From Quarks to Cosmos with AI - Carnegie-Mellon University, July 15, 2021
https://events.mcs.cmu.edu/qtc2021/

Applications of Differentiable Programming to Fundamental Physics Research: Status and Perspectives

Tommaso Dorigo

INFN, Sezione di Padova

MODE

INFN
Istituto Nazionale di Fisica Nucleare

Abstract

Take the chain rule of differential calculus, model your system with continuous functions, add overparametrization and an effective way to navigate stochastically through the parameter space in search of an extremum of an utility function, and you have all it takes to find an optimal solution to even the hardest optimization problem. Deep learning, nowadays “differentiable programming”, is boosting our reach to previously intractable problems.

I will look at the status of applications of differentiable programming in research in particle physics and related areas, and make a few observations of where we are heading.

Contents of this lecture

- The road to AI: a 5-slide AI history, and a few crucial definitions
- Future challenges of fundamental research: the status of deep learning for particle physics today
- Informing machines of our real objectives: differentiable programming solutions
- The frontier: end-to-end optimization of experimental design

What is Intelligence ?

Before we discuss *artificial* intelligence, and how we are getting there today, we should agree on what **intelligence** is – and what isn't.

- The term «intelligence» comes from Latin "*intelligo*"

  → to comprehend, to perceive

…but that does not help much. Consider the literature for help: we find that notable definitions differ significantly, also in relation to what can be general and what is specific of human beings:

- "*The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment*" [Wechsler 1944]
- "*The unique propensity of human beings to change or modify the structure of their cognitive functioning to adapt to the changing demands of a life situation*" [Feuerstein 1990]

But also, and more useful to us, are more abstract definitions such as:

- "*Intelligence measures an agent's ability to achieve goals in a wide range of environments*" [Legg 2007]
- "*Intelligence is goal-directed adaptive behavior*" [Sternberg 1982]

What is Artificial Intelligence ?

Curiously, a precise definition of Artificial Intelligence is not less challenging than that of Intelligence at large

- AI: *"Intelligence demonstrated by machines"* – devices that have the ability to perceive the environment and take actions toward achieving some goal

  The above puts automatic wipers and thermostats in the AI category…

Patricia McCorduck describes the AI effect: «*AI is what hasn't been done yet*» [Mc Corduck 2004]

→We continue to raise the bar as we get accustomed to progress in artificial intelligence research:

- Self-driving cars? Just a smart NN with lots of input data
- Speech recognition? Only an algorithm, not real AI
- Facebook targeted ads, gmail text completion? Specialized data-crunching, little more than nearest-neighbor techniques

At the basis of this is the perception that we, as sentient beings, «**know**» what we are doing. But we, too, are machines!

A Five-Slide History of AI : 1 – the Origins

The idea of artificial intelligence in ancient history is only found connected with that of creating life from unanimated matter:

- *Talos*, an automated guardian of Crete
- *Galatea*, a statue made by king and sculptor Pygmalion
- The *Golem*, from Jewish folklore

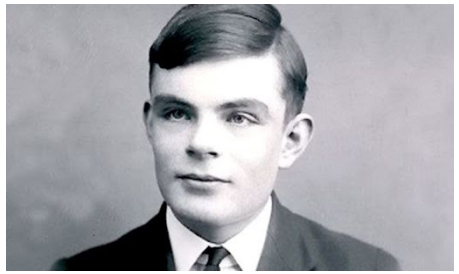
A different path emerged much later, with the idea of *automata* and real examples built by craftsmen

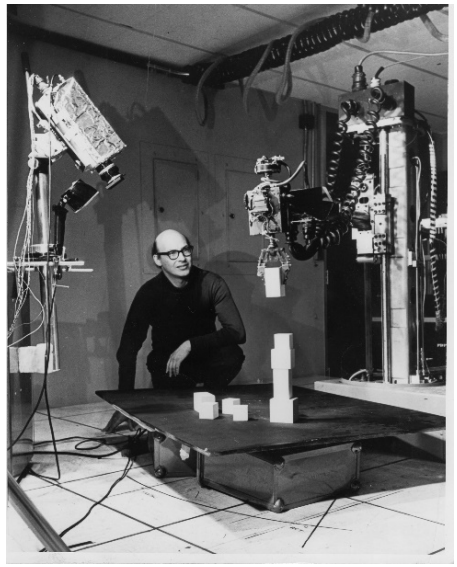
A significant development of the underpinnings of AI was produced by SciFi writers in the late XIX and early XX centuries

The study of mathematical logic provided even more concrete bases, with a demonstration that any form of mathematical reasoning could be mechanized [Church 1936; Turing 1936]

AI History: 2 – Before the Birth

Important milestones toward a recognition of the possibility of artificial thinking:

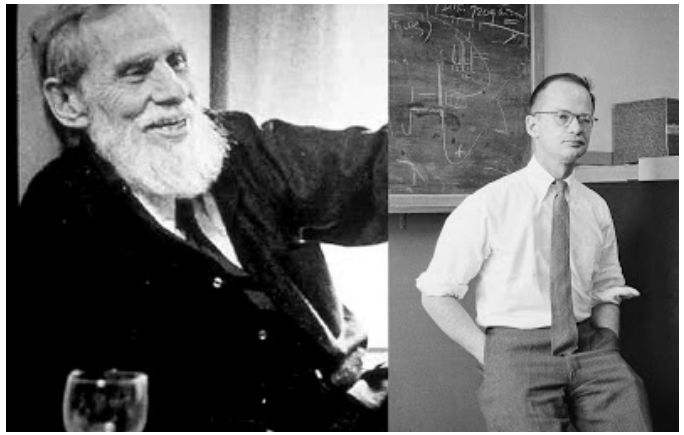


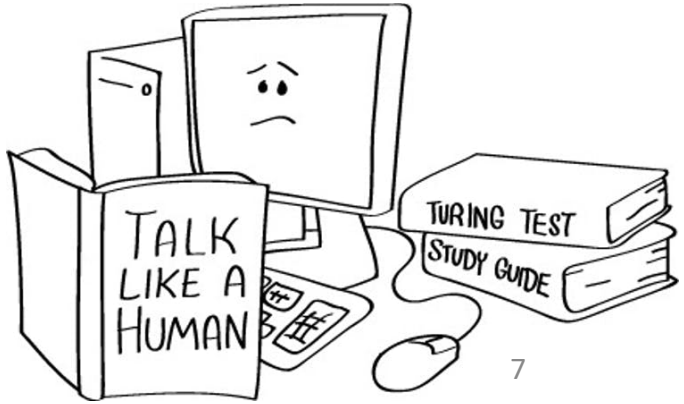


- In 1943 Walter Pitts and Warren McCullock showed how artificial neurons could be arranged in networks to perform simple logic [Pitts 1943](the basic unit term «psychon» never caught up)

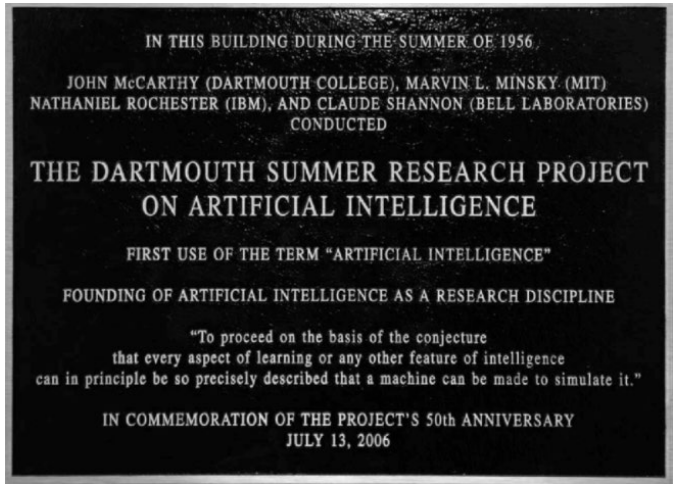
- In 1950 Turing laid down the bases to discuss about AI, proposing the Turing test («imitation game») to express how intelligent thinking could be defined [Turing 1950]

- Marvin Minsky took inspiration from Pitts and McCullock to build the first neural network in 1951





AI History: 3 – The Modern Origins



In 1956 a landmark workshop was organized in Dartmouth by Minsky and John McCarthy. At the workshop were laid the bases of fundamental research in **artificial intelligence** –a name coined there

The paradigm that was born in Dartmouth was that "*every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it*"

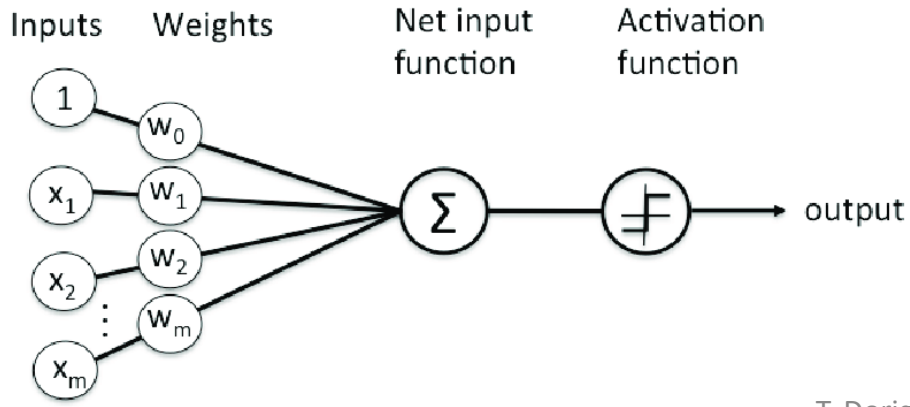
Simulation!

- A sham object: **counterfeit** [Merriam-Webster]
- The action of pretending: **deception** [Oxford Languages]
- The production of a **computer model** of something [OL]

AI History: 4 – Summers …

Following Dartmouth, AI received large funding and ideas thrived. New applications were found and problems solved; enthusiasm reigned:

- 1958: "*Within ten years a digital computer will be the world's chess champion; […] within ten years a digital computer will discover and prove an important new mathematical theorem*." [Simon 1958]
- 1967: "*Within a generation … the problem of creating 'artificial intelligence' will substantially be solved*." [Minsky 1967]



Frank Rosenblatt invented in 1958 the perceptron, a single-layer NN element [Rosenblatt 1958], and predicted this would bring a breakthrough: "*a perceptron may eventually be able to learn, make decisions, and translate languages*".

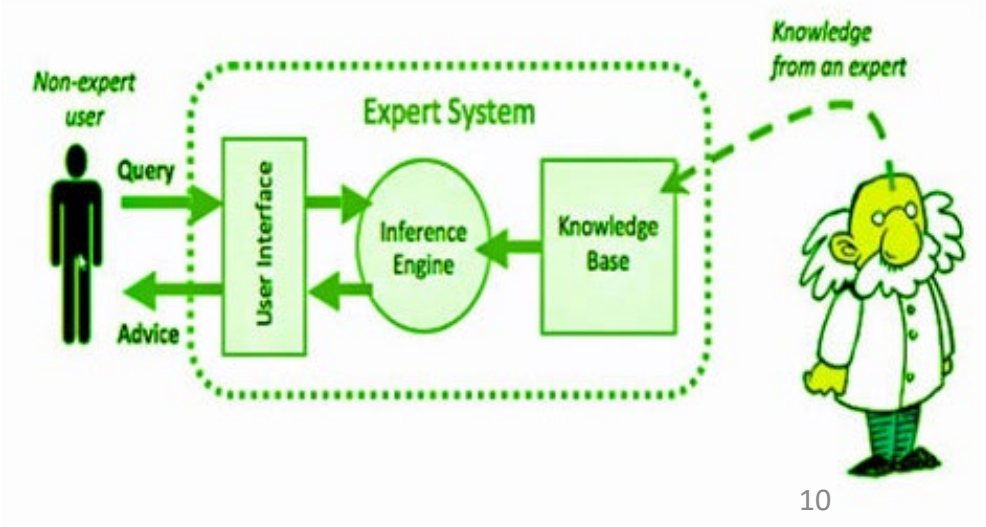
… And Winters

A 1969 book by Minsky and Papert [Minsky 1969] threw serious doubt on the capability of such systems. This had a huge influence on the general perception of AI.

In the early 1970s the lack of results given the high expectations, and the lack of sufficient computing power to handle anything but toy applications, started a period of almost zero funding and academic interest (the «AI Winter»).

A short revival of interest in AI research came in the 1980s, with the focus on «**expert systems**» - computer programs that answer questions or solve problems about a specific domain of knowledge, using logical rules derived from the knowledge of experts.

But that was also short-lasting, due also to the high cost of specialized hardware and a unfavourable economic juncture → **second AI winter**



AI History 5: Machine Learning → AI

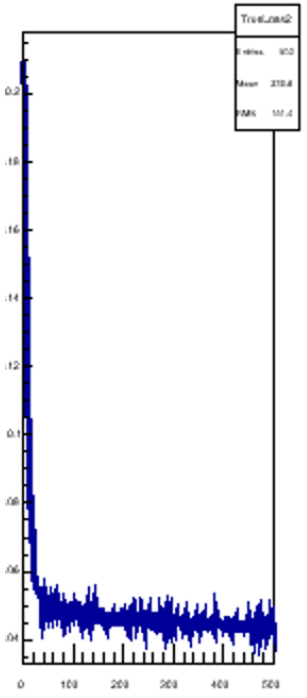
From the 90ies the path to AI took a different direction, eased by increases in cheap CPU, with development of Statistical Learning techniques (*e.g.,* decision trees) and breakthroughs in neural networks research

- [Mitchell 1997] provides a succinct definition of what Machine Learning is: "*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E*"

The development of complex neural networks, **and their adaptation to different problems**, brought many successful applications, where we may now arguably identify artificial intelligence at work

1997: Deep Blue defeats the chess World Champion G. Kasparov

A Digression: Statistical vs Machine Learning



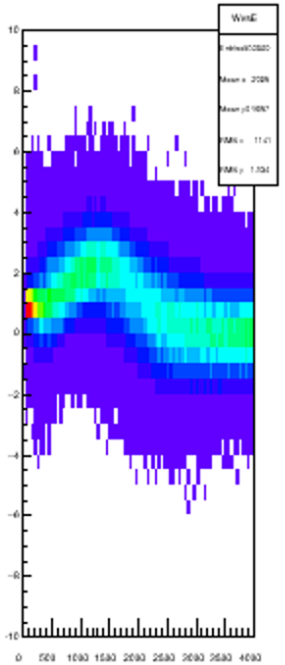
While the progress of AI research was going through the above described roller-coaster, statisticians were happily using tools that we today have repurposed into modern machine learning ones – no revolution, from their perspective

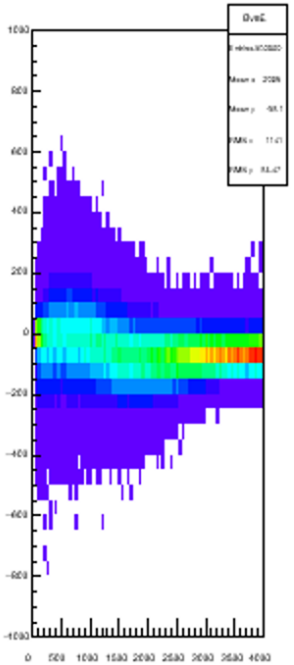
Indeed, one can really just as well do «learning» with old-fashioned, century-old tools. Take a k-nearest neighbor algorithm, for example: a classic SL tool.

In a recent study (to be published soon) I showed* how the regressor can be overparametrized by assigning a weight w_i and a bias b_i to every training event i=1...N:

$$\hat{x} = \sum_{i=1}^{k} x_i \quad \rightarrow \quad \hat{x} = \sum_{i=1}^{k} (w_i x_i + b_i)$$

This introduces 2N = O(1M) tunable parameters, which can be optimized by stochastic gradient descent to the minimum of a loss function exactly as a neural network does. Nonlinearity is introduced through locality of multi-D metric for distance.





BTW - Can We Agree on What "Learning" Is?

Maybe this really is the relevant question. Not idle to answer it, as by clarifying what our goal is we take a step in the right direction. Artificial intelligence requires machines to learn new tricks from experience!

Some **definitions of Learning**:

- the process of acquiring new, or modifying existing, knowledge, behaviors, skills, values, or preferences
- the acquisition of knowledge or skills through study, experience, or being taught.
- becoming aware of something, by information or from observation.
- Also, still valid is Mitchell's 1997 definition, given two slides back

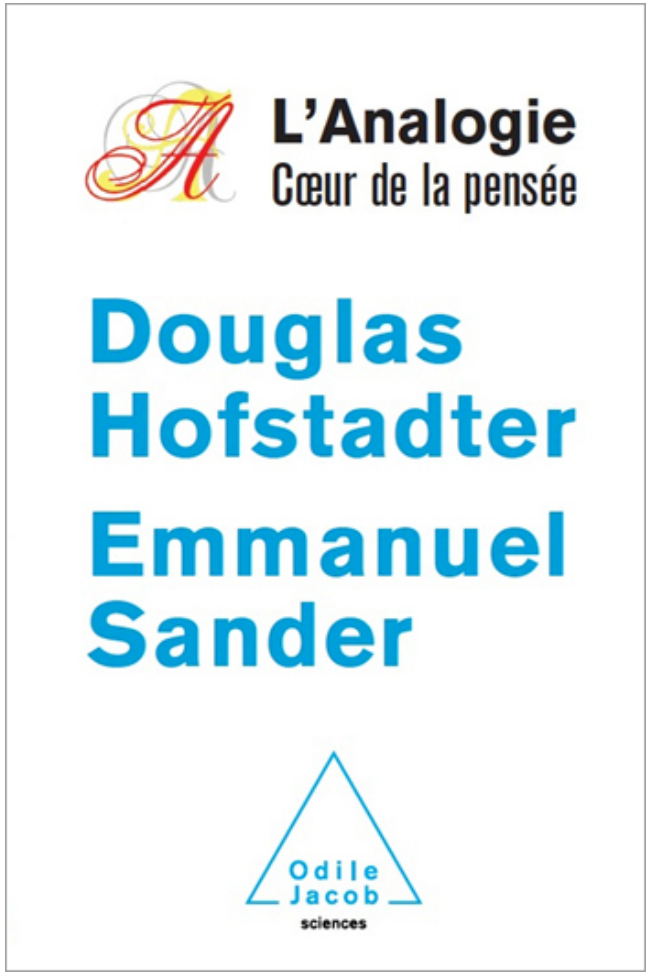
In general: learning involves **adapting our response** to stimuli via a continuous inference. The inference is done by comparing new data to data already processed. How is that done? **By the use of the mechanism called analogy**.

Analogies: The Heart of Thinking

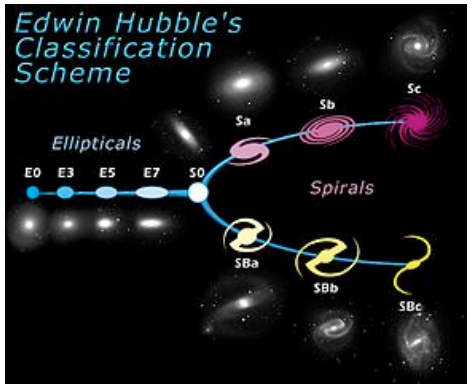
Analogies: arguably the building blocks of our learning process. We learn new features in a unknown object by **analogy** to known features in **similar** known objects
True for language, tools, complex systems – EVERYTHING WE INTERACT WITH

... But before we even start to use analogy for inference or guesswork, we (sometimes implictly) **classify** unknown elements in their equivalence class!

In a sense, **Classification** is even more "fundamental" in our learning process than is **Analogy.** Or maybe we should say that classification is the key ingredient in building analogies.

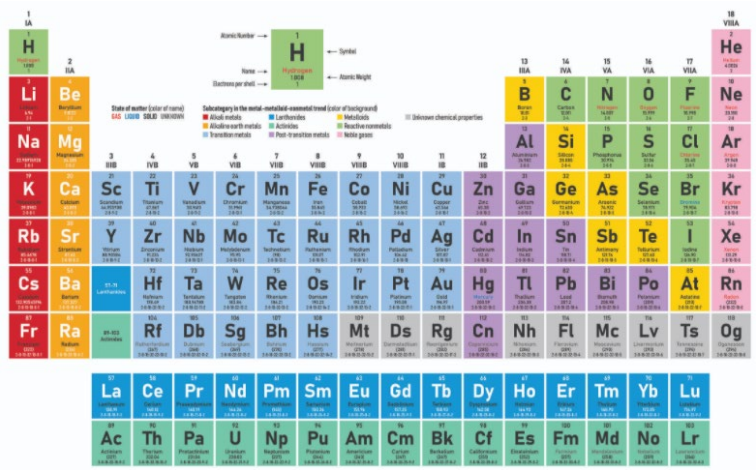


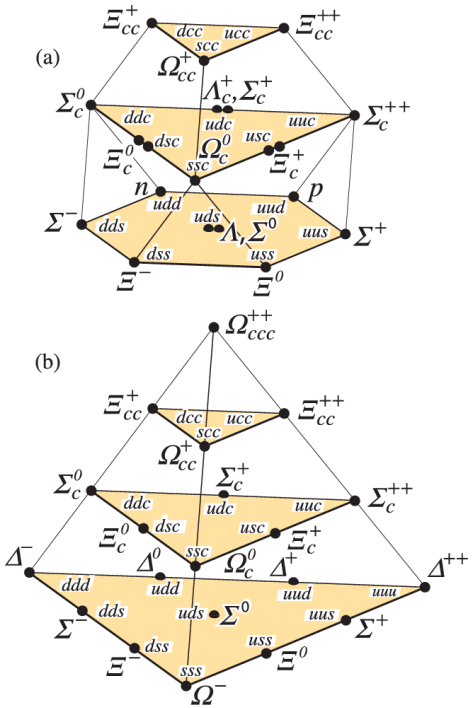
Classification and the Scientific Method

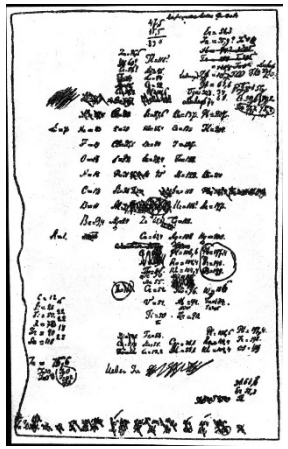


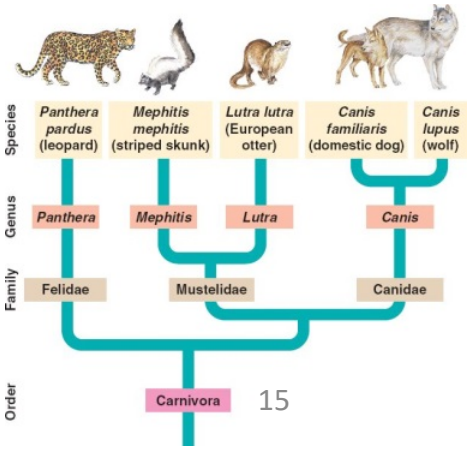
Countless examples may be made of how breakthroughs in the understanding of the world were brought by classifying observations

- Mendeleev's table, 1869
- Galaxies: Edwin Hubble, 1926
- Evolution of species: Darwin, 1859 (classification of organisms was instrumental in allowing a model of the evolution, although he used genealogy criteria rather than similarity)
- Hadrons: Gell-Mann and Zweig, 1964









It is thus no surprise that the advancements in supervised learning are bringing huge benefits to fundamental physics research today

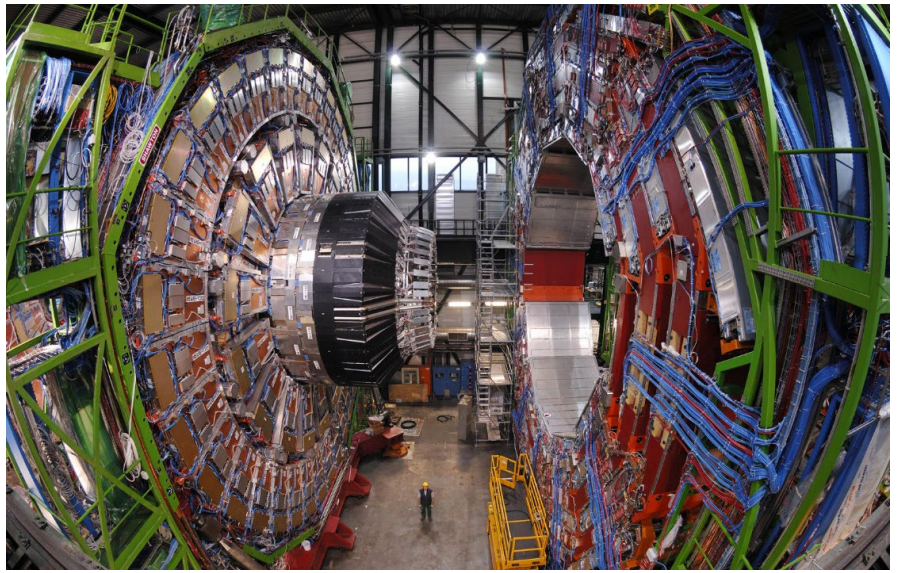
High-Energy Physics and AI

While the use of AI is common to all research areas of fundamental physics [incl. high-energy physics (HEP), astro-HEP, nuclear, neutrino,...], **let us focus on HEP examples here**

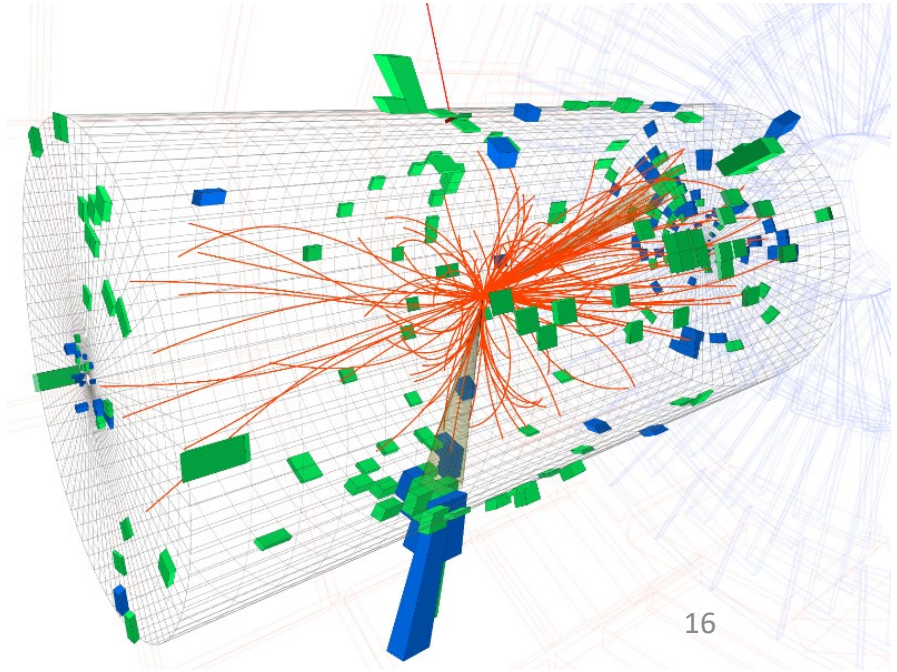
In HEP we study the inner structure of matter by colliding high-energy particle beams. Reactions produce complex «events».

- We need to reconstruct the reactions from the information we collect of all secondary interactions between produced particles and detection elements → inference, pattern recognition, dimensionality reduction
- We need to identify rare signals in enormous datasets dominated by backgrounds → classification
- We want to measure physical quantities with the utmost precision → regression, optimized inference
- We search for unpredicted structures in data → unsupervised learning: clustering, anomaly detection

In these and other applications, AI tools are fundamental helpers.



Above: the CMS experiment at CERN; below: computer reconstruction of particle signals produced in a collision



AI directions in fundamental physics

The main directions of research in today's particle physics research, taking the LHC as the playground, include:

- identification of the flavour of hadronic jets
- identification of hadronic decays within fat boosted jets
- event reconstruction in dense environments
- generative models and fast simulation of showers
- incorporation of nuisance parameters in statistical inference

What do these applications have in common?

They can all be tackled with deep learning algorithms

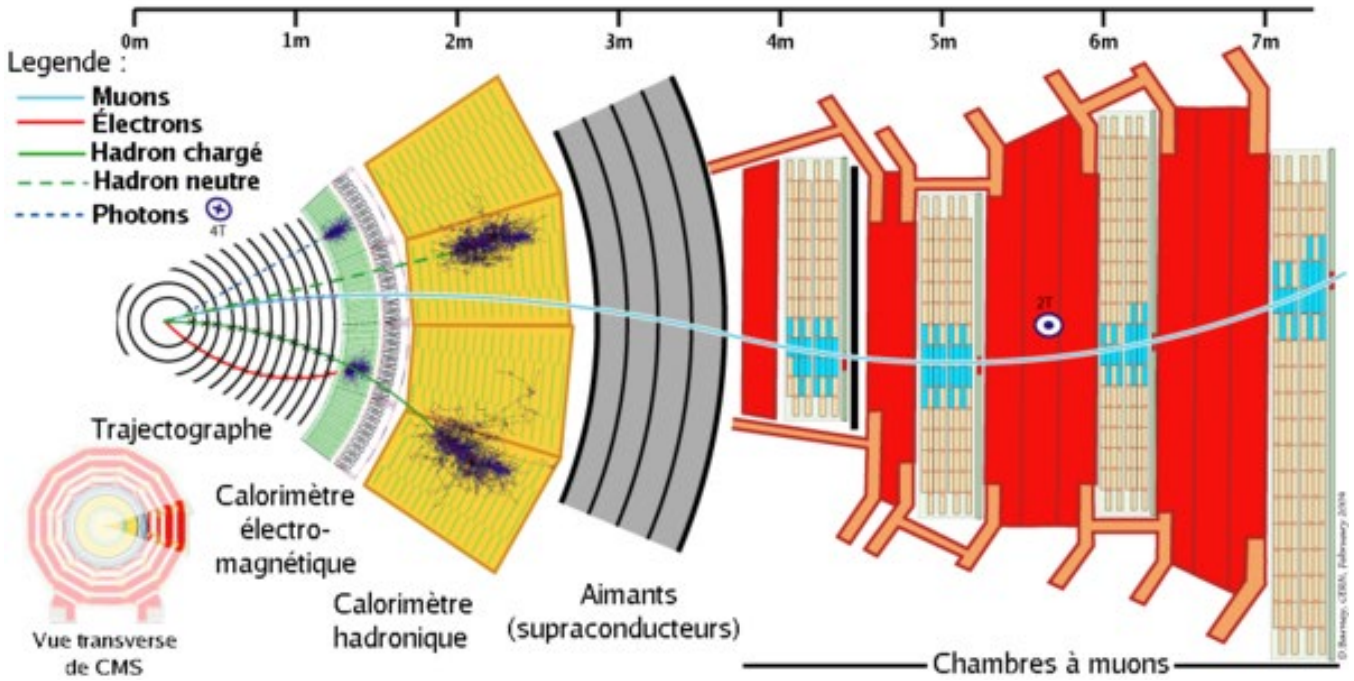
No time to discuss in detail the above applications and their state of the art. I will nit-pick a bit, as I wish to have a chance to discuss what is really at the frontier: end-to-end optimization of detector design

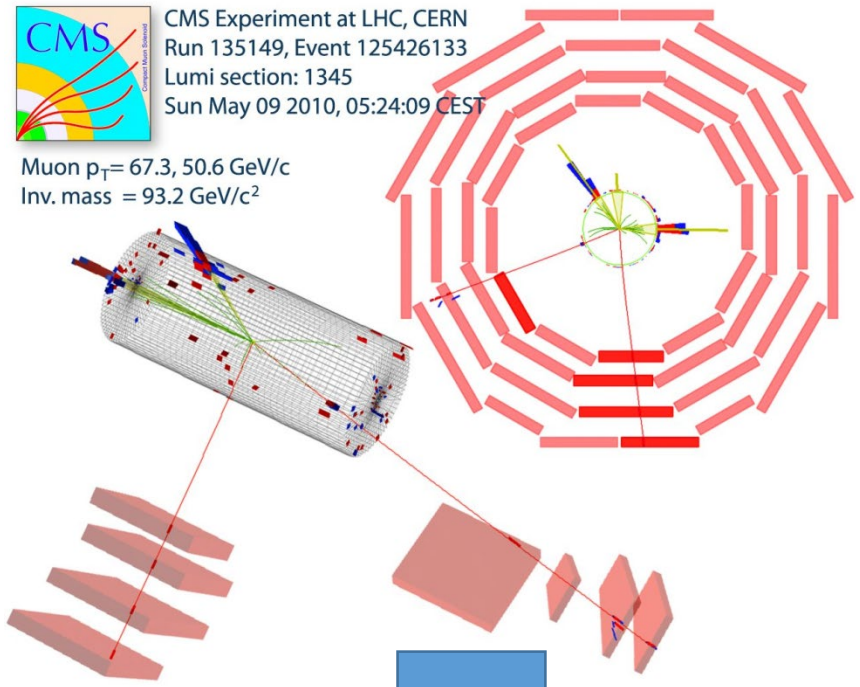
Cooking up good summaries

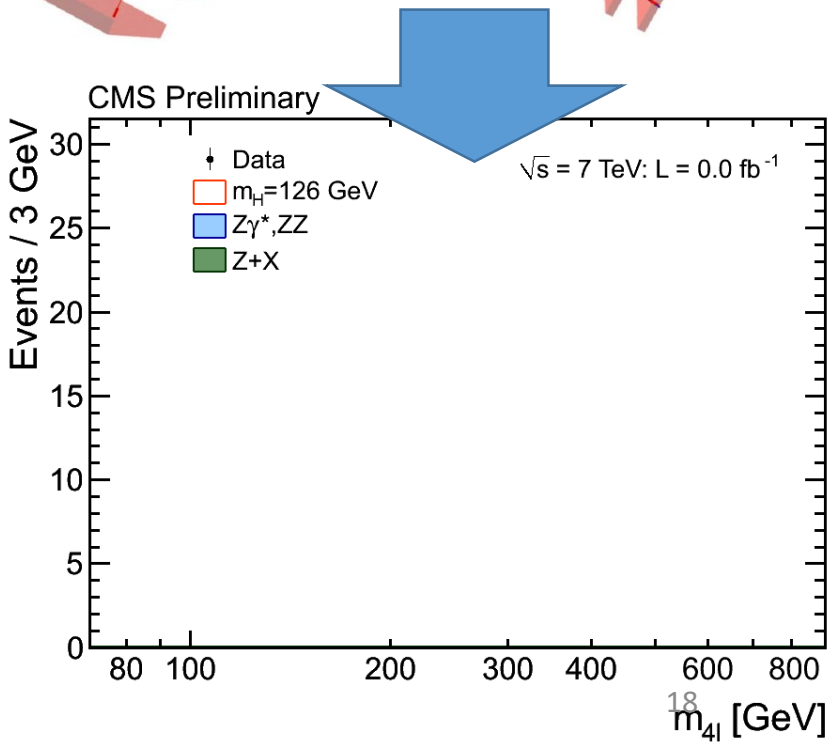
The reconstruction of O(100M) electronic signals and its decoding into high-level features that describe what particles whizzed through, and with what energy and direction, is a huge dimensionality reduction task.

The challenge is to retain information as inference is performed on a low-D space of summary statistics...

Below: a section of the CMS detector, highlighting different signals left by particles







The inverse problem: simulation-based inference

The quantum nature of physical processes at colliders (particle production, decay, interaction with matter) makes their observation a quintessential probabilistic process

→ The extraction of information from data must be formulated in statistical terms with the inclusion of all steps, to infer theory parameters from final-state observables: reconstruction → identification → measurement → inference

→Classification, hypothesis tests, and regression are all based on the specification of a statistical model **p(x|θ)** that describes the probability to observe **x**, given parameters **θ** of a underlying theory

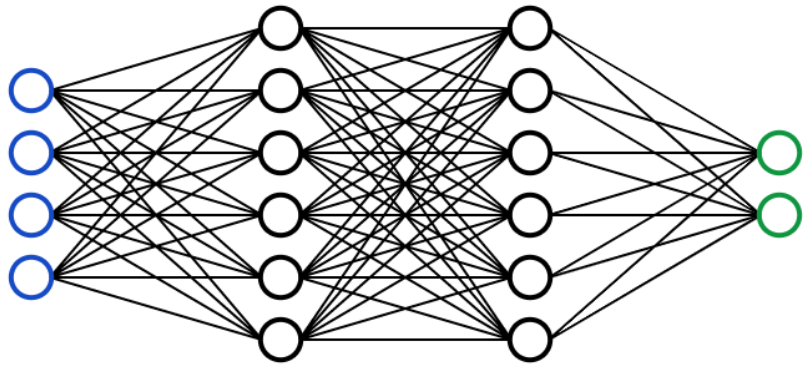
The high dimensionality of x makes the direct derivation of p impossible. We have learned to rely on simulation (forward generation of examples of **x** given a choice of **θ**) based on a reduction of the dimensionality of the problem → simulation-based inference

This way, though, we lose information, unless we find ways to retain the *sufficiency* of the summary statistic on which we base inference on the value of **θ**.

This topic was covered excellently by world-expert Kyle Cranmer – see his lecture from July 12th

The rise of deep learning in HEP

DNNs use skyrocketed in HEP after 2012, when they were used for the Higgs discovery (2012 is also the turning point in the imagenet challenge). A **true paradigm change**!



Further evidence of the benefit of ML tools for HEP was given by the Kaggle Higgs challenge [Kaggle 2014], with 1800 teams participating (physicists, statisticians, computer scientists).

Task: separate H→ττ decays from backgrounds

The most effective solution was based on a pool of DNNs, with emphasis on cross-validation

Alternative methods commonly used in HEP (xgboost, Bayesian NN, etc.) were beaten soundly

Solution	Score
Gabor Melis (DNN pooling)	**3.806**
MultiBoost	3.405
TMVA boosted trees	3.200
Naive Bayesian classifier	2.060
1D cut-based selection	1.535

equiv. to 6 times more data!

Looking deeper into the best solution

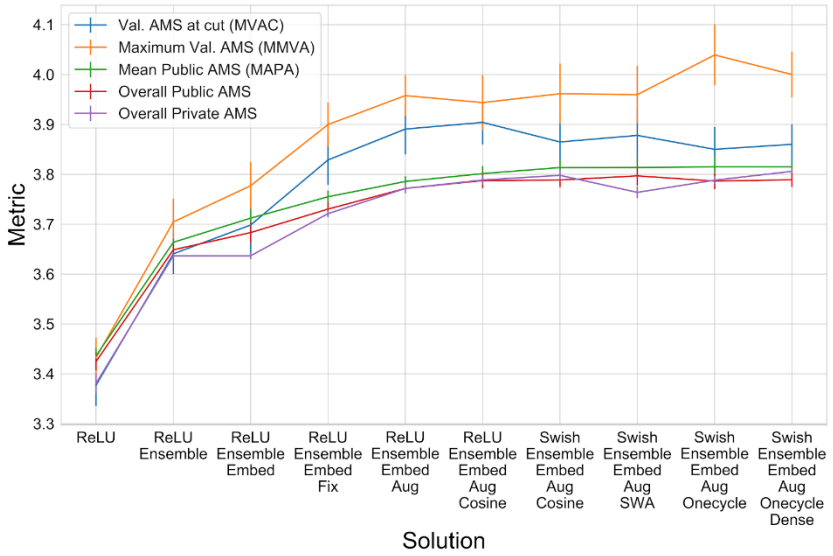
In a recent study[Strong 2020], Giles C. Strong reproduced Gabor Melis' performance with a similar but much more GPU-efficient setup.

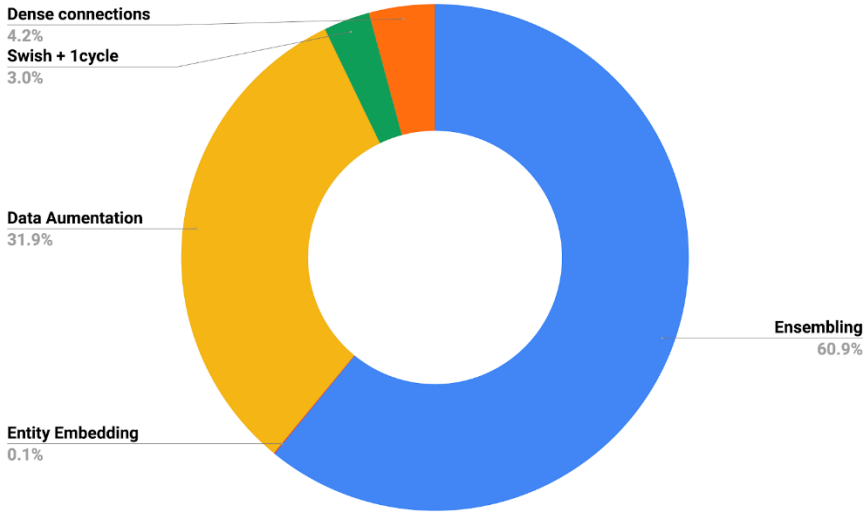
He could thus study the ingredients which were the most useful to improve the solution: Data augmentation + Swish activation + Densely connected layers = more performant model

- Require fewer models in ensemble to achieve same performance (10 versus 70)
- Ensemble much quicker to apply at inference time

- Fewer models + 1-cycle schedule = quicker to train
- Accounting for difference in GPU (Titan → 1080 Ti) processing power, new solution:
  - 13x quicker to train on GPU, 7x quicker on laptop CPU
  - 33x quicker to apply on GPU, 3x quicker on laptop CPU

	Our solution	1st place	2nd place	3rd place
Method	10 DNNs	70 DNNs	Many BDTs	108 DNNs
Train-time (GPU)	8 min	12 h	N/A	N/A
Train-time (CPU)	14 min	35 h	48 h	3 h
Test-time (GPU)	15 s	1 h	N/A	N/A
Test-time (CPU)	3 min	???	???	20 min
Score	3.806 ± 0.005	3.80581	3.78913	3.78682



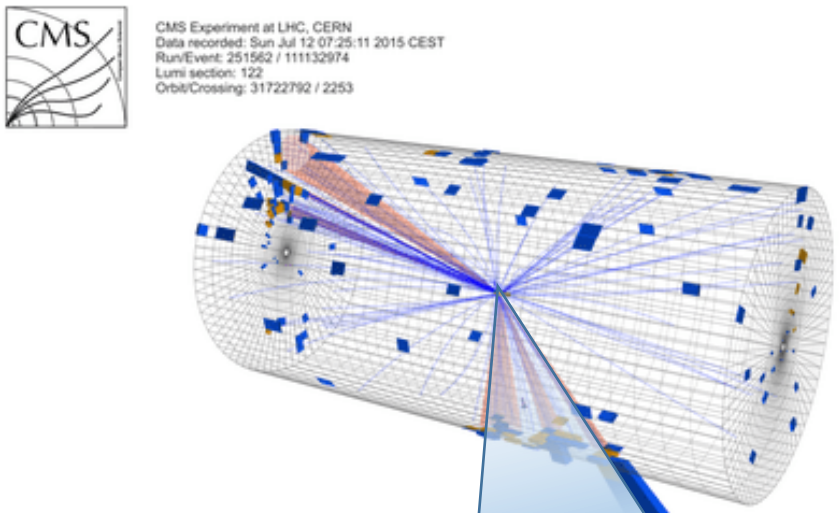


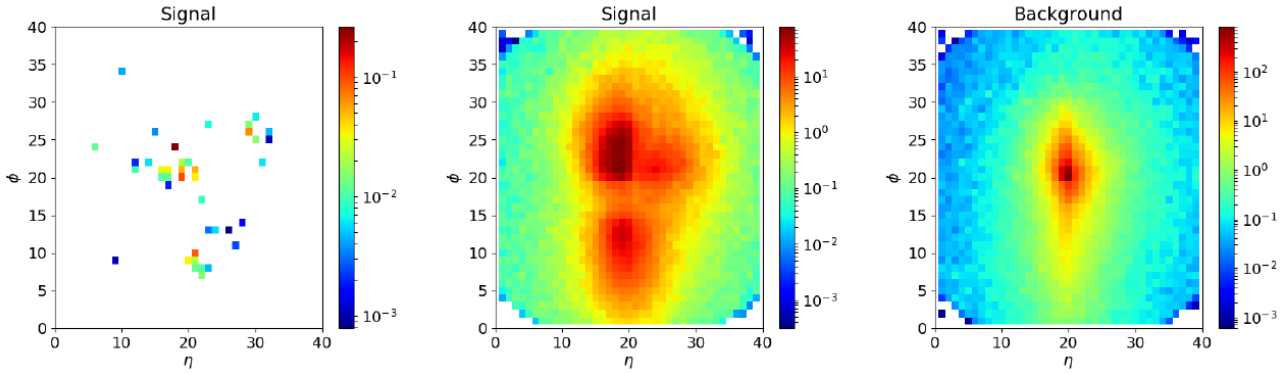
New ways to detect particles: boosted jets

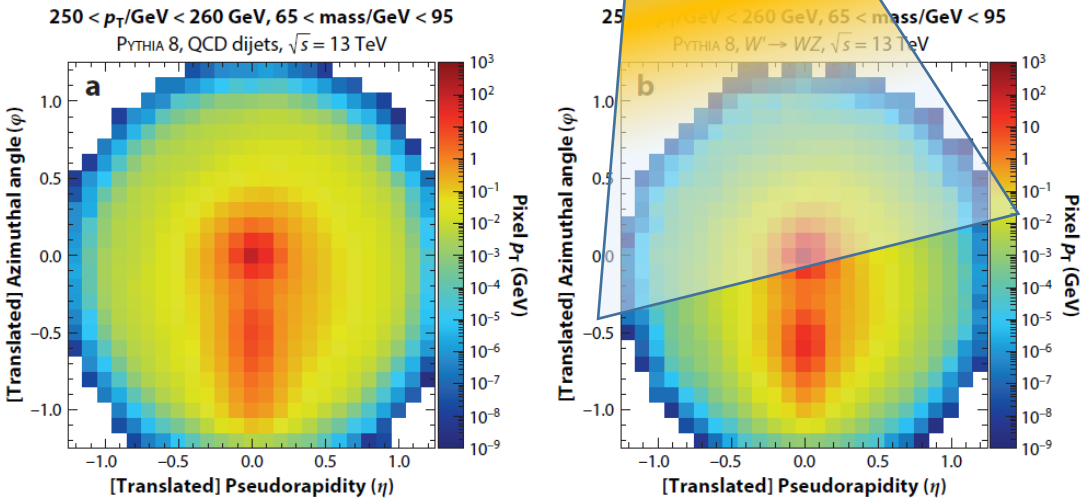
The W boson (1983) and the top quark (1995) were discovered using decays involving electrons and muons, as the more frequent quark decays are difficult to distinguish from QCD-induced backgrounds

We realized at the LHC that if top, W, Z, H particles have high energy, the jets of hadrons they produce in their quark decays are distinguishable from QCD through their inner structure – something which requires highly granular calorimeters which ATLAS and CMS do posssess.

Today the most sensitive searches for new massive particles benefit from CNN-powered imaging techniques [Kasieczka 2019]







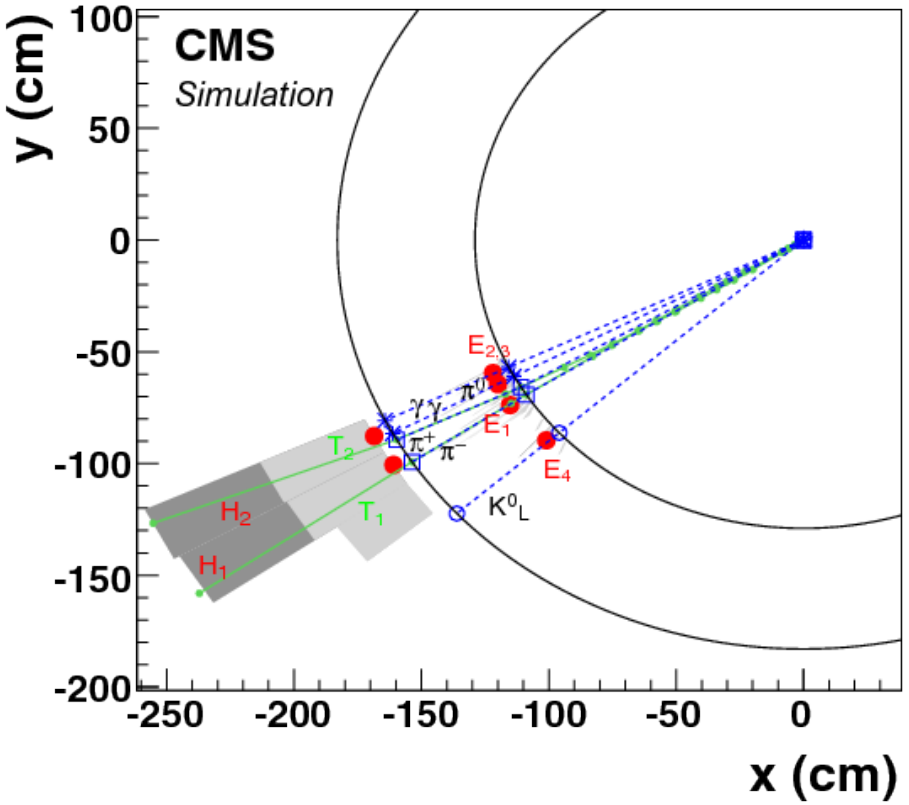
Left: true pattern of particles in a top quark jet. Center: observed average distribution for top quark decays. Right: average shape of QCD jets.

Tracking Particles in Dense Environments

CMS reconstructs particles emerging from collision events using a technique called «particle flow»[10], where every charged as well as neutral particle is assigned energy deposits and hits

- This is challenging when there is overlap of many collisions in the same «bunch crossing» – but that's precisely where we are heading with HL-LHC
  - To study deeper in the structure of matter we need more luminosity, as explorers in a dark cavern

To preserve the performance of our reconstruction methods in O(200) pileup conditions, Jan Kieseler has proposed a technique[Kieseler2020] based on object condensation, showing increases in the performance of particle flow method (see below)



Object condensation

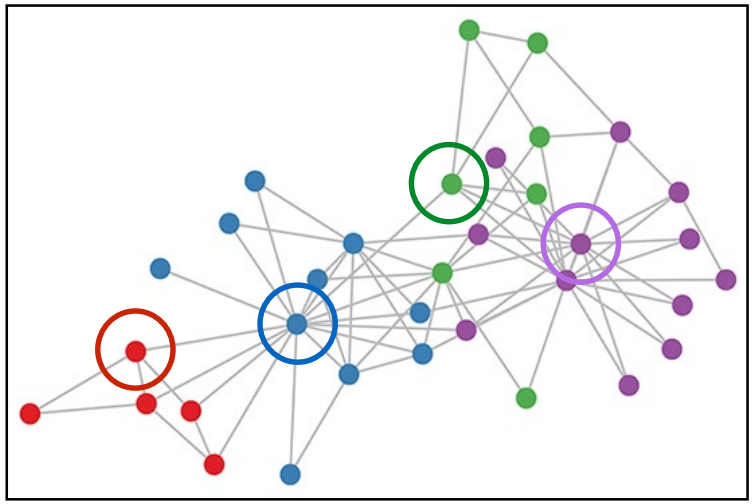
Aim:

- Determine object properties (e.g. particles 4-momenta, ID)
- One shot: no seeding
- **Aggregate all object properties in representative 'condensation point'**
- Allow for fractional/ambiguous assignments

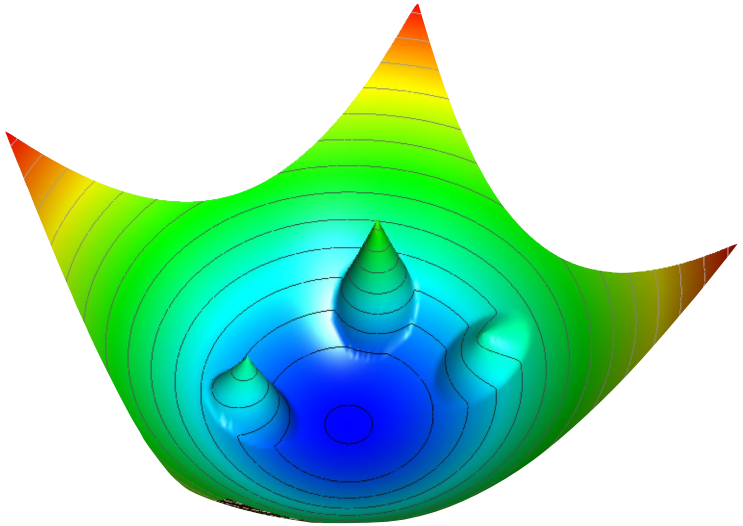
- Define truth:
  - Assign each vertex to one object (e.g. highest fraction)
  - Assign all object properties to each assigned vertex

- Predict per vertex
  - Object properties are collective of all clustered points
  - Confidence β of assignment
  - Cluster coordinates x

Define charge, attractive and repulsive potential which allows to condense objects in loss minimum



J.Kieseler, arxiv:2002.03605, EPJC



Attractive potential for an object in the presence of three other objects nearby

Slide contents are courtesy Jan Kieseler

Object condensation / 2 – Example on image data

Proof of principle using images with large overlaps (<=90%)

- Condensation, object ID
- Rather simple CNN

TASK: identify objects of different kind, associate each to a condensation point representative of the object

Inference

- Start with highest β vertex, collect points
- Get object properties
- Repeat until $\beta_{min} \cong 0.1$

Image+condensation points | Pixel class | Cluster coordinates

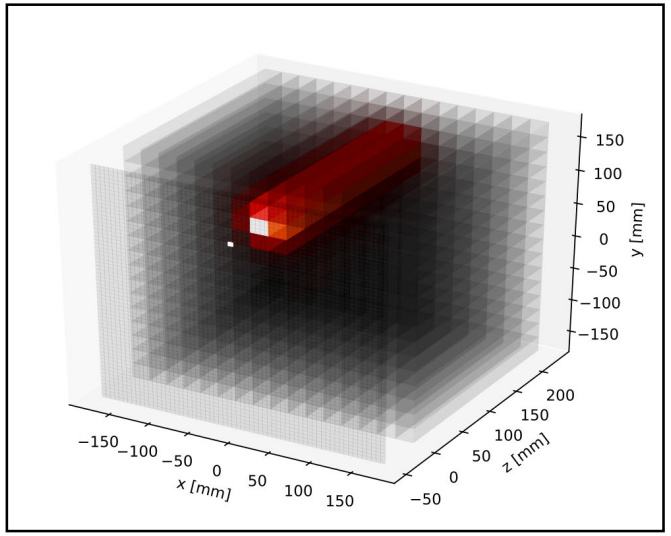
Slide contents are courtesy Jan Kieseler

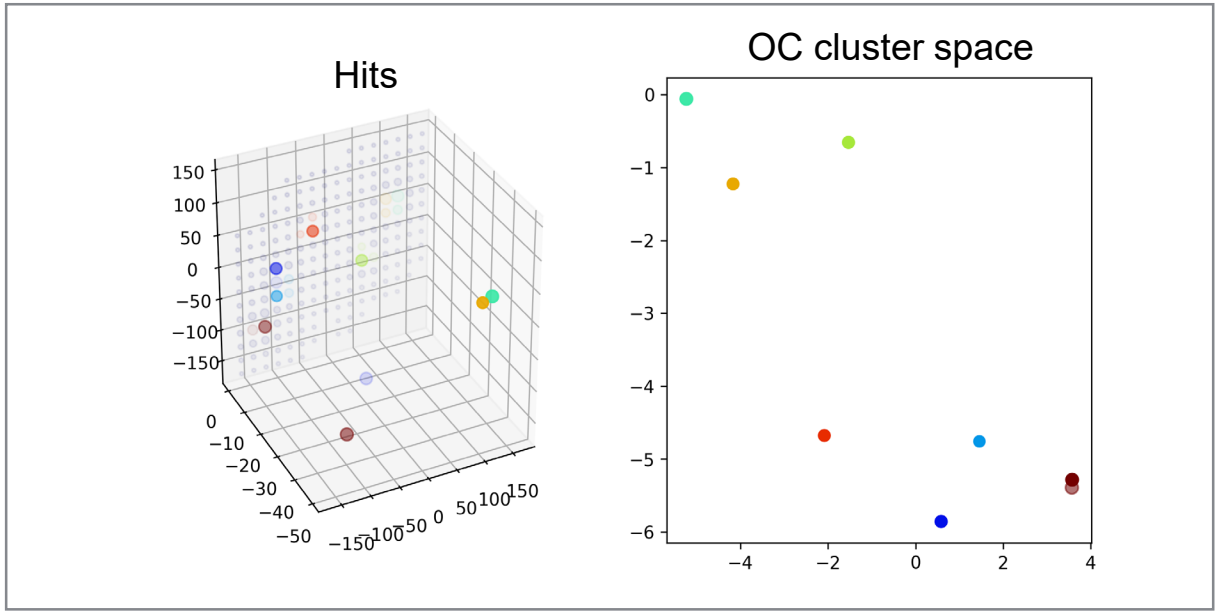
OC 3 - Application to particle reconstruction

Try it on a simplified shower: only consider EM objects

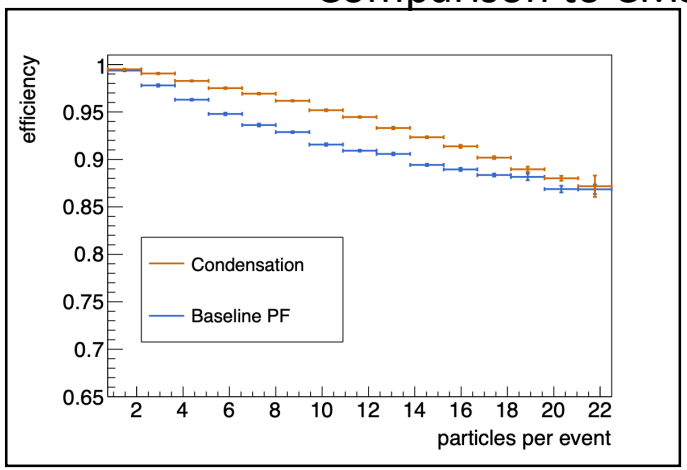
Shoot up to 9 particles per event;

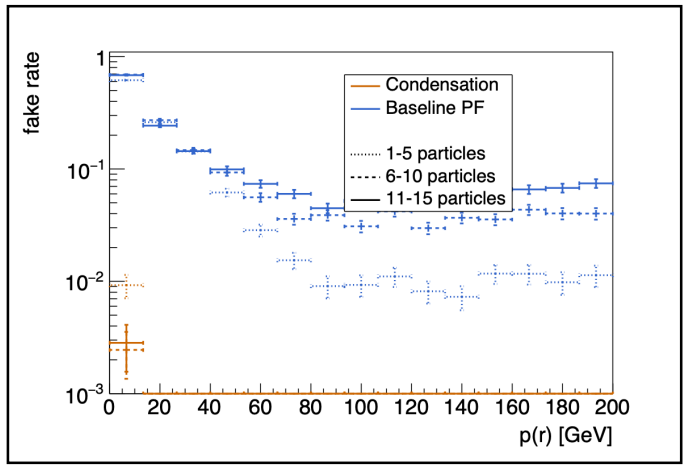
approximate tracker by one layer + truth momentum smearing





Comparison to CMS-like Particle flow:





→ Excellent **extrapolation properties** beyond training conditions

Low fake rate, and fakes only at low energies

Improved single particle resolution

Slide contents are courtesy Jan Kieseler

Incorporation of Nuisances in Classification

Realizing the misaligment between the dimensionality reduction techniques (operated by NN, e.g. in classifying signal vs background based on observed event features) operated in data analysis and the true goal of the analyses (*e.g.*, «smallest total uncertainty in parameter x»), a number of deep learning techniques have flourished in the past five years to

- parametrize effect of nuisances as input to NN
- decorrelate summary statistics from sensitive variable
- penalize the loss to be insensitive of model variations
- make the loss of a NN aware of the final objective and method of extraction
- etcetera

Impossible to summarize these here (I barely managed in a 1h seminar here at CMU a few months ago), but for those interested P. de Castro and I produced a review [De Castro 2020] (soon to be published in the book «Artificial Intelligence for HEP» by World Scientific)

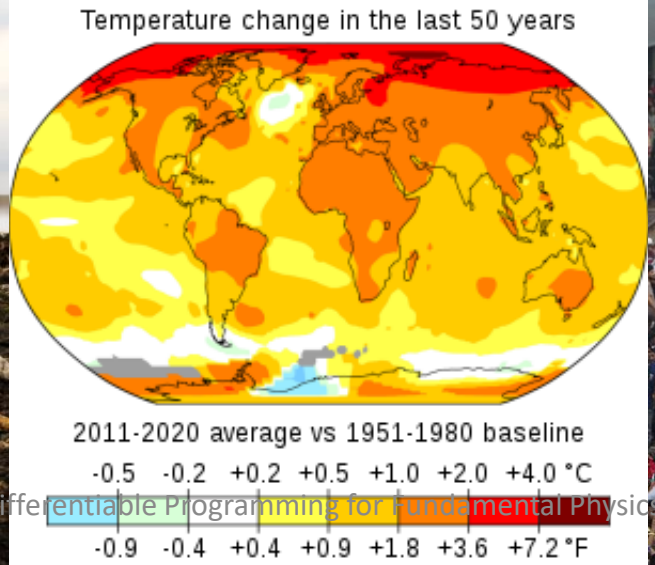
Future Challenges: Introduction/1

The recent publication of the 2020 update of the European Strategy for Particle Physics (EUSUPP) [1] encourages feasibility studies for new large, long-term projects which will once again push our technological skills to their limits.

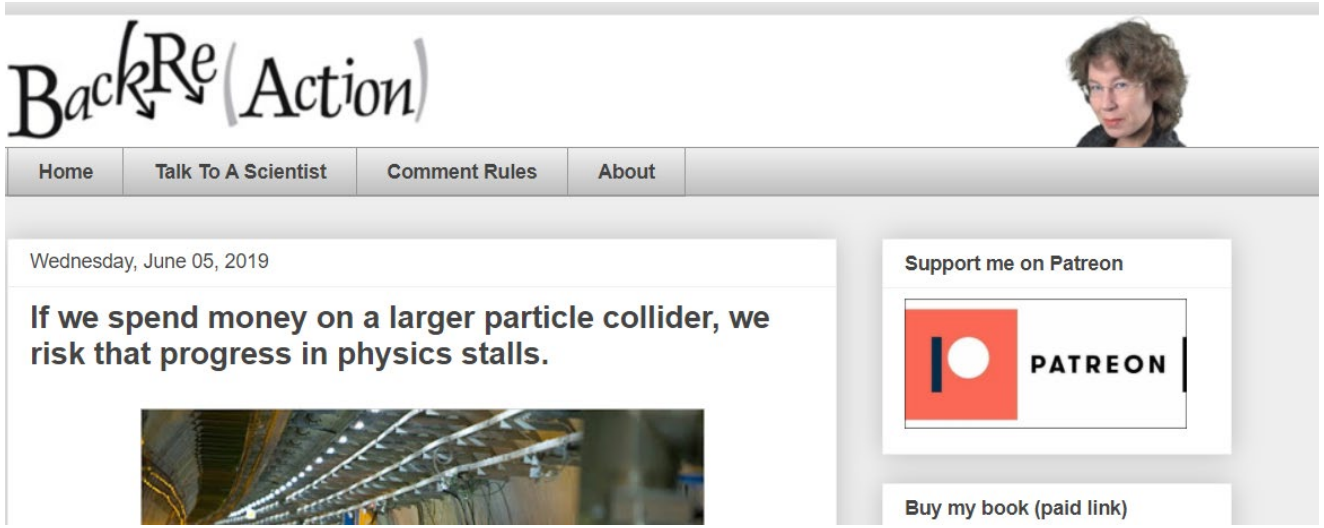
Future Challenges: Introduction/2

The 2020 update of the European Strategy for Particle Physics (EUSUPP) [1] encourages feasibility studies for new large, long-term projects which will once again push our technological skills to their limits.

At the same time, humanity faces unprecedented global challenges (climate change, pandemics, overpopulation) which demand resources to seek solutions through applied science innovations, rather than investing in fundamental research.



Future Challenges: Introduction/3



Furthermore, there are indications that wide sectors of society no longer consider the furthering of our understanding of matter at the smallest distance scales, or other projects that require large and coordinated effort and significant funding, a top priority [2].

In this situation, **ensuring the maximum exploitation of any resources spent on fundamental research is a moral imperative,** and it may be a key to ensure that the long-term projects envisioned by the EUSUPP may be undertaken and sustained.

The Status Quo

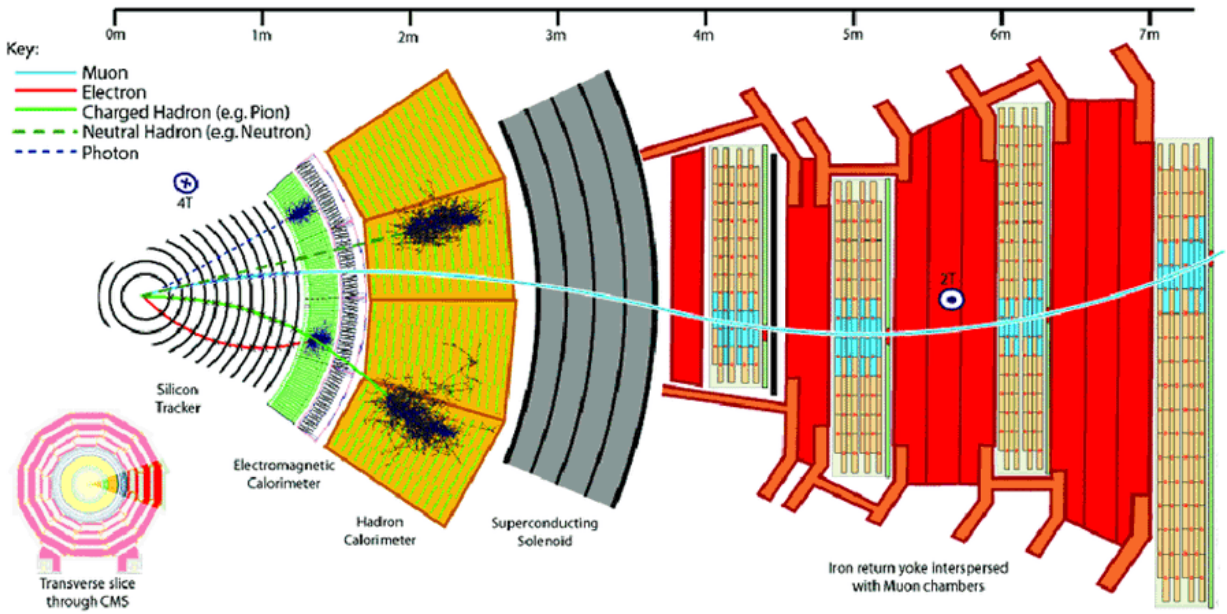
In the past 50+ years the design of new detectors systematically leveraged the most performant available technologies, often fostering significant further advancements and spin-offs [3].

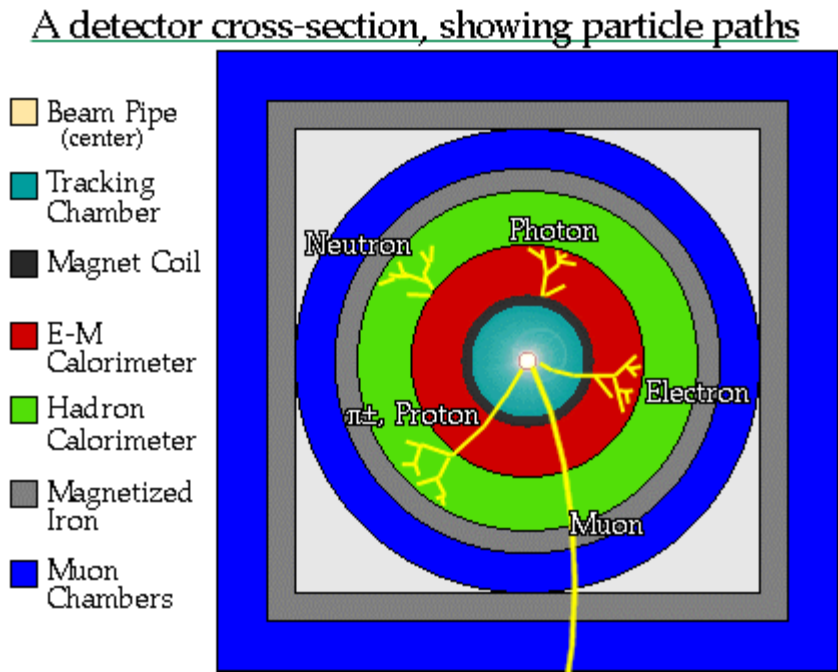
Yet one observes that **the crucial underlying global paradigms** of experimental design **have remained mostly unchallenged** across decades.

- **"Track first, destroy later"**
- **Redundancy in detection systems**
- **Symmetrical layouts**

→ **No guarantee of optimality**

While these choices have served us very well for a long time, they **do not directly maximize a high-level utility function**, such as the highest discovery reach for a physical process, or measurement precision for a given physics parameter.





Optimal for what?

The reason why detectors are complex is not only that the studied physics is complex: a lot has to do with Science being a demanding job. We want to study *everything* and do it *better* than previously

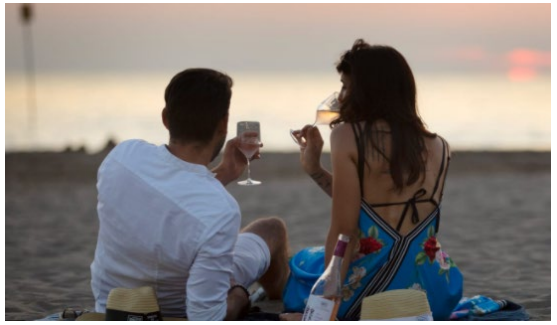
So, what does it mean for a detector to be *optimal*?

What loss function do we aim to minimize?

Does it make sense to speak of a single utility function?

Concerning the last question: I am convinced that it does, and I will try to convince you in the next few slides.

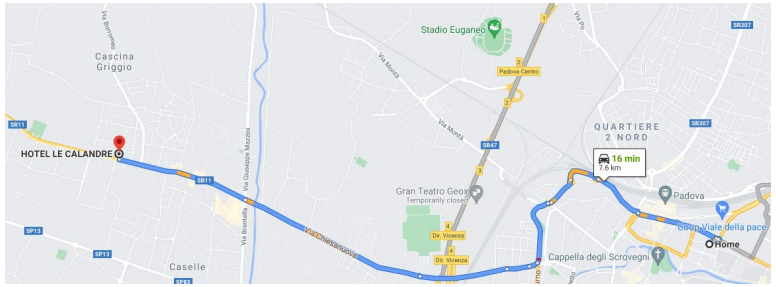
Recipe for a perfect dinner







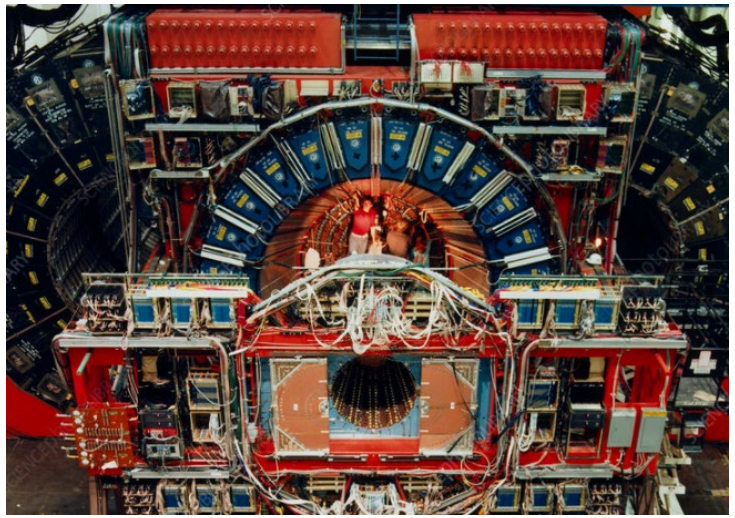


We are not alien to confidently taking complex decisions in a multi-objective space. We actually do it routinely...

Of course, we are not deterred by knowing that our optimization target is not universal

Recipe for a perfect trigger

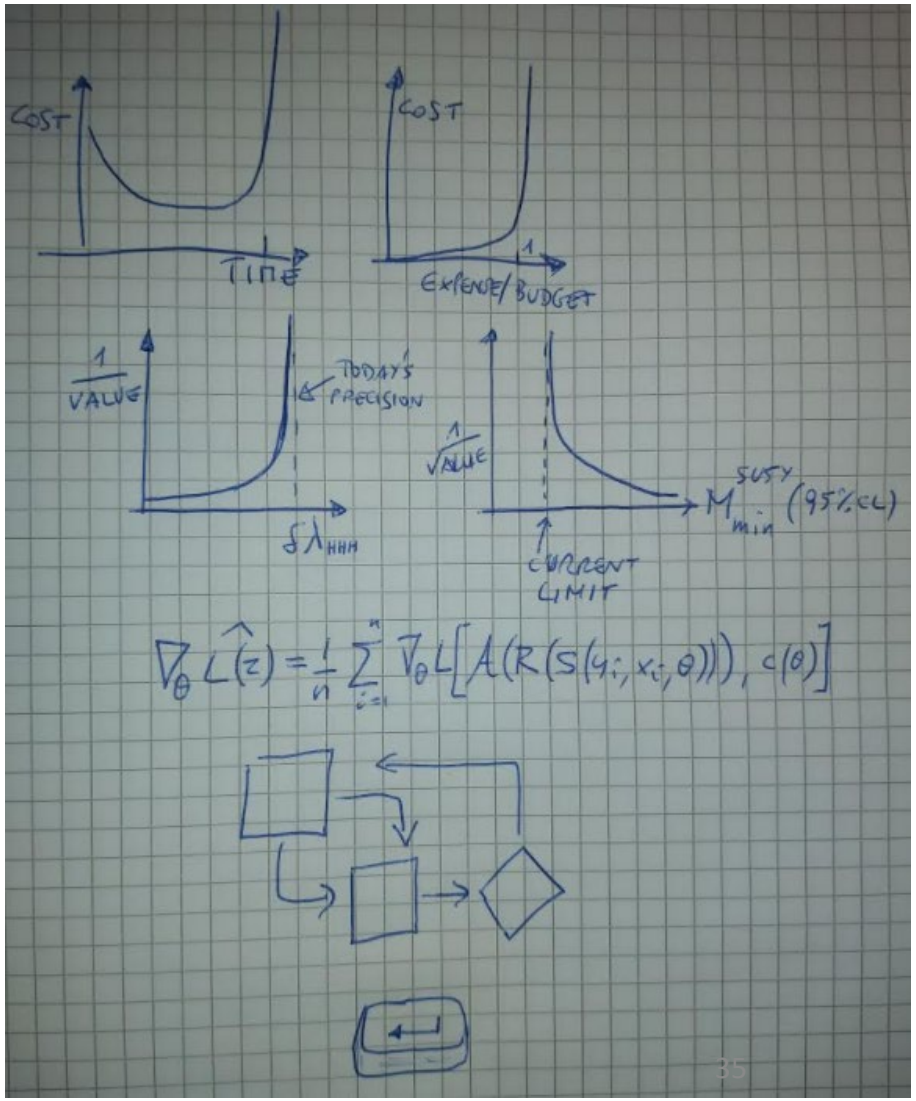
Similarly, we are actually *used* to create multi-target optimization strategies, e.g. when we allocate resources for the trigger menu of a collider detector.



Consider CDF, Run 1 (1992-96): taking in a rate of 300 kHz of proton-antiproton collisions and having to select 50 Hz of writable data created some of the most heated scientifically-driven, rationally motivated, painfully well argumented debates I ever listened to. The top quark had to be discovered, but it was not the only goal of the experiment...

Recipe for a perfect detector

1. Assess your total **budget** and **time**-to-completion
2. Model as a steep function the **cost** of overriding budget or time
3. Assess the **scientific impact** of each achievable scientific results, optionally *as a continuous function of their precision*
4. **Create a differentiable model** of the geometry, the components, the information-extraction procedures, and the utility function
5. Construct a pipeline with those modules, enabling backpropagation and gradient descent functionality
6. Let the chain rule of differential calculus do the hard work for you



We'll get there, but first let's see why we need to.

Makeshift surrogates of objectives

When we design the sensors for a tracking device, operate choices on budget allocations, define requirements for the various resolutions of detection elements, or choose composition and layout of active and passive material of calorimeter cells, **we are implicitly trying to find an optimal working point in a loosely-constrained feature space of hundreds of dimensions**. Such a task is clearly super-human.

Because of that, we set our aim on makeshift surrogates of our real objectives.

- *E.g.*, we might desire our objective to be "*the highest precision on the Higgs boson self-couplings our budget can ensure*", but all we can do is stick to useful proxies suggested by past experience, and rather focus on the "*highest achievable energy resolution for isolated photons*", ignoring the rest of the parameter space
- In a neutrino detector this would instead sound as "the highest precision on $\boldsymbol{\theta_{13}}$ we can get", when the focus becomes instead maximizing the number of interactions and reducing the background level.
- Our simulations only allow us to probe the result of specific choices, not to map interdependencies and find extrema of an utility.

Makeshift surrogates of objectives

When we design the sensors for a tracking device, operate choices on budget allocations, define requirements for the various resolutions of detection elements, or choose composition and layout of active and passive material of calorimeter cells, **we are implicitly trying to find an optimal working point in a loosely-constrained feature space of hundreds of dimensions**. Such a task is clearly super-human.

Because of that, we set our aim on makeshift surrogates of our real objectives.

- *E.g.*, we might desire our objective to be "*the highest precision on the Higgs boson self-couplings our budget can ensure*", but all we can do is stick to useful proxies suggested by past experience, and rather focus on the "*highest achievable energy resolution for isolated photons*", ignoring the rest of the parameter space.
- In a neutrino detector this would instead sound as "the highest precision on θ_{13} we can get", when the focus becomes instead maximizing the number of interactions and reducing the background level.
- Our simulations only allow us to probe the result of specific choices, not to map interdependencies and find extrema of an utility.

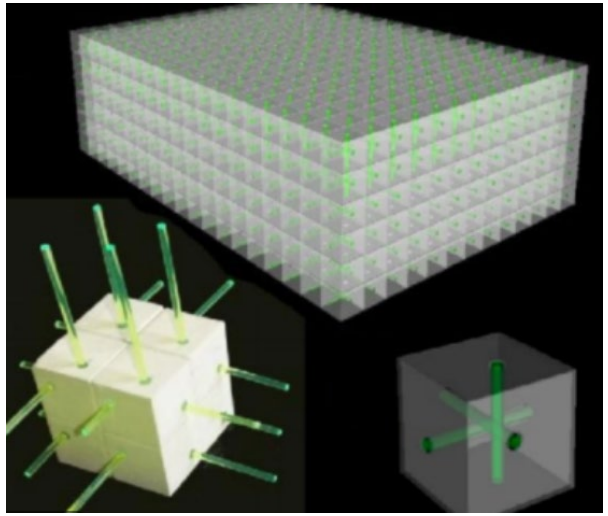
Evolving from this modus operandi to directly goal-informed decisions enables potentially enormous performance gains.

The design space is large – no, larger

New technological advancements are crucially enabling a better optimization of our instruments by reducing the cost of complex layouts.

- 3D printing of scintillation detectors is being explored for neutrino physics [4]
- Very thin layouts of resistive AC-coupled silicon detector elements provide large gains in spatial and temporal resolution [5].

→ **The geometry space has become larger and more complex to explore.**



Higher-performance demands are also rising as, *e.g.*,

- Tracking in dense environments requires AI solutions
- Hitting the neutrino floor (*e.g.* at SuperCDMS) may require new paradigms for DM searches
- As we move fundamental physics research to space, payload and power consumption become driving constraints
- Boosted jet tagging at high p_T –all the rage for NP searches at the LHC –demands us to invest in more granular, higher-performance hadron calorimeters

A hybrid calorimeter?

As particle flow techniques allow the tracing of individual particles and the complete reconstruction of dense, collimated jets, we must have more of that:

- Optimizing the design of a detector for a long-timescale project based on reconstruction capabilities which will be available in the future [11] seems the right thing to do, if we can pull that off
- Integrating tracking and calorimetry layers may improve the «image» reconstruction of energetic hadronic jets, shown to be crucial for high-mass new particles
- Measuring muon energy from their radiative loss in a dense environment using convolutional neural networks may extend the use of muon probes of NP in multi-TeV region (see *infra*)
- Nuclear interactions have always been dreaded in a tracker, but in a granular calorimeter they may strengthen particle-ID (using probabilistic information coming from nuclear cross sections of different species)
  →Of special interest to a number of applications

Muon energy measurement in a calorimeter?

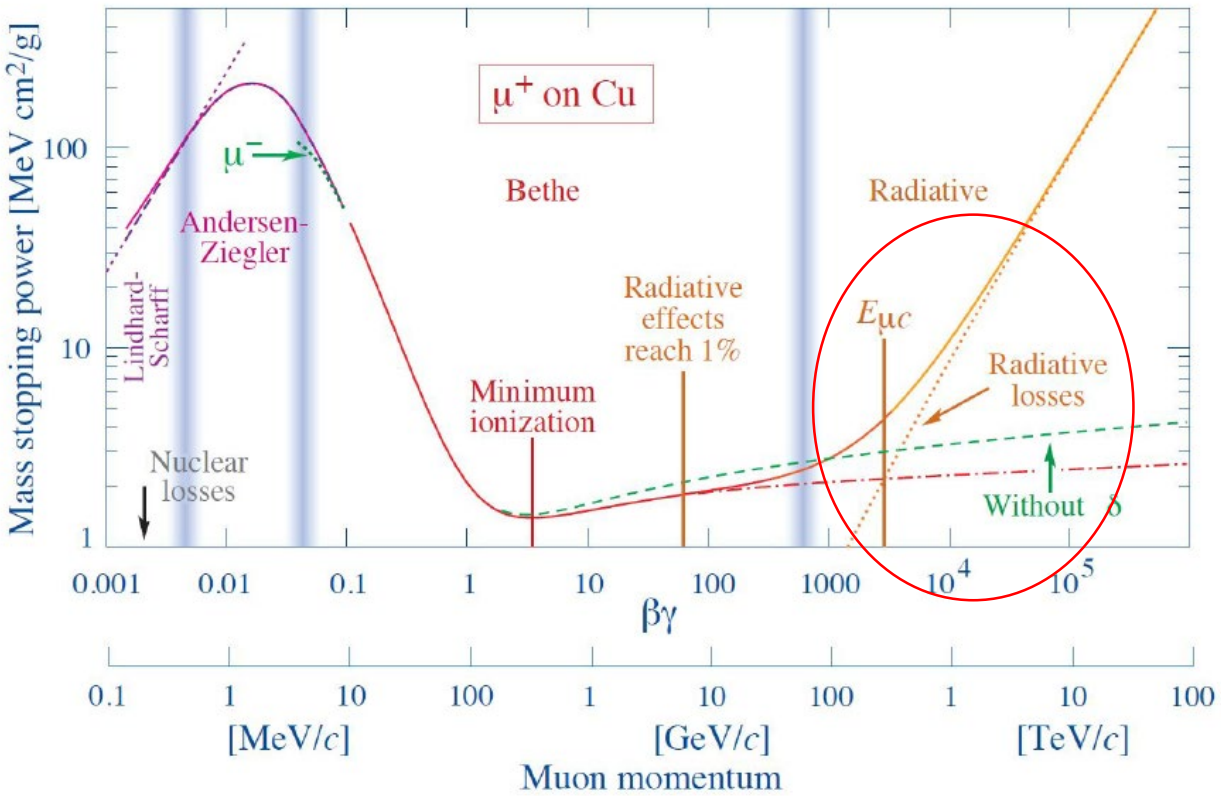
Muons interact with matter by ionization, pair production, bremsstrahlung, and photonuclear reactions. The E loss is dominated by the high-end of the Landau distribution (knock-on electrons).

E loss is very modest and stochastic, so we have to rely on magnetic bending for inference on muon momentum

Bending measurements break down at TeV energies: in B=2T, a 1-TeV muon is deflected by a mm by traveling for 2 metres

Resolution scales linearly with momentum, because **P=qBR** for a charge q in a field B, with R radius of curvature.

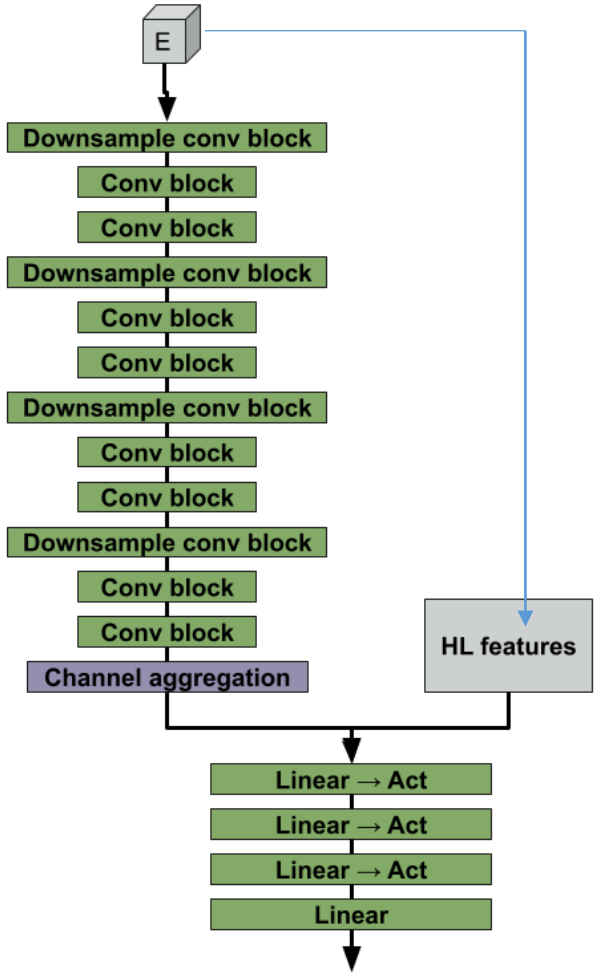
E.g., in ATLAS **σ(p)/p = 0.1 - 0.2 p** depending on angle etc.



Above: mass stopping power for positive muons in Cu, showing the radiative energy loss onset above 1 TeV

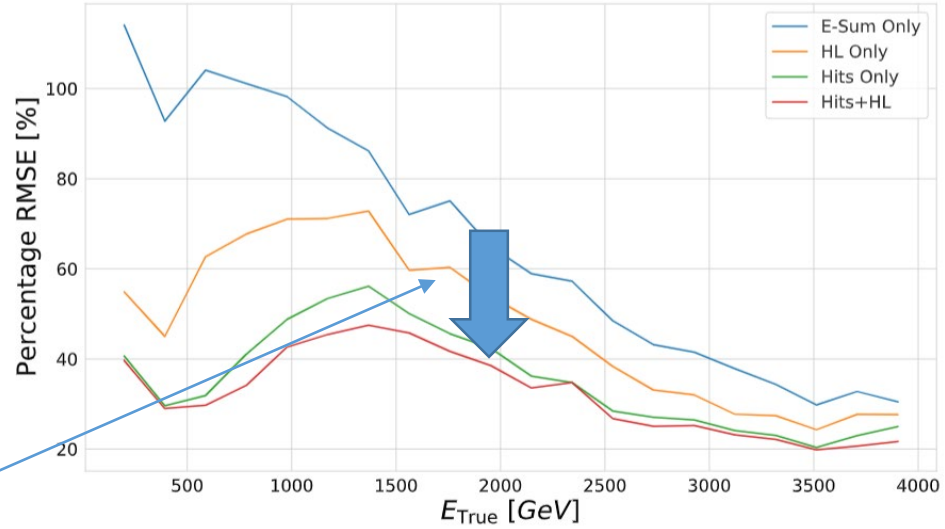
Results of Regression with CNNs

To see how well we can measure high-energy muons in a calorimeter we used a customized deep learning architecture, which combines convolutional blocks and dense layers using both high-level features and raw «image-like» energy deposits in 3D space



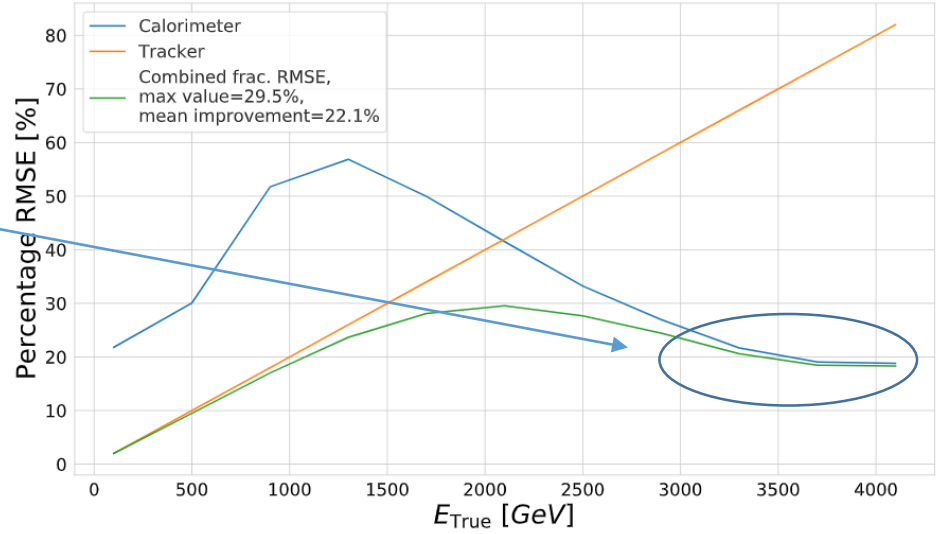
Results of Regression with CNNs

To see how well we can measure high-energy muons in a calorimeter we used a customized deep learning architecture, which combines convolutional blocks and dense layers using both high-level features and raw «image-like» energy deposits in 3D space



Of relevance is the point that the pattern of radiation deposits contains information useful to regress to true muon energy

Results (right) show that one can reach 20% resolution with radiation loss information at 3-4 TeV, a region where curvature information (here assumed is a relative momentum resolution of 20% from magnetic bending at 1 TeV) becomes useless



How large are the gains of a full optimization?

I recently provided a clear example [13] of how experimental design as is carried out today leaves ample room for improvement from the systematic study of even seemingly irrelevant choices for, *e.g.*, the placement of active and passive material in a simple detector.

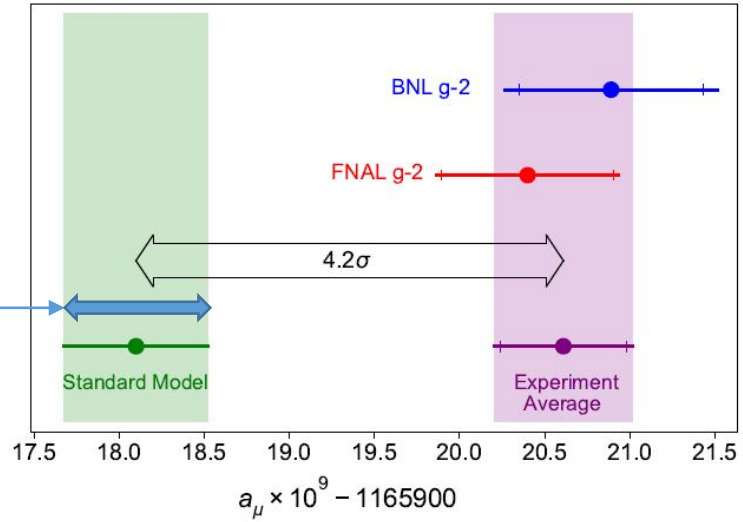
The chance of doing so was offered by my refereeing work of the detector proposed by the MUonE collaboration [14], which aims at determining with high precision the cross section of elastic muon-electron scattering.

In the cited study I demonstrated, through the direct exploration of the parameter space of detector geometry, how **large gains** in suitable utility functions (related to the resolution in the event q^2) can be obtained by moving away from choices dictated by past experience

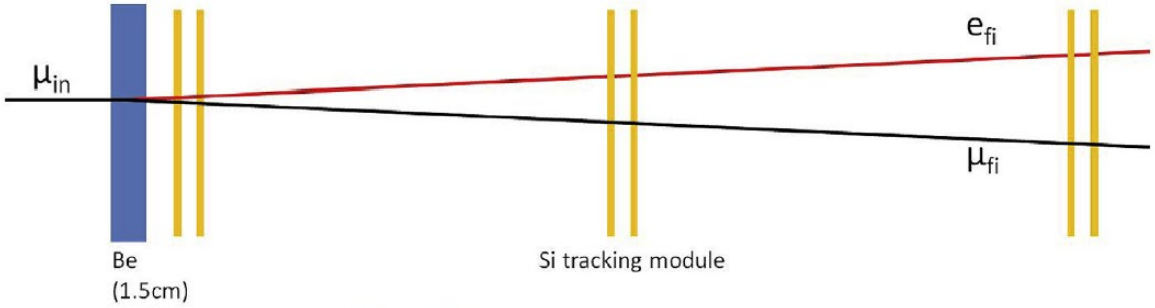
One example of geometry optimization: MUonE

MUonE [14] aims to determine with high precision the muon-electron elastic scattering differential cross section, to extract hadronic contributions and reduce the systematics of the g-2 muon anomaly

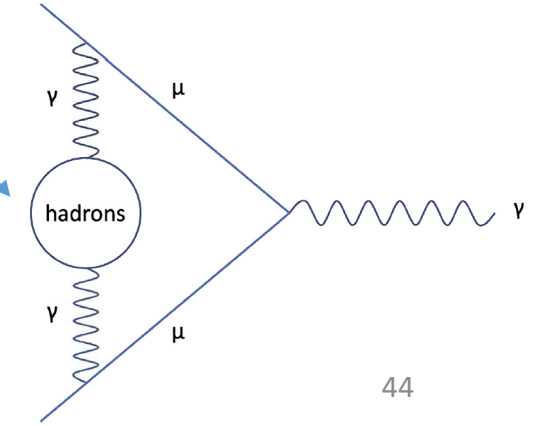
The experiment must be sensitive to hadronic loop effects particularly at high energy, where a 10^{-4} measurement may substantially improve the theoretical understanding of the g-2 value



Above: the largest known anomaly in the Standard Model to date, g-2



Above: layout of one of 40 1m-long stations



Right: A virtual hadronic loop

MUonE optimization

After unsuccessfully asking MUonE members to study alternative layouts, I wrote a simulation of muon scatterings and of event reconstruction, and then optimized the geometry of the apparatus.

This proved how **a factor of 2 improvement in the relevant metric could be achieved** without increase in detector cost.

ELSEVIER

Physics Open
Volume 4, September 2020, 100022

Geometry optimization of a muon-electron scattering detector

Tommaso Dorigo [1]

Show more

https://doi.org/10.1016/j.physo.2020.100022

Abstract

A high-statistics determination of the differential cross section of elastic muon-electron scattering as a function of the transferred four-momentum squared, $d\sigma_{el}(\mu e \to \mu e)/dq^2$, has been argued to provide an effective constraint to the hadronic contribution to the running of the fine-structure constant, $\Delta\alpha_{had}$, a crucial input for precise theoretical predictions of the anomalous magnetic moment of the muon. An experiment called "MUonE" is being planned at the north area of CERN for that purpose. We consider the geometry of the detector proposed by the MUonE collaboration and offer a few suggestions on the layout of the passive target material and on the placement of silicon strip sensors, based on a fast simulation of elastic muon-electron scattering events and the investigation of a number of possible solutions for the detector geometry. The employed methodology for detector

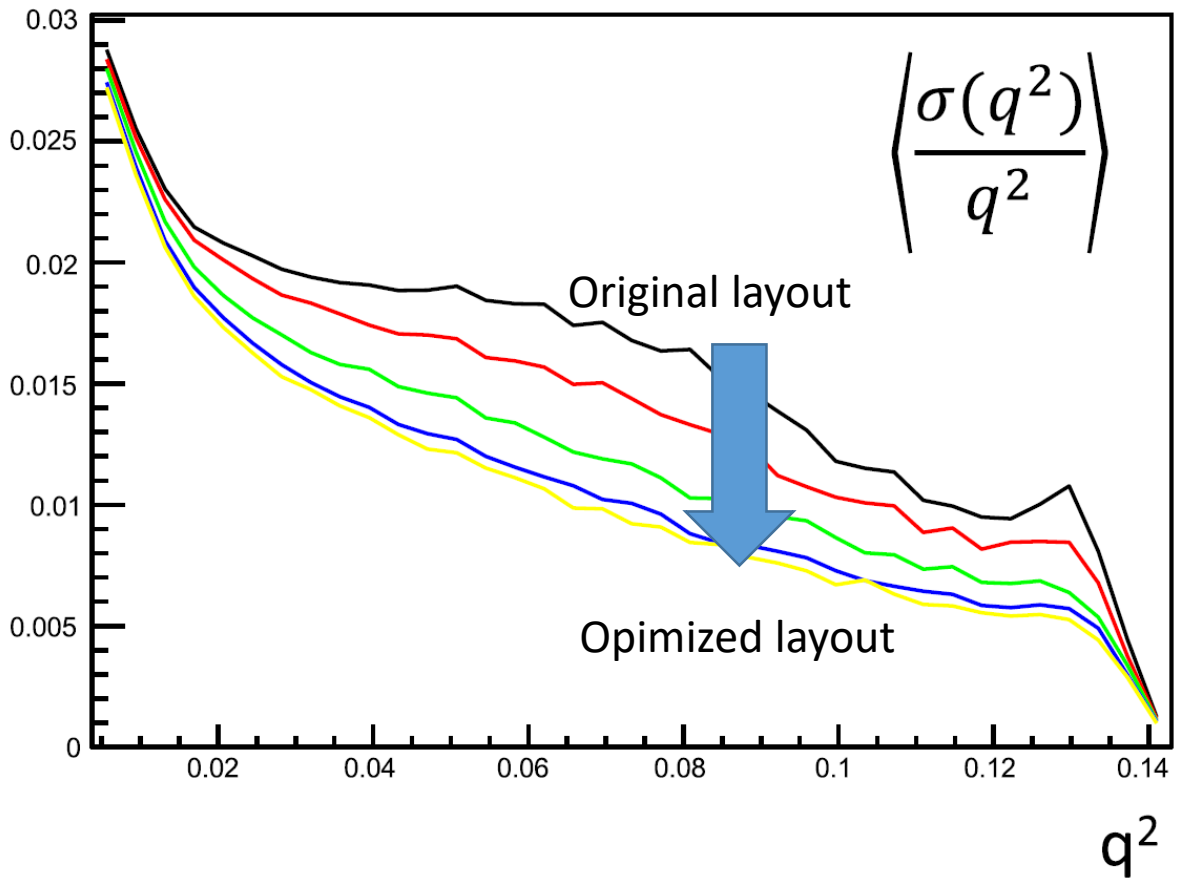
MUonE optimization

After unsuccessfully asking MUonE members to study alternative layouts, I wrote a simulation of muon scatterings and of event reconstruction, and then optimized the geometry of the apparatus.

This proved how **a factor of 2 improvement in the relevant metric could be achieved** without increase in detector cost.

We can only guess how large are the gains in the final experimental objectives possible if a fully differentiable model is created for detectors of significantly higher complexity than MUonE.

→ **My guess: huge.**



Above: relative resolution in event q^2 for different configurations (the higher, black line is the original proposal by the MUonE coll.)

Computer science to the rescue

Progress in CS redefined performance standards of our technologies.

We can today identify AI ingredients in, *e.g.*, language translation, speech recognition, self-driving vehicles.

→ Of course, that AI is not general but application-specific: its potential of providing new solutions to old tasks **depends on our ability to create the right interfaces**.

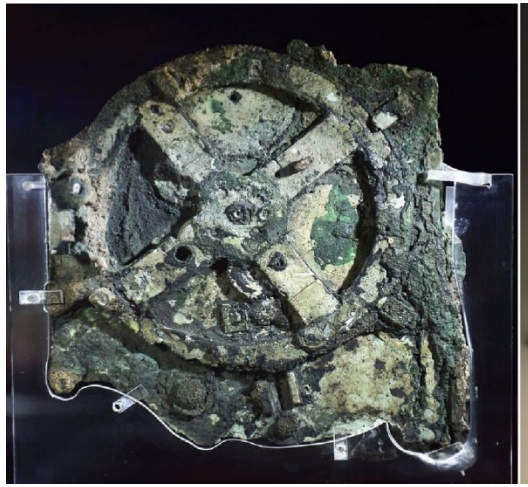
A new paradigm shift is offered by differentiable programming [15], which eases the systematic search of minima of arbitrarily complex multi-dimensional functions; by casting the whole problem in a differentiable framework a full end-to-end optimization becomes possible.

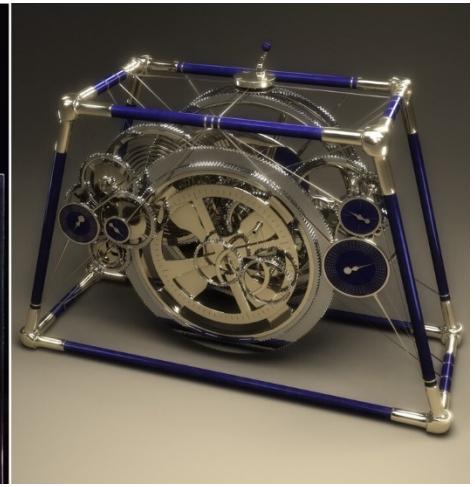
When technology exists, and is not used…

The 1901 discovery of a wrecked ship in front of Antikythera, a Greek island south of the Peloponnese, brought evidence that the ancient Greeks could craft very complex instruments to predict eclipses and the position of planets, moon and sun at the turn of a crank

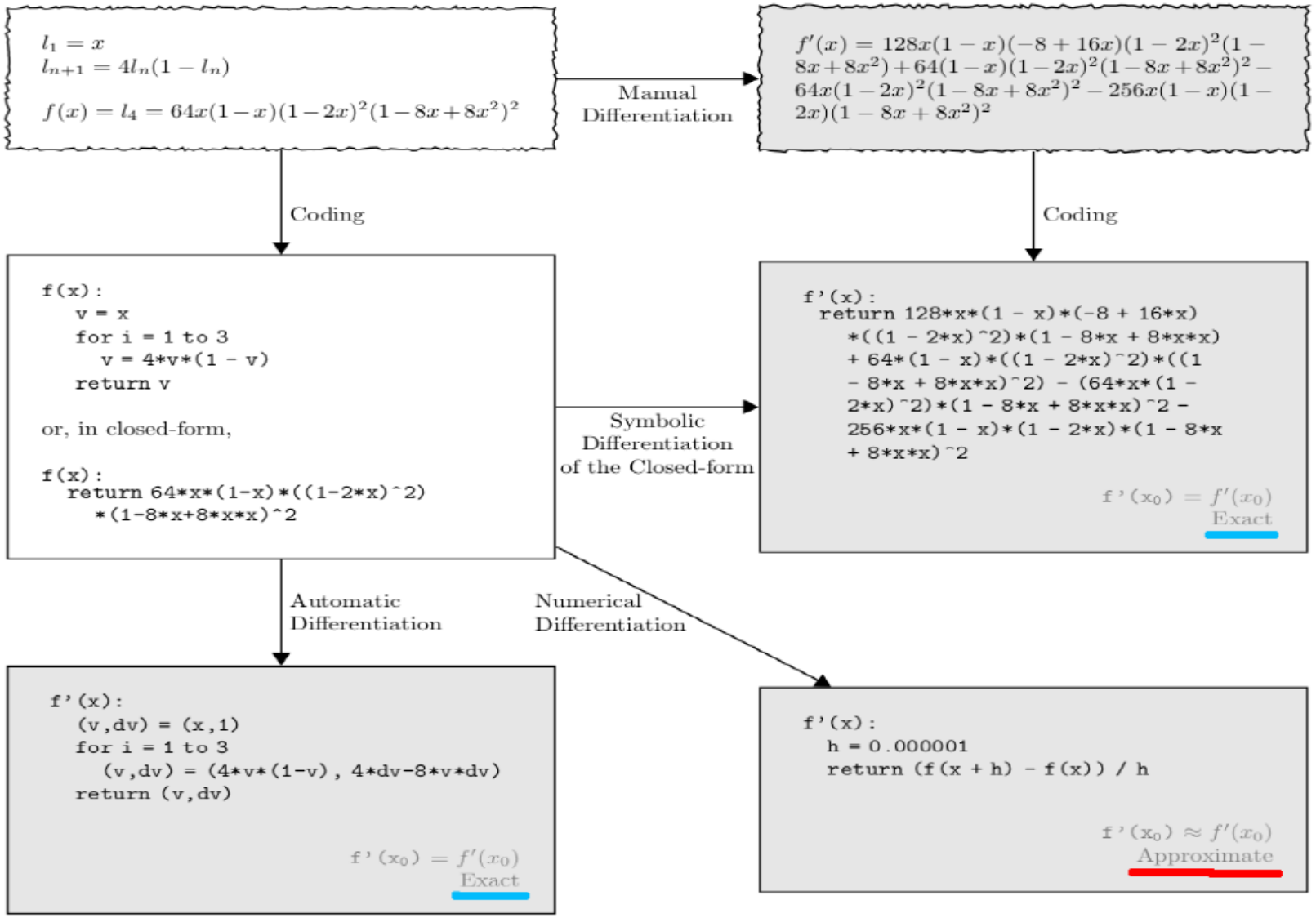
It has been argued that such technological marvels could have spurred a revolution that would have brought us to the technology level we're at now in 300 years, but it did not – there were not enough use cases to attract those early inventor's wit, nor resources to support basic research.

I'm not saying we run the same risks as the Greeks (and humanity) of 2500 years ago, but revolutions do not start by themselves – we have to take the new technology and adapt it to our use cases!





Automatic Differentiation



Wouldn't it be nice if you coded a problem and the dependence of variables in a program, and the language took care of figuring out how functions vary depending on parameters, and carry out the complicated task of propagating derivatives around?

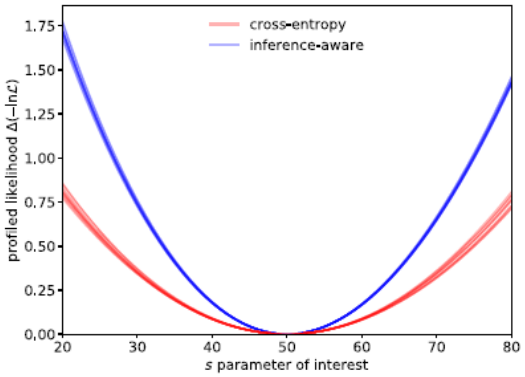
Those of us who have done this manually can't be happier by seeing the rise of Pytorch, TF, etc.

Image from [Baydin 2018]
Günes Baydin et al, JMLR 18 (2018) 1–43

INFERNO

As an example of what differentiable programming can do for fundamental research, I designed with **P. de Castro** an innovative algorithm [16] using automatic differentiation to construct a loss function that **directly targets the end result of the analysis** instead of an intermediate one.

If the loss function is constructed to incorporate the effects of nuisance parameters on the measurement objective, virtual optimality of the classification task and **large improvements in precision** can be achieved



(b) profile-likelihood comparison

Left: profile likelihood on the parameter of interest for a neural network with (blue) and without (red) the feedback on effect of nuisances provided by INFERNO

ELSEVIER

Computer Physics Communications
Volume 244, November 2019, Pages 170-179

INFERNO: Inference-Aware Neural Optimisation

Pablo de Castro, Tommaso Dorigo

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.cpc.2019.06.007

Abstract

Complex computer simulations are commonly required for accurate data modelling in many scientific disciplines, making statistical inference challenging due to the intractability of the likelihood evaluation for the observed data. Furthermore, sometimes one is interested on inference drawn over a subset of the generative model parameters while taking into account model uncertainty or misspecification on the remaining nuisance parameters. In this work, we show how non-linear summary statistics can be constructed by minimising inference-motivated losses via stochastic gradient descent such that they provide the smallest uncertainty for the parameters of interest. As a use case, the problem of confidence interval estimation for the mixture coefficient in a multi-dimensional two-component mixture model (i.e. signal vs background) is considered, where the proposed technique clearly outperforms summary statistics based on probabilistic classification, a commonly used alternative which does not account for the presence of nuisance parameters.

A study of muon shielding in SHIP

In another seminal work [17] by MODE members in HSE, local generative surrogates of the gradient of the objective function were proven to allow for the minimization by SGD and a strong reduction in muon background fluxes in the SHIP experiment

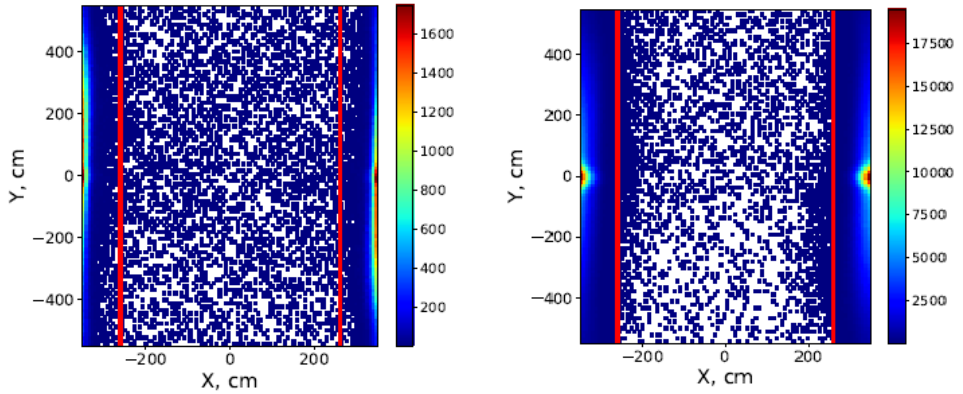
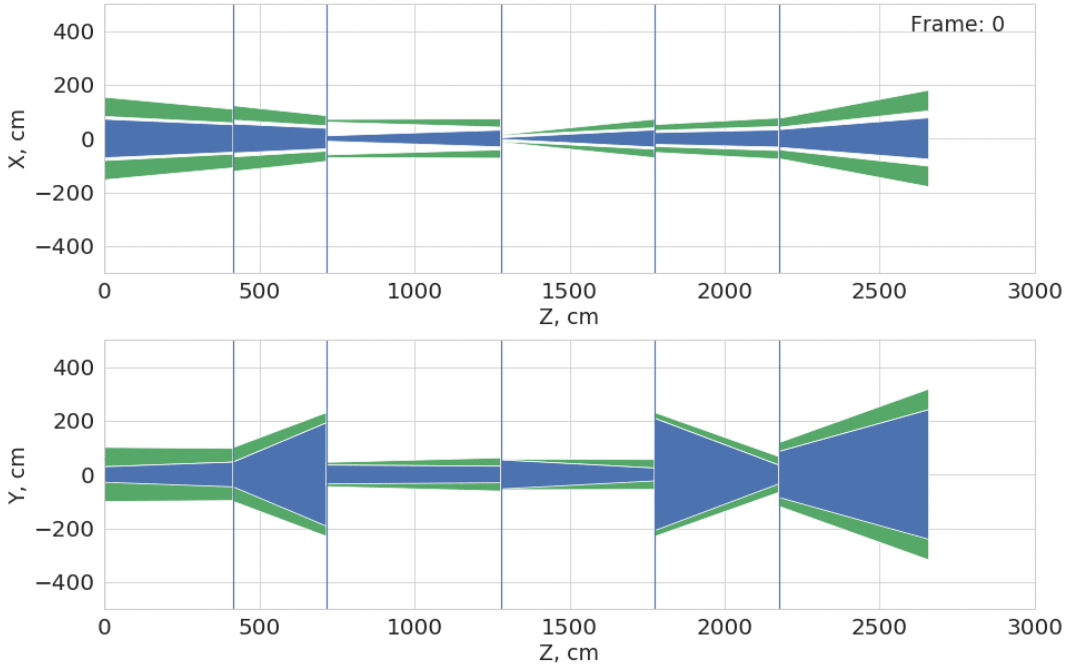
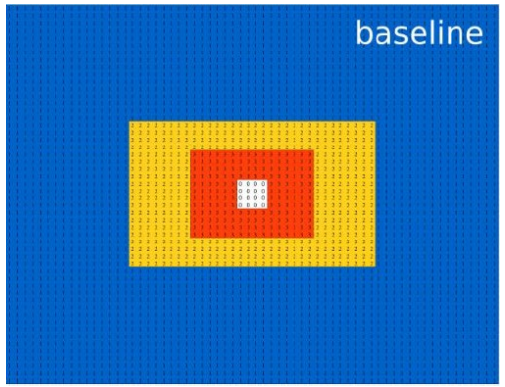


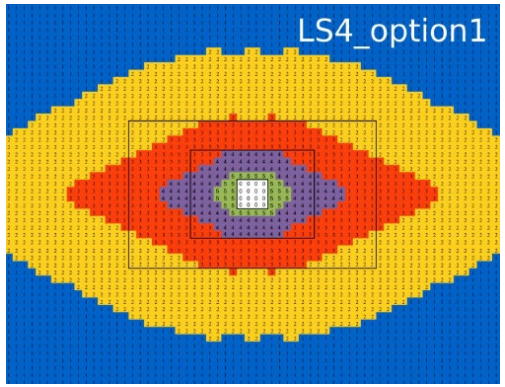
Figure 7. Muon hits distribution in the detection apparatus (depicted as red contour) obtained by Bayesian optimization (Left) and by L-GSO (Right), showing better distribution. Color represents number of the hits in a bin.



Geometry optimization at work in **real time**!

LHCb EM calorimeter optimization

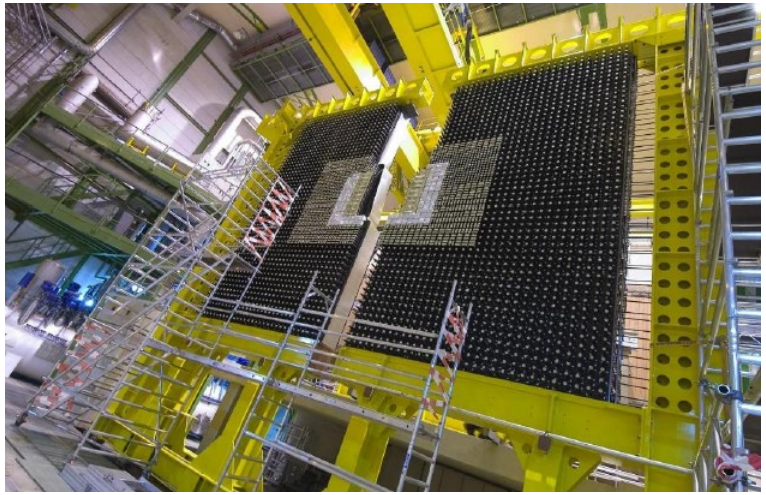


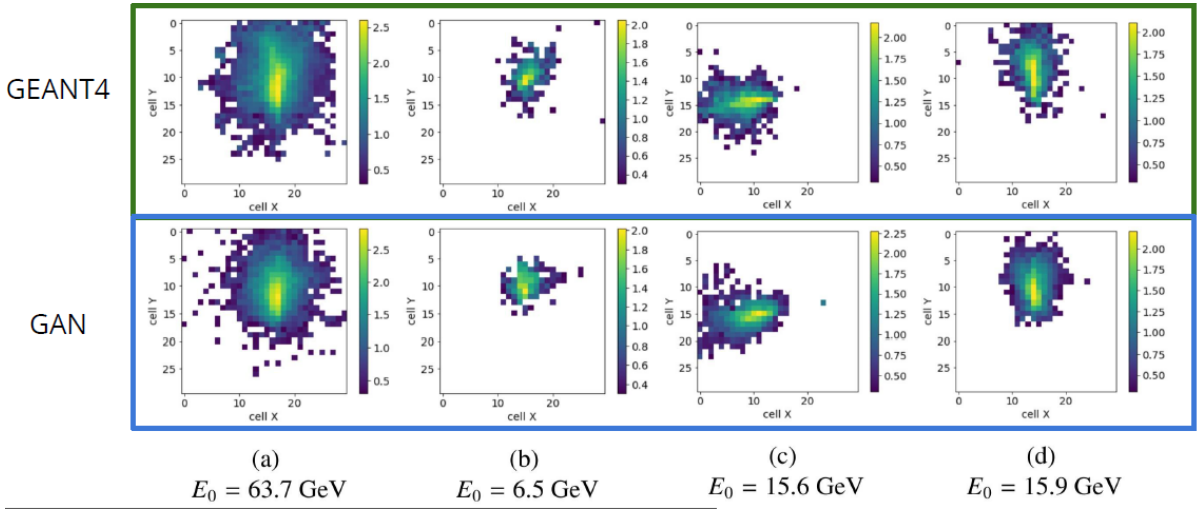


MODE members who collaborate with LHCb (**Alexey Boldyrev, Fedor Ratnikov, Andrey Ustyuzhanin, Denis Derkach**) are studying the optimization of the electromagnetic calorimeter for the LHCb upgrade, using a differentiable model of the detector

Creating a valid model of showers is a hard problem, which can be solved with GANs

The model allows to investigate how to best arrange modules of three different kinds optimizing cost and performance





Machine-Learning Optimized Design of Experiments

MODE Collaboration

https://mode-collaboration.github.io

A. G. Baydin[5], A. Boldyrev[4], K. Cranmer[8], P. de Castro Manzano[1], T. Dorigo[1], C. Delaere[2], D. Derkach[4], J. Donini[3], A. Giammanco[2], J. Kieseler[7], G. Louppe[6], L. Layer[1], P. Martinez Ruiz del Arbol[9], F. Ratnikov[4], G. Strong[1], M. Tosi[1], A. Ustyuzhanin[4], P. Vischia[2], H. Yarar[1]

1 INFN, Sezione di Padova (and associates from Padova and Naples Universities), Italy
2 Université Catholique de Louvain, Belgium
3 Université Clermont Auvergne, France
4 Laboratory for big data analysis of the Higher School of Economics, Russia
5 University of Oxford
6 Université de Liege
7 CERN
8 New York University
9 IFCA

Realigning design choices and ultimate goals

The target of **MODE** is to design a scalable, versatile architecture that can provide end-to-end optimization of particle detectors, proving it on a number of different applications across different domains.

The key observation is that while the detectors of HEP, astro-HEP etc. are quite varied and their purposes are diverse, the inference extraction process, and therefore its modeling, share marked similarities across the board

-- like the gears of the Antikythera mechanism!

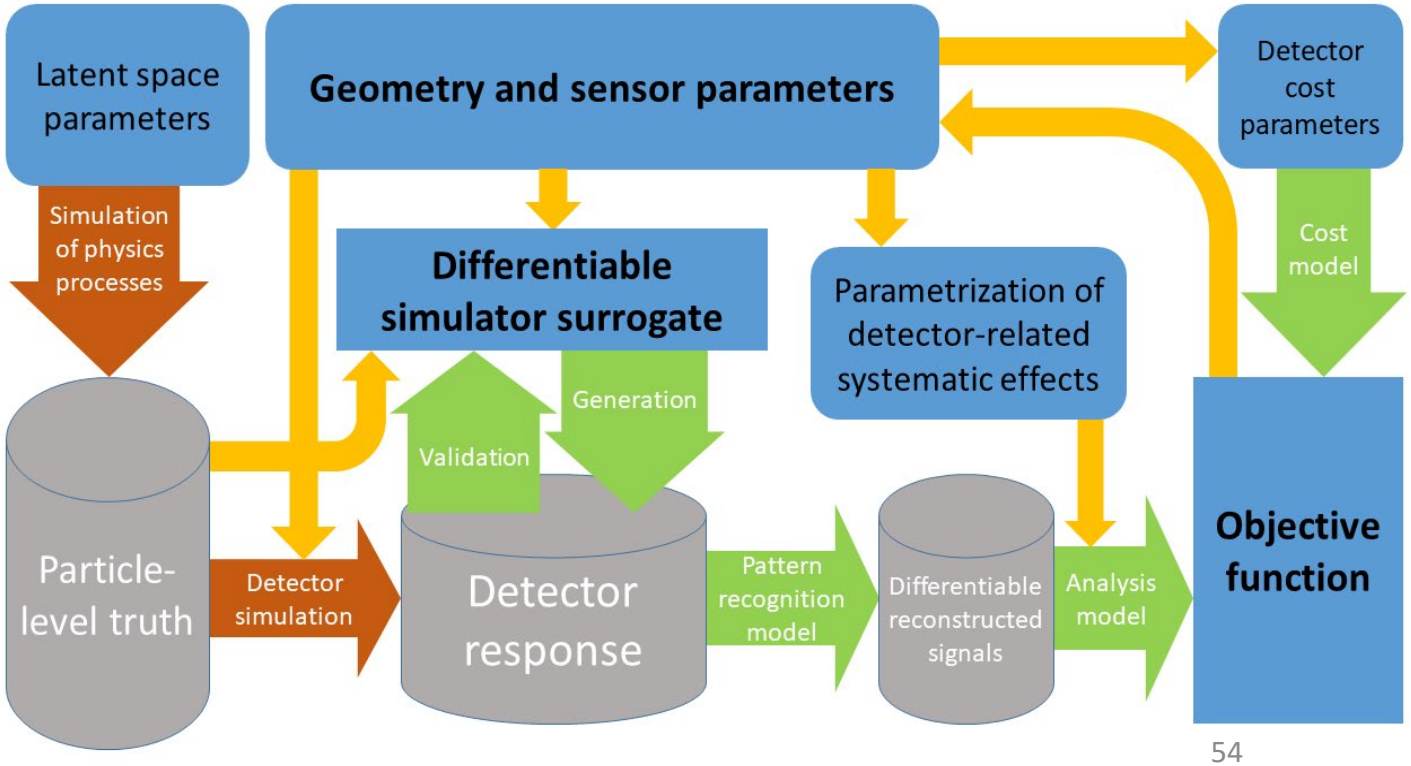
→ If we "solve" a few problems we may construct a library of solutions and exploit the universality of the architecture and its modularity, re-using modeling efforts

Initial study cases:

- Optimization of inference → in progress
- Muon tomography detector optimization [18] → in progress
- Hybrid calorimeter design integrating tracking layers → starting

Other use cases being considered include:

- Muon collider detector shielding [20];
- Optimization of detectors for air Cherenkov showers (SWGO)
- plus many more



Four-slide description of a differentiable model for design optimization - 1

An end-to-end detector design optimization task can be briefly formalized in the following way.

We start with a simulation of the physics processes of relevance for the considered application, which generates a multi-dimensional, stochastic input variable **x**, distributed with a PDF **f(x)**.

The input is turned by the simulation of the detection apparatus into sensor readouts **z** distributed with a PDF

$$\mathbf{p(z|x,\theta)},$$

which constitute the observed low-level features of the physical process; readouts **z** depend through **p( )** on parameters **θ** that describe the physical properties of the detector and its geometry.

Four-slide description of a differentiable model for design optimization - 2

The observations **z** are used by a reconstruction model **R( )** that produces high-level features

$$\zeta(\theta) = R[z,\theta,\nu(\theta)]$$

(*e.g.* particle four-momenta), by employing knowledge of the detector parameters as well as a model of the detector-driven nuisance parameters **ν(θ)** which affect the pattern recognition task.

In turn, high-level features **ζ(θ)** constitute the input of a further, less dramatic, dimensionality reduction, the data analysis step: this is typically performed by a classifier or regressor **NN( )** powered by a neural network.

Once properly trained for the task at hand, the network produces a low-dimensional summary statistic

$$s = NN[\zeta(\theta)]$$

with which inference can finally be carried out to produce the desired goal of the experiment.

Four-slide description of a differentiable model for design optimization - 3

In general, one may formally specify the problem of identifying optimal detector parameters as that of finding estimators $\widehat{\theta}$ that satisfy

$$\widehat{\theta} = arg \min_{\theta} \int L[NN(\zeta), c(\theta)] p(z|x, \theta) f(x) dx dz$$

c(θ) is a function modeling the cost of the considered detector layout of parameters **θ**, and the loss function **L[NN,c]** is constructed to appropriately weight the result of the measurement in terms of its desirable goals, as well as to obey cost constraints and other use-case-specific limitations.

Since in the cases of interest the PDF **p(z|x,θ)** is not available in closed form –the considered models are implicit–, we must rely on forward simulation: we approximate $\widehat{\theta}$ with a sample of n events:

$$\widehat{\theta}_a = arg \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L\left[NN\left(R(z_i)\right), c(\theta)\right]$$

where z_i is distributed as $F(x_i,\theta)$ to emulate **p(z|x,θ)** as x_i is sampled from its PDF **f( )** by the simulator. One may thus obtain an estimate of the loss function and the detector parameters which minimize it.

Four-slide description of a differentiable model for design optimization - 4

It has been shown how in applications such as those of our interest it is viable to approximate the non-differentiable stochastic simulator **F( )** with a local surrogate model,

$$z = S(y,x,\theta),$$

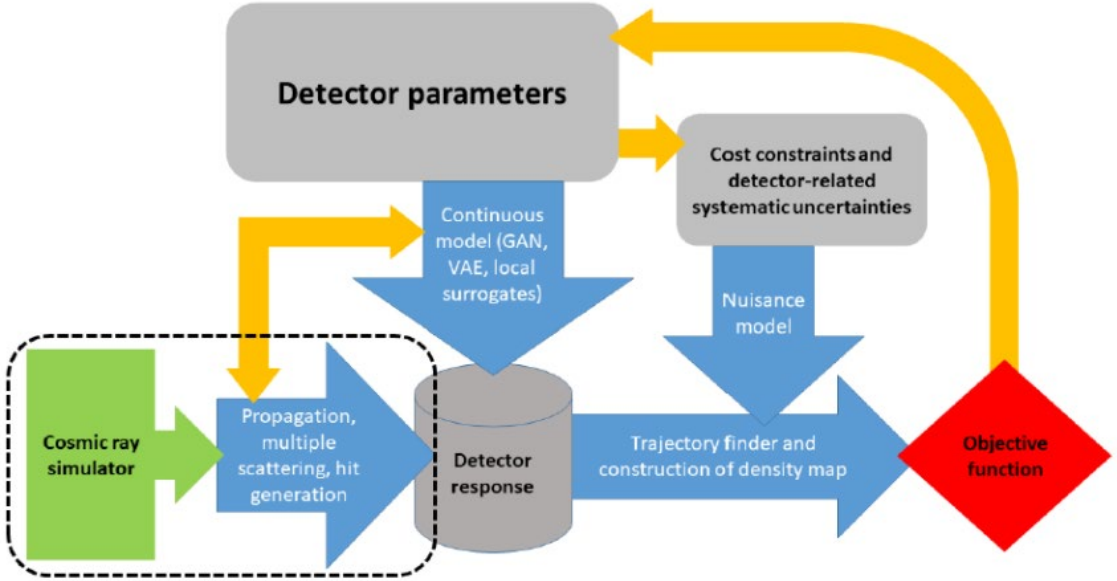
that depends on a parameter **y** describing the stochastic variation of the approximated distribution[XX]. This allows to descend to the minimum of the approximated loss $\widehat{L(z)}$ by following its surrogate gradient

$$\nabla_\theta \widehat{L(z)} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta L[NN\left(R\left(S(y_i, x_i, \theta)\right)\right), c(\theta)].$$

The above recipe requires one to learn the differentiable surrogate **S( )**: this is a task liable to be carried out independently from the optimization procedure.

The modular structure of a differentiable pipeline modeling the optimization cycle allows the user to turn on and off specific parts of the chain, helping the system in its exploration of the feature space.

And for the time being…



«Simple» use case: muon tomography. We need no surrogate of a simulator, yet all other pieces of the puzzle still need to be carved and set in.

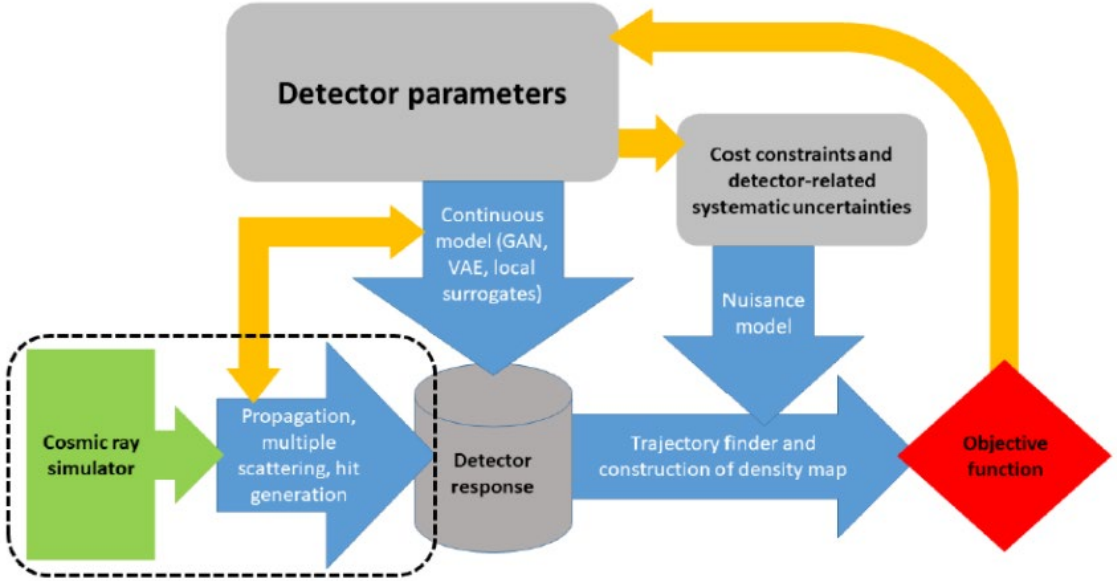
Muon tomography consists in exploiting the flux of cosmic rays that impinge constantly on the Earth's surface, as a source of data with which to produce inference (or directly imaging) on the contents of unknown volumes of material

Applications are countless:

- study core of volcanos
- prevent smuggling of dangerous materials in containers
- study wear of industrial apparata (e.g. steel pipes, furnace material)
- archaeological prospections
- and many more

Two main methods for inference:
1 – exploit dependence of scattering angle of muons, θ, on atomic number of material, Z
2 – use incoming momentum spectrum of muons and absorption of parts of it by thick matter layers to infer thickness (in 3D, with multiple scans)

And for the time being…



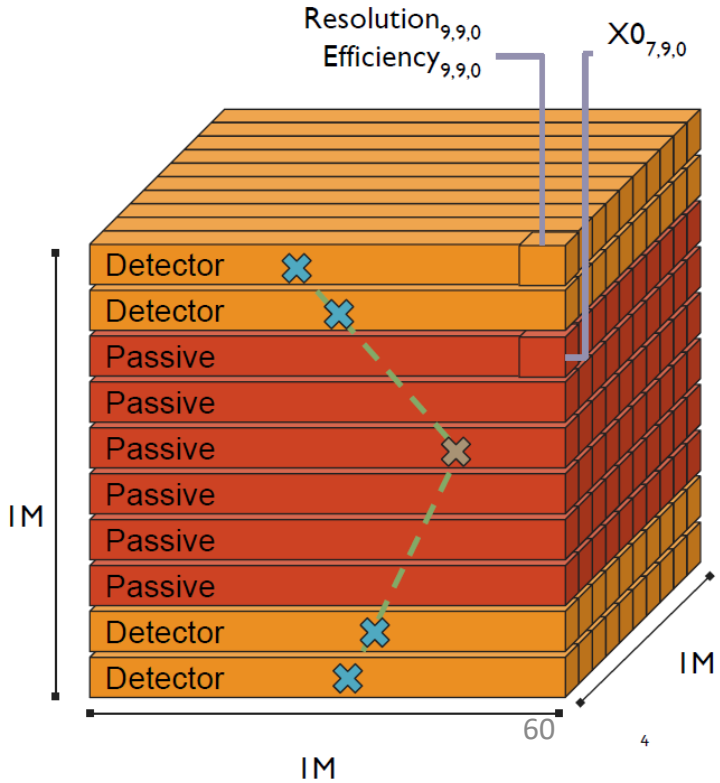
«Simple» use case: muon tomography. We need no surrogate of a simulator, yet all other pieces of the puzzle still need to be carved and set in.

For a simple test, we model a scanned volume of unknown density (*e.g.* one including a high-Z block of 0.5x0.1x0.1 m^3 somewhere inside a 0.6x1x1m^3 of low-Z material)

The system «learns» how to compromise cost and precision to optimize the inference on the Z map, and where detector elements are less useful

Still under development, but promising first results already achieved (see next slide)

Right: scheme of the modeled apparatus (graph courtesy **G. C. Strong**)



VERY preliminary results

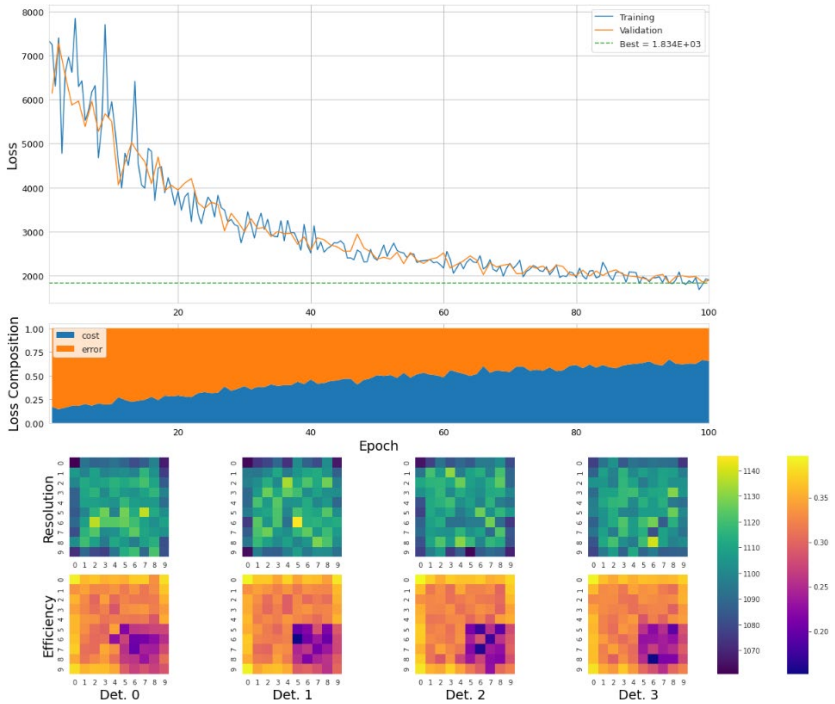
The code and results shown have been produced by **Giles C. Strong**

These graphs show the result of a run of 100 epochs training, followed by a prediction with 100k muons

First proof of principle of correct training of a differentiable model of a schematic muon tomography apparatus.

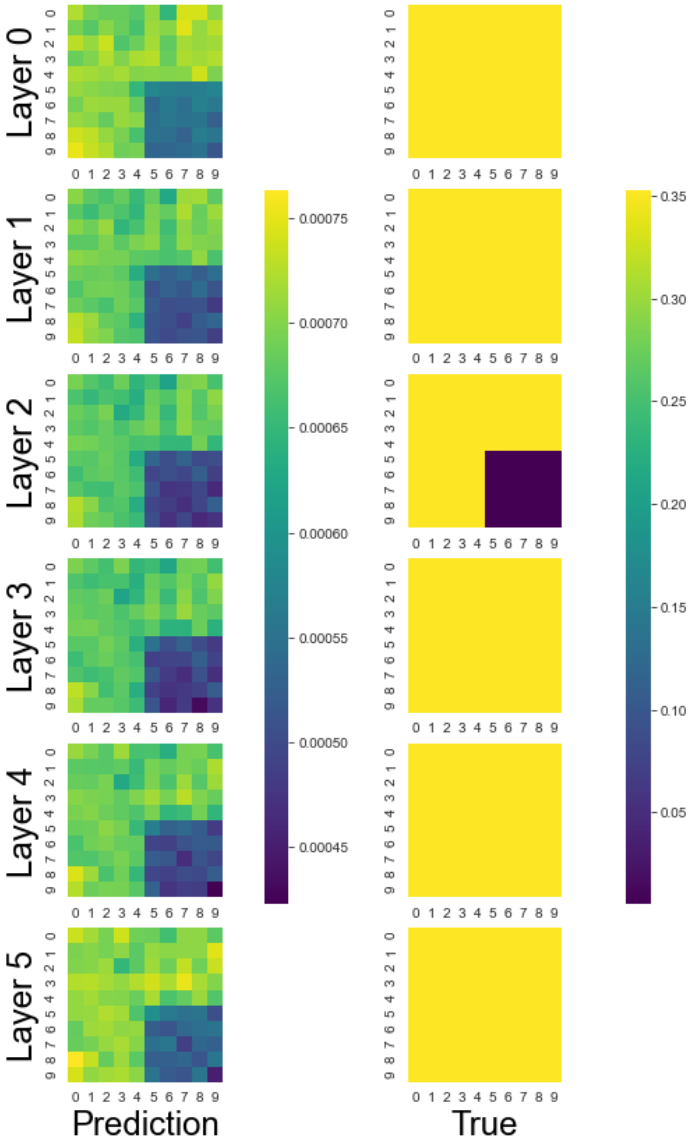
The **loss** is a combination of detector cost (itself a function of sensors efficiency and resolution) and RMSE on rad length estimate

Still a long way to go, but an important milestone for this use case



Above, top to bottom: loss, loss composition, resolution map, and efficiency map of detection elements after minimization.

Right: predicted and true X_0 of passive volume



Status and next steps

An article describing the MODE program has been published last month in Nuclear Physics News International [21]

- A more detailed "white paper" on differentiable programming for detector design optimization is being drafted

We are organizing a workshop on *"Differentiable Programming for Design Optimization"* on September 6-8 2021 in Louvain-la-Neuve, to allow interested scientists to join and discuss together the means and the possible applications

- Extra support for this activity is provided by IRIS-HEP and JENAA

You are welcome to join, propose applications, become a member of MODE

Every solved application = a publication AND added knowledge base on solving these hard problems!

Day 1, morning:

8.30-9.00 Introduction and goals of the workshop (T. Dorigo, INFN-Padova)
9.00-11.00 Progress in Computer Science (chair: Atilim Gunes Baydin, U. Oxford)
- Keynote lecture on deep learning
- Keynote lecture on differentiable programming applications to design optimization
- other contributions by CS experts

11.00-11.30 break
11.30-13.30 Applications in muon tomography and hadron therapy (chair: Andrea Giammanco, U. Cath. Louvain)

Day 1, afternoon:

14.30-16.30 Applications and requirements for particle physics (chair: Pietro Vischia, U. Cath. Louvain)
16.30-17.00 break
17.00-19.00 Applications and requirements in astro-HEP (chair: Roberto Ruiz de Austri Bazan, IFIC Valencia)

Day 2, morning:

9.00-11.00 Applications and requirements for neutrino detectors (chair: Kazuhiro Terao, Columbia U.)
11.00-11.30 break
11.30-13.30 Applications and requirements in nuclear physics experiments (chair: Gian Michele Innocenti, CERN)

Day 2, afternoon:

14.30-15.30 Round table discussion, formation of working groups
15.30-17.00 (Parallel) Specification of tasks and goals of the working groups
17.00-17.30 break
17.30-19.00 Organization of future activities and closing of the workshop (chair: T.Dorigo, INFN-Padova)

1st MODE Workshop, September 6-8, 2021

Funded by **IRIS-HEP** and **JENAA** (Appec-NuPecc-ECFA)

Will bring together computer scientists and physicists to focus on synergies with our use cases, and discuss cutting-edge techniques for simulation, inference, modeling for differentiable programming applications

Sessions targeting HEP, Astro-HEP, Nuclear, and Neutrino communities

Still accepting registrants at https://indico.cern.ch/event/1022938/

NPNI article

A short «advertising» article describing the plans of the MODE collaboration has been published in Nuclear Physics News International (paper version out this fall, already available online[21])

Nuclear Physics News

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gnpn20

Toward Machine Learning Optimization of Experimental Design

Atılım Güneş Baydin, Kyle Cranmer, Pablo de Castro Manzano, Christophe Delaere, Denis Derkach, Julien Donini, Tommaso Dorigo, Andrea Giammanco, Jan Kieseler, Lukas Layer, Gilles Louppe, Fedor Ratnikov, Giles C. Strong, Mia Tosi, Andrey Ustyuzhanin, Pietro Vischia & Hevjin Yarar

To cite this article: Atılım Güneş Baydin, Kyle Cranmer, Pablo de Castro Manzano, Christophe Delaere, Denis Derkach, Julien Donini, Tommaso Dorigo, Andrea Giammanco, Jan Kieseler, Lukas Layer, Gilles Louppe, Fedor Ratnikov, Giles C. Strong, Mia Tosi, Andrey Ustyuzhanin, Pietro Vischia & Hevjin Yarar (2021) Toward Machine Learning Optimization of Experimental Design, Nuclear Physics News, 31:1, 25-28, DOI: 10.1080/10619127.2021.1881364

To link to this article: https://doi.org/10.1080/10619127.2021.1881364

Published online: 30 Mar 2021.

Submit your article to this journal

THANK YOU FOR YOUR ATTENTION!

References

- **Wechsler 1944**: *D. Wechsler, The measurement of adult intelligence,* Baltimore: Williams & Wilkins (1944) ISBN 978-0-19-502296-4
- **Feuerstein 1990**: *R. Feuerstein et al., Dynamic assessments of cognitive modifiability,* ICELP Press (1990)
- **Legg 2007**: S. Legg and M. Hutter, *A Collection of Definitions of Intelligence* , Proc. AGI Workshop 206, **157** (2007). ISBN 9781586037581
- **Sternberg 1982**: R.J. Sternberg and W. Salter, *Handbook of human intelligence,* Cambridge University Press (1982) ISBN 978-0-521-29687-8
- **McCorduck 2004**: P. McCorduck, *Machines who think*, AK Peters (2004), ISBN 1-56881-205-1
- **Church 1936**: A. Church, *An Unsolvable Problem of Elementary Number Theory,* American Journal of Mathematics **58** (2) 345 (1936), doi:10.2307/2371045
- **CMS 2012**: CMS collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B 716 (2012) 30, doi:10.1016/j.physletb.2012.08.021
- **Turing 1937**: A.M. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem*, *Proceedings of the London Mathematical Society*, 2, **42**, 230 (1936), doi:10.1112/plms/s2-42.1.230
- **Pitts 1943**: W.S. McCulloch and W. Pitts, *A logical calculus of the ideas immanent in nervous activity,* Bulletin of Mathematical Biophysics. **5** (4) 115 (1943), doi:10.1007/bf02478259
- **Turing 1950**: A.M. Turing, *Computing Machinery and Intelligence,* Mind **LIX** (236) 433 (1950), doi:10.1093/mind/LIX.236.433

References /2

- **Simon 1958**: H.A. Simon and A. Newell, *Heuristic Problem Solving: The Next Advance in Operations Research*, Operations Research, **6** (1958), doi:10.1287/opre.6.1.1
- **Minsky 1967**: M. Minsky, *Computation: Finite and Infinite Machines*, Englewood Cliffs, N.J.: Prentice-Hall (1967)
- **Rosenblatt 1958**: F. Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review **65** (6) 386–408 (1958), doi:10.1037/h0042519
- **Minsky 1969**: M. Minsky and S. Papert, *Perceptrons, an introduction to computational geometry*, MIT press (1969), ISBN: 9780262630221.
- **Mitchell 1997**: T. Mitchell, *Machine Learning*, McGraw Hill. (1997). ISBN 978-0-07-042807-2.
- **Kaggle 2014**: https://www.kaggle.com/c/higgs-boson
- **Strong 2020**: Giles C. Strong, "*On the impact of selected modern deep-learning techniques to the performance and celerity of classification models in an experimental high-energy physics use case*", arXiv: 2002.01427 [physics.data-an] (2020).
- **Kasieczka 2019**: G. Kasieczka et al., *The Machine Learning Landscape of Top Taggers*, SciPost Phys. 7, 014 (2019).
- **Kieseler 2020**: J.Kieseler, arxiv:2002.03605, EPJC
- **De Castro 2020**: arXiv:2007.09121(stat.ML), to be published by World Scientific.
- **Dorigo 2019**: P. de Castro Manzano and T. Dorigo, INFERNO: Inference-Aware Neural Optimization, Computer Physics Comm. 244 (2019) 170, 10.1016/j.cpc.2019.06.007
- **Baydin 2018**: Günes Baydin et al, JMLR 18 (2018) 1–43.
- **XiLINK 2020**: https://www.xilinx.com/publications/powered-by-xilinx/cerncasestudy-final.pdf

References / 3

1. The European Strategy Group, "*2020 Update of the European Strategy for Particle Physics*," CERN, Geneva, https://europeanstrategy.cern/, doi: http://dx.doi.org/10.17181/ESU2020Deliberation, 2020.

2. The topic has been discussed widely in blogs and other social media. *E.g.*, see Sabine Hossenfelder's discussion here: http://backreaction.blogspot.com/2019/06/if-we-spend-money-on-larger-particle.html, or see Alessandro Strumia's guest post at the same site: http://backreaction.blogspot.com/2020/06/guest-post-who-needs-giant-new-collider.html. Also see the New York Times op-ed by S. Hossenfelder, https://www.nytimes.com/2019/01/23/opinion/particle-physics-large-hadron-collider.html

3. F. Hartmann, *Evolution of Silicon Sensor Technology in Particle Physics*, Springer Tracts in Modern Physics, 2017, ISBN 978-3-319-64436-3; P. Delpierre, *A history of hybrid pixel detectors, from high energy physics to medical imaging*, J. INST. 9 (2014) C05059, doi: 10.1088/1748-0221/9/05/C05059; P. Allport, *Applications of silicon strip and pixel-based particle tracking detectors*, Nature Reviews Physics 1 (2019) 567–576, doi: https://doi.org/10.1038/s42254-019-0081-z; G. Case, *New trends in silicon detector technology*, JINST 15 (2020), https://iopscience.iop.org/article/10.1088/1748-0221/15/05/C05057; M. Garcia-Sciveres and N. Wermes, *A review of advances in pixel detectors for experiments with high rate and radiation*, Rep. Prog. Phys. 81 (2018) 066101, doi: http://doi.org/10.1088/1361-6633/aab064.

References / 4

4. Y. Mishnayot *et al*., "*3D Printing of Scintillating Materials*", arXiv:1406.4817[cond-mat.mtrl-sci] (2014); Y. Abreu *et al*. (SoLID Collaboration), "*A novel segmented-scintillator antineutrino detector*", JINST 12 (2017) P04024, doi: 10.1088/1748-0221/12/04/P04024.

5. M. Mandurrino *et al*., "*Demonstration of 200 --, 100 --, and 50 micron Pitch Resistive AC Coupled Silicon Detectors (RSD) With 100% Fill Factor for 4D Particle Tracking*," IEEE Electron Device Letters 40, 11 (2019) 1780, arXiv:1907.03314[physics.ins-det]; M. Mandurrino *et al*., "*Analysis and numerical design of Resistive AC Coupled Silicon Detectors (RSD) for 4D particle tracking*", Nucl. Instr. Meth. A959 (2020) 163479, doi: 10.1016/j.nima.2020.163479; M. Tornago et al., "*Resistive AC-Coupled Silicon Detectors: Principles of operation and first results from a combined analysis of beam test and laster data*", NIM A 1003 (2021) 165319, https://www.researchgate.net/publication/343095980 .

6. CMS Collaboration, "*The Phase-2 Upgrade of the CMS Endcap Calorimeter*", CERN-LHCC-2017-023 (2018), https://cds.cern.ch/record/2293646?ln=en .

References / 5

7. J. Repond *et al.*, "*Hadronic Energy Resolution of a Combined High Granularity Scintillator Calorimeter System*", JINST 13 (2018) P12022, doi: 10.1088/1748-0221/13/12/P12022; C. Adloff *et al.*, "*Tests of a particle flow algorithm with CALICE test beam data*", JINST 6 (2011) P07005, doi: 10.1088/1748-0221/6/07/P07005.

8. P.W. Cattaneo *et al.*, "*CaloCube: a novel calorimeter for high-energy cosmic rays in space*", JINST 12 (2017) 06, C06004, doi: 10.1088/1748-0221/12/06/C06004 .

9. E. Hoogeboom *et al.*, "*HexaConv*", arXiv:1803.02108[cs.LG] (2018)).

10. M. Sirunyan *et al.* (CMS Collaboration), "*Particle-flow reconstruction and global event description with the CMS detector*", JINST 12 (2017) P10003, doi: 10.1088/1748-0221/12/10/P10003 .

References / 6

11. See *e.g.* S.R. Qasim, **J. Kieseler**, Y. Iiyama, and M. Pierini, "*Learning representations of irregular particle-detector geometry with distance-weighted graph networks*", Eur. Phys. J. C79 (2019) 7,608, doi: 10.1140/epjc/s10052-019-7113-9; **J. Kieseler**, "*Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph and image data*", arXiv:2002.03605[physics.data-an] (2020); J. Alimena, Y. Iiymana, **J. Kieseler**, "*Fast convolutional neural networks for identifying long-lived particles in a high-granularity calorimeter*", arXiv:2004.10744 [hep-ex] (2020).

12. **T. Dorigo, J. Kieseler, L. Layer** and **G. Strong**, "*Muon Energy Measurement from Radiative Losses in a Calorimeter for a Collider Detector*", http://arxiv.org/abs/2008.10958 (2020).

13. **T. Dorigo**, "*Geometry Optimization of a Muon-Electron Scattering Detector*," Physics Open 4 (2020) 100022, arXiv:200200973[physics.ins-det], doi: 10.1016/j.physo.2020.100022.

14. G. Abbiendi *et al.*, "*Letter of Intent: The MUonE Project*", CERN-SPSC-2019-026/SPSC-I-252 (2019).

References / 7

15. **A. Güneş Baydin**, B.A. Pearlmutter, A.A. Radul, and J.M. Siskind, "*Automatic Differentiation in Machine Learning: a Survey*", Journal of Machine Learning Research (JMLR) 18 (153) (2018) 1, http://jmlr.org/papers/v18/17-468.html; Y. LeCun, Y. Bengio, and G. Hinton, "*Deep learning*", Nature, 521(7553) (2015) 436, doi: 10.1038/nature14539.

16. **P. de Castro Manzano** and **T. Dorigo**, "*INFERNO: Inference-Aware Neural Optimization*", Comp. Phys. Commun. 244 (2019) 170; Arxiv:1806.04743v2 [stat.ml] (2018), doi: 10.1016/j.cpc.2019.06.007.

17. S. Shirobokov, **A. Ustyuzhanin, A. Güneş Badyin** *et al.*, "*Differentiating the Black-Box: Optimization with Local Generative Surrogates*", arXiv:2002.04632v1[cs.LG] (2020).

18. S. Wuyckens, **A. Giammanco**, P. Demin, and E. Cortina Gil, "*A Portable muon telescope based on small and gas-tight Resistive Plate Chambers*", Phil. Trans. Royal Soc. A377 (2019) 2137, arXiv:1806.06602v2[physics.ins-det] (2018), doi: 10.1098/rsta.2018.0139.

References / 8

19. P. Giubilato *et al*., "*iMPACT: innovative pCT scanner*", IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC) IEEE (2015), https://ieeexplore.ieee.org/abstract/document/7581240; S. Mattiazzo, *et al*., "*Advanced proton imaging in computed tomography*", Radiation protection dosimetry 166.1 (2015) 388, https://academic.oup.com/rpd/article/166/1-4/388/1612689; S. Mattiazzo *et al*., "*The iMPACT project tracker and calorimeter*", Nucl. Instr. Meth. A845 (2017) 664, https://www.sciencedirect.com/science/article/pii/S0168900216303412; N. Pozzobon *et al*. "*Calorimeter prototyping for the iMPACT project pCT scanner*", Nucl. Instr. Meth. A936 (2019) 1, https://ui.adsabs.harvard.edu/abs/2019NIMPA.936....1P/abstract.

20. N. Bartosik *et al*., "*Preliminary Report on the Study of Beam-Induced Background Effects at a Muon Collider*," arXiv:1905.03725[hep-ex] (2019); N. Bartosik *et al.*, "*Detector and Physics Performance at a Muon Collider*," J. Inst. 15 (2020) P05001, doi: 10.1088/1748-0221/15/05/P05001.

References / 9

21. **A.G. Baydin** et al., *"Toward Machine Learning Optimization of Experimental Design"*, submitted to Nuclear Physics News International, 2021.

22. W. Sakumoto et al., "*Measurement of TeV muon energy loss in iron*",Phys. Rev. D45, 9 (1992) 3042.

23. The IceCUBE Collaboration, "*Muon energy reconstruction and atmospheric neutrino spectrum unfolding with the IceCube detector"*, proc. 30th International Cosmic-Ray Conference (2007), Merida, Mexico, arXiv:0711.0353v1

24. **P. de Castro Manzano**, https://github.com/pablodecm/paper-inferno

25. **G. C. Strong**, https://gilesstrong.github.io/pytorch_inferno/

Backup

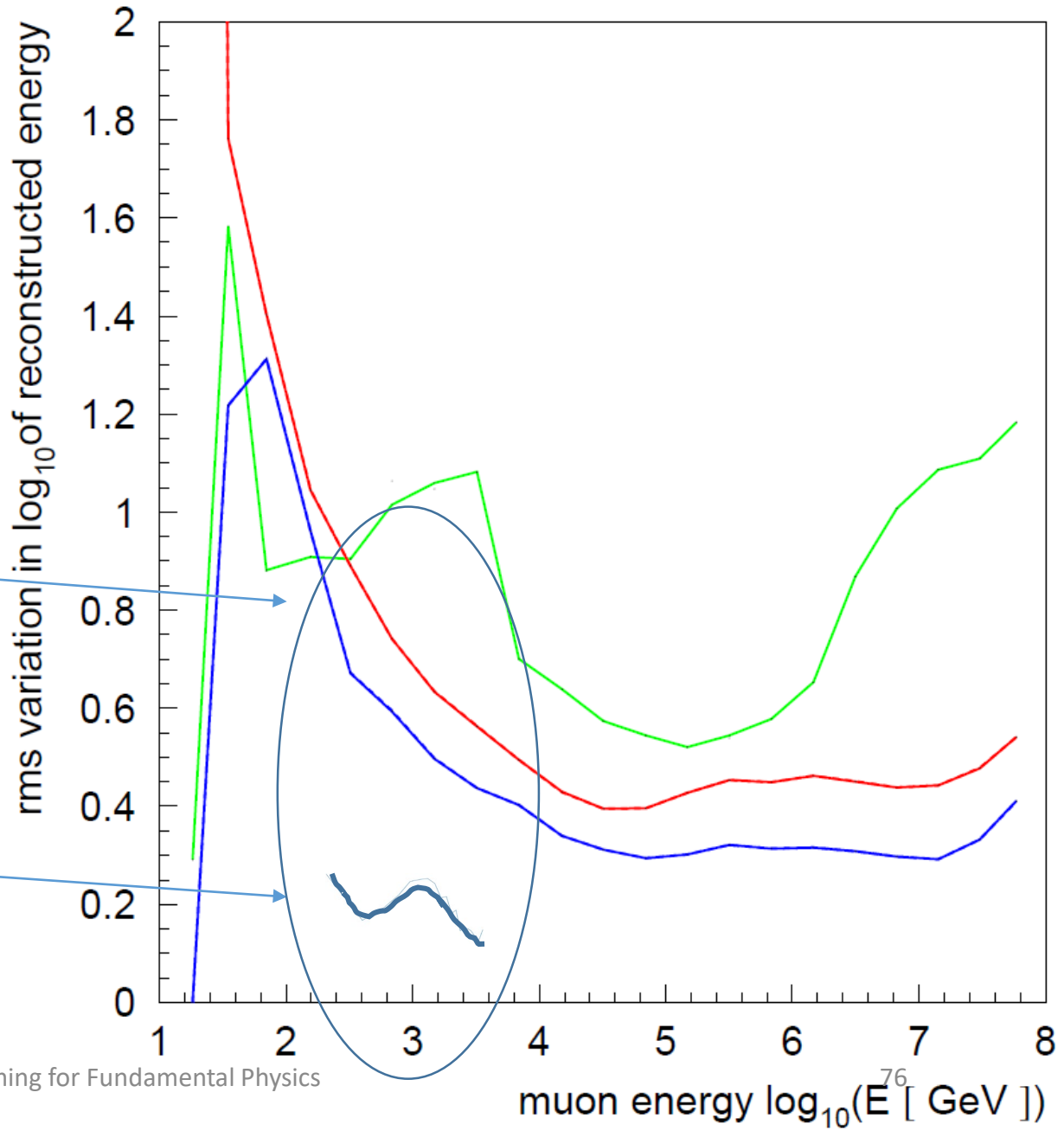
How good is that, BTW?

Measuring multi-TeV muons has been an issue before LHC experiments started to consider it

The resolution of muons traversing 1.5km of ice (=3850 X_0) in IceCUBE has been determined with three different methods in[23].

Although of course the problem is very different, I have not resisted the temptation to overlay to the graph on the right the ballpark of the resolution we achieve with a 2m-long lead tungstate calorimeter (=225 X_0) + CNN reconstruction

→ **3x better**



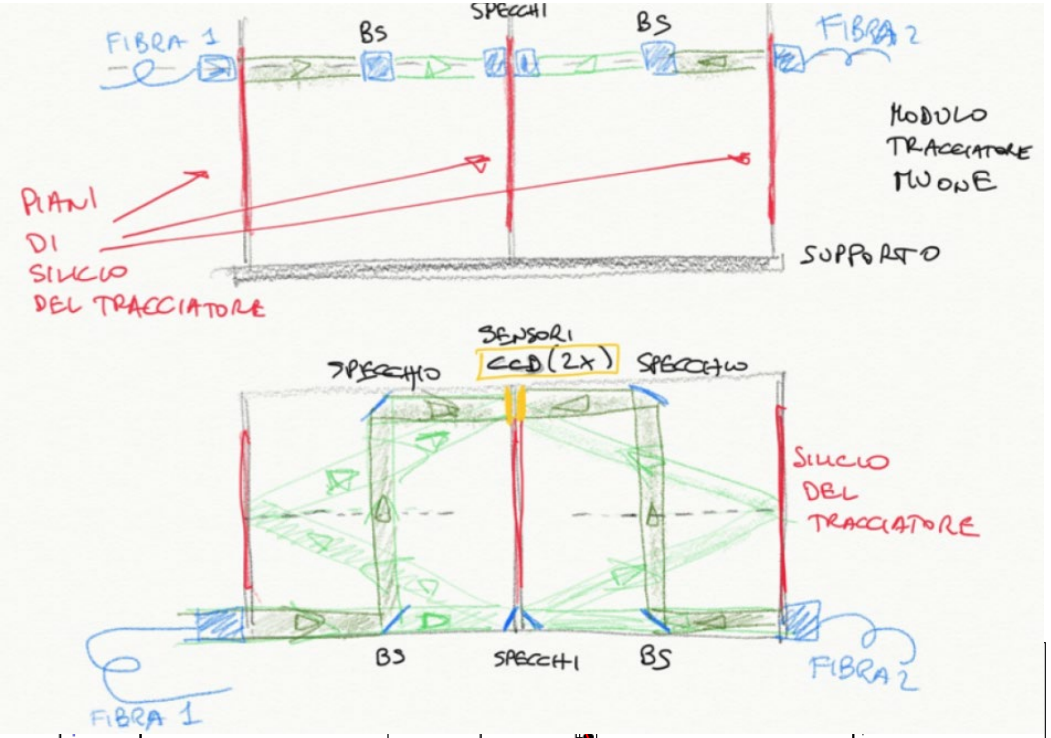
Speaking of systematic uncertainties,

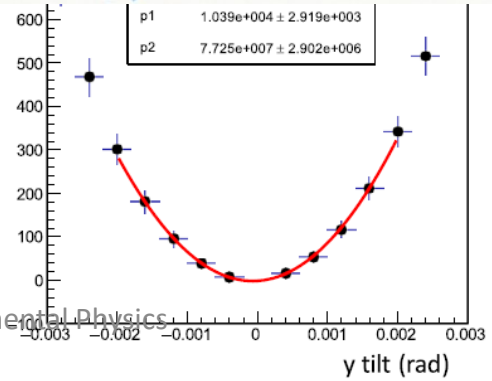
MUonE correctly identified the need for locating the scattering vertex to within 10μm along the beam axis (it has a strong impact on the q^2 resolution), and proceeded to design a very fancy holographic laser system, to be mounted on each station (=40 systems) to monitor the sensors locations

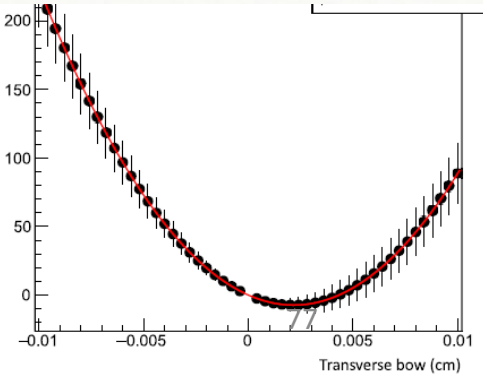
Cost: **several hundred kEuros**

As a by-product of the modeling of detector + information extraction process, the optimization study showed that with 5' of muon beam data, the location, tilt and bow of all detector and target elements **can be determined with O(1μ) accuracy by a global fit to the vertex**!

This is an example of the dividends that the study of a full model of (physics)+(detector)+(reco method)+(inference extraction) can provide







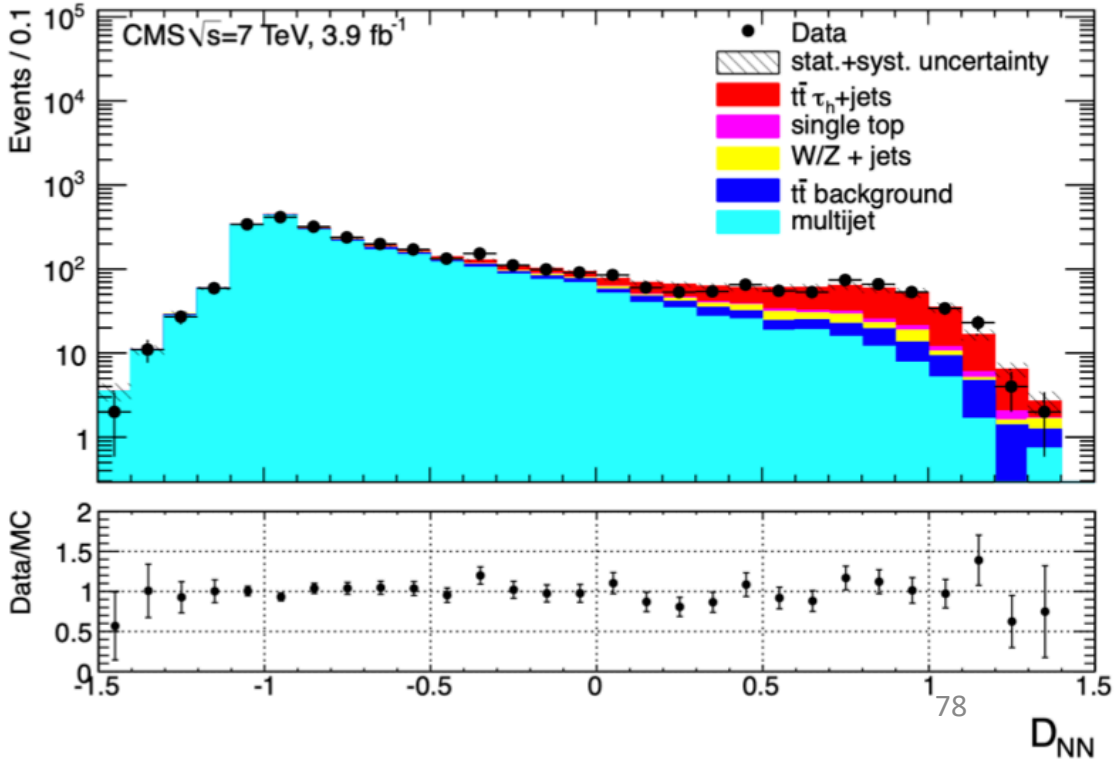
Replication of tt → τ_h + jets (TOP-11-004) with CMS Open Data

Goal: test and benchmark INFERNO in a realistic CMS analysis

- Replicate tt cross section measurement in tau + jets channel: 2011 CMS analysis that uses a Neural Network and is dominated by systematic uncertainties

σ(tt+X)=156 ± 12 (stat.) ± 33 (sys.) ± 3 (lumi) pb

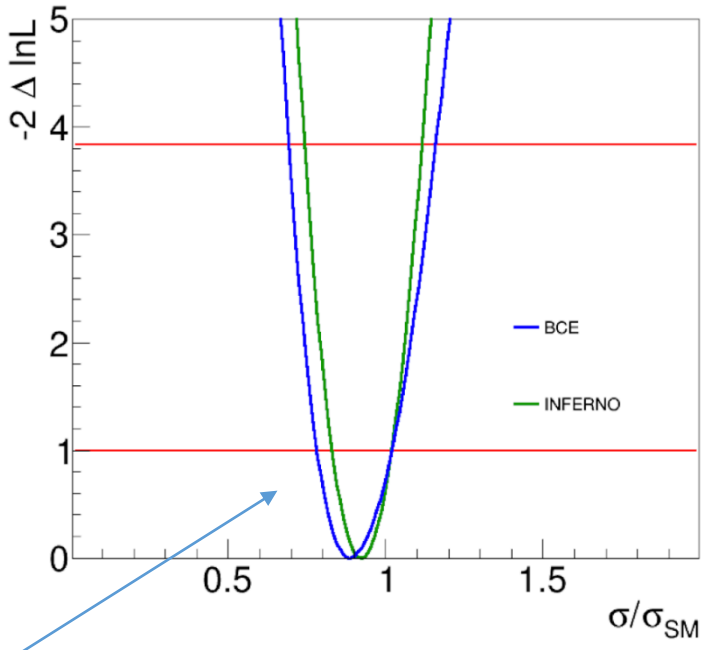
Source	Rel. uncert. [%]
W/Z/tt backgr. cross section uncert.	±3
Top-quark mass	±3
Renormalization/factorization scale	±2
Parton matching	±3
PDF	±5
τ_h trigger efficiency	±7
Pileup	$^{+5}_{-1}$
τ_h energy scale	±7
τ_h identification	±9
Jet energy scale	±11
Jet energy resolution	±2
Unclustered E_T^{miss}	±7
b-tagging	±3
Multijet background reweighting	±5
Syst. uncert.	±21
Stat. uncert. from fit and MC samples	±8
Stat. uncert. from trigger	±0.4
Total stat. uncert.	±8

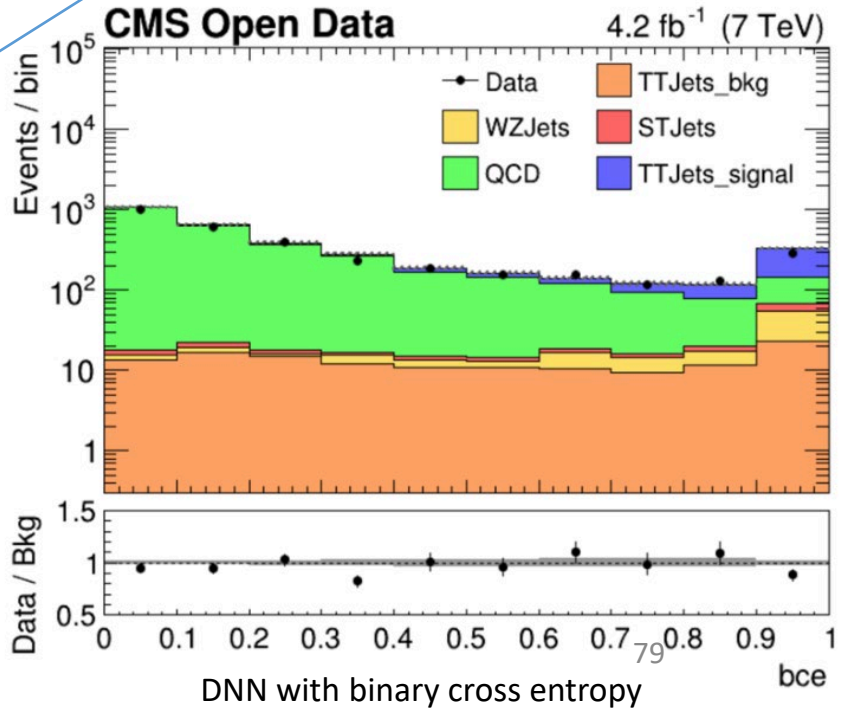


Measurement of tt → τ_h + jets with INFERNO

Results of tt → τ_h + jets reproduced with CMS Open Data

- First successful trainings with INFERNO
- Indication that INFERNO outperforms BCE also with realistic data





DNN with binary cross entropy

Realigning our design choices to future AI

A point which cannot be stressed enough is that if we design today something that will operate 10 or 20 years in the future, we need to account for the pattern recognition capabilities of future automated systems

In 20 years, will we use a Kalman filter to reconstruct trajectories in our trackers, or photon energy and direction in our calorimeters?

No, we won't. We will employ AI technology, streamlined by a decade of consolidation in similar tasks.

Shouldn't we then build those devices by considering how AI technology could best exploit them? If we do not, we will suffer a misalignment of our design choices and the future capabilities of the software we will end up using.

How to get around this problem?

We can and should try to model increasingly performant pattern recognition in our optimization loops, and verify whether there are discontinuities in the solutions space.

It is not going to be easy, but it is IMHO absolutely necessary to start getting equipped.

Speaking of calorimeters…

High granularity has become a compelling requirement.

- The CMS HGCAL detector [6] is a step in that direction; its design will improve by a large margin usable information about the showers, development, pointing, and composition.
- For different reasons, similar developments and improvements are planned for other projects (*e.g.,* CALICE [7] or CaloCUBE [8]).

However, an end-to-end optimization of the design of such instruments has not been attempted yet; **nor have models of the future potential of machine learning in pattern recognition been considered so far in the design phase**

As a telling example, the HGCAL detection elements are arranged in a hexagonal symmetry which offers construction benefits but significantly complicates the most common imaging techniques for shower reconstruction. While solutions to this issue do exist (*e.g.*, see [9]), this is an example of misalignment between design and potential exploitation.

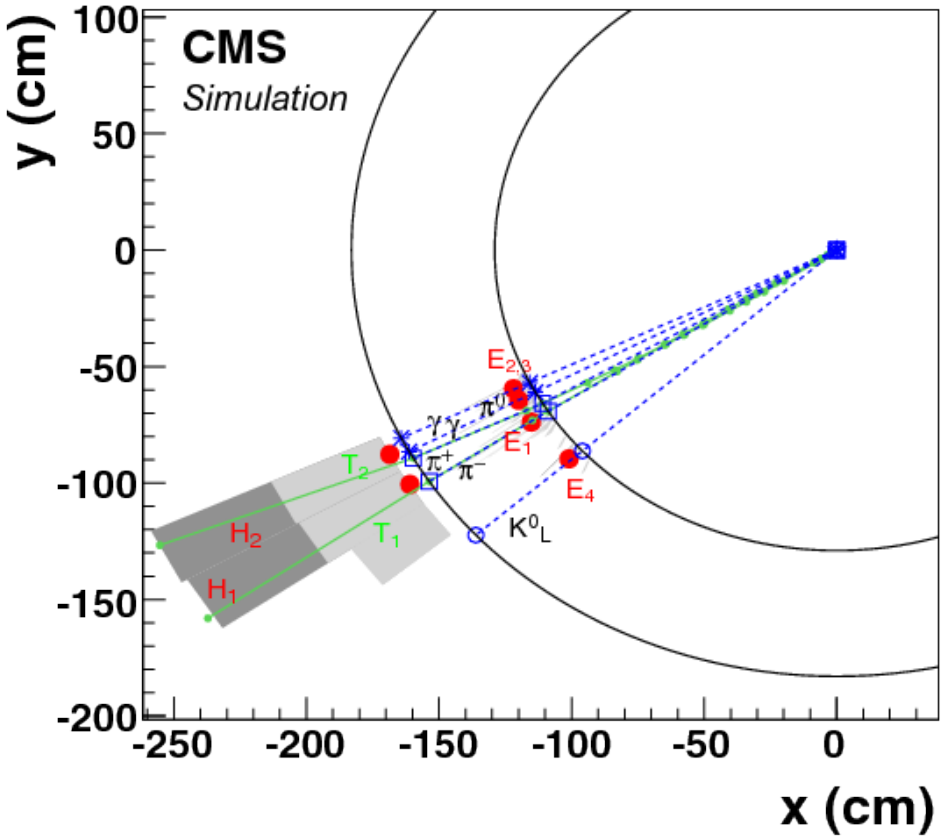
Serendipitous alignment: the case of HCAL

Take the LHC experiments for a telling example. CMS was originally endowed with a less performant hadron calorimeter than ATLAS. But hadron calorimetry ended up being crucial for a number of new physics searches involving boosted jets.

CMS regained the lost ground through the high performance of its "particle flow" reconstruction algorithm [10].

This was only possible thanks to the high magnetic field integral of CMS, which spreads out charged particles of different momenta within jets, easing their matching to calorimeter deposits.

This *post-hoc* exploitation of the solenoid characteristics, whose original specifications were rather driven by compactness and transverse momentum resolution, is a **striking example of how the combined search of hardware and software solutions may be proficuous** to inform the optimization of a modern particle detector.

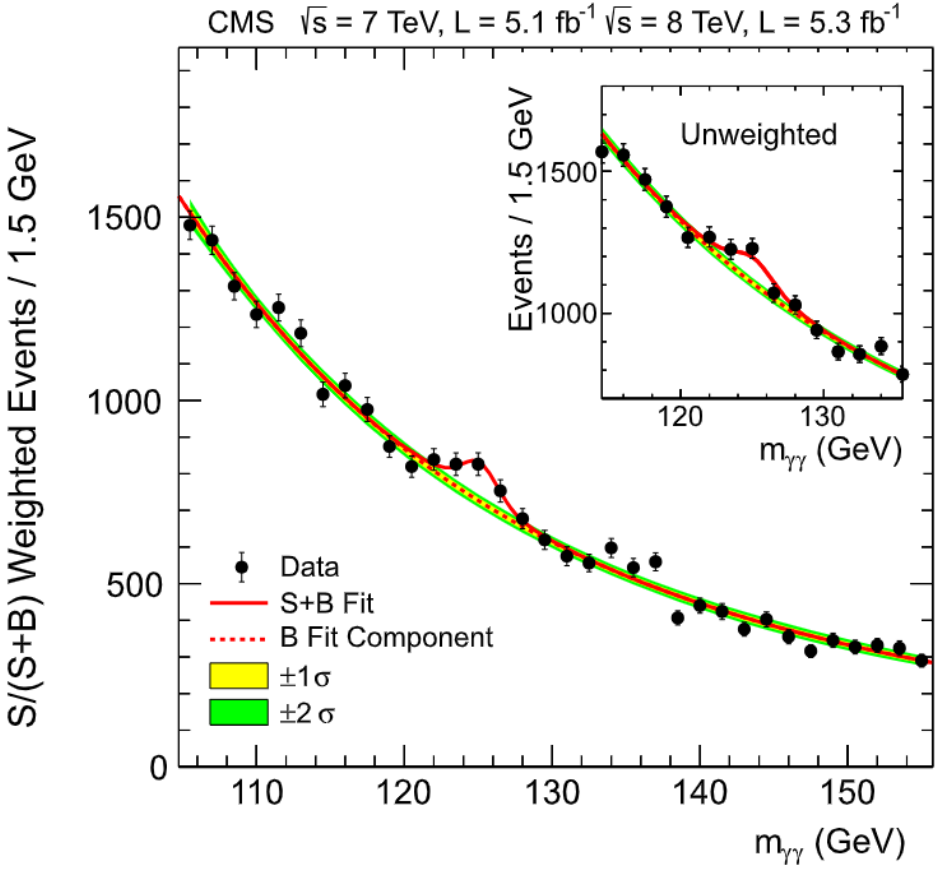


Applicazioni AI in HEP: la scoperta del bosone di Higgs

Per identificare un segnale della produzione e decadimento del bosone di Higgs in due fotoni [CMS 2012], CMS ha fatto largo uso di **Boosted Decision Trees** per

- calibrare l'energia di fotoni
- stimare l'incertezza su E
- selezionare il vertice di produzione di H
- identificare i fotoni del decadimento H→γγ
- separare il segnale dai fondi

Si è trattato di un vero e proprio paradigm shift, in quanto fino ad allora l'uso di metodi MVA era stato molto limitato in HEP



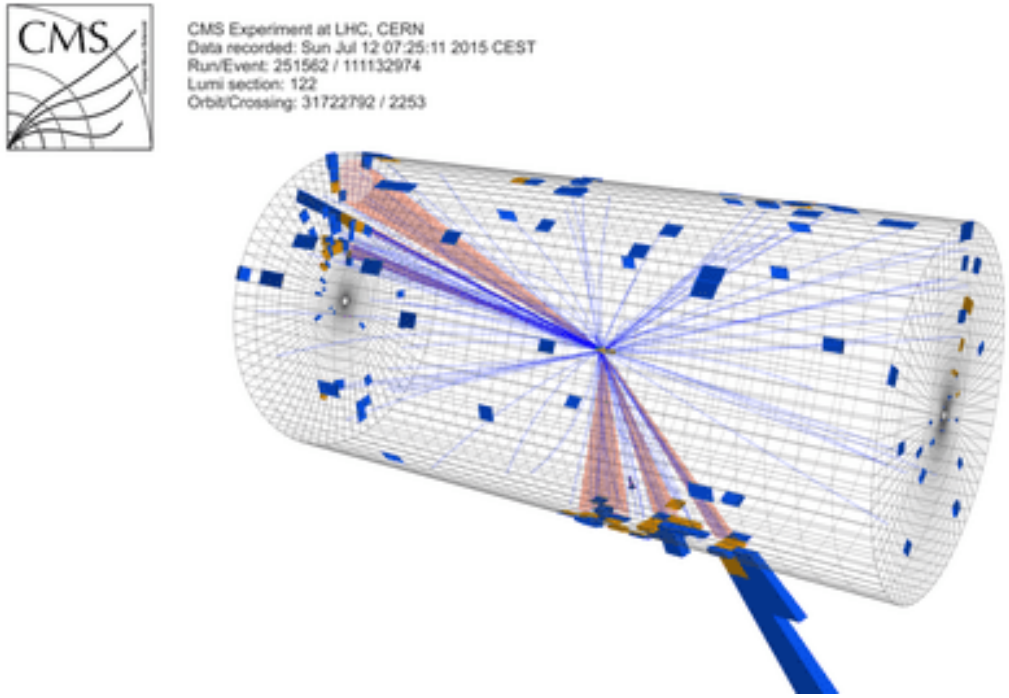
Sopra: spettro di massa ricostruita di coppie di fotoni di CMS, con un fit al background (verde) e segnale (in rosso)

Background: boosted decays and fat jets

Energetic LHC collisions may produce heavy objects with large momentum (top quarks, or W, Z, H bosons). When these decays, they usually yield a collimated stream of particles – a single hadron jet.

A number of techniques allow the extraction of features sensitive to the heavy object decay

The point is however that high granularity and effective identification of constituents in dense environments has become unavoidable



Above: a top-pair decay produces two fat jets, where the individual subjects are visible

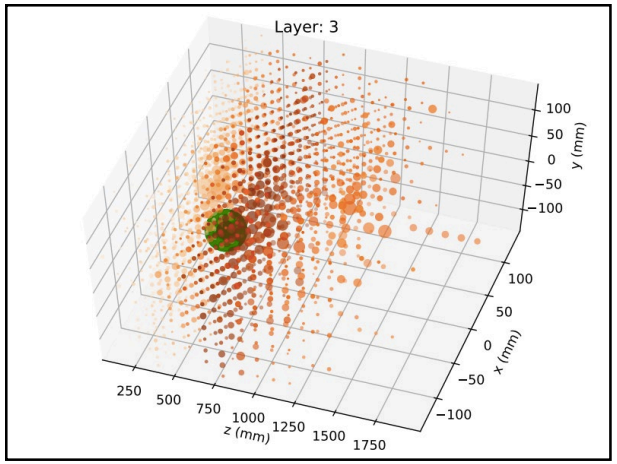
Condensate and predict

$$\breve{V}_k(x) = ||x - x_\alpha||^2 q_{\alpha k}, \text{ and}$$

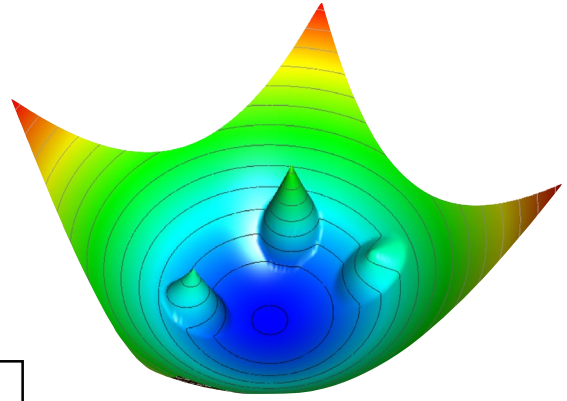
$$\hat{V}_k(x) = \max(0, 1 - ||x - x_\alpha||) q_{\alpha k}.$$

Maximum charge vertex for object k

- Maximum β/charge vertices *are* center points *
- Encourage network to select one representative point per object k



$$L_\beta = \frac{1}{K} \sum_k (1 - \beta_{\alpha k}) + s_B \frac{1}{N_B} \sum_i^N n_i \beta_i,$$



$$L_p = \frac{1}{\sum_{i=0}^N (1 - n_i) \operatorname{arctanh}^2 \beta_i} \sum_{i=0}^N L(t_i, p_i)(1 - n_i) \operatorname{arctanh}^2 \beta_i$$

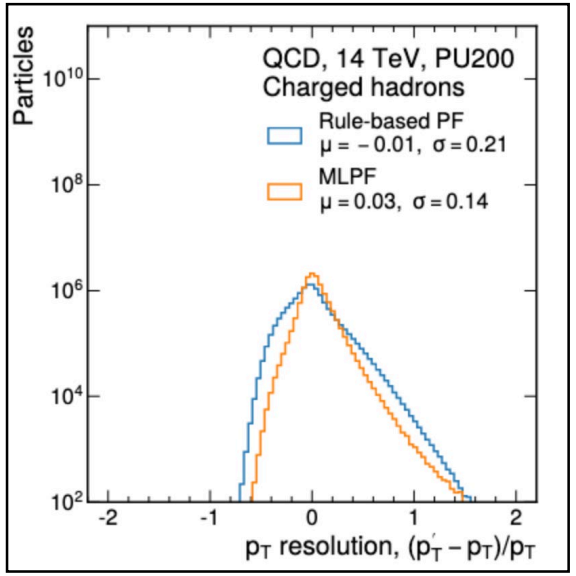
- *Very natural approach for dynamic graph NN*

*NB: Removes saddle point for large N

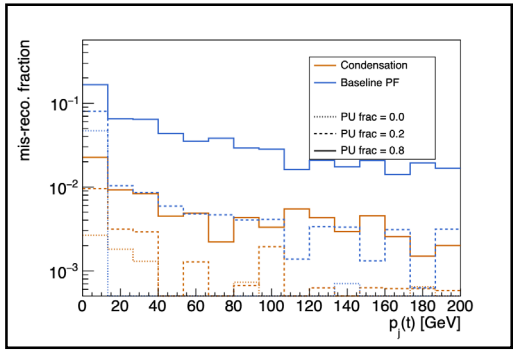
JK, arxiv:2002.03605, EPJC

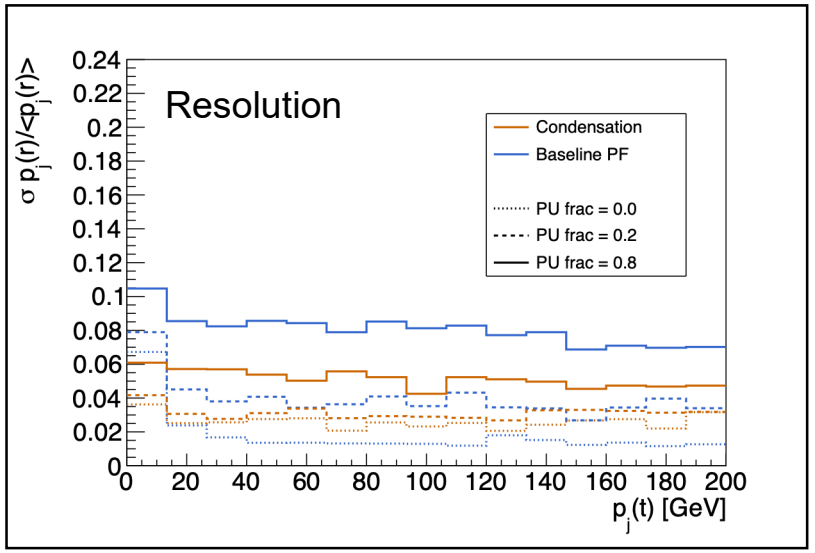
Comparison on other variables

- Consistent observations also for hadrons using Delphes and comparable PF DNN approach



J. Pata et al, arxiv:2101.08578, EPJC





- Cumulative quantities: jets
  - Standard PF does very well for 0 PU fraction (built-in energy conservation)
  - With higher PU fraction identification of **individual particles** way more important: **object condensation starts to be better**, in particular at low momenta

Jan Kieseler

Jan Kieseler, arxiv:2002.03605, EPJC

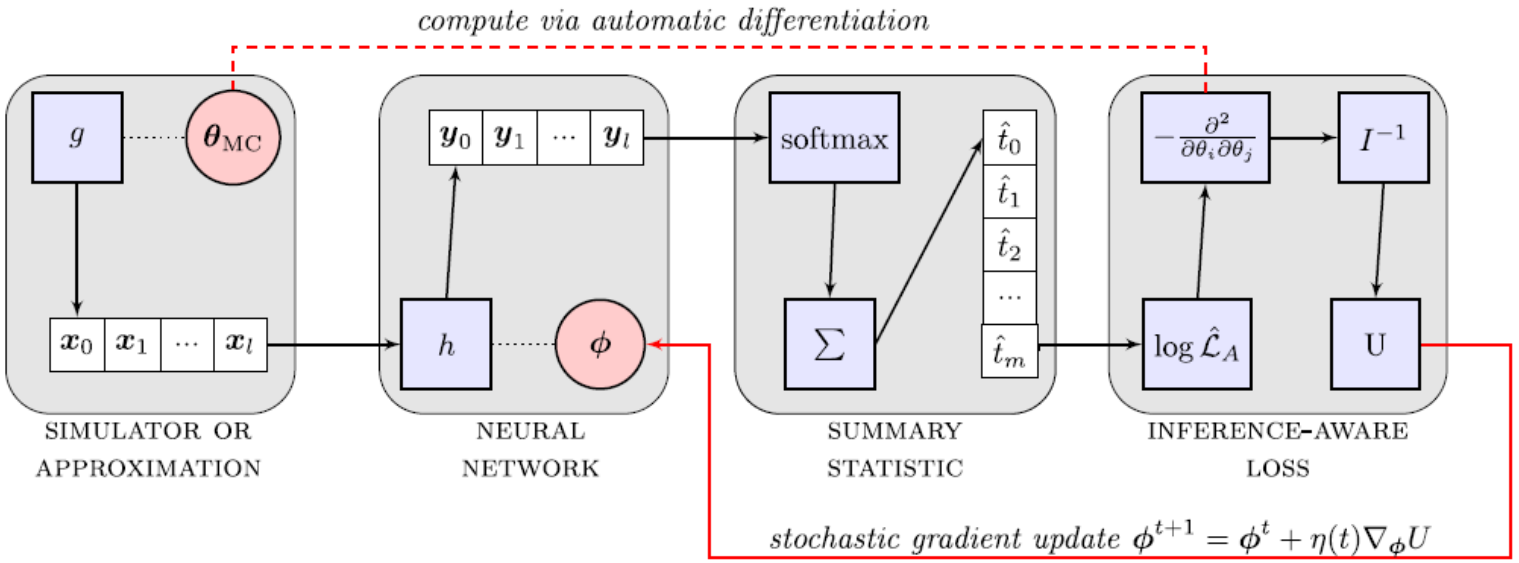
The frontier: «inference-aware» analysis methods

In HEP, the signal/noise discrimination is an intermediate step in the analysis chain: using signal-enriched data we measure relevant physical parameters, striving for the highest precision

Imprecise models used in training the classifier affect the measurement with nuisance parameters. The related systematic uncertainties are only dealt with after the training is completed → there is a misalignment between the classifier objective and the experimental goal

In a recent development [Dorigo 2019] employing differentiable programming techniques we constructed a feedback loop to inform the DNN of the true final outcome of the measurement process, such that the loss function may account for it in the training phase.

→large gains in the ultimate (stat+syst) precision of resulting inference!

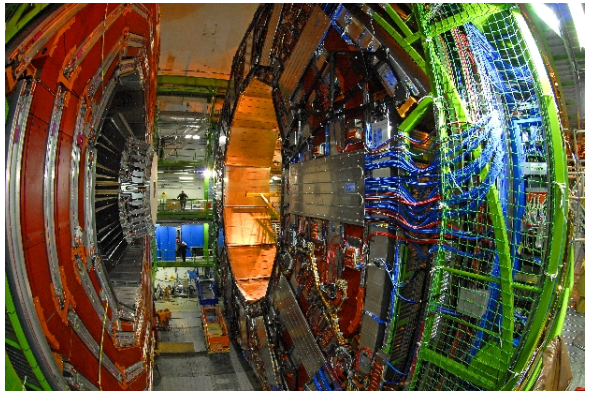


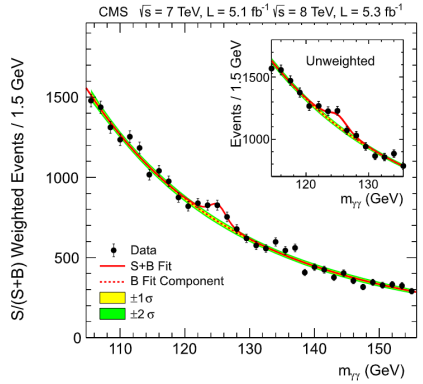
Trigger primitives

LHC produces 40 million sets of O(100) collisions every second in the detectors, which record the resulting information in O(10^7) digital electronic channels → an online selection is required to pick and store the most interesting collisions: the **trigger**

The huge reconstruction task of making sense of 150 Tb/s, to be carried out in a few microseconds per event, is today dealt with by neural networks encoded in FPGA [Xilink 2020]

This is a highly specialized application, as it requires extremely low latency and very high data fluxes





Serendipitous alignment: the case of HCAL

Take the LHC experiments for a telling example. CMS was originally endowed with a less performant hadron calorimeter than ATLAS. But hadron calorimetry ended up being crucial for a number of new physics searches involving boosted jets.

CMS regained the lost ground through the high performance of its "particle flow" reconstruction algorithm [10].

This was only possible thanks to the high magnetic field integral of CMS, which spreads out charged particles of different momenta within jets, easing their matching to calorimeter deposits.

This *post-hoc* exploitation of the solenoid characteristics, whose original specifications were rather driven by compactness and transverse momentum resolution, is a **striking example of how the combined search of hardware and software solutions may be proficuous** to inform the optimization of a modern particle detector.