Embedded ML Software

PyTorch

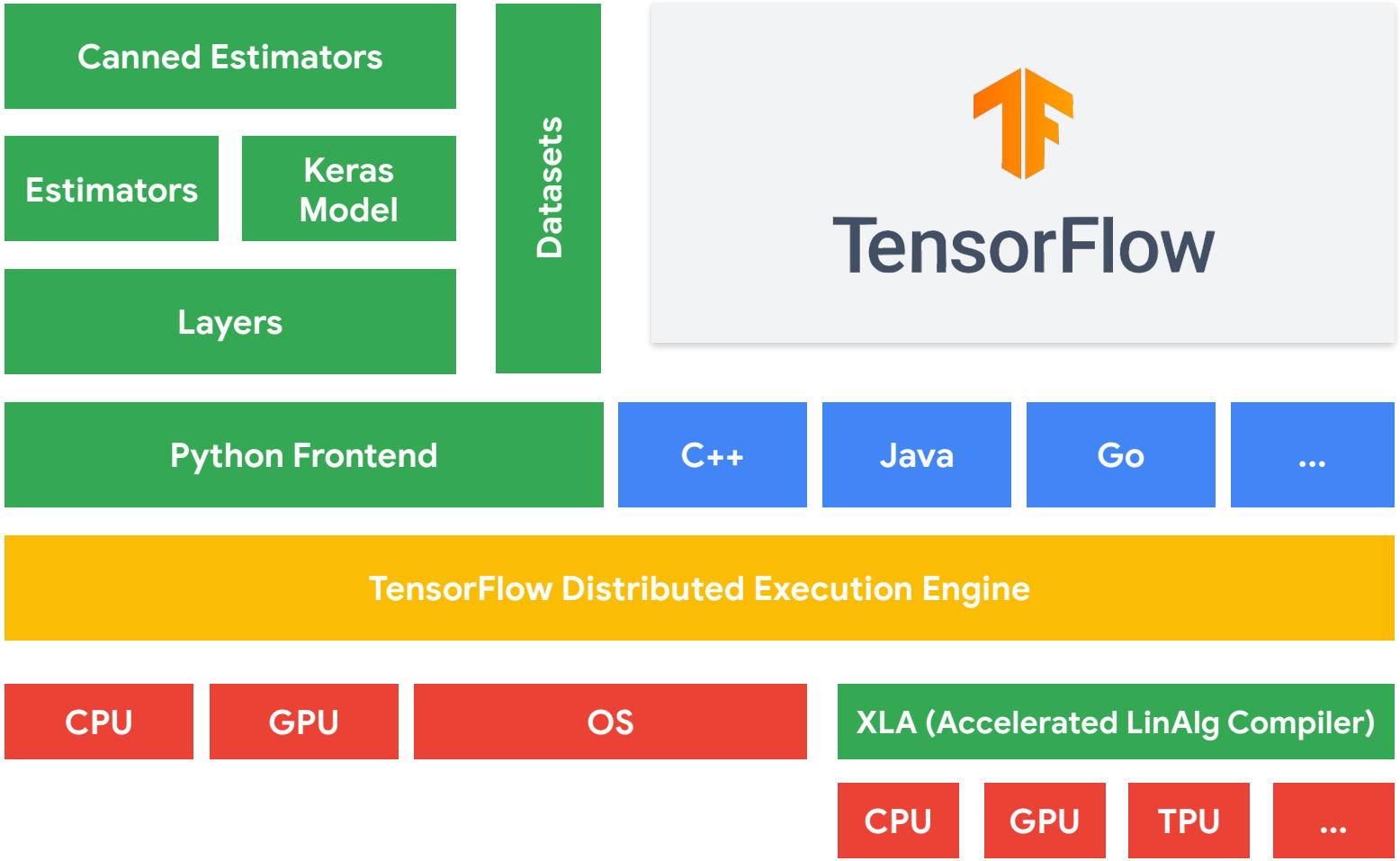
TensorFlow

Caffe

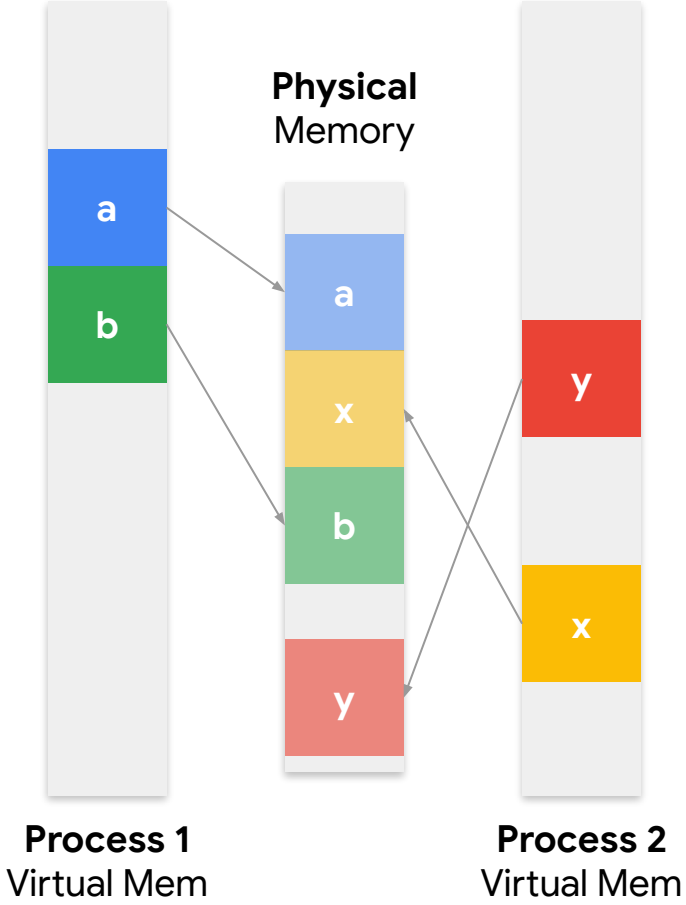
scikit learn

theano

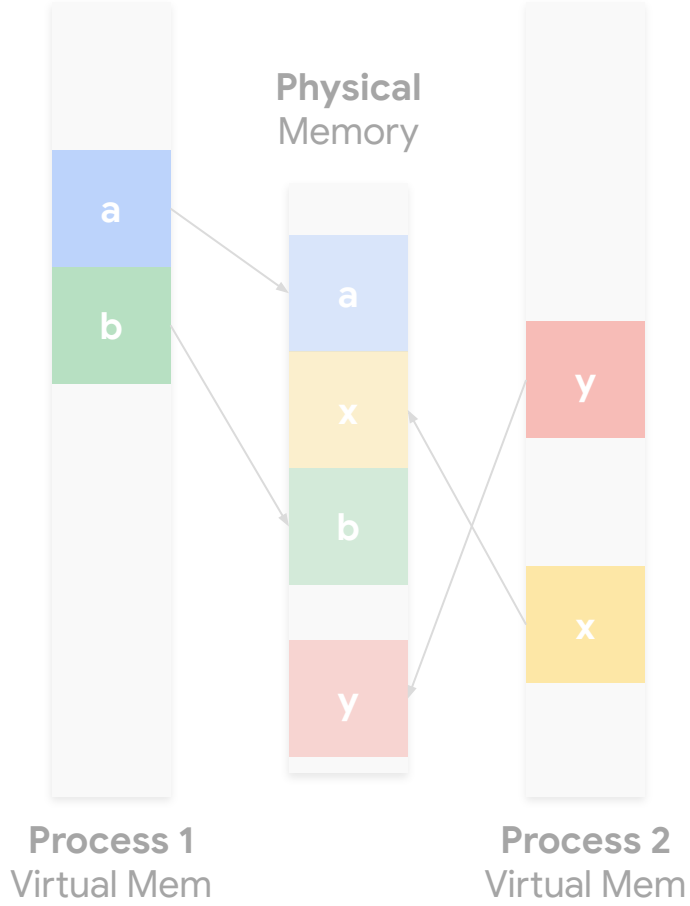
torch

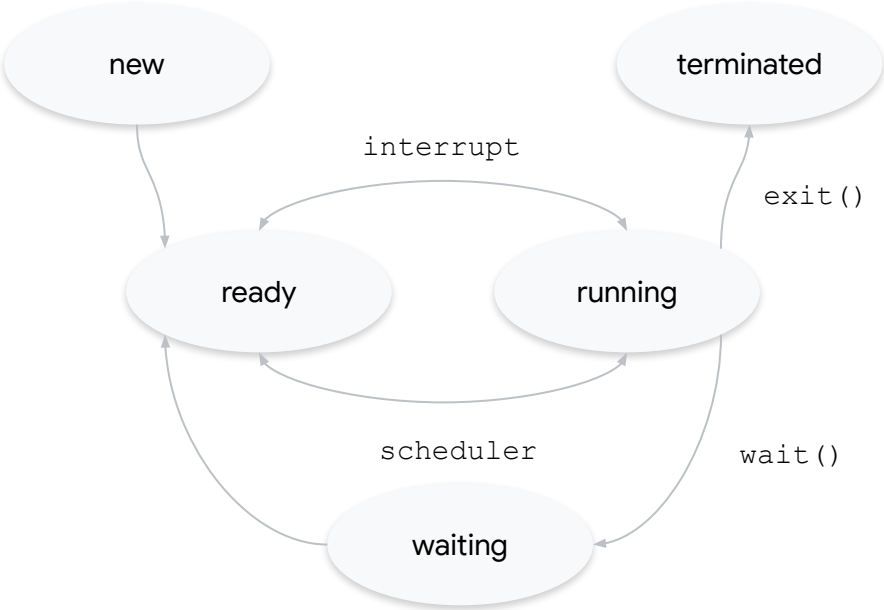

Canned Estimators
Estimators
Keras Model
Layers
Datasets
TensorFlow
Python Frontend
C++
Java
Go
...
TensorFlow Distributed Execution Engine
CPU
GPU
OS
XLA (Accelerated LinAlg Compiler)
CPU
GPU
TPU
...


mac
OS
iOS
Linux
Mobile OS


Physical
Memory
a
b
a
x
b
y
y
x
Process 1
Virtual Mem
Process 2
Virtual Mem


new
terminated
interrupt
exit()
ready
running
scheduler
wait()
waiting
process control





process control

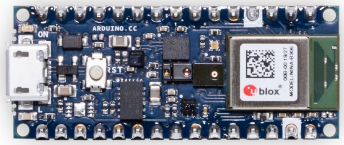
Embedded Systems

Arduino
BLE Sense 33

Himax
WE-I Plus EVB

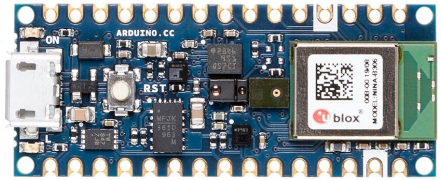
SparkFun
Edge 2

Espressif
EYE



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

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



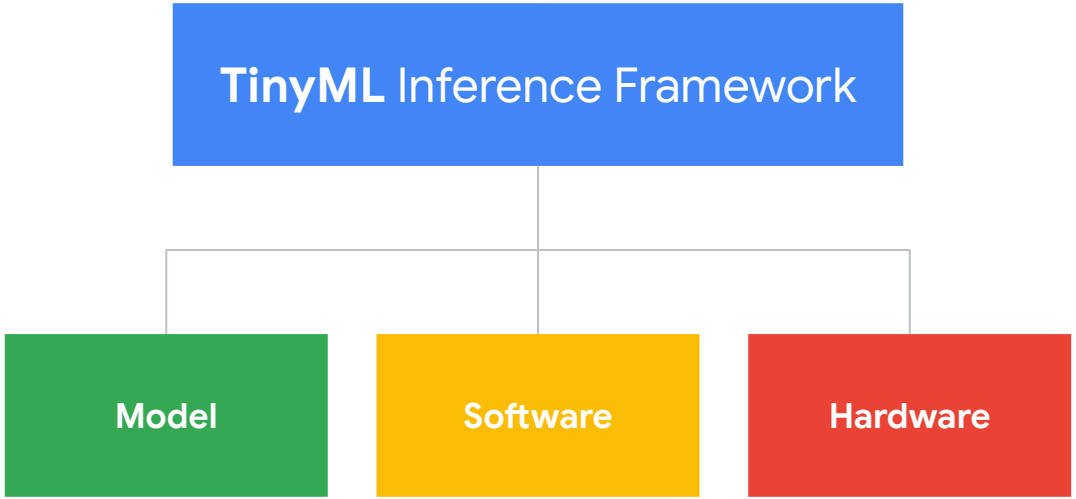
Less memory

Limited OS support

Lower compute power

Only focused on *inference*

?


TinyML Inference Framework
Model
Software
Hardware

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

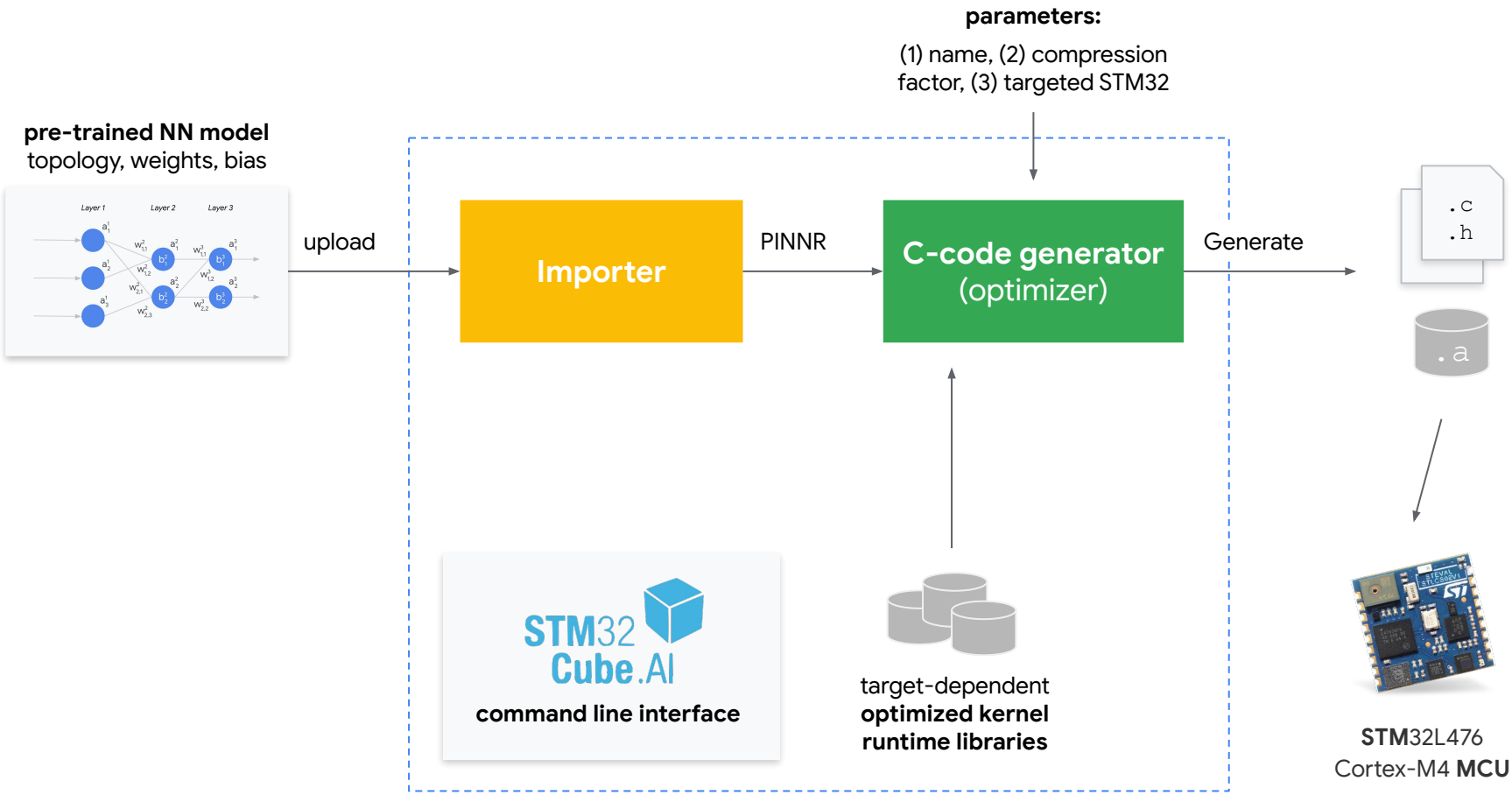
“Tiny” Machine Learning **Frameworks**

Arduino
BLE Sense 33

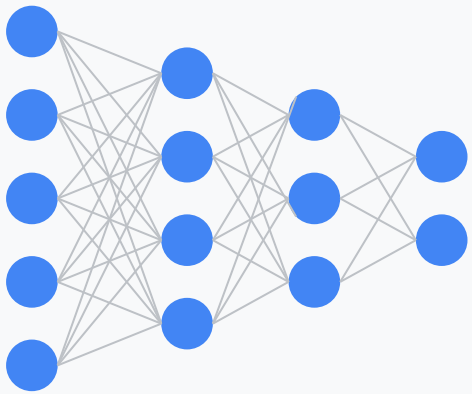
Himax
WE-I Plus EVB

SparkFun
Edge 2

Espressif
EYE


parameters:
(1) name, (2) compression factor, (3) targeted STM32
pre-trained NN model
topology, weights, bias
Layer 1
Layer 2
Layer 3
upload
Importer
PINNR
C-code generator
(optimizer)
Generate
.c
.h
.a
STM32
Cube.AI
command line interface
target-dependent
optimized kernel
runtime libraries
STM32L476
Cortex-M4 MCU

Model



*3 dense **layers***
*321 **parameters***
***No** quantization or compression*

STM32 Cube.AI

27kB Flash
5kB RAM
77uS Inference Time

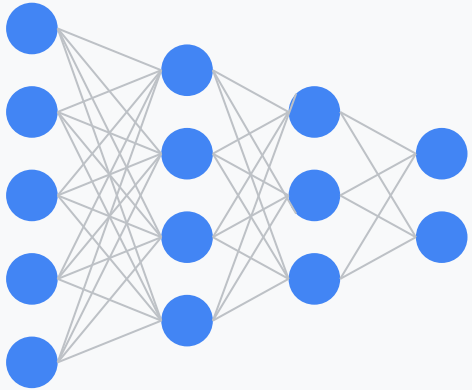
Closed Source

TensorFlow Lite

50kB Flash
4.7kB RAM
104uS Inference Time

Open Source

Model



*3 dense **layers***
*321 **parameters***
***No** quantization or compression*

STM32 Cube.AI

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

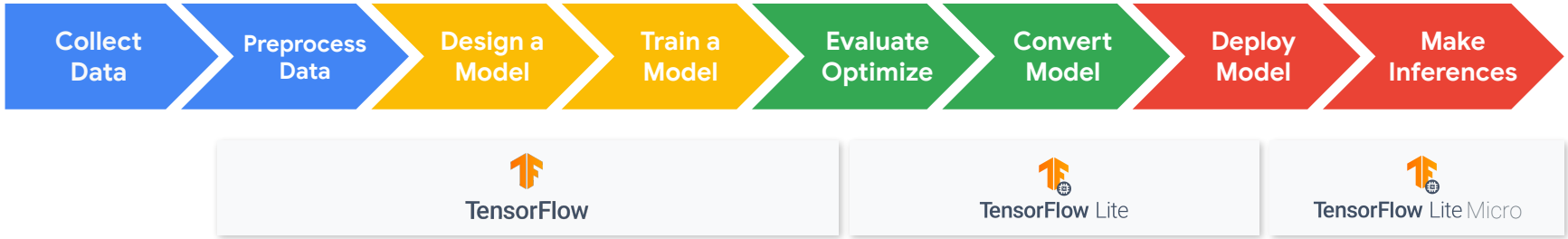
TensorFlow Lite

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

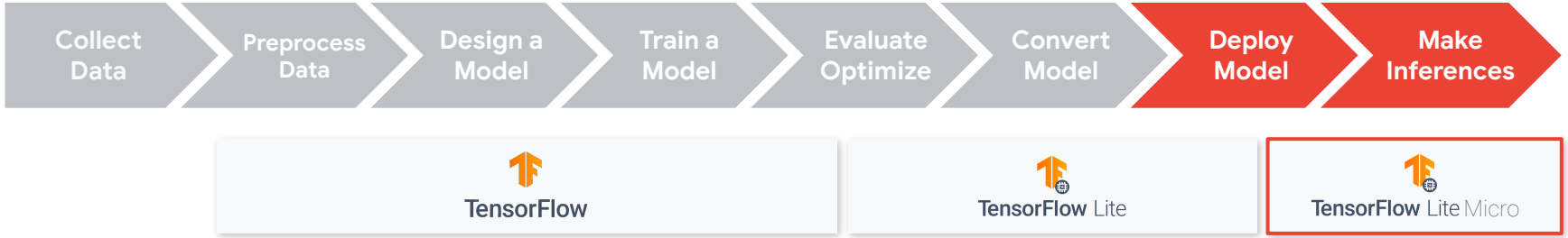
Choosing Frameworks

Hardware, IDE	Training/Embedded	other
compute **memory** constraints	**training framework** **embedded framework** (you'll export to)	**documentation** **sample code** (for use case) **personal experience**

Choosing Frameworks



Choosing Frameworks



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