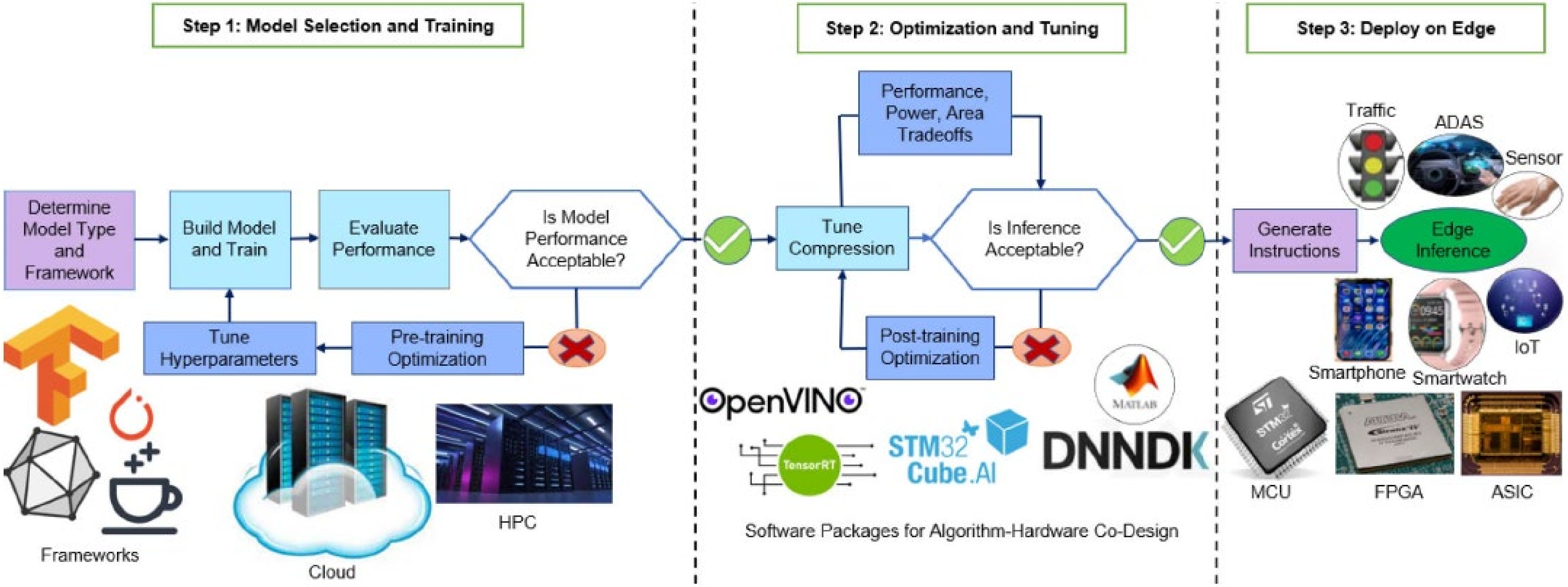
Uncited images created by the author using [Adobe Illustrator]

INF2009 – Edge Computing and Analytics [2024/25 T2]

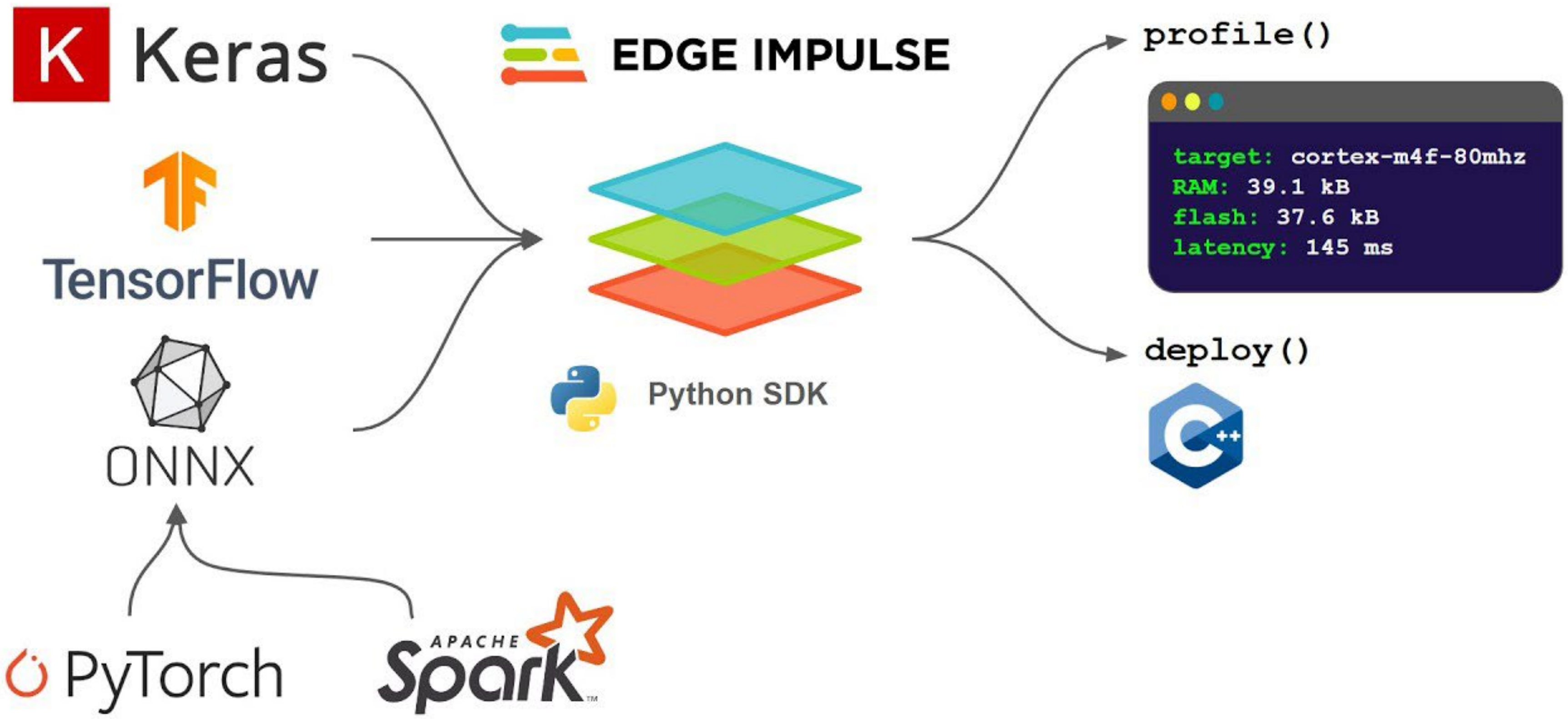
Edge Analytics – Tools, Challenges and Future Perspective

[Revisit] Training to Inference Framework



Shuvo, Md Maruf Hossain, Syed Kamrul Islam, Jianlin Cheng, and Bashir I. Morshed. "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review." Proceedings of the IEEE 111, no. 1 (2022): 42-91.

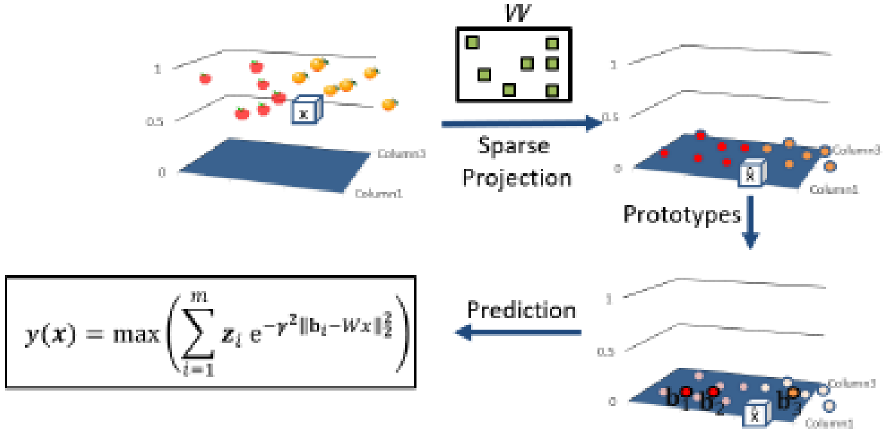
Tool for End→End Edge Computing



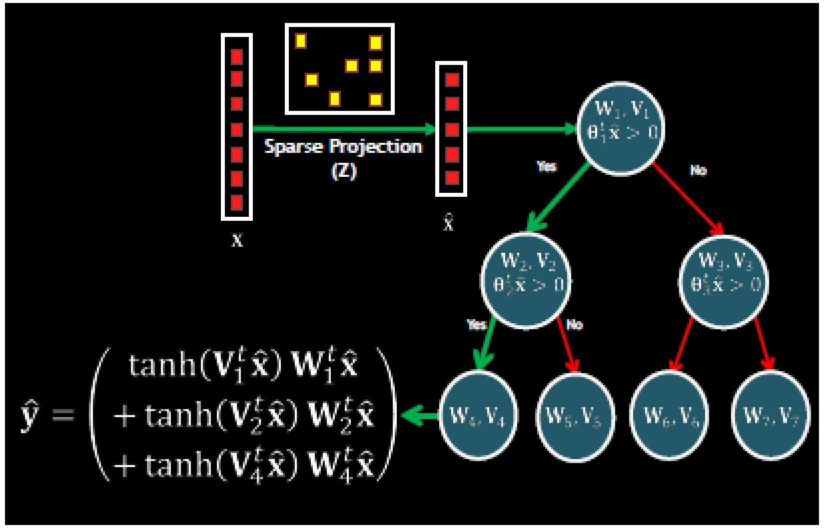
[Getting Started with the Edge Impulse Python SDK]()

(Software) Tools for Edge Analytics

ProtoNN - light weight (<16kB)
k-nearest neighbors (kNN)



Bonsai – light weight (<2 kB)
Regressor



Gupta, Chirag, Arun Sai Suggala, Ankit Goyal, Harsha Vardhan Simhadri, Bhargavi Paranjape, Ashish Kumar, Saurabh Goyal, Raghavendra Udupa, Manik Varma, and Prateek Jain. "Protonn: Compressed and accurate knn for resource-scarce devices." In International conference on machine learning, pp. 1331-1340. PMLR, 2017.

Kumar, Ashish, Saurabh Goyal, and Manik Varma. "Resource-efficient machine learning in 2 kb ram for the internet of things." In International conference on machine learning, pp. 1935-1944. PMLR, 2017.

https://github.com/Microsoft/EdgeML/wiki/Algorithms

(Software) Tools for Edge Analytics

MicroPython puts an implementation of Python 3.x on a microcontroller or embedded system

TensorFlow Lite, a mobile library for deploying models on mobile, microcontrollers and other edge devices

PyTorch

ExecuTorch is an end-to-end solution for enabling on-device inference capabilities across mobile and edge devices including wearables, embedded devices and microcontrollers.

ONNX

ONNX makes it easier to access hardware optimizations

https://micropython.org/
https://www.tensorflow.org/lite
https://onnx.ai/
https://pytorch.org/executorch-overview

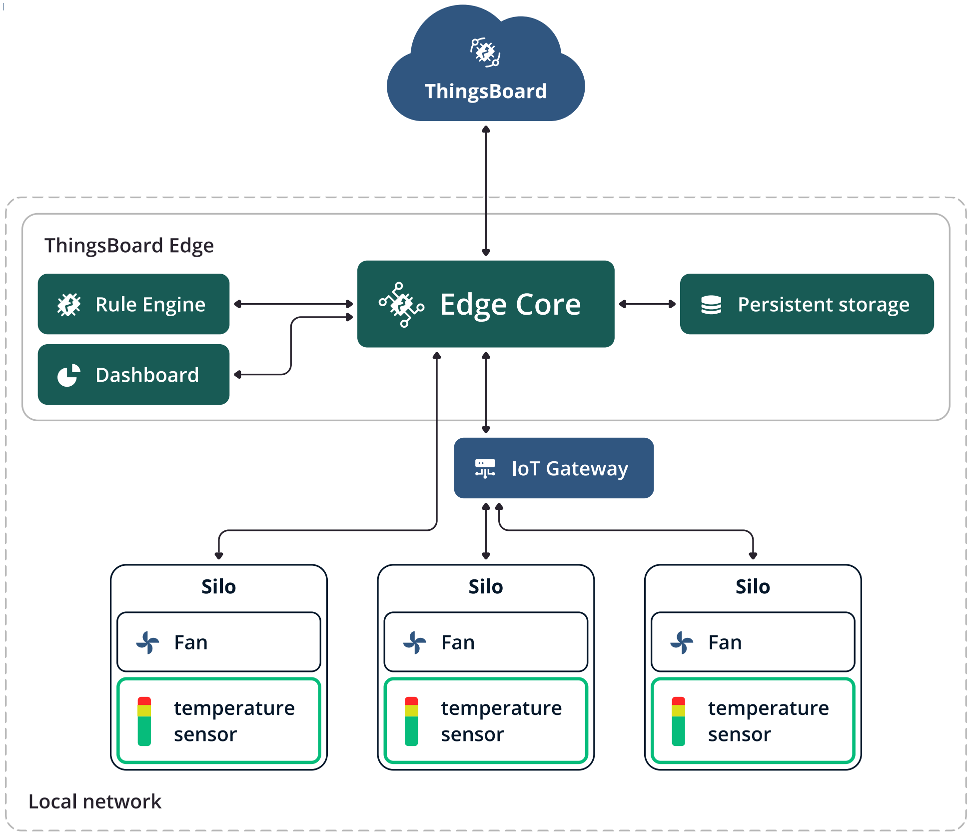
(Software) Tools for Edge Computing

Software for Edge Inference	Supported Frameworks	Supported Edge Devices
Intel OpenVINO Toolkit	Caffe, TensorFlow, ONNX	Intel® CPU, Integrated Graphics, Neural Compute Stick 2, Movidius™ VPUs, and FPGAs
Matlab Deep Learning HDL Toolbox	Kerns, TensorFlow	Xilinx Zynq®-7000 ZC706, UltraScale+TM MPSoC ZCU102, Intel Arria® IO SoC
XCUBE-AI	Keras, TensorFlow Lite, ONNX standard format	STM32 Arm® Cortex®-M-based MCU
AMD (XILINX) DNNDK	Caffe and TensorFlow	Xilinx® Zynq®-7000 and Zynq UltraScale+™ MPSoC
NVIDIA TensorRT	TensorFlow, MATLAB, ONNX	Tesla P4, Tesla VI00, Drive PX2, Jetson TX2, NVIDIA DLA
CEVA Deep Neural Network (CONN)	Caffe, TensorFlow, ONNX	CEVA-XM Vision Processor, NeuPro, and SensPro
Qualcomm® Neural Processing SDK	Caffe/Caffe2, TensorFlow ONNX	Qualcomm® Snapdragon mobile chips (Hexagon™ DSPs, AdrenoTM GPUs, Kryo™ CPUs)
Cadence Stratus HLS	TensorFlow, Caffe	RTL/FPGA
Embedded Learning Library (ELL)	Microsoft CNTK, Darknet, ONNX	Raspberry Pi, Arduino, micro:bit

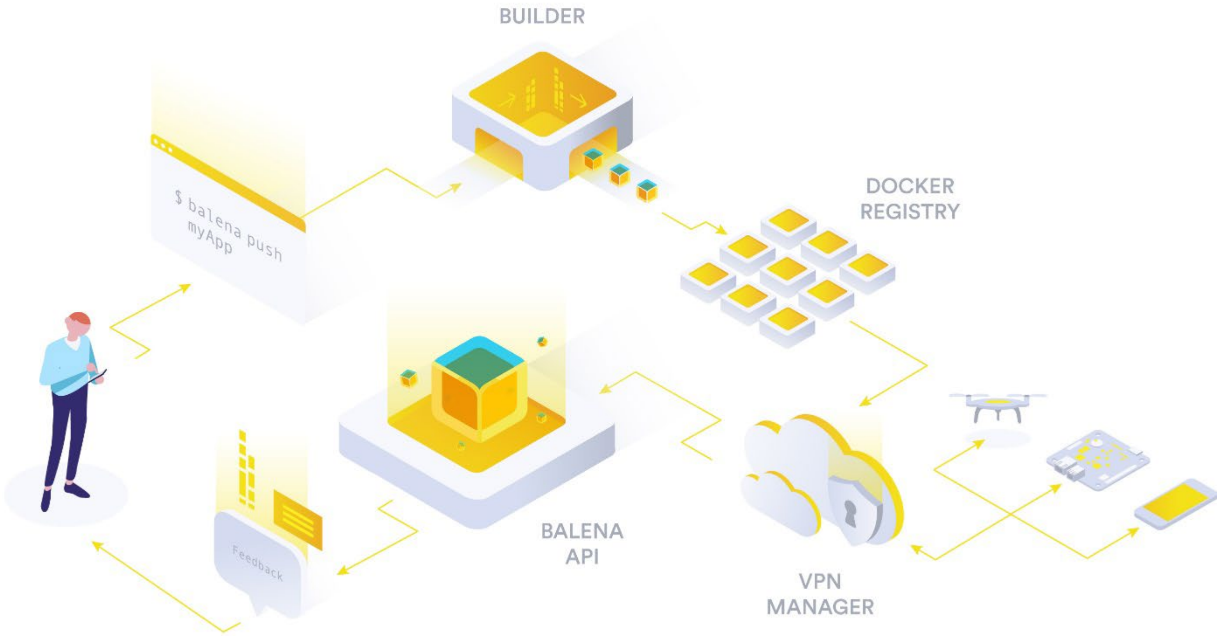
Shuvo, Md Maruf Hossain, Syed Kamrul Islam, Jianlin Cheng, and Bashir I. Morshed. "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review." Proceedings of the IEEE 111, no. 1 (2022): 42-91.

Edge Fleet Management

Thingsboard



Balena



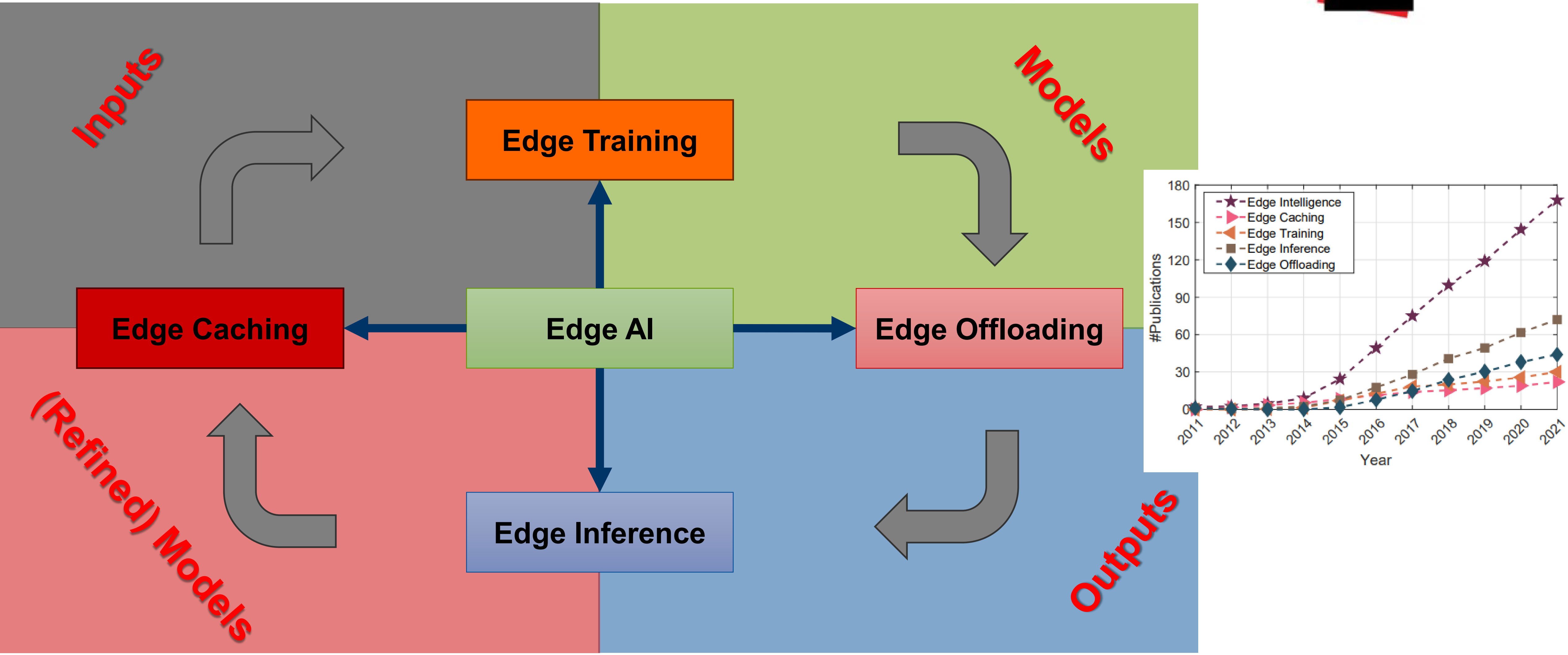
https://thingsboard.io/products/thingsboard-edge/
https://docs.balena.io/learn/welcome/primer/

Challenges/Opportunities for Edge Computing

- ✓ Adaptability to Data Heterogeneity
  - ❑ Robustness to sensing environments → Edge specific augmentations
- ✓ Automatic Mapping of DL to Hardware
  - ❑ Available tools for mapping are less efficient (Edge Impulse is playing a role here)
- ✓ Developing Benchmarks
  - ❑ Proper benchmark datasets and models are required
- ✓ Automatic, Joint, and Edge Aware Compression
  - ❑ Developing an automatic compression technique
- ✓ Algorithm–Hardware Codesign
  - ❑ Neural Accelerators to Handle Sparsity
- ✓ Neural Architecture Search for Edge Inference
  - ❑ NN architecture tuned to specific hardware (e.g., ProxylessNAS is an option)
- ✓ Training on the Edge
- ✓ Increased Demand of Communication Resources
- ✓ Explainability in Edge Inference

Shuvo, Md Maruf Hossain, Syed Kamrul Islam, Jianlin Cheng, and Bashir I. Morshed. "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review." Proceedings of the IEEE 111, no. 1 (2022): 42-91.

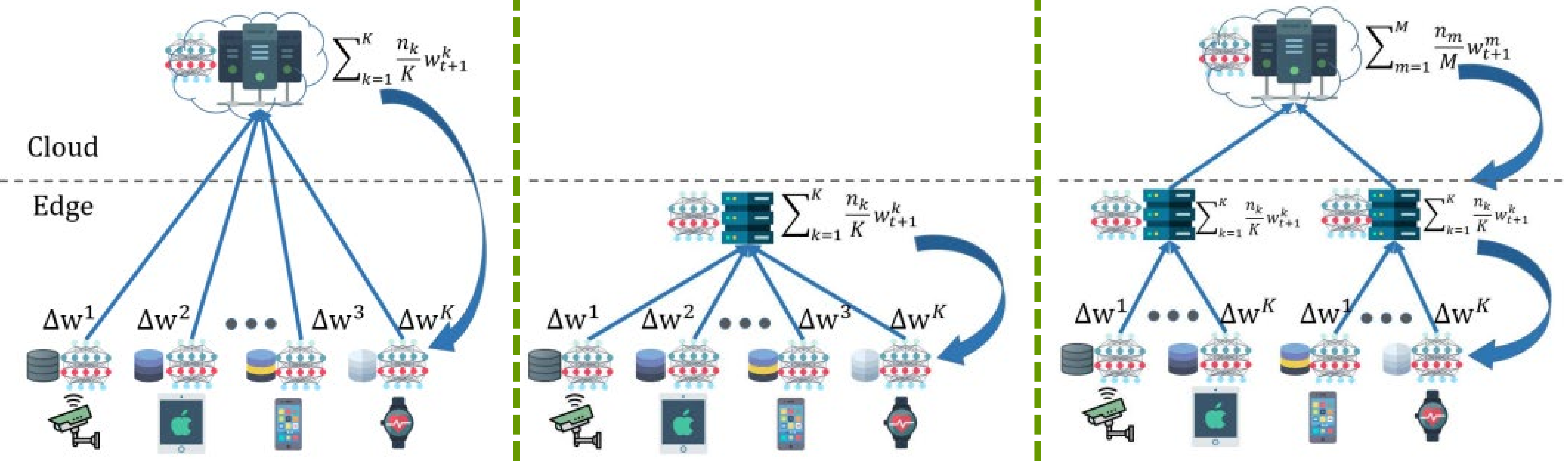
Future of Edge Computing - Edge AI



Xu, Dianlei, Tong Li, Yong Li, Xiang Su, Sasu Tarkoma, Tao Jiang, Jon Crowcroft, and Pan Hui. "Edge intelligence: Empowering intelligence to the edge of network." Proceedings of the IEEE 109, no. 11 (2021): 1778-1837.

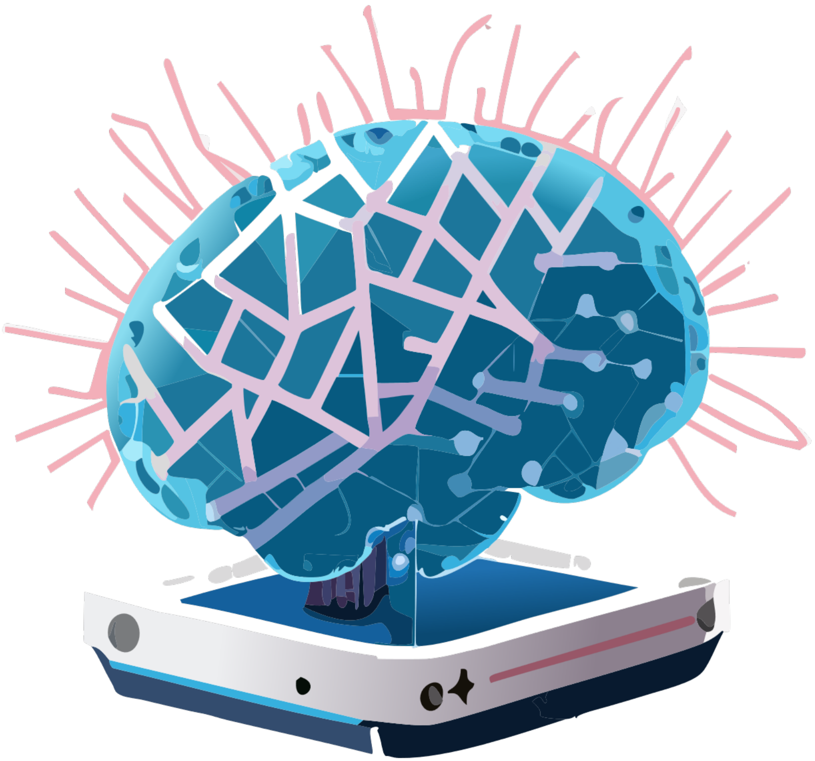
Future of Edge Computing – Edge-Cloud Learning

Cloud-based Federated Learning | **Edge-based Federated Learning** | **Hierarchical Federated Learning**



Xu, Dianlei, Tong Li, Yong Li, Xiang Su, Sasu Tarkoma, Tao Jiang, Jon Crowcroft, and Pan Hui. "Edge intelligence: Empowering intelligence to the edge of network." Proceedings of the IEEE 109, no. 11 (2021): 1778-1837.

Summary



- *Edge computing and analytics is a key component in the future of AI*

- *Hardware Architectures, Software Frameworks and Communication Technologies are key to the success of Edge AI*

Contact: mahesh.panicker@singaporetech.edu.sg