AI Application Processing Requirements

Low

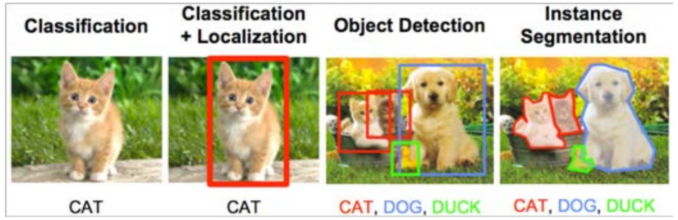
e-AR sensor
Heart rate sensor
Wound healing sensor
Local Processing Unit
Blood pressure sensor
Blood glucose sensor
EMG sensor

- Sensor analysis
- Activity Recognition (motion sensors)
- Stress Analysis or Attention Analysis

Medium

- Audio & sound
- Speech Recognition
- Object detection

High



- Computer Vision
- Multiple Objects Detection/Classification/Tracking
- Speech Synthesis

STM32

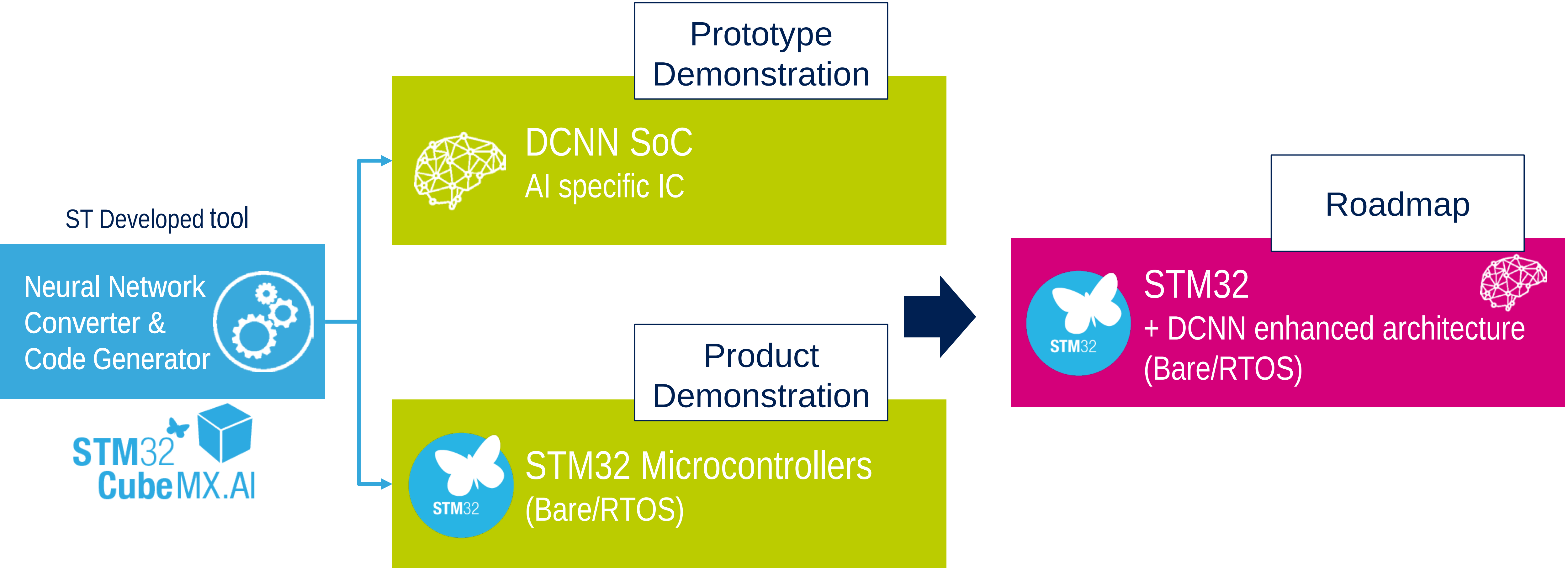
From IP embedded in MCU/MPU to dedicated SOC

STM32 CubeMX.AI

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

- Mandatory to support advanced Audio and Video complex use cases.

ST Solutions for Artificial Intelligence



ST Developed tool

Neural Network Converter & Code Generator

STM32 CubeMX.AI

Prototype Demonstration

DCNN SoC
AI specific IC

Product Demonstration

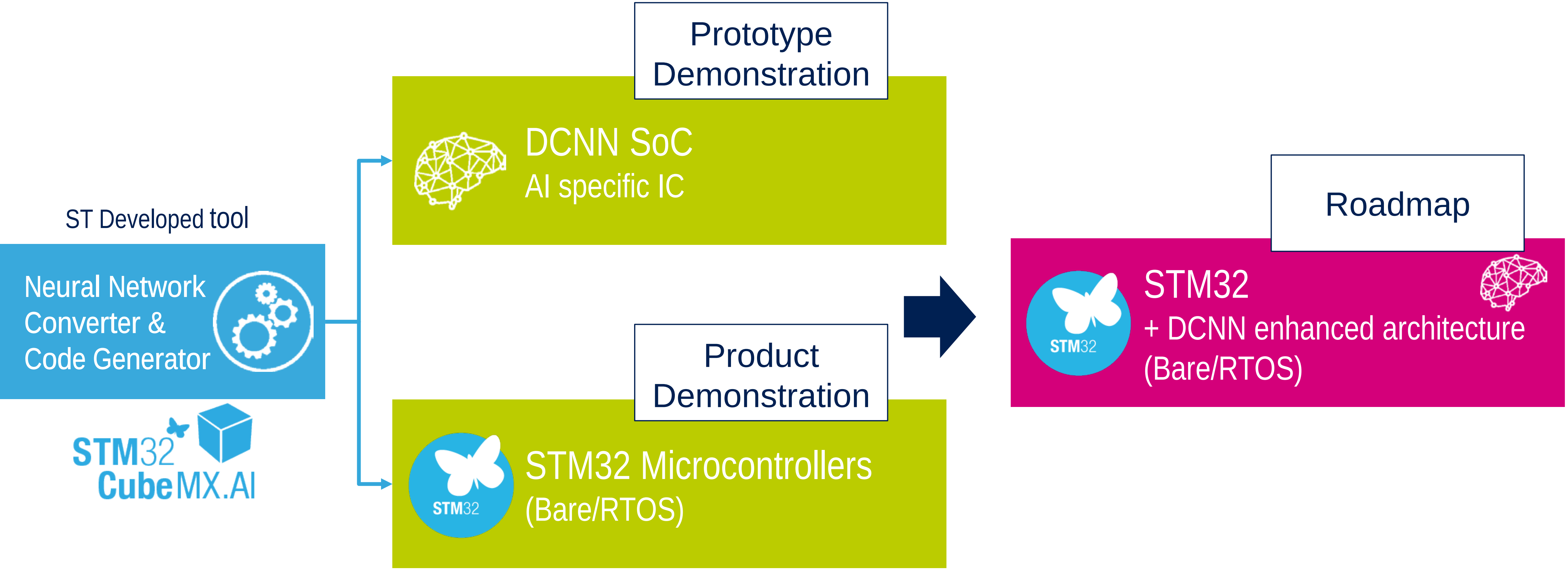
STM32 Microcontrollers
(Bare/RTOS)

Roadmap

STM32
+ DCNN enhanced architecture
(Bare/RTOS)

DCNN = Deep Convolutional Neural Network

ST Solutions for Artificial Intelligence



ST Developed tool

Neural Network Converter & Code Generator

STM32 CubeMX.AI

Prototype Demonstration

DCNN SoC
AI specific IC

Product Demonstration

STM32 Microcontrollers
(Bare/RTOS)

Roadmap

STM32
+ DCNN enhanced architecture
(Bare/RTOS)

DCNN = Deep Convolutional Neural Network

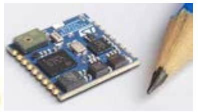
ST Enables AI on STM32

Benefits of Deep Learning now available across all STM32 portfolio

Input your Framework dependent, Pre-Trained neural network into **STM32Cube**MX.AI

Automatic and fast generation of an STM32-optimized library

STM32CubeMX.AI guarantees interoperability with state-of-the-art Deep Learning Design Frameworks

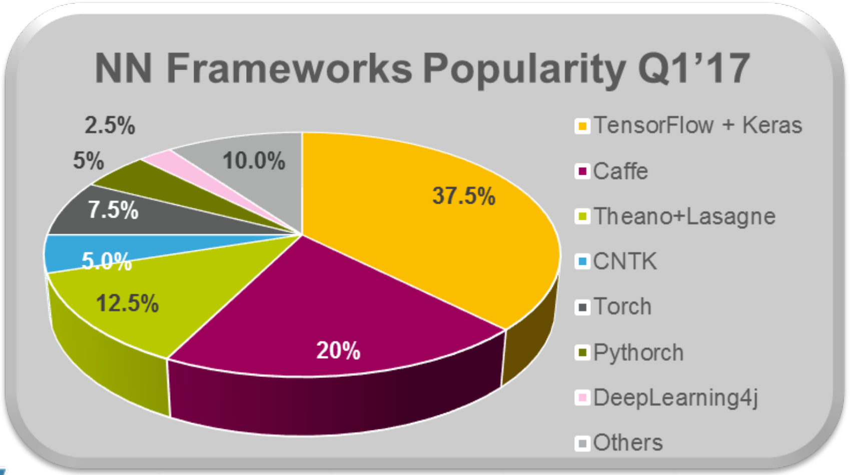


Neural Networks Ecosystem

Benefits of Deep Learning now available across all STM32 portfolio

How to create a Neural Network?

1. Define a problem
2. Collect/find/label data
3. Define topology
4. Design with



Off-the-Shelf Deep Learning Design Frameworks are available to train and create Artificial Neural Networks

Pre-Trained Neural Network Representation
(Creation tool specific format)

Neural Networks

Available Now for STM32

Benefits of Deep Learning now available across all STM32 portfolio

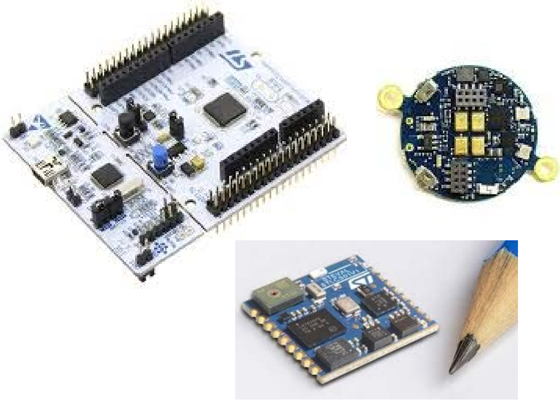
Off-the-shelf :
Pre-trained Neural
Network Model

Deep Learning
Framework dependent

STM32 CubeMX.AI

Embedded Solution
Optimized Neural
Network Code
generated for STM32

Deep Learning SW Solution



... brings your AI-based innovation to the existing STM32 Portfolio

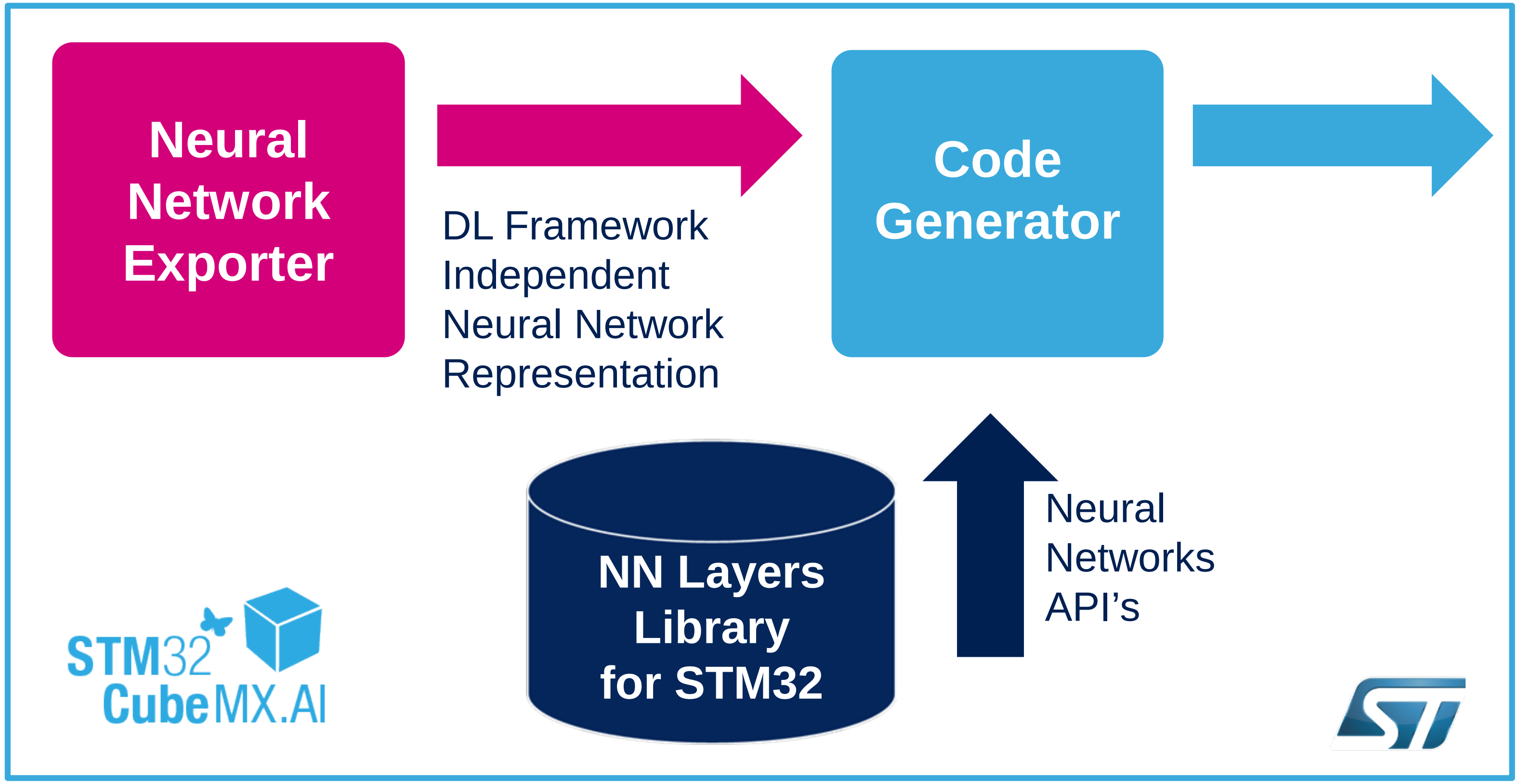
life.augmented

STM32 CubeMX.AI : Architecture

Benefits of Deep Learning now available across all STM32 portfolio

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

Deep Learning
Framework dependent



Neural Network Exporter

DL Framework
Independent
Neural Network
Representation

Code Generator

NN Layers Library for STM32

Neural
Networks
API's

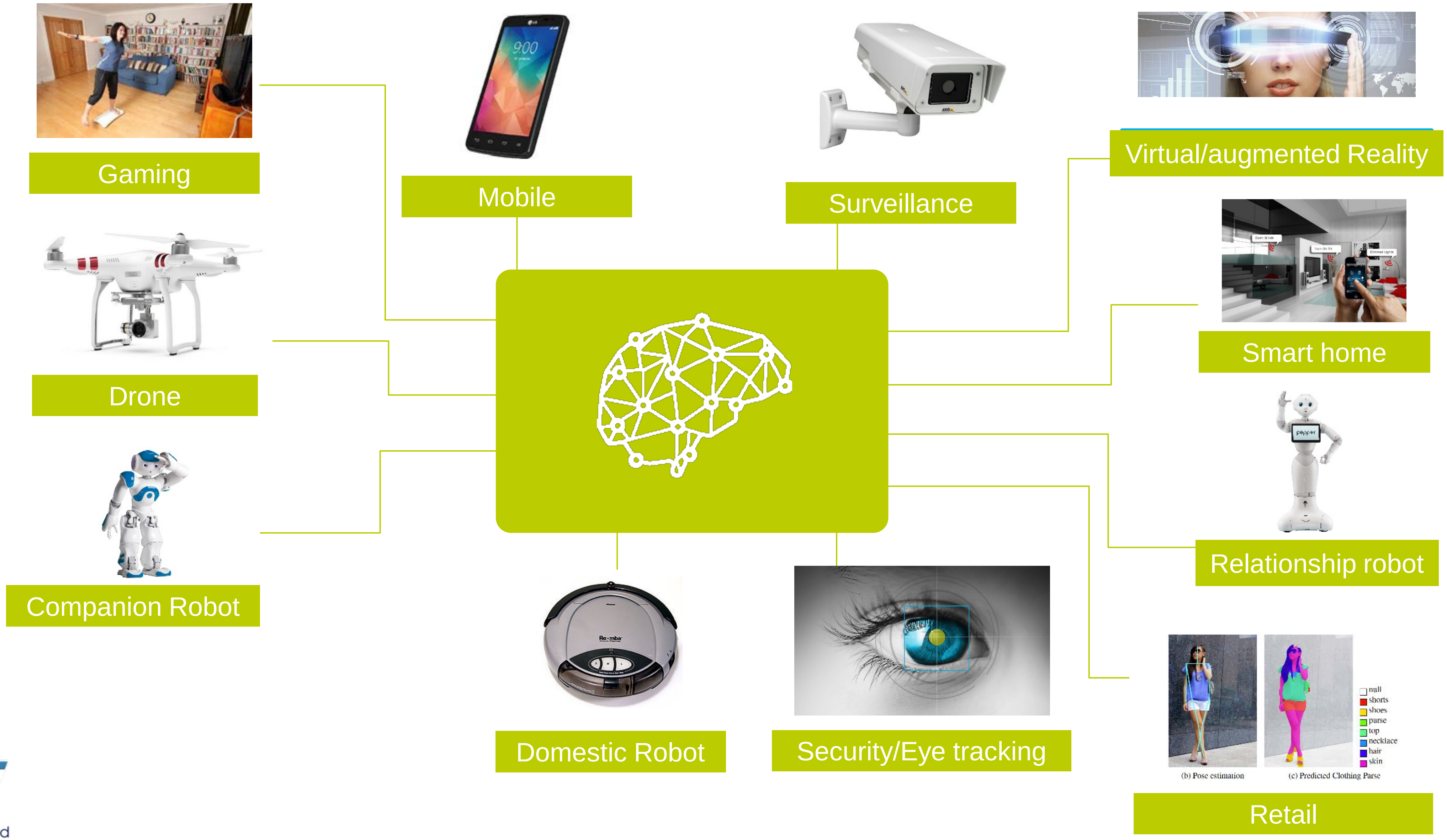
STM32 CubeMX.AI

Embedded Solution
Optimized Neural
Network Code
generated for STM32

This optimized STM32 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

Artificial Intelligence is Everywhere



Gaming

Mobile

Surveillance

Virtual/augmented Reality

Smart home

Drone

Relationship robot

Companion Robot

Domestic Robot

Security/Eye tracking

Retail

life.augmented

Artificial Intelligence for Everything

Analysis

Where am I ?

- Scene classification (audio, video, environmental sensors)

Which objects are in the scene, where are they?

- Video object detection/classification

What am I doing?

- Activity recognition (audio, video, inertial sensors)

What's happening?

- Event recognition (audio, video, inertial sensors, environmental sensors).

User Interaction

- Command detection (audio)
- Speech Recognition (audio)
- Gesture Recognition (inertial sensors, video)
- User identification and mood detection (audio, video)

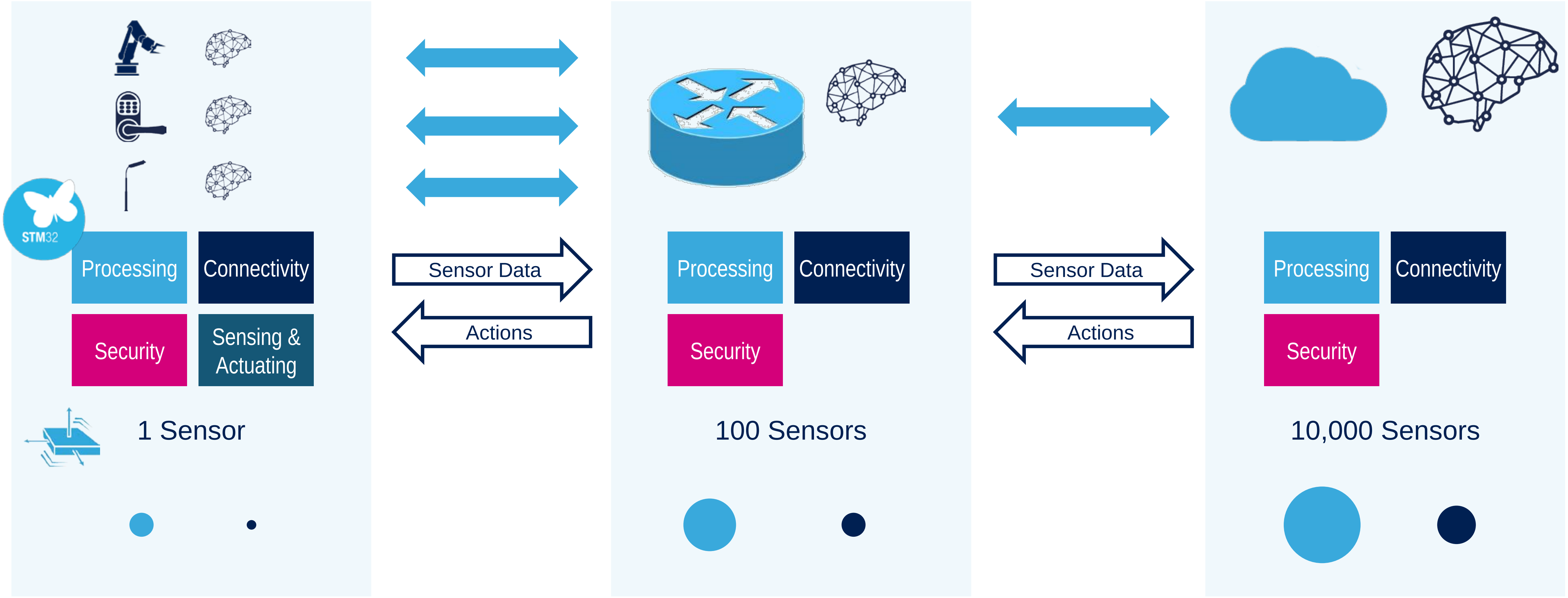
Continuous Learning

- How can I detect unpredictable, unclassified events in dynamic environments?
- Recurrent networks (audio, video, inertial sensors, environmental sensors)

Distributed Artificial Intelligence is a Must

to Increase Systems Efficiency

NODES



STM32

Processing

Connectivity

Security

Sensing & Actuating

1 Sensor

Sensor Data

Actions

Processing

Connectivity

Security

100 Sensors

Sensor Data

Actions

Processing

Connectivity

Security

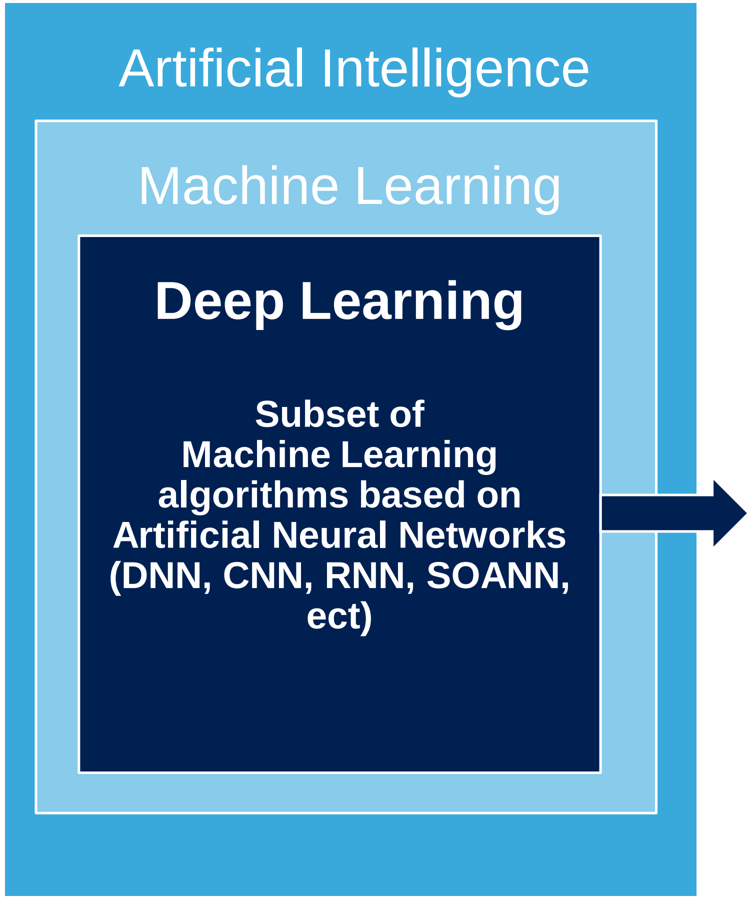
10,000 Sensors

● Processing Requirements

● Connectivity/Bandwidth Requirements

Neural Networks are Key

for Intelligent Nodes



Artificial Intelligence

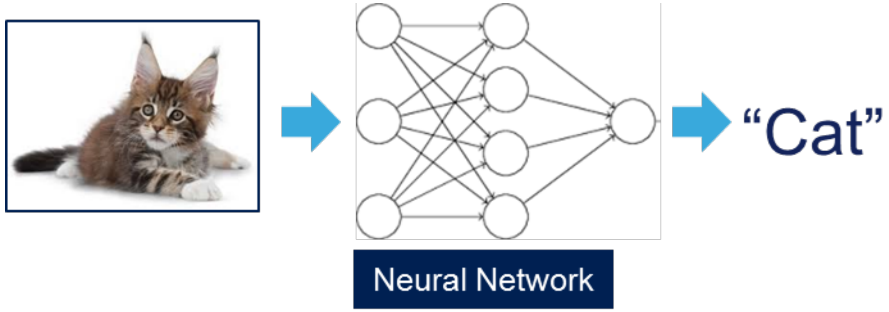
Machine Learning

Deep Learning

Subset of Machine Learning algorithms based on Artificial Neural Networks (DNN, CNN, RNN, SOANN, ect)

What is an Artificial Neural Network?

- Models for input-output transfer function approximation inspired by biological neural networks.
- Capable of modeling and processing highly time varying and non linear relationships between inputs and outputs.
- Exponentially faster and more efficient than traditional computer processing models for typical AI uses like detection, classification, prediction.



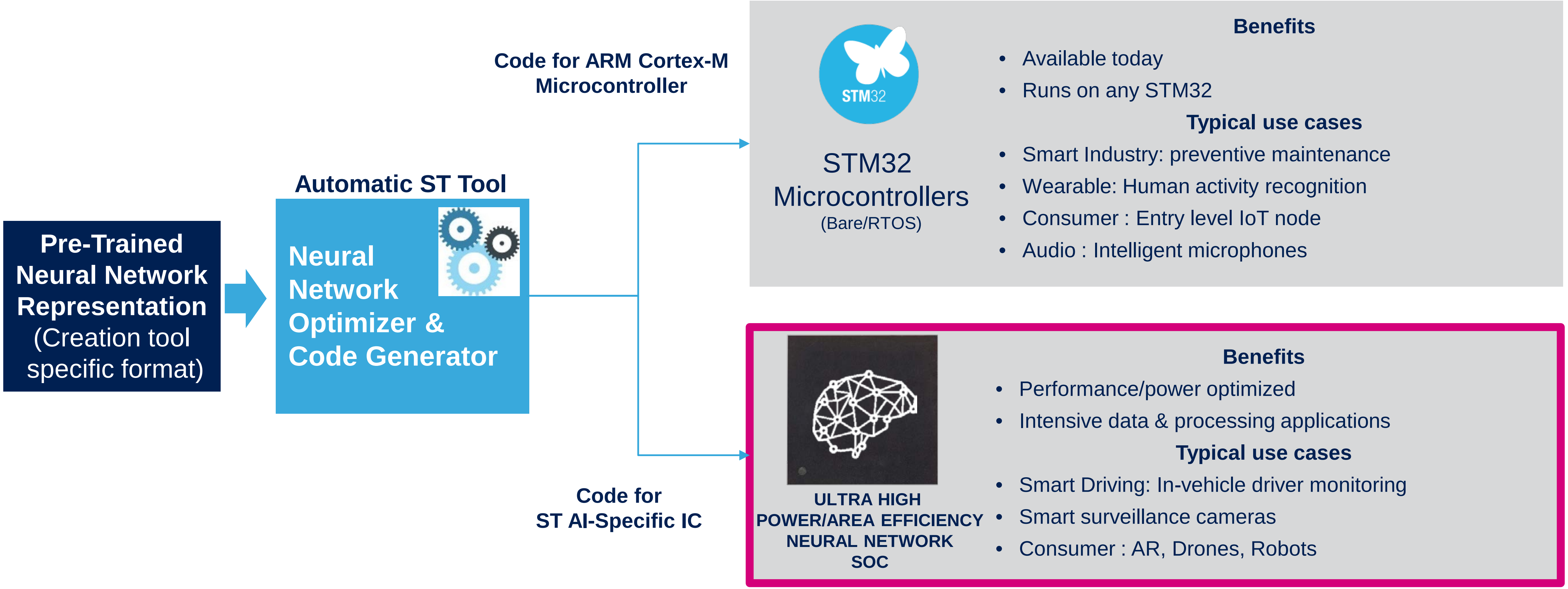
Example: A neural network trained to classify an object in a picture.

How to design ANN's?

Off-the-Shelf Deep Learning Design Frameworks are available to design and train Artificial Neural Networks

Pre-Trained Neural Network Representation
(Creation tool specific format)

ST Enables AI @ the Edge



Pre-Trained Neural Network Representation (Creation tool specific format)

Automatic ST Tool

Neural Network Optimizer & Code Generator

Code for ARM Cortex-M Microcontroller

STM32 Microcontrollers
(Bare/RTOS)

Benefits

- Available today
- Runs on any STM32

Typical use cases

- Smart Industry: preventive maintenance
- Wearable: Human activity recognition
- Consumer : Entry level IoT node
- Audio : Intelligent microphones

Code for ST AI-Specific IC

ULTRA HIGH POWER/AREA EFFICIENCY NEURAL NETWORK SOC

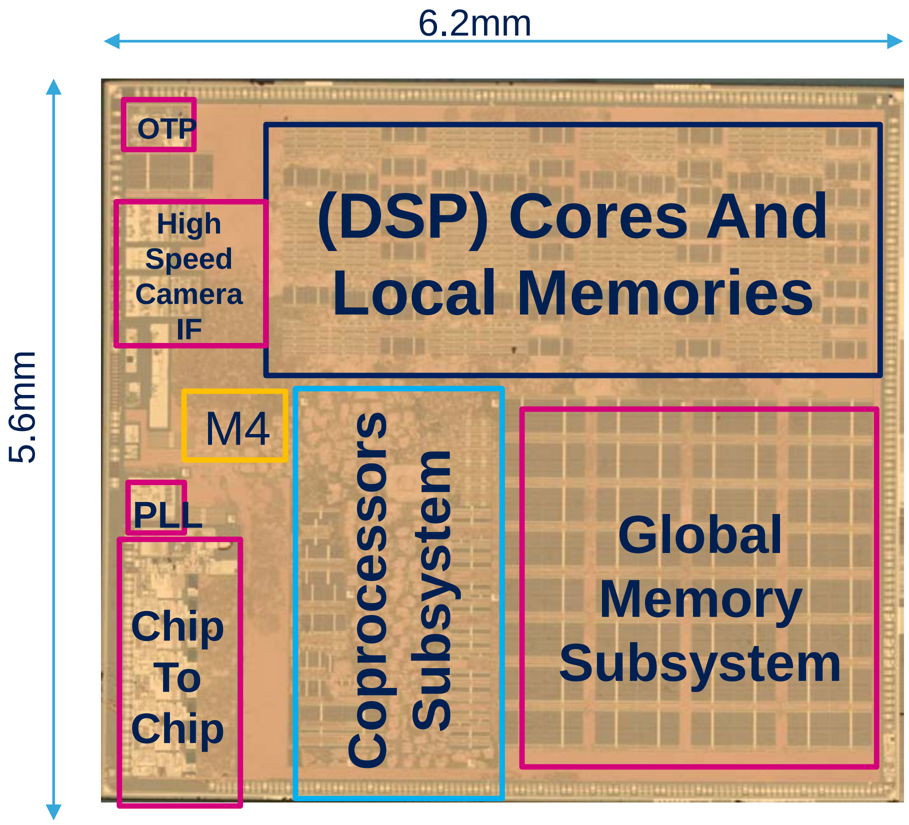
Benefits

- Performance/power optimized
- Intensive data & processing applications

Typical use cases

- Smart Driving: In-vehicle driver monitoring
- Smart surveillance cameras
- Consumer : AR, Drones, Robots

HW Accelerated Deep Learning SoC



A configurable, scalable and design time parametric **Convolutional Neural Network Processing Engine**

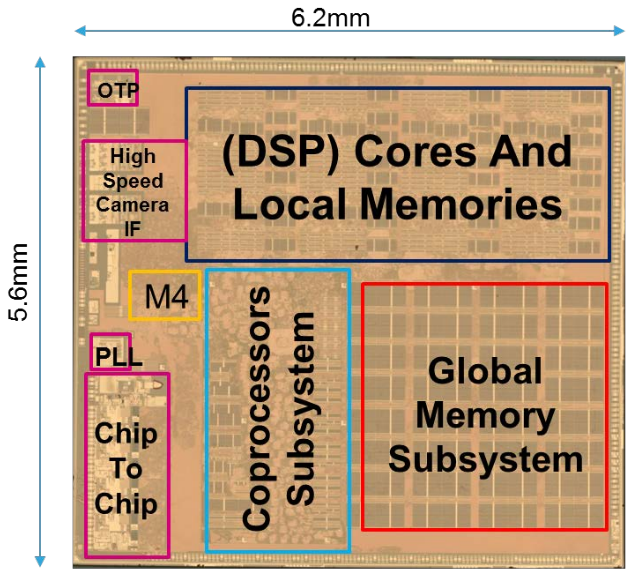
DCNN Convolutional Layers accounts for more than 90% DCNN operations, hence **8 Convolution HW Accelerators** allow high efficiency in area vs GOPS vs power consumption

In addition to **ARM Cortex M4, 8 DSP Clusters** allow both programmability and flexible mapping of diversified, custom DCNN's

Embedded Memory enables further reduction of power consumption required by IOT applications.

life.augmented

DCNN SoC Test Chip Main Features



(*) 1 MAC defined as 2 OPS (ADD + MUL)

Technology	FD-SOI 28nm
Package	FBGA 15x15x1.83
Clock freq	200MHz – 1.175GHz
Supply voltages	0.575V – 1.1V digital – 1.8V I/O
On-chip RAM	4x1 MB (Global), 8x192 KB (DSP), 128 KB (Host)
Host	ARM® Cortex®-M4
DSPs Nr	16
Peak DSP performance **(1.175GHz, 1.1V)**	75 GOPS (dual 16b MAC loop)
Convolutional Accelerators Nr	8
CA size (including local memory)	0.27 sqmm
Max Tot CAs performance **(1.175GHz, 1.1V)**	676 GOPS
Power (**) **@200MHz, 0.6V, 2 CAs**	37.5mW @ 10 FPS
Power (**) **@1GHz, 1V, 8 CAs**	600mW @ 60FPS (not optimized yet)
CAs Peak Efficiency **@200Mhz, 0.575V (Alexnet)**	2.9 TOPS/W