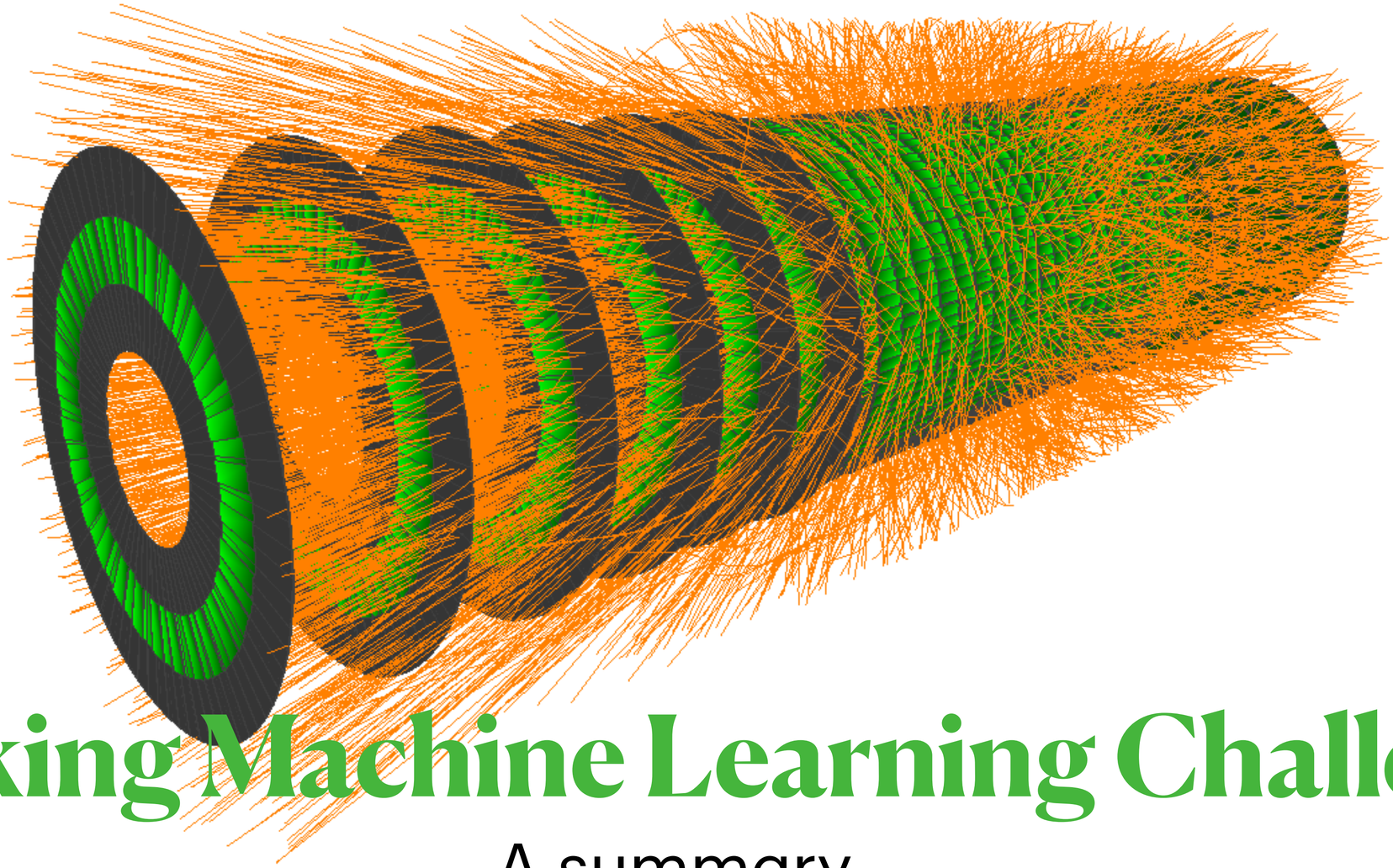
ICHEP 2020 | PRAGUE 28 July 2020 to 6 August 2020, virtual conference



Tracking Machine Learning Challenge

A summary

@SaltyBurger

A. Salzburger (CERN) for the organisers

Organisation

Phase 1

kaggle

focus on accuracy

Apr 13, 2018 → Aug 14, 2018

Phase 2

CodaLab

accuracy & speed

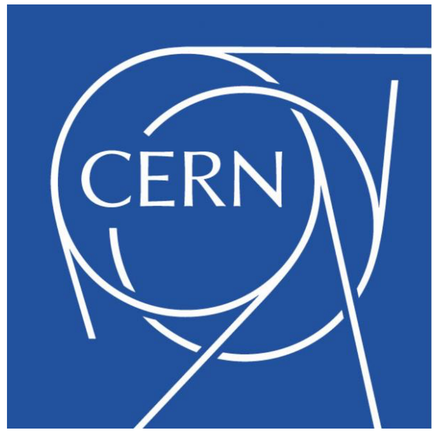
Sep 07, 2018 → Nov 12, 2018

Organisation team

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)

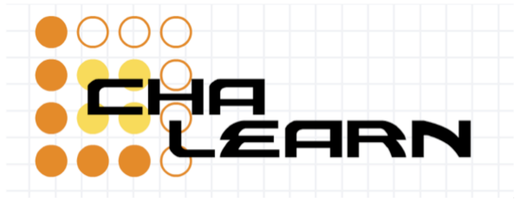
Summary publications TrackML Particle Tracking Machine Learning Challenge [Phase 1][Phase 2]

Sponsors



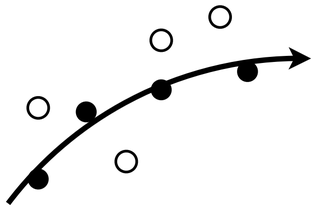
UNIVERSITÉ DE GENÈVE

Paris-Saclay Center for Data Science



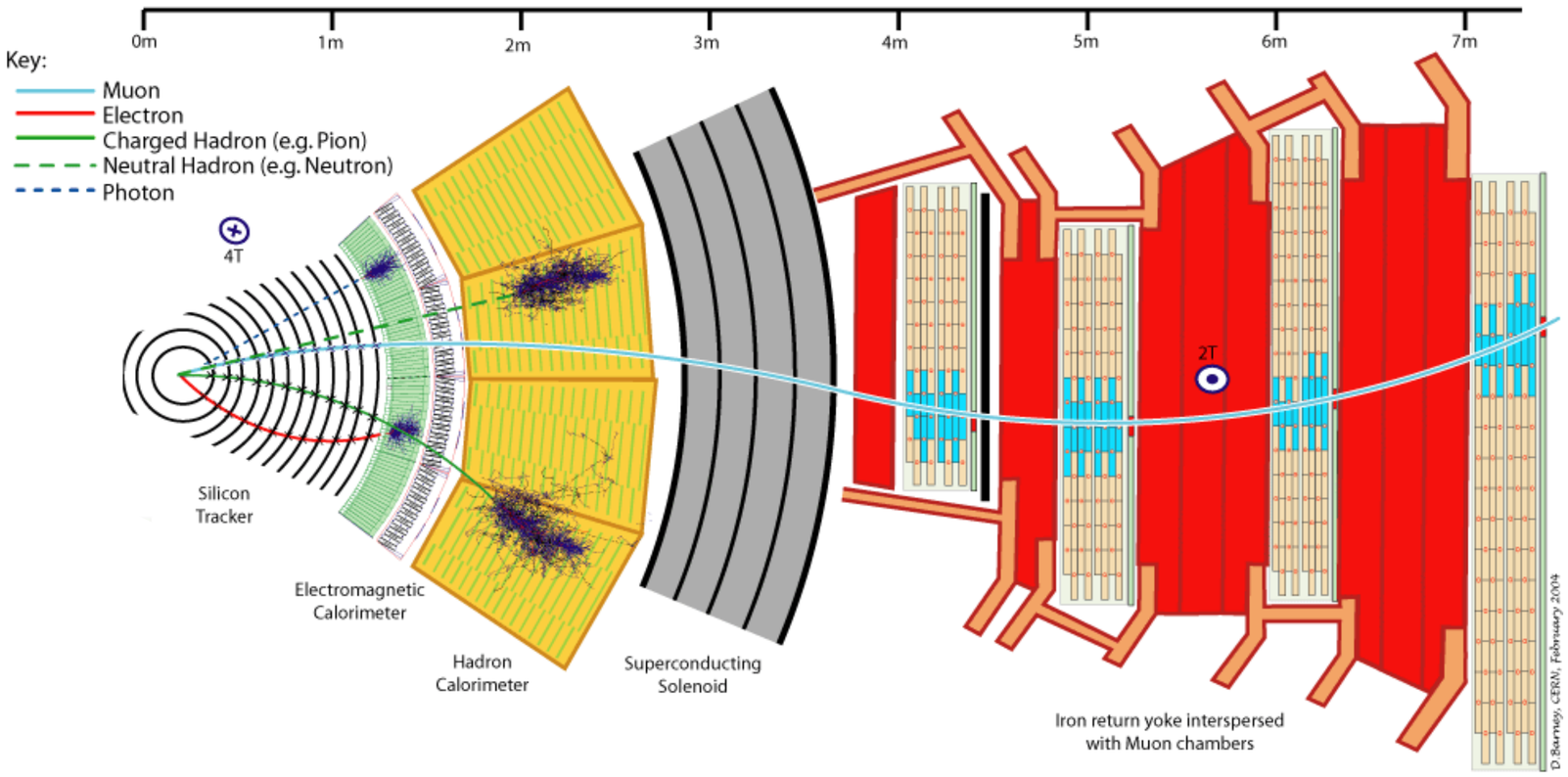
iris hep

Institute for Research & Innovation in Software for High Energy Physics

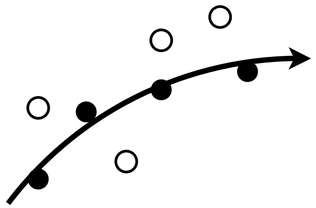


Particle Tracking

Trajectory and vertex finding in tracking detectors

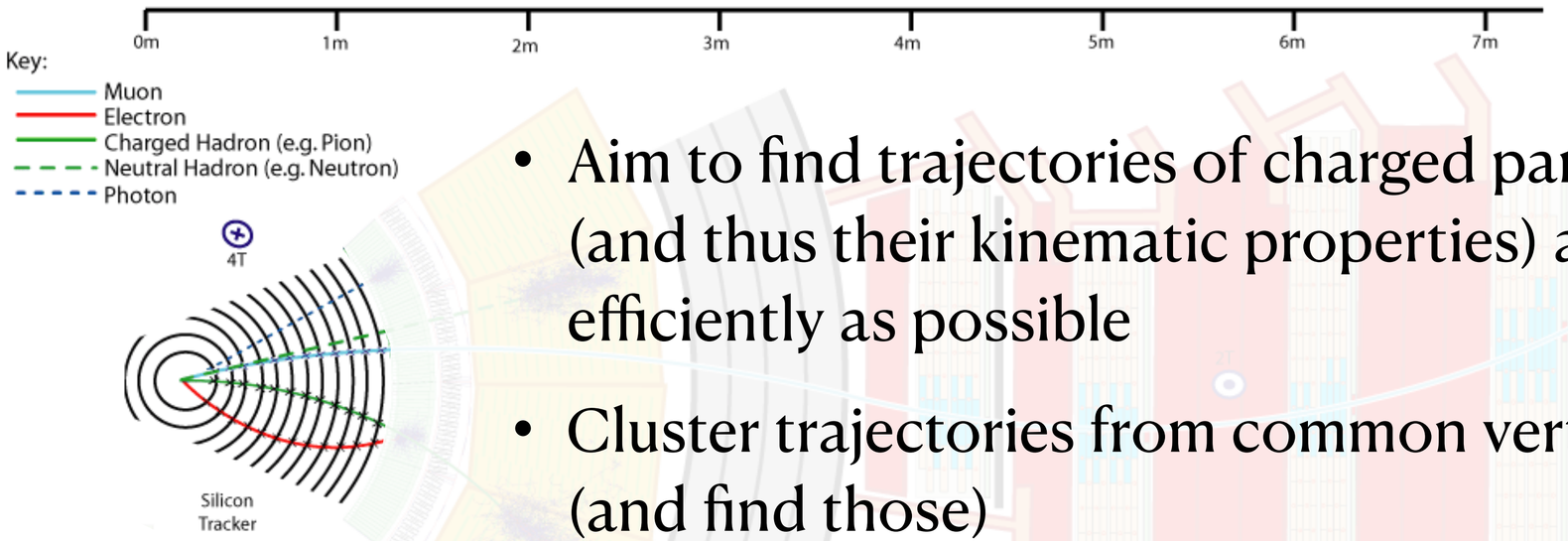


Slice of the CMS detector



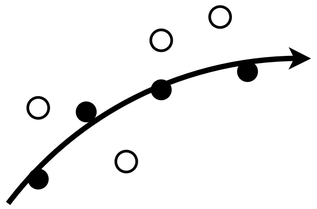
Particle Tracking

Trajectory and vertex finding in tracking detectors



- Aim to find trajectories of charged particles (and thus their kinematic properties) as efficiently as possible
- Cluster trajectories from common vertices (and find those)
- If possible, first particle identification
- Avoid: fake/ghost trajectories, duplicates, ...

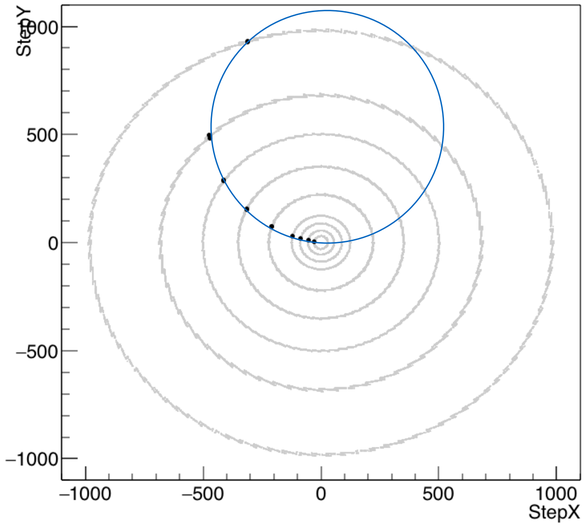
Slice of the CMS detector



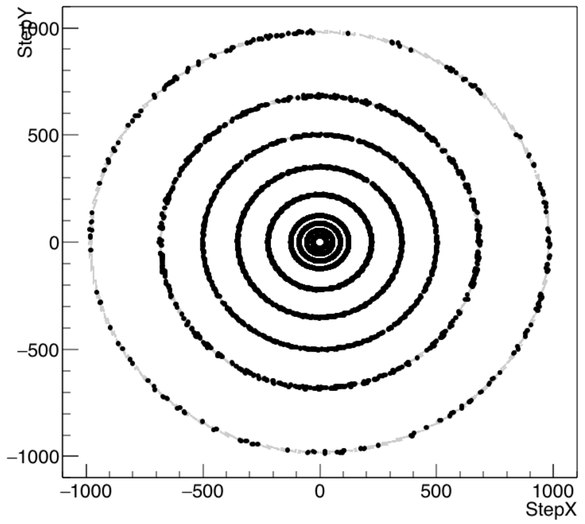
Particle Tracking

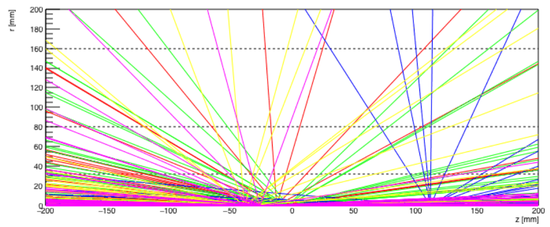
Complexity with increasing pile-up

Hits from one particle

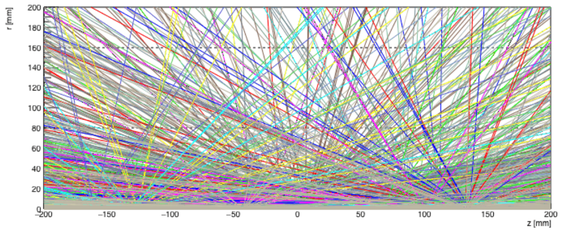


Fraction of hits of a <μ>~200 event

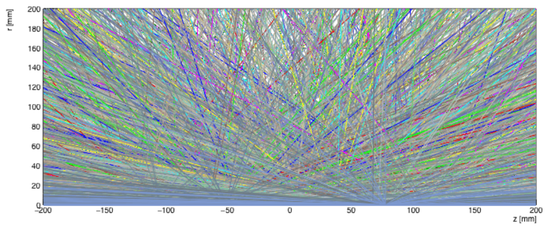




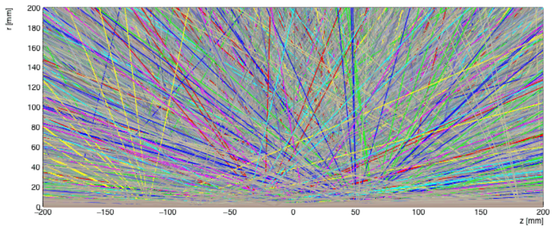
LHC Run-1, <μ> ~ 5



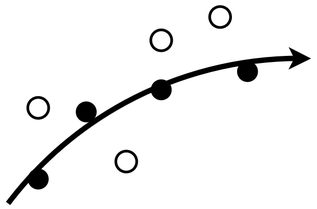
LHC Run-2, <μ> ~ 20



HL-LHC, <μ> ~ 200

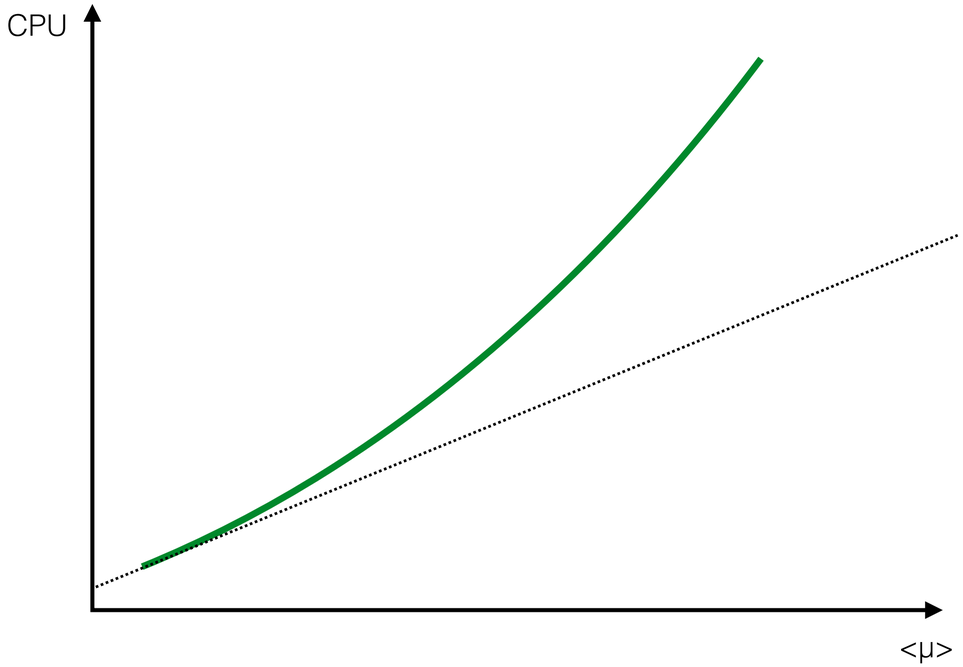


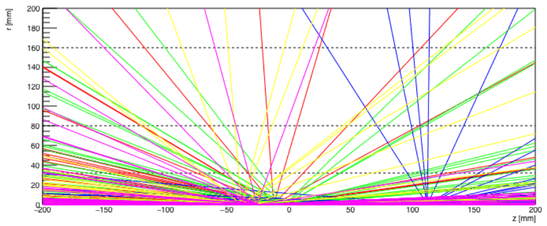
FCC-hh (25ns) <μ> ~ 1000



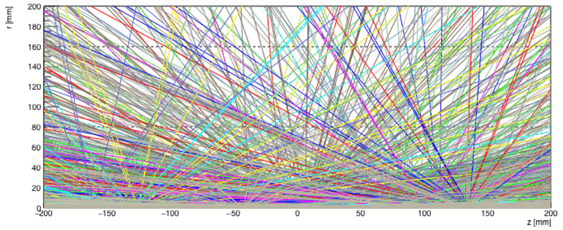
Particle Tracking

A combinatorial / computational challenge

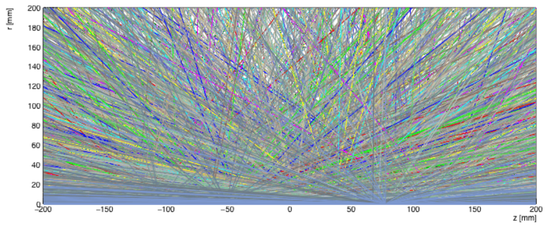




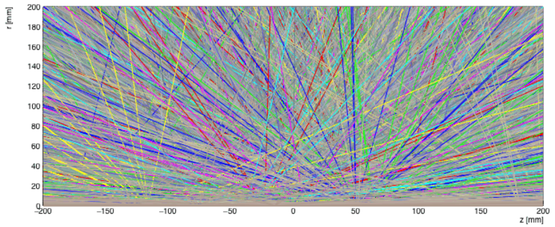
LHC Run-1, <μ> ~ 5



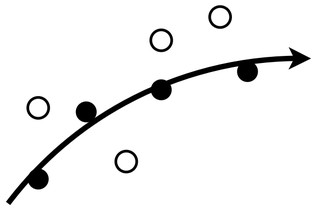
LHC Run-2, <μ> ~ 20



HL-LHC, <μ> ~ 200



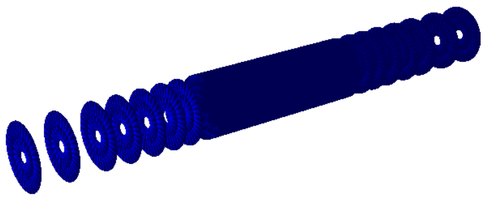
FCC-hh (25ns) <μ> ~ 1000

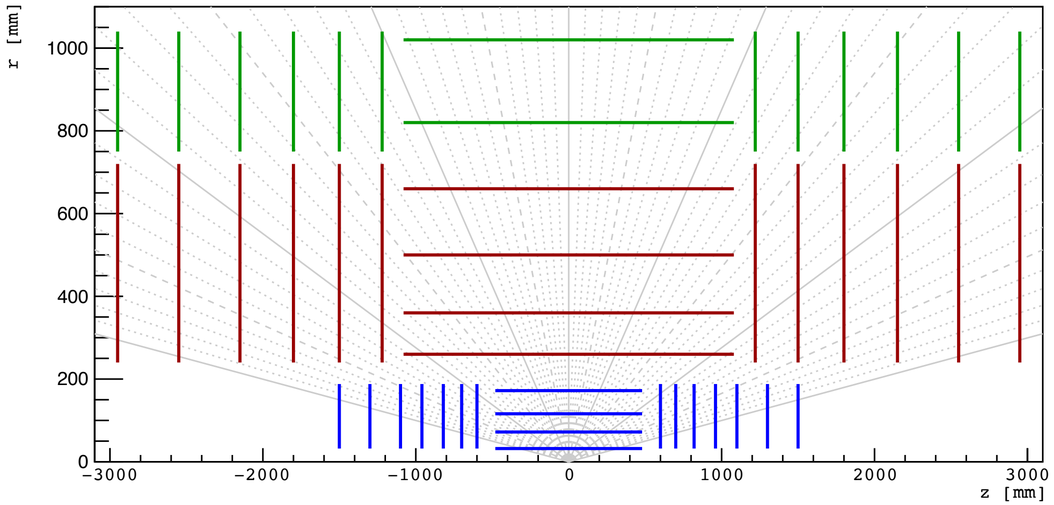


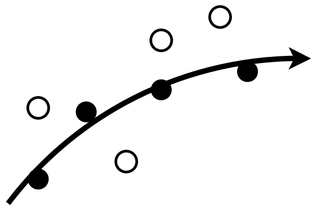
The Challenge

The TrackML detector - Innermost Pixel System

4 barrel-layer central pixel system
7 EC discs (both sides)

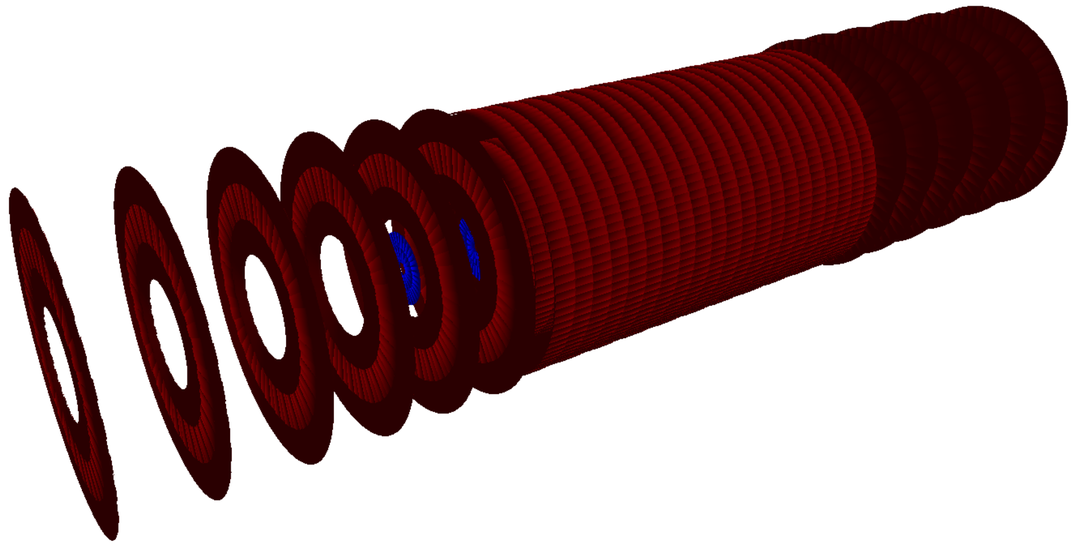




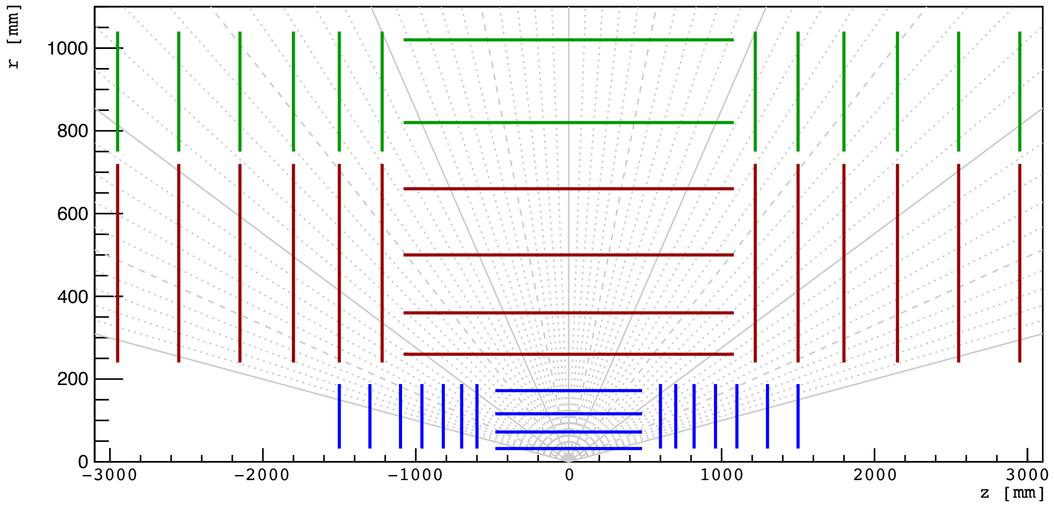


The Challenge

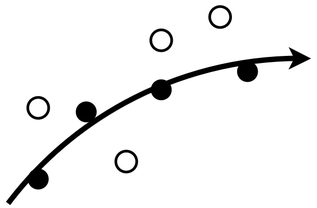
The TrackML detector - Short Strip System



4 barrel-layer central strip system
6 EC discs (both sides)

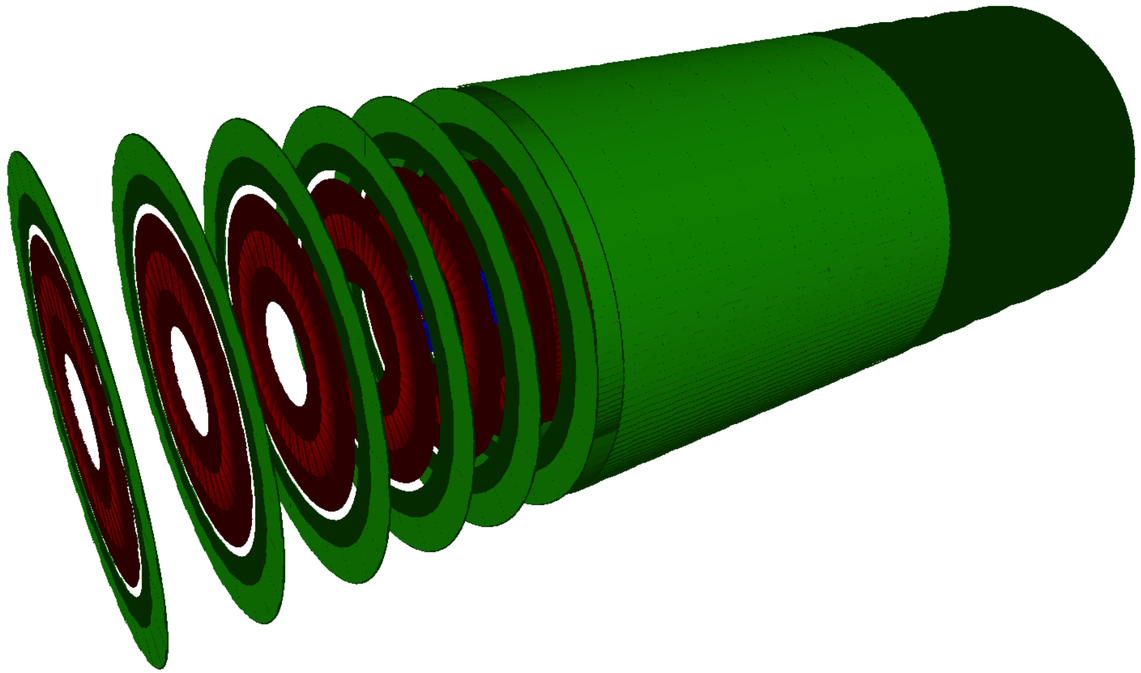


Coverage | η | < 3

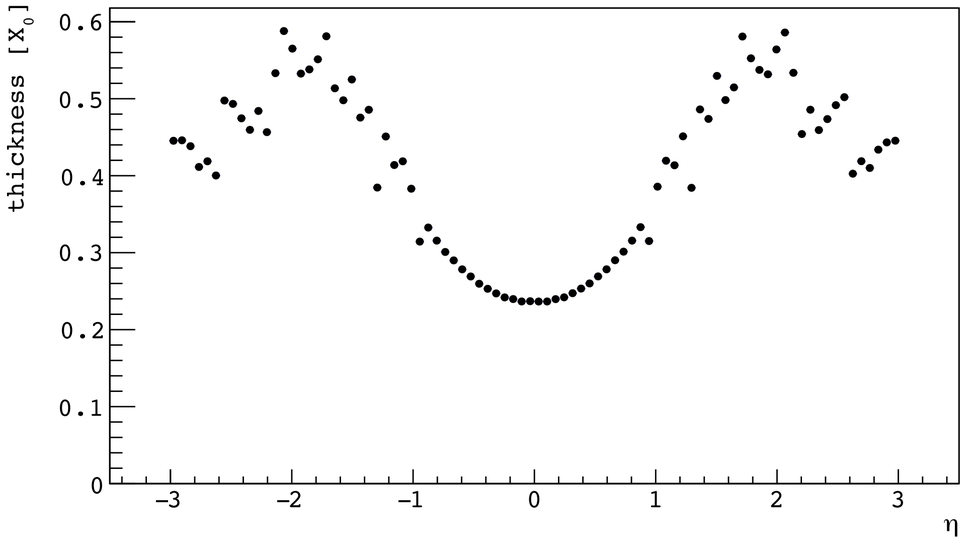


The Challenge

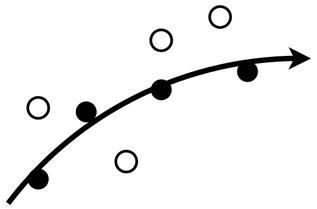
The TrackML detector - Long Strip System



2 barrel-layer outer strip system
6 EC discs (both sides)

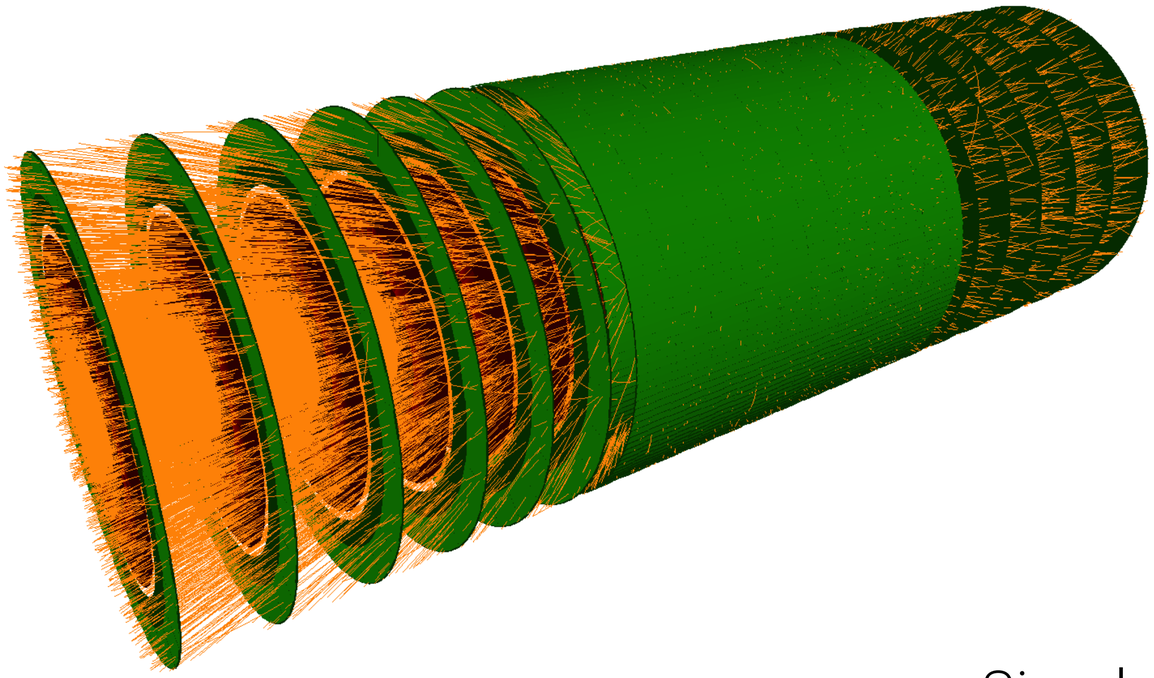


Optimistic material budget



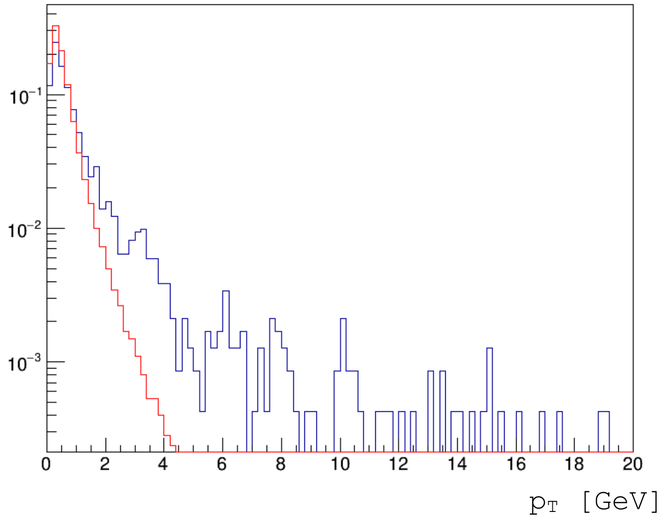
The Challenge

A TrackML event



Signal event: top quark pair production

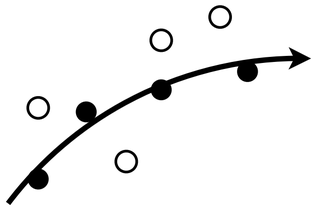
Overlaid pile-up: <μ> = 200



Simulated with acts fast track simulation

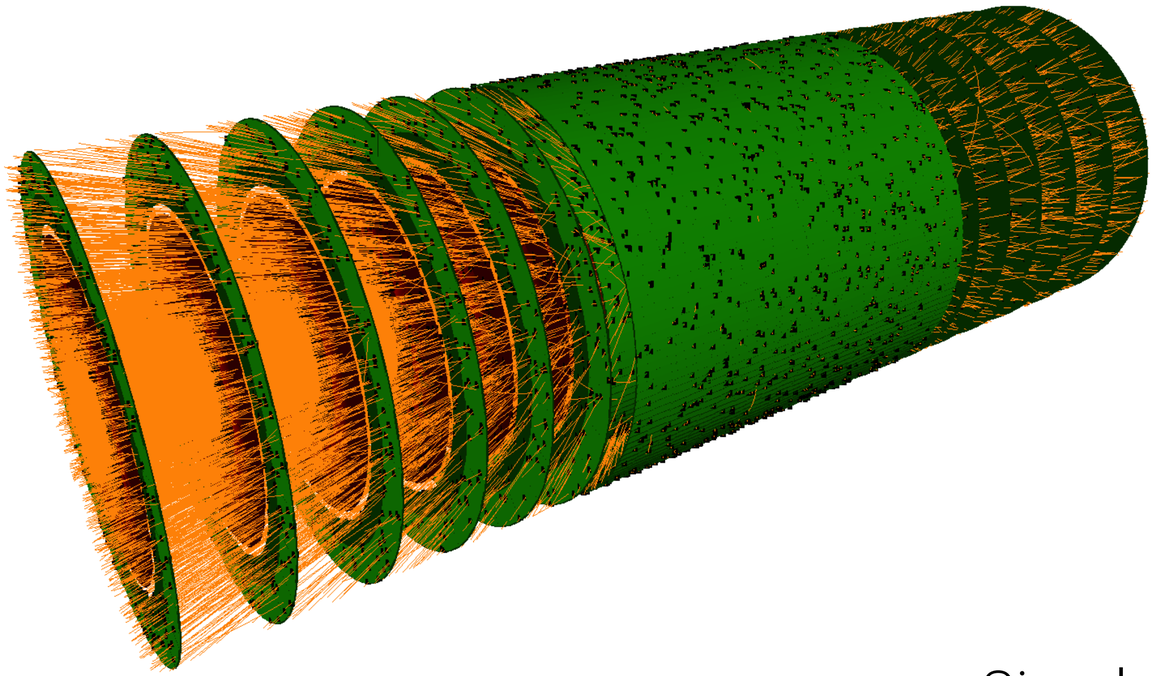
→ 10k particles, 100k hits / event

acts Link to [repository]



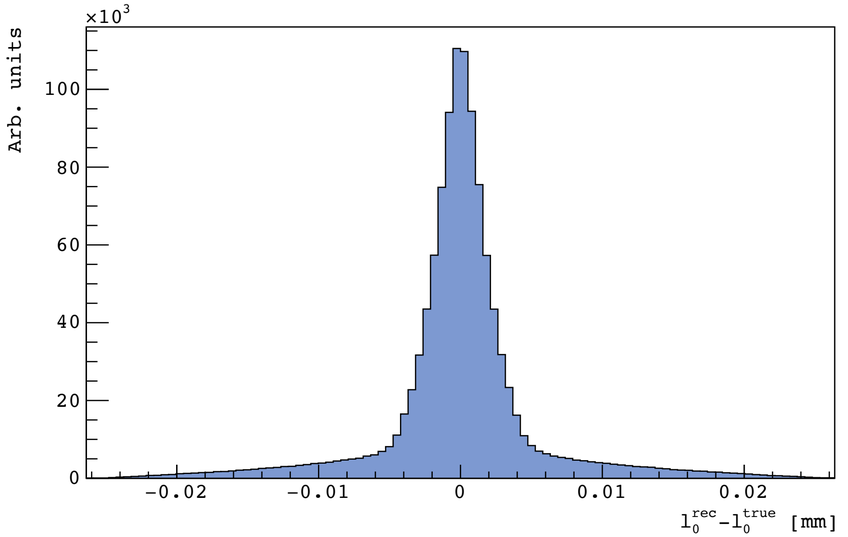
The Challenge

A TrackML event



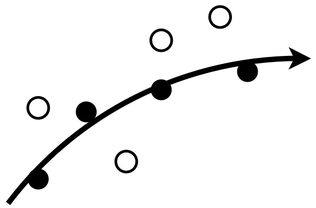
Quasi-realistic resolution

- Using geometric digitisation approach
- Non-gaussian hit residuals



Simulated with a:ts fast track simulation

a:ts Link to [repository]

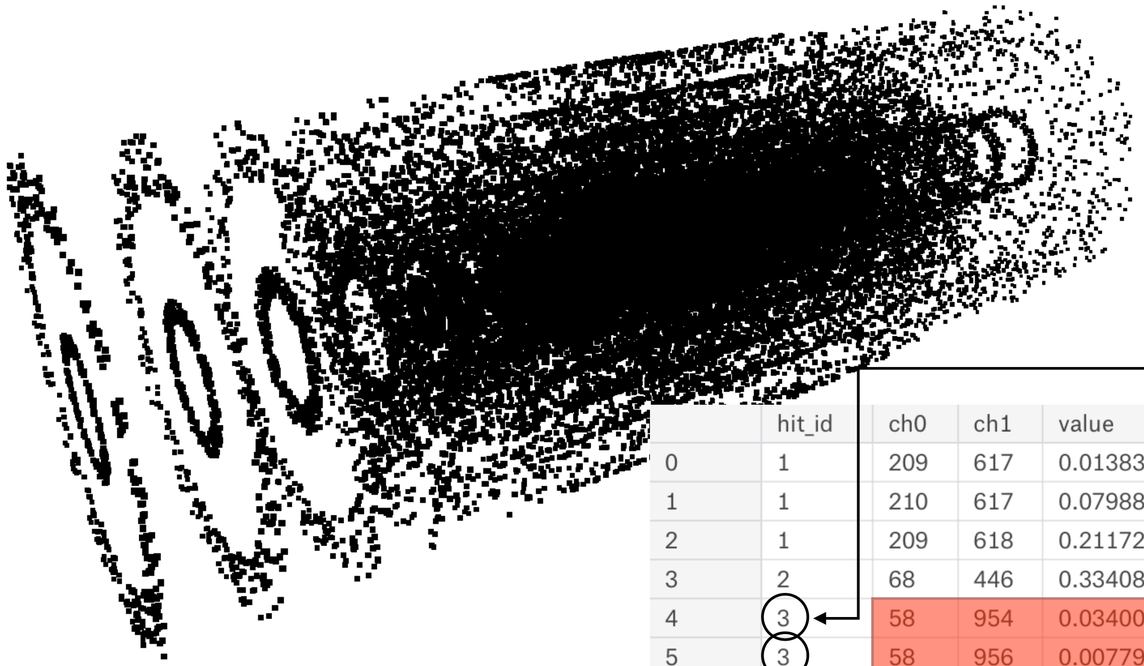


The Challenge

A TrackML event - The Dataset

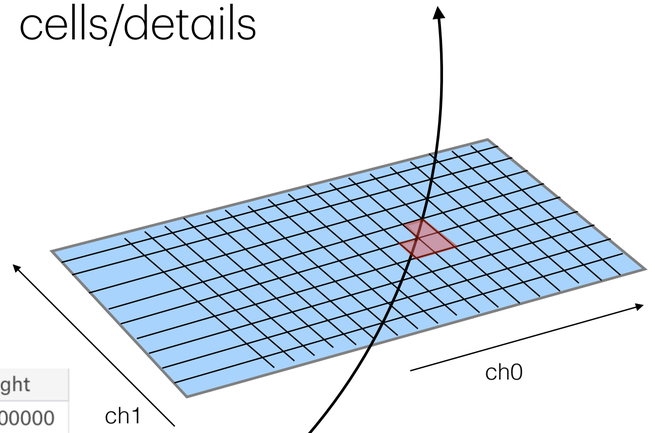
hits

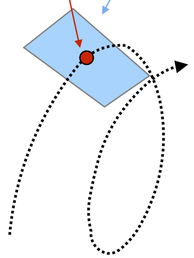
	hit_id	x	y	z	volume_id	layer_id	module_id
0	1	-64.409897	-7.163700	-1502.5	7	2	1
1	2	-55.336102	0.635342	-1502.5	7	2	1
2	3	-83.830498	-1.143010	-1502.5	7	2	1
3	4	-96.109100	-8.241030	-1502.5	7	2	1
4	5	-62.673599	-9.371200	-1502.5	7	2	1
5	6	-57.068699	-8.177770	-1502.5	7	2	1
6	7	-73.872299	-2.578900	-1502.5	7	2	1
7	8	-63.853500	-10.868400	-1502.5	7	2	1
8	9	-97.254799	-10.889100	-1502.5	7	2	1
9	10	-90.292900	-3.269370	-1502.5	7	2	1
10	11	-59.182999	-0.670508	-1502.5	7	2	1



	hit_id	ch0	ch1	value
0	1	209	617	0.013832
1	1	210	617	0.079887
2	1	209	618	0.211723
3	2	68	446	0.334087
4	3	58	954	0.034005
5	3	58	956	0.007798
6	3	60	951	0.019897

cells/details



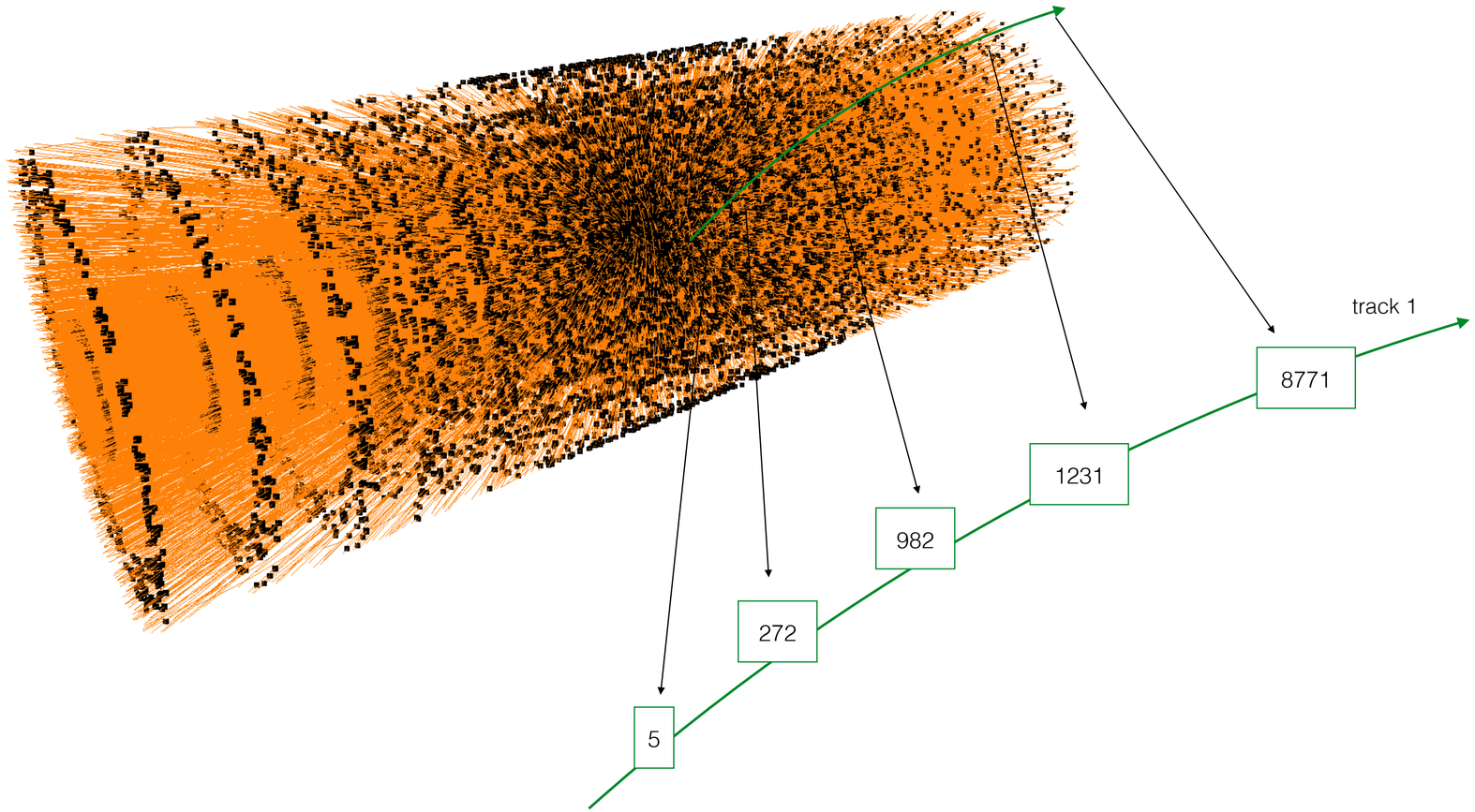


truth

	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000	0.000000

The Challenge

A TrackML event - A solution



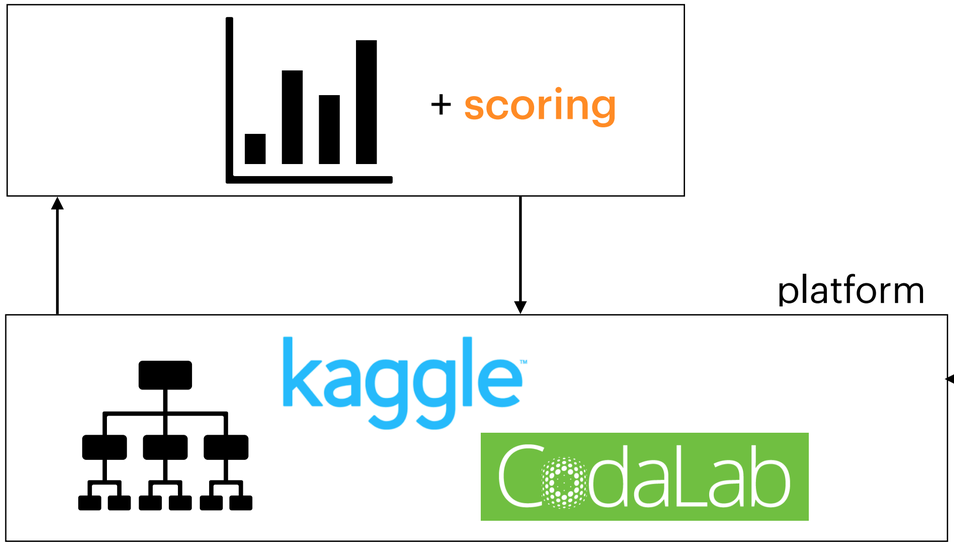
solution.csv

	hit_id	track_id
1	5	1
2	272	1
3	982	1
4	1231	1
5	18771	1

labelled set of hits

Submission

And scoring

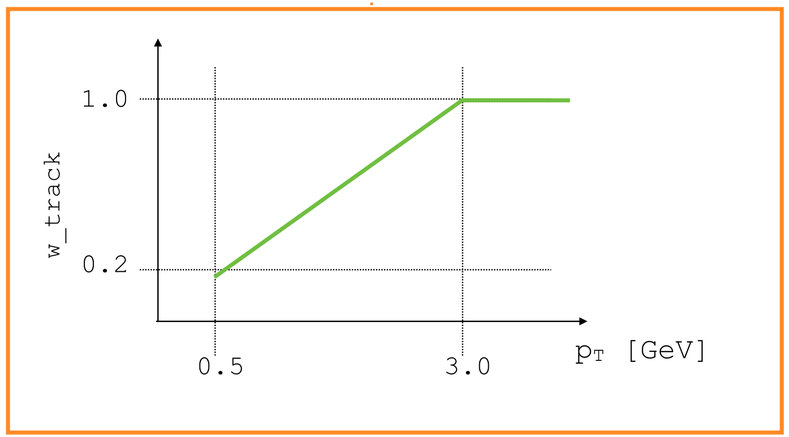


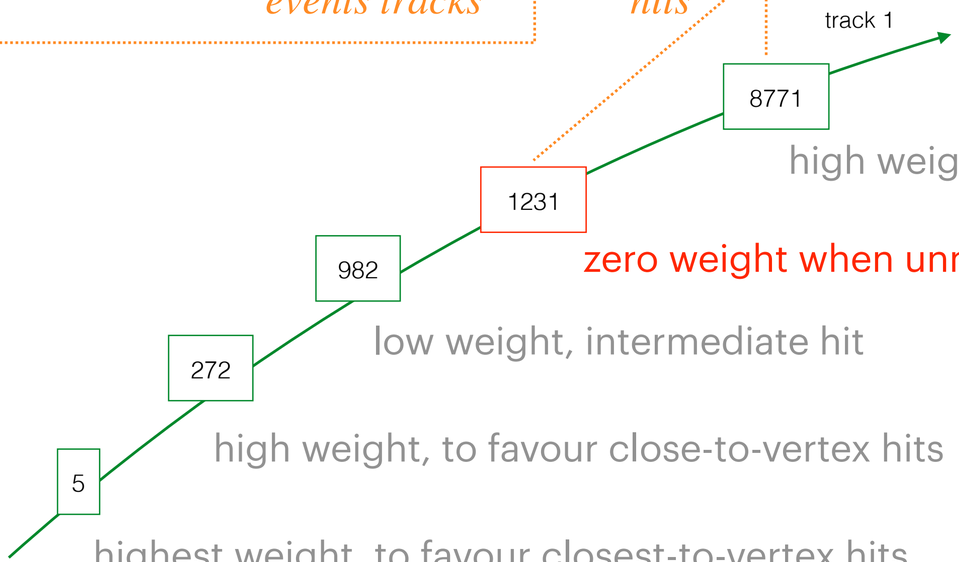
solution.csv

participant A

	hit_id	track_id
1	5	1
2	272	1
3	982	1
4	1231	1
5	18771	1

$$score = \sum_{events} \sum_{tracks} w_{track} \sum_{hits} w_{hit}^{matched}$$





high weight, to favour track completeness

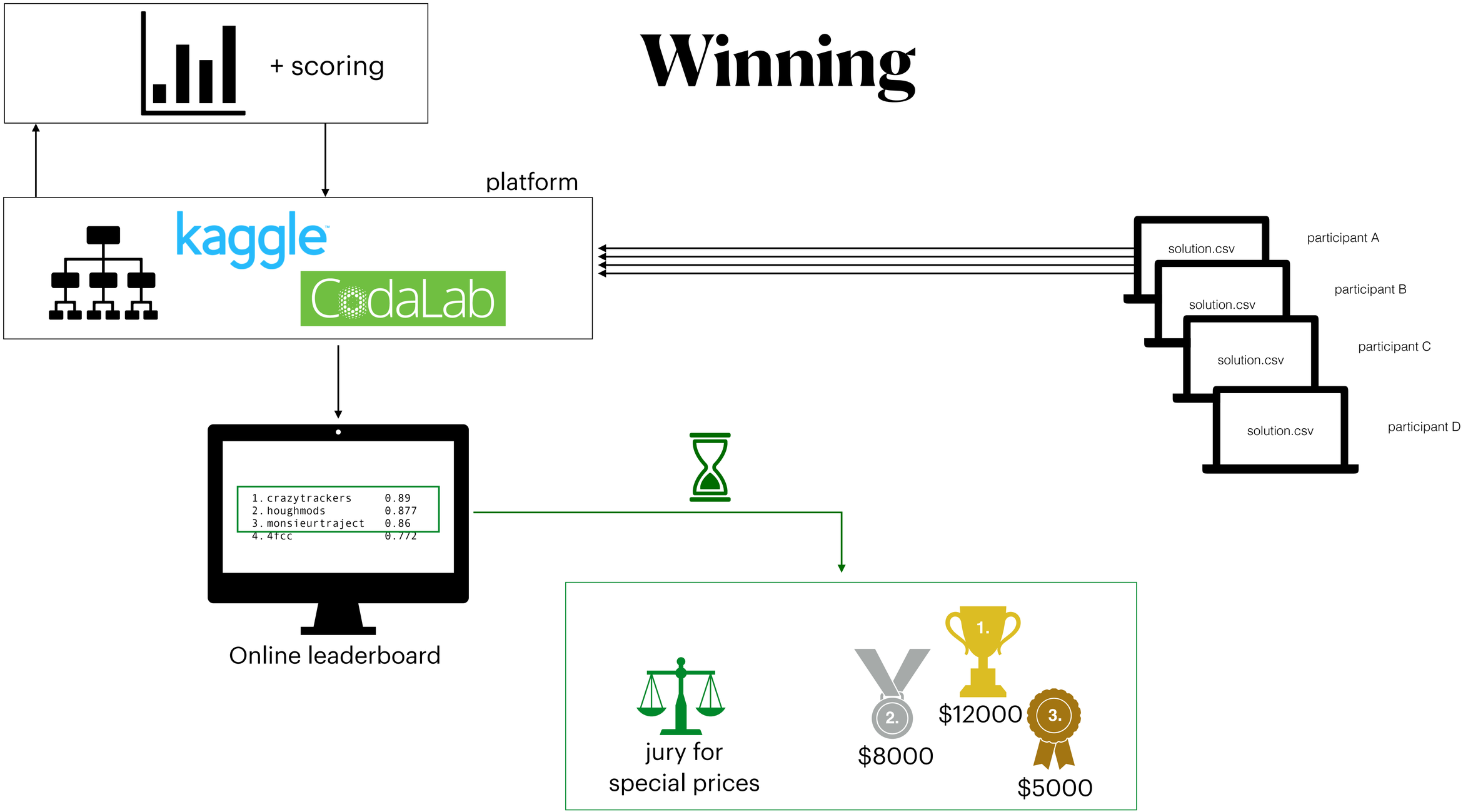
zero weight when unmatched with truth

low weight, intermediate hit

high weight, to favour close-to-vertex hits

highest weight, to favour closest-to-vertex hits

Winning


+ scoring
platform
kaggle
CodaLab
solution.csv
participant A
solution.csv
participant B
solution.csv
participant C
solution.csv
participant D
1. crazytrackers 0.89
2. houghmods 0.877
3. monsieurtraject 0.86
4. 4fcc 0.772
Online leaderboard
jury for
special prices
1.
2.
3.
$12000
$8000
$5000

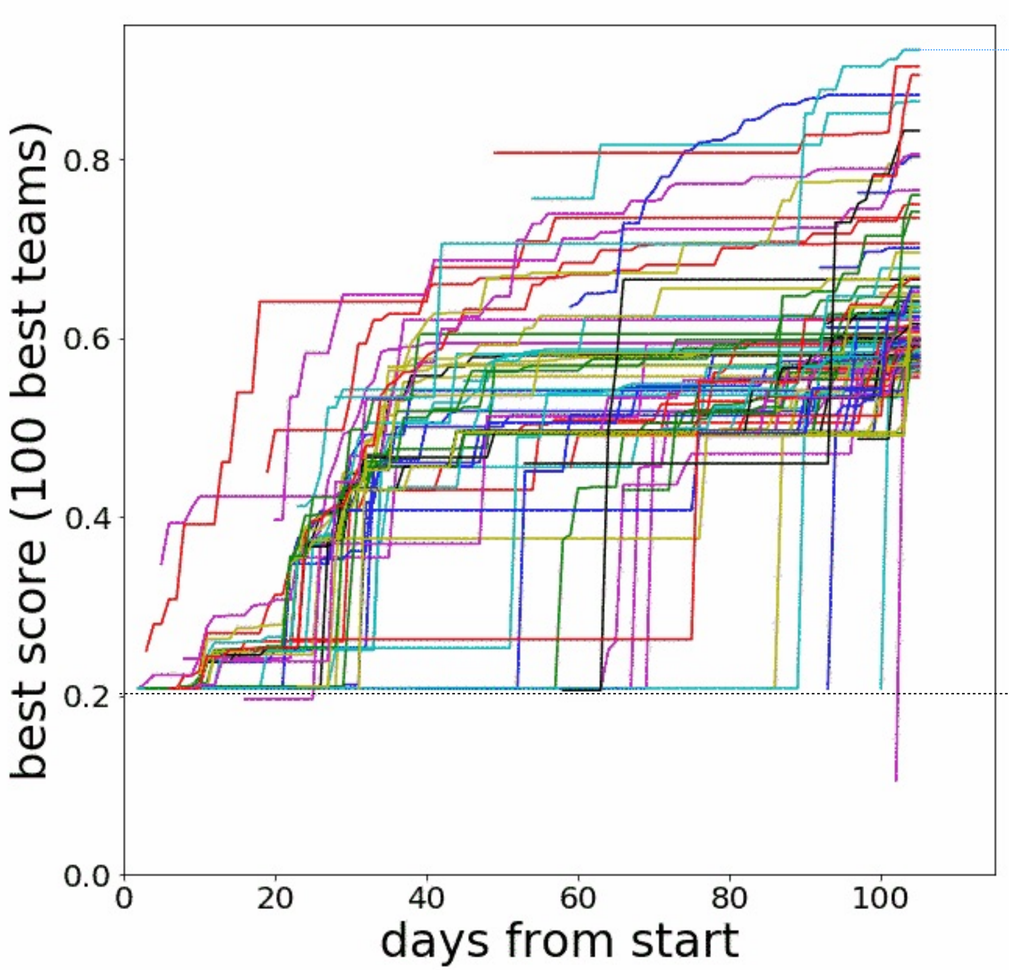
Phase 1

Accuracy Challenge

kaggle

Apr 13, 2018 → Aug 14, 2018

Phase 1 - Score Evolution



0.92

score of the the starter kit (DBScan)

At final date:

656	776	5,837
Teams	Competitors	Entries

Phase 1 - Final Leaderboard

Public Leaderboard | Private Leaderboard

The private leaderboard is calculated with approximately 71% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Refresh

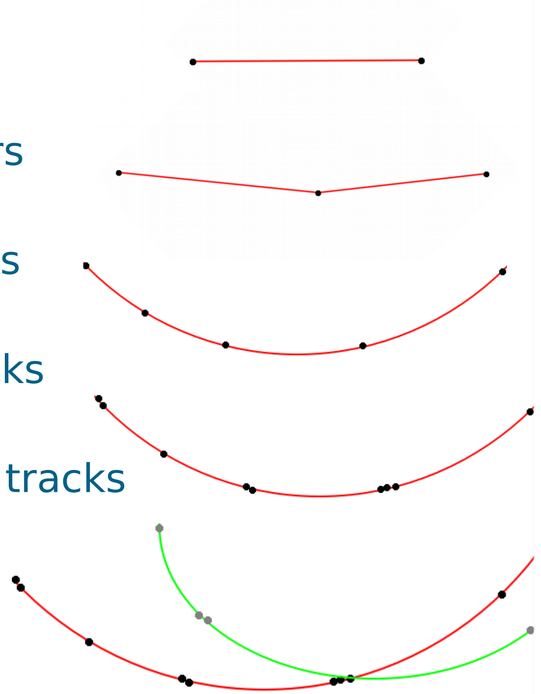
In the money | Gold | Silver | Bronze

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	Top Quarks			0.92182	10	2mo
2	—	outrunner			0.90302	9	2mo
3	—	Sergey Gorbunov			0.89353	6	2mo
4	—	demelian			0.87079	35	2mo
5	—	Edwin Steiner			0.86395	5	2mo
6	—	Komaki			0.83127	22	2mo
7	—	Yuval & Trian			0.80414	56	2mo
8	—	bestfitting			0.80341	6	2mo

Phase 1 Top Quarks

Main steps

- Select promising pairs
  - 7 million / 0.99
- Extend pairs to triples
  - 12 million / 0.97
- Extend triples to tracks
  - 12 million / 0.95
- Add duplicate hits to tracks
  - 12 million / 0.96
- Assign hits to tracks
  - 90% of hits / 0.92



- Logistic regression for track candidate pruning

- Pure C++, some scikit-learn for training

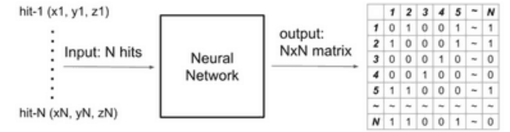
Findings

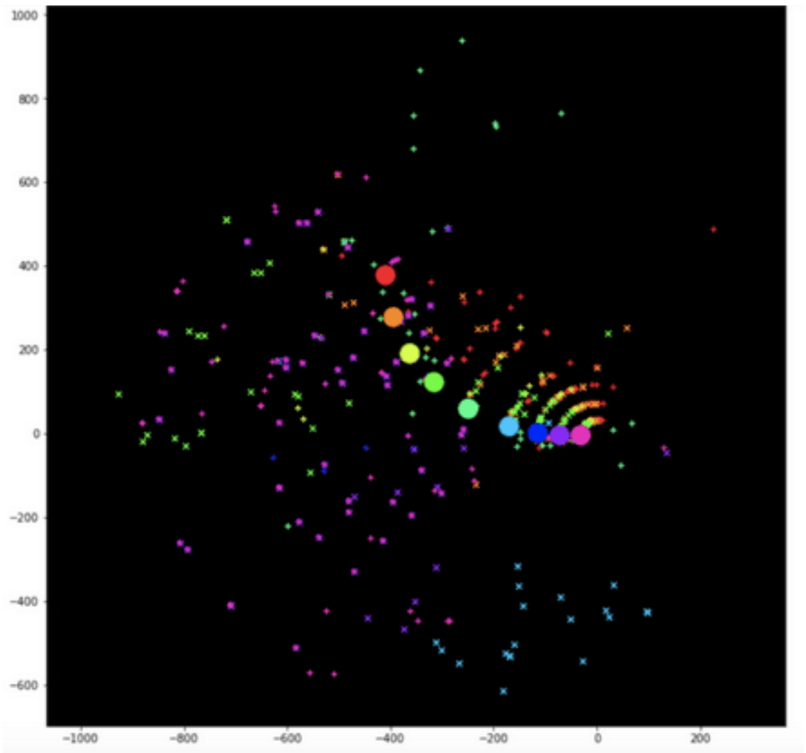
- No magic formula
  - We won because we were fast to try out and implement many ideas and got the details right
    - I once earned 0.03 (0.85→0.88) from fixing a tuning parameter
  - In other words: combination of many factors

Author	Johan S. Wind
<Wall time>/evt	7m 17s
Peak memory	2.78 Gb

Phase 1 2. Outrunner

- ML approach using python & keras
  - event with N hits
  - predict N x N relationships between hits, connect pairs when their probability is 1
- Training
  - 5 hidden layers with 4k - 2k - 2k - 1k nodes
  - 27 input features
- Prediction & Reconstruction
  - predict pairwise relationship probability
  - finding highest probability pair & add pairwise

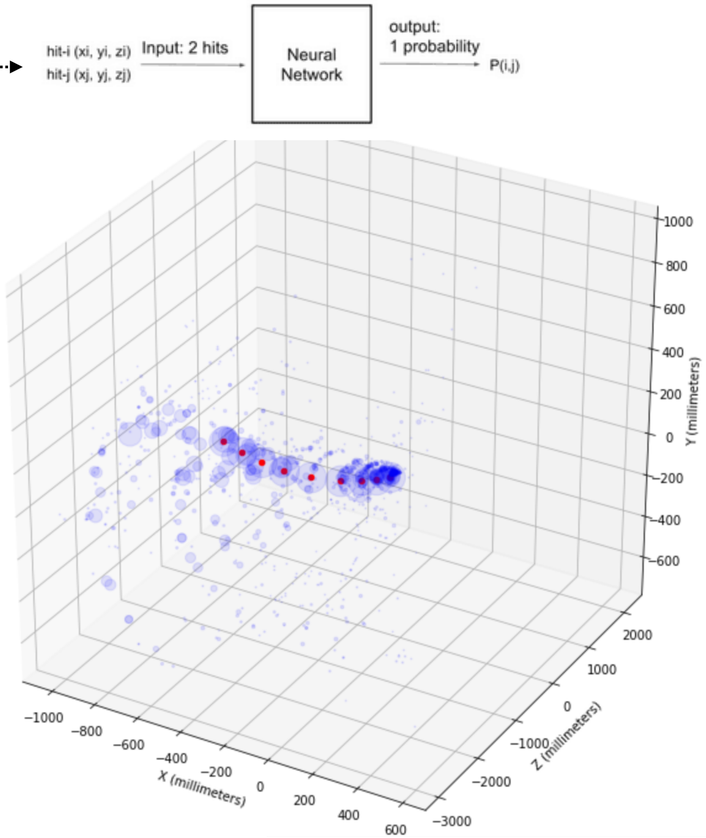




Author	Pei-Lien Chou
<Wall time>/evt	~1 day

Phase 1 2. Outrunner

- ML approach using python & keras
  - event with N hits
  - predict N x N relationships between hits, connect pairs when their probability is 1
- Training
  - 5 hidden layers with 4k - 2k - 2k - 1k nodes
  - 27 input features
- Prediction & Reconstruction
  - **predict** pairwise relationship probability
  - **finding** highest probability pair & **add pairwise**

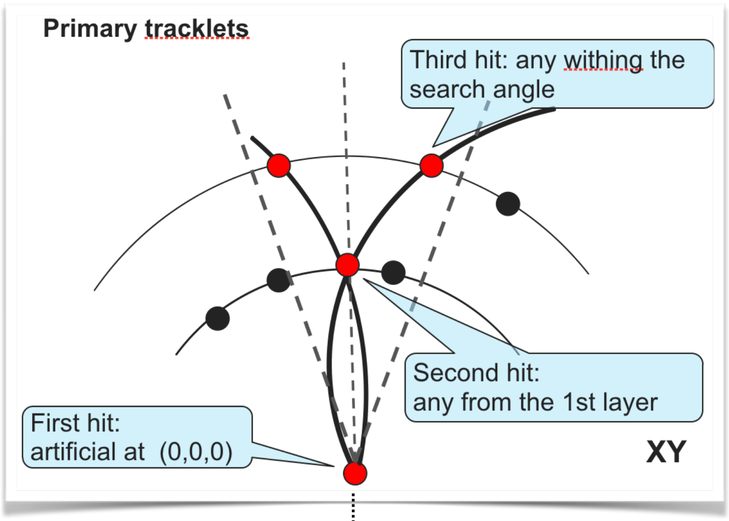


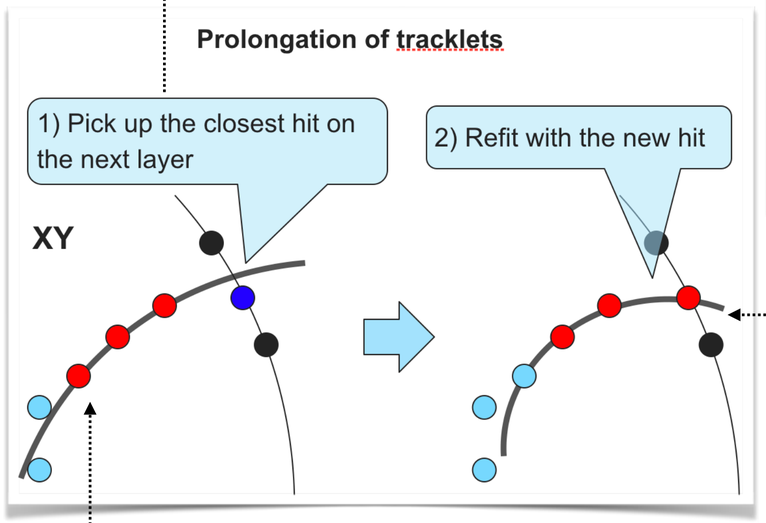
Author	Pei-Lien Chou
<Wall time>/evt	~1 day

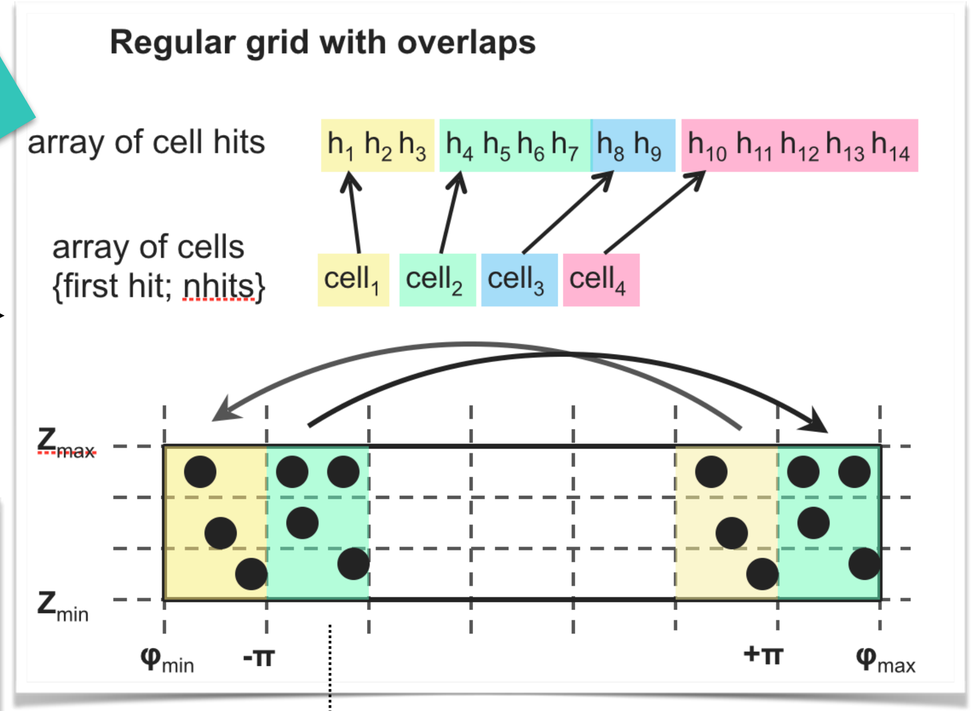
Phase 1 3. Sergey Gorbunov

HEP

- Combinatorial approach based on track following
- No search branches
- Simple track model (Helix)







C++ CPP	Author	[Sergey Gorbunov]
	<Wall time>/evt	~1.2 mins

Phase 1 Jury Prizes

Innovation prize

Yuval Reina & Trian Xylouris
Marginalized Hough transform with machine learning classifier

Clustering prize

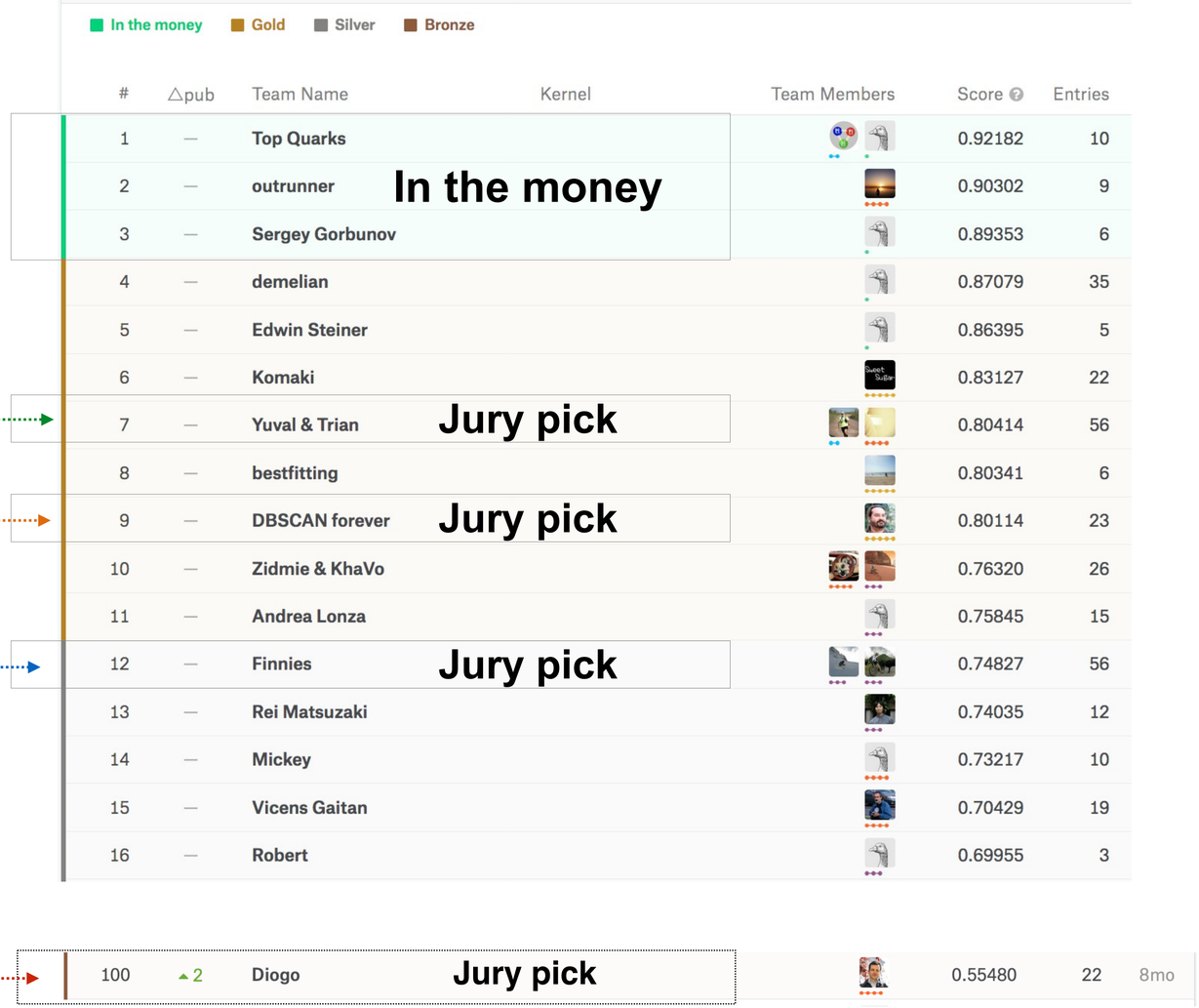
Jean-Francois Puget (kaggle grandmaster)
DBScan clustering with iterative Hough transform

Deep Learning prize

Nicole & Liam Finnie
DBScan seeding and LSTM track Building

Organizer's pick

Diogo R. Ferreira
Innovative pattern matching

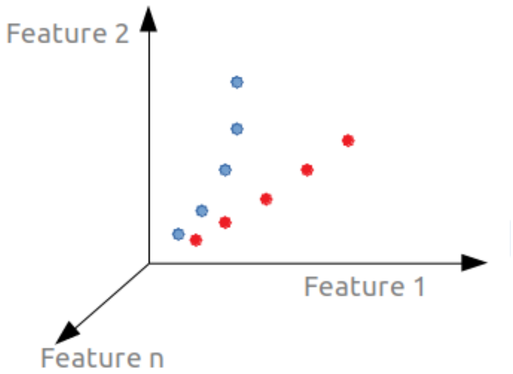


Phase 1 DL Prize

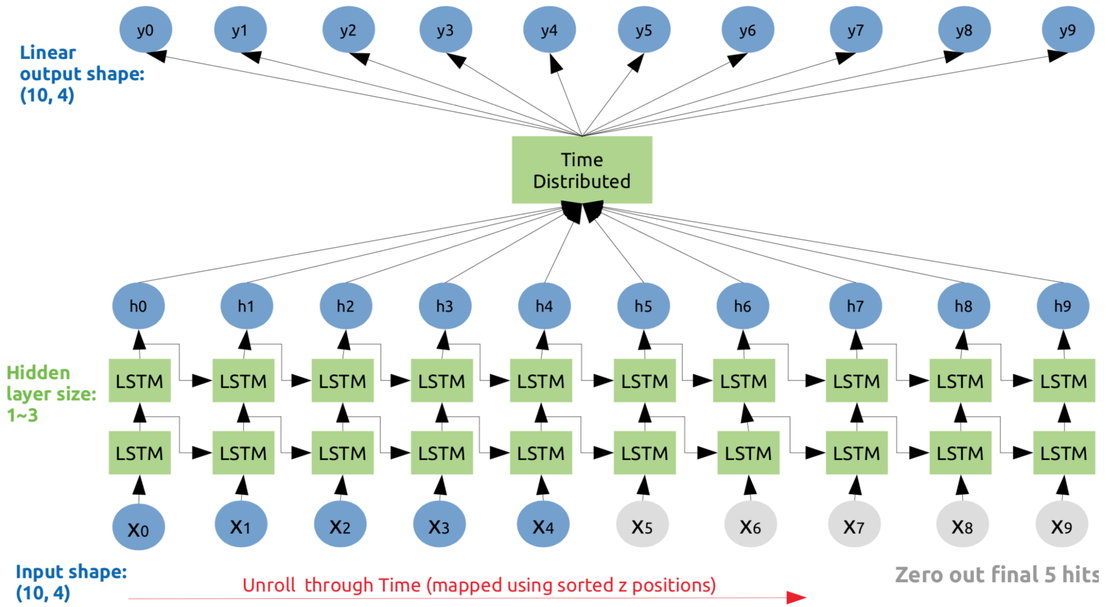
HEP

Seeding → **Inference & Assembling** → **Fitting**

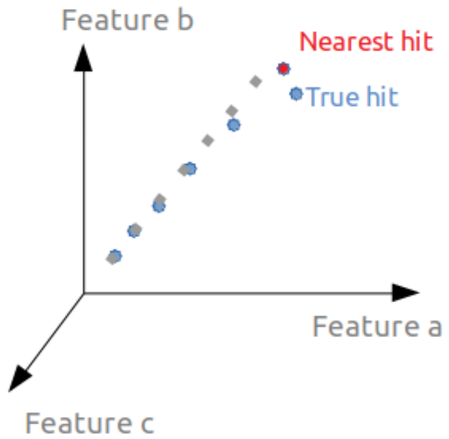
DBScan | Recurrent NN | KNN-tree



10 output sets of hit quadruplets



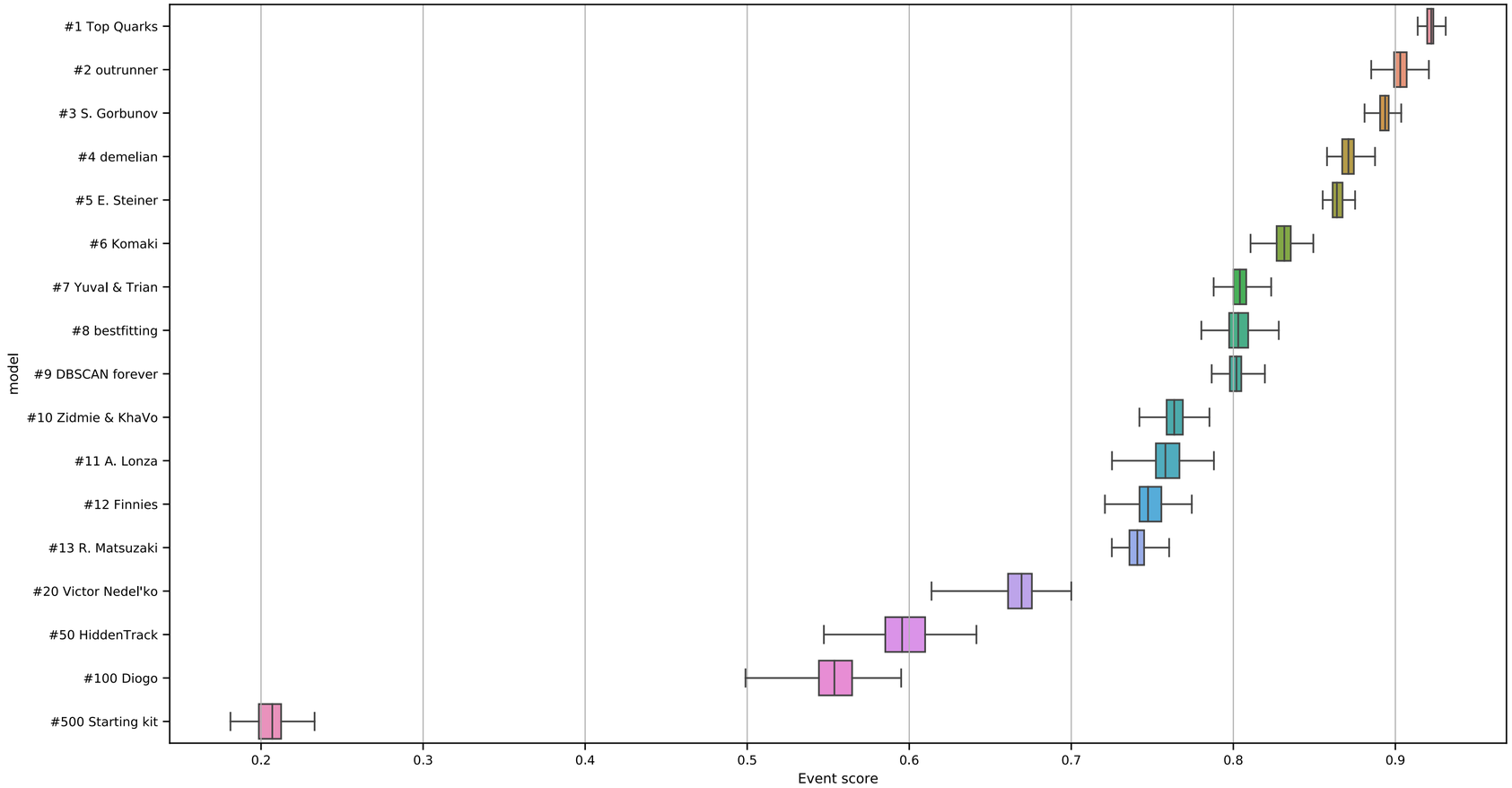
$(\phi, r, z, z/r)$ 5 sets of hit quadruplets



Authors	L. & N. Finnie

Phase 1 Aftermath - Score Stability

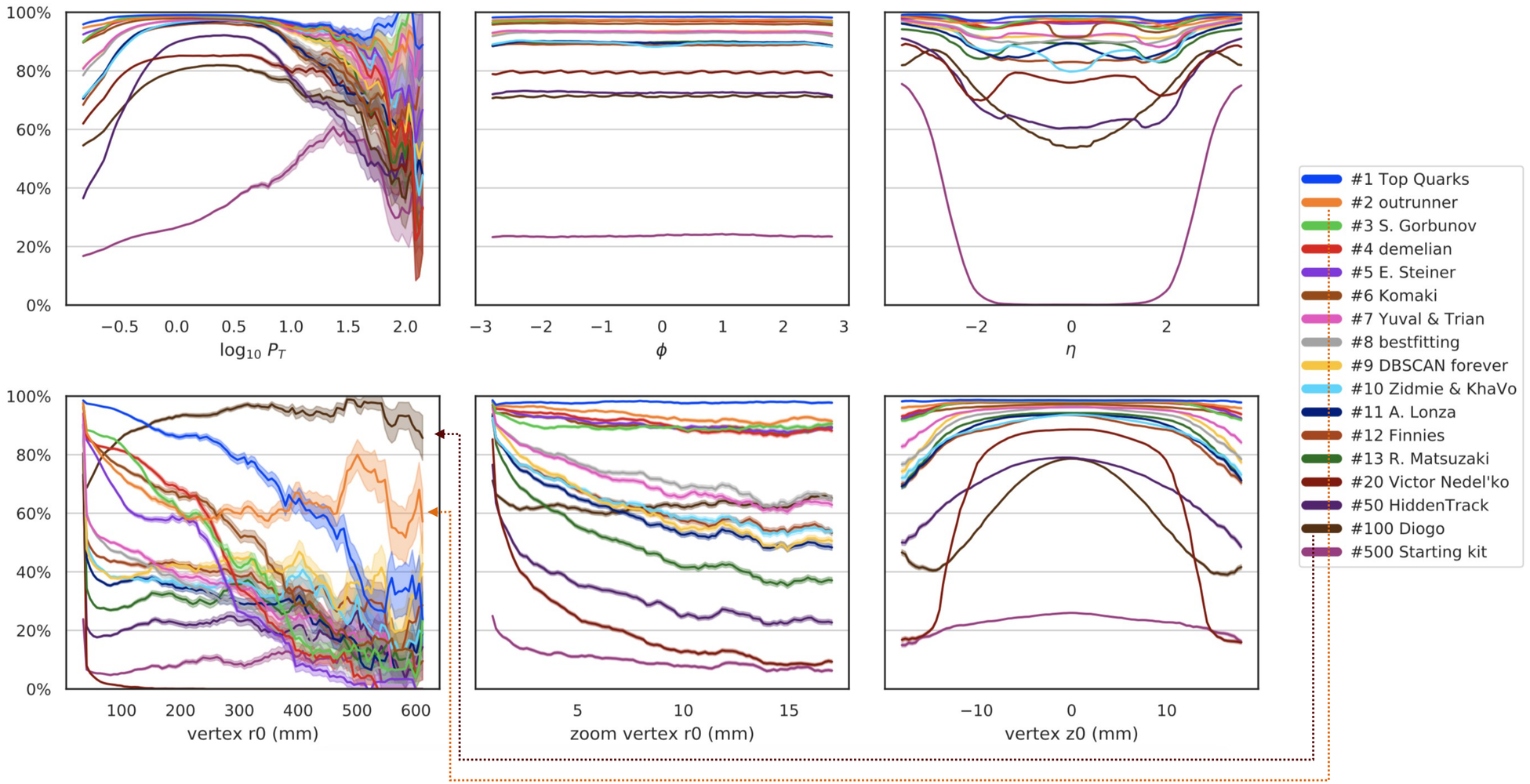
Score is excellent measure of performance.



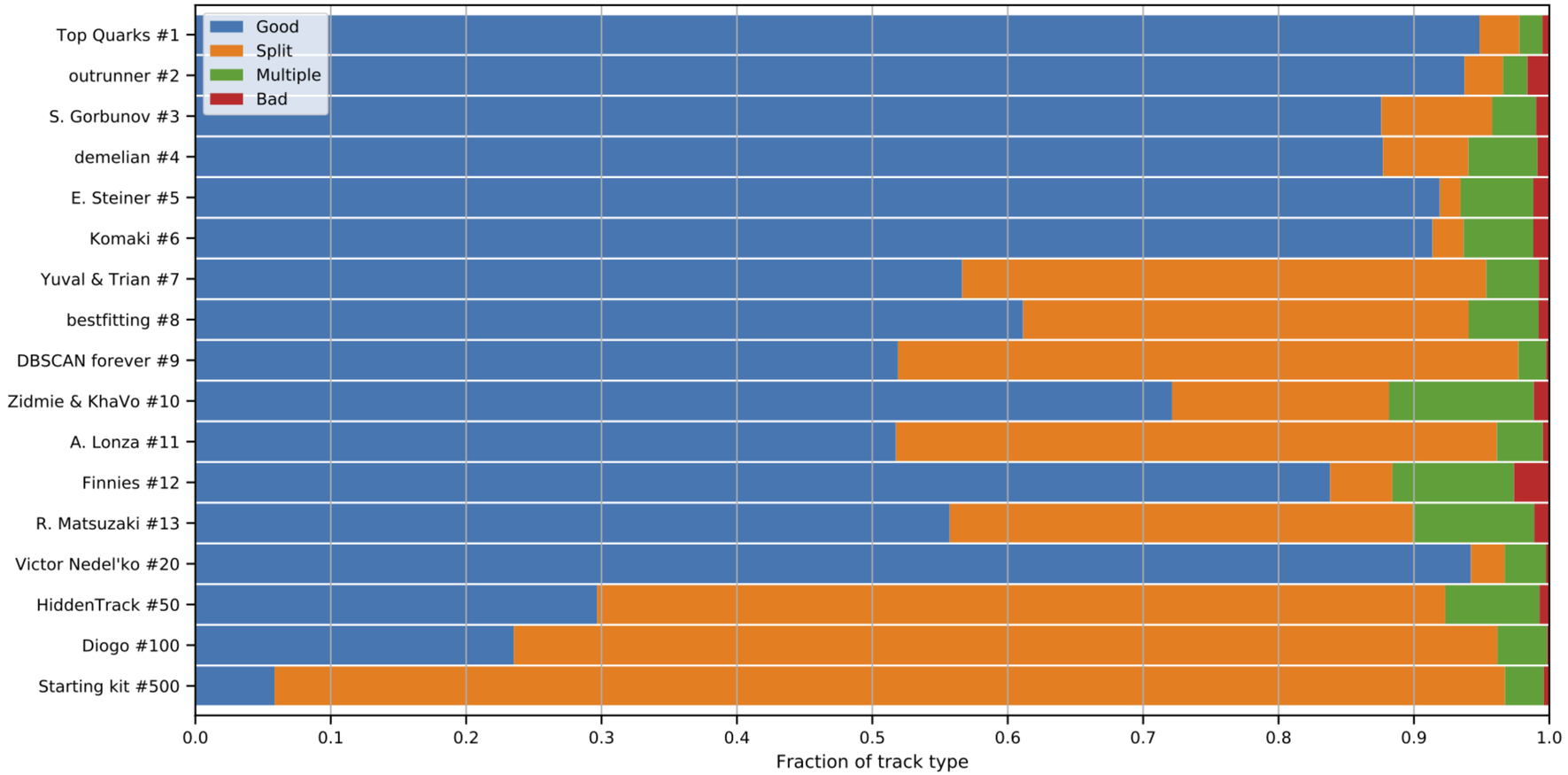
Score quartiles/extrema of submitted solutions.

Phase 1 Aftermath - Tracking Efficiency

High score means High tracking efficiency



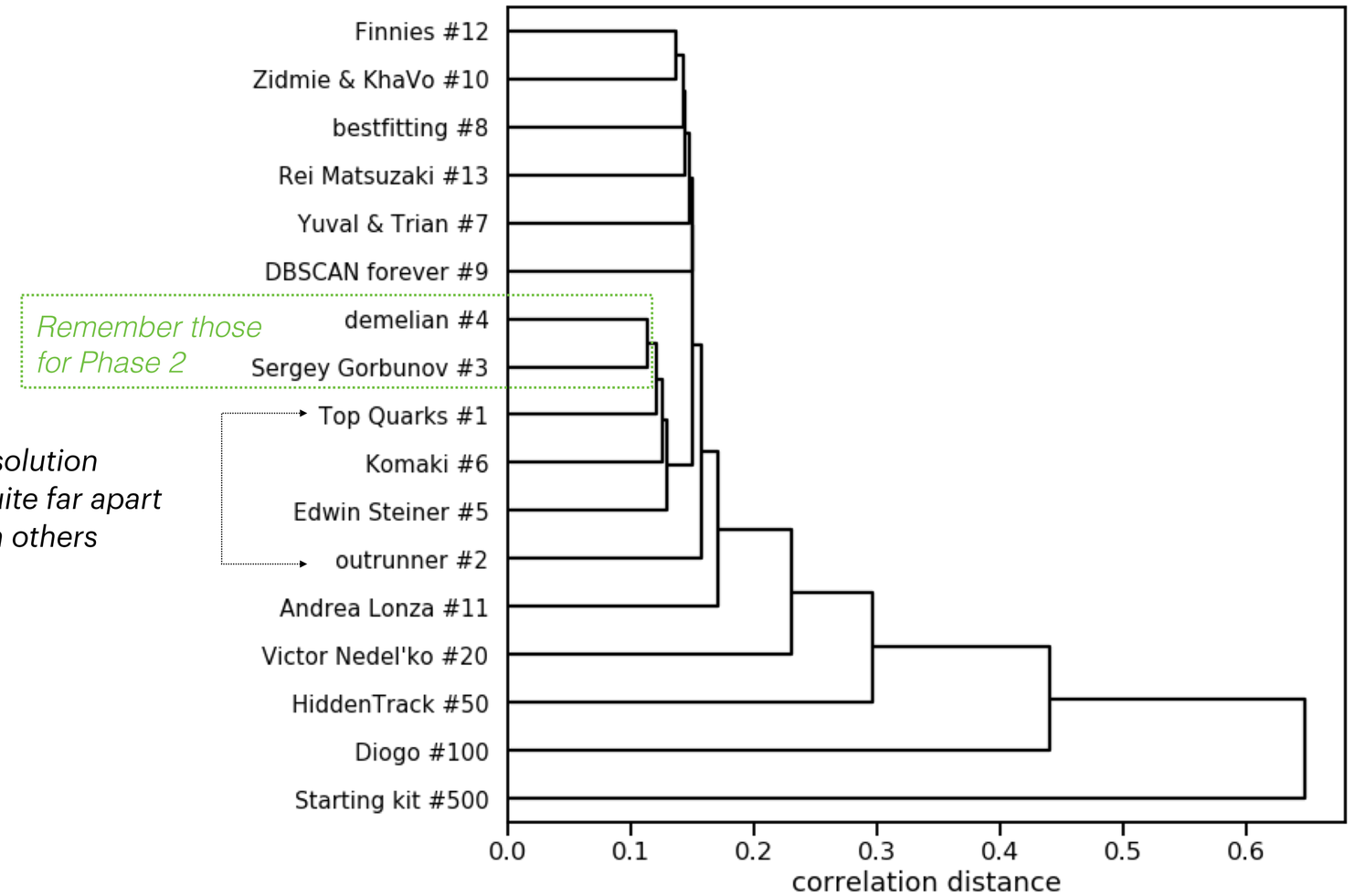
Phase 1 Aftermath - Track classes



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%

Phase 1 Aftermath - Score correlations

High score means High tracking efficiency



Remember those for Phase 2

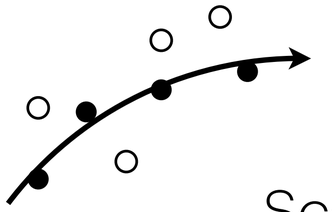
NN solution is quite far apart from others

Phase 2

Accuracy & Speed Challenge

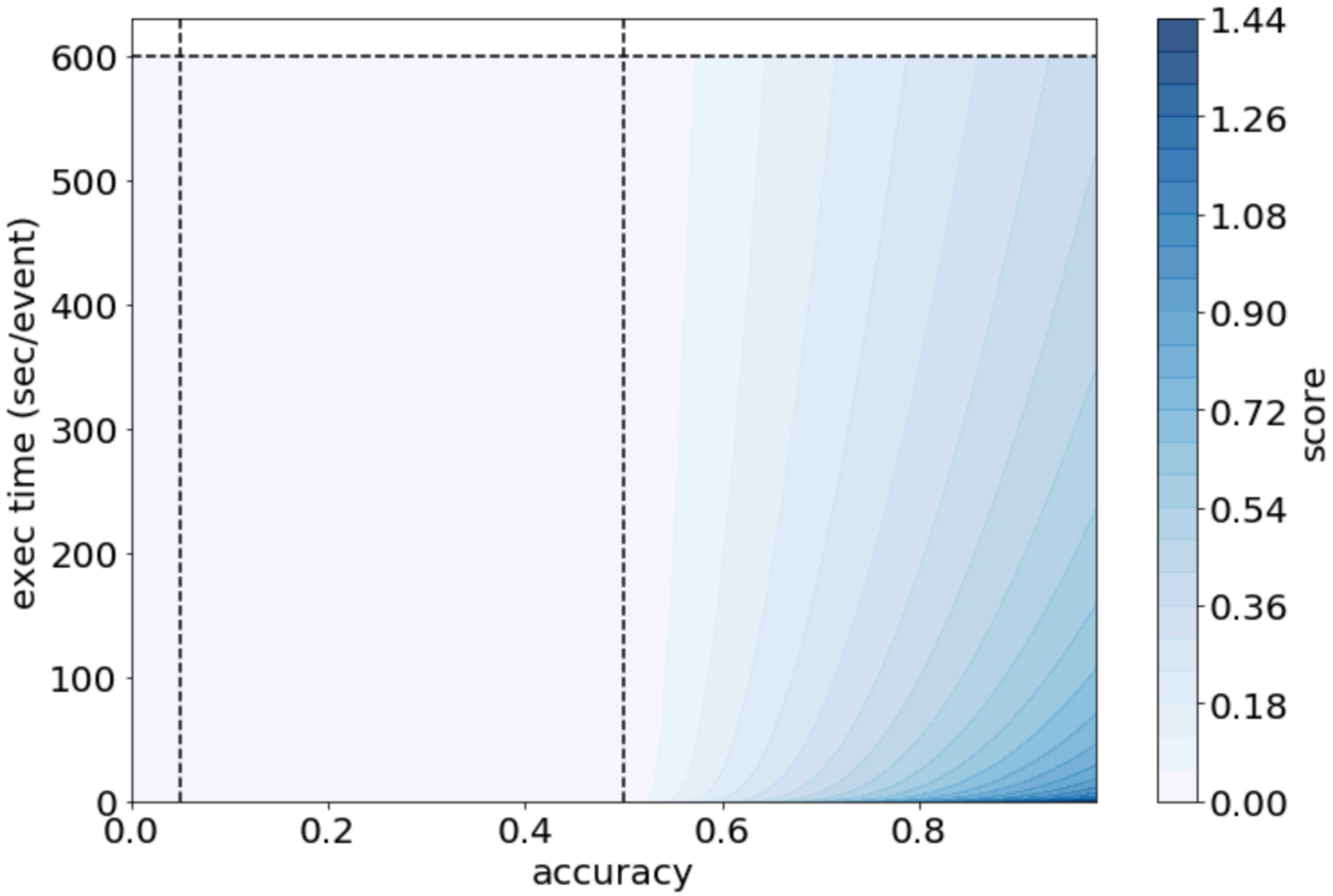
CodaLab

Sep 07, 2018 → Nov 12, 2018



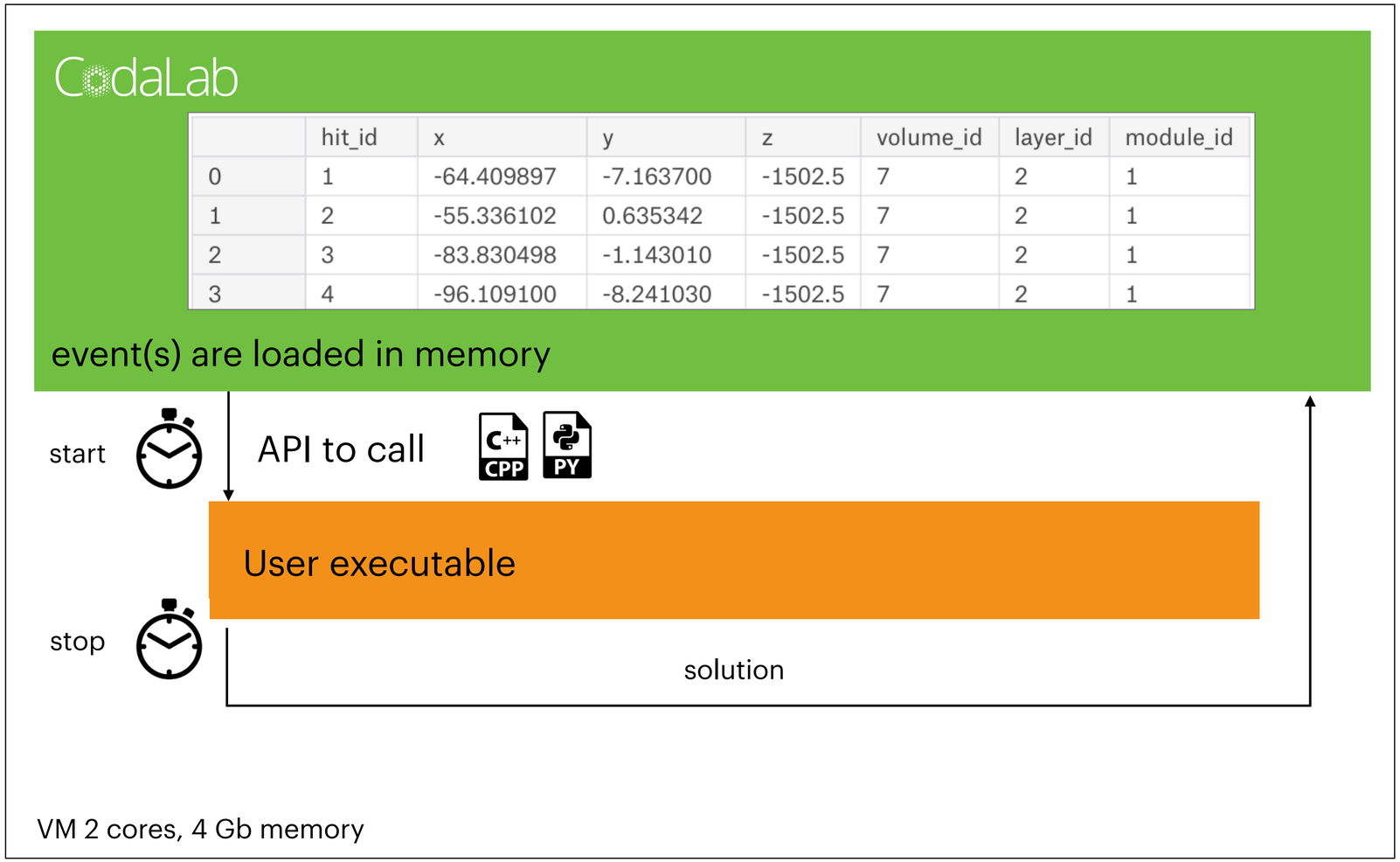
Phase 2 - Modifications

Score built from two components + restricted to primary particles only

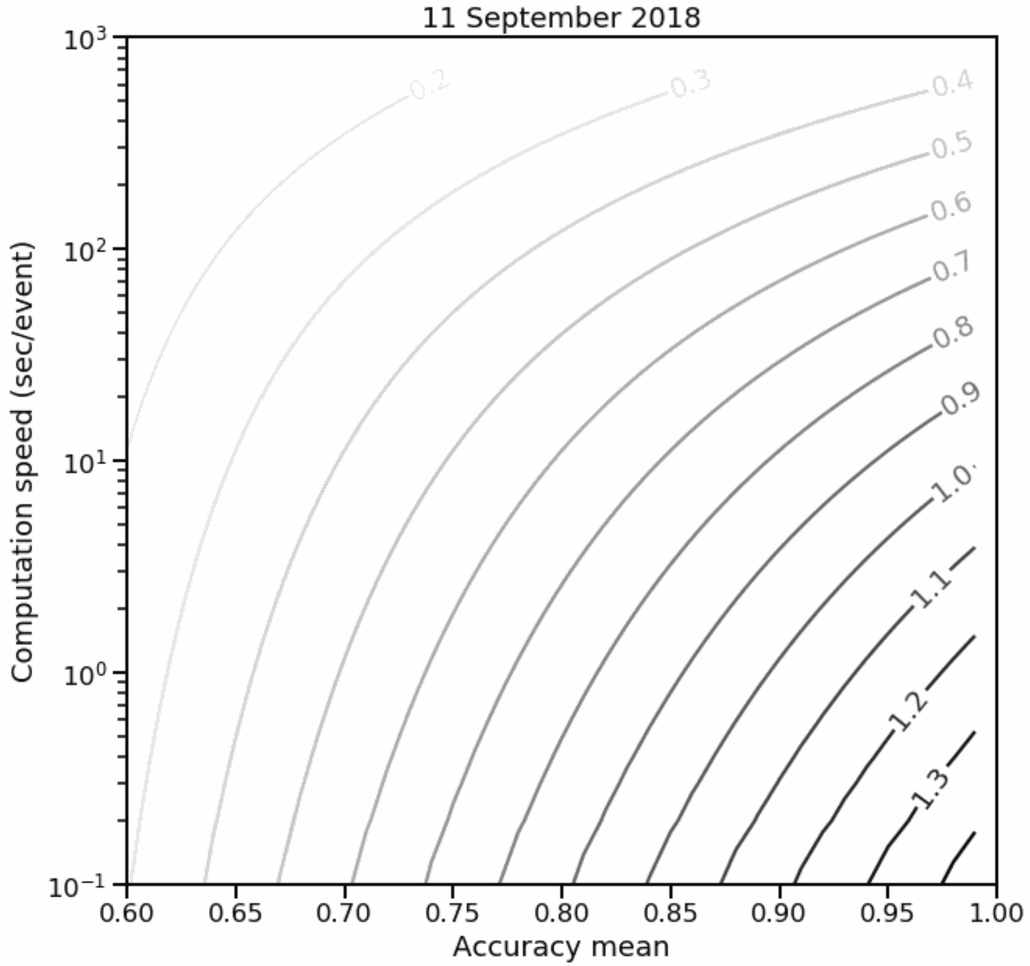


Phase 2 - Modifications

Runtime control



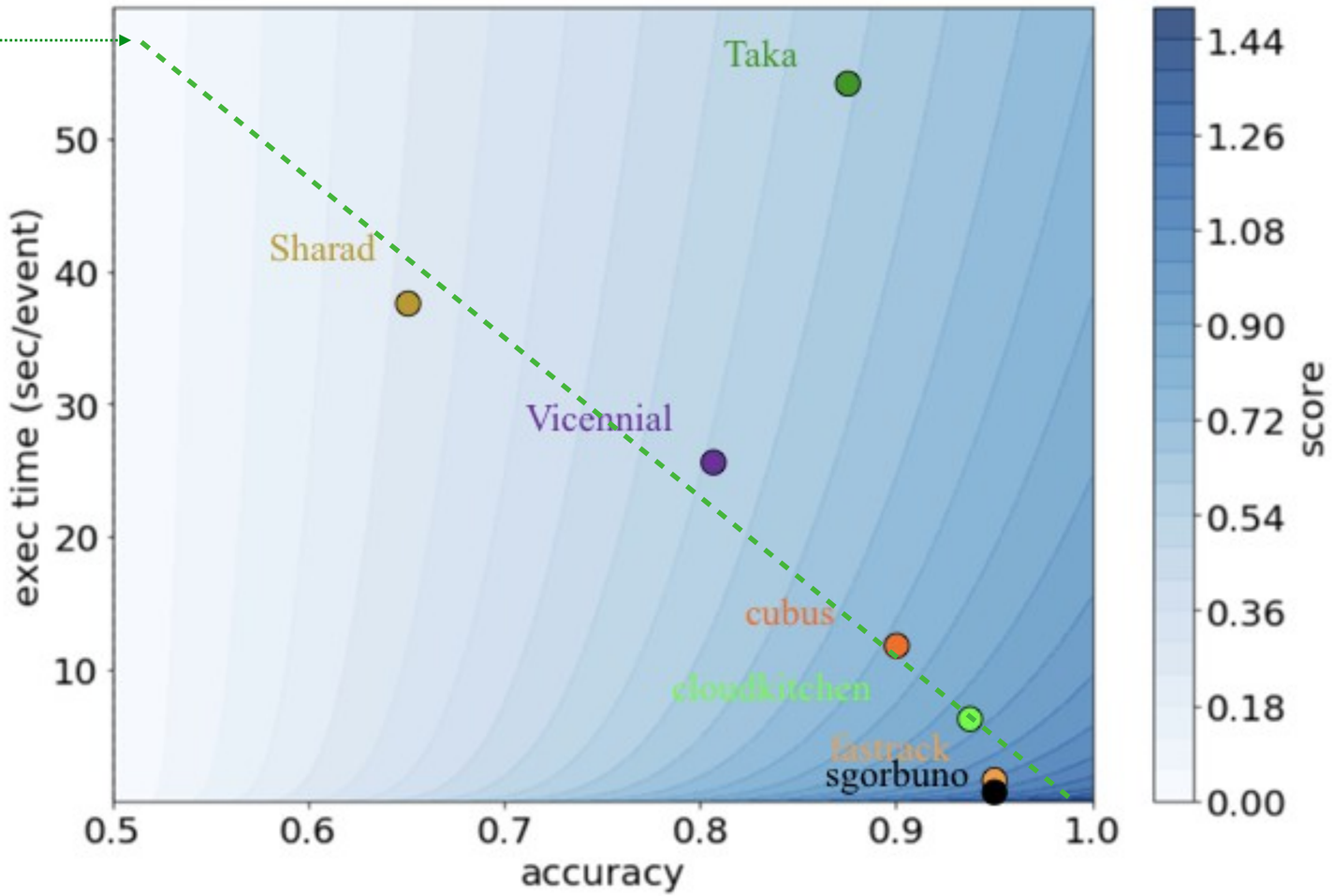
Phase 2 - Score Evolution



Phase 2 - Final Score Map

Correlation accuracy / speed

Fastest solution are
in general also the most
accurate one!



Phase 2 - Final Leaderboard

Only 7 teams made the cut!

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)
				0.0000					

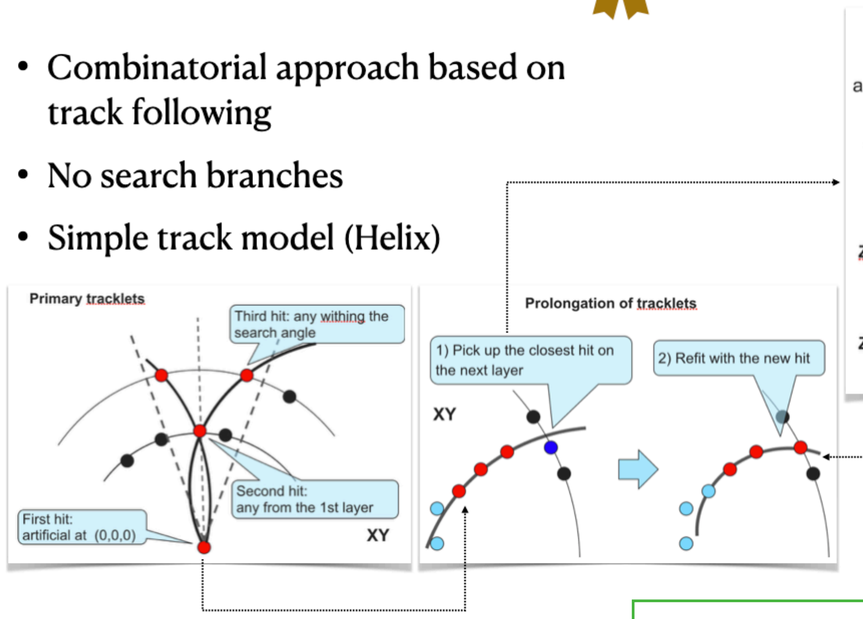
Phase 2 1. Sergey Gorbunov

HEP

- "Mikado" algorithm
  - Based on 3rd solution of Phase 1
- Modifications
  - Runs iteratively in 80 passes with hit removal
  - Search branches are enabled
  - Parameters for each pass optimised (semi-automated):
    O(10k) parameters optimized
- High Accuracy
  - Reaching 0.944 (primary particles only)

Phase 1 3. Sergey Gorbunov

- Combinatorial approach based on track following
- No search branches
- Simple track model (Helix)



Author	[Sergey Gorbunov]
<Wall time>/evt	~1.2 mins

Author	[Sergey Gorbunov]
<Wall time>/evt	0.56s
Peak memory	0.1 Gb

Phase 2 FASTrack

- Iterative algorithm
  - Based on 4th solution of Phase 1

- Algorithm outline
  - using measurement shapes to predict track inclination
  - segment based track following
    - Using connection graph pre-built from Detector.csv
    - Run Cellular Automaton (CA), OpenMP parallelized
    - Candidate building: graph traversal with applied KF
    - Combinatorial kalman filtering

- High Accuracy
  - Reaching 0.944 (primary particles only)

Iteratively done 3 times
with hit removal

Author	[Sergey Gorbunov]
<Wall time>/evt	1.1 s
Peak memory	0.6 Gb

+ OpenMP

Phase 2 3. Cloud kitchen

- Partly reused concept
  - Based on winning solution of Phase 1
  - Achieved accuracy: 0.93

Phase 1 1. Top Quarks

Main steps

- Select promising pairs
  - 7 million / 0.99
- Extend pairs to triplets
  - 12 million / 0.97
- Extend triples to tracks
  - 12 million / 0.95
- Add duplicate hits to tracks
  - 12 million / 0.96
- Assign hits to tracks
  - 90% of hits / 0.92

- Logistic regression for track candidate pruning
- Pure C++, some scikit-learn for training

Findings

- No magic formula
- Because we were fast to try
- Implement many ideas and got
- 0.03 (0.85→0.88) from
- parameter
- combination of many
- factors

Triplet finder

NN3

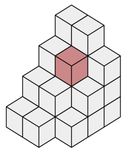
Doublet finder

NN1 NN2

Author	Johan S. Wind
<Wall time>/evt	7m 17s
Peak memory	2.78 Gb

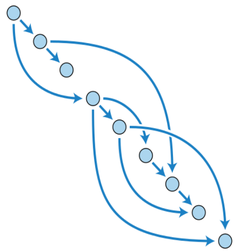
hits

sorted in voxels



organised in direct acyclic graphs (DAG)

pre-trained on ~25 events ground truth



DAGs are used to fast navigate through voxel space

$\pm z$ graph set

$\eta - \phi$ graph set

Threaded

Author	[Marcel Kunze]
<Wall time>/evt	~7 s
Peak memory	0.79 Gb

Key Lessons

Phase 1

- Huge interest in/outside community
  - Reputation of **kaggle** platform helped
- Worth investing in a good score
  - Solution ranking reflects physics performance
- HEP code is/remains competitive
  - Alternative solutions become powerful
- Domain knowledge does help

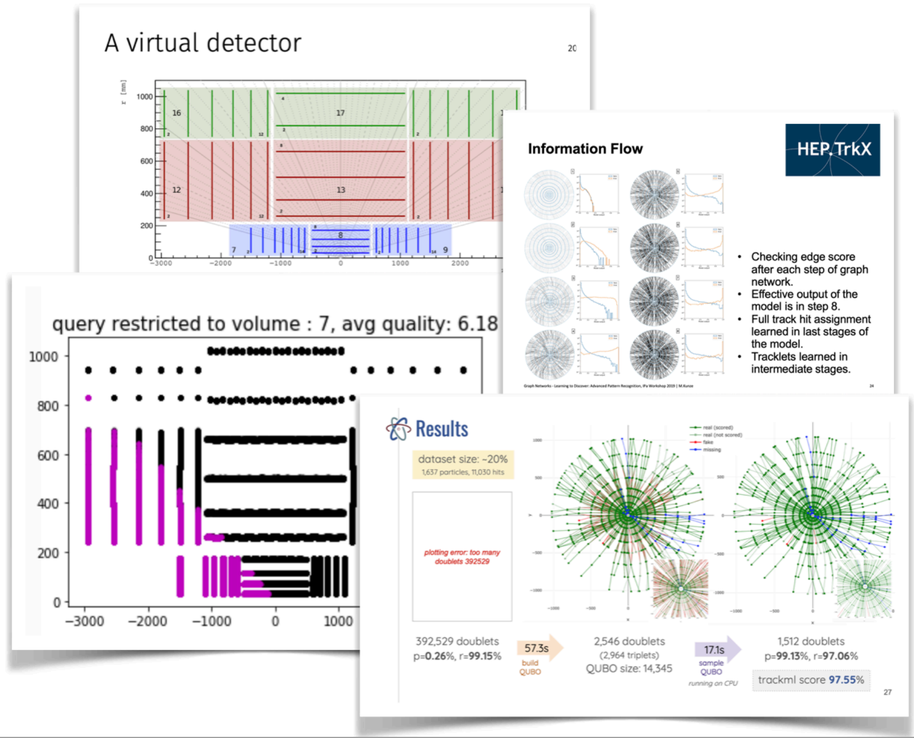
Phase 2

- Entry barrier was (too) high
  - Predefined API makes it more restrict
  - Less known/prominent **coda lab** platform
- HEP code is currently the best option
  - Can still be improved (is ongoing)
- Alternative solutions lack ~1 order of magnitude in CPU behind
  - Worth watching, and certainly not everything was tried

Spin-offs & Remains

Dataset used in many R&D projects

From this workshop only



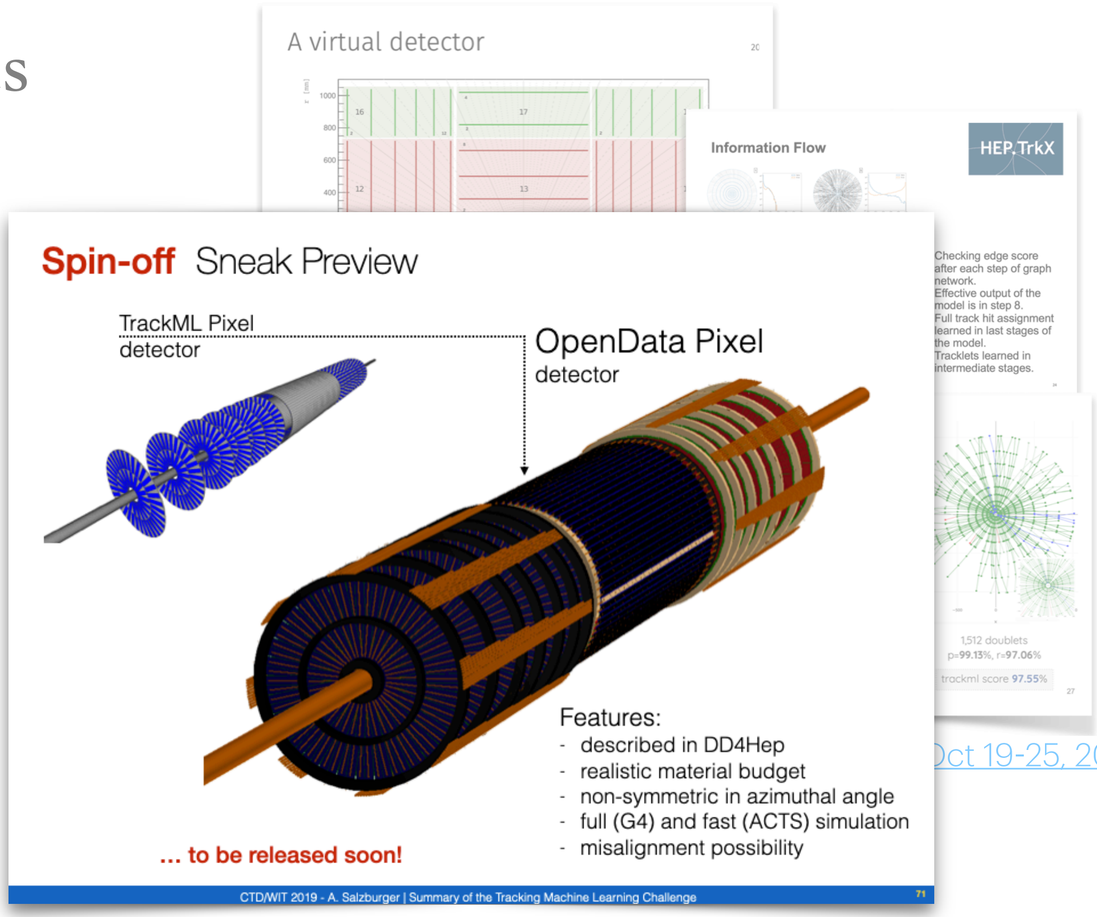
Slide from [IPA, Oct 19-25, 2019]

Spin-offs & Remains

Dataset used in many R&D projects

Detector being evolved into an Open Data Detector project

From this workshop only



Oct 19-25, 2019]

More information: [CHEP2018] [CTD2019]

Spin-offs & Remains

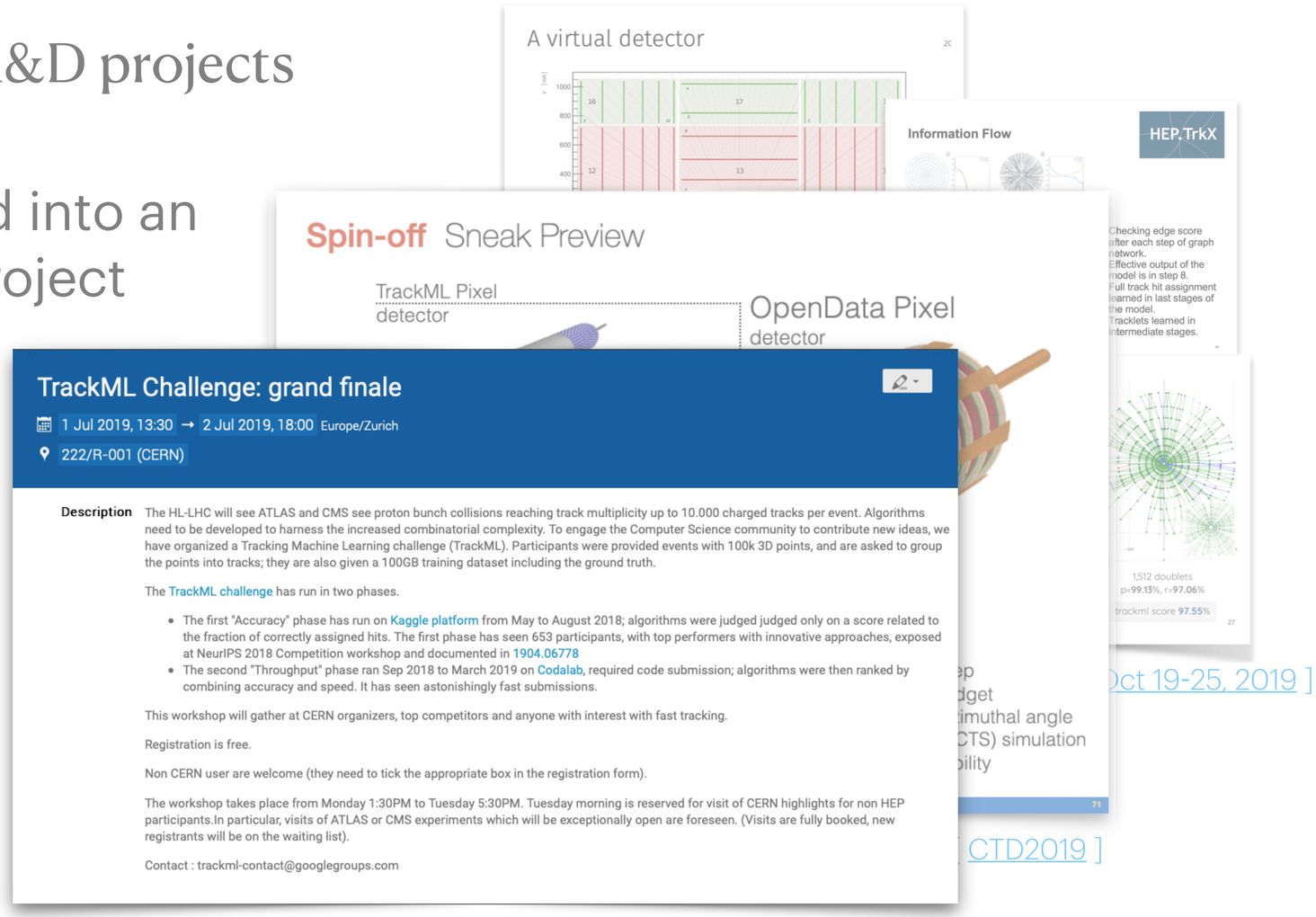
Dataset used in many R&D projects

Detector being evolved into an Open Data Detector project

Community

- Competition track at [NeurIPS2018]
- [TrackML Grand Finale], CERN, 2019
- [IPA Workshop], Paris, 2019
- Conference talks, seminars, etc.

From this workshop only

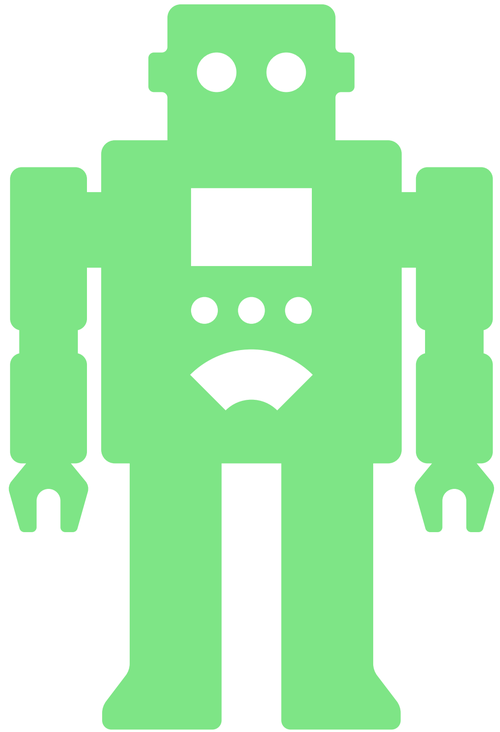


Conclusions

- TrackML project started with an idea in 2015
  - Based on experience of the HiggsML challenge
  - Took some 2-3 years in making, performed through 2018
  - Ran very successfully

- Huge resonance inside/outside the field
  - Connected DS/ML and HEP Tracking community
  - Would we have liked to see more ML ? Probably.

- Spin-offs will remain
  - Dataset is still at great use in R&D projects
  - Open Data Detector to come for refined development/testing



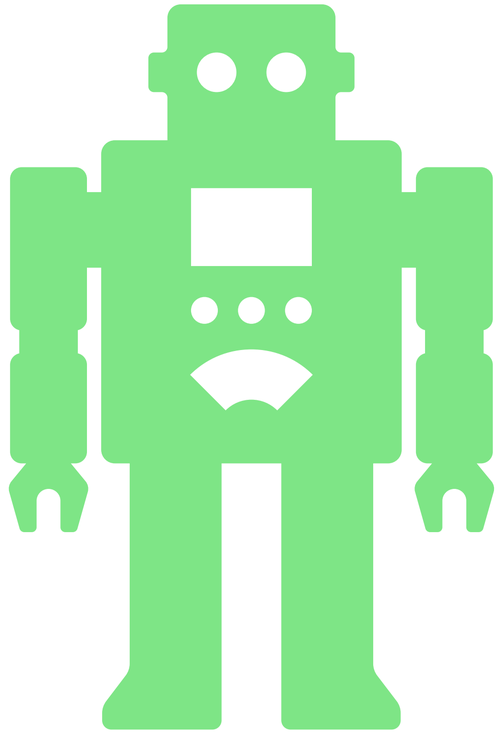
trackml.contact@gmail.com

https://sites.google.com/site/trackmlparticle/

@trackmllhc

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

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



Backup Material

Phase 1 Organisers Prize

- Algorithm outline

  - First step is a route data bank building
    Geometry identifier (module, layer, volume)
    used to pre-build route patterns,
    route is a sequence of modules

    assuming training set contains all possible patterns

  - Second step is hit matching
    searching through all possible routes and check
    if you have hits on each module
    this defines a track candidate