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

CERN seminar, 7th Mar 2018

Outline

- Why a Tracking challenge now ?
- HiggsML challenge recap
- Simulation
- Metric
- Conclusion

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

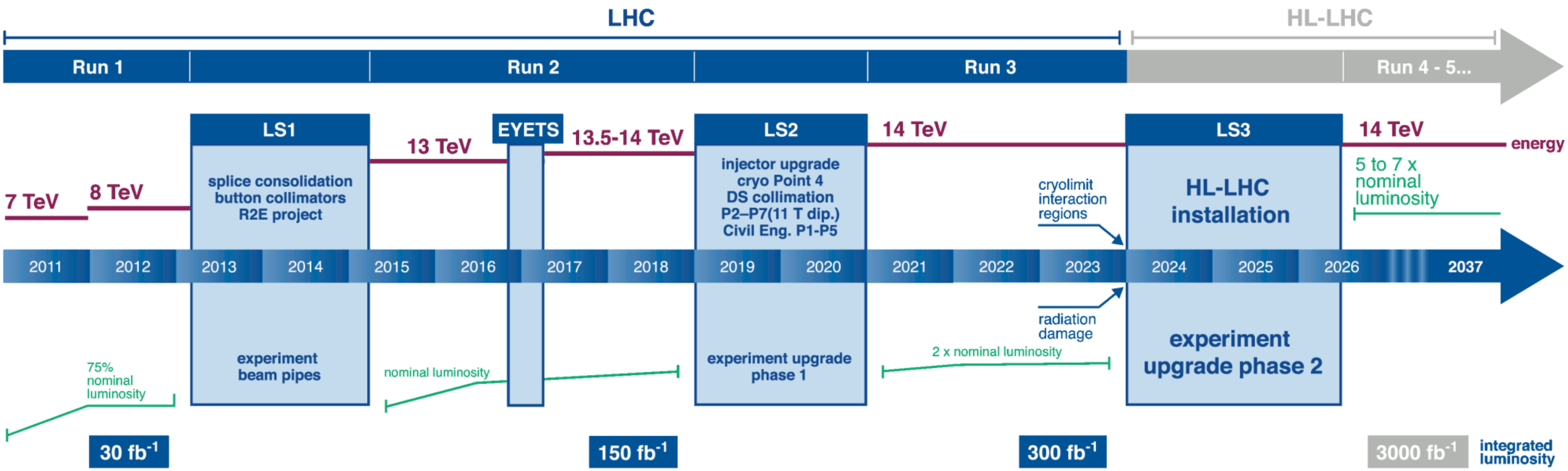
Inria
INVENTEURS DU MONDE NUMÉRIQUE

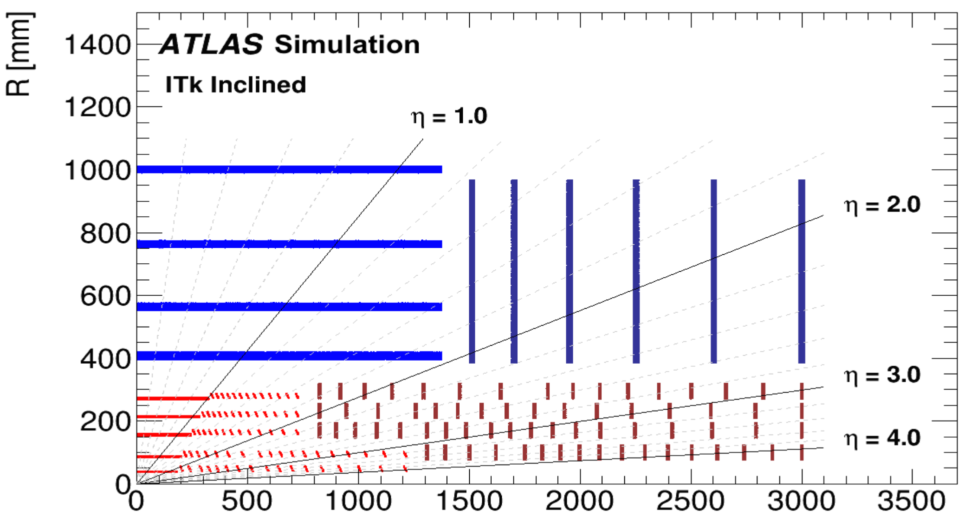
Paris-Saclay
Center for Data Science

CERN
openlab

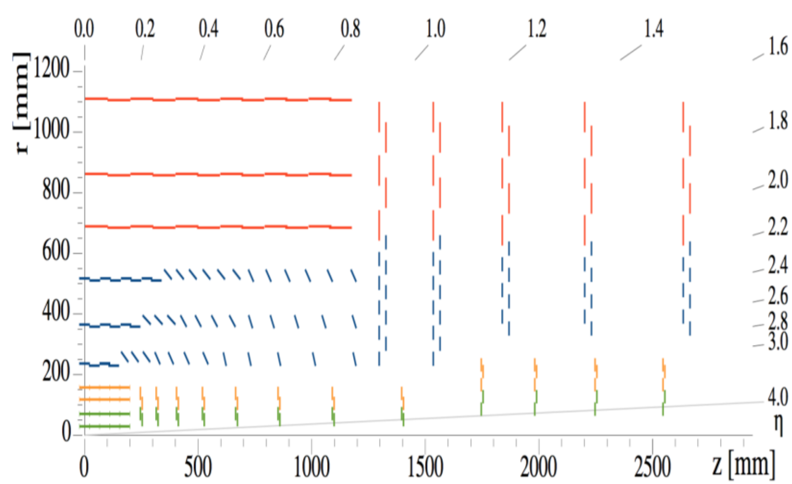
LHC tracking

HL-LHC upgrade



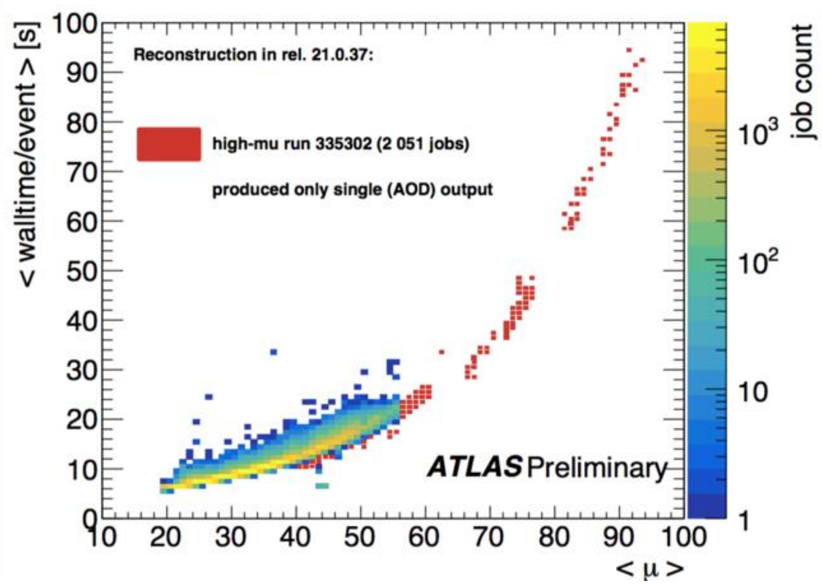


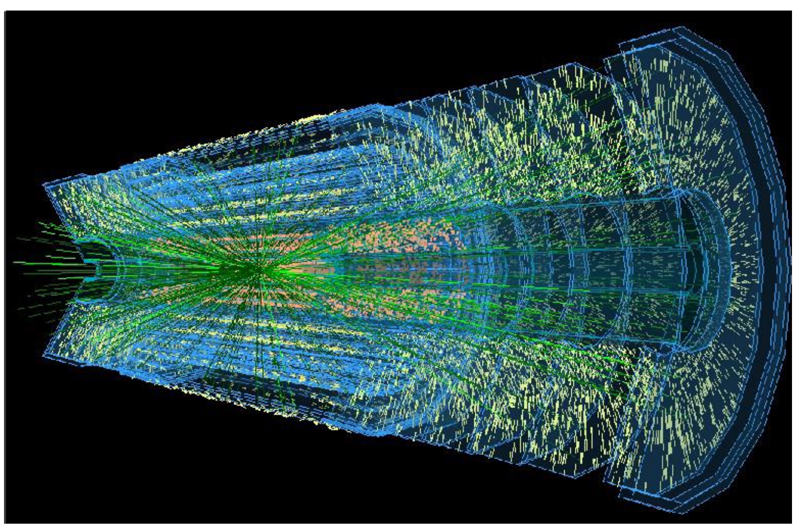
CMS

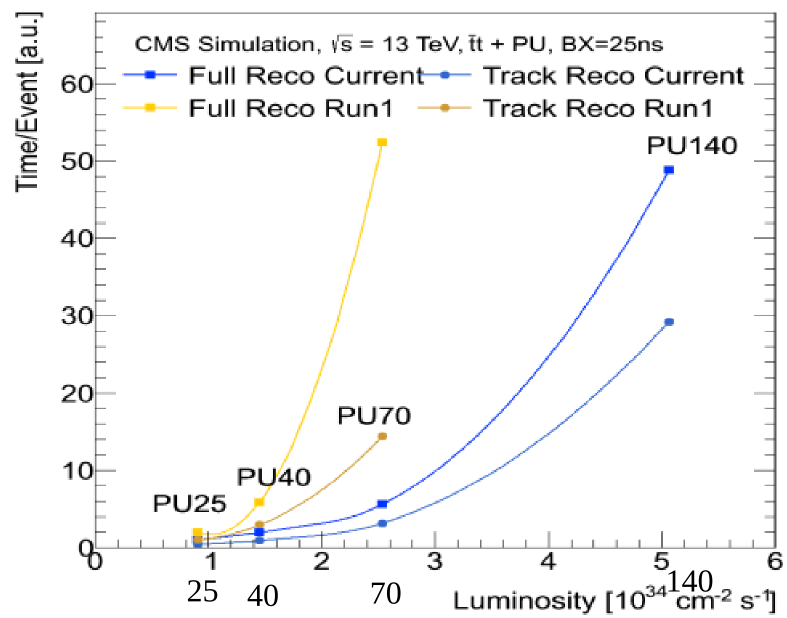


Tracking crisis

- 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

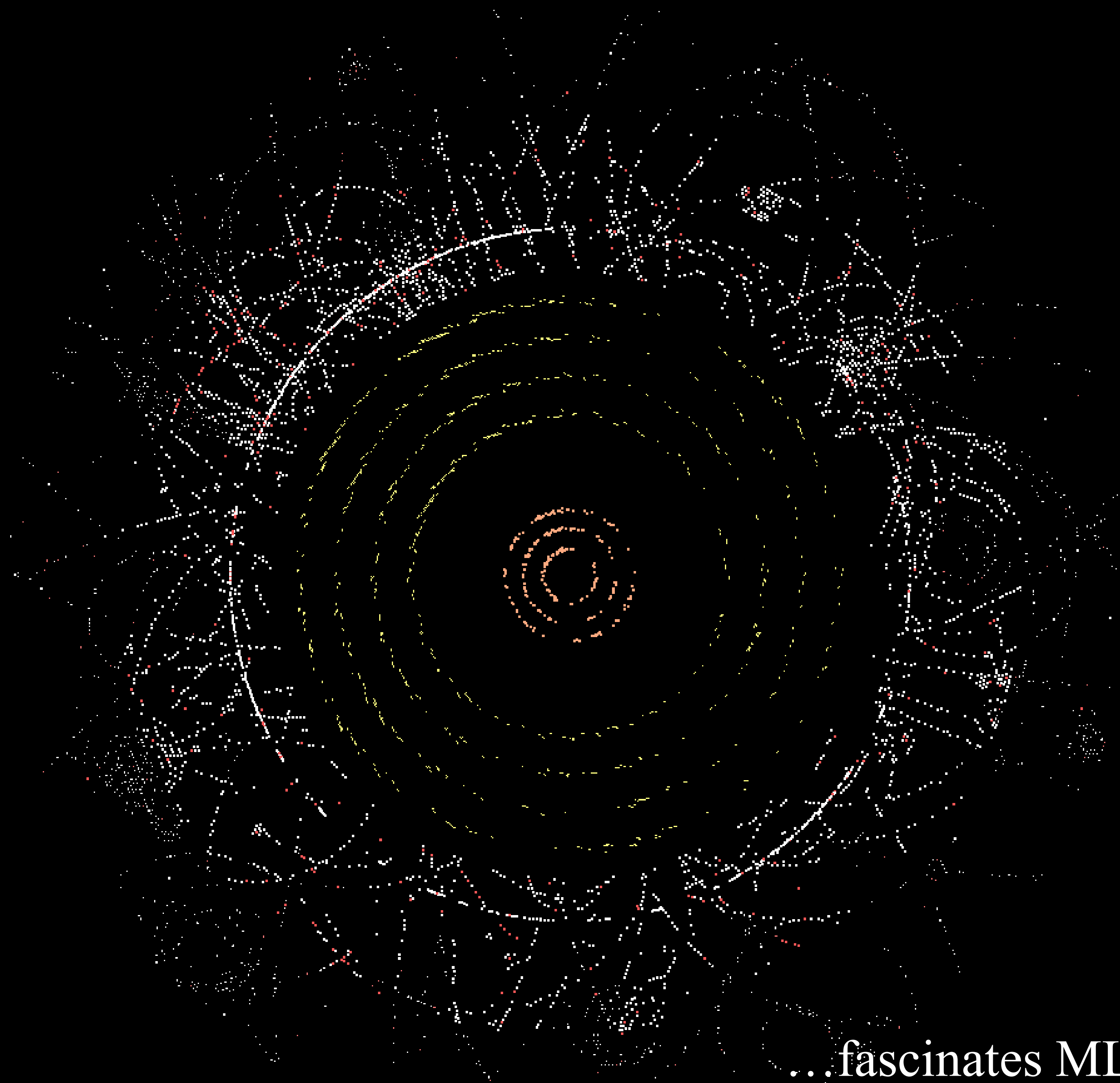




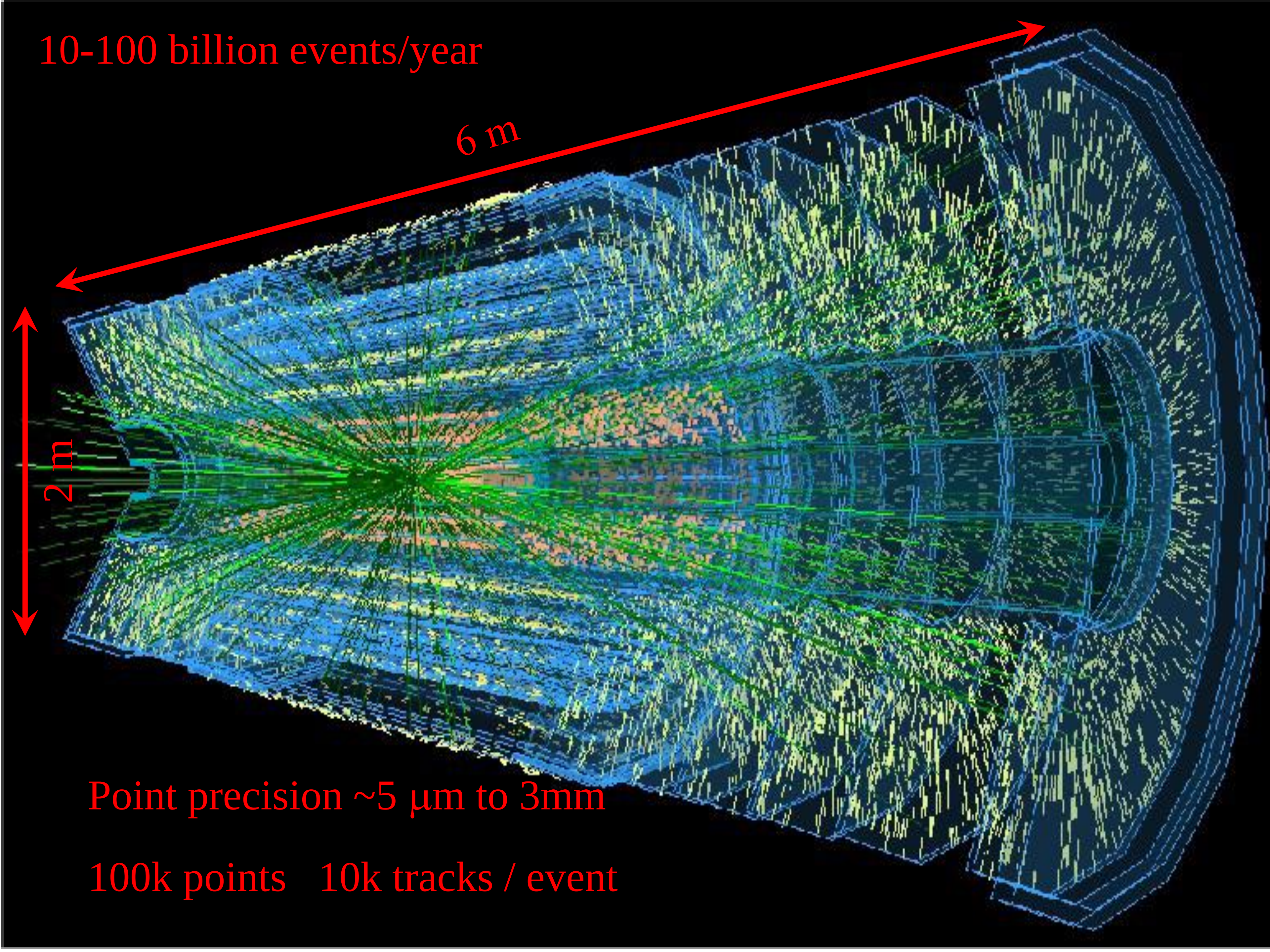


Particle Tracking

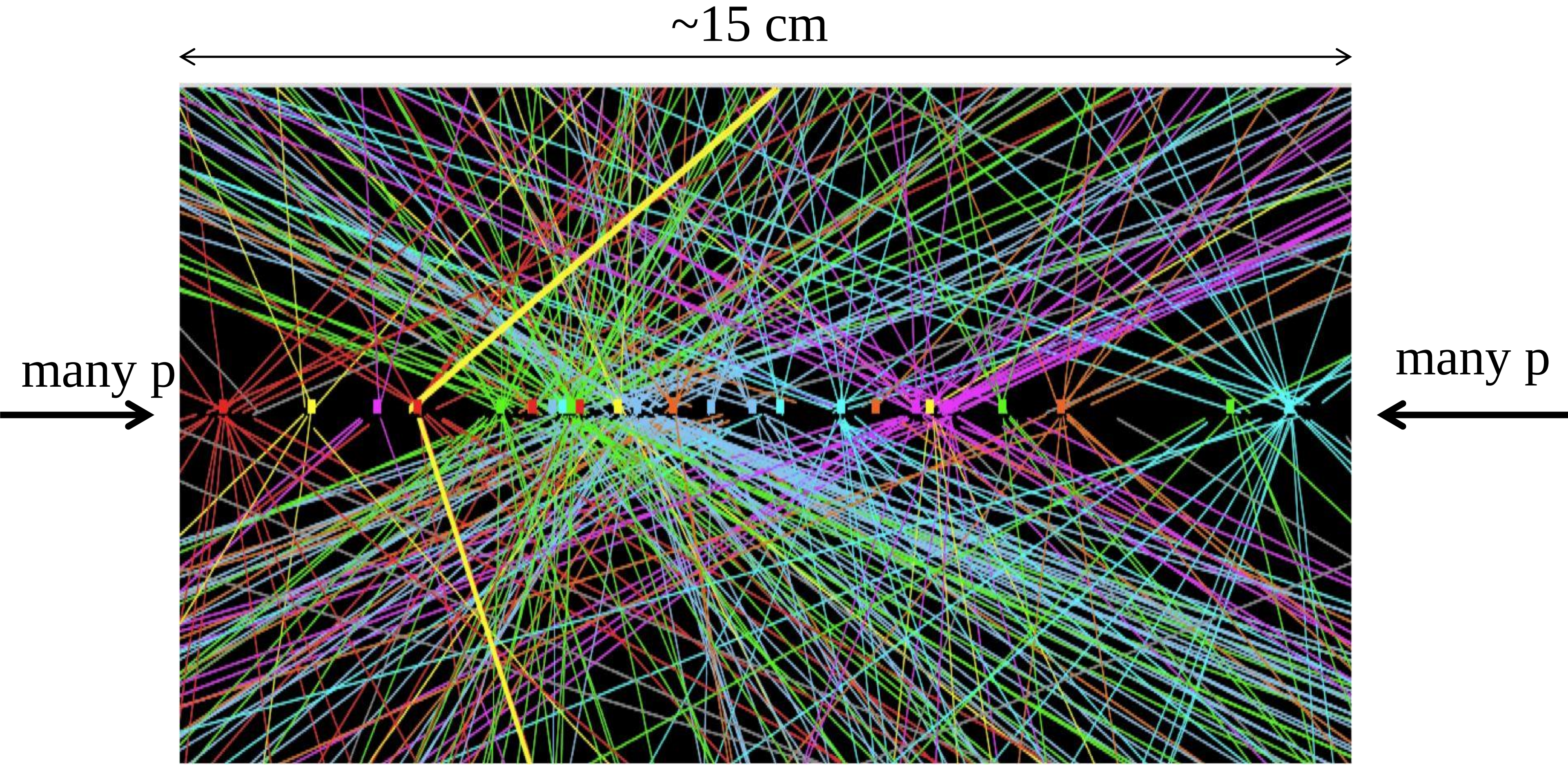
LHC tracking…



…fascinates ML scientists

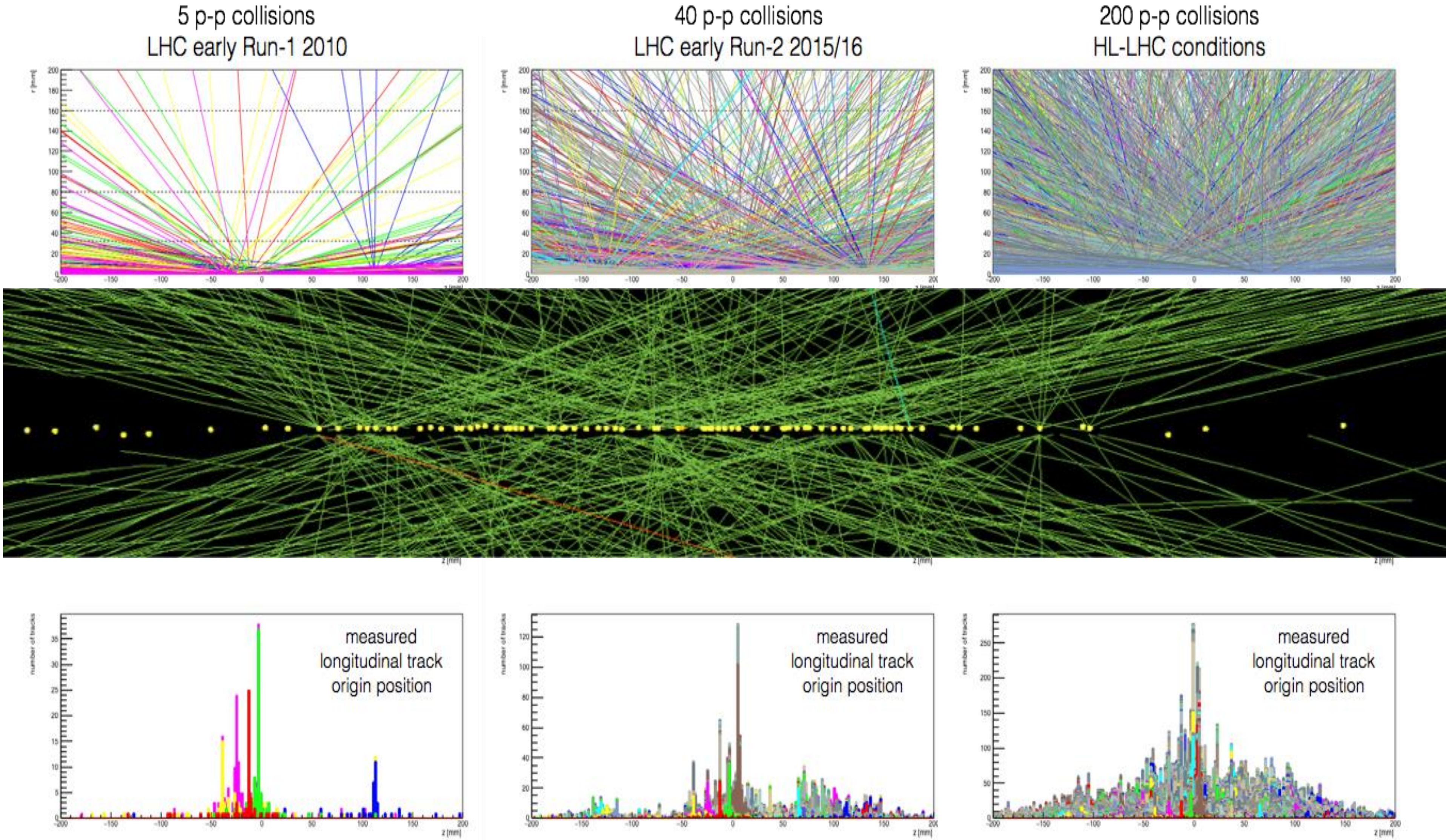

10-100 billion events/year
6 m
2 m
Point precision ~5 μm to 3mm
100k points 10k tracks / event

Bunch collision



Current situation: 20 parasitic collisions
High Lumi-LHC : 200 parasitic collisions

Pile-up



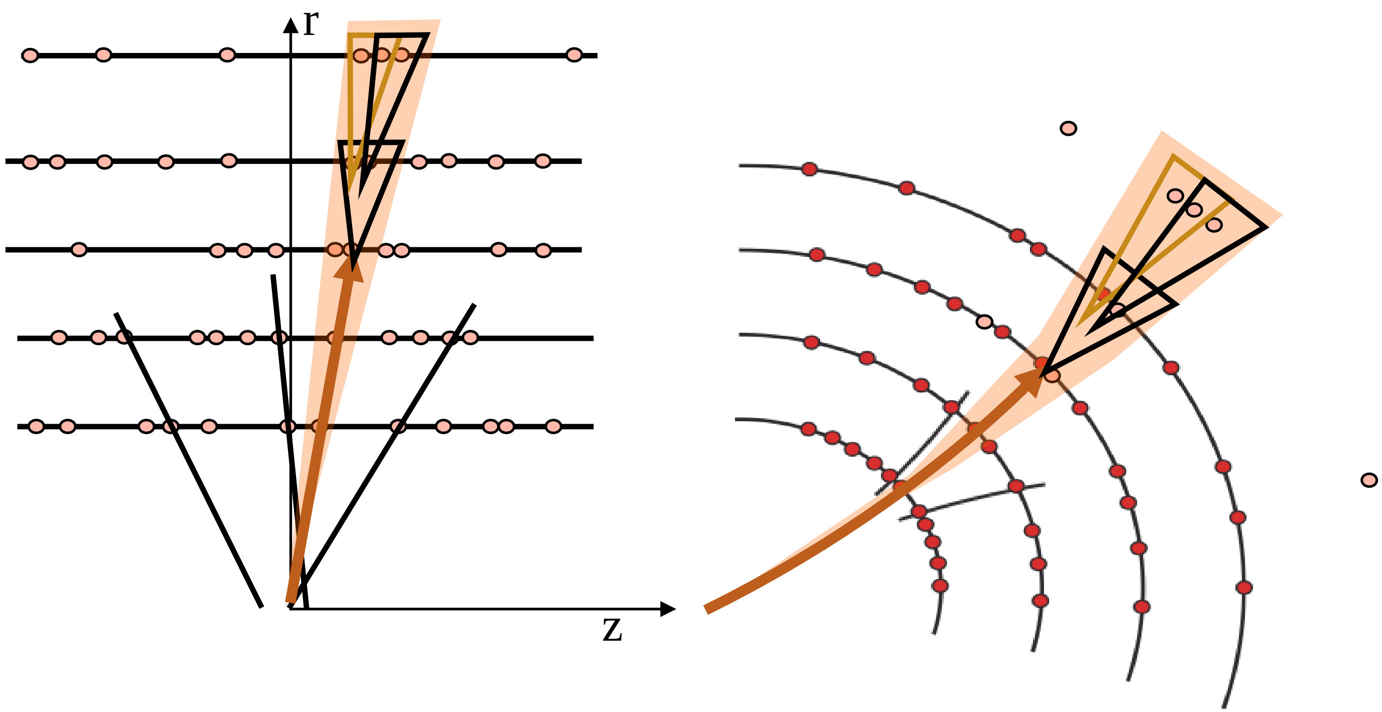
5 p-p collisions
LHC early Run-1 2010
40 p-p collisions
LHC early Run-2 2015/16
200 p-p collisions
HL-LHC conditions
measured
longitudinal track
origin position
measured
longitudinal track
origin position
measured
longitudinal track
origin position

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

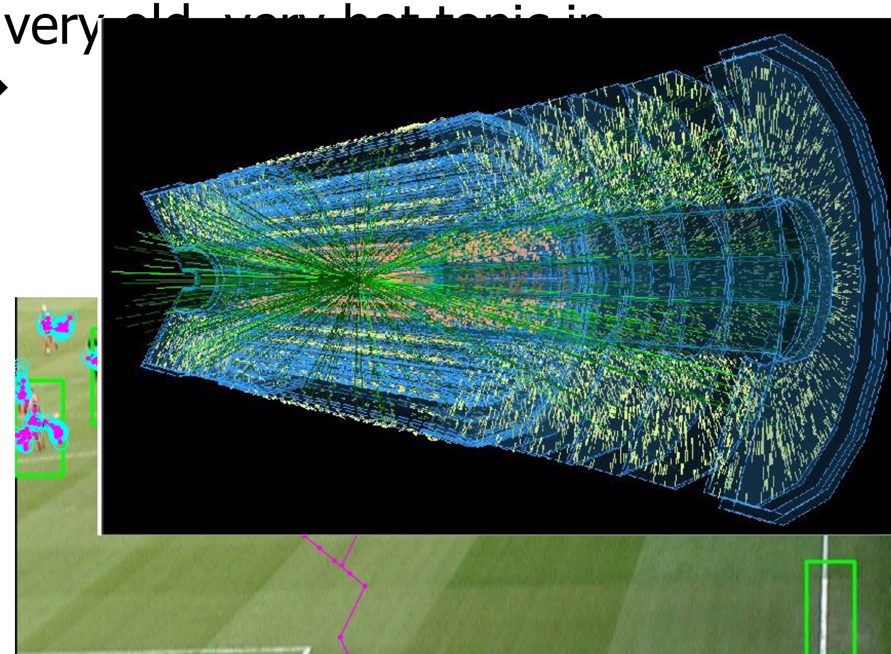
Current Algorithms

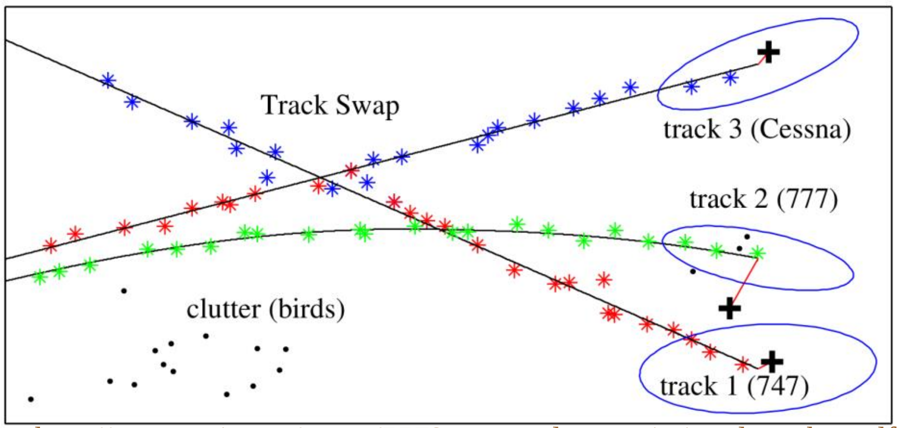
- Pattern : connect 3D points into tracks
- Essentially combinatorial approach
- Tracks are (not perfect) helices pointing (approximately) to the origin
- Challenge : explore completely new approaches
- (not part of the challenge : given the points, estimate the track parameters)



Pattern recognition in ML

- Pattern recognition, tracking, is a very
  Artificial Intelligence : examples➔

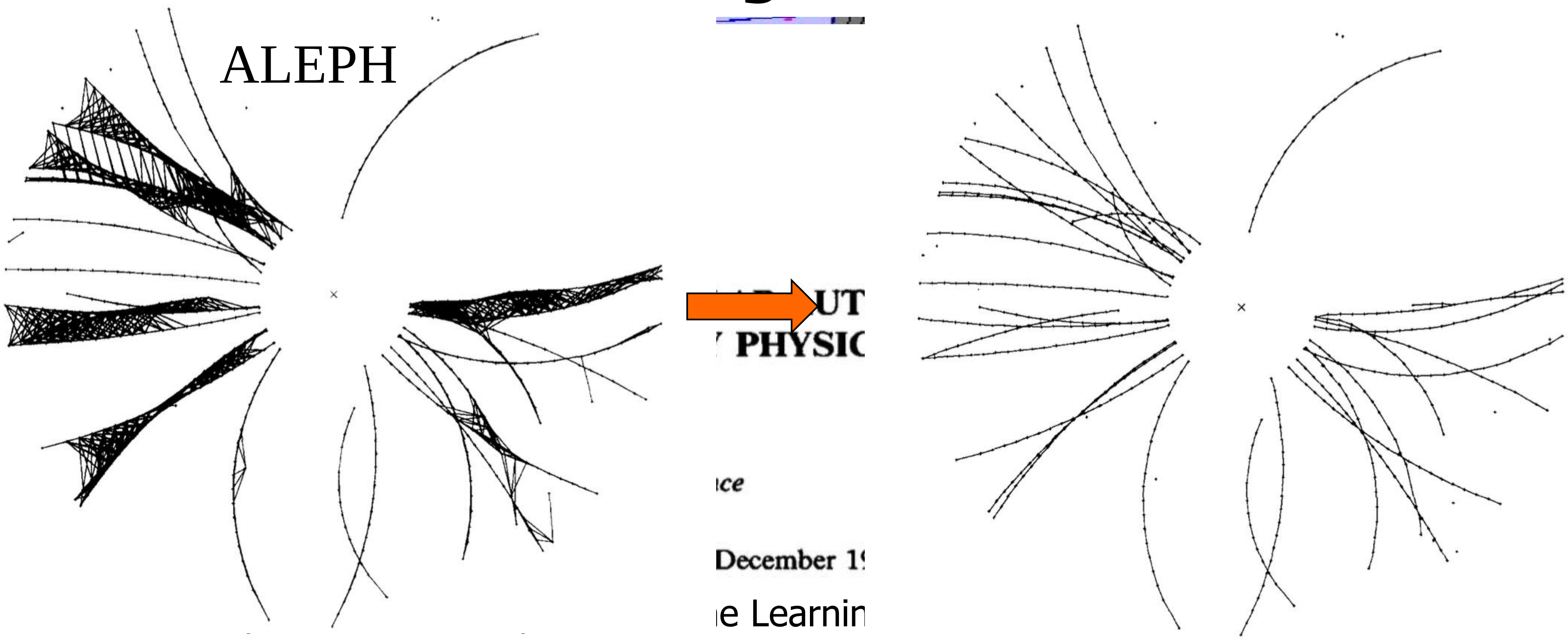




http://papers.nips.cc/paper/5572-a-complete-variational-tracker.pdf

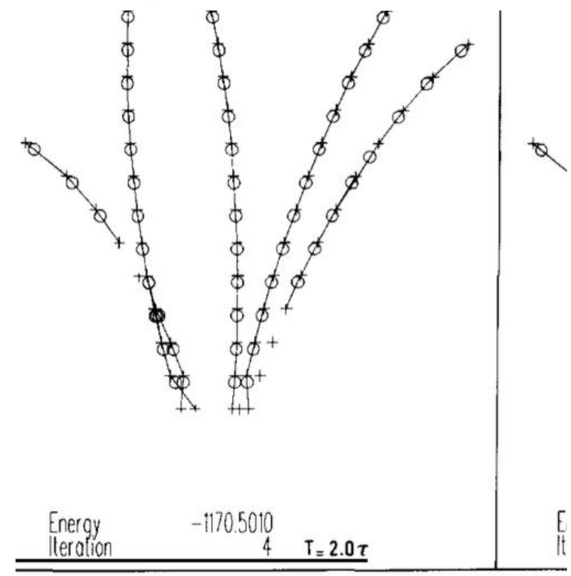
- Note that these are real-time applications, with CPU constraints
- Worry about efficiency, "track swap",…
- But no on-the-shelf algorithm will solve our problem
- (in fact a few lines calling DBScan in sk-learn does find some tracks)

An early attempt



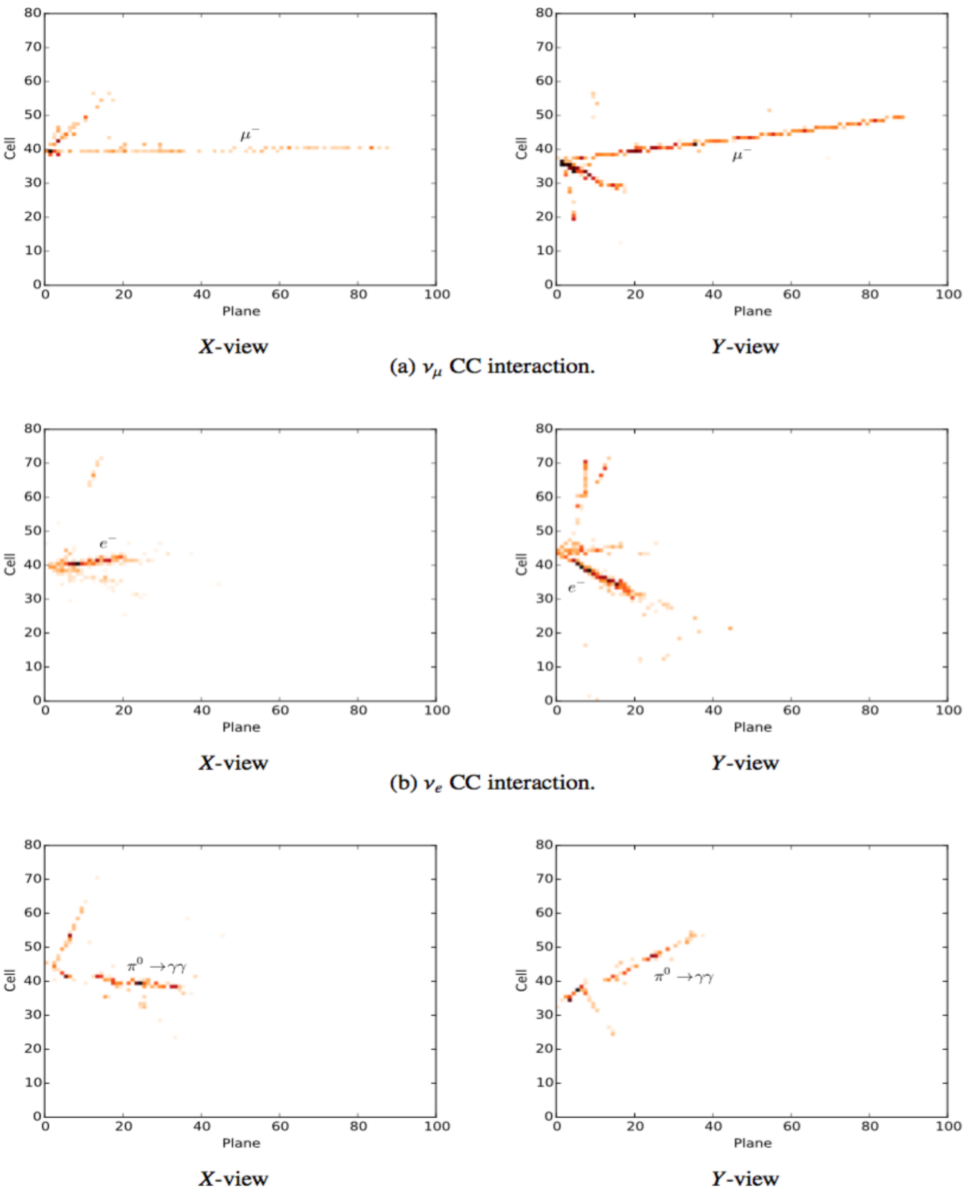
known

- 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



(a) ν_μ CC interaction.

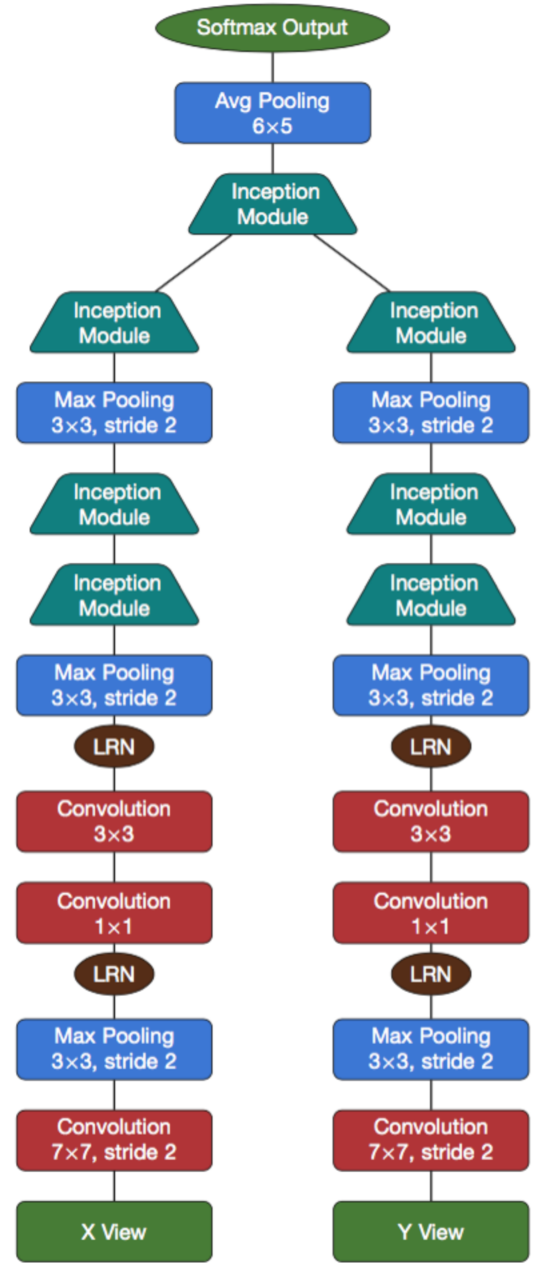
(b) ν_e CC interaction.

(c) NC interaction.

Neutrino interaction classification

Using Convolutionnal Neural Network

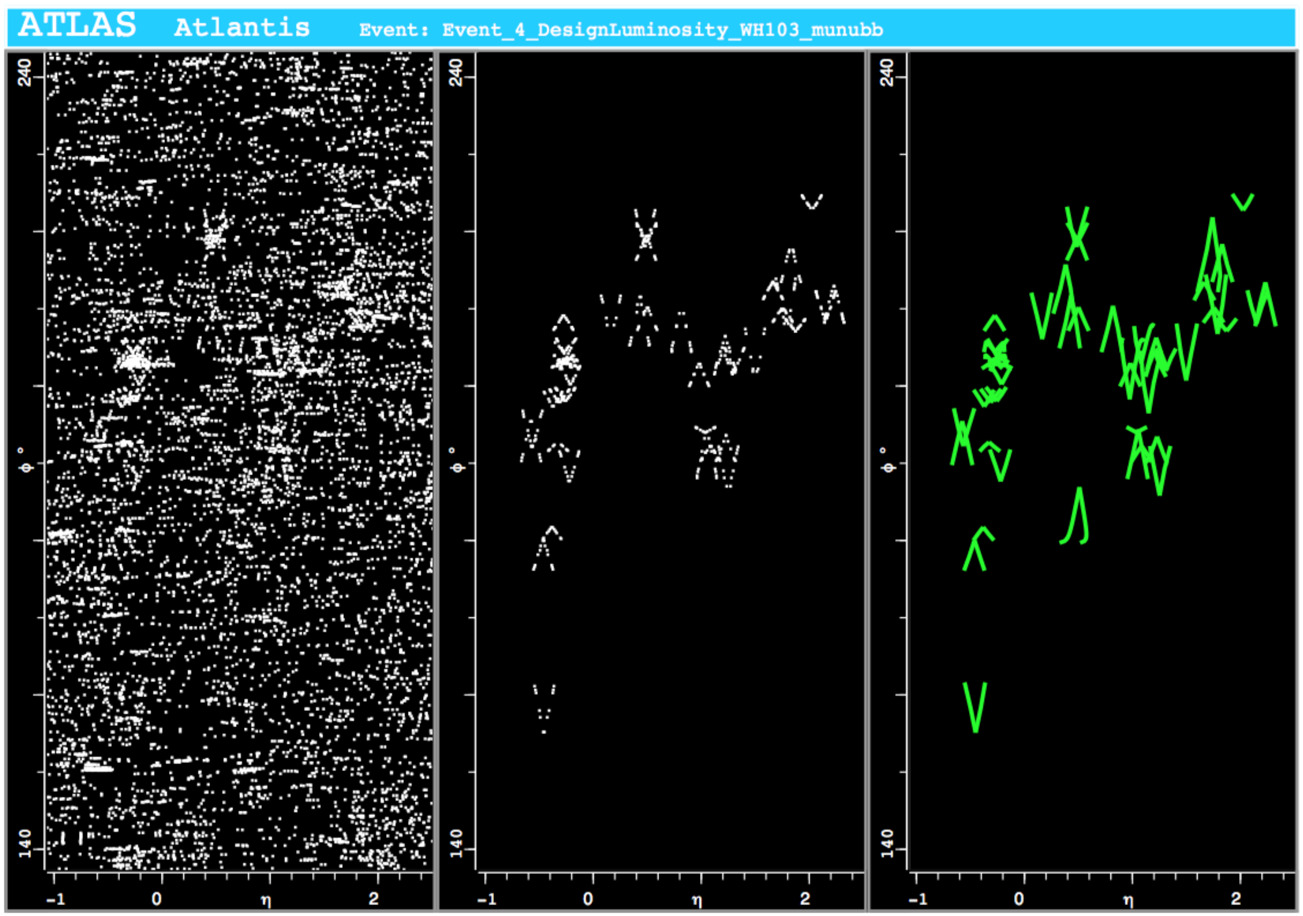
No attempt to separate individual tracks.



V plots

- Tracks are not visible by eye
- How ever they are with a clever projection :Eta phi projection with $\delta\eta = +/- \varepsilon(r_{max} - r)$
- See G. Taylor Nucl. Inst. and Meth. A 549 (2005) 183–187

CHEP04



20 parasitic collisions

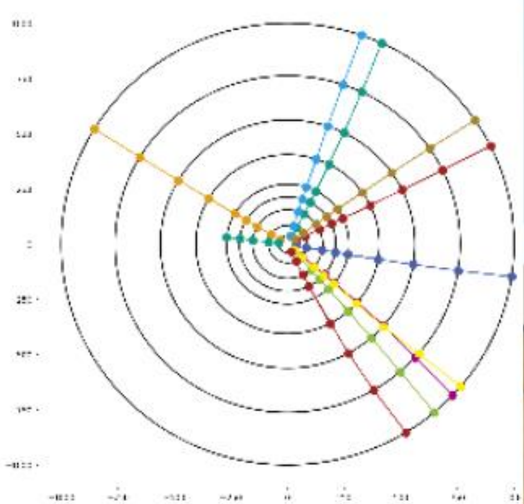
TrackML Ramp


- A simplified tracking challenge setup on RAMP (Center for Data Science Paris-Saclay platform, Balazs Kégl)
- A (non completely trivial) 2D simulation with ~10 tracks instead of 3D/10.000 tracks
- Run as a 40 hours hackathon during CTDWIT 6-9th March 2017 LAL-Orsay
- Allowed to validate robustness a scoring variable and show richness of possible algorithms: combinatorial (HEP baseline), conformal mapping, MCTS, LSTM (See also S. Farrell et al paper accepted by NIPS 2017 “Deep Learning for Physical Science”

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



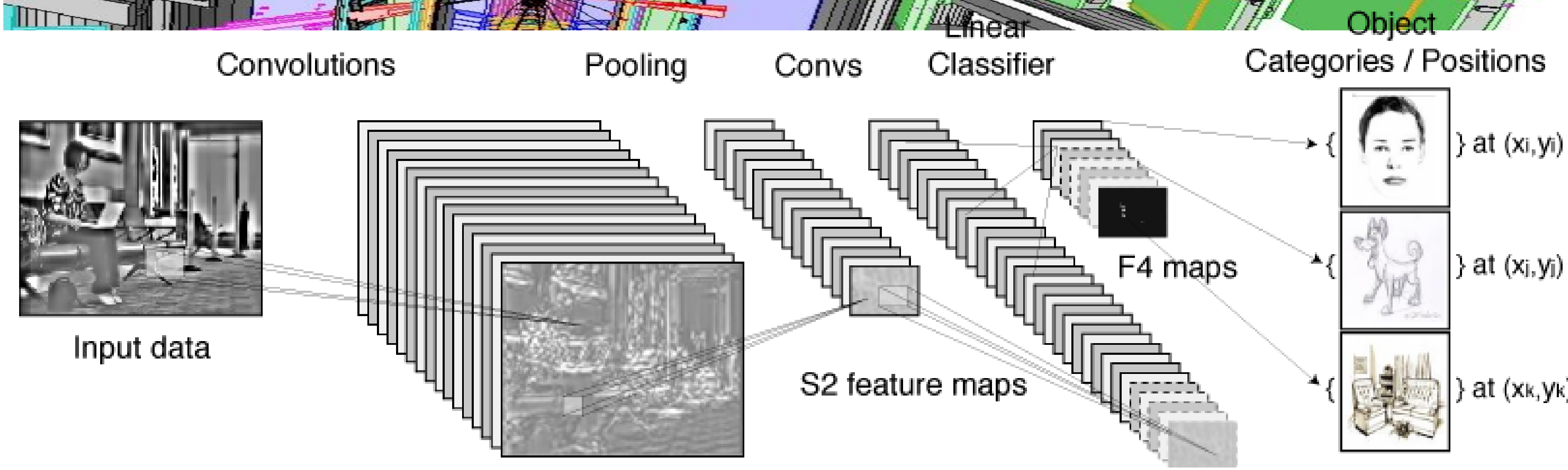

1

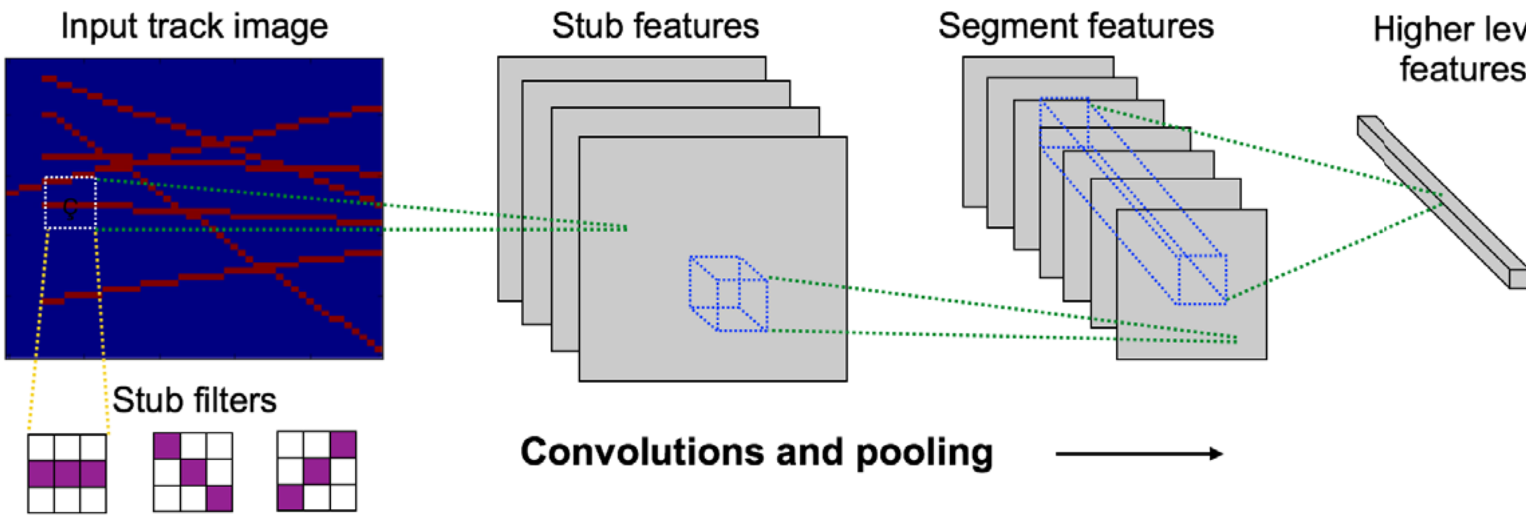
David Rousseau
@dhpmrou

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

À l'origine en anglais

Convolution NN

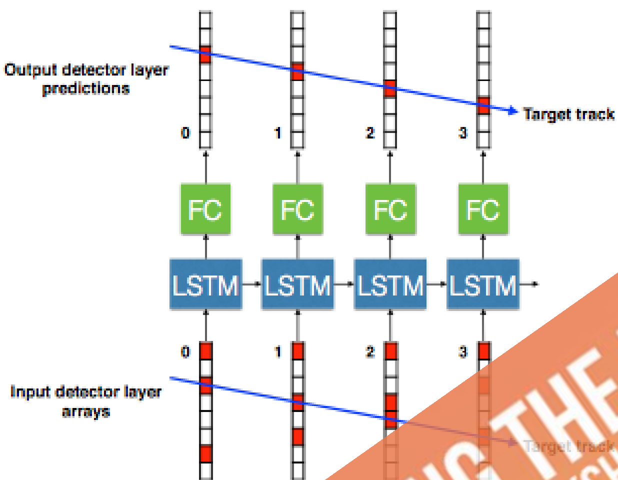


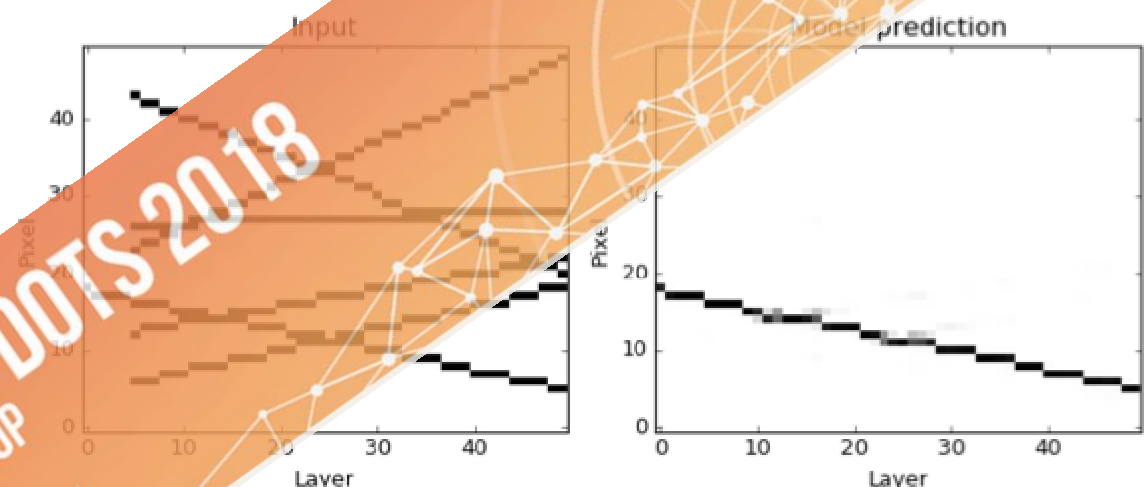


See: Farrel S. et al, The HEP.TrkX Project: deep neural networks for HL-LHC online and offline tracking, EPJ Web of Conferences 150, 00003 (2017)

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)

2014 HiggsML challenge recap

HiggsML in a nutshell

- (see JMLR proceedings http://proceedings.mlr.press/v42/cowa14.html)
- ATLAS Htautau MC analysis ntuple released
- Competition on kaggle to optimise Higgs selection : https://higgsml.lal.in2p3.fr
- 1785 teams (1942 people) have participated (participation=submission of at least one solution)
  - (6517 people have downloaded the data)
  - ➔most popular challenge on the Kaggle platform (until spring 2015)
  - 35772 solutions uploaded
- 136 forum topics with 1100 posts

What data did we release ?

- From ATLAS full sim Geant4 MC12 production
- 30 variables
- Signal is H→tautau, Background a mixture of : Z, top, W
- Based on November 2013 ATLAS Htautau conf note ATLAS-CONF-2013-108
- Preselection for lep-had topology : single lepton trigger, one lepton identified, one hadronic tau identified
- →800.000 events (all that was available):
  - 250.000 training data set
  - 550.000 test data set without label and weight
- Reproduces reasonably well (~20%) content of 3 highest sensitivity bins (x 2 categories) in conf note
- (some background and many correction factors deliberately omitted so that the sample cannot be used for physics, only for machine learning studies)

Dataset

- Permanently available and usable by anyone (also non ATLAS) on CERN Open Data:
- http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014
- ASCII csv file, with mixture of Higgs to tautau (lephad) signal and corresponding backgrounds, from official GEANT4 ATLAS simulation
- Weight and signal/background (for training dataset only)
- weight (fully normalised)
- label : « s » or « b »
- Conf note variables used for categorization or BDT:
  - DER_mass_MMC
  - DER_mass_transverse_met_lep
  - DER_mass_vis
  - DER_pt_h
  - DER_deltaeta_jet_jet
  - DER_mass_jet_jet
  - DER_prodeta_jet_jet
  - DER_deltar_tau_lep
  - DER_pt_tot
  - DER_sum_pt
  - DER_pt_ratio_lep_tau
  - DER_met_phi_centrality
  - DER_lep_eta_centrality

- Primitive 3-vectors allowing to compute the conf note variables (mass neglected),
- 16 independent variables:
  - PRI_tau_pt
  - PRI_tau_eta
  - PRI_tau_phi
  - PRI_lep_pt
  - PRI_lep_eta
  - PRI_lep_phi
  - PRI_met
  - PRI_met_phi
  - PRI_met_sumet
  - PRI_jet_num (0,1,2,3, capped at 3)
  - PRI_jet_leading_pt
  - PRI_jet_leading_eta
  - PRI_jet_leading_phi
  - PRI_jet_subleading_pt
  - PRI_jet_subleading_eta
  - PRI_jet_subleading_phi
  - PRI_jet_all_pt

Real life vs challenge

Real life

1. Systematics (and data vs MC)
2. 2 categories x n BDT score bins
3. Background estimated from data (embedded, anti tau, control region) and some MC
4. Weights include all corrections. Some negative weights (tt)
5. Potentially use any information from all 2012 data and MC events
6. Few variables fed in two BDT
7. Significance from complete fit with NP etc…
8. MVA with TMVA BDT

Challenge

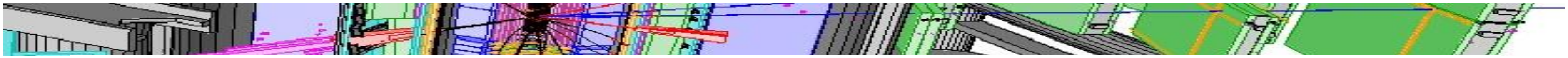
1. No systematics
2. No categories, one signal region
3. Straight use of ATLAS G4 MC
4. Weights only include normalisation and pythia weight. Neg. weight events rejected.
5. Only use variables and events preselected by the real analysis
6. All BDT variables + categorisation variables + primitives 3-vector
7. Significance from "regularised Asimov"
8. MVA "no-limit"

Simpler, but not too simple!

Final leaderboard

#	Δrank	Team Name ‡ model uploaded * in the money		Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Gábor Melis ‡ *	7000$ « deep » learning	3.80581	110	Sun, 14 Sep 2014 09:10:04 (-0h)
2	↑1	Tim Salimans ‡ *	4000$ BDT ensemble	3.78913	57	Mon, 15 Sep 2014 23:49:02 (-40.6d)
3	↑1	nhlx5haze ‡ *	2000$	3.78682	254	Mon, 15 Sep 2014 16:50:01 (-76.3d)
4	↑38	ChoKo Team		3.77526	216	Mon, 15 Sep 2014 15:21:36 (-42.1h)
5	↑35	cheng chen		3.77384	21	Mon, 15 Sep 2014 23:29:29 (-0h)
6	↑16	quantify		3.77086	8	Mon, 15 Sep 2014 16:12:48 (-7.3h)
7	↑1	Stanislav Semenov & Co (HSE Yandex)		3.76211	68	Mon, 15 Sep 2014 20:19:03
8	↓7	Luboš Motl's team	Best physicist	3.76050	589	Mon, 15 Sep 2014 08:38:49 (-1.6h)
9	↑8	Roberto-UCIIIM		3.75864	292	Mon, 15 Sep 2014 23:44:42 (-44d)
10	↑2	Davut & Josef		3.75838	161	Mon, 15 Sep 2014 23:24:32 (-4.5d)
45	↑5	crowwork ‡	HEP meets ML award XGBoost authors Free trip to CERN	3.71885	94	Mon, 15 Sep 2014 23:45:00 (-5.1d)
782	↓149	Eckhard	TMVA expert, with TMVA improvements	3.49945	29	Mon, 15 Sep 2014 07:26:13 (-46.1h)
991	↑4	Rem.		3.20423	2	Mon, 16 Jun 2014 21:53:43 (-30.4h)
		simple TMVA boosted trees		3.19956		

Why challenges work ?



MOTIVATION OF ORGANIZING CONTESTS: EXTREME VALUE

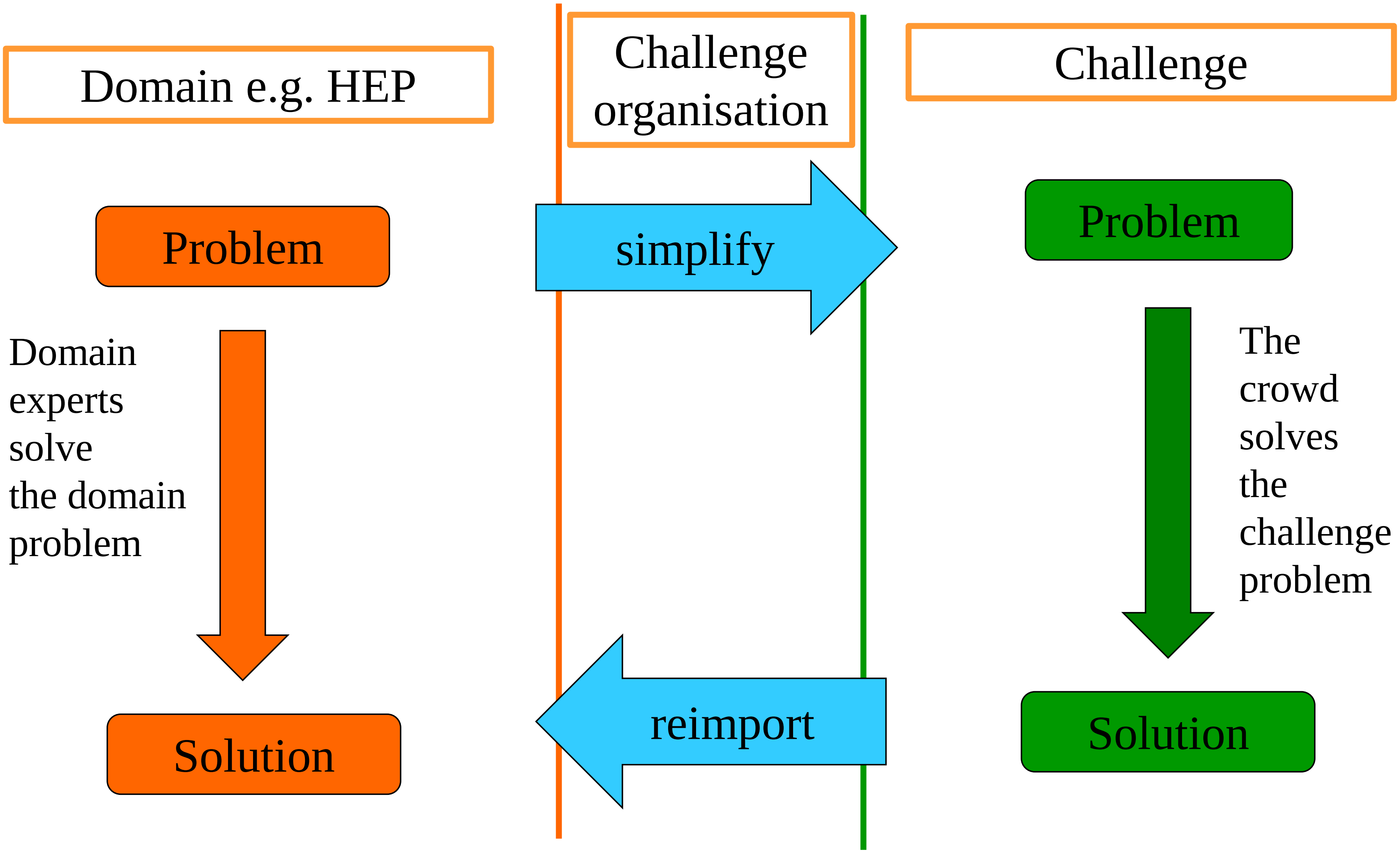
Courtesy : Lakhani 2014

Experts are highly skilled, trained -> more focused, performed solution, low variety

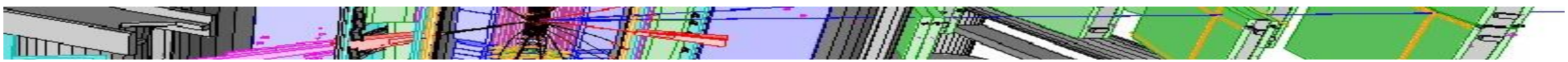
OI is suitable for a variety of nonconvential surprising ideas that are « far » from traditional expertise -> high volatility

Not just ML, but a general trend: Open Innovation

From domain to challenge and back

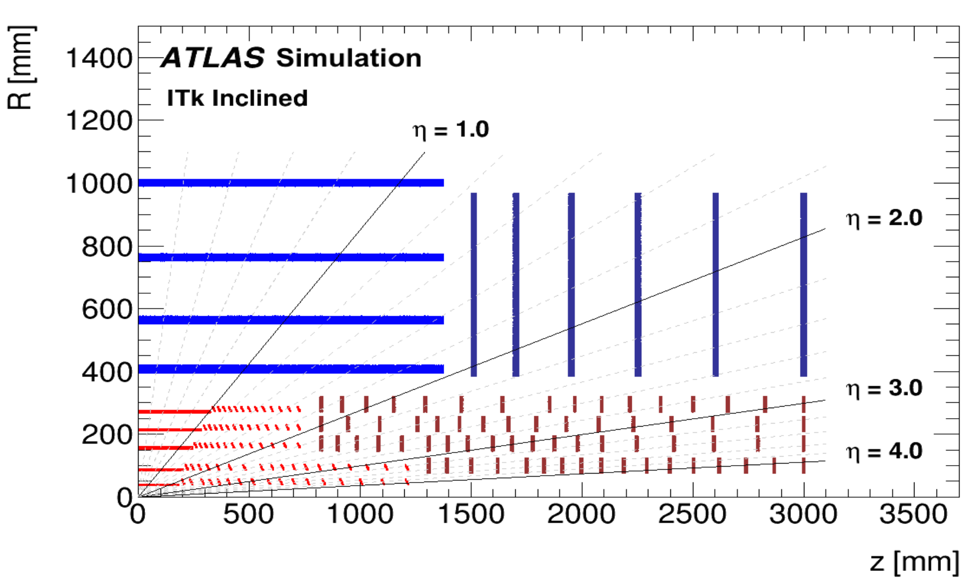


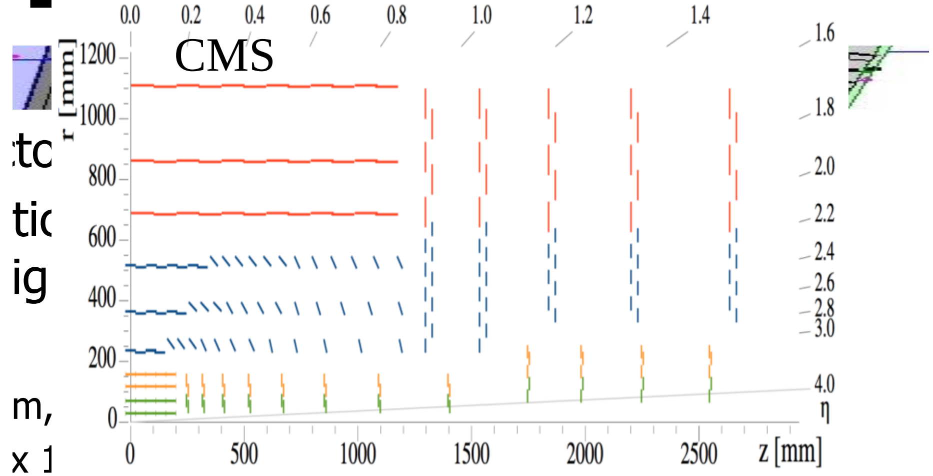
The tracking challenge



In a nutshell

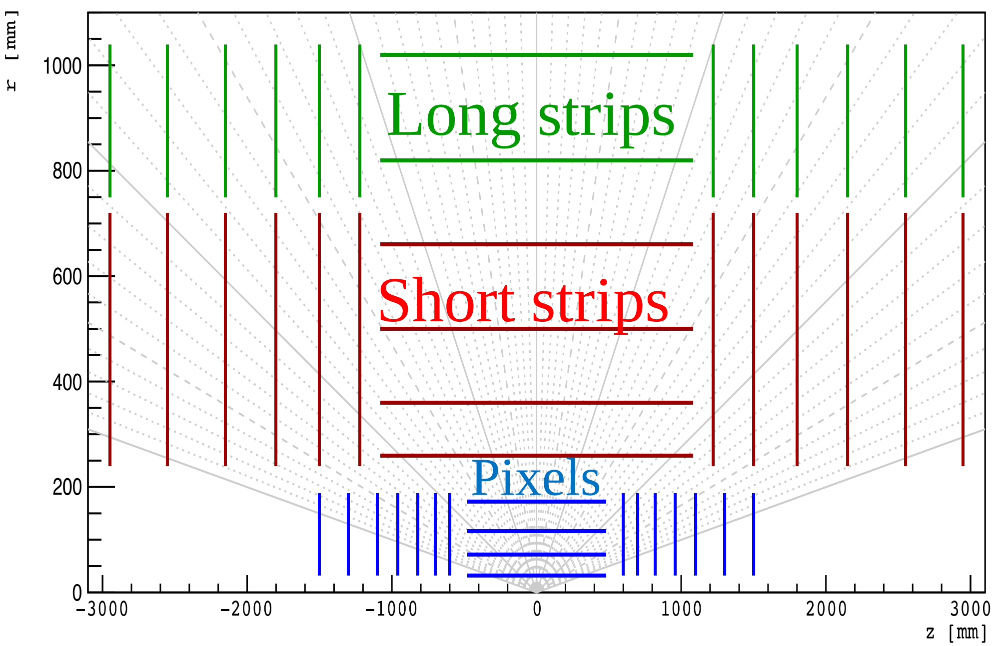
- Accurate simulation engine (ACTS https://gitlab.cern.ch/acts/acts-core) to produce realistic events
  - One file with list of 3D points
  - Ground truth : one file with point to particle association
  - Ground truth auxiliary : true particle parameter (origin, direction, curvature)
  - Typical events 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

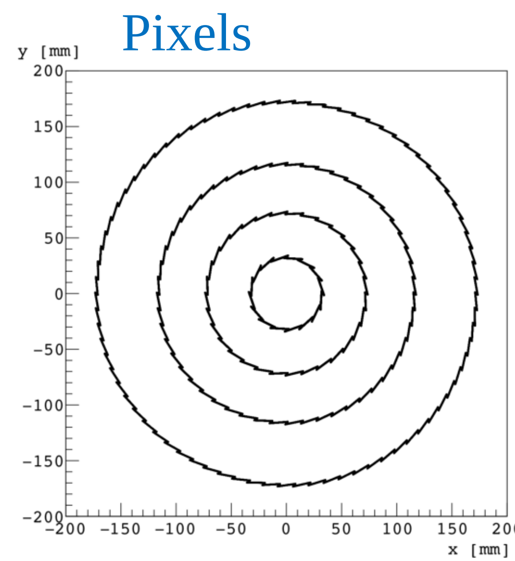




tc
tic
ig
m,
x 1

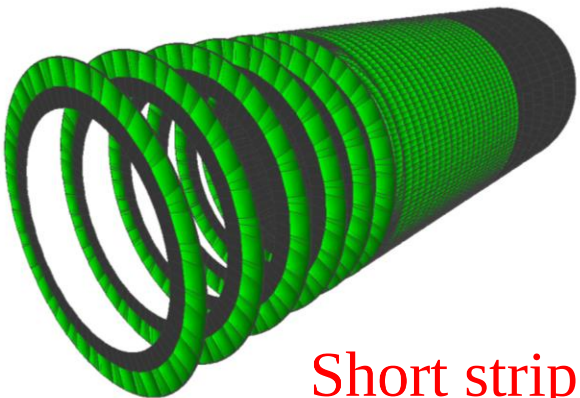
x 10.8 mm, digital clusters: σ ~ 25 µm x 3.1mm



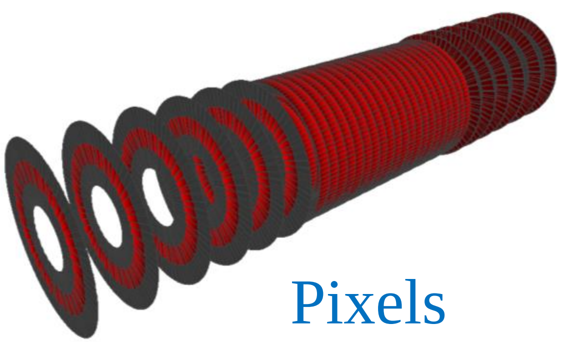


Detector : layout

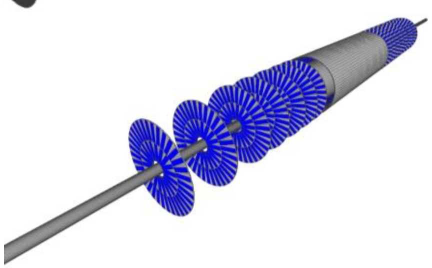
Long strips

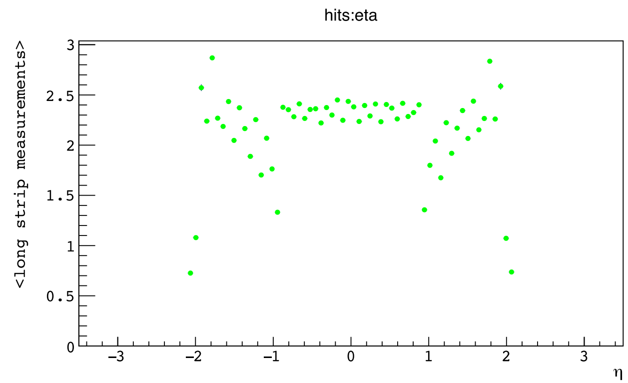


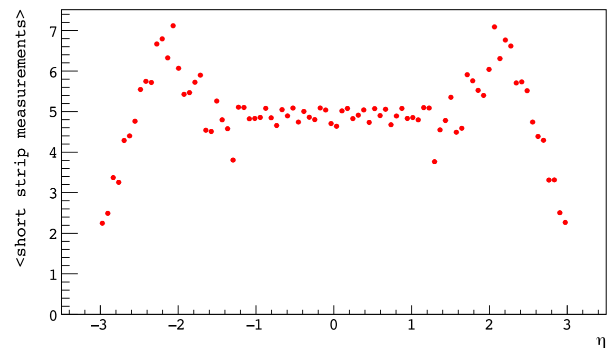
Short strips

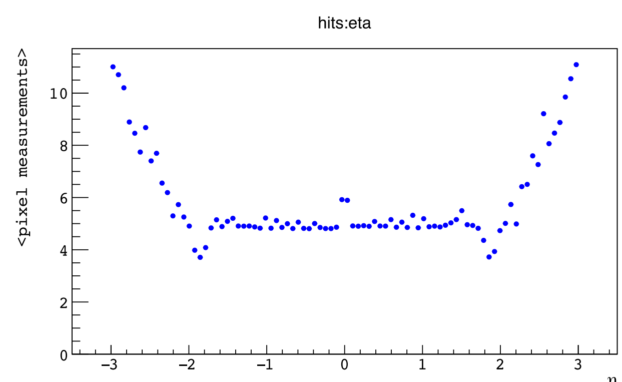


Pixels



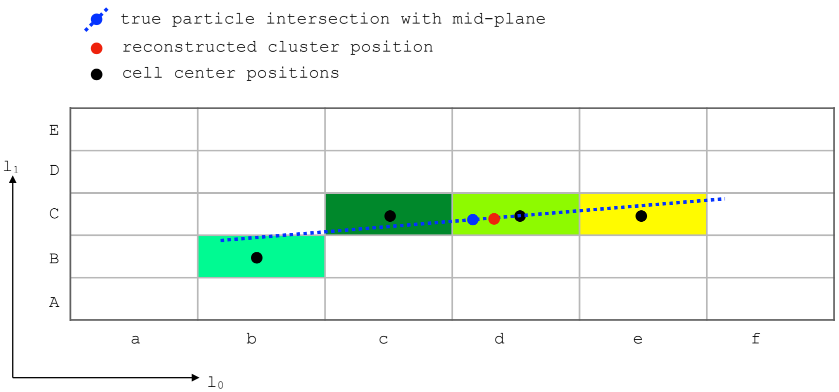


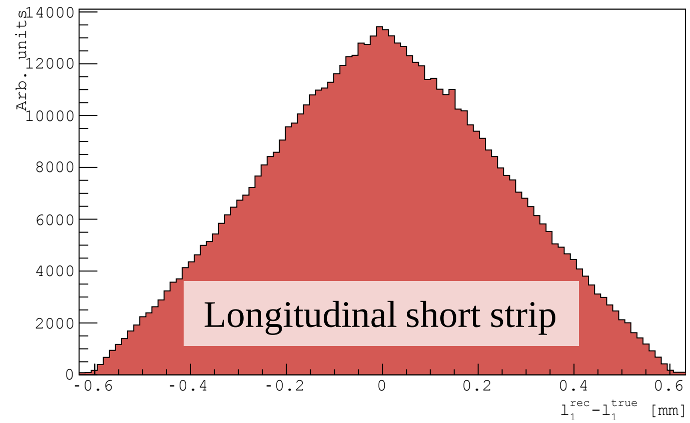




~12 points per tracks

Detector resolution





Clustering : analog in Pixel, digital in Strips

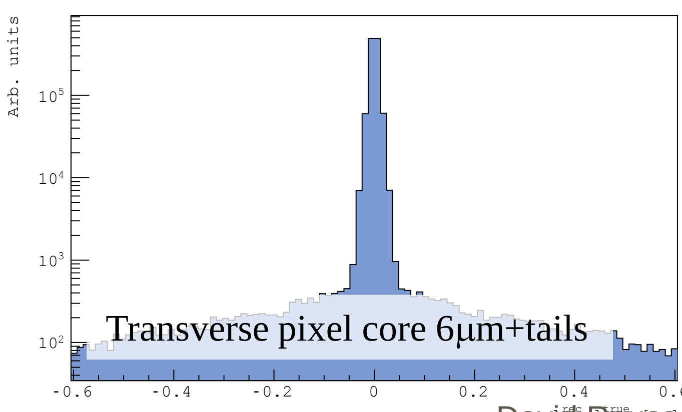
Different pitches

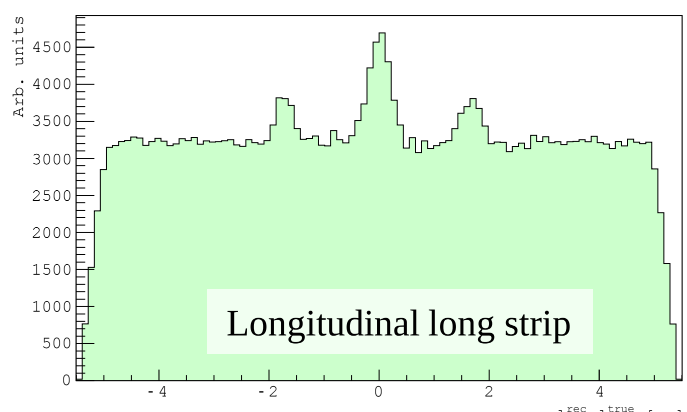
➔very different residuals (see examples)

➔we'll let participants figure out given $(x,y,z)_{measured}$ ⇔ $(x,y,z)_{true}$

Non trivial simplification : one true track ⇔one reco hit (except for 1% inefficiency)

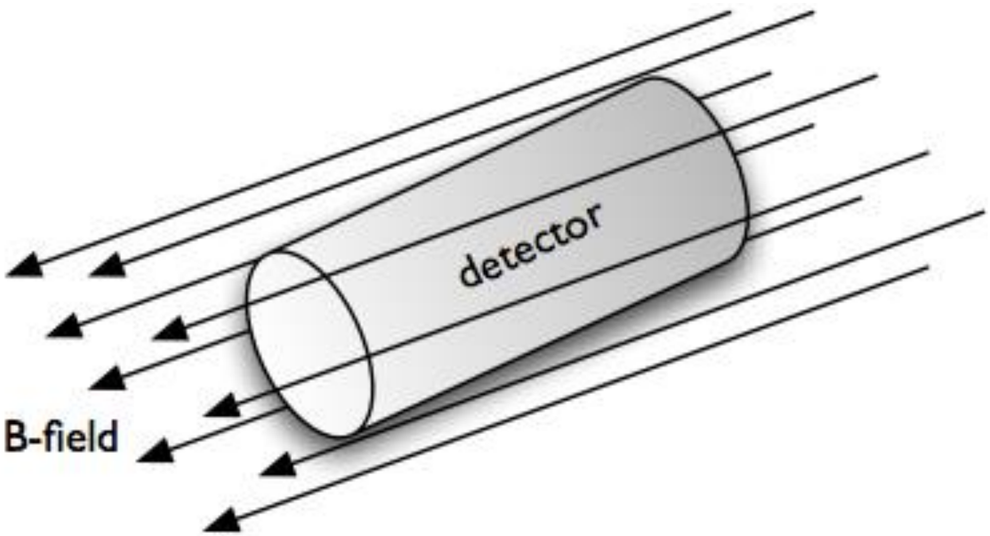
=>no hit merging/splitting

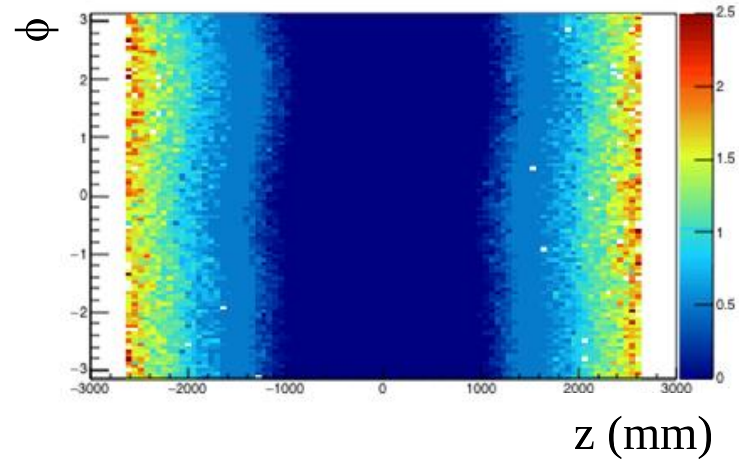


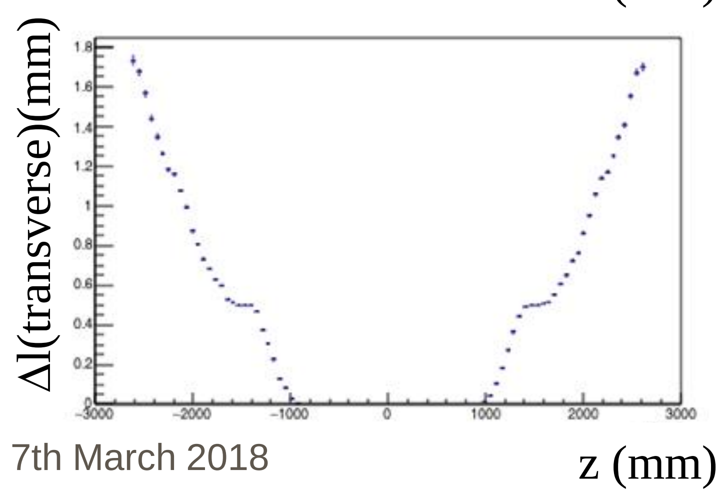


Magnetic Field

- If B field uniform➔tracks are perfect helices (except for MS)
- However ATLAS/CMS magnetic field not perfectly uniform (Solenoïd too shorts, Tilt)
- ➔Event simulated with ATLAS field map
- ➔systematic departure from perfect helix reaches ~1mm at middle radius at high rapidity
- ➔broken azimuthal symmetry
- ➔taking this into account not mandatory to get started, but ultimately needed
- We don't provide the field map to participant

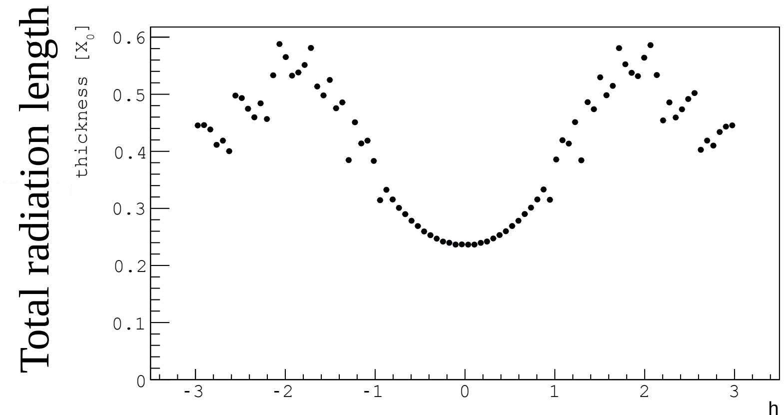


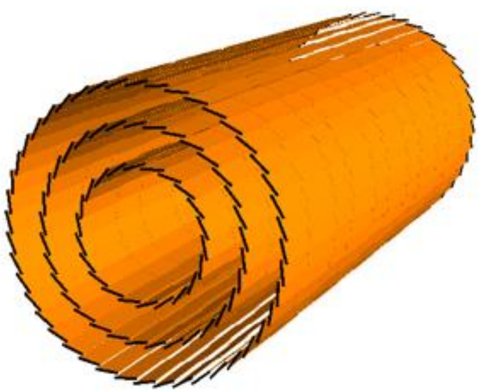




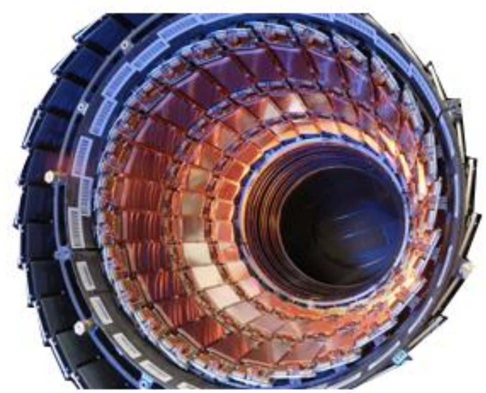
Material

- Per layer : Radiation length : 2-3%, Interaction length : 1%
- As uniform cylinder and slabs, no attempt for detailed electronics, services description





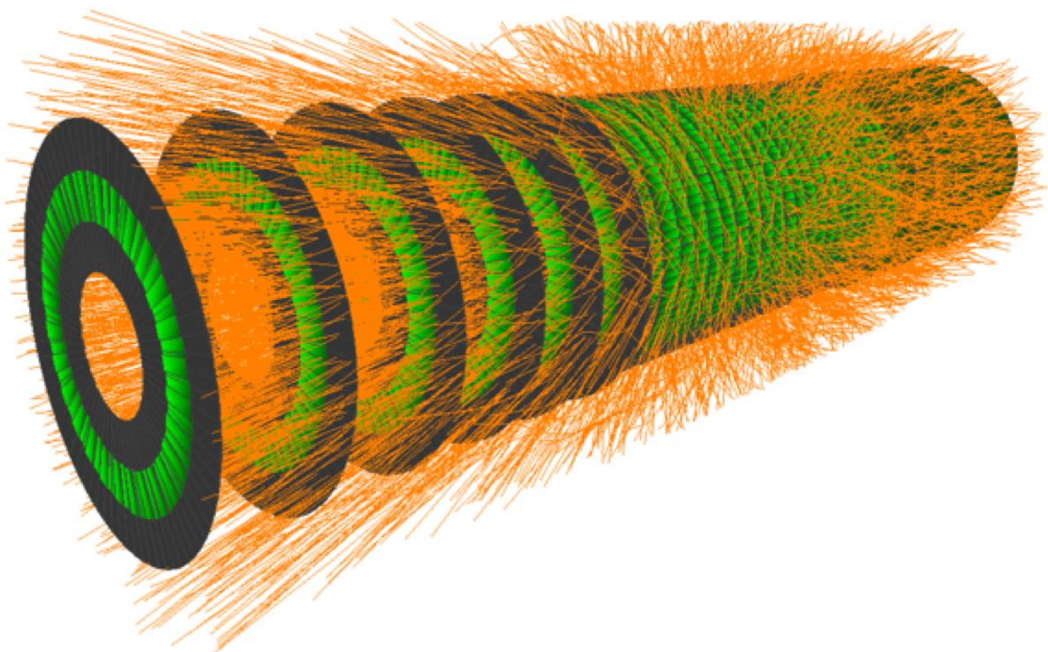
ideal

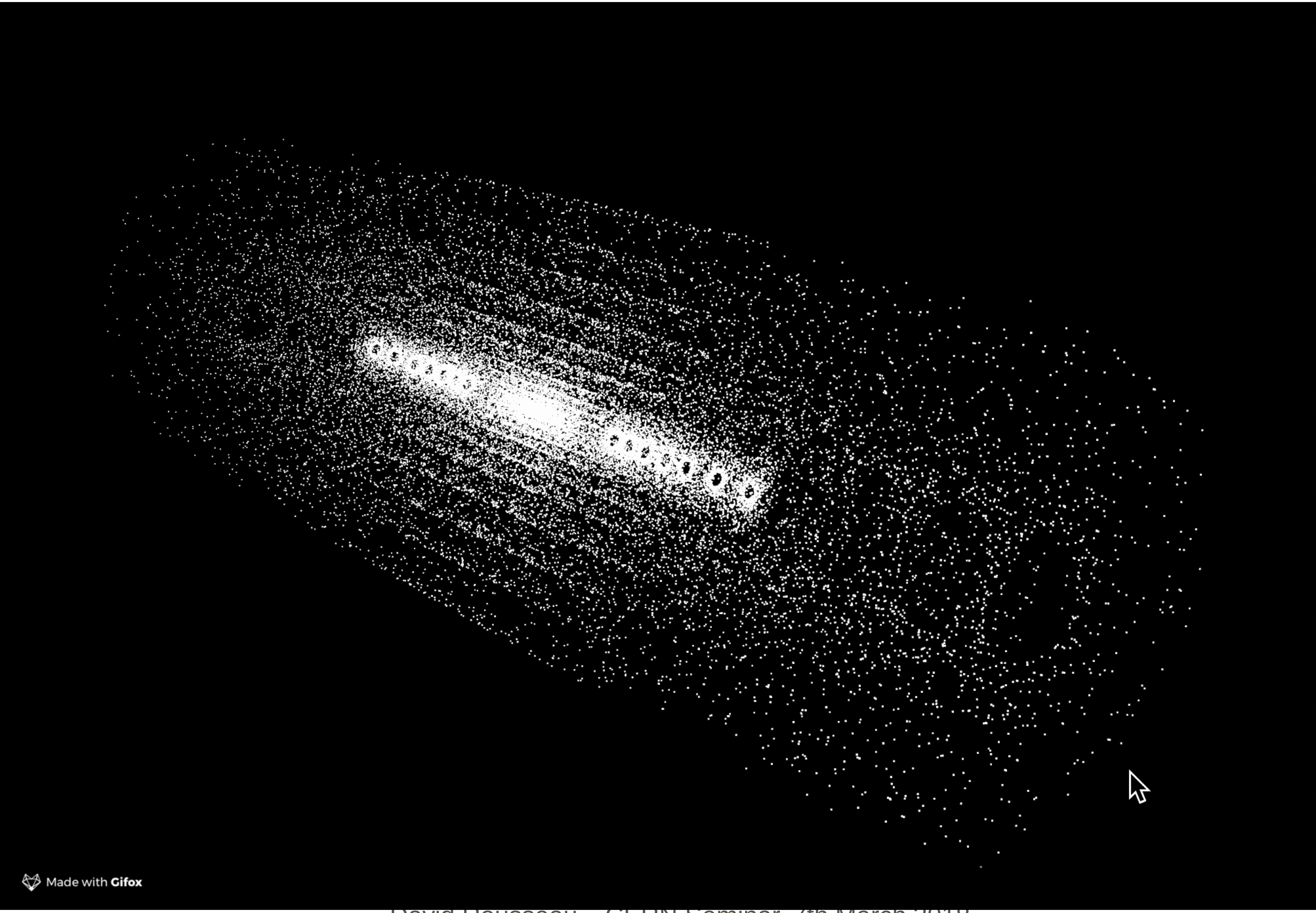


real

Event simulation

- Pythia tt-bar event
- Overlaid with Poisson(200) Pythia minimum bias
- Luminous region : gaussian σ_z=5.5 cm, transverse σ=15μm
- 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	58562600635465728	-63.972698	-3.72889	-1502.5	-0.342366	-0.001899	-7.83544	0.018565
1	2	1035829975879 51616	-40.287201	2.84328	-1502.5	-0.366049	0.013878	-13.55470	0.035088
2	3	1080880403243 33568	-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

		origin vertex (mm)			momentum (GeV)			charge
	particle_id	vx	vy	vz	px	py	pz	q
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

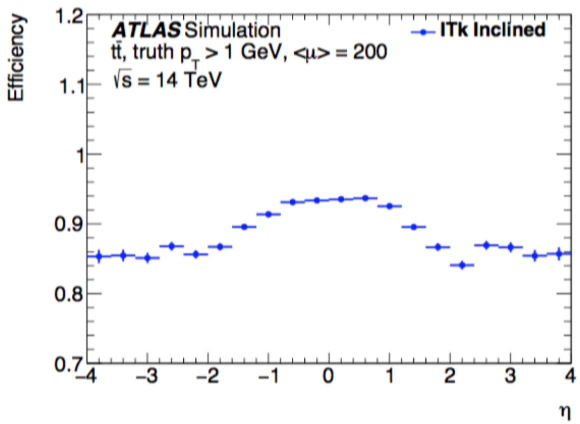
(note : we do not ask participant to reconstruct these track parameters but these could be useful latent variables)

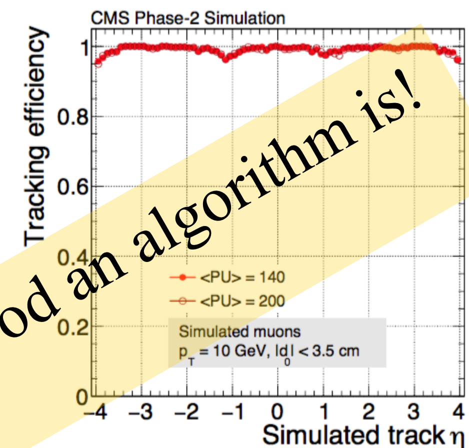
❑ (static)Detector file

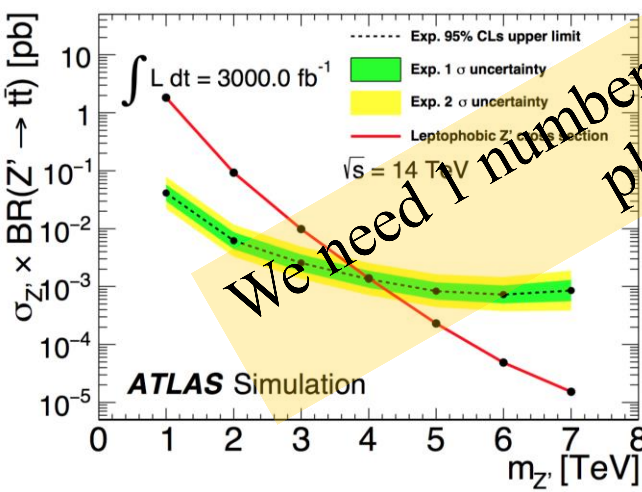
				center position (mm)			3x3 rotation matrix			
	volume_id	layer_id	module_id	cx	cy	cz	rot_xu	rot_xv	rot_xw	ro
0	6	2	1	-65.7965	-5.17830	-1502.5	0.078459	-0.996917	0.0	-0.99
1	6	2	2	-139.8510	-6.46568	-1502.0	0.046183	-0.998933	0.0	-0.99
2	6	2	3	-138.6570	-19.34190	-1498.0	0.138156	-0.990410	0.0	-0.99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

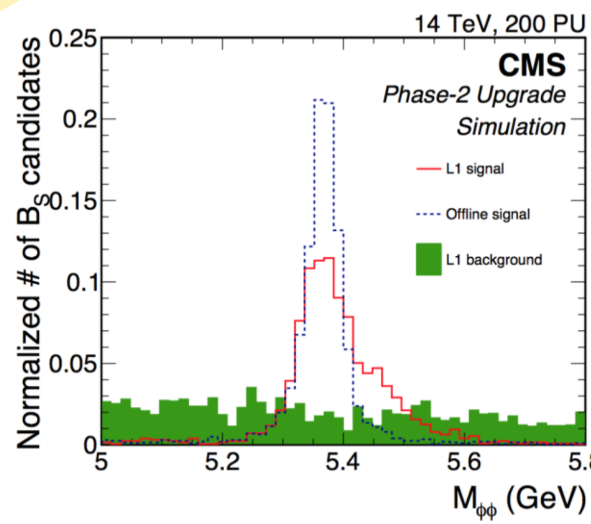
Score

- CMS tracker TDR : Chapter 6 expected performance 31 pages 58 figures
- ATLAS Si strip TDR Chapter 4 ITk Performance and Physics Benchmark Studies 54 pages 80 figures





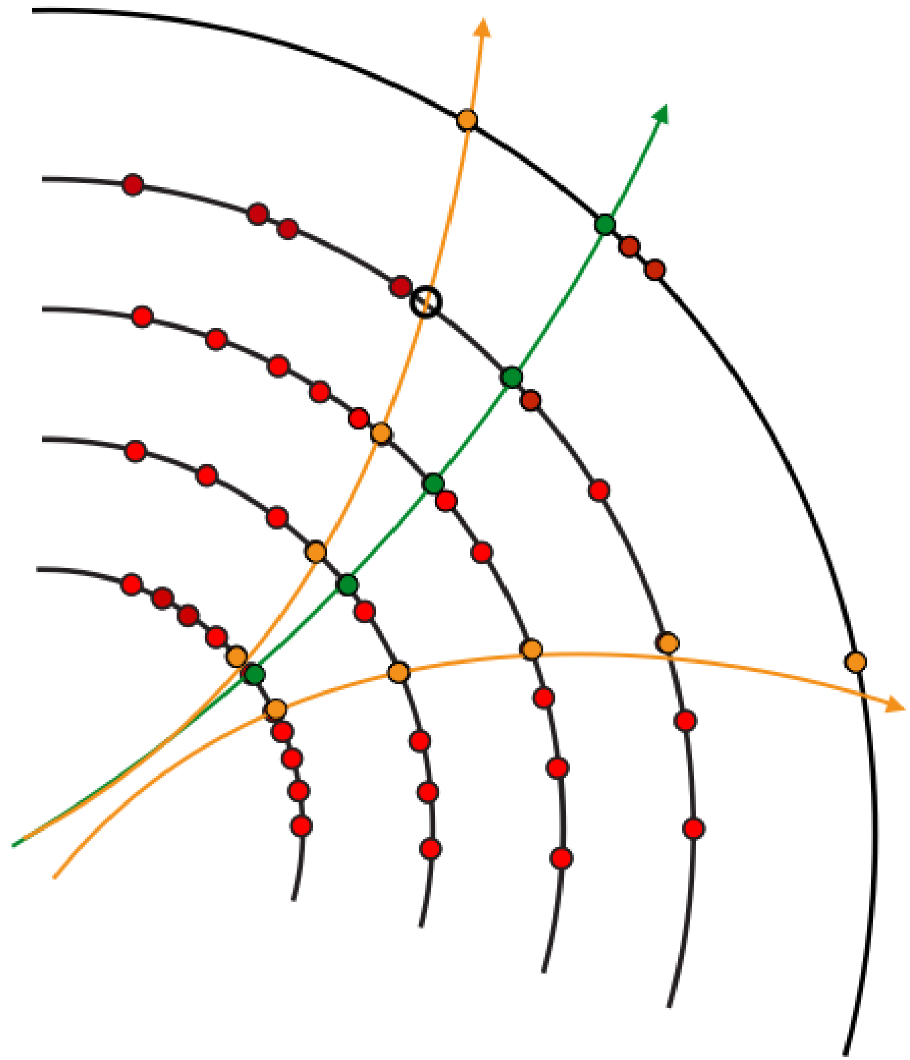




We need 1 number to specify how good an algorithm is!
plus CPU time

Track evaluation

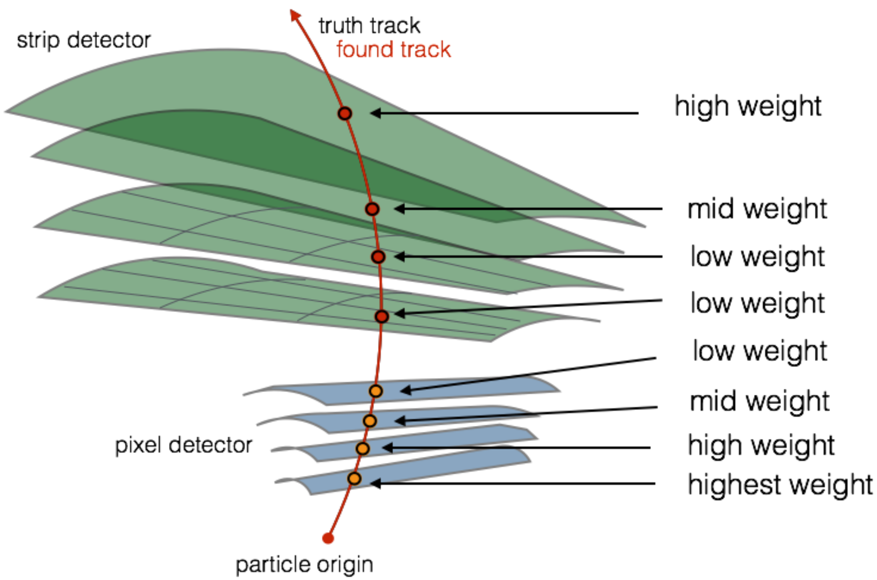
good track	not so good track
many compatible hits	short tracks
completeness	holes
uniqueness	shared hits
low χ^2/ndf	bad fit quality, outliers
small impact parameter (for primaries)	
clusters are compatible	

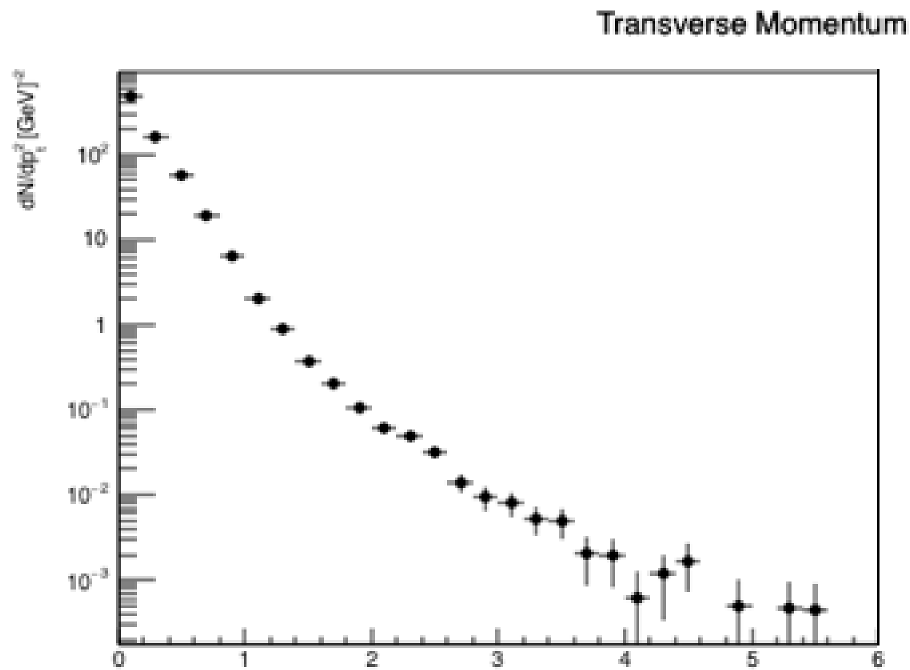


Hit weighting

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

Weighted track score



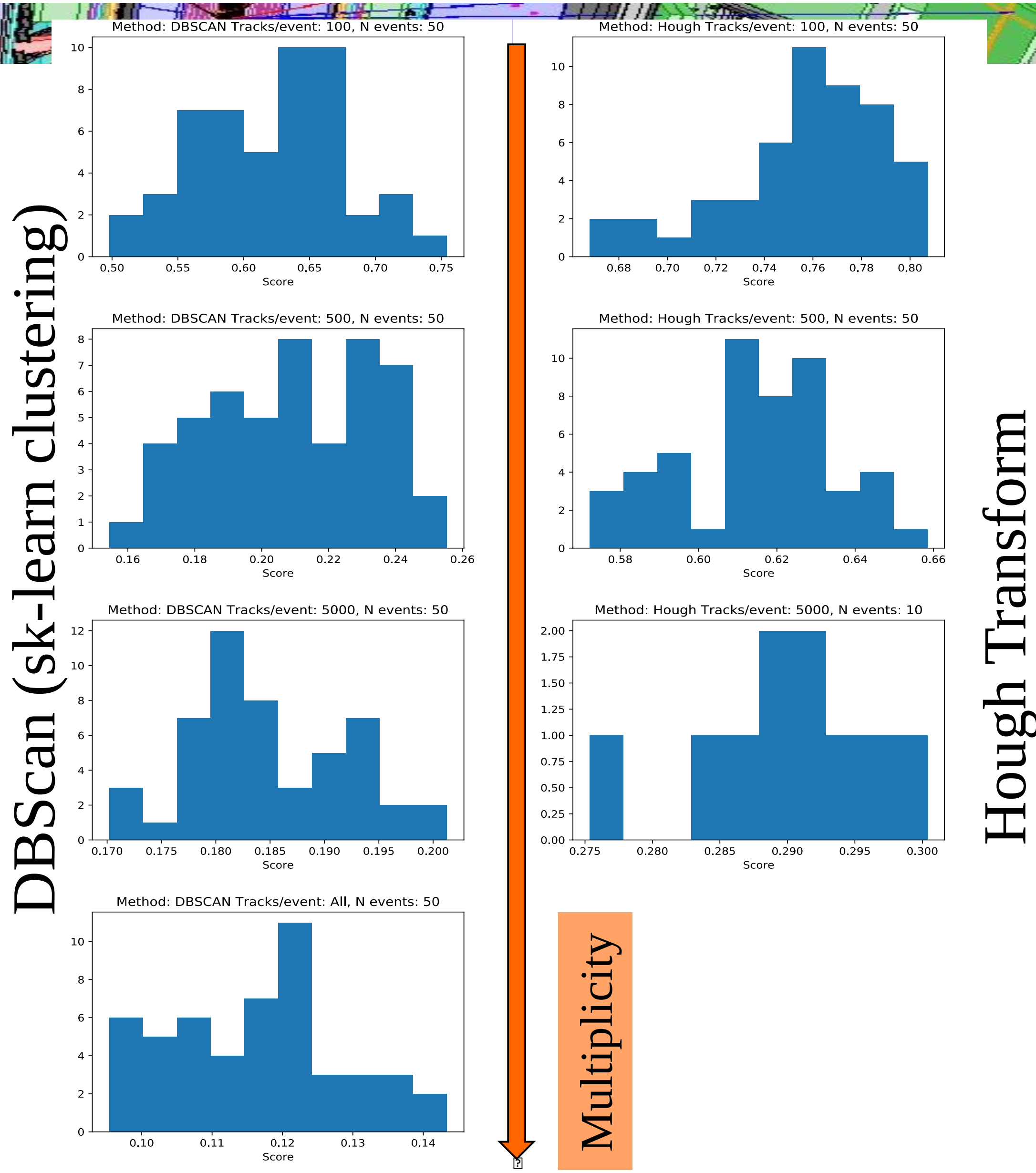


- Weight$_{order}$: more emphasis on first and last hits
- Weight$_{pt}$: more emphasis on high pT tracks
- Weight=0 for noise hits or hits from particle with <=3 hits

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%

Attempt with 2 simple algs



Real life vs challenge

1. Wide type of physics events
2. Full 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)

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 have decided to 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 March-June 2018
  - 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
- Discussion with Kaggle being finalised : they want to run the TrackML challenge and are even ready to sponsor the prize money
- Prizes :
  - From leaderboards of both phases: 8k$ 5k$ 2k$
  - 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

Events

- Challenge Schedules
  - **March to June** Run challenge Accuracy phase
  - **July to October** : Run challenge Throughput phase
- Conference/workshops
  - Connecting The Dots 20-22nd March 2018 Seattle hackathon
  - July 2018 : Accuracy Phase accepted as an official competition for the IEEE World Congress on Computational Intelligence at Rio de Janeiro
  - July 2018 : (submitted) as a talk at CHEP Sofia and ICHEP Seoul
  - December 2018 : (submitted) Throughput Phase as a NIPS 2018 competition and workshop
  - Spring 2019 : grand finale workshop at CERN with prize delivery

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
- HEP physicists more than welcome to participate* : on one's own, or good opportunity to team-up with a friendly ML scientist on your campus !
  - *CERN employees can participate but not claim any price, per Kaggle rule