Unlocking AI potential in embedded systems through efficient development and optimization

Daniel Wang

Technical Marketing Manager

STMicroelectronics

The explosion of AI-enabled devices is accelerating the inference shift from the cloud to the tiny edge

Inference

Cloud-centric

On device-centric

Tiny edge-centric

Enabled by a different class of hardware and software

Inferring at the edge brings substantial benefits

Ultra-low latency
Real-time applications

Reduced data transmission
Generate meaningful information

Enhanced privacy and security
No data sharing in the cloud

Sustainable on energy
Low data, low power

Lower cost of inference to enable a new class of operations

Tiny edge-centric

Opening a new range of embedded AI applications

Upgrade existing devices with AI-based services

Arc Fault Detection

Predictive Maintenance

Battery Management

Reduce BoM by shifting from MPUs to MCUs

People Detection

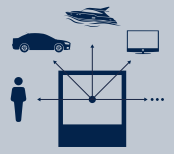
Sound Analysis

Speech Recognition

Offer advanced services without cloud costs

Object Classification

Pose Estimation



Object Segmentation

A few KOPS → Tens of TOPS

Edge AI application examples

AI addresses a wide variety of projects

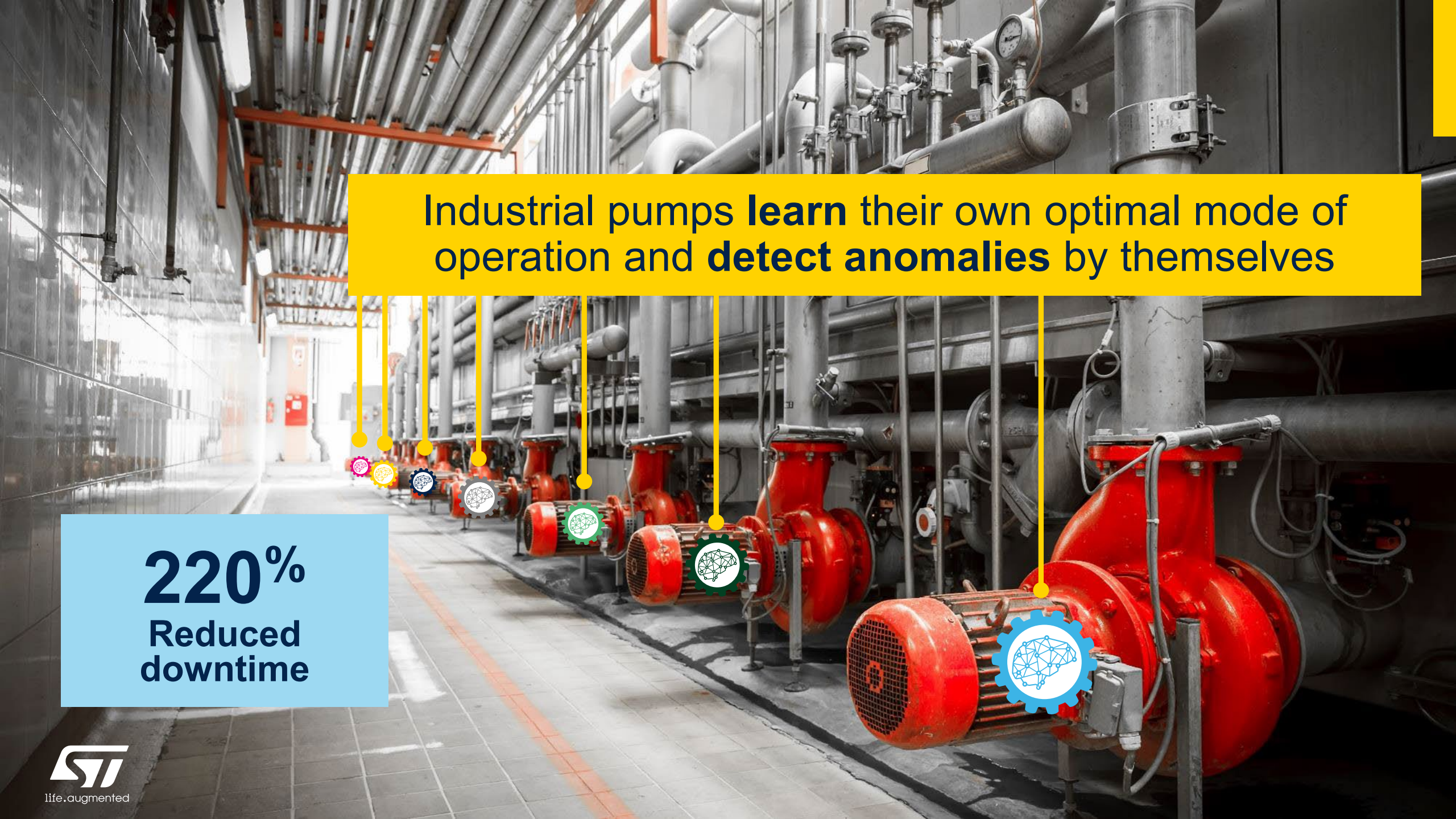
Smart city	Smart buildings	Smart home	Energy	Industry 4.0	Healthcare
Energy optimization Traffic management Public safety Public transportation	Energy optimization Access control Safety Predictive maintenance	Outdoor cameras Babycams Smart doorbells Home appliances Energy management Lawn mowers Vacuum robots	Solar panels Breaker's control EV chargers Smart meters Fraud detection	Predictive maintenance Tools safety Environment monitoring	Wearables Patient monitoring Fall detection Predictive maintenance

Battery management

Arc fault detection

Face / object recognition

Anomaly detection


Industrial pumps **learn** their own optimal mode of operation and **detect anomalies** by themselves
220%
Reduced downtime
life.augmented

A washing machine uses **advanced motor control algorithms** to weigh clothes and optimize water, detergent, and energy used

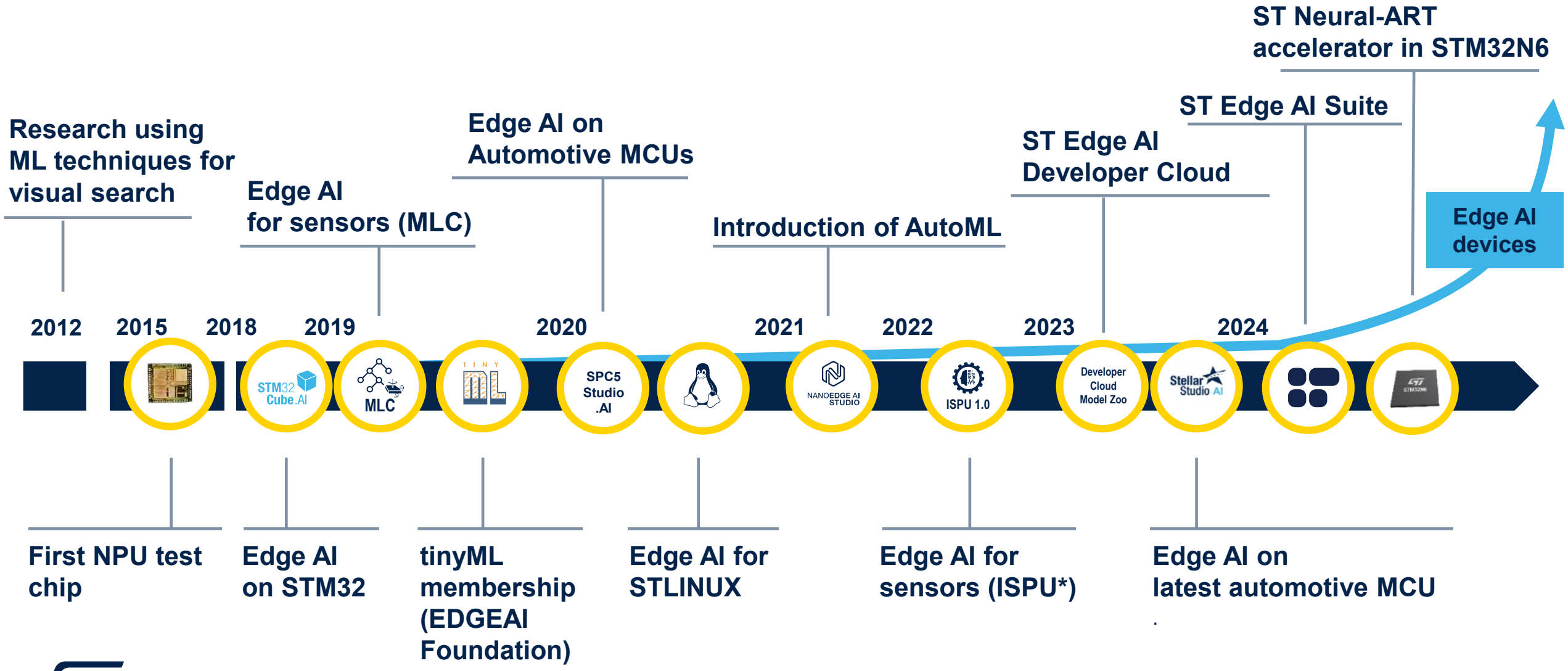
~15-40%
Energy saving per washing cycle

Leader in white-goods

Production starting in 2024 for **millions** units

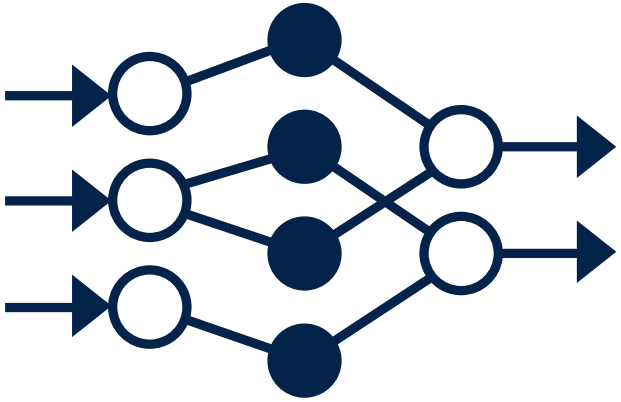
ST’s investment in edge AI enabling broad adoption

10+ years of research, development, and deployment



*Intelligent sensor processing unit

The challenge of deploying embedded AI



AI expertise

Data

Memory footprint

Inference time

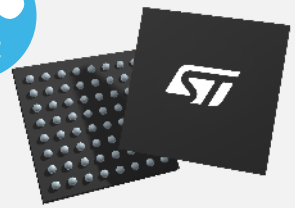
Power consumption

SW development

Deploying embedding AI on microcontrollers presents significant challenges

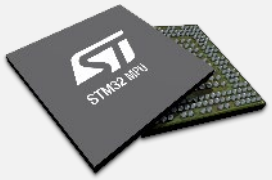
A comprehensive approach to help developers accelerate their product transformation

STM32



STM32 MCUs

STM32N6 MCU



STM32 MPUs

STM32MP2 MPU

Stellar

SPC5

Stellar MCUs

SPC5 MCUs

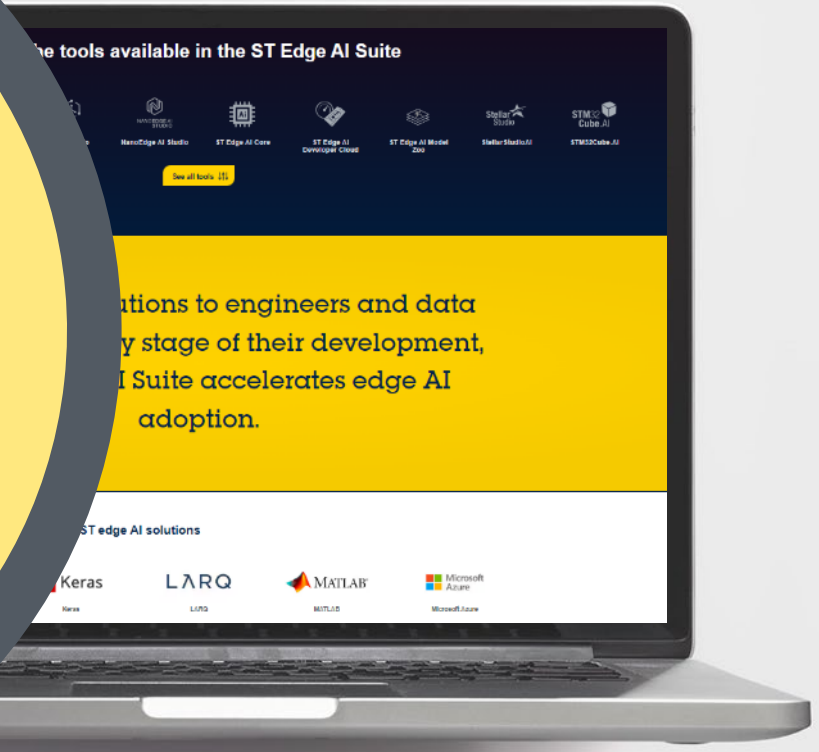
Sensors

MEMS with **MLC and ISPU**

Hardware platforms

Software tools

ST Edge AI Suite



Click on the products to find out more

With embedded AI acceleration

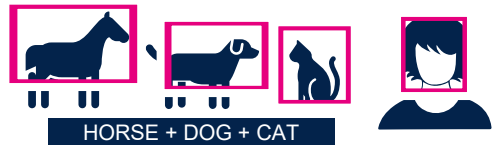
STM32 product offering

Enabling major edge AI technologies

SENSING (TIME SERIES)

AUDIO

VISION



Software tools for any user profile

SENSING
AD

NANOEDGE AI STUDIO

User-friendly AutoML tool for STM32 MCUs

SENSING
AUDIO
VISION

STM32 Cube.AI

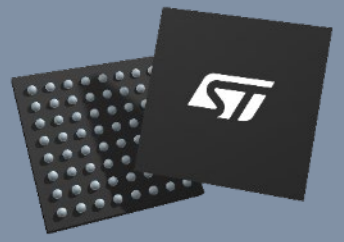
AI model optimizer and code generator for STM32 MCUs

SENSING
AUDIO
VISION

X-LINUX-AI

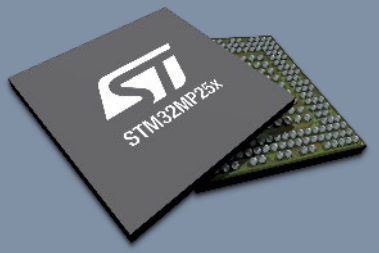
A complete AI framework for STM32 MPUs

Large choice of general purpose & accelerated hardware



STM32 MCUs

STM32N6 MCU with AI acceleration



STM32MP1 & STM32MP2 MPUs

STM32 portfolio

LONGEVITY 10 YEARS COMMITMENT

MPU

- **STM32MP1** — Up to 1 GHz Cortex-A7, 209 MHz Cortex-M4
- **STM32MP2** — Dual 1.5 GHz Cortex-A35, 400 MHz Cortex-M33

High-performance MCUs

- **STM32F7** — 1,082 CoreMark, 216 MHz Cortex-M7
- **STM32H7** — Up to 3,224 CoreMark, Up to 600 MHz Cortex -M7, 240 MHz Cortex -M4
- **STM32N6** — 3,360 CoreMark, 800 MHz Cortex -M55, Neural processing unit
- **STM32F2** — Up to 398 CoreMark, 120 MHz Cortex-M3
- **STM32F4** — Up to 608 CoreMark, 180 MHz Cortex-M4
- **STM32H5** — Up to 1,023 CoreMark, 250 MHz Cortex-M33

Mainstream MCUs

- **STM32F3** — 245 CoreMark, 72 MHz Cortex-M4
- **STM32G4** — 569 CoreMark, 170 MHz Cortex-M4

Mixed-signal MCUs

- **STM32C0** — 114 CoreMark, 48 MHz Cortex M0+
- **STM32F0** — 106 CoreMark, 48 MHz Cortex-M0
- **STM32G0** — 142 CoreMark, 64 MHz Cortex-M0+
- **STM32F1** — 177 CoreMark, 72 MHz Cortex-M3

Ultra-low-power MCUs

- **STM32L0** — 75 CoreMark, 32 MHz Cortex-M0+
- **STM32U0** — 140 CoreMark, 56 MHz Cortex-M0+
- **STM32L4** — 273 CoreMark, 80 MHz Cortex-M4
- **STM32U3** — 393 CoreMark, 96 MHz Cortex-M33
- **STM32L4+** — 409 CoreMark, 120 MHz Cortex-M4
- **STM32L5** — 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+
- **STM32WB0** — 64 MHz Cortex-M0+
- **STM32WB** — 216 CoreMark, 64 MHz Cortex-M4, 32 MHz Cortex-M0+ ●
- **STM32WBA** — 407 CoreMark, 100 MHz Cortex-M33

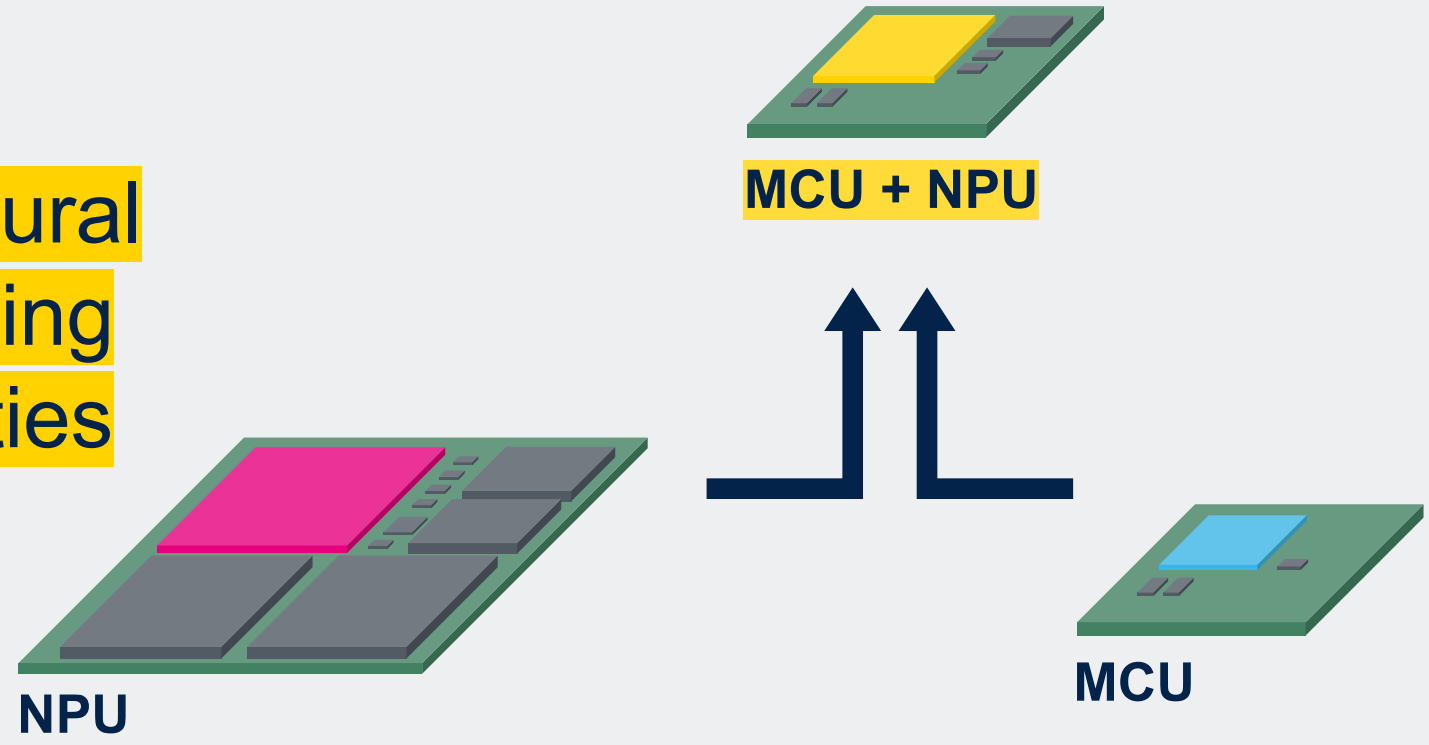
Latest product generation ● Radio coprocessor only ▭ New series or lines introduced in 2025

The first high-performance STM32 MCU with AI acceleration

Take the best of both worlds!

Benefit from extended Neural Network computing power while maintaining the advantages of the MCU

High Neural Processing capabilities



Small footprint

Lower power

Lower cost

Lower BOM

Faster boot/wkup

STM32N6 feature overview

Dedicated embedded neural processing unit (NPU)

- 600 GOPS NPU
- 3 TOPS/W power consumption

Arm® Cortex®- M55 core

- 1280 DMIPS / 3360 CoreMark
- New DSP extensions (MVE)

Embedded RAM

- 4.2 Mbytes of embedded RAM for real-time data processing and multitasking

Computer vision pipeline

- Parallel and MIPI CSI-2 camera module I/F
- Dedicated image processor (ISP)

Extended multimedia capabilities

- 2.5D graphics accelerator
- H.264 encoder, JPEG encoder/decoder

Extended security features

- Arm® TrustZone® for the Cortex®-M55 core and the NPU
- Target certifications SESIP3, PSA L3

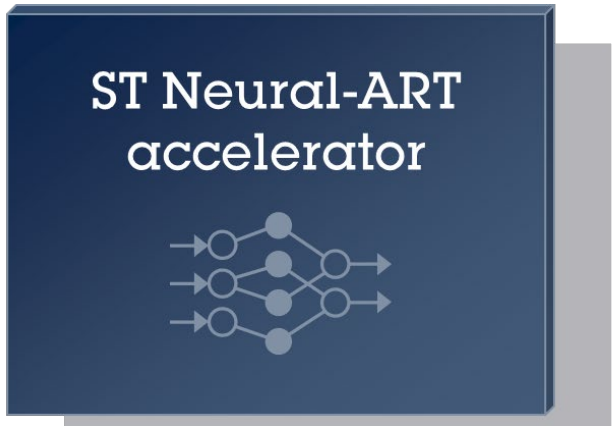
* 600 GOPS NPU vs 1 GOPS NN peak processing capabilities on STM32H7

STM32N6 General purpose MCU with optional Edge AI

General Purpose Line
STM32N645/655

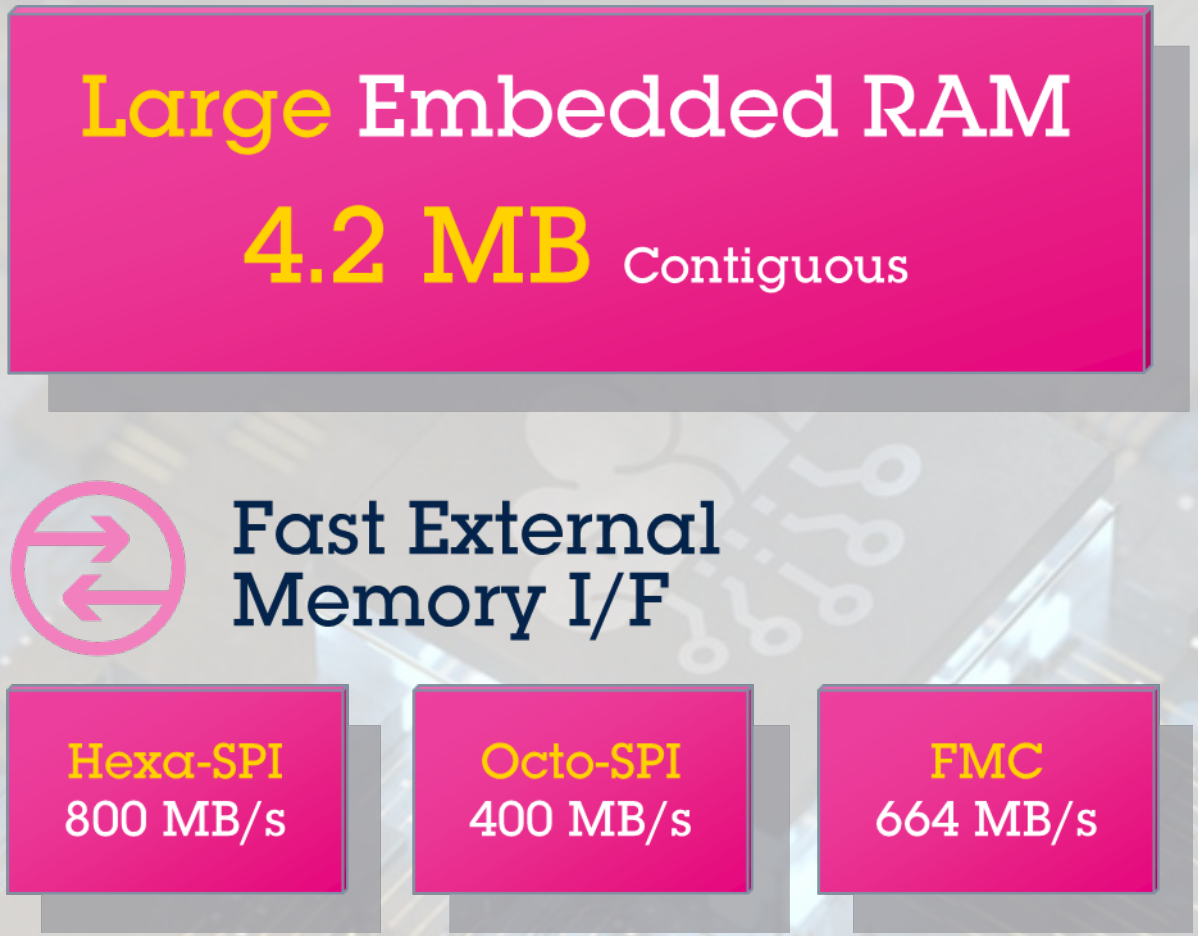
- High Performance, featured product
- Cortex M55 @ 800MHz leading Edge MCU core
- Large embedded memory
- High bandwith I/F to external memories
- Multimedia capabilities

Artificial Intelligence Line
STM32N647/657



- Engineered for edge AI applications
- Optimized for intensive AI algo like vision, audio
- Fully integrated into the STM32 ecosystem for a seamless development experience

Optimize your application with the large embedded memory



Large contiguous embedded memory

- Ideal for Neural Network or graphic applications
- External RAM becomes optional

Fast serial I/F for external memories

- Allows the use of fast and cost-effective memory
- Hexa-SPI for fast access to RAM
- Octo-SPI for secured FLASH

Flash-less configuration

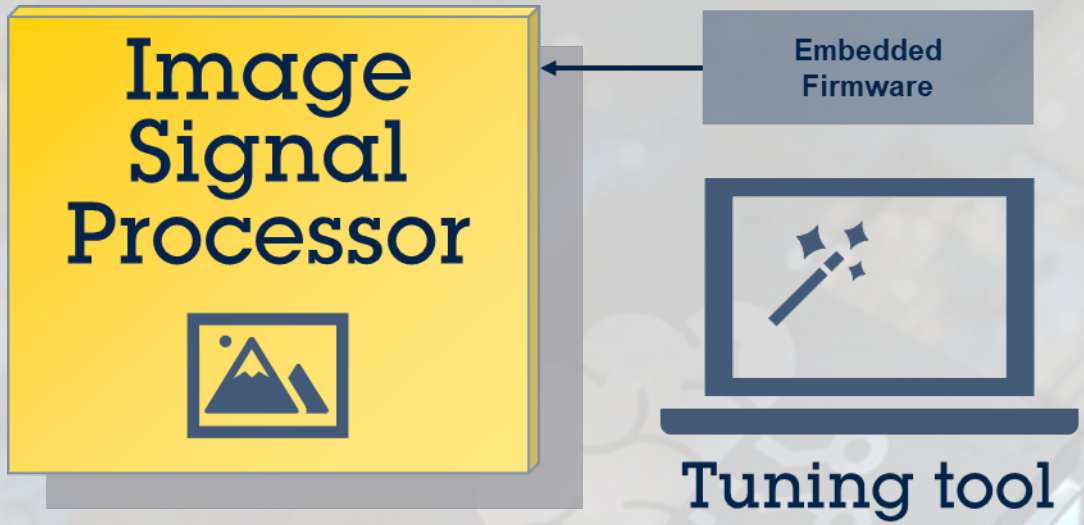
Flexible Memory Controller

- PSRAM, SDRAM, NOR, NAND

Improved security with on-the-fly encryption

- HW accelerated crypto engine on all interfaces

Geared toward computer vision



Extensive camera I/F

Dedicated Image Signal Processor (ISP)

- Dimensioned for 5 Mpixel camera @ 30 fps
- Generate 3 different outputs from the same input

Tool suite to manage ISP tuning

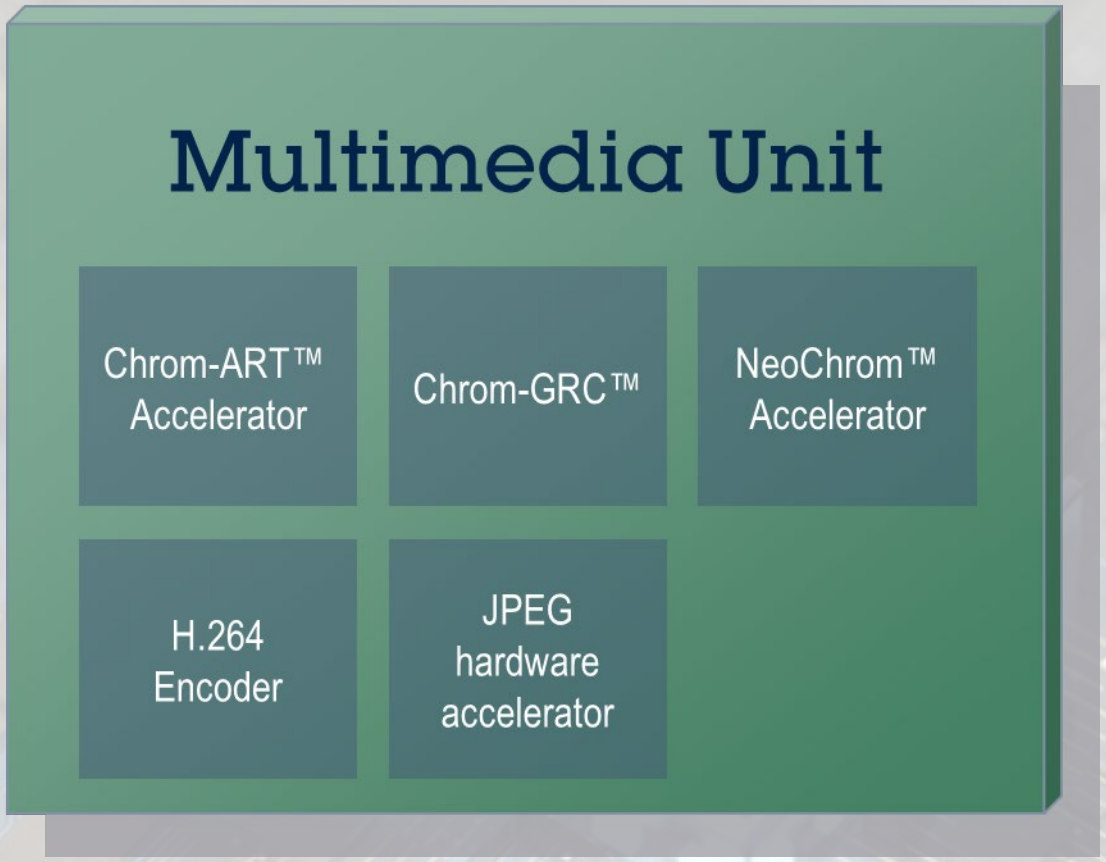
Embedded Firmware on arm Cortex

- 2A for auto white-balance and auto exposure
- Image processing library

Multiple camera interfaces

- MIPI CSI-2, 2 lanes interface
- 16-bit parallel interface

Extended multimedia capabilities



Chrom-ART™ Accelerator

- 2D graphics acceleration

Chrom-GRC™

- Graphic Resource Cutter for non-square displays

NeoChrom™ Accelerator

- 2.5D acceleration for advanced drawing
- Perspective correction and texture mappings

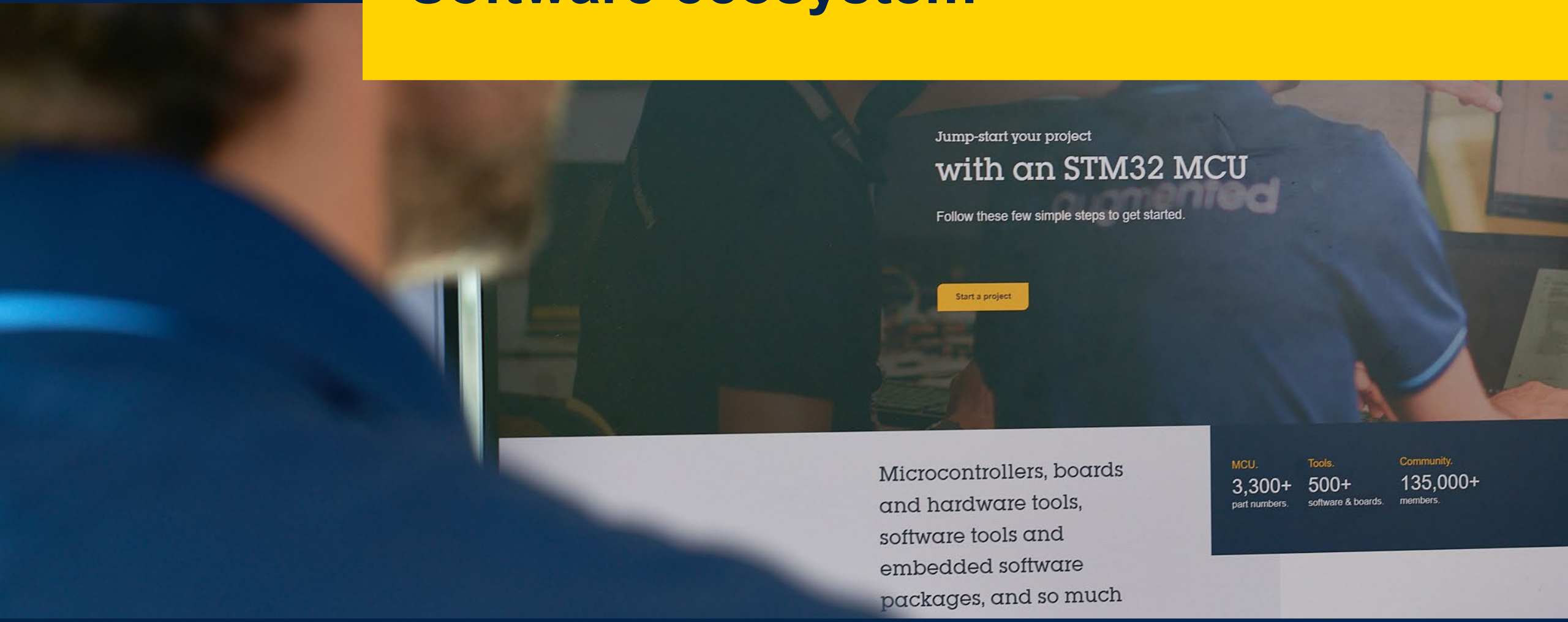
H.264 Encoder

- 720p/1080p @ 30 fps

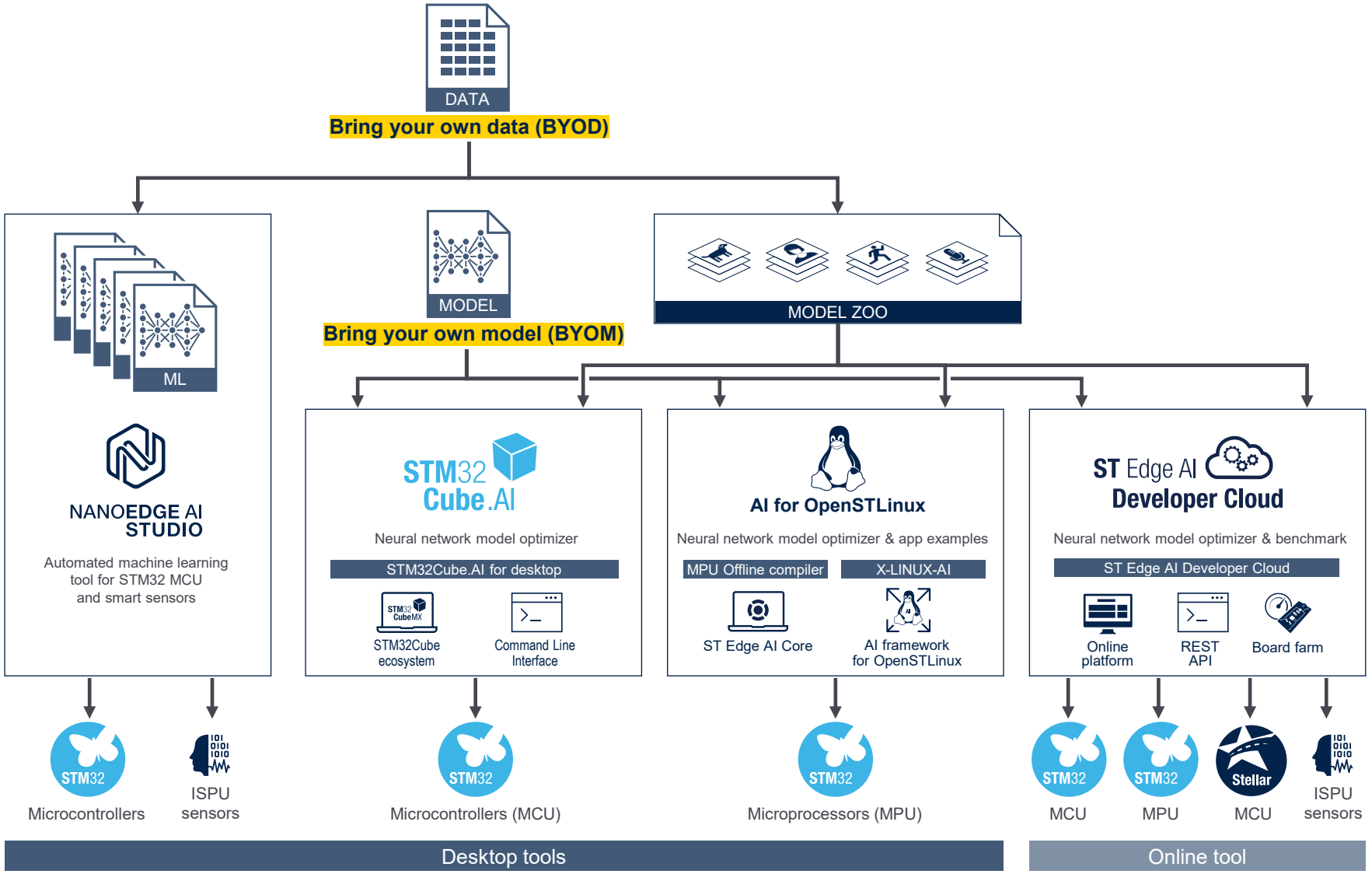
JPEG hardware accelerator

- JPEG compression and decompression
- High quality motion JPEG video playback

Software ecosystem

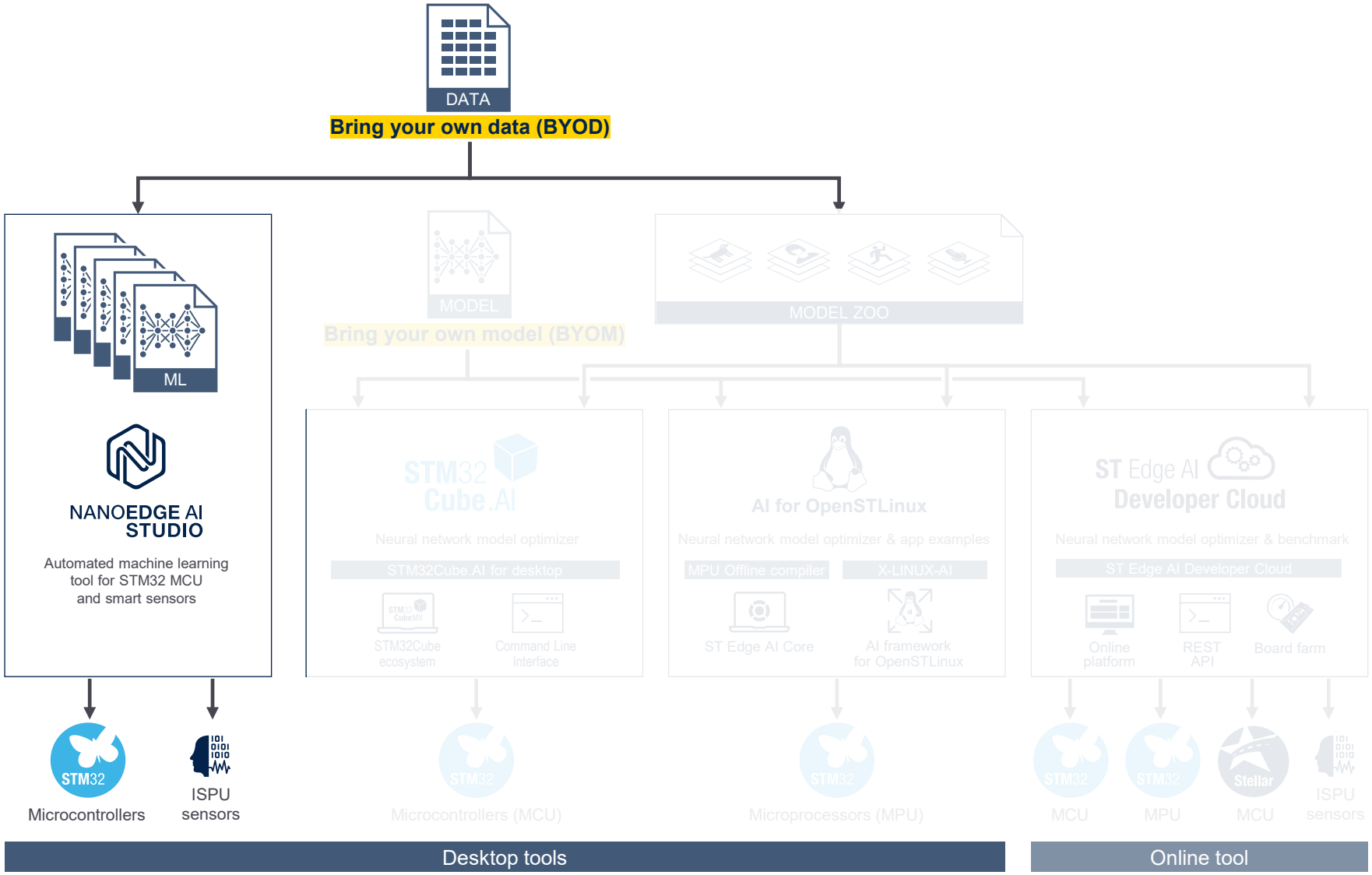


A broad collection of free tools



NanoEdge AI Studio, the AutoML tool

Software tool: NanoEdge AI Studio



Simplified edge AI development workflow

Deployment of NanoEdge AI Studio libraries, the market reference AutoML tool, is completely free for unlimited quantities on any STM32

… and available on any Arm® Cortex® -M MCU*

NANOEDGE AI STUDIO

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

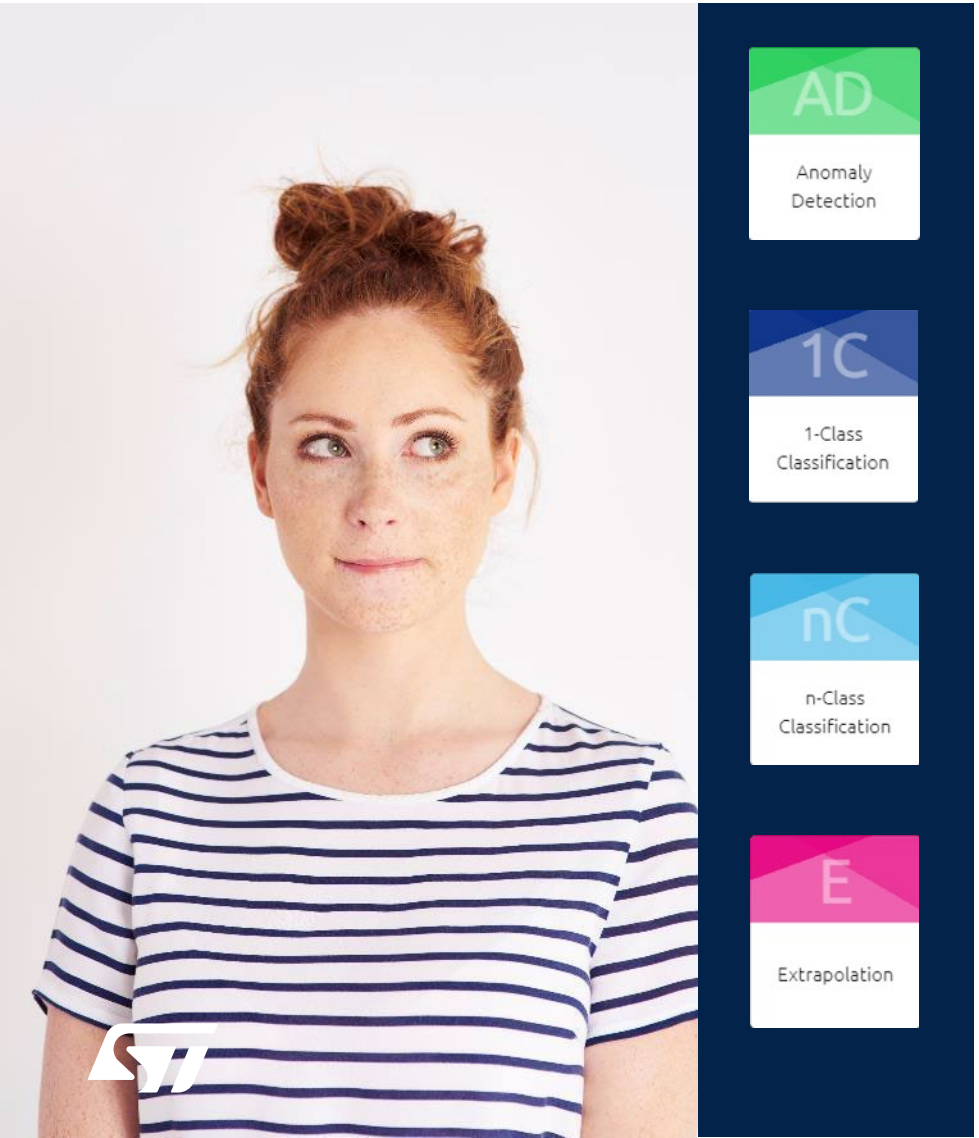
On-device learning capability to fine-tune a deployed solution without retraining

Bring your own data approach:
no need to create edge AI models

**under a special license agreement*

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

NanoEdge AI Studio workflow

Create your dataset
And set up your project

Project parameters
max ram, max flash, microcontroller ref.

011001
100111
110100
111000

Dataset
Collection of data from sensors

Analyze, choose, and validate the best ML processing

NANOEDGE AI STUDIO

Automated Machine Learning tool
for STM32 MCU and smart sensors

STM32

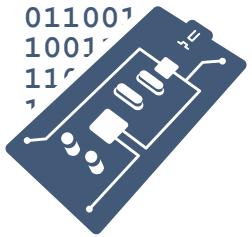
ISPU Sensors

Generate an optimized library

.a Library

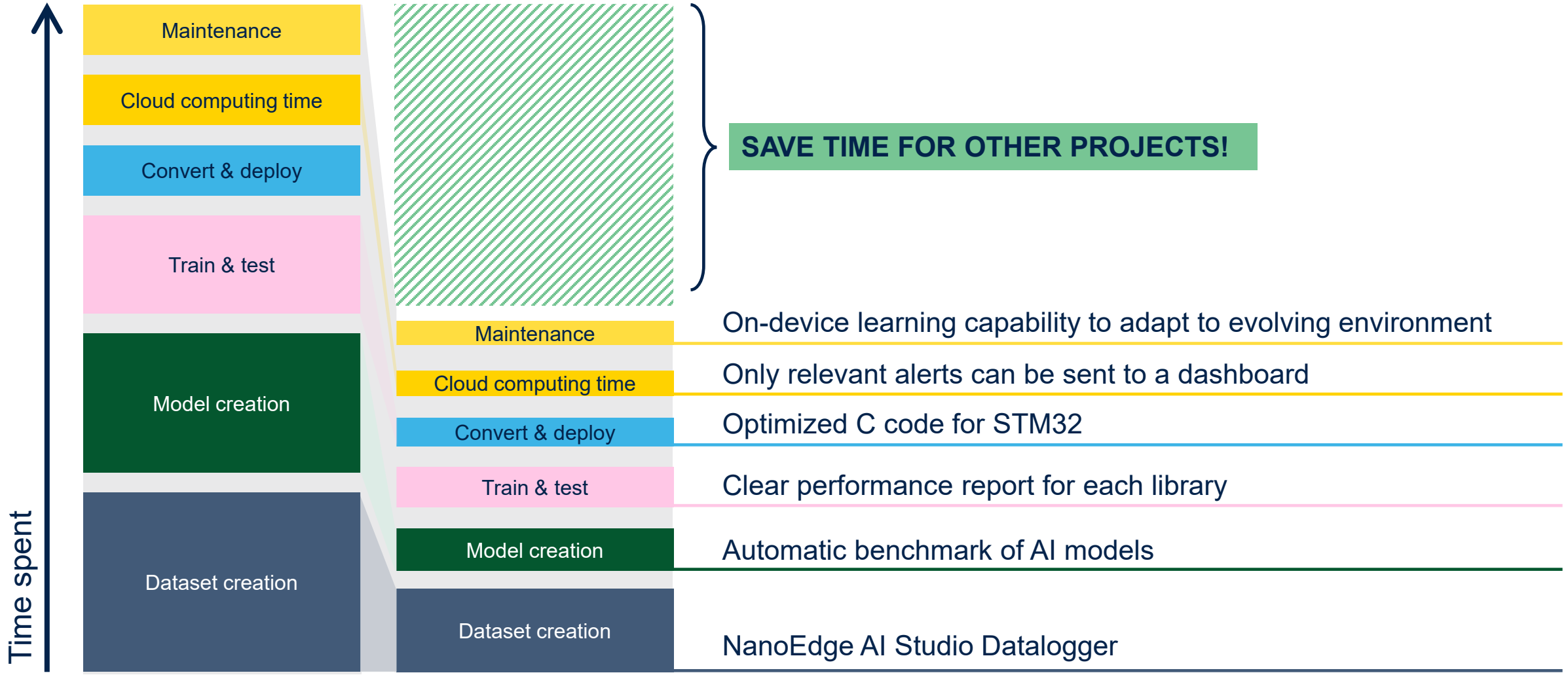
Precompiled library (.a) to link to your main code

Train your model directly on the device



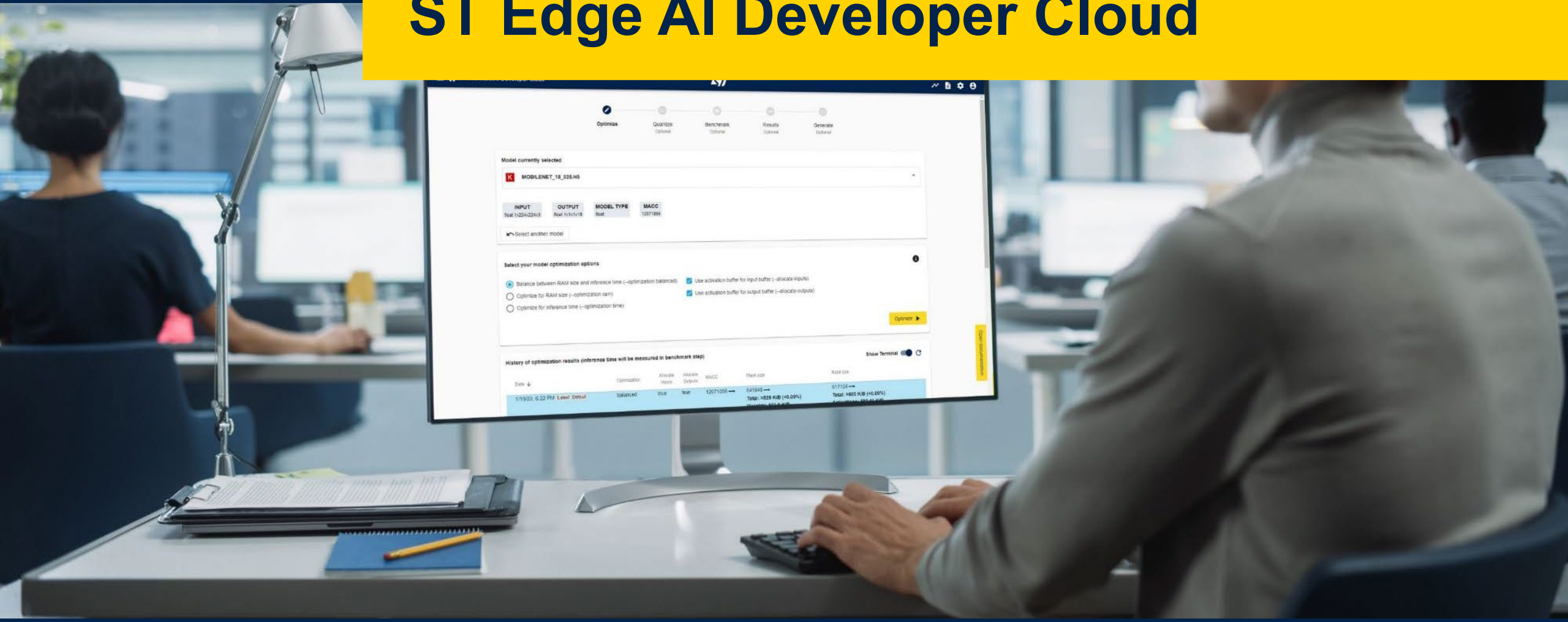
On-device learning
(for anomaly detection)

AI solutions development flow enhanced with NanoEdge AI Studio

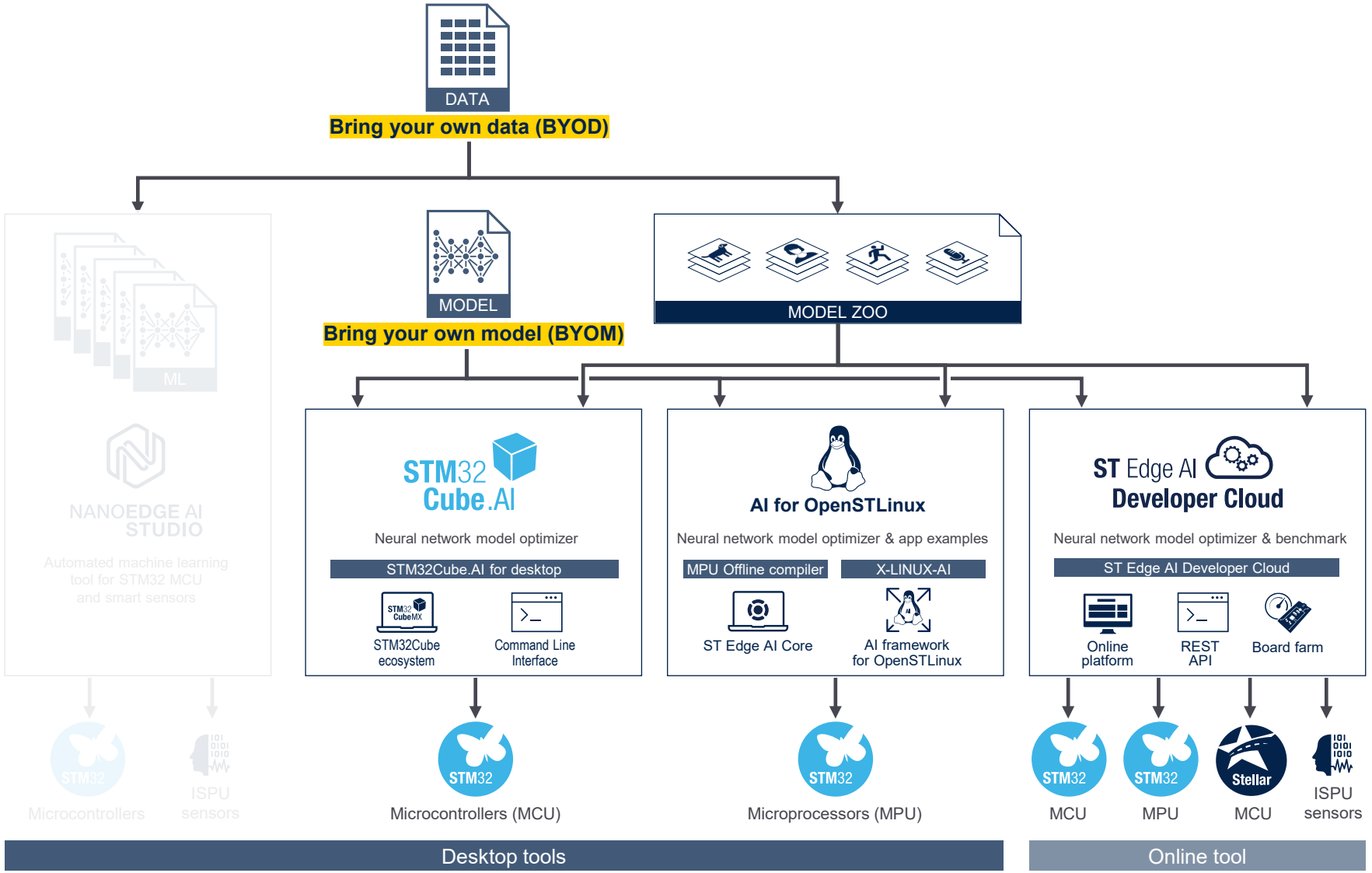


NANOEDGE AI STUDIO

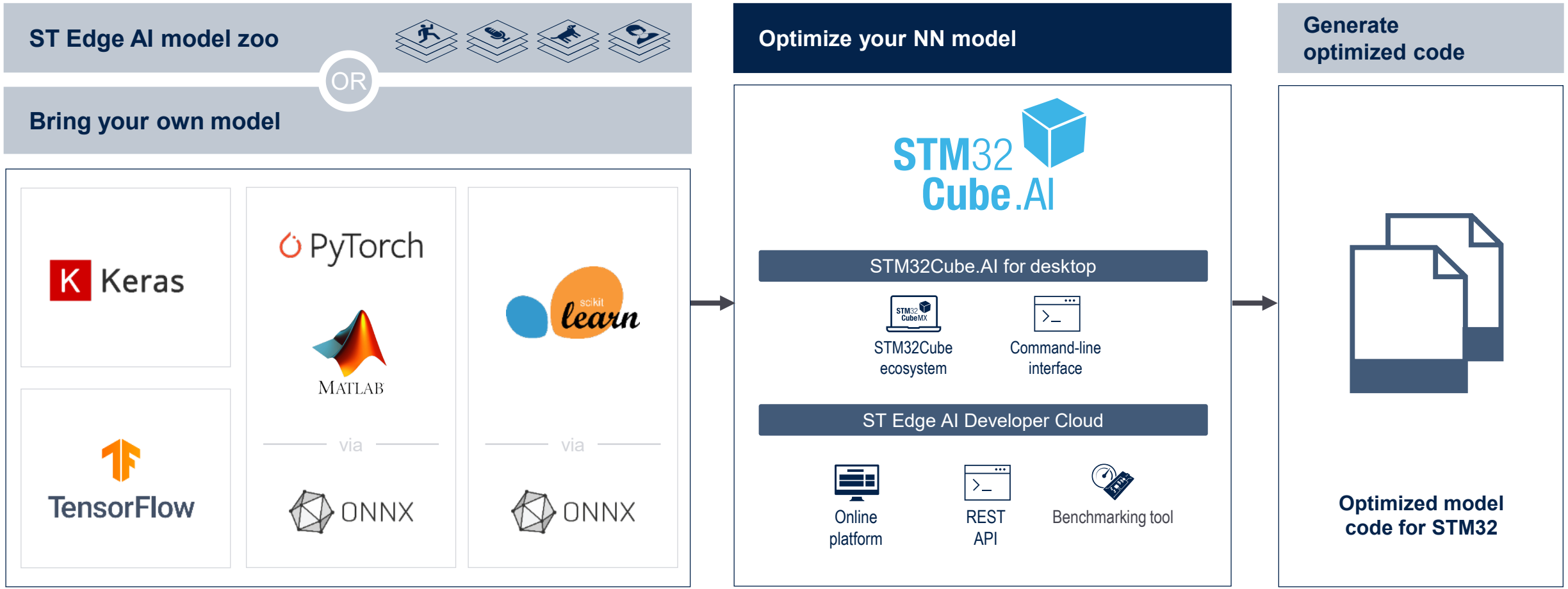
STM32Cube.AI &
ST Edge AI Developer Cloud



Software tool: STM32Cube.AI



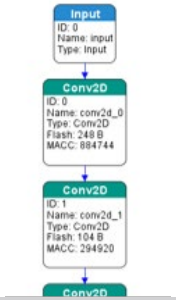
One tool – two versions to deploy AI on STM32



The 3 pillars of STM32Cube.AI

Graph optimizer

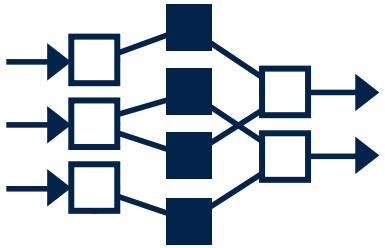
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target hardware architectures



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

Quantized model support

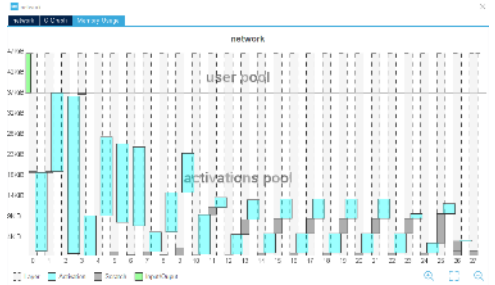
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- 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.

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

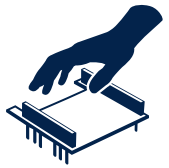
Object detection

Computer vision

Hosted on GitHub

< | > **Model training scripts**

- Scripts to generate and validate

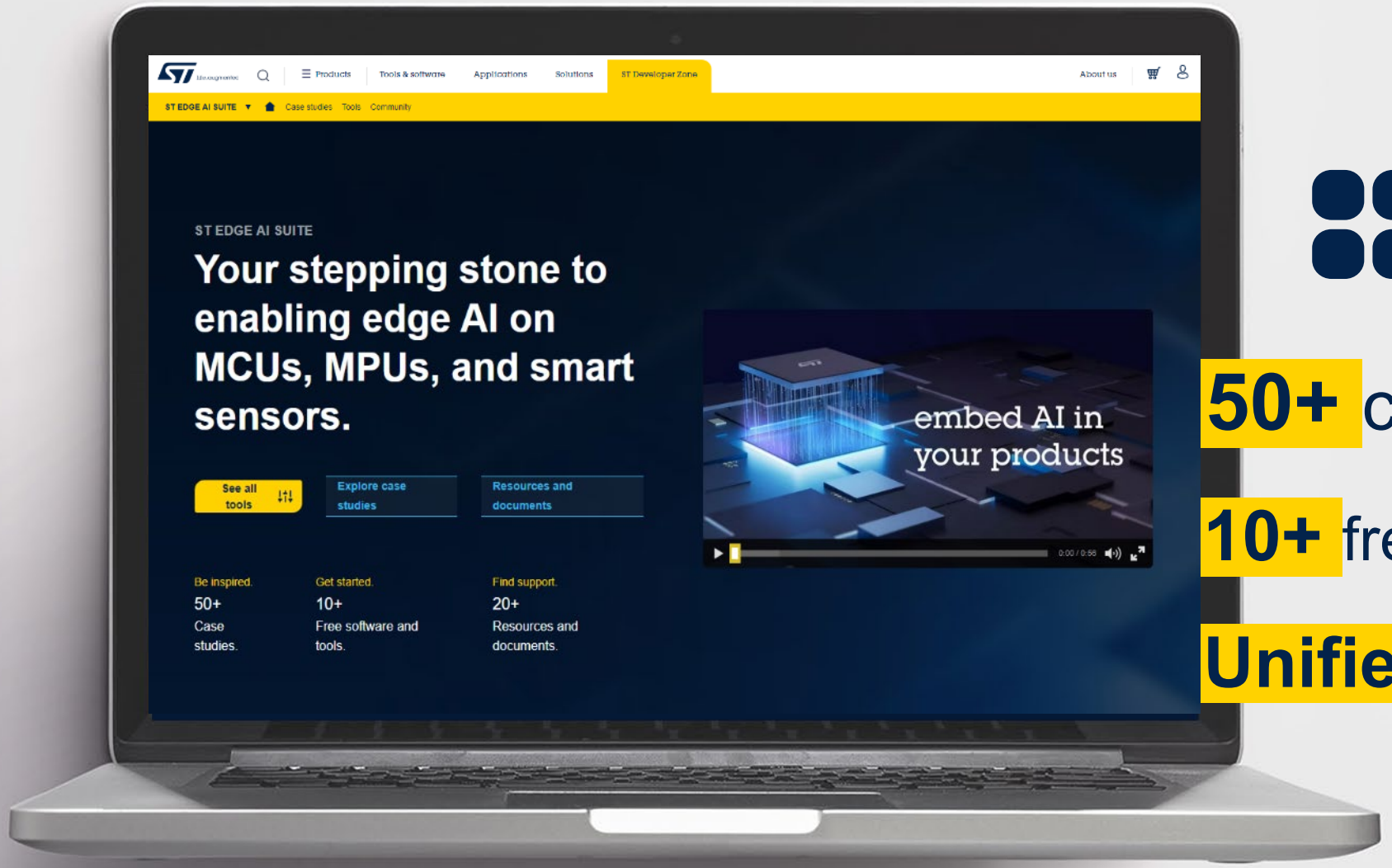


Application code example

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

ST Edge AI Suite

By bringing solutions to engineers and data scientists at every stage of their development, the ST Edge AI Suite accelerates edge AI adoption.



ST Edge AI **Suite**

50+ case studies

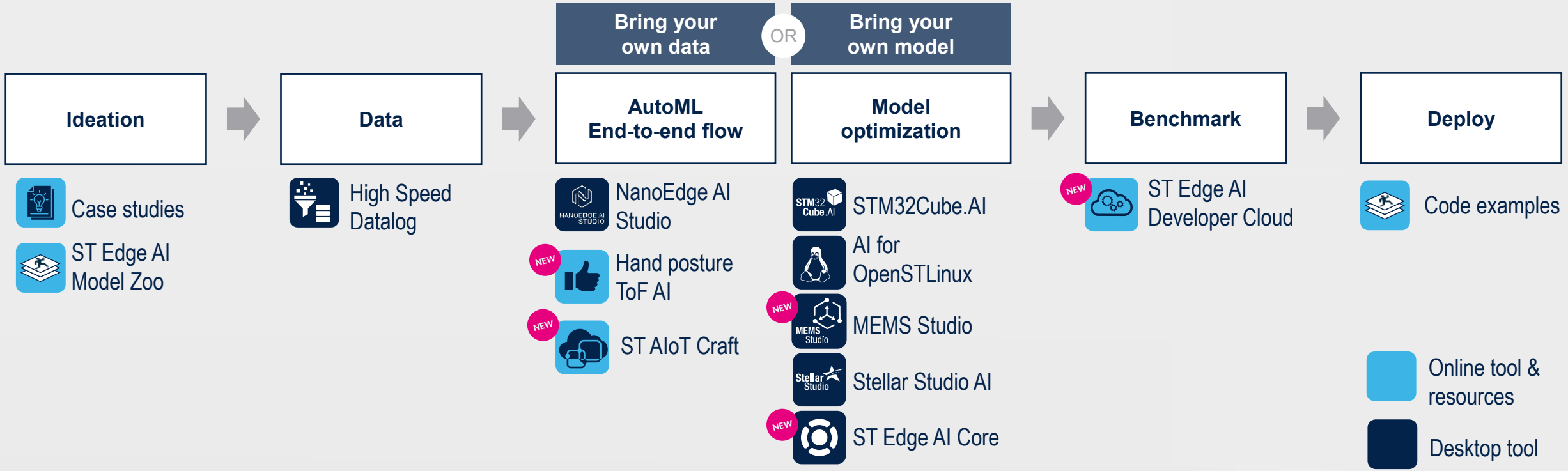
10+ free software tools

Unified AI core technology

ST Edge AI Suite

Free tools to run edge AI on MCUs, MPUs, and smart sensors

Find the tools you need to optimize and deploy machine learning algorithms, from data collection to final deployment on hardware.



Our technology starts with You

Find out more at www.st.com