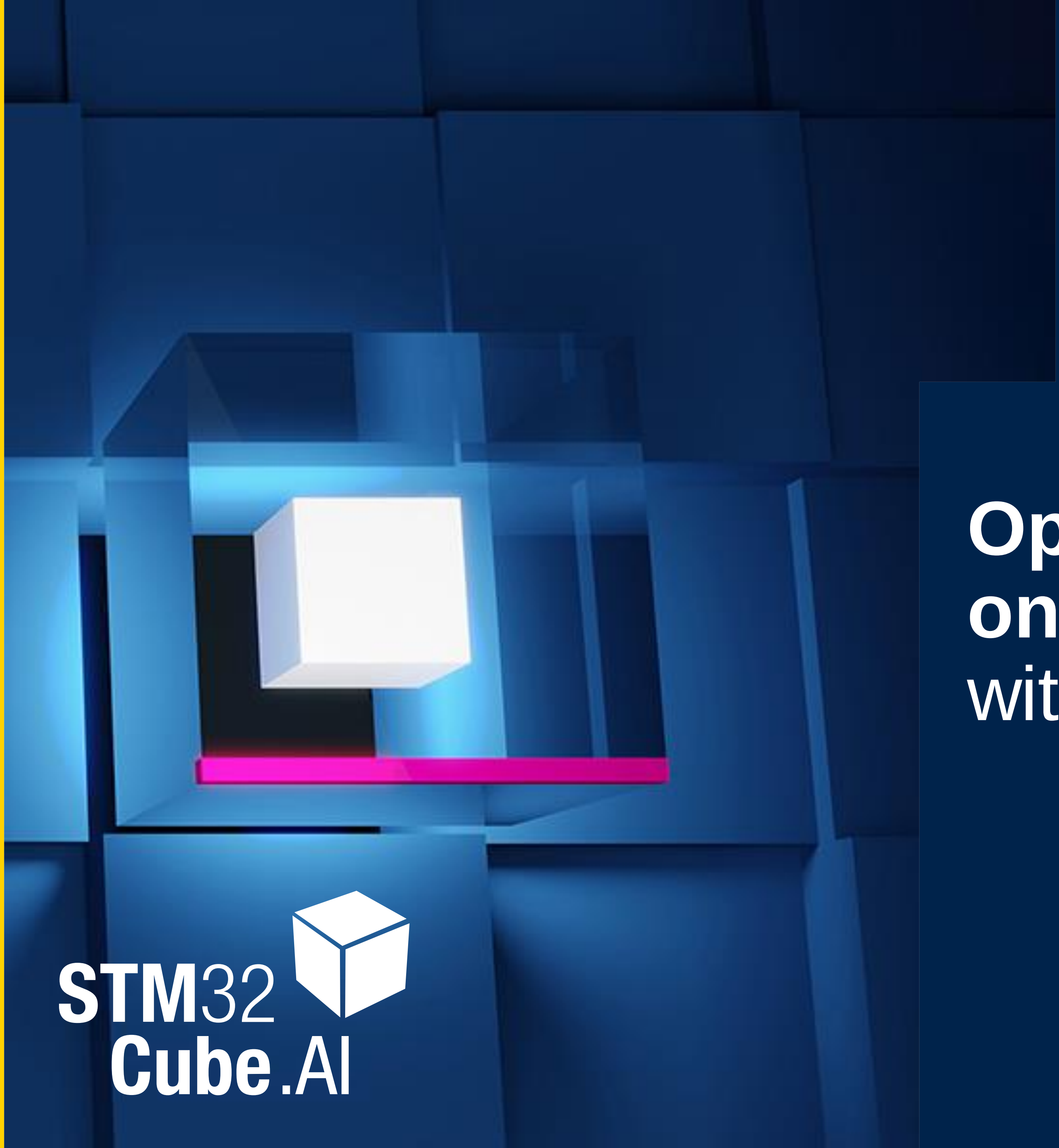


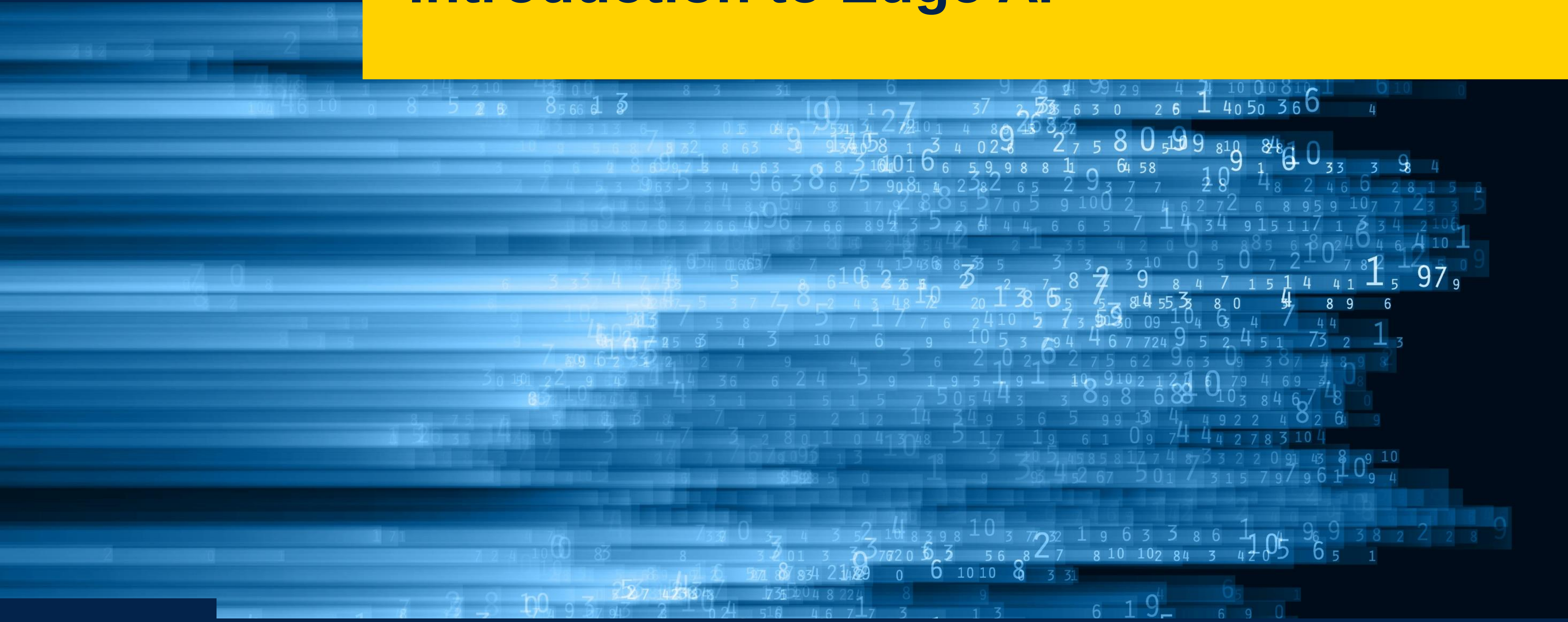
life.augmented

Optimize neural networks on STM32

with STM32Cube.AI

STM32

Introduction to Edge AI



Signals turning into data

Embedded applications will collect more data in the future

Growing demand for data-driven insights

Increasing use of sensors

Proliferation of IoT devices

AI is offering the best approach to process this growing amount of data

Algorithms and **predefined models** to analyze data and make predictions or decisions

Machine learning algorithms to automatically **learn** patterns and relationships **from the data**

Traditional approaches show their limitations:

- when dealing with **large datasets**
- when the **phenomena are too complex**

AI-based data processing offers a more flexible and powerful approach to analyzing and making decisions from large data collection

The raise of Edge AI

Ultra-low latency
Real-time applications

Reduced data transmission
Generate meaningful information

Enhanced privacy and security
No data sharing in the cloud

Power efficiency
Low-data / Low-power

Improved accuracy
analyze data from a wide range of sensors and sources

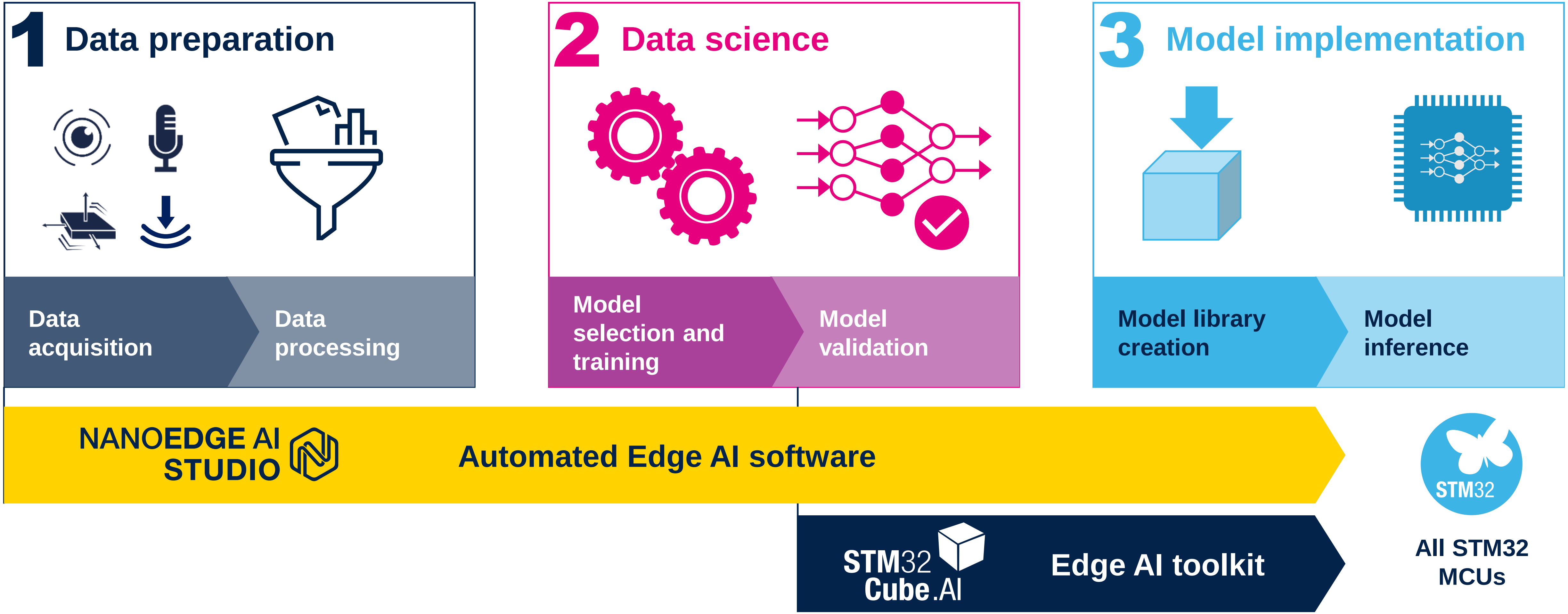
Edge AI will benefit many application domains:

Industrial maintenance
Condition monitoring
Predictive maintenance

Control systems
From home heating systems to industrial machines

Internet of Things (IoT)
smart cities, smart buildings, connected homes, and industrial automation

AI development workflow – ST software offering



1 Data preparation

- Data acquisition
- Data processing

2 Data science

- Model selection and training
- Model validation

3 Model implementation

- Model library creation
- Model inference

NANOEDGE AI STUDIO — Automated Edge AI software

STM32 Cube.AI — Edge AI toolkit

STM32

All STM32 MCUs

life.augmented

ST ecosystem ease your AI to reach production level

	STM32 Cube.AI **Edge AI toolkit** for model optimization on STM32	NANOEDGE AI STUDIO **Automated ML software** for end-to-end Edge AI solution design on STM32
Key benefits	✓ Get optimized C-code from your trained model ✓ Desktop and online versions ✓ Benchmark service on remote hardware (online version) ✓ On-device performance validation	✓ The easiest way to integrate AI into your system ✓ Save resources and development cost ✓ Reach the highest performance with the automated model finder embedded in the tool
Application domain	All	Time series (except voice and speech)
Business model	Free of charge	Free for prototyping on STM32 dev boards Production requires right of use

A proven technology adopted by multiple clients

NANOEDGE AI STUDIO

INDUSTRIAL | DEMO

Fan anomaly detection based on vibrations

Learn to detect abnormal behavior at the edge on a vibrating machine.

nke WATTECO

NANOEDGE AI STUDIO

INDUSTRIAL | CUSTOMER

AI solution for industrial predictive maintenance with NKE Watteco

Predictive maintenance solution for industrial equipment.

CROUZET

NANOEDGE AI STUDIO

TRANSPORTATION | CUSTOMER

AI solution for monitoring automatic doors with Crouzet

Predictive maintenance on motors for automatic door motors.

NANOEDGE AI STUDIO

INDUSTRIAL | DEMO

Anomaly detection in an electric motor

Current sensing to detect abnormal behaviors in motors.

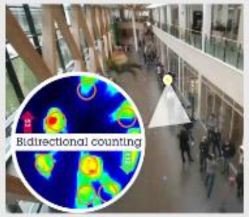
OXYtronic

NANOEDGE AI STUDIO

INDUSTRIAL | CUSTOMER

AI solution for industrial predictive maintenance with Oxytronic

Predictive maintenance solution for industrial equipment.



Schneider Electric

STM32 Cube.AI

SMART OFFICE | CUSTOMER

People flow counting Sensor with Schneider Electric

An innovative approach to measure people flows using an in-house thermal sensor.

STM32 Cube.AI

SMART CITY | DEMO

Acoustic scene classification

Identify different environments (indoor, outdoor, in-car) using a simple microphone.

STM32 Cube.AI

WEARABLES | DEMO

Human Activity Recognition

Easily identify 5 different activities with a 3D accelerometer.

STM32 Cube.AI

INDUSTRIAL | DEMO

People presence detection (visual wake word)

Human detection on high-performance MCU.

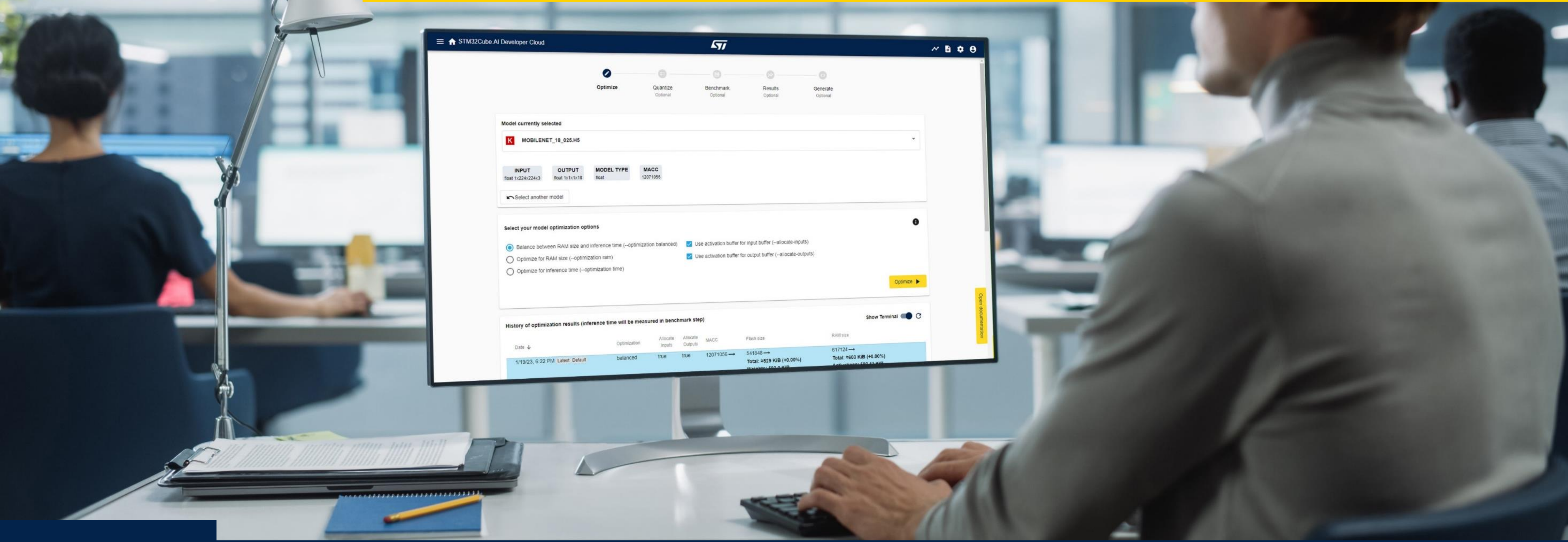
STM32 Cube.AI

INDUSTRIAL | DEMO

Aftermarket wireless digit reader

Equip meters with aftermarket wireless & low-power readers.

STM32Cube.AI
AI optimization tool for STM32 portfolio



Model currently selected

MOBILENET_18_025.H5

INPUT	OUTPUT	MODEL TYPE	MACC
float 1x224x224x3	float 1x1x1x18	float	12071056

Select another model

Select your model optimization options

● Balance between RAM size and inference time (--optimization balanced)
○ Optimize for RAM size (--optimization ram)
○ Optimize for inference time (--optimization time)

☑ Use activation buffer for input buffer (--allocate-inputs)
☑ Use activation buffer for output buffer (--allocate-outputs)

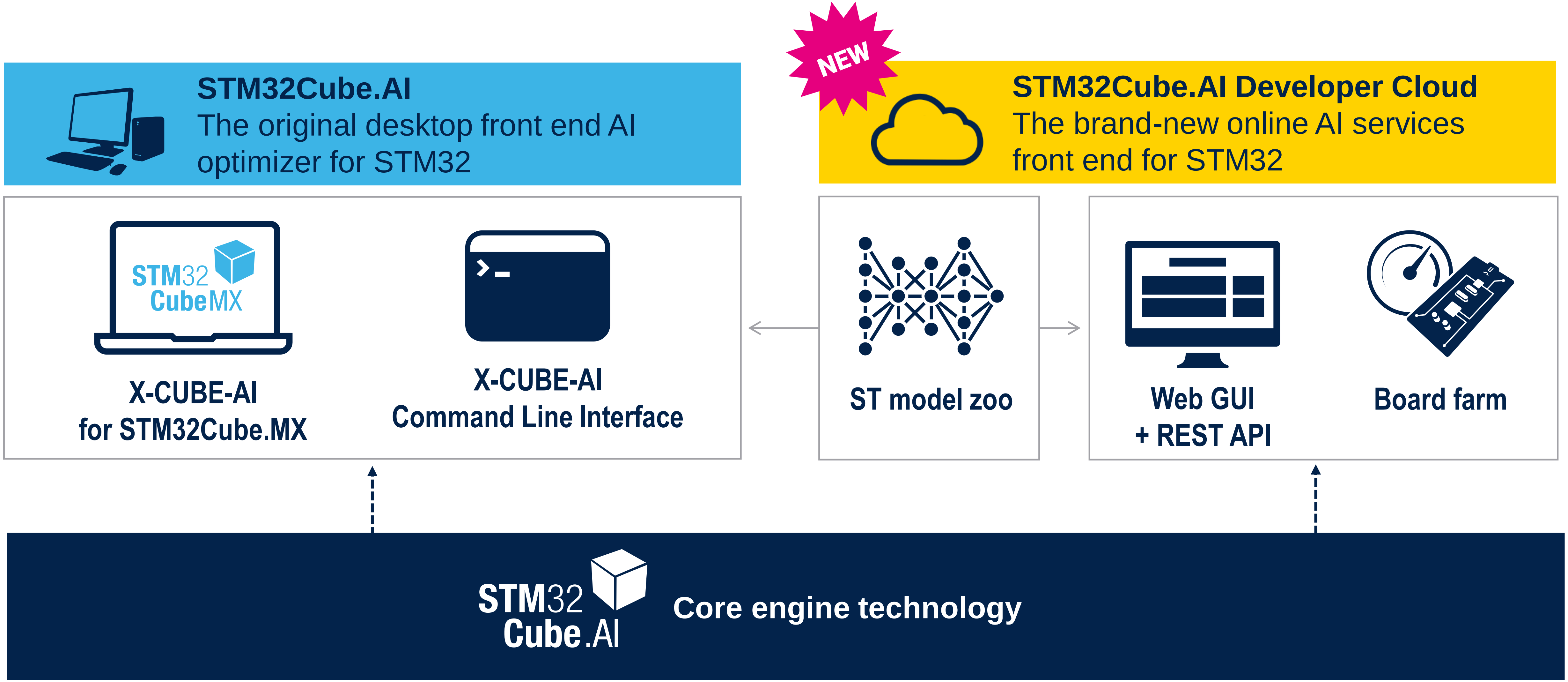
History of optimization results (inference time will be measured in benchmark step)

Show Terminal

Date ↓	Optimization	Allocate Inputs	Allocate Outputs	MACC	Flash size	RAM size
1/19/23, 6:22 PM Latest Default	balanced	true	true	12071056 ⟶	541848 ⟶ Total: ≈529 KiB (+0.00%)	617124 ⟶ Total: ≈603 KiB (+0.00%)

life.augmented

STM32Cube.AI overview



NEW

STM32Cube.AI
The original desktop front end AI optimizer for STM32

STM32Cube.AI Developer Cloud
The brand-new online AI services front end for STM32

X-CUBE-AI
for STM32Cube.MX

X-CUBE-AI
Command Line Interface

ST model zoo

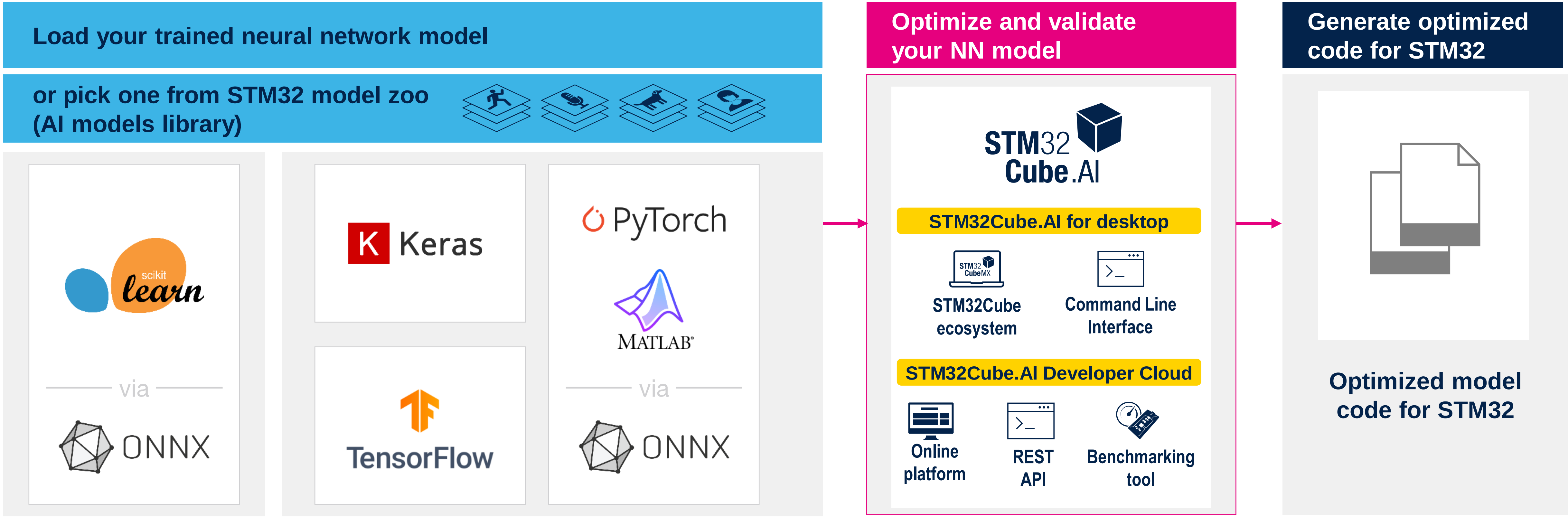
Web GUI
+ REST API

Board farm

STM32 Cube.AI

Core engine technology

One tool – two versions to deploy AI on STM32



Load your trained neural network model

or pick one from STM32 model zoo (AI models library)

Optimize and validate your NN model

STM32Cube.AI for desktop

- STM32Cube ecosystem
- Command Line Interface

STM32Cube.AI Developer Cloud

- Online platform
- REST API
- Benchmarking tool

Generate optimized code for STM32

Optimized model code for STM32

life.augmented

STM32Cube.AI
Get the best AI performance on STM32

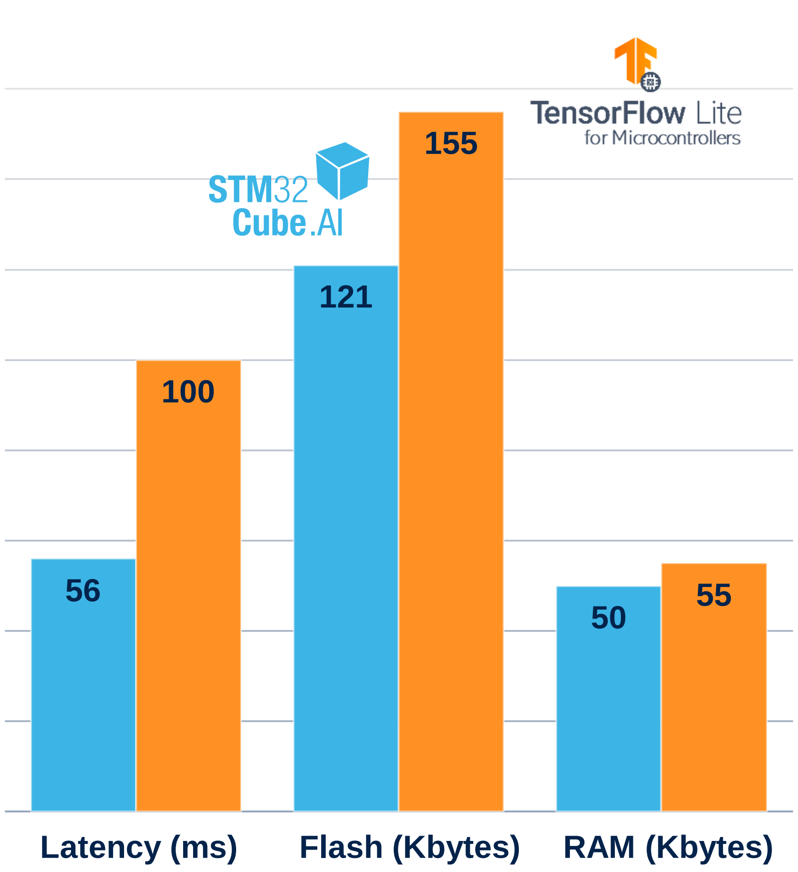
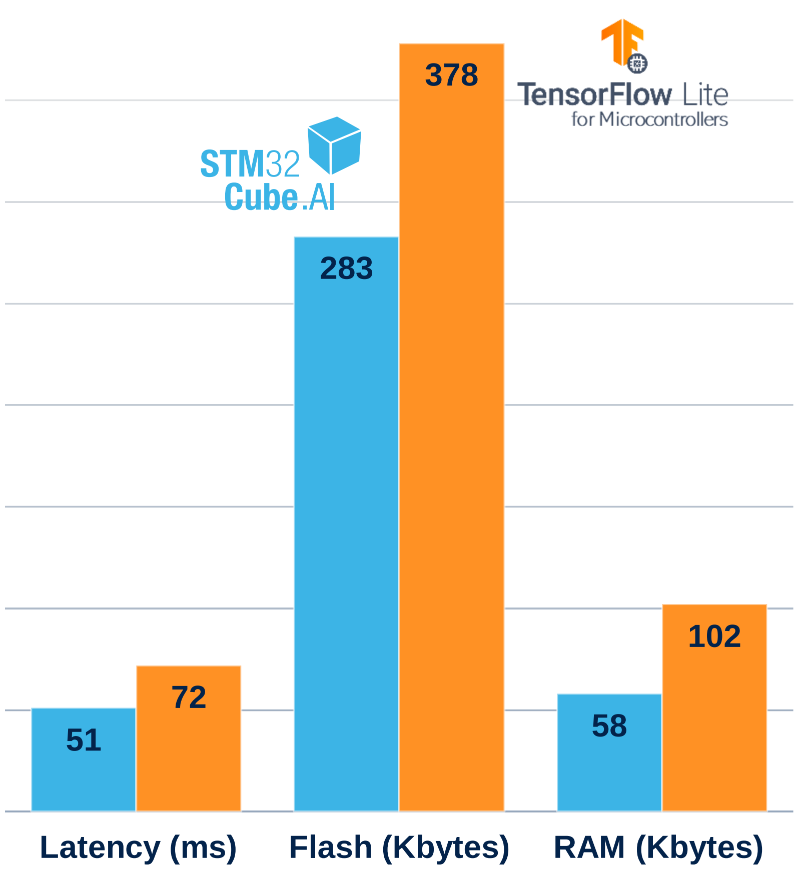


Image Classif v1.0 MLPerf Tiny



Visual Wake Word v1.0 MLPerf Tiny

UP TO

60 %

faster inference time*

UP TO

20 %

space freed-up in flash and RAM*

HW Target: STM32H7A3
Flash: 2 Mbytes
RAM: 1.4 Kbytes
Freq: 280 MHz

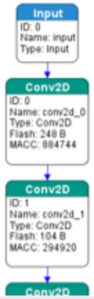
SW Version:
X-Cube.AI v 7.3.0
TFLm v2.10.0

** versus TensorFlow Lite for microcontroller*

The 3 pillars of STM32Cube.AI

Graph optimizer

Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



Input
ID: 0
Name: input
Type: Input

Conv2D
ID: 0
Name: conv2d_0
Type: Conv2D
Flash: 248 B
MACC: 884744

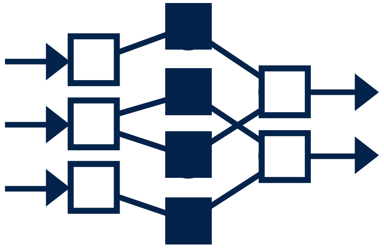
Conv2D
ID: 1
Name: conv2d_1
Type: Conv2D
Flash: 104 B
MACC: 294920

Conv2D

- Auto graph rewrite
- Node/operator fusion
- Layout optimization
- Constant-folding…
- Operator-level info to fine-tune memory footprint and computation

Quantized model support

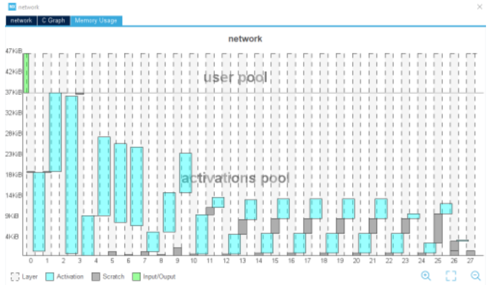
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8 or mixed-precision
- Minimum loss of accuracy
- Code validation on target
  - Latency
  - Accuracy
  - Memory footprint

Memory optimizer

Optimize memory allocation to get the best performance while respecting the constraints of your embedded design

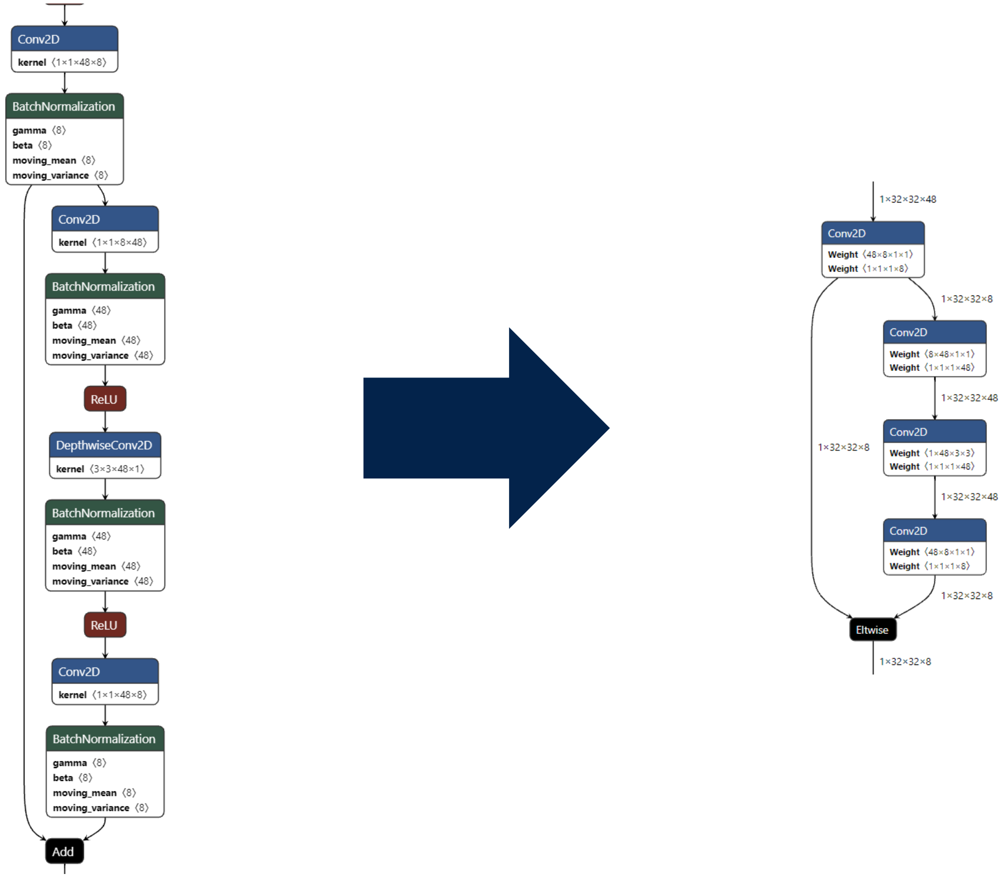


- Memory allocation
- Internal/external memory repartition
- Model-only update option

STM32Cube.AI is **free of charge**, available both in graphical interface and in command line.

Graph optimizer

Squeeze your graph to fit into an MCU!



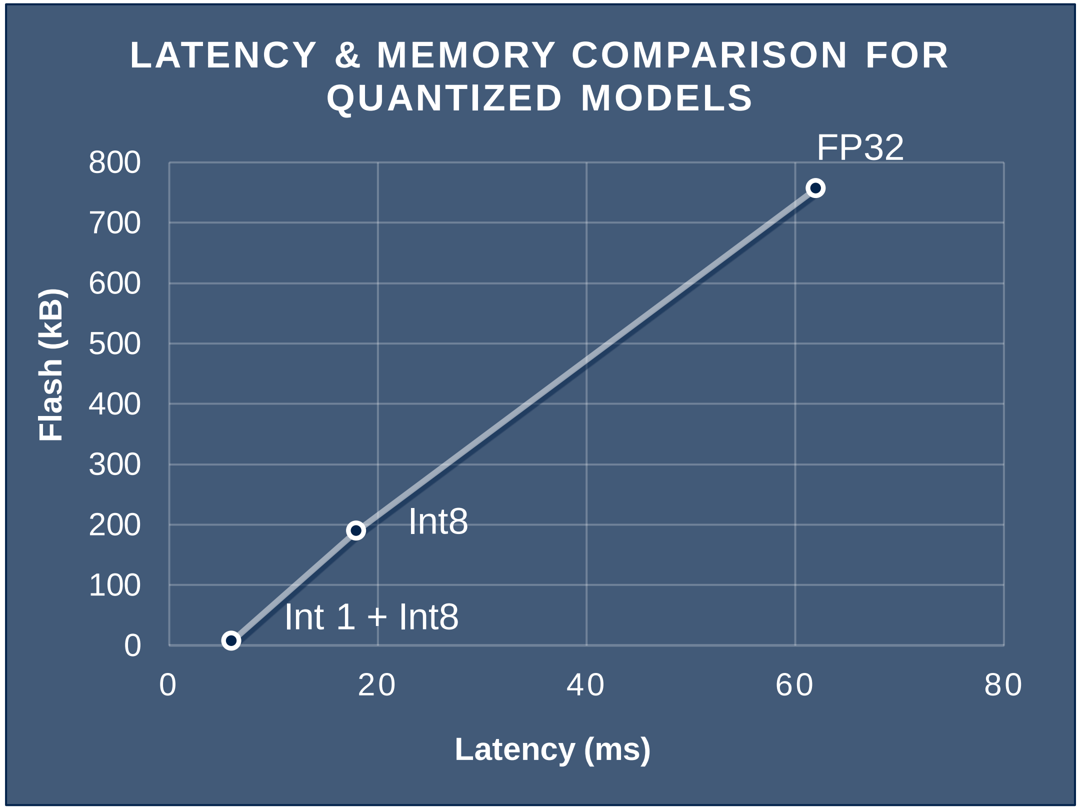
Fully automated process in the STM32Cube.AI workflow

- Your original graph is optimized at the very early stage for optimal integration into STM32 MCU/MPU
- Loss-less conversion

life.augmented

Quantized model support

Simply use quantized networks to reduce memory footprint and inference time



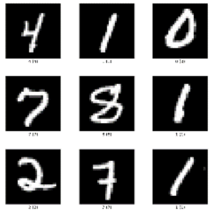
STM32Cube.AI support quantized Neural Network models with **all parameter formats**:

- FP32
- Int8
- Mixed binary Int1 to Int8 (Qkeras*, Larq.dev*)

**Please contact edge.ai@st.com to request the relevant version of STM32Cube.AI*

HW Target: NUCLEO-STM32H743ZI2
Model: Low complexity handwritten digit reading
Freq: 480 MHz
Accuracy: >97% for all quantized models

Tested database: MNIST dataset

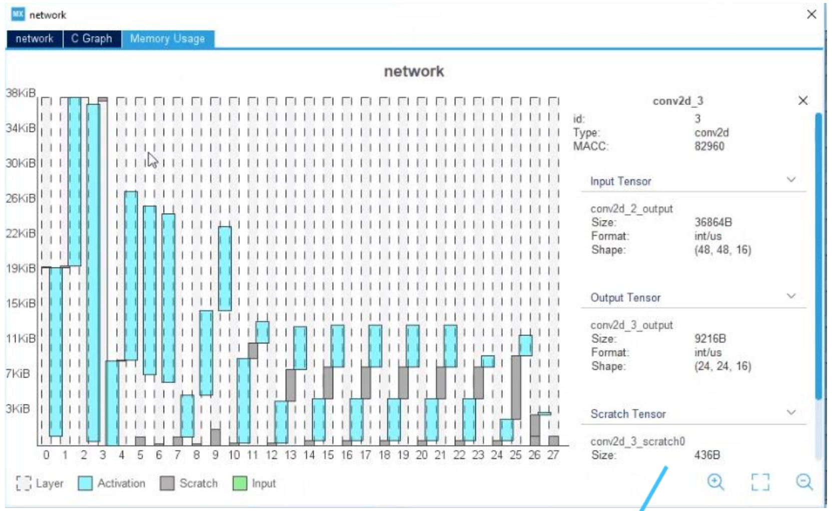


MNIST dataset

life.augmented

Memory optimizer

Optimize performance easily with the memory allocation tool



conv2d_3

id:	3
Type:	conv2d
MACC:	82960

Input Tensor

conv2d_2_output

Size:	36864B
Format:	int/us
Shape:	(48, 48, 16)

Output Tensor

conv2d_3_output

Size:	9216B
Format:	int/us
Shape:	(24, 24, 16)

Scratch Tensor

conv2d_3_scratch0

Size:	436B

Model RAM consumption per layer

- Easily identify most critical layers

Model memory allocation

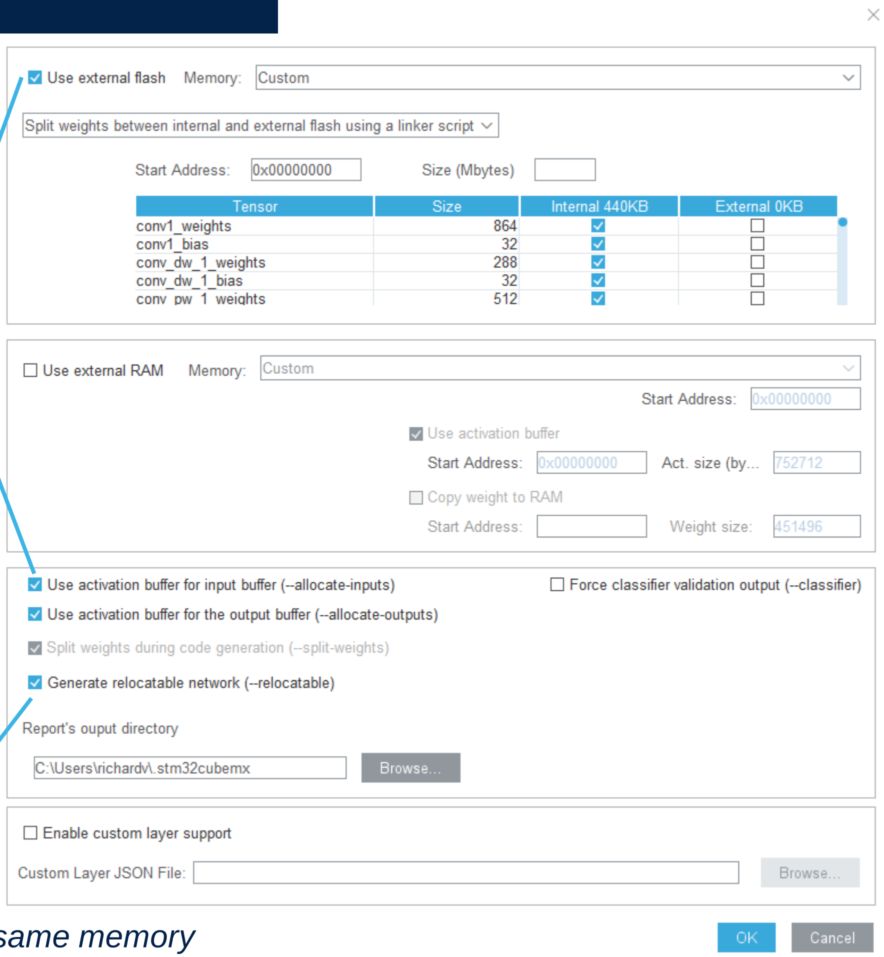
- Set your external memory
- Map in non-contiguous internal flash section
- Partition internal vs external flash memories

Re-use model input buffer to store activation data*

- Minimize RAM requirements

Relocatable network

- A separate binary is generated for the library and the network to enable standalone model upgrade



☑ Use external flash Memory: Custom

Split weights between internal and external flash using a linker script

Start Address: 0x00000000 Size (Mbytes):

Tensor	Size	Internal 440KB	External 0KB
conv1_weights	864	☑	□
conv1_bias	32	☑	□
conv_dw_1_weights	288	☑	□
conv_dw_1_bias	32	☑	□
conv_pw_1_weights	512	☑	□

□ Use external RAM Memory: Custom

Start Address: 0x00000000

☑ Use activation buffer

Start Address: 0x00000000 Act. size (by... 752712

□ Copy weight to RAM

Start Address: Weight size: 451496

☑ Use activation buffer for input buffer (--allocate-inputs)

□ Force classifier validation output (--classifier)

☑ Use activation buffer for the output buffer (--allocate-outputs)

☑ Split weights during code generation (--split-weights)

☑ Generate relocatable network (--relocatable)

Report's ouput directory

C:\Users\richardv\.stm32cubemx

□ Enable custom layer support

Custom Layer JSON File:

** Requires input and activation buffers in same memory*

life.augmented

Performance benchmarking made simple
STM32Cube.AI Developer Cloud

The unique possibility to evaluate the performance of models remotely, on real STM32 boards

Get the real inference time from optimized models running on STM32

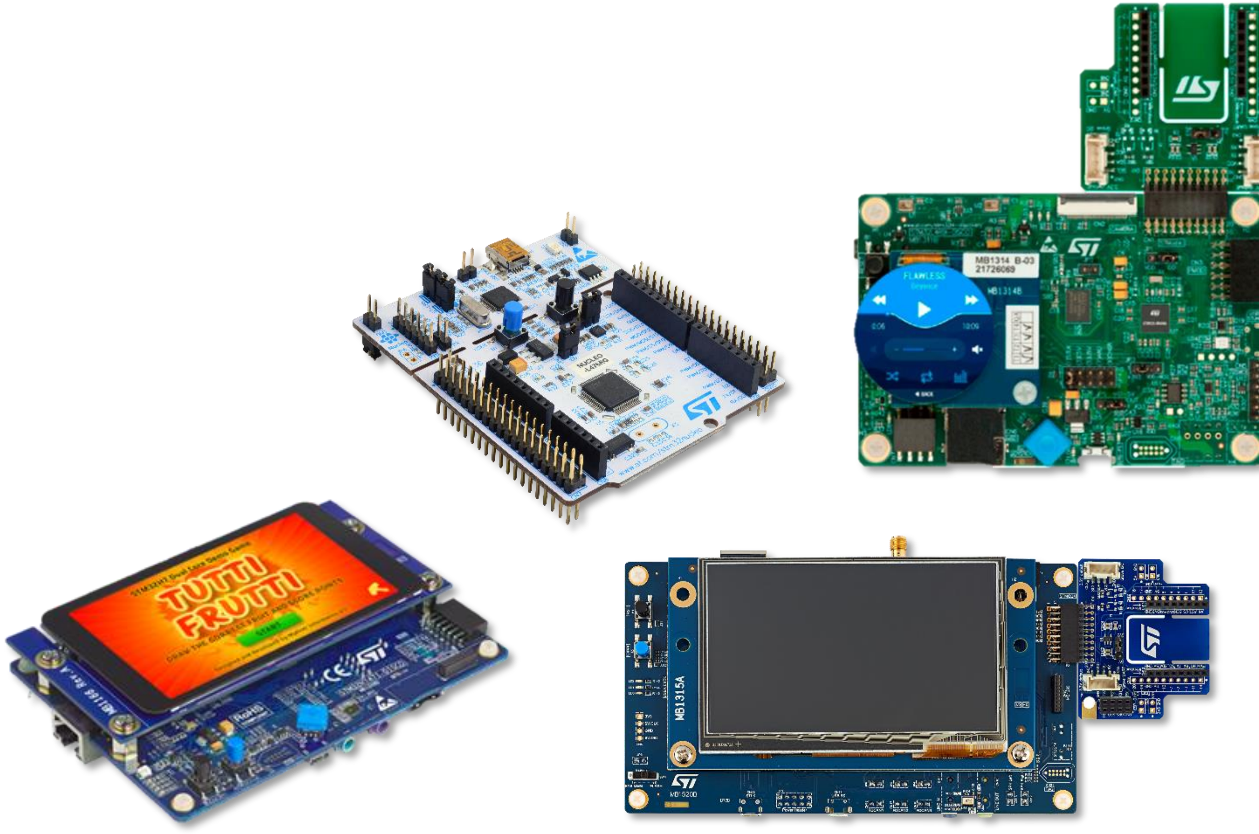
Benchmark models on a large variety of STM32 boards

- Find the most appropriate board for your application

NEW

Get access to the most recent devices

- A board farm is constantly updated with the latest available boards



Start with edge AI optimized models
STM32 model zoo

A collection of application-oriented models optimized for STM32

Human activity

Motion Sensing

Image classification

Computer vision

Audio event detection

Audio classification

Object detection

Computer vision

Hosted on Github

< I >

Model training scripts

- Scripts to generate and validate

Application code example

- Designed to host optimized NN models
- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation

We provide everything to kick off your project

Design documentation

Getting started

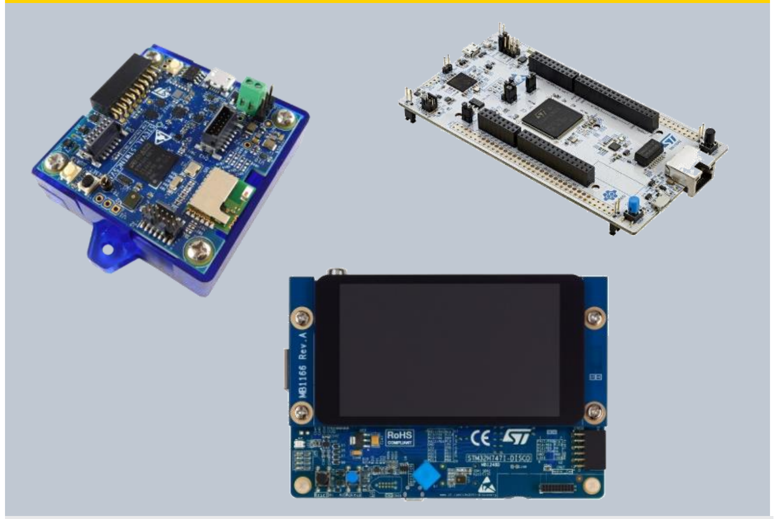
Be guided step-by-step to learn STM32 ecosystem

Development zone

Get started on application development and project sharing

- **Wiki by ST** is a great forum to learn and start developing AI on STM32!
- Videos of application examples
- Massive Open Online Course (MOOC)

Hardware and software tools



- Evaluation platforms for STM32 MCU/MPU
- Extra sensor boards
- Full software suite

Support & Updates

- **ST Community**: STM32 ML & AI group
- Distributor certified FAE
- Support center
- Newsletter

What’s new in STM32Cube.AI v8.0.0?

v8.0.0

Bringing a higher degree of versatility with STM32Cube.AI

# ONNX quantized models' support

Introducing the support of **ONNX Tensor-oriented file format (QDQ):**

- ONNX models quantized with ONNX runtime post-training quantization.

Up-to-date and improved code generation

- **Support for TensorFlow 2.11 models**
- **Support Keras.io 2.11**
- **Support ONNX Runtime 1.13.1**
- New kernel performance improvements.

life.augmented

Making Edge AI accessible to all STM32 portfolio

Take advantage of STM32Cube.AI on all STM32 series

High Perf MCUs			STM32**F2** Up to 398 CoreMark 120 MHz Cortex-M3	STM32**F4** Up to 608 CoreMark 180 MHz Cortex-M4	STM32**F7** 1082 CoreMark 216 MHz Cortex-M7	■ STM32**H7** Up to 3224 CoreMark Up to 550 MHz Cortex -M7 240 MHz Cortex -M4
Mainstream MCUs			STM32**F3** 245 CoreMark 72 MHz Cortex-M4	■ STM32**G4** 569 CoreMark 170 MHz Cortex-M4		***Mixed-signal MCUs***
	■ STM32**C0** 114 CoreMark 48MHz Cortex M0+	STM32**F0** 106 CoreMark 48 MHz Cortex-M0	■ STM32**G0** 142 CoreMark 64 MHz Cortex-M0+	STM32**F1** 177 CoreMark 72 MHz Cortex-M3		
Ultra-low Power MCUs	STM32**L0** 75 CoreMark 32 MHz Cortex-M0+	STM32**L1** 93 CoreMark 32 MHz Cortex-M3	STM32**L4** 273 CoreMark 80 MHz Cortex-M4	STM32**L4+** 409 CoreMark 120 MHz Cortex-M4	STM32**L5** 443 CoreMark 110 MHz Cortex-M33	■ STM32**U5** 651 CoreMark 160 MHz Cortex-M33
Wireless MCUs			■ STM32**WL** 162 CoreMark 48 MHz Cortex-M4 48 MHz Cortex-M0+	■ STM32**WB** 216 CoreMark 64 MHz Cortex-M4 32 MHz Cortex-M0+ ●		

■ Latest product generation

life.augmented

Don't go alone

We have created a network of companies to support you

Partner Program

Trust our **authorized partners** to ensure the success of your project. Learn more at st.com/stm32ai

Wish to discuss a co-development partnership for ML/AI projects? Contact us at edge.ai@st.com

Releasing your creativity

- /STM32
- @ST_World
- community.st.com
- stm32ai.st.com
- wiki.st.com/stm32
- github.com/STMicroelectronics
- Videos
- STM32Cube.AI blog articles

STM32

life.augmented

Our technology starts with You

Find out more at stm32ai.st.com

life.augmented