Queen Mary
University of London

DR ADRIAN BEVAN

PRACTICAL MACHINE LEARNING

CLASSIFICATION PROBLEMS WITH NEURAL NETWORKS

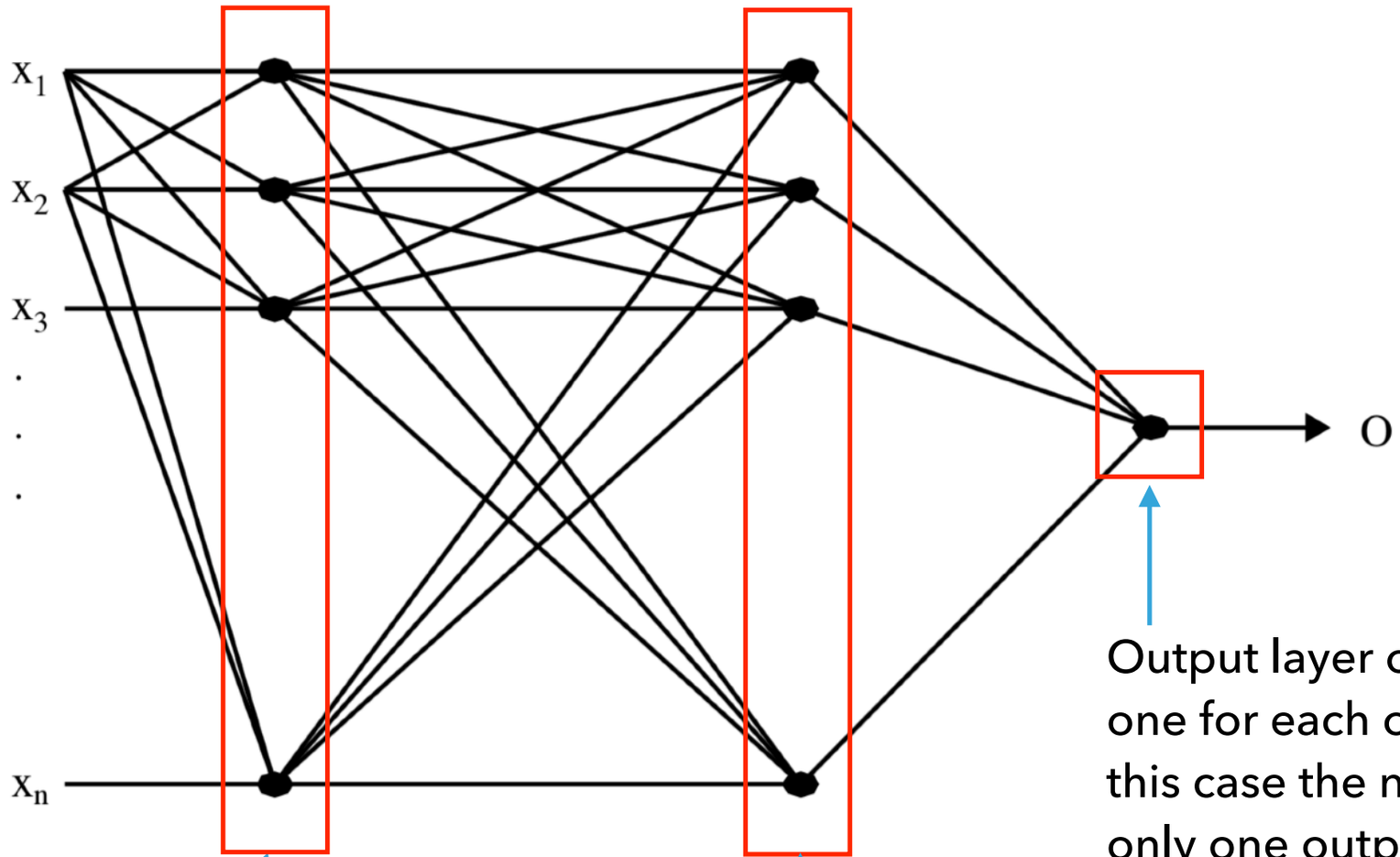
LECTURE PLAN

- Classification
- Examples
  - MNIST
    - Introduce the softmax activation function for multiclass output problems.
  - Particle physics example: Spin-parity assignment of the Higgs boson using $H \rightarrow \tau^+\tau^-$
- Summary

QMUL Summer School: https://www.qmul.ac.uk/summer-school/
Practical Machine Learning QMplus Page: https://qmplus.qmul.ac.uk/course/view.php?id=10006

CLASSIFICATION

- From the introductory NNs Lecture



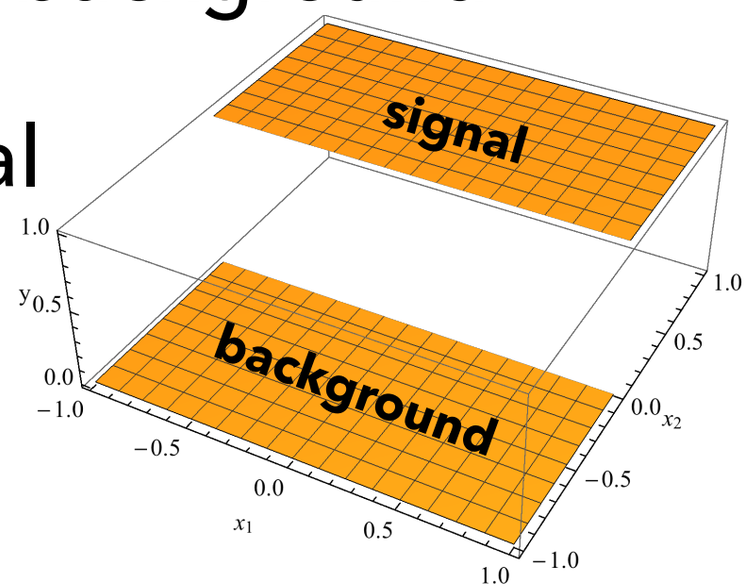
Output layer of perceptrons; one for each output type. In this case the network has only one output.

Input layer of n perceptrons; one for each dimension of the input feature space

Hidden layer of some number of perceptrons, M; at least one for each dimension of the input feature space.

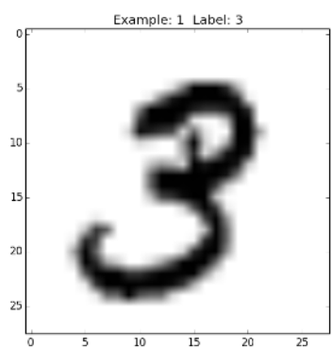
CLASSIFICATION

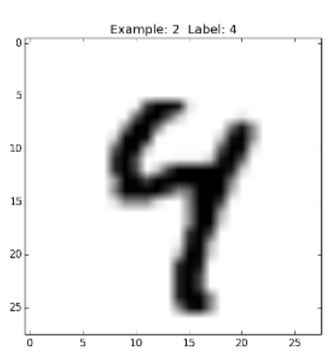
- The final perceptron in a network can be used to assign a type to a data example.
  - e.g. consider a binary activation function that gives an all or nothing response with a data sample that contains two types of event (signal and background)
    - Let nothing correspond to one type: background
    - Let all correspond to the other: signal
- Multi-class output is discussed shortly.

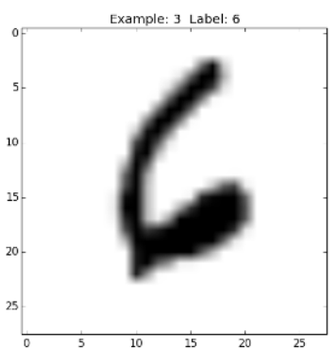


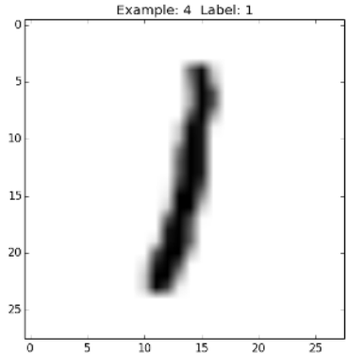
EXAMPLES: MNIST

- For more than two classification output types we need to have N_{type} perceptrons in the final output layer.
  - Each output perceptron has an all or nothing response that classifies if a training example is classified as that type or not.
  - e.g. the numbers 1, 2, 3, … 9, 0 [MNIST example]









- If we have a complete set of possible outcomes then we can use this constraint to reduce the number of perceptrons to N_{type}.
  - Assumes that the default classification for one category is given by an example not being classified as any of the others.

http://yann.lecun.com/exdb/mnist/

EXAMPLES: MNIST

- 60000 training examples
- 10000 test examples
- These are greyscale images (one number required to represent each pixel)
  - Renormalise [0, 255] on to [-1, 1] or [0, 1] for processing*.
- Each image corresponds to a 28x28 pixel array of data.
  - For an MLP this translates to 784 features.

* Depends on which activation function is being used.
http://yann.lecun.com/exdb/mnist/

EXAMPLES: MNIST

- The N_{type} = 10 perceptrons are used to make the following decisions:
  - The number 1 vs not the number 1
  - The number 2 vs not the number 2
  - The number 3 vs not the number 3
  - The number 4 vs not the number 4
  - The number 5 vs not the number 5
  - The number 6 vs not the number 6
  - The number 7 vs not the number 7
  - The number 8 vs not the number 8
  - The number 9 vs not the number 9
  - The number 0 vs not the number 0

For those with a statistical background, this is like a null hypothesis and an alternative hypothesis.

The null hypothesis provides a specific response/expectation.

The alternative hypothesis is the complement of the null.

In this context you classify an example as a specific type, or you provide a decision that it is not that type.

We will see more of the MNIST data when talking about convolutional neural networks.

http://yann.lecun.com/exdb/mnist/

EXAMPLES: MNIST

- An alternative representation is to use a softmax activation function to encode the 10 outputs in a single function.

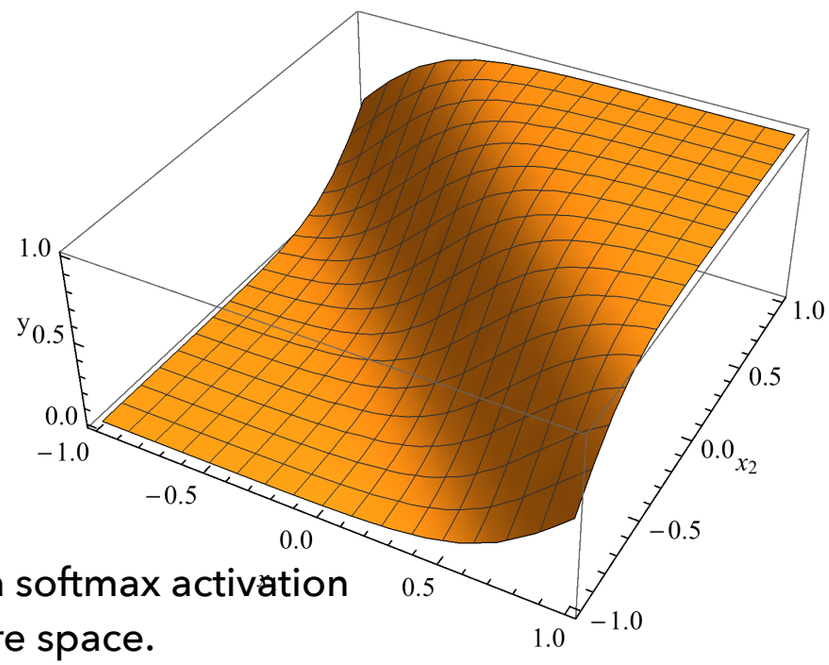
$$f_j(x) = \frac{e^{w_j^T x}}{\sum_{i=1}^{N} e^{w_i^T x}}$$

i is the index for the output classification type

The score for the i^{th} output is normalised by the sum of outputs.

$$f(x) = \frac{1}{\sum_{i=1}^{N} e^{w_i^T x}} \begin{bmatrix} e^{w_1^T x} \\ e^{w_2^T x} \\ \vdots \\ e^{w_N^T x} \end{bmatrix}$$

$f_i(x)$ is normalised to lie in the range [0, 1]



Example of the i^{th} output of a softmax activation function for a 2D input feature space.

- Can convert output to {0, 1}.

http://yann.lecun.com/exdb/mnist/

EXAMPLES: MNIST

- LeCun's website lists a number of complicated ways to train a neural network to solve this problem.
- Recent advances in computing have allowed the use of GPU's have meant that MLPs have been applied to the MNIST data, and have produced good results: error rate of 0.35% (Ciresan et al [1]).

ID	architecture (number of neurons in each layer)	test error for best validation [%]	best test error [%]	simulation time [h]	weights [milions]
1	1000, 500, 10	**0.49**	0.44	23.4	1.34
2	1500, 1000, 500, 10	**0.46**	0.40	44.2	3.26
3	2000, 1500, 1000, 500, 10	**0.41**	0.39	66.7	6.69
4	2500, 2000, 1500, 1000, 500, 10	**0.35**	0.32	114.5	12.11
5	9×1000, 10	**0.44**	0.43	107.7	8.86

[1] Neural Computation, Volume 22, Number 12, December 2010
http://yann.lecun.com/exdb/mnist/

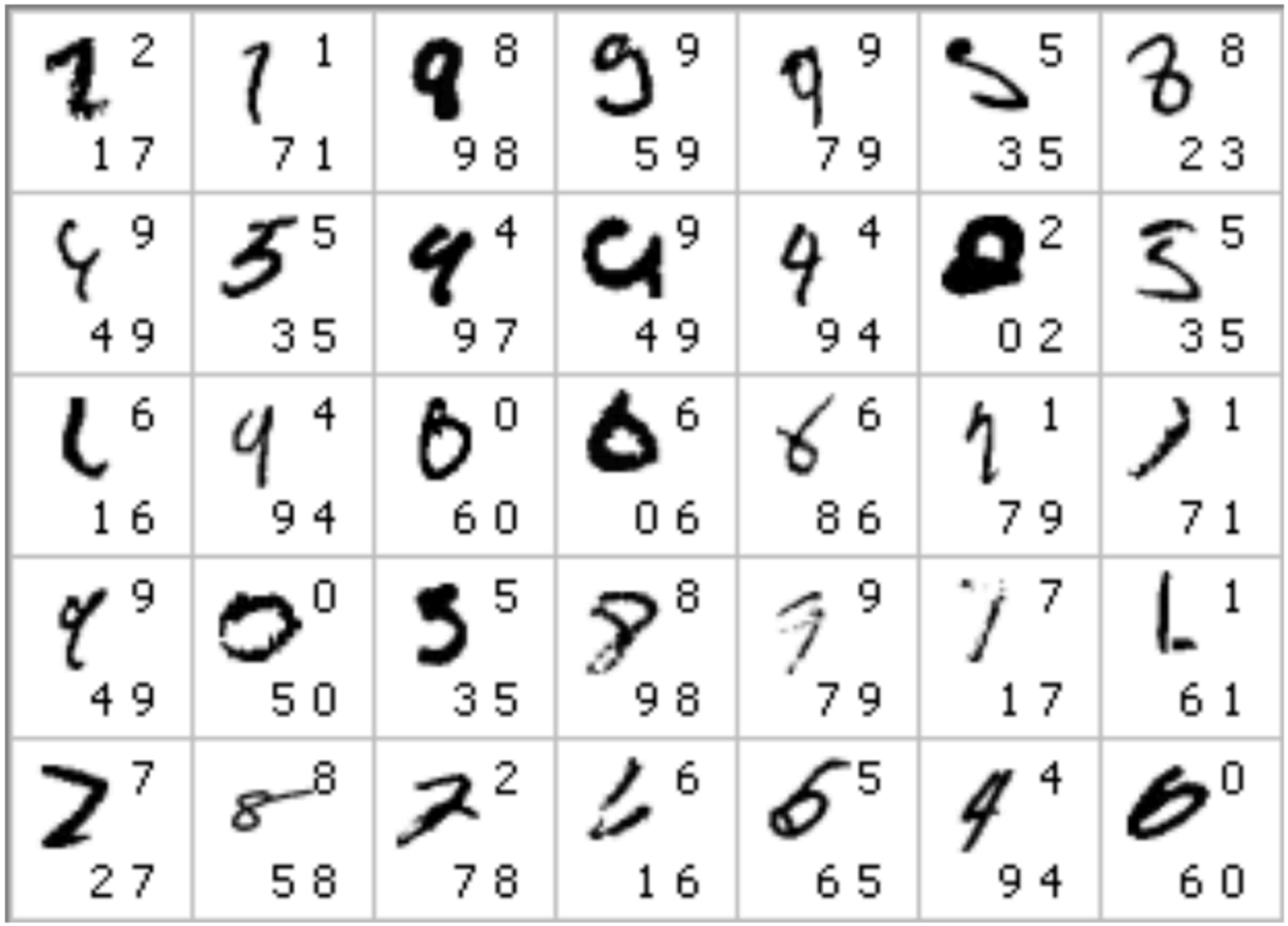
EXAMPLES: MNIST

- 35 training examples were mis-classified by the best NN architecture

Correct label

NUMBER

Two most likely predictions (L, R)



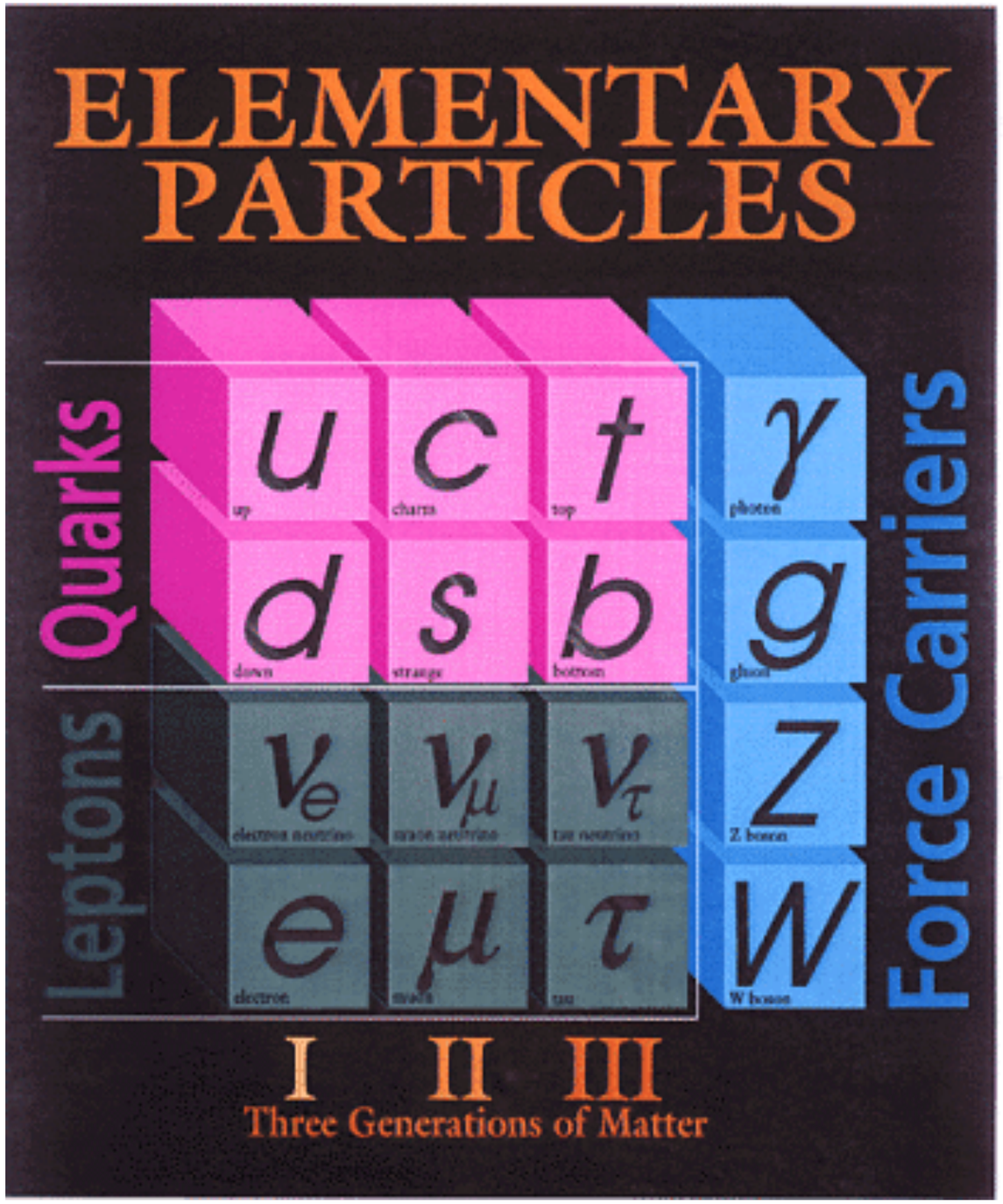
[1] Neural Computation, Volume 22, Number 12, December 2010
http://yann.lecun.com/exdb/mnist/

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

- Classification is used in a variety of particle physics scenarios:
  - Make trigger decisions [all or nothing selection]
  - Identify types of particle [provide labels for events to map against charged particle type]
  - Make perform a hypothesis test on some data (e.g. spin-parity analysis of the Higgs boson)
- First we introduce a small amount of particle physics to set the context.

EXAMPLES: SPIN–PARITY OF THE HIGGS BOSON

- Atoms are made up of a nucleus that is surrounded by one or more electrons.
- The nucleus contains neutrons and protons - each of these is made of up and down quarks.
- The force carriers play a role in:
  - g: binding nuclear material;
  - γ: binding electrons to atoms.



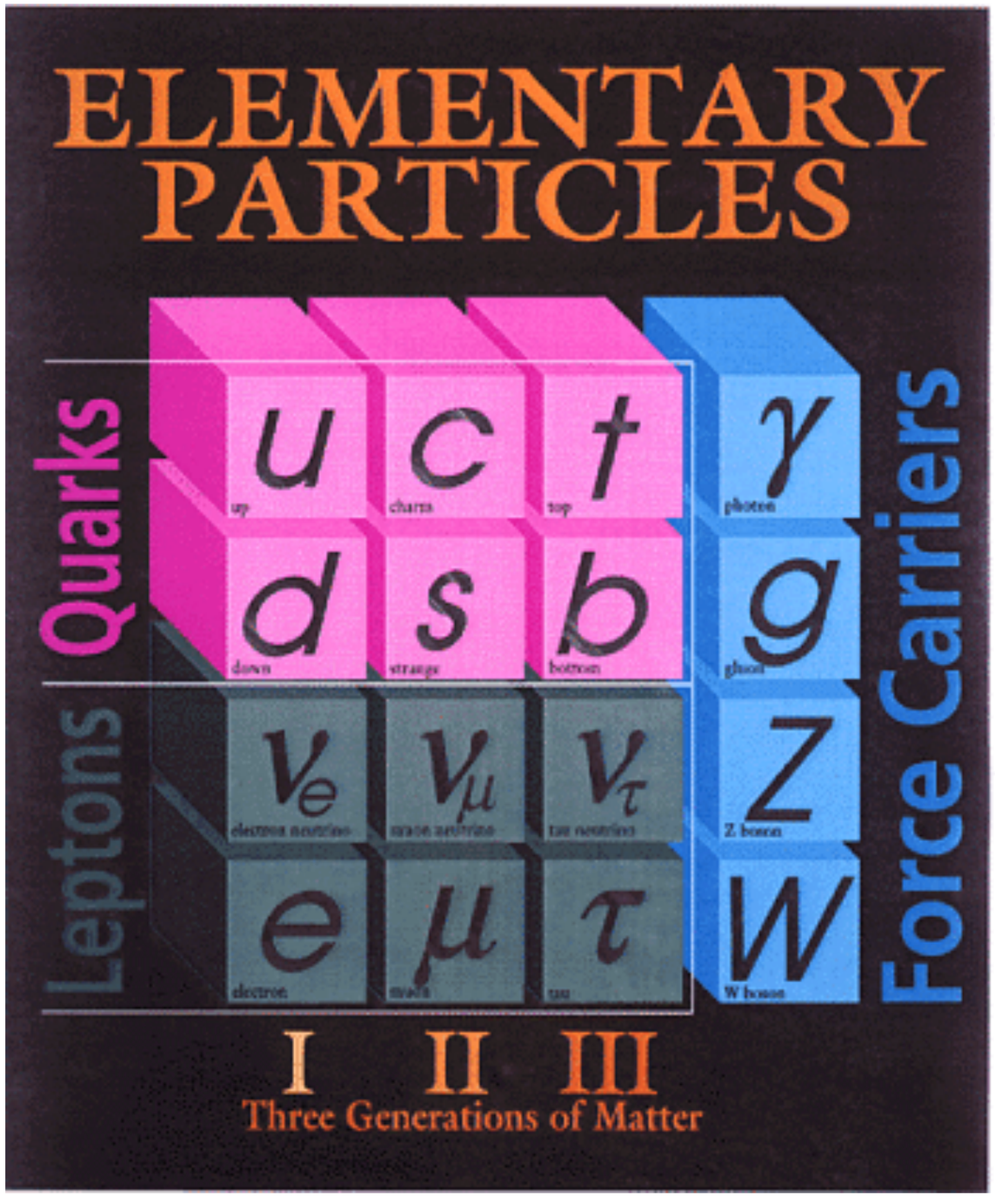
EXAMPLES: SPIN–PARITY OF THE HIGGS BOSON

- Antiparticles exist for these particles*.
- Combinations of:

$$q\bar{q}$$

$$qqq$$

- can form new particles that live for a short period of time.
  - Lifetime depends on mass, but is typically 10^{-12}-10^{-15} s.

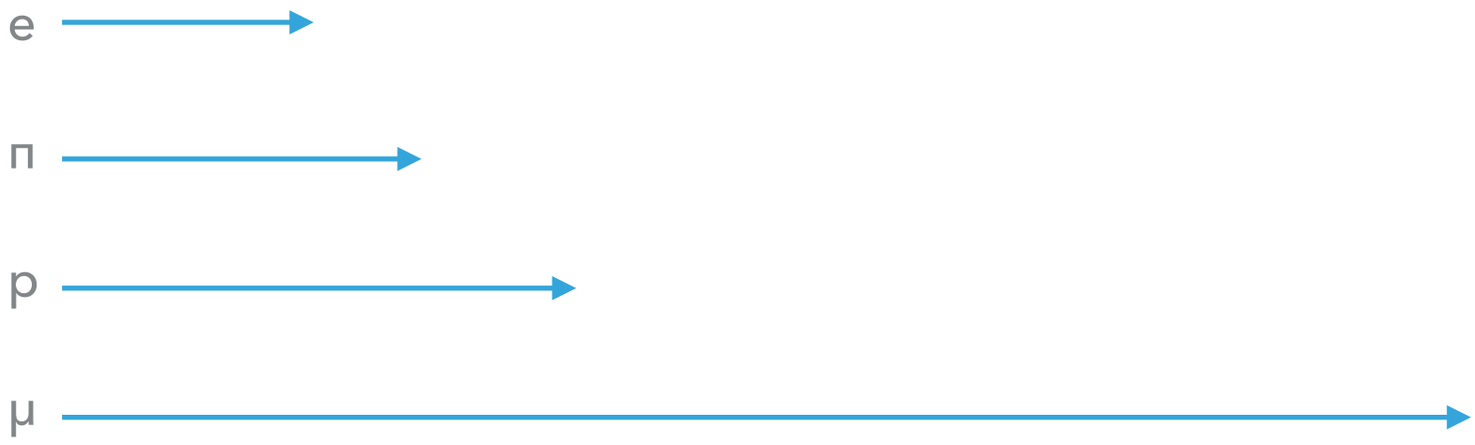


* The photon, γ, is its own antiparticle.

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

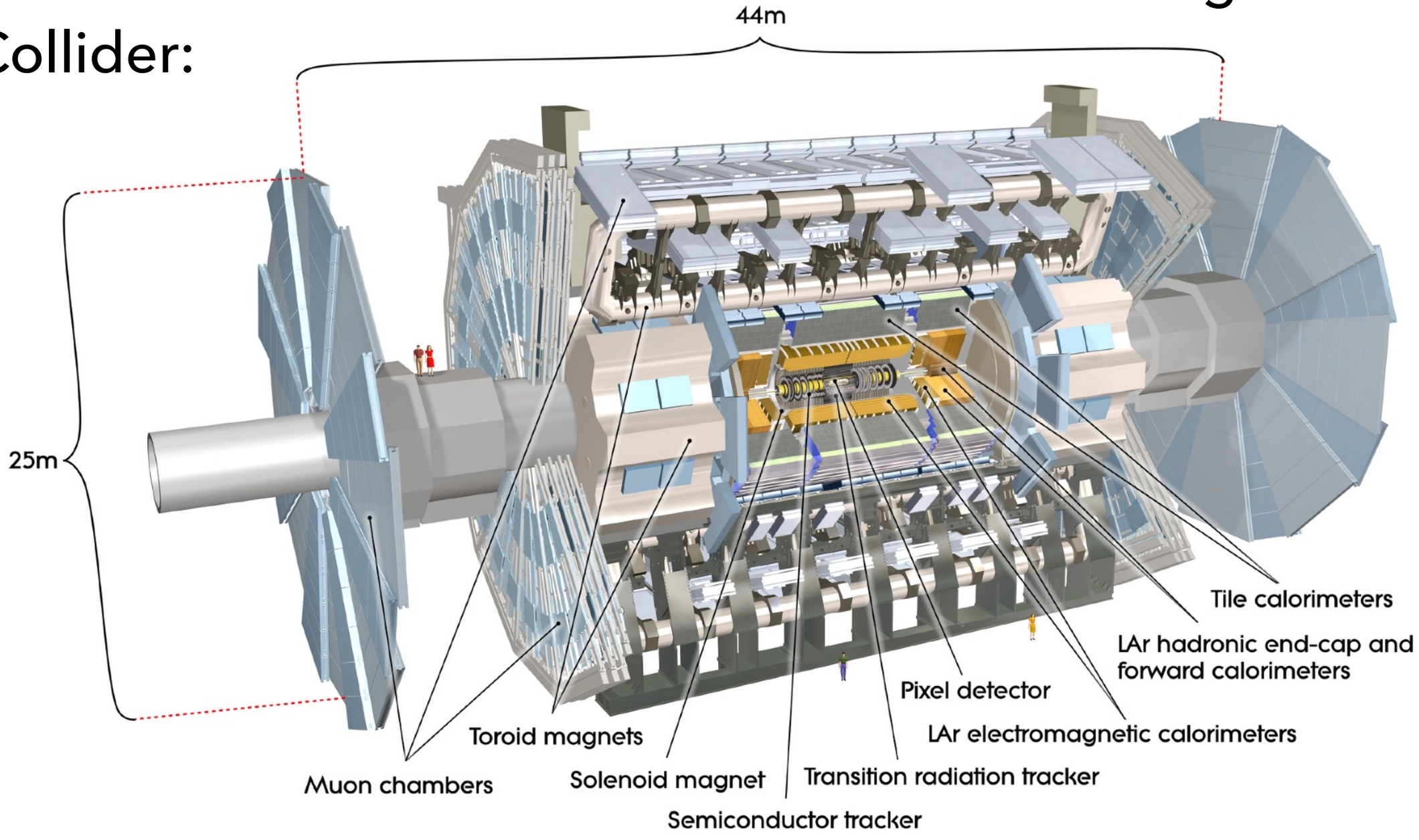
- We build large instruments to detect charged particles.
- There are sets of similar objects to be identified and we can use machine learning and AI to assist in this task.
- e.g. consider charged particles that we reconstruct as "tracks" (c.g. vapour trail indicating the passage of a plane).

Schematic (not to scale) of the penetrating power of charged particles passing through material.



EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

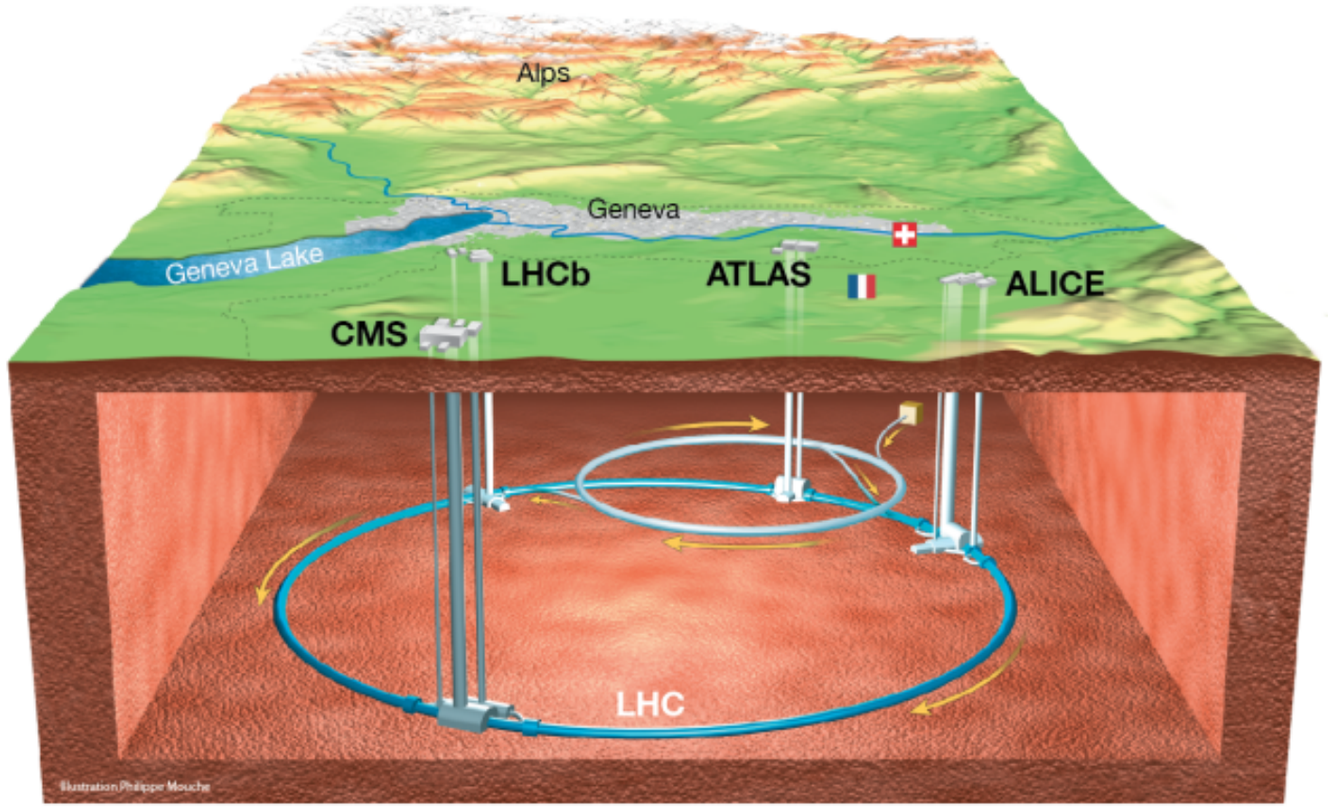
- Lets consider the ATLAS detector at CERN's Large Hadron Collider:



Each part of the detector system provides information (electronic pulses) every 25ns.

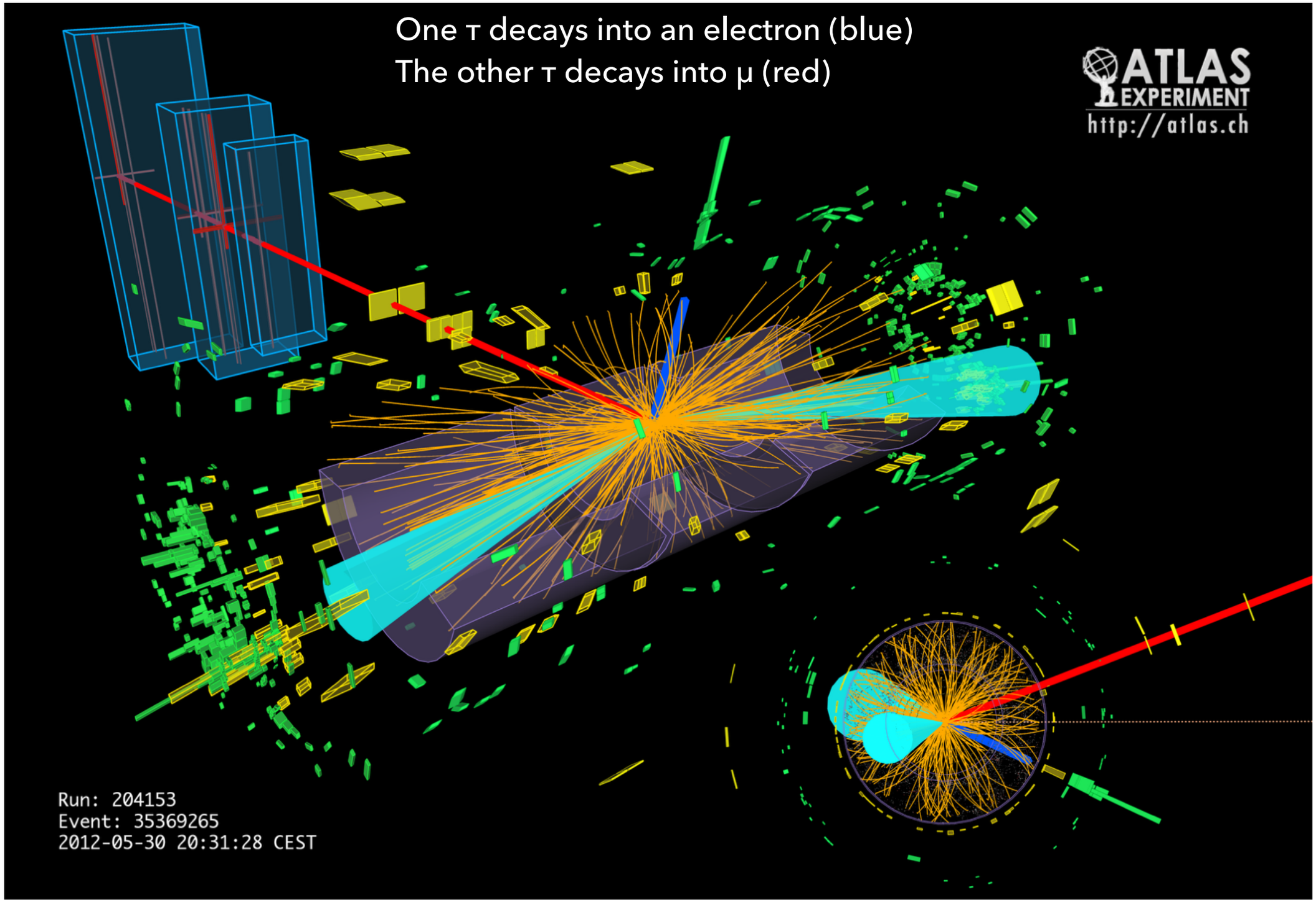
EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

- Protons are brought together in the heart of the ATLAS detector at the "interaction point".



- Many things can happen in the pp collision; we are only interested in very rare events and use pre-selection to identify them for further analysis.

EXAMPLES: SPIN–PARITY OF THE HIGGS BOSON



- This is a $H \to \tau^+\tau^-$ candidate event (example)

EXAMPLES: SPIN–PARITY OF THE HIGGS BOSON

- The Standard Model of particle physics predicts one Higgs boson, and that is a scalar particle.
  - Scalar means that the spin quantum number is zero and that the wave function of the particle ψ (related to the probability distribution) is an even function.
- ψ is a complex number and the probability is given by

$$P = \psi\psi^* = |\psi|^2$$

EXAMPLES: SPIN–PARITY OF THE HIGGS BOSON

- Spin and parity are quantum numbers associated with fundamental particles.
  - Spin is a degree of freedom, which takes the value J=0 for the Standard Model Higgs boson.
  - Parity is a degree of freedom related to the wave function of the particle;
    - P = +1: Even Parity, means the wave function is even.
    - P = −1: Odd Parity, means the wave function is odd.
- One question that needs to be addressed is:
  - What is J^P for the "Higgs boson" discovered at CERN in 2012?
  - Is it a scalar ($J^P = 0^+$), pseudo scalar ($J^P = 0^-$), or some mixture of states?
- To think in terms of hypothesis testing we can test if the Higgs boson is a scalar particle or not; where the not option includes the pseudo scalar and mixture options.

Józefowicz, Richter-Was, Was https://arxiv.org/abs/1608.02609

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

- The decay used for this classification problem is
$$H \rightarrow \tau^+\tau^-$$
- We will encounter data for this channel later in the course.
  - The τ lepton can decay into pairs of quarks, $q\bar{q}$ leptons (e, μ) and neutrinos.
  - The decay to a final state including a ρ or a_1 meson can be used to distinguish between the 0^+ and 0^- or mixture hypotheses.
  - The feature space used for the problem is reconstructed from 2 and 3 body final states of each τ decay used to reconstruct the Higgs candidate.

Józefowicz, Richter-Was, Was https://arxiv.org/abs/1608.02609

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

- For background, τ leptons can decay in many ways

Modes with one charged particle

Γ_1	particle$^-$ ≥ 0 neutrals $\geq 0K^0\nu_\tau$ ("1-prong")		(85.35 ±0.07) %	S=1.3
Γ_2	particle$^-$ ≥ 0 neutrals $\geq 0K^0_L\nu_\tau$		(84.71 ±0.08) %	S=1.3
Γ_3	$\mu^-\overline{\nu}_\mu\nu_\tau$	[a]	(17.41 ±0.04) %	S=1.1
Γ_4	$\mu^-\overline{\nu}_\mu\nu_\tau\gamma$	[b]	(3.6 ±0.4) $\times 10^{-3}$	
Γ_5	$e^-\overline{\nu}_e\nu_\tau$	[a]	(17.83 ±0.04) %	
Γ_6	$e^-\overline{\nu}_e\nu_\tau\gamma$	[b]	(1.75 ±0.18) %	
Γ_7	$h^-\geq 0K^0_L\ \nu_\tau$		(12.06 ±0.06) %	S=1.2
Γ_8	$h^-\nu_\tau$		(11.53 ±0.06) %	S=1.2
Γ_9	$\pi^-\nu_\tau$	[a]	(10.83 ±0.06) %	S=1.2
Γ_{10}	$K^-\nu_\tau$	[a]	(7.00 ±0.10) $\times 10^{-3}$	S=1.1
Γ_{11}	$h^-\geq 1$ neutralsν_τ		(37.10 ±0.10) %	S=1.2
Γ_{12}	$h^-\geq 1\pi^0\nu_\tau$ (ex.K^0)		(36.58 ±0.10) %	S=1.2
Γ_{13}	$h^-\pi^0\nu_\tau$		(25.95 ±0.09) %	S=1.1
Γ_{14} ρ	$\pi^-\pi^0\nu_\tau$	[a]	(25.52 ±0.09) %	S=1.1
Γ_{15}	$\pi^-\pi^0$non-$\rho(770)\nu_\tau$		(3.0 ±3.2) $\times 10^{-3}$	
Γ_{16}	$K^-\pi^0\nu_\tau$	[a]	(4.29 ±0.15) $\times 10^{-3}$	
Γ_{17}	$h^-\geq 2\pi^0\nu_\tau$		(10.87 ±0.11) %	S=1.2
Γ_{18}	$h^-2\pi^0\nu_\tau$		(9.52 ±0.11) %	S=1.1
Γ_{19}	$h^-2\pi^0\nu_\tau$ (ex.K^0)		(9.36 ±0.11) %	S=1.2
Γ_{20} a_1	$\pi^-2\pi^0\nu_\tau$ (ex.K^0)	[a]	(9.30 ±0.11) %	S=1.2

Modes with three charged particles

Γ_{56}	$h^-h^-h^+\geq 0$ neutrals $\geq 0K^0_L\nu_\tau$		(15.20 ±0.08) %	S=1.3
Γ_{57}	$h^-h^-h^+\geq 0$ neutrals ν_τ (ex. $K^0_S\to\pi^+\pi^-$) ("3-prong")		(14.57 ±0.07) %	S=1.3
Γ_{58}	$h^-h^-h^+\nu_\tau$		(9.80 ±0.07) %	S=1.2
Γ_{59}	$h^-h^-h^+\nu_\tau$ (ex.K^0)		(9.46 ±0.06) %	S=1.2
Γ_{60}	$h^-h^-h^+\nu_\tau$ (ex.K^0,ω)		(9.42 ±0.06) %	S=1.2
Γ_{61}	$\pi^-\pi^+\pi^-\nu_\tau$		(9.31 ±0.06) %	S=1.2
Γ_{62} a_1	$\pi^-\pi^+\pi^-\nu_\tau$ (ex.K^0)		(9.02 ±0.06) %	S=1.1
Γ_{63}	$\pi^-\pi^+\pi^-\nu_\tau$ (ex.K^0), non-axial vector		< 2.4 %	CL=95%
Γ_{64}	$\pi^-\pi^+\pi^-\nu_\tau$ (ex.K^0,ω)	[a]	(8.99 ±0.06) %	S=1.1
Γ_{65}	$h^-h^-h^+\geq 1$ neutrals ν_τ		(5.39 ±0.07) %	S=1.2
Γ_{66}	$h^-h^-h^+\geq 1\pi^0\nu_\tau$ (ex. K^0)		(5.09 ±0.06) %	S=1.2
Γ_{67}	$h^-h^-h^+\pi^0\nu_\tau$		(4.76 ±0.06) %	S=1.2
Γ_{68}	$h^-h^-h^+\pi^0\nu_\tau$ (ex.K^0)		(4.57 ±0.06) %	S=1.2
Γ_{69}	$h^-h^-h^+\pi^0\nu_\tau$ (ex. K^0, ω)		(2.79 ±0.08) %	S=1.2
Γ_{70}	$\pi^-\pi^+\pi^-\pi^0\nu_\tau$		(4.62 ±0.06) %	S=1.2
Γ_{71}	$\pi^-\pi^+\pi^-\pi^0\nu_\tau$ (ex.K^0)		(4.48 ±0.06) %	S=1.2
Γ_{72}	$\pi^-\pi^+\pi^-\pi^0\nu_\tau$ (ex.K^0,ω)	[a]	(2.70 ±0.08) %	S=1.2
Γ_{73}	$h^-\rho\pi^0\nu_\tau$			
Γ_{74}	$h^-\rho^+h^-\nu_\tau$			
Γ_{75}	$h^-\rho^-h^+\nu_\tau$			

- For this problem we are interested in:

$$\tau^\pm \to \rho^\pm\nu \to \pi^\pm\pi^0\nu$$

$$\tau^\pm \to a_1^\pm\nu \to 3\pi\nu$$

43.8% of the possible decays of the τ fit into these categories.

Józefowicz, Richter-Was, Was https://arxiv.org/abs/1608.02609

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

- Feature space used:
  - Invariant mass of the intermediate resonances (ρ, a_1).
  - Acoplanarity: angle between the decay planes of the products of each τ decay.
  - y (energy asymmetry variables):

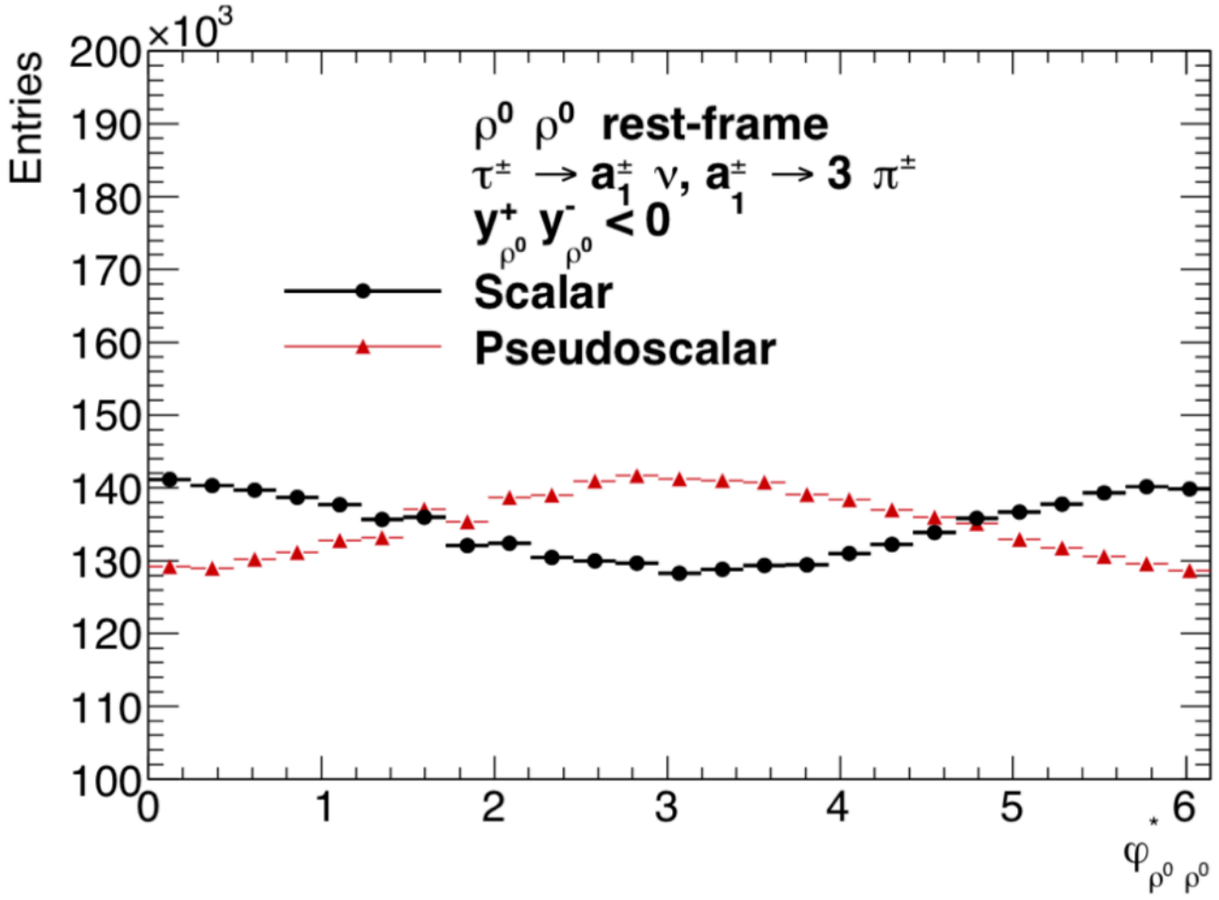
$$y^{\pm}_{\rho^0} = \frac{E^{\pi^+} - E^{\pi^-}}{E^{\pi^+} + E^{\pi^-}}, \quad y^{\pm}_{a_1} = \frac{E^{\rho^0} - E^{\pi^{\pm}}}{E^{\rho^0} + E^{\pi^{\pm}}} - \frac{m^2_{a_1} - m^2_{\pi^{\pm}} + m^2_{\rho^0}}{2m^2_{a_1}}$$

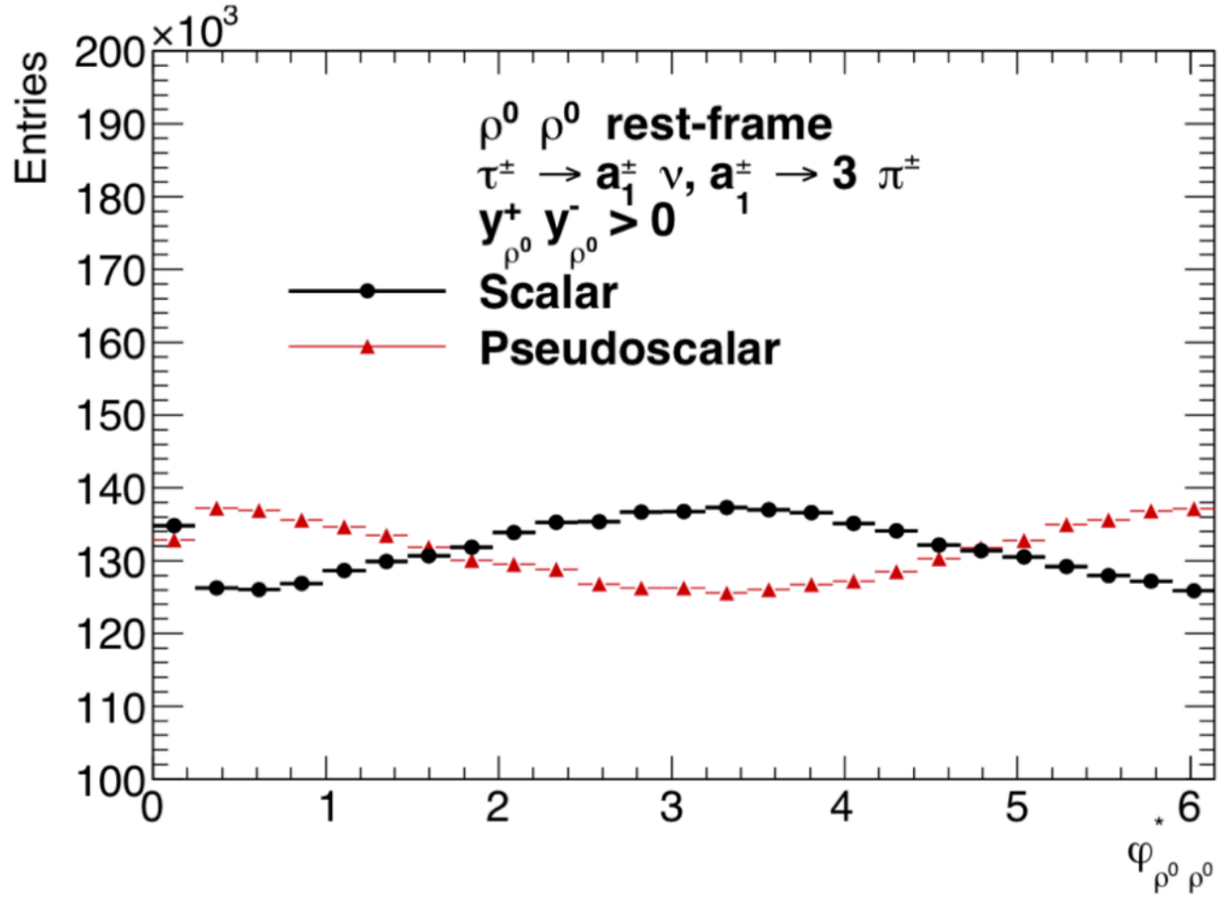
  - 4-momenta of visible decay products and intermediate resonances: $p^{\mu} = (E, \underline{p})$.

Józefowicz, Richter-Was, Was https://arxiv.org/abs/1608.02609

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

- Acoplanarity distributions for $H \to \tau^+\tau^-$ resulting in 2ρ's





Józefowicz, Richter-Was, Was https://arxiv.org/abs/1608.02609

EXAMPLES: SPIN–PARITY OF THE HIGGS BOSON

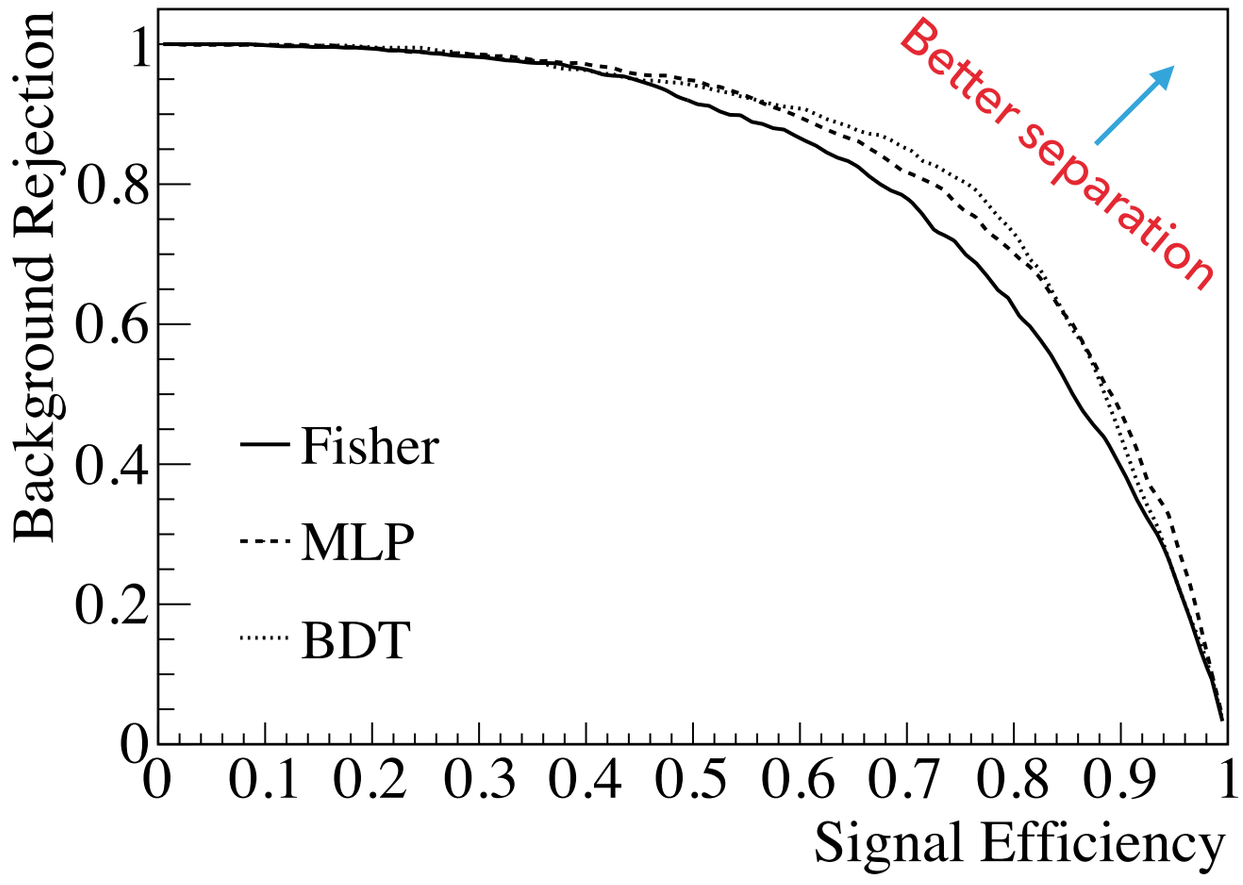
- Model: A deep neural network with 6 layers each with 300 perceptrons and a single output perceptron in the 7th layer.
  - Output perceptron uses a sigmoid activation function to compute y_h; all other perceptrons use a ReLU activation function.
- Use the Adam optimiser, with batch optimisation and dropout to improve hyperparameter training.
- Use the area under the ROC curve (AUC) as the metric to compare models. Theoretically best result is AUC = 0.782.
- Optimise a loss function that is based on the probability of each of the possible outcomes for each event:

$$-\ln p(y|y_h) = -(y=0)\ln(y_h) - (y=1)\ln(1-y_h)$$

Józefowicz, Richter-Was, Was https://arxiv.org/abs/1608.02609

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

- Aside: ROC: Receiver Operating Characteristic
  - A plot of the signal vs noise.
    - Integrate the normalised distributions of signal and background predictions as a function of model output; and plot 1-background efficiency vs signal efficiency.
    - Illustration here is for 3 different model types; we have encountered the Fisher discriminant. The MLP will be covered later in the course, and the BDT is described elsewhere [1].



[1] For example see https://pprc.qmul.ac.uk/~bevan/teaching/ATLAS-UK-ML.html

EXAMPLES: SPIN-PARITY OF THE HIGGS BOSON

Features/variables	Decay mode: $\rho^{\pm} - \rho^{\mp}$ $\rho^{\pm} \to \pi^0\,\pi^{\pm}$	Decay mode: $a_1^{\pm} - \rho^{\mp}$ $a_1^{\pm} \to \rho^0\pi^{\pm}$, $\rho^0 \to \pi^+\pi^-$ $\rho^{\mp} \to \pi^0\,\pi^{\mp}$	Decay mode: $a_1^{\pm} - a_1^{\mp}$ $a_1^{\pm} \to \rho^0\pi^{\pm}$, $\rho^0 \to \pi^+\pi^-$
$\varphi^*_{i,k}$	1	4	16
$\varphi^*_{i,k}$ and y_i, y_k	3	9	24
$\varphi^*_{i,k}$, 4-vectors	25	36	64
$\varphi^*_{i,k}$, y_i, y_k and m_i, m_k	5	13	30
$\varphi^*_{i,k}$, y_i, y_k, m_i, m_k and 4-vectors	29	45	78

Features/variables	Decay mode: $\rho^{\pm} - \rho^{\mp}$ $\rho^{\pm} \to \pi^0\,\pi^{\pm}$	Decay mode: $a_1^{\pm} - \rho^{\mp}$ $a_1^{\pm} \to \rho^0\pi^{\mp}$, $\rho^0 \to \pi^+\pi^-$ $\rho^{\mp} \to \pi^0\,\pi^{\mp}$	Decay mode: $a_1^{\pm} - a_1^{\mp}$ $a_1^{\pm} \to \rho^0\pi^{\pm}$, $\rho^0 \to \pi^+\pi^-$
True classification	0.782	0.782	0.782
$\varphi^*_{i,k}$	0.500	0.500	0.500
$\varphi^*_{i,k}$ and y_i, y_k	0.624	0.569	0.536
4-vectors	0.638	0.590	0.557
$\varphi^*_{i,k}$, 4-vectors	0.638	0.594	0.573
$\varphi^*_{i,k}$, y_i, y_k and m_i^2, m_k^2	0.626	0.578	0.548
$\varphi^*_{i,k}$, y_i, y_k, m_i^2, m_k^2 and 4-vectors	0.639	0.596	0.573

- 0.5 is a random prediction, 1.0 is a perfect prediction. These models are all somewhere in between, and none match the ideal best possible prediction.

Józefowicz, Richter-Was, Was https://arxiv.org/abs/1608.02609

SUMMARY

- We have discussed the issue of making a decision given the output of a neural network.
  - Binary classification problem using a single perceptron in the output layer.
  - Multiple perceptrons in the output layer for a multi-class output.
    - Use of softmax activation functions as an alternative to multiple perceptrons in the output layer.
- Two examples have been discussed:
  - The MNIST benchmark handwriting recognition problem.
  - Determining the J^P of the Higgs boson using $H \rightarrow \tau^+\tau^-$ decays.

SUGGESTED READING

- Discussion of event classification in text books
  - MacKay: *Information theory, inference and learning algorithms*
    - Chapter: V
  - C. Bishop: *Neural Networks for Pattern Recognition*
    - Chapter: 1
  - C. Bishop: *Pattern Recognition and Machine Learning*
    - Chapter: 1
  - T. Hastie, R. Tibshirani, J. Friedman, *Elements of statistical learning*
    - Chapter: 2, 4, 9,11
- Examples of classifiers used in particle physics:
  - Bialas, Nemeth, Richter-Wąs, *A multi-instance deep neural network classifier: application to Higgs boson CP measurement*, arXiv:1803.00838.
  - Madrazo, Cacha, Iglesias, Lucas, *Application of a Convolutional Neural Network for image classification to the analysis of collisions in High Energy Physics*, arXiv:1708.07034.
  - Abrahão et al., *Novel event classification based on spectral analysis of scintillation waveforms in Double Chooz*, arXiv:1710.04315.
  - Józefowicz, Richter-Was, Was, *Potential for optimizing Higgs boson CP measurement in $H \rightarrow \tau\tau$ decay at LHC and ML techniques*, arxiv:1608.02609.