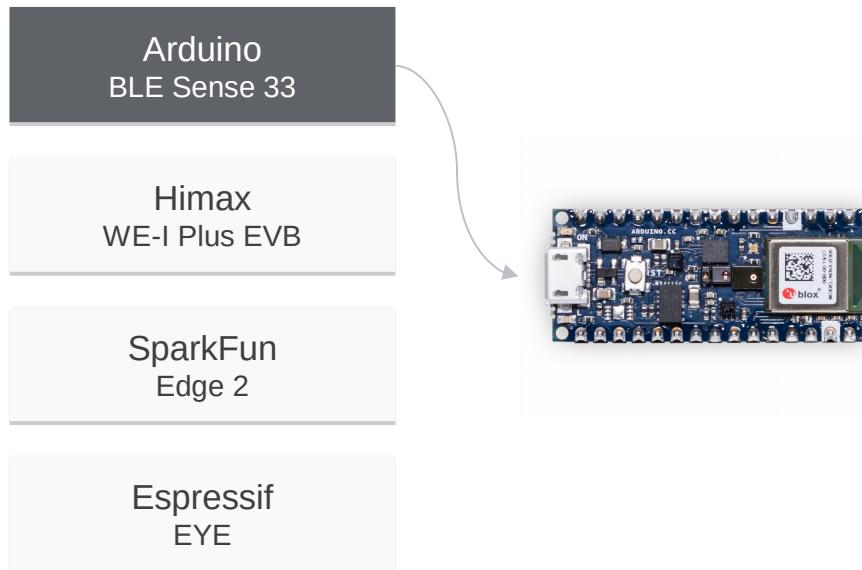
Arduino  
BLE Sense 33

Himax  
WE-I Plus EVB

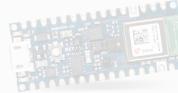
SparkFun  
Edge 2

Espressif  
EYE

# Embedded Systems



	<b>Microprocessor</b>	<b>&gt;</b>	<b>Microcontroller</b>
Platform			
Compute	1GHz–4GHz	<b>~10X</b>	1MHz–400MHz
Memory	512MB–64GB	<b>~10000X</b>	2KB–512KB
Storage	64GB–4TB	<b>~100000X</b>	32KB–2MB
Power	30W–300W	<b>~1000X</b>	150µW–23.5mW

	<b>Microprocessor</b>	<b>&gt;</b>	<b>Microcontroller</b>
Platform			
Compute	1GHz–4GHz	<b>~10X</b>	1MHz–400MHz
Memory	512MB–64GB	<b>~10000X</b>	2KB–512KB
Storage	64GB–4TB	<b>~100000X</b>	32KB–2MB
Power	30W–300W	<b>~1000X</b>	150µW–23.5mW



Less memory

Limited OS support

Lower compute power

Only focused on *inference*



# **TinyML** Inference Framework



# ML Framework Design Checklist

## Model

Support for training?	Yes	No
Support for inference?	Yes	No
How many ops?	few	

# ML Framework Design Checklist

## Model

Support for training?

Yes

No

Support for inference?

Yes

No

How many ops?

few

## Software

Need quantization/optimization tools?

Yes

No

Can rely on virtual memory support?

Yes

No

# ML Framework Design Checklist

## Model

Support for training?

Yes

No

Support for inference?

Yes

No

How many ops?

few

## Software

Need quantization/optimization tools?

Yes

No

Can rely on virtual memory support?

Yes

No

## Hardware

Support a diverse range of processor hardware on a device?

Yes

No

Need to support many different platforms and architectures?

Yes

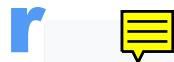
No

# “Tiny” Machine Learning **Frameworks**



TensorFlow Lite Micro

**uTenso**



**STM32**  
**Cube.AI**

**Arduino**  
**BLE Sense 33**

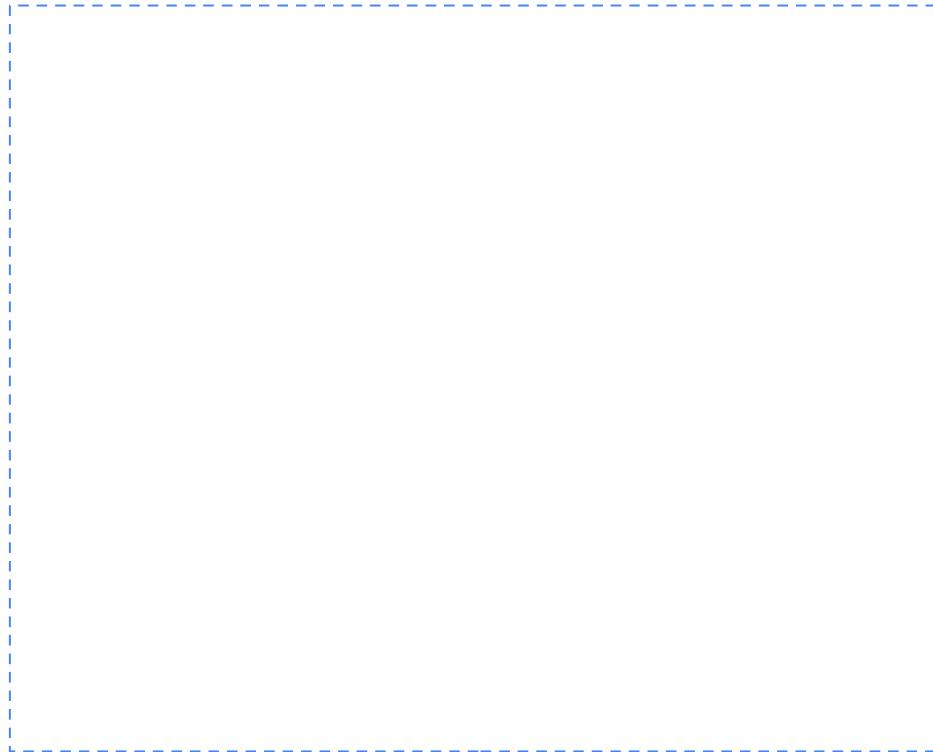
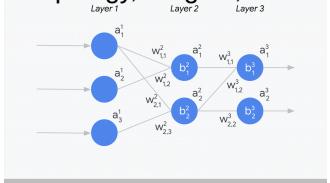
**Himax**  
**WE-I Plus EVB**

**SparkFun**  
**Edge 2**

**Espressif**  
**EYE**

## pre-trained NN model

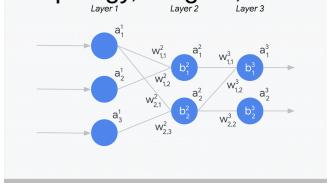
topology, weights, bias



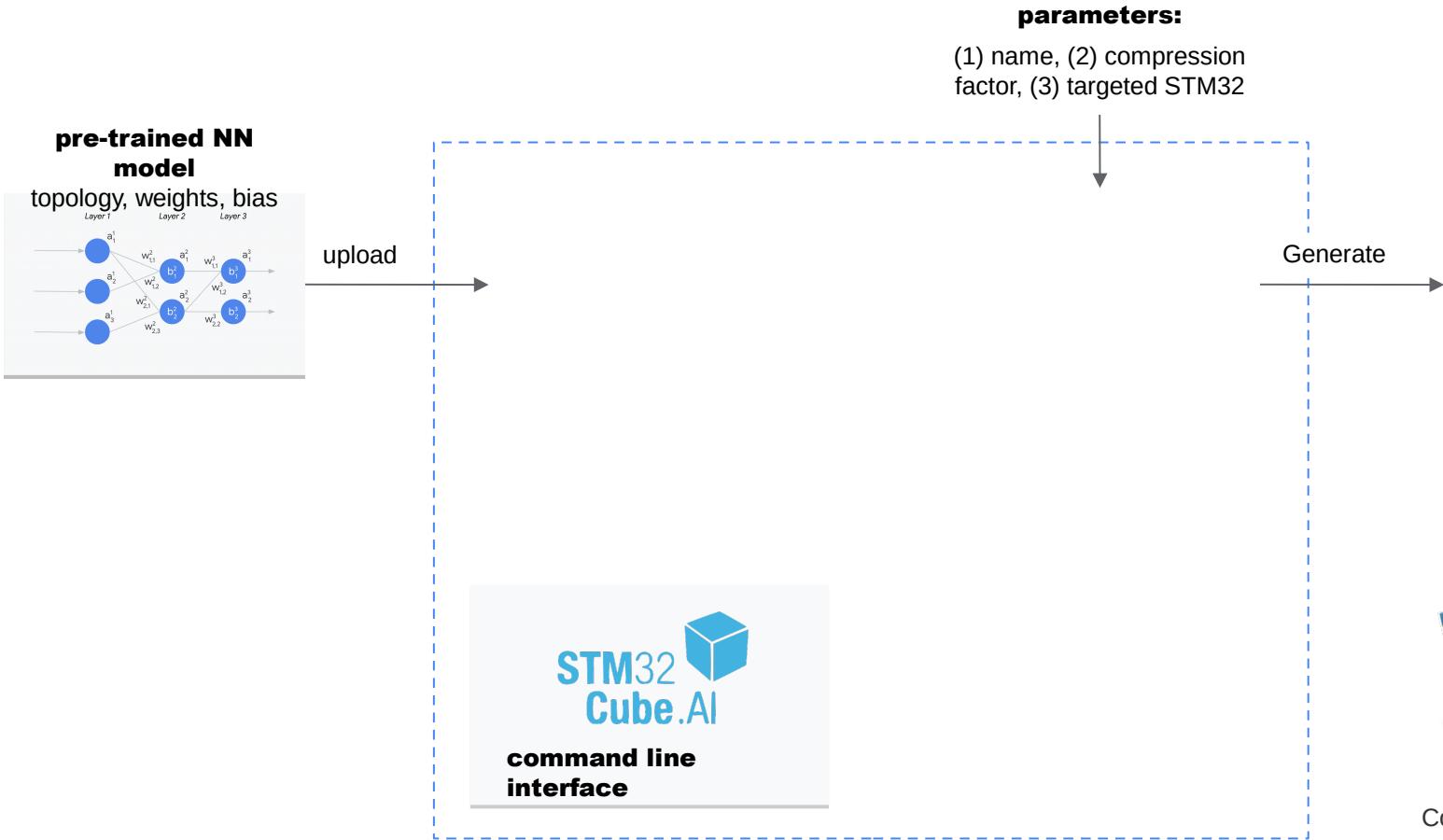
**STM32L476**  
Cortex-M4 **MCU**

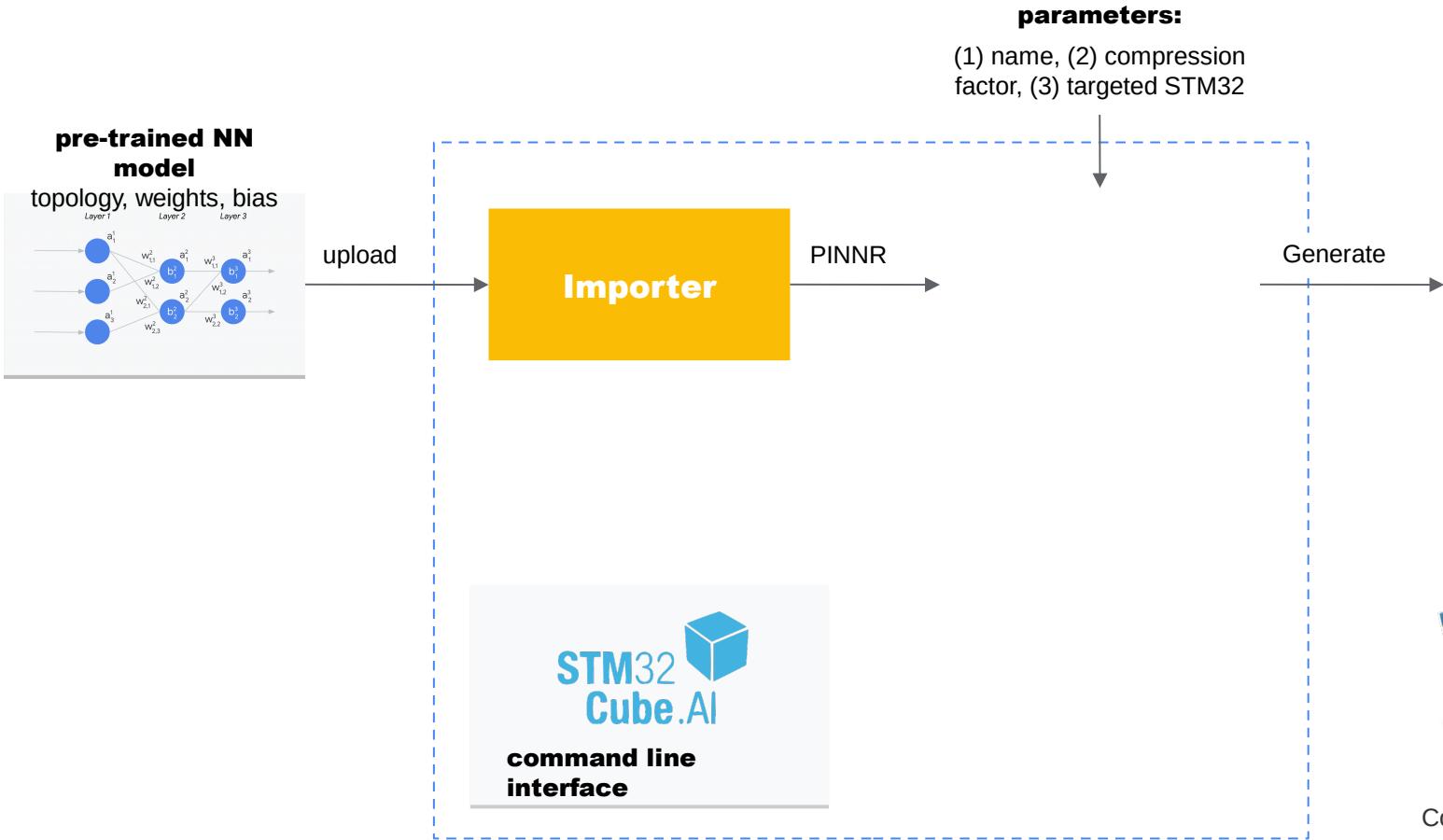
## pre-trained NN model

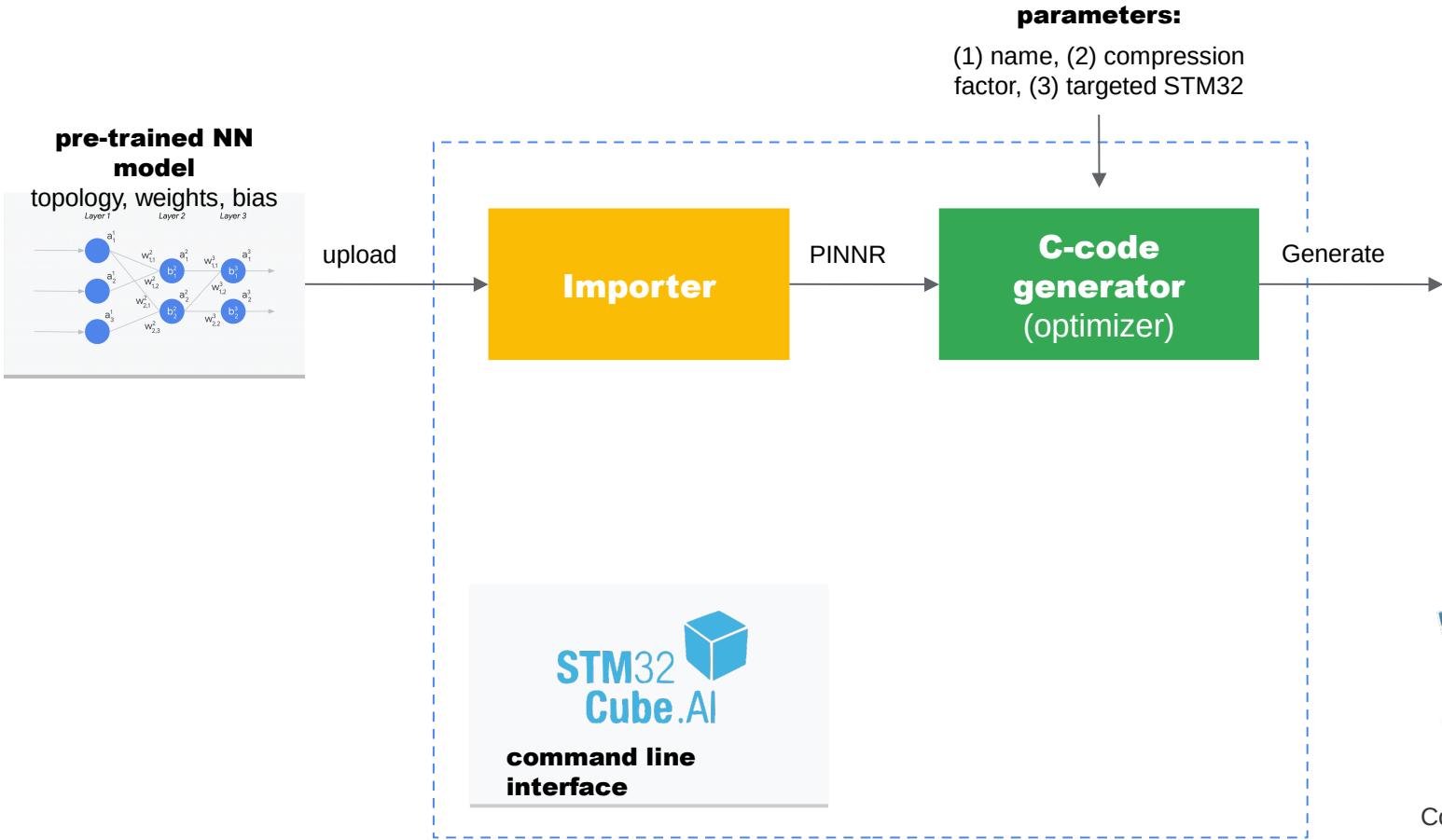
topology, weights, bias

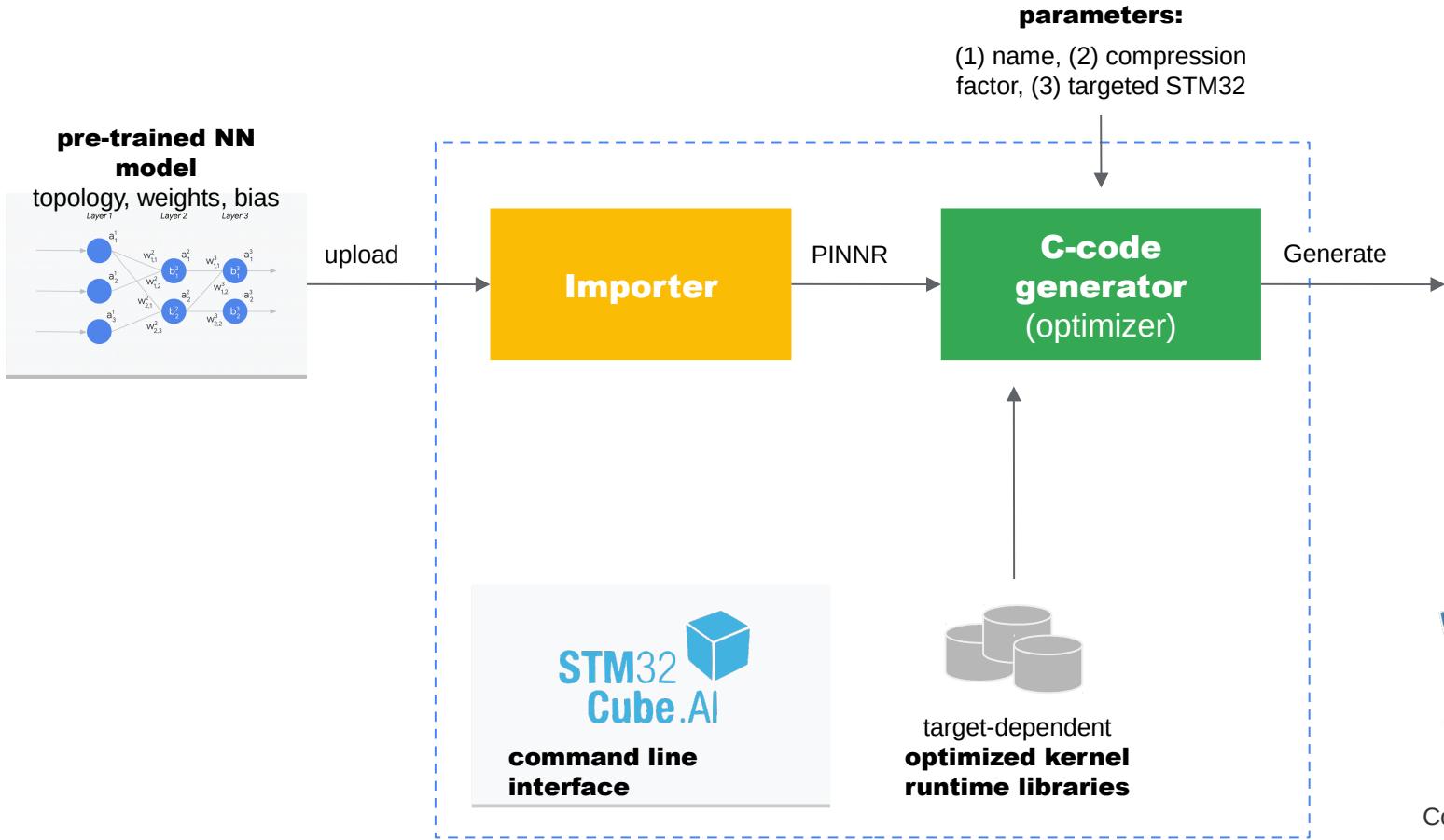


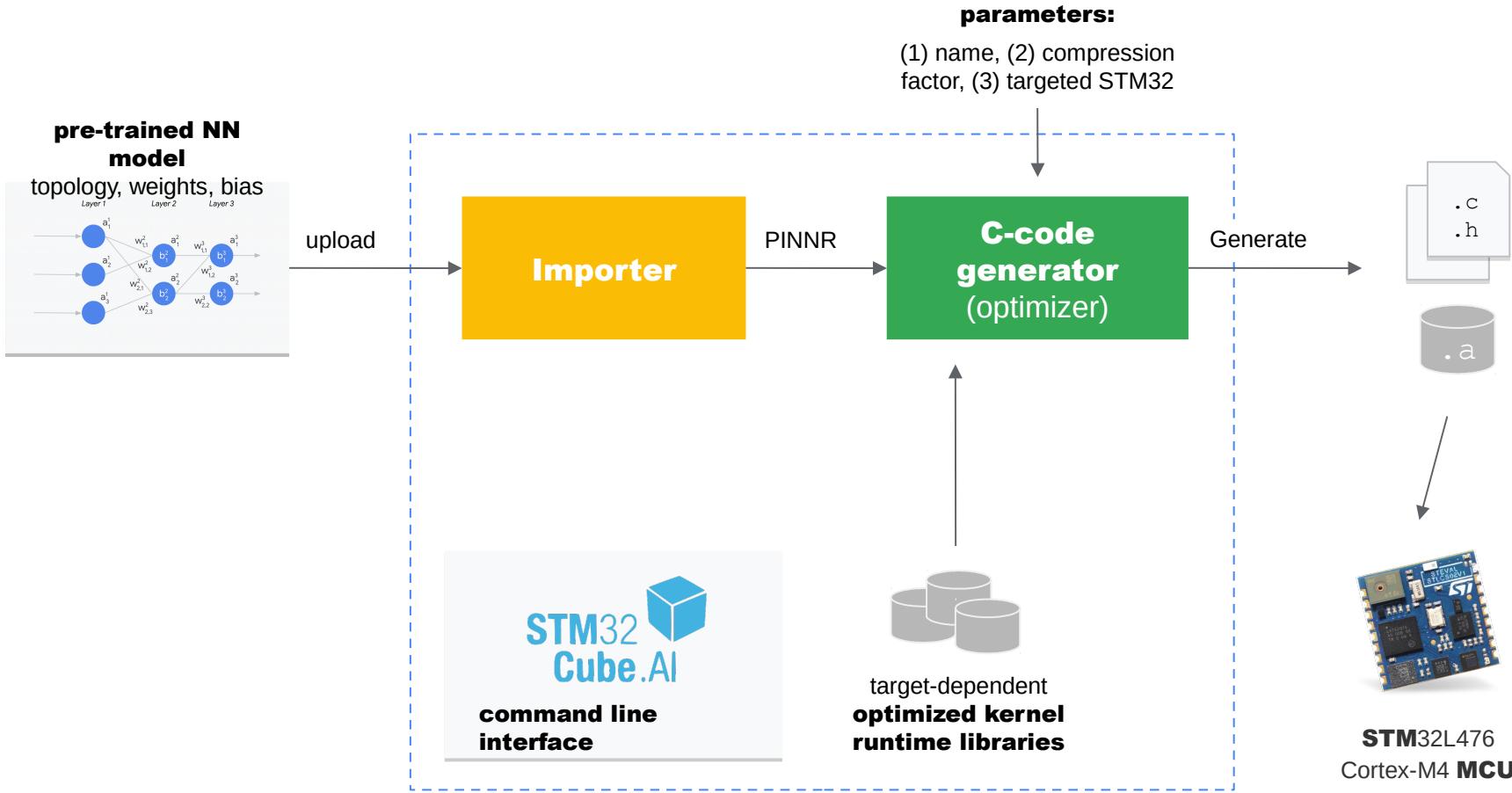
**STM32L476**  
Cortex-M4 **MCU**



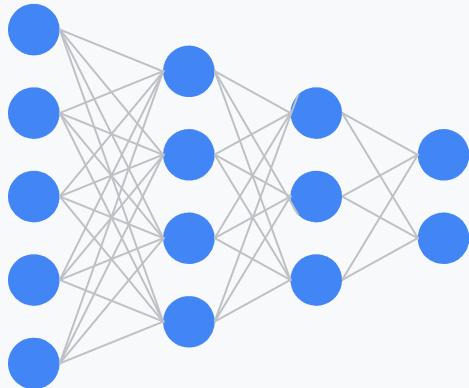








## Model



3 dense **layers**  
321 **parameters**  
**Nb** quantization or compression



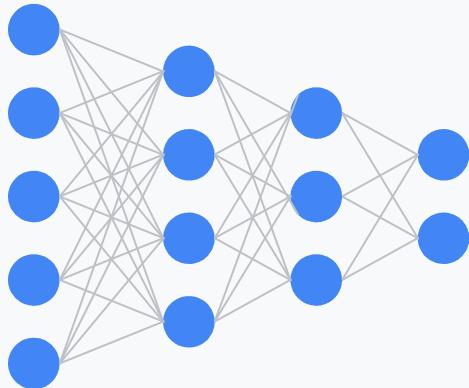
**27kB** Flash  
**5kB** RAM  
**77uS** Inference  
Time  
**Closed** Source



**TensorFlow Lite**

**50kB** Flash  
**4.7kB** RAM  
**104uS** Inference  
Time  
**Open** Source

## Model



3 dense **layers**  
321 **parameters**  
**Nb** quantization or compression



**27kB** Flash  
**5kB** RAM  
**77uS** Inference

Time

**Closed** Source



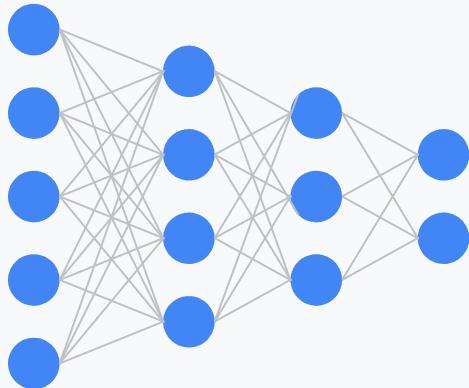
**TensorFlow Lite**

**50kB** Flash  
**4.7kB** RAM  
**104uS** Inference

Time

**Open** Source

## Model



3 dense **layers**  
321 **parameters**  
**Nb** quantization or compression



**27kB** Flash  
**5kB** RAM  
**77uS** Inference  
Time  
**Closed** Source



**TensorFlow Lite**

**50kB** Flash  
**4.7kB** RAM  
**104uS** Inference  
Time  
**Open** Source

# Choosing Frameworks

## Hardware, IDE

**compute**  
**memory** constraints

## Training/Embedded

**training framework**  
**embedded framework** (you'll  
export to)

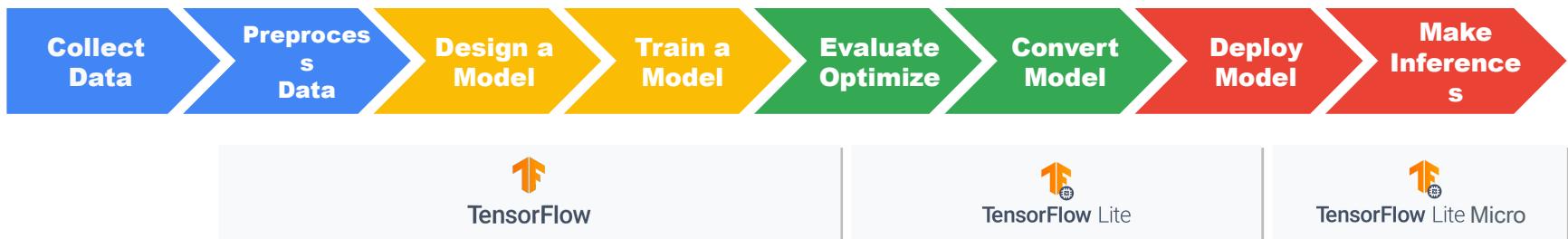
## other

**documentation**

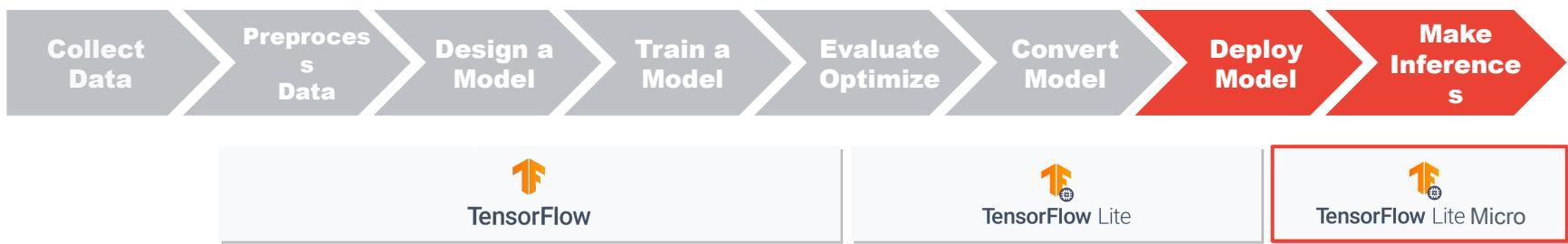
**sample code**  
(for use case)

**personal experience**

# Choosing Frameworks



# Choosing Frameworks



**Pete Warden**, Technical Lead, TensorFlow  
Mobile and Embedded Team, Google.