TrackML Introduction

David Rousseau

LAL-Orsay

rousseau@lal.in2p3.fr @dhpmrou

TrackML Grand Finale workshop

CERN, 1-2 Jul 2019

Who and How

- Organisation: Jean-Roch Vlimant (Caltech), Vincenzo Innocente, Andreas Salzburger (CERN), Sabrina Amrouche, Tobias Golling, Moritz Kiehn (Geneva University), David Rousseau, Yetkin Yilmaz (LAL-Orsay), Paolo Calafiura, Steven Farrell, Heather Gray (LBNL), Vladimir Vava Gligorov (LPNHE-Paris), Laurent Basara, Cécile Germain, Isabelle Guyon, Victor Estrade (LRI-Orsay), Edward Moyse (University of Massachussets), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex, HSE)

Platforms:

kaggle Phase 1

CodaLab Phase 2

accuracy

throughput

sponsors

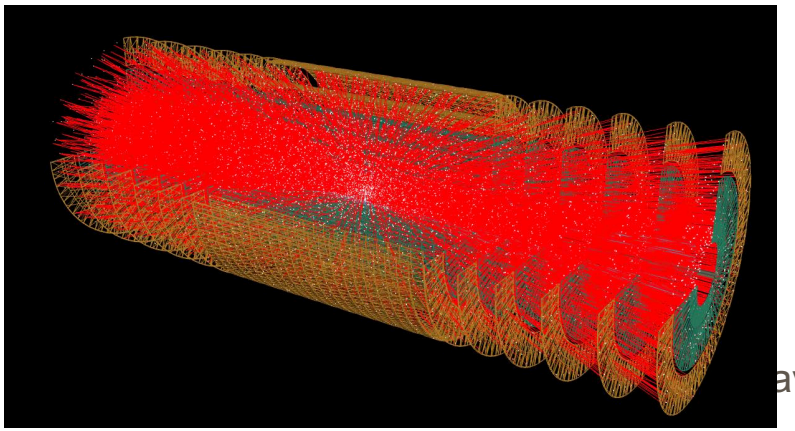
Talk today!

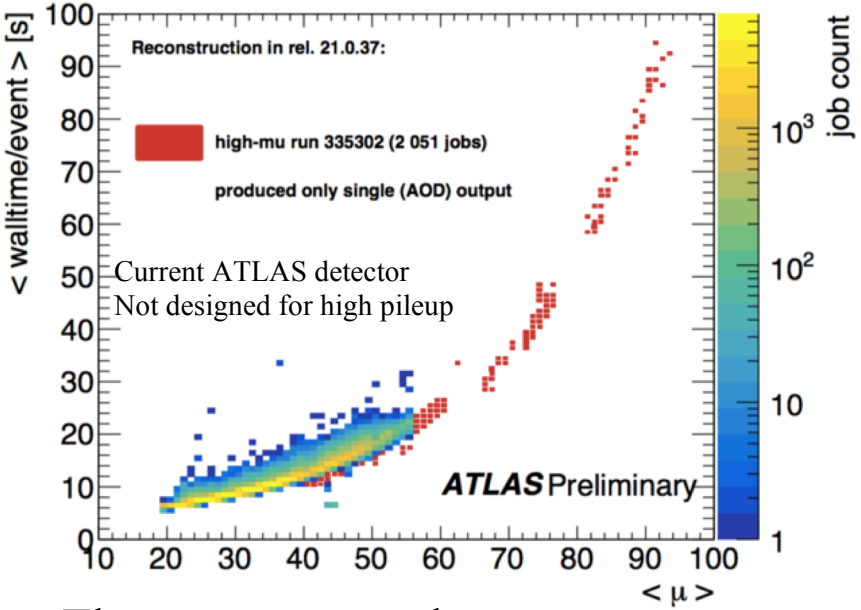
Talk today!

Talk today!

Tracking crisis

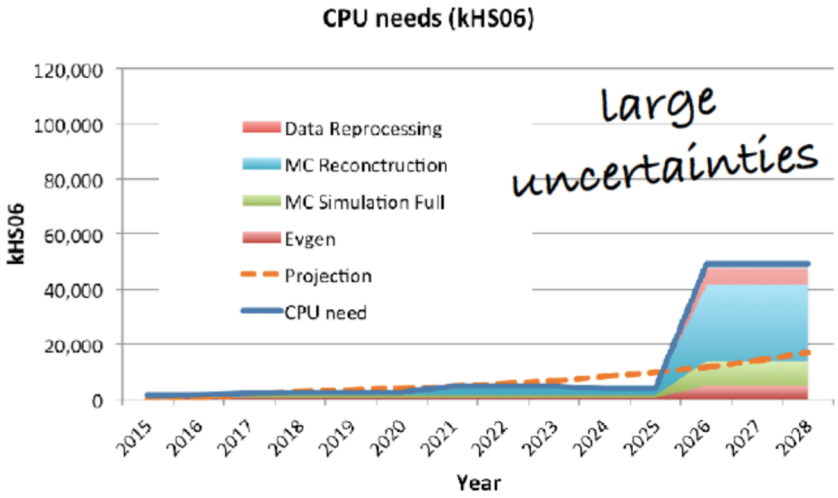
- Tracking (in particular pattern recognition) dominates reconstruction CPU time at LHC
- HL-LHC (phase 2) perspective : increased pileup :Run 1 (2012): <>~20, Run 2 (2015): <>~50,Phase 2 (2025): <>~200
- CPU time quadratic/exponential extrapolation (difficult to quote any number)
- Large effort within 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
- ➔challenge TrackML !!!

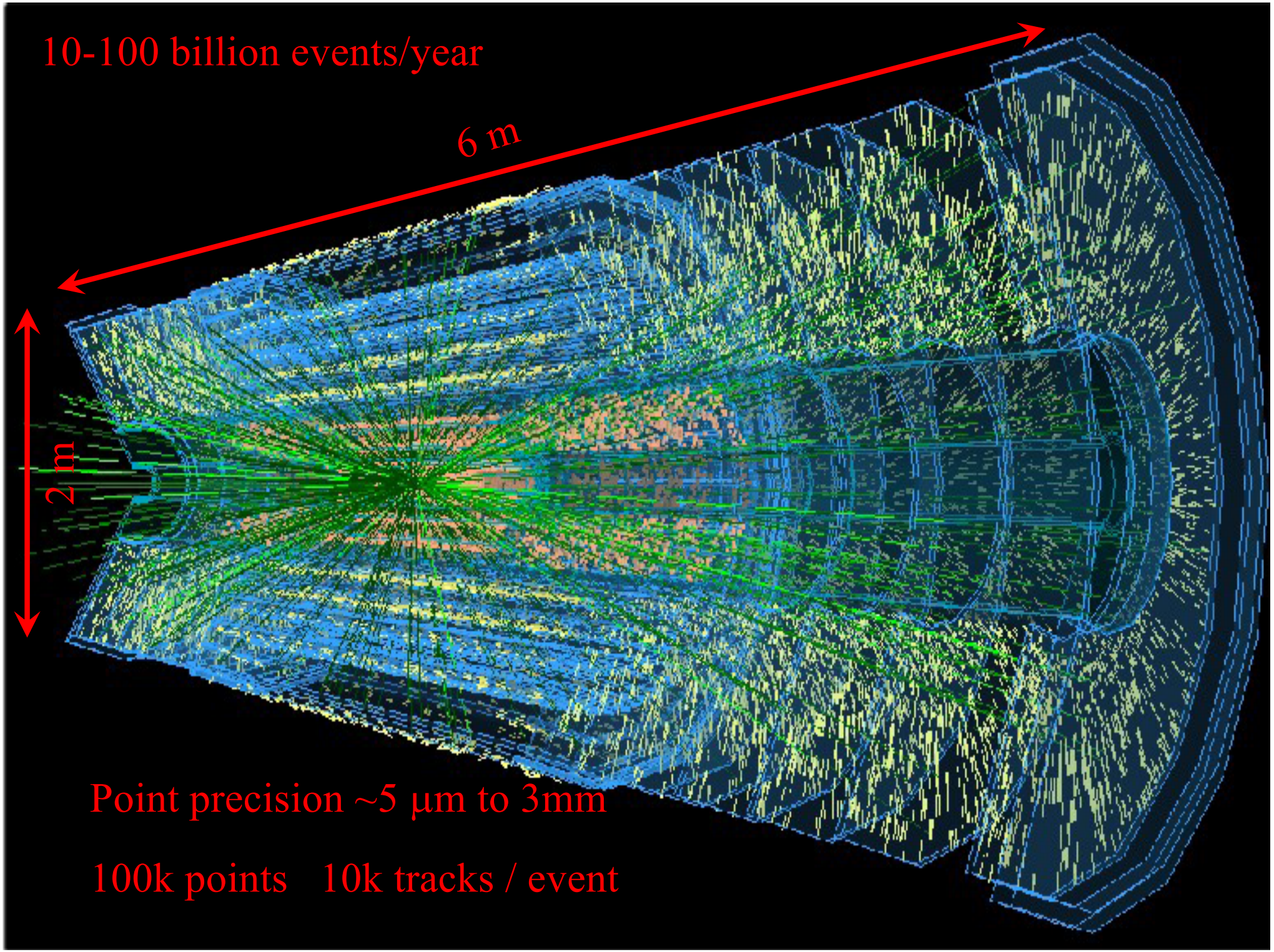




200

There are newer plots, same message




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

From domain to challenge and back

Domain e.g. HEP

Challenge organisation

Challenge

~years

Problem

simplify

Problem

Domain experts solve the domain problem

~ months

The crowd solves the challenge problem

~years

Solution

reimport

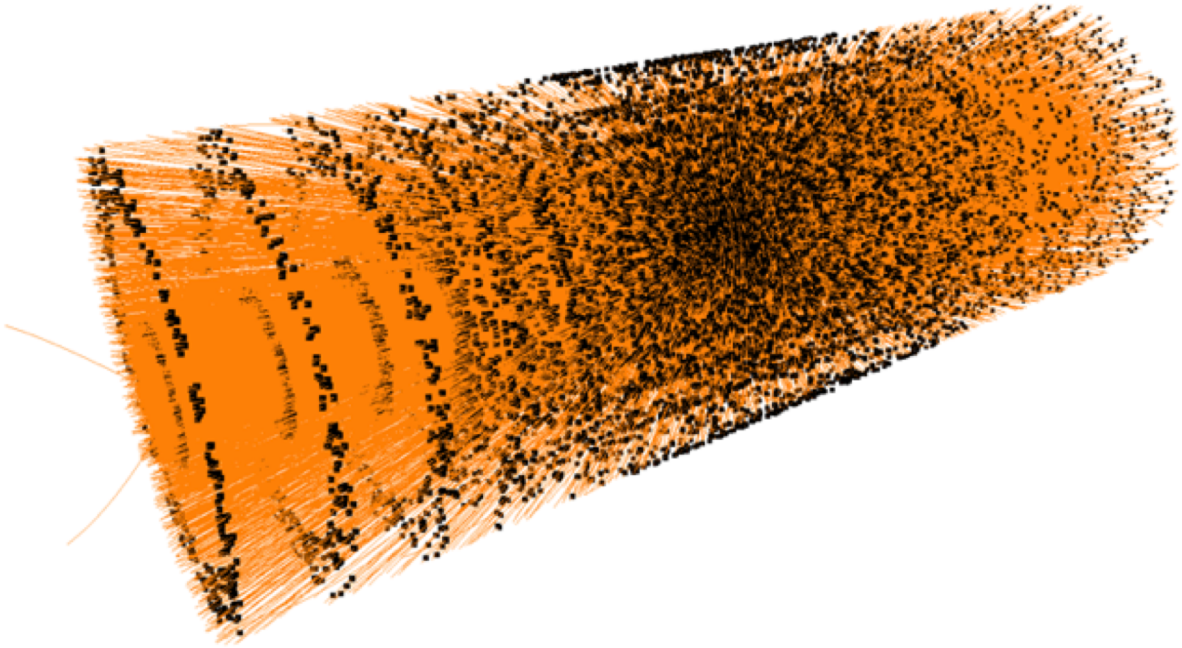
Solution

TrackML 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 10k events, 0.1 billion tracks, 1 billion points, ~100GByte
- Accuracy phase (May to August 2018) on Kaggle
  - 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
  - Score : fraction of points correctly grouped together
  - Evaluation on test sample with per-mille precision on 100 event
- Throughput phase Sep to Mar 2019 on Codalab
  - Strong CPU incentive

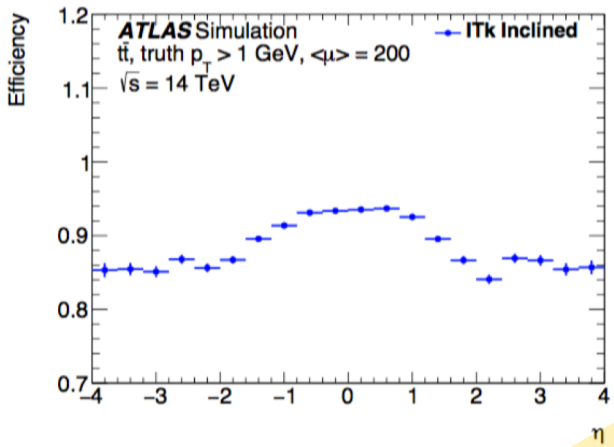
Dataset

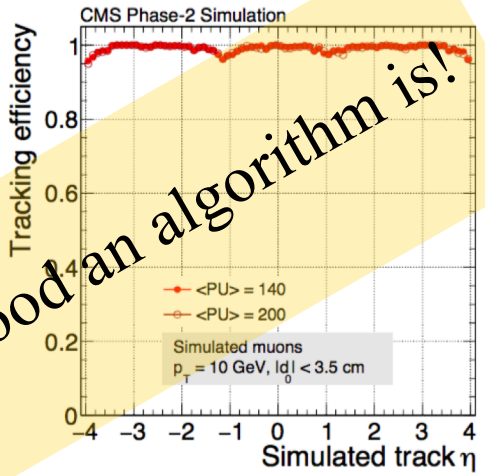
Much more in Andreas' talk

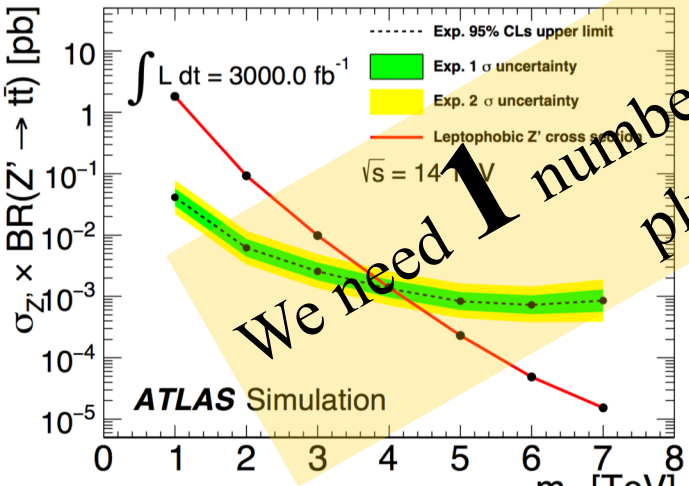


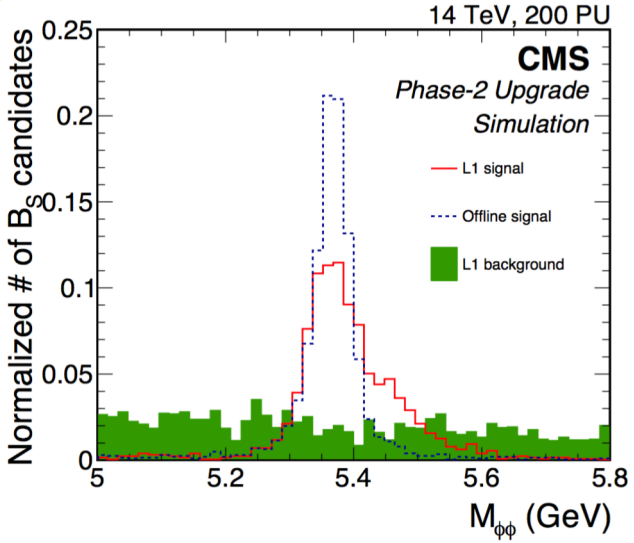
Score

- 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 pages 80 figures







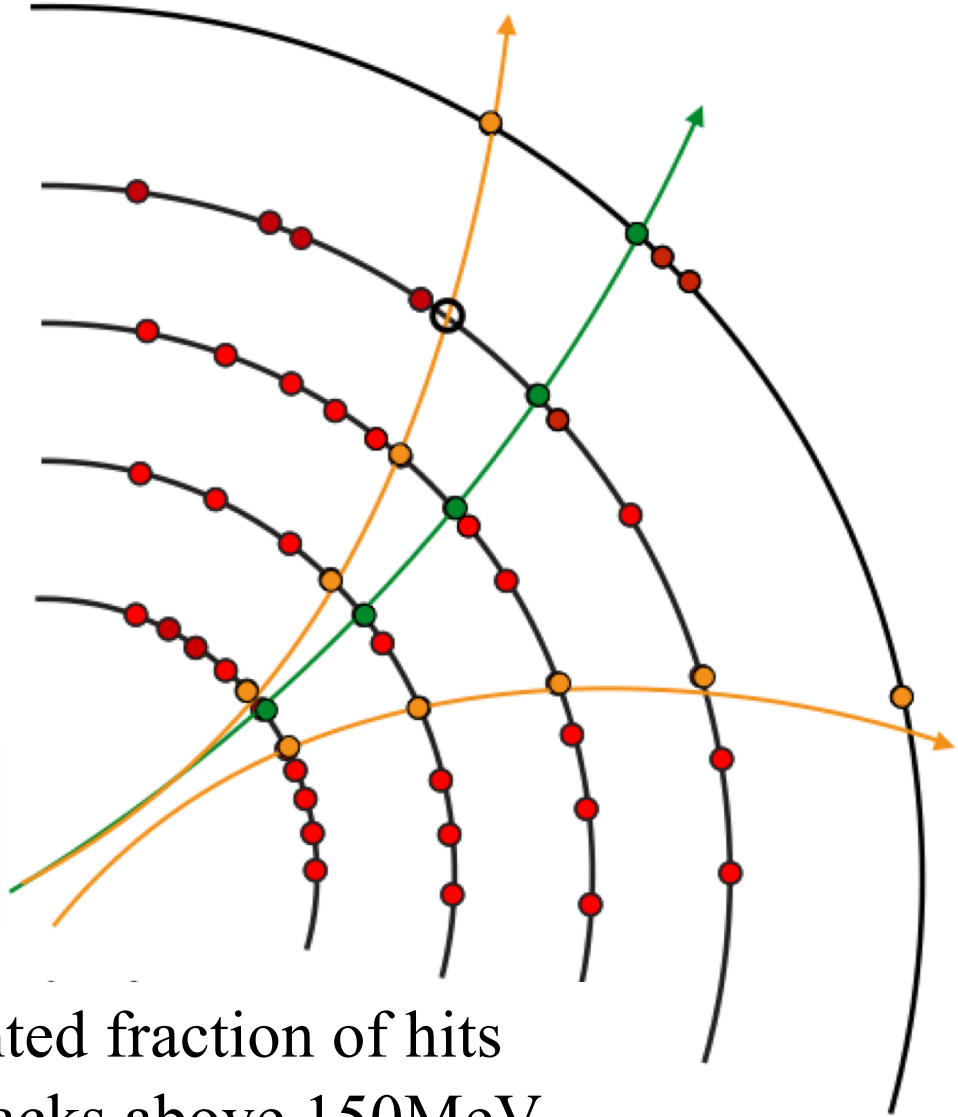


We need **1** number to specify how good an algorithm is!
plus CPU time

Score (2)

HEP tradition : track based 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	



Big decision : score is ~ « the weighted fraction of hits correctly associated ». Include all tracks above 150MeV

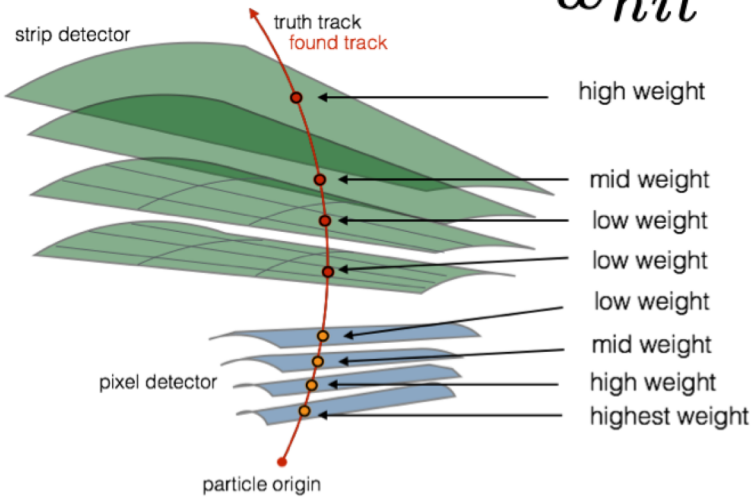
Score in more details

- List intersection hits of reco tracks and true particles
- Intersection should have more than 50% of the hit of the reco track and 50% of the true particle, then:

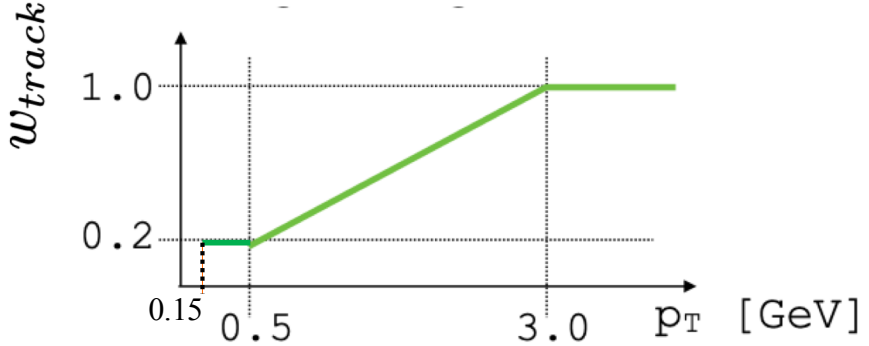
$$score = \frac{1}{N_{evts}} \sum_{evts} \sum_{intersection\ hits} w_{track} \times w_{hit}$$

Reminiscent of $\frac{\cap}{\cup}$

w_{hit}



Perfection : score=1.



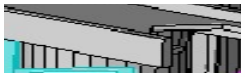
Real life vs challenge

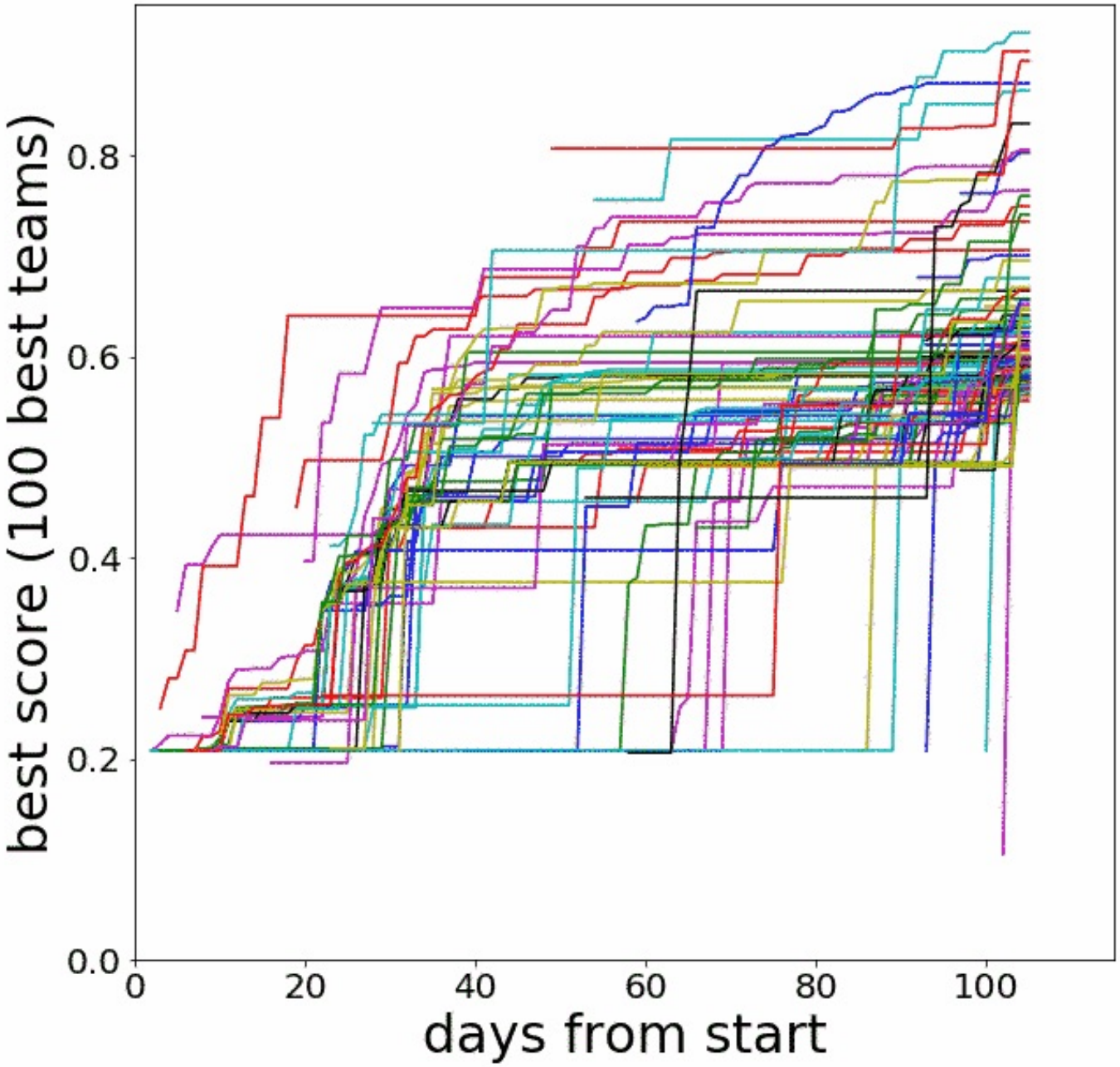
1. Wide type of physics events
2. Full detailed Geant 4 / data
3. Detailed dead matter description
4. Complex geometry (tilted modules, double layers, misalignments…)
5. Hit merging
6. Allow shared hits
7. Output is hit clustering, track parameter and covariance matrix
8. Multiple metrics (see TDR's)

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!

Evolution of leaderboard



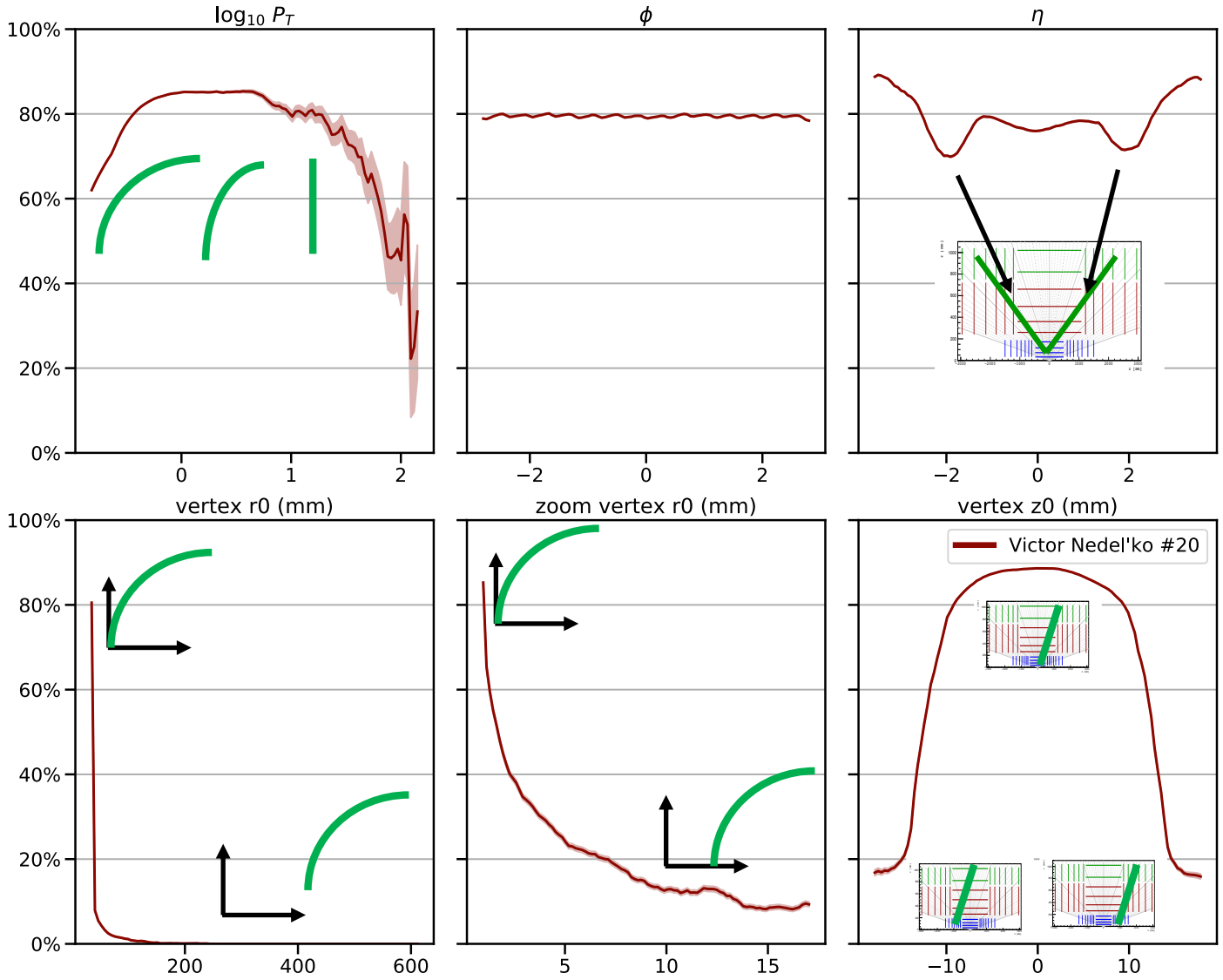


Final Leaderboard

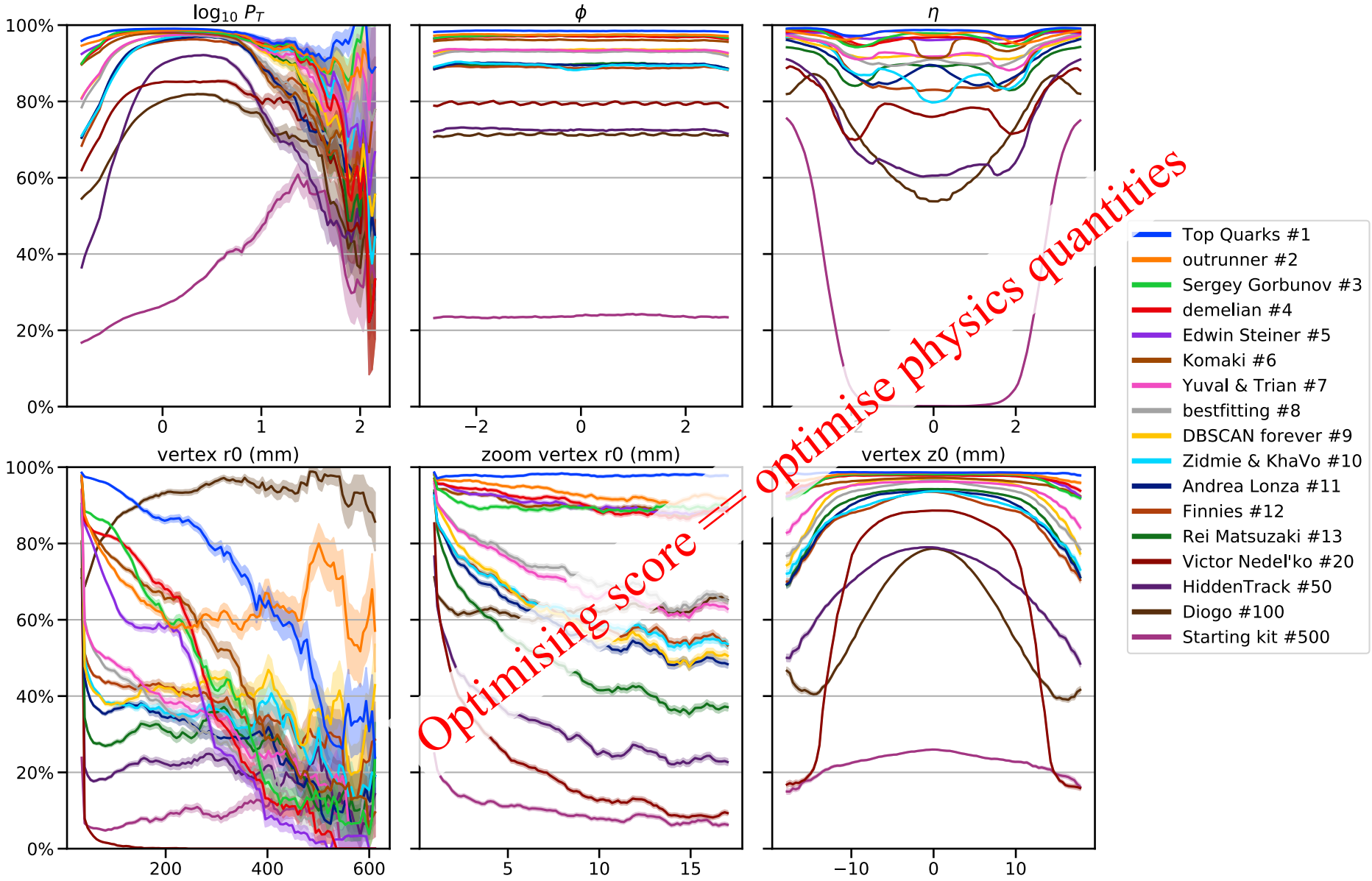
1	—	Top Quarks	Covered by Moritz tomorrow	0.92182	10	19d
2	—	outrunner	Covered by Moritz tomorrow	0.90302	9	18d
3	HEP	Sergey Gorbunov	Talk today	0.89353	6	18d
4	HEP	demelian	Talk today	0.87079	35	1mo
5	—	Edwin Steiner		0.86395	5	18d
6	—	Komaki		0.83127	22	18d
7	—	Yuval & Trian	Talk tomorrow	0.80414	56	18d
8	—	bestfitting		0.80341	6	18d
9	—	DBSCAN forever		0.80114	23	18d
10	—	Zidmie & KhaVo		0.76320	26	18d
11	—	Andrea Lonza		0.75845	15	18d
12	—	Finnies	Talk tomorrow	0.74827	56	18d
13	—	Rei Matsuzaki		0.74035	12	18d
14	—	Mickey		0.73217	10	2mo
15	—	Vicens Gaitan		0.70429	19	1mo
16	—	Robert		0.69955	3	21d
17	—	Yuval-CPMP tribute band		0.69364	20	20d
18	—	N. Hi. Bouzu		0.67573	9	22d
19	—	Steins;Gate		0.66763	12	19d
20	▲1	Victor Nedel'ko		0.66723	4	2mo

Only public LB to private LB rank change

Efficiency #20 Nedel'ko



Efficiency all



Throughput Phase

Now participants submit their software…

… and are evaluated on accuracy AND speed !

Launched 6th Sep 2018 until 12th March 2019 on Codalab

Codalab Schematic

CodaLab

	hit_id	x	y	z	volume_id	layer_id	module
0	1	-64.409897	-7.163700	-1502.5	7	2	1
1	2	-55.336102	0.635342	-1502.5	7	2	1
2	3	-83.830498	-1.143010	-1502.5	7	2	1
3	4	-96.109100	-8.241030	-1502.5	7	2	1

event(s) are loaded in memory

start

API to call python/C++API

User executable

stop

solution

Measure both accuracy score and evaluation time

VM 2 cores, 4 Gb memory

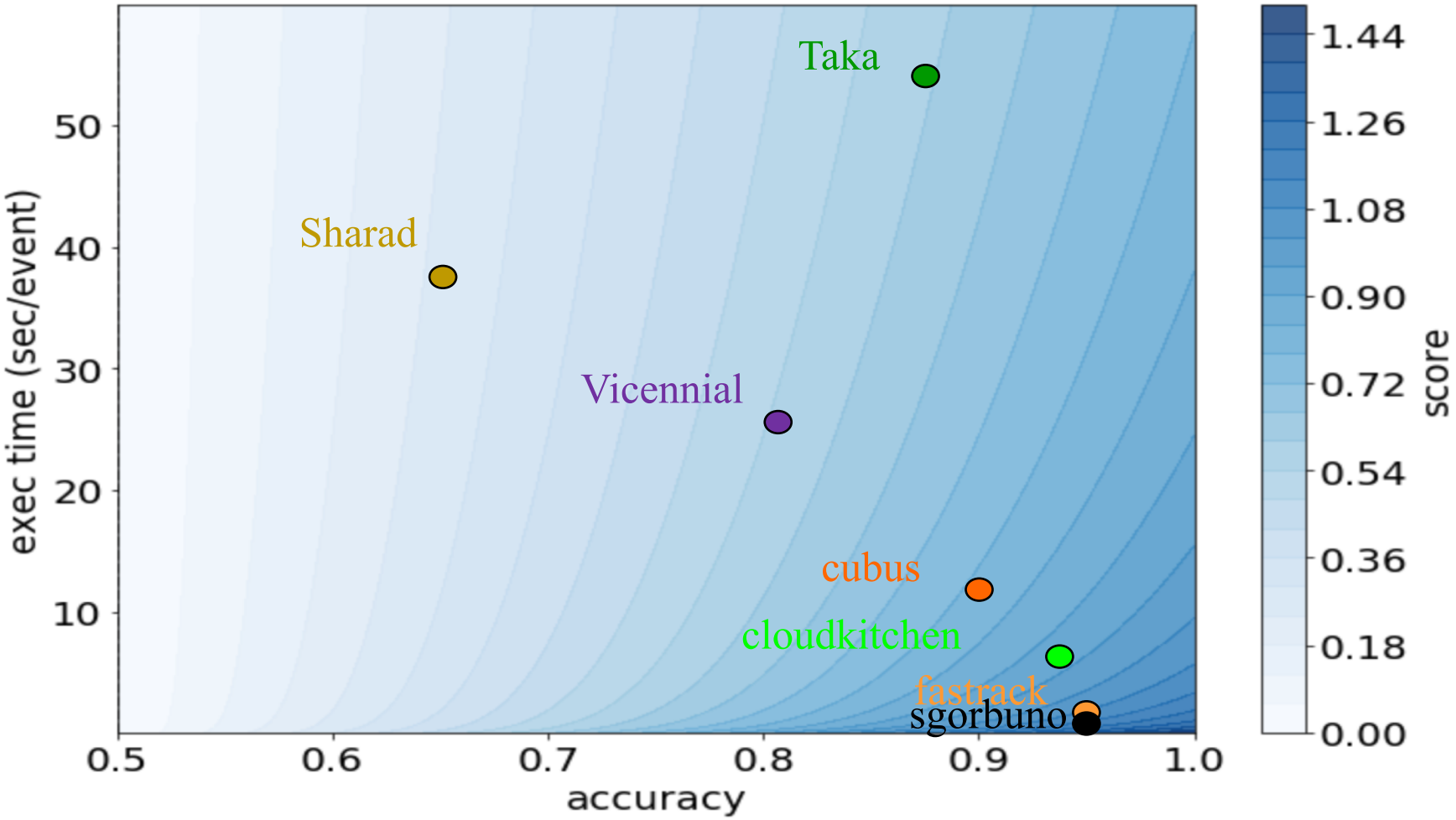
More in Victor's talk tomorrow

Updated dataset

- Fixed bug in electron multiple scattering (was much too large)
- …and layer material thickness (was twice too large)
- Changed the scoring to only include primary track (no secondaries originating far from the origin)

Throughput results

- *Ranking score including time* :
  - 0 if time >600 s or accuracy <50%
  - $\sqrt{log}(1 + 600/time) * (accuracy - 0.5)^2$



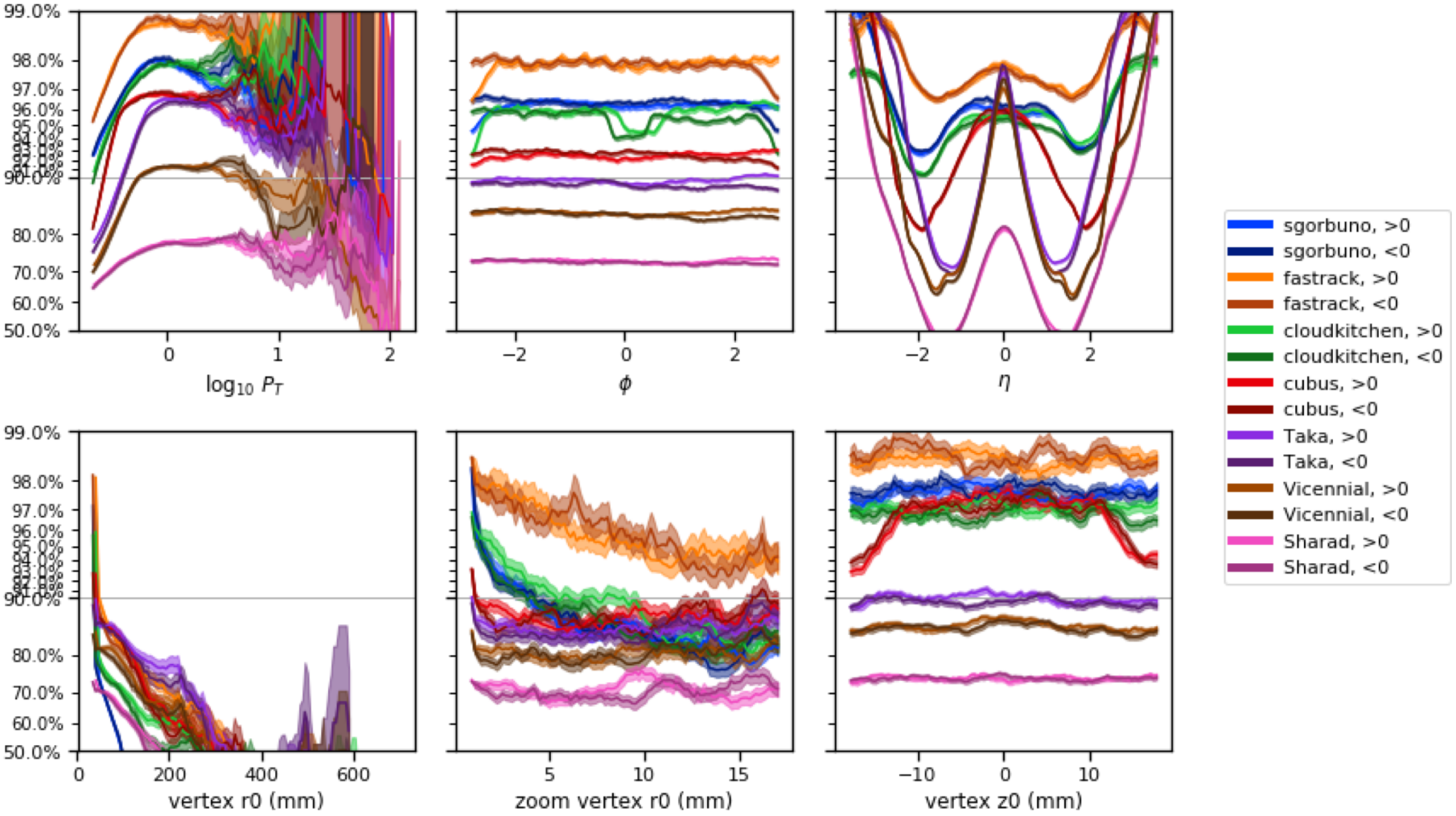
Throughput phase LB

RESULTS

Private leaderboard

#	User	Entries	Date of Last Entry	score ▲	accuracy_mean ▲	accuracy_std ▲	computation time (sec) ▲	computation speed (sec/event) ▲	Duration ▲
1	sgorbuno HEP	9	Talk today	1.1727 (1) 1.16	0.944 (2) 0.943	0.00 (14)	28.06 (1)	0.56 (1) 0.60	64.00 (1)
2	fastrack HEP	53	Talk today	1.1145 (2) 1.12	0.944 (1) 0.944	0.00 (15)	55.51 (16)	1.11 (16) 1.00	91.00 (6)
3	cloudkitchen	73	Talk today	0.9007 (3) 0.897	0.928 (3) 0.927	0.00 (13)	364.00 (18)	7.28 (18) 7.41	407.00 (8)
4	cubus	8	09/13/18	0.7719 (4) 0.770	0.895 (4) 0.895	0.01 (9)	675.35 (19)	13.51 (19) 13.7	724.00 (9)
5	Taka	11	01/13/19	0.5930 (5)	0.875 (5)	0.01 (12)	2668.50 (23)	53.37 (23)	2758.00 (13)
6	Vicennial	27	02/24/19	0.5634 (6)	0.815 (6)	0.01 (10)	1270.73 (20)	25.41 (20)	1339.00 (10)
7	Sharad	57	03/10/19	0.2918 (7)	0.674 (7)	0.02 (4)	1902.20 (22)	38.04 (22)	1986.00 (12)
8	WeizmannAI	5	03/12/19	0.0000 (8)	0.133 (11)	0.01 (11)	88.08 (17)	1.76 (17)	124.00 (7)
9	harshakoundinya	2	03/12/19	0.0000 (8)	0.085 (13)	0.01 (6)	49.22 (8)	0.98 (8)	86.00 (3)
10	iWit	6	03/10/19	0.0000 (8)	0.082 (15)	0.01 (8)	48.23 (3)	0.96 (3)	85.00 (2)
				0.0000					

Efficiency



Not quite as good as for accuracy phase: do not spend time when not worth it

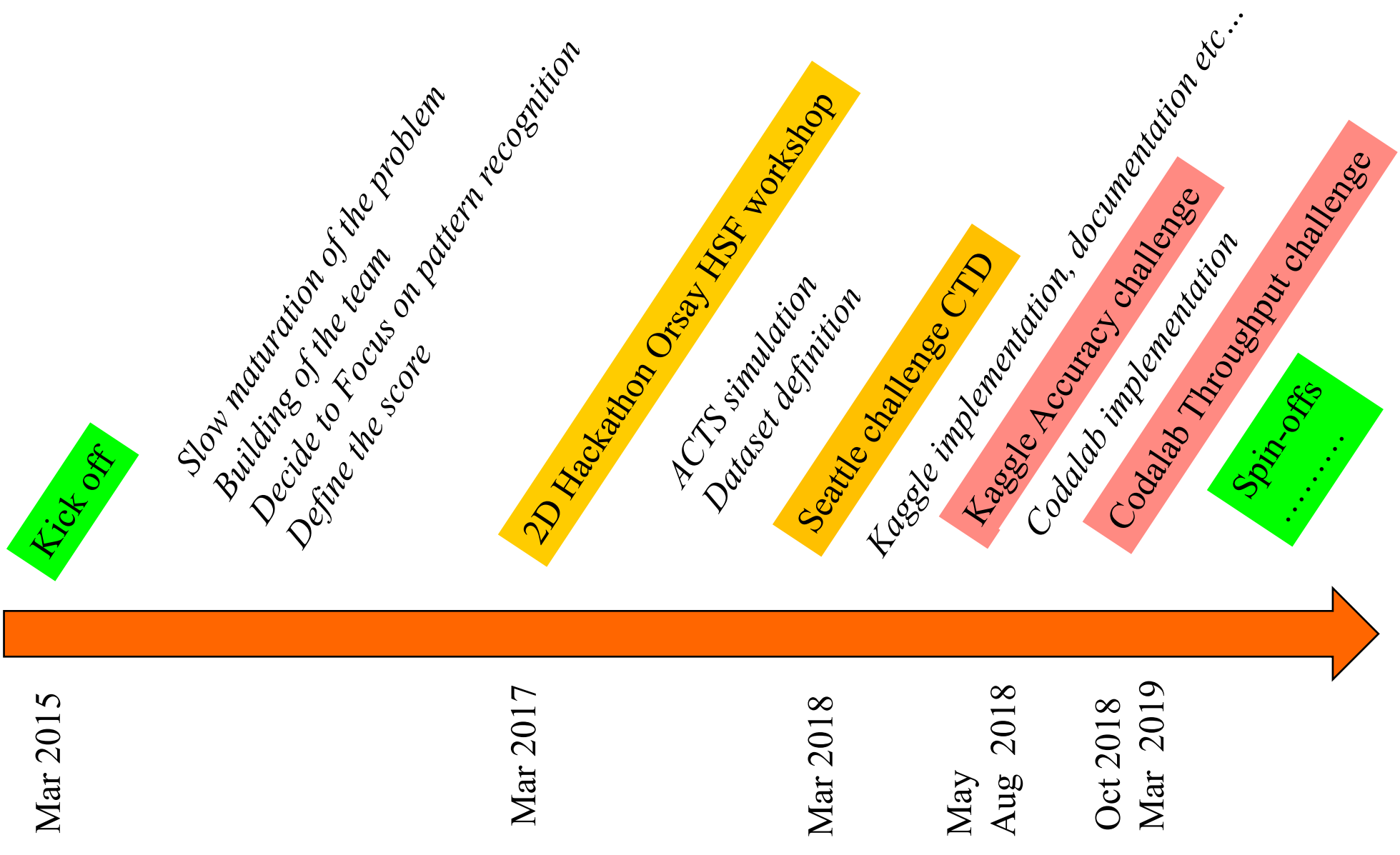
Useful links

- Contact : trackml.contact@gmail.com https://sites.google.com/site/trackmlparticle Twitter : @trackmllhc
- Accuracy phase @ Kaggle : https://www.kaggle.com/c/trackml-particle-identification
  - ➔chapter in the NeurIPS 2018 Competition book arXiv:1904.06778 , small revision on-going
- Throughput phase @ Codalab : https://competitions.codalab.org/competitions/20112
  - Write-up to be finalised

TrackML Conference talks

- Connecting The Dots 2015 Seattle
- Connecting The Dots 2016 Vienna
- CHEP 2016 Okinawa
- Connecting The Dots / Intelligent Trackers 2017 Orsay
- NeurIPS 2017 Los Angeles CiML workshop
- Connecting The Dots 2018 Seattle
- CHEP 2018 Sofia
- WCCI 2018 Rio de Janeiro
- ICHEP 2018 Seoul
- IEEE NSSMIC 2018 Sidney
- IEEE eScience 2018 Amsterdam
- NeurIPS 2018 Montreal Competition workshop
- ACAT 2019 Saas-Fe
- Connecting The Dots 2019 Valencia
- EPS 2019 Ghent
- CHEP 2019 Adelaïde
- …and much more workshops and seminars….

TrackML timeline



Monday 1st July

13:30 → 13:50 **Introduction : overview of the problem and the two phases of the challenge** 20m
Speaker: David Rousseau (LAL-Orsay, FR)

13:55 → 14:15 **Nvidia sponsor talk** 20m
Speaker: Peter Messmer (Nvidia)

14:20 → 14:40 **TrackML open dataset and CTD/WIT Valencia highlights** 20m
Speaker: Andreas Salzburger (CERN)

14:45 → 15:05 **TrackML Throughput #1 (and Accuracy #3)** 20m
Speaker: Sergey Gorbunov (Johann-Wolfgang-Goethe Univ. (DE))

15:10 → 15:30 **TrackML Throughput #2 (and accuracy #4)** 20m
Speaker: Dmitry Emeliyanov (Science and Technology Facilities Council STFC (GB))
TheTrackML_works... TheTrackML_works...

15:35 → 15:55 **Coffee break** 20m

16:00 → 16:20 **TrackML Throughput #3** 20m
Speaker: Marcel Kunze (Heidelberg University)

16:25 → 16:45 **Spin-off : TrackML Hololens visualisation** 20m
Speakers: Tobias Isenberg (Inria), Xiyao Wang

16:50 → 17:10 **UniGe sponsor talk : Hardware efficient meshes in computational fluid dynamics** 20m
Speaker: Jonas Latt (Université de Genève)

17:15 → 17:35 **Kaggle sponsor talk (remote)** 20m
Speaker: Walter Reade (Kaggle)

17:40 → 18:00 **Spin-off : Hep.QPR/QMLQCF tracking with quantum computing with TrackML dataset** 20m
Speaker: Jean-Roch Vlimant (California Institute of Technology (US))

Tuesday 2nd July

13:30 → 13:50 **Throughput accuracy : overview of solutions not covered** — 20m
Speaker: Moritz Kiehn (Universite de Geneve (CH))

13:55 → 14:10 **Trackml challenge implementation in Codalab** — 15m
Speaker: Mr Victor Estrade (LRI)

14:15 → 14:35 **Throughput accuracy HepML prize (Yuval and Reina)** — 20m
Speaker: Yuval Reina

14:40 → 15:00 **TrackML Accuracy NeurIPS invite : LSTM by the Finnies** — 20m
Speakers: Liam Finnie, Nicole Finnie (nicole.lin@gmail.com)

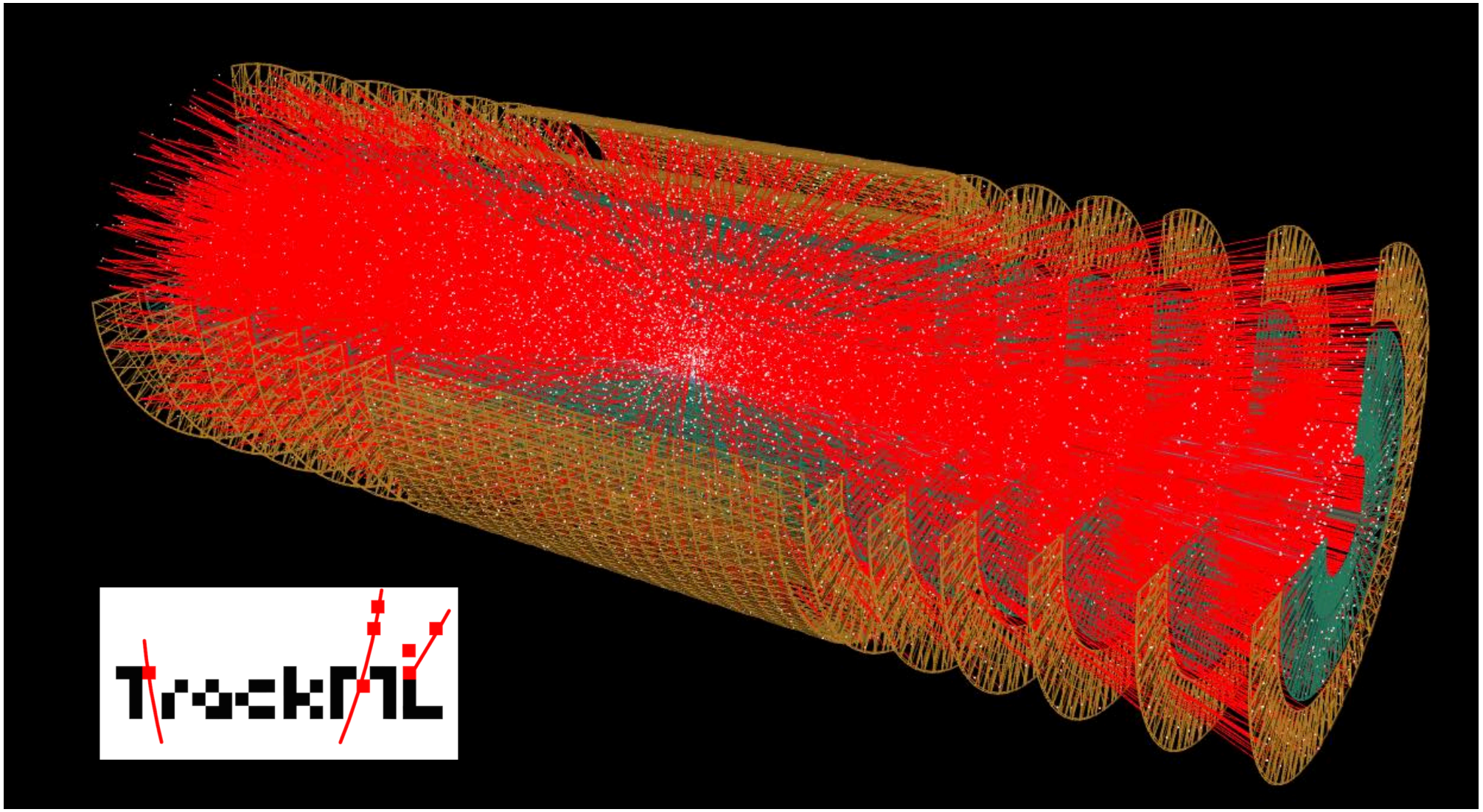
15:05 → 15:25 **Spin-off Similarity hashing and learning for tracks reconstruction** — 20m
Speaker: Sabrina Amrouche (Université de Geneve (CH))

15:30 → 15:50 **Spin-off : Hep.TrkX** — 20m
Speaker: Jean-Roch Vlimant (California Institute of Technology (US))

15:55 → 16:15 **Coffee break** — 20m

16:15 → 16:35 **Future usage of TrackML dataset** — 20m
Speaker: Andreas Salzburger (CERN)

16:40 → 17:30 **FInal discussion** — 50m



TrackML Grand Finale workshop

CERN, 1st-2nd July 2019