CogX

Machine Learning in Physics

Using large datasets in the unknown

Adrian Bevan

The Alan Turing Institute

Queen Mary University of London

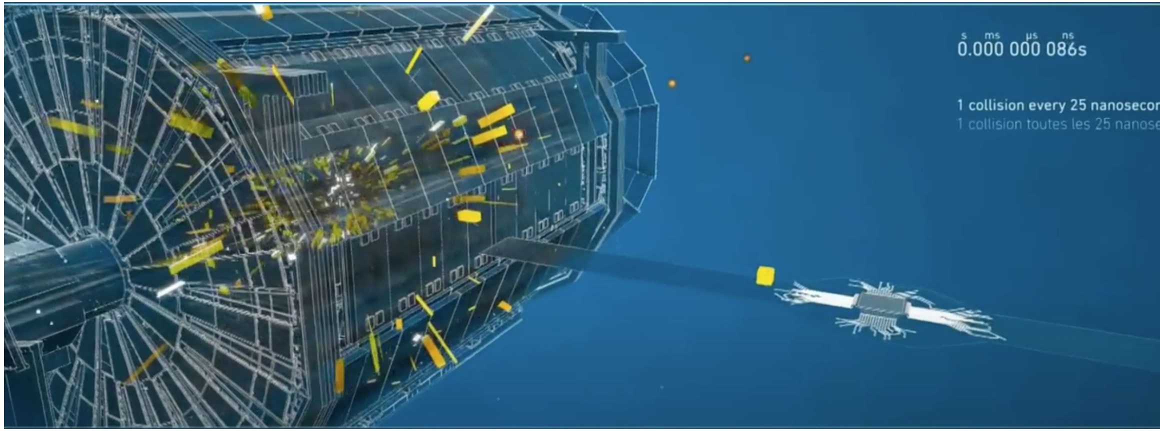
Machine Learning in Physics: Fundamental Science

Scale of big science projects continues to increase with time.

Modern particle physics experiments are 7 storeys tall.

Every 25ns a snapshot of the experiment is taken.

40M pictures of the detector every second.


s ms µs ns
0.000 000 086s
1 collision every 25 nanosecon
1 collision toutes les 25 nanose

CERN Video shown at CogX:

More videos like these can be found on the ATLAS experiment YouTube channel: https://www.youtube.com/user/TheATLASExperiment

Machine Learning in Physics: Fundamental Science

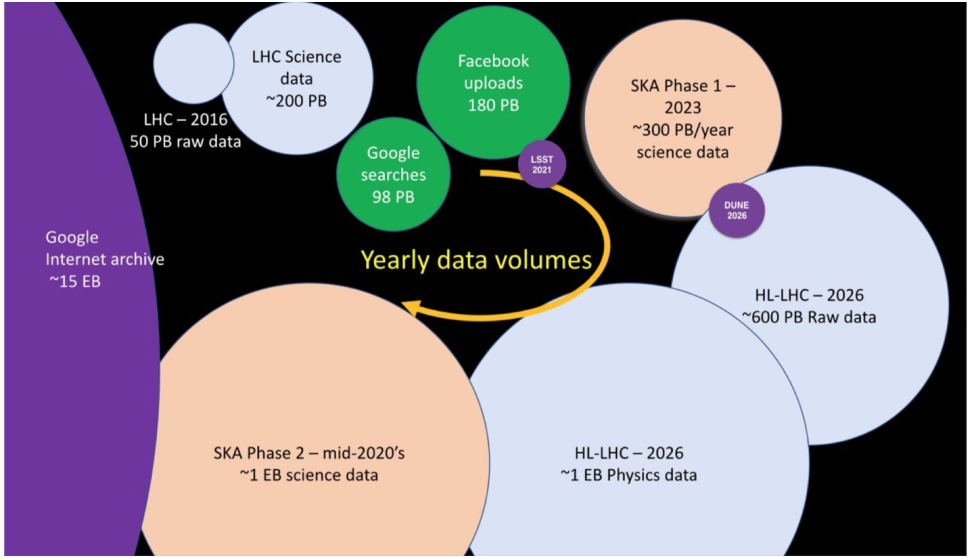
Experiment design, construction and operational lifetime runs over decades.

Initial goals lead to deeper questions.

Unrelated novel questions arise all the time that scientists need to be agile and adapt.

Data samples growing from ~10’s Tb in the late 90’s to Eb scale in ~mid 2020’s.

HEP Computing and ML roadmaps have been written for the coming decade.



E. Sexton-Kennedy: https://inspirehep.net/literature/1749123

1Eb = 1000 Pb = 1,000,000Tb

Machine Learning in Physics: Fundamental Science

The existing paradigm does not scale; we need to be creative in finding a solution.

How can machine learning help?

Event selection: Develop an understanding of how to incorporate in real time analysis of data: e.g. using FPGAs.

Event reconstruction: Use ML to reconstruct objects that feed into end of chain analysis; deep learning is already starting to help here.

End of chain analysis: identify the 1 in a billion examples of interest to statistically analyse.

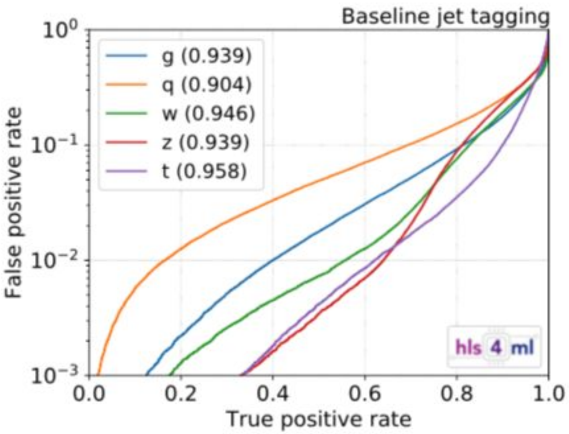
For all but the simplest problems we will need to employ machine learning broadly to help us make accurate and reliable predictions.

Event selection

Machine Learning in Physics: Fundamental Science

Trigger decisions to record events must be made quickly in collider experiments.

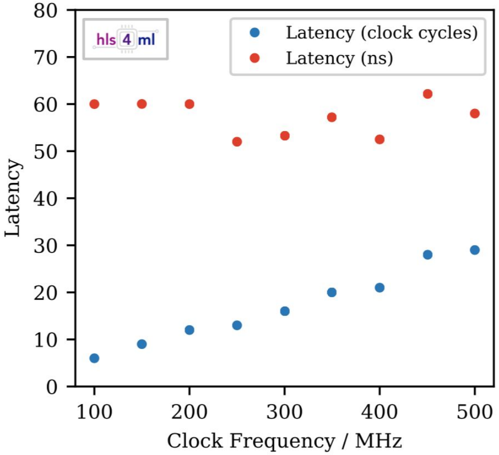
FPGA hardware provides a basis for implementations with 00’s ns latency.



(Left) Efficient binary and tertiary NN implementations of a deep network [16:64:32:32:5] on an FPGA, using TensorFlow, Keras & PyTorch.

(Right) Decision tree implementation for models learned from SciKit Learn XGBoost and TMVA.

Sub 100 ns latency achieved for inference for both methods.



Results shown using a Xilinx Virtex Ultrascale 9+ FPGA

Summers et al., https://arxiv.org/abs/2002.02534
Guglielmo et al., https://arxiv.org/abs/2003.06308

Machine Learning in Physics: Fundamental Science

Use deep learning to replace human effort for discovery science.

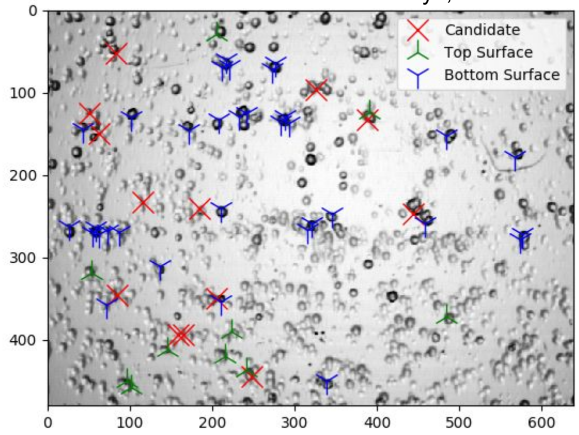
MoEDAL uses nuclear track detectors (NTDs) to search for new types of particle: e.g. magnetic monopoles.

Nanoscopic damage in the NTDs from particles created in collisions at the LHC is amplified by chemical etching.

Search for microscopic holes that could indicate the passage of new types of charged particles through the material.

Holes equivalent to 1/5th the size of a football pitch on the surface of the Earth.

AB+J. Hays, L. Milward



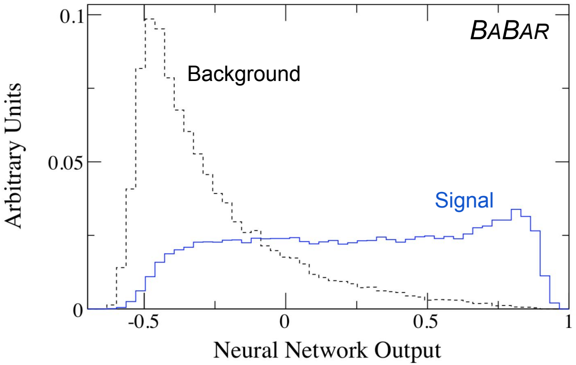
- Scalability
- Accuracy
- Reliability
- Understanding

A. Bevan, Phil.Trans.Roy.Soc.Lond. A 377 (2019) 2161, 20190392

Machine Learning in Physics: Fundamental Science

Algorithms are used to predict how signal like an example is; used for classification or inference.

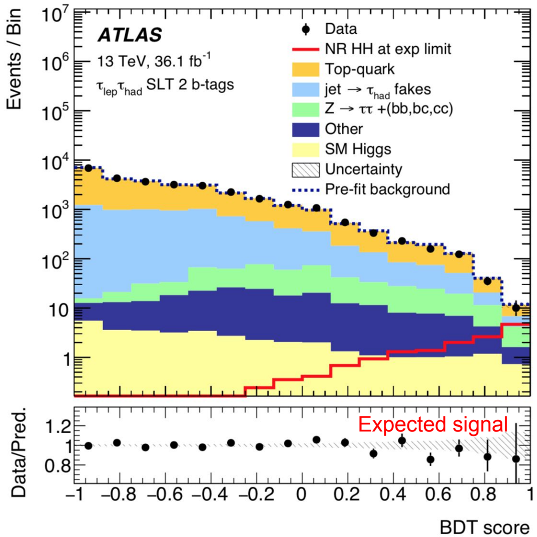
Distributions used in (un)binned likelihood fits to extract signal for analysis.



MLP output transformed and used as an input feature to a 10D unbinned maximum likelihood fit.

~700 signal : 4x10^8 background in the subsequent fit to data used to measure a parameter from the signal.

BaBar Collaboration, Phys. Rev .D76, 052007 (2007)



BDT output used as a feature in a 1D binned likelihood fit.

Search for a rare decay, aim is to separate signal from the dominant background.

ATLAS Collaboration, Phys. Rev. Lett. 121, 191801 (2018)

Machine Learning in Physics: Challenges

What is the best algorithm to use for the problem?

How do we decide which algorithm is the best for a given problem; balancing Occam's Razor against model performance in the choice of ML is not trivial?

How reproducible is the result?

Bias or large variance results in large errors on physical measurements.

Is the model overtrained?

An overtrained model affects discovery potential (linked to reproducibility).

What is the model really doing?

Does the function learned address the question we intended?

G. Carleo et al., Rev. Mod. Phys. 91, 045002 (2019).
A. Bevan, Challenges of Machine Learning for Physics, Physics Challenges for Machine Learning and Network Science, Sept 2019, QMUL, UK

Machine Learning in Physics: What’s missing?

Most Universities in the UK currently don’t teach ML formally to students, some Universities globally have set a goal of every student taking least one data science module during their degree.

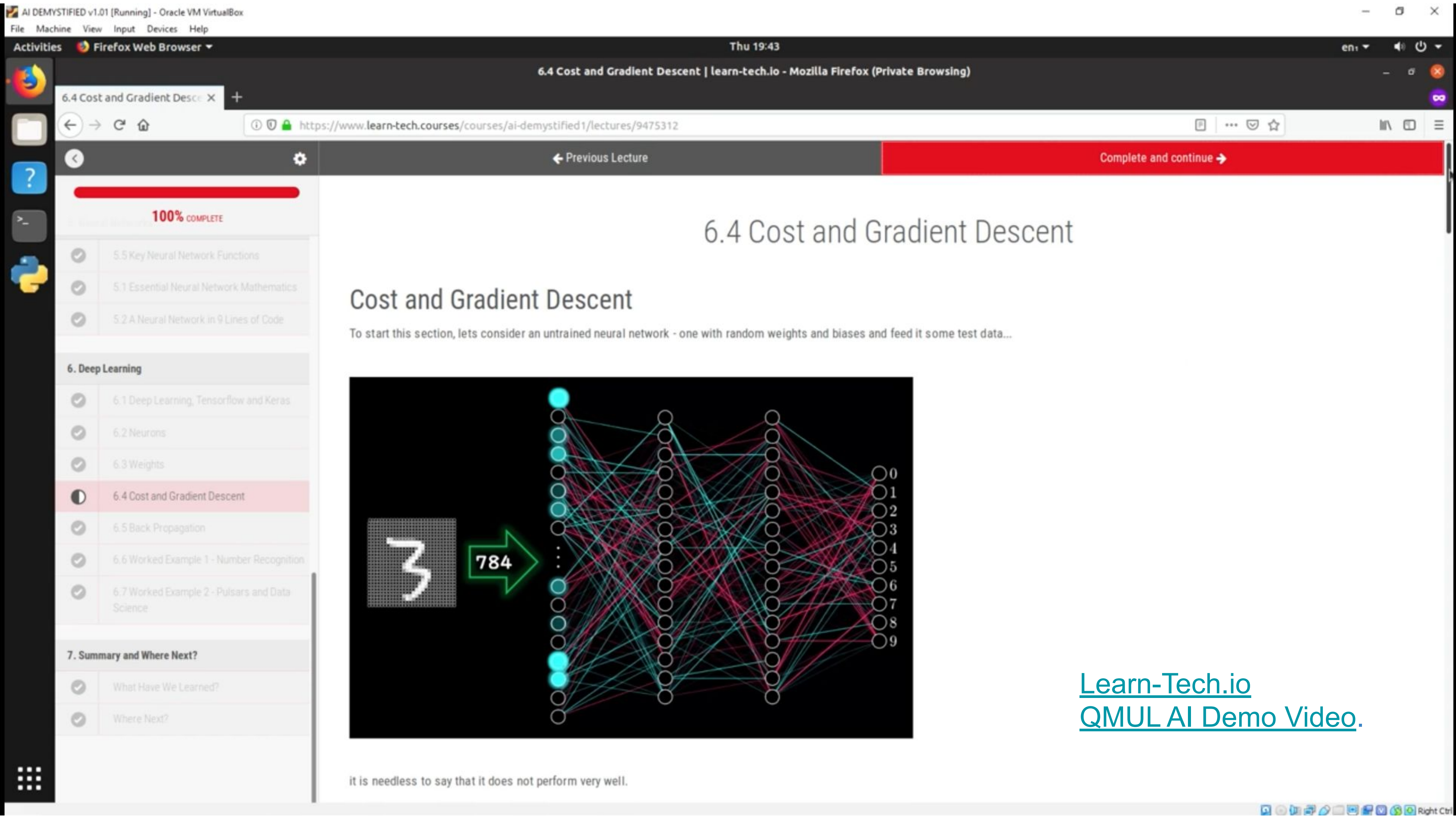
We need to change the way we educate students to do better science:

Data science as part of the curriculum for school, undergraduate and graduate students.

Quality training for a rapidly developing subject that is not core to the discipline is a challenge.

- Lots of information on the web, but not all of it is reliable.
- The ATI have developed the Turing Way handbook for reproducible data science.
- Companies such as Learn-Tech.io are working in this space: e.g. the AI Demystified training course: QMUL AI Demo Video.

… and don’t leave the wider population behind.


AI DEMYSTIFIED v1.01 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Firefox Web Browser
Thu 19:43
en
6.4 Cost and Gradient Descent | learn-tech.io - Mozilla Firefox (Private Browsing)
6.4 Cost and Gradient Desc
https://www.learn-tech.courses/courses/ai-demystified1/lectures/9475312
Previous Lecture
Complete and continue
100% COMPLETE
5.5 Key Neural Network Functions
5.1 Essential Neural Network Mathematics
5.2 A Neural Network in 9 Lines of Code
6. Deep Learning
6.1 Deep Learning, Tensorflow and Keras
6.2 Neurons
6.3 Weights
6.4 Cost and Gradient Descent
6.5 Back Propagation
6.6 Worked Example 1 - Number Recognition
6.7 Worked Example 2 - Pulsars and Data Science
7. Summary and Where Next?
What Have We Learned?
Where Next?
784
0
1
2
3
4
5
6
7
8
9
Right Ctrl

6.4 Cost and Gradient Descent

Cost and Gradient Descent

To start this section, lets consider an untrained neural network - one with random weights and biases and feed it some test data...

it is needless to say that it does not perform very well.

Learn-Tech.io
QMUL AI Demo Video.

Machine Learning in Physics: Summary

ML has been used for decades in physics, 1989 is the earliest downloadable paper with ML in the title in INSPIRE.

The past 10 years has seen a significant increase in the use of ML for analysis & more recently with deep learning.

The next 10 years: We need to see a **deeper education of physicists in data science and ML**:

Widespread use of all types of ML algorithm to a wide range of discovery problems.

A chance for developing countries to participate research in big science projects, where data is open or made available; with benefits from upskilling their local communities in data science.

Develop a deeper understanding of algorithm predictions in the context of discovery.

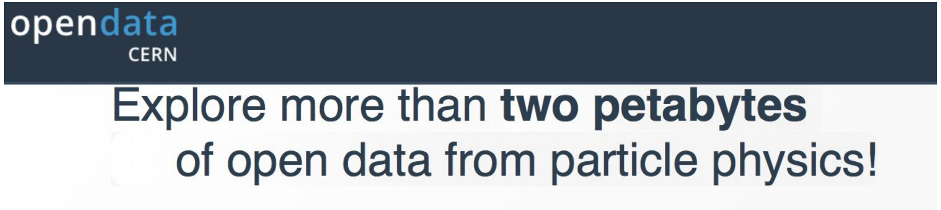
Explainability and interpretability is key for scientists to be able to widely adopt methods; this is not just a CS research problem.

Thank you for listening...

Machine Learning in Physics: Data

Want to analyse particle physics data?

See the CERN Open Data Project: http://opendata.cern.ch



opendata CERN

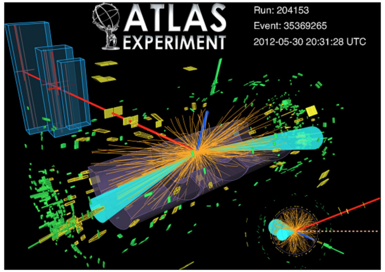
Explore more than **two petabytes** of open data from particle physics!

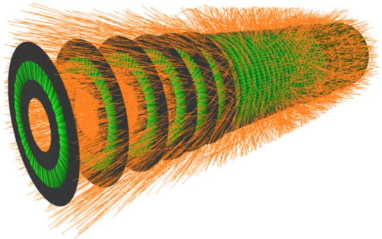
There are currently three Kaggle data challenges that use HEP data:

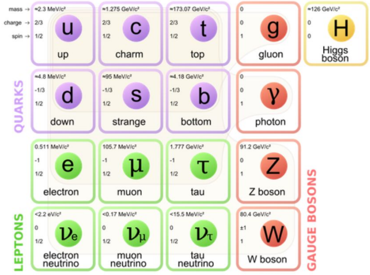
Higgs: https://www.kaggle.com/c/higgs-boson

TrackML: https://www.kaggle.com/c/trackml-particle-identification

Flavours of Physics: https://www.kaggle.com/c/flavours-of-physics






mass →
charge →
spin →
≈2.3 MeV/c² 2/3 1/2 u up
≈1.275 GeV/c² 2/3 1/2 c charm
≈173.07 GeV/c² 2/3 1/2 t top
0 0 1 g gluon
≈126 GeV/c² 0 0 H Higgs boson
≈4.8 MeV/c² -1/3 1/2 d down
≈95 MeV/c² -1/3 1/2 s strange
≈4.18 GeV/c² -1/3 1/2 b bottom
0 0 1 γ photon
0.511 MeV/c² -1 1/2 e electron
105.7 MeV/c² -1 1/2 μ muon
1.777 GeV/c² -1 1/2 τ tau
91.2 GeV/c² 0 1 Z Z boson
<2.2 eV/c² 0 1/2 νe electron neutrino
<0.17 MeV/c² 0 1/2 νμ muon neutrino
<15.5 MeV/c² 0 1/2 ντ tau neutrino
80.4 GeV/c² ±1 1 W W boson
QUARKS
LEPTONS
GAUGE BOSONS

The Standard Model of elementary particles

The Alan Turing Institute

Queen Mary University of London

Machine Learning in Physics: Particles

Matter is made up of fundamental particles:

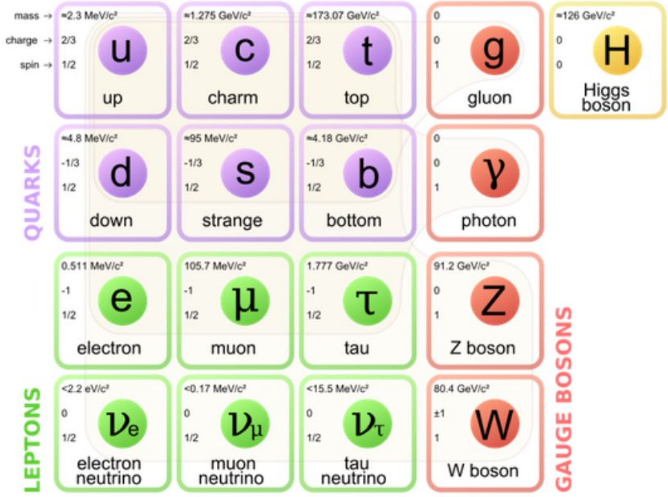
- Quarks
- Leptons
- Bosons: Force carriers and the Higgs

Interactions between these particles govern the sub-atomic behaviour of the Universe.

- Known particles don't provide a fully consistent explanation of nature

E.g. Matter and its counterpart anti-matter were created in equal amounts in the big bang.

- We live in a matter dominated Universe.
- Matter-antimatter differences don't explain the missing antimatter.
- The answer to this problem, sadly, is not 42.



The Standard Model of elementary particles

Machine Learning in Physics: Sli.do questions:

1. **What do you think the most important barrier to overcome in adopting ML for physics discovery in the coming decade is?**
   a. Resources
   b. Tools
   c. Education
   d. Other

2. **What do you think the most important challenge is with applying ML for physics discovery in the coming decade?**
   a. Choice of model
   b. Reproducibility
   c. Training models
   d. Explainability and interpretability

3. **Is the use of an overtrained model in a publication just bad science or scientific misconduct?**
   a. Yes
   b. No
   c. Possibly
   d. Don’t know

The Alan Turing Institute

Queen Mary University of London