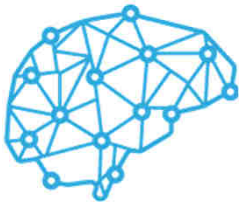


Artificial Neural Network Mapping Made Simple with the STM32Cube.AI

Markus Mayr
Product Marketing Manager, MCU

life.augmented

Technology Tour 2019
Minneapolis, MN | October 24

Artificial Intelligence (AI)

- AI is a superset of all the studies where machines mimic cognitive "human" capabilities. For example:
  - Interaction with the environment
  - Knowledge representation
  - Perception
  - Learning
  - Computer vision
  - Speech recognition
  - Problem solving and many more.
- Main ingredients
  - Computer science
  - Statistics
  - Mathematics

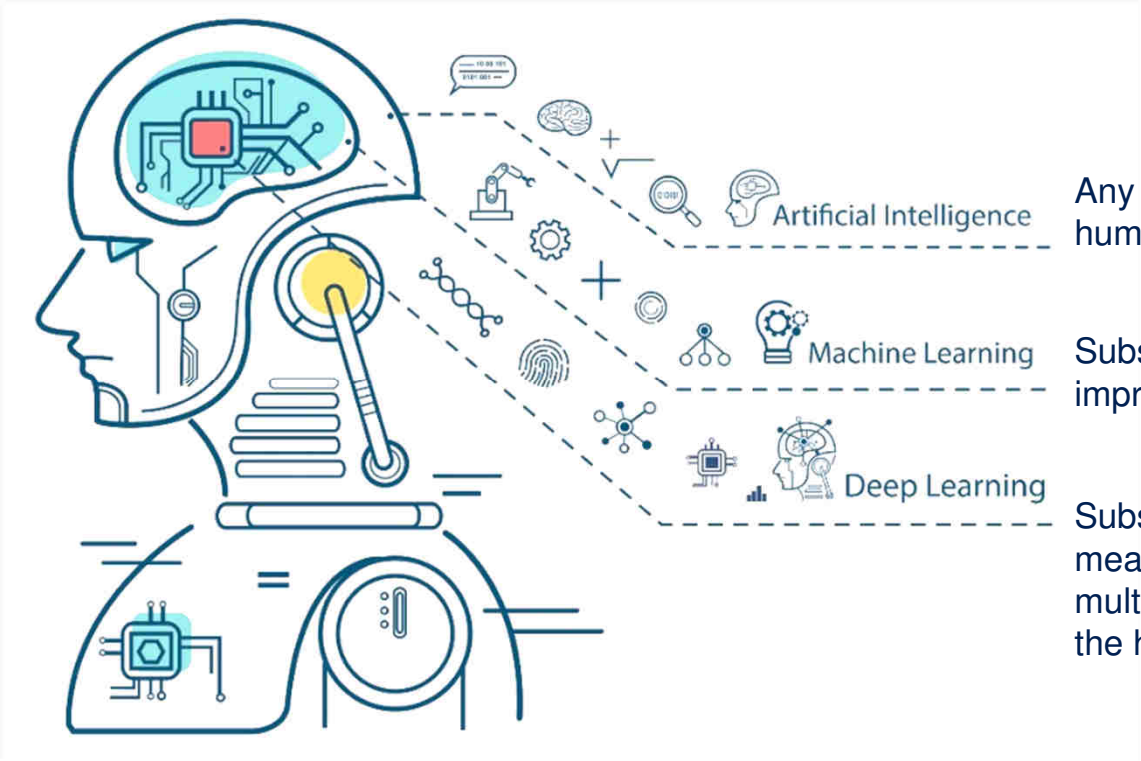
life.augmented

Artificial Intelligence (AI)

- **Main use cases in our everyday life:**
  - Face/voice recognition
  - Autonomous driving
  - Stock market trading strategy
  - Disease symptom detection
  - Predictive maintenance
  - Handwriting recognition
  - Content distribution on social media
  - Fraudulent credit card transaction
  - Translation engines
  - Shopping suggestions

life.augmented

Some definitions



Any technique that enables computer to mimic human behavior

Subset of AI. Algorithms and methodologies that improve over-time through learning from data

Subset of ML. Learning algorithms that derive meaning out of data, by using a hierarchy of multiple layers that mimic the neural networks of the human brain

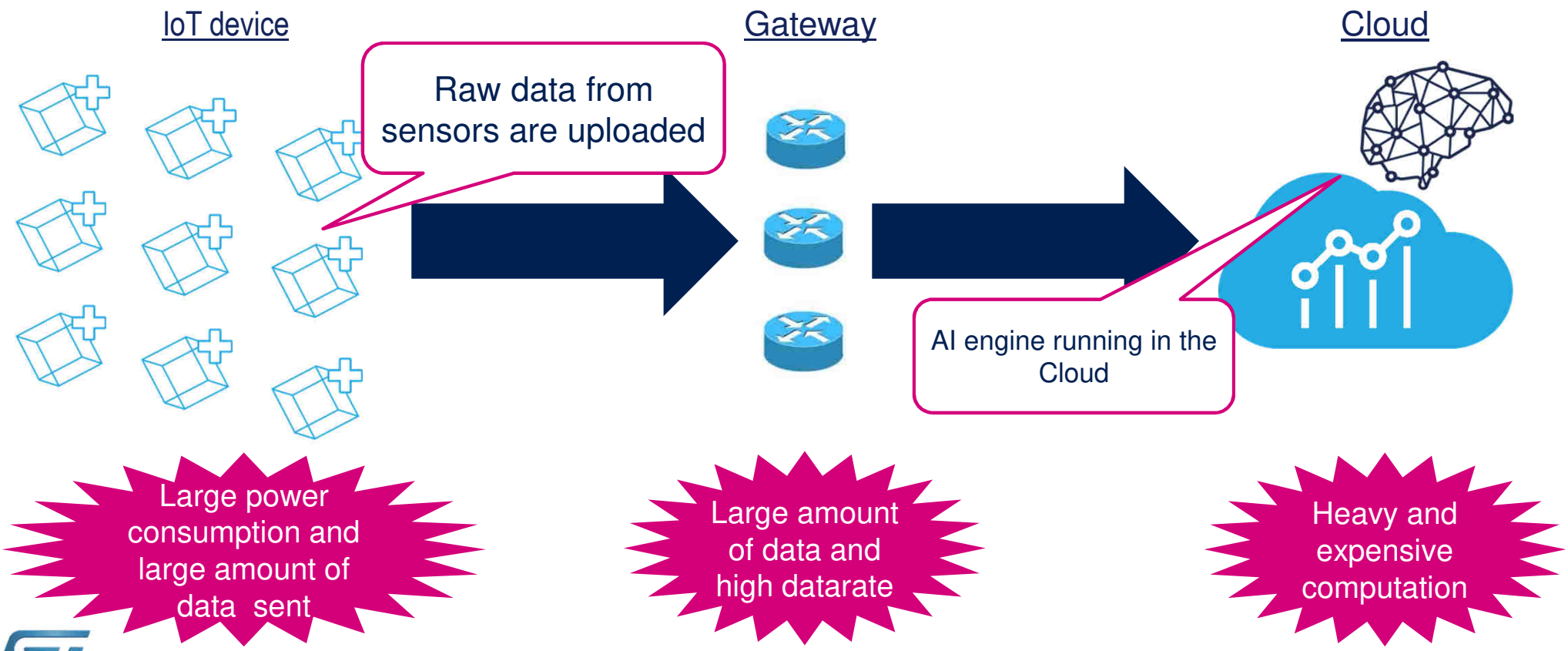
Why Deep Learning is so Important

- Convolutional Deep Neural Networks outperform alternative methods on a number of tasks:

Problem	Dataset	Best Accuracy w/o CNN	Best Accuracy with CNN	Diff
Object classification	ILSVRC	**73.8%**	**95.1%**	**+21.3%**
Scene classification	SUN	**37.5%**	**56%**	**+18.5%**
Object detection	VOC 2007	**34.3%**	**60.9%**	**+26.6%**
Fine-grained class	200Birds	**61.8%**	**75.7%**	**+13.9%**
Attribute detection	H3D	**69.1%**	**74.6%**	+5.5%
Face recognition	LFW	**96.3%**	**99.77%**	**+3.47%**
Instance retrieval	UKB	**89.3%** (CDVS: 85.7%)	**96.3%**	+7.0%

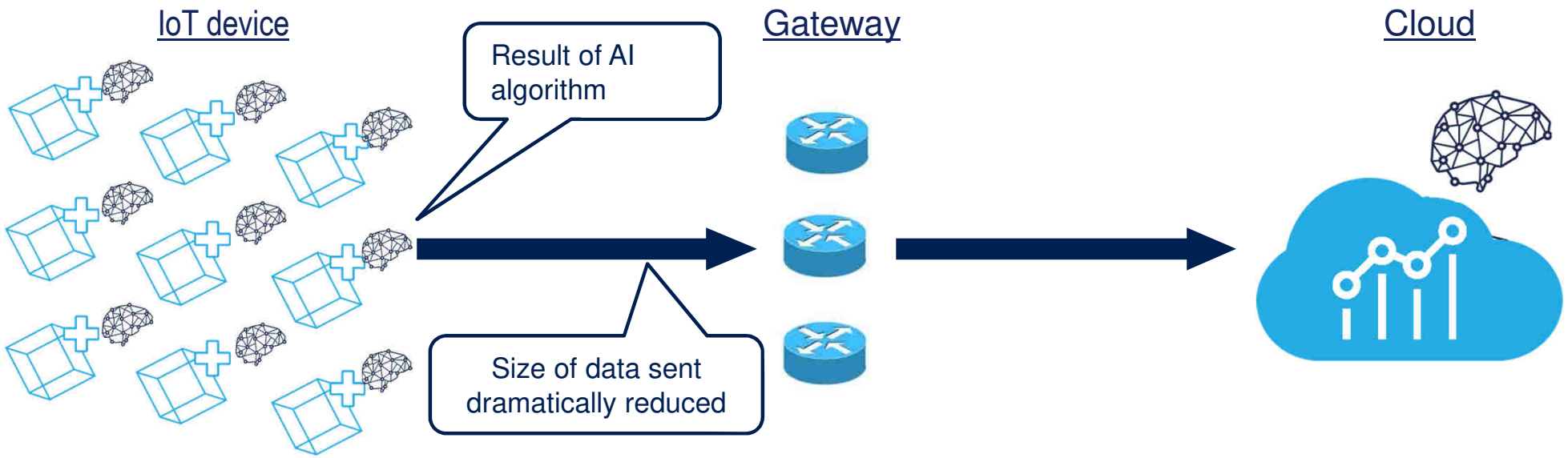
life.augmented

AI Cloud Computing



life.augmented

AI Edge Computing (Embedded)



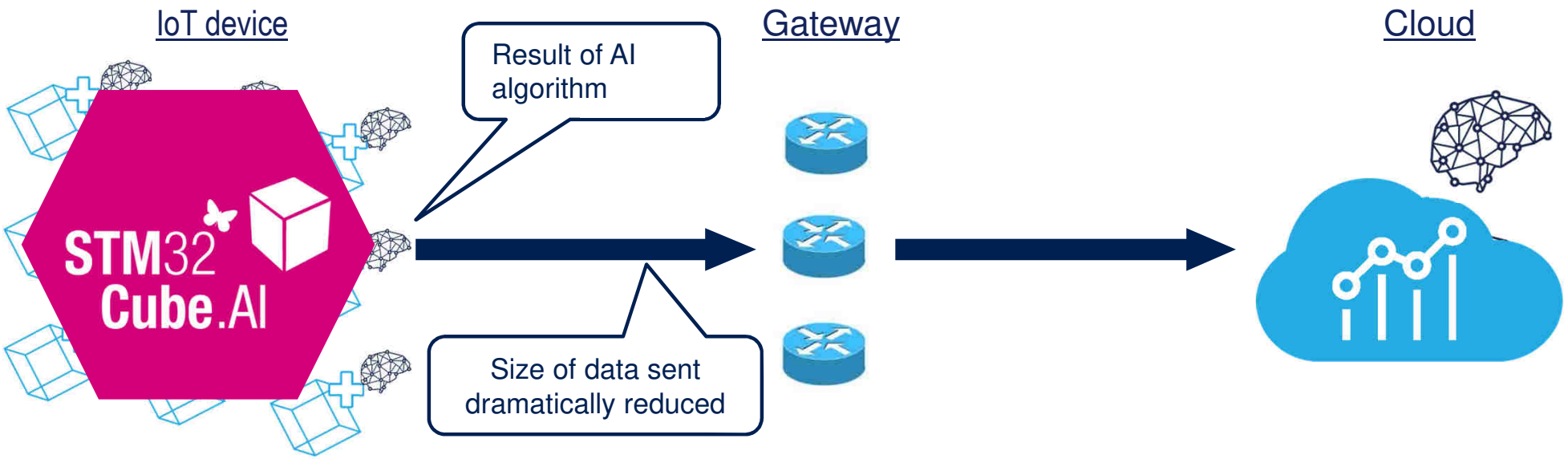
Low power consumption and small amount of data

Small amount of data and low datarate

Medium computation

life.augmented

AI Edge Computing (Embedded)



Low power consumption and small amount of data

Small amount of data and low datarate

Medium computation

life.augmented

Distributed AI

High Bandwidth

High centralized computing power

Potentially high latency

Reduced bandwidth

Lower centralized computing power

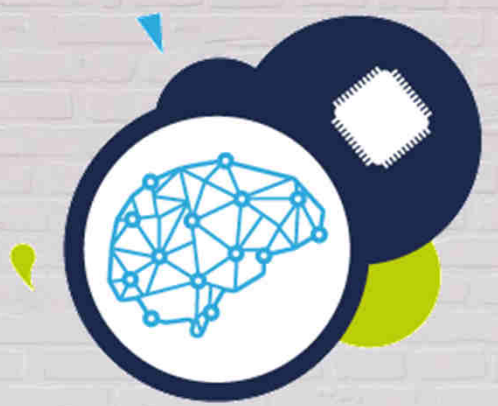
Real-time response

Preserved Privacy

life.augmented

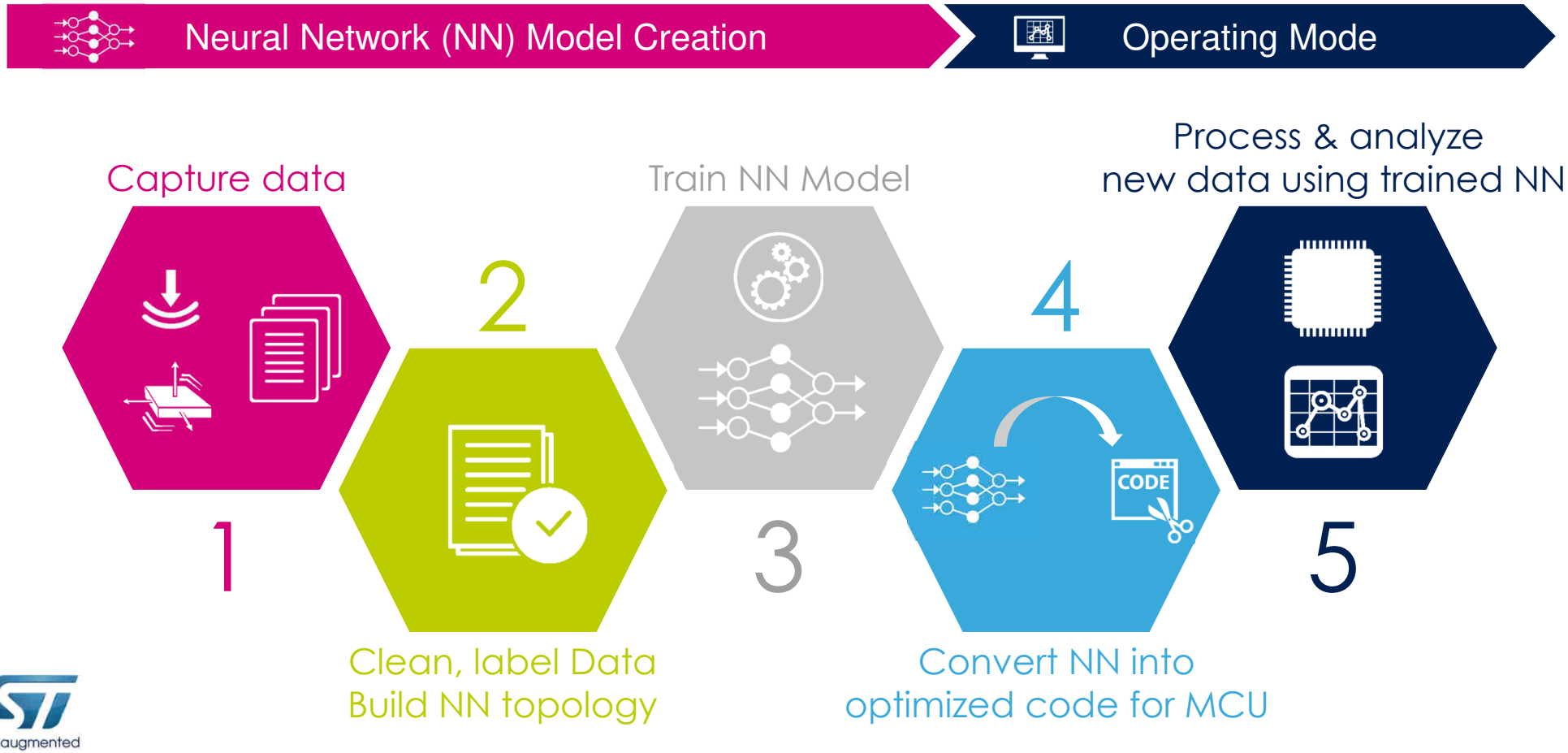
Neural Networks on STM32

Simple, fast, optimized



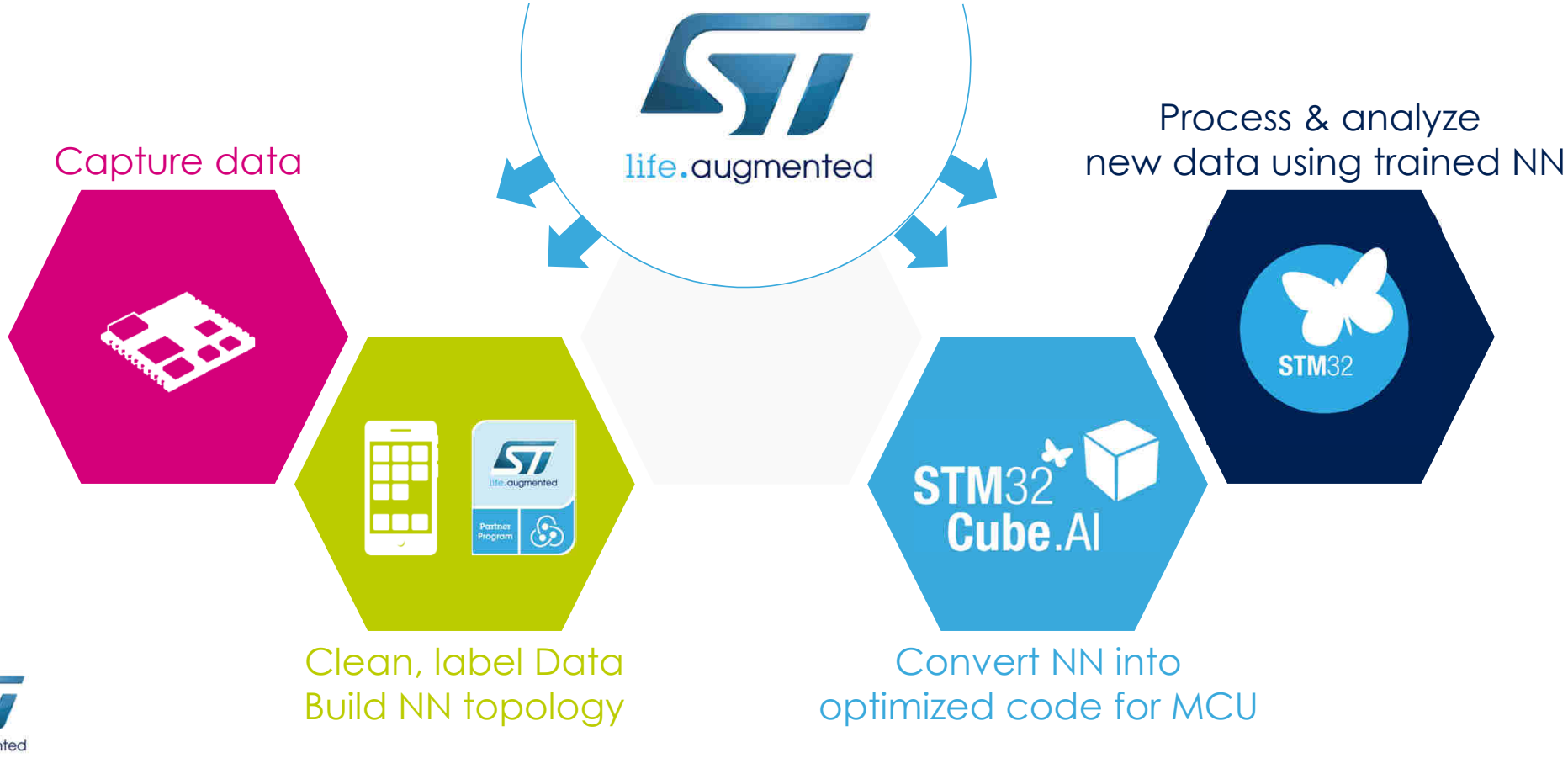
STM32 Cube.AI

The Key Steps Behind Neural Networks



life.augmented

ST Toolbox for Neural Networks



life.augmented

STM32 Solutions for Embedded AI

Extensive toolbox to easily create your AI application

STM32 Cube.AI

AI extension for STM32CubeMX
to **map pre-trained Neural Networks**

Software examples for quick prototyping
Audio, Motion and Vision Function packs
On **ST development Hardware**

STM32 **Community** with dedicated
Neural Network topic

STM32 AI Partner Program
with dedicated Partners providing
Machine or Deep Learning engineering services

life.augmented

STM32 AI Typical Applications

Low

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

Medium

- Audio & sound
- Speech Recognition
- Object detection

High

- Object detection / classification / tracking
- Natural Language Understanding / Speech Synthesis

10s MOPs **GOPs** **0.5-1 TOPs** **1-2 TOPs**

MCU → From IP embedded in MCU/MPU to dedicated SOC

SAGEMCOM

Industrial Maintenance Module

life.augmented

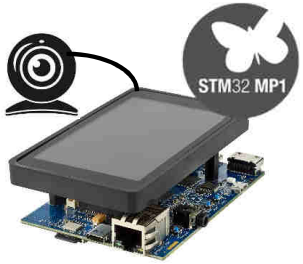
Smart Shoes
Human Activity Tracking

nnf
no new folk studio

Object & Gesture Detection device

neonode

OpenMV Machine Vision Camera



STM32MP1
Discovery board

STM32CubeMX Extension

AI Conversion Tool

Input your framework-dependent, pre-trained Neural Network into the **STM**32**Cube**.AI conversion tool

Automatic, fast generation of an STM32-optimized library

STM32**Cube**.AI offers interoperability with state-of-the-art Deep Learning design frameworks

Train NN Model

* TensorFlow used as a Keras backend. Not all operators accessible to MCUs

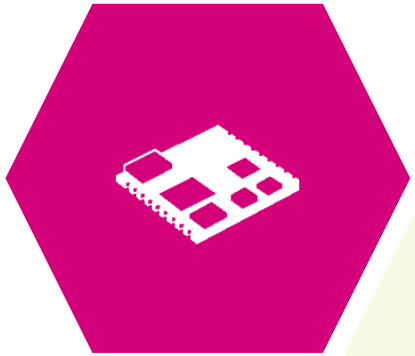
Convert NN into optimized code for MCU

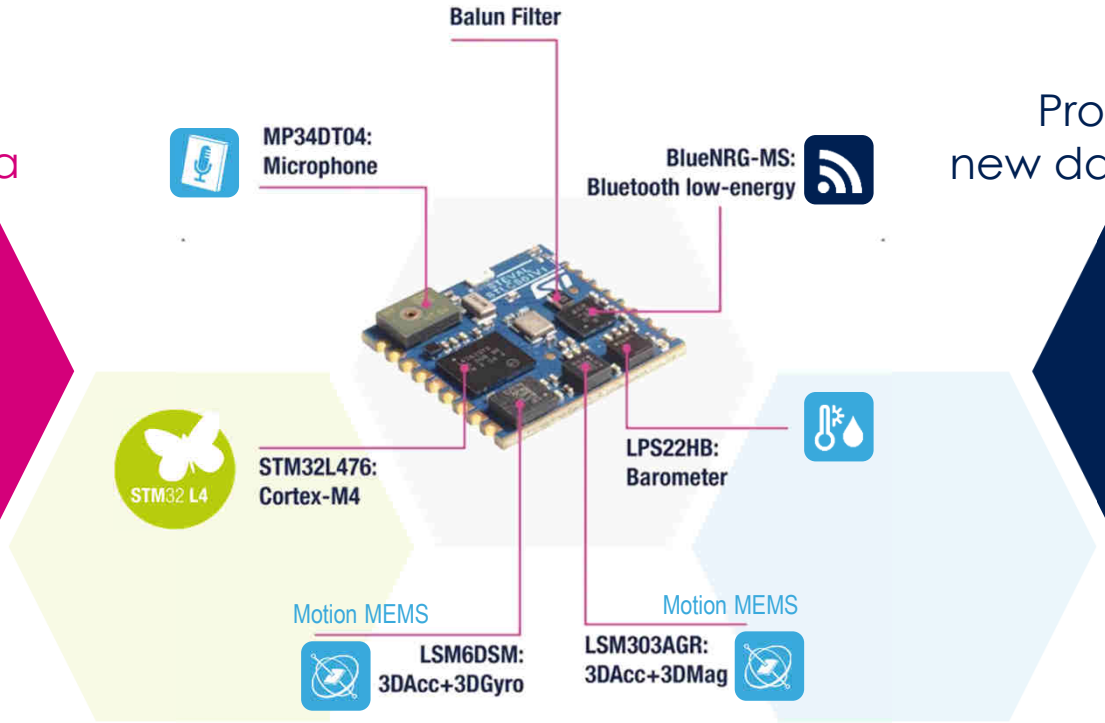
Process & analyze new data using trained NN

life.augmented

Form Factor Hardware
to Capture and Process Data

Capture data





Process & analyze
new data using trained NN

www.st.com/SensorTile
www.st.com/SensorTile-edu

life.augmented

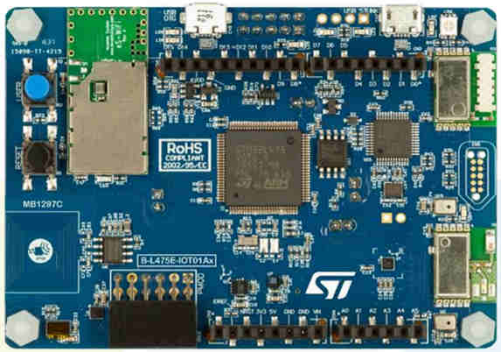
Form Factor Hardware

AI IoT Node for More Connectivity

Capture Data → Label Data → Train NN → STM32 Cube.AI → Run on STM32

Capture data

+ Sub-1GHz, Dynamic NFC Tag, Wi-Fi



Process & analyze new data using trained NN

More debug capabilities

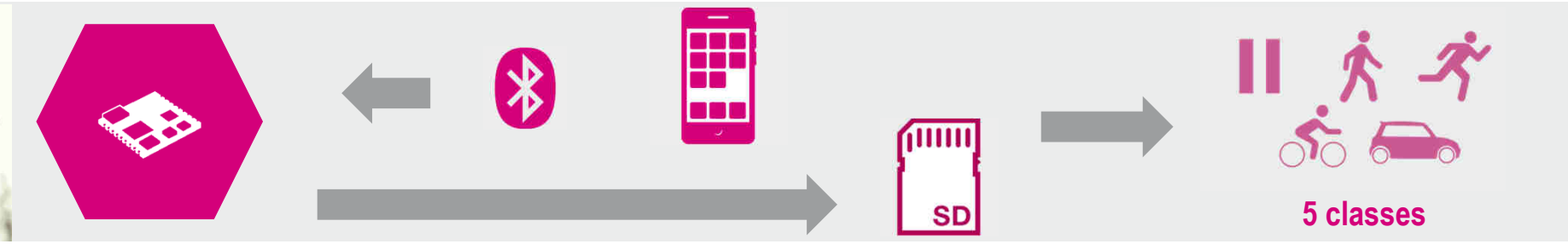
- Integrated ST-Link/V2.1
- PMOD extension connector
- Arduino Uno extension connectors

https://www.st.com/iotnode

life.augmented

Human Activity Recognition (HAR)

Motion Example in FP-AI-SENSING1 Package

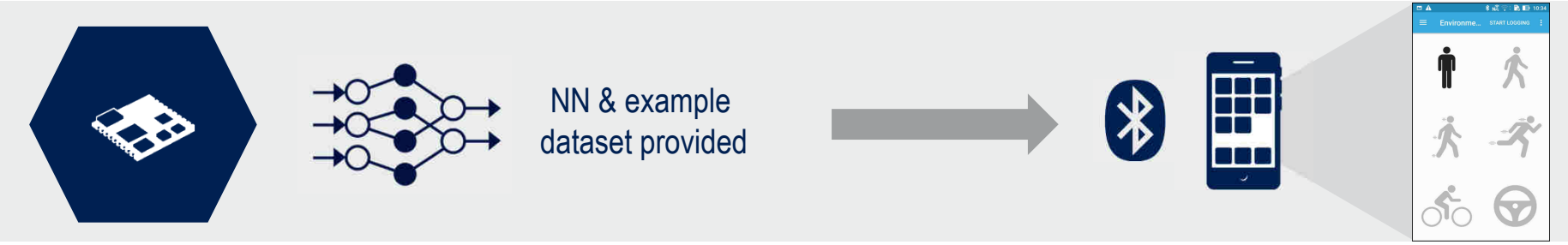


Embedded **motion**

Labelling controlled by smartphone application

Data stored on the device SD card for future **learning**

Stationary, walking, running, biking, driving



Embedded **motion** pre-processing

Inferences running on the microcontroller

Inference result displayed on mobile app

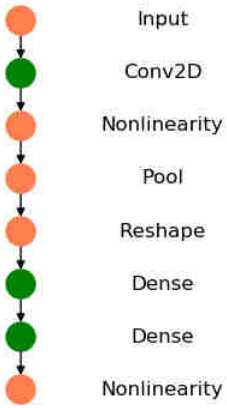
life.augmented

Human Activity Recognition
IGN (5 classes)

Use Case #2: HAR Human Activity Recognition Ignatov on SensorTile

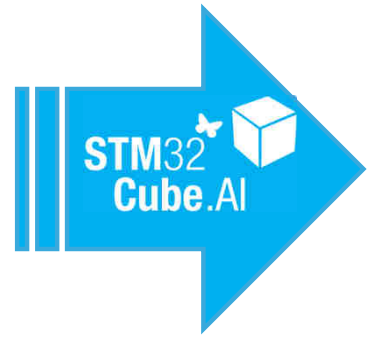
Neural Network

- Derived from a published paper Keras model
- ST proprietary dataset of 2.4M samples



Implementation

- Exploits 3-axis accelerometer data
- 5 classes: stationary, walking, running, biking, driving
- Pre/Post-processing: filtering gravity, reference rotation, temporal filter



STM32 Cube.AI NN

- Computational complexity 14k MACC
- Memory footprint: 1.8 KB RAM, 12 KB Flash

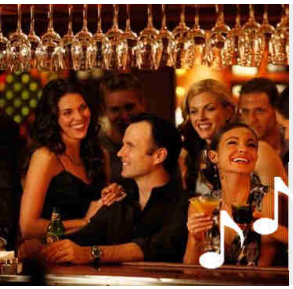
Performance on Sensor Tile

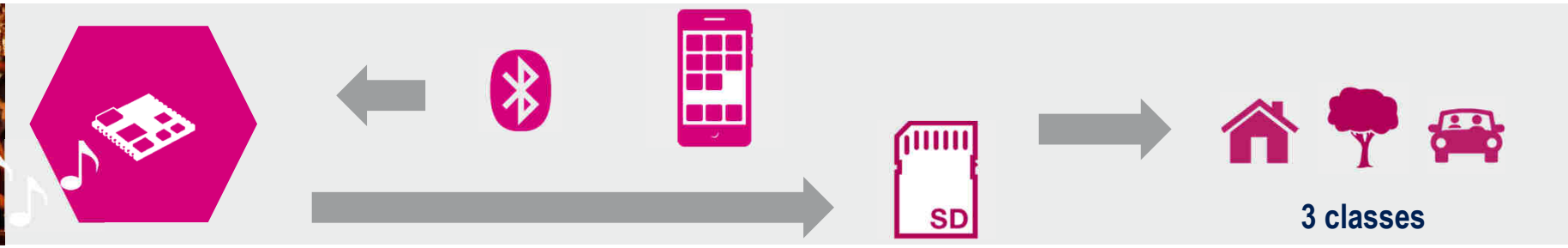
- STM32L476 80MHz Cortex-M4F
- Use case: 1 classification/sec
- Pre/Post-processing: 0.02 MHz NN processing: 0.35 MHz
- Power consumption (1.8 V)
  - System: 580 uA (with optim BLE)
  - STM32: 510 uA

life.augmented

Audio Scene Classification (ASC)

Audio Example in FP-AI-SENSING1 Package



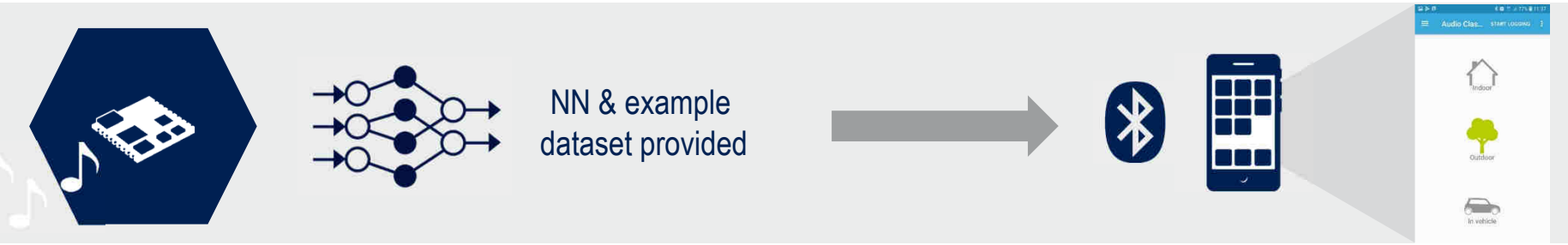


Embedded **audio**

Labelling controlled by smartphone application

Data stored on the device SD card for future **learning**

Indoor, Outdoor, In vehicle



Embedded **audio** pre-processing

Inferences running on the microcontroller

Inference result displayed on mobile app

Image Classification

Vision Example in FP-AI-VISION1 Package

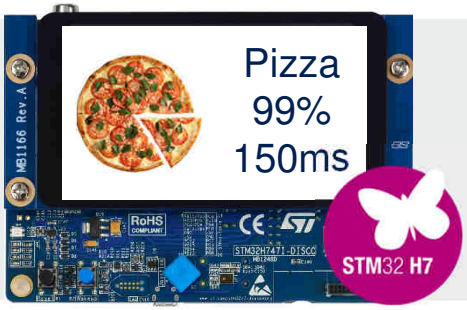
Enjoy the food classification demo

- Default demo based on 18 classes (224x224 RGB pictures)
- Several camera image output size possible

Full end-to-end optimized software example

- from camera acquisition to image pre-processing before feeding the NN
- Multiple memory mapping possibilities to optimize and test impact on performances
- Retrain this NN with your own dataset
- Quantize your trained network to optimized inference time and memory usage

NN & example dataset provided



Embedded **image** pre-processing (SW) on the STM32H747

Inferences running on the microcontroller

Inference result displayed on STM32H747 Discovery board LCD display

life.augmented

Food Recognition

Fast Downsampling MobileNet Food Recognition on STM32H747 Dual-Core Discovery board

Neural Network

- FD-MobileNet topology from public paper applied to food
- Dataset

Implementation Details

- Uses Camera either in continuous or one shot mode
- Floating Point or mixed model Floating/Fix Point
- 18 food classes

STM32 Cube.AI NN

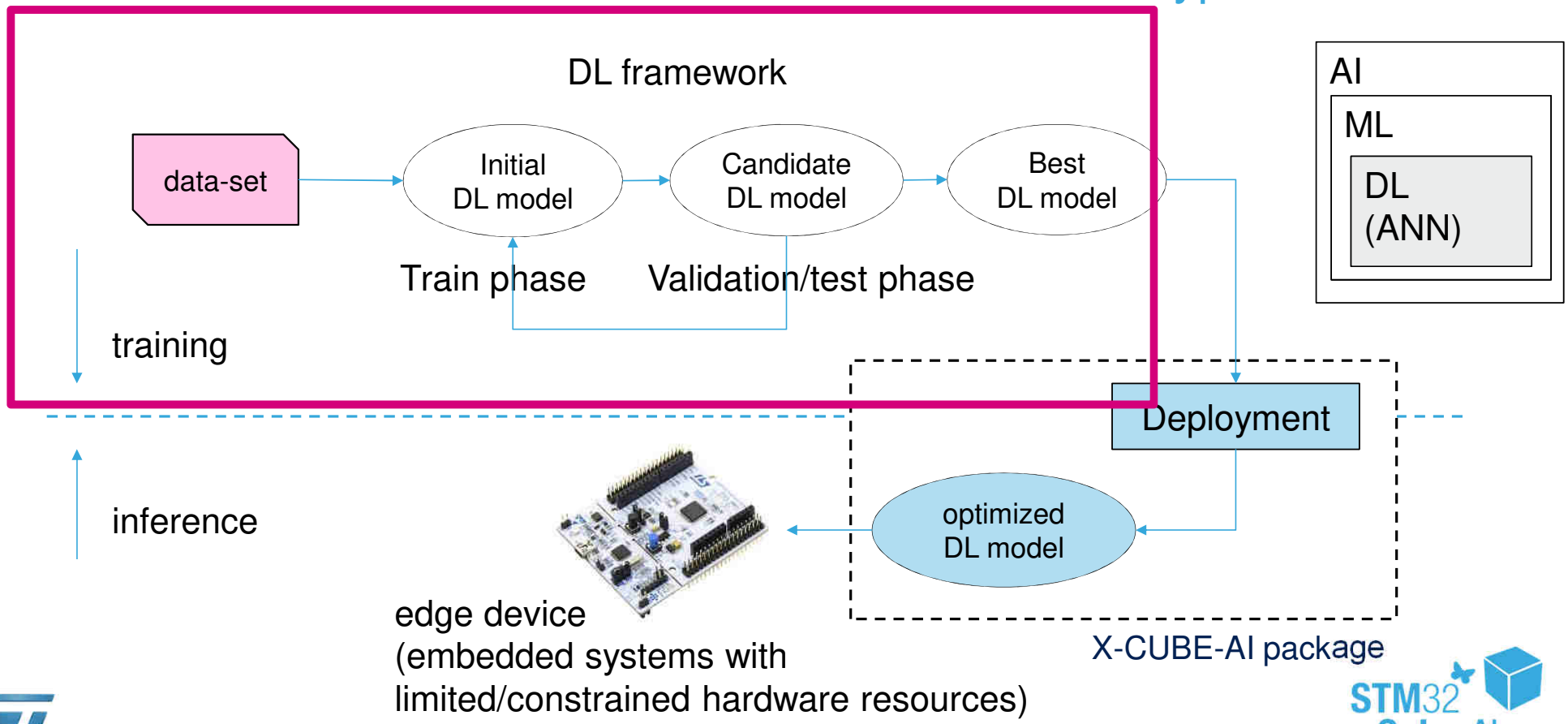
- Memory footprint: 205 KB RAM, 191 KB Flash

Performance on STM32H747

- 1 inference per image
- STM32H747 400 MHz Cortex-M7F
- Mix model Fix/Floating Point
  - 6.2 MHz / 150 ms per inference
  - Accuracy: 78.8%

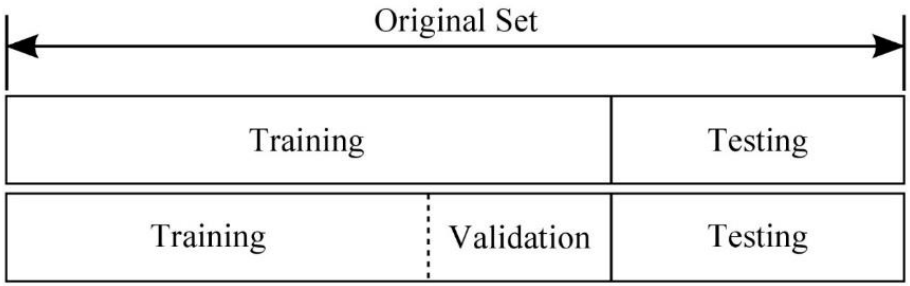
X-CUBE-AI Positioning

in a typical DL flow



Learning datasets

- One of the difficulties of working with NN is to have a big dataset. In general 90% of the time spent for a Deep Learning project is related to the creation of the Dataset.
- For Image Processing and few other applications there are already available big databases (with labels), both free or paid basis. On other hand for new and “niche” applications the designer needs to create the database, that becomes an high value asset.
- The dataset is then divided into
  - Training set
    - Training set
    - Validation set
  - Testing set



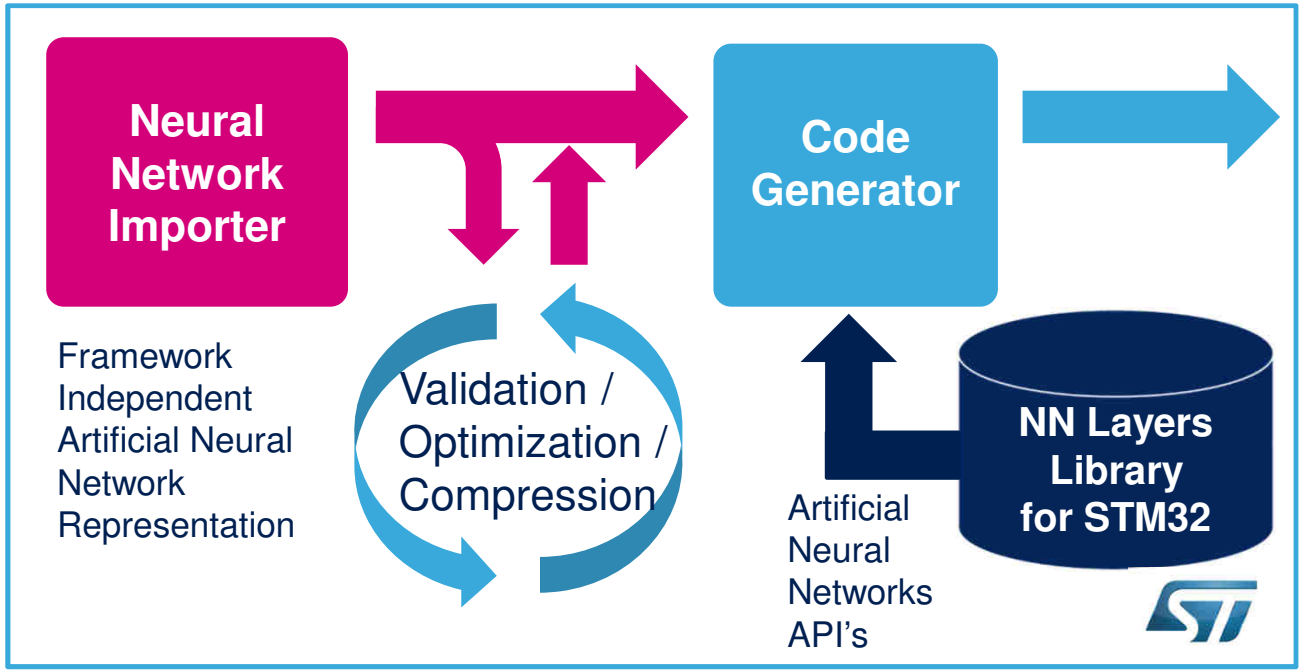
Data-driven approach

- *We provide many examples of each class to the computer and then develop **machine learning algorithms** that look at these examples and **learn about the visual appearance** of each class.*

X-Cube-AI : Architecture

Off-the-shelf :
Pre-trained Artificial Neural Network Model

Deep Learning Framework dependent

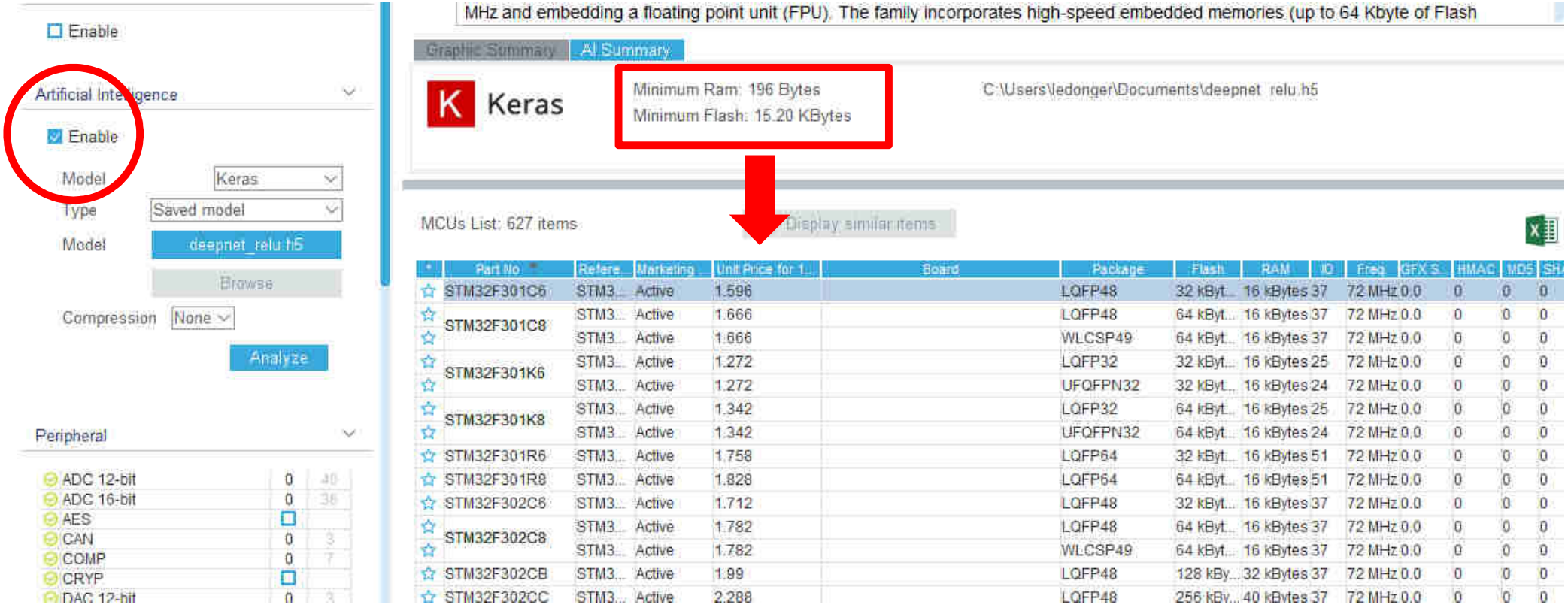


Embedded Solution
Optimized Artificial Neural Network Code generated for STM32

This optimized STM32 Artificial neural network model can be included into the user project (using KEIL, IAR, OpenSTM32) and can be compiled and ported onto the final device for field trials

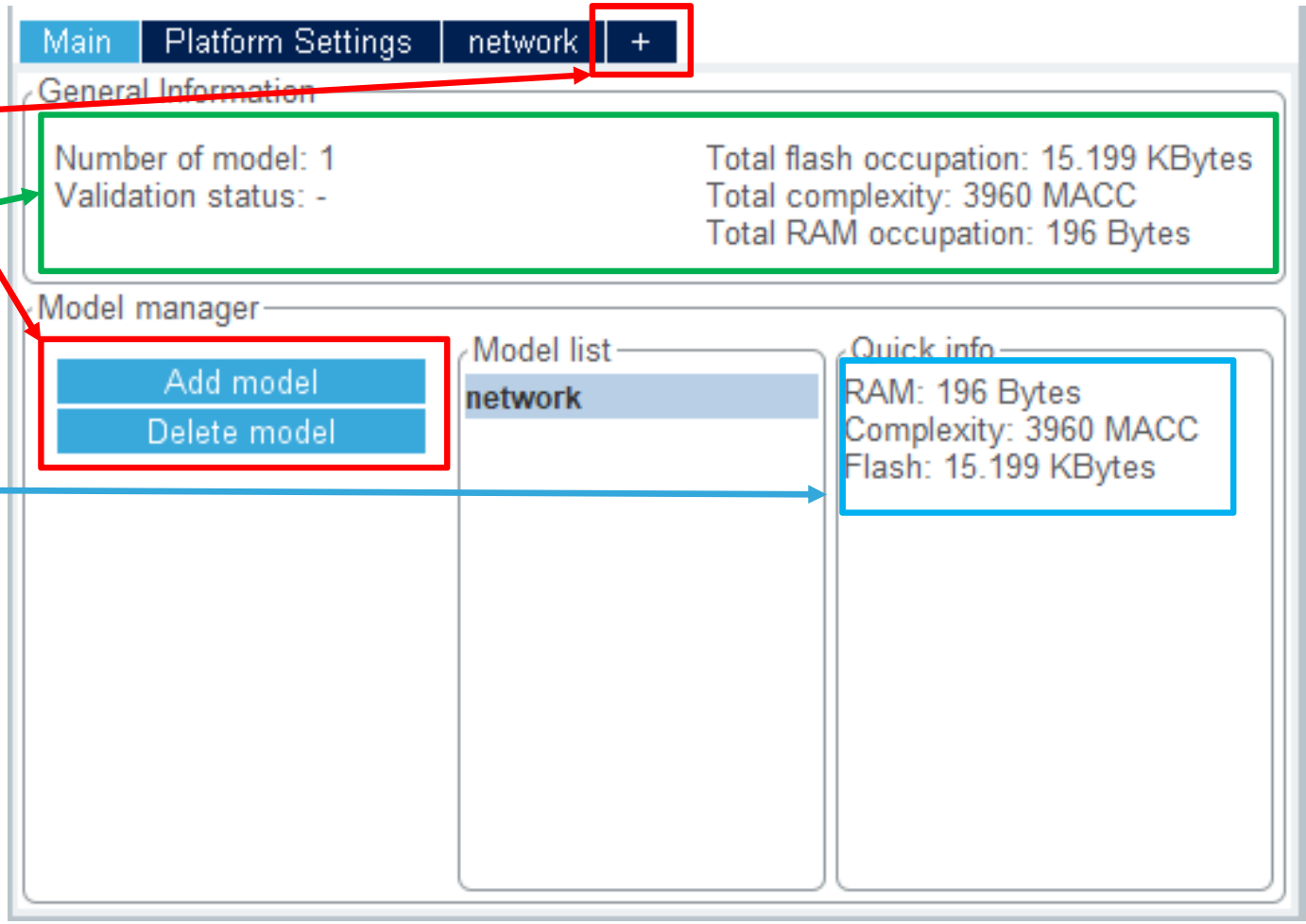
life.augmented

MCU Finder AI Filter



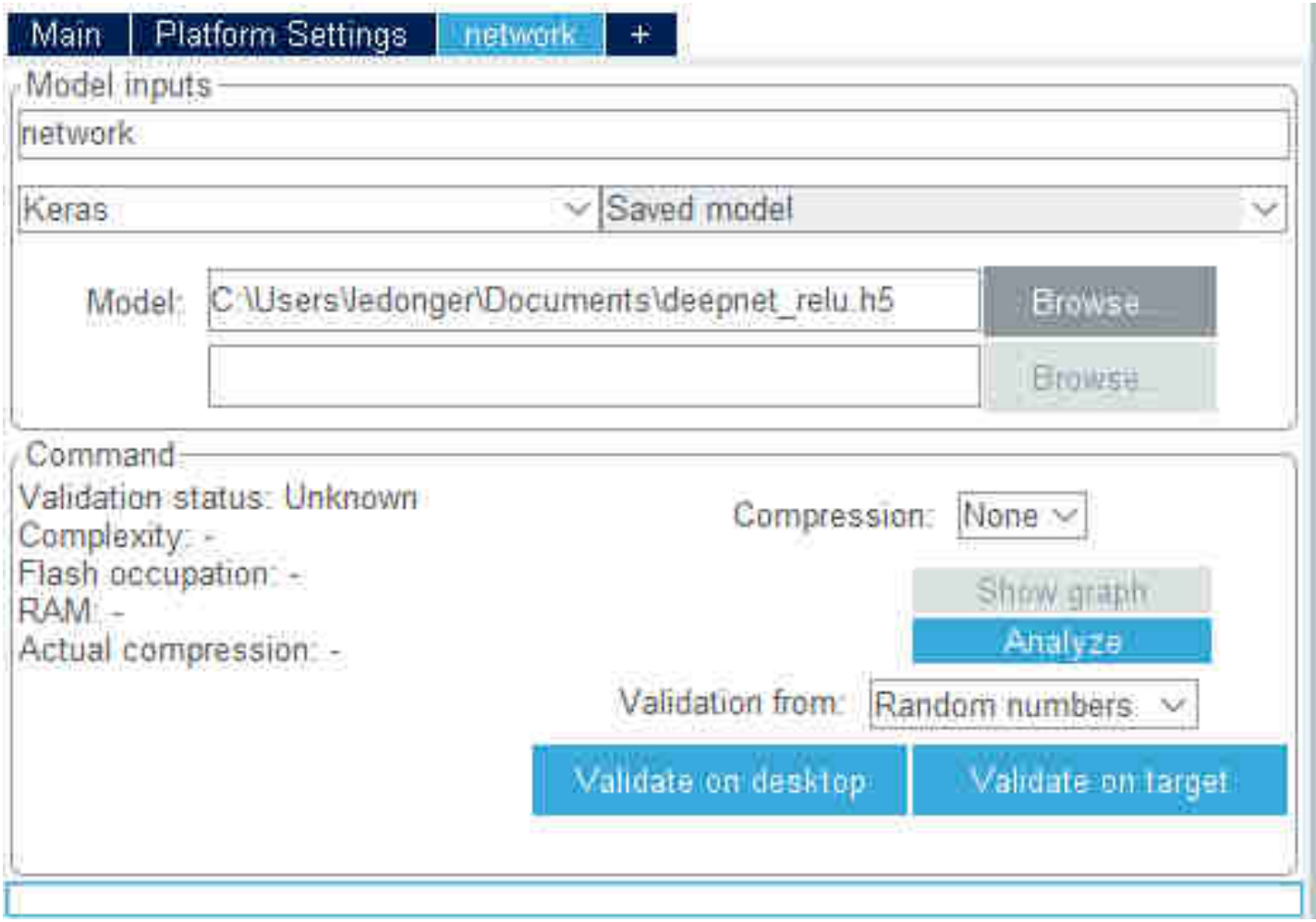
X-Cube-AI Main tab

- Add/Delete models
- Get general information
- Have a quick look at different models



X-Cube-AI Detailed View

- Perform **analysis** to compute the model size, get an image of the network and the complexity
- Perform **validation on desktop**
- Perform **validation on target**
- Set a **compression** to reduce the model size (By reducing the accuracy of the model)



Making AI Accessible Now

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

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

Compatible with **Machine Learning** Partner ecosystems

2007	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
World 1st Cortex-M MCU STM32 F1	World 1st Cortex-M Ultra-low-power STM32 L1	1st High Perf. 120 MHz, 90nm STM32 F2	1st High Perf. Cortex-M4 168 MHz STM32 F4	1st Mixed Signal DSP + Analog Cortex-M4 STM32 F3 STM32 F0 Entry Cost Cortex-M0	Entry Cost Ultra-low-power STM32 L0	World 1st Cortex-M7 STM32 F7	Leadership Ultra-low-power Cortex-M4 STM32 L4 447 ULPBench™ #1 ULP	#1 Performance 2400 CoreMark STM32 H7 STM32 L4+ Ultra-low-power Excellence		Mainstream Cortex-M0+ MCUs Efficiency at its best! STM32 G0 STM32 WB Dual-core, multi-protocol and open radio	First STM32 MPU Dual Cortex-A7 + Cortex-M4 STM32 meets Linux STM32 MP1 STM32 L5 Introduction of Cortex-M33 Excellence in ULP with more security

More than 60,000 customers

Over 4 Billion STM32 shipped since 2007

life.augmented

STM32Cube.AI Roadmap

2019 | Feb | Mar | Apr | May | Jun | Jul | Aug | Sept | Oct | Dec | **2020** | Jan | Feb | Mar | Apr

STM32 Cube.AI

- Market Introduction
- Floating Point Support

- Additional layers

- Fixed Point Quantization
- Command line interface
- UI Improvements
- Additional layers

- Integer Arithmetic Quantization
- Advanced External Flash Support

- ONNX introduction
- Additional layers
- Debug improvements

NN Training Tools Supported

Keras, ConvNetJS (Deep Learning in your browser), Lasagne, Caffe

+ TensorFlow Lite
- Keras Fixed Point
- TensorFlow Lite Float

+ TensorFlow Lite
- TensorFlow Lite Integer arithmetic quantization

Keras + ONNX: mxnet, Microsoft CNTK, Chainer, PyTorch

New layers & operators addition

life.augmented

STM32 Solutions for AI

More Than Just the STM32Cube.AI

An extensive toolbox to support easy creation of your AI application

AI extension for STM32CubeMX

To map pre-trained Neural Networks onto the STM32

Function packs for Quick prototyping

Audio and motion examples

SensorTile reference hardware

To run inferences or data collection

… And more coming!

STM32 Community with dedicated Neural Networks topic

Mobile phone application

To collect and label data

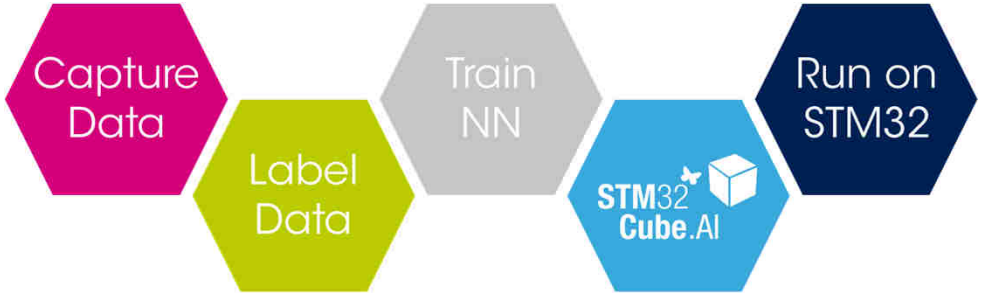
To display the result of inference processing on the STM32

ST Partner Program with a dedicated group of Partners providing Neural Networks engineering services

Data scientists and Neural network architects

For More Information

www.st.com/STM32CubeAI



life.augmented