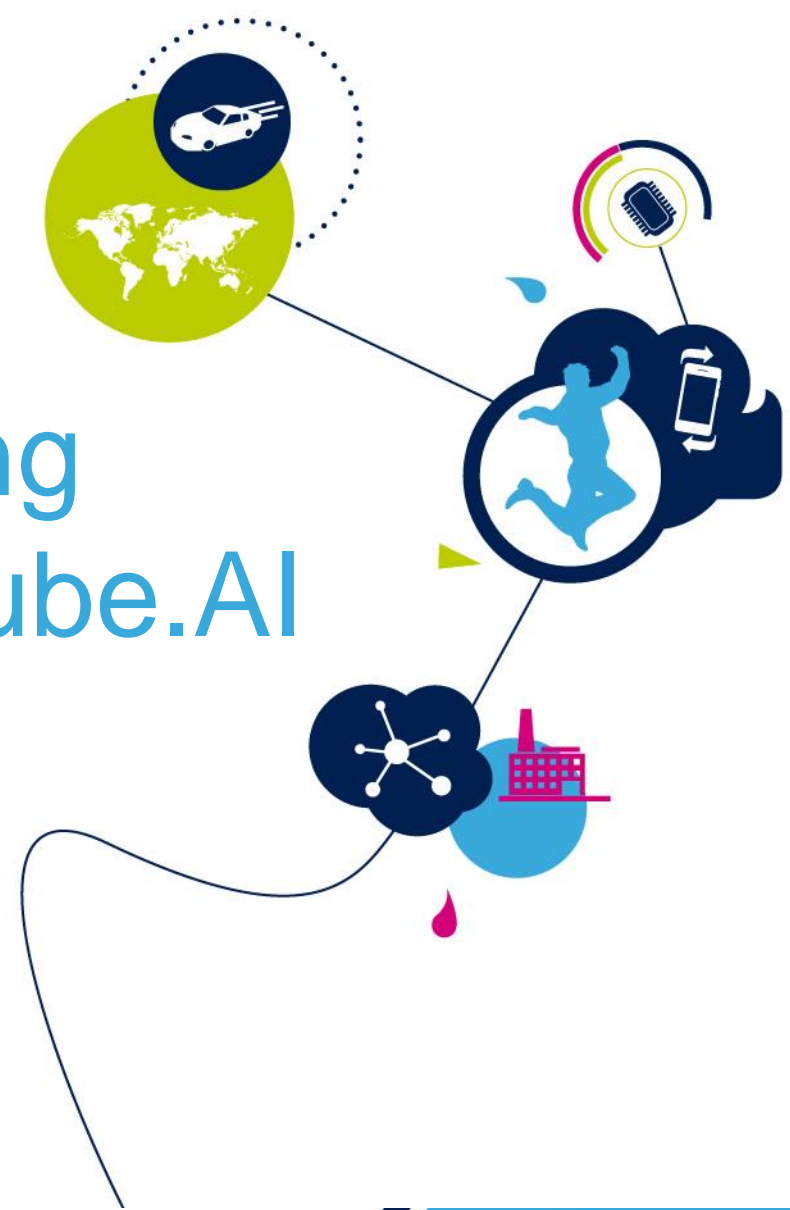




Artificial Neural Network Mapping Made Simple with the STM32Cube.AI

Markus Mayr
Product Marketing Manager, MCU



**ST Developers
Conference**

September 12th, 2019
Santa Clara Convention Center - Mission City Ballroom
Santa Clara, CA



Artificial Intelligence (AI)

2

- 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



Artificial Intelligence (AI)

3

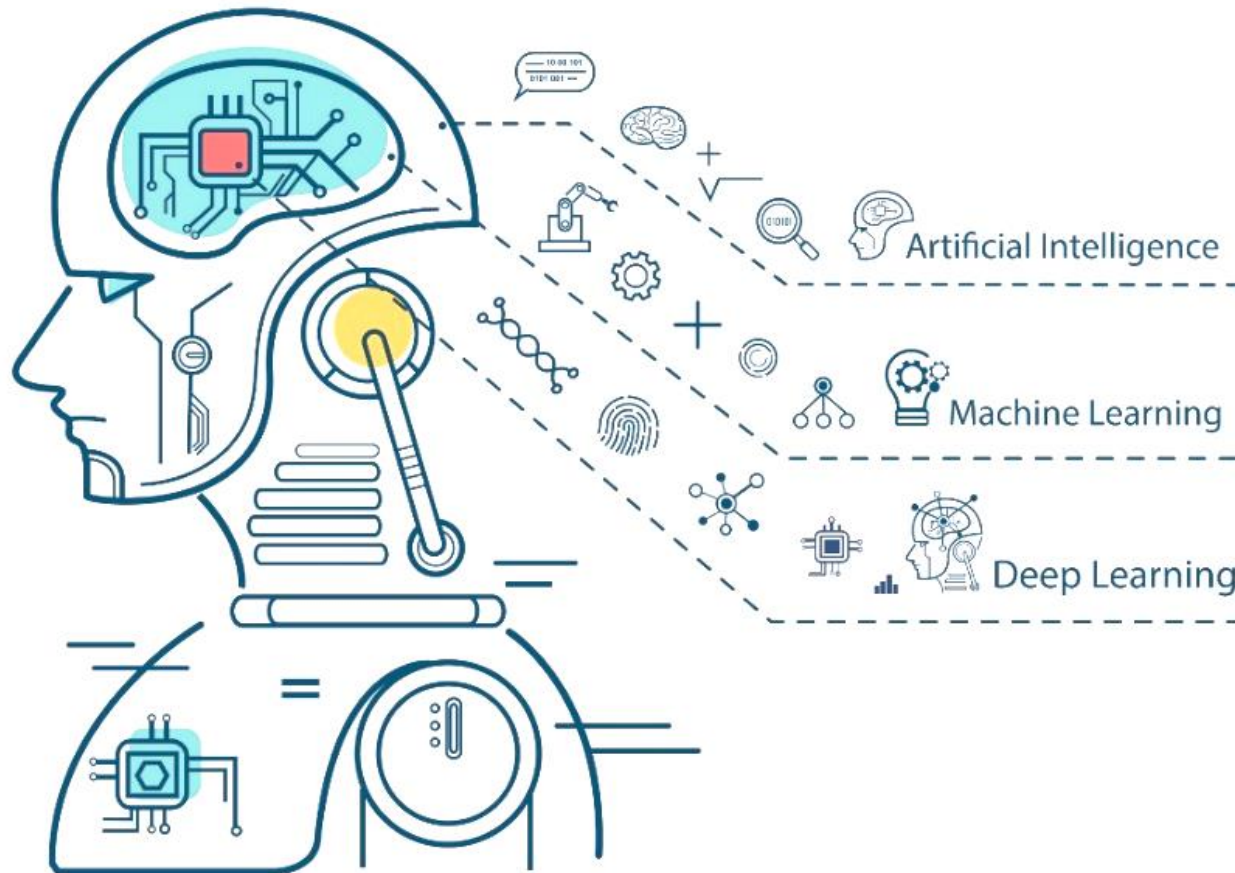
- **Main use cases in our everyday life:**

- 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



Some definitions

4



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

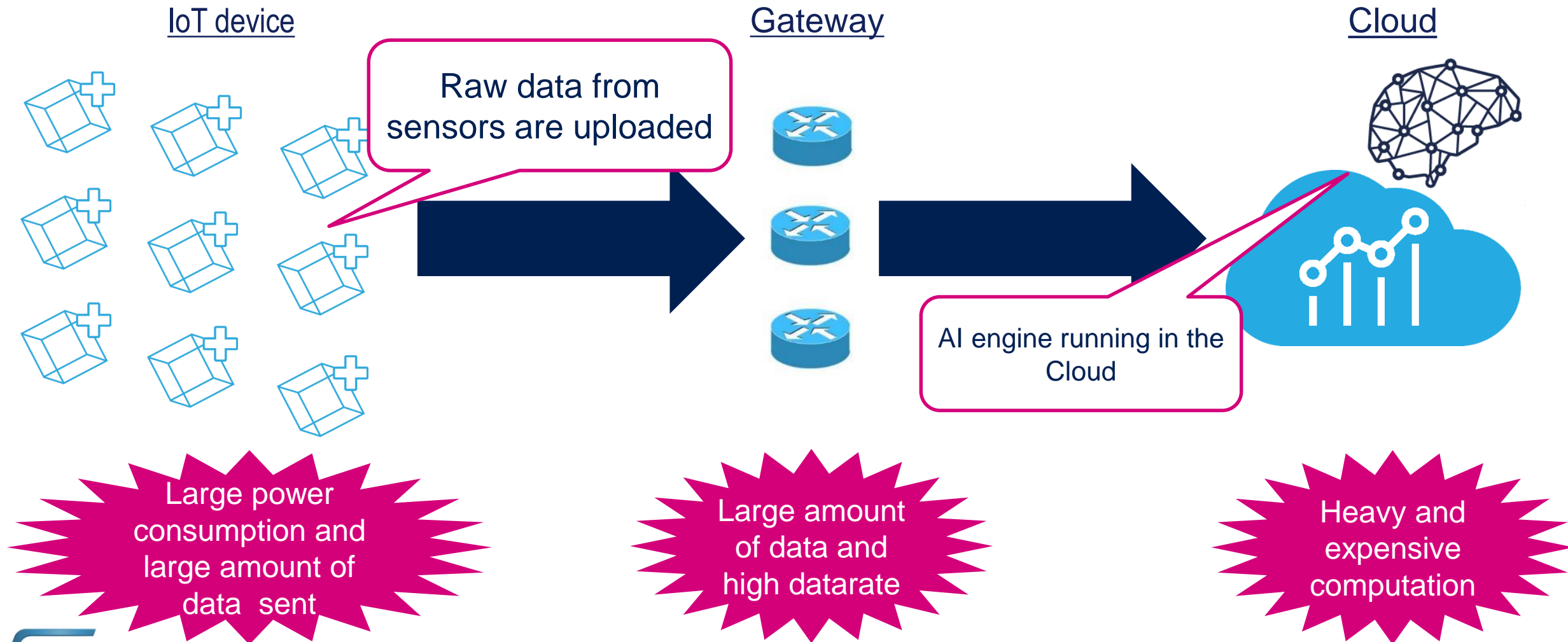
5

- 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%

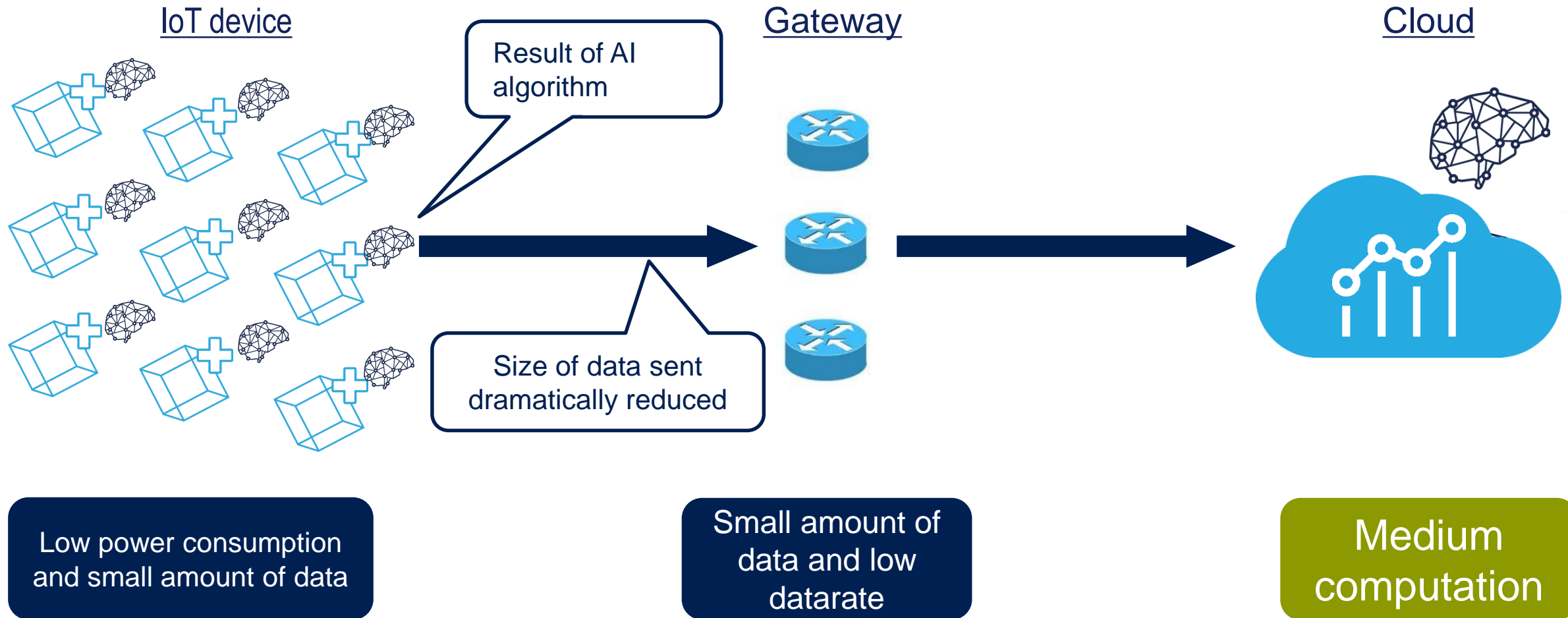
AI Cloud Computing

6



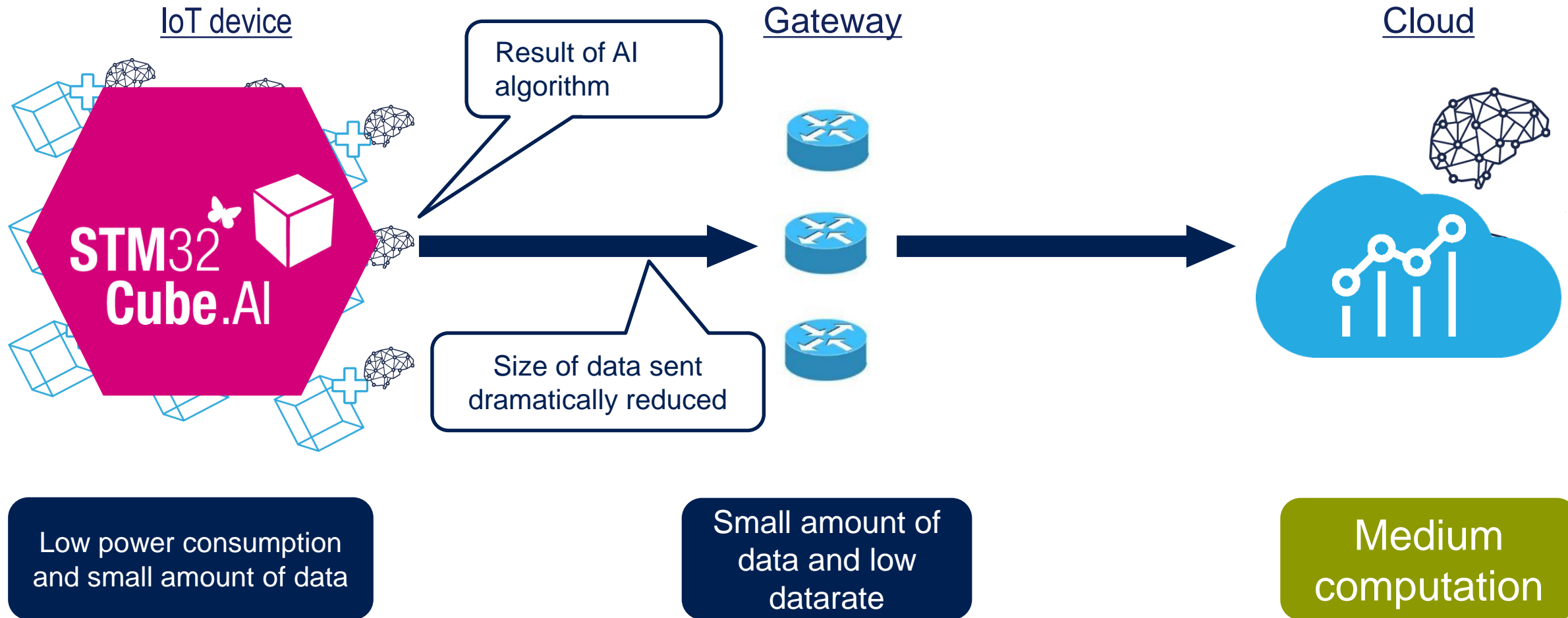
AI Edge Computing (Embedded)

7



AI Edge Computing (Embedded)

8

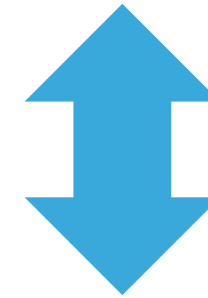


Distributed AI

9



High Bandwidth
High centralized computing power
Potentially high latency



Reduced bandwidth
Lower centralized computing power
Real-time response
Preserved Privacy

Neural Networks on STM32

Simple, fast, optimized



STM32[🦋] 
Cube.AI



The Key Steps Behind Neural Networks

11



Neural Network (NN) Model Creation



Operating Mode

Capture data



1

2



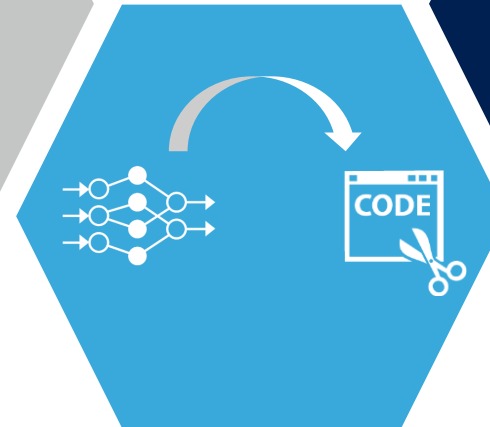
Clean, label Data
Build NN topology

Train NN Model



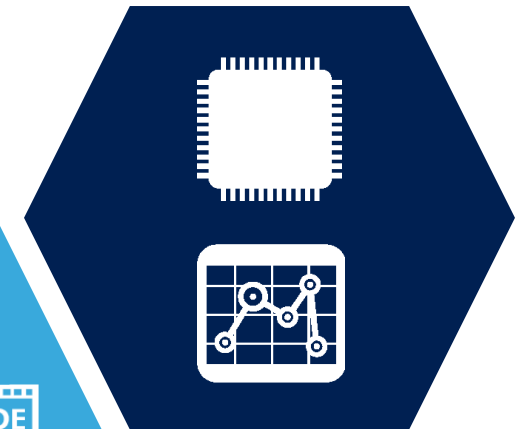
3

4



Convert NN into
optimized code for MCU

Process & analyze
new data using trained NN

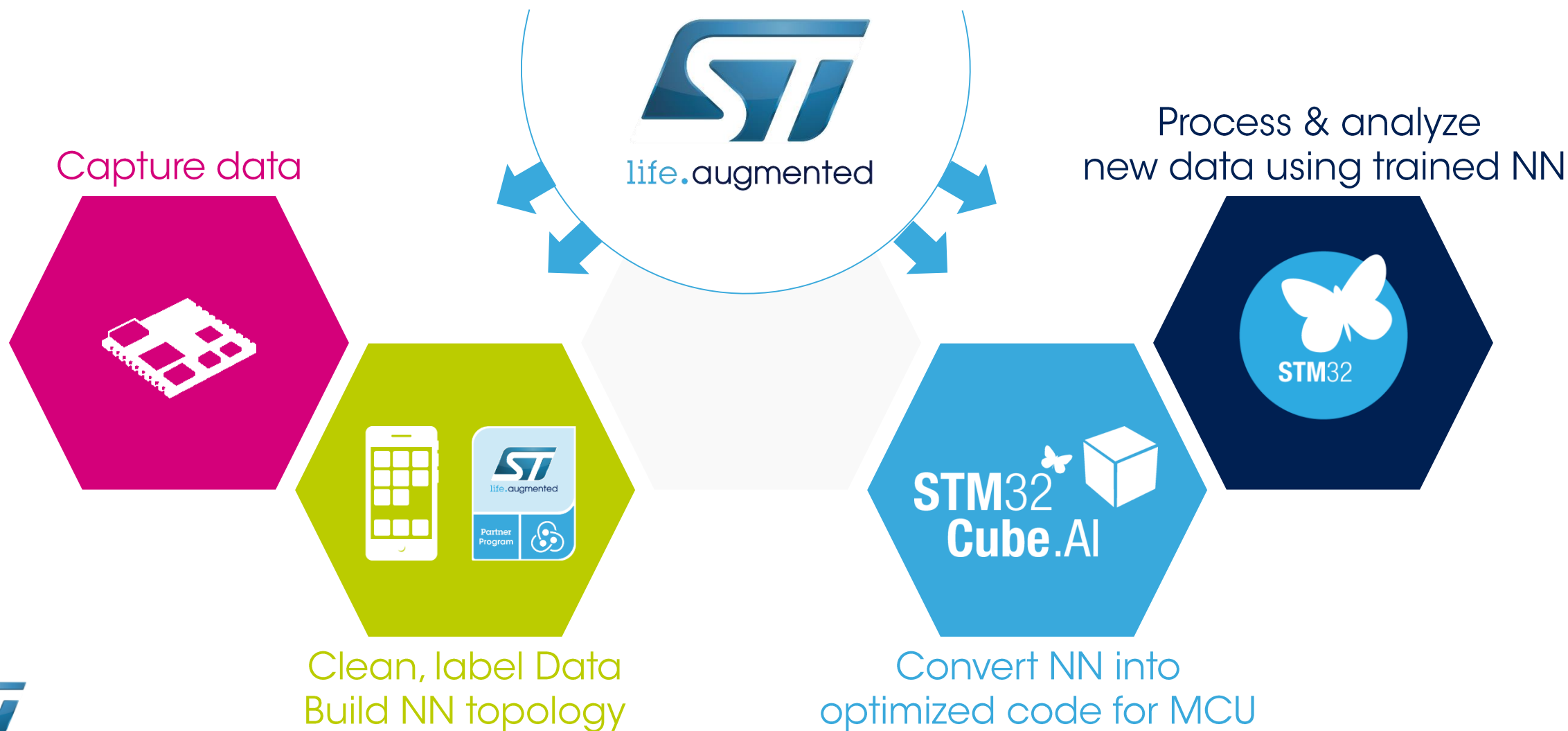


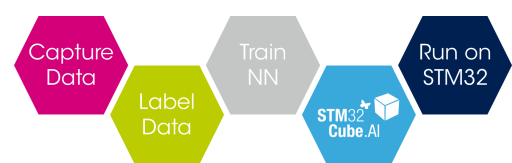
5



ST Toolbox for Neural Networks

12





STM32 Solutions for Embedded AI

Extensive toolbox to easily create your AI application

13

Neural Networks on STM32
Simple, fast, optimized



STM32Cube.AI



STM32Cube.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**



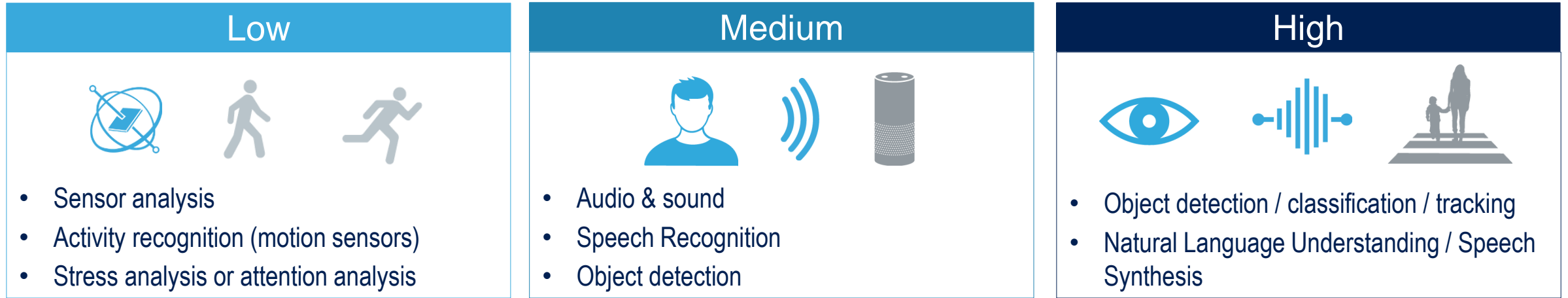
STM32 **Community** with dedicated
Neural Network topic



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

STM32 AI Typical Applications

14



10s MOPs

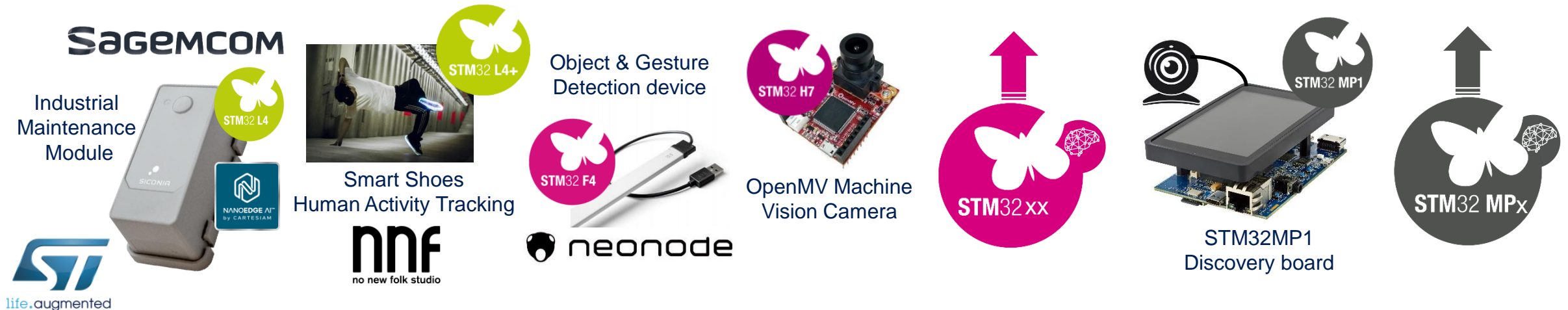
GOPs

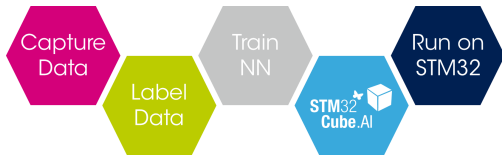
0.5-1 TOPs

1-2 TOPs

MCU

From IP embedded in MCU/MPU to dedicated SOC





STM32CubeMX Extension

AI Conversion Tool

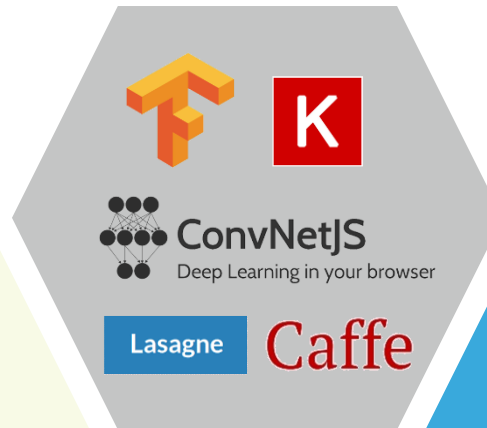
15

Input your framework-dependent, pre-trained Neural Network into the **STM32Cube.AI** conversion tool

Automatic, fast generation of an STM32-optimized library

STM32Cube.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

Process & analyze new data using trained NN

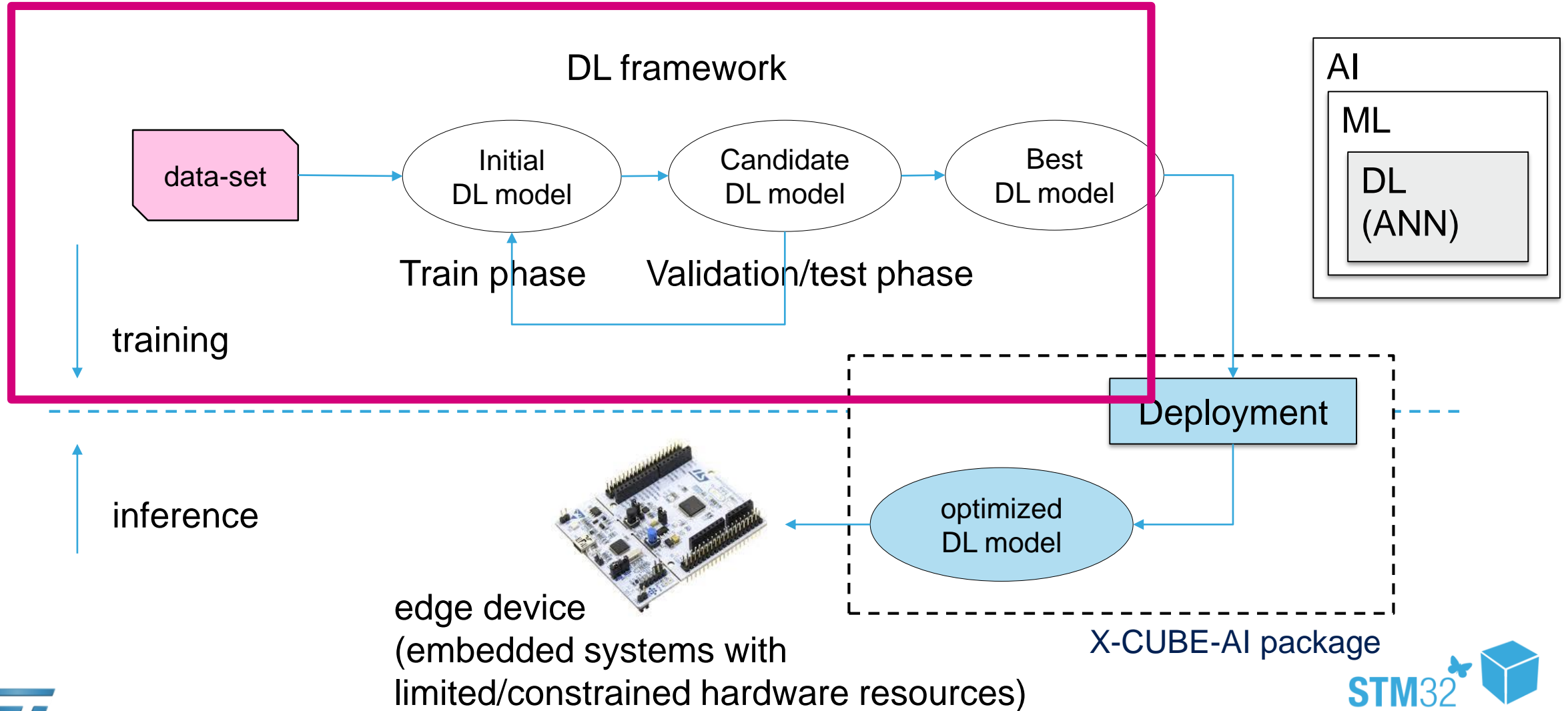


Convert NN into optimized code for MCU

X-CUBE-AI Positioning

in a typical DL flow

16



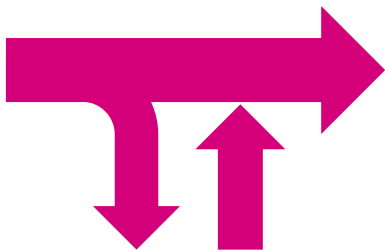
Off-the-shelf :
Pre-trained Artificial
Neural Network Model

Deep Learning
Framework dependent



**Neural
Network
Importer**

Framework
Independent
Artificial Neural
Network
Representation



**Code
Generator**



Embedded Solution
Optimized Artificial
Neural Network Code
generated for STM32

Artificial
Neural
Networks
API's

**NN Layers
Library
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

MHz and embedding a floating point unit (FPU). The family incorporates high-speed embedded memories (up to 64 Kbyte of Flash

Graphic Summary AI Summary



Minimum Ram: 196 Bytes
Minimum Flash: 15.20 KBytes

C:\Users\ledonger\Documents\deepnet_relu.h5

MCUs List: 627 items

Display similar items

*	Part No	Refere...	Marketing ...	Unit Price for 1...	Board	Package	Flash	RAM	IO	Freq.	GFX S...	HMAC	MD5	SH
☆	STM32F301C6	STM3...	Active	1.596		LQFP48	32 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F301C8	STM3...	Active	1.666		LQFP48	64 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F301C8	STM3...	Active	1.666		WLCSP49	64 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F301K6	STM3...	Active	1.272		LQFP32	32 kByt...	16 kBytes	25	72 MHz	0.0	0	0	0
☆	STM32F301K6	STM3...	Active	1.272		UFQFPN32	32 kByt...	16 kBytes	24	72 MHz	0.0	0	0	0
☆	STM32F301K8	STM3...	Active	1.342		LQFP32	64 kByt...	16 kBytes	25	72 MHz	0.0	0	0	0
☆	STM32F301K8	STM3...	Active	1.342		UFQFPN32	64 kByt...	16 kBytes	24	72 MHz	0.0	0	0	0
☆	STM32F301R6	STM3...	Active	1.758		LQFP64	32 kByt...	16 kBytes	51	72 MHz	0.0	0	0	0
☆	STM32F301R8	STM3...	Active	1.828		LQFP64	64 kByt...	16 kBytes	51	72 MHz	0.0	0	0	0
☆	STM32F302C6	STM3...	Active	1.712		LQFP48	32 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F302C8	STM3...	Active	1.782		LQFP48	64 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F302C8	STM3...	Active	1.782		WLCSP49	64 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F302CB	STM3...	Active	1.99		LQFP48	128 kBy...	32 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F302CC	STM3...	Active	2.288		LQFP48	256 kBy...	40 kBytes	37	72 MHz	0.0	0	0	0

☐ Enable

Artificial Intelligence

☒ Enable

Model

Keras

Type

Saved model

Model

deepnet_relu.h5

Browse

Compression

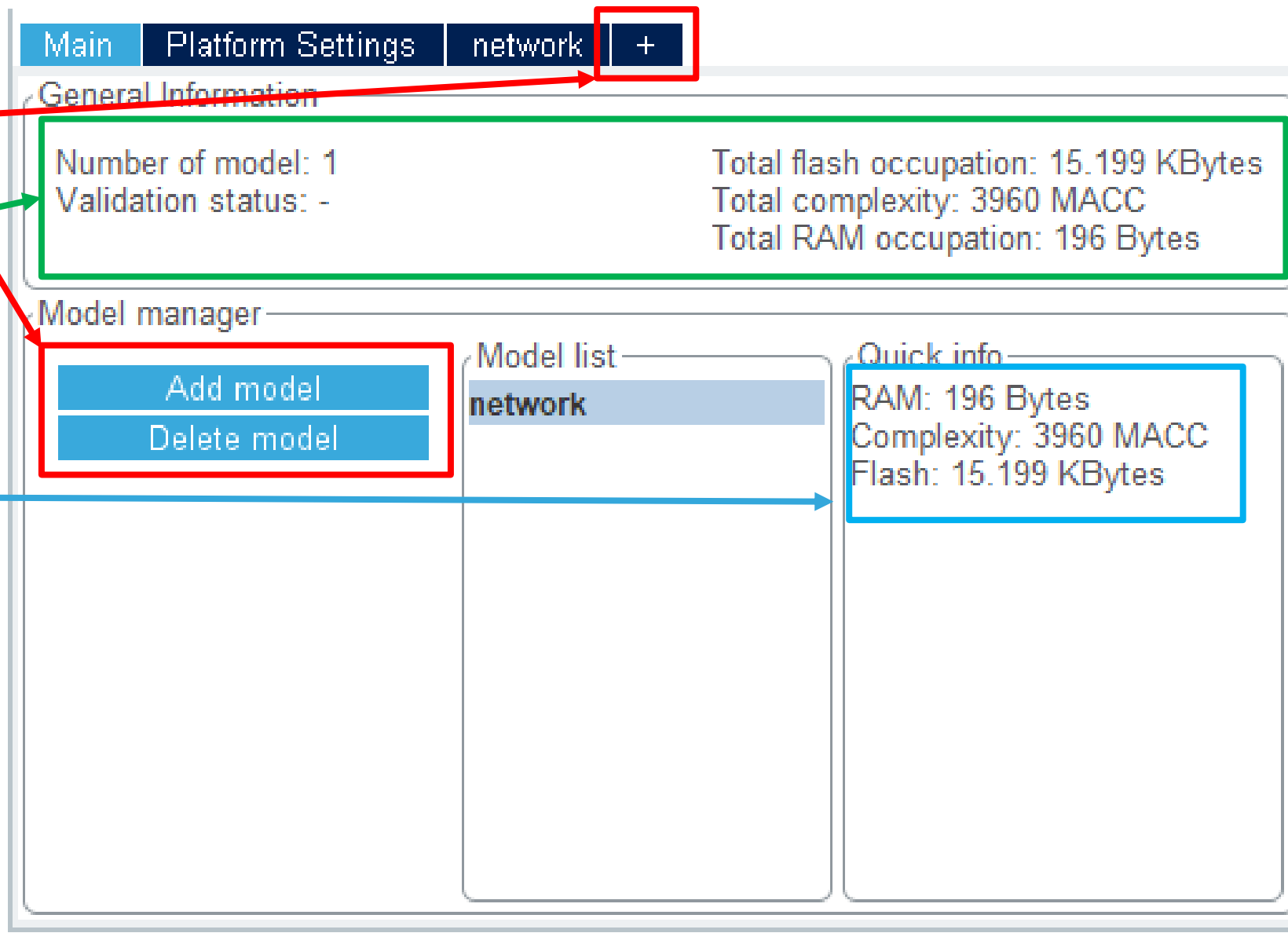
None

Analyze

Peripheral

✓ ADC 12-bit	0	40
✓ ADC 16-bit	0	36
✓ AES	<input type="checkbox"/>	
✓ CAN	0	3
✓ COMP	0	7
✓ CRYP	<input type="checkbox"/>	
✓ DAC 12-bit	0	3

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



The screenshot shows the X-Cube-AI Main tab interface. At the top, there are tabs: 'Main', 'Platform Settings', 'network', and a '+' button. The 'Main' tab is selected. Below the tabs, there is a 'General Information' section with a green border. It contains the following text:

- Number of model: 1
- Validation status: -
- Total flash occupation: 15.199 KBytes
- Total complexity: 3960 MACC
- Total RAM occupation: 196 Bytes

Below the 'General Information' section is a 'Model manager' section. It contains a red-bordered box with two buttons: 'Add model' and 'Delete model'. To the right of the 'Model manager' section is a 'Model list' section with a blue header 'network'. To the right of the 'Model list' section is a 'Quick info' section with a blue border. It contains the following text:

- RAM: 196 Bytes
- Complexity: 3960 MACC
- Flash: 15.199 KBytes

Annotations include:

- A red arrow pointing from the 'Add/Delete models' list item to the 'Add model' and 'Delete model' buttons.
- A green arrow pointing from the 'Get general information' list item to the 'General Information' section.
- A blue arrow pointing from the 'Have a quick look at different models' list item to the 'Quick info' section.
- A red box around the '+' button in the top navigation bar.
- A red box around the 'Add model' and 'Delete model' buttons.
- A blue box around the 'Quick info' section.

- 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)

Main
Platform Settings
network
+

Model inputs
network

Keras
Saved model

Model: C:\Users\ledonger\Documents\deepnet_relu.h5
Browse...

Browse...

Command

Validation status: Unknown
Complexity: -
Flash occupation: -
RAM: -
Actual compression: -

Compression: None

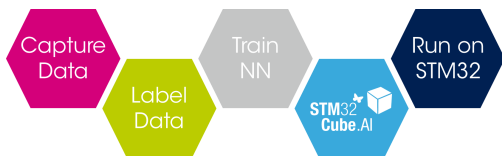
Show graph

Analyze

Validation from: Random numbers

Validate on desktop

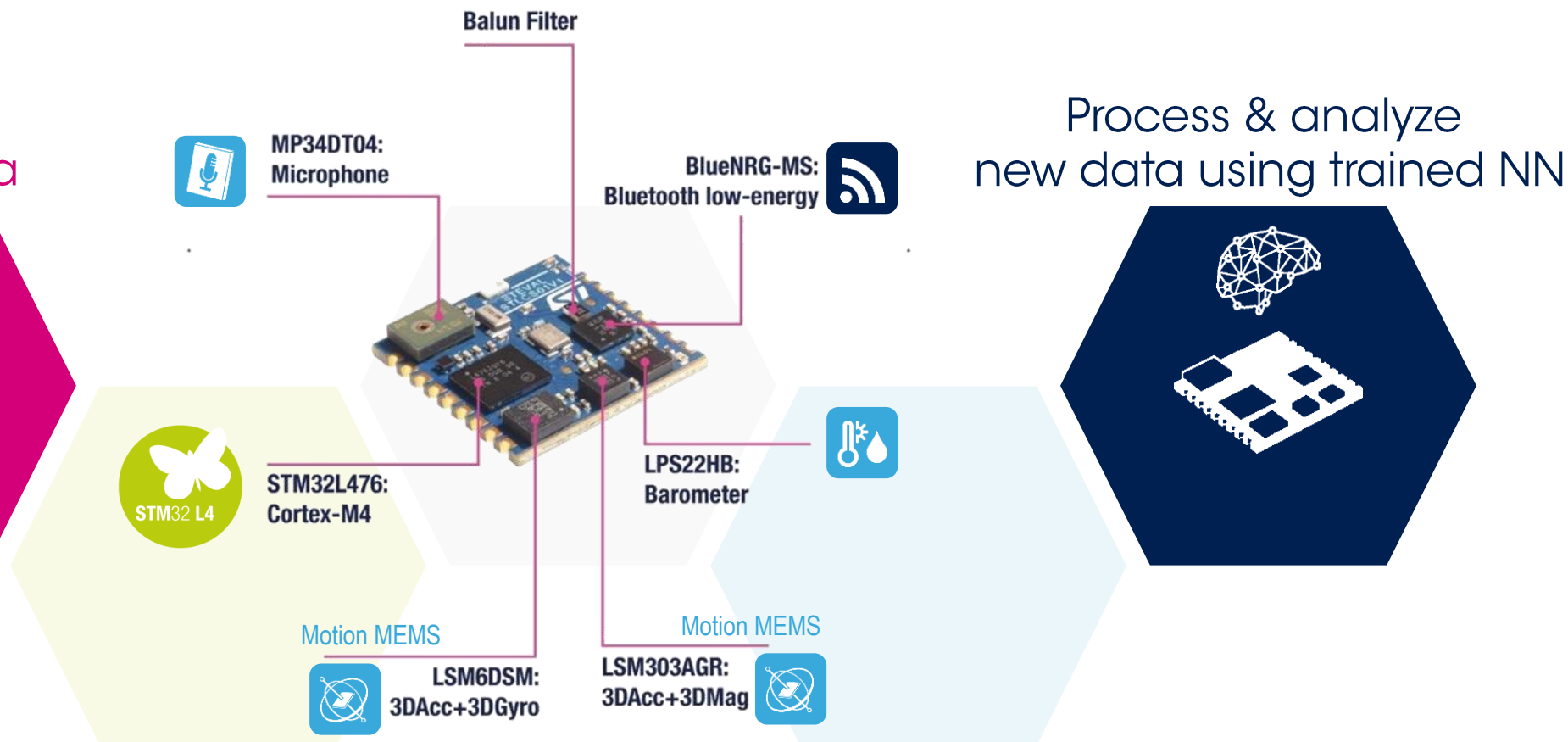
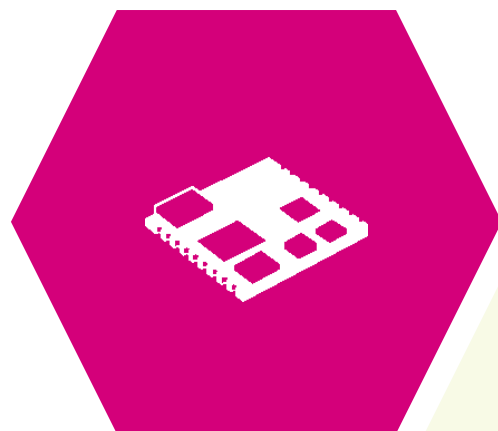
Validate on target

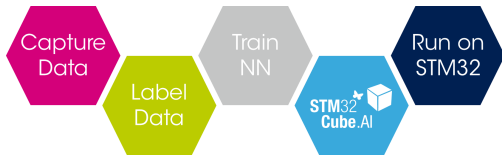


Form Factor Hardware to Capture and Process Data

21

Capture data



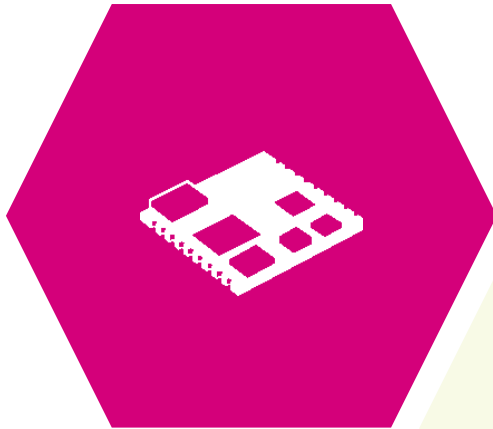


Form Factor Hardware

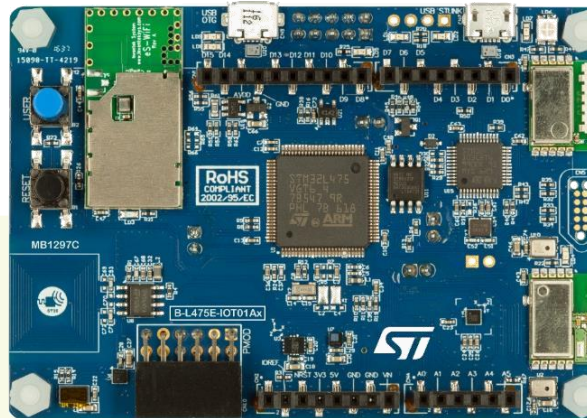
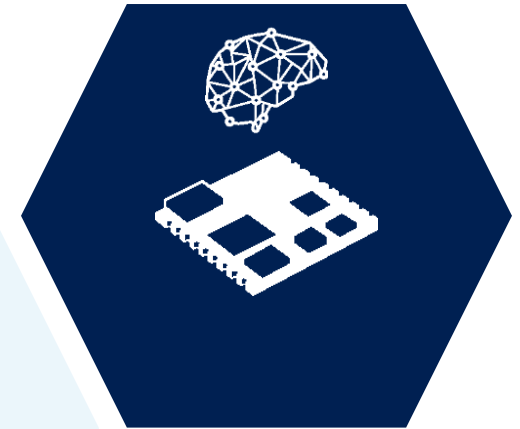
AI IoT Node for More Connectivity

22

Capture data

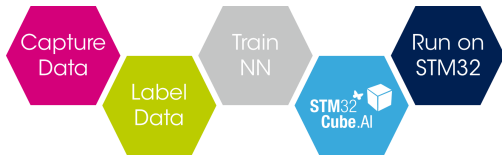


Process & analyze
new data using trained NN



More debug capabilities

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



Collecting Data & Architecting a NN Topology

23

Services provided by Partners

ST tools to support

Capture data



Clean, label Data
Build NN topology



ST BLE Sensor mobile phone application

Collect and label data from ST SensorTile.



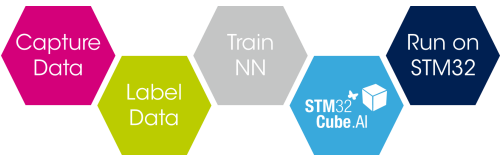
ST BLE
Sensor



Selected partners

Neural Networks engineering services support.
Data scientists and Neural network architects.

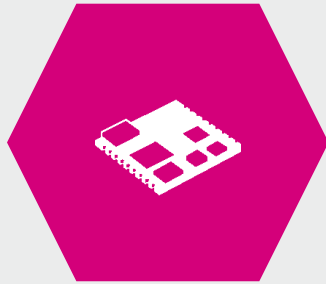
www.st.com/STM32CubeAI#Partners?



Human Activity Recognition (HAR)

Motion Example in FP-AI-SENSING1 Package

24



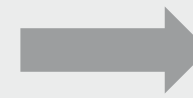
Embedded **motion**



Labelling controlled
by smartphone application

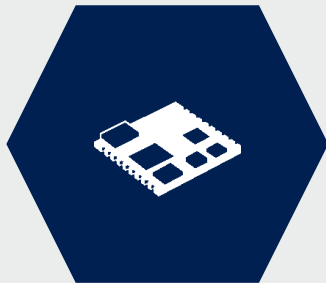


Data stored on the device
SD card for future **learning**



5 classes

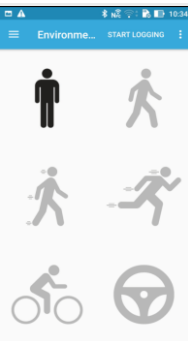
Stationary, walking, running,
biking, driving



Embedded **motion**
pre-processing



NN & example
dataset provided



Inference result
displayed on mobile app



Human Activity Recognition

IGN (5 classes)

25

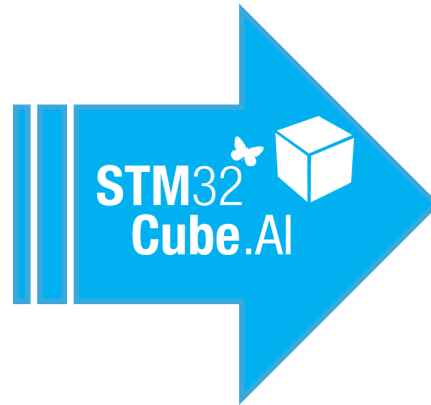
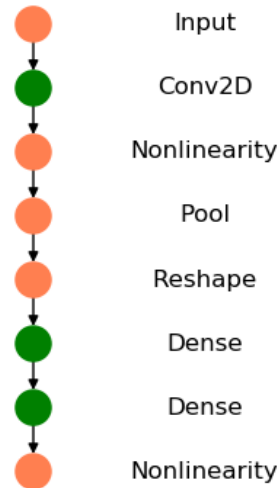
Use Case #2: HAR Human Activity Recognition Ignatov on SensorTile

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

Making AI Accessible Now

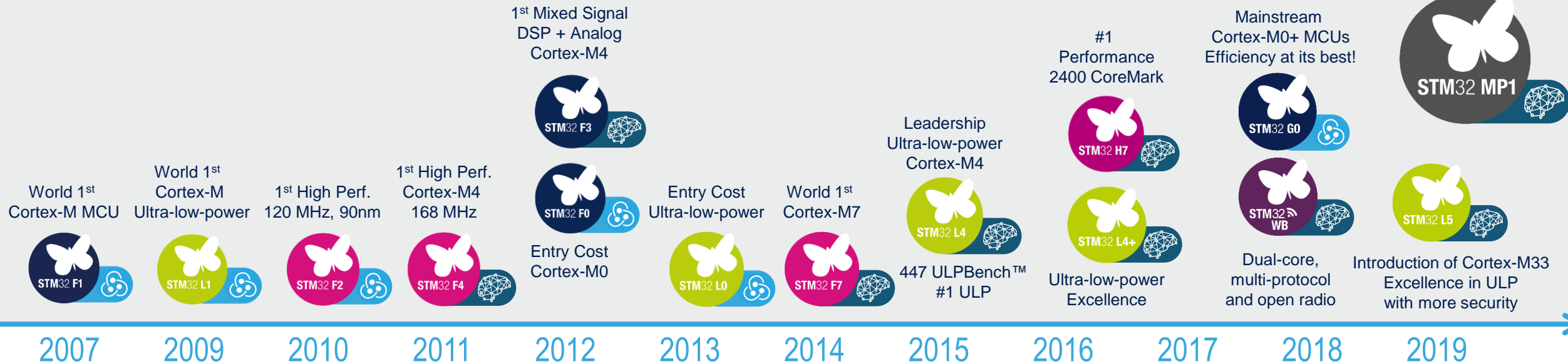
27

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

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

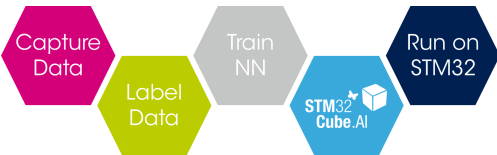


Compatible with **Machine Learning**
Partner ecosystems



More than 60,000 customers

Over 4 Billion STM32 shipped since 2007



STM32Cube.AI Roadmap

28

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



Lasagne

Caffe

Deep Learning in your browser

+ TensorFlow Lite

- Keras Fixed Point
- TensorFlow Lite Float

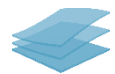
+ TensorFlow Lite

- TensorFlow Lite Integer arithmetic quantization



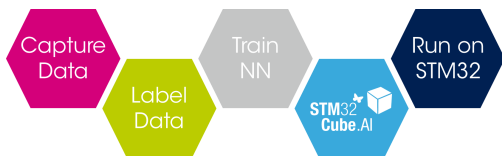
+

ONNX { mxnet, Chainer, PyTorch, Microsoft CNTK }



New layers & operators addition





STM32 Solutions for AI

More Than Just the STM32Cube.AI

29

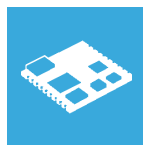
An extensive toolbox to support easy creation of your AI application

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

30



www.st.com/STM32CubeAI



Capture
Data

Label
Data

Train
NN

STM32
Cube.AI

Run on
STM32