AI Tools in HEP (and beyond)

Federica Legger

Digital Twins for Nuclear and Particle physics - NP-Twins 2025

Istituto Nazionale di Fisica Nucleare

Once upon a time

Istituto Nazionale di Fisica Nucleare

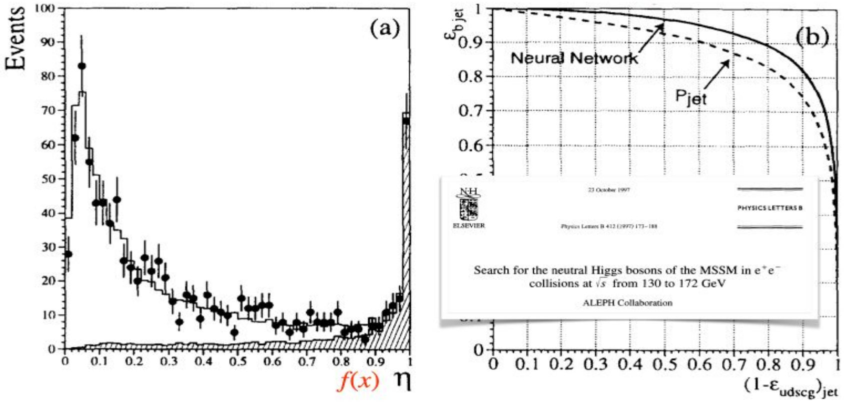


Fig. 2. (a) The output η of the neural network b tag for radiative returns to the Z for 161 GeV $q\bar{q}$ Monte Carlo (histogram) compared to the data at 161 GeV (points). The shaded region shows the contribution from generated b-jets. (b) The performance of the neural network b tag (solid line) for Monte Carlo events, presented in terms of the efficiency for identifying b-jets versus the efficiency for rejecting light quark jets. The performance of the single most powerful b tagging input variable to the neural network is shown for comparison (dashed curve).

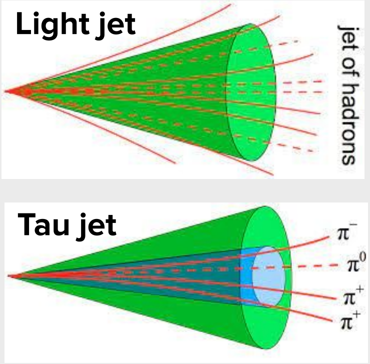
The ALEPH Collaboration,
Search for the neutral Higgs bosons of the MSSM in e+e− collisions at √s from 130 to 172 GeV
Phys.Lett.B 412 (1997) 173-188

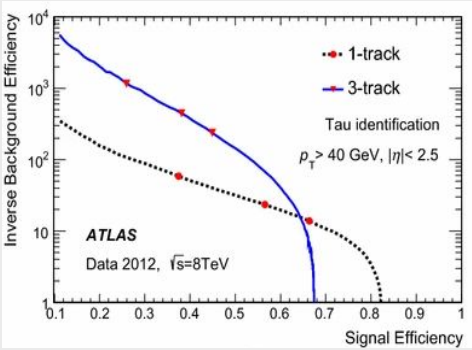
- **TMVA** - Toolkit for Multivariate Data Analysis, arXiv:physics/0703039
- TMVA fully integrated in ROOT in 2013

- ML mainly for classification and regression tasks
  - decision trees, support vector machines, cellular automata, perceptrons

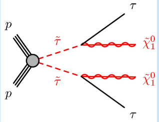
Particle Identification, Rare Processes

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

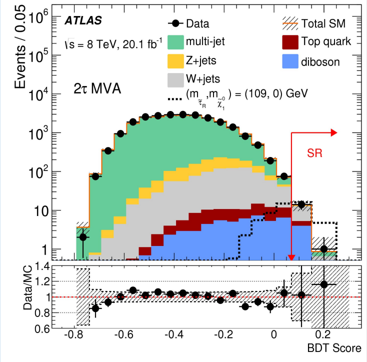


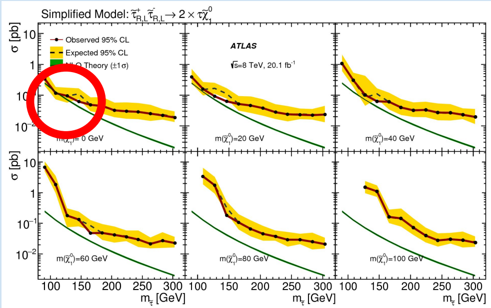


- **Search for direct stau production**
  - **BDT** with low and high level variables



Stau mass 109 GeV





The ATLAS Collaboration, *Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at √s=8 TeV*, EPJC 75: 303 (2015)

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

AI applications for LHC today

- At the LHC we are **resource-limited** everywhere!
  - trigger and analysis level
  - generation, simulation, reconstruction, tracking

- **Deep Learning (DL)** may help to save resources and extend the physics reach

- Also for: **detector, operation and data quality, computing**

- Challenges:
  - **Sparse data**: traditional NN (CNN, RNN) may work but at a cost, GNNs, Transformers
  - **Custom edge computing**: inference needs to run everywhere (FPGA, custom chips, grid)
  - **Real time:** inference within 1 μs (trigger boundary)

ML Workflows in HEP

- Commonalities with “standard” HEP data analysis:
  - access to **exascale datasets**
    - distributed storage
  - access to **heterogeneous computing**
    - accelerators, GPUs
  - **Reproducibility** and reusability within large user communities
    - share data, software and code
  - Possibly **interactive** analysis

“Traditional” grid-based computing model (WLCG) not ideal for ML

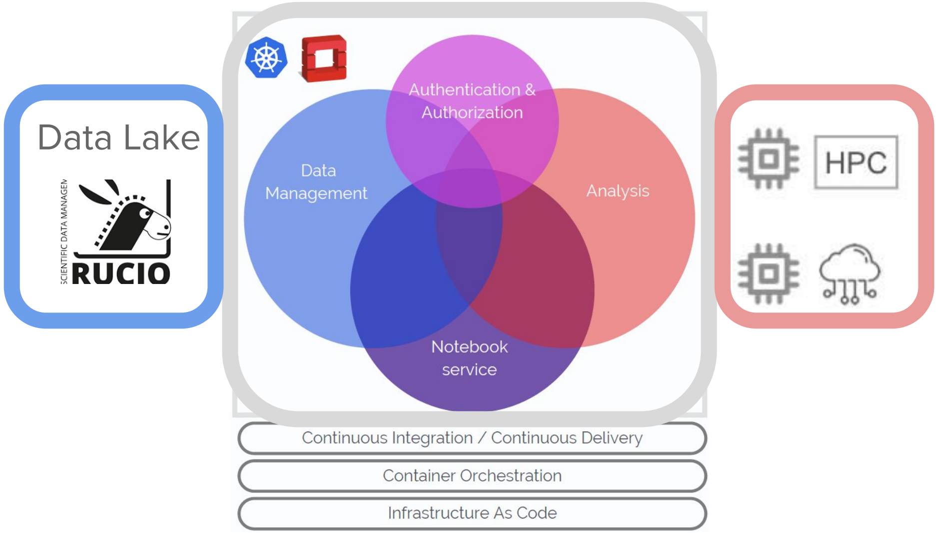
WLCG: Worldwide LHC Computing Grid

The Virtual Research Environment (VRE)

Istituto Nazionale di Fisica Nucleare

- An **open source** analysis platform for researchers to develop, share and reproduce scientific results

- Provides access to:
  - data and software
  - computing resources
  - In compliance with FAIR* standards



- VRE developed by the *ESCAPE EU project*
- Similar efforts referred to as **Analysis Facility***

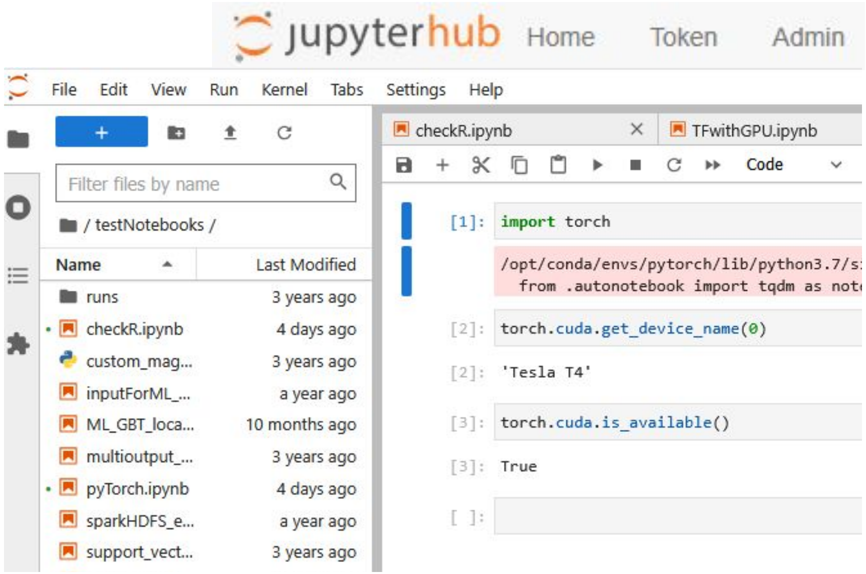
* findable, accessible, interoperable, reproducible

* *White paper*

The Yoga Cluster At INFN Turin

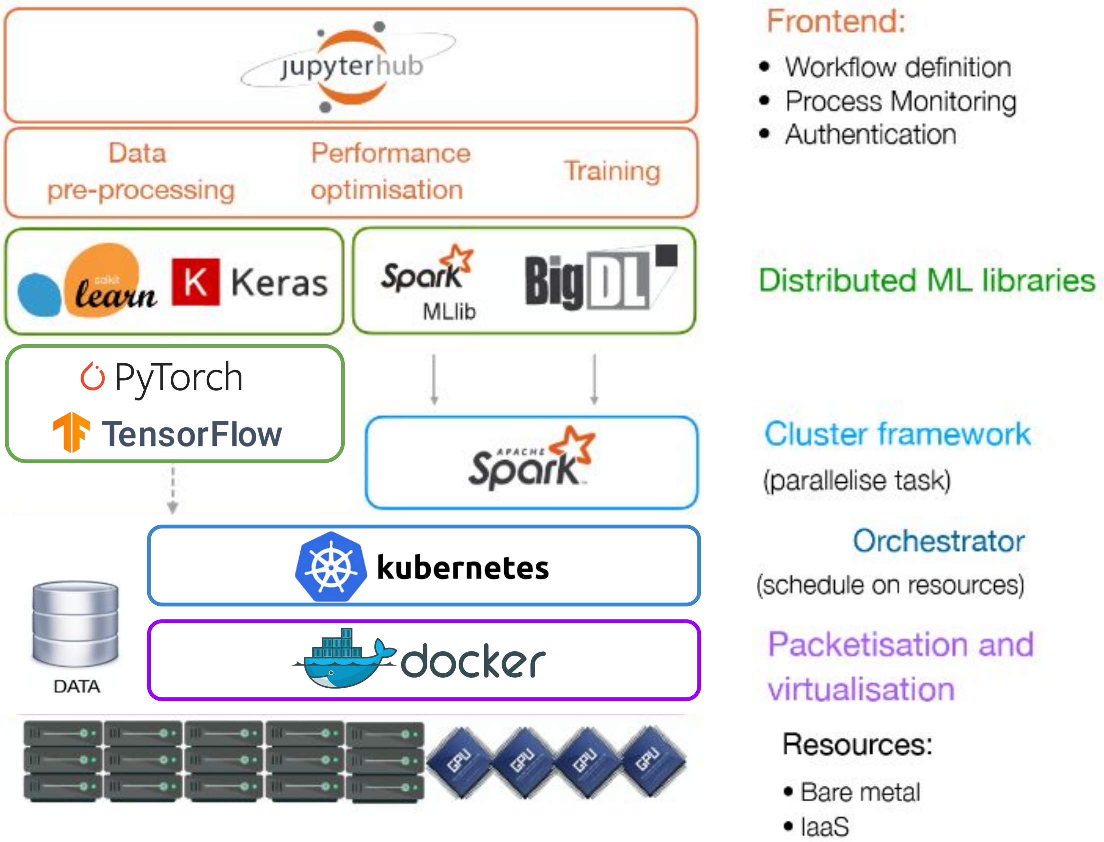
- **Jupyter Hub for local users** (INFN and University)
  - ML/AI, multi-core simulations

- Jupyter notebooks are:
  - open-source web application
  - interactive
- create and share documents:
  - visualisations
  - narrative text
  - live code
  - equations



S. Bagnasco, G. Fronzé, F. Legger, S. Lusso, S. Vallero, *Delivering a machine learning course on HPC resources*, EPJ Web of Conferences 245, 08016 (2020)

The Yoga Cluster: MLaaS - ML as a Service

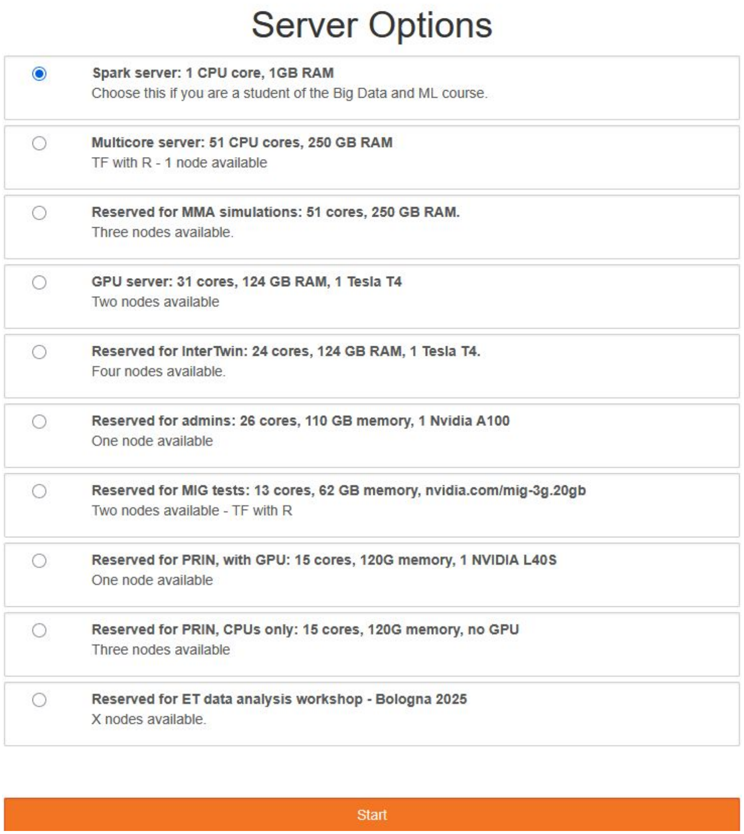


- Hardware:
  - 240 cores
  - 2 TB of RAM
  - 1 Gbps Ethernet
  - 4.6 TB ephemeral disk
  - 6.5 TB shared gluster storage
  - **6 Nvidia Tesla T4**
  - **2 Nvidia A100**
  - **1 Nvidia Grace Hopper**

- Being expanded with PNRR resources

The Yoga Cluster: projects

- On-going:
  - Spark cluster for students of the “**Big Data and Machine Learning**” course for Ph.D students at University of Turin
  - **Intertwin**, development of the Virgo Digital Twin for transient noise characterization
  - S. Argirò, F. Oberto, Burnup analysis in LFR (**Lead Fast Reactor**).
  - P. Angelino, S. O'Toole, PAGEpy - Predictive Analysis of **Gene Expression** with Python
  - R. Bonino et al, **PRIN SKYNET, Deep Learning for Astroparticle Physics**

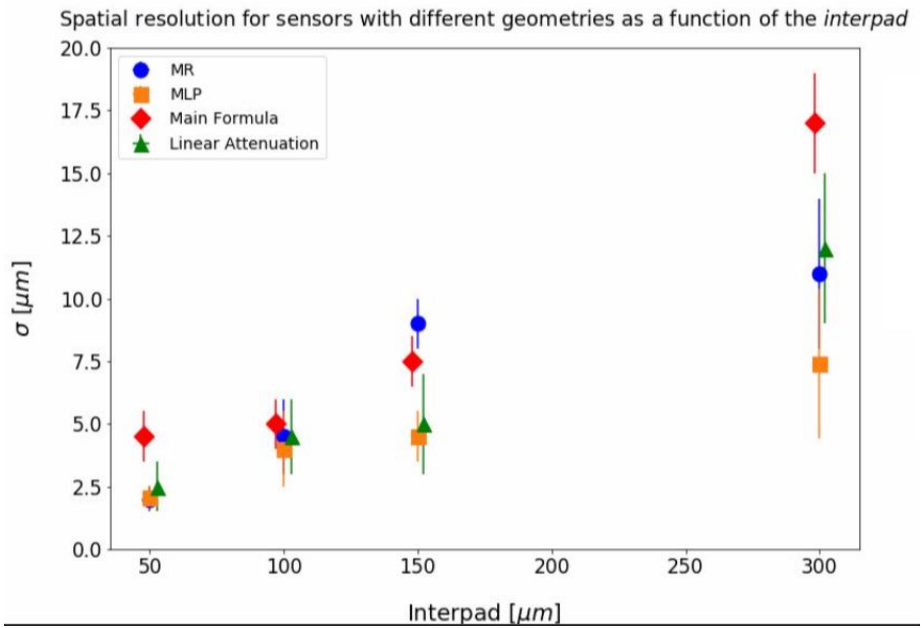


Detector Development: RSD

Developed on Yoga

Istituto Nazionale di Fisica Nucleare

- 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



F. Siviero et al., *First application of machine learning algorithms to the position reconstruction in Resistive Silicon Detectors*, JINST 16 P03019 (2021)

The InterTwin Project

Istituto Nazionale di Fisica Nucleare

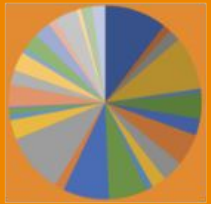
- EU-funded project, https://www.intertwin.eu/
- Aim: design and build a prototype of an interdisciplinary **Digital Twin (DT) Engine**, based on a co-designed Blueprint Architecture

36 months
from
Sep 22
To
Aug 25

Budget

12 M Euros

30 Partners
with
1500 PMs
Coordinated by
EGI



2 + 2 + 3 DT
Use Cases

from

HEP
Astro
Climate
Environment

- **Digital twin**: a virtual representation (typically a simulation or a ML model) of a system, updated from real-time data

InterTwin: the Virgo Digital TWIN

- sensitivity of GW interferometers is limited by noise
- DT aims to realistically simulate and detect transient noise (glitches) quasi-real time
- Final goal: veto and (later) de-noise

- GW signal → strain
- transient noise (**glitches**) → auxiliary channels and strain

- map glitches from auxiliary channels to strain
- **deep generative models** to capture non-linear structures in the data

- **Veto** events containing glitches
- **Noise subtraction** from the strain channel

L. Asprea, E. Cellini, F. Legger, A. Romano, F. Sarandrea, S. Vallero, *GlitchFlow, a Digital Twin for transient noise in Gravitational Wave Interferometers,* presented at CHEP 2024

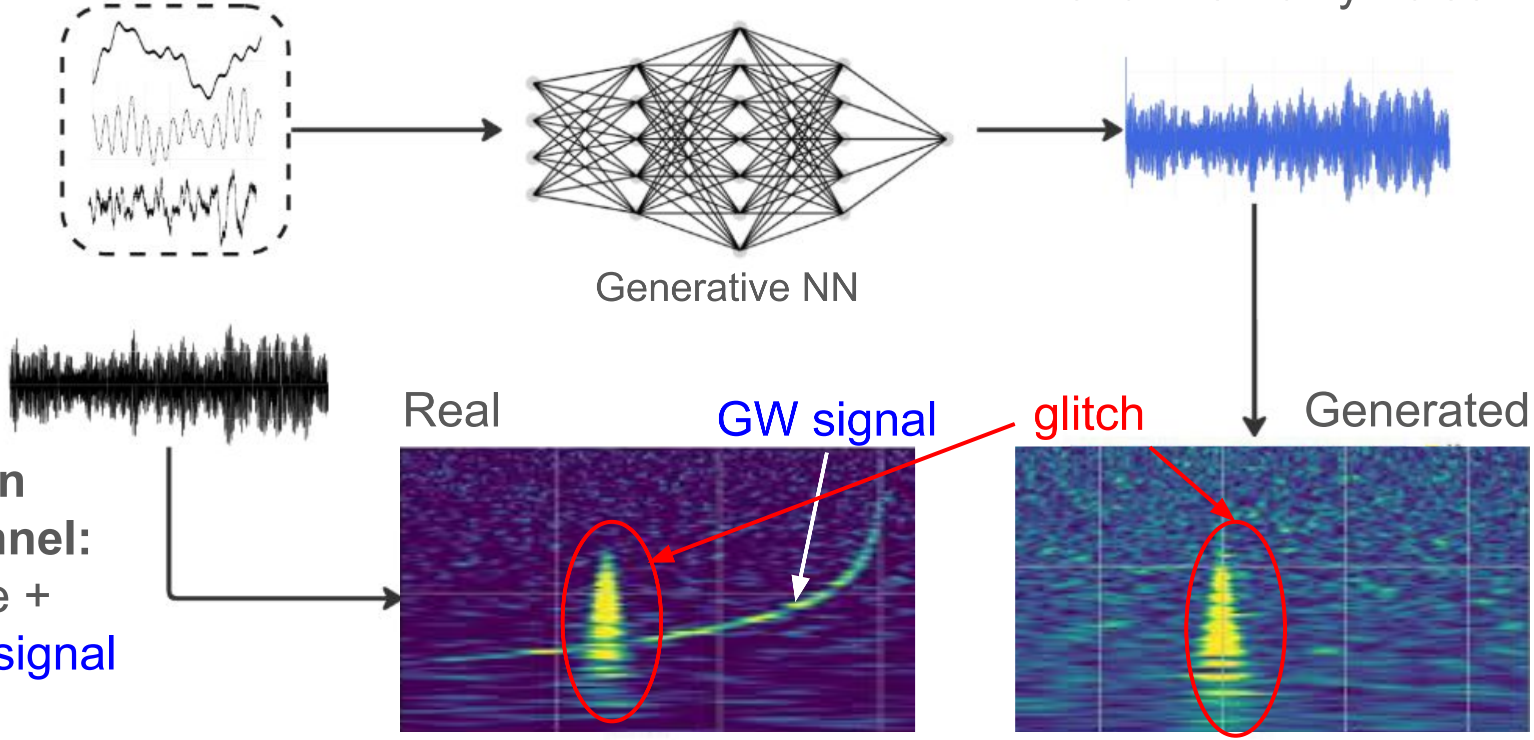
The Virgo DT: GlitchFlow

Developed on Yoga

INFN Istituto Nazionale di Fisica Nucleare

auxiliary channels: only noise

generated strain channel: only noise



strain channel: noise + GW signal

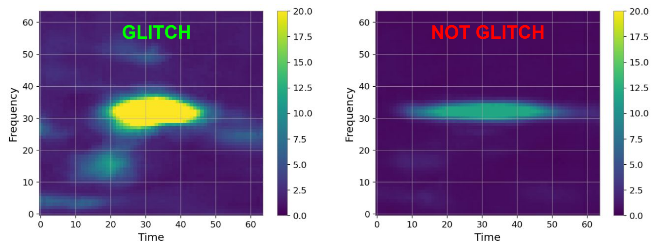
The Virgo DT: results

Developed on Yoga

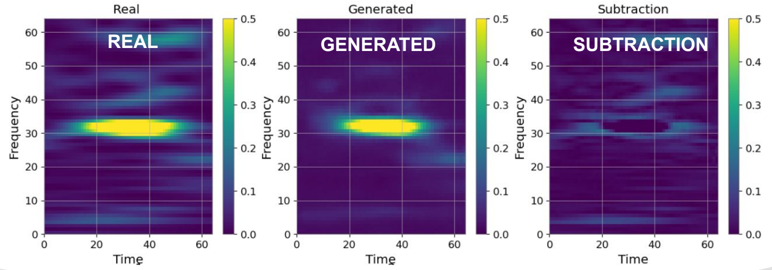
- **Glitch definition**: Cluster of at least 10 pixels with SNR above threshold
  - This choice mimics actual alert mechanisms used by Virgo (Omicron)
- Use Clustering mechanism as Classifier on generated data
  - Test set: 1083 Glitches, 536 empty background

	SNR 8	SNR 10	SNR 15
Glitch generation Accuracy Model is able to correctly generate a glitch given control channels	99.0%	98.7%	97.8%
Glitch position Accuracy Model generates glitches with right time and frequency. Intensity is saturated at trigger SNR	25.9%	99.0%	99.6%

Glitch generation

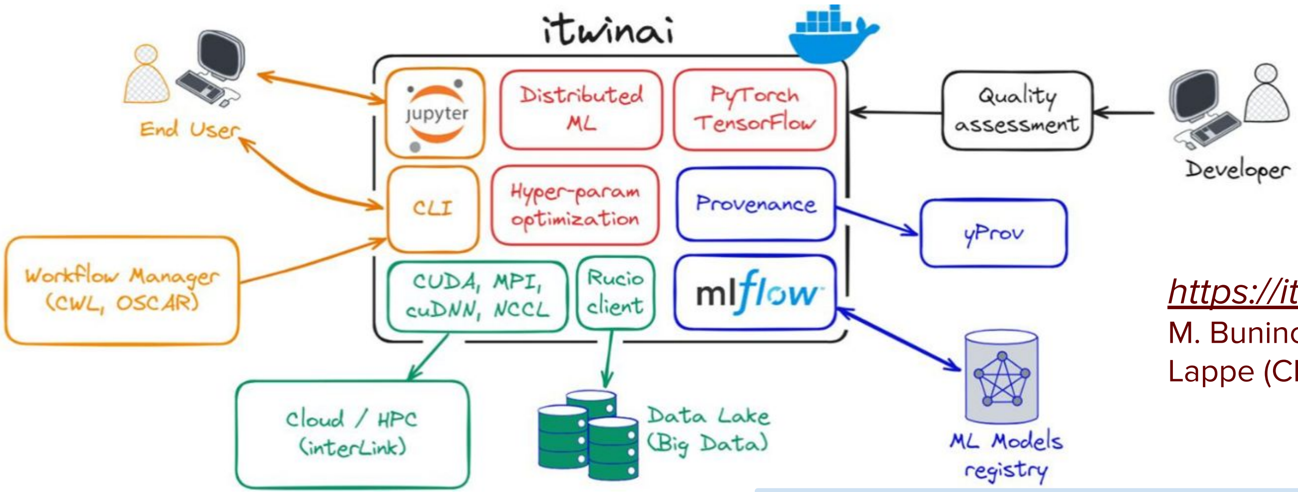


Glitch position



InterTwin: the itwin-ai framework

Istituto Nazionale di Fisica Nucleare



https://itwinai.readthedocs.io/latest/

M. Bunino (CERN), J. Saether (CERN), A. Lappe (CERN), M. Girone (CERN)

- developers define and load config files and training scripts;
- provides common interface for AI through a CLI;

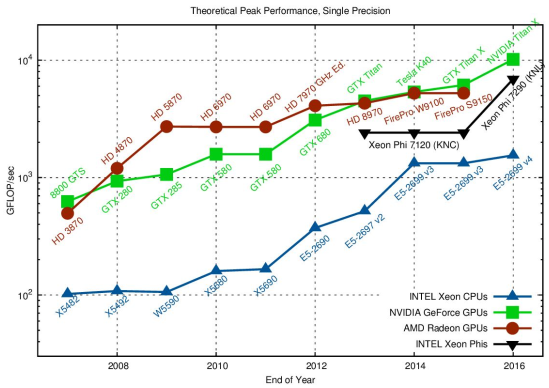
- integrates with PyTorch, TensorFlow, and other tools for distributed ML and hyperparameter optimization;
- Models and logs managed by MLFlow tracking server.

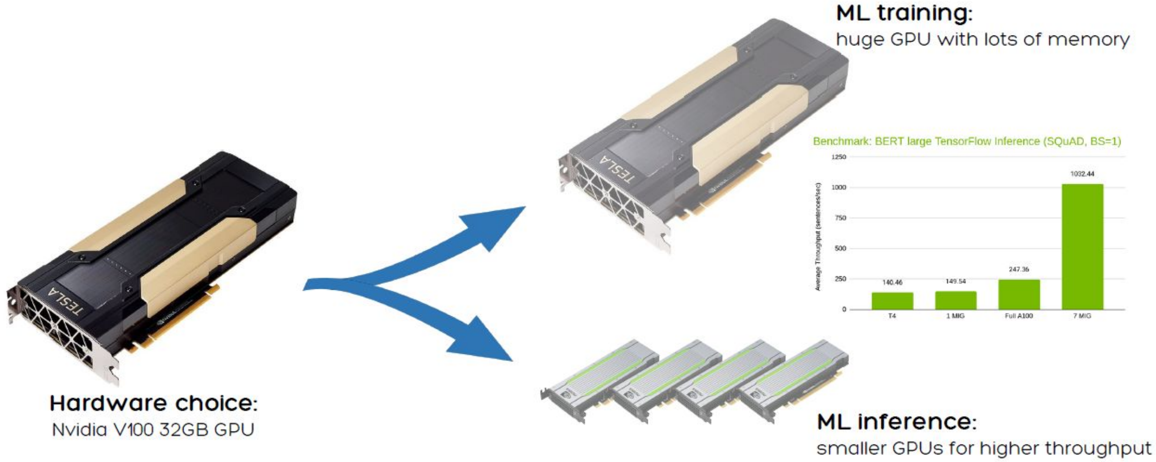
OpenForBC

Developed on Yoga

Istituto Nazionale di Fisica Nucleare

- Open For Better Computing
  - funded by 2021 INFN *Research4Innovation* (R4I) call
  - Promote use of GPUs for scientific applications





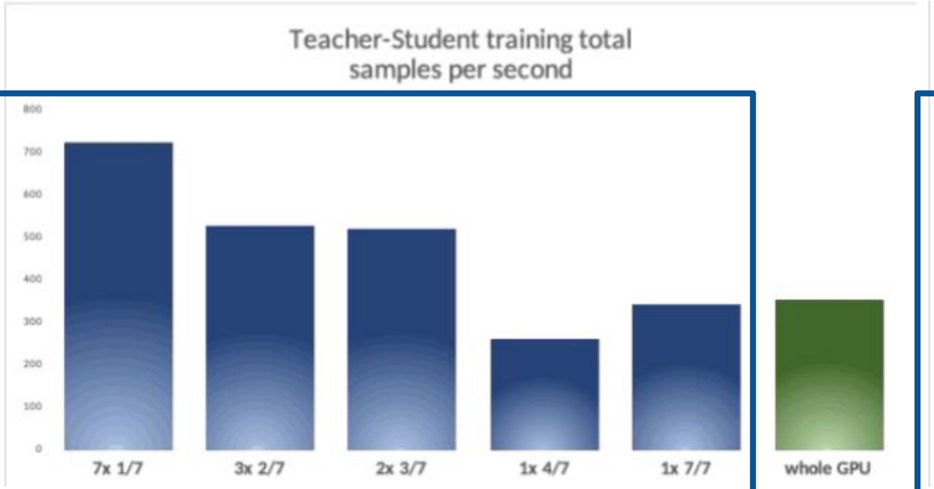
Open ForBC

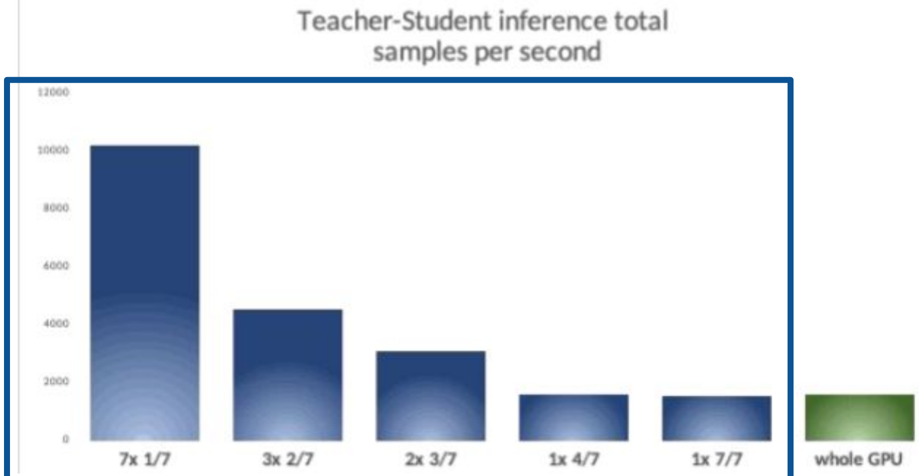
Effortless GPU partitioning in Linux KVM

OpenForBC: Results

Developed on Yoga

Istituto Nazionale di Fisica Nucleare



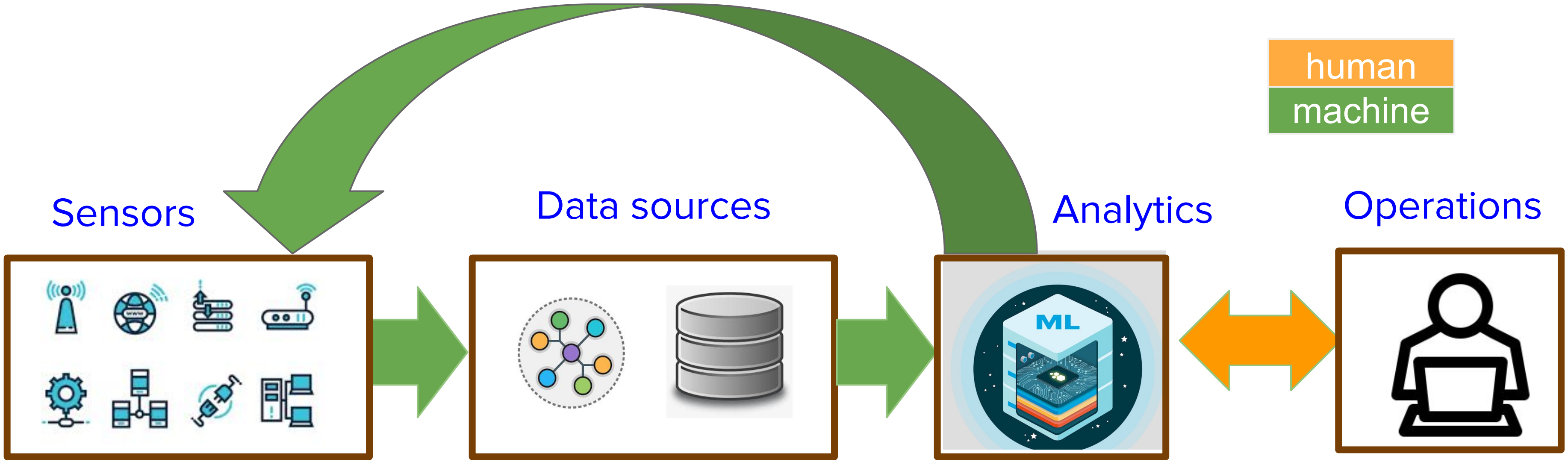


- peak throughput computed as the sum of the average throughput of all creatable partitions given a specific profile
- All creatable partitions have been allocated and loaded with computation

F. Legger, G. Fronzè, *OpenForBC, the GPU partitioning framework*, *POS (2022) 221, DOI:10.22323/1.414.0221*

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 (2019-2022)

OI INFN Istituto Nazionale di Fisica Nucleare

- Targets **reduction of operational costs** of distributed computing infrastructure through smart automation

- Exploits anomaly detection in time series, natural language processing (NLP) and clusterization techniques

- **Use case:** Worldwide LHC Computing Grid (WLCG)
- **Metrics:** reduction of number of tickets, number of operators, time to solve, user satisfaction

- **Bonus:** Increase resource utilisation efficiency
- increase uptime, less resources wasted => Save CO2

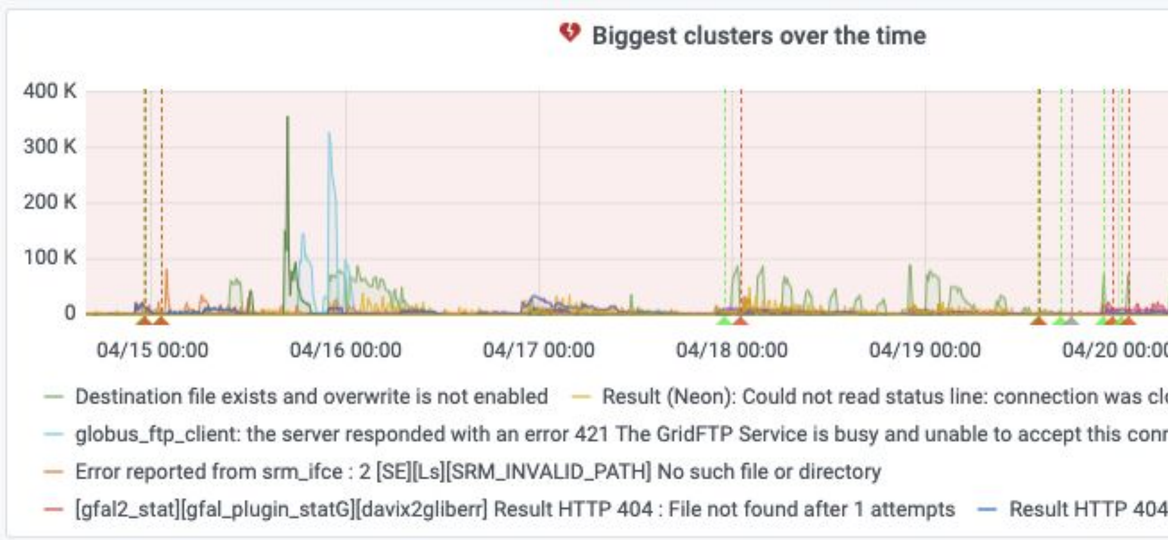
Operational Intelligence for Distributed Computing Systems for Exascale Science, EPJ Web Conf., 245 (2020) 03017

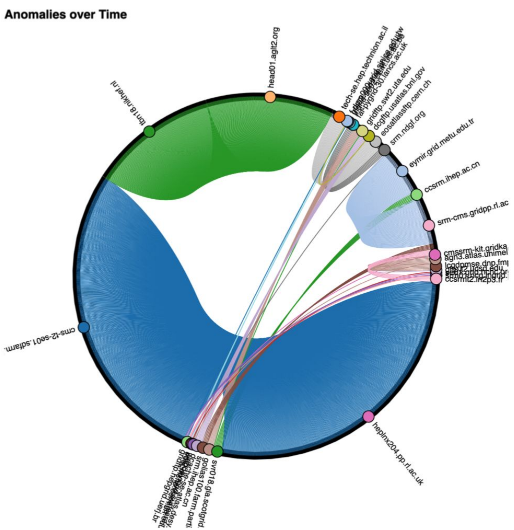
https://operational-intelligence.web.cern.ch/

FTS Log Analysis

- **FTS (File Transfer Service)** distributes LHC data across WLCG
- Billions of files transferred per days => hundred thousands errors

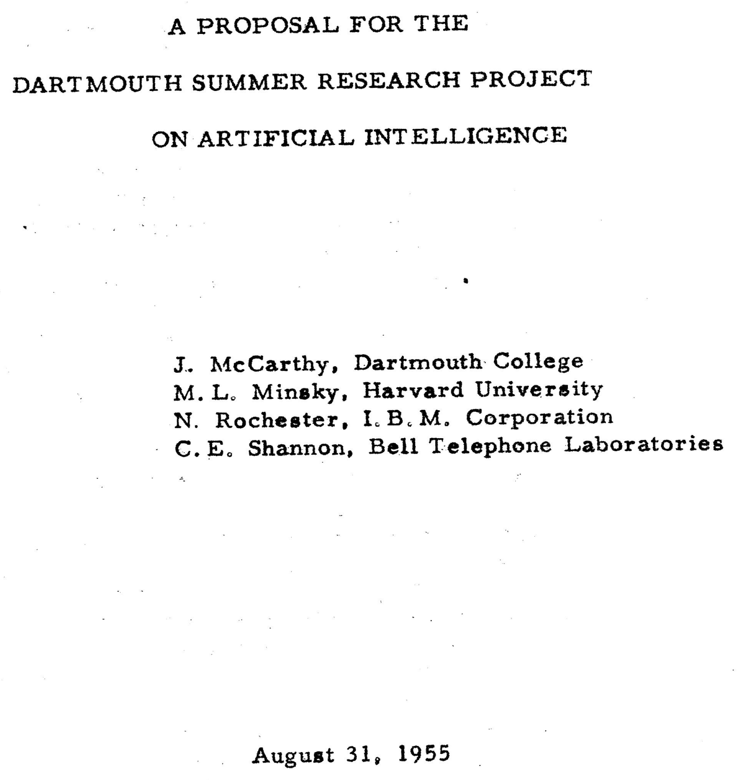
- Help operation teams analyse multiple transfer errors
- **Clusterize** similar error messages, detect **anomalous links**





Outlook

- HEP 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

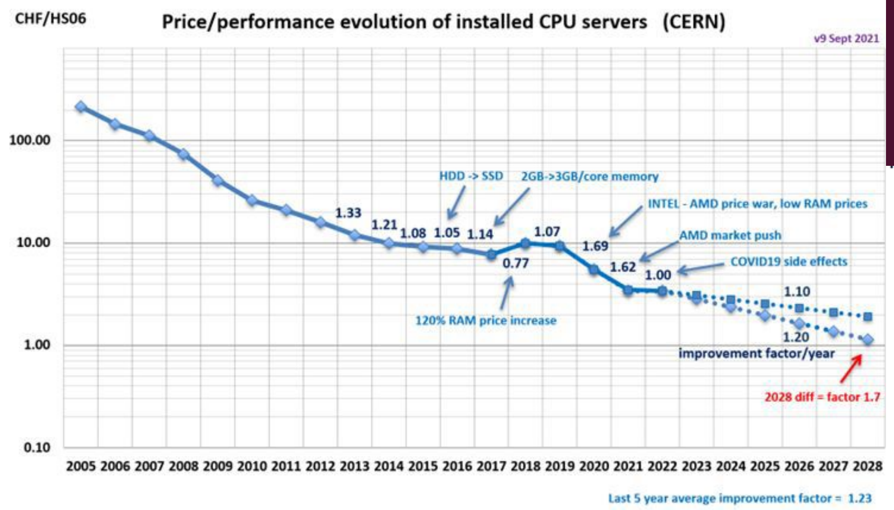
BACKUP

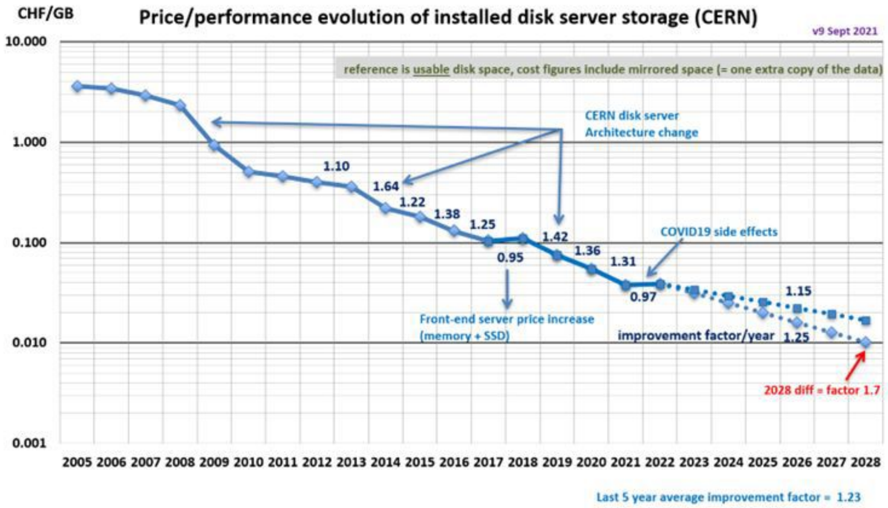
Challenges

- Hardware cost evolution
- Semiconductors shortage
- Energy cost

Vs:

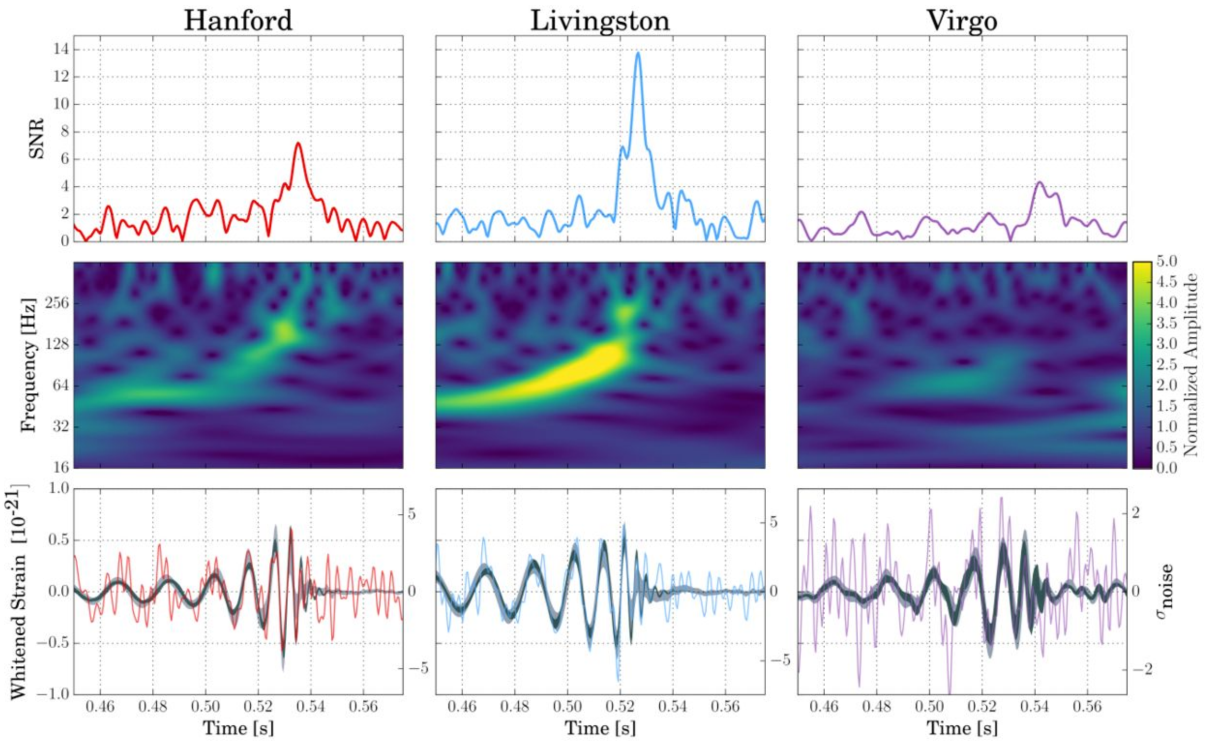
- **LHC Run 3**: ALICE x100 recorded events, LHCb x30 throughput
- **LHC Run 4**: ATLAS and CMS x10 luminosity, x5 event size





Detection of Gravitational Waves

- based on the strain measurement (deformation of the interferometer arms)
- interferometer status and environmental conditions are monitored in the auxiliary channels

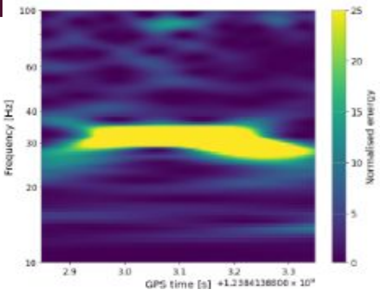


GW170814, PRL 119, 141101 (2017)

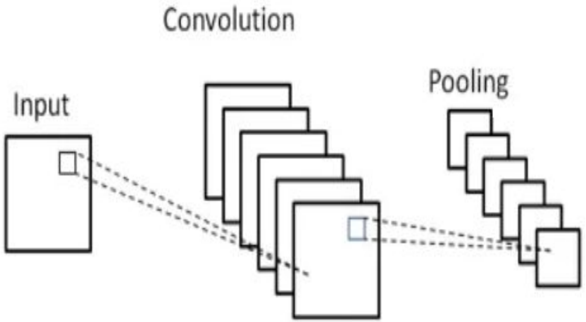
GlitchFlow: Generative Model

- Code in Pytorch
- L1-Loss= Σ |Generated Output - Target Output|

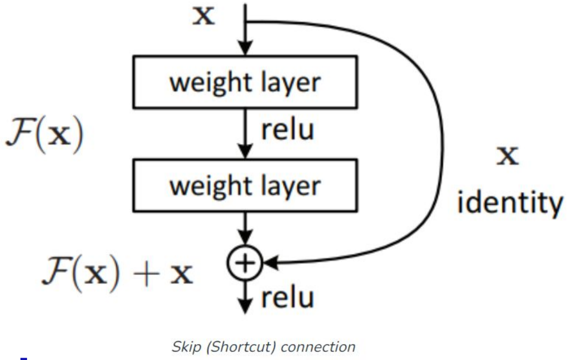
- **Input:** 2 aux channels



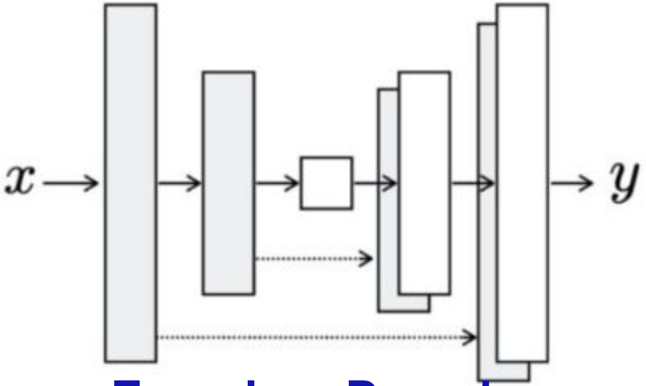
CNN



ResNet (residual block)



U-Net



Decoder

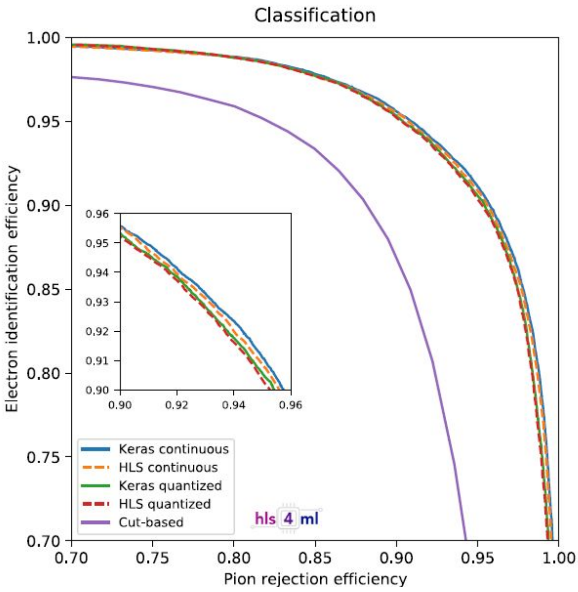
Encoder - Decoder

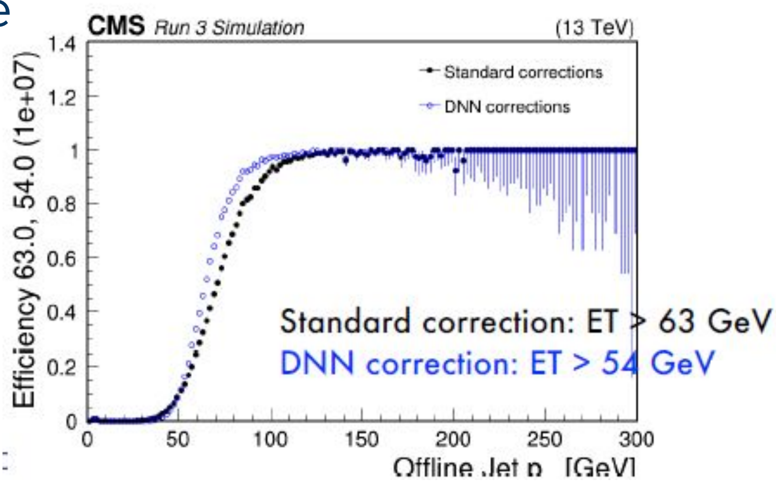
The Yoga Cluster: past Projects

- N. Cibrario, *Joint machine learning and analytic track reconstruction for X-ray polarimetry with gas pixel detectors*, A&A, 674 (2023) A107
- E. Angelino (PhD Thesis, chapter 5) *Calibration, simulation and analysis of the low-energy region in XENON project for direct dark matter search*
- M. Branchesi, U. Dupletsa, B. Banerjee, S. Ronchini
  - *A&A 690 (2024), A362*
  - *JCAP 07 (2023) 068*
  - *Astron.Comput. 42 (2023) 100671 (2023)*
  - *A&A 665, A97 (2022)*
  - *A&A 678, A126 (2023)*
- L. Tabasso (bachelor thesis), *Valutazione delle Prestazioni di Codice per Analisi di Onde Gravitazionali su Architetture HPC Innovative*
- F. Legger, G. Fronzè, *OpenForBC, the GPU partitioning framework*, *POS (2022) 221, DOI:10.22323/1.414.0221*
- F. Siviero et al., *First application of machine learning algorithms to the position reconstruction in Resistive Silicon Detectors*, *JINST 16 P03019*
- E. Grasso (bachelor thesis), *Machine Learning - Un progetto di alternanza scuola lavoro*
- M. Olocco (master thesis), *Natural Language Processing techniques for error message analysis in WLCG data transfers*

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

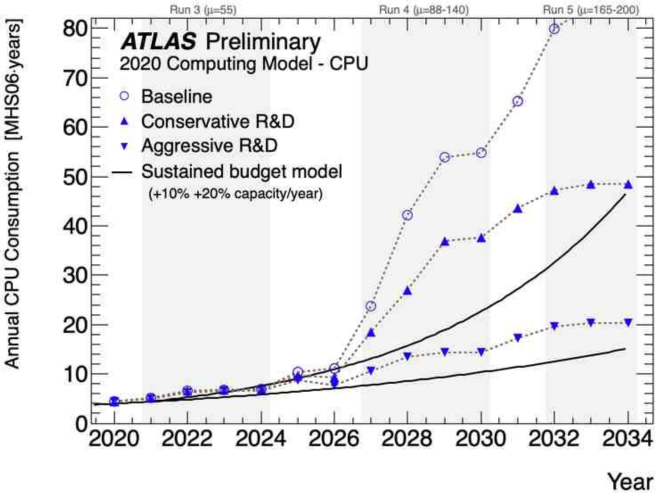


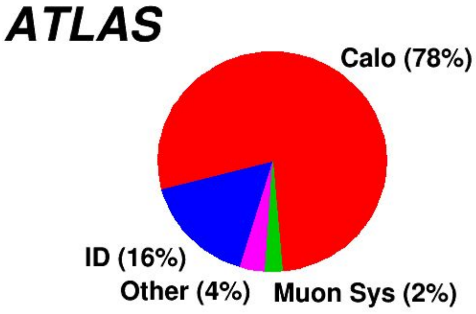


Distance-Weighted Graph Neural Networks on FPGAs for Real-Time Particle Reconstruction in High Energy Physics, arXiv:2008.03601

Deep Learning for simulation

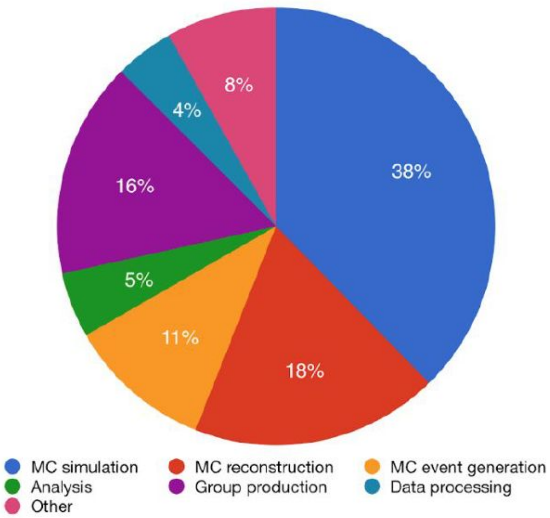
- Computing demands increase nonlinearly with increasing pileup
- LHC Run 2: full detector simulation (Geant4) took ~40% of grid CPU resources for CMS & ATLAS
- Calorimeter simulation most CPU intensive



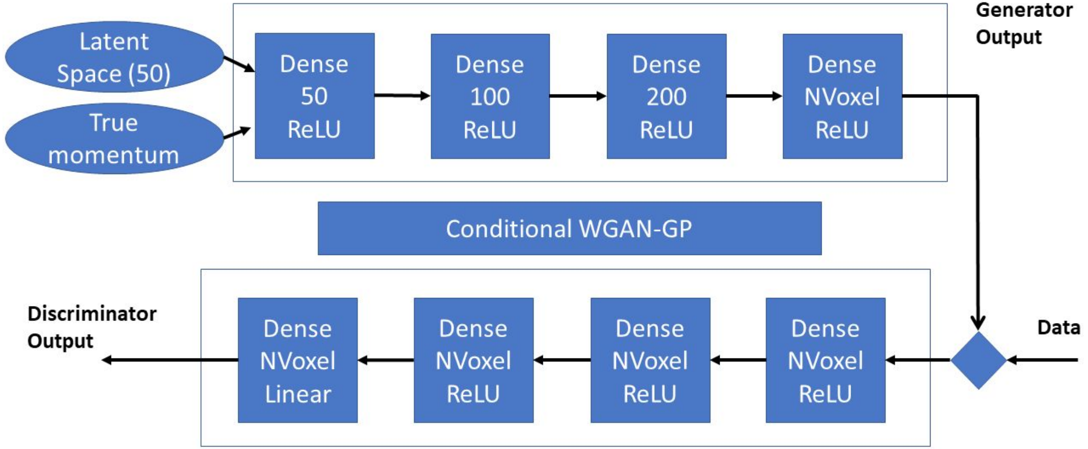


Subdetector CPU fraction for 50 ttbar events
MC16 Candidate Release

Wall clock consumption per workflow

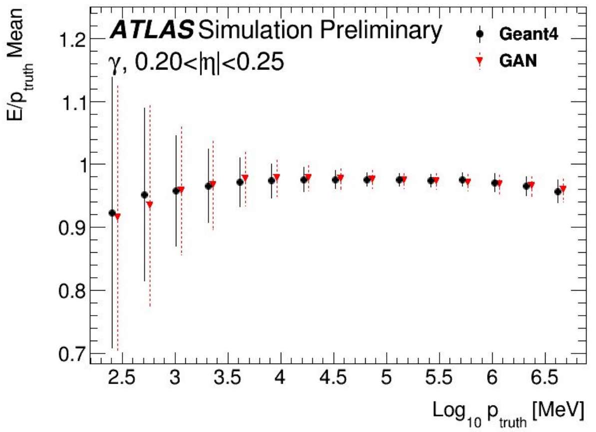


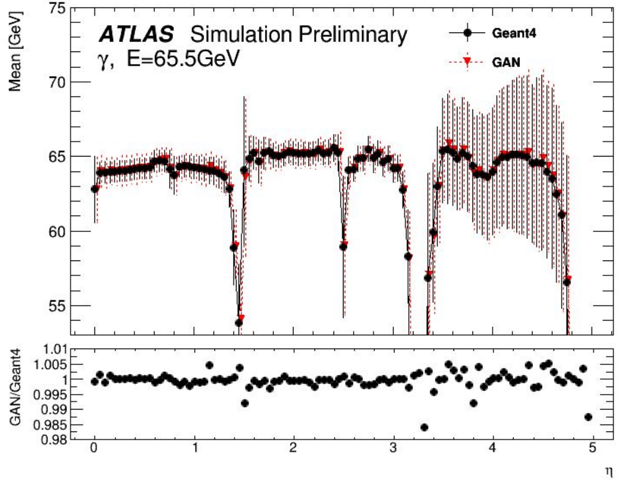
ATLAS FastCaloGan



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

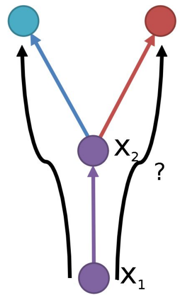
Fast simulation of the ATLAS calorimeter system with Generative Adversarial Networks, ATL-SOFT-PUB-2020-006 (2020)

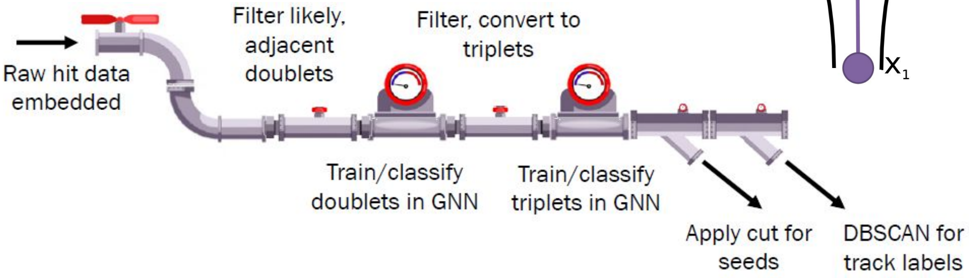


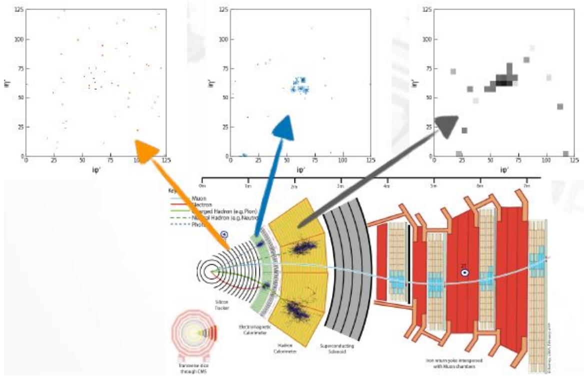


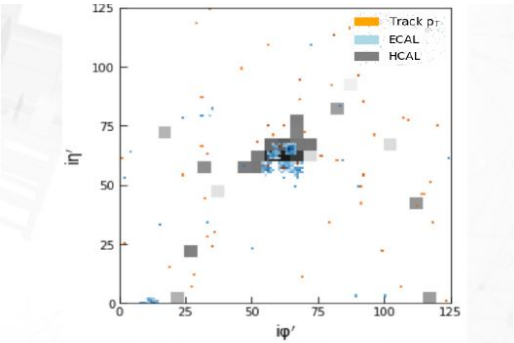
New Trends in Reco

- 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





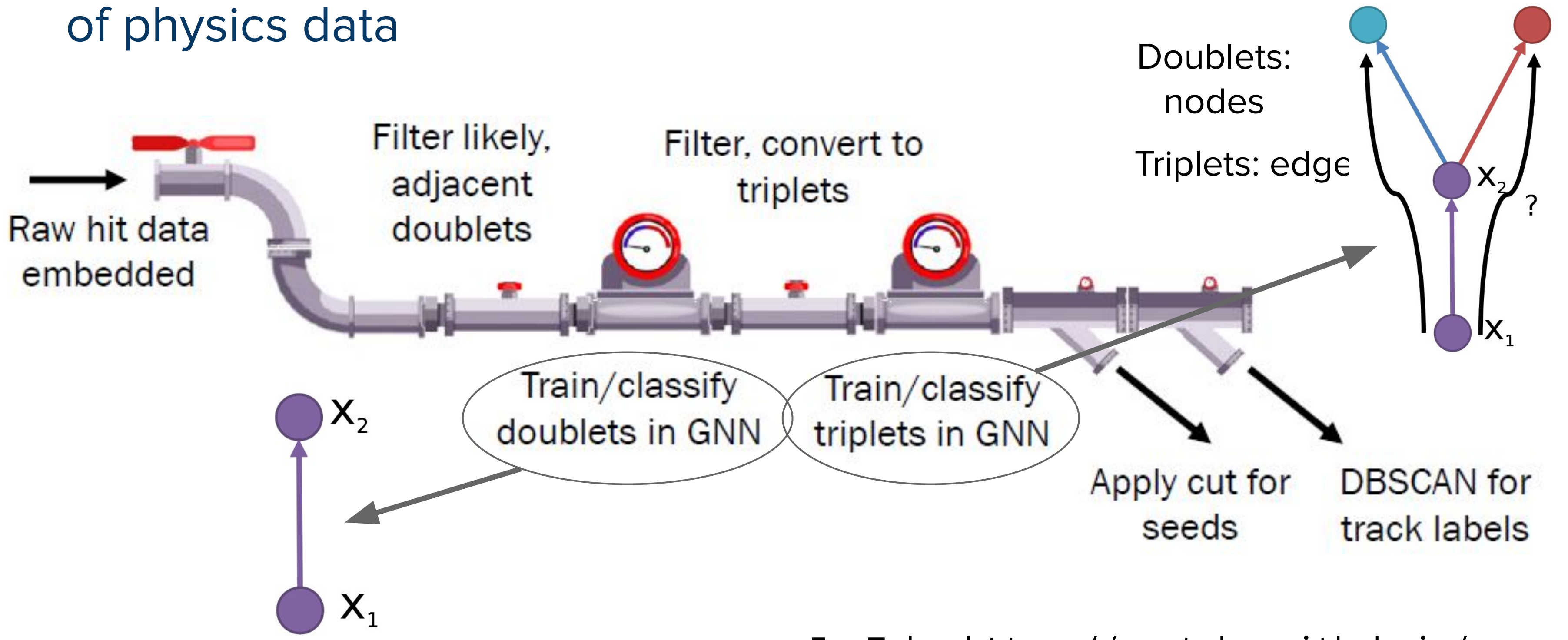




Gluon-jet image → 3 channels: track p_T, ECAL & HCAL

GNN for Tracking

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



ExaTrkx https://exatrkx.github.io/

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)