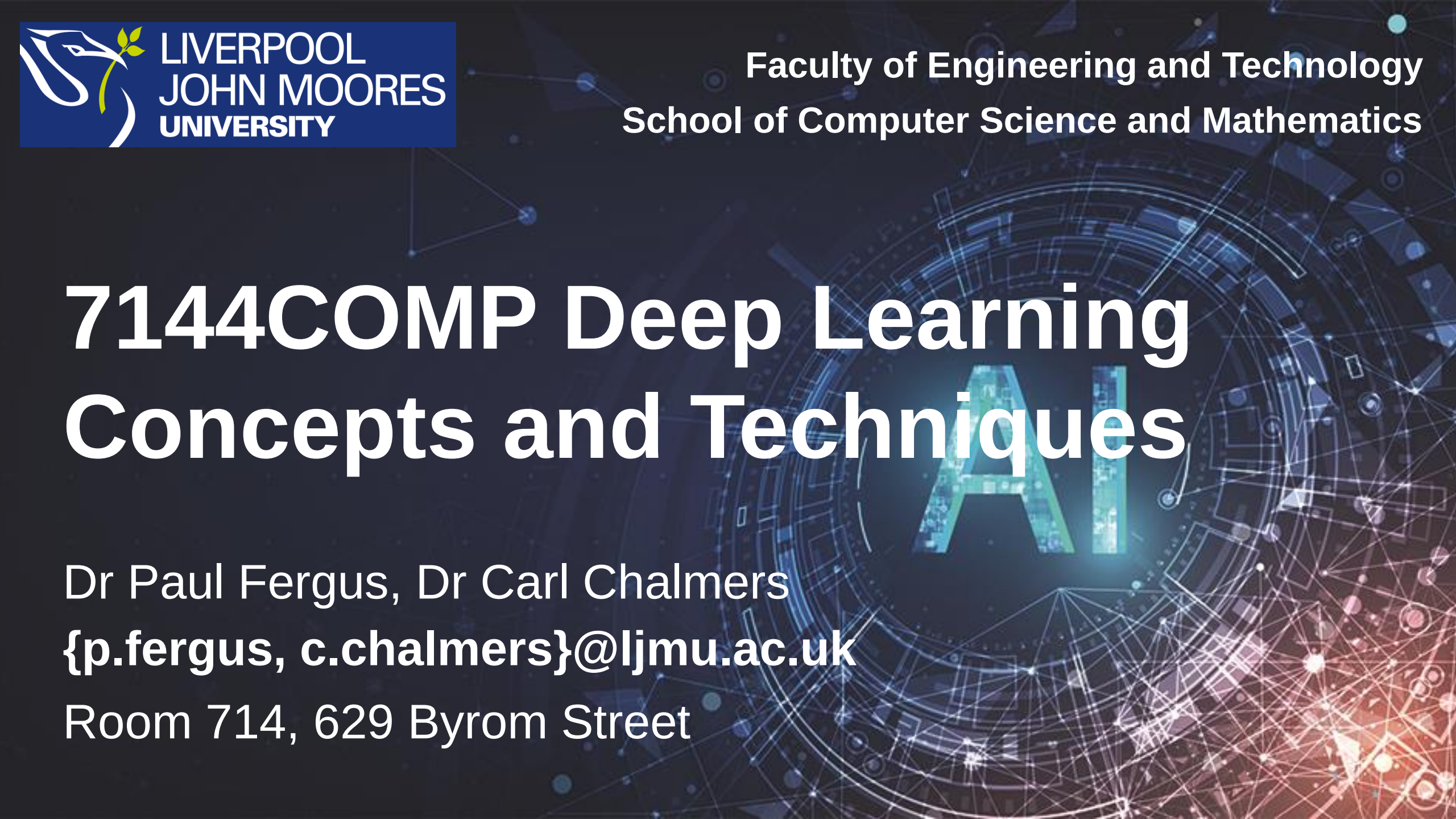
LIVERPOOL JOHN MOORES UNIVERSITY

Faculty of Engineering and Technology
School of Computer Science and Mathematics

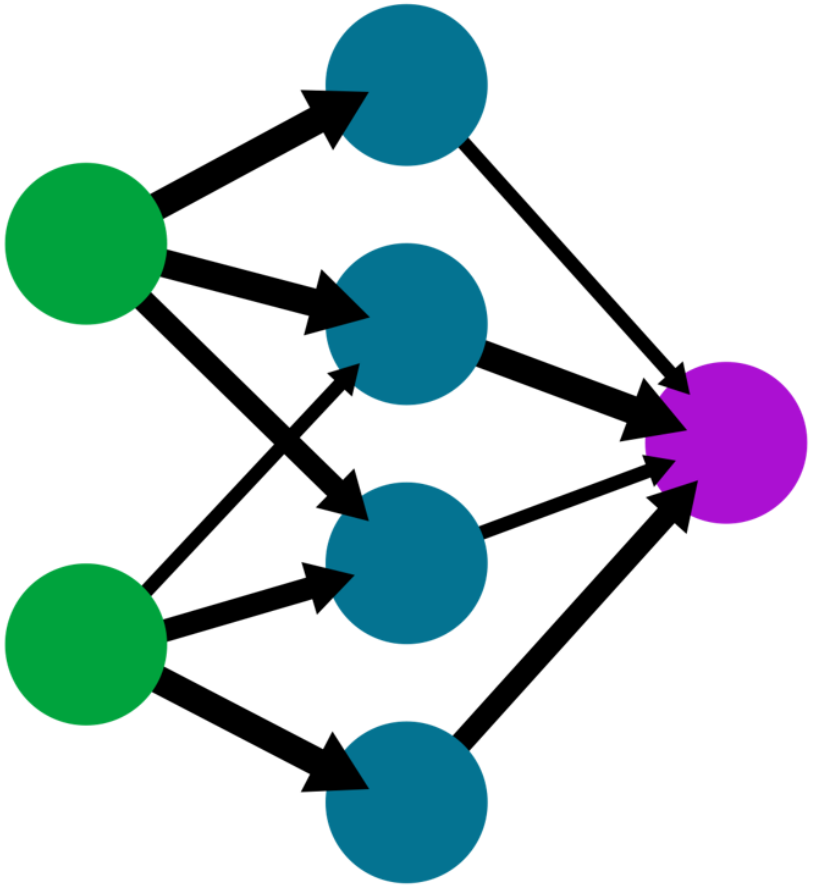
# 7144COMP Deep Learning Concepts and Techniques




Dr Paul Fergus, Dr Carl Chalmers
**{p.fergus, c.chalmers}@ljmu.ac.uk**
Room 714, 629 Byrom Street

# Lecture 6

## Image Classification and Object Detection Part 1



# In this session…

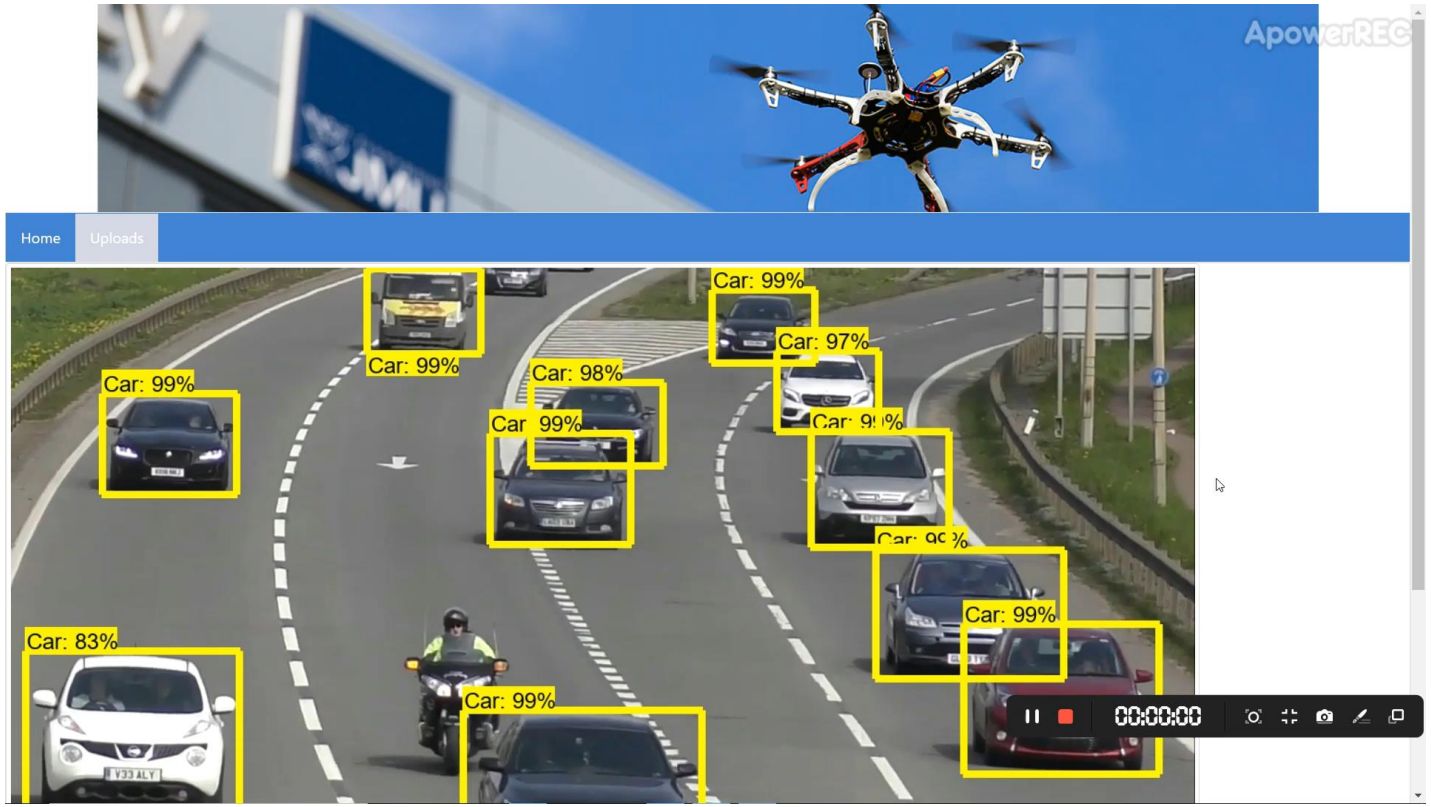
- We will cover:
  - Computer vision overview
  - Historical context
  - Why computer vision is useful
  - Data-Driven Approach (Collect a dataset of images and labels)
  - Famous datasets
  - Hardware Accelerated Deep Learning (CPU vs GPU)
  - Training and Associated Hardware (Development Systems, Training Systems, Inferencing Systems)
  - Tensor Processing Unit (TPU)
  - Other Hardware Considerations
  - Distributed Training
  - Model Parallelism

# Computer Vision Overview

- One of the most significant use cases for DL is image classification and object detection which aims to replicate one of the most important senses humans have
- The influx of both data and compute capability has enabled the rapid growth and adoption of computer vision applications
- This is arguably one of the most significant technological revolutions that has demonstrated signification impact across multiple domains including manufacturing, healthcare, security and autonomous vehicles

# Computer Vision Overview

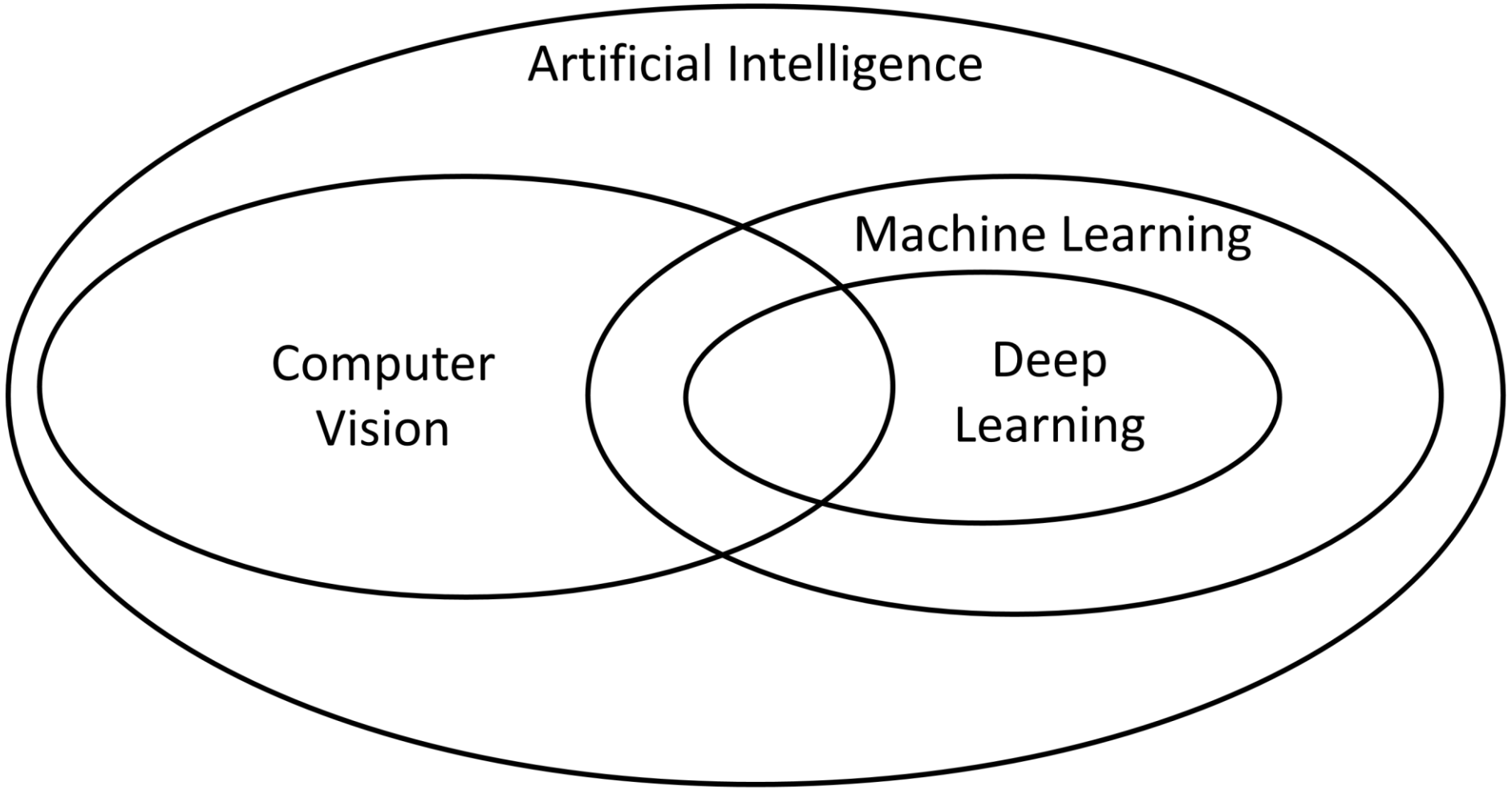
- These vision systems are becoming increasingly integrated within the fabric of many vision applications currently deployed such as CCTV to understand and contextualise information for situational awareness



# Computer Vision Overview

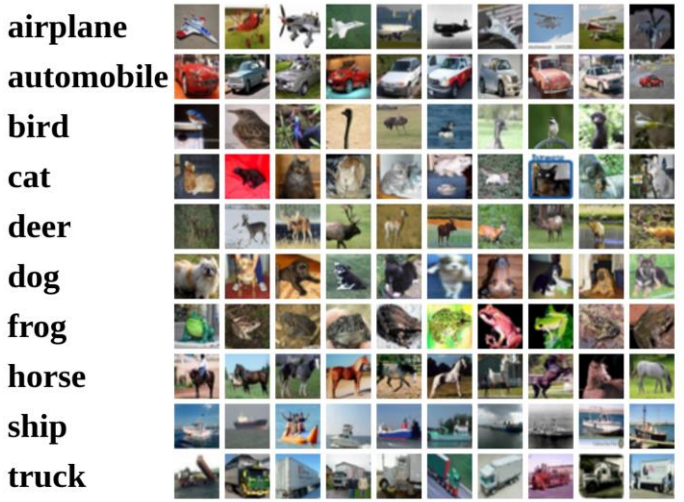
- Many aspects of computer vision have been research topics for many years however DL is regarded as a disruptive technology so much so that most computer vision applications now incorporate many aspects of deep learning
- This is not to say that many of the traditional tools and techniques are redundant instead we would see them forming part of the application pipeline where DL and the traditional computer vision tools work in conjunction
- DL is very good at extracting the important features which allows them to detect and differentiate between different object objects. However more complex tools found in computer vision would be required for much more detailed analysis
- For example, packages such as OpenCV have a rich collection of non-DL algorithms to support the complex requirements for images analysis and processing

# Computer Vision Overview



# Famous Deep Learning Datasets

- MNIST
  - 10 classes
  - Digits 0 to 9
  - 28x28 grayscale images
  - 50k  training images
  - 10k test images
- CIFAR10
  - 10 classes
  - 50k training images (5k per class)
  - 10k testing images (1k per class) 32x32 RGB images



# Famous Deep Learning Datasets

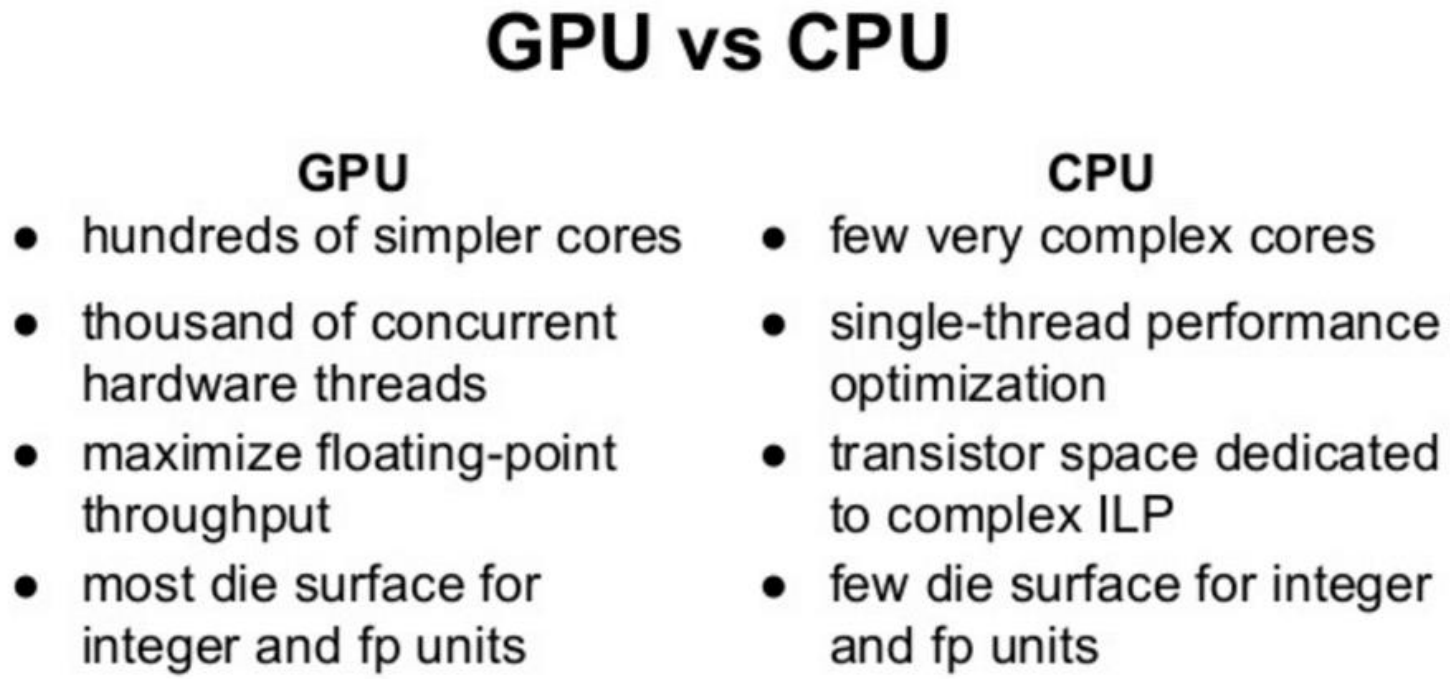
- ImageNet
  - 1000classes
  - 1.3M training images (1.3K per class) 50K validationimages (50 per class) 100K test images (100 per class)
- COCO
  - 1.5 million object instances
  - 80 object categories
  - Object segmentation


# Hardware Accelerated Deep Learning

- Although you can train and inference (host) DL models using CPUs (which is supported in most frameworks) it incredibly inefficient and slow
- Simple model architectures can perform reasonably well however those which are comprised of hundreds of layers require alternative hardware such as Graphics Processing Units (GPU's)
- CPU's are designed to handle a wide range of tasks which can be processed very quickly
- However, tasks can only be processed sequentially, and this means there is an inherent bottle neck in CPU computation when many calculations are required

# Hardware Accelerated Deep Learning

- GPU's addresses this limitation by running such calculations in parallel

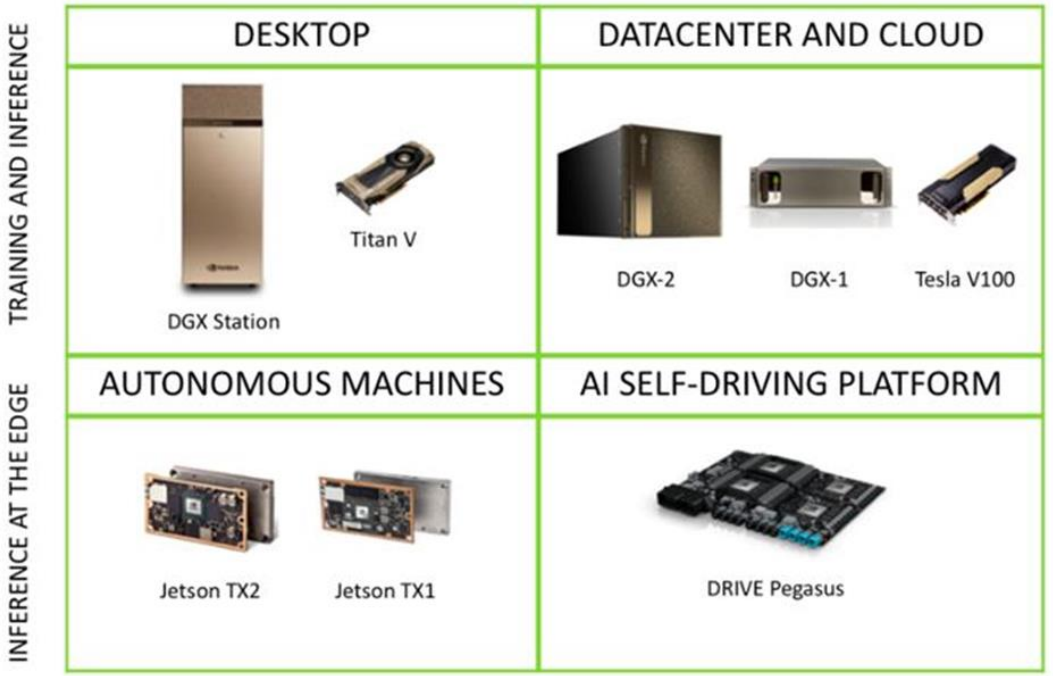


# Hardware Accelerated Deep Learning

- In the context of training a DL model where both a forward pass and backpropagation are undertaken large matrix multiplications are required
- This process is the most computationally intensive part of the neural network which is made up of multiple matrix multiplications
- Using a CPU to train large DL models would take too long
- For example, in an object detection model that such as ResNet50 with 23 million trainable parameters it could take weeks if not months for the model to converge. To deal with this level of complexity you would require a GPU

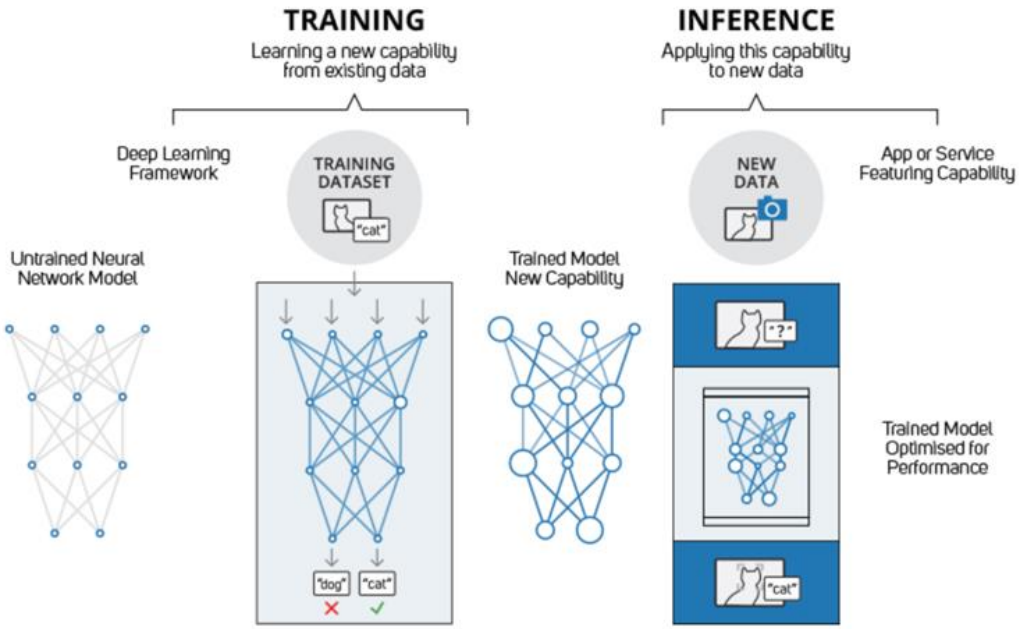
# Hardware Accelerated Deep Learning

- A GPU allows us to perform all of the operations at the same time instead of doing each one after another hence why they are used for training and hosting DL models
- There are a wide variety of different GPUs and selecting the right hardware is important when developing data intensive applications such as object detection
- The cost can range from £800 to over £10000 per card and some heavy-duty models may require several cards to support both the memory and compute requirements



# Training and Associated Hardware

- Whilst accelerated hardware is one of the most important components for DL the hardware requirements can vary significantly depending on the application such as development, training or inferencing
- Each application has specific hardware requirements which needs careful consideration



# Development Systems

- During the development stage DL workstations using GPU accelerators enable rapid performance for developing and debugging your Deep Learning and Machine Learning models. During this development stage you can identify potential models and prototype different architecture before training on a data centre solution. Some of the common GPU's used in this stage at the time of writing includes:
  - GeForce RTX 2080, 8GB RAM, 2944 cores
  - GeForce RTX 2080 Ti, 11GB RAM, 4352 cores
  - Quadro RTX 6000, 24GB RAM, 4608 CUDA Cores, 576 Tensor Cores, 72 RT Cores
  - Quadro RTX 8000, 48GB RAM, 4608 CUDA Cores, 576 Tensor Cores, 72 RT Cores
  - Quadro GV100, 32GB RAM, 5120 CUDA Cores, 640 Tensor Cores

# Training Systems

- Once initial model testing and selection has been undertaken full scale training can be undertaken on large scale training system. Some of the most common GPU based datacentre configurations at the time of writing includes:
  - NVIDIA DGX-1, 8 * Tesla V100s (each one has 5120 CUDA Cores, 640 Tensor Cores, and 32GB RAM)
  - NVIDIA DGX-A100, 8 * A100s (each one has 6912 CUDA Cores, 432 TF Tensor Cores, 40GB RAM)
  - NVIDIA DGX-2, 16 * Telsa V100s (each one has 5120 CUDA Cores, 640 Tensor Cores, and 32GB RAM)

# Inferencing Systems

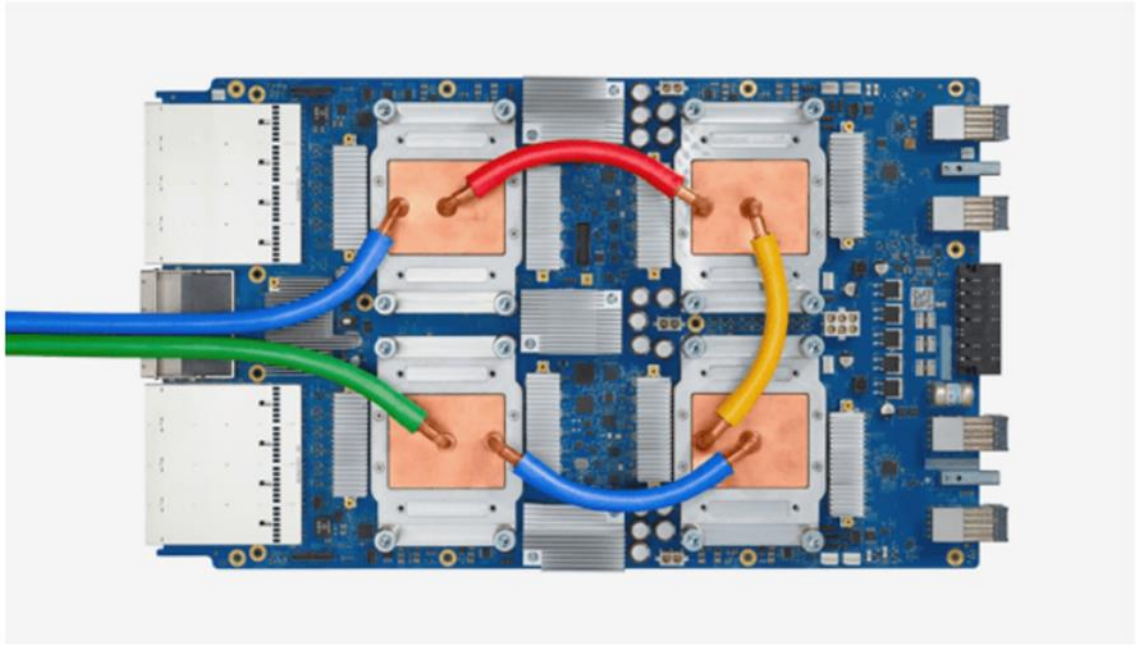
- Once an AI model has been trained, the resulting streamlined neural network can then be presented with new data for it to identify - whether it be to recognise and diagnose medical images, identify spoken words or predict habits – these results, dependant on the model's training, is called inference
- The first approach when inferring looks at parts of the neural network that don't get activated after it's trained. These sections just aren't needed and can be "pruned" away
- The second approach looks for ways to fuse multiple layers of the neural network into a single computational step. It's akin to the compression that happens to a digital image

# Inferencing Systems

- Designers might work on these huge, beautiful, million pixel-wide and tall images, but when they go to put it online, they'll turn into a jpeg. It'll be almost exactly the same, indistinguishable to the human eye, but at a smaller resolution
- Similarly, with inference you'll get almost the same accuracy of the prediction, but simplified, compressed and optimized for runtime performance
- As the needs for inference are different form training, a dedicated set of hardware is recommended, often required if the place where inferring is to be done – such as remote cameras, drones or autonomous mobile vehicles. Some of the most common inferencing hardware at the of writing includes:
  - Tesla T4, 2560 CUDA Cores, 320 Turing Tensor Cores, 16GB RAM.
  - Jetson AGX Xavier, 512 Core Volta GPU, 16GB RAM
  - Intel Arria 10 FPGA Card

# Tensor Processing Unit (TPU)

- An alternative to using the GPU for training is to use TPU's. The first TPU's which are manufactured by google first come available in 2016 – version 4 being the latest TPU release
- TPU's offer a wide range of advantages which includes:
  - Accelerated performance of liner algebra computation which is extensively used in ML.
  - Minimised time to accuracy when training large and complex neural network models (convergence)
  - Model which previously took weeks to train can converge in hours on TPU's
  - They scale across multiple nodes



# Tensor Processing Unit (TPU)

- A TPU is known as an application-specific integrated circuit (ASIC) and is specifically designed for DL/ML tasks
- TPUs are 15 to 30 times faster than GPUs in inferencing and delivered a 30–80 times improvement in TOPS/Watt measure. In machine learning training, the Cloud TPU is more powerful in performance (180 vs. 120 TFLOPS) and four times larger in memory capacity (64 GB vs. 16 GB of memory) than Nvidia's best GPU Tesla V100
- The improved performance of TPU's is partially due to the fact that parameters are loaded from memory into a matrix of multipliers and added. As each multiplication is executed, the result will be passed to the next multipliers while taking summation at the same time
- So, the output will be the summation of all multiplication result between data and parameters. During the whole process of massive calculations and data passing, no memory access is required

# Other Hardware Considerations

- There are a few additional hardware considerations which are worth paying attention to. Firstly, GPU's can be power hungry and require a Power Supply Unit (PSU) which is capable of meeting the requirements of the GPU
- Low voltage can cause training to fail or affect the performance of your GPU as they can be throttled by the driver
- One important thing to be aware of is that even if a PSU has the required wattage, it might not have enough PCIe 8-pin or 6-pin connectors
- Certain types of hardware like the Tesla T4 are prone to thermal throttling if they are not cooled efficiently. You need to check with the manufacturer for both power and cooling requirements

# Other Hardware Considerations

- With system memory (RAM) the general rule of thumb is to spec the system with the same amount of memory (RAM) of your highest GPU memory.
- If you are using large datasets you need to ensure that there is enough hard drive space with adequate throughput. One of the most common problems you are likely to encounter is GPU out of memory.
- Here you can either replace your GPU with one that has lager memory capacity or add additional GPU’s as long as they are the same make and model as you have sufficient PSU capability

# Reading For Directed Study Week

- Conservation AI: Live Stream Analysis for the Detection of Endangered Species Using Convolutional Neural Networks and Drone Technology. https://arxiv.org/ftp/arxiv/papers/1910/1910.07360.pdf
- Collaborative Pressure Ulcer Prevention: An Automated Skin Damage and Pressure Ulcer Assessment Tool for Nursing Professionals, Patients, Family Members and Carers https://arxiv.org/ftp/arxiv/papers/1808/1808.06503.pdf
- Object Detection with Deep Learning: A Review https://arxiv.org/pdf/1807.05511.pdf
- Deep learning https://towardsdatascience.com/12-papers-you-should-read-to-understand-object-detection-in-the-deep-learning-era-3390d4a28891

# Next Session

- Image Classification and Object Detection Part 2
- Object Recognition
- Image Classification
- Object Detection
- Semantic Segmentation
- Object Segmentation
- Skip connections
- Data Augmentation (Horizontal Flips, Random Crops and Scales, Colour Jitter)