Machine and Deep Learning Applications in ~~High Energy Physics~~ ATLAS and CMS

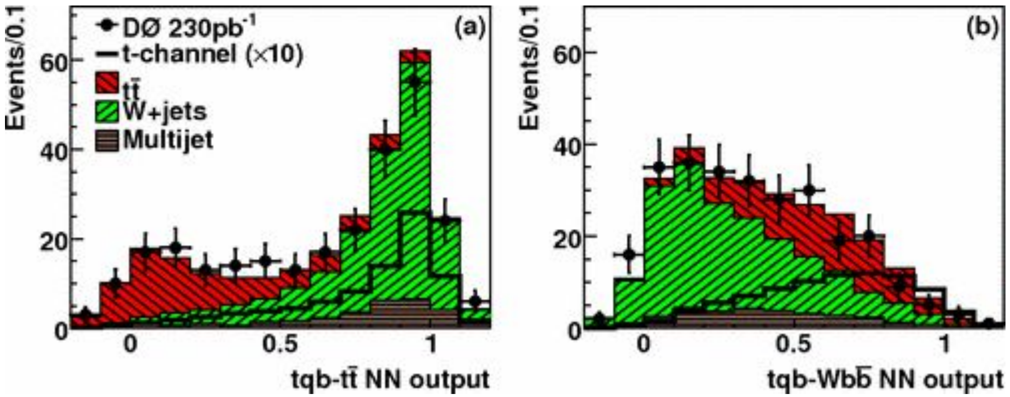
Terzo incontro di fisica con ioni pesanti alle alte energie (IFIPAE) 2021

Federica Legger

Once upon a time

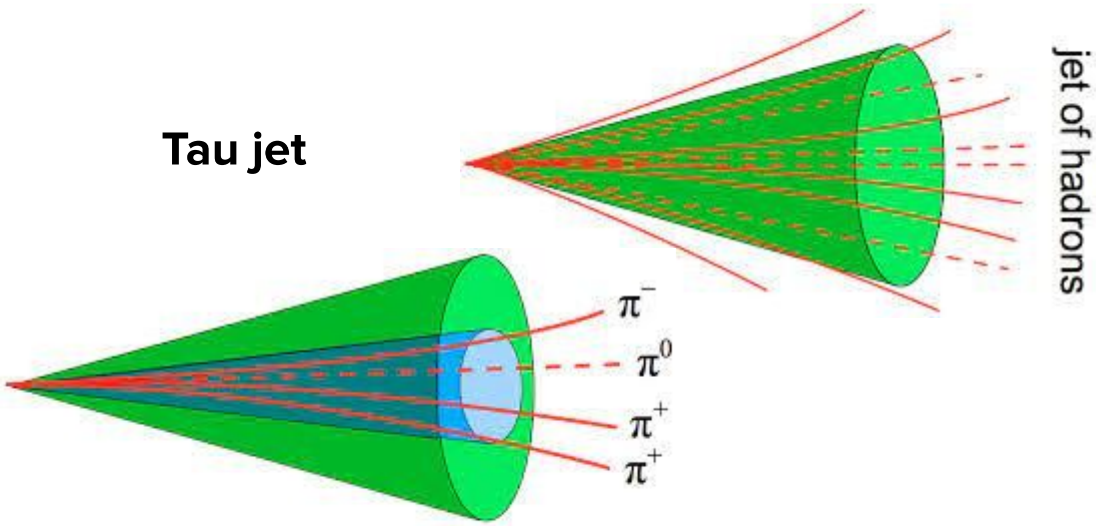
- TMVA - Toolkit for Multivariate Data Analysis, arXiv:physics/0703039
- TMVA fully integrated in ROOT in 2013
- Mainly for classification and regression tasks
  - boosted decision trees, support vector machines, cellular automata, multilayer perceptrons, …

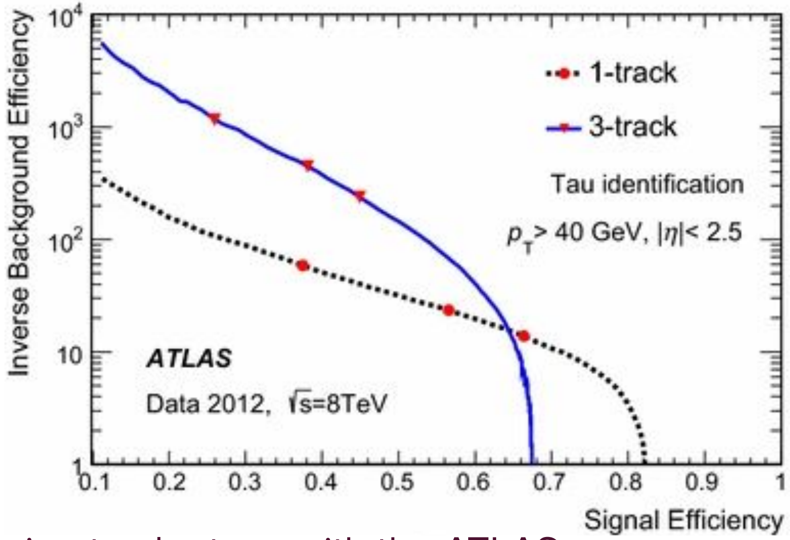
Multivariate searches for single top quark production with the DØ detector, Phys.Rev.D75:092007 (2007)



Particle Identification: Tau ID

- Hadronically decaying **tau leptons** vs jet of hadrons with **Boosted Decision Trees (BDT)**

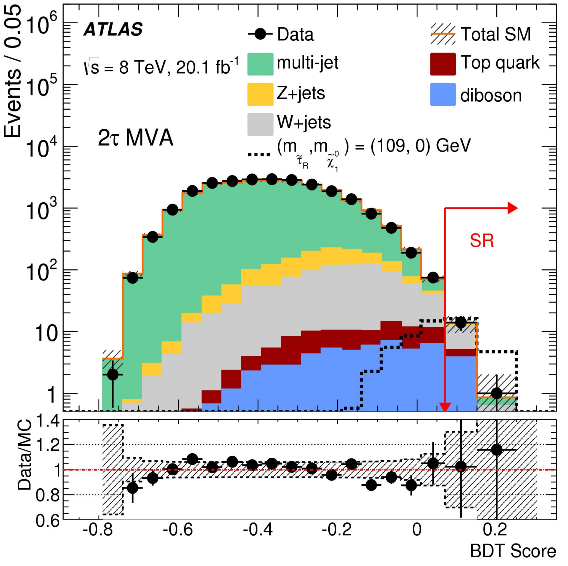


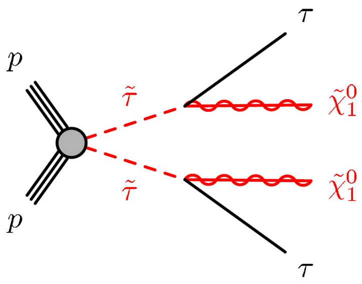


Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at $\sqrt{s}$=8 TeV, EPJC 75: 303 (2015)

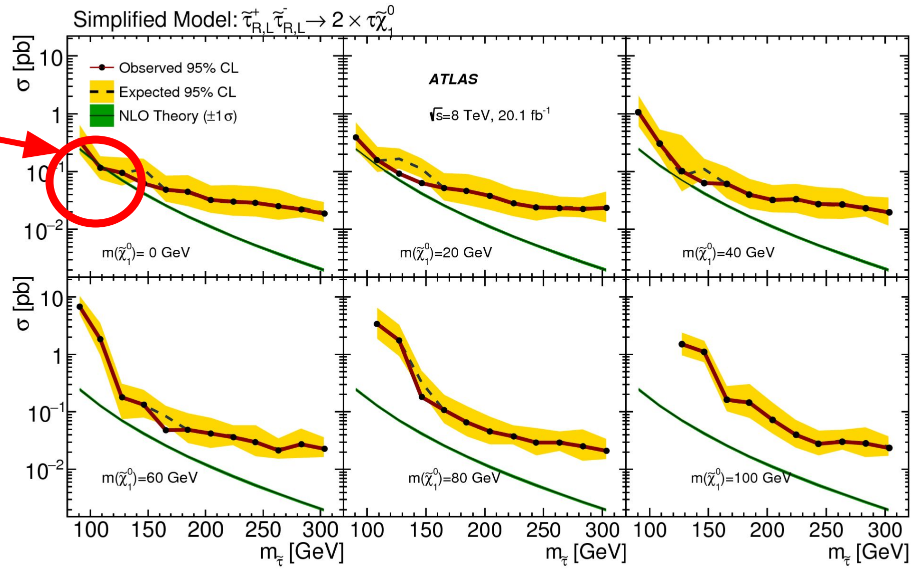
Rare Processes: Direct Stau Production

- **BDT** with low and high level variables





Stau mass 109 GeV

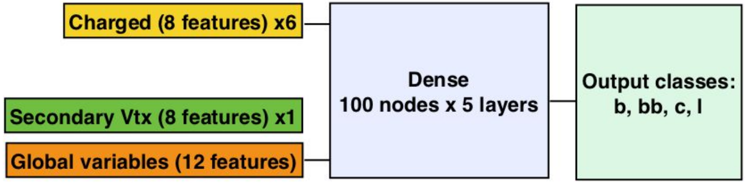


Search for the electroweak production of supersymmetric particles in √s=8 TeV pp collisions with the ATLAS detector, Phys. Rev. D 93, 052002 (2016)

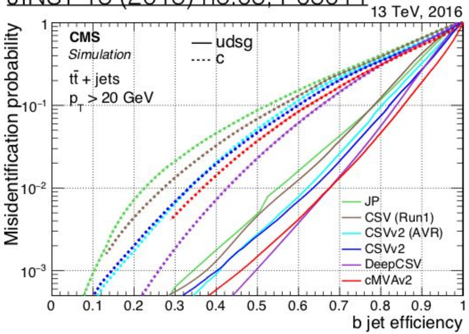
Going down the deep road - B tagging at CMS

DeepCSV:

- fully-connected layers
- Multi classification



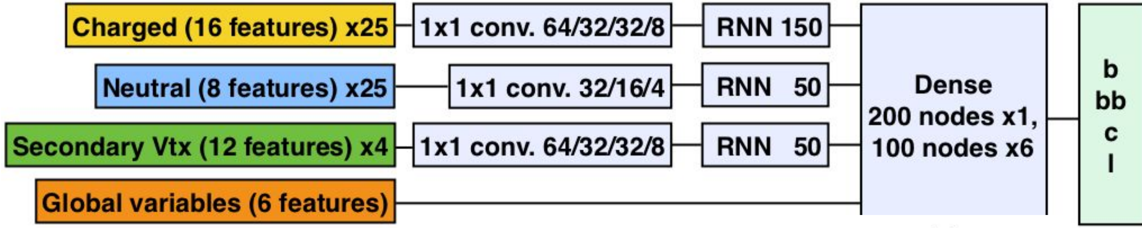
JINST 13 (2018) no.05, P05011

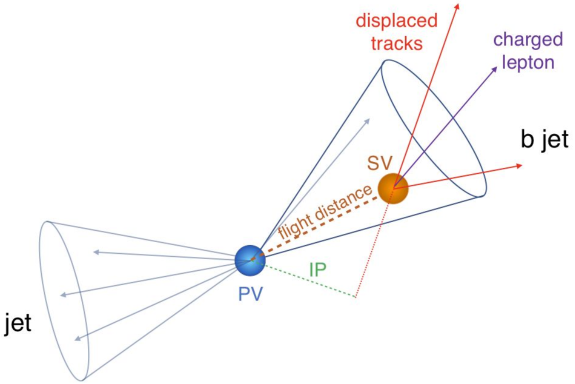


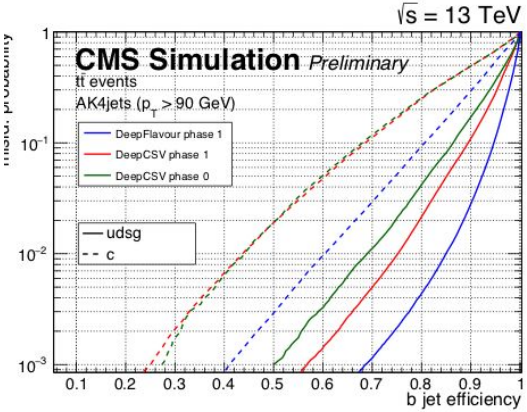
Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV, [JINST **13** P05011](#) (2018)

DeepJet:

- Convolutional layers learn compact feature representation (automatic feature engineering)
- RNN extract information from each set of features



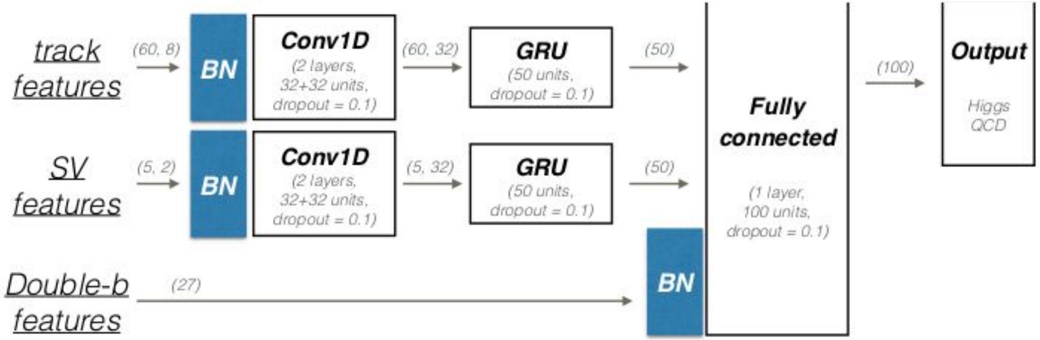


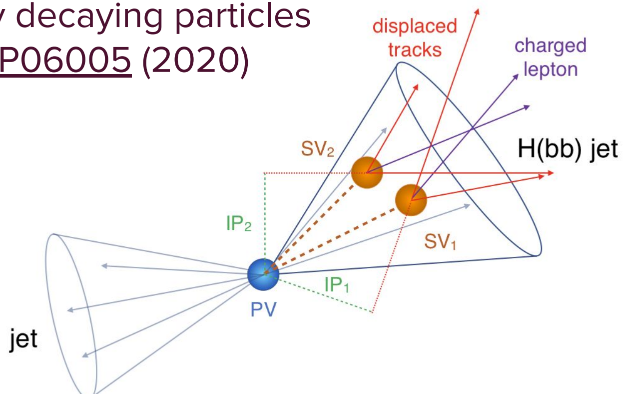


Higgs (Double B) Tagging

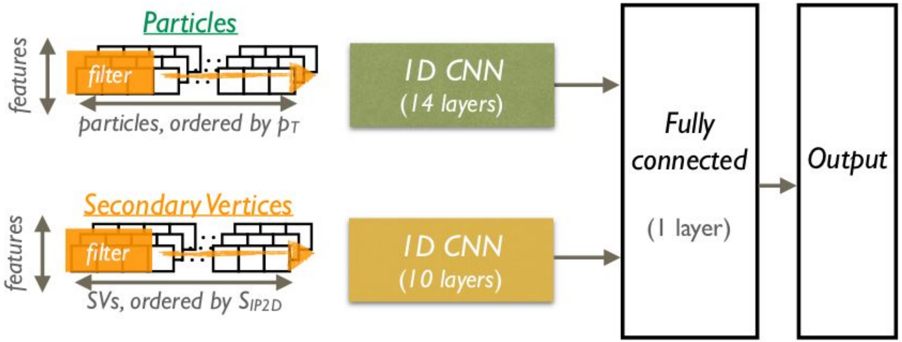
Deep Double B

Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques, JINST 15 P06005 (2020)

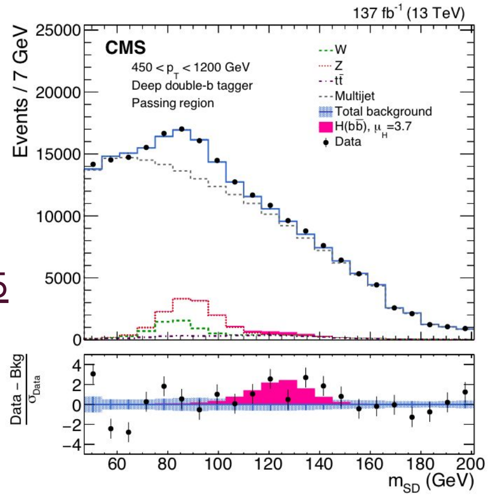




Deep AK8

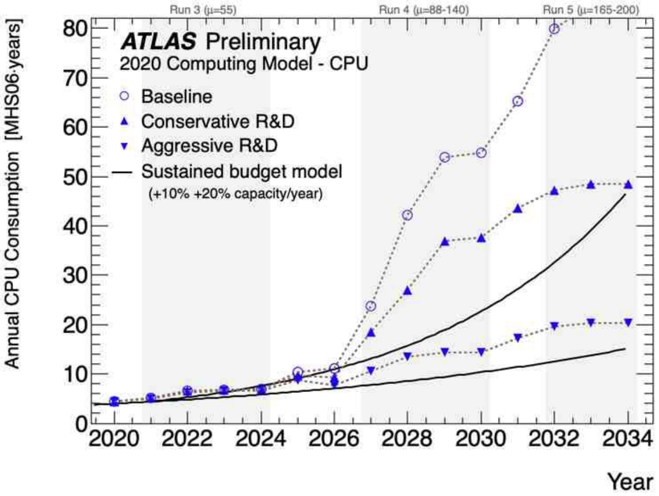


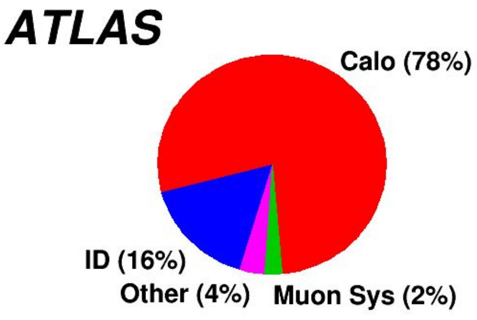
Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in pp collisions at $\sqrt{s}$= 13 TeV, JHEP 12 085 (2020)



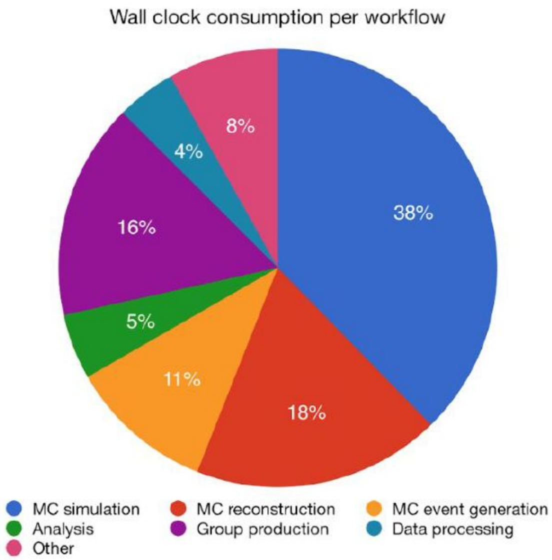
Deep Learning for simulation

- Computing demands increase nonlinearly with increasing pileup
- LHC Run 2: full detector simulation (Geant4) took ~40% of grid CPU resources for CMS & ATLAS
- Calorimeter simulation most CPU intensive

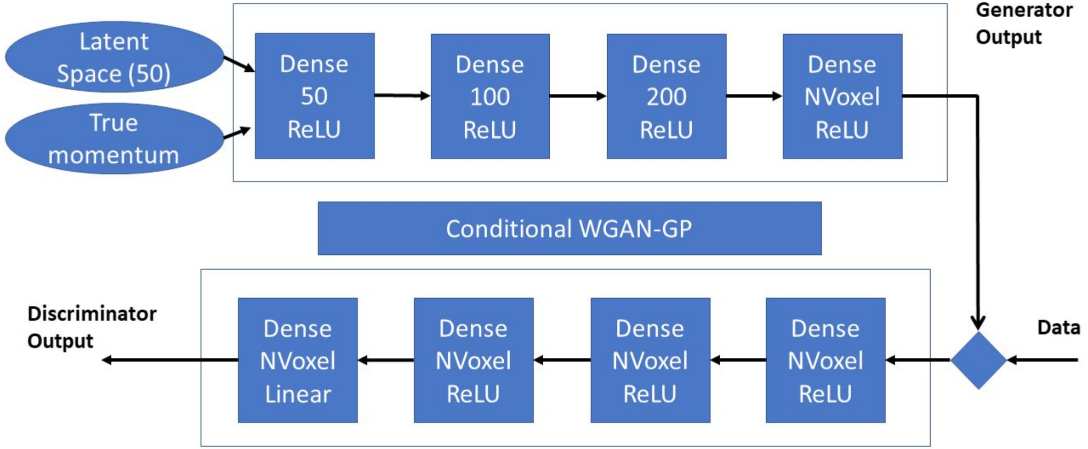


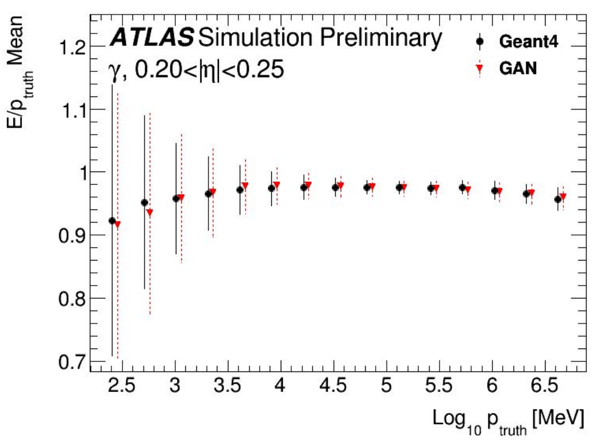


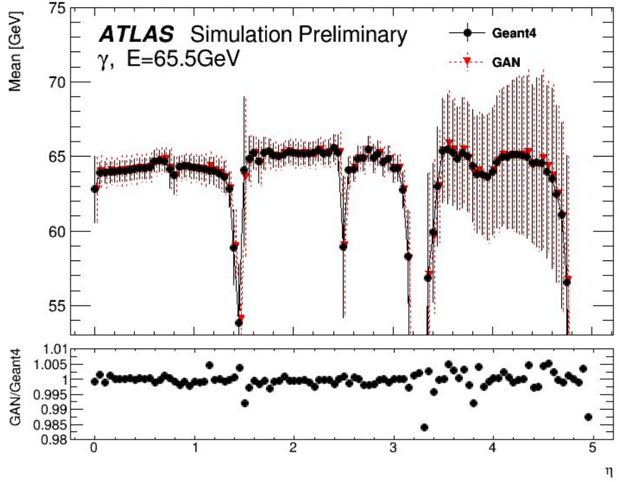
Subdetector CPU fraction for 50 ttbar events
MC16 Candidate Release



ATLAS FastCaloGan





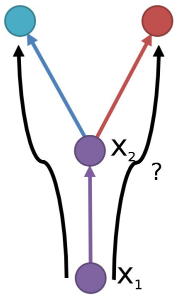


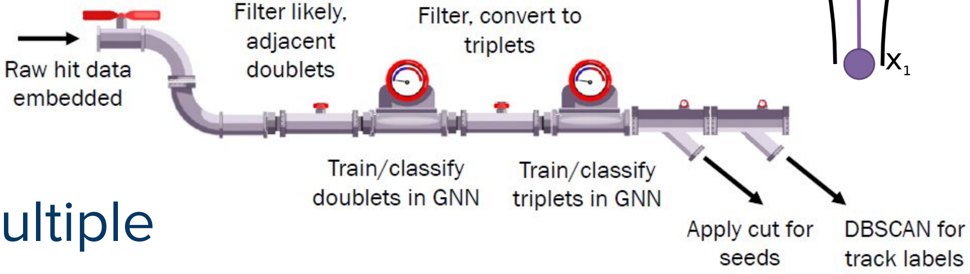
G	50 (Input latent Space), 50, 100, 200, NVoxel (pid and η dependent)
D	NVoxel, NVoxel, NVoxel, NVoxel, 1
Activation function	ReLU (in all layers)
Optimiser	Adam [21]
Learning Rate	10^{-4}
β	0.5
Batchsize	128
Training ratio (D/G)	5
Gradient penalty λ	10

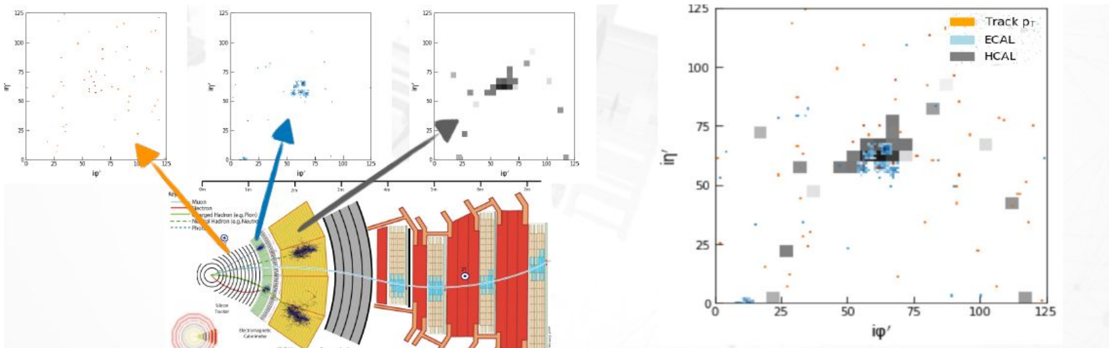
Fast simulation of the ATLAS calorimeter system with Generative Adversarial Networks, ATL-SOFT-PUB-2020-006 (2020)

New Trends in Reco

- Ideal applications for **graph neural networks:**
  - **Hit clouds** in Calorimeters: point cloud of energy deposits
  - Tracking
  - Jet tagging
- **End-to-end** reconstruction of multiple particles simultaneously







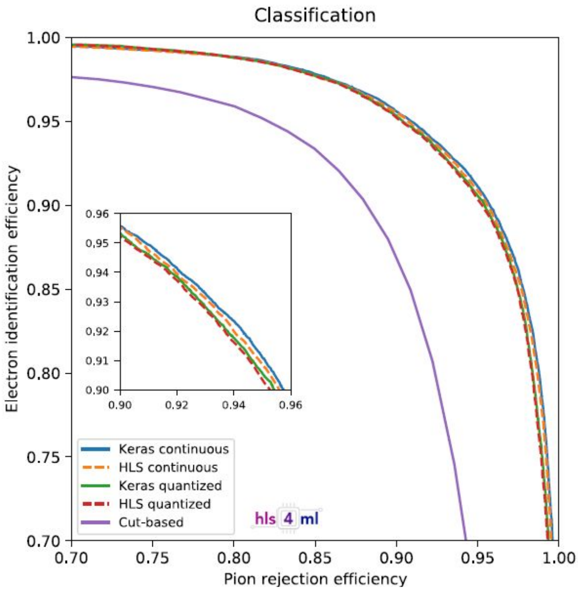
Gluon-jet image → 3 channels: track p_T, ECAL & HCAL

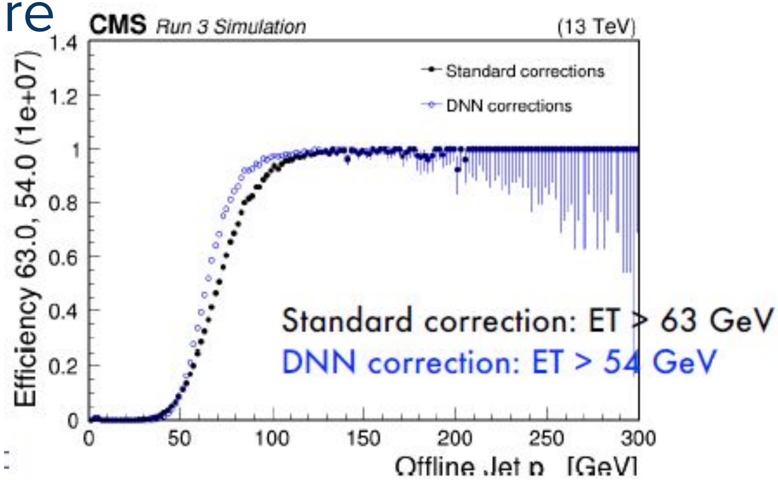
The future

- Run Anomaly detection in the trigger
  - Variational autoencoders for new physics mining at the Large Hadron Collider, J. High Energ. Phys. 2019, 36 (2019)
- Improve unfolding with invertible networks: detector ⇔ high level variables
  - Invertible networks or partons to detector and back again, SciPost Phys. 9, 074 (2020)
- Use attention to mitigate combinatorics in ttbar events: Network output should be invariant under permutations of the input jet order
  - SPANet: Generalized Permutationless Set Assignment for Particle Physics using Symmetry Preserving Attention, arXiv:2106.03898 (2021)

Deep Learning on FPGAs: HLS4ML

- Tool to deploy NNs to FPGA
  - reads as input models trained on standard DL libraries
  - implements common ingredients (layers, activation functions, etc)
- Uses HLS softwares to provide a firmware implementation of a given network
  - Pruning
  - Quantization

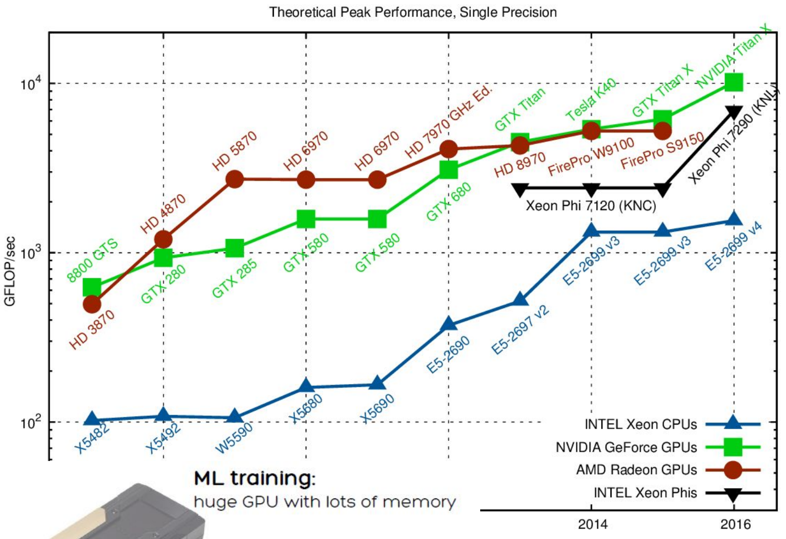


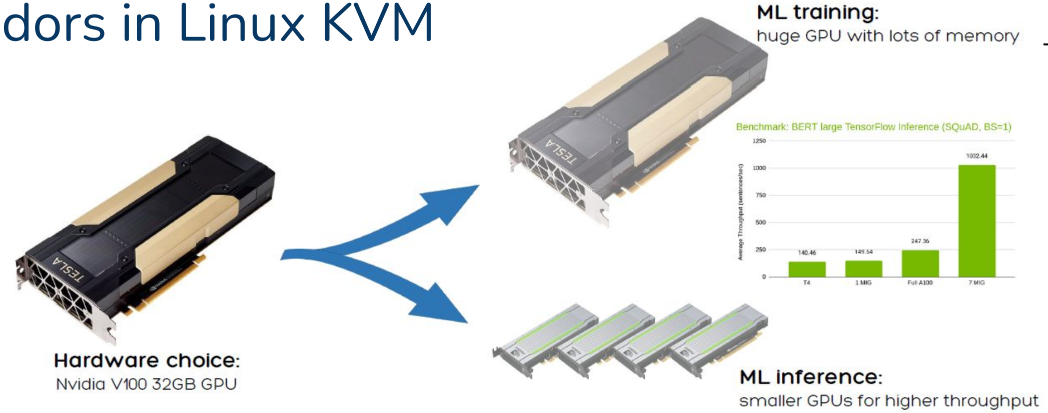


Distance-Weighted Graph Neural Networks on FPGAs for Real-Time Particle Reconstruction in High Energy Physics, arXiv:2008.03601

OpenForBC

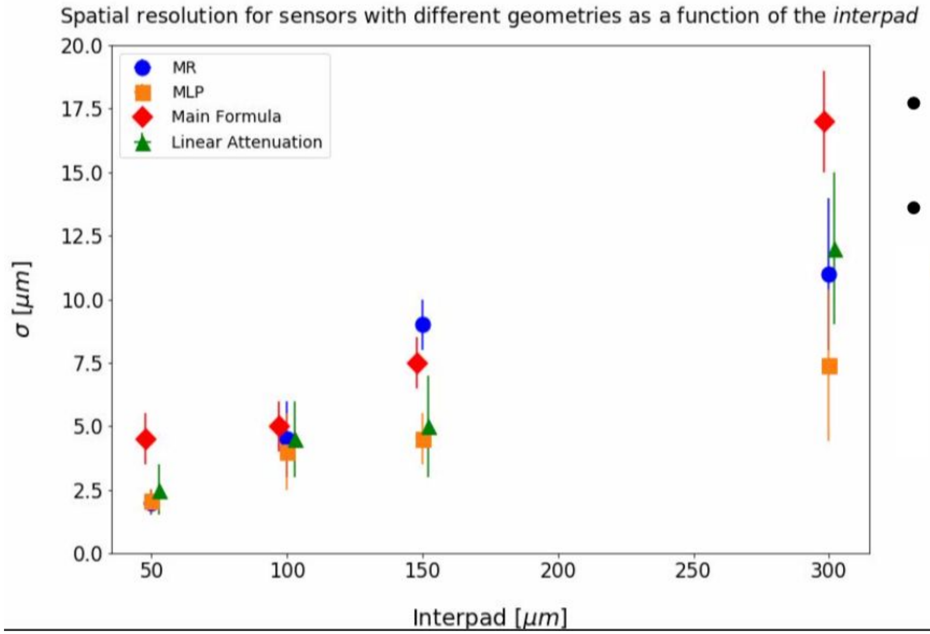
- Open For Better Computing
  - Project funded by 2021 INFN Research4Innovation (R4I) call
  - Promote use of GPUs for scientific applications
- Effortless GPU partitioning for hardware from different vendors in Linux KVM





Detector Development: RSD

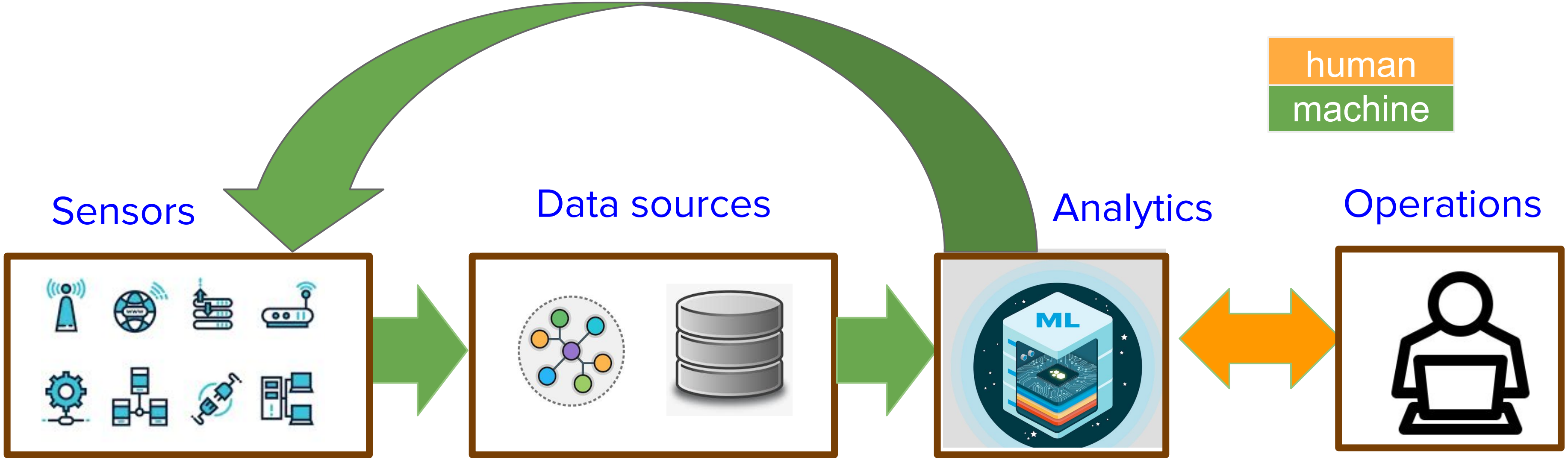
- RSDs (Resistive AC-Coupled Silicon Detectors): silicon sensors based on LGAD (Low-Gain Avalanche Diode)
  - Signal is seen over several pixels
- **Multi-Output regression (MR)** and **Multi-layer Perceptron (MLP)** models using various amplitudes as input to predict hit position



First application of machine learning algorithms to the position reconstruction in Resistive Silicon Detectors, JINST 16 P03019 (2021)

Smart Infrastructure

- Inspired from **S.M.A.R.T.** (Self-Monitoring, Analysis and Reporting Technology) for hard drives
- Predictive vs reactive maintenance for complex infrastructure
- Can be detectors, computing centers, factories, IOT



Operational Intelligence

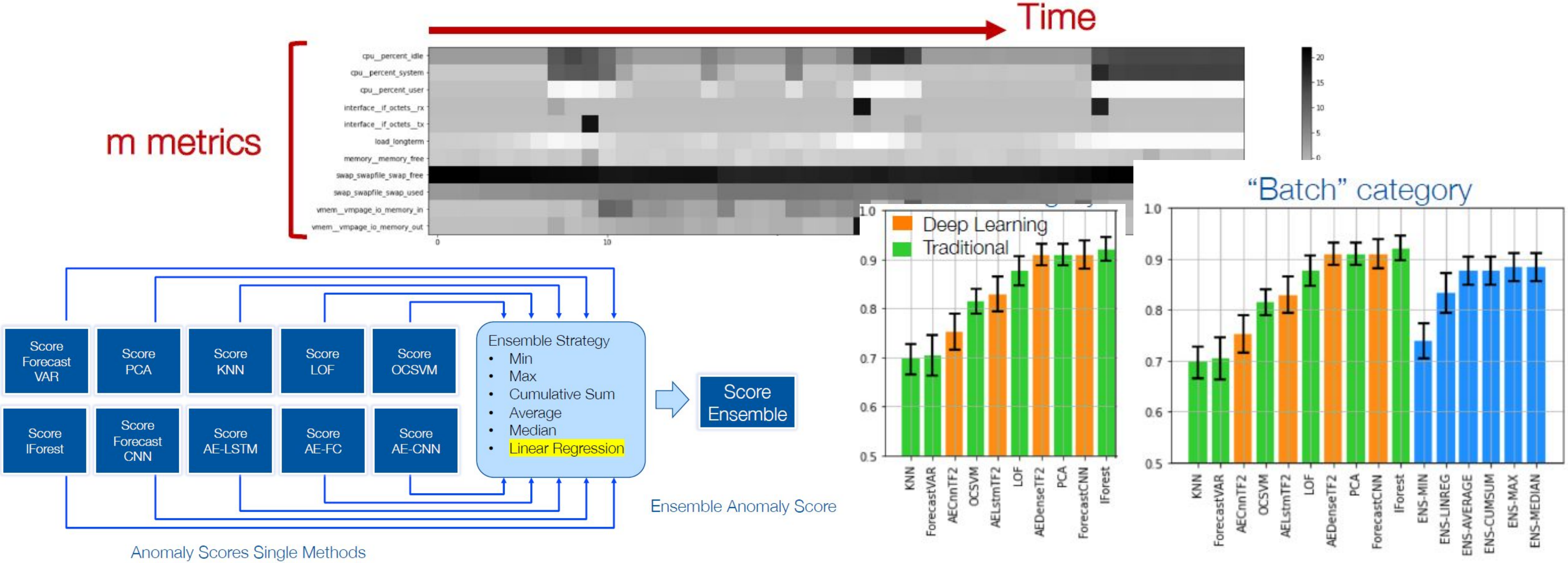
- Targets **reduction of operational costs** of distributed computing infrastructure through smart automation
  - Use case: Worldwide LHC Computing Grid (WLCG)
  - Metrics: reduction of number of tickets, number of operators, time to solve, user satisfaction
- Exploits anomaly detection in time series, natural language processing (NLP) and clusterization techniques
- Bonus: Increase resource utilisation efficiency => increase uptime, less resources wasted => “Green” development

https://operational-intelligence.web.cern.ch/

Operational Intelligence for Distributed Computing Systems for Exascale Science, EPJ Web Conf., 245 (2020) 03017

Anomaly Detection in the CERN Cloud

- Aims to identify problematic nodes in the CERN cloud
- Metrics are encoded as images or vectors according to model



Anomaly detection in the CERN cloud infrastructure, EPJ Web Conf 251, 02011 (2021)

Summary

- HEP is using MVA methods (aka ML) since > 20 years
- DL entered the scene with jet tagging
  - Now successfully used in analysis (S/B, jet to parton assignment,...)
  - Anomaly detection, attention, GANs, ... for Run 3 and beyond
- At the LHC we are resource-limited at L1, HLT and offline
  - DL may be a way to save resources and extend physics reach
    - **Sparse data**: traditional NN (CNN, RNN) may work but at a cost
    - **Custom edge computing**: inference needs to run everywhere (FPGA, custom chips, grid)
    - **Real time:** inference within 1 μs (trigger boundary)
- Many more applications: detector, computing, ...

Outlook

- Physics applications of ML and DL in a wide range of domains, and growing
- Many challenges ahead:
    - Keep the pace with AI research
        - Nowadays mainly driven by industry, science should not stand behind!
    - Foster the use of common tools/technologies
    - Exploit heterogeneous hardware
    - Deploy to production

A PROPOSAL FOR THE

DARTMOUTH SUMMER RESEARCH PROJECT

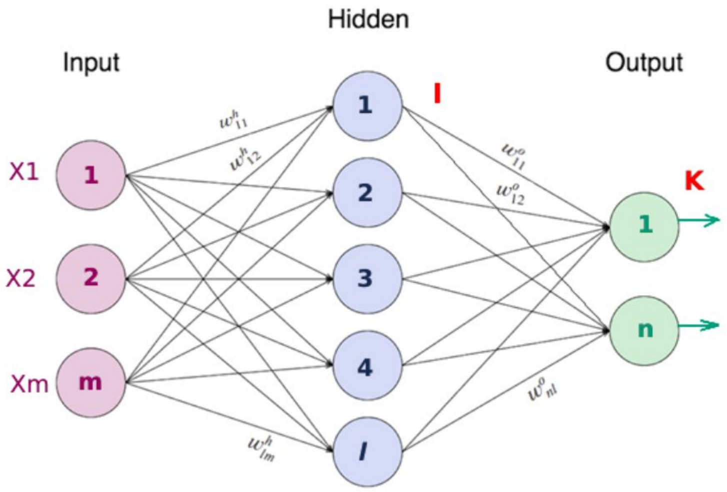
ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I. B. M. Corporation
C. E. Shannon, Bell Telephone Laboratories

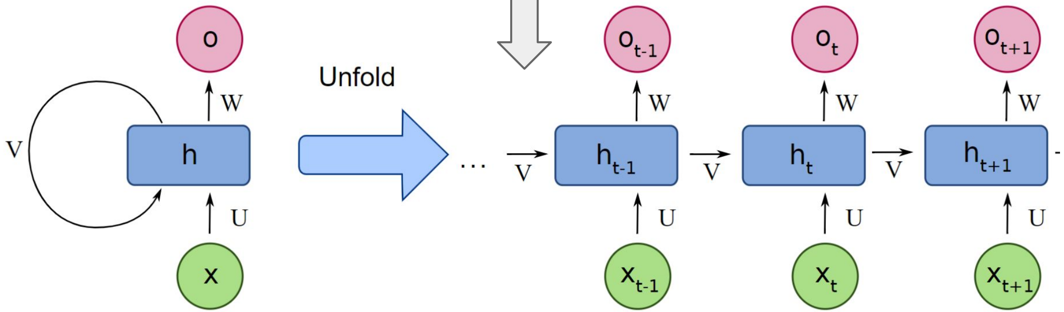
August 31, 1955

BACKUP

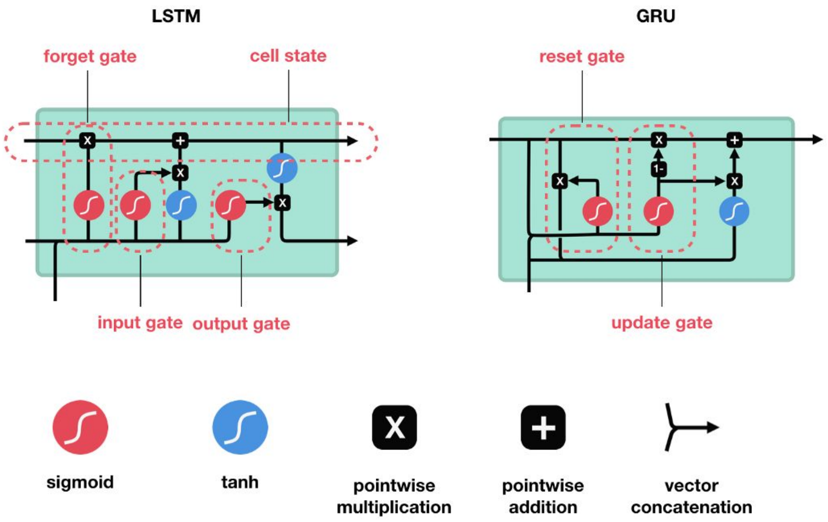
- Feed-forward Neural Network (FFNN)



- Recurrent Neural Network (RNN)



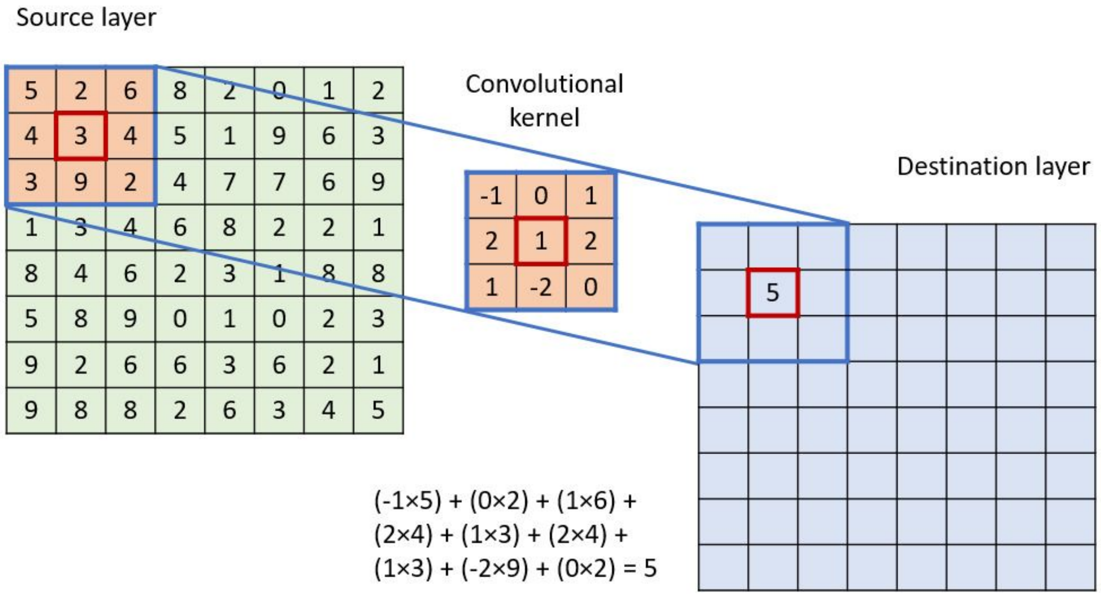
- Long short term memory(LSTM)
- Gated Recurrent Unit (GRU)

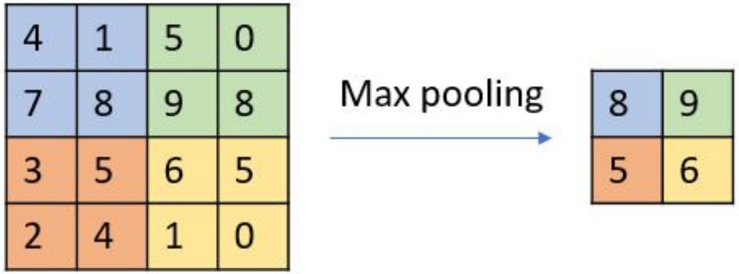


Convolutional Neural Networks (CNN)

Istituto Nazionale di Fisica Nucleare

- **Convolutional** layer: two functions produce a third that describes how the shape of one is changed by the other
- **pooling** layer: reduce dimensionality



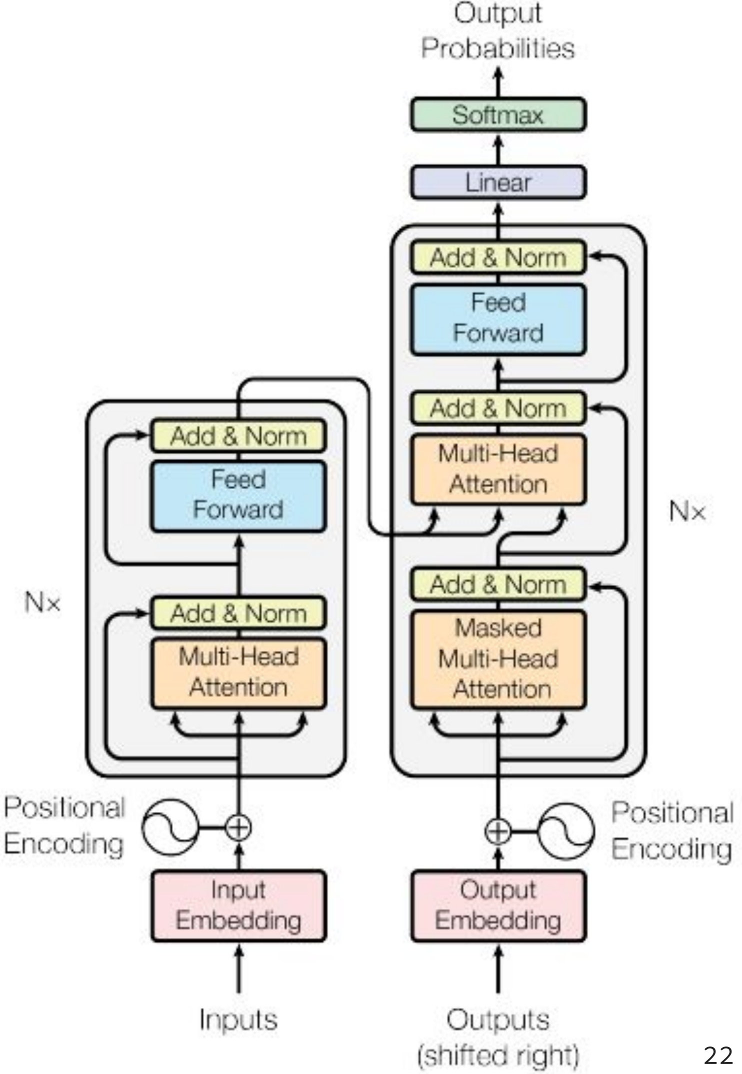


Transformers (2017)

- All you need is attention
- Self-**attention: query, key, value**:
  - the output is a weighted sum of the values, where the weight assigned to each value is determined by the dot-product of the query with all the keys:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

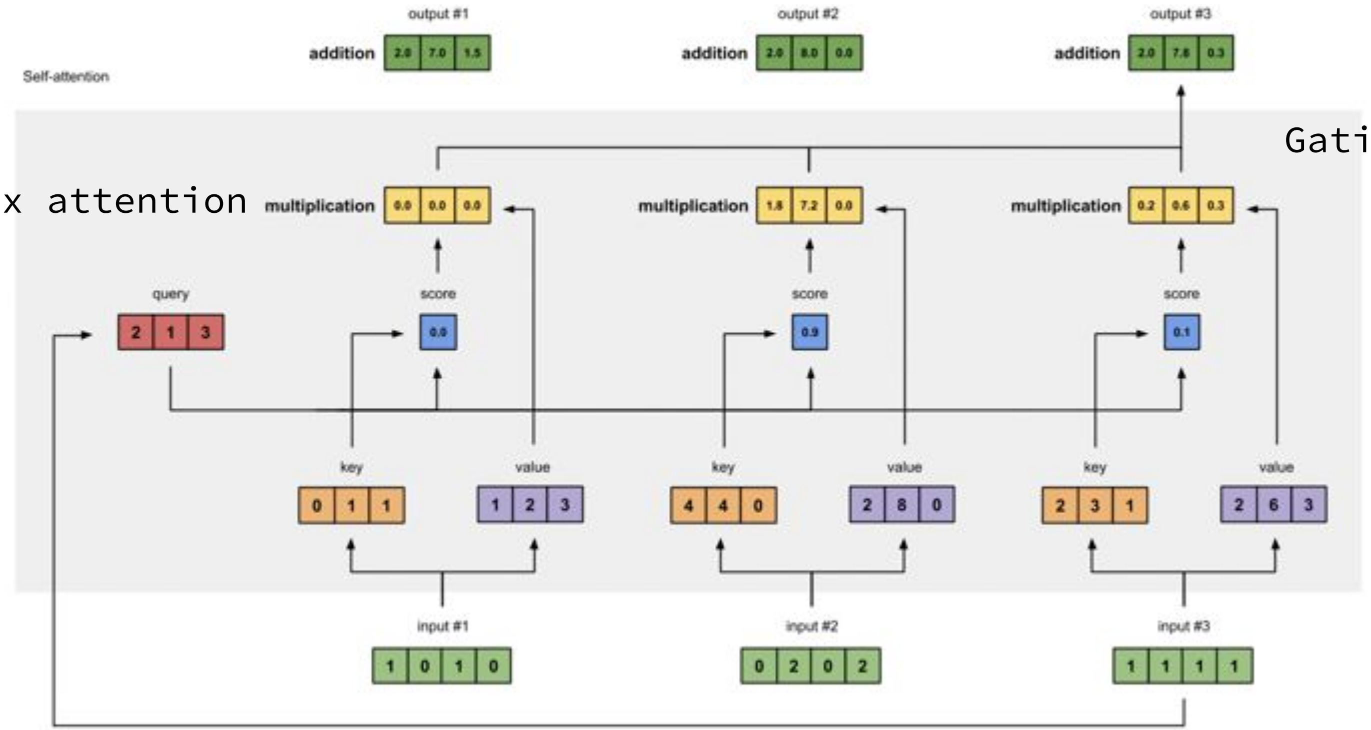
arXiv:1706.03762 [cs.CL]



Attention

Gating of the V

SoftMax attention

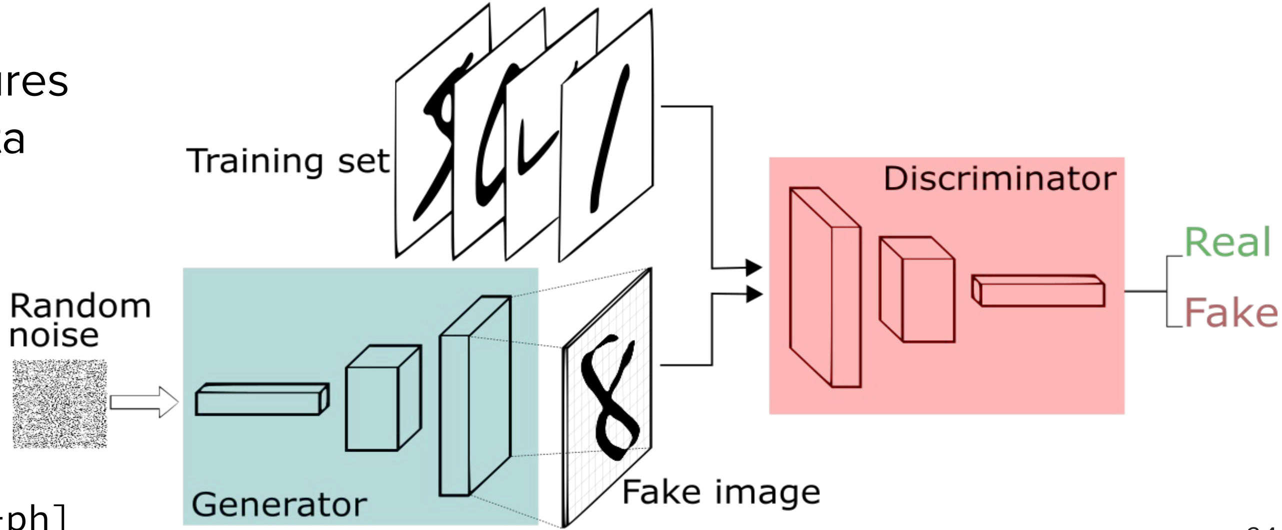


Generative Adversarial Network - GAN (2014)

Istituto Nazionale di Fisica Nucleare

- Double network: generator net and discriminator net
  - generator produces samples close to training samples
  - discriminator differentiates samples from generator and training set
  - training until discriminator can no longer distinguish

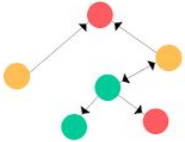
Wasserstein distance: measures the distance between the data distribution observed in the training dataset and the distribution observed in the generated examples.

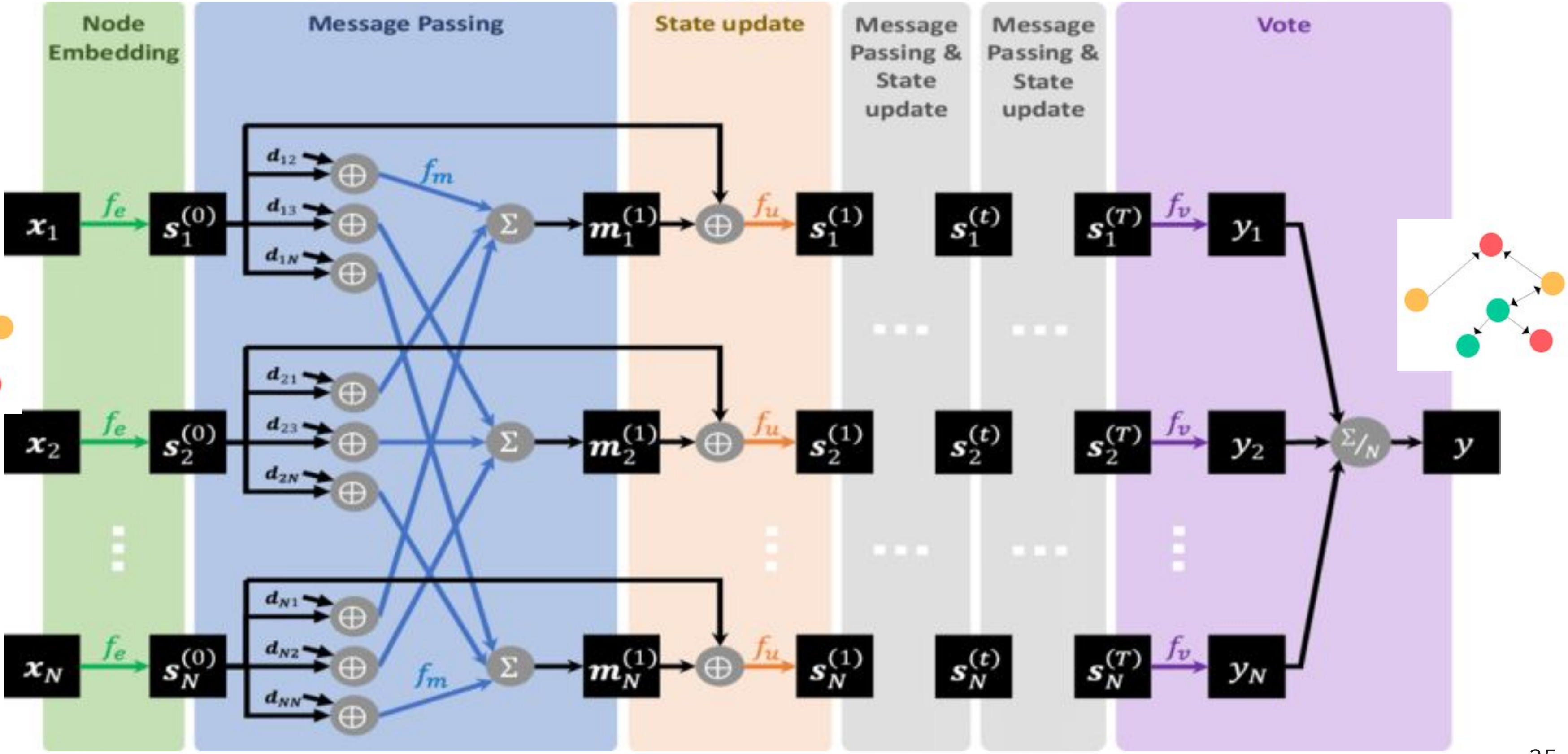


arXiv:2101.06250v1 [quant-ph]

Graph Neural Network (GNN)

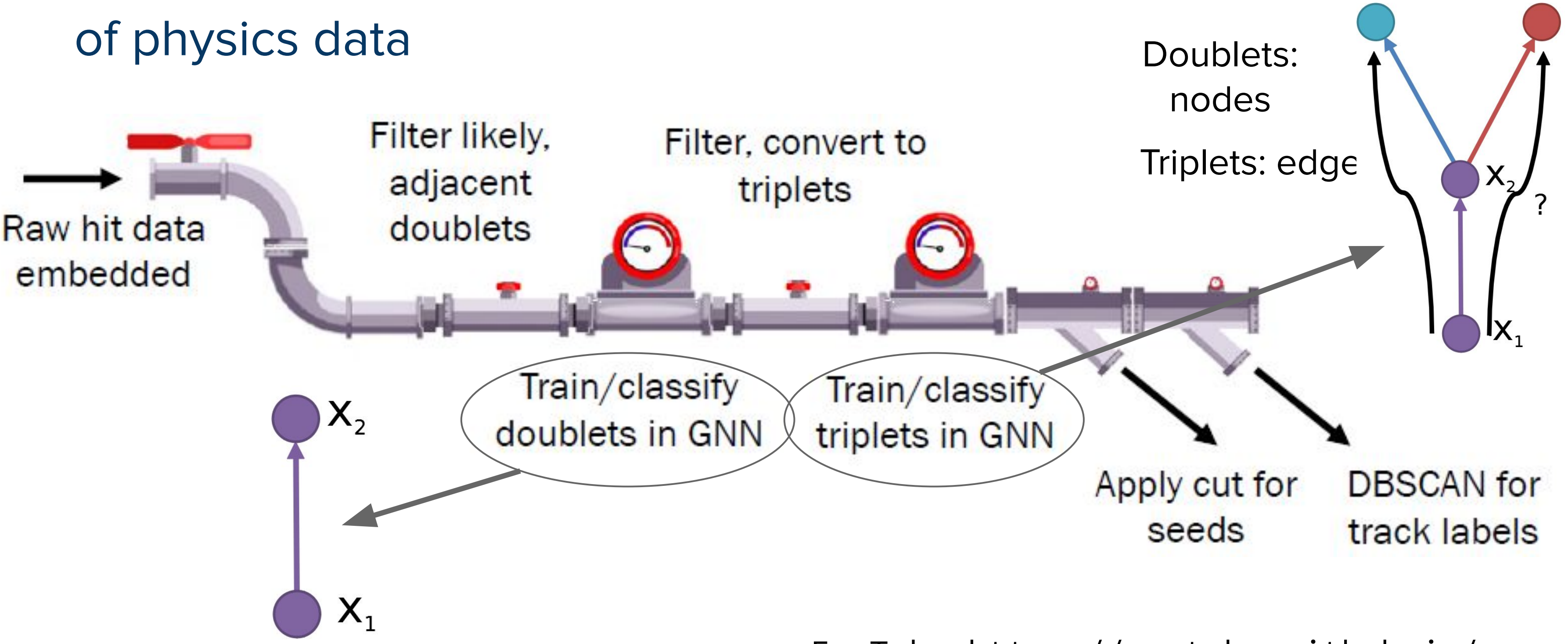
- Node prediction
- Edge prediction
- Graph prediction





GNN for Tracking

- Graphs can capture the sparsity, manifold, relational structures of physics data



Artificial General Intelligence

- **Common sense**:
  - Current systems may be easily fooled by just slight changes in the input data (for example image taken from another viewpoint)
  - Embed coordinate systems, whole-part relationship (capsules)
- **Abstract concepts**:
  - Current models may be able to distinguish between a jet and a tau, but do not know what a particle is
- **Creativity**:
  - Current models highly specialised and engineered to solve specific problems

[Murray Shanahan, Geoff Hinton]

Self Supervised Learning

- Supervised learning needs many labeled data
- Reinforced learning:
  - Not practical to train in real world (when no simulation is available)
  - takes longer than an average human for a machine to learn a new task
- **Self supervised learning:** Predict everything from everything else - learn representations, rather than learning specific tasks
  - Very large networks trained with large amount of data
  - Fill_ing the bl_anks - Word2Vec, Transformer architecture for NLP
  - Not (yet) so successful for continuous problems (image, video)

[Yann LeCun]

Consciousness Prior

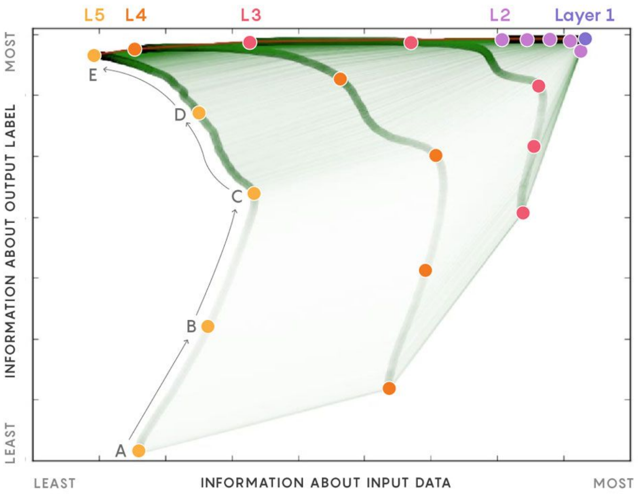
- Current deep learning:
  - **System 1:** fast, unconscious task solving
- Future deep learning:
  - **System 2:** slow, conscious task solving like reasoning, planning
- How?
  - Learn by predicting in abstract space
  - Learn representations (low dimensional vector), derived using attention from a high dimensional vector
  - The prior: the factor graph (joint distribution between a set of variables) is sparse

$$P(S) = \frac{\prod_j f_j(S_j)}{Z}$$

Yoshua Bengio, arXiv:1709.08568 [cs.LG]

Information Bottleneck

- Hidden layers represent a Markov chain of topologically distinct representations
  - Information about the inputs decreases along the hidden layers
  - I(X, h_1) > ...> I(X, h_i) > I(X, h_i+1)
- In the first epochs, the network is trained to fully represent the input data; then, it learns to forget the irrelevant details by compressing the representation of the input



A **INITIAL STATE:** Neurons in Layer 1 encode everything about the input data, including all information about its label. Neurons in the highest layers are in a nearly random state bearing little to no relationship to the data or its label.

B **FITTING PHASE:** As deep learning begins, neurons in higher layers gain information about the input and get better at fitting labels to it.

C **PHASE CHANGE:** The layers suddenly shift gears and start to "forget" information about the input.

D **COMPRESSION PHASE:** Higher layers compress their representation of the input data, keeping what is most relevant to the output label. They get better at predicting the label.

E **FINAL STATE:** The last layer achieves an optimal balance of accuracy and compression, retaining only what is needed to predict the label.

arXiv:1503.02406 [cs.LG]