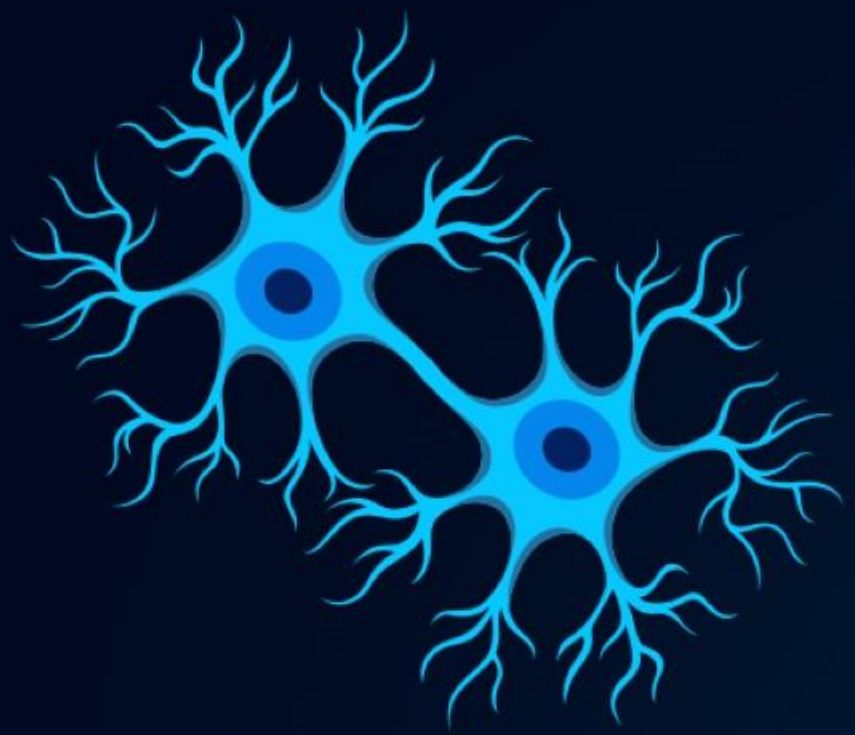


EDGE AI

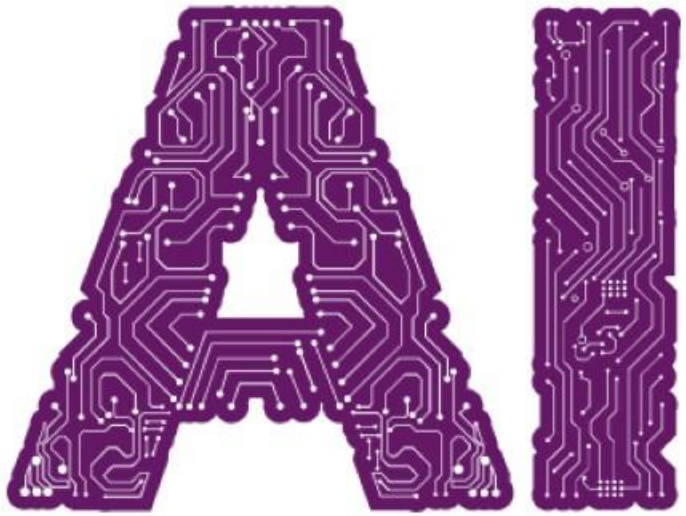
STATE OF PLAY

- Research and innovation

November 25, 2022
EFECS 2022, Workshop on Edge AI, Amsterdam

Dr. Ovidiu Vermesan
Chief Scientist, SINTEF

EDGE AI



Edge AI

Edge AI combines AI, IoT and edge computing technologies and provides real-time collection, processing, analytics and decision-making. The applications that accelerate the edge AI developments are autonomous systems (e.g., autonomous vehicles, internet of robotic things, swarm colonies of things), smart metering, predictive maintenance, video surveillance, intelligent buildings, etc.

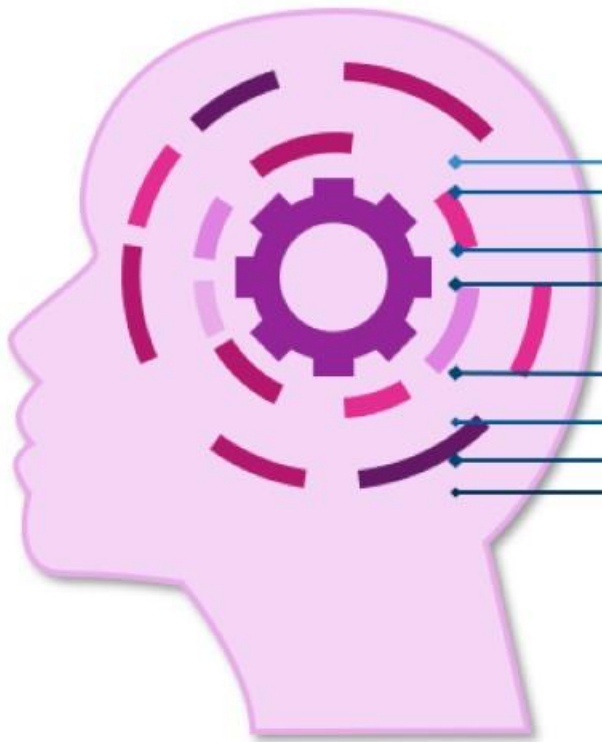
01 Reduces latency, bandwidth by introducing cost-effective and efficient low power HW/SW/AI solutions allowing processing and analysing data locally.

02 End-to-end edge AI architecture from device to system. HW/SW/AI co-design methodologies and the use of multiple toolchains, frameworks and platforms,

03 Verification, validation and testing in heterogenous and dynamic environments.

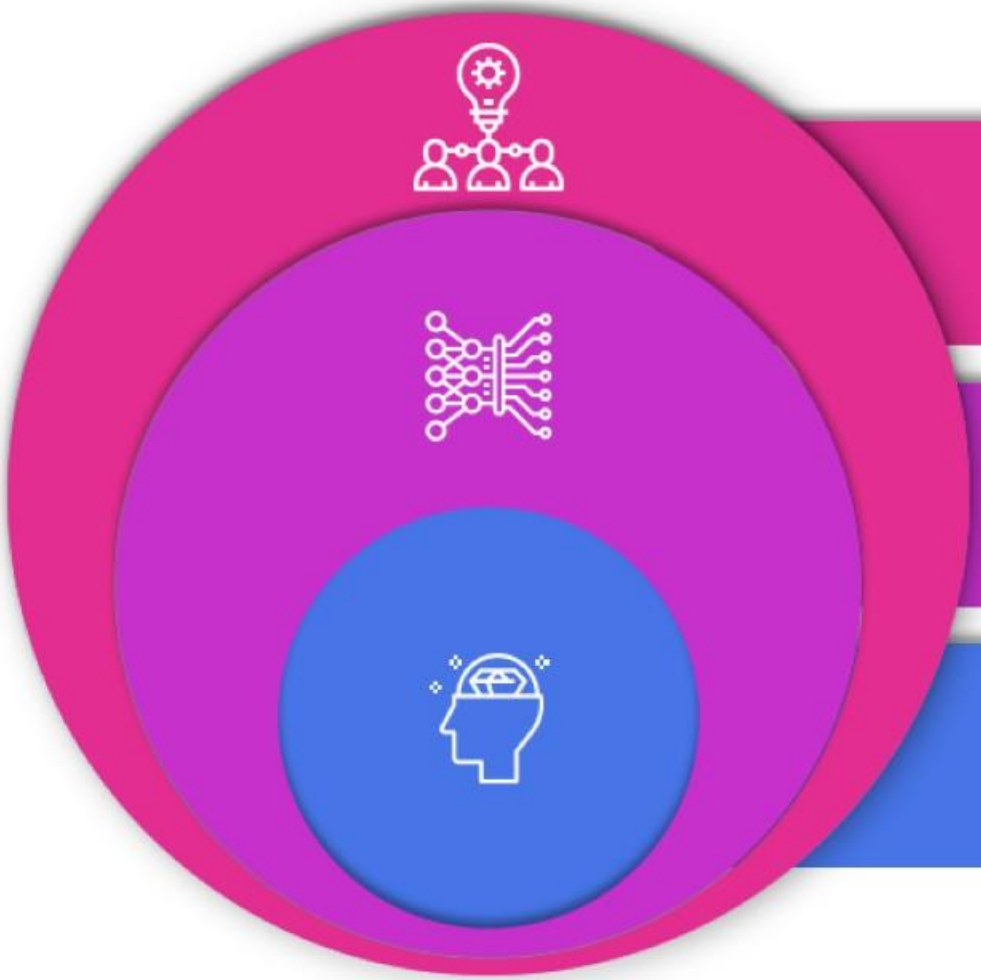
04 Deploy edge AI-powered applications that perform complex tasks in various industrial sectors.

EDGE AI - MULTIDISCIPLINARY ISSUES



- Dynamic, Hybrid and Heterogenous Environments
- Software-Hardware-AI Engineering Design
- Edge AI System and Application Architecture
- Overall Edge AI System Operation
- Edge AI Modeling
- Edge AI Data and Model Operations
- Trustworthy Edge AI: Dependability Engineering
- Trustworthy Edge AI: Engineering Ethics

EDGE AI DOMAINS IN FOCUS



Edge Artificial Intelligence
Technique and methods which enables processing units to mimic inteligent behavior at the edge.

Machine Learning
Subset of AI techniques which use statistical methods to enable devices to improve with experiences.

Deep Learning
Subsets of ML which make the computation of multi-layer neural networks feasible.

EDGE AI HARDWARE SOFTWARE MARKET

Edge AI Hardware Market

- Edge AI hardware market includes components (processor, memory, sensors/actuators, etc.), devices (smartphones, gateways, industrial PLCs, cameras, robots, wearables, etc.), systems (micro-servers, hybrid systems) used in training and inference by users in different industries (consumer electronics, smart home, automotive, government, aerospace, defence, healthcare, industrial, construction, etc.)
- The global edge AI hardware market size valued at **USD 6,88 billion in 2020** and projected to reach **USD 38,87 billion by 2030** - compound growth rate (CAGR) of 18,8%.
- Growth of the global edge AI hardware market driven by emergence of AI coprocessors for edge computing, the advancements in heterogeneous computing architectures, the increased functionalities and features of IoT devices and their applications. The growth is accelerated by increase in real-time low latency on edge devices and major efforts to reduce the power consumption, increase the energy efficiency, reduce the size and cost.

Source: MARKETS and MARKETS 2022, Allied Market Research 2021

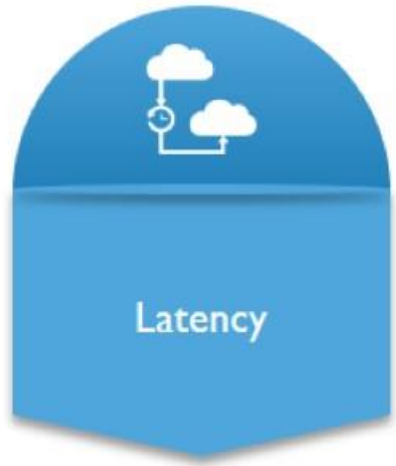
EDGE AI HARDWARE SOFTWARE MARKET

Edge AI Software Market

- Edge AI software market includes components (solution and services), data sources (video- and image recognition, mobile data), organization size, vertical (energy and utilities, manufacturing and life sciences).
- The edge AI software market size is projected to grow from **USD 0,8 billion in 2022** to **USD 3,1 billion by 2027** – CAGR compound growth rate of 28,9%.
- Growth of the global software edge AI market driven by the advances in sensing/actuating, processing capabilities, heterogenous computing architectures and the development of many connectivity solutions (wireless and cellular). The edge AI software developments are not yet mature with the development of many solutions that are not interoperable, scalable and trustworthy. The lack of toolchains and benchmarking methods and techniques is still an issue. The integration of HW/SW/AI co-design requires future efforts.

Source: MARKETS and MARKETS 2022, Allied Market Research 2021

EDGE AI PROCESSING



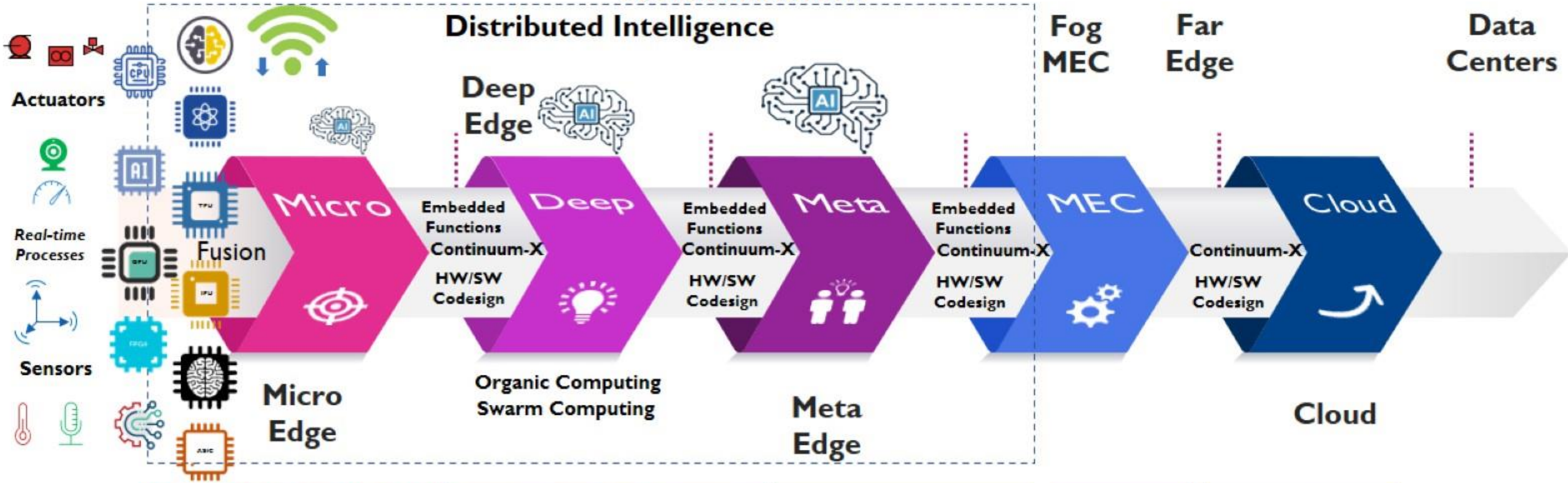
Processing at the Edge

COMPUTING CONTINUUM – INTELLIGENT EDGE GRANULARITY

Scalability, efficiency, adaptability, transparency,
Dependability, trustworthiness.
Over-the-air updates/upgrades.

Applications

OS and Orchestration



Distributed Data Pipeline

DSPs, FGPAs, CPUs, GPUs, ASICs Network Processing Unit (NPU), Intelligent Processing Unit (IPU). Tensor Processing Unit (TPU), Reduced Instr. Set Computer RISC-V, Neuromorphic.

Computing units (industrial processing, panel units, etc.) , network computing units (intelligent routers, switches, gateways and other communications hardware), intelligent controllers (PLCs, RTUs, DCS).

Micro and clustered servers to handle compute intensive tasks / workloads (e.g., high-end CPUs, GPUs, FGPAs, etc.), on premises edge computing, local edge.

Multi-access edge computing (MEC) infrastructure. Fog processing and platforms

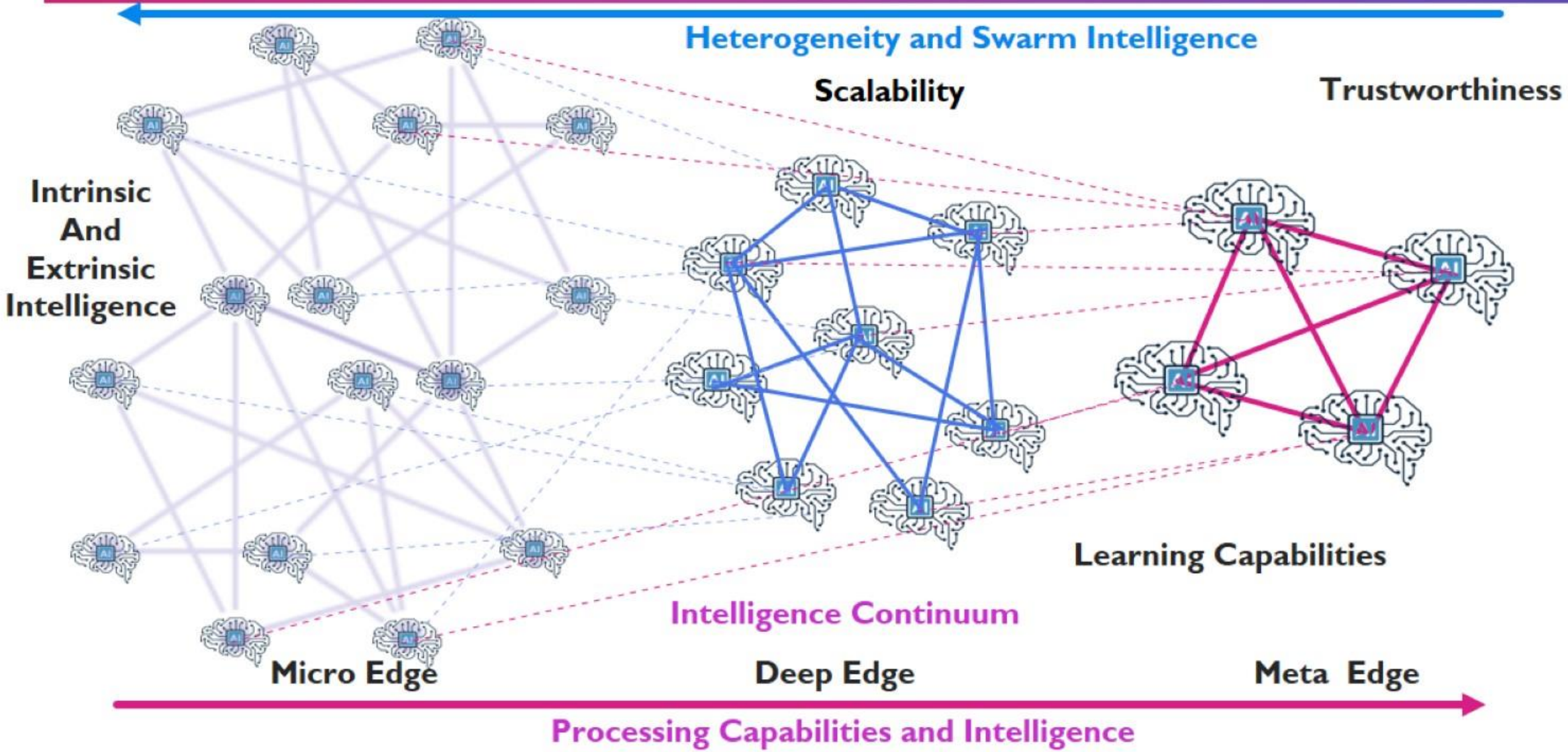
Cloud Infrastructure. Local, regional and national data centres. Federation of clouds and data centres.

O. Vermesan

COMPUTING CONTINUUM

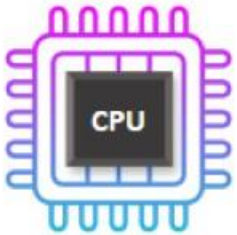
- Processing capabilities, resources, range
  - Micro-edge latency below 1 ms, range from mm to 15 m
  - Deep-edge latency below 2-5 ms, range up to 1 km
  - Meta-edge latency below 10 ms, range up to 50 km
  - MEC latency 10-5 ms, range up to 75 km
  - Fog latency 10-20 ms, range up to 100 km
  - Far-edge latency 20-50 ms, range up to 200 km
  - Cloud and data centers latency 50-100 ms, range up to 1000 km.

DISTRIBUTED INTELLIGENCE

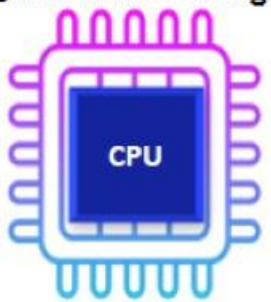


EDGE HARDWARE EVOLUTION

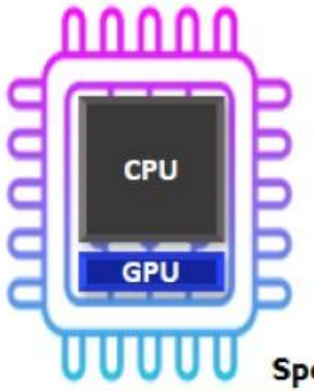
Single core



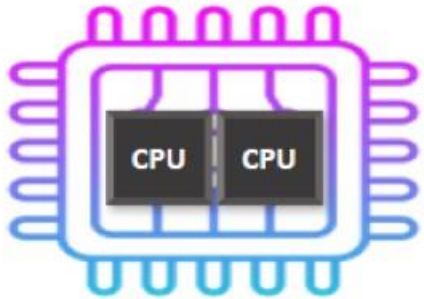
Larger core
(Quantitative change)



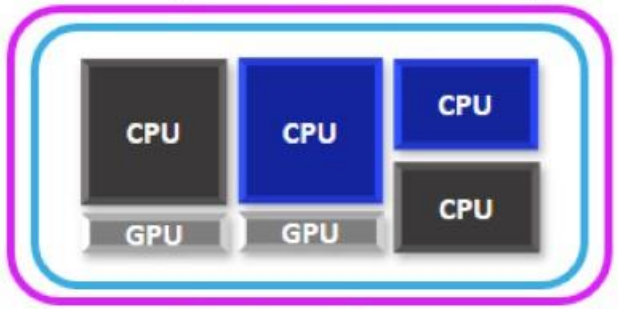
Specialisation



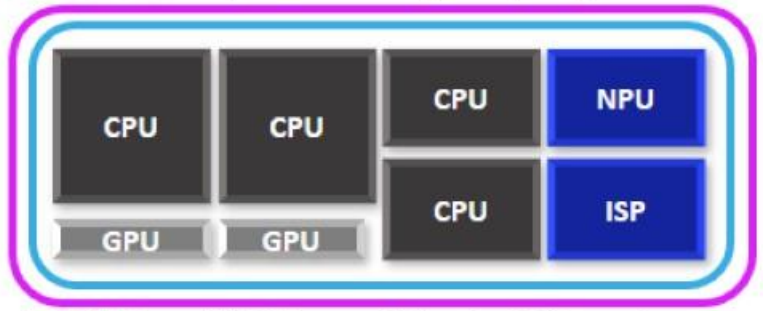
Multi Core



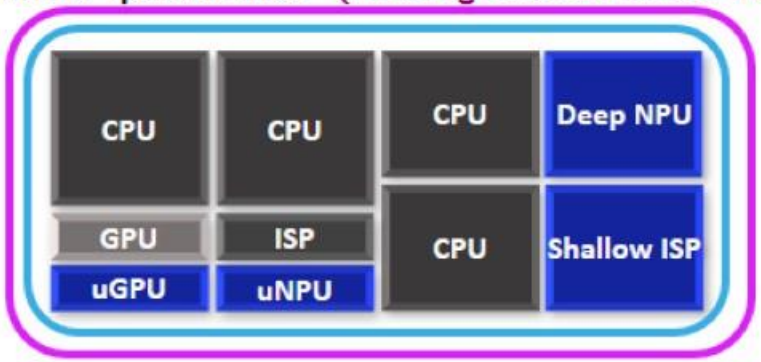
Heterogeneous Multiprocessing



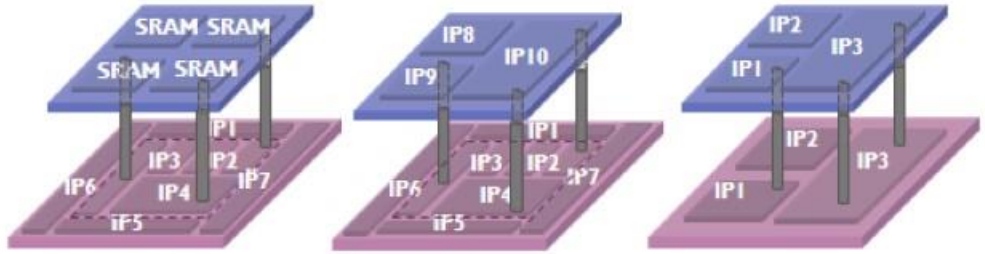
Specialization (NPU, ISP)

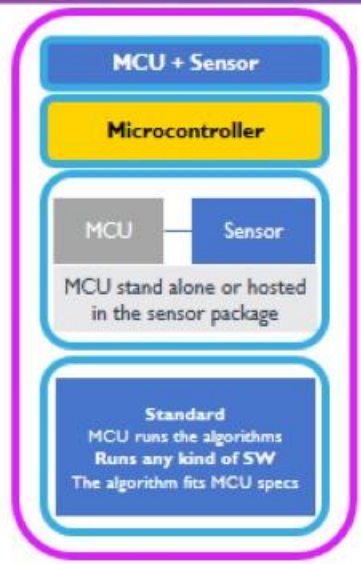


Further specialization - (Heterogenous GPUs and NPUs)

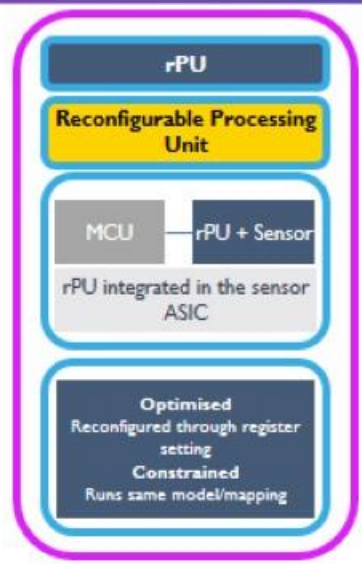


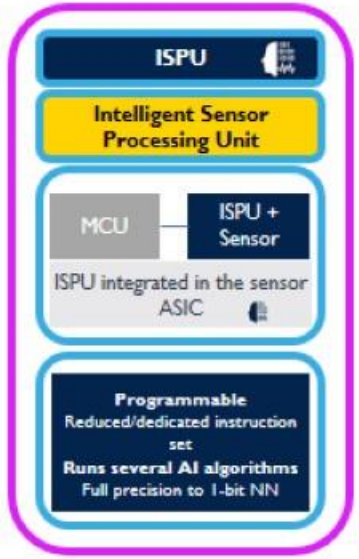
NPU 100x more efficient than CPUs on tasks such convolution and matrix multiplication
Neuromorphic architectures (Spiking NN, Convolutional NN)



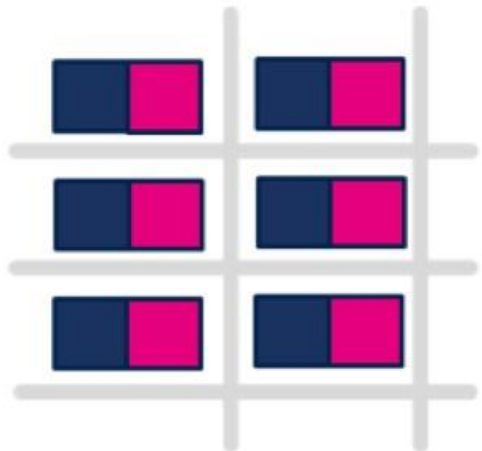


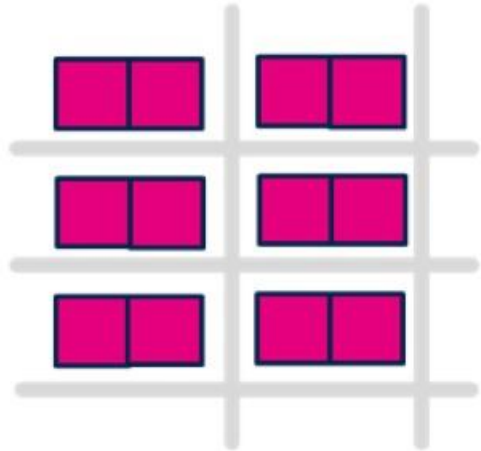
Different Edge AI development platforms, frameworks, workflows for ML, DL on embedded devices across edge continuum (micro-, deep-, meta-edge).





EDGE AI MEMORY PARADIGM





Traditional Computing

Data flows through a single combine bus

0.1 TOPS/Watt

Near-memory Computing

Shorter distance for data flow

9.6 TOPS/Watt

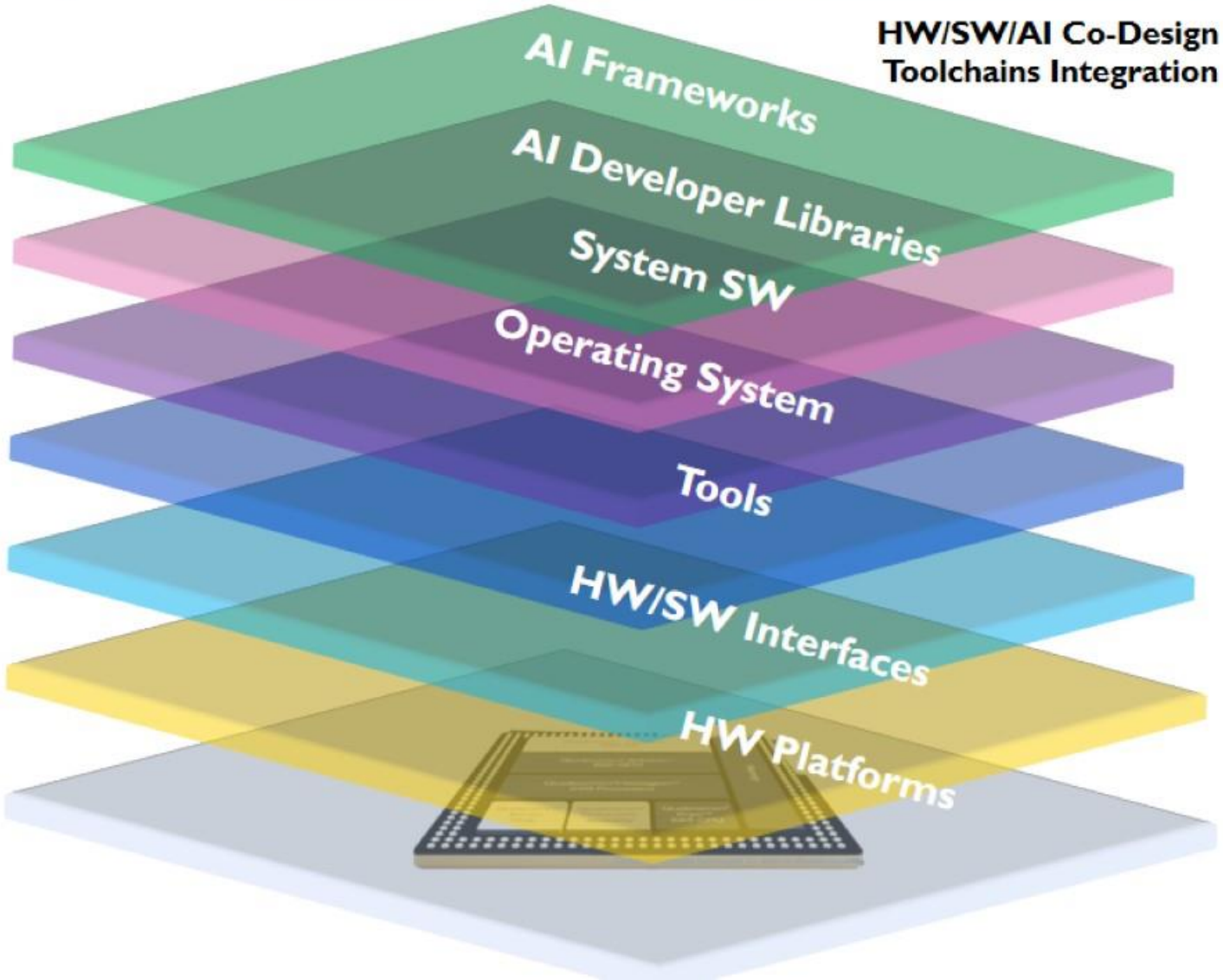
In-memory Computing

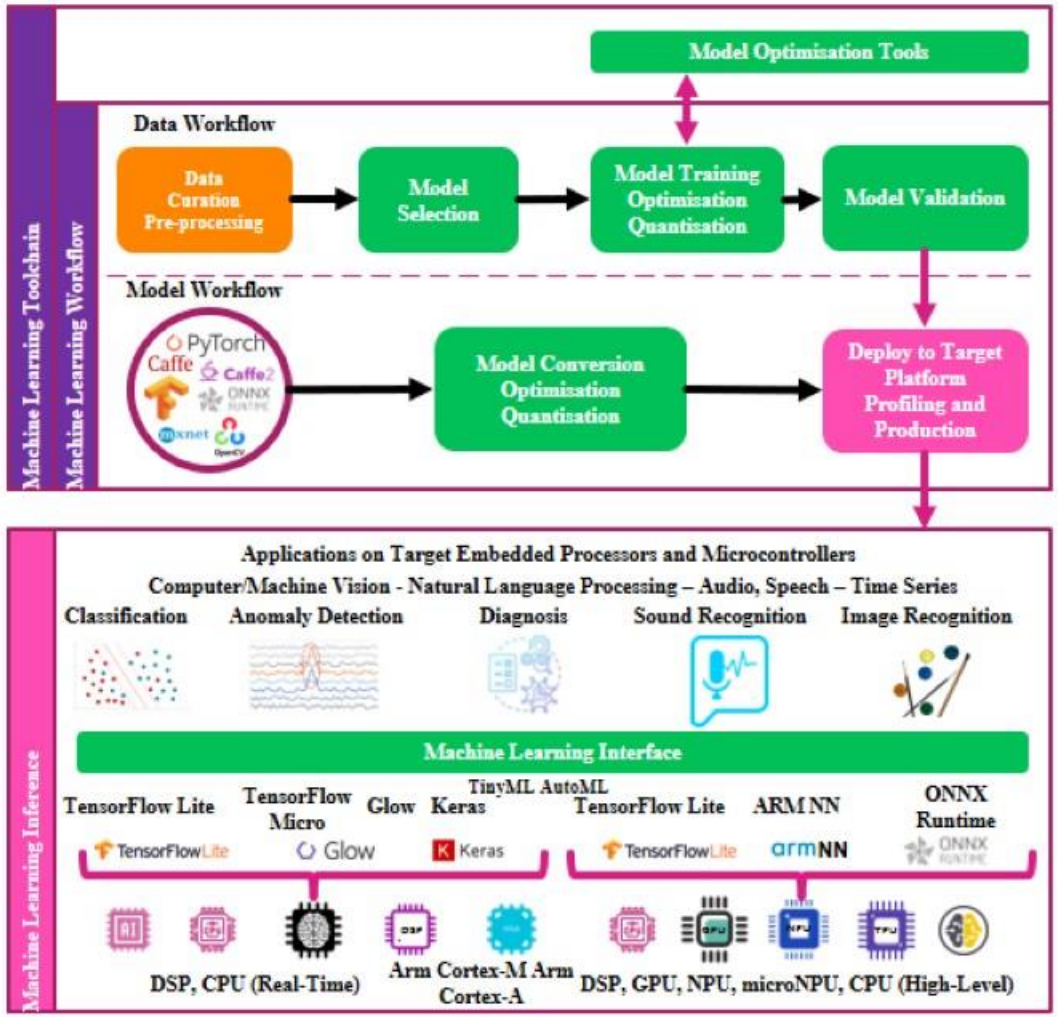
The compute and memory functions are merged

28 TOPS/Watt

TOPS - tera operations per second

EDGE AI HW/SW DESIGN AND INTEGRATION

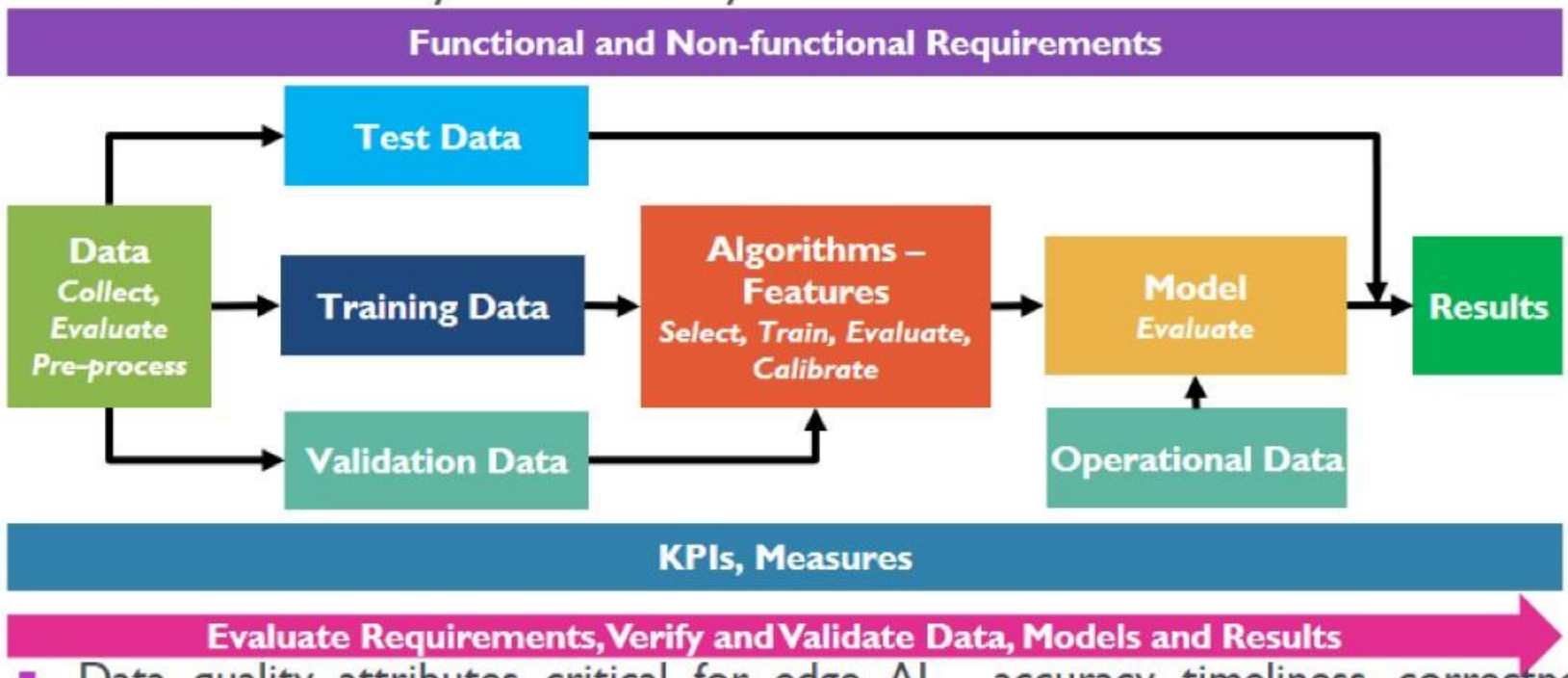


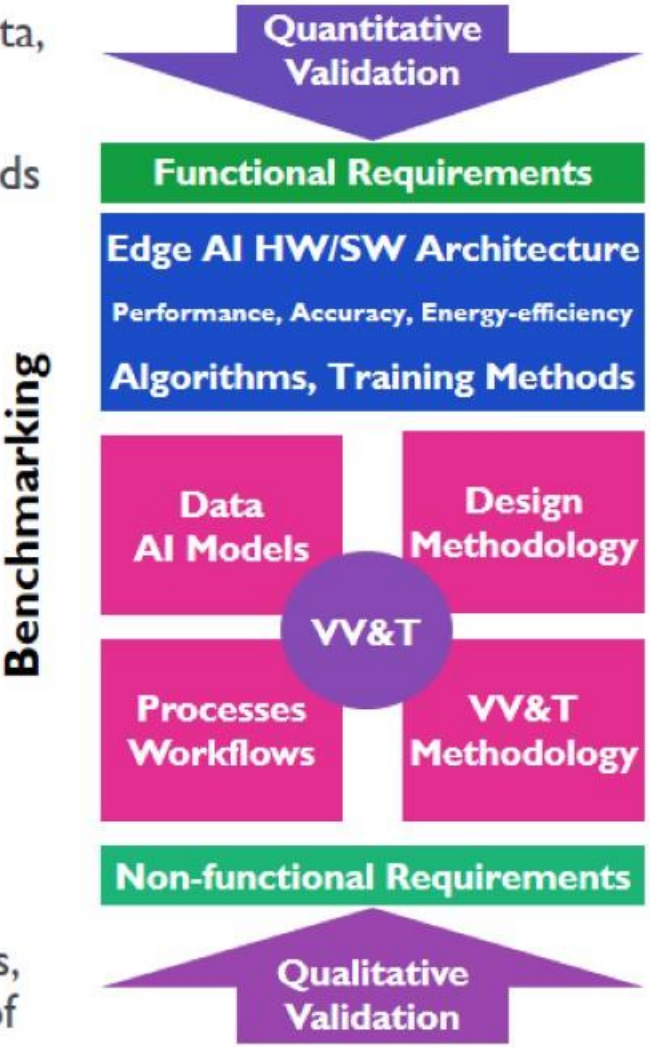


AI and Math libraries, compilers, virtual platforms, profilers and debuggers, programming languages, core libraries, system interfaces, SoC, SoM, accelerator drivers, emulation support

VERIFICATION, VALIDATION, TESTING EDGE AI SYSTEMS

- Edge AI verification validation and testing require approaches and solutions at data, model and system level beyond those for Cloud AI and conventional systems.
- Edge AI life cycle workflow requires to combine the SW/HW engineering methods with the data and system level analysis.

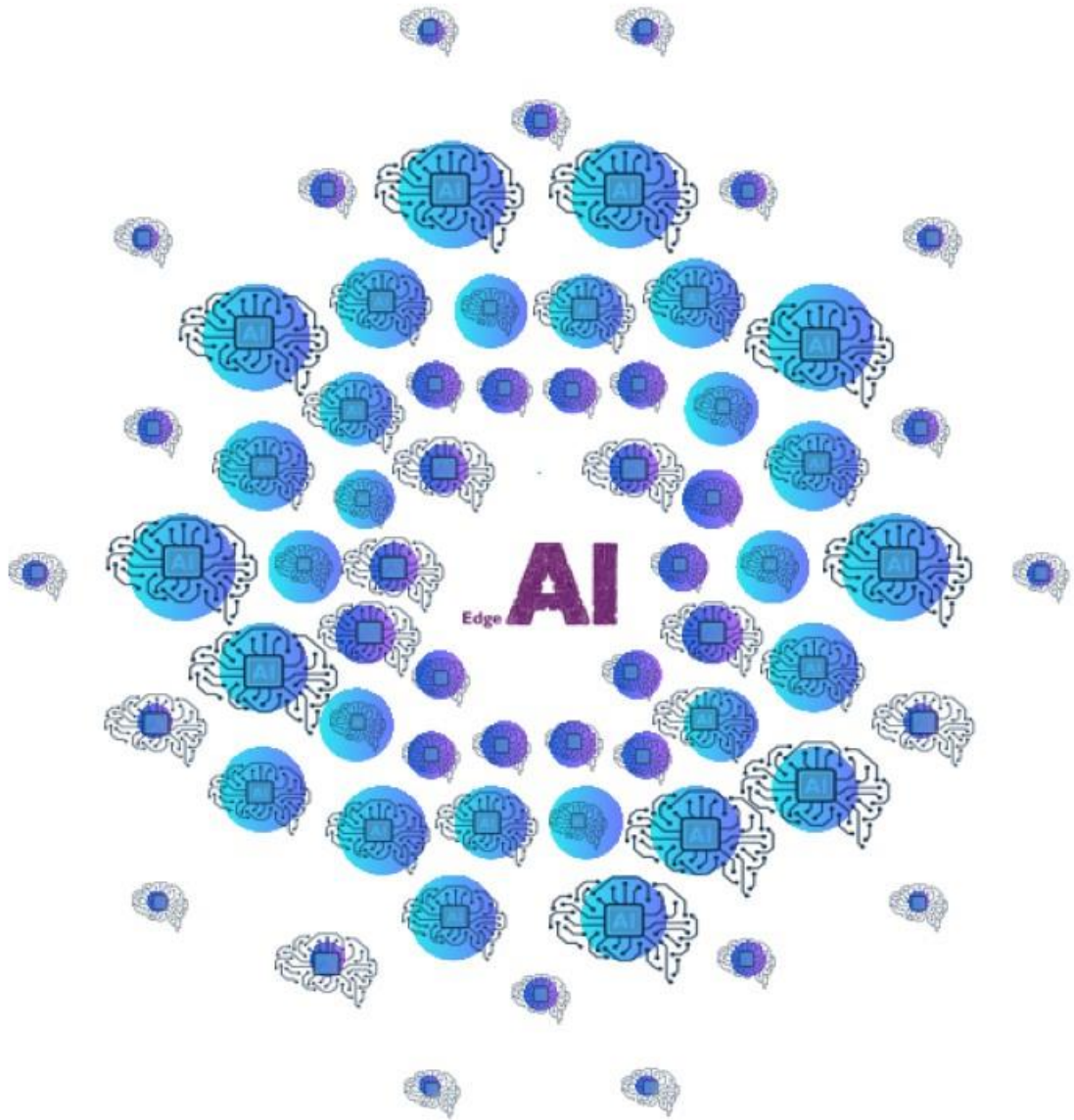




- Data quality attributes critical for edge AI - accuracy, timeliness, correctness, consistency, usability, security, privacy, accessibility, accountability, scalability, lack of bias. The attributes are part of a larger edge AI non-functional requirements set.

TRENDS & CHALLENGES

- Self-organising edge AI systems
- Self-replicating edge AI systems that enables a set of AI-based algorithms to produce other tools.
- Improving edge AI solutions accuracy rates on reduced datasets.
- Edge AI ML models trained on reduced datasets and optimised to accomplish specific tasks or maximise specified reward functions.
- New edge AI methods for increasing the capacity for generalising or transferring learning from a training task to a new one.
- Edge AI energy-efficient federated learning.
- New heterogeneous edge AI energy-efficient hardware architectures and circuits combining new computing paradigms,
- Edge AI HW/SW/AI integrated tools, interoperable frameworks and platforms,
- Edge AI collective intelligence- combining intrinsic and extrinsic intelligence solutions,
- Self-supervised learning, deep reinforcement learning for edge AI
- Edge AI meta-learning, continual-learning and multitask learning



THANK YOU

Ovidiu.Vermesan@sintef.no