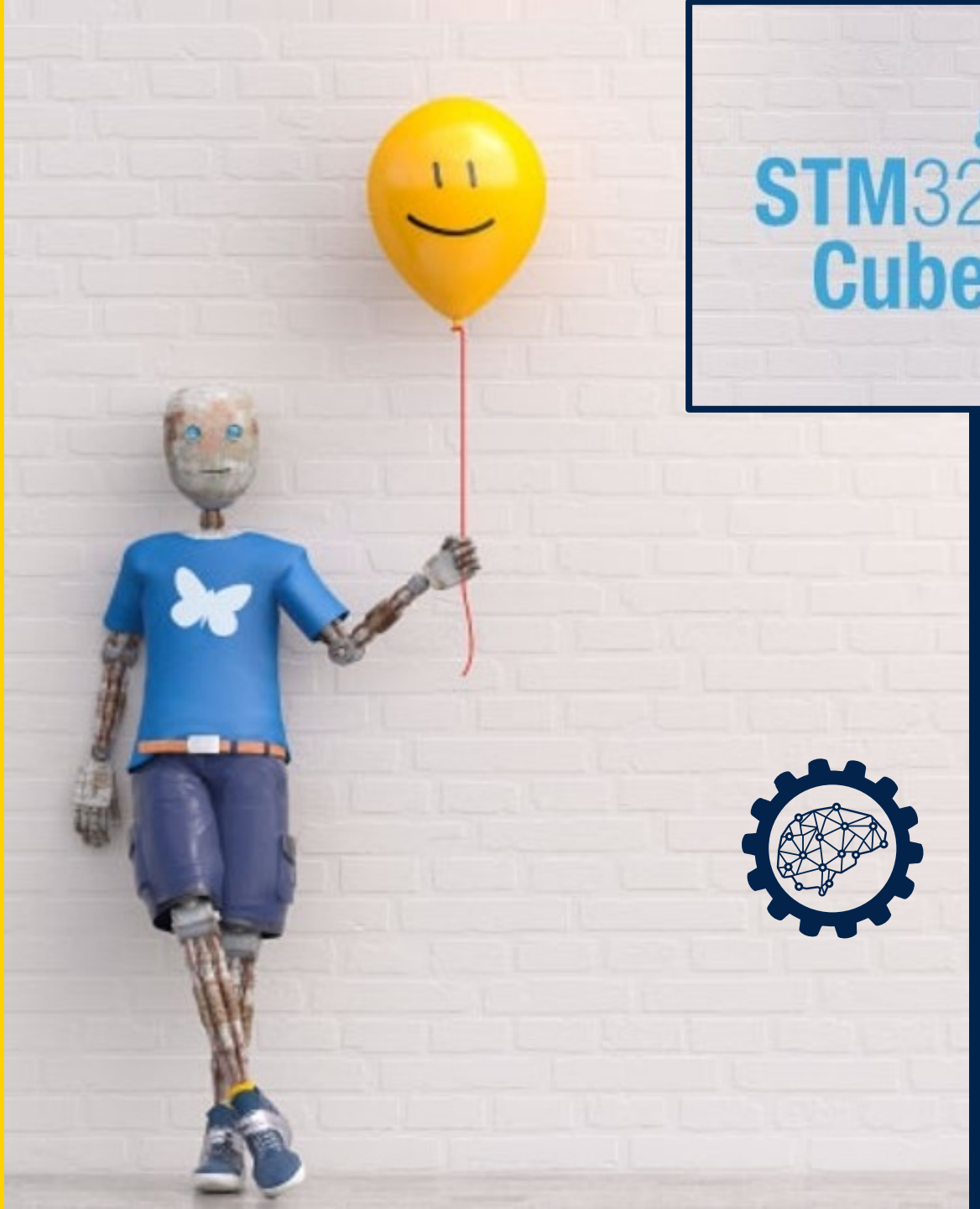


STM32 Cube.AI

life.augmented

Methodology, tools and perspectives to achieve neural networks fast deployment on STM32:

Image classifier case study using Nucleo IoT node

Danilo Pau
Technical Director, IEEE & ST Fellow
System Research and Applications

Dec 9th, 2021

Key steps for Supervised Deep Learning

Neural Network (NN) Model Creation | Operating Mode

1

2

Clean, label data
Build NN topology

Conceive &
Train NN Model

3

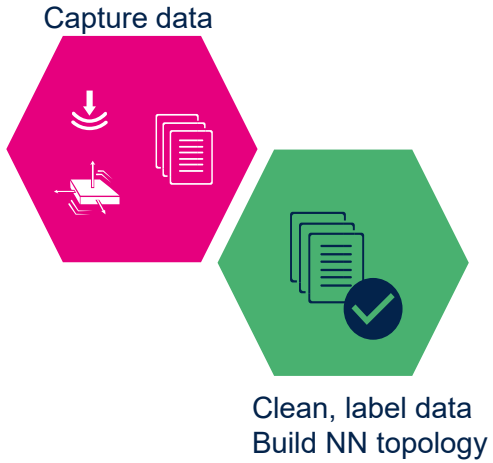
4

Convert NN into
optimized code for MCU

Deploy application
on the field

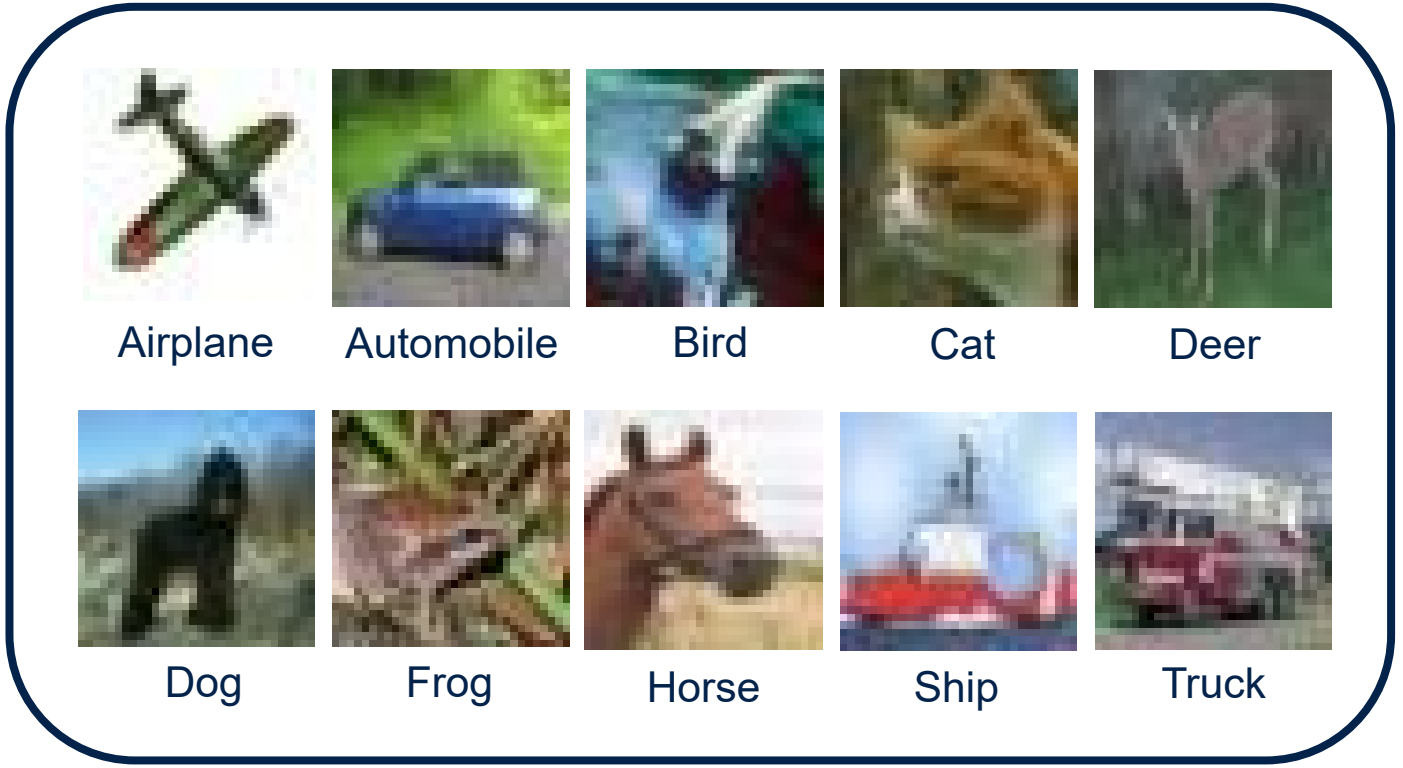
5

life.augmented



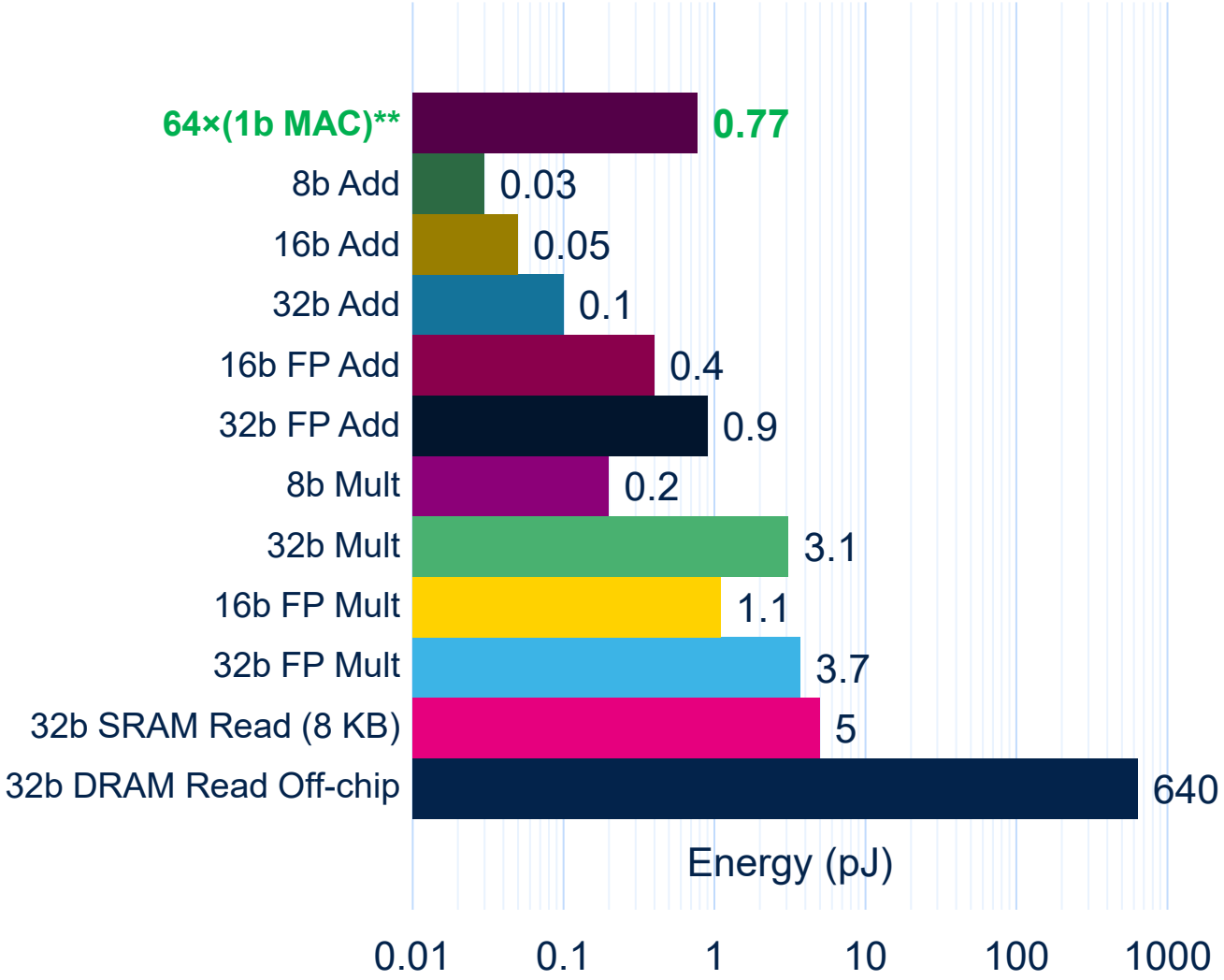
CIFAR-10 dataset

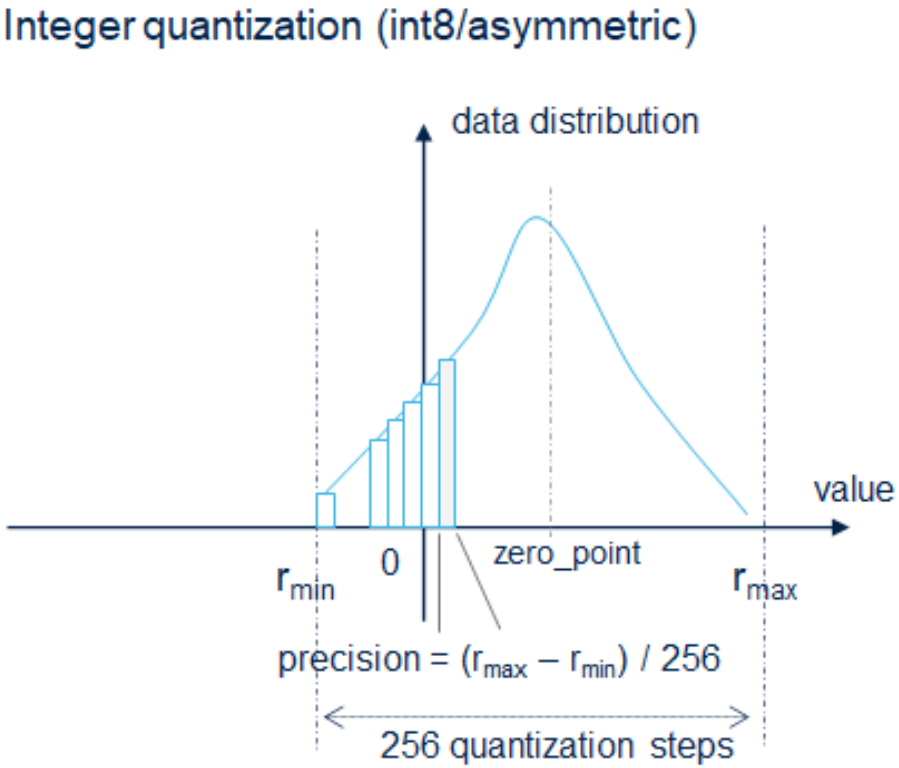
- **CIFAR-10** is a labeled subset of a 80 million tiny images dataset
- **60,000 8-bit color images**
  - **50,000 training** images
  - **10,000 test** images
- **32x32** pixel images
- **10 classes** each with 6,000 images



life.augmented

INT8 representation





*Energy values from: Mark Horowitz, “Computing Energy Problem”, ISSCC 2014;
Area values from: Synthesis with Design Compiler under TSMC 45 nm std_cells
**Agrawal et al., “Xcel-RAM: Accelerating Binary Neural Networks in High-Throughput SRAM Compute Arrays”, 2018

$$r = (q - zero_point) \times scale \qquad q = \frac{r}{scale} + zero_point$$

life.augmented

Most adopted AI development flow

TensorFlow Model

32-bit float numbers

Train

TensorFlow Model
(.pb file)

Export

Frozen Graph
(.pb file)

Convert

TensorFlow Lite
(.tflite file)

8-bit int 32-bit float

FP32/INT8

INT8 Quantization aware training

Check-points

or

Keras Model

32-bit float numbers

Train

Keras Model
(.h5 file)
32-bit float numbers

Keras to TensorFlow Lite converter

TensorFlow Lite converter

Post training INT8 quantization

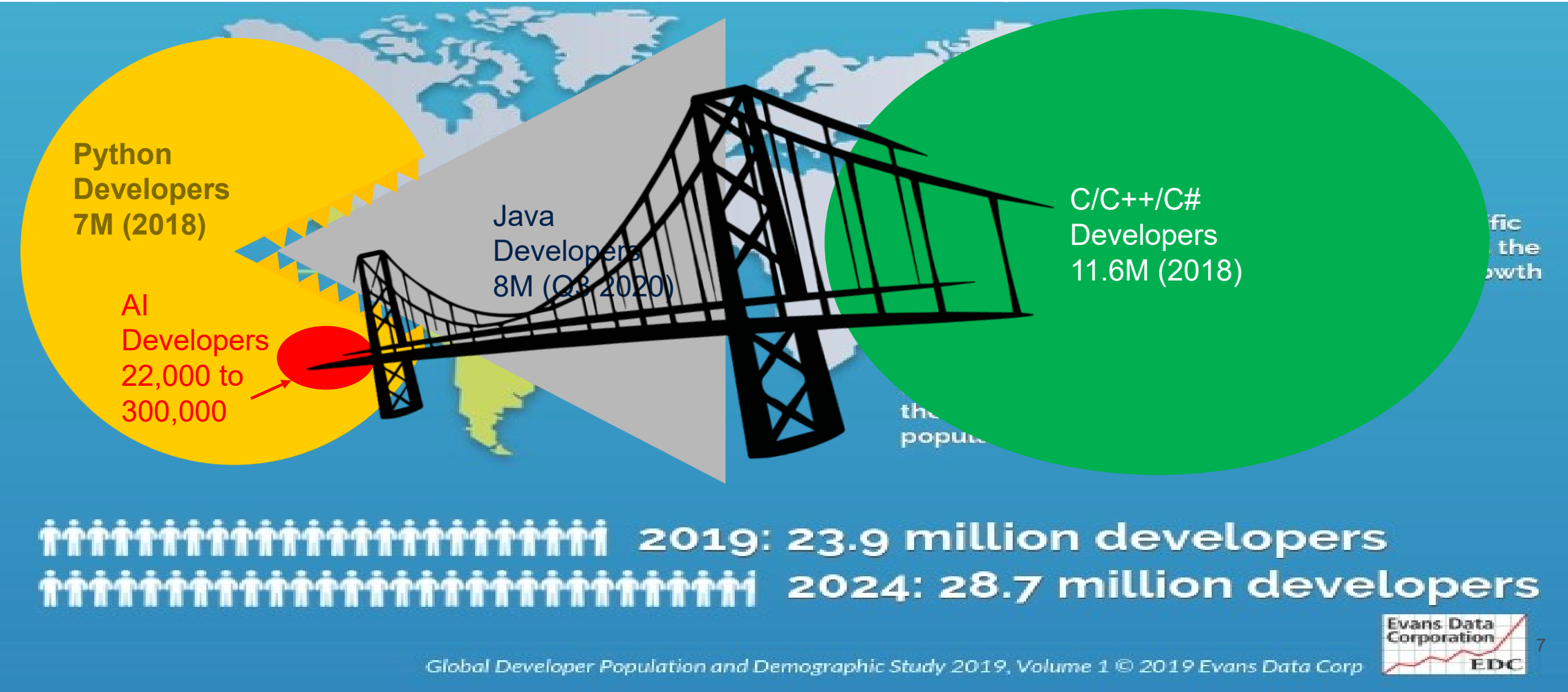
Conceive &
Train NN Model

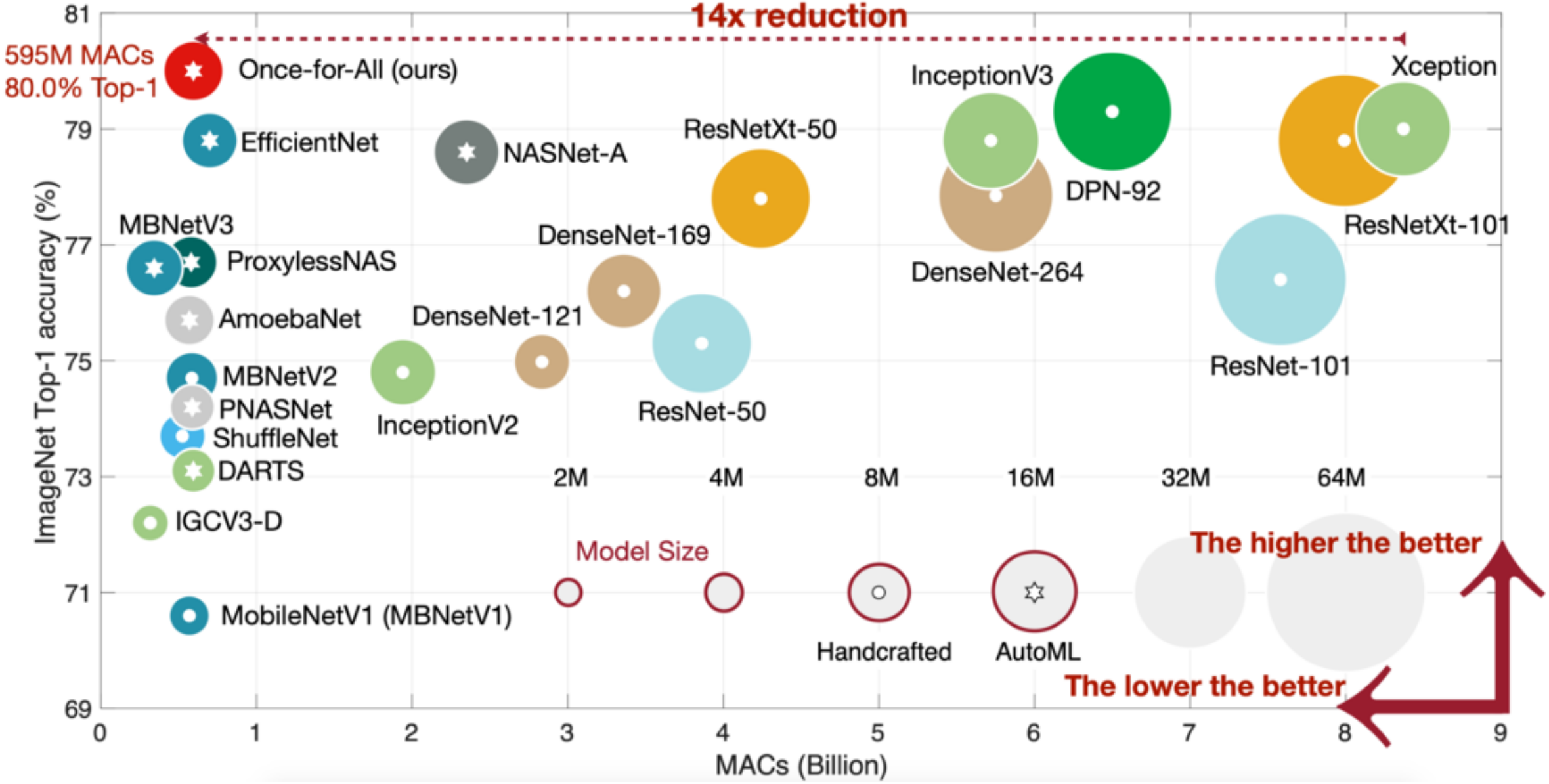
How to productively deploy pre trained Neural Networks on off-the shelf micro controllers and sensors ?

3

life.augmented

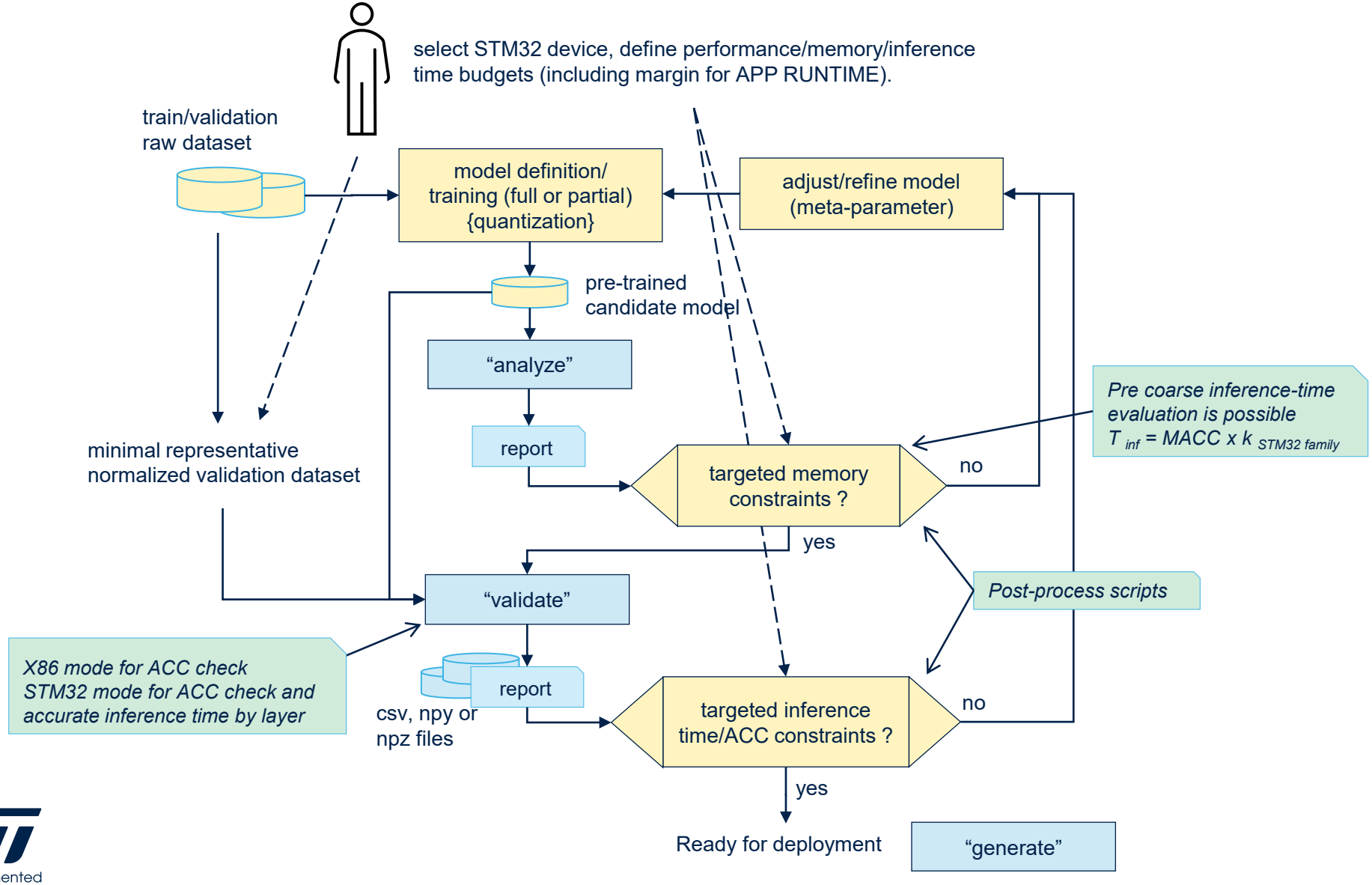
How to bridge the AI and embedded communities?



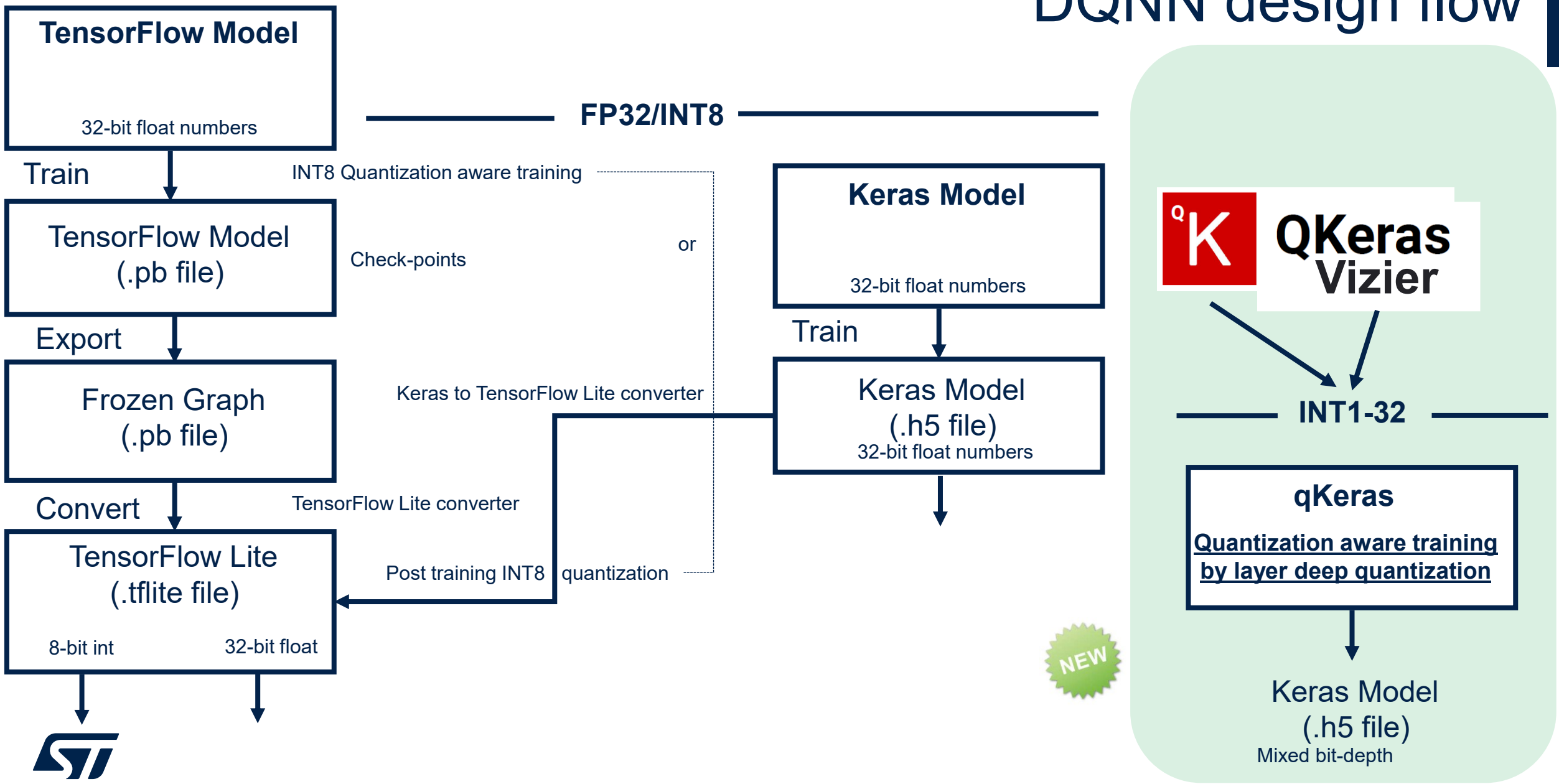


https://medium.com/@culurciello/analysis-of-deep-neural-networks-dcf398e71aae

NAS flow to support MCU deployability



DQNN design flow



life.augmented

Founding requirements

Interoperability
Pre-trained Neural Network models
Deep Learning framework dependent

Everybody else

ONNX
Caffe2 PyTorch
LibSVM MATLAB Tengine Chainer
mxnet PaddlePaddle
scikit learn Cognitive Toolkit dmlc XGBoost

Google

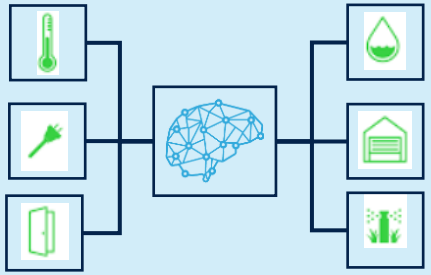
Keras
TensorFlow Lite

Sensors and OS Agnostic

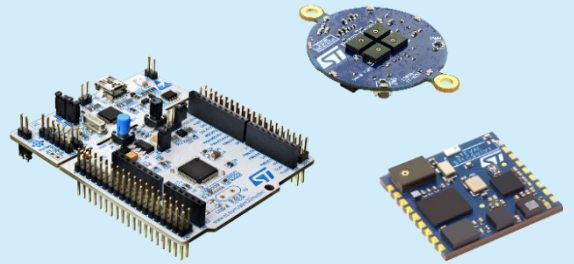
RTOS

STM32 Cube.AI
SPC5 Studio.AI

AI Hub
Multiple Neural Networks



STM32 Deep Learning SW Solution



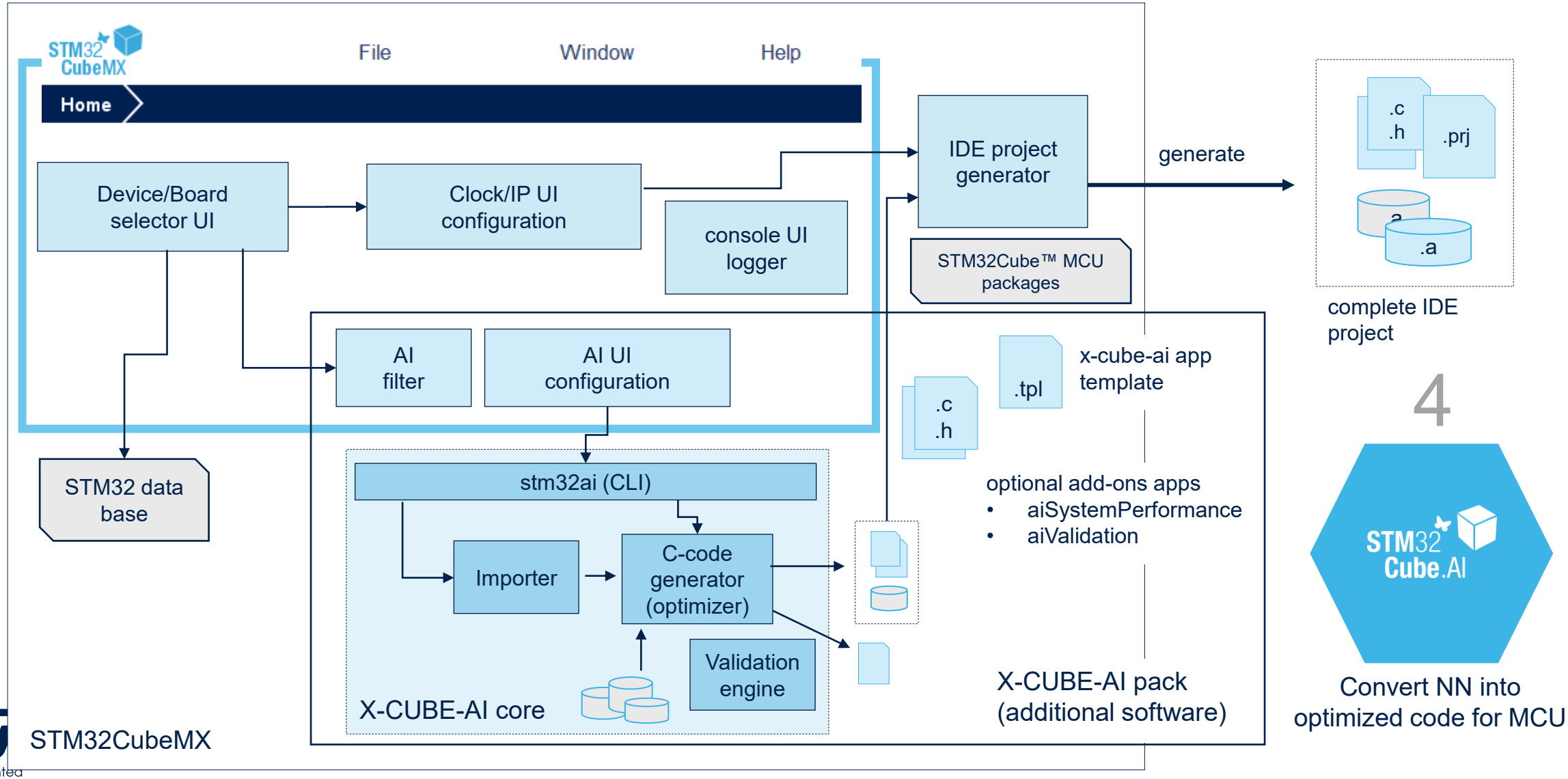
Choose your IDE
Compiler and Debugger
Framework Independent

IAR SYSTEMS, arm KEIL, ac6, atollic, ARDUINO
arm MBED, TASKING, SEGGER, GCC, SysProgs
HI-TECH C COMPILERS, CooCox, SYSTEM, emPROG, Green Hills SOFTWARE

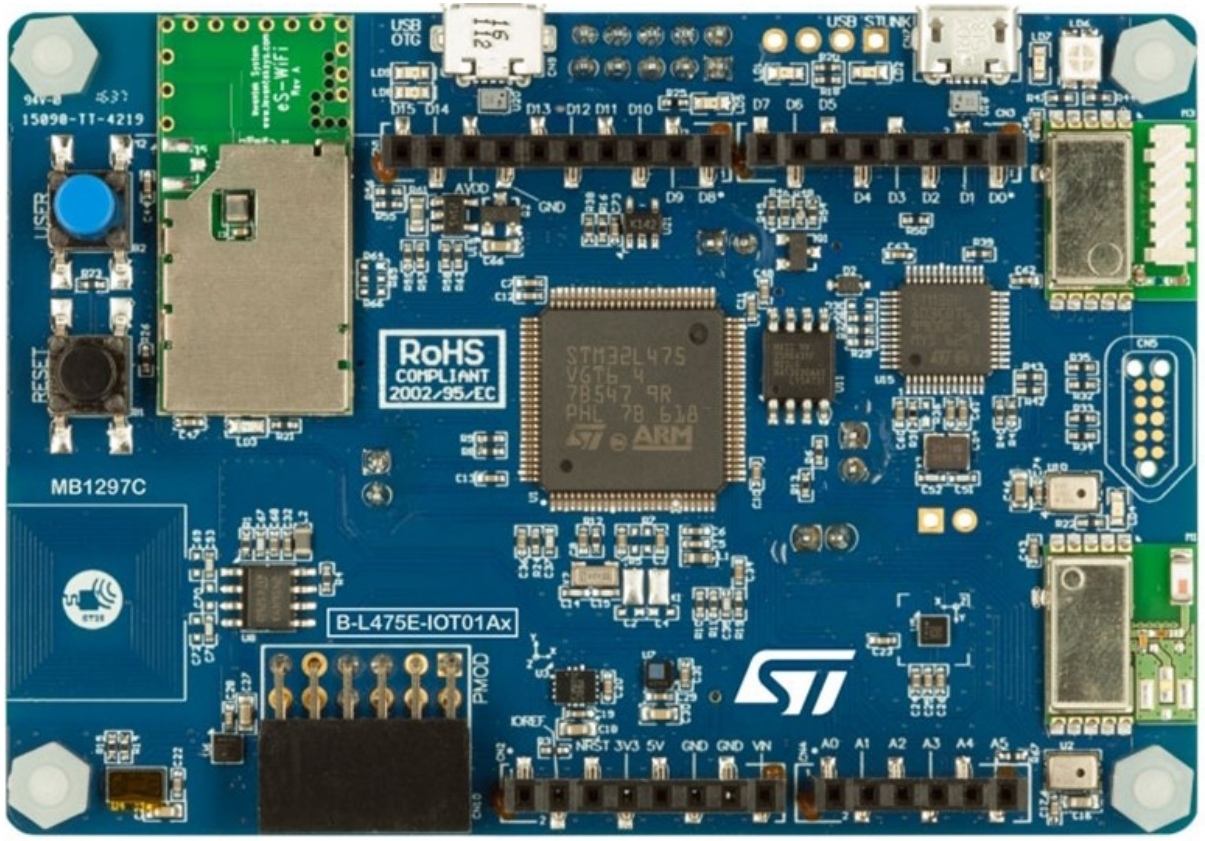
SPC5

life.augmented

X-CUBE-AI package
a STM32CubeMX additional sw



STM32L4 Discovery kit IoT node



- Arm® Cortex®-M4 core with 1 Mbyte of Flash memory and 128 Kbytes
- 64-Mbit Quad-SPI Flash memory
- Bluetooth® V4.1 module
- Sub-GHz (868 MHz or 915 MHz) low-power-programmable RF module
- 802.11 b/g/n compliant Wi-Fi® module
- Dynamic NFC tag based on M24SR with its printed NFC antenna
- 2 digital omnidirectional microphones
- Capacitive digital sensor for relative humidity and temperature
- High-performance 3-axis magnetometer
- 3D accelerometer and 3D gyroscope
- 260-1260 hPa absolute digital output barometer
- Time-of-Flight and gesture-detection sensor

Q&A

Thank you

danilo.pau@st.com

© STMicroelectronics - All rights reserved.
ST logo is a trademark or a registered trademark of STMicroelectronics International NV or its affiliates in the EU and/or other countries.
For additional information about ST trademarks, please refer to www.st.com/trademarks.
All other product or service names are the property of their respective owners.

life.augmented