An Introduction to Edge AI for Data Scientists

Ann-Christin Bette

2024-05-14

What if your device was capable of …

…detecting abnormal patterns in your heart rate, other than just RHR or HRV?

…detecting movements & repetitions during your workout with highest accuracy?

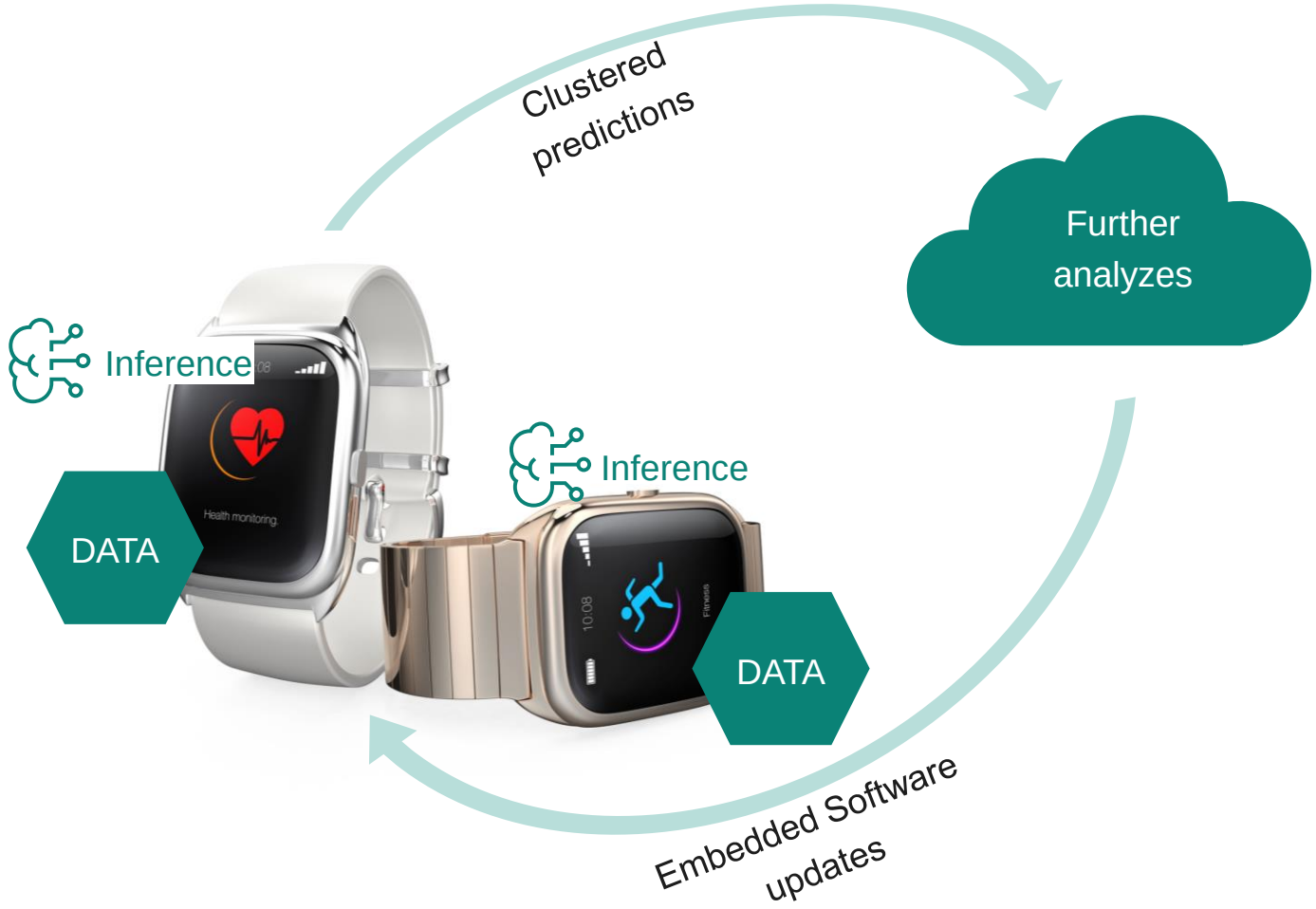
…giving indications of your health based on snoring, coughing or other sounds?

What if your device was limited to Cloud AI?

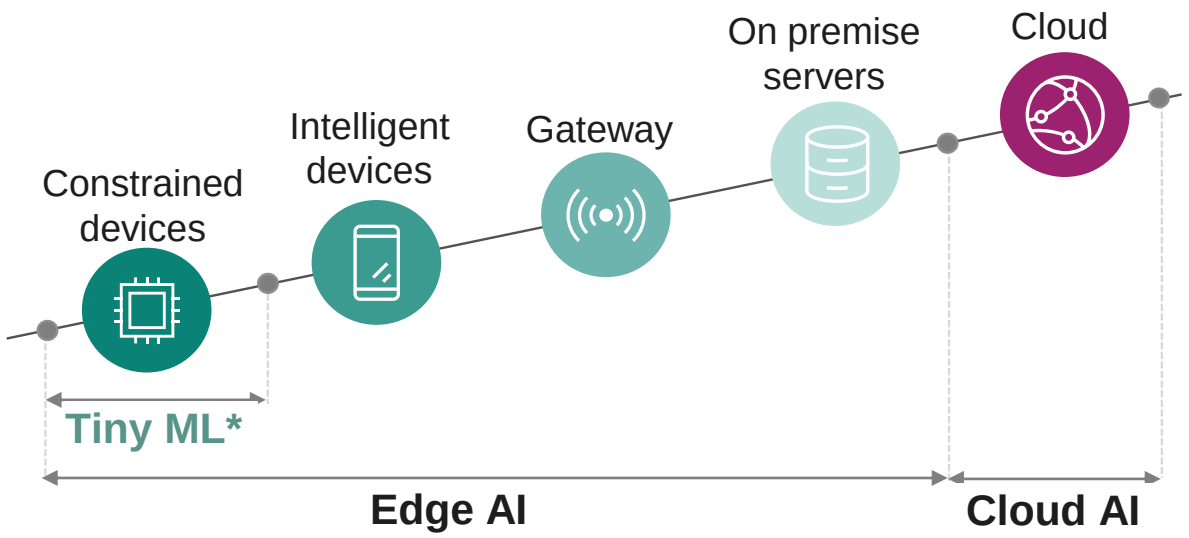
- Data privacy
- Power efficiency
- System reliability
- Latency
- Functional security
- Cost efficiency

What if your device could work with Edge AI?

- Data privacy
- Power efficiency
- System reliability
- Latency
- Functional security
- Cost efficiency

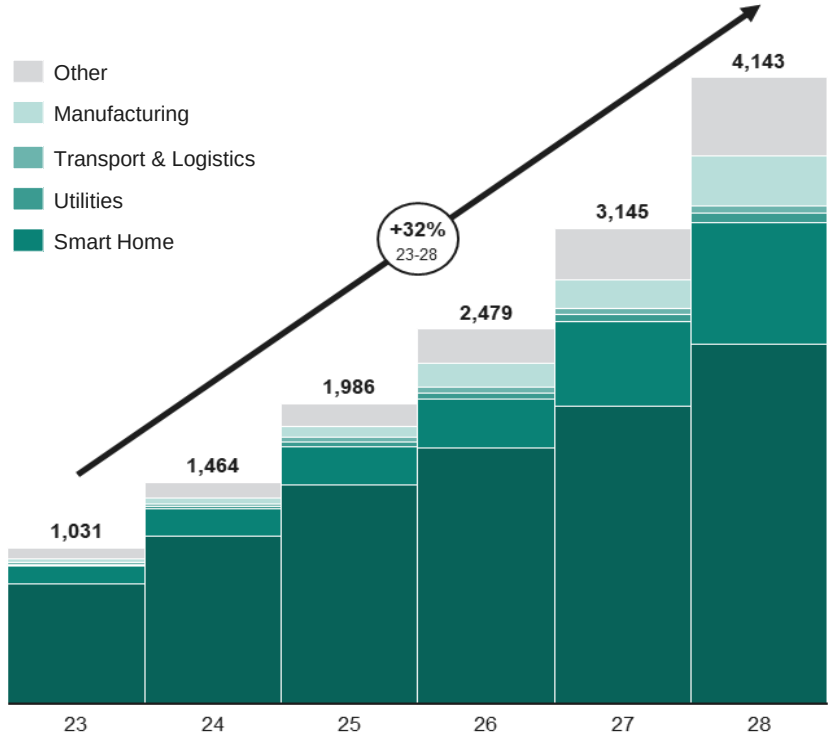


AI moves directly into the devices…



* Tiny ML = Tiny Machine Learning
Refers to AI inference on microcontrollers

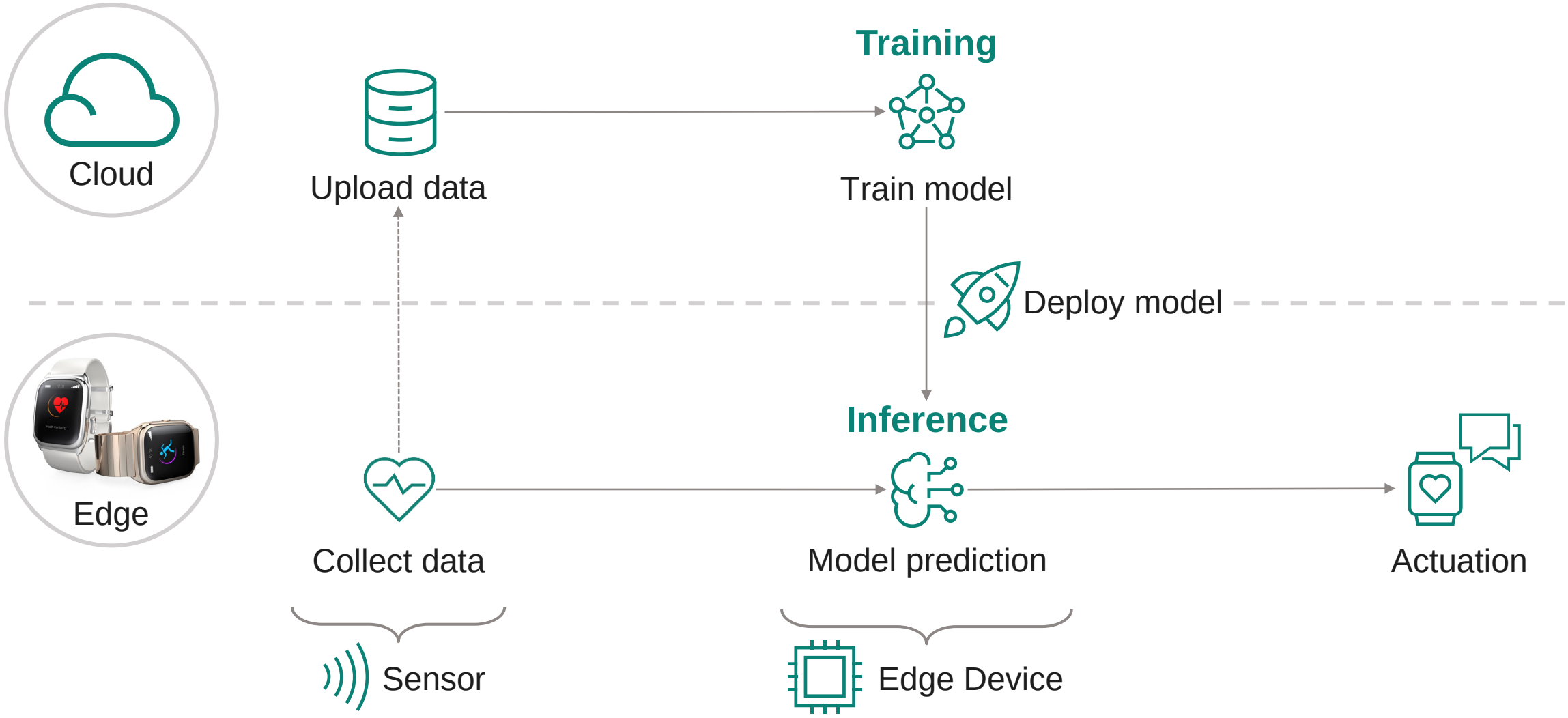
… which impacts a variety of markets



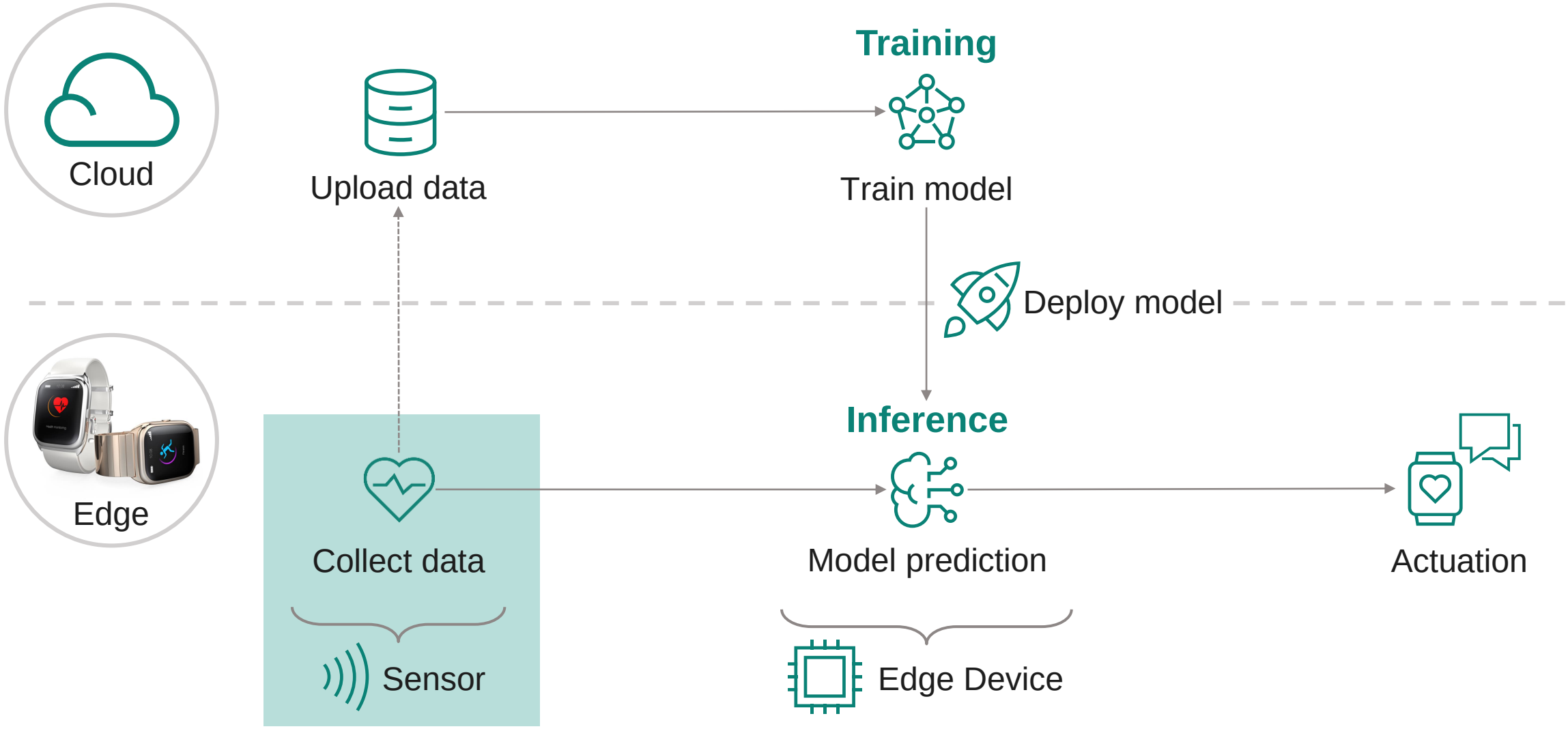
Total shipments of Tiny ML devices by vertical (millions)[1]

Source: ABI Artificial Intelligence and Machine Learning (Q1/2023);

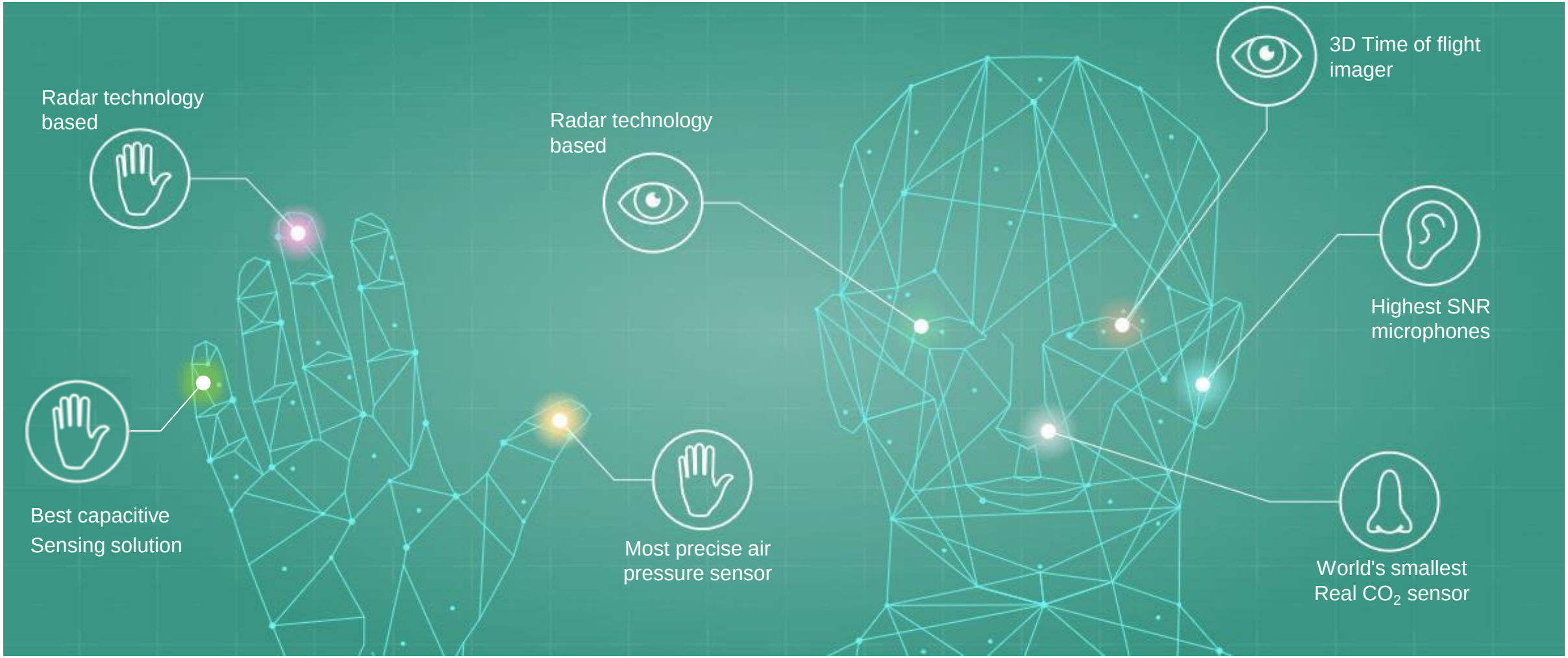
AI at the Edge – Workflow



AI at the Edge – Workflow



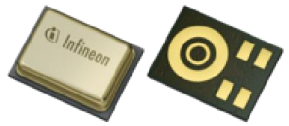
Our intuitive sensors are enabling Edge AI – Giving things the human sense



Infineon's XENSIV™ Sensors – Adding simulated human senses to digital systems to make our lives easier and better

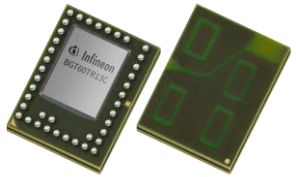
Smart ears

Microphone

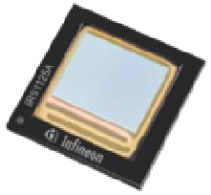


Smart eyes

mmW RADAR

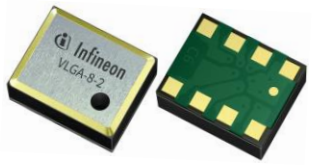


3D Time-of-Flight



Smart feel

Pressure



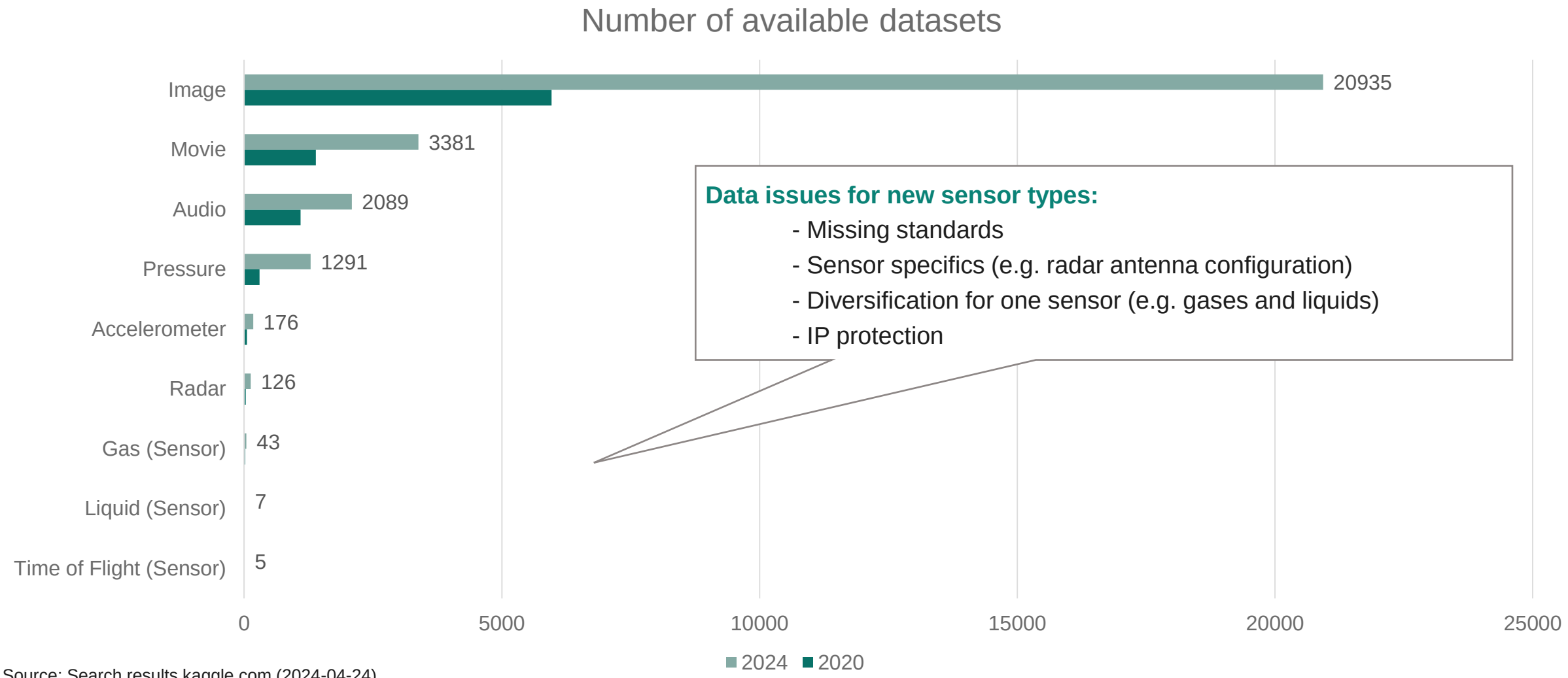
Smart nose

Environmental

Sensor fusion – Combining multiple sensors to improve data quality and confidence

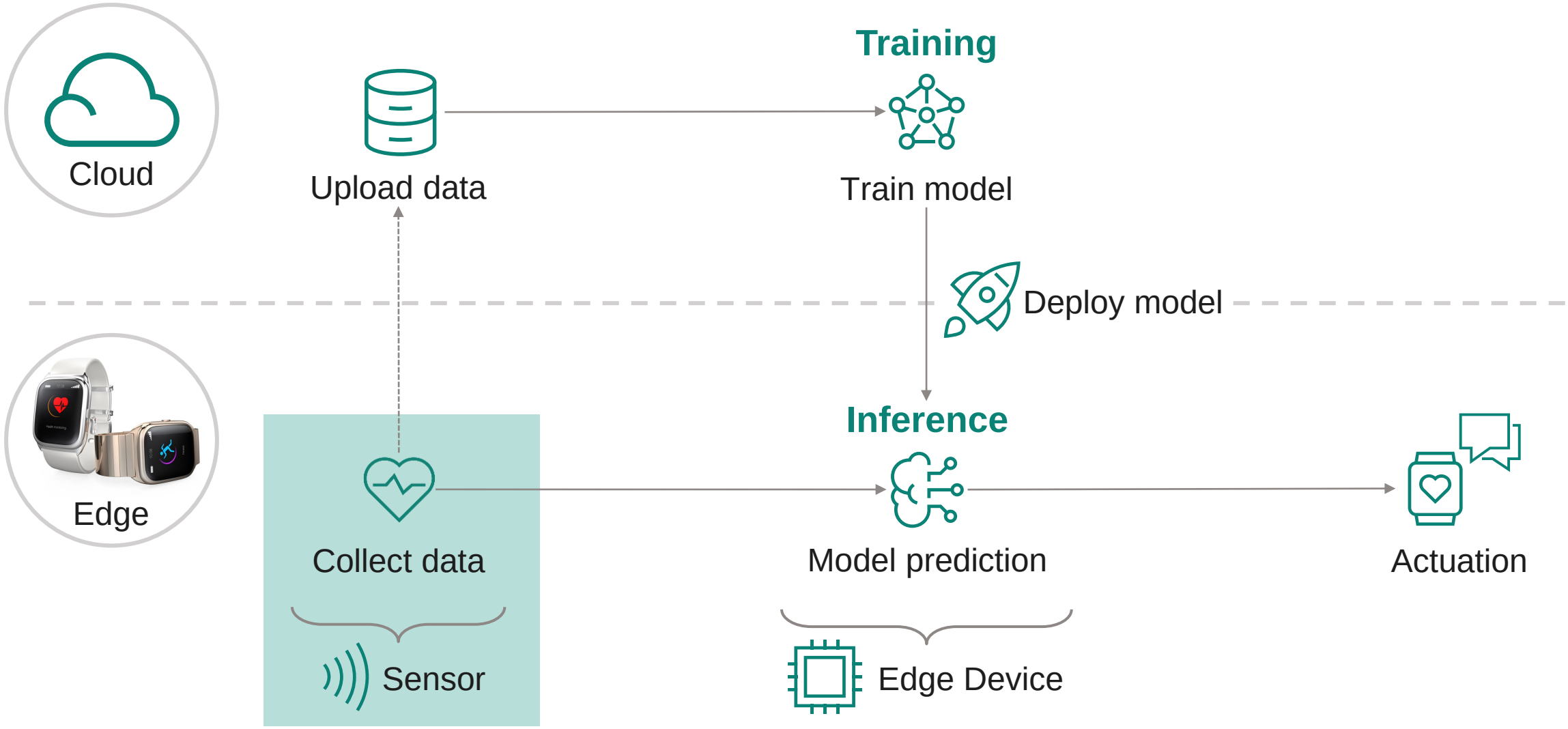
Infineon XENSIV™ sensors are exceptionally precise thanks to industry-leading technologies. They are the perfect fit for various customer applications in automotive, industrial and consumer markets.

Public data sets beyond audio and vision are sparsely available

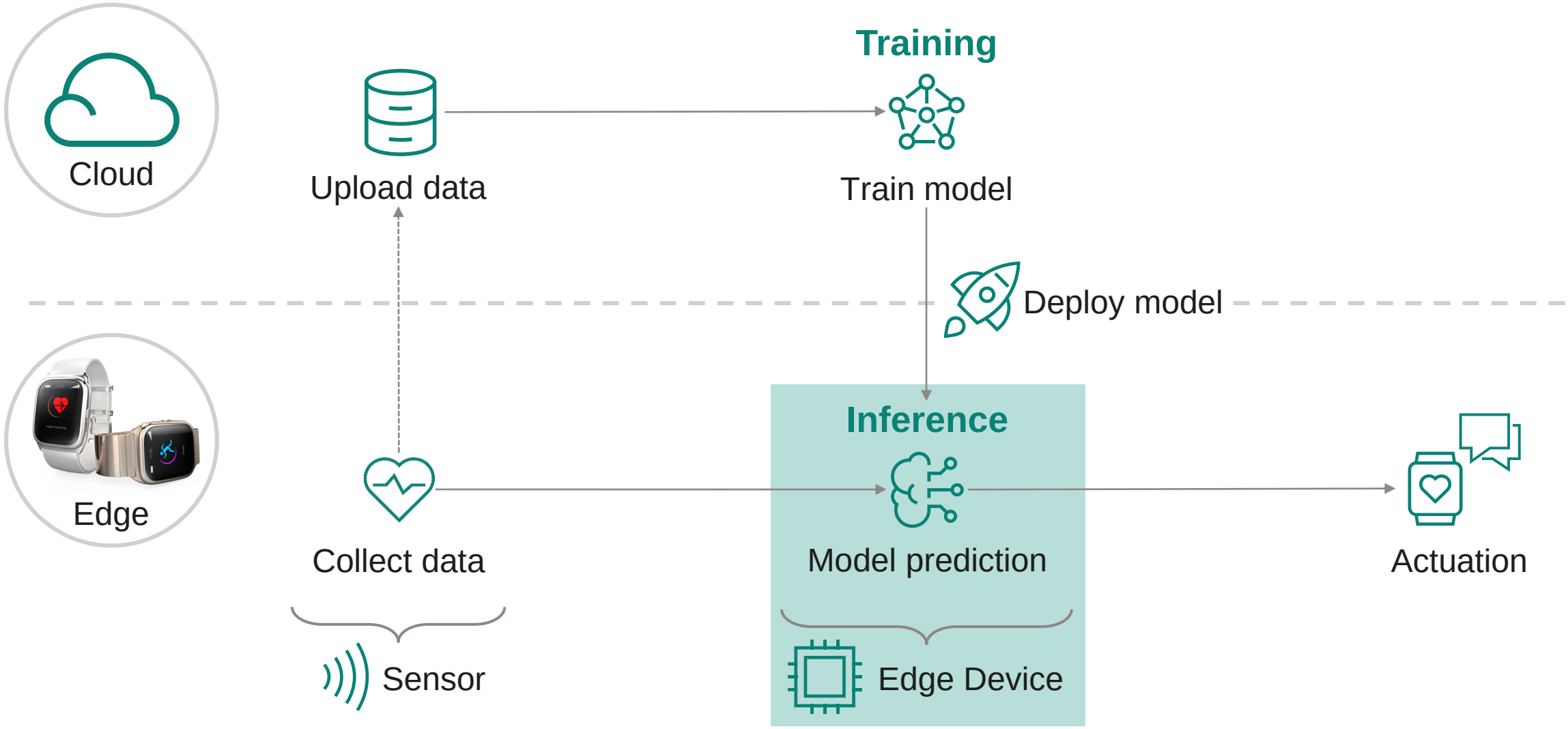


Source: Search results kaggle.com (2024-04-24)

AI at the Edge – Workflow



AI at the Edge – Workflow



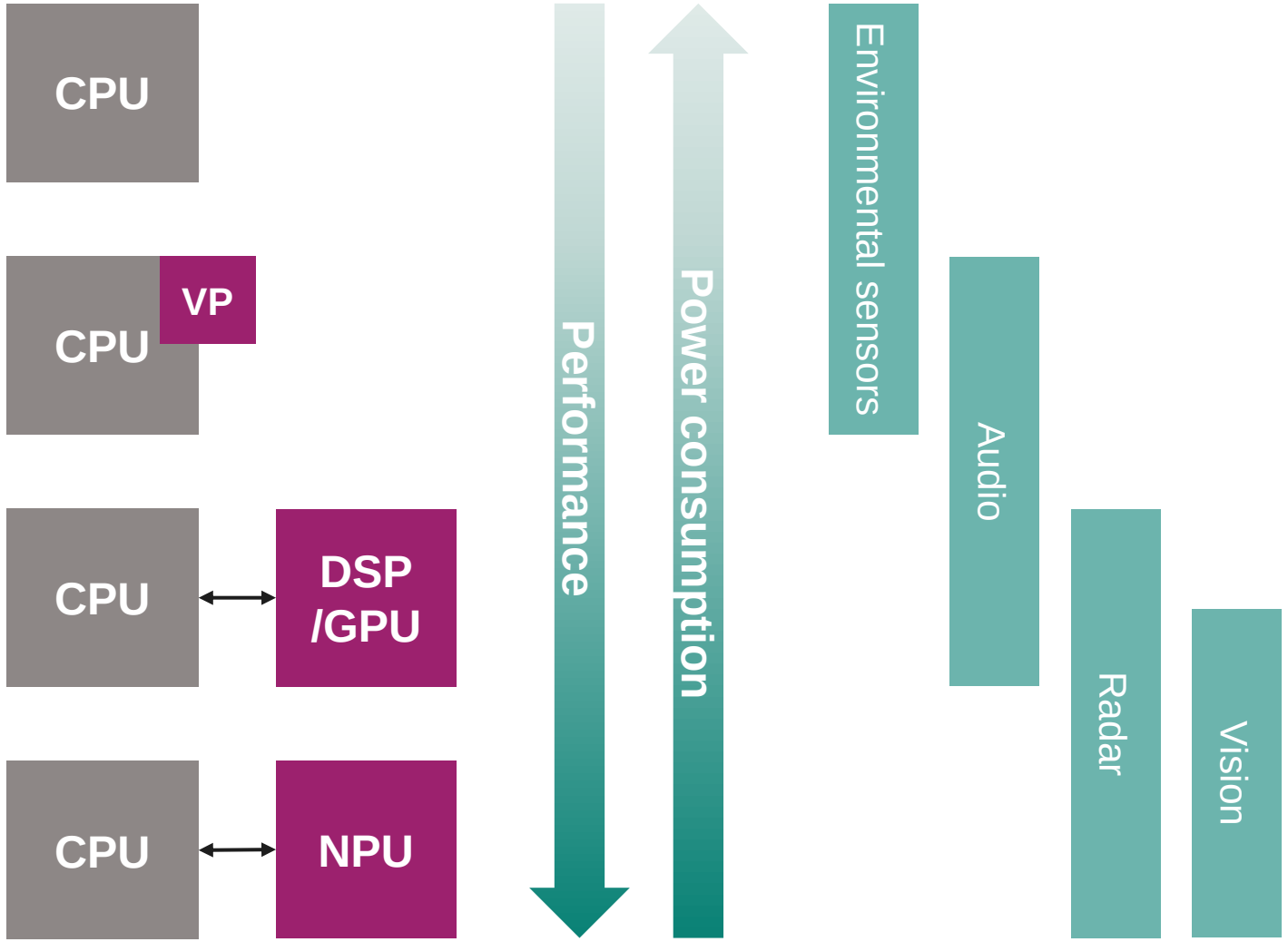
Ways to accelerate neural networks in silicon according to the respective field of application

CPU only
AI in software

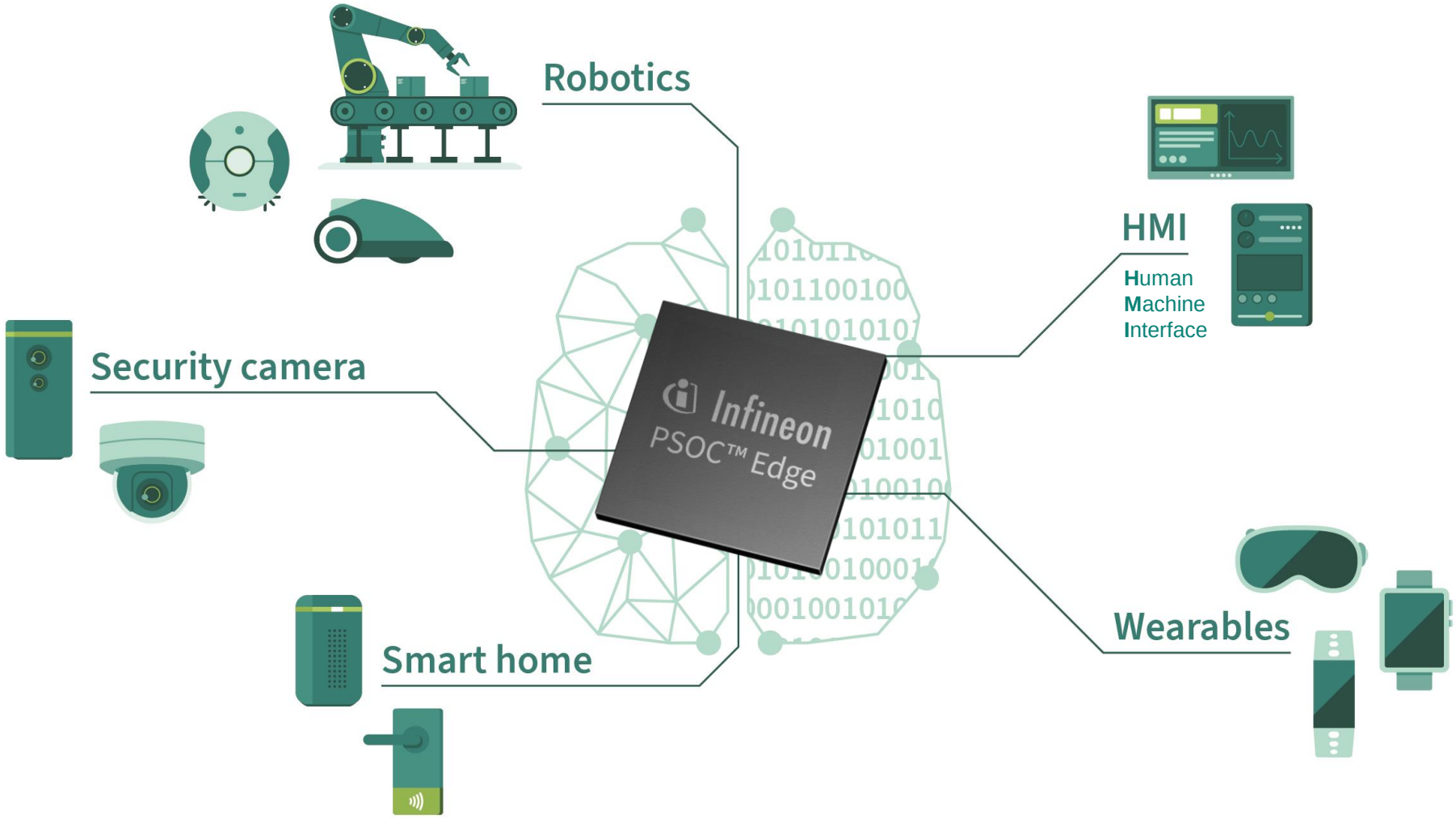
CPU with extensions
Vector processing extensions

CPU + DSP/GPU
Digital Signal Processor
Graphics Processing Unit

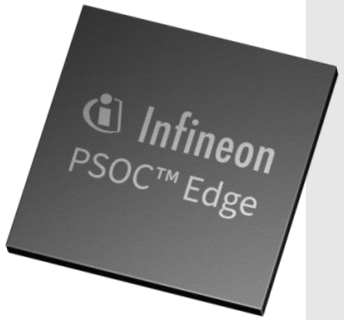
CPU + NPU
Specialized
Neural Processing Unit



PSoC™ Edge
Enabling a new generation of responsive ML Edge devices



PSoC™ Edge
Enabling a new generation of responsive ML Edge devices



System Resources

- Power Management
- Clock Management

Memory System

- Embedded SRAM
- Embedded PRAM
- External Interface Memory

Low Power CPU System

- Arm® Cortex®-M33
- Infineon NNLite

High Performance CPU System

- Arm® Cortex®-M55
  - FPU
  - Helium™ DSP
- Arm® Ethos®-U55

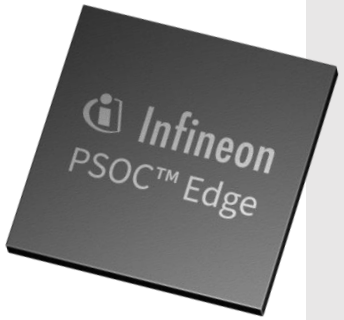
Security System

- Secure Enclave

Peripherals & IO

- Serial Comm.
- Audio
- Display

PSoC™ Edge
Enabling a new generation of responsive ML Edge devices



System Resources

Power Management

Clock Management

Memory System

Embedded SRAM

Embedded PRAM

External Interface Memory

Low Power CPU System

Arm® Cortex®-M33

Infineon NNLite

High Performance CPU System

Arm® Cortex®-M55

FPU

Helium™ DSP

Arm® Ethos®-U55

Security System

Secure Enclave

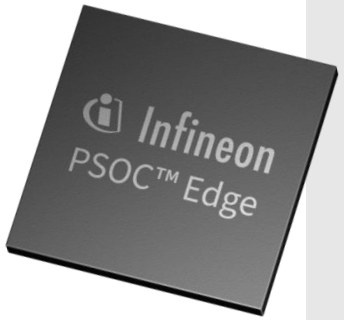
Peripherals & IO

Serial Comm.

Audio

Display

PSoC™ Edge
Enabling a new generation of responsive ML Edge devices



System Resources

Power Management

Clock Management

Memory System

Embedded SRAM

Embedded PRAM

External Interface Memory

Low Power CPU System

Arm® Cortex®-M33

CPU

Infineon NNLite

NPU

High Performance CPU System

Arm® Cortex®-M55

FPU

Helium™ DSP

CPU

VP

Arm® Ethos®-U55

NPU

Security System

Secure Enclave

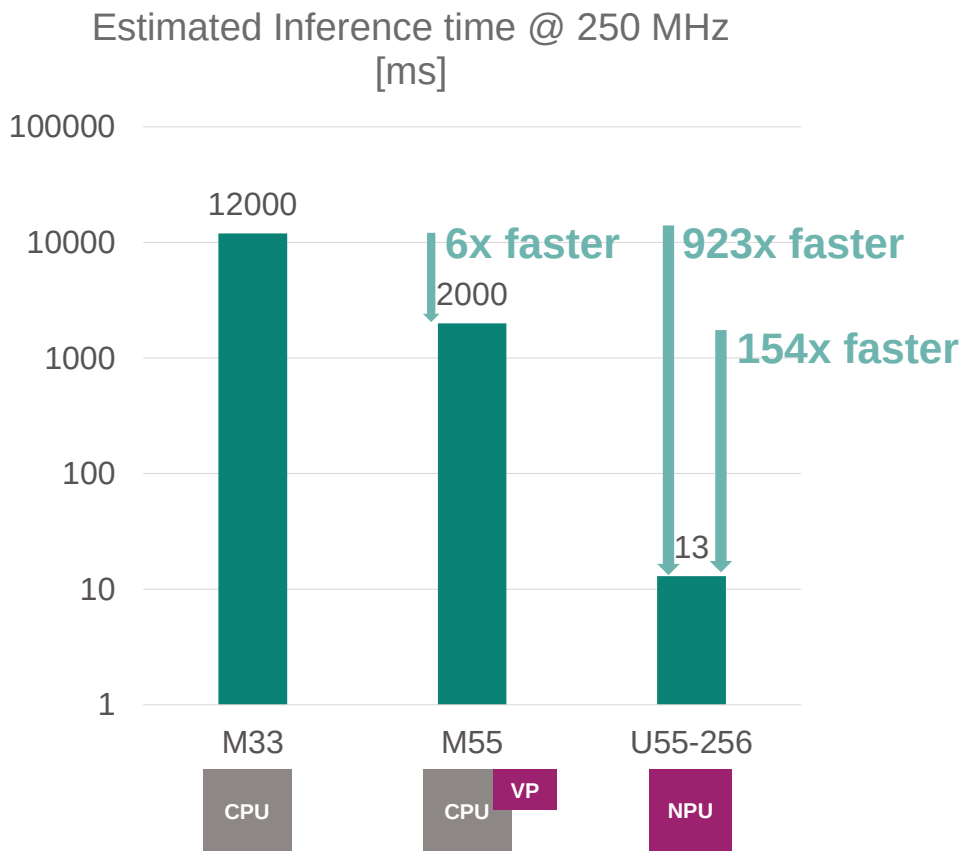
Peripherals & IO

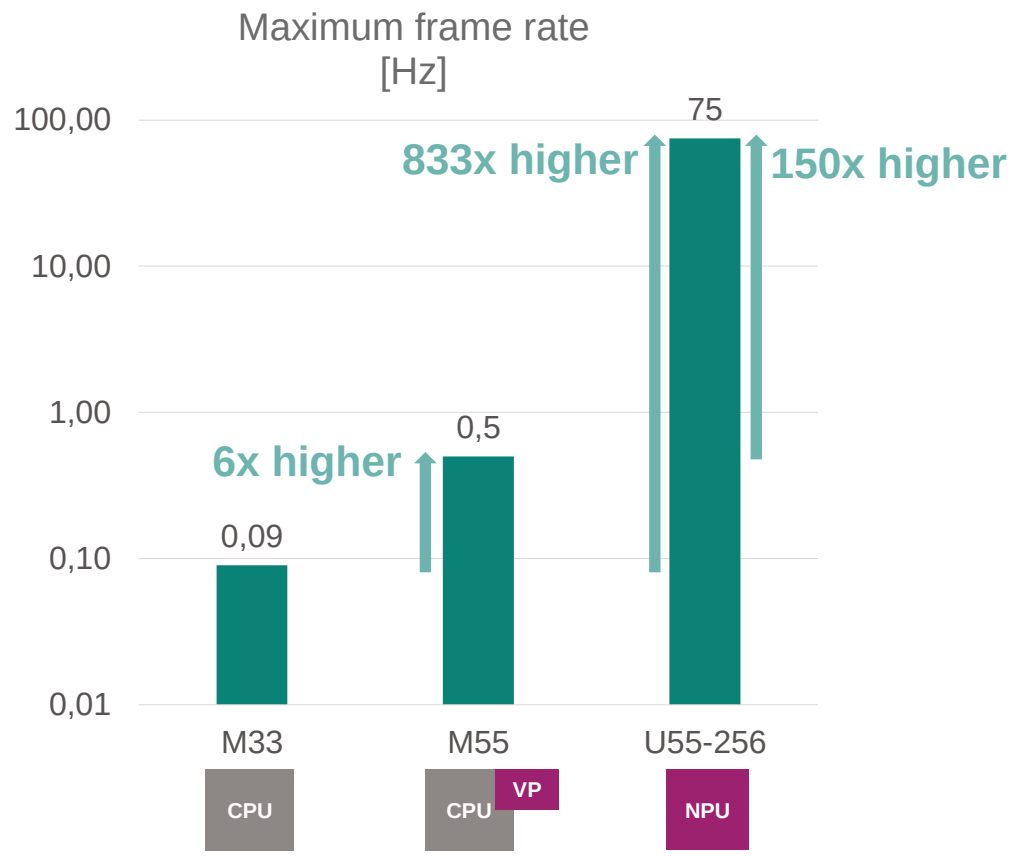
Serial Comm.

Audio

Display

Inference and frame rate benchmark of Arm Cortex-M33, -M55 and Ethos-U55 for MobileNet_v1_1.0_224_quant





Data obtained by Infineon

Memory and storage benchmark of PSoC Edge, Apple A13 Bionic and Nvidia V100 for ResNet-50 and MobileNetV2

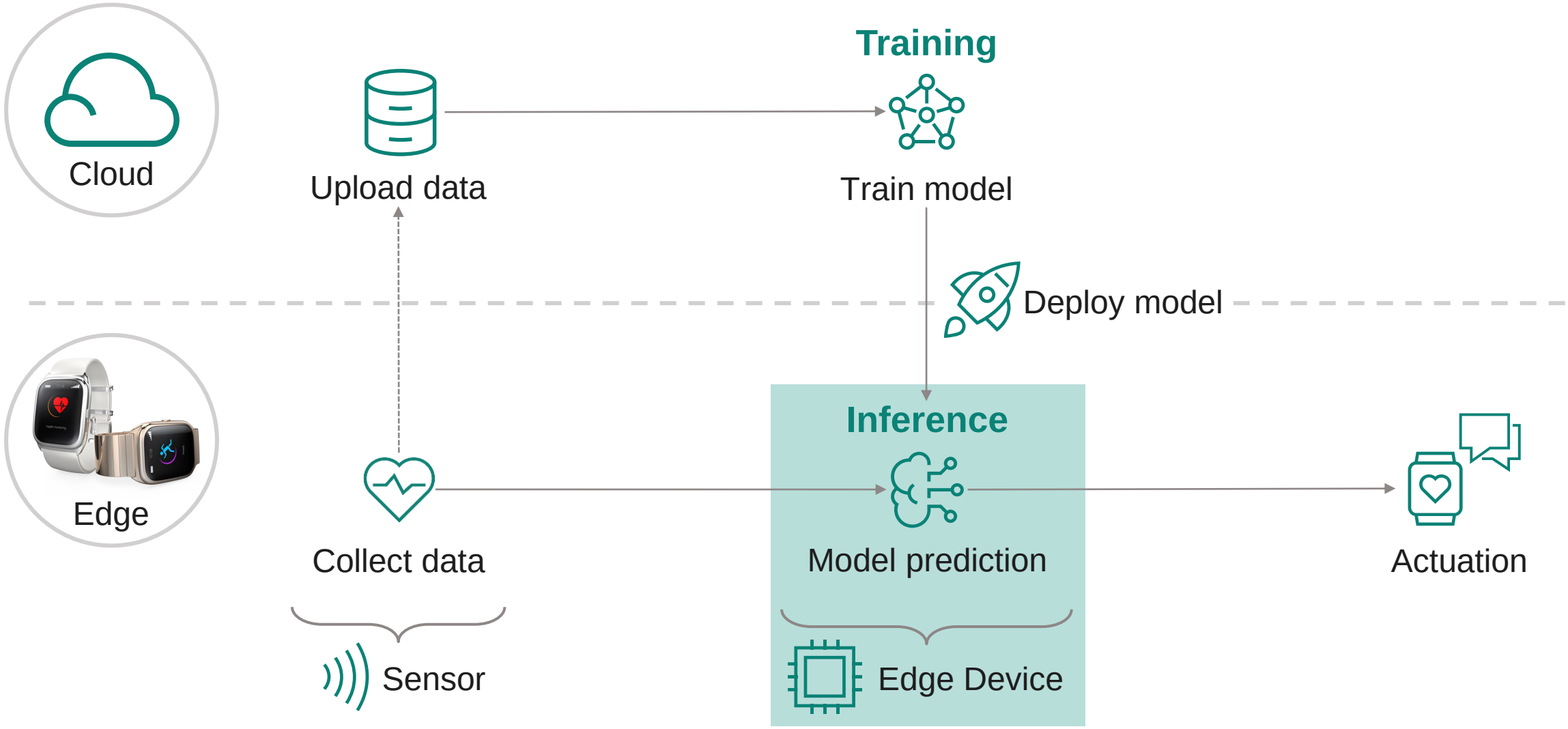
	Cloud AI (Nvidia V100)	**Mobile AI** (A13 Bionic)	**Tiny ML** (PSoC Edge)
Memory (RAM)	16 GB	4GB	<5 MB*
Storage (SSD/Flash)	TB~PB	>64 GB	A few MB*

	ResNet-50 (float32)	**MobileNetV2** (float32)	**MobileNetV2** (int8)
Memory (RAM)	7.2 MB	6.8 MB	1.7 MB
Storage (SSD/Flash)	102 MB	13.6 MB	3.4 MB

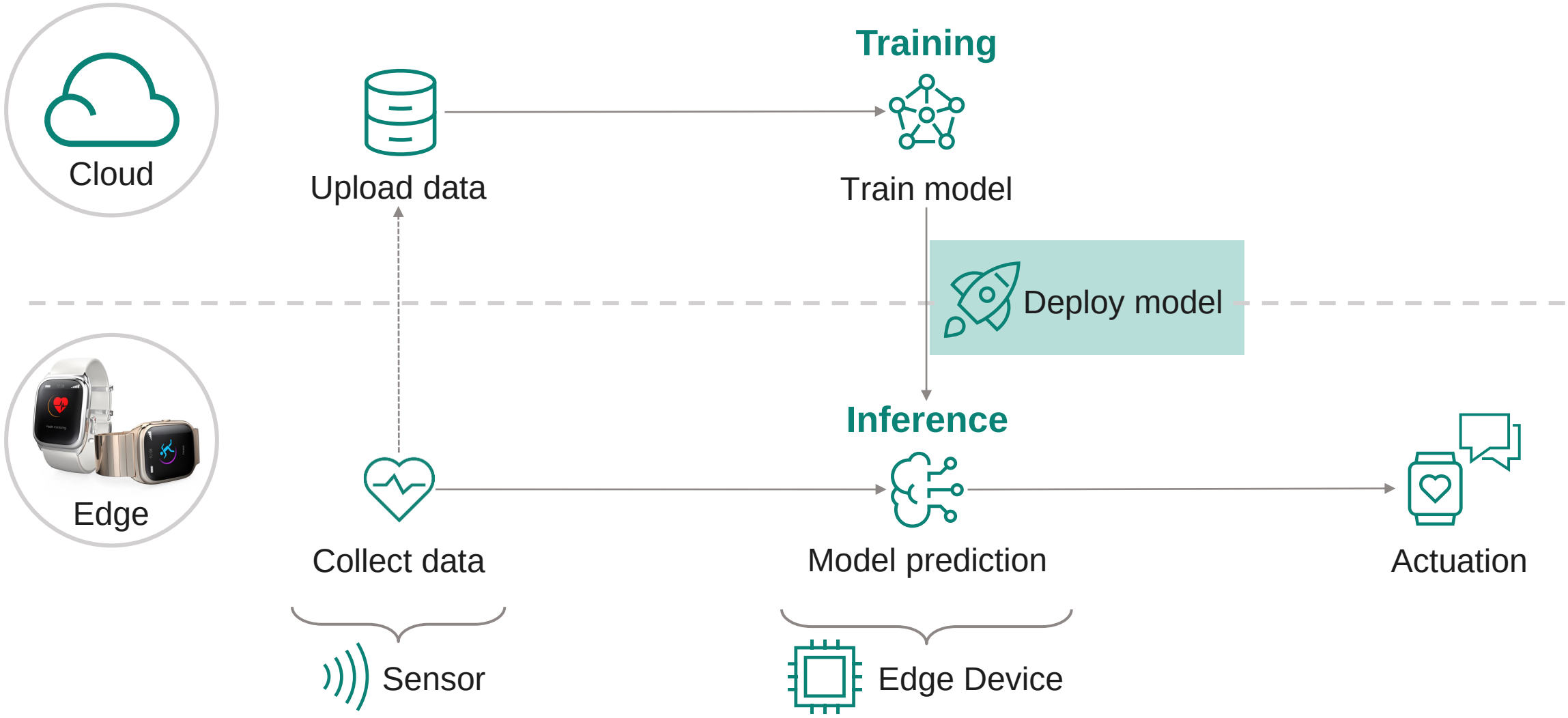
* Data added by Infineon

Source: J. Lin, L. Zhu, W. -M. Chen, W. -C. Wang and S. Han, "Tiny Machine Learning: Progress and Futures [Feature]," in IEEE Circuits and Systems Magazine, vol. 23, no. 3, pp. 8-34, thirdquarter 2023, doi: 10.1109/MCAS.2023.3302182

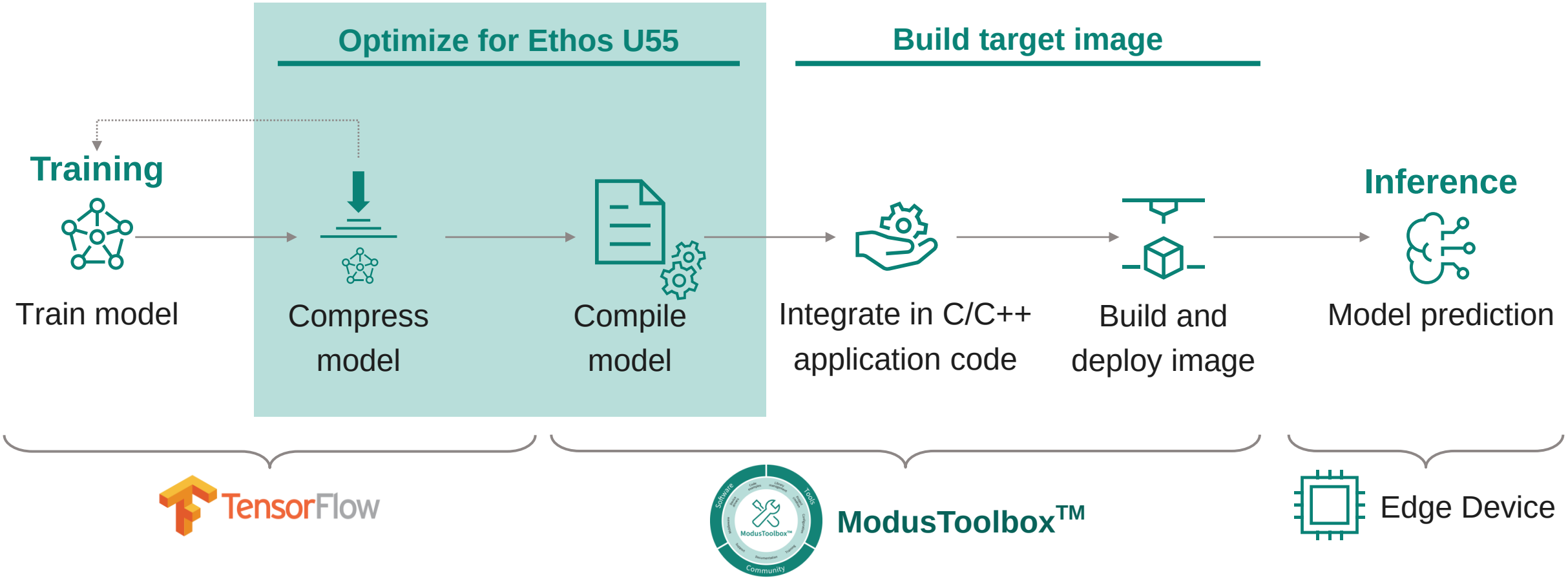
AI at the Edge – Workflow



AI at the Edge – Workflow



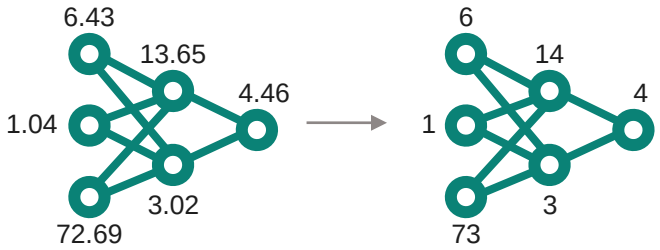
End-to-end model deployment flow with Ethos-U55



Model optimization: Two common compression techniques

Applied during or
after training phase

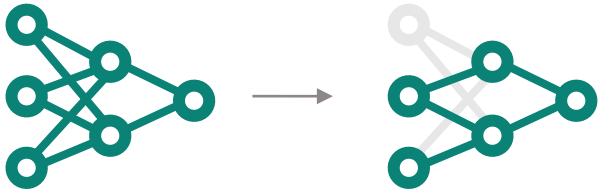
Goal: Reduce the
size of the model



Quantization:
Floating point to 8-bit
Integer conversion

„Costs of Compression“:
Decreased model size leads to…

…decreased memory size
…decreased inference time
…decreased power consumption
…decreased accuracy

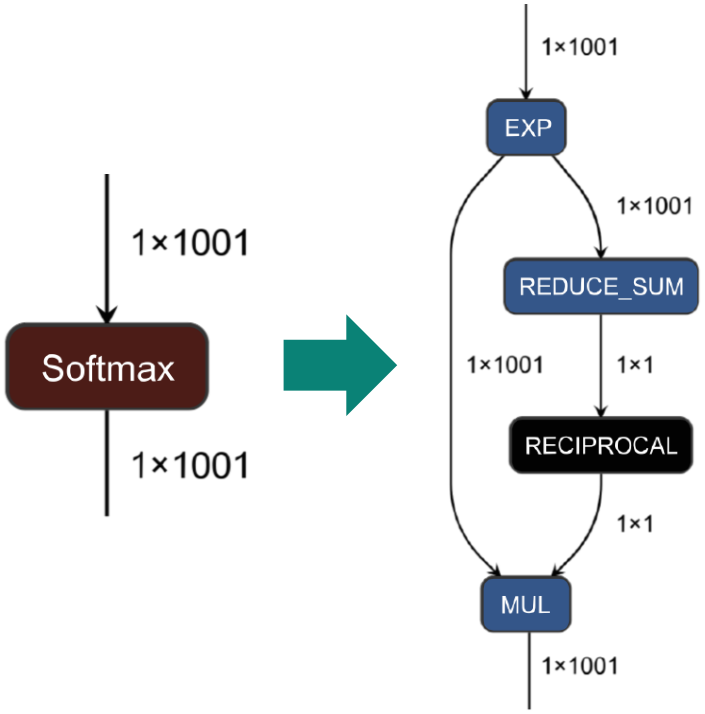


Pruning:
Remove unnecessary
weights and nodes

Model optimization: Model compiler

Applied after training phase on compressed model

Goal: Convert the model into machine readable code



Graph level optimization:
A set of graph-level operators for model layer fusion and mapping

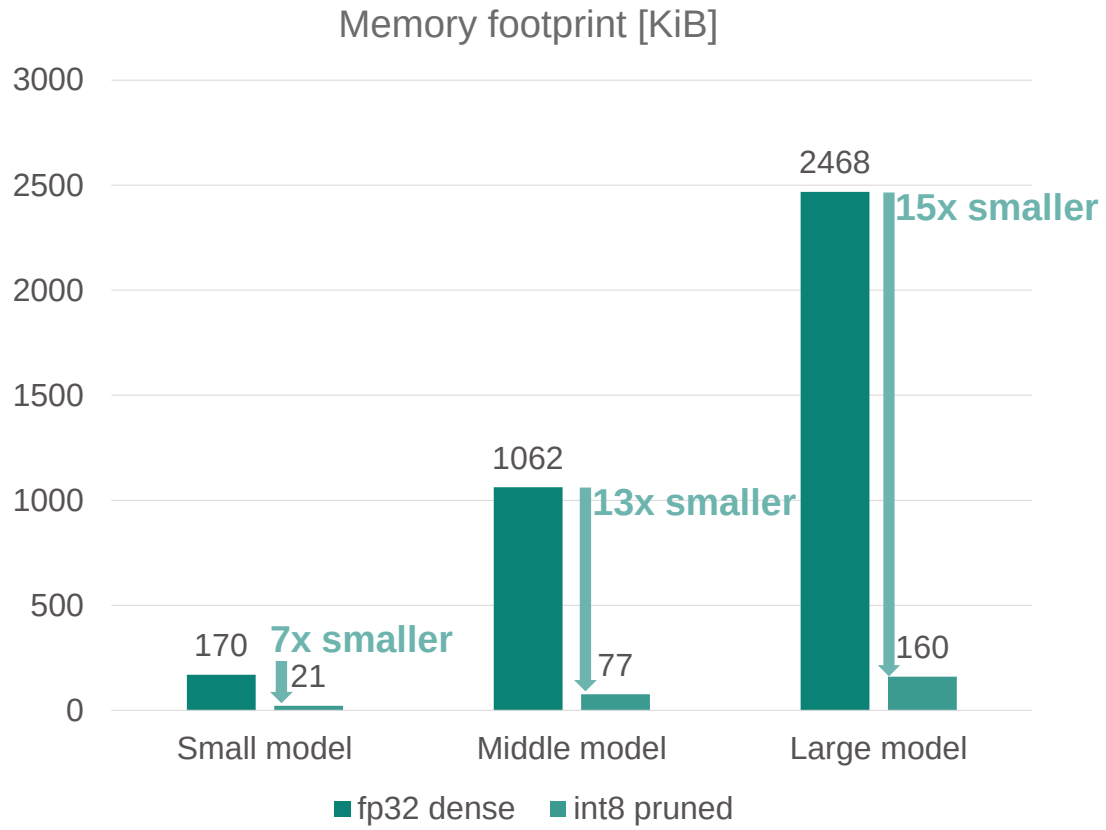
Tensor level optimization:
A set of tensor-level operators for loop unrolling, vectorization, parallelization…

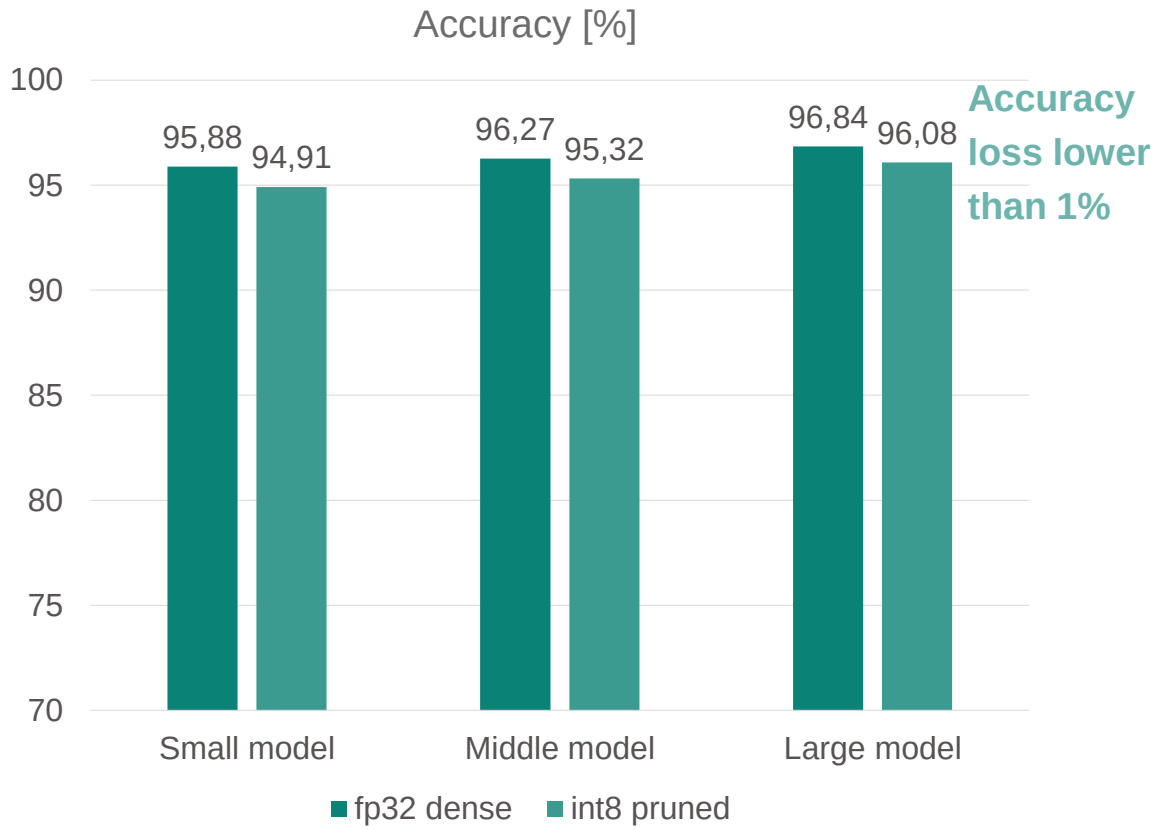
„Costs of model compilers“:

- Hardware dependence:
  The choice of compiler depends on the hardware
- Compiler support:
  Layer fusion and mapping not supported for all types of layers

Memory and accuracy benchmark for dense and compressed (quantized and pruned) keyword spotting models

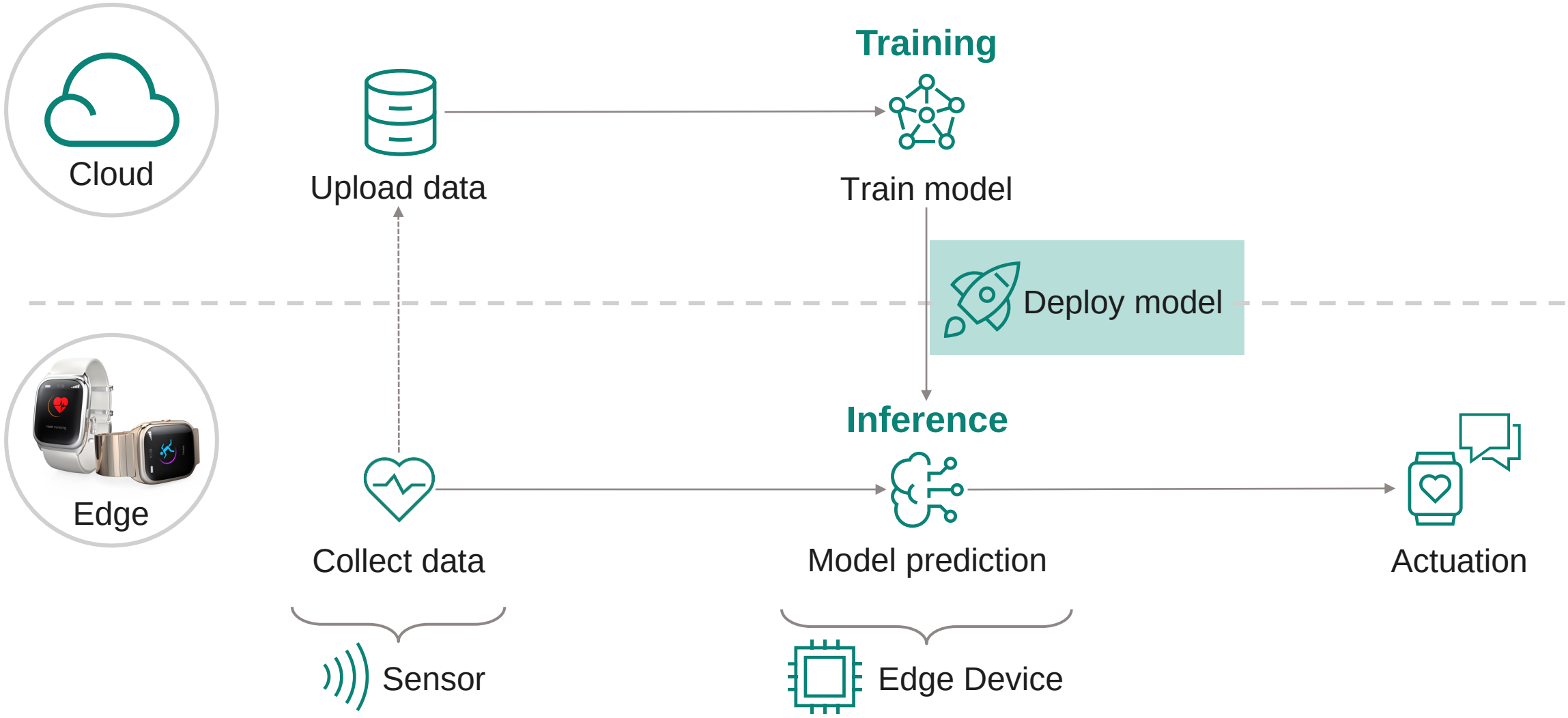
Dataset: Google speech commands



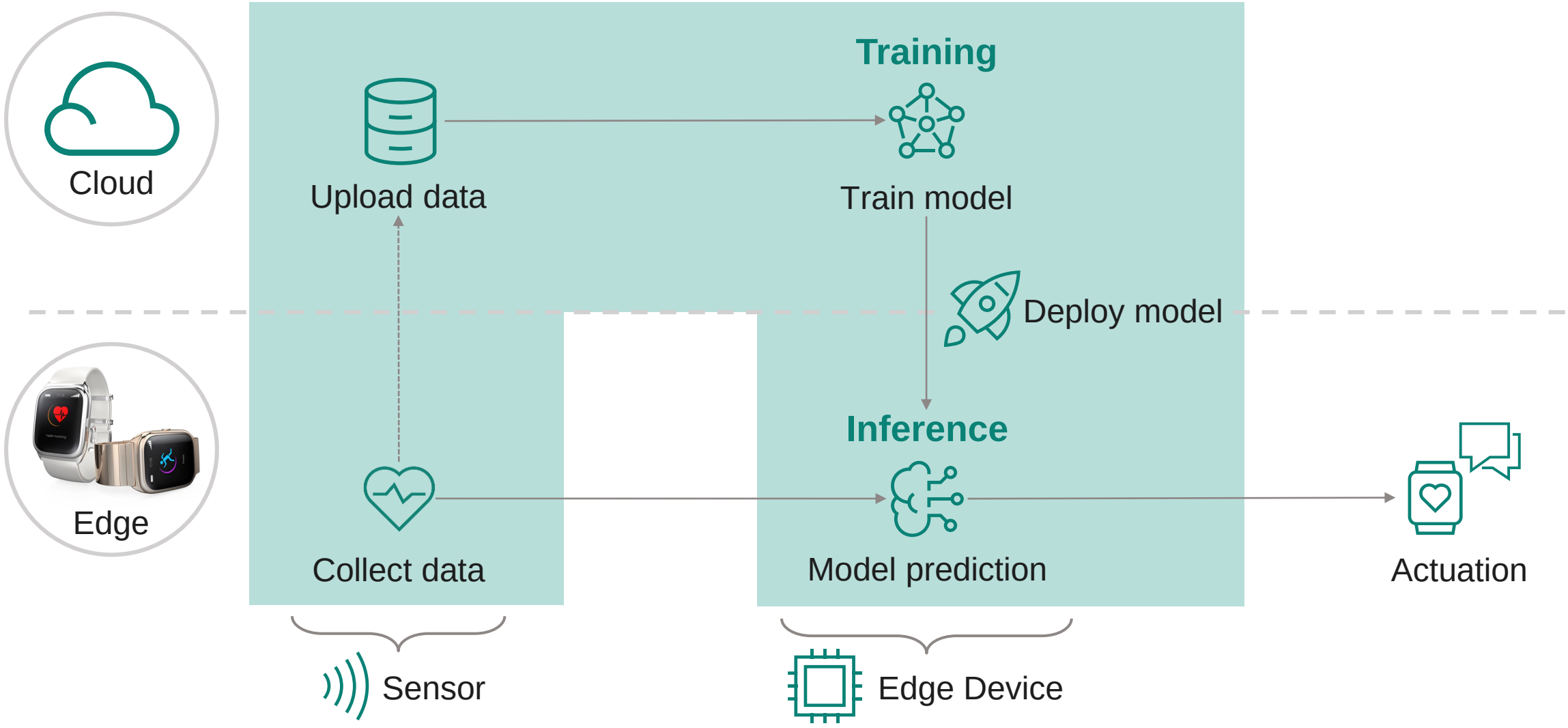


Source: E. Trommer, B. Waschneck and A. Kumar, "dCSR: A Memory-Efficient Sparse Matrix Representation for Parallel Neural Network Inference," 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD), Munich, Germany, 2021, pp. 1-9, doi: 10.1109/ICCAD51958.2021.9643506.

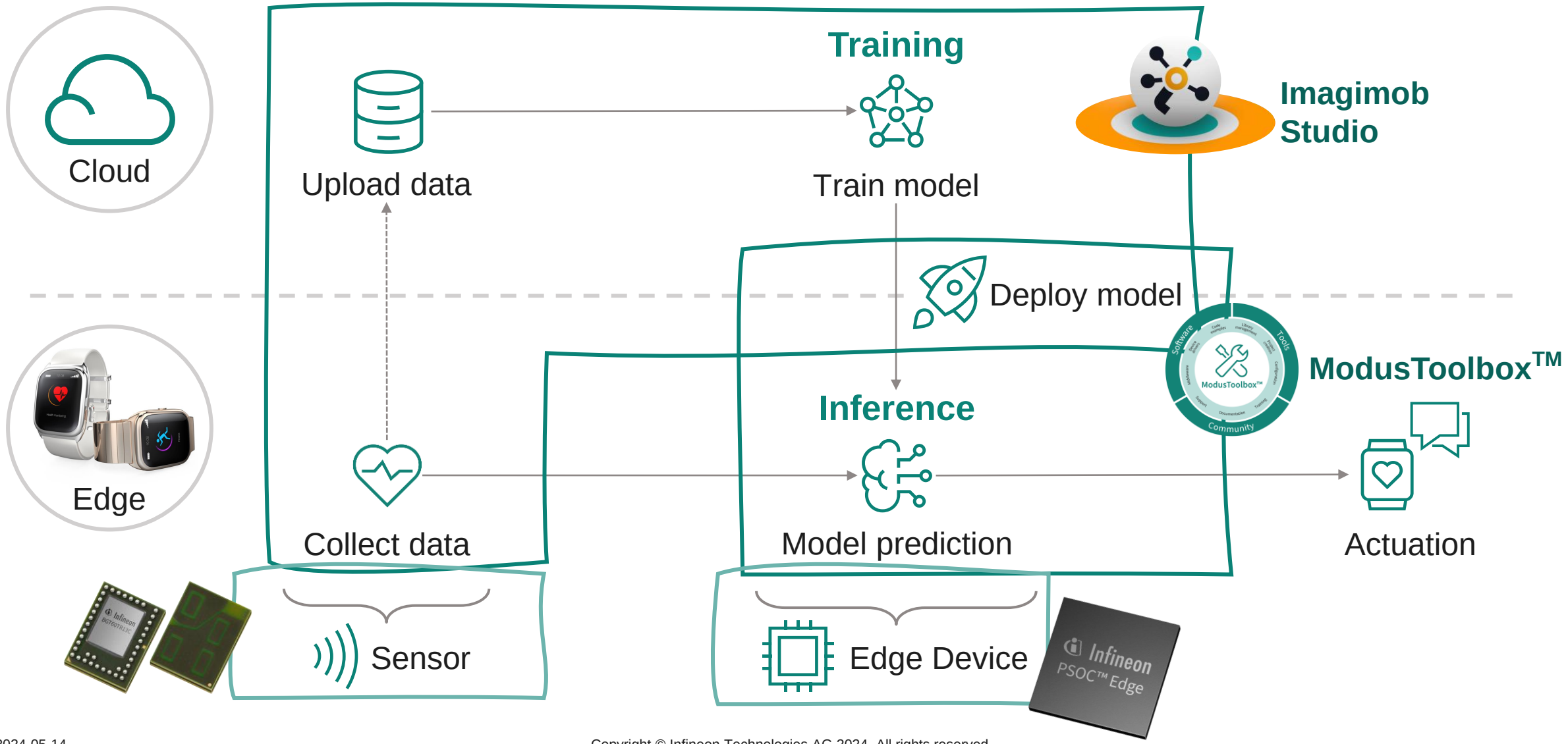
AI at the Edge – Workflow



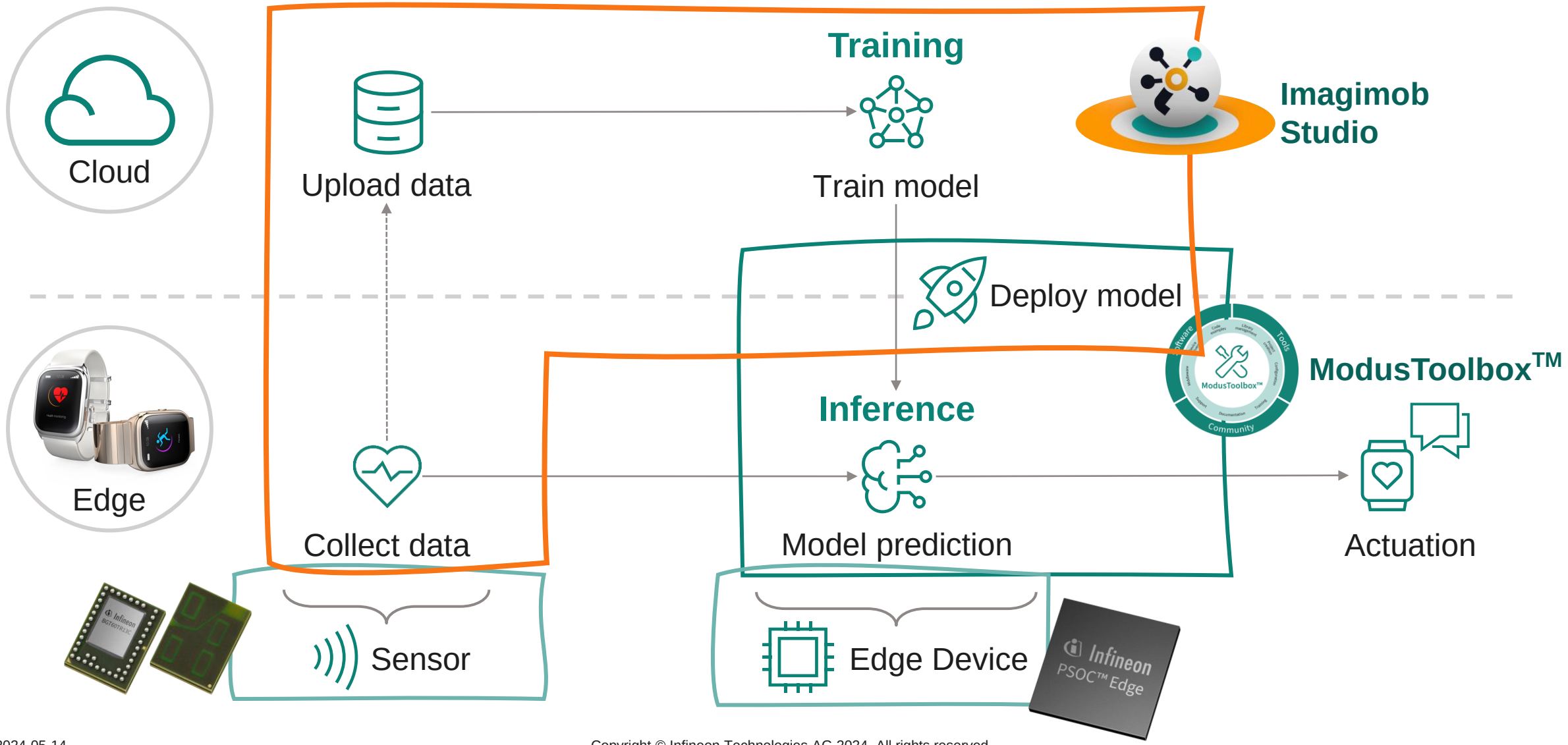
AI at the Edge – Workflow



AI at the Edge – Infineon Offering for Consumer IoT



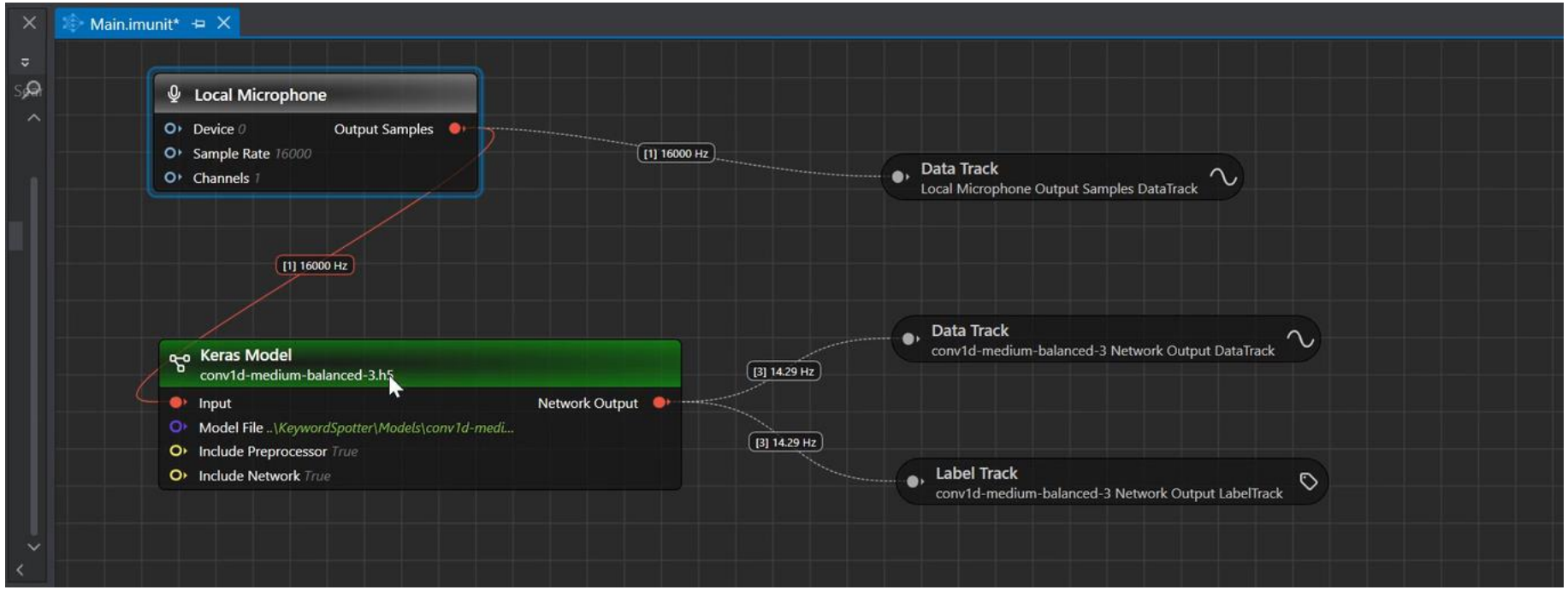
AI at the Edge – Infineon HW/SW Solution for Consumer IoT



Introduction to Imagimob (Videolink for external)

https://e.video-cdn.net/share?video-id=FW5JRqijjDXxSe_VCwGcbZ&player-id=2t2W2ykrDB_RisZ1QorEJU&channel-id=101646

Imagimob Studio – Graph UX is an intuitive interface to visualize the end-to-end machine learning workflow as graphs



Ready for the edge: Our offering of production quality, ready to deploy ML models lets you easily add new features to your device

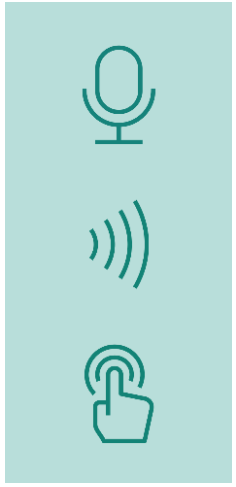
The fastest way of taking edge AI to market: Ready to go, fully trained, and comprehensively tested machine learning model that are ready for production.

Coughing Detection

Siren Detection

Snoring Detection

Baby Cry Detection



Acoustic Radar CAPSENSE™

Available Now

Coming Soon

Find out more or get your model: https://www.imagimob.com/ready-models

- **Fully trained and tested models**
- **Production quality – developed by experts**
- **High accuracy**
- **Easy to integrate, ready to commercialize**

Key Take-aways

- Edge AI enables the creation of completely new applications
- Several advantages over Cloud AI, e.g. low latency and power efficiency
- Biggest challenge for models: memory-efficiency while maintaining high accuracy
- New generation of microcontrollers are optimized for Edge AI applications

Questions & answers


Infineon