tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Better productivity leveraging AI community driven interoperability"

Danilo Pietro Pau- STMicroelectronics

July 20, 2021

TINY ML

www.tinyML.org

TINY ML TALKS webcast

tinyML Talks Sponsors

arm

tinyML Strategic Partner

Deeplite

EDGE IMPULSE

tinyML Strategic Partner

emza visual sense

tinyML Strategic Partner

GREENWAVES TECHNOLOGIES

tinyML Strategic Partner

LatentAI Adaptive AI for a Smarter Edge

tinyML Strategic Partner

HOTG

tinyML Strategic Partner

maxim integrated™

Qeexo

Qualcomm

tinyML Strategic Partner

Reality AI®

tinyML Strategic Partner

SensiML™

tinyML Strategic Partner

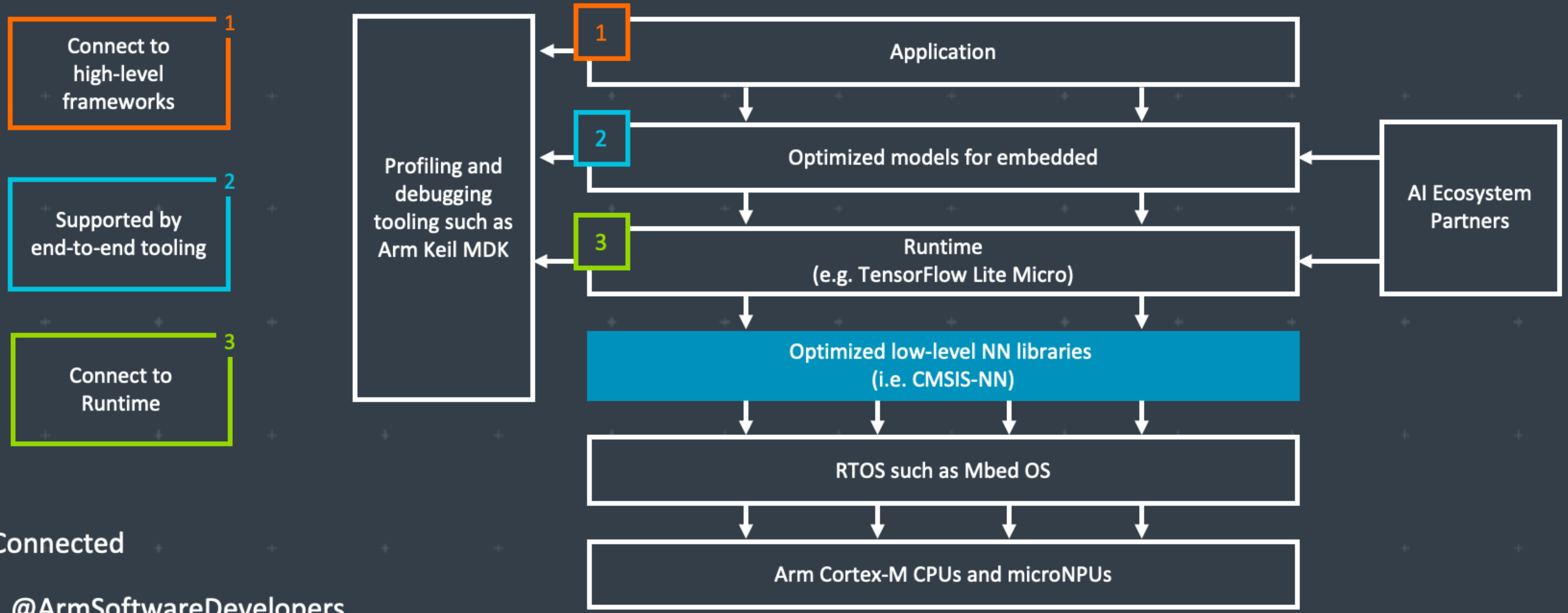
SynSense

SYNTIANT

tinyML Strategic Partner

Additional Sponsorships available – contact Olga@tinyML.org for info

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

arm

Deeplite

WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

Automatically compress SOTA models like MobileNet to <200KB with **little to no drop in accuracy** for inference on resource-limited MCUs

Reduce model optimization trial & error from weeks to days using Deeplite's **design space exploration**

APPLY NOW

Deploy more models to your device without sacrificing performance or battery life with our **easy-to-use software**

BECOME BETA USER **bit.ly/testdeeplite**

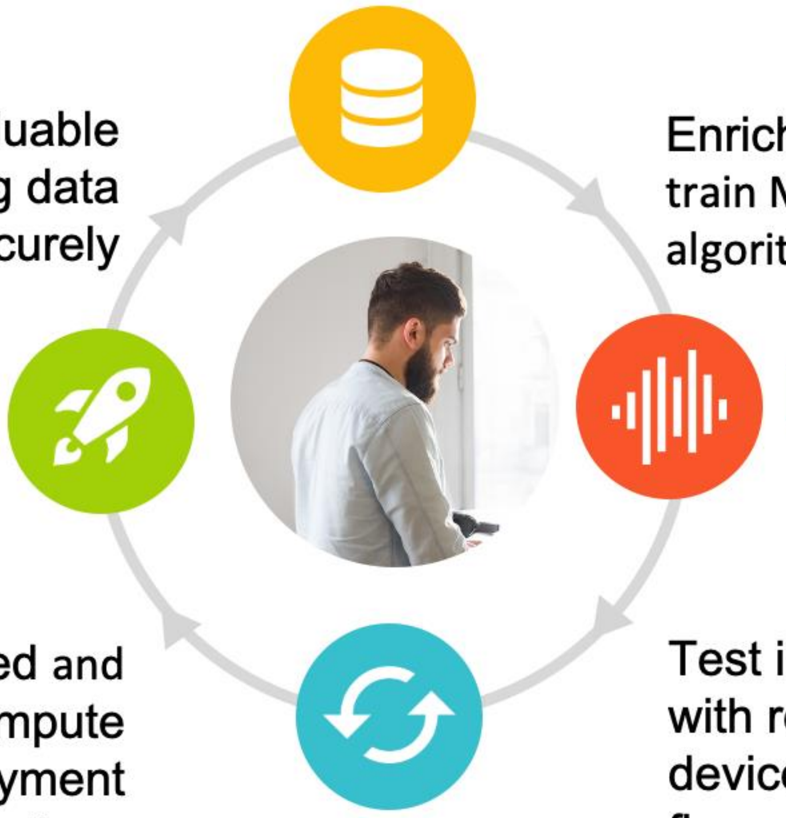
mobilityXlab arm AI 100 CBINSIGHTS

TinyML for all developers

C++ library

Arduino library

WebAssembly



Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

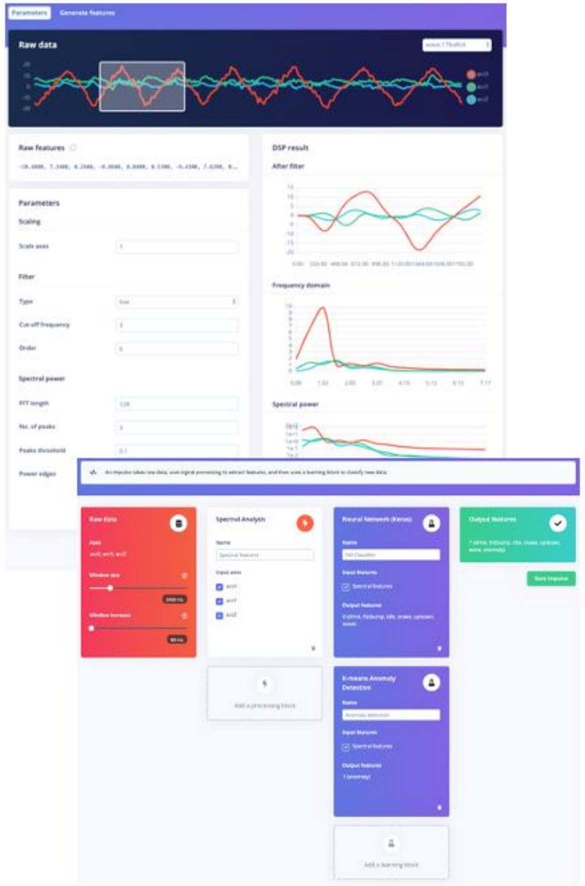
Open source SDK

Impulse

Embedded and edge compute deployment options

Test

Test impulse with real-time device data flows

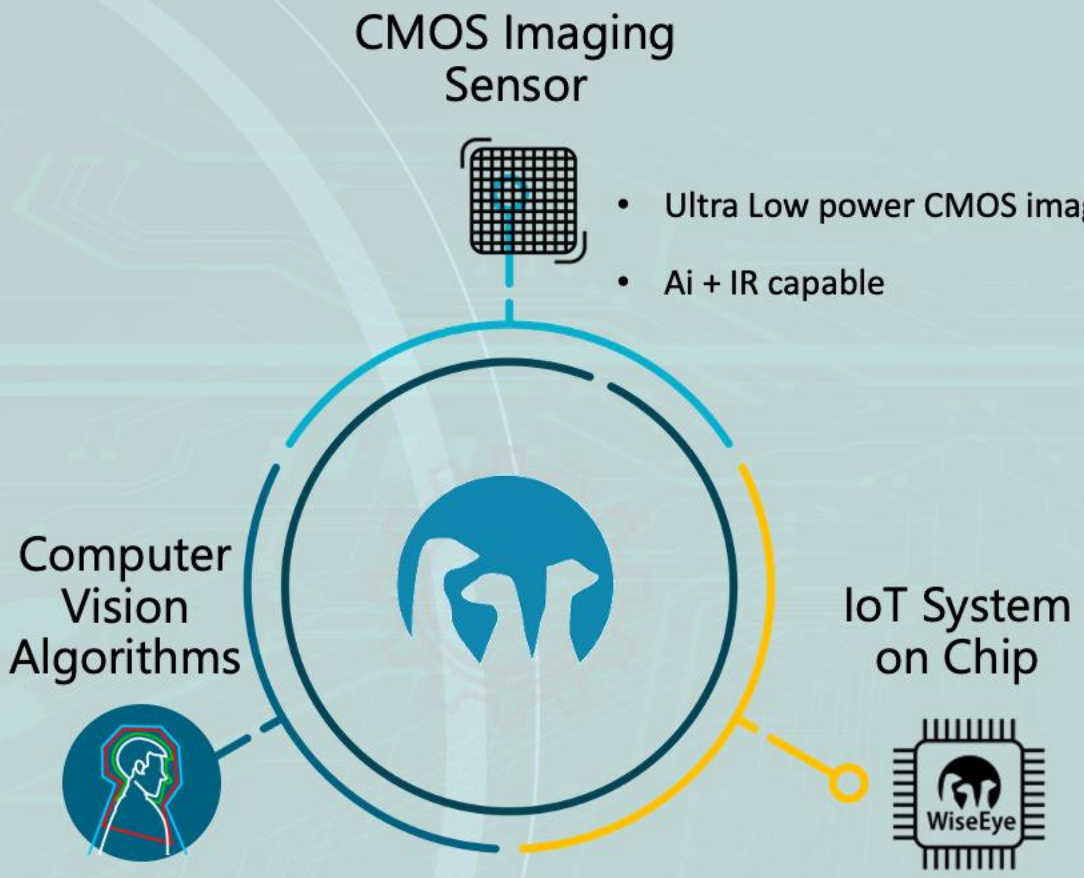


www.edgeimpulse.com

emza
visual sense

The Eye in IoT

Edge AI Visual Sensors



CMOS Imaging Sensor

- Ultra Low power CMOS imager
- Ai + IR capable

Computer Vision Algorithms

- Machine Learning algorithm
- <1MB memory footprint
- Microcontrollers computing power
- Trained algorithm
- Processing of low-res images
- Human detection and other classifiers

IoT System on Chip

- Machine Learning edge computing silicon
- <1mW always-on power consumption
- Computer Vision hardware accelerators

info@emza-vs.com

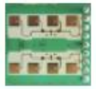
Enabling the next generation of Sensor and Hearable products to process rich data with energy efficiency

Visible Image

Sound

IR Image

Radar



Bio-sensor

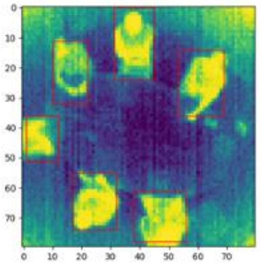
Gyro/Accel

GREENWAVES TECHNOLOGIES

Wearables / Hearables

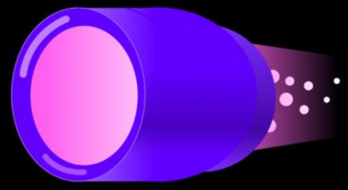
Battery-powered consumer electronics

IoT Sensors

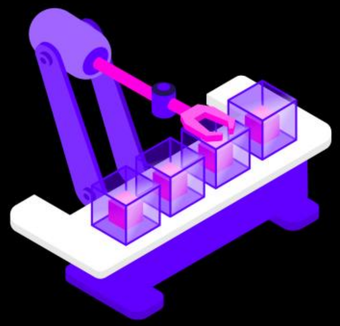


Distributed infrastructure for TinyML apps

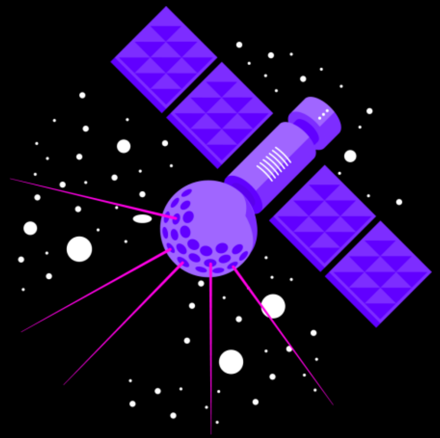
HOTG
Decoupling intelligence



Develop at warp speed



Automate deployments



Device orchestration

HOTG is building the distributed infrastructure to pave the way for AI enabled edge applications

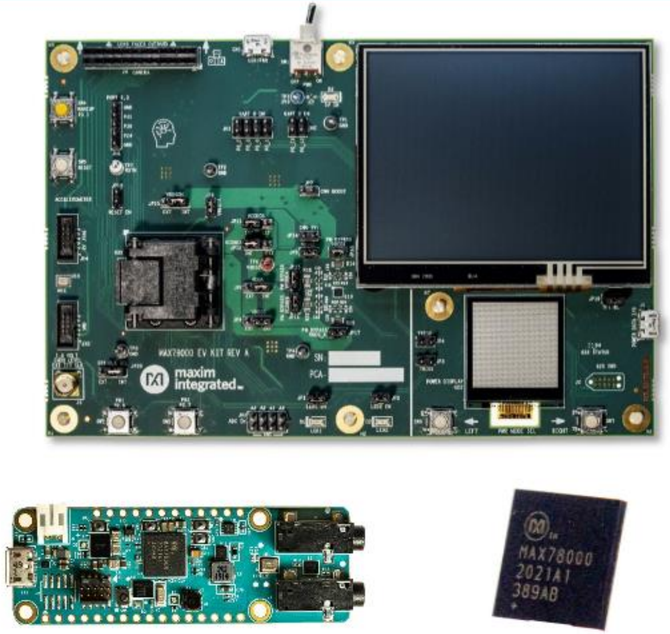
LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

Maxim Integrated: Enabling Edge Intelligence

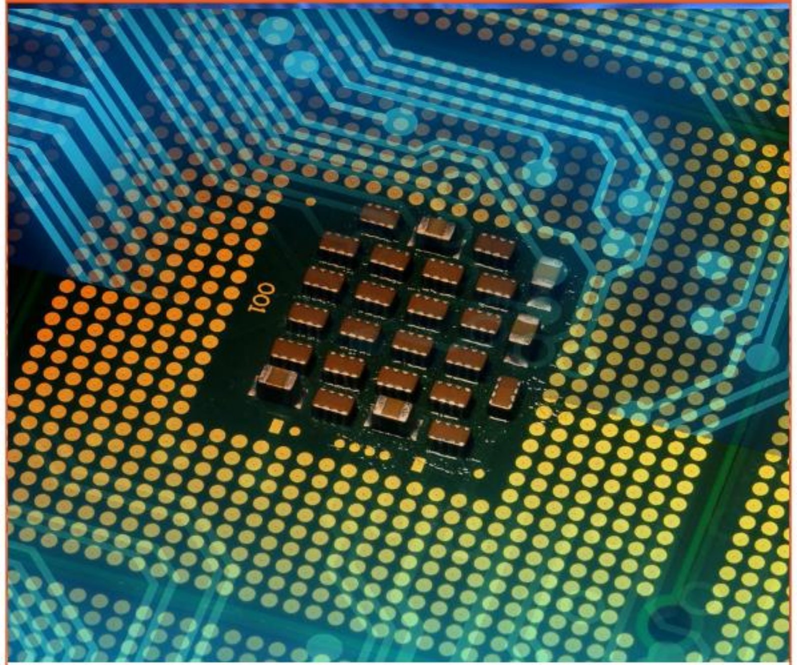
Advanced AI Acceleration IC



The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

www.maximintegrated.com/MAX78000

Low Power Cortex M4 Micros



Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

www.maximintegrated.com/microcontrollers

Sensors and Signal Conditioning

Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

www.maximintegrated.com/sensors

Qeexo AutoML

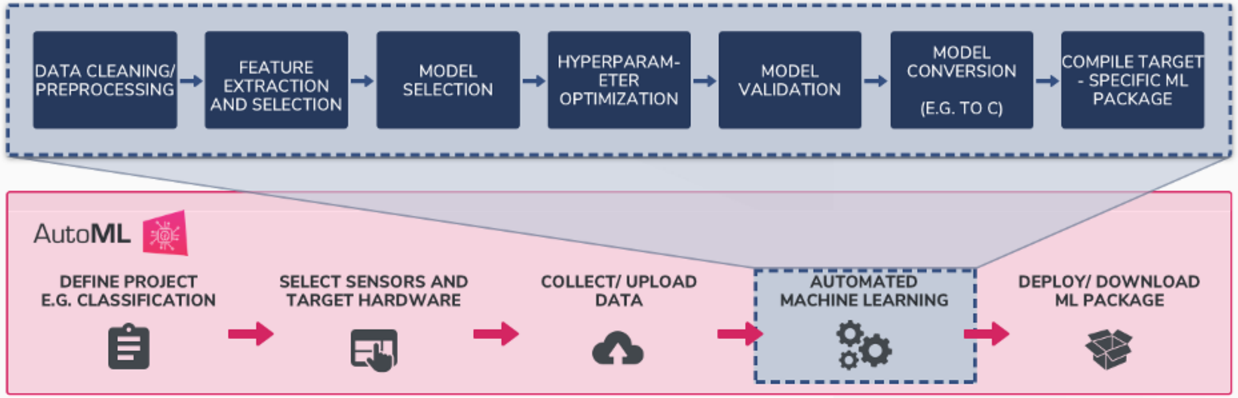
Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

CES INNOVATION AWARDS 2021 HONOREE

Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

End-to-End Machine Learning Platform



For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai info@reality.ai @SensorAI Reality AI

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions

- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars "see with sound"

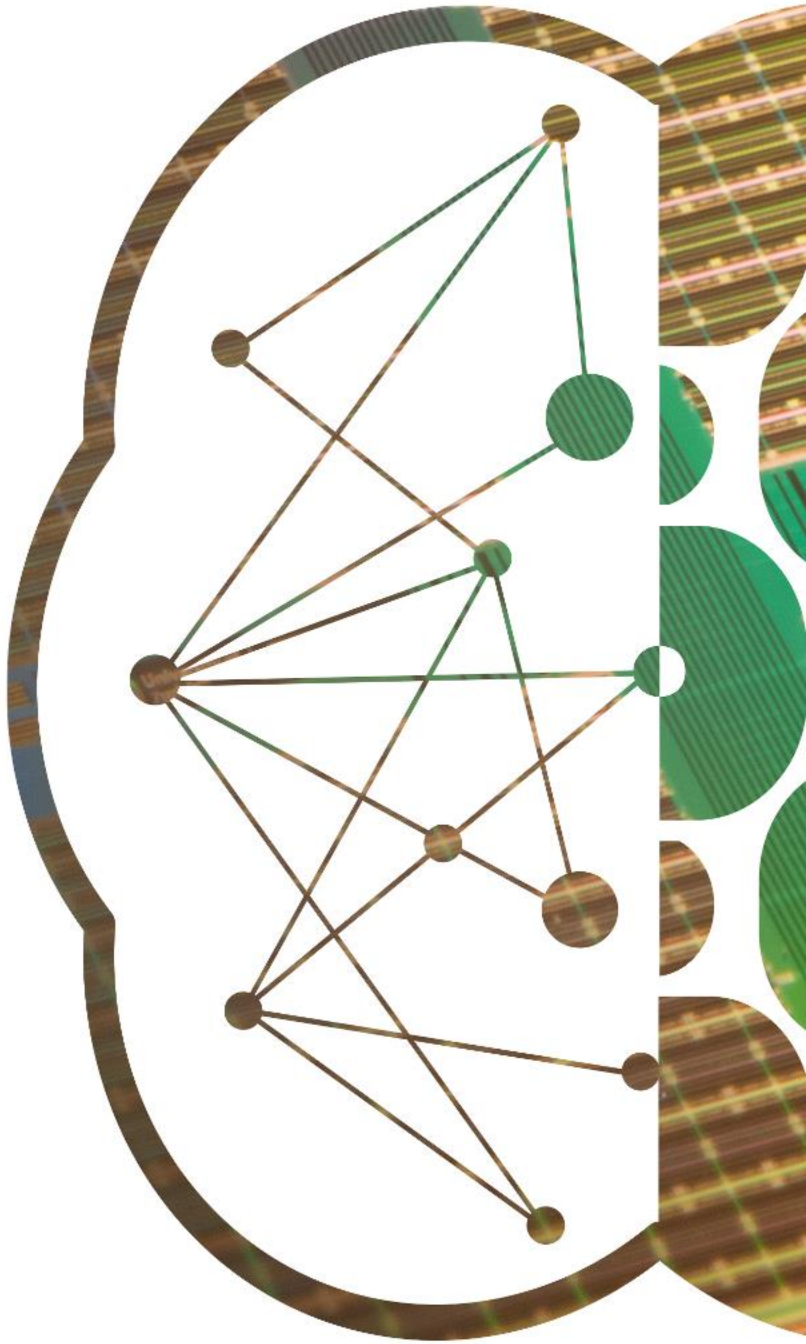
Reality AI Tools® software

- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

SynSense

SynSense builds **sensing and inference** hardware for **ultra-low-power** (sub-mW) **embedded, mobile and edge** devices. We design systems for **real-time always-on smart sensing**, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai



SYNTIANT

Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com

@Syntiantcorp

TINY ML TINY ML FOUNDATION collaboration with hackster.io AN AVNET COMMUNITY

Focus on:
(i) developing new use cases/apps for tinyML vision; and (ii) promoting tinyML tech & companies in the developer community

arm
EDGE IMPULSE
Himax
Qualcomm
intel

Submissions accepted until August 20th, 2021
Winners announced on September 1, 2021 ($6k value)
Sponsorships available: *sponsorships@tinyML.org*

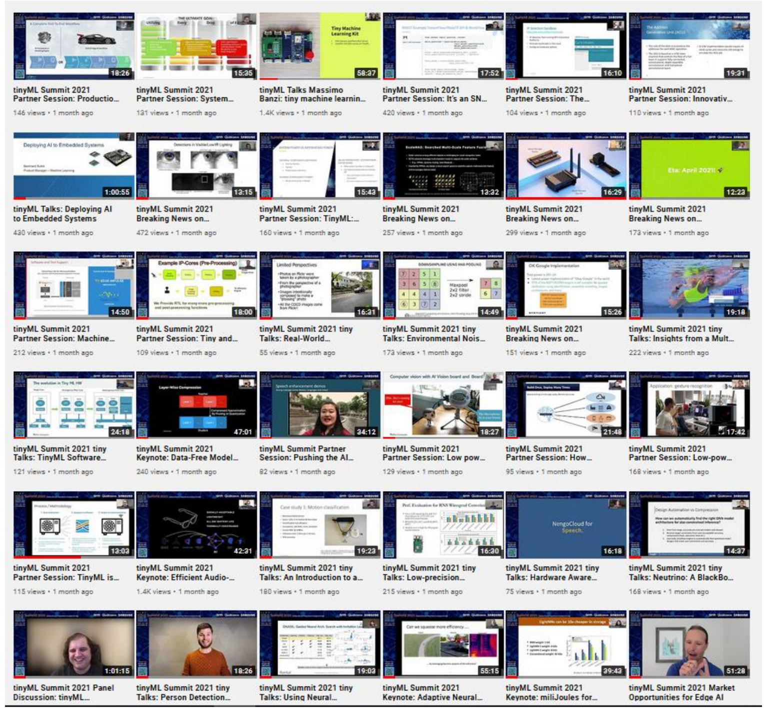
LATTICE SEMICONDUCTOR
PixArt Imaging Inc. SEE. FEEL. TOUCH.
SynSense
LUXonis

https://www.hackster.io/contests/tinyml-vision

Successful tinyML EMEA 2021

TINY ML EMEA Virtual JUNE 7-10, 2021

- Videos are available on www.youtube.com/tinyML
- **4** days of tinyML excitement
  - **2** tutorials
  - **5** keynotes
  - **15** tinyTalks
  - **7** lightning talks
  - **3** panel discussions & networking
  - **16** papers in the Student Forum
  - **4** partner sessions
  - **16** sponsoring companies
- **58** speakers, **1687** registered attendees!



250 videos with 121k views as of July 10, 2021

YouTube

Next tinyML Talks

Date	Presenter	Topic / Title
Tuesday, August 3	**Vikram Shrivastava,** Sr. Director, IoT Marketing, Knowles Corporate	Dedicated Audio Processors at the Edge are the Future of AI

Webcast start time is 8 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting

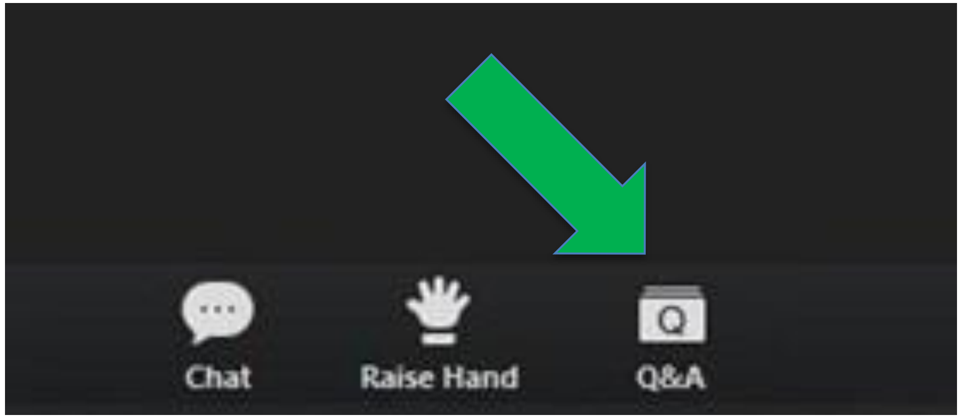
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions



Danilo Pau

One year before graduating from the Polytechnic University of Milan in 1992, Danilo PAU joined STMicroelectronics, where he worked on HDMAC and MPEG2 video memory reduction, video coding, embedded graphics, and computer vision. Today, his work focuses on developing solutions for deep learning tools and applications. Since 2019 Danilo has been an IEEE Fellow. Currently serves as a member of IEEE Region 8 Action for Industry and Member of the Machine Learning, Deep Learning and AI in the CE (MDA) Technical Stream Committee IEEE Consumer Electronics Society (CESoc). With over 80 patents, 104 publications, 113 MPEG authored documents and 39 invited talks/seminars at various worldwide Universities and Conferences, Danilo's favorite activity remains mentoring undergraduate students, MSc engineers and PhD students from various universities.

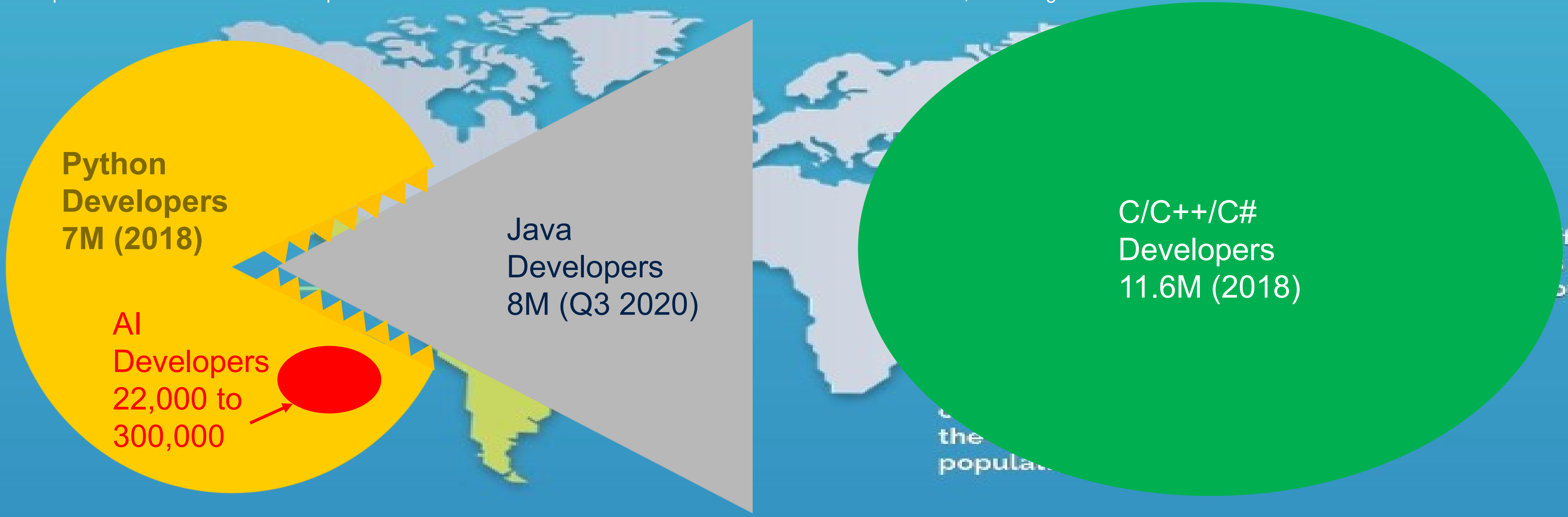
Better productivity leveraging AI community driven interoperability

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

July 20, 2021

Global Developer Population and Demographic Study 2019, Vol 1

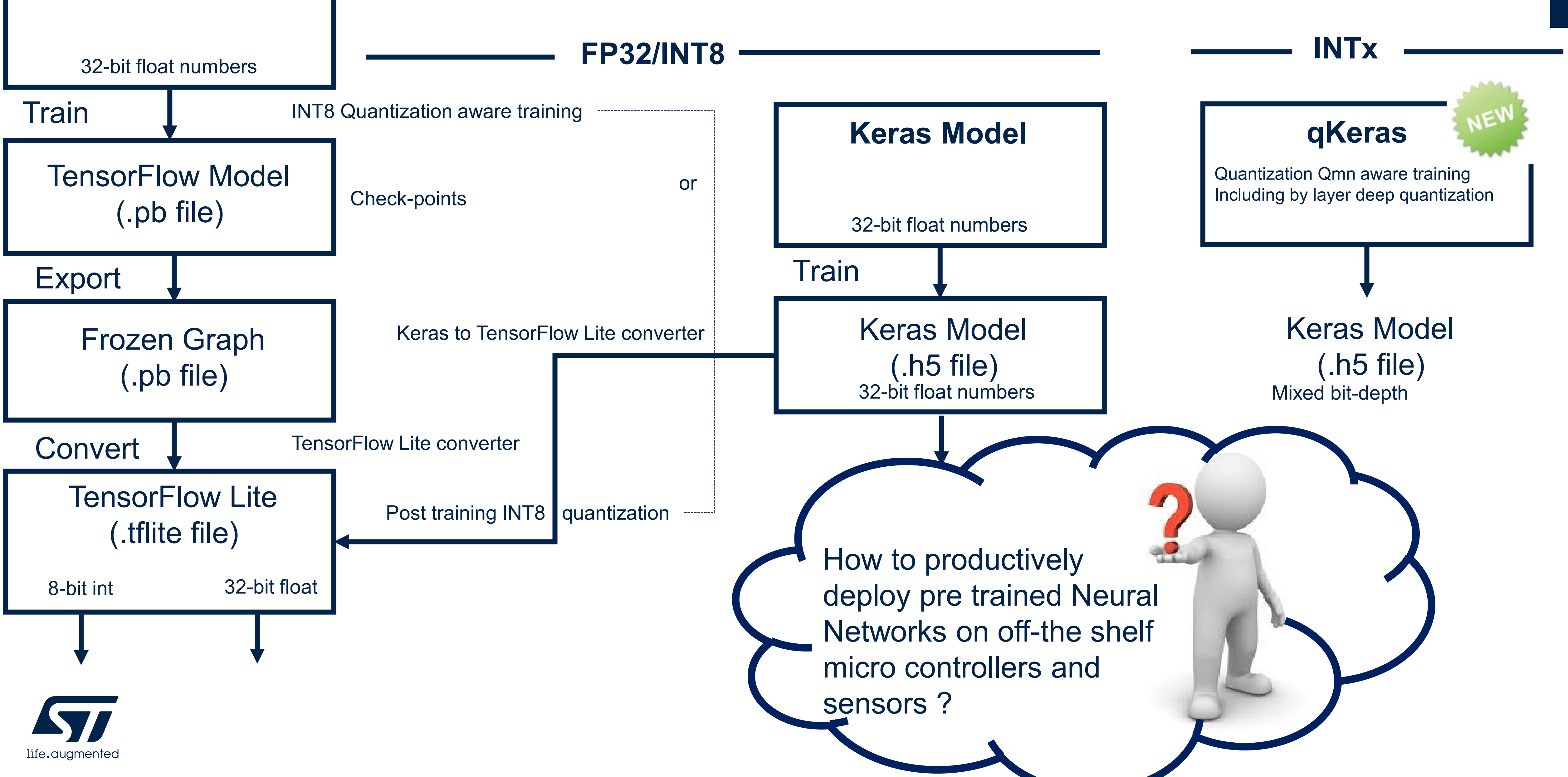
Source https://www.daxx.com/blog/development-trends/number-software-developers-world
*https://www.stateofai2019.com/chapter-6-the-war-for-talent/#:~:text=Estimates%20of%20the%20number%20of,AI%20originated%20in%20academia.



2019: 23.9 million developers
2024: 28.7 million developers

Evans Data Corporation EDC

Most adopted AI development flow



life.augmented

Conditions for Deep Learning on MCUs

0 $$\text{Accuracy}_{\text{neural networks}} > \text{Accuracy}_{\substack{\text{traditional machine learning} \\ or\ hand-crafted\ algorithm}}$$

1 $$\text{RAM}_{\text{MCU}} > \text{Activations}_{\text{Network}}$$

2 $$\text{ROM}_{\text{MCU}} > \text{Weights}_{\text{Network}}$$

3 $$\frac{\text{NN complexity}_{\text{FLOPS}}}{\text{MCU}_{\text{FLOPS/s}}} < \frac{\text{samples in a window}}{\text{f}_{\text{sensor sampling rate/s}}}$$

"Embedded Real-Time Fall Detection with Deep Learning on Wearable Devices," *2018 21st Euromicro Conference on Digital System Design (DSD)*, Prague, 2018, pp. 405-412, doi: 10.1109/DSD.2018.00075.

life.augmented

Making AI Accessible on STM32

Leader in Arm® Cortex®-M 32-bit General Purpose MCU

MPU
- STM32**MP1** — 4158 CoreMark; 800 MHz Cortex -A7; 209 MHz Cortex -M4

High Perf MCUs
- STM32**F2** — 398 CoreMark; 120 MHz
- STM32**F4** — 608 CoreMark; 180 MHz
- STM32**H7** — 3224 CoreMark; 240 MHz Cortex -M4; 480 MHz Cortex -M7
- STM32**F7** — 1082 CoreMark; 216 MHz

Mainstream MCUs
- STM32**F0** — 106 CoreMark; 48 MHz
- STM32**G0** — 142 CoreMark; 64 MHz
- STM32**F1** — 177 CoreMark; 72 MHz
- STM32**F3** — 245 CoreMark; 72 MHz
- STM32**G4** — 550 CoreMark; 170 MHz

Ultra-low Power MCUs
- STM32**L0** — 75 CoreMark; 32 MHz
- STM32**L1** — 93 CoreMark; 32 MHz
- STM32**L5** — 424 CoreMark; 110 MHz
- STM32**U5** — 651 CoreMark; 160 MHz
- STM32**L4** — 273 CoreMark; 80 MHz
- STM32**L4+** — 409 CoreMark; 120 MHz

Wireless MCUs
- STM32**WL** — 161 CoreMark; 48 MHz
- STM32**WB** — 216 CoreMark; 64 MHz

Arm® Cortex® core: -M0 | -M0+ | -M3 | -M33 | -M4 | -M7 | dual -A7& -M4

Compatible with **Machine Learning** Partner ecosystems

Compatible with **Deep Learning** **STM32Cube.AI** ecosystem

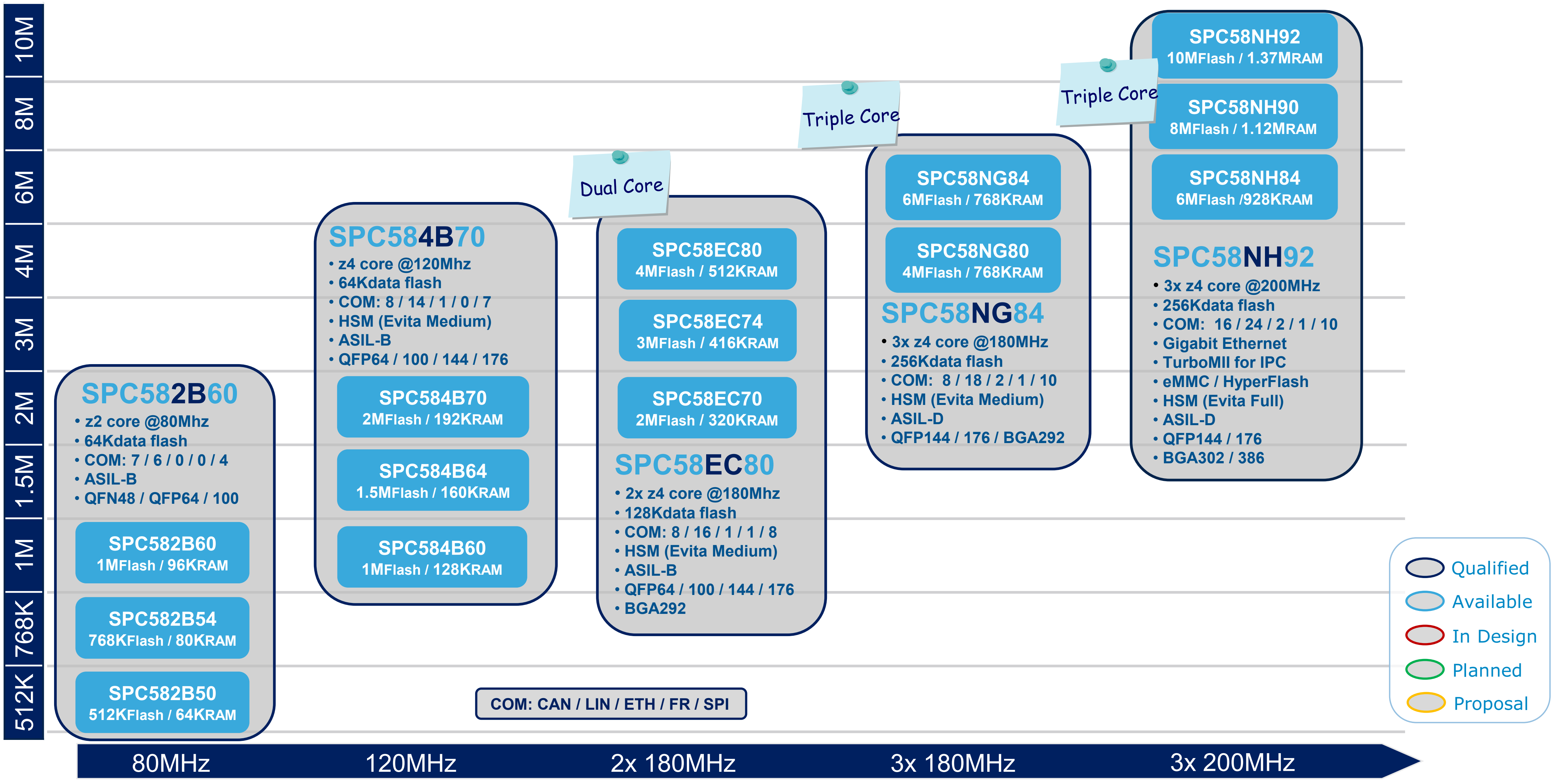
STM32 Cube.AI

More than 40,000 customers

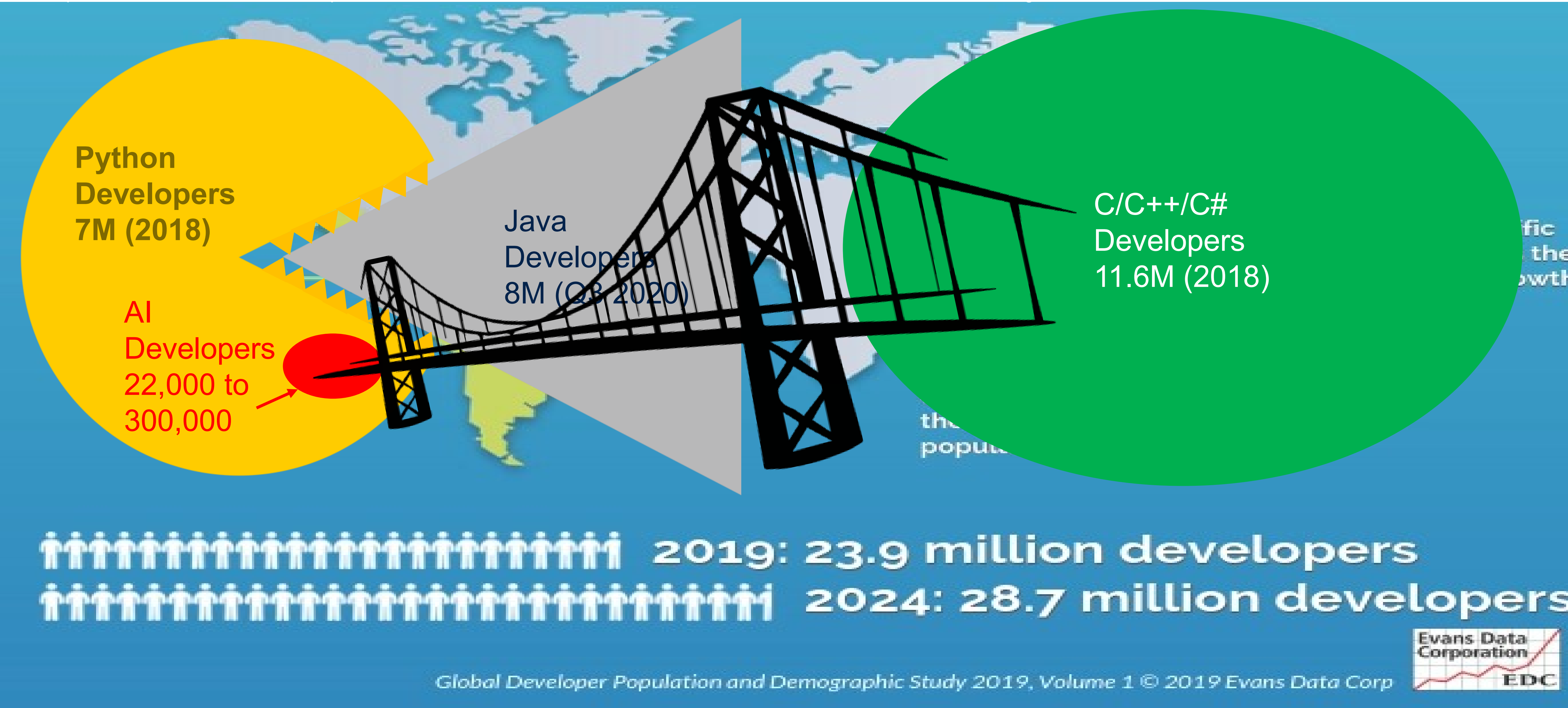
Over 4 Billion STM32 shipped since 2007

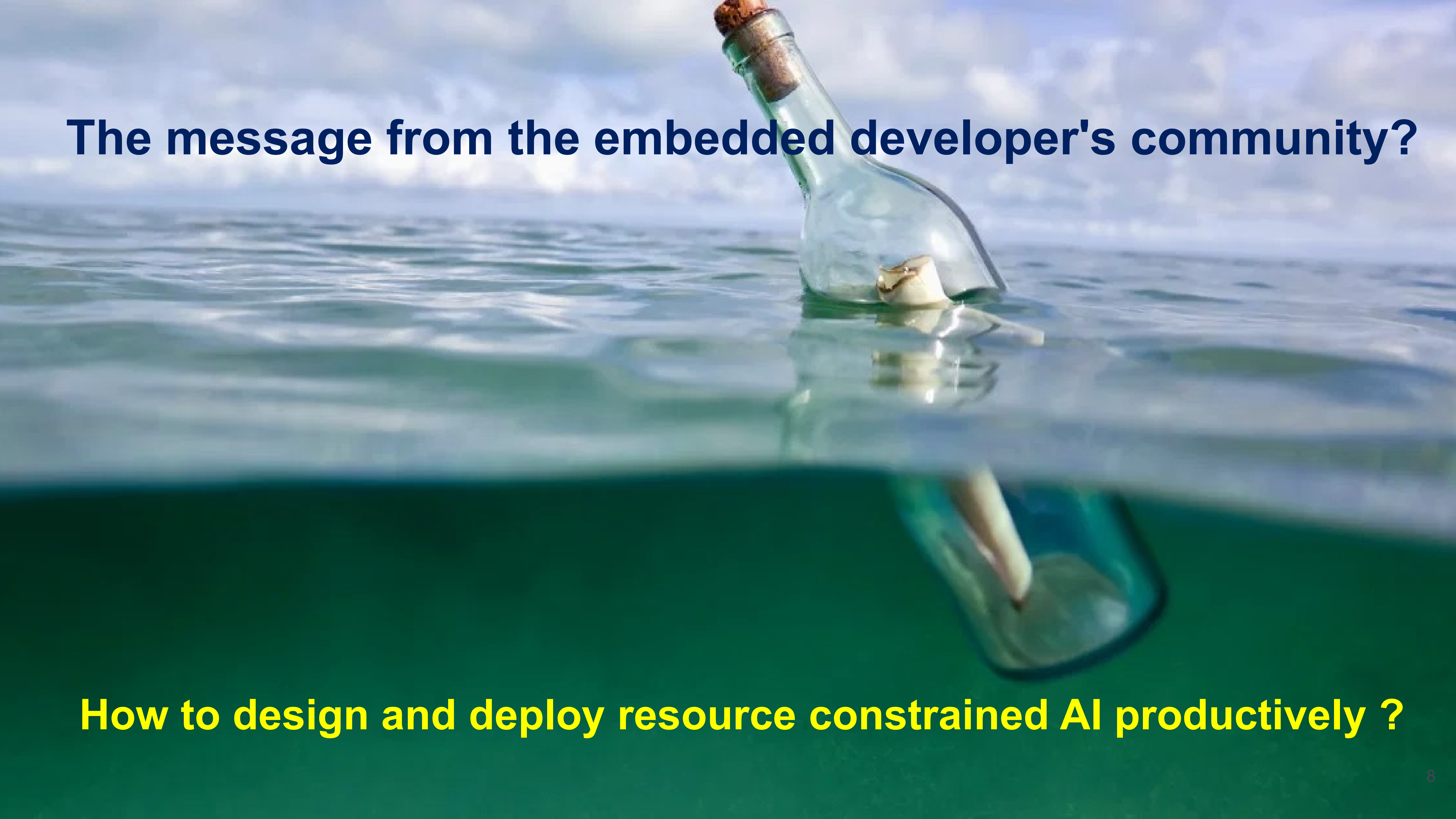
life.augmented

and on SPC58 - Chorus Family



How to bridge the AI and embedded communities?



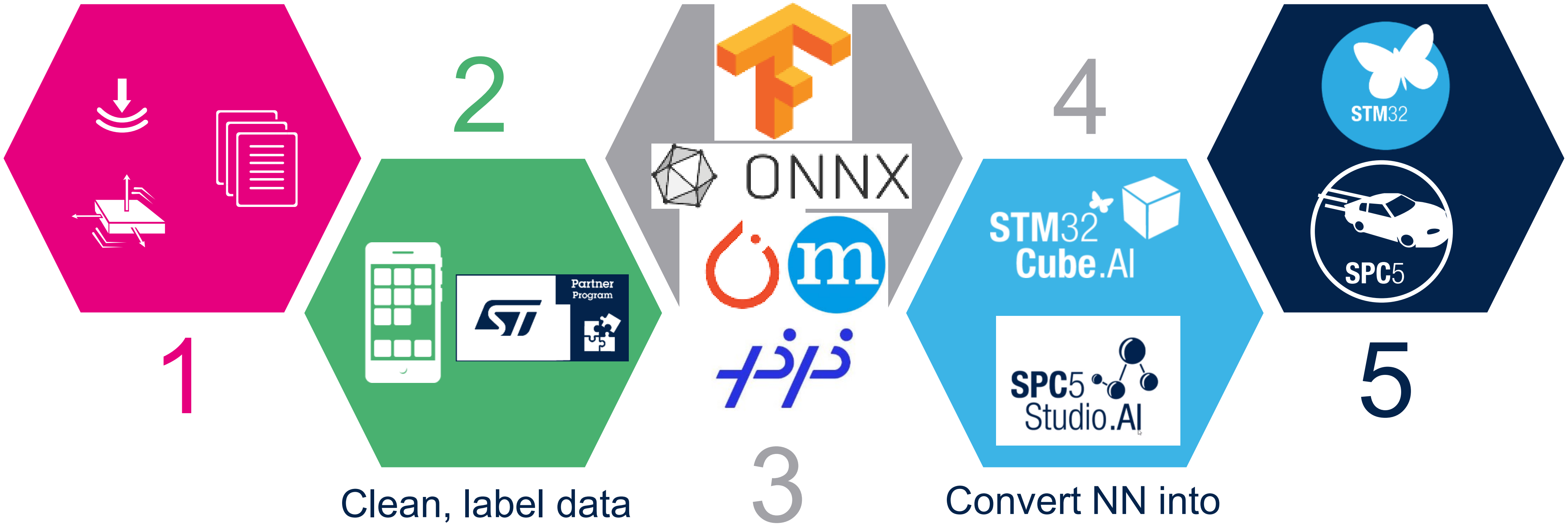


The message from the embedded developer's community?

How to design and deploy resource constrained AI productively ?

Key steps for Supervised Deep Learning

Neural Network (NN) Model Creation | Operating Mode



1

2 Clean, label data
Build NN topology

3 Conceive &
Train NN Model

4 Convert NN into
optimized code for MCU

5 Deploy application
on the field

life.augmented

Interoperability
Pre-trained Neural Network models
Deep Learning framework dependent

Interoperability met across levels !

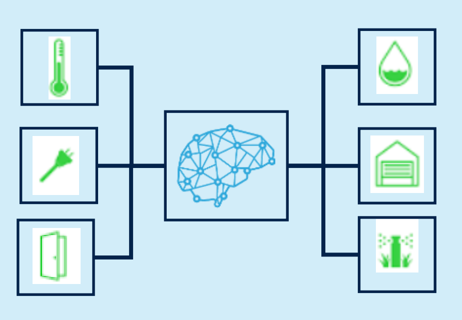
Everybody else

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

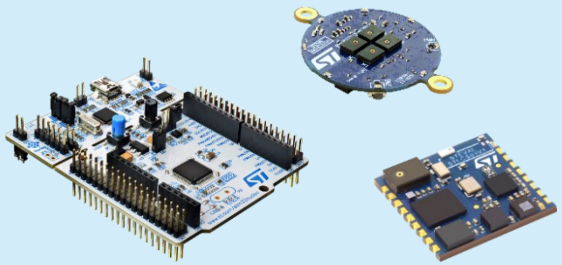
Google

Keras
TensorFlow Lite

AI Hub
Multiple Neural Networks



STM32
Deep Learning SW Solution



Sensors and OS Agnostic

RTOS

STM32 Cube.AI

SPC5 Studio.AI

SPC5

Choose your IDE
Compiler and Debugger
Framework Independent

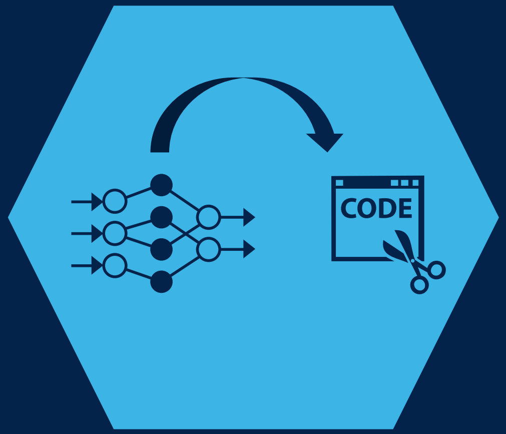
IAR SYSTEMS, arm KEIL, ac6, atollic, ARDUINO, arm MBED, TASKING, SEGGER, GCC, SYSPROGS, HI-TECH C COMPILERS, CooCox, iSYSTEM, EMPROG, Green Hills SOFTWARE

life.augmented

STM32Cube.AI

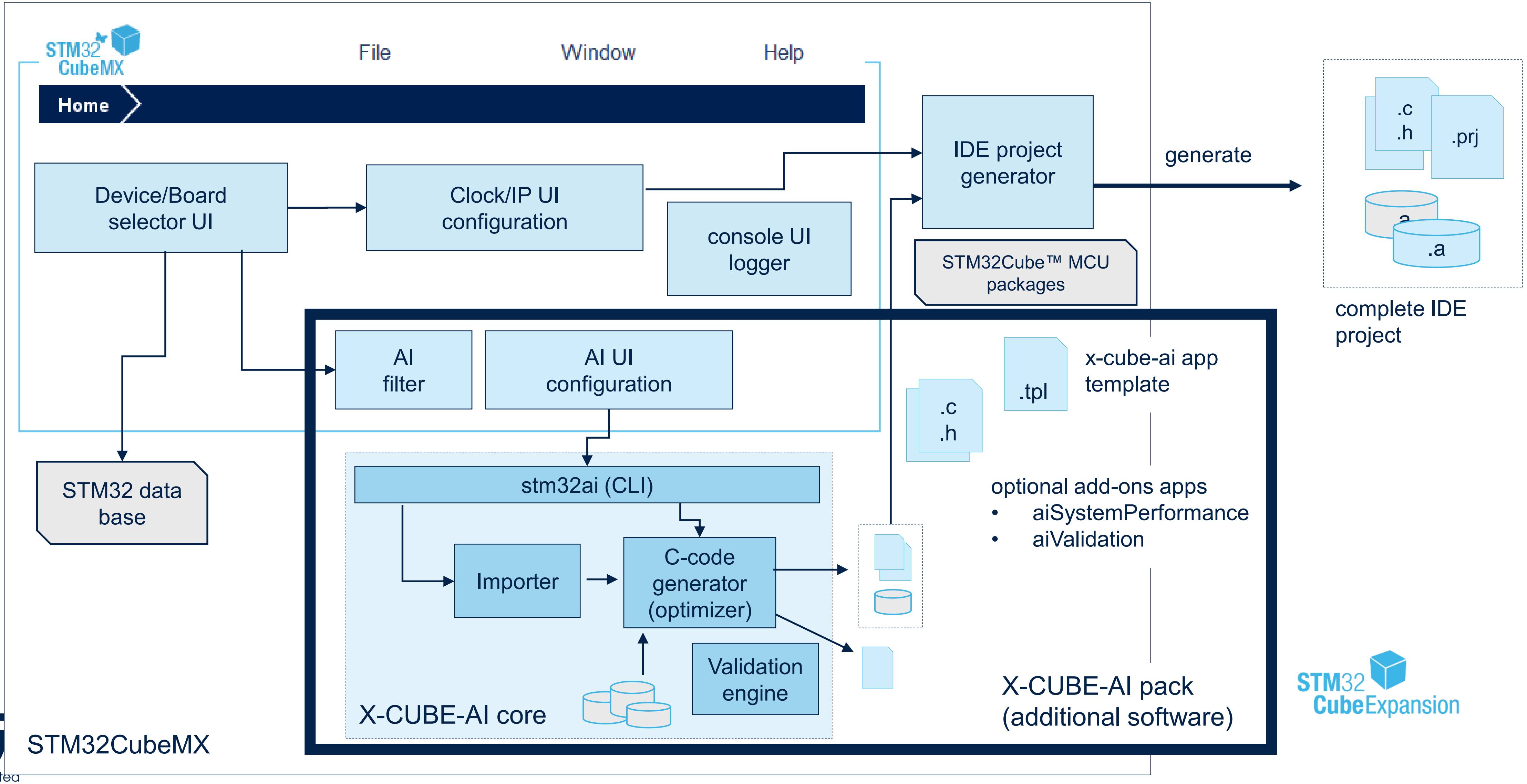
STM32 Cube.AI

4



Convert NN into optimized code for execution

X-CUBE-AI package
a STM32CubeMX additional sw



life.augmented

X-CUBE-AI user flow

Load ANN and filter MCUs

Select MCU and create .ioc project

ANN model characterization

Run Analyze

Run Validation on x86

Run Validation on MCU

Either 1-click or via IDE

Select and Run System Performance

via IDE

Generate Application Template

Integrate application specific code

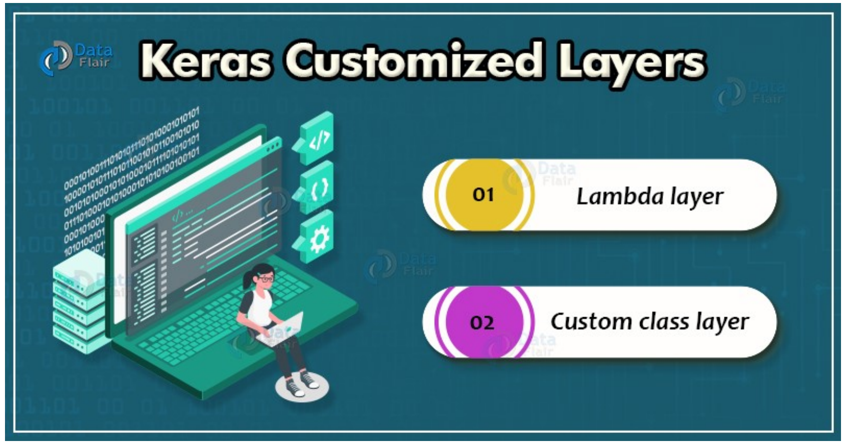
Run Field Tests

life.augmented

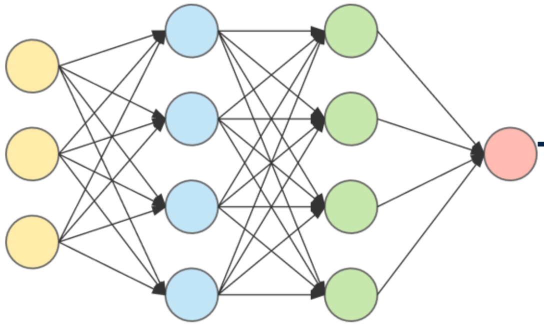
```
C:\Users\danilo pau\OneDrive - STMicroelectronics\ai>stm32ai_v700 supported-ops
Neural Network Tools for STM32AI v1.5.1 (STM.ai v7.0.0-RC8)
246 operators found
 Abs (ONNX), ABS (TFLITE), Acos (ONNX), Acosh (ONNX), Activation (KERAS), ActivityRegularization (KERAS),
 Add (KERAS), Add (ONNX), ADD (TFLITE), AlphaDropout (KERAS), And (ONNX), ARG_MAX (TFLITE), ARG_MIN (TFLITE),
 ArgMax (ONNX), ArgMin (ONNX), ArrayFeatureExtractor (ONNX), Asin (ONNX), Asinh (ONNX), Atan (ONNX),
 Atanh (ONNX), Average (KERAS), AVERAGE_POOL_2D (TFLITE), AveragePool (ONNX), AveragePooling1D (KERAS),
 AveragePooling2D (KERAS), BATCH_TO_SPACE_ND (TFLITE), BatchNormalization (KERAS), BatchNormalization (ONNX),
 Bidirectional (KERAS), Cast (ONNX), CAST (TFLITE), Ceil (ONNX), CEIL (TFLITE), Clip (ONNX), Concat (ONNX),
 Concatenate (KERAS), CONCATENATION (TFLITE), Constant (ONNX), Conv (ONNX), Conv1D (KERAS), Conv2D (KERAS),
 Conv2DTranspose (KERAS), CONV_2D (TFLITE), ConvTranspose (ONNX), Cos (ONNX), COS (TFLITE), Cosh (ONNX),
 Cropping1D (KERAS), Cropping2D (KERAS), CustomFloorDiv (KERAS), CustomFloorMod (KERAS), CustomPow (KERAS),
 CustomReshape (KERAS), CustomShape (KERAS), CustomUnpack (KERAS), Dense (KERAS), DEPTHWISE_CONV_2D (TFLITE),
 DepthwiseConv2D (KERAS), DEQUANTIZE (TFLITE), DequantizeLinear (ONNX), Div (ONNX), DIV (TFLITE), Dropout (KERAS),
 Dropout (ONNX), ELU (KERAS), Elu (ONNX), ELU (TFLITE), Equal (ONNX), EQUAL (TFLITE), Erf (ONNX), Exp (ONNX),
 EXP (TFLITE), EXPAND_DIMS (TFLITE), Flatten (KERAS), Flatten (ONNX), Floor (ONNX), FLOOR (TFLITE), FLOOR_DIV (TFLITE),
 FLOOR_MOD (TFLITE), FULLY_CONNECTED (TFLITE), GaussianDropout (KERAS), GaussianNoise (KERAS), Gemm (ONNX),
 GlobalAveragePool (ONNX), GlobalAveragePooling1D (KERAS), GlobalAveragePooling2D (KERAS), GlobalMaxPool (ONNX),
 GlobalMaxPooling1D (KERAS), GlobalMaxPooling2D (KERAS), Greater (ONNX), GREATER (TFLITE), GREATER_EQUAL (TFLITE),
 GreaterOrEqual (ONNX), GRU (KERAS), HARD_SWISH (TFLITE), HardMax (ONNX), HardSigmoid (ONNX), Identity (ONNX),
 InputLayer (KERAS), InstanceNormalization (ONNX), KaldiNormLayer (KERAS), L2_NORMALIZATION (TFLITE),
 LabelEncoder (ONNX), LEAKY_RELU (TFLITE), LeakyReLU (KERAS), LeakyRelu (ONNX), Less (ONNX), LESS (TFLITE),
 LESS_EQUAL (TFLITE), LessOrEqual (ONNX), LOCAL_RESPONSE_NORMALIZATION (TFLITE), Log (ONNX), LOG (TFLITE),
 LOG_SOFTMAX (TFLITE), LOGICAL_AND (TFLITE), LOGICAL_NOT (TFLITE), LOGICAL_OR (TFLITE), LOGISTIC (TFLITE),
 LogSoftMax (ONNX), LpNormalization (ONNX), LRN (ONNX), LSTM (KERAS), LSTM (ONNX), LSTM (TFLITE), MatMul (ONNX),
 Max (ONNX), MAX_POOL_2D (TFLITE), Maximum (KERAS), MAXIMUM (TFLITE), MaxPool (ONNX), MaxPooling1D (KERAS),
 MaxPooling2D (KERAS), Mean (ONNX), MEAN (TFLITE), Min (ONNX), Minimum (KERAS), MINIMUM (TFLITE), MIRROR_PAD (TFLITE),
 Mul (ONNX), MUL (TFLITE), Multiply (KERAS), Neg (ONNX), NEG (TFLITE), Not (ONNX), Or (ONNX), PACK (TFLITE),
 Pad (ONNX), PAD (TFLITE), PADV2 (TFLITE), Permute (KERAS), PlaceholderCustomLayer (KERAS), Pow (ONNX),
 POW (TFLITE), PReLU (KERAS), PRelu (ONNX), PRELU (TFLITE), QLinearConv (ONNX), QLinearMatMul (ONNX),
 QUANTIZE (TFLITE), QuantizeLinear (ONNX), Reciprocal (ONNX), REDUCE_ANY (TFLITE), REDUCE_MAX (TFLITE),
 REDUCE_MIN (TFLITE), REDUCE_PROD (TFLITE), ReduceL1 (ONNX), ReduceL2 (ONNX), ReduceMax (ONNX), ReduceMean (ONNX),
 ReduceMin (ONNX), ReduceProd (ONNX), ReduceSum (ONNX), ReduceSumSquare (ONNX), ReLU (KERAS), Relu (ONNX),
 RELU (TFLITE), RELU6 (TFLITE), RELU_N1_TO_1 (TFLITE), RepeatVector (KERAS), Reshape (KERAS), Reshape (ONNX),
 RESHAPE (TFLITE), Resize (ONNX), RESIZE_BILINEAR (TFLITE), RESIZE_NEAREST_NEIGHBOR (TFLITE), Round (ONNX),
 ROUND (TFLITE), Rsqrt (ONNX), RSQRT (TFLITE), Selu (ONNX), SeparableConv1D (KERAS), SeparableConv2D (KERAS),
 Shape (ONNX), SHAPE (TFLITE), Sigmoid (ONNX), Sign (ONNX), Sin (ONNX), SIN (TFLITE), Sinh (ONNX), Slice (ONNX),
 SLICE (TFLITE), Softmax (KERAS), Softmax (ONNX), SOFTMAX (TFLITE), Softplus (ONNX), Softsign (ONNX),
 SPACE_TO_BATCH_ND (TFLITE), SpatialDropout1D (KERAS), SpatialDropout2D (KERAS), SPLIT (TFLITE), Sqrt (ONNX),
 SQRT (TFLITE), SQUARE (TFLITE), Squeeze (ONNX), SQUEEZE (TFLITE), STRIDED_SLICE (TFLITE), Sub (ONNX),
 SUB (TFLITE), Subtract (KERAS), Sum (ONNX), SUM (TFLITE), SVMClassifier (ONNX), SVMRegressor (ONNX),
 Tan (ONNX), Tanh (ONNX), TANH (TFLITE), TensorFlowOpLayer (KERAS), ThresholdedReLU (KERAS), ThresholdedRelu (ONNX),
 Tile (ONNX), TILE (TFLITE), TimeDistributed (KERAS), Transpose (ONNX), TRANSPOSE (TFLITE), TRANSPOSE_CONV (TFLITE),
 TreeEnsembleRegressor (ONNX), UNIDIRECTIONAL_SEQUENCE_LSTM (TFLITE), UNPACK (TFLITE), Unsqueeze (ONNX),
 Upsample (ONNX), UpSampling1D (KERAS), UpSampling2D (KERAS), Xor (ONNX), ZeroPadding1D (KERAS), ZeroPadding2D (KERAS),
 ZipMap (ONNX)
```

life.augmented

Some advanced features



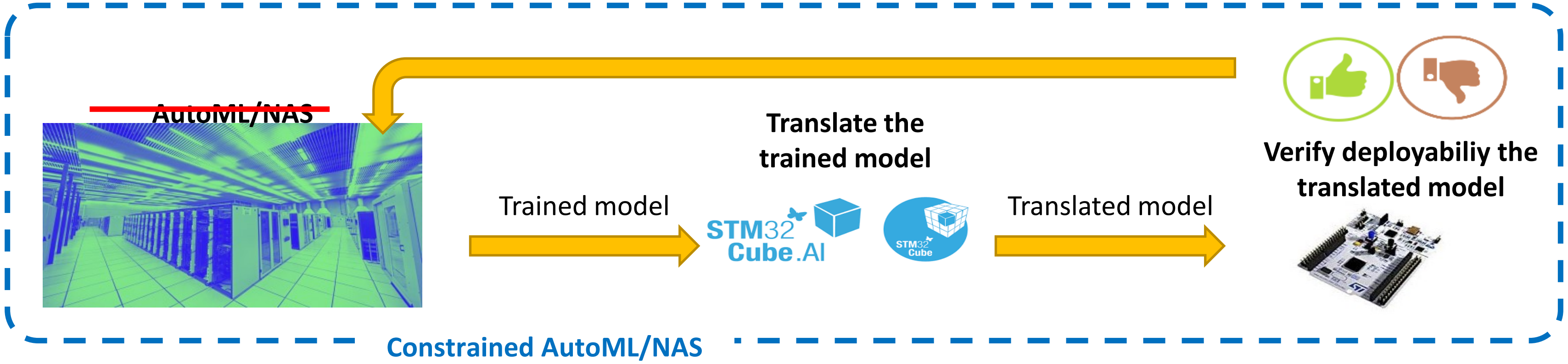
= to express developer creativity, create differentiation, your IPs
Available since X-CUBE-AI v6



ONNX-ML

= mix neural and machine learning layers
Available since X-CUBE-AI v7

Automated deployable design
Through X-CUBE-AI CLI



life.augmented

Dataset of sodium chloride sterile liquid in bottles for intravenous administration and fill level monitoring
Journal: Data in Brief
Danilo Pietro Pau; Luca Simonetta; Bipin Kumar; Prashant Namekar; Gauri Dhande, Dec 2020

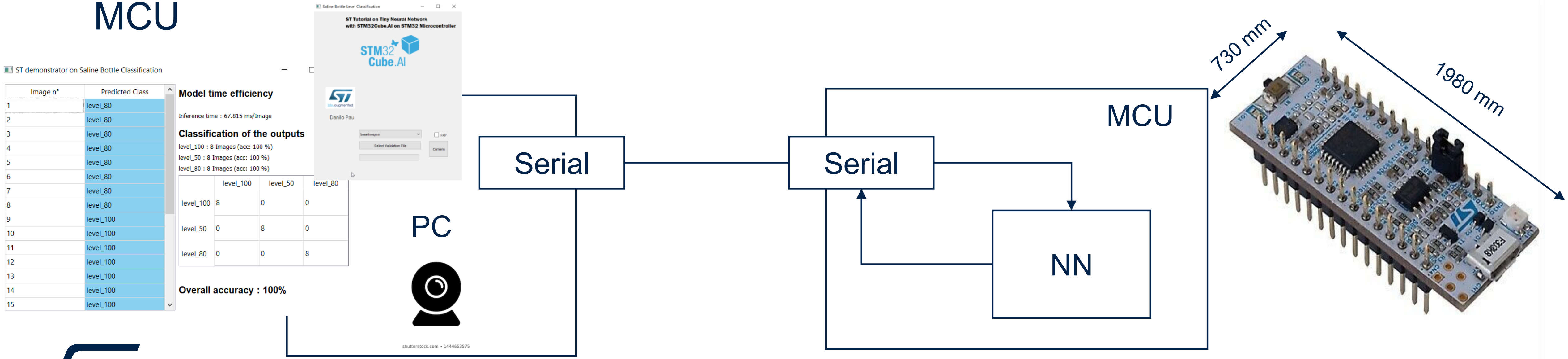
Saline bottle image classification

	480MHz 2MB FLASH 1MB RAM	MACC	PARAMS	ROM KB	RAM KB	TIME/INFERENCE ms	ACCURACY
1	Baseline FP32	20,863,600	1,247,267	4.76 MB	140	NA	95.3%
2	Baseline FP32 compression=4	20,863,085	1,247,267	1.38 MB	140	212.46	same as 1
3	Baseline INT8	20,776,316	1,247,267	1.19 MB	40.94	69.6	same as 1
4	Reduced FP32	6,707,837	67,891	265.2	48.25	75.29	96%
5	Reduced INT8	6,669,699	67,891	67.11	17.53	35.9	same as 4
6	Tiny1 FP32	451,045	11,547	45.11	17.75	6.51	93.2%
7	Tiny1 INT8	451,051	11,547	11.64	8.45	2.43	same as 6
8	Tiny1DW FP32	253,384	9,419	35.67	18.12	3.84	90.84%
9	Tiny1DW INT8	246,772	9,131	9.46	9.65	1.55	-0.43 vs 8
10	Tiny2 FP32	396,625	7,671	29.96	17.75	6	92.1%
11	Tiny2 INT8	387,687	7,671	7.75	8.37	2.3	+0.43 vs 10
12	Tiny3 Sep FP32	342,749	4,910	19.5	35	6.7	93.5%
13	Tiny3 Sep INT8	342,755	4,910	5.5	9.64	2.63	-0.44 vs 12
14	Tiny4 Sep FP32	43,672	1,126	4.57	12.2	1.4	84.55%
15	Tiny4 Sep INT8	39,023	1,126	1.46	3.16	0.821	same as 14

life.augmented

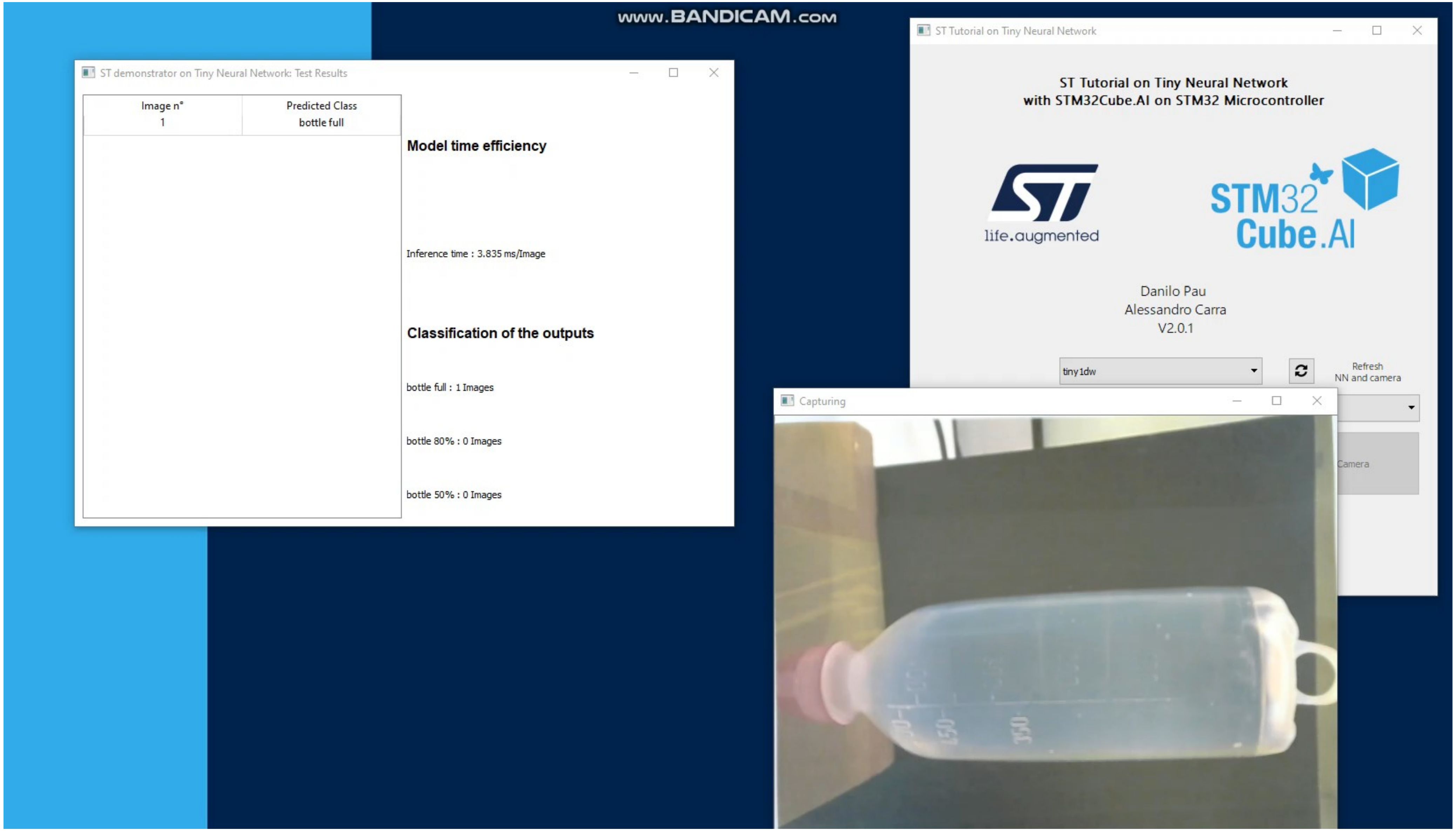
Affordable POC for early IoT practitioners

- Attach a MUCLEO-STM32 to the laptop:
  - STM32 MCU runs the model generated with X-CUBE-AI «validation on target» app
  - Image Sensor is attached to PC (webcam), python script reads sensor data and sends to the NUCLEO-STM32, which process and send back results to GUI
  - Communication through a serial port emulated on USB
  - Data information is encoded on PC with the STM32 binary protocol and decoded on MCU

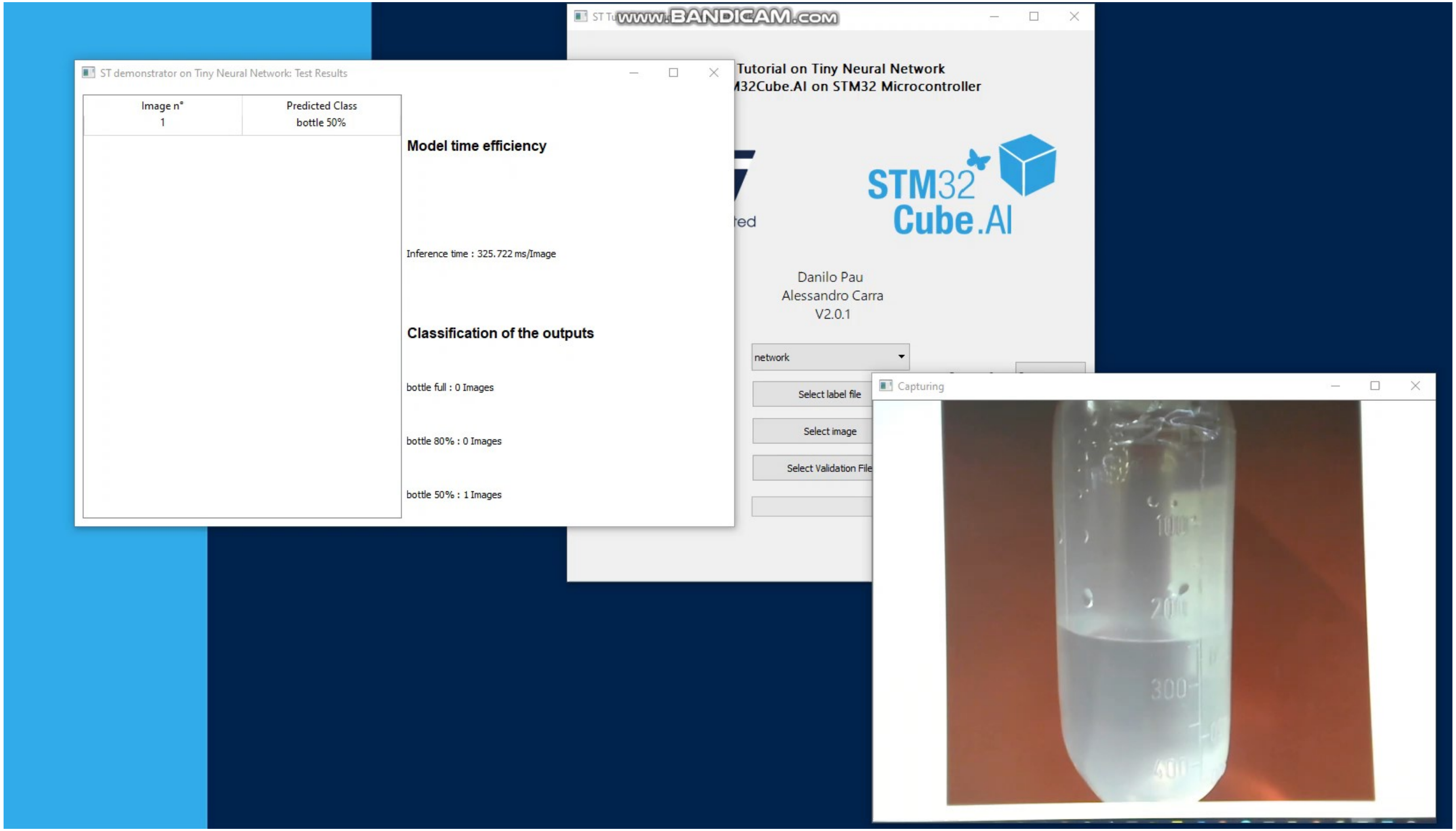


X-CUBE-AI built-in application

Affordable POC for early IoT STM32H7 practitioners



Affordable POC for early IoT STM32F3 practitioners



(Tiny) Function Packs

Simple, fast, optimized

STM32 Cube.AI

STM32 Embedded Software Packages

Part Number	Manufacturer	Description
X-LINUX-AI	ST	STM32 MPU OpenSTLinux Expansion Pack for AI computer vision application
FP-AI-SENSING1	ST	STM32Cube function pack for ultra-low power IoT node with artificial intelligence (AI) application based on audio and motion sensing
FP-AI-VISION1	ST	STM32Cube function pack for high performance STM32 with artificial intelligence (AI) application for Computer Vision
FP-AI-NANOEDG1	ST	Artificial Intelligence (AI) condition monitoring function pack for STM32Cube
FP-AI-FACEREC	ST	Artificial Intelligence (AI) face recognition function pack for STM32Cube
FP-AI-CTXAWARE1	ST	STM32Cube function pack for ultra-low power context awareness with distributed artificial intelligence (AI)

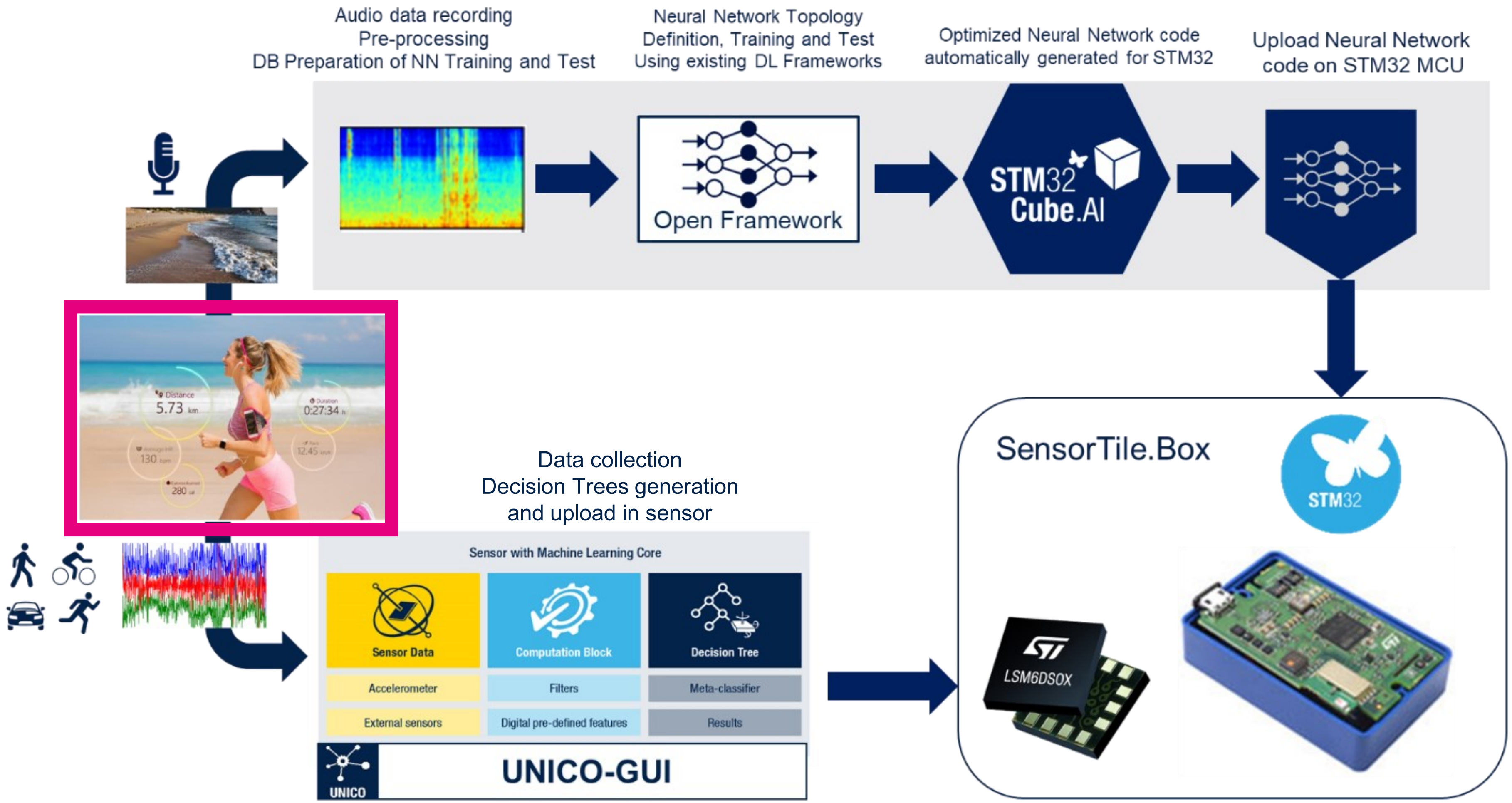
https://www.st.com/content/st_com/en/ecosystems/stm32-ann.html

Context awareness: https://www.youtube.com/watch?v=l_XqYFci5PE

FP-AI-CXTAWARE1: the best power system saving solution

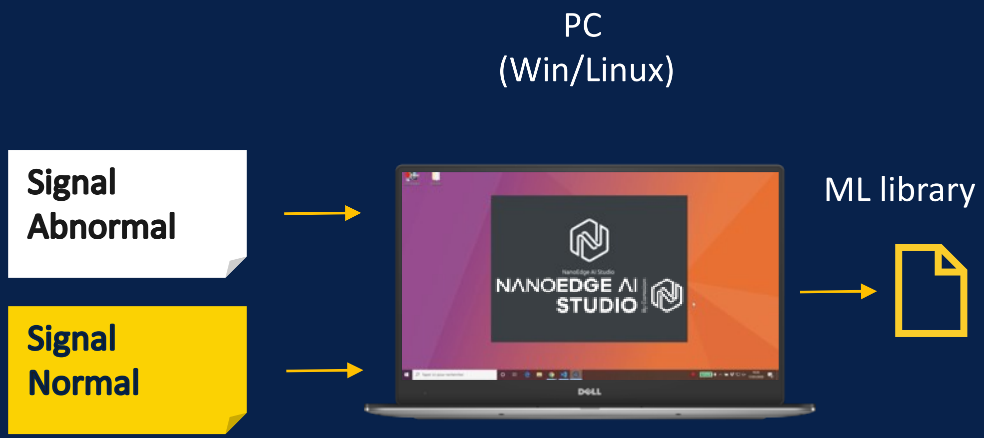
MCU
Audio Scene Classification

Sensor
Human Activity Recognition



Step 1 : **(PC Side)**
Creation of an ANOMALY DETECTION Machine Learning library

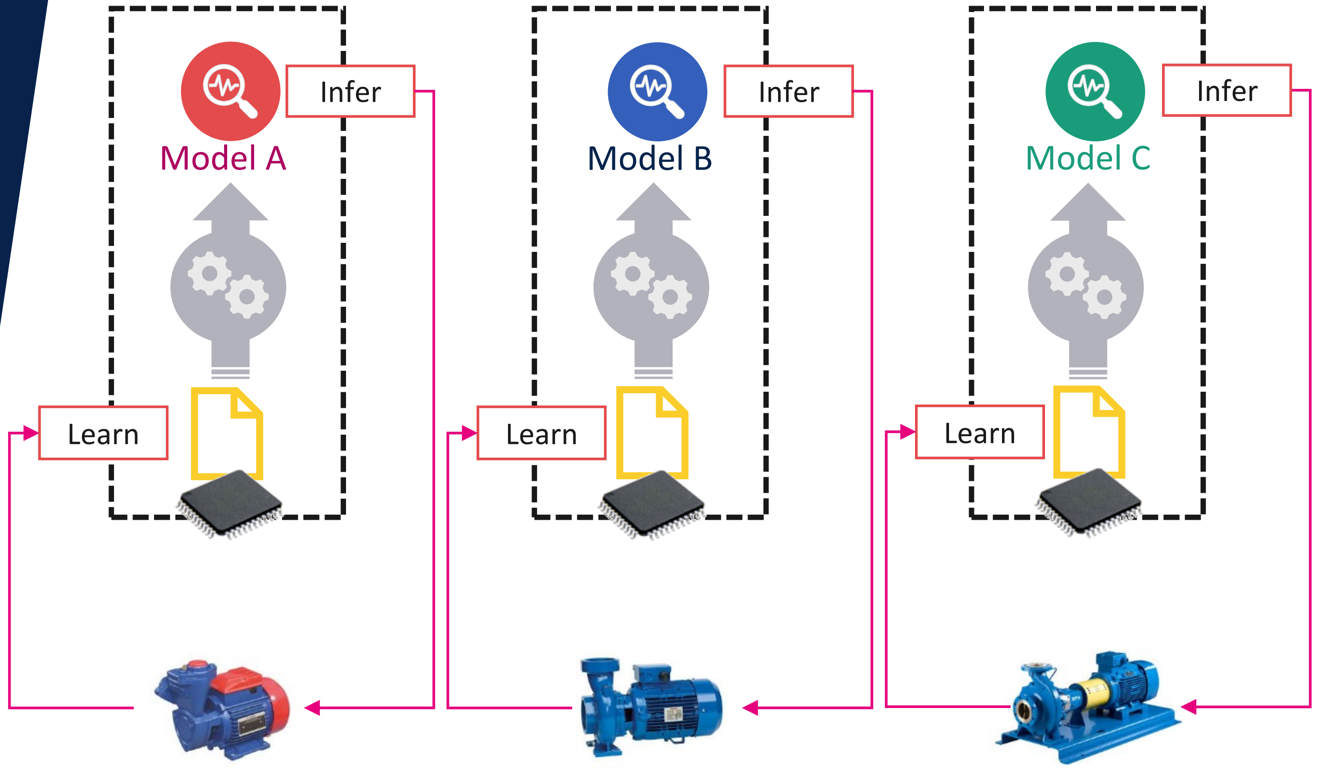
In just a few steps, create a machine learning library***, custom to your project, and based on a small amount of data captured using your sensor.**



*A NanoEdge AI **Machine Learning Library is a self learning engine, that will train a ML model, inside the Microcontroller and based on locally acquired data/signal.**

Step 2 : **(MCU Side)**
Use of an ANOMALY DETECTION Machine Learning library

Each library learns the engine on which it is placed and then analyses it by comparing the learning done locally with the new signals that are coming.



NanoEdge AI by Cartesiam

NanoEdge AI is a static AI library
for embedded c software
running on any Arm Cortex MCU

Learning and inference
done at the edge.

No pre-trained neural network needed

All work (learning and inference)
executed inside the STM32 MCU

NANOEDGE AI™
by CARTESIAM

CARTESIAM.AI
ARTIFICIAL INTELLIGENCE AT THE EDGE

life.augmented

SPC5 Studio.AI

AI beyond autonomous-driving

Multiple scenarios

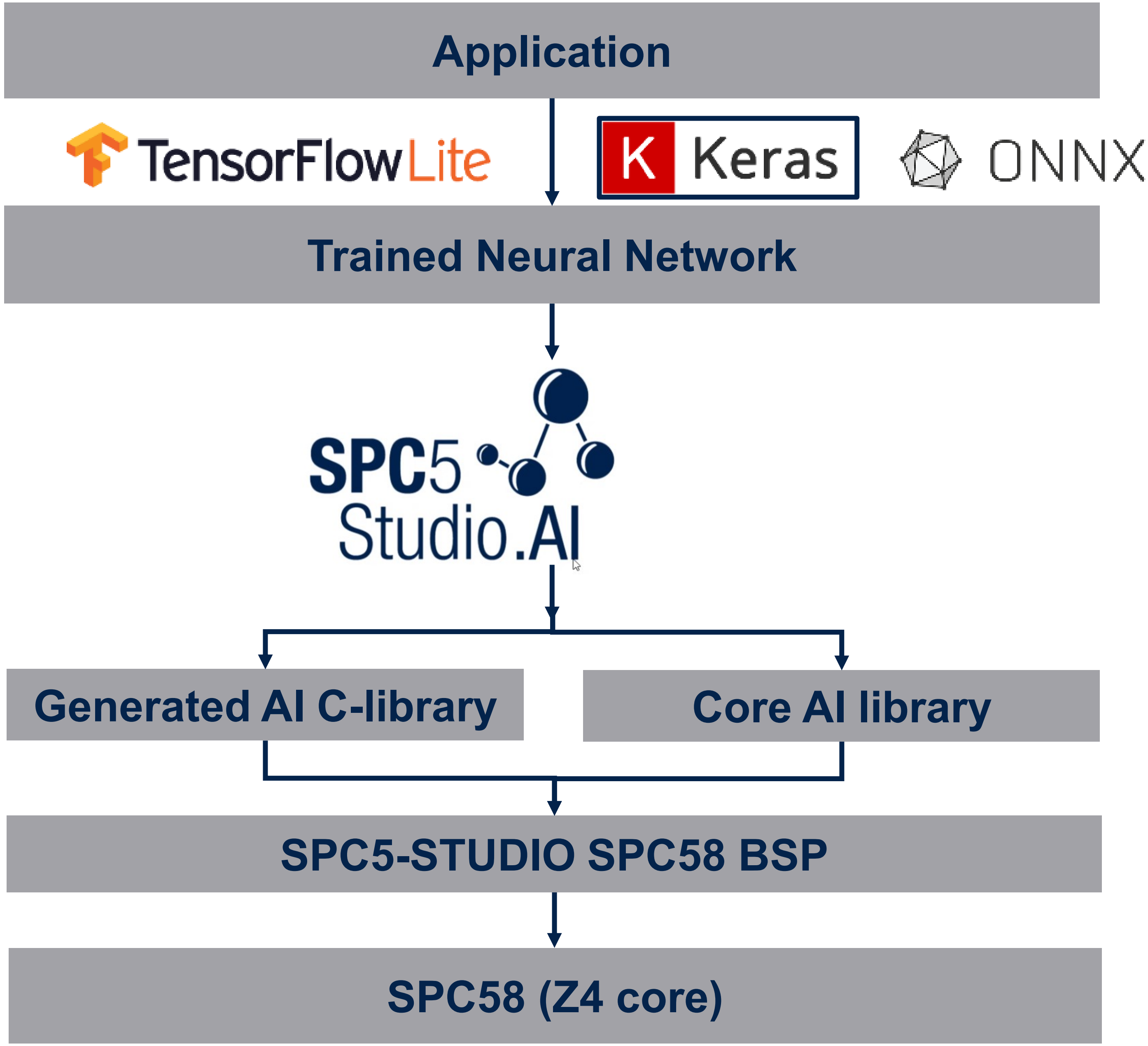
Electrification
- Battery Management: State of Health and Charge
- Hybrid: Efficient Propulsion Mix
- Transmission: Improved Torque Control

Predictive Maintenance
- Early Detection of anomalies
- Dynamic recalibration to improve robustness

Vehicle Security
- Hacking, Malicious Attacks
- Physical Vehicle Tampering

In-Cabin Behaviour Monitoring
- Emergency conditions: Drowsiness, Illness, “Backseat Child”
- Adaptive Comfort settings & driving mode (Sport / Eco)
- Driver identification and monitoring

Sensor Augmentation
- NOx emission prediction
- Virtual sensors (Vehicle Attitude)
- Sensor degradation compensation



- OpenSource IDE based on ECLIPSE
- Full MISRA 2012 compliant register level access (RLA) low level drivers
- MISRA 2012 checking for customer code
- Visual MCU's pins configuration, run modes and full clock tree configuration with automatic constraints checking
- FreeRTOS support
- Software examples for discovery kits and premium evaluation boards covering most used peripherals
- Compilers: GCC, GHS (green hills software), HighTec EDV-System
- **Artificial Intelligence plug-in; generated software easily portable across different boards**

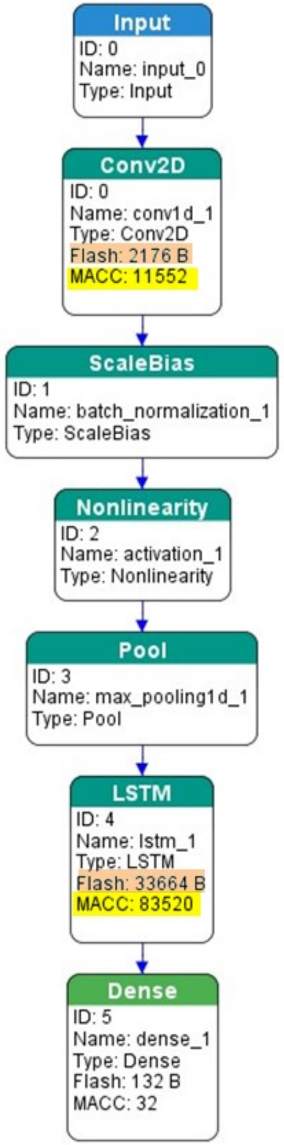
1) Intrusion detection
2) Battery charge prediction
3) Battery lifetime prediction
4) Audio de-noising

Battery Management with Long Short-Term Memory (LSTM)

Example Neural Net provided with SPC5 Studio.AI

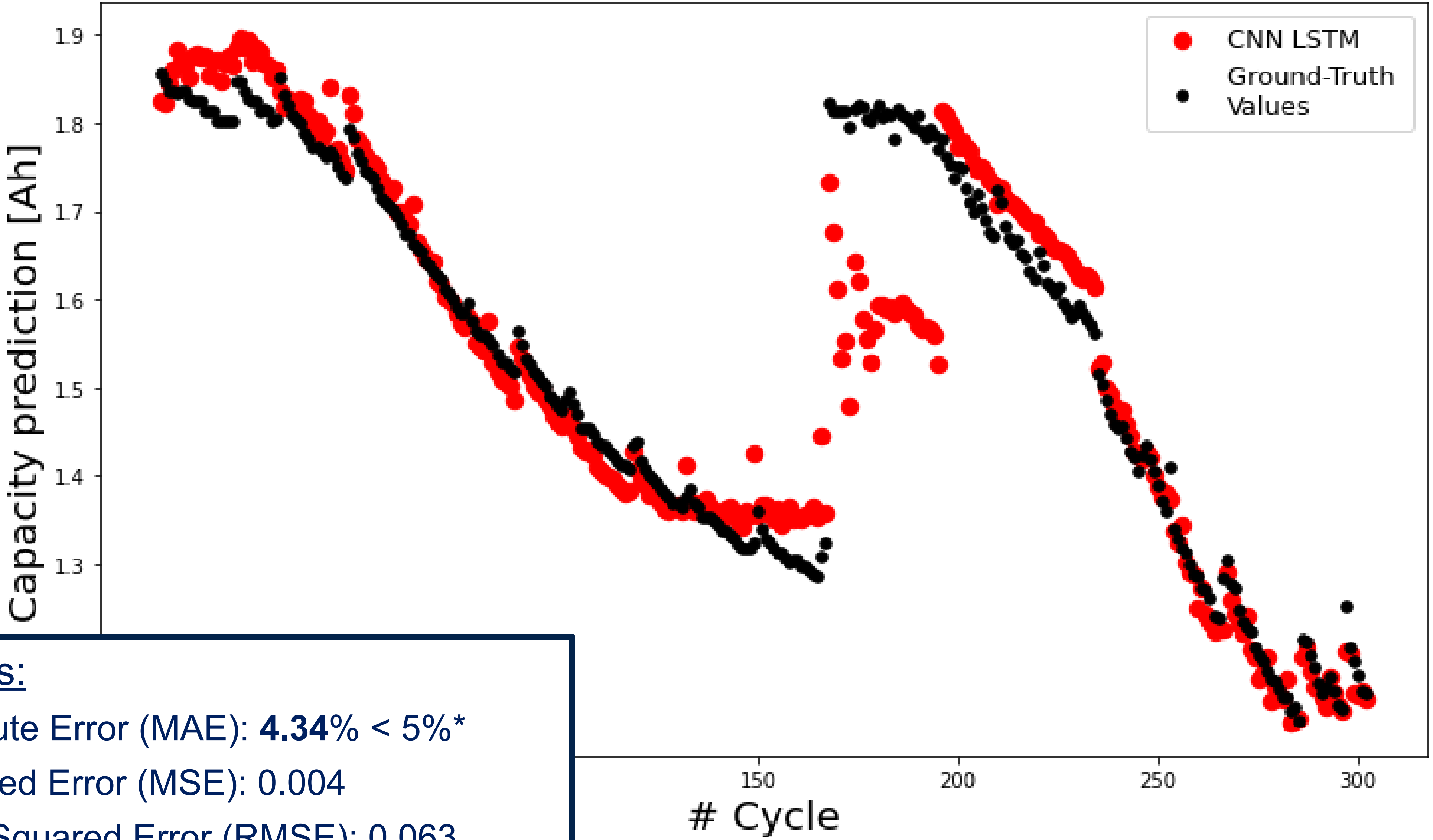
Actual Performance on Chorus family of automotive MCUs

Measured with SPC5 Studio.AI rel. 2.0.0, 32b-floating-point			
Device	**Flash used (kB)**	**RAM used (kB)**	**Inference time (ms)**
SPC584B 1 (of 1) z4 core @120MHz	149.05	6.66	6.3791
SPC58EC 1 (of 2) z4 @180MHz	152.35	6.79	4.3906
SPC58EG 1 (of 3) z4 @180MHz	154.46	6.91	4.398
SPC58NH 1 (of 3) z4 @200MHz	153.98	6.85	3.7923



Dataset http://ti.arc.nasa.gov/project/ prognostic-datarepository.

life.augmented



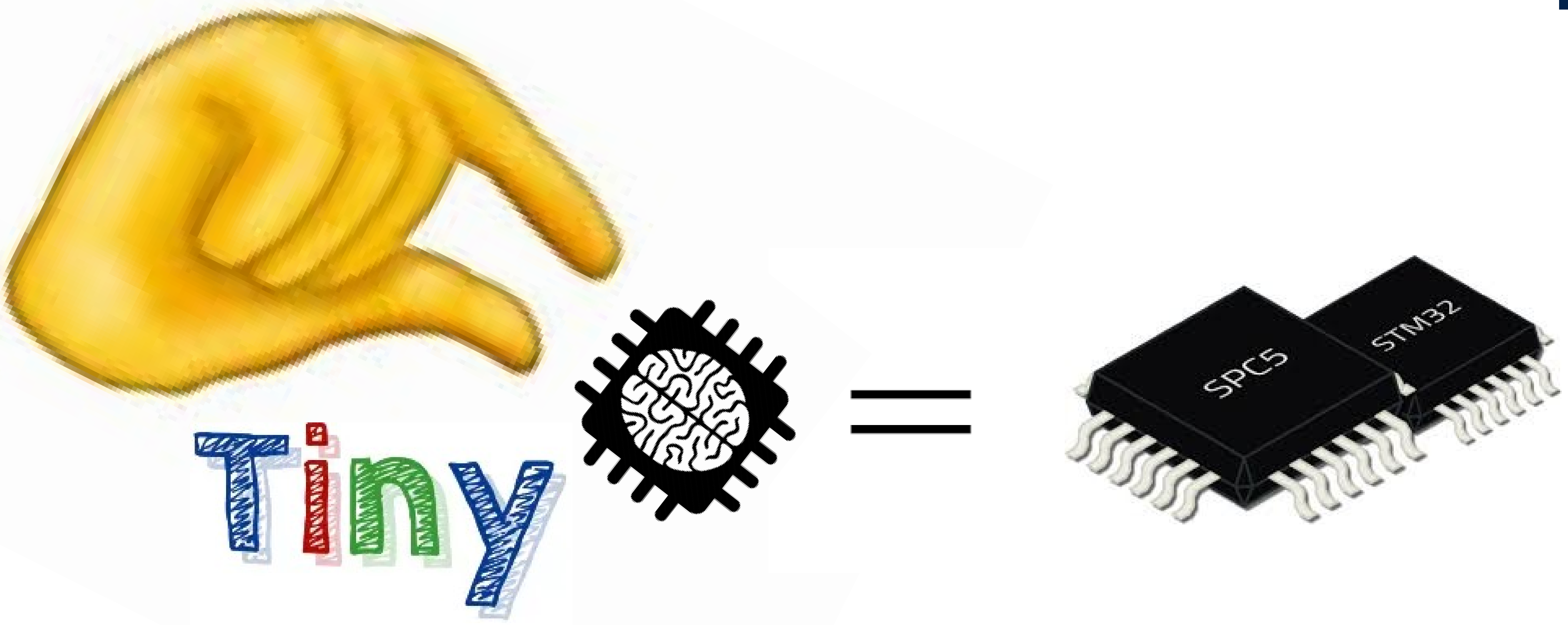
Testing results:

- Mean Absolute Error (MAE): **4.34**% < 5%*
- Mean Squared Error (MSE): 0.004
- Root Mean Squared Error (RMSE): 0.063

* EVs acceptable SoH error range

Dataset http://ti.arc.nasa.gov/project/ prognostic-datarepository.

Design ML



Q&A

Thank you

danilo.pau@st.com

life.augmented

Copyright Notice

www.tinyML.org