Machine Learning applications at IN2P3

ANF Machine Learning - 21-25 Septembre 2020, Orsay

Julien Donini
LPC / Université Clermont Auvergne

Introduction

ML has been present since a long time in many research fields of the **IN2P3**

Since a few years **small revolution** with **modern** ML librairies and infrastructure

Access to ML more ‘democratic’ and widespread than before

A **lot of work** in this field: here just a **few** hand-picked results

To have a more **complete overview** of ML activities in the past year

- Prospectives Machine Learning de l’IN2P3 (oct. 2019)
- Journées Machine Learning et Physique Nucléaire (oct. 2019)
- IN2P3/IRFU Machine Learning workshop (jan. 2020)

Outline

Overview of ML activities @ IN2P3

- Detector / accelerator
- ML for HEP analyses
- Nuclear physics
- Astrophysics

Training and schools

Conclusion

Detectors / accelerators

Detector / accelerator

Detector design

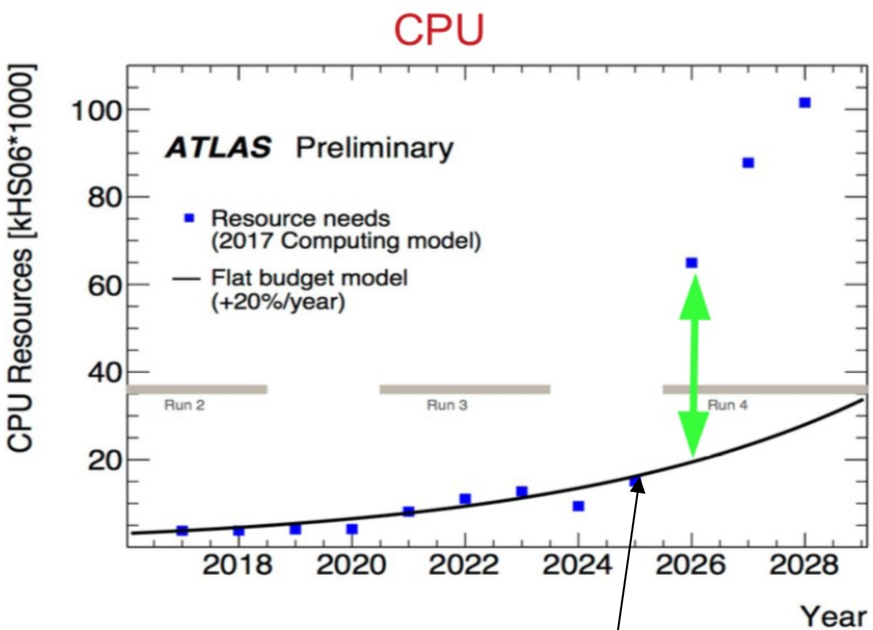
ML for Accelerator developments

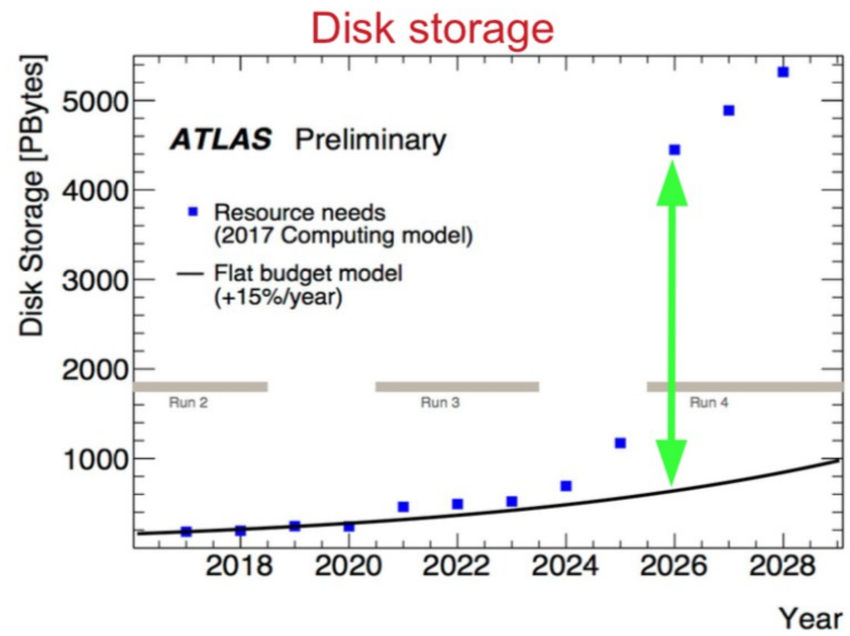
- ML for ThomX **experiment**

Simulation

- Simulation of ATLAS **calorimeter** with GAN’s

Fast simulation for High-Lumi LHC



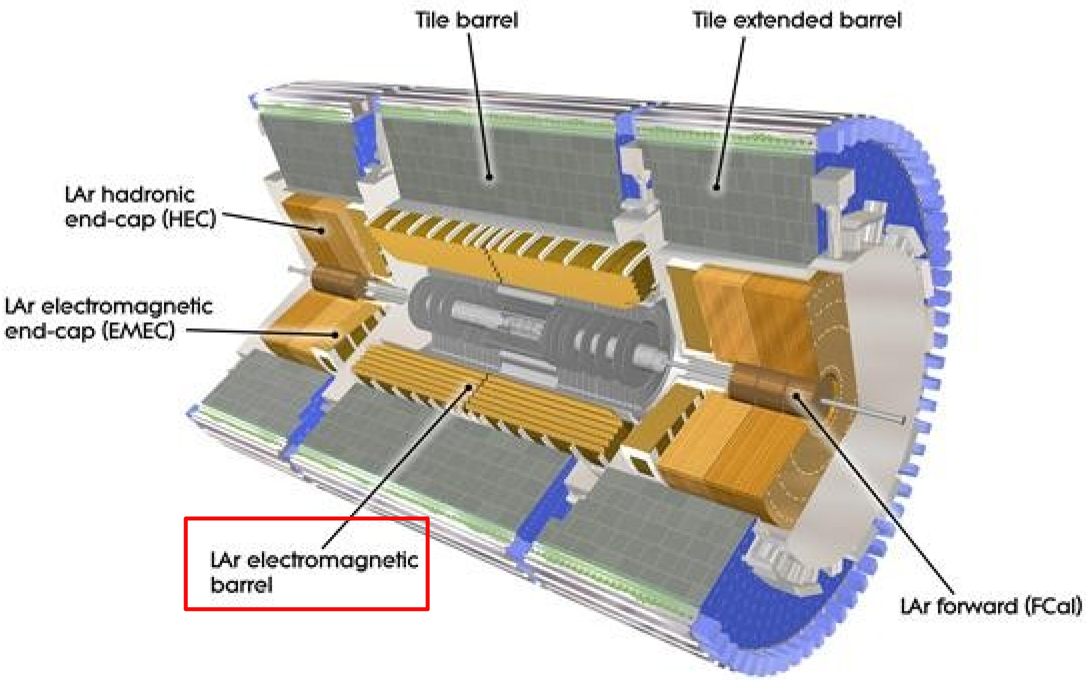


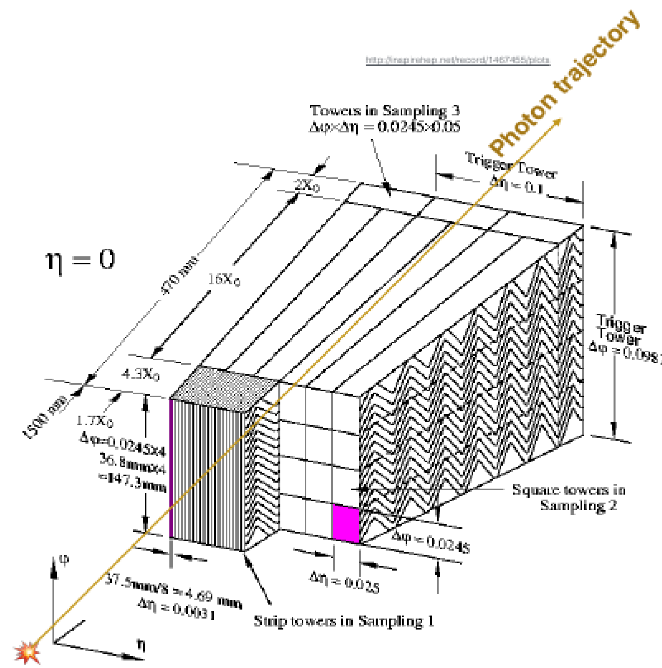
Dominated by : calorimeter simulation and tracking

ML used to design fast simulation algorithms

GAN for simulation for ATLAS

Simulation of liquid-argon electromagnetic calorimeter response with GAN's





Particle goes through 4 layers :

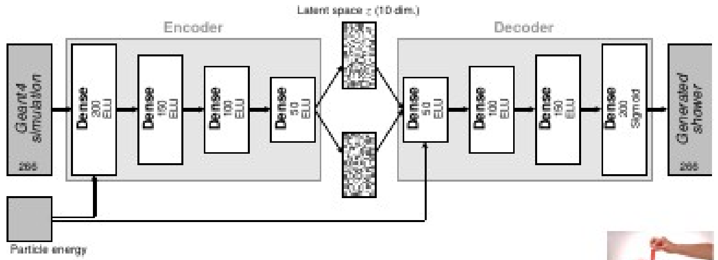
0. Pre-Sampler : (7x3) Some energy deposit
1. Strips: (56x3) Very granular in η; more energy deposit
2. Middle: (7x7) Thickest layer, maximum energy deposit
3. Back: (4x7) Little Energy deposits

First results (2018)

D. Rousseau, A. Ghosh (LAL), G. Louppe (U Liège) et al.

ATL-SOFT-PUB-2018-001 and update ATLAS-SIM-2019-004

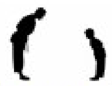
VAE:



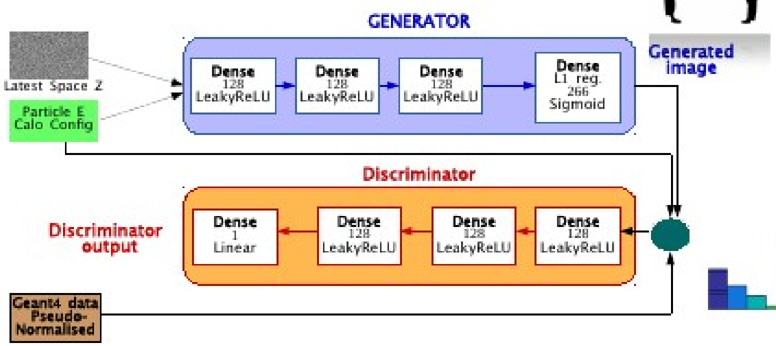
100 epochs, 2 mins, CPU

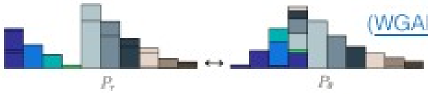
Flat vector of 266 cells are the output of both generators

50k 'epochs', 7 hours training, 1 GPU



GAN:





Keras TensorFlow

Training dataset:

- Single **photon** samples from Geant4
- 88000 events
- 9 **discrete energy points** : {1, 2, 4, 8, 16, 32, 65, 131, 262} GeV
- $0.20 < |\eta| < 0.25$
- 4 electromagnetic calorimeter layers

Not an ideal training dataset

Data preprocessing

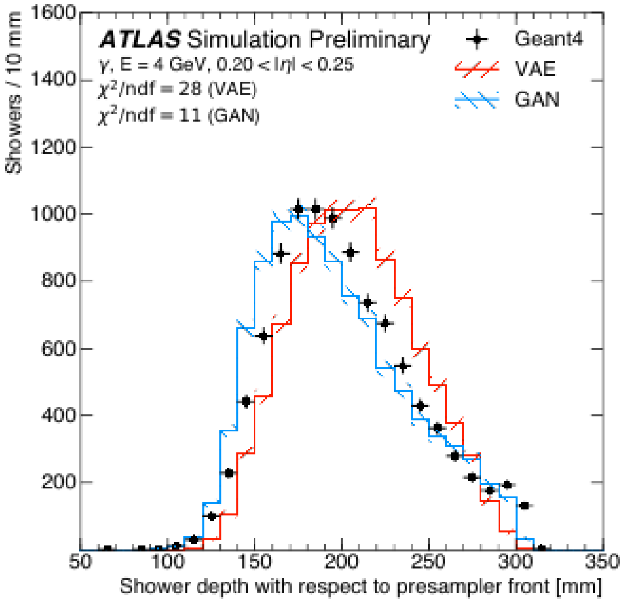
- Negative energies set to 0
- Mirror $\eta < 0$

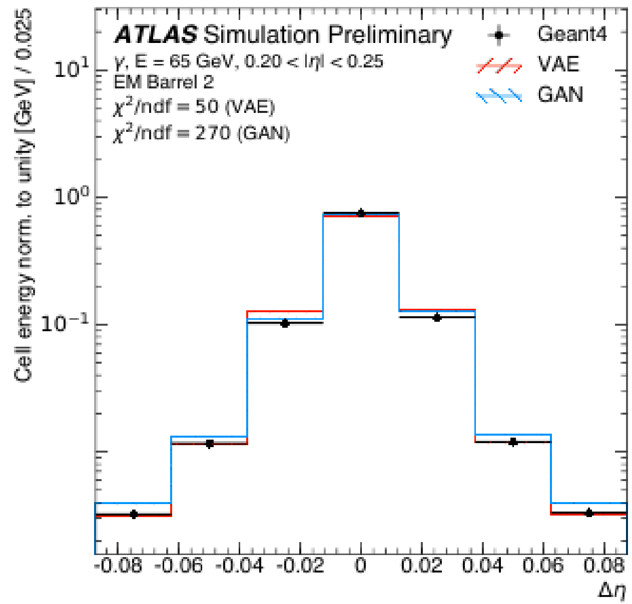
(WGAN-GP, Improved WGAN-GP nightmare on Ke...

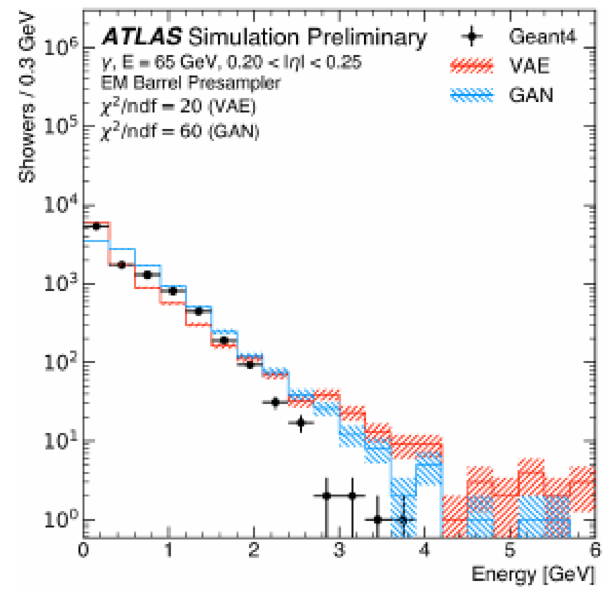
Slide taken from Aishik Ghosh talk 01/20

First results (2018)

[ATL-SOFT-PUB-2018-001](#) and update [ATLAS-SIM-2019-004](#)



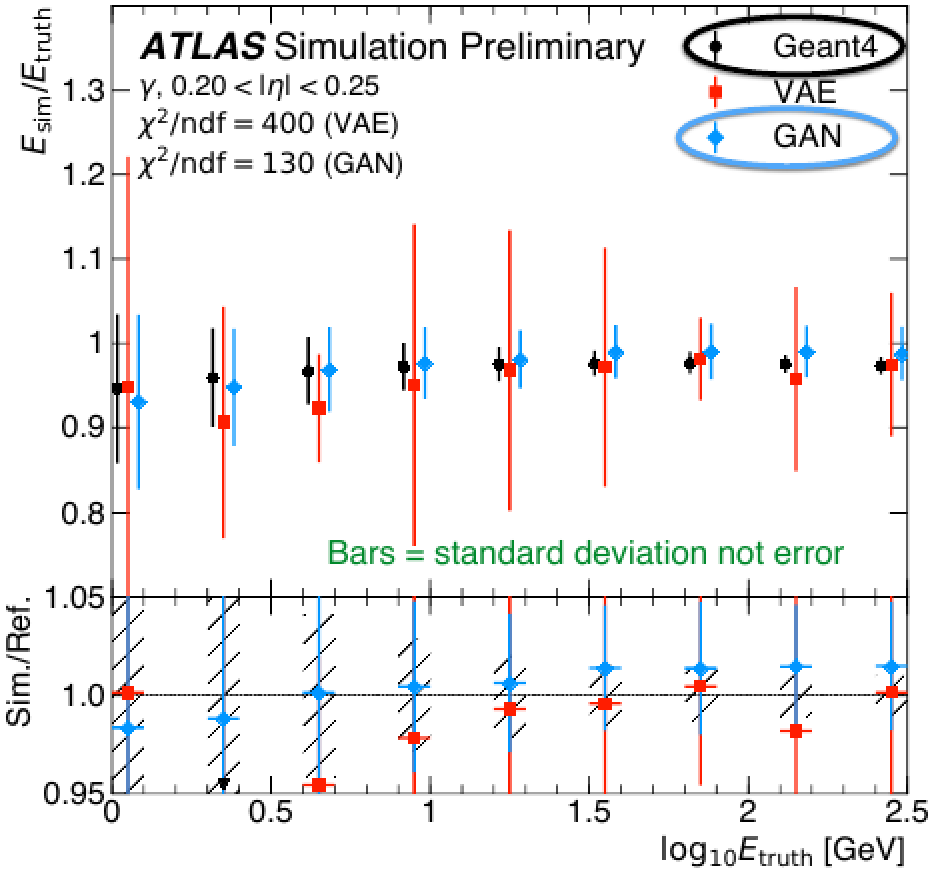




- < 1 ms instead of ~10 s to reconstruct an object
- x100 gain for a full event
- But some limitations: energy resolution, etc

First results (2018)

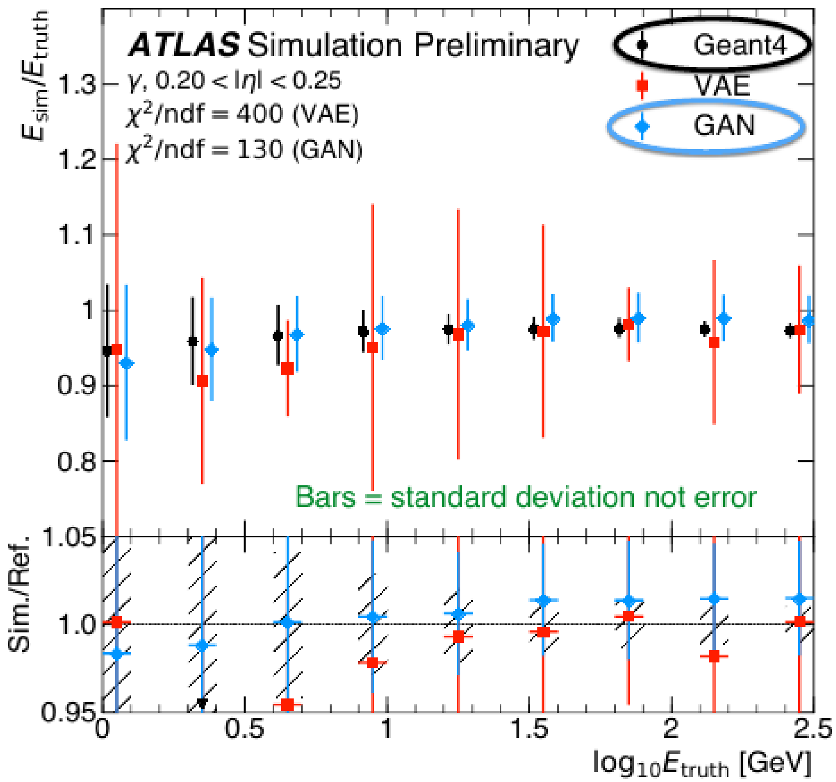
Problem: cannot model well detector resolution

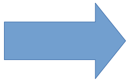


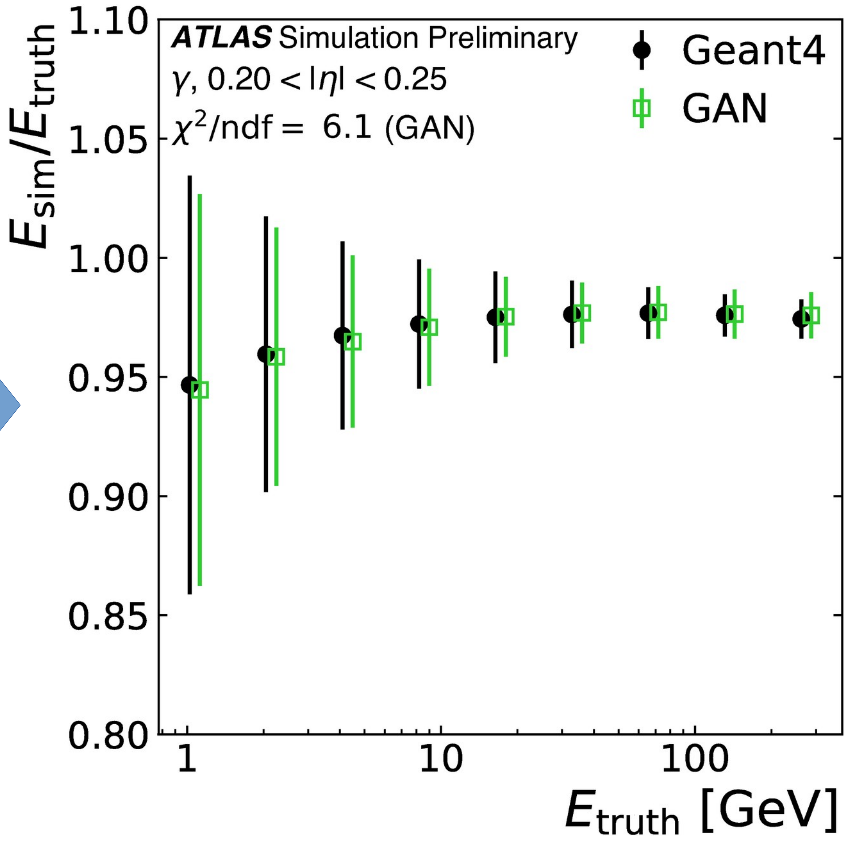
GAN gets the means but not the widths of the energies

Improvements

New GAN architecture + conditioning (energy, position, geometry)
→ improved energy resolution, particle position







Impressive progresses but probably still a long road to go

(More details here)

Machine learning for accelerators

OWLE: The **O**ne **W**orld particL**e** acc**E**lerator colloquium and seminars

General trend in ML for accelerators

Recently people from **PSI** (SLAC, DESY,CERN, MIT) started series of online seminars for physics of accelerators, and in particular **ML for accelerators**.

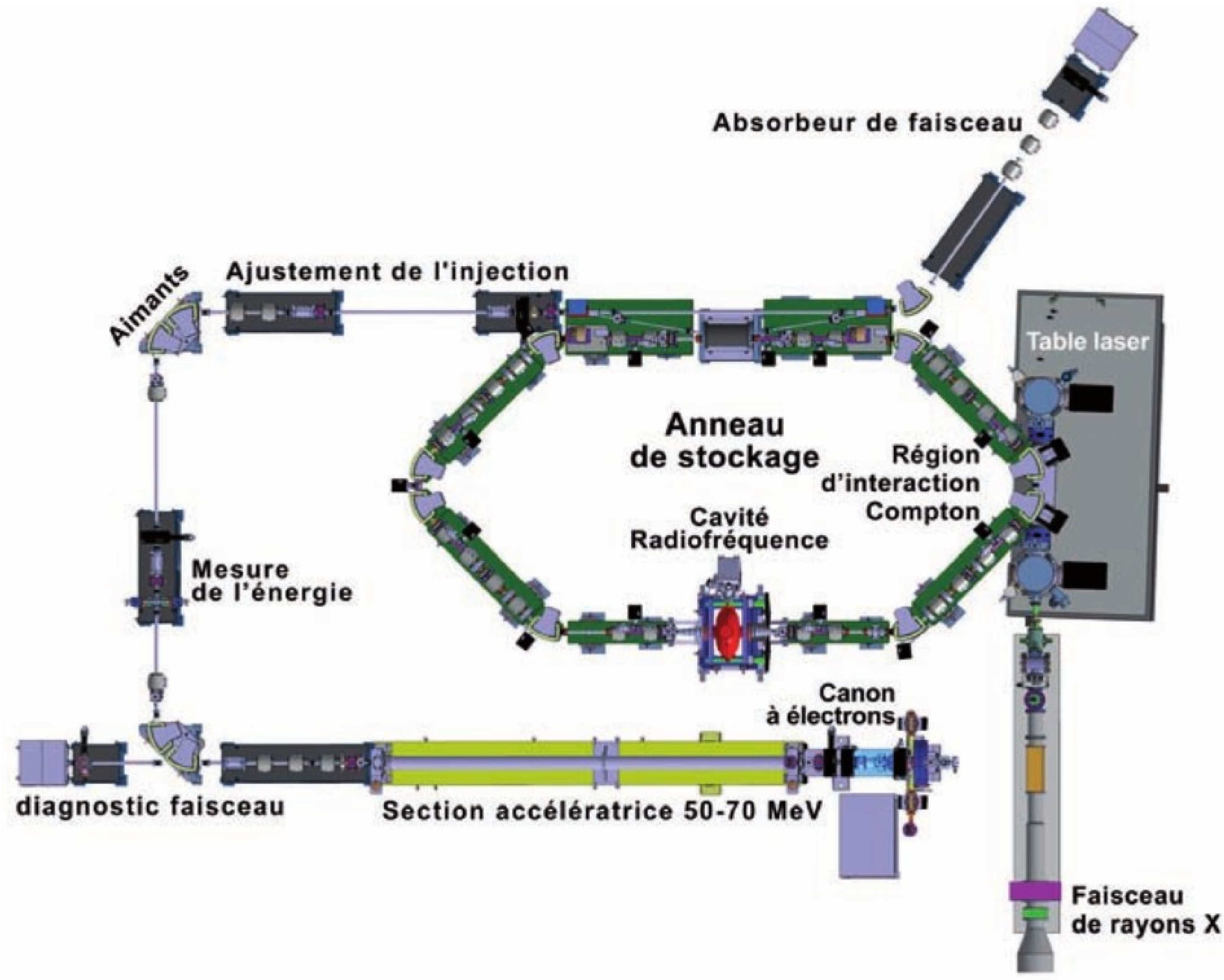
The **OWLE-Colloquium** is aimed at giving researchers a platform to share research and development results of very broad interest.

The **OWLE-ML seminar series** has a topical focus on machine learning and **experimental demonstration of AI-ML**.

https://sites.google.com/view/owle/home

Machine learning for accelerators

The ThomX project: high intensity and energy X-ray source produced by compton interaction of photons (laser) and electron (accelerator ring)

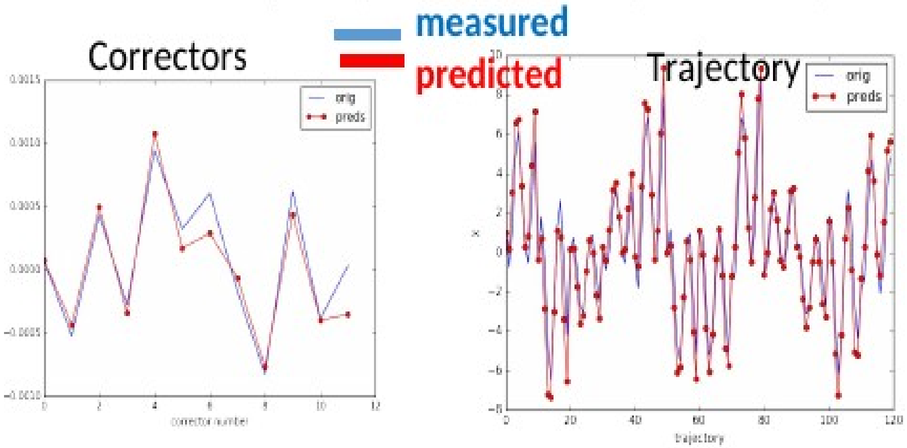


Supervised learning for accelerators

H. Guler, V. Kubytskyi et al. (IJCLAB)

ThomX RING : Single particle Trajectory (several turns)

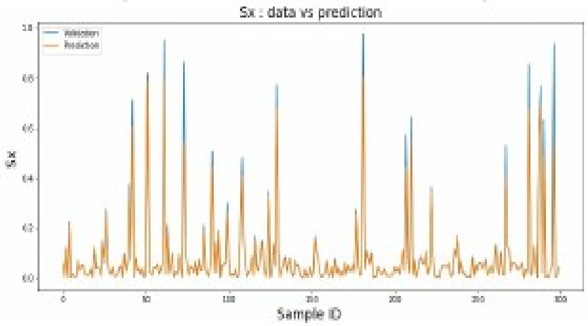
1. Control parameter: Corrector magnets. 12 independent variables in transverse horizontal/vertical planes.
2. Measured: trajectory/n-turns/orbit represented by 120 variables (12 BMPs x 10 Turns).



Model trained on simulation to predict correctors based on the trajectory input.
XGBRegressor + MultiOutputRegressor, or NN

ThomX LINAC : Reproduce beam longitudinal dynamics from simulation

1. Control parameter: Solenoid, RF phases, laser parameters (10 parameters)
2. Measured beam parameters (size, emittance, …) : 6 observables
3. Retrieve parameters from beam profile (CNN)

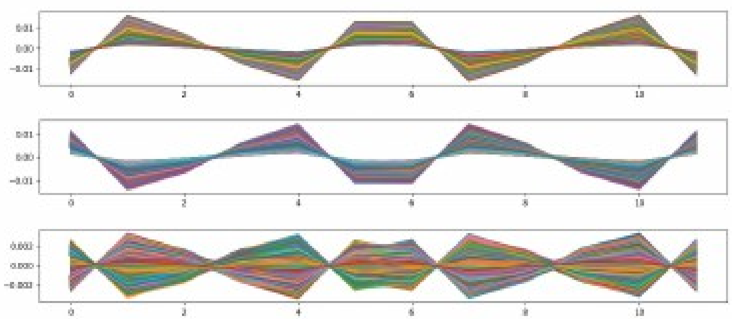


- Model trained on simulation (slow simulation 10 min per configuration)
- Neuronal network model for scalar data
- CNN for images data

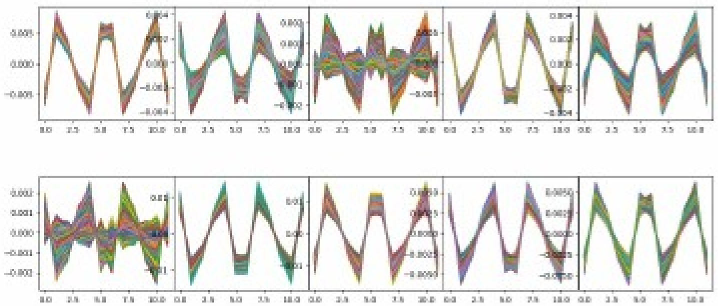
Unsupervised and reinforcement learning

Classification of trajectories with K-Means.
Same dataset is easily regrouped to few categories/classes.

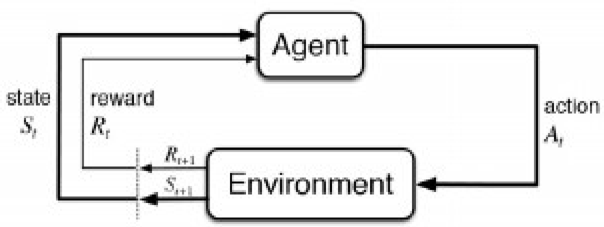
3 categories



10 categories



Reinforcement Learning using model based on NN
Find special beam characteristics (example : minimum size, emittance etc)



OpenAI Gym environment used together with **Stable Baseline** and Tensorflow.

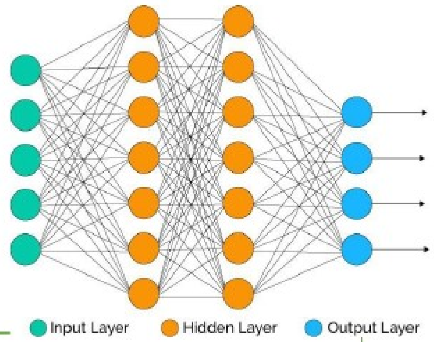
Different models benchmarked (DQN, DDQN, **TD3**, ...) with proper policy.

Could find beam minimum size after less than 20 epocs

Machine learning for accelerators

Typical workflow

1. Formulation of the problem
2. Preparing dataset and "understanding of the data"
3. ML model: development, training, improvement



Ongoing studies:
- prediction of correctors magnets currents
- Trajectory minimization
- Noise "filtering" in the data
- Model robustness
- Beta function reconstruction from TBT data
- Optimization of injection
- Orbit classification
- Failure/anomaly detection
- Inverse problems

Methods:
NN, CNN, XGBoost
RL

Tools:
Jupyter notebook, Keras, Tensorflow, PyTorch, OpenAI

Hardware:
CPU (x48), GPU(GV100, Laptop)

When NN is not learning, search why :
- Dataset, more datapoints in trajectory
- NN architecture: layers, depth, activation
- learning rate
- Normalisation
- Optimizer
- Add noise

ML for HEP analyses

ML for HEP analyses

Historically a vast playground for ML approaches – many IN2P3 contributions

Object reconstruction, particle identification, calibration

Event classification, regression

Phenomenology and theory

Real time analysis and triggering

Treatment of uncertainties

Data reduction

Search for anomalies

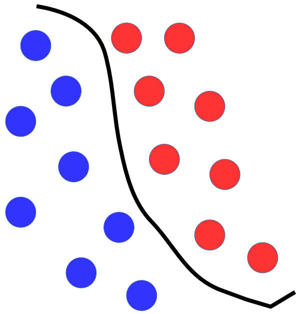
- Searches for new physics at LHC

Fundamental parameter inference

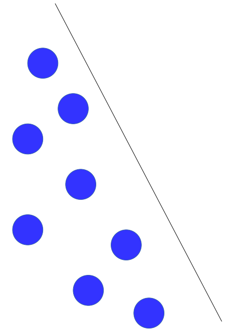
- Likelihood free inference

Anomaly detection

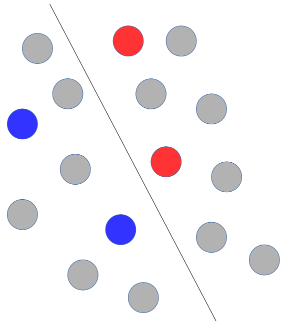
Supervised (labels)
DNN, BDT, SVM



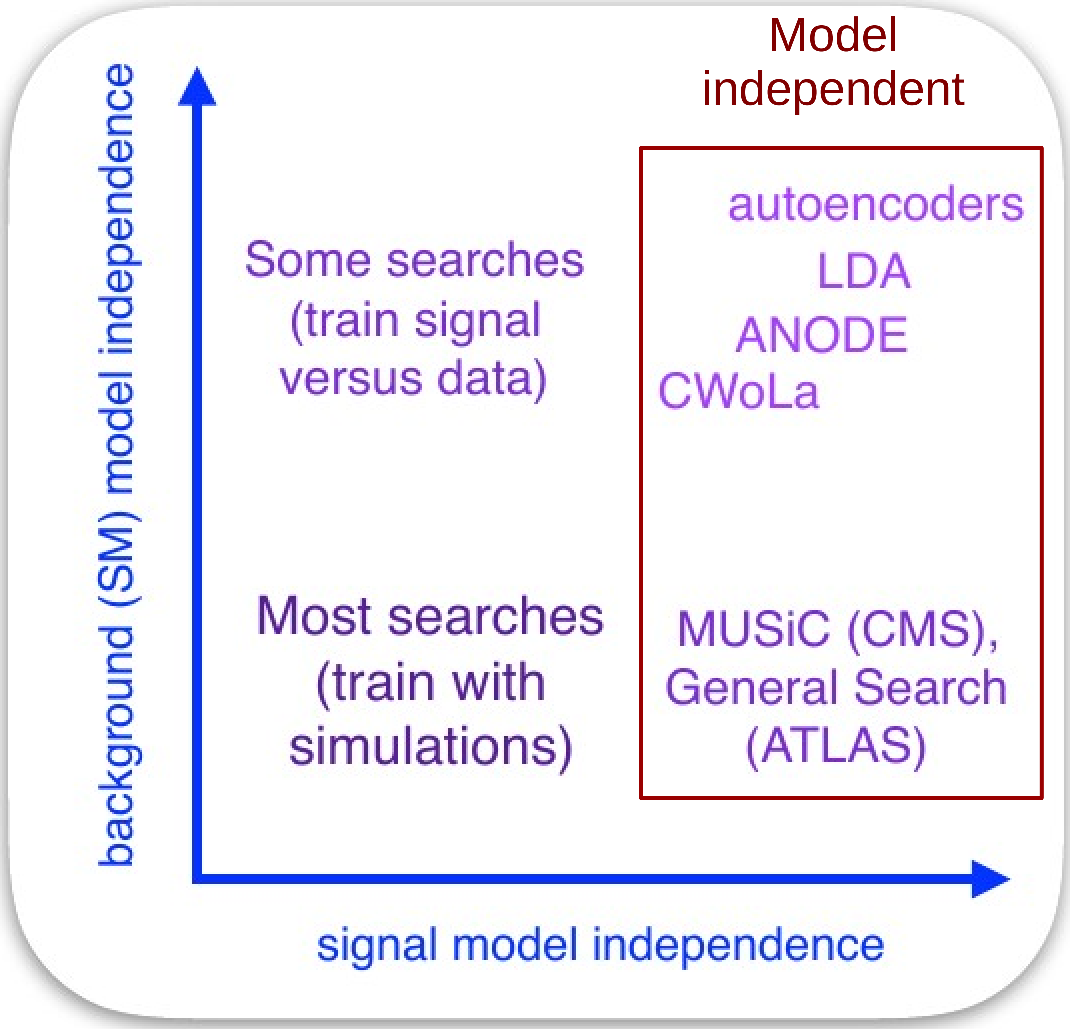
Unsupervised
(no labels)
SVM-1class, **AE**, VAE,
WAE, **GAN-AE**,...



Semi-supervised
(some labels)
triplet NN,...



Searches for New Physics at LHC



B. Nachman, D. Shih, arxiv:2001.04990

LHC Olympics challenge

Anomaly detection challenge using simulated data

Despite an impressive and extensive effort by the LHC collaborations, there is currently no convincing evidence for new particles produced in high-energy collisions. Goal is to ensure that the LHC search program is sufficiently well-rounded to capture “all” rare and complex signals.

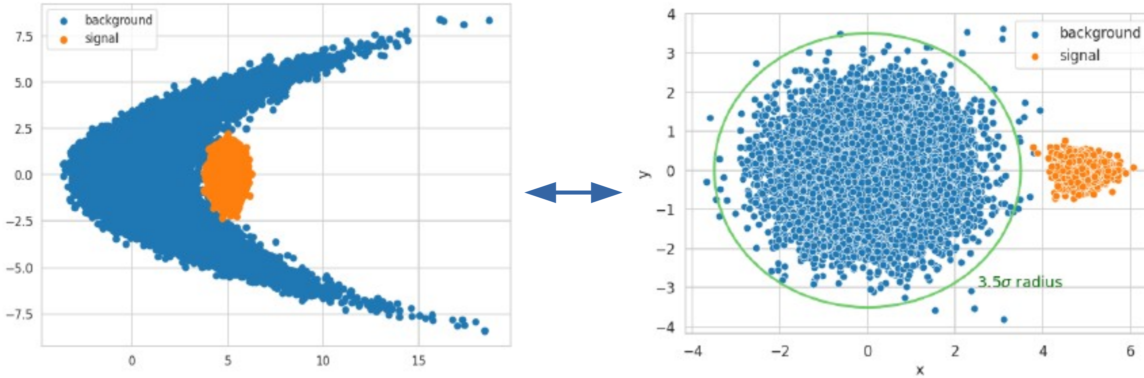
Two editions in 2020 (Winter and Summer): https://lhco2020.github.io/homepage/

Event-level anomaly detection methods

L. Vaslin, I. Dinu, J. Donini (LPC Clermont)

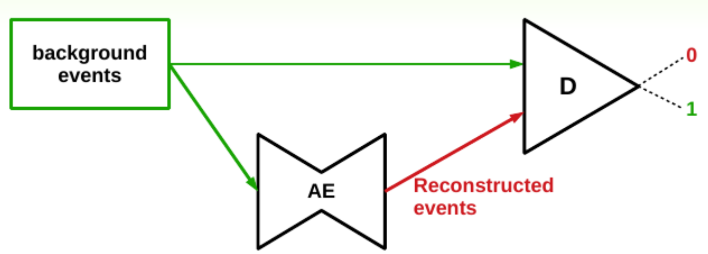
1. Normalizing flow for anomaly detection

Determine bijective transformations between background data and multivariate Gaussian



2. Autoencoders and generative models

Increase performances of AE using a GAN-like architecture

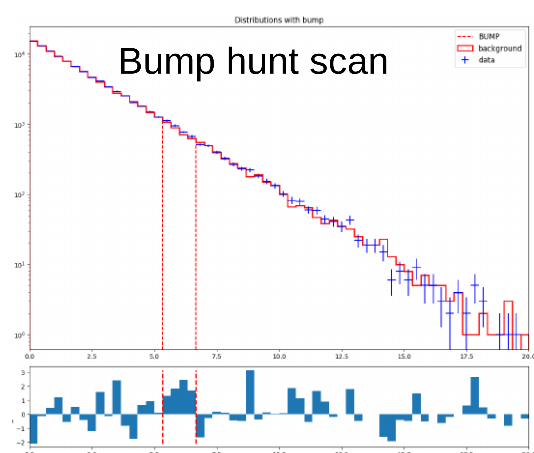


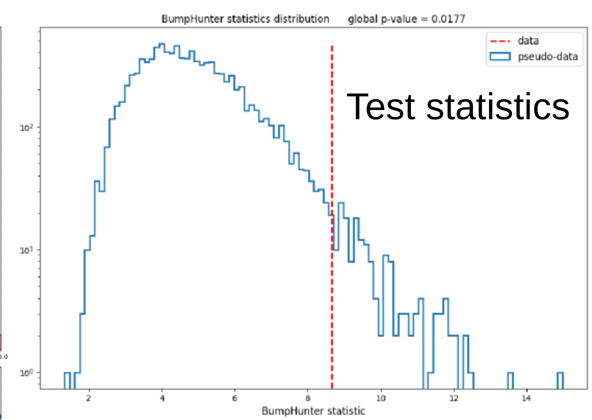
3. pyBumpHunter project

Python implementation of the popular BumpHunter algorithm

Code here: pyBumpHunter

Implementation in scikit-hep

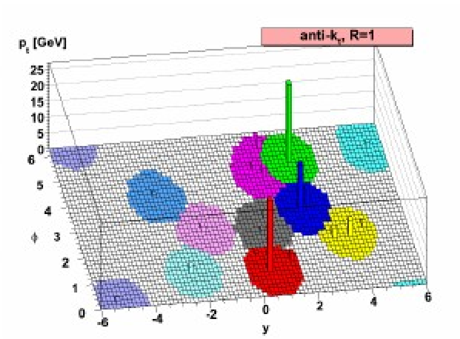




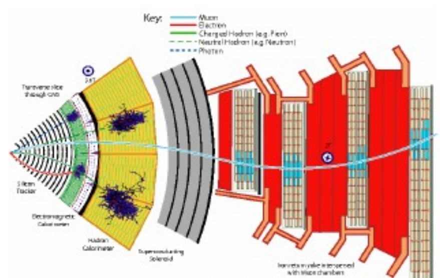
Likelihood free inference

Latent variables

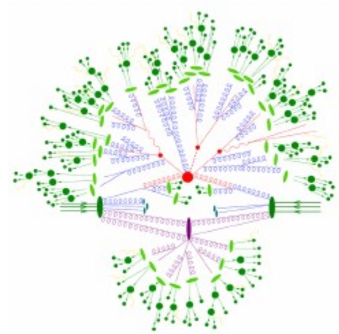
Observables		Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
x	$\leftarrow$	$z_d \leftarrow$	$z_s \leftarrow$	$z_p \leftarrow$	θ



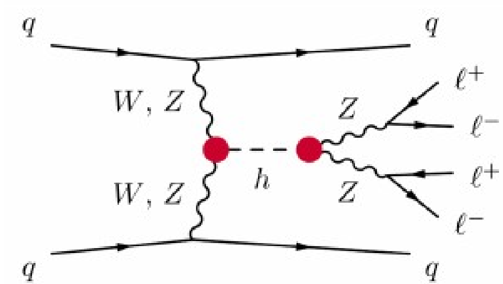
[M. Cacciari, G. Salam, G. Soyez 0802.1189]



[CMS]



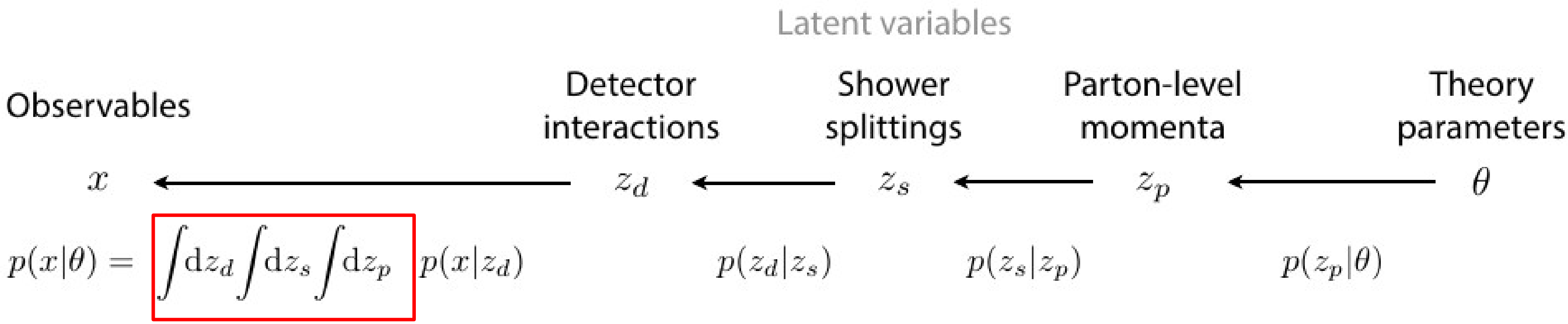
[F. Krauss]



Evolution

For details see J. Brehmer slides

Likelihood free inference

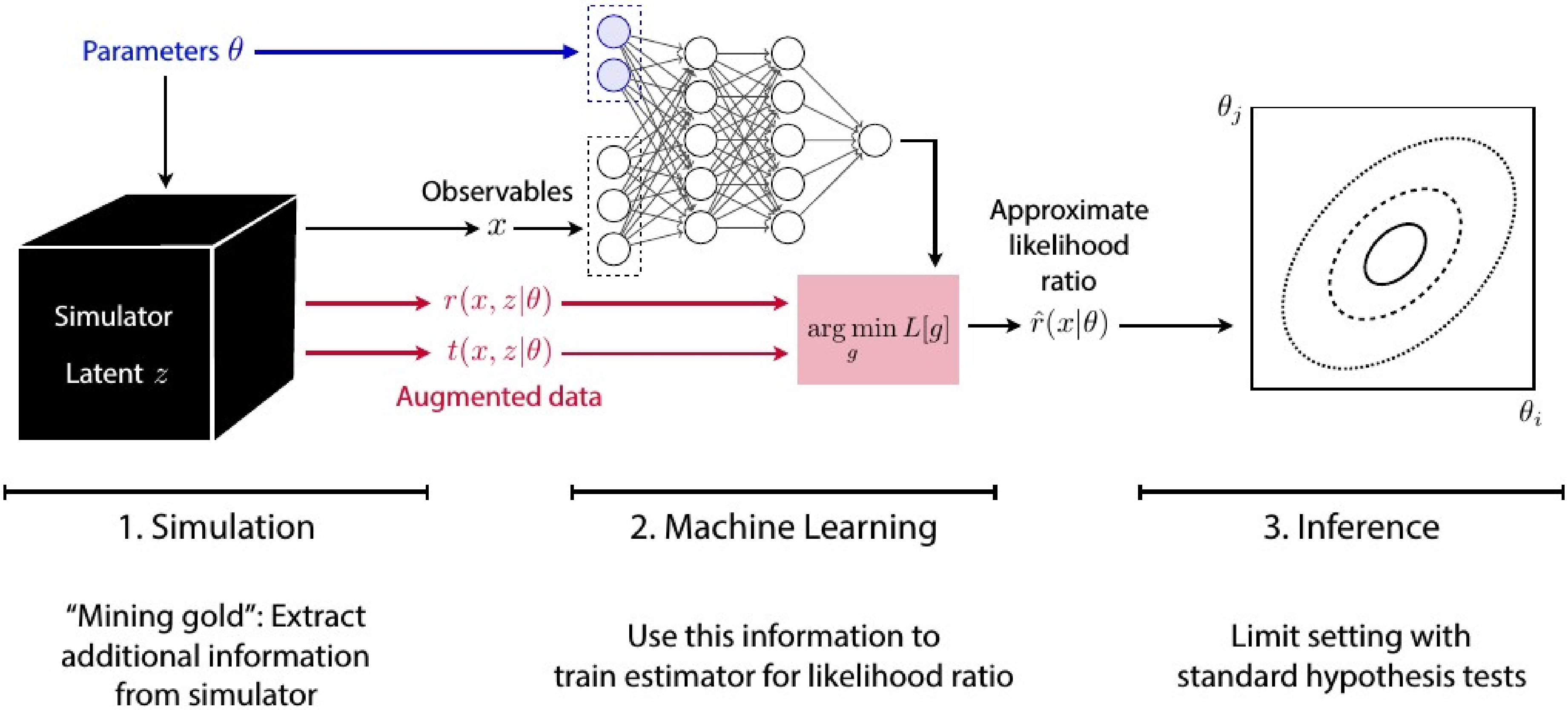


Impossible to calculate integral over enormous space

→ analysis at LHC generally rely on other approches (collect data in form of histograms, etc)

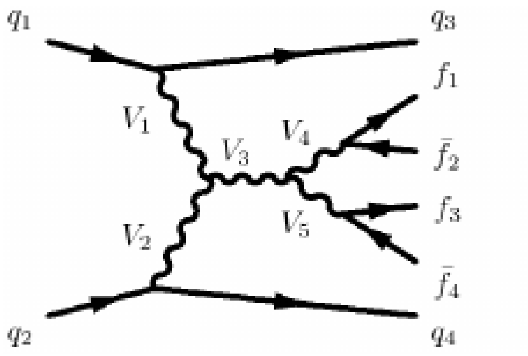
Likelihood free inference

Approach: use more informative targets to regress for a neural network

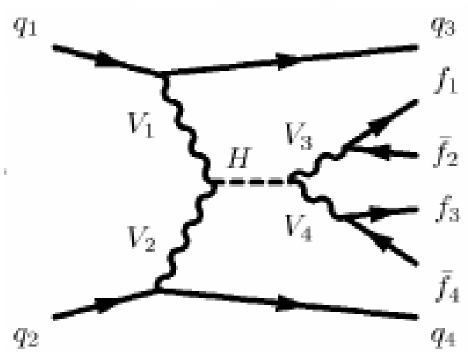


Likelihood free inference

Full example using this approach : Measuring Quantum Interference in the Off-shell Higgs to 4 Leptons (see A. Ghosh presentation)



Vector Boson Propagator
(Background)



Higgs Propagator
(Signal)

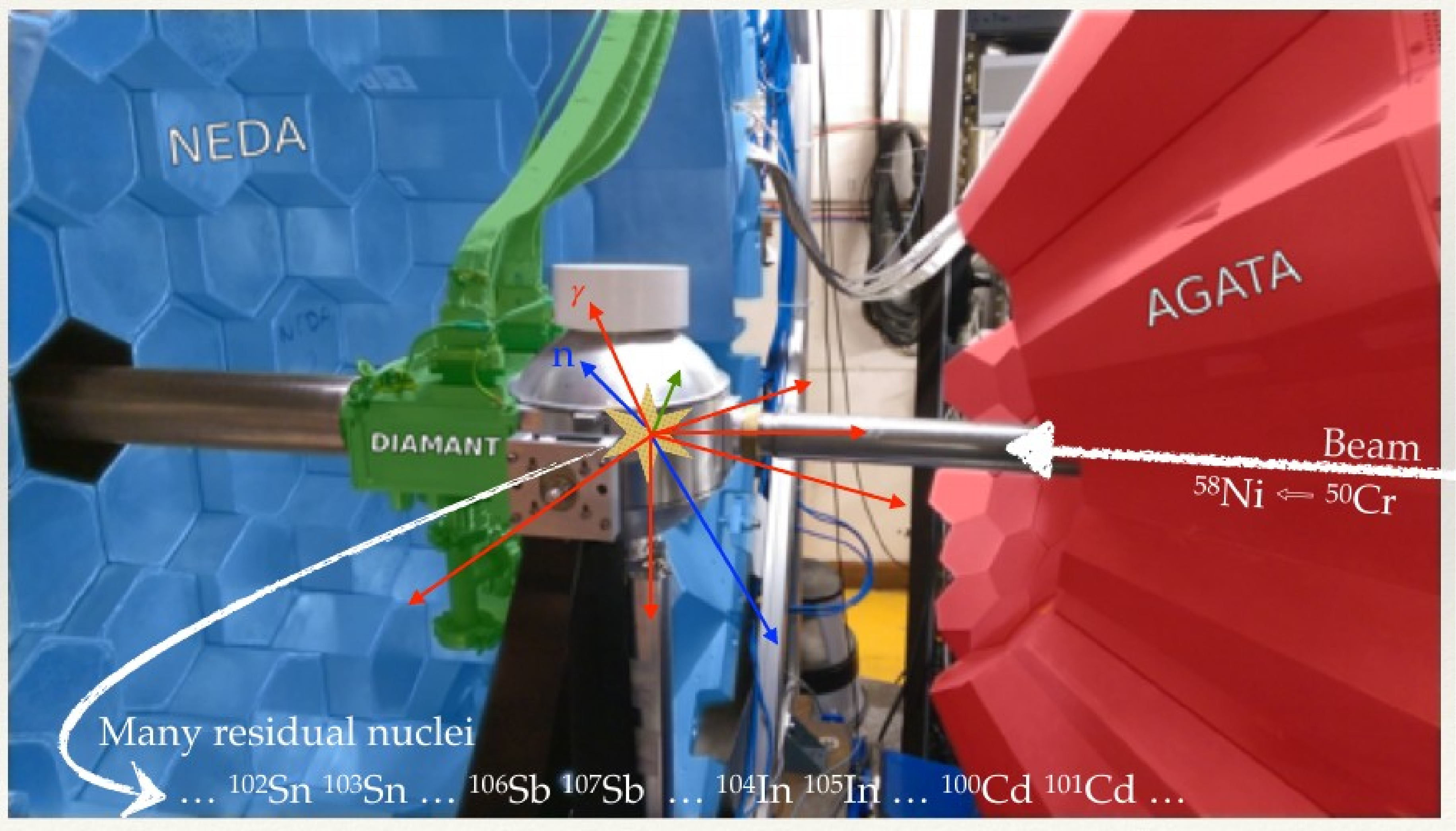
Aim: directly learn the likelihood using ML – results seem promizing

Nuclear physics

Pulse shape discrimination with NEDA

O. Stezowski et al. (IP2I)

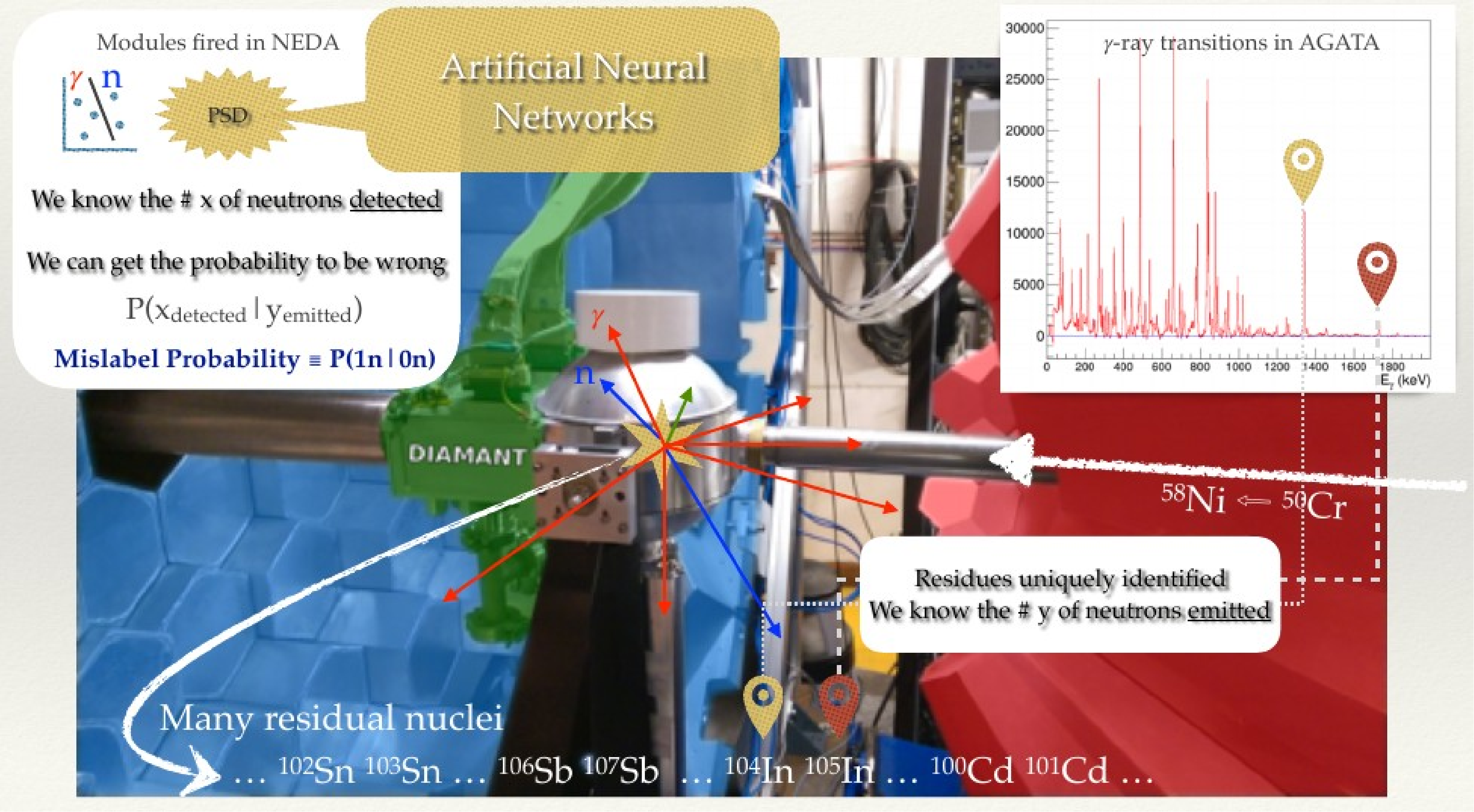
Data from an experiment AGATA + NEDA + DIAMANT in coincidence [GANIL 2018]



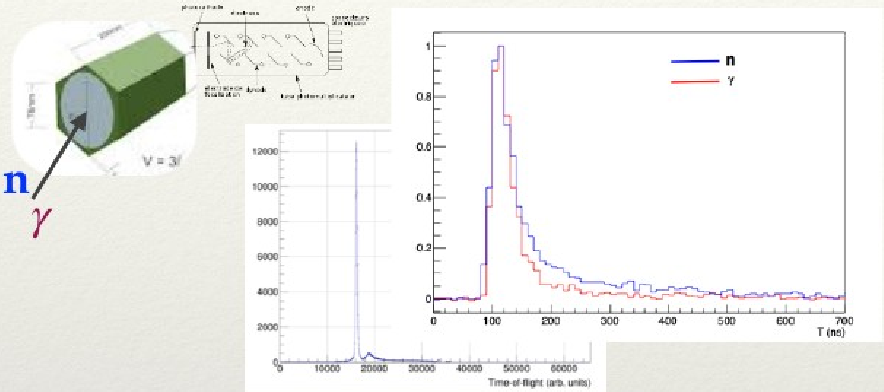
(Slides O. Stezowski)

Pulse shape discrimination with NEDA

Data from an experiment AGATA + NEDA + DIAMANT in coincidence [GANIL 2018]



Pulse shape discrimination with NEDA



Inputs used for the **Discrimination :**
the waveform - the amplitude - the time of flight

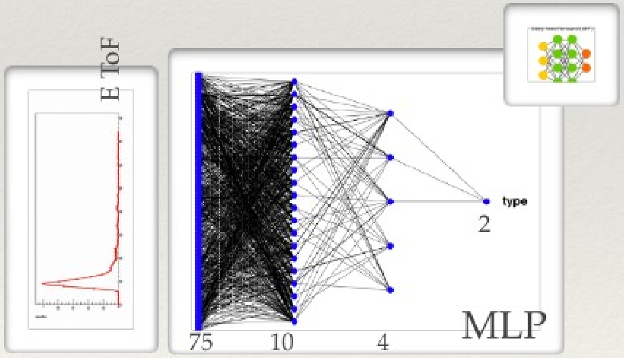
Common parametrisation of the signal

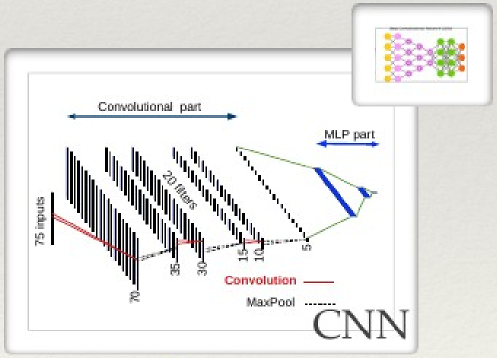
$$s(t) = A\,[\exp(-t/td1) - \exp(-t/tr) + R^{*}(\exp(-t/td2) - \exp(-t/tr))] \text{ if } t > T0$$

A amplitude = energy
td1, **td2**, **tr** independent of γ and **n**

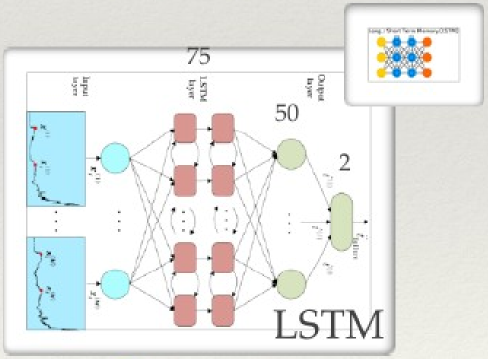
T0 relies on how the signal is captured
R depends of the type of the particle

Three different Artificial Neural Network architectures tested : MLP / LSTM / CNN





👍 pattern



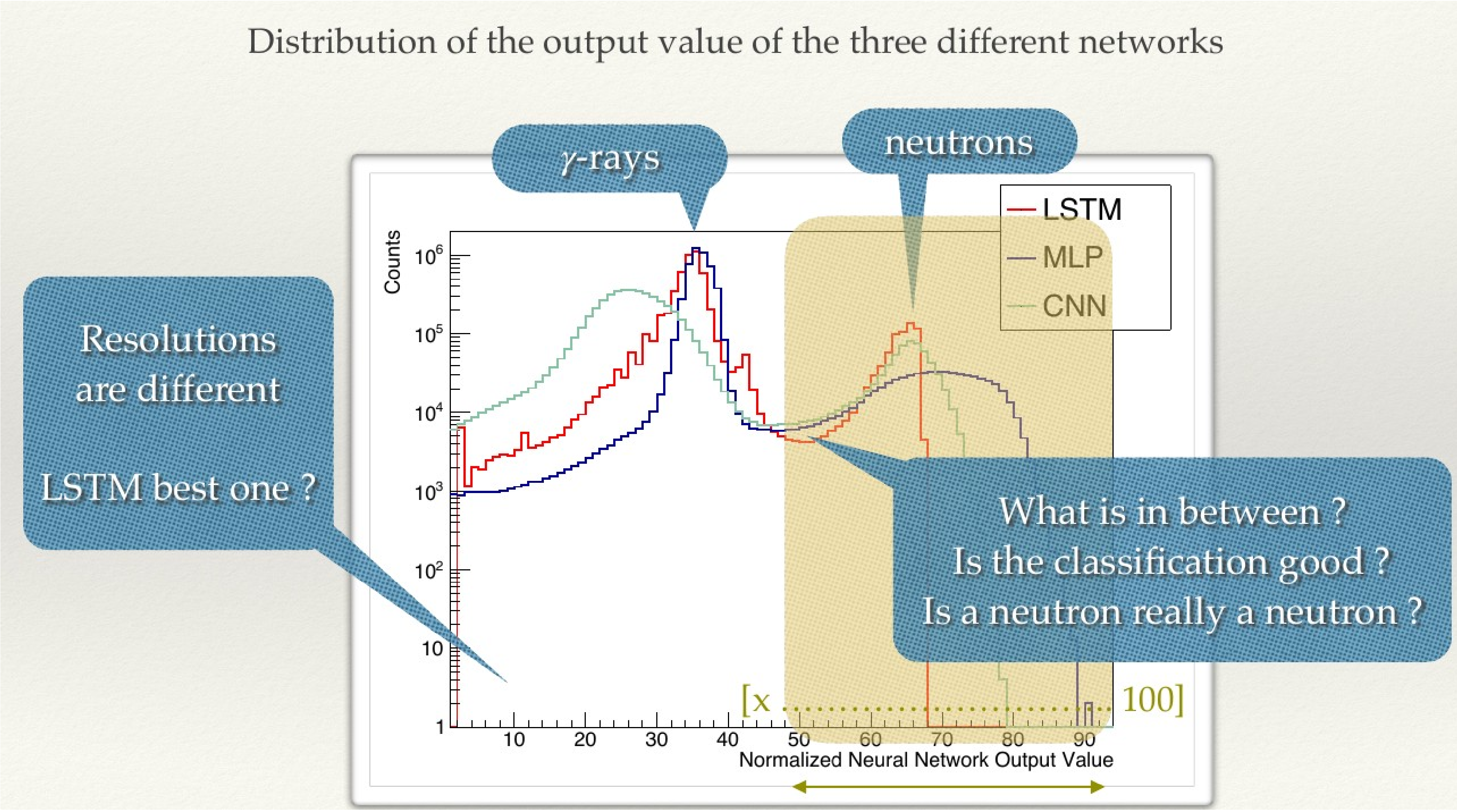
👍 time series

R&D NEDA
discrimination for low energy better that classical methods *
Implementation with ROOT - mono thread / CPU
➥ Tensorflow / multi CPU / GPU

* Ronchi et al., NIMA 610 (2009) 534–539

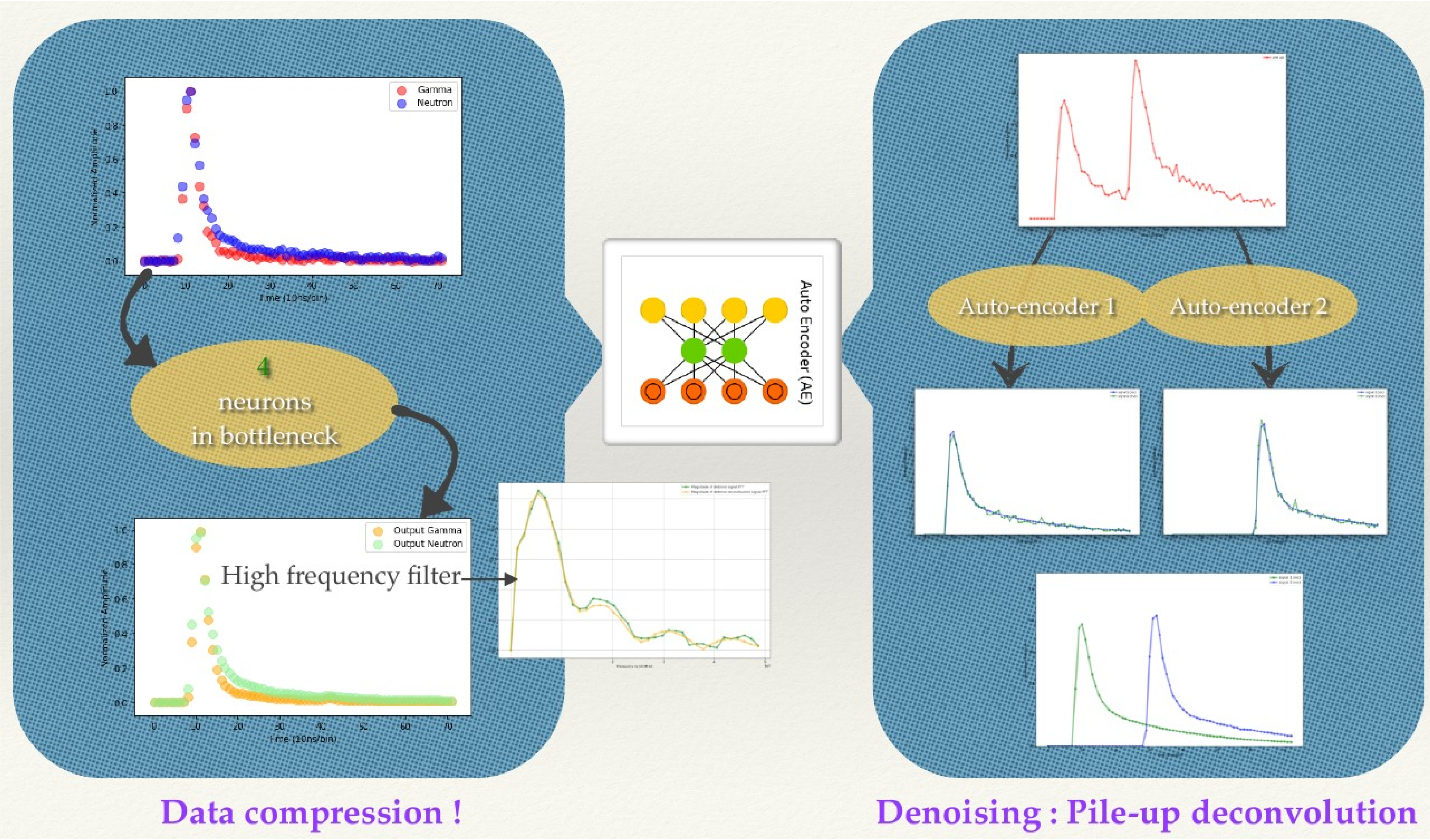
Number of parameters
MLP: 814, LSTM: 10502, CNN: 7042

Pulse shape discrimination with NEDA



First steps in using Machine Learning for data processing, 3 ANN architectures studied

Signal and autoencoders



Auto encoders into the game for compression / de-noising

Astrophysics

Astrophysics

Deep Learning

- Image analysis:
  - Characterization of gamma-ray events in CTA - LAPP
  - Photometry of blended galaxies with deep learning - APC
  - Photometric redshift estimation - CPPM
  - *Identification of tumors through real time imaging - IMNC*

- Events analysis:
  - Single detector glitches and signal identification in VIRGO - APC
  - Waveform reconstruction and characterization in LISA- APC
  - Classification of time-series from astronomical transients - LPC, CPPM

- Signal separation:
  - Generate pure EE/BB power spectra from CMB - APC
  - Deblending of galaxies with VAEs - APC
  - Galaxy signal / noise separation - CPPM

For a complete review see talk E. Ishida (journées prospectives IN2P3): here

Training and schools

Collaborations with CS/maths

ML collaborations @ IN2P3

- Common project, co-supervision of PhD, post-doc
- Example of **(past) local** collaborations :
  - **LPC** and LIMOS/ISIMA (CS), LMBP (maths)
    - LSST (astronomical time series), ATLAS (anomaly detection), LHCb (bayesian learning)
  - **LPNHE** and Sorbonne (maths): ATLAS (fuzzy number systems)
  - **LAL** and LRI (CS): ATLAS (TrackML, Syst. Aware Training)
  - **CPPM** and LIS (CS): ATLAS (ttH), Cosmology (deep learning)
  - **LAPP** and LISTIC (CS): CTA (deep learning)
  - …
- **International** collaborations: EU-projects with non-academics partners, ...

Obvious advantage in collaborating with ML experts but some caveats:

- Speaking same **language** & getting familiar with vast stat **litterature**
- Question of access to **confidential** experimental data and **authorship**
- **Publication** in journal of CS/math field
- Produce outcome **relevant** to collaborator

Training and schools

Being able to apply ML to practical problems requires understanding underlying statistical concepts and ML algorithms.

- **Target**: students (Master, PhD), staff IN2P3

Training courses exist in several universities / labs

- In general Master degree level some also open to staff for continuous training
  - Ex: Diplome Universitaire Data Scientist
- Training CNRS formation entreprise
  - Ex: Introduction to ML and Deep learning

Schools / workshops

- IN2P3 School of Statistics (organized every 2 years since 2008)
- Workshop CCIN2P3: GPU and deep learning

Uncovered needs should trigger specific training actions

Diplôme Universitaire Data Scientist (UCA)

Formation de l'Université Clermont Auvergne – partenariat CCIN2P3

8 semaines de cours réparties sur l'année

Ouvert à la formation continue

1) Analyse données, **python** (25h)

2) **Statistiques** avancées (20h)

3) **Data Mining** (20h)

4) **Machine learning** (20h)

5) **Deep Learning** (20h)

6) Langage **R** (20h)

7) Séries **temporelles** (20h)

8) **Data engineering** (30h)
Centre calcul CNRS, Lyon

Toutes les informations disponibles sur le site de la formation

Conclusions

Usage of “traditional” **ML** since many years within IN2P3

Many resarch field at IN2P3 moved to **modern** ML approaches

Fast growing **expertise** on ML at IN2P3 but **training** is important

Huge research potential and many **opportunities**

Continuity and **support** is essential to maintain activities

Challenge: scalability, optimization, integration to experimental software

BACKUP

ML @ IN2P3

1. **Detectors & accelerators**
2. **Simulation**

Detector design

- Use ML to optimize detector design (LPNHE)

Color code

Advanced
Studies
Interest

ML for Accelerator developments

- **Accelerator** tuning, lasers, virtual detectors (LAL)
- NN for particle **accelerator** operations and optimization (LPSC)

Simulation

- Simulation of ATLAS **calorimeter** with GAN's (LAL)
- MC sample **reweighting** in ATLAS (LPNHE)
- NN to simulate **fuel evolution** in nuclear reactors (IPNO)
- BDT's for multidim **reweighting** between MC (LAL)
- Gaussian Processes to **smooth MC** stat fluctuations (LAL)

ML @ IN2P3

3. Object Reconstruction, Identification, and Calibration

Several contributions:

- **Tracking** ML challenge for LHC (LAL)
- **b-tagging** algorithms with BDT's for ATLAS (CPPM)
- **Particle identification** for LHCb (LPNHE)
- **Position reconstruction** of particles for med app (IMNC)
- Reconstruction **calorimeter** objects with CNN, RNN for LHCb (LAL)
- DNN to optimize **jet reconstruction** using RNN for ATLAS (LPSC)
- RNN for **tau ID** and QCD rejection for CMS (IP2I)
- Reco position, tracking **gamma** for nuclear app. (IP2I)
- Full **Event interpretation** algorithm with DNN, Belle 2 (IPHC)
- DNN for **calo reco** and transfert to FPGA for L1 ATLAS trigger (CPPM)

Color code

Advanced
Studies
Interest

5. Uncertainty Assignment

Contributions:

- **Systematic** aware training (LAL)
- ML tools for handling **uncertainties** ATLAS (LPNHE)

Color code

Advanced
Studies
Interest

6. Learning the Standard Model – searches for anomalies

Color code

Advanced
Studies
Interest

Contributions:

- Search for **anomalies** (LPC)

ML @ IN2P3

7. **Matrix Element Method with ML** → **Uncovered ?**

8. **Theory Applications**

- LPSC: ML activities for HEP **phenomenology** (LPSC)

7. **Computing Resource Optimization**

- **CCIN2P3**

Color code

Advanced
Studies
Interest