

# Artificial Neural Network Mapping Made Simple with the STM32Cube.AI

Markus Mayr  
Product Marketing Manager, MCU



Technology Tour 2019

Toronto, Canada | May 29



# Artificial Intelligence (AI)

2

- AI is a superset of all the studies where machines mimic cognitive capabilities like humans. 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

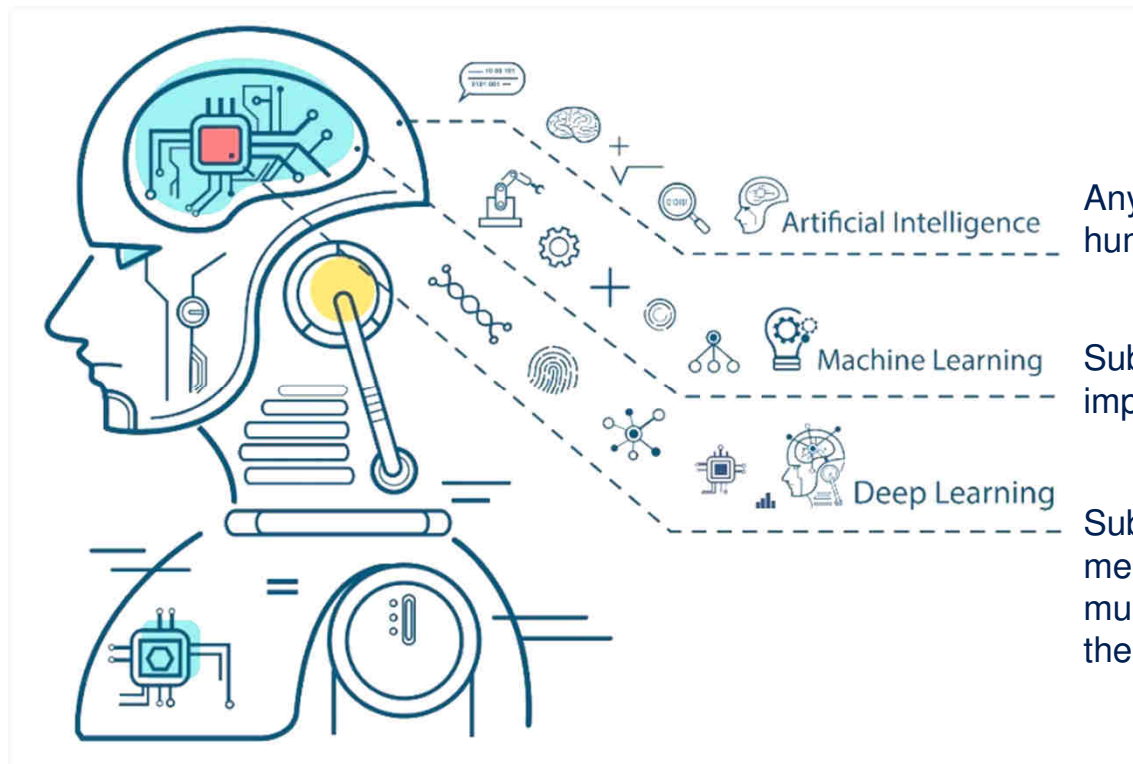
- **Main use cases in our everyday life:**

- 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 technic which enables computer to mimic human behavior

Subset of AI, algorithms and methodologies to 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

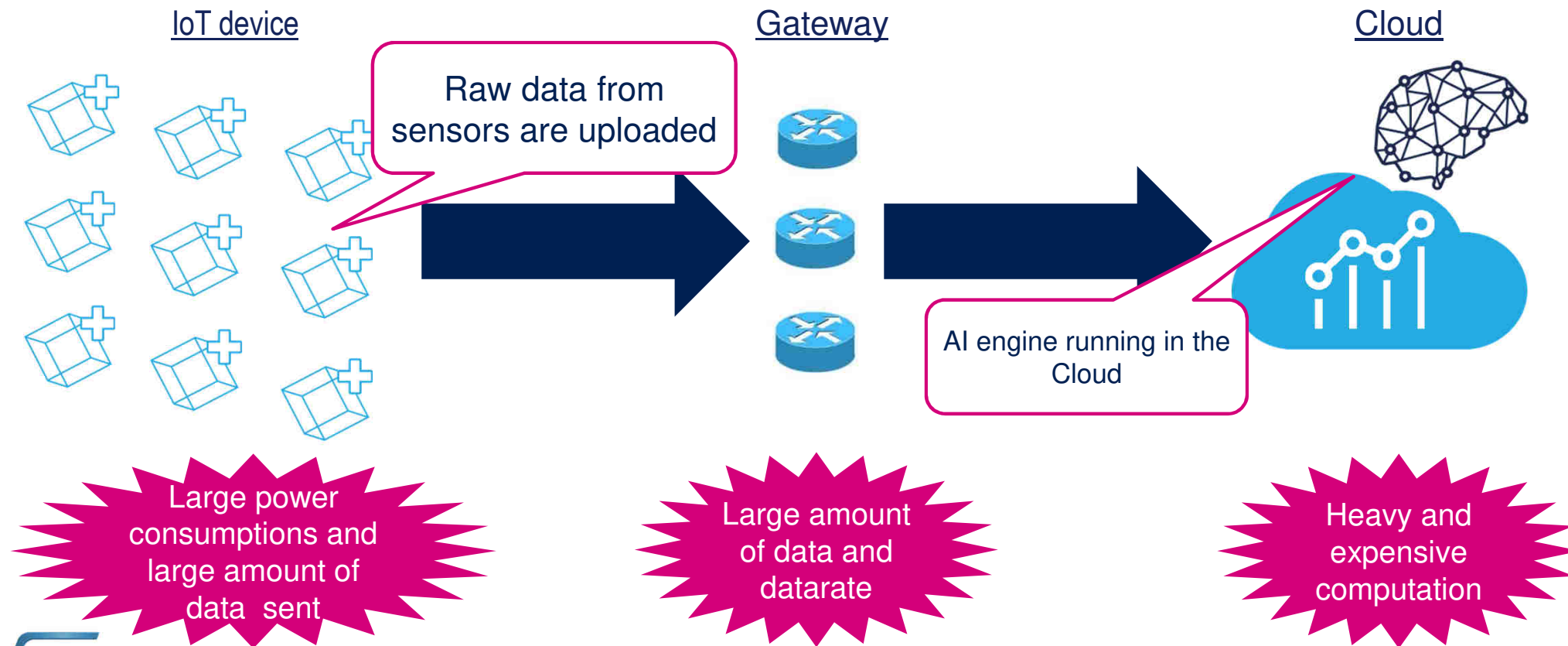
- Convolutional Deep Neural Networks outperform previous methods on a number of tasks:

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%

May 2015

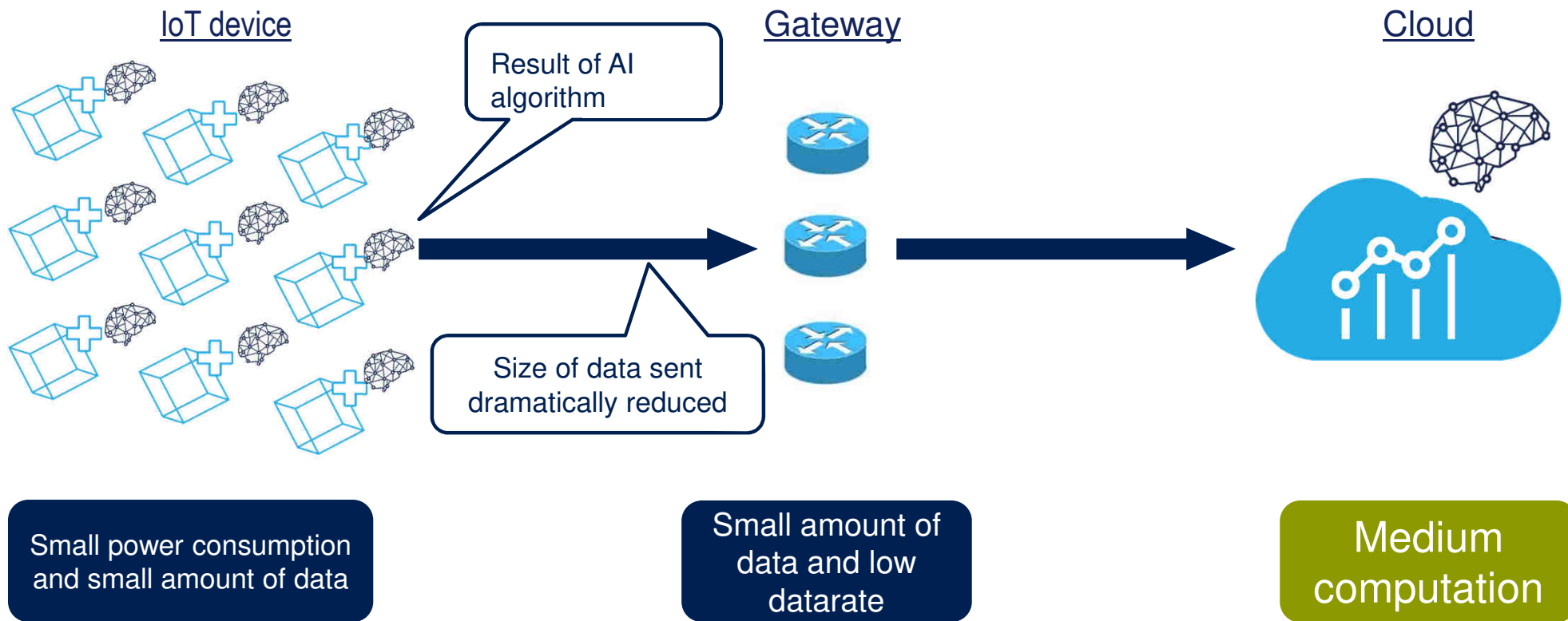
# AI Cloud computing

6



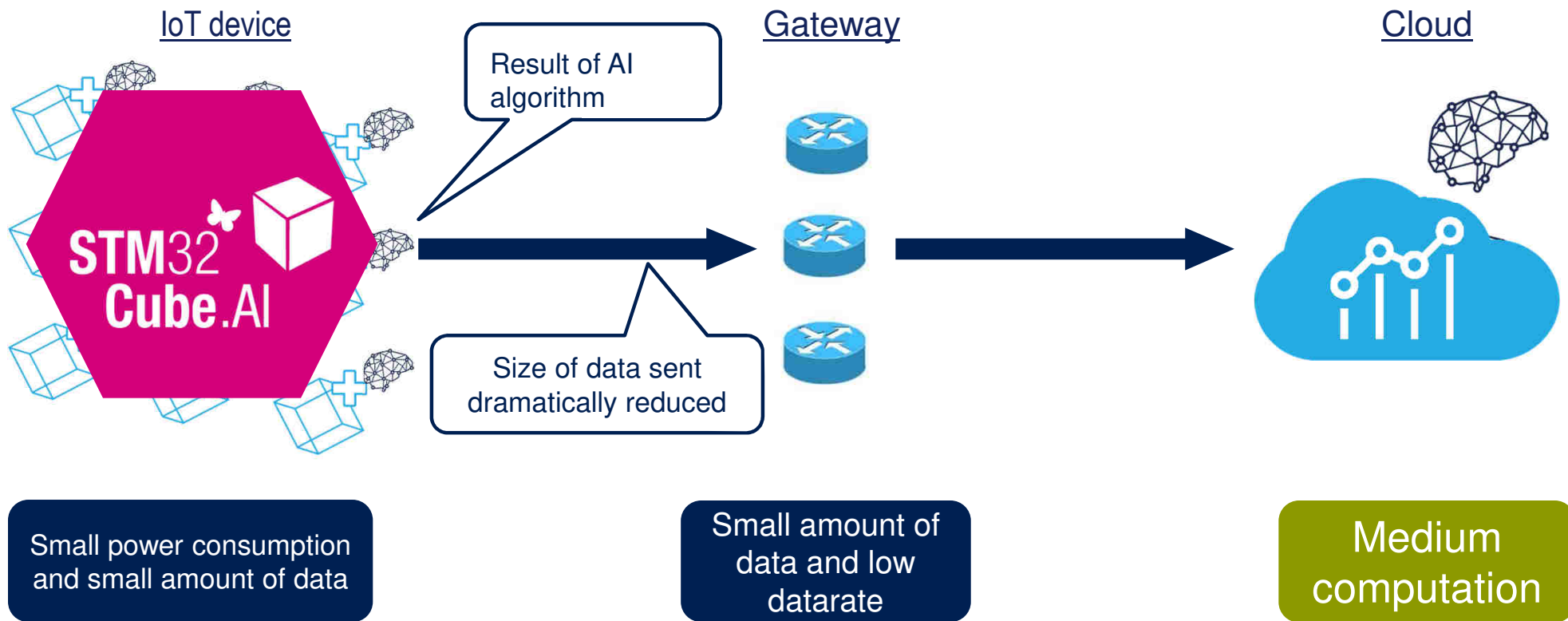
# AI Edge computing (Embedded)

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  
Preserving Privacy



# Artificial Intelligence and STM32

## Application trend

10

### Sensors



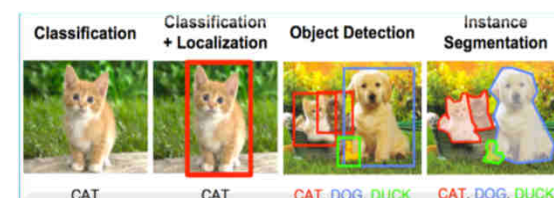
- Activity Recognition with Inertials (DCNN, ESN, LSTM)
- Stress Analysis or Attention Analysis (DCNN, SON), etc

### Audio Processing



- Speech Recognition (DeepSpeech, Wave2Letter)
- Speech Synthesis (WaveNet, Tacotron)

### Video Analysis



- Classification (Alexnet, Inception, VGG)
- Detection (Yolo, SSD)

STM32



Dedicated AI hardware needed



- Audio use cases with individual commands
- Classic motion sensor use cases

- Video analysis cannot be done in timely manner with MCU
- Advanced Audio use cases with Natural language understanding not yet accessible for MCUs

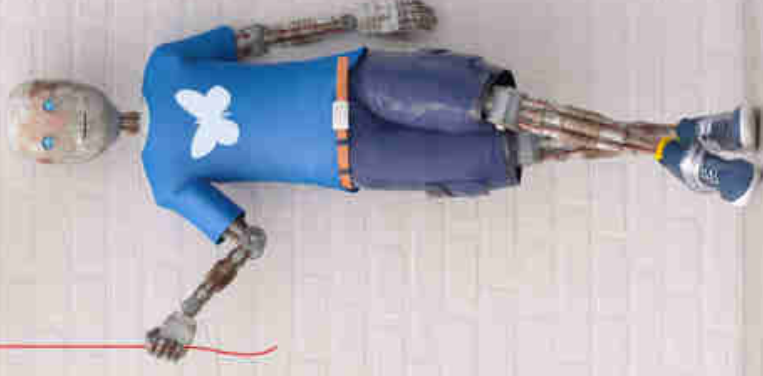


# Neural Networks on STM32

## Simple, fast, optimized



 **STM32**  
**Cube.AI** 



# The Key Steps Behind Neural Networks

12



Neural Network (NN) Model Creation



Operating Mode

Capture data



1

2



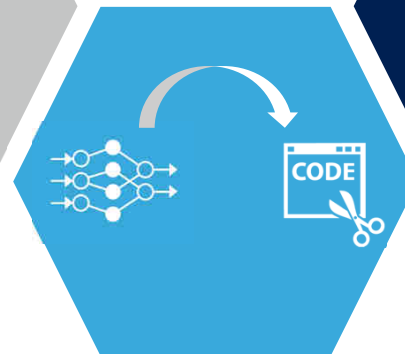
Clean, label Data  
Build NN topology

Train NN Model



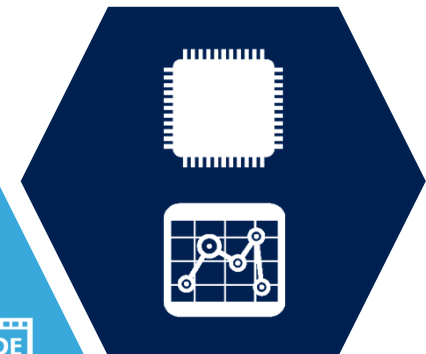
3

4



Convert NN into  
optimized code for MCU

Process & analyze  
new data using trained NN

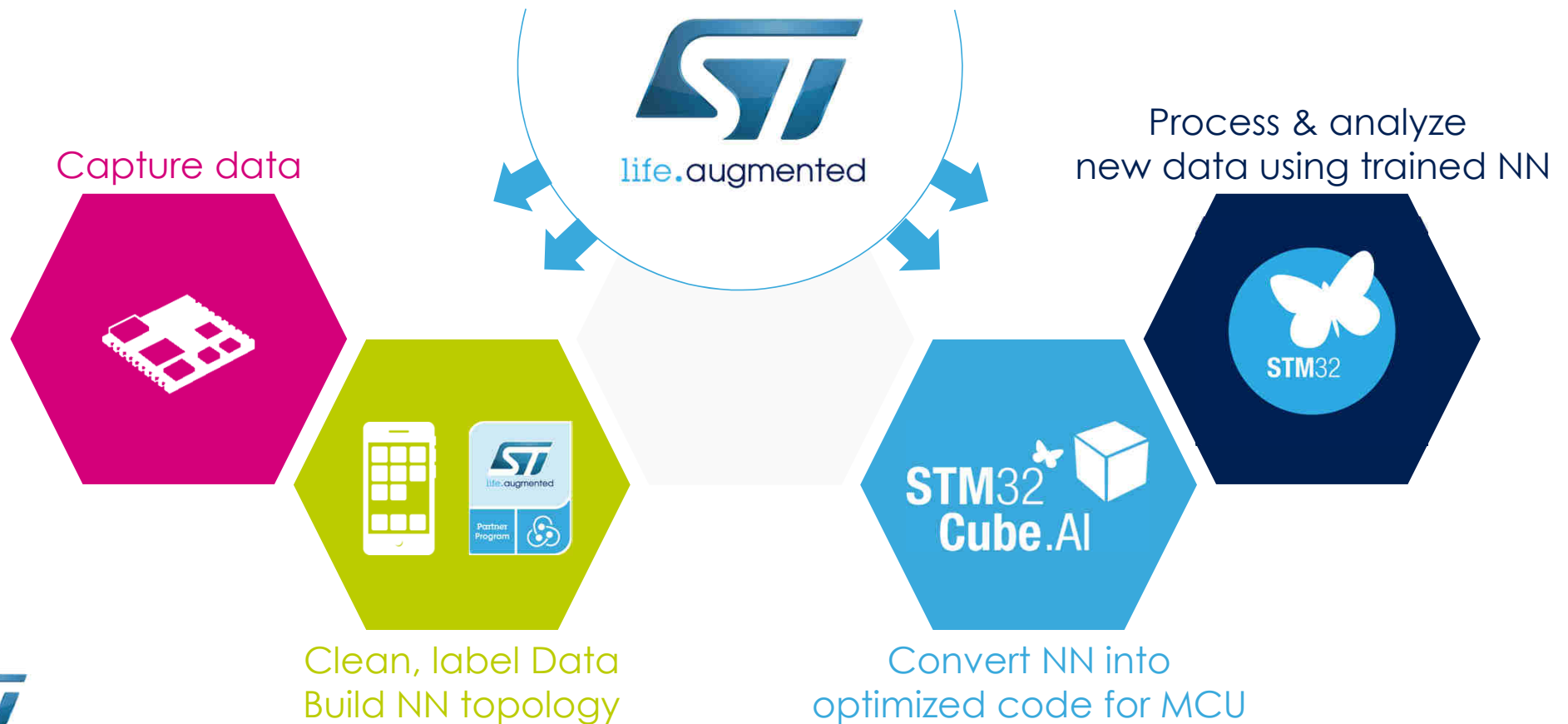


5



# ST Toolbox for Neural Networks

13





# STM32CubeMX Extension

## AI Conversion Tool

14

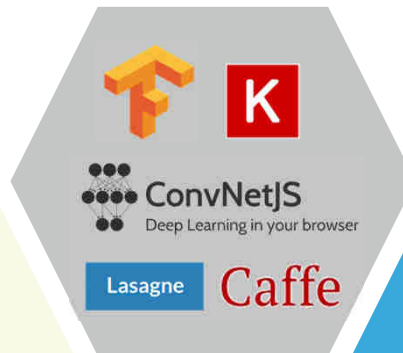
Input your framework-dependent, pre-trained Neural Network into the **STM32Cube.AI** conversion tool

Automatic and 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

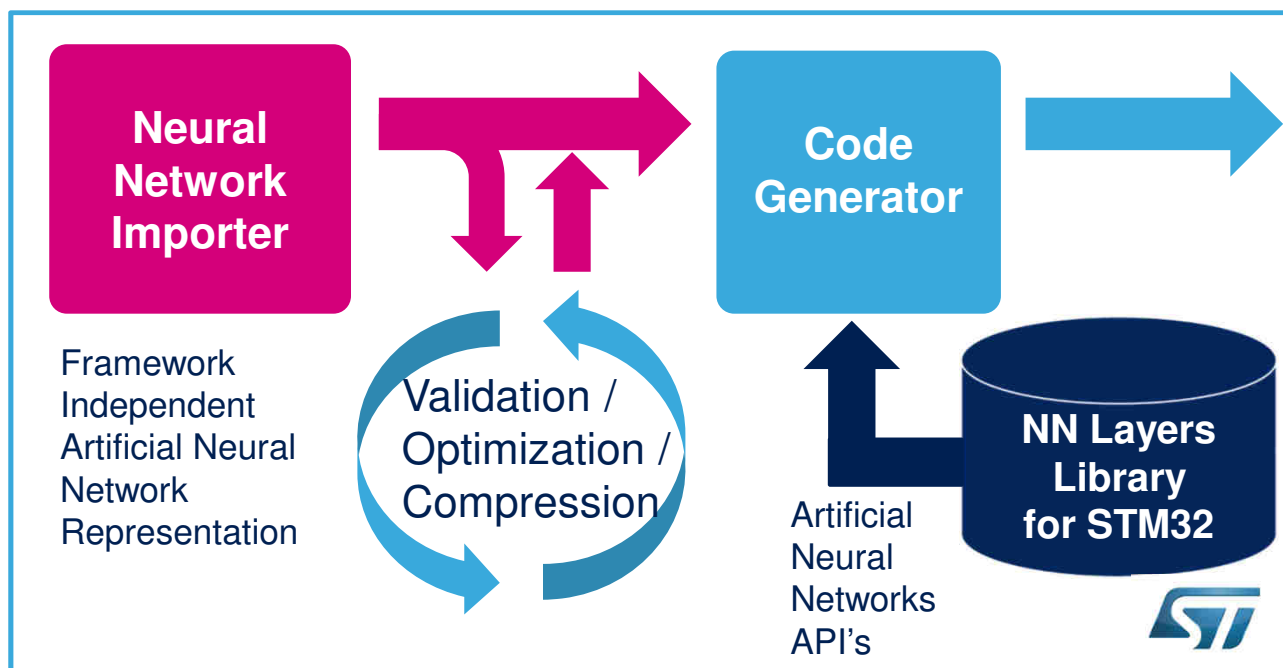


# X-Cube-AI : Architecture

15

**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



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	HMAL	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
☆	STM32F301K6	STM3...	Active	1.272		WLCSP49	64 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F301K8	STM3...	Active	1.272		LQFP32	32 kByt...	16 kBytes	25	72 MHz	0.0	0	0	0
☆	STM32F301R6	STM3...	Active	1.342		LQFP32	64 kByt...	16 kBytes	25	72 MHz	0.0	0	0	0
☆	STM32F301R8	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
☆	STM32F302CB	STM3...	Active	1.782		WLCSP49	64 kByt...	16 kBytes	37	72 MHz	0.0	0	0	0
☆	STM32F302CC	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

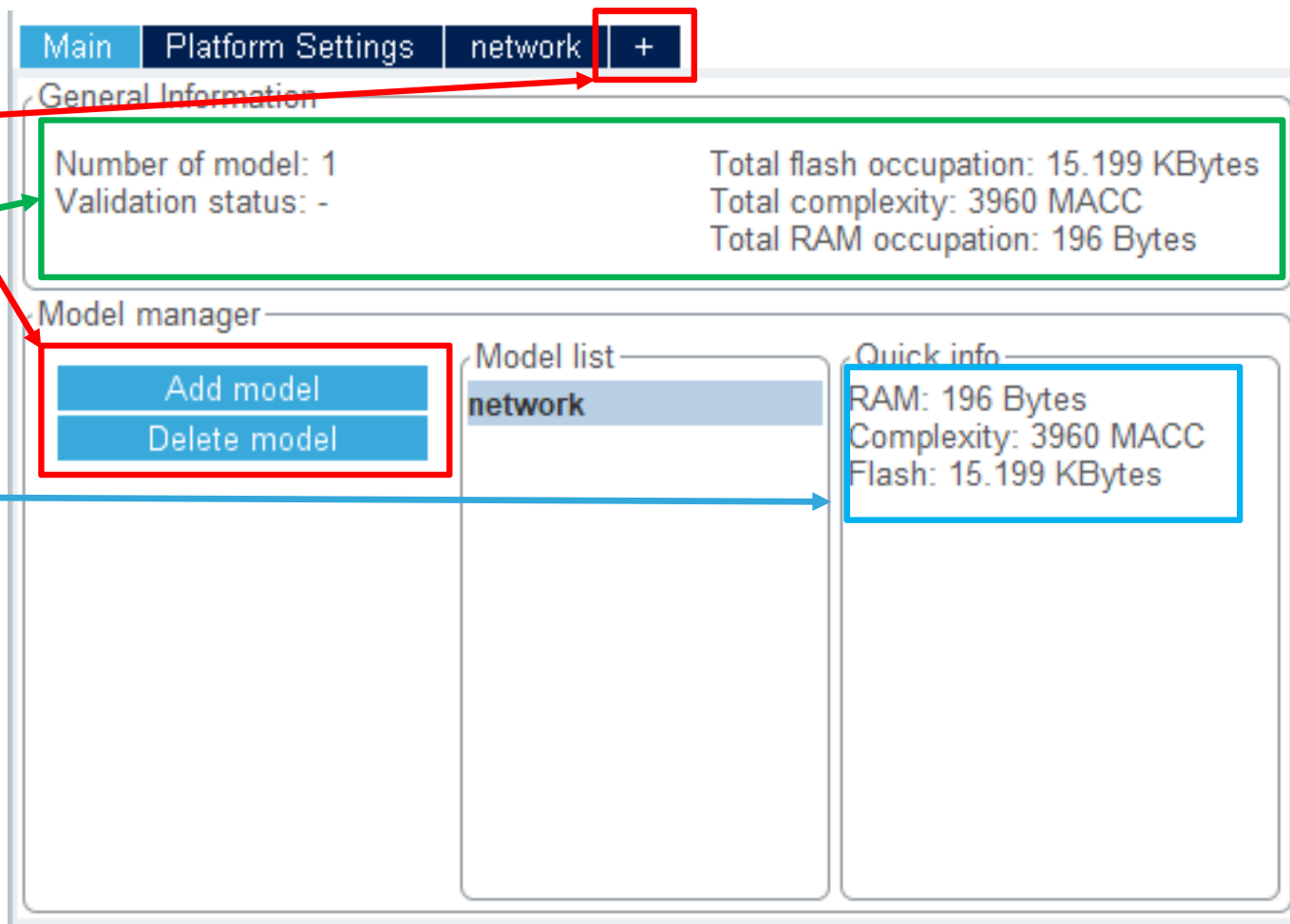
Analyze

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 on different models



The screenshot shows the X-Cube-AI Main tab interface. The top navigation bar includes 'Main', 'Platform Settings', 'network', and a '+' button. The 'General Information' section displays model statistics and resource usage. The 'Model manager' section contains 'Add model' and 'Delete model' buttons. The 'Model list' shows the 'network' model. The 'Quick info' section provides a summary of resources.

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

Model manager	
Add model	Model list network
Delete model	

Quick info
RAM: 196 Bytes
Complexity: 3960 MACC
Flash: 15.199 KBytes

# X-Cube-AI Detailed View

18

- 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

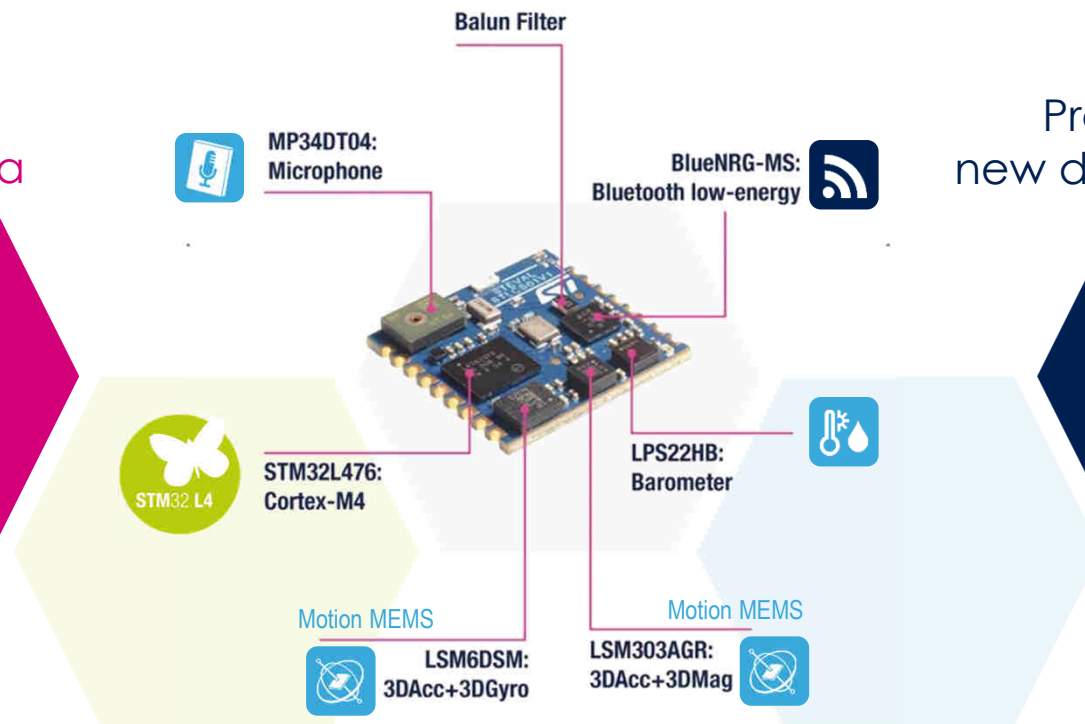
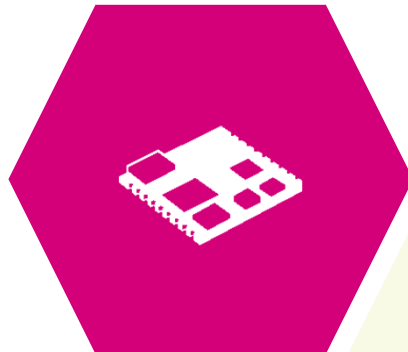
Validate on target



# Form Factor Hardware to Capture and Process Data

19

Capture data



Process & analyze  
new data using trained NN



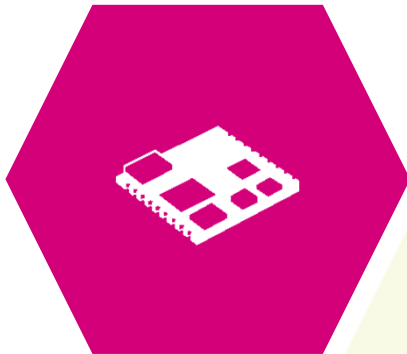


# Form Factor Hardware

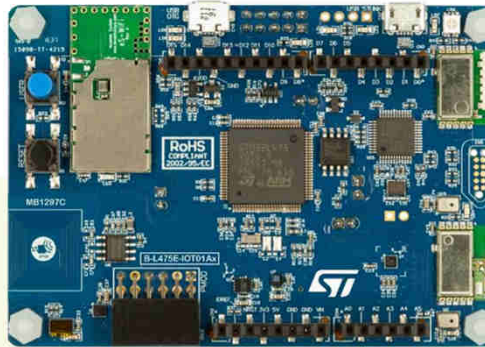
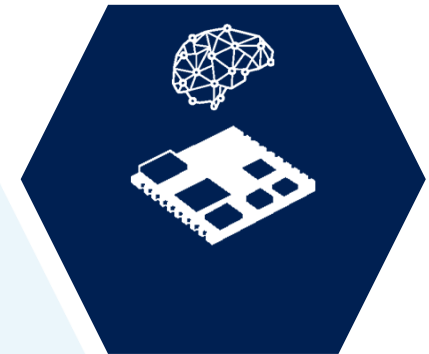
## AI IoT Node for More Connectivity

20

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

21

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 the SensorTile.



ST BLE  
Sensor



**Selected partners**

Neural Networks engineering services support.

Data scientists and Neural network architects.

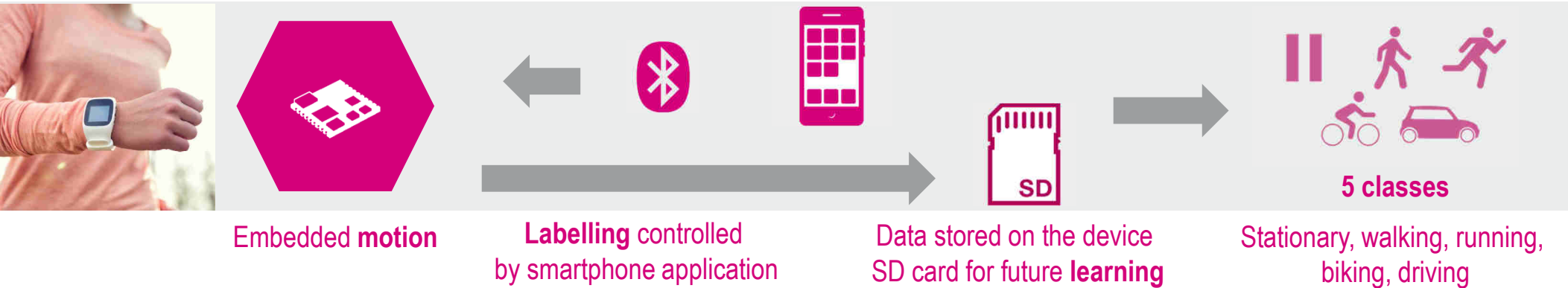
[www.st.com/STM32CubeAI#Partners?](http://www.st.com/STM32CubeAI#Partners?)



# Human Activity Recognition (HAR)

## Motion Example in FP-AI-SENSING1 Package

22

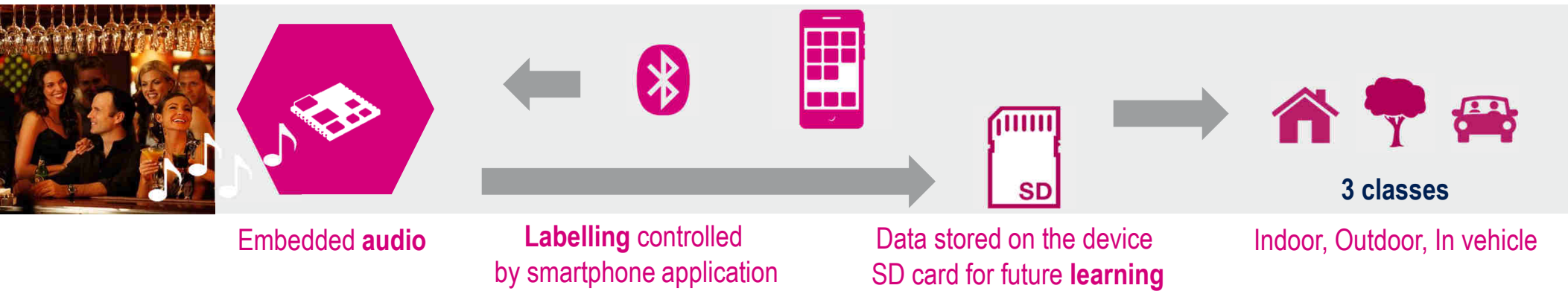




# Audio Scene Classification (ASC)

## Audio Example in FP-AI-SENSING1 Package

23







# Making AI Accessible Now

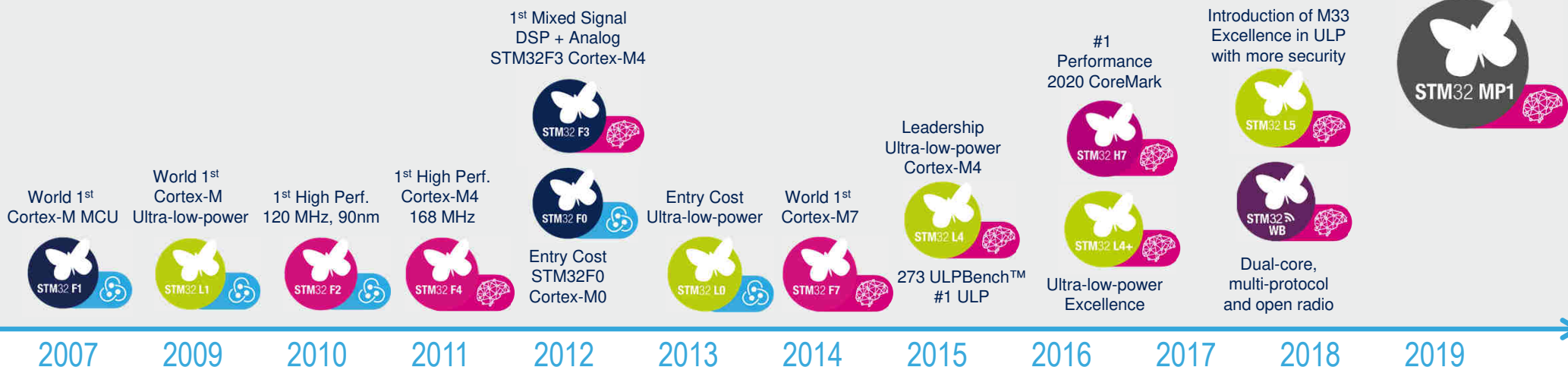
24

## Leader in Arm® Cortex®-M 32-bit General Purpose MCU

Compatible with STM32Cube.AI ecosystem



Compatible with Partner Machine Learning ecosystems



More than 40,000 customers

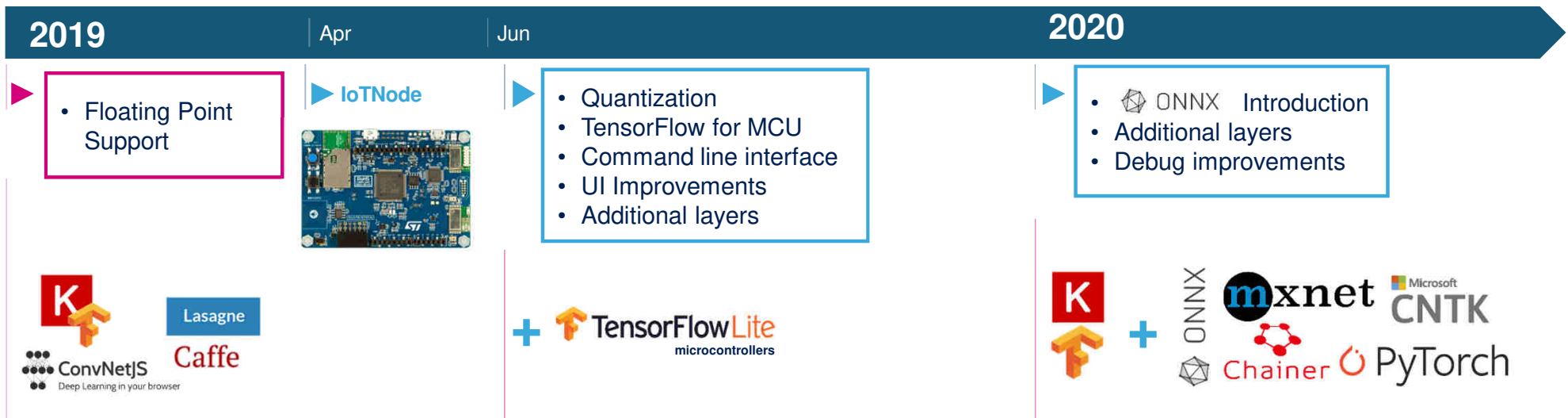
Over 4 Billion STM32 shipped since 2007





# STM32Cube.AI Roadmap

25





# ST Toolbox for Neural Networks

## More Than Just a Conversion Tool

26



Resources

- Function packs for **quick prototyping**
- **Audio** and **motion** examples



- STM32 **Community** for **support** and **idea** exchange
- **Dedicated** topic for Neural Networks

Process & analyze  
new data using trained NN



Convert NN into  
optimized code for MCU



# STM32 Solutions for AI

## More Than Just the STM32Cube.AI

27

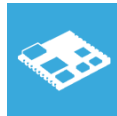
An extensive toolbox to support easy creation of your AI application

**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!

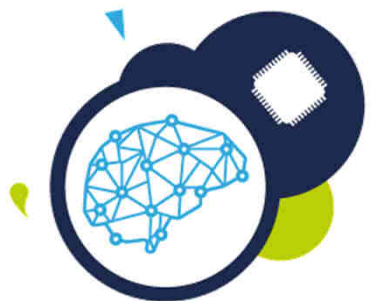


**STM32 Community** with dedicated Neural Networks topic

**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

28



[www.st.com/STM32CubeAI](http://www.st.com/STM32CubeAI)



Capture  
Data

Label  
Data

Train  
NN

STM32  
Cube.AI

Run on  
STM32