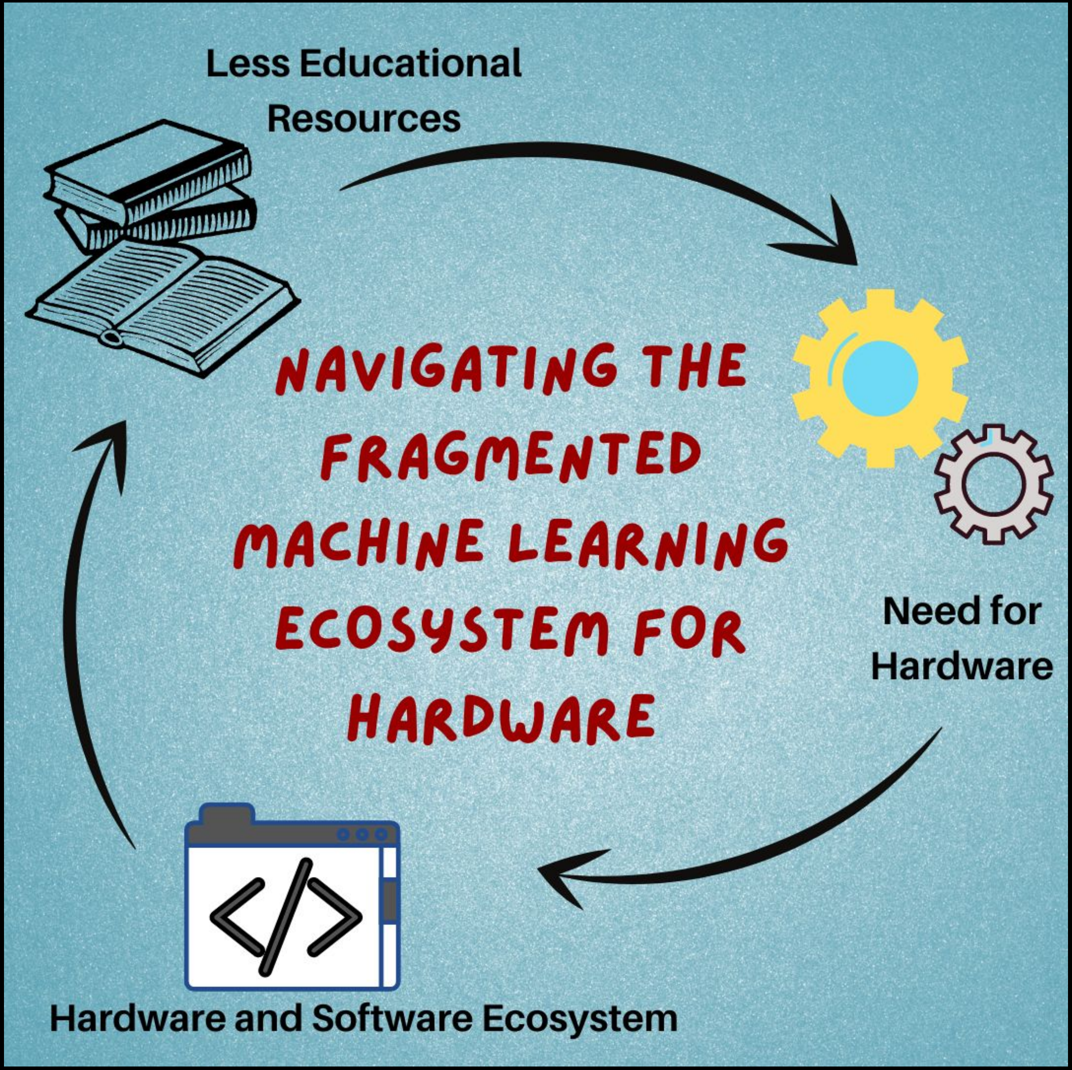
Navigating the Fragmented Machine Learning Ecosystem for Hardware Devices

Vaidheeswaran Archana



NAVIGATING THE FRAGMENTED MACHINE LEARNING ECOSYSTEM FOR HARDWARE

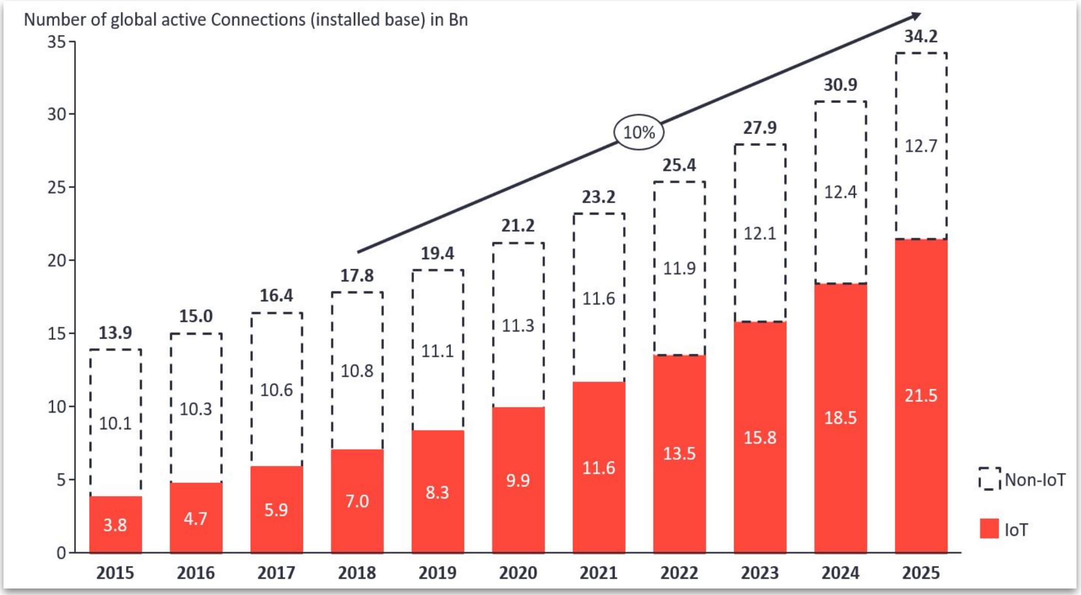
Less Educational Resources

Need for Hardware

Hardware and Software Ecosystem

Internet of Things

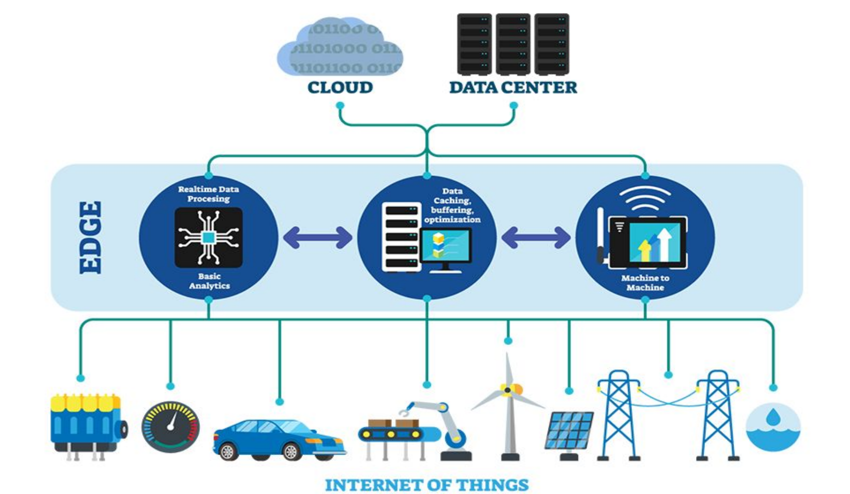
- IoT devices have a microcontroller and sensors
- They can collect and transfer data, as well as perform simple tasks
- IoT devices have become popular over the last few years
- They are used as smart home and fitness devices, but are also being used in industrial settings
- However **99% of IoT sensor data is discarded**



State of IoT; iot-analytics.com

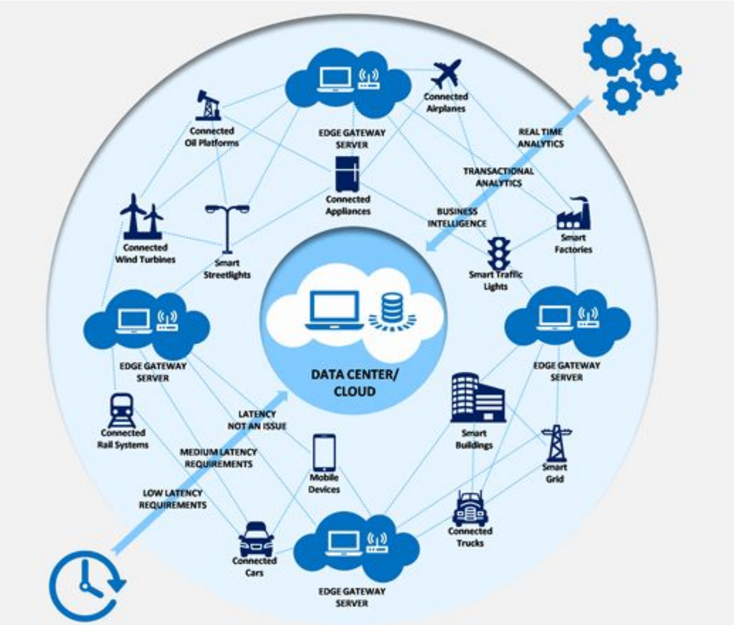
Internet of Things and Edge Computing

- Most DL models are deployed on cloud servers
- Running models closer to where the data is being generated is called edge computing



Edge Computing Use Cases; innovationatwork.ieee.org

Edge Computing



1– Increase in IoT devices causes an increase in cloud dependency

2-Need edge devices which have their own data centres

3-The computing that is performed in those data centers is Edge computing

4- Application: Security Cameras, Self Driving Cars

Advantages of Edge Computing

- Reduced Latency
- Reduced Internet Bandwidth
- Increased Security
- Reduction in Dependence on Cloud Services

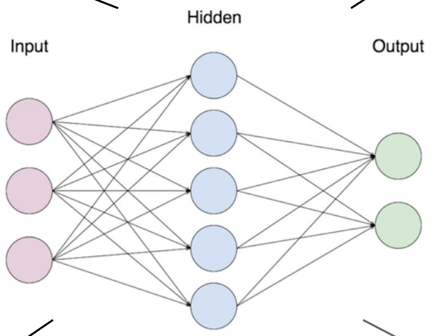
Edge Computing Hardware

- **Microprocessors**
  - General Purpose
  - More Powerful
  - Consume More Power
- **Microcontrollers**
  - For Simpler Tasks
  - Less Computationally Powerful
  - Consumes Less Power
- **Accelerators**
  - Specialised for one task
  - Smaller and more power efficient
  - Computationally powerful, but can only perform a specific task

Need for Edge Computing Algorithms

<u>Power constraints:</u> Neural networks require massive amount of computational power and energy to execute on CPU and GPUs

<u>Composed of Floating Point Values:</u> Neural Networks are generally trained to preserve accuracy and not speed



<u>Memory Constraints:</u> Laptops and PC come with at least 4GB of RAM whereas Raspberry Pi 3 has 1GB of RAM

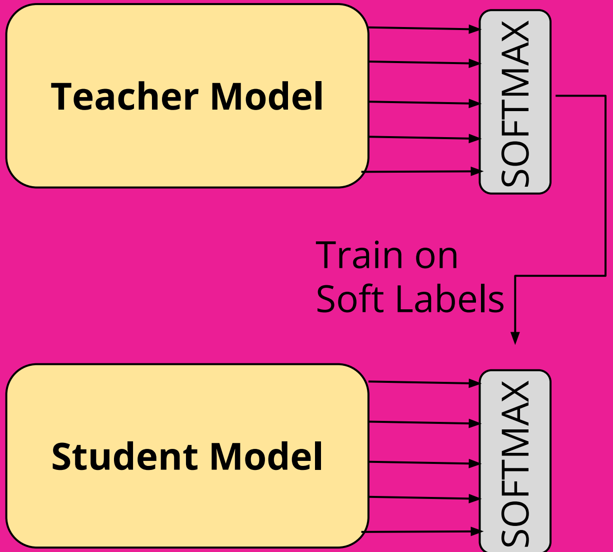
<u>Inference Efficiency:</u> We need model that takes less time for inference

Edge Computing Frameworks

- TensorFlow Lite
  - Quantization
  - Pruning
  - Weight Clustering
  - Support for EdgeTPU
- OpenVINO
  - Quantization
  - Intermediate Representations
  - Support for NCS and other Intel Hardware
- Others
  - ONNX
  - PyTorch
  - DeepStream

Knowledge Distillation

- Works by transferring the knowledge learned by a large teacher model to a smaller student model
- The student model is easier to execute
- The student model can be trained with unlabelled data
- Student models can often achieve similar or more accuracy than the teacher



Teacher Model
SOFTMAX
Train on Soft Labels
Student Model
SOFTMAX