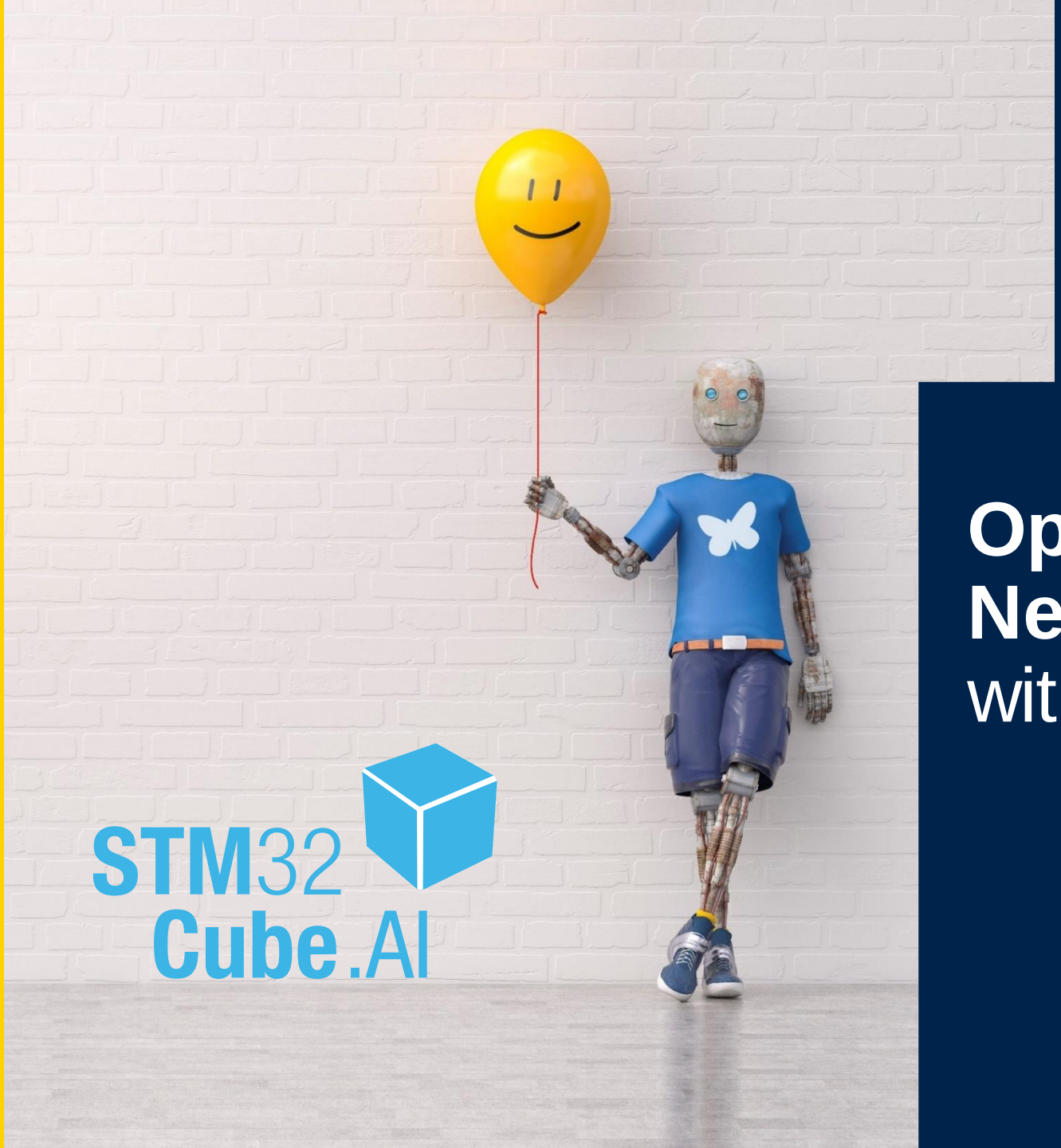


life.augmented

Optimized Neural Networks on STM32 with STM32Cube.AI

STM32

Introduction to Edge AI

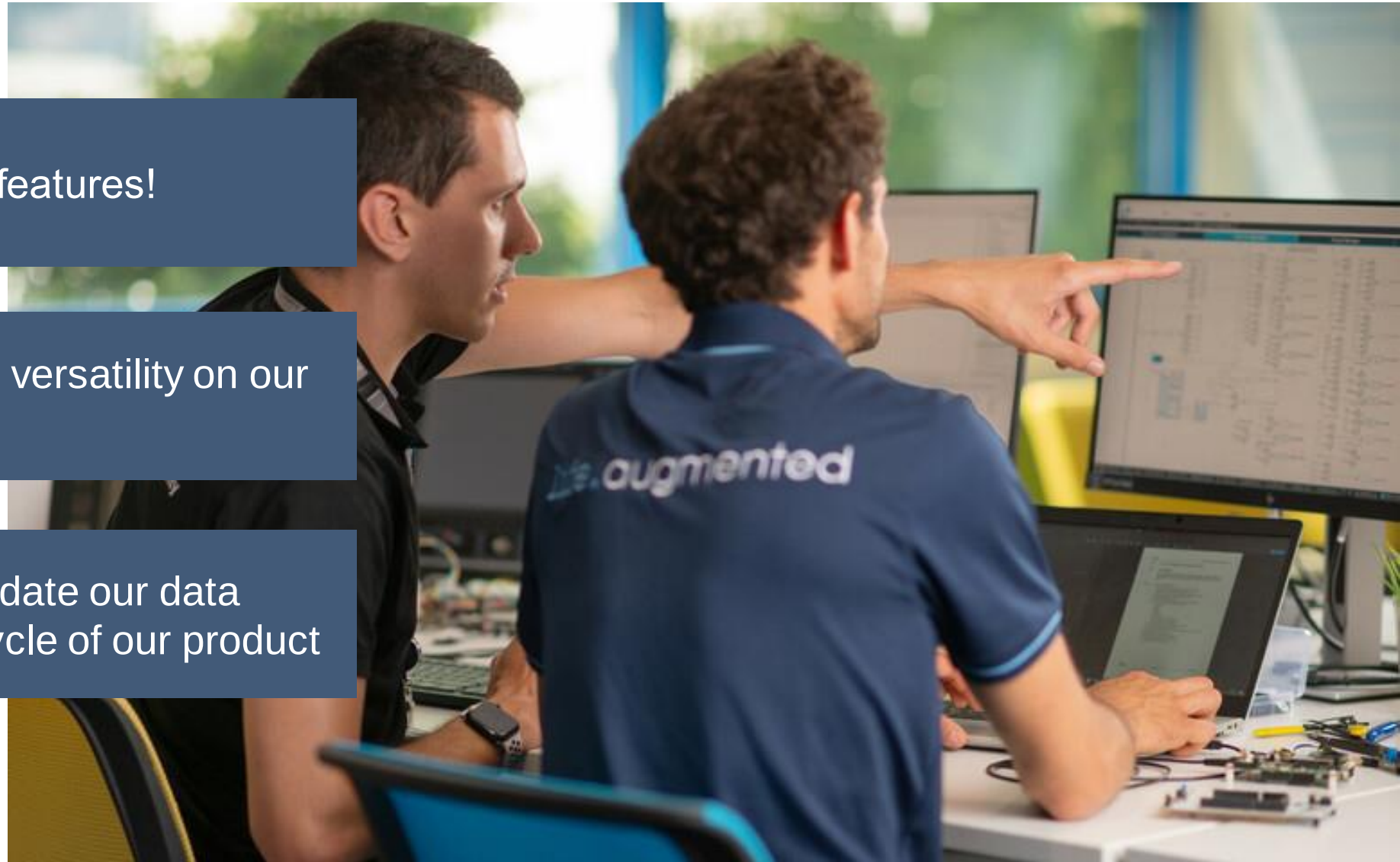
life.augmented

What is the market telling us?

> We're looking for new features!

> We want to have more versatility on our data analysis

> We want to be able update our data analysis over the lifecycle of our product



life.augmented

life.augmented

Product development new paradigm

From rule-based engineering to data-driven engineering

Standard programming
Handcrafted rules based on experience

- Requires digital signal processing skills
- Manual feature extraction?
- Need to rewrite if environment evolves

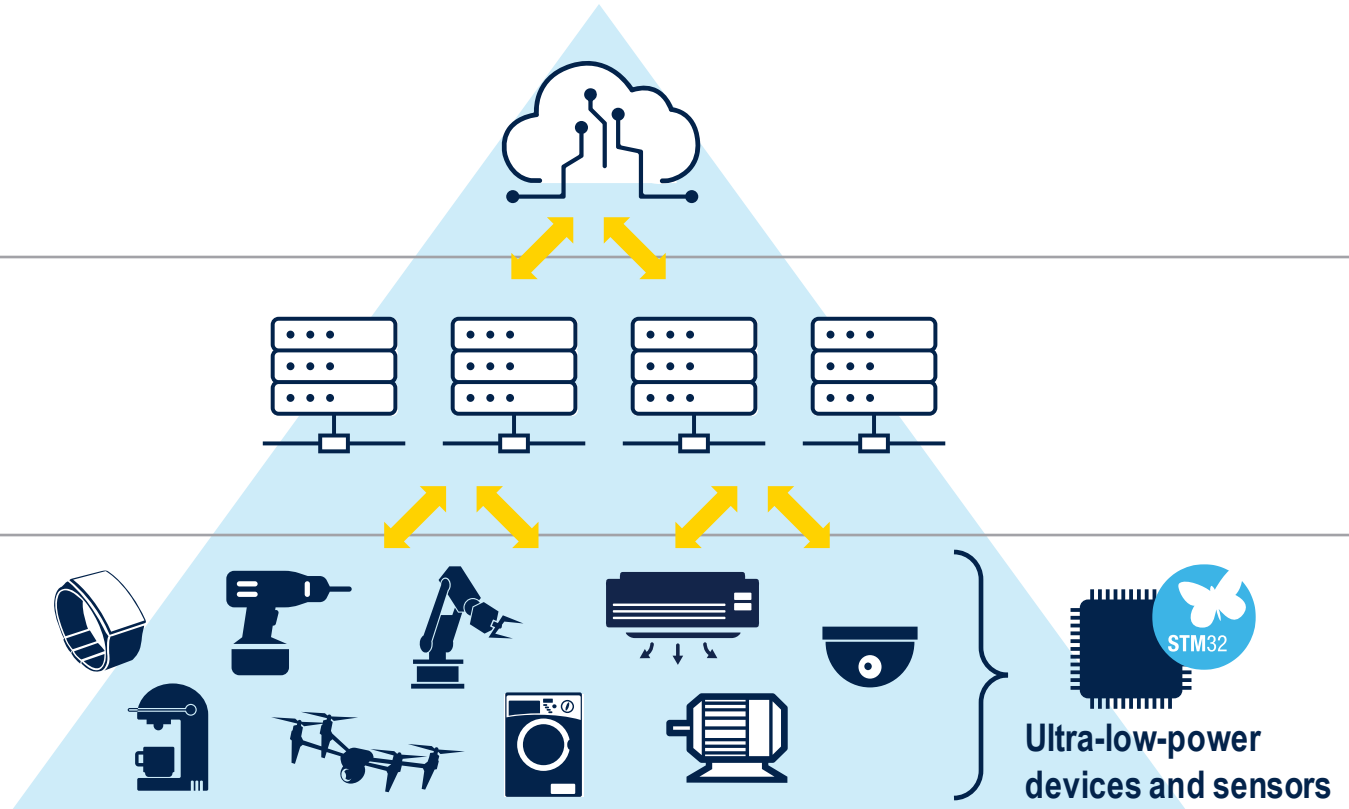
Machine Learning
Rules learnt from real-world data

- Generate code from real-world observations
- Automated feature extraction?
- Re-learn from data if environment evolves

life.augmented

Distributed Artificial Intelligence approach

Leverage billions of devices at the Edge!



Data Center Cloud

Analytics, storage, compute

Edge Nodes

IoT gateways, micro datacenters

Edge Things

Real time, local processing

life.augmented

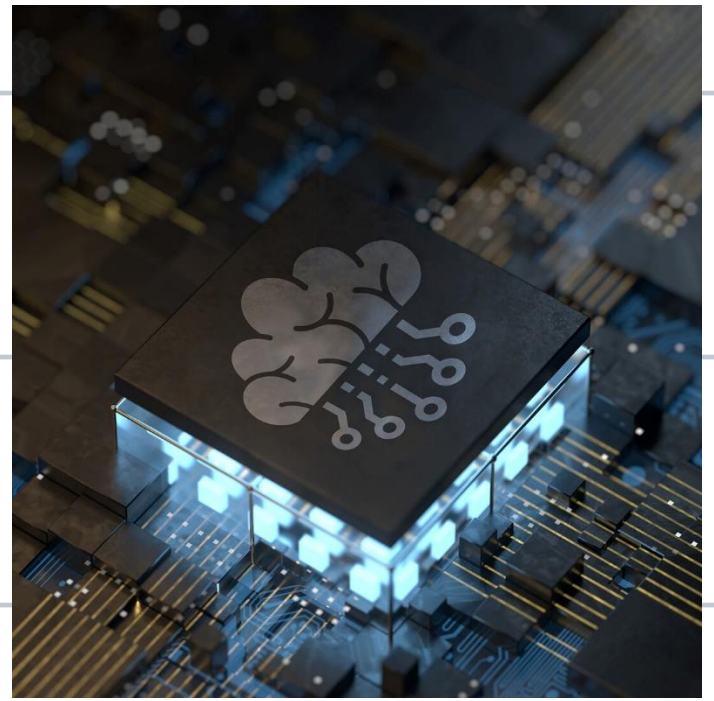
Artificial intelligence at the Edge

Moving part of Artificial Intelligence closer to the data acquisition brings several benefits

Ultra-low latency
Real-time applications

More reliability

Security of data
No sharing in the cloud



Privacy by design
GDPR compliant

Sustainable on energy
Low-power consumption

Better user experience

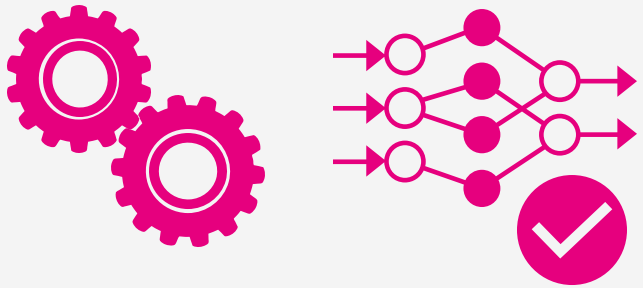
life.augmented

AI development workflow

1 Data Preparation

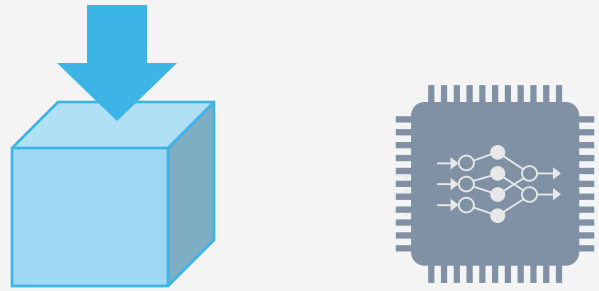
- Data acquisition
- Data processing

2 Data Science



- Model selection and training
- Model testing

3 Model Implementation



- Model library creation
- Model inference

life.augmented

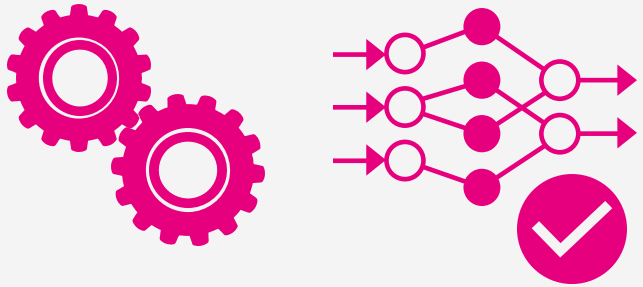
AI development workflow – STM32Cube.AI

1 Data preparation

Data acquisition | Data processing

Data logging and curation tools

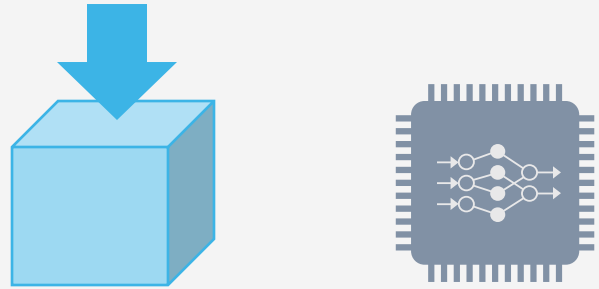
2 Data Science



Model selection and training | Model validation

scikit learn · Keras · TensorFlow Lite · PyTorch · ONNX

3 Model Implementation



Model Optimization | Deployment

Edge AI toolkit

STM32 Cube.AI · STM32

life.augmented

AI development workflow – NanoEdge AI

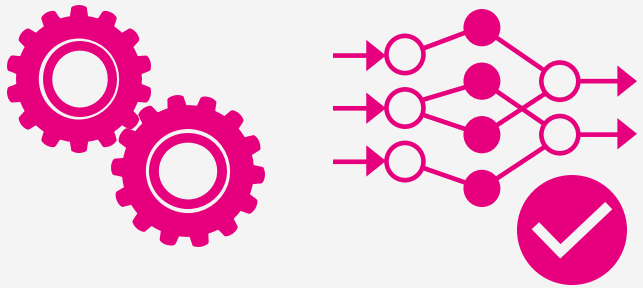
1 Data preparation

Data acquisition

Data processing

Data logging tools

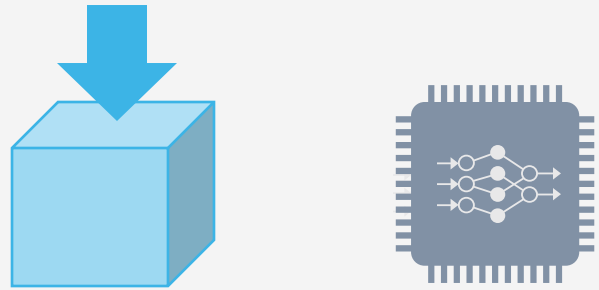
2 Data Science



Model selection and training

Model testing

3 Model Implementation



Model library creation

Model inference

Automated Edge AI Software

NANOEDGE AI STUDIO

STM32

life.augmented

ST ecosystem simplifies your AI development to reach production

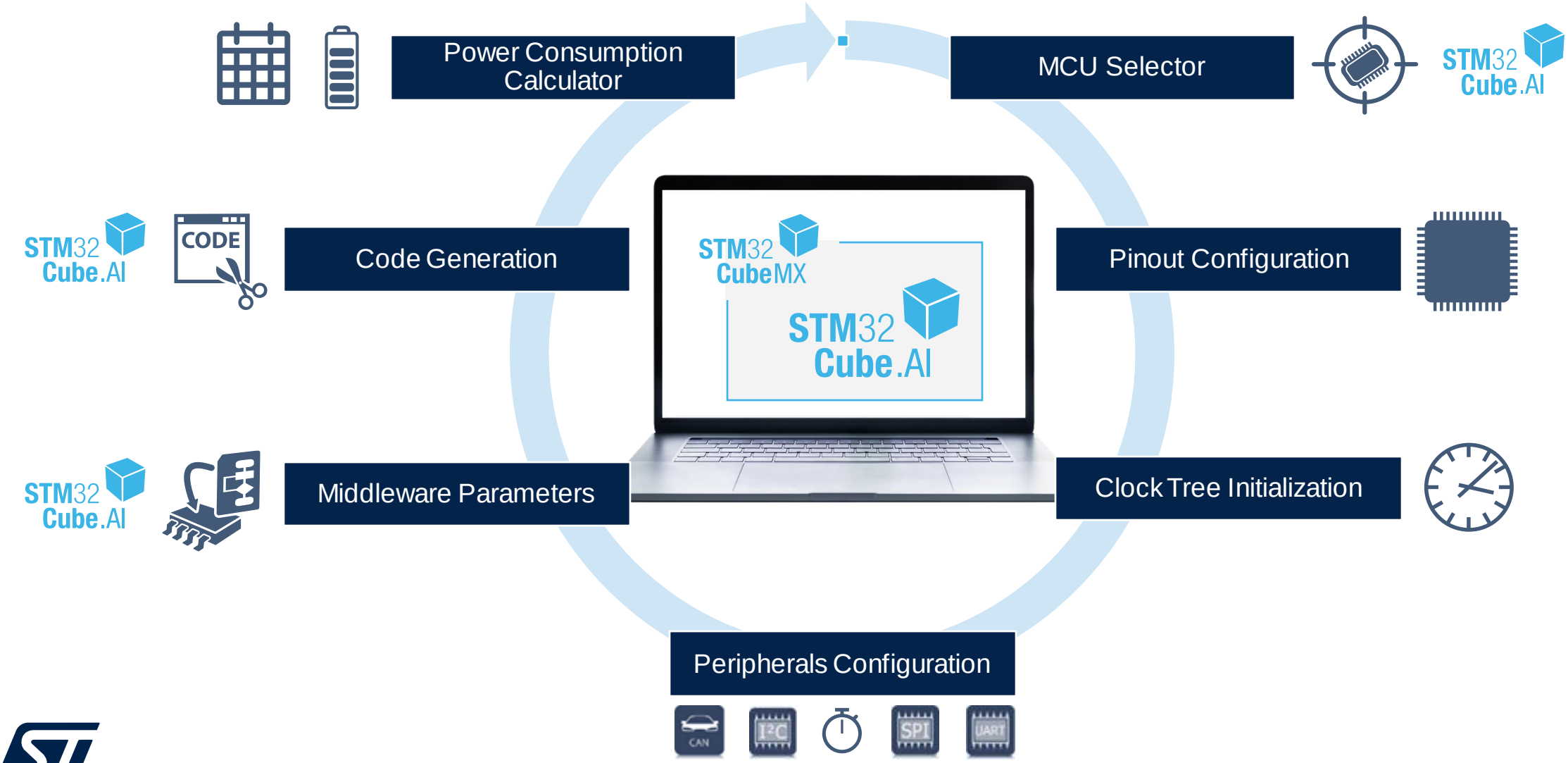
	Edge AI toolkit for model optimization on STM32	**Automated ML Software** for end-to-end Edge AI solution design on STM32
Key benefits	✓ Get from your pre-trained model the best compromise between performance and compression ✓ On device performance validation to identify best STM32 candidate	✓ Save resource and development cost ✓ The easiest way to integrate AI into your system ✓ Reach the highest performance with the Automated model finder
Application domain	Application domain agnostic	All kind of sensing domains (vision and speech not supported)
Business model	Free of charge	License + Royalties

life.augmented

STM32Cube.AI

life.augmented

STM32Cube.AI
The STM32CubeMX expansion pack for ML



life.augmented

A tool to seamlessly integrate AI in your projects

Machine Learning

scikit learn

ONNX

Deep Learning

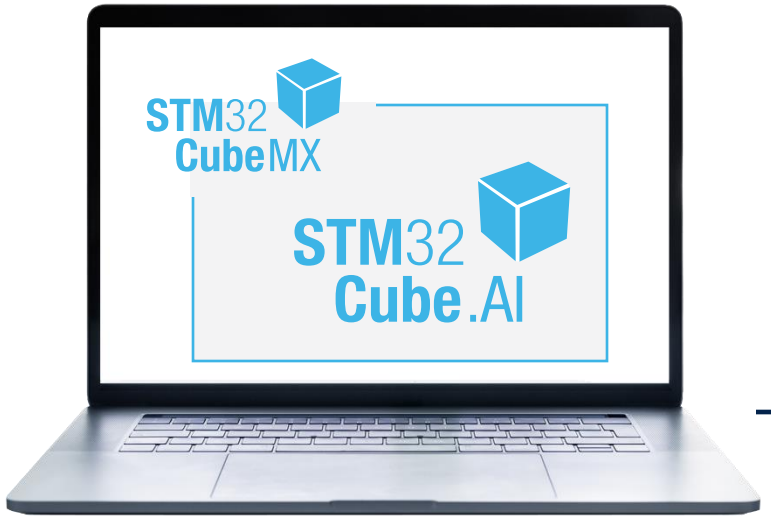
MATLAB

PyTorch

TensorFlow Lite

Keras

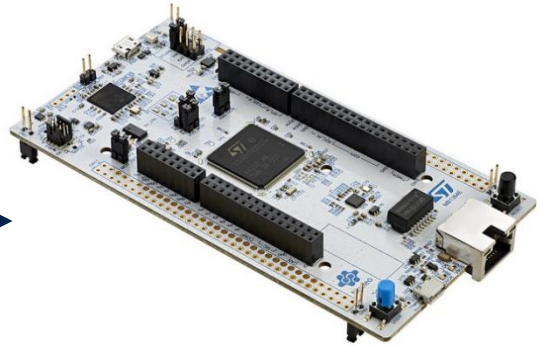
ONNX



Select MCU & upload your model

Optimize and validate

Generate project and deploy



HIGH PERFORMANCE EDGE AI PRODUCT

life.augmented

STM32 comprehensive AI ecosystem

Applications

Applicative Examples
(Function Packs)

Pre and post Processing libraries

Frameworks

NANOEDGE AI STUDIO

scikit learn

MATLAB

PyTorch

TensorFlow Lite

ONNX

Keras

STM32 Cube.AI

AI Model convertor

Graph optimizer

Quantizer

Memory optimizer

Edge Hardware

STM32

STM32 series

Discovery kit

STM32 Nucleo board

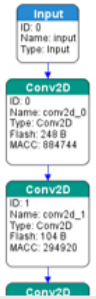
Camera add-on

life.augmented

The 3 pillars of STM32Cube.AI

Graph optimizer

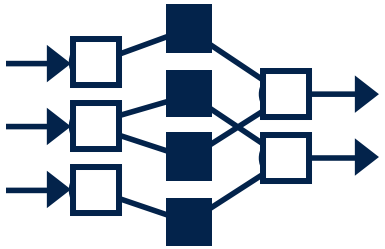
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



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

Quantized model support

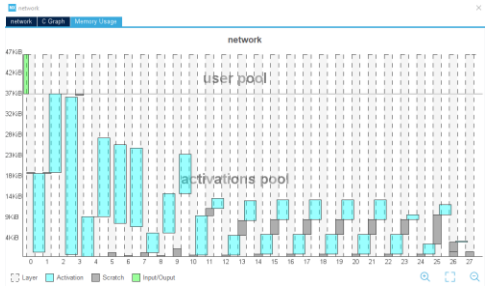
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8
- Minimum loss of accuracy
- Code validation on target
  - Latency
  - Accuracy
  - Memory usage

Memory optimizer

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



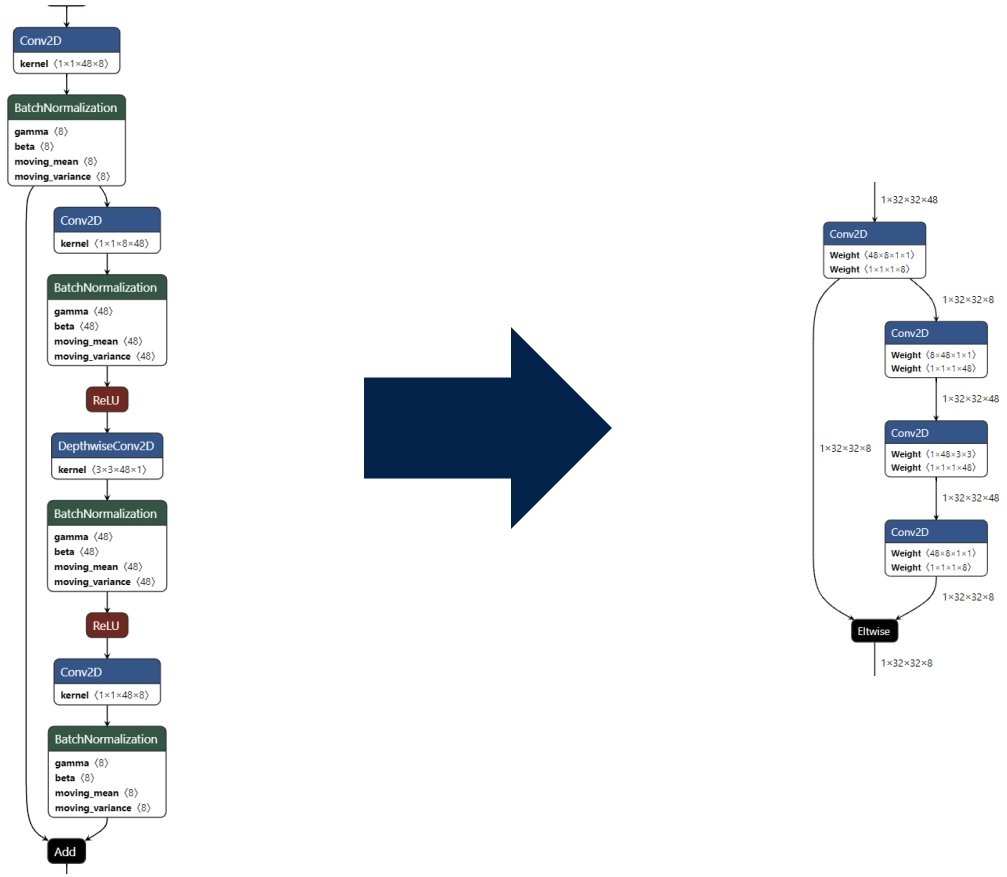
- Memory allocation
- Internal/external memory repartition
- Model-only update option

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

life.augmented

Graph optimizer

Reduce your graph to fit into an MCU!



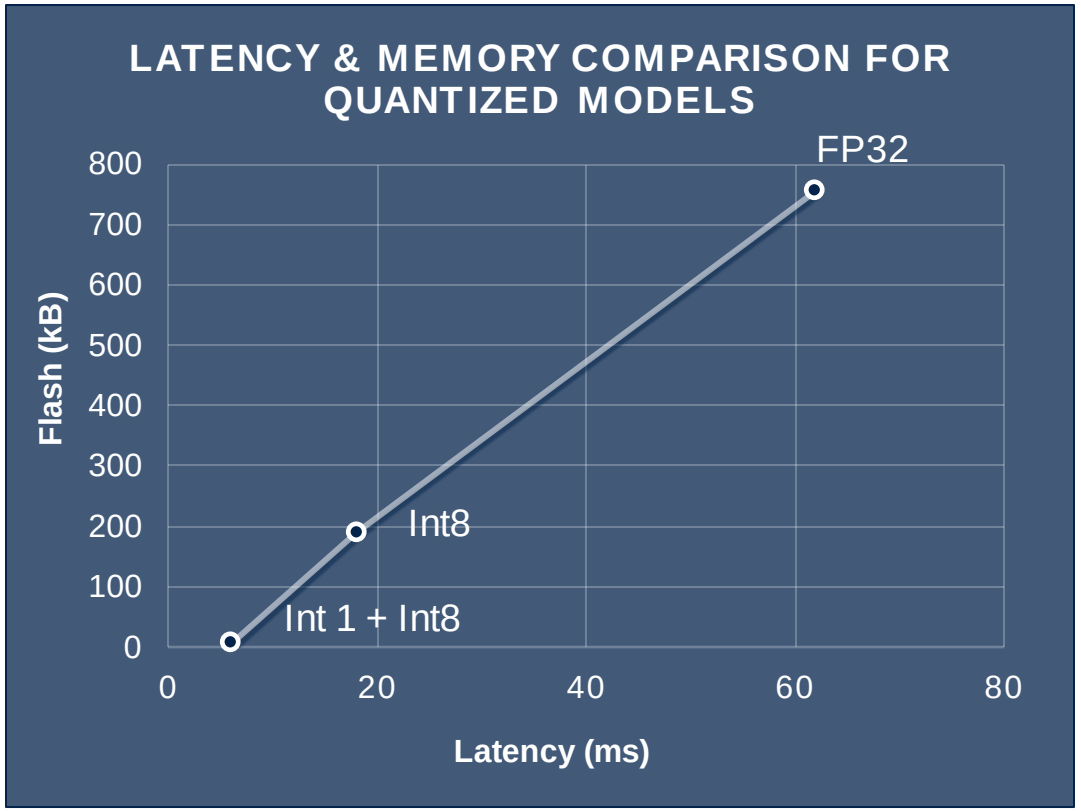
Fully automated process in the STM32Cube.AI workflow

- 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



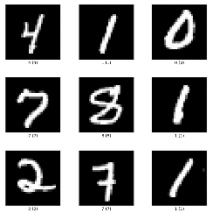
STM32Cube.AI support quantized Neural Network models with **all parameter formats**:

- 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

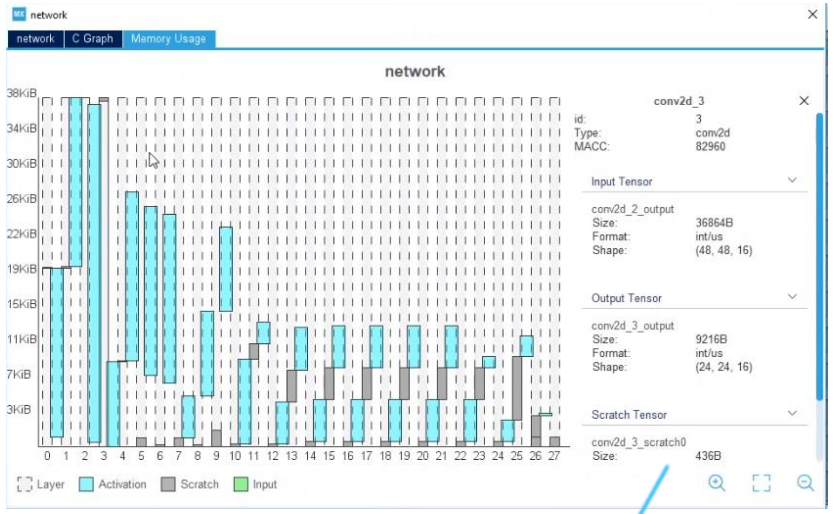


MNIST dataset

life.augmented

Memory optimizer

Optimize performance easily with the memory allocation tool

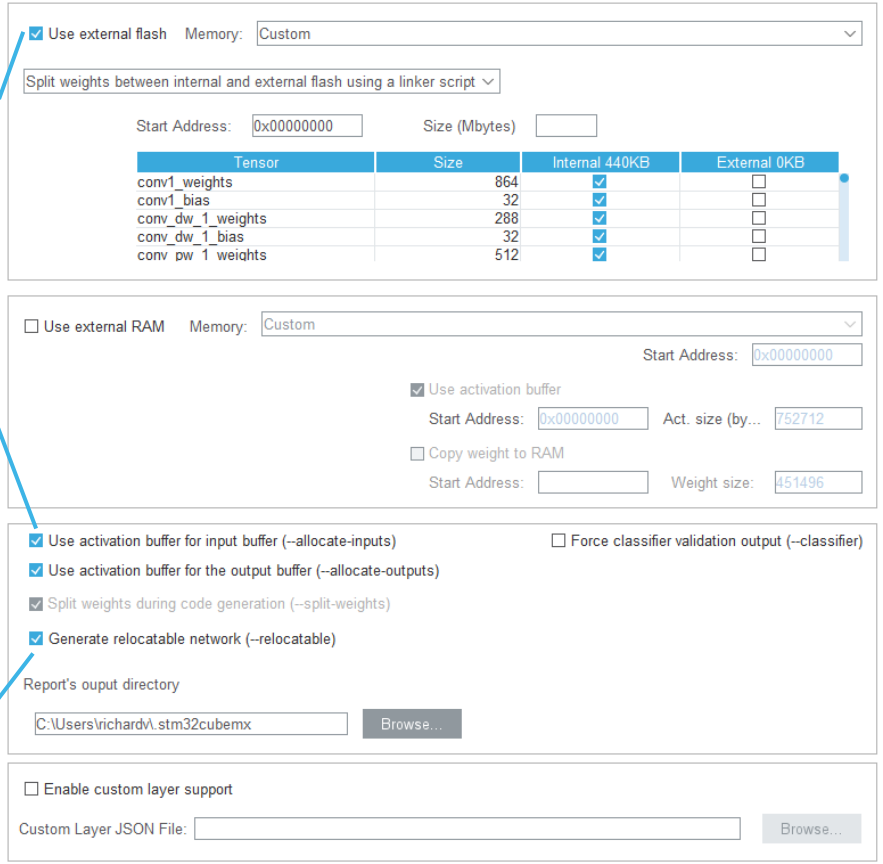


Model RAM consumption per layer

- 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

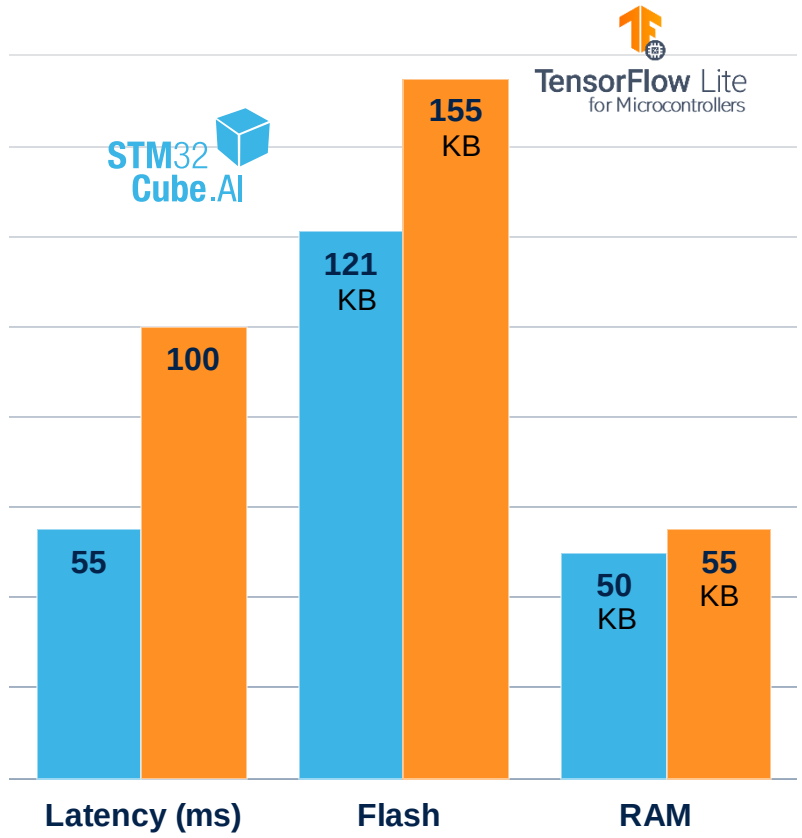
- A separate binary is generated for the library and the network to enable standalone model upgrade

** Requires input and activation buffers in same memory*

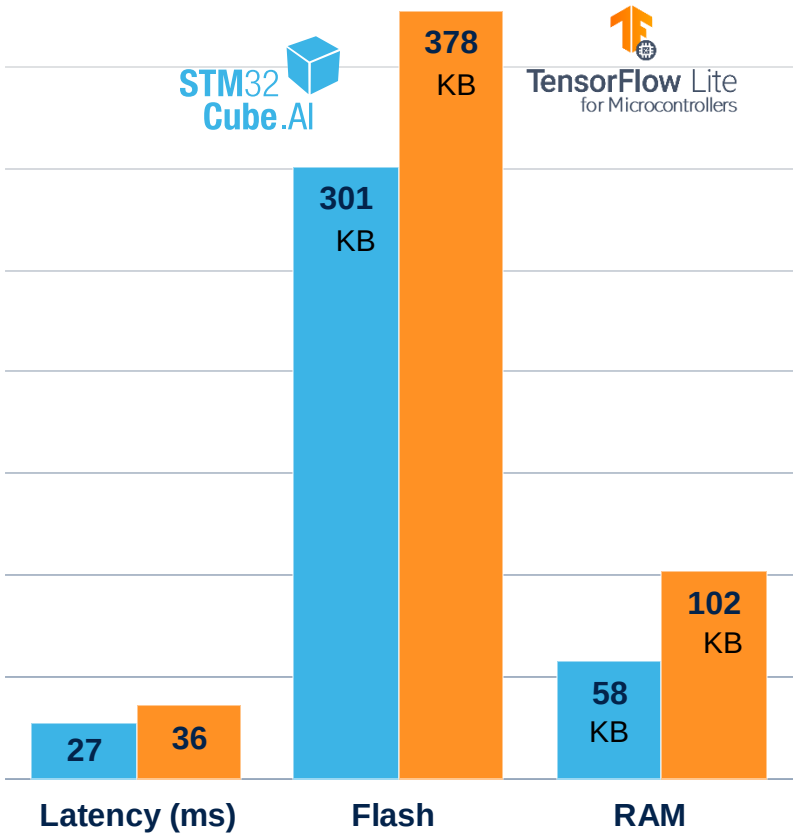
life.augmented

STM32Cube.AI
Get the best performance on STM32

Image Classif v0.7



Visual Wake Word v0.7



HW Target: STM32H723
Flash: 1Mbyte
RAM: 564 Kbytes
Freq: 550 MHz

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

** the lower the better*

life.augmented

Making Edge AI accessible to all STM32 portfolio

STM32Cube.AI is compatible with all STM32 series

MPU						**STM32MP1** 4158 CoreMark Up to 800 MHz Cortex –A7 209 MHz Cortex –M4
High Perf MCUs	STM32**F3** 245 CoreMark 72 MHz Cortex-M4 Optimized for mixed-signal Applications	**STM32G4** 569 CoreMark 170 MHz Cortex-M4	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	**STM32H7** Up to 3224 CoreMark Up to 550 MHz Cortex -M7 240 MHz Cortex -M4
Mainstream MCUs	STM32**F0** 106 CoreMark 48 MHz Cortex-M0	**STM32G0** 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	**STM32U5** 651 CoreMark 160 MHz Cortex-M33
Wireless MCUs			**STM32WL** 162 CoreMark 48 MHz Cortex-M4 48 MHz Cortex-M0+	**STM32WB** 216 CoreMark 64 MHz Cortex-M4 32 MHz Cortex-M0+		

Latest product generation

life.augmented

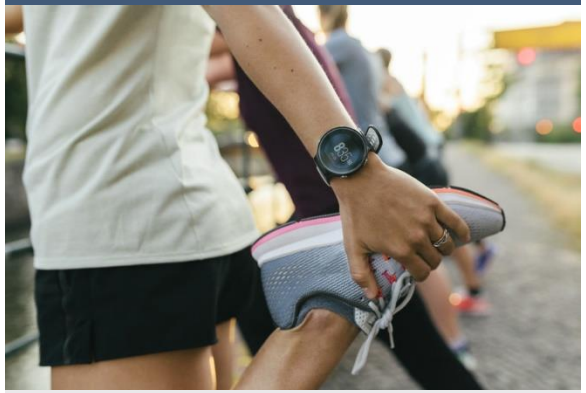
Integrate your ML models more easily with our application-oriented code examples

Time series-based monitoring

FP-AI-MONITOR1

- Predictive maintenance and much more sensor-monitoring apps
- Runs Libraries from NanoEdge™ AI Studio

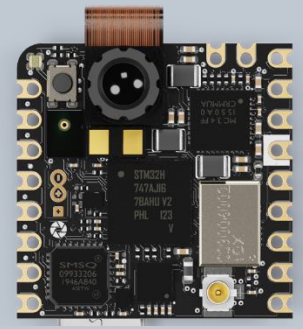
Audio and Sensing



FP-AI-SENSING1

- Human Activity Recognition
- Acoustic Scene Classification
- Data logging, labeling and result on BLE applications

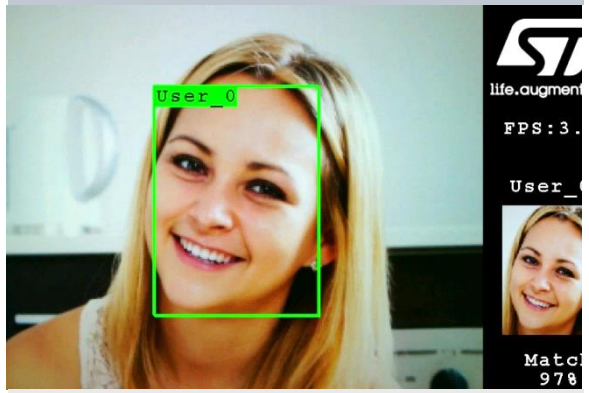
Computer Vision



FP-AI-VISION1

- Food recognition (CNN)
- Person presence detection (CNN)
- People counting (Object detection NN)
- Image processing Library

Face recognition



FP-AI-FACEREC1

- Face detection and recognition
- Fully functional without cloud connection

life.augmented

We provide everything to kick off your project

Design documentation

Getting started

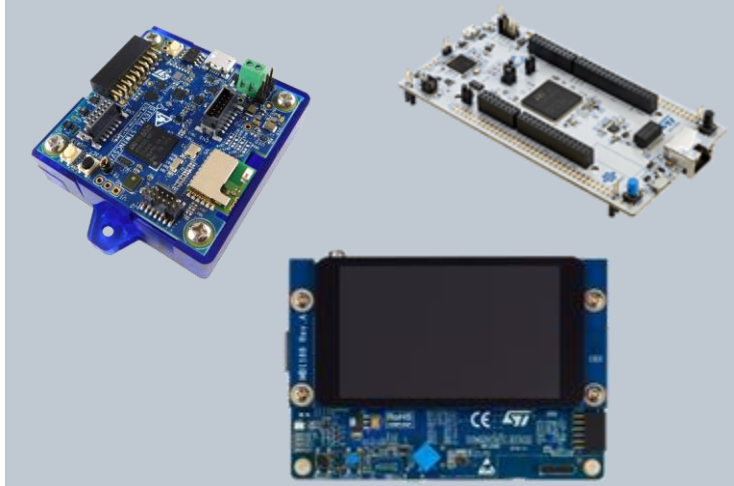
Be guided step-by-step to learn STM32 ecosystem

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

life.augmented

What's new in STM32Cube.AI v7.3.0?

Bringing higher degree of versatility with STM32Cube.AI optimization options

v7.3.0

Neural Network optimization options

Optimization options aim at:

- **Optimizing for *RAM*** to minimize RAM memory footprint
- **Optimizing for *Time*** to obtain the fastest inference processing
- ***Balanced* optimization** to get the best compromise of both.

Up-to-date and improved code generation

- **Support for TensorFlow 2.10 models**
- New kernel performance improvements.

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

life.augmented

Releasing your creativity

/STM32

@ST_World

community.st.com

www.st.com/STM32ai

wiki.st.com/stm32

github.com/STMicroelectronics

Videos

STM32Cube.AI blog articles

life.augmented

Our technology starts with You

Find out more at stm32ai.st.com/stm32-cube-ai/

life.augmented