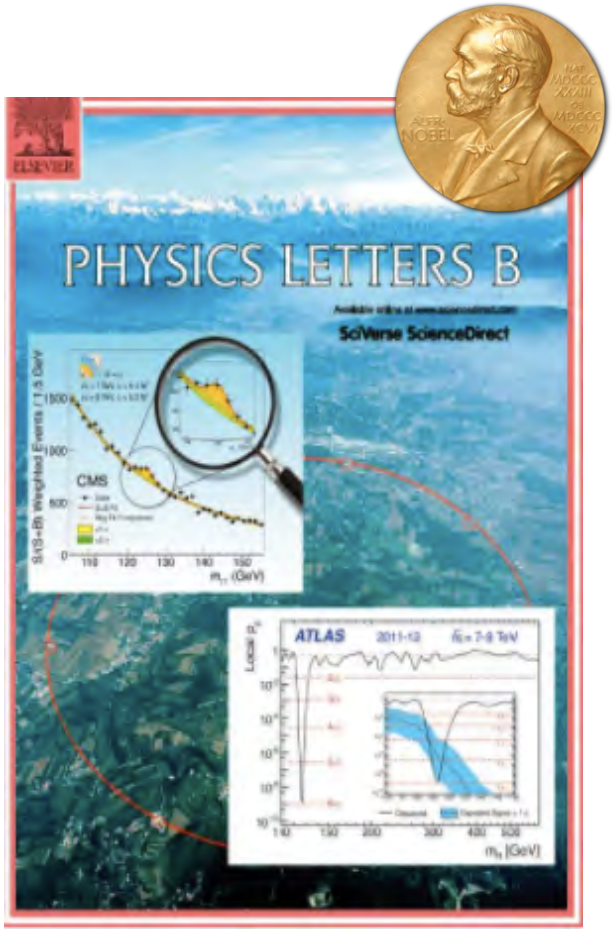
AI in Fisica delle Alte Energie

Prof. A. Zoccoli

Machine Learning nella scoperta del bosone di Higgs

Contributo alla ricerca del bosone di Higgs (2012)

Scoperta attesa - senza ML - intorno al 2015/16 (!)

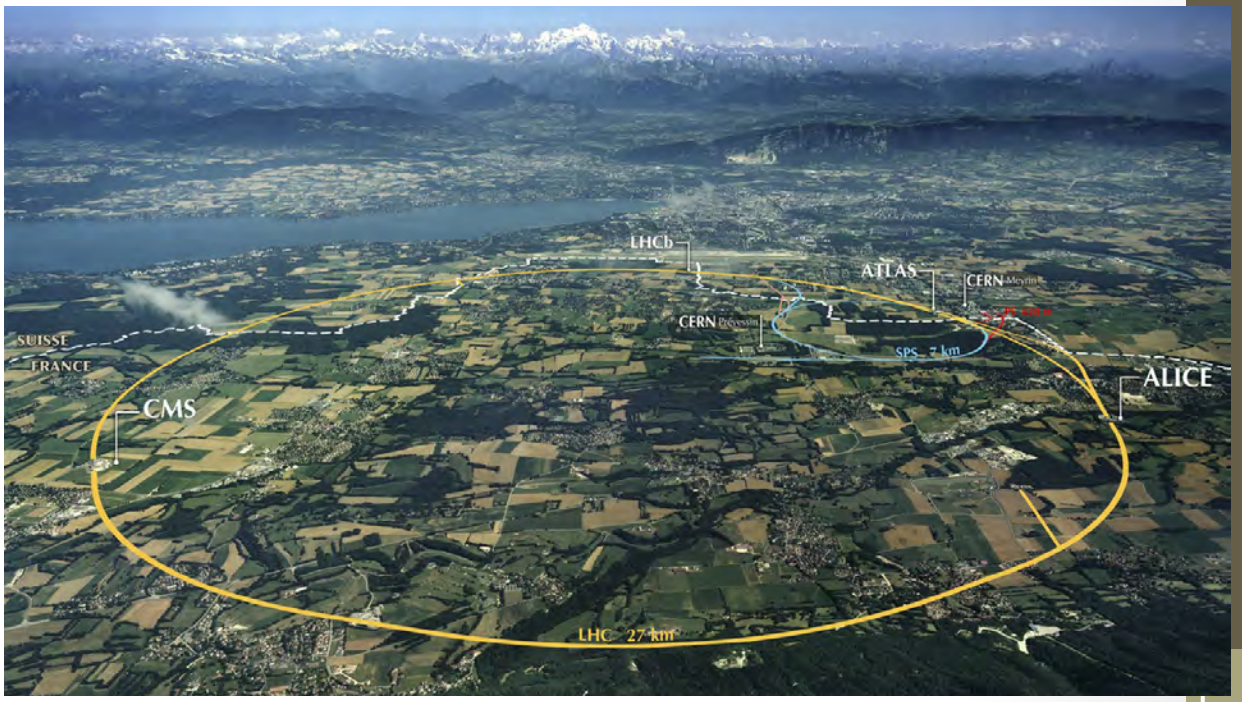


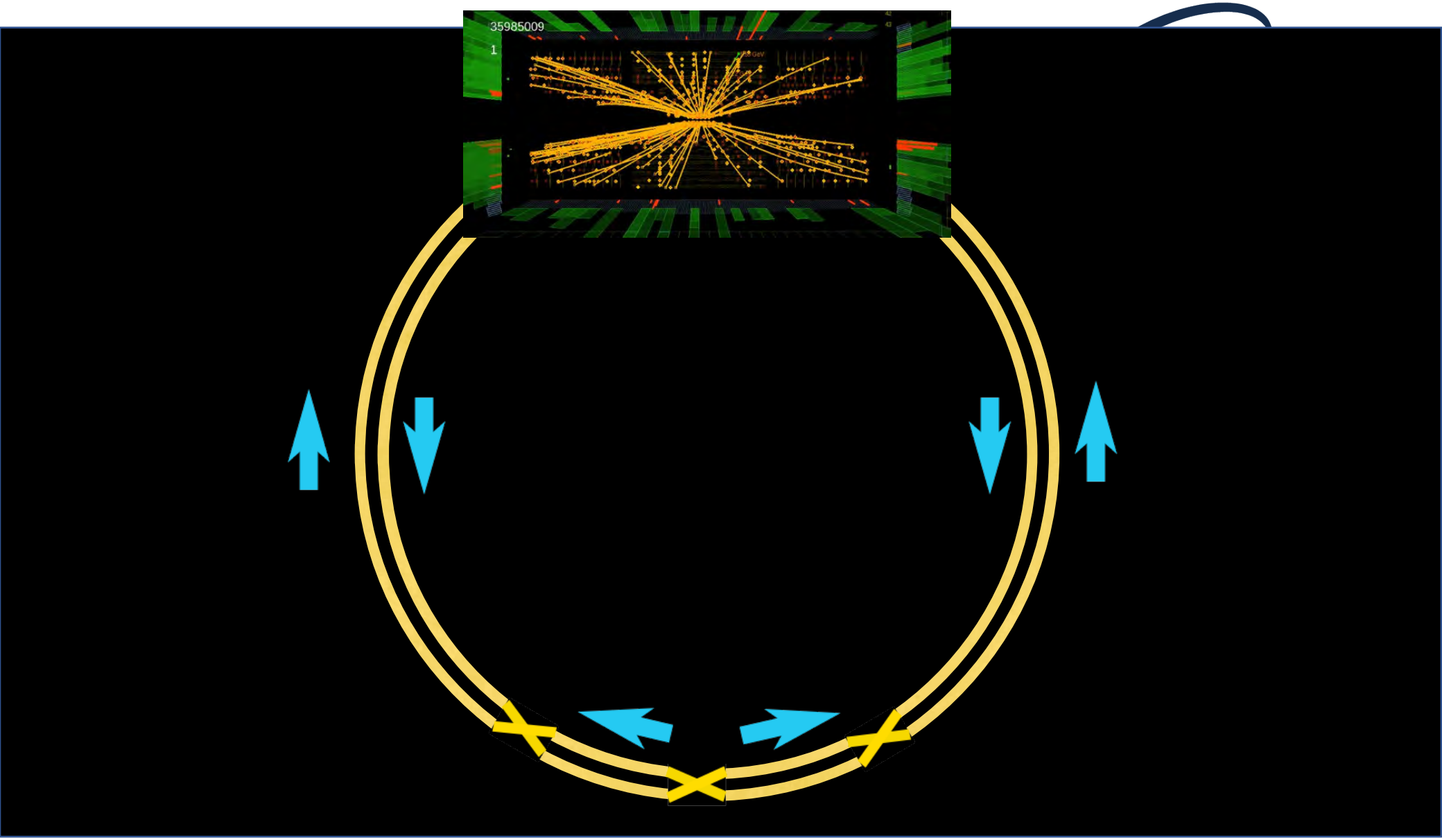
E.g. analisi dell'esperimento CMS nel canale H→ ɣ ɣ fortemente dipendente da **algoritmi di ML tradizionali**, come **boosted decision trees**, nel miglioramento delle risoluzioni sperimentali e nella selezione degli eventi

IL CERN di Ginevra

In order to create high energy densities we accelerate particles in opposite directions and make them collide one against the other

The CERN LHC accelerates protons. It has 27 km of circumference and is located in a tunnel about 100 m underground in the Geneva area

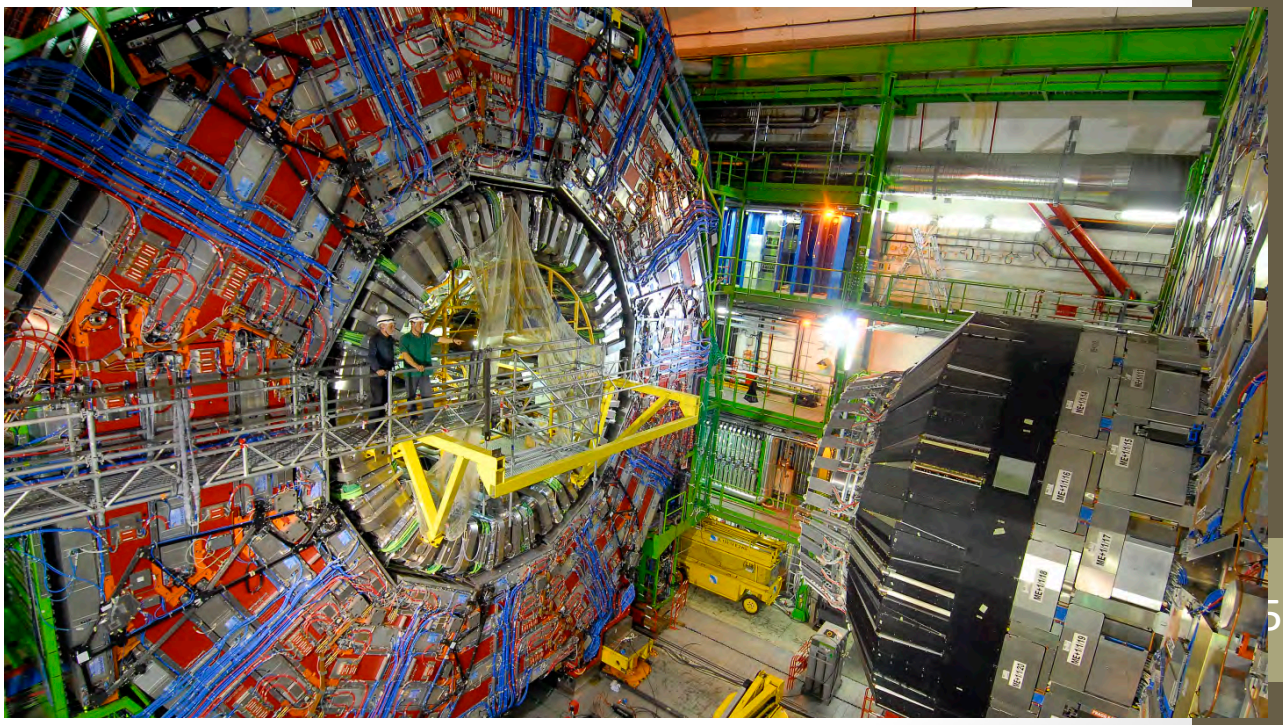



35985009
1

Particle detectors

Around collision points we have built particle detectors that can "see" the particle produced in the proton collision so that we can understand what happened.

Detectors have about 100 million channels that are acquired at each collision

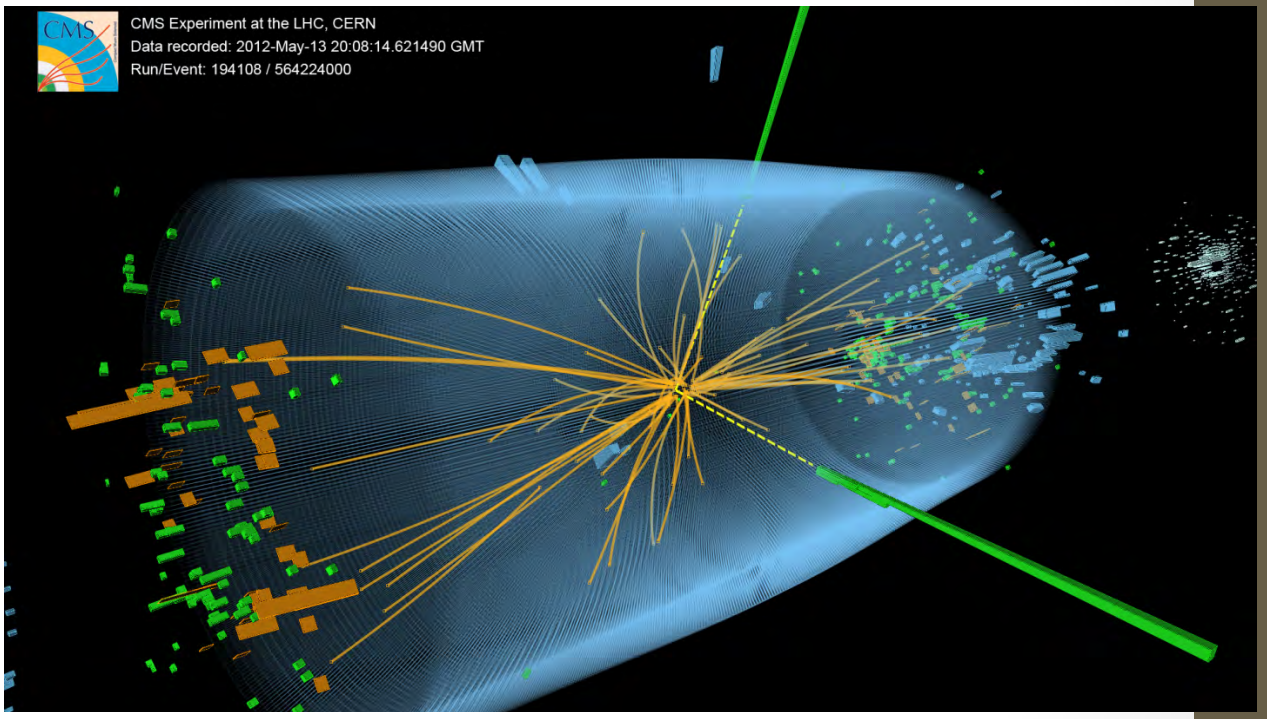


Collision events

We call "event" a single crossing of the proton bunches in the detector area.

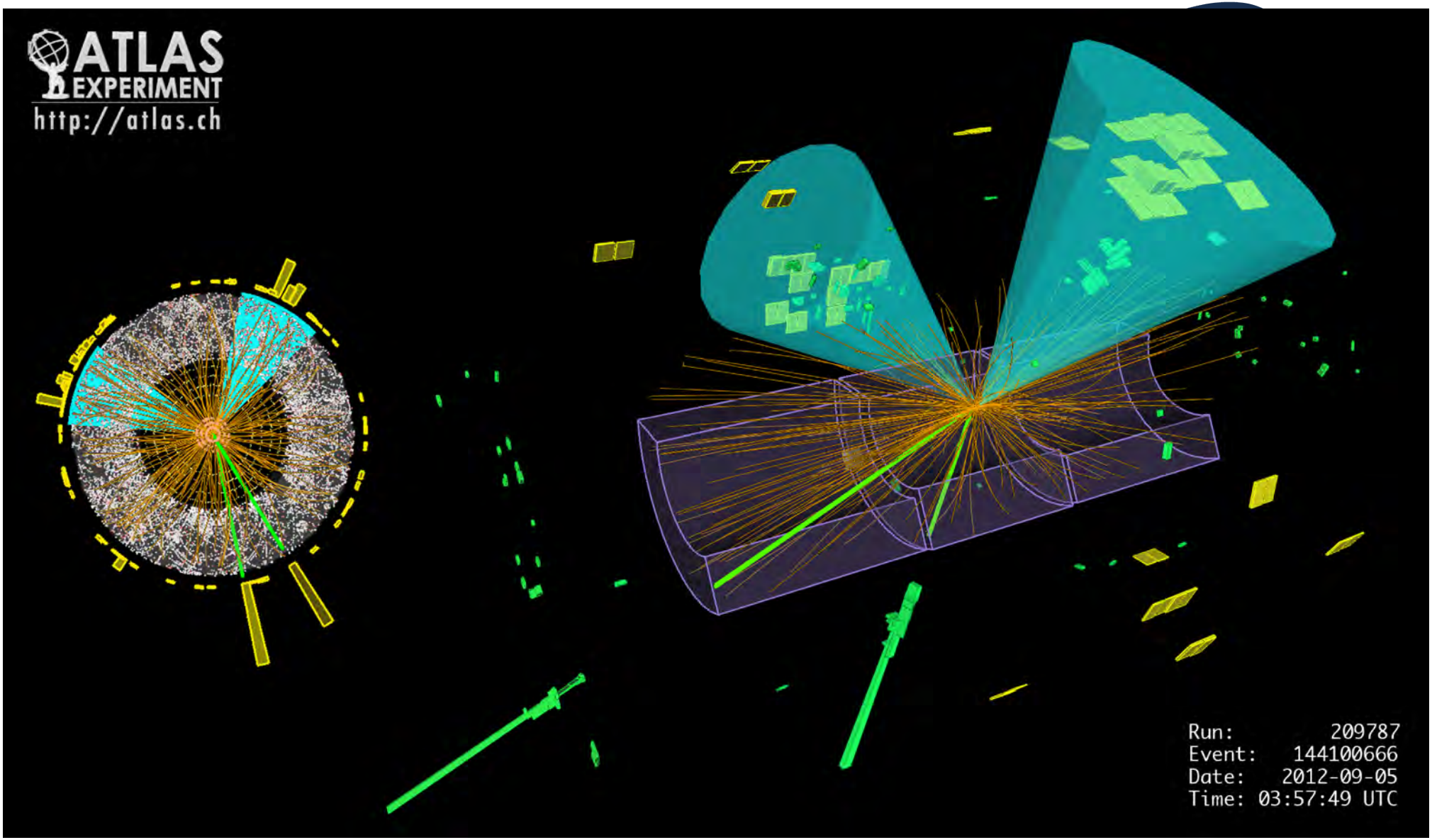
For each event we reconstruct the particles produced in the collisions.

There are 40 millions crossings per second



Some reference numbers

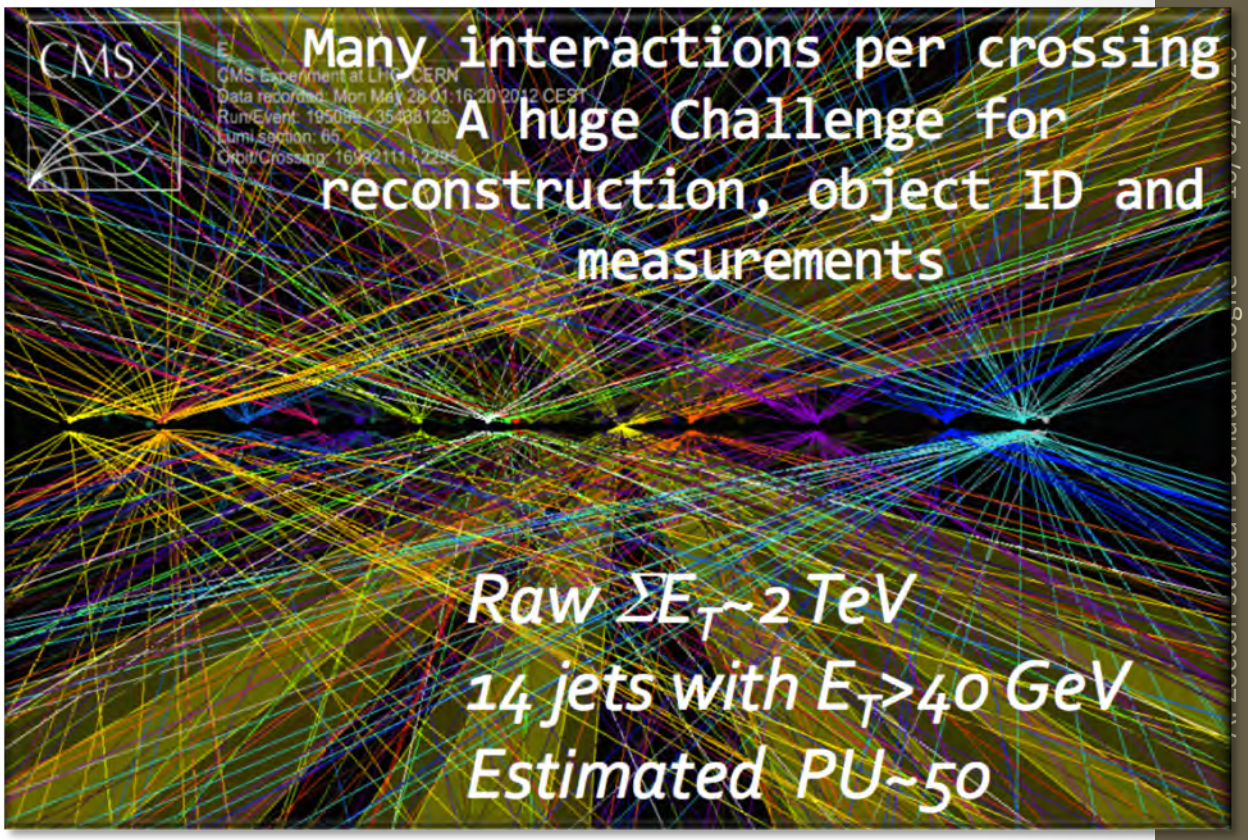
- *600 million collisions every second*
- *Only 1 in a million collisions is of interest*
- *Fast electronic preselection passes 1 out of 10 000 events and stores them on computer memory*
- *100 GB/s transferred to the experiment computing farm*
- *15 000 processor cores select 1 out of 100 of the remaining events*


ATLAS
EXPERIMENT
http://atlas.ch
Run: 209787
Event: 144100666
Date: 2012-09-05
Time: 03:57:49 UTC

Pile-up

If you're wondering why a bunch crossing rate of 40 MHz produces 600 collisions per second:

Every bunch crossing (event) there are on average 15 p-p collisions (AKA pileup)



Pileup is increased in 2017 to 50 and eventually to more than 150 in HL-LHC

…Computing infrastructure

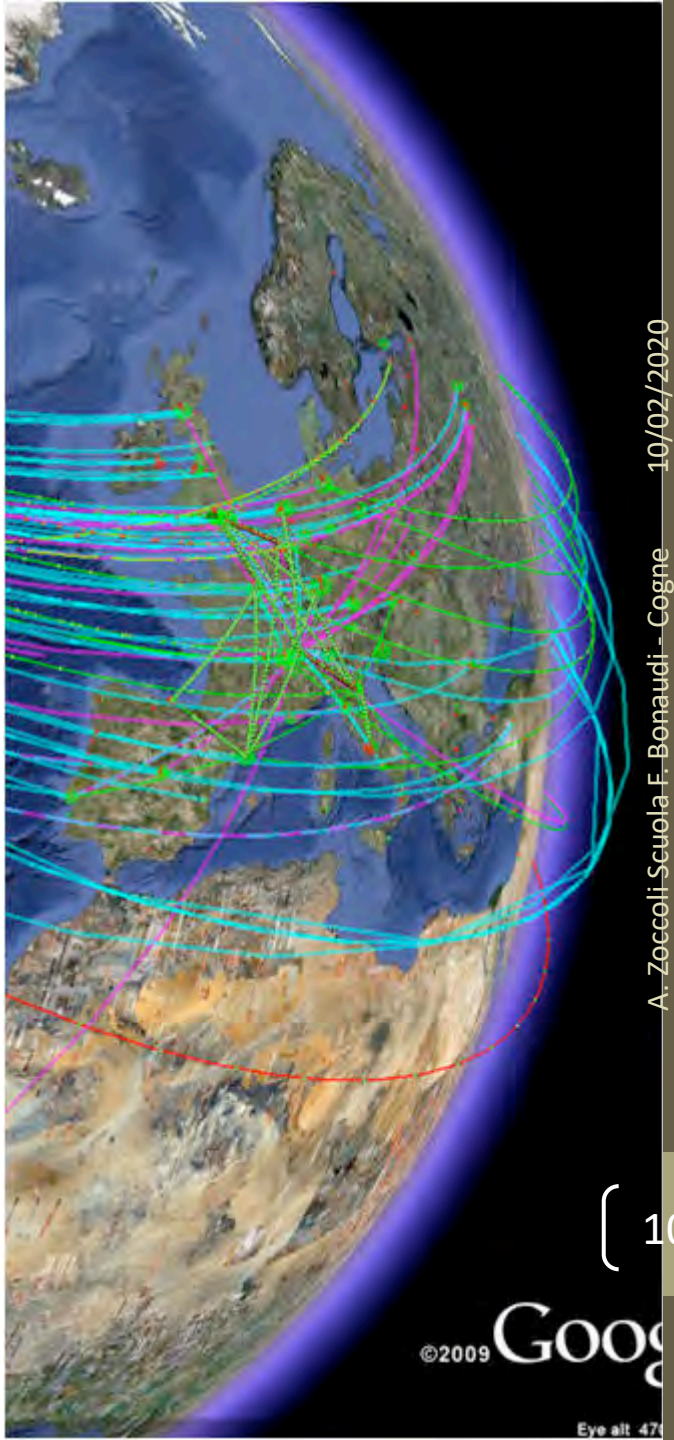
Global resources for 2020 are:

- *1.000.000 processor cores*
- *500.000 TB disk*
- *590.000 TB tape*
- *Dedicated network connections (from multiples of 10 Gb/s to multiples of 100 Gb/s)*

…and more available in collaborating institutes

More than 180 data centres in over 35 countries

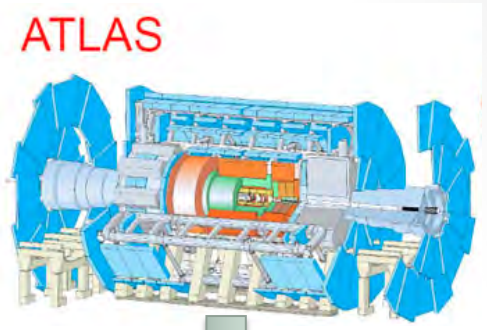
More than 8000 analysts all over the world



LHC Real data

LHC collisions

Decay of unstable particles

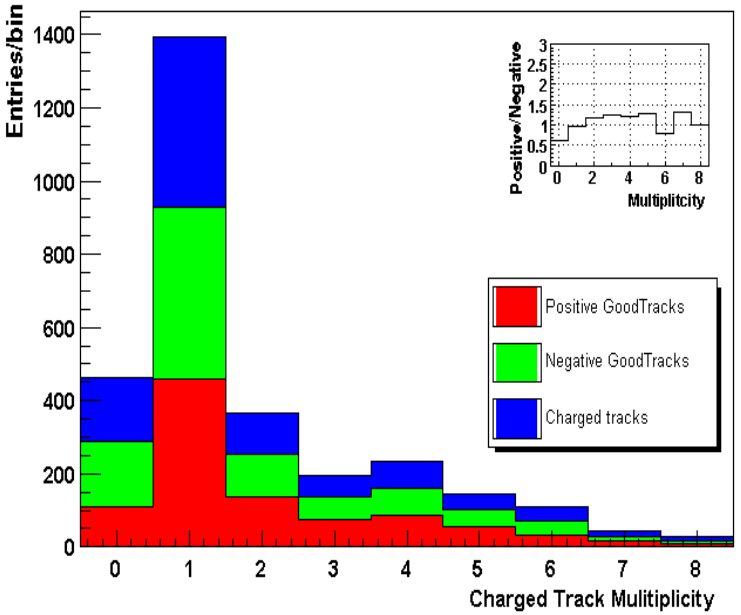


Detector electronics

Trigger

Reconstruction

Analysis



LHC Simulation

Not just real data form detectors!

Since it is not possible to use analytical solutions of physic processes going from the proton interactions to the final state particles, we use simulations based on Monte Carlo techniques

Events are generated according to theoretical models and then simulated in order to reproduce the detector behaviour and then treated in the same way of the real data

The simulated data sample is 1 to 2 times the real data sample

LHC Simulated data

Theoretical model

Simulation of decays of unstable particles

Pythia,…

Simulation of interactions particle-detector

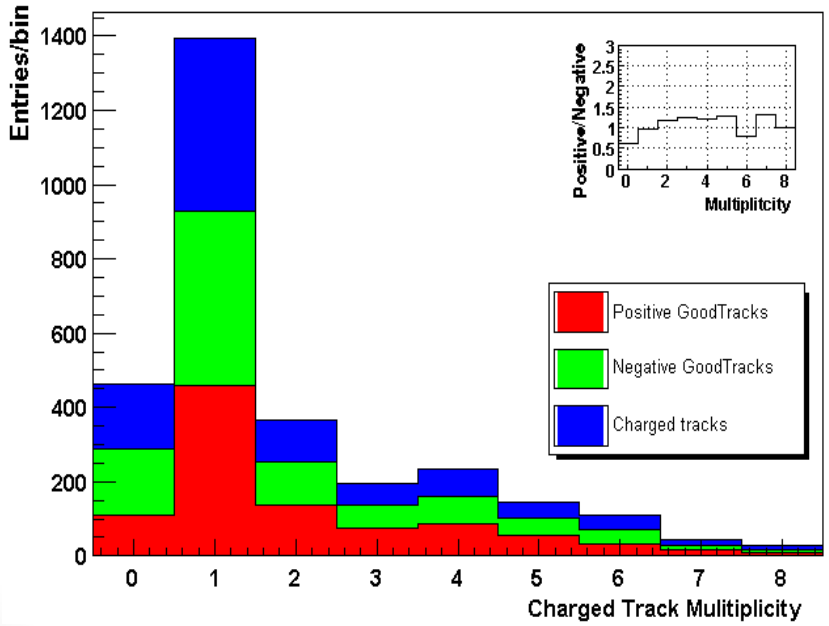
Geant4, …

Simulation of detector electronics

Trigger

Recontruction

Analysis



AI in Fisica delle Alte Energie (HEP)

Fino a pochi anni fa: ML “tradizionale”..

Use of field-specific knowledge for feature engineering
i.e. use physicist-designed high-level features as input to shallow algorithms

Data is organized differently from other typical ML tasks

Our data is very sparse, already zero suppressed,

Mapping our tasks to standard technologies often implies information loss

We try to adapt standard techniques to out use cases (recasting)

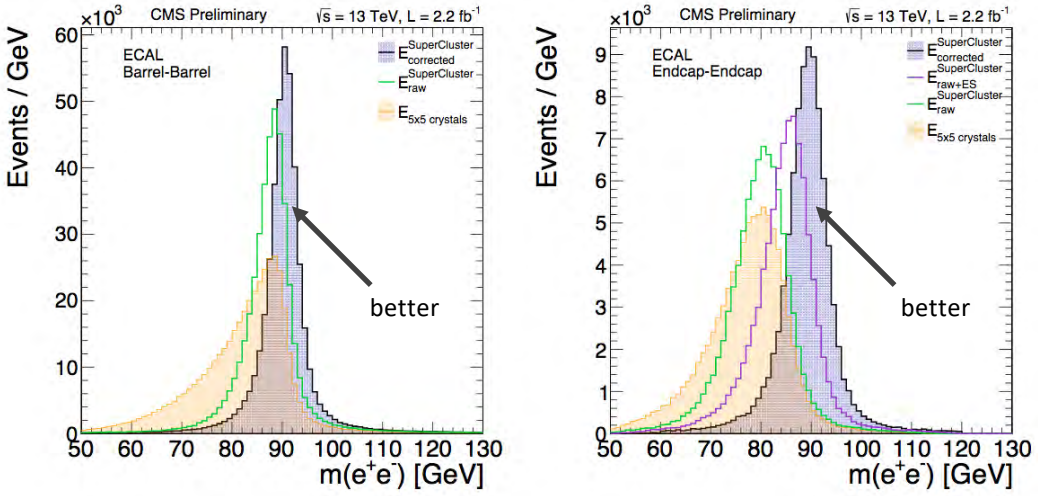
Examples of ML in HEP

- Minimization algorithms reimplemented with TensorFlow
- Convolutional or Recurrent Neural Networks used for jet tagging → Example 1
- Deep Neural Networks used for generic Signal/background discrimination
  - e.g. glitches detection in gravitational wave searches
- Generative Adversarial Networks used for event simulation

Particle properties: energy resolution

Using ML to improve the determination of particle properties is now commonplace in **all LHC experiments**

- E.g. energy deposited in calorimeters is recorded by many sensors, which are clustered to reconstruct the original particle energy. **CMS** is training **BDT**s to learn corrections using all information available in the various calorimeter sensors - thus resulting in a sizeable improvement in resolution

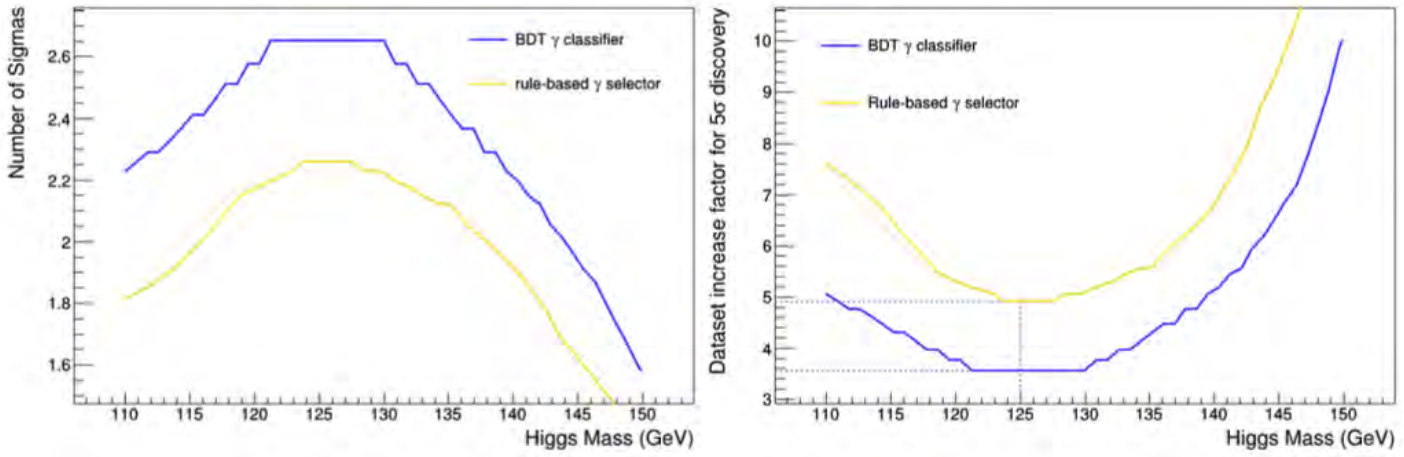


[2015 ECAL detector performance plots, CMS-DP-2015-057. Copyright CERN, reused with permission]

Improvements to the Z→e+e- energy scale and resolution from the incorporation of more sophisticated clustering and cluster correction algorithms (energy sum over the seed 5x5 crystal matrix, bremsstrahlung recovery using supercluster, inclusion of pre-shower energy, **energy correction using a multivariate algorithm**)

Discovery of the Higgs boson

ML played a key role in the discovery of the Higgs boson, especially in the diphoton analysis by CMS where ML (used to improve the resolution and to select/categorize events) increased the sensitivity by roughly the equivalent of collecting ~50% more data.



We were not supposed to discover the Higgs boson as early as 2012

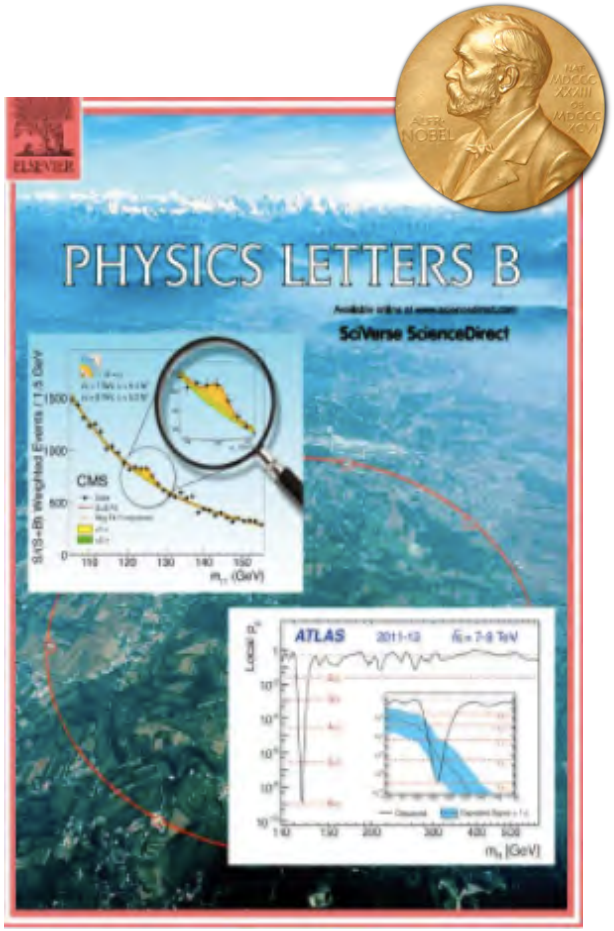
- Given how machine progressed, we expected discovery by end 2015 / mid 2016

We made it earlier thanks (also) to ML

Machine Learning nella scoperta del bosone di Higgs

Contributo alla ricerca del bosone di Higgs (2012)

Scoperta attesa - senza ML - intorno al 2015/16 (!)

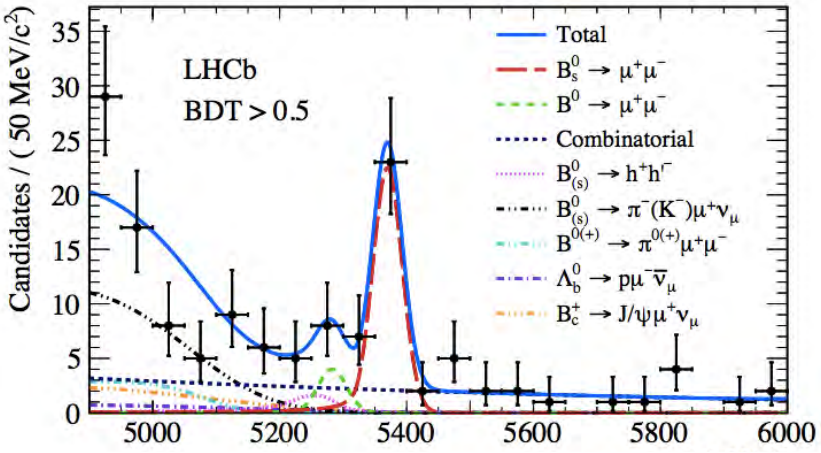


E.g. analisi dell'esperimento CMS nel canale H→ ɣ ɣ fortemente dipendente da **algoritmi di ML tradizionali**, come **boosted decision trees**, nel miglioramento delle risoluzioni sperimentali e nella selezione degli eventi

High-precision tests of the SM

CMS and LHCb were the first to find evidence for the $B^0_s \to \mu^+\mu^-$ decay with a combined analysis (as rare as ~ 1 / 300 billion pp collisions..)

- BDTs used to reduce the dimensionality of the feature space - excluding the mass - to 1 dimension, then an analysis was performed of the mass spectra across bins of BDT response
- decay rate observed is consistent with SM prediction with a precision of ~25%, placing stringent constraints on many proposed extensions to the SM
- To obtain the same sensitivity without ML by LHCb as a single experiment would have required ~4x more data. Just one of many examples of high-precision tests of the SM at the LHC where ML can dramatically increase the power of the measurement



Mass distribution of the selected $B^0 \to \mu^+\mu^-$ candidates with BDT > 0.5.

[arXiv: 1703.05747]

In Fisica delle Alte Energie (HEP)

Fino a pochi anni fa: ML “tradizionale”..

use of field-specific knowledge for feature engineering
i.e. use physicist-designed high-level features as input to shallow algorithms

Da qualche anno: Reti neurali (con molteplici architetture)

use of full high-dimensional feature space to train cutting-edge ML algorithms (e.g. DNNs). As in computer vision and NLP, growing effort in HEP too to skip the feature-engineering step

Simulation

Physics-based full simulation modelling in HEP (with GEANT 4 as the state of the art) is very computationally demanding

- e.g. for LHC, the large samples to be generated for future experimental runs and the increase in luminosity will exacerbate the problem, prohibitive also for GEANT

This already sparked the development of approximate, Fast Simulation solutions to mitigate this computational complexity - especially relevant in calorimeter showers simulations

Promising alternatives for Fast Simulation may be built on recent progress in high fidelity fast generative models

- e.g. Generative Adversarial Networks (**GAN**s) and Variational AutoEncoders (**VAE**s)
- ability to sample high dimensional feature distributions by learning from existing data samples

A simplified first attempt at using such techniques in simulation saw orders of magnitude increase in speed over existing Fast Simulation techniques, of which **all HEP experiments** would largely benefit

- not yet reached the required accuracy, though

Perhaps more towards >2020, but promising.

Trigger

Crucial trade-off in algorithm complexity and performance under strict inference time constraints

E.g. ATLAS/CMS each only keep about 1 in every 100 000 events, and yet their data samples are each still about 20 PB/yr

- ML algorithms have already been used very successfully for rapid event characterisation
- adoption depth vary across experiments, but the increasing event complexity at HL-LHC will require more sophisticated ML solutions and its expansion to more trigger levels

A critical part of this work will be to understand which ML techniques allow us to maximally exploit future computing architectures

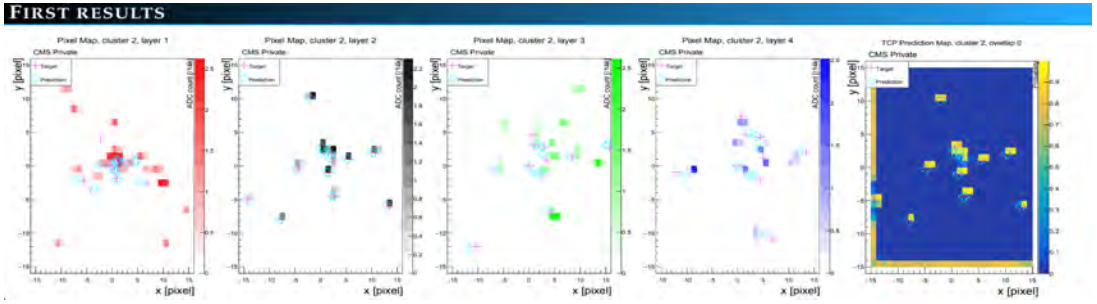
Tracking

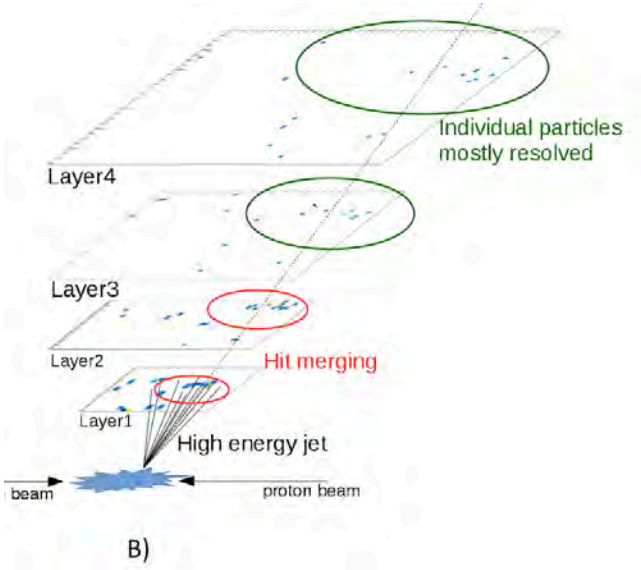
Pattern recognition has always been a computationally challenging step

- e.g. the HL-LHC environment makes it an extremely challenging task

Adequate ML techniques may provide a solution that scales linearly with LHC intensity.

Several efforts in the HEP community have started to investigate sophisticated ML algorithms for track pattern recognition on many-core processors.





Vertices, jets reconstruction

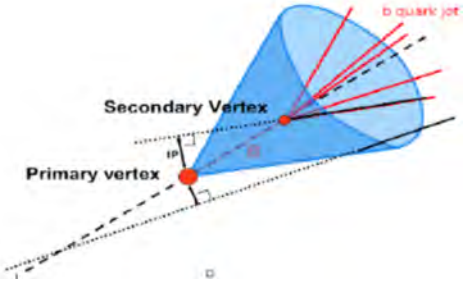
Graph networks are probably a game changer for us

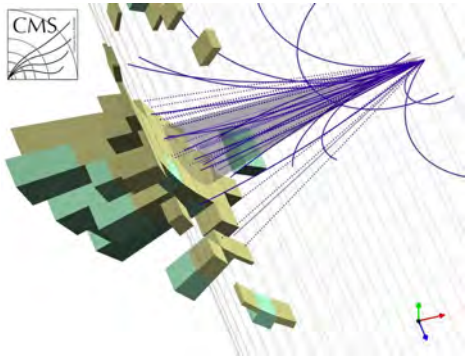
Vertexing

- Input for vertexing is very nicely represented with a graph (e.g. each node is a track and each edge has properties such as track to track distance)

Jet clustering and tagging

- Rather than sequential processing we could process the particle graph (each particle is connected with neighbor particles with some QCD motivated metric) and have hard process Feynman diagram (a graph!) as the target



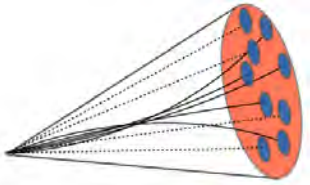


CNN (Convolutional Neural Networks)

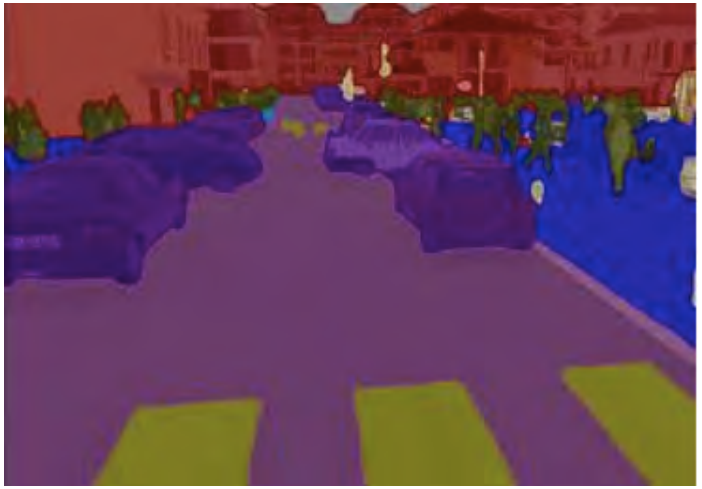
(reti neurali convoluzionali)

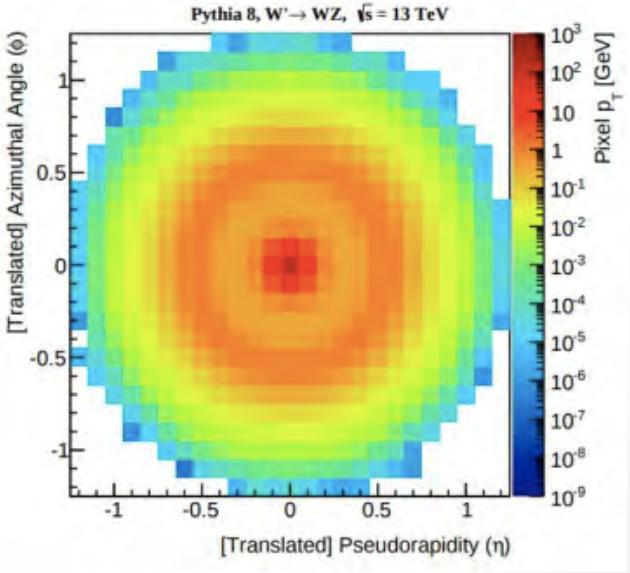
Metodo: livelli di convoluzione che estraggono (filtri) caratteristiche delle immagini

Fisica delle Alte Energie all'INFN:
3D imaging nei rivelatori,
classificazione di eventi, ..



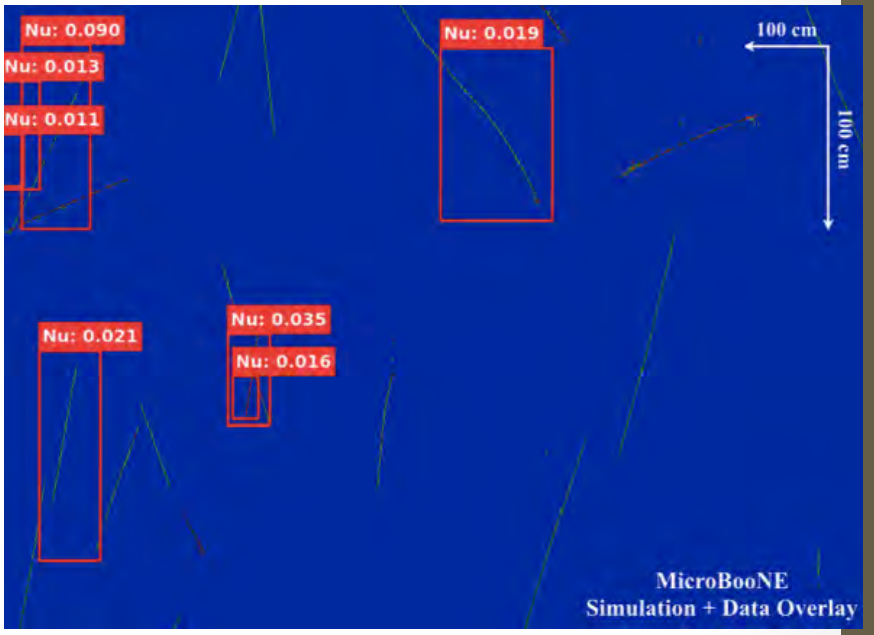
Industria:
Largo utilizzo in "computer vision"





Somiglianze notevoli tra applicazioni diverse (**Fisica** e non)

Rilevazione di **aeroporti** da immagini satellitari (metodo: **CNN**)



Rilevazione di **neutrini** su eventi di cosmic background (metodo: **CNN**)

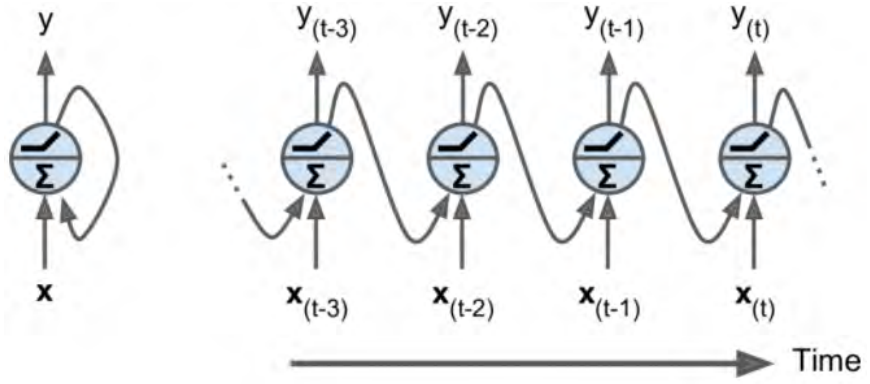
RNN (Recurrent Neural Networks)

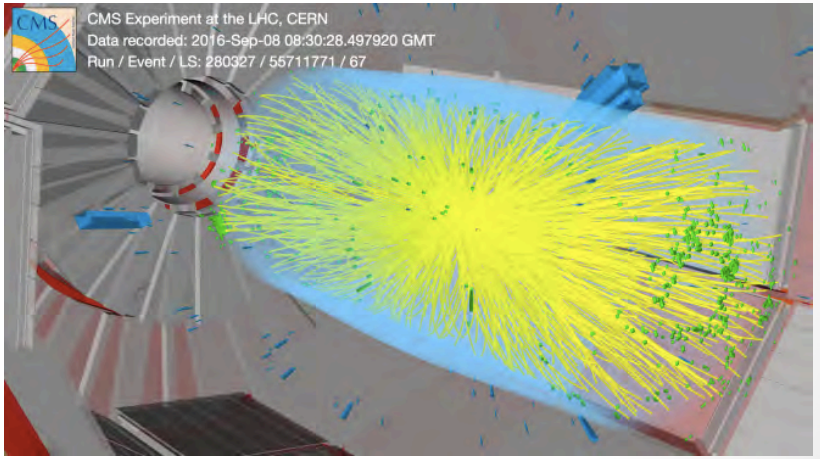
(reti neurali ricorrenti)

Metodo: aggiunta di connessioni di feedback, più istanti temporali considerati insieme

Industria:
gestione di “time series”
(audio, video, natural language processing)

Fisica delle Alte Energie all’INFN:
classificatori capaci di processare segnali complessi, o input di lunghezza variabile nel tempo (tracce, particelle, ..)





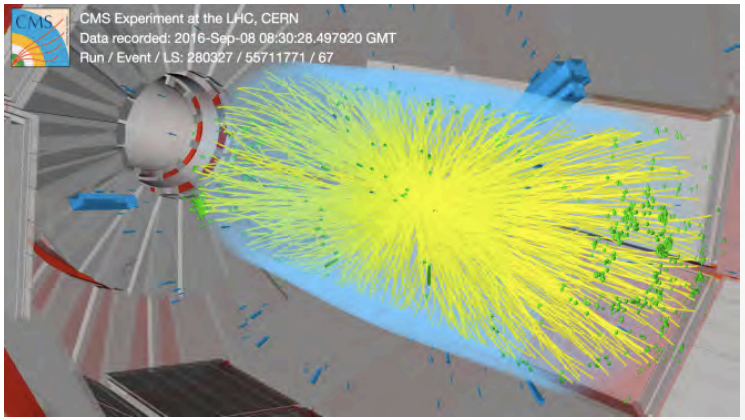
GAN (Generative Adversarial Networks)

(reti generative antagoniste)

Metodo: “training” di 2 reti neurali in competizione, per imparare come generare nuovi dati con la stessa distribuzione di quelli usati in fase di training

Industria:
Image editing, Data generation, Security, ..

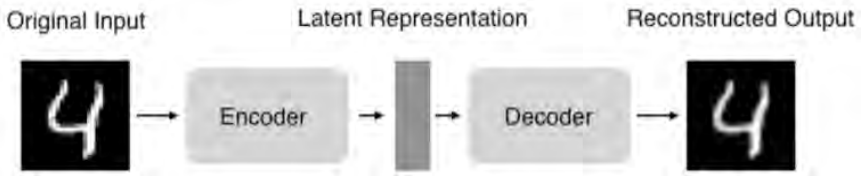
Fisica delle Alte Energie all’INFN: Generare la risposta dei rivelatori (promettente alternativa ai tradizionali programmi di simulazione)



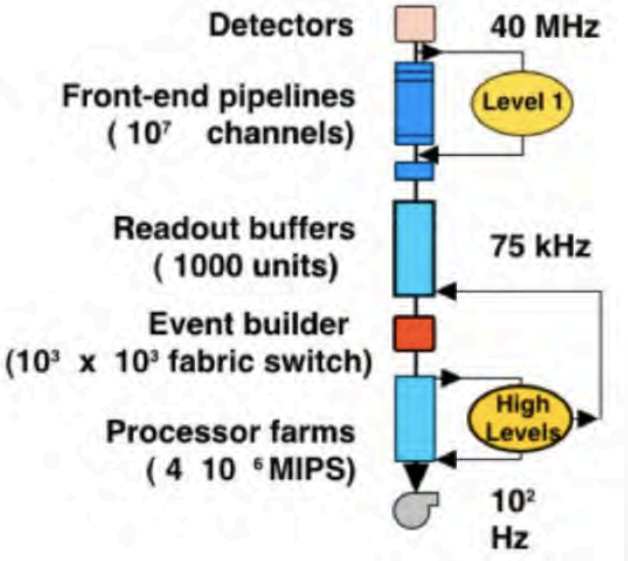
(V)AE (Variational Autoencoders)

Metodo: comprimere i dati (“encoder”) in uno spazio di variabili “latenti” e ricostruirlo (“decoder") generando così nuovi dati (NB: è **unsupervised ML**)

Industria:
Riduzione della dimensionalità, denoising..

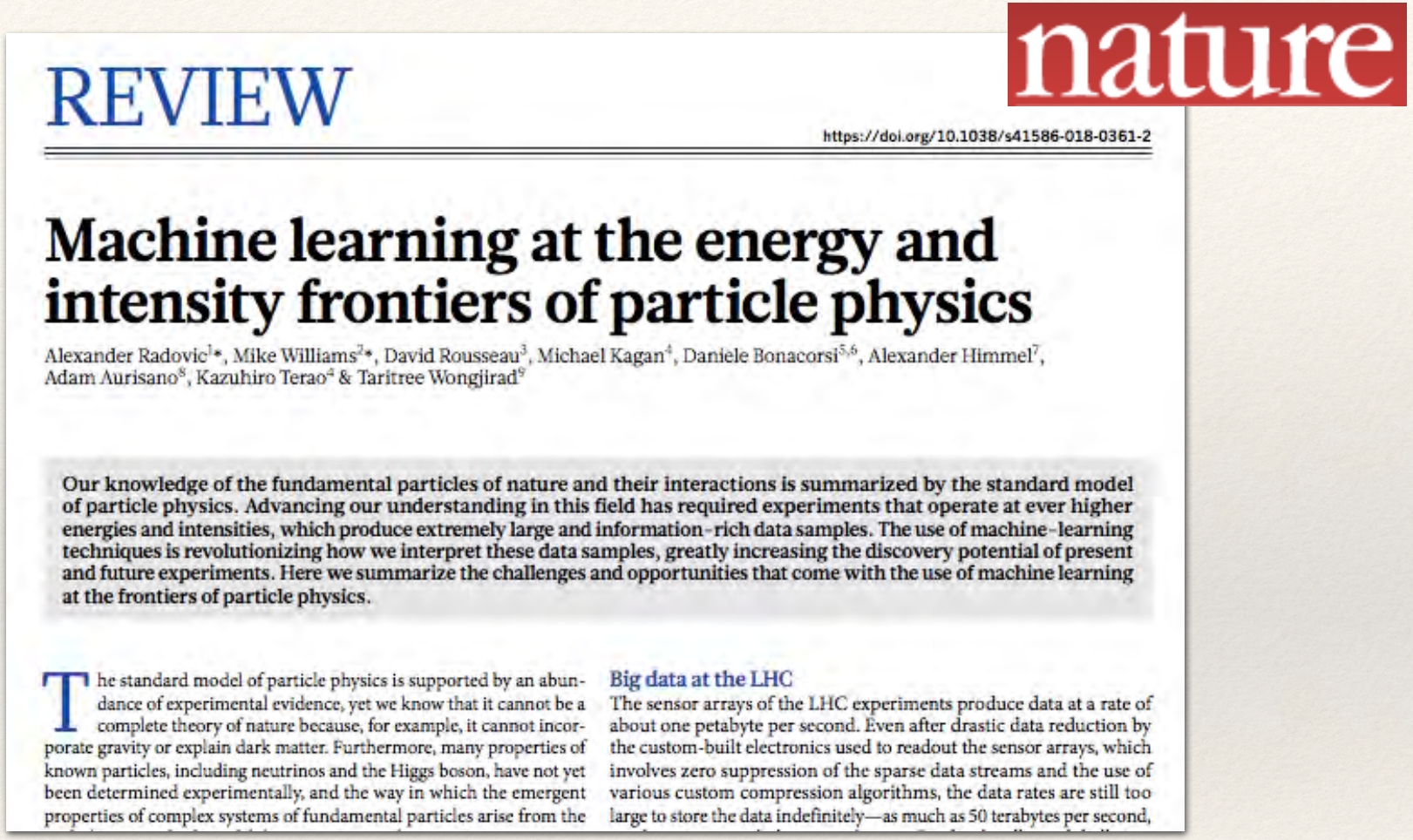


Fisica delle Alte Energie all’INFN: Isolare potenziale nuova fisica come eventi “outlier” di distribuzioni note



Grazie !

(quick) reading material



nature

REVIEW

https://doi.org/10.1038/s41586-018-0361-2

Machine learning at the energy and intensity frontiers of particle physics

Alexander Radovic[1]*, Mike Williams[2]*, David Rousseau[3], Michael Kagan[4], Daniele Bonacorsi[5,6], Alexander Himmel[7], Adam Aurisano[8], Kazuhiro Terao[4] & Taritree Wongjirad[9]

Our knowledge of the fundamental particles of nature and their interactions is summarized by the standard model of particle physics. Advancing our understanding in this field has required experiments that operate at ever higher energies and intensities, which produce extremely large and information-rich data samples. The use of machine-learning techniques is revolutionizing how we interpret these data samples, greatly increasing the discovery potential of present and future experiments. Here we summarize the challenges and opportunities that come with the use of machine learning at the frontiers of particle physics.

The standard model of particle physics is supported by an abundance of experimental evidence, yet we know that it cannot be a complete theory of nature because, for example, it cannot incorporate gravity or explain dark matter. Furthermore, many properties of known particles, including neutrinos and the Higgs boson, have not yet been determined experimentally, and the way in which the emergent properties of complex systems of fundamental particles arise from the

Big data at the LHC

The sensor arrays of the LHC experiments produce data at a rate of about one petabyte per second. Even after drastic data reduction by the custom-built electronics used to readout the sensor arrays, which involves zero suppression of the sparse data streams and the use of various custom compression algorithms, the data rates are still too large to store the data indefinitely—as much as 50 terabytes per second,

https://www.nature.com/articles/s41586-018-0361-2