

DR ADRIAN BEVAN

PRACTICAL MACHINE LEARNING

RESOURCES



LECTURE PLAN

- ▶ Introduction
- ▶ Hardware considerations
- ▶ Benchmarking example: MoEDAL
- ▶ Summary
- ▶ Suggested reading

QMUL Summer School:

<https://www.qmul.ac.uk/summer-school/>

Practical Machine Learning QMplus Page:

<https://qplus.qmul.ac.uk/course/view.php?id=10006>



INTRODUCTION

- ▶ The purpose of this session is to discuss the different types of resources that you may wish to use with machine learning algorithms, and the consequences of those choices.
- ▶ There are different constraints on the use of an algorithm that will lead to different choices of hardware.
 - ▶ What is your application?
 - ▶ Do you have a hardware constraint (e.g. mobile phone, games console, etc.)?
 - ▶ How much time do you have to make a decision?

QMUL Summer School:

<https://www.qmul.ac.uk/summer-school/>

Practical Machine Learning QMplus Page:

<https://qplus.qmul.ac.uk/course/view.php?id=10006>



HARDWARE CONSIDERATIONS

- ▶ The platform you want to deploy your algorithm on will dictate what you can and can not do.
 - ▶ e.g spell checking algorithms on mobile phones have small (~3Mb) memory footprint requirements c.f. scientific applications where 10's Gb of memory can be available.
 - ▶ What type of computation are you performing?
 - ▶ Are rounding errors critical for your application?
 - ▶ Is error checking on memory critical for your application?
 - ▶ Is your model complex (large numbers of HPs or computations)?

HARDWARE CONSIDERATIONS

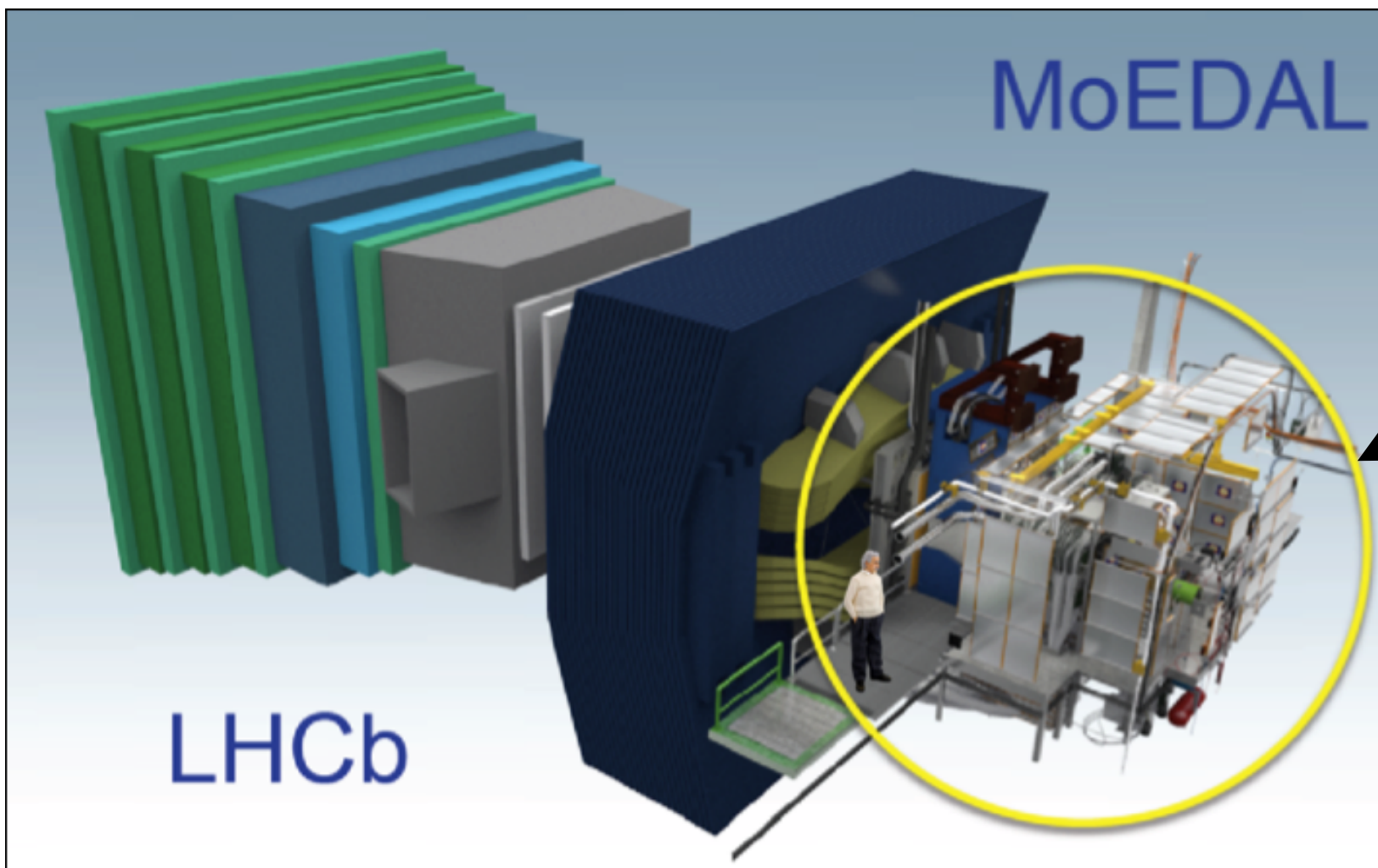
- ▶ Computing resources have evolved significantly with time.
 - ▶ The world's first programmable digital electronic computer was a thermionic valve based system called Colossus, constructed at Bletchley Park during WW2.
 - ▶ The use of transistors (in the 1950's) for computing revolutionised the capabilities of humans; and played an important role in ensuring that humans stepped foot on the moon.
 - ▶ The Apollo Guidance Computer (AGC) had a memory footprint of 2048 16-bit words (4096 bytes = 0.004Mb) and a clock speed of 2 MHz.
 - ▶ Computers typically have several-tens of Gb of memory and operate at GHz clock speeds: i.e. several thousand times faster than the AGC with a few million times the memory.

<https://bletchleypark.org.uk>

<http://www.tnmoc.org>

BENCHMARKING EXAMPLES: MOEDAL

- ▶ The benchmarking examples here include work drawn from our research: looking for new particles at the LHC.



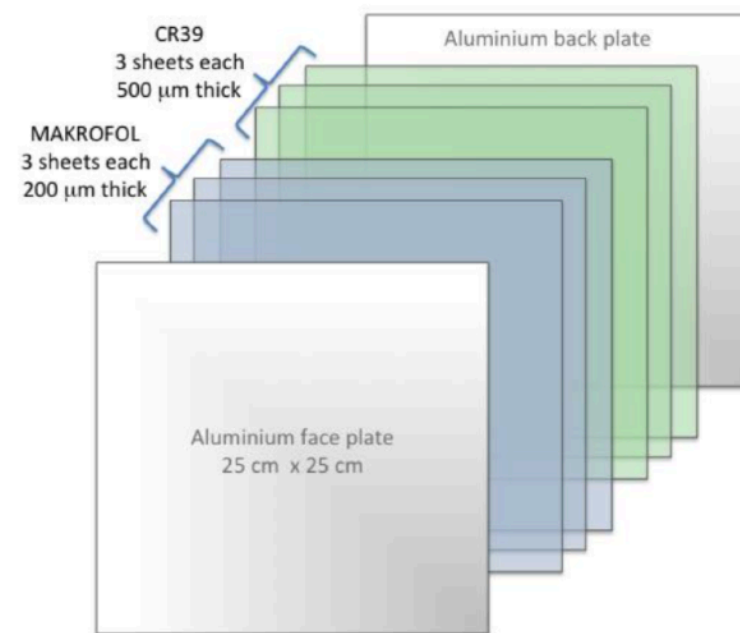
MoEDAL shares the detector cavern with the LHCb experiment.

pp collisions occur in the heart of the MoEDAL detector.

Passive instrumentation is placed around this region.

BENCHMARKING EXAMPLES: MOEDAL

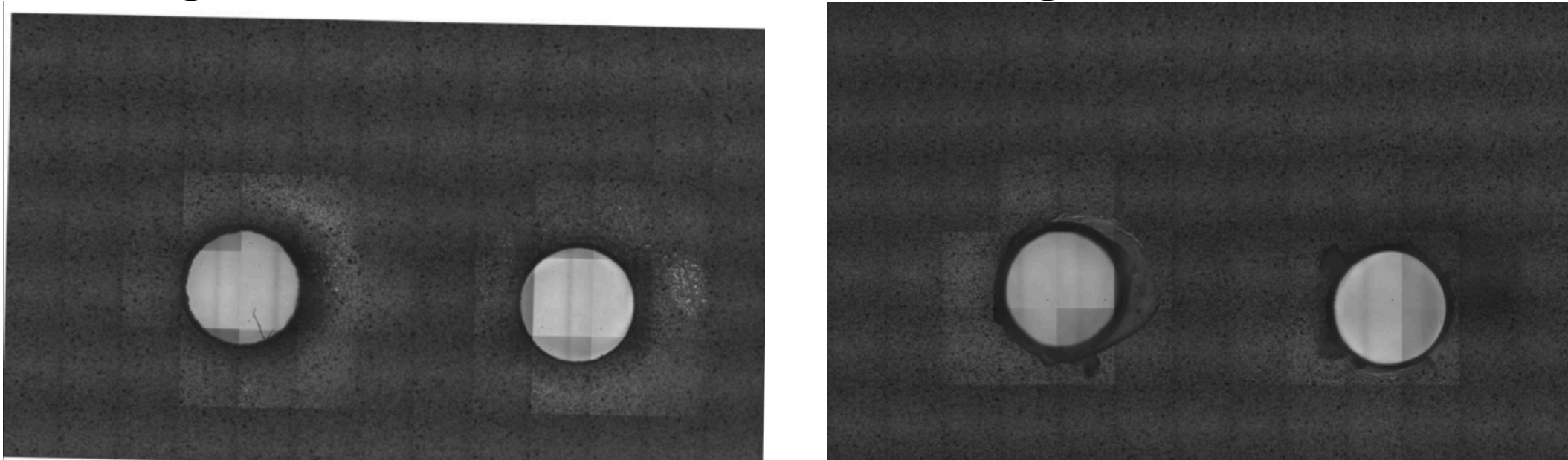
- ▶ The nuclear track detector system of MoEDAL consists of boxes of plastic: Macrofoil and CR39.



- ▶ Charged particles traversing a sheet of plastic will break polymer chains.
- ▶ Chemically etching the plastic after exposure to the LHC results in holes where charged particles have passed through the material.

BENCHMARKING EXAMPLES: MOEDAL

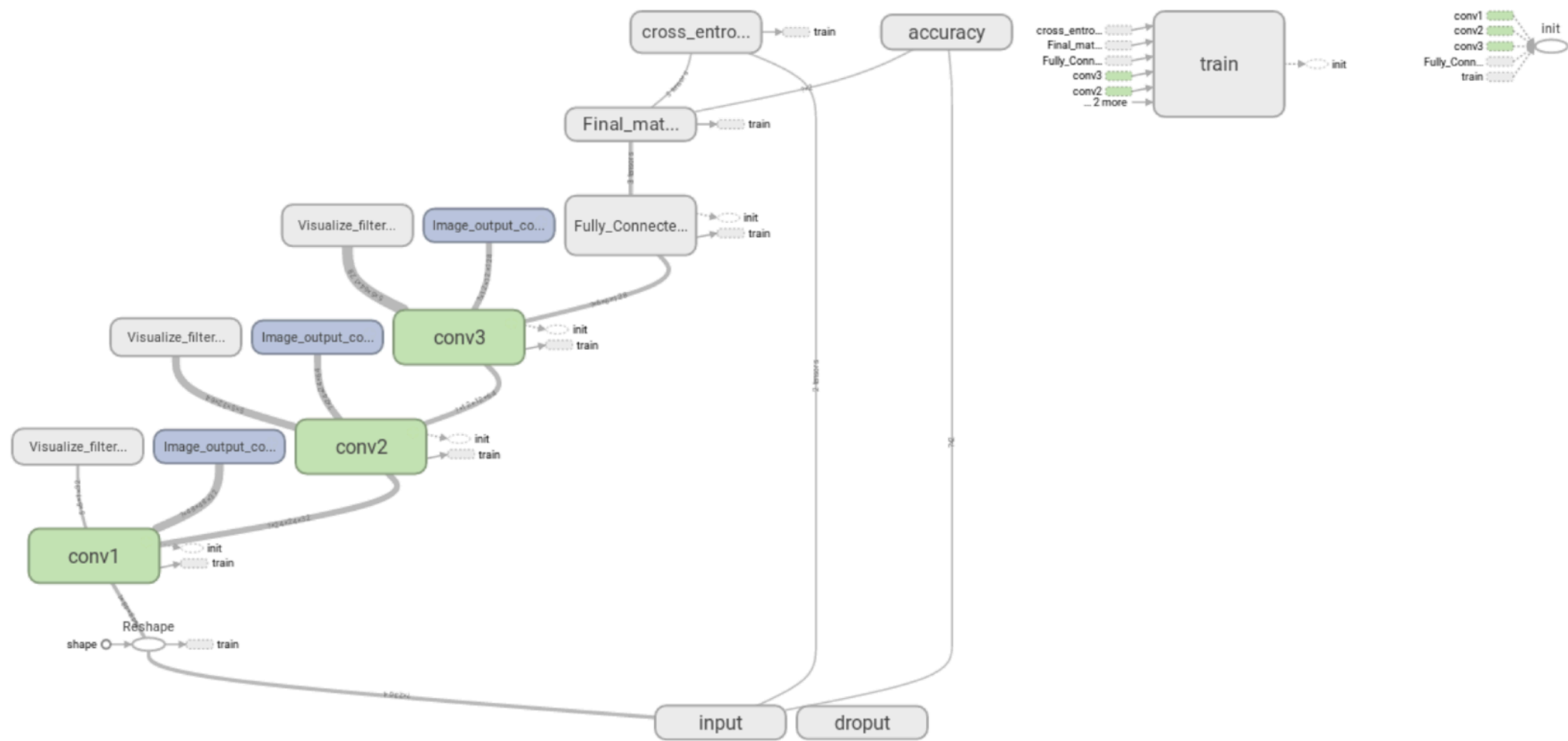
- ▶ The plastic then needs to be scanned for subsequent analysis. The images used for this benchmarking tests are below:



- ▶ Not enough training data to train a CNN.... so we oversample the data to generate pseudo samples for training.
- ▶ Signal: holes
- ▶ Background: elsewhere

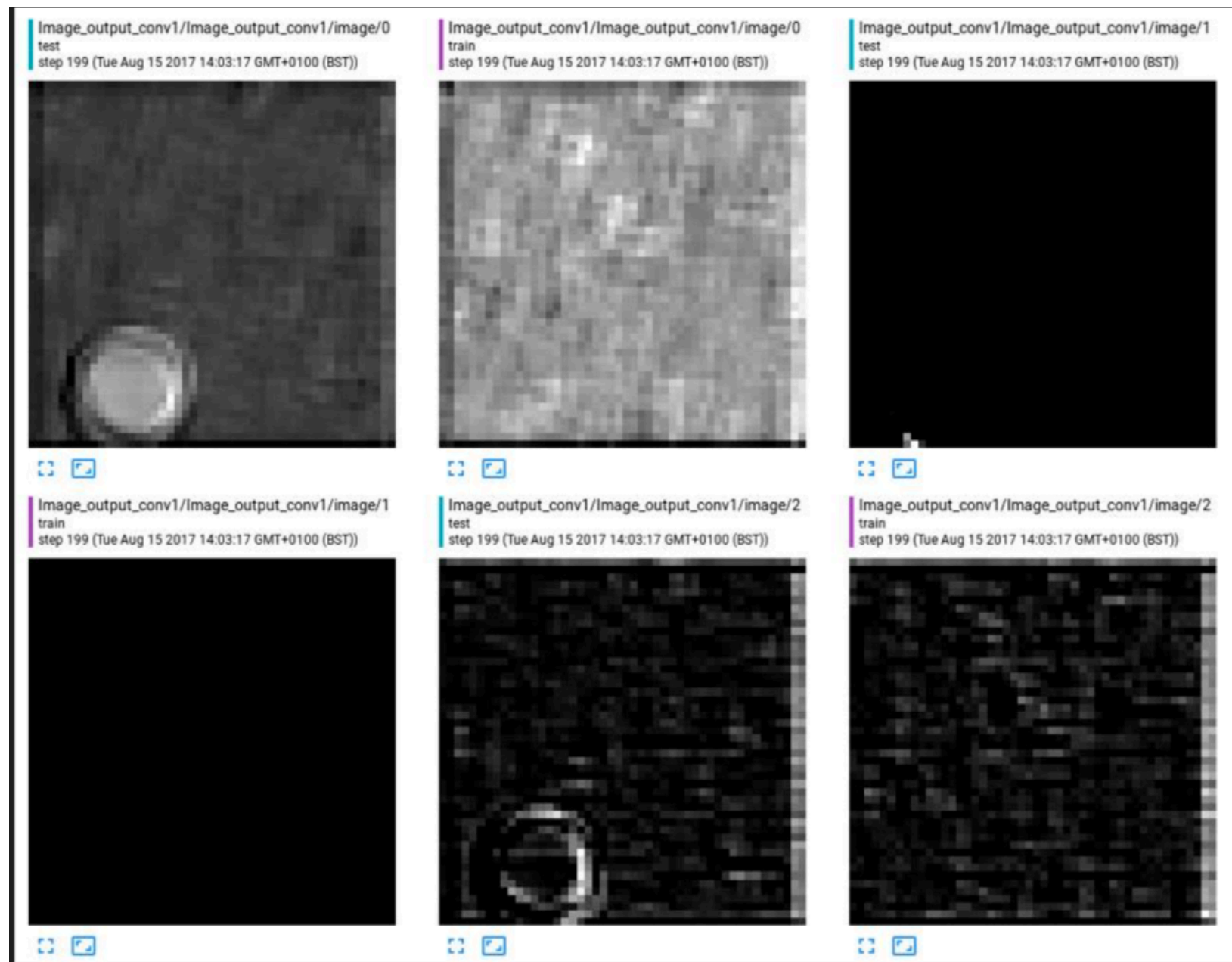
BENCHMARKING EXAMPLES: MOEDAL

- ▶ Model graph: 3 conv layers and a fully connected layer



BENCHMARKING EXAMPLES: MOEDAL

▶ Example convolution images



Different convolution filters pick up on different features in the data.

Some identify holes.

Others identify the random noise of background etching on the surface of the plastic.

This information is combined in the final decision made by the network.

BENCHMARKING EXAMPLES: MOEDAL

▶ Hardware used:

▶ NVidia K40 (12 Gb of RAM)

Number of cores: 2880

Memory bandwidth: 288 GB/sec

Memory speed: 3.0 GHz

235W graphics card power

▶ NVidia GeForce 1080 Ti (12 Gb of RAM)

Number of cores: 3584

Memory bandwidth: 484 GB/sec

Memory speed: 11Gbps

250W graphics card power

Dedicated professional GPU card

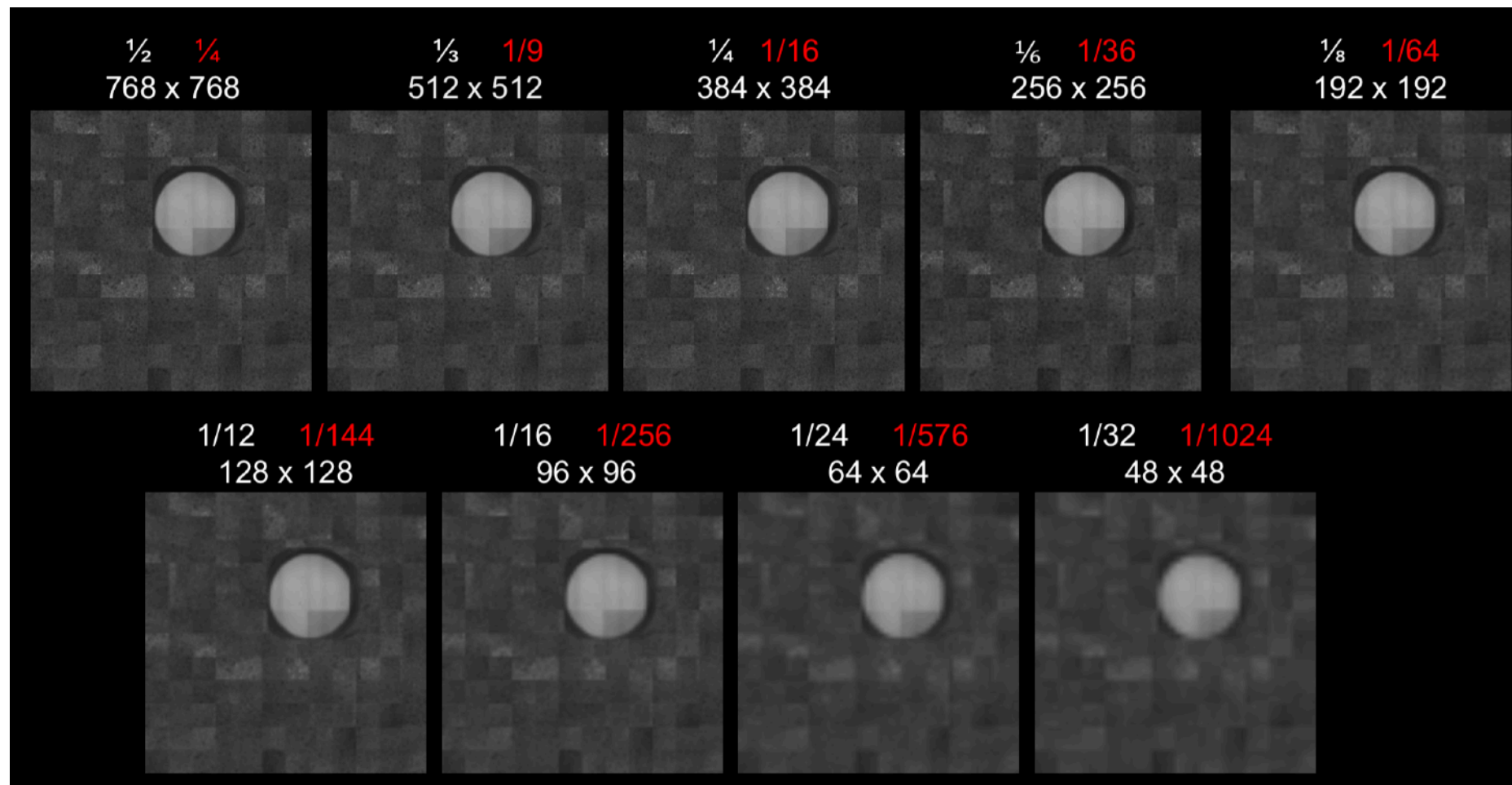


High end gaming GPU card



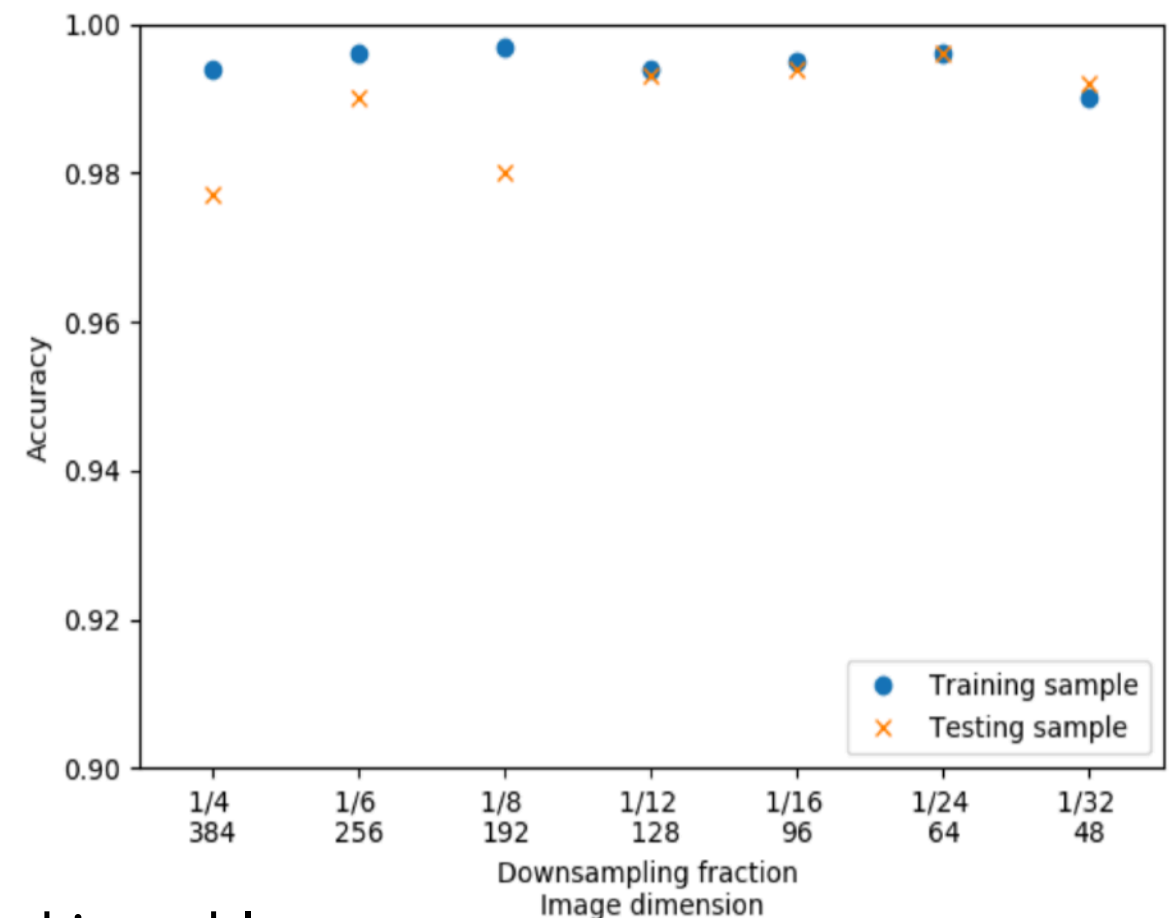
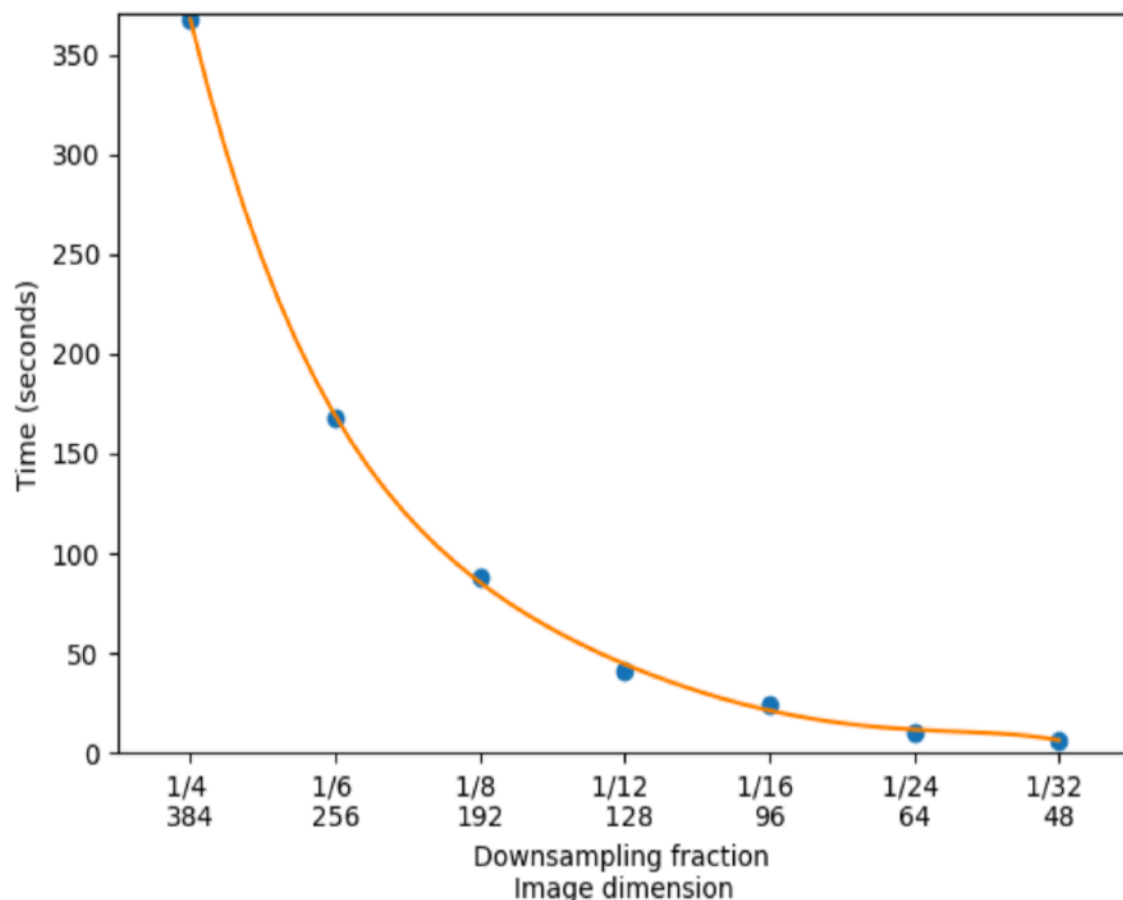
BENCHMARKING EXAMPLES: MOEDAL

- ▶ Model performance for this problem has been tested with different sized images from 48x48 pixels up to 768x768.



BENCHMARKING EXAMPLES: MOEDAL

- ▶ For this problem we find that the downsampling of information does not affect the accuracy of the model significantly.
- ▶ We would not be able to have performed this work effectively using a single core (e.g. the hardware you have been using).

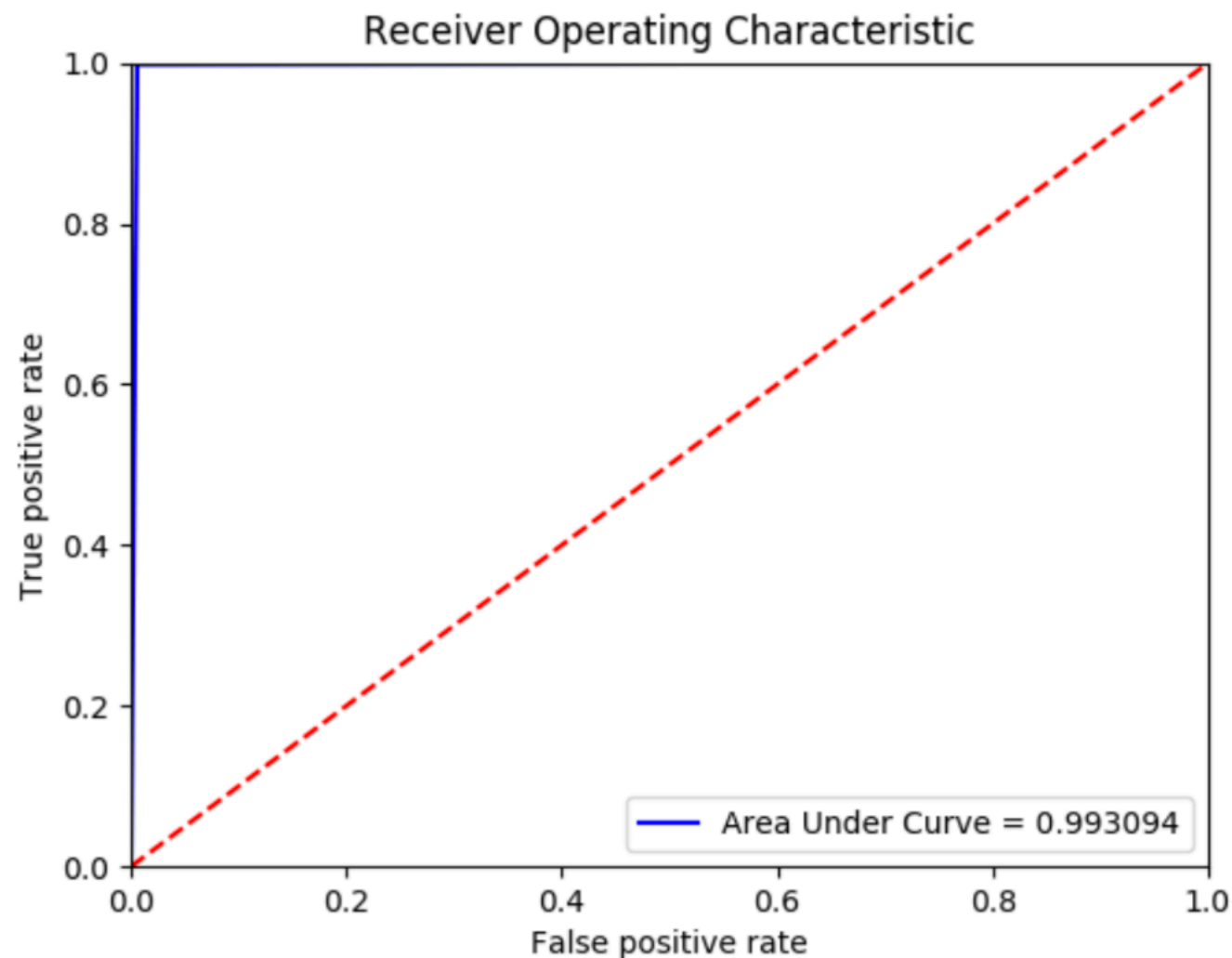


1080Ti is much faster (and cheaper) than the K40 for this problem

Approx 4x faster training performance with the T1080 c.f. the K40 for this problem.

BENCHMARKING EXAMPLES: MOEDAL

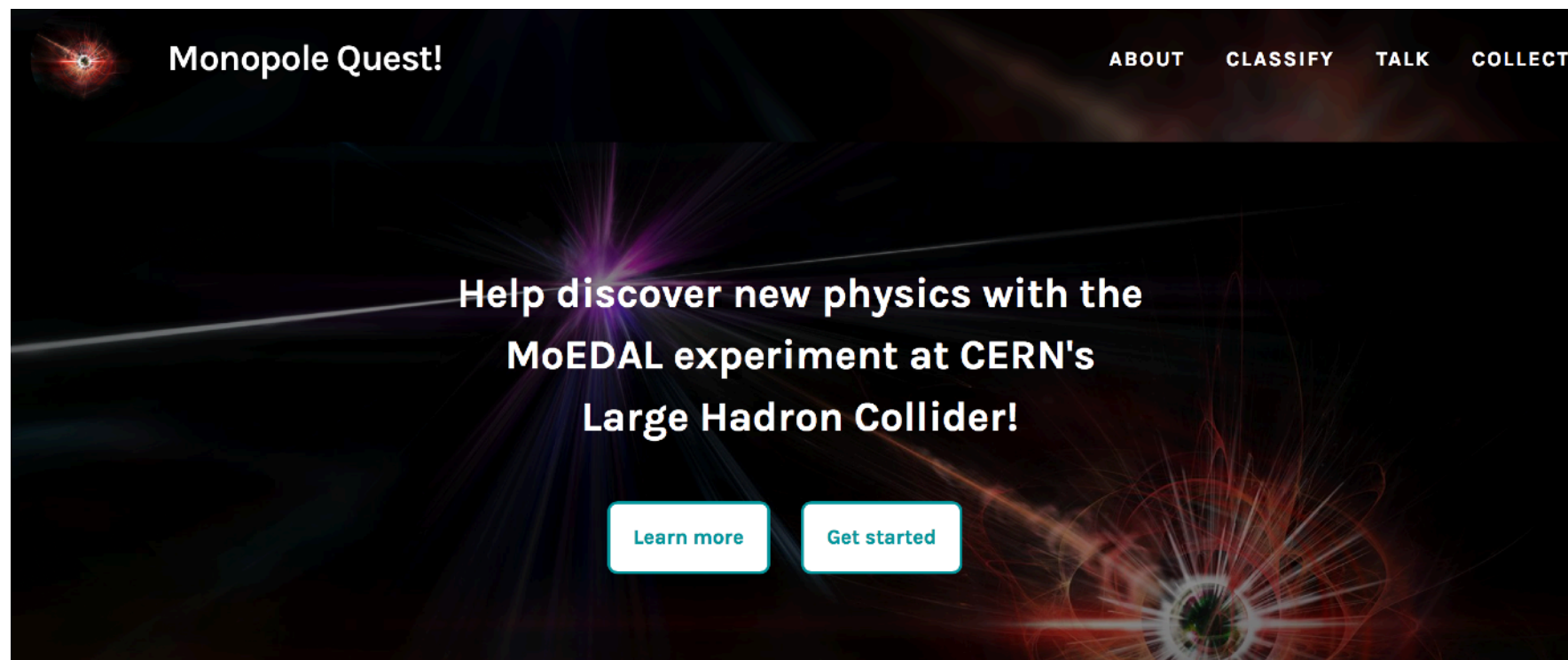
- ▶ This problem is easy for the CNN to address, so the ROC curve is not particularly informative - the AUC is 0.993.



BENCHMARKING EXAMPLES: MOEDAL

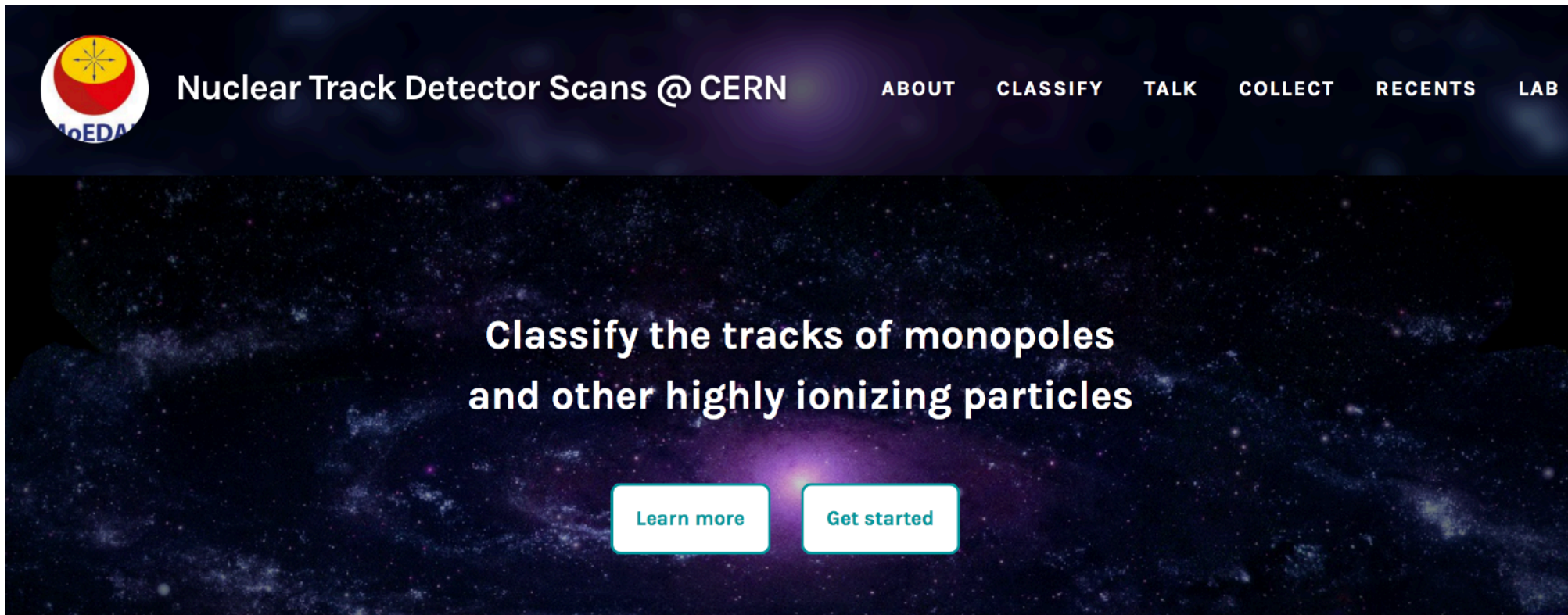
- ▶ This example was the first wave of our attempt to solve this problem.
- ▶ We now have a zooniverse project where the general public are able to help us identify what the signal and background are.
- ▶ Removes our need for simulation or image oversampling.
- ▶ See the following URL for details:

<https://www.zooniverse.org/projects/twhyntie/monopole-quest>



BENCHMARKING EXAMPLES: MOEDAL

- ▶ We are working on an update of the original zooniverse project.
- ▶ Currently this is a members only access (discussion is required within MoEDAL to allow us to recruit the general public).
- ▶ We can sign interested people up who want to contribute to classification.
- ▶ If you are interested in this please talk with Tom Charman, who will be able to register you for the project.



The screenshot shows the MoEDAL website interface. At the top left is the MoEDAL logo, a red and yellow circle with a white starburst. To its right is the text "Nuclear Track Detector Scans @ CERN". Further right is a navigation menu with the items: ABOUT, CLASSIFY, TALK, COLLECT, RECENTS, and LAB. The main content area has a dark blue background with a starry space pattern. In the center, the text reads "Classify the tracks of monopoles and other highly ionizing particles". Below this text are two buttons: "Learn more" and "Get started".

HARDWARE: FIELD PROGRAMMABLE GATE ARRAYS

- ▶ FPGAs are a popular fast configurable integrated circuit.

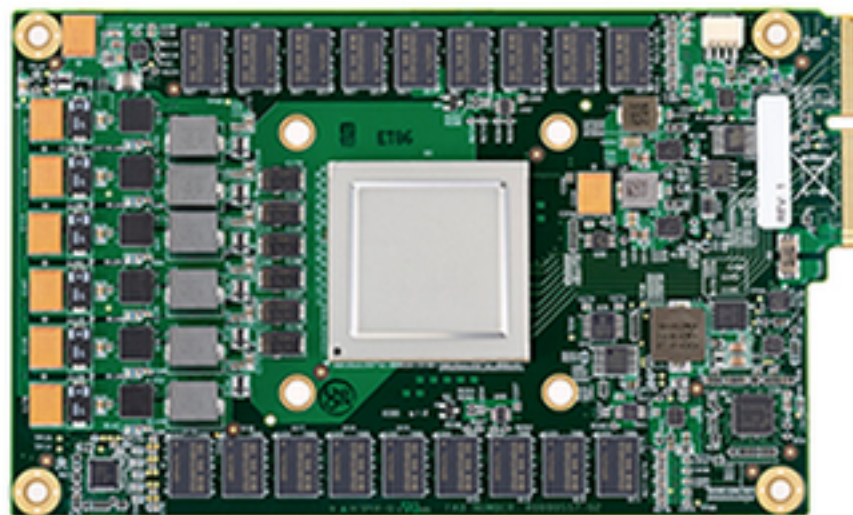
Wikipedia states: https://en.wikipedia.org/wiki/Field-programmable_gate_array

FPGAs contain an array of [programmable logic blocks](#), and a hierarchy of "reconfigurable interconnects" that allow the blocks to be "wired together", like many logic gates that can be inter-wired in different configurations. [Logic blocks](#) can be configured to perform complex [combinational functions](#), or merely simple [logic gates](#) like [AND](#) and [XOR](#). In most FPGAs, logic blocks also include [memory elements](#), which may be simple [flip-flops](#) or more complete blocks of memory.^[1] Many FPGAs can be reprogrammed to implement different [logic functions](#),^[2] allowing flexible [reconfigurable computing](#) as performed in [computer software](#).

- ▶ In the past few years, toolkits have started to appear to allow software engineers and data scientists embed AI technology in FPGA's: e.g.
 - ▶ [Intel's FPGA can be used in Azure for AI](#)
 - ▶ Other companies are also producing AI interfaces to hardware; e.g. see [XILINX](#).

HARDWARE: TENSOR PROCESSING UNITS

- ▶ AI is driven by linear algebra. The types of mathematical computation that are required to optimise and compute an AI model prediction are very specific.
- ▶ Google has developed hardware specifically optimised for this kind of computation: they call it a Tensor Processing Unit (TPU).



(left) Google's first TPU board (right) deployed in a Google data centre

<https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu>



SUMMARY

- ▶ We've touched upon some issues related to the choice of hardware resources for different problems, and looked at the application to an example from the LHC.
 - ▶ Some models are problematic to train on conventional single core (or even multi core CPUs).
 - ▶ GPUs provide significant acceleration in performance that can turn hour or days of work into minutes.
 - ▶ New ways of using AI have emerged: TPUs and FPGAs, that will ultimately change the way that we do computing, and use AI in the real world.
- ▶ Thankfully there are toolkits that have been developed to address the issue of efficiently using hardware, and this means the user can focus on model building, without the need to understand how to program a TPU/GPU/FPGA in detail.



SUGGESTED READING

- ▶ Using TensorFlow with GPU's is beyond the scope of this course, however the interested student might wish to take a look at the following:
 - ▶ https://www.tensorflow.org/guide/using_gpu
- ▶ There are two routes to using GPUs for machine learning; the first is to use tools, like TensorFlow, that have been written by computer scientists to efficiently use parallel resources.
- ▶ The second (much more challenging route) is to learn how to write code to directly address these issues. Hardware vendors provide libraries to help users do this; for example NVIDIA gives us a CUDA framework to work with, and other providers such as IBM have their own offerings.
- ▶ There are now also cloud resources that mean users don't need to invest in hardware at the outset (although a simple GPU system can be set up for a few £k to get started with using such resources).
- ▶ Also see other links in these slides and:
 - ▶ <https://www.ibm.com/it-infrastructure/power/accelerated-computing>
 - ▶ <https://www.nvidia.com/en-us/deep-learning-ai/>