Advances in Machine Learning in High Energy Physics
Deep Learning, GAN and more

David Rousseau

LAL-Orsay

rousseau@lal.in2p3.fr

AGATA/GRETINA tracking arrays meeting

4th April 2018

ML in HEP

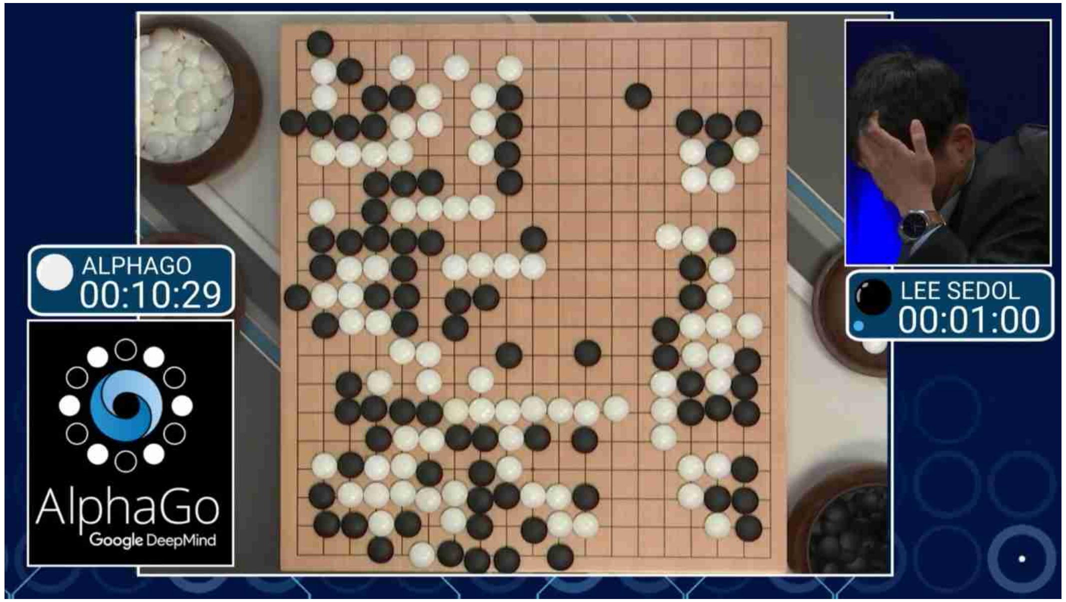
- Use of Machine Learning (a.k.a Multi Variate Analysis as we call it) already at LEP somewhat, much more at Tevatron (Trees)
- At LHC, Machine Learning used almost since first data taking (2010) for reconstruction and analysis
- In most cases, Boosted Decision Tree with Root-TMVA, on ~10 variables
- For example, impact on Higgs boson sensitivity at LHC:

Analysis	Data taking year	No ML sensitivity	ML sensitivity	Relative data gain
CMS $H\to\gamma\gamma$	2011-2012	2.2	2.7	51%
ATLAS $H\to\tau^{+}\tau$	2011-2012	2.5	3.4	85%
ATLAS $VH\to bb$	2011-2012	1.9	2.5	73%
ATLAS $VH\to bb$	2015-2016	2.8	3.0	15%
CMS $VH\to bb$	2011-2012	1.4	2.1	125%

➔~50% gain on LHC running

ML in HEP

- Meanwhile, in the outside world :

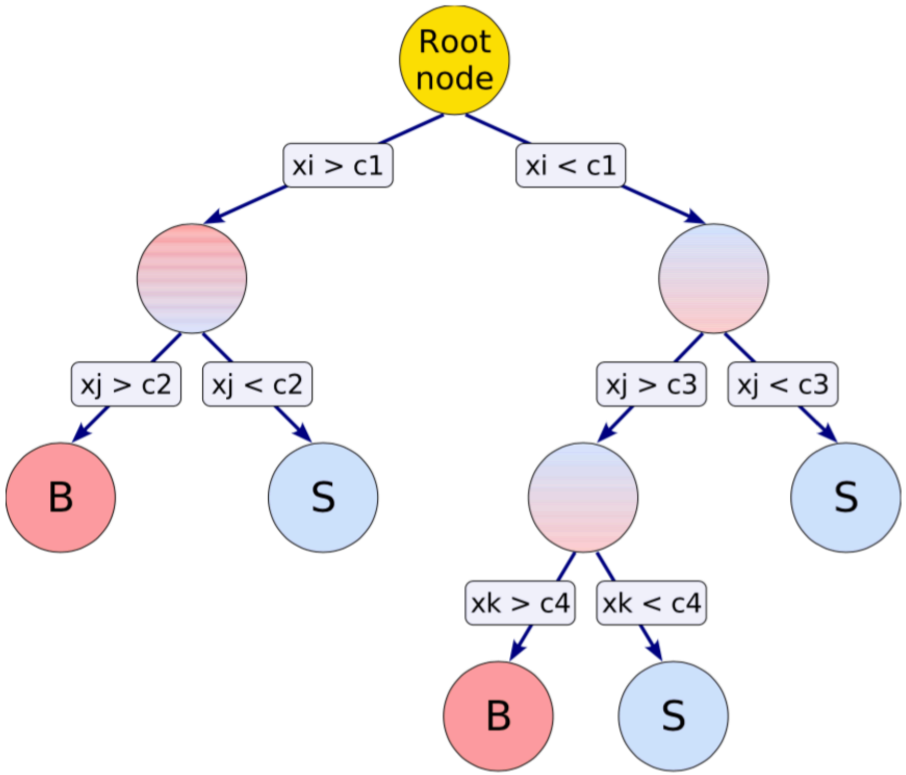


- “Artificial Intelligence” not a dirty word anymore!
- We (in HEP) have realised we’re been left behind! Trying to catch up now…

- This talk on very selected promising use of advanced ML in HEP

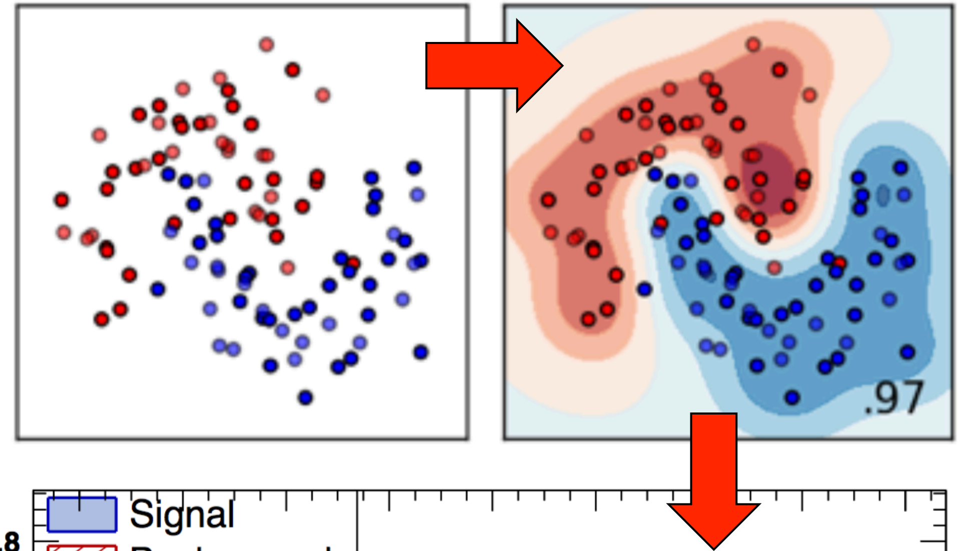
ML Basics

BDT in a nutshell



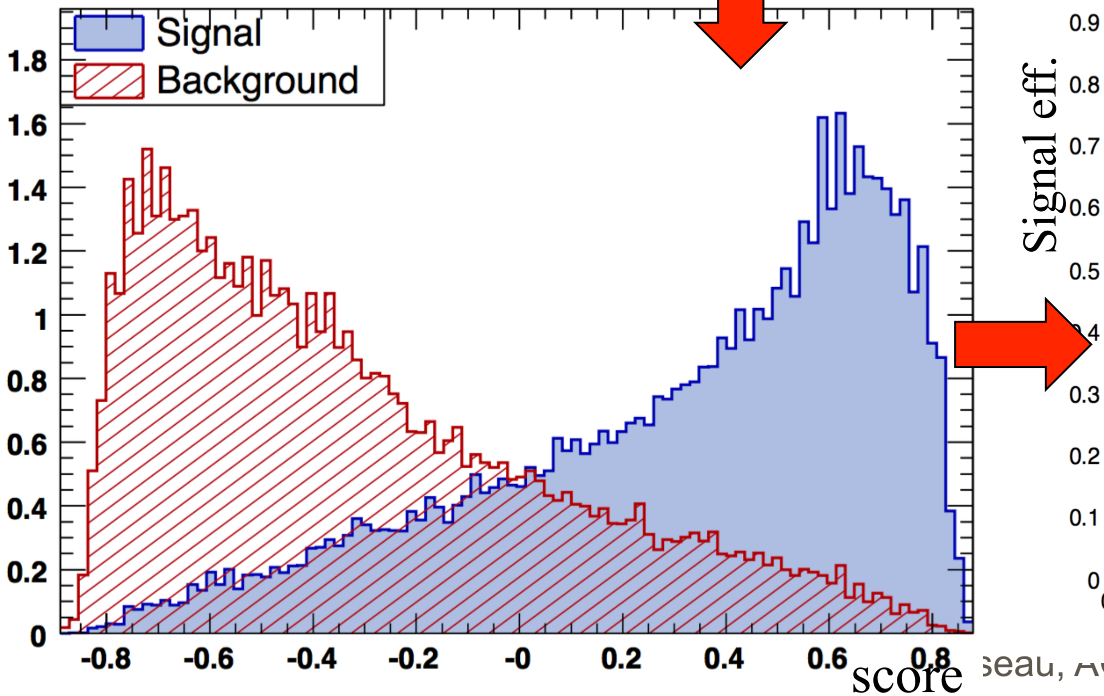
- Single tree (CART) <1980
- AdaBoost 1997 : rerun increasing the weight of misclassified entries →Boosted Decision Trees (Gradient BDT, random forest…)

Classifier basics

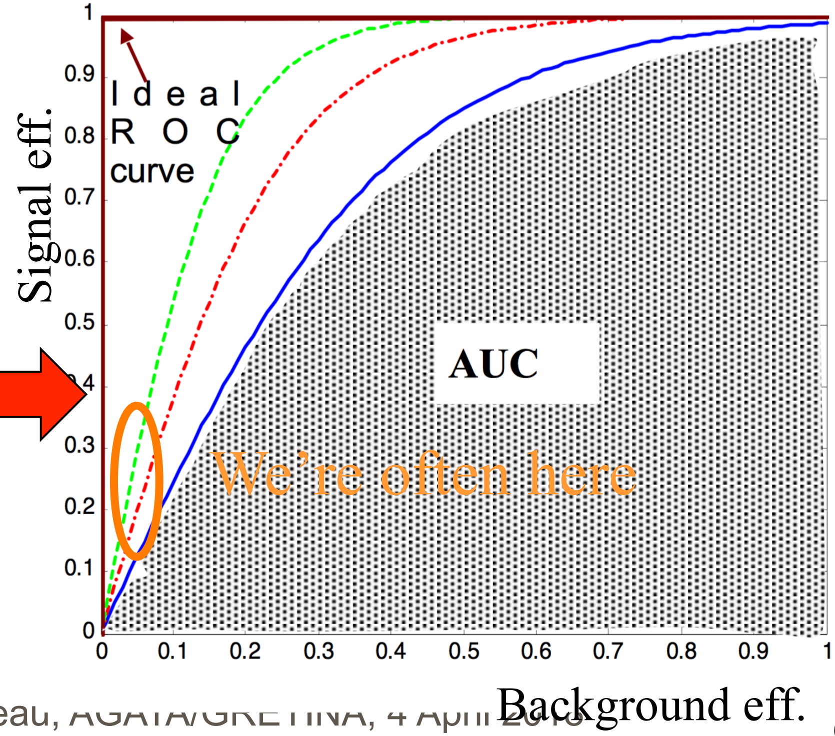


Train on Signal and Background Monte-Carlo
➔learn the separation between S and B distribution
Apply on test sample
Apply on data

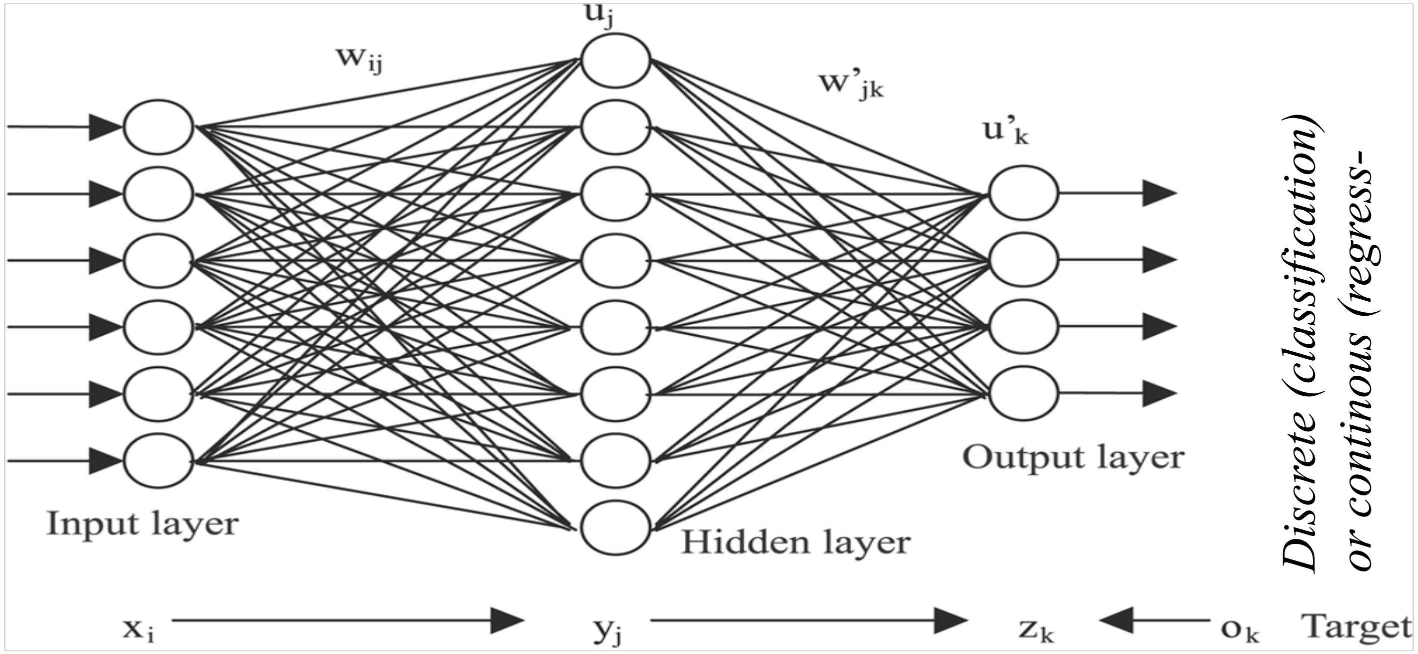
Note: instead of *classifying* 0 or 1, can *regress* !



AUC : Area Under the (ROC) Curve

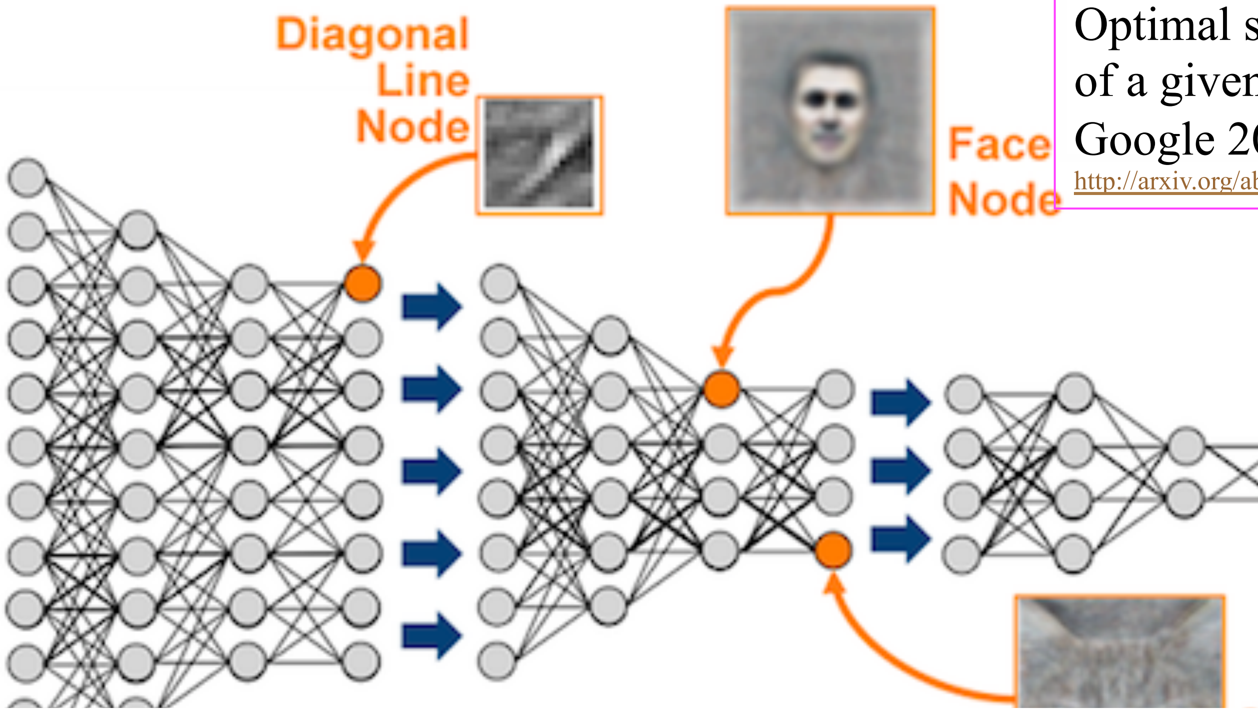


Neural Net in a nutshell

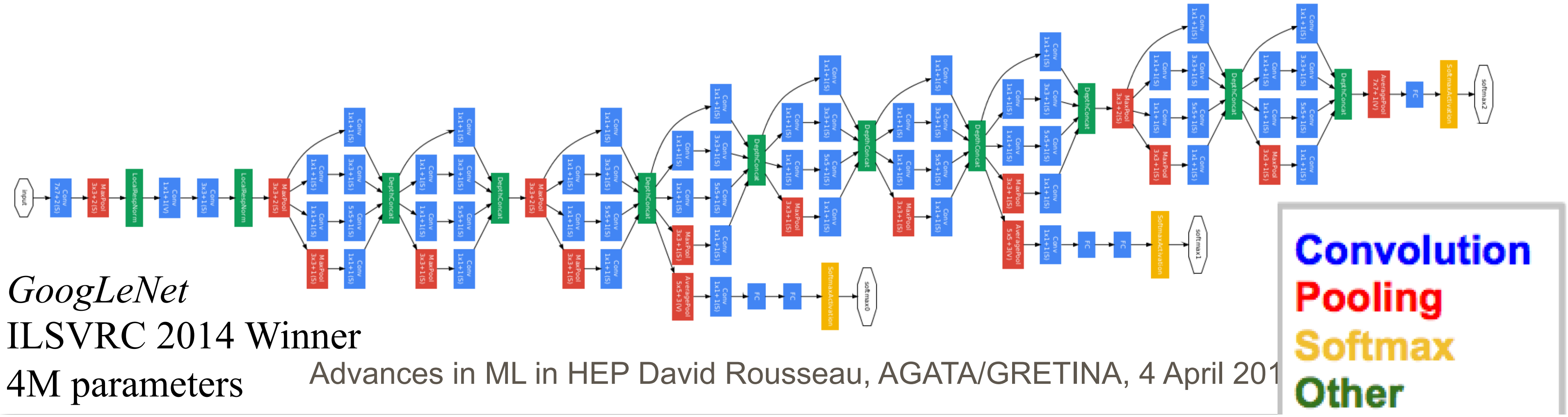


- Neural Net ~1950!
- But many many new tricks for learning, in particular if many layers (also ReLU instead of sigmoïd activation)
- "Deep Neural Net" up to 100 layers
- Computing power (DNN training can take days even on GPU)

Deep learning



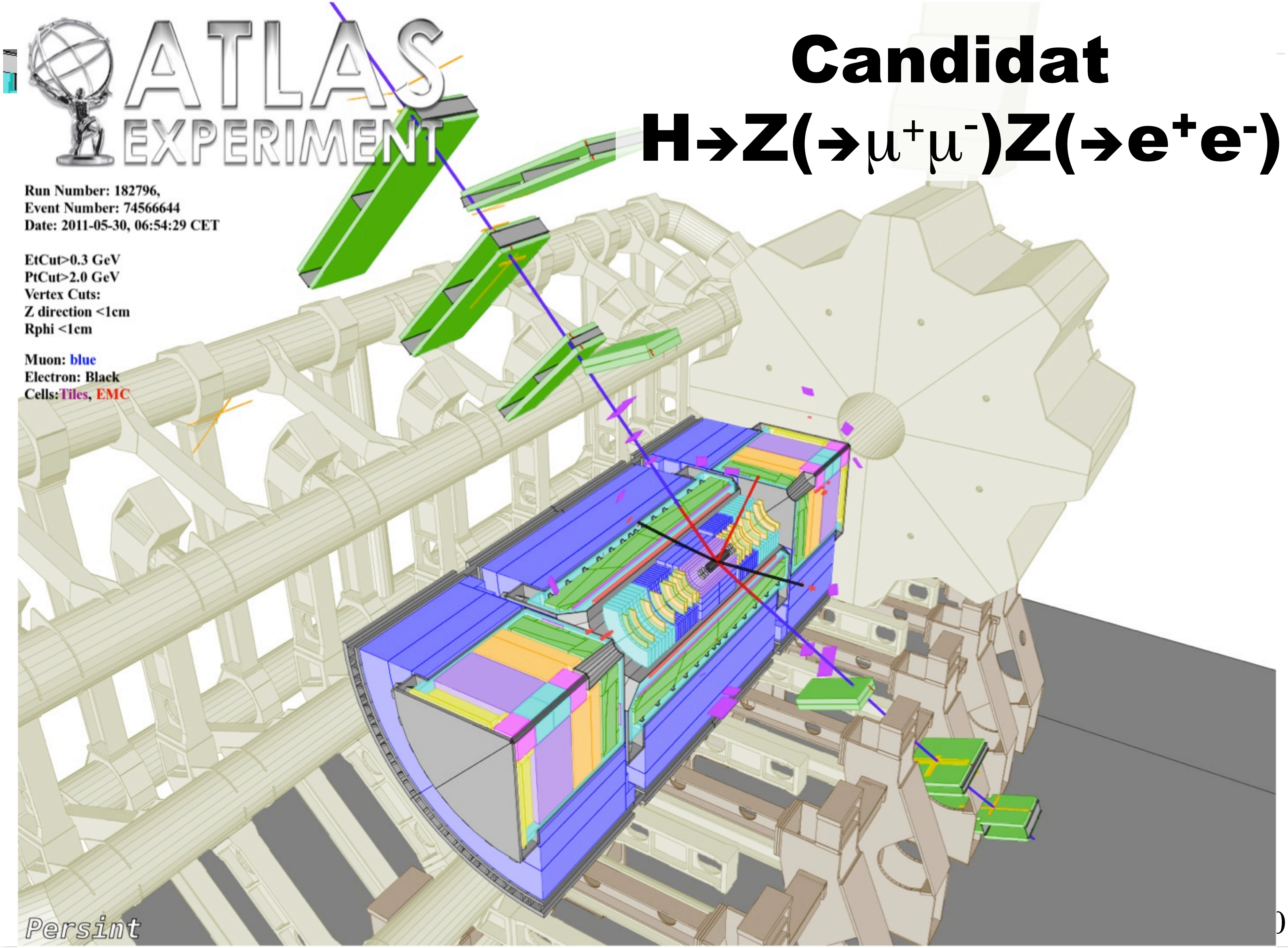
Optimal stimulus of a given neuron
Google 2012
http://arxiv.org/abs/1112.6209



GoogLeNet
ILSVRC 2014 Winner
4M parameters

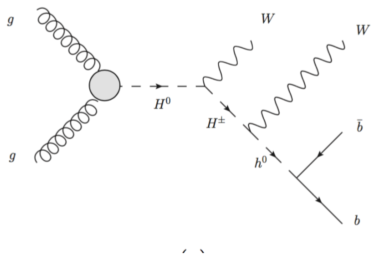
ML in analysis

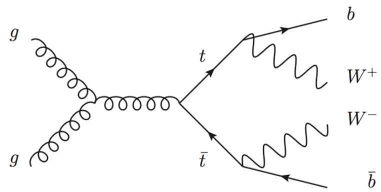
Candidat $H \rightarrow Z(\rightarrow \mu^+\mu^-)Z(\rightarrow e^+e^-)$

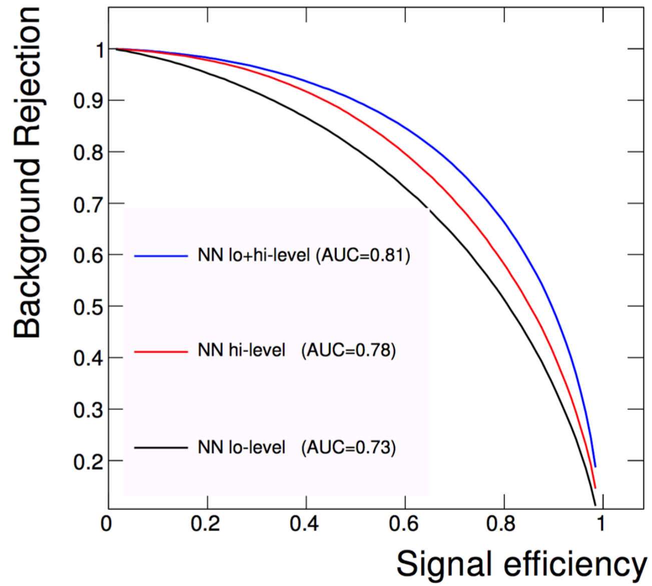


Deep learning for analysis

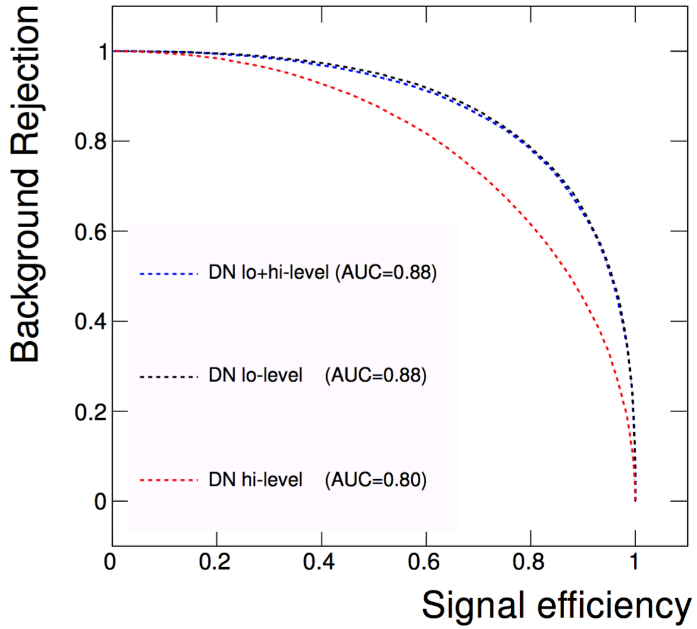
1402.4735 Baldi, Sadowski, Whiteson







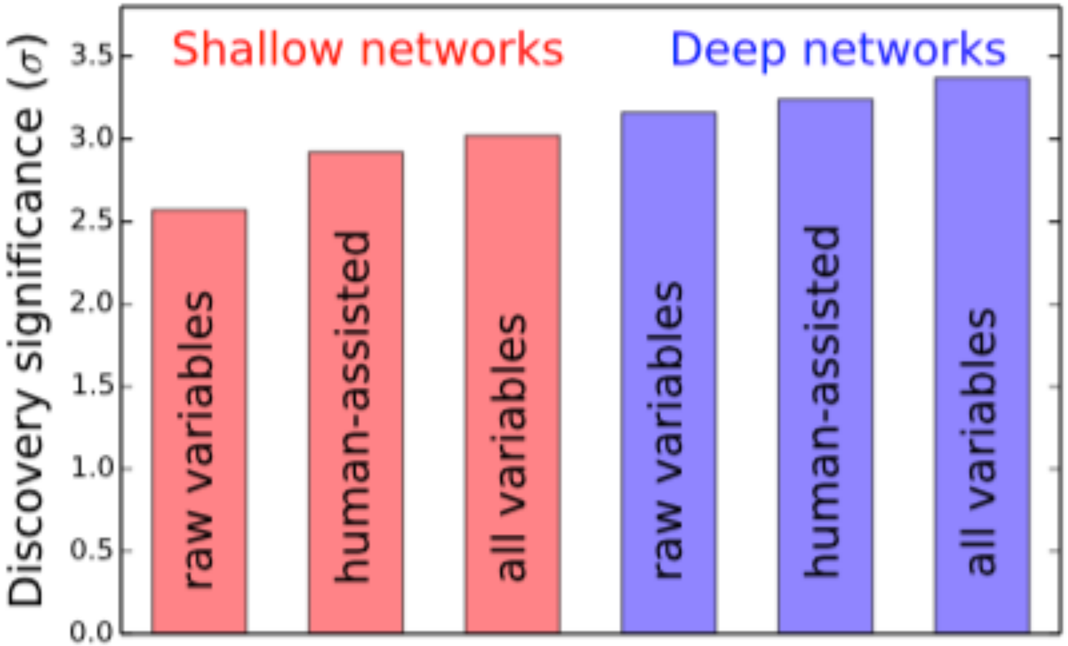
- MSSM at LHC : $H^0 \rightarrow WWbb$ vs $tt \rightarrow WWbb$
- Low level variables:
  - 4-momentum vector
- High level variables:
  - Pair-wise invariant masses
- Deep NN outperforms NN, and does not need high level variables
- DNN learns the physics ?



Deep learning for analysis (2)

1410.3469 Baldi Sadowski Whiteson

- H tautau analysis at LHC: H→tautau vs Z→tautau
  - Low level variables (4-momenta)
  - High level variables (transverse mass, delta R, centrality, jet variables, etc…)



- Here, the DNN improved on NN but still needed high level features
- Both analyses with Delphes fast simulation
- ~100M events used for training (>>100* full G4 simulation in ATLAS)

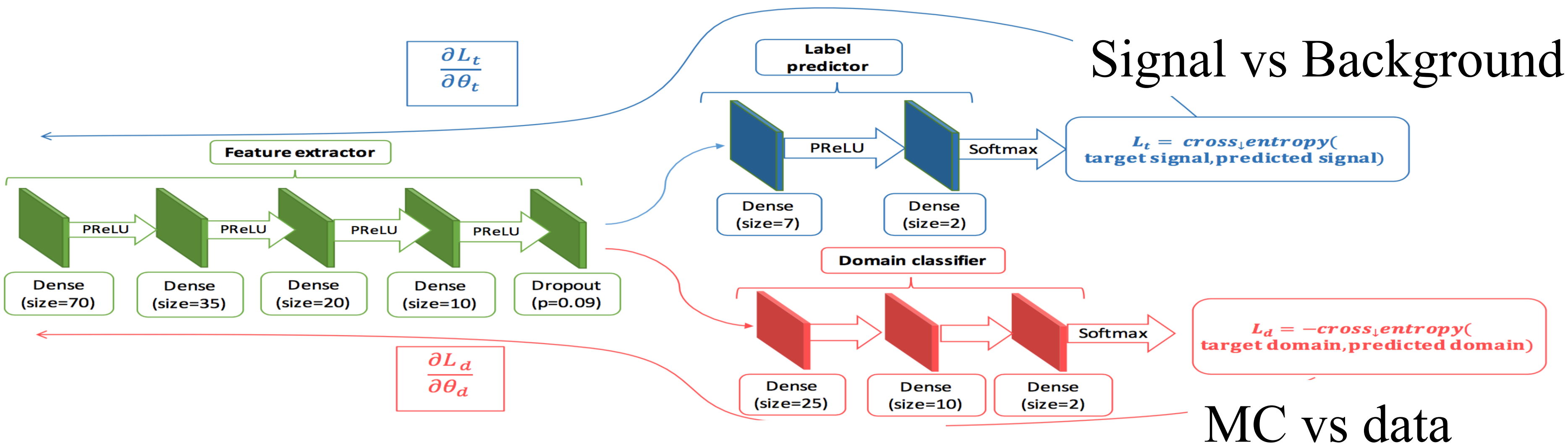
Systematics-aware training

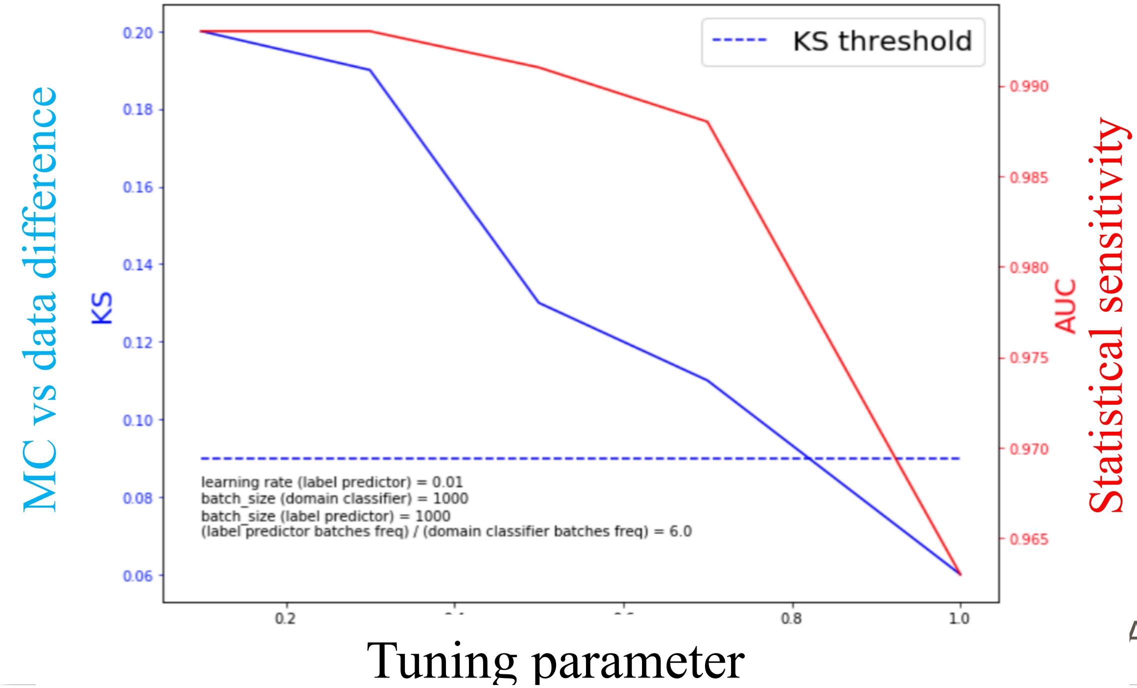
- Our experimental measurement papers typically ends with
  - measurement = m ± σ(stat) ± σ(syst)
  - σ(syst) systematic uncertainty : known unknowns, unknown unknowns…
- Name of the game is to minimize quadratic sum of :

  σ(stat) ±σ(syst)
- ML techniques used so far to minimise σ(stat)
- Impact of ML on σ(syst) or even better global optimisation of σ(stat) ± σ(syst) is an open problem
- Worrying about σ(syst) untypical of ML in industry
- However, a hot topic in ML in industry: *transfer learning*
- E.g. : train image labelling on a image dataset, apply on new images (different luminosity, focus, angle etc…)
- For HEP : we train with Signal and Background which are not the real one (MC, control regions, etc...)➔source of systematics

Syst Aware Training: adversarial

Inspired from 1505.07818 Ganin et al : ACAT 2017 Ryzhikov and Ustyuzhanin



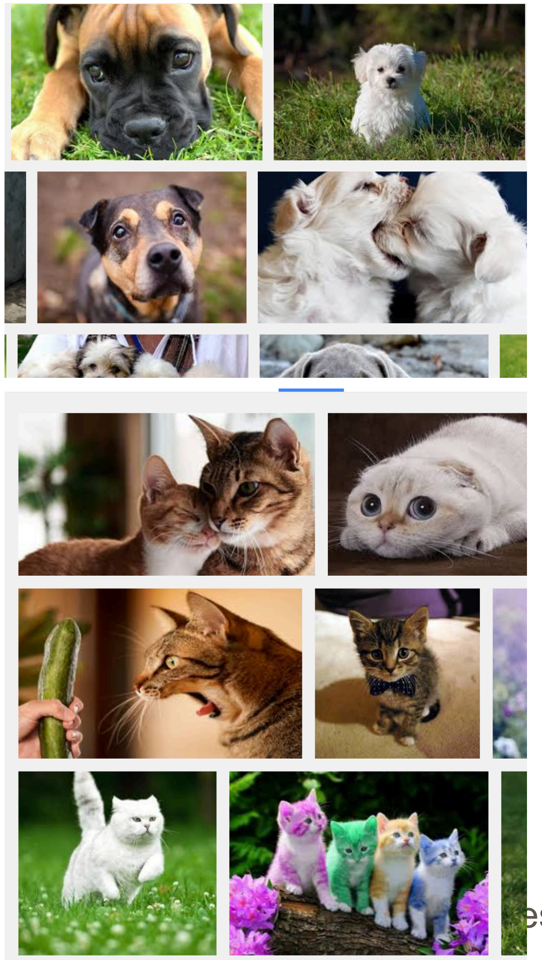


ML in reconstruction

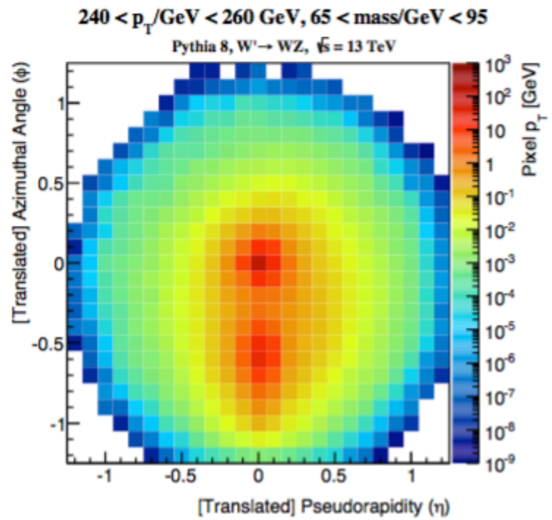
Jet Images

[arXiv 1511.05190](https://arxiv.org/abs/1511.05190) de Oliveira, Kagan, Mackey, Nachman, Schwartzman

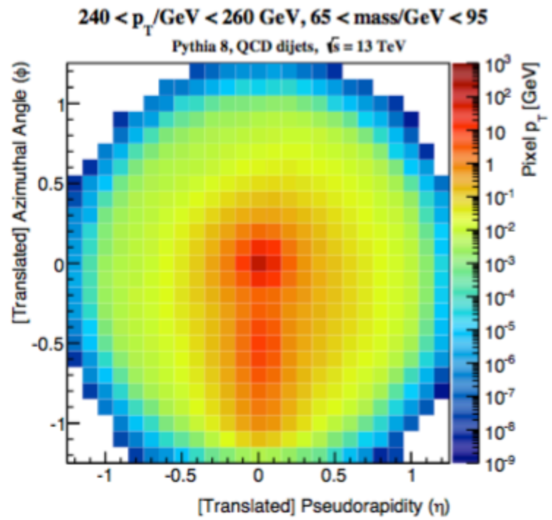
- Distinguish boosted W jets from QCD
- Particle level simulation
- Average images:



Boosted W→qq jet

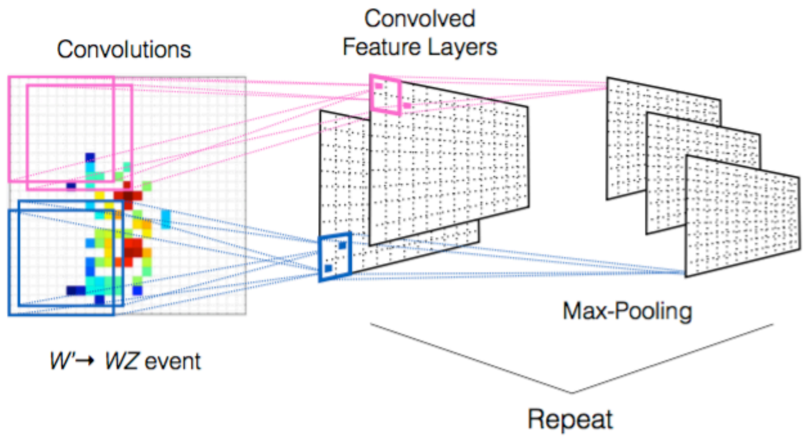


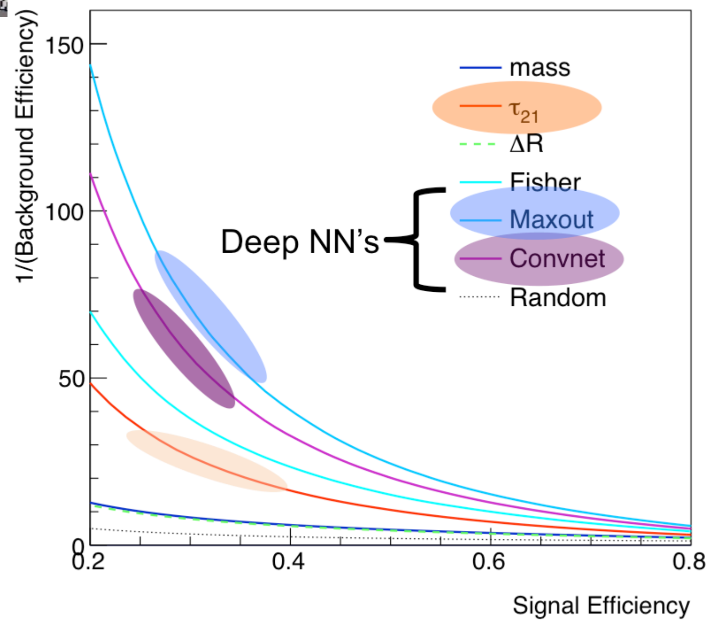
QCD



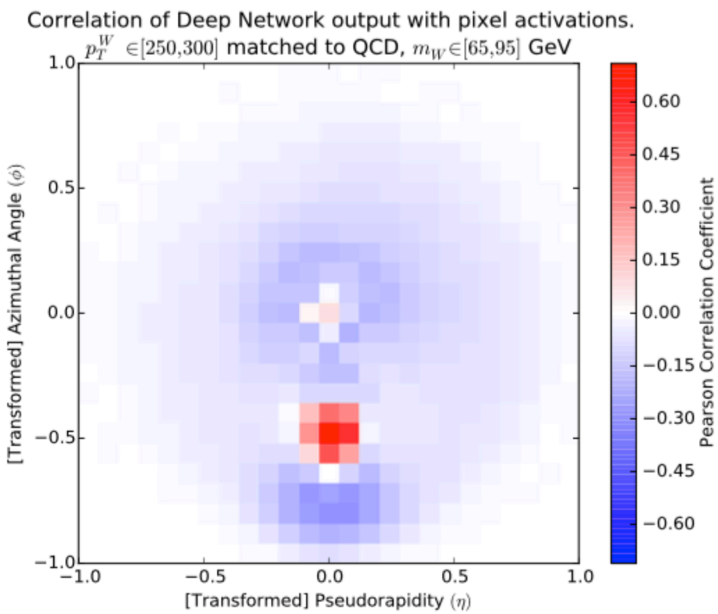
Jet Images : Convolution NN

arXiv:1511.05190

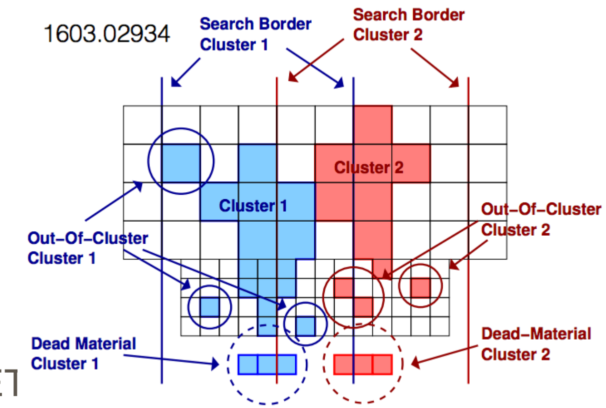




- Variables build from CNN outperform the more usual ones



- What the CNN sees (the "cat" neurone")
- Now need proper detector and pileup simulation
- ➔3Dimension



End to end Learning

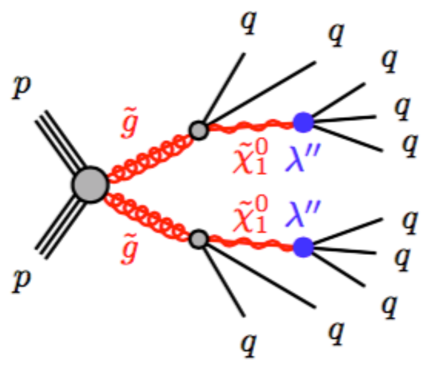
End to end learning

[Bhimji et al, 1711.03573]

- Train directly for signal on « raw » event ?
- Start from RPV Susy search

ATLAS-CONF-2016-057

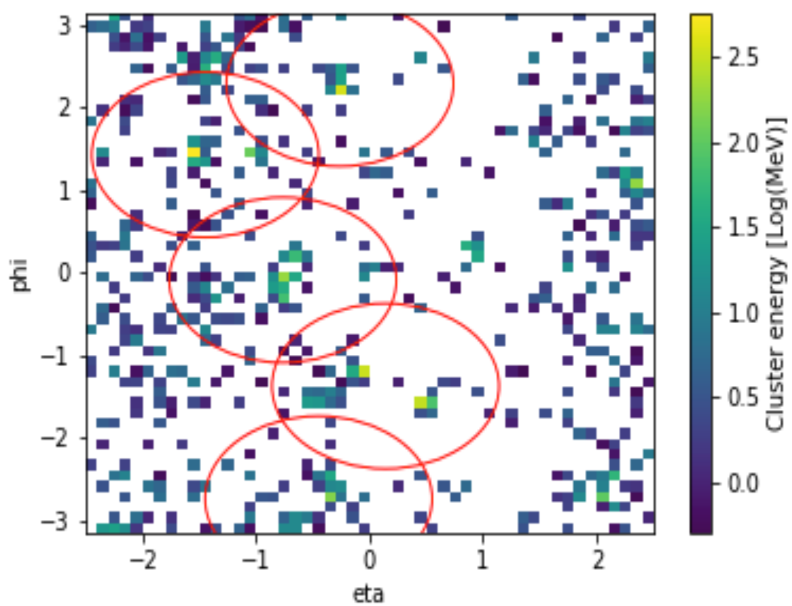
- Fast Simulated events with Delphes



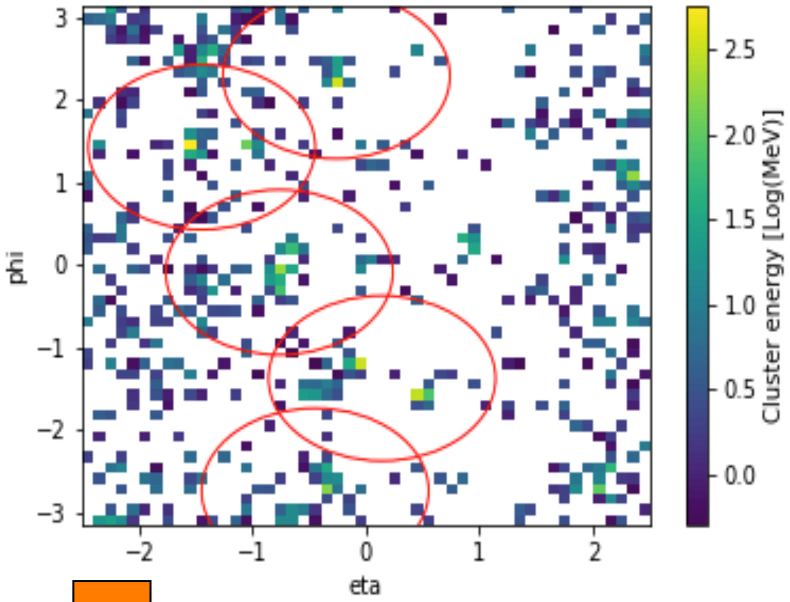
(b) gluino cascade decay

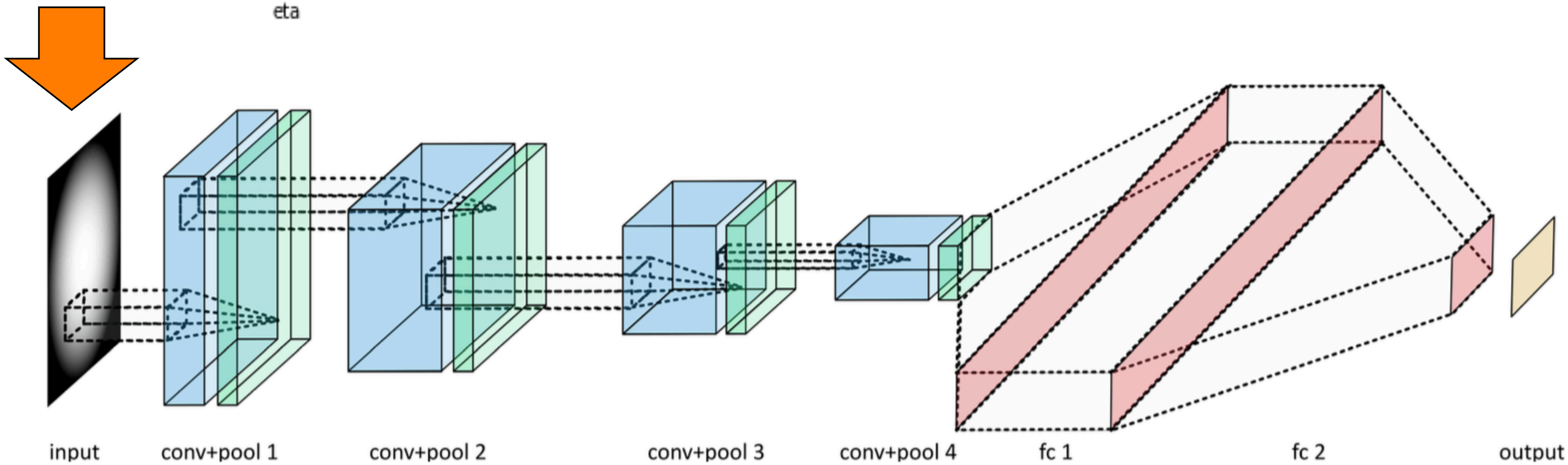
- Project energies on 64x64 ηxϕ grid
- Compare with usual jet Reconstruction and physics Analysis variables such as:

$$M_J^\Sigma = \sum_{\substack{p_T > 200\,\mathrm{GeV} \\ |\eta| \le 2.0}}^{4} m^{\mathrm{jet}}$$

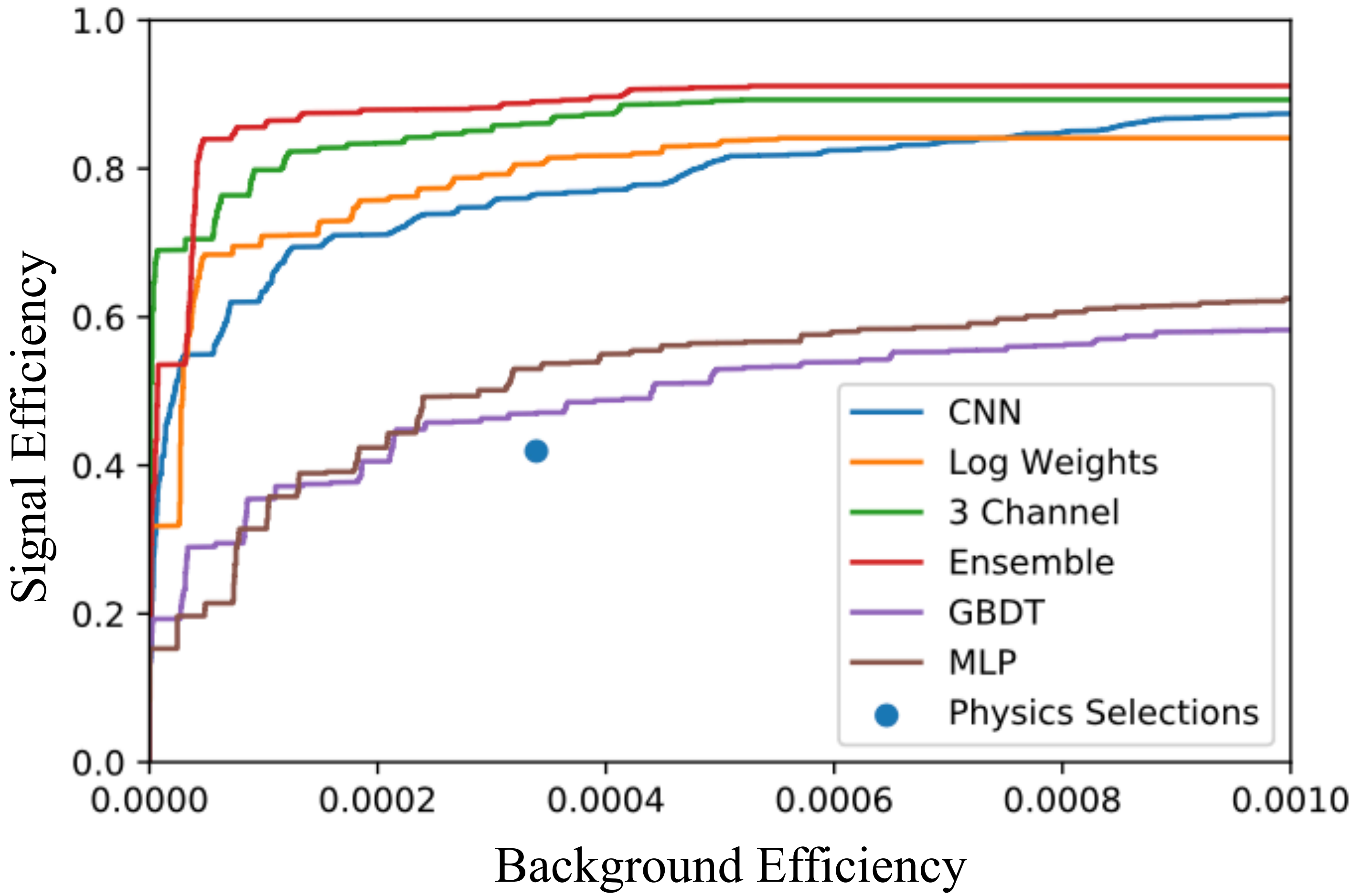


End to end learning (2)





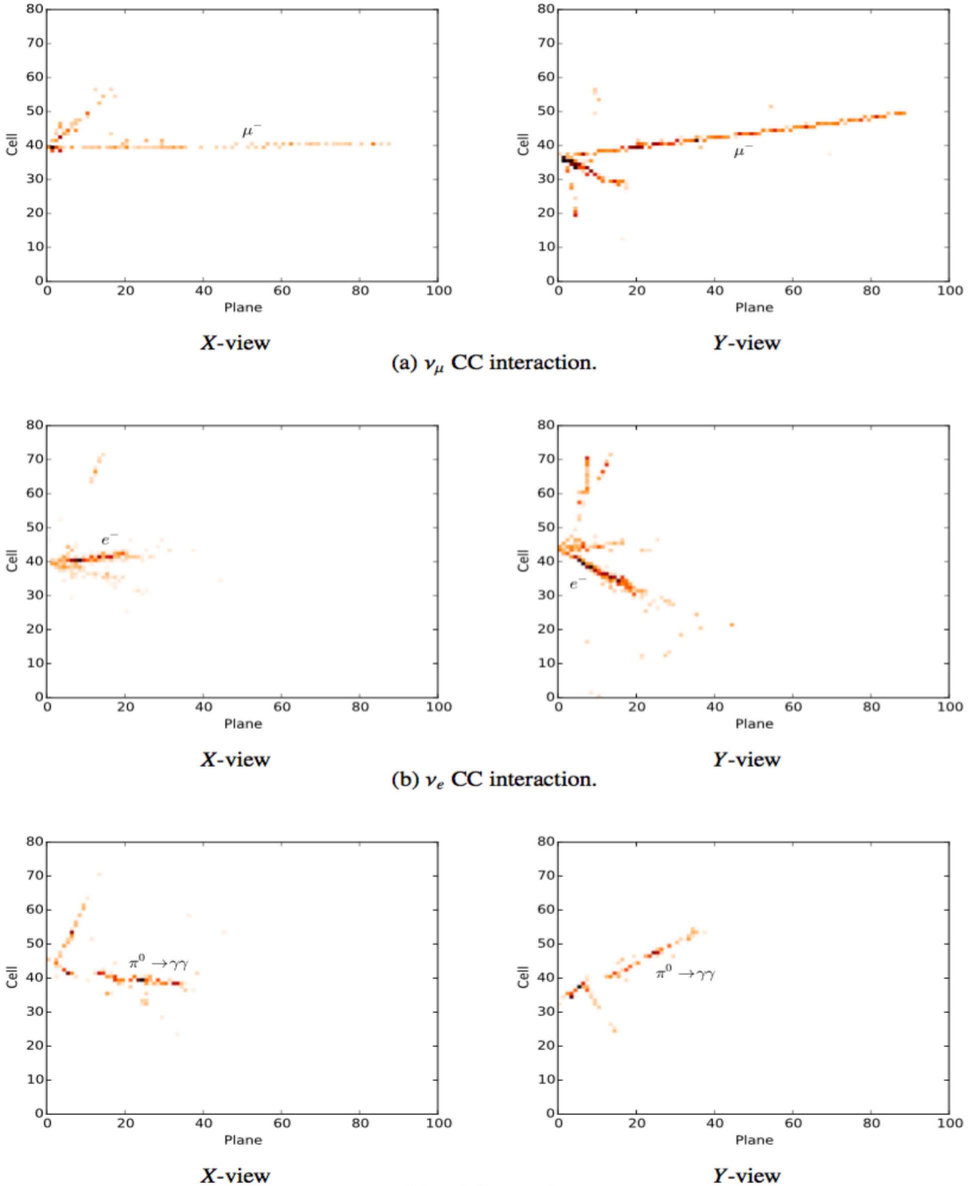
End to end learning (3)



- >x2 gain over BDT/shallow network using physics variable and 5 leading jet 4-momenta
- ➔CNN extract information from energy grid which is lost in the jets ?
- Not sure they should compare to applying DL on the jets

A recent attempt : NOVA

arXiv 1604.01444 Aurisano et al



(a) ν_μ CC interaction.

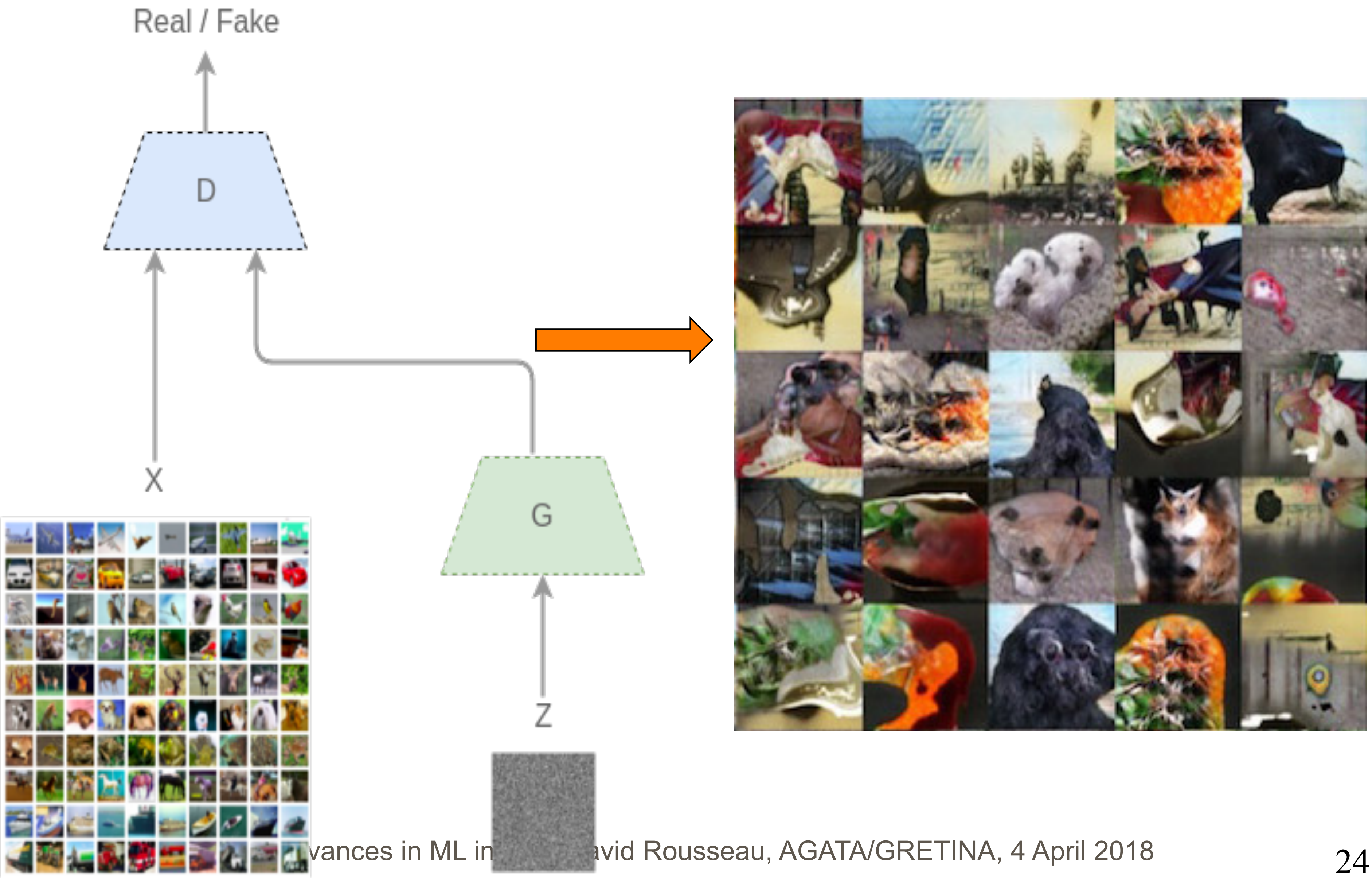
(b) ν_e CC interaction.

(c) NC interaction.

Neutrino interaction classification
Using Convolutionnal Neural Network (GoogleNet)
Actually used for analysis

ML in simulation

Generative Adversarial Network



Condition GAN

Text to image

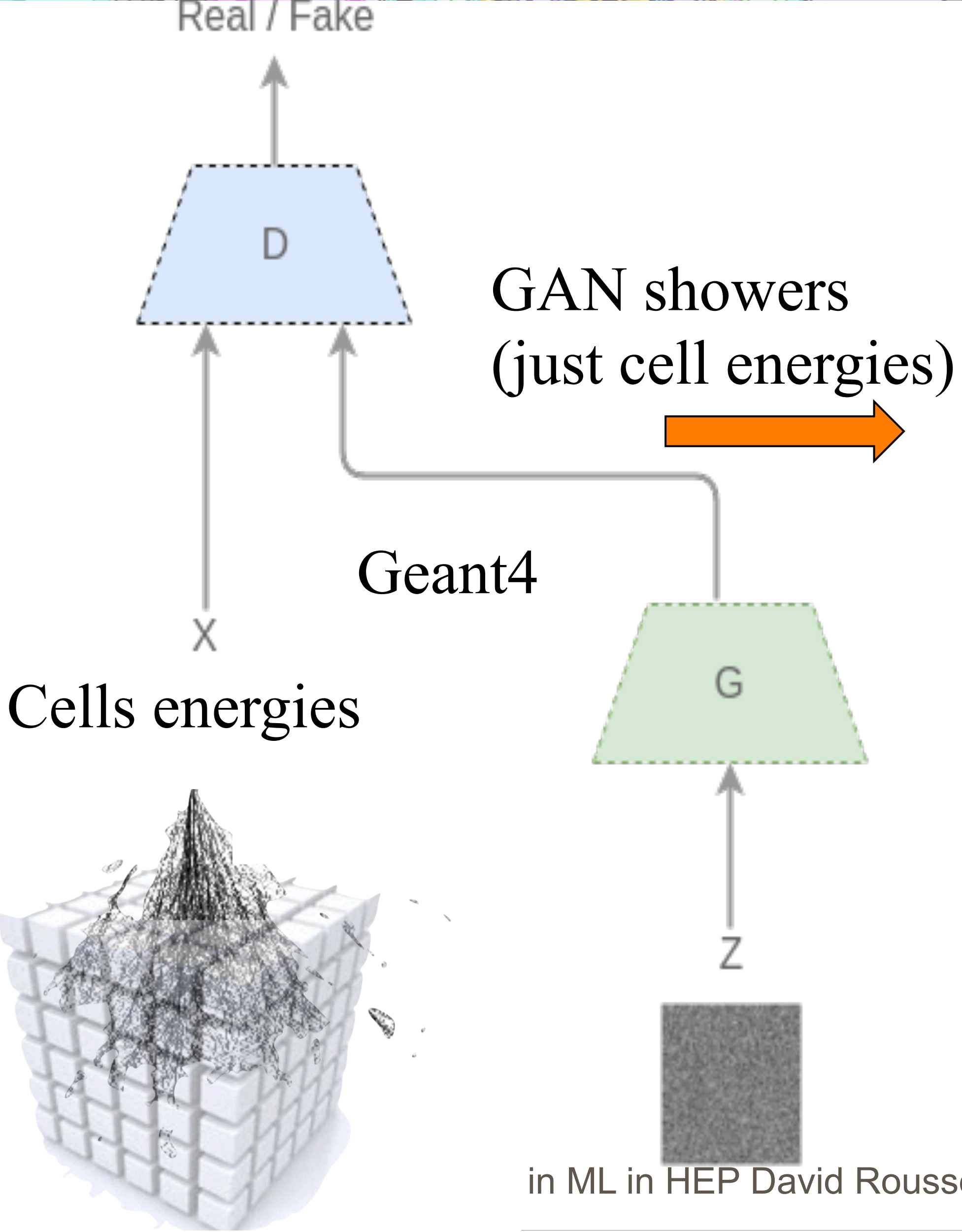
this small bird has a pink breast and crown, and black primaries and secondaries.

this magnificent fellow is almost all black with a red crest, and white cheek patch.

the flower has petals that are bright pinkish purple with white stigma

this white and yellow flower have thin white petals and a round yellow stamen

GAN for simulation

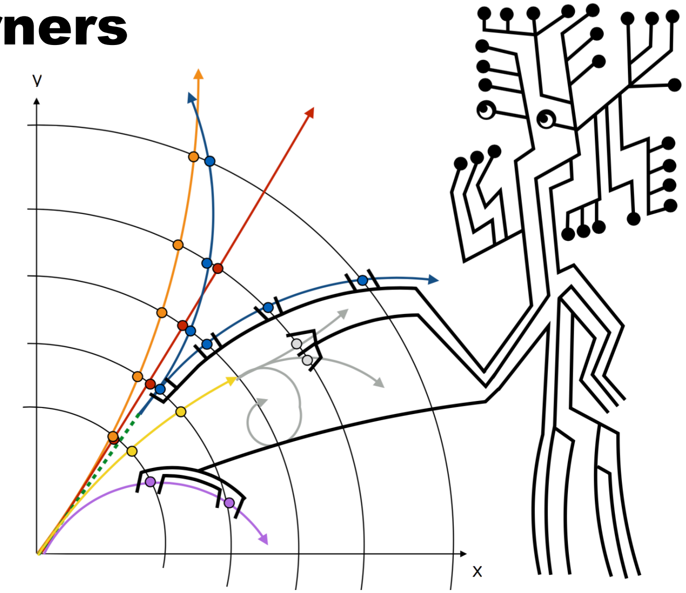


- Half of LHC grid computers (~300.000 cores) are crunching Geant4 simulation 24/24 365/365
- …while LHC experiments are collecting more and more events
- ➔reducing CPU consumption of simulation is very important

- Imagine training a GAN on single particle showers of all types and energies
- Then when an event is simulated it would ask for GAN showers on request (superfast by 3-4 order of magnitude)
- Would replace current fast simulation, frozen shower libraries….
- Just an idea until recently, but see https://arxiv.org/abs/1705.02355 ,also GeantV team is looking into this
- If/when it works, would require large GPU clusters

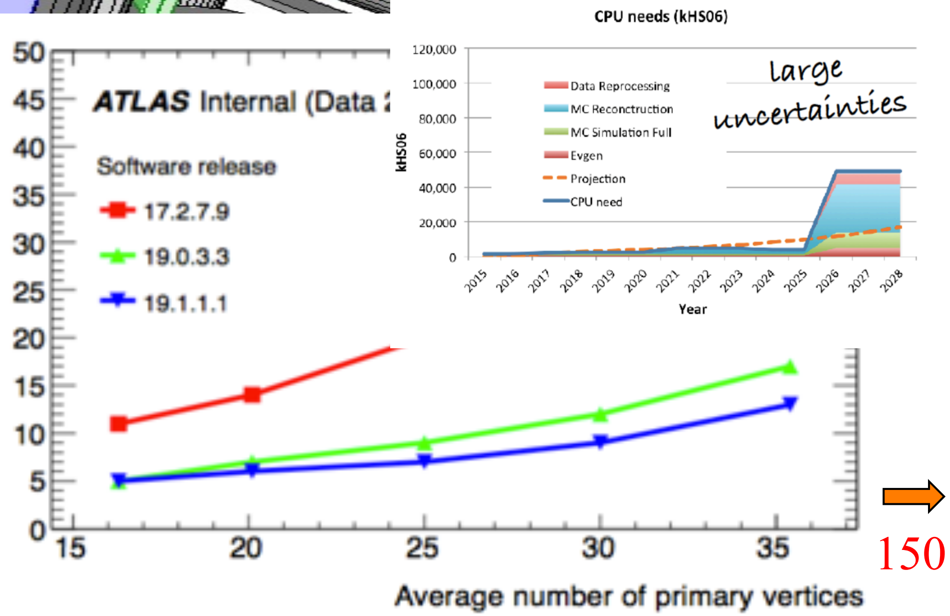
Tracking Machine Learning challenge 2018

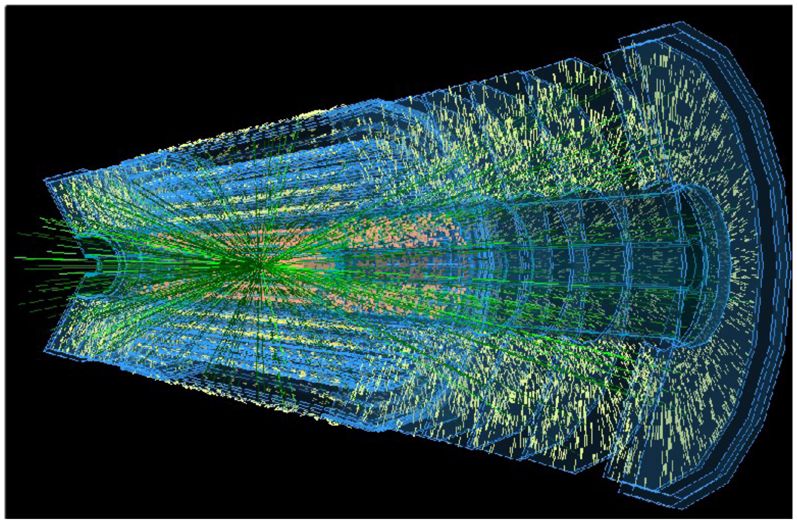
A collaboration between ATLAS and CMS physicists, and Machine Learners

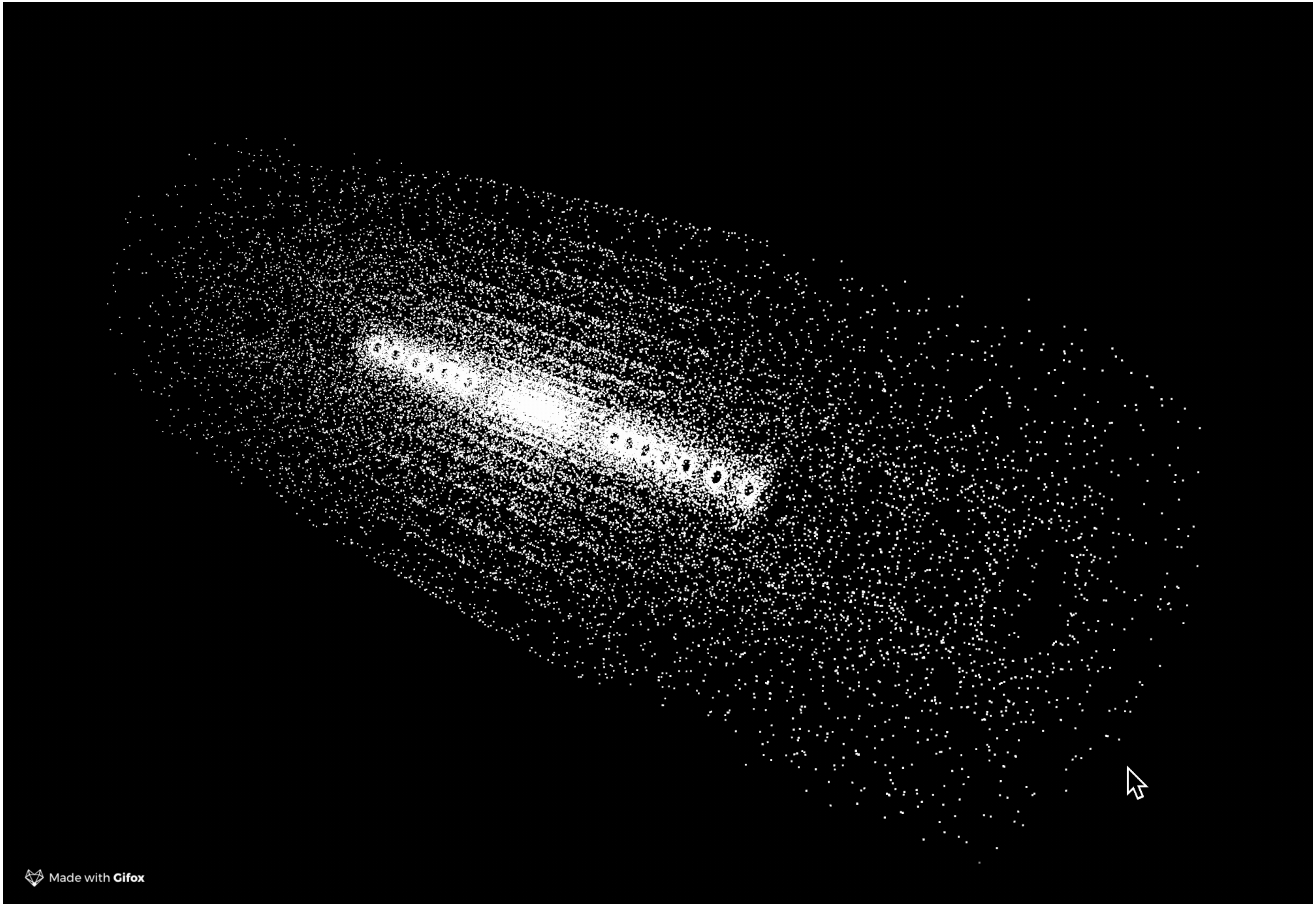


TrackML : Motivation

- Tracking (in particular pattern recognition) dominates reconstruction CPU time at LHC
- HL-LHC (phase 2) perspective : increased pileup :Run 1 (2012): <>~20, Run 2 (2015): <>~30,Phase 2 (2025): <>~150
- CPU time quadratic/exponential extrapolation (difficult to quote any number)
- Large effort within HEP to optimise software and tackle micro and macro parallelism. Sufficient gains for Run 2 but still a long way for HL-LHC.
- >20 years of LHC tracking development. Everything has been tried?
  - Maybe yes, but maybe algorithm slower at low lumi but with a better scaling have been dismissed ?
  - Maybe no, brand new ideas from ML (i.e. Convolutional NN)
- ➔Tracking challenge to be launched on Kaggle this April 2018
- Follow us on @trackmllhc !






Made with Gifox

Wrapping-up

ML Collaborations

- Many of the new ML techniques are complex→difficult for HEP physicists alone
- ML scientists (often) eager to collaborate with HEP physicists
  - prestige
  - new and interesting problems (which they can publish in ML proceedings)
- Takes time to learn common language
- Note : Yandex Data School of Analysis (with ~10 ML scientists) now a bona fide institute of LHCb
- Access to experiment internal data an issue, but there are ways out→more and more Open Dataset
- Very useful/essential to build HEP - ML collaborations : study on shared dataset, thesis (Computer Science or HEP)
- There is always a friendly Machine Learner on a campus!