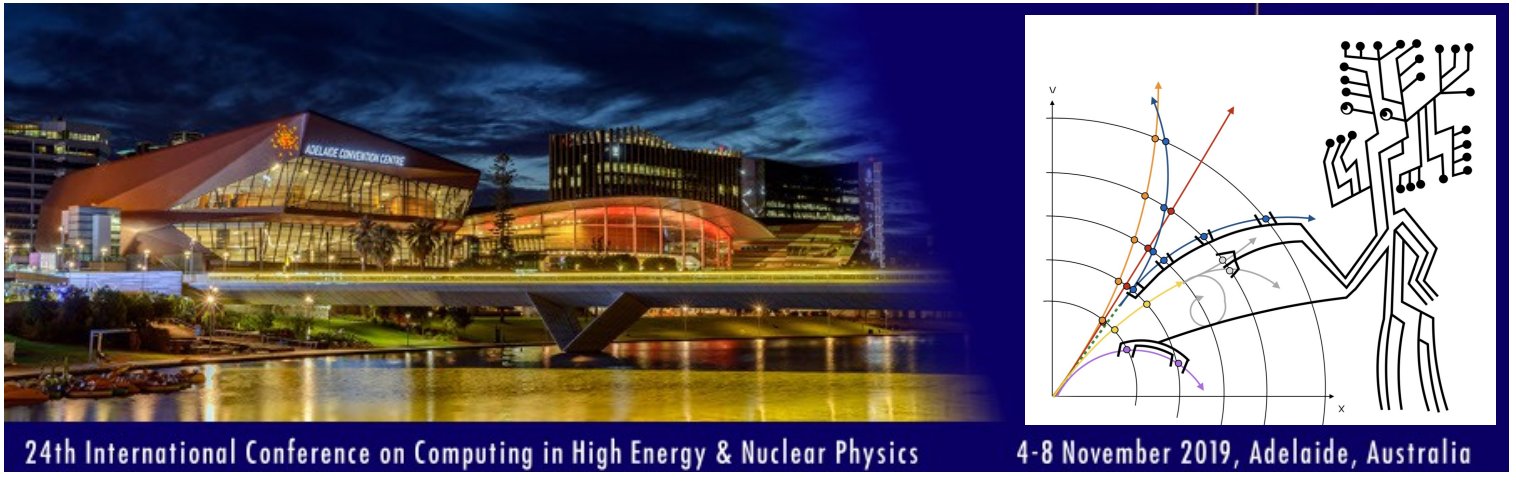
TrackML : a Tracking Machine Learning Challenge



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

kaggle

NVIDIA

UNIVERSITÉ DE GENÈVE

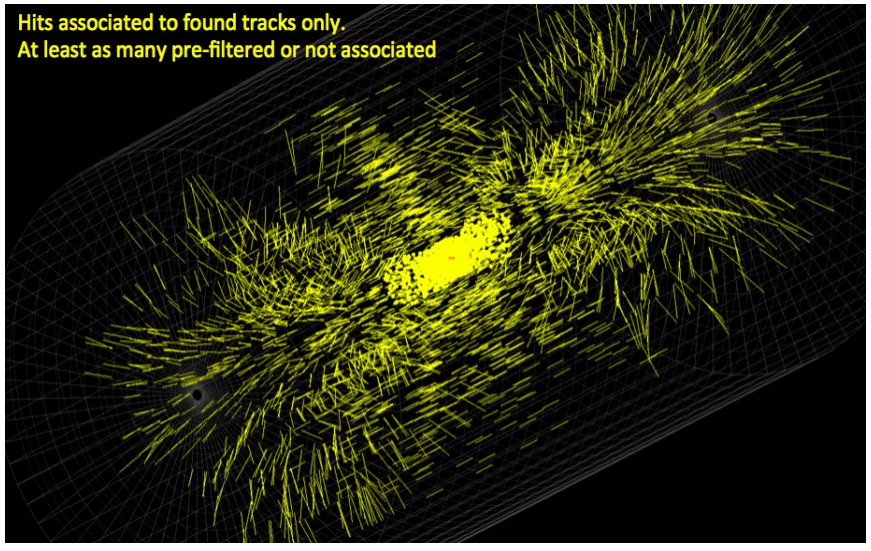
Paris-Saclay Center for Data Science

IRIS-HEP

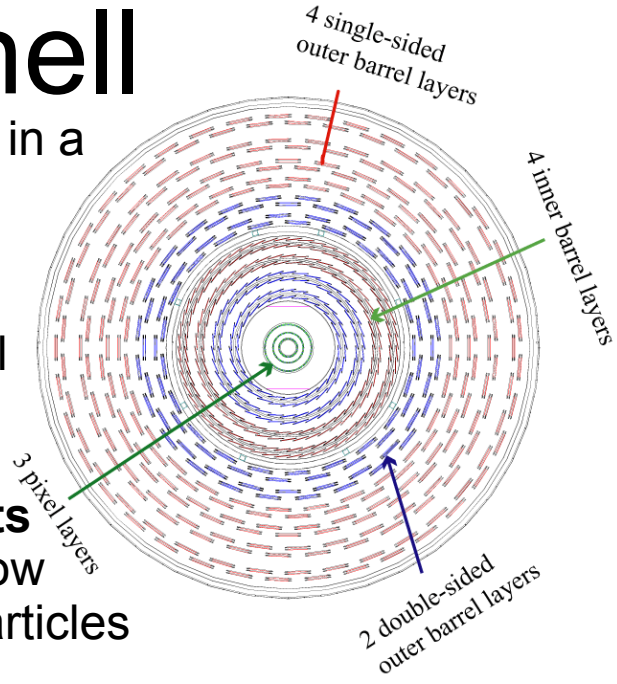
Outline

- Forewords on Tracking and Machine learning
- Accuracy Phase
- Throughput Phase
- Outlook

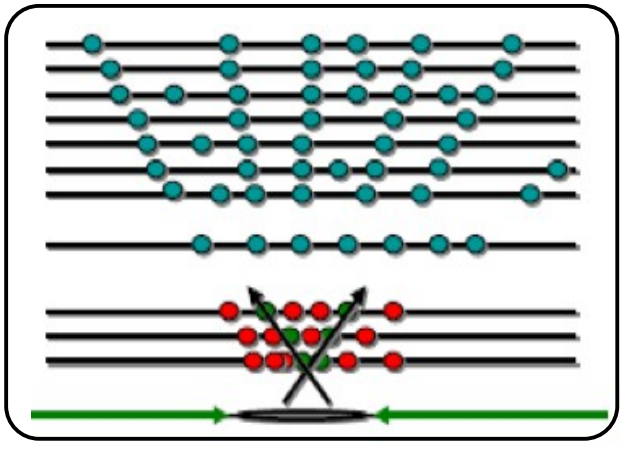
Tracking in a Nutshell



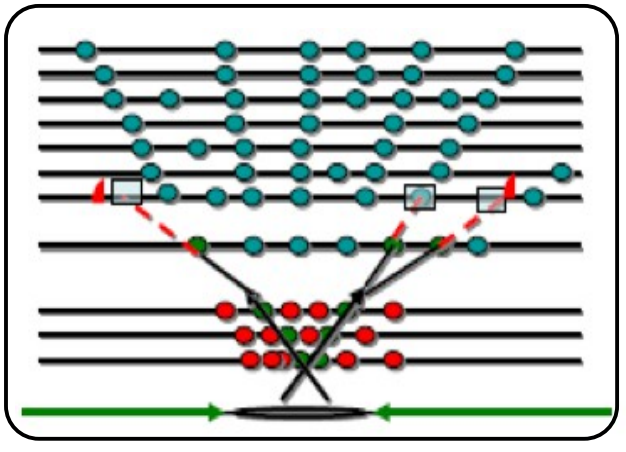
- Particle trajectory bended in a solenoidal magnetic field
- Curvature is a proxy to momentum
- Particle ionize silicon pixel and strip throughout several concentric layers
- **Thousands of sparse hits**
- Lots of hit pollution from low momentum, secondary particles

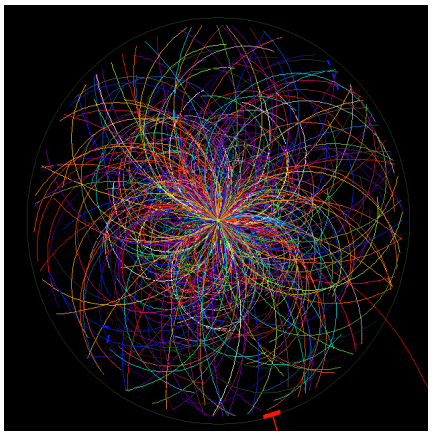


Seeding



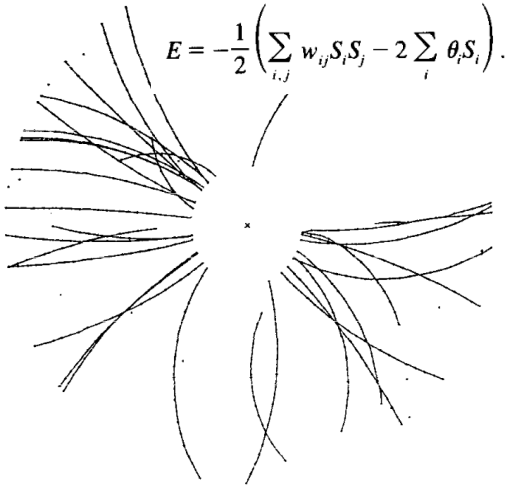
Kalman Filter





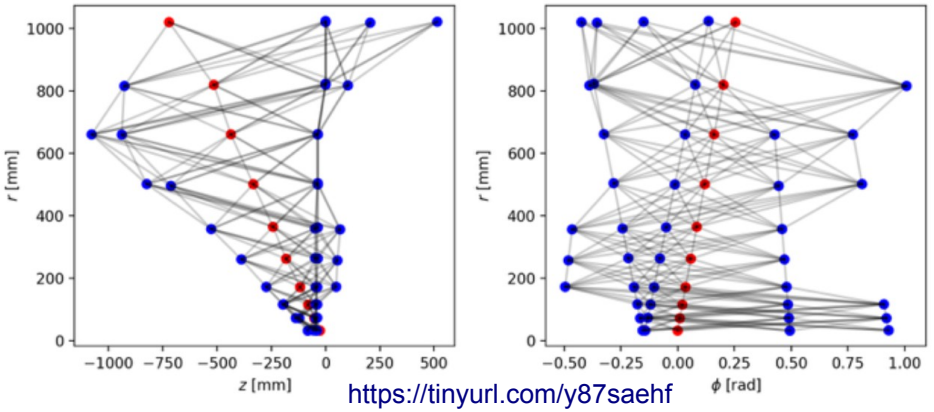
- **Explosion in hit combinatorics** in both seeding and stepping pattern recognition
- **Highly computing consuming task** in extracting physics content from LHC data

Tracking with Machine Learning

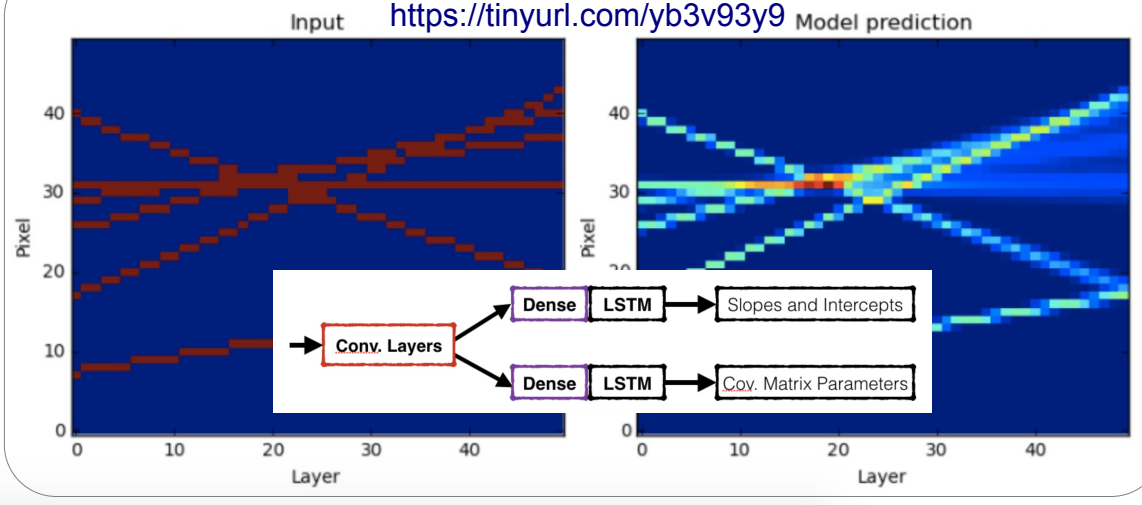


$$E = -\frac{1}{2}\left(\sum_{i,j} w_{ij} S_i S_j - 2\sum_i \theta_i S_i\right).$$

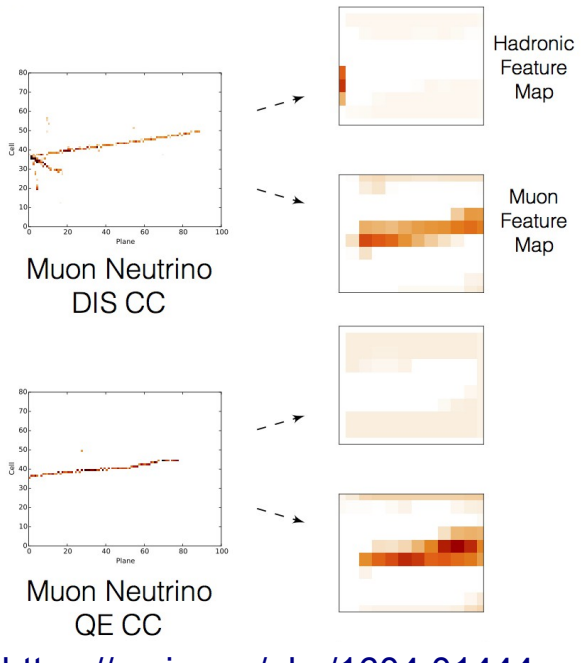
Fig. 4. Tracks in the ALEPH TPC reconstructed with a Hopfield net [13]. https://tinyurl.com/y9swquw6



https://tinyurl.com/y87saehf



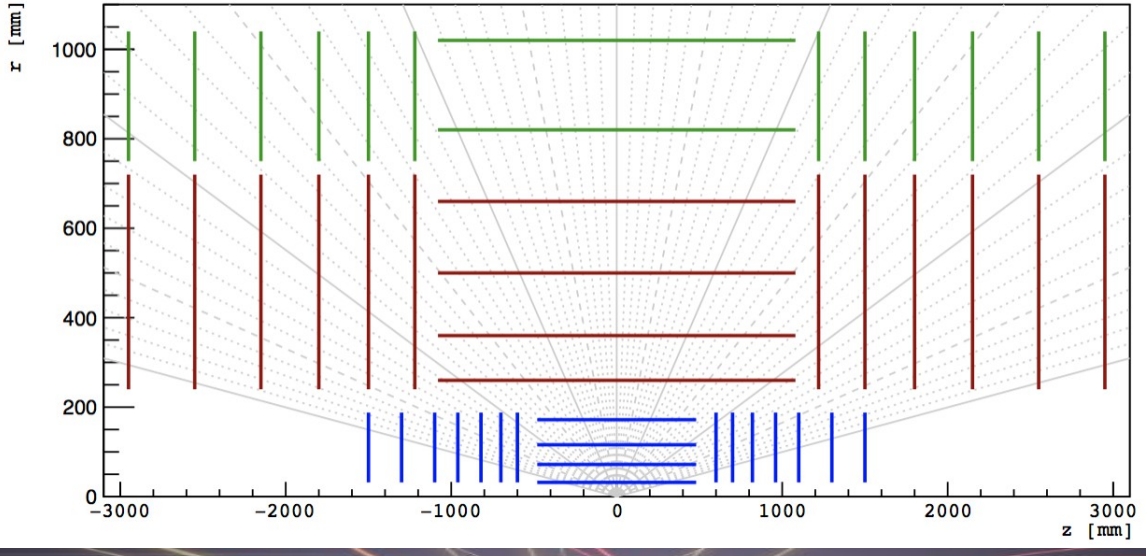
https://tinyurl.com/yb3v93y9



https://arxiv.org/abs/1604.01444

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



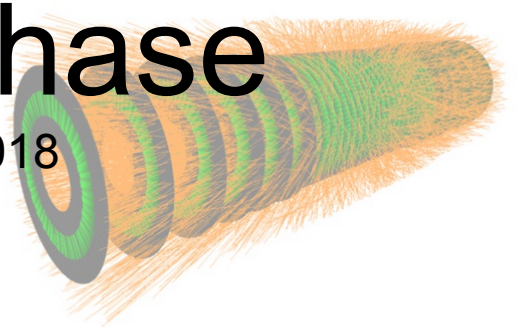
Overview Data Kernels Discussion Leaderboard Rules **Join Competition**

Overview

Description
Evaluation
Timeline
Prizes
About The Sponsors

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

While [illegible] 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.

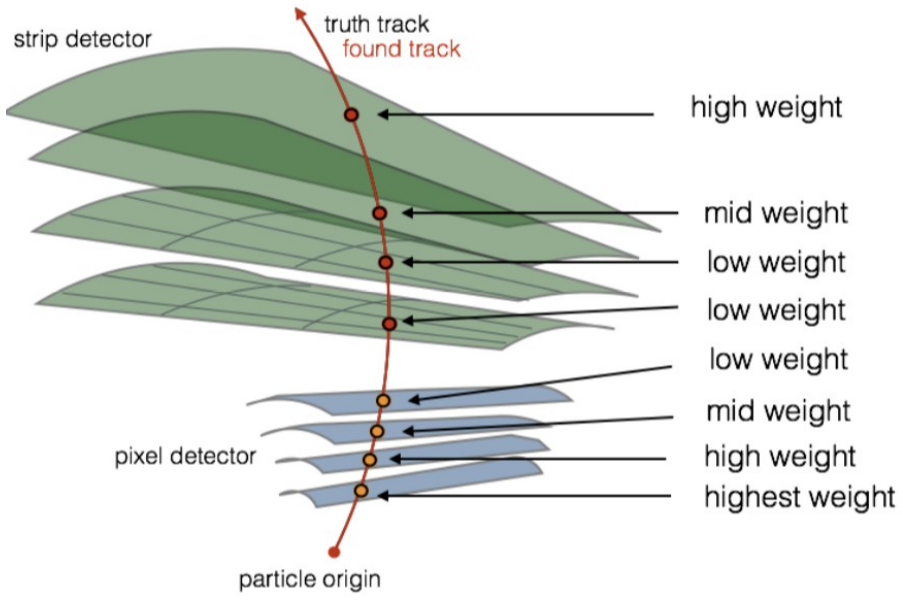
To help address this problem, a team of Machine Learning experts and physics scientists working at CERN (the world largest high energy physics laboratory), has partnered with Kaggle and prestigious sponsors to answer the question: can machine learning assist high energy physics in discovering and characterizing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

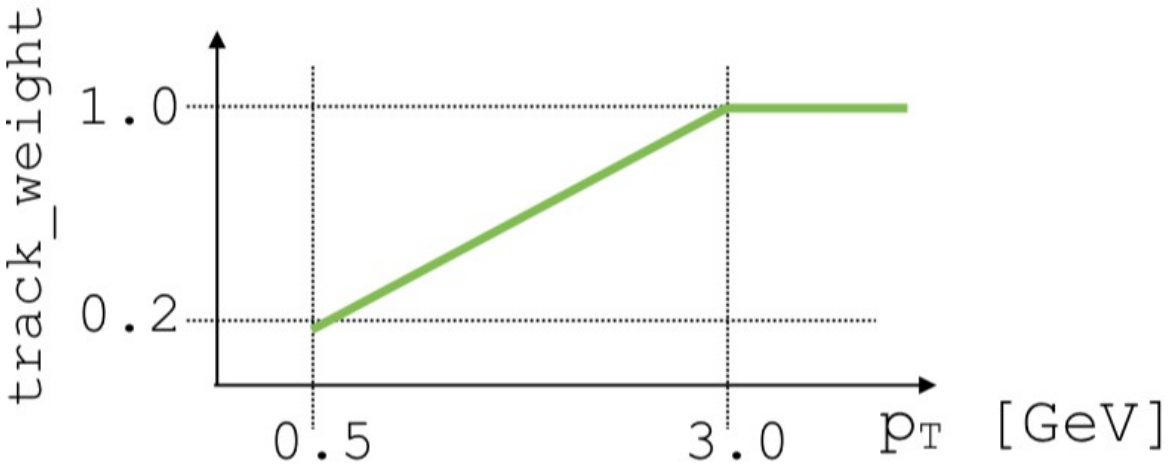
Accuracy Phase

April 30 – August 13, 2018

https://www.kaggle.com/c/trackml-particle-identification
https://arxiv.org/abs/1904.06778

Scoring : Accuracy Phase

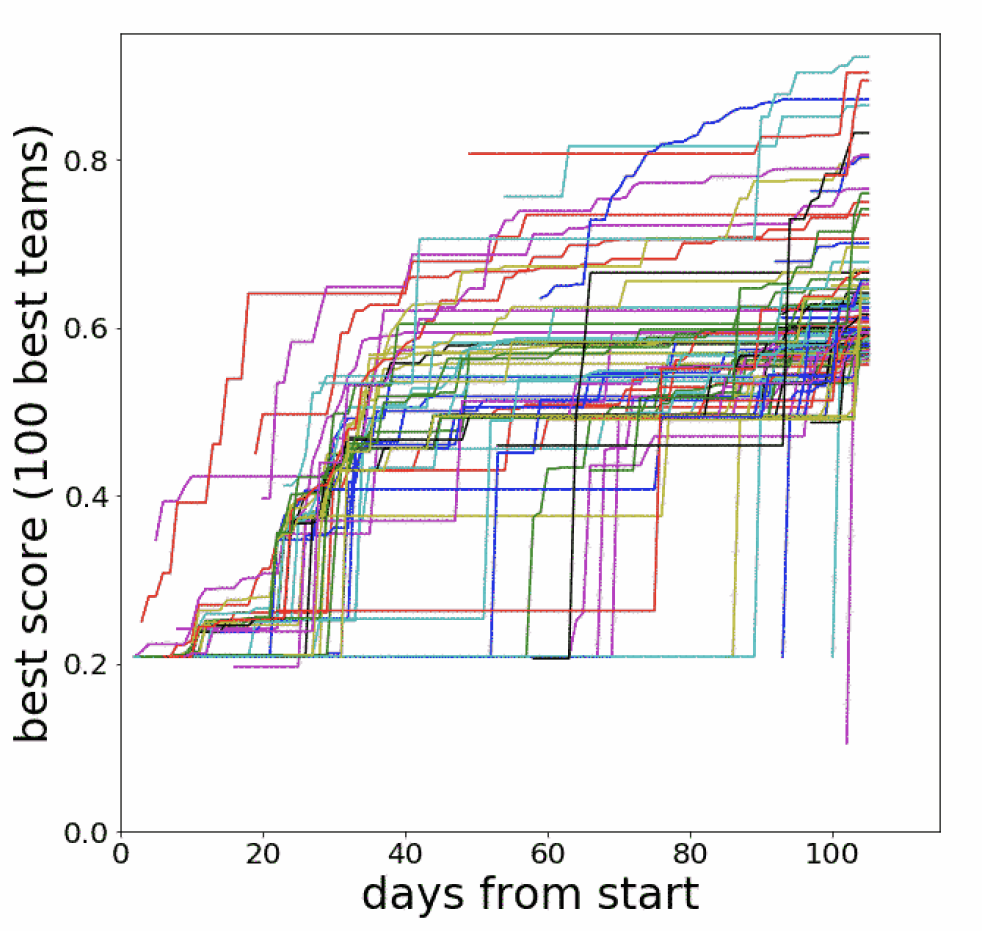




- At least 50% hits from the same ground truth particle
- At least 50% hits of the ground truth particle in the track
- Sum of weights (w_{order} x w_{pT}) of truth matched hits

- 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

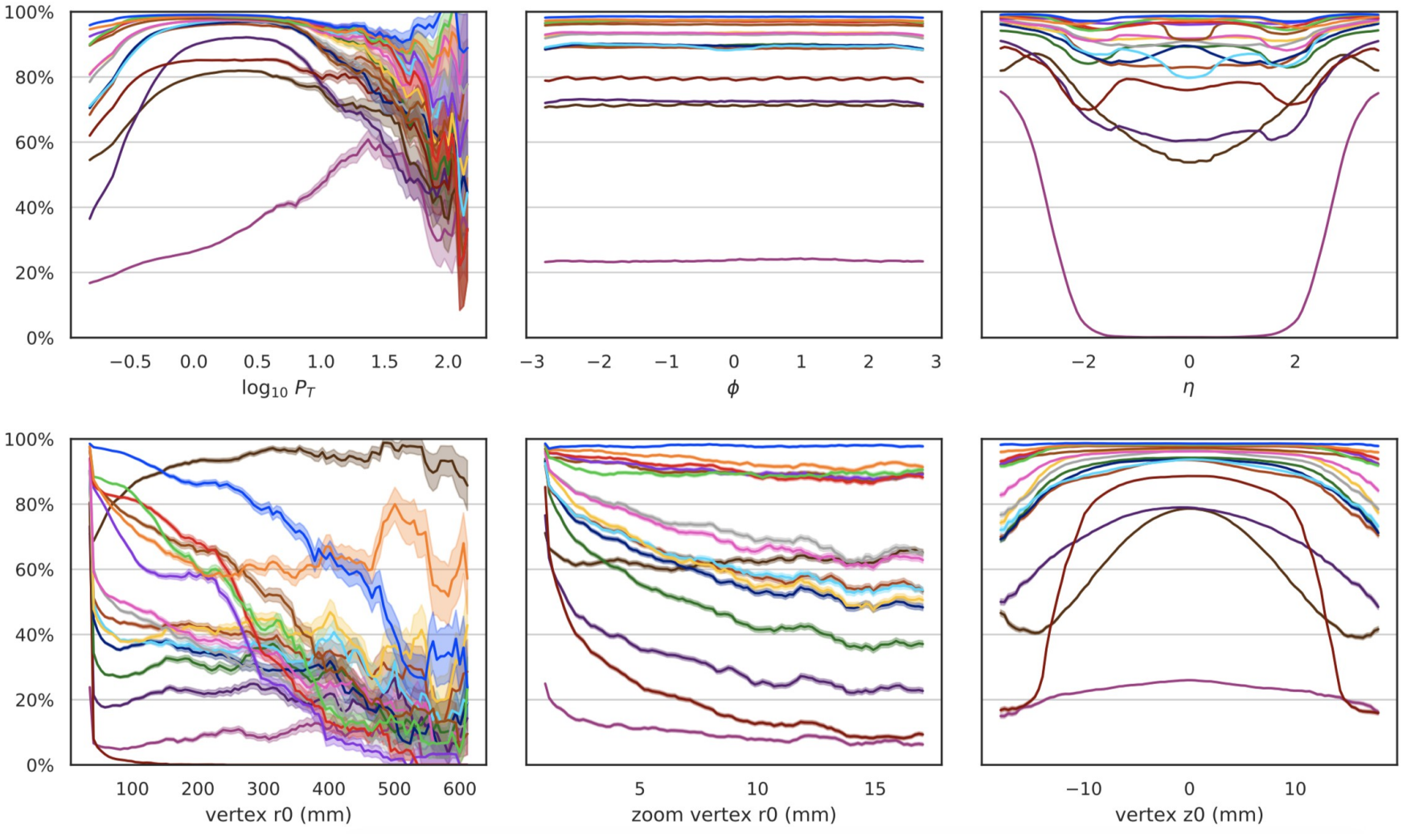
Jury pick

Jury pick

Jury pick

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

Physics Performance



Highest score correlates well with the tracking efficiency

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

- **Third : *Sergey Gorbunov***
  - Sergey Gorbunov is a physicist, expert in tracking
  - Iterative steps, 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**

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

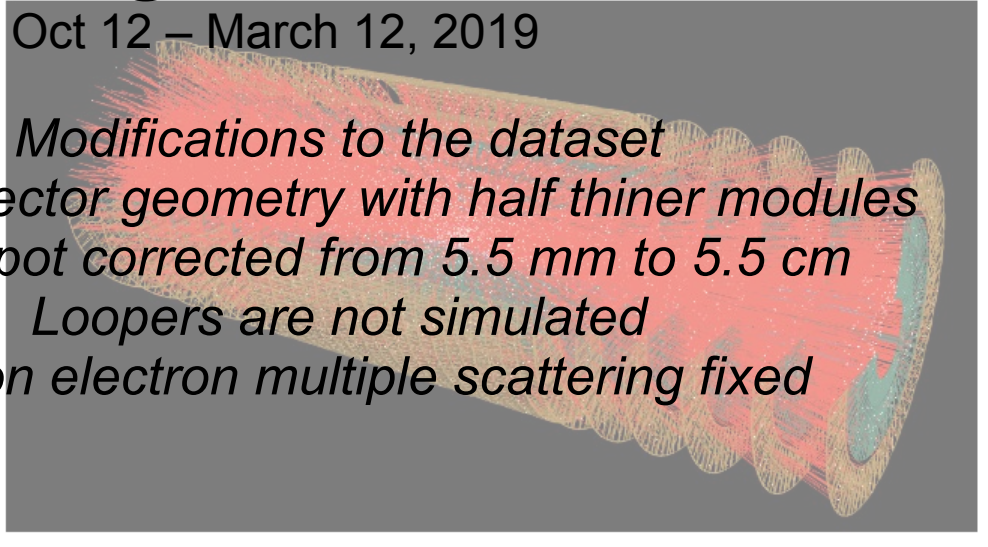
Welcome!

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

Throughput Phase

Oct 12 – March 12, 2019



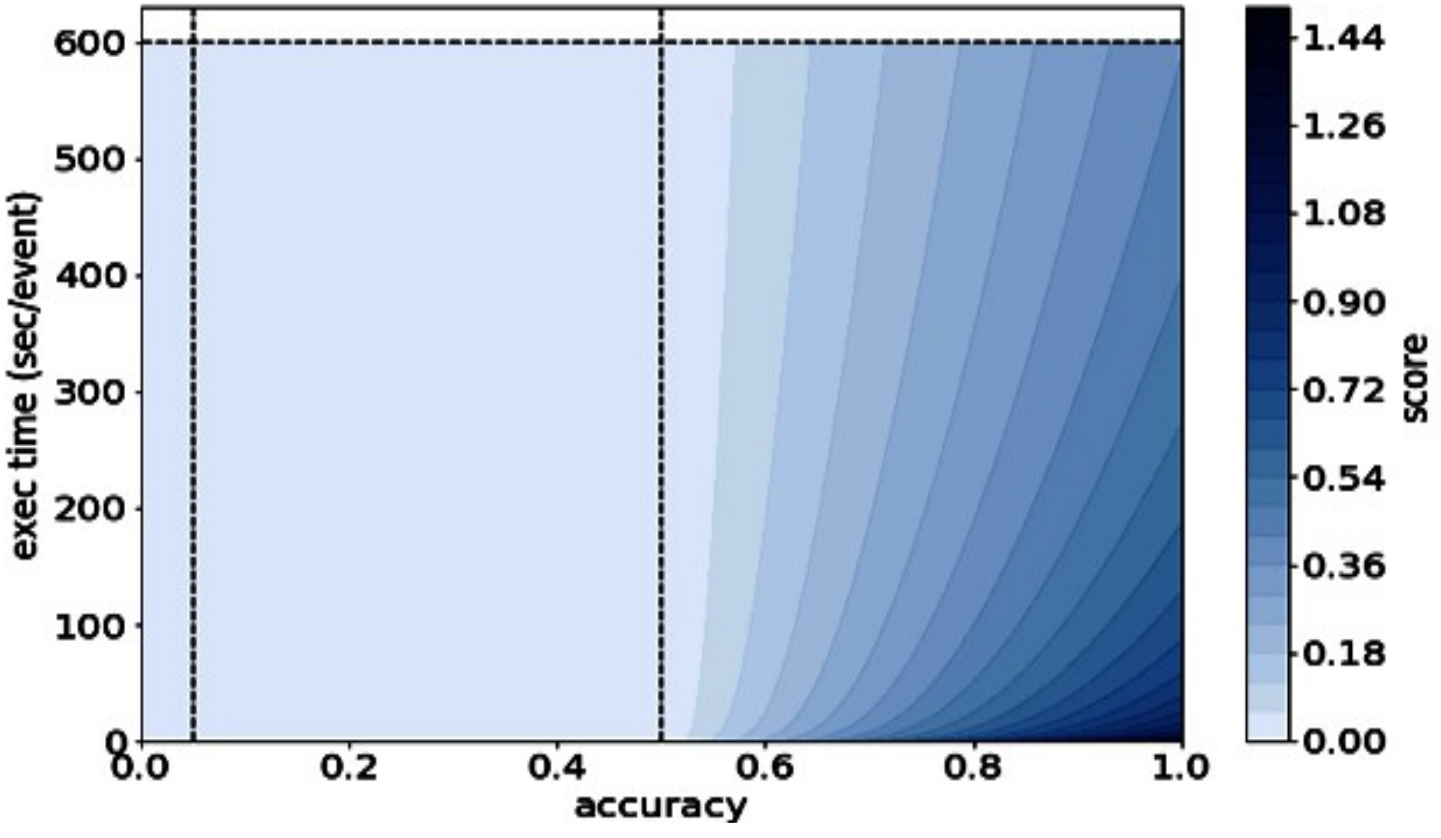
Modifications to the dataset
Same detector geometry with half thiner modules
Beamspot corrected from 5.5 mm to 5.5 cm
Loopers are not simulated
Bug on electron multiple scattering fixed

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

Scoring : Throughput Phase

Zero score if time > 600s and accuracy < 0.5

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



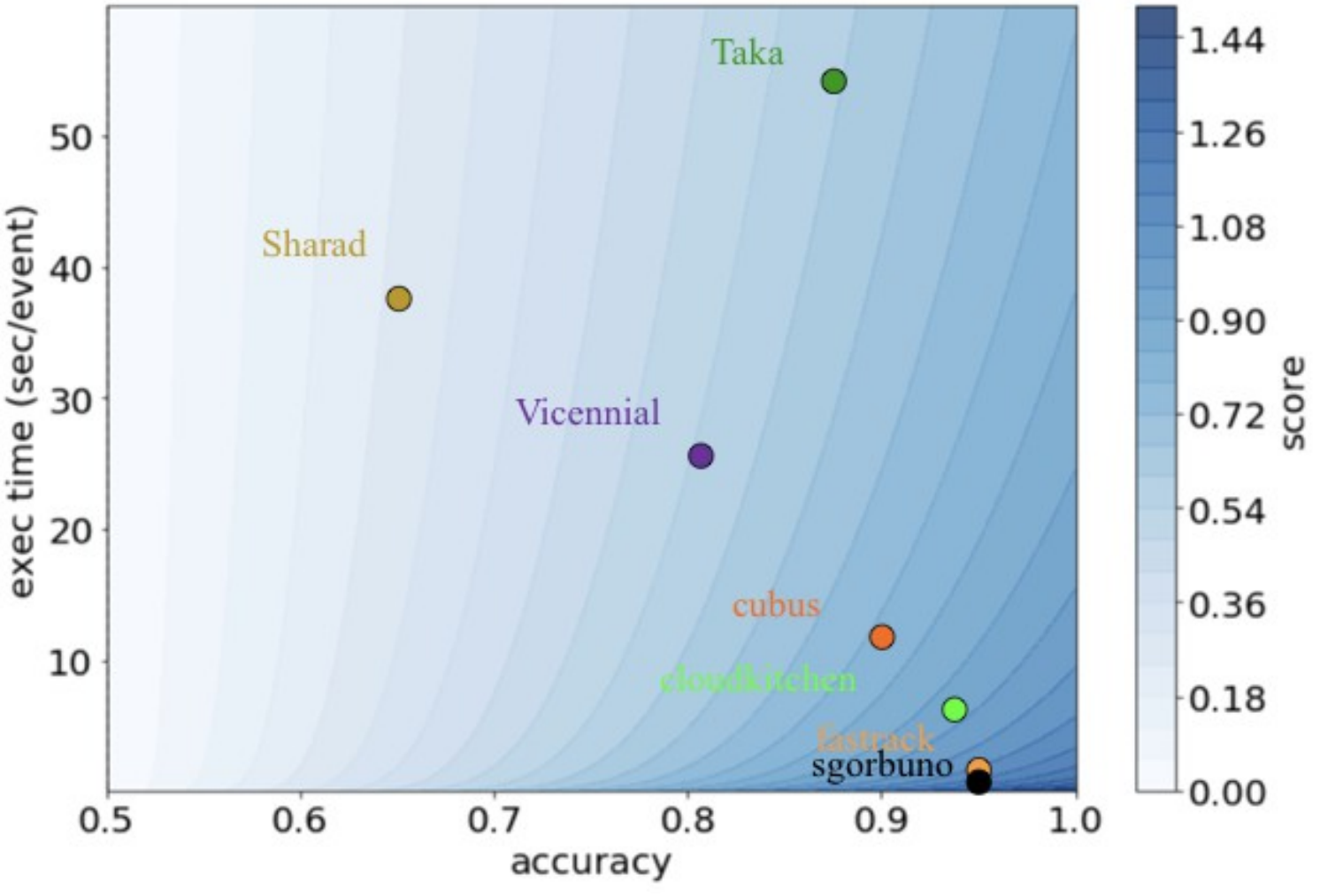
Leaderboard

HEP people

PH+CS

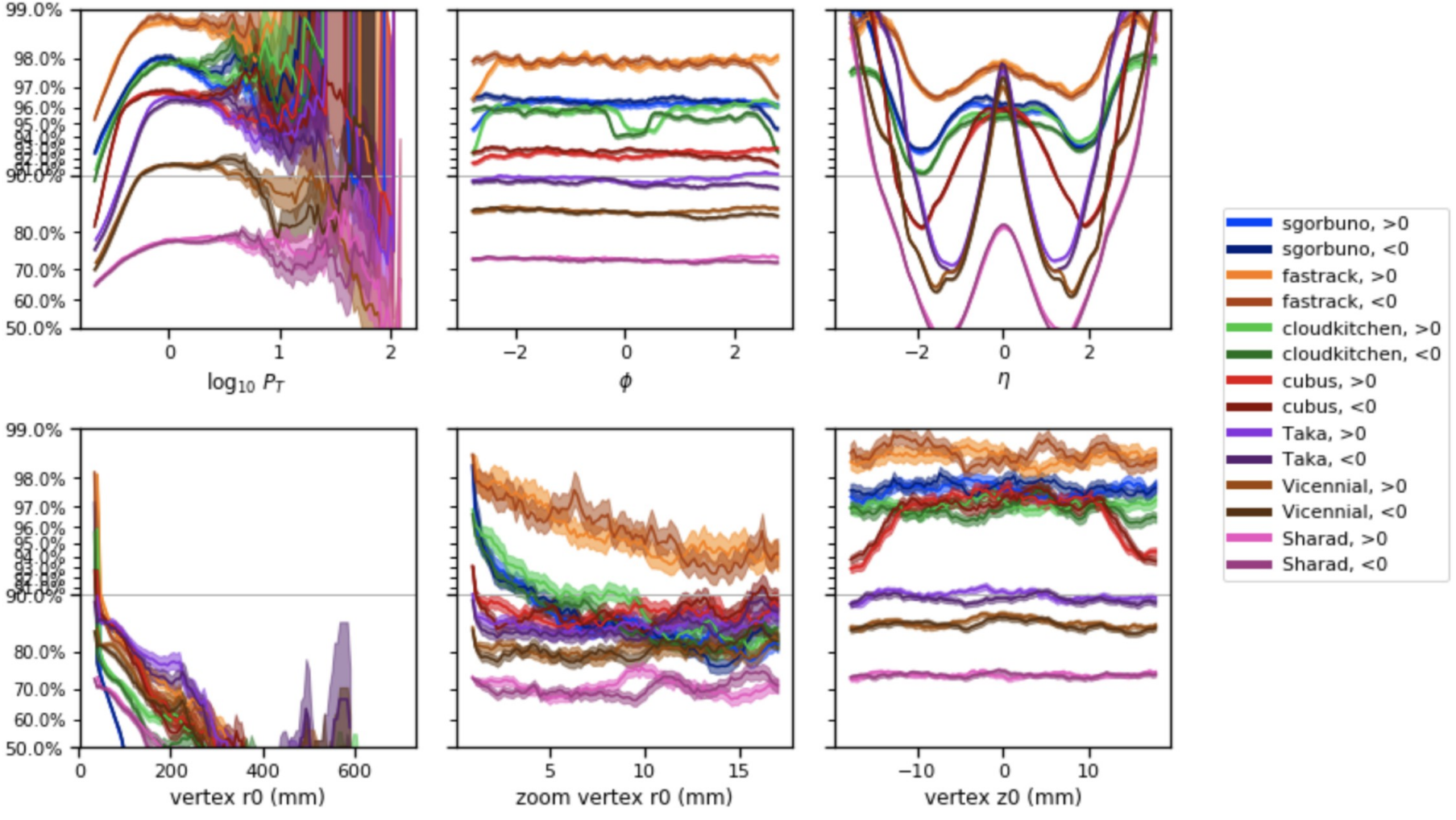
RESULTS									
#	User	Entries	Date of Last Entry	score ▲	accuracy_mean ▲	accuracy_std ▲	computation time (sec) ▲	computation speed (sec/event) ▲	Duration ▲
1	sgorbuno	9	03/12/19	1.1727 (1)	0.944 (2)	0.00 (14)	28.06 (1)	0.56 (1)	64.00 (1)
2	fastrack	53	03/12/19	1.1145 (2)	0.944 (1)	0.00 (15)	55.51 (16)	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)	13.51 (19)	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)
11	yangguo	1	03/01/19	0.0000 (8)	0.082 (15)	0.01 (8)	48.63 (4)	0.97 (4)	86.00 (3)

Time – Accuracy Decomposition



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

Physics Performance



Highest score correlates well with the tracking efficiency

Insights on Algorithms

- **First : *Sgorbuno***
  - Sergey Gorbunov is a physicist, expert in tracking
  - 3rd position at the accuracy phase
  - Iterative steps, triplet seeding, **trajectory following**

- **Second** : ***fastrack***
  - Dmitry Emeliyanov is a physicist
  - Graph representation of neighbors, cellular automaton, track following

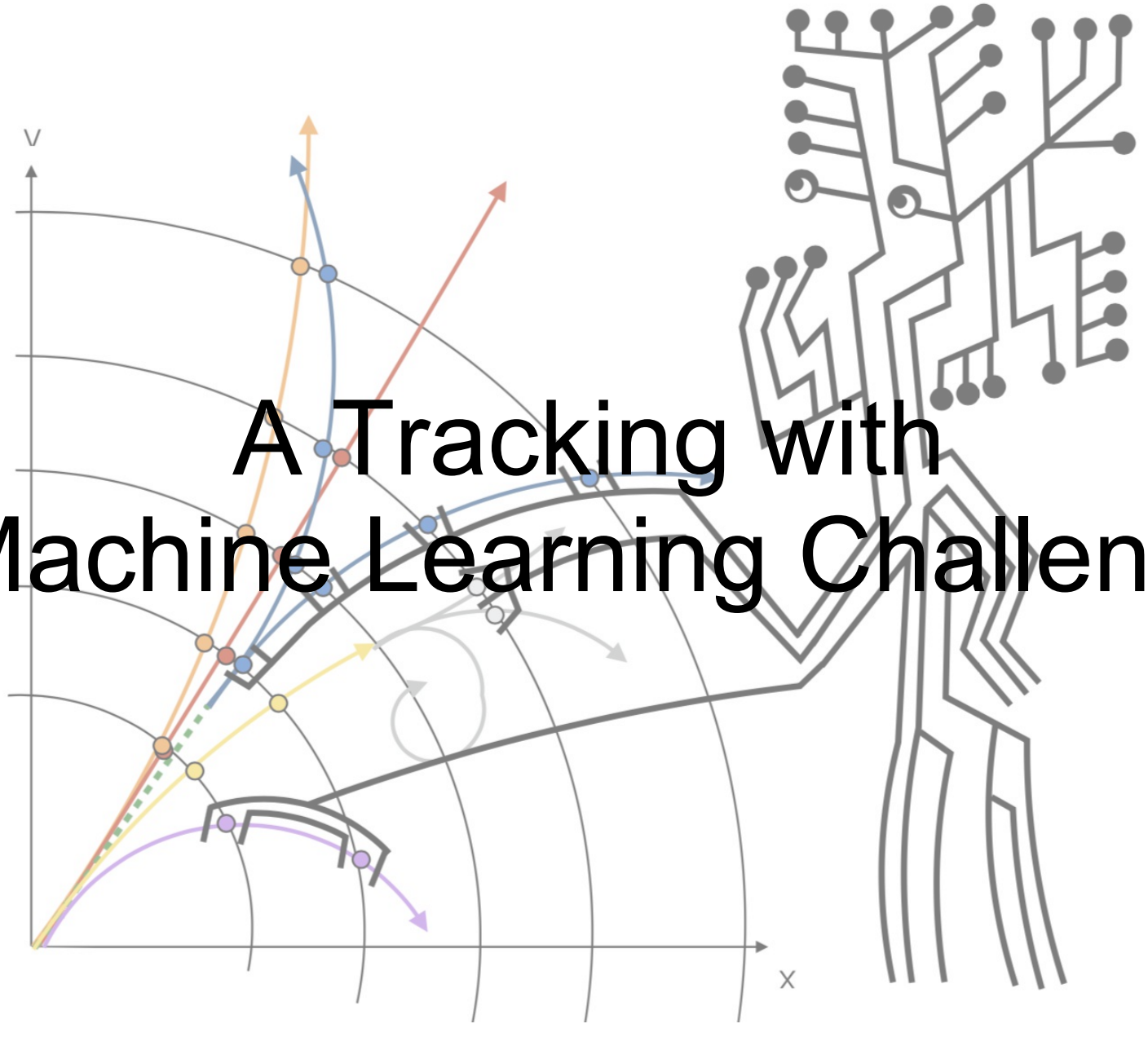
- **Third : *cloudkitchen***
  - Marcel Kunze is a former physicist,
  - Direct acyclic graph of voxels, pair and triplet classification + Top Quark solution of accuracy phase (**trajectory following**, track cleaning, all with **machine learning** for quality selection)

Summary & Outlooks

- Challenges are hard to prepare
  - Finding the right metric is key.
- Timing the code is hard
  - Reason to use codalab. Kaggle will step up
- Challenges are fun to run and to participate to
  - The TrackML Challenge concluded at the July1-2 Grand Finale at CERN https://indico.cern.ch/event/813759/
- Understanding the solutions will take time
  - Several proposed algorithms are being implemented in ACTS for benchmarking
- The challenge datasets were used in several papers
  - A consolidated tracking challenge dataset will be made public on CERN open data portal.

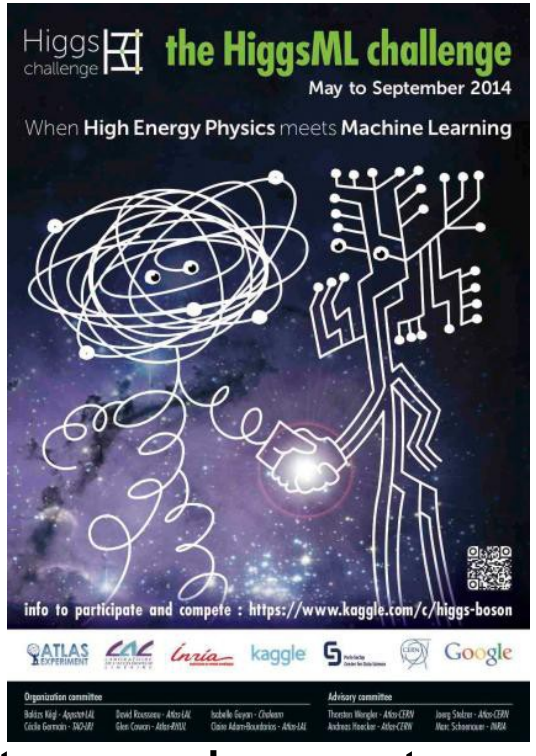

TrackML
sponsors
CERN
kaggle
NVIDIA
UNIVERSITÉ DE GENÈVE
Paris-Saclay Center for Data Science
INSTITUTE DATAIA
Data Science, Intelligence & Society
CHA LEARN
IRIS-HEP
CERN openlab
Inria
PyTorch
cnrs
erc
Fermilab
COMMON GROUND
inqnet
Alliance for Quantum Technologies
UNIVERSITE PARIS SUD
université PARIS-SACLAY

Extra Material



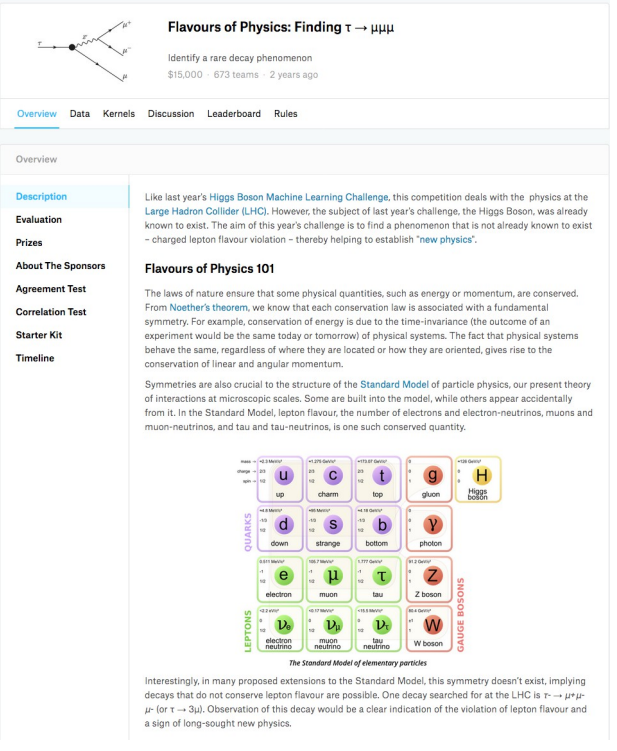
A Tracking with Machine Learning Challenge

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

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