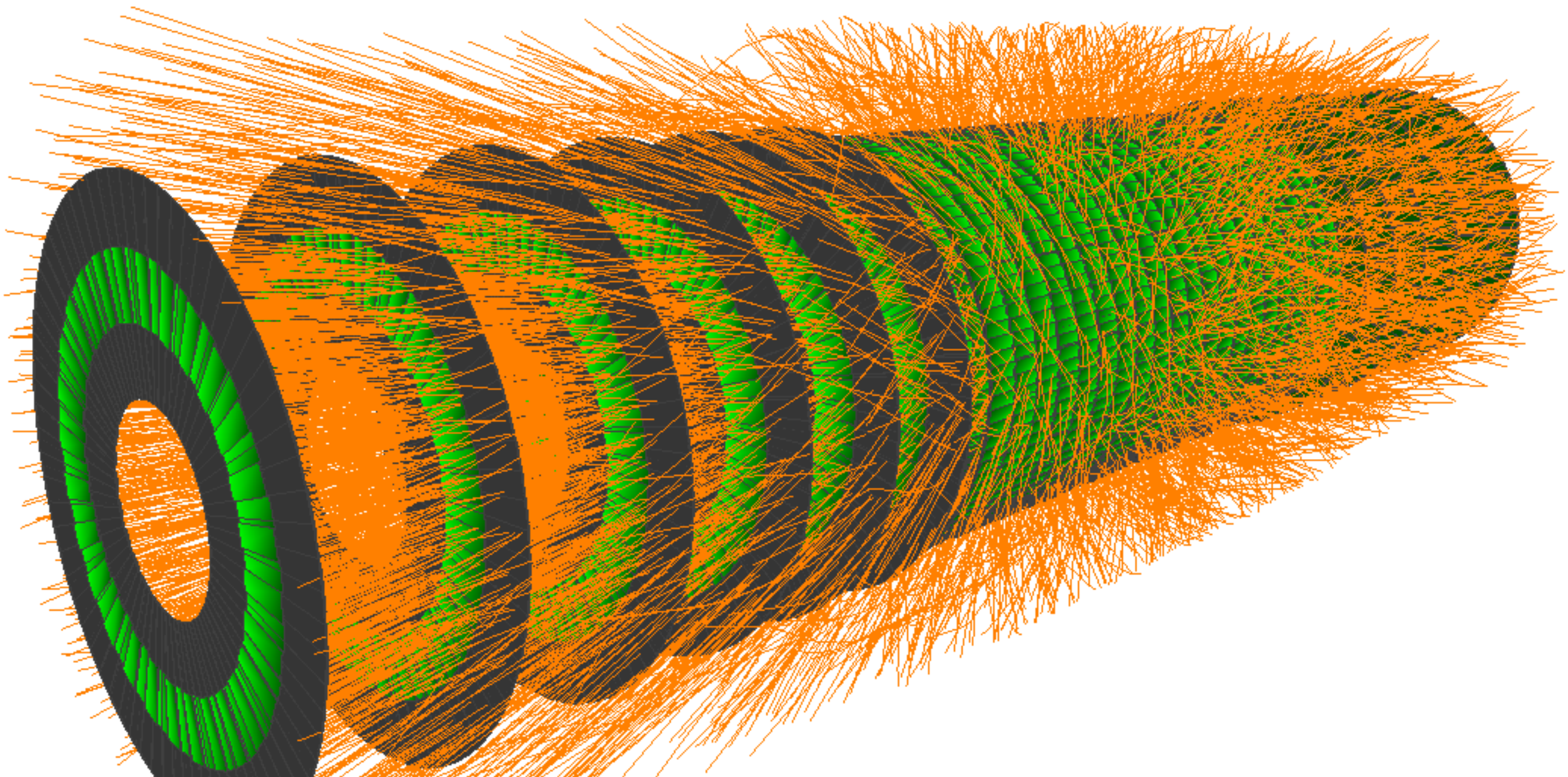


Tracking Machine Learning Challenge - a half way through review

@trackmllhc

A. Salzburger (CERN)
@SaltyBurger

TrackML Who and How

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), Cécile Germain, Victor Estrade(LRI-Orsay),Edward Moyse (University of Massachussets), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex, HSE)

Partners:

kaggle

CodaLab

Sponsors:

NVIDIA

UNIVERSITÉ DE GENÈVE

CHALEARN

Caltech

Fermilab

CERN openlab

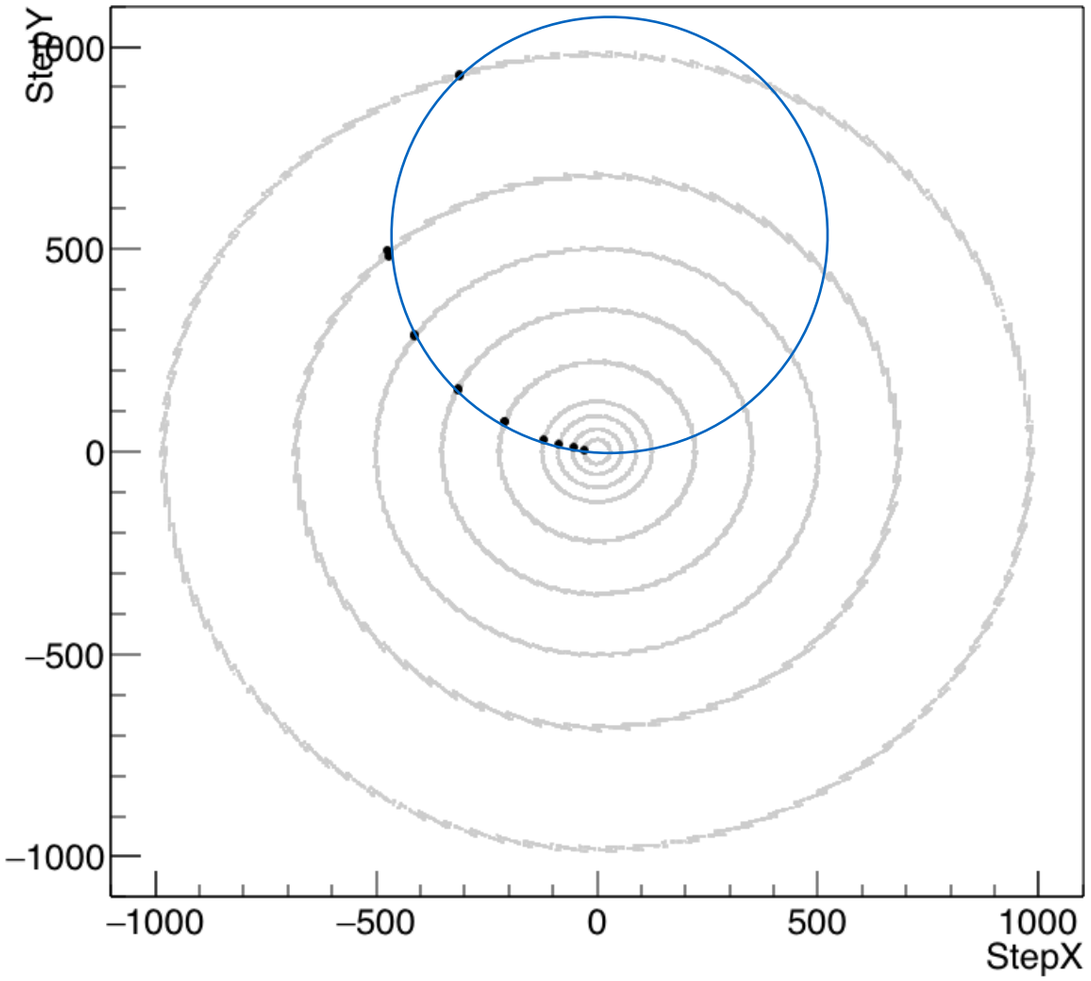
Paris-Saclay Center for Data Science

Inria INVENTORS FOR THE DIGITAL WORLD

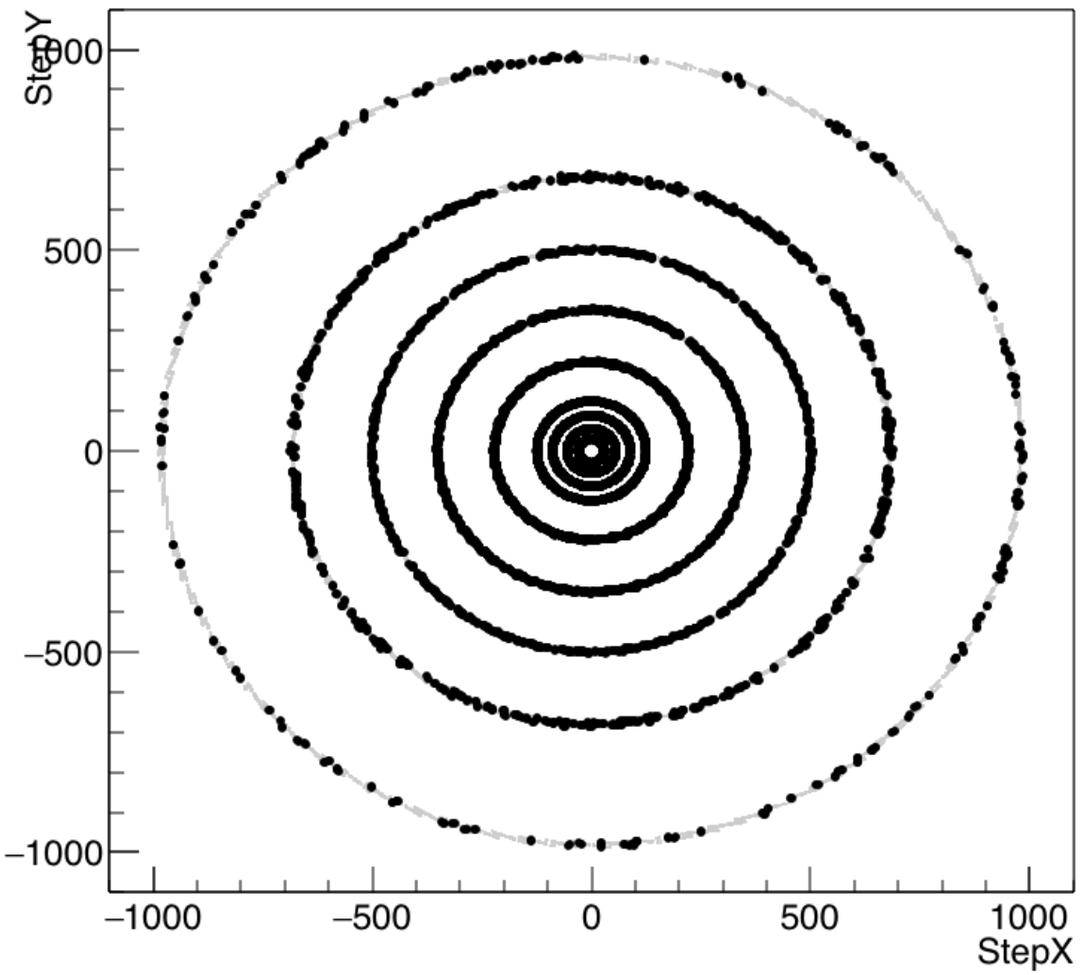
COMMON GROUND

PyTorch

Introduction Charged particles

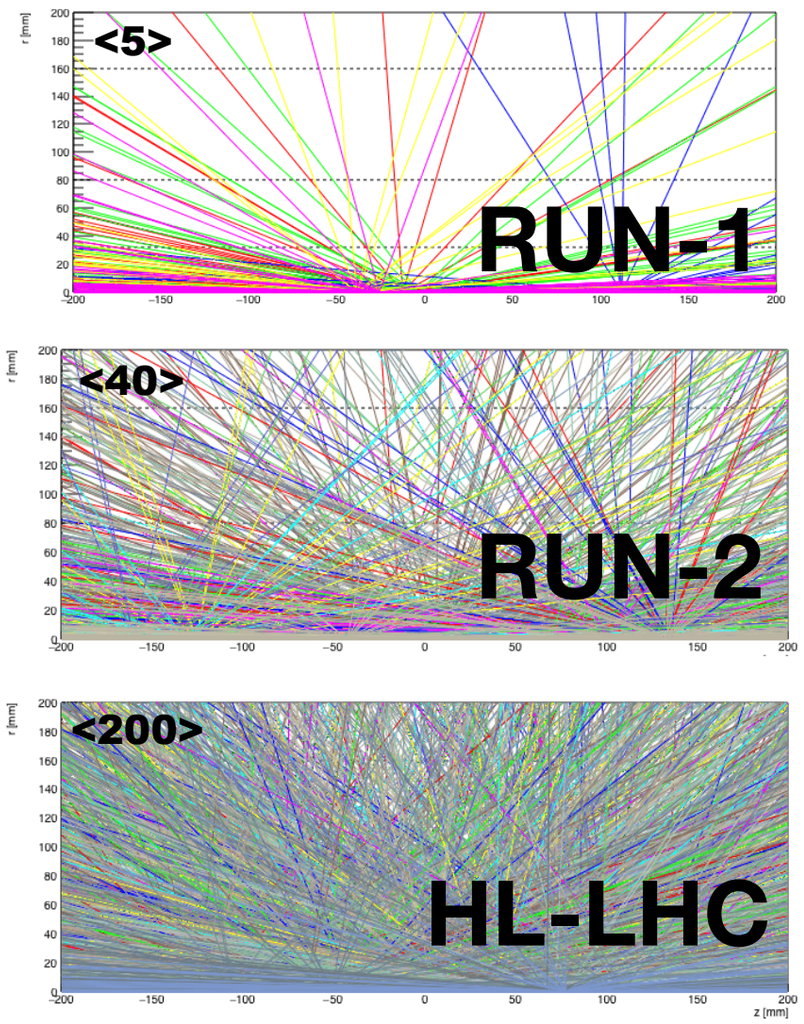


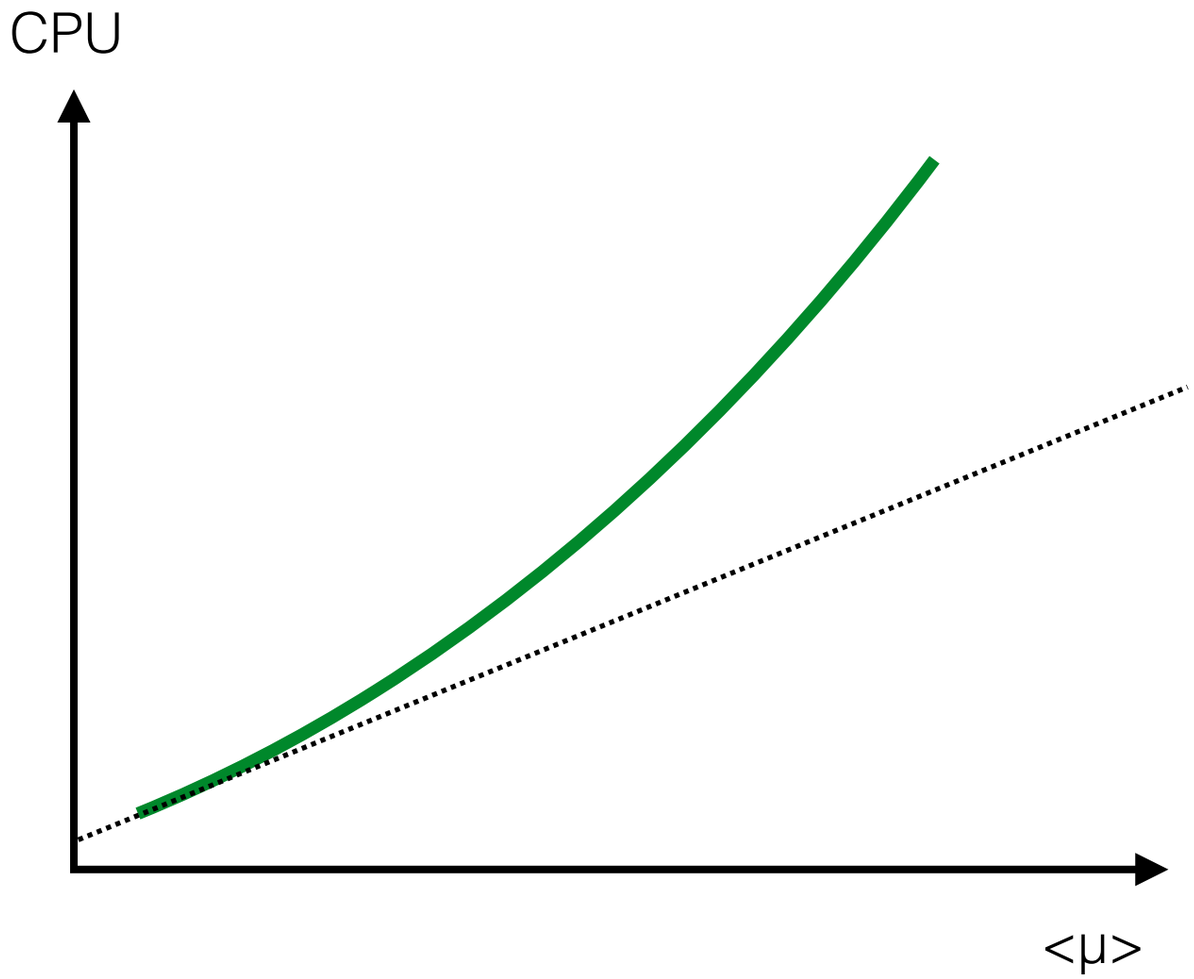
hits from 1 particle



fraction of hits
from particles
in 200 pile-up events

Tracking at LHC, HL-LHC and FCC-hh

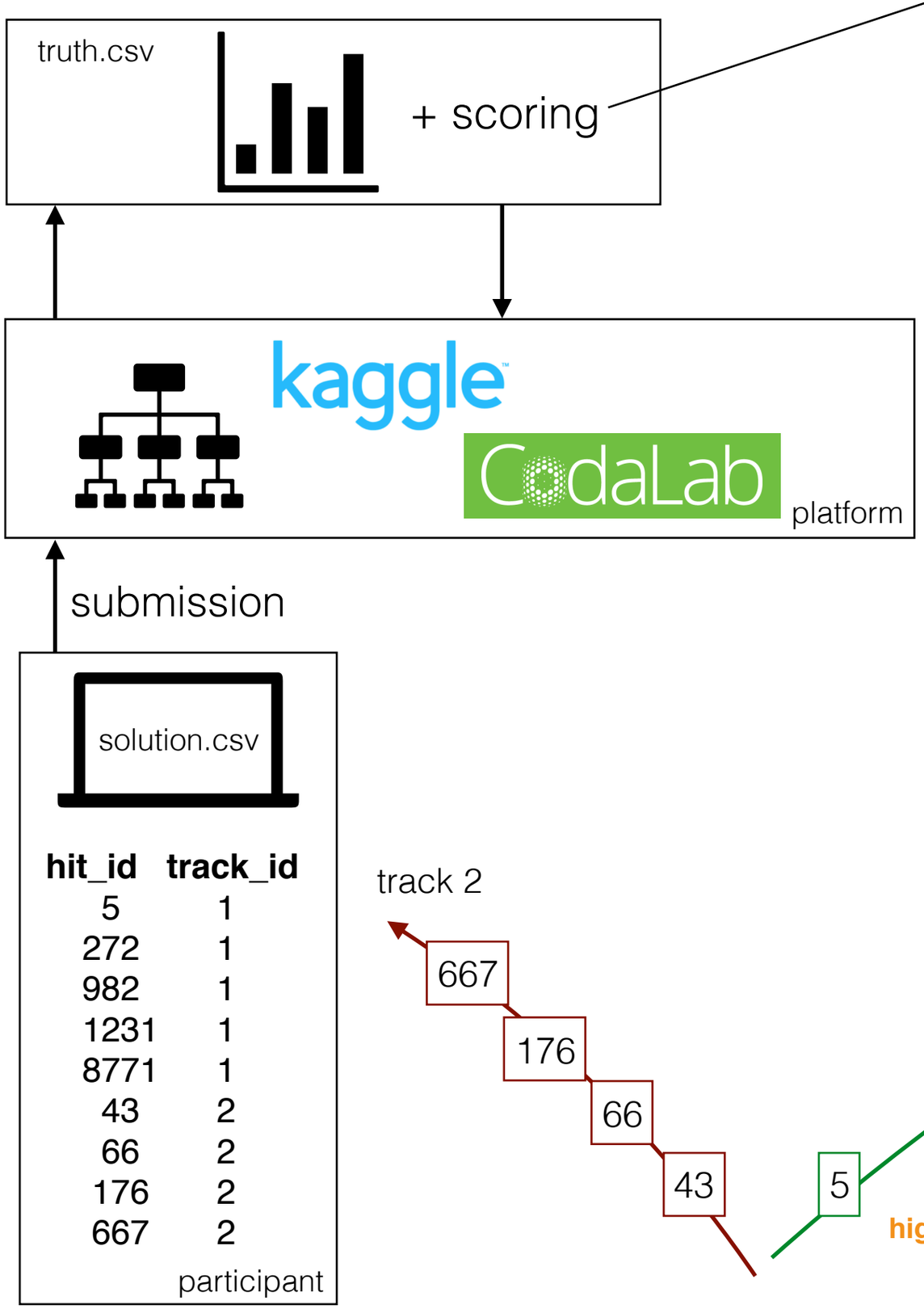




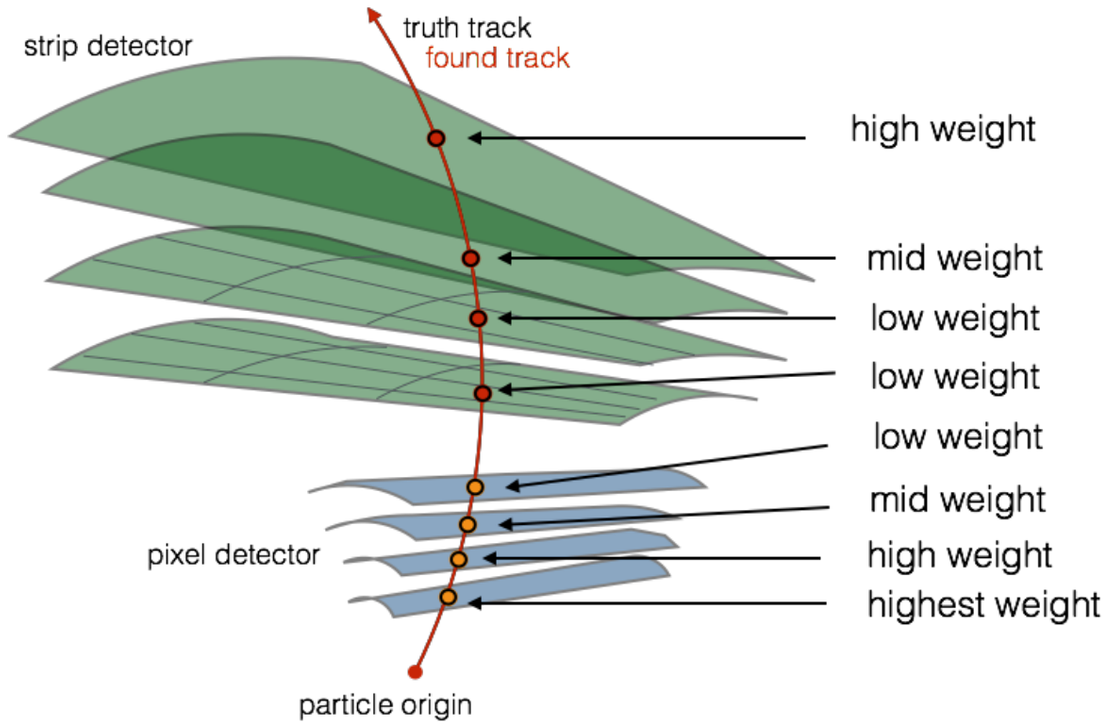
Pattern recognition is due to combinatorial behavior main CPU driver

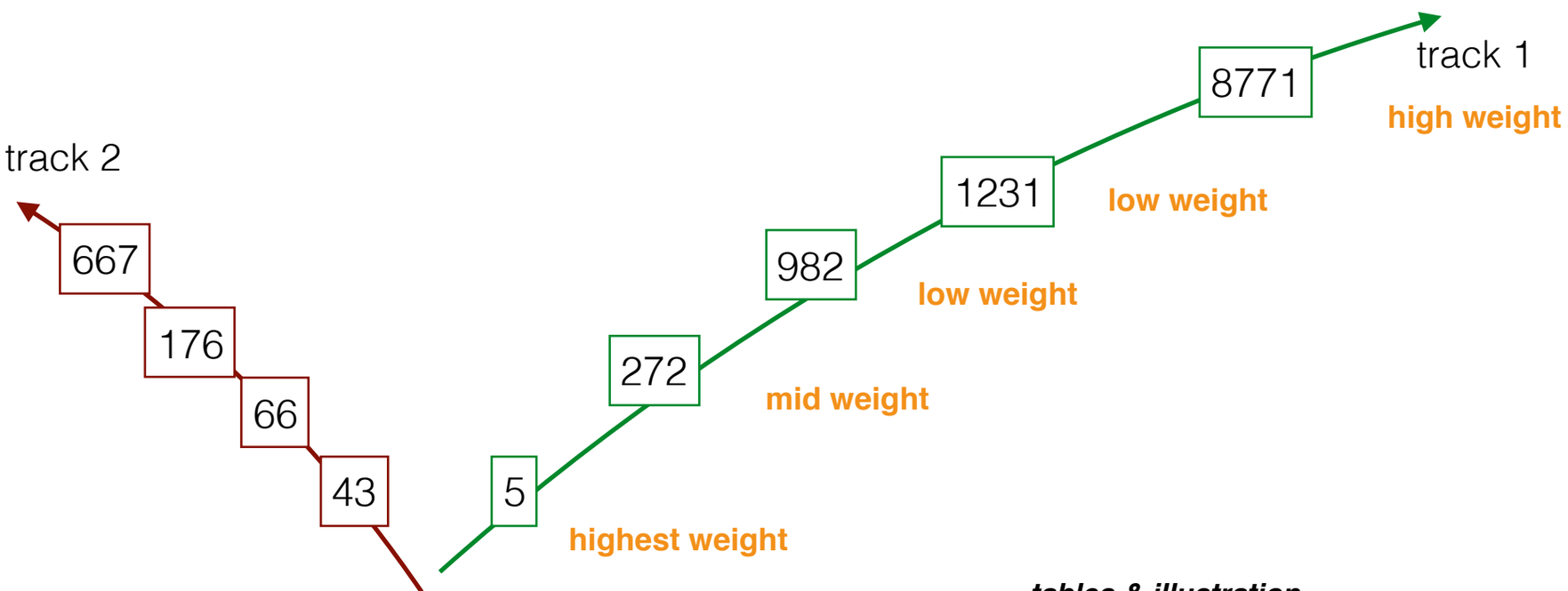
- many improvements done, but still missing a factor 5-10 from trigger rate increase

Submission



hits on track have **weights**

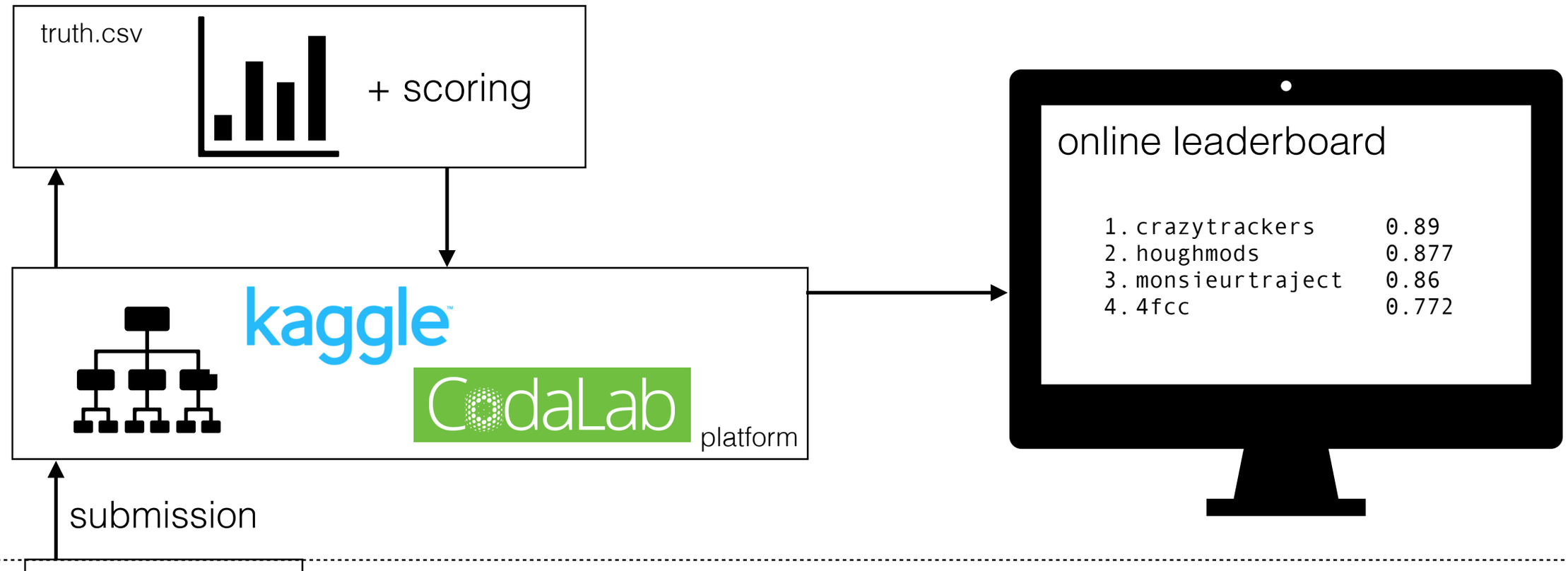




tables & illustration

(top) csv file format for validation hit dataset

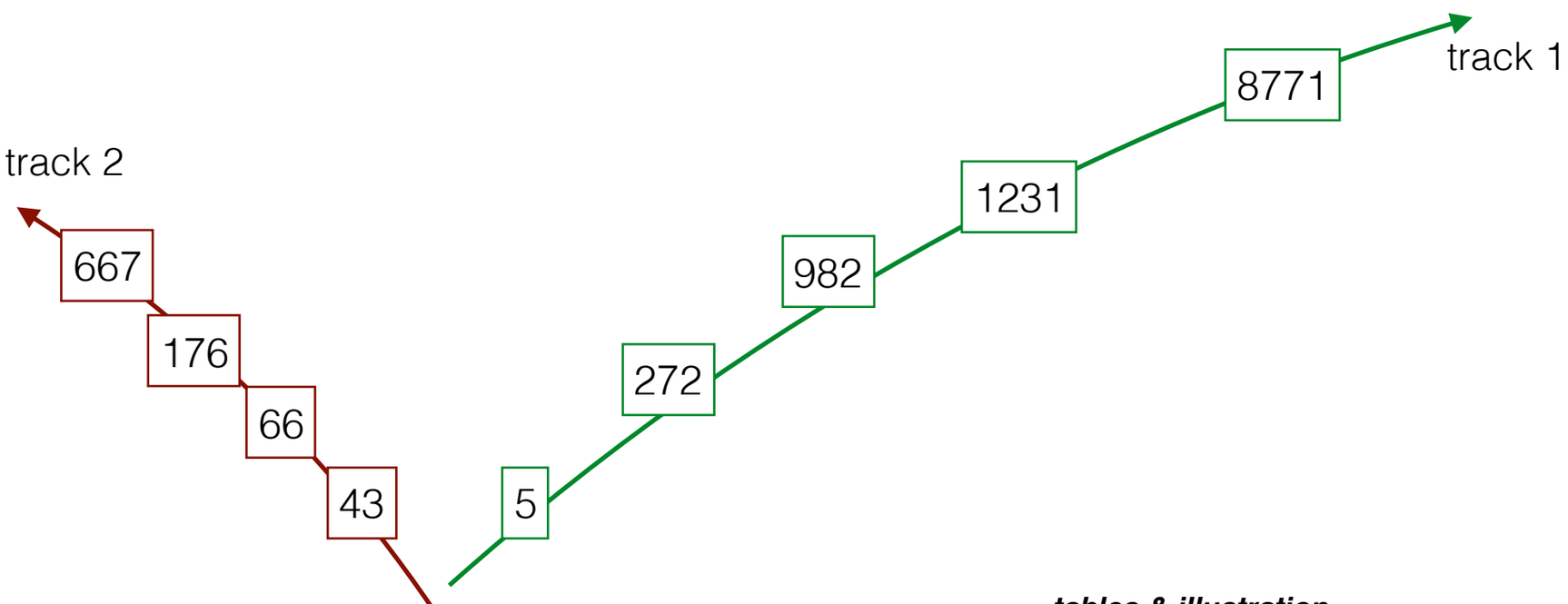
Submission & scoring



solution.csv

hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

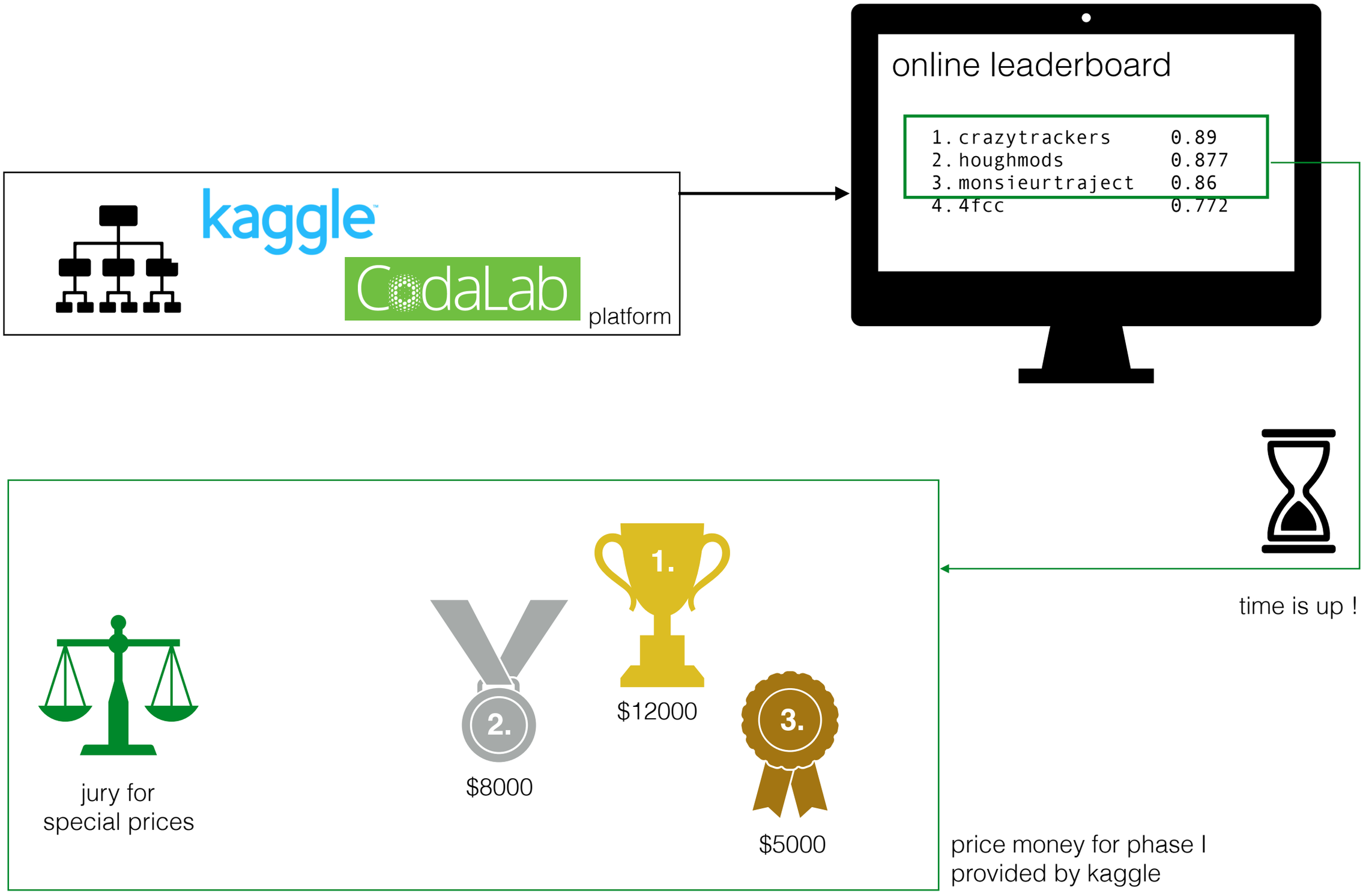
participant



tables & illustration

(top) csv file format for validation hit dataset

Winning


kaggle
CodaLab
platform
online leaderboard
1. crazytrackers 0.89
2. houghmods 0.877
3. monsieurtraject 0.86
4. 4fcc 0.772
time is up !
jury for
special prices
1.
2.
3.
$12000
$8000
$5000
price money for phase I
provided by kaggle

The challenge in 2 phases

phase 1: **accuracy phase**

phase 2: **throughput phase**

▶ Current	Next	End
Development	Final	Competition Ends
Sept. 7, 2018, midnight UTC	Nov. 5, 2018, 11:59 p.m. UTC	Nov. 12, 2018, 11:59 p.m. UTC

Phase 1 Accuracy

kaggle™

Phase 1 Winners

Public Leaderboard | **Private Leaderboard**

The private leaderboard is calculated with approximately 71% of the test data.

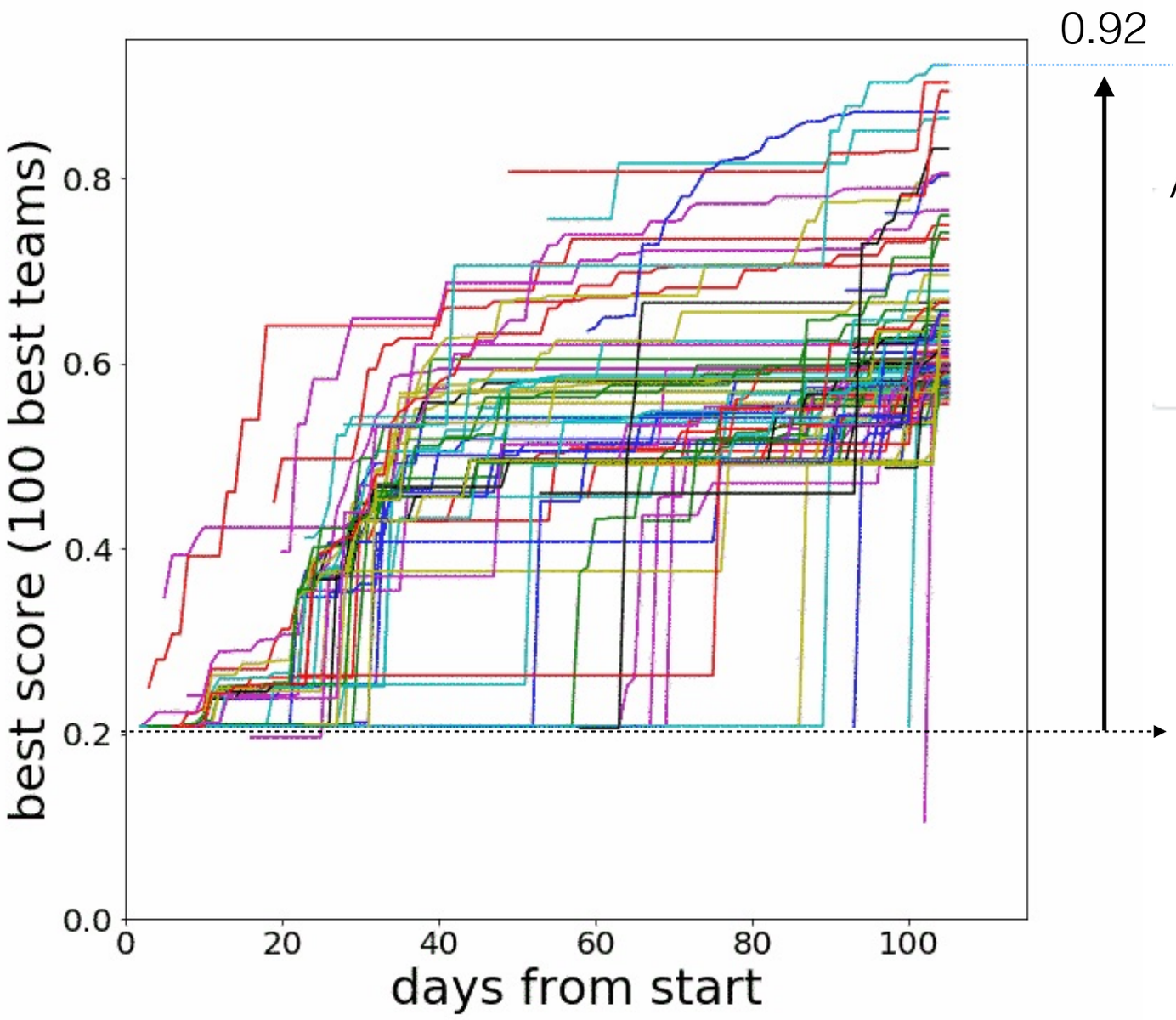
This competition has completed. This leaderboard reflects the final standings.

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 Evolution of score over time



At final date:

656	776	5,837
Teams	Competitors	Entries

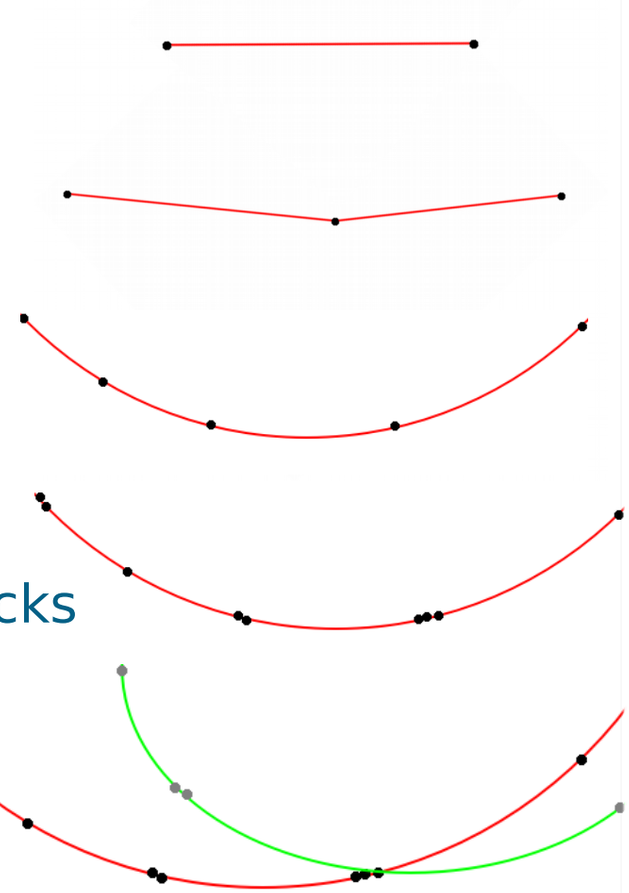
score of the the starter kit (DBScan)

Phase 1 Top Quarks

	Wall clock time	Peak memory usage
Average	7m17s	2.78GB
Max	11m20s	4.07GB

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

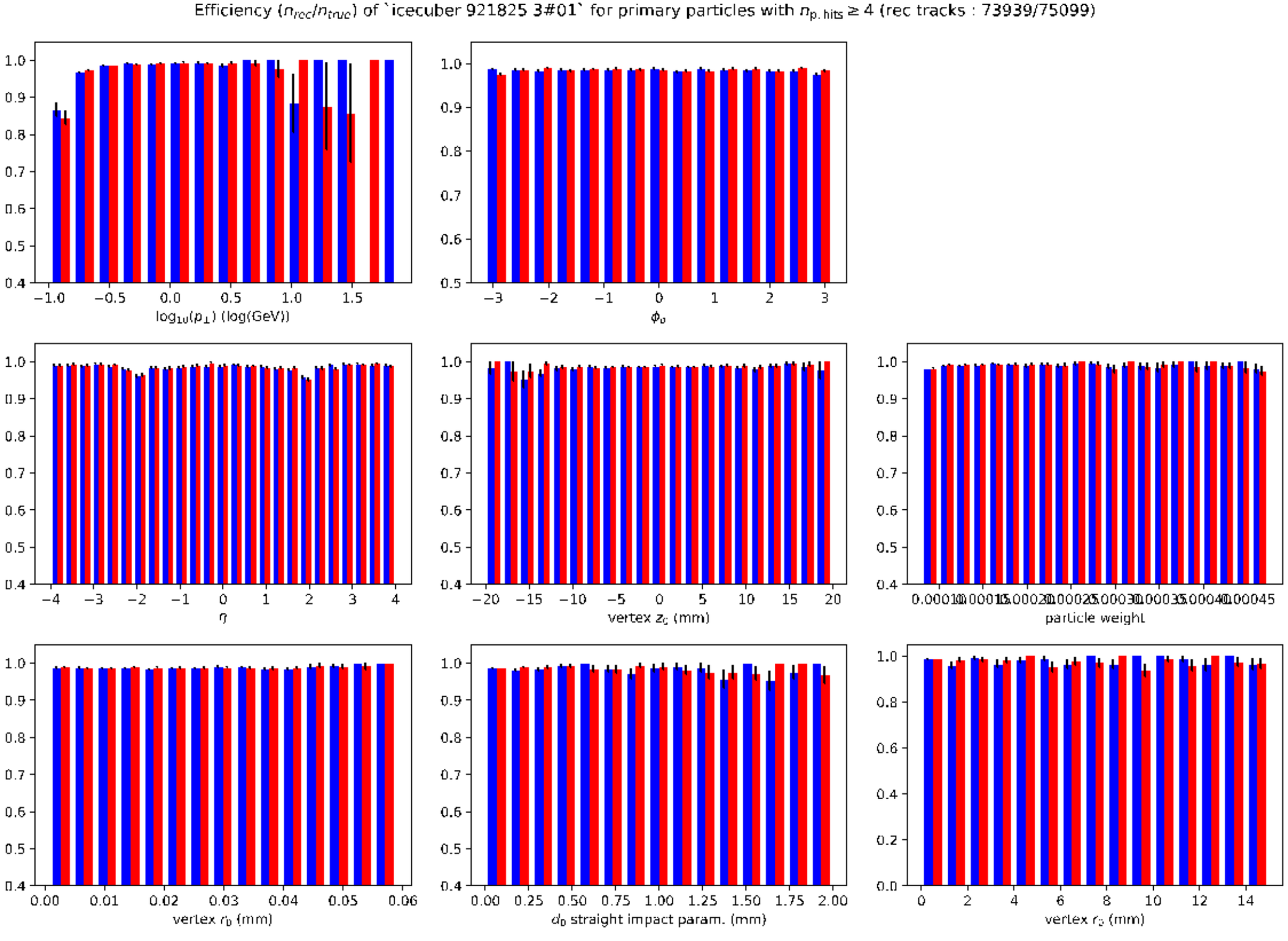


Findings

- 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

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

Phase 1 Top Quarks



Phase 1 outrunner

"Wall clock time"
~1 day/event

Pure ML approach using python & Keras

- Event with **N** hits
- predict **N x N** relationships between hits, connect pairs when their probability is 1 (rather than 0)

Training:

- **5** hidden layers with **4k - 2k - 2k - 2k - 1k**
- **27** input variables per pair:
  x, y , z, counts, sum(cells.value) per hit
  two unit vectors per hit for direction from cell information
  4 parameters for linear (z_0) and helical compatibility

Prediction:

- predict relationship probability

Reconstruct

- starting from one hit, find highest probability pair, then add pairwise hits
- test new hit for compatibility

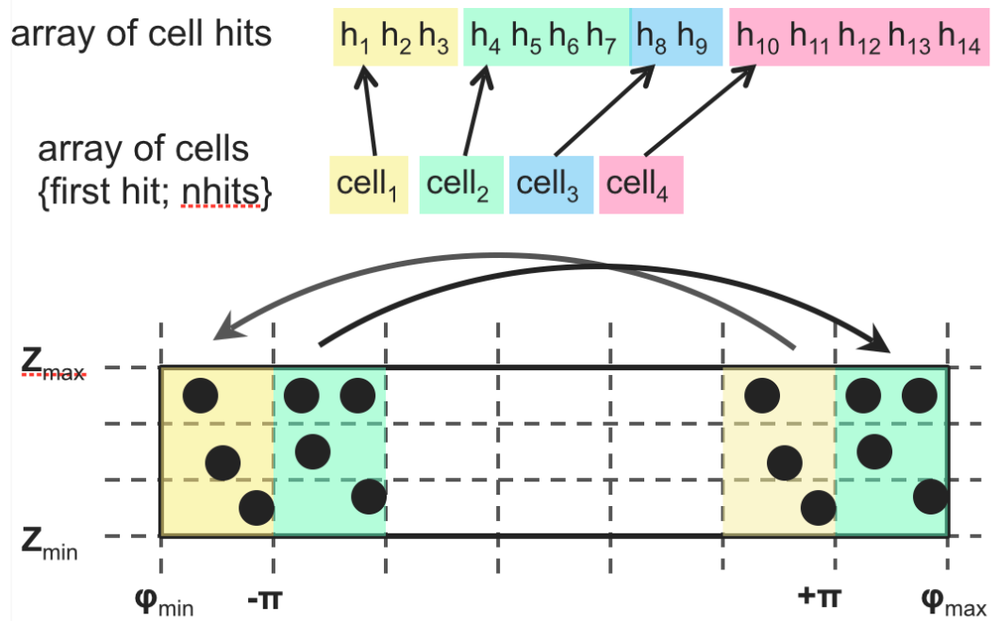
Phase 1 Sergey Gorbunov

Execution time
1.2 min on single core 2.6 GHz CPU

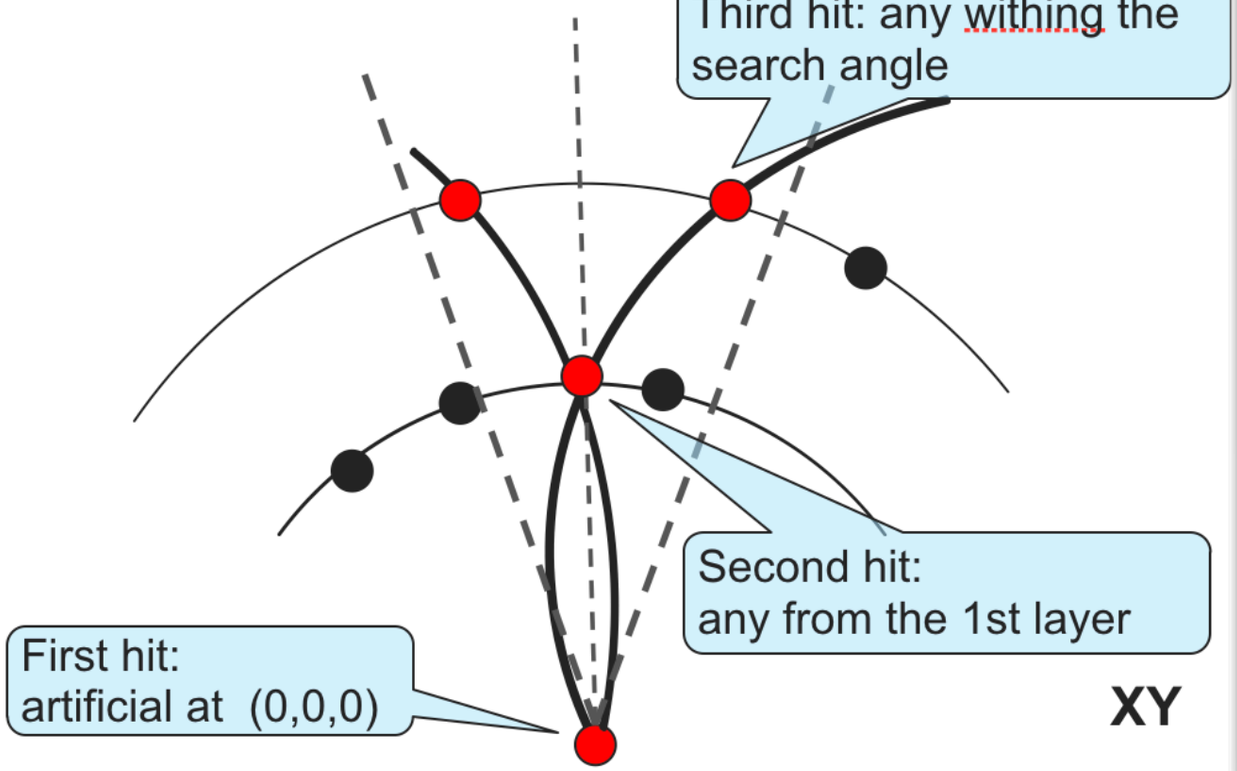
Summary

- A combinatorial algorithm, based on the track following method
- No search branches
- Simple track model: local 3-hit helix
- Fast data access

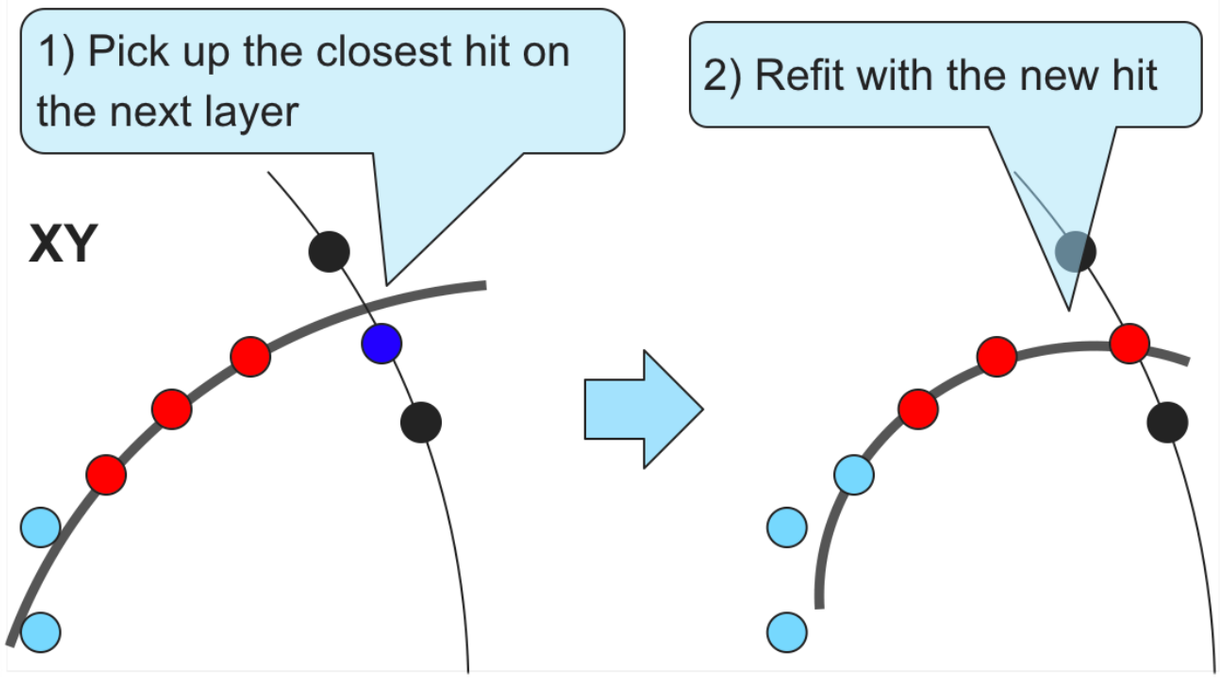
Regular grid with overlaps



Primary tracklets



Prolongation of tracklets



Phase 1 Special prices … to be announced

4	—	**demelian**	0.87079	35	2mo
5	—	**Edwin Steiner**	0.86395	5	2mo
					2mo
					2mo
					2mo
					2mo
					2mo
					4mo
15	—	**Vicens Gaitan**	0.70429	19	2mo
16	—	**Robert**	0.69955	3	2mo
17	—	**Yuval-CPMP tribute band**	0.69364	20	2mo

The TrackML **International Advisory Committee** composed of High Energy Physicists and Computer Scientists Markus Elsing, Frank Gaede, Alison Lowndes, Maurizio Pierini, Danilo Rezende and Marc Schoenauer (as detailed in https://sites.google.com/site/trackmlparticle/international-advisory-committee), has examined the contributions submitted.

Preliminary conclusion reached, recipients will be announced shortly.

Phase 1 Some lessons to be learned

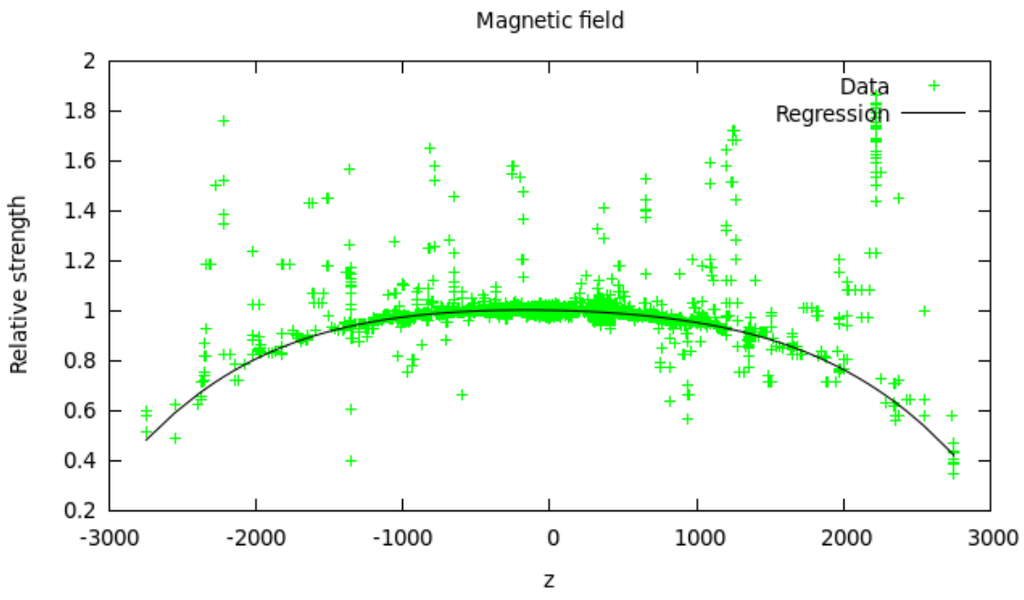
The threshold has been scary

- for many outside the field the simple size of the dataset was frightening
- even though there were many many teams !!!

Domain knowledge is important

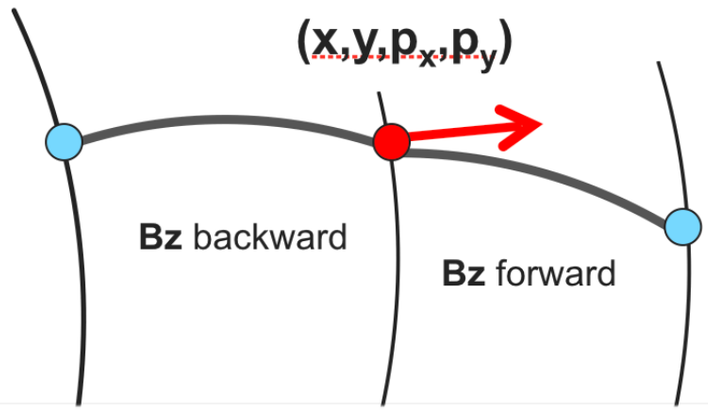
- put some physics helps :-)
- we did not give the magnetic field (on purpose)
- 2 out of three front-runners estimated the magnetic field

- Plot magnetic field strength



Fit of the magnetic field

- Use particle truth to estimate forward and backward field for each hit
- For each layer fit the field values with a poynom



Phase 1 Some lessons to be learned

The threshold has been scary

- for many outside the field the simple size of the dataset was frightening
- even though there were many many teams !!!

Domain knowledge is important

Background knowledge

- Very good slides for beginners
- Lecture of particles tracking
- Full helix equations for ATLAS - All equations you need!
- Diplom thesis of Andreas Salzburger (Wow, he started in this field as a CERN student already in 2001 :p)
- Doctor thesis of Andreas Salzburger
- CERN tracking software Acts - Sadly, we didn't have time to explore it :)

Andreas Salzburger **Competition Host** • just now • Options • Edit • Reply 0

Oh - you made me feel old now ... :-)

Thanks for participating and I hope you had fun in the challenge!!

Phase 2 Throughput

CodaLab

Phase 2 Dataset

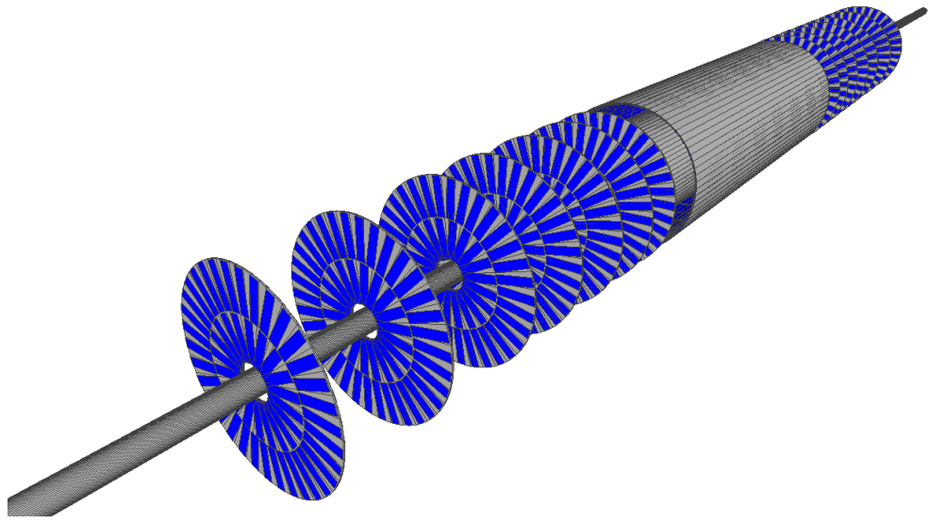
Detector remained unchanged

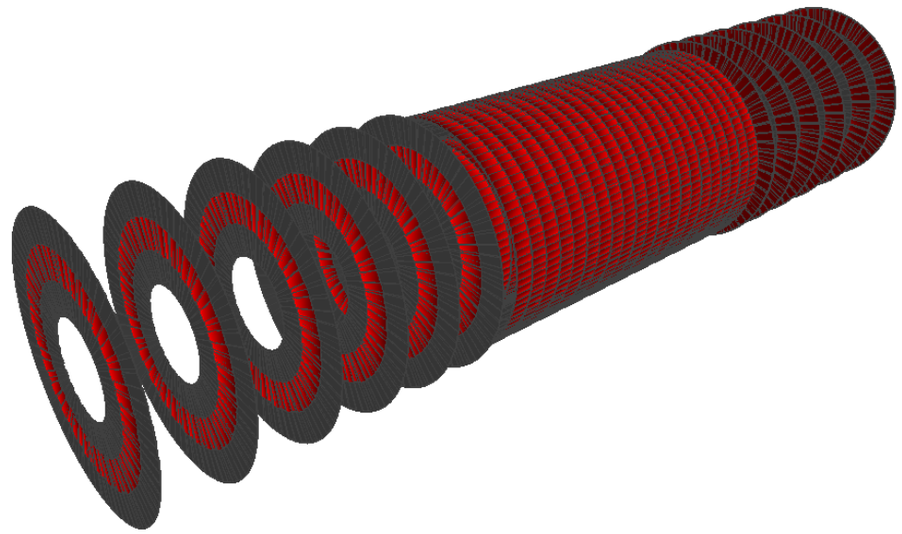
Dataset was slightly simplified

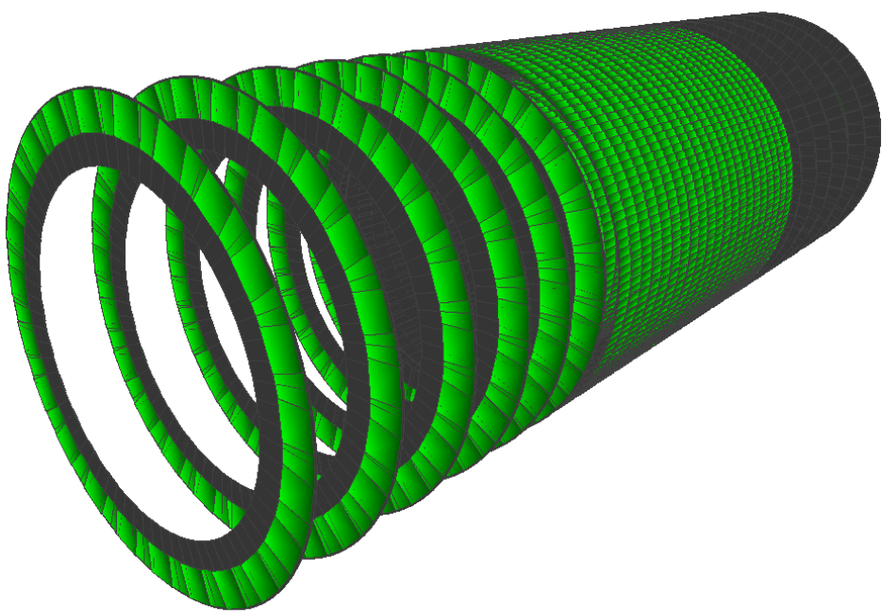
- only primary particles enter the scoring

Some “features” have been fixed

- module thickness is corrected (for cluster sizes)
- too narrow beam spot in the phase 1 (5.5 mm -> 5,5 cm)
- looping particles (present in phase 1) have been removed
- overshooting scattering for electrons (0.5 % effect in phase 1 dataset) has been fixed



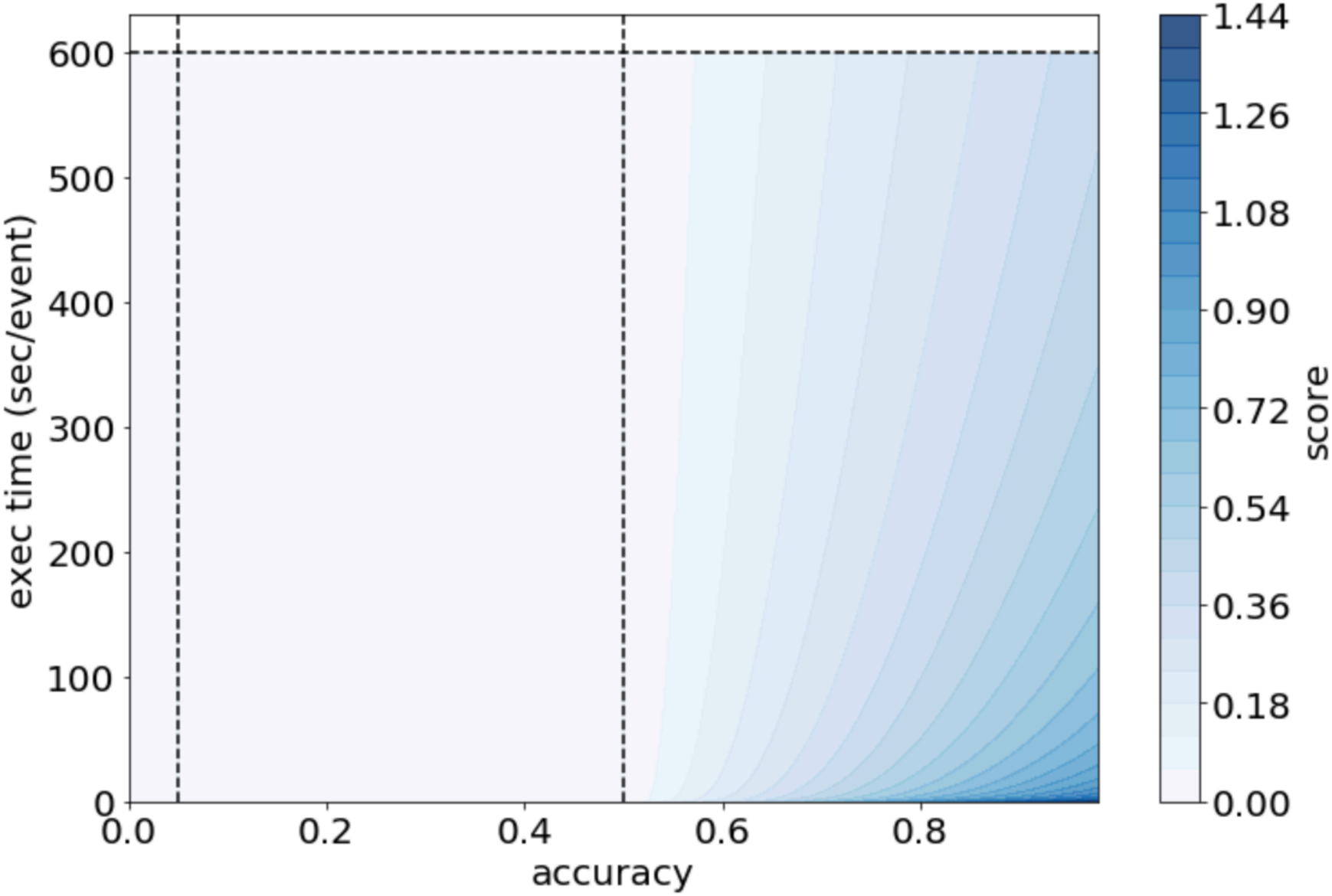




Phase 2 Scoring

Two-dimensional score folding accuracy & execution time

- needs a controlled environment for estimating the exec time robustly (special development done for codalab)

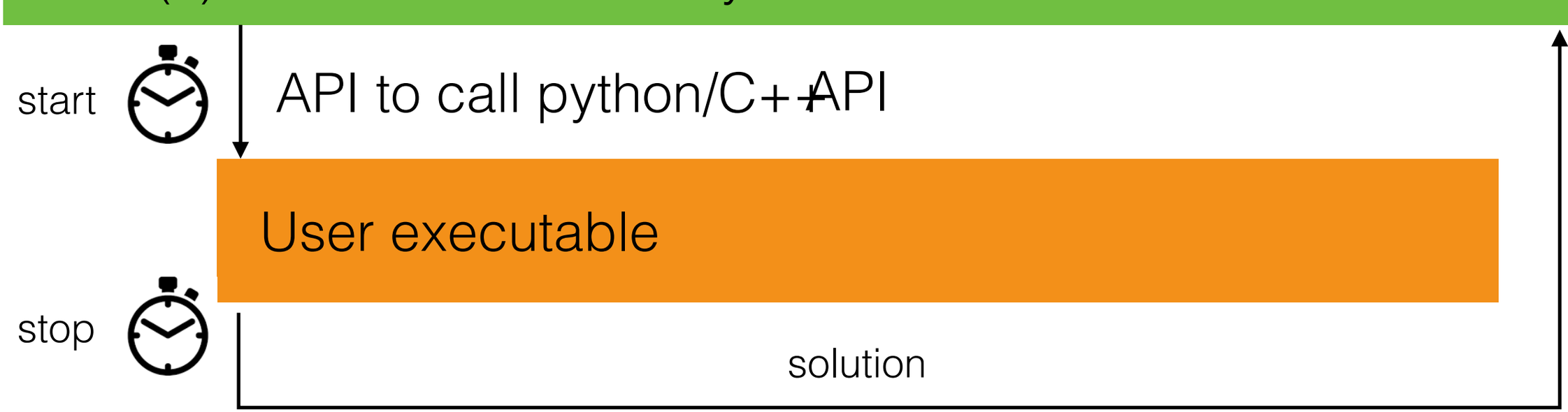


Phase 2 Setup - Controlling timing environment

CodaLab

	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

event(s) are loaded in memory



VM 2 cores, 4 Gb memory

Phase 2 Current submissions

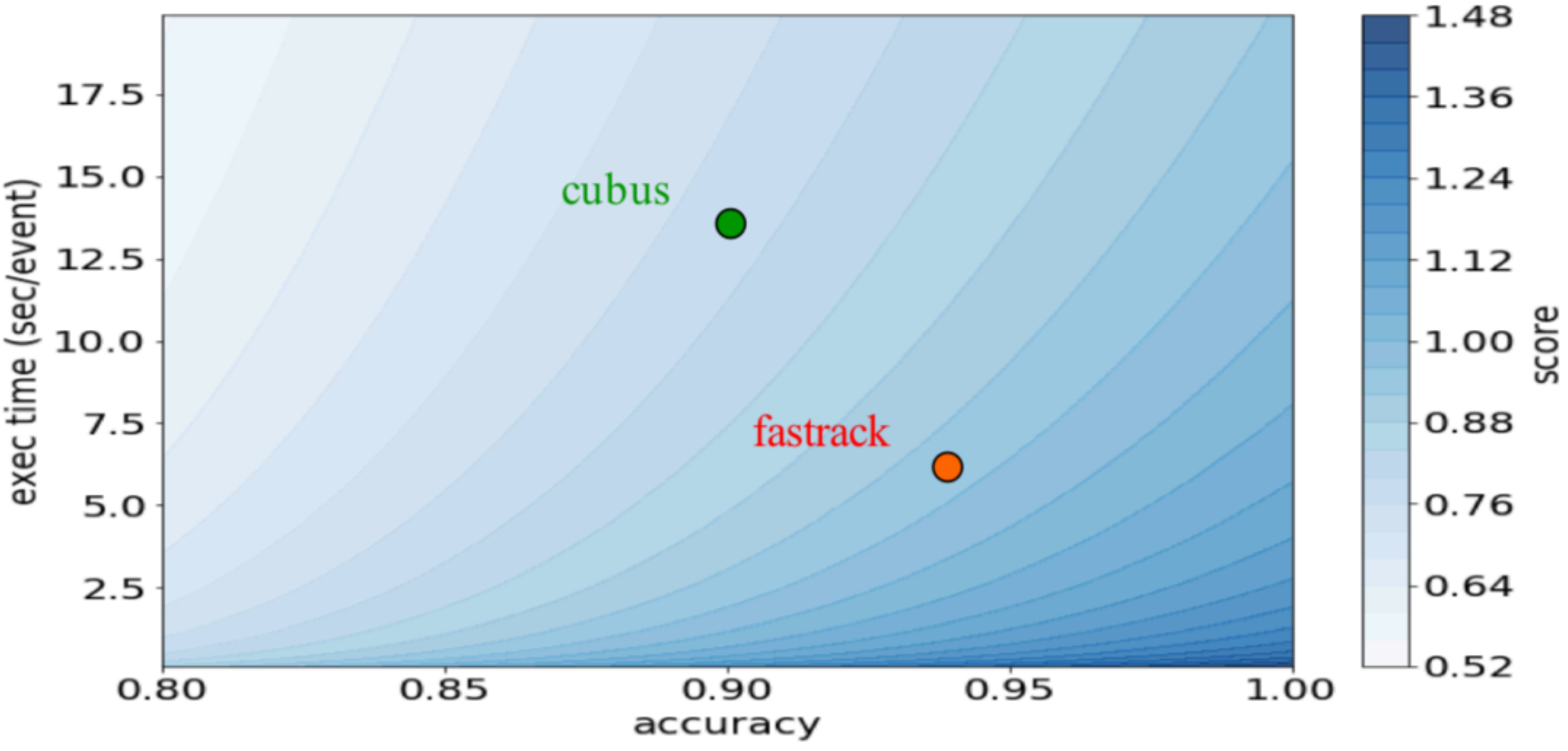
Participation has dropped dramatically compared to phase 1

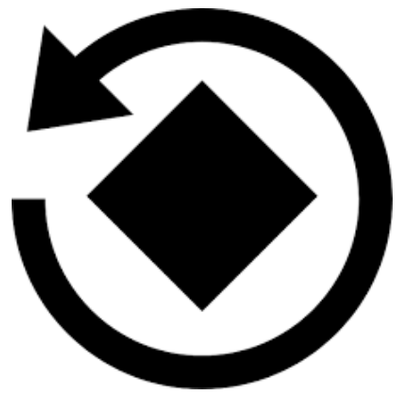
- currently only two active submitter
- a few more submissions but mainly with 0 score (starting kit submission)

RESULTS									
#	User	Entries	Date of Last Entry	score ▲	accuracy_mean ▲	accuracy_std ▲	computation time (sec) ▲	computation speed (sec/event) ▲	Duration ▲
1	fastrack	19	10/16/18	0.9328 (1)	0.94 (1)	0.00 (4)	321.02 (8)	6.42 (8)	362.00 (5)
2	cubus	8	09/13/18	0.7719 (2)	0.90 (2)	0.01 (3)	675.35 (9)	13.51 (9)	724.00 (6)
3	EdmonWales	1	10/14/18	0.0000 (3)	0.08 (4)	0.01 (2)	49.23 (4)	0.98 (4)	86.00 (3)
4	dcoldeira	1	10/13/18	0.0000 (3)	0.08 (4)	0.01 (2)	49.66 (7)	0.99 (7)	86.00 (3)
5	brunoseznec	1	10/08/18	0.0000 (3)	0.08 (4)	0.01 (2)	49.35 (5)	0.99 (5)	87.00 (4)
6	Taka	2	09/23/18	0.0000 (3)	0.08 (4)	0.01 (2)	48.13 (2)	0.96 (2)	84.00 (1)
7	mikhail94321	1	09/11/18	0.0000 (3)	0.18 (3)	0.02 (1)	4945.60 (10)	98.91 (10)	5080.00 (8)
8	droussea_naif	1	09/07/18	0.0000 (3)	0.08 (4)	0.01 (2)	48.21 (3)	0.96 (3)	85.00 (2)
9	Tester_91	1	09/07/18	0.0000 (3)	0.08 (4)	0.01 (2)	49.62 (6)	0.99 (6)	87.00 (4)
10	alexander_liao	12	10/04/18	-1.0000 (4)	0.00 (5)	0.00 (5)	-1.00 (1)	-0.02 (1)	3003.00 (7)

Phase 2 Current situation

The two submissions, however, are really fast





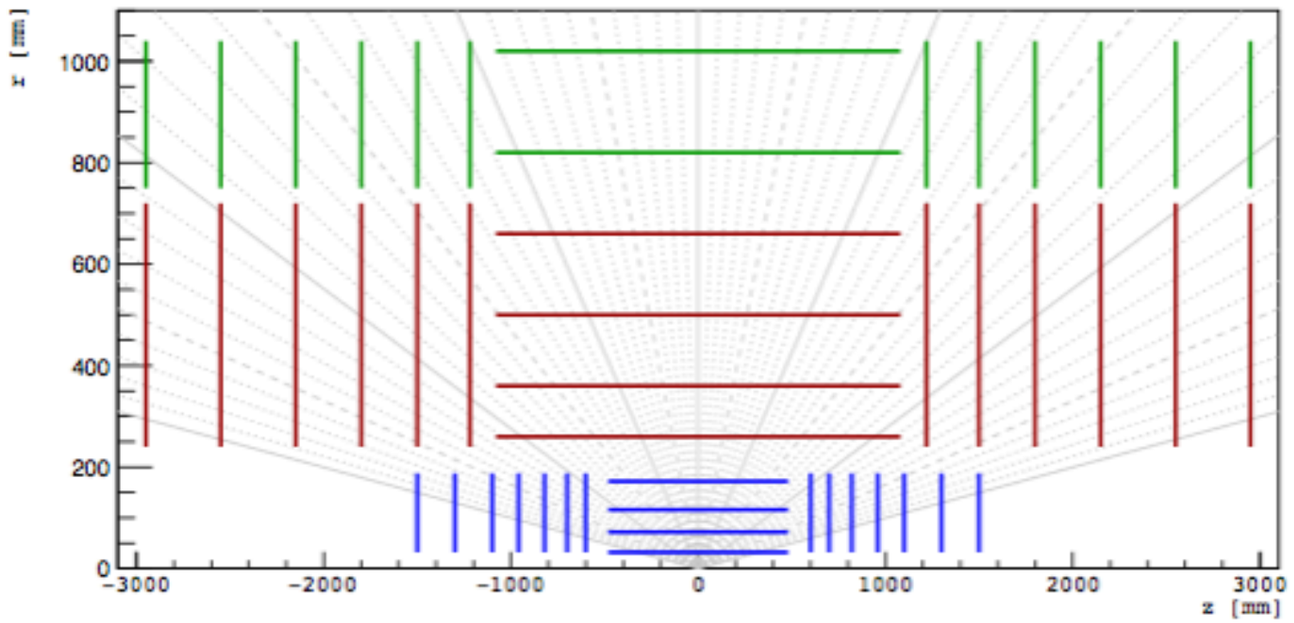
Spin-off Reference Dataset

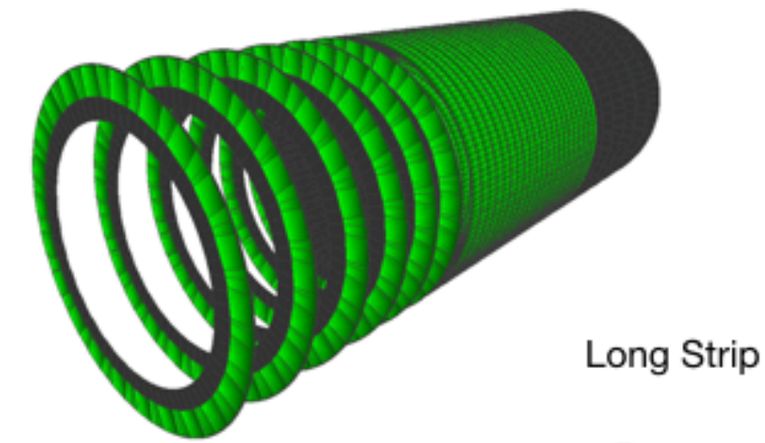
Spin-off

Reference detector&dataset

Common dataset for development within the community

- detector used for TrackML
- dataset produced with ACTS fast simulation
- **proposal**: iron out the few features we discovered & dataset (LHC/HL-LHC), publish on opendata CERN

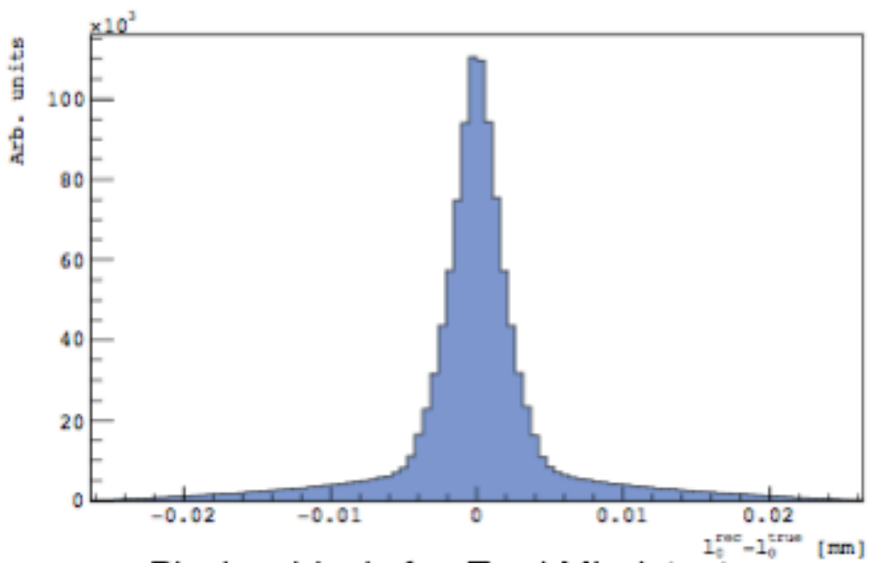




Spin-off

Reference detector&dataset

Quasi-realistic full silicon detector
- non-Gaussian measurements, with *realistic* cluster shapes
- *realistic* material budget
- main particle-material interactions



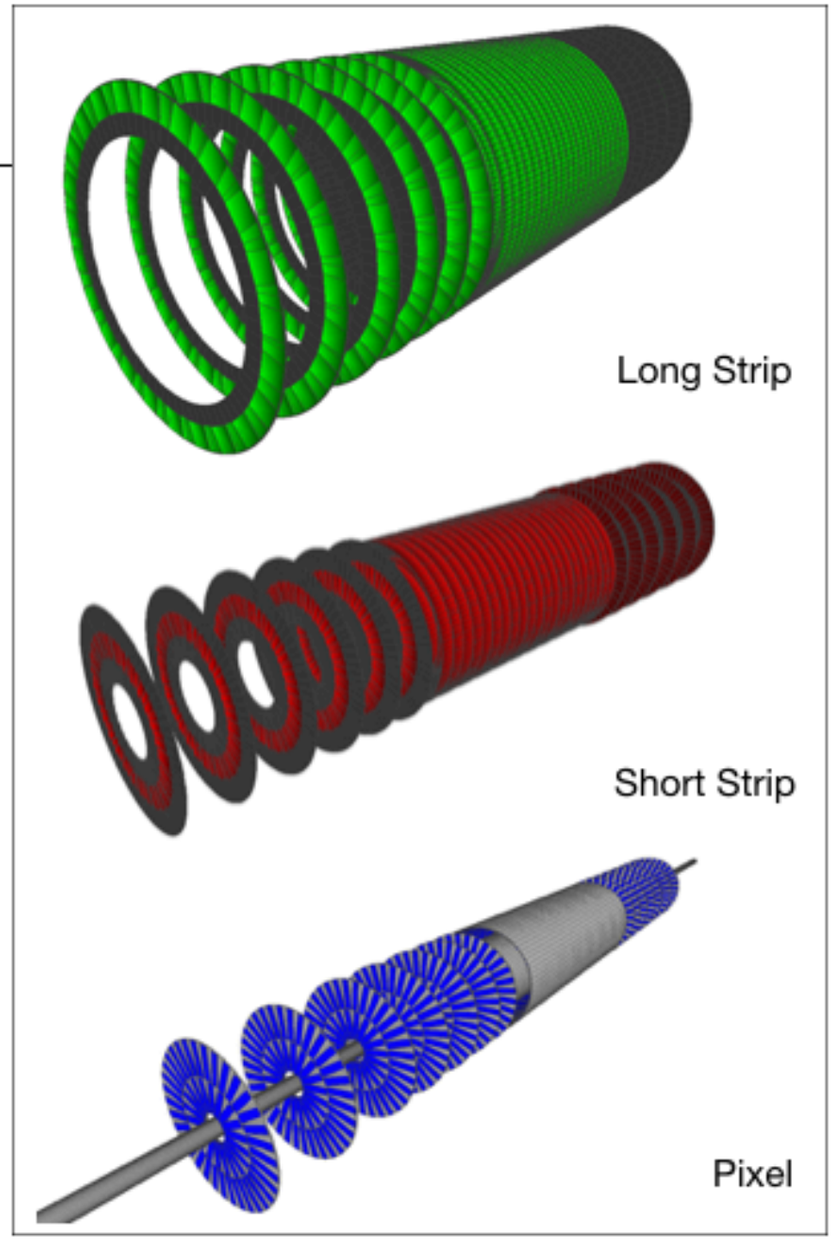
Pixel residuals for TrackML detector
(50 μm x 50 μm pixel size)

opendata CERN

development & testing ecosystem

Experiment SW ecosystem

publish



More Information

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

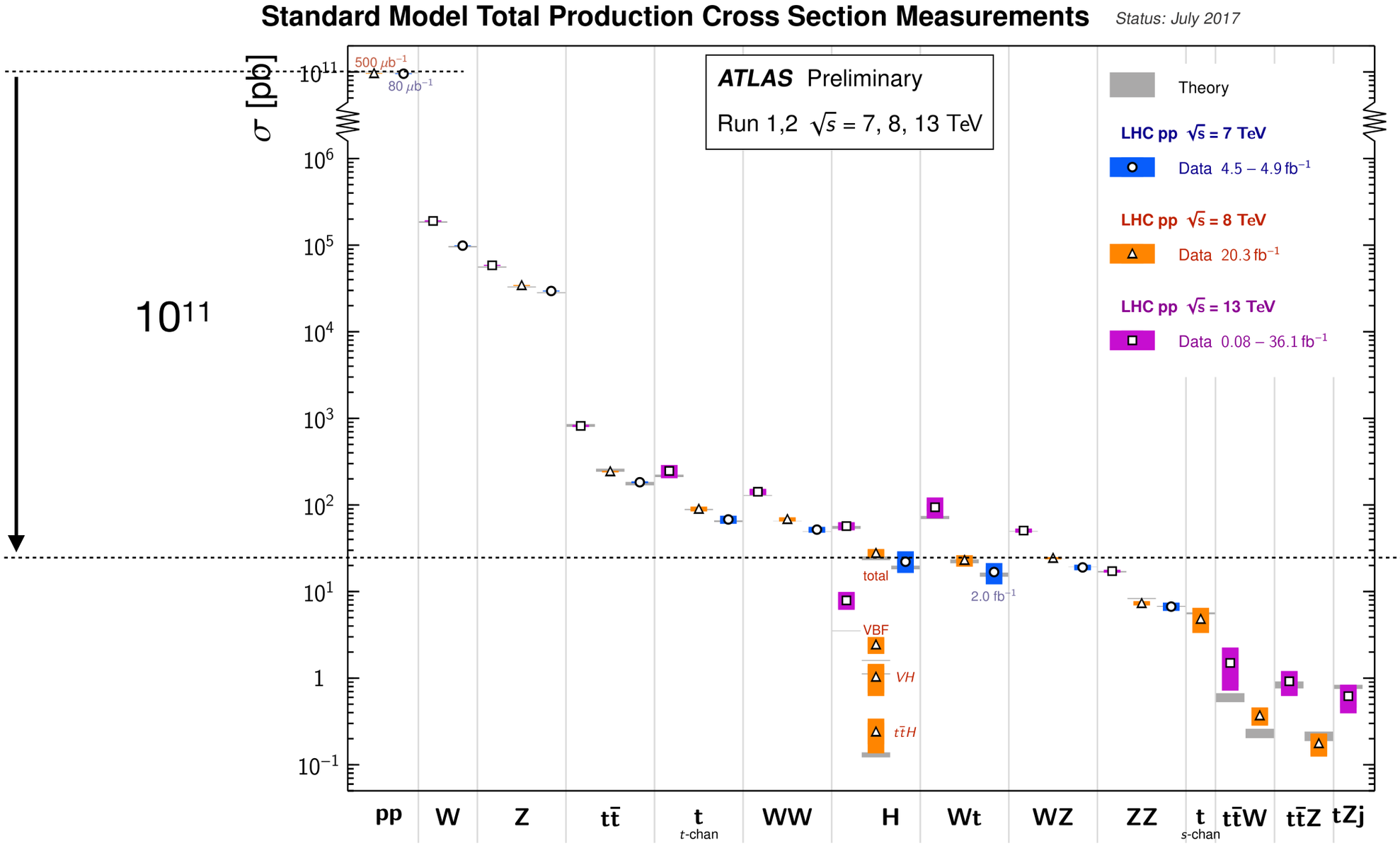
Phase 2 is running - you can still participate !!!

Backup slides

Introduction Physics

Focus on hadron colliders as the LHC

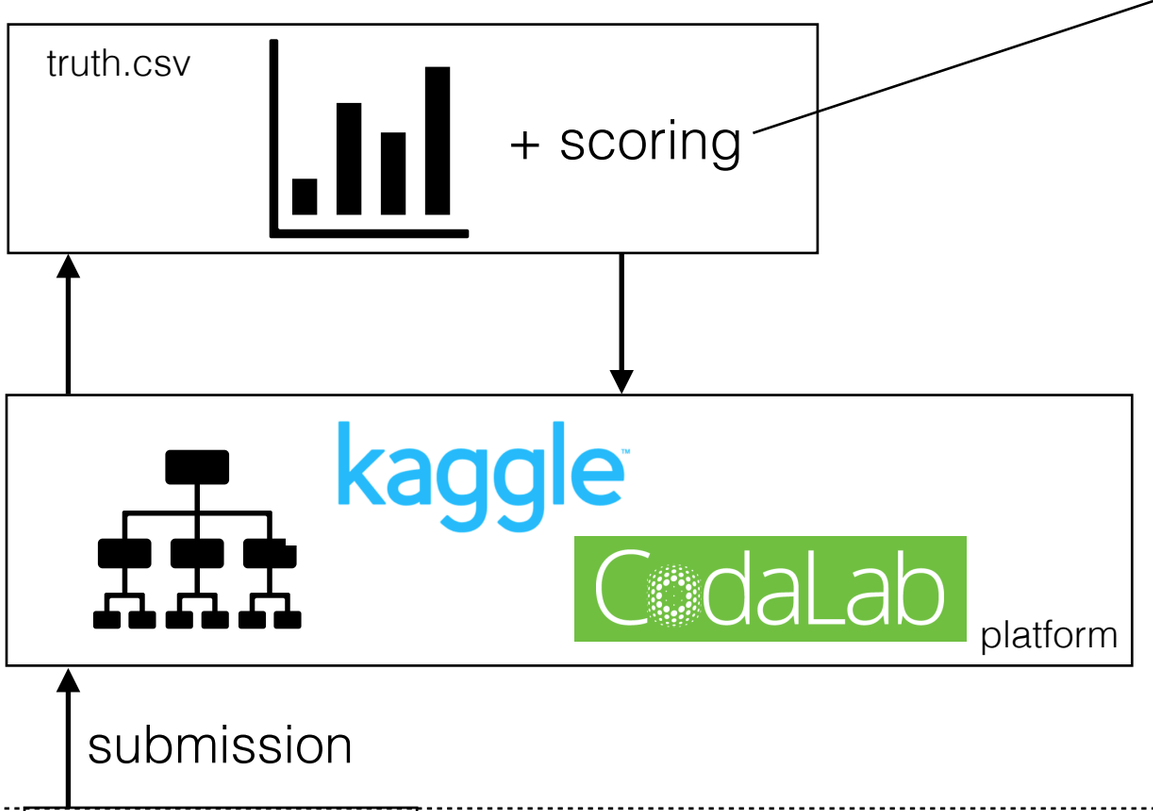
- High luminosity (HL-)LHC
- Future FCC-hh design study in preparation



10^{11}

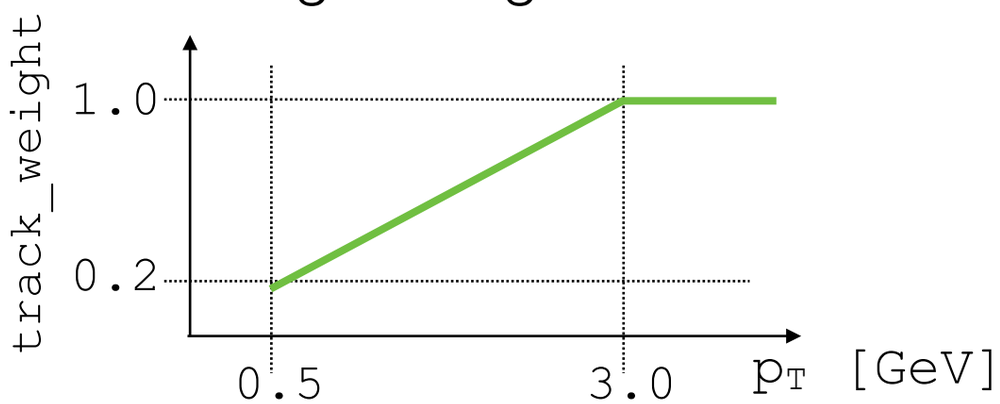
potentially new stuff to find

Submission & scoring (3)



$$\text{overall\_score} = \sum_{\text{events}} \sum_{\text{tracks}} \textbf{track\_weight} * \text{track\_score}$$

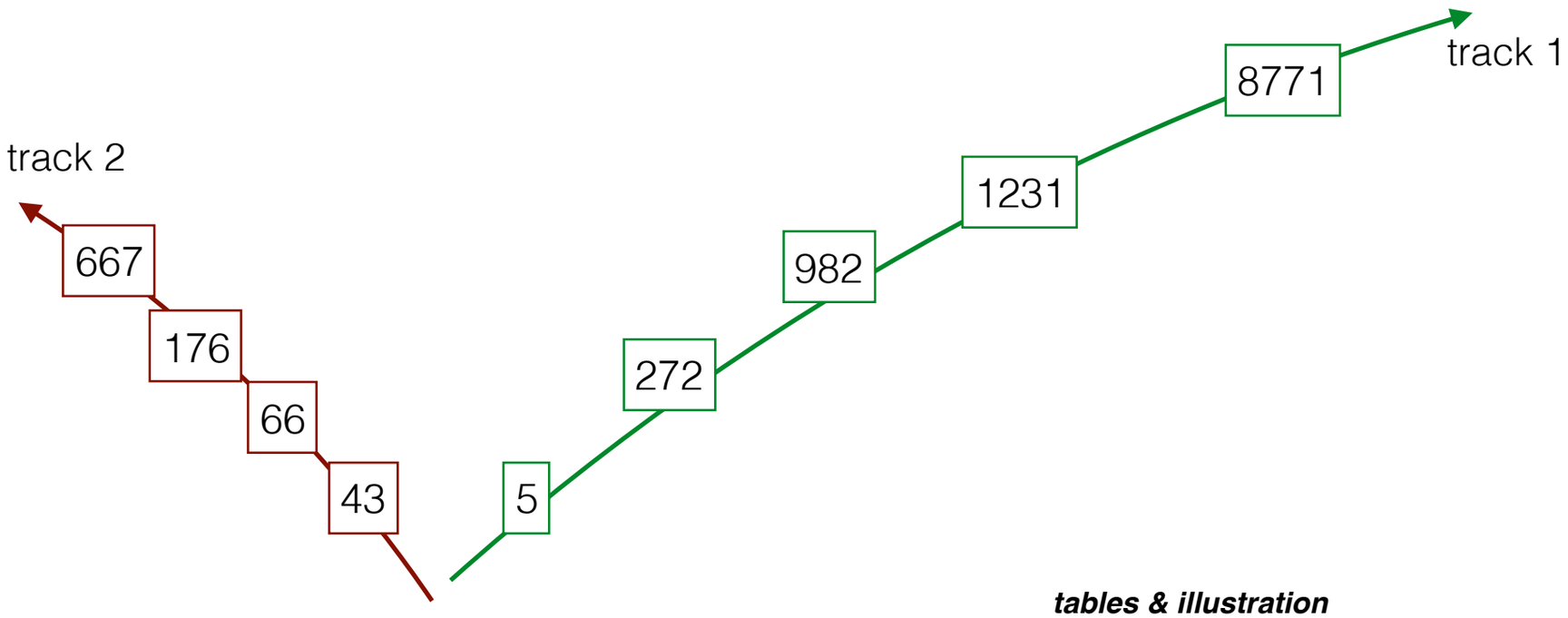
higher momentum
gives higher score:



solution.csv

hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

participant

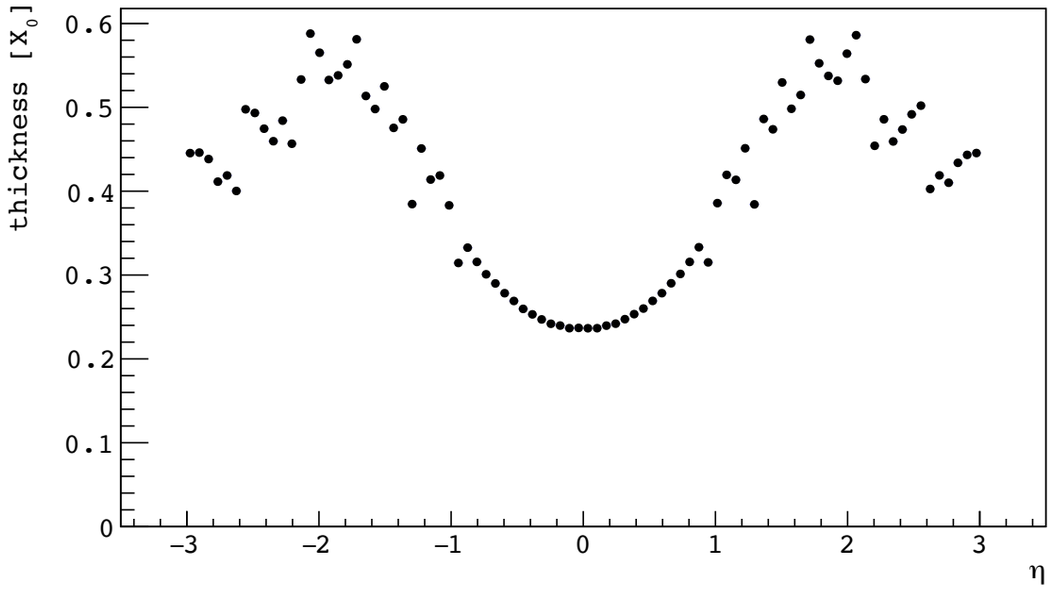


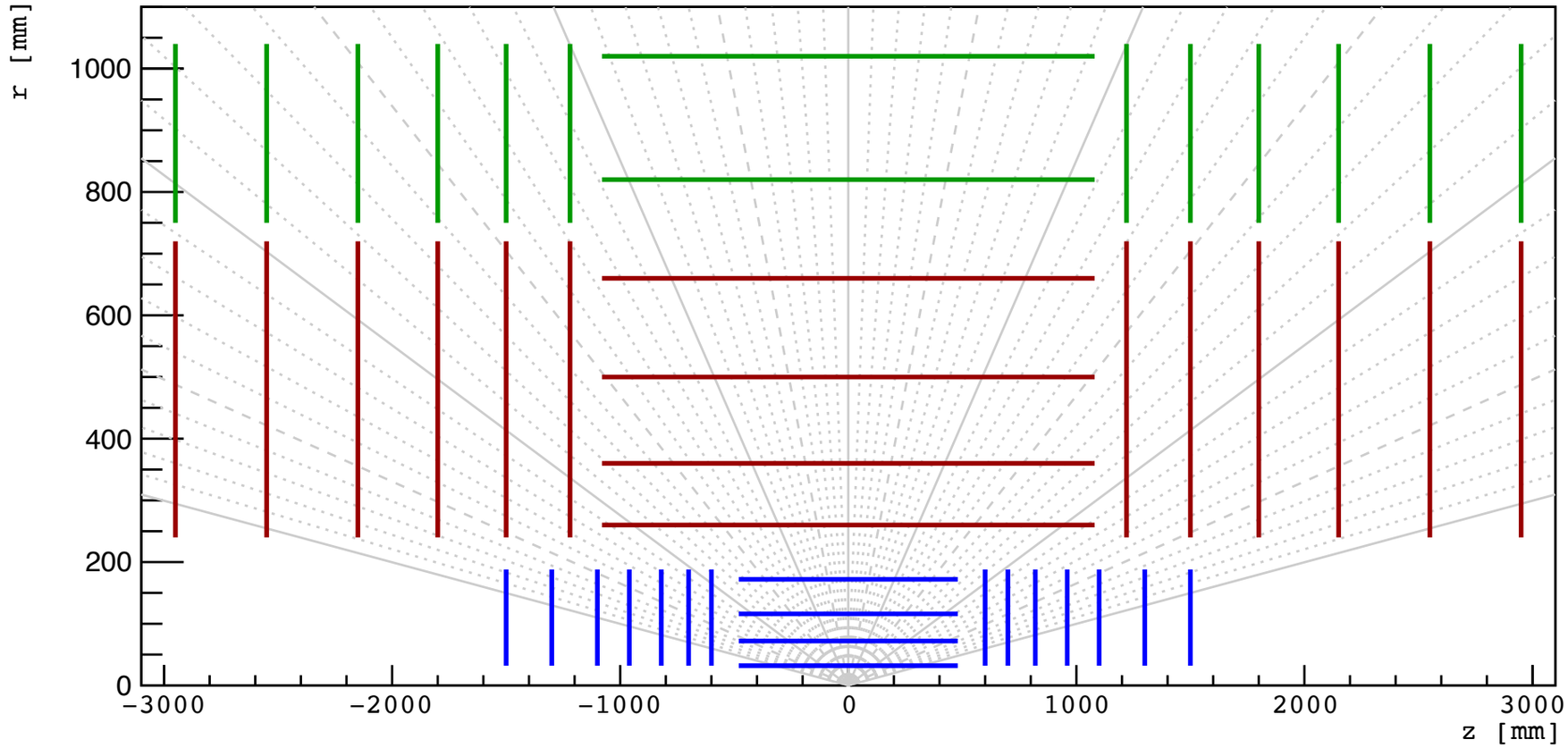
tables & illustration
(top) csv file format for validation hit dataset

The **detector**

Defined a Phase-2 like detector

- full silicon detector with realistic resolution, material budget, magnetic field
- composed as Pixel, short strip, long strip
- restricted to size of ~ ATLAS ID volume and |eta| < 3



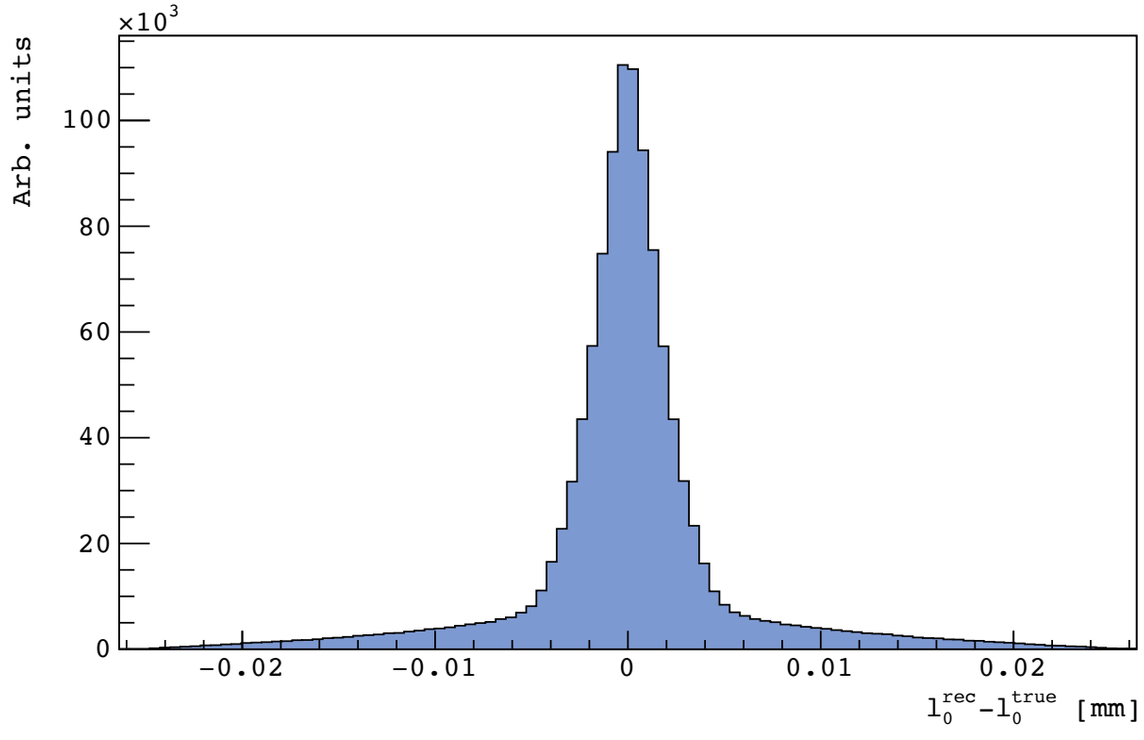


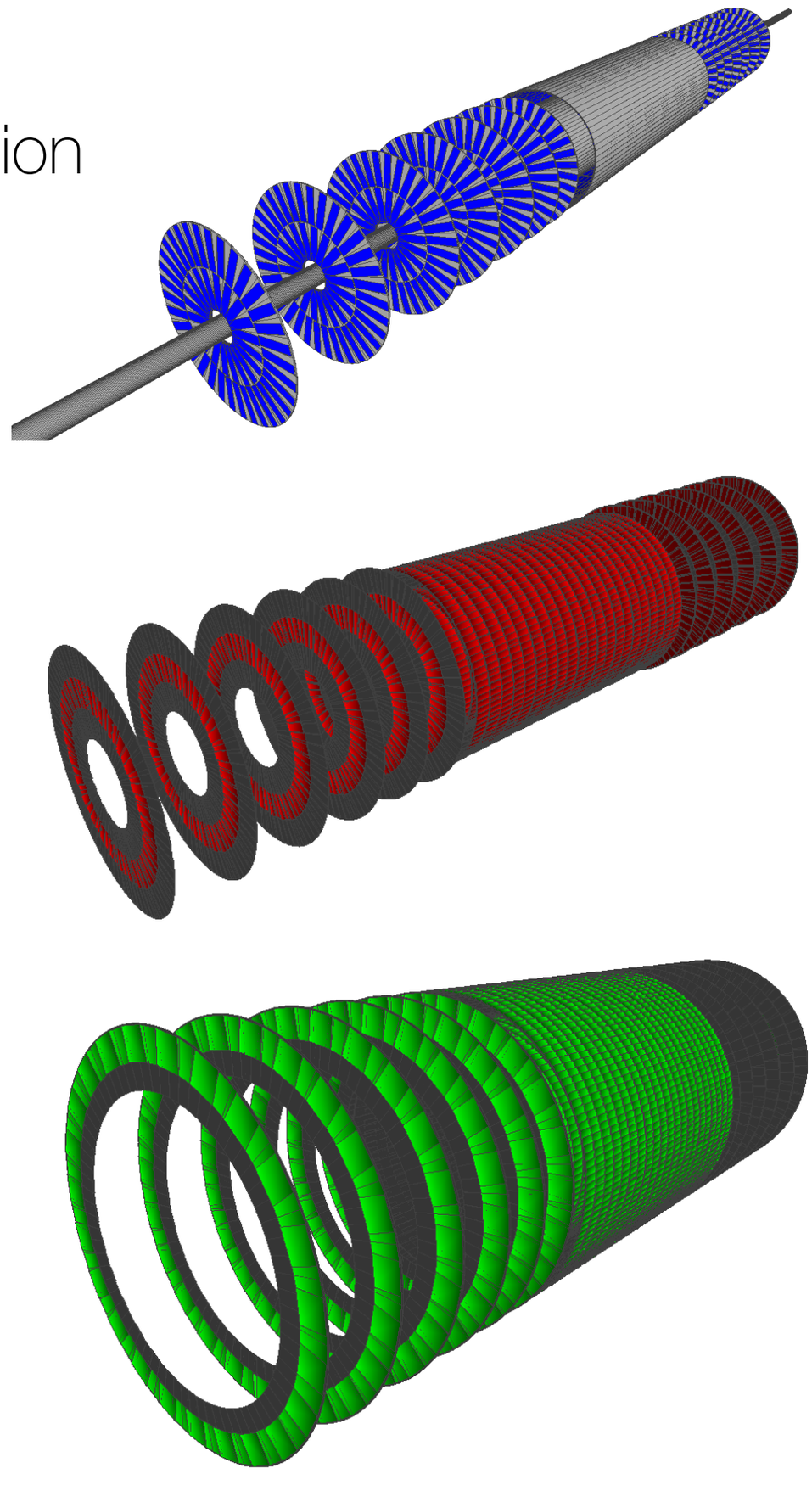
plot & image
(left) X0 distribution of the trackML detector
(right) longitudinal view of the trackML detector

The **detector** (2)

Dataset is simulation with ACTS fast simulation

- includes **multiple scattering, energy loss** and **hadronic interactions**
- includes **inefficiencies** and **noise/low momentum** particle hits
- includes pseudo-realistic **clustering model** (and hence resolutions)



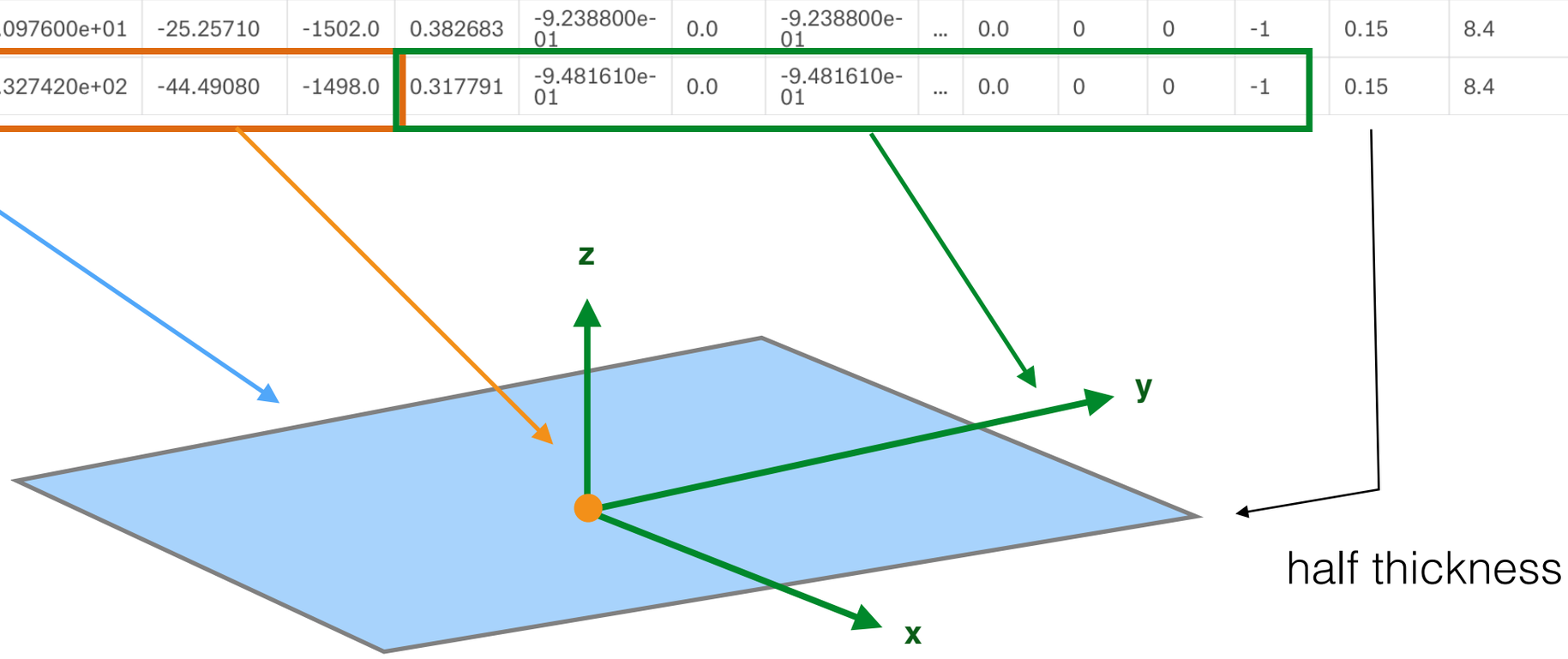


plot & images
(left) estimated pixel resolution distribution
(right) 3D view of pixel, short strip and long strip detector

The **detector** (3)

Detector description is given as .csv file

	volume_id	layer_id	module_id	cx	cy	cz	rot_xu	rot_xv	rot_xw	rot_yu	...	rot_yw	rot_zu	rot_zv	rot_zw	module_t	module_minhu	mod
0	7	2	1	-6.579650e+01	-5.17830	-1502.5	0.078459	-9.969170e-01	0.0	-9.969170e-01	...	0.0	0	0	-1	0.15	8.4	8.4
1	7	2	2	-1.398510e+02	-6.46568	-1502.0	0.046183	-9.989330e-01	0.0	-9.989330e-01	...	0.0	0	0	-1	0.15	8.4	8.4
2	7	2	3	-1.386570e+02	-19.34190	-1498.0	0.138156	-9.904100e-01	0.0	-9.904100e-01	...	0.0	0	0	-1	0.15	8.4	8.4
3	7	2	4	-6.417640e+01	-15.40740	-1498.0	0.233445	-9.723700e-01	0.0	-9.723700e-01	...	0.0	0	0	-1	0.15	8.4	8.4
4	7	2	5	-1.362810e+02	-32.05310	-1502.0	0.228951	-9.734380e-01	0.0	-9.734380e-01	...	0.0	0	0	-1	0.15	8.4	8.4
5	7	2	6	-6.097600e+01	-25.25710	-1502.0	0.382683	-9.238800e-01	0.0	-9.238800e-01	...	0.0	0	0	-1	0.15	8.4	8.4
6	7	2	7	-1.327420e+02	-44.49080	-1498.0	0.317791	-9.481610e-01	0.0	-9.481610e-01	...	0.0	0	0	-1	0.15	8.4	8.4



plot & image
(top) csv file format for the detector
(bottom) module center and orientation

The dataset - physics

Pythia configured with:

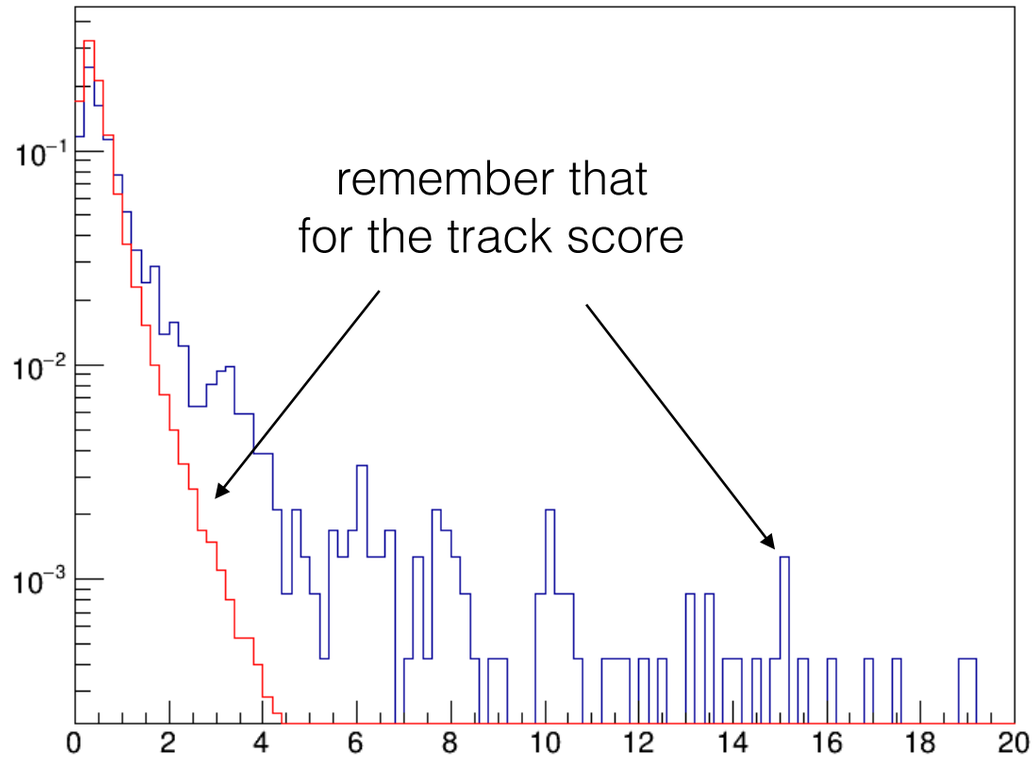
- HS: **“Top:gg2ttbar = on”**
- PU (@200): **“SoftQCD = on”**

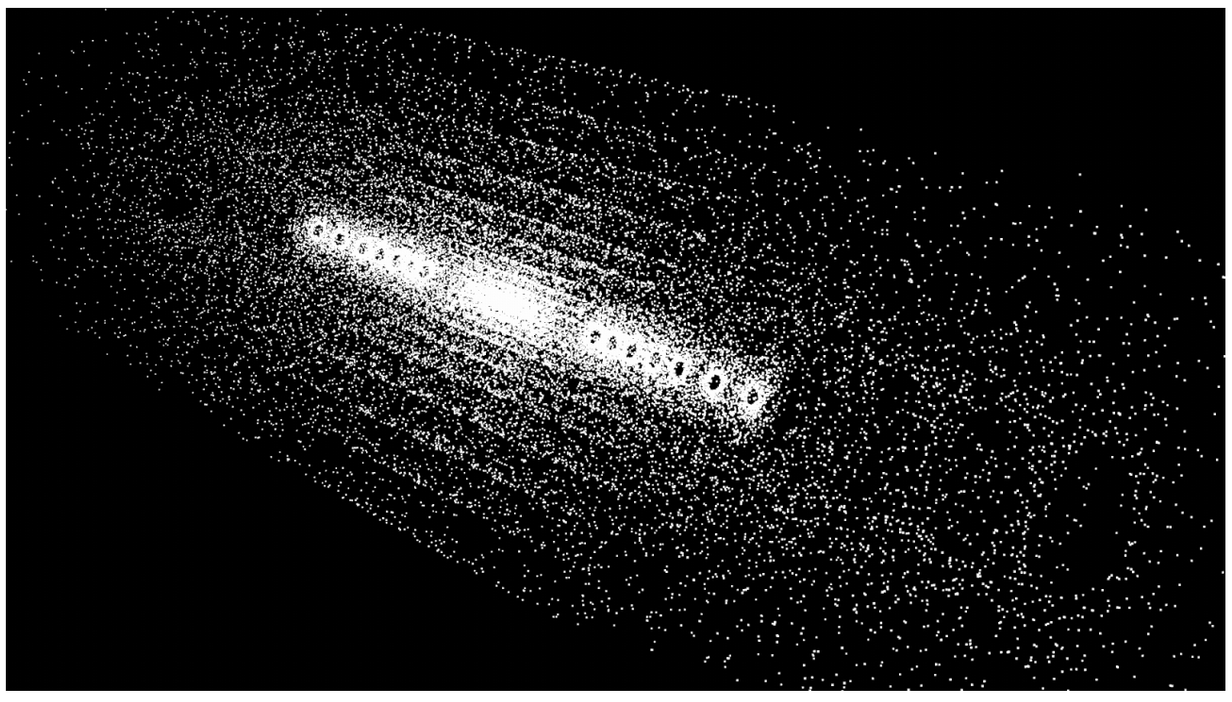
Smeared beam spot

- σ_Z = 5.5 mm, σ_T = 15 μm

Charged particles are simulated

- p_T > 150 MeV





large benchmark dataset (100s Gb)
to be released as CERN OpenData

plot & image
(top) transverse momentum distribution for hard scatter and pileup event
(bottom) hits produced in one single event

The training **dataset** - `eventXXXX-hits.csv`

	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

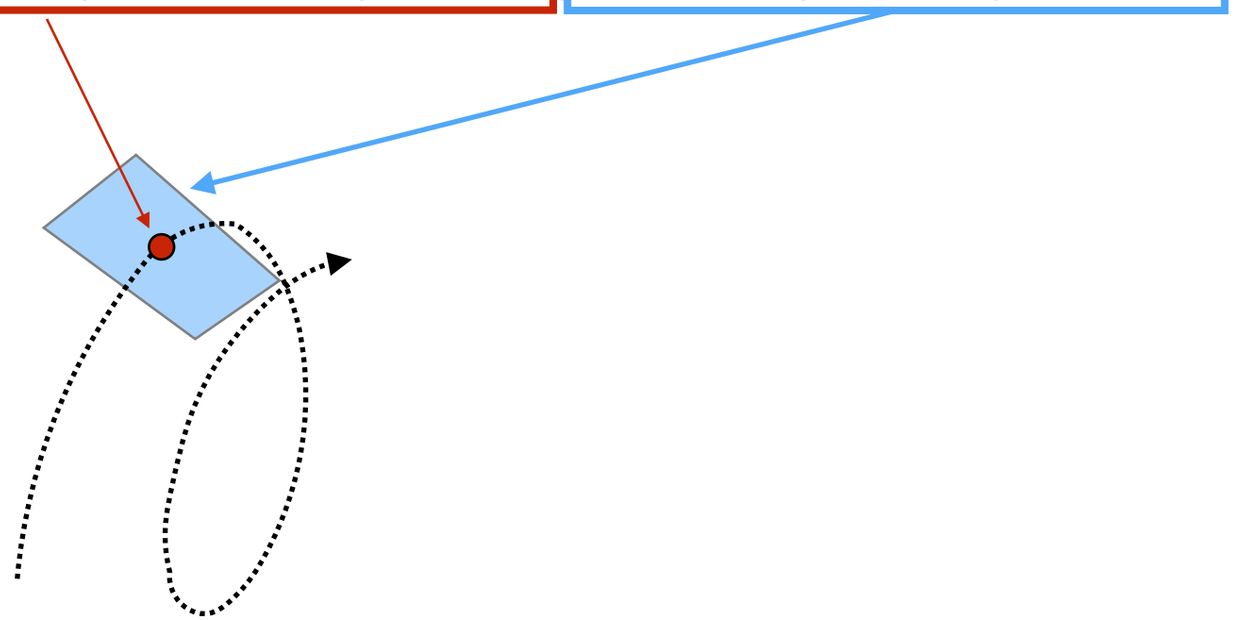


table & images
(top) csv file format for the hit file
(bottom) illustration of the hit information

The training **dataset** - `eventXXXX-cells.csv`

hits:

	hit_id	x	y	z	volume_id	layer_id	module_id
0	1	-64.409897	-7.163700	-1502.5	7	2	1

and cells:

link

	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

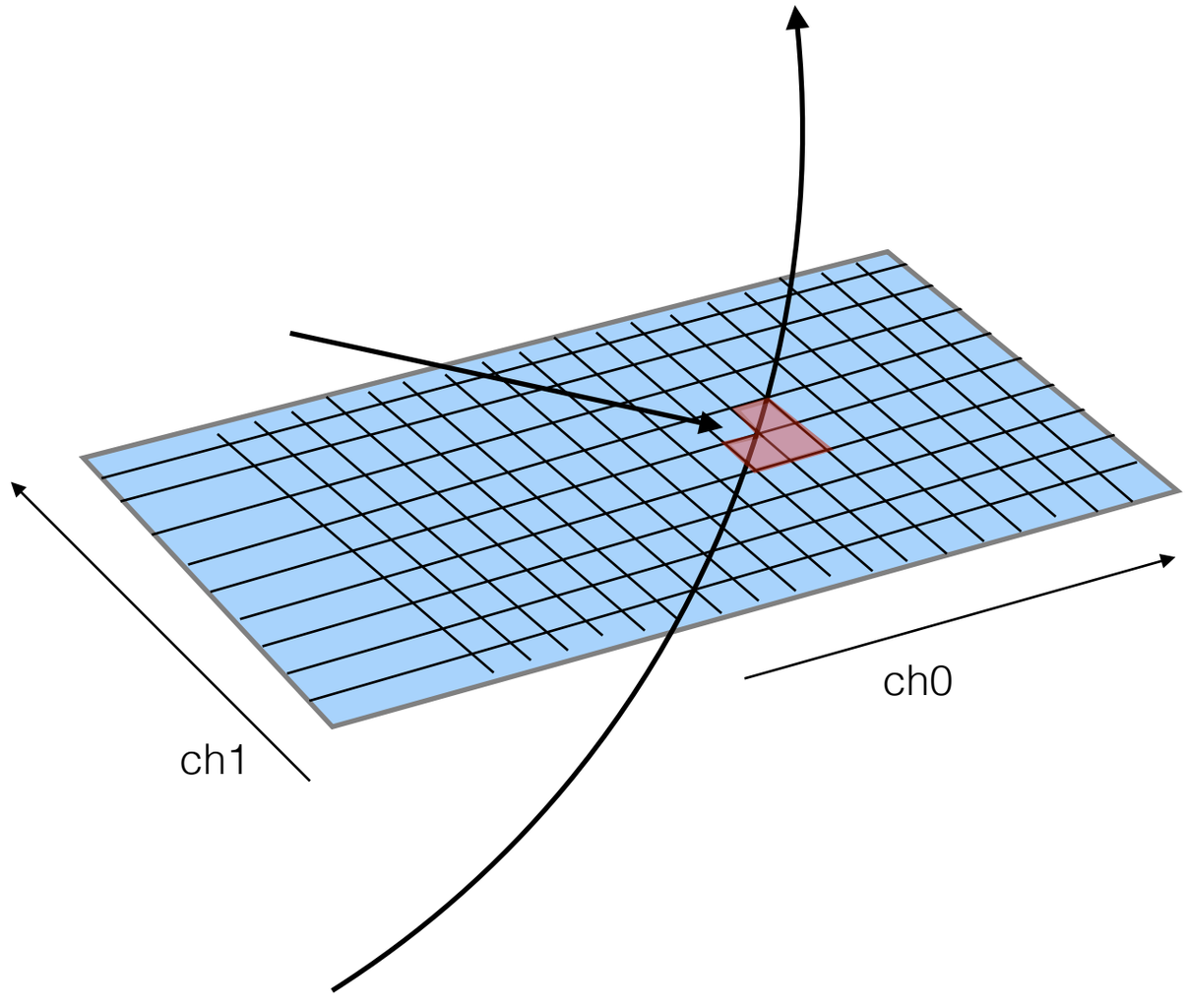


table & images
(top) csv file format for the hit file
(bottom left) csv file format of the cells information
(bottom right) cell information illustration

The training **dataset** - `eventXXXX-truth.csv`

hits:

	hit_id	x	y	z	volume_id
0	1	-64.409897	-7.163700	-1502.5	7
1	2	-55.336102	0.635342	-1502.5	7

reconstructed hit position

link

truth position/true momentum

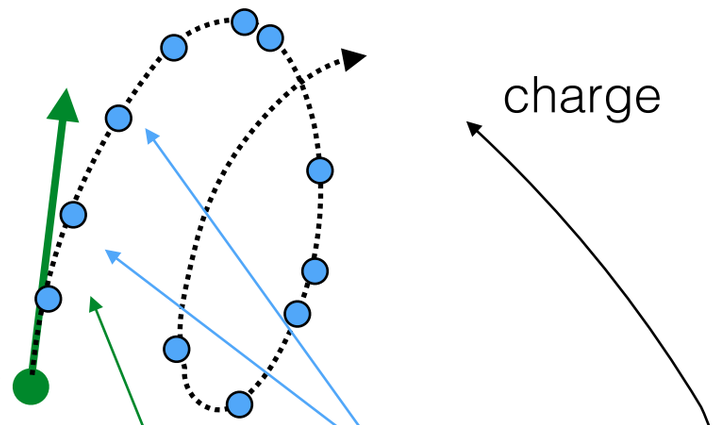
	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000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

noise hit
with 0 weight

hit weight
for scoring (see later)

tables
(top) csv file format for the hit file
(bottom) csv file format for the truth file

The training **dataset** - `eventXXXX-particles.csv`



charge

	particle_id	vx	vy	vz	px	py	pz	q	nhits
520	225257634377236 48	-0.015802	0.006381	1.16279	-0.56967	-0.011187	-15.496	1	10

link

	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	225257634377236 48	-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

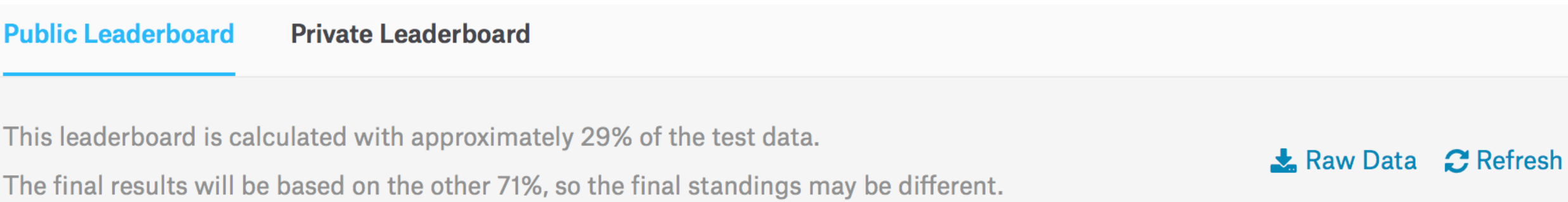
noise hit
with 0 weight

hit weight
for scoring (see later)

tables
(top) csv file format for the particle file
(bottom) csv file format for the truth file

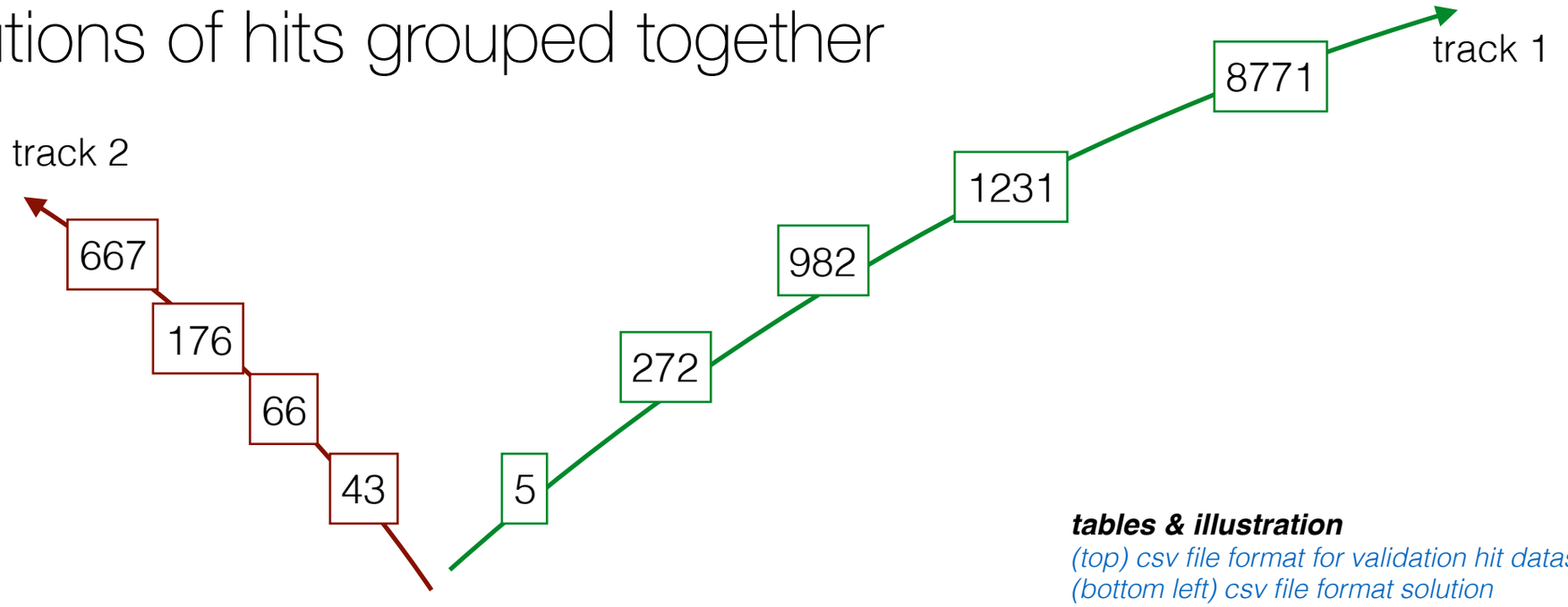
The validation **dataset & solution**

Independent but structurally identical hit dataset



We look for solutions of hits grouped together

hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2



tables & illustration
(top) csv file format for validation hit dataset
(bottom left) csv file format solution
(bottom right) track representation of solutions

Submission & scoring (2)

truth.csv

+ scoring

missing hits
reduce the **track score accordingly**

kaggle™

platform

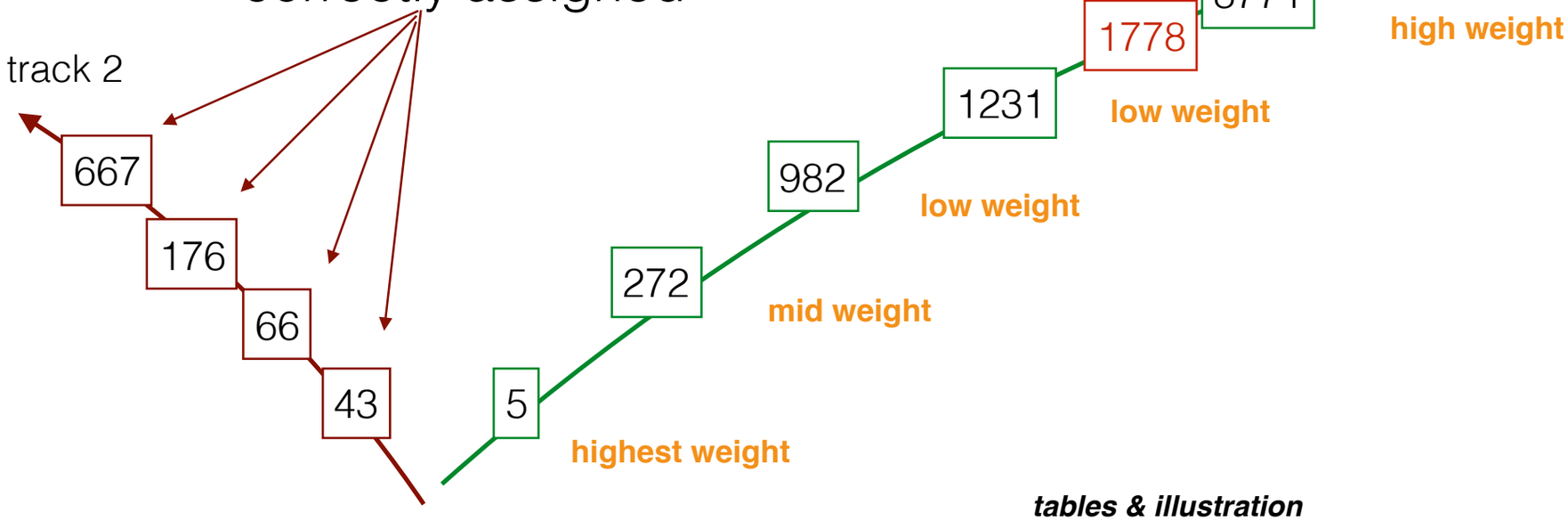
submission

solution.csv

hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

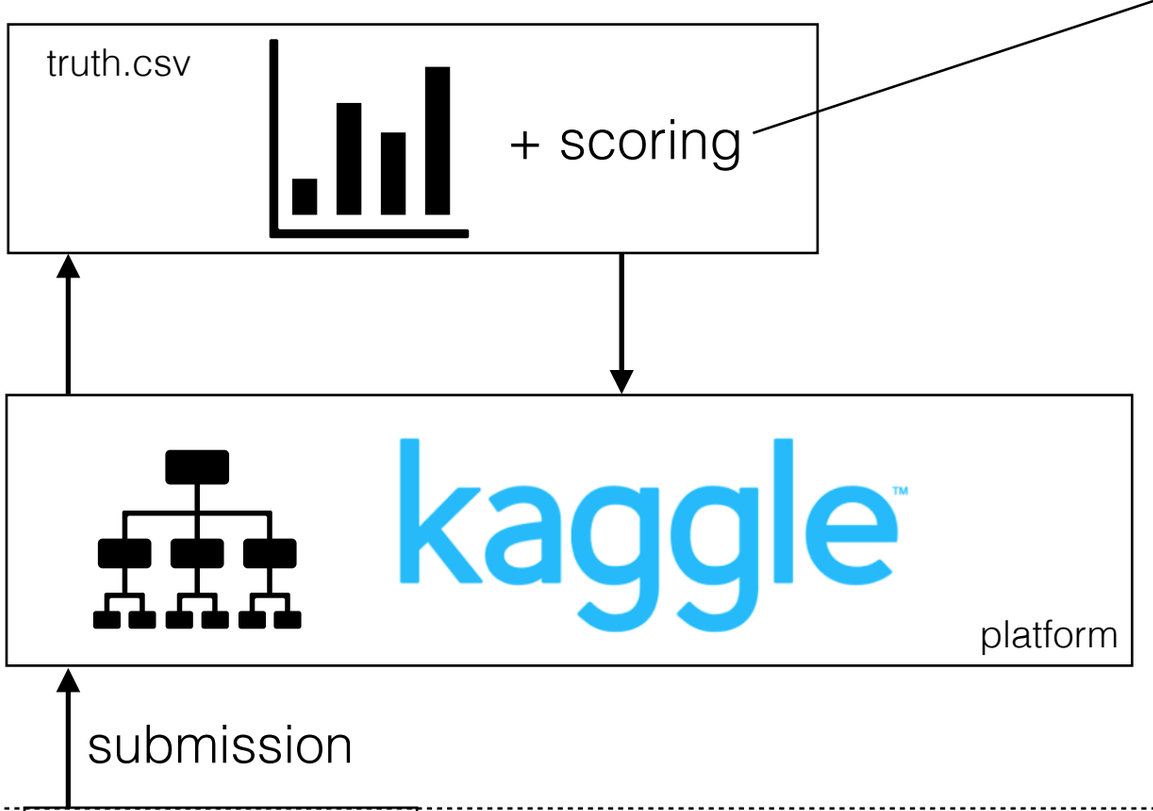
participant

garbage tracks will reduce overall event score, as hits will not be correctly assigned



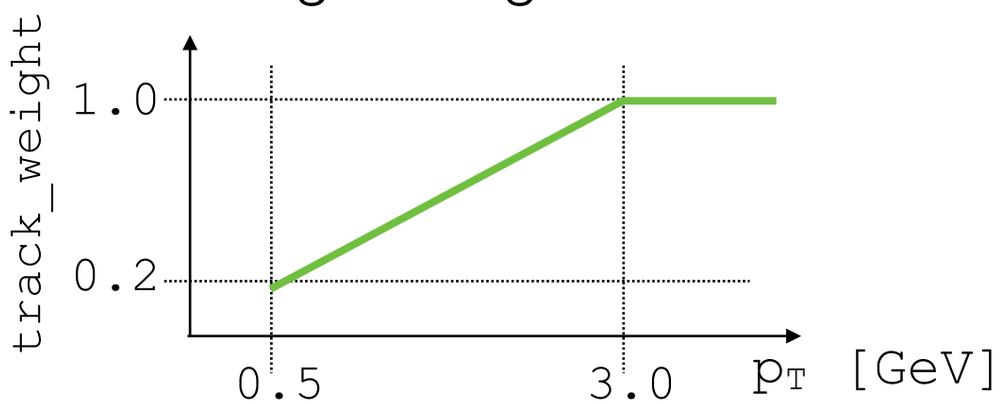
tables & illustration
(top) csv file format for validation hit dataset

Submission & scoring (3)



$$\text{overall\_score} = \sum_{\text{events}} \sum_{\text{tracks}} \texttt{track\_weight} * \texttt{track\_score}$$

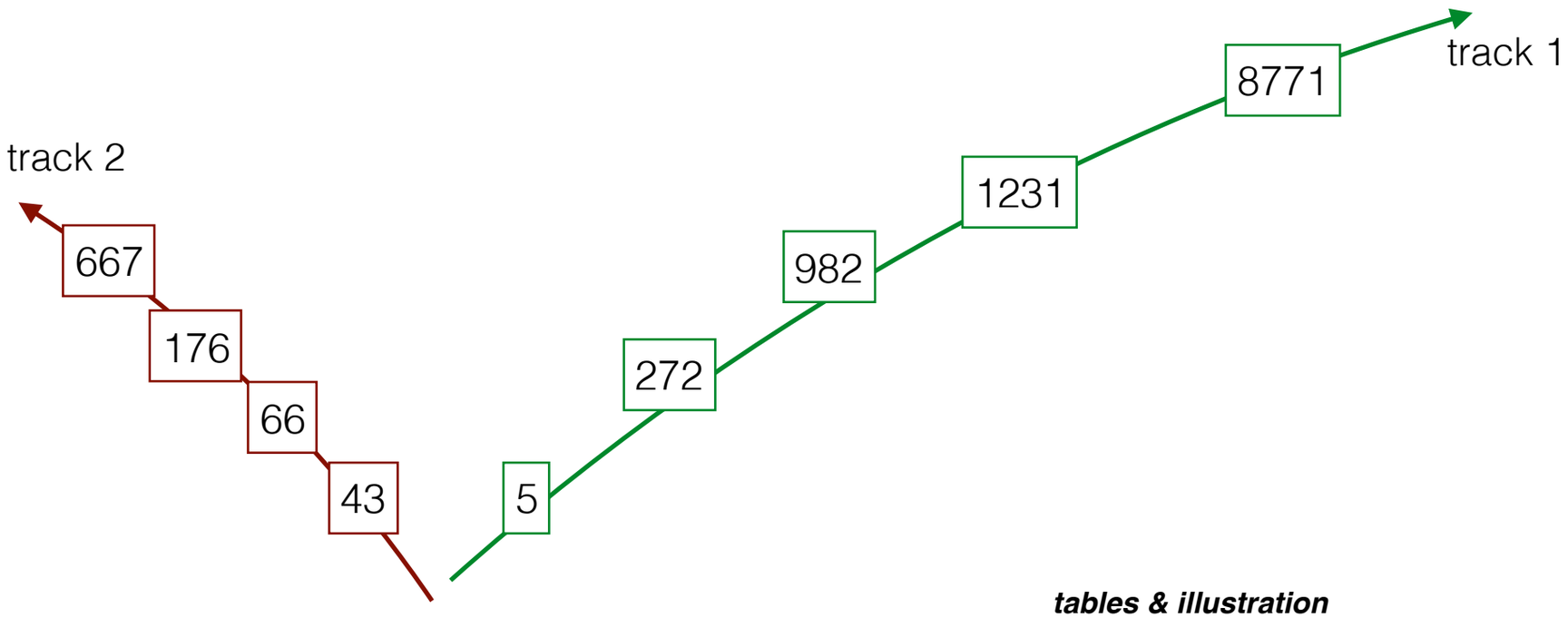
higher momentum
gives higher score:



solution.csv

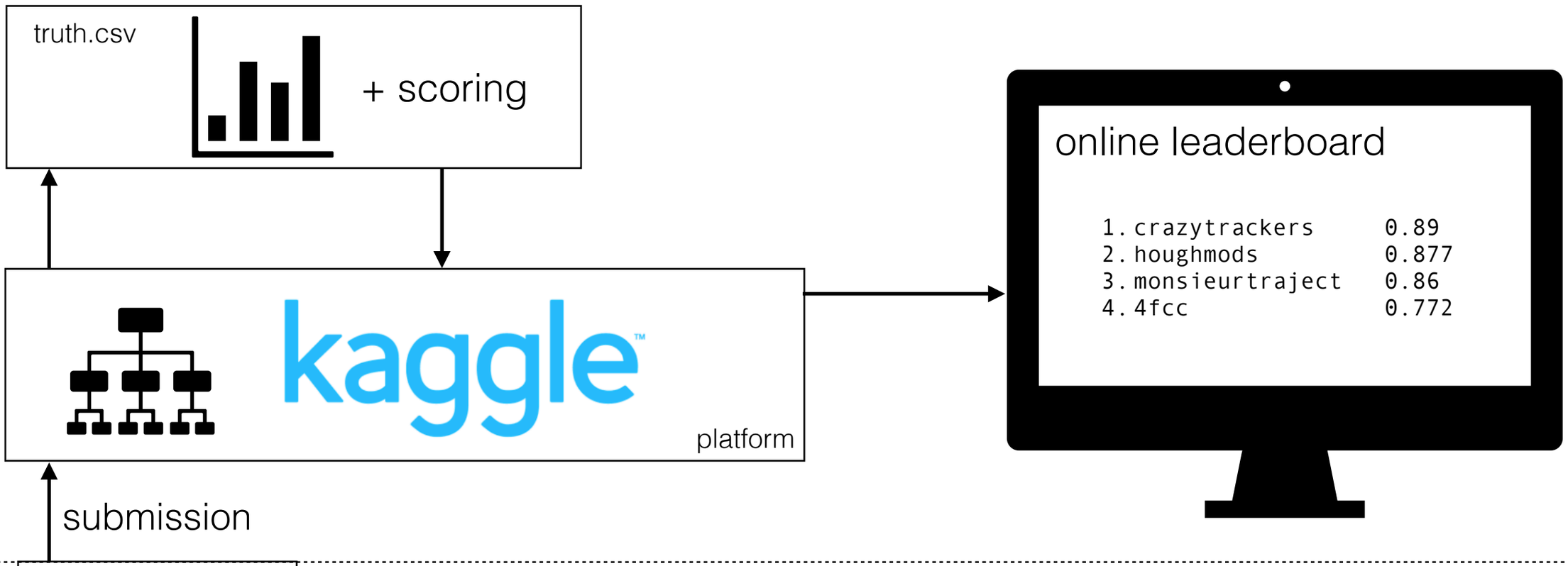
hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

participant



tables & illustration
(top) csv file format for validation hit dataset

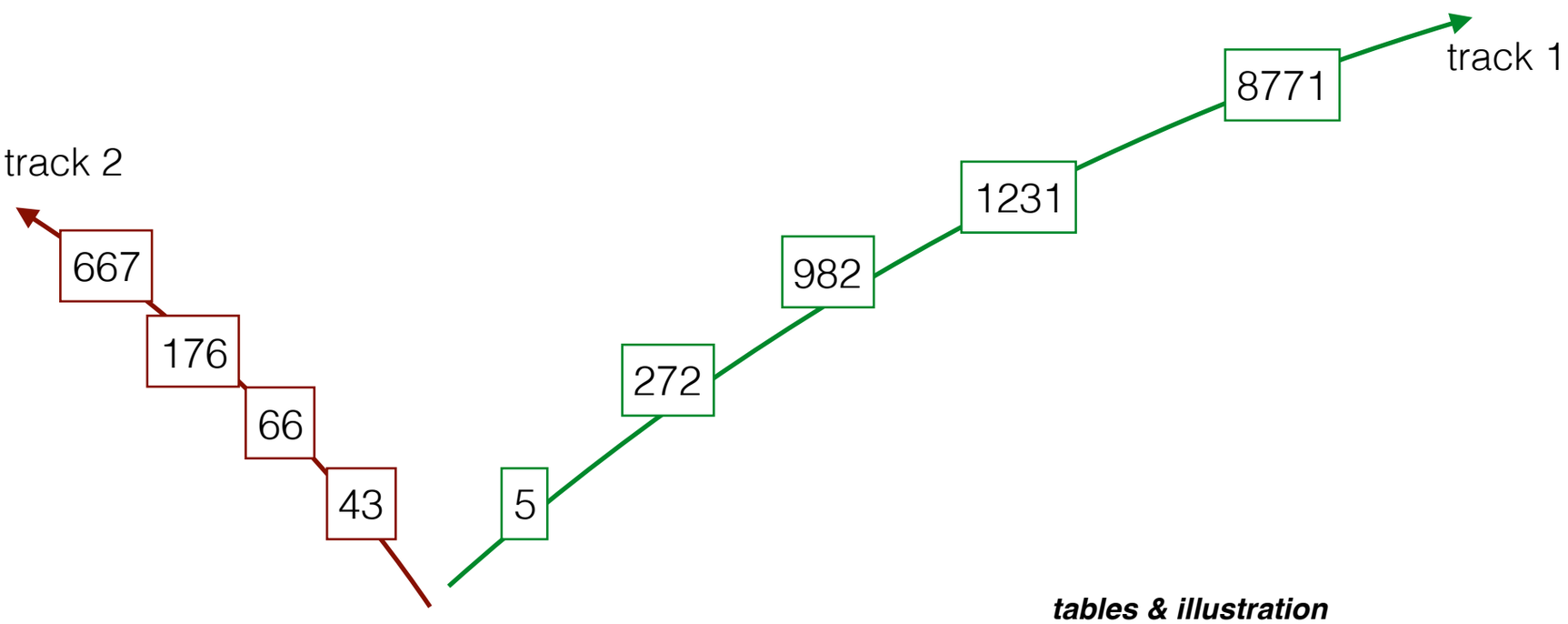
Submission & scoring (4)



solution.csv

hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

participant



tables & illustration

(top) csv file format for validation hit dataset