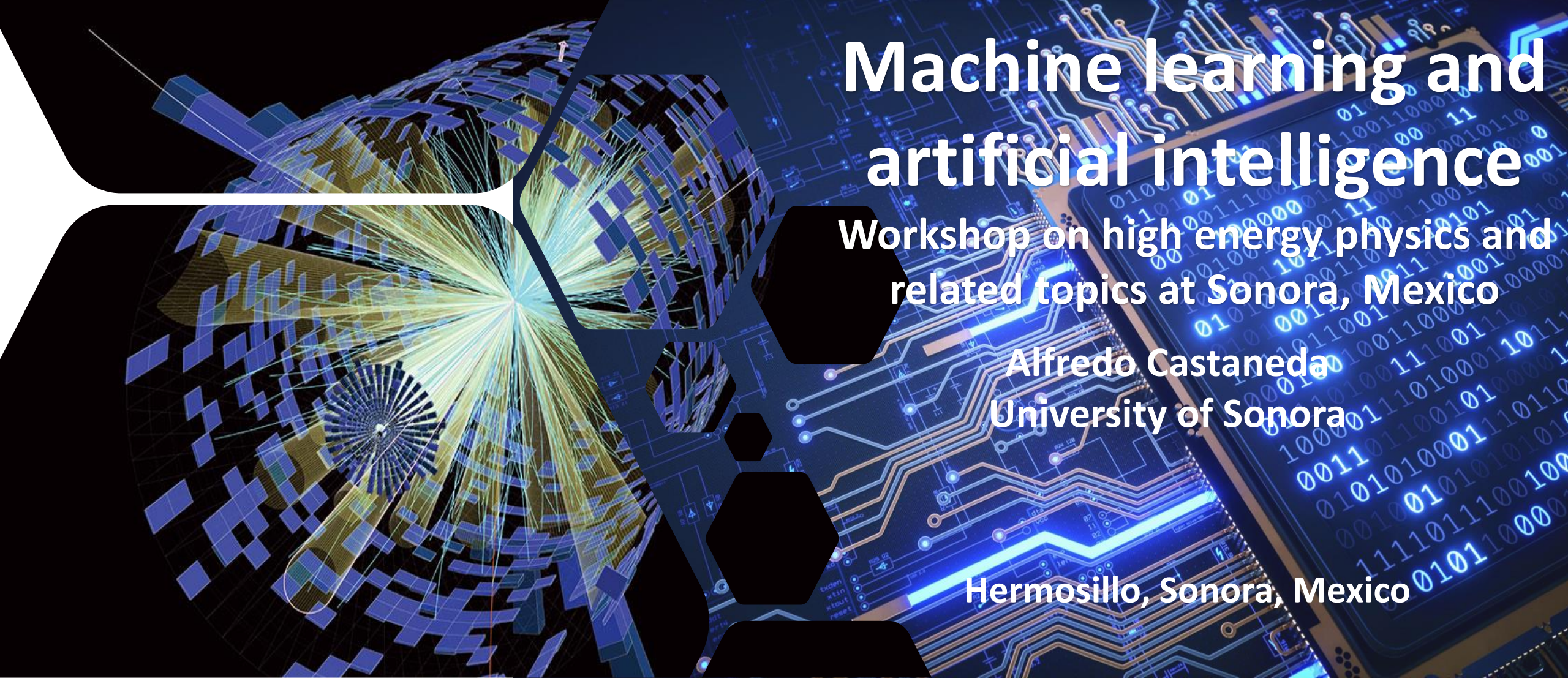
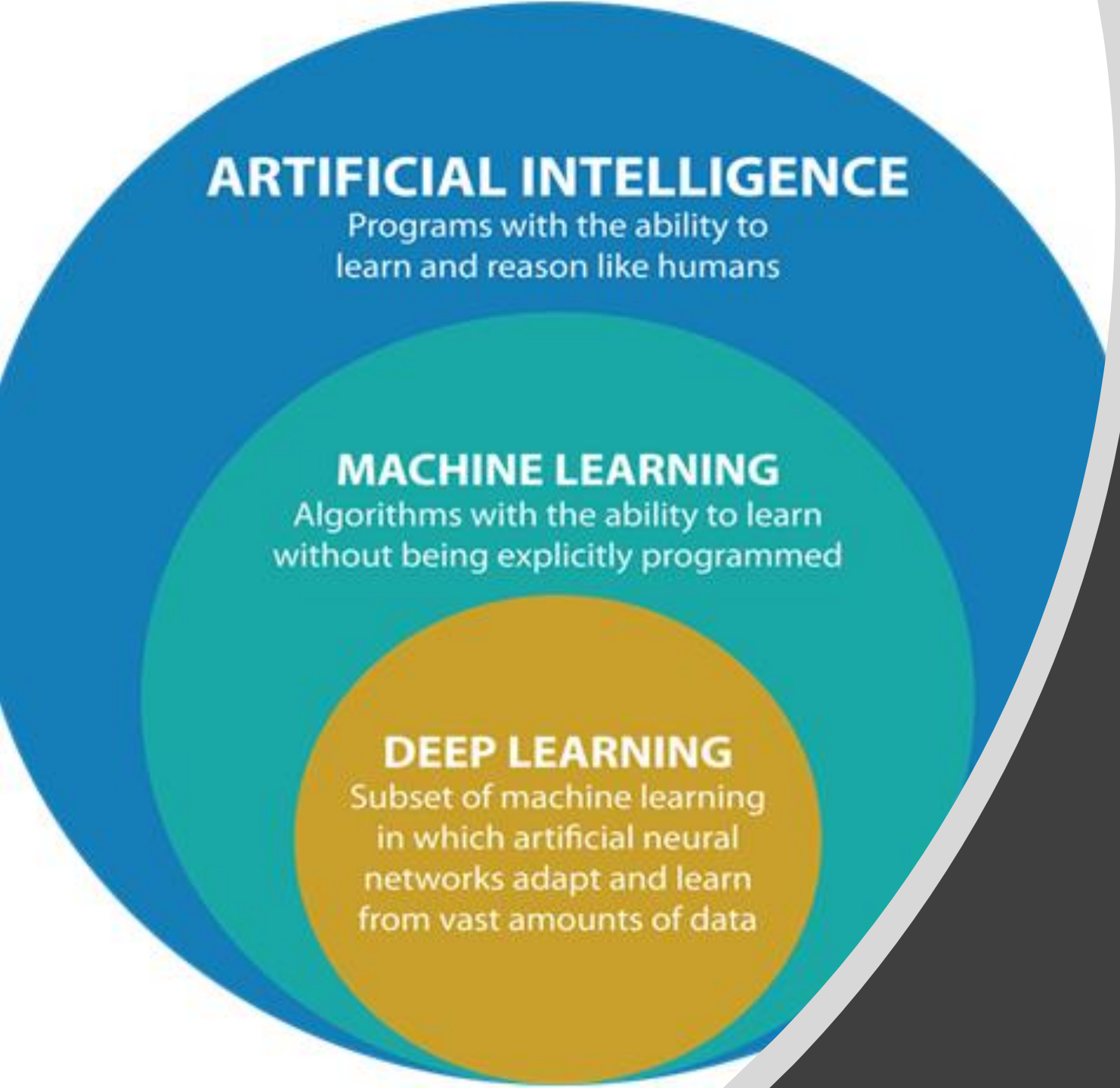


Machine learning and artificial intelligence

Workshop on high energy physics and related topics at Sonora, Mexico

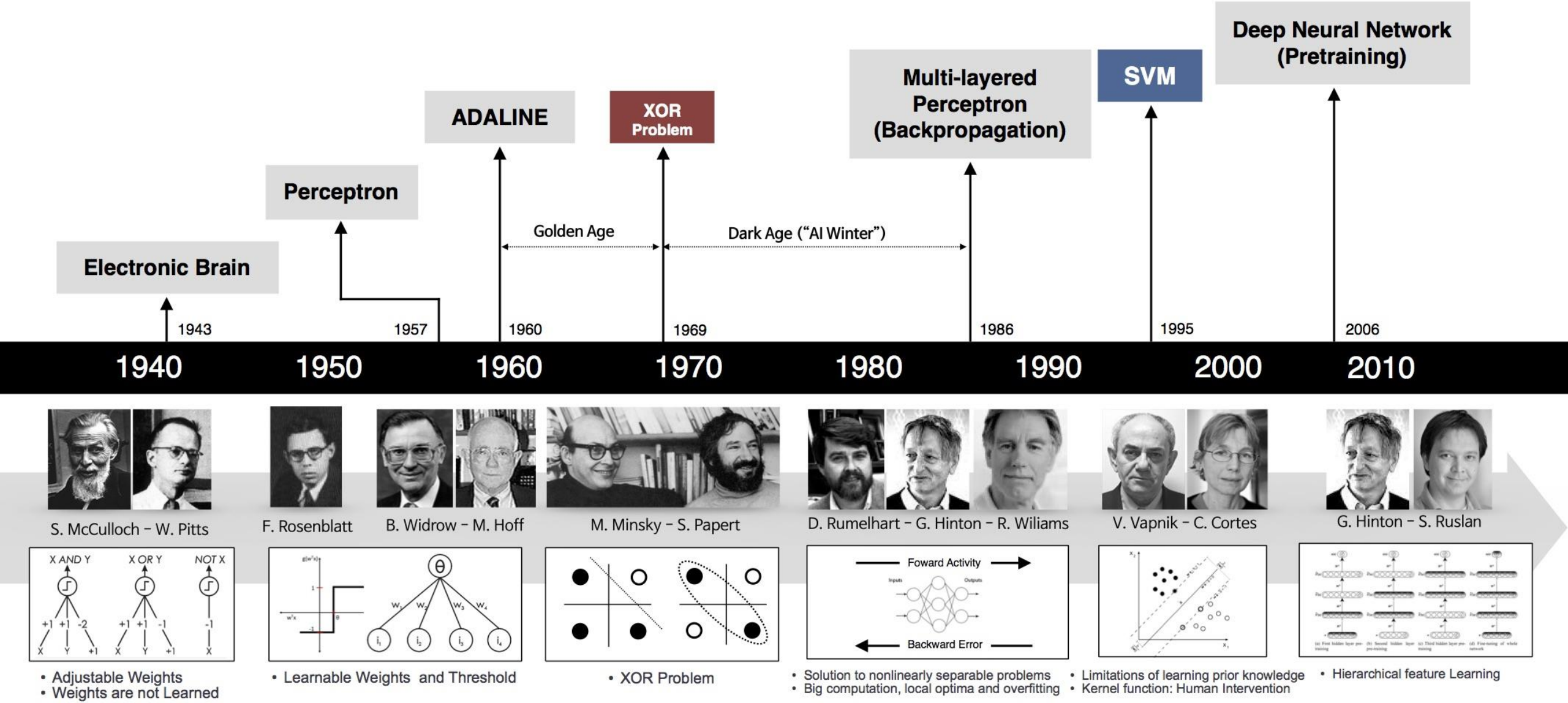
Alfredo Castaneda
University of Sonora

Hermosillo, Sonora, Mexico



What is AI/ML/DL?

DL development in time



- Adjustable Weights
- Weights are not Learned

- Learnable Weights and Threshold

- XOR Problem

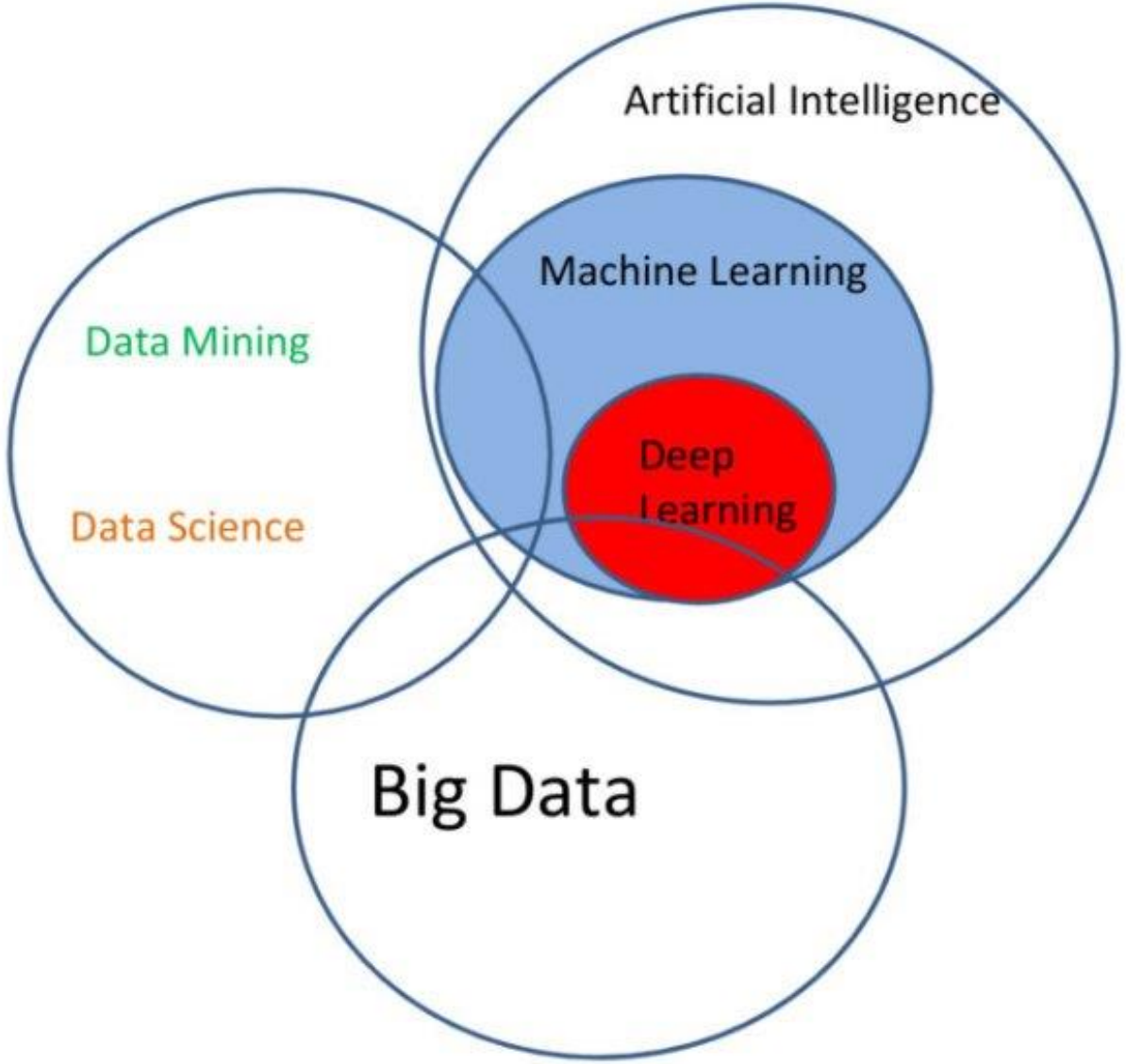
- Solution to nonlinearly separable problems
- Big computation, local optima and overfitting

- Limitations of learning prior knowledge
- Kernel function: Human Intervention

- Hierarchical feature Learning

What is the connection between big data and Deep learning

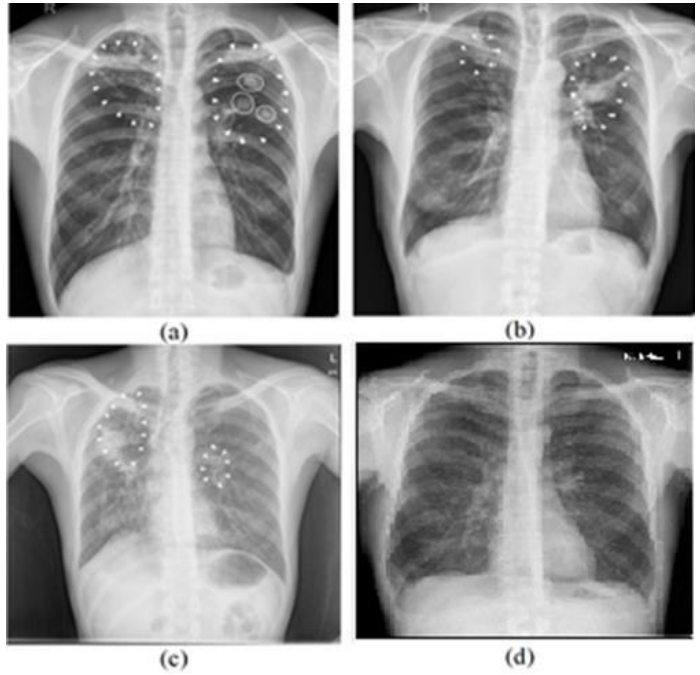
- Deep learning needs large datasets to train their neural networks
- HEP experiments produce large amounts of data each year
- Several task can be improved using DL, for instance classification of data, classifications of images, anomaly detection, etc..



Areas of opportunity for DL applications

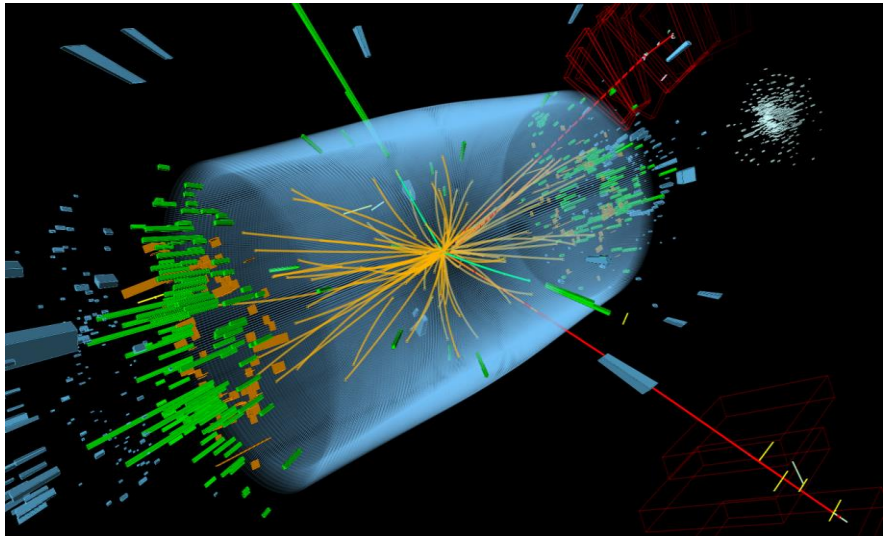
Medical Physics

Image classification for an early detection of different anomalies



High energy Physics

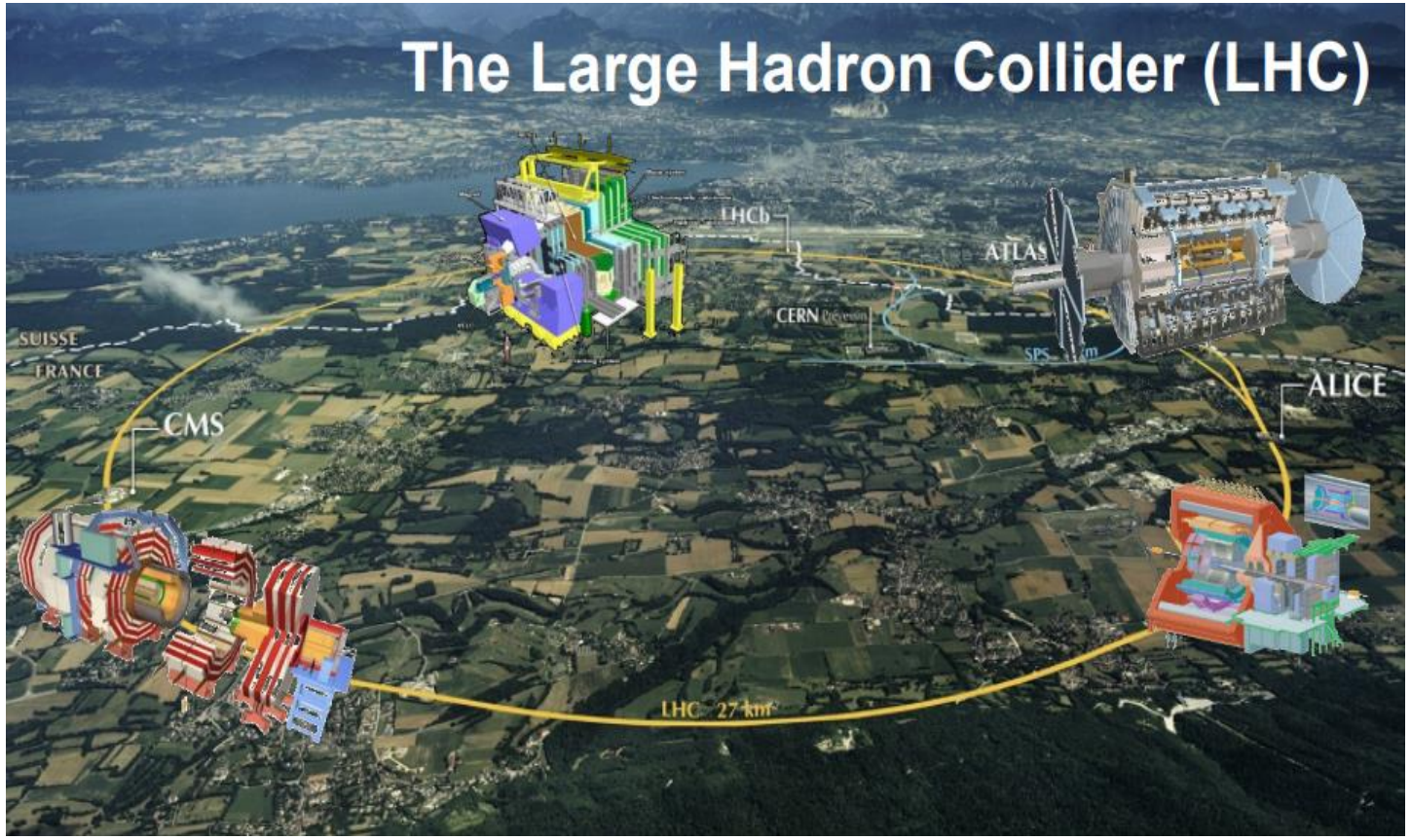
To optimize the data collection, particle identification, signal vs bkg separation

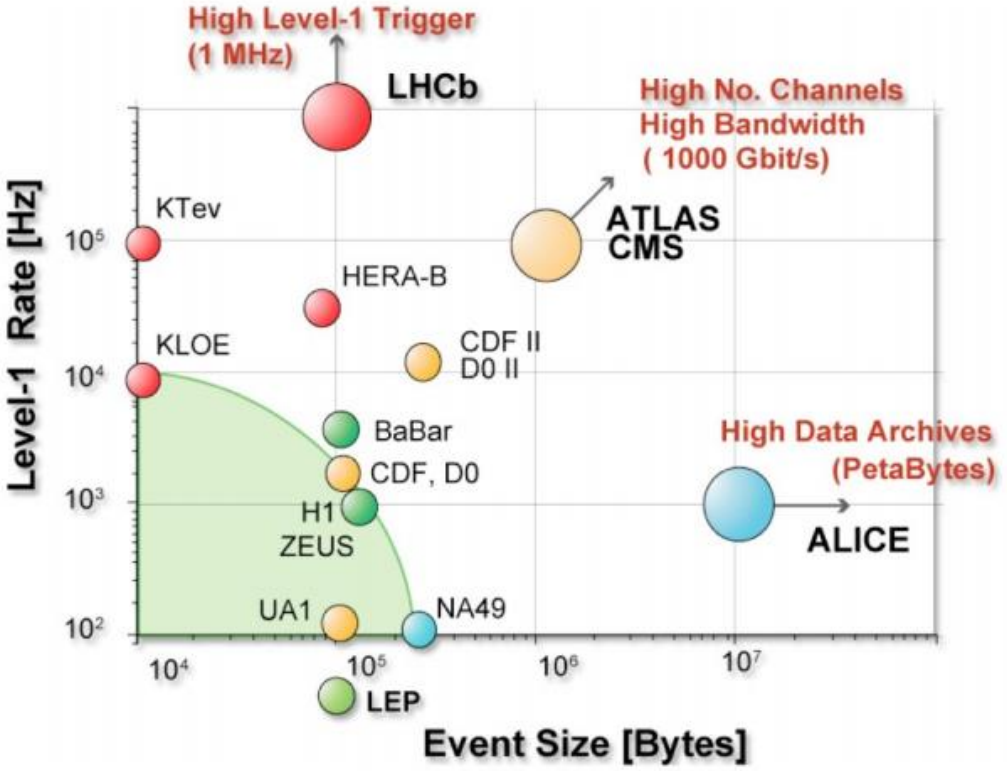


Fast electronics

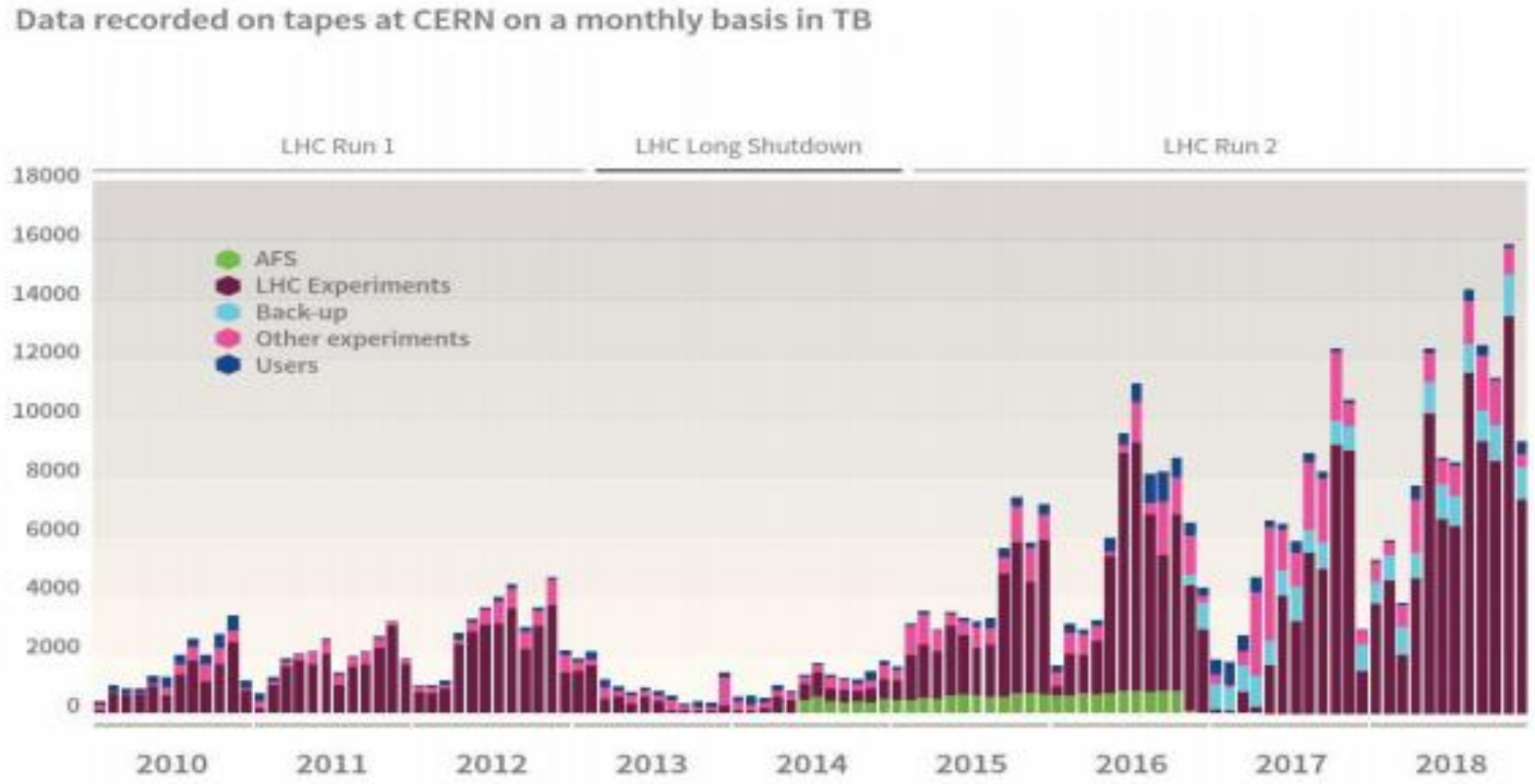
Smart signal processing used in different science and private sector fields

The Large Hadron Collider: A big data experiment



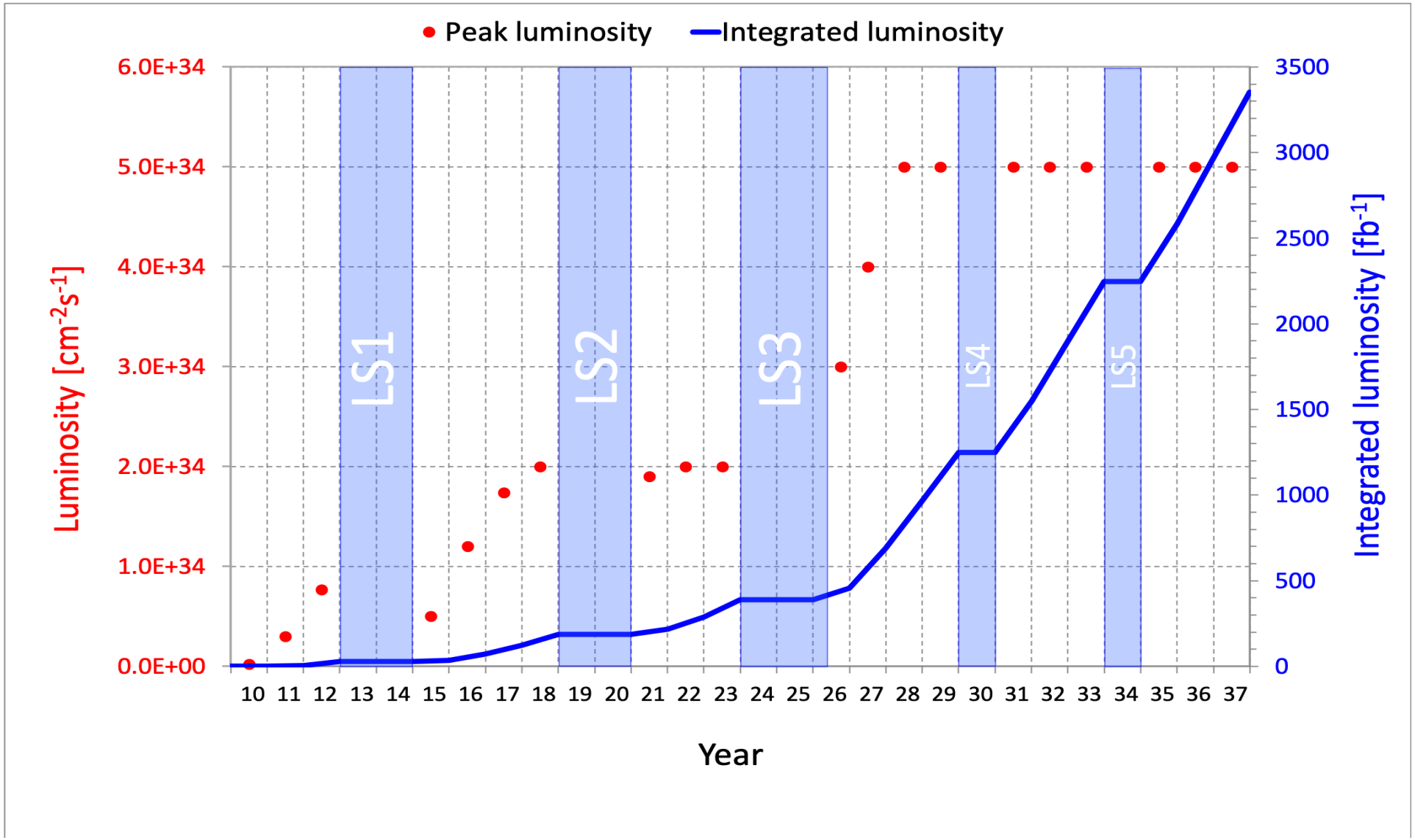


The Large Hadron Collider: A big data experiment



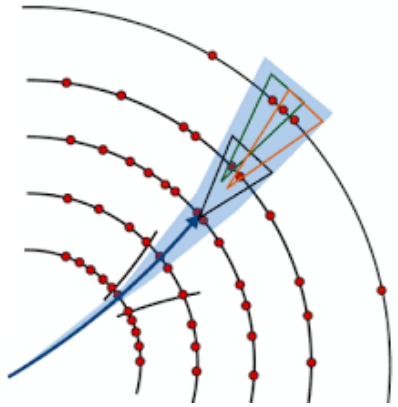
This plot shows the amount of data recorded on tape generated by the LHC experiments, the other experiments, various back-ups and users. In 2018, over 115 petabytes of data in total (including about 88 petabytes of LHC data) were recorded on tape, with a record peak of 15.8 petabytes in November.

Expectations for the HL-LHC

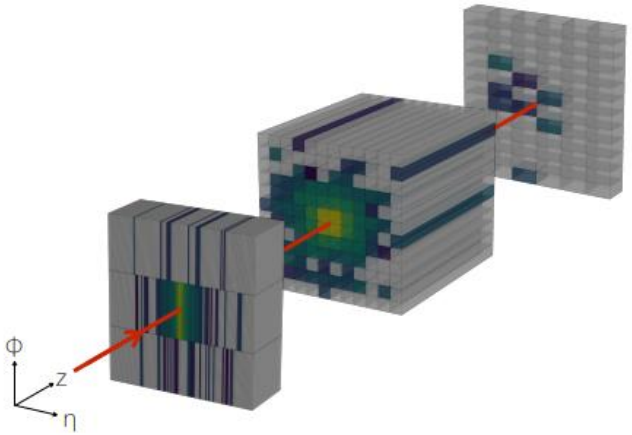


Where can DL be applied in HEP?

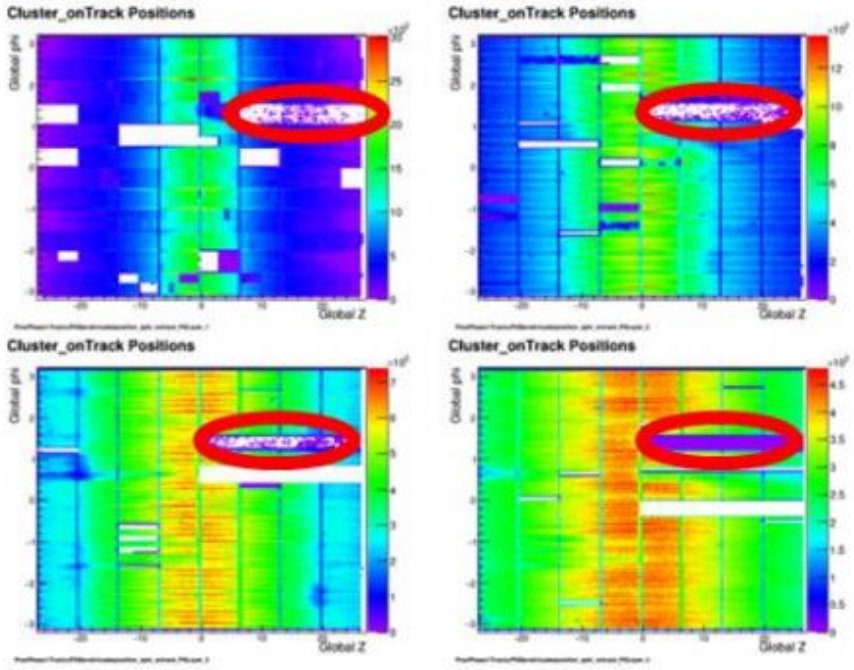
Particle reconstruction



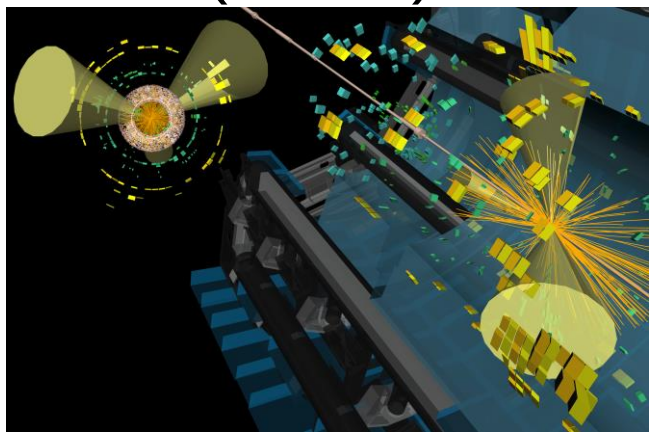
Simulation



Anomaly detection (Data quality monitoring)



Signal/bkg separation (offline)



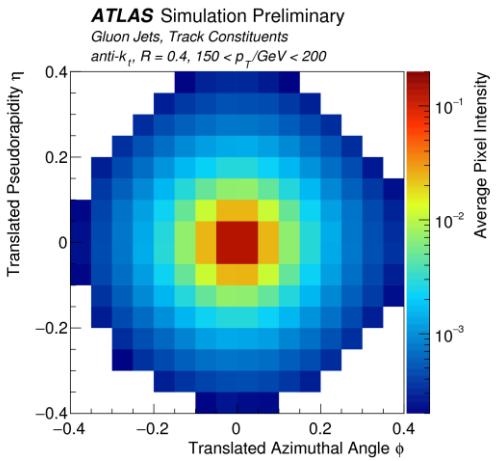
Online triggering

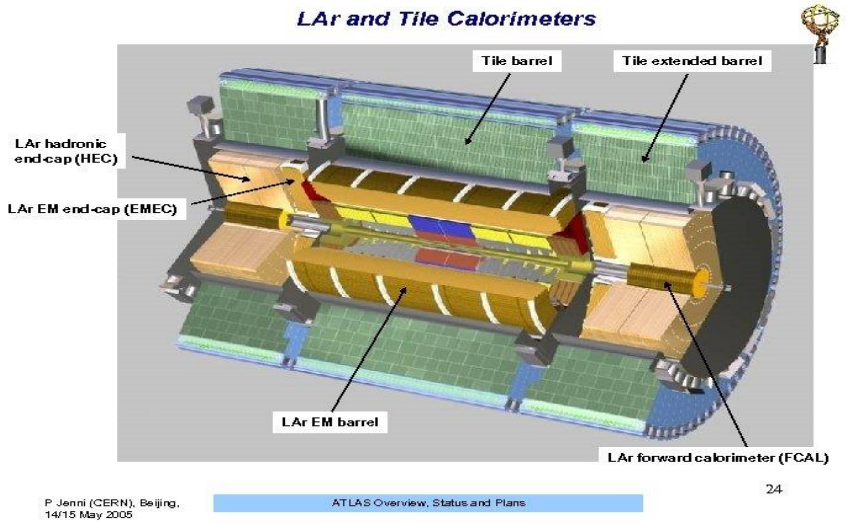
Where is the data coming from?

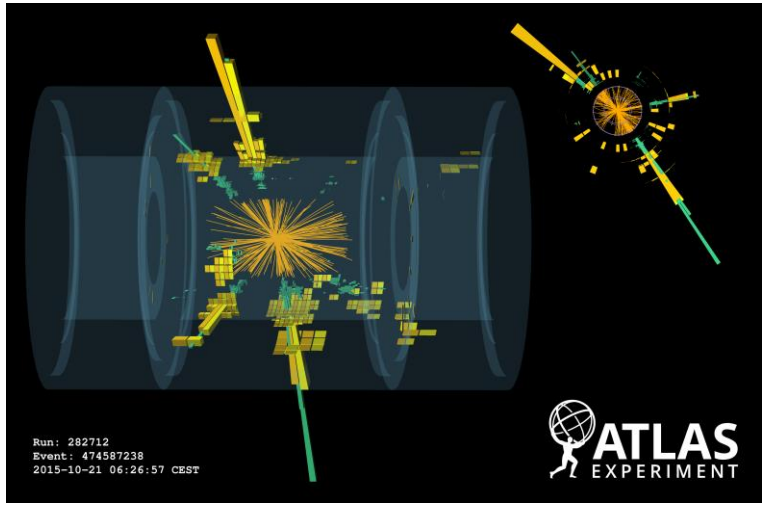
- 80 million electronic channels
- 10 petabytes/s of information
- ~1000 Mb/s raw data to tape
- 50 Pb of data per year writing to tape

Jet identification using CNN

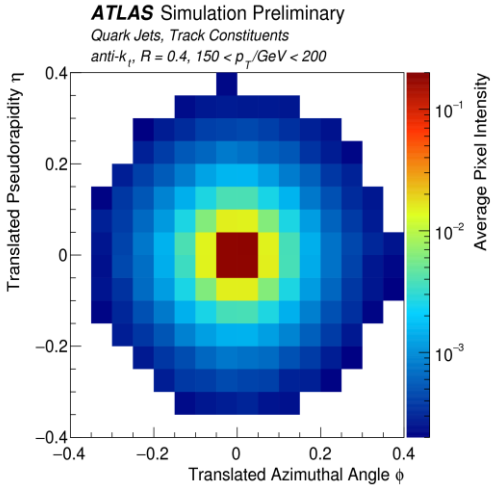
Gluon initiated jet

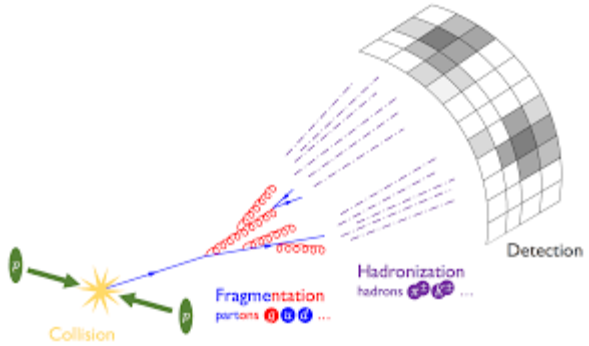


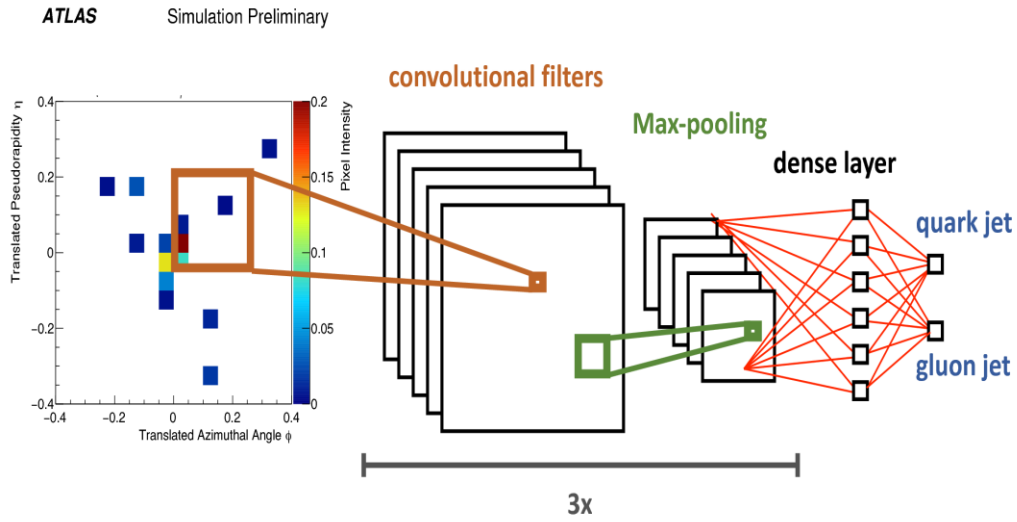




Quark initiated jet



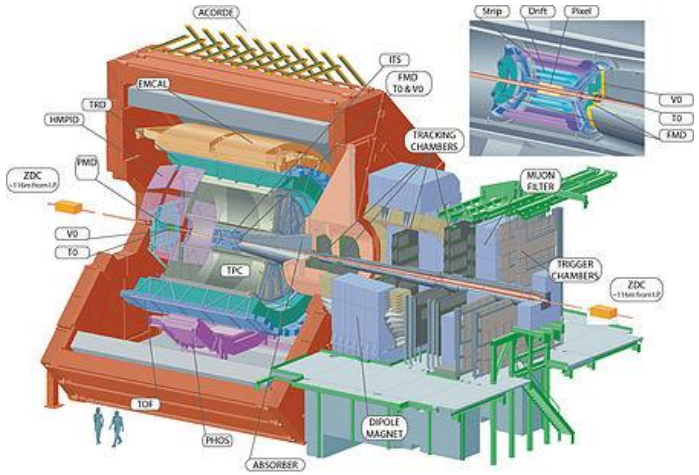


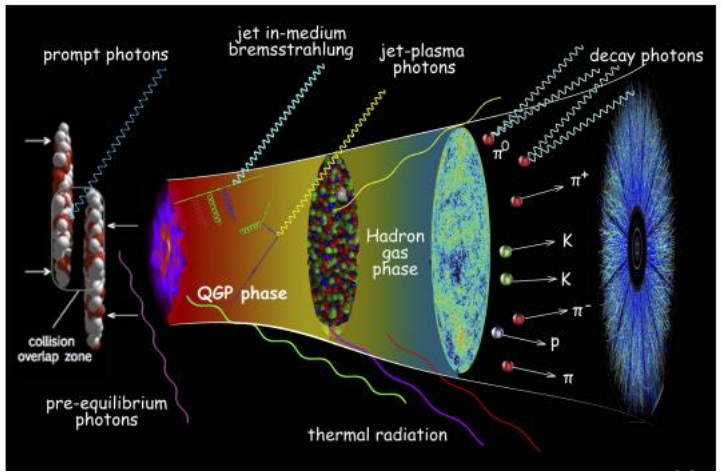


ATL-PHYS-PUB-2017-017

Charmed Baryon production in ALICE

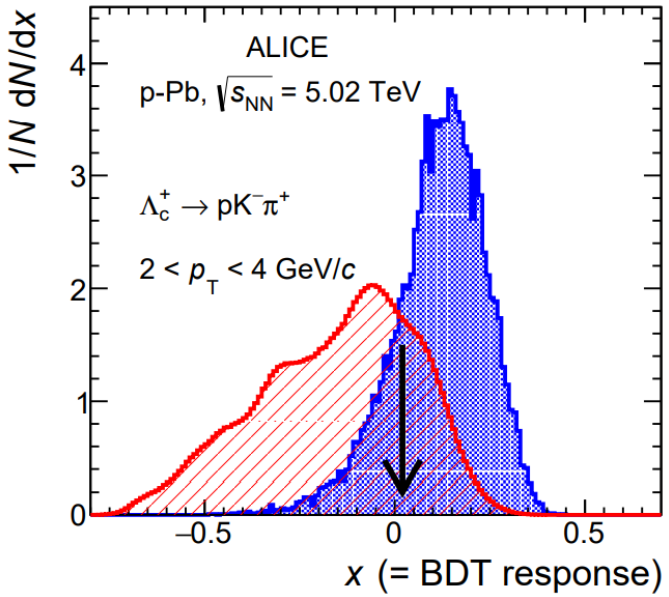
- ALICE is the dedicated heavy-ion experiment at the LHC
- Study of the Quark Gluon Plasma (QGP)

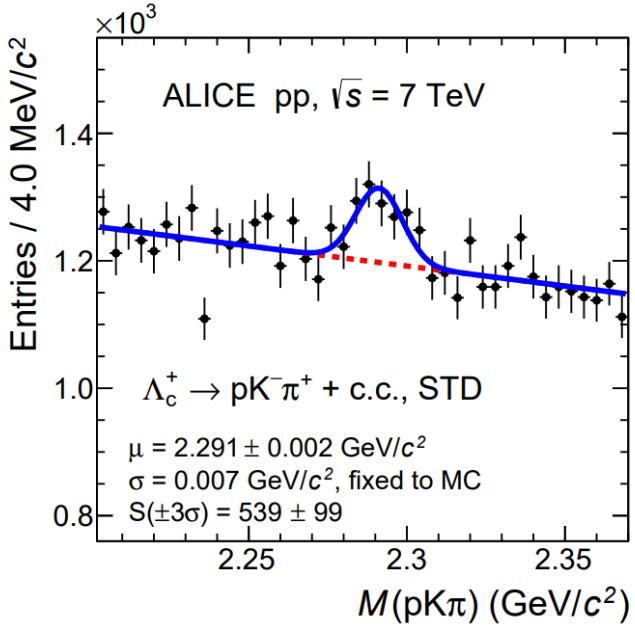




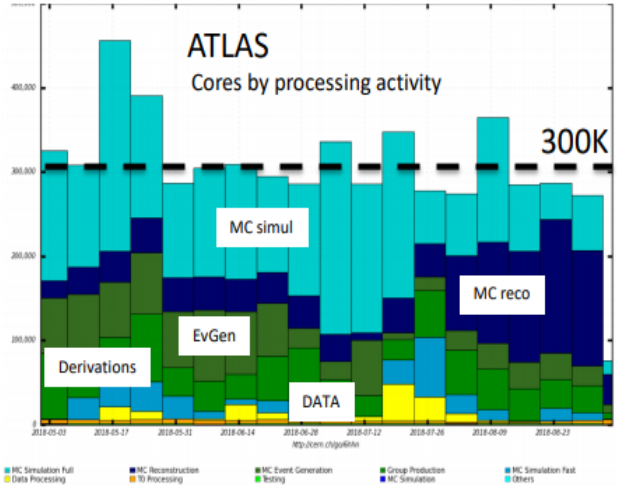
- Due to the short lifetime of the Λ_c baryons and statistical limitation the reconstruction was particularly challenging
- BDT were used to separate background from signal

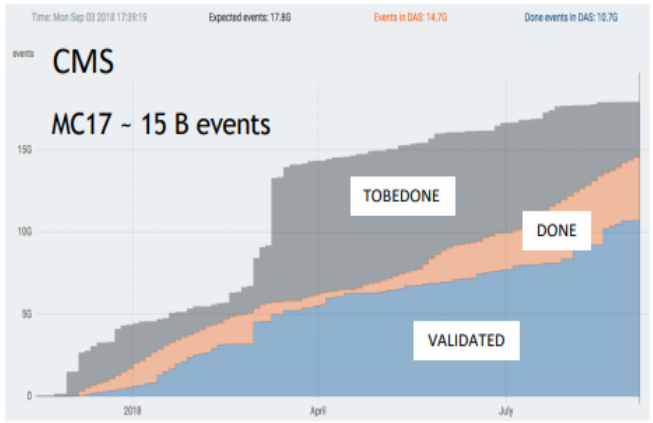
JHEP 1804 (2018) 108





Simulations in HEP



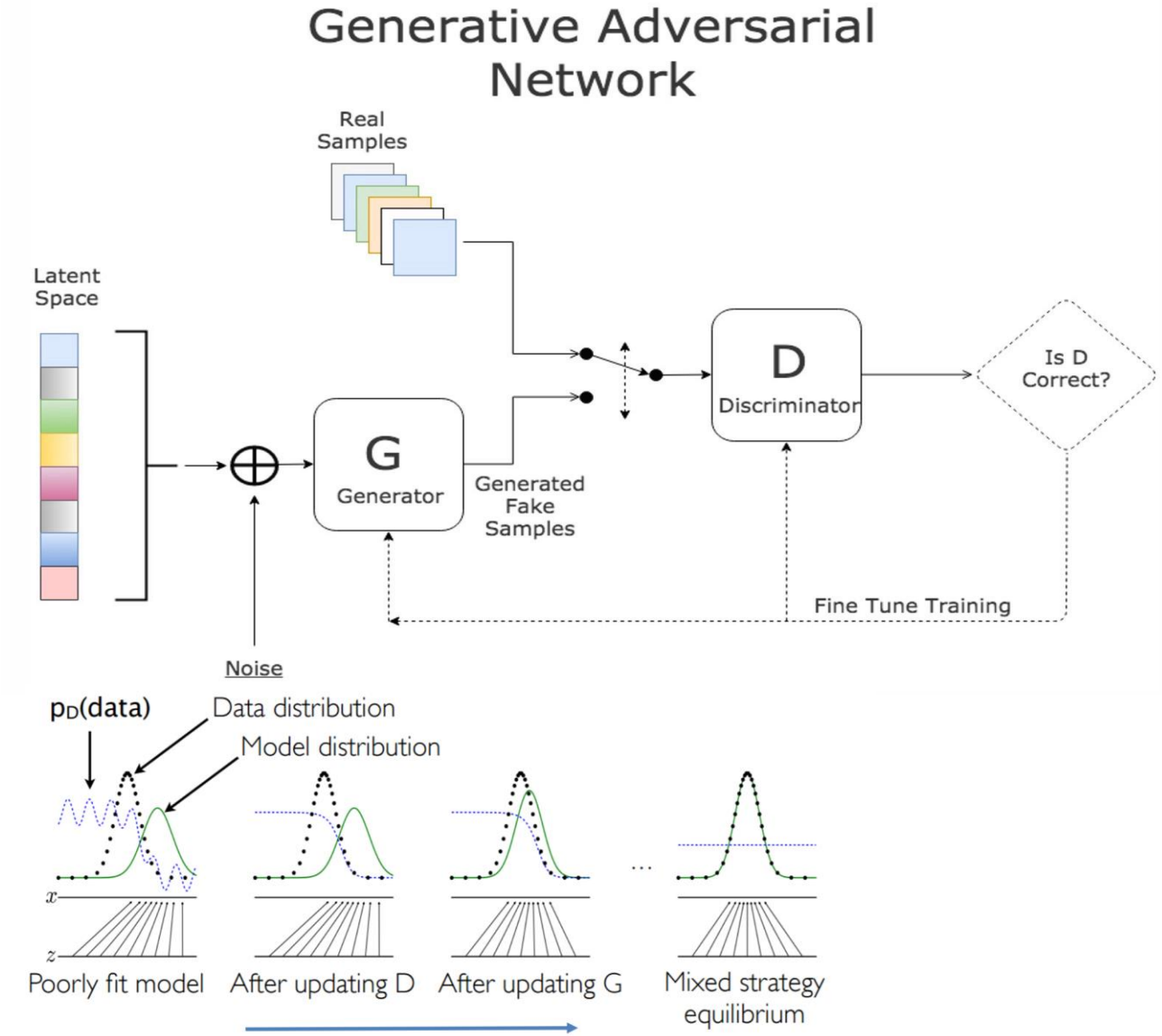


The majority of CPU cycles is spent in Monte Carlo production

For the past few years Monte Carlo simulations have represented more than 50% of the WLCG (LHC computing grid) workload

- Production of Monte Carlo simulation samples is a crucial task in HEP experiments
- Detailed simulation is needed to optimize the events selection for SM searches and new physics searches
- To measure the performance of new detector technologies
- Usually, samples with large statistic are needed in order to reduce the analysis uncertainties
- A lot of resources (computational, human) are needed to fulfill the requirements of the entire experimental collaborations

GANs for fast simulation



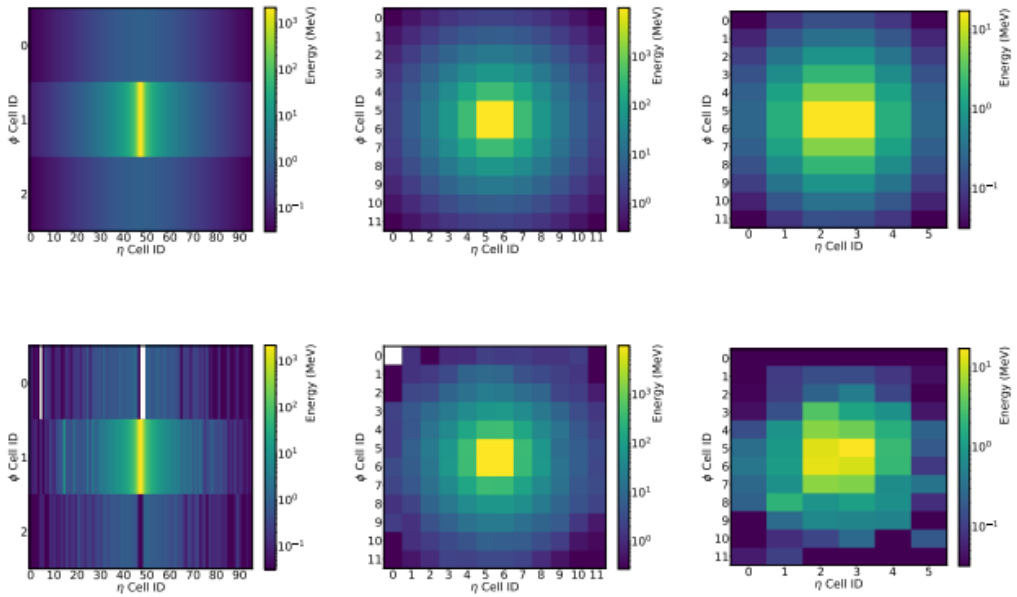
- Generative Adversarial Networks (GANs) are a type of deep neural networks with an architecture comprised of two nets, pitting one against the other ("adversarial")
- One neural network called "generator" generate new data instances, while the other "discriminator" evaluates the authenticity comparing with the training data
- The goal is that the generator is trained in a way that the discriminator will not identify if the information received is not the truth one

Introduced by Ian Goodfellow in 2014, Referring as the "most interesting idea in the last 10 years of ML"

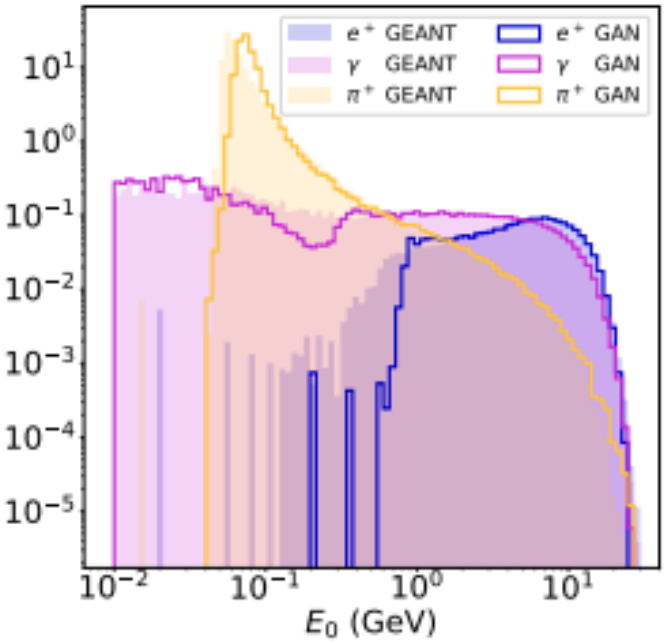
CaloGAN: a calorimetry fast simulation

- A simulation of the calorimeter using GANs (CaloGAN)
- From a series of simulated showers, the CaloGAN is tasked with learning from the simulated data distribution for gammas, e+ and pi+
- The training dataset is represented in image format by three figures of dimensions 3x96, 12x12 and 12x6, each representing the shower energy depositions

Average gamma Geant4 shower from Geant4 (top) and CaloGAN (bottom)



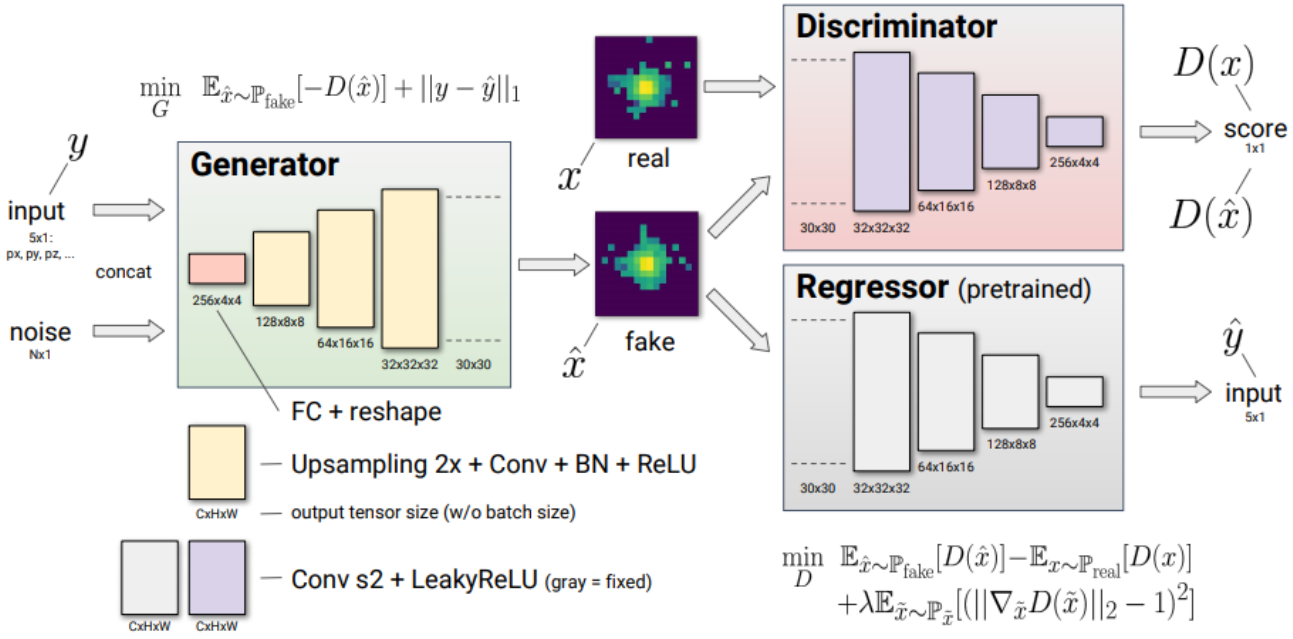
Shower shape variables for e+, gamma and pi+, comparison Geant4 and CaloGAN



Phys. Rev. D 97, 014021 (2018)

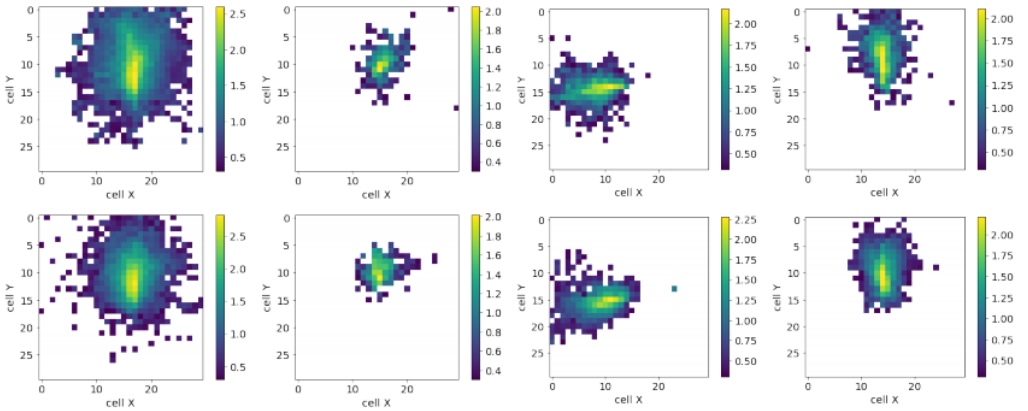
Calorimeter FastSim in LHCb

- New Deep Learning framework based on GANs
- Faster than traditional simulated methods by 5 orders of magnitude
- Reasonable accuracy
- This approach could allow physicist to produce enough simulated data needed during the HL-LHC

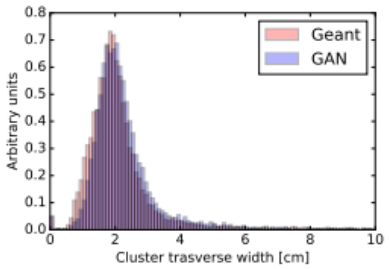


https://arxiv.org/abs/1812.01319

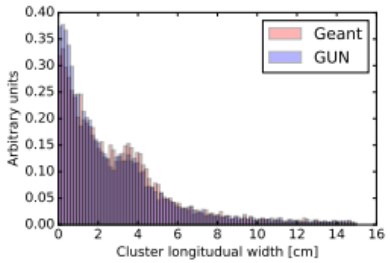
Showers generated with Geant4 (top) and showers simulated with GANs (bottom)



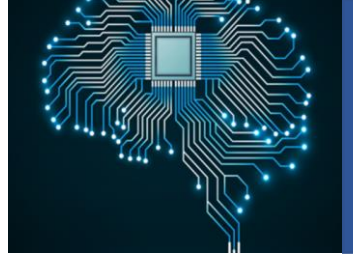
(a) E_0 = 63.7 GeV (b) E_0 = 6.5 GeV (c) E_0 = 15.6 GeV (d) E_0 = 15.9 GeV



(a) The transverse width of real and generated clusters



(b) The longitudinal width of real and generated clusters



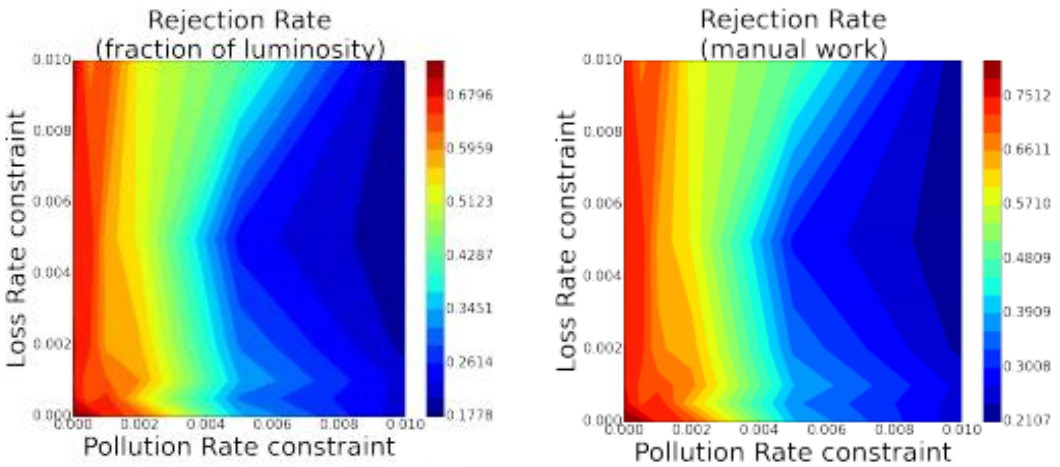
Data Quality monitoring

- Data quality monitoring is a crucial task for every large-scale High-Energy Physics experiment
- The system still relies on the evaluation of experts
- An automated system can save resources and improve the quality of the collected data
- A detector measures physical properties of proton collisions products
- When a subdetector exposes abnormal behavior, it is reflected in measured or reconstructed properties

Data Quality monitoring

- The primary goal of the system is to assist the Data Quality managers by filtering most obvious cases, both positive and negative
- The system objective is minimization of the fraction of data samples passed for the human evaluation (i.e. the rejection rate)
- The Evaluation of the algorithm was done using CMS experimental data published in the **CERN open data** portal

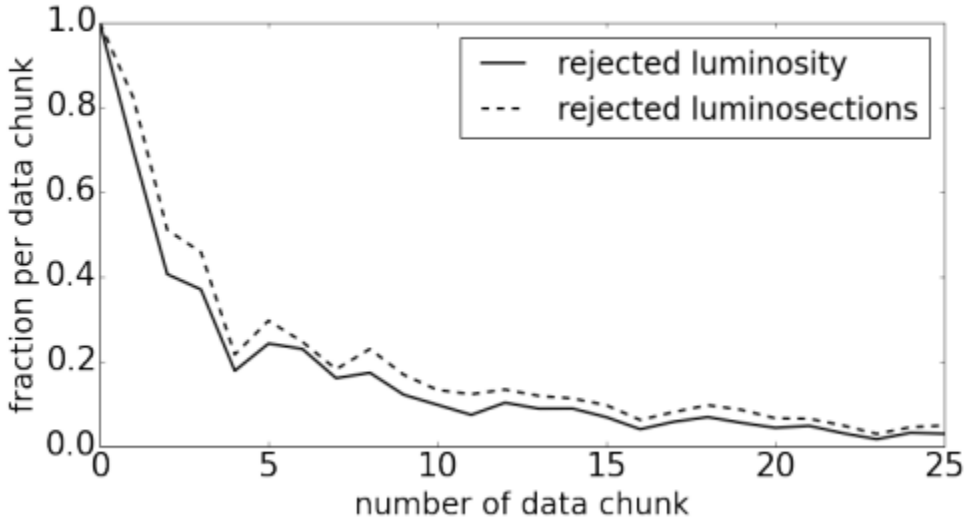
$$\text{Rejection Rate} = \frac{\text{Rejected}}{\text{Total}} \rightarrow \min,$$



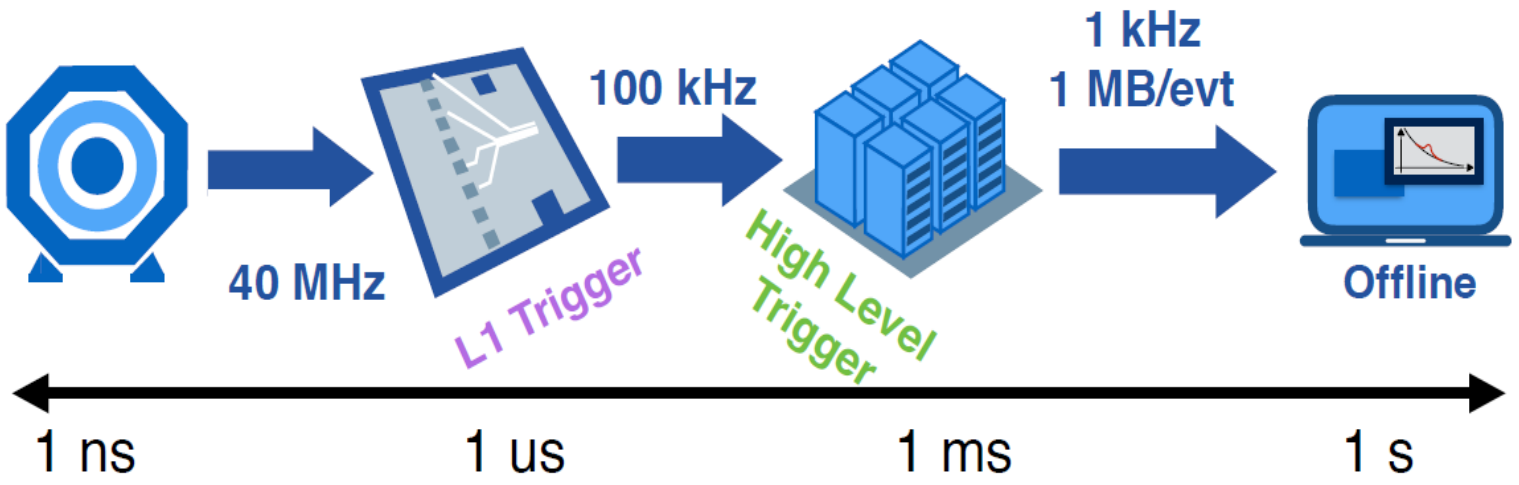
(b) Fraction of rejected luminosity.

(a) Fraction of rejected samples.

doi :10.1088/1742-6596/898/9/092041

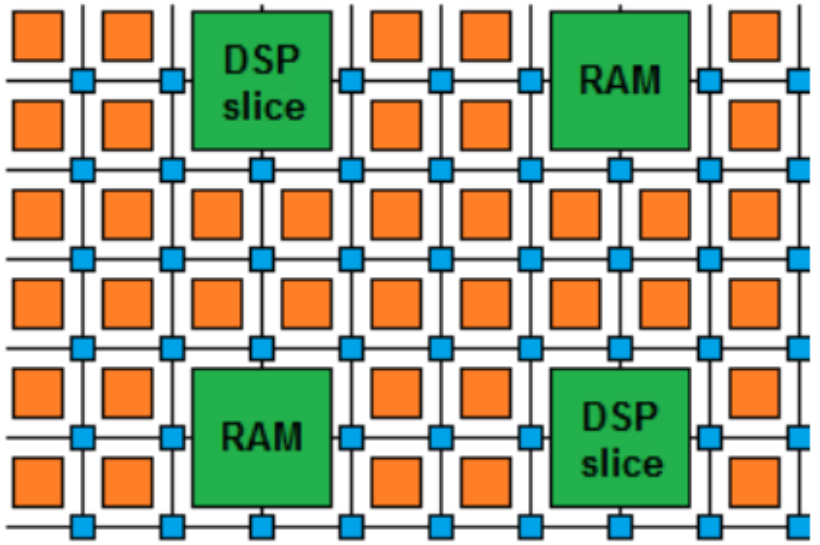


Data processing with FPGAs



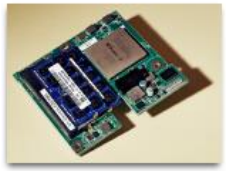
- L1 trigger decision (hardware, FPGA based) to be done in 4 microsec
- Could a ML model be fast enough to make a decision in such short time window?

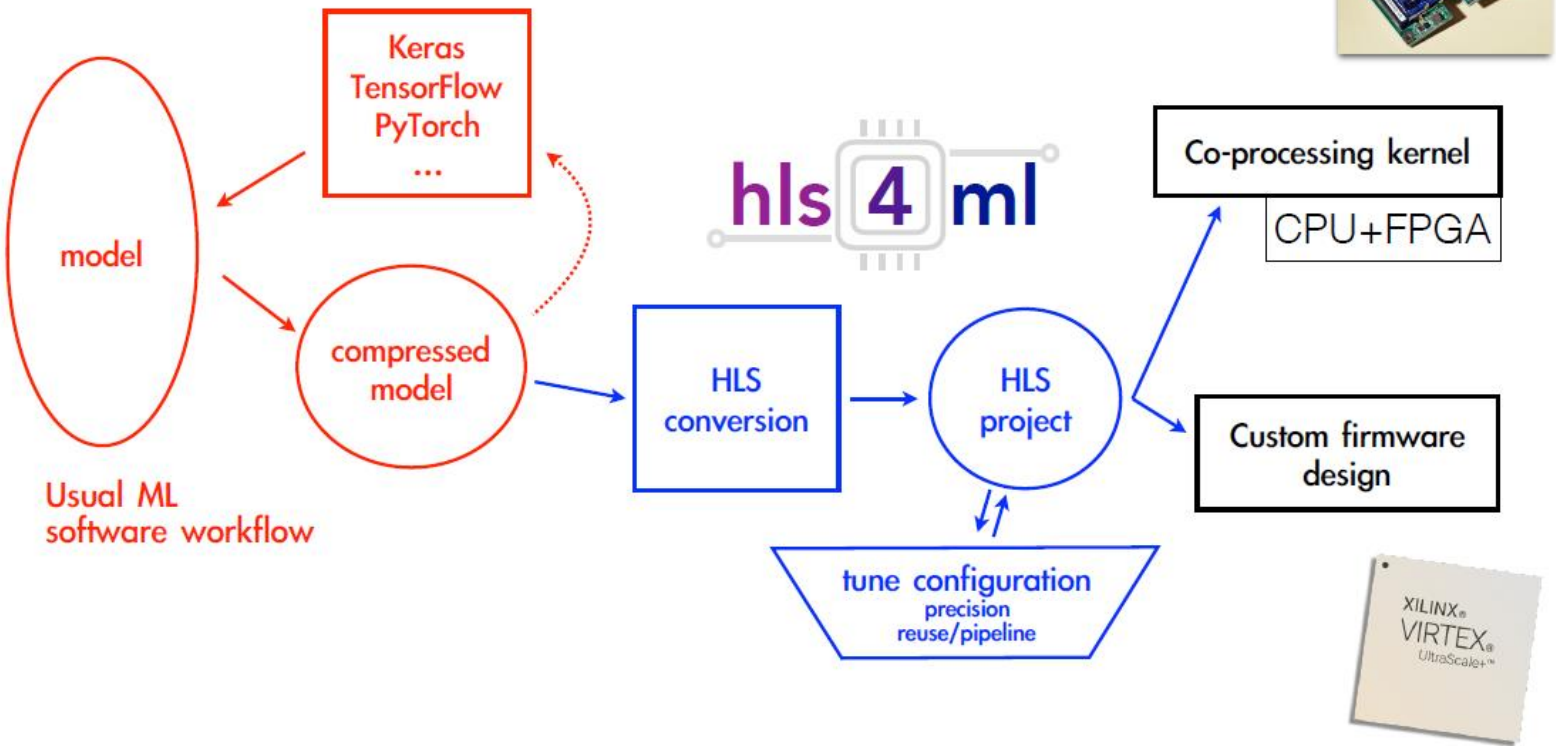
Field Programmable Gate Arrays (FPGAs)



- Reprogrammable fabric of logic cells embedded with DSPs, BRAMs, high speed, IO
- Low power consumption compared to CPU/GPU
- Massively parallel

ML interface for FPGAs





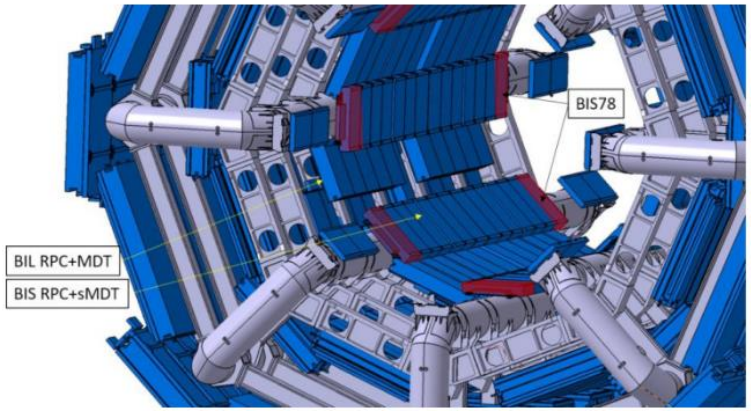
- Regular neural network model trained in CPU/GPU
- Model loaded in firmware via the hls4ml interface
- Decision made in the FPGA

https://arxiv.org/abs/1804.06913

FPGAs for Phase-2 ATLAS Muon barrel

- ATLAS Level-0 muon trigger will face a complete upgrade for HL-LHC
- New trigger processor
  - FPGA based system
- New trigger station
  - New RPC Layer
- Trigger algorithms improved
  - Need to be very fast and flexible

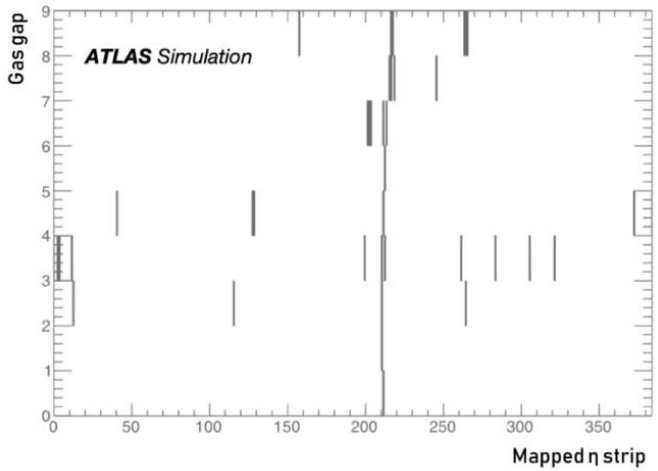
Neural Network a good candidate

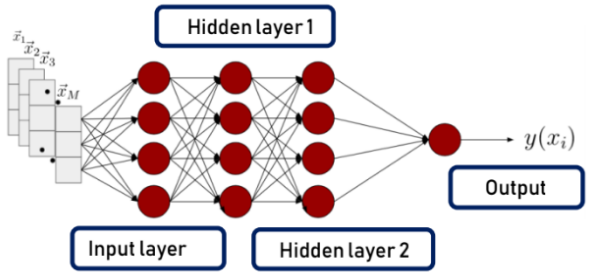


ATLAS-TDR-026

CNNs

From strip maps to images



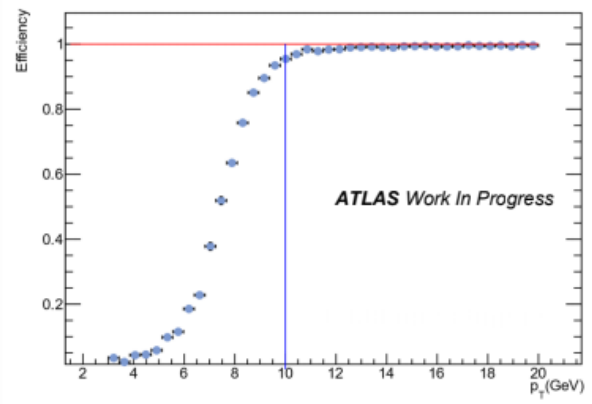


The CNN is trained to give a 5D output:

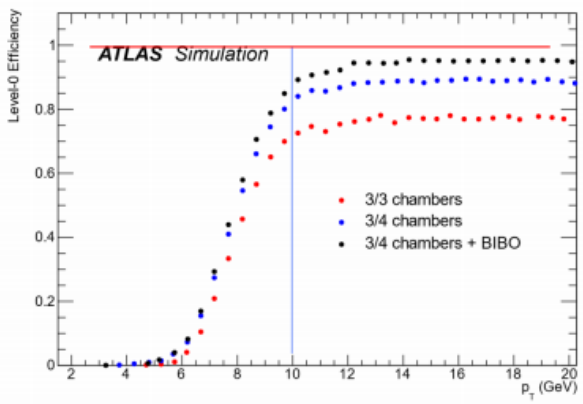
0 → 0/1/2 muons
-1 → more than 2 muons

$(p_T^{lead},\ \eta^{lead},\ p_T^{sublead},\ \eta^{sublead},\ N^{muons})$

CNN

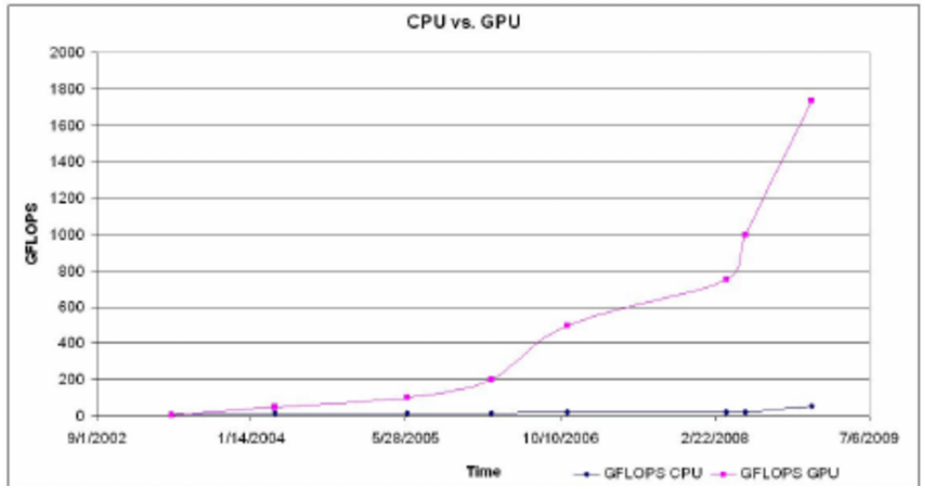


Standard Algorithm

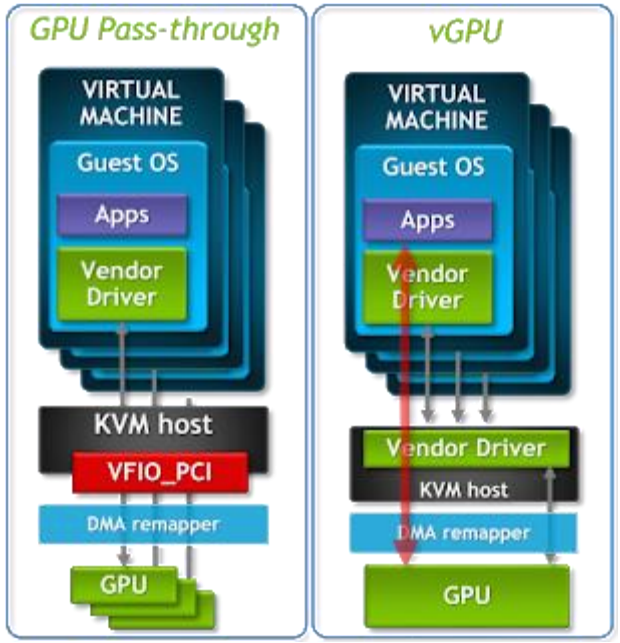


GPU resources at CERN

- Graphical Processing Units (GPUs) are fundamental for the training of models in machine learning
- With GPUs the training time can be reduced several orders of magnitude with respect to normal CPU
- HEP workloads can be benefit from from massive parallelism
- Current challenges as the TrackML motivates CERN to consider GPU provisioning in the OpenStack cloud

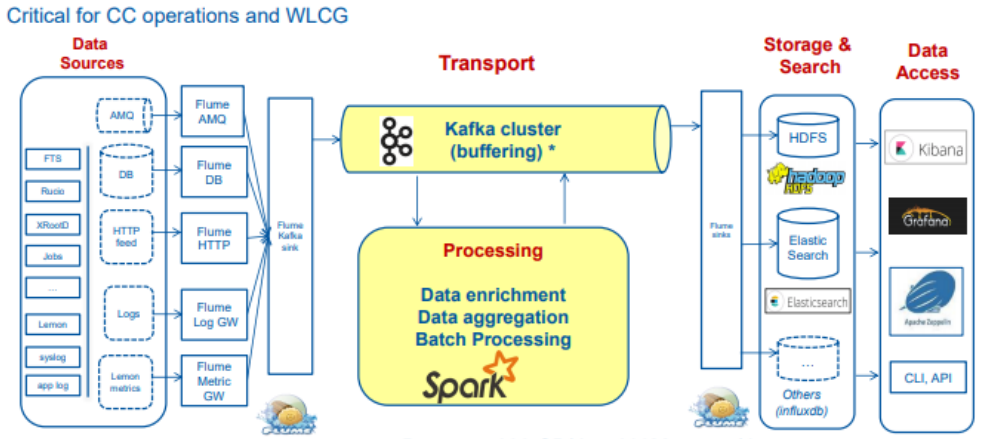


- GPUs can be accessible via virtual machines
- The user can access normal CUDA applications as tensorflow
- The resources are so far quite limited but is intended to grow in the coming years

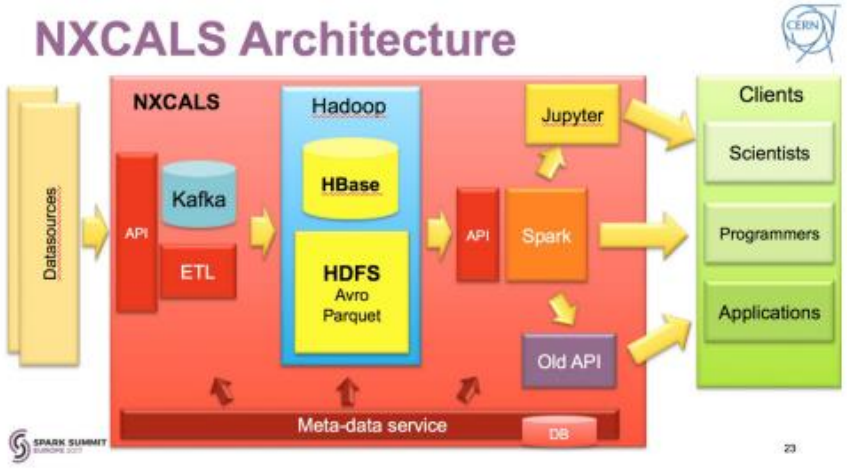


Projects in which computer scientists can be involved

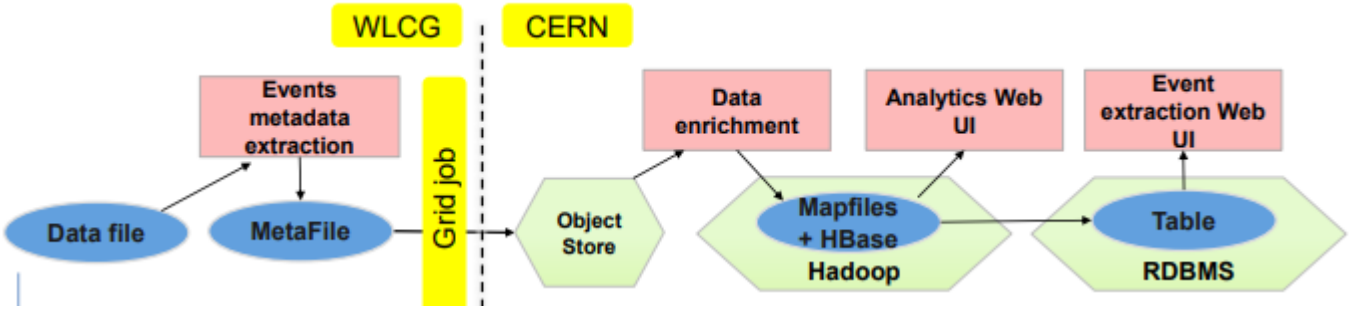
New CERN IT monitoring infrastructure



Next Generation archiver for accelerators logs



The ATLAS event index

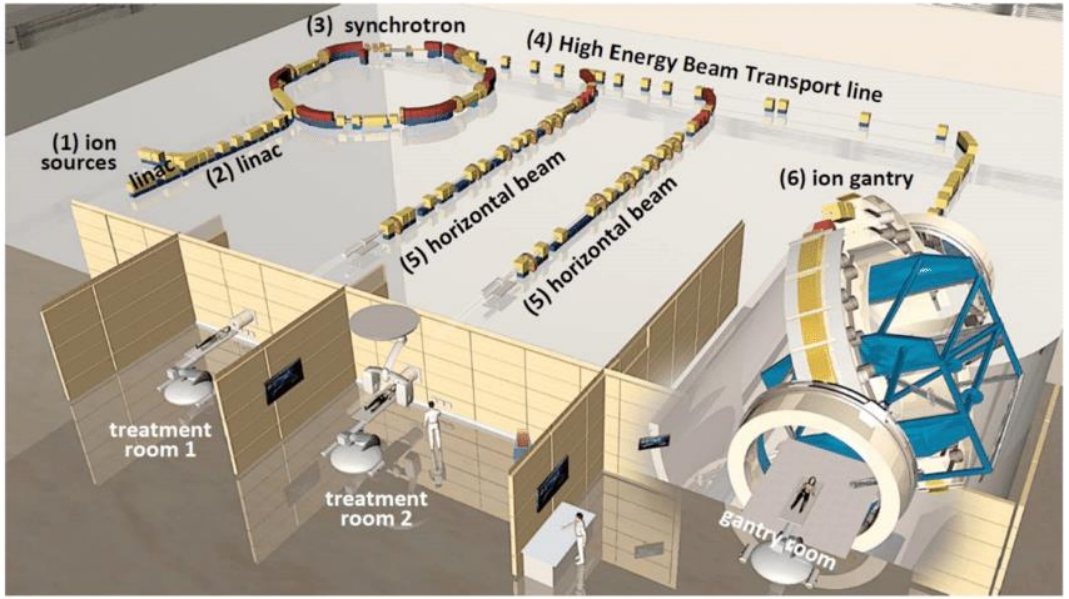


Swan services

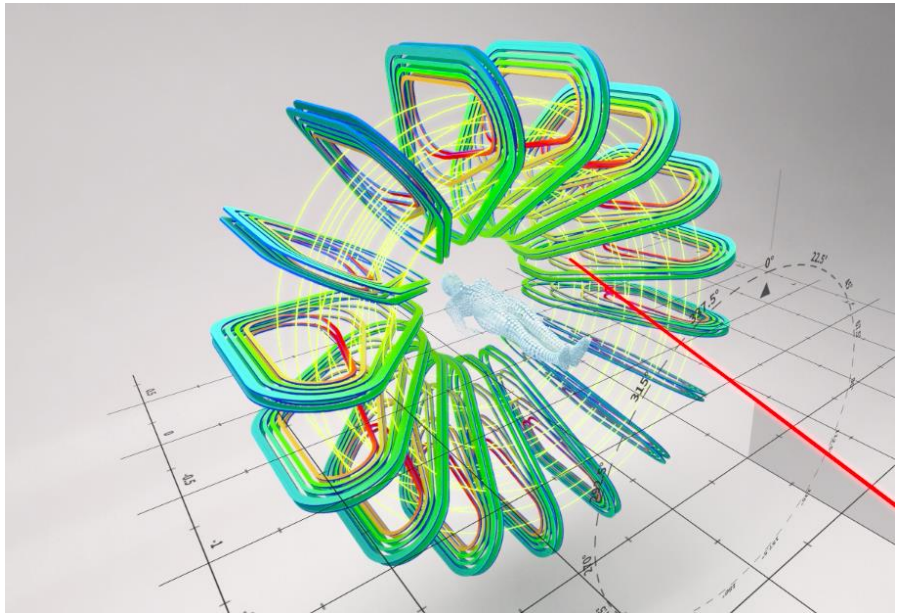
- Fully integrated with Sparks and Hadoop at CERN
- Modern, powerful and scalable platform for data analysis

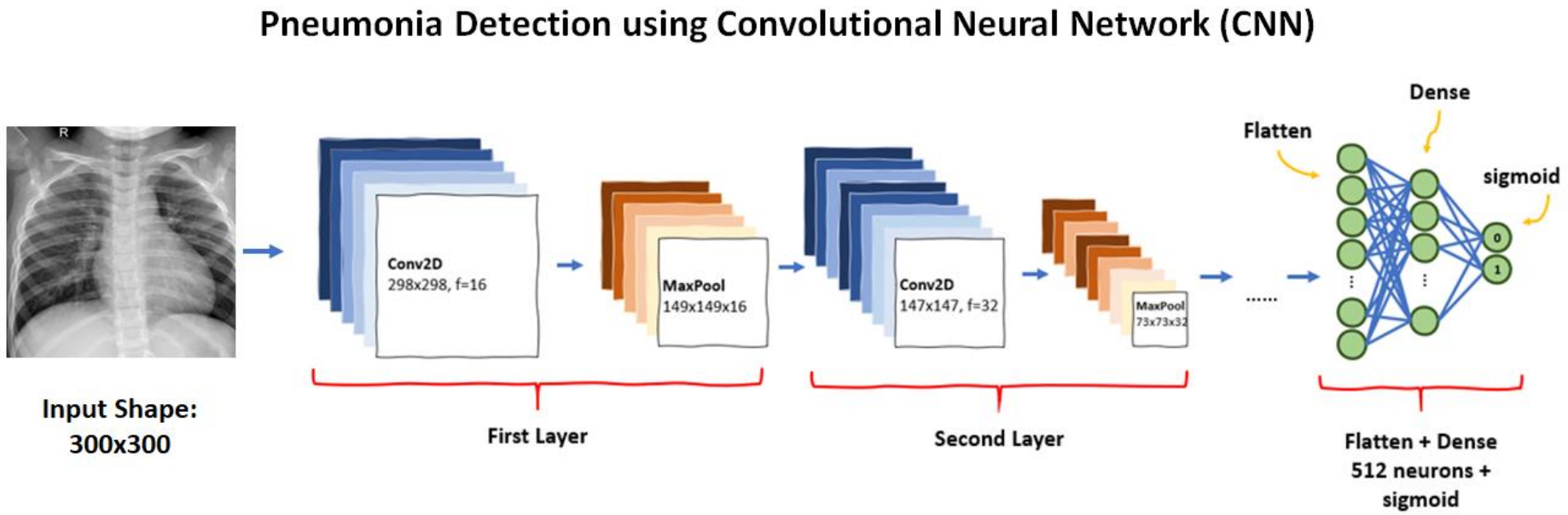
Medical physics applications connected with DL and HEP

Development of novel acceleration technologies for cancer treatment installation, with DL implementations similar to the LHC



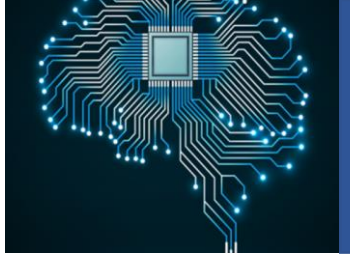
Development of simulation for testing new prototypes in medical physics (toroida magnet to focus the beam during radiation treatment)





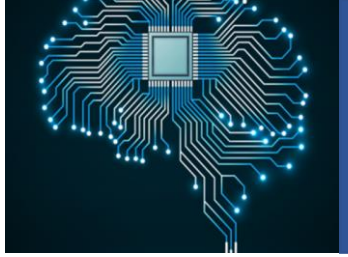
Anomaly detection using Convolutional neural networks

- One of the most successful ways to implement DL in medical physics is with image analysis
- Commonly medical images are (visually) analyzed by medical specialist
- This usually comes with a probability of a wrong diagnostic and late treatment
- Introducing a reliable model in DL has the potential to produce an early diagnostic and start the treatment



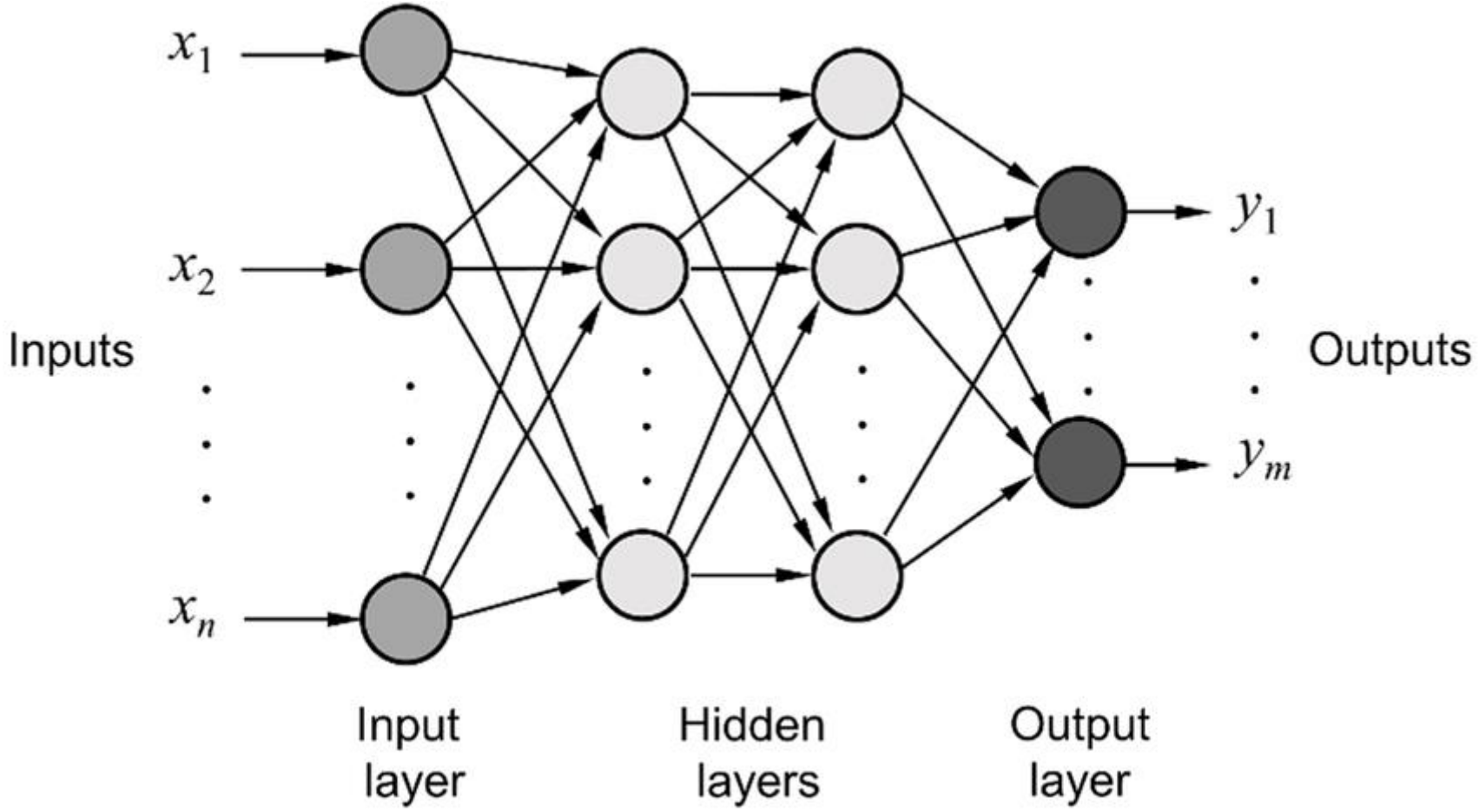
Conclusions and Perspectives

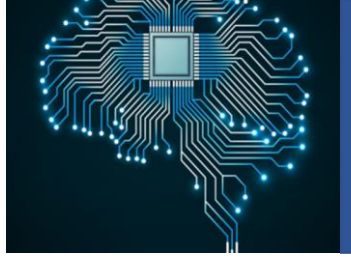
- ML and Bigdata development are (and will be) relevant for the development in different areas including HEP and medical physics
- There need to be a better communication and related projects between physicist, data scientists, computer and electronics engineers to develop multidisciplinary projects to advance not only fundamental science but also applied science with social impact
- Currently we have the pieces to start this projects
    - Academic programs (physics, medical physics, data science, engineering)
    - Installations (labs, computing)
    - Big projects to power such developments (CERN experiments)



Backup

Multilayer neural network





CERN openlab initiative

- Most of the new developments on big data and ML tools are now carried by the CERN openlab initiative
- CERN openlab is a public-private partnership that works to accelerate the development of cutting-edge ICT solutions for the worldwide LHC community
- Some of the ongoing projects are:

- Oracle cloud
- REST services, Javascript, and JVM performance
- Quantum computing for high-energy physics
- High-throughput computing collaboration
- Code modernization: fast simulation
- Intel big-data analytics
- Oracle big-data analytics
- Yandex data popularity and anomaly detection

CERN openlab