Université
Paris Cité

Machine Learning for/in Astroparticle Physics
Special focus on VHE Astronomy

–

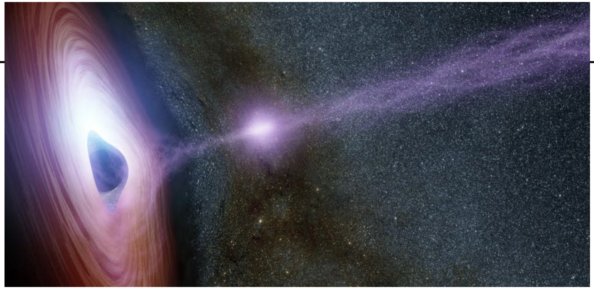
Yvonne Becherini
Université de Paris Cité
Laboratoire Astroparticule et Cosmologie
Data Intelligence Institute of Paris
Astrogamma.eu

List of subjects

- ➔ High Energy Astroparticle Physics: Very High Energy Gamma-Ray Astronomy
- ➔ Data flow in large experiments
- ➔ Feature-based ML and Deep Learning
- ➔ Examples of ML applications
- ➔ Visualization of data
- ➔ Data filtering
- ➔ Problems related to simulated data

Copyright: ESA/NASA, the AVO project and Paolo Padovani

Very High Energy Astrophysics



The goal is to understand

- ➔ The mechanisms of generation of energy in the Universe
- ➔ The creation and propagation of energetic particles in the Universe: gamma-rays, neutrinos, protons
- ➔ The nature of Dark Matter

Methods used

- ➔ Observation of phenomena through ultra-precise and ultra-sensitive particle detectors
- ➔ The analysis of the data acquired is often complex for one main reason: the signal searched is tiny, compared to a huge amount of background

Astroparticle Physics research

- Gamma-Ray Astronomy
- Neutrino Astronomy
- Gravitational waves
- Cosmic Rays

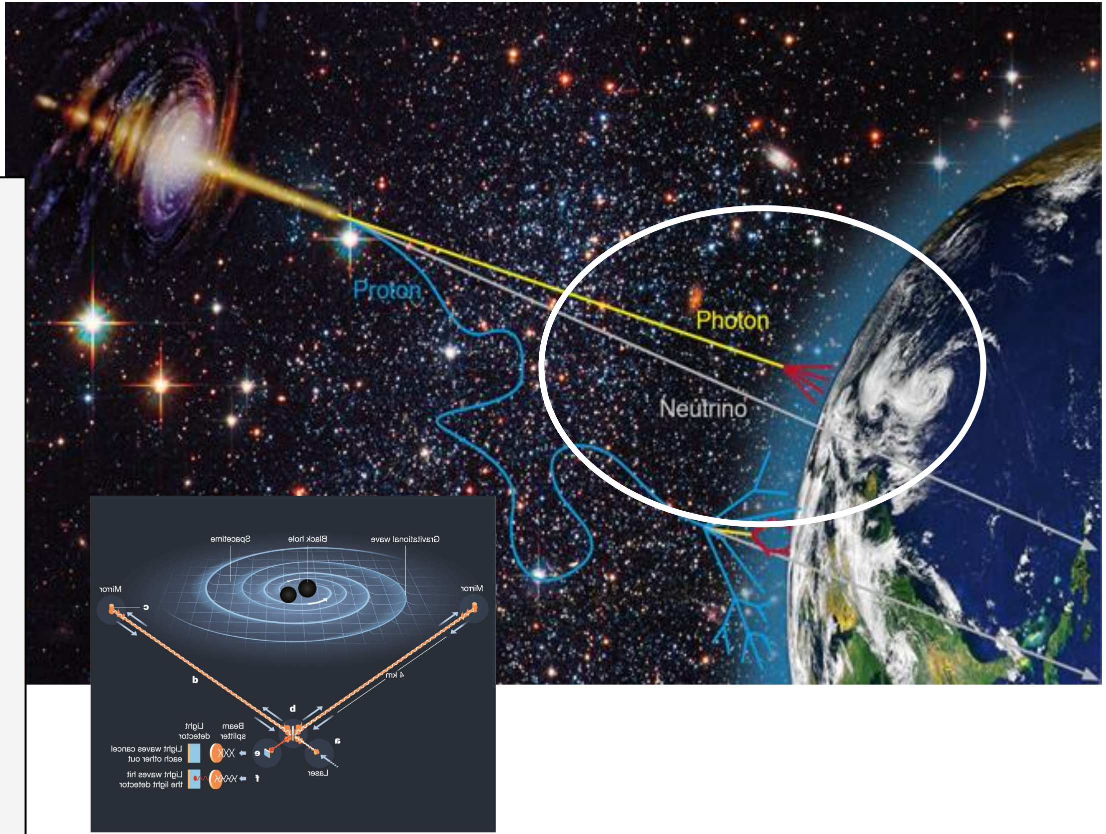
Multi-messenger Astronomy

Photons
- Travel in a straight line
- Origin of accelerated particles difficult to identify
- Limited Horizon

Neutrinos
- Travel in a straight line
- Difficult to detect, because they interact very weakly
- If neutrinos present, then accelerated protons

Cosmic rays
- Deviated by magnetic fields up to very high energy: do not point towards their source
- At very high energy: very rare, require very large detection surfaces

Gravitational waves
- Present only for certain types of phenomena



Understanding the Universe through
the detection of gamma rays & neutrinos
with imagers, trackers & calorimeters

For a better and faster performance
of the data analysis, through:

- Parameter regression
- Event classification
- Monte Carlo simulations augmentation

New frontiers:

- Event filtering
- Visualization
- Simulation refinement

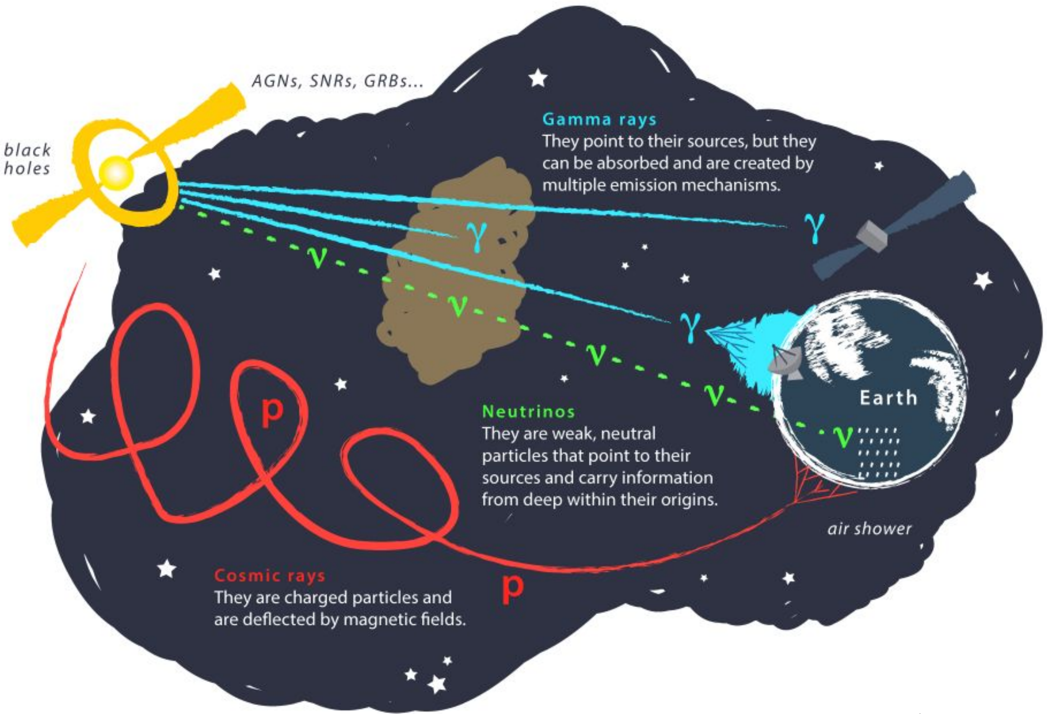
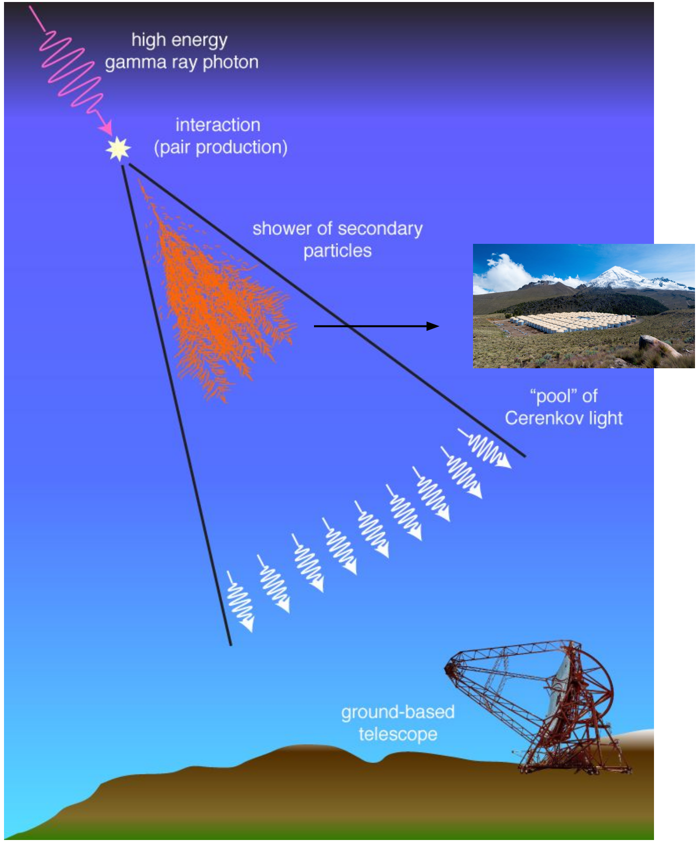


Image: Juan Antonio Aguilar and Jamie Yang. IceCube/WIPAC

Very High Energy Gamma Ray Astronomy Experiments

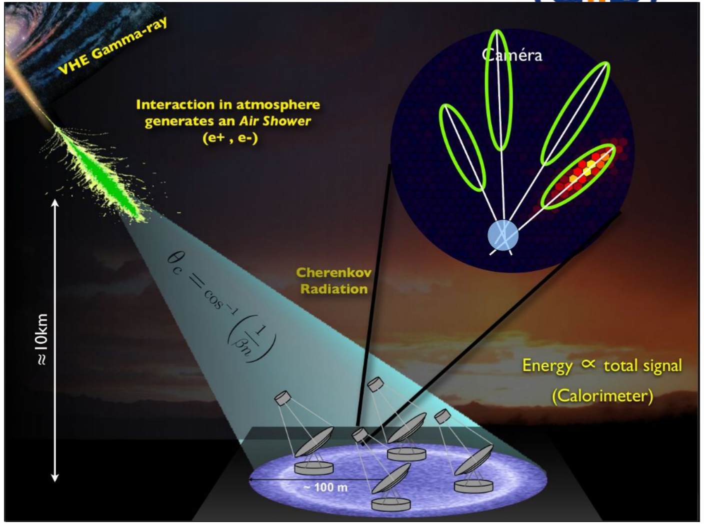
- Observations are made:
  - With Imaging Atmospheric Cherenkov Telescopes detecting the Cherenkov light generated in the atmosphere by the passage of highly relativistic charged particles
  - Wide field of view detector arrays: Surface detectors catching the particles in the atmospheric showers
- Gamma rays from astrophysical sources are rare, and at the same time we receive a huge amount of background events from cosmic rays (very similar)
- The amount of data generated can be huge: several TB per month

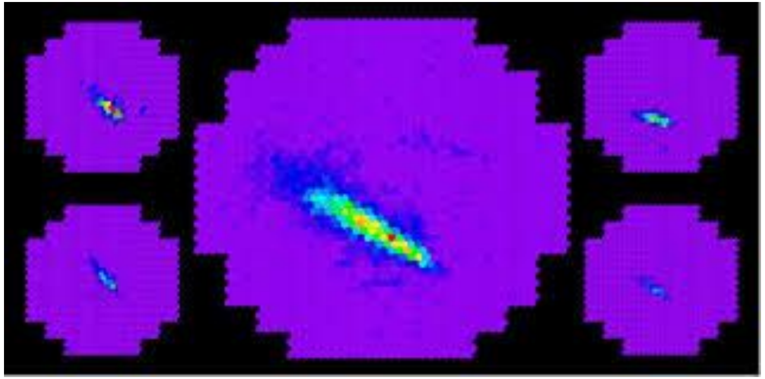


The H.E.S.S. telescopes

The HESS telescope array is located in Namibia and is a 5-tel array of Imaging Atmospheric Cherenkov Telescopes detecting the Cherenkov light created in the atmosphere by the passage of highly relativistic charged particles.

One of the most crucial steps in the analysis of data, is the suppression of the cosmic ray background to extract the “signal” of gamma-rays.

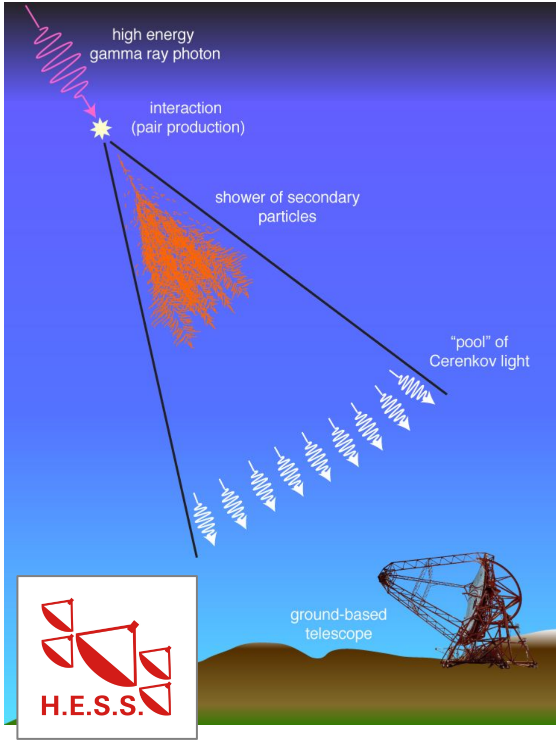




H.E.S.S.

H.E.S.S. : a very-high-energy gamma-ray observatory

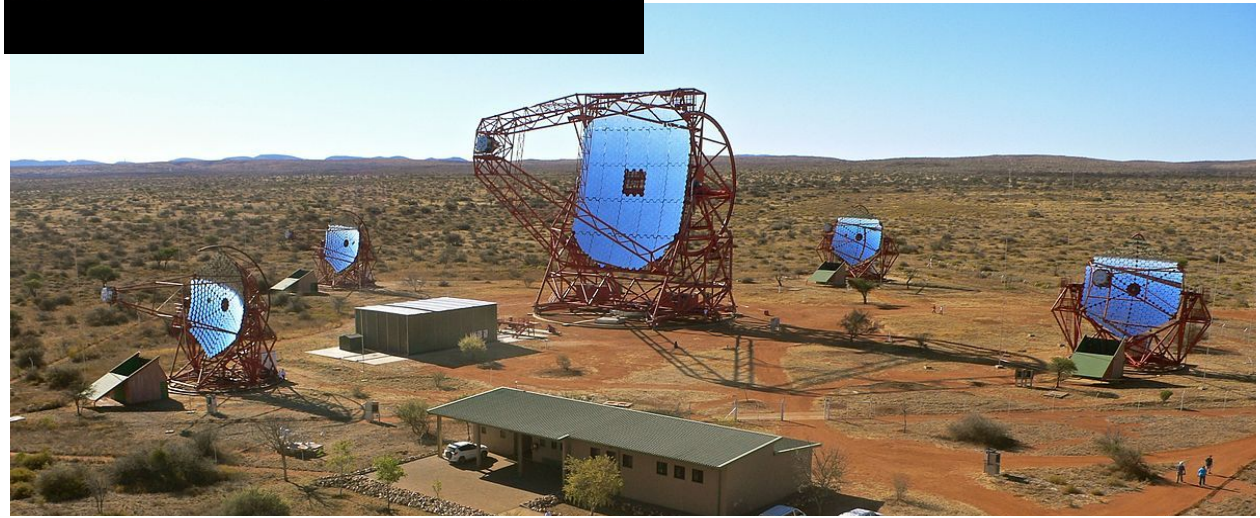
Energy interval
50 GeV-100 TeV
In activity since 2003 up to 2024



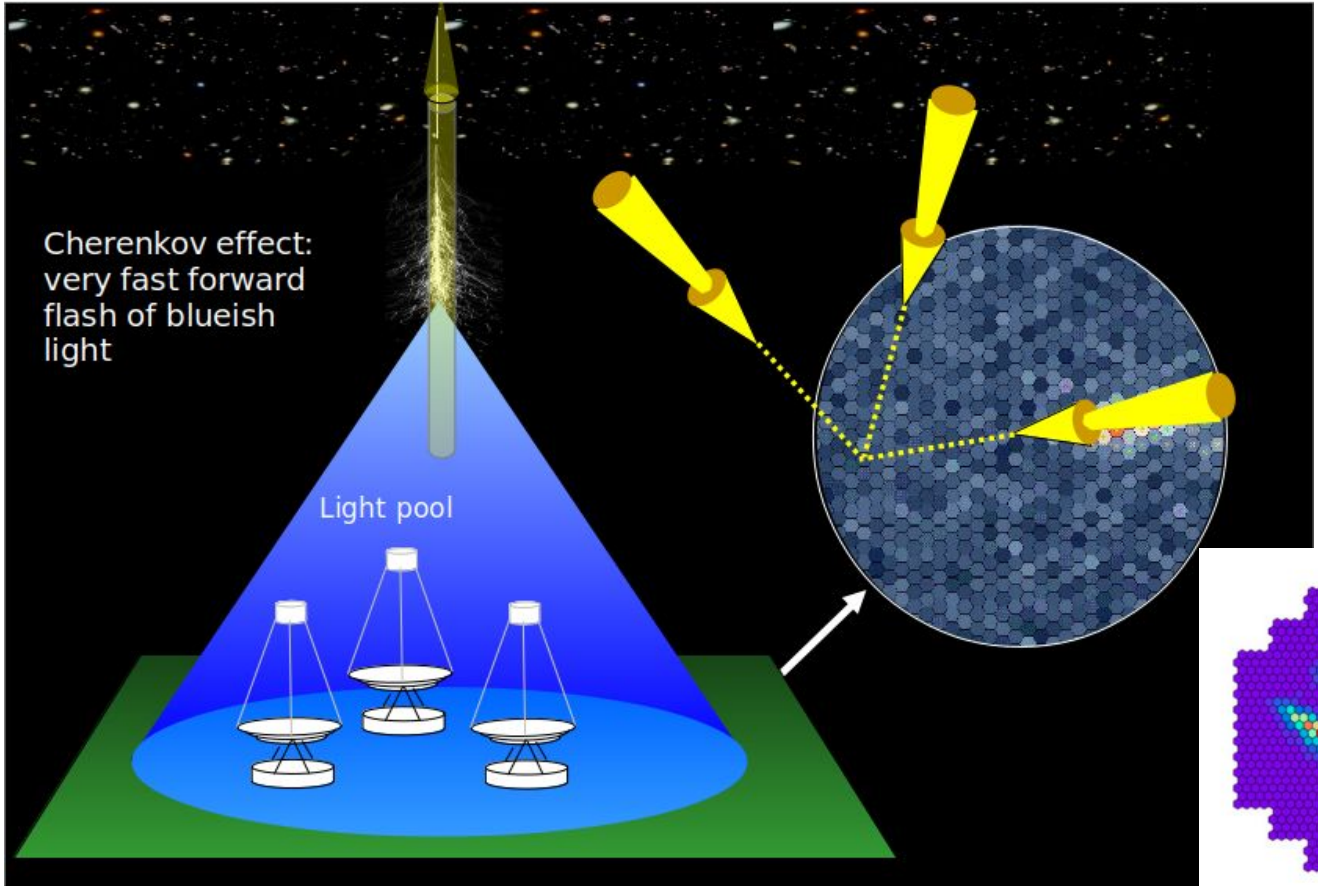
H.E.S.S. is one of the three VHE gamma-ray observatories of the current generation

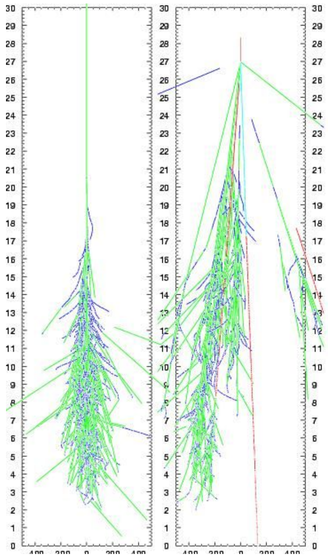
H.E.S.S. just had a party for its 20th anniversary

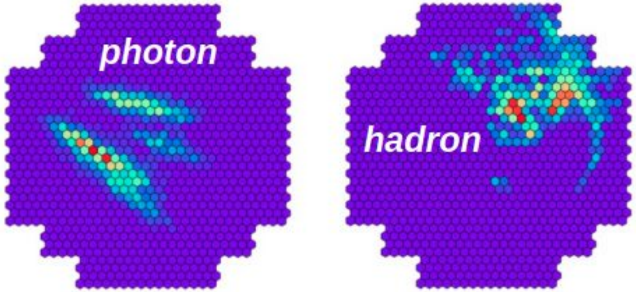
In operation in Namibia at an altitude of 1800 m



H.E.S.S. : gamma-ray detection method







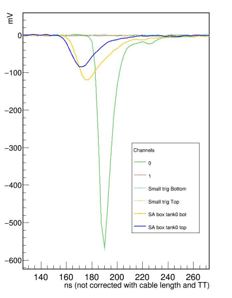
Preparing an analysis strategy in Astroparticle Physics (classic way, no ML)

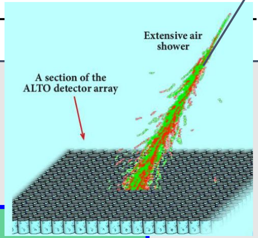
Signal Monte Carlo

Background Monte Carlo

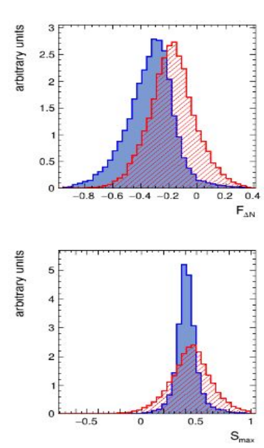
Background Real Data

Calibration





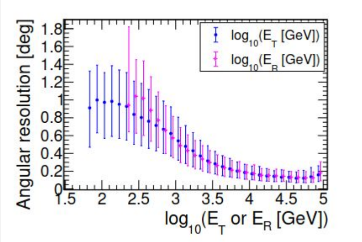
Reconstruction of arrival direction, Energy, etc for all events via the minimization of a function "Goodness of fit" approach



Signal over Background Discrimination i.e.
Fix the analysis cuts, through square cuts

Instrument Response Functions,
Angular & Energy resolutions,
Effective areas

In VHE Gamma-Ray Astronomy now customary that 2 independent analysis chains confirm the results

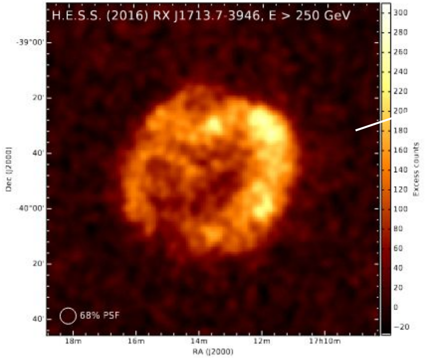


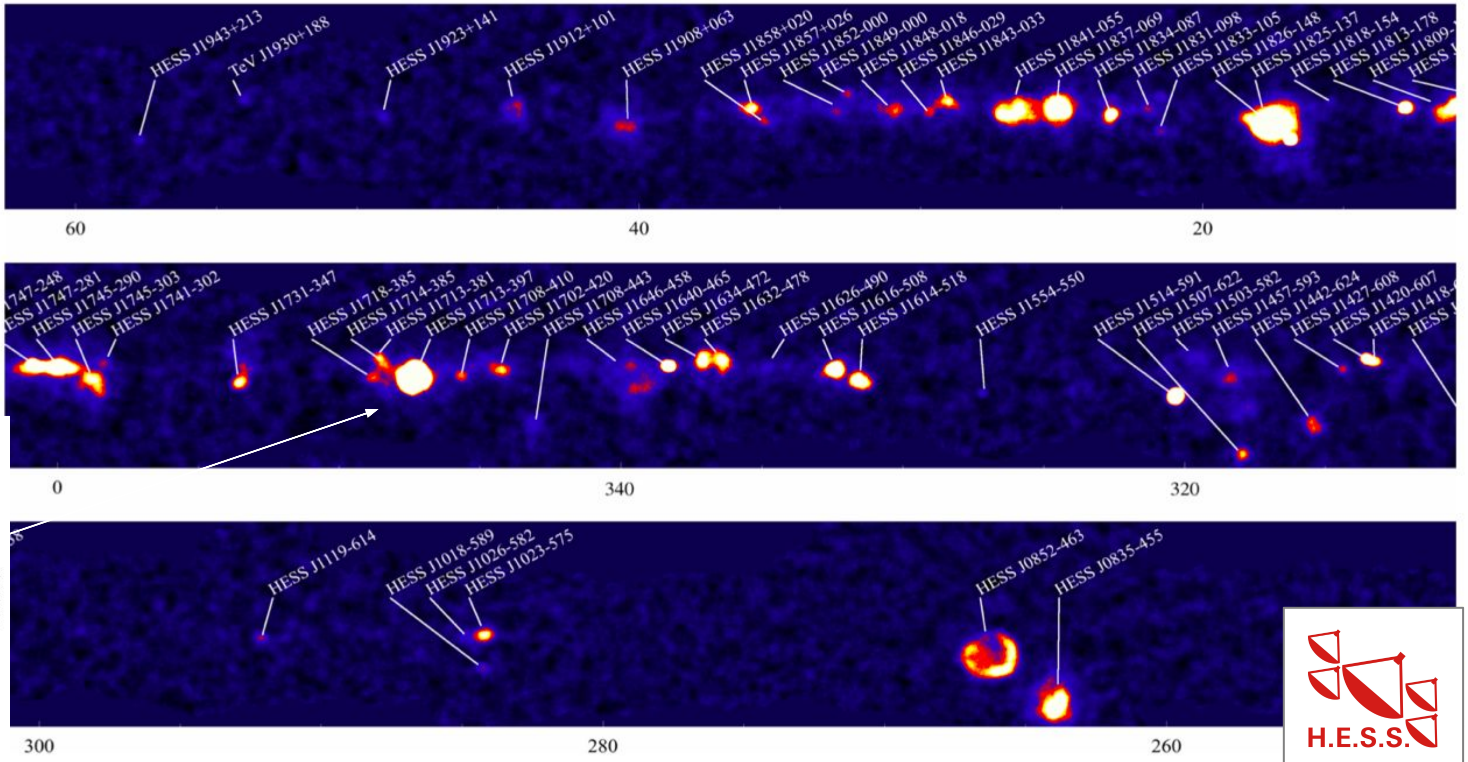
Preparation

Discovery of 252 sources of particle acceleration in our Galaxy and beyond

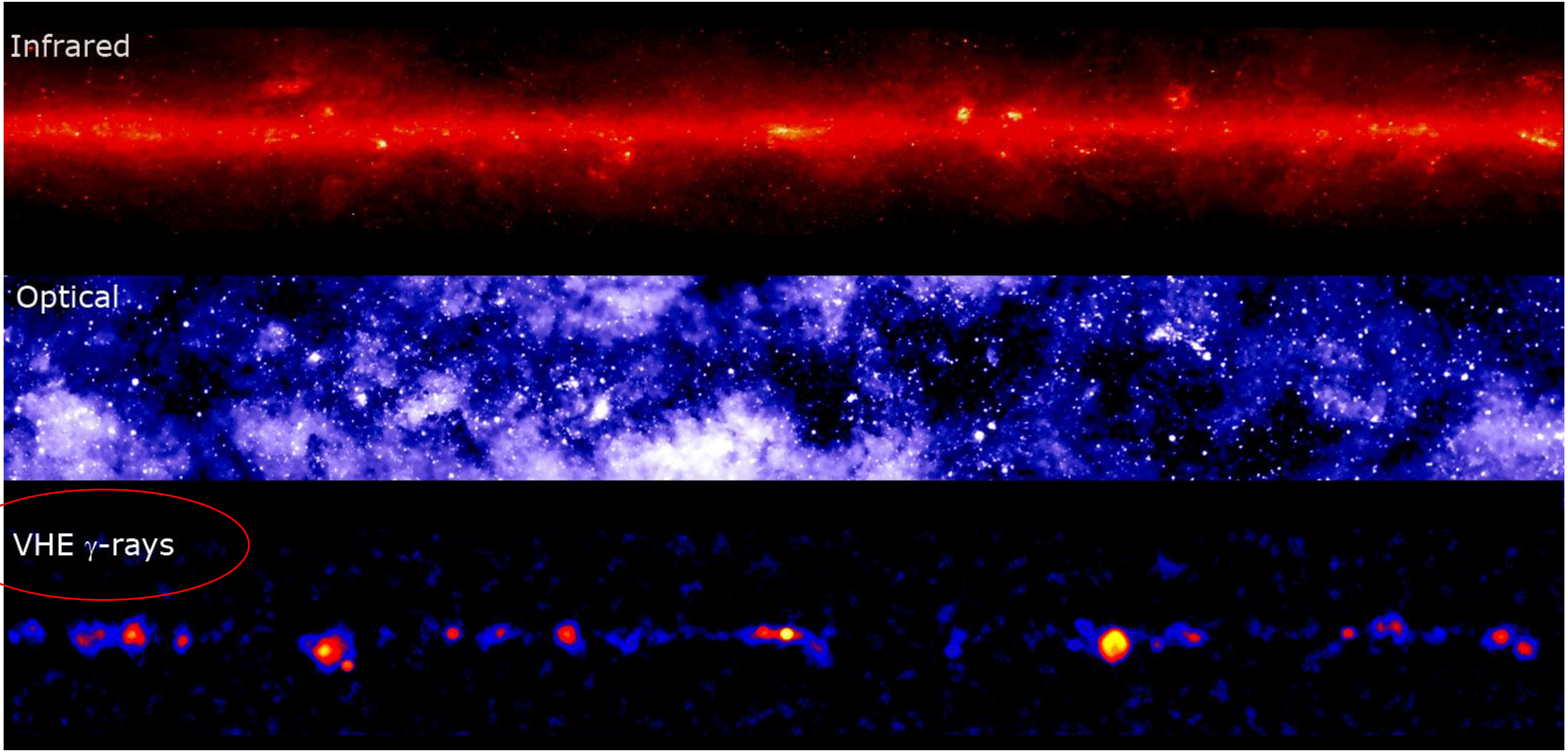
In our Galaxy :

→ Wind shocks in star-forming regions
→ Supernova Remnants
→ Pulsar and Pulsar Wind Nebulae
→ Binary systems, Novae

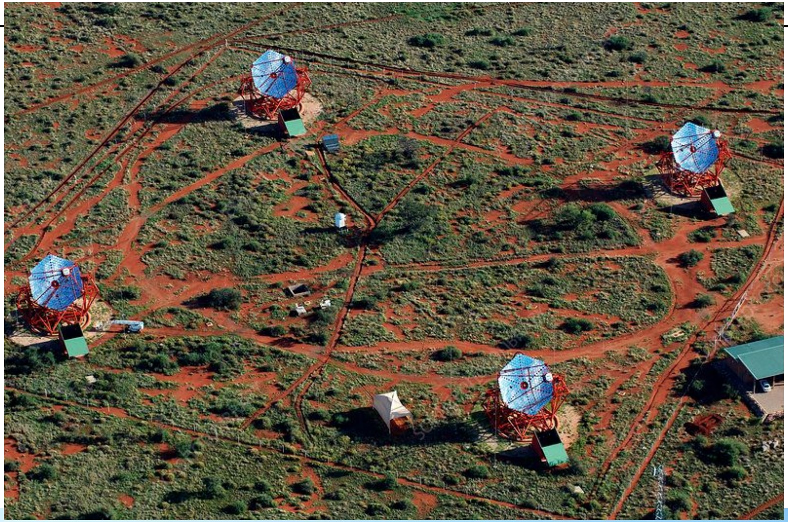




A new observational window onto the Universe



H.E.S.S. success story

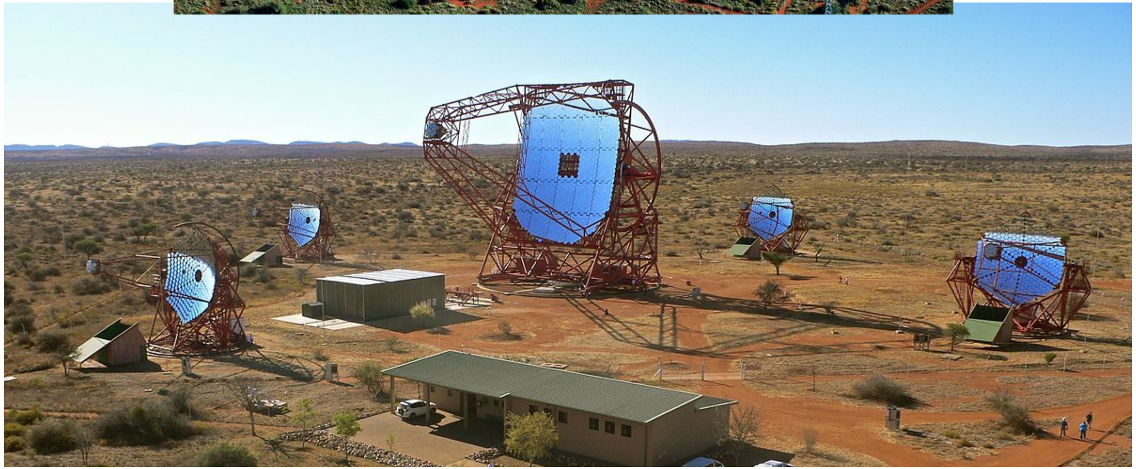


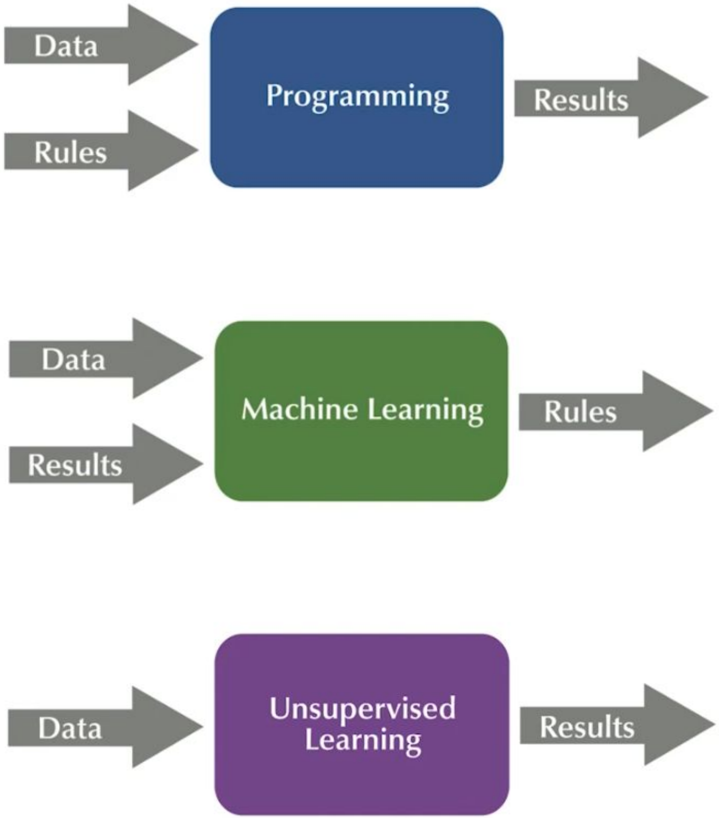
- Data taking began 2003-2004
- In 2008 the most luminous sources were already discovered
  with standard analysis methods
- Needed a boost in sensitivity
  to see more sources

Possibilities:

- Hardware: new telescope
- Software: Machine Learning
- Or both!

What I will show here is just
the gain driven by Machine Learning





Programming

- You collect a bunch of data, you apply some known rules, and you turn that set of data and rules into the results

Supervised Machine Learning

- We have the data and the results (the labels) and we input these into an ML model that produces the rules that we want for the programming

Unsupervised Learning

- We do not have rules nor labels in input, so here we only have the unlabelled data
- We want to output something about the structure of the data (how data cluster, how dense are the structures, or, we just want to reduce the dimensionality of data)

Leland McInnes

Preparing an analysis strategy in Astroparticle Physics (with ML)

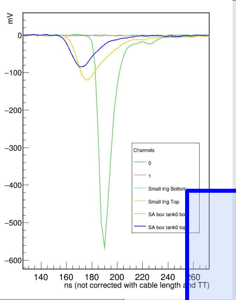
Signal Monte Carlo

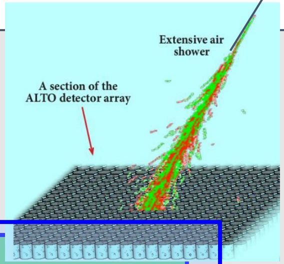
Unsupervised Learning

Background Monte Carlo

Background Real Data

Calibration

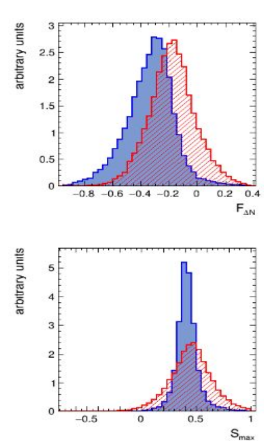




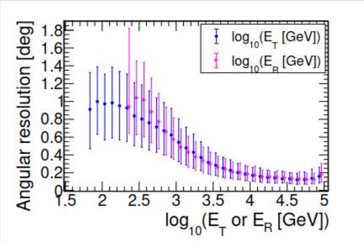
Reconstruction of arrival direction, Energy, etc for all events

Supervised Learning

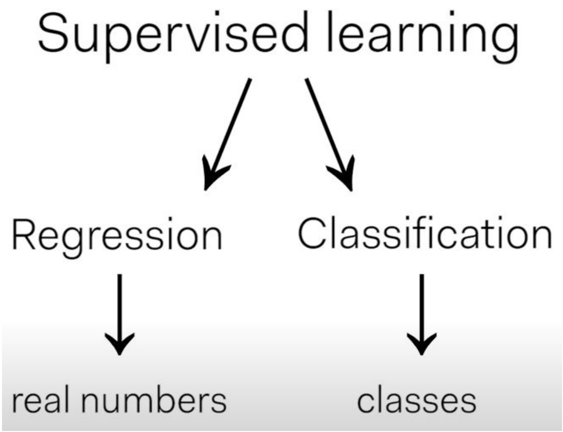
Signal over Background Discrimination
i.e. Fix the analysis cuts



Instrument Response Functions,
Angular & Energy resolutions,
Effective areas



Preparation



Supervised Learning has two main branches

- In Classification, the output of the model is Classes or Categories.
- In Regression, the output of the model is a real number

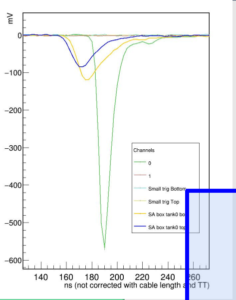
Signal Monte Carlo

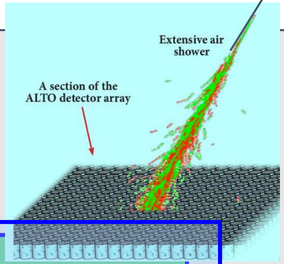
Unsupervised Learning

Background Monte Carlo

Background Real Data

Calibration





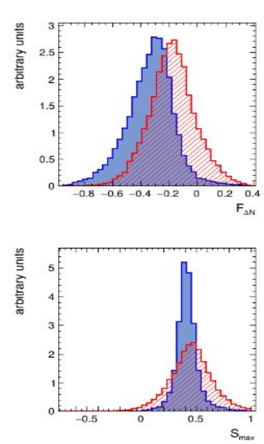
Regression

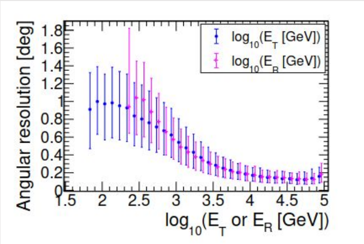
Reconstruction of arrival direction, Energy, etc for all events

Supervised Learning

Classification

Signal over Background Discrimination
i.e. Fix the analysis cuts



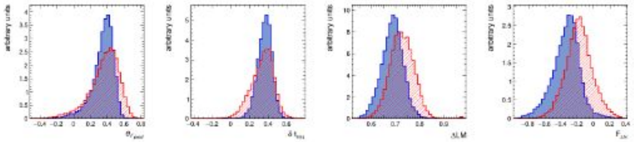


Instrument Response Functions, Angular & Energy resolutions, Effective areas

Preparation

Supervised Learning: Feature-based ML and Deep Learning

Feature-based Machine Learning ("Classic")



Classification

- Define parameters which differ between the signal searched and the background

Regression

- Search for correlations between the value to be regressed and the other parameters available

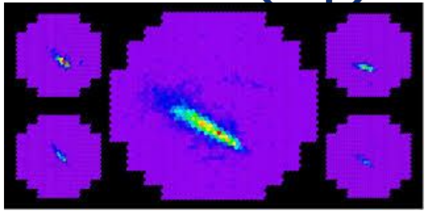
Advantages

- Quick implementation, simple to add new parameters

Disadvantages

- A lot of time spent in feature engineering, and important parameters might be missed

Deep Learning



Classification & Regression

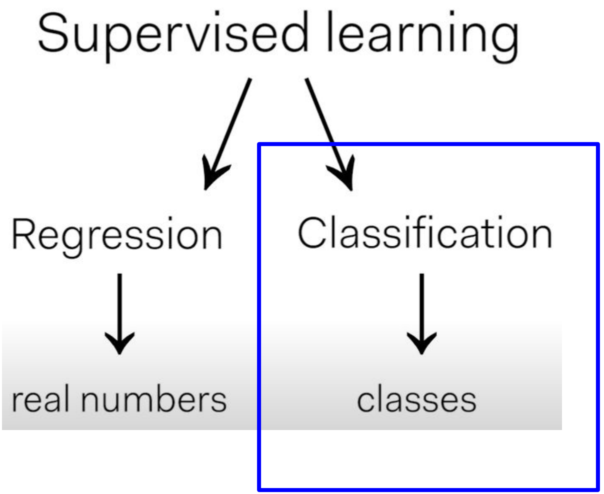
- Define the inputs: 2D maps, time series, graphs
- Define the output

Advantages

- No feature engineering needed, as the relevant parameters are learned internally by the NN

Disadvantages

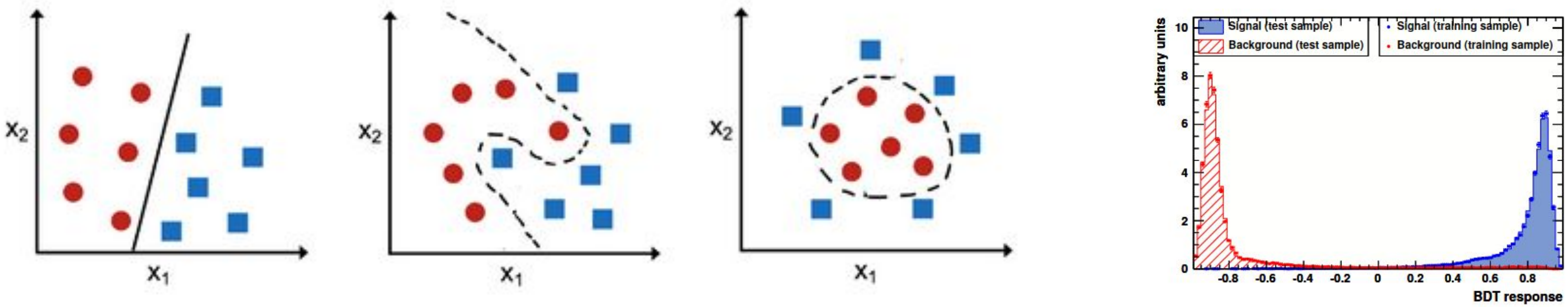
- Sometimes slow and needing GPUs for complicated tasks
- Need to be sure that the NN is learning what you want it to learn (check for pitfalls)



Supervised Learning has two main branches:

- In Classification, the output of the model is Classes or Categories.
- In Regression, the output of the model is a real number

Supervised Learning: Signal over Background separation (Binary Classification)



Extract a rare “signal” in the presence of a large amount of noise, which is equivalent to finding a needle in a haystack

Need to develop powerful methods to extract these rare events.

For most problems, after having cleaned properly the data, the answer is

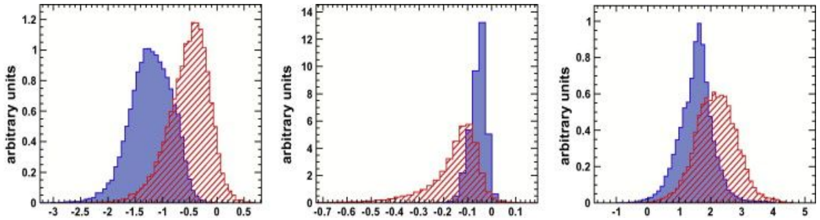
- event classification through user-defined features

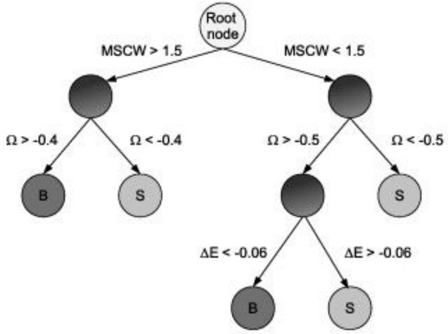
But if you wish to achieve a better classification performance in difficult phase-space regions (where the signal is very small, for instance), better to switch to Deep Learning

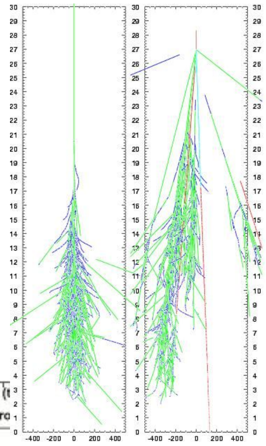
Y. Becherini et al., Astroparticle Physics (2011)   M. Senniappan, Y. Becherini et al, JINST (2021)

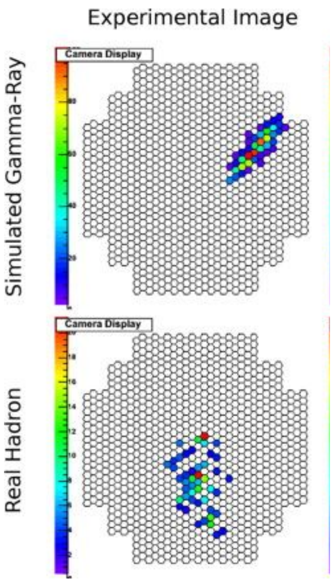
Gamma/Proton classification in H.E.S.S. data

- Supervised feature-based ML
- Classification of gamma-rays and protons using a set of user-defined input variables
- The algorithm performing the separation is the Boosted Decision Trees method.

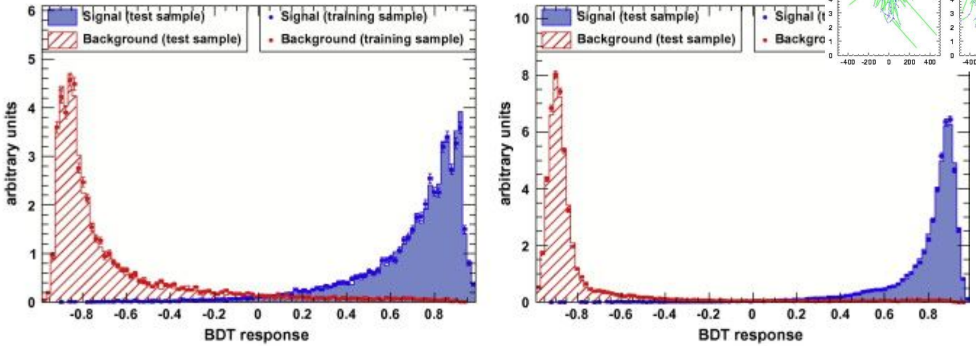




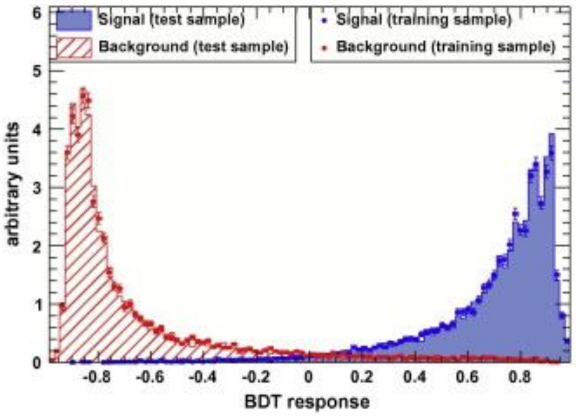


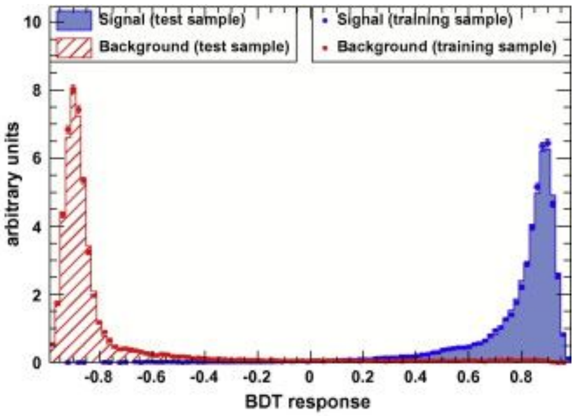


[“A new analysis strategy for detection of faint gamma-ray sources with Imaging Atmospheric Cherenkov Telescopes”, Astroparticle Physics, (2011)]



Gamma/Proton classification in H.E.S.S. data

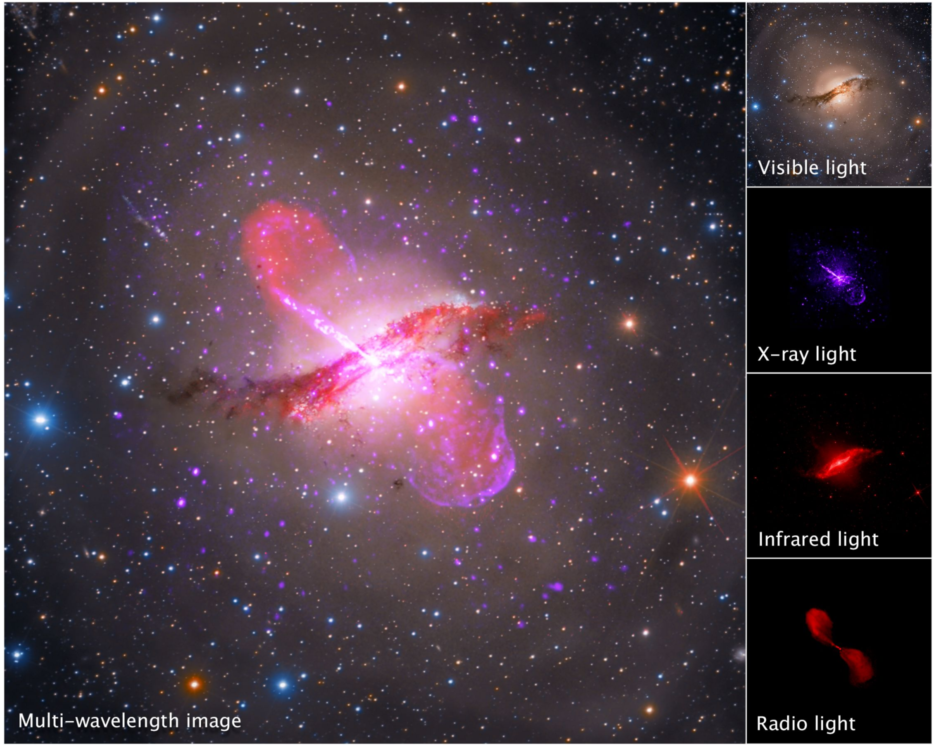




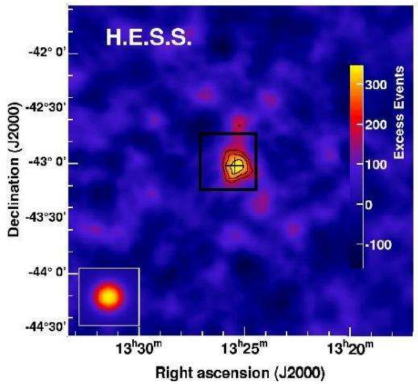
- The final response of the algorithm to an independent set of data (test data) allows defining an analysis cut before looking at the real data.
- The final analysis cut can be based on a desired gamma-ray efficiency. Example: if I say I will cut at 0.4 on the right plot, I will have 95% of gamma-rays and a contamination of less than 1% of protons.
- When the analysis cuts are frozen, you are then allowed to look at the real data.

Implications of having a better event classification

The Active Galactic Nucleus
Centaurus A - A very weak source (110 hours of obs)

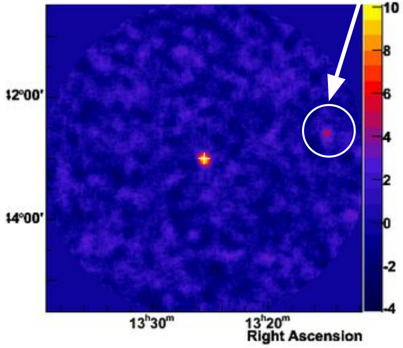


2008: discovery with a detection significance of 5σ with standard analyses (no ML)



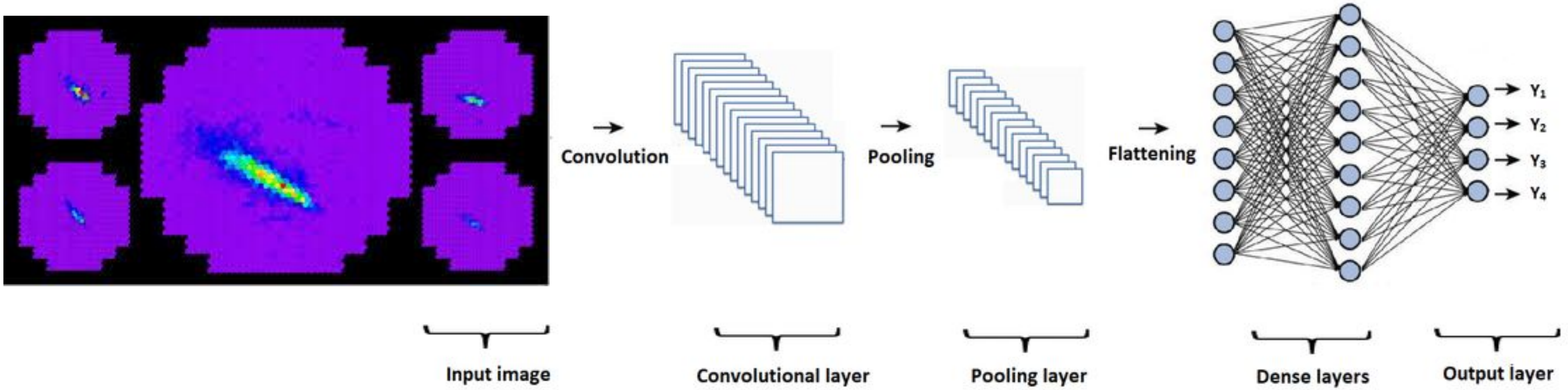
2016: Re-analysis of data using supervised ML 9.8σ !

Appearance of a second source in the field of view!



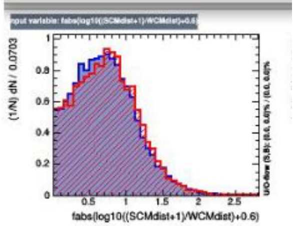
A clear gain in the detectability of **weak** gamma-ray emitting sources

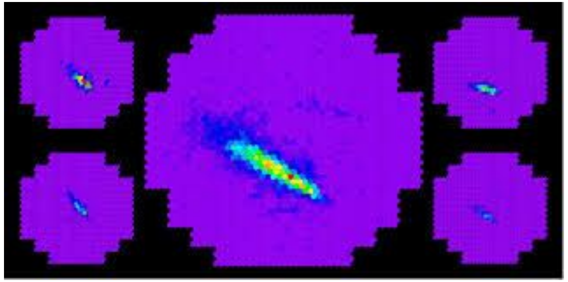
Source


Convolution
Pooling
Flattening
Y1
Y2
Y3
Y4
Input image
Convolutional layer
Pooling layer
Dense layers
Output layer

Programming versus Machine Learning versus Deep Learning for Classification

- For event classification, a clear gain in using Machine Learning versus the standard way of implementing square cuts in High Energy Physics. With relatively small effort, a factor of 2 in sensitivity can be reached.
- Extremely useful for the detection of sub-threshold sources (weak emitters)
- Standard Machine Learning working very well, but:
  - Need to perform Feature Selection
  - Only limitation is where the defined features do not catch any difference in the samples





- Deep Learning might help in these regions, where the human cannot see any difference between the groups
- Typically: regions of low signal, low energies
- Deep Learning might catch slight differences, where the human cannot
- Efforts ongoing, but results show that you need an ultra-wide representation of images
- CNNs have been developed, but they show to be sensitive to the night-sky-background

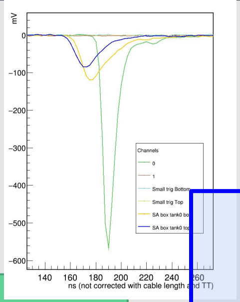
Parsons, R.D., Ohm, S.
Eur. Phys. J. C **80**, 363 (2020)

Signal Monte Carlo

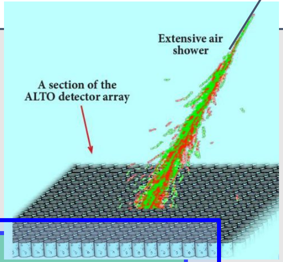
Unsupervised Learning
Generative Models

Background Monte Carlo

Background Real Data



Calibration



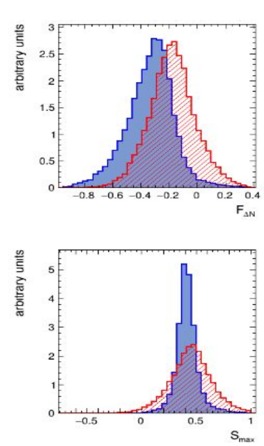
Regression

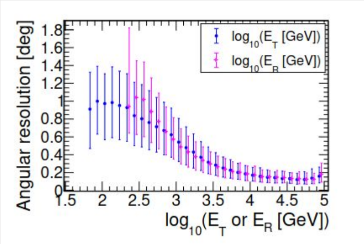
Reconstruction of arrival direction, Energy, etc for all events

Supervised Learning

Classification

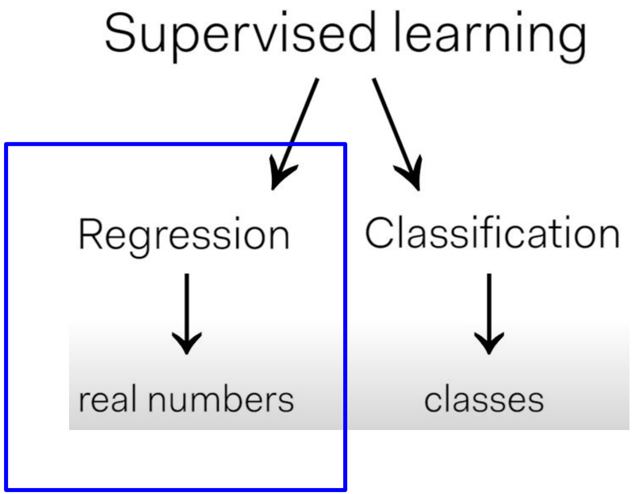
Signal over Background Discrimination
i.e. Fix the analysis cuts





Instrument Response Functions,
Angular & Energy resolutions,
Effective areas

Preparation



Supervised Learning has two main branches:

- In Classification, the output of the model is Classes or Categories.
- In Regression, the output of the model is a real number

Supervised Learning: Reconstruction of event parameters (Regression)

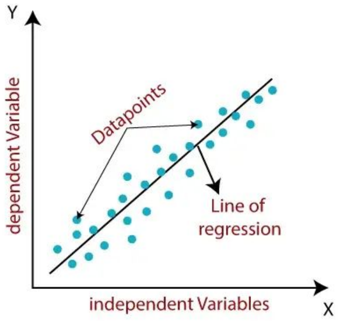
After data are calibrated, we need to perform the reconstruction of the kinematics of the gamma ray: incoming direction and energy

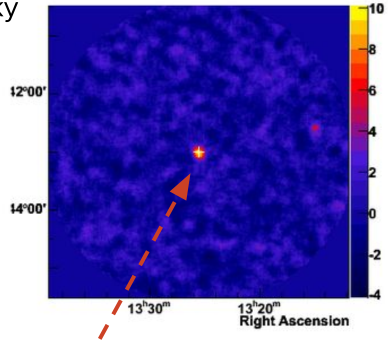
This can be done with algorithms or using ML regression

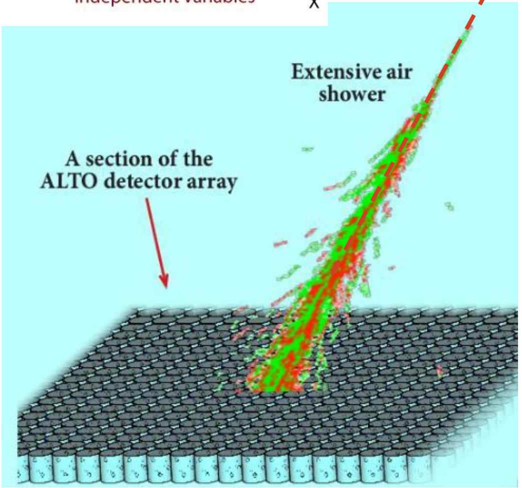
Regression is a method allowing to estimate the value of a variable associated with the signal (or the background).

With the help of simulations and data analysis, we can infer what the values expected for a particular event are

Astrophysical source in the sky





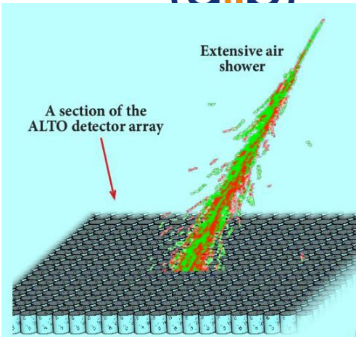


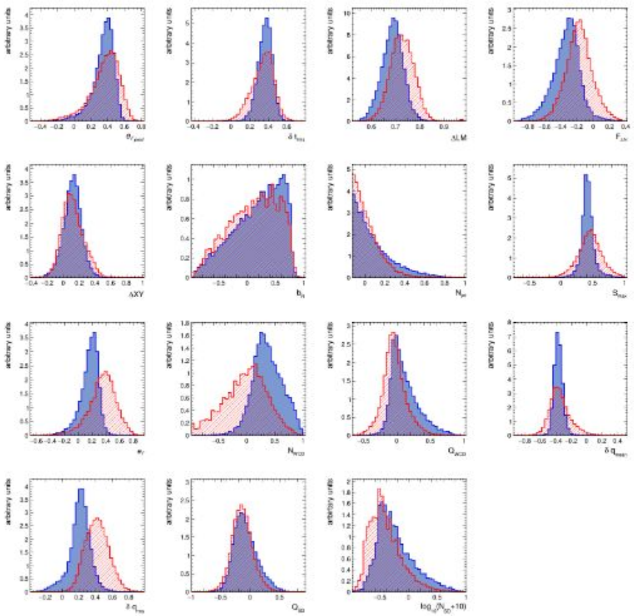
What the gamma ray would look like close to the detector

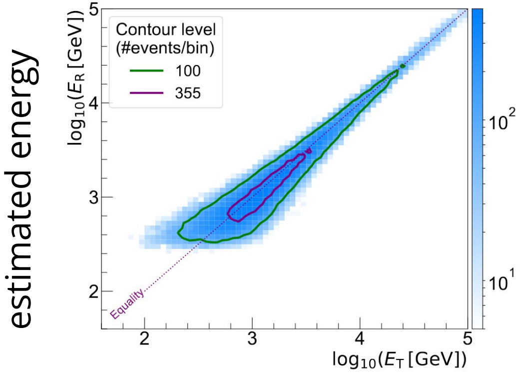
Regression of event parameters in the ALTO R&D project

Event energy and height of shower maximum
But also: arrival direction

➔ Standard procedure of **feature engineering** and selection and then neural networks (Multi Layer Perceptron)



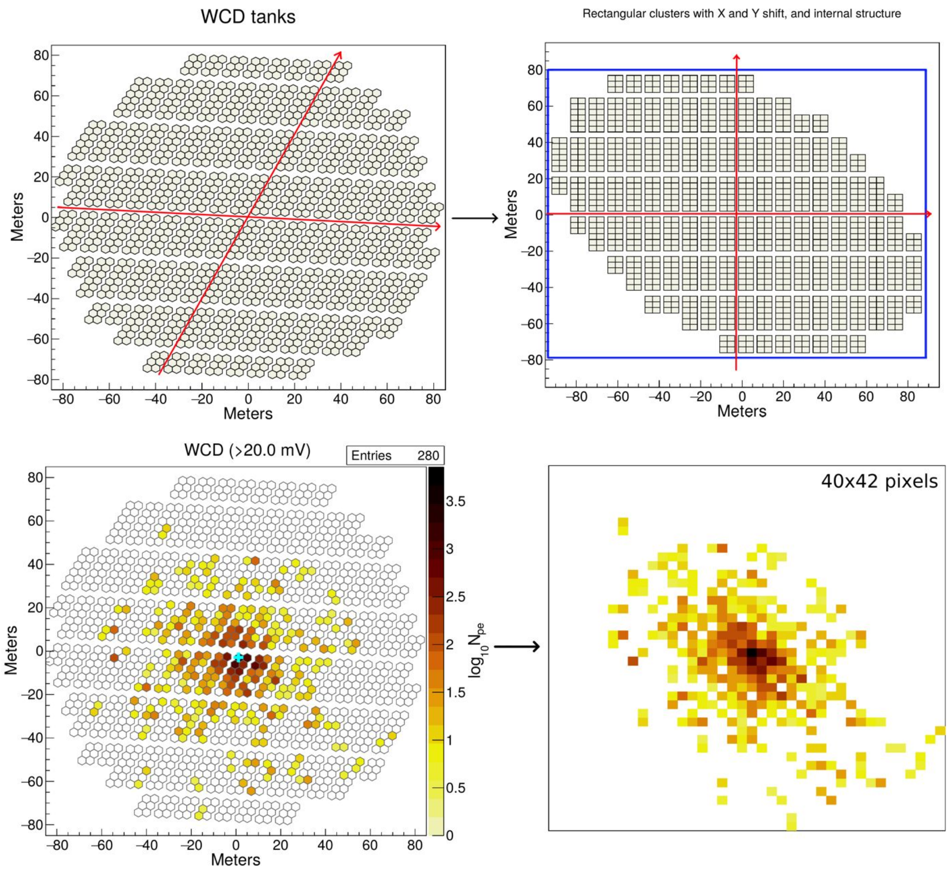




simulated energy

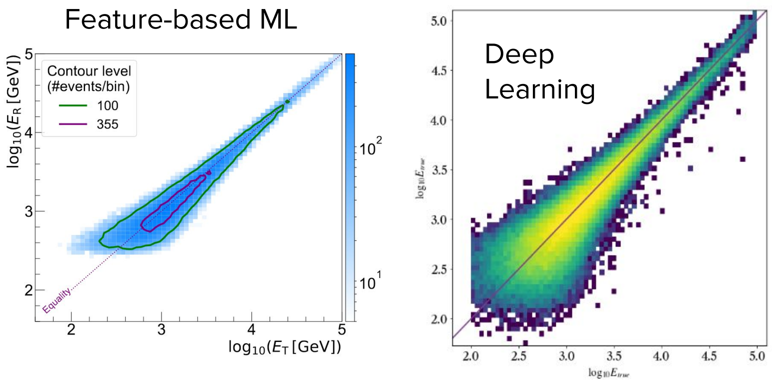
Better resolution on estimated parameters = smaller errors on physical results.

Regression of event parameters with Deep Learning



- ➔ **Deep Learning** of the images (“footprints”) of gamma rays in the detectors
- ➔ Input images contain calibrated footprints, converted to a rectangular array using the Oversampling method suitable for the Convolutional Neural Network

Feature-based ML

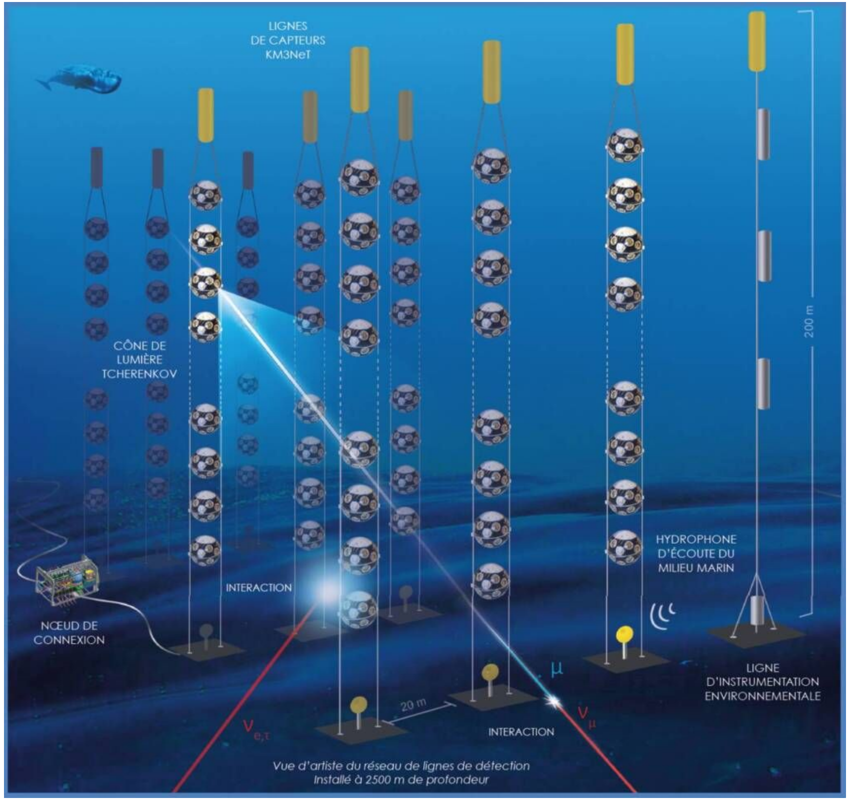


- A clear gain in performance in energy reconstruction seen with Machine Learning and Deep Learning

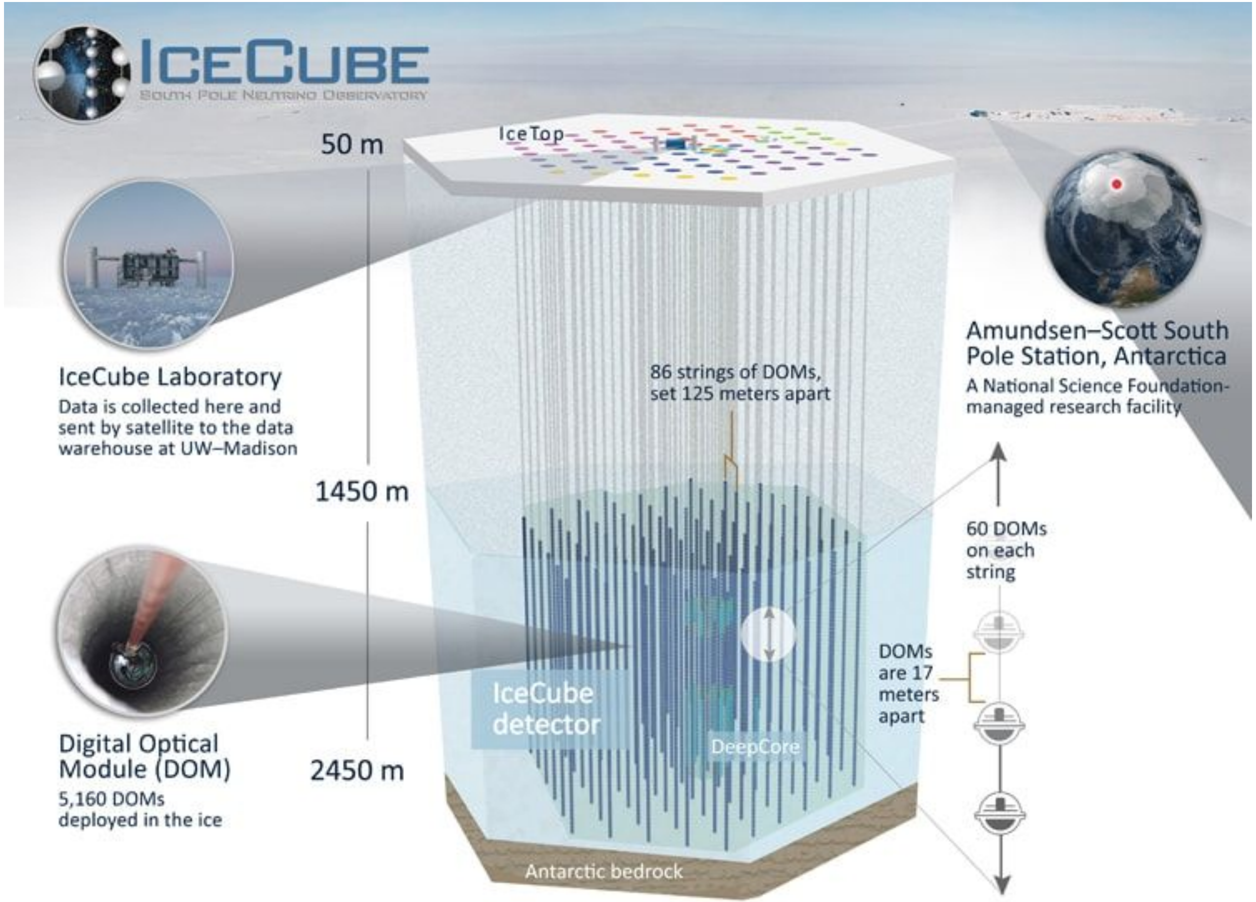
- The improvement in the incoming direction of gamma rays is still uncertain, as standard algorithm-based reconstruction methods are performing well

- Deep Learning And especially Graph Neural Networks (GNNs) for direction reconstruction in Neutrino Telescopes (IceCube, KM3NeT) seems much more promising, especially at the lowest neutrino energies

    - The uncertainty on the ice and water absorption and scattering
    - These are not known very well, so Deep Learning helps in modelling them

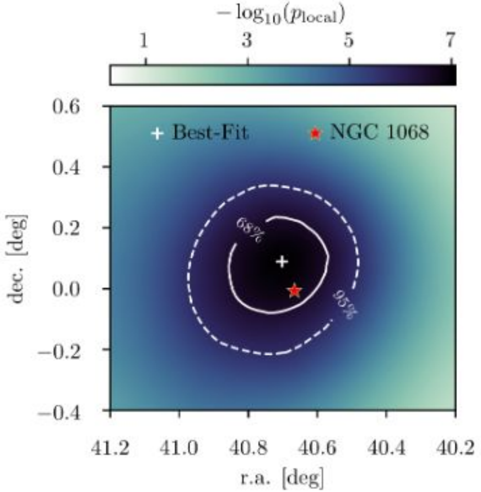


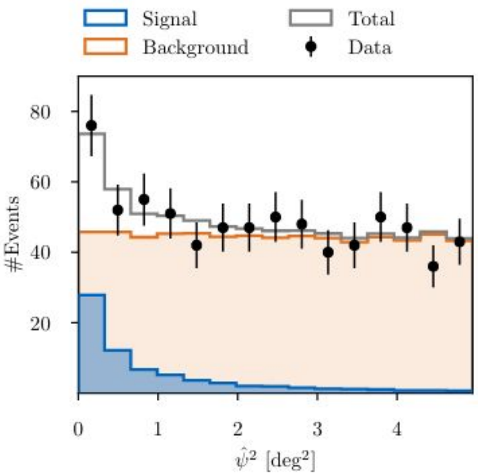
IceCube : a neutrino observatory in the antartic ice



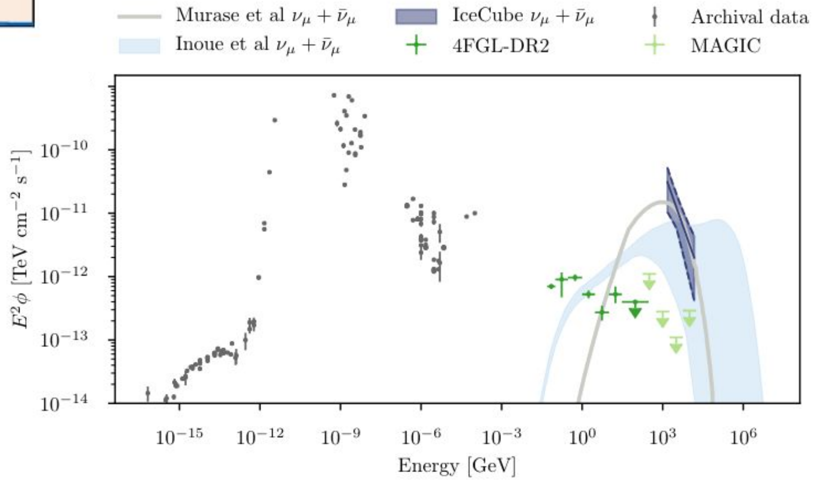
- In activity since 2009
- Size: 1 KM^3
- Results start to pop up after 10 year of data taking

Low energy neutrinos from NGC 1068





"Evidence for neutrino emission from the nearby active galaxy NGC 1068"
Science, 2022

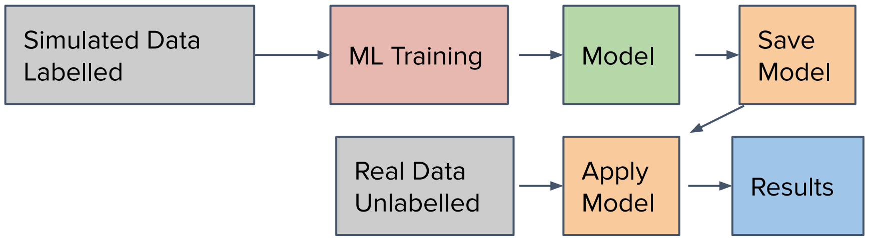


- Results from November 2022
- Signal significance 4.4 sigma
- The highest up to now

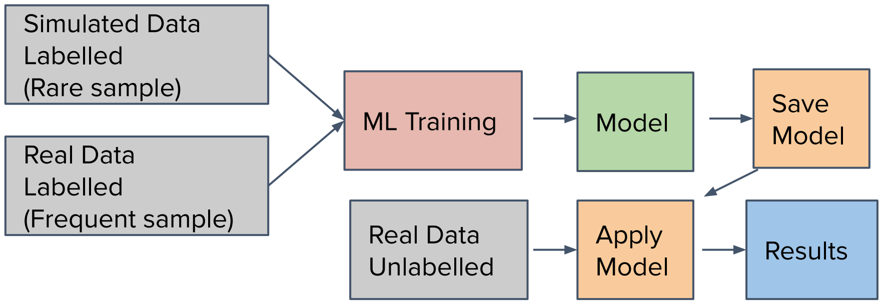
No Graph Neural Networks but
- Gradient boosted decision tree for angular errors (so standard ML)
- CNN for energy reconstruction (DL)

New frontiers: Problems related to simulated data

Scenario 1:
Train on labelled simulated data, apply model on real unlabelled data



Scenario 2:
Train on simulated data (Rare sample) and on real data (Frequent sample), then apply rules on all real unlabelled data



But if my Monte Carlo simulations are not representative of my real data?

New frontiers: Unsupervised Learning

Programming

- You collect a bunch of data, you apply some known rules, and you turn that set of data and rules into the results

Supervised Machine Learning

- We have the data and the results (the labels) and we input these into an ML model that produces the rules that we want for the programming

Unsupervised Learning

- We do not have rules nor labels in input, so here we only have the unlabelled data
- We want to output something about the structure of the data (how data cluster, how dense are the structures, or, we just want to reduce the dimensionality of data)

Leland McInnes

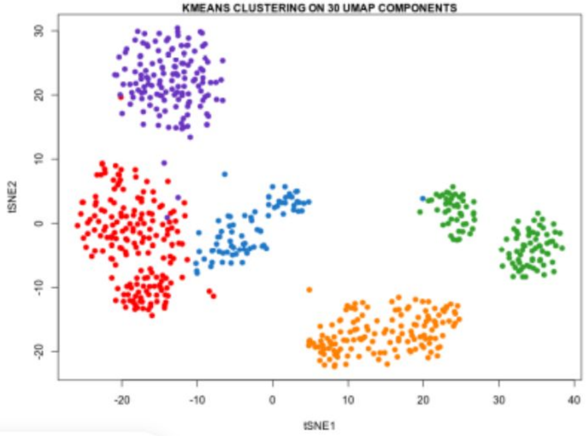
New frontiers: Unsupervised Learning & Generative Models

Unsupervised Learning

The system tries to learn with no supervision (unlabelled data)

Used for:

- Clustering
- Visualization and dimensionality reduction
- Dimensionality reduction & Clustering
- Anomaly Detection and novelty detection



Generative Models

The system learns dense representations of the input data, called latent representation

Autoencoders

- Learns to efficiently construct dense representations of the input data, called latent representation, useful for dimensionality reduction and Visualization purposes
- Acts a feature detector
- Can generate new data that looks very similar to the input data.

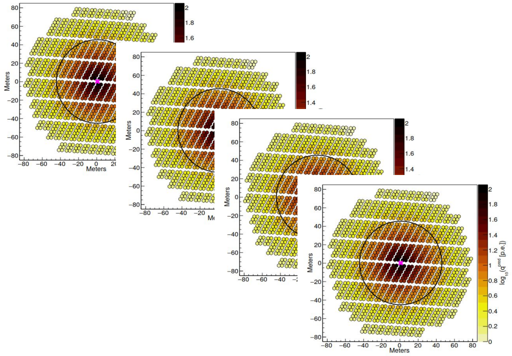
Generative Adversarial Networks (GANs)

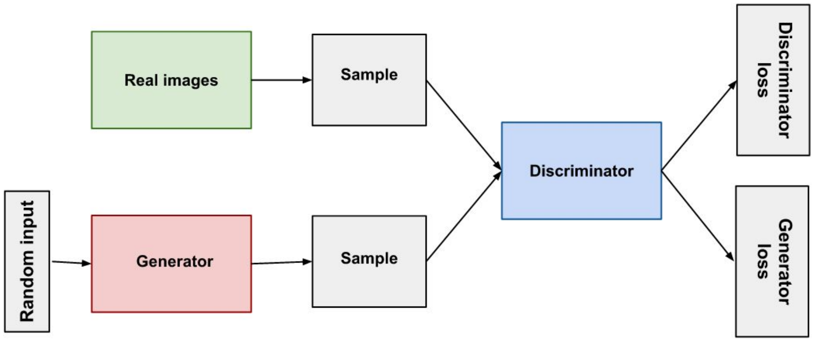
- Can very efficiently generate new data by using two neural networks
  - A generator which tries to generate data that looks similar to the training data and
  - A discriminator that tries to tell real data from fake data

In physics, in order to perform a measurement and to set up the data analysis strategy, we need to simulate the physical phenomena and the response of the measuring apparatus.

Important amount of computing time, as a lot of statistics needed and as they often cannot be parallelized.

One way of simulating more data is by adding “noise” to the existing data, generating more instances of the same data.





A GAN is an ML model in which two neural networks (the generative and the discriminator) compete with each other to become more accurate in their predictions.

The purpose of the generative model is to produce new data, while the discriminative model learns how to create images as similar as possible to the real data.

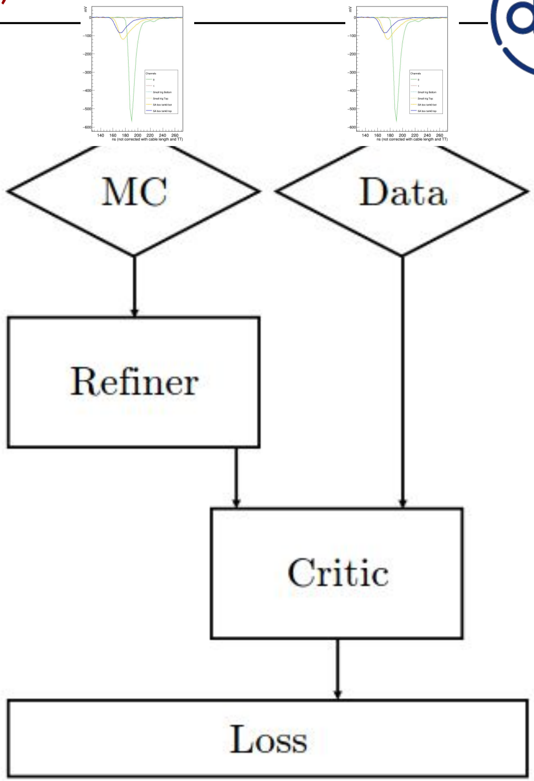
The result is a more and more accurate simulation of images

New frontiers: refinement of simulations (domain adaptation)

- Setting up a data analysis strategy is based on Supervised Learning with Monte Carlo simulations
- The analysis is then applied to real data
- What happens if the Monte Carlo simulations are not able to reproduce the real data? Domain shift. The performance of the analysis will be degraded.
- Lengthy study to improve the simulations

OR… Generative Adversarial Networks

- You allow the Monte Carlo “Refiner” NN to make small changes to the signals seen in the detectors
- The ”Critic” NN will decide, upon also having the real signals in input, if the simulations match the real data or not through the loss calculation
- Step by step, the NN learns how to optimize the simulations

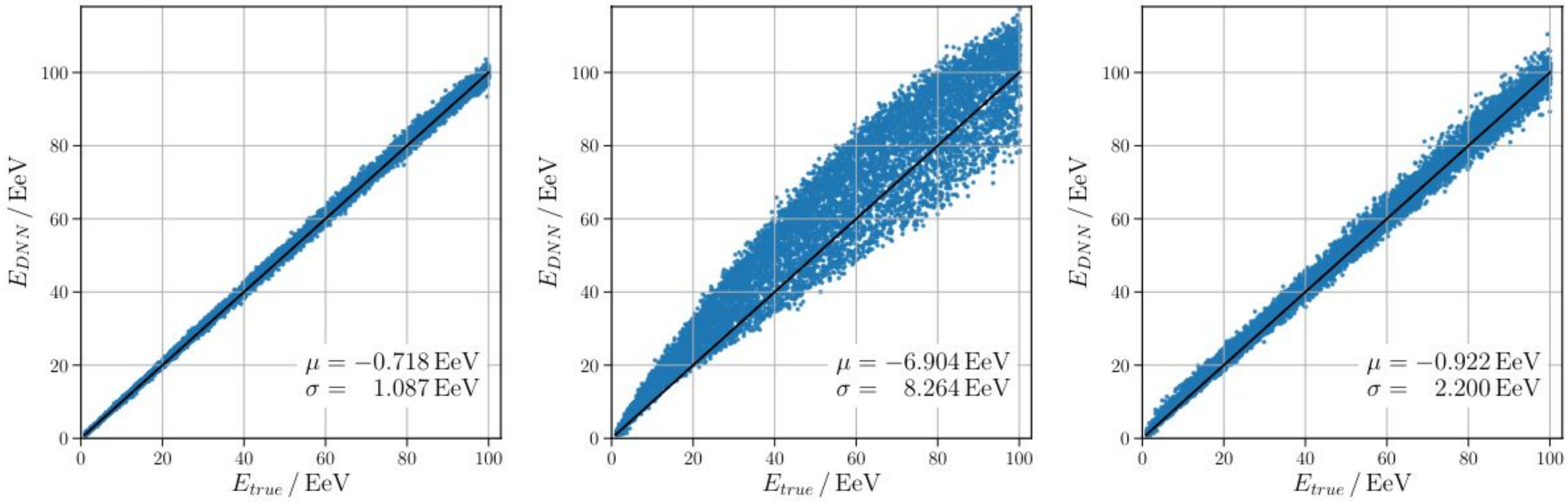


Erdmann, M., Geiger, L., Glombitza, J. *et al. Comput Softw Big Sci* **2**, 4 (2018). https://doi.org/10.1007/s41781-018-0008-x

New frontiers: refinement of simulations (domain adaptation)

Energy reconstruction in the Auger experiment

Erdmann, M., Geiger, L., Glombitza, J. *et al. Comput Softw Big Sci* **2**, 4 (2018). https://doi.org/10.1007/s41781-018-0008-x



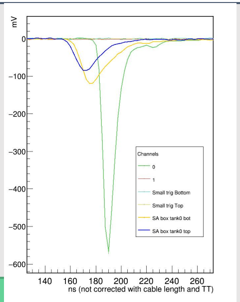
Benchmark:
Energy reconstruction trained on simulated data, and applied on another set of simulated data following the same distribution

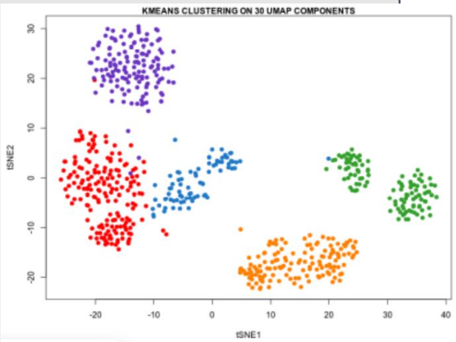
Application to real data
Poor generalization

Application to real data of the refined simulated data training

This shows that the refiner network is able to modify simulations to more accurately resemble the data distribution.

Unsupervised/
Supervised



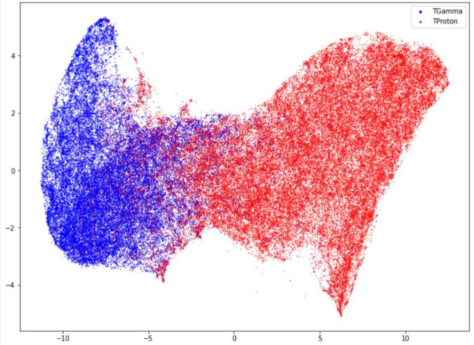


Data

Calibration

Visualization and Clustering

➜ Reducing dimensions from N to 2 or 3 for Data Visualization



Useful for:

- Visualizing how events cluster in 2D/3D to take decisions, as for instance filtering
- Comparing real data with simulations
- Identifying strange (unexpected) behaviour in data for monitoring

Preparation

Conclusions on Machine Learning for Very High Energy Astronomy

- “Classic” Machine Learning - a big YES, especially in classification tasks
- Deep Learning in gamma-ray astronomy - a big MAYBE - more research needed, especially to put it in massive production of results
- Very useful in neutrino telescopes due to the uncertainty of the refractive medium response
- A mixture of approaches, as for instance that given by Scientific Machine Learning, where you would mix previous knowledge with knowledge learned by the Neural Network might be more suited
- Machine Learning offers a wealth of possible improvements in data analysis
- There is no magic recipe:
  - For classification problems, feature-based ML largely sufficient
  - For Energy reconstruction DL helping, especially at low energies
  - For the regression of the arrival direction: hard to beat standard likelihood minimization
- New frontiers: unsupervised learning, simulation refinement