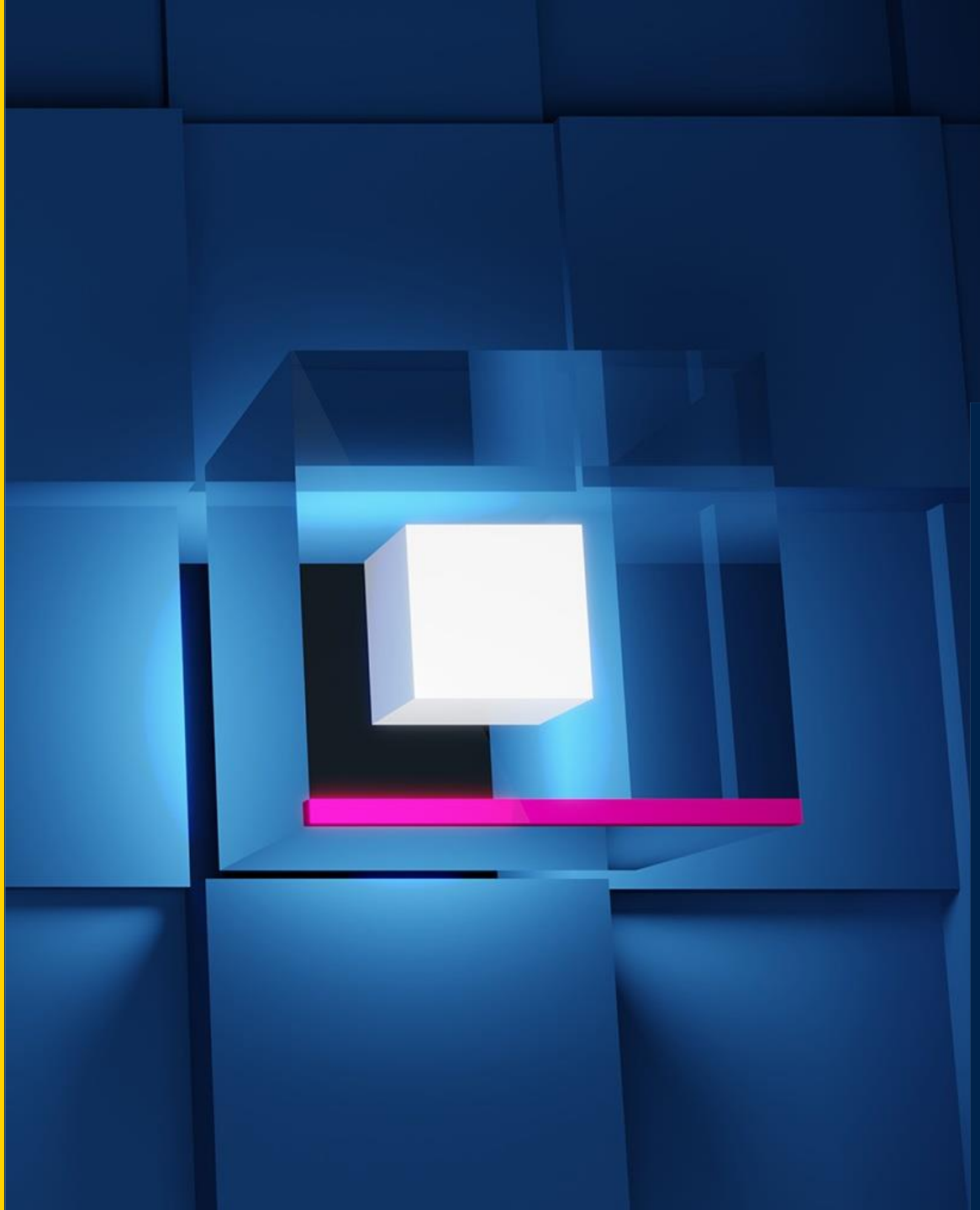


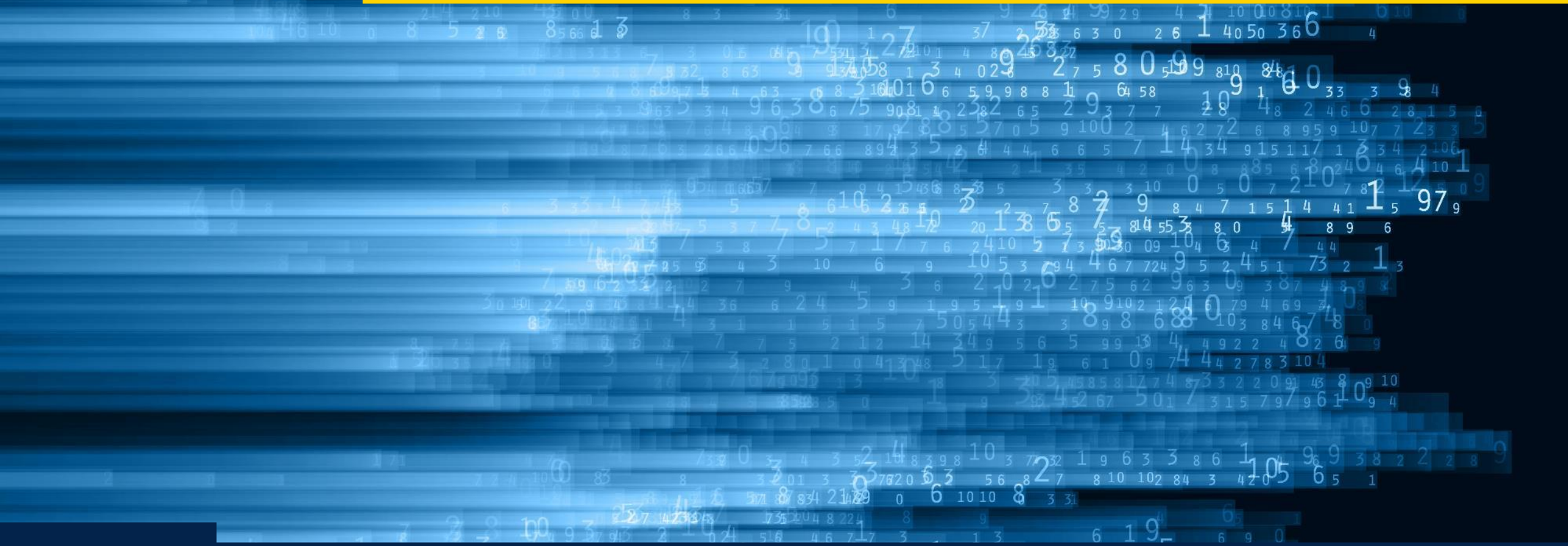
life.augmented

STM32 Enhanced AI Solutions and Ecosystem

STMicroelectronics

문현수 과장

Introduction to artificial intelligence at the edge



life.augmented

Signals turning into data

Embedded applications will collect more data in the future

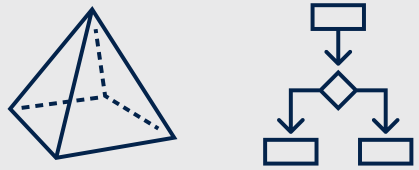
Growing demand for data-driven insights

Increasing use of sensors

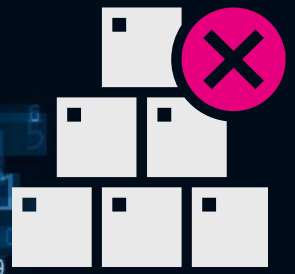
Proliferation of IoT devices

life.augmented

AI offers the best approach to process this growing amount of data



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



Traditional approaches have limitations:

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

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

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

life.augmented

The rise 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

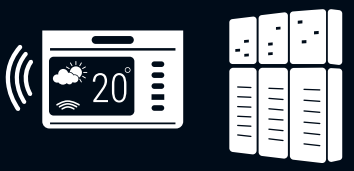
Power efficiency
Low-data / Low-power

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

Edge AI benefits many application domains:

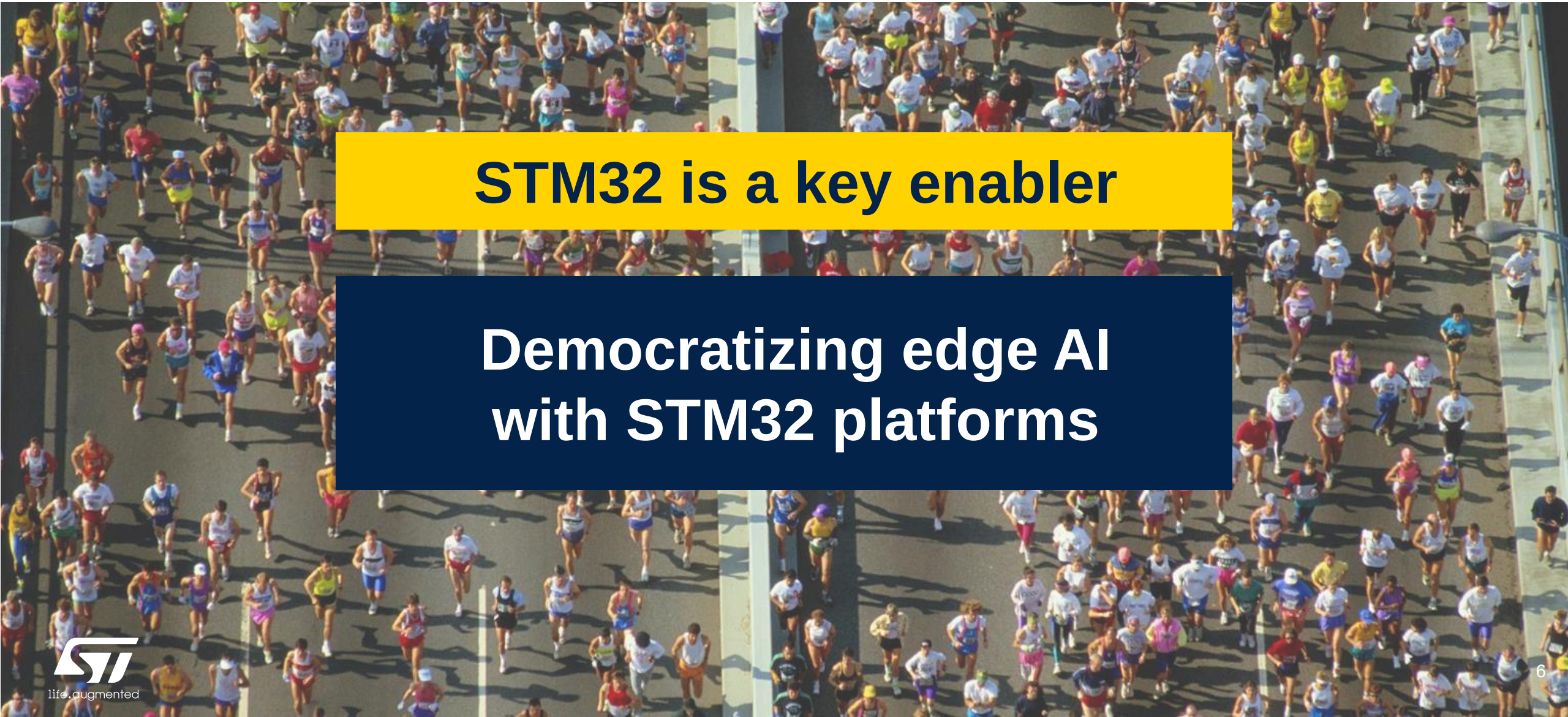
Industrial maintenance
Condition monitoring
Predictive maintenance

Control systems
From home heating systems to industrial machines



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

life.augmented

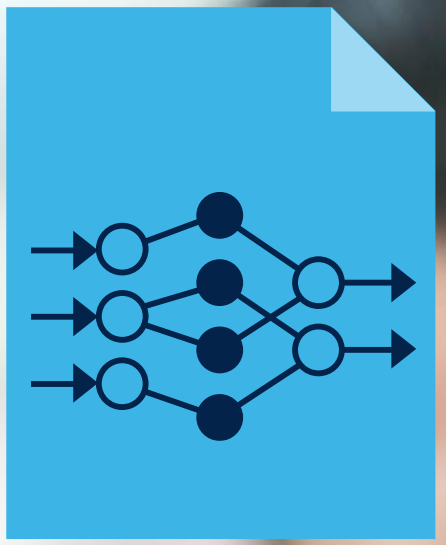


STM32 is a key enabler

Democratizing edge AI with STM32 platforms

life.augmented

STM32 Edge AI



?

Edge AI expertise

eSW development

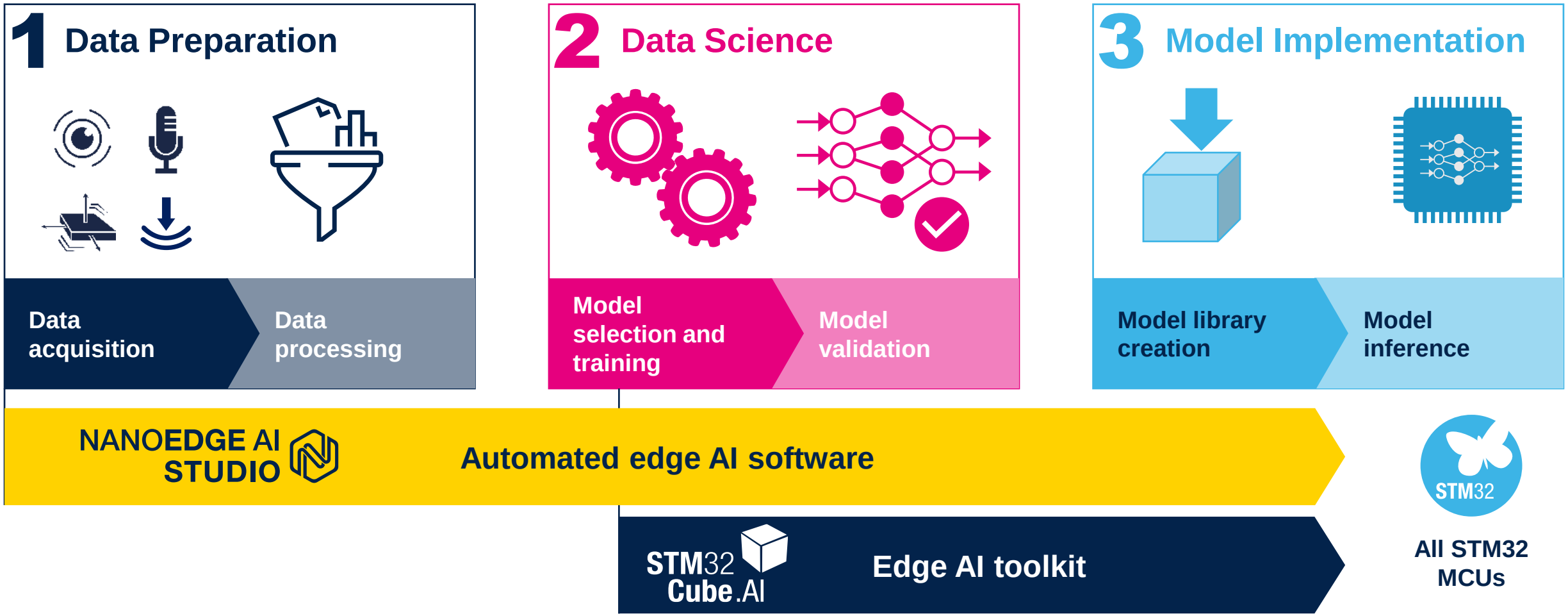
Data

Power consumption

Memory footprint

Inference time

Edge AI development workflow
ST software offering



STM32 product offering simplifies developers' approach to edge AI

- 3 products for embedded developers and data scientists

AD + LOD
SENSING

NANOEDGE AI STUDIO

User-friendly Auto-ML tool for STM32 MCUs

SENSING
AUDIO
VISION

STM32 Cube.AI

AI model optimizer and code generator for STM32 MCUs

SENSING
AUDIO
VISION

X-LINUX-AI

A complete AI framework for STM32 MPUs

- Covering a broad variety of applications

ANOMALY DETEC. + L.O.D.

- Anomaly detection
- Predictive maintenance
- Learning on device

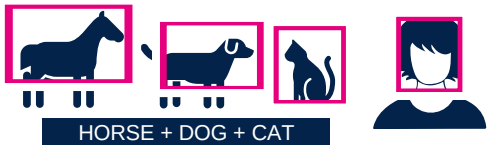
SENSING

- Sensor analysis
- Activity recognition (motion sensors)
- Stress analysis or attention analysis

AUDIO

- Audio (key word, scene detection)
- Speech (sentences) recognition
- Speech synthesis

VISION

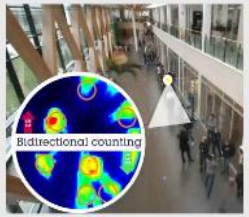


- Multiple object detection
- Face/object analysis (face detection)

life.augmented

stm32ai.st.com

A proven technology already widely adopted



SMART OFFICE | CUSTOMER

People flow counting Sensor with Schneider Electric

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

INDUSTRIAL | DEMO

Aftermarket wireless digit reader

Equip meters with aftermarket wireless & low-power readers.

SMART CITY | DEMO

Acoustic scene classification

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

INDUSTRIAL | DEMO

People presence detection (visual wake word)

Human detection on high-performance MCU.

SMART HOME | DEMO

Floor type detection for vacuum cleaners

Advanced solution for material recognition of floor type (hard or soft) enabled by AI technology.

INDUSTRIAL | DEMO

Fan anomaly classification based on ultrasound analysis

Neural Network classification based on a high-frequency analog microphone pipeline.

SMART BUILDING | DEMO

Hand posture recognition without camera module

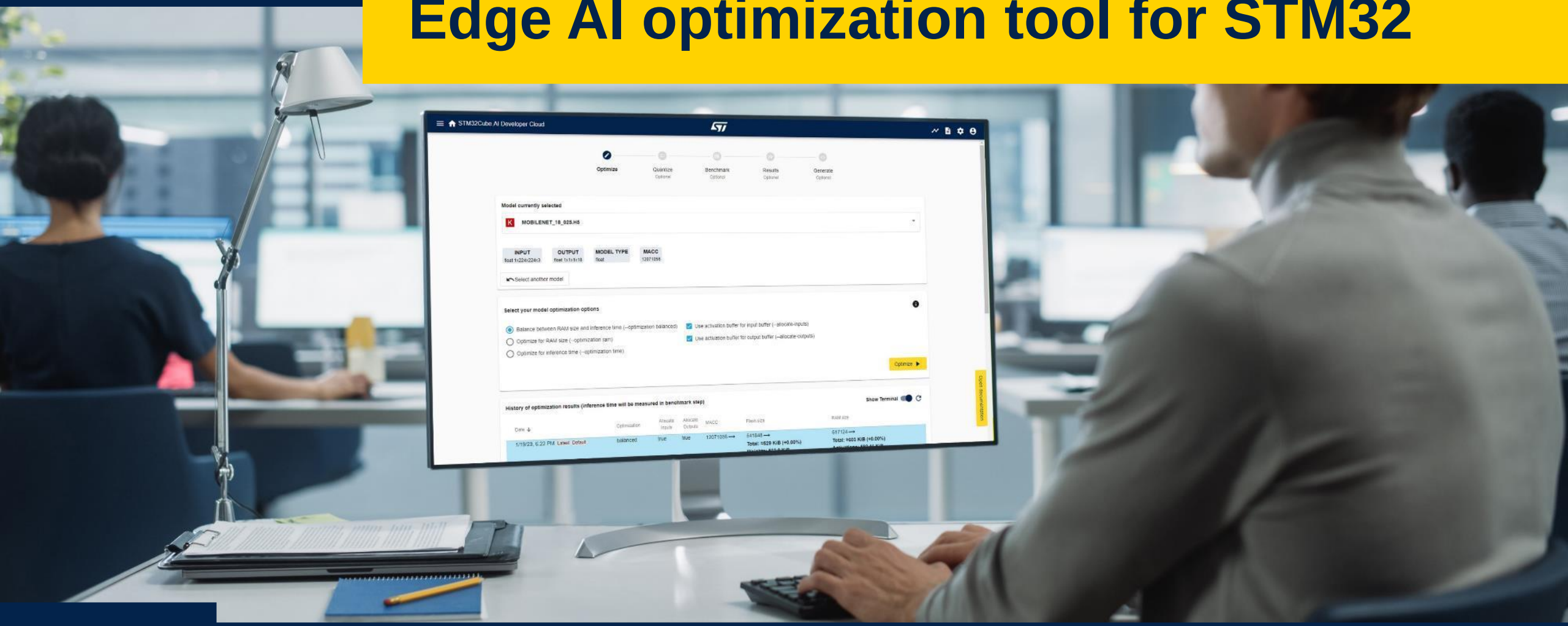
Hand posture recognition running on STM32F401 based on ST multizone Time-of-Flight ranging sensor.

WEARABLES | DEMO

Human Activity Recognition

Easily identify 5 different activities with a 3D accelerometer.

STM32Cube.AI
Edge AI optimization tool for STM32



life.augmented

STM32Cube.AI solutions

X-CUBE-AI
STM32CubeMX extension package

Developer Cloud
Online version of the tool

STM32 board farm
Online performance benchmarking facilities

STM32 Cube.AI

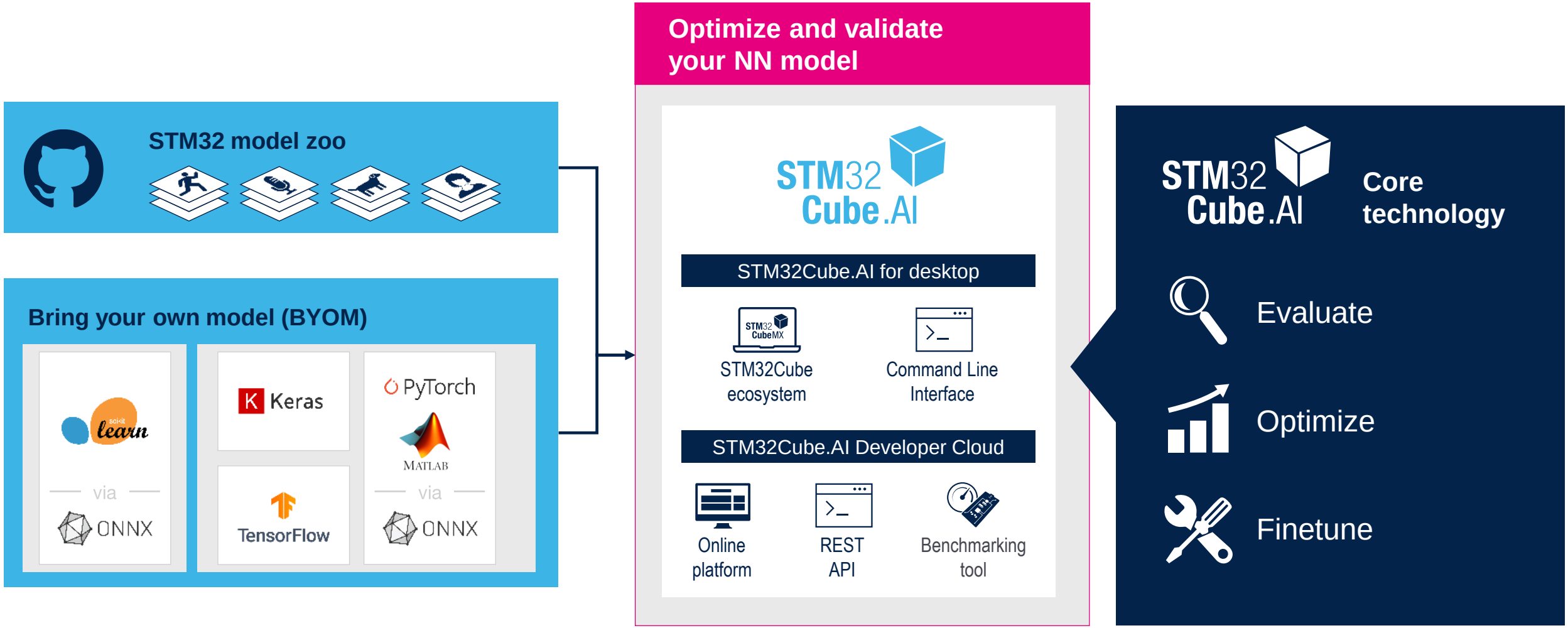
STM32 model zoo
Dedicated GitHub repo

Getting started
Tutorial + code examples

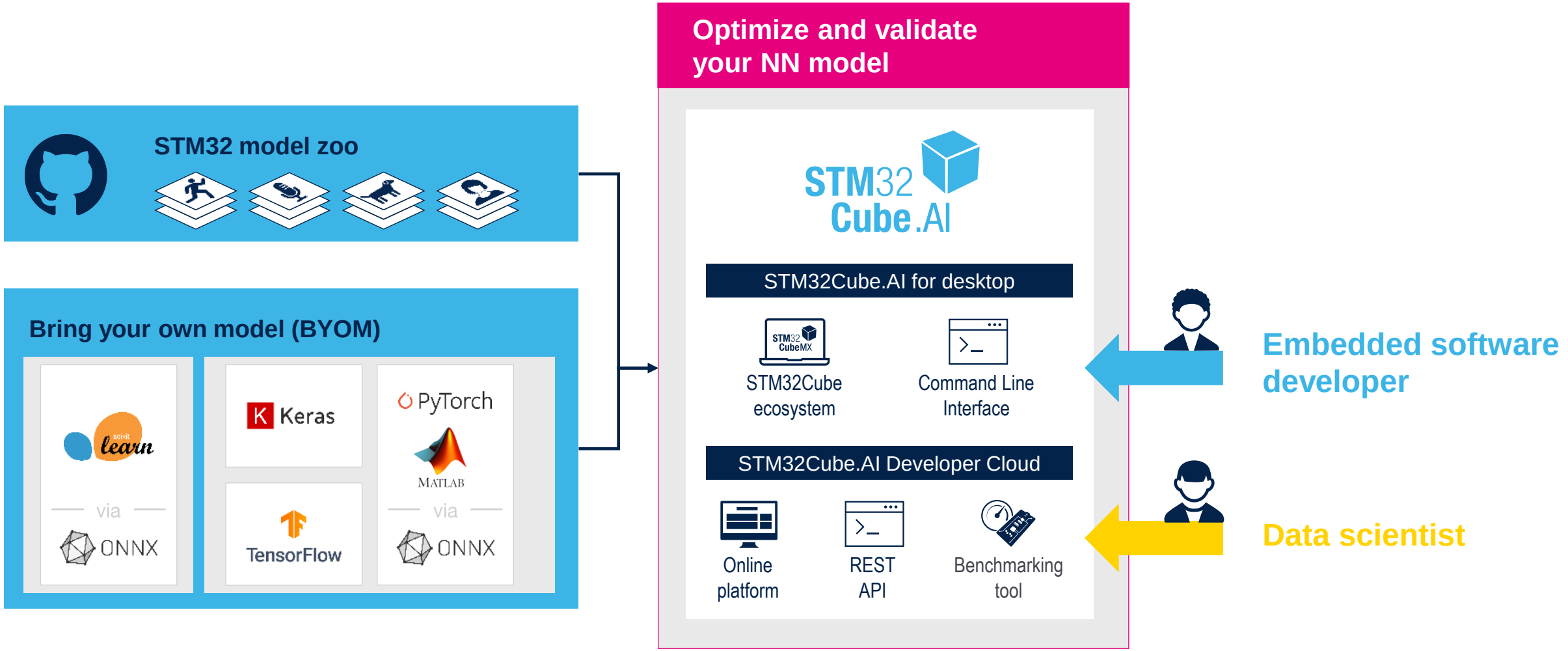
Partners
Dedicated partner ecosystem

life.augmented

STM32Cube.AI core technology



Two versions of the same tool depending on your profile



life.augmented

STM32 benchmarking tool

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

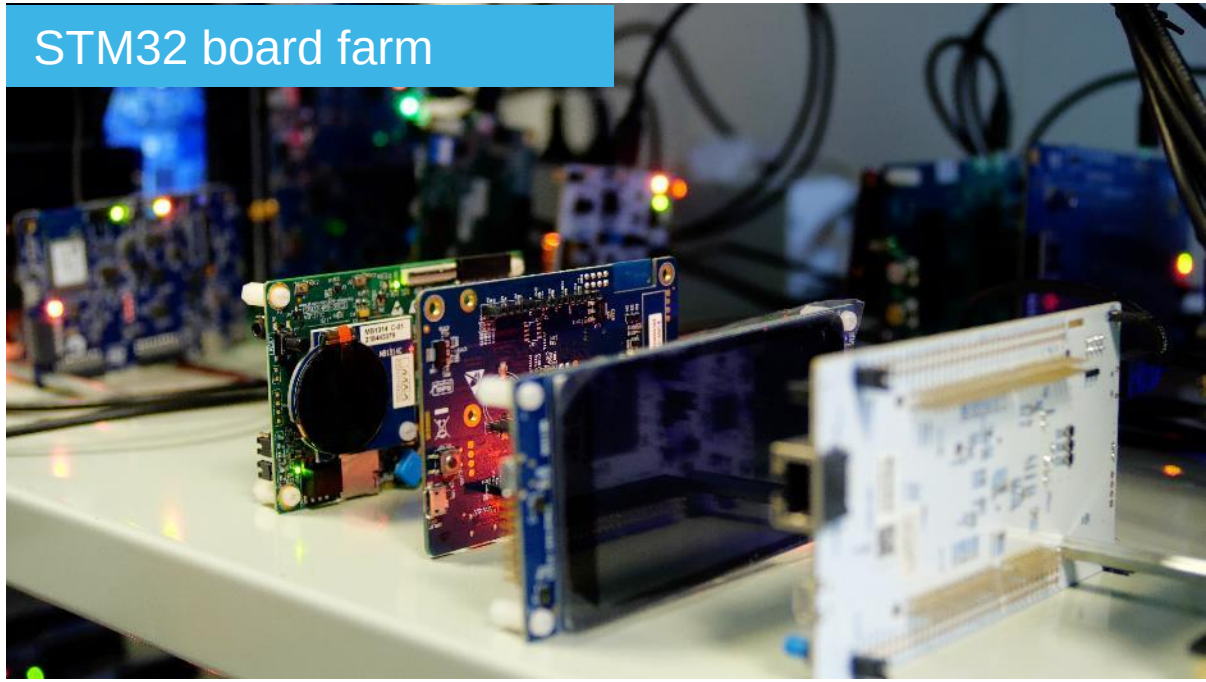
Benchmark models on a large variety of STM32 boards

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



life.augmented

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

Model training scripts

- Scripts to generate and validate

Getting started application packages

- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation

life.augmented

STM32Cube.AI can run on all STM32 series

Five product categories

Wireless MCU

Short- and long-range connectivity

Ultra-low-power MCU

Mainstream MCU

High-performance MCU

32-bit general-purpose microcontrollers: from 75 to 3,224 CoreMark score

Embedded MPU

32- and 64-bit microprocessors





Enabling edge AI solutions

Scalable security

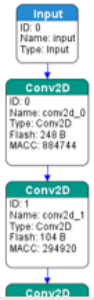
MPU portfolio

MCU portfolio

The 3 pillars of STM32Cube.AI

Graph optimizer

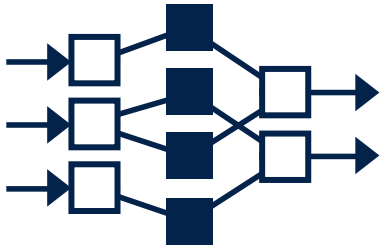
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



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

Quantized model support

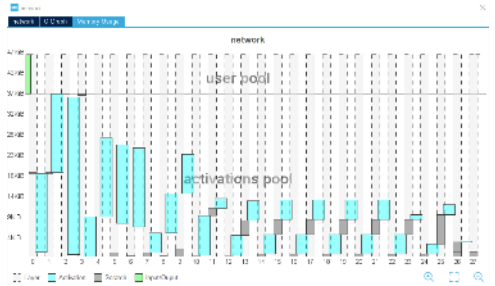
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- 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



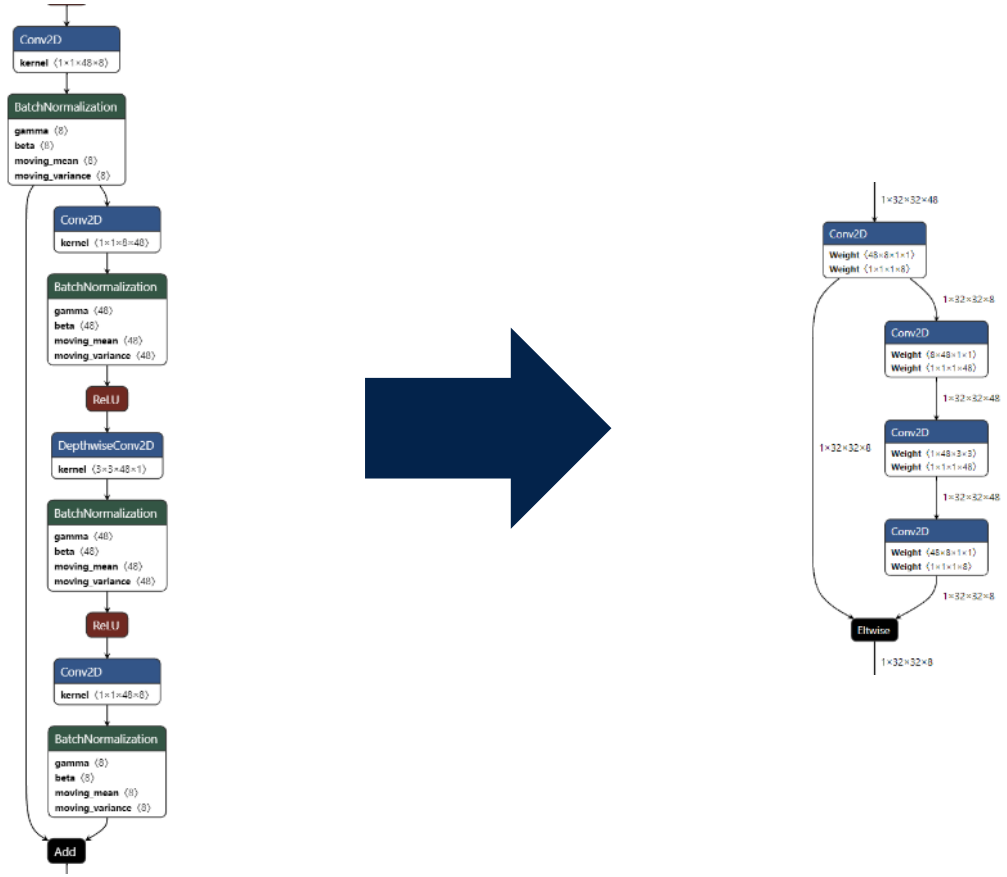
- Memory allocation
- Internal/external memory repartition
- Model-only update option

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

life.augmented

Graph optimizer

Squeeze your graph to fit into an MCU!



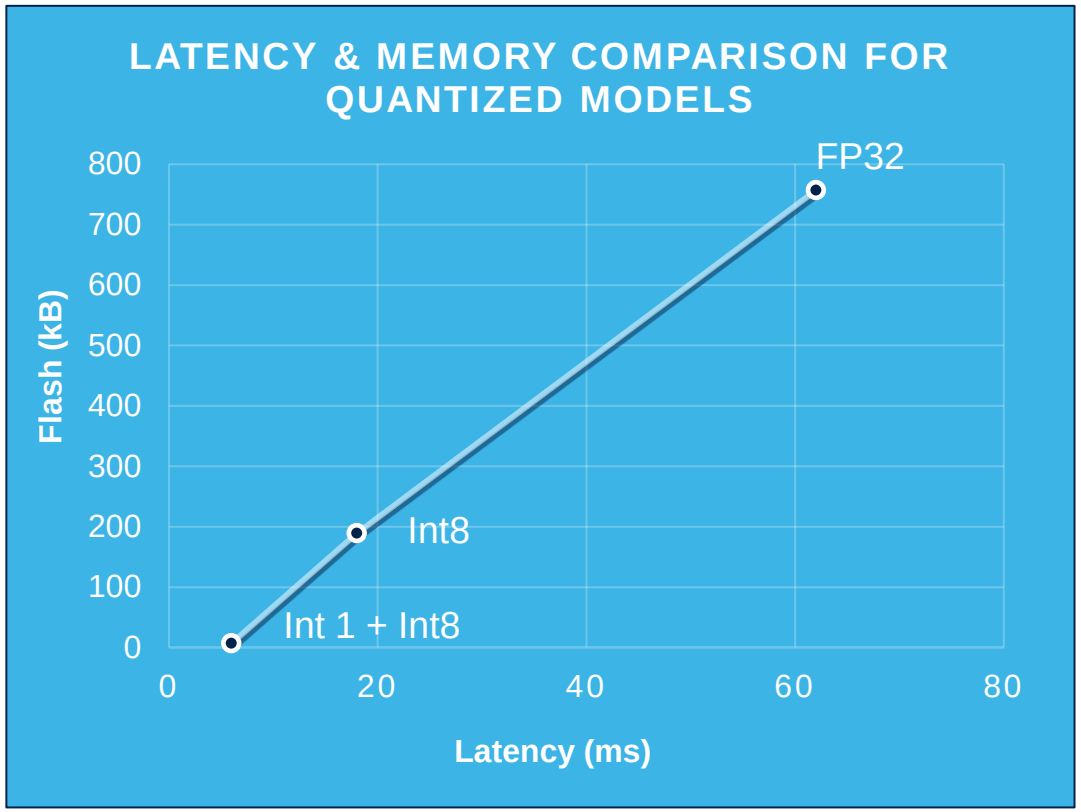
Fully automated process in the STM32Cube.AI workflow

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

life.augmented

Quantized model support

Simply use quantized networks to reduce memory footprint and inference time



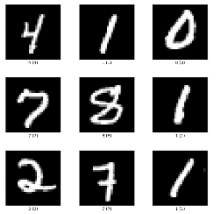
STM32Cube.AI supports quantized neural network models with **all parameter formats**:

- 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

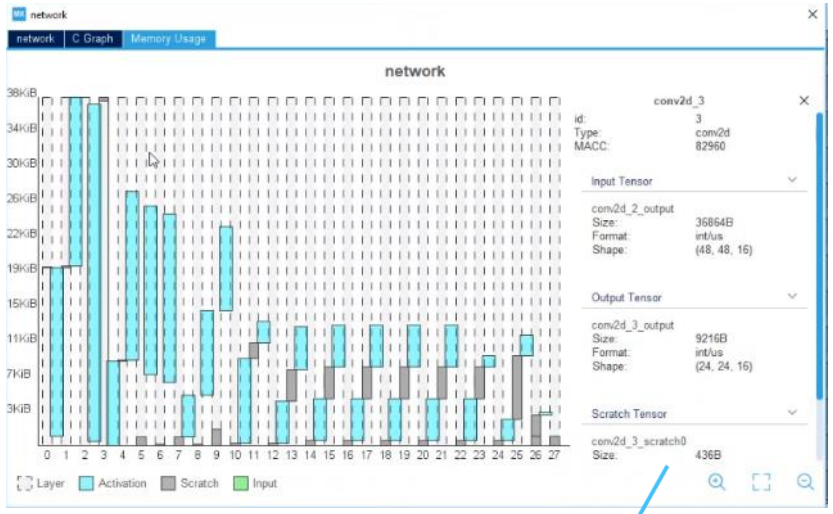


MNIST dataset

life.augmented

Memory optimizer

Optimize performance easily with the memory allocation tool



Model RAM consumption per layer

- Easily identify the most critical layers

Model memory allocation

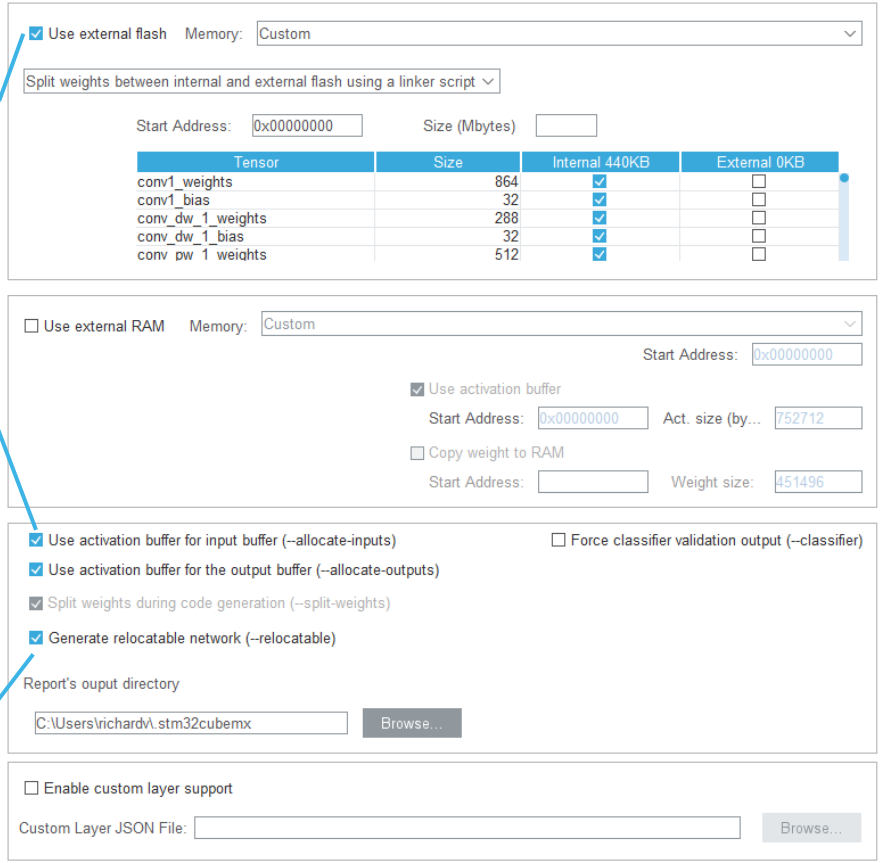
- Set your external memory
- Map in non-contiguous internal flash section
- Partition internal vs external flash memories

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

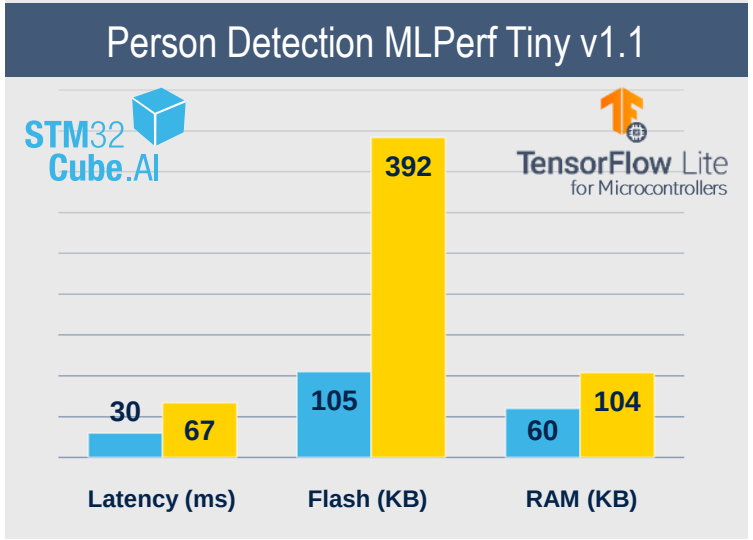
What's new in STM32Cube.AI v8.1.0?

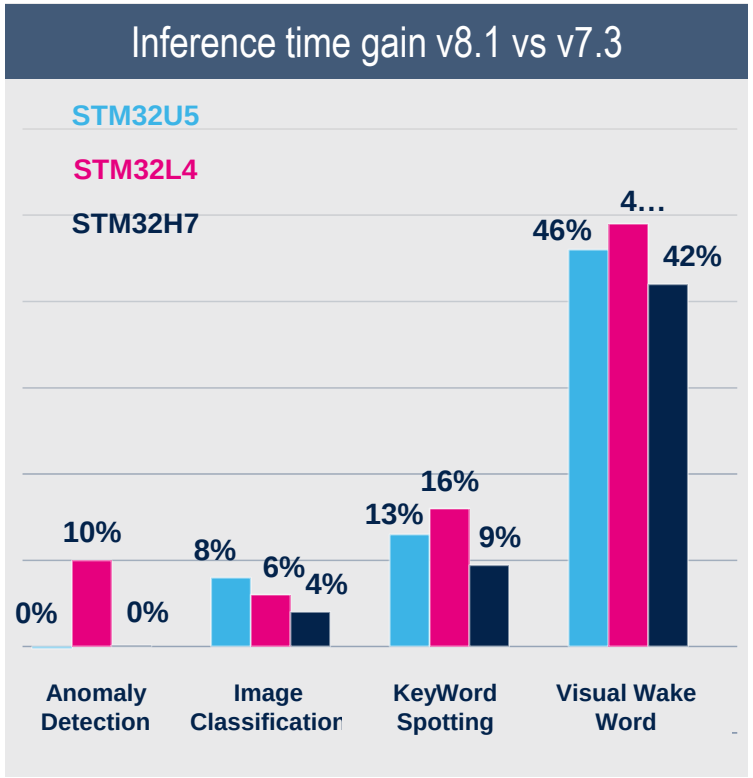
Pushing STM32Cube.AI performance further

- **Improved performance** in NN models optimization and generation
- **Improved support** of ONNX features and operators
- **Implemented** unsupported layers / operators

UP TO
60 %
faster inference time*
a

UP TO
70 %
space freed-up in FLASH and RAM*



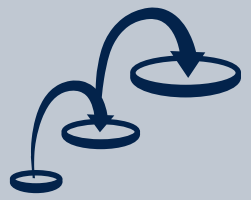


* versus TensorFlow Lite for microcontroller using STM32H7A3

life.augmented

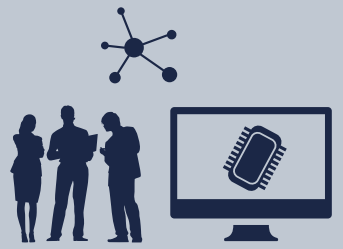
We provide everything to kick off your project

Design documentation



Getting started

We guide you step-by-step to start with the STM32 ecosystem

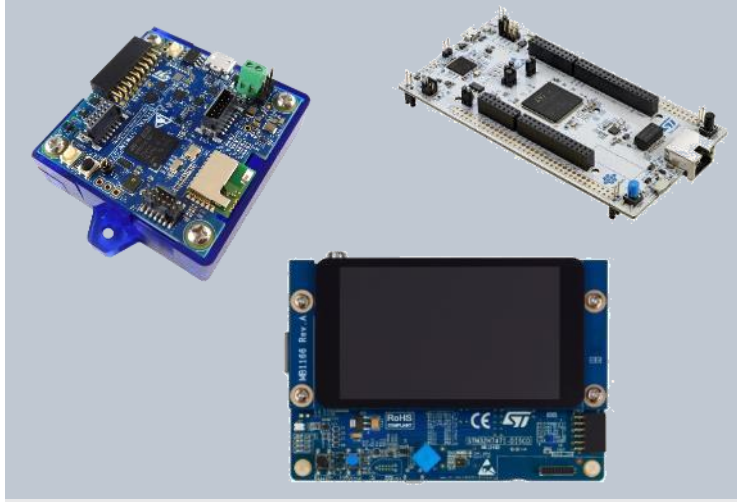


Development zone

Get started on application development and project sharing

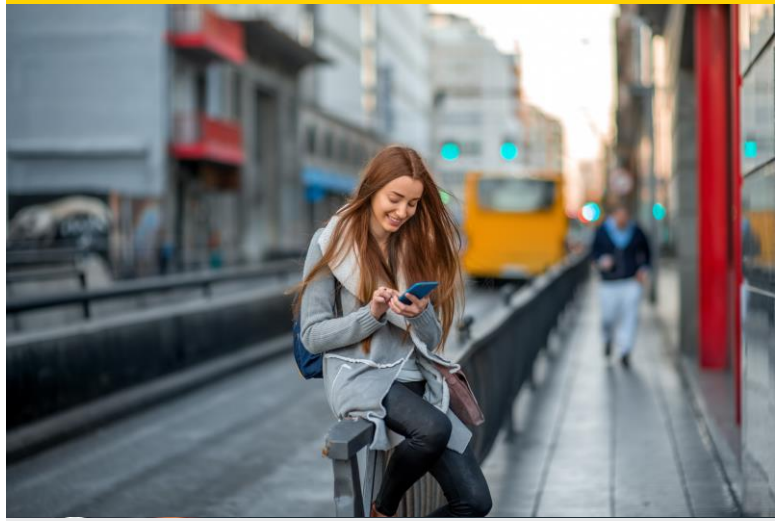
- **Wiki by ST** is a great forum to learn and start deploying edge AI on STM32!
- Videos of application examples
- Massive open online courses (MOOCs)

Hardware and software tools



- Evaluation platforms for STM32 MCUs and MPUs
- Additional sensor boards
- Full software suite

Support and updates



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

life.augmented

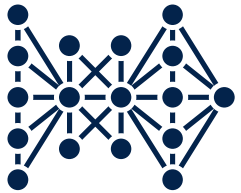
Introducing STM32Cube.AI Developer Cloud

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

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

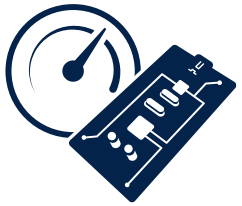
X-CUBE-AI for STM32Cube.MX

X-CUBE-AI Command Line Interface



ST Model Zoo

Web GUI + REST API

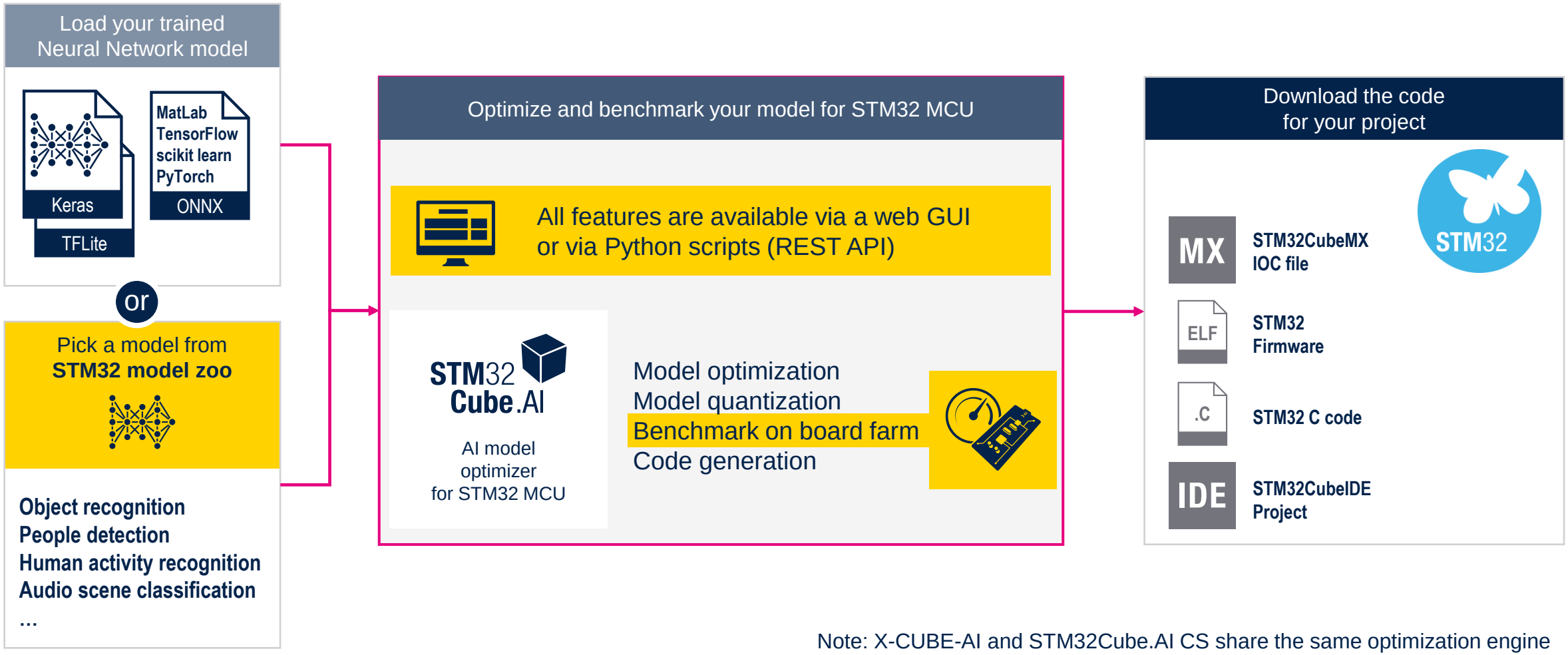


Board farm

STM32Cube.AI **Core engine technology**

life.augmented

Seamlessly integrate AI in your STM32 projects

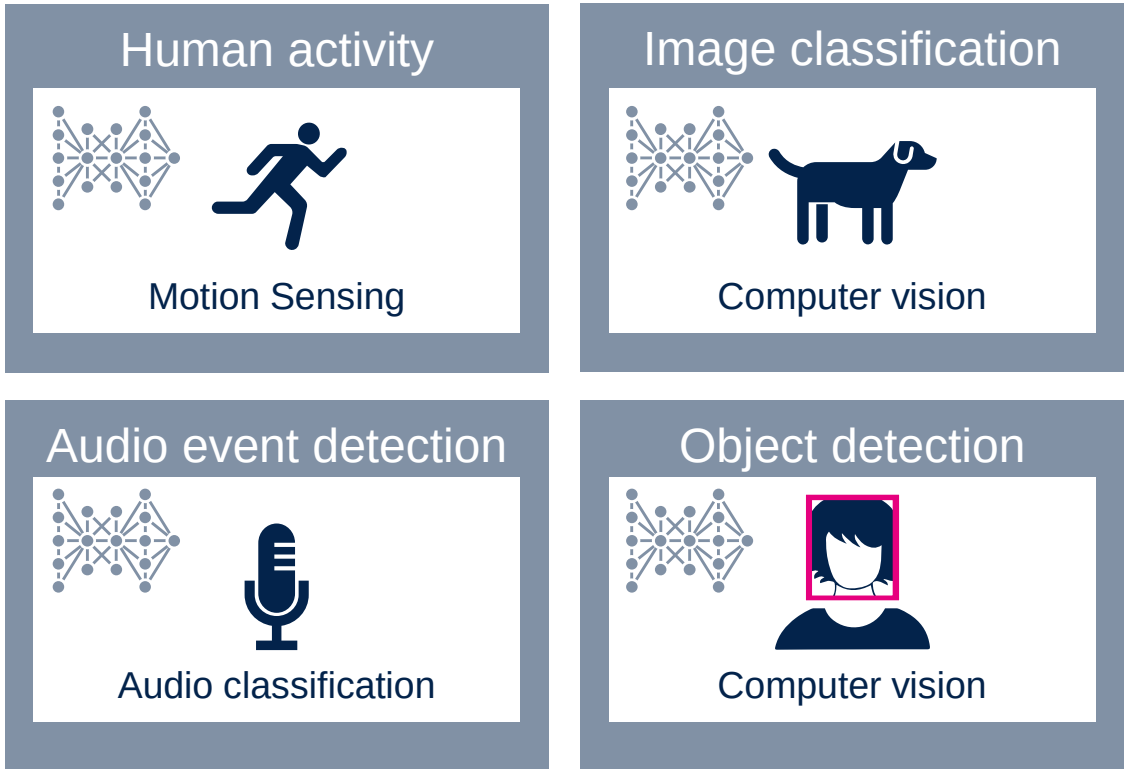


Note: X-CUBE-AI and STM32Cube.AI CS share the same optimization engine

life.augmented

STM32 model zoo

A collection of application-oriented models optimized for STM32



Hosted on Github

< | > Model training scripts

- Scripts to generate and validate

Getting started application packages

- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation

life.augmented

STM32Cube.AI Developer Cloud Board Farm

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

Get the actual inference time for optimized models

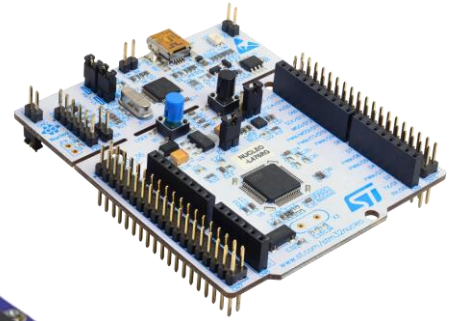
Benchmark models on a large variety of STM32 boards

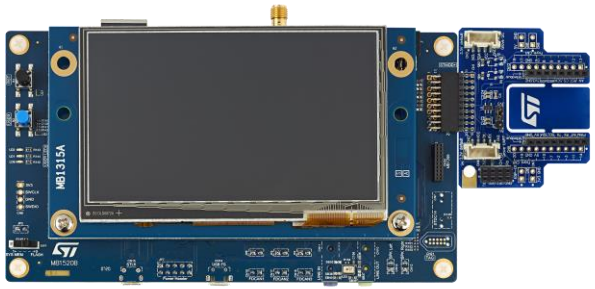
- Find the most appropriate board for your application

NEW

Get access to the most recent devices

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





life.augmented

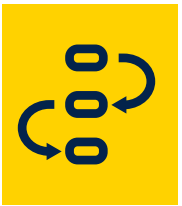
STM32Cube.AI Developer Cloud benefits

STM32Cube.AI Developer Cloud

Save Time
- Immediately start by selecting model from ***STM32 model zoo***
- No code/Low code solution with no installation required

Make the best choice on hardware selection
- Benchmarking service via the ***ST board farm***
- Save workload and money

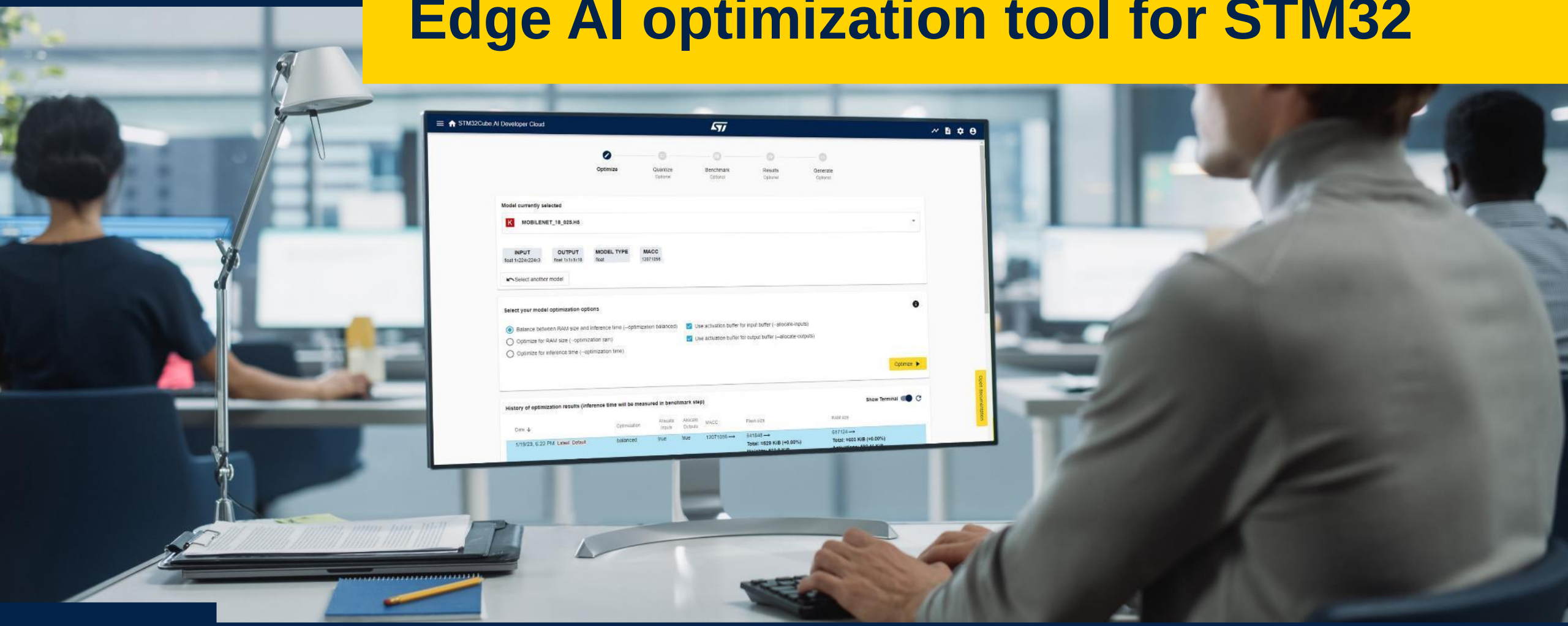
Maximize AI performance on STM32
- Benefit from the STM32Cube.AI performance



Ease integration in your ML workflow
- Use REST API and Python scripts to optimize your ML workflow

life.augmented

NanoEdge AI Studio
Edge AI optimization tool for STM32



life.augmented

Simplify your AI development workflow

NanoEdge AI Studio, an automated ML design solution

NANOEDGE AI STUDIO

The **best combination** for given data:
AI model, hyperparameters and preprocessing

On-device learning capability to fine-tune deployed solution without retraining

No need to create/train AI models in AI frameworks

life.augmented

Boost your productivity

Take advantage from **state of the art of ML**

AI technological breakthrough delivered in regular updates

Speed up your development time

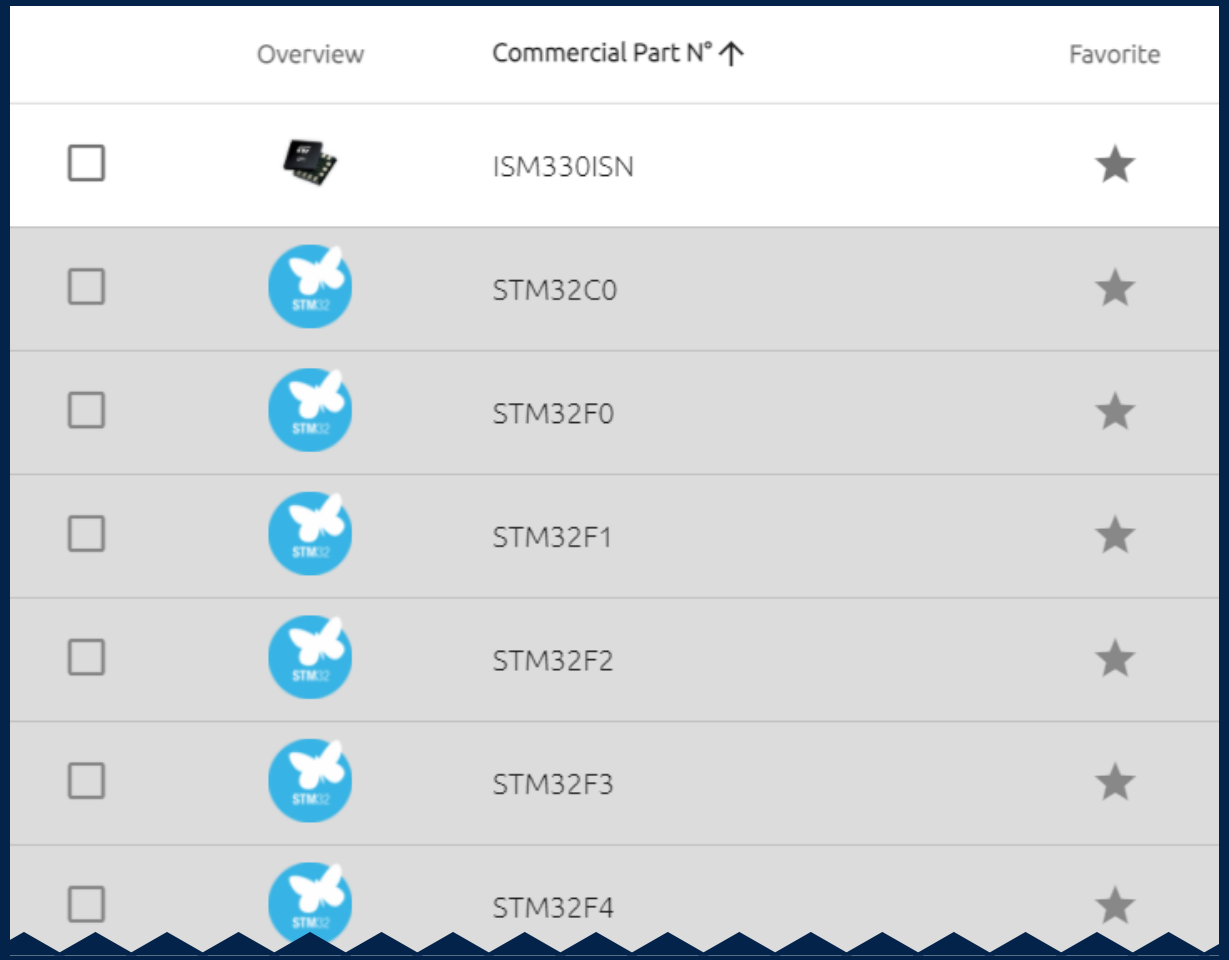
Benchmark thousands of ML combination in minutes

Generate **ultra optimized ML library** for every STM32

From 0.2 KB to 10 KB of memory footprint

life.augmented

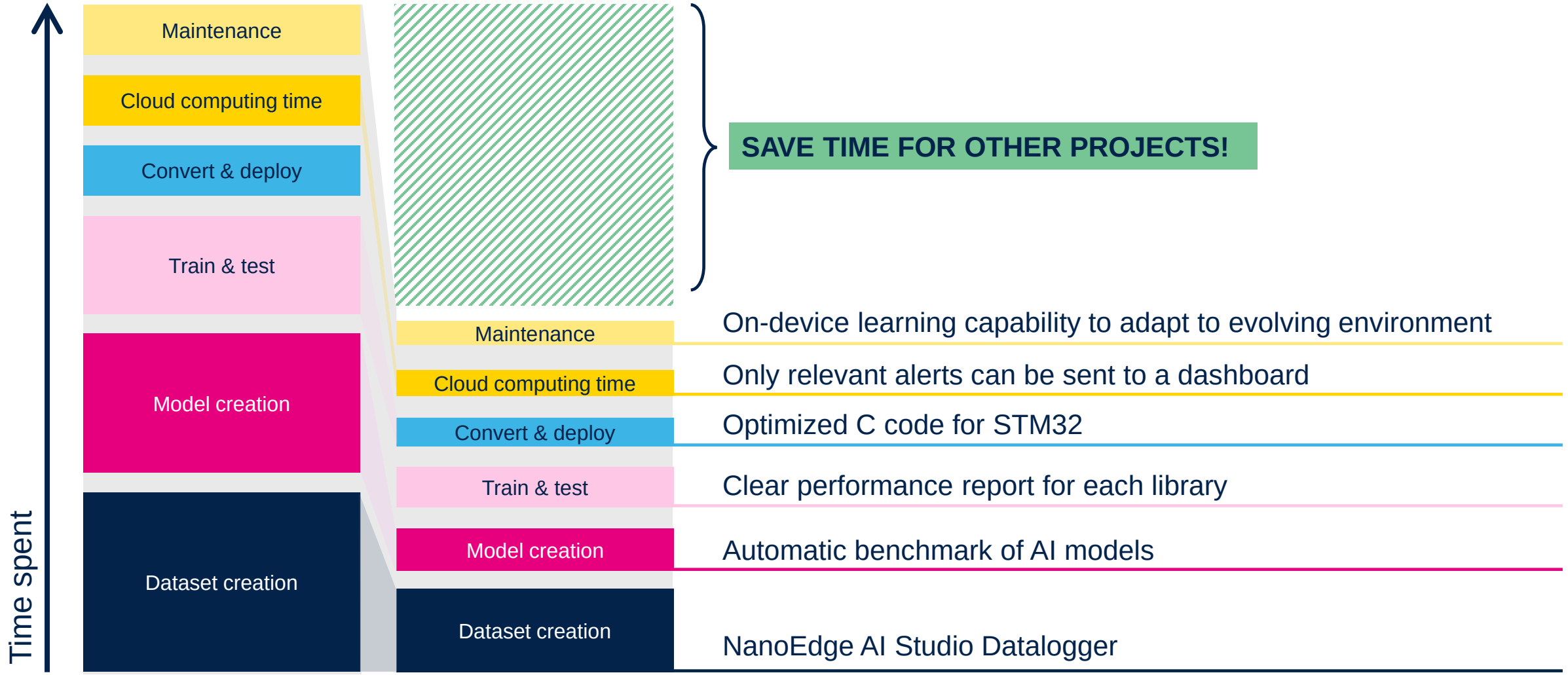
Create solutions for any STM32 MCU



The compact size of the library and its optimized code allows for **seamless integration with any STM32** microcontrollers.

life.augmented

AI solutions development flow enhanced with NanoEdge AI Studio



NANOEDGE AI STUDIO

life.augmented

State-of-the-art machine learning for smarter products

AD
Anomaly Detection

1C
1-Class Classification

nC
n-Class Classification

E
Extrapolation

> **I want to anticipate product failures**

> **I need to detect any outliers**

> **I want to identify the activity, the environment, the usage**

> **I need to predict future states**

life.augmented

An integrated workflow

Create your data set and setup your project

Project parameters
max ram, max flash, microcontroller ref

Dataset
Collection of data from sensors

01010101
01010101
01010101
01010101

Analyse, choose and validate the best ML processing

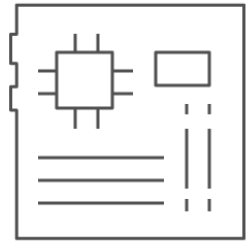
NANOEDGE AI STUDIO

NanoEdge™ AI Studio

Generate optimized library for STM32

Precompiled
library (.a) to link to your main code

Train your model directly on the device



on-device learning

life.augmented

All the tools needed for data collection

DL

Data Logger

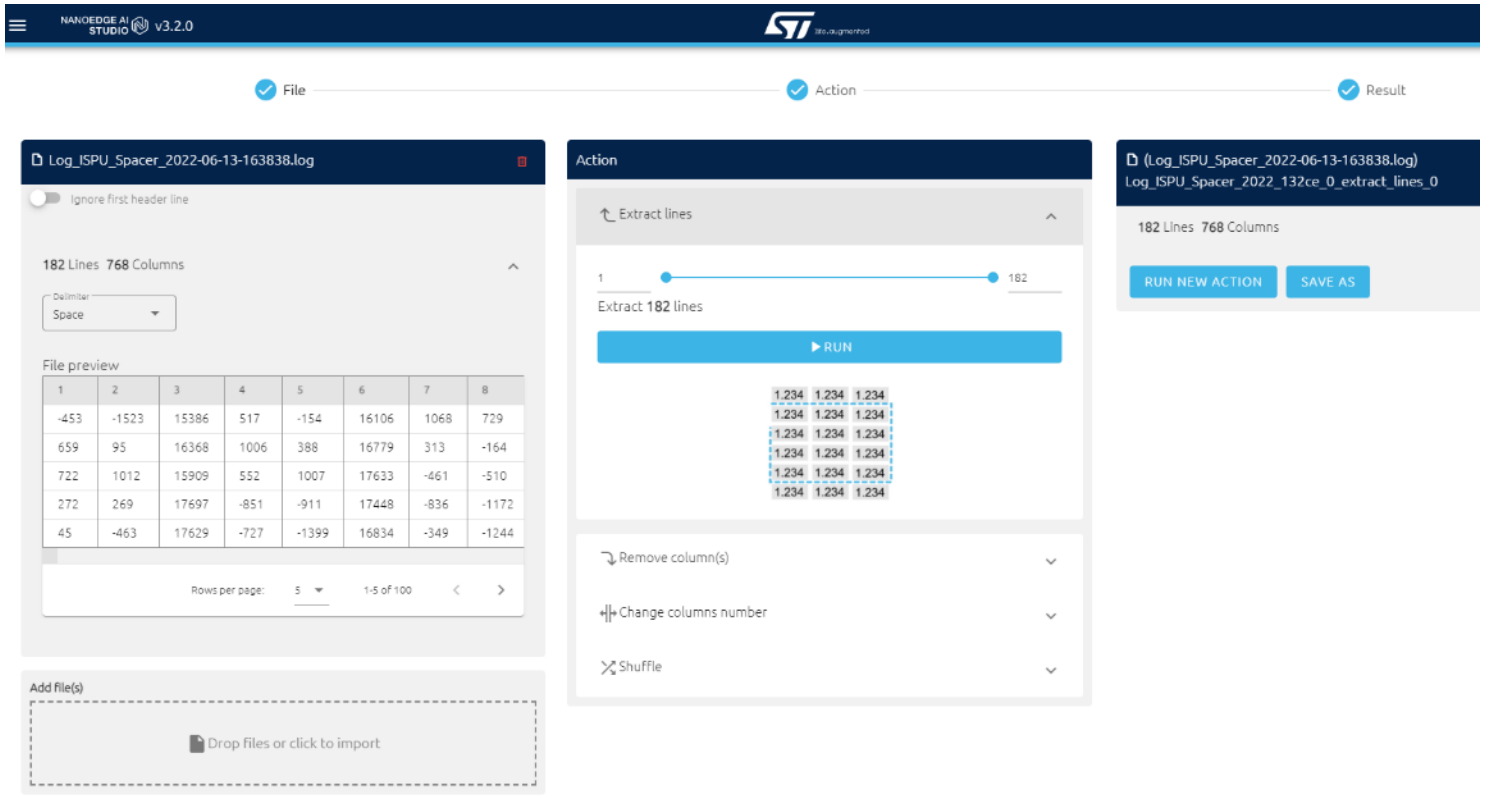
“I want the tools to easily create my dataset”

All the tools needed for data preparation

DM

Data Manipulation

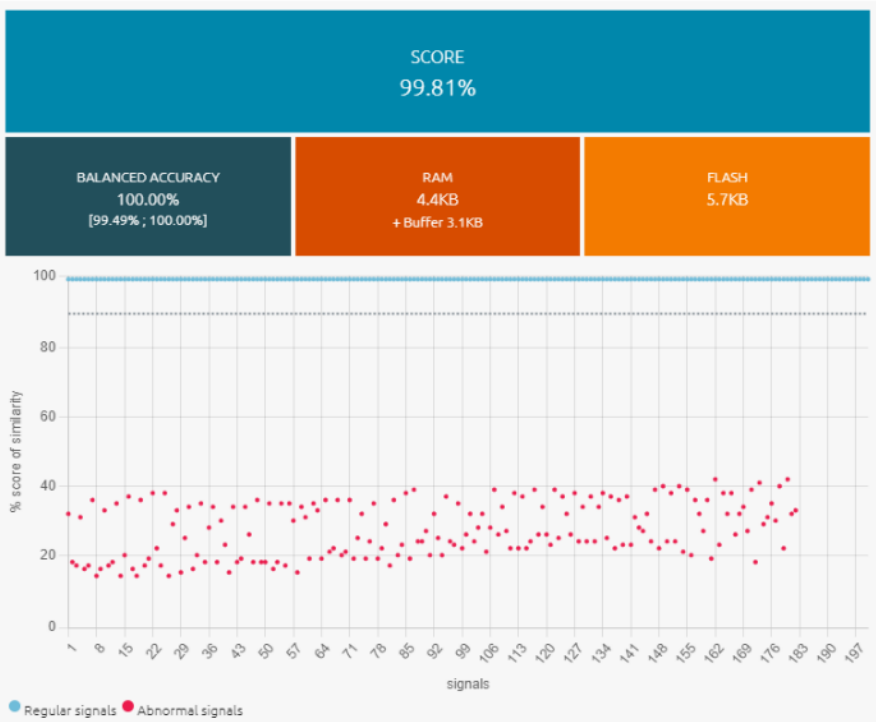
“I want to quickly and efficiently clean my data”

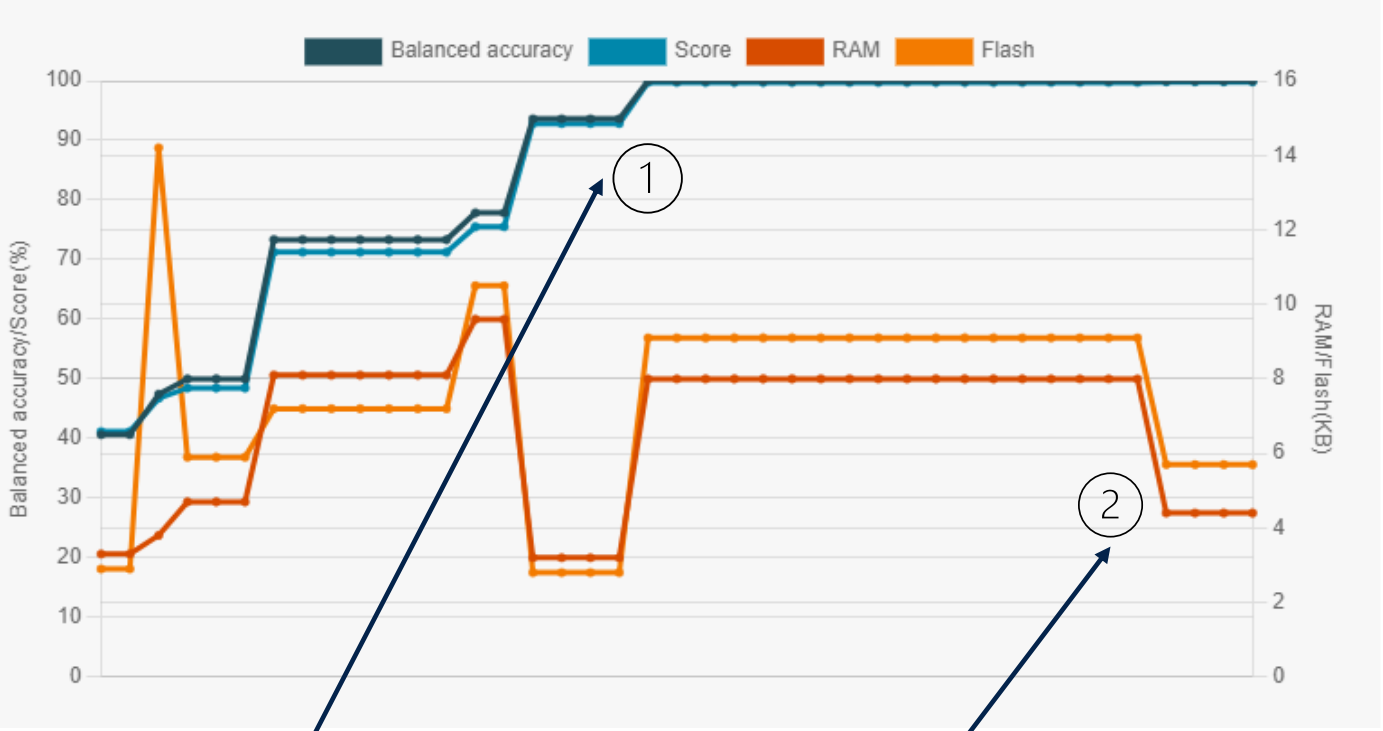


life.augmented

Get the best ML algorithm for your application

Internal benchmarking tool picks the best algorithm for your data



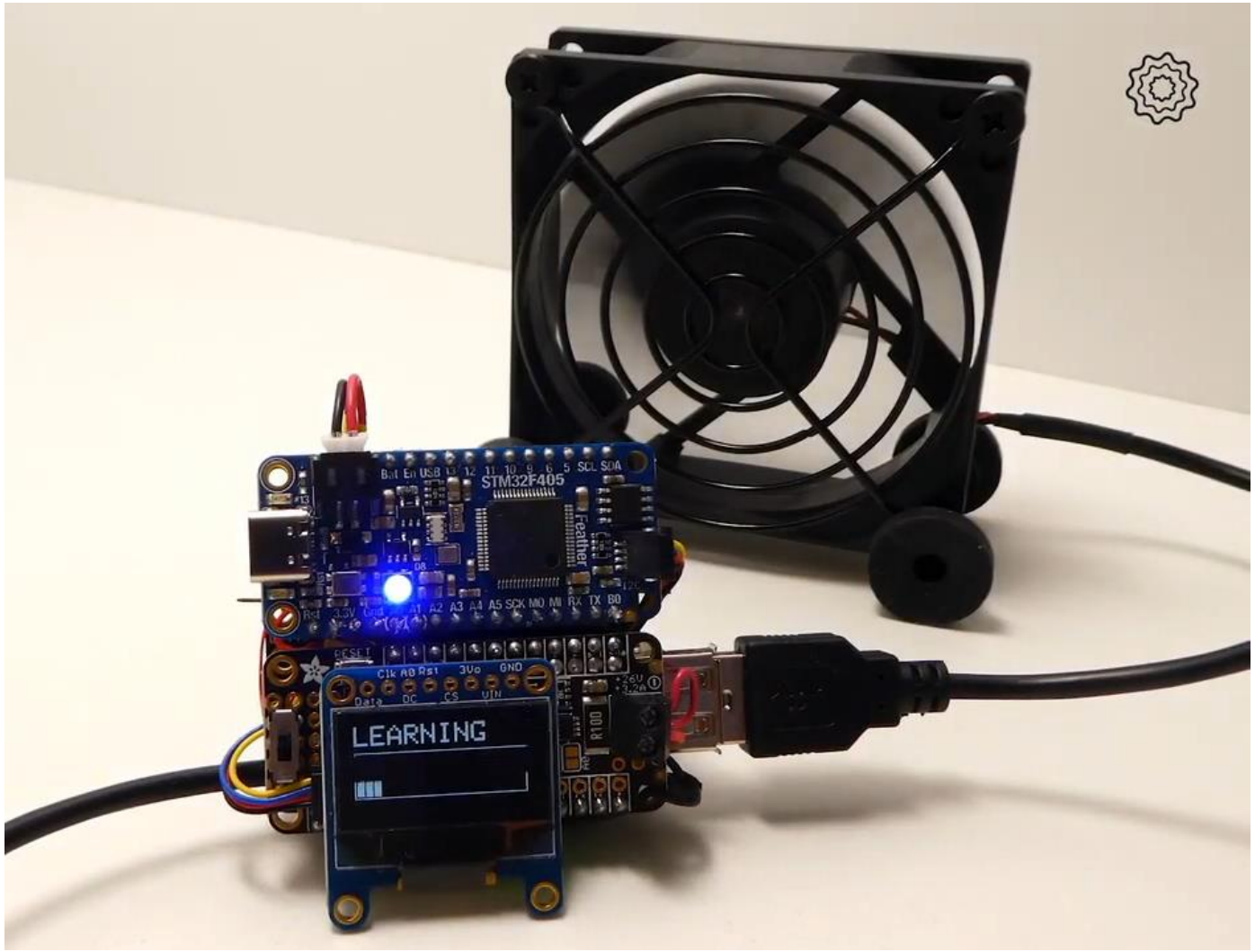


NanoEdge AI Studio improves the performance of the model..

.. And then optimizes it to reduce footprint and latency

life.augmented

On-device learning to learn on the edge an adapt to real environment



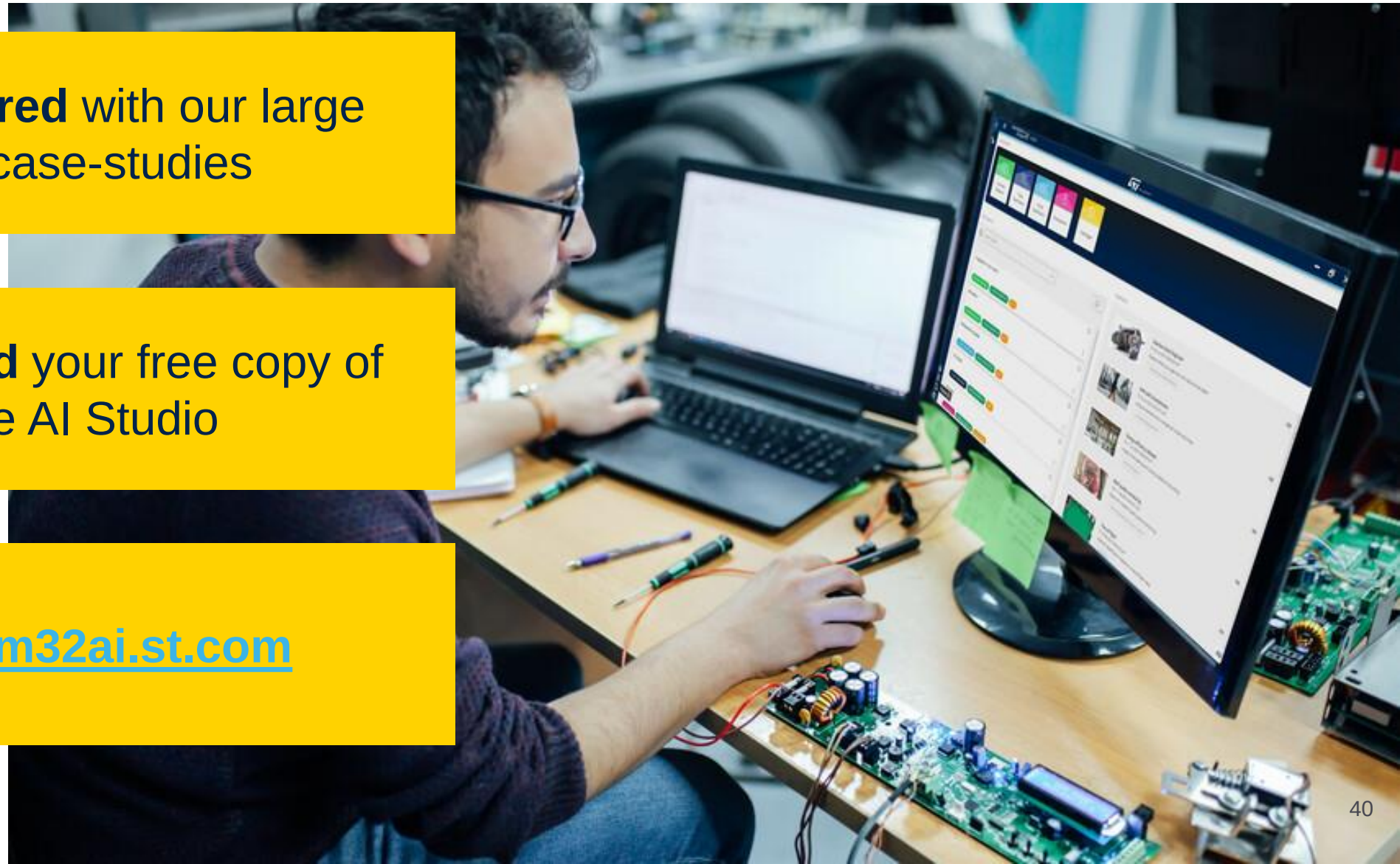
life.augmented

It’s time to convert your data into valuable information!

Get inspired with our large library of case-studies

Download your free copy of NanoEdge AI Studio

https://stm32ai.st.com



life.augmented

Our technology starts with You

Find out more at www.st.com

life.augmented