TrackML : a Tracking Machine Learning Challenge

ACAT, Saas-fee, 10-15 March 2019

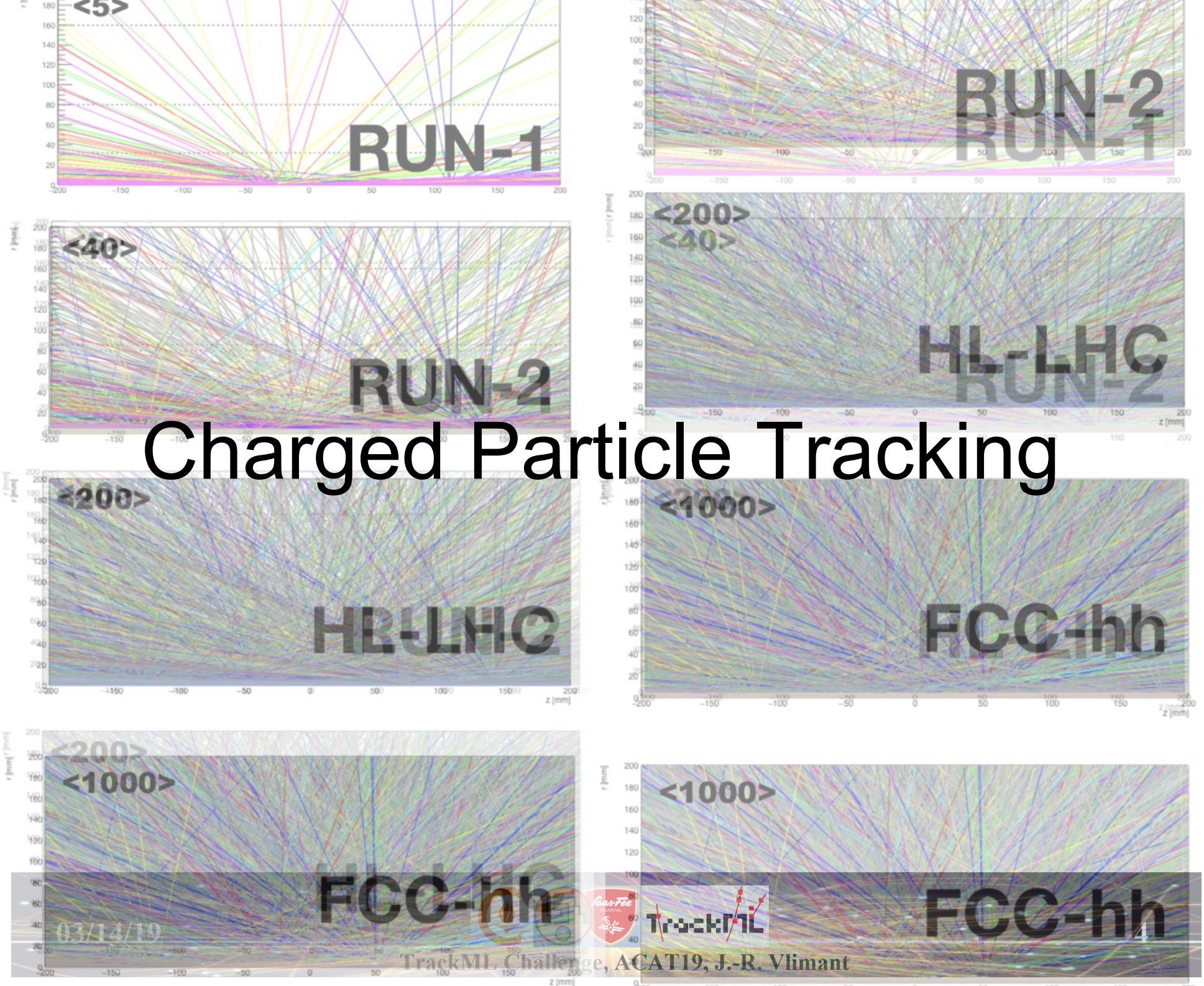
Jean-Roch Vlimant (Caltech), Vincenzo Innocente, Andreas Salzburger (CERN), Isabelle Guyon (ChaLearn), Sabrina Amrouche, Tobias Golling, Moritz Kiehn (Geneva University), David Rousseau, Yetkin Yilmaz (LAL-Orsay), Paolo Calafiura, Steven Farrell, Heather Gray (LBNL), Vladimir Vava Gligorov (LPNHE-Paris),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

sponsors

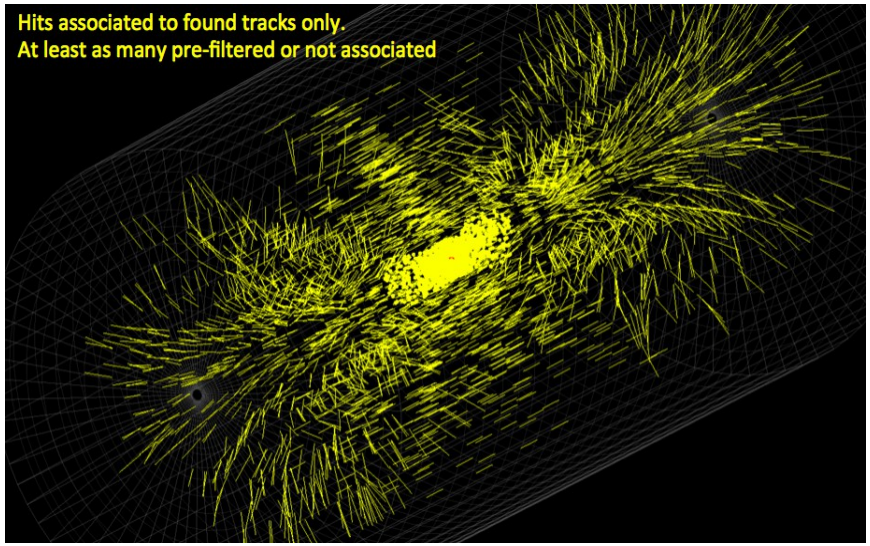
Outline

- Forewords on Charged Particle Tracking and Machine Learning
- A Tracking with Machine Learning Challenge
- Accuracy and Throughput Phases
- Lessons Learned and outlook

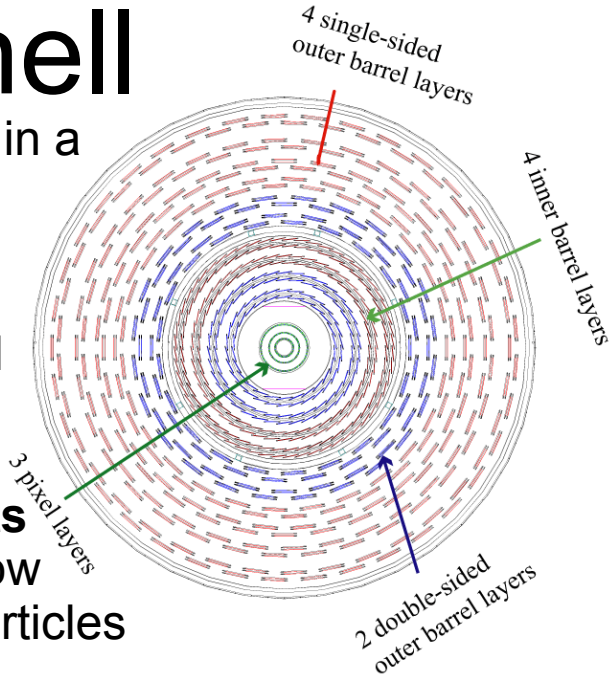
Charged Particle Tracking



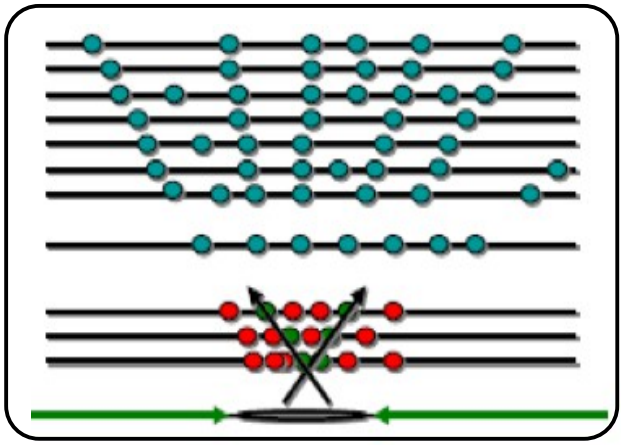
Tracking in a Nutshell



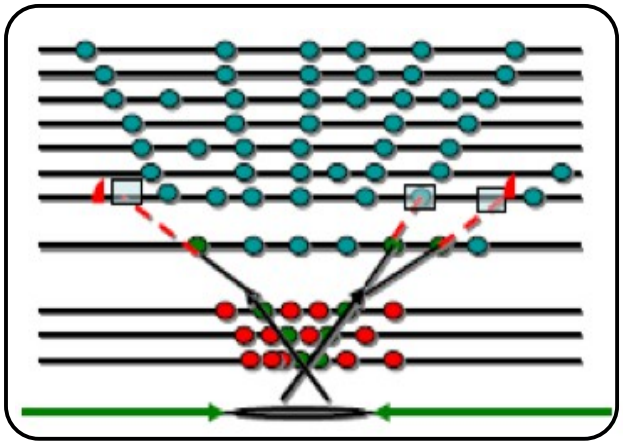
- Particle trajectory bended in a solenoidal magnetic field
- Curvature is a proxy to momentum
- Particle ionize silicon pixel and strip throughout several concentric layers
- **Thousands of sparse hits**
- Lots of hit pollution from low momentum, secondary particles

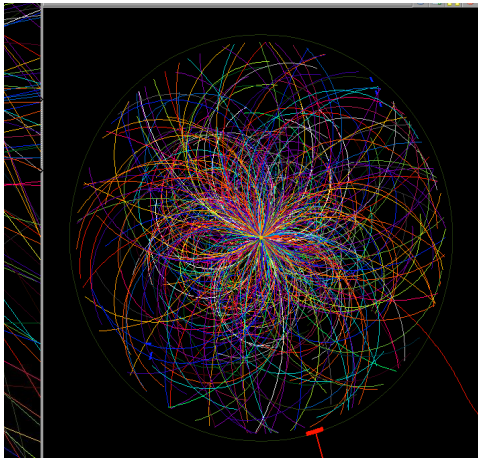


Seeding



Kalman Filter

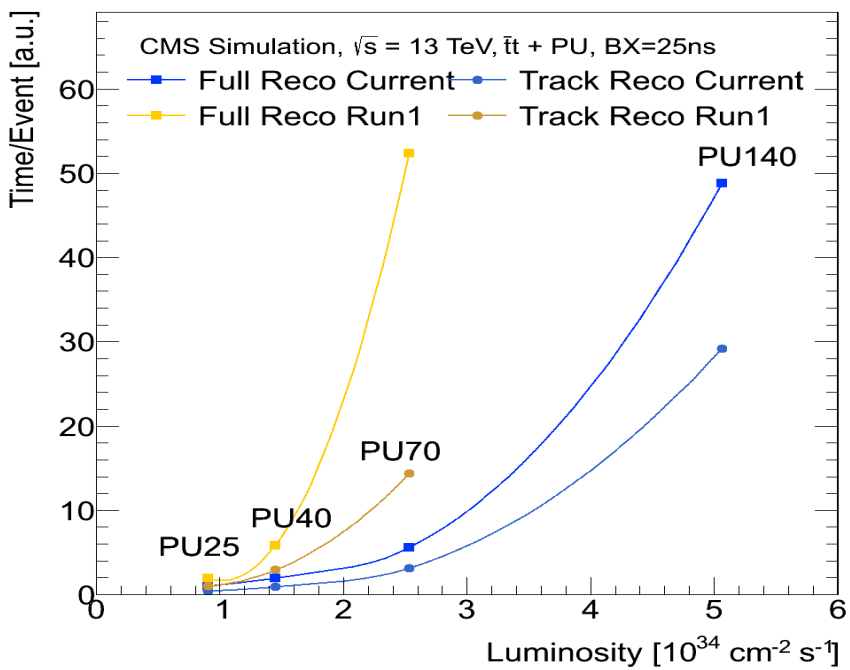


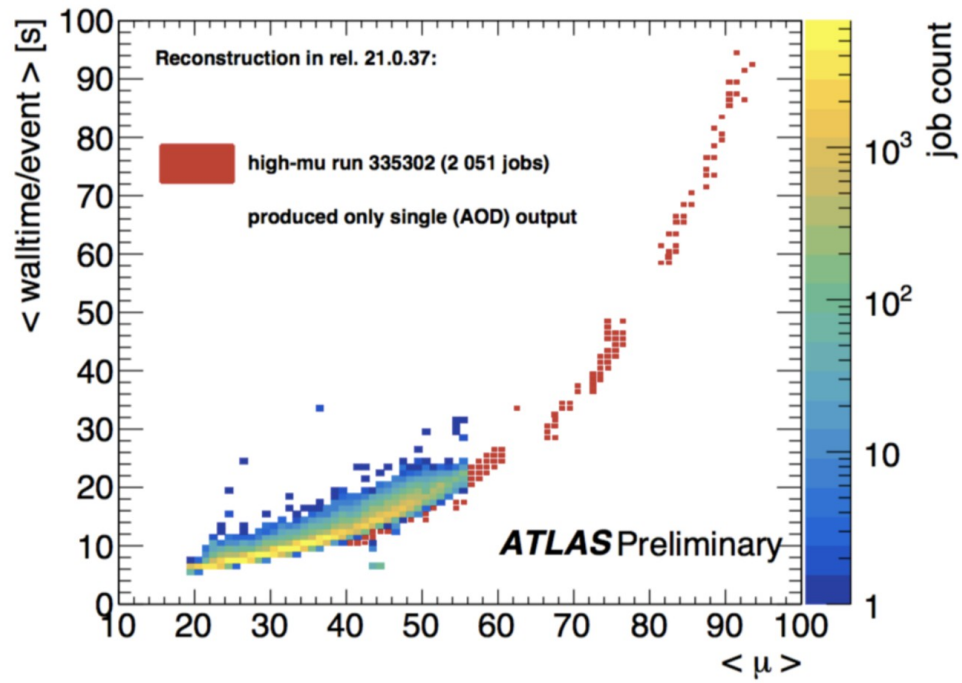


- **Explosion in hit combinatorics** in both seeding and stepping pattern recognition
- **Highly computing consuming task** in extracting physics content from LHC data

Cost of Tracking

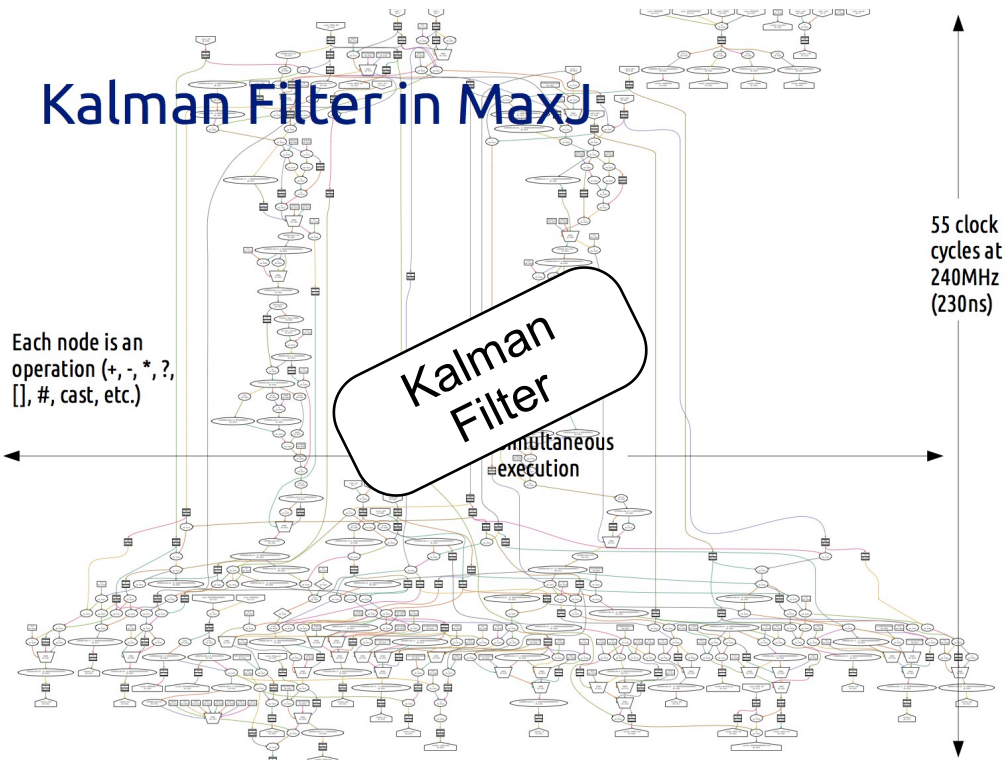
- Charged particle track reconstruction is one of the **most CPU consuming task** in event reconstruction
- Future computing budget flat at best
- Optimizations (to fit in computational budgets) **mostly saturated** and **long way to go** for HL-LHC
- Need factor **10-100 speed-up**

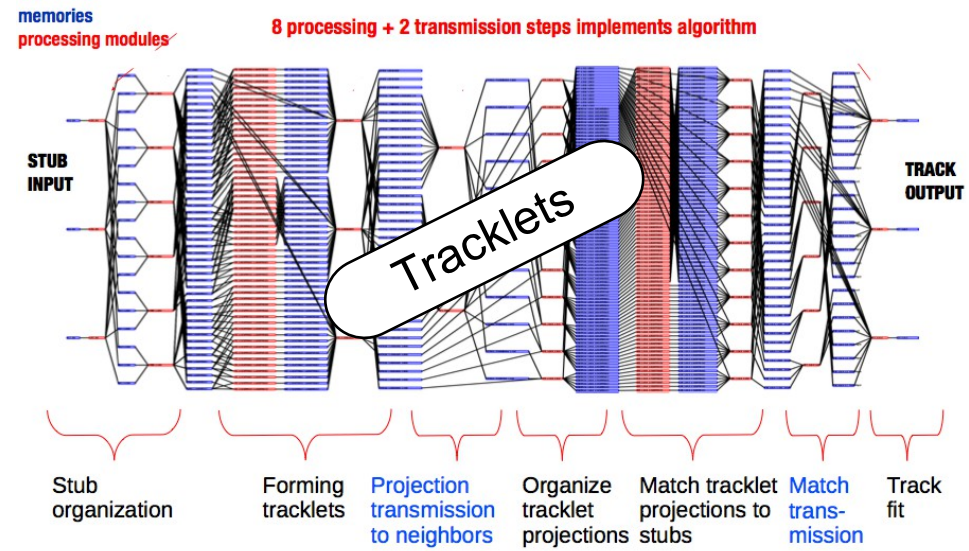




Fast Hardware Tracking

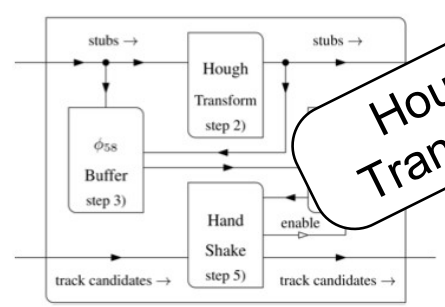
- Track trigger implementation for Trigger upgrades development on-going
- Several approaches investigated
- **Dedicated hardware is the key** to fast computation.
- **Not applicable for offline** processing unless by adopting heterogeneous hardware.





Firmware Implementation - Bin

- Each bin represents a $^{q}/_{p_T}$ column in the HT array



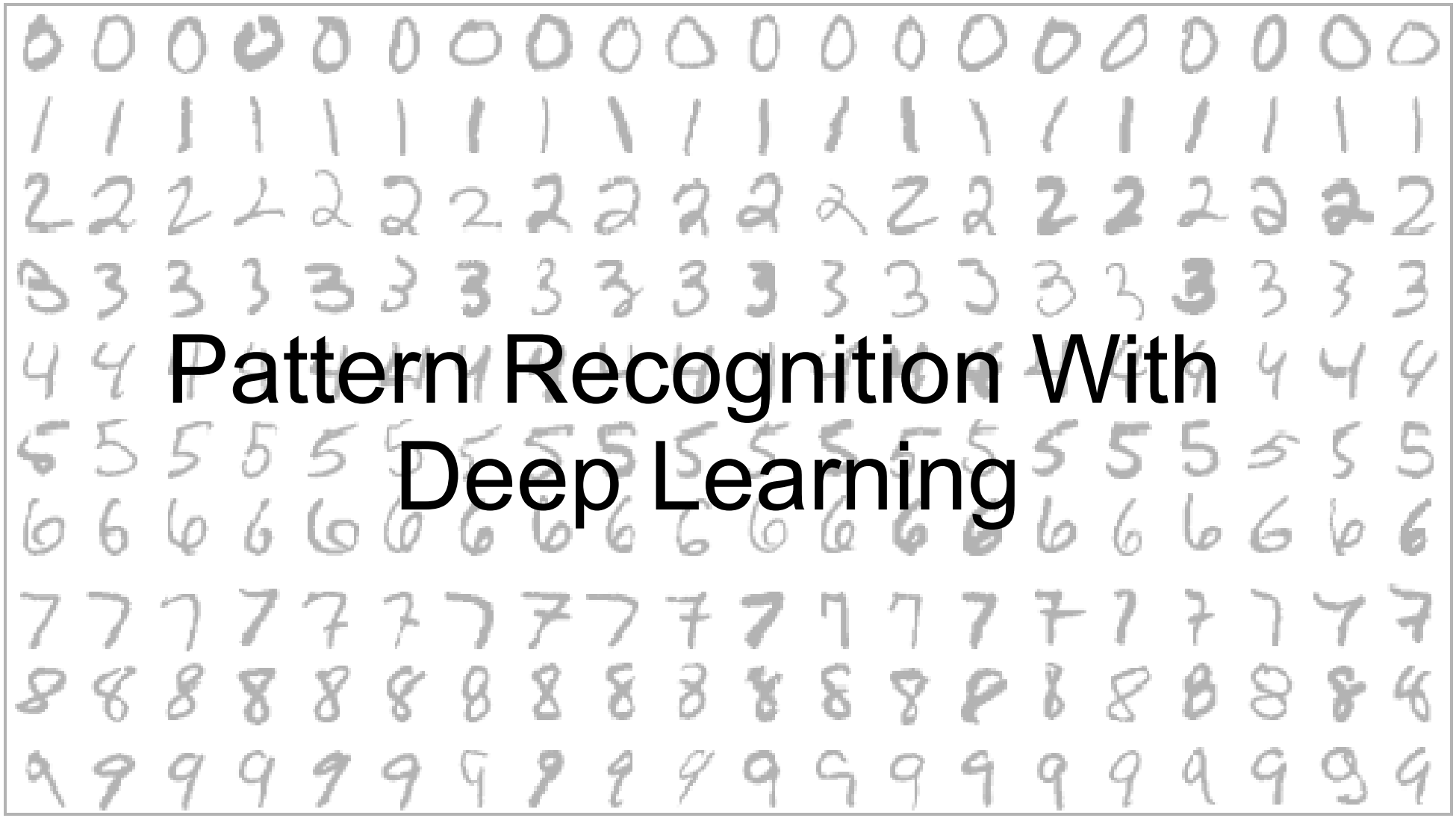
Transform:
- s ϕ_{58} at left boundary
- ulates ϕ_{58} at right boundary

er:
- Duplicates stubs if it belongs to two cells.

- Track Builder:
  - Sorts stubs in ϕ_{58} cells.
  - Marks ϕ_{58} cells with stubs in at least 4/5 layers.
- Hand Shake:
  - Controls read-out of candidates

See https://ctdwit2017.lal.in2p3.fr/

Pattern Recognition With Deep Learning



Scene Labeling

LeCun Seminar at CERN

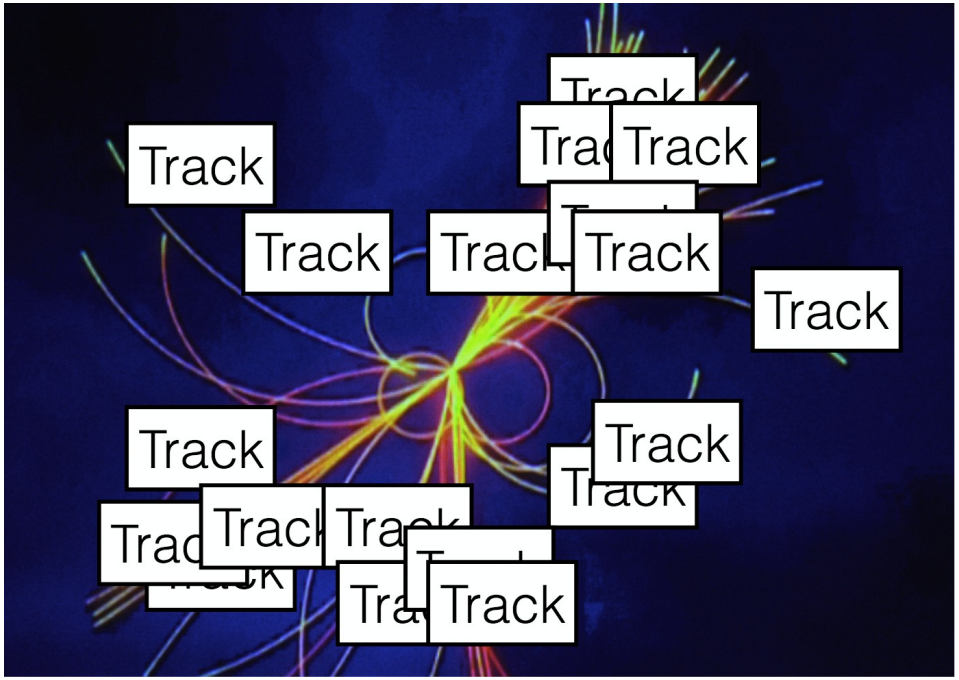


Photo by Pier Marco Tacca/Getty Images

- There exists recent work on applying machine learning and deep learning as we know it to the challenge of particle tracking :
  - Hopfield network : http://inspirehep.net/record/300646/
  - CNN in NOVA : https://arxiv.org/abs/1604.01444
  - HEP.TrkX : https://heptrkx.github.io/
  - TrackML RAMP : https://tinyurl.com/y84yd5hn
  - ... no golden solution yet

Some Early Solutions

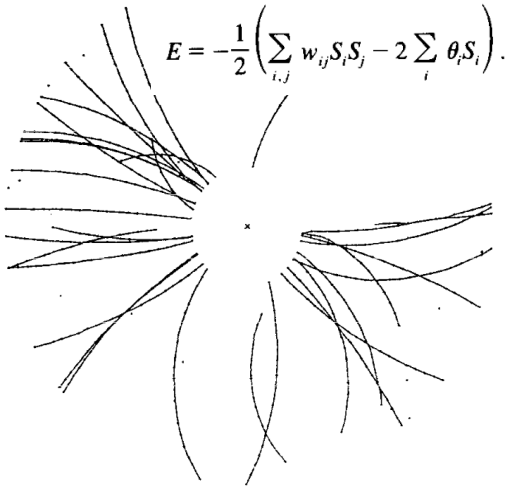
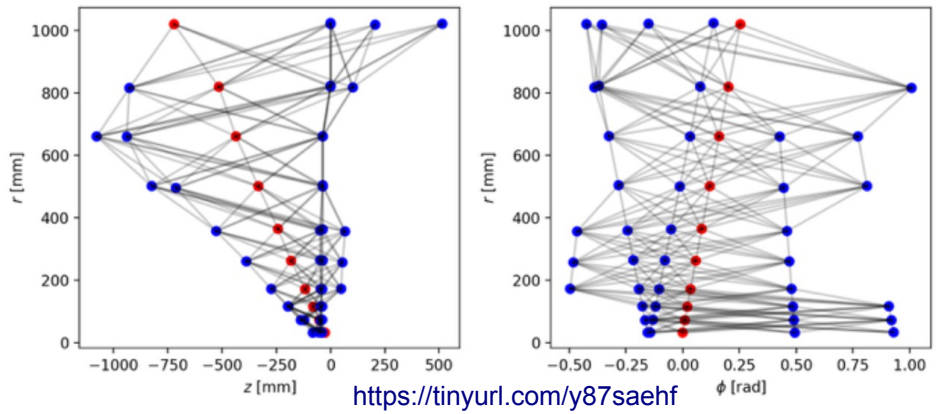
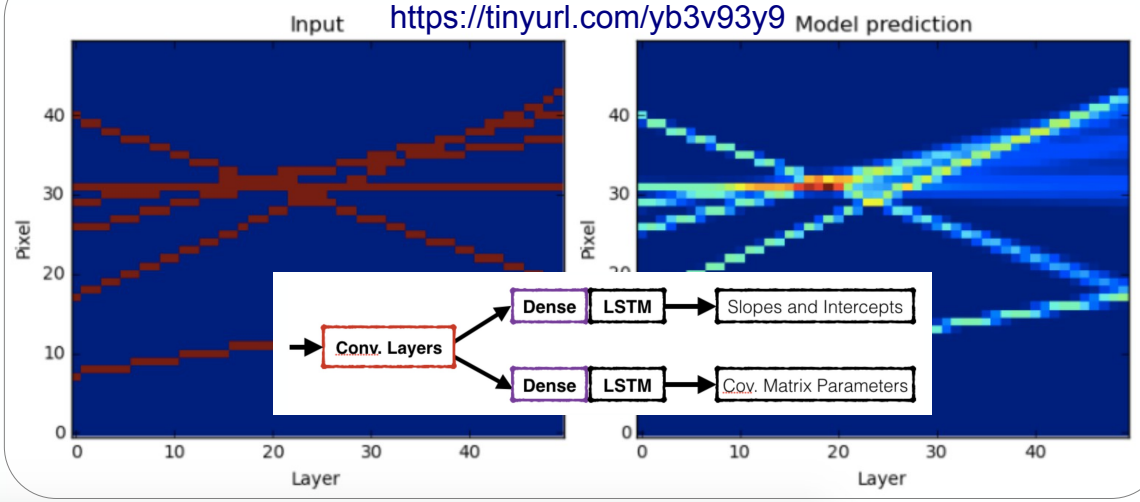


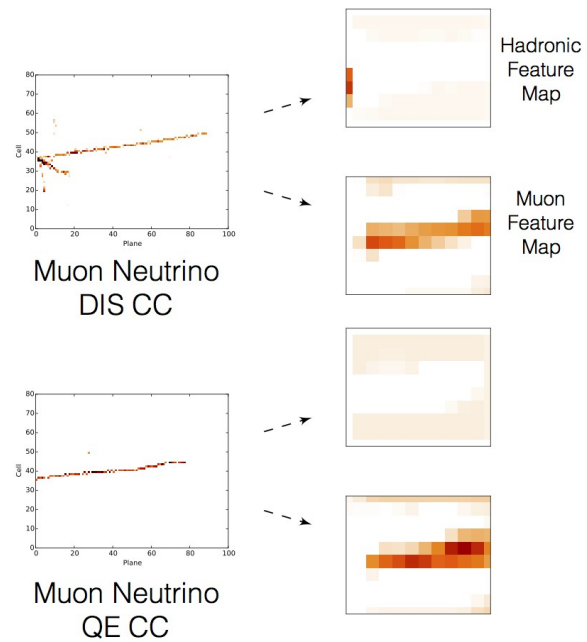
Fig. 4. Tracks in the ALEPH TPC reconstructed with a Hopfield net [13]. https://tinyurl.com/y9swquw6



https://tinyurl.com/y87saehf



https://tinyurl.com/yb3v93y9



https://arxiv.org/abs/1604.01444

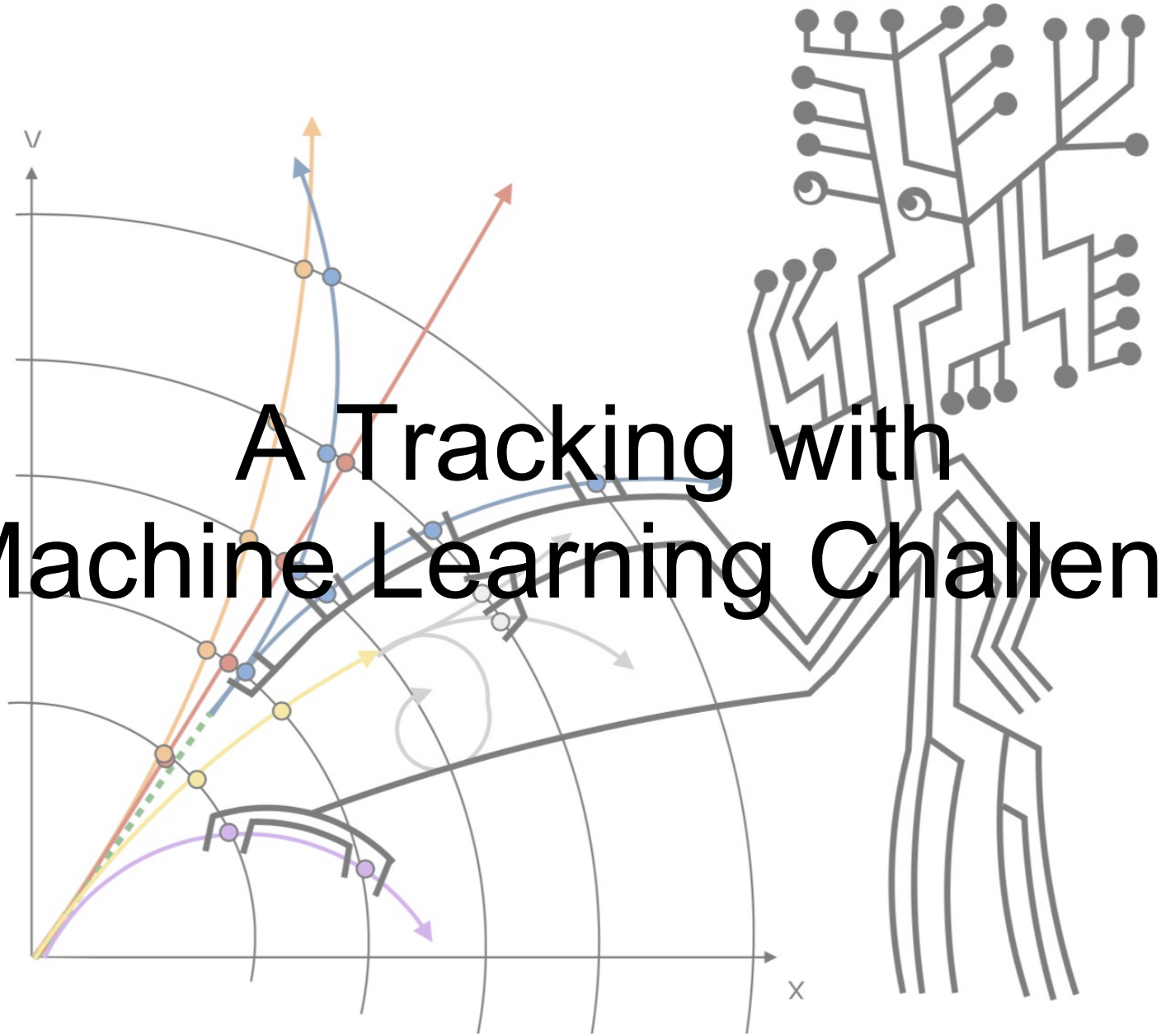
Bottom Lines

Current algorithms for tracking are highly performant physics-wise and scale badly computation-wise

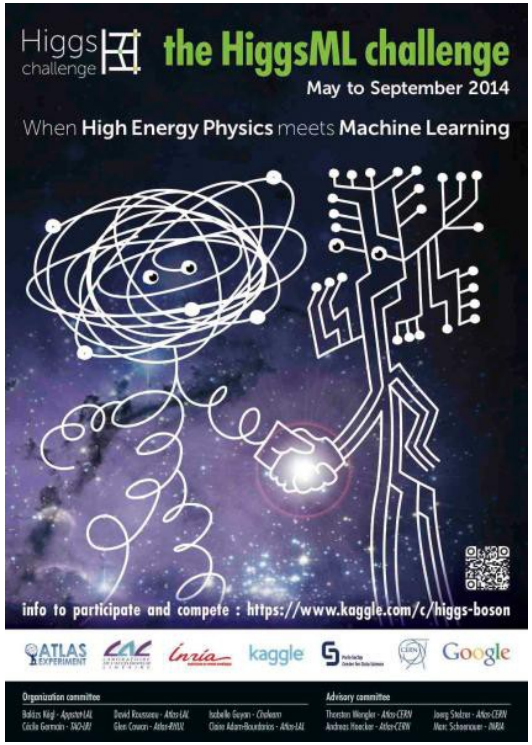
Faster implementations are possible with dedicated hardware

Reach out to machine learning community for new methods and possible solution, applicable on commodity hardware

A Tracking with Machine Learning Challenge

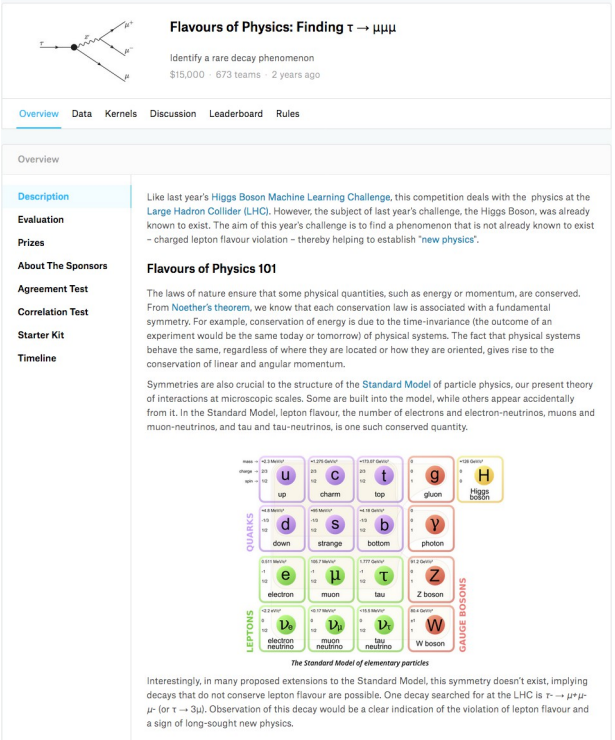


Previous Challenges



- 2000 teams. Largest competition at the time
- Winners went to DeepMind and OpenAI

https://www.kaggle.com/c/higgs-boson



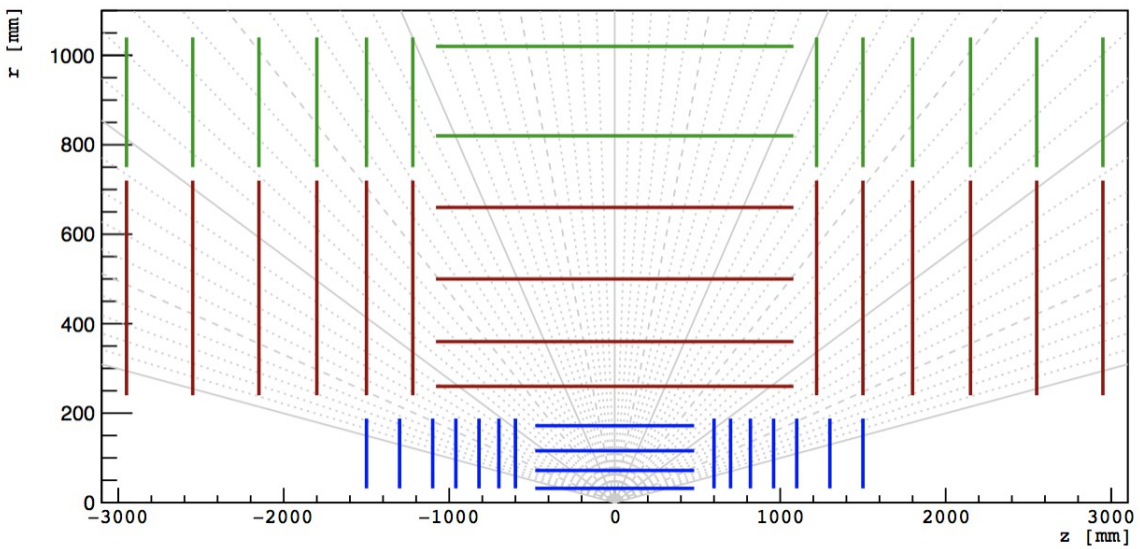
- 700 teams.
- Experienced data exploitation
- Some methods learned and re-applied later

https://www.kaggle.com/c/flavours-of-physics

The organizing team has participated in the organization of both events

Challenge Datasets

- Accurate simulation engine (ACTS) to produce realistic dataset
  - One file with list 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, ~100k hits/event)
- The goal of the challenge is to **assemble hits into tracks**
- Large training sample 100k events, 10 billion tracks ~100 GB
- Run in two phases
  - **Accuracy phase** on Kaggle platform
  - **Throughput phase** on Codalab



The Jury

Markus Elsing, CERN senior staff, group leader of the ATLAS computing and software group.
Frank Gaede senior physicist at DESY (Germany) is software coordinator for ILD.
Alison Lowndes is responsible for NVIDIA's Artificial Intelligence Developer Relations in the Europe, Middle East & Africa region.
Maurizio Pierini is a CERN physicist lead of the machine learning for Particle Physics ERC grant.
Danilo Rezende is Staff Research Scientist at Google DeepMind.
Marc Schoenauer is senior scientist at INRIA-Saclay
Svyatoslav Voloshynovskyy is associate professor at University of Geneva.

Reviewed the documentation made public by contestants of the challenge and decided on the level of innovation.

https://sites.google.com/site/trackmlparticle/international-advisory-committee

Featured Prediction Competition

TrackML Particle Tracking Challenge

High Energy Physics particle tracking in CERN detectors

CERN · 653 teams · 7 months ago

$25,000
Prize Money

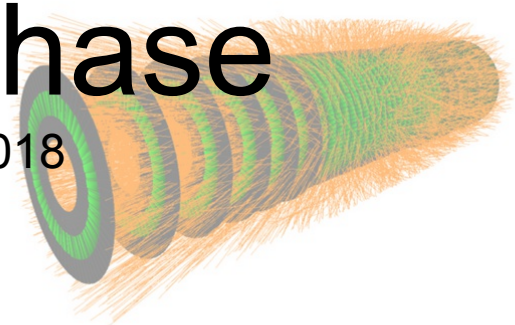
Overview | Data | Kernels | Discussion | Leaderboard | Rules | Join Competition

Overview

- Description
- Evaluation
- Timeline
- Prizes
- About The Sponsors

Accuracy Phase

April 30 – August 13, 2018



To explore what our universe is made of, scientists at CERN are colliding protons, essentially re-creating mini big bangs, and meticulously observing these collisions with intricate silicon detectors.

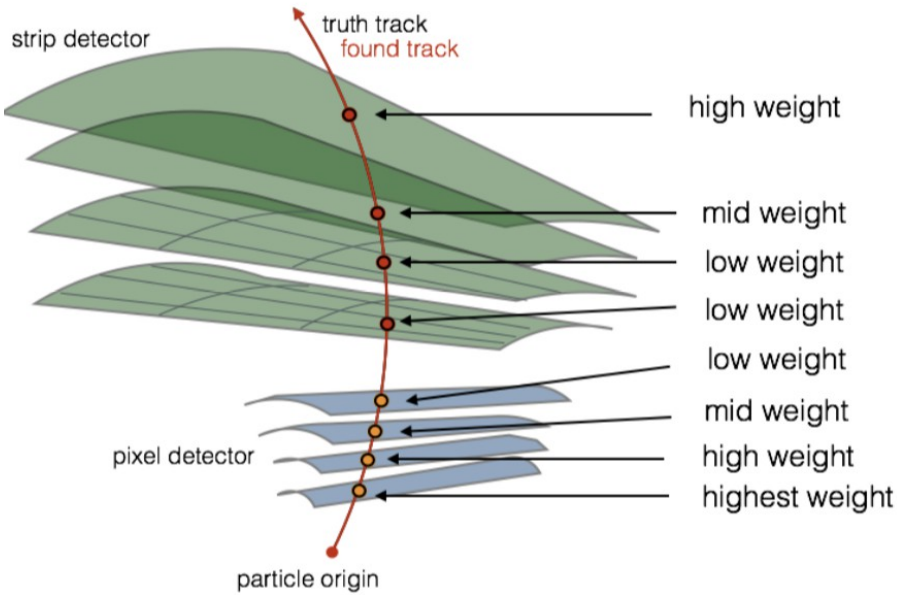
While simply creating the collisions and observations is already a massive scientific accomplishment, analyzing the enormous amounts of data produced from the experiments is becoming an overwhelming challenge.

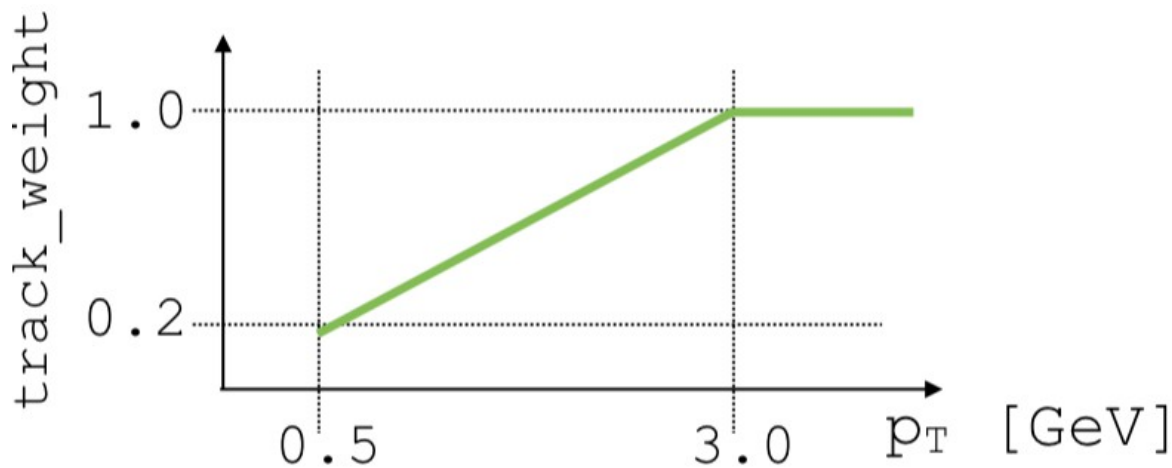
Event rates have already reached hundreds of millions of collisions per second, meaning physicists must sift through tens of petabytes of data per year. And, as the resolution of detectors improve, ever better software is needed for real-time pre-processing and filtering of the most promising events, producing even more data.

To help address this problem, a team of Machine Learning experts and physics scientists working at CERN (the world largest high energy physics laboratory), has partnered with Kaggle and prestigious sponsors to answer the question: can machine learning assist high energy physics in discovering and characterizing new particles?

https://www.kaggle.com/c/trackml-particle-identification

Scoring

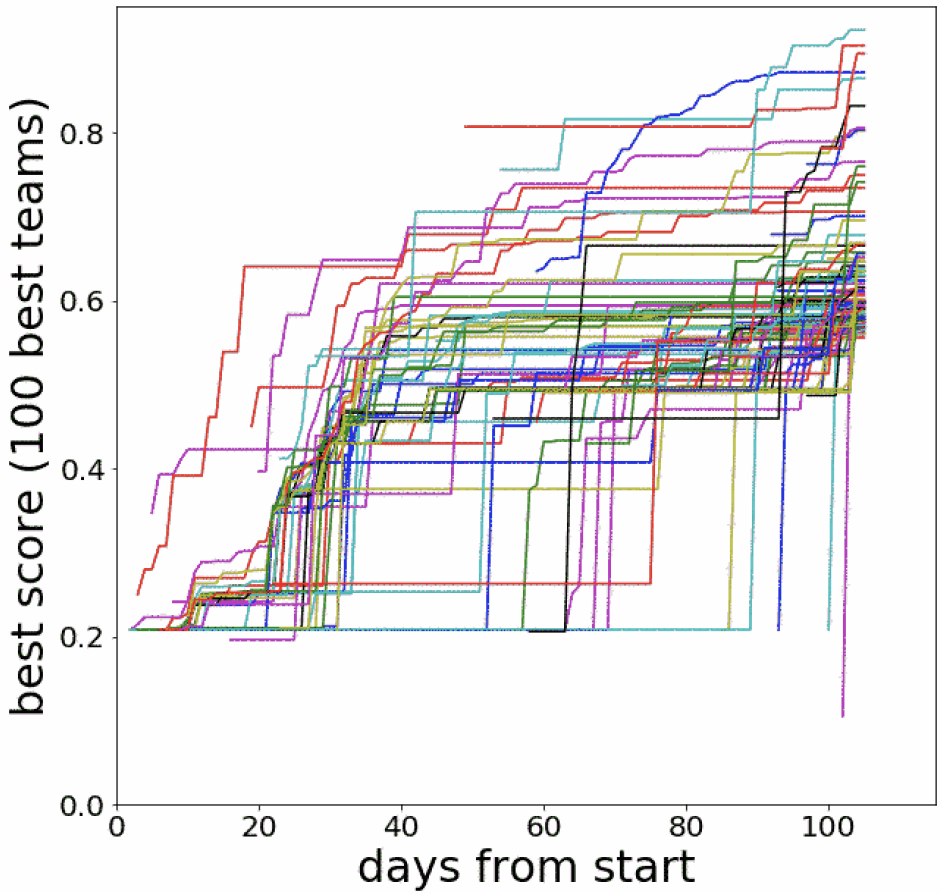




- At least 50% hits from the same ground truth particle
- At least 50% hits of the ground truth particle in the track
- Sum of weights (w_{order} x w_{pT}) of truth matched hits

- Score normalized to sum of weights : ideal score **is 1**
- 100 events used for scoring : precision ~0.1%

Final Leaderboard



In the money | Gold | Silver | Bronze

#	△pub	Team Name	Kernel	Team Members	Score	Entries
1	—	Top Quarks			0.92182	10
2	—	outrunner			0.90302	9
3	—	Sergey Gorbunov			0.89353	6
4	—	demelian			0.87079	35
5	—	Edwin Steiner			0.86395	5
6	—	Komaki			0.83127	22
7	—	Yuval & Trian			0.80414	56
8	—	bestfitting			0.80341	6
9	—	DBSCAN forever			0.80114	23
10	—	Zidmie & KhaVo			0.76320	26
11	—	Andrea Lonza			0.75845	15
12	—	Finnies			0.74827	56
13	—	Rei Matsuzaki			0.74035	12
14	—	Mickey			0.73217	10
15	—	Vicens Gaitan			0.70429	19
16	—	Robert			0.69955	3

In the money

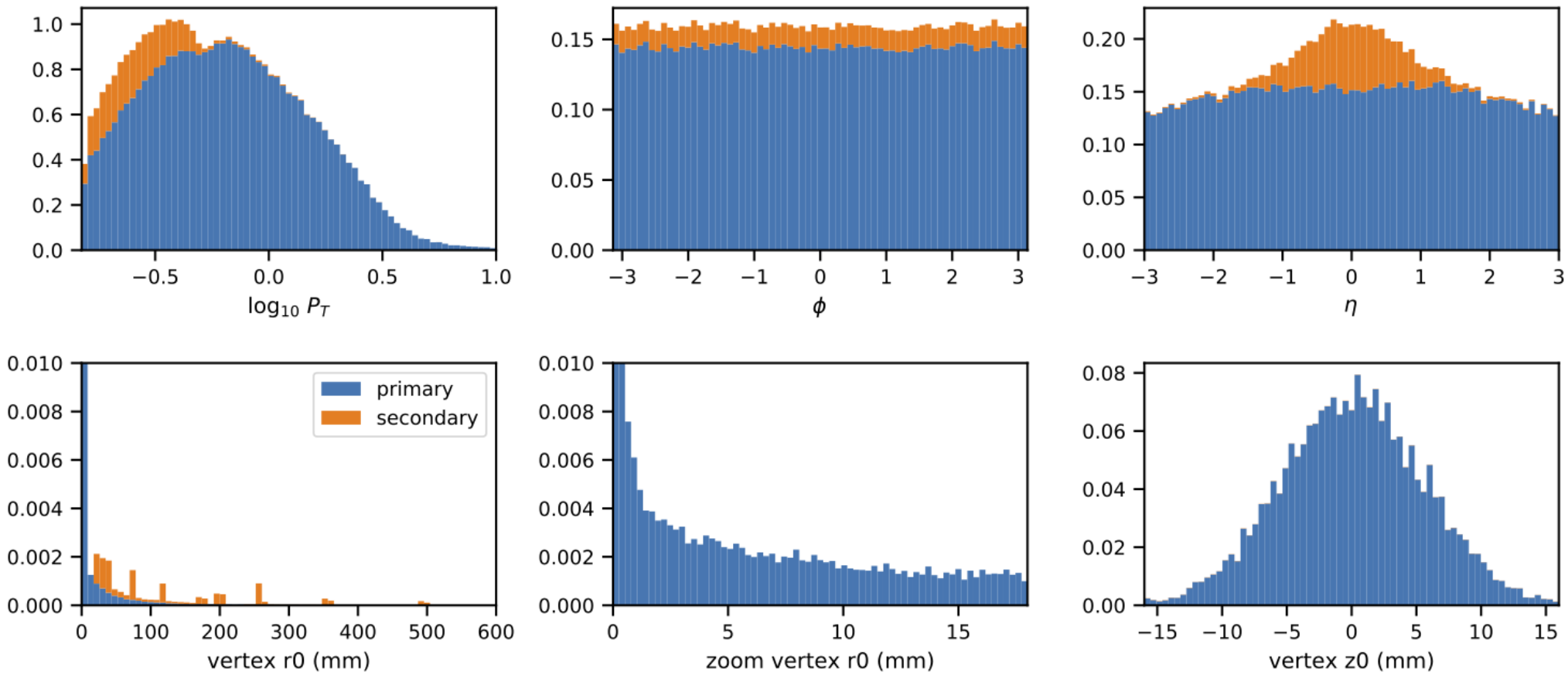
Jury pick

Jury pick

Jury pick

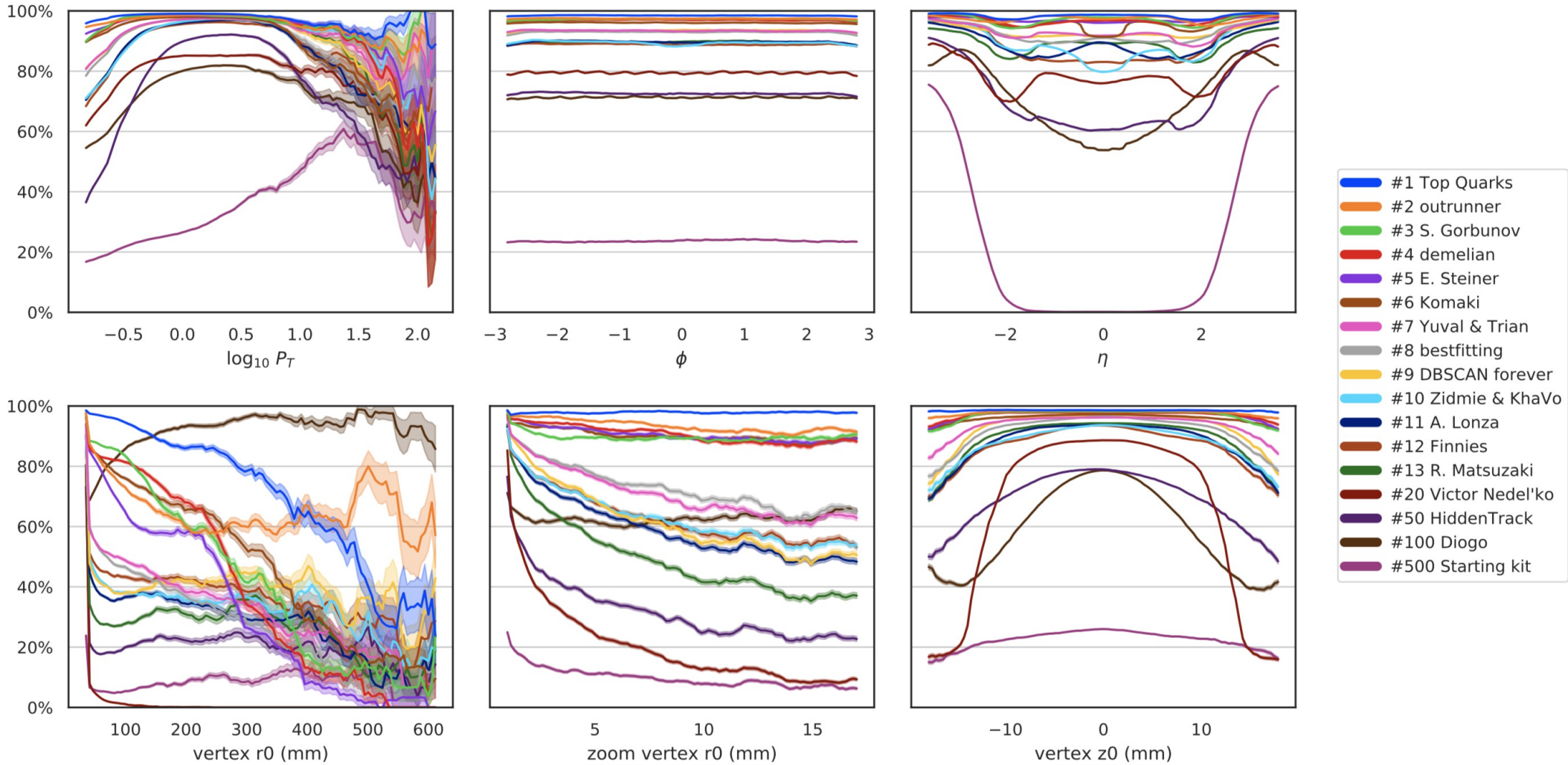
https://www.kaggle.com/c/trackml-particle-identification/leaderboard

The “Physics Score”



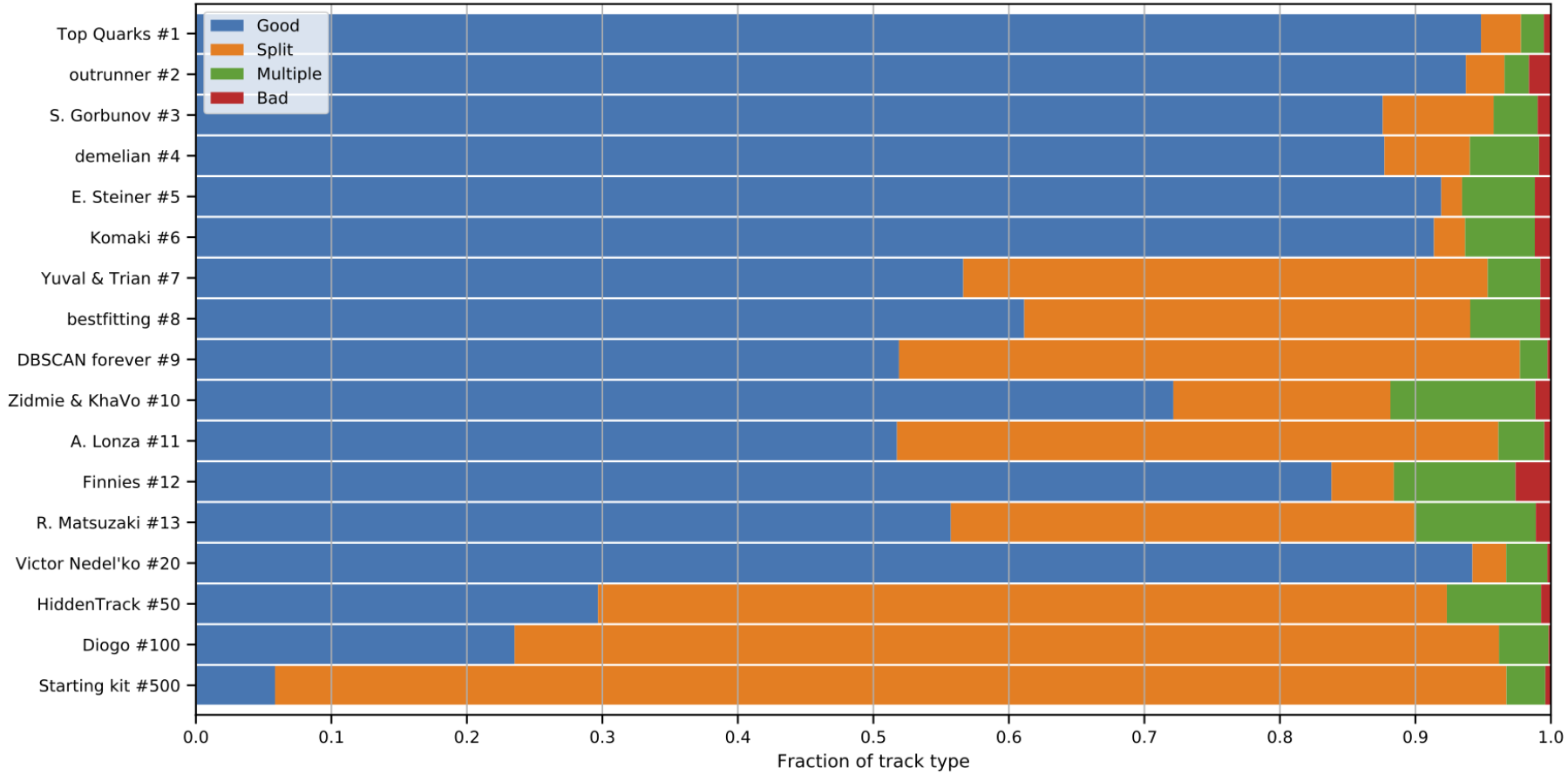
Weighted distribution of track parameters in the dataset

Physics Performance



Highest score correlates well with the tracking efficiency

Track Type



- **Good** track and particle purities above 50% (goes into the score)
- **Split** particle purity below 50%, track purity above 50%
- **Multiple** Particle purity above 50%, but track purity below 50%
- **Bad** Both below 50%

Insight on Algorithms 1/2

- **First : *Top Quarks***
  - Johan Sokrates is an industrial Mathematics master student
  - Pair seeding, triplet extension, **trajectory following**, track cleaning, all with **machine learning** for quality selection

- **Second :**
  - Pei-Lien Chou is a software engineer in image-based deep learning in Taïwan
  - **Machine learning** to predict the adjacency matrix

- **Third :**
  - Sergey Gorbunov is a physicist, expert in tracking
  - Triplet seeding, **trajectory following**

Insight on Algorithms 2/2

- **Jury Innovative prize**
  - Yuval Reina is an electronic engineer and Trian Xylouris is an entrepreneur
  - Marginalized Hough transform with **machine learning classifier**

- **Jury Clustering prize**
  - Jean-François Puget CPMP is a software engineer at IBM. He is both competition and discussion Kaggle grandmaster
  - **DBSCAN clustering** with iterative Hough transform

- **Jury Deep Learning prize**
  - Nicole and Liam Finnie are software engineers
  - DBSCAN seeding, **trajectory following with LSTM**

- **Organization pick**
  - Diogo R. Ferreira is a professor/researcher, focusing on data science and nuclear fusion
  - **Pattern matching**

Lesson Learned

- Preparing a challenge is a hard process
- Not too surprising to find trajectory following in winning solutions. Great to have machine learning added
- Noise-driven control of the combinatorial explosion
- Controlling efficiency along the algorithm
- Data driven estimation of the magnetic field
- Computational cost of deep learning
- Tuning of algorithm parameters

Throughput Phase

TrackML Throughput Phase

Organized by VictorEstrade - Current server time: March 7, 2019, 3:52 p.m. UTC
Reward $15,000

▶ Current	Next	End
Development	Final	Competition Ends
Sept. 7, 2018, midnight UTC	March 12, 2019, 11:59 p.m. UTC	March 12, 2019, 11:59 p.m. UTC

Learn the Details | Phases | Participate | Results | Public Submissions | Forums

- Overview
- Evaluation
- Terms and Conditions
- Prizes
- Sponsors, organisers and International Advisory Committee
- Timeline
- Step by step
- News
- Contact

Welcome!

This competitions is an official NeurIPS 2018 competition.

To explore what our universe is made of, scientists at CERN are colliding protons, essentially recreating mini big bangs, and meticulously observing these collisions with intricate silicon detectors. Event rates have already reached hundreds of millions of collisions per second, meaning physicists must sift through tens of petabytes of data per year. And, as the resolution of detectors improve, ever better software is needed for real-time pre-processing and filtering of the most promising events, producing even more data. To help address this problem, a team of Machine Learning experts and physics scientists working at CERN (the world largest high energy physics laboratory), has partnered with prestigious sponsors to answer the question: can machine learning assist high

Oct 12 – March 12, 2019

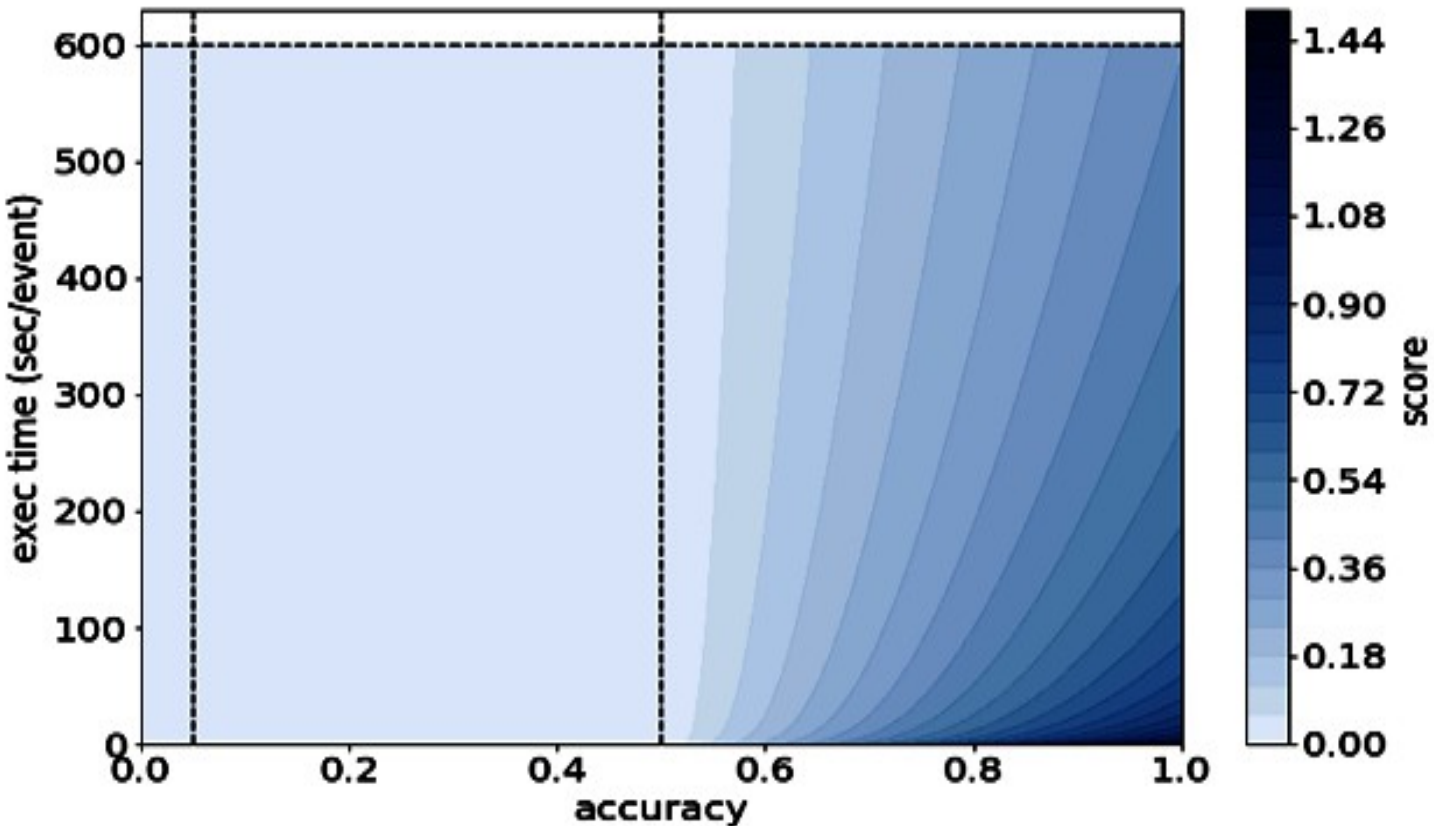
Modifications to the dataset
Same detector geometry with half thiner modules
Beamspot corrected from 5.5 mm to 5.5 cm
Loopers are not simulated
Bug on electron multiple scattering fixed

https://competitions.codalab.org/competitions/20112

Scoring

Zero score if time > 600s and accuracy < 0.5

$$\sqrt{\log\left(1+\frac{600}{time}\right)\times\left(accuracy-0.5\right)^2}$$



Leaderboard

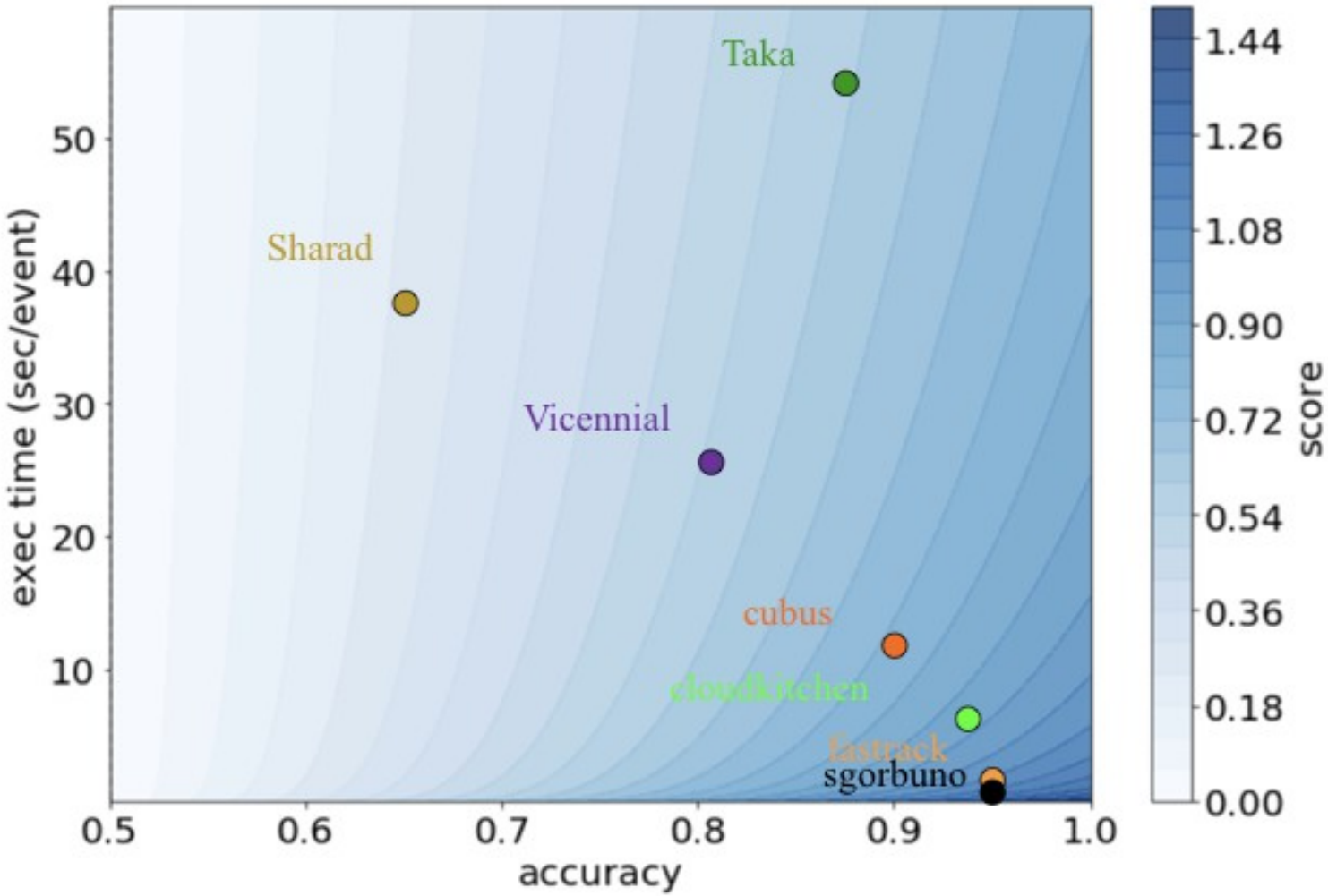
HEP people

PH+CS

RESULTS									
#	User	Entries	Date of Last Entry	score ▲	accuracy_mean ▲	accuracy_std ▲	computation time (sec) ▲	computation speed (sec/event) ▲	Duration ▲
1	sgorbuno	9	03/12/19	1.1727 (1)	0.944 (2)	0.00 (14)	28.06 (1)	0.56 (1)	64.00 (1)
2	fastrack	53	03/12/19	1.1145 (2)	0.944 (1)	0.00 (15)	55.51 (16)	1.11 (16)	91.00 (6)
3	cloudkitchen	73	03/12/19	0.9007 (3)	0.928 (3)	0.00 (13)	364.00 (18)	7.28 (18)	407.00 (8)
4	cubus	8	09/13/18	0.7719 (4)	0.895 (4)	0.01 (9)	675.35 (19)	[illegible]	724.00 (9)
5	Taka	11	01/13/19	0.5930 (5)	0.875 (5)	0.01 (12)	[illegible]	[illegible]	2758.00 (13)
6	Vicennial	27	02/24/19	0.5634 (6)	0.815 (6)	[illegible]	[illegible]	25.41 (20)	1339.00 (10)
7	Sharad	57	03/10/19	0.2918 (7)	0.67[illegible]	[illegible]	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)
11	yangguo	1	03/01/19	0.0000 (8)	0.082 (15)	0.01 (8)	48.63 (4)	0.97 (4)	86.00 (3)

Final leaderboard to be determined from more precie time measurement

Time – Accuracy Decomposition



Incidentally, best solutions are also best accuracy and best timing.
Software will be submitted and analyzed in the coming weeks.

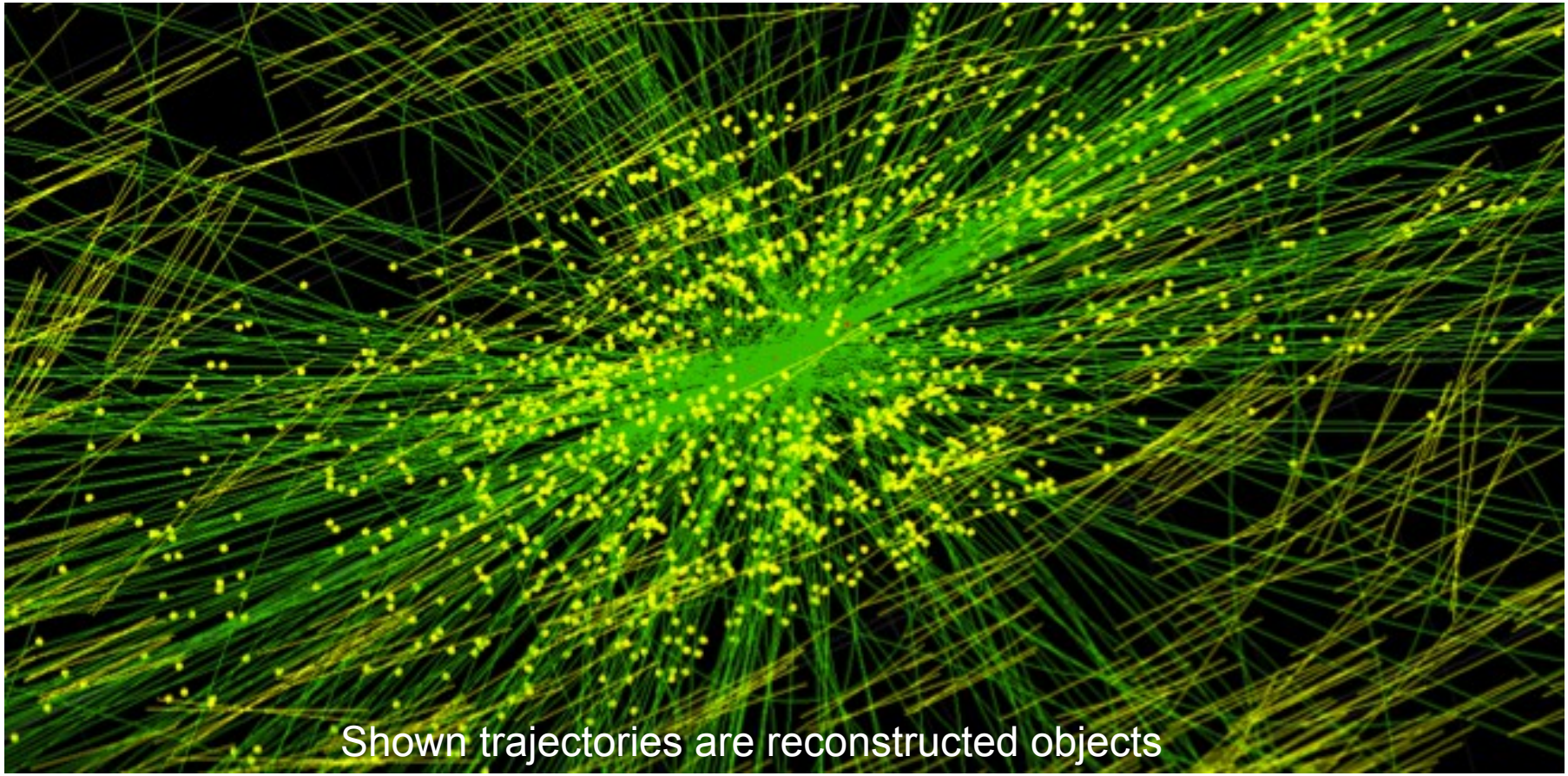
Summary & Outlook

- Active challenge on both platforms.
- Some lesson learned so far. Chapter in NeurIPS 2019 competition book.
- Traditional trajectory following was used. Some machine learning technique presented.
- Further work required to extract the novel ideas proposed during the challenge. Ongoing work to port it in ACTS.
- Final version of the dataset will be on the cern open data portal. Already used in several benchmarking.
- Stay tune for the aftermath of the throughput phase.

sponsors

Extra Material

Complexity and Ambiguity



The future is with **x10 more hits**

Scene Labeling

Farabet et al. ICML 2012, PAMI 2013