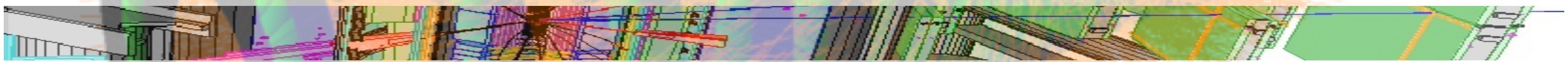
TrackML : Tracking Machine Learning challenge



David Rousseau (LAL-Orsay, U Paris-Saclay)
(rousseau@lal.in2p3.fr),

with Paolo Calafiura, Steven Farrell, Heather Gray (LBNL-Berkeley), Jean-Roch Vlimant (CalTech), Yetkin Yilnaz (LAL), Cécile Germain (LAL/LRI), Isabelle Guyon (ChaLearn, U Paris Saclay), Vincenzo Innocente, Andreas Salzburger (CERN), Tobias Golling, Moritz Kiehn, Sabrina Amrouche (U Geneva), Vava Gligorov (LPNHE-Paris), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex)

Special thanks for the preparation of the slides : Andreas Salzburger, Jean-Roch Vlimant

IN2P3 ML workshop, CC-Lyon, 29th Mar 2018

Who are we ?

Paolo Calafiura, Steven Farrell, Heather Gray (LBNL-Berkeley), Jean-Roch Vlimant (CalTech), Cécile Germain (LAL/LRI U Paris Saclay), Isabelle Guyon (ChaLearn, U Paris Saclay), David Rousseau, Yetkin Yilnaz (LAL Orsay U Paris Saclay), Vincenzo Innocente, Andreas Salzburger (CERN), Tobias Golling, Moritz Kiehn, Sabrina Amrouche (U Geneva), Vava Gligorov (LPNHE-Paris), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex)

- Particle physics tracking experts from three large CERN experiments on the LHC ATLAS, CMS and LHCb
- Machine Learning scientists
- Some of us have organised challenges on Kaggle
  - The Higgs Machine Learning challenge 2014 (proceedings of NIPS 2014 workshop)
  - Flavour of Physics challenge 2015
- We have been preparing this new challenge since 3 years…

Partners

kaggle

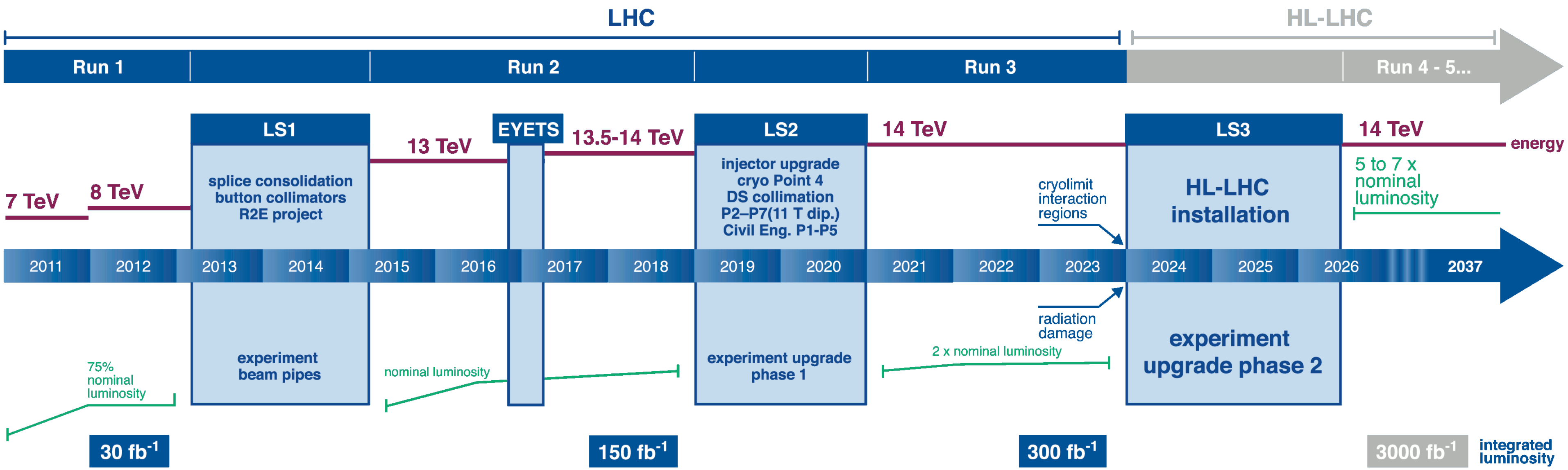
Inria
INVENTEURS DU MONDE NUMÉRIQUE

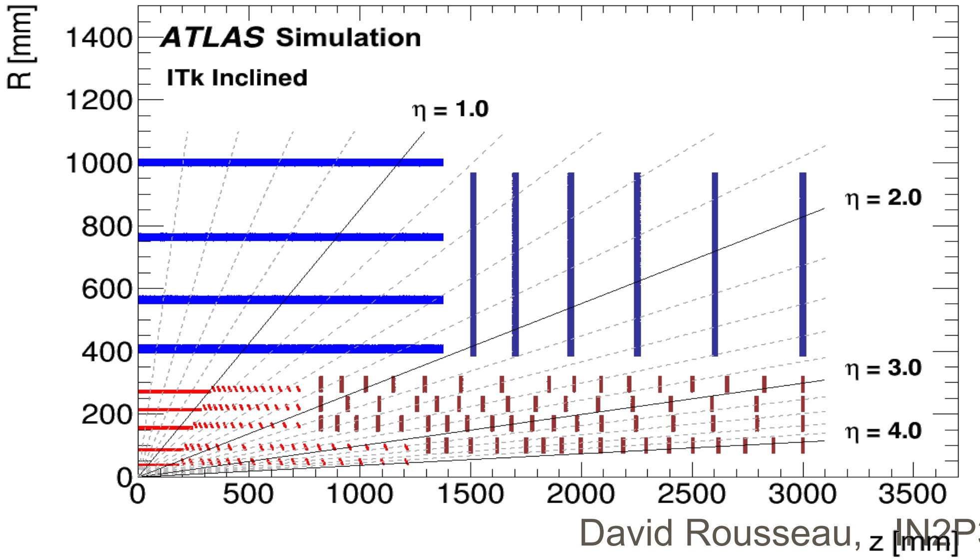
Paris-Saclay
Center for Data Science

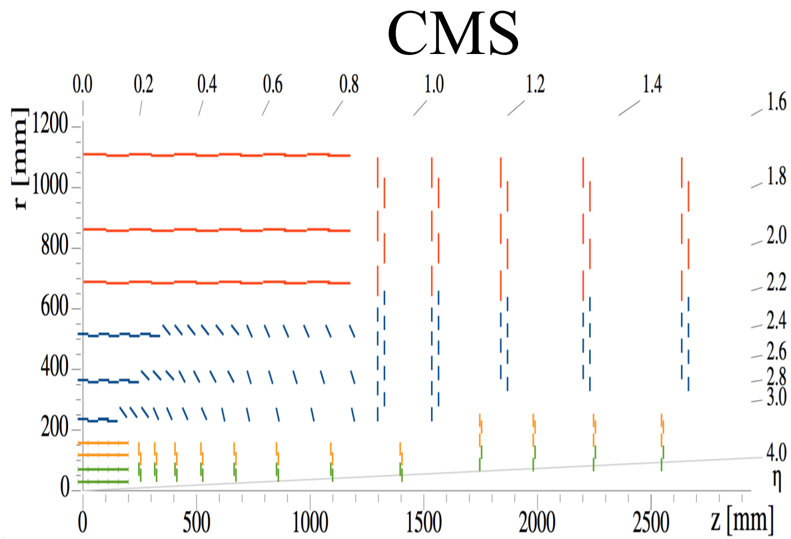
CERN openlab

LHC tracking

HL-LHC upgrade

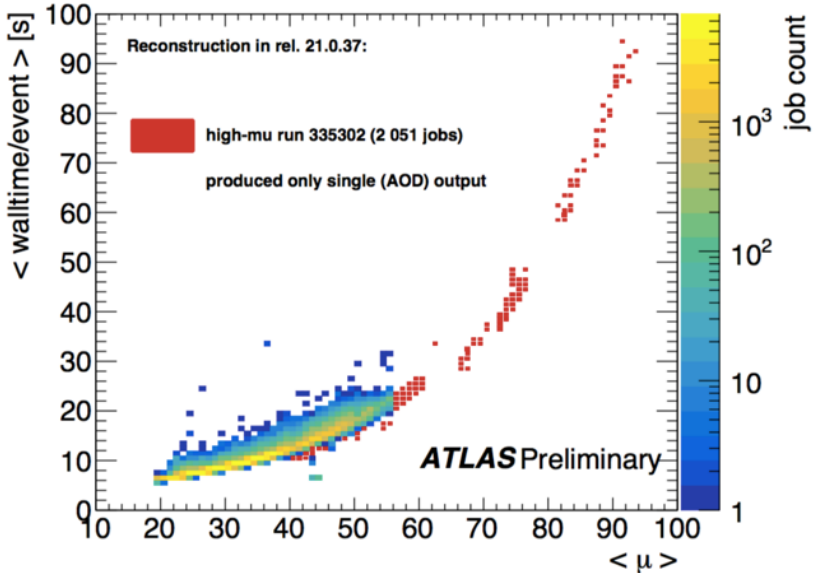


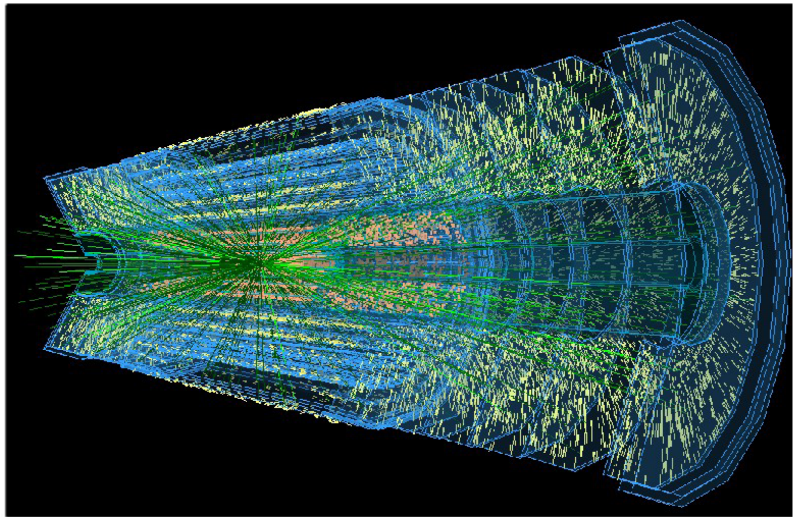


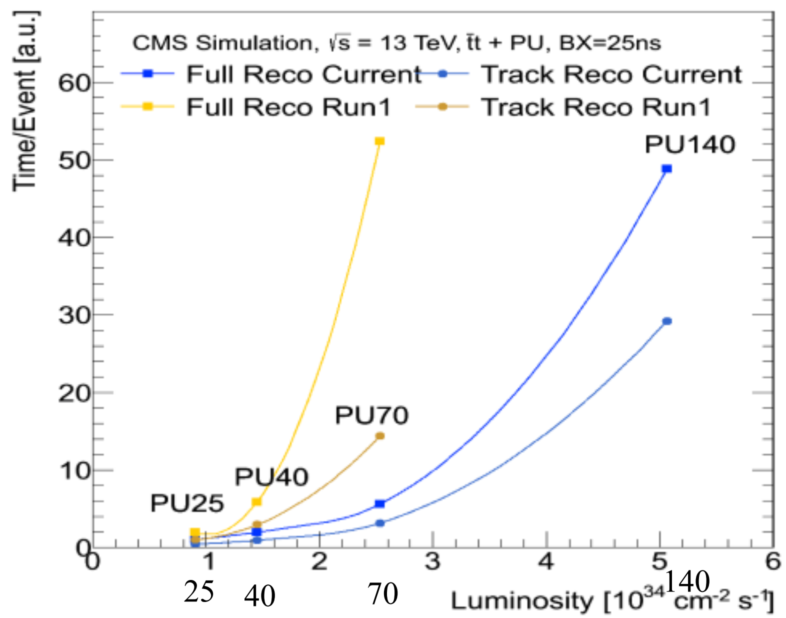


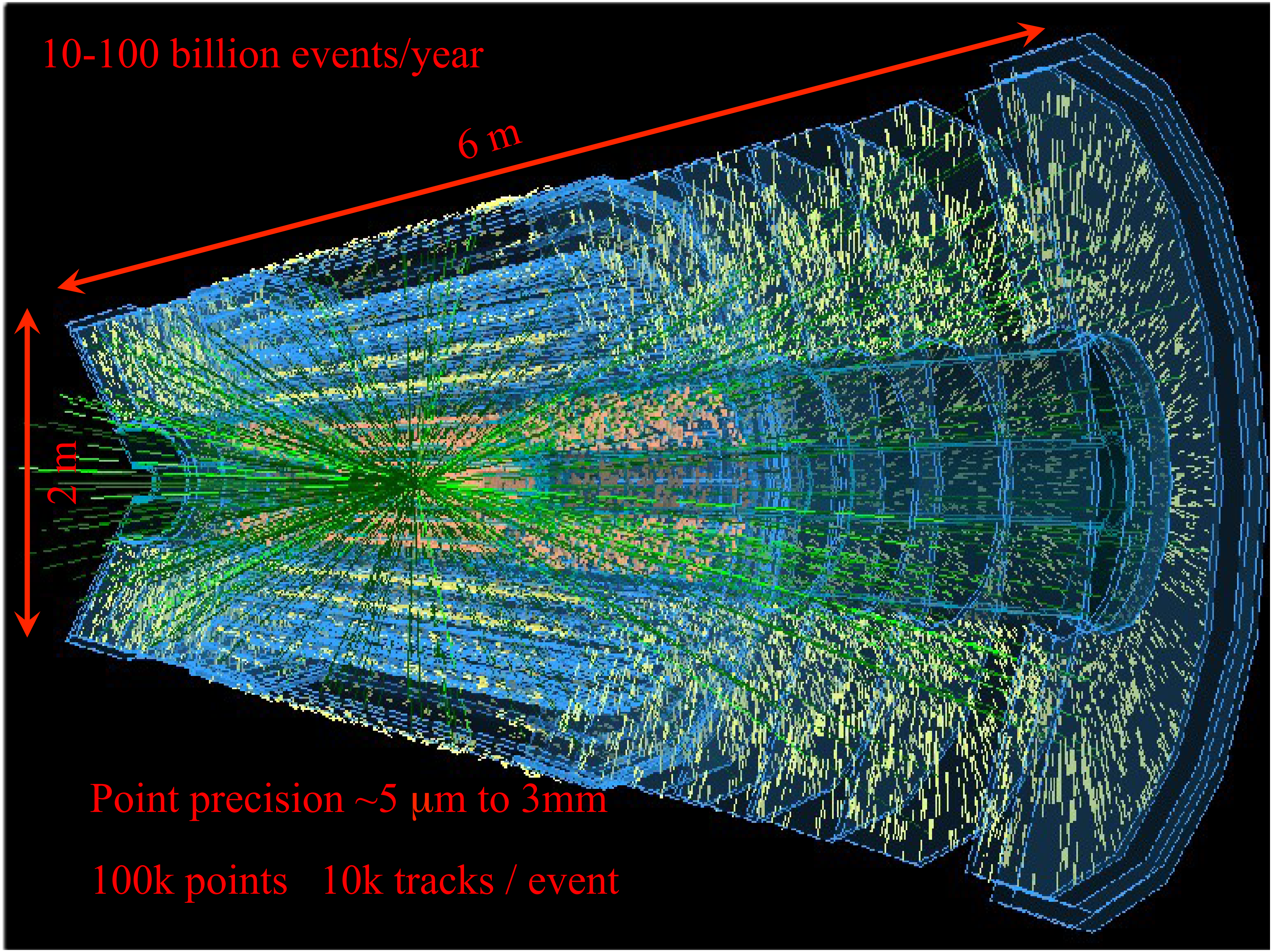
Tracking crisis

- Tracking (in particular pattern recognition) dominates reconstruction CPU time at LHC
- High Luminosity-LHC perspective : increased rate of parasitic collisions from 40 (2017) to 200
- CPU time of current software quadratic/ exponential extrapolation (difficult to quote any number)
- (current software give sufficiently good results in terms of accuracy, but x10 too slow)
- Distant future FCC-hh would reach 1000

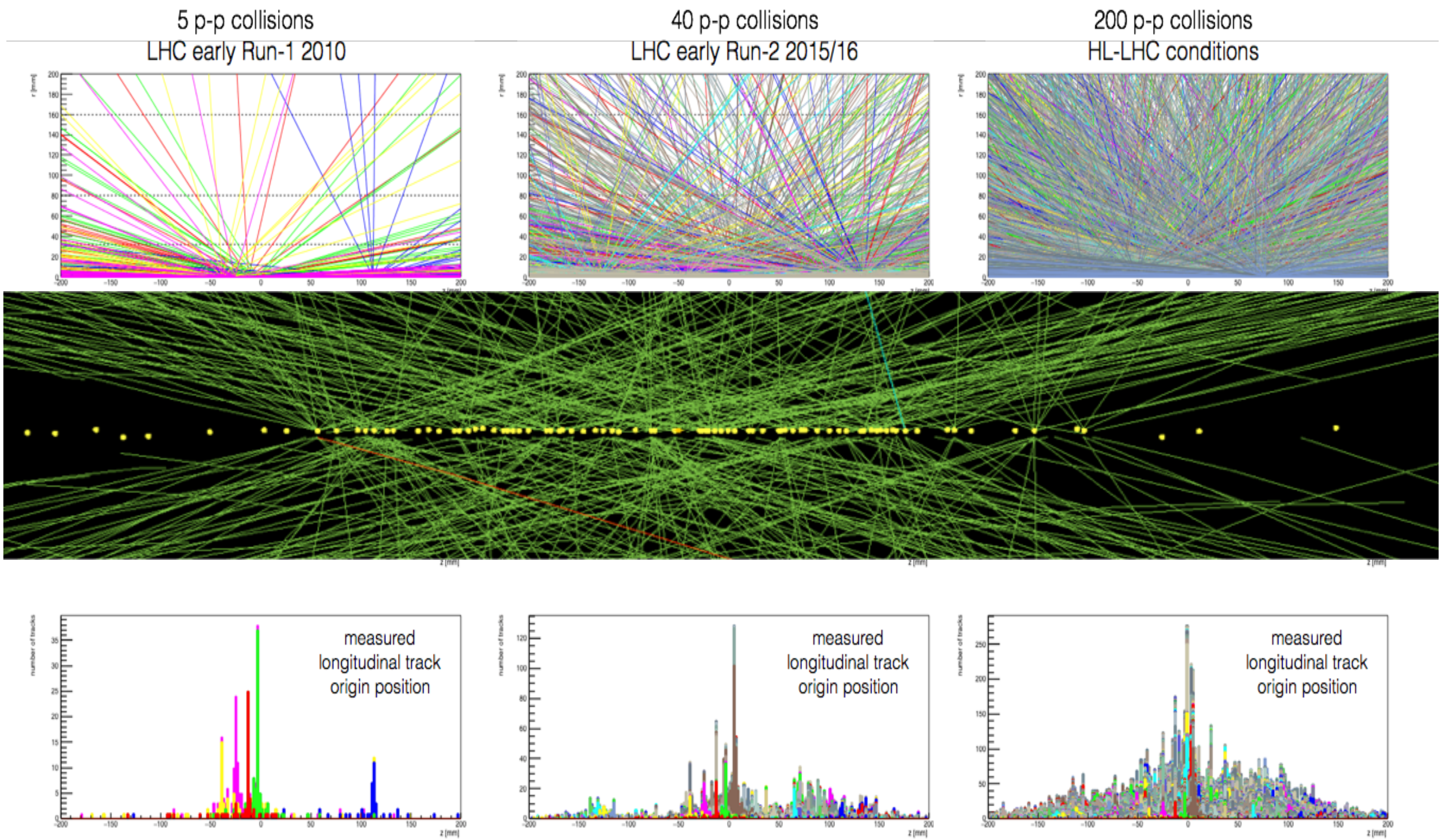







10-100 billion events/year
6 m
2 m
Point precision ~5 μm to 3mm
100k points 10k tracks / event

Pile-up


5 p-p collisions
LHC early Run-1 2010
40 p-p collisions
LHC early Run-2 2015/16
200 p-p collisions
HL-LHC conditions
measured
longitudinal track
origin position
measured
longitudinal track
origin position
measured
longitudinal track
origin position

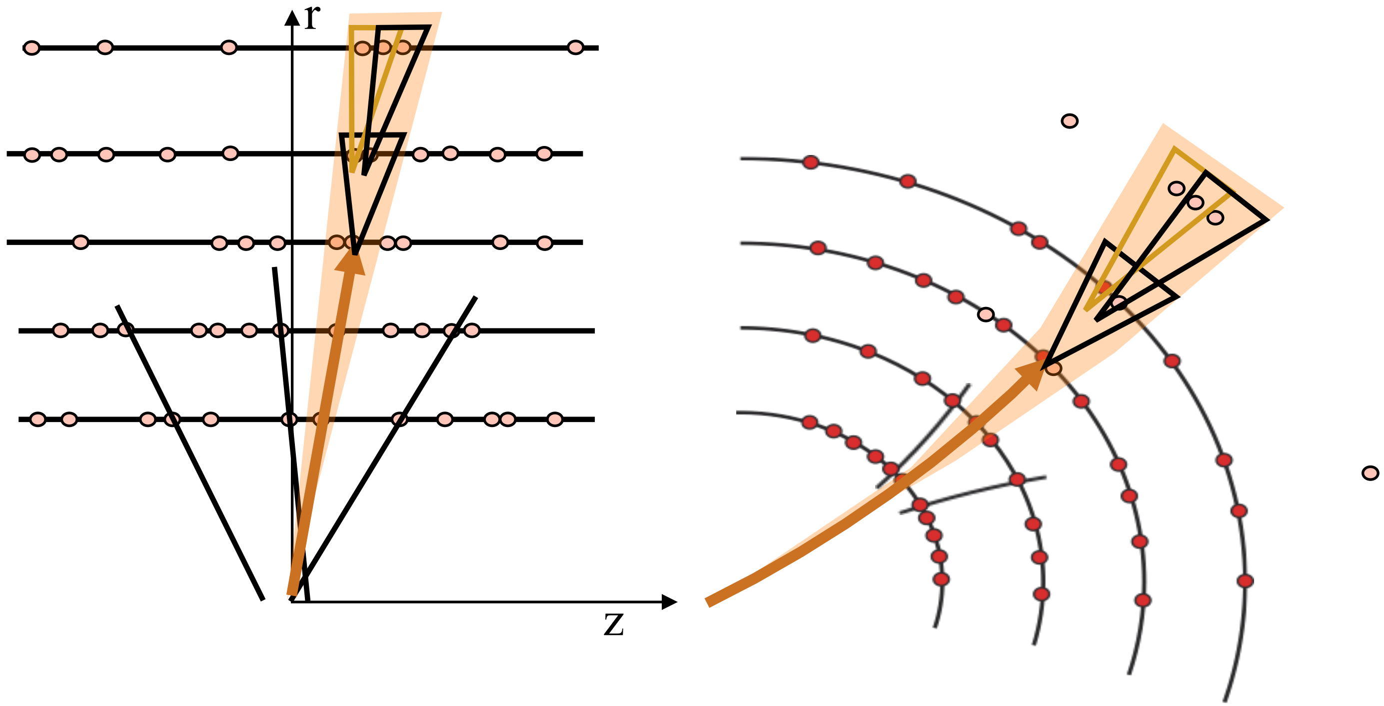
Motivation

- LHC experiments future computing budget flat (at best) (LHC experiments use 300.000 CPU cores on the LHC world wide computing grid)
- Installed CPU power per $==€==CHF expected increase factor <10 in 2025
- Experiments plan on increase of amount of data recorded (by a factor ~10)
- ➔HighLumi reconstruction to be as fast as current reconstruction despite factor 10 in complexity
- ➔requires very significant software CPU improvement, factor ~10
- Large effort to optimise current software and tackle micro and macro parallelism
  - Also development of dedicated hardware for fast tracking
- >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
- Need to engage a wide community to tackle this problem

Particle Tracking algorithms

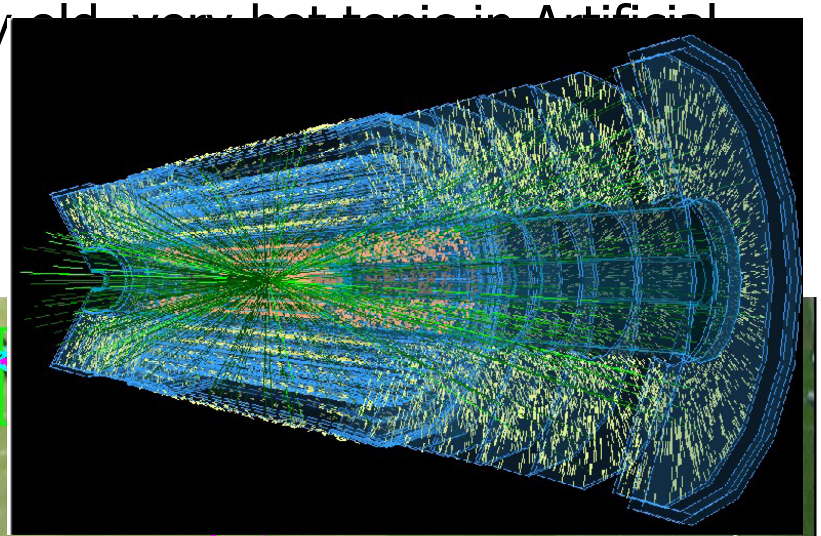
Current Algorithms

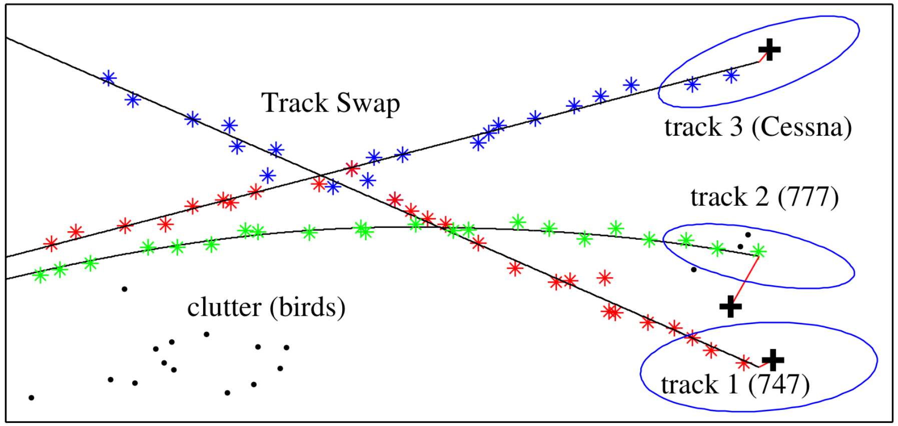
- Pattern : connect 3D points into tracks
- Essentially combinatorial approach
- Tracks are (not perfect) helices pointing (approximately) to the origin
- Challenge : explore completely new approaches
- (not part of the challenge : given the points, estimate the track parameters)



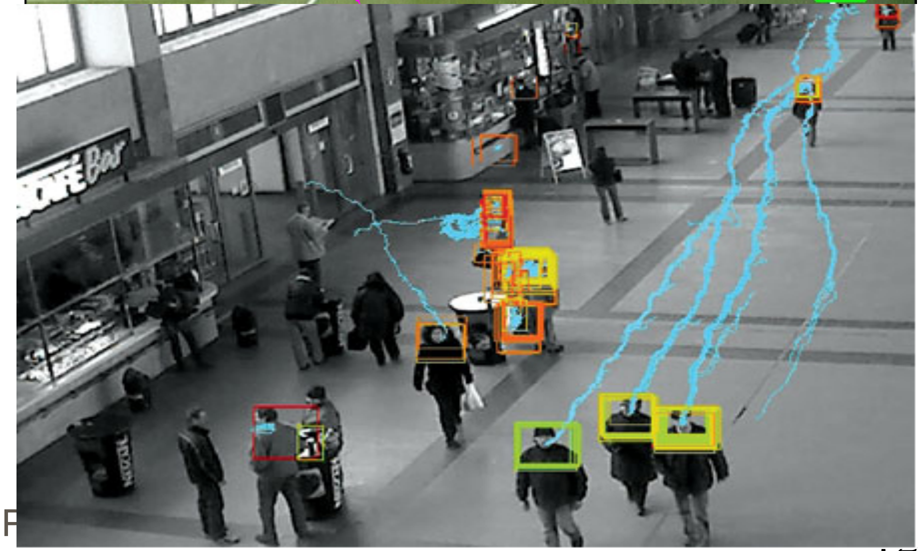
Pattern recognition in ML

- Pattern recognition, tracking, is a very
Intelligence : examples→



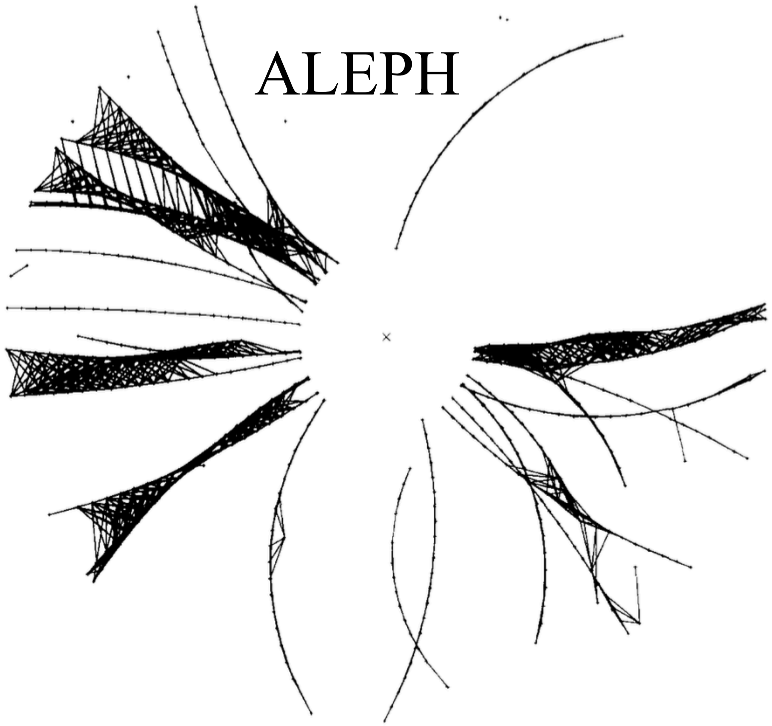


http://papers.nips.cc/paper/5572-a-complete-variational-tracker.pdf



- Note that these are real-time applications, with CPU constraints
- Worry about efficiency, "track swap",…
- But no on-the-shelf algorithm will solve our problem
- (in fact a few lines calling DBScan in sk-learn does find some tracks)

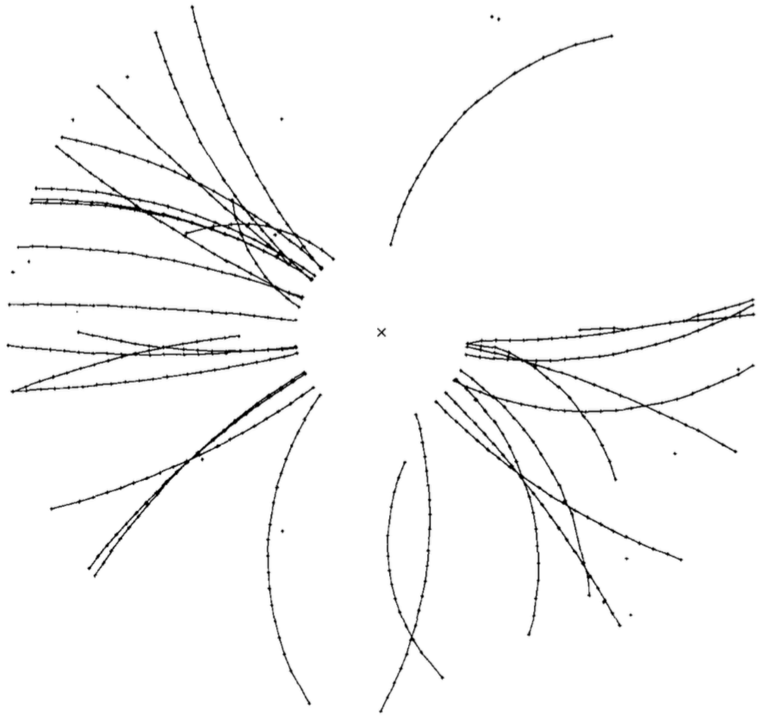
An early attempt



UT
PHYSIC

ice

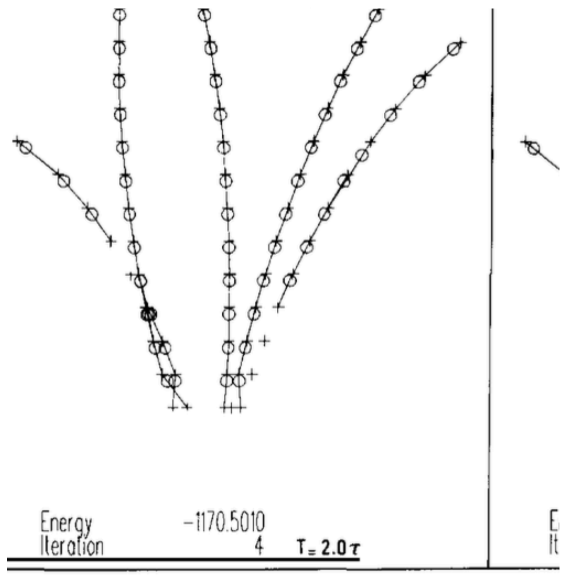
December 1



e Learnir

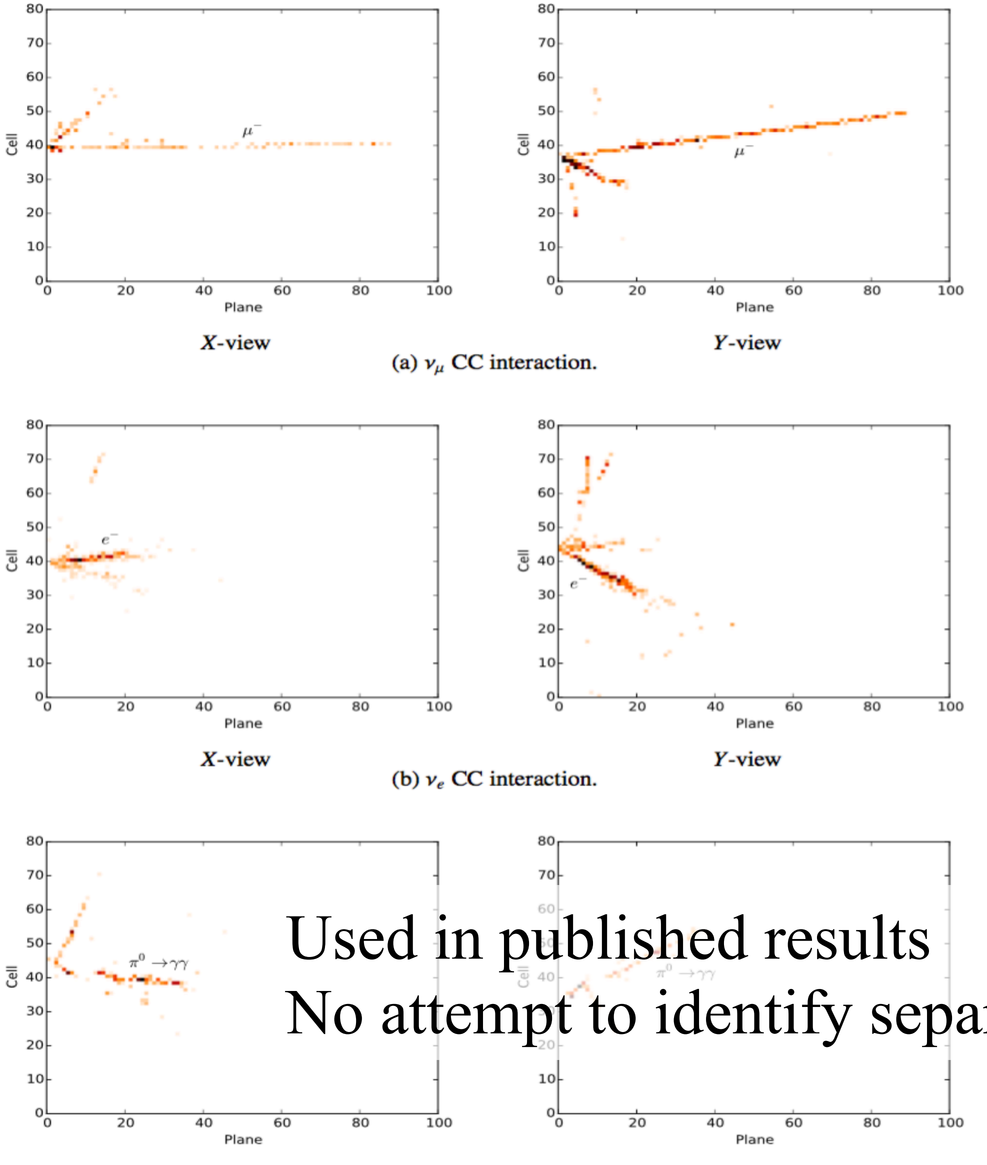
known

- Losely inspired from Traveling Salesman Problem with NN by Hopfield & Tank Biological Cybernetics 52 (1985) 141. or with Minimal Tree Span Cassel & Kowalski Nucl Inst; and Meth 185 (1981) 235
- (large litterature since, e.g. Neural Combinatorial Optimization with reinforcement learning, Bello et al Google Brain 1611.0994)
- Full implementation in ALEPH Stimpfl & Garrido (1990) Computer Physics Comm. 64 (1991) 46.
- However never deployed



A recent attempt : NOVA

arXiv 1604.01444 Aurisano et al



X-view Y-view
(a) ν_μ CC interaction.

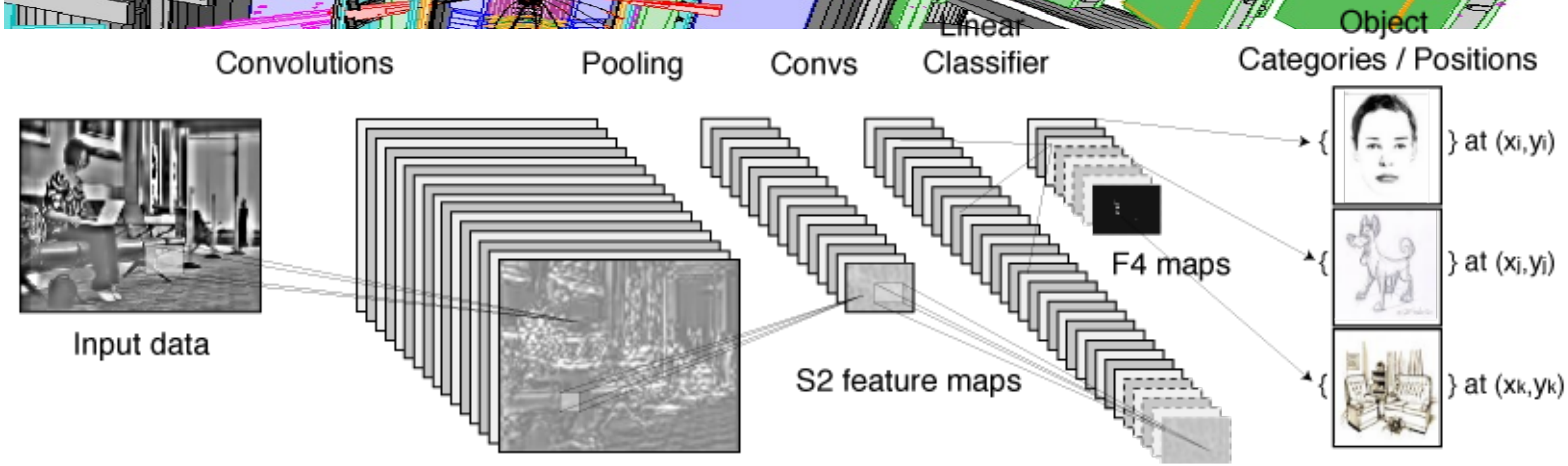
X-view Y-view
(b) ν_e CC interaction.

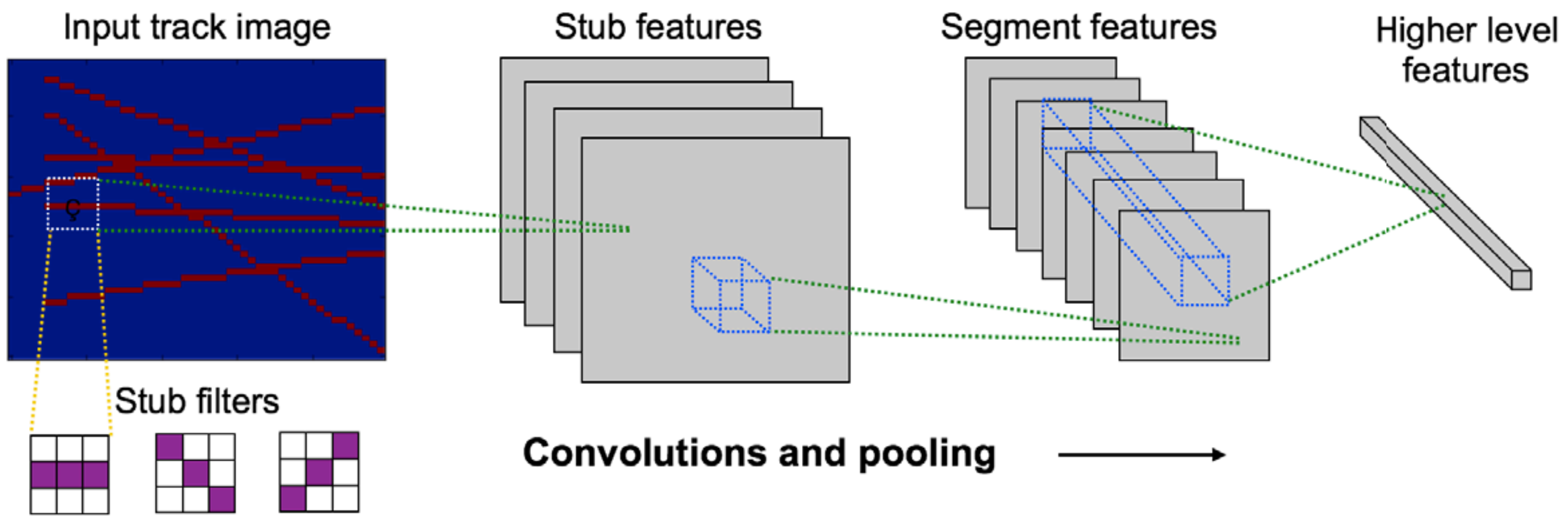
X-view Y-view
(c) NC interaction.

Neutrino interaction classification
Using Convolutionnal Neural Network
No attempt to separate individual tracks.

Used in published results
No attempt to identify separate tracks

Convolution NN

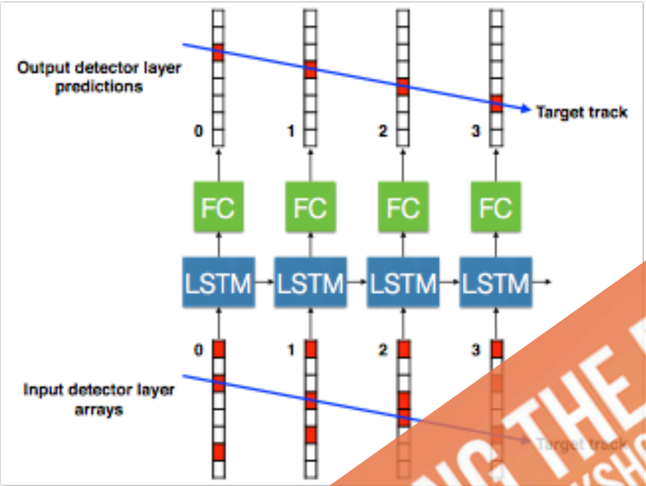


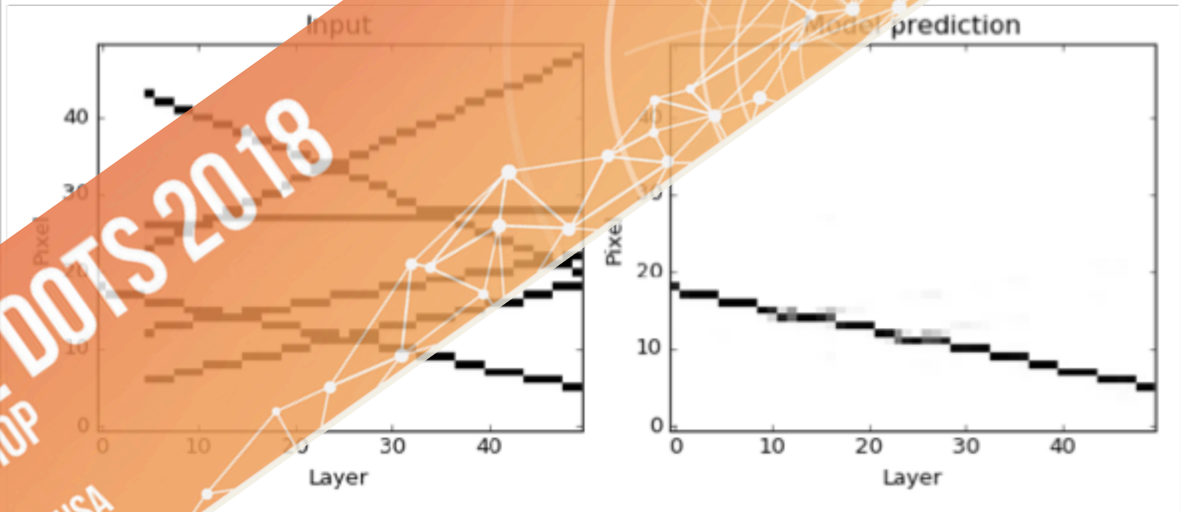


See:
Farrel S. et al, The HEP.TrkX Project: deep neural networks for HL-LHC online and offline tracking, EPJ Web of Conferences 150, 00003 (2017)

RNN

Long Short Term Memory (LSTM)





See:
Farrel S. et al, The HEP.TrkX Project: deep neural networks for HL-LHC online and offline tracking, EPJ Web of Conferences 150, 00003 (2017)

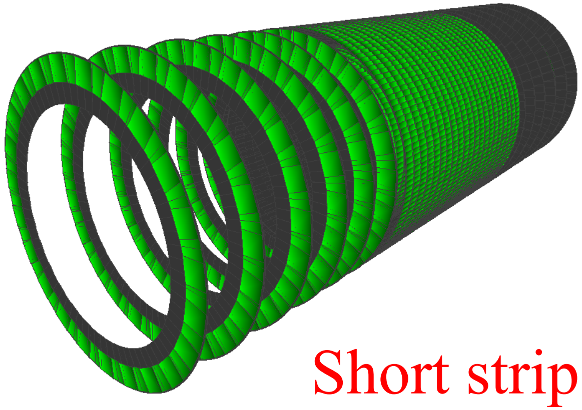
The tracking challenge

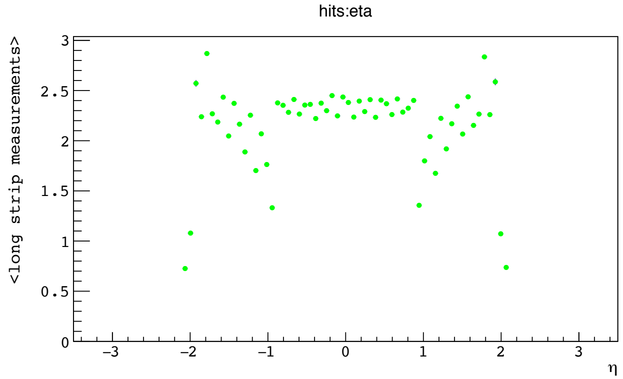
In a nutshell

- Accurate simulation engine (ACTS https://gitlab.cern.ch/acts/acts-core) to produce realistic events
  - One file with list of 3D points
  - Ground truth : one file with point to particle association
  - Ground truth auxiliary : true particle parameter (origin, direction, curvature)
  - Typical events with ~200 parasitic collisions (~10.000 tracks/event)
- Large training sample 100k events, 10 billion tracks ~100GByte
- Participants are given the test sample (with usual split for public and private leaderboard) and run the evaluation to find the tracks
- They should upload the tracks they have found
  - A track is a list of 3D points
  - (do not consider estimation of particle parameter)
  - Score : fraction of points correctly grouped together
  - Evaluation on test sample with per-mille precision on 100 event

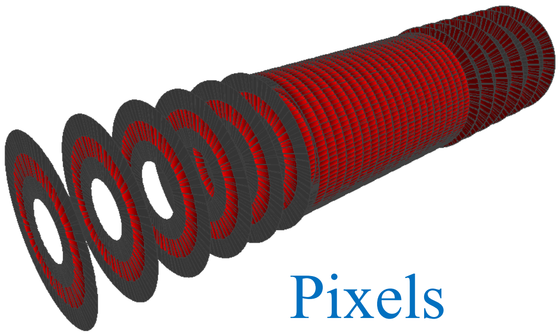
Detector : layout

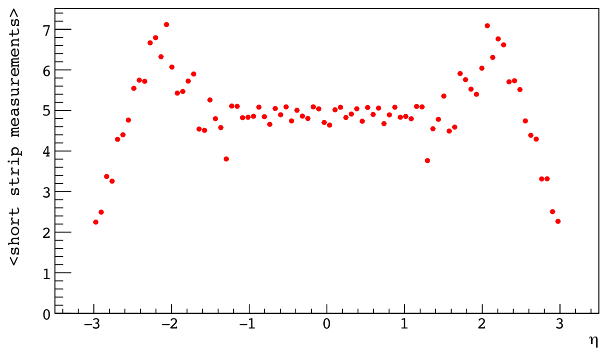
Long strips





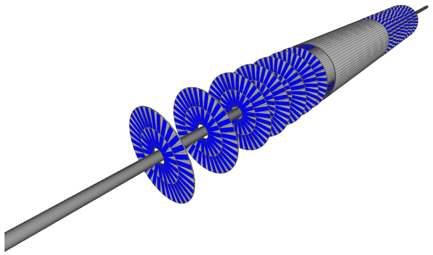
Short strips

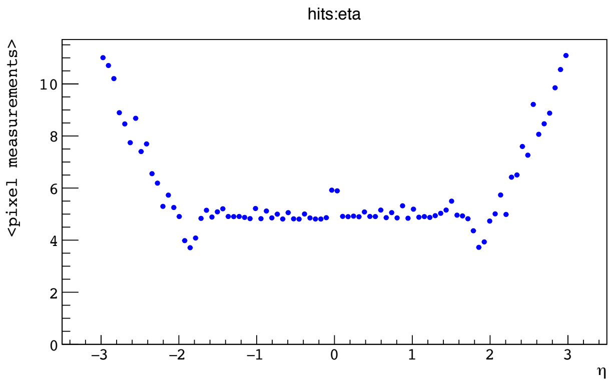




~12 points per tracks

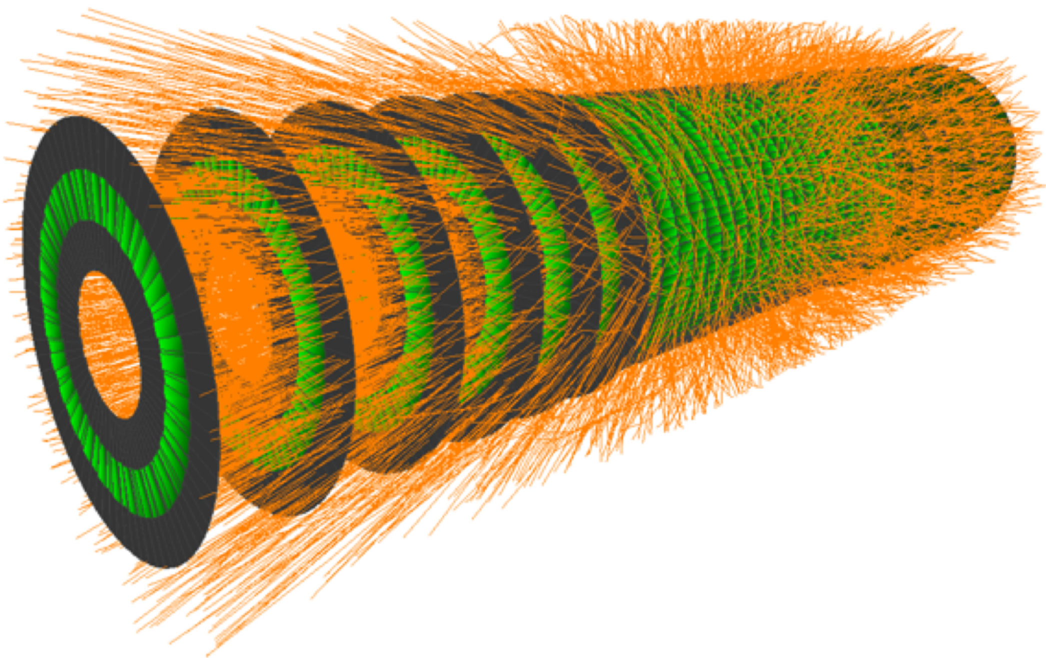
Pixels

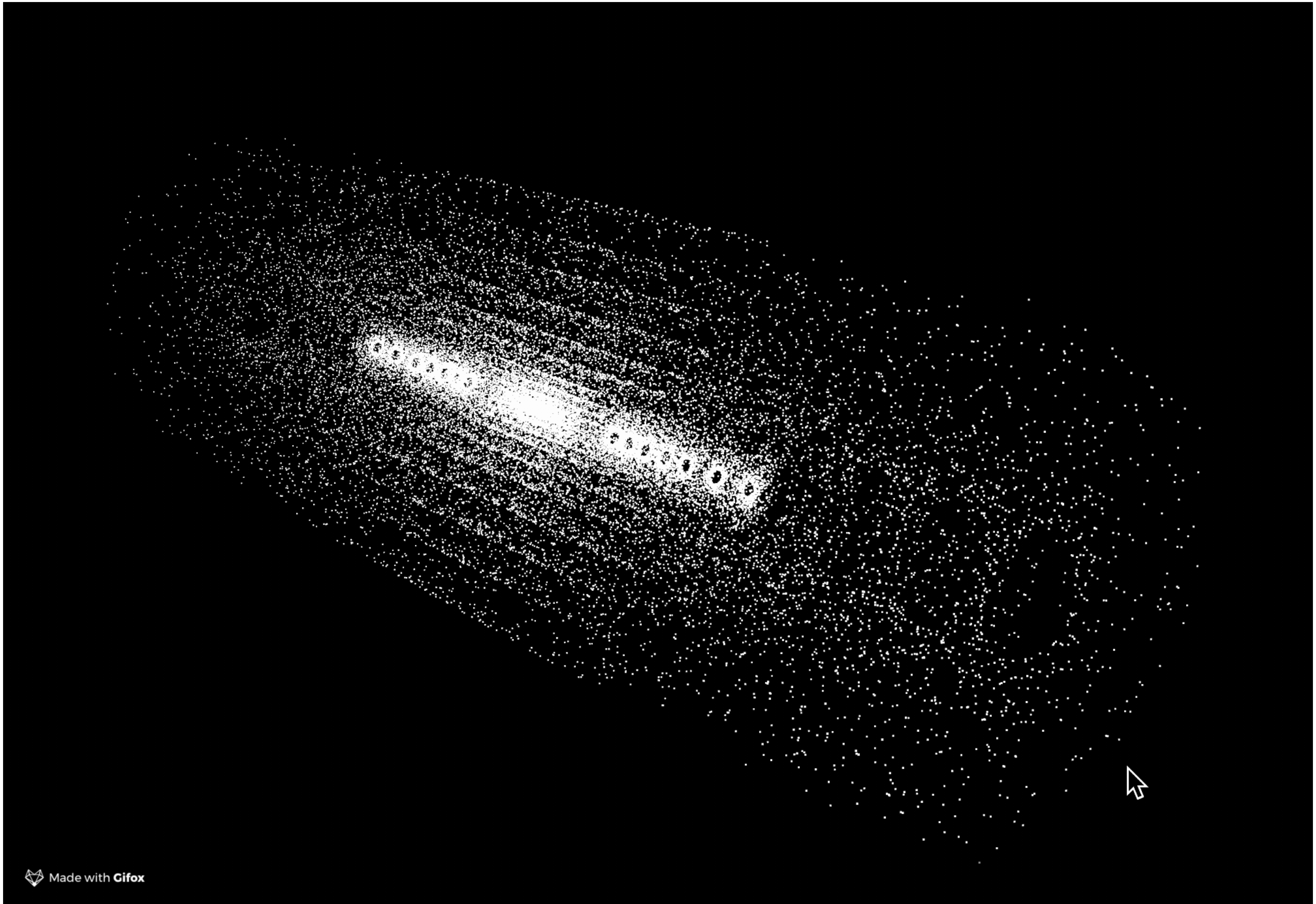




Event simulation

- Typical LHC event simulated
  - Pythia tt-bar event
  - Overlaid with Poisson(200) Pythia minimum bias
  - ~10'000 tracks
- Most tracks are coming from a central region: gaussian σ_z=5.5 cm, transverse σ=15μm, some from a larger cylinder
- 15% of random hits
- Trajectories are deterministic, except for Multiple Scattering, Energy Loss and hadronic interaction




Made with Gifox

Datasets

❑ Hit file

(measured position mm) (pixel location and charge)

	hit_id	volume_id	layer_id	module_id	x	y	z	ncells	pixels
0	1	7	2	1	-63.9659	-3.70513	-1502.5	1	[[141, 605, 0.297491]]
1	2	7	2	1	-40.2738	2.82386	-1502.5	1	[[48, 176, 0.291861]]
2	3	7	2	1	-88.1049	-11.72380	-1502.5	1	[[263, 1044, 0.327308]]
3	4	7	2	1	-39.7041	-8.71702	-1502.5	1	[[279, 182, 0.327097]]
4	5	7	2	1	-30.4918	-8.19262	-1502.5	1	[[283, 18, 0.258165]]

❑ Truth file

(true position mm particle momentum GeV)

	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	585626006354657280	-63.972698	-3.72889	-1502.5	-0.342366	-0.001899	-7.83544	0.018565
1	2	103582997587951616	-40.287201	2.84328	-1502.5	-0.366049	0.013878	-13.55470	0.035088
2	3	108088040324333568	-88.089600	-11.72360	-1502.5	-0.550128	-0.041929	-9.22279	0.018542
3	4	108090926542356480	-39.712601	-8.71581	-1502.5	-0.363936	-0.094646	-14.01150	0.035088
4	5	108103502206599168	-30.470400	-8.18647	-1502.5	-0.413489	-0.123403	-20.65790	0.000000

Datasets

❑ Particle file

	particle_id	vx	vy	vz	px	py	pz	q
		origin vertex (mm)			momentum (GeV)			charge
0	4503805785800704	-0.021389	-0.012618	-0.624757	38.907001	-16.146099	-84.311096	-1
1	4504011944230912	-0.021389	-0.012618	-0.624757	-0.661993	0.118267	249.181000	1
2	4504080663707648	-0.021389	-0.012618	-0.624757	0.821614	0.954217	0.948994	-1
3	4504149383184384	-0.021389	-0.012618	-0.624757	0.300791	0.080450	2.656530	1
4	4504218102661120	-0.021389	-0.012618	-0.624757	-0.552250	-0.481988	-0.888733	1

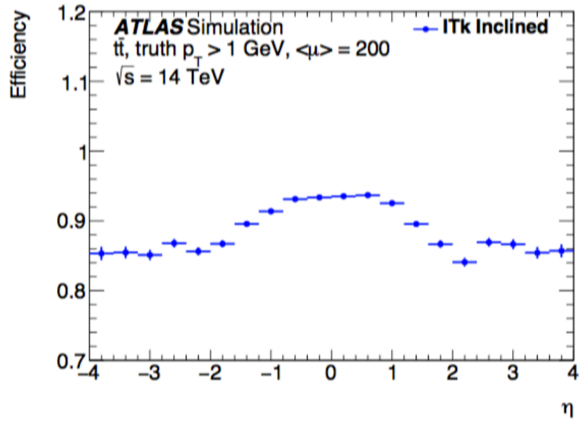
(note : we do not ask participant to reconstruct these track parameters but these could be useful latent variables)

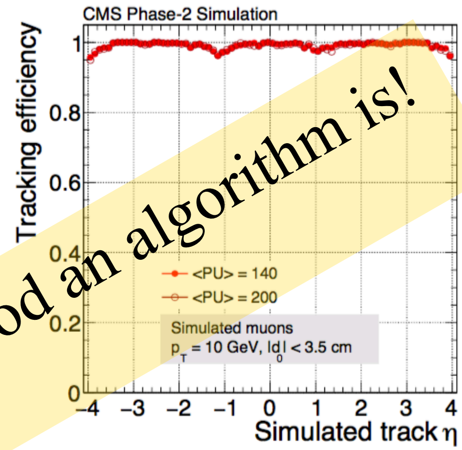
❑ (static)Detector file

	volume_id	layer_id	module_id	cx	cy	cz	rot_xu	rot_xv	rot_xw	ro
				center position (mm)			3x3 rotation matrix			
0	6	2	1	-65.7965	-5.17830	-1502.5	0.078459	-0.996917	0.0	-0.99
1	6	2	2	-139.8510	-6.46568	-1502.0	0.046183	-0.998933	0.0	-0.99
2	6	2	3	-138.6570	-19.34190	-1498.0	0.138156	-0.990410	0.0	-0.99
3	6	2	4	-64.1764	-15.40740	-1498.0	0.233445	-0.972370	0.0	-0.97

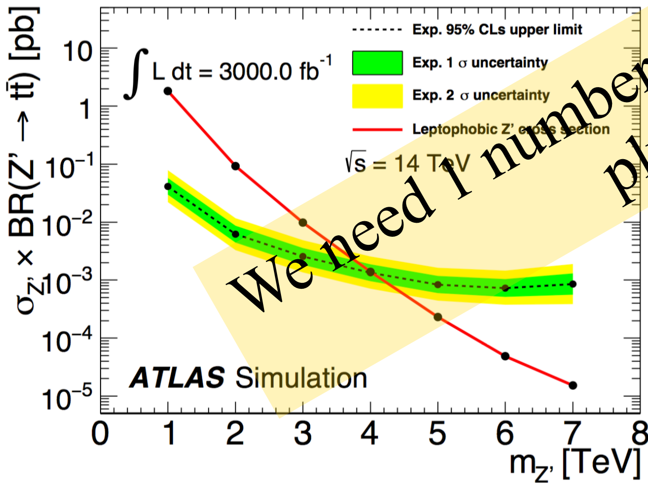
Score

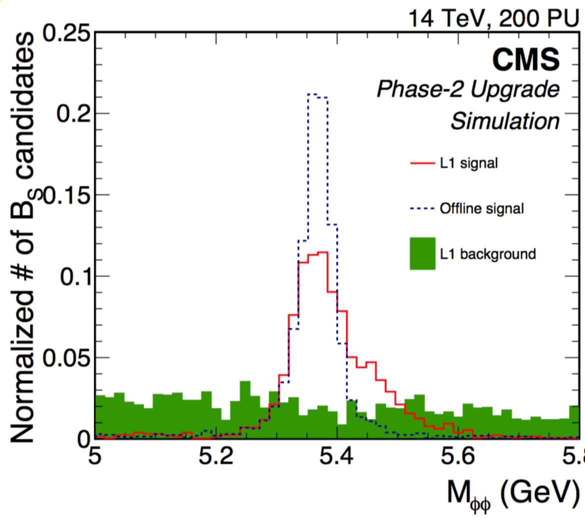
- [2017 CMS tracker Technical Design Report](#) : Chapter 6 expected performance 31 pages 58 figures
- [ATLAS Si strip Technical Design Report](#) Chapter 4 ITk Performance and Physics Benchmark Studies 54 page…





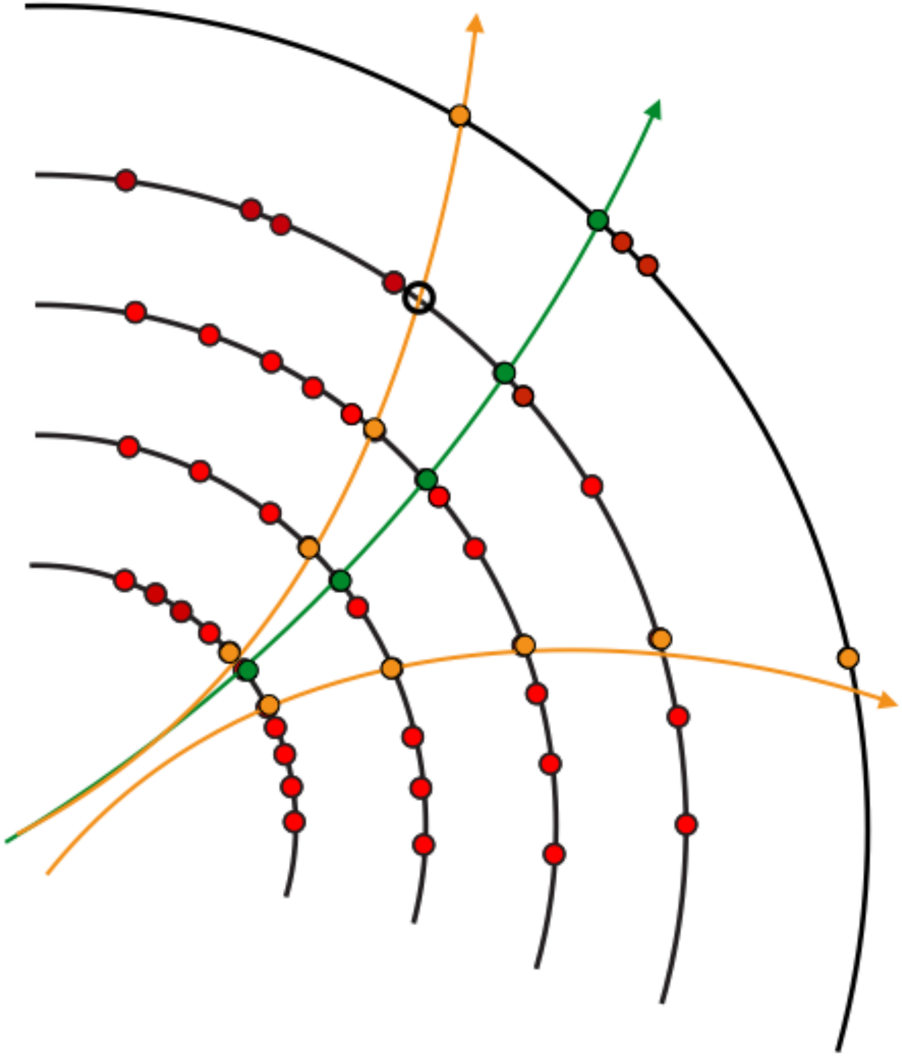
We need 1 number to specify how good an algorithm is!
plus CPU time





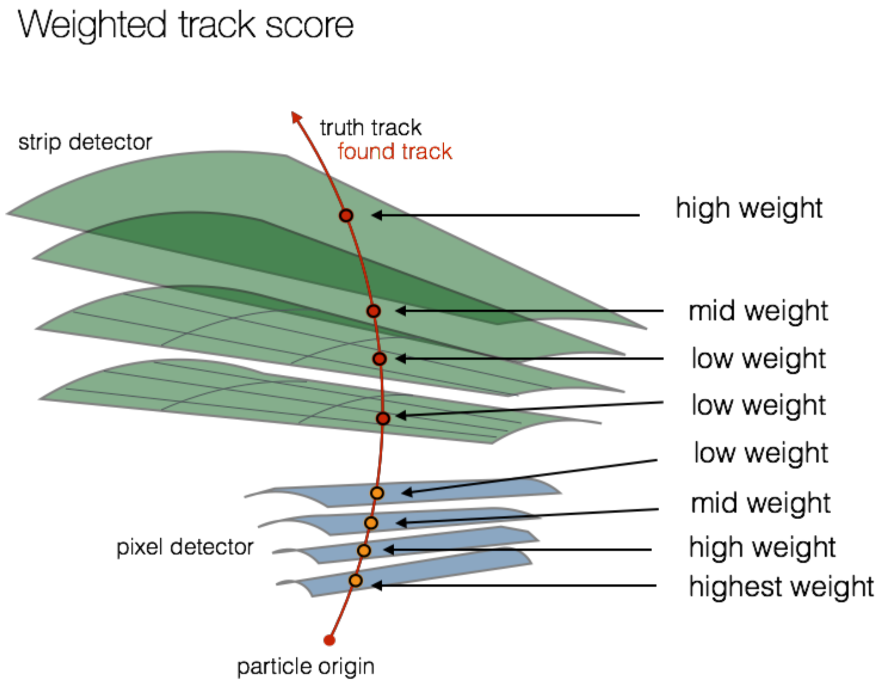
Track evaluation

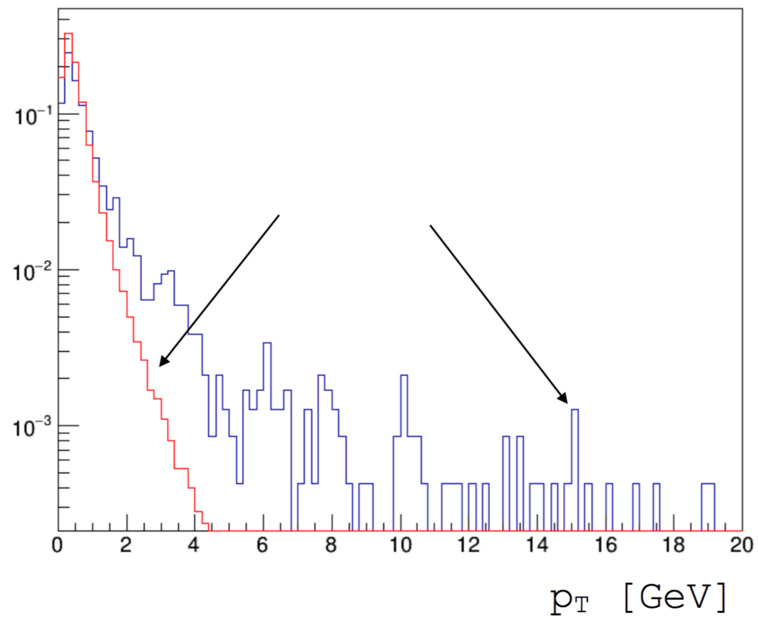
good track	not so good track
many compatible hits	short tracks
completeness	holes
uniqueness	shared hits
low χ^2/ndf	bad fit quality, outliers
small impact parameter (for primaries)	
clusters are compatible	



Hit weighting

- Define : weight=$weight_{order}$ x $weight_{pt}$



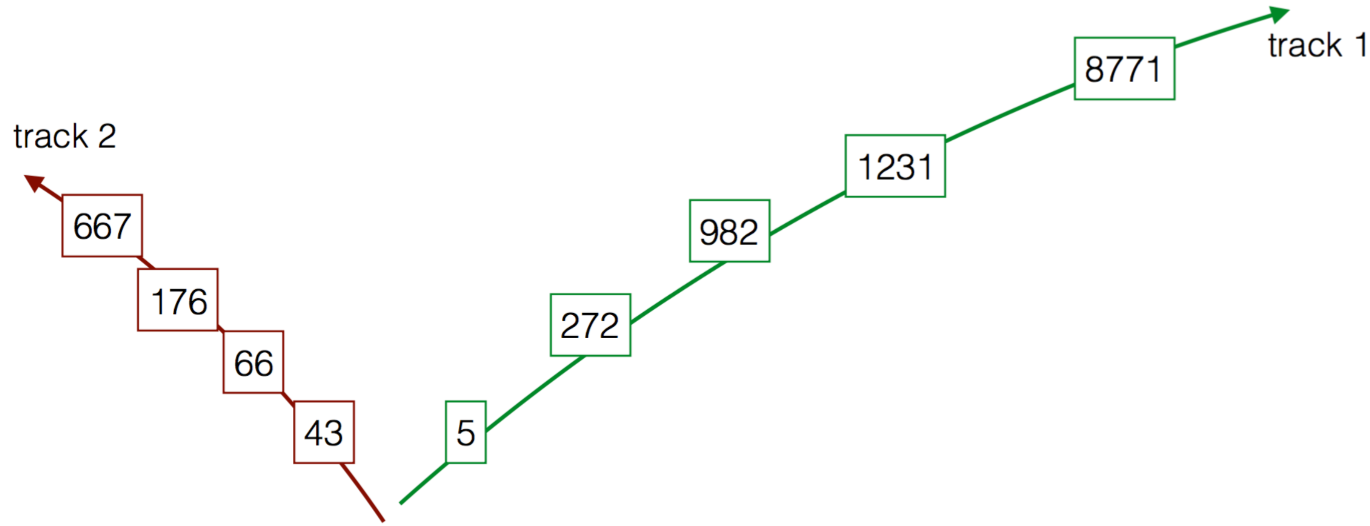


- $Weight_{order}$: more emphasis on first and last hits
- $Weight_{pt}$: more emphasis on high pT tracks
- Weight=0 for noise hits or hits from particle with <=3 hits

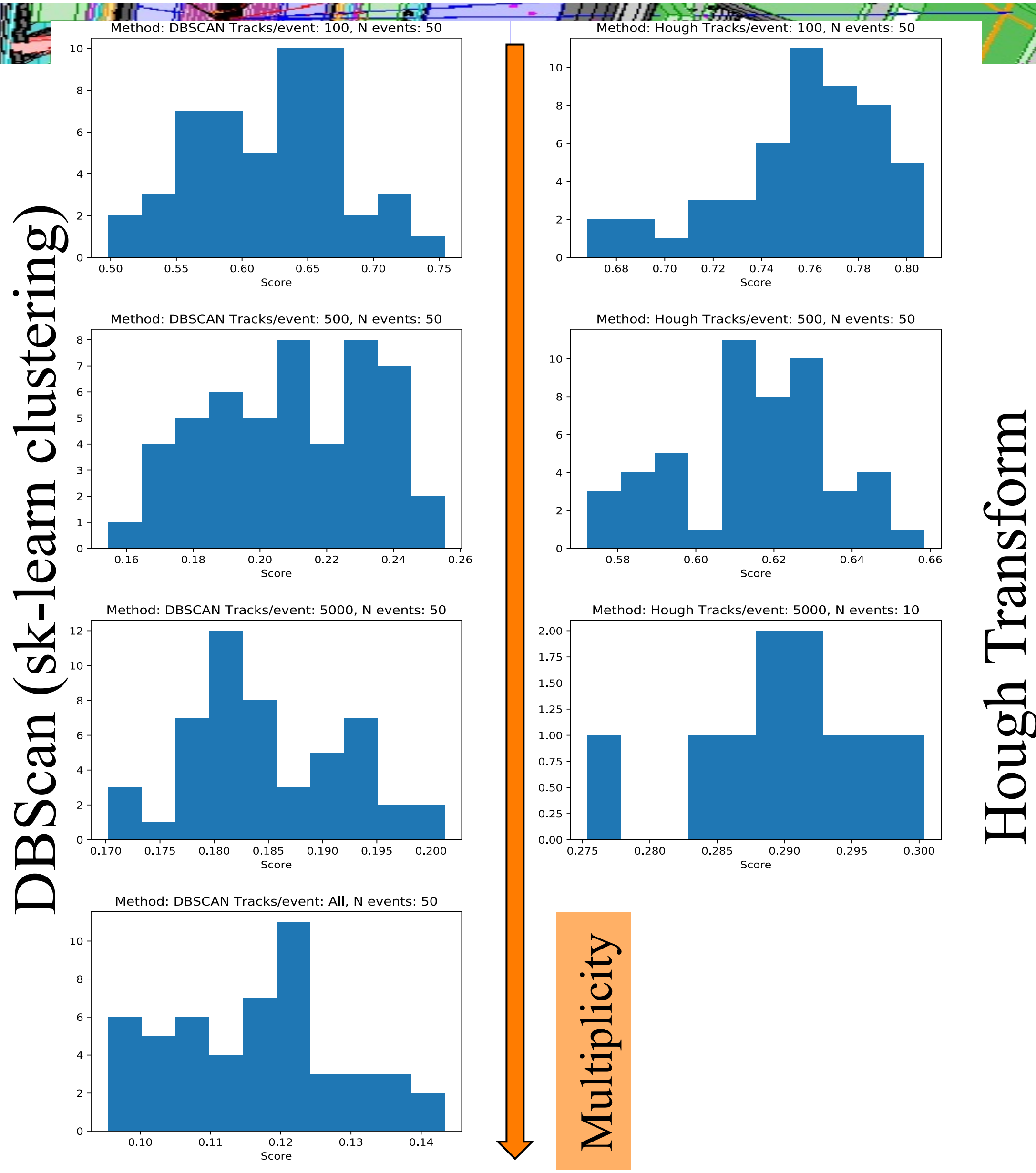
Track scoring

- Overall scoring defined at hit level
- Loop on reco tracks
  - Require >50% of hits from same true particle
  - Require >50% of hits from this true particle in this reco track
  - At this point 1⇔1 relationship between true and reco tracks
  - Sum the weights of the intersection (hits belonging both to true and reco track)
- Event score normalised to the sum of weights of all the hits
  - ➔ ideal algorithm has score==1.
- Final score averaged of 100 events➔statistical precision ~0.1%

hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2



Attempt with 2 simple algs



Real life vs challenge

Real life

1. Wide type of physics events
2. Full detailed Geant 4 / data
3. Detailed dead matter description
4. Complex geometry (tilted modules, double layers, misalignments…)
5. Hit merging
6. Allow shared hits
7. Output is hit clustering, track parameter and covariance matrix
8. Multiple metrics (see TDR's)

Challenge

1. One event type (ttbar)
2. ACTS (MS, energy loss, hadronic interaction, solenoidal magnetic field, inefficiency)
3. Cylinders and slabs
4. Simple, ideal, geometry (cylinders and disks)
5. No hit merging
6. Disallow shared hits
7. Output is hit clustering
8. Single number metrics

Simpler, but not too simple!

Challenge phases

- We will run in two phases
  - Accuracy Phase : focus only on accuracy, no CPU incentive
    - Goal is to expose innovative algorithms
    - Training time unlimited
    - Evaluation time unlimited
    - To run April-June 2018 on Kaggle
  - Throughput Phase: focus on CPU, preserving accuracy
    - Goal is to expose the fastest algorithms
    - Training time (still) unlimited
    - Require the challenge platform to run the algorithm evaluation within fully reproducible controlled environment (VM with x86 processor with 2GB memory, but do not exclude a GPU track in addition)
    - To run in July-October 2018 (NOT on Kaggle)
    - Official NIPS 2018 competition
- Prizes :
  - From leaderboards of first phase: 8k$ 5k$ 2k$ (from Kaggle)
  - From jury examining the algorithms: what are the more likely to be beneficial to HEP ? Invitation to NIPS workshop (if confirmed) and to CERN workshop
  - (Looking for more sponsors, academic or private)

Conclusion

- Setting up TrackML : a particle tracking challenge
- Goal is to involve ML community in overhauling core algorithms of CERN LHC experiments.
  - Looking for new approaches rather than hyper-optimised (HEP) approaches
- Very large training dataset ~100GB
  - Will be released (CERN Open Data portal most likely) after the challenge
- Wealth of possible ML techniques (NN, CNN, RNN, Reinforcement learning, clustering techniques, MCTS…) ... which makes it all the more interesting
- Separate Accuracy phase (most accurate algorithm) and Throughput phase (fastest algorithm to reach similar accuracy)
- Sponsorship more or less OK for Accuracy Phase, still looking for ~40k€ for Throughput phase
- Contact : trackml.contact@gmail.com
- More details, news, etc… : https://sites.google.com/site/trackmlparticle/ , twitter @trackmllhc
- We've beeing accepted as a NIPS 2018 competition (Throughput phase)