Introduction to Edge AI

Ireland AI Roadshow 2025

George Dickey

Global Technology Leader

George Dickey

Global Technology Leader High-end Processing & AI

- Deep technical background
  - Digital ASIC design
  - FPGA
  - Embedded Software & DSP
  - Artificial Intelligence

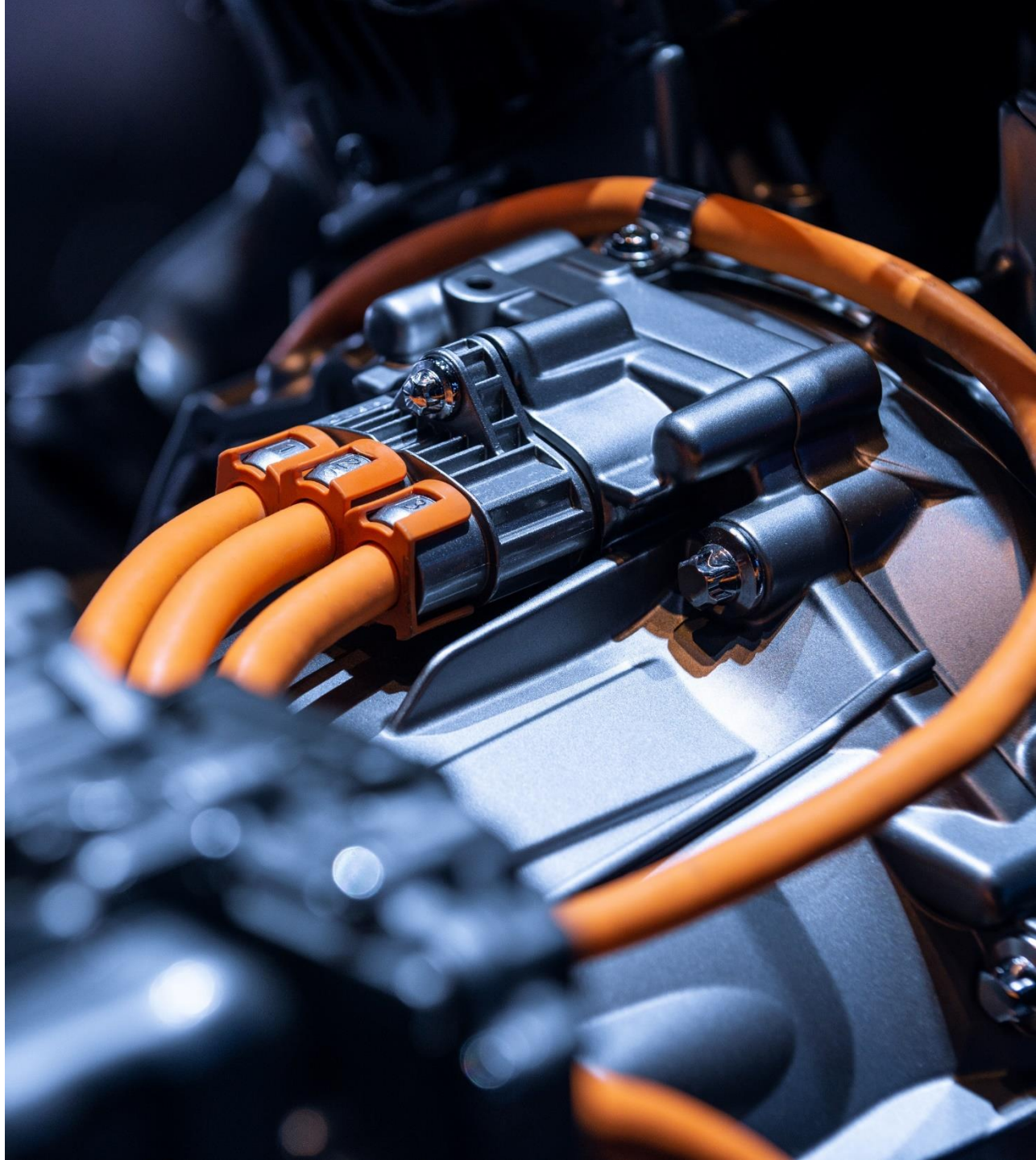
- Across multiple roles
  - Designer
  - Engineering Manager
  - CTO

- In different scaled organisations
  - Independent Consultant
  - SME
  - Multinational

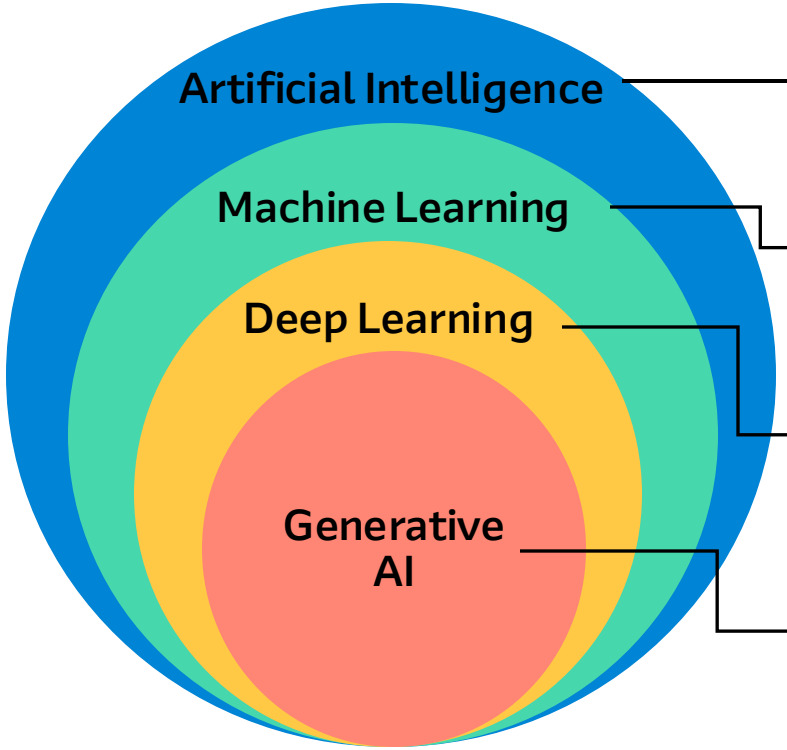
Agenda

Artifical Intelligence

- What is AI and why use it?
- Why at the Edge?
- Device Selection
- Final thoughts



What is an AI model?



Solves a task that mimics human behavior

Input → Analyze → Output

Trains a computer to solve a specific AI problem

Data → Stats & Math Algorithms → Prediction

Uses neural networks to solve an AI problem

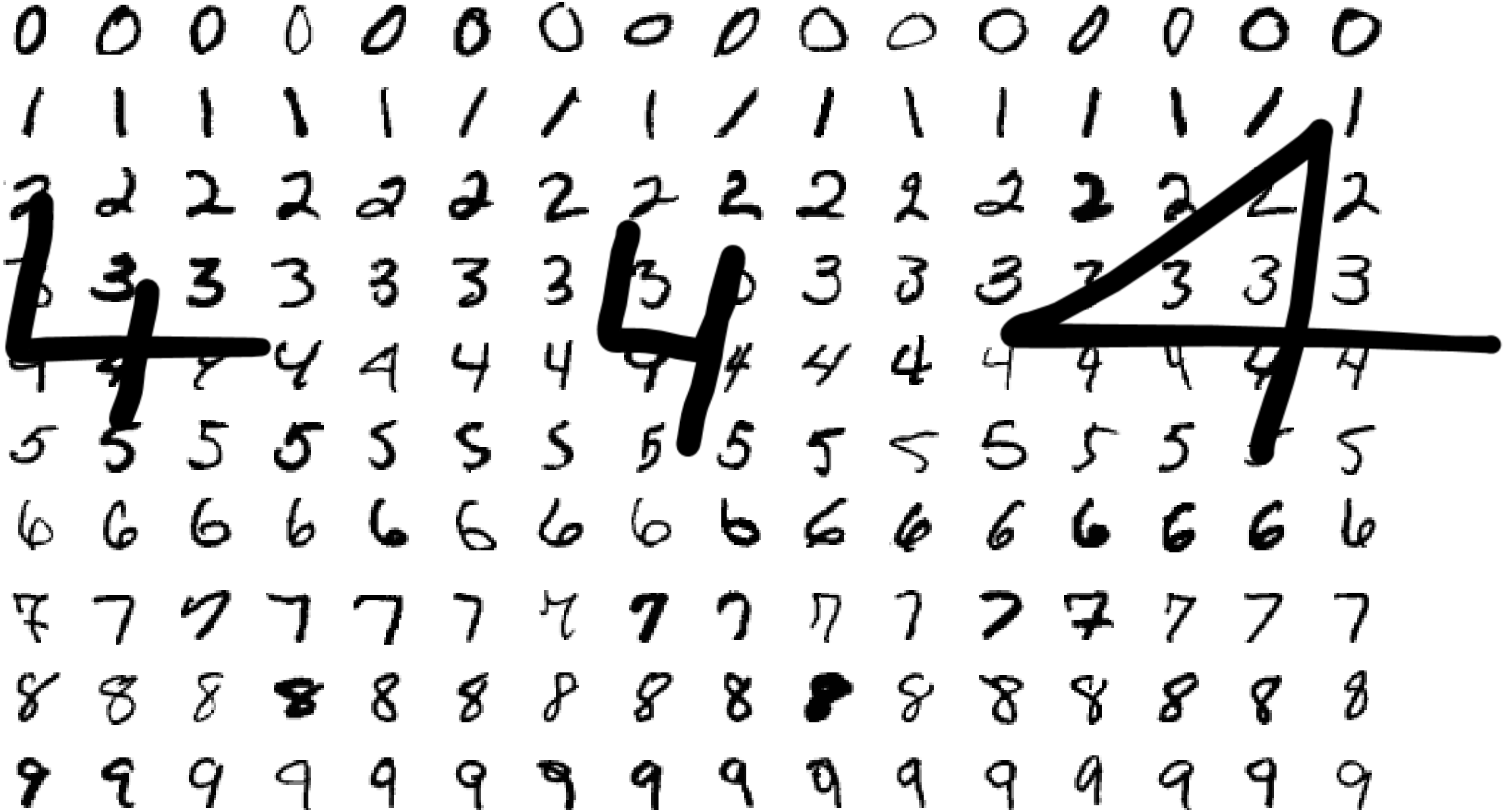
Data → Neural Networks → Prediction

Learns patterns and trends from the training data using neural networks, understands the context to create new content that mimics human generated content

Data (very large dataset) → Neural Networks → Understand Context and Generate New and Unique Content

Edge AI is primarily Deep Learning

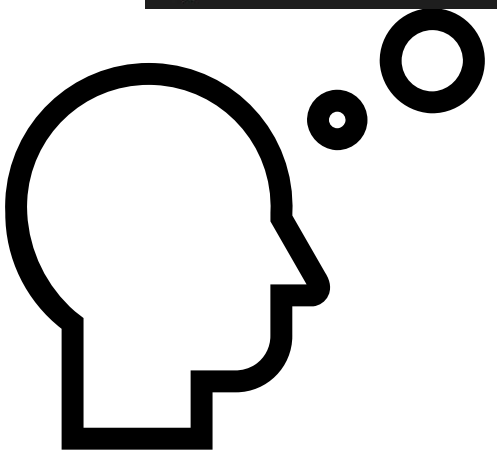
Why use AI?



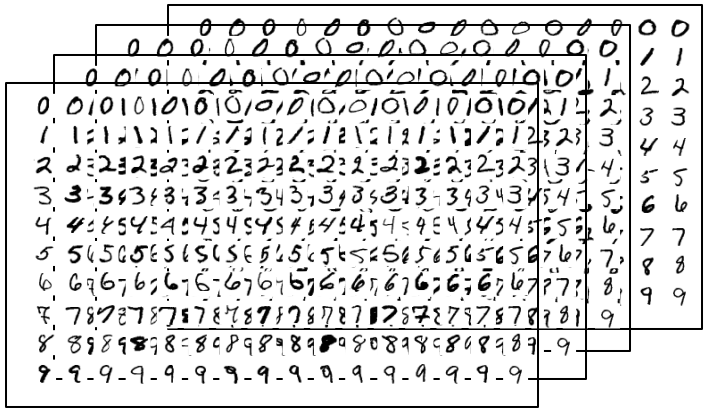
Design Engineer

```
ublic class StarMap extends PApplet

    public void setting
    {
        size(800, 800);
    }

    public void setup()
    {
        loadData();
    }

    public void loadDat
    {
        Table table = l
    }
```

```
loadTable(String filename) : Table
loadTable(String filename, String opti
loadImageFormats : String[]
looping : boolean
launch(String... args) : Process
lerp(float start, float stop, float am
lerpColor(int c1, int c2, float amt) :
lerpColor(int c1, int c2, float amt, i
listFiles(File base, String... options
listFiles(String path, String... optio
listPaths(String path, String... optio
loadBytes(File file) : byte[]
```



Deep Learning

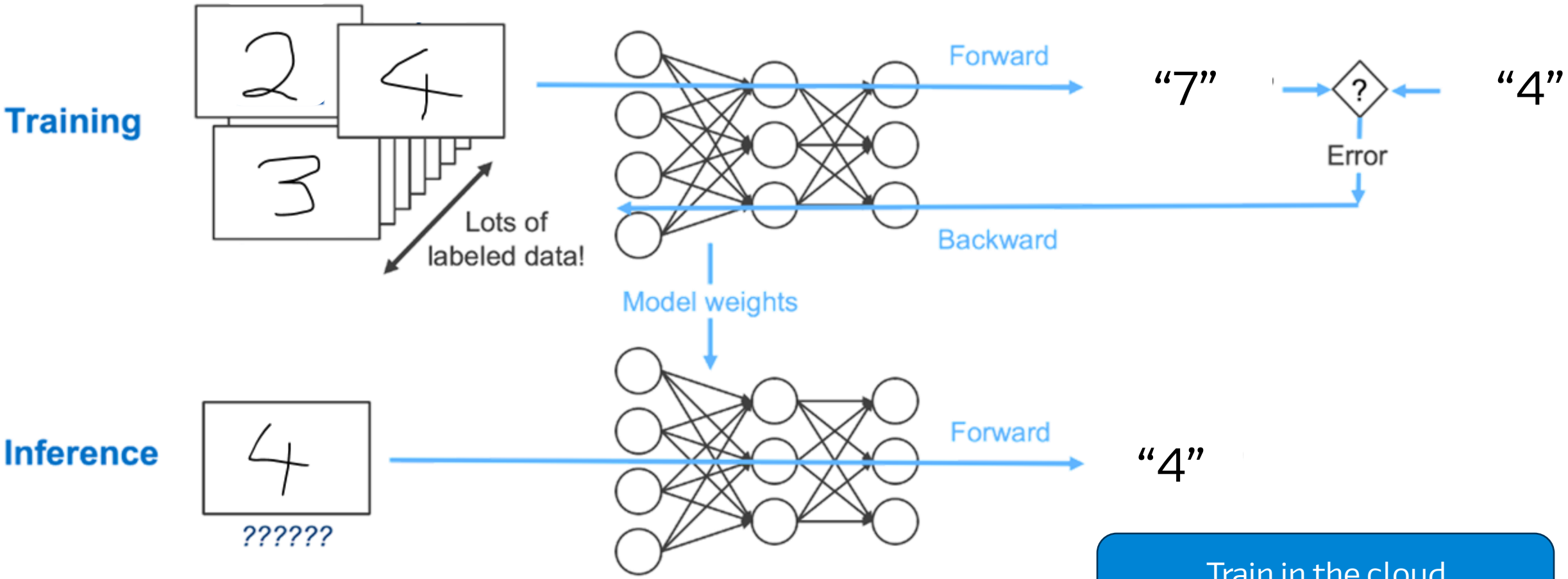


Data

Compute

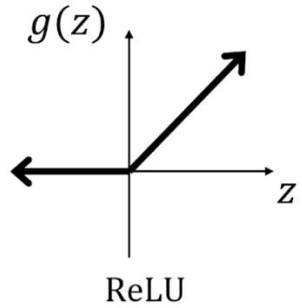
Deep Learning solves complex problems by using lots of data and compute resources rather than human derived algorithms

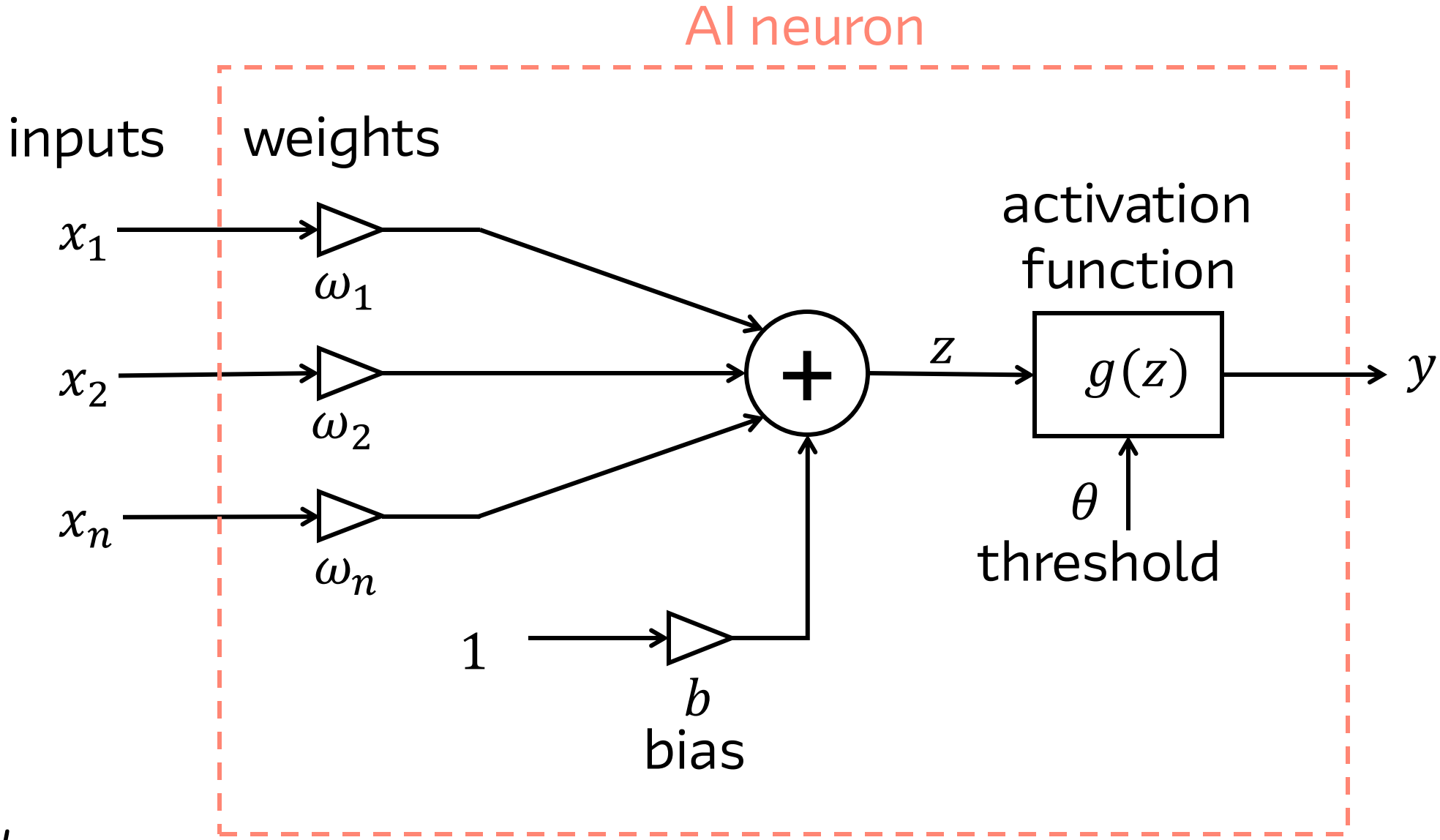
Training and Inference



Train in the cloud
Infer at the edge

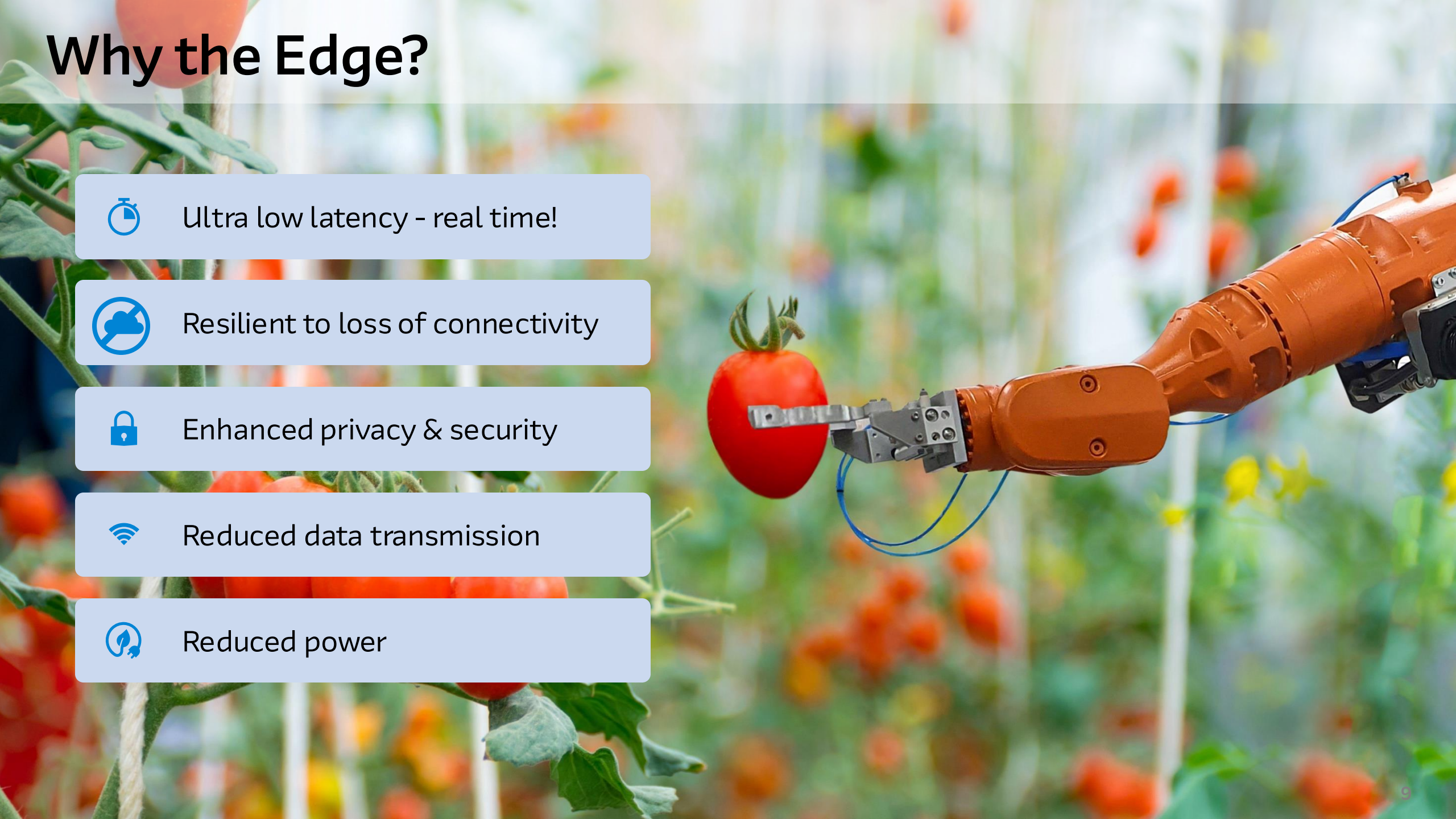
It's all just maths



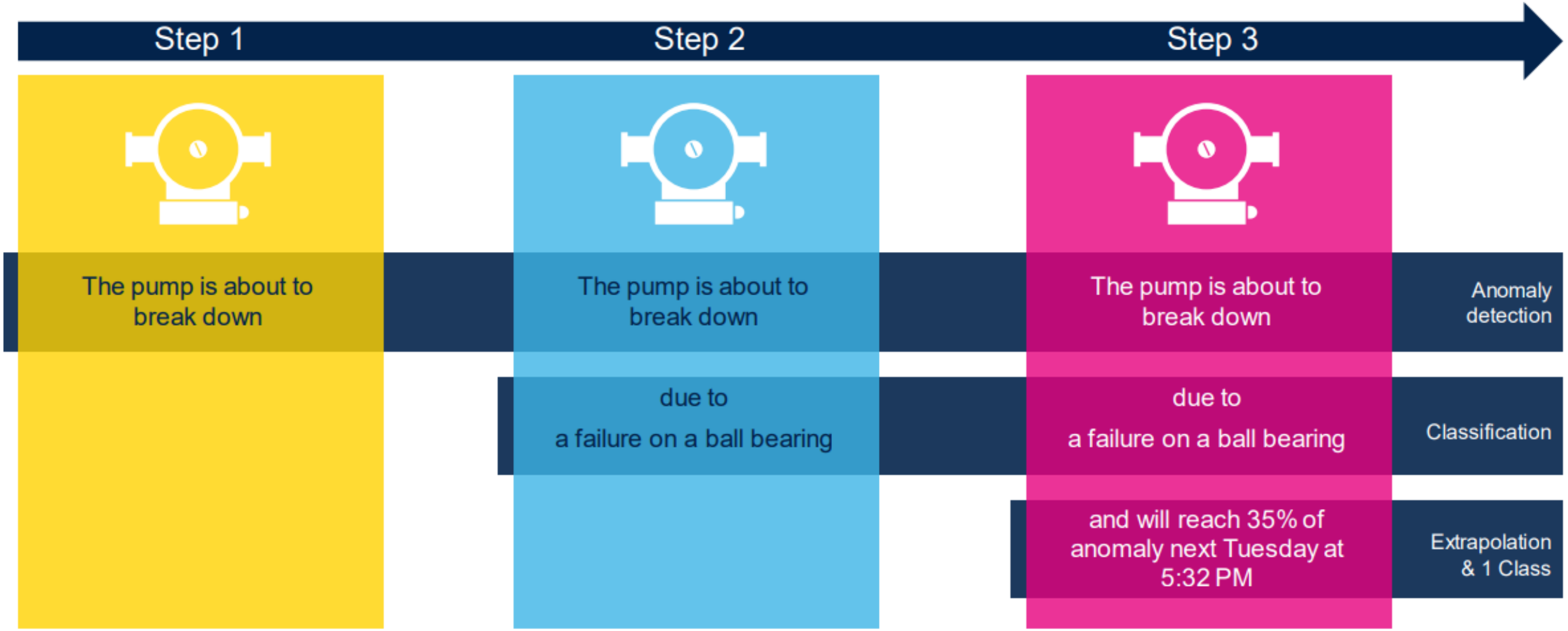


Why the Edge?

- Ultra low latency - real time!
- Resilient to loss of connectivity
- Enhanced privacy & security
- Reduced data transmission
- Reduced power

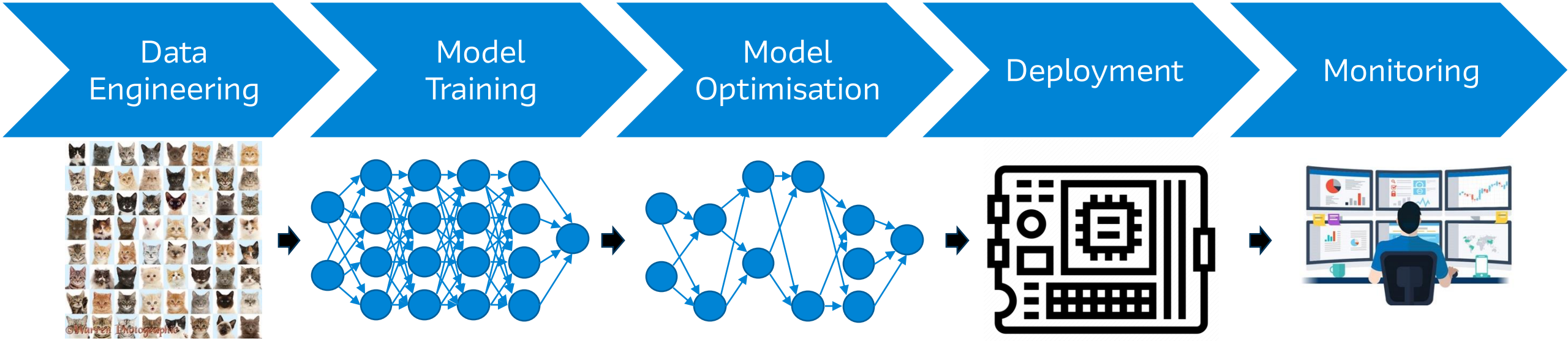


Predictive Maintenance



AI/ML development Flow

The Five steps to realize AI



Specify, collect, clean and prepare data for model training and testing

Selecting, training and testing the model

Optimise the model for the edge, to improve performance and reduce footprint

Deploying and running the model on HW platform

Monitoring the model for drift, basis and accuracy

Dataset Collection

Open problem

- Environment, factors beyond control
- Unknown, various situations or objects
- Difficult tail cases

Closed problem

- Controlled environment
- Well-defined, known objects
- Easier to address tail cases

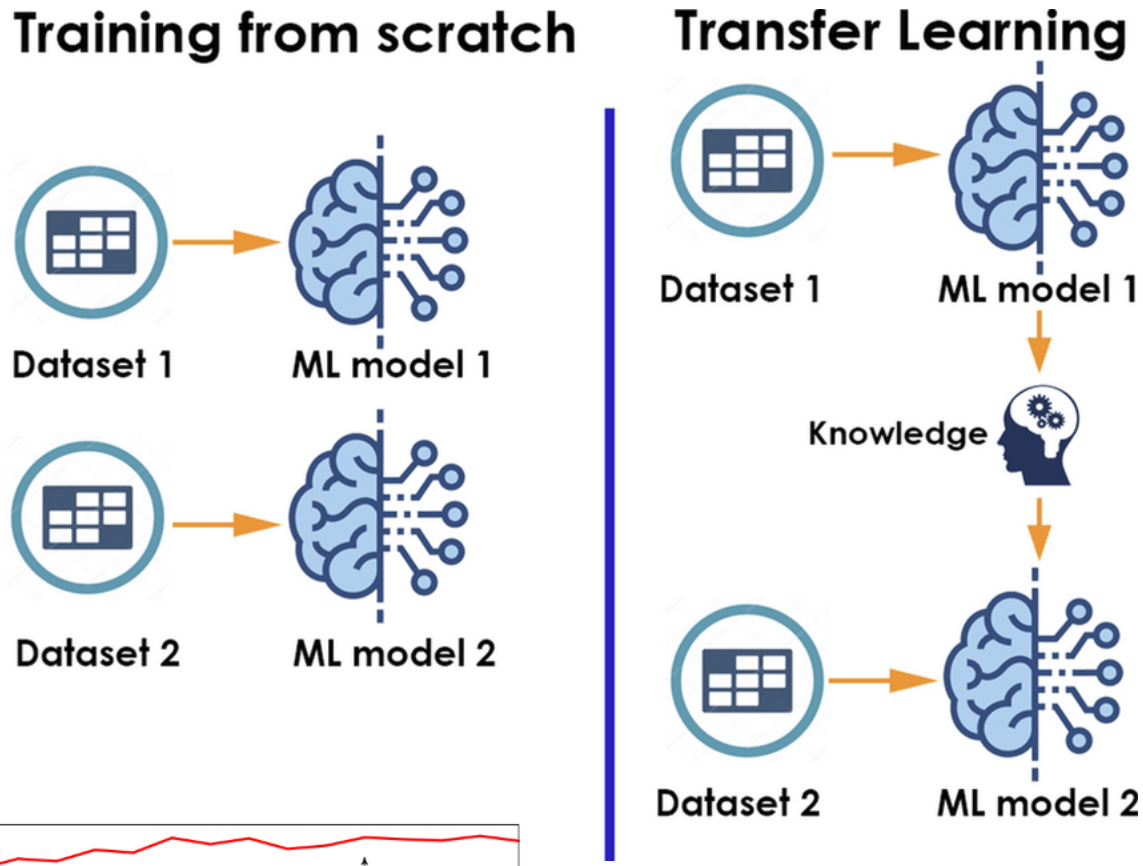
Transfer Learning

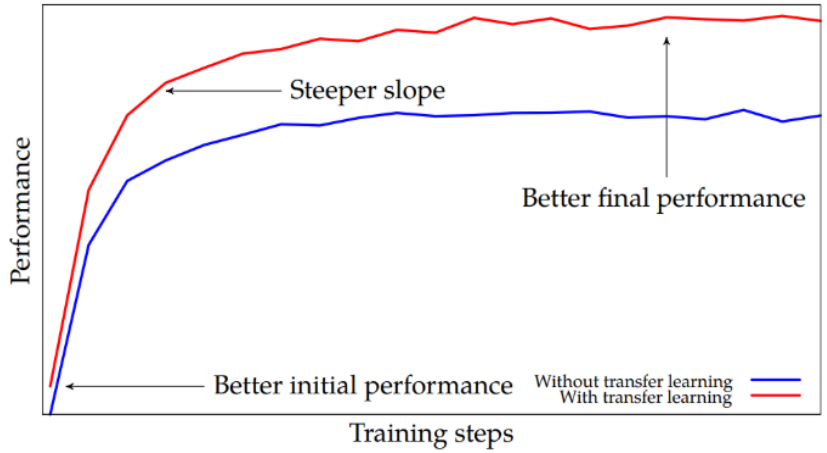
Why

- Reduced Training Time
- Improved Performance
- Data Efficiency

How

- Start with a pre-trained model (**baseline**)
- Lock all but the last few layers (or add additional layers)
- Re-train the unlocked layers of the mode
- Unlock and fine-tune the whole model



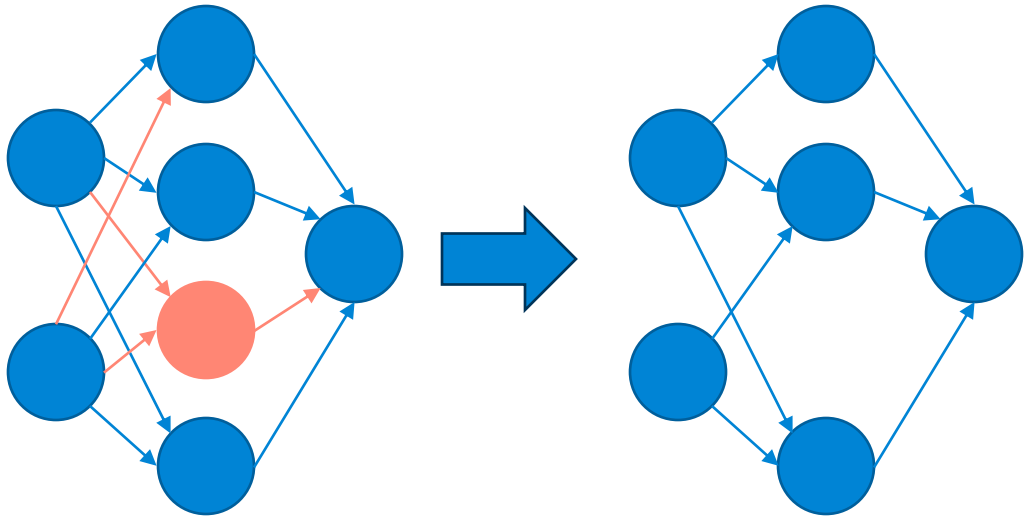


Quantization

- Typically floating point (FP32) to integer (Int8)
- Int4, ternary and even binary are being researched for further reduction in memory and complexity

Pruning

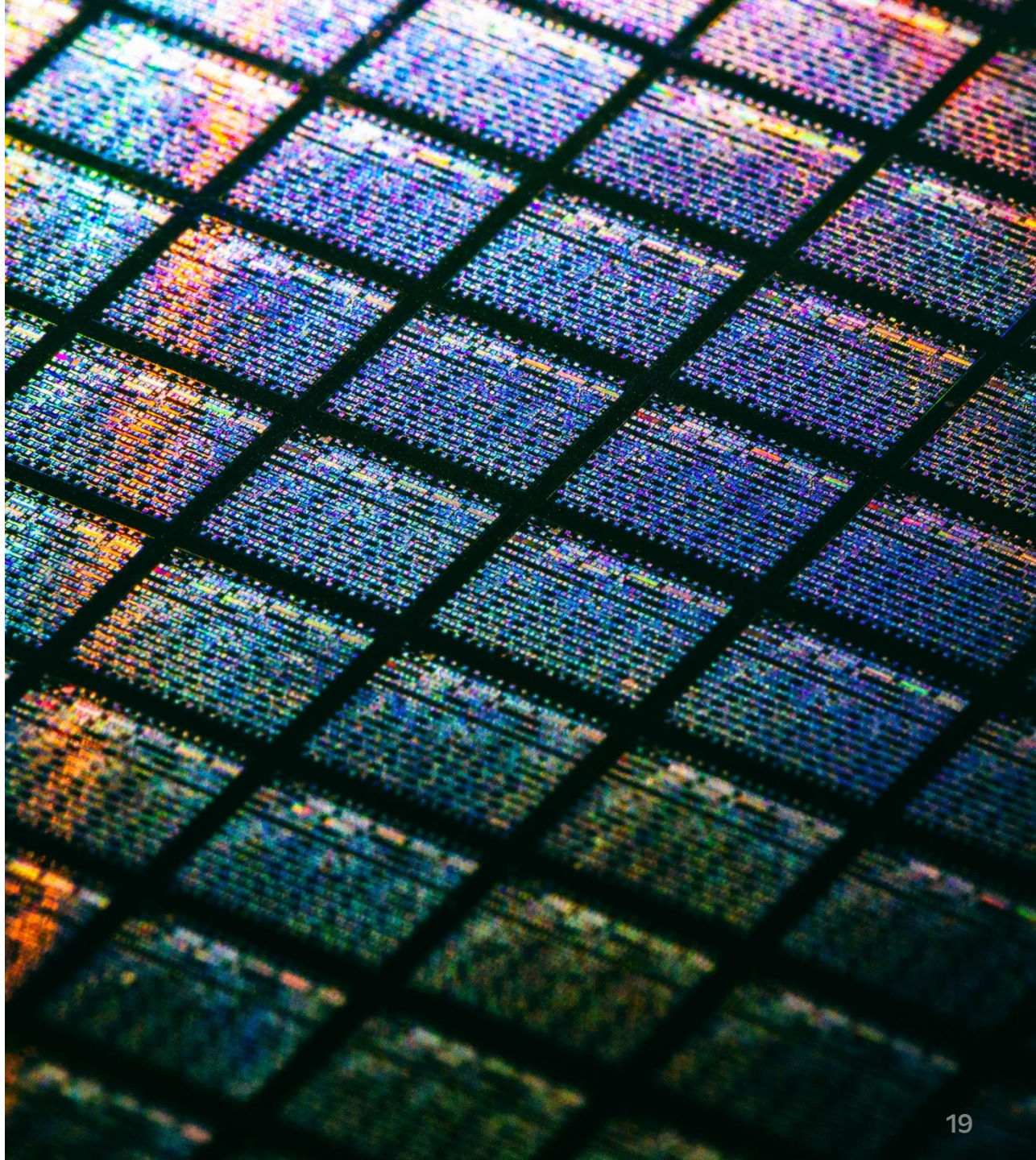
- Connections with weights that are close to 0
- Nodes whose output is close to 0

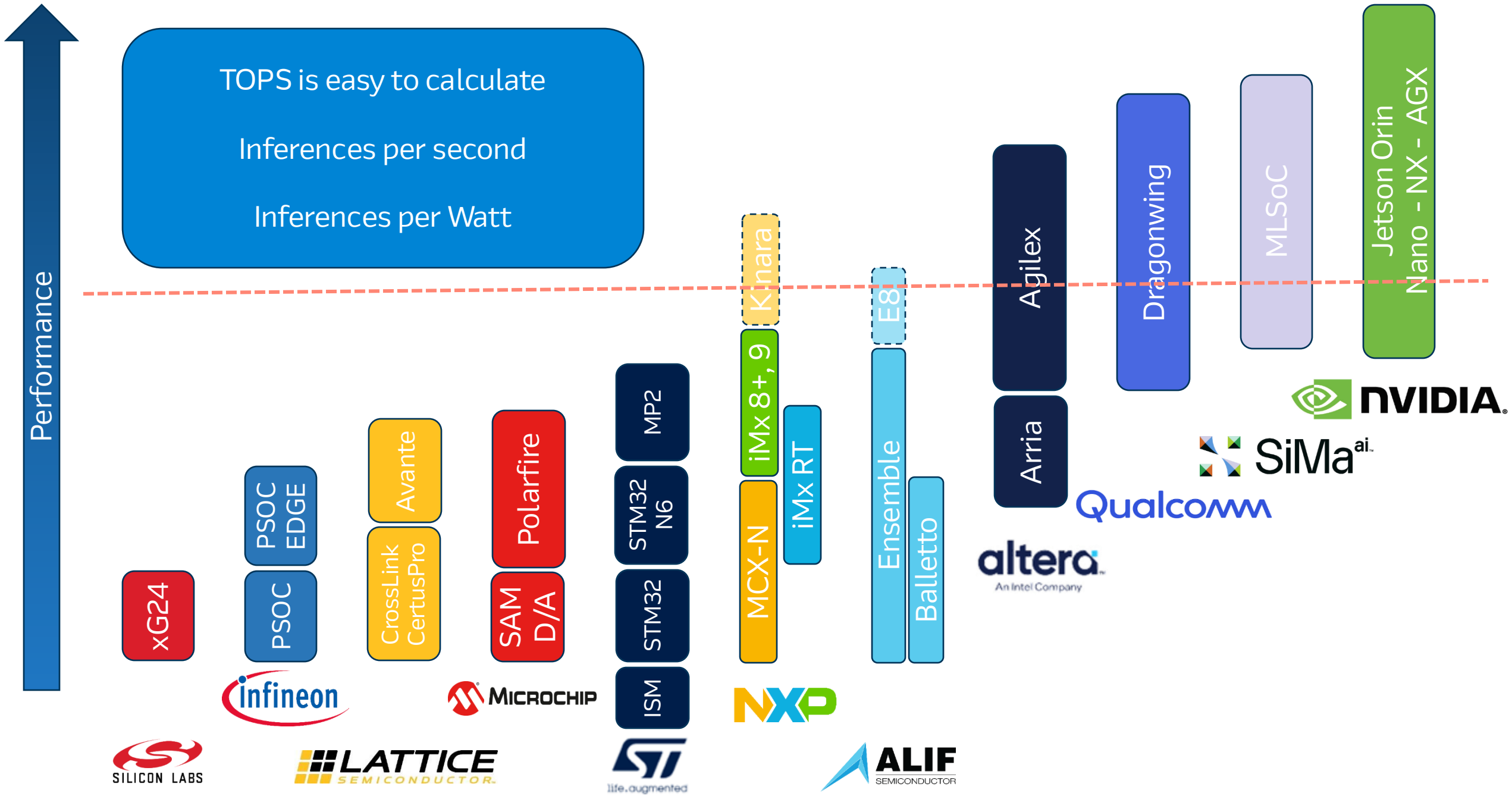


Device selection

Factors

- Performance – GOPS / TOPS
- Through put
- Latency
- AI acceleration
- Development Tools
- Application specific cores
- I/O
- Power
- Cost
- Community





Illustrative, not to scale

Neural Processing Unit (NPU)

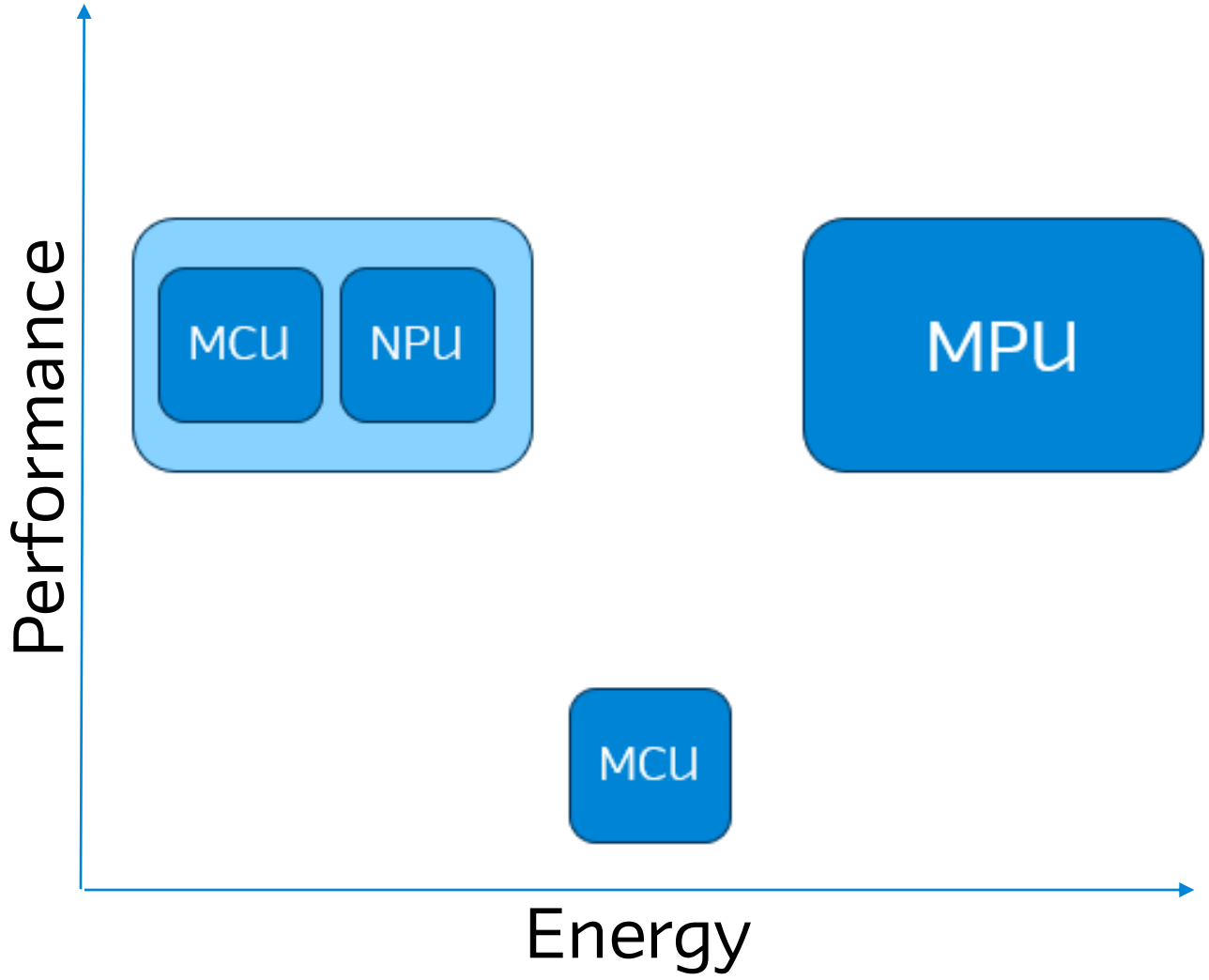
An NPU is dedicated hardware that is very efficient at AI operations

For AI tasks an NPU

- Provides higher performance
- For less energy

Microcontroller with NPU vs Microprocessor

- Similar AI performance
- Less power
- Less memory
- Lower total BOM cost



Neural Processing Unit (NPU)

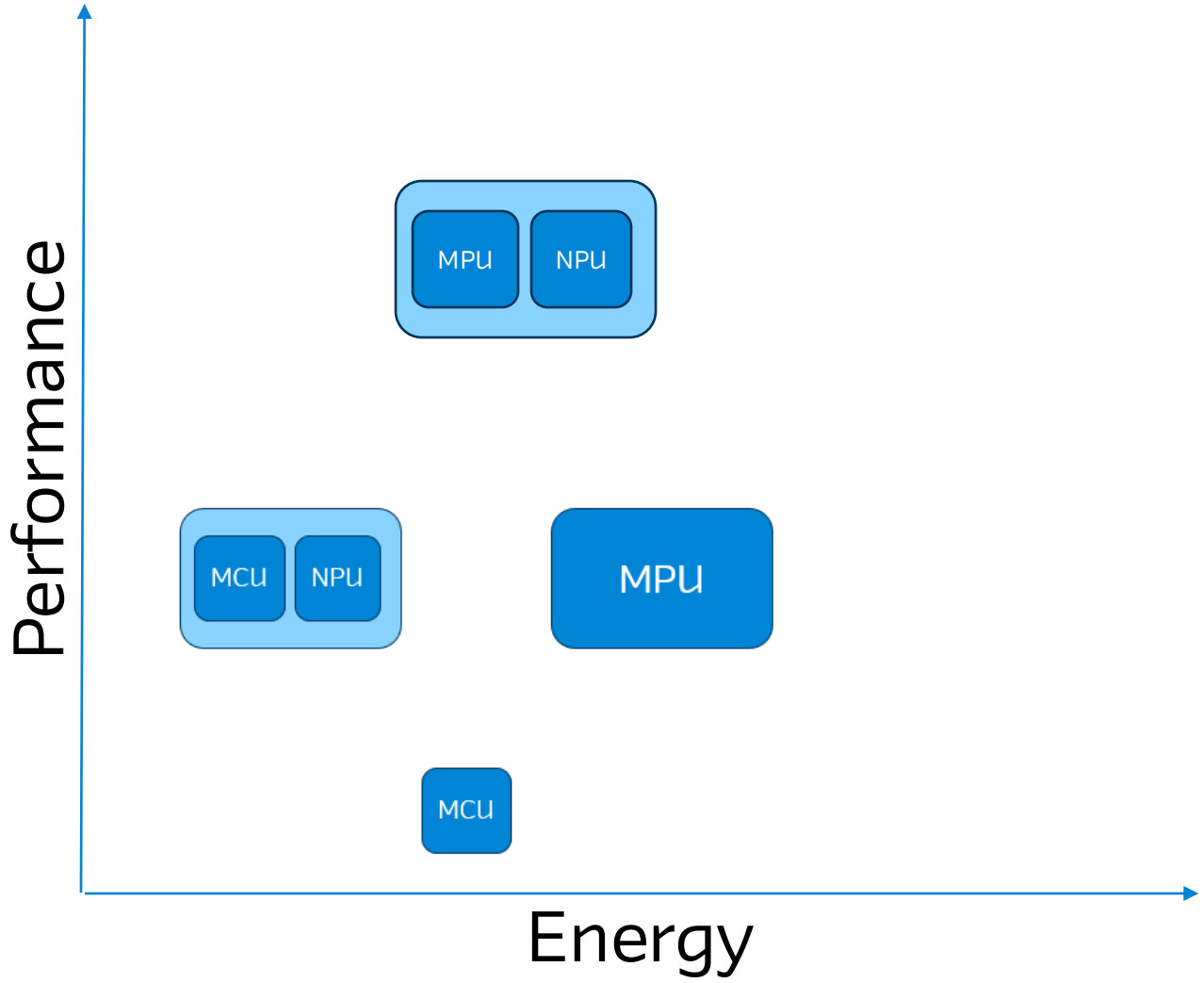
An NPU is dedicated hardware that is very efficient at AI operations

For AI tasks an NPU

- Provides higher performance
- For less energy

Microprocessor with NPU vs Microprocessor

- Higher AI performance
- Less power
- Less memory
- Lower total BOM cost

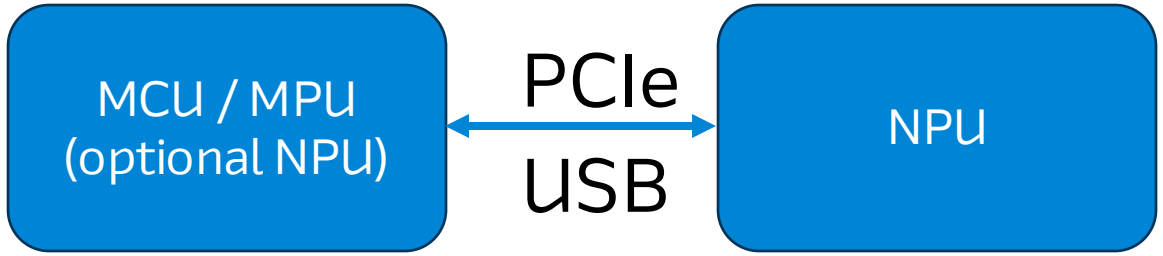


Neural Processing Unit (NPU)

An NPU is dedicated hardware that is very efficient at AI operations

For AI tasks an NPU

- Provides higher performance
- For less energy

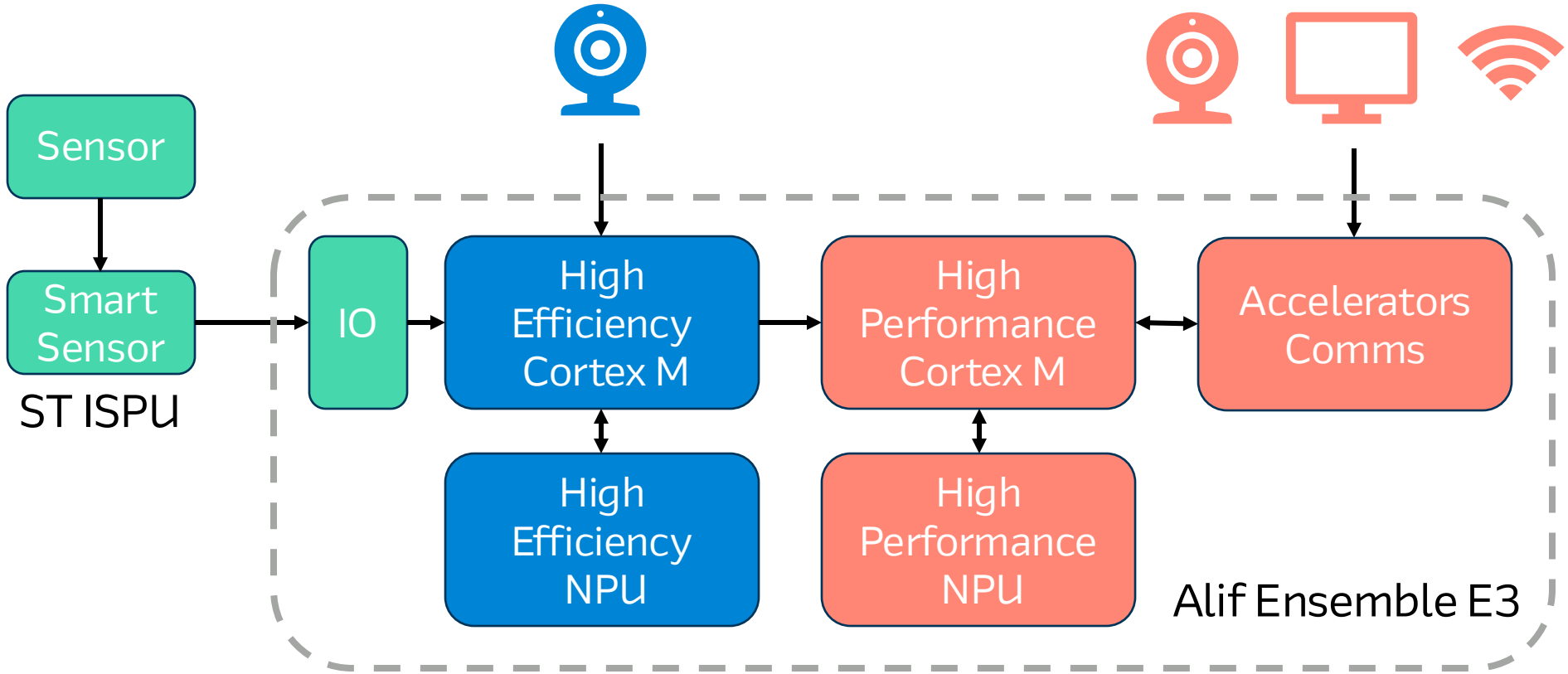


MCU / MPU with external NPU

- Substanially higher AI performance
- Low power per inference

Hardware and multiple models

It’s all about sleeping when you can



Always On

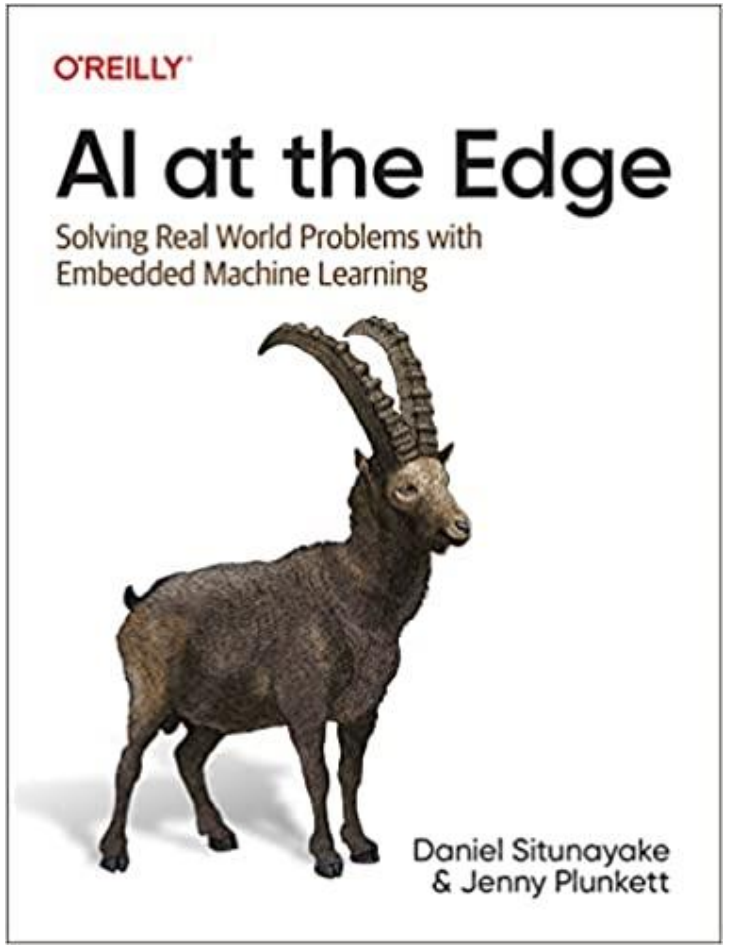
High Efficiency

High Performance

Final thoughts

For your embedded AI journey

- Have a go! Edge Impulse
- ***Transfer Learning*** - Make use of pre-trained models
- Embedded Engineers
  - Use an end-to-end software studio if possible
- AI Specialists
  - AI tools for data engineering and model training
  - Supplier tools for optimization and deployment
- Use existing hardware (devkits, SOMs) where you can, especially for prototyping.
- Recommended Book - AI at the Edge, O'Reilly



We can help!

Arrow, driving innovation forward!

Thank You!

arrow.com/edge-ai

arrow.com/edge-ai-webinars

linkedin.com/in/gdickey