ANIE FEDERAZIONE

CONFINDUSTRIA

Ubiquitous AI

Smart and secure IoT as enabler contributors boosting servitization approaches

Claudio Marchisio

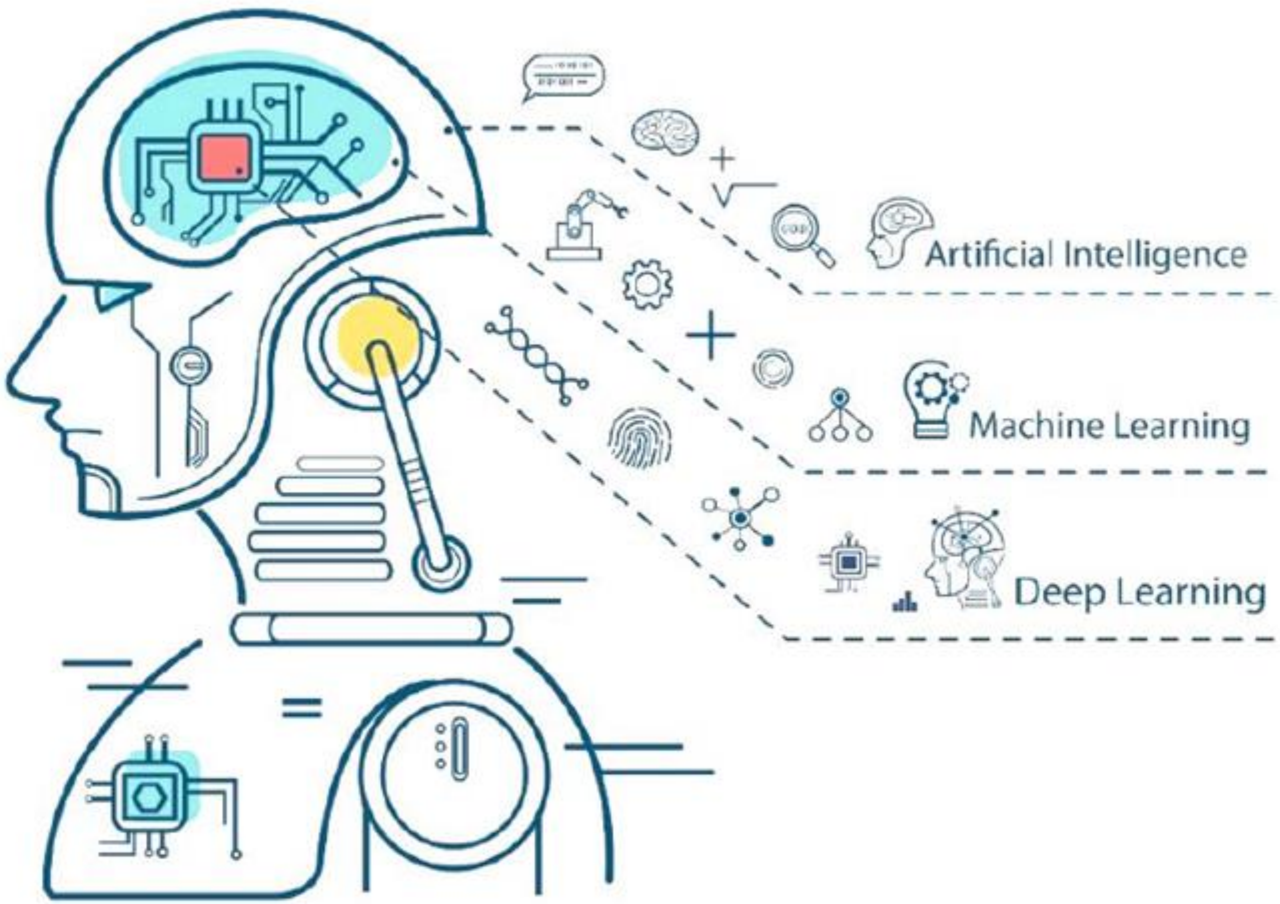
STMicroelectronics

System Research & Applications / AI SW & Tools

Organizzato da ANIE SERVIZI INTEGRATI

Edge AI

Artificial Intelligence
Some definitions



- AI: any technique which enables a computer to mimic human behavior
- ML is a subset of AI that provides systems the ability to automatically learn and improve from data without being explicitly programmed
- DL is a subset of ML, utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning

life.augmented

ANIE FEDERAZIONE
FEDERAZIONE NAZIONALE IMPRESE ELETTROTECNICHE ED ELETTRONICHE
CONFINDUSTRIA

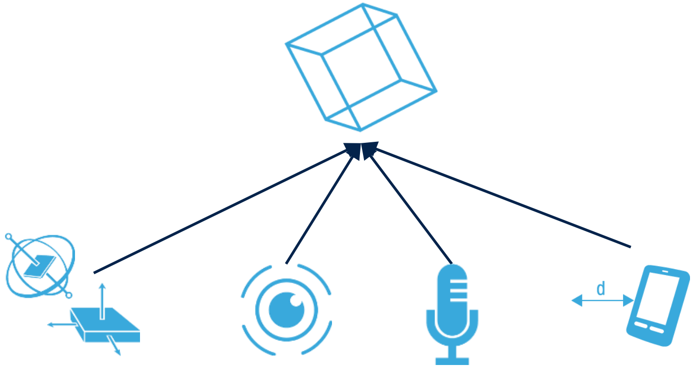
The building blocks of the IoT

	Processing	Security	Sensing & Actuating	Connectivity	Conditioning & Protection	Motor Control	Power & Energy Management
Smart Things Smart Home & City Smart Industry	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

life.augmented

What is Edge AI?

- **Edge AI** means that AI algorithms are executed locally on a hardware device to process the data generated by the attached sensors
- An **Edge AI device** processes data, extracts information and takes decisions without a connection
- Anyhow a connection (*IoT nodes*) is useful to:
  - receive FW updates, update Deep Learning models, etc.
  - send (reduced) data and results of local processing
  - receive commands

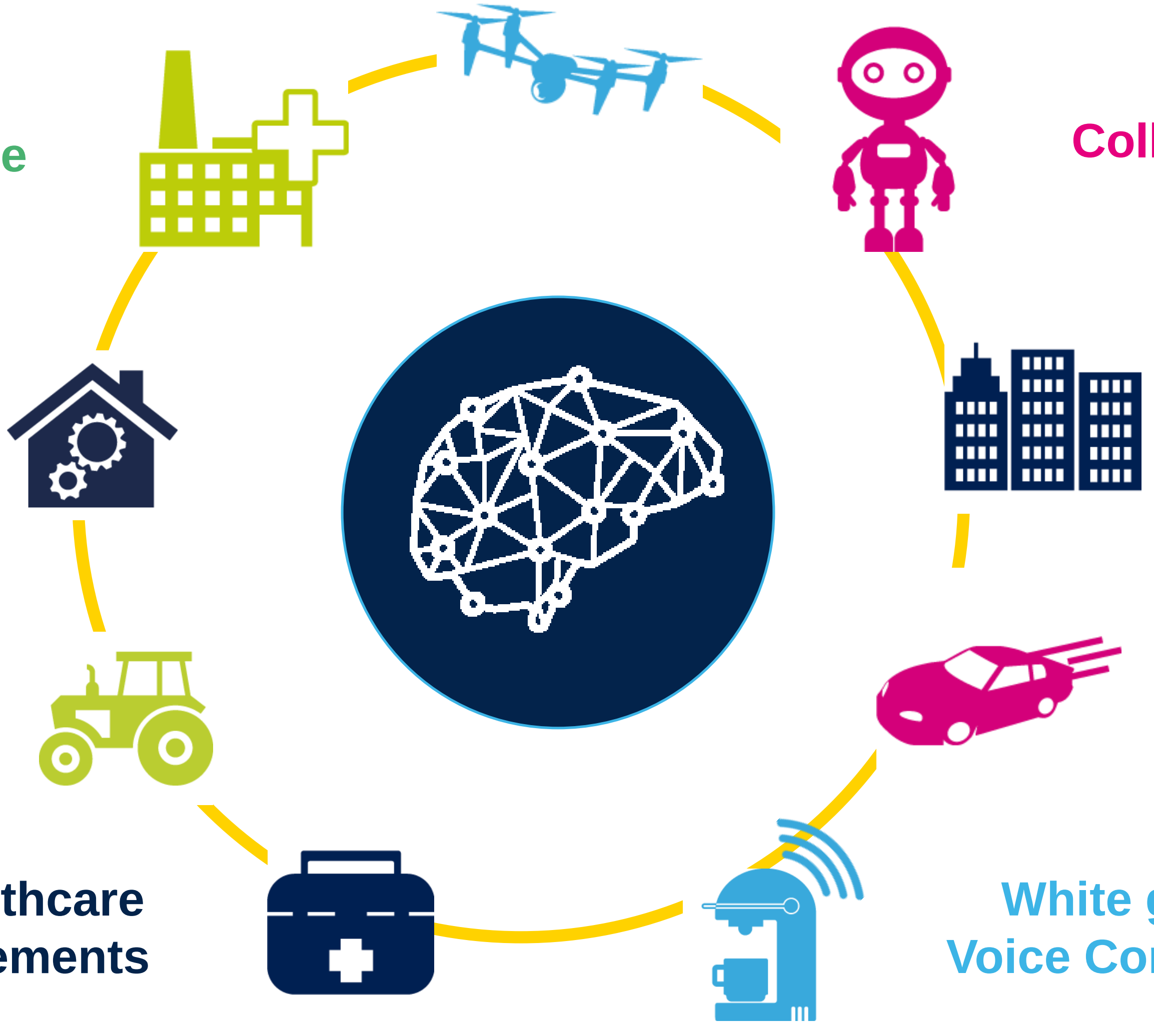


Edge AI means Ubiquitous AI

Drones
Driving & Landing

Smart Industry
Predictive maintenance

Robots
Collision Detection



Smart Home
Event Detection

Smart Buildings
Energy saving

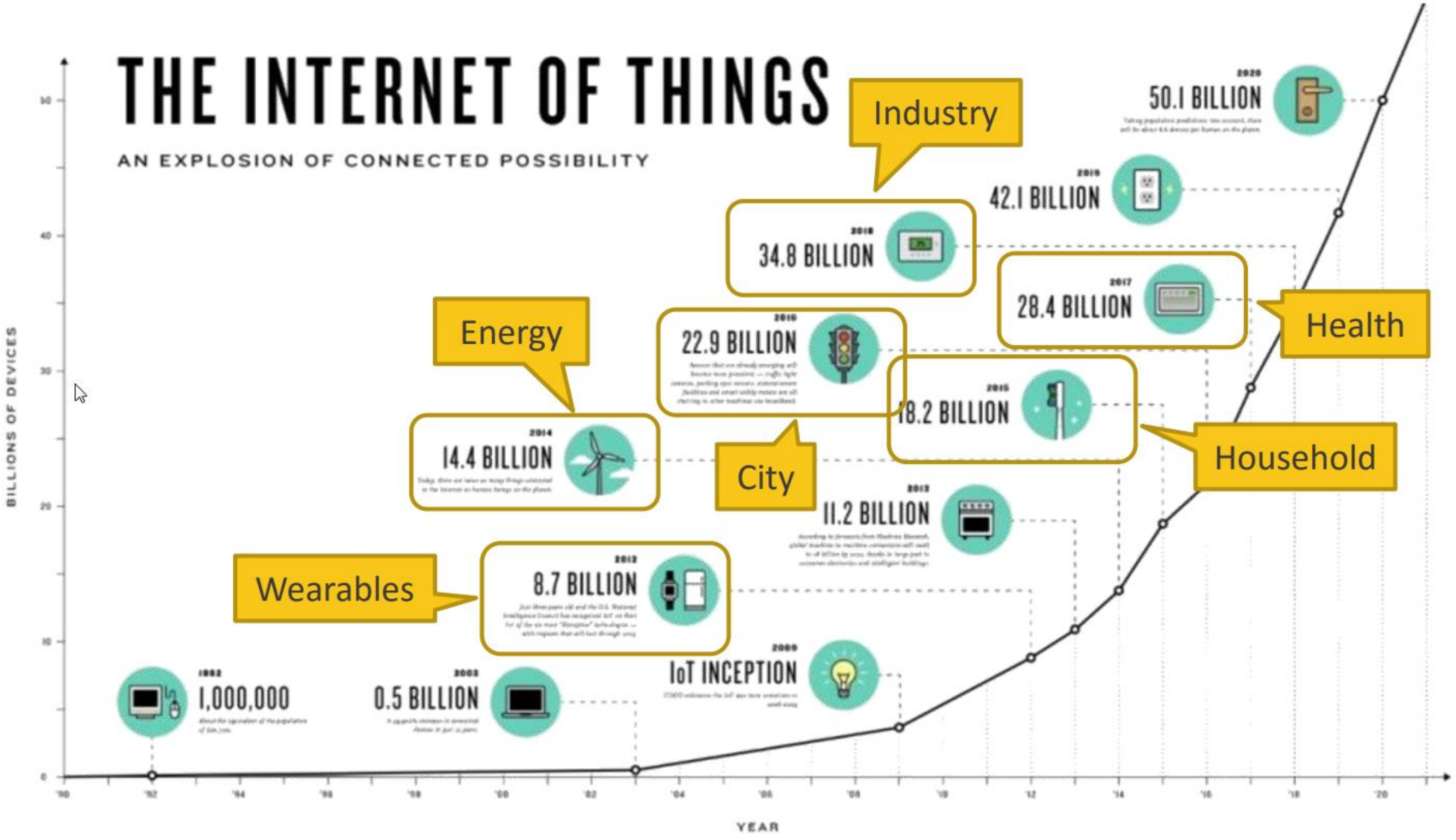
Smart Farming
Optimizing Conditions

Self-driven Cars
Autonomous Driving

Personal Healthcare
Body measurements

White goods
Voice Commands

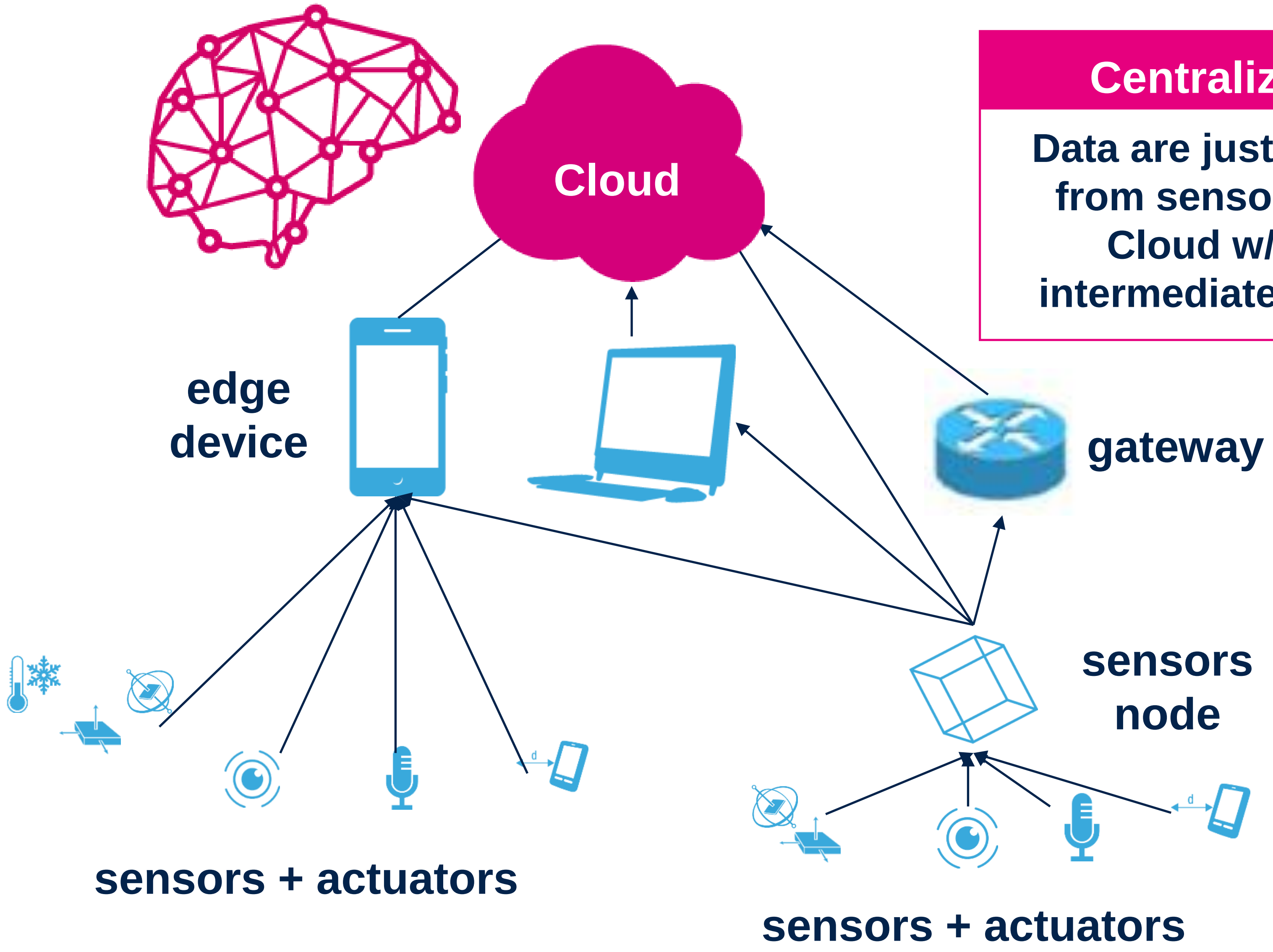
Edge AI is driven by IoT



Source: https://www.slideshare.net/KNIMESlides/advanced-analytics-for-the-internet-of-things-restocking-rental-bike-stations

Centralized approach
AI in the Cloud

- ↓ Responsiveness
- ↑ Bandwidth
- ↓ Privacy
- ↓ Security
- ↓ Energy Saving

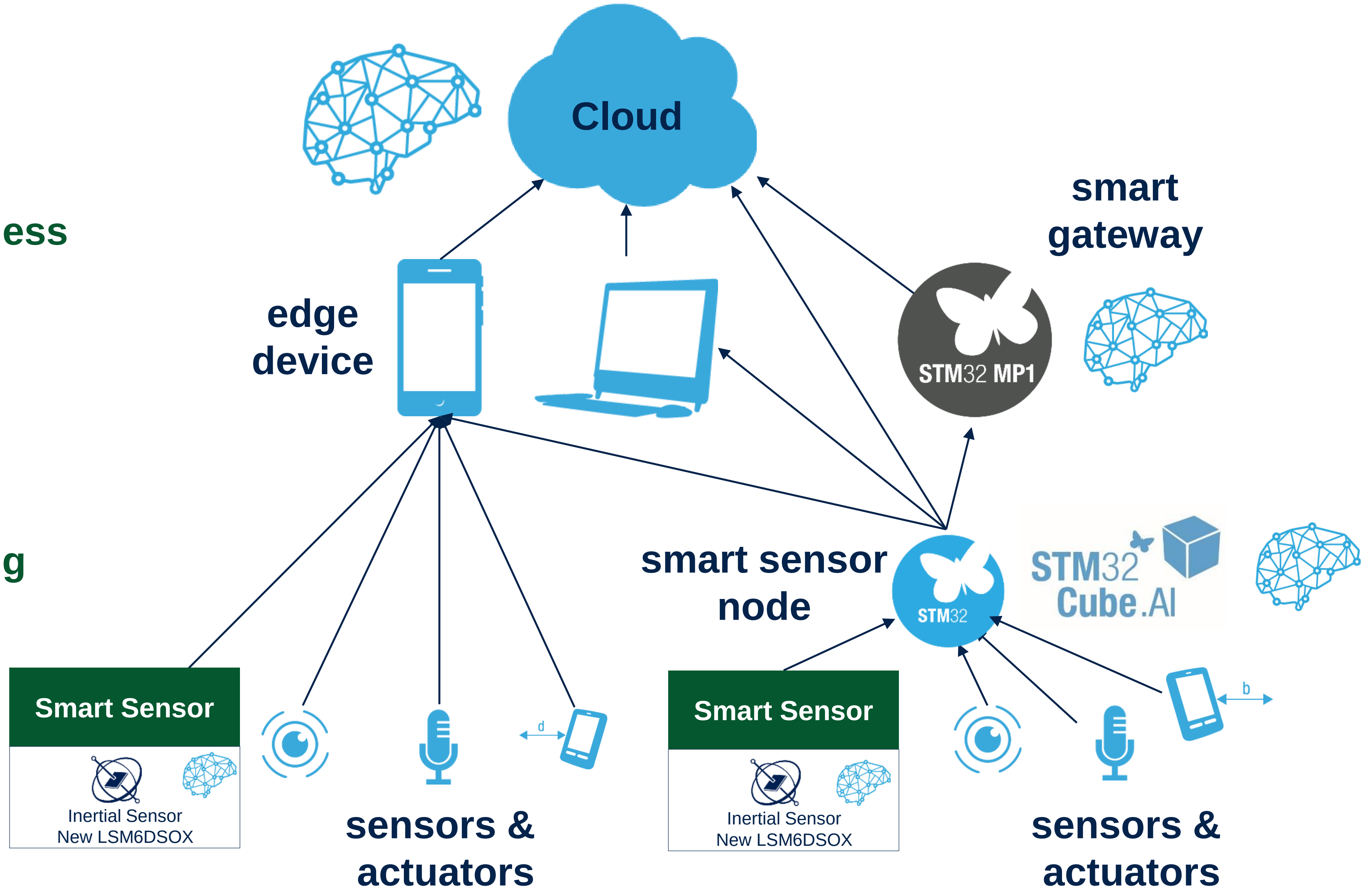


Centralized AI

Data are just provided from sensors to the Cloud w/o any intermediate analysis

Edge AI

- Responsiveness (up)
- Bandwidth (down)
- Privacy (up)
- Security (up)
- Energy Saving (up)



Distributed AI from Edge to Cloud

Edge	Cloud

Endpoint

MCU
- Real time
- Data privacy
- Micro Decisions (AI)
- Micro Actuation

↑ Data ↑ Data ↑ Event

Sensor
- Sensing

Smart sensor
- Sensing
- Preprocessing
- Micro classification

Machine level decisions

→ Sensor Data (Option)
→ Event detection
← Actions

Gateways

MCU / MPU
- Data or Event integration
- Data Privacy
- Meso Decisions (AI)
- Meso Actuation


Industrial

Home

Room or building-level decisions

→ Event detection
← Actions

Cloud

- Fleet monitoring
- Macro Decisions (AI)
- Macro Actuation

Cities or factory-level decisions

Edge AI computing technologies

STM32 Cube.AI

The right HW for the right application

- Simple event detection
- Multi-sensors decisions
- Activity recognition
- Context awareness
- Speech Recognition
- Computer Vision

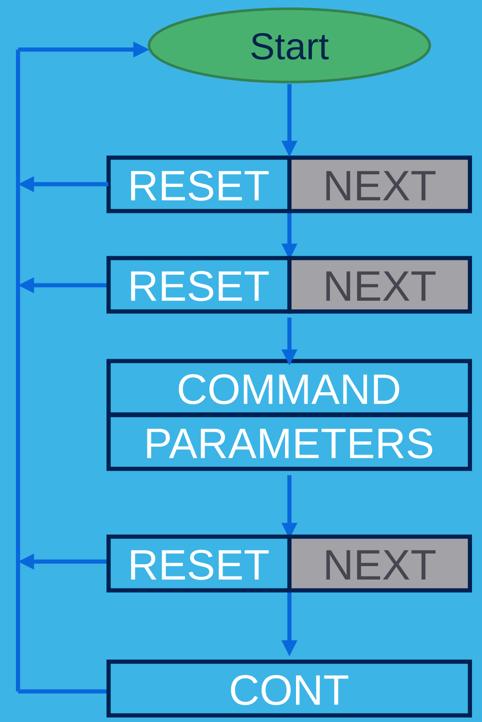
application complexity

HW resources

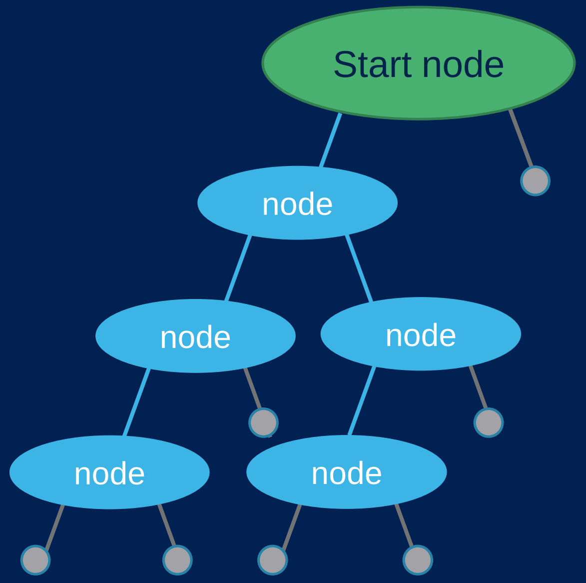
- Smart Sensors — Smart Sensor
- Smart MCU — STM32 Cube.AI
- Smart Sensors & Smart MCU — Smart Sensor & STM32 Cube.AI
- MCU + NPU (Neural Processing Unit) — STM32
- Smart MPU — STM32 MP1

Add intelligence *INSIDE* the sensor

Finite State Machine

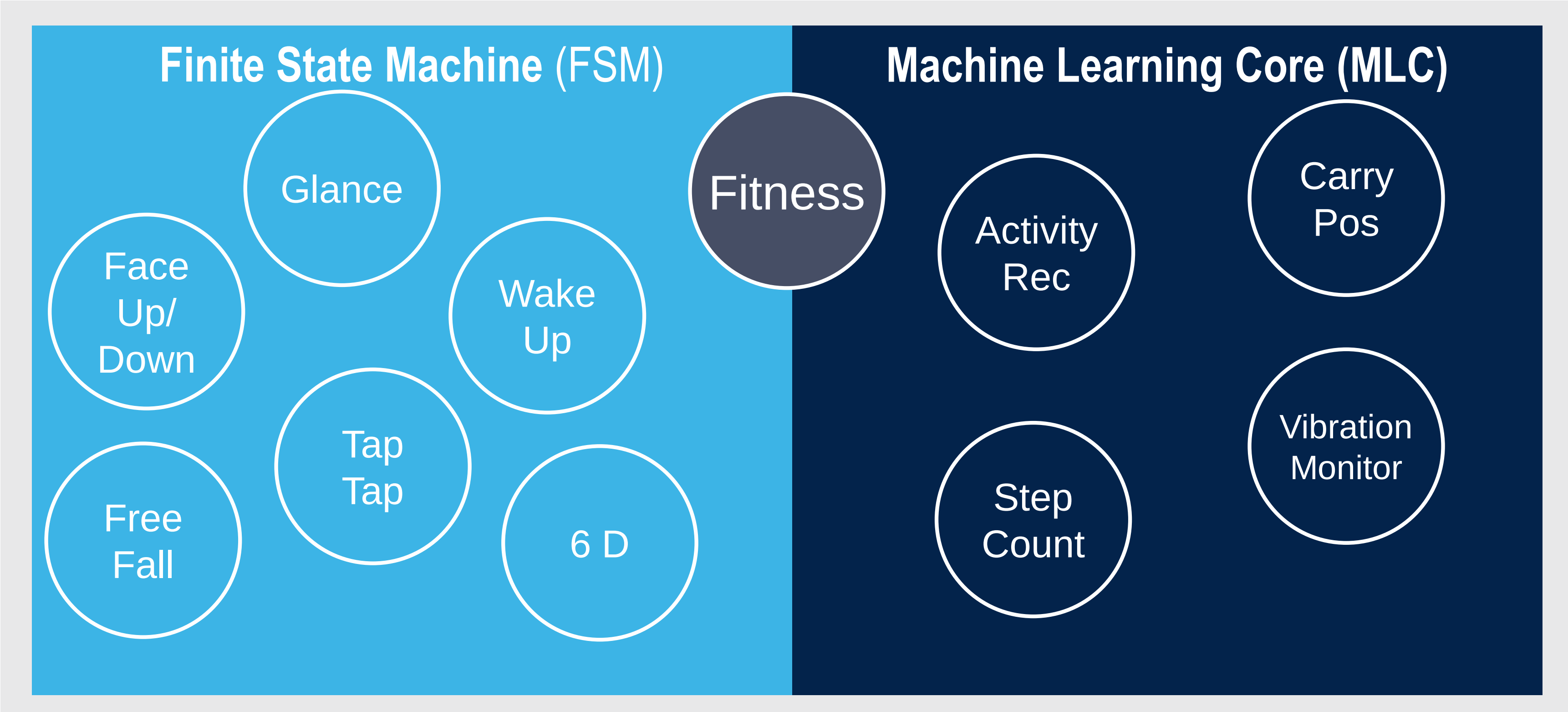


Machine Learning Core

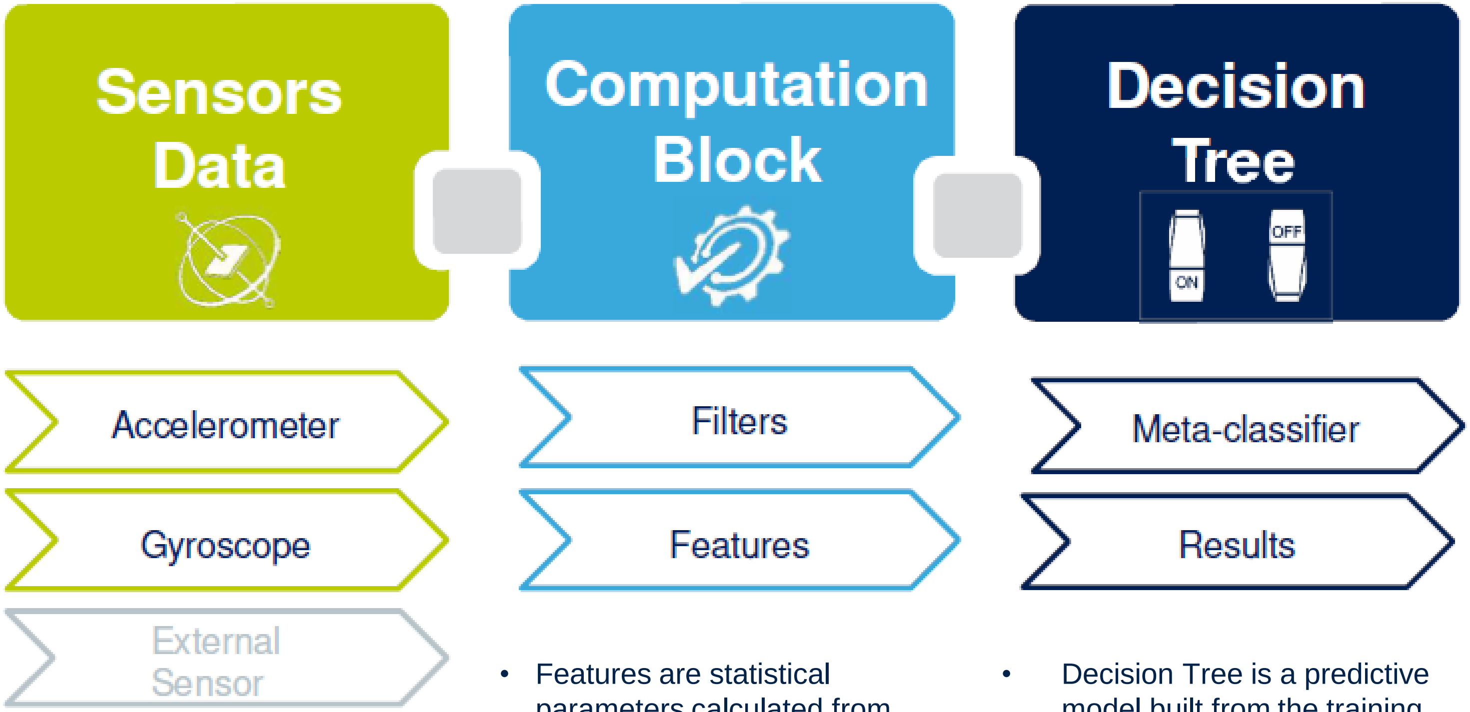


FSM and MLC use cases

Finite State Machine and Machine Learning Core

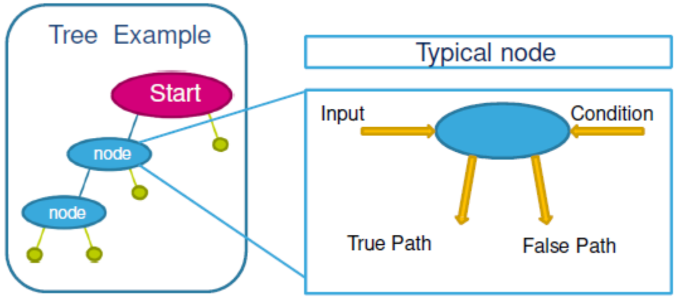


MLC
Block diagram



- Features are statistical parameters calculated from
  - Input data (accX, accY, gyroX, gyroY, …)
  - Filtered data (high-pass on accZ, band-pass on accY
- Examples:
  - Mean
  - Variance
  - Energy
  - …

- Decision Tree is a predictive model built from the training data



Activity recognition use case

10 to 1000 time energy saving by running MLC on Sensor vs. MCU/AP

How it works in 5 simple steps and with an intuitive use case:

1

User defines **Classes** to be recognized

Capture data

2

Collect, clean and label data **Logs** according the classes

Label data

3

Define **Features** that best characterize the identified classes

Build decision tree

4

Machine Learning tools generate program based on **Logs** and **Features**

Embed decision tree

5

Configure the LSM6DSOX and **run** the application

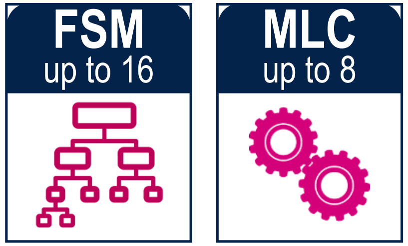
Process new data

WEKA The University of Waikato

Smart Sensor + Smart MCU

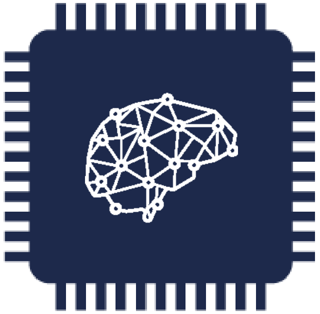
Smart Sensor
with Machine Learning Core

Smart STM32
second level of AI processing



Inertial Sensor
New LSM6DSOX

Raw Data →
Event Decision →
← FSM and MLC Re-configuration



Deep Learning
Neural Networks
Machine Learning

- Best ultra-low-power sensing at high performance:
  - 550µA (gyroscope and accelerometer)
    ➔ 200µA less than closest competitor
  - 20~40µA (Accelerometer only for HAR)
- Efficient Finite State Machines: 3µA
- Configurable Machine Learning Core: 1~15µA

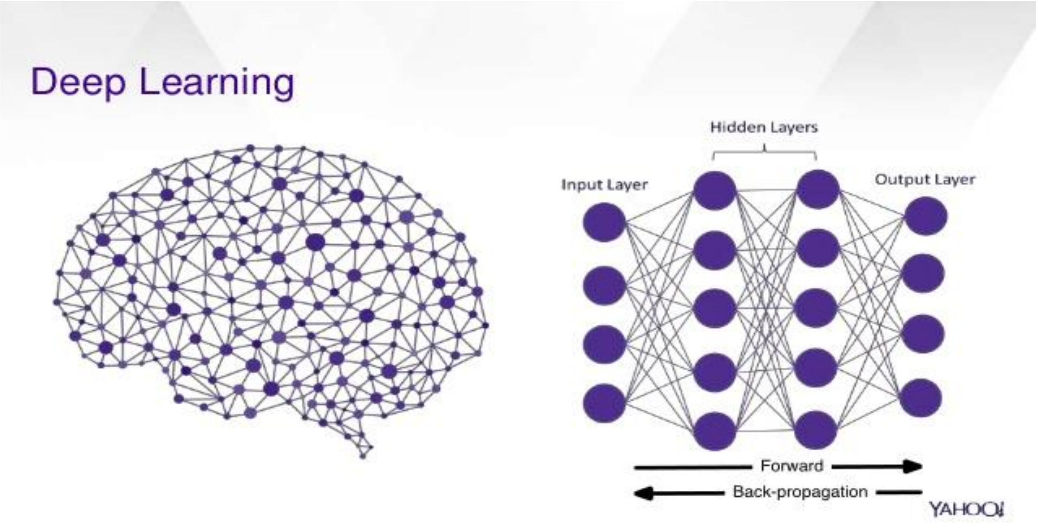
- More advanced and complex NNs
- Decisions on multiple sensors
- NN input can be sensor data and/or sensor Machine Learning decisions
- Multiple Neural Networks support
- Actuation & communication

STM32 Cube.AI

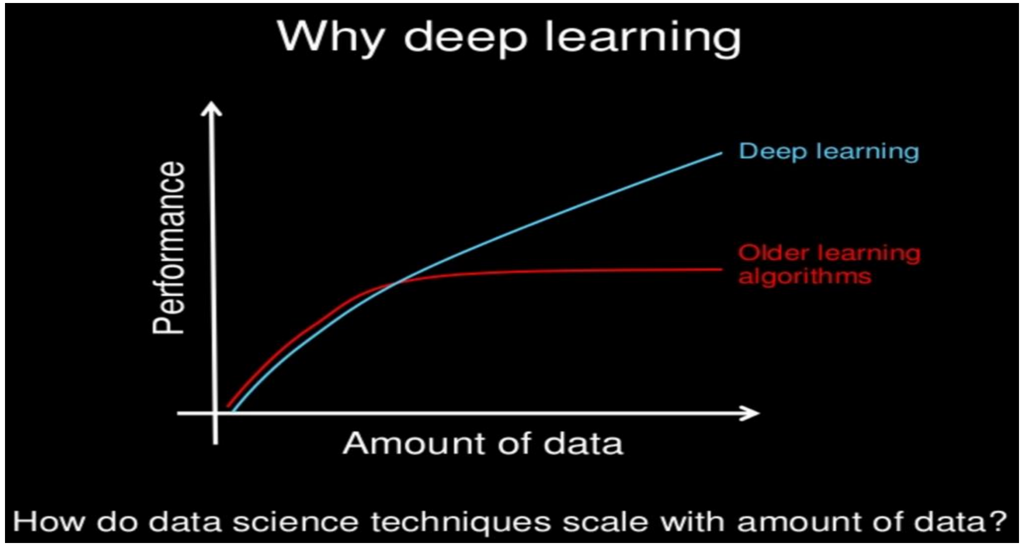
life.augmented

Deep Learning (DL)

- Deep Learning is ML using Neural Networks
  - Inspired by biological neural networks
  - Deep because of having many intermediate learning steps
  - Lots and lots of data is required



Advantages	Disadvantages
Autonomous Learning of data patterns & relationships	Large Datasets
High Accuracy	High Computational Requirements
Easy Improvement & Fine Tuning	Weak theoretical explanation
Adaptive Solutions	Black box (for most people)



FEDERAZIONE NAZIONALE IMPRESE ELETTROTECNICHE ED ELETTRONICHE

CONFINDUSTRIA

Why Deep Learning is so important

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

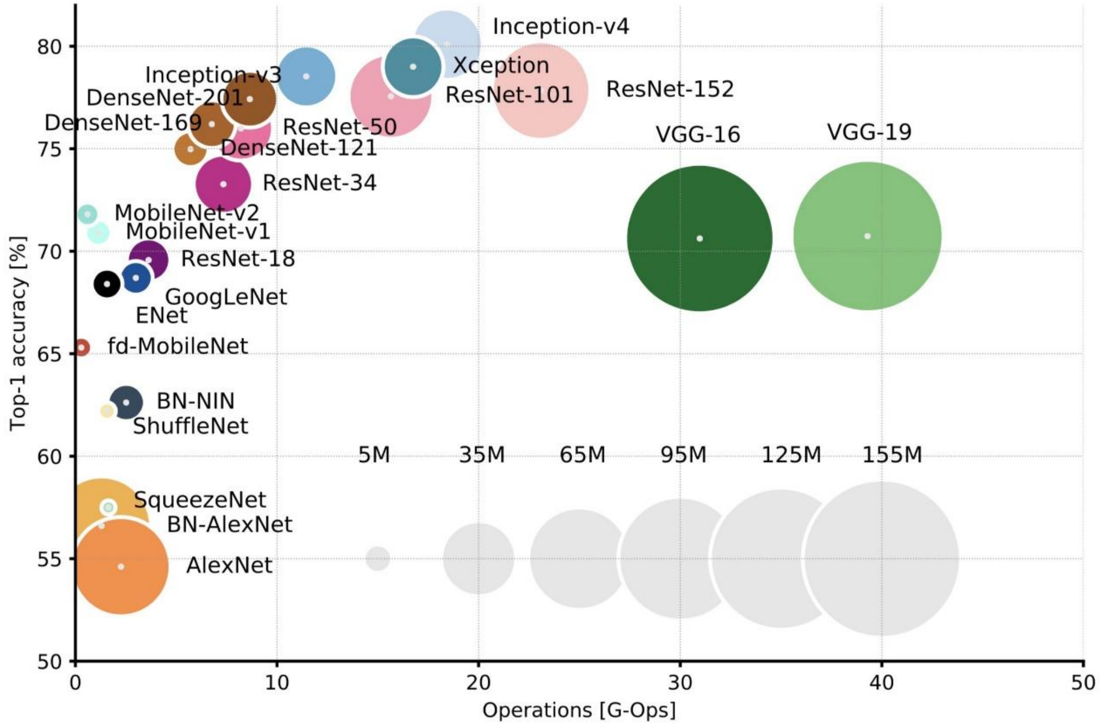
Problem	Dataset	Best Accuracy w/o CNN	Best Accuracy with CNN	Diff
Object classification	ILSVRC	**73.8%**	**95.1%**	**+21.3%**
Scene classification	SUN	**37.5%**	**56%**	**+18.5%**
Object detection	VOC 2007	**34.3%**	**60.9%**	**+26.6%**
Fine-grained class	200Birds	**61.8%**	**75.7%**	**+13.9%**
Attribute 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

life.augmented

Neural Networks on MCU
A real challenge

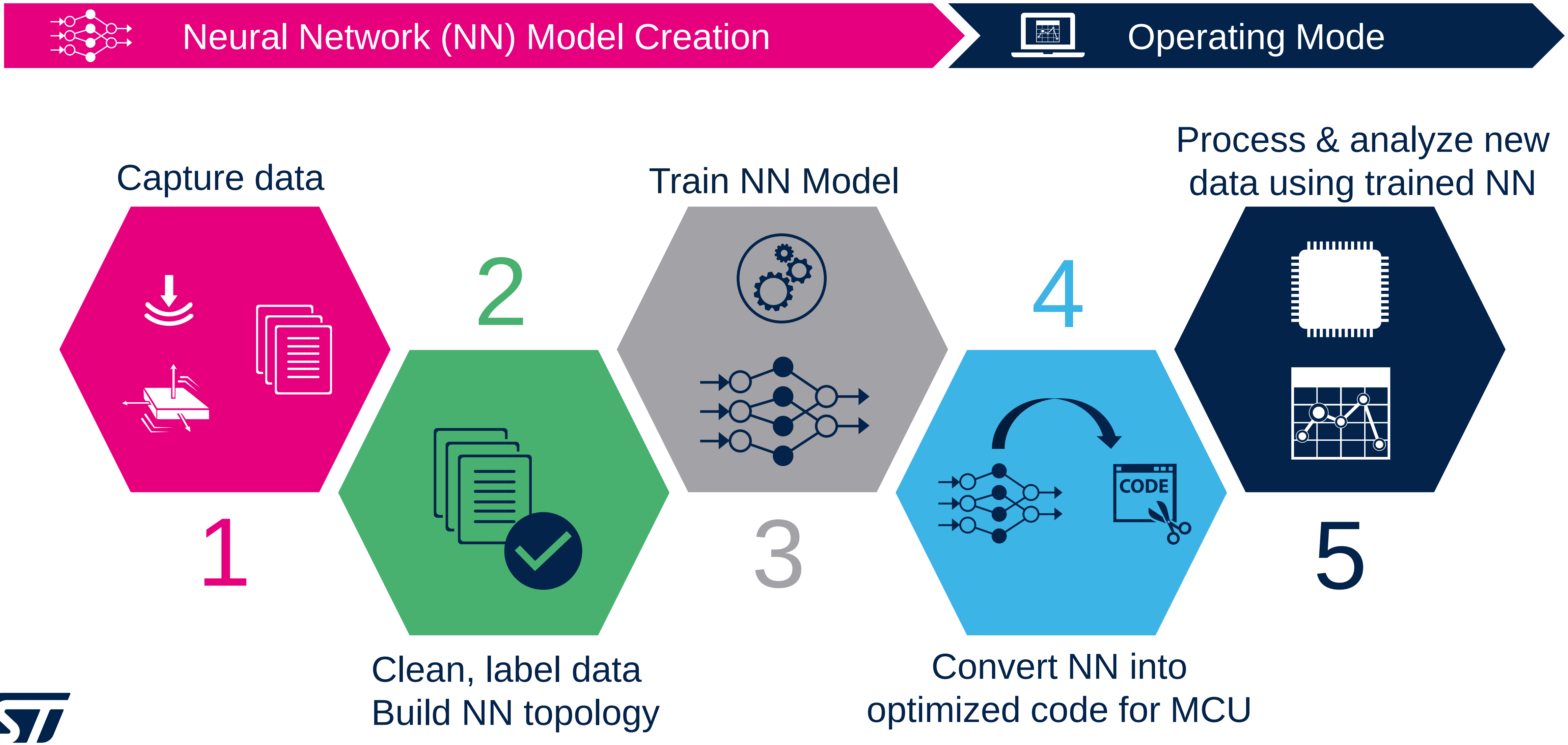
- A MCU usually has:
  - Limited Non Volatile Memory (< 2 MB) → take care of # weights
  - Limited SRAM (< 1 MB) → take care of data size & activations
  - Low frequency (< 500 MHz): take care of # ops for time constrained applications
- The challenging task is to fit these constraints with good quality results !
- An example: image classification task
  - Gops should be very low
  - #weights should be very low
  - MobileNet-v2 is still high in terms of #weights and cannot run in real time, i.e., @30fps
  - We implemented FD-MobileNet (on 18 foods) on STM32H7
    - Memory footprint: 205 KB SRAM, 191 KB Flash !
    - Inference time 150 ms



Source: https://medium.com/@culurciello/analysis-of-deep-neural-networks-dcf398e71aae

life.augmented

The key steps behind Neural Networks



life.augmented

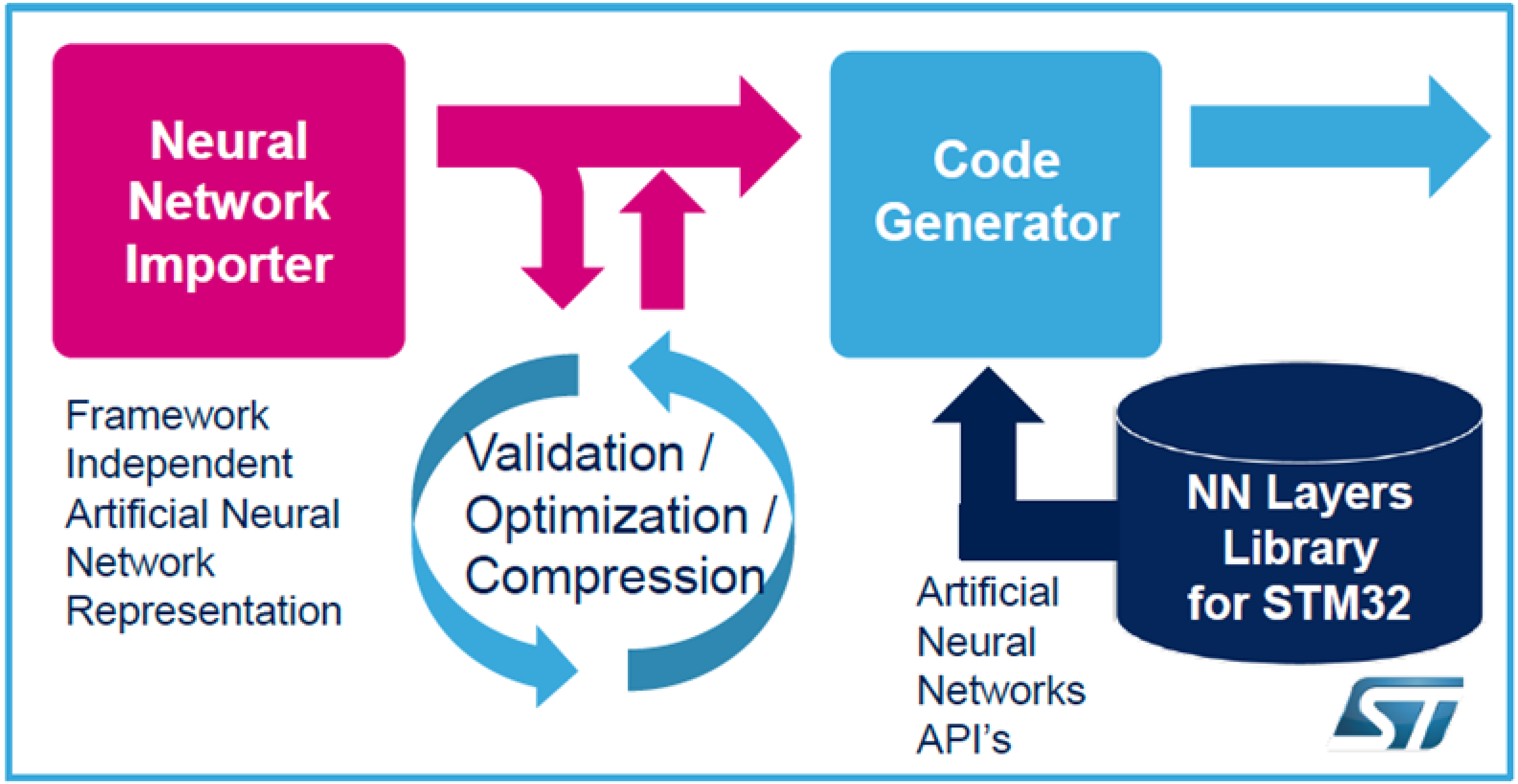
Neural Networks on MCU
The ST's way

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

Deep Learning Framework dependent

Caffe K TensorFlow Lite ONNX MATLAB PyTorch

And more



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

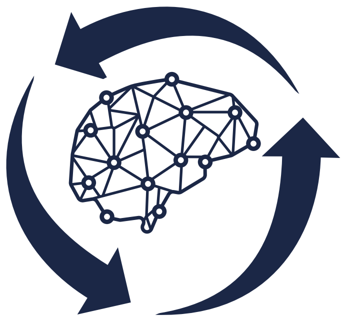
Optimize after initial deployment

Continuous learning &
Over-the-air update

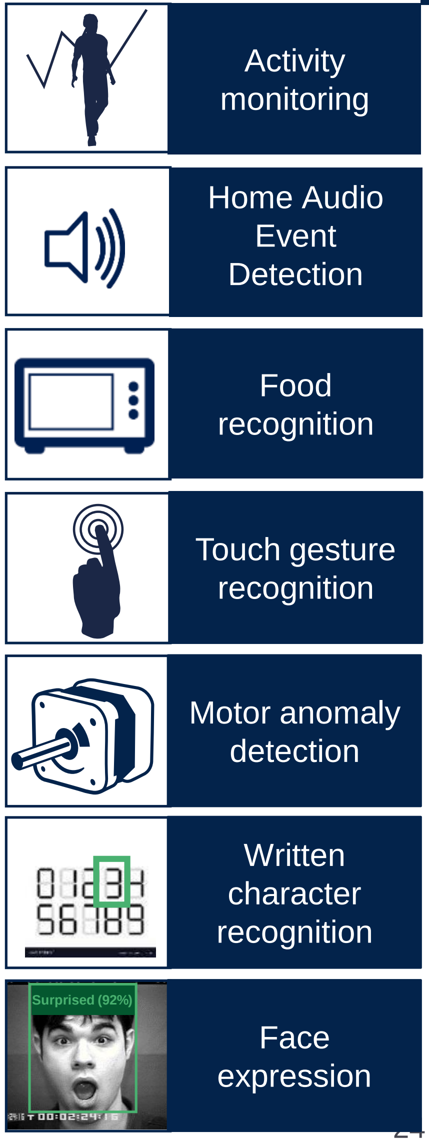
STM32
Cube.AI

Continuous learning in local

Refine Neural Network to local conditions

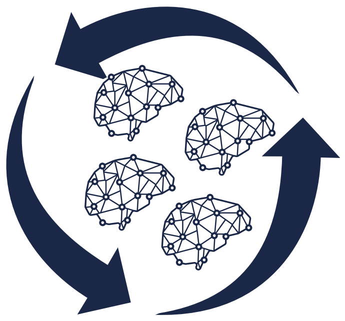


- Learning at the edge to specialize NN to local sensor input: personalize to a particular user, home/factory environment…
- Supervised learning: user inputs feedback to re-enforce classification or regression output
- Unsupervised learning: classification with high probability is used for re-enforcement, output labels are guessed from inferences in same temporal window

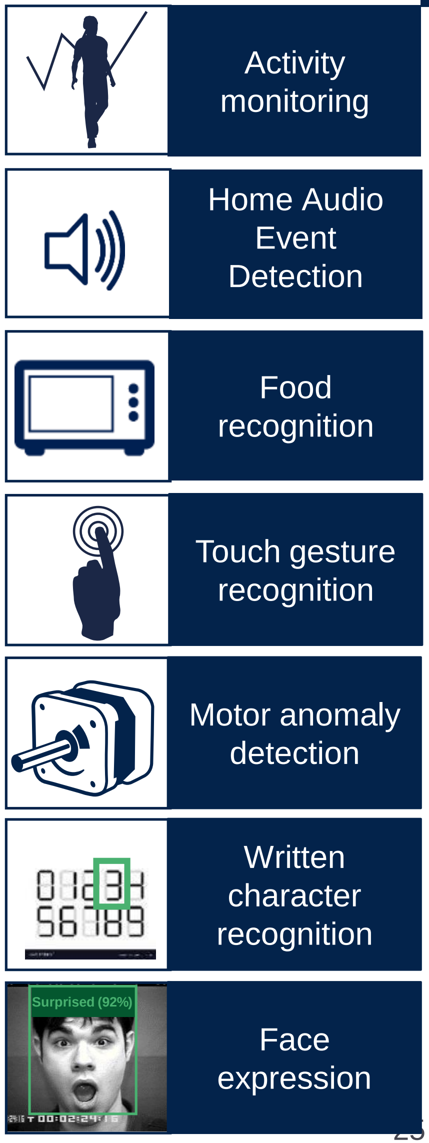


Federated learning

Refine Neural Network from learning of all local conditions

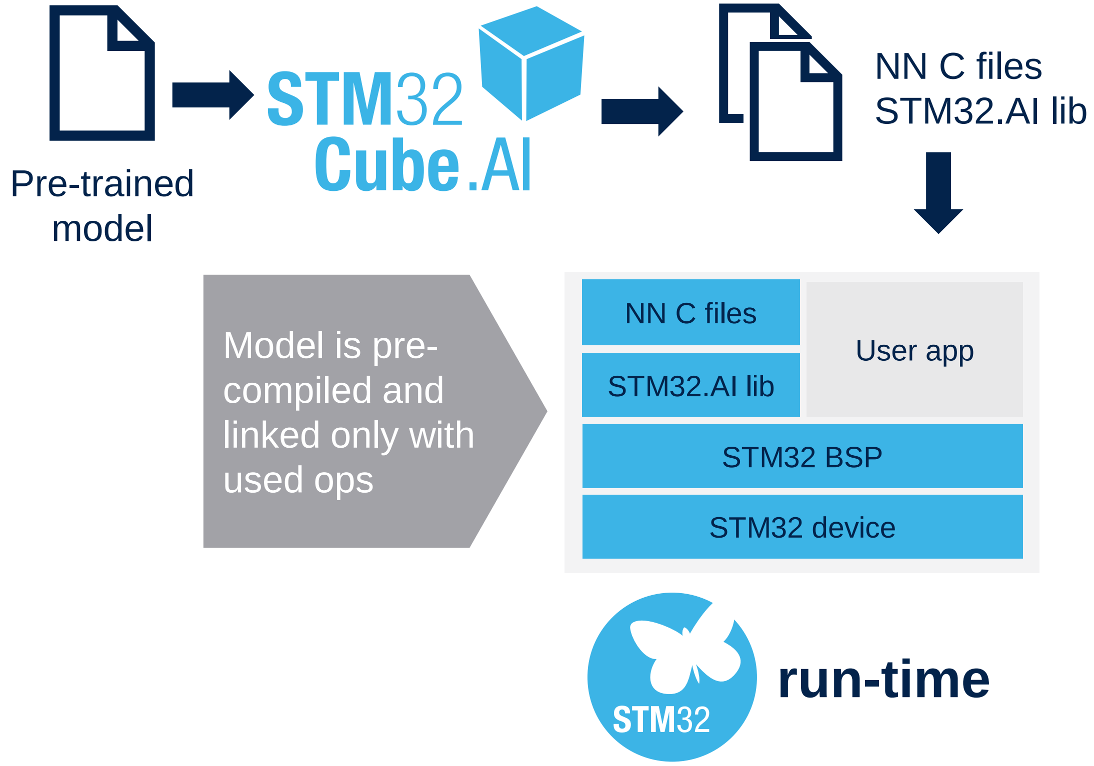


- Learn at the edge from different conditions, send learning cumulated over batch of data to the Cloud regularly
- Cloud cumulates learnings from all sensors: builds learning from diverse dataset without full raw data transfer over connectivity
- Generalization is ensured by federated learning of many diverse examples



life.augmented

NN update over the air



Update weights

Using STM32Cube.AI v5.0.0, model weights are stored in specific table of NN C files

- Fine tune weights (when learnt in the Cloud), independently of the FW

Update topology

From STM32Cube.AI v6.0.0, Neural Networks files and AI library are stored in specific sections. Topology can be updated without full FW update

- Add a new class
- Add an extra layer
- Add new operators

Application markets

Condition monitoring and predictive maintenance

STM32 Cube.AI

life.augmented

Predictive Maintenance

A Smart Industry hot topic

Preventive Maintenance

Scheduled maintenance tasks based on a time schedule – don't care of the actual status of the equipment

Advantages

- Simple to plan

Drawbacks

- Maintenance may happen too late (or too early)
- Maintenance may not be necessary

Condition Based Maintenance

Maintenance is based on the estimated conditions of the machine, typically monitored through inspection or sensors

Advantages

- Maintenance only takes place when necessary

Drawbacks

- Maintenance only after machine begins to show signs of failure

Predictive Maintenance

Maintenance actions predicted in advance based on monitoring combined with a dynamic predictive model for failure analysis

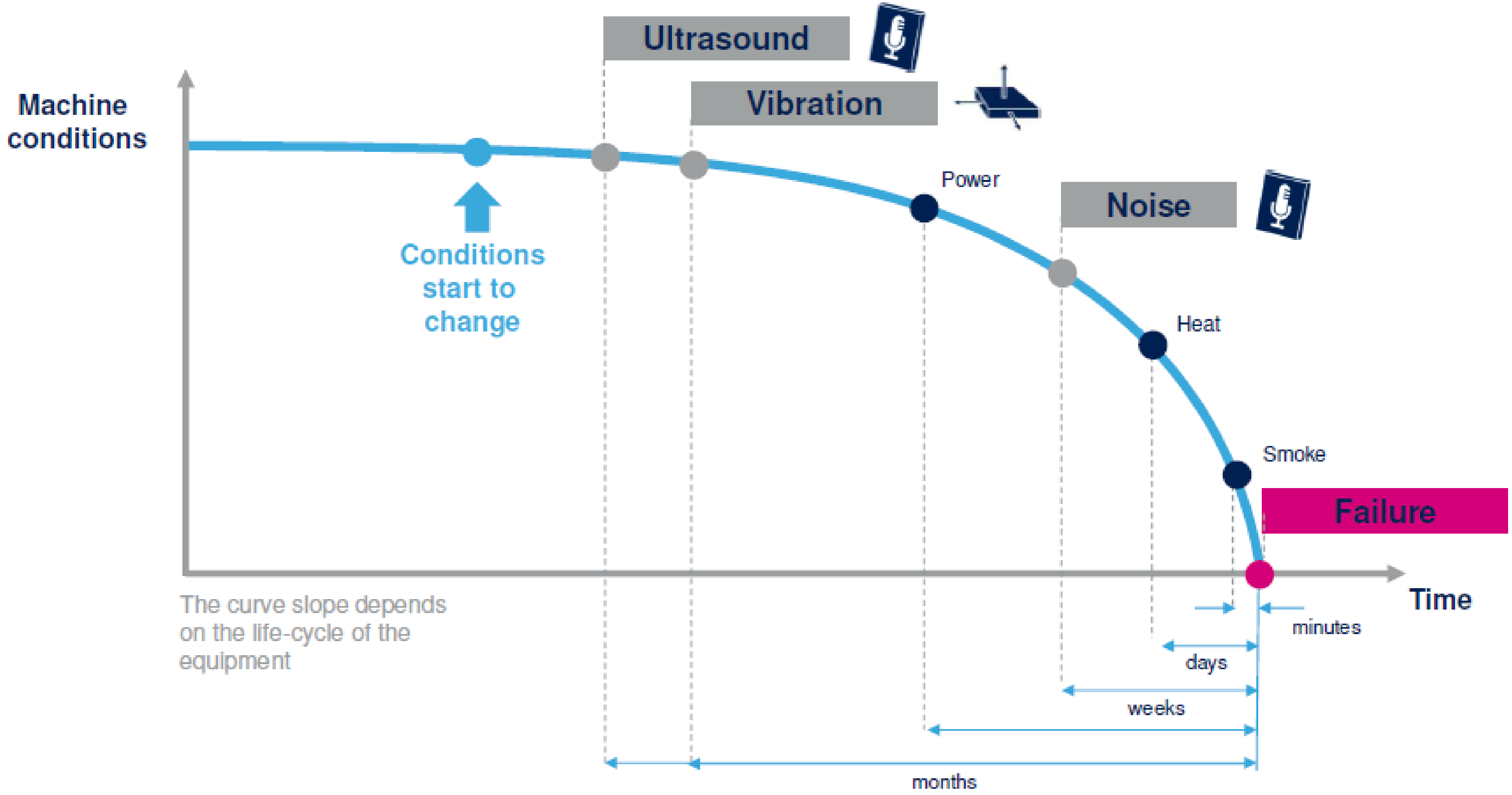
Advantages

- Maintenance optimized for machine life and production efficiency

Drawbacks

- Requires complex overall system

Predictive Maintenance
Microphones and inertial sensors



life.augmented

Predictive Maintenance
Benefits

Reduced lost production time
Maintenance on the production line only when needed and at the optimal time

Longer machine lifetime/lower effective cost
Replacing the minimum amount of parts before failure causes damage to others

Faster and more efficient repair
Optimized workers interventions and minimum labor for parts replacement

Increased safety
Prevents failures that could be dangerous for workers before they happen

Aggregated figures from different sources (Accenture, McKinsey, ST)

life.augmented

Condition monitoring & predictive maintenance applications

New equipment (greenfield)

Integration possible with power supply and existing sensors

In-field maintenance (retrofit)

Battery-powered simplifies installation

Industrial

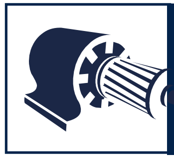
- Manufacturing and Process Automation
- Power and energy
- Home appliances and Smart Building Automation

Automotive and Transportation

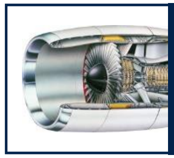
- Vehicles
- Railways
- Infrastructures

Home Appliance

- Lighting
- Washing Machine
- Vacuum cleaner, Air conditioning



Industrial motor vibration monitoring



Engine TTF vibration monitoring

Motor current monitoring

Crankshaft rotation monitoring

Pipe flow monitoring

Temperature, humidity, gas monitoring

acoustic monitoring

life.augmented

Application markets

Sound event and context awareness

STM32 Cube.AI

Sound-based applications

Complex audio front-end not always required for sound detection

Sound event classification

- Home alarms
- Glass break
- Machine anomaly

Sound event detection

- Vehicle count based on Doppler
- Sound direction detection with multiple MICs

Sound wake-up

- Sound detection to wake up more advanced algorithms

Context awareness

- Indoor/outdoor environment
- Vehicular environment
- Factory activity
- Human presence detection

Vehicle counting

Intrusion detection

Engine break detection

Audio scene classification

Building occupancy

Crowd panic detection

Wake-up / orient camera based on sound

life.augmented

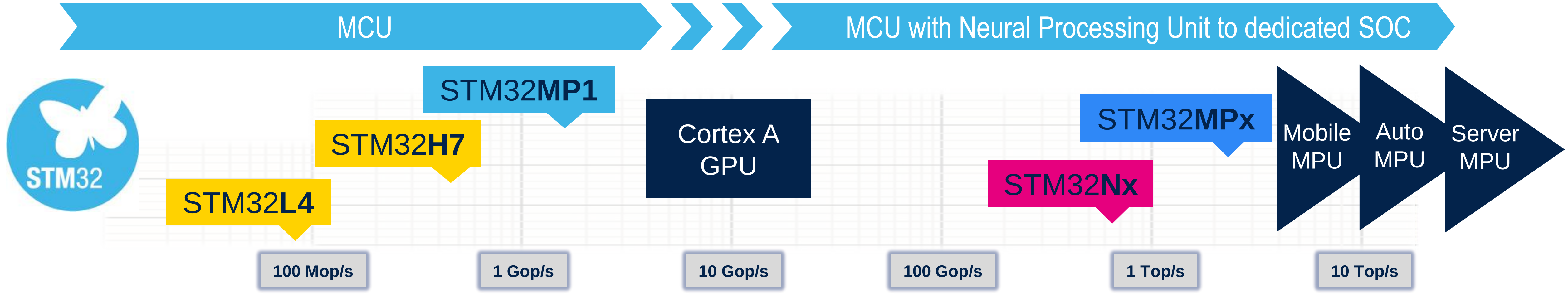
Application markets

Computer Vision

STM32 Cube.AI

Computer Vision
Processing requirements

Low	Medium	High
▪ Static low-res images ▪ Known object position ▪ Good light conditions	▪ Low frame rate / QVGA images ▪ Open environment ▪ Medium light conditions	▪ High rate Video ▪ High resolution ▪ Adaptive light conditions



life.augmented

Vision based applications

IoT object often rely on batteries, as main power is mostly not available

Replacing batteries is not sustainable in IoT, low power is a MUST

Computer Vision for voice interfaces

- Gaze detection
- Gesture recognition
- Person direction detection

Predictive maintenance

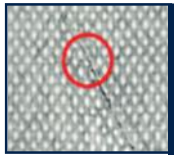
- Smoke detection
- Product defect detection
- Spilled liquid detection

Visual wake-up

- Person detection to wake up more advanced algorithms

Classification and recognition

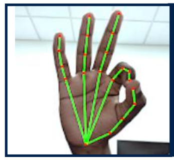
- Recognize pests, weeds, disease in fields
- Characters and digits recognition
- Ingested camera imaging
- Simple image classification
- Texture, fabric recognition
- Visual biometrics


Defect detection

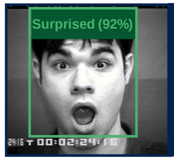

Crop disease detection & classification


Thermal camera


Meter aftermarket


Gesture recognition


Person detection



Face expression

life.augmented

Servitization

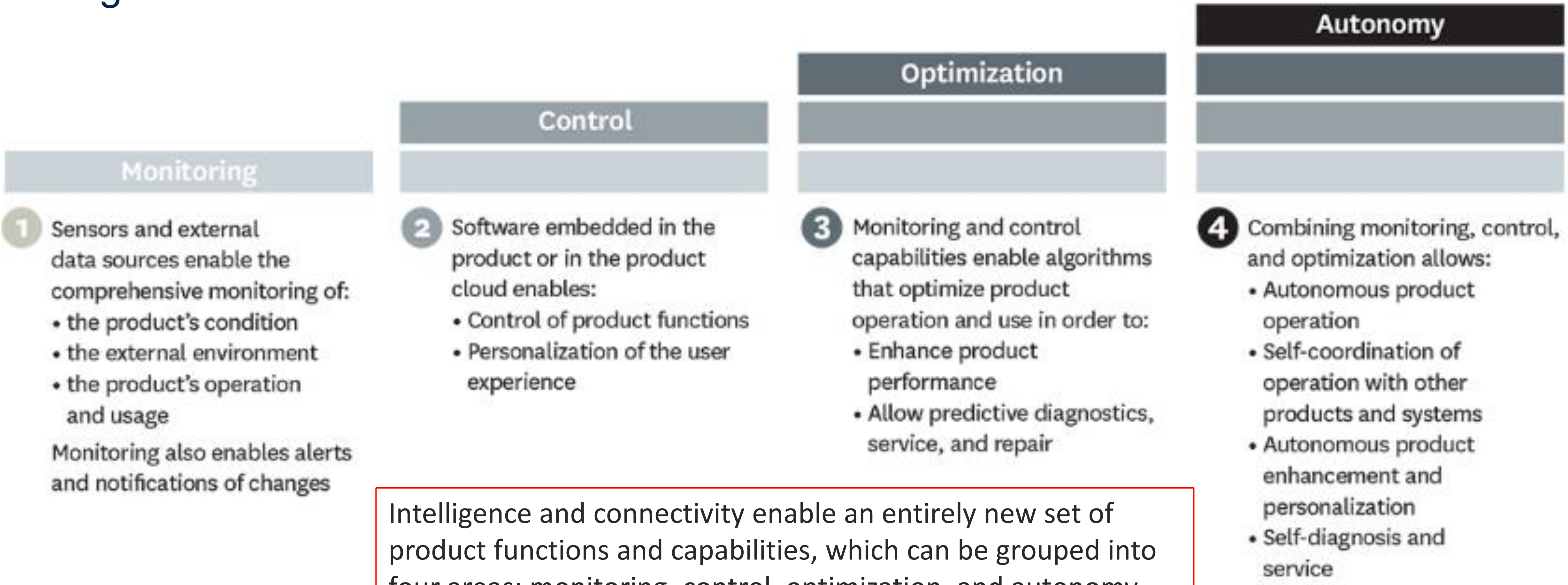
Servitization

- **Servitization** means the joint offering of products and services, or better “product-service systems”
- Companies can better create value by moving from the sale of products to the sale of systems consisting of products and services
- **Servitization** enables companies to differentiate their offer from that of their competitors and at the same time increase customer loyalty over time
- Some examples:
  - IBM selling Cloud services (like *Watson* and *Quantum Computing*), instead of selling hardware (mainframes)
  - Rolls-Royce selling "power-by-the-hour“, instead of selling aircraft engines

Vandermerwe, S., & Rada, J (1988) "Servitization of Business: Adding Value by Adding Services", European Management Journal, 6(4), 314–324

Edge AI and servitization

- Edge AI is a clear enabler of this “service transformation”



Intelligence and connectivity enable an entirely new set of product functions and capabilities, which can be grouped into four areas: monitoring, control, optimization, and autonomy.

Source: M.E. Porter and J.E. Heppelmann; https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

life.augmented

Thank you

life.augmented