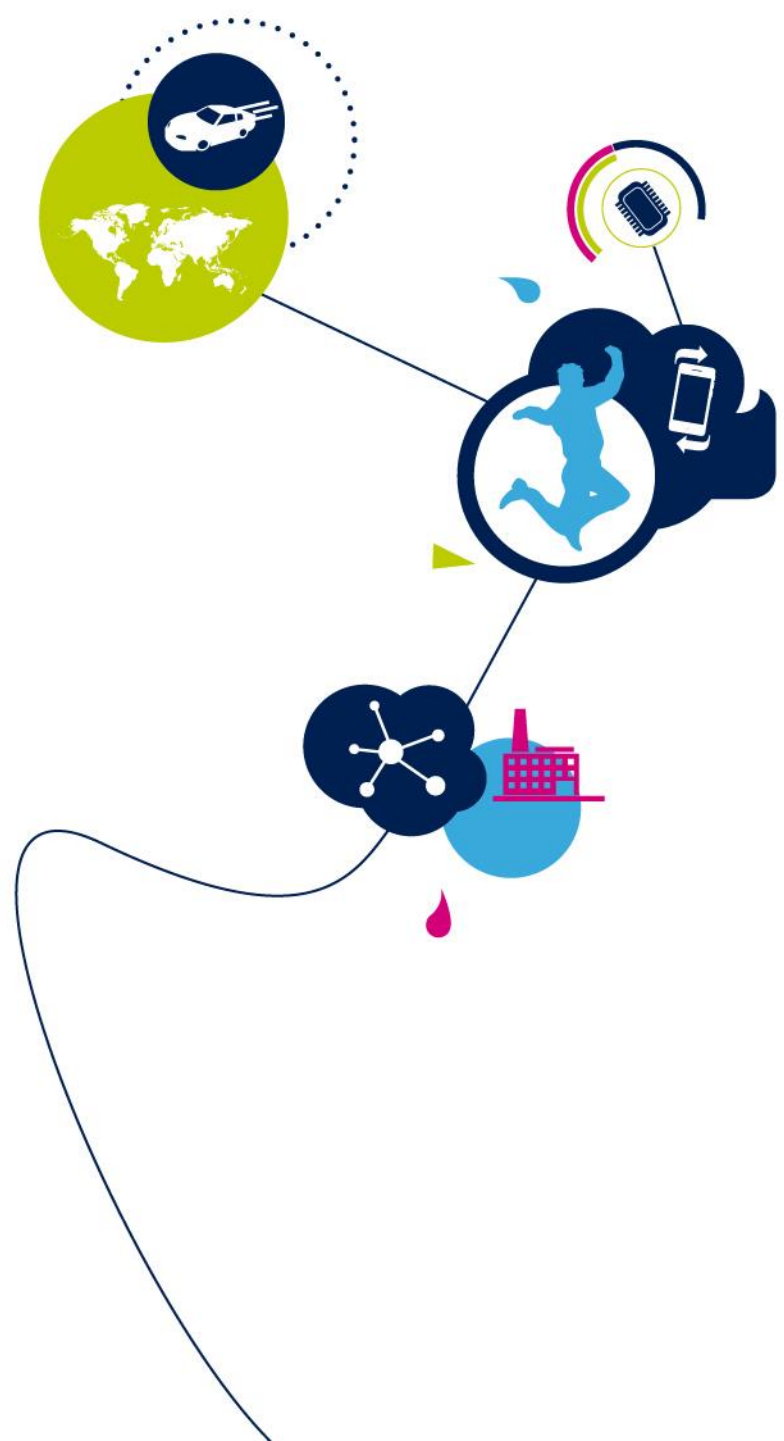
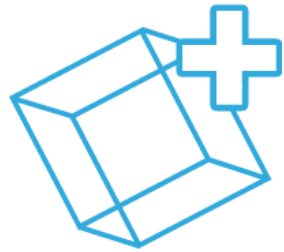
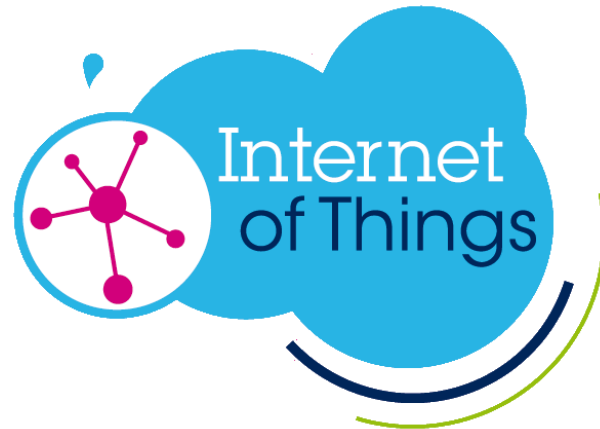


AI Edge Computing – New Paradigm for IoT

Franck Martins

Head of Strategic Technical Marketing – Asia Pacific Region
STMicroelectronics





Smart Things



Smart Home & City

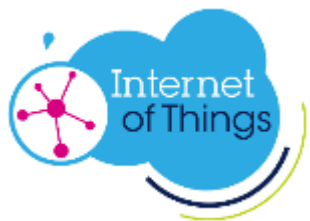


Smart Industry

















Smart Driving





The Building Blocks of the IoT

3

	Processing	Security	Sensing & Actuating	Connectivity	Conditioning & Protection	Motor Control	Power & Energy Management
Smart Things							
Smart Home & City	Ultra-Low Power to High Performance	Scalable security solutions	Full range of sensors and actuators	10 cm to 10 km	Nano Amps to Kilo Amps	Power conversion Monitoring Drivers	Nano Watt to Mega Watt
Smart Industry							



Artificial Intelligence Quick Overview



Artificial Intelligence

Human intelligence exhibited by machines

5

AI is a superset of all the studies to replicate human reasoning with computer systems and is used everyday in our life

- Face / voice recognition
- Autonomous driving
- Stock market trading strategy
- Disease symptom detection
- Predictive maintenance
- Hand writing recognition
- Content distribution on social media
- Fraudulent credit card transaction
- Translation engines
- Suggested shopping
- ...





Artificial Intelligence

Human intelligence exhibited by machines

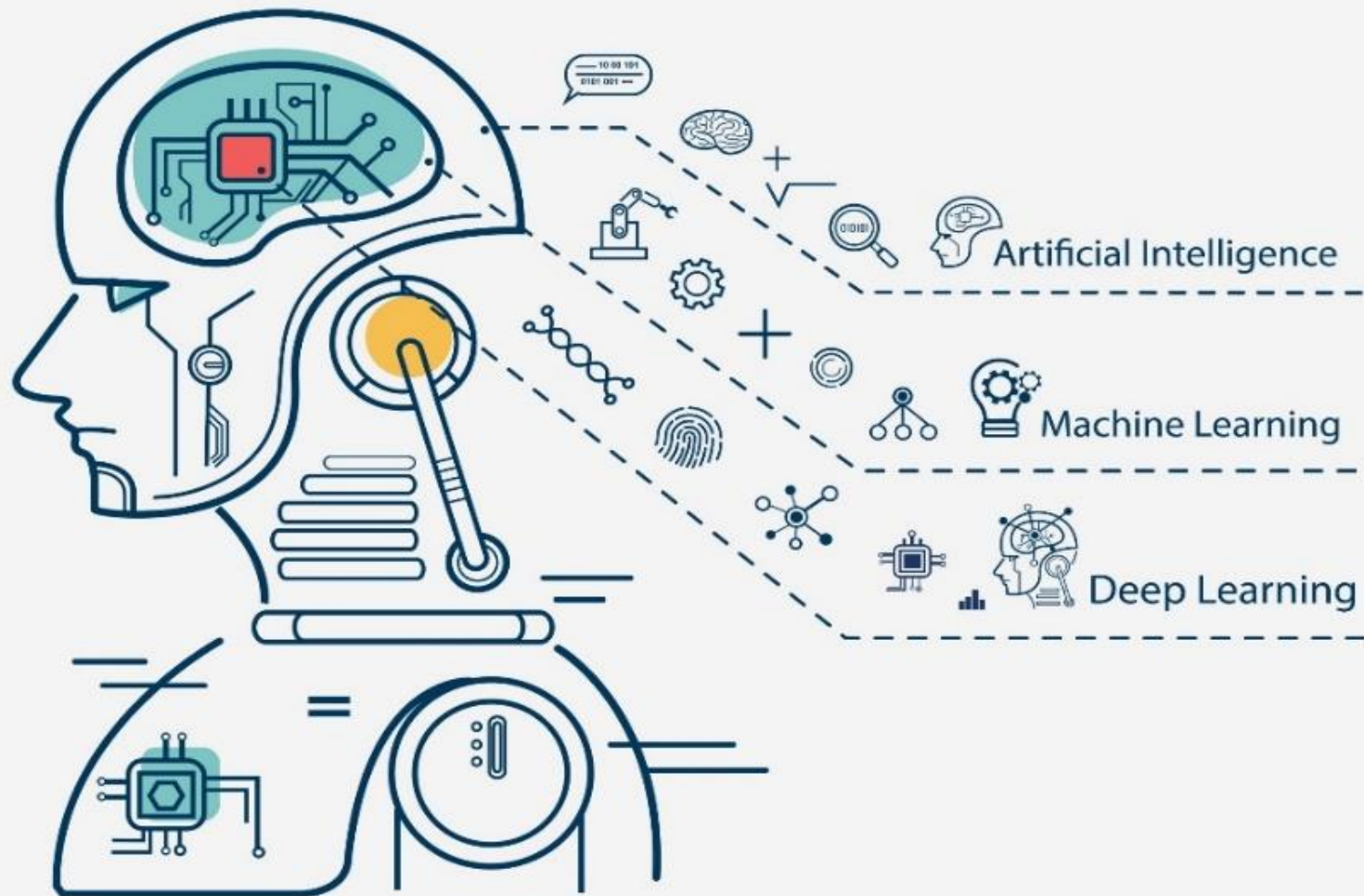
Machine learning

An approach to achieve A.I.

Deep learning

Subset of machine learning algorithm based on artificial neural networks

6



“The science and engineering of making intelligent machines”
(John McCarthy)

Machine learning is a sub-branch of A.I. and is the field of computer science that gives computers the ability to learn without being explicitly programmed

Deep learning is machine learning based on Artificial Neural Networks



Artificial Intelligence

Human intelligence exhibited by machines

Machine learning

An approach to achieve A.I.

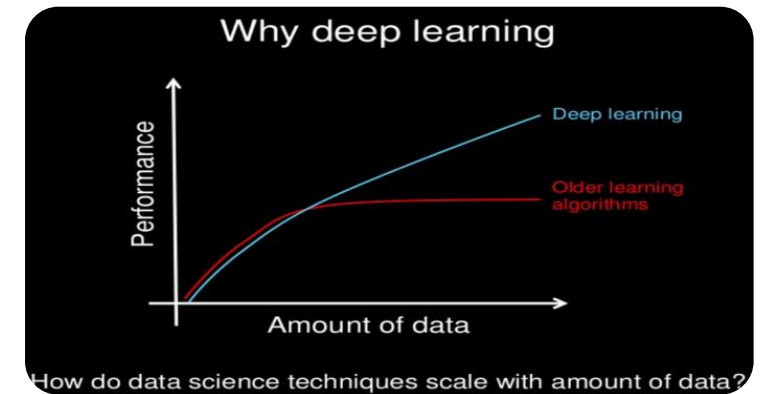
Deep learning

Subset of machine learning algorithm based on artificial neural networks

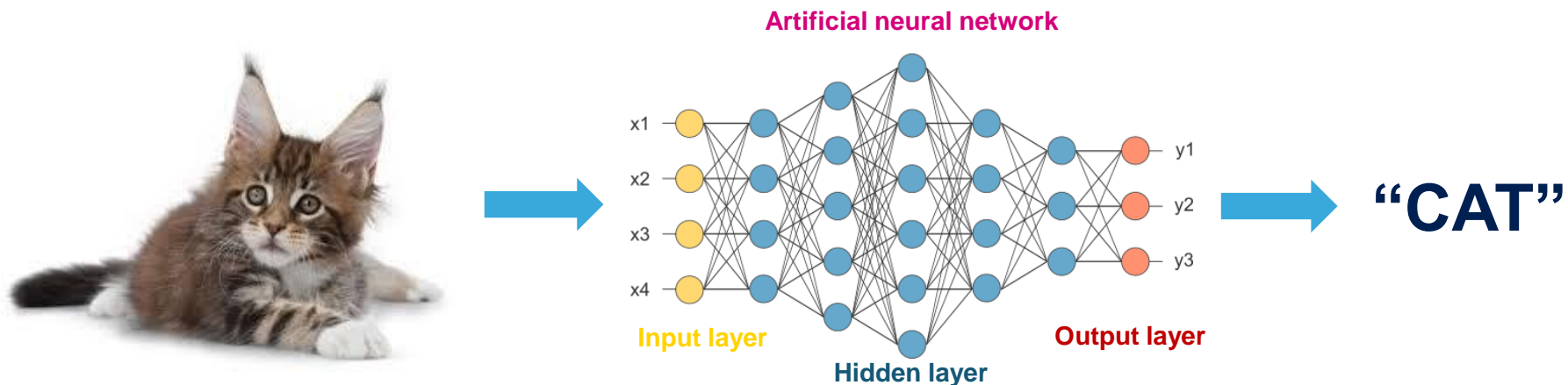
7

- Convolutional Neural Networks are efficient for classification
 - CNN are exponentially more accurate and efficient than traditional computer processing models for AI use cases like recognition, identification and classification tasks

Problem	Dataset	Best accuracy without CNN	Best accuracy using CNN	Difference
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%



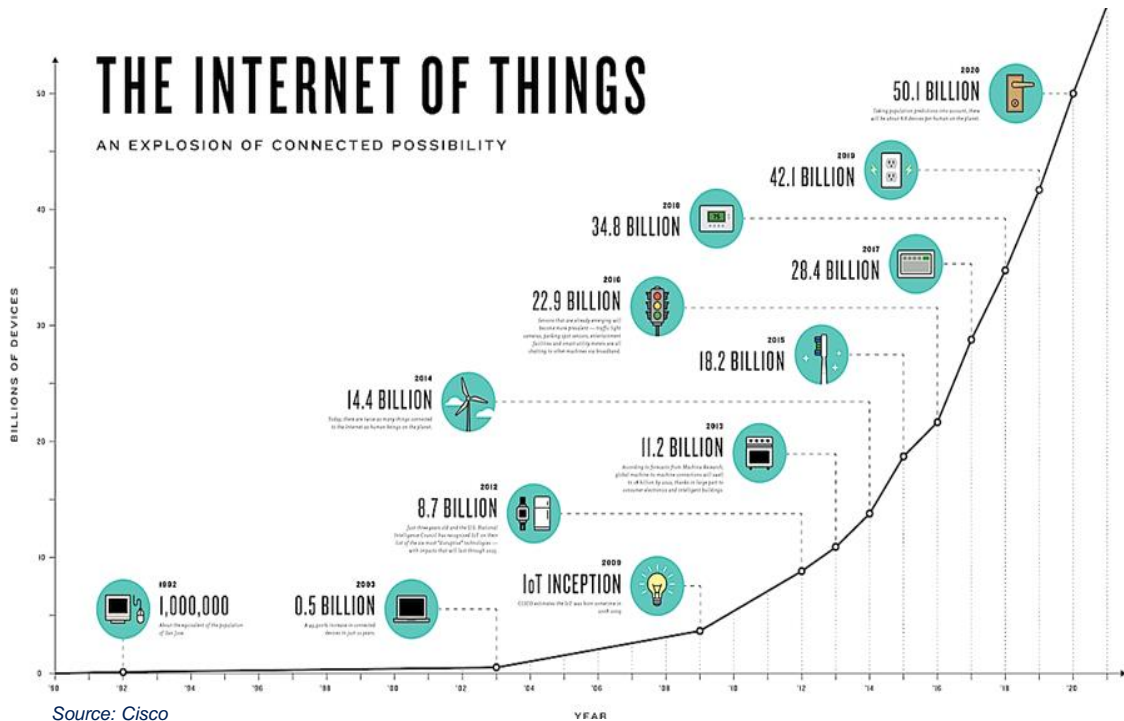
- Example: Neural network that has been trained to recognize an animal from a picture





IoT Pushes AI to the Edge

8



The world is producing excessive amounts of "unstructured data" that need to be reconstructed (IBM's CTO Rob High)

Source: Tractica



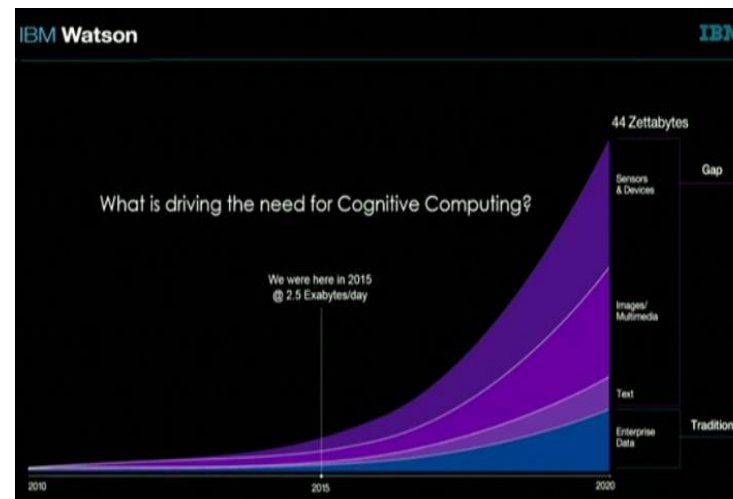
1 Billion cameras WW (2020)
30B Inference/sec



30 images per second
200ms latency



50% of world at less than 8mbps
Only 73% 3G/4G availability WW



Source: IBM

Since 2015, roughly 2.5 Exabyte of data are being generated per day. Projection shows a 44 Zettabytes of data per day by 2020.

"A PC will generate 90 megabytes of data a day, an autonomous car will generate 4 terabytes a day, a connected plane will generate 50 terabytes a day."

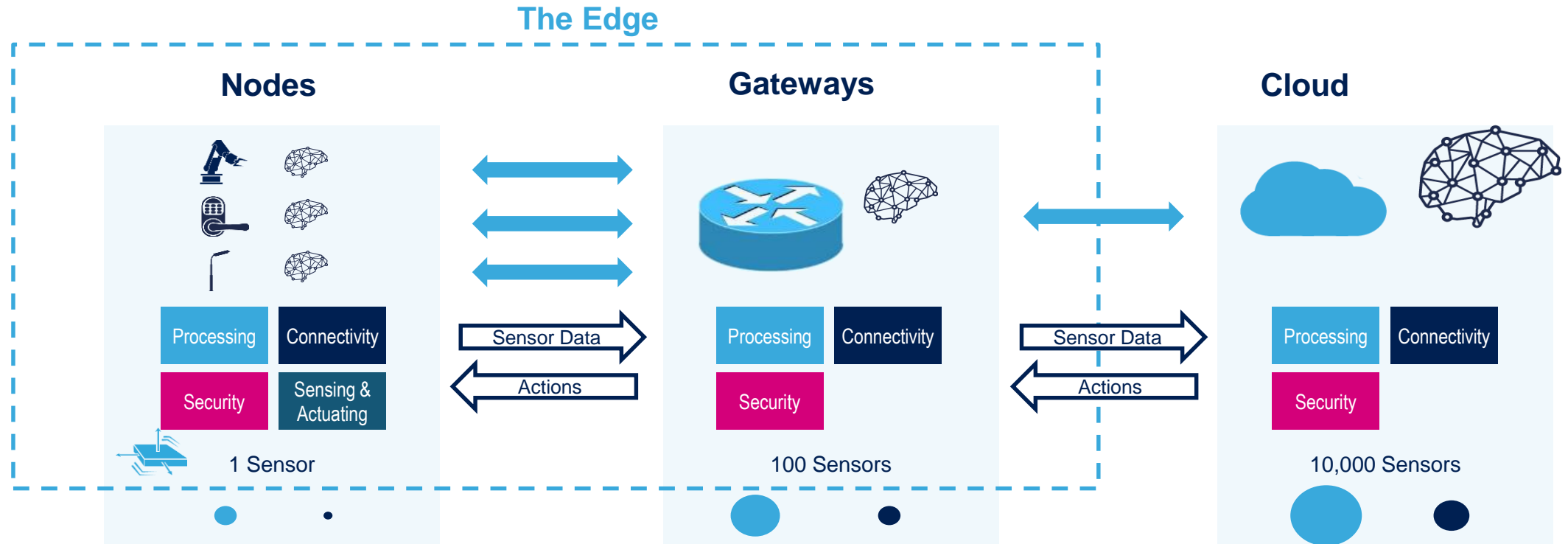
Source: Samsung HBM



The Edge Will Beat The Cloud

9

“Accelerating AI at the edge is critical in enabling Arm’s vision of connecting a trillion IoT devices.” – Rene Haas (VP ARM)



Billions of IoT devices and associated reams of data will make the distribution of AI to the Edge an absolute necessity, some of the major benefits will be:

- Real-time processing to ensure low-latency response (safety issue)
- Connectivity (Cloud) availability and bandwidth
- Data privacy and Security
- Power consumption
- Data sorting, filtering, pre-processing at Edge before Cloud
- Offload cloud processing



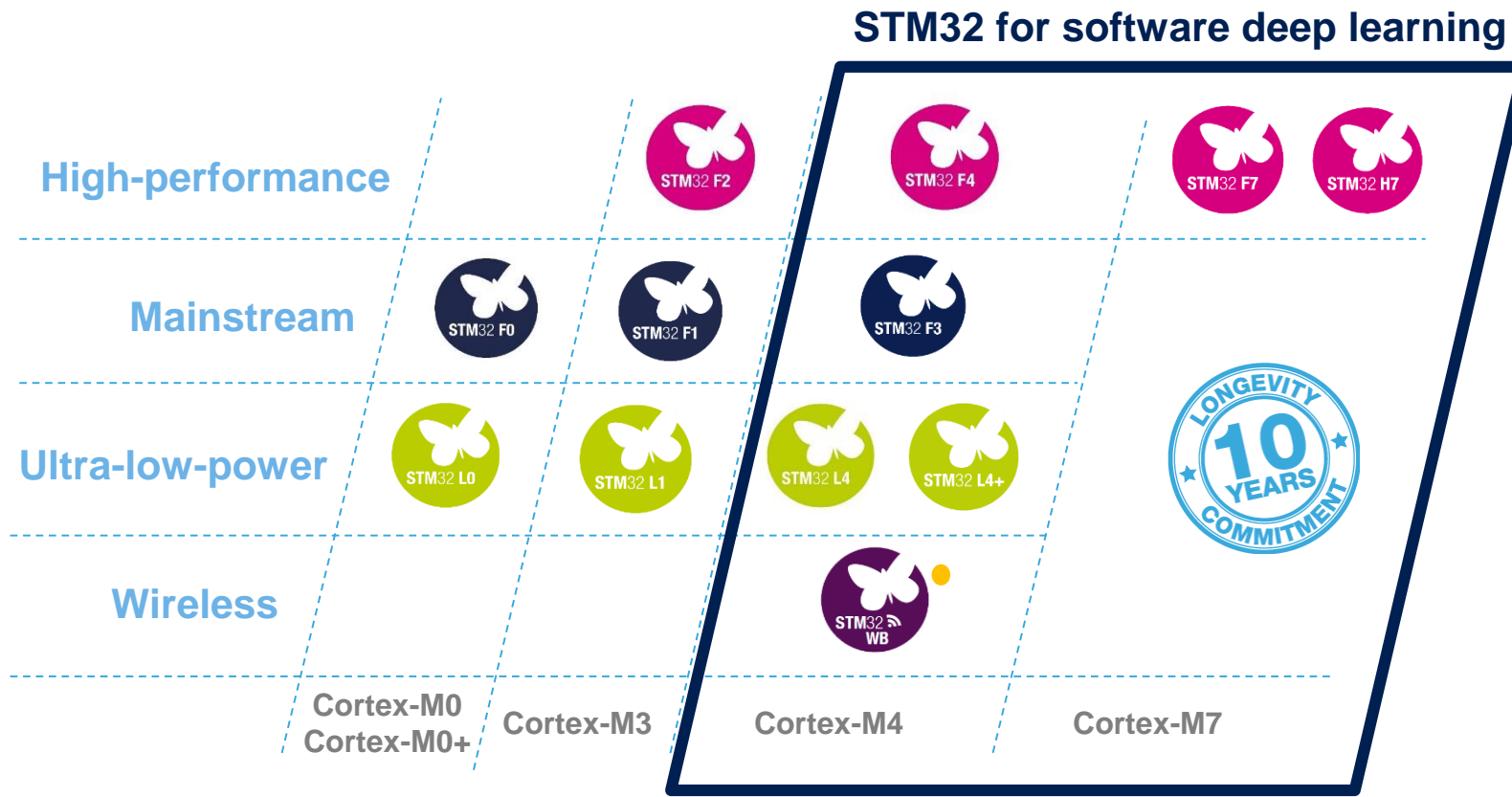
Artificial Intelligence Neural Networks on STM32 Microcontrollers



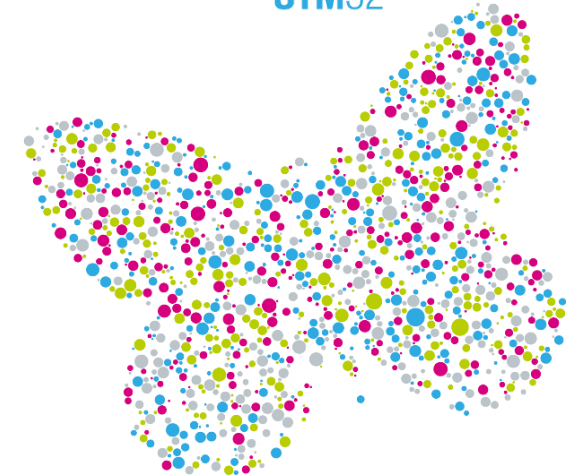
Artificial Intelligence STM32 Solutions

11

12 product series / more than 50 product lines



10
years of
STM32



More than 40,000 customers
3 Billion STM32 shipped since 2007



AI Application Processing Requirements

12

Low



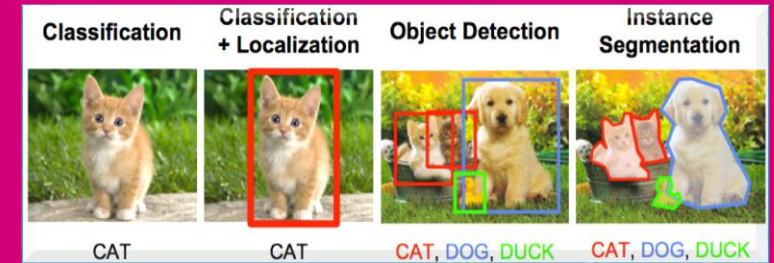
- Sensor analysis
- Activity recognition (motion sensors)
- Stress analysis or predictive maintenance

Medium



- Audio and sound
- Speech recognition
- Object detection

High



- Computer vision
- Multiple object detection, classification, tracking
- Speech synthesis

STM32
(hundreds MOPs)

From IP embedded in MCU/MPU to dedicated SoC
(GOPs to TOPs)



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

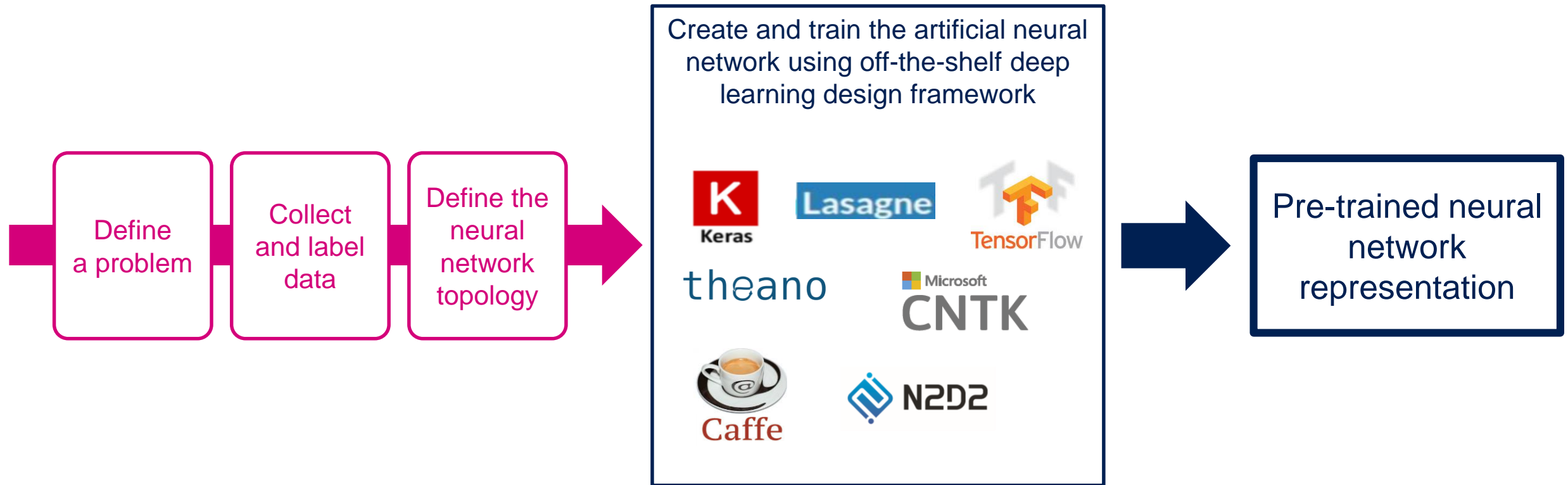
- Mandatory to support complex Audio and Video use cases.



How to Create a Neural Network?

13

Training





Neural Network Implementation on STM32

14

ST has developed a specific tool called **STM32CubeMX.AI** which brings AI based innovation to the existing STM32 portfolio

Training



Off-the-shelf

Pre-trained neural network model
Deep learning framework dependent

Conversion



Embedded Solution

Optimized neural network for STM32
Code generated

Inference



Software Deep Learning solution

STM32CubeMX.AI brings:

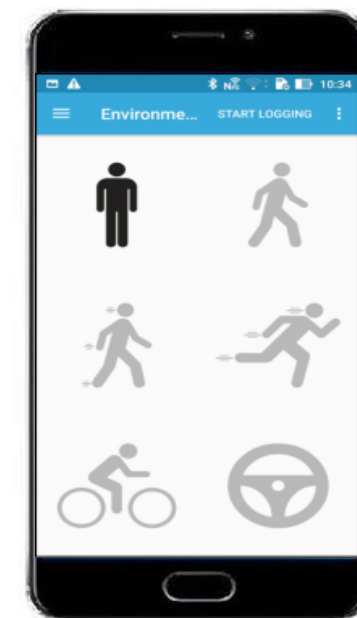
- Increasing programmers productivity
- Interoperable with off-the-shelf deep learning tools
- Allowing best use of constrained processing and memory resources



Human Activity Recognition

15

- Able to detect 5 classes
 - Stationary, walking, running, cycling & driving
 - Based on 3-axis accelerometer data only
- Neural network design
 - ST proprietary CNN
 - ST training / testing database (2.4Millions samples collected)
- STM32CubeMX.AI neural network detail
 - Complexity => 69067 MACC
 - Memory footprint => RAM 4.13KB / Flash 5.92KB
- ST platform
 - STM32L476 / 80MHz ultra low power Cortex-M4
- Details on implementation
 - Pre / post processing included => filtering, gravity compensation and temporal filter (8.38ms)
 - 1 activity classification per second





Keyword Spotting

16

- Keyword spotting
 - Audio system wake up
 - Such has Amazon “Alexa” trigger word
 - 16KHz audio data sampling rate
 - Neural network design
 - ST proprietary CNN
 - ST training / testing database (1540 “Marvin” keyword samples, 1240 background noise samples, 1440 non-keyword samples)
 - STM32CubeMX.AI neural network details
 - Complexity => 83629 MACC
 - Memory footprint => RAM 1.02KB / Flash 320.52KB
 - ST platform
 - STM32L476 (80MHz ultra low power Cortex-M4)
 - Implementation details
 - Neural network inference time + MFCC pre-processing => 25.5ms

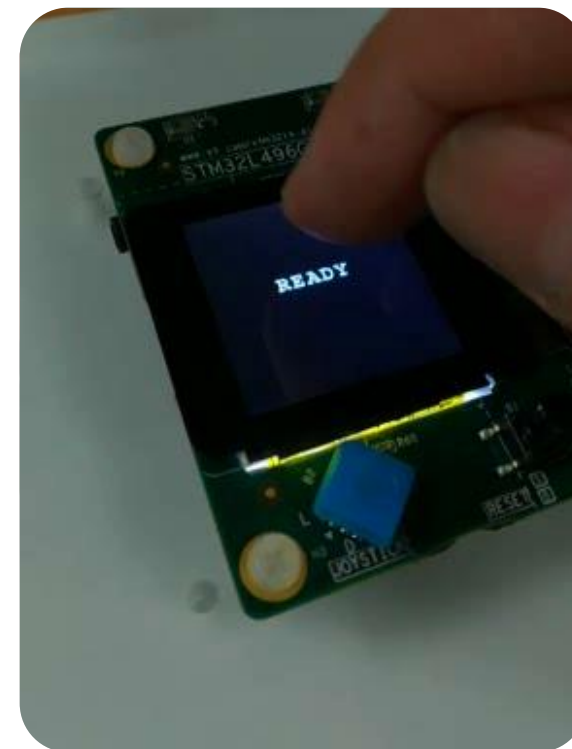




Alpha Numeric Character Recognition

17

- Letters and numbers detections drawn with fingertip on a touch screen display
 - Smart-watch size display creating a message and executing a specific action (i.e. “Call Mum” triggers a specific phone call)
- Neural network design
 - ST proprietary CNN
 - EMNIST training database (36 classes)
- STM32CubeMX.AI neural network details
 - Complexity => 2.27M MACC
 - Memory footprint => RAM 23KB / Flash 633KB
- ST platform
 - STM32L496 (80MHz ultra low power Cortex-M4)
- Implementation details
 - Neural network inference time => 286ms

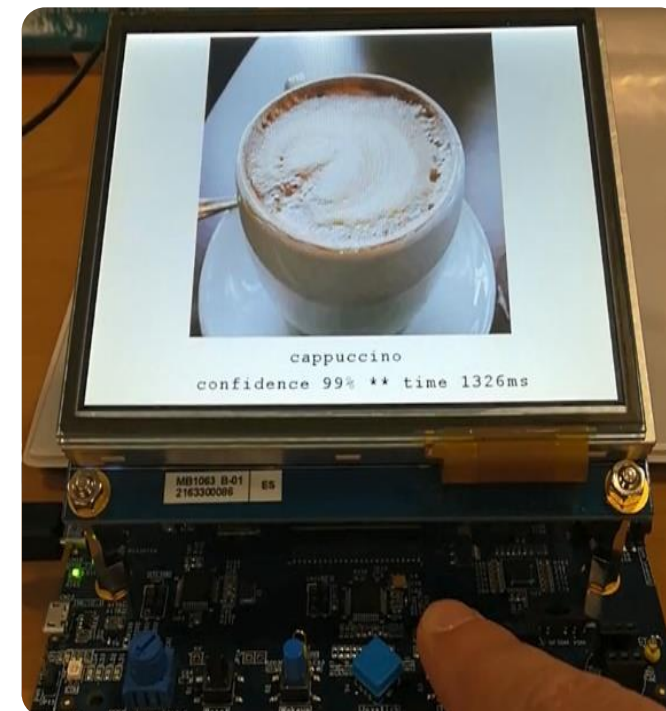




Food Classification (MobileNet tuned by ST)

18

- Neural network design
 - Off-the-Shelf NN => https://github.com/fchollet/deep-learning-models/releases/download/v0.6/mobilenet_2_5_224_tf.h5
 - ST retrained database (with 250 images / 224x224)
- Able to detect 18 classes
- STM32CubeMX.AI neural network details
 - Complexity => 39.2M MACC
 - Memory footprint => RAM 1.61MB / Flash 859KB
- ST platform
 - STM32H743XIH6 (400MHz high performance Cortex-M7)
- Implementation details
 - No cropping / no downscaling done by STM32H7
 - Neural network inference => 1.32sec/frame





life.augmented