Advances in Machine Learning in experimental High Energy Physics

David Rousseau
LAL-Orsay
rousseau@lal.in2p3.fr

Chinese Academy of Science visit to CC-IN2P3
8th Sep 2017

Outline

(note: I co-organise the ATLAS Machine Learning Forum and the IN2P3 ML project)

- ML in analysis
- ML in reconstruction/simulation
- ML challenges
- Wrapping up

Focus on applications rather than details of the techniques

ML in HEP

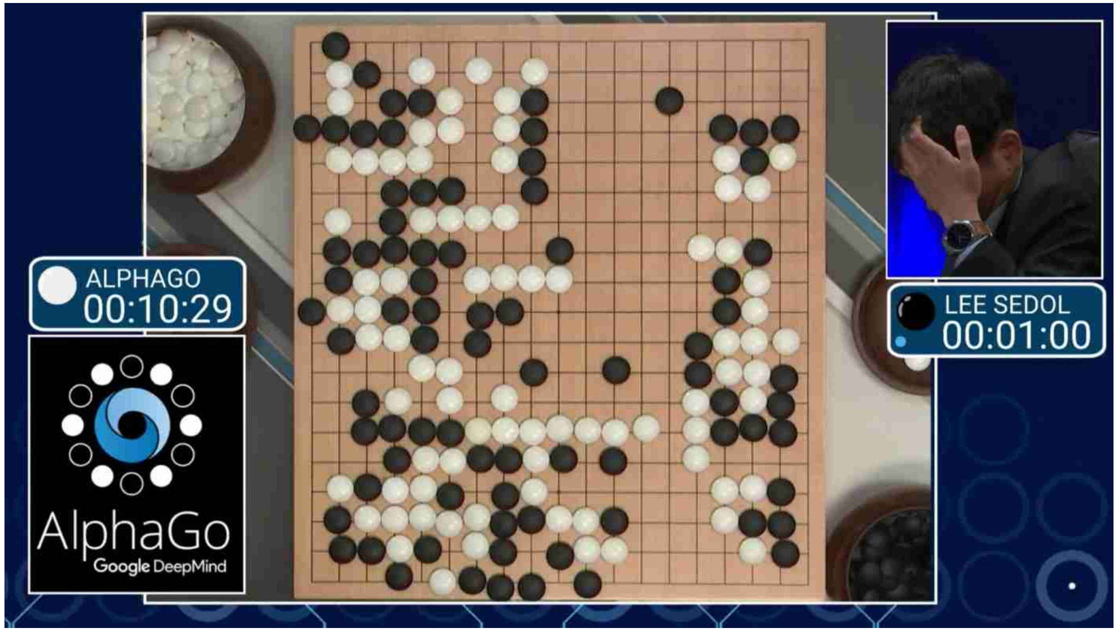
- Use of Machine Learning (a.k.a Multi Variate Analysis as we call it) already at LEP somewhat, much more at Tevatron (Trees)
- At LHC, Machine Learning used almost since first data taking (2010) for reconstruction and analysis
- In most cases, Boosted Decision Tree with Root-TMVA, on ~10 variables
-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

analysis	data taking year	no ML sensitivity	ML sensitivity	ML data gain
ATLAS $H \rightarrow \gamma\gamma$ [16]	2011-2012	4.3	-	-
CMS $H \rightarrow \gamma\gamma$ [17]	2011-2012	?	2.7	?
ATLAS $H \rightarrow \tau^+\tau^-$ [18]	2012	2.5	3.4	85%
CMS $H \rightarrow \tau^+\tau^-$ [19]	2012	3.7	-	-
ATLAS $VH \rightarrow bb$ [20]	2012	1.9	2.5	73%
ATLAS $VH \rightarrow bb$ [21]	2015-2016	2.8	3.0	15%
CMS $VH \rightarrow bb$ [22]	2012	1.4	2.1	125%
CMS $VH \rightarrow bb$ [23]	2015-2016	-	2.8	-

➔~50% gain on LHC running

ML in HEP

- Meanwhile, in the outside world :



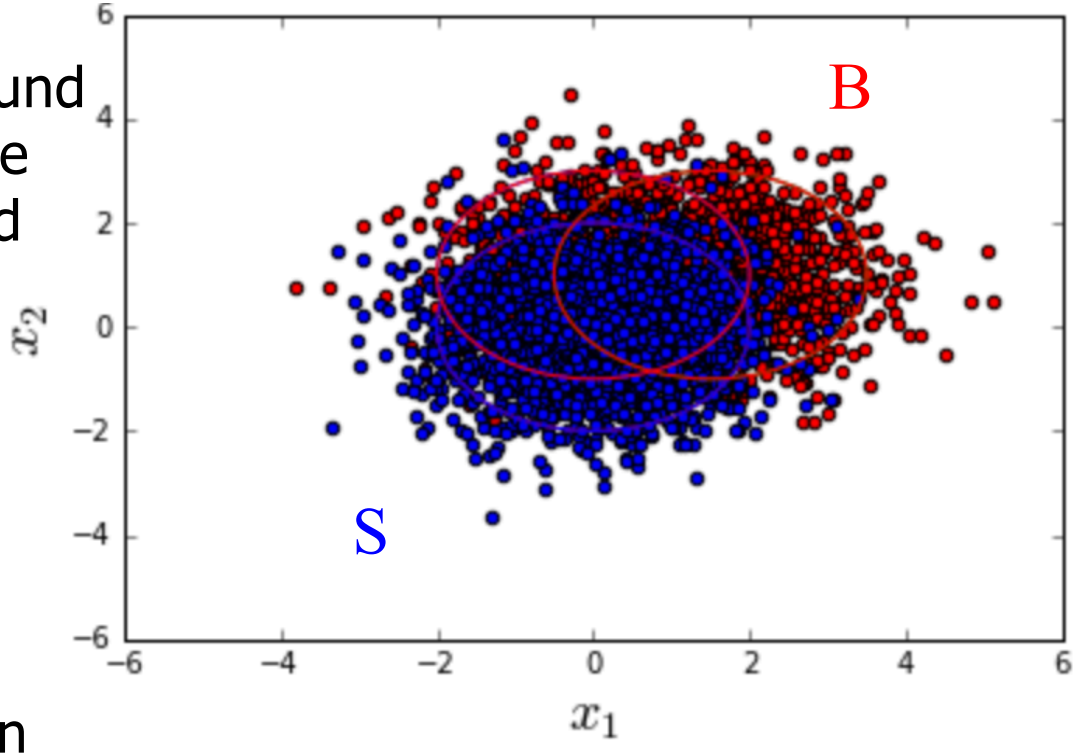
- “Artificial Intelligence” not a dirty word anymore!
- We’ve realised we’re been left behind! Trying to catch up now…

Multitude of HEP-ML events

- HiggsML Challenge, summer 2014
  - ➔HEP ML NIPS satellite workshop, December 2014
- Connecting The Dots, Berkeley, January 2015
- Flavour of Physics Challenge, summer 2015
  - ➔HEP ML NIPS satellite workshop, December 2015
- DS@LHC workshop, 9-13 November 2015
- LHC Interexperiment Machine Learning group
  - Started informally September 2015, gaining speed
  - IML workshop @CERN 20-22 March 2017
- Moscou/Dubna ML workshop 7-9th Dec 2015
- Heavy Flavour Data Mining workshop, 18-21 Feb 2016
- Connecting The Dots, Vienna, 22-24 February 2016
- Hep Software Foundation workshop 2-4 May 2016 at Orsay, M
- Connecting The Dots, LAL-Orsay, 6-9 March 2017
- DS@HEP workshop @FNAL 8-12 May 2017
- ACAT conference Seattle, Sep 2017

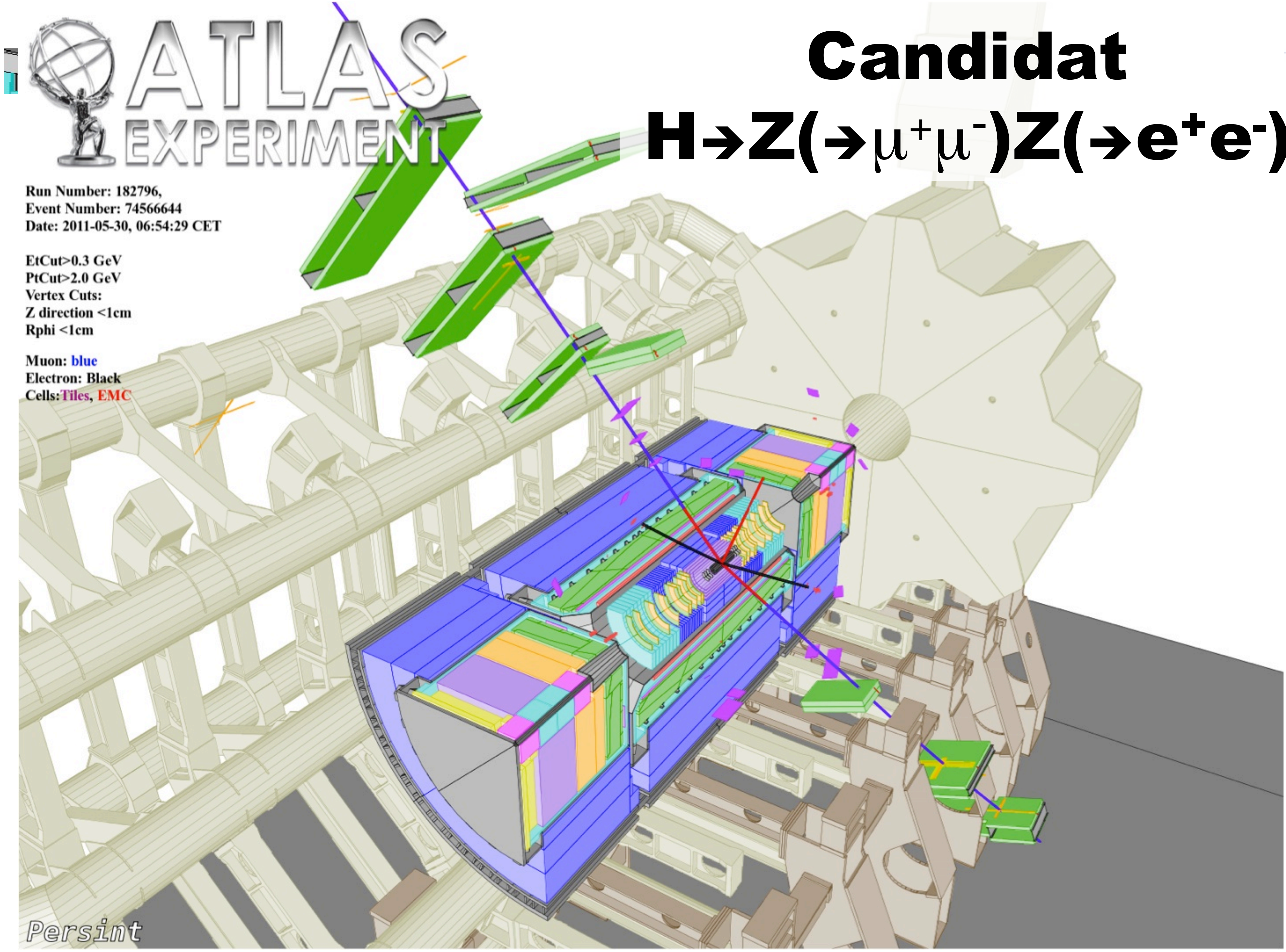
No miracle

- ML (nor Artificial Intelligence) does not do any miracles
- For selecting Signal vs Background and underlying distributions are known, nothing beats Likelihood ratio! (often called "bayesian limit"):
  - $L_S(x)/L_B(x)$
- OK but quite often L_S L_B are unknown
  - \+ x is n-dimensional
- ML starts to be interesting when there is no proper formalism of the pdf
- ➔mixed approach, if you know something, tell your classifier instead of letting it guess

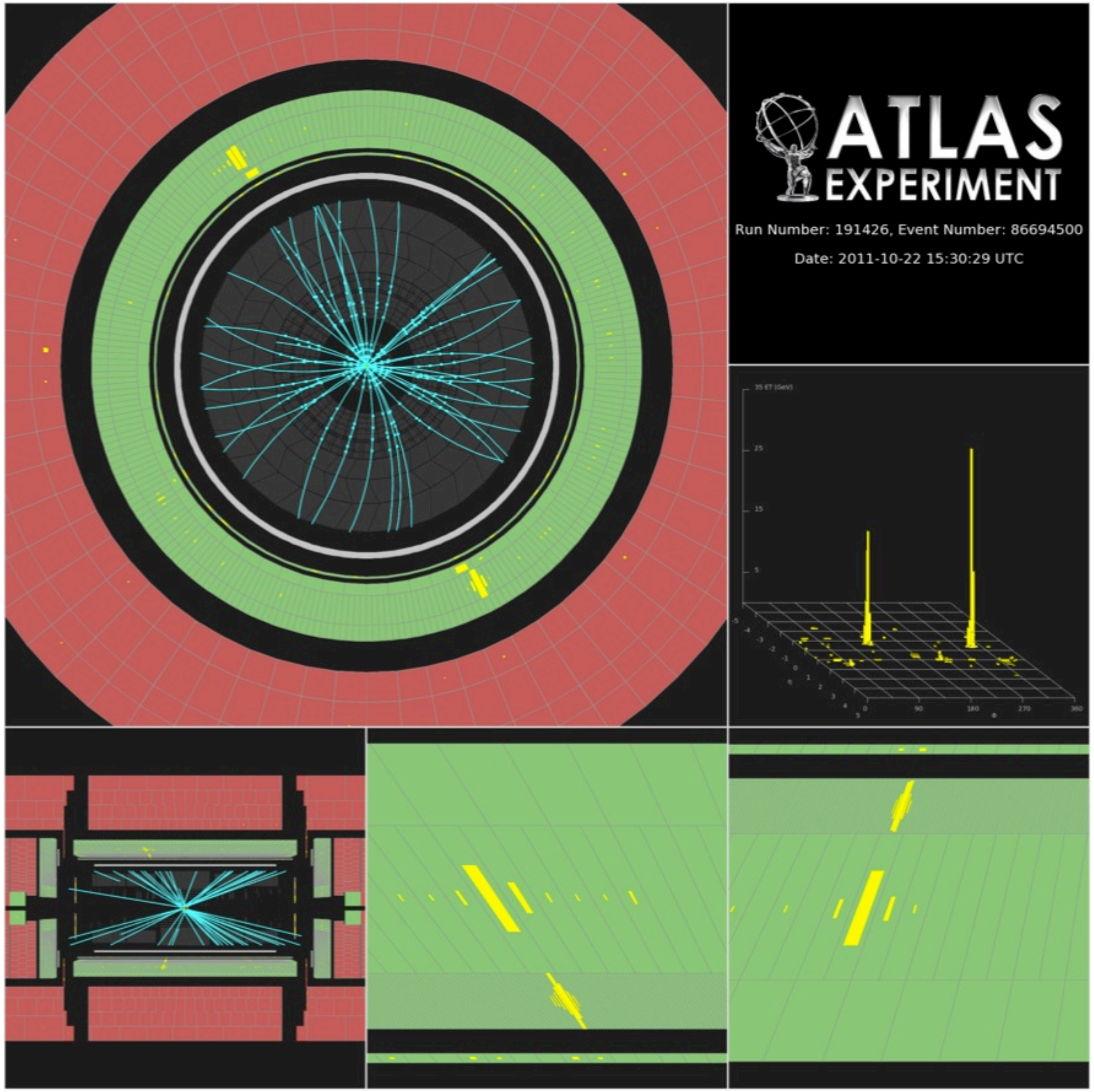


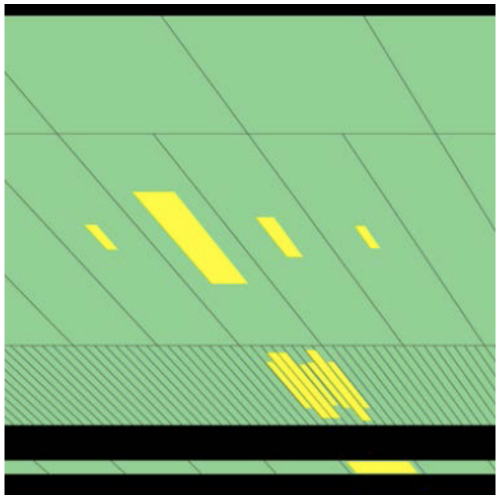
ML in analysis

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



Candidat H→ gamma gamma

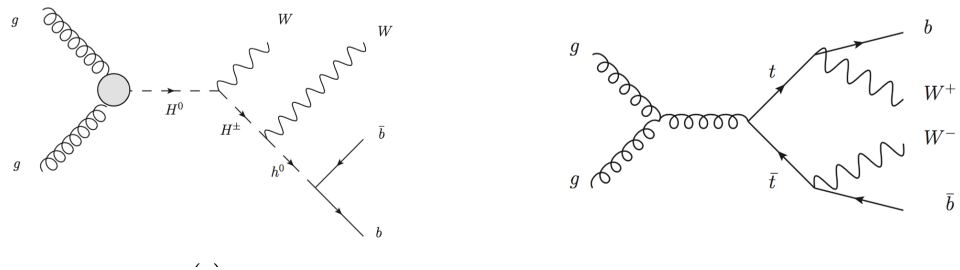


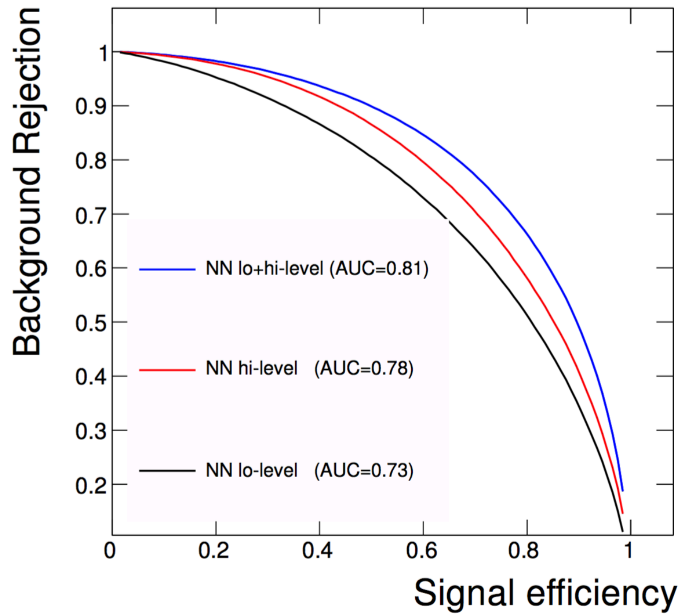


Neutral pion

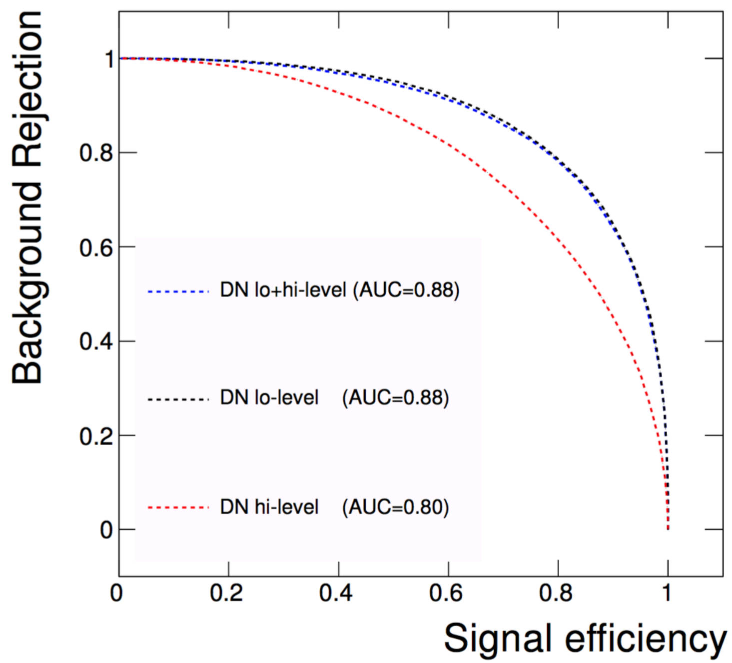
Deep learning for analysis

1402.4735 Baldi, Sadowski, Whiteson





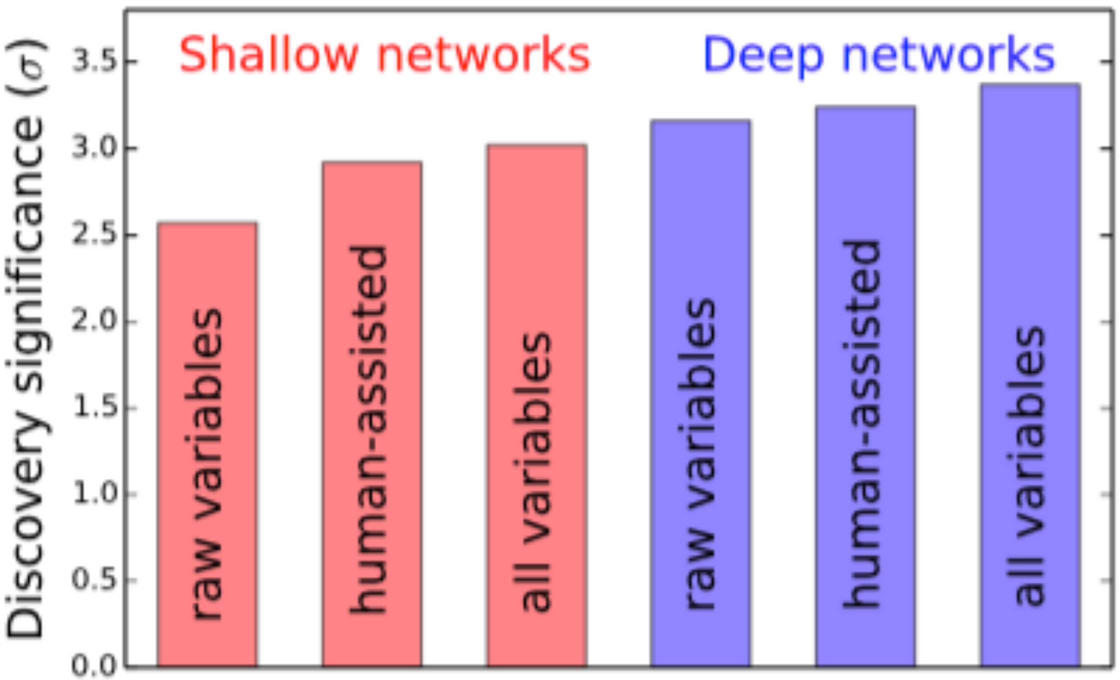
- MSSM at LHC : $H^0 \rightarrow WWbb$ vs $tt \rightarrow WWbb$
- Low level variables:
  - 4-momentum vector
- High level variables:
  - Pair-wise invariant masses
- Deep NN outperforms NN, and does not need high level variables
- DNN learns the physics ?



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
- ~10M events used for training (>>10* full G4 simulation in ATLAS)

Systematics-aware training

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

$$\sigma(\text{stat}) \pm \sigma(\text{syst})$$

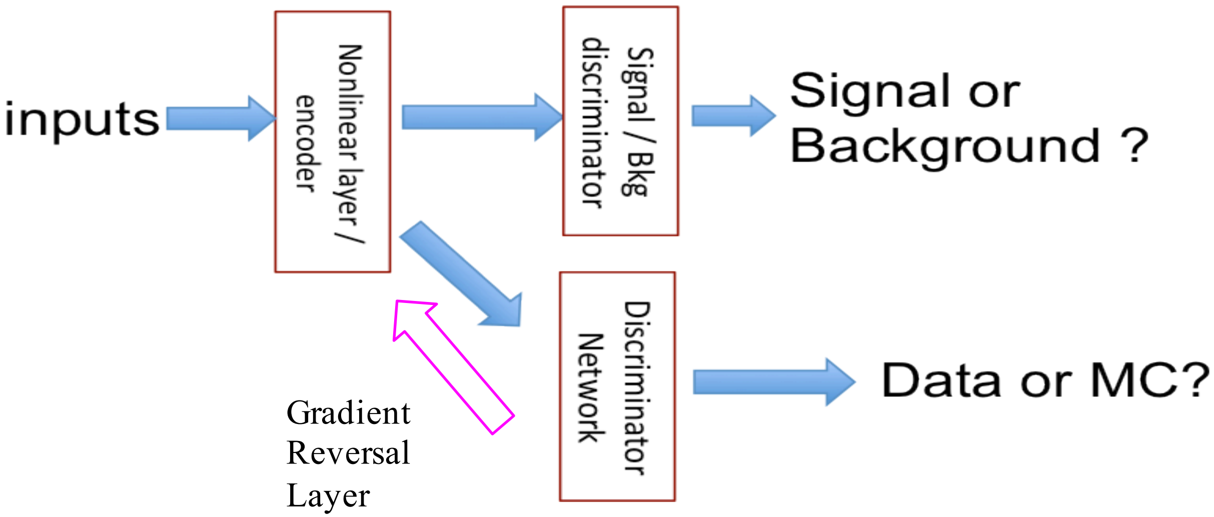
- ML techniques used so far to minimise σ(stat)
- Impact of ML on σ(syst) or even better global optimisation of σ(stat) $\pm$ σ(syst) is an open problem
- Worrying about σ(syst) untypical of ML in industry

Systematics aware training

- 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
- One possible approach (many on-going)

Adversarial neural networks

Adapted from : 1505.07818 Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, Lempitsky



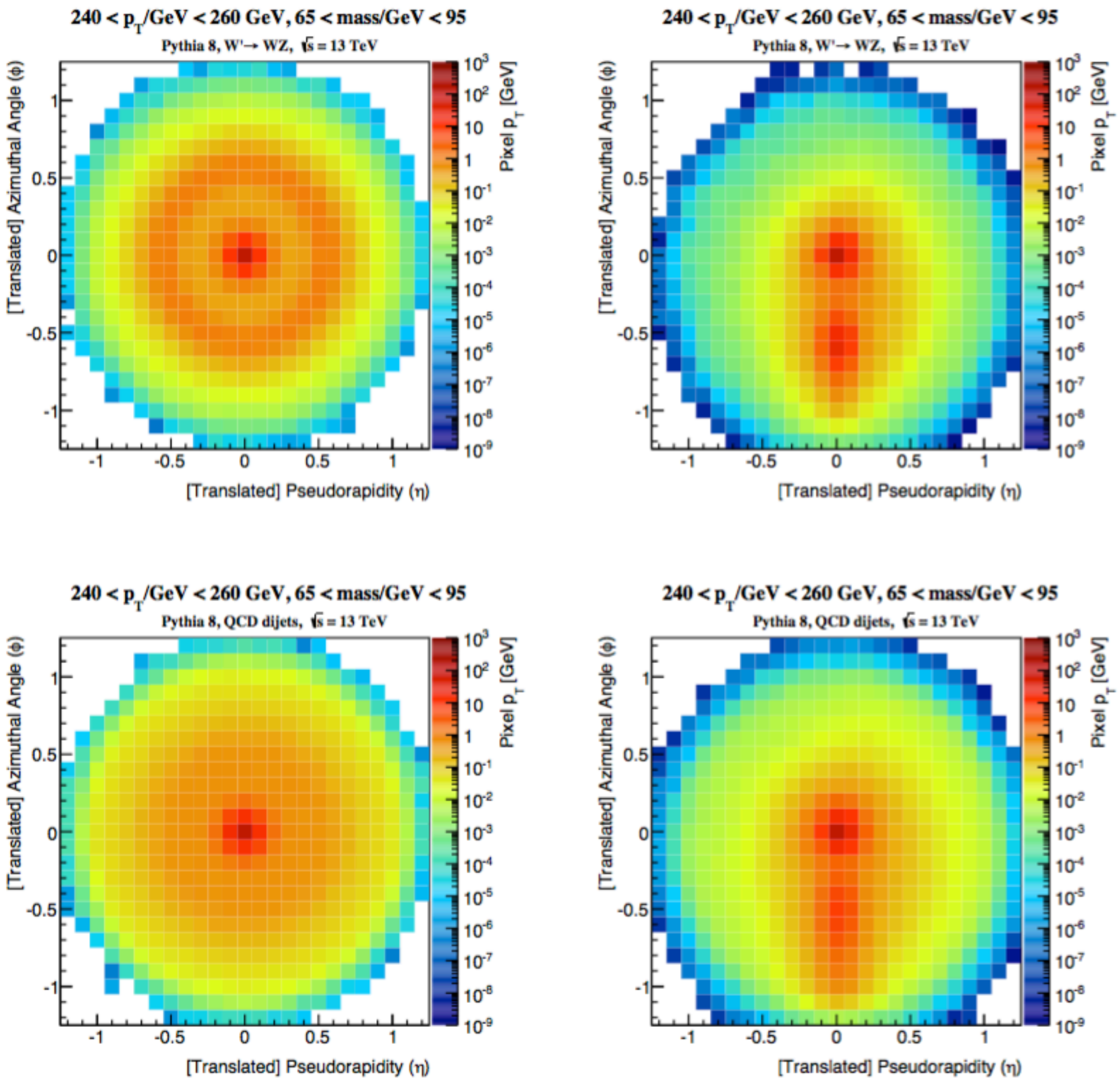
See ACAT 2017 Ryzhikov and Ustyuzhanin

ML in reconstruction

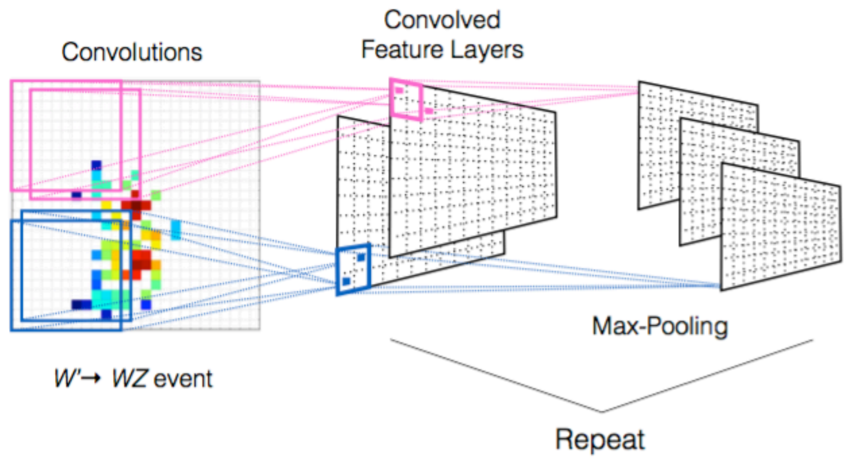
Jet Images

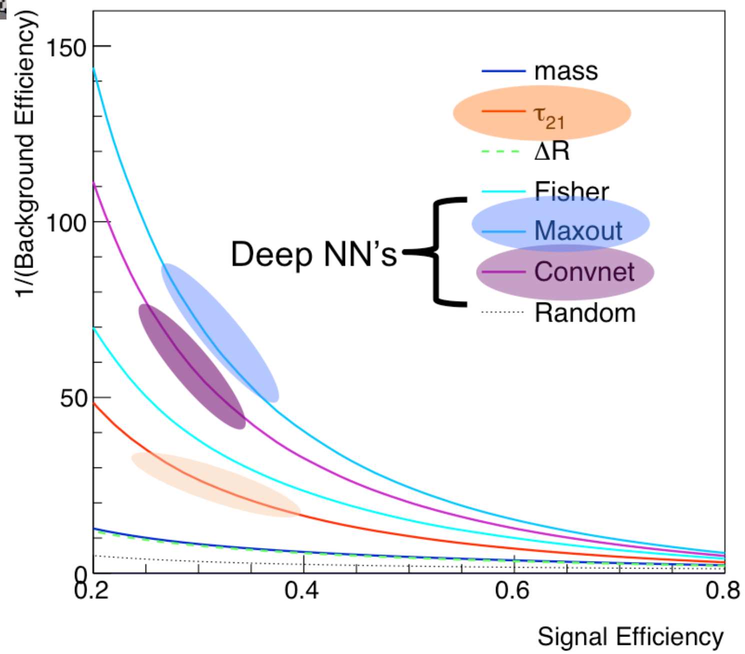
arXiv 1511.05190 de Oliveira, Kagan, Mackey, Nachman, Schwartzman

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

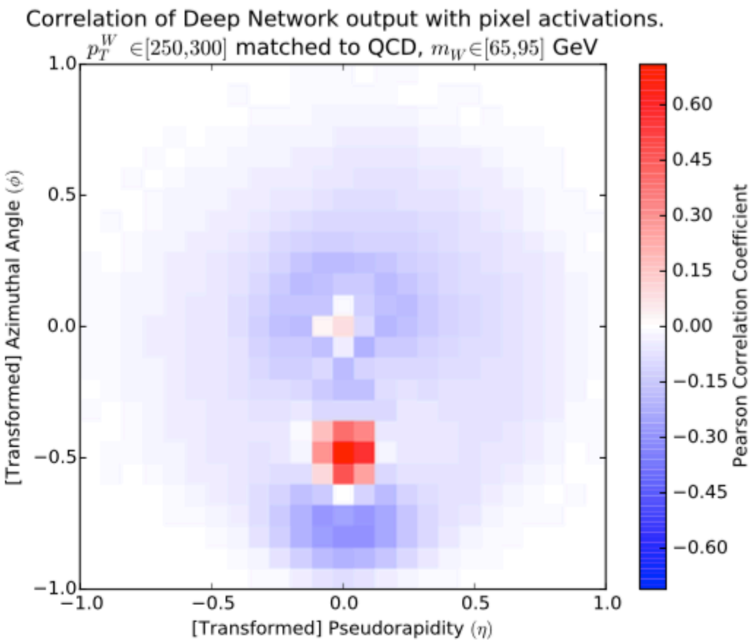


Jet Images : Convolution NN



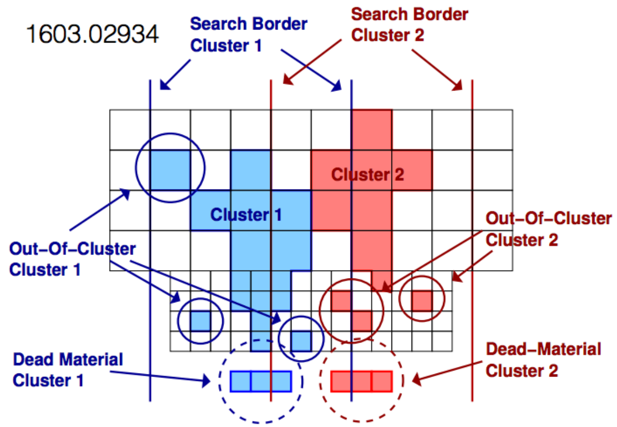


- Variables build from CNN outperform the more usual ones



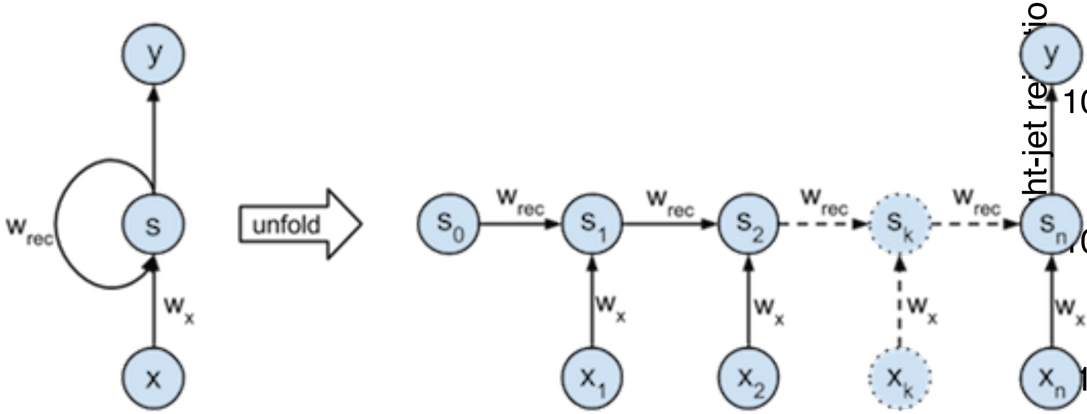
- What the CNN sees (the "cat" neurone")
- Now need proper detector and pileup simulation
- ➔3Dimension

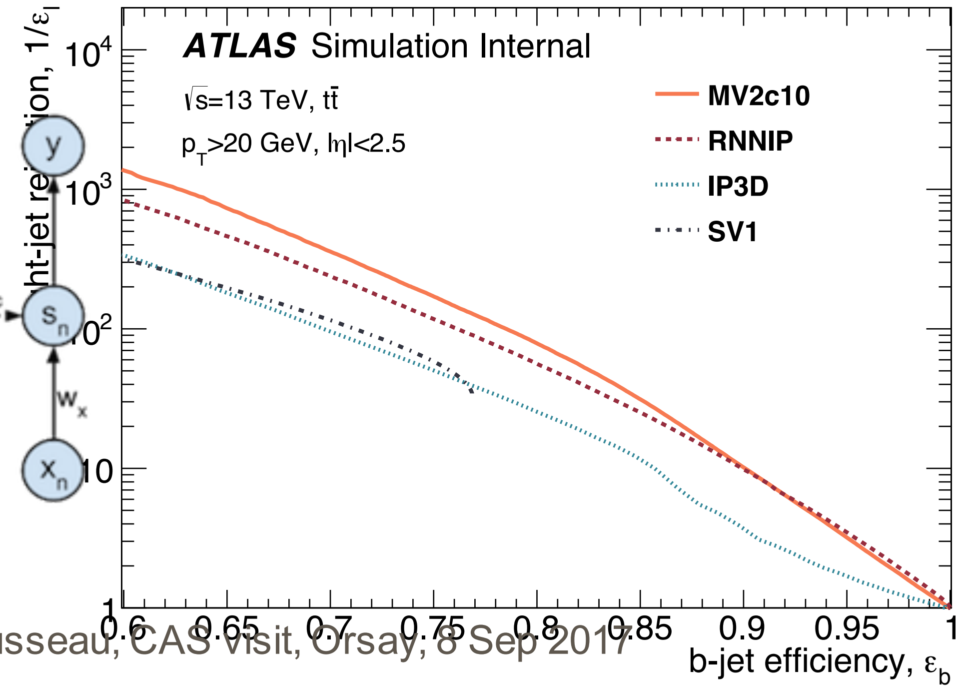
1603.02934



RNN for b tagging

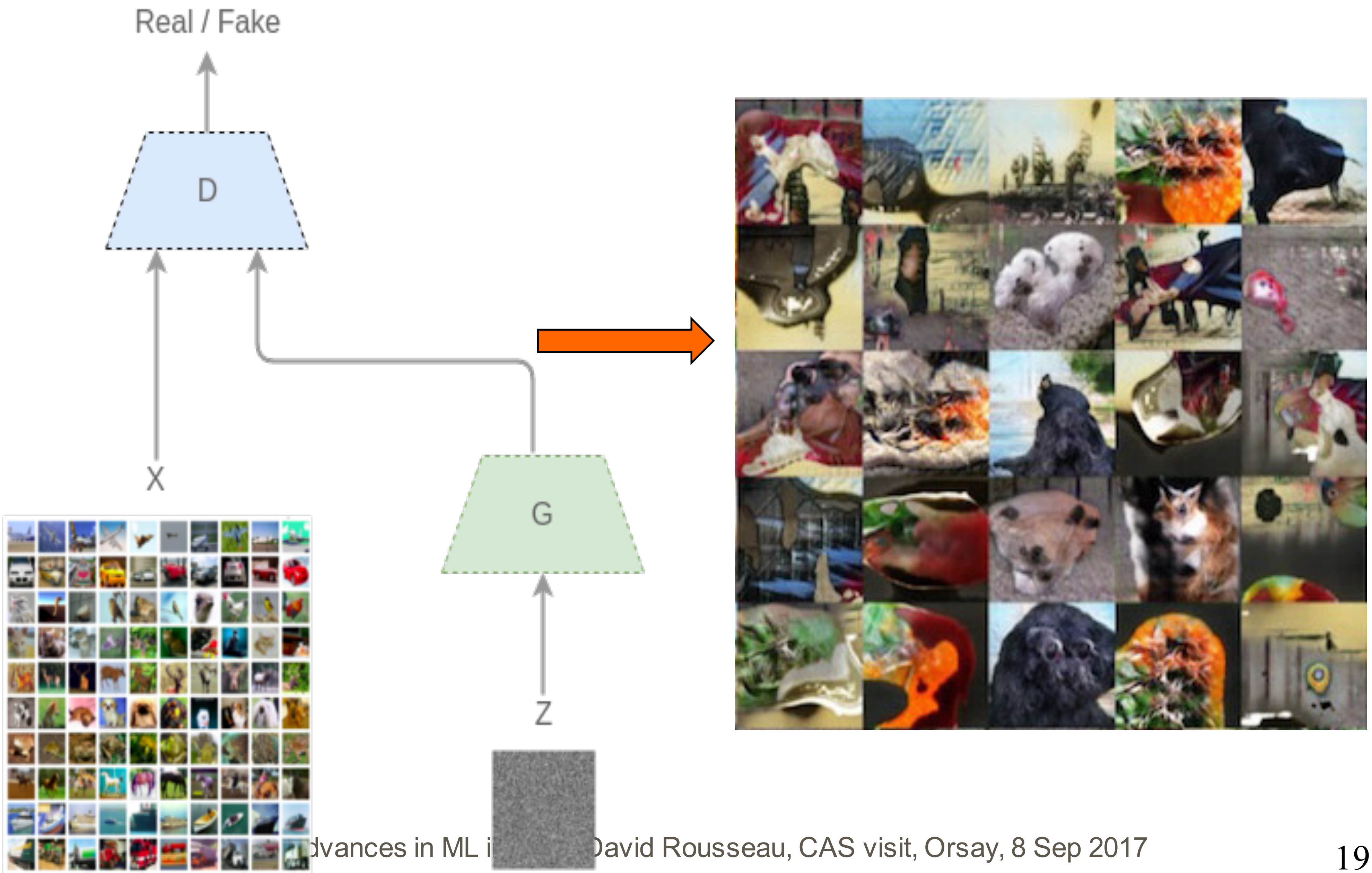
- BDT and usual NN expect a fix number of input. What to do when the number of inputs is not fixed like the tracks for b-quark jet tagging ?
- Recurrent neural networks have seen outstanding performance for processing sequence data
  - Take data at several "time-steps", and use previous time-step information in processing next time-steps data
- For b-tagging, take list of tracks in jet and feed into RNN
  - Basic track information like d0, z0, pt-Fraction of jet, …
  - Physics inspired ordering by d0-significance
- RNN outperforms other IP algorithms
  - No explicit vertexing, still excellent performance
  - First combinations with other algorithms in progress
- Learning on sequence data may be important in other places!
  - Combining tracks with clusters? Track to vertex mat





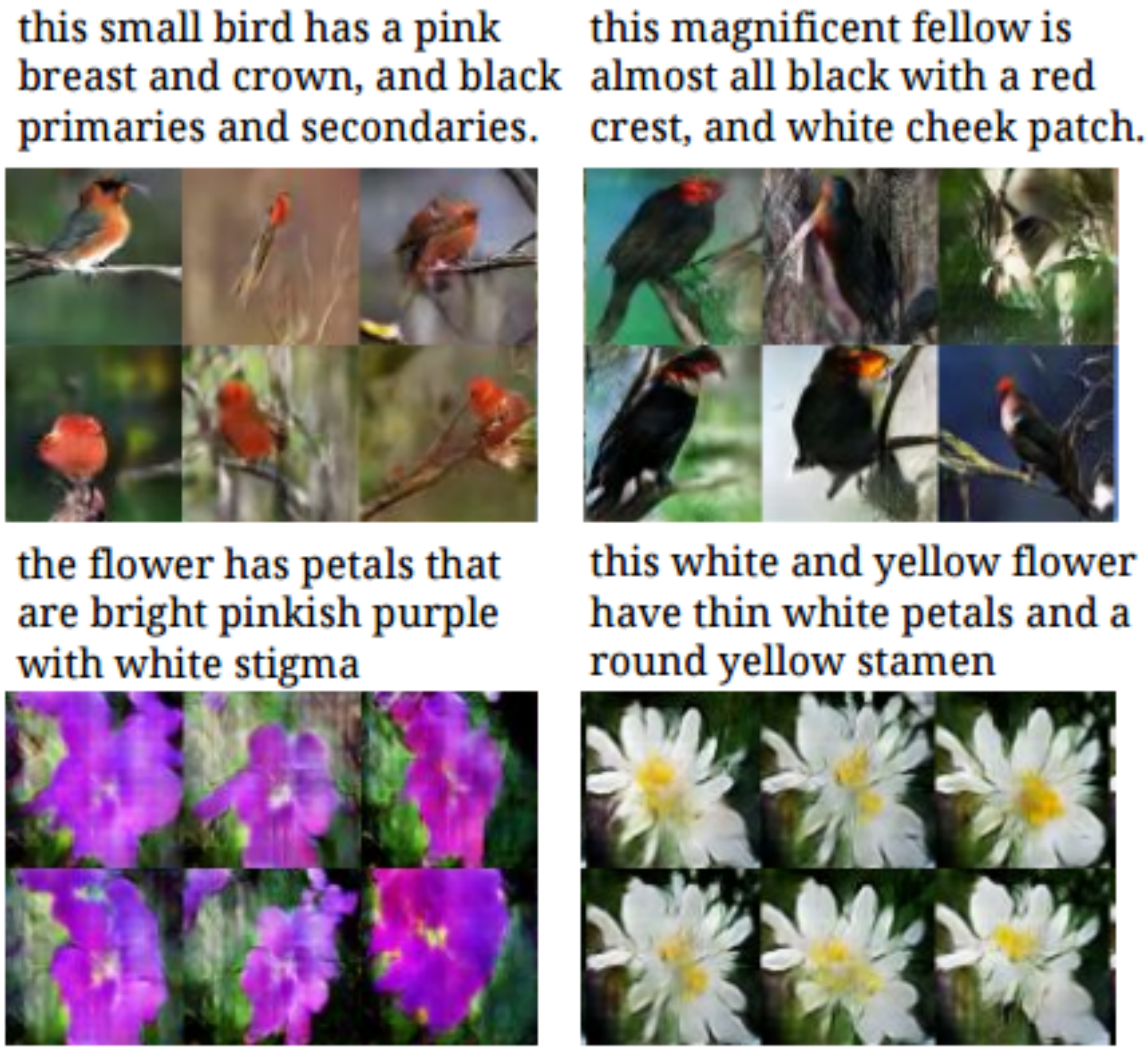
ML in simulation

Generative Adversarial Network

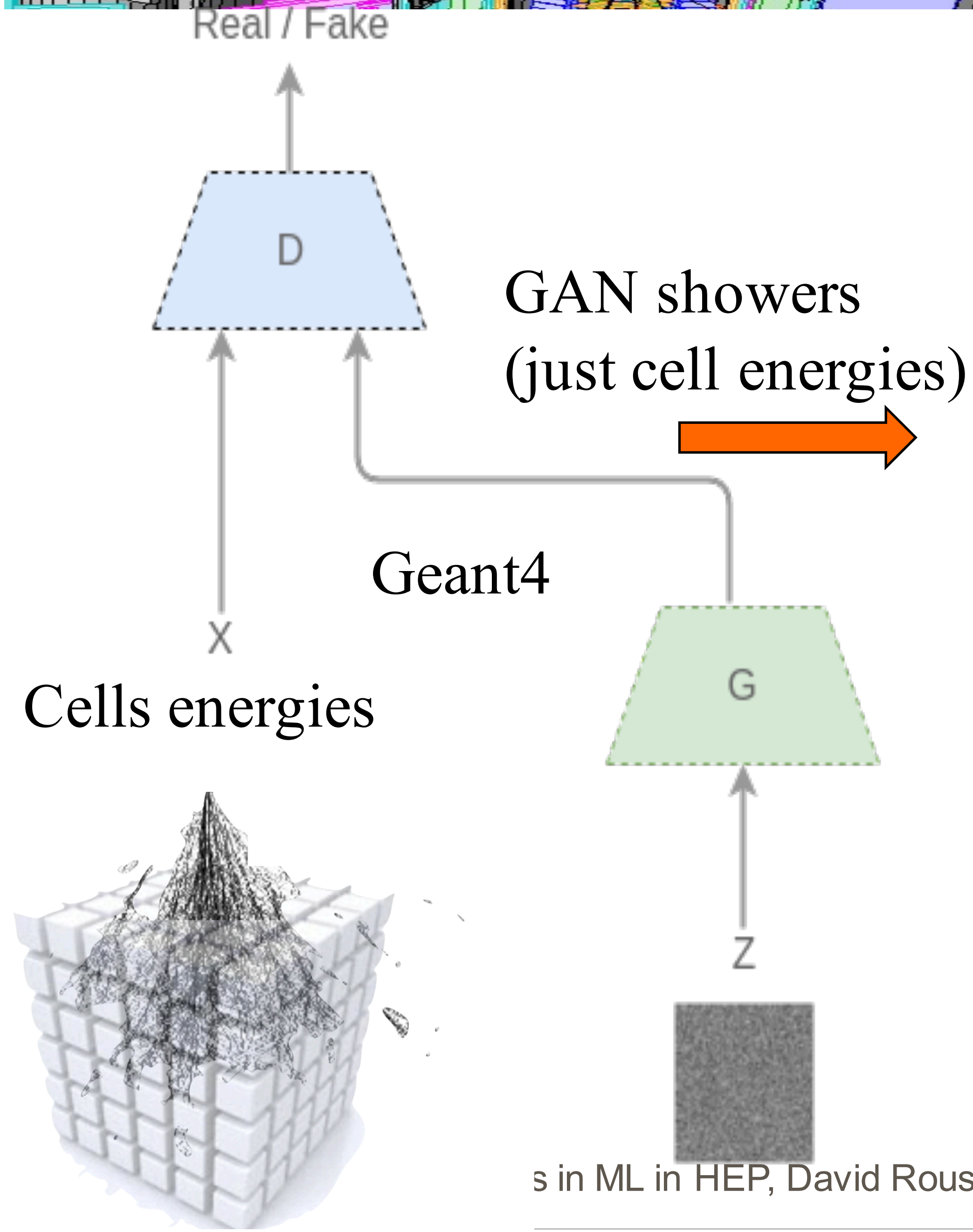


Condition GAN

Text to image



GAN for simulation

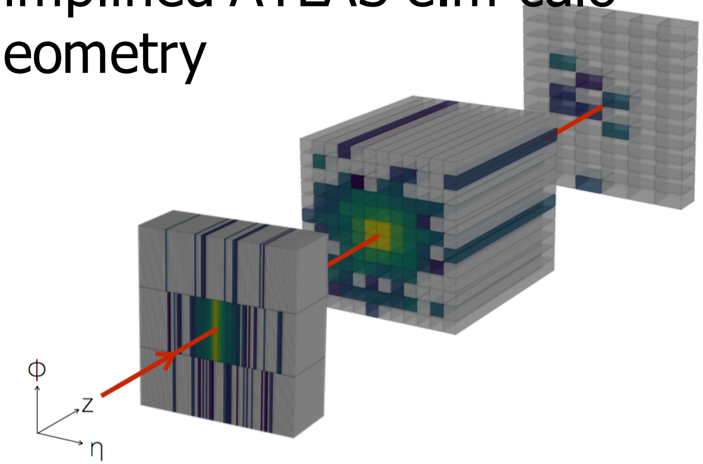


- Half of LHC grid computers (~300.000 cores) are crunching Geant4 simulation 24/24 365/365
- …while LHC experiments are collecting more and more events
- ➔reducing CPU consumption of simulation is very important

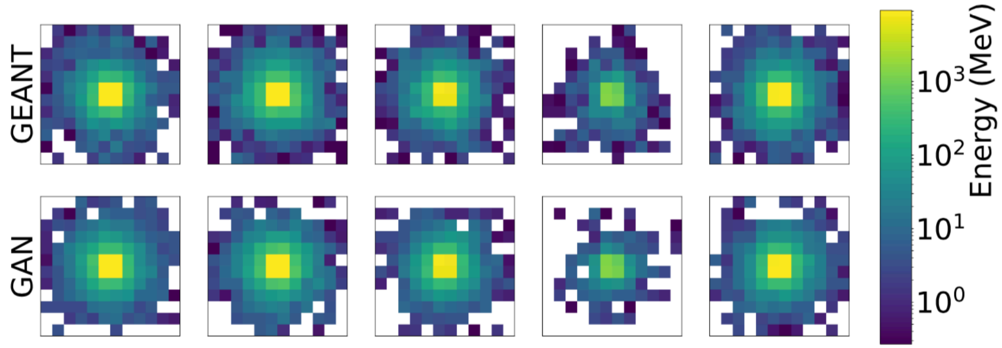
- Imagine training a GAN on single particle showers of all types and energies
- Then when an event is simulated it would ask for GAN showers on request (superfast by 3-4 order of magnitude)
- Would replace current fast simulation, frozen shower libraries….
- Just an idea until recently, but see https://arxiv.org/abs/1705.02355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

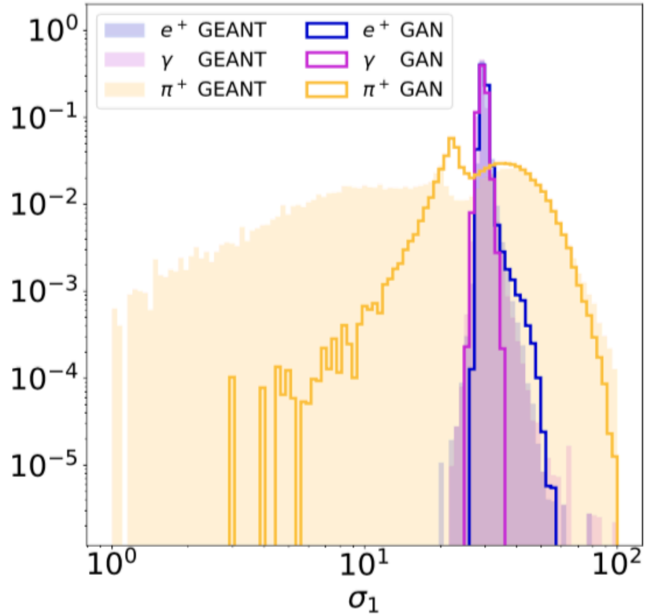
CaloGAN

Simplified ATLAS e.m calo geometry



[Paganini et al.]

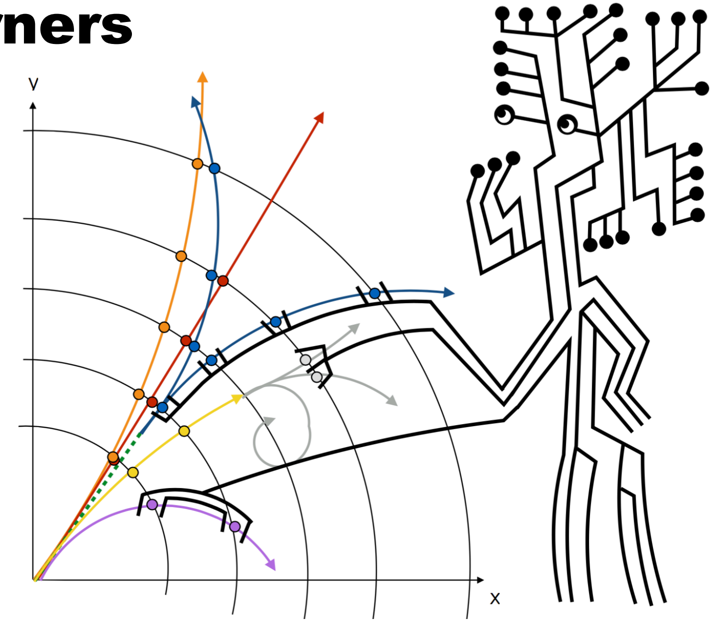




- σ_1:width in Middle layer
- One of many physics variable examined
- Pion more difficult
- ➔very promising

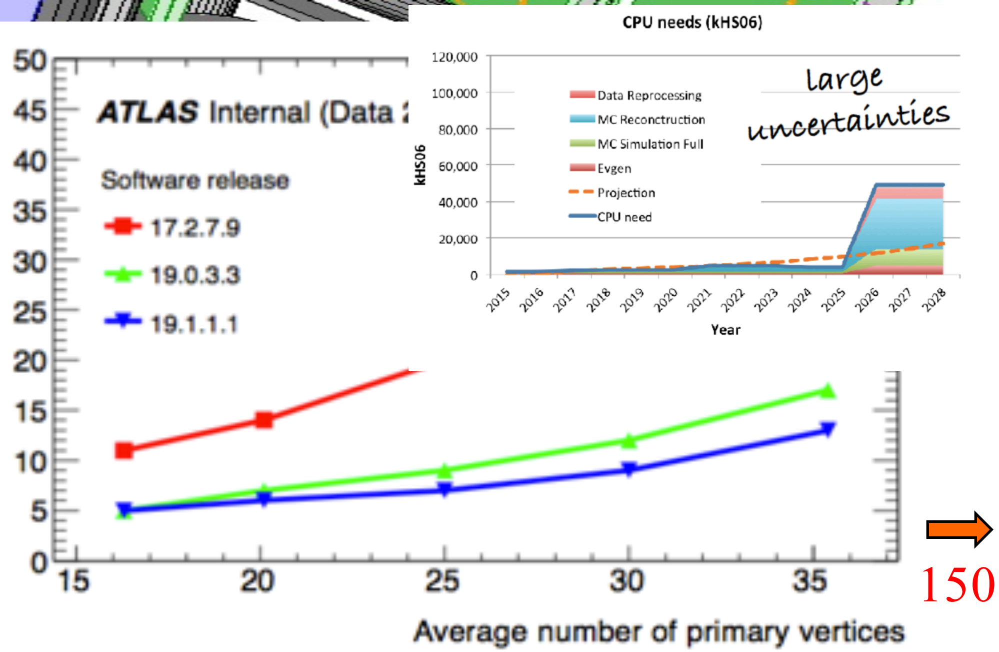
Towards a Future Tracking Machine Learning challenge

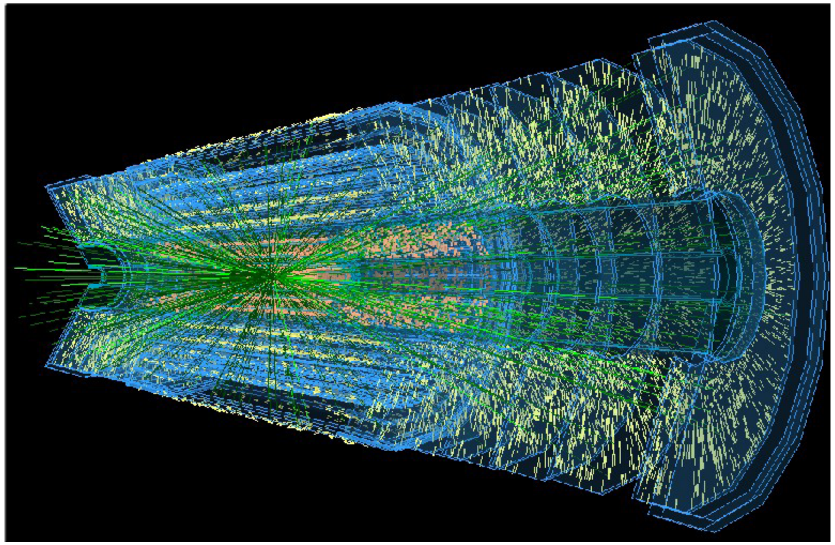
A collaboration between ATLAS and CMS physicists, and Machine Learners



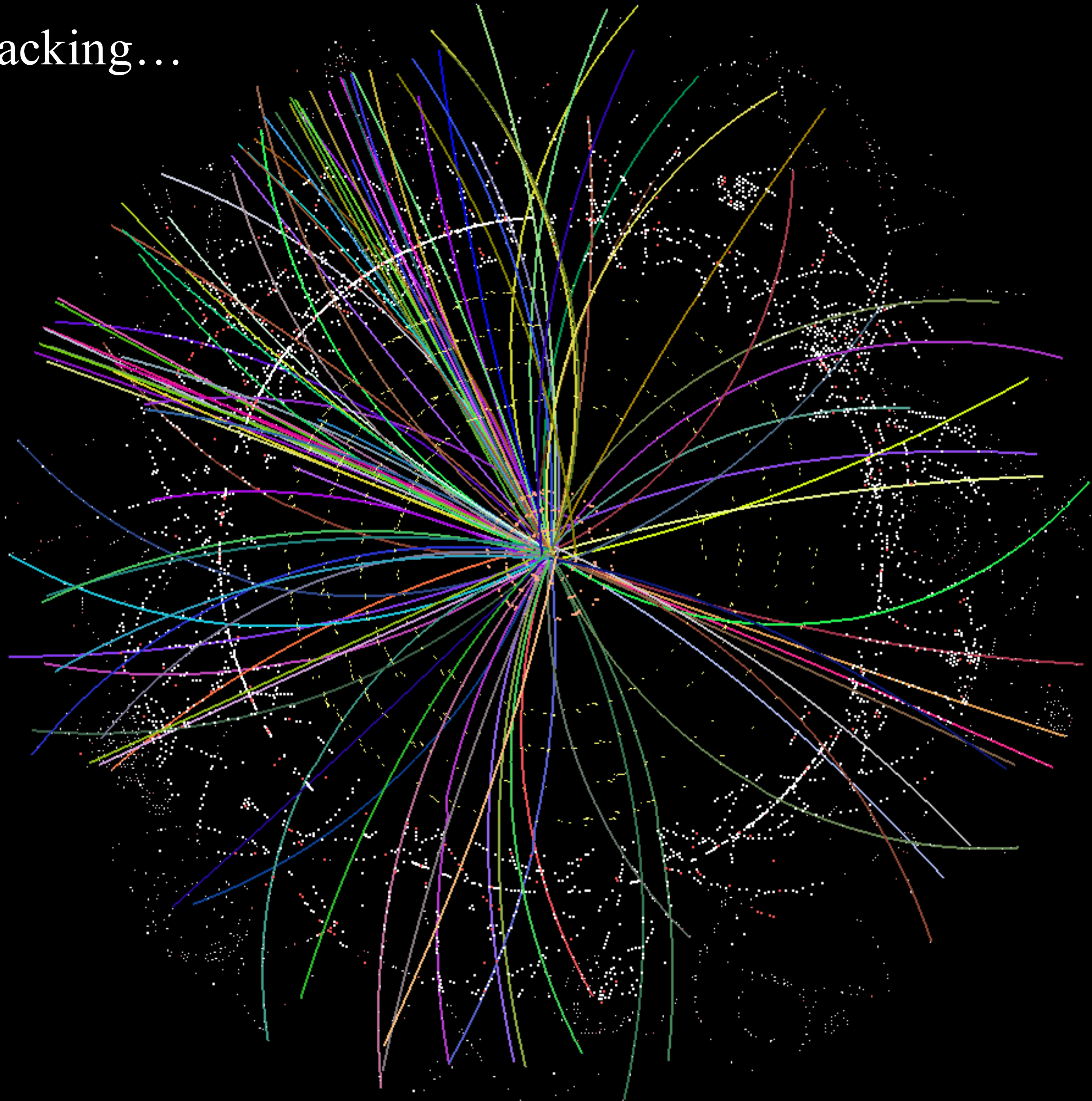
TrackML : Motivation

- See details [DR talk at CTD2016]
- 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)

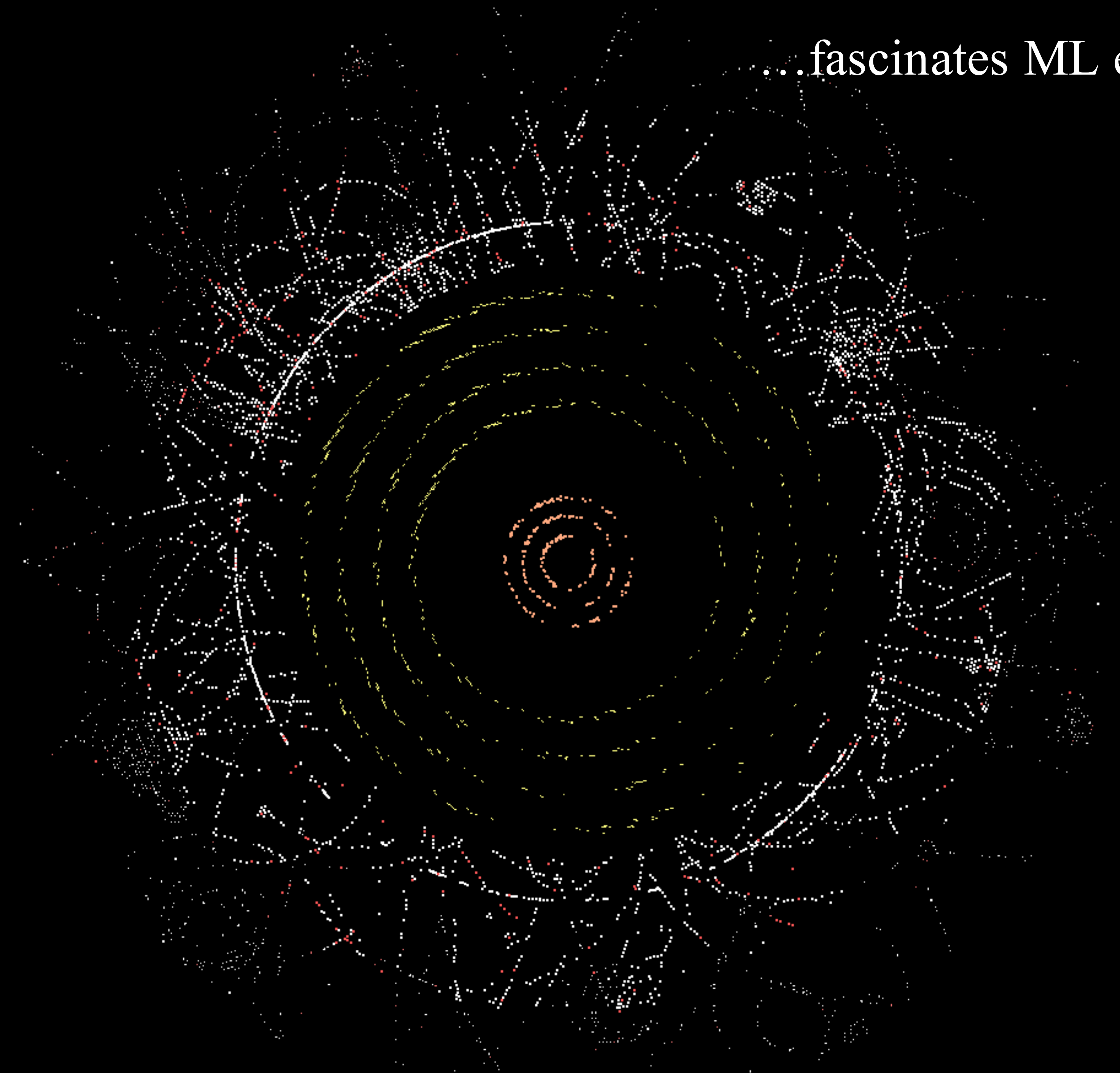




HEP tracking…



…fascinates ML experts

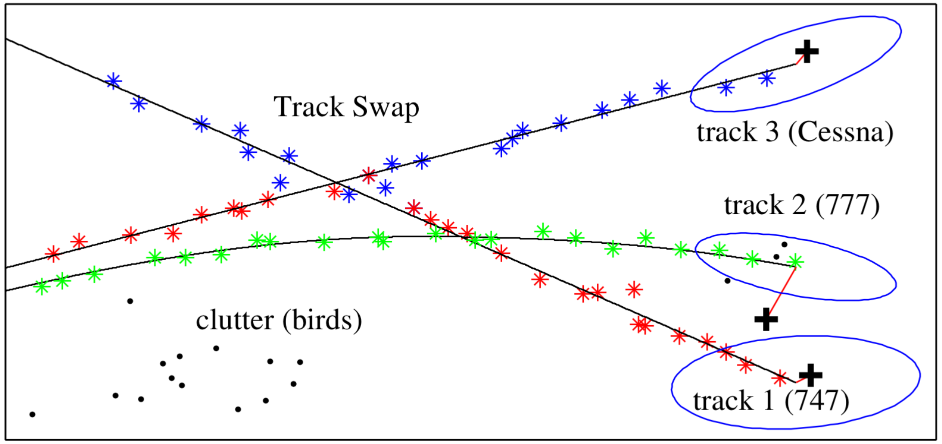


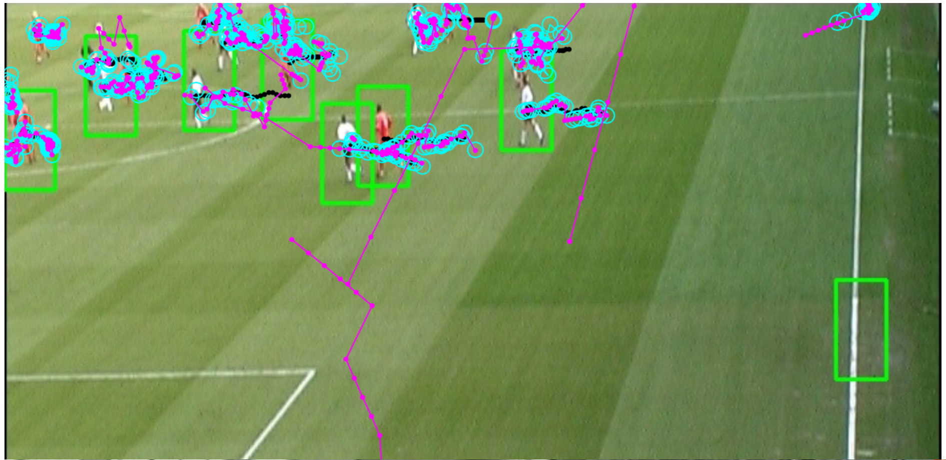
TrackML : engaging Machine Learners

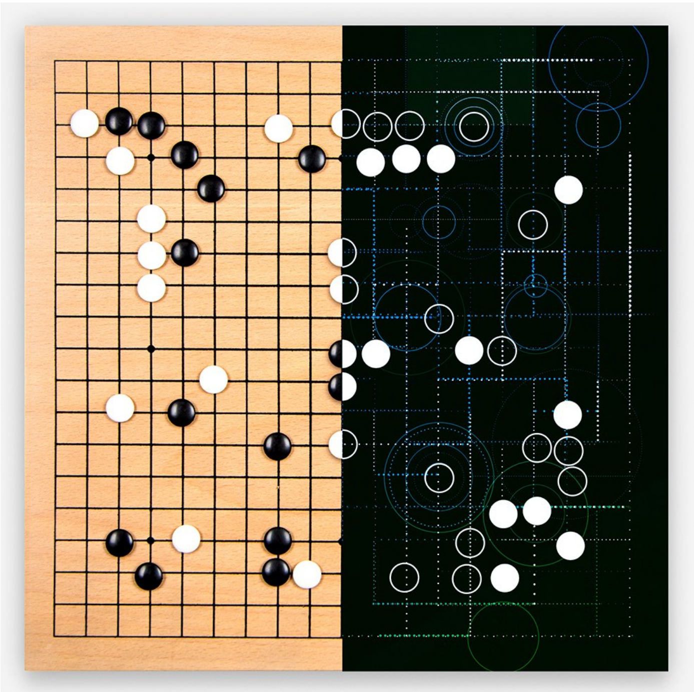
- Suppose we want to improve the tracking of our experiment
- We read the literature, go to workshops, hear/read about an interesting technique (e.g. ConvNets, MCTS…). Then:
  - Try to figure by ourself what can work, and start coding → traditional way
  - Find an expert of the new technique, have regular coffee/beer, get confirmation that the new technique might work, and get implementation tips → better
- …repeat with each technique...
- Much much better:
  - Release a data set, with a benchmark, and have the expert do the coding him/herself
  - → he has the software and the know-how so he'll be (much) faster even if he does not know anything about our domain at the beginning
  - → engage multiple techniques and experts simultaneously (e.g. 2000 people participated to the Higgs Machine Learning challenge) in a comparable way
  - → even better if people can collaborate
  - → a challenge is a dataset with a benchmark and a buzz
  - Looking for long lasting collaborations beyond the challenge
- Focus on the pattern recognition : release list of 3D points, challenge is to associate them into tracks fast. Use public release of ATLAS tracking (ACTS) as a simulation engine and starting kit

Pattern recognition

- Pattern recognition is a very old, very hot topic in Artificial Intelligence,
- Note that these are real-time applications, with CPU constraints

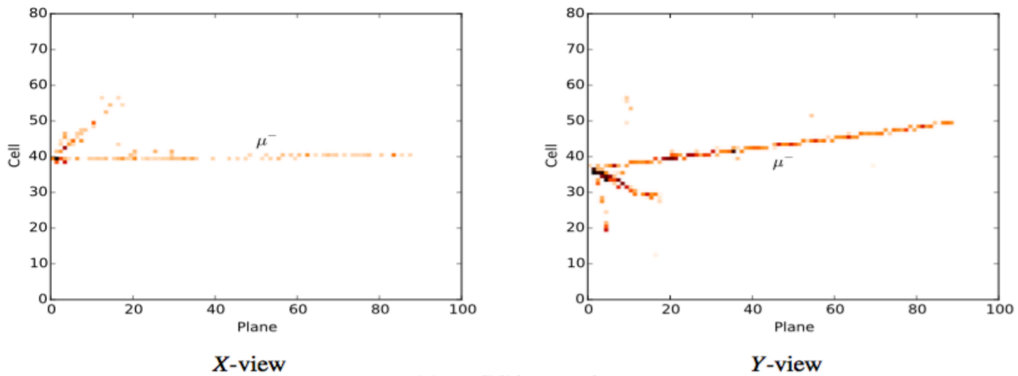




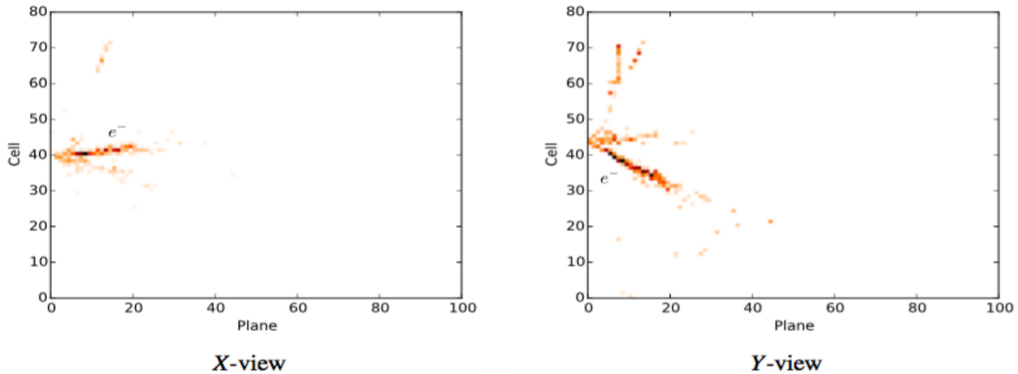


A recent attempt : NOVA

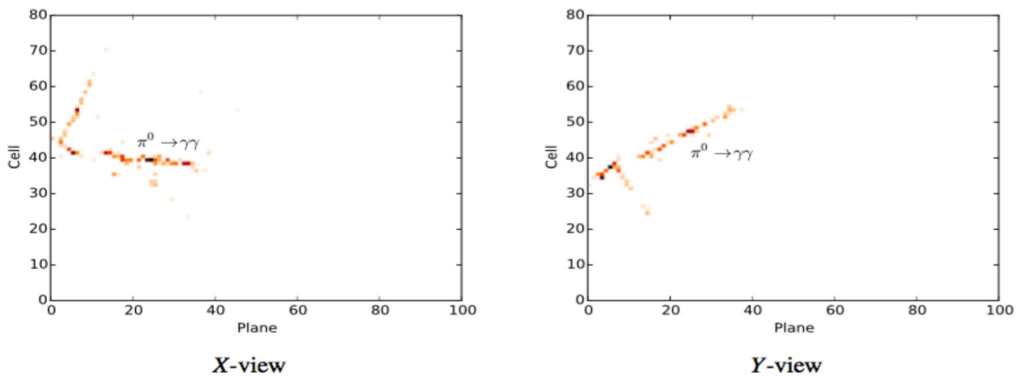
arXiv 1604.01444 Aurisano et al



(a) ν_μ CC interaction.

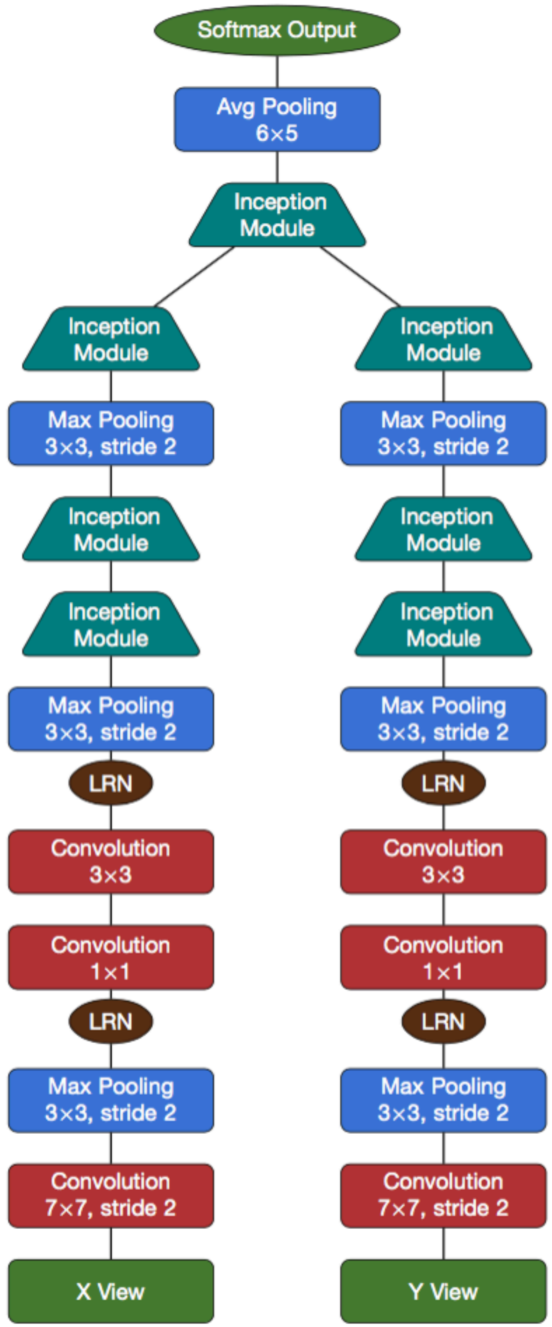


(b) ν_e CC interaction.



(c) NC interaction.

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



CTDWIT 2017 2D tracking Hackathon

CTDWIT 6-9th March 2017 LAL-Orsay

- Very simplified 2D simulation with HL-LHC ATLAS layout (circular detectors, multiple scattering, inefficiency, stopping tracks)

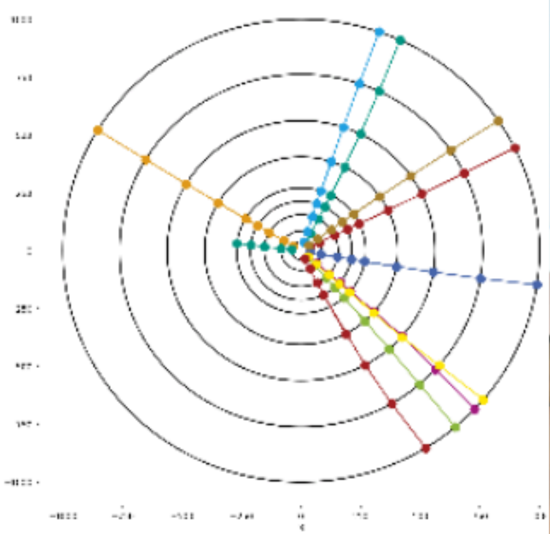
EPJ Web Conf., 150 (2017) 00015

- Run on RAMP platform
- 30 people (tracking experts mostly) for 2 hours in the same room, plus 36 hours till the end of the conference
- Winner is a Monte Carlo Tree Search algorithm (used in Go algorithms before and also by Alpha-Go)
- Runner-up a "real" ML algorithm : Long Short Term Memory

Belle II Experiment @belle2collab · 15 min

Congrats to four #Belle2 PhD students for winning the Tracking Challenge at this year's Connecting the DotsD Conference! #ctdwit #hackathon

À l'origine en anglais



1

David Rousseau
@dhpmrou

.@SteveAFarrell winner of #CTDWIT TrackMLRamp 2D #hackathon at @LALOrsay in the ML category. Congrats !

À l'origine en anglais

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
- Access to experiment internal data an issue, but there are ways out
- Note : Yandex Data School of Analysis (with ~10 ML scientists) now a bona fide institute of LHCb
- 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!

Open Data

- Public dataset are essential to collaborate (beyond talking over beer/coffee) on new ML techniques with ML experts (or even physicists in other experiments)
  - can share without experiments Non Disclosure policies
- Some collaborations built on just generator data (e.g. Pythia) or with simple detector simulation e.g. Delphes
  - good for a start, but inaccurate
- Effort to have better open simulation engine (e.g. Delphes 4-vector detector simulation, ACTS for tracking)
- UCI dataset repository has some HEP datasets
- Role of CERN Open Data portal:
  - We (ATLAS) initially saw its use for outreach purposes (CMS has been more open on releasing data)
  - But after all, ML collaboration is a kind of scientific outreach
  - ➔ATLAS uploaded there in 2015 the data from Higgs Machine Learning challenge (essentially 4-vectors from full G4 ATLAS simulation Higgs->tautau analysis)
  - ATLAS consider releasing more datasets dedicated to ML studies

Conclusion

- We (in HEP) are analysing data from multi-billion € projects➔should make the most out of it!
- Recent explosion of novel (for HEP) ML techniques, novel applications for Analysis, Reconstruction, Simulation, Trigger, and Computing
- Some of these are ~easy, most are complex: open source software tools are ~easy to get, but still need (people) training, know-how
- More and more open datasets/simulators
- More and more HEP and ML workshops, forums, schools, challenges
- More and more direct collaboration between HEP researchers and ML researchers
- HEP will need more and more access to (GPU) training resources
- Never underestimate the time for :
  - (1) Great ML idea➔
  - (2) …demonstrated on toy dataset➔
  - (3) …demonstrated on real experiment analysis/dataset ➔
  - (4) …experiment publication using the great idea