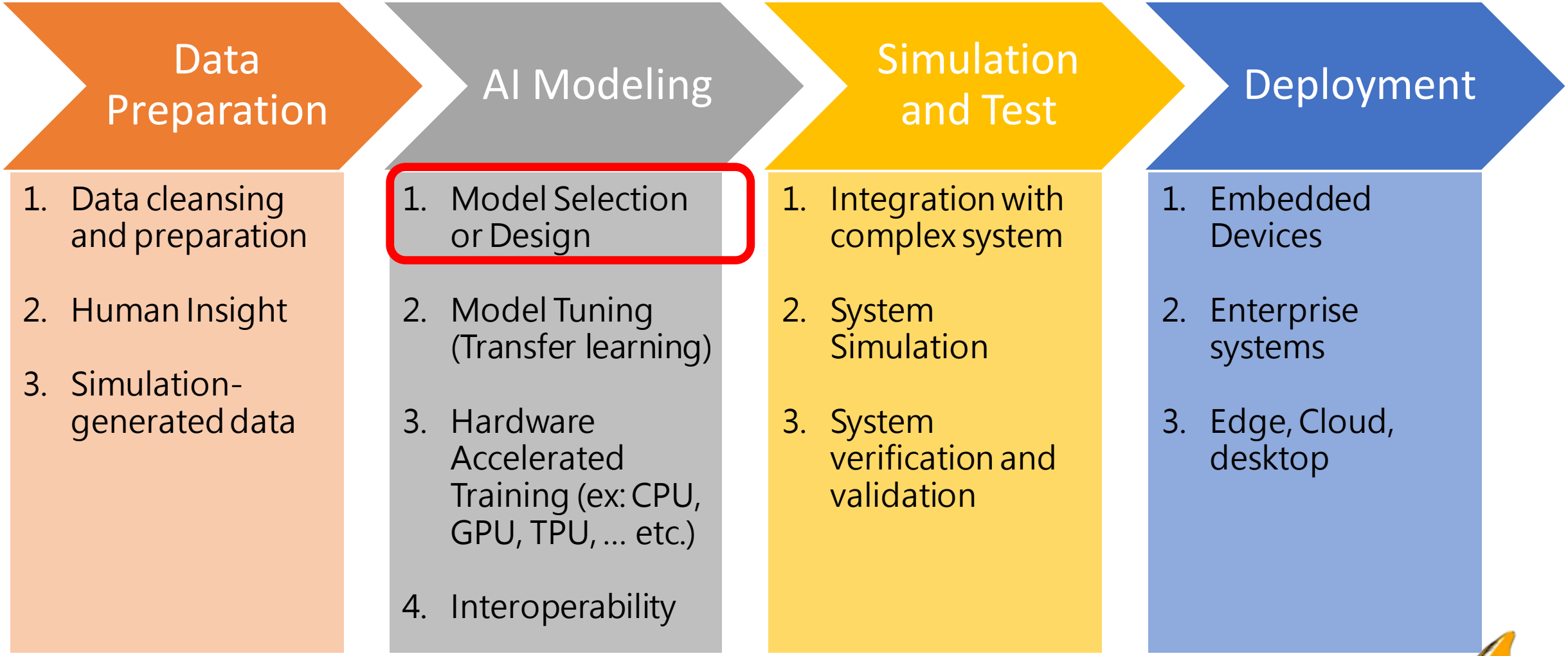
如何在電腦或是Edge Device部屬和inference模型-onnx

黃志勝
義隆電子 人工智慧研發部
國立陽明交通大學 AI學院 合聘助理教授
台北科技大學 業師

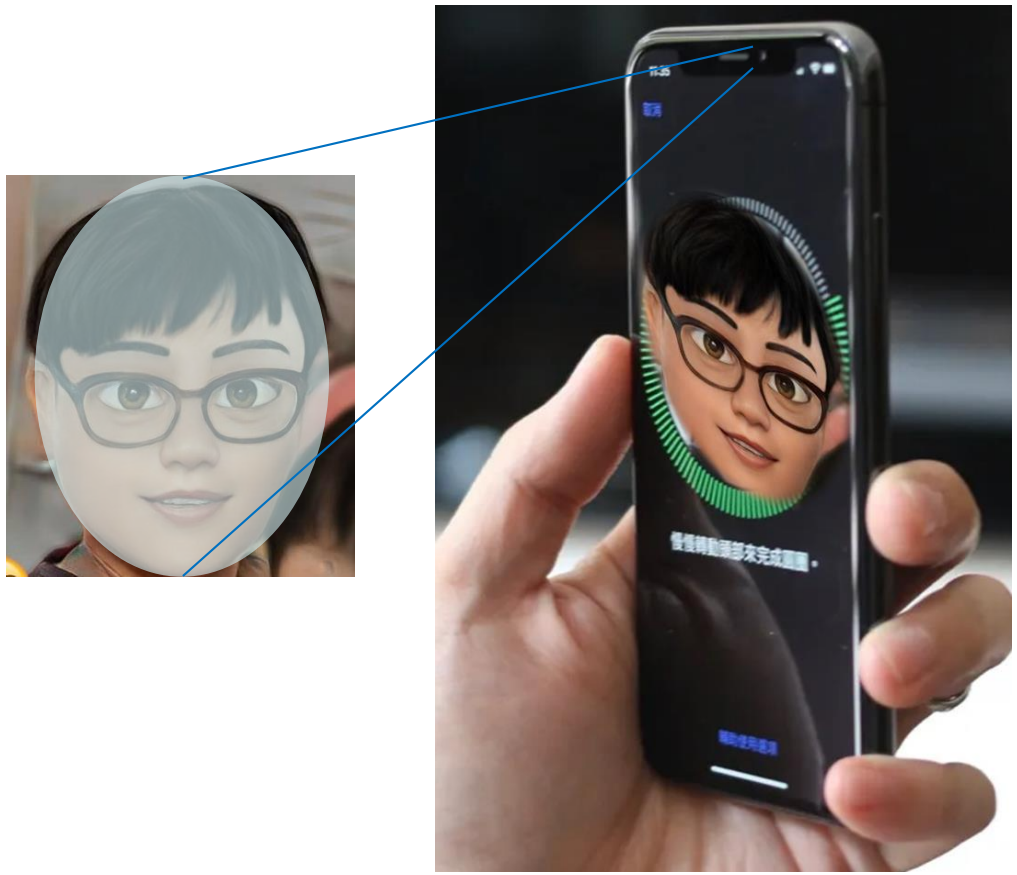
AI Is More Than a Model: Four Steps to Complete Workflow Success

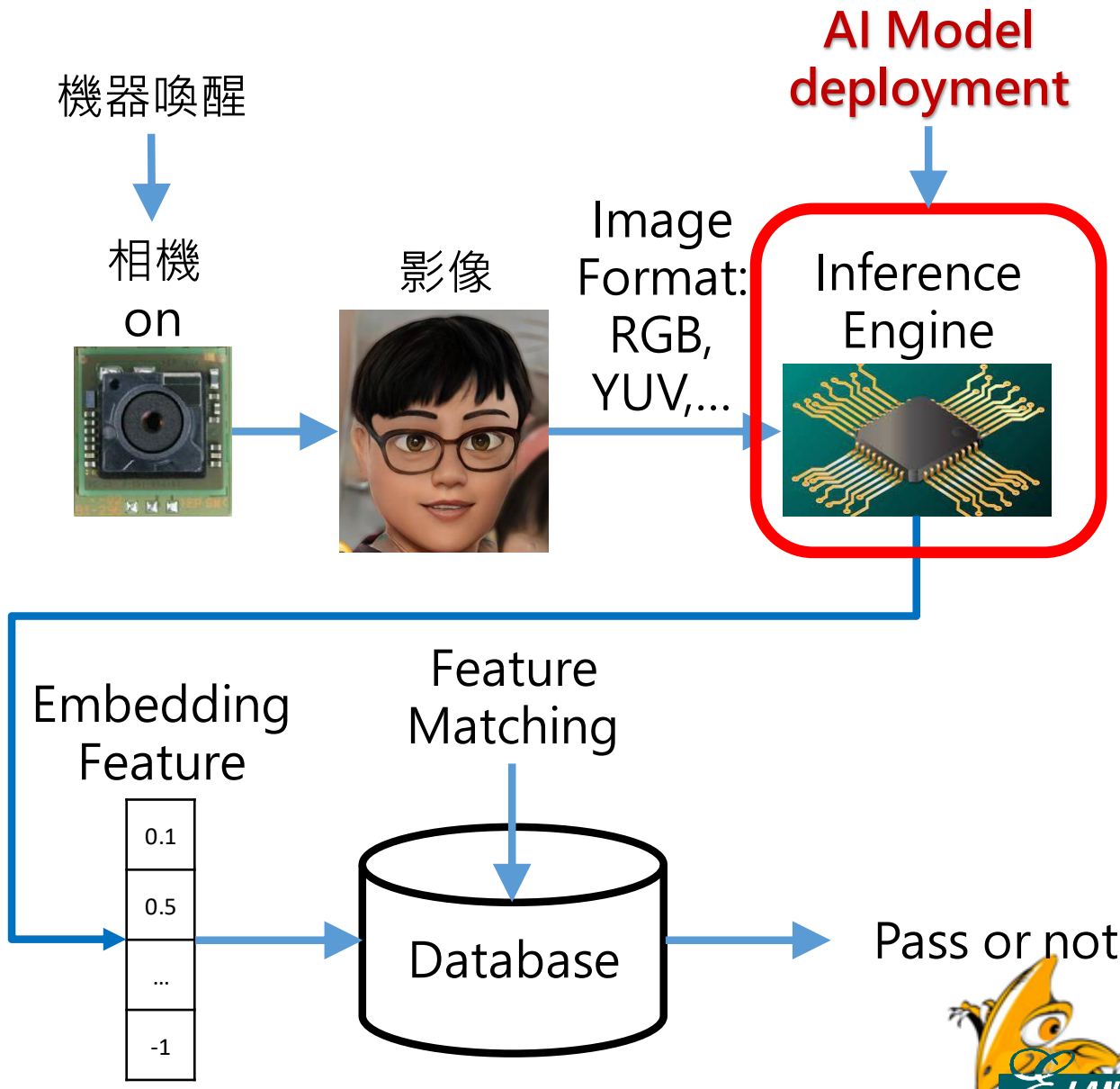


Modeling is an important step in the workflow, but the model is not the end of the journey

https://reurl.cc/n5oMd1

Example: AI 人臉辨識





人工智慧應用成功三大要素

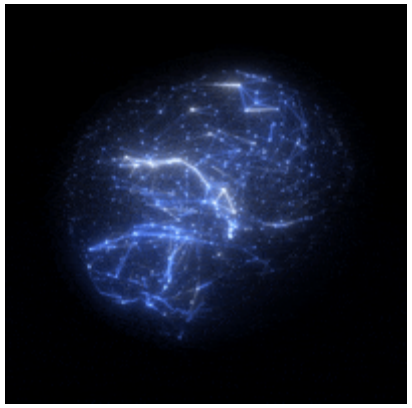
DATA

- Definition
- Collection
- Labeling
- Selection
- Augmentation

Computing Resource

Cloud or Edge?

AI



Learning Algorithm

Which AI Algorithm?

AI 要邊緣運算還是雲端運算

- 高效能計算
- 非同步(高延遲)結果推論
- 巨量資料傳輸 (影像傳輸封包過大)
- 成本過高

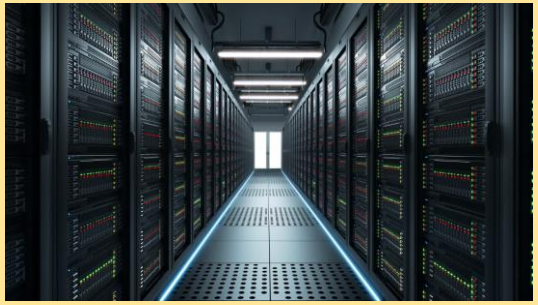
Sources

Image　Voice　Sensors i.e. Gyro

- 低延遲
- 高隱私
- 較低成本

邊緣運算:
運算設備盡量靠近資料取得來源。

雲端運算
(Cloud Computation)

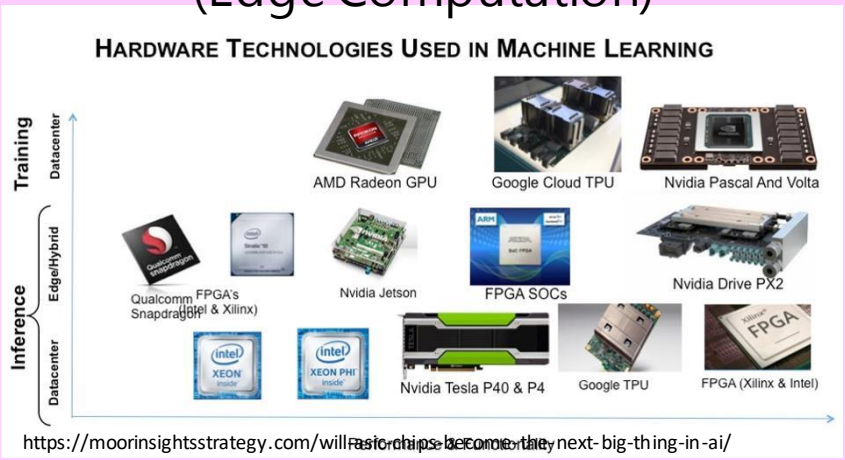


資料儲存(data storage)
模型訓練(Model Train)
模型推論(Model Inference)

模型部屬
(Model Deployment)

上傳偵測結果
(Meta information upload)

邊緣運算
(Edge Computation)

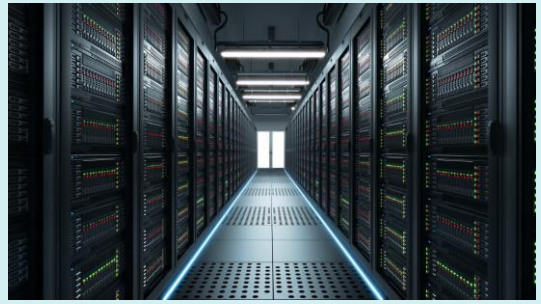


模型推論(Model Inference)

AI 要**邊緣運算**還是**雲端運算**

Camera: 30FPS (1920*1920)

雲端運算



上傳原始影像需要100MB的傳輸
(一分鐘約5.8GB，一天8.2TB)
降低解析傳輸一天傳輸量仍需約0.5TB。

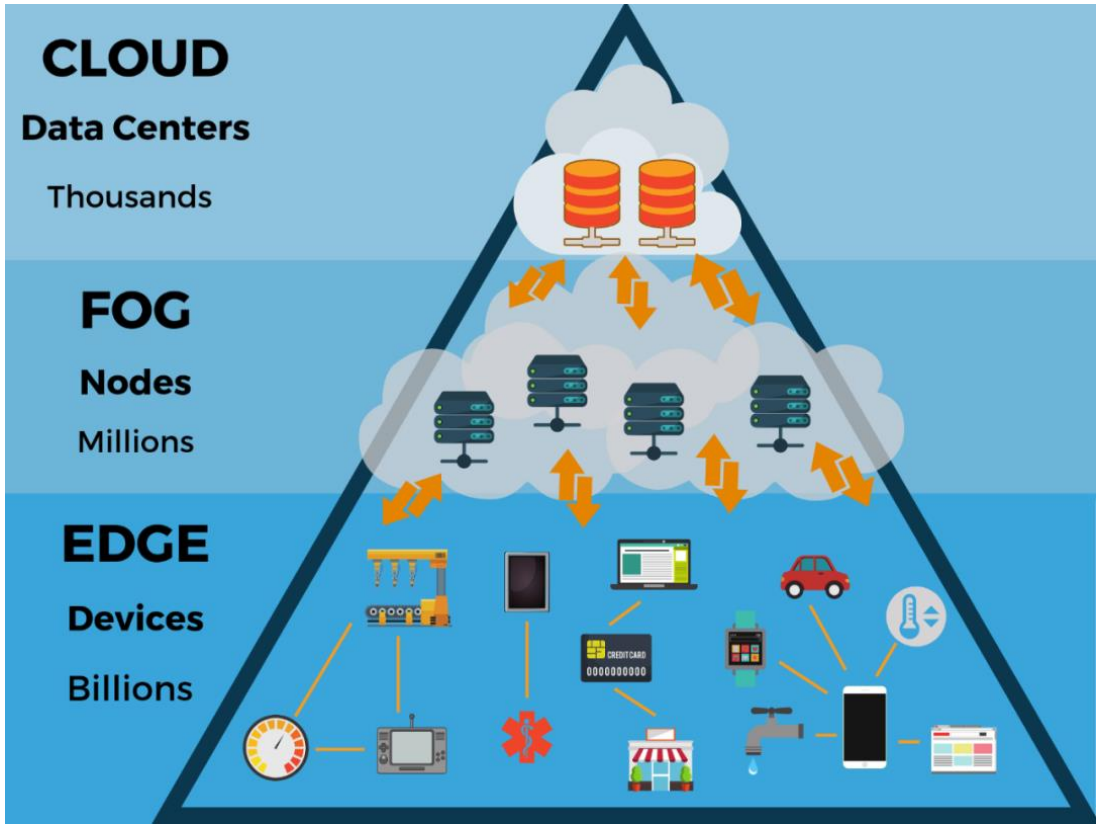
監控中心

車流分析及時數據

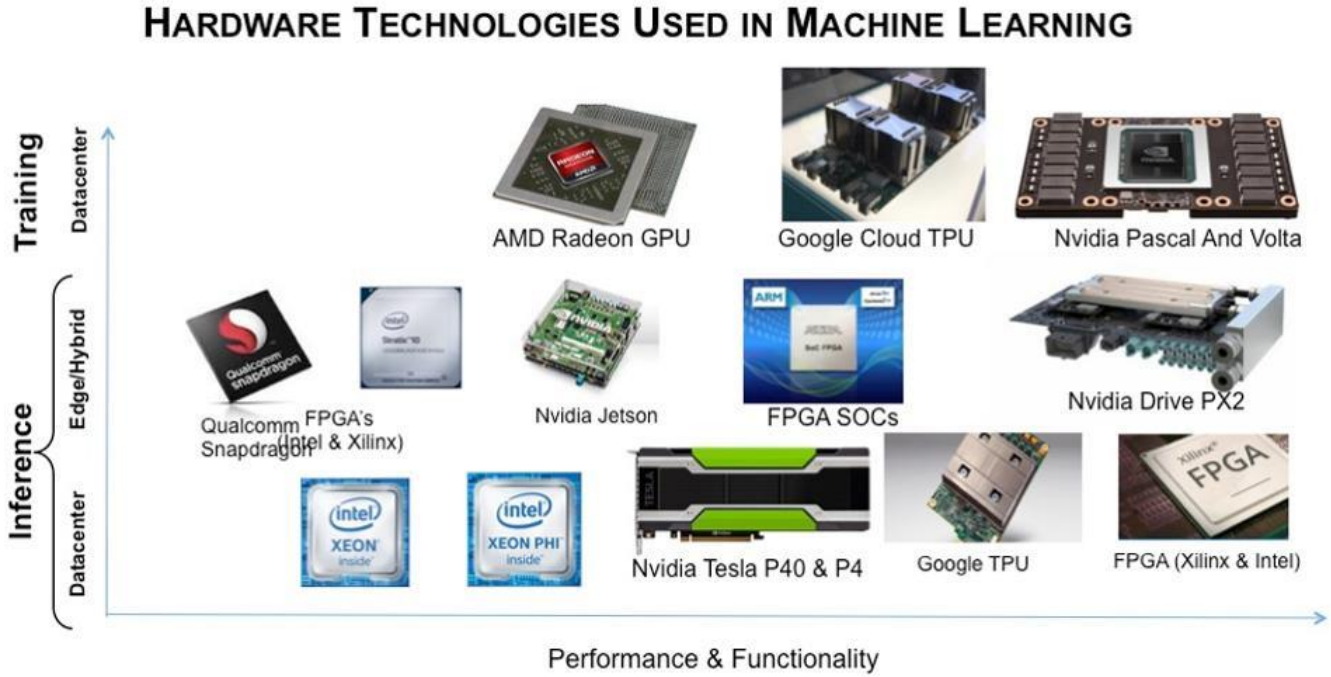
邊緣運算

上傳
每分鐘車流數據

AI Computing Resource



https://read01.com/OA4Na3P.html#.X6FL94gzaUk



https://moorinsightsstrategy.com/will-asic-chips-become-the-next-big-thing-in-ai/

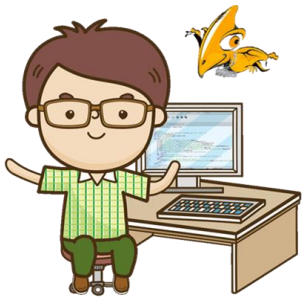
不同應用用不同的平台

邊緣運算好處:低功耗、資料傳輸(延遲)、儲存和隱私。

模型訓練和部屬

Data Collection

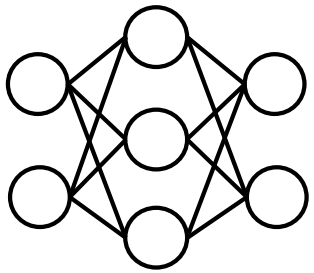
Manual



Labeled Database

Training

AI Model



伺服器訓練

Deployment

Edge Device

手機

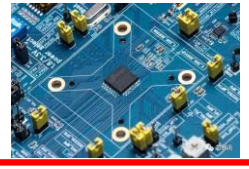
邊緣運算GPU

18 cm x 18 cm
GPU: 87 mm x 50 mm

10 cm x 9 cm
GPU:70 mm x 45 mm

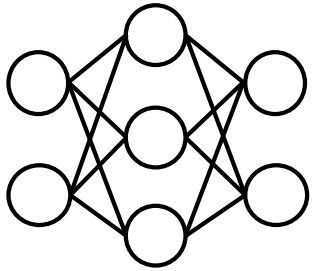
邊緣運算ASIC晶片

8 mm x 8 mm



模型訓練框架

AI Model



TensorFlow

PyTorch

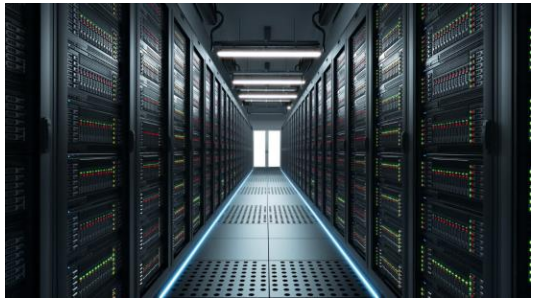
mxnet

Chainer

Keras

Caffe2

伺服器訓練



模型訓練和部屬

Edge Device

手機

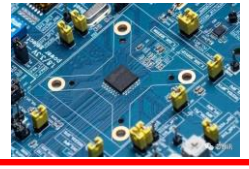
邊緣運算GPU

18 cm x 18 cm
GPU: 87 mm x 50 mm

10 cm x 9 cm
GPU:70 mm x 45 mm

邊緣運算ASIC晶片

8 mm x 8 mm



模型怎麼在不同的運算機器上跑

OS: Windows, Linus, RTOS

AI engine: Pytorch, Tensforflow, Caffe, Mxnet,...

Inference NN accelerator: CMSIS-NN, TensorFlow Lite, Openvino, TensorRT, Onnxruntime,...

不同的訓練架構要怎麼互通

是否有相同的運算架構做這樣的事情

PyTorch

TensorFlow

mxnet

Chainer

Keras

Caffe2

ONNX

Edge Device

手機

邊緣運算GPU

18 cm x 18 cm
GPU: 87 mm x 50 mm

10 cm x 9 cm
GPU:70 mm x 45 mm

邊緣運算ASIC晶片

8 mm x 8 mm

ONNX

Edge Device

手機

邊緣運算GPU

18 cm x 18 cm
GPU: 87 mm x 50 mm

10 cm x 9 cm
GPU:70 mm x 45 mm

邊緣運算ASIC晶片

8 mm x 8 mm

FP-64/32/16
INT8

ONNX RUNTIME

TensorRT

OpenVINO

arm
CMSIS-NN

Micro Edge & Mini Edge

- **算力限制 (Computation Limitation)**
單位時間計算能力

- **記憶體限制 (Memory Limitation)**
模型參數量、精度(FP64/32/16, INT1/2/4/8)

- **功耗限制(Power Consumption Limitation)**
主晶片、周邊元件功耗

- **開發框架限制**

OS: Windows, Linus, RTOS
AI engine: Pytorch, Tensforflow, Caffe, Mxnet,…
Inference NN accelerator: CMSIS-NN, TensorFlow Lite, Openvino, TensorRT, Onnxruntime,…

- **價格限制**

Raspberry Pi 4
(ARM Cortex-A72)

Arduino Portenta H7
(ARM M4+M7)

ARM (CMSIS-NN)

Google (TensorFlow Lite)

Intel (OpenVINO)

Nvidia Jetson (TensorRT)

Kneron AI SoC

Reference: http://omnixri.blogspot.com/2021/05/20210515edge-ai.html

ONNX

ONNX [編輯]

維基百科，自由的百科全書

ONNX（英語：Open Neural Network Exchange）是一種針對機器學習所設計的開放式的文件格式，用於存儲訓練好的模型。它使得不同的人工智慧框架（如Pytorch、MXNet）可以採用相同格式存儲模型數據並交互。 ONNX的規範及代碼主要由微軟，亞馬遜，Facebook和IBM等公司共同開發，以開放原始碼的方式託管在Github上。[2][3][4] 目前官方支持加載ONNX模型並進行推理的深度學習框架有： Caffe2, PyTorch, MXNet，ML.NET，TensorRT 和 Microsoft CNTK，並且 TensorFlow 也非官方的支持ONNX。

https://github.com/onnx/onnx
https://zh.wikipedia.org/wiki/ONNX

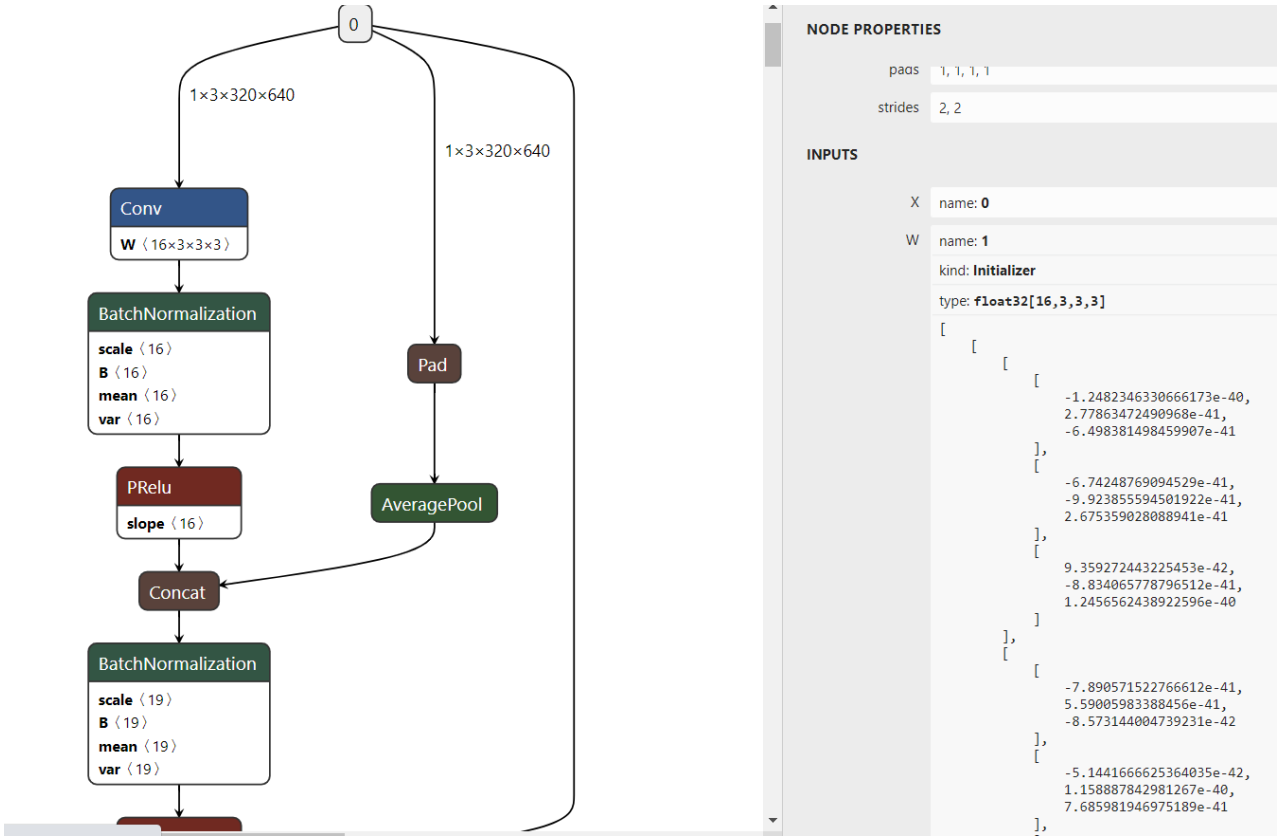
ONNX Structure

ONNX Model

graph

- Input (each layer shape)
- Node (parameters & computation)
- Initializer (weight & bias)
- Output

netron



https://netron.app/

Demo (pytorch to onnx)

- https://github.com/TommyHuang821/NTUT_EdgeAICourse/blob/main/main_pytorch_imageclassification_onnx.ipynb

- https://github.com/TommyHuang821/NTUT_EdgeAICourse/blob/main/main_pytorch_objectdetection_onnx.ipynb