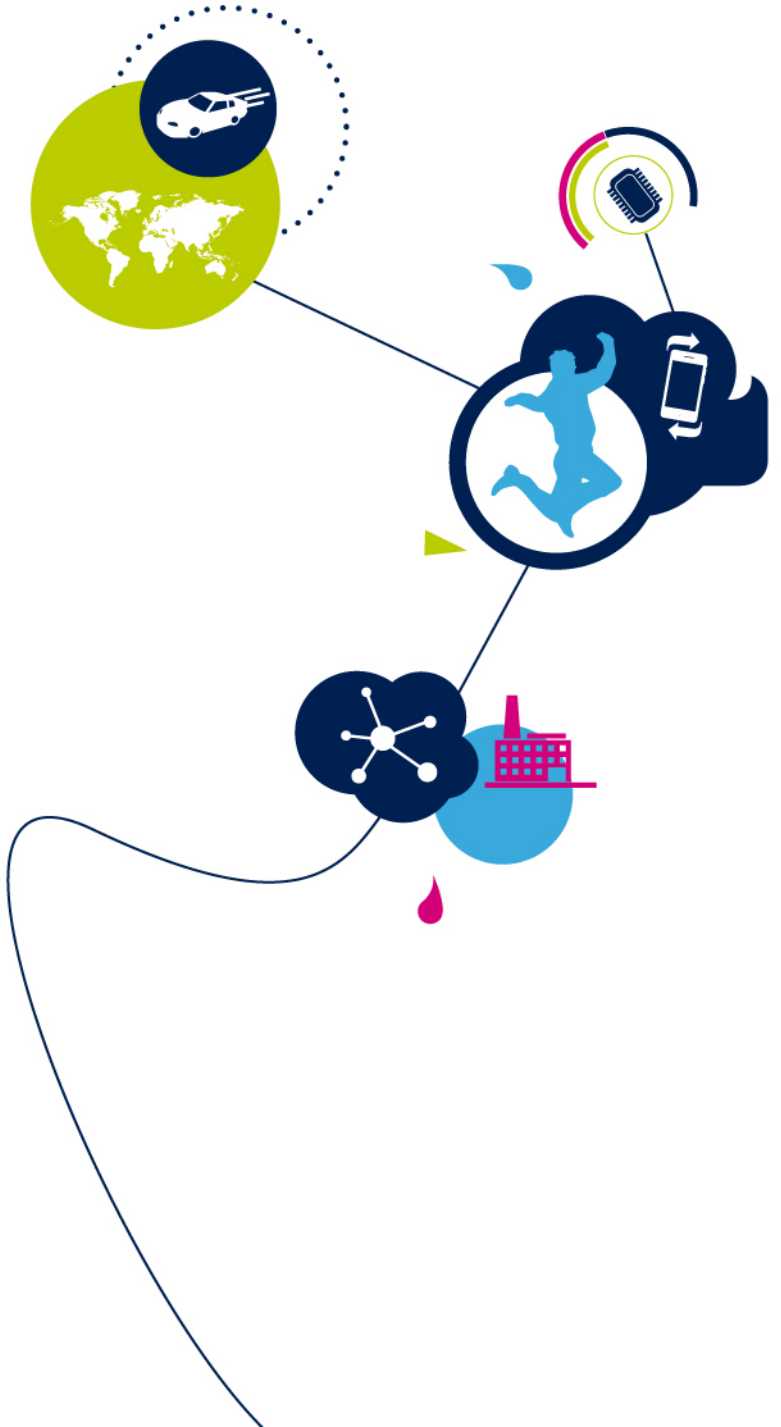
Artificial Neural Networks on Resource-Constrained Devices

Markus Mayr
Product Marketing Manager, MCU



life.augmented

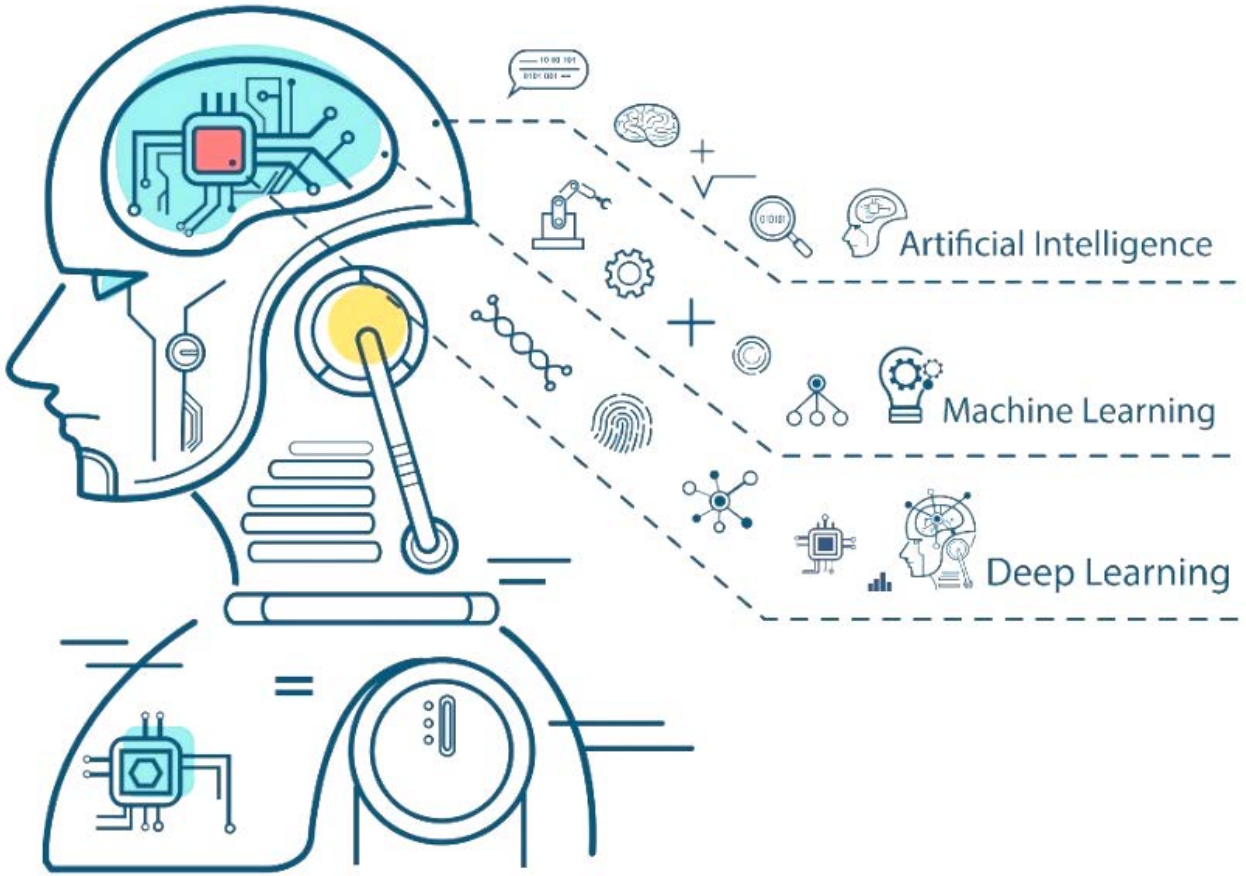
Artificial Intelligence (AI)

- AI allows machines to mimic cognitive capabilities of humans. Examples:
  - 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)

- **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

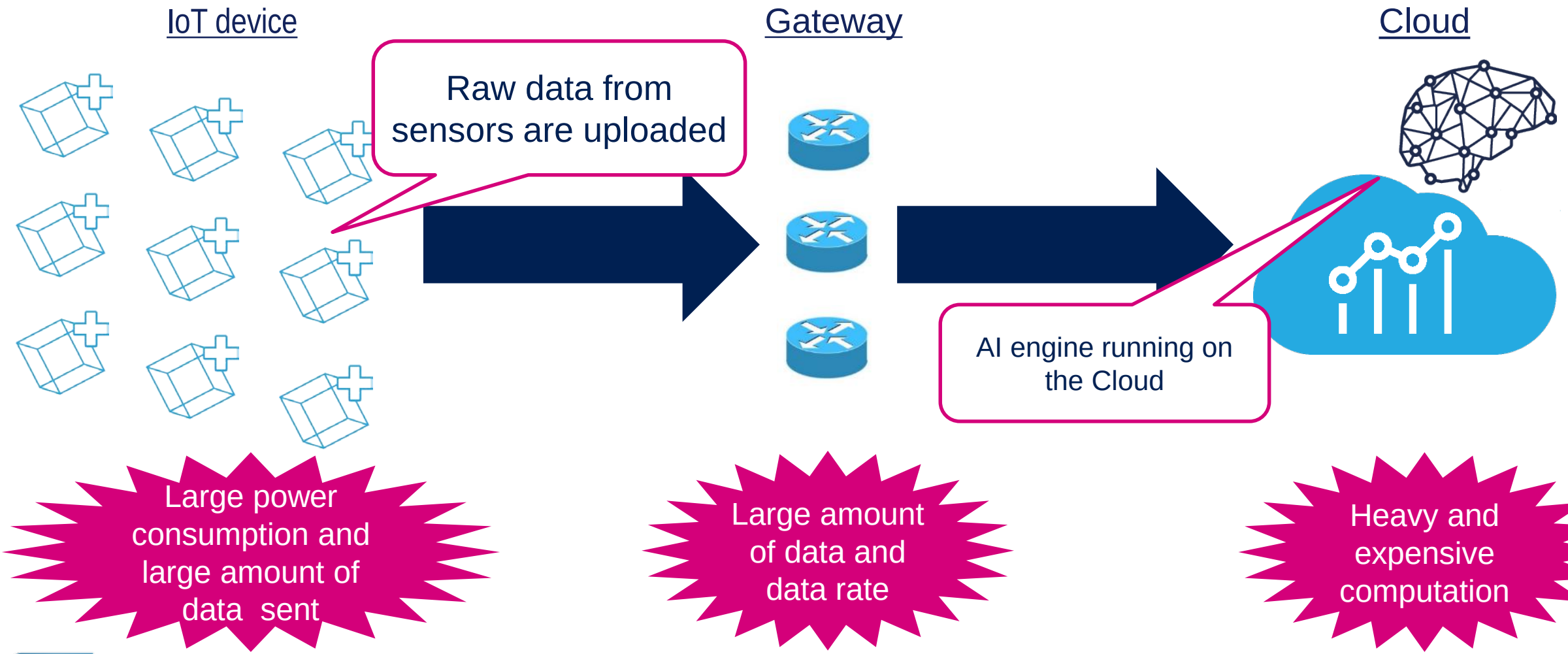


Any technique which enables a computer to mimic human behavior

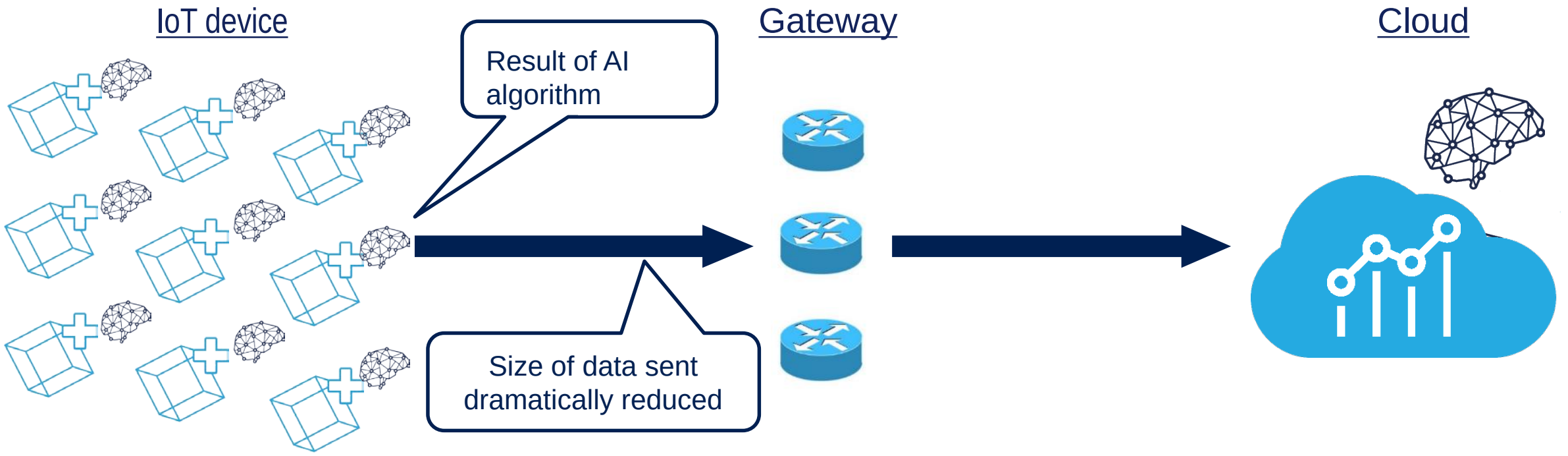
Subset of AI, algorithms and methodologies to improve over-time through learning from data

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

AI Cloud computing



AI Edge computing



Small power consumption and small amount of data

Small amount of data and low data rate

Medium computation

life.augmented

Distributed AI: Holistic approach

High Bandwidth

High centralized computing power

Potentially high latency

Reduced bandwidth

Lower centralized computing power

Real time response

Preserving Privacy

life.augmented

AI Edge Computing on MCU

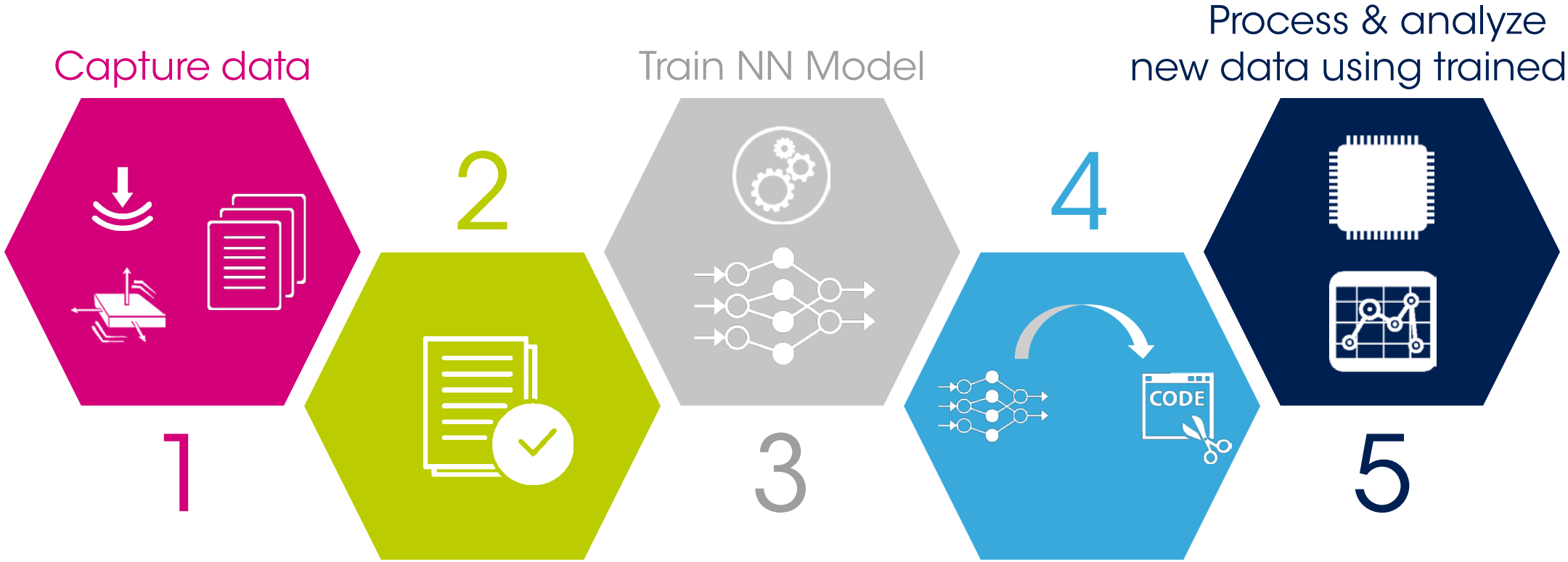
- More efficient end-to-end solutions are possible by switching from a centralized to a distributed system
- The objective of the AI Edge computing is:
  - To reduce the amount of data sent to the cloud
  - To decrease latencies due to network delays & outages
  - To improve system response time
  - Sensitive data is not sent to the cloud for privacy/security
- AI and deep learning allow low power solutions close to the sensor enabling true edge computing

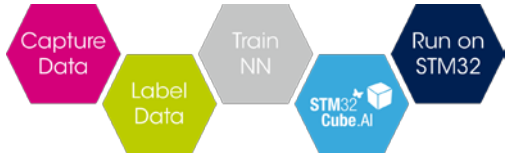
AI on MCUs – How does it work?

- Most MCUs today do not have the memory and processing power to run complex learning algorithms and create Deep Neural Networks
- However, MCUs can run the DNNs themselves, provided they are optimized for MCUs
- Dedicated tools such as the STM32Cube.AI can optimize DNNs for the use with MCUs such as the STM32 family:
  - A pre-trained NN Model (Caffe, Keras, Lasagne, TensorFlow, etc.) and convert it into MCU code
  - The code is optimized to adapt it to the memory, processing and power capabilities of an MCU
  - The generated code can be 10x smaller than the original with negligible loss of accuracy
  - The functionality of the adapted DCNN can be checked and adjusted

The Key Steps Behind Neural Networks

Neural Network (NN) Model Creation | Operating Mode





ST Toolbox for Neural Networks

life.augmented

Capture data

Clean, label Data
Build NN topology

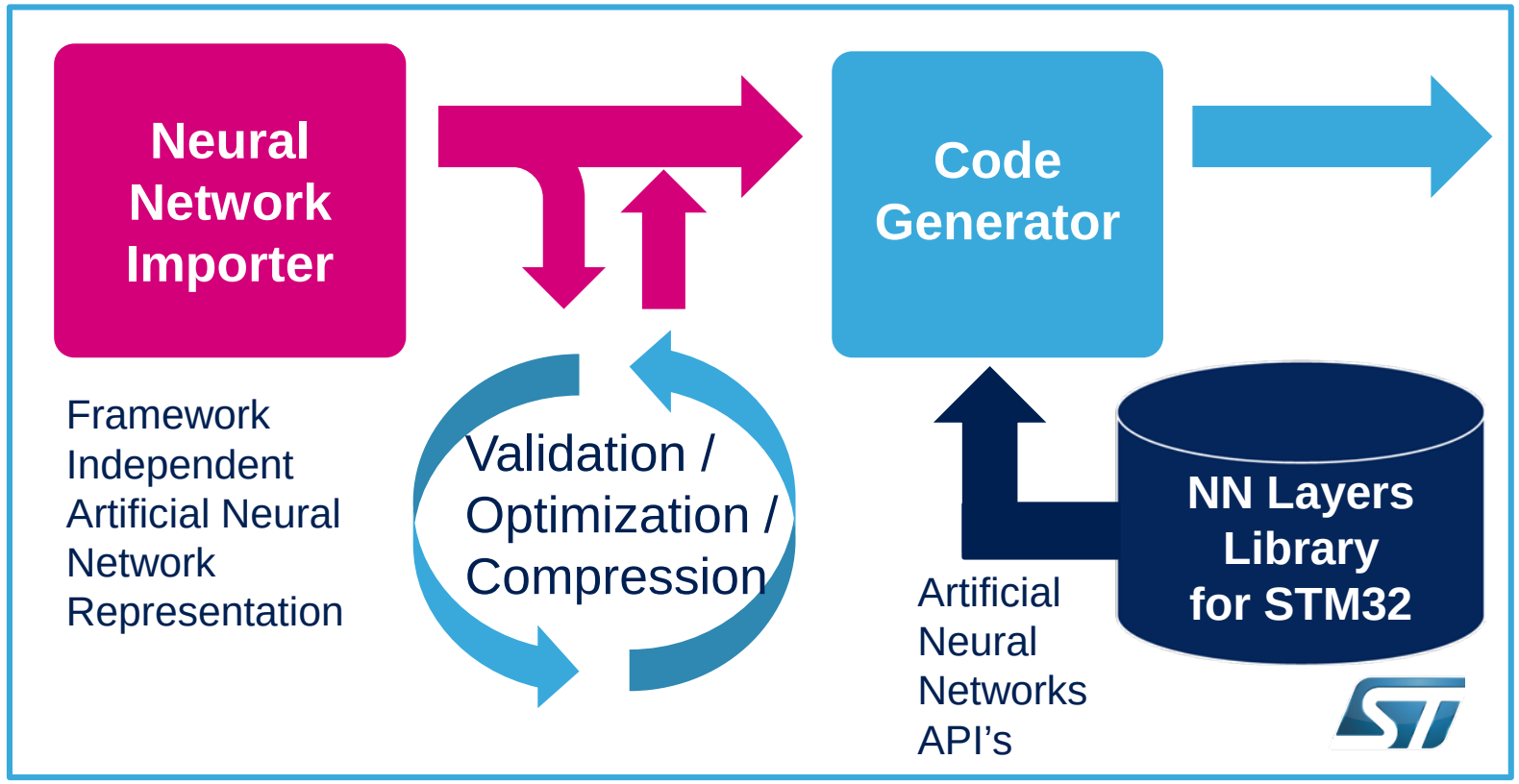
Convert NN into
optimized code for MCU

Process & analyze
new data using trained NN

STM32Cube-AI : Architecture

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

Deep Learning Framework dependent



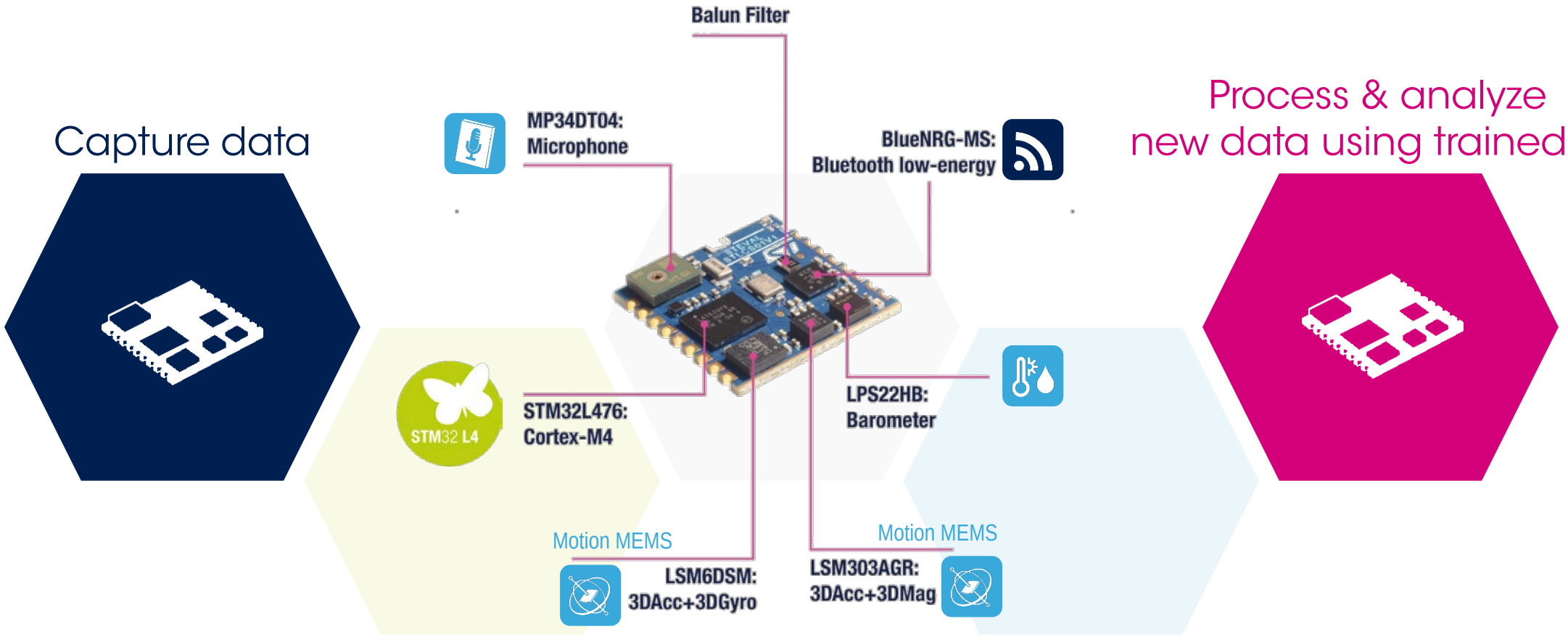
Embedded Solution
Optimized Artificial Neural Network Code generated for STM32

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

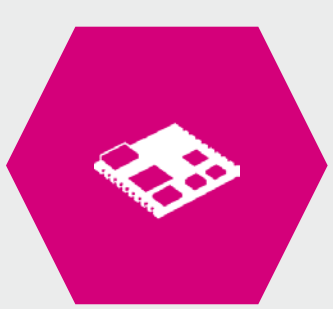
Requirements of different NN examples

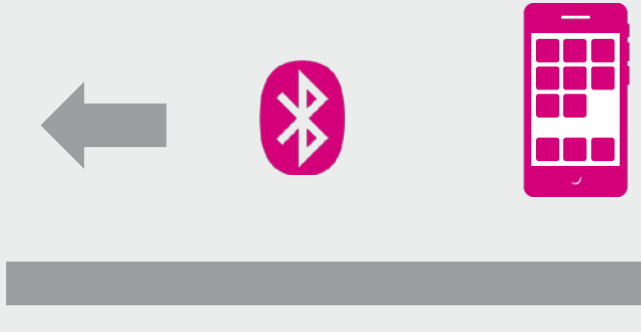
NN project	Output classes	Memory footprint	Complexity (MACC)	NN Model	Dataset
Human Activity Recognition GMP (Accel. 3-axis input)	5	4 KB RAM 6 KB Flash	69k	ST proprietary CNN Lasagne model	ST proprietary dataset of 2.4M samples
Human Activity Recognition IGN (Accel. 3-axis input)	5	1.7 KB RAM, 12 KB Flash	14k	Derived from a published paper Keras model	ST proprietary dataset of 2.4M samples
Human Activity Recognition IGN (Accel. 3-axis input)	4	1.7 KB RAM, 5.4 KB Flash	14k	Derived from a published paper Keras model	Wisdom public dataset
Acoustic Scene classification (Mic. 16KHz input)	3	18KB RAM, 31KB Flash	517k	ST proprietary CNN Keras model	ST proprietary dataset of 22h53m audio samples

Form Factor Hardware
to Capture and Process Data



Example: Human Activity Recognition





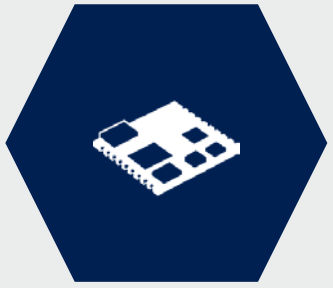
5 classes

Embedded **motion**

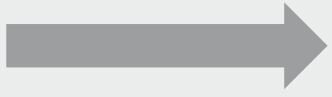
Labelling controlled
by smartphone application

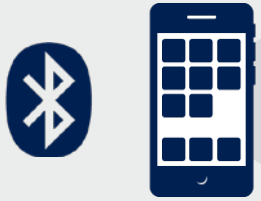
Data stored on the device
SD card for future **learning**

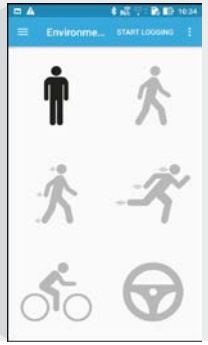
Stationary, walking, running,
biking, driving



NN & example
dataset provided





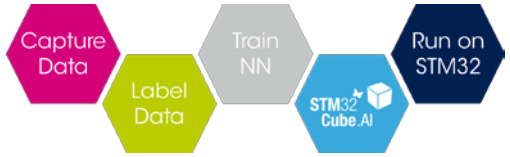


Embedded **motion**
pre-processing

Inferences running
on the microcontroller

Inference result
displayed on mobile app

life.augmented



STM32 Solutions for AI

More Than Just the STM32Cube.AI

An extensive toolbox to support easy creation of your AI application

AI extension for STM32CubeMX

To map pre-trained Neural Networks onto the STM32

STM32 Cube.AI

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

life.augmented

For more Information

www.st.com/STM32CubeAI