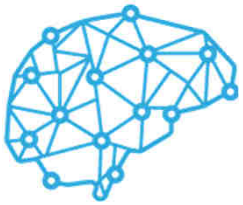


Artificial Neural Network Mapping Made Simple with the STM32Cube.AI

Markus Mayr
Product Marketing Manager, MCU

life.augmented

Technology Tour 2019
Vancouver, BC | September 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

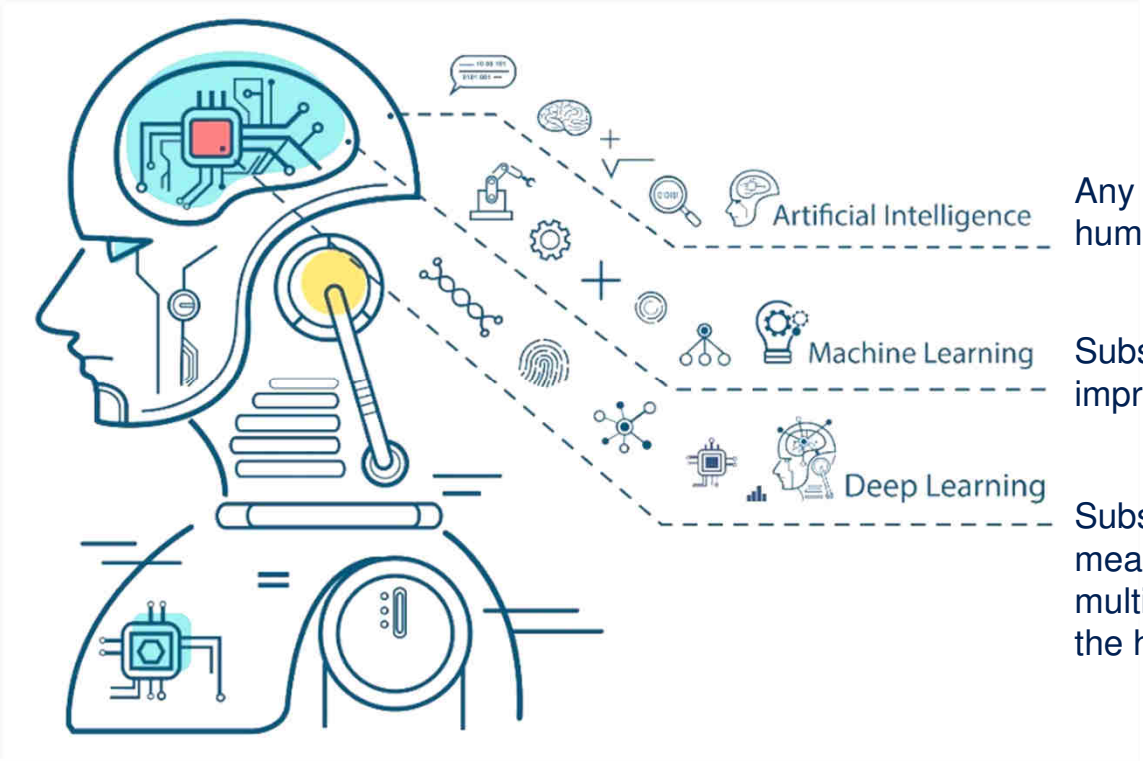
life.augmented

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

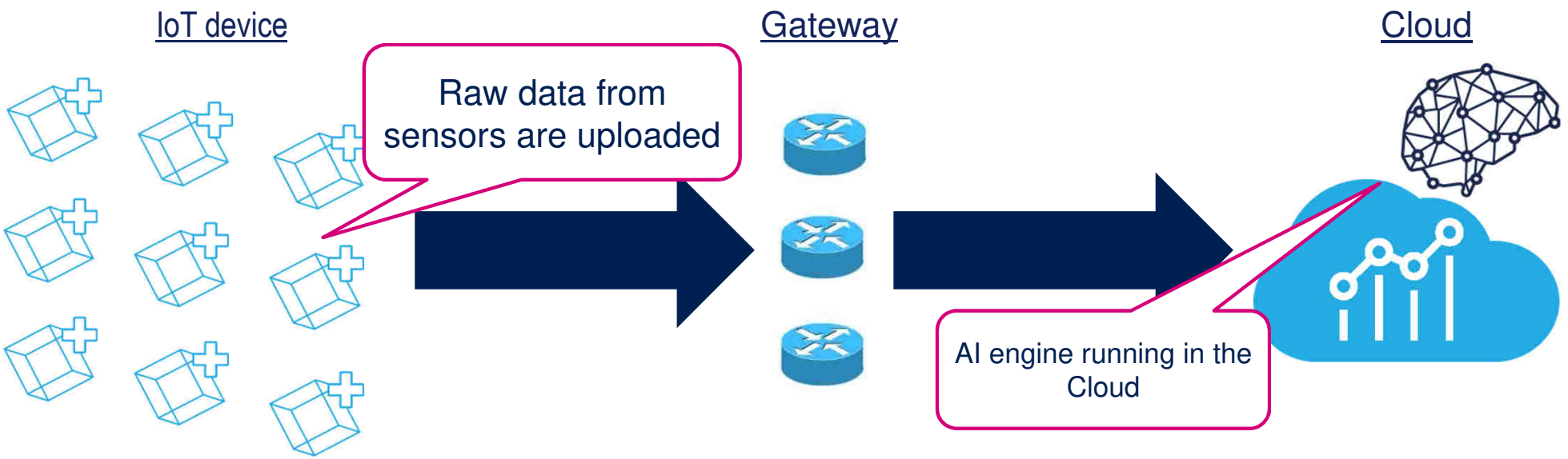
Subset of ML. Learning algorithms that derive meaning out of data, by using a hierarchy of multiple layers that mimic the neural networks of the human brain

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%

AI Cloud Computing



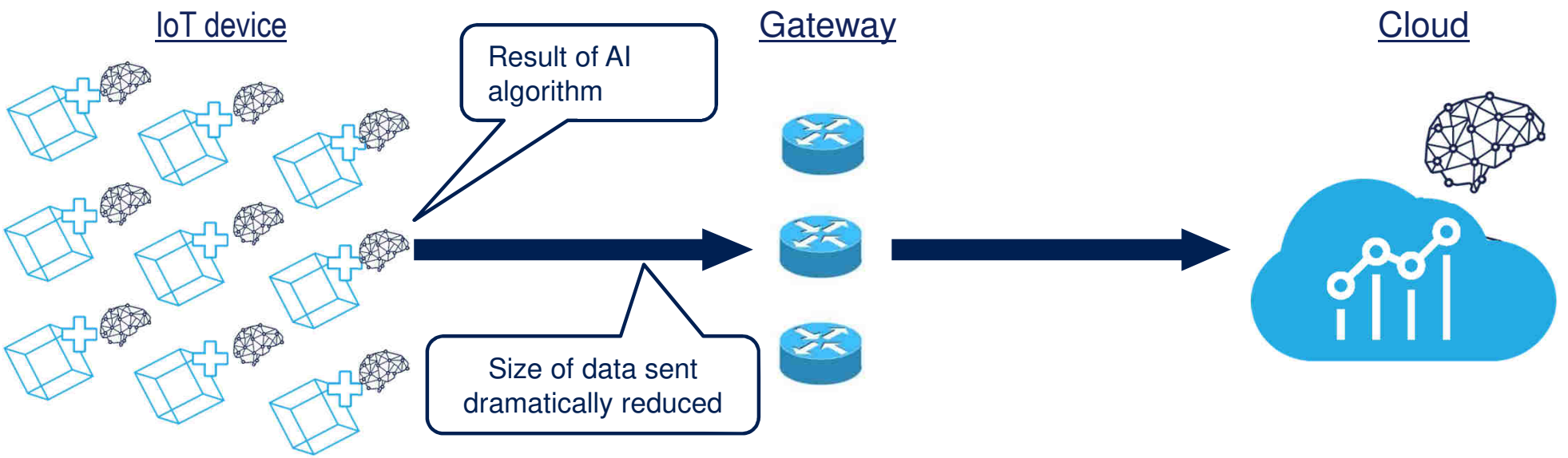
Large power consumption and large amount of data sent

Large amount of data and high datarate

Heavy and expensive computation

life.augmented

AI Edge Computing (Embedded)



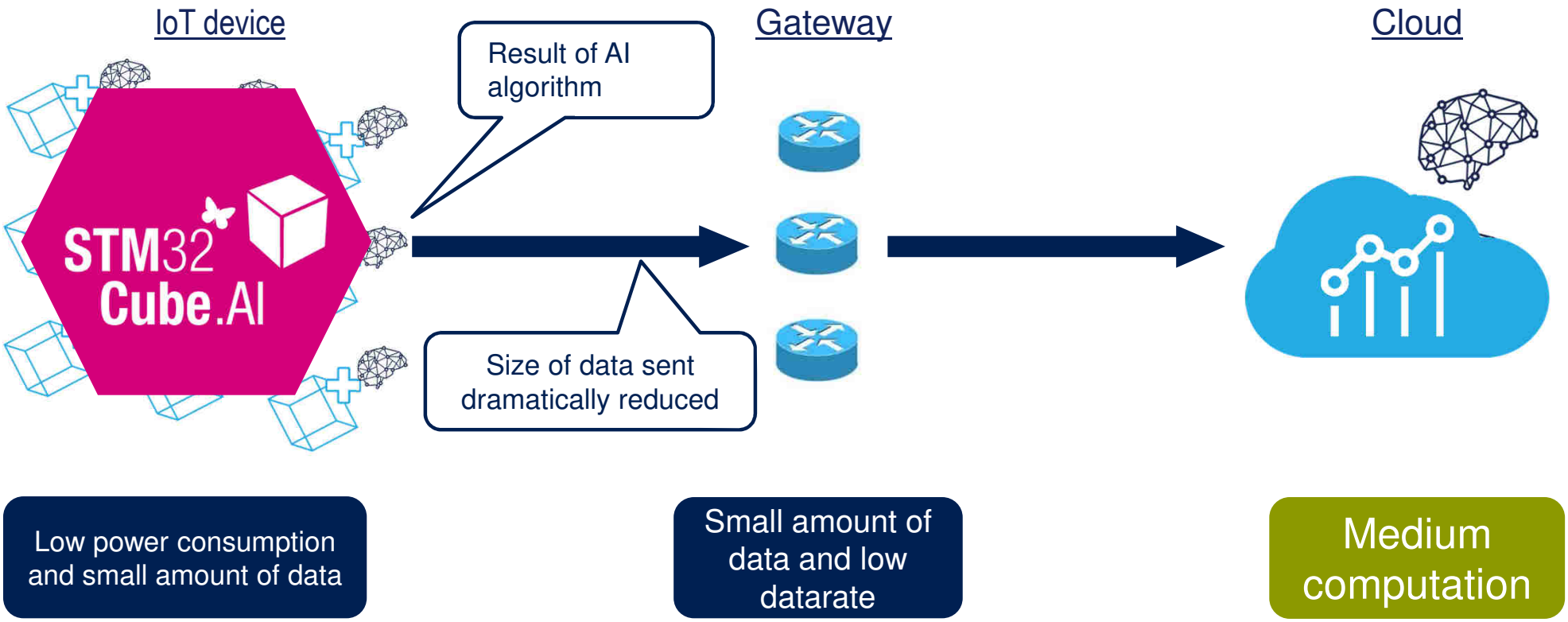
Low power consumption and small amount of data

Small amount of data and low datarate

Medium computation

life.augmented

AI Edge Computing (Embedded)



life.augmented

Distributed AI

High Bandwidth

High centralized computing power

Potentially high latency

Reduced bandwidth

Lower centralized computing power

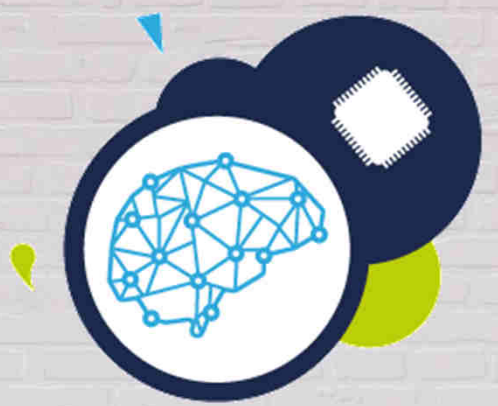
Real-time response

Preserved Privacy

life.augmented

Neural Networks on STM32

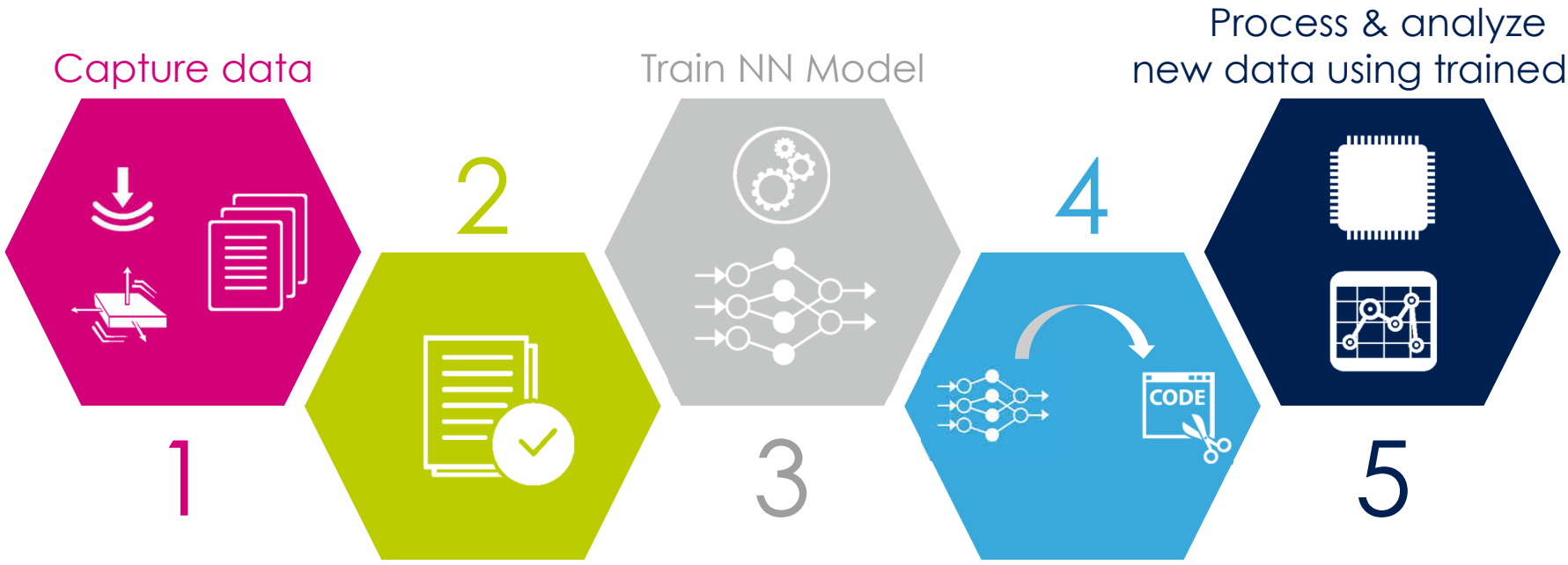
Simple, fast, optimized



STM32 Cube.AI

The Key Steps Behind Neural Networks

Neural Network (NN) Model Creation | Operating Mode



life.augmented

ST Toolbox for Neural Networks

life.augmented

Capture data

Process & analyze new data using trained NN

STM32

STM32 Cube.AI

Clean, label Data
Build NN topology

Convert NN into optimized code for MCU

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	Medium	High
• Sensor analysis • Activity recognition (motion sensors) • Stress analysis or attention analysis	• Audio & sound • Speech Recognition • Object detection	• 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

Smart Shoes
Human Activity Tracking

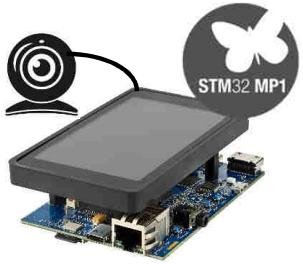
nnf
no new folk studio

Object & Gesture Detection device

neonode

OpenMV Machine Vision Camera

STM32xx



STM32MP1
Discovery board

STM32 MPx

life.augmented

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

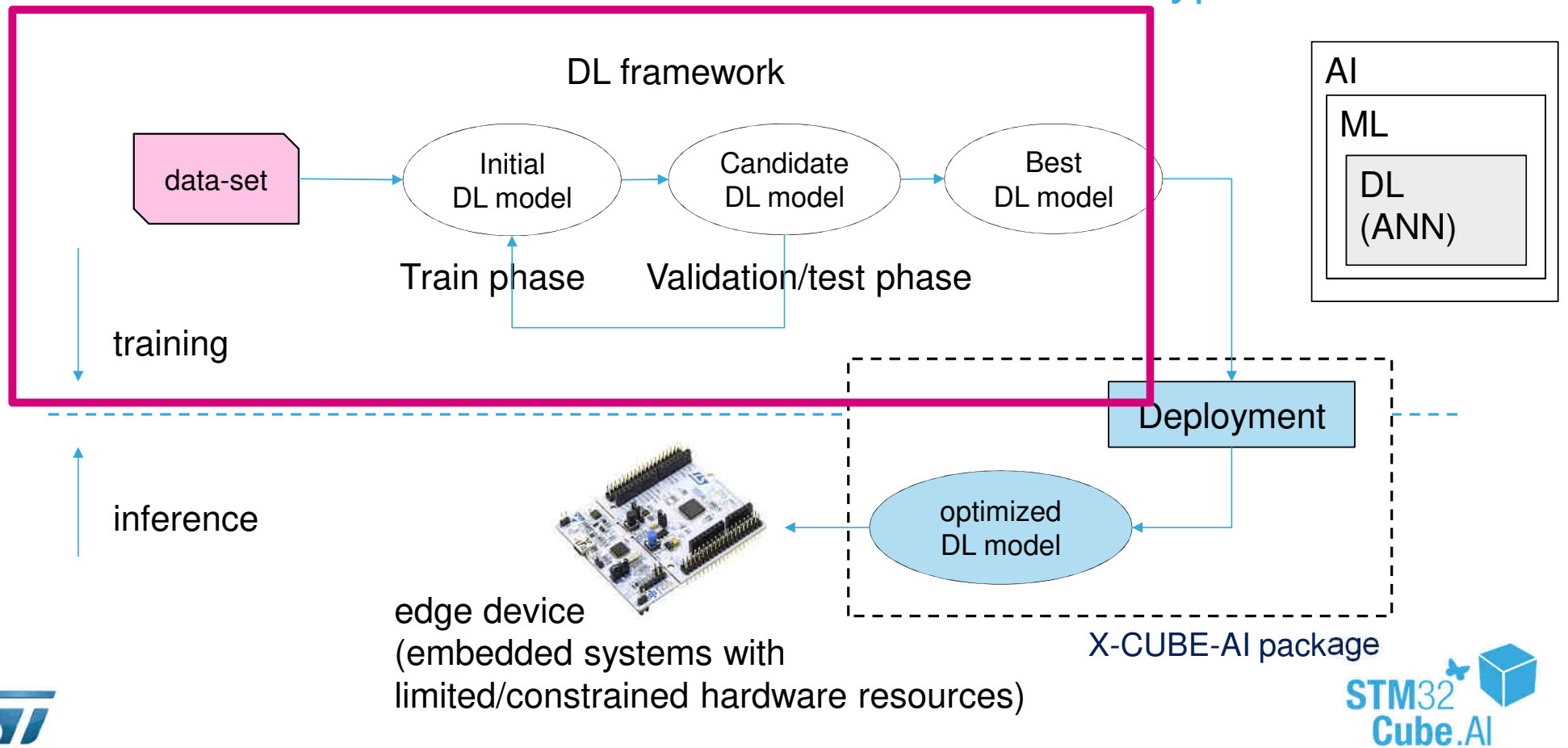
Convert NN into optimized code for MCU

Process & analyze new data using trained NN

life.augmented

X-CUBE-AI Positioning

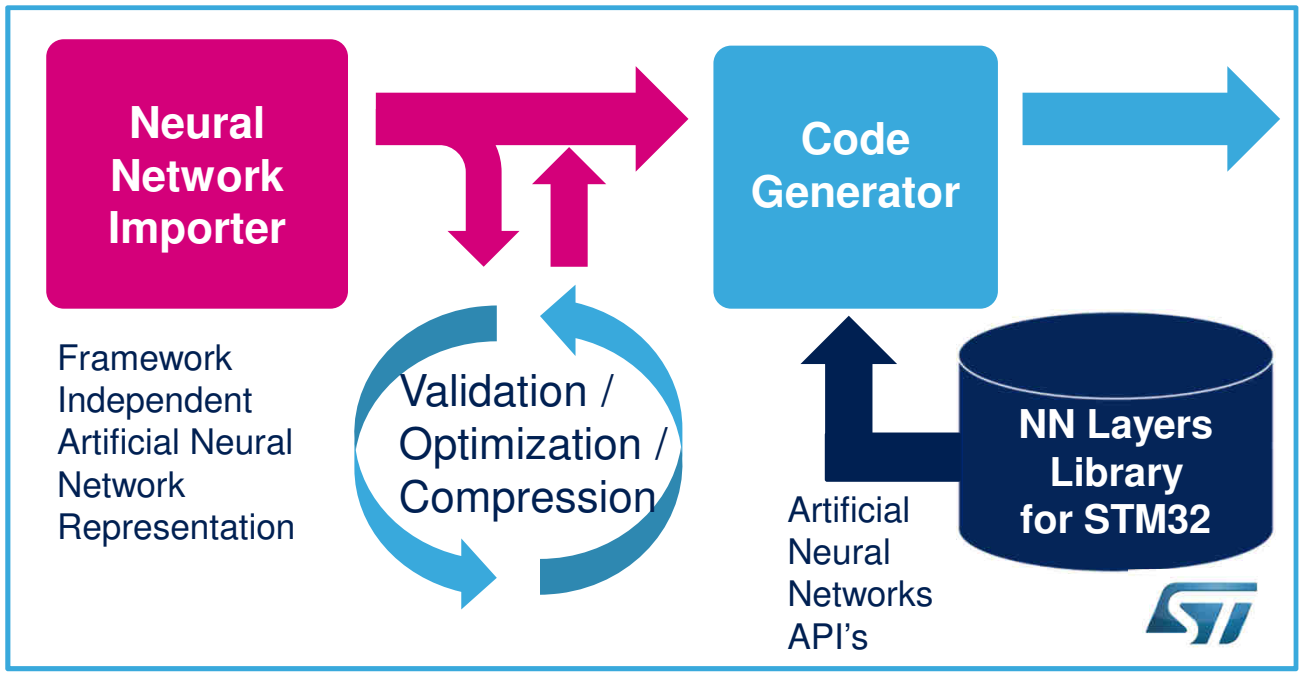
in a typical DL flow



X-Cube-AI : Architecture

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

Deep Learning Framework dependent

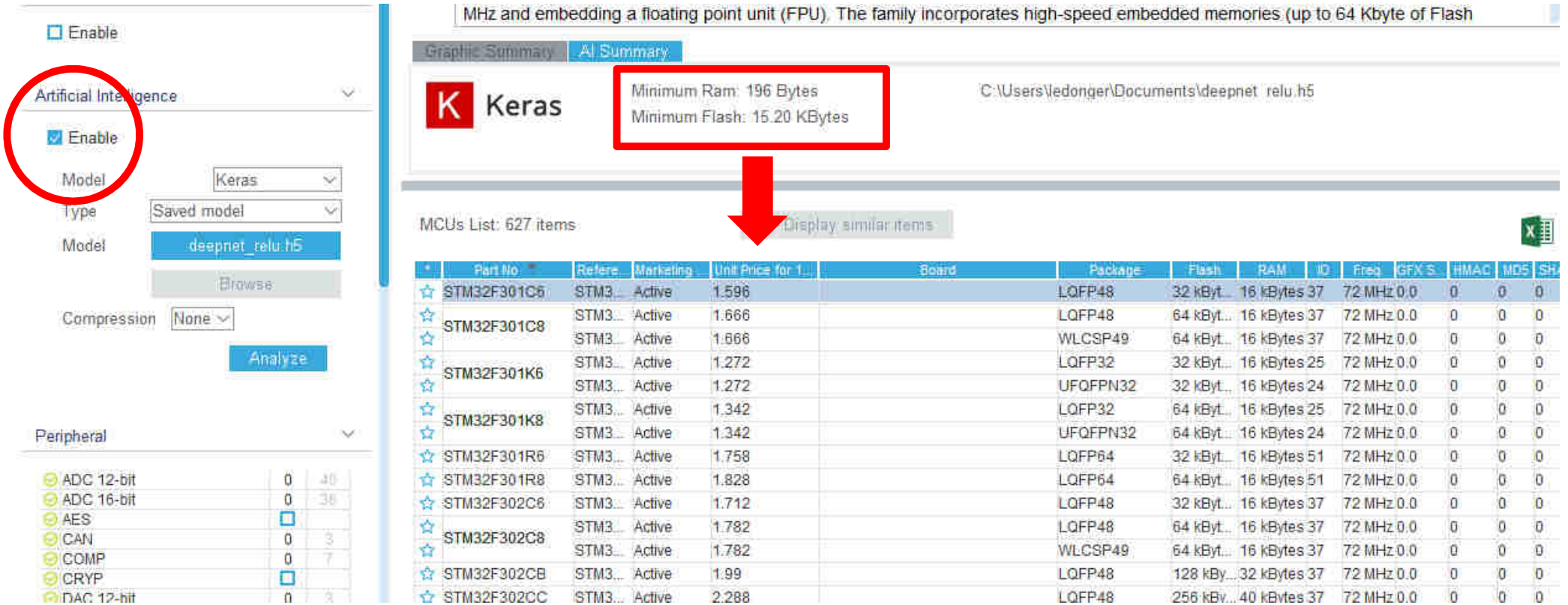


Embedded Solution
Optimized Artificial Neural Network Code generated for STM32

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

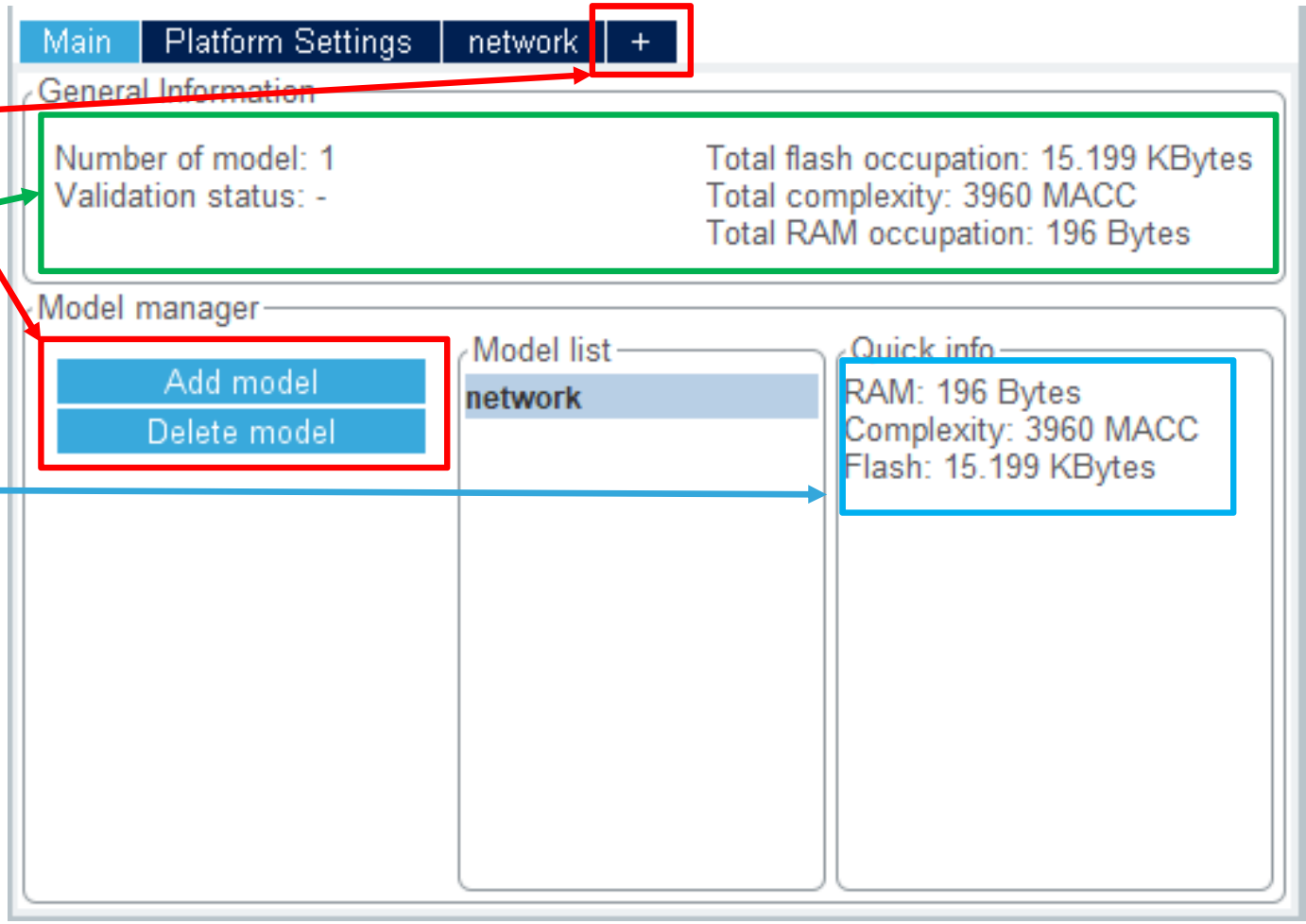
life.augmented

MCU Finder AI Filter



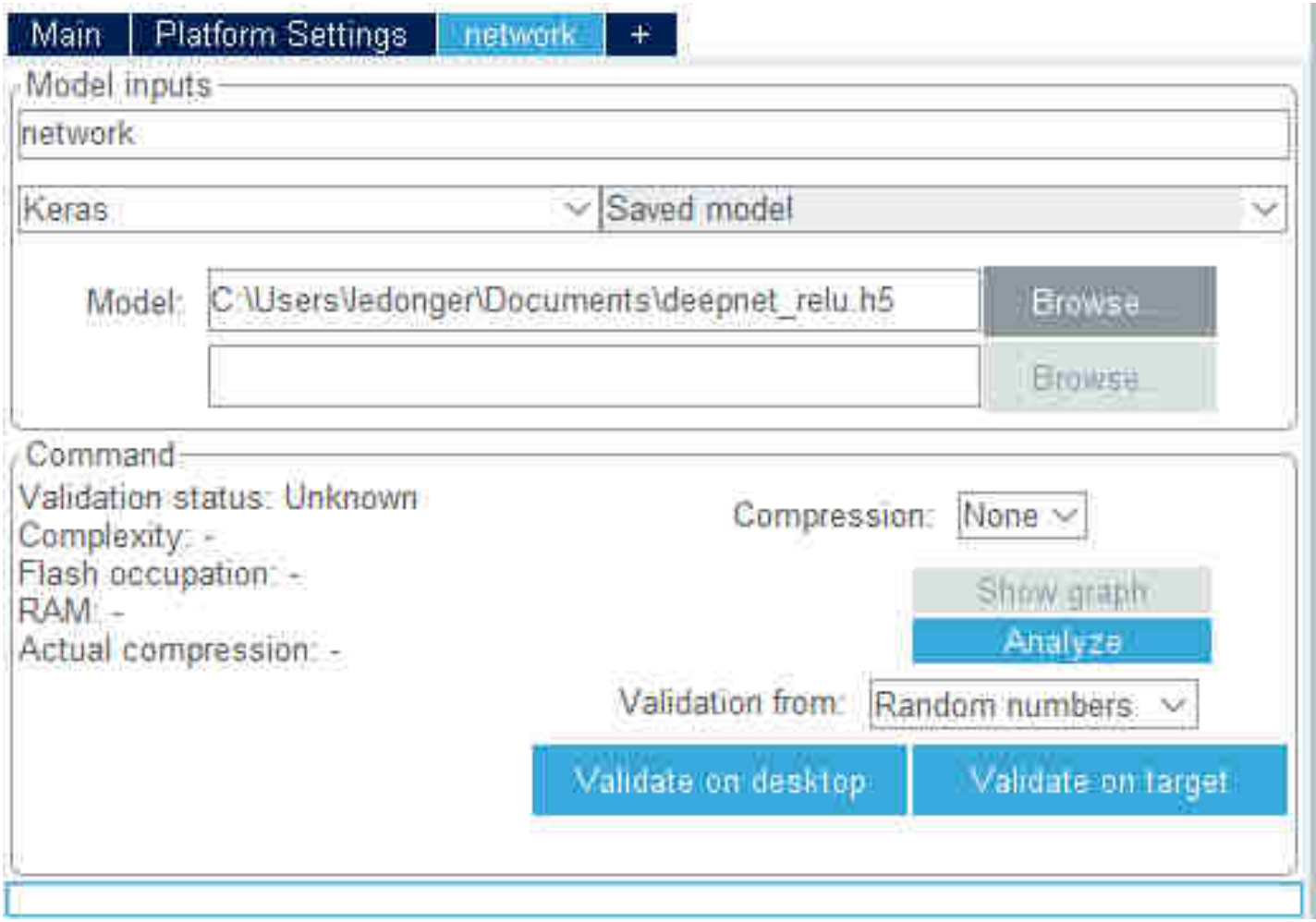
X-Cube-AI Main tab

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



X-Cube-AI Detailed View

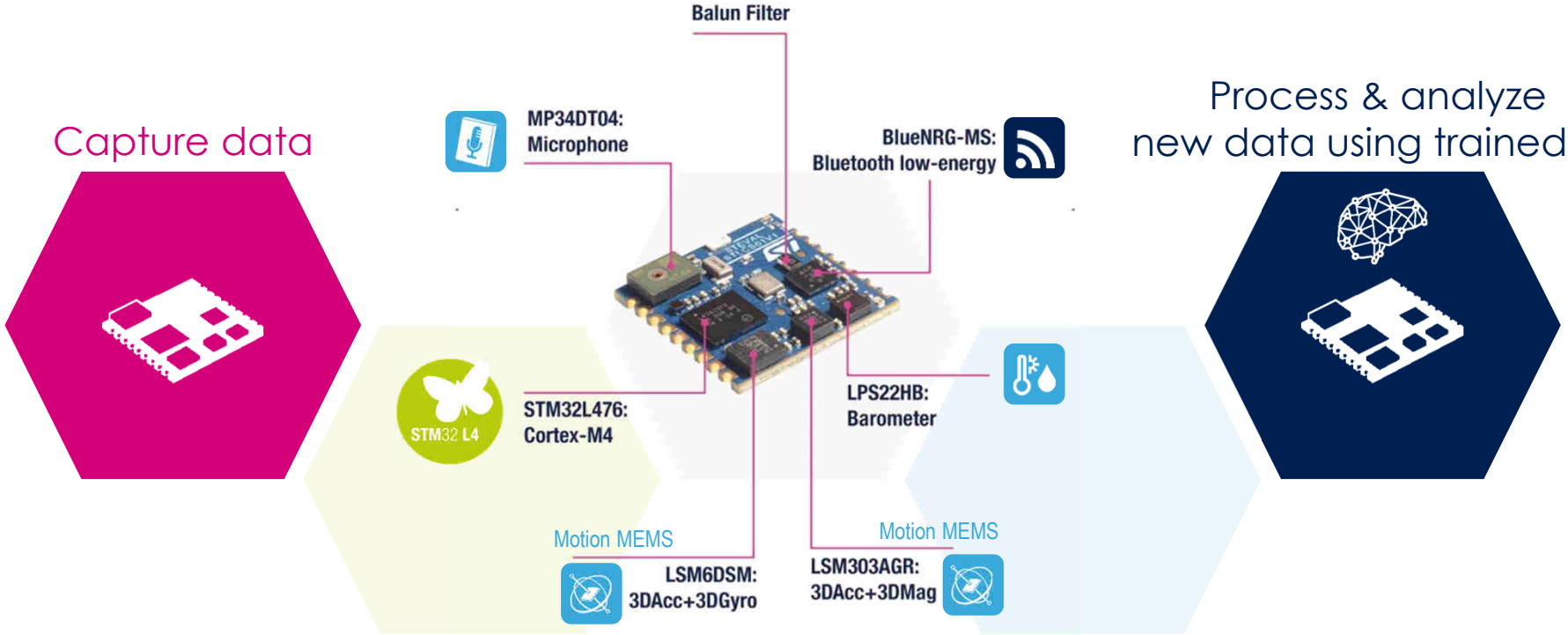
- Perform **analysis** to compute the model size, get an image of the network and the complexity
- Perform **validation on desktop**
- Perform **validation on target**
- Set a **compression** to reduce the model size (By reducing the accuracy of the model)



Form Factor Hardware
to Capture and Process Data

Capture data

Process & analyze new data using trained NN



life.augmented

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

Form Factor Hardware

AI IoT Node for More Connectivity

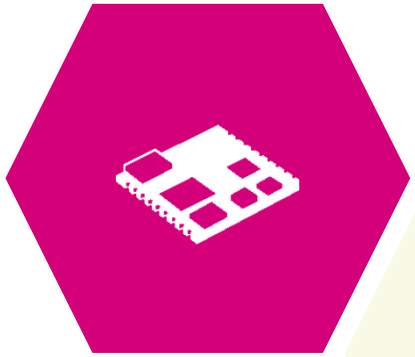
Sub-1GHz

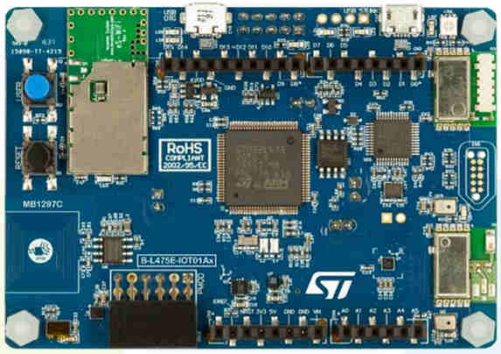
Dynamic NFC Tag

\+

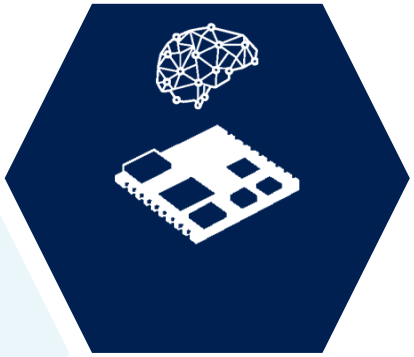
Wi-Fi

Capture data





Process & analyze new data using trained NN



More debug capabilities

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

life.augmented

https://www.st.com/iotnode

Collecting Data
& Architecting a NN Topology

Services provided by Partners

Capture data

Clean, label Data
Build NN topology

ST tools to support

ST BLE Sensor

ST BLE Sensor mobile phone application

Collect and label data from ST SensorTile.

Selected partners

Neural Networks engineering services support.

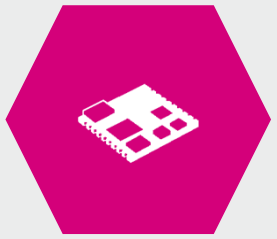
Data scientists and Neural network architects.

www.st.com/STM32CubeAI#Partners?

life.augmented

Human Activity Recognition (HAR)

Motion Example in FP-AI-SENSING1 Package



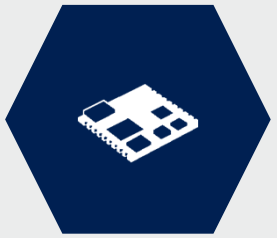
5 classes

Embedded **motion**

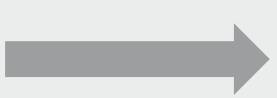
Labelling controlled by smartphone application

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

Stationary, walking, running, biking, driving



NN & example dataset provided



Embedded **motion** pre-processing

Inferences running on the microcontroller

Inference result displayed on mobile app

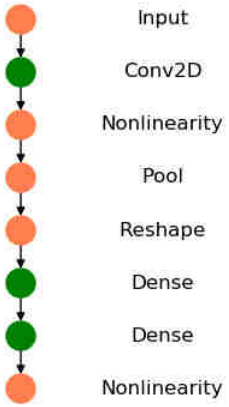
life.augmented

Human Activity Recognition
IGN (5 classes)

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

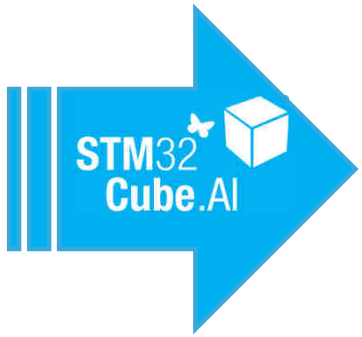
Neural Network

- 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

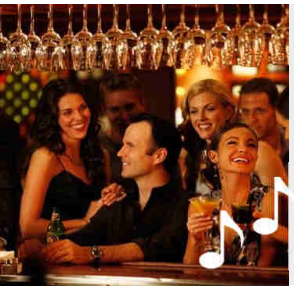
Performance on Sensor Tile

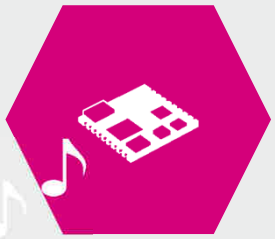
- STM32L476 80MHz Cortex-M4F
- Use case: 1 classification/sec
- Pre/Post-processing: 0.02 MHz NN processing: 0.35 MHz
- Power consumption (1.8 V)
  - System: 580 uA (with optim BLE)
  - STM32: 510 uA

life.augmented

Audio Scene Classification (ASC)

Audio Example in FP-AI-SENSING1 Package





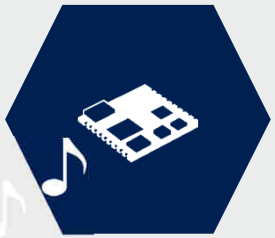
3 classes

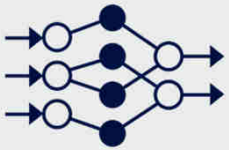
Embedded **audio**

Labelling controlled
by smartphone application

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

Indoor, Outdoor, In vehicle





NN & example
dataset provided

Embedded **audio**
pre-processing

Inferences running
on the microcontroller

Inference result
displayed on mobile app

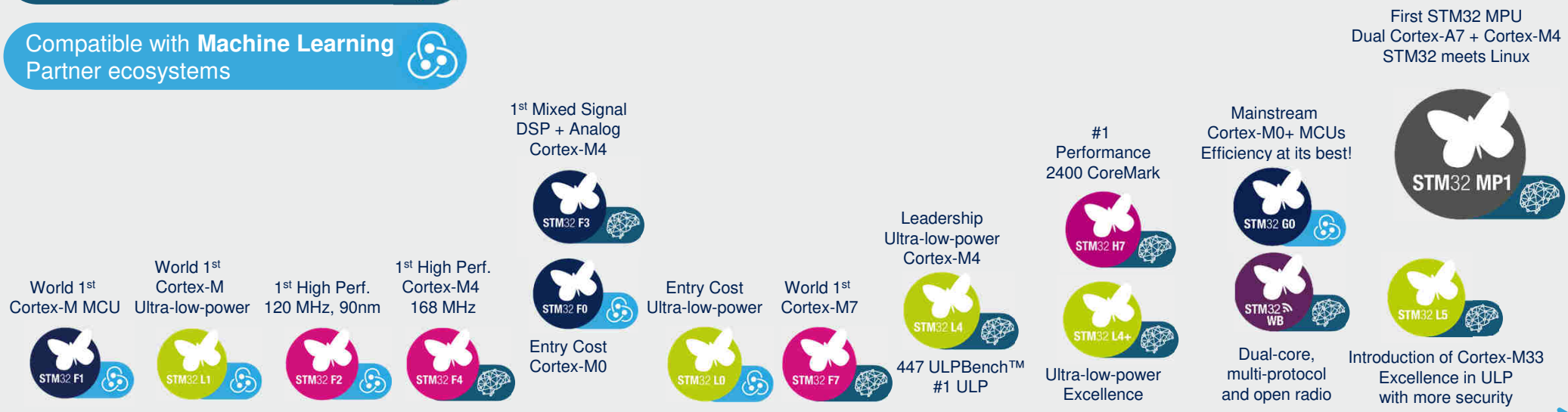
life.augmented

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



life.augmented

More than 60,000 customers

Over 4 Billion STM32 shipped since 2007

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, Lasagne, Caffe, ConvNetJS Deep Learning in your browser

+ 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

STM32 Cube.AI

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

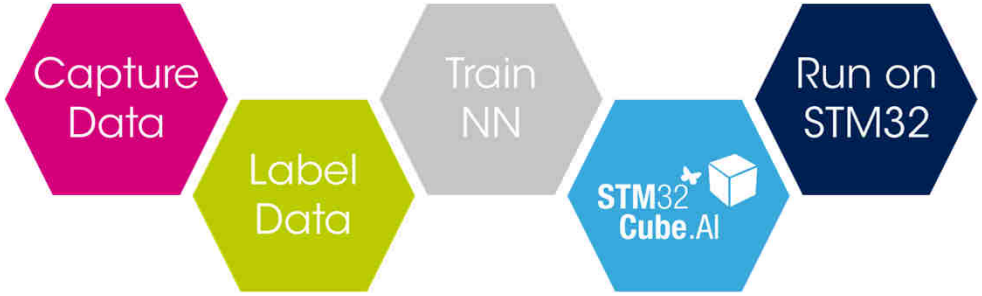
ST Partner Program with a dedicated group of Partners providing Neural Networks engineering services

Data scientists and Neural network architects

life.augmented

For More Information

www.st.com/STM32CubeAI



life.augmented